A

Aboriginal Rights

1. *Aboriginal Rights*

Aboriginal rights commonly are understood to be the rights of the original peoples of a region that continue to exist notwithstanding the imposition of power over them by other peoples. The term came into common usage in the 1970s, but is likely to be superseded by the term 'Indigenous rights' which gained precedence in the 1990s. Here, the two terms are considered synonyms. While seemingly straightforward, the definition of Aboriginal rights is complex and contested. This article discusses the origin and development of the concept as well as the term, and the ambiguities associated with its contemporary usage as well as those concerning the definition of 'Aboriginal peoples.' It concludes with a brief account of the history of scholarship respecting Aboriginal rights, and alternatives to the concept of rights to understand and resolve relationships between Aboriginal peoples and states.

2. *Defining Aboriginal Rights*

The concept of Aboriginal rights has existed at least since the beginnings of the period of European colonization. It originated in the political and legal system of those who colonized and poses the question of what rights rest with the original population after colonization. In this context, the term used by the British was 'Native rights.' Aboriginal rights were seen to differ from one Aboriginal group to another. Practical considerations such as the ability to resist colonial rule played an important role in determining rights (Reynolds 1999). The rationale for their determination was inevitably ethnocentric reasoning, in which similarity to European religions, customs, or economies played a crucial role (Bennett 1978).

Since the Second World War, and especially since the United Nations Declaration on *The Granting of Independence to Colonial Countries and Peoples* in 1960 (United Nations 1960), the concept of Aboriginal rights has undergone a significant shift. In Africa and Asia, colonies once ruled by European powers became independent states. Here, 'Aboriginal (or Indigenous) rights' is no longer used to describe the rights of former colonized populations, but rather the rights of peoples who now form a small and relatively powerless fragment of the state's population, such as the hunting peoples in Botswana or the scheduled tribes in India. In Scandinavia, it is applied to the rights of the Sami, albeit as an aspect of customary rather than common law. 'Aboriginal rights' is also used in Latin America, including situations where Indigenous peoples form a majority of the population in a state but do not control the state's cultural, political, and legal organization, for example, the Maya in Guatemala.

Most commonly, in the postdecolonization period 'Aboriginal Rights' is used to describe the rights of Aboriginal peoples who form a minority of the population in states founded by settlers of European origin, especially in English-speaking states with common law traditions, such as Australia, Canada (while the term 'Aboriginal rights' generally excludes treaty rights' in Canadian usage, as used here, it includes 'treaty rights'), and New Zealand. In these countries, the definition of these rights, as in the past, has been determined largely from the perspective of the state's political and legal regime. The extent and nature of these rights, as so defined, has taken different forms at different times. At some moments, states have defined Aboriginal rights as transitory, pertaining to presumptions, such as 'backwardness,' which they assumed would eventually disappear. At other times, as in Australian, Canadian, and US policies up to the 1970s, states have taken an assimilationist perspective, asserting that the future of Aboriginal peoples rests with their full integration into the general population and, concomitantly, with the disappearance of any special status ascribed to them. At other moments, governments have seen Aboriginal rights as akin to rights of ethnic or cultural minorities. They have also sought to depict Aboriginal rights as arising out of unique circumstances and are thus not comparable, for example, to colonial relations described in United Nations Declarations.

By the 1990s, Aboriginal rights had come to be understood (within the dominant legal and political regimes of settler states) as substantial legal protection to pursue a traditional way of life free from state interference. In Australia, New Zealand, and Canada, Aboriginal rights are considered to include ownership to tracts of traditional lands. States may also accept a wider range of rights. For example, Canada recognizes that Aboriginal rights include religious traditions, the pursuit of economic activities such as hunting and fishing (even using contemporary technology), and governmental powers (but only those recognized

through formal agreements with the Crown). In Canada, the rights of Indigenous peoples receive protection through a clause in its 1982 Constitution (Hogg 1997). The Aboriginal people of New Zealand, the Maori, have their rights guaranteed by the 1840 Treaty of Waitangi (Orange 1987). In the United States, Aboriginal peoples' sovereignty as "domestic dependent nations" is protected through judicial interpretation of the Constitution (Wilkinson 1987). Other states which protect Aboriginal rights do so through the common law or, as in the Philippines, by legislation (*Philippine Natural Resources Law Journal* 1999). In every case, ultimate jurisdiction over Aboriginal rights rests with the state. For example, in the United States, Aboriginal rights are under the plenary authority of Congress.

Indigenous peoples, especially in states with common law traditions, have adopted the concept and often the term, 'Aboriginal rights,' to describe their relationship with the state. They have developed at least three approaches to defining the scope of these rights. Indigenous peoples who advance the first approach see Aboriginal rights largely as rights to pursue a way of life on their traditional territories and under self-government structures with minimal interference from the state, but within the context of existing state sovereignty and ultimate jurisdiction. By accepting state sovereignty, this orientation is closely analogous to international definitions of rights of ethnic and cultural minorities (e.g., United Nations 1979).

The second approach advanced by Indigenous peoples follows the one developed in the international arena for rights of colonized peoples. The work of the Working Group on Indigenous Populations of the United Nation's Sub-Commission on Prevention of Discrimination and Protection of Minorities of the Economic and Social Council has been of major import. One result is the 'The Draft Declaration on the Rights of Indigenous Peoples.' In language that echoes the 1960 Declaration, this Declaration states that: 'Indigenous peoples have the right to self-determination. By virtue of that right they freely determine their political status and freely pursue their economic, social and cultural development' (United Nations 1994).

All parties are keenly cognizant of the usage of the 'right of self-determination' in both Declarations. Some states seek to differentiate between the right to self-determination in the two cases. Specifically, they assert that unlike the Declaration on Colonized Peoples and the Covenants, self-determination in the Draft Declaration does not sanction the redrawing of state boundaries. That is, as the New Zealand statement of 7 December 1998 to the Working Group on the Draft Declaration states (United Nations 1998):

> … any right to self-determination included in this Declaration shall not be construed as authorising or encouraging any action which would dismember or impair, totally or in part, the territorial integrity or political unity of sovereign and independent States, possessed of a government representative of the whole people belonging to the territory, without distinction as to race, creed or colour.

At the same time, many Indigenous parties consider the expression of 'self-determination' in both Declarations identical and thus see their situation as mirroring that of colonized peoples under the 1960 Declaration. In effect, this view extends the purview of that Declaration to situations of internal colonialism.

The third approach adopted by Indigenous peoples suggests that, while incorporating aspects of ethnic and minority rights on the one hand and those of colonized peoples on the other, Aboriginal rights represent something different, especially in their expression. This view is grounded on the premise that Indigenous peoples have a right to self-determination identical to that of colonized peoples. However, the resolution of this situation is not through the redrawing of political borders. Rather, they advocate a reconfiguration of political relations within the existing state, but not, as with ethnic and minority rights, within existing state polity. Instead, reconfiguration necessitates changes in the political relationship from a situation where one party dominates the other to one based on a form of political 'mutuality' (e.g., Indian Brotherhood of the Northwest Territories 1977, Orange 1987). This view of Aboriginal rights is founded on the principle of 'sharing' between peoples and is often described as established through 'treaty relationships.' In some cases, as with certain treaties in Canada, the Indigenous party understands that such a treaty relationship was established by mutual agreement at the time the treaty was negotiated. Here, the objective is to oblige the state to honor those agreements. It is an approach to resolving issues of relationship that would not necessitate the creation of new states or the redrawing of existing political borders. Only if it proved impossible to reconfigure the state in such a manner would the Aboriginal right to self-determination be expressed as the right of colonized peoples to political independence from the existing state.

Is the term 'Aboriginal rights' transitional? Perhaps. In one view, it is similar to ethnic and minority cultural rights. From another perspective, Aboriginal rights, in principle, cannot be differentiated from the rights of colonized peoples as expressed in the 1960 United Nations Declaration. Aboriginal rights from these viewpoints will be salient only as long as the political relationship between Indigenous peoples and states remains unresolved, and thereafter will be equivalent to either rights of minority cultures or rights of colonized peoples. However, the third stream of thought provides a definition of Aboriginal rights which differentiates them from those of ethnic minorities or colonized peoples. This differentiation is

located not in the origin of the rights, but in their expression. Here, the concept of Aboriginal rights is unique. It serves as a conceptual framework to reconfigure political relations between Indigenous peoples and the states within which they find themselves that promotes rather than suppresses the fact that peoples with different cultures and histories live together and share the same political space. In this sense, the concept, if not the term, may find broader application to other political situations where multiple ethnonational communities exist within the same state.

3. *Defining 'Aboriginal'*

In common usage the term 'Aboriginal' as in 'Aboriginal peoples' refers to the original people of a territory and is used to contrast that population with those who came later, especially after the invasions and colonial expansions during the past 500 years. In certain countries, the term 'Aboriginal' has a specific legal definition. For example, 'Aboriginal people' in Canada is defined constitutionally as Indians, Inuit, and Metis (the latter includes the group of people descended from marriages between Indians and settlers). In other countries, 'Aboriginal' refers to specific groups, such as in Australia where it denotes the Aboriginal people of that continent only.

Terms for particular Aboriginal peoples, such as Cree or Navajo, generally denote national identity. However, the referent for the collective term 'Aboriginal' is not that obvious. The literature lists four possibilities. The first is national identity, that is, Aboriginal peoples are defined as a collectivity of nations of people. But which nations are included and on what basis are they differentiated from other nations? The second definition is the 'original' peoples of an area, that is, those who have lived in a territory since time immemorial, or at least for a long, long time. But, does this mean that it is appropriate to refer to the Germans, the English, and other such national groups as Aboriginal? The third is way of life. In this view, 'Aboriginal peoples' are those groups who practice and have certain cultural values that encompass a particular relationship to land, certain spiritual ties, and certain kinds of internal social relationships. The typical features of Aboriginal societies are idealized to include economic norms such as gathering and hunting as well as forms of cultivation that only intrude minimally on the environment, a sense of intimate ties through spiritual relations to the land and all living and nonliving things related to it, and political systems that promote harmony rather than division. The extent to which this is always a realistic portrayal of such groups is doubtful and raises questions: Should people not live up to these ideals or cease to live up to them, are they still Aboriginal? Equally, can groups, that in other respects could not qualify as Aboriginal, become so defined solely because they make conscious efforts to incorporate such 'Aboriginal' values and practices into their lives? The fourth perspective on 'Aboriginal' is political location. Accordingly, Aboriginal refers to those national groups that are original to an area, do not have state power, and are likely in a subordinate position within existing states. Were this the case, would the achievement of state power mean that these groups are no longer Aboriginal?

Such questions may remain unanswered for a long time, in part due to the complexity of the term itself. But more importantly, the term is contested because at present the description of who constitutes Aboriginal persons and peoples may well carry with it certain internationally recognized rights. For example, there are states such as China and India that would prefer to exclude from the definition peoples who have not been incorporated into settler states through colonial processes. A definition has been developed through the United Nations Working Group on Indigenous Peoples 'for the purposes of international action that may be taken affecting their [Indigenous populations] future existence.' It is the current consensus definition. How long it remains such will ultimately depend upon political processes such as those discussed above. This definition states (Cobo 1986):

> Indigenous communities, peoples and nations are those which, having a historical continuity with pre-invasion and pre-colonial societies that developed on their territories, consider themselves distinct from other sectors of the societies now prevailing in those territories, or parts of them. They form at present non-dominant sectors of society and are determined to preserve, develop and transmit to future generations their ancestral territories, and their ethnic identity, as the basis of their continued existence as peoples, in accordance with their own cultural patterns, social institutions and legal systems .... On an individual basis, an Indigenous person is one who belongs to these Indigenous populations through self-identification as Indigenous (group consciousness) and is recognized and accepted by these populations as one of its members (acceptance by the group).

Neither the total population identified as Aboriginal by this definition nor the total number of peoples included within it can be calculated with precision; however, both are large. For example, the 1997 Working Group included Aboriginal peoples' representatives from over 25 countries from all inhabited continents. The Aboriginal peoples of these countries alone represent a population of over 100 million people (Burger 1990).

4. *Scholarship on Aboriginal Rights*

Scholarship respecting Aboriginal peoples and their rights has played an important role in Western thought since the beginnings of European colonization. One

orientation, now discredited but of great historical importance, used ethnocentric reasoning that compared Aboriginal societies and their rights with those in Europe in order to advance Western economic, social, political, and legal thought as well as to justify imperialism. Important early exemplars of this approach included Locke, Hobbes, Adam Smith, Rousseau, and Blackstone. A second orientation that continues to the present began with such as sixteenth Century Spanish scholars Vittoria, Las Casas, and Sepulveda (Dickason 1984). This approach focuses on identifying the specific rights of Aboriginal peoples under the Law of Nations and international law. A third orientation, which developed in the period between the two World Wars, concentrated specifically on the rights of 'Native' peoples within colonial legal and political systems, especially in Africa and Asia. As illustrated in the work of British social Anthropologists, the dominant paradigm, functionalism, stressed that despite cultural differences, Aboriginal societies were rational and their rights should be protected in colonial law (e.g., Malinowski 1945). Still, this scholarship was in harmony with the dominant political philosophy of the time—Indirect Rule—and did not question the ultimate authority of colonial powers in these regions.

In the postdecolonization period, scholarship on Aboriginal rights has focused largely on the situation in settler states such as Canada, the United States, Australia, and New Zealand. Three strains of research predominate. The first developed from political and legal philosophy. It concerns the nature and extent of Aboriginal rights in the abstract, asking whether they are unique or an aspect of other kinds of rights. The second, also largely researched in political science and law, questions the extent to which the state can be reshaped to accommodate the legal, economic, and political rights of Aboriginal peoples in a manner consistent with democratic principles. The third stream, which includes research in anthropology, law, political science, and other social science disciplines, discusses the legitimacy of the assertion of Aboriginal rights as special rights within the state. This approach questions the extent to which cultural differences, human rights, or the history of colonialism form legitimate grounds for assertions concerning Aboriginal rights (see *Rights: Legal Aspects*). In recent years, the scope of research has broadened to include work on Aboriginal rights and relations in countries such as in Africa and Asia that are not settler states with European origins. These streams of research have led to a rethinking of: the history of settler states (e.g., Deloria and Lytle 1984, Reynolds 1983, Williams 1990); the position of Indigenous peoples in International law (e.g., Barsh 1986, Crawford 1988); and the nature of democratic institutions and citizenship (e.g., Kymlicka 1989, Tully 1995). It is also bringing to the fore conflicts between grounding Aboriginal rights as a universal human right and as a right based on cultural difference (e.g., Wilson 1997) (see *Fundamental Rights and Constitutional Guarantees*).

The relationship between Aboriginal peoples and states has been a central focus of scholarly knowledge in many Aboriginal societies. Recently, scholarship from the viewpoint of Aboriginal societies has come into the Western academic literature, contributing to all aspects of inquiry concerning Aboriginal rights. One significant dimension concerns how specific Aboriginal peoples understand their relationships to settlers, to settler states, and to other peoples in general (e.g., Treaty 7 et al 1996). Often, these discussions are framed on the basis of sharing or treaty making between peoples rather than rights of peoples. As such, they provide a useful alternative to rights-based discourse as a means to conceptualize and reform relationships between Aboriginal peoples and states (Asch 2001).

See also: Australian Aborigines: Sociocultural Aspects; Cultural Evolution: Overview; Cultural Psychology; Cultural Rights and Culture Defense: Cultural Concerns; Discrimination; Ethnic Cleansing, History of; Ethnic Groups/Ethnicity: Historical Aspects; Fundamental Rights and Constitutional Guarantees; Gay/Lesbian Movements; Human Rights, Anthropology of; Human Rights, History of; Human Rights in Intercultural Discourse: Cultural Concerns; Human Rights: Political Aspects; Postcolonial Law; Rights

Bibliography

Asch M 2001 Indigenous self-determination and applied anthropology in Canada: Finding a place to stand. *Anthropologica* **43**(2): 201–7

Barsh R L 1986 Indigenous peoples: An emerging object of international law. *American Journal of International Law* **80**: 369–85

Bennett G 1978 *Aboriginal Rights in International Law*. Anthropological Institute for Survival International, London

Burger J 1990 *The GAIA Atlas of First Peoples*. Robertson McCarta, London

Cobo J R M 1986 *Study on the Problem of Discrimination Against Indigenous Populations*, Vol. vs. United Nations Publication E/CN.4/Sub.2/1986/7/Add.4, p. 29, paras 378 and 379

Crawford J (ed.) 1988 *The Rights of Peoples*. Clarendon Press, Oxford, UK

Deloria V Jr, Lytle C 1984 *The Past and Future of American Indian Sovereignty*. Pantheon, New York

Dickason O P 1984 *Myth of the Savage and the Beginnings of French Colonialism in the Americas*. University of Alberta Press, Edmonton, Canada

Hogg P W 1997 *Constitutional Law of Canada*, 4th edn. Carswell, Scarborough, ON

Indian Brotherhood of the Northwest Territories (Dene Nation) 1977 Dene declaration. In: Watkins M (ed.) 1977 *Dene Nation: The Colony Within*. University of Toronto Press, Toronto, Canada

Kymlicka W 1989 *Liberalism, Community, and Culture*. Oxford University Press, Oxford, UK

Malinowski B 1945 *The Dynamics of Culture Change: An Inquiry into Race Relations in Africa*. Yale University Press, New Haven, CT
Orange C 1987 *The Treaty of Waitangi*. Allen & Unwin, Wellington, New Zealand
Philippine Natural Resources Law Journal 1999 *Philippines Indigenous Peoples Rights Act* 1997
Reynolds H 1983 *The Other Side of the Frontier: Aboriginal Resistance to the European Invasion of Australia*. Penguin, Melbourne
Reynolds H 1999 *Why Weren't We Told?* Viking, Ringwood, Victoria
Treaty 7 Elders and Tribal Council with Hildebrandt W Rider D F Carter S, 1996 *The True Spirit and Original Intent of Treaty 7*. McGill-Queen's University Press, Montreal, Canada
Tully J 1995 *Strange Multiplicity: Constitutionalism in an Age of Diversity*. Cambridge University Press, New York
United Nations 1960 Declaration: *The Granting of Independence to Colonial Countries and Peoples*. General Assembly Resolution 1514(XV), December 14, 1960
United Nations 1979 *Study on the Rights of Persons Belonging to Ethnic, Religious and Linguistic Minorities* (F Capotorti, Special Rapporteur), UN Sales No. E.78.XIV.1 (1979) 16–26
United Nations 1994 *Draft United Nations Declaration on the Rights of Indigenous Peoples* as agreed upon by the members of the UN Working Group on Indigenous Populations at its eleventh session, Geneva, July 1993. Adopted by the UN Subcommission on Prevention of Discrimination and Protection of Minorities by its resolution 1994/45, August 26, 1994. UN Doc. E/CN.4/1995/2/sub.2/1994/56, at 105
United Nations 1998 Commission on Human Rights Working Group on the Draft United Nations Declaration on the Rights of Indigenous Peoples, December 7 1998. *New Zealand Statement*
Wilkinson C F 1987 *American Indians, Time and the Law*. Yale University Press, New Haven, CT
Williams R A Jr 1990 *The American Indian in Western Legal Thought*. Oxford University Press, Oxford, UK
Wilson R A (ed.) 1997 *Human Rights, Culture and Context: Anthropological Perspectives*. Pluto Press, London

M. Asch

Absolutism, History of

The term 'absolutism' first came into use in the nineteenth century. It describes a form of rule and government which evolved in the early modern period and dominated seventeenth- and eighteenth-century Central and Western Europe (though not England) during the formation of modern states.

The term itself, much like the period during which it prevailed, is by no means sharply defined. It has been less accepted in France and England than in Germany, where it also led to problems in the later decades. Any attempt to define absolutism must begin with an explanation of what it was not, which was 'absolute' in the sense of unlimited power. Intensive research into modern European history has brought to light the complex and multitiered requirements and conditions needed for the development of an absolute monarchy. It also revealed the limits of an absolute sovereign and highlighted the differences between the claim and reality of such rule.

1. Terminology and Theory

In a historical context, absolutism is a relatively new word. The term was rarely used at the turn of the seventeenth and eighteenth centuries, gained popularity in the nineteenth, but did not really thrive until later. What is meant by absolutism is a specific type of monarchy, which played an important role in seventeenth- and eighteenth-century Europe, the reality of which, however, is only conditionally described by the word itself. Characteristics of an absolute monarchy include the concentration of state power on a monarch who is not encumbered by other persons or institutions, and who can enforce his or her sovereignty with the instruments of legislation, administration, taxation, and a standing army, and who is also the final arbiter of the courts. That is the ideal definition, though one must distinguish between theory and concrete practical limitations. Even an absolute despot was bound by divine right, the country's fundamental laws, customs of inheritance, representation of the state abroad, and the preservation of law domestically.

The jurist Jean Bodin (1530–96) first used the term *potestas absoluta*, meaning the highest sovereign, who, independent of any institutions, is subject to no laws. Jaques Benigne Bossuet (1624–1707) most adamantly supported the theory of 'divine right': especially in regard to Louis XIV, Bossuet wrote that God is the source of royal power, making it absolute and independent of any temporal control. In his book *Patriarcha: or the Natural Power of Kings*, Robert Filmer (1588–1653), influenced by the Civil War, postulated that all royal power derives from God and interpreted that to mean unlimited paternal authority. Thomas Hobbes (1588–1679), on the other hand, wrote that a sovereign's power derives from a social contract, the unlimited and irrevocable transferal of natural human rights to a higher authority for protection from the natural condition, which is a state of perpetual war of all against all.

In the late seventeenth century, discussions on limiting monarchic power, influenced especially by the English Revolution, began in earnest. The struggle for power between parliament and the king led to civil war, the abolition of the monarchy, and then its restoration. It culminated in the 'Glorious Revolution' of 1688/9 and an unwritten constitution, which called for a 'king in parliament.' In France, ruled by an absolute monarch, the political discussion on absolutism would remain theoretical for another century.

John Locke (1632–1704) was not so much interested in limiting absolute rule, which hadn't taken hold in England, as in eliminating it. Based on a social contract, which stipulates that the power of the state derives from the unanimous agreement of free and equal men (civil society), the consent of the governed and majority rule, Locke wrote that all forms of government are characterized by their exercise of power. Locke introduced the principle of separation of power with checks and balances. Montesquieu (1689–1755) confronted the reality of the *monarchie absolue* with the concept of *monarchie limitée*, citing it as the most moderate form of government. The sovereign remains the source of state and civil authority, though it is implemented by intermediaries (*pouvoirs intermédiares*): the aristocracy, the professional associations and guilds (as institutions), the high courts and city magistrates. State power is also kept in check by the separation of the legislative, executive, and the judiciary. 'When the law making and law enforcement powers are united in the same person,' wrote Montesquieu, 'there can be no liberty.' Montesquieu saw England as the country whose constitution guaranteed these liberties, though he also recognized the historical and political circumstances that led up to it.

The various forms of absolutist rule in Europe cannot be defined by a set of theories, nor can they be limited to specific periods. Eberhard Weis has distinguished between early, courtly, classical, and enlightened absolutism (Weis 1985). These differences describe a development in time, but in reality they were not clearly drawn. One cannot claim that 'classical absolutism' did nothing in the way of political reform, nor that 'enlightened absolutism' can be classified as especially progressive or modern.

2. *Politics*

What we have come to call absolutism was a form of government that developed along with the creation of modern territorial states in Europe. It gained in strength as the power of the states was centralized and intensified and the monarch's rule was legitimized with the aim of stabilizing peace and security within and among the states.

This policy led to religious peace and internal state formation. An early example of an absolute monarchy, or rather a monarchy with absolute claims, is that of Phillip II of Spain (1556–98). Convinced of the divine right of his reign and his responsibility toward God, Phillip II ruled over a huge conglomerate that had grown though inheritance, marriage, and conquest. This conglomerate was comprised of politically, socially, and culturally diverse countries and territories, which all kept their rights and institutions. They were held together by the Spanish crown and represented by the bearer of the crown, whose power was unlimited in principle, though often limited in practice and not met entirely without resistance.

In practice, absolute rule was realized by expanding and securing territory through treaties, alliances, war with other states, the possession of military potential—preferably a standing army (*miles perpetuus*)—and the financial means to support it. Financial and taxation policy, central to absolutist rule, became the impetus for employing a complex bureaucracy. The ostentatious presentation of the power vested in the monarch was characteristic of this policy, including the royal court and court ceremonies.

Absolutism can rightly be seen as the dominant political tendency in seventeenth- and eighteenth-century Europe. The period between the Peace of Westphalia (1648) and the Peace of the Pyrenees (1659) on the one hand, and the French Revolution (1789) on the other, can roughly be seen as the 'age of absolutism.' But absolutism did not flourish everywhere in Europe: The Republic of Venice, the Swiss Confederation, the Dutch Republic, the elected monarchy in Poland, or the many principalities and religious states of the Holy Empire were counterexamples. The only case in Europe where absolute despotism was introduced as law was Denmark in 1661.

Sweden's introduction in parliament (*Riksdag*) of an absolutist regimen in 1686 was replaced with the 'Age of Liberty,' an almost preparliamentary system, just 40 years later. In England, the Stuart dynasty's attempts at absolute despotism were thwarted by parliament. After the Civil War, the king's execution, and the abolition and resurrection of the monarchy, the end of the Glorious Revolution (1688/89) resulted in a parliamentary monarchy with a 'king in parliament.' During the same period, absolute despotism was enjoying its high point in France. Louis XIV was a model and ideal of absolutism and a royal court thronging around the monarch.

The dissimilar results of the two cases, both of which were precipitated by heavy fighting—between the crown and nobility in France, in England between the crown and parliament presenting the entire country—points to the importance of the differing historical political and social circumstances under which the rise, stall, and failure of absolutism occurred.

For a long time, historians predominantly dealt with absolutism only under the aspect of its importance in the development of the modern state. Absolute despotism was seen to have promoted the political, social, and economic development of Europe by strengthening the state, suppressing moves for independence by powerful nobles, establishing and ensuring religious hegemony, expanding the administration, supporting trade and early industry and eliminating the political influence of professional associations and guilds. This view underestimates the dependence of the monarchy on older local and regional institutions; institutions that protected not

only their corporate rights but also the rights of individuals. The numerous *pouvoirs intermédiares*, which, according to Montesquieu, placed boundaries on freedoms of the monarch, did forfeit some of their political significance under absolute despotism, but they remained constituent elements of the monarchy.

The encroachment on castes and corporate bodies, on the privileges of the aristocracy and the church, as well as their integration into the state, and the attempt in Germany to achieve independence from the constitution and high courts, comprised only a part of absolutist policies. The organization and presentation of despotic rule was more important. This included the creation of institutions and offices dependent on the monarch, the creation by monarchs of an administration subject only to them, to include the regional level, organizing the government and recruiting its staff—not only from the nobility, but also from the bourgeoisie—and the establishment of a civil servant nobility.

Finance and taxation posed the biggest problem for absolute monarchs, especially the collection and redistribution of taxes. This is where the contradictions and limits of absolutism can be seen most clearly. The tax privileges of the aristocracy were left untouched, income and spending not balanced. France could not do away with patronage and private tax collectors. Prussia developed a tightly controlled and vigorously pursued tax system, supported by the fusion of the military and provincial treasuries and the implementation of a tax administration on the municipal level, which, relative to the population, favored an oversized army, in sharp contrast to the rigorous frugality in other areas and the representational culture of the court.

Among the notable consequences of absolutist policy were the beginnings of planned trade and economic policies following the principles of mercantile theory (initially in France) or cameralism, which in Germany was developed into administrative theory and later into the academic science of public administration. The advance of early and preindustrial manufacturing disappointed early expectations. Lagging behind England, it nevertheless precipitated the economic development on the continent. In the long term, absolutism also had an elementary effect on day-to-day life through the increasing regulation by the state, the introduction of compulsory schooling, the church maintaining state tasks (e.g., the census, the declaration of state edicts via the pulpit). This was especially true in Protestant rural areas, where the church had a profound influence on people's behavior, preaching the virtues of hard work and obedience toward state authority. Obedience to the state, both directly and indirectly, became ubiquitous. The fact that this educational process was by no means linear or similarly successful everywhere, should not lead one to doubt its influence in general. To call it *Sozialdisziplinierung*, as Gerhard Oestreich did, is only correct if one does not see the state as strict enforcer of discipline, but rather in the sense of getting the public used to government regulation, recognizing its usefulness and growing demands on the state. Under absolutist rule, court proceedings were often decided against peasants and citizens, which frequently led to difficulties enforcing laws and regulations at the local level.

Patronizing the sciences, most spectacularly by naming academics to the royal academies of science, which themselves were founded not only in expectation of research, but to increase the cultural esteem of the royal house. The constant display of power and dignity, especially at the royal court, and at which France's Louis XII exceeded, was an exercise of power. That made Versailles, both symbolically and concretely, France's cultural center. It was not only a place to which all those looked who were seeking office and rank, decoration and pardon, or commissions and rewards, but it was also where one looked to confirm and strengthen one's national pride.

The exaggeration of the *monarchie absolue* could not last. The entire system of absolutism was in danger of becoming paralyzed, losing respect and approval if it lacked the ability to conform, develop, and reform within the boundaries of the social and economic situation, or if it lacked the political consciousness, as in France, where, after failed attempts in the second half of the eighteenth century, the monarchy proved unable to take the decisive step away from absolutism and toward a constitutional monarchy.

3. *Enlightened Absolutism*

The term 'enlightened absolutism' (*Aufgeklärter Absolutismus*, *despotisme éclair*) remains controversial. German historians, particularly, have used the term to distinguish a later period of absolutism in German-speaking Europe from classical absolutism that existed in France before the revolution of 1789 and which had prevented reform. According to this view, the governments of several of the German states, influenced by the Enlightenment, which began in the mid-eighteenth century, left the idea of classical absolutism behind them. They began the transition from absolute despotism to an administrative state and a reformed monarchy.

Examples of this phenomenon were seen in the Prussia of Frederick II (1740–86) as well as in Austria during the reign of Maria Theresia and Joseph II (1740–90). There was no solid foundation of theory to support enlightened despotism. It was simply the practice of enlightened monarchs directed by their understanding of the task to modernize the political, economic, and social circumstances of their countries, and to abandon hindering laws, institutions, and traditions. In achieving this goal, they often justified using an approach even more rigorous than under

classical despotism. For Frederick II, the monarchy was an office imposed on him by a social contract which could neither be taken from him, nor from which he could escape. He considered himself the highest servant of the state and saw it as his duty to act in the best interest and well-being of the state and his subjects, but also to decide what those interests were and what means would be used to achieve them. Frederick II also determined just how much influence the Enlightenment should have on Prussian politics. His absolutist regimen was underlined and reinforced by the paring down of the royal court, his personal command and control, and increasingly rigid adherence to a monarchic military and administrative state as well as state control of the professional associations, guilds, and mercantile policies.

Did 'enlightened absolutism' possess the ability to reform itself? Was the Enlightenment a practical philosophy, a rational way of thinking, an emancipating mentality—was it the motivating factor behind absolutist monarchies' reform policies? What did it accomplish?

Much of what is often ascribed to the Enlightenment—that is, the influence of the Enlightenment on the interpretations and actions of monarchs, their ministers, and councils—actually goes back to the earlier intentions and foundations of governments in the seventeenth century. It is also based on the concept that welfare, security, and happiness are the aim and duty of politics, which was anything but new. The justification and legitimization of reform policies, their scope and aims as well as their methods of execution, were new. The policies and their success depended more on the conditions in the states of the latter half of the eighteenth century than they did on energy, providence, and political savvy. It should be noted that enlightened reform was undertaken in states that, in the eyes of the reformers, were underdeveloped (i.e., Spain, Portugal, Naples), where reforms were extended and accelerated (Austria, Tuscany), or where the reforms had to be more systematic and thorough (Prussia). The suppression of the Jesuit Order (1773) helped invigorate reform policy in Catholic countries. In other countries, reorganization was brought on by the loss of territory (Galicia and Lodomeria to Austria after the first Polish division, Silesia to Prussia after the 1st Silesian War). In many other places, the improvement of trade and commerce drove politics, as did, when the monarch was weak, the vigor of leading ministers.

With variances in the different states, the reforms applied to criminal and procedural law and prisons as well as to the standardization and codification of civil law in the various regions. The Prussian code '*Landrecht*' (1794) demanded a state-controlled school system, increased agricultural efficiency, the influx and settlement of foreigners, administrative reform, and improvements in health care and care for the poor.

The introduction of Austria's 'Toleration Patent' under Joseph II (1782) was considered a good example of enlightened reform policy. His abridgement of the (Catholic) Church's influence, closing down of monasteries and convents and attempts to improve the education of priests were met with both approval and resistance. Disciplinary measures affecting the public's religious life (limiting the number of holidays and pilgrimages) led to opposition.

The military constitution and the interests of the nobility set limits to the state protection of the peasants against the encroachment by large landowners. The semiservitude of East-Elbe peasants, for example, was left untouched. This contrasts with Austria, where the nobility was even taxed. In general, one can say that in the period of 'enlightened absolutism' the privileges of the nobility remained intact. Full emancipation of peasants was achieved only in Sweden and Denmark, and late there as well.

The reform policies of 'enlightened absolutism' was entirely an authoritarian affair. The monarch and government were sure of their actions. In most cases, they did not seek the approval of the *pouvoirs intermédiares* or of their subjects, for whom they did look out, since a free 'public' was considered an important condition of enlightened, liberal policy.

Enlightened absolutism or 'reform absolutism,' as it is sometimes called (E. Hinrichs), was often more decisive than 'classic' absolutism could be. It met very little institutional resistance. Its problem was that it left untouched the structure of privileged society, which kept tight control of the measure, flexibility, and extent of reform, and even disrupted or ended it. The aristocracy might also, as was the case with Joseph II, drive reform forward with a hasty flood of laws. The attempt has been labeled 'revolution from above.' Joseph II had to retract many of his decrees before his death. His brother and successor, Leopold II (1790–2), Grand Duke of Tuscany, a model enlightened monarch and one of Joseph's critics, had to rescind even more reforms in light of unrest aimed at Joseph in the Netherlands and Hungary and the French Revolution.

Nowhere did enlightened absolutism lead directly to a constitutional monarchy. But since the reform path had already been taken and sustained despite much resistance—the counterreformation had begun even before the French Revolution—could they be expected to continue? Did reforms, in those countries where they were instituted, prevent a revolution or even make it unnecessary? Was reform perhaps prevented or blocked by the radicalization and expansion of the revolution?

That is certainly one of the reasons the drive toward revolutionary change was relatively weak in Germany. Many of the German states, including the religious ones, were ruled by moderate, enlightened governments already prepared to make reforms. The reform policies from 1800 to 1820 were made necessary by Napoleon's intervention in Germany. In many

respects the policies were directly linked to the reforms of enlightened absolutism and Bonapartism. With the exception of the most important German states, Austria and Prussia, the reforms resulted in an early form of constitutionalism in many states. The emerging restoration jammed up, but did not end, that process. The importance of the executive over the legislative in Germany, the 'constitutional monarchy' which can be characterized as 'state' or 'bureaucratic absolutism,' remained even after the 1848 revolution.

4. Current State of Research

The actuality of 'absolutism' research is based on the:

(a) difficulty depicting, in the era after World War I, the history of modern states as a continuous development;

(b) results of prodigious international research into the history of representative (class) and parliamentary institutions;

(c) discussion of 'early modern period' as a phase between the Middle Ages and modern times;

(d) analysis of the historical circumstances and meaning of absolutism in the creation of European states;

(e) interpretation of France's political *longue durée* and its structures;

(f) debate on enlightened or reform absolutism and its modernity.

The question has arisen, during the course of research and debate, of whether one can still speak about absolute monarchy, absolute rule, and an 'age of absolutism' or if one should differentiate further and include content-related terms. Recent research into the daily life, mentality, and behavioral history of the early modern period has shown how distant the world of that time is in which elements of the modern period developed. We must also reinforce comparative research, the various conditions under which absolutism thrived, was laid claim to or failed, and its consequences in the development of state and society, even today.

See also: Enlightenment; Hobbes, Thomas (1588–1679); Locke, John (1632–1704); Montesquieu, Charles, the Second Baron of (1689–1755); Sovereignty: Political; State, History of

Bibliography

Blänker R 1992 'Absolutismus' und frühmoderner Staat. Probleme und Perspektiven der Forschung. In: Vierhaus R (ed.) *Frühe Neuzeit, Frühe Moderne*. Göttingen, Germany, pp. 48–75

Duchhardt H 1989 *Das Zeitalter des Absolutismus*. Munich, Germany

Gagliardo J A 1967 *Enlightened Despotism*. London

Gerhard D (ed.) 1969 *Ständische Vertretungen in Europa im 17. und 18. Jahrhundert*. Göttingen, Germany

Hartung F, Mournier R 1955 Quelques problèmes concernant la monarchie absolue. In: *Relazoni del x congresso internationale de science stjoriche, romea 1955*. Florence, Italy, Vol. 4, pp. 1–55

Hinrichs E 2000 *Fürsten und Mächte. Zum Problem des europäischen Absolutismus*. Göttingen, Germany

Hintze O 1970 *Staat und Verfassung. Gesammelte Abhandlungen zur allgemeinen Verfassungsgeschichte*, 2nd edn. Göttingen, Germany

Köpeczi B et al. 1985 *L'absolutisme Éclairé*. Budapest, Hungary

Kopitzsch F (ed.) 1976 *Aufklärung, Absolutismus und Bürgertum in Deutschland*. Munich, Germany

Krieger L 1976 *An Essay on the Theory of Enlightened Despotism*. Chicago

Kruedener J 1973 *Die Rolle des Hofes im Absolutismus*. Stuttgart, Germany

Kunisch J 1979 *Staatsverfassung und Mächtepolitik. Zur Genese von Staatskonflikten im Zeitalter des Absolutismus*. Berlin

Lehmann H 1980 *Das Zeitalter des Absolutismus. Gottesgnadentum und Kriegsnot*. Stuttgart, Germany

Oestreich G 1968 Strukturprobleme des europäischen Absolutismus. *Vierteljahrschrift für Wirtschafts und Sozialgeschichte* **55**: 329–47

Oestreich G 1977 *Friedrich Wilhelm I. Preußischer Absolutismus, Merkantilismus, Militarismus*. Göttingen, Germany

Raeff M 1983 *The Well-Ordered Police State: Social and Institutional Change through Law in the Germanies and Russia 1660–1800*. New Haven, CT

Vierhaus R 1966 Absolutismus. In: *Sowjetsystem und demokratische Gesellschaft. Eine vergleichende Enzyklopädie*. Freiburg, Germany, Vol. 1, pp. 17–37

Vierhaus R 1984 *Staaten und Stände. Vom Westfälischen Frieden bis zum Hubertusburger Frieden. 1648–1763*. Berlin

Vogler G 1996 *Absolutistische Herrschaft und ständische Gesellschaft. Reich und Territorien von 1648–1790*. Stuttgart, Germany

Weis E 1985 Absolutismus. In: *Staatslexikon der Görres Gesellschaft*, 7th edn., d. 1. Freiburg, Germany, pp. 37–41

Zöllner E (ed.) 1983 *Österreich im Zeitalter des Aufgeklärten Absolutismus*. Vienna

R. Vierhaus

Academic Achievement: Cultural and Social Influences

The term 'school achievement' or 'academic achievement' encompasses many aspects of students' accomplishments in school, including progress in core academic subjects—mathematics, science, language, arts, and social studies, as well as in subjects that are emphasized less frequently in contemporary curricula, such as athletics, music and the arts, and commerce. Because of the emphasis placed on the core subjects, it is to the literature in the core subjects that reference is made most frequently in discussions of research and

philosophies of education. Little attention has been paid to achievements in personal and social spheres (Lewis 1995).

1. Obstacles to Discussion

Discussions of the influence of culture and social factors on academic achievement must deal with a complex and highly contentious area of research and social policy. The area is complex because many factors are involved and contentious because the discussion often relies on differences in the personal experience of the participants rather than on carefully conducted research. Unfortunately, these obstacles sometimes allow untrained advocates of many different positions in education to gain a ready audience among policy makers and the general public.

2. Special Problems

A good deal of attention is given in this article to methodological factors that pose special problems for research dealing with the influence of cultural and social factors on academic achievement. Emphasis has been placed on academic achievement rather than school achievement because the available information has been directed primarily at attempts to explain success in core academic subjects. Consequently, the first major section of this article deals with important methodological factors. The second section is concerned with interpretations of some of the most commonly obtained results and the final section is oriented toward policy issues that emerge in these areas of concern.

3. Cultural Factors in Education and Social Policy

A commonly asked question is why cultural factors, such as beliefs, attitudes, and practices which distinguish members of one culture from others should be of general interest. They are of interest primarily because of what can be learned about the researchers' own culture (Stevenson and Stigler 1992). By placing practices from one culture in juxtaposition with those of other cultures, everyday events suddenly demand attention and concern; events once considered to be novel or unique become commonplace. For example, five-year-olds in one culture may be able to solve mathematics problems that their peers in a second culture are able to solve only after several years of formal instruction. Proposing cultural factors to account for such a difference in academic achievement leads to new perceptions of the capabilities of kindergarten children and of how cultures differ in their efforts at explanation.

Another important contribution of comparative studies of academic achievement is to clarify the characteristics of different systems of education. What may be routine in terms of what is expected in one culture may be regarded as an exciting innovation by members of another culture. As a result of such discoveries, comparative studies of academic achievement have received increasing attention since the 1970s (Paris and Wellman 1998).

4. Comparative Studies

Led primarily by the International Association for the Evaluation of Educational Achievement (IEA), some of the most extensive studies of school achievement have taken place since the 1970s. These have included the First and Second International Mathematics Study, and more recently, the Third International Mathematics and Science Study (TIMSS) a project that involved over 500,000 students in 41 countries (Beaton et al. 1996, Martin et al. 1997). These studies provide a basis for discussing methodological issues involved in research on cultural factors related to academic achievement. Results have been similar in the three IEA studies and in other, smaller comparative studies: Western students were out-scored by students from many countries, especially those from East Asia, including Japan, Hong Kong, Singapore, and South Korea (Stevenson and Lee 1998). The only high scoring Western country was the Czech Republic.

The results were startling to the low-ranking cultures; they aroused a heightened concern for issues in education policy and education reform throughout the West. Policy makers have attempted to explain the performance of students in countries such as Germany and the United Kingdom, which had exerted leadership in research and application in mathematics and science for many years, and in countries such as the United States, where there have been large investments in education for many years.

4.1 Interpreting Differences

However, rather than focusing on Western weaknesses, it is more useful to go on to other, more productive discussions. Primary among these is a consideration of the successes and weaknesses of the explanations offered by different cultures in their efforts to clarify the bases of their students' low scores. Because many of the comparisons made in TIMSS involved Germany, Japan, and the United States, reference is made to these countries in discussing comparative studies. A review of the literature is likely to suggest explanations such as the following.

4.1.1 Common explanations.

(a) *Motivation*. Western students are involved less intensely in the achievement tests than are students in other cultures and are less likely to attend closely to the problems they are to solve.

(b) *Homework*. Western students are assigned less homework than are students in East Asian cultures, thereby depriving Western students of the extensive practice and review that are available in East Asia.

(c) *Time at school*. The school day and the school year are shorter in Western cultures, creating an unfair advantage to cultures that provide their students with greater opportunities to learn.

(d) *Heterogeneity of populations*. The population of many Western countries is assumed to be more diverse than those of East Asian cultures, resulting in a disproportionate representation of low-achieving students.

(e) *Divorce and poverty*. Many Western students lack harmonious and healthy psychological and social environments at home. As a result of parental disinterest and lack of support, Western students are less able to obtain pleasure and satisfaction from their experiences at school.

5. *Problems in Evaluation*

Advocates of particular viewpoints are able to continue to propose factors such as these as explanations of the differences in achievement precisely because the proposals are so difficult to evaluate. What measures serve as reliable indices of a culture's emphasis on mathematics and science? What evidence is there that home life is less healthy in Western than in other cultures? Does time devoted to clubs and extracurricular activities rather than to academic study provide a partial account for the longer school days found in some cultures? A search of the literature would reveal little firm data to support the usefulness of these measures as reliable explanations of differences in academic achievement.

5.1 *Designing Comparative Studies*

The main problem with these explanations of cultural differences in academic achievement is that they have not been subjected to careful scrutiny, especially in the area of methodology. It is useful, therefore, to discuss some of the methodological considerations that merit attention in comparative studies of academic achievement.

5.2 *Selection of Participants*

Common criteria are required for the selection of participants in comparative studies of different cultures, for comparisons across cultures are valid only if they are based on representative samples of the members of the cultures included in the research.

5.3 *Tests*

It is obviously unfair to test students with questions that cover information not yet discussed in their classroom lessons. One appropriate index of culture-fair materials can be obtained from analyses of the content of the participants' textbooks.

5.4 *Questionnaires*

Items which are not developed within the context of the cultures participating in the research run the risk of introducing bias into the interpretation of differences in academic learning. A problematic practice is the tendency to rely on questionnaires that have been constructed in Western or other cultures, translated and back-translated, and then adopted as the research instruments in a comparative study. Translation and back-translation may be helpful with relatively simple concepts but they are often inadequate as sources of information about the psychological and cultural variables in studies involving several different cultures. Creating instruments in different languages that are truly comparable in content and nuances of meaning is extremely difficult and requires the simultaneous participation of persons with high levels of skill in each of the languages as well as in the terminology of the social sciences.

5.5 *Interviews*

Questionnaires have an important role to play in the rapid collection of large amounts of data, but one-on-one structured interviews are likely to be a more fertile ground for obtaining insights into cultural phenomena. Large-scale studies seldom have the necessary time or funds to permit such interactions with more than a subsample of the population of participants included in the study, thereby limiting the range of respondents and number of topics that can be included.

5.6 *Ethnographies*

Ethnographies, like interviews, typically are conducted with subsamples of the groups being studied, but the ability to observe and to participate in daily activities reduces the possibilities of misunderstandings, low motivation, distractibility, and other problems that may accompany the use of methods where such participation is not possible.

6. Using Computers

The introduction of computers has made it possible to conduct new types of observational analyses. Although video cameras have been used effectively in educational settings for several decades, analysis of the videotapes continues to be cumbersome.

6.1 Computer Programs for Observational Records

Computers change the manner in which observational records can be taped and analyzed, thereby greatly extending the number and reliability of observations that can be included in a study. Rather than have a single observer compile a narrative or time-sampling record of activities, permanent records made under comparable conditions are readily available for detailed analyses. The combination of compact disks for recording video images, the attachment of translations of the audio portion, and the rapid location of images illustrating concepts and processes introduces vastly expanded opportunities for the creation of reliable observations of everyday behavior in social and academic settings (Stigler and Hiebert 1999).

6.2 Statistical Analyses

A second consequence of the use of powerful computers is in statistical analyses. In general, our knowledge of correlates of academic achievement was limited because of the impossibility of processing large amounts of data. It has become routine to consider sets of data that would have been impossible to handle with the computer and recording capabilities available only a few decades ago.

6.3 Choosing a Method

There is no consensus concerning the method that is most appropriate for studies of cultural and social variables. The most vigorous argument pits quantitative methods represented in tests and questionnaires against an opposing view that seeks more frequent collection of qualitative, descriptive information.

In the end, the appropriateness of each method depends on the types of research questions being asked. TIMSS adopted all of the methods: tests, questionnaires, interviews, ethnographies, and videotaped observations. The inclusion of so many methods was partially a response to criticisms that comparative studies had not yielded test items appropriate for evaluating achievement in different cultures nor had they offered more than superficial explanations of the bases of different levels of academic achievement. Although the integration of information obtained from the various methods remains a time-consuming task, the use of various methods vastly expands the possibilities for understanding the influence of social and cultural variables on students' achievement.

7. Culture and Policy

The ultimate purpose of research dealing with the influence of culture on academic achievement is to provide evidence for maintaining the *status quo*, for instituting new policies, or for modifying older ones. In realizing these purposes, policies are translated into action. Because of methodological problems and the resulting tentativeness of conclusions derived from research involving cultural and social phenomena, discussions of education policy often involve defenses of opposing views. Several examples illustrate some of the sources of disagreement.

7.1 Nature and Nurture

As in accounts for many psychological functions, explanations for successful achievement have tended to rely on both innate biological and acquired environmental variables (Friedman and Rogers 1998). Positions that emphasize the role of biology typically discuss innate differences among persons in attributes such as intelligence, personality, and motivation. In contrast, those that emphasize the influence of experience are more likely to consider the contributions of social status, home environments, and child-rearing practices. As research has progressed in studies of achievement, an interaction view of causality has been adopted. According to this view, the effects of innate and acquired factors are considered to be interactive, such that the influence exerted by each factor depends on the status of the second factor. For example, the low ability of slow learners may be compensated for by diligent study to produce average achievement, whereas neither factor alone may provide a sufficient explanation. Throughout East Asia, practices continue to be guided by the environmentalism contained in still influential Confucian principles. In contrast, Western explanations have tended to depend increasingly on innately determined factors in their interpretations of the bases of academic achievement.

7.2 Tracking

One of the most difficult dilemmas facing educators is the question of how classrooms should be organized. In some cultures it is believed that children should be separated into ability groups early in their education; members of other cultures believe this should occur later in the child's life. In Germany, for example, students attend general-purpose schools through the

fourth grade. Following this, the rapid learners who aspire to attend university are admitted to *Gymnasien*, academically oriented secondary schools with high standards. Arrangements are made for their slower learning peers to attend schools that provide a less demanding curriculum but offer opportunities to gain practical experience that will qualify them for employment after their graduation. In China and Japan, on the other hand, the separation of students into different tracks does not occur until the students enter high school. Behind these beliefs is the assumption that academic success is strongly dependent on the child's motivation and diligence, qualities that Germans believe can be gauged by the time children are 10 years old. Japan and China reject this assumption and suggest that it is impossible to evaluate the child's interest and potential for academic work until after the student has experienced the more demanding years of junior high school.

7.3 Education Standards

Another difficult decision that must be made by education authorities concerns the level of achievement for which the academic curricula are constructed and the manner in which academic performance is evaluated. To whom should education standards be addressed? Should officials who supervise the construction of the curriculum and of evaluation aim at average students or should the standards be more demanding and establish levels of performance toward which all students should aspire?

One argument against increasing standards demanded of all students is that they will experience heightened stress and anxiety, or in the worst case, resort to suicide. Studies provide little support for this argument. Students in high-achieving countries may complain about the need to work harder but they display little evidence of heightened stress. When asked about their reaction to high academic demands they point to the fact that all students are expected to improve, thus, it is a shared requirement for which they have the support of their families and of society. Western students, with more competing goals and weaker support for academic achievement, report more frequent stress related to their school work.

Other arguments are put forth against the adoption of standards that exceed the capabilities of low achievers. Even though members of many societies may decide to emphasize the education of the average child, they are faced with the problem of exceptionally high and low achievers. High achievers are of concern because if they are placed in special classes they gain the advantage of qualifying for prestigious programs and obtaining admission to schools with high standards, excellent teachers, and up-to-date facilities. Attention to low achievers is also required if all students are to attain their maximal potential. Research on learning and teaching practices in different cultures may offer suggestions for innovative solutions to this dilemma.

7.4 Attributions for Success

Highly significant differences appear among the choices of students in different cultures when they are asked to explain the sources of high levels of academic achievement. Students are asked to explain the most important sources of achievement—studying, natural ability, difficulty of the task, or luck. East Asian students are more likely to choose studying than are Western students and Western students are more likely to choose innate ability than are East Asian students. If the alternatives are modified to include studying, having a good teacher, home environment, or innate ability, the positive consequences of studying are again emphasized by East Asian students. Western students are more likely to choose 'having a good teacher.' Whether this is because the quality of teachers actually does differ to a greater degree in Western cultures than in East Asia or because Western students are unwilling to assume responsibility for their performance, nearly all variants of the attributions about which the questions are asked produce significant cultural effects.

8. Conclusion

It is clear that academic achievement is tied closely to social and cultural factors operating within each society. Relying both on indigenous beliefs, attitudes, and practices as well as those borrowed from other cultures, information is obtained in studies comparing different societies that is of local as well as more universal significance. Thus, the ultimate goal of attempting to understand antecedents and correlates of academic achievement requires familiarity with practices that lie both within and among cultures. It seems unlikely, however, that demands for high achievement will be met by a nation's schools until the bodies of research dealing with these phenomena are understood more thoroughly.

Study of practices that exist within cultures demonstrating high levels of academic achievement may be especially fruitful. This does not mean that differences among cultures are ignored; it does mean that broader consideration of successful practices may lead to advances in performance by children and youths at all levels of ability and from a much broader range of societies.

See also: Cross-cultural Study of Education; Cultural Diversity, Human Development, and Education; Educational Policy: Comparative Perspective; Educational Systems: Asia; Motivation, Learning, and

Instruction; School Achievement: Cognitive and Motivational Determinants; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values

Bibliography

Beaton A E, Mullis I V S, Martin M O, Gonzalez D L, Smith T A 1996 *Mathematics Achievement in the Middle School Years*. TIMSS International Study Center, Boston

Friedman R C, Rogers K B 1998 *Talent in Context*. American Psychological Association, Washington, DC

Lewis C C 1995 *Educating Hearts and Minds*. Cambridge University Press, New York

Martin M O, Mullis I V S, Beaton A E, Gonzalez E J, Smith T A, Kelly D L 1997 *Science Achievement in the Primary School Years*. TIMSS International Study Center, Boston

Paris S G, Wellman H W (eds.) 1998 *Global Prospects for Education: Development, Culture, and Schooling*. American Psychological Association, Washington, DC

Stevenson H W, Lee S Y 1998 An examination of American student achievement from an international perspective. In: Ravitch D (ed.) *Brookings Papers on Education Policy*. Brookings Institution, Washington, DC, pp. 7–52

Stevenson H W, Stigler J W 1992 *The Learning Gap*. Summit, New York

Stigler J W, Hiebert J 1999 *The Teaching Gap*. Free Press, New York

H. W. Stevenson

Academic Achievement Motivation, Development of

Over the years, psychologists have proposed many different components of academic motivation (see Weiner 1992 for full discussion of history of this field). Historically, this work began with efforts to understand and formalize the role of the basic need of achievement for human drive, the introduction of the idea of competence motivation, and early work on expectancies and social learning. Developmentalists such as Vaugh and Virginia Crandall, Battle, and Heckhausen translated these ideas into a developmental framework for studying the origins of individual differences in achievement motivation (e.g., Battle 1966, V C Crandall 1969, V J Crandall et al. 1962, Heckhausen 1968). Sarason and his colleagues elaborated the concept test anxiety, developed measures, and outlined a developmental theory to explain the origins of individual differences in this critical component of academic achievement motivation (e.g., Sarason et al. 1960, Hill and Sarason 1966).

Through this early period, the focus was on achievement motivation as a drive and need. With the cognitive revolution of the 1960s, researchers shifted to a much more cognitive view of motivation. Largely through the work of Weiner, attribution theory became the central organizing framework (see Weiner 1992). This article falls in this cognitive tradition.

Eccles et al. (1998) suggested that one could group these various components under three basic questions: Can I succeed at this task? Do I want to do this task? Why am I doing this task? Children who develop positive and/or productive answers to these questions are likely to engage their school work and to thrive in their school settings more than children who develop less positive and/or noneffectual answers.

1. Can I Succeed?

Eccles and her colleagues' expectancy—value model of achievement-related choices and engagement, (see Eccles et al. 1998) is depicted in Fig. 1. Expectancies and values are assumed to directly influence performance, persistence, and task choice . Expectancies and values are assumed to be influenced by task-specific beliefs such as perceptions of competence, perceptions of the difficulty of different task, and individuals' goals and self-schema. These social cognitive variables, in turn, are influenced by individuals' perceptions of other peoples' attitudes and expectations for them, by their own interpretations of their previous achievement outcomes, and by their affective memories of, or affective expectations about, similar tasks. Individuals' task-perceptions and interpretations of their past outcomes are assumed to be influenced by socializer's behavior and beliefs, by their own histories of success and failure, and by cultural milieu and unique historical events.

Bandura (1997) proposed a social cognitive model of motivated behavior that also emphasizes the role of perceptions of efficacy and human agency in determining individuals' achievement strivings. He defined self-efficacy as individuals' confidence in their ability to organize and execute a given course of action to solve a problem or accomplish a task. Bandura proposed that individuals' efficacy expectations (also called perceived self-efficacy) are determined by: previous performance (people who succeed will develop a stronger sense of personal efficacy than those who do not); vicarious learning (watching a model succeed on a task will improve one's own self-efficacy regarding the task); verbal encouragement by others, and the level of one's physiological reaction to a task or situation.

Bandura (1997) proposed specific development precursors of self-efficacy. First, through experiences controlling immediate situations and activities, infants learn that they can influence and control their environments. If adults do not provide infants with these experiences, they are not likely to develop as strong a sense of personal agency. Second, because self-efficacy requires the understanding that the self produced an

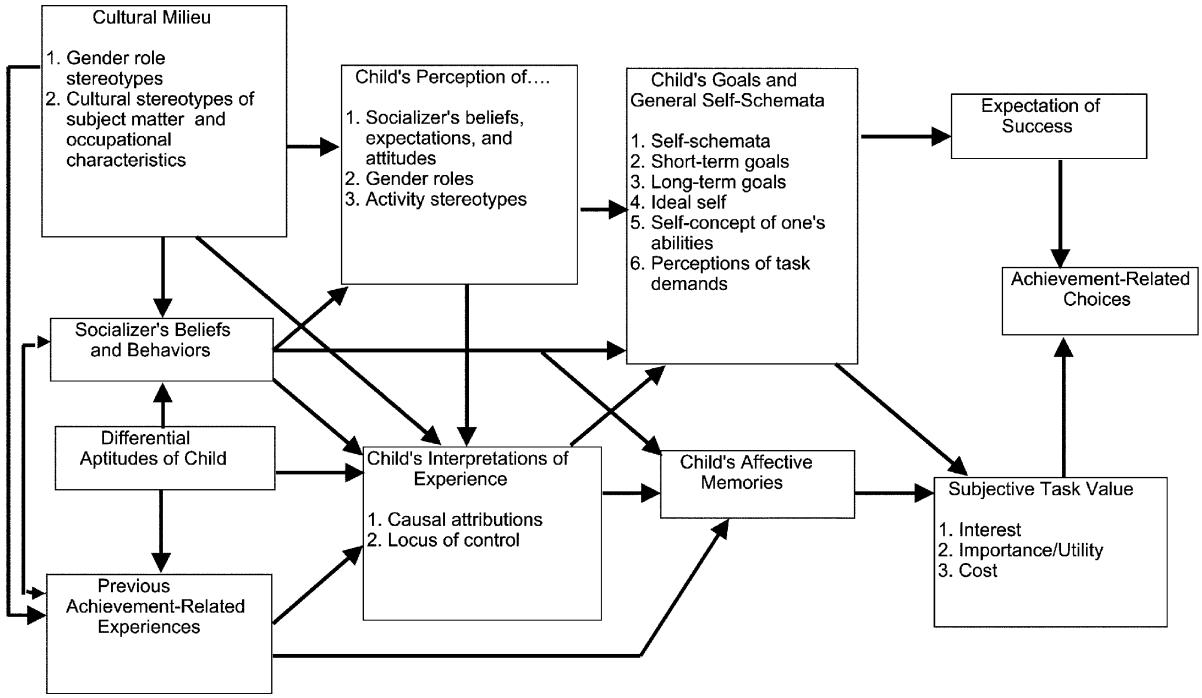

Figure 1
Model of Achievement Goals

action and an outcome, Bandura argued that a more mature sense of self-efficacy should not emerge until children have at least a rudimentary self-concept and can recognize that they are distinct individuals—which happens sometime during the second year of life. Through the preschool period, children are exposed to extensive performance information that should be crucial to their emerging sense of self-efficacy. However, just how useful such information is likely depends on the child's ability to integrate it across time, contexts, and domains. Since these cognitive capacities emerge gradually over the preschool and early elementary school years, young children's efficacy judgments should depend more on immediate and apparent outcomes than on a systematic analysis of their performance history in similar situations.

2. *The Development of Competence-related/ Efficacy Beliefs*

2.1 *Changes in Children's Understanding of Competence-related Beliefs*

Nicholls asked children questions about ability, intelligence, effort, and task difficulty, and how different levels of performance can occur when children exert similar effort (e.g., Nicholls 1990). He found four relatively distinct levels of reasoning: Level One (ages 5 to 6)—effort, ability, and performance are not clearly differentiated in terms of cause and effect; Level Two (ages 7 to 9)—effort is seen as the primary cause of performance outcomes; Level Three (ages 9 to 12)—children begin to differentiate ability and effort as causes of outcomes; Level Four—adolescents clearly differentiate ability and effort. They understand the notion of ability as capacity and believe that ability can limit the effects of additional effort on performance, that ability and effort are often related to each other in a compensatory manner, and, consequently, that a successful outcome that required a great deal of effort likely reflects limited ability.

2.2 *Change in the Mean Level of Children's Competence-related Beliefs*

Children's competence-related beliefs decline across the school years (see Eccles et al. 1998). To illustrate, in Nicholls (1979) most first graders (6 years old) ranked themselves near the top of the class in reading ability, and there was essentially no correlation between their ability ratings and their performance level.

In contrast, the 12-year-olds' ratings were more dispersed, and their correlation with school grades was 70 or higher. Expectancies for success also decrease during the elementary and secondary school years. In most laboratory-type studies, 4- and 5- year old children expect to do quite well on a specific task, even after repeatedly failing (Parsons and Ruble 1977). Across the elementary school years, the mean levels of children's expectancies for success both decline and become more sensitive to both success and failure experiences.

These studies suggest that most children begin elementary school with quite optimistic ability-related self-perceptions and expectations, and that these beliefs decline rather dramatically as the children get older. In part this drop reflects the initially high, and often unrealistic, expectations of kindergarten and first-grade children. Other changes also contribute to this decline—changes such as increased exposure to failure feedback, increased ability to integrate success and failure information across time to form expectations more closely linked with experience, increased ability to use social comparison information, and increased exposure to teachers' expectations.

Some of these changes are also linked to the transition into elementary school. Entrance into elementary school and then the transition from kindergarten to first grade introduces several systematic changes in children's social worlds. First, classes are age stratified, making within-age-ability social comparison much easier. Second, formal evaluations of competence by 'experts' begin. Third, formal ability grouping begins usually with reading group assignment. Fourth, peers have the opportunity to play a much more constant and salient role in children's lives. Each of these changes should impact children's motivation. Parents' expectations for, and perceptions of, their children's academic competence are also influenced by report card marks and standardized test scores given out during the early elementary school years, particularly for mathematics (Alexander and Entwisle 1988).

There are significant long-term consequences of children's experiences in the first grade, particularly experiences associated with ability grouping and within class differential teacher treatment. For example, teachers use a variety of information to assign first graders to reading groups including temperamental characteristics like interest and persistence, race, gender, and social class. Alexander, et al. (1993) demonstrated that differences in first-grade reading group placement and teacher-student interactions have a significant effect (after controlling for initial individual differences in competence) on motivation and achievement several years later. Furthermore, these effects are mediated by both differential instruction and the impact of ability-group placement on parents' and teachers' views of the children's abilities, talents, and motivation (Pallas et al. 1994).

3. *Theories Concerned With the Question 'Do I Want to Do This Task?'*

3.1 *Subjective Task Values*

Eccles et al. (1983) outlined four motivational components of subjective task value: attainment value, intrinsic value, utility value, and cost. Attainment value is the personal importance of doing well on the task. Intrinsic value is the enjoyment the individual gets from performing the activity, or the subjective interest the individual has in the subject. Utility value is how well a task relates to current and future goals, such as career goals. Finally, they conceptualized 'cost' in terms of the negative aspects of engaging in the task (e.g., performance anxiety and fear of both failure and success), as well as both the amount of effort that is needed to succeed and the lost opportunities resulting from making one choice rather than another.

Eccles and her colleagues have shown that ability self-concepts and performance expectancies predict performance in mathematics and English, whereas task values predict course plans and enrollment decisions in mathematics, physics, English, and involvement in sport activities even after controlling for prior performance levels (see Eccles et al. 1998). They have also shown that values predict career choices.

3.2 *Development of Subjective Task Values*

Eccles and their colleagues have documented that even young children distinguish between their competence beliefs and their task values. They have also shown that children's and adolescents' valuing of certain academic tasks and school subjects decline with age.

Although little developmental work has been done on this issue, it is likely that there are differences across age in which of the components of achievement values are most dominant motivators. Wigfield and Eccles (1992) suggested that interest is especially salient during the early elementary school grades. If so, then young children's choice of different activities may be most directly related to their interests. And if young children's interests shift as rapidly as their attention spans, it is likely they will try many different activities for a short time each before developing a more stable opinion regarding which activities they enjoy the most. As children get older the perceived utility and personal importance of different tasks likely become more salient, particularly as they develop more stable self-schema and long-range goals and plans.

A third important developmental question is how children's developing competence beliefs relate to their developing subjective task values? According to both the Eccles et al. model and Bandura's self-efficacy theory, ability self-concepts should influence the de-

velopment of task values. Mac Iver et al. (1991) found that changes in junior high school (ages 11–13) students' competence beliefs over a semester predicted changes in children's interests much more strongly than vice versa. Does the same causal ordering occur in younger children? Wigfield (1994) proposed that young children's competence and task-value beliefs are likely to be relatively independent of each other. This independence would mean that children might pursue some activities in which they are interested regardless of how good or bad they think they are at the activity. Over time, particularly in the achievement domain, children may begin to attach more value to activities on which they do well, for several reasons: first, through process associated with classical conditioning, the positive affect one experiences when one does well should become attached to the activities yielding success. Second, lowering the value one attaches to activities that one is having difficulty with is likely to be an effective way to maintain a positive global source of efficacy and self-esteem. Thus, at some point the two kinds of beliefs should become more positively related to one another.

3.3 Interest Theories

Closely related to the intrinsic interest component of subjective task value is the work on 'interest' (Renninger et al. 1992). Researchers in this tradition differentiate between individual and situational interest. Individual interest is a relatively stable evaluative orientation towards certain domains; situational interest is an emotional state aroused by specific features of an activity or a task. The research on individual interest has focused on its relation to the quality of learning. In general, there are significant but moderate relations between interest and text learning. More importantly, interest is more strongly and positively related to indicators of deep-level learning (e.g., recall of main ideas, coherence of recall, responding to deeper comprehension questions, representation of meaning) than to surface-level learning (e.g., responding to simple questions, verbatim representation of text). The research on situational interest has focused on the characteristics of academic tasks that create interest. Among others, the following text features arouse situational interest: personal relevance, novelty, and comprehensibility.

3.4 Developmental Changes in Interest

Several researchers have found that individual interest in different subject areas at school declines continuously during the school years. This is especially true for the natural sciences (see Eccles et al. 1978). These researchers have identified changes in the following instructional variables as contributing to these declines: clarity of presentation, monitoring of what happens in the classroom, supportive behavior, cognitively stimulating experiences, self-concept of the teacher [educator vs. scientist], and achievement pressure.

3.5 Intrinsic Motivation Theories

Over the last 25 years, studies have documented the debilitating effects of extrinsic incentives on the motivation to perform even inherently interesting activities (Deci and Ryan 1985). This has stimulated interest in intrinsic motivation. Deci and Ryan (1985) argue that intrinsic motivation is maintained only when actors feels competent and self-determined. Deci and Ryan (1985) also argue that the basic needs for competence and self-determination play a role in more extrinsically motivated behavior. Consider, for example, a student who consciously and without any external pressure selects a specific major because it will help him earn a lot of money. This student is guided by his basic needs for competence and self-determination but his choice of major is based on reasons totally extrinsic to the major itself. Finally, Deci and Ryan postulate that a basic need for interpersonal relatedness explains why people turn external goals into internal goals through internalization.

3.6 Developmental Changes in Intrinsic Motivation

Like interest and subjective task value intrinsic motivation declines over the school years (see Eccles et al. 1998), particularly during the early adolescent years (which coincide in many countries with the transition into upper-level educational institutions). Such changes lead to decreased school engagement. The possible origins of these declines have not been studied but are likely to be similar to the causes of declines in expectations, ability-related self-confidence and interest—namely, shifts in the nature of instruction across grade levels, cumulative experiences of failure, and increasing cognitive sophistication.

4. Why Am I Doing This?

The newest area of motivation is goal theory. This work focuses on why the children think they are engaging in particular achievement-related activities and what they hope to accomplish through their engagement. Several different approaches to goal theory have emerged. For instance, Schunk (1991) focuses on goals' proximity, specificity, and level of challenge and has shown that specific, proximal, and somewhat challenging goals promote both self-efficacy and improved performance. Other researchers have defined and investigated broader goal orientations.

Nicholls and his colleagues (Nicholls 1990) defined two major kinds of motivationally relevant goal patterns or orientations: ego-involved goals and task-involved goals. Individuals with ego-involved goals seek to maximize favorable evaluations of their competence and minimize negative evaluations of competence. Questions like 'Will I look smart?' and 'Can I outperform others?' reflect ego-involved goals. In contrast, with task-involved goals, individuals focus on mastering tasks and increasing their competence. Questions such as 'How can I do this task?' and 'What will I learn?' reflect task-involved goals. Dweck and her colleagues provide a complementary analysis distinguishing between performance goals (like ego-involved goals), and learning goals (like task-involved goals) (Dweck and Leggett 1988). Similarly, Ames (1992) distinguishes between the association of performance (like ego-involved) goals and mastery goals (like task-focused goals) with both performance and task choice. With ego-involved (or performance) goals, children try to outperform others, and are more likely to do tasks they know they can do. Task-involved (or mastery-oriented) children choose challenging tasks and are more concerned with their own progress than with outperforming others.

4.1 Development of Children's Goals

To date there has been surprisingly little empirical work on how children's goals develop. Nicholls (1990) documented that both task goals and ego goals are already developed by second graders. However, Nicholls also suggested that the ego-goal orientation becomes more prominent for many children as they get older, in part because of developmental changes in their conceptions of ability and, in part, because of systematic changes in school context. Dweck and her colleagues (Dweck and Leggett 1988) also predicted that performance goals should get more prominent as children go through school, because they develop a more 'entity' view of intelligence as they get older and children holding an entity view of intelligence are more likely to adopt performance goals.

It is also likely that the relation of goals to performance changes with age due to the changing meaning of ability and effort. In a series of studies looking at how competitive and noncompetitive conditions, and task and ego-focused conditions, influence pre- and elementary-school-aged children's interests, motivation, and self-evaluations, Butler (e.g., 1990) identified several developmental changes. First, competition decreased children's subsequent interest in a task only among children who had also developed a social-comparative sense of ability. Competition also increased older, but not younger, children's tendency to engage in social comparison. Second, although children of all ages engaged in social comparison, younger children seemed to be doing so more for task mastery reasons, whereas older children did so to assess their abilities. Third, whereas, 5, 7, and 10 year-old children's self-evaluations were quite accurate under mastery conditions, under competitive conditions 5- and 7-year-olds inflated their performance self-evaluations more than 10-year-olds.

5. The Development of Motivational Problems

5.1 Test Anxiety

Performance anxiety has been an important topic in motivational research from early on. In one of the first longitudinal studies, Hill and Sarason (1966) found that test anxiety both increases across the elementary and junior high school years and becomes more negatively related to subsequent grades and test scores. They also found that highly anxious children's achievement test scores were up to two years behind those of their low anxious peers and that girls' anxiety scores were higher than boys'. Finally, they found that test anxiety was a serious problem for many children.

High anxiety emerges when parents have overly high expectations and put too much pressure on their children (Wigfield and Eccles 1989). Anxiety continues to develop in school as children face more frequent evaluation, social comparison, and (for some) experiences of failure; to the extent that schools emphasize these characteristics, anxiety become a problem for more children as they get older.

5.2 Anxiety Intervention Programs

Earlier intervention programs emphasized the emotionality aspect of anxiety and focused on various relaxation and desensitization techniques. Although these programs did succeed in reducing anxiety, they did not always lead to improved performance, and the studies had serious methodological flaws. Anxiety intervention programs linked to the worry aspect of anxiety focus on changing the negative, self-deprecating thoughts of anxious individuals and replacing them with more positive, task-focused thoughts. These programs have been more successful both in lowering anxiety and improving performance.

5.3 Learned Helplessness

Dweck and her colleagues initiated an extensive field of research on academic learned helplessness. They defined learned helplessness 'as a state when an individual perceives the termination of failure to be independent of his responses' (Dweck and Goetz 1978, p. 157). They documented several differences between helpless and more mastery-oriented children's respon-

ses to failure. When confronted by difficulty (or failure), mastery-oriented children persist, stay focused on the task, and sometimes even use more sophisticated strategies. In contrast, helpless children's performance deteriorates, they ruminate about their difficulties, often begin to attribute their failures to lack of ability. Further, helpless children adopt an 'entity' view that their intelligence is fixed, whereas mastery-oriented children adopt an incremental view of intelligence.

In one of the few developmental studies of learned helpless behavior, Rholes et al. (1980) found that younger children did not show the same decrements in performance in response to failure as some older children do. However, Dweck and her colleagues' recent work (Burhans and Dweck 1995) suggests that some young (5- and 6-year-old) children respond quite negatively to failure feedback, judging themselves to be bad people. These rather troubling findings show that negative responses to failure can develop quite early on.

What produces learned helplessness in children? Dweck and Goetz (1978) proposed that it depends on the kinds of feedback children receive from parents and teachers about their achievement outcomes, in particular whether children receive feedback that their failures are due to lack of ability. In Hokoda and Fincham (1995), mothers of helpless third-grade children (in comparison to mothers of mastery-oriented children) gave fewer positive affective comments to their children, were more likely to respond to their children's lack of confidence in their ability by telling them to quit, were less responsive to their children's bids for help, and did not focus them on mastery goals.

5.4 *Alleviating Learned Helplessness*

There are numerous studies designed to alleviate learned helplessness by changing attributions for success and failure so that learned helpless people learn to attribute failure to lack of effort rather than to lack of ability (see Fosterling 1985). Various training techniques (including operant conditioning and providing specific attributional feedback) have been used successfully in changing children's failure attributions from lack of ability to lack of effort, improving their task persistence, and performance.

Self-efficacy training can also alleviate learned helplessness. Schunk and his colleagues (Schunk 1994) have studied how to improve low-achieving children's academic performance through skill training, enhancement of self-efficacy, attribution retraining, and training children how to set goals. A number of findings have emerged from this work. First, the training increases both children's performance and their sense of self-efficacy. Second, attributing children's success to ability has a stronger impact on their self-efficacy than does either effort feedback, or ability and effort feedback. Third, training children to set proximal, specific, and somewhat challenging goals enhances their self-efficacy and performance. Fourth, training that emphasizes process goals (analogous to task or learning goals) increases self-efficacy and skills. Finally, combining strategy training, goal emphases, and feedback to show children how various strategies relate to their performance has a strong effect on subsequent self-efficacy and skill development.

6. *Summary*

In this article, a basic model of achievement motivation was presented and discussed. Developmental origins of individual differences in students' confidence in their ability to succeed, their desire to succeed, and their goals for achievement were summarized. To a large extent individual differences in achievement motivation are accounted for by these three beliefs. Most importantly, lack of confidence in one's ability to succeed and extrinsic (rather than intrinsic) motivation are directly related to the two major motivational problems in the academic achievement domain: test anxiety and learned helplessness. Specific interventions for these two motivational problems were discussed. Future research needs to focus on interconnections among the various aspects of achievement motivation. For example, how is confidence in one's ability to master academic tasks related to individuals' desire to master these tasks and to the extent to which the individual is intrinsically motivated to work towards mastery? More work is also needed on the impact of families, schools, and peers on the development of confidence, interest, and intrinsic motivation. Exactly how can parents and teachers support the development of high interest and high intrinsic motivation to work hard to master academic tasks? Finally, we need to know a lot more about the motivational factors that underlie ethnic and gender group differences in academic achievement patterns.

See also: Academic Achievement: Cultural and Social Influences; Motivation: History of the Concept; Motivation, Learning, and Instruction; Motivation and Actions, Psychology of; School Achievement: Cognitive and Motivational Determinants; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values; Test Anxiety and Academic Achievement

Bibliography

Alexander K L, Entwisle D 1988 Achievement in the first two years of school: Patterns and processes. *Monographs of the Society for Research in Child Development* **53** (2, Serial No. 218)

Alexander K L, Dauber S L, Entwisle D R 1993 First-grade classroom behavior: Its short- and long-term consequences for school performance. *Child Development* **64**: 801–3
Ames C 1992 Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology* **84**: 261–71
Battle E 1966 Motivational determinants of academic competence. *Journal of Personality and Social Psychology* **4**: 534–642
Bandura A 1997 *Self-efficacy: The exercise of control*. Freeman, New York
Burhans K K, Dweck C S 1995 Helplessness in early childhood: The role of contingent worth. *Child Development* **66**: 1719–38
Butler R 1990 The effects of mastery and competitive conditions on self-assessment at different ages. *Child Development* **61**: 201–10
Crandall V C 1969 Sex differences in expectancy of intellectual and academic reinforcement. In: Smith C P (ed.) *Achievement -related Motives in Children*. Russell Sage Foundation, New York, pp. 11–74
Crandall V J, Katkovsky W, Preston A 1962 Motivational and ability determinants of young children's intellectual achievement behavior. *Child Development* **33**: 643–61
Deci E L, Ryan R M 1985 *Intrinsic Motivation and Self-Determination In Human Behavior*. Plenum Press, New York
Dweck C S, Goetz T E 1978 Attributions and learned helplessness. In: Harvey J H, Ickes W, Kidd R F (eds.) *New Directions in Attribution Research*. Erlbaum, Hillsdale, NJ, Vol. 2
Dweck C S, Leggett E 1988 A social-cognitive approach to motivation and personality. *Psychological Review* **95**: 256–73
Eccles J S, Wigfield A, Schiefele N 1998 Motivation to succeed. In: Eisenberg N (Vol. ed.), Demon U (series ed.) *Handbook of Child Psychology*, 5th edn. Wiley, New York, Vol. 3, pp. 1017–95
Eccles P, Adler T F, Futterman R, Goff S B, Kaczala C M, Meece J L, Midgley C 1983 Expectancies, values, and academic behaviors. In: Spence J T (ed.) *Achievement and Achievement Motivation*. H. Freeman, San Francisco, pp. 75–146
Fosterling F 1985 Attributional retraining: A review. *Psychological Bulletin* **98**: 495–512
Heckhausen H 1968 Achievement motivation research: Current problems and some contributions towards a general theory of motivation. In: Arnold W J (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, pp. 103–74
Hill K T, Sarason S B 1966 The relation of test anxiety and defensiveness to test and school performance over the elementary school years: A further longitudinal study. *Monographs for the Society for Research in Child Development* **31**(2, Serial No. 104)
Hokoda A, Fincham F D 1995 Origins of children's helpless and mastery achievement patterns in the family. *Journal of Educational Psychology* **87**: 375–85
Mac Iver D J, Stipek D J, Daniels D H 1991 Explaining within-semester changes in student effort in junior high school and senior high school courses. *Journal of Educational Psychology* **83**: 201–11
Nicholls J G 1979 Development of perception of own attainment and causal attributions for success and failure in reading. *Journal of Educational Psychology* **29**: 94–9
Nicholls J G 1990 What is ability and why are we mindful of it? A developmental perspective. In: Sternberg R J, Kolligian J (eds.) *Competence Considered*. Yale University Press, New Haven, CT
Pallas A M, Entwisle D R, Alexander K L, Stluka M F 1994 Ability-group effects: Instructional, social, or institutional? *Sociology of Education* **67**: 27–46
Parsons J E, Ruble D N 1977 The development of achievement-related expectancies. *Child Development* **48**: 1075–9
Renninger K A, Hidi S, Krapp A (eds.) 1992 *The Role Of Interest in Learning and Development*. Erlbaum, Hillsdale, NJ
Rholes W S, Blackwell J, Jordan C, Walters C 1980 A developmental study of learned helplessness. *Developmental Psychology* **16**: 616–24
Sarason S B, Davidson K S, Lighthall F F, Waite R R, Ruebush B K 1960 *Anxiety In Elementary School Children*. Wiley, New York
Schunk D H 1991 Self-efficacy and academic motivation. *Educational Psychologist* **26**: 207–31
Schunk D H 1994 Self-regulation of self-efficacy and attributions in academic settings. In: Schunk D H, Zimmerman B J (eds.) *Self-Regulation of Learning and Performance*. Erlbaum, Hillsdale, NJ
Weiner B 1992 *Human Motivation: Metaphors, Theories, and Research*. Sage, Newbury Park, CA
Wigfield A 1994 Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review* **6**: 49–78
Wigfield A, Eccles J S 1989 Test anxiety in elementary and secondary school students. *Educational Psychologist* **24**: 159–83
Wigfield A, Eccles J 1992 The development of achievement task values: A theoretical analysis. *Developmental Review* **12**: 265–310

J. S. Eccles and A. Wigfield

Academy and Society in the United States: Cultural Concerns

Over the past half-century, the American research university system has become the finest in the world. (Rosovsky 1990). Whether we measure relative standing by scientific discoveries of major importance rewarded by Nobel Prizes, the balance of intellectual migration as assessed by flows of students to American universities from abroad, or estimates of the impact of scholarly papers through citation counts, there can be little doubt that the best private and public American research universities dominate the upper tier of the world's educational institutions.

The best of American universities remain part of one societal institution where an open, free exchange of ideas is still possible and where the content of unpopular ideas is still protected reasonably well from political influence and formal sanctions. They are places that create opportunities for social mobility. They are also places where the biases and presuppositions of students and faculty are challenged; they encourage fundamental critical reasoning skills; they provide fertile soil for the development of new scholarly ideas and scientific and technological discovery; and they remain places that, at their best,

combine research, teaching, and a commitment to civic responsibility.

Throughout this period of American ascendancy, which we will define as the period from the end of World War II to the present, the educational system has undergone profound changes that reflect dynamics in the broader society. This essay is about the dynamic tensions in the relationship between the academy and the larger society and some of their consequences. These dynamic and reciprocal relations have transformed both the academy and society in ways that have benefited both, but not without creating structural strain in universities and at times in the larger society. Universities have always been places in which creative tensions have sparked change and advance. Today, the tensions are found as much with other institutions as within the academy's own borders.

Not only is the academy embedded deeply in American society but the reciprocal interactions between it and the larger culture create forces that may alter the traditional structures and normative codes of conduct associated with the great research universities. In short, the historical quasi-independence of universities from the larger cultural context during the period from their formation in the 1870s to the Second World War has been replaced by a close linkage between universities and colleges and the nation's other institutions.

The most noteworthy linkages that have had a transforming effect on the academy of the past decade have emerged from the changing relations between government, industry, and universities. Consider only a few of the extraordinarily positive outcomes of the linkages that have grown stronger over the past 50 years and some of the derivative consequences of those linkages that have posed problems for universities. Since the creation of the National Science Foundation and the National Institutes of Health in the late 1940s and early 1950s, the federal government has created the fuel that has propelled small universities into large producers of scientific and technical knowledge. This has transformed universities in a positive way. However, that linkage with government has simultaneously introduced inordinate levels of bureaucracy at universities that accompany an inordinate number of government regulations and compliance requirements that increase costs and, at times, operate to undermine some of the academy's traditional values. The impact of these features of change has done more than anything else to move universities from ecclesiastical models of organization to bureaucratic organizational forms.

Similarly, relationships with industry have enabled universities to bring knowledge more directly to the marketplace, have led to new medical treatments and new technologies of value to the larger society, and consequently have opened new streams of revenue to universities. Yet these new linkages have also created significant tensions within the academy about the norms that should govern research and about the ownership of intellectual property.

1. *Dynamic Change in the Academy*

1.1 *The Rise of Meritocracy*

Perhaps the greatest demographic shift in the American academy in the last 50 years is the increased realization of the ideal of meritocracy among the student and faculty populations. Today, the campuses of American universities and colleges mirror the faces of the nation and in many cases the faces of the peoples of the world. This heterogeneity is a recent phenomenon. In the 1950s, students at elite universities and colleges of the United States were predominantly white and Christian. In a typical class at an Ivy League institution, the diverse faces of both American society and the world were largely absent. Today, students from minority backgrounds often represent from 30 to 40 percent of the student population. Students from other nations frequently outnumber American graduate students in top quality Ph.D. programs. Increased campus diversity has led to student demands for changes in the curriculum, often linked to identity politics. Campus efforts by students to have faculty and courses related to their own ethnic or racial identities, and to create new departments of ethnic studies, have produced tensions at universities that reflect tensions in the larger society.

Efforts by universities and colleges to become more diverse in both their student and faculty populations had led to substantial controversy over the method of classification and selection of students for admissions and faculty members for appointments. Issues such as whether undergraduate admissions officers can or should take racial identity (along with other factors) into account when considering applicants for admission have become questions of substantial public and legal debate and have led to controversial state and federal policies (Bowen and Bok 1997).

The American 'revolutions' in areas of race, ethnicity, gender, and religion have had a transforming impact on the organization of the academy. Doors of opportunity have opened to groups which, as late as the 1960s, could only dream of higher education and the possibilities of the economic and social rewards attainable through this traditional avenue of social mobility. The impact of these changes has also been seen in the changing demographic profile of professors. For example, fields that had been largely closed to women and minorities have now been opened to them.

1.2 *The Changing Size, Complexity, and Composition of the Academy*

The American system of higher education has grown rapidly over the past 50 years and continues to expand.

According to the National Center for Educational Statistics, the number of institutions of higher education rose from 2,000 to 3,595 between 1960 and 1990. The number of full-time students grew from about 400,000 to 6.5 million, the percentage of women undergraduates moved from 37 to 51 percent, and the percentage of students from designated minority backgrounds increased from 12 to 28 percent. The number of doctoral degrees awarded annually increased from about 10,000 to 38,000 and the total number of faculty members grew from roughly 281,000 to close to 987,000 (Kerman 1997). There are now more than 300 universities offering a substantial number of PhD degrees. About 60 of these are considered major institutions.

1.3 Organizational Consequences of Growth and Complexity

Even if the names of the great educational institutions are the same today as 50 years ago, the institutions themselves are vastly different from what they were in the mid-twentieth century. For example, Columbia University had an operating budget of $57 million for 15 schools in 1959; 42 years later, that annual operating budget (for the same number of schools), which doubled about every 10 years, stood at approximately $2 billion (source: Columbia University Operating Plan and Capital Budget 2001–02). The pattern of growth and increased complexity has been the same in most other distinguished research universities. Not only has the size of these budgets grown. The distribution of revenues and expenses has shifted markedly during the period. Today, the set of health science schools and research institutes represent as much as half or more of the total expense budget of many research universities. The increases in federal funding for biomedical research and in physician practice plan revenues represent the two fastest growing sectors of the university budget. The budget of the traditional core of universities—the arts and sciences disciplines and undergraduate colleges—represents a smaller proportion of the total budget.

The patterns of change experienced by research universities over the past half-century have led to substantial convergence between the public and private universities. The public research universities, which are major contributors to the research base of the nation and are more likely than in the past to share preeminence with their sister private institutions, receive an increasing share of government funding. Less than 20 percent of the operating budgets of major state institutions come from state allocations, although they benefit greatly by capital investments by the states in the infrastructure needed for modern scientific and engineering research. Similarly, the private universities, which have long depended on the kindness of their alumni and friends, also receive large portions of their budgets from state and federal research agencies. In short, the private universities are becoming more public and the publics are becoming more private.

1.4 Consequences for the Growth of Knowledge

Increased size and complexity of universities has produced a set of positive and negative consequences for the growth of knowledge. Many of the great research universities, imitating industrial organizations, have become highly decentralized budgetary entities. It remains unclear whether the financial and budgetary organization act as fetters on the development of new ideas and knowledge systems. Universities may well be migrating, without the appropriate awareness of potential unintended negative consequences, toward becoming large 'holding companies', without any substantial unifying central core or force. Research universities at the beginning of the twenty-first century often resemble states in which central authority is weak and baronies (the schools), each looking out for its own interests, strong. In short, larger institutional priorities that require fund-raising rarely trump local priorities. Unifying features of research universities (the 'uni' in universities) may represent endangered species that are at risk of extinction at many universities, despite some recent efforts to reverse this trend.

Some counter-forces within the academy are operating against this pattern of extreme decentralization of authority. These may be found in the emergent and dynamic intellectual interests within the university faculties for studying research problems that intrinsically require expertise from multiple schools and disciplines. These forces are organic and are growing from within the belly of the university. The current budgetary and organizational model conflicts increasingly with the way knowledge is growing. Faculty members are finding the most interesting and challenging problems, such as understanding global climate change or understanding human genomics, at the interstices among departments and schools of the university. They are even experimenting with forming new multidisciplinary groups (virtual departments) and trying to find ways of collaborating across the barriers of disciplinary 'foreign languages.'

Free intellectual trade among departments and schools is often impeded in precisely those areas where the pressure toward multidisciplinary collaboration is greatest. The successful nineteenth-century structures for organizing the research university along departmental and disciplinary lines continue to fulfill important functions and will endure for the foreseeable future. But in the twenty-first century, they also represent limits on unusual and productive teaching and research collaborations that link people with different angles of vision born out of their disciplinary training. Deans, who oversee self-contained budgetary

schools, often place disincentives on multischool collaborations in teaching or research that have negative consequences on their own bottom line. Ironically, those bureaucratic and organizational structures that enhance the efficiencies and capabilities within schools tend to impede the ability of universities to maximize the use of talented faculty who are less committed to an organizational straight-jacket and who want to participate in scholarship without borders.

The growth in complexity and in organizational structure has produced tensions within the academy over issues of governance and decision-making. When universities were more like colleges, they could organize decision-making as faculty members and administrators as if they were 'a company of equals.' This is no longer possible. Yet, the culture of the academy resists the idea of a division of corporate responsibility, and rightly so for matters of curricula, hiring, and promotion. Faculty, who may like decentralized control of budgets and decentralized authority, have problems with corporate decision-making models. In prosperous times for higher education, when endowments are growing, when general inflation is low, when demand for educational programs is high, and when the budgets of national supporters of basic and applied research are growing rapidly, the structural problems of decision-making at universities are muted and fade into the background. However, when the economy of the university is at risk, as happens periodically, it is increasingly difficult to make hard choices among competing programmatic alternatives, to limit new initiatives, and to eliminate academic programs that are either moribund or no longer major academic priorities.

Decentralized control and lack of clarity about criteria add further complication to the task of making and carrying out tough choices in periods of limited growth for universities, especially when coupled with the question of legitimacy of the right of academic administrators (who are themselves professors) to make decisions about retrenchment as well as growth.

1.5 'Corporatizing' of the University Presidency

The organizational structure of the academy and its continual need for life-supporting resources necessary for maintaining excellence, has had an effect on the voice of leadership at these institutions. The chief executive officers of these institutions, whether presidents or chancellors, are spending increasing amounts of time on resource acquisition—whether by cultivating prospective donors, lobbying elected officials or leaders of key federal agencies with budgetary control over resources, or seeking positive publicity for their institutions. This often involves heroic effort by leaders who are expected to meet resource targets in much the same way as their corporate counterparts are expected to provide positive earnings reports each quarter. The division of intellectual labor at universities has, as a consequence, become more sharply divided. The provosts are the chief academic officers. While they have far lower profiles than the presidents and share academic policymaking with them, they and the deans have as much to do with the academic quality and strategic planning for the university as the presidents.

The need for resources has also placed enormous pressure on academic leaders to remain silent on matters of public interest and consequence, rather than take a public position on controversial issues which might resonate negatively with external audiences capable of influencing the allocation of resources to universities. These economic concerns may have done more to muzzle university presidents than any other factor. If an outspoken president of a major research university takes a public position with which a public official is in vehement disagreement, there is the risk that the stream of resources to the university will be jeopardized. Consequently, research universities, despite their impact on societal welfare, have become institutions without a strong public voice. They are less often attached to large corporations or foundations through memberships on boards of trustees and their presidents are less often appointed to lead large national commissions with mandates for inquiries into areas of national interest.

This is not to say that, in the mythical golden past, university presidents were unconstrained in formulating public positions and public policy. President Harper of the University of Chicago surely must have thought about whether his public positions would offend John D. Rockefeller, the chief benefactor of the University of Chicago during the later part of the nineteenth and early part of the twentieth century. The historical record would likely show that President Elliot of Harvard was not a totally independent agent and often had to consider the consequences of his public positions on Harvard's efforts to attract private donors. Nonetheless, the complexity of the task of gathering resources and the implications of differences of opinion between university presidents and external supporters increase the probability that sitting university presidents of the great research institutions will not speak strongly.

Although Elliot, Harper and their colleagues were undoubtedly influenced by the values of 'friends' of the university, the fact is that their constituencies—the folks that they had to be careful not to offend—were a much more homogeneous group of people. Thus the synaptic potential for offense was less varied and less complex. With the growth in diversity among the alumni of research universities, as well as within student bodies and faculty, and of external political leadership at all levels of government, there are more opportunities for a university president to offend a powerful group, no matter what position is taken.

These 'kings' continue to find that they are faced with a conflict of interest as a result of their dual role: as an individual member of the institution and as leader of the same institution. This is perhaps one reason why presidents of major research universities relinquish their office more quickly than in the past.

Universities, then, face an ironic and unintended outcome of the positive role that they have played in increasing the changing face of American society. The achievements of diversity in the academy and throughout American society have produced various groups occupying positions of wealth, power, and influence. The positive effects of these developments have simultaneously produced the unintended effect of muting voices at the very institutions that most value diversity. This may also be why highly intelligent university leaders (who often have strong private opinions on all manner of important societal issues) often appear to students and other external audiences as corporate bureaucrats rather than as people with ideas and opinions.

2. *Resources Matter: The Larger Picture*

While it plainly is not the only factor that determines institutional success, the existence of plentiful resources that can be applied toward that end remains a key ingredient to success. The amount of research university endowments grew at a rapid pace during the 1980s and 1990s —a period of unprecedented wealth creation in the United States. In the year 2000, Harvard University's endowment was roughly $19 billion compared with approximately $371 million in 1960; Yale's endowment today is roughly $10 billion; Princeton's about $8 billion; Columbia's and Chicago's slightly above and below $4 billion respectively. While most other universities have not experienced the same level of absolute growth as these, increased donations and the law of compound interest have contributed to a growing inequality of wealth among private research universities. Improved rates of interest and of return on endowments are apt to exacerbate rather than attenuate these inequalities over time, especially since about 85 percent of growth in endowments can be attributed to the appreciated value of endowments through investments, rather than to a rise in the number of gifts to the endowment.

Since prominence or prestige, whether national rankings of academic departments or the quality of professional schools, is strongly correlated with universities' levels of endowment, this growing inequality is cause for growing concern. How could the widening gap in resources influence the distribution of quality among American research universities? Will the academy become like baseball where fewer and fewer clubs can truly play in the competitive game or have a shot at winning a championship on opening day? Finally, what effects would reducing the number of distinguished universities have on the output of these universities—and with what consequences for the larger society?

3. *Resources Matter: Competitive vs. Cooperative Strategies of Growth*

Throughout much of their history, American research universities have been engaged in intense competition to establish their preeminence. Those at the top of the research pyramid continue to compete fiercely for the students with the greatest potential and for faculty members whose achievements will bring credit to the institution. Today, the most prominent faculty members are highly mobile, identifying themselves more often with their professional discipline or specialty than with their home institution, and act increasingly as academic free agents. The market for those who bring quality and prestige to a department or institution, the academic stars, has driven the cost of maintaining preeminence in many fields to staggering levels.

Competition has, for the most part, proved to be a productive force for American higher education. The top universities have had to provide the resources needed to permit the development of creative science, technology and scholarship, as well as clinical research. Students, faculty, and the larger society have benefited from periodic revolutions in the technical means and modes of scientific and scholarly production that have been financed by universities through private and public monies. Those research universities that have fared well in the competition are today's most distinguished institutions of higher learning. They also have had the largest impact on the growth of knowledge and have been the most successful in obtaining external resources to continue their missions. At these universities, there is now a technological imperative which justifies, in part, the continual quest for resources. Falling from the ranks of the most distinguished only multiplies the cost of trying to reclaim that territory, so that institutions are loath to retreat from the competition, even for short periods of time. The cost of quality, in short, continues to spiral upwards.

At what point will rising costs become dysfunctional for the system of higher education? Will there be sufficient resources, even at the most well endowed institutions, to continue in this mode of intense institutional competition, without cutting down on their range of activities? There are signs that leaders of leading research universities are seriously beginning to consider encouraging the number of cooperative efforts and joint ventures that represent enhancements for each of the institutions. One early effort is the creative sharing of books, journals, and data among libraries; another, the sharing of instructional cost

associated with teaching esoteric foreign languages; and, a third, experimenting with joint Ph.D. programs. Also, law schools, business schools, science, and engineering programs at major American universities are forming joint degree programs with major institutions in the United Kingdom, Europe, and Asia. These joint ventures permit universities to gain enormous additional strength and quality at minimal additional costs. Finally, joint ventures are already underway for the creation of new forms of learning, course content and on-line degree programs through the exploitation of new digital technologies. Although it is unlikely that the American academic community will witness full-scale mergers in the foreseeable future, it is probable that we will see a more balanced mix between competitive and cooperative strategies of growth over the next several decades.

4. The Dynamic Interaction between the Worlds of Government, Industry, and Research Universities

In his manifesto, *Science: The Endless Frontier*, Vannevar Bush (1945) outlined for President Harry Truman a scenario for American preeminence in post-World War II science and technology. Using taxpayer dollars, the government would support basic and applied research through newly created agencies (the National Science Foundation and the National Institutes of Health) and would create a mechanism for training an elite group of younger scientists and engineers who would gain advanced training in the laboratories of the nation's best academic scientists and engineers (Cole et al. 1994).

Over the past half-century since the creation of this partnership, we have seen the greatest growth in scientific and technological knowledge since seventeenth century England and arguably the greatest period of expansion in scientific knowledge in history. This partnership turned into a great American success story. It transformed the American research university and may represent the most significant example, along with the GI Bill, of how the larger society has influenced change in the academy.

While still largely in place, the partnership has become far more complex over the past 20 years, than in the immediate post-war period. For example, in 1982 the Bayh–Dole Act assigned intellectual property rights for discoveries emerging from federally funded research to the universities concerned, thus creating incentives for the translation of discoveries into useful products by bringing them to the market more rapidly and for developing technology transfer mechanisms. These incentives have worked, as universities have generated a new revenue stream to support its research and teaching mission through intellectual property licensing agreements with pharmaceutical companies or through the creation of small incubator companies. The Bayh–Dole Act was one factor that heightened awareness within universities of the value of the knowledge they created. Consequently, who 'owns' intellectual property becomes an important matter for discussion and policymaking within the research university. These new commercial activities at universities have produced normative dilemmas about free access to and utilization of knowledge.

Potential revenues from intellectual property have begun to redefine the relationship between universities and industry, as well as those within universities and closer institutional ties are clearly being established. Major companies are investing substantial research dollars in genome and other biomedical projects, in exchange for limited or exclusive rights to the intellectual property resulting from those research efforts. There are many examples of university-based discoveries that have contributed to improved treatment of disease and illness. The fruits of this increased interaction can be easily documented. However, potential problems from these closer ties are also becoming clear. The traditional normative code of science and academic research enjoined scientists and scholars to choose problems and create works without consideration of personal financial gain. This normative code may have represented an ideal never fully approximated in practice, but it was internalized and did guide behavior.

Today, the boundaries between pure academic research and work that is intended to bring significant financial returns to the scholar (often resulting from work that will also benefit the larger society) have become blurred. Increasingly, scientists are starting their own companies with the goal of bridging the divide between pure science and technological products. In short, there is increasing tension at universities between the value placed on open communication of scientific results and the proprietary impulses that lead scientists and engineers to consider the market value of discoveries that could lead universities and individuals to withhold knowledge from the public domain.

Many salient questions can be asked about these new relationships. At what point are members of the academy more often absent than present at their universities—people used to create prestige without attending to students or colleagues? Will the desire to reap the considerable rewards that could be gained from the sale of intellectual property lead to institutional blindness about conflicts of interest? Will tenure decisions be affected by the value of a scientist's patents as opposed to his scholarship? Will professors have their most talented graduate and postdoctoral students work on problems most likely to lead to patents and licenses, rather than the scientifically most significant problem for the field?

The boundaries between university and commercial ventures are apt to become even more complex in light of the extraordinary biological discoveries that will be

made in the twenty-first century. Universities will continue to try to separate potential conflicts between research and for-profit businesses. However, there will be substantial pressure to alter the normative code of inquiry so that faculty members and universities can take advantage of the changing market value of their intellectual property.

Universities, and the society in which they are embedded, are also being transformed by the revolution in new digital media communications. Hundreds of millions of people around the world are now users of the Internet. They use it to access information and, with increasing frequency, for electronic commerce. In due course, these new media technologies will be widely used for educational purposes. Higher education at many levels is apt to be affected profoundly by the development of new digital media. In fact, some prognosticators claim that the new media will make brick and mortar universities obsolete. That is unlikely, and it is very unlikely that the new digital media will undermine the quality of the best American colleges and universities or replace the functions that they fulfill. The new digital media technology does have, however, great potential to bring ideas and electronic courses of leading scholars and scientists to people around the world who would not otherwise be able to access this information or courses on demand regardless of place or time.

In short, the new digital media will revolutionize the distribution of knowledge, just as electronic books and courses will revolutionize the mode of academic production. Today, about 90 percent of all scholarly monographs sell fewer than 800 copies (half of which were purchased by libraries). Tomorrow, the same authors of high quality monographs and research studies will be able to reach audiences of 8,000 or 80,000 or more on the Internet. Realizing that the education market is huge in the United States and around the world, entrepreneurs are creating for-profit educational businesses. Few of these educational experiments are yet of high quality and fewer still profitable, but the revolutionary change has just begun.

Under pressure to explore new revenue sources and to defend itself against private internet educational companies threatening to compete with universities for virtual space and students, research universities face a 'prisoner's dilemma.' In the absence of good information about what their competitors are intending to do, should they move rapidly and at great expense, to occupy the high end educational space on the web and act as a first mover, or should they wait and examine the evolving terrain while creating for themselves high quality educational content and technological knowledge tools which facilitate the use of these materials by their own students, faculty, and alumni?

The new digital technologies are less likely to make profound changes in the research process at universities. While digital media will help scholars in their search for information published in electronic journals, and they may foster institutional collaborations, they are unlikely to substitute for close interaction found in laboratory collaborations or for the interpersonal associations so critical to the creative process. The technology is thus far more important for the distribution of knowledge than for its production.

Changes in the means of academic production of knowledge will put new pressure on traditional relationships between faculty members and their universities. For example, questions are already surfacing about ownership of intellectual property. Who should own the works that are created in digital form by members of the faculty? What are the rights and responsibilities of faculty members in the process of creating new digital course content? Should full-time faculty members be permitted to create courses for new digital businesses? Where do conflicts of commitment begin and end for members of a university's faculty? If there is income generated by universities from the sale of courses or other forms of knowledge and information, how should the faculty and the university share in the distribution of those revenues? What are appropriate and inappropriate uses of the University's name when used by faculty members creating new digital products for other institutions? Clearly, the introduction of these new digital media will require new definitions of the traditional relationships and roles of faculty, students, technology experts, and university administrative leaders.

Many of these attributes of the American academy result from its dynamic connection to the nation's other institutions. The academy has responded to the changing needs of the larger society and it has done much to influence social change and rapid economic growth in America. Correlatively, exogenous movements in the broader society create challenges for the academic world. There is little reason to believe that this dynamic interaction and its attendant tensions will change in the decades to come.

See also: Intellectual Transfer in the Social Sciences; Policy Knowledge: Universities; Scientific Academies, History of; Scientific Disciplines, History of; Universities, in the History of the Social Sciences

Bibliography

Bowen W G, Bok D 1997 *The Shape of the River*. Princeton University Press, Princeton, NJ

Bush V 1945 *Science–The Endless Frontier*. Republished by the National Science Foundation, Washington, DC, 40th Anniversary Edition

Cole J R, Barber E, Graubard S 1994 *The Research University in a Time of Discontent*. Johns Hopkins University Press, Baltimore, MD

Kerman A (ed.) 1997 *What Happened to the Humanities?* Princeton University Press, Princeton, NJ
Rosovsky H 1990 *The University. An Owners Manual.* W. W. Norton, New York

J. R. Cole

Acceptance and Change, Psychology of

When intervening, applied psychology is always oriented toward change procedures in one sense of the term 'change.' It is one thing, however, to try to change a particular piece of psychological content; it is another to change the very meaning and purpose of change efforts themselves. This distinction has been addressed in many different ways in the various traditions in psychology: first-order change versus second-order change, changes in form versus function, changes in content versus changes in context, and several others. The terms 'acceptance' and 'change' can be added to the list of distinctions oriented toward the same basic issue. This article seeks to differentiate 'acceptance' and 'change,' to define different kinds of acceptance, and to discuss what is known about the relative value of these two classes of approaches in given settings.

1. *Change*

The ordinary approach to difficult psychological content (e.g., troublesome thoughts, unpleasant bodily sensations, negative feelings, ineffective overt behavior) is to target these events for deliberate change. Change efforts of this kind are 'first-order.' That is, they are designed to reduce the frequency of difficult psychological content, to alter its intensity or other aspects of its form, or change the events that give rise to such content.

Most intervention procedures from across the many psychological traditions (behavioral, humanistic, psychodynamic, cognitive, and biological) are change-oriented in this sense. The targets, rationales, and techniques may differ, but the strategy is the same. For example, systematic desensitization may seek to reduce anxiety directed toward given stimuli, while cognitive restructuring may seek to alter irrational thoughts, but at a higher level of abstraction both procedures are designed to alter the form, frequency, or situational sensitivity of difficult psychological content.

2. *Acceptance*

Acceptance can also be conceptualized as a kind of change, but it is of a different variety: it is second-order change, metachange, or contextual change. According to the *Oxford English Dictionary*, etymologically 'acceptance' comes from a Latin root that means 'to take in' or 'to receive what is offered.' There are four primary dictionary definitions of the term, each of which have parallels in psychology:

(a) To receive willingly or with consent;

(b) To receive as sufficient or adequate, hence, to admit;

(c) To take upon oneself, to undertake as a responsibility;

(d) To receive with favor.

The first sense of the term refers in psychology to a deliberate openness, mindfulness, or psychological embracing of experience. Perhaps most emphasized originally in the humanistic traditions (but also in each of the other main traditions), this kind of *active, embracing acceptance* involves deliberate actions taken to heighten contact with psychological content, or to remove the barriers to such contact. Meditative procedures, body work, or experiential procedures, exemplify this kind of acceptance.

The second sense of the term in psychology refers to an acknowledgement of psychological events. This kind of *passive, acknowledging acceptance* involves the dissolution of barriers to admitting to states of affairs. For example, working to convince a patient of the presence of an illness or psychological disorder, or of the need to change, or working to help a grieving person to admit the permanence of death, exemplify this kind of acceptance.

The third sense of the term in psychology refers to *taking responsibility* for events, past, present, or future. Self-control, self-management, or consciousness raising exemplify this kind of acceptance. The fourth sense of the term in psychology refers to *affirmation* or *approval.*

So defined, passive acceptance and responsibility are components of virtually all psychological (and many physical) change procedures. Most healthcare interventions are based on an initial acknowledgement of the need for intervention (e.g., the drug addict must first admit to an addiction; the cancer patient must first admit to having cancer). In most problem areas (though not all, e.g., times when a 'positive halo' is helpful), a failure to acknowledge difficulty leads to ineffective remediation of that difficulty. For example, a hospitalized psychotic person who fails to admit to the need for treatment is likely not to comply with medication regimen and is thus more likely to be rehospitalized. Similarly, in all procedures that require the patient's active participation, accepting responsibility is key to follow-through and to eventual success.

Conversely, the fourth sense of the term (approval) rarely applies to healthcare interventions. It needs to be mentioned, however, because patients often have a hard time distinguishing other types of acceptance from approval. For example, an abused person may have a hard time accepting the fact of abuse in the quite reasonable sense that the perpetrator's actions should not be condoned.

Active acceptance is not a component of many psychological or physical change procedures, though there is increasing evidence of its utility, particularly with severe, chronic, or treatment-resistant problems. Thus, this is the kind of acceptance that most requires analysis and the kind that seems most innovative. Except as noted, in the remainder of this article, 'acceptance' refers to active acceptance.

3. *Domains of Acceptance and Change*

There are several domains of acceptance and change, psychologically speaking. We can break these down into personal domains on the one hand—including personal history, private events, overt behavior, and self—and social or situational domains on the other. Active acceptance is not appropriate in all of these domains.

3.1 *Personal History*

Humans are historical organisms and when looking at their problems it is the most natural thing for people to imagine that if their history had been different then their problems would have been different. In some sense, this is literally true, but since time and the human nervous system always goes 'forward,' in the sense that what comes after includes the change from what went before, it is not possible to change a history. All that can be done is to build a history from here.

A person who has been raped, for example, may imagine that it would be better not to have been raped. Unable to accomplish this, the person may try to repress the incident, pretend that it did not happen, or even pretend that it happened to someone else. The trauma literature shows, however, that these avoidant coping strategies are extremely destructive and lie at the very core of trauma (Follette et al. 1998) (see *Post-traumatic Stress Disorder*). Conversely, acceptance of one's personal history (not in the sense of approval) seems clearly necessary and healthy.

3.2 *Private Events*

A second area is that of private events, such as emotions, thoughts, behavioral predispositions, and bodily sensations. Here, the picture is more complex. Some private events surely can be changed deliberately, and it might be quite useful to do so. For example, a person might feel weak or ill, and by going to a doctor discover the source of these difficulties and ideally have them treated successfully. Minor anxieties may be readily replaced by relaxation. If emotions and thoughts repeatedly persist in the face of competently conducted change techniques, however, it may be time to give acceptance strategies a try. Furthermore, psychologically produced private events very often do not respond well to first-order change efforts, both because they are produced by extensive histories and because change efforts might paradoxically increase them (see Hayes et al. 1996 for a review).

A number of studies have demonstrated, for instance, that when subjects are asked to suppress a thought they later show an increase in this suppressed thought when compared to subjects who are not given suppression instructions (Wegner and Pennebaker 1993). More recent literature shows similar effects for some kinds of emotions and bodily sensations. For example, attempts to suppress feelings of physical pain tend to increase the length of time pain persists and to lower the threshold for pain (Cioffi and Holloway 1993).

The culture can be very supportive of first-order change practices in this area, sometimes to the point of repression. For example, it is not uncommon for a person facing the sadness associated with a death in family to be told to think about something else, to 'get on with life,' to focus on the positive things, to take a tranquilizer, or to otherwise avoid the sadness that naturally comes with the death of a loved one. Thus, it is not surprising that most patients who seek help with a psychological problem will cast that problem in terms of supposedly needed changes in emotions, thoughts, or other private events.

3.3 *Overt Behavior*

Most often, but not always, deliberate change efforts are useful and reasonable in the overt behavioral domain. Except in the sense of admission or responsibility, there is no reason to 'accept' maladaptive behavior (which in this context, would probably mean approval).

3.4 *Sense of Self*

The fourth area is the area of self. If we limit the senses of 'self' that are relevant here to those that involve knowing by the person involved, there are three senses of self to examine: self as the content of knowing, as the process of knowing, and as the context of knowing.

3.4.1 The conceptualized self. Patients invariably have a story about their problems and the sources of those problems. If this story is accepted by the client as the literal truth, it must be defended, even if it is unworkable. 'I am a mess because of my childhood' will be defended even though no other childhood will ever occur. 'I am not living because I am too anxious' will lead to efforts to change anxiety even if such efforts have always been essentially unsuccessful. Thus, acceptance of a conceptualized self, held

as a literal belief, is rarely desirable. If the story is negative, accepting it is tantamount to adopting a negative point of view that, furthermore, is to be defended. If the story is positive, facts that do not fit the tale must be distorted.

3.4.2 Self as a process of knowing. Self as the process of knowing is necessary for humans to live a civilized life. Our socialization about what to do in life situations is tied to the process of verbal knowing. For example, a person who is alexithymic will not know how to describe behavioral predispositions in emotional terms. Thus, active acceptance seems very appropriate in the area of the ongoing process self-knowledge. In most conditions, it is desirable to 'know thyself.'

3.4.3 Self as context. The final aspect of self is consciousness *per se*—that is, knowing from a consistent locus or perspective. Active acceptance is clearly beneficial here. Any attempt to disrupt continuity of consciousness (e.g., through dissociation) is almost universally harmful (Hayes et al. 1996).

3.5 Social and Situational Domains

Social and situational domains present some of the same complexities in personal domains. When we are considering the domains relevant to other persons, we can consider the acceptance of other's personal history, private events, overt behavior, or sense of self.

Once again, acceptance of others' personal history seems to be the only reasonable course available, since history is not changeable except by addition to what is. Acceptance of others' overt behavior is sometimes called for, but often change is equally appropriate. Acceptance is called for when the efforts to change overt behavior of others undermines other features of the relationship that are important, or when the behavior itself is relatively unimportant. For example, it may not be worth the effort it would take to prevent a spouse from leaving underwear on the bathroom floor. But it is also possible to err on the other side of this issue when the behaviors are not trivial. For example, a spouse may be unwilling to face the possibility of rejection and may fail to request changes in a loved one. Often this avoidance is in the name of the relationship, but in fact it contributes to a dishonest relationship.

In the area of situations, first-order change efforts are usually called for unless the situation is unchangeable. A person dealing with an assuredly fatal disease, or with the permanent disability of a child, is dealing with an unchangeable situation, and acceptance is the only reasonable course of action.

4. When Acceptance is Useful

Acceptance seems called for when one of five things occurs. First, *the process of change contradicts the outcome*. For example, if a person tries to earn self-acceptance by change, a paradox is created. Somebody may believe that he or she will be an acceptable person when they change, but the very fact that the person needs to change reconfirms the fact that they are not acceptable now.

The second instance when acceptance is called for is when *change efforts leads to a distortion, or unhealthy avoidance of, the direct functions of events*. For example, a person with a difficult childhood may insist that their childhood was always a happy one, even if it means that they cannot recall the events of their childhood clearly.

A third situation in which acceptance is called for is one in which *social change efforts disrupts the social relation or devalues the other*. Repeatedly trying to get a spouse to change a minor habit, for example, may create an aversive atmosphere that undermines the relationship itself.

A fourth situation is one in which the *outcome ultimately cannot be rule governed*. Deliberately trying to be spontaneous is doomed to failure because spontaneity does not occur by following rules, and deliberate change efforts are always efforts in rule following at their core.

The final situation has already been mentioned: *the event is unchangeable*.

5. Methods of Acceptance

Acceptance methods have always existed in psychology, but they have been largely embraced by relatively nonempirical traditions (e.g., gestalt, humanistic, and psychoanalytic traditions). Little actual data on their impact was collected until behavioral and cognitive psychologists began to explore them as well. In the modern era these more empirical forms of psychology have attempted to work out when and for whom acceptance or change methods would be most effective (see Hayes et al. 1994 for a collection of such authors). Many of these procedures have now been empirically supported, at least to a degree. Among many others, these include:

Interoceptive exposure. Deliberately creating contact with feared bodily states is among the more effective methods with several forms of anxiety disorders (Barlow et al. 1989).

Eastern traditions, such as mindfulness meditation. Mindfulness meditation is known to be helpful in several areas, such as in the acceptance of urges in substance abuse as a component of relapse prevention (Marlatt 1994) (see *Relapse Prevention Training*), the acceptance of emotions in personality disorders (Linehan 1993), or the acceptance of chronic pain (Kabat-

Zinn 1991). Mindfulness is at its essence an active acceptance procedure because it is designed to remove the barriers to direct contact with psychological events.

Social acceptance. Jacobson and his colleagues (Koener et al. 1994) have improved success in behavioral marital therapy by working on acceptance of the idiosyncrasies of marital partners as a route to increased marital satisfaction.

Cognitive defusion. Acceptance and commitment therapy (Hayes et al. 1999) is a procedure designed to increase emotional acceptance in part by undermining cognitive fusion with literal evaluations.

Emotional exposure. Emotion-focused therapy (Greenberg et al. 1993) is an experiential approach that has shown good results with couples and various psychological problems by increasing emotional exposure and acceptance (see *Experiential Psychotherapy*).

6. Conclusion

An anxious, depressed, angry, or confused individual usually thinks that these states need to be changed before a healthy and successful life can be lived. A growing body of evidence suggests otherwise. In the context of psychological acceptance, fearsome content is changed functionally, even if no change occurs in its form or its frequency. When one deliberately embraces difficult psychological content, one has transformed its function from that of an event that can cause avoidance, to that of an event that causes observation and openness. The paradox is that as one gives up on trying to be different one immediately becomes different in a very profound way. Stated another way, active acceptance is one of the most radical change strategies in the psychological intervention armamentarium.

See also: Attitude Change: Psychological; Dialectical Behavior Therapy

Bibliography

Barlow D H, Craske M G, Cerny J A, Klosko J S 1989 Behavioral treatment of panic disorder. *Behavior Therapy* **20**: 261–82

Cioffi D, Holloway J 1993 Delayed costs of suppressed pain. *Journal of Personality and Social Psychology* **64**: 274–82

Follette V M, Ruzek J I, Abueg F F 1998 *Cognitive Behavioral Therapies for Trauma*. Guilford Press, New York

Greenberg L S, Rice L N, Elliott R 1993 *Facilitating Emotional Change: The Moment-By-Moment Process*. Guilford Press, New York

Hayes S C, Jacobson N S, Follette V M, Dougher M J (eds.) 1994 *Acceptance and Change: Content and Context in Psychotherapy*. Context Press, Reno, NV

Hayes S C, Strosahl K D, Wilson K G 1999 *Acceptance and Commitment Therapy: An Experiential Approach to Behavior Change*. Guilford Press, New York

Hayes S C, Wilson K W, Gifford E V, Follette V M, Strosahl K 1996 Experiential avoidance and behavioral disorders: A functional dimensional approach to diagnosis and treatment. *Journal of Consulting and Clinical Psychology* **64**: 1152–68

Kabat-Zinn J 1991 *Full Catastrophe Living*. Delacorte Press, New York

Koener K, Jacobson N S, Christensen A 1994 Emotional acceptance in integrative behavioral couple therapy. In: Hayes S C, Jacobson N S, Follette V M, Dougher M J (eds.) *Acceptance and Change: Content and Context in Psychotherapy*. Context Press, Reno, NV, pp. 109–18

Linehan M M 1993 *Cognitive-behavioral Treatment of Borderline Personality Disorder*. Guilford Press, New York

Marlatt G A 1994 Addiction and acceptance. In: Hayes S C, Jacobson N S, Follette V M, Dougher M J (eds.) *Acceptance and Change: Content and Context in Psychotherapy*. Context Press, Reno, NV, pp. 175–97

Wegner D M, Pennebaker J W (eds.) 1993 *Handbook of Mental Control*. Prentice-Hall, Englewood Cliffs, NJ

S. C. Hayes

Access: Geographical

1. Definition and Meaning

Access in a geographical context is the quality of having interaction with, or passage to, a particular good, service, facility, or other phenomenon that exists in the spatiotemporal world. For example, access may be based on measuring the distance or travel time between where residents live (housing units) and the facilities they need (e.g., medical facilities, shops, workplaces). Access is also a relative concept that varies according to the level of opportunity afforded at the destination. Assessments of access (or lack of access) are made meaningful by comparing access in one zone (or for one type of individual) with access in (or for) another. If goods are spatially specific, geographical access typically involves one or more origins and one or more destinations and the distance between them.

1.1 Alternative Definitions

While the above definition of access is common, there are other concepts of access that should be identified. One emerging view is that the notion of access must be redefined for the information age, whereby transactions take place in virtual as opposed to physical space or some hybrid form (NCGIA 1998). Part of this interest relates to the notion of varying levels of access to information technologies and how this variation effects matters of equity in a wide variety of ways. But there is also an attempt to understand how information technology has changed accessibility patterns by changing the geographic locations of people and the built environment that sustains them. If infor-

mation technologies affect patterns of land use, for example, such technologies indirectly affect accessibility patterns that are determined by land use configurations. Another complexity is that access does not have to be viewed as a positive phenomenon. Access to goods can also be negative, as in the case of environmentally hazardous areas, dilapidated buildings, or other services and facilities considered to have an excessive number of negative externalities. When these negative costs associated with access are analyzed, the issue is one of environmental justice, and whether there are discriminatory patterns of negative access along racial or economic lines. Studies have shown that access to negative conditions in the environment is often higher among low-income groups (Bowen et. al. 1995).

Some views of access are not based on distances between two or more locations in space, but may instead be based on social factors, cultural barriers, or ineffective design. For example, there may be barriers to access based on whether or not an individual possesses a certain subjectively defined level of 'citizenship' (Staeheli and Thompson 1997). In addition to exclusionary practices that prohibit certain groups from 'free' access to a given good, there may be problems inherent to the good, service or place itself. For example, public parks may or may not be designed appropriately to deter crime by incorporating defensible space techniques, which in turn may significantly impact access to that space.

1.2 Spatial Equity and Access

Access defined on the basis of spatial distributions invokes the concept of spatial equity. The issue is one of who has access to a particular good or service and who does not, and whether there is any pattern to these varying levels of access. Spatial equity can be defined as equality, in which everyone receives the same public benefit (i.e., access), regardless of socioeconomic status, willingness to pay, or other criteria. Alternatively, access equity may vary according to indicators such as poverty, race, or the nature of the service being provided.

In a distance-based analysis, the purpose of research on access might address the question of whether access to a particular good is discriminatory. Such inquiries might entail, more specifically, an examination of the extent to which there is a spatial pattern to varying levels of access, and whether that spatial pattern varies according to spatially-defined socioeconomic groups (Talen 1998). For example, do people of color have to travel further to gain access to public goods than others?

Over the past several decades, researchers have examined patterns of accessibility to certain services and the spatial relationship between service deprivation and area deprivation (Knox 1978, Pacione 1989). Geographers have explored regional and local variations in access to, among other things, recreational amenities, secondary education, public playgrounds, and child care (see Talen 1998).

In addition to exposing differentials in accessibility, there is the quest to discover why certain patterns of access exist. Factors implicated include urban form, organizational rules, citizen contacts, politics, and race. Until recently, spatial inequity has been explained predominantly by the notion of unpatterned inequality (Mladenka 1980). This is the idea that although there is inequality in people's access to services and facilities, there is no evidence that there is a clear discriminatory pattern to it. In the absence of patterned inequality, some argue, it is difficult to attach blame to those responsible for the existing distributional pattern. Current critiques of this theory (Miranda and Tunyavong 1994) focus on the failure to take the political process properly into account, and the problem of variable definition.

1.3 Normative Views of Access

Interrelated to equity considerations, the concept of access has taken on a normative role. Specifically, access and its aggregate, accessibility, are increasingly seen as important criteria of well-designed urban environments. The promotion of access through planning and design is seen as a way of counter balancing the decentralizing forces of metropolitan expansion. Access to facilities, goods and services in a spatial sense is what differentiates urban sprawl from compact city form. In short, some urban forms inherently have better access: development patterns that are low-density and scattered necessarily diminish accessibility because facilities tend to be far apart and land uses are segregated (Ewing 1997). For locally oriented populations, accessibility to urban services is crucial because distance is not elastic (Wekerle 1985). This is particularly true for populations who rely on modes of transport other than the automobile (e.g., the elderly and the poor).

Current models of normative urban pattern give geographical access a prominent and defining role, and access is viewed as having a direct impact on quality of life. Physical proximity defines access, and it retains importance despite the increase in nonspatial forms of interaction occurring via virtual networks. Most importantly, access can be improved through design. Kevin Lynch (1981) made this connection early on, and held 'access' as a key component of his theory of ideal urban form. His view of access was highly qualitative, since he viewed it as integral to the 'sensuous' quality and symbolic legibility of place. In this same genre, New Urbanists have developed a specific town planning manifesto based on enhancing access at the level of region (by promoting a variety of transportation alternatives), metropolis (by promoting compact urban form), and neighborhood (by promoting mixed uses and housing density).

2. The Measurement of Access

The majority of studies of geographic access assume that access is a positive phenomenon, and that access is based in part on some measurement of distance in space. If access is being considered as something desirable, the impediments to access—friction or blockage of the opportunity to interact or the right to enter—must be factored in. Right of entry assumes that a transaction must occur between the consumer and the good, service, or facility. More important in terms of definitions of access is that these transactions have a cost associated with them. Geographical interest in access therefore is often focused on these transaction costs. The interest is often methodological—how can these transaction costs be measured? Empirical investigations—who pays a higher transaction cost for access and why—invoke the issue of spatial equity discussed above.

2.1 Factors Affecting the Measure of Access

There are five classes of factors affecting the measure of access. The first two are simply the spatial locations of points of origin and points of destination. Usually the points of origin refer to housing locations, and points of destination involve entities that can be spatially referenced, such as schools or places of employment.

The third factor is the travel route and its distance between an origin(s) and destination(s). This involves not only the distance between two or more points, but the qualities of the route and the mode of travel that occurs on that route. Factors that effect the route include topography, design speed, number of lanes of traffic, and mode. For pedestrian access, perceived safety, sidewalk quality, and traffic volumes are important factors. Measuring the distance along a route can be based on the shortest distance between destination and origin, or can be more complex and involve a variety of spatial networks.

Another factor affecting the measurement of access has to do with the attributes of the individuals who seek access. Characteristics of individuals are usually derived using the characteristics of a given spatial unit, such as a census block (the degree of disaggregation of the spatial unit varies widely). Factors that might affect access include socioeconomic status, age, gender, and employment status. Certain assumptions can be made about the attractiveness or relevance of travel to certain facilities (and the likely mode of travel) based on these characteristics. The frictional effect of the available travel mode is also likely to be predicated on the characteristics of residents. For example, lack of bus service may adversely impact access for low-income individuals but have only a marginal effect on higher income groups.

A final class of factors involves the destination(s), specifically the amount, type, and quality of a given destination (e.g., facility). These attributes determine the attractiveness of a destination for consumers, and therefore affects how access to it is measured.

2.2 Types of Measures

Accessibility is measured in a variety of ways, and there can be significant variation in the resulting measurement depending on which method is used. Traditional measures assess the 'cumulative opportunities' of a given location (Handy and Niemeier 1997). There can be counts of the number of facilities within a given spatial unit or range, or measures based on average travel cost and minimum distance. Alternatively, a gravity potential measure can be used in which facilities are weighted by their size (or other characteristic) and adjusted for the frictional effect of distance. Another type of accessibility measure is based on random utility theory and measures access on the basis of the desirability or utility of a set of destination choices for an individual (Handy and Niemeier 1997).

Finally, accessibility measures can be based on individual rather than place access (see Hanson and Schwab 1987, Kwan 1999). This approach, which uses travel diaries to determine destination choices and linkages between them in an individual's daily pattern of movement, has two important advantages. First, multipurpose trips can be factored in and therefore interdependencies in trip destinations can be taken into account. Second, individual space–time constraints can be included in the evaluation of differences in personal accessibility.

See also: Discrimination; Discrimination, Economics of; Discrimination: Racial; Justice, Access to: Legal Representation of the Poor; Location: Absolute/Relative; Spatial Equity

Bibliography

Bowen W M, Salling M J, Haynes K E, Cyran, E J 1995 Toward environmental justice: Spatial equity in Ohio and Cleveland. *Annals of the Association of American Geographers* **85**: 641–63

Ewing R 1997 Is Los Angeles-style sprawl desirable? *Journal of the American Planning Association* **63**: 107–26

Handy S L, Niemeier D A 1997 Measuring accessibility: An exploration of issues and alternatives. *Environment and Planning A* **29**: 1175–94

Hanson S, Schwab M 1987 Accessibility and intraurban travel. *Environment and Planning A* **19**: 735–48

Knox P L 1978 The intraurban ecology of primary medical care: Patterns of accessibility and their policy implications. *Environment and Planning A* **10**: 415–435

Kwan M-P 1999 Gender and individual access to urban opportunities: A study using space-time measures. *Professional Geographer* **51**: 210–27

Lynch K 1981 *Good City Form*. MIT Press, Cambridge, MA

Miranda R A, Tunyavong I 1994 Patterned inequality? Reexam-

ining the role of distributive politics in urban service delivery. *Urban Affairs Quarterly* **29**: 509–34
Mladenka K R 1980 The urban bureaucracy and the Chicago political machine: Who gets what and the limits to political control. *American Political Science Review* **74**: 991–98
NCGIA 1998 *Measuring and Representing Accessibility in the Information Age*. Varenius Conference held at Pacific Grove, CA, November 20–22
Pacione M 1989 Access to urban services—the case of secondary schools in Glasgow. *Scottish Geographical Magazine* **105**: 12–18
Staeheli L A, Thompson A 1997 Citizenship, community and struggles for public space. *The Professional Geographer* **49**: 28–38
Talen E 1998 Visualizing fairness: Equity maps for planners. *Journal of the American Planning Association* **64**: 22–38
Wekerle G R 1985 From refuge to service center: Neighborhoods that support women. *Sociological Focus* **18**: 79–95

E. Talen

Accidents, Normal

Normal Accident Theory (NAT) applies to complex and tightly coupled systems such as nuclear power plants, aircraft, the air transport system with weather information, traffic control and airfields, chemical plants, weapon systems, marine transport, banking and financial systems, hospitals, and medical equipment (Perrow 1984, 1999). It asserts that in systems that humans design, build and run, nothing can be perfect. Every part of the system is subject to failure; the design can be faulty, as can the equipment, the procedures, the operators, the supplies, and the environment. Since nothing is perfect, humans build in safeguards, such as redundancies, buffers, and alarms that tell operators to take corrective action. But occasionally two or more failures, perhaps quite small ones, can interact in ways that could not be anticipated by designers, procedures, or training. These unexpected interactions of failures can defeat the safeguards, mystify operators, and if the system is also 'tightly coupled' thus allowing failures to cascade, it can bring down a part or all of system. The vulnerability to unexpected interactions that defeat safety systems is an inherent part of highly complex systems; they cannot avoid this. The accident, then, is in a sense 'normal' for the system, even though it may be quite rare, because it is an inescapable part of the system.

Not all systems are complexly interactive, and thus subject to this sort of failure; indeed, most avoid interactive complexity if they can, and over time become more 'linear,' by design. (The jet engine is less complex and more linear than the piston engine.) And not all complexly interactive systems are tightly coupled; by design or just through adaptive evolution they become loosely coupled. (The air traffic control system was more tightly coupled until separation rules and narrow routes or lanes were technically feasible, decoupling the system somewhat.) If the system has a lot of parts that are linked in a 'linear' fashion the chances of unanticipated interactions are remote. An assembly line is a linear system, wherein a failure in the middle of the line will not interact unexpectedly with a

Complex systems	*Linear systems*
Proximity	Spatial segregation
Common-mode connections	Dedicated connections
Interconnected subsystems	Segregated subsystems
Limited substitutions	Easy substitutions
Feedback loops	Few feedback loops
Multiple and interacting controls	Single purpose, segregated controls
Indirect information	Direct information
Limited understanding	Extensive understanding
Tight coupling	*Loose coupling*
Delays in processing not possible	Processing delays possible
Invariant sequences	Order of sequences can be changed
Only one method to achieve goal	Alternative methods available
Little slack possible in supplies, equipment, personnel	Slack in resources possible
Buffers and redundancies are designed-in, deliberate	Buffers and redundancies fortuitously available
Substitutions of supplies, equipment, personnel limited and designed-in	Substitutions fortuitously available

Figure 1
Characterstics of the two major variables, complexity and coupling

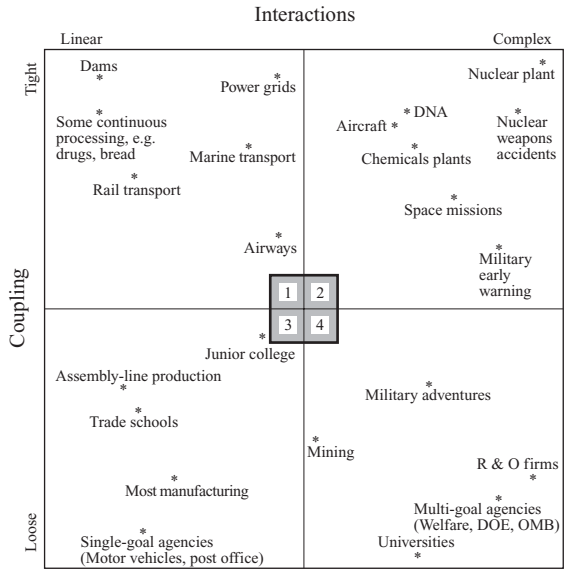

Figure 2
Interaction/coupling chart showing which systems are most vulnerable to system accidents

failure near the end, whereas a chemical plant will use waste heat from one part of the process to provide heat to a previous or later part of the process. A dam is a linear system; a failure in one part is comprehensible and though it may be one that is not correctable, making an accident inevitable, the system characteristics are not the cause of the failure; a component simply failed. But a dam is tightly coupled, so the component failure cannot be isolated and it precipitates the failure of other components. A university is an example of a complexly interactive system that is not tightly coupled. Substitutes can be found for an absent teacher, another dean for an absent dean, or to retract or delay a mistaken decision, the sequencing of courses is quite loose and there are alternative paths for mastering the material. Unexpected interactions are valued in a university, less so in the more linear vocational school, and not at all in the business school teaching typing. Figure 1 summarizes some of the characteristics of the two major variables, complexity and coupling.

NAT has a strong normative content. It emerged from an analysis of the accident at the Three Mile Island nuclear power plant in Pennsylvania in 1979. Much of the radioactive core melted and the plant came close to breaching containment and causing a disastrous escape of radioactivity. The catastrophic potential of that accident, fortunately not realized, prompted inquiry. It appeared that elites in society were causing more and more risky systems with catastrophic potential to be built, and just trying harder was not going to be sufficient to prevent catastrophes. Though people at all levels in the company running the Three Mile Island plant did not appear to have tried very hard to prevent accidents, the more alarming possibility was that even if they had, an accident was eventually inevitable, and thus a catastrophe was possible. Other systems that had catastrophic potential were also found to be both complexly interactive and tightly coupled. Figure 2 arrays these two variables in a manner that suggests which systems are most vulnerable to system accidents. The catastrophic potential of those in the upper right cell is evident. The policy implications of this analysis is that some systems have such extensive catastrophic potential (killing hundreds with one blow, or contaminating large amounts of the land and living things on it), that they should be abandoned, scaled back sharply to reduce the potential, or completely redesigned to be more linear in their interactions and more loosely coupled to prevent the spread of failures.

Normal Accidents reviews accidents in a number of systems. The Three Mile Island (TMI) accident was the result of four failures, three of which had happened before (the fourth was a failure of a newly installed safety device), all four of which would have been handled easily if they had occurred separately, but could not when all four interacted in unforseen ways. The system sent correct, but misleading, indications to the operators, and they behaved as they had been trained to do, which made the situation worse. Over half of the core melted down, and had it not been for the insight of a fresh arrival some two hours into the accident, all of the core could have melted, causing a breach of containment and extensive radioactive releases. Several other nuclear power plant accidents appear to have been system accidents, as opposed to the much more common component failure accidents, but were close calls rather than proceeding as far as that at TMI. Several chemical plant accidents, aircraft accidents, and marine accidents are detailed that also fit the definition, and though there were deaths and damages, they were not catastrophic. In such linear systems as mining, manufacturing, and dams the common pattern is not system accidents but preventable component failure accidents.

One of the implications of the theory concerns the organizational dilemma of centralization versus decentralization. Some processes still need highly complex interactions to make them work, or the interactions are introduced for efficiency reasons; tight coupling may be required to ensure the most economical operation and the highest throughput speed. The CANDO nuclear reactors in Canada are reportedly more forgiving and safer, but they are far less efficient than the 'race horse' models the USA adopted from the nuclear navy. The navy design did not require huge outputs with continuous, long-term 'base load' operation, and was smaller and safer; the electric power plant scaled up the design to an unsafe level to achieve economies of scale.

Tight coupling, despite its associated economies, requires centralized decision making; processes are fast and invariant, and only the top levels of the system have a complete view of the system state. But complex interactions with uncertainty call for decentralized decision making; only the lower-level operators can comprehend unexpected interactions of sometimes quite small failures. It is difficult, and perhaps impossible, to have a system that is at the same time centralized and decentralized. Given the proclivities of designers and managers to favor centralization of power over its decentralization, it was a fairly consistent finding that risky systems erred on the centralization side and neglected the advantages of decentralization, but it was also clear that immediate, centralized responses to failures had their advantages. No clear solution to the dilemma, beyond massive redesign and accompanying inefficiencies, was apparent.

A few noteworthy accidents since the 1984 publication of *Normal Accidents* have received wide publicity: the Challenger space shuttle, the devastating Bhopal (India) accident in Union Carbide's chemical plant, the Chernobyl nuclear power plant explosion in the former USSR, and the Exxon Valdez oil tanker accident in Alaska. (These are reviewed in the Afterword in a later edition of *Normal Accidents* (Perrow 1999).) None of these were truly system accidents; rather, large mistakes were made by designers, management, and workers in all cases, and all were clearly avoidable. But the Bhopal accident, with anywhere from 4,000 to 10,000 deaths, prompted an important extension of Normal Accident Theory. Hundreds of chemical plants with the catastrophic potential of Bhopal have existed for decades, but there has been only one Bhopal. This suggests that it is very hard to have a catastrophe, and the reason is, in a sense, akin to the dynamics of system accidents. In a system accident everything must come together in just the right way to produce a serious accident; that is why they are so rare. We have had vapor clouds with the explosive potential to wipe out whole suburbs, as in the case of a Florida suburb, but it was night and no cars or trucks were about to provide the spark. Other vapor clouds have exploded with devastating consequences, but in lightly populated rural areas, where only a few people were killed. The explosion of the Flixborough chemical plant in England in 1974 devastated the plant and part of the nearby town, but as it was a Saturday few workers were in the plant and most of the townspeople were away shopping.

Warnings are important. There was none when the Vaiont dam in Italy failed and 3,000 people died; there was a few hours warning when the Grand Teton dam failed in the USA and only a few perished. Eighteen months after Bhopal another Union Carbide plant in West Virginia, USA, had a similar accident, but not as much of the gas was released, the gas was somewhat less toxic, and few citizens were about (though some 100 were treated at hospitals). (Shortly before the accident the plant had been inspected by the Occupational and Safety and Health Administration and declared to be very safe; after the accident they returned and found it to be 'an accident waiting happen' and fined Union Carbide.) Such is the role of retrospective judgment in the accident investigations Perrow 1999.)

To have a catastrophe, then, requires a combination of such things as: a large volume of toxic or explosive material, the right wind direction or presence of a spark, a population nearby in permeable dwellings who have no warning and do not know about the toxic character of the substance, and insufficient emergency efforts from the plant. Absent any one of these conditions and the accident need not be a catastrophe. The US government, after the Union Carbide Bhopal and West Virginia accidents, calculated that there had been 17 releases in the US with the catastrophic potential of Bhopal in 20 years, but the rest of the conditions that obtained at Bhopal were not present (Shabecoff 1989). The difficulty of killing hundreds or thousands in one go may be an important reason why elites continue to populate the earth with risky systems.

A number of developments appear to have increased the number of these 'risky systems' and this may account for the attention the scheme has received. Disasters caused by humans have been with us for centuries, of course, but while many systems started out in the complex and coupled quadrant, almost all have found ways to increase their linearity and/or their loose coupling, avoiding disasters. We may find such ways to make nuclear power plants highly reliable in time, for example. But the number of risky systems has increased, their scale has increased; so has the concentration of populations adjacent to them; and in the USA more of them are in privatized systems with competitive demands to run them hotter, faster, bigger, and with more toxic and explosive ingredients, and to operate them in increasingly hostile environments. Recent entries might be global financial markets, genetic engineering, depleted uranium, and missile defenses in outer space, along with others that are only now being recognized as possibilities, such as hospital procedures, medical equipment, terrorism, and of course, software failures.

NAT distinguishes system accidents, inevitable (and thus 'normal') but rare, from the vastly more frequent component failure accidents. These could be prevented. Why do component failure accidents nevertheless occur even in systems with catastrophic potential? Three factors stand out: the role of production pressures, the role of accident investigations that are far from disinterested, and the 'socialization of risk' to the general public. The quintessential system accident occurs in the absence of production pressures; no one did anything seriously wrong, including designers, managers, and operators. The accident is

rooted in system characteristics. But the opportunities for small failures that can interact greatly increases if there are production pressures that increase the chances of small failures. These appear to be increasing in many systems, and not just in complex/tightly coupled ones, as a result of global competition, privatization, deregulated markets, and the failure of government regulatory efforts to keep up with the increase in risky systems. Accidents have been rising in petrochemical plants, for example, apparently because their growth has not included growth of unionized employees. Instead, work is contracted out to non-union contractors with inexperienced, poorly trained and poorly paid employees, and they do the most risky work at turnaround and maintenance times. The fatalities in the contractor firms are not included in safety statistics of the industry, but counted elsewhere (Kochan et al. 1994).

A second reason preventable accidents are not prevented in risky systems is the 'interested' nature of the investigations. Operators—those at the lowest level, though this includes airline pilots and officers on the bridge of ships—are generally the first to be blamed, though occasionally there is a thorough investigation that moves the blame up to the management and the design levels. If operators can be blamed then the system just needs new or better trained operators, not a thorough overhaul to change the environment in which operators are forced to work. Operators were blamed at TMI for cutting back on high pressure injection, but they were trained to do that; the possibility that 'steam voids' could send misleading information and there could be a zirconium–water interaction was not conceived by designers; indeed, the adviser to the senior official overseeing the recovery effort, the Governor of Pennsylvania, was told it would not happen. Furthermore, if conditions A and B are found to be present after the accident, these conditions are blamed for it. No one investigates those plants that had conditions A and B but did not have an accident, suggesting that while A and B may be necessary for an accident, they are sufficient; unrecognized condition C may be necessary and even sufficient, but is not noted and rectified.

A third reason for increases in accidents may be the 'socialization of risk.' A large reinsurance company found that it was making more money out of arbitraging the insurance premium it was collecting from many nations: making money by transferring the funds in the currency the premium was paid in to other currencies that were slightly more valuable. They enlarged the size of the financial staff doing the trading and cut the size of their property inspectors. The inspectors, lacking time to investigate and make adequate ratings of risk on a particular property, were encouraged to sign up overly risky properties in order to increase the volume of premiums available for arbitraging. More losses with risky properties occurred, but the losses were more than covered by the gains made in cross-national funds transfers. The public at large had to bear the cost of more fires and explosions ('socializing' the risk). Insurance companies have in the past promoted safe practices because of their own interest in not paying out claims; now some appear to make more on investing and arbitraging premiums than they do by promoting safety. Open financial markets, and the speed and ease of converting funds, appear to interact unexpectedly with plant safety.

Normal Accident Theory arose out of analyzing complex organizations and the interactions of organizations within sectors (Perrow 1986). Recent scholarship has expanded and tightened the organizational aspects of the theory of normal accidents. Scott Sagan analyzed accidents and near misses in the United States' nuclear defense system, and pointed to two aspects of NAT that needed emphasis and expansion: limited or bounded rationality, and the role of group interests (Sagan 1993).

Because risky systems encounter much uncertainty in their internal operations and their environments, they are particularly prone to the cognitive limits on rationality first explored by Herbert Simon, and elaborated by James March and others into a 'garbage can' model of organizations, where a stream of solutions and problems connect in a nearly random fashion under conditions of frequent exit and entry of personnel and difficult timing problems (March and Olsen 1979). Sagan highlights the occasions for such dynamics to produce unexpected failures that interact in virtually incomprehensible ways.

The second feature that deserved more emphasis was the role of group interests, in this case within and among the many organizations that constitute the nuclear defense system. These interests determined that training was ineffective, learning from accidents often did not occur, and lessons drawn could be counterproductive. Safety as a goal lost out to group interests, production pressures, and 'macho' values. In effect, Sagan added an additional reason as to why accidents in complex/coupled systems were inevitable. The organizational properties of bounded rationality and group interests are magnified in risky systems making normal safety efforts less effective.

A somewhat competing theory of accidents in high-risk systems, called High Reliability Theory, emphasizes training, learning from experience, and the implanting of safety goals at all levels (Roberts 1990, La Porte and Consolini 1991, Roberts 1993). Sagan systematically runs the accidents and near misses he found in the nuclear defense system by both Normal Accident Theory and High Reliability Theory and finds the latter to be wanting. Sagan has also developed NAT by exploring the curious association of system accidents with redundancies and safety devices, arguing that redundancies may do more harm than good (Sagan 1996).

NAT touched on social-psychological processes and

cognitive limits, but this important aspect of accidents was not as developed as much as the structural aspects. Building on the important work of Karl Weick, whose analysis of the Tenerife air transport disaster is a classic (Weick 1993), Scott Snook examines a friendly fire accident wherein two helicopters full of UN peacekeeping officials were shot down by two US fighters over northern Iran in 1991 (Snook 2000). The weather was clear, the helicopters were flying an announced flight plan, there had been no enemy action in the area for a year, and the fighters challenged the helicopters over radio and flew by them once for a preliminary inspection. A great many small mistakes and faulty cognitive models, combined with substantial organizational mismatches and larger system dynamics caused the accident, and the hundreds of remedial steps taken afterwards were largely irrelevant. In over 1,000 sorties, one had gone amiss. The beauty of Snooks' analysis is that he links the individual, group, and system levels systematically, using cognitive, garbage can, and NAT tools, showing how each contributes to an understanding of the other, and how all three are needed. It is hard to get the micro and the macro to be friends, but he has done it.

Lee Clarke carried the garbage can metaphor of organizational analysis further and looked at the response of a number of public and private organizations to the contamination by dioxins of a 18-story government building in Binghamton, NY (Clarke 1989). Organizations fought unproductively over the cause of the accident, the definition of risk involved, the assignment of responsibility, and control of the cleanup. While the accident was a simple component failure accident, the complexity of the organizational interactions of those who could claim a stake in the system paralleled the notion of interactive complexity, and their sometimes tight coupling led to a cascade of failures to deal with it satisfactorily. An organizational 'field' can have a system accident, as well as an organization. Clarke followed this up with an analysis of another important organizational topic related to disasters (Clarke 1999). When confronted with the need to justify risky activities for which there is no experience—evacuating Long Island in New York in the event of nuclear power plant meltdown; protecting US citizens from an all-out nuclear war; protecting sensitive waterways from massive oil spills—organizations produce 'fantasy documents' based on quite unrealistic assumptions and extrapolations from minor incidents. With help from the scientific community and organizational techniques to co-opt their own personnel, they gain acceptance from regulators, politicians, and the public to launch the uncontrollable. It is in the normative spirit of *Normal Accidents*.

Widespead remediation apparently saved us from having a world-wide normal accident when the year 2000 rolled around and many computers and embedded chips in systems might have failed, bringing about interactive errors and disasters. But even while extensive remediation saved us, something else was apparent: the world is not as tightly coupled as many of us thought. Though there were many 'Y2K' failures, they were isolated, and the failures of one small system (cash machines, credit card systems, numerous power plants, traffic lights and so on) did not interact in a catastrophic way with other failed systems. A few failures here and there need not interact in unexpected ways, especially if everyone is alert and watching for failures, as the world clearly was as a result of all the publicity and extensive testing and remediation. It was a very reassuring event for those who worry about the potential for widespread normal accidents.

One lesson is that NAT is appropriate for single systems (a nuclear plant, an airplane, or chemical plant, or part of world-wide financial transactions, or feedlots and live-stock feeding practices) that are hard-wired and thus tightly coupled. But these single systems may be loosely coupled to other systems. It is even possible that instead of hard-wired *grids* we may have a more 'organic' form of dense *webs* of relationships that overlap, parallel, and are redundant with each other, that dissolve and reform continuously, and present many alternative pathways to any goal. We may find, then, undersigned and even in some cases unanticipated alternatives to systems that failed, or pathways between and within systems that can be used. The grid view, closest to NAT, is an engineering view; the web is a sociological view. While the sociological view has been used by NAT theorists to challenge the optimism of engineers and elites about the safety of the risky systems they promulgate, a sociological view can also challenge NAT pessimists about the resiliency of large system (Perrow 1999).

Nevertheless, the policy implications of NAT are not likely to be challenged significantly by the 'web' view. While we have wrung a good bit of the accident potential out of a number of systems, such as air transport, the expansion of air travel guarantees catastrophic accidents on a monthly basis, most of them preventable but some inherent in the system. Chemical and nuclear plant accidents seem bound to increase, since we neither try hard enough to prevent them nor reduce the complexity and coupling that make some accidents 'normal' or inevitable. New threats from genetic engineering and computer crashes in an increasingly interactive world can be anticipated. Lee Clarke's work on fantasy documents shows how difficult it is to extrapolate from experience when we have new or immensely enlarged risky systems, and how tempting it is to draw ridiculous parallels in order to deceive us about safety (Clarke and Perrow 1996, Clarke 1999).

It is also important to realize how easily unwarranted fears can be stimulated when risky systems proliferate (Mazur 1998). Formulating public policy when risky systems proliferate, fears abound, production pressures increase, and the costs of accidents can be 'socialized' rather than borne by the systems, is

daunting. We can always try harder to be safe, of course, and should; even civil aviation has seen its accident rate fall, and commercial air travel is safer than being at home, and about as safe as anything risky can be. But for other systems—nuclear plants, nuclear and biological weapons, chemical plants, water transport, genetic engineering—there can be policy attention to internalizing the costs of accidents, making risk taking expensive for the system; downsizing operations (at some cost to efficiency); decoupling them (there is no engineering need for spent fuel rod storage pools to sit on top of nuclear power plants, ready to go off like radioactive sparklers with a power failure or plant malfunction); moving them away from high-population areas; and even shutting some down. The risks systems to operators may be bearable; those to users and innocent bystanders less so; those to future generations least of all.

NAT was an important first step for expanding the study of accidents from a 'operator error,' single failure, better safety, and more redundancy viewpoint that prevailed at the time *Normal Accidents* was published. It questioned all these and challenged the role of engineers, managers, and the elites that propagate risky systems. It has helped stimulate a vast literature on group processes, communications, cognition, training, downsizing, and centralization/decentralization in risky systems. Several new journals have appeared around these themes, and promising empirical studies are appearing, including one that operationalizes effectively complexity and coupling for chemical plants and supports and even extends NAT (Wolf and Berniker 1999). But we have yet to look at the other side of systems: their resiliency, not in the engineering sense of backups or redundancies, but in the sociological sense of a 'web-like' interdependency with multiple paths discovered by operators (even customers) but not planned by engineers. NAT, by conceptualizing a system and emphasizing systems terms such as interdependency and coupling and incomprehensibility, and above all, the role of uncertainty, should help us see this other, more positive side.

See also: Islam and Gender; Organizational Behavior, Psychology of; Organizational Culture, Anthropology of; Risk, Sociological Study of; Risk, Sociology and Politics of

Bibliography

Clarke L 1989 *Acceptable Risk? Making Decisions in a Toxic Environment*. University of California Press, Berkeley, CA

Clarke L 1999 *Mission Improbable: Using Fantasy Documents to Tame Disaster*. University of Chicago Press, Chicago

Clarke L, Perrow C 1996 Prosaic organizational failure. *American Behavioral Scientist* **39**(8): 1040–56

Kochan T A, Smith M, Wells J C, Rebitzer J B 1994 Human resource strategies and contingent workers: The case of safety and health in the petrochemical industry. *Human Resource Management* **33**(1): 55–77

La Porte T R, Consolini P M 1991 Working in practice but not in theory. *Journal of Public Administration Research and Theory* **1**: 19–47

March J G, Olsen J P 1979 *Ambiguity and Choice in Organizations*. Universitesforleget, Bergen, Norway

Mazur A 1998 *A Hazardous Inquiry: The Rashomon Effect a Love Canal*. Harvard University Press, Cambridge, MA

Perrow C 1984 *Normal Accidents*. Basic Books, New York

Perrow C 1986 *Complex Organizations: A Critical Essay*. McGraw Hill, New York

Perrow C 1999 *Normal Accidents with an Afterword and Postscript on Y2K*. Princeton University Press, Princeton, NJ

Roberts K 1993 *New Challenges to Understanding Organizations*. Macmillan, New York

Roberts K H 1990 Some characteristics of one type of high-reliability organization. *Organization Science* **1**: 160–76

Sagan S D 1993 *Limits of Safety: Organizations Accidents, and Nuclear weapons*. Princeton University Press, Princeton NJ

Sagan S D 1996 *When Redundancy Backfires: Why Organizations Try Harder and Fail More Often*. American Political Science Association Annual Meeting, San Francisco, CA

Shabecoff P 1989 Bhopal disaster rivals 17 in US. *New York Times*, New York

Snook S 2000 *Friendly Fire: The Accidental Shootdown of US Black Hawks Over Northern Iraq*. Princeton University Press, Princeton, NJ

Weick K E 1993 The vulnerable system: An analysis of the Tenerife air disaster. In: Roberts K (ed.) *New Challenges to Understanding Organizations*. Macmillan, New York, pp. 73–98

Wolf F Berniker E (1999) Complexity and tight coupling: A test of Perrow's taxonomy in the petroleum industry. *Journal of Operations Management*

Wolf F 2001 Operationalizing and testing accident theory in petrochemical plants and refineries. *Production and Operations Management* (in press)

C. Perrow

Accountability: Political

Political accountability is the principle that governmental decision-makers in a democracy ought to be answerable to the people for their actions. The modern doctrine owes its origins to the development of institutions of representative democracy in the eighteenth century. Popular election of public officials and relatively short terms of office were intended to give the electorate the opportunity to hold their representatives to account for their behavior in office. Those whose behavior was found wanting could be punished by their constituents at the next election. Thus, the concept of accountability implies more than merely the tacit consent of the governed. It implies both

mechanisms for the active monitoring of public officials and the means for enforcing public expectations.

1. Accountability and Responsibility

When the doctrines and institutions of representative democracy were originally developed, the term most commonly used to capture what we now mean by accountability was 'responsibility.' As the editor of a standard edition of *The Federalist Papers* notes, '[r]esponsibility is a new word that received its classic definition in the ratification debate [over the proposed Constitution of 1787] and, especially, in the pages of *The Federalist*. Although the term had appeared sporadically in eighteenth-century British politics, it was in America in the 1780s that it achieved its lasting political prominence' (Hamilton et al. 1999, p. xxii). In *The Federalist*, however, 'responsibility' carried several meanings, only one of which is synonymous with 'accountability.' The virtual replacement of the broader term by the narrower one in modern political discourse is indicative of powerful trends in democratic thought.

In his essays on the presidency Alexander Hamilton described responsibility as equivalent to accountability. A plural executive, he argued, tends 'to conceal faults and destroy responsibility' because the people do not know whom to blame for misconduct or poor stewardship of the affairs of state. Hamilton contrasted England, where the king was legally 'unaccountable for his administration' with 'a republic where every magistrate ought to be personally responsible for his behavior in office.' The American president was responsible (i.e., accountable) to the people through the electoral process and, in serious cases of misconduct, through impeachment, conviction, and removal (Hamilton et al. 1999, pp. 395–97).

Yet *The Federalist* also includes a somewhat broader and more subtle understanding of responsibility. 'Responsibility', James Madison wrote in an essay on the proposed Congress, 'in order to be reasonable, must be limited to objects within the power of the responsible party, and in order to be effectual, must relate to operations of that power, of which a ready and proper judgment can be formed by the constituents.' Here Madison distinguished governmental measures 'which have singly an immediate and sensible operation' from others that depend 'on a succession of well-chosen and well-connected measures, which have a gradual and perhaps unobserved operation.' A 'reasonable' understanding of governmental responsibility would recognize the value of legislators exercising their discretion and judgment in promoting the long-term well being of the country. Put differently, legislators act responsibly when they behave in this way, despite the fact that their constituents may not recognize the value of their acts, at least for some time (Hamilton et al. 1999, pp. 351–2).

Hamilton and Madison each followed these discussions with powerful statements that when the people push for prejudiced, irresponsible, or unjust measures, their elected officials have a 'duty' to resist the popular desires until reason can regain its hold on the people (Hamilton et al. 1999, pp. 352, 400). Public officials have a responsibility at times to act against public opinion in the interest of the public good. This notion of responsibility goes well beyond accountability, at least as commonly understood. As one modern commentator notes, responsibility in this broader sense 'is concerned with results.' The responsible public official 'takes care that the results are correct 'and 'good for many. Such responsibility is the most interesting kind because it goes beyond questions of accountability and obligation in any simple meaning' (Blitz 1998). It follows that responsibility as accountability may at times conflict with responsibility in the broader sense of acting to promote the good of many.

2. The Accountability of Elected Officials

Under the first constitution of the USA, the Articles of Confederation (1781–1788), the delegates to the Congress were appointed by the state legislatures for one-year terms. Under Article V, each state retained the authority 'to recall its delegates, or any of them, at any time within the year.' Thus, the nation's legislators were entirely accountable to those who appointed them. Under the Constitution of 1787, however, neither the popularly elected members of the House of Representatives with their two-year terms nor the members of the Senate, elected by the state legislatures for six-year terms, were recallable between elections. Here the Constitution's framers sacrificed some amount of accountability in order to promote the independent judgment and the collective deliberation of legislators. Secure in office for either two or six years, lawmakers could act for the public good as they came to understand it without the fear that unpopular actions would result in their immediate dismissal. Because of the difference in the length of terms, this principle was expected to apply with much greater force to senators than to representatives.

As early as the First Congress (1789–1791) under the new Constitution, some in the House and Senate sought to remedy what they took to be deficiencies in the accountability of national lawmakers. Members of the House, for example, proposed an amendment to the Constitution that would guarantee the right of the people 'to instruct their representatives', thereby binding them to a certain course of action. Representative Elbridge Gerry of Massachusetts argued that it was 'absurd to the last degree' to say that 'sovereignty resides in the people' but to deny the right of the people 'to instruct and control their Representatives.' But others, such as Representative Thomas Hartley of Pennsylvania, argued that 'the great end of

meeting [in a legislature] is to consult for the common good' and thus to transcend 'local or partial view[s].' The proposed amendment was defeated by a 4–1 margin.

In the new Senate the proponents of greater accountability argued that under a proper understanding of the Constitution senators were already obliged to follow instructions from the state legislatures that appointed them. Indeed, from the very beginning some of these legislatures instructed their senators as to how to vote on specific bills. Although this practice continued for some decades, there was no way for state legislatures to enforce compliance with their instructions. At best they could refuse to reappoint a senator when his term expired. The adoption in 1913 of the direct popular election of senators (Seventeenth Amendment) effectively ended any issue of accountability to state legislatures, while it also promoted the accountability of senators to the people of their states.

It was in the 1960s and 1970s that public pressure to make Congress more accountable for its behavior reached its peak. The institution responded with a variety of 'government in the sunshine' reforms, including: (a) the opening up of nearly all committee meetings to public scrutiny; (b) the televising of floor debates in the House and Senate; and (c) new requirements that most votes in committee and on the floor be recorded and made available to the public. Although these changes have been popular, some members of the institution as well as some scholars have questioned whether greater accountability has been good for legislative deliberation within Congress. In the early 1980s, for example, after a decade of conducting its mark-up (bill drafting) sessions in public, the tax-writing House Ways and Means Committee returned to closed sessions. Committee members had become convinced that greater accountability to constituents and interest groups had made it increasingly difficult for the members to take actions that imposed costs on their supporters, however conducive such measures might be to the broader public good. In the language of *The Federalist*, these legislators promoted responsible behavior, even if this came at the cost of some direct accountability.

3. *The Accountability of the Bureaucracy*

The issue of political accountability is not limited to elected officials. Modern democratic states are characterized by large bureaucracies whose members have no direct electoral link to the populace. How are these thousands of individuals, whose actions directly affect the citizenry in a myriad of ways, held to account for what they do? In the USA President Andrew Jackson (1829–37) and Senator Daniel Webster engaged in a classic debate on this issue. Jackson's position was that the president himself was 'responsible for the entire action of the executive department.' Executive officials, 'mere instrument[s] of the Chief Magistrate in the execution of the laws', were accountable to the nation for their behavior through the combination of their subservience to the president and his direct accountability to the people. The president, Jackson maintained, is 'accountable at the bar of public opinion for every act of his Administration.' Webster, by contrast, decried this 'undefined, undefinable, ideal responsibility to the public judgment.' Executive branch officials were not primarily 'the President's agents' but 'agents of the laws.' It was the law that 'define[d] and limit[ed] official authority', that 'assign[ed] particular duties to particular public servants', and that 'define[d] those duties.' And it is Congress, of course, that makes the law.

In practice the movement for a more accountable bureaucracy has traveled down these two distinct paths. One path is the enhancement of presidential control of the bureaucracy through (a) the centralization of executive branch budget making; (b) White House clearance of legislative proposals; and (c) the development of personnel policies to increase the influence of the president's appointees over career civil servants. The other is the formalization of Congress's oversight function, stemming from the Legislative Reorganization Act of 1946 which required that 'each standing committee exercise continuous watchfulness of the execution [of the laws] by the administrative agencies.' One of the ironies of this movement to enhance governmental accountability is that even as Congress has formalized and expanded its oversight activities, it has increasingly delegated policy-making responsibilities to the executive branch. Scholars have argued that the members of legislative institutions have strong political incentives to create new bureaucracies and empower them to make the controversial policy decisions. The bureaucrats 'make the hard choices' and the lawmakers 'disclaim any responsibility for harm done' (Schoenbrod 1993, p. 8, Fiorina 1989, pp. 40–7). When constituents complain, lawmakers eagerly intervene on their behalf with the bureaucrats. In this way Congress undermines genuine democratic accountability while giving the appearance of increasing bureaucratic accountability.

4. *Accountability and Other Goals*

The modern movement for greater political accountability has led to such reforms as requirements that governmental agencies do their business in public, that the public have access to government information, and that the public be given formal opportunities to testify or comment on proposed administrative rules and regulations. More than ever before modern democratic governments are open to the scrutiny of the media, of interest groups, and of the broader public. The tendency throughout has been to view and pursue accountability as an end in itself, as an

unmitigated good. Yet while accountability is necessary to ensure that democratic government is faithful to the interests of those it serves, it is also in some tension with other conditions and values of sound governance, such as the exercise of informed discretion by decision-makers and the promotion of sound deliberation about common ends.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Delegation of Power: Agency Theory; Political Representation; Representation: History of the Problem; Responsibility: Philosophical Aspects

Bibliography

Aberbach J D 1990 *Keeping a Watchful Eye: The Politics of Congressional Oversight*. Brookings Institution, Washington, DC

Arnold R D 1990 *The Logic of Congressional Action*. Yale University Press, New Haven, CT

Blitz M 1998 Responsibility and Public Service. In: Lawler P A, Schaefer R M, Schaefer D L (eds.) *Active Duty: Public Administration as Democratic Statesmanship*. Rowman and Littlefield, Lanham, MD

Burke E 1774 Speech to the electors of Bristol. In: *The Writings and Speeches of Edmund Burke*, *II*. Little, Brown and Company, Boston, pp. 89–98

Fiorina M P 1989 *Congress: Keystone of the Washington Establishment*, 2nd edn. Yale University Press, New Haven, CT

Friedrich C J (ed.) 1960 Nomos 3: Responsibility. *Yearbook of the American Society of Political and Legal Philosophy*. Liberal Arts Press, New York

Hamilton A, Madison J, Jay J 1999 *The Federalist Papers*. Rossiter C (ed.) with a new Introduction and Notes by Kesler C. Mentor Book, New York

Light P C 1993 *Monitoring Government: Inspectors General and the Search for Accountability*. Brookings Institution, Washington, DC

Schoenbord D 1993 *Power without Responsibility: How Congress Abuses the People through Delegation*. Yale University Press, New Haven, CT

J. M. Bessette

Action Planning, Psychology of

1. Basic Concepts

1.1 Action

The concept of action refers to *the intended behavior of an agent*. Different prototypical types of actions can be distinguished: 'goal-directed action' aims at the attainment of an end state (such as repairing a bicycle), and to a great extent is under cognitive control. 'Intuitive actions' are more or less spontaneously performed without much conscious thought and awareness (e.g., many acts in face-to-face social interaction). In experience- (or process-) oriented actions' it is the performance process itself, and not the end state, that is important (such as skiing or dancing), and they may produce the experience of 'flow' (Csikszentmihalyi 1990). 'Acts committed in the heat of passion' are not under full cognitive control, but seem to erupt in states of strong affective pressures (as is the case in many violent crimes). Finally, there are long-term 'projects' (such as constructing a house or achieving an academic degree) which are composed of many consecutive actions of various forms. These types of actions seem to correspond to social prototypes, in the sense that cultural concepts of them exist. However, pure forms of action prototypes are rarely observed because most actions in real life contain elements of several prototypes. In this article, we concentrate on *goal-directed action of human individuals* since the features of 'planning' are most important and clear in this context.

Most researchers in the field of action consider it as characteristically human; therefore, basic assumptions about the human nature have led to different concepts of action, with different authors and schools emphasizing different aspects. Some emphasize the commonsense nature of an action (Heider 1967); others conceive it as an ex-post interpretation of behavior (Lenk 1978); still others concentrate on its logical nature (Smedslund 1997) or emphasize its symbolic meaning in a cultural context (Boesch 1991). Approaches that discuss action from a motivational and a regulation perspective have received the most attention in the field and have elicited the most empirical work. Both approaches refer to goal-directed action, seeing it as events in a systemic context and emphasize its cognitive, if not rational, nature. The motivational approach addresses the problem of how a person decides, among the many possibilities, on a *specific goal or action*, and how, once chosen, an action is maintained and *carried through* (Heckhausen 1991). The regulation approach concentrates on the question of *how* the attainment of a given goal is achieved (Cranach et al. 1982, Hacker 1998). Starting with these two approaches, we add the notion that individual action tends also to be socially steered and controlled.

As a summary, we define action as the behavior of an individual human agent (or actor) which is directed and (at least partly) consciously aspired, wanted, planned, and steered in order to achieve a specific goal.

1.2 Planning

In order to attain a goal, an actor needs energy, so the mind–body system can be set in motion. He or she also

needs direction, or steering, towards the intended end-state. We therefore distinguish between *energizing processes* and *steering processes*. Mental processes can serve either one, or both, functions. Thus, we distinguish between 'decision' and 'resolve' (William James' 'fiat'). The first refers to the choice between alternatives and has a steering function, whereas the second energizes the action as an initiating command.

In a broad sense, the term 'planning' is used with regard to all mental activities that serve action steering (e.g. Miller et al. 1960). For example, this is the way the term 'plan' is used in the method of 'plan–analysis' in psychotherapy (Caspar 1995). In this case, planning contains elements of goals *and* elements of the means to attain the goals. In the context of action-regulation theory and some related approaches, planning is used in a narrower sense and is seen as one part of a more or less ordered action steering cycle composed of anticipatory cognitive representations of the operations, steps, rules, and procedures of goal attainment. Planning therefore is the *program of the course of action*.

2. Planning as 'Steering'

2.1 Expectancy Theories of Behavior

The basic idea that an action should be executed if its expected results are of high value and if the action appears to be a likely mean to realize them was formulated by Blaise Pascal and David Bernoulli in the seventeenth and eighteenth centuries. However, its psychological formulation is based on Kurt Lewin's (1951) field theory. It has been further developed by John W. Atkinson, Heinz Heckhausen and many others (see Heckhausen 1991). The expectancy theories of motivation, which are one of the most influential branches of motivation theory, have been differentiated in many details and have led to a great variety of experimental studies. Their assumptions explain how people choose a specific action among several possibilities: (a) in the course of their life, people develop (and adopt) values that may be formulated as goals; (b) certain situations contain hints ('incentives') about the possibility of realizing a goal through an action; (c) the incentive is related to a certain *expectancy* that the goal can be attained; and (d) if the value-based incentive and the expectancy are strong enough, an *intention* is formed to realize the action (in a suitable situation). For example, if a person values energy saving and believes that insulating windows helps to achieve this goal, he or she should be inclined to install better windows in his or her house. This normally means that some other possible personal goals may have to be sacrificed.

On the basis of expectancy and value assumptions, the 'theory of planned behavior' (Ajzen 1991), which is a development of the 'theory of reasoned action' by Marty Fishbein and Icek Ajzen, tries to explain the influence of attitudes on behavior. Value and expectancy produce an *attitude* with regard to a specific behavior. However, even a positive attitude is normally not enough to elicit an action. The attitude becomes an *intention* to act, when two other conditions are satisfied: first, the person must hold a normative belief which conforms to the action, and he or she must be motivated to comply to that norm ('*subjective norm*'). Second, the person must perceive that he or she can in fact execute the action and is able to reach the goal ('*perceived behavior control*'). In our example, the house-owner knows that insulated windows conform to the standards of the neighborhood, and that he or she has enough money, time, organizational talent, and persistence to realize this project.

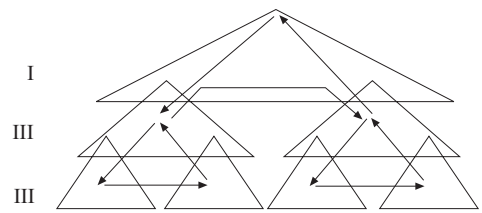

Figure 1
The hierarchical–sequential organization of the action

2.2 Action Regulation Theory

The decision to purchase new windows is only the beginning of the action. The house-owner must now decide what kind of windows to choose, which contractor to commission with the job, and when and how to install the new windows. A host of complicated cognitive operations and their execution (not to speak of their emotional aspects and the necessary volitional acts) are required. These processes are treated by *the action regulation theory* (for a more comprehensive description, see Frese and Zapf 1994) that has been developed in the context of work psychology (e.g., Hacker 1998) and in social psychology (Cranach et al. 1982). Psychological action theory is based on assumptions from *activity theory*, developed by Russian psychologists (e.g., S. L. Rubinstein and A. L. Leontjew) combined with a *cybernetic approach* (e.g., Miller et al. 1960).

Actions are seen as goal-directed *units of work activity* that are consciously and voluntarily steered. Actions are hierarchically and sequentially organized. The hierarchical organization refers to the nested units of goals, subgoals, and subsubgoals of an action. In order to attain a goal, the related subgoals must be attained before the agent can proceed to the next superordinate unit (the sequential characteristic of action) (Fig. 1). The model includes information transfer from lower to higher levels of action, but also

on the horizontal axis, and therefore represents a weak form of hierarchy. In order to facilitate the application of the model, various authors have introduced the concept of *levels of regulation* which contains the notion of a quasistandardized hierarchy. Units on a given level are assumed to have certain features in common. In his original model, Hacker (1998) proposed the 'intellectual,' the 'perceptive conceptual,' and the 'sensorimotor' levels of action regulation, each level having different functions and qualities. All levels of action regulation include steering activities together with the actual action execution. Higher-level processes are more comprehensive and more likely to require conscious monitoring of the operations whereas lower-level processes are more likely to become automatic.

The sequential dimension of an action has been further differentiated, and cyclical regulation processes have been described. On each level, successful action requires a prototypical cycle of regulatory functions: A cycle starts with either *goal determination* or *situational orientation*; the latter contains the two components of *orientation about the environment* and *orientation about the agent's own state*. After a goal has been determined, the *development of an action plan* follows. These components are integrated into a relatively stable cognitive representation, known as the *operative representation system* (Miller et al.'s 'image'), which directs the *execution*. Action execution is constantly monitored with regards to deviations from plans and goal attainment (*execution control*). The regulatory cycle is accomplished with the *consumption of the result* and a *final evaluation of the action*. If we return to our example: after deciding to install new windows (goal determination), our agent seeks information about the prices of different windows on the market, possible contractors, as well as his or her own budget (*orientation about environment and own state*). He or she must then decide what windows to purchase and which contractor to entrust with the work. The next step requires that the details of the procedure to replace the old windows must be worked out (*plan development*). Execution of the work follows according to the plan in several steps, each consisting of a number of sub- and subsubtasks, which must be worked through in a specific sequential order (first, take out the old windows before installing the new ones). It has been found that actions which follow an ideal cyclical order tend to be more successful.

3. *Planning in the Narrower Sense*

Besides the general notion of behavior regulation and steering of behavior, planning can also be considered as the anticipatory, cognitive construction of the action and its steps, that is, the cognitive *development of the action program*. Steering in its prototypical form includes planning in this sense as one of the important regulatory activities. Planning in the narrower sense is basically a behavior stream that is imagined and thought, but not executed. Planning consists of conditions, possible executions, and possible outcomes of an action (Dörner 1990). As conditions can vary, several plans can be made with regard to one outcome. Planning can start at the status quo (when I cover furniture before installing the windows), but also start from the end state (when will the contractor come) and be recursive. Backwards planning has been found to be less frequent, but more effective in certain types of problem solving. (Dörner 1990).

Planning seems to have obvious benefits. It could therefore be assumed that human actors normally carefully plan ahead before acting. However, individuals and groups often do not do so or they show satisfying tendencies (Simon 1957), in that they elaborate only very rudimentary action plans. In spite of this, they somehow reach their goals. In this context, the following questions are important: what is the advantage of planning, and when is planning necessary?

(a) The possibility of elaborating a complete plan before acting depends very much on the information available. If the situation is very clear from the beginning and complete information is available, the elaboration of a complete plan or action program is possible. If the situation is dynamic and new information is acquired during the action, *conditional planning* or *rolling planning* is more indicated.

(b) *Different kinds of tasks require different amounts of planning*. For example, in sports, a 5,000 m race requires a lot of strategic and adaptive planning from all of the competitors. In contrast, the execution of a pole-vault requires that the essential activities are automated to a high degree in order to lead to a smooth and successful execution. The essential parts of the activity must therefore be trained and practiced in order not to require planning; conscious monitoring could even disturb the perfect execution.

(c) *Automation of behavior execution*, which involves the absence of conscious and detailed planning, is normal when an action is frequently executed. This is especially true for the lower levels of regulation. An obvious example of this is talking, where we may plan on a general level what to say, but the details of composing the sentences, and even more so the phonological production, are highly automated and not consciously controlled. However, conscious monitoring of normally automated parts (the so-called 'emergency function of consciousness') will often be resumed if difficulties arise—for example, we perceive and correct speech slips. Automatic actions have been characterized as situation-specific, integrating different operations, requiring less feedback and fewer decisions and leading to more parsimonious movements (Semmer and Frese 1985).

(d) For many situations *preformed plans exist*, even on a societal level. For example, there exists a widely

shared general sequence of action steps, a 'script' (Schank and Abelson 1977), of how to eat in a restaurant, containing a general plan that can be specified and adapted for a specific situation. If 'scripts' exist, planning effort is reduced. People also tend to establish *personal scripts by planning on stock, for example they imagine or fantasize about situations* and behaviors during activities that do not demand their full attention (for example, they may plan the installation of the windows during a bus ride).

(e) There is great variation in the individual inclination and capability to plan. Personality dimensions like 'action versus state orientation' (Kuhl and Beckmann 1994) or 'action styles' like planfulness (Frese et al. 1987) are important indicators of the amount of planning done by individuals.

Taking all these complications into account, there is still a lot of evidence that planning is strongly related to successful action. Studies of experts (Ericsson and Lehmann 1996) and very successful workers ('superworkers') show that *intellectual penetration* of the task (Hacker 1996) and the development of more complete operative image systems that serve as a more complete basis for planning are crucial for expert performance. In addition, the study of thought processes during computer-simulation tasks (Dörner 1990) and of the action-related communication of successful teams (Tschan and von Cranach 1996) lead to similar conclusions.

4. *Inefficient Planning and Planning Mistakes*

Even if, generally speaking, planning is a necessary prerequisite to many actions and leads to more successful goal attainment, planning can be inefficient or may even fail.

Besides the planning inefficiencies that may occur if the informational basis is too small or if the planning is based on false assumptions or contains illogical or wrong 'if–then' relations, planning that is too detailed or too general may be inefficient. Too detailed planning, often done to reduce insecurity, may lead the person to get stuck and hinder execution. Another disadvantage is the concentration on details and specific aspects of a more complex task and the neglect of other important aspects. Plans that are too general may fail to include important specifications and conditions to their execution or they may be incomplete in other respects, and are therefore difficult to carry out (Dörner 1990). In dealing with complex, dynamic and non-transparent situations, typical planning inefficiencies include the misjudgment of the importance of certain information, which narrows down the searching space and misses valuable information, the neglect of side-effects of planned actions, and the failure to perceive possible delayed effects which may produce unforeseen, unwanted secondary results.

Error research distinguishes between slips (errors in action execution) and mistakes (errors in the intention). The latter include planning mistakes. Mistakes of steering activities have been conceptualized in a hierarchical model which corresponds to the levels of regulation described above (Reason 1990; for another important classification, see Frese and Zapf 1994). The model distinguishes *knowledge-based* regulation, where new plans are elaborated. Failures in the elaboration of correct plans on this level include limitations of workspace, limitations of handling complexity, and numerous biases, e.g., illusory correlation, halo effect, and confirmation bias. On the *rule-based* level, people faced with a new situation choose from a 'pool of available plans' or rules and apply them to a new situation. Mistakes on this level include the misapplication of a good rule or plan to a new situation and the application of a wrong plan to a specific situation. Reason (1990) summarizes the main mechanisms that produce errors as *similarity matching* (looking for and applying something that is already known) and *frequency gambling* (using knowledge which has already been frequently used) (see also Kahnemann et al. 1982).

See also: Action Theory: Psychological; Activity Theory: Psychological; Attitudes and Behavior; Motivation and Actions, Psychology of

Bibliography

Ajzen I 1991 The theory of planned behaviour. *Organizational Behavior and Human Decision Processes* **50**: 179–211

Boesch E E 1991 *Symbolic Action Theory and Cultural Psychology*. Springer-Verlag, Berlin

Caspar F 1995 *Plan Analysis. Towards Optimizing Psychotherapy*. Hogrefe, Seattle, WA

von Cranach M, Kalbermantten U, Indermühle K, Gugler B 1982 *Goal Directed Action*. Academic Press, London

Csikszentmihalyi M 1990 *Flow: The Psychology of Optimal Experience*. Harper & Row, New York

Dörner D 1990 The logic of failure. *Philosophical Transactions of the Royal Society of London, Series B* **327**: 463–73

Ericsson K A, Lehmann A C 1996 Expert and exceptional performance: evidence of maximal adaptation to task constraints. *Annual Review of Psychology* **47**: 273–305

Frese M, Stewart J, Hannover B 1987 Goal orientation and planfulness: action styles as personality concepts. *Journal of Personality and Social Psychology* **52**: 1182–94

Frese M, Zapf D 1994 Action as the core of work psychology: a German approach. In: Triandis H C, Dunnette M D, Hough L M (eds.) *Handbook of Industrial and Organizational Psychology*. Consulting Psychologists Press, Palo Alto, CA, Vol. 4, pp. 271–340

Hacker W 1996 *Diagnose von Expertenwissen. Von Abzapf- (Broaching-) zu Aufbau- Re-construction- Konzepten*. Akademie Verlag, Berlin

Hacker W 1998 *Allgemeine Arbeitspsychologie. Psychische Regulation von Arbeitstätigkeiten*. Hans Huber, Berne

Heckhausen H 1991 *Motivation and Action*. Springer, Berlin

Heider F 1967 *The Psychology of Interpersonal Relations*. Science Editions, New York
Kahnemann D, Slovic P, Tversky A 1982 *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK
Kuhl J, Beckmann J 1994 *Volition and Personality: Action Versus State Orientation*. Hogrefe & Hober publishers, Göttingen Germany
Lenk H 1978 Handlung als Interpretationskonstrukt. Entwurf einer konstituenten- und beschreibungstheoretischen Handlungsphilosophie. In: Lenk H (ed.) *Handlungstheorien Interdisziplinär*. W Fink, Munich, Germany
Lewin K 1951 *Field Theory in Social Science; Selected Theoretical Papers*, 1st edn. Harper & Row, New York
Miller G A, Galanter E, Pribram K H 1960 *Plans and the Structure of Behavior*. Holt, New York
Reason J 1990 *Human Error*. Cambridge University Press, Cambridge, UK
Schank R C, Abelson R P 1977 *Scripts, Plans, Goals and Understanding*. Erlbaum, Hillsdale, NJ
Semmer N, Frese M 1985 Action theory in clinical psychology. In: Frese M, Sabini J (eds.) *Goal Directed Behavior: the Concept of Action in Psychology*. Erlbaum, Hillsdale, NJ, pp. 296–310
Simon A A 1957 *Models of Man*. Wiley, New York
Smedslund J 1997 *The Logical Structure of Psychological Common Sense*. Erlbaum, London
Tschan F, von Cranach M 1996 Group task structure, processes and outcome. In: West M (ed.) *Handbook of Work Group Psychology*. Wiley, Chichester, UK, pp. 95–121

M. von Cranach and F. Tschan

Action Theory: Psychological

Action theory is not a formalized and unitary theory agreed upon by the scientific community, but rather a unique perspective, narrative or paradigm. Although this perspective varied in saliency during the history of psychology, it has been in existence since the very beginning of psychology in the nineteenth century both in Europe and North America. In Germany, Brentano, a teacher of Freud's, focused 1874 on intentionality as a basic feature of consciousness leading to the concept of 'acts of consciousness.' Ten years later, Dilthey distinguished between an explanation of nature and an understanding of the mind/soul, a dichotomy, which paved the way for the ongoing discourse on the dichotomy of explanation and understanding. In 1920 Stern criticized the mainstream psychology of his time because it neglected *intentionality* and also *cultural change* as a created framework for human development. In Paris, Janet wrote his dissertation about '*Automatisme*' in 1889. This was the beginning of an elaborated action theoretical system of neuroses (Schwartz 1951). In North America James developed a sophisticated theory of action at the end of the nineteenth century that anticipated a remarkable amount of action theory concepts (Barbalet 1997). Münsterberg, a disciple of Wundt, proposed action as the basic unit of psychology instead of sensations at the turn of the century. These early traditions were overruled by the neopositivistic logic of explanation expounded by the Vienna circle in philosophy and behaviorism in psychology. They were taken up in philosophy by Wittgenstein's 'language games' that are different in natural science and the humanities. In psychology action theory terms have increased in importance again during the 1960s to the 1990s.

In fact, in recent times human action (or aspects of it) is taken as a framework for analysis and/or research in many branches of psychology. This is true for 'basic science': in theories of motivation (e.g., Gollwitzer 1990), problem solving (e.g., Dörner and Kaminski 1988), ontogenetic development (e.g., Oppenheimer and Valsiner 1991), social psychology (von Cranach 1991), and particularly in cultural psychology (Boesch 1991). And it is true for 'applied domains': in clinical psychology (Schwartz 1951), educational psychology (Bruner 1996), organizational psychology or psychology of work (Hacker 1986), and sport-psychology (Kaminski 1982). Under an action theory perspective the boundaries between these domains become fuzzy. Cultural psychology, for instance, becomes an integrated enterprise, which is developmental as well as cognitive, affective, and motivational (cf. Boesch 1991; Cole 1990).

Beyond this diversity of action based theories in psychology, human action is also focused upon in other human sciences. It is particularly reflected upon in philosophy (Lenk 1984), and has a long tradition in sociology (Parsons 1937/1968, Weber 1904) and anthropology (Shweder 1990).

Finally, a 'second tradition' exists, basically equivalent to action theories, and with partly the same roots in Janet's work (Eckensberger 1995): the Russian activity theory in the tradition of Vygotsky, Luria, Leontiev, its most famous representatives. In the US this framework is particularly elaborated and applied by Cole, Rogoff, Valsiner, Wertsch; in Germany by Holzkamp.

1. *Attributes of Actions*

Considering the breadth of action theoretical frameworks, it is not surprising that the issues studied are not identical and that the terminology is not coherent or fully agreed upon by different authors or traditions.

1.1 *Action as an Analytical Unit*

From an analytical perspective it appears necessary to note that: To act does not mean to behave (although some authors consider actions as a particular subtype

of behavior); To speak of an action instead of behavior implies the following features:

(a) *Intentionality*: Broadly speaking, it means that sentences, symbols, but also mental states refer to something in the world (Searle 1980). Intentionality therefore occurs, if a subject (called *agency*) refers to the world. Agencies refer to the world by *acting* with reference to the world, by *experiencing* it (they think, feel, perceive, imagine, etc.) and by speaking about it: The latter is called '*speech act*.' An intentional state thus implies a particular content and a psychic mode (a subject can think that it rains, wish that it rains, *claim* that it rains, etc., where rain is the content, thinking, wishing, and claiming are modes). The intent of the action is the intentional state of an action; the intended consequence or goal is content. This implies what is also called 'futurity' (Barbalet 1997) or future orientation of an action. Although some try to explain actions by interpreting these intentions as causes of actions there is agreement at present that actions cannot epistemologically be explained by (efficient material) causes, but have to be understood in terms of their reasons (cf. Habermas 1984; von Wright 1971). This leads to serious problems, if psychology is understood as a natural science, which basically interprets events in terms of causes. It follows that aections are not necessarily observable from the outside. If they are, one also uses the term '*doing*' (Groeben 1986). However, allowing something to happen as well as refraining from doing something are also actions (von Wright 1971).

(b) *Control over the action*: It is assumed that action involves the free choice to do something (A or B), to let something happen, or to refrain from doing something. This condition is strongly related to the (subjective perception) of free will. Although the control aspect is sometimes also expanded to include the intended effects of an action, these two aspects should be distinguished, because the effects of an action can be beyond the control of the agency, although the decision to act itself was controlled.

(c) The *basic structure* of an action which aims at some effect (to bring something about) is the following: Analytically, it is assumed that the means applied to carry out an action follow rationally from the intentions, i.e., they can be justified or made plausible by the agency. They are chosen on the basis of finality (in order to reach a goal). If they are applied, the result is some change, and this change leads causally to some consequences. Those consequences of actions that represent the goal are intended, others are unintended. To let fresh air into a room (intended goal) one opens the window (does something), after having opened the window it is open (result) and lets fresh air in (intended), but the room may become cold (unintended).

(d) So actions in principle are *conscious activities* of an *agency*. The agency can reflect upon (a) their actions as well as (b) upon themself as an agency. This is why Eckensberger (1979) proposed interpreting action theories as a 'theory family' based upon the self-reflective subject or agency. This position is related to the basic issue of whether or not homo sapiens has a special position in nature, because this species is the only one that can decide not to follow natural laws. Once more this poses a serious problem for psychology as a natural science.

(e) There are *different types of actions*: (i) If directed at the physical/material world, and aimed at bringing about some *effect* (also 'letting things happen': von Wright 1971) or suppressing some effect, they are called 'instrumental actions'; (ii) But if actions are directed at the social world, i.e., at another agency B, they cannot (causally) 'bring something about' in B, but have to be coordinated with B's intentions. Therefore, agency B's intentions have to be understood (interpreted by agency A). This presupposes a communicative attitude (Habermas 1984). This type of action is consequently called a 'communicative action.' If this orientation not only implies understanding B, but also respecting B's intentions, this is clearly a moral action. If B's intentions are simply used for A's benefit it is a strategic action (Habermas 1984). Interestingly, in non-Western philosophies/religions (Hinduism, Buddhism, Confucianism) this 'adaptive attitude' and respect for the 'non-A' is extended to include the plant and animal world. So one may distinguish between two action types which aim at A's control of the environment (instrumental and strategic), and two action types which aim at harmonizing A with the environment (communicative and adaptive actions).

Although an agency is in principle considered autonomous, actions are not arbitrary but follow rules (of prudence as well as of social/cultural conventions and/or expectations). This tension between autonomy and heteronomy is basic to all action theories that also focus on the social/cultural context of actions (cf. Parsons 1937/1968). One tries to resolve this tension, however, by assuming that cultural rules and their alternation are also man-made, although the implied intentionality of cultural rules/norms may 'get lost' in time. In principle within this theoretical frame an action links the actor and his/her environment (see James 1897/1956), and cultures are considered intentional worlds (Shweder 1990) or action fields (Eckensberger 1995, Boesch 1991).

1.2 Actions as Empirical Units

The most recent and comprehensive review of action related to research in (developmental) psychology is given by Brandtstädter (1998):

1.2.1 Hierarchy of goals. There is considerable agreement among researchers that empirically actions do not just have one goal but many. They can

be seen as forming a chain or a hierarchy. To read an article may have the goal of understanding a particular problem and may be considered an action. Reading individual characters on paper may be taken as subactions or elements of an action (called 'actemes' by Boesch (1991)). Yet, reading the article can also be embedded in a larger set of goals (e.g., passing an examination), and may even be part of overarching far-reaching goals like becoming famous (called 'fantasms' by Boesch (1991)). These hierarchies are particularly elaborated in the application of action theory to work and sport settings (i.e., in instrumental contexts). But they are also relevant to communicative actions.

The fact that actions are meaningful to an agency implies that it is exactly this meaning which has to be identified empirically. This calls for hermeneutic methods, because actions have to be interpreted. Harré (1977) calls for an ethogenic approach (ethogenics literally mean 'meaning-giving'). This does not just refer to the dichotomy between qualitative and quantitative methods in psychology, but is a basic methodical feature derived from the theoretical model of an action (it should be noted, however, that no science can do without interpretation).

Beyond the structural aspects of actions, the course of actions is particularly relevant in empirical contexts. This is divided into action phases The number and features of these action phases differ, however: while, e.g., Boesch (1991) distinguishes three action phases (beginning phase, its course, and end) others, e.g., Heckhausen (see Gollwitzer 1990), propose four phases (a predecision phase, a preactional phase, the action phase (doing), and a postaction phase). Here, the decision to act plays an important role (Heckhausen uses the metaphor of crossing the Rubicon). In all these phases there is interplay between cognitive, affective, and energetic aspects of action.

Affects determine the 'valence' of a goal (and therefore of the environment in general), but as actions have more than one goal, goals are also 'polyvalent' (Boesch 1991). Additionally, affects also evaluate the course of an action (dealing with barriers, impediments during the action) and its end (was the action successful or not). These impediments basically increase consciousness, and thus, regulatory processes are of special interest in empirical research. They are basically coping processes dealing with occurring affects (external or primary control, actional, or secondary control). From a systematic point of view regarding these regulatory processes as 'secondary actions' is attractive because they are in fact 'action oriented actions' (Eckensberger 1995).

All questions relating to an agency are of particular empirical interest. First, the consciousness of actions is discussed differently. While some authors claim that consciousness is a necessary aspect of an action (which also implies the methodical possibility of asking actors about their actions), others claim that only the potential self-reflectivity of an agency (and a specific action) is crucial (Eckensberger 1979). This not only implies that a self-reflective action may be a rare event (during a day) but also that actions can turn into automatisms, etc. yet still remain actions. This calls for the analysis of the development of actions. Development therefore is a genuine and crucial dimension in many action theories (as microprocess or actual genesis, as ontogenesis, and as social/cultural change). Second, the development of the agency is a focus of research. Here, studies on self-development become relevant. Of particular interest in this context is the agency's perception of being able to act (called action potential or communicative competency) as a triggering or incitement condition for agency development. Third, the development of agency can itself be considered an action, as a project of identity development which has a goal and which may fail (Brandtstädter 1998). Eckensberger (1995), therefore, proposed calling these identity projects, which have agency related action structures, 'tertiary actions.'

The structural components distinguished above (intentions, finality, causality, etc.) have also become rather central empirical research topic. In fact, the expanding research on '*theory of mind*' (cf. Bartsch and Wellman 1995), and scripts (Nelson 1981) can be interpreted systematically as a program aiming at the questions of whether or not, and at what age, children can think in terms of action structures (distinguish between causal and intentional states, etc.). This strategy has also been applied to the development of moral judgments by Eckensberger and Reinshagen (1980), when analyzing arguments used in moral dilemmas in terms of action structures. Thus, most research programs on social cognition can be (re)interpreted in terms of action theory. Since the action links an agency with the (social and nonsocial) environment (see above), the action is the overlap between the internal and external action field. The internal action field is formed during ontogenetic experiences in the sense that actions are internalized as operations (in Piagetian sense) and normative rules (Turiel 1998) or categories, which for instance develop from (action bound) taskonomies to (generalized) taxonomies. These developments as well as control theories, individual rule systems (logic, understanding of morality, law, conventions) and ideas of the self as agency constitute the internal action field. The external action field, which is understood as culture, provides opportunities and constraints for actions, but it also attributes value to actions. Rituals as a cultural proffer of organized action clusters and myths (as complements of fantasms on the cultural level) are just as important as personal processes of construction (active production of order in the Piagetian sense). Like actions (the action field) can also have different levels of comprehensiveness and be organized hierarchically. According to Boesch (1991), for instance, the external action field of culture can be subdivided

into action spheres (like occupation or family) and action domains (like the office or kitchen). Both, the internal and external action field acquire their affective meaning (valence) via actions.

2. Action Theory as Opportunity for Developing an Integrated Psychological Theory

The uniqueness of the action theory approach to humans not only poses problems for the definition of psychology as a natural science, but also entails the possibility of developing an integrated theory, which not only interrelates different developmental dimensions (actual genesis, ontogenesis, and cultural change, see above) but also resolves most of the 'classical splits' in psychology (Overton 1998), like body/mind, nature/culture, cognition/affects.

The physiological bases for actions as well as the phylogenetic emergence of 'self-reflectivity' in nature can both be understood as 'enabling conditions' for human actions (Harré 1977). Cognitions and affects are integral parts of human actions and their development.

See also: Action Planning, Psychology of; Activity Theory: Psychological; Motivation and Actions, Psychology of; Personality and Conceptions of the Self; Self-concepts: Educational Aspects; Self-conscious Emotions, Psychology of; Self-development in Childhood; Self-esteem in Adulthood; Self: History of the Concept

Bibliography

Barbalet J M 1997 The Jamesian theory of action. *The Sociological Review* **45**(1): 102–21

Bartsch K, Wellman H M 1995 *Children Talk About the Mind*. Oxford University Press, New York, Oxford

Boesch E E 1991 *Symbolic Action Theory and Cultural Psychology*. Springer, Berlin

Brandtstädter J 1998 Action perspectives on human development. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology, Vol. 1: Theoretical Models of Human Development*. Wiley, New York, pp. 807–63

Bruner J 1996 *The Culture of Education*. Harvard University Press, Cambridge, MA

Cole M 1990 Culture psychology: A once and future discipline? In: Berman J J (ed.) *Nebraska Symposium on Motivation 1989 'Cross-cultural Perspectives.'* University of Nebraska Press, Lincoln, NE, pp. 273–335

von Cranach M 1991 Handlungsfreiheit und Determination als Prozeß und Erlebnis. [Action freedom and determination as process and experience.] *Zeitschrift für Sozialpsychologie* **22**: 4–21

Dörner D, Kaminski G 1988 Handeln—Problemlösen—Entscheiden. In: Immelmann K, Scherer K R, Vogel C, Schmoock P (eds.) *Psychobiologie. Grundlagen des Verhaltens*. Gustav Fischer Verlag, Stuttgart, New York, pp. 375–414

Eckensberger L H 1979 A metamethodological evaluation of psychological theories from a cross-cultural perspective. In: Eckensberger L H, Lonner W, Poortinga Y H (eds.) *Cross-cultural Contributions to Psychology*. Swets & Zeitlinger, Lisse, pp. 225–75

Eckensberger L H 1995 Activity or Action: Two different roads towards an integration of culture into psychology? *Culture & Psychology* **1**: 67–80

Eckensberger L H, Reinshagen H 1980 Kohlbergs stufentheorie der entwicklung des moralischen Urteils: Ein Versuch ihrer Reinterpretation in Bemgsrahmen handlungstheoretischer konzepte. In: Eckensberger L H, Silbereisen R K (eds.) *Entwicklung Sozialer Kognitionen*. Klett-Cotta, Stuttgart, Germany, pp. 65–131

Gollwitzer P M 1990 Action phases and mind-sets. In: Higgins E T, Sorrentino R M (eds.) *Handbook of Motivation and Cognition: Foundations of Social Behavior*. Guilford Press, New York, vol. 2, pp. 53–92

Groeben N 1986 *Handeln, Tun, Verhalten als Einheiten Einer verstehend-erklärenden Psychologie*. Francke, Tübingen, Germany

Habermas J 1984 Erläuterungen zum Begriff des kommunikativen Handelns. In: Habermas J (ed.) *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns* Suhrkamp Verlag, Frankfurt am Main, Germany, pp. 571–606

Hacker W 1986 *Arbeitspsychologie*. Deutscher Verlag der Wissenschaft, Berlin

Harré R 1977 The ethogenic approach: Theory and practice. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, p. 10

James W 1897 [1956] *The Will to Believe and Other Essays in Popular Philosophy*. Dover Publications, New York

Kaminski G 1982 What beginner skiers can teach us about actions. In: von Cranach M, Harré R (eds.) *The Analysis of Action*. Cambridge University Press, Cambridge, UK, pp. 99–114

Lenk H 1984 *Handlungstheorien interdisziplinär*. Fink, Munich, Germany

Nelson K 1981 Social cognition in a script framework. In: Flavell J H, Ross L (eds.) *Social Cognitive Development*. Cambridge University Press, Cambridge, UK

Oppenheimer L, Valsiner J (eds.) 1991 *The Origins of Action. Interdisciplinary and International Perspectives*. Springer, New York

Overton W F 1998 Developmental psychology: Philosophy, concepts, and methodology. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology, Vol. 1: Theoretical Models of Human Development*, 5th edn. Wiley, New York, pp. 107–88

Parsons T 1937 [1968] *The Structure of Social Action*. McGraw-Hill Book Co., New York

Schwartz L 1951 *Die Neurosen und die dynamische Psychologie von Pierre Janet*. Benno Schwabe & Co., Basle, Switzerland

Searle J R 1980 The intentionality of intention and action. *Cognitive Sciences* **4**: 47–70

Shweder R A 1990 Cultural psychology: What is it? In: Stigler J W, Shweder R A, Herdt G (eds.) *Cultural Psychology: Essays on Comparative Human Development*. Cambridge University Press, Cambridge, MA, pp. 1–43

Turiel E 1998 The development of morality. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology, Vol. 3: Social, Emotional, and Personality Development* 5th edn. Wiley, New York, pp. 863–932

Valsiner J 1997 The legacy of Ernst E. Boesh in cultural psychology. *Culture and Psychology* **3**: 243–51

Weber M 1904 Die 'Objektivität' sozialwissenschaftlicher und sozialpolitischer Erkenntis. In: Winkelmann M (ed.) *Gesammelte Aufsätze zur Wissenschaftslehre von Max Weber, 1951.* Mohr, Tübingen, Germany
von Wright G H 1971 *Explanation and Understanding*. Cornell University Press, Ithica, NY

L. H. Eckensberger

Action, Collective

Collective action is the means individuals use to pursue and achieve their values when individual action is not possible or likely to fail. Collective action theory is studied in all the social sciences: In economics, it is the theory of public goods and of collective choice (Stevens 1993); in political science, it is called 'public choice' (Mueller 1989); in sociology, it is linked to rational choice, collective behavior, and social movement theory. When markets fail because of imperfect competition, externalities, transaction costs, collective goods provision, and some other reasons, institutions and organizations—governments, political parties, corporations, universities, churches, kinship, social movements, etc.—structure collective action and allocate resources through nonmarket methods. Among these institutions have been conventions, ethical codes, morality, and norms which contribute to efficiency and welfare in social transactions (Arrow 1974). In the broadest sense, collective action theory seeks to explain the origins, evolution, and varieties of nonmarket institutions.

Most collective action is undertaken by organizations that initiate, coordinate, control, and reward individuals' participation in a joint enterprise. In a narrow sense, the theory of collective action deals with the noncoerced, voluntary provision of collective goods, the groups and organizations that provide them, participation and contribution in their pursuit, and contentious actions against targets that resist collective goods attainment. The groups and organizations are interest groups, civic associations, advocacy groups, dissidents, social movements, insurgents, and more transitory social formations such as crowds.

Collective and mass phenomena which result from many individuals pursuing personal goals in spatial and temporal proximity, as in a migration, a baby boom, or the fluctuations of public opinion, have been viewed as aggregations of individual choices and beliefs. Nevertheless, when there are strong externalities and when individuals choose strategically, collective action theory provides powerful insights about aggregation dynamics. Schelling (1978) has shown that housing choices in a mixed-race residential neighborhood can lead to more extreme patterns of racial segregation than the racial preferences of the majority of people in both groups. Similarly Boudon (1982) showed how French higher education reforms meant to increase the opportunities for working-class youth led to the perverse effect of increasing it for affluent youth. Unanticipated consequences, positive and negative bandwagons, unstable equilibria, critical mass, and threshold effects are common consequences of collective actions and central to the theory (Marwell and Oliver 1993).

1. *Collective Behavior*

Collective behavior refers to fads, panics, crazes, hostile crowds, riots, cults, moral panics, cargo cults, witch-hunts, Ghost Dance, and the like. The conventional explanation assumed a variety of social psychological and psychodynamic processes such as consciousness of kind, herd instinct, imitation, contagion, and regression. Observers were struck by the spontaneity and volatility, the emotional-expressive and transitory character of such behaviors in contrast to normatively structured everyday routines. Collective behavior was thought to result from extreme deprivation and threat perception in extraordinary situations when norms and expectations fail to guide action. The best known theorist in this tradition was LeBon ([1895] 1960), who postulated three 'laws' of crowd behavior: mental unity, loss of rational and moral faculties, and hero worship.

The problems with LeBon's and kindred theories of collective behavior is their highly selective character and disregard for alternative explanations. For the same episodes of crowd behavior in the French Revolution that LeBon described, Rude (1959) showed that they were atypical of crowds and many could be explained as purposive action without assuming unproven social psychological processes. Later theorists showed that uniform behavior and mental unity are due to selective convergence of predisposed participants, and that much variance of behavior occurs, ranging from engagement by hardcore activists to standing around by curious bystanders. Rather than amorality, emergent norms structure crowd behavior. Irrational crowd behavior results from the n-person, single game, prisoner's Dilemma aspect of some collective behavior, as in panics of escape (Turner and Killian 1987, Brown 1965).

Because of these shortcomings in the conventional view, collective behavior has been explained with collective action theory, even violent, destructive and bizarre collective behavior, such as lynch mobs, riots, and the witch-hunts of early modern Europe. Southern US lynch mobs in 1880–1920 were structured, ritualized, and predictable (Tolnay and Beck 1995). To be sure, some collective behavior manifests a lot of emotion, we-feeling, hate, fears, violence, and unusual beliefs, yet participants do respond to the benefits and

costs of actions, as they do in other situations. Charismatic leaders are not needed to organize collective behavior. Schelling (1978) has demonstrated how convergence and coordinated behavior by strangers comes about without prior leadership, organization and communication. Tilly (1978) has studied culturally learned and maintained repertoires of collective action of ordinary people against elites and the state. Because of empirical findings and explanations based on rational choice, collective behavior has been gradually integrated with collective actions theory.

Other theories are also being pursued. For Melucci (1996), emotions and identity seeking in small groups eclipses rationality and strategic interaction. Smelser (1963) defines collective behavior as mobilization on the basis of generalized beliefs which redefine social action. Unlike ordinary beliefs, generalized beliefs are about the existence of extraordinary forces, threats, conspiracies, wish fulfillment, utopian expectations, and their extraordinary consequences for adherents. Yet much institutionalized behavior, e.g., miracles in organized religion, the born again Christian conversion, have similar attributes, and the beliefs that inspire political crowds and movements tend to express conventional principles of popular justice, authority, solidarity, and equity (Rude 1959, Tilly 1978). Some political ideologies like xenophobic nationalism, the racist ideology of Nazis against Jews, and religious beliefs like the heresy of witchcraft in early modern Europe, as well as lesser moral panics, are full of threats and conspiracies, and instigate violent, cruel, and fatal actions against thousands of innocent people. Such belief systems and resulting collective actions have been explained with reference to elite manipulation and framing of mass opinion through the communications media, social control of citizens and repression of regime opponents, and conformity to the majority and one's peers (Oberschall 1993, Jenkins 1998). The most notorious genocides including the Holocaust were well planned and thoroughly organized by regime elites fully in command of the state apparatus, the military, the agents of social control, and the citizenry (Fein 1993).

2. *Interest Groups*

Interest groups, labor unions, professional and voluntary associations were assumed to form automatically from the common interest shared among a category of persons in reaction to deprivation and blocked goal attainment (Truman 1958). In the pathbreaking *Logic of Collective Action* where he applied the theory of public goods to public affairs, Olson (1965) argued that although individuals have a common interest in a collective good, each has a separate, individual interest of contributing as little as possible and often nothing to group formation and collective good attainment, i.e., enjoy the benefit and let others pay the cost. Because collective goods possess jointness of supply—if they are provided at all, they must be provided to contributors and non-contributors alike—collective action is subject to free-rider tendencies. Because many public, cultural, and social issues center on collective goods—pollution free environment, social and political reforms such as non-discrimination in employment and the right to abortion, humanitarian causes, charities, listener supported music stations—Olson's deduction that many collective goods will not be provided voluntarily and some only in suboptimal amount has far ranging implications.

Olson's conclusions applied strictly only to large groups of potential beneficiaries. In small groups, especially when one member has a large interest and obtains a large share, the collective good will be provided by such an interested person, though in suboptimal amount, while the others free-ride. This put the accent on patrons and sponsors in collective good provision, and the exploitation of the great by the freeriding small in alliances and coalitions. In an intermediate size population composed of a federation of small groups, a very important category in real world situations and applications, there is some likelihood of contributions to collective good attainment. In large populations, free-rider tendencies dominate. To overcome them, and because voluntary associations do not have the means to coerce contributions as the state can coerce its citizens to pay taxes, groups and leaders induce participation and contribution by offering selective incentives: these are individual benefits that non-contributors do not get, from leadership in the group to travel and insurance discounts, many other material and financial benefits, and some non-material solidarity incentives and social standing.

Free-rider tendencies are especially strong when the collective good is fixed in supply and diminishes as the size of the beneficiaries grows, i.e., freeriders actually diminish the amount of collective good available to contributors, as is the case with tax cheaters in a locality. For many non-market issues (humanitarian, social reform, environmental), the collective good is not subject to crowding, e.g., lower air pollution benefits all regardless of their numbers. Olson showed how the properties of the collective good, and not the psychological disposition and attitudes of individuals, explain differences in relationship between beneficiaries and contributors, recruitment and exclusion strategies of the group, and applied these insights to labor unions and labor laws.

Olson's theory was designed for economic interest groups though much of his model is broadly applicable. Because the accent is on obstacles to collective good attainment and on freeriding, one expects fewer voluntary associations than are actually observed in

democratic countries with freedom of association (Hirschman 1982). His conclusions were based on several assumptions: individuals don't think strategically and don't influence one another's decisions to contribute; only primary beneficiaries have an interest in contributing, and not humanitarians, ideologues, and identity seekers; the beneficiary population is composed of isolated individuals lacking a group and network structure; the production function of resources to collective good is linear and without thresholds, and the good itself continuous and divisible, not lumpy and indivisible; selective incentives are material, and not moral, ideological, and social; potential contributors and beneficiaries differ only in their interest in the collective good—their utility function—and not on many other variables that make for variation in the cost of contribution and the expected benefit.

Collective action theorists have modified and relaxed these assumptions to suit particular topics and applications in political science, sociology, and other disciplines. In particular the theory was restated as an iterated, open ended, N-person prisoner's Dilemma (PD) and integrated with the vast theoretical and experimental PD literature (Hardin 1982, Lichbach 1995). A major empirical test confirmed some parts of Olson's theory (Walker 1983). Walker studied over 500 voluntary associations concerned with national public policy. Citizen groups promoting social reform, environmental and ideological causes were frequently sponsored and maintained by patrons and sponsors, most often foundations, corporate philanthropy, and government agencies in search of a citizen constituency. Many had few material selective incentives to offer members, unlike professional and occupational interest groups, and made instead moral appeals to a conscience constituency who were not primary beneficiaries of the collective good. New communications technologies—direct mail, WATTS long distance telephone lines (and most recently email and the Internet), coupled with postal rates and tax laws favorable to nonprofits, have greatly reduced the costs of organizing and of communication with members and the public (Zald and McCarthy 1987).

A further advance has come from Ostrom (1990, 1998) and her associates' integration of collective action theory with empirical findings from case studies all over the world of member-managed common-pool resources (CPRs). CPRs such as fisheries, irrigation systems and shared groundwater basins, common pastures and forests, share a 'tragedy of the commons' PD, yet can under some circumstances be exploited and managed by beneficiaries in limited fashion without degrading or exhausting the CPR. Thus there is an alternative between privatization of the CPR and surrendering control to the state. Humans create and learn rules and norms, and adapt them to solving their problems. They learn to trust each other and abide by reciprocity norms when their actions are monitored and when compliance influences their reputations. Thus a PD is transformed into an assurance game. Trust, reciprocity, reputation, positive feedback in face to face groups, and a long time horizon make for contingent cooperation that overcomes social dilemmas (PDs) and free-riding. Much research is advancing collective action theory along these lines.

3. *Social Movements*

Social movements consist of groups, organizations, and people who advocate and promote a cause or issue and associated collective goods. They confront opponents, who are frequently governments and privileged groups. They use a mixture of institutional and unconventional means of confrontation, at times even coercive and violent means. In recent decades a large variety of movements have been studied: nationalist, ethnic, separatist, anticolonial, peace, democracy, human rights, environmental, ethnic minority, civil rights, women's rights and feminist, animal rights, labor, peasant, student, for and against abortion, temperance, antismoking, religious revival, religious fundamentalist, and so on. Many social, political, and cultural changes have resulted in part from social movements, even when they have failed in the short term. Olson's work, a reinterpretation of the Nazi movement based on new historical scholarship, and the social turmoil and explosion of popular participation in protests in the 1960s stimulated a reconceptualization of social movement theory.

Olson's assumptions were modified to suit the social contexts of collective action while adhering to the core of Olson's thinking. Sociologists embedded participation in social networks and groups. Moral, social, and ideological incentives were added to the material selective incentives, which put the accent on nonbeneficiary contributors and a conscience constituency. Seeking and fulfilling an identity through protest was added. Between spontaneous, unstructured crowds and social movement organization loosely structured collective action was discovered (a variety of the federated group). Participation became variable with a division of labor between activists, part-timers, supportive conscience constituency, and a sympathetic by-stander public. Issues were socially constructed through framing; varieties of contentious actions by challengers against opponents were studied, some of which were learned and culturally transmitted in protest repertoires. The trigger mechanism provided by small groups of activists and dissidents for large-scale collective action diffusion was discovered. Production functions for collective goods became nonlinear as well as linear, depending on the character of the collective good and the tactic of confrontation. Just as important, there was an outpouring of empirical research based on observation, social surveys of participants, case studies and comparative studies

from history, the testimony of movement leaders and rank-and-file participants, news coverage, systematic content analysis of news media, video documentaries, and much else, from a growing number of countries and especially the USA and Western Europe, with many researchers and writers using the same terminology, concepts, viewpoints, hypotheses, and methods (Diani and Eyerman 1992). Although some maintained that the Europeans adhered to 'new social movement theory' (Dalton and Kuechler 1990) and the US sociologists to 'resource mobilization' (Zald and McCarthy 1987), the difference was a matter of emphasis and only slight theoretical import. These and other labels obscure the fundamental unity and coherence of collective action theory.

Social movement theory operates at two levels simultaneously, the micro and the macro, both of which have a static and a dynamic dimension (Obserschall 1993). At the macro level, there are four conditions for initiation and continuance of social movements: (a) discontent, when institutionalized relief fails; (b) beliefs and values that filter, frame, and transform discontent into grievances calling for action (Snow et al. 1986); (c) capacity to act collectively, or mobilization (Gamson 1975, Tilly 1978); (d) opportunity for successful challenge, or political opportunity. Some analysts have emphasized one or another of these dimensions, as Gurr (1970) did with relative deprivation, a grievance variable; Zald and McCarthy (1987) did with social movement organizations, a mobilization variable; and McAdam (1982) and Tarrow (1993) did with political process. In actual studies, these and other theorists address all four dimensions. Many case studies and comparative studies of social movements can be accomodated within this four-dimensional approach, though causal theories of discontent, grievance, ideology, mobilizing capacity, and political opportunity are still incomplete.

At the micro level, the decision to participate in collective action is based on the value of the collective good to beneficiary multiplied by the probability of attainment, a subjective estimate. This term and selective incentives constitute the expected benefit. On the cost side there are opportunity costs and costs of participation, including expected costs of arrest, injury, blacklisting, and the like. Participation is chosen when net benefit is positive (Klandermans 1997, Oberschall 1993, Opp 1989). Because probability of attainment and some costs are a function of the expected number of participants, there is strong feedback among individual decisions. Dramatic shifts in net benefit can occur in a short time, which precipitates cascades of joining, or negative bandwagons. Empirical research, and in particular survey research based on participants' responses compared to nonparticipants, confirms the micro model for a variety of social movements and several countries (Klandermans 1997).

The dynamic theory of challenger–target confrontations, though richly dealt within case studies, lacks a developed theory. McAdam (1983) has demonstrated innovation, learning, and adaptation by both challengers and targets in confrontation sequences. Oberschall (1993) has shown that social control in contentious confrontations gives rise to new issues and grievances, e.g., police brutality, that mobilize new participants, and that news media coverage of protests can make for rapid protest diffusion by signaling focal points and issues for convergence and providing vicarious expectations of participant numbers. Confrontation dynamics show promise for analysis with game theory and simulation (Marwell and Oliver 1993) but much remains to be done.

4. *Norms and Institutions*

Together with the new institutional economics (North 1990) transaction cost theory (Williamson 1975) cooperation theory (Axelrod 1984), and public choice, rational choice/rational actor theory in sociology seeks to explain norms, institutions, group formation, social organization, and other products of collective action, from elementary principles. The most ambitious effort to date is Coleman (1990). The elementary units of analysis are actors, resources, interests, and control. From these both systems of exchange and authority relations and structures are constructed, based on the right to control resources and other actors' actions, e.g., a norm to which the actors conform. The demand for norms arises when actors create externalities for one another, yet a market in rights of control cannot be easily established. The realization of norms occurs when social relationships among actors enable them to share the benefits and costs of sanctioning. Among system properties discussed by Coleman are agency, social capital, and trust. Slowly evolved and inherited institutions such as kinship and the family (Ben-Porath 1980) and village communities (Popkin 1979) can be understood with these theories. Ascriptive relationships, as in kinship, are instances of specialization by identity, when individuals transact only with the same person or small groups in bilateral monopoly. This mode of transacting enables huge asymmetric investments in other human beings, as with raising children by parents, that are not expected to be paid back for a long time. To break out of kin encapsulation with limited opportunities for specialization, kin groups forge horizontal alliances through marriage and build political leadership through these networks. A perverse aggregate consequence, under conditions of extreme resource scarcity and competition, discovered by Banfield (1958) in a South Italian district, is 'amoral familism'—an instance of PD—which impedes community wide civic organization and leadership, and hinders social change. There is no civic culture. Among

other topics studied are sharing groups (Lindenberg 1982), state formation (Levi 1988) and varieties of religious organization and behavior (Iannaccone 1988).

The major social inventions of modernity for Coleman are roles, offices, and corporate actors—from the colonial trade joint stock companies and chartered towns to the modern corporation, labor unions, and professional associations—which are resistant to people mortality and turnover and allow investment in and transacting with a corporate venture and between corporate actors, not just specific persons. Corporate actors create entirely new opportunities for social change as well as problems of governance, agency, and asymmetries of power. They generate hitherto unknown modes of impersonal trust based on certified skills and professional ethics codes. These and other related topics on the evolution of institutions, based on rational choice and collective action theory, are published in *Rationality and Society* and similar journals. Together with advances in the other social sciences, in evolutionary biology, and in cognitive and evolutionary psychology, collective action theory is becoming a part of an integrated and comprehensive social science with a wide reach.

See also: Action, Theories of Social; Coleman, James Samuel (1926–95); Collective Behavior, Sociology of; Disasters, Sociology of; Fashion, Sociology of; Interest Groups, History of; Olson, Mancur (1932–98); Panic, Sociology of; Rational Choice Theory in Sociology; Social Movements, Sociology of; Violence, History of; Violence: Public

Bibliography

Arrow K J 1974 *The Limits of Organization*. 1st edn. Norton, New York

Axelrod R 1984 *The Evolution of Cooperation*. Basic Books, New York

Banfield E C 1958 *The Moral Basis of a Backward Society*. Free Press, Glencoe, IL

Ben-Porath Y 1980 The F-connection: Families, friends and firms and the organization of exchange. *Population and Development Review* **6**(1): 1–30

Boudon R 1982 *The Unintended Consequences of Social Action*. Macmillan, London

Brown R 1965 *Social Psychology*. Free Press, New York

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Dalton R J, Kuechler M 1990 *Challenging the Political Order: New Social and Political Movements in Western Democracies*. Oxford University Press, New York

Diani M, Eyerman R (eds.) 1992 *Studying Collective Action*. Sage, London

Fein H 1993 *Genocide, a Sociological Perspective*. Sage, London

Gamson W A 1975 *The Strategy of Social Protest*. Dorsey Press, Homewood, IL

Gurr T R 1970 *Why Men Rebel*. Princeton University Press, Princeton, NJ

Hardin R 1982 *Collective Action*. Johns Hopkins University Press, Baltimore, MD

Hirschman A O 1982 *Shifting Involvments*. Princeton University Press, Princeton, NJ

Iannaccone L 1988 A formal model of church and sect. *American Journal of Sociology* **94**(Supplement): S241–268

Jenkins P 1998 *Moral Panic*. Yale University Press, New Haven, CT

Klandermans B 1997 *The Social Psychology of Protest Action*. Blackwell, Oxford, UK

Le Bon G 1960 *The Crowd*. Viking, New York

Levi M 1988 *Of Rule and Revenue*. University of California Press, San Francisco

Lichbach M I 1995 *The Rebel's Dilemma*. University of Michigan Press, Ann Arbor, MI

Lindenberg S 1982 Sharing groups: Theory and suggested applications. *Journal of Mathematical Sociology* **9**: 33–62

Marwell G, Oliver P 1993 *The Critical Mass in Collective Action*. Cambridge University Press, Cambridge, UK

McAdam D 1982 *Political Process and the Development of Black Insurgency, 1930–1970*. University of Chicago Press, Chicago

McAdam D 1983 Tactical innovation and the pace of Insurgency. *American Sociological Review* **48**(6): 735–54

Melucci A 1996 *Challenging Codes*. Cambridge University Press, Cambridge, UK

Mueller D C 1989 *Public Choice II*. Cambridge University Press, Cambridge, UK

North D C 1990 *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, UK

Oberschall A 1993 *Social Movements, Ideologies, Interests and Identities*. Transaction, New Brunswick, NJ

Olson Jr M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA

Ostrom E 1990 *Governing the Commons. The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK

Ostrom E 1998 A behavioral approach to the rational choice theory of collective action. *American Political Science Review* **92**(1): 1–22

Opp K -D 1989 *The Rationality of Political Protest*. Westview Press, Boulder, CO

Popkin B 1979 *The Rational Peasant*. University of California Press, Barkeley

Rude G 1959 *The Crowd in the French Revolution*. Clarendon Press, Oxford, UK

Schelling T C 1978 *Micromotives and Macrobehavior* 1st edn. Norton, New York

Smelser N 1963 *The Theory of Collective Behavior*. Free Press of Glencoe, New York

Snow D, Rochford E, Worden S, Benford R 1986 Frame alignment processes, micromobilization, and movement participation. *American Sociological Review* **51**(August): 464–481

Stevens J B 1993 *The Economics of Collective Choice*. Westview Press, Boulder, CO

Tarrow S 1993 *Power in Movements*. Cambridge University Press, Cambridge, UK

Tilly C 1978 *From Mobilization to Revolution*. Addison Wesley, Reading, MA

Tolnay S E, Beck E M 1995 *A Festival of Violence*. University of Illinois Press, Urban IL

Turner R H, Killian L M 1987 *Collective Behavior* 3rd edn. Prentice Hall, Englewood Cliffs, NJ

Truman D 1958 *The Governmental Process*. Knopf, New York
Walker J L 1983 The origins and maintenance of interest groups in America. *American Political Science Review* **77**(2): 390–406
Williamson O E 1975 *Markets and Hierarchies*. Free Press, New York
Zald M N, McCarthy J D 1987 *Social Movements in an Organizational Society*. Transaction Books, New Brunswick, NJ

A. R. Oberschall

Action, Theories of Social

1. The Interpretive Theory of Social Action

Social action has become an important topic in sociological theory under the influence of the great German sociologist Max Weber. To him, 'social action, which includes both failure to act and passive acquiescence, may be oriented to the past, present, or expected future behavior of others' (Weber 1922, p. I). To Weber, explaining a social phenomenon means analyzing it as the effect of individual actions. He says explicitly in a letter addressed the year of his death to a friend, the marginalist economist Rolf Liefmann: 'sociology too must be strictly individualistic as far as its methodology is concerned' (Mommsen 1965, p. 44). The 'too' means that sociology should, according to Weber, follow the same principle as economics, a principle later christened 'methodological individualism' by Joseph Schumpeter, and later popularized by Friedrich Hayek and Karl Popper. This principle states simply that any *collective* phenomenon is the outcome of *individual* actions, attitudes, beliefs, etc. To methodological individualists, as Weber, a crucial step in any sociological analysis is to determine the causes of individual actions. Weber introduces then a crucial second postulate: that the causes of any action lie in the meaning to the actor of his action. Thus, the cause responsible for the fact that I look on my right and my left before crossing a street is that I want to avoid the risk of being hit by a car. To the operation aiming at retrieving the meaning to the actor of his action, Weber gives a name: *Verstehen*, to understand. Given the importance of the *Verstehen* postulate, Weber calls the style of sociology resting upon these two postulates 'comprehensive' sociology. To Weber, by contrast with notably Dilthey, the notion of 'comprehension' characterizes exclusively individual actions, attitudes or beliefs.

Weber (1922) has proposed in famous pages of his posthumous work *Economy and Society* a distinction between four main types of actions. Actions can be inspired by instrumental rationality (*Zweckrationalität*): when an actor does X because he perceives X as an adequate way of reaching a goal G. They result from axiological rationality (*Wertrationalität*) when an actor does X because X is congruent with some value he endorses. Actions are 'traditional' (*traditionell*) when they are oriented to the fact that such actions have been regularly performed in the past, and are perceived as recommended by virtue of that fact. Finally, an action is 'affective' (*affektuell*) when it is inspired by some feeling or generally emotional state of the subject.

2. The Functional Theory of Social Action

An important contribution to the theory of social action is Parsons' *The Structure of Social Action* (1937), a work where the American sociologist attempts to combine some seminal ideas on social action developed by Weber, Durkheim, Pareto, and Alfred Marshall. Parsons devotes much attention to the point that, to Weber, action is defined as oriented to the behavior 'of others.' He is notably concerned by the idea that social actors are embedded in systems of social roles. To him, roles rather than individuals should be considered as the atoms of sociological analysis. This shift from individuals to roles was inspired to Parsons by his wish of combining the Weberian with the Durkheimian tradition, individual actions with social structures. The most popular aspect of Parsons' theory is his typology of the 'pattern variables.' These 'pattern variables' are a set of four binary attributes by which all roles can in principle be characterized. Thus, the role of a bank clerk is 'specific' in the sense where his relation to his customers is limited to well-defined goals, by contrast with the role of, say, 'mother,' that is 'diffuse.' The role of mother is 'ascribed,' while the role of clerk is 'achieved.' The former is 'particularistic' in the sense where it deals with specific individuals; the latter is 'universalistic': the clerk is supposed to apply the same rules indistinctly to all customers. Ralf Dahrendorf (1968) saw in the Parsonian theory a definition of the *homo sociologicus* and a proper basis for making sociology a well-defined discipline, resting on a well-defined set of postulates. While economics sees the *homo oeconomicus* as moved by his interests and as able of determining rationally the best ways of satisfying them, the Parsonian *homo sociologicus*was described as moved, not only by interests, but by the norms and values attached to his various roles. Merton (1949) developed ideas close to Parsons's, insisting on the norms and values attached to roles but also on the ambiguities and incompatibilities generated by the various roles an individual is embedded in. It must be recognized, though, that the idea according to which the parsonian *homo sociologicus* would guarantee to sociology foundations as solid as the *homo oeconomicus*to economics has never gained recognition. More precisely, while most sociologists accept the idea

that norms, beside interests should be taken into account in the explanation of action, they doubt that the parsonian *homo sociologicus* can be expressed in a form able to generate deductive theories as precise and powerful as the *homo oeconomicus*.

The skepticism toward the Parsonian theory of action that appeared in the 1960s results not only from this theoretical consideration but also from conjunctural circumstances. In the 1960s, the so-called 'functional' theory, a general and vague label that covered then the sociology with a Parsonian inspiration, became strongly attacked. Critical sociologists objected to 'functionalism' in that it would contribute to legitimate the existing social institutions, while the main objective of sociology should be to criticize them. To this unfair objection another, equally unfair, was added: that functionalism would not be scientifically fruitful. Functionalism provides a useful theoretical framework to develop a sociological theory of stratification, of the legitimacy of institutions, and of other social phenomena. But it is true that it did not succeed in providing a theoretical basis from which sociological research could develop cumulatively. By contrast with the *homo oeconomicus*, the *homo sociologicus* of the functionalist tradition failed to generate a well-identified research tradition.

3. *The Utilitarian Theory of Social Action*

Neither critical theory nor other more recent sociological movements, as ethnomethodology or phenomenology, succeeded in providing a solid basis for a theoretical consensus among sociologists. The 'balkanized' character of sociological theory incited some sociologists to propose to identify the *homo sociologicus* with the *homo oeconomicus*. This proposal was motivated by the fact that the model of the *homo oeconomicus* had actually been applied successfully to several kinds of problems belonging traditionally to the jurisdiction of sociology. Thus, the so-called 'theory of opportunities' rests upon the postulate that criminal behavior can be analyzed as a maximizing behavior. The economist G. Tullock (1974) had shown that differential data about crime could notably be accounted for by a theory close to the theory of behavior used by neoclassical economists. G. Becker, another economist, proposed to analyze social discrimination along the same line. In *Accounting for Tastes*, Becker (1996) analyzes addiction as resulting from cost-benefit considerations and claims that the 'rational choice model,' namely the model of man proposed by neoclassical economists, is the only theory able to unify the social sciences. This general idea had been developed by J. Coleman (1990) in his *Foundations of Social Theory*.

The idea of explaining social action by the 'utilitarian' (in Bentham's sense) postulates is not new. Classical sociologists use it occasionally. Thus, in his *The Old Regime and the French Revolution*, Tocqueville ([1856] 1986) explains that the underdevelopment of French agriculture at the end of the eighteenth century, at a time when British agriculture knows a phase of quick modernization, is the effect of landlords' absenteeism. As to the latter, it results from the fact that the French landlords were better off socially when they bought a royal office than when they stayed on their land. The French centralization meant that many royal offices were available and brought prestige, power, and influence to those who filled them. In Britain by contrast, a good way of increasing one's influence was to appear as an innovative gentleman farmer and by so doing getting local and eventually national political responsibilities. So, Tocqueville's landowners make their decisions on the basis of a cost-benefit analysis, along the line of the 'rational choice model.' The social outcome is different in the two contexts because the parameters of the two contexts are different. But Tocqueville uses this model exclusively on subjects where it seems to account for historical facts.

The utilitarian postulates defended by rational choice modelists were not only occasionally used by Tocqueville, they had also been treated as universally valid by some theorists, notably Marx and Nietzsche and their followers. To Marx, and still more to Marxians, individual actions and beliefs should be analyzed as motivated by class interests, even though the final role of his interests can remain unrecognized by the actor himself ('false consciousness'). To Nietzsche, and still more to Nietzscheans, individual actions and beliefs should be analyzed as motivated by their positive psychological consequences on the actor himself. Thus, to Nietzsche, the Christian faith developed originally among the lower classes because of the psychological benefits they could derive from endorsing a faith that promised Paradise to the weak and the poor. In his *Essays in the Sociology of Religion*, Weber (1920–1) is critical toward such theories: 'my psychological or social interests can draw my attention on an idea, a value or a theory; I can have a positive or a negative prejudice toward them. But I will endorse them only if I think they are valid, not only because they serve my interests.' Weber's position has the advantage of making useless the controversial 'false consciousness' theory. As rightly stressed by Nisbet (1966), the ideas of 'false consciousness' in the Marxian sense (the concept itself being due to F. Mehring) and of 'rationalization' in the Freudian sense have become commonplace; they postulate highly conjectural psychological mechanisms, though.

The utilitarian approach proposed by rational choice theorists owes little to this Marxian–Nietzschean tradition. The motivation of 'rational choice theorists' resides rather in the fact that the postulates used by neoclassical economics explain many social phenomena of interest to sociologists. Moreover, they make possible the use of the math-

ematical language in sociological theory building. Above all, they provide final explanations without black boxes.

While the 'rational choice' approach is important and can be effectively used on many subjects, its claim to be the theoretical ground on which sociology could be unified is unjustified. Its limits are more and more clearly recognized by economists. Thus, Bruno Frey (1997) has shown that under some circumstances people are more willing to accept unpalatable but collectively beneficial outcomes than they are to accept outcomes for which they receive compensation.

Generally, a host of social phenomena appear as resistant to any analysis of the 'rational choice' type as the example of the so-called 'voting paradox' suggests. As in a national election a single vote has a practically zero influence on the outcome, why should a 'rational' voter vote?

Ferejohn and Fiorina (1974) have proposed considering the paradox of voting as similar in its structure to Pascal's bet: as the issue of the existence of God is crucial, even if the probability that God exists is supposed close to zero, I have an interest in betting that He exists. Pascal's argument is relevant in the analysis of attitudes toward risk. Thus, it explains why it is not necessary to force people to take an insurance against fire: the cost of the insurance is small and the importance to me of the damages being compensated in the case my house would burn is great, so that I would normally subscribe. That the same argument can be realistically used in the case of voting behavior is more controversial, notably because actual voters often show a very limited interest in the election.

Overbye (1995) has offered an alternative theory: people would vote because nonvoting would be negatively regarded, so that nonvoting would entail a cost. But, rational people should see that any individual vote fails to influence the outcome of an election; why then should they consider nonvoting as bad?

Another theory claims that people vote because they estimate in a biased fashion the probability of their vote being pivotal. The bias must be so powerful, however, that such an assumption appears as *ad hoc*.

Another theory, also resting on the 'rational choice model,' submits that people vote because they like to vote. In that case, the cost of voting being negative, the paradox disappears. Simple as it is, the theory introduces the controversial assumption that voters would be victims of their 'false consciousness,' since they do not see that they just like to vote and believe that they vote for some higher reasons. Moreover, this theory does not explain why the turnout is variable from one election to another.

Actually, no theory using the basic postulates of the 'rational choice model' appears as convincing. The good explanation is that people vote because they believe that democracy is a good regime, that elections are a basic institution of democracy, and that one should vote as long as one has the impression that a policy or a candidate are better than alternative ones. This is an example of what Weber has called 'axiological rationality.'

4. *The Cognitive Theory of Social Action*

The theory of action characteristic of neoclassical economics and used by 'rational choice' theorists was made more flexible by H. Simon. His study of decisions within organizations convinced him that decision-makers take 'satisficing' rather than 'optimizing' decisions: because of the costs of information, stopping a deliberation process as soon as one has discovered a 'satisfying' decision can be more rational than exploring further the field of possible decisions. A chess player could in principle determine the best next move. Actually, this would entail a huge number of computations. So, he will use rather rules of thumb. H. Simon qualified this type of rationality as 'bounded.' His contributions stress the crucial point that social action includes an essential cognitive dimension and invite sociologists to drop the postulate of neoclassical economics (used for instance in game theory) according to which social actors would be fully informed when they take their decisions.

Experimental cognitive psychology has also contributed to the sociological theory of action. It has shown that ordinary knowledge is often 'biased,' as in the case where respondents are confronted with a situation where they have to estimate the probabilities of alternative events. Thus, in an experiment, subjects are invited to guess the outcome of a heads-and-tails game with a biased coin, where heads and tails have a probability of coming out, respectively, of 0.8 and 0.2 and where the subjects are informed about it. The experiment reveals that most people guess 'heads' and 'tails,' respectively, with probabilities of 0.8 and 0.2. Now, by so doing, they are worse off than if they would predict 'heads' all the time, since they would then win on average eight times out of 10, while with their preferred strategy the probability of winning is $(0.8 \times 0.8) + (0.2 \times 0.2) = 0.68$. Rather than talking of 'biases' in such cases, it is perhaps more illuminating to make the assumption that, when people are faced with problem solving situations, they try to build a *theory*, satisfying to their eyes, but depending of course of their cognitive resources. In that case, people use the theory that, since they are asked to predict a sequence of events 'heads' or 'tails,' a good strategy is to use the law governing the actual sequence generated by the experimenter. So, while wrong, their answer may be analyzed as understandable, since inspired by a theory which in other situations would be valid. False scientific theories also generally result from *understandable* systems of reasons. Priestley did not believe in the phlogiston theory because he was affected by some cognitive 'bias,' but because strong

reasons convinced him of the existence of the phlogiston.

Fillieule (1996) has rightly contended that the theory of sociological action should take seriously the meaning of the notion of rationality as defined not only by neoclassical economics but by the philosophy of science as well. In the vocabulary of the philosophy of science, an actor is 'rational' when he endorses a theory because he sees it as grounded on strong reasons. Durkheim (1912) maintains in his *Elementary Forms of Religious Life* that scientific knowledge and ordinary knowledge differ from one another in degree rather than nature. Even religious and magical beliefs as well as the actions generated by these beliefs should be analyzed in the same fashion as scientific beliefs: primitive Australians have strong reasons to believe what they believe. One can call this type of rationality, evoked by Durkheim as well as by philosophers of science, 'cognitive rationality.'

Applications of this notion are easily found. In the early phase of the industrial revolution in Britain, the Luddites destroyed their machines because they thought that machines destroy human work and generate unemployment. Their action was grounded on a belief and the belief on a theory. They endorsed the theory because it is grounded on strong reasons: a machine is effectively designed and built with the purpose of increasing productivity by substituting mechanical for human work. So, other things equal, when a machine is introduced in a factory, it destroys effectively some amount of human work. But other things are not equal: in an economic system as a whole, human work is needed to conceive, build, maintain, and modernize the new machine, so that on the whole the new machine can create more work than it destroys. Whether this is actually the case is an entirely empirical question. But, at a local level, the workers have strong reasons to believe that the introduction of new machines are a threat to employment.

Taking 'cognitive' rationality into account, beside the instrumental type of rationality used in the 'rational choice model,' is essential to a realistic theory of social action. As stressed by Weber as well as Durkheim, beliefs are a normal ingredient of social action. Now, beliefs cannot generally be explained by the 'rational choice model': I generally do not believe that X is true because believing so serves my interests, but because I have strong reasons for so believing. The dominant status in contemporary sociology of the instrumental-utilitarian conception of rationality incorporated in the 'rational choice model' has the effect that the powerful intuition of classical sociologists according to which, first, explaining beliefs should be a main concern in the sociological theory of action and, second, beliefs should be analyzed as endorsed by social actors because they have strong reasons for endorsing them, appears actually as neglected.

Normative and axiological beliefs, beside representational beliefs, are also a crucial ingredient of social action. Weber's distinction between instrumental and axiological rationality introduces the crucial idea that normative beliefs *cannot always* be analyzed as the product of instrumental rationality nor *a fortiori* by the contemporary 'rational choice model,' which considers exclusively instrumental rationality. Boudon (1998) has submitted that a fruitful interpretation of the notion of 'axiological rationality' would be to consider that axiological beliefs are legitimated in the mind of actors because the latter see them as grounded on strong reasons. Axiological rationality would then be a variant, dealing with prescriptive rather than descriptive beliefs, of the 'cognitive' type of rationality.

Axiological rationality is responsible notably for the evaluations people bear on situations they are not involved in. The 'rational choice model' cannot for instance account for the opinions of people on a topic such as death penalty, because most people are obviously not personally concerned with the issue. They have strong convictions on the subject, though. Should we consider these convictions as irrational since the 'rational choice model' is unable to account for them, or decide rather to follow and elaborate on the classical sociological theory of rationality?

See also: Action, Collective; Bounded Rationality; Coleman, James Samuel (1926–95); Critical Theory: Contemporary; Functionalism, History of; Functionalism in Sociology; Interests, Sociological Analysis of; Norms; Parsons, Talcott (1902–79); Rational Choice Theory in Sociology; Sociology: Overview; Utilitarian Social Thought, History of; Utilitarianism: Contemporary Applications; Voting, Sociology of; Weber, Max (1864–1920)

Bibliography

Becker G S 1996 *Accounting for Tastes*. Harvard University Press, Cambridge, MA

Boudon R 1998 Social mechanisms without black boxes. In: Hedström P, Swedberg R (eds.) *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge University Press, New York, pp. 172–203

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Dahrendorf R 1968 Homo Sociologicus. In: Dahrendorf R (ed.) *Essays in the Theory of Society*. Stanford University Press, Stanford, CA, pp. 19–87

Durkheim E 1912 *Les Formes Élémentaires de la Vie Religieuse*. F. Alcan, Paris

Ferejohn J A, Fiorina M P 1974 The paradox of not voting: a decision theoretic analysis. *American Political Review* **68**: 525–36

Fillieule R 1996 Frames, inference, and rationality: some light on the controversies about rationality. *Rationality and Society* **8**: 151–65

Frey B S 1997 *Not Just for the Money, an Economic Theory of Personal Motivation*. Edward Elgar, Cheltenham, UK
Merton R K 1949 *Social Theory and Social Structure: Toward the Codification of Theory and Research*. Free Press, Glencoe, IL
Mommsen W 1965 Max Weber's political sociology and his philosophy of world history. *International Social Science Journal* **17**(1): 23–45
Nisbet R A 1966 *The Sociological Tradition*. Basic Books, New York
Overbye E 1995 Making a case for the rational, self-regarding, 'ethical' voter ... and solving the 'paradox of not voting' in the process. *European Journal of Political Research* **27**: 369–96
Parsons T 1937 *The Structure of Social Action: A Study in Social Theory with Special Reference to a Group of Recent European Writers*, 1st edn. McGraw-Hill, New York
de Tocqueville A [1856] 1986 L'Ancien régime et la révolution. In: Tocqueville A de (ed.) *De la démocratie en Amérique, Souvenirs, l'Ancien Régime et la Révolution*. Laffont, Paris
Tullock G 1974 Does Punishment Deter Crime? *The Public Interest* **36**: 103–11
Weber M 1920–1 *Gesammelte Aufsätze zur Religionssoziologie*. Mohr, Tübingen, Germany, Vol. 1
Weber M 1922 *Wirtschaft und Gesellschaft*, 4th edn. Mohr, Tübingen, Germany, 2. Vols.

R. Boudon

Activity Theory: Psychological

1. Introduction: Activity, Action, Operation

Psychology has traditionally distinguished various classes of behaviors, e.g., reflexes, affective responses, and goal-oriented activities. The special nature of goal-oriented activities is clearest when they are contrasted with the behavior of human beings who are not, or not yet, able to orient their activities to goals, i.e., mentally retarded people or those with major injuries to the frontal lobes of the brain. Those activities that are not organized towards goals are typically characterized as trial and error, impulsively and unreflectively driven, without direction and orientation, and without examination of the consequences of alternatives. Goal-oriented selection of programs, known or yet to be developed, is lacking. The particular steps of activity are not oriented towards goal implementation. They are neither *integrated parts* of linear sequences of steps nor *subordinated parts* of a hierarchical plan. Hence they are all perceived to be of equal importance for goal implementation. Furthermore, there is no anticipating comparison between the given state and a desired goal state. Finally, a prospective evaluation of consequences is lacking (see also *Action Planning, Psychology of*). There are distinctions to be drawn among the concepts of activity, action, and operation.

Activities are motivated and regulated by higher-order goals and are realized through actions that are themselves relatively independent components of each activity.

Actions differ from each other with respect to their specific goals. Actions may themselves be decomposed into their subordinate components, the operations. Operations are described as subordinate because they do not have goals of their own. Operations can be taken to be movement patterns or, in the case of mental activities, elementary cognitive operations.

The concept of a psychology of activities has been, since the mid-twentieth century, central to the tradition especially of Russian and German psychology. There are many points of agreements with, but also important differences between, the orientations of the leading research groups, particularly those of Leontjev (1979)—a student of Vygotski—Rubinstein (1961), and Tomaszewski (1981). The philosophical foundations of Marxism, the psychological findings of Lewin (1926), the psychophysiological results of Bernstein (1967), the neuropsychological results of Luria (1973), and suggestions of cybernetics, particularly of Systems Theory, have all contributed to the development of this concept. The basic idea of this concept is that activity cannot adequately be researched in stimulus–response terms. The elements or 'building blocks' of even primitive and unchallenging real-life activities are not just responses or reactions, but goal-oriented actions (Hacker 1985a, 1985b). Goal orientation, however, does not mean a strictly top-down planned activity (von Cranach et al. 1982). Instead, goal-oriented real-life activity is 'opportunistically' organized, which means that people are trying to accomplish goals by a kind of 'muddling through' with some planned episodes. A modern review from the very special point of view of linking cognition and motivation to behavior was presented by Gollwitzer and Bargh (1996), and a more general review was written by Frese and Sabini (1985).

The concept of goal-oriented activities and actions is a relational one that relates at least five components:

(a) the anticipated and desired result, represented as the goal;

(b) the objects of the activities (e.g., raw materials), which typically have their own laws governing how they can be transformed from a given state into the desired one;

(c) transformations of the physical or social world (e.g., nailing), requiring the expenditure of energy and the use of information (i.e., the *actual change* of the objects without which there would be only an unimplemented intention);

(d) the acting person, with her/his ability to have an impact on, and attitudes toward, the processes; these processes, in turn, act back on the person; and

(e) the means needed for, and the contextual conditions of, the activities.

This relational framework offers a guideline for task and activity analysis, especially the analysis of the

knowledge base which is necessary in order to implement a task.

2. Sequential and Hierarchical Organization of Activity

There are good theoretical reasons to describe the structure of the mental processes and representations regulating activity as being simultaneously sequentially (or cyclically) and hierarchically organized.

The sequential organization of activity control is a cyclical one in-as-much as the steps or stages of activity may be described in terms of control loops. An example of such a description is the Test–Operate–Test–Exit (TOTE) model (Miller et al. 1960). We will come back to these 'stages of control' in a less abstract manner in Sect. 4.

The concept of the hierarchical organization can help to explain the different levels of consciousness of the mental processes and representations that regulate activities. One can distinguish between (a) processes that one is normally not able to process consciously (breathing, e.g., while speaking), (b) processes that one is able to regulate consciously, but is not obliged to have in consciousness (e.g., formulating complete sentences while speaking), and (c) processes that have to be represented in consciousness (e.g., complex inferences).

The hierarchy of these levels means that 'higher' or conscious ('controlled') levels include and determine 'lower' ('automated') ones (Fig. 1).

Since the cyclically organized phases (e.g., in terms of TOTE units) are simultaneously hierarchically nested one in another, action control will be simultaneously organized sequentially and hierarchically.

Following this model of action control, a few characteristics will follow:

(a) superordinate levels of control with higher consciousness and a broader range of impact *include* subordinate ones;

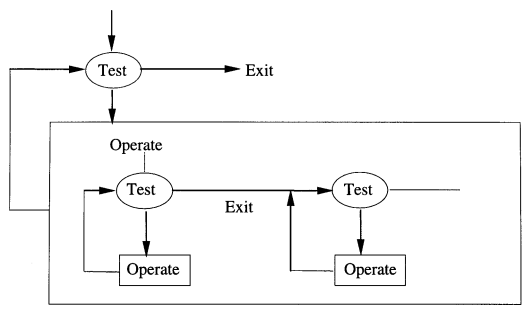

Figure 1
Hierarchic organization of the cyclical control loop (TOTE) units

(b) superordinate levels *determine* subordinate ones;

(c) superordinate levels *delegate* details, thus saving mental capacity;

(d) subordinate levels obtain a *relative autonomy* and the possibility of a bottom-up impact on higher ones (von Cranach et al. 1982).

Following the above hierarchical organization, the regulation of movements or motor operations is a dependent component of the superior goal-oriented actions. There are several consequences of this approach, as follows.

First the *meaning of a task* for the subject/actor will determine the structure of the operations involved, as was shown by neuropsychological case studies (Luria 1973). The outstanding Russian physiologist Bernstein (1967, p. 70) more generally stressed, 'What kind of motor response will be analyzed..., only the meaning of the task and the anticipation... of the result are the invariant parameters which determine a fixed program or a program reorganized during implementation that both step by step will govern the sensory corrections.' The temporal and spatial parameters of the motor response do not offer the inevitable invariant parameters since several variations of a motor response may have the same result.

Following Bernstein's (1967) cited notion along with the meaning of the task, the *anticipated result* of an action is regulating the accomplishment of the action. This anticipation will become the goal in-as-much as it will be combined with the intention to implement the anticipated result. Generalizing this line of thinking on the goal-directed anticipative control of motor operations as components of actions, von Weizsäcker (1947, p. 139) stressed, 'In a motor response the effect will not be determined necessarily by its components, but mainly the motor process will be governed by its effect.'

3. Control of Activity by Goals and Other Mental (or Memory) Representations

Activity and actions are controlled by anticipations, i.e., the goals, which may form a hierarchy of super- and subordinate (or partial) goals. Goals are anticipations of the future results; motivationally, they are intentions to achieve these results by the person's own effort; from the point of view of memory they are the desired values, to be stored until the goal is completely achieved; emotionally, they are the starting points of specific task-inherent emotions (perception of success, failure, or flow), and from a personality point of view, goals and goal achievement are measures for self-assessment. Strings or hierarchies of goals are often reorganized into plans. Within a plan, the individual partial goal becomes a dependent component—a means—of a superordinate goal. For this reason, the sequence of individual goals is rationalized, and measures for goal achievement are integrated. For the

correspondences and differences within the conceptualization of goals in action regulation, see, for instance, Broadbent (1985), Kuhl and Atkinson (1986), and Locke (1991).

Furthermore, actions are controlled and led by mental representations of the conditions of action execution (e.g., the options of the machinery used). Action-guiding mental representations are a specific (e.g., response-compatible) type of mental representation. Some of these mental representations are not conscious ones (for further aspects, see *Action Planning, Psychology of*).

4. Stages of Control

Five stages are needed to describe the mental regulation of goal oriented activities (Tomasczewski 1981): (a) goal setting, i.e., redefinition of the task as the individual's goal, derived from his/her motives; (b) orientation towards the conditions of execution in the environment and in the actor's memory representations; (c) construction, or reproduction, of sequences of subgoals and the necessary measures; (d) decisions among particular versions of execution, if there is freedom to choose (autonomy); and (e) control of the execution by comparing the immediate results with the stored goals and possibly the plans. This control-loop shows the above-mentioned cyclical structure of action regulation. Functional units have to be represented in working memory, at least for the duration of execution: these units consist of goals, measures (programs), and the feedback or comparison processes just mentioned (Hacker 1985a, 1985b). Relationships between action theory and control theory are mentioned by Carver and Scheier (1990), among others.

From a motivational point of view, Heckhausen (1980) developed a sophisticated model of the sequential stages of goal-oriented activity, the Rubicon model. The main idea is that the action-preparing steps or stages show some different mental features in comparison with the action-accomplishing steps. The crucial transition from preparation to accomplishment is the specific decision actually to start the implementation ('to cross the Rubicon').

The above mentioned mental or memory representations are inevitable for the control of activity. All phases of activity are guided by them. It is useful to classify these representations into three types: (a) the goals or desired values; (b) the representations of the conditions of implementation; and (c) the representations of actions themselves, i.e. of the required patterns of operations, transforming the given state into the desired one.

The goal as the essential kind of internal representation, i.e., the anticipated result of the action, is the indispensable invariant regulator of every goal-directed process.

Goals are relatively stable memory representations that act as the necessary desired values during the implementation of an action. In the feedback processes mentioned, the actual state attained is compared with the goal as the required state.

5. Complete vs. Partialized Activity

As has been pointed out, activity will normally include, from a sequential or control loop point of view, preparation (goal setting, plan development, decision making), organization (co-ordination with other persons), execution of the intention, and checking the results in comparison with the stored goals. Checking, thus, produces the feedback closing the circle of the control loop or TOTE unit (Hacker 1986).

Correspondingly, from a hierarchical point of view, the regulating processes and representations of these phases, e.g., automated or intellectually controlled ones, are operating simultaneously at the aforementioned different levels of consciousness. The approach of 'complete vs. partialized (work) activity' enters here. This approach was launched as an ethical impulse by the German humanist philosopher Albert Schweitzer, discussing Western industrial culture (Schweitzer 1971). One can designate those activities as 'complete' that not only include routinized execution operations but also offer the opportunity for preparatory cognitive steps (e.g., goal development, decision making on the action programs), for checking the results, and for co-operation in terms of participating in the organization. Hence these 'complete' or 'holistic' activities are complete from the point of view of both the hierarchy of regulation levels and the cyclical control units. These sequentially (or cyclically) and hierarchically complete activities will offer the crucial option of learning, as opposed to the loss of skills and abilities in simple and narrow activities. Incomplete activities are sometimes called partialized activities.

6. Decision Latitude and Intrinsic Motivation

Decision latitude or autonomy is the most important variable of complete activity. Complete activity offers the decision latitude which is necessary for self-set goals, which are the prerequisite of comprehensive cognitive requirements of a task, and of intrinsic task motivation, i.e., motivation by a challenging task content ('task challenge').

Starting from the approach of decision latitude, Frese (1987) proposed a Theory of Control and Complexity of activity. He argued that one should distinguish among control over one's activity, complexity of activity, and complicatedness of activity: control here is defined in terms of the possibility to decide on the goals of activities and on the kind of goal attainment. Sometimes the amount of control is called decision latitude. Complexity refers to decision necessities, and complicatedness to those decision necessi-

ties that are difficult to control and socially and technologically unnecessary. Control should be enhanced, complexity should be optimized, and complicatedness should be reduced—in as much as working activities are concerned. Control should be increased at the expense of the other features because it has positive long-term consequences on performance, impairment, and well-being. The clearest finding of the Frese studies was that control has a moderating effect on the relationship between stressors and psychosomatic complaints at work: for the same amount of stressors given, complaints are lower with higher control.

7. *Activity Theory and Errors*

Activity theorists define errors as the nonattainment of an activity goal. The comparison of the activity outcome with the goal determines whether the goal has been achieved or whether further actions have to be accomplished. If an unintended outcome occurs, an error will be given. Consequently, a definition of an error based on Activity Theories integrates three aspects: (a) errors will only appear in goal-directed actions; (b) an error implies the nonattainment of the goal; (c) an error should have been potentially avoidable (Frese and Zapf 1991). Frese and Zapf (1991) developed an error taxonomy based on a version of Action Theory. This taxonomy and other comparable ones are inevitable in the examination of causes of errors and faults as a prerequisite of error prevention. Error prevention became an inevitable concern in modern technologies, e.g., the control rooms of nuclear power plants.

8. *Future Development*

Psychological theories are often short-term fashions, which are just launched and then soon pass away. Action Theories—there is still no 'one' Action Theory—seem to be an integrative long-term approach which is still developing especially with the development of subapproaches. Action Theories are still more a heuristic broad-range framework than final theories. Just this, however, might be their future advantage. The integrative power of Action Theories will bridge some inter-related gaps: the gap between cognition and action, the gap between cognition and motivation, and even the gap between basic and applied approaches—the last by fostering a dialogue between general (cognitive, motivational) psychology and the 'applied' disciplines (Hacker 1993). Perhaps its most challenging contribution might be to promote a reintegration of the breaking-off disciplines of psychology.

See also: Action Planning, Psychology of; Action Theory: Psychological; Activity Theory: Psychological; Intrinsic Motivation, Psychology of; Mental Representations, Psychology of; Motivation and Actions, Psychology of; Motivation: History of the Concept; Motivational Development, Systems Theory of; Vygotskij, Lev Semenovic (1896–1934); Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Bernstein N A 1967 *The Coordination and Regulation of Movements*. Pergammon, Oxford, UK

Broadbent D E 1985 Multiple goals and flexible procedures in the design of work. In: Frese M, Sabini J (eds.) *Goal Directed Behavior*. Erlbaum, Hillsdale, NJ, pp. 105–285

Carver C S, Scheier M F 1990 Principles of self-regulation. In: *Handbook of Motivation and Cognition: Foundations of Social Behavior*, Guildford Press, New York, Vol. 2, pp. 3–52

Frese M 1987 A theory of control and complexity: implications for software design and integration of computer systems into the work place. In: Frese M, Ulich E, Dzida W (eds.) *Human Computer Interaction in the Work Place*. Elsevier, Amsterdam, pp. 313–36

Frese M, Sabini J (eds.) 1985 *Goal Directed Behavior: The Concept of Action in Psychology*. Erlbaum, Hillsdale, NJ

Frese M, Zapf D (eds.) 1991 *Fehler bei der Arbeit mit dem Computer. Ergebnisse von Beobachtungen und Befragungen im Bürobereich*. Huber, Berne [Errors in Computerized Work]

Gollwitzer P M, Bargh J A (eds.) 1996 *The Psychology of Action—Linking Cognition and Motivation to Behavior*. Guilford Press, New York

Hacker W 1985a On some fundamentals of action regulation. In: Ginsburg G P, Brenner J, von Cranach M (eds.) *Discovery Strategies in the Psychology of Action* European Monographs in Social Psychology, Vol. 35. Academic Press, London, pp. 63–84

Hacker W 1985b Activity—a fruitful concept in psychology of work. In: Frese W, Sabini J (eds.) *Goal Directed Behavior*. Erlbaum, Hillsdale, NJ, pp. 262–84

Hacker W 1986 Complete vs. incomplete working tasks—a concept and its verification. In: Debus G, Schroiff W (eds.) *The Psychology of Work Organization*. Elsevier, Amsterdam, pp. 23–36

Hacker W 1993 Occupational psychology between basic and applied orientation—some methodological issues. *Le Travail Humain* **56**: 157–69

Heckhausen H 1980 *Motivation und Handeln*. Springer, Berlin [Motivation and Action]

Johnson-Laird P N 1983 *Mental Models. Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge, MA

Kuhl J, Atkinson J W 1986 *Motivation, Thought and Action. Personal and Situational Determinants*. Praeger, New York

Leontjev A N 1979 *Tätigkeit, Bewußtsein, Persönlichkeit*. Volk und Wissen, Berlin [Activity, Mind, Personality]

Lewin K 1926 Untersuchungen zur Handlungs- und Affektpsychologie. *Psychologische Forschung* **7**: 295–385 [Studies of the Psychology of Action and Affect]

Locke E A 1991 Goal theory vs. control theory: Contrasting approaches to understanding work motivation. *Motivation and Emotion* **15**: 9–28

Luria A R 1973 *The Working Brain*. Allen Lane, London

Miller G A, Galanter E, Pribram K-H 1960 *Plans and the Structure of Behavior*. Holt, New York

Rubinstein S L 1961 *Sein und Bewußtsein*. Akademie-Verlag, Berlin [Reality and Mind]
Schweitzer A 1971 Verfall und Wiederaufbau der Kultur. In: Schweitzer A (ed.) *Gesammelte Werke*. Union-Verlag, Berlin [Decline and Reconstruction of Culture], pp. 77–94
Tomaszewski T 1981 *Zur Psychologie der Tätigkeit*. Deutscher Verlag der Wissenschaften, Berlin [Psychology of Activity]
von Cranach M, Kalbermatten U, Indermühle K, Gugler B 1982 *Goal Directed Action*. Academic Press, London
von Weizsäcker V 1947 *Der Gestaltkreis*. Thieme, Stuttgart, Germany [The Gestalt Loop]

W. Hacker

Actor Network Theory

The term 'actor network theory' (ANT) combines two words usually considered as opposites: actor and network. It is reminiscent of the old, traditional tensions at the heart of the social sciences, such as those between agency and structure or micro- and macro-analysis. Yet, ANT, also known as the sociology of translation, is not just another attempt to show the artificial or dialectical nature of these classical oppositions. On the contrary, its purpose is to show how they are constructed and to provide tools for analyzing that process. One of the core assumptions of ANT is that what the social sciences usually call 'society' is an ongoing achievement. ANT is an attempt to provide analytical tools for explaining the very process by which society is constantly reconfigured. What distinguishes it from other constructivist approaches is its explanation of society in the making, in which science and technology play a key part.

This article starts by presenting the contribution of ANT to science and technology studies and then shows how this approach enables us to renew the analysis of certain classical problems in social theory.

1. Technosciences Revisited by ANT: Sociotechnical Networks

Spawned in the 1970s by the sociology of scientific knowledge, science studies strives, on the basis of empirical research, to explain the process in which scientific facts and technical artifacts are produced, and hence to understand how their validity and efficacy are established and how they are diffused. It follows two main approaches. The first has remained faithful to the project aimed at providing a social explanation for scientific and technical content (Collins 1985). The second, illustrated by ANT, has denied this possibility and embarked on a long-term undertaking to redefine the very object of social science.

For the promoters of ANT, the social explanation of scientific facts and technical artifacts is a dead end. 'Providing a social explanation means that someone is able to replace some object pertaining to nature by another pertaining to society' (Latour 2000). A scientific fact is thus assumed to be shaped by interests, ideologies, and so on, and a technological artifact crystallizes and reifies social relations of domination or power. Now, as research on scientific practices in laboratories and the design of technical artifacts shows, this conception, in which nature is dissolved in society, is no more convincing than the more traditional and cautious one in which the two are totally separate.

1.1 From World to Words

Let us enter a laboratory to observe the researchers and technicians at work. The laboratory is an artificial setting in which experiments are organized. The objects on which these experiments are performed, such as electrons, neutrinos, or genes, have been put in a situation in which they are expected to react or prove recalcitrant. It is the possibility of producing a discrepancy between what an entity is said to do and what it actually does that motivates the researcher to perform the experiment. This raises the question of the mysterious adequacy between words and things, between what one says about things and what they are.

To this classic philosophical question, ANT offers an original answer based on the notion of inscription (Latour and Woolgar 1986). Inscriptions are the photos, maps, graphs, diagrams, films, acoustic or electric recordings, direct visual observations noted in a laboratory logbook, illustrations, 3-D models, sound spectrums, ultrasound pictures, or X-rays as arranged and filtered by means of geometric techniques. All these inscriptions are produced by instruments. Researchers' work consists of setting up experiments so that the entities they are studying can be made 'to write' in the form of these inscriptions, and then of combining, comparing, and interpreting them. Through these successive translations researchers end up able to make statements about the entities under experimentation.

Inscription is two-sided. On the one hand it relates to an entity (e.g., an electron, gene, or neutrino) and, on the other, through combination with other traces or inscriptions, relates to propositions that have been tested by colleagues. Instead of positing a separation between words and things, ANT tracks the proliferation of traces and inscriptions produced in the laboratory, which articulate words and things. The analysis of this articulation leads to the two complementary concepts of a network and circulation.

Circulation should be understood in a traditional sense. The map drawn up by a geologist, based on readings in the field; the photos used to follow the trajectories identified by detectors in a particle accelerator; the multicolored strips stacked on a chrom-

atograph; the tables of social mobility drawn up by sociologists; the articles and books written by researchers: all these circulate from one laboratory to the next, from the research center to the production unit, and from the laboratory to the expert committee which passes it on to a policy maker. When a researcher receives an article written by a colleague, it is the genes, particles, and proteins manipulated by that colleague in her or his own laboratory that are present on the researcher's desk in the form of tables, diagrams, and statements based on the inscriptions provided by instruments. Similarly, when political decision makers read a report that asserts that diesel exhaust fumes are responsible for urban pollution and global warming, they have before them the vehicles and atmospheric layers that cause that warming. We thus move away from classical epistemology, which opposes the world of statements, on the one hand, and that supposedly 'other' world (more or less real) of the things to which the statements refer and which, in a sense, constitutes the context of those statements, on the other. Referents do not lie outside the world of statements; they circulate with them and with the inscriptions from which they are derived.

By circulating, inscriptions articulate a network qualified as sociotechnical because of its hybrid nature (Latour 1987). The sociotechnical network to which the statement 'the hole in the ozone layer is growing' belongs, includes all the laboratories working directly or indirectly on the subject, eco-movements, governments that meet for international summits, the chemical industries concerned, and the parliaments that pass laws, as well as the chemical substances and the reactions they produce, and the atmospheric layers concerned. The statement 'the ozone layer is disappearing due to the use of aerosols' binds all these elements, both human and nonhuman. At certain points in these networks we find translation centers, which capitalize on all the inscriptions and statements. Inscriptions are information and consequently those centers are able to act at a distance on elements without bringing them in for good. But inscriptions can also be accumulated and combined: new unexpected connections are produced that explain why the translation centers are endowed with the capacity for strategic and calculated action; they can conceive of certain states of the world (e.g., without a hole in the ozone layer), and identify and mobilize the elements with which to interact to produce the desired state. Such strategic action is possible only because the sociotechnical network exists. Action and network are thus two sides of the same reality; hence, the notion of an actor network.

1.2 Black-boxing Collective Action

Technology can be analyzed in the same way. The social explanation of technological artifacts raises the same difficulties as that of scientific facts. Once again, it is by jettisoning the idea of a society defined *a priori*, and replacing it by sociotechnical networks that ANT avoids the choice between sociological reductionism, on one hand, and positing a great divide between techniques and societies, on the other.

Consider a common artifact such as the automobile. Its phenomenal success is probably due to the fact that it enables users to extend the range and variety of actions they can successfully undertake, freeing them to travel about without having to rely on anyone else. Thus, autonomous users endowed with the capacity to decide where they want to go, and to move about as and when they wish, are 'inscribed' in the technical artifact itself, the automobile (Akrich 1992). Paradoxically, the driver's autonomy stems from the fact that the functioning of the automobile depends on its being but one element within a large sociotechnical network. To function, it needs a road infrastructure with maintenance services, motorway operating companies, the automobile manufacturing industry, a network of garages and fuel distributors, specific taxes, driving schools, traffic rules, traffic police, roadworthiness testing centers, laws, etc. An automobile is thus at the center of a web of relations linking heterogeneous entities, a network that can be qualified as sociotechnical since it consists of humans and nonhumans (Callon et al. 1986).

This network is active, which again justifies the term actor network. Each of the human and nonhuman elements comprising it participates in a collective action, which the user must mobilize every time he or she takes the wheel of his or her automobile. In a sense the driver then merges with the network that defines what he or she is (a driver-choosing-a-destination-and-an-itinerary) and what he or she can do. When the driver turns the ignition key of a Nissan to go meet a friend on holiday at Lake Geneva, the driver not only starts up the engine, but also triggers a perfectly coordinated collective action. This action involves: the oil companies that refined the oil, distributed the petrol, and set up petrol stations; the engineers who designed the cylinders and valves; the machines and operators who assembled the vehicle; the workers who laid the concrete for the roads; the steel that withstands heat; the rubber of the tires that grip the wet road; the traffic lights that regulate the traffic flow, and so on. We could take each element of the sociotechnical network to show that, human or nonhuman, it contributes in its own way to getting the vehicle on the road. This contribution, which was progressively framed during the establishment of the sociotechnical network, is not reducible to a purely instrumental dimension. In its studies of technological innovation, ANT stresses the ability of each entity, especially nonhuman ones, to act and interact in a specific way with other humans or nonhumans. The automobile—and this is what defines it as a technical artifact—makes it possible, in a place and at a point in time, to

use a large number of heterogeneous elements that silently and invisibly participate in the driver's transportation. We may call these elements 'actants,' a term borrowed from semiotics, which highlights the active nature of the entities comprising the network. We could also say that this collective action has been black-boxed in the form of an artifact—here, an automobile. When it moves, it is the whole network that moves.

Sometimes, however, black boxes burst open. Thus, the role of these actants becomes explicitly visible when failures or incidents occur: petrol transporters go on strike; war breaks out in the Middle East; a road collapses; taxes increase the price of petrol in a way considered unacceptable; environmental standards curb the use of internal combustion engines; a driver's concentration flags; alloys tear because they are not resistant enough to corrosion; automobile bodies rip open on impact. At these times, the collective action becomes visible and all the actants who contributed to the individual and voluntary action of the driver are unveiled (Jasanoff 1994, Wynne 1988). But it is during the historical constitution of these sociotechnical networks, that is, during the conception, development, and diffusion of new technical artifacts, that all the negotiations and adjustments between human and nonhuman actants, preceding the black-boxing, most clearly appear. And it is to such processes of constitution that ANT directs its attention (Law 1987).

In the cases of both science and technology, the notion of sociotechnical networks is at the heart of the analysis. ANT has put a considerable effort into analyzing the process of construction and extension of these networks. Concepts such as 'translation,' '*interessement*' (a term borrowed from French) and 'the spokesperson' have been developed to explain the progressive constitution of these heterogeneous assemblages (Callon 1986). To account for either scientific facts or technical artifacts ANT refuses to resort to a purely social explanation, for ANT replaces the purity of scientific facts and technical artifacts with a hybrid reality composed of successive translations. These networks can be characterized by their length, stability, and degree of heterogeneity (Callon 1992; Bowker and Star 1999). This viewpoint necessarily challenges traditional conceptions of the social, an issue we shall now examine.

2. *Making up Hybrid Collectives*

For ANT, society must be composed, made up, constituted, established, maintained, and assembled. There is nothing new about this assertion, as such; it is shared by many constructionist currents. But ANT differs from these approaches in the role it assigns to nonhumans in the composition of society. In the traditional view, nonhumans are obviously present, but their presence resembles that of furniture in a bourgeois home. At best, when these nonhumans take the form of technical artifacts, they are necessary for the daily life they facilitate; at worst, when they are present in the form of statements referring to entities such as genes, quarks, or black holes, they constitute elements of context, a frame for action. To the extent that they are treated as lying outside the social collective or as instrumentalized by it, nonhumans are in a subordinate position. Similarly, when the topic of analysis is institutions, organization, or rules and procedures, social analysts assume that these meta-individual realities are human creations, like the technical artifacts that supplement daily life. The social sciences are founded on this great divide between humans and nonhumans, this ontological asymmetry that draws a line between the social and the nonsocial.

However, the past two decades of science and technology studies have caused this division to be called into question. Moreover, as we have seen, in the laboratory nonhumans act and, because they can act, they can be made to write and the researcher can become their spokesperson. Similarly, technical artifacts can be analyzed as devices that at some point capitalize on a multitude of actants, always temporarily. Society is constructed out of the activities of humans and nonhumans who remain equally active and have been translated, associated, and linked to one another in configurations that remain temporary and evolving. Thus, the notion of a society made of humans is replaced by that of a collective made of humans and nonhumans (Latour 1993). This reversal has numerous consequences. We shall stick to a single example, that of the distinction between macro and micro levels, which has been replaced by framed and connected localities.

Does a micro level exist? The answer seems obvious. When our motorist takes to task another motorist who refused him right of way, or when he receives a traffic fine, he enters into interactions with other perfectly identifiable individual actors. Generally speaking nothing else than interactions between individuals has ever been observed. Yet it seems difficult to simply bracket off realities like institutions or organizations that obviously shape and constrain the behavior of individual agents, even when they are considered as the unintentional outcome of the aggregation of numerous individual actions. To avoid this objection (and the usual solutions that describe action as simultaneously structuring and structured), ANT introduces the notion of locality, defined as both framed and connected.

Interactions, like those between motorists arguing with each other after an accident, or between them and the traffic policeman who arrives on the scene, take place in a frame that holds them. In other words, there are no interactions without framing to contain them. The mode of framing studied by ANT extends that analyzed by Goffman, by emphasizing the active part

played by nonhumans who prevent untimely overflowing. The motorists and traffic officers are assisted, in developing their argument about how the accident occurred, by the nonhumans surrounding them. Without the presence of the intersection, the traffic lights that were not respected, the traffic rules that prohibit certain behaviors, the solid lines that 'materialize' lanes, and without the vehicles themselves that prescribe and authorize certain activities, the interaction would be impossible, for the actors could give no meaning to the event and, above all, could not properly circumscribe and qualify the incident itself.

This framing which constrains interactions by avoiding overflowing is also simultaneously a connecting device. It defines a place (that of the interaction) and at the same time connects it to other places (where similar or dissimilar accidents have taken place, where the policemen go to write up reports, or where these reports land up, etc.). All the elements that participate in the interaction and frame it establish such connections for themselves. The motorist could, for example, invoke a manufacturing defect, the negligence of a maintenance mechanic, a problem with the traffic signals, the bad state of the road, the traffic officer's lack of training, etc. Suddenly the circle of actants concerned has become substantially bigger. Through the activities of the traffic officer, the automobile, and the infrastructure which all together frame interactions and their implications, other localities are associated with those of the accident: the multiple sites in which automobile manufacturers, the networks of garage owners, road maintenance services, and police training schools act. Instead of microstructures, there are now locally framed interactions; instead of macrostructures, there are connected localities, because framing is also connecting.

With this approach it is possible to avoid the burdensome hypothesis of different levels, while explaining the creation of asymmetries, i.e., of power relations, between localities. The more a place is connected to other places through science and technology, the greater its capacity for mobilization. The translation centers where inscriptions and statements converge provide access to a large number of distant and heterogeneous entities. The technical artifacts present in these far-off places ensure the distant delegation of the action decided in the translation center. On the basis of the reports and results of experiments it receives, a government can, for example, decide to limit CO_2 emissions from cars to a certain level. As a translation center it is in a position to establish this connection between the functioning of engines and the state of pollution or global warming. It sees entities and relations that no one else can see or assemble. But the application of this decision implies, among other things, the setting up of pollution-control centers and the mobilization of traffic officers to check that the tests have been performed, and if necessary to fine motorists, on the basis of a law passed by parliament. Thus, the action decided by the translation center mobilizes a large number of human and nonhuman entities who actively participate in this collective and distributed action. Just as the motorist sets in motion a whole sociotechnical network by turning the ignition key, so the minister for the environment sets in motion an elaborately constructed and adjusted network by deciding to fight pollution. The fact that a single place can have access to other places and act on them, that it can be a translation center *and* a center for distant action—in short, that it is able to sum up entire sociotechnical networks—explains the asymmetry that the distinction between different levels was supposed to account for.

3. *ANT : An Open Building Site*

ANT is an open building site, not a finished and closed construction (Law and Hassard 1999). It is itself more an inspirational frame than a constraining theoretical system (Star and Griesemer 1989, Singleton and Michael 1993, Star 1991, Lee and Brown 1994, Mol and Law 1994. Moreover, many points remain controversial. Its analysis of agency (and, in particular, the symmetry it postulates between humans and nonhumans) has been strongly criticized (Collins and Yearley 1992). For ANT this principle of symmetry is not a metaphysical assertion but a methodological choice which facilitates the empirical study of the different modalities of agency, from strategic to machine-like action. In all cases, agency is considered to be distributed and the forms it takes are linked to the configuration of sociotechnical networks. The opposition between structure and agency is thus overcome. In the 1990s, researchers inspired by ANT have moved into new fields such as organization studies (Law 1994) and the study of the formation of subjectivity or the construction of the person (Law 1992). After including nonhumans in the collective, ANT strives to analyze how socialized things participate, particularly through animate bodies, in the creation of subjectivities (Akrich and Berg, in press).

In parallel with work on the role of the hard sciences and technology in the construction of collectives, ANT also analyses the contribution of social science to the creation of society. It notes that the social sciences are no more content with just offering an analysis of a supposed society than the natural sciences are content just to describe a supposed nature. This point has been made in detail for economics. If we consider an extended definition of economics, including accounting, marketing, management science, etc., it is possible to study how a social science (here, economics) helps to format markets and economic agents such that organized modern markets are embedded in economics (Callon 1998). This approach, extended to the other social sciences such as sociology, psychology, anthropology, or political science, should

facilitate better understanding of the process through which society tends to think of itself as distinct from its environment, and that of its internal differentiation. It is by refusing to countenance, on a methodological level, the great divides postulated by the sciences (both natural and social), that ANT is in a position to explain, on a theoretical level, the role of the sciences in their construction and evolution.

See also: Constructivism/Constructionism: Methodology; Scientific Knowledge, Sociology of; Social Constructivism; Technology, Social Construction of

Bibliography

Akrich M 1992 The description of technical objects. In: Bijker W, Law J (eds.) *Shaping Technology/Building Society. Studies in Sociotechnical Change*. MIT Press, Cambridge, MA, pp. 205–24

Bowker G C, Star S L 1999 *Sorting Things Out. Classification and Its Consequences*. MIT Press, Cambridge, MA

Callon M 1986 Some elements for a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In: Law J (ed.) *Power, Action and Belief. A New Sociology of Knowledge?* Routledge and Kegan Paul, London, pp. 196–229

Callon M 1992 The dynamics of techno-economic networks. In: Coombs R, Saviotti P, Walsh V (eds.) *Technological Change and Company Strategies*. Academic Press, London, pp. 72–102

Callon M (ed.) 1998 *The Laws of the Markets*. Blackwell, London

Callon M, Law J et al. (eds.) 1986 *Mapping the Dynamics of Science and Technology. Sociology of Science in the Real World*. Macmillan, London

Collins H 1985 *Changing Order. Replication and Induction in Scientific Practice*. Sage, London

Collins H, Yearley S 1992 Epistemological chicken. In: Pickering A (ed.) *Science as Practice and Culture*. University of Chicago Press, Chicago, pp. 301–26

Jasanoff S (ed.) 1994 *Learning From Disaster: Risk Management After Bhopal*. University of Pennsylvania Press, Philadelphia

Latour B 1987 *Science In Action. How to Follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge, MA

Latour B 1993 *We Have Never Been Modern*. Harvester Wheatsheaf, Hemel Hempstead, UK

Latour B 2000 When things strike back. A possible contribution of 'science studies' to the social sciences. *British Journal of Sociology* **51**: 105–23

Latour B, Woolgar S 1986 *Laboratory Life. The Construction of Scientific Facts*. Princeton University Press, Princeton, NJ

Law J 1987 Technology and heterogeneous engineering: The case of Portuguese expansion. In: Bijker W E, Hughes T P, Pinch T (eds.) *The Social Construction of Technological Systems. New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, MA, pp. 111–34

Law J 1992 Notes on the theory of the actor network theory: Ordering, strategy and heterogeneity. *Systems Practice* **5**: 379–93

Law J 1994 *Organizing Modernities*. Blackwell, Oxford, UK

Law J, Hassard J (eds.) 1999 *Actor Network Theory and After*. Blackwell, Oxford, UK

Lee N, Brown S 1994 Otherness and the actor network: The undiscovered continent. *American Behavioral Scientist* **37**(6): 772–90

Mol A, Law J 1994 Regions, networks and fluids: Anemia and social topology. *Social Studies of Science* **24**(4): 641–71

Singleton V, Michael M 1993 Actor networks and ambivalence: General practitioners in the UK cervical screening program. *Social Studies of Science* **23**: 227–64

Star S L 1991 Power, technologies and the phenomenology of conventions: On being allergic to onions. In: Law J (ed.) *A Sociology of Monsters. Essays on Power, Technology and Domination*. Routledge, London, pp. 26–56

Star S L, Griesemer J 1989 Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939. *Social Studies of Science* **19**: 387–420

Wynne B 1988 Unruly technology: Practical rules, impractical discourses and public understanding. *Social Studies of Science* **18**: 147–67

M. Callon

Adaptation, Fitness, and Evolution

The adaptation, or adaptedness, of organisms to their environments is a central concept in evolutionary biology. It is both a striking phenomenon demanding explanation and an essential feature of the mechanisms underlying the patterns of evolutionary stasis and change alike. The organism–environment interaction which the adaptation concept embodies is the causal driver of the process of evolution by natural selection. Its nature, role in the evolutionary concept structure, and limitations must all be understood if a clear view of evolution is to be possible. In particular, adaptation's distinctness from and relation to the concept of fitness must be seen clearly. Only thus can evolution by natural selection, a central and perhaps the only 'natural law' peculiar to the life sciences (Rosenberg 2000, Watt 2000), be properly understood.

1. Adaptation's Identity and its Distinction from Fitness

If no concept is more central to evolution by natural selection than adaptation, then also none has been more debated. All the basic features of its definition are found in the work of Darwin, but progress in unfolding its full scope and implications continues to be made even at present.

Biological evolution, as distinct from cultural evolution (though often interwoven with it; e.g., Cavalli-Sforza and Feldman 1983), is manifested as change in the genetic composition of populations over time. Therefore, some genetic terminology is needed at the

outset. A 'gene' is a functionally coherent sequence of bases in nucleic acid, usually DNA except for some viruses, determining or influencing some biological structure and/or function. An 'allele' is one possible sequence of a gene, determining one alternative state of gene action. Many organisms, including most animals, carry two copies of each gene (and are thus termed 'diploid'). 'Genotype' refers to the whole heritable composition of a creature, whether viewed gene-by-gene (e.g., carrying two copies of the same allele, hence a 'homozygote,' or one copy each of two different alleles, hence a 'heterozygote') or more broadly up to the whole 'genome' which includes all genes. 'Phenotype' refers to the expressed structure and function of the organism as it develops via interactions of its genotype with the environment in which development takes place. Present understanding of the complexities of the evolutionary process requires this terminology to avoid ambiguity and confusion.

1.1 Basic Definitions of Adaptation

As a general concept, adaptation or adaptedness is best defined as an extent or degree of matching or suitedness between the heritable features (heritable functional phenotypes) of organisms and the environments in which they occur. It finds direct expression in the effectiveness with which organisms perform their characteristic biological tasks—osmoregulation, locomotion, capturing food, evading predators, etc.—in their environments. As such, its states are in principle quantitatively measurable, or at least orderable, rather than only qualitatively organized.

This general definition is exemplified in many parts of Darwin's writing, notably in the Introduction to his *Origin of Species* (1859) where a woodpecker appears as an exemplar 'with its feet, tail, beak, and tongue, so admirably adapted to catch insects under the bark of trees.' Here, the phenotypic states of these morphological characters, modified as compared to simpler and more general forms found in other birds, are related to their functional performance consequences in acquiring food resources which other birds, lacking those specific adaptive phenotypic states, cannot reach.

Adaptation also refers to the process of successively descended, modified phenotypes becoming more suited, 'better adapted,' to particular environments under the action of natural selection on variation in those phenotypes. Palaeontology provides strong evidence for such local improvement in adaptedness over time, as in the escalation of predator and prey attack and defense morphologies in marine invertebrates (Vermeij 1987). Recently, real-time experiments have shown such adaptive improvement directly, in the evolution of a bacterial stock in novel culture conditions over periods of 10^4 generations: a stock at an early stage in the process, if samples are frozen for later reactivation, is found to be inferior in performance to its own, manifestly better adapted, descendants sampled from late in the history of the stock (Lenski and Travisano 1994).

The historical nature of this successive evolutionary refinement of adaptive state has led to much debate over when a phenotypic feature may be termed 'an' adaptation and when not, i.e., how far it has been specifically selected for its current state of function in its environment. If, as stated above, states of adaptation differ quantitatively, then any viable phenotype represents *some* level of adaptation, and this debate loses urgency. Further, calling a phenotype 'an' adaptation only if it is the best available at a given time (as advocated by, e.g., Reeve and Sherman 1993) would require continuous redefinition as newer alternatives arise, and seems to offer no compensating advantage.

1.2 Elaborations of the Basic Concept

Gould and Vrba (1982) extended and refined definitions of adaptation in useful ways. In their terminology, 'aptation' describes the primary, historically unmodified relation of suitedness between phenotype and environment—that of any newly arisen variant, positive or negative in its functional effects. They regarded 'adaptation,' in this context, as the successive refinement of phenotypic suitedness by selection of newer variants, and coined the term 'exaptation' for the co-opting of a phenotypic feature by selection for a new function, as in the modification of skull-jaw joint bones toward the ossicles of vertebrate ears (e.g., Romer 1955). The exaptation/adaptation distinction poses problems of discrimination (how much change under a new selection pressure is needed before a phenotype of *ex*aptive origin is recognizable as presently *ad*aptive? Reeve and Sherman 1993), and also emphasizes that we are dealing with quantitative scales of variation, not alternate qualitative categories. Often the Gould–Vrba terms are not used unless the distinctions are pertinent to the issue at hand, and otherwise 'adaptation' is used as a generally inclusive term.

Another important augmentation of the adaptation concept is the work of Laland et al. (1996, 1999) in modeling 'niche construction.' This term refers to the active modification of environments by organisms in ways favorable to their own function, as emphasized by Lewontin (1983). It occurs in very diverse ways in different groups: for example, bacteria may release protease enzyme catalysts into their surroundings to aid foraging upon potential food items, while among multicellular animals beaver lodges and dams are a classic and dramatic example of such activities (quite aside from the obvious and often environmentally destructive capabilities of humans in this direction). Evolutionary models incorporating niche-construc-

tive feedbacks on organism–environment interactions may have very distinct properties from those not including such active forms of adaptation (Laland et al. 1996, 1999).

1.3 The Distinction Between Adaptation and Fitness

Alternative states of adaptation are the causes of evolutionary changes through their differences in organism–environment interaction and hence habitat-specific performances of these phenotypes. The performances of differently adaptive phenotypes, minute by minute or day by day, cumulatively affect how long they live, and how much they reproduce while alive. In short, adaptive differences among phenotypes alter their demographic parameters: survivorship ($= l_x$ of demography, where x denotes intervals of time) and male mating success or female fecundity ($= m_x$ of demography). These parameters are components of what, since the advent of mathematical population genetics, has been termed 'fitness' or 'Darwinian fitness' (though Darwin did not use the word in this way): the reproductive success of whole populations or of specific genotypes. Adaptation and fitness, then, are serially related concepts, but are in no sense the same.

In evolutionary genetics, fitness is usually measured as the net reproductive rate or replacement rate of organisms, whether an average value for a whole population or more narrow average values specific to carriers of particular genotypes. It is defined in 'absolute' terms as $R = \Sigma l_x m_x$ (e.g., Roughgarden 1979) under simple demographic conditions of non-overlapping generations and homogeneous reproductive periods (as, e.g., in annual plants or many insects). For complex demography in age-structured populations, the closest equivalent expression is λ (the leading eigenvalue of the demographic 'Leslie matrix'), a number which expresses complex interactions of age-specific survivorships and fecundities (e.g., Charlesworth 1994, McGraw and Caswell 1996). If either R or λ, as appropriate, is compared among genotypes by taking the ratio of each value to that of a chosen standard genotype, there result 'relative' genotypic fitnesses, wose value for the standard genotype is 1. Most evolutionary-genetic models use relative fitnesses for symbolic or numeric convenience.

It is essential to recognize that usage of the terms adaptation and fitness has changed dramatically since Darwin. He, Wallace, and other early evolutionists used 'fitness' as a synonym for 'adaptation,' and by 'survival' they often referred not to the demographer's life-cycle variable l_x but to 'persistence over long time periods.' Spencer's phrase 'survival of the fittest,' translated, thus meant 'the persistence through time of the best adapted.' Darwin had (necessarily) a clear view of the concept which an evolutionary geneticist now denotes by 'Darwinian fitness,' but he represented it by one version or another of a stock phrase (for which he had no summary term): '… the best chance of surviving and of procreating …' in the *Origin of Species* (1859). Failure to recognize these usage changes, and thus blurring of the sharp distinction between adaptation as cause and fitness as within-generation result, has led to no small confusion in later literature, including fallacious claims of an alleged circularity of evolutionary reasoning.

2. The Roles of Adaptation and of Fitness in Darwin's Argument for Natural Selection

From the inceptions of both Darwin's and Wallace's ideas of natural selection, differences in adaptation among heritable variants played the central, causal role in the process. Darwin formalized his argument in Chapter 4 of the *Origin*, especially in its first paragraph and its concluding summary, in such a way that it can be cast as a verbal theorem—as Depew and Weber (1995) make clear by judicious editing of Darwin's summary. Here it may be reformulated in modern terms.

To begin the argument, there are three points 'given' by direct observation: (a) organisms vary in phenotype; (b) some of the variants are heritable; and (c) some of these heritable variants are differently able to perform their biological functions in a given specific habitat, i.e., some are better adapted than others to that habitat. Then, Darwin's Postulate is that the better adapted, hence better performing variants in a habitat will survive and/or reproduce more effectively over their life cycles, i.e., have higher fitness, than other variants. Demography shows that greater reproduction of variants will cause their maintenance or increase of frequency in successive generations of a population. One may thus conclude that when the Postulate holds, the best adapted heritable phenotypes will persist and/or increase in frequency over time, thus realizing evolution by natural selection. This completes Darwin's Theorem.

The distinction between differences in diverse adaptive performances minute by minute or day by day in organisms' experience, and the resulting, cumulative appearance of fitness differences among them over their whole life spans is quite simply the difference between cause and effect. Its recognition is essential to keep straight the logic of natural selection, and to organize empirical studies of the process (Feder and Watt 1992, Watt 1994). The causal basis, in natural organism–environment interactions, of adaptive performance differences among genetic variants, and the transformation rules which translate those performance differences into fitness consequences, are now subjects of active and increasingly diverse study by evolutionary biologists.

3. *Alternatives to Adaptation in Evolution*

Adaptation is not ubiquitous, and natural selection is not all-powerful. 'Darwin's Theorem,' as summarized above, is not only empirically testable, but indeed may not hold in some well-defined circumstances. Two principal sources of limitation on the scope of adaptation are now considered.

3.1 *Neutrality*

As Darwin said in Chapter 4 of the *Origin*, 'Variations neither useful nor injurious would not be affected by natural selection …' The modern concept of neutrality (Kimura 1983, Gillespie 1991), which he thus described, is the null hypothesis for testing all causal evolutionary hypotheses. It occurs at each of the recursive stages of natural selection, as recognized by Feder and Watt (1992).

First, at the **genotype → phenotype** stage, genetic variants may differ in sequence but not in resulting function. For example, the 'degeneracy' of the genetic code often means that differences in DNA base sequence lead only to the same amino acid's insertion into a given position of a protein molecule. Alternatively, at least at some positions in proteins, substitution for one amino acid residue by a similar one, e.g., valine by isoleucine, may have little effect on the protein's function.

Next, at the **phenotype → performance** stage, functional differences among variants may not lead to performance differences among them, as other phenotypic mechanisms constrain or suppress their potential effects. For example, in the physiological reaction pathway used by bacteria to digest milk sugar (lactose), a twofold range of natural genetic variation in a phenotypic parameter (the V_{max}/K_m ratio) is observed for each of the protein catalysts, or enzymes, catalyzing the first two reactions. When these variants' resulting performances were measured under steady-state growth conditions, variants of the first enzyme in the pathway showed sizeable, reproducible differences, but no such effects were seen among variants of the second enzyme despite the similar size of their phenotypic differences—due to system constraints related to the position of the reactions in the pathway, analyzable by the theory of physiological organization (Watt and Dean 2000).

At the stage of **performance → fitness,** performance differences may not lead to corresponding fitness differences, e.g., if improved performance has less fitness impact above a threshold value of habitat conditions. For example, performance differences among feeding phenotypes (bill sizes and geometries) of Darwin's finches have minimal fitness impact when food is abundant in wet seasons, but have much more impact when food is scarce in dry seasons (Grant 1986).

Finally, at the stage of **fitness → genotype**, which completes the natural-selective recursion, small population size can allow random genetic drift to override fitness differences, as in the loss from small mouse populations of developmental mutant alleles ('t-system' variants) which should be in frequency equilibrium between haploid, gametic selection favoring them and their recessive lethality at the diploid, developing-phenotypic stage of the life cycle (Lewontin and Dunn 1960).

Because the usual statistical null hypothesis is that no treatment effect exists between groups compared, any adaptive hypothesis of difference between heritable phenotypes is *ipso facto* tested against neutrality by statistical testing.

Further, there is a subtler neutral hypothesis, that of association or 'hitchhiking': variants seeming to differ in fitness at a gene under study may be functionally neutral but genetically linked to an unobserved gene whose variants are the real targets of selection. But, 'hitchhiking' predicts that fitness differences seen among variants will not follow from any functional differences among them, so it is rejected when prediction from function to fitness is accurate and successful.

Indeed, where substantive adaptive difference exists among genetic variants in natural populations, neutral null hypotheses may be rejected under test at each of these levels, from phenotypic function to its predictable fitness consequences and the persistence or increase of the favored genotypes. This has been done, for example, for natural variants of an energy-processing enzyme in the 'sulfur' butterflies, *Colias* (Watt 1992). The explicit test of adaptive hypotheses against neutral nulls gives much of its rigor to such experimental study of natural selection in the wild (Endler 1986).

3.2 *Constraint*

Gould (e.g., 1980, 1989) has emphasized that many features of organisms may not result from natural selection at all but rather from various forms of constraint due to unbreakable geometric or physical properties of the universe at large or of the materials from which organisms are constructed, or other, more local biological limitations or conflicts of action.

Geometric or topological constraints may take a major hand in the form or function of organisms, e.g., in snail shells' form (Gould 1989) or in the fractional-power scaling of metabolic processes with body mass (West et al. 1997).

Selection among phenotypic alternatives at one time may entail diverse predispositions or constraints at later times. In one such case, the tetrapodal nature of all land-dwelling vertebrate animals (the bipedality of birds, kangaroos, and hominid primates is secondary) follows from the historical constraint that their ancestors, certain sarcopterygian fish, swam with two pairs

of oar-like ventral fins having enough structural strength *ab initio* that they could be exaptively modified into early legs (e.g., Gould 1980, Cowen 1995). In a more pervasive case, the evolved rules of diploid, neo-Mendelian genetics constrain many evolutionary paths. For example, if a heterozygous genotype is the best adapted, hence most fit, in a population, it can rise to high frequency in that population but cannot be the only genotype present because it does not 'breed true.'

Conflicts among different aspects of natural selection may constrain the precision of adaptation in diverse ways. As a case in point, adjustment of insects' thermoregulatory phenotypes may be held short of maximal or 'optimal' matching to average conditions in cold, but highly variable, habitats, because such 'averagely optimized' phenotypes would overheat drastically in uncommon but recurrent warm conditions (Kingsolver and Watt 1984). This illustrates the general point that environmental variance may sharply constrain adaptation to environmental means.

4. *Misdefinitions of Adaptation or Misconceptions of its Role*

Many misdefinitions of adaptation err by confusing it with fitness in one fashion or another. Much of this may originate in the usage changes, discussed earlier, between the early Darwinians and the rise of evolutionary genetics, such that 'fitness' ceased to be a synonym of adaptation and came to mean instead the 'best chance of surviving and of procreating' (e.g., Darwin 1859, p. 63). This entirely distinct concept is, as noted above, the cumulative demographic effect of adaptation. Some writers on evolutionary topics have been confused by inattention to these usage changes, but others have erred through conscious disregard or blurring of the adaptation–fitness distinction.

For example, Michod, despite early recognition of the separate nature of adaptation and fitness and of their antecedent–consequent relationship (Bernstein et al. 1983), has recently (Michod 1999) sought to collapse these concepts into different 'senses' of the single term 'fitness' to be used in different contexts to refer to both 'adaptive attributes' and their consequences in reproductive success. Authors may choose terminology for their own uses within some limits, but this usage is at best an ill-advised source of confusion, and at worst a mistaken conflation of distinct concepts.

A more subtle misconception was asserted by Lewontin (1983) in the course of an otherwise important argument for studying 'niche construction' (cf. above; Laland et al. 1996, 1999). Arguing that Darwin's adaptation concept implies 'passiveness' on the part of adapting organisms, he criticized it for allegedly implying that adaptation is like 'filing a key to fit a pre-existing lock.' But no such passivity is really in evidence. In Darwin's example already mentioned, the woodpecker's feeding 'strategy' actively transforms its environment compared to that experienced by more generally feeding birds, using a resource that its more generalized relatives do not even perceive. Further, Darwin's discussions (1859, Chap. 4) of co-adaptive mutualism between flowers and pollinators also show the active, indeed constructive nature of many adaptations. Pollinators not only obtain resource rewards from plants and spread their pollen during their foraging, but in so supporting the reproduction of their food sources, they increase their own future resource bases. Niche construction is therefore *an important form of* adaptation, not distinct from it or opposed to it.

Lewontin has also mis-stated the role of adaptation in the evolutionary process, arguing that 'three propositions'—variation, heritability, and differential reproduction—alone were sufficient to explain natural selection, and that the adaptation concept was gratuitously introduced into the argument by Darwin for sociological reasons (e.g., 1984). This claim is wrong, and has been extensively critiqued (e.g., Hodge 1987, Brandon 1990, Watt 1994, Depew and Weber 1995). Adaptation is the one element which distinguishes natural selection from artificial selection or from sexual selection alike. Without it, Lewontin's three propositions are sufficient only to define 'arbitrary' selection, wherein we do not know the cause of any differential reproduction of heritable variants. But the adaptive cause is, indeed, central to evolutionary change resulting from natural selection.

5. *Adaptationism and its Drawbacks*

Rose and Lauder (1996) identify adaptationism as '… a style of research … in which all features of organisms are viewed *a priori* as optimal features produced by natural selection specifically for current function.' Some, e.g., Parker and Maynard Smith (1990) or Reeve and Sherman (1993), hail the assumption of adaptiveness as a virtue, while others, e.g. Gould and Lewontin (1979), have attacked it as a vice. The question is: is it helpful, or legitimate, to assume that adaptation is ubiquitous?

First, is it true that adaptiveness is often assumed in practice? The usual null hypothesis in statistical testing is that there is no 'treatment effect.' Thus, any statistical test of adaptive difference among character states assumes *ab initio* that there is *no* such difference, i.e., that the character states in question are *neutral*. Only if this null hypothesis can be rejected according to standard decision rules is an effect recognized. All the null models of population genetics itself, beginning with the single-gene Hardy–Weinberg distribution, start with neutral assumptions. Tests of the population-genetic consequences of putatively adaptive differences in phenotypic mechanisms may or may not find departure from neutrality, but it is the routine 'starting point.'

Mayr (1988), among others, argues for testing all possible adaptive explanations for phenotypes before considering the 'unprovable' explanation of chance, i.e., neutral, origins. But this argument depends on a historicist approach to evolutionary studies. If one can instead analyze a phenotype by testing among neutrality, constraint, or adaptation with present-day experiments, historicism is no longer entailed. Even fossil structures unrepresented in living descendants can often be studied functionally by various means (Hickman 1988). A historical approach may sometimes be indispensable, but it is not the only one available to evolutionary biology.

As Gould (1980) observed, assuming the ubiquity of adaptation has a strong tendency to discourage attention to structural or constrained alternative explanations of phenotypes. It is not enough merely to test one specifically adaptive hypothesis about some phenotype against neutrality; one should consider all feasible alternative hypotheses, including the constraint-based one that a phenotype does make a non-neutral difference to performance and thence fitness, but does so in a particular way because no other is feasible or possible, rather than because it is 'optimized.'

Indeed, the strongest objection to adaptationism may be that attempting to use the optimal adaptedness of a phenotypic feature as a null hypothesis (an alternative sometimes suggested by adherents of this view) runs the serious risk of falling victim to 'the perils of preconception.' How can a scientist making such an attempt know that phenotypic function has been correctly identified, or that an appropriate adaptive hypothesis has been arrived at, to begin with (cf. Gordon 1992)?

As an illustrative cautionary tale, the behavior of certain 'laterally basking' butterflies in orienting their wings perpendicular to sunlight was at first guessed to be an adaptation to minimize their casting of shadows, hence to avoid predators' attention. More careful consideration shows that it does not do so! Parallel orientation of the closed wings to the solar beam gives true minimization of shadow. It was instead shown experimentally, with appropriate testing against neutral null hypotheses, that the perpendicular orientation behavior is adaptive, but in relation to thermoregulatory absorption of sunlight (Watt 1968).

Some users of adaptationist approaches do recognize these concerns, and construct optimizing models for test only within the confines of possible constraints or other alternative explanations (e.g., Houston and McNamara 1999). Nonetheless, the intellectual hazards of assuming adaptiveness of phenotypes seem to many to outweigh the possible advantages. Certainly, studies of adaptive mechanisms in diverse organisms have been executed successfully, achieving results which are both rigorous and generalizable, without this assumption (e.g. Lauder 1996, Watt and Dean 2000).

6. *The Future Study of Adaptation*

In brief, it is clear that mechanistic approaches to the study of adaptation in the wild are increasing in diversity, rigor, and effectiveness. Application of biomechanical approaches to the function of morphological adaptations (Lauder 1996), or of molecular approaches to study of adaptation in metabolism and physiology (Watt and Dean 2000), allow specific results to be obtained with much precision. At the same time, philosophical ground-clearing may reduce misunderstanding of adaptation or misapplication of the concept, and lead to greater effectiveness of specific work as well as greater possibilities for general insight (Brandon 1990, Lloyd 1994, Watt 2000). There has often been a tension between the use of well-studied 'model' systems which can maximize experimental power and the fascination with diversity which drives the study of evolution for many workers. Both have value for the study of adaptation, and the tension may be eased by the interplay of comparative and phylogenetic studies (Larson and Losos 1996) with genetics-based experimental or manipulative study of organism–environment interactions and their demographic consequences in the wild. This synergism of diverse empirical and intellectual approaches holds great promise for the widening study of adaptation as a central feature of evolution by natural selection.

See also: Body, Evolution of; Brain, Evolution of; Darwin, Charles Robert (1809–82); Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolution: Optimization; Evolutionary Selection, Levels of: Group versus Individual; Genotype and Phenotype; Natural Selection; Optimization in Evolution, Limitations of

Bibliography

Bernstein H, Byerly H C, Hopf F A, Michod R E, Vemulapalli G K 1983 The Darwinian dynamic. *Quarterly Review of Biology* **58**: 185–207

Brandon R N 1990 *Adaptation and Environment*. Princeton University Press, Princeton, NJ

Cavalli-Sforza L L, Feldman M W 1981 *Cultural Transmission and Evolution: a Quantitative Approach*. Princeton University Press, Princeton, NJ

Charlesworth B 1994 *Evolution in Age-structured Populations*, 2nd edn. Cambridge University Press, Cambridge, UK

Cowen R 1995 *The History of Life*, 2nd edn. Blackwell Scientific Publications, Oxford, UK

Darwin C 1859 *The Origin of Species*, 6th rev. ed. 1872. New American Library, New York

Depew D J, Weber B H 1995 *Darwinism Evolving*. MIT Press, Cambridge, MA

Endler J A 1986 *Natural Selection in the Wild*. Princeton University Press, Princeton, NJ

Feder M E, Watt W B 1992 Functional biology of adaptation. In: Berry R J, Crawford T J, Hewitt G M (eds.) *Genes in Ecology*. Blackwell Scientific Publications, Oxford, UK, pp. 365–92

Gillespie J H 1991 *The Causes of Molecular Evolution*. Oxford University Press, Oxford, UK
Gordon D M 1992 Wittgenstein and ant-watching. *Biology and Philosophy* **7**: 13–25
Gould, S J 1980 The evolutionary biology of constraint. *Daedalus* **109**: 39–52
Gould S J 1989 A developmental constraint in *Cerion*, with comments on the definition and interpretation of constraint in evolution. *Evolution* **43**: 516–39
Gould S J, Lewontin R C 1979 The spandrels of San Marco and the Panglossian paradigm. *Proceedings of the Royal Society of London* **B205**: 581–98
Gould S J, Vrba E S 1982 Exaptation—a missing term in the science of form. *Paleobiology* **8**: 4–15
Grant P R 1986 *Ecology and Evolution of Darwin's Finches*. Princeton University Press, Princeton, NJ
Hickman C S 1988 Analysis of form and function in fossils. *American Zoologist* **28**: 775–93
Hodge M J S 1987 Natural selection as a causal, empirical, and probabilistic theory. In: Kruger L, Gigerenzer G, Morgan M S (eds.) *The Probabilistic Revolution*. MIT Press, Cambridge, MA, Vol. 2, pp. 233–70
Houston A I, McNamara J M 1999 *Models of Adaptive Behaviour*. Cambridge University Press, Cambridge, UK
Kimura M 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK
Kingsolver J G, Watt W B 1984 Mechanistic constraints and optimality models: thermoregulatory strategies in *Colias* butterflies. *Ecology* **65**: 1835–39
Laland K N, Odling-Smee F J, Feldman M W 1996 The evolutionary consequences of niche construction: an investigation using two-locus theory. *Journal of Evolutionary Biology* **9**: 293–316
Laland K N, Odling-Smee F J, Feldman M W 1999 Evolutionary consequences of niche construction and their implications for ecology. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 10242–47
Larson A, Losos J B 1996 Phylogenetic systematics of adaptation. In: Rose M R, Lauder G V (eds.) *Adaptation*. Academic Press, New York, pp. 187–220
Lauder G V 1996 The argument from design. In: Rose M R, Lauder G V (eds.) *Adaptation*. Academic Press, New York, pp. 55–91
Lenski R E, Travisano M 1994 Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 6808–14
Lewontin R C 1983 Gene, organism, and environment. In: Bendall D S (ed.) *Evolution from Molecules to Men*. Cambridge University Press, Cambridge, UK, pp. 273–85
Lewontin R C 1984 Adaptation. In: Sober E (ed.) *Conceptual Issues in Evolutionary Biology*. MIT Press, Cambridge, MA, pp. 235–51
Lewontin R C, Dunn L C 1960 The evolutionary dynamics of a polymorphism in the house mouse. *Genetics* **45**: 705–22
Lloyd E A 1994 *The Structure and Confirmation of Evolutionary Theory*, 2nd edn. Princeton University Press, Princeton, NJ
Mayr E 1988 *Toward a New Philosophy of Biology*. Harvard University Press, Cambridge, MA
McGraw J B, Caswell H 1996 Estimation of individual fitness from life-history data. *American Naturalist* **147**: 47–64
Michod R E 1999 *Darwinian Dynamics*. Princeton University Press, Princeton, NJ
Parker G A, Maynard Smith J 1990 Optimality theory in evolutionary biology. *Nature* **348**: 27–33
Reeve H K, Sherman P W 1993 Adaptation and the goals of evolutionary research. *Quarterly Review of Biology* **68**: 1–32
Romer A S 1955 *The Vertebrate Body*, 2nd edn. Saunders, Philadelphia, PA
Rose M R, Lauder G V 1996 Post-spandrel adaptationism. In: Rose M R, Lauder G V (eds.) *Adaptation*. Academic Press, New York, pp. 1–8
Rosenberg A 2000 Laws, history, and the nature of biological understanding. *Evolutionary Biology* **32**: 57–72
Roughgarden J 1979 *Theory of Population Genetics and Evolutionary Ecology: an Introduction*. Macmillan, New York
Vermeij G J 1987 *Evolution and Escalation*. Princeton University Press, Princeton, NJ
Watt W B 1968 Adaptive significance of pigment polymorphisms in *Colias* butterflies. I. Variation of melanin pigment in relation to thermoregulation. *Evolution* **22**: 437–58
Watt W B 1992 Eggs, enzymes, and evolution—natural genetic variants change insect fecundity. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 10608–12
Watt W B 1994 Allozymes in evolutionary genetics: self-imposed burden or extraordinary tool? *Genetics* **136**: 11–16
Watt W B 2000 Avoiding paradigm-based limits to knowledge of evolution. *Evolutionary Biology* **32**: 73–96
Watt W B, Dean A M 2000 Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. *Annual Review of Genetics* **34**: 593–622
West G B, Brown J H, Enquist B J 1997 A general model for the origin of allometric scaling laws in biology. *Science* **276**: 122–26

W. B. Watt

Adaptive Preferences: Philosophical Aspects

'Adaptive preferences' refer to a process of preference change, whereby people's preferences are altered, positively or negatively, by the set of feasible options among which they have to choose. In the negative case ('unreachable grapes are probably sour anyway'), people value options less highly *ex post* of realizing that they are not feasible anyway than they valued those options *ex ante* of that realization. In the positive case ('forbidden fruit is sweeter'), they value more highly things *ex post* of realizing that they are beyond their grasp than they did *ex ante* of that realization.

Adaptive preferences of either sort threaten to violate normative canons of rational choice and undercut welfare theorems built around them (Elster 1983, Chap. 3). People's getting what they want makes them unambiguously better off, just so long as those preferences constitute fixed, independent standards of assessment. Where people alter their preferences in response to whatever they get (or did not get or could or could not get), just because that is what got (or did

not get or could or could not get), satisfying people's *ex ante* preferences does not necessarily make them better off *post hoc*. While adaptive preferences do not alter people's choice behavior, they do alter their evaluation of their chosen option relative to other infeasible ones, in that way affecting people's subjective welfare. Adaptive preferences also skew people's behavior in investigating new possibilities, making them more or less prone to being manipulated by inculcating misperceptions of what is or is not within the feasible set.

1. *Preference, Choice, and Welfare*

In the standard model of rational choice, normative decision theory prescribes that agents first produce a complete and consistent ranking over all conceivable options, then map the 'feasible set' onto that ranking, and finally choose the highest-ranked alternative (or pick among equal-highest ranked alternatives) that fall within the feasible set. Thus, microeconomic representations of consumer choice start by sketching 'indifference curves' representing the agent's preferences, then superimpose a 'budget line' (or 'production possibility frontier') on that, and finally identify where the budget line intersects the highest indifference curve as the rational choice.

When people proceed in this way, they maximize their subjective welfare (defined, tautologically, in terms of reaching the highest preference plateau they can), given their budget constraints. When groups of such individuals interact in free, perfectly competitive markets, the exchanges that they make similarly maximize the collective welfare of all concerned (defined in terms of Pareto-optimality: no one can be made better off without someone being made worse off).

1.1 *Preferences as Fixed, Independent Standards*

For those welfare conclusions to emerge, however, it is crucial that people's preferences form a fixed, independent standard of assessment. Suppose that people's preferences were not fixed but instead fluctuated randomly and with great frequency, so much so that we could be virtually certain that their preferences would have changed by the time the goods they have chosen were actually delivered. In the case of consumers so fickle as that, we have no reason to think that respecting their original preferences and delivering to them the goods they have chosen will leave them (individually or collectively) better off than any other course of action.

Suppose instead that people's preferences were not independent of ('exogenous to') the system that is supposed to be satisfying them. Suppose, for example, that people were infinitely adaptable and agreeable, thinking (like Dr Pangloss) that whatever happens is for the best and whatever they are allocated is *ipso facto* what they most want. Or suppose that people were infinitely impressionable, thinking that they most want whatever producers' advertising tells them they want (Gintis 1972). Where preferences are shaped in such ways by the same processes that are supposed to satisfy them, we once again have no reason to think that respecting people's original preferences will leave them (individually or collectively) better off than any other course of action. Presumably such adaptive or impressionable consumers could and would adjust their preferences in such a way that they would like equally well anything else they were allocated (Sunstein 1993).

1.2 *Relaxing that Requirement*

To insist that people's preferences never change is asking too much. Clearly, people's preferences do change all the time, at least at the margins. Respecting their expressed preferences still seems the most likely way to maximize their welfare, individually and collectively, just so long as those preference changes are not too large or too frequent.

So too is it too much to insist that people's preferences never change in ways endogenous to the process that is supposed to be satisfying them. Satisfying one preference causes yet another to come to the fore. The more experience people have of a certain good, the more they tend to like (or dislike) it, in part simply because they have more information about it, and in part because they have more 'consumption capital' that interacts with the good to enhance people's enjoyment (or otherwise) of that good (Stigler and Becker 1977, Becker 1996). More generally, people's preferences are socially inculcated and culturally transmitted, with the same underlying processes generating a demand for certain cultural forms and individual traits while at the same time ensuring a supply of them (Bowles 1998). Radical economists are rightly suspicious of such processes. But, again, so long as the causal processes shaping preferences operate at sufficient distance from the processes satisfying them—and especially if preferences, once formed, tend to be relatively impervious to subsequent influence by those same forces (Lerner 1972)—then perhaps we might still suppose that respecting people's expressed preferences is the most likely way to maximize their welfare, individually and collectively.

2. *Adapting Preferences to Possibilities*

People's altering their preferences in response to their perceived possibilities similarly threatens to prevent preferences from functioning as fixed, independent standards of the sort which could reliably ground welfare judgments.

People who get what they want are better off in consequence; but people who want what they get, just because that is what they got, are not unambiguously better off. They would have been happy (perhaps equally happy, if we can talk in such cardinal-utility terms) with whatever they got.

By the same token, people who can get what they want are better off (better off, that is, than they would be if they could not get what they wanted). But people who want what they can get, just because that is what they can get, are not unambiguously better off. They would have been happy (perhaps equally happy) with whatever they could get. Conversely, people who do not want what they can get, just because that is what they can get, would have been unhappy (perhaps equally unhappy) with whatever they could get. Preferences that adapt in either of these ways, either positively or negatively to possibilities, thus seem to undercut the status of preferences as the sorts of fixed, independent standards which can reliably ground welfare judgments.

2.1 Intentional Adaptation

In general, adaptability is something to be desired. It helps us to be individually well-adjusted and evolutionarily fit as a species. Adapting our future choices in light of past experiences is the essence of learning. Adapting our choices to what we expect others to do is the essence of strategic rationality (see *Game Theory*).

On some accounts, adapting your preferences to your possibilities might be desirable in some of the same ways. Stoics, Buddhists, and others have long advised that the best way to maximize your happiness is to restrain your desires, confining them to what you already have or can easily get (Kolm 1979). Theorists of self-control sometimes describe that process in terms of a game of strategy, whereby one's 'higher self' adaptively responds to the anticipated reactions of the 'base self' (Elster 1979, Chap. 2, Schelling 1984). Athletic trainers and social reformers, in contrast, often advise us to set our aspirations just beyond what realistically we believe we can obtain.

Those are cases of preferences that are intentionally adaptive. There, the individuals concerned deliberately and self-consciously attempt to alter their own preferences in certain directions. Intentionally adaptive preferences are in that respect akin to other instances of deliberate, self-conscious preference formation: people's striving to overcome unwanted addictions, build their character, or cultivate their tastes. It is unobjectionable for people to try to shape or reshape their own preferences in these or any of various other ways.

Preferences which are adapted unintentionally to possibilities are potentially more problematic, precisely because they can claim no warrant in the agent's will. Individuals who find themselves unself-consciously adapting their preferences to their circumstances are being controlled by their environment rather than controlling it. They no longer fully qualify as 'sovereign artificers' choosing their own way in the world. They no longer qualify fully as external sources of value, independent assessors of the worth of alternative states of the world.

2.2 The Irrelevance of Adaptation to Choice

Suppose people judge the feasible set correctly. What they think is impossible really is impossible, and what they think is possible really is possible. Suppose furthermore that the feasible set is given exogenously, and the agents themselves can do nothing to alter its contents.

In that case, there is no reason to think that the adaptiveness of people's preferences to their possibilities does anything to alter their choices. If people's preferences are positively adaptive, they will prefer options more strongly if they are in their feasible set than they would have preferred those same options if they were not in their feasible set; and conversely if people's preferences are negatively adaptive. Adaptation of either sort changes the relative ranking of options in the feasible set to options outside the feasible set. But neither sort of adaptation changes the relative rankings of options all of which are within the feasible set.

Adaptive preferences, in effect, just introduce a constant inflator (in the case of positive adaptation: deflator, in the case of negative adaptation) which applies equally to all options in the feasible set. Since all options in the feasible set are marked up (or down) by the same multiplier, feasible options' rankings relative to one another remains unaltered. And since choice can only be among feasible options, which option is in fact chosen is unaltered by either form of adaptation of preferences to possibilities.

2.3 Adaptation and Subjective Welfare

Even if adaptive preferences do not cause people to do (i.e., choose) differently, they nonetheless cause people to feel differently about their choices. People who positively adapt preferences to possibilities will think themselves fortunate to have been able to choose from a good set of options. People who negatively adapt preferences to possibilities will think themselves unfortunate to have been forced to choose from a bad set of options. Each group thinks as it does, not because of anything to do with the content of the set of options, but merely because those were the options that were indeed available to them.

Expressed in terms smacking of cardinal utilities, we might put the point this way. Suppose people were asked to put a cash value on various sets of options,

without being told which among them is possible and which is not. Upon being told which were possible and which were not, positively adaptive people would increase the value they attribute to those sets which are feasible and they would decrease the value they attribute to those sets that are infeasible. Negatively adaptive people would do the opposite.

Expressed in terms of merely ordinal utility rankings, the same point might be put this way. Suppose people were asked to rank order various sets of options, without being told which among them is possible and which is not. Upon being told which were possible and which were not, those sets of options which are indeed feasible would rise in the rankings of positively adaptive people and those sets of options which are not feasible would fall in their ranking. The opposite would be true among negatively adaptive people.

Differential evaluations of possible and impossible options can never directly manifest themselves in revealed choices, since there is never any opportunity actually to choose between possible and impossible options. But those differential evaluations might manifest themselves behaviorally in more indirect ways. Positively adaptive preferences tend to make people generally more content with their world, negatively adaptive ones make people generally more discontent with it.

People who are discontent tend, in turn, to be unhappy in themselves and unforthcoming in cooperative endeavors, and content people conversely. Adaptive preferences, by contributing to those more general personal dispositions, can thus have an indirect effect on individual and collective welfare, even if they do nothing to alter people's actual choices.

3. *Adaptive Preferences with Nonfixed Possibilities*

The conclusion that adaptive preferences make no difference to people's actual choice depends on the assumptions that possibilities are known and that they cannot be altered. Where either of those assumptions fails to be met, adaptive preferences really can make a difference, not just to how people feel about their choices, but to how they actually choose.

3.1 *Altering the Feasible Set*

Suppose that there is something that people, individually or collectively, can do to alter the possibilities before them. Suppose that they can invest in research and development into some new technology, for example.

It is assumed conventionally that rational choosers ought always prefer expanding their feasible set. There is substantial variability over time in the information upon which an individual's choices are based and the circumstances in which they are made—as well as, of course, variability over time in the individual's preferences themselves. Owing to variability in all those dimensions, the same individual might rationally choose different options at different times and the availability of different options is itself valuable in consequence (Arrow and Fisher 1974).

Those arguments for valuing the expansion of the feasible set are made independently of any consideration of how preferences might vary with past choices or future possibilities. Suppose, now, that people's preferences are strongly and positively endogenous, with previous experience leading us to seek yet more experiences of the same sort in future. That removes at least one of the reasons for valuing a range of options wider than merely continuing along the same path. (Other reasons may however remain: varying circumstances or information may mean that, in future, we will need to pursue some different path to secure the same sort of experience.)

Preferences that adapt to possibilities complicate the story still further. People with positively adaptive preferences tend, by definition, to be relatively more satisfied with their existing set of options than they would be were their preferences nonadaptive. That fact would make them relatively less anxious to seek out some new options than they would if their preferences were nonadaptive. People with negatively adaptive preferences represent the converse case: being relatively more dissatisfied with their existing options, they would be inclined to invest relatively more heavily in the search for new options (even if they would also tend to downgrade those new options, in turn, immediately upon their being discovered and added to the feasible set).

But of course the possibility of discovering new possibilities is itself one of the many possibility before people. People with positively adaptive preferences are inclined to mark up the value of the possibility of discovering new possibilities, just because it is possible. Those with negatively adaptive preferences are inclined to mark down its value for the same reason.

That latter set of considerations tends to push people in the opposite direction from the first. People with positively adaptive preferences would value new possibilities less (because they are not presently possible), but they would value the possibility of discovering new possibilities more (because that possibility is itself presently possible); and people with negatively adaptive preferences conversely. The joint effect of those two opposing tendencies might be to leave people of both inclinations roughly 'adaptively neutral' with respect to the search for new options. Alternatively, people might simply learn to differentiate between their appreciation of having and of using possibilities to discover new possibilities.

Less subtly, and more straightforwardly, people with positively adaptive preferences might adopt the

simple rule of thumb that, 'Possibilities are good, and more possibilities are better.' They would be led by that rule to seek out new possibilities, not because there is anything wrong with their present possibilities and not because presently impossible options hold any particular allure, but merely because possibilities themselves are what is to be maximized. What people with negatively adaptive preferences want is not the converse (to minimize possibilities): instead, what they want is to make possible the presently impossible (even knowing that they will downgrade the value of those options immediately upon their becoming possible). In practice, that might amount to much the same, a rule of maximizing possibilities being broadly desirable from either positively or negative adaptive perspectives.

The differences between positively and negatively adaptive preferences is more clear cut when it comes to restricting rather than expanding the feasible set. People whose preferences are positively adaptive value relatively highly their existing options and would be reluctant to see any reduction in them. People whose preferences are negatively adaptive attach relatively more value to options that they do not have; and they would be relatively more indifferent to reductions in their existing options, which they value less highly (in the paradoxical limiting case, watching with indifference as their feasible set is extinguished altogether).

3.2 Uncertainty Concerning the Feasible Set

Suppose, next, that people do not know with complete confidence what is within the feasible set. There are some things that are certainly inside that set, and some other things that are certainly outside it. But there are various other things that may be inside or out.

People with positively adaptive preferences will mark up the value of things that might be possible, compared to that which they know with confidence to be impossible. They do not mark up the value of 'maybe possible' options as much as they mark up the value of options that they know with confidence to be possible, to be sure. But the sheer fact that those options are somewhere in the penumbra of the feasible set makes them relatively more attractive to people with positively adaptive preferences than they otherwise would have been. People with negatively adaptive preferences will display the converse pattern, marking down the value of things in the penumbra of the feasible set.

Here again, the fact that it is possible that something is possible is itself a possibility, and people with positively adaptive preferences should respond positively (and those with negative adaptives ones negatively) to that possibility as any other. But there is surely something mad about applying a double inflator (or deflator) to merely possible possibilities. If it is good that something is possible, then what is good is that it is possible tout court. It is not doubly good that it merely possibly possible. It is not so much the possibility as such but the optionality—the eligibility for choice—that positively adpative preferences value (and negatively adaptive ones disvalue).

4. Manipulating Perceptions of the Feasible Set

Advertisers and other 'hidden persuaders' famously attempt to manipulate people's choices by shaping their perceptions of the relative desirability of various options before them. Shaping people's preferences is one fairly direct way to shape their choices. Much the same effect can also be produced indirectly, by shaping their perceptions of the feasible set.

Perceptions of what is possible, jointly with our preferences, determine our choices. That which is impossible is rightly regarded as beyond the bounds of rational choice. But our information about what is or is not possible for them to do is rarely perfect, and shaping people's perceptions of the possibilities and impossibilities facing them is one effective way of manipulating their choices (Goodin 1982, Chap. 7).

That trick works to shape the choices of rational choosers, quite generally, since all rational agents choose merely from among the options they perceive to be open to them. Some people, however, adapt not just their choices to their possibilities but also their preferences to their possibilities; and that makes them more (in the case of positively adpative preferences) or less (in the case of negatively adaptive ones) easily prey to that trick.

People whose preferences are positively adaptive inflate the value of options perceived to be in their feasible set, relative to ones that are not. If they are persuaded that something is not possible anyway, then by virtue of that very fact the value of that option falls in their estimation. Because of that, in turn, they will suffer less regret at not being able to pursue that option, they will be less inclined to search for ways to make that option possible after all, and so on. And because of that, there is less risk of them discovering that their perception of that option as being impossible is in fact in error.

People with negatively adaptive preferences constitute the converse case, valuing particularly highly options they perceive as impossible. Regretting and resenting their impossibility as they do, such people are more likely to seek ways of rendering those options possible. That makes it more likely for them to discover that their perception of the option's impossibility is in error, thus exposing the manipulative fraud.

See also: Decision Biases, Cognitive Psychology of; Heuristics for Decision and Choice; Risk: Theories of Decision and Choice; Utility and Subjective Pro-

bability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies; Well-being (Subjective), Psychology of

Bibliography

Arrow K J, Fisher A C 1974 Environmental preservation, uncertainty and irreversibility. *Quarterly Journal of Economics* **88**: 312–9

Becker G S 1996 *Accounting for Tastes*. Harvard University Press, Cambridge, MA

Bowles S 1998 Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature* **36**: 75–111

Elster J 1979 *Ulysses and the Sirens*. Cambridge University Press, Cambridge, UK

Elster J 1983 *Sour Grapes*. Cambridge University Press, Cambridge, UK

Gintis H 1972 A radical analysis of welfare economics and individual development. *Quarterly Journal of Economics* **68**: 572–99

Goodin R E 1982 *Political Theory and Public Policy*. University of Chicago Press, Chicago

Kolm S-C 1979 La philosophie bouddhiste et les 'hommes économiques.' *Social Science Information* **18**: 489–588

Lerner A P 1972 The economics and politics of consumer sovereignty. *American Economic Review* (*Papers and Proceedings*) **62**: 258–66

Schelling T C 1984 *Choice and Consequences*. Harvard University Press, Cambridge, MA

Stigler G J, Becker G S 1977 De gustibus non est disputandum. *American Economic Review* **67**: 76–90

Sunstein C R 1993 Endogenous preferences, environmental law. *Journal of Legal Studies* **22**: 217–54

R. E. Goodin

Additive Factor Models

Sternberg's (1969) Additive Factor Method is one of the major ways of analyzing response times. The goal is to learn about mental processes and how they are organized. To use it, the experimenter manipulates experimental variables called factors (e.g., brightness, discriminability), while a person performs a task (e.g., naming a digit). The person executes processes such as perceiving and deciding (processes are actions, not processors). Assume the processes are executed one after the other, in series, each process stopping before its successor starts. The time to complete the task, the response time, is the sum of the durations of the individual processes. A factor selectively influences a process if changing the factor changes the duration of that process, leaving durations of the other processes unchanged. If the combined effect on mean response time of changing two factors together is the sum of the effects of changing them separately, the factors are called additive factors. In early applications, experimental results were interpreted as follows. (a) If two factors are additive, each factor selectively influences a different process. (b) If two factors are not additive, at least one process is influenced by both factors. This entry discusses the validity and current use of the method for response times, extensions to other measures such as accuracy and evoked potentials, and extensions to operations other than addition.

A common strategy in science is to isolate components by taking an object apart. Obviously, processing in the human brain cannot be studied this way, so methods are needed for analyzing the intact system. The main method of this type for response times is the Additive Factor Method.

1. Response Time and Serial Processes

Shwartz et al. (1977) provide an illustrative example. An arrow pointing rightward or leftward was presented. Response was with a button on the right or left. The experimenter manipulated three factors. First was intensity: in some trials the arrow was bright, in others, dim. Second was similarity: the arrow pointed distinctly rightward or leftward, or indistinctly. Third was compatibility: for some participants, the arrow pointed toward the correct response button, for others, away. In an analysis of variance, the factors had additive effects on mean response time (RT). The authors concluded that the mental processes required for the task were executed in series, and that each factor selectively influenced a different process.

1.1 Selective Influence

Intuitively, a factor selectively influences a process when changing the level of the factor changes only that process, leaving other processes unchanged. It is implicitly assumed that changing the level of the factor also leaves the arrangement of the processes unchanged.

For processes in series, the mean RT is the sum of the means of the individual process durations (whether or not processes communicate or their durations are stochastically independent). A factor selectively influencing a process may change its mean duration. It leaves the marginal distributions of other processes unchanged, and hence leaves their means unchanged. Therefore, if two factors selectively influence two different processes, the change in RT they produce when combined is the sum of the changes they produce individually.

Measures other than mean RT require stronger assumptions. A common assumption is that process durations are stochastically independent, that is, the joint distribution of the process durations is the product of their marginal distributions. Then a factor

selectively influences a process if changing the level of the factor changes the marginal distribution of the process, does not change the marginal distribution of any other process, and leaves process durations independent.

1.2 Response Time Cumulative Distribution Functions

The mean gives only part of the response time information. The summation test (Ashby and Townsend 1980, Roberts and Sternberg 1992) examines distribution functions. Consider a task requiring process a followed by process b. Suppose when the level of a factor changes from 1 to 2, the duration of process a changes from random variable A_1 to random variable A_2. Likewise, suppose when the level of another factor changes from 1 to 2, the duration of process b changes from random variable B_1 to random variable B_2. When the first factor is at level i and the second factor is at level j, denote the condition as (i, j) and the response time as T_{ij}. Then,

$$T_{ij} = A_i + B_j; \quad i, j = 1, 2$$

so

$$T_{11} + T_{22} = A_1 + B_1 + A_2 + B_2 = T_{12} + T_{21} \qquad (1)$$

Ashby and Townsend (1980), assuming stochastic independence, proved that the distributions of $(T_{11} + T_{22})$ and $(T_{12} + T_{21})$ are the same.

Roberts and Sternberg (1992) developed an innovative test of this. For a given participant, every response time in condition (1, 1) is paired with every response time in condition (2, 2). (That is, the Cartesian product of the set of response times in condition (1, 1) and the set of response times in condition (2, 2) is formed.) Every such pair of response times is added. The empirical cumulative distribution of these sums is an estimate of the cumulative distribution of the composite random variable $(T_{11} + T_{22})$. Similarly, the cumulative distribution of the composite random variable $(T_{12} + T_{21})$ is estimated. These two estimates are predicted to be close, and were found to be remarkably close in a number of experiments analyzed by Roberts and Sternberg (1992).

1.3 Counterexamples

It might seem at first that if two factors have additive effects on RT, then processing occurs in two disjoint time intervals, each factor changing the duration of a different time interval. The conclusion does not follow. Counterexamples include a dynamic system model of Townsend and Ashby (1983) and (approximately) McClelland's (1979) cascade model.

Despite the counterexamples, there are conditions under which additive effects of factors on response time imply the existence of random variables whose sum is the response time, and such that changing the level of one of the factors changes only one of the random variables in the sum (see below). It is natural to say the random variables are the durations of processes in series, but the mechanism producing the random variables is not implied.

With inductive reasoning, one can say an empirical finding of additive factors supports the statement that the factors selectively influence processes in series. For strong support, evidence for additivity in many circumstances is needed. Additivity must occur at all levels of the factors said to be additive, and at all levels of other factors also. The statistical power must be high.

If factors are not additive, it is tempting to conclude that they do not selectively influence different processes. If (a) processes are serial and (b) each factor selectively influences a different process, then the factors will indeed have additive effects on mean RT. But an interaction may indicate that the processes are not serial. If the task requires completion of parallel processes, the RT is the maximum of the process durations, not the sum (Sternberg 1969, Townsend and Ashby 1983). Factors selectively influencing parallel processes will have interactive effects on RT (see *Network Models of Tasks*).

2. Other Measures and Process Arrangements

Using experimental factors selectively to influence mental processes was so successful for RT (e.g., Sanders 1980) that it was extended to other dependent measures. Three will be discussed: accuracy, evoked potentials, and rate.

2.1 Accuracy: Log Probability Correct

Call a process correct if its output is right for its given input. Suppose the probability of a correct response for the task is the product of the probabilities that each individual process is correct. This assumption is plausible for serial processes, and for other arrangements also. (Stochastic independence is stronger, requiring multiplicative rules for all outcomes, correct and incorrect.) Then the log of the probability of a correct response is the sum of the logs of the probabilities that the individual processes are correct. Hence, factors selectively influencing different processes will have additive effects on log percent correct (Schweickert 1985).

If the probability P of a correct response is large, a test can be based on the fact that for large P, $\log_e P$ is approximately equal to $1 - P$, the error probability. Then additivity for log probability correct implies

approximate additivity for error probability. Schwartz et al. (1977), in the arrow identification experiment described above, found additive effects of the three factors on error probability.

Another test, not requiring large P, can be based on a log linear model. Predictions of this log linear model agree well with the observations of Schwartz et al. (1977), see Schweickert (1985).

2.2 Accuracy: Probability Correct and Multinomial Processing Trees

Serial processes are not the only possibility, as Townsend (1972) emphasized. Sometimes a correct response can be made in one of two mutually exclusive ways. For example, in an immediate serial recall experiment, after a list of words is presented, suppose each word can be correctly recalled via a speech-like representation (articulatory loop) or via a semantic representation, but not both. The probability of correctly recalling a word is the probability of correct recall via the articulatory loop plus the probability of correct recall via the semantic representation. Then two factors selectively influencing the two ways of producing a correct response will have additive effects on probability correct.

In a relevant experiment by Poirier and Saint-Aubin (1995), participants sometimes repeated an irrelevant word aloud during list presentation. Suppose then the articulatory loop is not as likely to lead to a correct response. Sometimes the words in a list were from the same semantic category, sometimes from different categories. Suppose with different categories the semantic representation is not as likely to lead to a correct response. The factors of repeating aloud and semantic similarity had additive effects, supporting the interpretation of mutually exclusive ways to respond correctly.

Recall of a word can be represented as a multinomial processing tree. Processing starts at the root node. Each branch represents the outcome of a process, correct or incorrect. Terminal nodes represent responses, correct or incorrect. Additivity is predicted by a tree with one path leading from the root to a correct response via the speech-like representation and another path via the semantic representation. If, instead, a path leads to a correct response using both representations, the result would be an interaction (Schweickert 1993).

2.3 Evoked Potentials and Parallel Processes: The Additive Amplitude Method

During mental processing neurons change electric fields and the changes in potential (voltage) can be measured at points on the scalp. The potential measured at any point in space is the sum of the potentials reaching that point from all sources, the basis of what Kounios calls the Additive Amplitude Method. Consider two mental processes executing simultaneously. If each of two factors selectively influences a different process, their combined effect on potential will be the sum of their individual effects, at every point at which potential is measured, throughout the duration of the processing.

Kounios and Holcomb (1992) presented sentences such as 'NO DOGS ARE FURNITURE.' Participants responded with the truth value. Sometimes the subject and predicate were related, sometimes not. Sometimes the subject was an exemplar and the predicate a category, sometimes the reverse. The two factors had additive effects on potential at each electrode site, throughout the interval from 300 to 500 ms after the predicate was presented. A brief interpretation is that the potential in this interval reflects two different parallel processes having synchronized neural firing, one for semantics and one for hierarchy.

2.4 Rates and Timers: The Multiplicative Factors Method

Roberts (1987) considered sequences of responses, such as repeated lever presses, controlled by an internal timer which emits pulses at the rate r per second. The pulses are sent to a filter permitting a fraction f of the pulses to be sent to another filter, which permits a fraction g of the pulses to be sent to the responder. Responses are made at the rate rfg. Suppose one factor changes the fraction of pulses sent on by the first filter, and another factor changes the fraction of pulses sent on by the second filter. The factors will have multiplicative effects on rate.

Roberts (1987) gives example data from Clark (1958). Rats pressed a lever for food. Different groups received different variable interval reward schedules, and rats were tested at different times after feeding. Reward schedules and testing times had multiplicative effects on the rate of lever pressing, as predicted.

3. Generalization: Selective Influence and Other Combination Rules

When processes are not independent, what does it mean to selectively influence one? This question was considered first by Townsend and Ashby (1983), and most recently by Dzhafarov. The work considerably extends the scope of selective influence, and quite general combination rules can be tested.

When factor α is at level i and factor β is at level j, let A_{ij} be the duration of process a and B_{ij} be the duration of process b. Let $A \cong$ mean A has the same distribution as $\cong$. Factors α and β selectively influence the durations of processes a and b, respectively, if there are independent random variables P_1 and P_2, and functions G and H such that

$$A_{ij} \cong G(P_1, P_2; i) \quad \text{and} \quad B_{ij} \cong H(P_1, P_2; j) \qquad (2)$$

and some additional technical conditions are met (Dzhafarov in press). As A_{ij} depends only on i, and B_{ij} depends only on j, denote them as A_i and B_j, respectively. Because A_i and B_j are functions of the same random variables, P_1 and P_2, they may be stochastically dependent. One form of dependence, relevant here, is perfect positive stochastic interdependence. Suppose there is a single random variable P, uniformly distributed between 0 and 1, such that

$$A_i \cong G(P; i) \quad \text{and} \quad B_j \cong H(P; j)$$

Then A_i and B_j are said to be perfectly positively stochastically interdependent, written $A_i \parallel B_j$. (A random variable's distribution can always be written as the distribution of its quantile function applied to a uniformly distributed random variable. Required here is that for A_i and B_j, it is the same uniform random variable.) Expressions such as $T_{11} \parallel T_{22}$ are defined analogously.

Now consider any binary operation ◆ which is associative, commutative, strictly monotonic, and continuous in both arguments, or is maximum or minimum. Let T_{ij} be the random variable observed when factor α is at level i and factor β is at level j. Then there exist random variables A_1, A_2, B_1, and B_2 such that

$$T_{11} \cong A_1 \blacklozenge B_1 (A_1 \parallel B_1)$$
$$T_{12} \cong A_1 \blacklozenge B_2 (A_1 \parallel B_2)$$
$$T_{21} \cong A_2 \blacklozenge B_1 (A_2 \parallel B_1)$$
$$T_{22} \cong A_2 \blacklozenge B_2 (A_2 \parallel B_2)$$

if and only if

$$T_{11} \blacklozenge T_{22} \cong T_{12} \blacklozenge T_{21} (T_{11} \parallel T_{22} \quad \text{and} \quad T_{12} \parallel T_{21})$$

For gist, consider Equation (1) with ◆ substituted for $+$. For details and uniqueness, see Dzhafarov and Schweickert (1995).

The thrust of theoretical work is toward statements of the form above, that a model with certain properties accounts for the data if and only if the data satisfy certain conditions. Usually, the model is not unique, that is, under the same conditions, other models with radically different properties may also account for the data (e.g., Townsend 1972). One can state only that the brain produces data indistinguishable from the predictions of a model with certain properties. One cannot validly conclude that brain and model operate the same way. Methods based on selective influence cannot overcome this limitation. What they can provide is the conclusion that the brain behaves as if made of separately modifiable components (Sternberg 1998). They also provide explicit relations between changes in the experiment, the model, and the data. Without these relations, an analysis of the entire system would be uninterpretable.

See also: Information Processing Architectures: Fundamental Issues; Measurement Theory: Conjoint Measurement Theory

Bibliography

Ashby F G, Townsend J T 1980 Decomposing the reaction time distribution: Pure insertion and selective influence revisited. *Journal of Mathematical Psychology* **21**: 93–123

Clark F C 1958 The effect of deprivation and frequency of reinforcement on variable-interval responding. *Journal of Experimental Analysis of Behavior* **1**: 221–8

Dzhafarov E N 2001 Unconditionally selective dependence of random variables on external factors. *Journal of Mathematical Psychology* **45**: 421–51

Dzhafarov E N, Schweickert R 1995 Decomposition of response times: An almost general theory. *Journal of Mathematical Psychology* **39**: 285–314

Kounios J 1996 On the continuity of thought and the representation of knowledge: Electrophysiological and behavioral time-course measures reveal levels of structure in semantic memory. *Psychonomic Bulletin & Review* **3**: 265–86

Kounios J, Holcomb P 1992 Structure and process in semantic memory: Evidence from brain related potentials and reaction time. *Journal of Experimental Psychology—General* **121**: 459–79

McClelland J L 1979 On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review* **86**: 287–330

Poirier M, Saint-Aubin J 1995 Memory for related and unrelated words: Further evidence on the influence of semantic factors in immediate serial recall. *Quarterly Journal of Experimental Psychology, Series A* **48**: 384–404

Roberts S 1987 Evidence for distinct serial processes in animals: The multiplicative-factor method. *Animal Learning & Behavior* **15**: 135–73

Roberts S, Sternberg S 1992 The meaning of additive reaction-time effects: Tests of three alternatives. In: Meyer D, Kornblum S (eds.) *Attention and Performance XIV*. MIT Press, Cambridge, MA, pp. 611–54

Sanders A F 1980 Stage analysis of reaction processes. In: Stelmach G E, Requin J (eds.) *Tutorials in Motor Behavior*. North Holland, Amsterdam

Schweickert R 1985 Separable effects of factors on speed and accuracy: Memory scanning, lexical decision and choice tasks. *Psychological Bulletin* **97**: 530–46

Schweickert R 1993 A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition* **21**: 168–75

Shwartz S P, Pomerantz J R, Egeth H 1977 State and process limitations in information processing: An additive factors analysis. *Journal of Experimental Psychology—Human Perception and Performance* **3**: 402–10

Sternberg S 1969 The discovery of processing stages: Extensions of Donders' method. In: Koster W G (ed.) *Attention and Performance II, Acta Psychologica* **30**: 276–315

Sternberg S 1998 Discovering mental processing stages: The method of additive factors. In: Scarborough D, Sternberg S (eds.) *Methods, Models, and Conceptual Issues: An Invitation to Cognitive Science*, 2nd edn. MIT Press, Cambridge, MA, pp. 703–864

Townsend J T 1972 Some results concerning the identifiability of parallel and serial processes. *British Journal of Mathematical & Statistical Psychology* **25**: 168–99

Townsend J T, Ashby F G 1983 *Stochastic Modeling of Elementary Psychological Processes*. Cambridge University Press, Cambridge, UK

R. Schweickert

Administration in Organizations

Administration in organizations—sometimes referred to as administrative science—is a mid-twentieth century construct. Social sciences consider management issues and practices as a specific field of inquiry. It deals with action and action taking in social units which follows some particular purpose: public agencies, firms, and voluntary associations (see *Bureaucracy and Bureaucratization*). How far is it possible within such rational entities to mobilize resources and people so as to achieve some degree of compatibility and some level of efficiency between differentiated tasks and between heterogenous logics of action? The challenge is to offer a set of theories and informations that explain and even predict behaviors and outcomes. Two main approaches of management emerge: the organization as an arena for strategic behavior and the organization as a moral community.

1. *From Principles to Concepts*

Modern management and organization thinking is rooted in the industrial revolution of the 1700s. How to organize and control complex economic and technical ventures such as factories has led the professions of mechanical engineering, industrial engineering, and economics to formulate prescriptions. What is often called the classical theory was dominant well into the 1940s. Its basic assumptions are that organizations exist to accomplish economic goals, that they act in accordance with rational criteria of choice, and that there exists one best way to solve a problem. Some of its leading figures are well known, such as Taylor (1911) an American practicing manager or Fayol (1949), a French engineer. Such a classical school claimed that administration was a matter of science. Action guidelines can be derived from universally applicable principles, whatever the type of organization is. Models and procedures are provided such as centralization of equipment and labor in factories, specialization of tasks, unity of command, and financial incentives based upon individual productivity. While Fayol was handling the issues of how to manage a firm as a whole, Taylor was defining expertise about how to get the individual worker organized. Optimism prevailed: managers have to learn a set of principles, to get them translated into procedures by experts, and, with the additional help of control and discipline, employees' behaviors will conform.

A strong attack was launched after World War II challenging such oversimplistic mechanistic views of administration. The rebellion against the classical school was led by organizational theorists trained in sociology and in political science.

Simon (1946) emerges as a pioneer and perhaps as its best known figure. In his opinion, the principles as defined by Taylor, Fayol, and others are instead mere proverbs: they are neither true nor false. He criticizes explicitly and rather violently the relevance of the principles approach. Specialization of the tasks, span of control, unity of command led to impasses, according to Simon. They are conflicting and inconsistent with most situations administrations face. With equal logic they are applicable in diametrically opposed ways to the same sets of circumstances. Therefore, in order to become a really scientific theory, administration in organizations has to substitute concepts to principles and make them operational.

In a subsequent book, Simon (1947) lays the ground for administration as a specific field of inquiry. He sketches a conceptual framework, which meaning corresponds to facts or situations empirically observable. He questions, for instance, the relevance of the principle of rationality. In organizations, even if they are purposive, an individual does not have the intellectual capacity to maximize, and they are also vulnerable to the surrounding social and emotional context. What human beings do is to satisfice: they try to find trade-offs between preferences and processes, they do the best they can where they are. Human and organizational decisions are subject to bounded rationalities. Simon also shows that efficiency is not a goal shared the same way by everyone in the organization, including the managers, and which can be defined *ex ante*. It should be a research question, starting from the hypothesis that the individuals or the organizations themselves carry a specific definition of what is good or correct from an efficiency point of view. In more general terms, contexts vary, and they make a difference.

Simon follows Max Weber's perspectives: administration belongs to the domain of rational action. Firms or public agencies are organizations driven by purposes. But managers rely upon the mediation of an organized setting in order to implement goals, purposes, or values. Therefore, the organization simultaneously provides a resource and becomes a constraint, managers experience it as a solution as well as a problem. Simon underlines the necessity for social sciences to approach management as a field aimed at understanding the nature of empirical phenomena. Its primary goal is not to formulate solutions for action but to consider action as a problem under scrutiny.

Practicing managers could nevertheless rely upon relevant findings and apply such a body of knowledge—or part of it—to enlighten problem solving. Such an agenda is structured around the study of the actual functioning of organizations. In a more specific way, Simon defines decision-making processes—or action—as the center of the scientific discipline of management. Any decision or action can be studied as a conclusion derived by the organization or by an individual from a set of premises. Some premises are factually grounded: they link a cause to an effect. Therefore, they are subject to a test by experience. Other premises are of a different nature: they are value grounded, made out of norms or ethical references. In this case they are not checkable empirically. While both categories are not separable in action, analysts have to separate them and focus upon factual premises only.

Firms and public agencies should also be treated as open organizations. They do not and cannot exist as self-contained islands within society and the market. They are linked to specific environments. The relationships which are structured between the inside and the outside play a very important function. Where and how an organization is embedded, what is exchanged, are phenomena which have an impact on the inner functioning as well as on the environment. A major theoretical breakthrough was offered by a sociologist, Philip Selznick (1949). The concept of co-optation which he elaborates describes how an organization gains support for its programs within the local communities where its execution agencies operate. An empirical study is offered by Selznick about an American federal agency, Tennessey Valley Authority. Co-optation refers specifically to a social process by which an organization brings outside groups and their leaders into its policy-making process, enabling such elements to become allies, not a threat to its existence and its mission. Bringing the environment back in solves a major difficulty the classical approach would not consider, especially when dealing with public administrations. Two of his founders, Woodrow Wilson and Frank J. Goodrow, had been calling for a theory of management which should make a dichotomy between politics and administration, between the elaboration of the policy of the state and the execution of that will. Selznick suggests that such a postulate should become a research question. He also proposes that, beside organizational phenomena as such, science should consider institutionalization dynamics, which means how values and norms are diffused, appropriated, and what impacts they have on managerial action-taking.

2. *Managing Arenas for Strategic Behavior*

Simon's agenda was recognized only in the 1960s as a milestone. It paved the way for what could be called the behavioral revolution in the field of administration. In the early part of the twenty-first century, it is still one of the most influential schools in business education and public management.

During the 1950s progress was made basically around the Carnegie Institute of Technology and under the leadership of Simon himself. With March he reviewed the studies of bureaucracies developed by social scientists such as Robert K. Merton, Philip Selznick, and Alvin W. Gouldner. Various models of bureaucratic behavior were formalized and compared (March et al. 1958). In highly proceduralized organizations, whether private or public, individuals and groups do not remain passive: they reinterpret rules and procedures, they play with and around them, they use them for secondary purposes of their own such as increasing their autonomy inside the hierarchical line of authority or as bargaining their participation to the organization. At an organizational level, management by rules favors or induces dysfunctional processes. Managers relying upon formalization and procedure are trapped in vicious circles. In order to fight unintended consequences of such tools for action, they reinforce formal rules.

The Carnegie School also addressed and criticized the theory of the firms as defined by orthodox microeconomics. Organizational decision making is the focal point. Is utility maximization the main function that business firms do, in fact, achieve? Cyert and March (1963) studied how coalitions are structured and activated inside a company around action taking and choice processes. Negotiations occur through which coalitions impose their demands on the organizational objective. Simon's conception is demonstrated as being applicable to economic actors: satisficing is a much more powerful concept to explain their strategic decisions than maximizing economic profit. Such is specifically the case of pricing in an oligopolistic market. In other terms, some characteristics of organizational structure determine rational behavior.

The implications of such a perspective are essential. From a knowledge point of view, conflict is a basic attribute of any organization. Business firms and public agencies as well are not monolithic entities sharing one common purpose. They behave as pluralistic systems in which differentiated and even antagonistic interests float around, conflict, or cooperate. They look like political coalitions between subgroups (March 1962). From an action-taking or managerial perspective, organizations require their leaders to develop skills that are less analytical than behavioral. Administrations are close to political brokers, negotiating and bargaining inside their organizations being a crucial task to fulfill. A firm looks like an arena for strategic microbehaviors, a collection of subunits pursuing separate goals. The role of management is to structure inducements so that each individual subunit identifies its interests with

those of the firm and, thereby, contributes to its mission.

In the 1960s, the behavioral approach has widened internationally and given birth to a stream of organizational researches about decision-making, power, and efficiency.

Allison (1971) studies the same event—the USA presidential handling of the 1962 Cuban missile crisis—comparing three different paradigms about decision making. An organizational process model, which is clearly derived from the Carnegie School approach, completed with a so-called governmental political model, which deals with partisan politics and presidential tactics on the public-opinion scene, shows a superior ability than a rational actor or classical model to explain how John F. Kennedy addressed the challenge and which outcomes were elaborated, despite the manifest of theory game based techniques used by the executive. Lindblom (1959) too takes a hard look at the rational model of choice. He rejects the notion that most decisions are made by total information processes and suggests that synoptic approaches provide self-defeating strategies for action. Instead, he sees the whole policy-making processes as being dependent upon small instrumental decisions that tend to be made in a disjointed order or sequence in response to short-term political conditions. Such a muddling through view prescribes managers to make small changes at a time and at the margin, not focusing too much and explicitly about content, whenever it is possible and, if needed, making some minor concession (two steps forward and one step backward). March and Olsen (1976) develop a garbage-can model of choice. Choices are characterized by ambiguity about goals, intentions, technologies, causation, participation, and relevance. What is a problem for actor A looks more like a solution for actor B, formal opportunities for choice look after problems to handle, decisions are made without being considered by the participants as being made. Such anarchic contexts occur in specific organizational settings such as bureaucracies and very loosely formalized structures. Nobody is really in control of the process and decisions are experienced as random-based outcomes. The implications of such a model for top managers are that they should not use quantitative tools as instruments of government or intervene in tactical ways but keep their hands free for what they consider as fundamental issues and use as action tools two basic vehicles: the selection of their immediate subordinates and a redesign of the formal structures of their organization.

Power phenomena are viewed as key variables for understanding and managing. Crozier (1963) offers a perspective that helps inter-relating micro-processes—such as the behavior of single actors—and macroprocesses—such as the functioning of the whole organization. Individuals and groups are pursuing rational strategies: they try to fulfill goals that are structured by the specific and local context within which they act daily. Asymetric interdependance relationship link them together: some are more dependant than others. Those who control a source of uncertainty from which others depend control power bases and are able, in exchange of their good will, to set up the rules of the game. In other terms, organizational functioning and change derives from the social-regulation processes as induced by the actors who, at various levels of the pyramid, try to make their specific and heterogenous strategies or logics of action compatible. From a managerial point of view, such a comprehensive framework implies that management is about the art and skills to reallocate uncertainties and power inside the organization, therefore, to structure interests inducing the actors to cooperate or not. Bower (1970) applies such a perspective to stretegic investment planning in a giant corporation. Allocating capital resources is a process that requires management to identify the various organizational components such as routines, parochialism, attention to issues, and discretionary behaviors of action controlling major uncertainties.

A third major critical examination of the classical school made by organizational theorists deals with rationality and efficiency. Landau (1969) argues that redundancy within a firm or a public agency is not a liability—or a symptom of waste and inefficiency—but a fundamental mechanism of reliability. Duplication and overlap provide solutions for action-taking in general. The breakdown of one part does not penalize the whole system. The arrogance of a subsystem controlling a monopoly on a problem or a function is diminished. Duplication and overlap may create political conflicts; they also generate conditions for communication, exchange, and co-operation. They lower risks. Organizations are not self-evaluating entities. They tend to substitute their own knowledge to the information generated by their environment. Economic efficiency and optimality as defined by economists are normative enterprises. In reality, management is much more related to failure avoidance and to fault or error analysis in a world where total control of events remains an impossible task to fulfill.

3. Moral Community Building as an Alternative Approach

While the behavioral revolution successfully challenges basic postulates upon which the classic theories of organization (a set of principles) and economics (optimality) are grounded, it nevertheless assumes that management and administration handle a firm or a public agency through an economy of incentives. Incentives are the rewards and sanctions imposed by leaders, and they generate behaviours. Well-designed incentives—whether financial or organizational—align individual goals and collectively produce man-

agerially desired action. Designed poorly, incentives may produce subunit conflict and poor firm performance. Implicit in this view of the organization is the assumption that organizational actors, either persons or subunits, possess preferences and influence resources which include position or office, functional or professional expertise, side payments, and the like. The relative salience of influence resources may be viewed as the weights that should be attached to predictions of the effects of influence attempts.

Another view treats preferences as endogenous. While it still assumes that organizational actors hold resources that may drive decision-making, it differs from the first view, however, by relaxing the assumption of strong preferences for specific action outcomes. The role of administration is to take actions that are designed to help structure more or less plastic preferences. Mechanisms include leadership, especially charismatic leadership, ideology, socialization, recruitment, and environmental constituencies to which individuals have personal or professional loyalty. The organization is understood and managed as a moral community. Common to all these mechanisms is the attempt to foster identifications and loyalties, the normative order providing the backbone of the organization. Management is about forging and changing values, norms, and cognitive characteristics: it may also have to do with preaching and educating.

The role of management in structuring preferences is documented by a set of literature on missionary, professional, and community organizations. Institutionalization as studied by Selznick (1949) offers a vehicle to mobilize an organization for meaning and action. Knowledge, or interpretations in action, structures the community. The theoretical roots of such an approach relate to two different social sciences traditions. Shils (1975) identifies in each society the existence of a center or a central zone that is a phenomenon of the realm of values and beliefs as well as of action. It defines the nature of the sacred and it embodies and propounds an official religion, something that transcends and transfigures the concrete individual existence, the content of authority itself. The periphery in mass society is integrated through a process of civilization. Anthropologists such as Geertz (1973) demonstrate that culture as a collectively sustained symbolic structure is a means of 'saying something of something,' Through emotions common cognitive schemes or common meanings are learned: they provide an interpretative fonction, a local reading of a local experience, which is a story the participants tell themselves about themselves.

More recent contributions have laid down perspectives focused specifically around organizations and their administration. Various processes within firms could actually play the role of center: brainstorming sessions, informal encounters, networks linking persons across departments and units, socialization mechanisms of newcomers, etc. Strong centers can create rigidity phenomena in terms of cognitive blindness, the firm as a community being unable to catch signals emitted by its environment. Daft and Weick (1984) propose a model of organizations as interpretation systems that stresses their sociocognitive characteristics more than the economic ones. Interpretation is the process through which information is given meaning and actions are selected and fulfilled. Kogut and Zander (1996) treat firms as organizations that represent social knowledge of coordination and learning: identity lies at the heart of such social systems, which implies a moral order as well as rules for exclusion.

See also: Closed and Open Systems: Organizational; Conflict: Organizational; Industrial Sociology; Intelligence: Organizational; Learning: Organizational; Management: General; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations: Authority and Power; Organizations, Sociology of

Bibliography

Allison G T 1971 *Essence of Decision: Explaining the Cuban Missile Crisis*. Little, Brown, Boston

Bower J L 1970 *Managing the Resource Allocation Process*. Harvard University Press, Boston

Crozier M 1963 *The Bureaucratic Phenomenon*. University of Chicago Press, Chicago

Cyert R M, March J G 1963 *A Behavioral Theory of the Firm*. Prentice Hall, Englewood Cliffs, NJ

Daft R L, Weick K E 1984 Toward a model of organizations as interpretation systems. *Academy of Management Review* **2**: 284–95

Fayol H 1949 *General and Industrial Management*. Pitman, London

Geertz C 1973 *The Interpretation of Culture*. Basic Books, New York

Kogut B, Zander U 1996 What firms do? Coordination, identity and learning. *Organization Science* **7**: 502–18

Landau M 1969 Redundancy, rationality, and the problem of duplication and overlap. *Public Administration Review* **29**: 349–58

Lindblom C E 1959 The science of 'muddling through.' *Public Administration Review* **19**: 79–88

March J G 1962 The business firm as a political coalition. *Journal of Politics* **24**: 662–78

March J G, Olsen J P 1976 *Ambiguity and Choice in Organizations*. Universitetsforlaget, Bergen, Norway

March J G, Simon H A, Guetzkow H 1958 *Organizations*. Wiley, New York

Selznick P 1949 *TVA and the Grass Roots*. University of California Press, Berkeley, CA

Shils E 1975 *Center and Periphery*. University of Chicago Press, Chicago

Simon H A 1946 The proverbs of administration. *Public Administration Review* **6**: 53–67

Simon H A 1947 *Administrative Behavior*. MacMillan, New York
Taylor F W 1911 *The Principles of Scientific Management*. Harper & Brothers, New York

J.-C. Thoenig

Administrative Law

Administrative law refers to the body of laws, procedures, and legal institutions affecting government agencies as they implement legislation and administer public programs. As such, the scope of administrative law sweeps broadly. In most countries, bureaucratic agencies make up the largest part of the governmental sector and generate most of the decisions having a direct impact on citizens' lives. Administrative law governs agency decisions to grant licenses, administer benefits, conduct investigations, enforce laws, impose sanctions, award government contracts, collect information, hire employees, and make still further rules and regulations.

Administrative law not only addresses a wide and varied array of government actions, it also draws its pedigree from a variety of legal sources. Administrative law, as a body of law, is part constitutional law, part statutory law, part internal policy, and, in some systems, part common law. The organization and structure of administrative agencies can be shaped by constitutions or statutes. The procedures used by these agencies can be dictated by constitutional law (such as to protect certain values such as due process), by generic procedural statutes (such as the US Administrative Procedure Act), or by statutes addressing specific substantive policy issues such as energy, taxation, or social welfare. As a result, administrative procedures can vary significantly across agencies, and even within the same agency across discrete policy issues.

Administrative law, in all its varied forms, speaks ultimately to how government authority can and ought to be exercised. By directing when and how governmental power can be employed, administrative law of necessity confronts central questions of political theory, particularly the challenge of reconciling decision-making by unelected administrators with democratic principles. The study of administrative law is characterized in part by prescriptive efforts to design rules that better promote democratic and other values, including fairness, effectiveness, and efficiency. At its core, administrative law scholarship seeks to understand how law can affect the behavior of governmental officials and organizations in such a way as to promote important social objectives. As such, administrative law is also characterized by positive efforts to explain the behavior of governmental organizations and understand how law influences this behavior. A specific emphasis in administrative law scholarship is placed on the empirical study of how courts influence administrative policy. Although administrative law scholarship has a rich tradition of doctrinal analysis, the insights, and increasingly the methods, of social science have become essential for achieving an improved understanding of how administrative law and judicial review can affect democratic governance.

1. *Administrative Law and Democracy*

Administrative agencies make individual decisions affecting citizens' lives and they set general policies affecting an entire economy, but they are usually headed by officials who are neither elected nor otherwise directly accountable to the public. A fundamental challenge in both positive and prescriptive scholarship has been to analyze administrative decision-making from the standpoint of democracy. This challenge is particularly pronounced in constitutional systems such as the United States' in which political party control can be divided between the legislature and the executive branch, each seeking to influence administrative outcomes. Much work in administrative law aims either to justify administrative procedures in democratic terms or to analyze empirically how those procedures impact on democratic values.

A common way of reconciling decision-making by unelected administrators with democracy has been to consider administrators as mere implementers of decisions made through a democratic legislative process. This is sometimes called the 'transmission belt' model of administrative law (Stewart 1975). Administrators, under this model, are viewed as the necessary instruments used to implement the will of the democratically controlled legislature. Legislation serves as the 'transmission belt' to the agency, both transferring democratic legitimacy to administrative actions and constraining those actions so that they advance legislative goals.

As a positive matter, the 'transmission belt' model underestimates the amount of discretion held by administrative officials. Laws require interpretation, and in the process of interpretation administrators acquire discretion (Hawkins 1992). Legislation often does not speak directly to the varied and at times unanticipated circumstances that confront administrators. Indeed, legislators may sometimes lack incentives for making laws clear or precise in the first place, as it can be to their electoral advantage to appear to have addressed vexing social problems, only in fact to have passed key tradeoffs along to unelected administrators. For some administrative tasks, particularly monitoring and enforcing laws, legislators give administrators explicit discretion over how to allocate their agencies' resources to pursue broad legislative goals.

Scholars disagree about how much discretion legislators ought to allow administrative agencies to

exercise. Administrative minimalists emphasize the electoral accountability of the legislature, and conclude that any legislative delegations to agencies should be narrowly constructed (Lowi 1979). The expansionist view emphasizes most administrators' indirect accountability to an elected executive and contends that legislatures themselves are not perfectly representative, especially when key decisions are delegated internally to committees and legislative staff (Mashaw 1985). While disagreement may persist over the amount of authority to be delegated to agencies, in practice administrative agencies will continue to possess considerable discretion, even under relatively restrictive delegations.

The study of administrative procedure takes it as given that agencies possess discretion. The aim is to identify procedures that encourage administrators to exercise their discretion in reasonable and responsive ways. A leading approach has been to design administrative procedures to promote interest group pluralism (Stewart 1975). Transparent procedures and opportunities for public input give organized interests an ability to represent themselves, and their constituencies, in the administrative process. Such procedures include those providing for open meetings, access to government information, hearings and opportunities for public comment, and the ability to petition the government. Open procedures are not only defended on the grounds of procedural fairness, but also because they force administrators to confront a wide array of interests before making decisions, thus broadening the political basis for administrative policy. These procedures may also protect against regulatory capture, a situation which occurs when an industry comes to control an agency in such a way as to yield private benefits to the industry (Stigler 1971).

A more recent analytic approach called 'positive political economy' seeks to explain administrative procedures as efforts by elected officials to control agency outcomes (McCubbins et al. 1987). Administrative law, according to this approach, addresses the principal–agent problem confronting elected officials when they create agencies or delegate power to administrators. The problem is that administrators face incentives to implement statutes in ways not intended by the coalition that enacted the legislation. It is difficult for legislators continually to monitor agencies and in any case the original legislators will not always remain in power. Analysts argue that elected officials create administrative procedures with the goal of entrenching the outcomes desired by the original coalition. Such procedures can be imposed by the legislative as well as executive branch, and they include formal procedures for legislative review and veto, general requirements for transparency and interest group access, and requirements that agencies conduct economic analysis before reaching decisions.

A recent area of empirical debate has emerged in the United States over which branch of government exerts most control over administrative agencies. The resulting evidence has so far been mixed, as might be expected, since most agencies operate in a complicated political environment in which they are subject to multiple institutional constraints. Indeed, the overall complexity of administrative politics and law presents a major challenge for social scientists seeking to identify the effects of specific kinds of procedures under varied conditions. The recent positive political economy approach advances a more nuanced analytical account of democratic accountability than the simple 'transmission belt' model of administrative law, but the ongoing challenge will be to identify with still greater precision which kinds of procedures, and combinations of procedures, advance the aims of democratic accountability as well as other important social values.

2. *Courts and Administrative Law*

As much as the connections between elected officials and administrators have been emphasized in administrative law, the relationship between courts and administrators has figured still more prominently in the field. Even when administrative procedures are created through legislation, the enforcement of such procedures often remains with judicial institutions. Courts have also imposed their own additional procedures on agencies based on constitutional and sometimes common law principles. As with democratic issues, scholarly attention to the role of the courts has both prescriptive and positive aspects.

The main prescriptive focus has been on the degree to which courts should defer to the decisions made by administrative agencies. Much doctrinal analysis in administrative law acknowledges that administrative agencies' capacity for making technical and policy judgments usually exceeds that possessed by courts. Even in legal systems with specialized administrative courts, agency staff often possess greater policy expertise than judges, not to mention that administrators are probably more democratically accountable than tenured judges. These considerations have long weighed in favor of judicial deference to administrative agencies. On the other hand, it is generally accepted that some credible oversight by the courts bolsters agencies' compliance with administrative law and may improve their overall performance. The prescriptive challenge therefore has been to identify the appropriate strategies for courts to take in overseeing agency decision-making.

This challenge typically has required choosing a goal for judicial intervention, a choice sometimes characterized as one between sound technical analysis or an open, pluralist decision-making process (Shapiro 1988). Courts can defer to an agency's policy judgment, simply ensuring that the agency followed transparent procedures. Or courts can take a careful look at the agency's decision to see that it was based on a

thorough analysis of all relevant issues. The latter approach is sometimes referred to as 'hard look' review, as it calls for judges to probe carefully into the agency's reasoning. Courts also face a choice about whether to defer to agencies' interpretations of their own governing legislation instead of imposing judicial interpretations on the agencies. Prescriptive scholarship in administrative law seeks to provide principled guidance to the judges who confront these choices.

Judicial decisions are influenced in part by legal principles. Empirical research has shown, for example, that after the US Supreme Court decided that agencies' statutory interpretations deserved judicial deference, lower courts made a significant shift in favor of deferring to agency interpretations (Schuck and Elliott 1990). Nevertheless, just as administrators themselves possess residual discretion, so too do judges possess discretion in deciding how deferential to be. Other empirical research suggests that in administrative law, as in other areas of law, political ideology also helps explain certain patterns of judicial decision-making (Revesz 1997).

In addition to empirical research on judicial decision-making, the field of administrative law has been concerned centrally with the impact of judicial review on agency decision-making. Normative arguments about judicial review typically depend on empirical assumptions about the effects courts have on the behavior of administrative agencies. Indeed, most legal scholarship in administrative law builds on the premise that judicial review, if employed properly, can improve governance (Sunstein 1990, Edley 1990). The effects often attributed to judicial review include making agencies more observant of legislative mandates, increasing the analytic quality of agency decision-making, and promoting agency responsiveness to a wide range of interests. Administrators who know that their actions may be subjected to review by the courts can be expected to exercise greater overall care, making better, fairer, and more responsive decisions than administrators who are insulated from direct oversight.

Notwithstanding the beneficial effects of courts on the administrative process, legal scholars also have emphasized increasingly courts' potentially debilitating effects on agencies. It has widely been accepted, for example, that administrators in the United States confront a high probability that their actions will be subject to litigation. Cross-national research suggests that courts figure more prominently in government administration in the USA than in other countries (Brickman et al. 1985, Kagan 1991). The threat of judicial review has been viewed as creating significant delays for agencies seeking to develop regulations (McGarity 1992). In some cases, agencies have been said to have retreated altogether from efforts to establish regulations. The US National Highway Traffic Safety Administration (NHTSA) is usually cited as the clearest case of this so-called 'ossification' effect, with one major study suggesting that NHTSA has shifted away from developing new auto safety standards in order to avoid judicial reversal (Mashaw and Harfst 1990). Other research, however, indicates that the threat of judicial interference in agency decision-making has generally been overstated. Litigation challenging administrative action in the United States occurs less frequently than is generally assumed (Harrington 1988, Coglianese 1997), and some research indicates that agencies can surmount seemingly adverse judicial decisions to achieve their policy objectives (Jordan 2000).

Concern over excessive adversarialism in the administrative process persists in many countries. Government decision makers worldwide are pursuing collaborative or consensus-based processes when creating and implementing administrative policies. In the USA, an innovation called negotiated rulemaking has been used by more than a dozen administrative agencies, specifically in an effort to prevent subsequent litigation. In a negotiated rulemaking, representatives from government, business, and nongovernmental organizations work toward agreement on proposed administrative policies (Harter 1982). In practice, however, these agreements have not reduced subsequent litigation, in part because litigation has ordinarily been less frequent than generally thought (Coglianese 1997). Moreover, even countries with more consensual, corporatist policy structures experience litigation over administrative issues, often because lawsuits can help outside groups penetrate close-knit policy networks (Sellers 1995). In pluralist systems such as the USA, litigation is typically viewed as a normal part of the policy process, and insiders to administrative processes tend to go to court at least as often as outsiders (Coglianese 1996).

Courts' impact on the process of governance has been and will remain a staple issue for administrative law. In order to understand how law can have a positive influence on governing institutions within society, it is vital to examine how judicial institutions affect the behavior of government organizations. Empirical research on the social meaning and behavioral impact of litigation in an administrative setting has the potential for improving prescriptive efforts to craft judicial principles or redesign administrative procedures in ways that contribute to more effective and legitimate governance.

3. *The Future of Administrative Law*

Administrative law lies at several intersections, crossing the boundaries of political theory and political science, of public law and public administration. As the body of law governing governments, the future of administrative law rests in expanding knowledge about how law and legal institutions can advance core

political and social values. Democratic principles will continue to dominate research in administrative law, as will interest in the role of courts in improving administrative governance. Yet administrative law can and should expand to meet new roles that government will face in the future. Ongoing efforts at deregulation and privatization may signal a renegotiation of the divisions between the public and private sectors in many countries, the results of which will undoubtedly have implications for administrative law. Administrative law may also inform future governance in an increasingly globalized world, providing both normative and empirical models to guide the creation of international administrative institutions that advance both public legitimacy and policy effectiveness. No matter where the specific challenges may lie in the future, social science research on administrative law will continue to support efforts to design governmental institutions and procedures in ways that increase social welfare, promote the fair treatment of individuals, and expand the potential for democratic decision making.

See also: Civil Law; Democracy; Dispute Resolution in Economics; Disputes, Social Construction and Transformation of; Environment Regulation: Legal Aspects; Governments; Judicial Review in Law; Law and Democracy; Legislatures: United States; Legitimacy; Litigation; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Occupational Health and Safety, Regulation of; Public Administration: Organizational Aspects; Public Administration, Politics of; Rechtsstaat (Rule of Law: German Perspective); Regulation and Administration

Bibliography

Brickman R, Jasanoff S, Ilgen T 1985 *Controlling Chemicals: The Politics of Regulation in Europe and the United States*. Cornell University Press, Ithaca, NY

Coglianese C 1996 Litigating within relationships: Disputes and disturbance in the regulatory process. *Law and Society Review* **30**: 735–65

Coglianese C 1997 Assessing consensus: The promise and performance of negotiated rulemaking. *Duke Law Journal* **46**: 1255–349

Edley C F Jr 1990 *Administrative Law: Rethinking Judicial Control of Bureaucracy*. Yale University Press, New Haven, CT

Harrington C 1988 Regulatory reform: Creating gaps and making markets. *Law and Policy* **10**: 293

Harter P J 1982 Negotiating regulations: A cure for malaise. *Georgetown Law Journal* **71**: 1–118

Hawkins K (ed.) 1992 *The Uses of Discretion*. Oxford University Press, Oxford, UK

Jordan W S 2000 Ossification revisited: Does arbitrary and capricious review significantly interfere with agency ability to achieve regulatory goals through informal rulemaking? *Northwestern University Law Review* **94**: 393–450

Kagan R A 1991 Adversarial legalism and American government. *Journal of Policy Analysis and Management* **10**: 369–406

Lowi T J 1979 *The End of Liberalism: The Second Republic of the United States*. W. W. Norton, New York

Mashaw J L 1985 Prodelegation: Why administrators should make political decisions. *Journal of Law, Econ., and Organization* **1**: 81

Mashaw J L, Harfst D L 1990 *The Struggle for Auto Safety*. Harvard University Press, Cambridge, MA

McCubbins M, Noll R, Weingast B 1987 Administrative procedures as instruments of political control. *Journal of Law, Econ., and Organization* **3**: 243

McGarity T O 1992 Some thoughts on 'deossifying' the rulemaking process. *Duke Law Journal* **41**: 1385–462

Revesz R L 1997 Environmental regulation, ideology, and the D.C. Circuit. *Virginia Law Review* **83**: 1717–72

Schuck P H, Elliott E D 1990 To the Chevron station: An empirical study of federal administrative law. *Duke Law Journal* **1990**: 984–1077

Sellers J M 1995 Litigation as a local political resource: Courts in controversies over land use in France, Germany, and the United States. *Law and Society Review* **29**: 475

Shapiro M 1988 *Who Guards the Guardians? Judicial Control of Administration*. University of Georgia Press, Athens, GA

Stewart R B 1975 The reformation of American administrative law. *Harvard Law Review* **88**: 1667–813

Stigler G J 1971 The theory of economic regulation. *Bell Journal of Econ. and Manag. Sci.* **2**: 3

Sunstein C R 1990 *After the Rights Revolution: Reconceiving the Regulatory State*. Harvard University Press, Cambridge, MA

C. Coglianese

Adolescence, Psychiatry of

For psychiatrists, adolescence holds a particular fascination. A period of rapid and often dramatic growth and development, adolescence presents countless challenges to those faced with making the transition from childhood to adulthood. While the vast majority of adolescents make this transition without any need for help from mental health professionals, others are not so fortunate. The psychiatric disorders of this period present treatment providers with unique challenges, because as with adolescence itself, they represent a distinctive mixture of childhood and adult components. However, adolescent psychiatric patients are frequently capable of making tremendous progress, and treating adolescents is often as rewarding as it is challenging.

In this article, we will briefly review the major psychiatric disorders affecting adolescents. We have divided the article into six sections, each representing a separate area of psychopathology. In each section, we list the major disorders, focusing on definitions, etiology, and treatment. For the purposes of this article, we have decided to follow the diagnostic criteria of the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (*DSM-IV*, American Psychiatric Association 1994). We based this decision on the fact that the *DSM-IV* is the most

widely used diagnostic system not only in the United States, but in the world as a whole (Maser et al. 1991).

1. *Anxiety Disorders*

Everyone experiences anxiety; this is especially true for adolescents. Yet whereas normal levels of anxiety can be adaptive, helping individuals avoid danger and promoting development, too much anxiety can become problematic. When anxiety hinders, rather than helps an individual function in the world, it is considered pathological. Disorders that involve pathological levels of anxiety are referred to as 'anxiety disorders,' and they are quite common among adolescents. The most prevalent adolescent anxiety disorders are: *separation anxiety disorder*; *generalized anxiety disorder*; *obsessive-compulsive disorder*; *social phobia*; and *panic disorder*.

1.1 *Separation Anxiety Disorder*

The essential feature of separation anxiety disorder (SAD) is excessive anxiety concerning separation from home or attachment figures. While this disorder is frequently associated with younger children, it is not uncommon among adolescents. There is some evidence to suggest that this disorder is more prevalent in lower socio-economic groups, and that it occurs most frequently in girls (Last et al. 1987).

The presentation of SAD varies with age. Younger children tend to express their anxiety through specific fears of harm coming to their attachment figures. The child might worry that a parent will be kidnapped, or attacked by a burglar. Adolescents, alternatively, frequently deny anxiety about separation; as a result diagnosis becomes more difficult. Among adolescents, separation anxiety is often expressed through school refusal and recurrent somatic complaints (Francis et al. 1987).

Behavioral interventions, psychodynamic psychotherapy, and pharmacotherapy have all been shown to be effective treatment interventions for SAD. Frequently, a multimodal approach (one that uses various interventions depending on the characteristics of the patient and the pathology involved) is indicated. Medication alone is not recommended (American Academy of Child and Adolescent Psychiatry 1997).

1.2 *Generalized Anxiety Disorder*

Previously referred to as overanxious disorder, generalized anxiety disorder (GAD) is characterized by excessive anxiety and worry. Adolescents who suffer from this disorder tend to worry about everything: their competence, their appearance, their health, even potential disasters such as tornados or nuclear war. This anxiety often causes significant impairment in social and school functioning. Other manifestations such as sleep disturbances and difficulty concentrating are common. Rates of GAD appear to be the same for boys and girls (Last et al. 1987).

The treatment of choice for GAD among adolescents is psychotherapy—including cognitive, psychodynamic, and family systems approaches. While some studies have shown symptom reduction following treatment with anti-anxiety medications (e.g., Kutcher et al. 1992), results have been equivocal with regard to the overall effectiveness of pharmacotherapy in treating GAD.

1.3 *Obsessive-compulsive Disorder*

Obsessive-compulsive disorder (OCD) is characterized by recurrent obsessions and/or compulsions of sufficient severity to significantly disrupt an individual's day-to-day functioning. Common obsessions include themes of contamination, of harming others whom one cares about, and of inappropriate sexual urges. Common compulsions (which often serve to ameliorate the anxiety brought about by obsessive thoughts) include the repeated washing of hands, counting tiles on a floor or ceiling, and checking to make sure that a door is locked. One need not have both obsessions and compulsions in order to receive a diagnosis of OCD. The disorder is more common among girls than boys.

OCD frequently goes undiagnosed among adolescents (Flament et al. 1988). This is particularly unfortunate given the fact that both cognitive-behavioral therapy and pharmacotherapy have been shown to be effective in treating the disorder (American Academy of Child and Adolescent Psychiatry 1999). While psychodynamic psychotherapy can be useful in helping an adolescent deal with the effects of the disorder on his or her life, it has not been shown to be effective in treating the disorder itself.

1.4 *Social Phobia*

Adolescents with social phobia are more than just shy. Their fear of interacting with new persons, or of being embarrassed in a social situation, provokes such intense anxiety that they often avoid such situations altogether. It is not uncommon, for instance, for persons with social phobia to avoid eating in public, using public restrooms, or even writing in public. The anxiety associated with social phobia is often experienced somatically: racing pulse; profuse sweating; and lightheadedness are all physical manifestations associated with the disorder.

As with specific phobias, the treatment of choice for social phobia is behavior therapy. There is also some evidence to suggest that medications such as *Prozac* (Fluoxetine) may be effective in reducing the symptoms of social phobia among adolescents (Birmaher et al. 1994).

1.5 Panic Disorder

The essential feature of panic disorder is the presence of panic attacks, which involve the sudden, unexpected onset of symptoms such as palpitations, sweating, trembling, shortness of breath, and fears of 'going crazy.' Panic disorder usually first appears during adolescence, with an average age of onset of between 15 and 19 years (Von Korff et al. 1985). Panic disorder may or may not be accompanied by agoraphobia, the fear of being in places or situations from which escape might be difficult. Panic disorder with agoraphobia can be especially debilitating, as adolescents with this disorder may become virtual prisoners, afraid to leave their homes even to go to school.

Panic disorder appears to occur more frequently in girls than in boys (Whitaker et al. 1990). Children who carry an earlier diagnosis of separation anxiety disorder are at increased risk of developing panic disorder (Biederman et al. 1993).

The treatment for panic disorder in adolescents is similar to that of panic disorder in adults: cognitive and behavioral therapy, and pharmacotherapy. Unfortunately, these treatments have not been well researched in adolescent populations, and clinicians are often forced to extrapolate from studies that have been conducted on adults.

2. Mood Disorders

As the name suggests, the essential feature of a mood disorder is a disturbance in mood. Mood disorders may involve depressed mood, elevated mood, or both. While normal fluctuations in mood are a regular occurrence for all of us, when an individual's ability to function is hindered by prolonged periods of depressed or elevated mood, he or she may be suffering from a mood disorder.

Many people hold the misperception that adolescence is inevitably a time of great emotional turmoil. Research has demonstrated that this is not the case (e.g., Offer and Schonert-Reichl 1992). In fact, when an adolescent experiences prolonged periods of 'moodiness,' he or she may in fact be suffering from a mood disorder. The most common mood disorders of adolescence are *major depressive disorder* and the *bipolar disorders*.

2.1 Major Depressive Disorder

The essential feature of major depressive disorder (MDD) in adolescents is a period lasting at least two weeks during which there is either depressed or irritable mood or a loss of interest or pleasure in nearly all activities. Changes in sleep and/or appetite, decreased energy, and feelings of worthlessness and guilt are also often present. Suicidal thoughts, plans, and attempts are not uncommon.

Table 1
US suicide rates[a] by race and sex for 15–19 year-olds, 1994–97

Race		1997	1996	1995	1994
White	M	15.97	16.27	18.17	18.43
	F	3.51	3.81	3.25	3.53
African-American	M	11.36	11.46	13.74	16.56
	F	2.75	1.82	2.30	2.44
Other	M	13.87	16.70	13.23	16.73
	F	3.19	4.86	3.48	6.03

Source: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control (revised July 14, 1999).
a Per 100,000.

Rates of depressed mood and clinical depression rise significantly during adolescence. By middle to late adolescence, rates of MDD approach those seen in adult populations. Depression is much more likely to occur in girls than boys—one study found that among 14–16 year-olds, girls with depression outnumbered boys by a ratio of 5:1 (Kashani et al. 1987). Such figures may overestimate actual gender differences in adolescent depression, however, because depressed boys are more likely than girls to express their depression by demonstrating behavior problems.

Cognitive-behavioral therapy, psychodynamic psychotherapy, family therapy, group therapy, and pharmacotherapy are all used to treat adolescent depression. Unfortunately, outcome studies on the above treatments in adolescent populations are sparse, and caution is indicated in applying adult-based research findings to adolescent populations. Given the relatively high mortality rates associated with depression and suicide (Bell & Clark 1998), it is critical that clinicians identify adolescents with depression and make appropriate interventions in a timely manner.

With regard to adolescent suicide, it is troubling to note that rates of completed suicide within this age group in the United States tripled between 1950 and 2000 (Rosewater and Burr 1998). However, the most recent statistics available suggest that this trend may have reversed itself (see Table 1). Major risk factors for adolescent suicide include psychiatric illness—especially the presence of a mood disorder—and substance abuse. Efforts to prevent adolescent suicide include school-based programs and teacher training to identify at risk youth.

2.2 Bipolar Disorders

There are two major bipolar disorders: bipolar I disorder and bipolar II disorder. Both disorders involve the presence of mania. Mania is defined as a period of abnormally elevated, expansive, or irritable mood. When an individual experiences mania in its severest form, called a manic episode, he or she usually requires psychiatric hospitalization. In its less severe

form, referred to as a hypomanic episode, functioning is impaired, but not to such an extent that a hospitalization is necessitated. Bipolar I disorder requires the existence of a manic episode, and bipolar II disorder requires the existence of a hypomanic episode as well as a major depressive episode.

Bipolar disorders in adolescents are much less common than the other mood disorders. While there is some evidence to suggest that bipolar disorders occur more frequently in girls than in boys (e.g., Krasa and Tolbert 1994), other studies have failed to replicate these findings.

A multimodal approach, incorporating psychosocial and psychoeducational interventions as well pharmacological ones, is the treatment of choice for adolescent bipolar disorder. Persons with bipolar disorders frequently have high rates of noncompliance with treatment, perhaps because they miss the 'high' associated with periods of mania. As a result, relapse prevention (including family involvement with treatment and frequent follow-ups) is a crucial aspect of treating bipolar disorder in adolescents.

3. *Eating Disorders*

Many people 'watch their weight,' and adolescents (primarily adolescent girls) are no different. In fact, adolescents are particularly concerned about appearance and acceptance; as a result they are especially affected by societal pressures to be thin. When a healthy awareness of one's ideal weight gives way to excessive measures in order to lose weight and keep it off, one is at risk of acquiring an eating disorder. There are two primary eating disorders: *anorexia nervosa* and *bulimia nervosa*.

3.1 *Anorexia Nervosa*

The essential features of anorexia nervosa include a distorted body image, an intense fear of gaining weight, and a refusal to maintain a minimally normal body weight. Postmenarchal women with this disorder are often amenorrheic (they stop having regular menstrual periods). Up to 95 percent of patients with anorexia nervosa are females who tend to be both white and middle to upper-middle class (Herzog and Beresin 1997).

If treated soon after onset, anorexia nervosa in adolescents has a relatively good prognosis. However, if not treated it may become a chronic condition that carries with it serious and even life-threatening consequences. Treatment approaches for this disorder include behavioral modification techniques, family therapy, and pharmacotherapy. Insight-oriented psychotherapy can be useful, but not during acute phases. In severe cases, psychiatric and medical hospitalizations are often required.

3.2 *Bulimia Nervosa*

Persons suffering from bulimia nervosa alternate binge eating with purging (vomiting, use of laxatives). As with anorexia nervosa, bulimia nervosa primarily affects females (9 out of 10 sufferers are women). Treatment of bulimia nervosa among adolescents differs from that of anorexia nervosa primarily in that the former rarely requires hospitalization.

4. *Psychotic Disorders*

Psychotic disorders are among the most devastating psychiatric illnesses. They involve a disturbance in an individual's ability to perceive the world as others do. Persons suffering from psychotic disorders may experience auditory hallucinations ('hearing voices'), delusions, and disordered thinking, among other symptoms.

Psychotic disorders are rare among adolescents, but they can and do occur in this age group. The most common psychotic disorder seen in adolescent patients is *schizophrenia*.

4.1 *Schizophrenia*

Schizophrenia involves what are referred to as 'positive' and 'negative' symptoms. Positive symptoms include hallucinations, delusions, and disorganized thinking. Negative symptoms include diminished emotional expressiveness, decreased productivity of thought and speech, and difficulty initiating goal-directed behaviors. As a result of these symptoms, adolescents with schizophrenia have a great deal of difficulty functioning in the world. Their personal hygiene may deteriorate, their ability to perform in school may decrease, and their relationships with other people may suffer.

Initial symptoms of schizophrenia in adolescents may appear rapidly or slowly. There is some evidence to suggest that a slow onset is associated with a worse prognosis. The disorder occurs more frequently in boys than girls before the age of 14. As mentioned previously, adolescent schizophrenia is rare, with only 4 percent of adult schizophrenics having developed the disorder before the age of 15 (Tolbert 1996).

The recent introduction of new antipsychotic medications has greatly changed the way schizophrenia is treated. While by no means a cure-all, these drugs provide many schizophrenics with relief from their symptoms without sedating them to such an extent that they are unable to function. Furthermore, thanks largely to these medications, persons with schizophrenia require fewer hospitalizations than they once did, and are often able to lead productive lives. Nevertheless, schizophrenia is more often than not a chronic condition that requires intensive treatment in order to keep its symptoms under control. In addition

to medication, supportive individual and group psychotherapy are important components of treating schizophrenia.

5. Disruptive Behavior Disorders

When adolescents are disruptive, their behavior tends to attract the attention of their parents and teachers. As a result, disruptive behavior disorders (DBDs) are a common reason for referral to child and adolescent psychiatrists. The essential features of DBDs are aggression, poor self-regulation, and excessive opposition to authority. The most common DBDs in adolescents are: *attention-deficit hyperactivity disorder*; *conduct disorder*; and *oppositional defiant disorder*.

5.1 Attention-deficit Hyperactivity Disorder

Attention-deficit hyperactivity disorder (ADHD) involves a chronic pattern of distractibility and hyperactivity or impulsivity that causes significant impairment in functioning. The presentation of this disorder varies with age. For instance, whereas younger children with ADHD may display more overt signs of agitation, adolescents may experience an internal feeling of restlessness. Other signs of adolescent ADHD include an inability to complete independent academic work and a propensity toward risky behaviors (automobile and bicycle accidents are not uncommon among adolescents with ADHD).

ADHD is a highly prevalent disorder. Some have estimated that it may account for as much as 50 percent of the patients in child psychiatry clinic populations (Cantwell 1996). Prevalence of the disorder appears to peak during late childhood/early adolescence.

The treatment strategy of choice for adolescent ADHD is a multimodal approach that incorporates psychosocial interventions with pharmacotherapy. Parent management training, school-focused interventions, and individual therapy are all important elements of such an approach. Additionally, central nervous system stimulants have proven effective in treating the symptoms of ADHD, with response rates exceeding 70 percent (Cantwell 1996).

5.2 Conduct Disorder

The essential feature of conduct disorder (CD) is a persistent pattern of violating the rights of others and disregarding societal norms. Common behavior associated with CD include aggressive or violent conduct, theft, and vandalism.

CD is one of the most prevalent forms of psychopathology in adolescents, and it is the single most common reason for referrals of adolescents for psychiatric evaluations. As with ADHD, boys with CD outnumber girls with the disorder. This gender gap becomes less pronounced toward late adolescence.

Table 2
US homicide rates[a] by race and sex for 15–19 year-olds, 1994–97

Race		1997	1996	1995	1994
White	M	10.92	12.07	14.30	14.98
	F	2.87	2,92	3.91	3.43
African-American	M	85.09	99.96	109.32	134.20
	F	10.57	12.94	16.37	15.11
Other	M	17.18	16.91	24.22	28.58
	F	3.61	4.41	3.48	2.90

Source: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control (revised July 14, 1999).
a Per 100,000.

The most effective treatments for CD are family interventions and psychosocial interventions (such as problem-solving skills training). Preventative measures aimed at reducing levels of CD have shown promise, but more research is needed before their effectiveness can be fully assessed (Offord and Bennett 1994).

5.3 Oppositional Defiant Disorder

When adolescent defiance toward authority figures becomes excessive and out of control, it may in fact represent a syndrome called oppositional defiant disorder (ODD). This disorder involves persistently negativistic and hostile attitudes toward those in authority, as well as spiteful, vindictive, and generally defiant behavior. In order to qualify for a diagnosis of ODD, an adolescent's behavior must clearly be excessive in comparison with that which is typically observed in his or her peers.

Treatment for ODD is similar to that for CD, utilizing a multimodal approach that emphasizes both individual and family therapy as well as psychosocial interventions.

Diagnoses of CD and ODD are frequently associated with juvenile delinquency. As a result, a high percentage of adolescents with these disorders find themselves involved with the criminal justice system. Yet as the emphasis in the American criminal justice system shifts from rehabilitation to punishment, fewer and fewer inmates receive the mental health services that they need. This is particularly unfortunate, because a number of programs aimed at reducing juvenile delinquency have been demonstrated to be effective in doing so (Zigler et al. 1992).

As the number of incarcerated juveniles increases, so do the number of crimes committed by adolescents. The overall level of arrests for juveniles in 1996, for instance, was 60 percent higher than it was in 1987 (Snyder 1997). However, as with adolescent suicide, the number of adolescent homicides in the United States has declined since then (see Table 2).

6. Substance Use Disorders

In recent years, there has been an increased effort on the part of many governments to curb adolescent substance use and abuse. The results of such efforts have been mixed. Alcohol is currently the most widely used psychoactive substance among adolescents, with marijuana being a close second.

There are many treatment philosophies when it comes to adolescent substance abuse. Inpatient programs, outpatient programs, and residential treatment centers have all been shown to be effective methods of treating substance use disorders. Recently, a greater emphasis has been placed on preventative measures, including school-based programs aimed at educating students on the dangers of substance abuse. However, the effectiveness of such programs has yet to be adequately demonstrated.

7. Summary

In this article we reviewed six areas of adolescent psychopathology: anxiety disorders; mood disorders; eating disorders; psychotic disorders; disruptive behavior disorders; and substance use disorders. We looked at the major diagnoses in each category, focusing on definitions, etiology, and treatment. In the bibliography, we indicate those readings that are recommended to those wishing to learn more about the topics covered

See also: Adolescent Behavior: Demographic; Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Adolescent Vulnerability and Psychological Interventions; Antisocial Behavior in Childhood and Adolescence; Child and Adolescent Psychiatry, Principles of; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Mental Health Programs: Children and Adolescents; Obesity and Eating Disorders: Psychiatric; Socialization in Adolescence; Substance Abuse in Adolescents, Prevention of; Suicide; Youth Culture, Sociology of

Bibliography

American Academy of Child and Adolescent Psychiatry 1997 Practice parameters for the assessment and treatment of children and adolescents with anxiety disorders. *Journal of the American Academy of Child and Adolescent Psychiatry* **36** (10 Suppl.): 69S–84S

American Academy of Child and Adolescent Psychiatry 1999 Practice parameters for the assessment and treatment of children and adolescents with obsessive-compulsive disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, **37**(10 Suppl.): 27S–45S

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders,* 3rd edn. (*DSM-III*). American Psychiatric Association, Washington, DC

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders,* 4th edn. (*DSM-IV*). American Psychiatric Association, Washington, DC

Bauman A, Phongsavan P 1999 Epidemiology of substance use in adolescents: prevalence, trends, and policy implications. *Drug and Alcohol Dependence* **55**: 187–207

Bell C C, Clark D C 1998 Adolescent suicide. *Pediatric Clinics of North America* **45**(2): 365–80

Biederman J, Rosenbaum J F, Bolduc-Murphy E A, Faraone S V, Hirshfeld C J, Kagan J 1993 A 3-year follow-up of children with and without behavioral inhibition. *Journal of the American Academy of Child and Adolescent Psychiatry* **32**(4): 814–21

Birmaher B, Waterman S, Ryan N, Cully M, Balach L, Ingram J, Brodsky M 1994 Fluoxetine for childhood anxiety disorders. *Journal of the American Academy of Child and Adolescent Psychiatry* **33**(7): 993–9

Bravender T, Knight J R 1998 Recent patterns of use and associated risks of illicit drug use in adolescents. *Current Opinion in Pediatrics* **10**: 344–9

Cantwell D P 1996 Attention deficit disorder: a review of the past 10 years. *Journal of the American Academy of Child and Adolescent Psychiatry* **35**(8): 978–87

Flament M F, Whitaker A, Rapoport J L, Davies M, Zaremba Berg C, Kalikow K, Sceery W, Shaffer D 1988 Obsessive compulsive disorder in adolescence: an epidemiological study. *Journal of the American Academy of Child and Adolescent Psychiatry* **27**(6): 764–71

Francis G, Last C G, Strauss C C 1987 Expression of separation anxiety disorder: the roles of age and gender. *Child Psychiatry and Human Development* **18**: 82–9

Herzog D B, Beresin E V 1997 Anorexia nervosa. In: Weiner J M (ed.) *Textbook of Child and Adolescent Psychiatry*, 2nd edn. American Psychiatric Press, Washington, DC, pp. 543–561

Kashani J, Beck N C, Hoeper E, Fallahi C, Corcoran C M, McAlister J A, Rosenberg T K, Reid J C 1987 Psychiatric disorders in a community sample of adolescents. *American Journal of Psychiatry* **144**(5): 584–9

Krasa N R, Tolbert H A 1994 Adolescent bipolar disorder: a nine year experience. *Journal of Affective Disorders* **30**: 175–84

Kutcher S P, Reiter S, Gardner D M, Klein R G 1992 The pharmacotherapy of anxiety disorders in children and adolescents. *Psychiatric Clinics of North America* **15**(1): 41–67

Last C G, Hersen M, Kazdin A E, Finkelstein R, Strauss C C 1987 Comparison of DSM-III separation anxiety and overanxious disorders: demographic characteristics and patterns of comorbidity. *Journal of the American Academy of Child and Adolescent Psychiatry* **26**: 527–31

Last C G, Strauss C C, Francis G 1987 Comorbidity among childhood anxiety disorders. *Journal of Nervous and Mental Disease* **175**: 726–30

Maser J D, Klaeber C, Weise R E 1991 International use and attitudes toward DSM-III and DSM-III-R: growing consensus in psychiatric classification. *Journal of Abnormal Psychology* **100**: 271–9

Offer D, Schonert-Reichl K A 1992 Debunking the myths of adolescence: findings from recent research. *Journal of the American Academy of Child and Adolescent Psychiatry* **31**(6): 1003–14

Offord D R, Bennett K J 1994 Conduct disorder: long-term outcomes and intervention effectiveness. *Journal of the Ameri-*

can Academy of Child and Adolescent Psychiatry **33**(8): 1069–78
Rosewater K M, Burr B H 1998 Epidemiology, risk factors, intervention, and prevention of adolescent suicide. *Current Opinion in Pediatrics* **10**: 338–43
Snyder H 1997 Juvenile arrests 1996. Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention, US Department of Justice, Washington, DC
Tolbert H A 1996 Psychoses in children and adolescents: a review. *Journal of Clinical Psychiatry* **57** (Suppl. 3): 4–8
Von Korff M R, Eaton W W, Keyl P M 1985 The epidemiology of panic attacks and panic disorder: results of three community studies. *American Journal of Epidemiology* **122**: 970–81
Whitaker A, Johnson J, Shaffer D, Rapoport J, Kalikow K, Walsh B T, Davies M, Braiman S, Dolinsky A 1990 Uncommon troubles in young people: prevalence estimates of selected psychiatric disorders in a nonreferred adolescent population. *Archives of General Psychiatry* **47**: 487–96
Wiener J M 1997 Oppositional defiant disorder. In: Weiner J M (ed.) *Textbook of Child and Adolescent Psychiatry*, 2nd edn. American Psychiatric Press, Washington, DC, pp. 459–63
World Health Organization 1994 *Lexicon of Psychiatric and Mental Health Items* (*ICD-10*). World Health Organization, Geneva, Switzerland
Zigler E, Taussig C, Black K 1992 Early childhood intervention: a promising preventative for juvenile delinquency. *The American Psychologist* **47**(8): 997–1006

D. Offer and D. Albert

Adolescence, Sociology of

Adolescence, as a stage in the life course, was not invented during the early decades of the twentieth century, as is sometimes suggested by cultural historians. It was, however, identified and institutionalized during the period when many Western societies were shifting from primarily agrarian to predominantly industrial economies. The extension of schooling and the emergence of a high paying labor market, accompanied by the disappearance of employment opportunities for youth, all contributed importantly to create a more distinct phase between childhood and adulthood—a period when parental control was relinquished and peer influence became more prominent. Prior to the twentieth century, youth remained an ambiguous and ill-defined period including children and teenagers or even young adults who remained semidependent well into adulthood.

1. The Discovery of Adolescence

The 'discovery' of adolescence does not imply that youths had not always experienced some of the features of adolescence—desiring greater autonomy, becoming more sensitive to peer influence, or questioning adult authority and limits. Moreover, puberty obviously has always occurred, with its potentially unsettling effects. From all accounts, teenagers and young adults before the twentieth century could be disruptive and, on occasion, threatening to the social order (Shorter et al. 1971, Kett 1977). However, the period of adolescence was not universally noted until after G. Stanley Hall popularized the term, helping to draw professional and public attention to this part of the lifespan. No doubt, too, the creation of developmental science in psychology, sociology, and anthropology helped to establish expectations, norms, and social understanding.

The idea that adolescence is especially problematic was presumed by Hall and his followers. Adolescence, Hall argued, has its source in the disjuncture of biology and culture. The asynchrony of physical development and social maturation introduces the cultural dilemma of managing youths who are physically but not socially adults. Relegated to a social category in which they are treated neither as children or adults, adolescents are inclined to turn away from the adult world, at least temporarily, and regard age peers as their natural allies. At the same time, in modern economies, parental oversight inevitably declines as the family relies increasingly on outside institutions, most notably the school and community. This process reinforces the power of peers, as youths are socially channeled into settings and institutions that generally do not afford the same level of social control provided inside the familial household.

Hall's hypotheses surely overstated what had occurred, but it brilliantly foreshadowed processes that came about in later decades. Moreover, the cultural construct of adolescence immediately took root in US society, where it became almost an ideology that helped to bring about the phenomenon of adolescence itself. In a society that already revered personal initiative and self-direction, it is easy to see why the US teenager emerged from the wings. This social category seemed almost culturally inevitable and developmentally imperative. Much of developmentally oriented social science merely reframed or institutionalized what was, to some extent, already familiar practice. Youths had always been granted some leeway to 'sow their wild oats' in rural communities. In both urban areas and rural communities, youths had long been apprenticed or boarded out. This was undoubtedly a time-honored way of managing difficult and economically unproductive children by sending them to the households of kin or neighbors for moral education as well as economic training (Morgan et al. 1993).

Adolescence as a life stage was more actively embraced in the USA and other Anglophone nations than in Continental Europe, where parental and community controls remain relatively high. Yet in Europe, too, the cultural construct of adolescence has taken hold to some extent. Differences in the salience of this age category across nations reveal the degree to

which adolescence has cultural, social, and political dimensions that are at least as important as the economic sources to which greater attention has been shown in social science research and discussions (Bachman et al. 1978).

Although 'discovered' in the first decade of the twentieth century, the study of adolescence did not take firm root until the middle part of the century, when legions of social scientists began to focus on the problematic features of this life stage in Western nations. Work in psychology by Karl Mannheim, Charlotte Buhler, Kurt Lewin, Roger Barker, and Gardner Murphy; in sociology by W. I. Thomas, Robert Lynd, Helen Lynd, and Willard Waller; and, in anthropology by Ruth Benedict and Margaret Mead, to mention but a few of the luminaries, helped to establish adolescence and early adulthood as a legitimate field of cross-disciplinary research in the second third of the century.

In the aftermath of World War II, however, a widespread estrangement occurred between European traditions of research (which were diminished for two or three decades after the war) and US social science. In the USA, fields of study became more specialized by discipline. This specialization is evident in the withdrawal of cultural anthropology from the examination of postindustrial societies, the narrowing of psychology to cognitive and often acontextual studies, and the limited attention given to the biological and psychological features of human development by sociologists. There were important exceptions such as the monumental program of research by Urie Bronfenbrenner and his students, the seminal writing of Erik Erikson that continued to focus on cross-cultural differences, and several classic field studies in sociology by Hollinghead and Redlick, Eaton and Weil, and Coleman in the middle decades of the twentieth century. But all of these writers bucked a trend toward disciplinary segregation that continued at least until the 1990s.

While specialization is likely to continue, efforts at international and disciplinary integration occurred in the latter decades of the twentieth century. The formation of cross-national collaborations and societies began to spring up in Europe and the USA. More and more, theory and research practice dictates against exclusive treatments by particular disciplines and demands a more holistic treatment that examines development in multiple contexts, using multiple methods including cross-national comparisons, ethnography, and historical research.

2. *The Demography of Adolescence*

In many respects, the history of adolescence reflects and responds to demographic changes that occurred during the twentieth century. Until the twentieth century, youths made a prolonged and often ill-coordinated transition from childhood to adulthood. Labor began early and independence was often delayed, depending on opportunities as much as skills. Accordingly, youths frequently resided with their parents or boarded with employers. Establishing one's own household often occurred after marriage, which itself was delayed by economic considerations and controlled to some extent by family interests.

In the twentieth century, the growth of industrial jobs liberated youths from kinship control and gradually created a labor market shaped by economic needs and demand for skills. Educational training became more necessary and parental influence became more indirect; it was achieved through parents' investment in schooling. The two world wars also lessened parental influence by exposing legions of young men to the authority of the state. In the USA, the benefits granted to veterans allowed large numbers of men to marry earlier than they otherwise might. This helped to spark the 'marriage rush' in the middle of the century that preceded the 'baby boom.' In a matter of two decades, the period between childhood and adulthood contracted dramatically. Youths married and formed families earlier than they ever had before. The time between schooling, leaving the parental household, and establishing one's own family shrank by several years in most Western countries from the 1940s to the 1960s. A widely read book in the USA declared that adolescence was vanishing (Friedenberg 1964). Social institutions, most notably the high school, co-opted youth, discouraging dissent and thwarting individual development. Youths were being 'oversocialized,' and prematurely inducted into adult society. The problem with adolescents in the 1950s was that there was not enough adolescence.

This idea quickly disappeared in the 1960s as the huge baby boom cohorts in all Western countries entered their teens. At the same time, the price of entering adulthood grew as the labor market began to require higher skills and jobs growth slowed. The growing need for credentials and the economic slowdown swelled the ranks of high schools and colleges, creating a heightened age-consciousness that was reflected in a growing youth culture. Commercialization and the expansion of media channels directed toward youth helped to shape a popular culture that divided youths from adults. Finally, the Vietnam War had a huge effect on both the culture and social identity of teenagers and young adults. Youth was prolonged and politicized in the late 1960s and early 1970s.

The rapid slowdown of Western economies in the 1970s and 1980s might have been expected to continue to the trend toward greater age-oriented political consciousness, but this did not occur. The end of the Vietnam War and the public reaction to the cultural era of the 1960s seemed to put an end to politically oriented youth culture. The aging baby boomers began to become more conservative as they entered the labor force and began to form families.

At the same time, youthful behavior was perceived to be increasingly problematic in the 1970s and 1980s. Concerns about idleness, nonmarital childbearing, crime, substance abuse, and mental health among adolescents and young adults increased during this period as social scientists, politicians, and the public discovered new arenas in which to make adolescence more problematic. It is not clear from available data whether actual behavioral trends reflect the growing public concern about problem behavior. From all available evidence, it appears that few linear trends in problem behavior can be detected from data sources. Certain manifestations of problems, such as delinquency or suicide, have increased and then fallen off. Moreover, these trends rarely are reproduced in all Western nations during the same time or in the same form. Therefore, it seems highly unlikely that there has been a steady increase in problem behaviors or a decrease in prosocial behavior in any country, much less throughout the West.

Public concern about youth, on the other hand, does appear to be steadily rising, at least judging from data in the USA. There is a growing perception that children and adolescents are being less well cared for and accordingly exhibiting more problems entering adulthood. Again, it is not at all evident that youths are any less committed to mainstream notions of success, or any less capable of or prepared for achieving adulthood than they were in earlier times. It is obvious, however, that the entrance to adulthood is coming later than ever before, if we mean by that the conjunction of leaving the parental household, entering the labor force, and establishing a family. In many countries, these status transitions are occurring later than they did a century ago and much later than at the middle of the twentieth century. Sociologically, adolescence has been extended well into the twenties, or it could be said that we are witnessing the reemergence of the period of semiautonomy that was common a century or more ago.

This has important economic, social, and psychological implications for the development of adolescents and young adults. Youths are more dependent on elders for economic support while simultaneous claims in support of the aging baby boom cohorts are being made. The extension of life, along with the unusually large size of the population at midlife and beyond, is straining both public and private resources and is likely to increase in the future. The possibility of generational conflict over resources is also bound to increase as societies face the difficulty of choosing to invest in increasing youth productivity or allocating resources to the care of the elderly.

Socially, the lengthening period of dependency has consequences for how much youths are willing or permitted to invest in the larger society. Youths not integrated, incorporated, or involved may lead to political and social alienation. The role of youth is focused more on consumption than production. Participation of young adults in society may shift to playing a symbolic role in the media rather than actually performing in arenas of social power. Thus, young people are visible (indeed overrepresented) in media portrayals and underrepresented in the labor force or in political bodies. This delay could have consequences on social commitments such as forming lasting bonds with a partner, having children, or engaging in civic activities such as voting or political activity.

Finally, the nature and quality of psychological maturation must change as individuals linger in the state of semiautonomy for longer periods. With youths either dependent on their families or society for social support, formation of selfhood becomes prolonged. One might expect adolescents and young adults to take on a more fluid and impermanent sense of identity. This has advantages for a job world in which flexibility and mobility are required, but it is likely to have some fallout for establishing relationships. Cohabitation permits and promotes flux in interpersonal commitments as individuals resist settling down until they are sure of who they are. But in withholding commitments, it becomes increasingly difficult to resolve the problem of identity. Thus, identity becomes a lifelong project rather than a stage of development that is more or less established by the entrance to adulthood.

If the end of adolescence has been delayed, its beginning has gradually moved up into the preteen years. The earlier onset of adolescence seems paradoxical, especially as most Western nations have given greater importance in the twentieth century to protecting children from the harsh realities of the workplace, sexual exploitation, and abuse from families. Viviana Zelizer describes this as the 'priceless child' phenomenon—the movement of children from precocious involvement in economic roles to a greater emphasis on emotional development and psychological investment by parents (Zelizer 1985).

Yet adolescence itself has become earlier, judging from the behavior of younger teens and preteens. The onset of sexual behavior occurs several years earlier on average than it did in the 1950s, when most teenagers waited until after or just before getting married to initiate coitus. There are several ways of explaining this downward trend: earlier ages of puberty owing to better nutrition and higher living standards; later age at marriage, making it more difficult for young people to delay; lower social control, as teens are less subject to parental monitoring; availability of contraception, making it possible to prevent some of the untoward consequences of sexual experimentation; and greater tolerance for premarital sex, as chastity is less prized.

Sex is just one indicator of an earlier adolescence. Dress, demeanor, and media consumption are other signs that preteens have been given over to expressions of adolescence. The tastes of preteens may be different from older teens, but they are surely even more

different from their parents. How has this come about even when parents are increasingly concerned about the corruption of their children's tastes and sensibilities?

Obviously, the commercialization of the media has played a part in cultivating younger audiences into market groups. The number of media outlets has grown, along with the marketing skills to identify and reinforce tastes and practices. The rise of computer literacy, no doubt, contributes to the growing sophistication of children in their preteens about features of the social world that were previously inaccessible to them. Another source of influence may be the schools and related agencies, such as the health care system and juvenile justice system, which have increasingly classified preteenagers into institutions for adolescents. In the USA, the growth of the middle school provides just such a case of grouping preteens with younger teenagers. This may well have fostered an earlier development of 'adolescence,' by which children sense that they should behave more autonomously. Finally, parents are ambivalent about these forces that have created an earlier adolescence for their children. While protective of their children's innocence, they are sensitive to the cultural cues that promote earlier development and generally are unprepared to resist the forces outside the family that foster early development. Indeed, many parents are encouraging of these forces because they do not want to see their children left behind. Earlier adolescence might be thought of as partly emanating from biology and partly from social systems governed by age-graded cultural and social norms that are malleable and adaptive to current conditions.

Not enough attention has been devoted to how such norms are influenced by public policies and private responses. The ways that adolescence interprets and responds to social and cultural signals is a frontier area for future researchers. How policies at the family, school, community, and societal level are instantiated in everyday practice is a promising topic for further research. Similarly, we need to examine how developmental processes themselves affect the reading of these social cues. For example, the ways in which younger and older adolescents interpret legal and social sanctions is a topic of great importance, especially in the USA, where legislators are putting into practice criminal statutes affecting increasingly younger children. Similarly, age of consent, labor laws regulating youth employment, political participation, and a number of other rules governing the timing of adultlike activities rest on developmental assumptions that have not been widely investigated.

See also: Adolescent Behavior: Demographic; Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Adolescent Vulnerability and Psychological Interventions; Adolescent Work and Unemployment; Adolescents: Leisure-time Activities; Antisocial Behavior in Childhood and Adolescence; Cognitive Development in Childhood and Adolescence; Counterculture; Delinquency, Sociology of; Gender-related Development; Generations, Sociology of; Hall, Granville Stanley (1844–1924); Identity in Childhood and Adolescence; Life Course: Sociological Aspects; Socialization in Adolescence; Teen Sexuality; Tolerance; Xenophobia; Youth Culture, Sociology of; Youth Movements

Bibliography

Bachman J G, O'Malley P M, Johnston J 1978 *Adolescence To Adulthood: Change and Stability in the Lives of Young Men*. Institute for Social Research, University of Michigan, Ann Arbor, MI

Carnegie Council on Adolescent Development 1995 *Great Transitions: Preparing Adolescents for a New Century*. Carnegie, New York.

Coleman J S 1974 *Youth: Transition to Adulthood*. Report on the Panel on Youth of the President's Science Advisory Commitee. University of Chicago Press, Chicago

Condran G A, Furstenberg F F 1994 Are trends in the well-being of children related to changes in the American family? Making a simple question more complex. *Population* **6**: 1613–38

Flacks R 1971 *Youth and Social Change*. Markham, Chicago

Friedenberg E Z 1964 *The Vanishing Adolescent*. Beacon Books, Boston

Furstenberg F F 2000 The sociology of adolescence and youth in the 1990s: A critical commentary. *Journal of Marriage and the Family* (in press)

Halls G S 1904 *Adolescence: Its Pyschology and its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion and Education*. D Appleton, New York

Jensor R, Colby A, Schueder R A 1996 *Ethnography and Human Development: Context and Meaning in Social Inquiry*. University of Chicago Press, Chicago

Kett J F 1977 *Rites of Passage: Adolescence in America, 1790 To the Present*. Basic Books, New York

Modell J 1989 Into One's Own: From Youth to Adulthood in the United States, 1920–1975. University of California Press, Berkeley

Moen P, Elder Jr G H, Lüscher K (eds.) 1995 *Examining Lives in Context*. American Psychological Association, Washington, DC

Morgan S P, McDaniel A, Miller A T, Preston S H 1993 Racial differences in household and family structure at the turn of the century. *American Journal of Sociology* **98**(4): 799–828

National Commission on Children 1991 *Beyond Rhetoric: A New American Agenda for Children and Families* US Government Printing Office, Washington, DC

Saporiti A, Sgritta G B 1990 *Childhood as a Social Phenomenon: National Report*. European Center, Vienna

Shorter E, Knodel J, Van De Waller E 1971 The decline of nonmarital fertility in Europe, 1880–1940. *Population Studies* **25**(3): 375–93

Zelizer V A 1985 *Pricing the Priceless Child: The Changing Social Value of Children*. Basic Books, New York

F. F. Furstenberg

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Adolescent Behavior: Demographic

1. From Standard Biography to Choice Biography

During young adulthood, young men and women are confronted with various life transitions and have to make decisions about their future. How long will they continue in full-time education, when will they look for a job, or will they combine work with schooling? Will they seek a partner, or choose to remain single? What are their attitudes towards starting a family of their own? This period in life is generally regarded as a first step towards adulthood in that it incorporates a move from dependence towards independence, in both financial and emotional terms as well as in terms of a young adult's social life. As such, it is an important life-course phase because each transition changes and determines the young adult's position within society.

Each society is characterized by a different set of normative life-course models which vary according to the gender and social background of the individuals living within that society. Individual life courses tend to mirror these socially established patterns to a certain extent, despite the fact that these patterns often include a considerable margin for individual choice. Some of the status passages are considered as desirable, others are considered risky and undesirable (Levy 1997). At the beginning of the twentieth century, many young adults had a limited range of behavioral options. For example, most young men and women left the parental home to marry. A smaller proportion left the parental home to take up employment. The timing of home leaving varied much more widely than it does today (Liefbroer and de Jong Gierveld 1995) and was influenced by parental encouragement or discouragement and prevailing family obligations and employment opportunities. Some adult children remained in the parental home, either because they were needed as caregivers or because they were designated as heirs to the family's land and property. The type of household these young adults started and the potential level of their household income were inextricably linked with the options that their paternal background provided. A young man born into a family of servants stood a strong chance of becoming a servant himself. Compared with their peers who lived at the beginning of the twentieth century, young adults born in the second half of the twentieth century had a much broader range of options and considerable freedom to choose the pattern and timing of their life transitions.

Sociologists summarize these developments by the opposing concepts of 'standard biography' (leaving home, followed by marriage, followed by childbirth) and 'choice biography' (the sequencing of transitions based on personal choice: e.g., leaving home, living alone, returning home, leaving home a second time, unmarried cohabitation, and marriage). Personal choice has become more or less obligatory (Giddens 1994).

2. Leaving the Parental Home; Changing Patterns of Determinants

Leaving home, the first of the young adult's transitions, is to be considered as a migratory movement triggered by other parallel life course careers (Mulder 1993). The most important parallel careers are union formation (marriage or unmarried cohabitation), higher education, and a change of work. An examination of home-leaving trends and determinants must therefore take account of young adults' preferences and behavioral patterns as regards these parallel careers.

The second half of the twentieth century witnessed profound sociostructural and cultural changes that influenced the parallel careers and home-leaving behavior of young adults. These changes included an improvement in the standard of living, a rise in educational levels and female labor force participation, the increasing equality of men and women, a decline in traditional and religious authority—including a slackening of parental supervision over the behavior of young adults—and the diffusion of individualization (Buchmann 1989, Lesthaeghe and Surkyn 1988, Liefbroer 1999, Van de Kaa 1987). As a result of these developments, young men and women increasingly prefer to have a period in life characterized by independence and the postponement of decisions that entail strong commitments (Mulder and Manting 1994). Union formation in general and marriage in particular, as well as parenthood, are decisions that are frequently postponed.

3. Leaving the Parental Home; Facts and Figures

Table 1 provides data on home leaving in selected European countries and in the US. The data were taken from the Family and Fertility Survey conducted in many countries under the auspices of the UN Economic Commission for Europe. The surveys used the same questionnaire modules in order to obtain the best possible comparability. The data used in this contribution relate to two birth cohorts: one from the early 1950s and one from the early 1960s. Data on later birth cohorts are incomplete as far as home leaving is concerned.

3.1 Leaving the Parental Home for the Purpose of Union Formation

Table 1 provides information about the reasons for leaving the parental home. In Spain, Italy, Portugal,

Table 1
Some characteristics of leaving home of young adults aged 15 years and over, for selected countries in Europe and the US

		Leaving home to marry[a] b.c. 1950–4	Leaving home to marry[a] b.c. 1960–4	Leaving home, union formation[a] b.c. 1950–4	Leaving home, union formation[a] b.c. 1960–4	Median age[b] at leaving home b.c. 1950–4	Median age[b] at leaving home b.c. 1960–4
Northern/Western Europe							
Sweden	F	6.7 (99 percent)	4.4 (97 percent)	35.3 (99 percent)	33.3 (97 percent)	21.0	21.1
FFS 1992/3	M	2.4 (98 percent)	1.7 (99 percent)	21.4 (98 percent)	25.4 (99 percent)	23.9	23.5
Finland	F	23.7 (99 percent)	8.8 (97 percent)	40.9 (99 percent)	45.2 (97 percent)	19.4	19.8
FFS 1989/90	M	8.6 (89 percent)	3.3 (85 percent)	29.2 (89 percent)	37.0 (85 percent)	20.7	22.0
Norway	F	7.2 (98 percent)	3.7 (98 percent)	10.3 (98 percent)	10.3 (98 percent)	19.5	19.2
FFS 1988/9	M	(.)[c]	2.3 (91 percent)	(.)[c]	7.9 (91 percent)	(.)[c]	21.0
France	F	55.9 (98 percent)	31.5 (96 percent)	63.2 (98 percent)	61.9 (96 percent)	20.5	20.0
FFS 1994	M	39.8 (95 percent)	14.6 (92 percent)	50.5 (95 percent)	50.0 (92 percent)	21.6	21.8
Germany	F	48.9 (98 percent)	29.8 (97 percent)	63.8 (98 percent)	58.1 (97 percent)	20.0	20.6
FFS 1991/2	M	30.8 (96 percent)	21.0 (91 percent)	43.1 (96 percent)	46.8 (91 percent)	21.8	22.3
Belgium/Flanders	F	84.0 (99 percent)	71.0 (95 percent)	88.8 (99 percent)	83.5 (95 percent)	21.1	21.5
FFS 1991/2	M	81.4 (94 percent)	64.6 (86 percent)	88.4 (94 percent)	82.1 (86 percent)	22.5	23.5
Netherlands	F	54.1 (98 percent)	26.7 (95 percent)	61.0 (98 percent)	55.8 (95 percent)	19.5	19.6
FFS 1993	M	46.0 (99 percent)	17.6 (97 percent)	58.3 (99 percent)	49.5 (94 percent)	21.4	21.8
Southern Europe							
Spain	F	81.7 (91 percent)	77.1 (89 percent)	83.1 (91 percent)	82.1 (89 percent)	23.1	23.3
FFS 1994/5	M	70.9 (84 percent)	59.7 (80 percent)	72.9 (84 percent)	70.7 (80 percent)	24.8	25.9
Italy	F	84.6 (89 percent)	78.4 (86 percent)	86.9 (89 percent)	83.8 (86 percent)	22.5	23.8
FFS 1995/6	M	73.6 (86 percent)	61.7 (75 percent)	75.6 (86 percent)	68.6 (75 percent)	25.6	27.5
Portugal	F	63.4 (89 percent)	61.4 (83 percent)	70.4 (89 percent)	70.8 (83 percent)	21.8	22.4
FFS 1997	M	57.2 (87 percent)	52.2 (77 percent)	61.9 (87 percent)	62.4 (77 percent)	23.7	24.9
Central Europe							
Hungary	F	62.1 (87 percent)	57.7 (83 percent)	65.8 (87 percent)	65.3 (83 percent)	21.4	21.3
FFs 1995/6	M	60.3 (83 percent)	53.5 (74 percent)	65.9 (83 percent)	67.0 (74 percent)	24.6	24.9
Czech Republic	F	70.2 (93 percent)	62.6 (84 percent)	79.8 (93 percent)	71.9 (84 percent)	18.9	18.8
FFS 1997	M	61.2 (95 percent)	57.5 (84 percent)	70.9 (95 percent)	71.7 (84 percent)	20.2	20.1
Slovenia	F	39.0 (89 percent)	33.7 (90 percent)	44.6 (89 percent)	51.8 (90 percent)	20.9	20.8
FFS 1994/5	M	28.3 (90 percent)	15.8 (86 percent)	33.9 (90 percent)	30.8 (86 percent)	21.2	20.8
Poland	F	59.8 (87 percent)	64.5 (75 percent)	61.2 (87 percent)	66.1 (75 percent)	22.3	22.5
FFS 1991	M	60.6 (81 percent)	65.0 (60 percent)	62.0 (81 percent)	66.3 (60 percent)	24.4	26.0
Latvia	F	26.6 (81 percent)	34.0 (78 percent)	30.8 (81 percent)	40.6 (78 percent)	21.4	21.5
FFS 1995	M	26.3 (82 percent)	31.5 (68 percent)	31.0 (82 percent)	41.0 (68 percent)	22.9	24.8
Lithuania	F	21.1 (84 percent)	24.0 (77 percent)	22.0 (84 percent)	25.0 (77 percent)	19.2	20.3
FFS 1994/5	M	14.0 (84 percent)	24.8 (81 percent)	14.9 (84 percent)	27.6 (81 percent)	18.9	20.5
US	F	37.0 (97 percent)	25.7 (96 percent)	40.0 (97 percent)	35.8 (96 percent)	18.7	18.8
FFS 1995	M	(.)[c]	(.)[c]	(.)[c]	(.)[c]	(.)[c]	(.)[c]

Source: Family and Fertility Surveys (FFS) (data analysis by Edith Dourleijn, NIDI). a In percentages of leavers; start marriage (or union formation, respectively) within ± 3 months of date of leaving home. Within brackets the percentages of young adults who had left home by the time of the FFS interview. b Median age in years. c No data available.

Belgium (Flanders), Poland, the Czech Republic, and Hungary, marriage was by far the most important reason why birth cohorts of the early 1950s left home. For cohorts born in the early 1960s, leaving home to marry was still the most dominant reason in Spain, Italy, Belgium (Flanders), Poland, and the Czech Republic, as well as, to a lesser extent, Portugal and Hungary. However, the percentage of home leavers born in the early 1960s in these countries who went on to marry decreased considerably compared with those born in the early 1950s. Marriage as a motive for leaving home decreased in *all* of the selected countries in northern, western, and southern Europe and in the US. The pattern for the Eastern European countries was more diverse. For example, the pattern in Hungary, the Czech Republic, and Slovenia, which were characterized by a decrease in marriage as a reason for leaving home, differed from that in Poland, Latvia, and Lithuania.

The decrease in marriage was coupled with an increase in unmarried cohabitation as a reason for leaving home in several countries, as can be seen by comparing columns in Table 1. This was very apparent in Finland and France, where more than 30 percent of the 1960–4 birth cohort combined leaving home with embarking on unmarried cohabitation. A comparison of the cohorts born in the early 1950s and early 1960s reveals that unmarried cohabitation cannot compensate in all countries for the declining importance of marriage. Union formation in general was a less important reason for leaving home for the birth cohorts of the early 1960s than those of the early 1950s in France, Belgium (Flanders), the Netherlands, Italy, Spain, and the US. In the northern European countries, however, the percentage of young adults leaving home for union formation was already low in the early 1950s, and continued to be low in the early 1960s. In some Central European countries leaving home for union formation increased. So, differences between regions mirror the ideas of the second demographic transition.

3.2 Leaving the Parental Home for Other Reasons

The percentage of young adults leaving the parental home to pursue postsecondary education in conjunction with living alone is increasing all over Europe. This in turn has led to a lowering of the home-leaving age.

Nowadays, leaving home to achieve personal freedom and independence—another current trend—is more apparent among young home leavers who do not wish to pursue postsecondary education. Most of them are (wholly or partly) financially independent due to their participation in the labor market. Young adults with a job leave home at a younger age than the unemployed (Cordón 1997).

Today, as in the past, some young adults leave the parental home at a fairly young age in order to extricate themselves from a difficult domestic situation, such as an unstable family structure (a stepfamily or one-parent family), or after the death of a parent (Goldscheider and Goldscheider 1998). Other 'negative' reasons for leaving home include friction with parents due to the parents' low income, or the unemployment of the child (Berrington and Murphy 1994).

Baanders (1998) emphasizes that the normative pressure exerted by the parents in terms of encouraging or discouraging young adults to leave the parental home is an important factor behind young adults' intentions, even among recent cohorts. Whether or not young adults are able to realize their preferences and intentions to leave will depend partly on their own resources (a job, income, or a partner able to provide these prerequisites), but also on the resources provided by the parents. In this context Goldscheider and DaVanzo (1989) distinguished between parents' location-specific resources (the number of siblings and the time the mother spends at home), and parents' transferable resources (income). De Jong Gierveld et al. (1991) expanded on this conceptual framework by distinguishing between the material and nonmaterial resources of the parents. The cross-tabulation of these two dimensions gives rise to four types of parental support that encourage young adults either to leave or to prolong their stay in the parental home.

3.3 Timing of Home Leaving

The reasons that trigger home leaving affect the timing of leaving home in various, and sometimes conflicting, ways. Mulder (1993) examined the trend towards postponement of home leaving for the purpose of union formation, concluding that there was an 'individualization effect' behind this trend accompanied by an additional (temporary) effect caused by unfavorable economic conditions that made it difficult to find affordable housing. The trend towards increasing participation in post-secondary education is proving to be a catalyst for young adults to leave home at relatively early ages. Leaving home to seek independence and live alone is becoming characteristic of young adults in more and more western and northern European countries, and this, too, is decreasing the age at home leaving.

The last two columns of Table 1 provide information about the timing of leaving home. In most of the northern and western European countries, the median age at leaving the parental home for cohorts born in the early 1950s and early 1960s is between 19 and 22 for young women, and between 20 and 24 for young men. Table 1 also indicates that in some European countries the end-result of both trends (postponing and bringing forward home leaving) is a stabilization of the median ages at home leaving for

the birth cohorts of the early 1950s and early 1960s (Mulder and Hooimeijer 1999).

A comparison of the cohorts born in the 1950s and 1960s shows that median ages at home leaving in Poland, Latvia, and Lithuania increased. More research is needed to gain insight into the mechanisms behind this trend.

In the southern European countries, median ages at leaving home were 22.5 years and over for young women, and ranged between 24.8 and 27.5 years for young men. In the southern European countries the overall trend was towards a rise in the age at home leaving, particularly among men. This trend is related to a persistence of traditional patterns, a fact which is borne out by the virtual absence of living alone and unmarried cohabitation among young adults. This southern European situation was analyzed in-depth in a special issue of the journal *Family Issues* (1997, pp. 18, 6), which stated that young adults in Italy and Spain prefer to lengthen the period of young adulthood through longer co-residence in the parental home. Some authors describe this period as a 'chronological stretching' of young adulthood (Rossi 1997).

4. *Successful Home Leaving and Returning?*

The growing acceptance of a reversal of former lifestyle decisions is one of the current life-course patterns. Returning home after leaving the parental home is one of these. Reasons to return include separation and divorce. The percentage of young adults who return to the parental home after leaving is high among those who left home at very young ages, and who left to start a non-family household (White and Lacy 1997). According to Zinnecker et al. (1996), about one in five young adults in the former West Germany—aged between 23 and 29—who currently live in the parental home can be characterized as returnees. Some authors attribute this to the disorderliness of the transitional period of young adulthood. Others contend that, by trying so hard to effect their own personal choices, young adults often end up making decisions that resemble decisions made by other young adults.

See also: Adolescent Development, Theories of; Adolescent Vulnerability and Psychological Interventions; Adolescent Work and Unemployment; Adolescents: Leisure-time Activities; Life Course in History; Life Course: Sociological Aspects; Puberty, Psychosocial Correlates of; Teen Sexuality; Teenage Fertility

Bibliography

Baanders A N 1998 Leavers, planners and dwellers; the decision to leave the parental home. Thesis, Agricultural University, Wageningen

Berrington A, Murphy M 1994 Changes in the living arrangements of young adults in Britain during the 1980s. *European Sociological Review* **10**: 235–57

Buchmann M 1989 *The Script of Life in Modern Society: Entry into Adulthood in a Changing World*. University of Chicago Press, Chicago

Cordón J A F 1997 Youth residential independence and autonomy. *Journal Family Issues* **18**: 6: 576–607

De Jong Gierveld J, Liefbroer A C, Beekink E 1991 The effect of parental resources on patterns of leaving home among young adults in the Netherlands. *European Sociological Review* **7**: 55–71

Giddens A 1994 Living in a post-traditional society. In: Beck U, Giddens A, Lash S (eds.) *Reflexive Modernization; Politics, Tradition and Aesthetics in the Modern Social Order*. Polity Press/Blackwell, Cambridge, UK, pp. 56–109

Goldscheider F K, DaVanzo J 1989 Pathways to independent living in early adulthood: marriage, semiautonomy and premarital residential independence. *Demography* **26**: 597–614

Goldscheider F K, Goldscheider C 1998 The effects of childhood family structure on leaving and returning home. *Journal of Marriage and the Family* **60**: 745–56

Lesthaeghe R, Surkyn J 1988 Cultural dynamics and economic theories of fertility change. *Population and Development Review* **14**: 1–45

Levy R 1997 Status passages as critical life-course transitions; a theoretical sketch. In: Heinz W R (ed.) *Theoretical Advances in Life Course Research*. Deutscher Studien Verlag, Weinheim, Germany, pp. 74–86

Liefbroer A C 1999 From youth to adulthood: Understanding changing patterns of family formation from a life course perspective. In: Van Wissen L J G, Dykstra P A (eds.) *Population Issues, an Interdisciplinary Focus*. Kluwer Academic, Dordrecht, The Netherlands, pp. 53–85

Liefbroer A C, de Jong Gierveld J 1995 Standardisation and individualization: the transition from youth to adulthood among cohorts born between 1903 and 1965. In: Van den Brekel J C, Deven F (eds.) *Population and Family in the Low Countries 1994*. Kluwer Academic, Dordrecht, The Netherlands, pp. 57–79

Mulder C H 1993 *Migration Dynamics: A Life Course Approach*. Thesis Publishers, Amsterdam

Mulder C H, Manting D 1994 Strategies of nest-leavers: 'settling down' versus flexibility. *European Sociological Review* **10**: 155–72

Mulder C H, Hooimeijer P 1999 Residential relocations in the life course. In: Van Wissen L J G, Dykstra P A (eds.) *Population Issues, an Interdisciplinary Focus*. Kluwer Academic, Dordrecht, The Netherlands, pp. 159–86

Rossi G 1997 The nestlings – why young adults stay at home longer: The Italian case Journals. *Family Issues* **18**: 6627–44

Van de Kaa D J 1987 Europe's second demographic transition. *Population Bulletin* **42**(-1): 1–57

Whyte L, Lacy N 1997 The effects of age at home leaving and pathways from home on educational attainment. *Journal of Marriage and the Family* **59**: 982–95

Zinnecker J, Strozda C, Georg W 1996 Familiengründer, Postadoleszente und Nesthocker – eine empirische Typologie zu Wohnformen junger Erwachsener. In: Buba H P, Schneider N F (eds.) *Familie: zwischen gesellschaftlicher Prägung und individuellem Design*. Westdeutscher Verlag, Opladen, Germany, pp. 289–306

J. de Jong Gierveld

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Adolescent Development, Theories of

Adolescence, the second decade of the human life cycle, is a transitional period that bridges childhood and adulthood. Because the nature of the transition is multifaceted, writers interested in adolescence have, over the years, addressed many different aspects of development during this period, including biological development, cognitive development, emotional development, and social development. The purpose of this article is to provide brief summaries of the major theoretical viewpoints and highlight the central contributions each has made to understanding development during the adolescent decade.

1. *Views of Adolescence in History*

Although scientific theorizing about adolescence did not appear until the beginning of the twentieth century, philosophers and educators have written about this period of development for centuries. Early writings on the period pointed to the youthful energy and vitality of the teenage years, and often depicted adolescents as both enthusiastic and impulsive. From the beginning, adolescence has been portrayed as a period of potential difficulty, either for the young person, who was presumed to have difficulty coping with the challenges inherent in the transition to adulthood, or for adults, who were presumed to have difficulty in controlling and reining in the adolescent's energy and impulses. This notion—that adolescence is a potentially difficult period for adolescents and for those around them—is a recurrent theme throughout most theoretical writings on the period. Accordingly, much of what has been written about adolescence has a strong problem orientation, with theorists either attempting to explain why adolescence is as difficult as it is or offering accounts about how the period might be made less stressful and more pacific. Although this problem orientation is less pervasive today than was the case earlier in the twentieth century, the focus on adolescence as a time of difficulty persists in contemporary writings on this stage of development.

The multifaceted nature of adolescence—the fact that it has biological, psychological, and social components—has made the period the focus of attention to theorists from many different disciplines, including biology, psychology, sociology, history, and anthropology. Not surprisingly, theorists writing from different vantage points bring a variety of emphases to their discussion of the adolescent transition. Biologists and psychologists have emphasized the changing physical, intellectual, emotional, and social capabilities of the individual and have asked whether and in what ways individual functioning during adolescence differs from functioning during childhood or adulthood. Sociologists, anthropologists, and historians, in contrast, have focused on the transition of individuals into adult status and have posed questions about the nature of these changes in roles, rights, and responsibilities.

2. *Biological and Psychological Theories of Adolescence*

2.1 *G. Stanley Hall's Theory of Recapitulation*

Theorists who have taken a biological view of adolescence stress the hormonal and physical changes of puberty as driving forces that define the nature of the period. The most important theorist in this tradition was G. Stanley Hall (1904), considered the founder of the scientific study of adolescence. Hall, who was very much influenced by the work of Charles Darwin, the author of the theory of evolution, believed that the development of the individual paralleled the development of the human species, a notion referred to as his *theory of recapitulation*. Infancy, in his view, was equivalent to the time during human evolution when we were primitive, like animals. Adolescence, in contrast, was seen as a time that paralleled the evolution of our species into civilization. For Hall, the development of the individual through these stages was determined primarily by biological and genetic forces within the person, and hardly influenced by the environment.

The most important legacy of Hall's view of adolescence is the notion that adolescence is inevitably a period of *storm and stress*. He believed that the hormonal changes of puberty cause upheaval, both for the individual and for those around the young person. Because this turbulence was biologically determined, in Hall's view, it was unavoidable. The best that society could do was to find ways of managing the young person whose 'raging hormones' would invariably lead to difficulties. This, he believed, was analogous to the taming of the human species that occurred as civilization evolved.

Although most scientists no longer believe that adolescence is an inherently stressful period, the contemporary study of adolescence continues to emphasize the role that biological factors—hormonal changes, somatic changes, or changes in reproductive maturity—play in shaping the adolescent experience. Indeed, the study of the impact of puberty on adolescent psychosocial development has been, and continues to be, a central question for the field, with some theorists emphasizing the direct and immediate impact of puberty on adolescent psychological functioning, and others focusing on the timing of the adolescent's maturation relative to that of his or her peers.

2.2 Psychoanalytic Theories

In psychoanalytic theory, as in Hall's theory of recapitulation, adolescence is seen as an inherent time of upheaval triggered by the inevitable changes of puberty. According to Freud, the hormonal changes of puberty upset the psychic balance that had been achieved during the prior psychosexual stage, *latency*. Because the hormonal changes of puberty are responsible for marked increases in sexual drive, the adolescent was thought to be temporarily thrown into a period of intrapsychic crisis, and old psychosexual conflicts, long buried in the unconscious, were revived. Freud and his followers believed that the main challenge of adolescence was to restore a psychic balance and resolve these conflicts. Working through these conflicts was necessary, Freud believed, in order for the individual to move into what he described as the final and most mature stage of psychosexual development—the *genital stage*. It was not until this stage of development that individuals were capable of mature sexual relationships with romantic partners.

Freud's daughter, Anna Freud (1958), extended much of her father's thinking to the study of development during the second decade of life. Her most important work, entitled *Adolescence*, continued the tradition begun by Hall in casting adolescence as a time of unavoidable conflict and both intrapsychic and familial turmoil. According to her view, the revivification of early psychosexual conflicts, caused by the hormonal changes of puberty, motivated the adolescent to sever emotional ties to his or her parents and turn to peers as objects of sexual desire and emotional affection. She described adolescence as a period of 'normative disturbance' and argued that the oppositionalism and defiance many parents encountered in their teenagers was not only normal, but desirable. Indeed, Anna Freud believed that adolescents needed to break away from their parents in order to develop into healthy and mature adults, a process known as detachment.

Over time, psychoanalytic theories of adolescence came to place less emphasis on the process of detachment or the motivating role of puberty, and began to emphasize the psychological capacities that developed as the adolescent negotiated a path toward independence and adult maturity. Psychoanalytic theorists of adolescence in the second half of the twentieth century turned their attention away from the analysis of drives and focused instead on the skills and capabilities individuals developed in order to resolve inner conflicts and establish and maintain mature relationships with others, especially others outside the family. The three most important writers in this neoanalytic tradition are Peter Blos (1979), whose theory of adolescent development emphasizes the growth of emotional autonomy from parents, a process called individuation; Harry Stack Sullivan (1953), whose view of adolescence revolves around the young person's growing need and capacity for intimate, sexual relationships with peers; and Erik Erikson (1968), who focused on the adolescent's quest for a sense of identity. By far the most important of the three theorists has been Erik Erikson.

Erikson's theory of the life cycle proposed eight stages in psychosocial development, each characterized by a specific 'crisis' that arose at that point in development because of the interplay between the internal forces of biology and the unique demands of society. According to him, adolescence revolves around the crisis of *identity vs. identity diffusion*. The challenge of adolescence is to resolve the identity crisis successfully and to emerge from the period with a coherent sense of who one is and where one is headed. In order to do this, the adolescent needs time to experiment with different roles and personalities. Erikson believed that adolescents needed a period of time during which they were free from excessive responsibility—a psychosocial moratorium, as he described it—in order to develop a strong sense of identity. This vision of adolescence as a period during which individuals 'find themselves' through exploration and experimentation has been a longstanding theme in portrayals of adolescence in literature, film, and television. Indeed, Erikson's notion of the identity crisis is one of the most enduring ideas in the social sciences.

2.3 Cognitive-developmental Theories

In contrast to psychoanalytic and neoanalytic theorists, who emphasized emotional and social development during adolescence, cognitive-developmental theorists characterized adolescence in terms of the growth of intellectual capabilities. The most influential theorist in this regard has been Jean Piaget (Inhelder and Piaget 1958), whose theory of cognitive development dominated the study of intellectual growth—not only during adolescence, but during infancy and childhood as well—for most of the latter half of the twentieth century.

Piaget believed that, as individuals mature from infancy through adolescence, they pass through four stages of cognitive development, and that each stage is characterized by a type of thinking that is qualitatively distinct from that which characterized intellectual functioning in other stages. In Piaget's theory, adolescence marks the transition from the stage of concrete operations, during which logical reasoning is limited to what individuals can experience concretely, to the stage of *formal operations*, during which logical reasoning can be applied to both concrete and abstract phenomena. According to this theory, adolescence is the period during which individuals become fully capable of thinking in abstract and hypothetical terms, an achievement which engenders and permits a variety of new intellectual and social pursuits. The development of formal operational thinking in adolescence

has been posited to undergird adolescents' ability to grasp such diverse phenomena as algebra, scientific hypothesis testing, existential philosophy, satire, and principles of human motivation.

One well-known application of Piaget's theory to adolescent development is found in the work of Lawrence Kohlberg (1969). Like Piaget, Kohlberg believed that reasoning during adolescence is qualitatively different from reasoning during childhood. More specifically, Kohlberg believed that adolescents were capable of viewing moral problems in terms of underlying moral principles, like fairness or equity, instead of limiting their moral thinking to concrete rules and regulations. Similar applications of Piaget's work can be found in theories of adolescent decision-making, political thinking, interpersonal relationships, religious beliefs, and identity development.

3. *Sociological, Historical, and Anthropological Theories*

The emphasis within most biological and psychological theories of adolescence is mainly on forces within the individual, or within the individual's unique environment, in shaping his or her development and behavior. In contrast, sociological, historical, and anthropological theories of adolescence attempt to understand how adolescents, as a group, come of age in society and how coming of age varies across historical epochs and cultures.

3.1 *Sociological Theories*

Sociological theories of adolescence have often focused on relations between the generations and have tended to emphasize problems that young people sometimes have in making the transition from adolescence into adulthood, especially in industrialized society. Two themes have dominated these discussions. One theme, concerning the marginality of young people, emphasizes the difference in power that exists between the adult and the adolescent generations. Two important thinkers in this vein are Kurt Lewin (1951) and Edgar Friedenberg (1959), both of whom stressed the fact that adolescents were treated as 'second class citizens.' Contemporary applications of this viewpoint stress the fact that many adolescents are prohibited from occupying meaningful roles in society and therefore experience frustration, restlessness, and difficulty in making the transition into adult roles.

The other theme in sociological theories of adolescence concerns *intergenerational conflict*, or as it is more commonly known, 'the generation gap.' Theorists such as Karl Mannheim (1952) and James Coleman (1961) have focused not so much on the power differential between adults and adolescent, but the fact that adolescents and adults grow up under different social circumstances and therefore develop different sets of attitudes, values, and beliefs. This phenomenon is exacerbated by the pervasive use of *age-grading*—the separation of individuals on the basis of chronological age—within our social institutions, particularly schools. As a consequence of this age-segregation, there is inevitable tension between the adolescent and the adult generations. Some writers, like Coleman, have gone so far as to argue that adolescents develop a different cultural viewpoint—a 'counterculture'—that may be hostile to the values or beliefs of adult society.

Although sociological theories of adolescence clearly place emphasis on the broader context in which adolescents come of age, rather than on the biological events that define adolescence, there is still a theme of inevitability that runs through their approach. Mannheim, for example, believed that because modern society changes so rapidly, there will always be problems between generations because each cohort comes into adulthood with different experiences and beliefs. Similarly, Lewin believed that marginality is an inherent feature of adolescence because adults always control more resources and have more power than young people.

3.2 *Historical and Anthropological Theories*

Historians and anthropologists who study adolescence share with sociologists an interest in the broader context in which young people come of age, but they take a much more relativistic stance. Historical perspectives, such as those offered by Glen Elder (1974) or Joseph Kett (1977), stress the fact that adolescence as a developmental period has varied considerably from one historical era to another. As a consequence, it is impossible to generalize about such issues as the degree to which adolescence is stressful, the developmental tasks of the period, or the nature of intergenerational relations. Historians would say that these issues all depend on the social, political, and economic forces present at a given time. These forces may result in very different adolescent experiences for individuals who are members of different *cohorts*, or groups of people who come of age at a similar point in historical time. Even something as central to psychological theories of adolescence as Erikson's notion of the adolescent 'identity crisis,' historians argue, is a social invention that arose because of industrialization and the prolongation of schooling. They suggest that before the industrial revolution, when most adolescents followed in their parents' occupation, crises over identity did not exist, or were a privilege of the extremely affluent.

One group of theorists has taken this viewpoint to its logical extreme. These theorists, called inventionists, argue that adolescence is entirely a social invention (Bakan 1972). This position is in stark contrast to that adopted by the biological and psychological theorists discussed earlier, who view adolescence as a

biologically determined reality. Inventionists believe that the way in which we divide the life cycle into stages—drawing a boundary between childhood and adolescence, for example—is nothing more than a reflection of the political, economic, and social circumstances in which we live. They point out that, although puberty has been a feature of development for as long as humans have lived, it was not until the rise of compulsory education that we began treating adolescents as a special and distinct group. They also note that where we draw the line between adolescence and adulthood vacillates with political and economic vicissitudes. This suggests that social conditions, not biological givens, define the nature of adolescent development.

A similar theme is echoed by anthropologists who have written about adolescence, the most important of whom were Ruth Benedict (1934) and Margaret Mead (1928). Anthropologists have examined how different cultures structure the transition to adulthood and the nature and meaning of different sorts of *rites of passage*, the ceremonies used to mark the transition. Based on their cross-cultural observations of the transition into adulthood, Benedict and Mead concluded that societies vary considerably in the ways in which they view and structure adolescence. As a consequence, these thinkers viewed adolescence as a culturally defined experience—stressful and difficult in societies that saw it this way, but calm and peaceful in societies that had an alternative vision. Benedict, in particular, drew a distinction between continuous and discontinuous societies. In *continuous societies* (typically, nonindustrialized societies with little social change), the transition from adolescence to adulthood is gradual and peaceful. In *discontinuous societies* (typically, industrialized societies characterized by rapid social change), the transition into adulthood is abrupt and difficult.

4. Contemporary Status of Theories of Adolescence

Over time, the prominence of the 'grand' theories of adolescence discussed in this article has waned somewhat, as scholars of adolescence have oriented themselves more toward understanding very specific aspects of adolescent development and have become less interested in general, broad-brush accounts of the period. Today's scholars of adolescence are less likely to align themselves consistently with single theoretical viewpoints and more likely to borrow from multiple theories that may derive from very different disciplines. As such, contemporary views of adolescence attempt to integrate central concepts drawn from a wide range of biological, psychological, sociological, historical, and anthropological perspectives. The emphasis in these integrative and eclectic approaches has been on understanding the way in which the social context in which young people mature interacts with the biological and psychological influences on individual development.

See also: Adolescence, Psychiatry of; Adolescence, Sociology of; Infant and Child Development, Theories of; Social Competence: Childhood and Adolescence; Socialization in Adolescence

Bibliography

Bakan D 1972 Adolescence in America: From idea to social fact. In: Kagan J, Coles R (eds.) *Twelve to Sixteen: Early Adolescence*. Norton, New York

Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston

Blos P 1979 *The Adolescent Passage: Developmental Issues*. International Universities Press, New York

Coleman J S 1961 *The Adolescent Society: The Social Life of the Teenager and its Impact on Education*. Free Press of Glencoe, New York

Elder G H 1974 *Children of the Great Depression: Social Change in Life Experience*. University of Chicago Press, Chicago, IL

Erikson E H 1968 *Identity: Youth and Crisis*. Norton, New York

Freud A 1958 Adolescence. *Psychoanalytic Study of the Child* **13**: 255–78

Friedenberg E Z 1959 *The Vanishing Adolescent*. Beacon Press, Boston

Hall G S 1904 *Adolescence*. Appleton, New York

Inhelder B, Piaget J 1958 *The Growth of Logical Thinking From Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures*. Basic Books, New York

Kett J F 1977 *Rites of Passage: Adolescence in America, 1790 to the Present*. Basic Books, New York

Kohlberg L 1969 Stage and sequence: The cognitive-developmental approach to socialization. In: Goslin D (ed.) *Handbook of Socialization Theory and Research*. Rand-McNally, Chicago

Lewin K 1951 *Field Theory in Social Science: Selected Theoretical Papers*. Harper, New York

Mannheim K 1952 The problem of generations. In: Mannheim K (ed.) *Essays on the Sociology of Knowledge*. Oxford University Press, New York, pp. 276–322

Mead M 1928 *Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilisation*. Morrow, New York

Sullivan H S 1953 *The Interpersonal Theory of Psychiatry*. Norton, New York

L. Steinberg

Adolescent Health and Health Behaviors

1. Definition of Adolescent Health Behavior

The topic of adolescent health behavior comprises two related areas. One concerns behaviors that may create threats to health during adolescence; the other concerns behaviors that place individuals at increased risk for chronic diseases in adulthood that have behavioral

components (cardiovascular disease or cancer). Research on adolescent health behavior determines factors related to increased risk for maladaptive behaviors such as cigarette smoking (*risk factors*), and variables that decrease risk for these behaviors (*protective factors*) and also may operate to reduce the impact of risk factors (*buffering effects*). This article considers factors related to adolescents' substance use, sexual behavior, violence, and suicide risk (see *Health Behaviors*).

2. Intellectual Context

The study of adolescent health behavior arose from epidemiological research during the 1950s and 1960s, which demonstrated that mortality from chronic disease in adulthood, such as heart attack or cancer, was related to factors such as substance use, dietary patterns, and life stress. Combining these findings with results from studies on longitudinal tracking of behavioral and physiological risk factors gave recognition to the concept that risk status began to develop earlier. This concept suggested a focus on studying health-related behaviors and dispositions at younger ages (e.g., cigarette smoking, hostility) so as to indicate early preventive approaches that would result in better health status over the long term (see *Adolescent Development, Theories of*).

The focus of this article is guided by data on causes of mortality during adolescence and young adulthood. United States data for 1997 show that for persons 15–24 years of age the most common causes of mortality are: accidents (prevalence of 35.4 per 100,000 population), homicide (15.8), and suicide (11.3). These prevalences are much greater than rates of mortality from congenital and infectious diseases. The most common causes of mortality for persons 25–44 years of age are: accidents (30.5 per 100,000), malignant neoplasms (25.8), heart disease (18.9), suicide (14.4), AIDS (13.4), and homicide (9.9). Though the relative rates for different causes could vary across countries, these statistics draw attention to the fact that three causes (accidents, violence, and suicide) are leading causes of mortality during both adolescence and adulthood. Thus, inquiring into factors related to these outcomes is of primary importance for the study of adolescent health.

The conditions that place adolescents at risk for adverse outcomes span several domains. Substance use has been linked to all the major causes of mortality during adolescence, including accidents, violence, and suicide (Goreczny and Hersen 1999). Chronic emotional distress (including depression and anger) is implicated in risk for substance use and violence (Marlatt and VandenBos 1997). Unprotected sexual intercourse, a factor in adolescent pregnancy, HIV infection, and sexually transmitted diseases, is related to many of the factors that predict adolescent substance use (S. B. Friedman 1998). The chapter focuses on research relevant to these conditions, noting also that various risk behaviors tend to occur together (e.g., substance use and unprotected sex).

3. Dominant Theories and Changes in Emphasis Over Time

3.1 Early Theoretical Models

Work on adolescent health was based initially in *cognitive and attitudinal* approaches to prediction of behavior (see *Health Behavior: Psychosocial Theories*). The Health Belief Model, developed by Rosenstock and Becker, assumed health behavior to be a function of the perceived benefits of the behavior, the costs (health, social, or economic) associated with the behavior, and the perceived barriers to engaging in the behavior. This model was used sometimes to generate research strategies for maladaptive behaviors, but this did not work out too well, as knowledge measures typically failed to correlate with behavior and prevention programs based on 'scare tactics' failed to deter smoking. Attitudinal approaches such as the Theory of Reasoned Action, developed by Fishbein and Azjen, posited that favorable attitudes would lead to intentions to perform a given behavior, and intentions would then be related to occurrence of the behavior. While intentions at one time point predicted behavior at subsequent time points, this model did not generate predictions about how attitudes develop in the first place, and prevention programs aimed at attitude modification had mixed results (Millstein et al. 1993).

The *social influence model* proposes that many behaviors are acquired through observing and modeling the behavior of influential others. This model was articulated by Richard Evans in primary prevention programs that used filmed or live models who demonstrated how to deal with situations where peers offered cigarettes and applied pressure to smoke. This approach showed preventive effects for adolescent smoking onset and made this model influential for a generation of smoking prevention programs (Ammerman and Hersen 1997). However, some studies showed reverse effects among students who were already smoking, suggesting that factors beside social pressure should be considered.

The *problem behavior model*, developed by Richard and Shirley Jessor, viewed adolescent substance use as a socially defined deviant behavior linked to rejection of conventional values as represented by family, school, and legal institutions. This model proposed that attitudes tolerant of deviance are related to several domains of variables, including personality, poor relationship with parents, low commitment to conventional routes to achievement (e.g., getting good grades), and affiliation with peers who were engaging in deviant behaviors (Jessor 1998). The problem

behavior model proposed that individuals who reject conventional values will demonstrate independence through adopting multiple deviant behaviors (e.g., heavy drinking, marijuana use, precocious sexuality). This model accounted for the observed intercorrelation of problem behaviors. Preventive implications were less clear, since it was not obvious whether to focus on self-esteem, deviant attitudes, family relationships, school performance, or deviant peer groups. This theory influenced several types of preventive interventions.

3.2 Changes in Focus Over Time

In the 1990s there have been several conceptual changes in approaches to understanding adolescent health behavior. The range of variables shown to predict substance use and other health-related behaviors has led to statistical models emphasizing how risk or protection arises through an interplay of contributions from individual, environmental, and social factors (Sussman and Johnson 1996). Thus, recent research on adolescent health has tested multiple domains of predictors, recognizing that health-risk behavior is not due to any single cause but rather arises from combinations of factors (Drotar 1999, Wills et al. 2001).

A specific change is in the conceptualization of social influence processes. Previously it was assumed that adolescent smoking, for example, was attributable to explicit pressure applied by smoking peers to unwilling (or ambivalent) targets. However, research has shown that some individuals have relatively favorable perceptions of smokers and perceive relatively greater frequency and/or acceptance of smoking among their peers and family members; it is these individuals who are most likely to adopt smoking. Similar processes have been demonstrated for teenage alcohol use and sexual behavior (Gibbons and Gerrard 1995). Hence there is more attention to how health behaviors may be motivated by social perceptions about persons who engage in those behaviors.

Earlier theories gave relatively little attention to emotional factors, but recent evidence shows that problem behavior occurs more often among adolescents experiencing higher levels of anxiety/depression, subjective stress, or anger (Marlatt and VandenBos 1997). Current theories have drawn out the linkages of emotional variables to patterns of peer affiliations and willingness to use substances if an opportunity presents itself. This question has drawn attention to why some individuals experience difficulty in controlling their emotions and behavior, and several theoretical models include self-control as a central concept, studying the consequences of poor control for health-risk behavior (Wills et al. 2001).

Finally, recent research has shown that characteristics measured between 3 and 8 years of age predict substance use in early adolescence or mental health problems in adulthood; these include temperamental characteristics, coping and self-control skills, and aggressive behavior. The observation that characteristics measurable in childhood predict health-related behavior at later ages has led to theories aimed at understanding how early temperament attributes are related to development of risk for complex problem behaviors (Wills et al. 2000a).

4. Emphases in Current Research

This section summarizes current knowledge about variables that are related to adolescent health behaviors. Listing a variable as a predictive factor does not indicate that it is the sole cause—or even necessarily a strong cause—of a behavior. A given variable may be, statistically, a weak predictor of a behavior but a combination of variables can be a strong predictor of the behavior. Also, buffering effects are common in health behavior, for example a person with high life stress might experience few adverse health outcomes because he/she also had a high level of social support. Thus there is no 'magic bullet' that can be used to tell who smokes and who doesn't. To predict health behavior with much confidence, one would have to assess multiple variables and to consider the balance between levels of risk factors and levels of protective factors.

This section considers predictor variables that are relevant for each of the health behaviors noted previously, because there are substantial intercorrelations among risk behaviors and a substantial degree of commonality in the predictors for the various outcomes. One would not expect, for example, that all adolescent smokers would have high rates of sexual intercourse; but separate lists of predictors for the various behaviors would have a high degree of overlap (DiClemente et al. 1996). The reader can understand the predictive context for a particular health behavior in more detail through reading some of the references at the end of this chapter.

4.1 Demographic Variables

Substance use among US adolescents varies by gender (males typically showing higher rates), ethnicity (higher rates among Caucasians, lower rates among African–Americans), family structure (higher rates in single-parent families), religiosity (lower rates among persons who attend a religious organization), and parental socioeconomic status (SES), with higher rates among adolescents from lower-SES families. However, effects of demographic variables may change over time, as has happened with cigarette smoking, with boys and girls reversing positions in US data

from the 1970s through the 1980s (Johnston et al. 1999). Socioeconomic effects also are complicated, with different patterns for various indices of smoking and alcohol use. It is found typically that effects for demographic variables are mediated through other risk and protective factors listed subsequently, though ethnicity tends to have direct effects, which indicates specific cultural influences. Demographic effects observed in US data could differ in other cultures, so in other countries, reference to local data is warranted.

4.2 Environmental Variables

Environmental variables are a type of influence that may operate independent of other characteristics. The limited data available suggest elevated risk for substance use in neighborhoods with lower income and higher crime rates, and where the neighborhood is perceived by residents as dangerous and/or neglected by the local government. Noting the existence of a relationship, however, fails to characterize the great variability in effects of environmental variables. For example, a large proportion of persons growing up in poverty areas may go through adolescence showing little or no substance use. This is believed to occur because processes in families and educational systems provide protective effects that offset the risk-promoting potential of the environment.

4.3 Dispositional Constructs

Temperament dimensions of attentional orientation, the tendency to focus attention on a task and avoid distraction, and positive emotionality, the tendency to smile or laugh frequently, are indicated as protective factors. Indicated as risk factors are activity level, the tendency to move around frequently and become restless when sitting still, and negative emotionality, the tendency to be easily irritated and become intensely upset. It appears that these variables act through affecting the development of generalized self-control ability (Wills et al. 2000a). Two other constructs have been related to substance use, drunken driving and sexual behavior in adolescence. Novelty seeking reflects a tendency to need new stimuli and situations frequently and to become bored easily (Wills et al. 1999a). Sensation seeking is a related construct in which high scorers prefer intense sensations (e.g., loud music) and are characterized as spontaneous and disinhibited (Zuckerman 1994).

4.4 Supportive Family Relationships

A positive relationship with parents is an important protective factor with respect to several health behaviors. High supportiveness is present when adolescents feel that they can talk freely with parents when they have a problem, and the parents will provide emotional support or practical assistance when it is needed. Family support and communication are consistently observed to have buffering effects, reducing the impact of risk factors such as poverty and negative life events. Family support acts through multiple pathways, associated among adolescents with better self-control, more value on achievement, and less acceptance of substance use (Wills et al. 1996).

4.5 Conflictual Relationships

A conflictual relationship with parents, involving disagreements and frequent arguments, is a significant risk factor for various adolescent problem behaviors. Family conflict and family support vary somewhat independently in the population, and various combinations of support and conflict can be observed. It should be noted that some argumentation between parents and children is normal, as teenagers establish autonomy from their parents and work toward their own identity; but a high level of conflict, in the absence of protective factors, is potentially problematic. Young persons who feel rejected by their parents look for other sources of acceptance and approval and tend to gravitate into groups of deviance-prone peers, which can lead to detected problem behavior and further family conflict, pushing the adolescent into increasing disengagement from the family and greater involvement in deviant peer groups. Note that there is no invariant relation between family conflict and serious child abuse (physical or sexual). It cannot be assumed that adolescent problem behavior necessarily indicates a history of child abuse, or conversely that severe child abuse is necessary to create risk for problem behavior. The noteworthy fact is that a high level of arguments and criticism by parents is a significant risk factor, and ameliorating family conflict is an important focus for counseling and prevention programs (Peters and McMahon 1996).

4.6 Good Self-control

A central protective factor is the construct of good self-control, also termed 'planfulness' or 'executive functions.' It is measured by several attributes involved in the planning, organizing, and monitoring of behavior. For example, soothability is the ability to calm oneself down when excited or upset; problem solving involves an active approach to coping with problems through getting information, considering alternatives, and making a decision about solving the problem. Good self-control contributes to emotional balance, helps to promote valuable competencies (e.g., academic competence), and ensures that problems get resolved rather than worsening and accumulating over time (Wills et al. 2000b). Individuals with good self-control also seem to be more discerning in their choice

of companions, so that they inhabit a social environment with more achievement-oriented peers and fewer deviant peers.

4.7 Poor Self-control

Poor self-control, also termed 'disinhibition' or 'behavioral undercontrol,' is based on a core of related characteristics. For example, impatience is the tendency to want everything as soon as possible; impulsiveness is the tendency to respond to situations quickly, without giving much thought to what is to be done. Irritability (e.g., 'There are a lot of things that annoy me') is sometimes included as part of this complex. It should be recognized that poor self-control is not simply the absence of good control; the two attributes tend to have different antecedents and different pathways of operation. Poor self-control is related to less involvement in school and seems to be a strong factor for bringing on negative life events. In later adolescence, poor control is related to perceiving substance use or sexual behavior as useful for coping with life stresses, an important factor in high-risk behavior (Wills et al. 1999a). Good self-control has been found to buffer the effects of poor control, so level of poor control by itself is not as predictive as the balance of good and poor control systems (see *Health: Self-regulation*).

4.8 Aggressiveness and hostility

Though anger-proneness is correlated strongly with indices of poor self-control, it is discussed separately because aggression or conduct disorder in children has been studied as a risk factor for problem behaviors. Physical aggression is highly stable over time from childhood onwards, leading to difficulties with parental socialization and peer social relationships, and aggressive tendency is a strong predictor of substance use and other problem behaviors. A predisposition to respond aggressively in interpersonal situations is conducive to injury through violent encounters and hence makes this an important factor in adolescents' morbidity as well as injury to others (Goreczny and Hersen 1999). It should be noted that overt physical aggression is only part of a predictive syndrome that also involves impulsiveness and negative affect, and diagnostic studies sometimes show the majority of children with conduct disorder also have depressive disorder. The combination of high levels of aggression and depression (referred to as comorbid disorder) is a particular risk factor for both substance abuse and suicide in adolescents. However, clinical-level disorder is not a necessary condition for risk, as simple measures of irritability or hostility predict substance use in adolescence and adulthood (Wills et al. 2001). Thus, the core characteristic for risk is probably a complex of attributes that reflect difficulty in regulating reactions to irritation or frustration.

4.9 Academic Involvement

Low involvement in school is a notable risk factor for adolescent substance use and other problem behaviors. This may be reflected in negative attitudes toward school, low grades, poor relationships with teachers, and a history of discipline in school (Wills et al. 2000a). The effect of academic involvement is independent of characteristics such as SES and family structure, though it is related to these to some extent. The reasons for its relationship to problem behavior are doubtless complex. Low involvement in school can be partly attributable to restlessness or distractibility, which make it difficult to adjust to the classroom setting, or to aggressive tendencies that make it difficult for the child to keep friends. Disinterest in getting good grades may derive from a social environment that devalues conventional routes to achievement or a conflictual family that does not socialize children to work toward long-term goals. It should be noted that many adolescents have one bad year in school but do better in subsequent years, without adverse effect; but a trajectory of deteriorating academic performance and increasing disinterest in school could be predictive of subsequent problems such as frequent substance use.

4.10 Negative life events

An accumulation of many negative life events during the previous year has been implicated in adolescent substance use and suicide. The events may be ones that occur to a family member (e.g., unemployment of a parent) or ones that directly involve the adolescent him/herself (e.g., loss of a friend). The pathways through which life events are related to substance use have been studied to some extent. Negative events are related to increased affiliation with deviant peers, apparently because experiences of failure and rejection predispose the adolescent to disengage from conventional institutions and spend more time with peers who are themselves frustrated and alienated. Another pathway is that negative events elevate perceived meaninglessness in life, thus setting the stage perhaps for drug use as a way to restore feelings of control. Life events also may prime the need for affect regulation mechanisms to help deal with the emotional consequences of stressors (Wills et al. 1999a).

4.11 General Attitudes, Norms, and Perceived Vulnerability

Persons who perceive the behavior as relatively accepted in some part of their social circle, and/or who

perceive they are less vulnerable to harmful consequences, are more likely to engage in problem behavior (Wills et al. 2000a). Attitudes do not have to be totally favorable in order to create risk status; for example, attitudes about cigarette smokers tend to be negative in the adolescent population, but those who have relatively less negative attitudes are more likely to smoke (Gibbons and Gerard 1995). The source of attitudes and norms is not understood in detail. It is likely that some variance is attributable to attitudes communicated by parents, that influential peers also communicate norms about substance use (which may differ considerably from those held by parents), and that media advertising also communicates images about substance use and sexual behavior. Attitudes may be shaped to some extent by dispositional characteristics; for example, high novelty-seekers tend to view substance use as a relatively desirable activity and perceive themselves as less vulnerable to harmful effects of tobacco and alcohol. Ongoing studies conducted among US teenagers have shown that rates of marijuana use vary inversely with the level of beliefs in the population about harmful effects of marijuana. The source of these beliefs has not been decisively linked to any specific source, but school-based preventive programs and government-sponsored counteradvertising are suggested as influential (see *Vulnerability and Perceived Susceptibility, Psychology of*).

4.12 Specific Attitudes and Efficacies

Specific attitudes and efficacies may be relevant for particular behaviors. In the area of sexual behavior and contraceptive use, for example, attitudes about sexuality and perceived efficacy for various types of contraceptive use are relevant (DiClemente et al. 1996). There are widely varying attitudes about condom use and differences across persons in the degree to which they feel comfortable in communicating with partners about condom use. Thus studies should always make an attempt to elicit and address specific beliefs and attitudes about a health behavior (Cooper et al. 1999). Perhaps the one general statement that can be made is about resistance efficacy, the belief that one can successfully deal with situations that involve temptation for a behavior (e.g., being offered a cigarette at a party). High resistance efficacy has been noted as a protective factor from relatively young ages. The source of this efficacy has not been extensively studied but some data relate high efficacy to a good parent-child relationship, to good self-control, and (inversely) to indices of anger and hostility (Wills et al. 2000a).

4.13 Emotional States

Negative emotional states including depression, anxiety, and subjective stress have been linked to adolescent substance use and other problem behavior. The source of the stress may be from recent negative events, from dispositional characteristics (e.g., neuroticism), or from living in a threatening environment. Some current evidence supports each of these perspectives. It is important to recognize that emotional states are linked to beliefs about oneself and the world. An individual reporting a high level of negative mood on a symptom checklist is also likely to endorse beliefs that they are an unattractive and unworthy person, that their current problems are uncontrollable, and that there is no clear purpose or meaning in their present life (Wills and Hirky 1996). Evidence has linked components from this complex (lack of control, pessimism, and perceived meaninglessness) to adolescent substance use and suicide risk, but the dynamic of the process is not well understood; it is conceivable that affect *per se* is less important than the control beliefs and world views that are embedded in the matrix of current emotions. Note that positive affect —which is not simply the absence of negative affect— is a protective factor for various problem behaviors and that positive affect has buffering effects, reducing the impact of negative affect on substance use (Wills et al. 1999a). Thus, in studying emotional states and risk status it is valuable to assess the balance of positive and negative affect for an individual.

4.14 Peer Relationships

Peer relationships are one of the most important factors in adolescent health behavior. Across a large number of studies it has been noted that adolescents who smoke, drink, fight, and/or engage in sexual behavior tend to have several other friends who do likewise (DiClemente et al. 1996, Marlatt and VandenBos 1997). Frequency and extent of use is the major consideration for risk status. For example, many teenagers will end up at times in situations where a friend is using some substance, but when many of a person's friends are smoking and drinking frequently, perhaps engaging in other illegal behaviors, and feeding a growing cycle of alienated beliefs, then concern would mount. Several aspects of the context of peer group membership should be considered. While engagement in peer group activity is normative for adolescents, it is when a person has high support from peers and low support from parents that substance use is particularly elevated. Also, there may be several different types of peer networks in a given school, including groups that are grade-oriented, athletes, persons identifying with painting or theater, and cliques focused around specific themes (e.g., skateboarding, 'heavy metal' rock music); hence the frequency of peer activity may be less important for risk than the types of peers and their associated behaviors. Appropriate responding to peer behavior has been a primary focus in prevention programs,

which have aimed to teach social skills for responding assertively in situations where an opportunity for a problem behavior occurs (Sussman and Johnson 1996).

4.15 Coping Motives

Persons may engage in a given behavior for different reasons, and the reasons have significant implications for adolescent health behavior. Problematic substance use and sexual behavior are prominent particularly among individuals for whom these behaviors are regarded as an important coping mechanism, perceived as useful for affect regulation and stress reduction (Wills et al. 1999a). This aspect distinguishes adolescents who use tobacco and alcohol at relatively low rates from those who use multiple substances at high rates and experience negative consequences because of inappropriate or dependent use. Coping motives for substance use are related to parental substance use, to poor self-control, and to a dispositional dimension (risk-taking tendency), hence are based in a complex biopsychosocial process involving self-regulation ability. Motivation concepts are also relevant for health-protective behavior such as condom use (Cooper et al. 1999).

5. Future Directions

This article has emphasized that meaningful risk is not predictable from knowledge about a single variable; rather it is the number of risk factors and their balance with protective factors that is most informative. The concept that health behavior is related to variables at several levels of analysis (environmental variables, personality variables, family variables, and social variables) has been emphasized, and the article has discussed how these variables interrelate to produce health promoting vs. problematic behavior.

Several future developments can be anticipated in this area. One is increasing use of theories that delineate how different domains of variables are related to problem behavior. For example, epigenetic models suggest how simple temperament characteristics are related over time to patterns of family interaction, coping, and social relationships, which are proximal factors for adolescent substance use and other behaviors (Wills et al. 2000b). Research of this type will be increasingly interdisciplinary, involving collaborations of investigators with expertise in developmental, social, and clinical psychology.

Another development is increasing integration of genetic research with psychosocial research. It has been known for some time that parameters relevant for cardiovascular disease (e.g., blood pressure and obesity) have a substantial heritable component; recent research also has shown substantial genetic contributions for liability to cigarette smoking and alcohol abuse/dependence. Although these health-related variables are shown related to genetic characteristics, there is little understanding of the physiological pathways involved (Wills et al. 2001). Current investigations are studying genes coding for receptors for neurotransmitters that have been linked to vulnerability to substance abuse and suicide, and identifying physiological and behavioral pathways for effects of genetic variation.

Finally, recent lifespan research has indicated that simple personality variables measured at early ages predict health-related outcomes over time (H. S. Friedman et al. 1995, Wills et al. 2001). Such research suggests investigations into whether early temperament characteristics are related directly to physiological pathways from the hypothalamic–pituitary axis, operating to dysregulate metabolic systems so as to create risk for cardiovascular disease and diabetes. Behavioral pathways are also possible in the observed longevity effects through liability for smoking or accident-proneness. Integrative research is suggested, using concepts from physiology and behavioral psychology to understand the mechanisms of psychosocial processes in premature mortality.

See also: Adolescence, Sociology of; Adolescent Development, Theories of; Adolescent Vulnerability and Psychological Interventions; Alcohol Use Among Young People; Childhood and Adolescence: Developmental Assets; Coping across the Lifespan; Drug Use and Abuse: Psychosocial Aspects; Health Behavior: Psychosocial Theories; Health Behaviors; Health Education and Health Promotion; Health Promotion in Schools; Self-efficacy; Self-efficacy and Health; Self-efficacy: Educational Aspects; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Substance Abuse in Adolescents, Prevention of

Bibliography

Ammerman R T, Hersen M (eds.) 1997 *Handbook of Prevention and Treatment with Children and Adolescents: Intervention in the Real World Context*. Wiley, New York

Cooper M L, Agocha V B, Powers A M 1999 Motivations for condom use: Do pregnancy prevention goals undermine disease prevention among heterosexual young adults? *Health Psychology* **18**: 464–74

DiClemente R J, Hansen W B, Ponton L E (eds.) 1996 *Handbook of Adolescent Health Risk Behavior*. Plenum, New York

Drotar D (ed.) 1999*Handbook of Research Methods in Pediatric and Clinical Child Psychology*. Kluwer Academic/Plenum Publishers, New York

Friedman H S, Tucker J S, Schwartz J E, Martin L, Tomlinson-Keasey C, Wingard D, Criqui M 1995 Childhood conscientiousness and longevity. *Journal of Personality and Social Psychology* **68**: 696–703

Friedman S B (ed.) 1998 *Comprehensive Adolescent Health Care*. Mosby, St. Louis, MO

Gibbons F X, Gerrard M 1995 Predicting young adults' health risk behavior. *Journal of Personality and Social Psychology* **69**: 505–17

Goreczny A J, Hersen M (eds.) 1999 *Handbook of Pediatric and Adolescent Health Psychology*. Allyn and Bacon, Boston, MA

Jessor R (ed.) 1998 *New Perspectives on Adolescent Risk Behavior*. Cambridge University Press, New York

Johnston L D, O'Malley P M, Bachman J G 1999 *National Survey Results on Drug Use from Monitoring the Future Study, 1975–1998*. National Institute on Drug Abuse, Rockville, MD

Marlatt G A, VandenBos G R (eds.) 1997 *Addictive Behaviors: Readings on Etiology, Prevention, and Treatment*. American Psychological Association, Washington, DC

Millstein S G, Petersen A C, Nightingale E O (eds.) 1993 *Promoting the Health of Adolescents*: New Directions for the Twenty-First Century. Oxford University Press, New York

Peters R D V, McMahon R J (eds.) 1996 *Preventing Childhood Disorders, Substance Abuse, and Delinquency*. Sage, Thousand Oaks, CA

Sussman S, Johnson C A (eds.) 1996 Drug abuse prevention: Programming and research recommendations. *American Behavioral Scientist* **39**: 787–942

Wills T A, Cleary S D, Shinar O 2001 Temperament dimensions and health behavior. In: Hayman L, Turner J R, Mahon M (eds.) *Health and Behavior in Childhood and Adolescence*. Erlbaum, Mahwah, NJ

Wills T A, Gibbons F X, Gerrard M, Brody G 2000a Protection and vulnerability processes for early onset of substance use: A test among African-American children. *Health Psychology* **19**: 253–63

Wills T A, Hirky A 1996 Coping and substance abuse: Theory, research, applications. In: Zeidner M, Endler N S (eds.) *Handbook of Coping*. Wiley, New York, pp. 279–302

Wills T A, Mariani J, Filer M 1996 The role of family and peer relationships in adolescent substance use. In: Pierce G R, Sarason B R, Sarason I G (eds.) *Handbook of Social Support and the Family*. Plenum, New York, pp. 521–49

Wills T A, Sandy J M, Shinar O 1999a Cloninger's constructs related to substance use level and problems in late adolescence. *Experimental and Clinical Psychopharmacology* **7**: 122–34

Wills T A, Sandy J M, Shinar O, Yaeger A 1999b Contributions of positive and negative affect to adolescent substance use: Test of a bidimensional model in a longitudinal study. *Psychology of Addictive Behaviors* **13**: 327–38

Wills T A, Sandy J M, Yaeger A 2000b Temperament and adolescent substance use: An epigenetic approach to risk and protection. *Journal of Personality* **68**(6): 1127–51

Zuckerman M 1994 *Behavioral Expressions and Biosocial Bases of Sensation Seeking*. Cambridge University Press, New York

T. A. Wills

Adolescent Injuries and Violence

Injuries are probably the most underrecognized public health problem facing the nation today. More adolescents in the United States die from unintentional injuries and violence than from all diseases combined. In 1998, 13,105 US adolescents aged 10 to 19 years died from injuries—the equivalent of more than one death every hour of every day (CDC 2001b). Injuries can affect dramatically an adolescent's development, social and physical growth, family and peer relations, and activities of daily living. The societal costs are enormous, including medical costs, lost productivity, and often extended welfare and rehabilitation costs. Because injury takes such a toll on the health and well-being of young people, approximately 50 Healthy People 2010 national health objectives address the reduction of injuries and injury risks among adolescents (US DHHS 2000). Families, schools, professional groups, and communities have the potential to prevent injuries to adolescents, and help youth to establish lifelong safety skills.

1. Unintentional Injury and Violence

An injury consists of unintentional or intentional damage to the body that results from acute exposure to thermal, mechanical, electrical, or chemical energy, or from the absence of such essentials as heat or oxygen. Injuries can be classified based on the events and behaviors that precede them, as well as the intent of the persons involved. Violence is the intentional use of physical force or power, threatened or actual, against oneself, another person, or a group or community, that either results in or is likely to result in injury, death, psychological harm, maldevelopment, or deprivation. Types of violence include homicide, assault, sexual violence, rape, child maltreatment, dating or domestic violence, and suicide. Unintentional injuries are those not caused by deliberate means, such as injuries related to motor vehicle crashes, fires and burns, falls, drowning, poisoning, choking, suffocation, and animal bites. These are often referred to as 'accidents,' although scientific evidence indicates that these events can be prevented.

Almost 72 percent of the 18,049 deaths of adolescents aged 10 to 19 years are attributed to only four causes: motor vehicle traffic crashes (34 percent), all other unintentional injuries (13 percent), homicide (14 percent), and suicide (11 percent) (CDC 2001b). Unintentional injuries, primarily those attributed to motor vehicle crashes, are the leading cause of death in the United States throughout adolescence. However, the relative importance of homicide and suicide increases from early to late adolescence. Homicide is the fourth leading cause of all deaths to US adolescents aged 10 to 14 years, and the second leading cause of death among adolescents aged 15 to 19 years. Suicide is the third leading cause of death among adolescents aged 10 to 19 years. From early to late adolescence, the number of suicides increases fivefold and the number of homicides increases eightfold.

Nonfatal injuries are even more common during adolescence. For every injury death in the United

States, approximately 41 injury hospitalizations occur, and 1,100 cases are treated in hospital emergency departments (CDC 1993). More than 7.4 million US adolescents aged 15 to 24 years suffer injuries requiring hospital emergency department visits annually (210.1 per 1,000 persons). Injuries requiring medical attention or resulting in restricted activity affect more than 20 million children and adolescents, and cost $17 billion annually in medical costs.

2. Priority Injuries

2.1 Motor Vehicle-related Injuries

More young people die in the United States from motor vehicle-related injuries than from any other cause. The majority of adolescent traffic-related deaths occur as motor vehicle occupants—60 percent of traffic-related deaths among adolescents aged 10–14 years and 86 percent of traffic-related deaths among those aged 15–19 years. In addition, 750,000 adolescents in the United States are victims of nonfatal motor vehicle injuries each year (Li et al. 1995). The likelihood that children and adolescents will suffer fatal injuries in motor vehicle crashes increases if alcohol is used. Teenaged male driver death rates are about twice those of females, and crash risk for both males and females is particularly high during the first years teenagers are eligible to drive. Traffic-related injuries also include those sustained while walking, riding a bicycle, or riding a motorcycle.

Collisions with motor vehicles are the causes of almost all bicycle-related deaths, hospitalizations, and emergency room visits among adolescents. Bicycles are associated with 60 adolescent deaths, 6,300 hospitalizations, and approximately 210,000 emergency room visits annually among US adolescents (Li et al. 1995), 90 percent of which are attributed to collisions with motor vehicles. Severe head injuries are responsible for 64 percent to 86 percent of bicycle-related fatalities. Children aged 10 to 14 years have the highest rate among all age groups of bicycle-related fatalities. In 1998, 463 US adolescents died as pedestrians, and 145 as motorcyclists (CDC 2001b).

2.2 Homicide, Assaults, and Interpersonal Violence

Adolescents are more likely than the general population to become both victims and perpetrators of violence. Between 1981 and 1990, the homicide rate among adolescents aged 10–19 years in the United States increased by 61 percent, while the overall rate in the population decreased by 2 percent (CDC 2001b). From 1990 to 1998, the homicide rate decreased by 31 percent among adolescents, and 34 percent in the overall population. Most adolescent homicide victims in the United States are members of minority ethnic and racial groups. In 1998, the homicide rate among males aged 10 to 19 years was 3.0 per 100,000 among white, non-Hispanic males; 6.4 per 100,000 among Asian/Pacific Islander males; 11.1 per 100,000 among American Indian/Alaskan Native males; 18.4 per 100,000 among Hispanic males; and 38.8 per 100,000 among black, non-Hispanic males (CDC 2001b). In 1998, adolescents aged 10 to 17 years accounted for one out of every six arrests for violent crimes in the United States (US DHHS 2001).

Females aged 18 to 21 years in the United States have the highest rate of rape or sexual assault victimization (13.8 per 1,000), followed by those aged 15 to 17 years (12 per 1,000) (Perkins 1997). More than one-half of female rape victims are less than 18 years of age (Tjaden and Thoennes 1998). Being raped before age 18 doubles the risk of subsequent sexual assault; 18 percent of US women raped before age 18 were also rape victims after age 18, compared with 9 percent of women not raped before the age of 18 (Tjaden and Thoennes 1998). Sexual violence is often perpetrated by someone known to the victim.

2.3 Suicide and Suicide Attempts

In 1998, 2054 adolescents aged 10 to 19 years completed suicide in the United States (CDC 2001b). One of the first detectable indications of suicide contemplation is suicidal ideation and planning. In 1999, 19 percent of US high school students had suicidal thoughts and 15 percent had made plans to attempt suicide during the preceding year (CDC 2000). Three percent of US high school students reported making a suicide attempt that had required medical treatment during the preceding year.

Depressive disorders, alcohol and drug abuse, family discord, arguments with a boyfriend or girlfriend, school-related problems, hopelessness, and contact with the juvenile justice system are commonly cited risk factors for suicide. Exposure to the suicide of others also may be associated with increased risk of suicidal behavior.

3. Settings for Adolescent Injuries

3.1 School-related Injuries

The most frequently treated health problem in US schools is injury. Between 10 percent and 25 percent of child and adolescent injuries occur on school premises, and approximately 4 million children and adolescents in the United States are injured at school each year (Posner 2000). Although the recent wave of school shootings in the United States has captured public attention, homicides and suicides are rare events at school: only 1 percent of homicides and suicides among children and adolescents occur at school, in

transit to and from school, or at school-related events (Kachur et al. 1996).

School-associated nonfatal injuries are most likely to occur on playgrounds, particularly on climbing equipment, on athletic fields, and in gymnasia. The most frequent causes of hospitalization from school-associated injuries are falls (43 percent), sports activities (34 percent), and assaults (10 percent). Male students are injured 1.5 times more often than female students (Di Scala et al. 1997). Middle and high school students sustain school injuries somewhat more frequently than elementary school students: 41 percent of school injury victims are 15 to 19 years of age, 31 percent are 11 to 14 years of age, and 28 percent are 5 to 10 years of age (Miller and Spicer 1998).

3.2 Sports-related Injuries

In the United States, more than 8 million high school students participate in school- or community-sponsored sports annually. More than one million serious sports-related injuries occur annually to adolescents aged 10 to 17 years, accounting for one-third of all serious injuries in this age group and 55 percent of nonfatal injuries at school (Cohen and Potter 1999). For those aged 13 to 19 years, sports are the most frequent cause of nonfatal injuries requiring medical treatment among both males and females.

Males are twice as likely as females to experience a sports-related injury, probably because males are more likely than females to participate in organized and unorganized sports that carry the greatest risk of injury, such as American football, basketball, gym games, baseball, and wrestling (Di Scala et al. 1997). Among sports with many female participants, gymnastics, track and field, and basketball pose the greatest risk of nonfatal injury. Among sports with male and female teams (such as soccer or basketball), the injury rate per player is higher among females than males (Powell and Barber-Foss 2000).

3.3 Work-related Injuries

Half of all US adolescents aged 16–17 years, and 28 percent of those aged 15 years, are employed. On average, these adolescents work 20 hours per week for about half the year. In 1992, more than 64,000 adolescents aged 14–17 years required treatment in a hospital emergency department for injuries sustained at work. In the United States, approximately 70 adolescents under 18 years of age die while at work every year (Cohen and Potter 1999).

Adolescents are exposed to many hazards at work, including ladders and scaffolding, tractors, forklifts, restaurant fryers and slicers, motor vehicles, and night-time work. In particular, motor vehicles and machinery are associated with on-the-job injuries and deaths. Night work is associated with an increased risk of homicide, which is the leading cause of death while at work for female workers of all ages.

4. Risk Behaviors Associated with Injury

4.1 Alcohol Use

Each month, half of US high school students drink alcohol on at least one day and 32 percent engage in episodic heavy drinking—consuming five or more drinks on a single occasion (CDC 2000). Alcohol use is associated with 56 percent of motor-vehicle-related fatalities among people in the United States aged 21–24 years, 36 percent of fatalities among those aged 15–20 years, and 20 percent of fatalities among children less than 15 years of age. Alcohol use is a factor in more than 30 percent of all drowning deaths, 14–27 percent of all boating-related deaths, 34 percent of all pedestrian deaths, and 51 percent of adolescent traumatic brain injuries.

Alcohol use is also associated with many adolescent risk behaviors, including using other drugs and delinquency, carrying weapons and fighting, attempting suicide, perpetrating or being the victim of date rape, and driving while impaired. In the United States, in a given month, 13 percent of high school students drive a motor vehicle after drinking alcohol, and 33 percent ride in a motor vehicle with a driver who has been drinking alcohol (CDC 2000).

4.2 Access to Weapons

In 1998, firearms were the mechanism of injury in 82 percent of homicides and 60 percent of suicides among adolescents aged 10 to 24 years in the United States (CDC 2001b). People with access to firearms may be at increased risk of both homicide and suicide (Kellermann et al. 1993). In approximately 40 percent of homes with both children and firearms, firearms are stored locked and unloaded (Schuster et al. 2000). In 1999, 17 percent of high school students reported carrying a weapon, such as a gun, knife or club, and nearly 5 percent reported carrying a firearm during the previous month. During the same time period, 7 percent carried a weapon on school property (CDC 2000).

4.3 Inadequate Use of Seat Belts and Helmets

Proper use of lap and shoulder belts could prevent approximately 60 percent of deaths to motor vehicle occupants in a crash in the United States (CDC 2001a). Motorcycle helmets may be 35 percent effective in preventing fatal injuries to motorcyclists, and 67

percent effective in preventing brain injuries. Proper bicycle helmet use could prevent up to 56 percent of bicycle-related deaths, 65 percent to 88 percent of bicycle-related brain injuries, and 65 percent of serious injuries to the face.

Adolescents are among the least frequent users of seat belts or helmets. In the United States, 16 percent of high school students claim that they never or rarely use seat belts when riding in a car driven by someone else. Of the 71 percent of US high school students who rode a bicycle in 1999, 85 percent rarely or never wore a bicycle helmet (CDC 2000). Peer pressure and modeling by family members may keep adolescents from using seat belts and bicycle helmets.

5. *Injury Prevention Strategies*

Most injuries to adolescents are both predictable and preventable. Preventing adolescent injuries requires innovations in product design and changes in environment, technology, behavior, social norms, legislation, enforcement, and policies.

Strategies such as product modifications (e.g., integral firearm locking mechanisms), environmental changes (e.g., placing soft surfaces under playground equipment), and legislation (e.g., mandating bicycle helmet use) usually result in more protection to a population than strategies requiring individual behavior change. However, behavioral change is a necessary component of even the most effective legislative, technological, automatic, or passive strategies (Gielen and Girasek 2001): even when seat belts or bicycle helmets are required by law, they must be used correctly and consistently to prevent injuries effectively.

While legislative strategies, such as graduated driver licensing laws to control adolescent driving behavior, or school policies to reduce violence, hold promise, they must be supported by parents and the public, and must be enforced by local authorities to be effective (Schieber et al. 2000). Other approaches, such as zero tolerance alcohol policies, primary enforcement safety belt use laws, enhanced law enforcement, lowered permissible blood alcohol levels, minimum legal drinking age laws, and sobriety checkpoints, have been shown to be effective in reducing motor vehicle-related injuries and death (CDC 2001a). To prevent youth violence, parent and family-based strategies, regular home visits by nurses, social-cognitive and skills-based strategies, and peer mentoring of adolescents have been recommended as 'best practices' (Thornton 2000).

The broad application of these and other health promotion strategies can lead to reductions in adolescent injury. It is not allegiance to a particular type of intervention but flexibility in combining strategies that will produce the most effective mix (Sleet and Gielen 1998). For example, to yield the desired result, legislation requiring the use of bicycle helmets should be accompanied by an educational campaign for teens and parents, police enforcement, and discounted sales of helmets by local merchants. Education and social skills training in violence prevention must be accompanied by changes in social norms, and policies that make the use of violence for resolving conflict less socially acceptable.

6. *Conclusions*

Injuries are the largest source of premature morbidity and mortality among adolescents in the United States. The four major causes of adolescent deaths are motor vehicle crashes, homicide, other unintentional injuries, and suicide. Risk-taking behaviors are an intrinsic aspect of adolescent development but they can be minimized by emphasizing strong decision-making skills, and through changes in the environment that facilitate automatic protection and encourage individual behaviors which result in increased personal protection. Interventions to reduce adolescent injuries must be multifaceted and developmentally appropriate, targeting environmental, product, behavioral, and social causes. Injury policy development; education and skill building; laws and regulations; family-, school-, and home-based strategies; and enforcement are important elements of a comprehensive community-based adolescent injury prevention program. The public health field cannot address the adolescent injury and violence problem effectively in isolation. Youth and families, schools, community organizations and agencies, and businesses should collaborate to develop, implement, and evaluate interventions to reduce the major sources of injuries among adolescents.

See also: Adolescence, Sociology of; Adolescent Behavior: Demographic; Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Adolescent Vulnerability and Psychological Interventions; Childhood and Adolescence: Developmental Assets; Disability: Psychological and Social Aspects; Injuries and Accidents: Psychosocial Aspects; Rape and Sexual Coercion; Risk, Sociological Study of; Suicide; Violence as a Problem of Health; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Gangs

Bibliography

CDC (Centers for Disease Control and Prevention) 1993 *Injury Mortality: National Summary of Injury Mortality Data 1984–1990*. National Center for Injury Prevention and Control, Atlanta, GA

CDC (Centers for Disease Control and Prevention) 2000 CDC Surveillance summaries: Youth risk behavior surveillance—United States, 1999. *MMWR*. **49**(SS-5): 1–94

CDC (Centers for Disease Control and Prevention) 2001a

Motor vehicle occupant injury: Strategies for increasing use of child safety seats, increasing use of safety belts, and reducing alcohol-impaired driving: A report on recommendations of the task force on community preventive services. *MMWR* **50**(RR-7): 1–13

CDC National Center for Injury Prevention and Control, Office of Statistics and Programming 2001b *Web-based Injury Statistics Query and Reporting System (WISQARS)*. NCHS Vital Statistics System. Online at http://www.cdc.gov/ncipc/wisqars. Accessed April 11

Cohen L R, Potter L B 1999 Injuries and violence: Risk factors and opportunities for prevention during adolescence. *Adolescent Medicine: State of the Art Reviews* **10**(1): 125–35

Di Scala C, Gallagher S S, Schneps S E 1997 Causes and outcomes of pediatric injuries occurring at school. *Journal of Health* **67**: 384–9

Gielen A C, Girasek D C 2001 Integrating perspectives on the prevention of unintentional injuries. In: Schneiderman N, Speers M A, Silva J M, Tomes H, Gentry J H (eds.) *Integrating Behavioral and Social Sciences with Public Health*. American Psychological Association, Washington DC

Kachur S P, Stennies G M, Powell K E, Modzeleski W, Stephens R, Murphy R, Kresnow M, Sleet D, Lowry R 1996 School-associated violent deaths in the United States, 1992–1994. *JAMA* **275**: 1729–33

Kellermann A, Rivara F P, Rushforth N B, Banton J G 1993 Gun ownership as a risk factor for homicide in the home. *New England Journal of Medicine* **329**: 1084–91

Li G, Baker S P, Frattaroli S 1995 Epidemiology and prevention of traffic-related injuries among adolescents. *Adolescent Medicine: State of the Art Reviews* **6**: 135–51

Miller T R, Spicer R S 1998 How safe are our schools? *American Journal of Public Health* **88**(3): 413–8

Perkins C A 1997 *Bureau of Justice Statistics Special Report: Age Patterns of Victims of Serious Violent Crime*. USDOJ publication no. NCJ 162031. US Department of Justice, Washington DC

Posner M 2000 *Preventing School Injuries: A Comprehensive Guide for School Administrators, Teachers, and Staff*. Rutgers University Press, New Brunswick, NJ

Powell J W, Barber-Foss K D 2000 Sex-related injury patterns among selected high school sports. *American Journal of Sports Medicine* **28**(3): 385–91

Schieber R A, Gilchrist J, Sleet D A 2000 Legislative and regulatory strategies to reduce childhood injuries. *The Future of Children* **10**(1): 111–36

Schuster M A, Franke T M, Bastian A M, Sor S, Halfon N 2000 Firearm storage patterns in US homes with children. *American Journal of Public Health* **90**(4): 588–94

Sleet D A, Gielen A C 1998 Injury prevention. In: Gorin S S, Arnold J (eds.) *Health Promotion Handbook*. Mosby, St Louis, MO

Thornton T N, Craft C A, Dahlberg L L, Lynch B S, Baer K 2000 *Best Practices of Youth Violence Prevention: A Source-Book for Community Action*. National Center for Injury Prevention and Control, Atlanta, GA

Tjaden P, Thoennes N 1998 *Prevalence, Incidence, and Consequences of Violence Against Women: Findings from the National Violence Against Women Survey, Research in Brief*. Publication no. (NCJ) 172837. National Institute of Justice and Centers for Disease Control and Prevention, Washington, DC

US DHHS (US Department of Health and Human Services) 2000 *Healthy People 2010* (Conference edition, 2 vols.). US DHHS, Washington DC

US DHHS (US Department of Health and Human Services) 2001 *Youth Violence: A Report of the Surgeon General*. US DHHS, Centers for Disease Control and Prevention, National Center for Injury Prevention and Control; Substance Abuse and Mental Health Services Administration, Center for Mental Health Services; and National Institutes of Health, National Institute of Mental Health, Rockville, MD

L. Barrios and D. Sleet

Adolescent Vulnerability and Psychological Interventions

Adolescents face special sources of vulnerability as they expand their lives into domains beyond their guardians' control. The magnitude of those risks depend on the challenges that their environment presents and teens' own ability to manage them. Psychological interventions seek to improve teens' coping skills, and to identify circumstances in which society must provide more manageable environments.

1. Assessing Personal Vulnerability

Adults often say that teens do reckless things because they feel invulnerable. If that is the case, then teens' perceptions resemble those of adults, for whom the phenomenon of unrealistic optimism is widely documented. In countries where such research has been conducted, most adults are typically found to see themselves as less likely than their peers to suffer from events that seem at least somewhat under their control. Moreover, people tend to exaggerate such control. Thus, for example, most adults see themselves as safer than average drivers, contributing to their tendency to underestimate driving risks. Unrealistic optimism can lead people to take greater risks than they would knowingly incur. As a result, exaggerated feelings of invulnerability can create actual vulnerability. Such feelings might also provide the (unwarranted) confidence that people need to persevere in difficult situations, believing, against the odds, that they will prevail (e.g., Kahneman et al. 1982, Weinstein and Klein 1995).

The relatively few studies assessing the realism of teens' expectations have typically found that, if anything, teens are somewhat less prone to unrealistic optimism than are adults (e.g., Quadrel et al. 1993). These results are consistent with the general finding that, by their mid-teens, young people have most of the (still imperfect) cognitive capabilities of adults (Feldman and Elliott 1990). How well teens realize this potential depends on how well they can manage their own affect and others' social pressure. For example, if

teens are particularly impulsive, then they are less likely to make the best decisions that they could. Conversely, though, if they feel unable to make satisfactory choices, then they may respond more impulsively, or let decision making slide until it is too late for reasoned thought.

The vulnerabilities created by imperfections in teens' (or adults') judgments depend on the difficulty of the decisions that they face. Some choices are fairly forgiving; others are not (von Winterfeldt and Edwards 1986). In very general terms, decisions are more difficult when (a) the circumstances are novel, so that individuals have not had the opportunity to benefit from trial-and-error learning, either about how the world works or about how they will respond to experiences, (b) the choices are discrete (e.g., go or stay home, operate or not) rather than continuous (e.g., drive at 100 or 110 kph, spend *x* hours on homework), so that one is either right or wrong (rather than perhaps close to the optimum), (c) consequences are irreversible, so that individuals need to 'get it right the first time,' and (d) sources of authority are in doubt, so that reliable sources of guidance are lacking (and, with them, the moral compass of social norms or the hard-earned lessons of prior experience).

In these terms, teens face many difficult decisions. In a brief period, young people must establish behavior patterns regarding drugs, sex, intimacy, alcohol, smoking, driving, spirituality, and violence, among other things—all of which can affect their future vulnerability and resiliency. These situations are novel for them, require making discrete choices, and portend largely irreversible consequences (e.g., addiction, pregnancy, severe injury, stigmatization). Adult guidance, even when sought, may not be entirely trusted—especially in nontraditional societies, where part of the 'work of adolescence' is learning to question authority. By contrast, adults who have reached maturity intact often have established response patterns, with trial and error compensating for cognitive limits.

A further source of vulnerability arises when people know less about a domain than they realize. Such overconfidence would reduce the perceived need to seek help, for individuals who lack the knowledge needed for effective problem solving. It could make tasks seem more controllable than they really are, thereby creating a condition for unrealistic optimism. As a result, even if they are just as (un)wary as adults, regarding the magnitude of the risks that they face, teens may make more mistakes because they just do not know what they are doing and fail to realize which situations are beyond their control.

A complementary condition arises when teens do things that adults consider reckless because they, the teens, feel unduly vulnerable. If their world feels out of control, then teens may take fewer steps to manage the situations confronting them. They may also see much worse long-term prospects to the continuity of their lives. If, as a result, they overly discount the future, short-term gains will become disproportionately valuable: There is less reason to protect a future that one does not expect to enjoy. There is also less reason to invest in their personal human capital, by doing homework, acquiring trades, looking for life partners, and even reading novels for what they reveal about chart possible life courses.

Teens might also discount the future if they felt that they might survive physically, but not in a form that they valued. They might fear being so damaged, physiologically or psychologically, that they wanted to enjoy themselves now, while they could and while it would be most valued. In addition to threats to well being that adults would recognize, teens might view adults' lives *per se* as less valued states. Although adults might view such rejection as immature, it could still represent a conscious evaluation. It would be fed by the images of a youth-oriented media and by directly observing the burdens borne by adults (health, economics, etc.).

A final class of threat to the continuity of life is finding oneself physically and mentally intact in a world that seems not worth living in. Such existential despair need not prompt suicidal thoughts to shorten time perspectives. Teens (like adults) might tend to live more for the moment, if they foresaw catastrophic declines in civil society, economic opportunity, or the natural environment—if each were critically important to them. Such discounting could parallel that associated with willingness to risk forfeiting a future that entailed great loss of physical vigor. Studies of unrealistic optimism typically ask participants to evaluate personal risk, relative to peers. One could feel relatively invulnerable, while living in a world that offers little overall promise.

2. *Assessing Actual Vulnerability*

Either overestimating or underestimating personal vulnerability can therefore lead teens to place disproportionate value on short-term benefits, thereby increasing their long-term vulnerability. The costs may be direct (e.g., injury) or indirect (e.g., failure to realize their potential). Adults concerned with teens' welfare have an acute need to understand the magnitude of these risks and convey them to teens (focusing on those places where improved understanding will have the greatest impact) (Schulenberg et al. 1997).

A first step toward achieving these goals is to create a statistical base estimating and tracking risks to young people. Many countries have such surveys which are, indeed, required for signatories to the UN Convention on the Rights of the Child. Estimating vulnerability (for teens or adults) has both a subjective and an objective component. The former requires identifying events so important that they constitute a

threat to the continuity of a life—sufficiently great that people would want to change how they lead their lives, if they thought that the event had a significant chance of happening. This is a subjective act because different teens, and different adults, will define significance (and, hence, 'risk' and 'vulnerable') differently. The UN Convention offers a very broad definition. In contrast, some public health statistics just look at deaths, whereas surveys might focus on the special interests of their sponsors (e.g., pregnancy, early school leaving, drug use).

Collecting statistics is 'objective,' to the extent that it follows accepted procedures, within the constraints of the subjectively determined definition. Different definitions will, logically, lead to different actions (Fischhoff et al. 1981).

Mortal risks to teens in developed countries are statistically very small. In the USA, for example, the annual death rate for 15-year-olds is about 0.04 percent (or 1 in 2,500). Of course, among those teens who died, the probability of going into their final year was much higher for some (e.g., those with severe illnesses), somewhat higher for yet others (e.g., those living in violence-ridden neighborhoods), and much lower for many others (e.g., healthy teens, living in favored circumstances, but victims of freak accidents). If perceived accurately, that probability might pass the significance threshold for some teens, but not for others. Whether it does might depend on whether teens adopt an absolute or a relative perspective (asking whether they are particularly at risk of dying). It might also depend on the time period considered. An annual risk of 0.1 percent might be seen as 2.5 times that of the average teen, or as a 1 percent risk in 10 years or a 2.5 percent risk, either of which could be seen as 'dying young.'

Although probability of death is relatively easy to estimate, it is a statistic that diverts attention away from adolescents. It treats all deaths as equal, unlike 'lost life expectancy' which gives great weight to the years lost when young people die prematurely. It also gives no direct recognition to physical or psychological conditions that might be considered severe enough to affect life plans. Some of these could precipitate early mortality (e.g., anorexia, diabetes, severe depression, cancer), others not (chronic fatigue, herpes). Estimating and weighting these conditions is much more difficult than counting deaths. However, it is essential to creating a full picture of the vulnerabilities facing teens, and which they must assess, in order to manage their lives effectively—within the constraints that the world presents them (Dryfoos 1992).

3. *Reducing Vulnerability*

Adults can reduce adolescents' vulnerability either by changing teens or by changing the world in which they live. Doing either efficiently requires not only assessing the magnitude of the threats appropriately, but also evaluating the feasibility of change. There is little point in worrying about big problems that are entirely out of one's control, or to exhorting teens to make good choices in impossible environments, or expecting behavioral sophistication that teens cannot provide. Where interventions are possible, they should, logically, be directed at the risk factors where the greatest change can be made at the least cost. ('Cost' here could mean whatever resources are invested, including individuals' time, energy, or compassion—as well as money.). Estimates of the effectiveness of interventions aimed at specific risk factors can be found in articles dedicated to particular risks, throughout this section of the Encyclopedia (e.g., drug abuse, depression).

This competition for resources, among different interventions aimed at reducing a specific vulnerability, parallels the competition for resources among vulnerabilities for the attention of researchers and practitioners (the topic of the previous section). Estimating the opportunities for reducing vulnerability raises analogous subjective and objective measurement questions. Providing answers is increasingly part of prioritizing research and treatment. When that occurs, researchers face pressure to analyze their results in terms of effect size (and not just statistical significance); administrators face pressure to justify their programs in such terms.

These pressures can promote the search for risk factors that run across problems, creating multiple vulnerabilities, and opportunities for broadly effective interventions. One such theoretical approach looks at problem behaviors that are precursors of troubled development. As such, they would provide markers for difficult developmental conditions, as well as targets for early intervention. The complementary approach seeks common sources of adolescent resilience, protecting them against adversity (Jessor et al. 1991, Fischhoff et al. 1998).

One multipurpose form of intervention provides skills training. These programs are roughly structured around the elements of the decisions facing teens. They help participants to increase the set of options available to them (e.g., by teaching refusal skills for gracefully reducing social pressures). They provide otherwise missing information (e.g., the percentages of teens actually engaging in risk behaviors). They help teens to clarify their own goals, and how likely those are to be realized by different actions. Delivered in a group setting, they might change teens' environment, by shaping their peers' expectations (Millstein et al. 1993, Fischhoff et al. 1999). Although such programs might focus on a particular risk behavior (e.g., sex, smoking), they teach general skills that might be generalized to other settings (Baron and Brown 1992).

These, and other, programs can be distinguished by the extent to which they adopt a prescriptive or empowering attitude. That is, do they tell teens what

to do or provide teens with tools for deciding themselves? The appropriate stance is partly a matter of social philosophy (what should be the relationship between adults and teens?) and partly a matter of efficacy (which works best?). Whereas the designers of an intervention may be charged with reducing a particular problem (e.g., smoking), teens must view it in the context of the other problems that they perceive (e.g., relaxation, social acceptance, weight gain).

These issues may play out differently across cultures and across time. As a result, this article has not offered universal statements about the scope and sources of adolescent vulnerability, or the preferred interventions. Rather, it has given a framework within which they can be evaluated: What consequences are severe enough to disrupt the continuity of adolescents' lives? How well are they understood by teens and their guardians? What missing information would prove most useful? How well can even the best-informed teens manage their affairs? Answering these questions requires integrating results from diverse studies regarding adolescents and their environments.

See also: Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Mental Health: Community Interventions; Substance Abuse in Adolescents, Prevention of; Vulnerability and Perceived Susceptibility, Psychology of

Bibliography

Baron J, Brown R V (eds.) 1991 *Teaching Decision Making to Adolescents*. L. Erlbaum Associates, Hillsdale, NJ

Dryfoos J G 1992 *Adolescents at Risk*. Oxford University Press, New York

Feldman S S, Elliott G R (eds.) 1990 *At the Threshold: The Developing Adolescent*. Harvard University Press, Cambridge, MA

Fischhoff B, Downs J S, Bruine de Bruin W B 1998 Adolescent vulnerability: a framework for behavioral interventions. *Applied and Preventive Psychology* **7**: 77–94

Fischhoff B, Lichtenstein S, Slovic P, Derby S L, Keeney R L 1981 *Acceptable Risk*. Cambridge University Press, New York

Fischhoff B, Crowell N A, Kipke M (eds.) 1999 *Adolescent Decision Making: Implications for Prevention Programs*. National Academy Press, Washington, DC

Jessor R, Donovan J E, Costa F M 1991 *Beyond Adolescence*. Cambridge University Press, New York

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York

Millstein S G, Petersen A C, Nightingale E O (eds.) 1993 *Promoting the Health of Adolescents*. Oxford University Press, New York

Quadrel M J, Fischhoff B, Davis W 1993 Adolescent (in)-vulnerability. *American Psychologist* **48**: 102–16

Schulenberg J L, Maggs J, Hurnelmans K (eds.) 1997 *Health Risks and Developmental Transitions During Adolescence*. Cambridge University Press, New York

von Winterfeldt D, Edwards W 1986 *Decision Analysis and Behavioral Research*. Cambridge University Press, New York

Weinstein N D, Klein W M 1995 Resistance of personal risk perceptions to debiasing interventions. *Health Psychology* **14**: 132–40

B. Fischhoff

Adolescent Work and Unemployment

In comparison to other postindustrial societies (Kerckhoff 1996), school-to-work transitions are relatively unstructured in North America. Most adolescents are employed in the retail and service sectors while attending high school. In European countries where apprenticeship is institutionalized (Germany, Austria, and Switzerland), adolescent employment is part of well-supervised vocational preparation programs (Hamilton 1990). In the developing countries, many adolescents leave school to take menial jobs in the informal economy.

While nonemployed students may be considered unemployed (if actively seeking employment), they are unlikely to acquire this social identity, since attendance in school normatively constitutes full engagement. High school dropouts are most vulnerable to unemployment, as the labor market favors applicants with higher degrees and strong technical skills. In addition to educational attainment, unemployment is affected by social and personal resources. Deficits in efficacy, work motivation and values, and poor mental health increase the risk. Unemployment, in turn, diminishes these psychological assets even more, and limits the acquisition of information-yielding contacts, engendering further disadvantage in the labor market (Mortimer 1994).

1. The Debate over Adolescent Work

There is lively controversy in the USA (Steinberg and Cauffman 1995) and Canada about whether working causes 'precocious maturity,' drawing youth away from school and developmentally beneficial 'adolescent' activities. The critics contend that employed teenagers are distracted from what should be the central focus of their lives—learning and achieving in school (Greenberger and Steinberg 1986, Steinberg et al. 1993). They argue that employed youth not only will have less time for homework and extracurricular activities, they will also come to think of themselves prematurely as adults and engage in behaviors that affirm this status. Many adolescent workers drink alcohol and smoke, legitimate behaviors for adults, but prohibited by law for minors. Such youth may chafe at what they perceive as dependent, childlike roles, such as that of the student, get in trouble in

school, and move quickly into full-time work. Finally, teenagers who work may be exposed to stressors and hazards which jeopardize their mental and physical health (NRC Panel 1998).

According to a more salutary perspective, working adolescents participate in an important adult social realm (Mortimer and Finch 1996). Although they are unlikely to be employed in the same jobs that they aspire to hold in adulthood, their jobs teach them valuable lessons about timeliness, responsibility, and what constitutes appropriate behavior in the workplace. Participating in the world of work can serve as an antidote to the isolation of young people in schools, from 'real world' adult settings. If employment fosters confidence about being able to succeed in a domain of great significance for the future adult 'possible self,' mental health and attainment could be enhanced.

Employment may encourage psychological engagement in the future prospect of working, interest in the rewards potentially available in adult work, and consideration of the occupations that would fit emerging interests and capacities. Work experiences can thus foster vocational exploration in a generation of youth described as 'ambitious but directionless' (Schneider and Stevenson 1999).

2. *The Empirical Evidence in North America*

Concerns about adolescent employment in the USA and Canada have generated much empirical research. Consistent with the critics' concerns, employment and hours of work are associated with problem behaviors, especially alcohol use (Mortimer et al. 1996), smoking, and other illicit drug use (Bachman and Schulenberg 1993), as well as minor delinquency. Of special interest is whether such problems herald long-term difficulties. In one longitudinal test, youth who worked intensively during high school were compared four years after high school with their counterparts who did little or no paid work. Because the other students had essentially 'caught up,' more frequent alcohol use among the more active workers was no longer manifest (Mortimer and Johnson 1998, McMorris and Uggen 2000).

Some studies find hours of work are linked to lower grade point averages (Marsh 1991); others find no significant association (Mortimer et al. 1996, Schoenhals et al. 1998). Youth who invest more time in work during high school manifest a small decrement in later educational attainment, while at the same time exhibiting more rapid acquisition of full-time work, more stable work careers, higher occupational achievement and earnings (NRC Panel 1998, Carr et al. 1996).

Since adolescents invest substantial time in work, typically about 20 hours per week, it may appear self-evident that academic work would suffer. But this line of reasoning is predicated on the supposition that work and school constitute a zero-sum game, with the amount of time devoted to work necessarily, and in equal measure, detracting from educational pursuits. However, Shanahan and Flaherty (2001) find that most working adolescents combine employment with many other involvements; relatively few focus on work to the neglect of school (or other activities). Other research shows that time spent watching television diminishes when adolescents work (Schoenhals et al. 1998). If adolescents make time to work by lessening their involvement in activities with little educational value, school performance could be maintained with little difficulty.

Instead of a mechanical zero-sum formula with respect to work and educational involvement, the adolescent should be viewed as an active agent, whose goals influence time allocation among diverse activities. In fact, children and adolescents with less interest in school and lower academic performance make more substantial subsequent investments in work and achieve higher quality work experiences than their more academically oriented peers.

Adolescents' self-concepts, values, and mental health, like those of adults, are responsive to work experiences, e.g., learning and advancement opportunities, supervisory relations, and work stressors (Mortimer and Finch 1996). Those enabled to develop skills on the job increase their evaluations of both intrinsic and extrinsic occupational rewards (Mortimer et al. 1996).

3. *Adolescent Work in the Context of Apprenticeship*

In countries that structure the school-to-work transition by apprenticeship, young (age 10–12 in Germany) children are channeled toward an academic track (the Gymnasium), preparatory to higher education (close to a third of the cohort) and professional and managerial occupations, or toward school programs eventuating in a three- to four-year apprenticeship placement (beginning at age 16 and 17) and vocational certification. (While Gymnasium students may hold odd jobs like their North American counterparts, their experience of work is quite unlike those who enter the vocational training system.)

The apprentice spends much of each week working in a firm, and one or two days in schoolwork linked to vocational training (Hamilton 1990). The amount of time spent at work, and the kinds of activities entailed, are set by the structure of the apprenticeship experience, not carved out by each adolescent (and employer) individually, as in North America.

Structured school-to-work experiences offer a context for the exercise of youth agency which is quite different from the unstructured North American setting. The young person's task is to acquire the best

apprenticeship placement, given that future life chances are at stake. Active exploration of the available possibilities has high priority. The fact that the vocational education and training system encompasses a broad range of occupations (in 1996, 498 in all, 370 of which required apprenticeships, and 128 required only school-based vocational education) instills motivation on the part of those who do not enter higher education to do well in school so as to optimize future prospects.

Because school and work experiences are integrated by design, and because the apprenticeship is part of a widely accepted early life course trajectory, there is no concern about young people 'growing up too soon' or getting into trouble as a result of working. On the contrary, the apprenticeship is a legitimate mode of entry to a desirable adult work role (Mortimer and Kruger 2000). Instead of a 'precocious adulthood,' apprenticeship fosters a biographical construction that motivates work effort and serves as a point of reference in evaluating career-related experiences (Heinz 1999). Concerns focus on the availability of sufficient placements for all who seek them, and the adequacy of the system in times of rapid change (Mortimer and Kruger 2000). While apprenticeship provides a bridge to adult work, the system of vocational credentials can restrict individual flexibility (and economic expansion) in a rapidly changing technological environment.

In Germany, only a small minority of adolescents (3–6 percent) participate neither in apprenticeship nor in higher educational preparation. These young people will be subject to high rates of unemployment, as they lack required qualifications in a highly regulated labor market.

4. *Adolescent Work in the Developing World*

In the developing countries, education is key to both fertility control and economic development. Gainful employment and schooling are in direct competition; adolescents (and children) who work typically cannot attend school, and are relegated to adult work in the informal economy, as street traders, day laborers, domestic workers, etc. (Mickelson 2000, Raffaelli and Larson 1999). Adolescents have little or no choice; family poverty propels them into the labor market where they are likely to encounter exploitative and health-threatening work conditions.

Children in the developing world who are not able to find gainful work are unemployed, like their adult counterparts. Indeed, the distinction between adolescent and adult, commonplace in the developed world, lacks clear meaning in a situation in which children must leave school to support their families economically. In these settings, more advantaged youth (especially boys) attend schools that are often more oriented to work opportunities abroad than to the local labor market. Adolescent work in developing societies provides neither an institutional bridge to desirable adult work, nor a source of occupational direction and anticipatory socialization while school and work are combined. While serving immediate economic needs (of the adolescent, the family, and the society), it restricts the acquisition of human capital through schooling that is sorely needed for economic development.

5. *Conclusion*

Adolescent work behavior, including employment and unemployment, must be understood within the broader context of the transition to adulthood. The debate over adolescent employment poses a fundamental question: can youth be incorporated into the adult world of work so as to enjoy the benefits which this exposure can entail without jeopardizing their educational and occupational prospects and placing them at risk? Future research should consider the societal conditions that make the more salutary outcomes more probable.

See also: Adolescence, Sociology of; Adolescent Development, Theories of; Career Development, Psychology of; Childhood and Adolescence: Developmental Assets; Cognitive Development in Childhood and Adolescence; Unemployment and Mental Health; Unemployment: Structural

Bibliography

Bachman J G, Schulenberg J 1993 How part-time work intensity relates to drug use, problem behavior, time use, and satisfaction among high school seniors: Are these consequences or merely correlates? *Developmental Psychology* **29**: 220–35

Carr R V, Wright J D, Brody C J 1996 Effects of high school work experience a decade later: Evidence from the National Longitudinal Survey. *Sociology of Education* **69**: 66–81

Entwisle D R, Alexander K L, Olson L S 2000 Early work histories of urban youth. *American Sociological Review* **65**: 279–97

Greenberger E, Steinberg L 1986 *When Teenagers Work: The Psychological and Social Costs of Adolescent Employment*. Basic Books, New York

Hamilton S F 1990 *Apprenticeship for Adulthood: Preparing Youth for the Future*. Free Press, New York

Heinz W 1999 Job-entry patterns in a life-course perspective. In: Heinz W (ed.) *From Education to Work: Cross-National Perspectives*. Cambridge University Press, Cambridge, UK, pp. 214–31

Kerckhoff A C 1996 Building conceptual and empirical bridges between studies of educational and labor force careers. In: Kerckhoff A C (ed.) *Generating Social Stratification: Toward a New Research Agenda*. Westview Press, Boulder, CO, pp. 37–58

Marsh H W 1991 Employment during high school: Character building or subversion of academic goals? *Sociology of Education* **64**: 172–89

McMorris B J, Uggen C 2000 Alcohol and employment in the transition to adulthood. *Journal of Health and Social Behavior*. **41**: 276–94

Mickelson R 2000 *Children on the Streets of the Americas: Homelessness, Education, and Globalization in the United States, Brazil, and Cuba*. Routledge, London

Mortimer J T 1994 Individual differences as precursors of youth unemployment. In: Petersen A C, Mortimer J T (eds.) *Youth Unemployment and Society*. Cambridge University Press, Cambridge, UK, pp. 172–98

Mortimer J T, Finch M 1996 *Adolescents, Work and Family: An Intergenerational Developmental Analysis*. Sage, Newbury Park, CA

Mortimer J T, Finch M D, Ryu S, Shanahan M J, Call K T 1996 The effects of work intensity on adolescent mental health, achievement and behavioral adjustment: New evidence from a prospective study. *Child Development* **67**: 1243–61

Mortimer J T, Johnson M K 1998 New perspectives on adolescent work and the transition to adulthood. In: Jessor R (ed.) *New Perspectives on Adolescent Risk Behavior*. Cambridge University Press, New York, pp. 425–96

Mortimer J T, Kruger H 2000 Pathways from school to work in Germany and the United States. In: Hallinan M (ed.) *Handbook of the Sociology of Education*. Kluwer Academic/Plenum, New York, pp. 475–97

NRC Panel, Committee on the Health and Safety Implications of Child Labor, National Research Council 1998 *Protecting Youth at Work: Health, Safety, and Development of Working Children and Adolescents in the United States*. National Academy Press, Washington, DC

Raffaelli M, Larson R (eds.) 1999 *Developmental Issues among Homeless and Working Street Youth*. Jossey-Bass, San Francisco

Schneider B, Stevenson D 1999 *The Ambitious Generation: America's Teenagers, Motivated but Directionless*. Yale University Press, New Haven, CT

Schoenhals M, Tienda M, Schneider B 1998 The educational and personal consequences of adolescent employment. *Social Forces* **77**: 723–61

Shanahan M J, Flaherty B 2001 Dynamic patterns of time use strategies in adolescence. *Child Development* **72**: 385–401

Steinberg L, Cauffman E 1995 The impact of employment on adolescent development. *Annals of Child Development* **11**: 131–66

Steinberg L, Fegley S, Dornbusch S M 1993 Negative impact of part-time work on adolescent adjustment: Evidence from a longitudinal study. *Developmental Psychology* **29**: 171–80

J. T. Mortimer

Adolescents: Leisure-time Activities

Adolescents spend many hours per week on various leisure activities—4 to 5 hours in East Asia, 5.5 to 7.5 hours in Europe, and 6.5 to 8 hours in North America. This data on leisure time for adolescents attending school is inversely related to time spent on school work; that is, adolescents in Asian countries do school work for the same length of time as those from North America engage in leisure (Larson and Verma 1999). Overall, leisure time in North America and Europe amounts to about 40 percent of waking hours, more than school and work combined.

By definition, leisure activities are chosen by the young in contrast to obligatory activities and are typically non-instrumental. Although some activities may be wasted time from a developmental standpoint, the majority of free time happens in contexts that are conducive to psychological development.

This paper focuses on time use in adolescence. In a historical perspective, such research is part of the interest shown in daily activities by sociologists, marketing researchers, and governments, since the early nineteen hundreds. The first descriptions of such behaviors regarding adolescence were provided by Barker and Wright (1951) who researched into behavior settings, thus foreshadowing the role of opportunities and individual choices in leisure activities.

1. *Time Spent on Leisure Activities*

Empirical research has distinguished two broad categories: media use (inactive leisure), including watching TV, reading, and listening to music; and active leisure, including working on hobbies, socializing with friends, and playing sports and games. (Where not otherwise stated, data are from a seminal paper on time budget studies by Larson and Verma (1999)).

Concerning media use, the most common activity is watching TV (about 2 hours daily) followed by reading (15 minutes in the US, 40 minutes in Europe and East Asia) and listening to music (about half an hour). TV watching serves as a form of relaxation and default activity when other options are not available, but does not seem to displace other, more socially valued activities. In the 1990s, however, TV watching has been challenged by computer games and new interactive media.

Active leisure concerns physically or mentally active undertakings, such as working on hobbies or socializing with peers. During adolescence, time spent conversing with friends (particularly via the telephone) increases rapidly, and chatting, particularly about the behavior of peers, becomes an important leisure activity (Silbereisen and Noack 1988). Such social interactions seem to be essentially spontaneous and mainly self-regulated.

In terms of more structured and adult-supervised leisure, such as participation in organizations and athletics, large differences exist between North America, Europe, and East Asia. Sports amount to at least one hour in the USA, compared to about half an hour per day or less in the other countries, with a declining trend across adolescence. The data for playing music and other structured activities (typically in groups) show the opposite national differences. Seen in a developmental framework, the degree of adult super-

vision inherent in the activities and contexts diminishes with age, opening up a vista for activities more or less self-chosen by the young and their peers (Hendry 1983).

2. *Company During Leisure Activities*

The main categories of companionship are family and peers. However, about 25 percent of waking hours is spent alone in their bedroom, which is typically a private space decorated with trophies signifying their emerging sense of self. Favorite leisure activities are listening to music, reading magazines, watching videos, and daydreaming.

The time spent with family declines from childhood through adolescence, parallel to the growing duties in school. Whereas in the US adolescents spend about 15 percent of waking hours with their family, in East Asia it is almost 40 percent. Such differences point to the role of cultural values: cohesion within the family rather than individualistic goal pursuit is central to collectivist orientations, and leisure within the vicinity of the family may be a reflection of the broader value systems.

Being together with peers is often associated with adolescent leisure. Indeed the figures for the US and Europe show that adolescents spend up to 30 percent of non-school waking hours going out to parties, attending discotheques, and other away-from-home activities (compared to less than half that for East Asia).

Time spent dating is in the order of one hour per day among adolescents in Europe and North America (Alsaker and Flammer 1999), whereas data on East Asian samples are close to one hour per week, again reflecting higher family control and lower appreciation of western-style self-regulated behaviors. On average, the company of other-sex peers in the US and Europe represents about twice the time adolescents spend with family. As peers select themselves on the basis of shared focal attributes (school achievement, substance use, etc.) and also have a mutual socialization influence, this large amount of time is particularly interesting for developmental consequences.

According to Raymore et al. (1999) leisure activities can be clustered into four groups that show only slight differences between genders. The 'positive-active' group is especially engaged in socially valued activities, like volunteering for community projects. Adolescents in the 'risky' group are more likely to do things for kicks (including substance use) and hang out with friends. The 'diffused' group spends little time in any activity, and thus has no clear preferences. The 'home-based' group applies predominantly to females and involves activities at home with family (including TV watching), whereas males engaged in frequent sports-related activities are over-represented in the 'jock' group. Across the transition to early adulthood, most of these clusters show remarkable stability (about 40 percent of the individuals remain in the same cluster).

However, leisure activities can also change dramatically. Social change such as the breakdown of the socialist countries in Europe is a case in point. The entire system of state-run youth organizations, which were major promoters of structured leisure activities, broke apart. As replacements came in slowly and were often commercial in nature, many lost their social networks and meeting places. Such events partly explain the upsurge of violent peer groups in former East Germany.

3. *Consequences for Psychosocial Development*

In a theoretical framework, leisure activities (particularly the active type involving peers) are an example of active genotype-environment correlation, where individuals seek out opportunities that match their personal propensities. Utilizing a twin design, Hur et al. (1996) analyzed the relative influence of genetic and environmental conditions on leisure time interests (which may differ from actual behavior). Whereas sports, music, and arts showed a strong genetic influence, TV viewing, dating, and various kinds of social activities were characterized by strong shared environmental influences. The latter points to the role of contextual opportunities. Moreover, the effect of shared environment was larger in adolescence than in young adulthood, indicating the new freedoms gained.

With regard to media use in particular, public concern is typically related to the exposure of adolescents to certain contents considered detrimental (sexuality plays a high role in TV programs and pop lyrics). However, as the decline in TV watching with age seems to indicate, such contents may not affect adolescents' behavior. Violent TV programs and videos are also often discussed because of their violence and their possible role in the development of aggression. Although the causal nexus is difficult to assess, studies in various nations show that exposure to media violence is prospectively related to aggression (Wartella 1995).

Media are also major forums for information and participation in popular culture that serve important roles in identity formation. Listening to music helps to forge important elements of one's identity; demonstrating shared preferences with others through behavioral style and outfit accessories helps location within an emerging social network.

A number of other functions of adolescent media use can be distinguished. Often it simply helps to make fun activities even more fun (e.g., driving around in a car with blasting rock music) or fulfills youths' propensity for high sensation (e.g., listening to heavy metal music). Media use may also serve as a general purpose coping strategy for calming down or over-

coming anger. Finally, media are used to connect to networks of peers sharing similar idols and values (Arnett 1995).

Some of the genres of music consumed by adolescents (e.g., rap, soul, heavy metal, and pop) represent the core of their taste culture as well as the more general youth culture. In psychological terms, two focal themes stand out—the expression of defiance toward authority and of love found or lost. Usually, the young select themselves into real or imaginary interactive groups, thereby achieving not only mood management (making bad things less bad, and good things even better), but also status and distinction as a cultural elite (in their own understanding) among peers (Zillman and Gan 1997). This is accomplished and reinforced by a set of other identity-providing attributes, such as particular attire, hairstyle, accessories, and mannerisms. Although often disruptive for psychosocial adaptation in the short term (e.g., acceptance of violence after exposure to rock music videos laden with defiance to authority), long-term negative effects are rare. This is probably due to the transitional nature of participation in such groups and their role in identity development.

The most relevant aspect of active leisure developmentally is the fact that such activities require initiative, planning, and organization of place, time, and content. The individuals themselves have to exercise control over their actions and regulate their emotions, and all this is accomplished in the company they choose. Adolescence in general is the time when individuals take the initiative concerning their own psychosocial growth, and consequently one should assume that leisure is purposively chosen to pursue development among other issues. Indeed, research in various countries has shown that adolescents entertain clear, age-graded conceptions of what they want to achieve, that they look for leisure locales suitable to pursue such goals, and that they are successful in this regard. Silbereisen et al. (1992) and Engels and Knibbe (2000) showed that adolescents who perceive a discrepancy between their current and hoped-for future state of romantic affairs seem to select leisure settings, such as discotheques, because they offer opportunities for contacts with the other gender. Moreover, once the mismatch between current and future state is resolved, their preferences change again, this time for more private encounters.

Certainly leisure activities do not affect all the developmental tasks of the second decade of life in the same ways. However, experiences in leisure have a carry-over effect to other arenas, such as occupational preparation and socialization. Research on entrepreneurs has shown that they took responsibility for others in out-of-school contexts from an early age (Schmitt-Rodermund and Silbereisen 1999). Thus, more structured active leisure, such as sports, may be more relevant for the future world of work, since they are organized by goals and standards of performance, include competition, and demonstrate that planned effort is effective.

The development of individual and collective agency seems particularly to profit from engaging in structural leisure activities. Heath (1999) reported changes in adolescents' conversations about their activities once they had entered youth organizations. Remarkably, they referred more often to issues such as monitoring, goal achievement, or adjusting their behavior. Positive effects on self-esteem and school achievement were also reported as an outcome of participation in civic activities (Youniss et al. 1997). However, some leisure activities may have negative long-term implications. Some team sports, for instance may foster accentuated gender roles and lead to excessive alcohol use (Eccles and Barber 1999).

4. *Future Directions*

In closing, time-budget studies have shed light on how, and to what effect, adolescents spend their leisure. Nonetheless, future research needs to address further the particular psychological qualities of the activities-in-context. Without such information, research on the selection of leisure experiences and the investigation of consequences for psychosocial development lack an understanding of the mediating links. New leisure activities (such as interactive media) and the role of societal change in general, represent another focus of much needed research.

See also: Adolescent Behavior: Demographic; Adolescent Development, Theories of; Leisure and Cultural Consumption; Leisure, Psychology of; Leisure, Sociology of; Media, Uses of; Popular Culture; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Sports, Psychology of

Bibliography

Alsaker F D, Flammer A 1999 *The Adolescent Experience. European and American Adolescents in the 1990s*. Erlbaum, Mahway, NJ

Arnett J J 1995 Adolescents' uses of media for self-socialization. *Journal of Youth and Adolescence* **24**: 519–33

Barker R G, Wright H F 1951 *One Boy's Day*. Harper and Row, New York

Eccles J S, Barber B L 1999 Student council, volunteering, basketball, or marching band: What kind of extracurricular involvement matters? *Journal of Adolescent Research* **14**: 10–43

Engels R C M E, Knibbe R A 2000 Alcohol use and intimate relationships in adolescence: When love comes to town. *Addictive Behaviors* **25**: 435–39

Heath S B 1999 Dimensions of language development: Lessons from older children. In: Masten A S (ed.) *Cultural Processes in Child Development: The Minnesota Symposium on Child Psychology*. Erlbaum, Mahwah, NJ, Vol. 29, pp. 59–75

Hendry L B 1983 *Growing Up and Going Out: Adolescents and Leisure*. Aberdeen University Press, Aberdeen, UK

Hur Y-M, McGue M, Iacono W G 1996 Genetic and shared environmental influences on leisure-time interests in male adolescents. *Personality and Individual Differences* **21**: 791–801
Larson R W, Verma S 1999 How children and adolescents spend time across the world: Work play, and developmental opportunities. *Psychological Bulletin* **125**: 701–36
Raymore L A, Barber B L, Eccles J S, Godbey G C 1999 Leisure behavior pattern stability during the transition from adolescence to young adulthood. *Journal of Youth and Adolescence* **28**: 79–103
Schmitt-Rodermund E, Silbereisen R K 1999 Erfolg von Unternehmern: Die Rolle von Persönlichkeit und familiärer Sozialisation [Entrepreneurial success: The role of personality and familial socialization]. In: Moser K, Batinic B (eds.) *Unternehmerisch erfolgreiches Handeln*. Hogrefe, Göttingen, Germany, pp. 116–43
Silbereisen R K, Noack P 1988 On the constructive role of problem behavior in adolescence. In: Bolger N, Caspi A, Downey G, Moorehouse M (eds.) *Persons in Context: Developmental Processes*. Cambridge University Press, Cambridge, MA, pp. 152–80
Silbereisen R K, Noack P, von Eye A 1992 Adolescents' development of romantic friendship and change in favorite leisure contexts. *Journal of Adolescent Research* **7**: 80–93
Wartella E 1995 Media and problem behaviors in young people. In: Rutter M, Smith D J (eds.) *Psychosocial Disorders in Young People: Time Trends and their Causes*. Wiley, New York, pp. 296–323
Youniss J, Yates M, Su Y 1997 Social integration: Community service and marijuana use in high school seniors. *Journal of Adolescent Research* **12**: 245–62
Zillmann D, Gan S 1997 Musical taste in adolescence. In: Hargreaves D J, North A C et al. (eds.) *The Social Psychology of Music*. Oxford University Press, Oxford, pp. 161–87

R. K. Silbereisen

Adoption and Foster Care: United States

1. Introduction

In the year 2000, more than $20 billion in federal, state, and local tax dollars was spent on public child welfare services in the USA. This does not include the hundreds of millions in private dollars (e.g., united ways, religious charitable agencies and organizations, foundations and privately arranged adoptions both in the US and abroad) that was spent on such services. Clearly, child welfare is a big business in the USA. This article focuses on two fundamental and important child welfare programs, adoptions, and foster care.

2. Trends in Foster Care

As indicated in Fig. 1, the rate of children entering the foster care system has been increasing steadily over the past decades (Schwartz and Fishman 1999). The most recent statistics indicate that at the end of the Adoption and Foster Care Analysis Reporting System's (AFCARS) six-month reporting period,

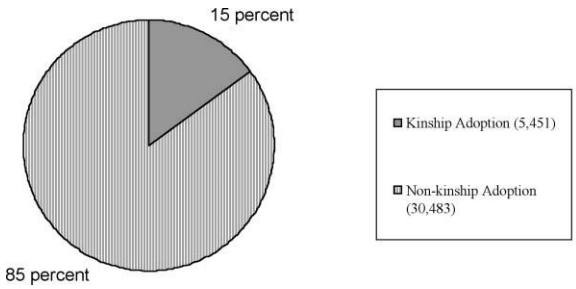

Figure 1
Foster children adopted by type of placement, 1998 (Total: 36,000). Source: US Department Health and Human Services, Adoption and Foster Care Analysis and Reporting System [www.acf.dhhs.gov]

there were 547,000 children in foster care in March 1999. It is generally agreed that this number is unacceptable and efforts are being made to reduce rising foster care caseloads. Initially, policies such as the Adoption Assistance and Child Welfare Act of 1980 encouraged family preservation with the expectation that children would not need to enter the foster care system or would return quickly to their families if those families were given supports (e.g., family services, counseling, parenting education). The law, however, was vague about how long services should be provided or when to make the determination that a child could not be returned to her/his family. As a result, a growing number of children either languished in foster care or entered the system as older children—making it more difficult for those children to get adopted. New policies that seek to arrest the rate of foster care caseloads recently have been implemented that mandate more timely permanent placements (preferably adoptions) for children in foster care and encourage both formal and informal kinship foster placements.

3. Current Responses

3.1 Adoption and Safe Families Act

As the foster care system grew to unmanageable proportions, the government looked for ways to ameliorate the situation. The Adoption and Safe Families Act, signed into law in November 1997, established a new benchmark for child welfare policy. Among other things, the new law authorized incentive payments to the States for increasing the number of foster children adopted in any given year. The law also required States to document their adoption efforts, lifted geographic barriers to cross-jurisdictional adoptions, and changed the timeline for permanency hearings from 18 to 12 months. Although the law continued and even expanded family preservation efforts under the auspices of the Safe and Stable

Families Program, it also mandated that States initiate termination of parental rights and approve an adoptive family for any child in the foster care system longer than 15 months. States are still required to make reasonable efforts to preserve and reunify families, but the new law makes exceptions to this requirement in cases where parents have been found guilty of chronic abuse, sexual abuse, murder of another child, felony assault resulting in bodily injury to a child, or termination of rights to a sibling (Child Welfare League of America 2000, Christian 1999).

By the end of 1998, 38 States enacted ASFA-related legislation. Since continued funding is contingent on such legislation, all States are expected to eventually comply. Although there are no specific plans to evaluate the impact of the Adoption and Safe Families Act, some level of review may be possible through the new Federal Monitoring Program, implemented in January 2000 and regulated by the Department of Health and Human Services (Testimony before House Ways and Means Committee 2000). This monitoring process is designed to assess the effectiveness and efficiency of the child welfare system overall by focusing on concrete outcomes.

3.2 President's Initiative on Permanency Planning

The Department of Health and Human Services, the Administration for Children and Families, the Administration on Children, Youth and Families, and the Children's Bureau recently collaborated on the development of an adoption and foster care initiative designed to promote State governance of permanency planning for children. The initiatives developed guidelines designed to assist States in their efforts to reform and revitalize their child welfare systems. Permanency, as used in the guidelines, means that 'a child has a safe, stable, custodial environment in which to grow up, and a life-long relationship with a nurturing caregiver' (Duquette et al. 1999). The initiative recommends that support services be put in place as soon as possible after a child enters State care and that these services constitute a reasonable effort to rehabilitate families for purposes of permanent reunification. The initiative waives the 'reasonable effort' requirement, first mandated by the Federal Adoption Assistance and Child Welfare Act of 1980, in cases where the parent is convicted of committing murder or specific crimes against children, where parental rights have previously been terminated, where children have been abandoned or severely abused, or when the parent voluntarily refuses services.

The guidelines regarding reasonable efforts reinforce the centrality of permanency in the new child welfare paradigm by insisting that such efforts include concurrent planning leading toward timely final placement. The initiative explicitly identifies adoption as the preferred method of permanent placement for those children who cannot be raised by their biological parents. The guidelines also support court-approved post-adoption contact agreements between adoptive and birth parents, but emphasize that such agreements do not endanger the irrevocability of the adoption contract itself. If adoption, for whatever reason, is not possible, the remaining options are identified as permanent guardianship, standby guardianship, and planned long-term living arrangements with a permanent foster family. Permanent guardianship is used primarily for children of 12 years or older and placement with a relative is preferred, particularly if the child has been in care of the relative for at least one year. Standby guardianship is used in cases where the birth parent is chronically or terminally ill. Long-term permanent foster care is the least-preferred method of placement; the guidelines support its use only in cases of children with serious disability and then only with older, established foster families.

3.3 Privatization

Although State contracting for delivery and/or management of child welfare services is not a new phenomenon, the original idea of partnership with non-profit organizations has been expanded to include contracting with for-profit companies. The child welfare system traditionally limited its use of private services, but a new trend toward Statewide privatization using a managed care approach was initiated in 1996 when Kansas began the process of privatizing its foster care system (Petr and Johnson 1999). The Kansas approach features a capitated, fixed-price pay system that promotes competition and operationalizes performance standards to enhance accountability (Eggers 1997). Attracted by the cost-saving potential of private enterprise and the possibility of improved performance, other States soon implemented their own privatization efforts. The Department of Social Services in Michigan partnered with the non-profit sector to reduce the length of time children wait for adoption to be finalized and North Dakota attributes its status as the State with the highest adoption rate in the country to its use of private contractors (Poole 2000). The State initiatives have been evaluated, but opinion remains divided as to actual outcomes. For every report detailing successful use of privatization methods and public-private partnerships, there is another report indicating failure. Part of the problem lies in the difficulty of establishing standard outcome measures, but work by the Child Welfare League and other child-focused agencies may eventually solve the problem and lay the foundation for rigorous outcome-based evaluations.

3.4 Kinship Care

Although research is inconclusive on whether children placed with relatives fair any better in terms of

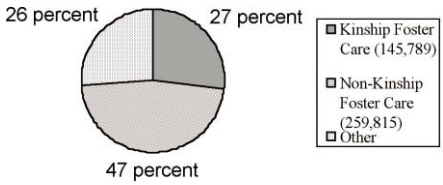

Figure 2
Children in foster care by type of placement, March 1999 (Total: 547,000). Source: US Department Health and Human Services, Adoption and Foster Care Analysis and Reporting System [www.acf.dhhs.gov]

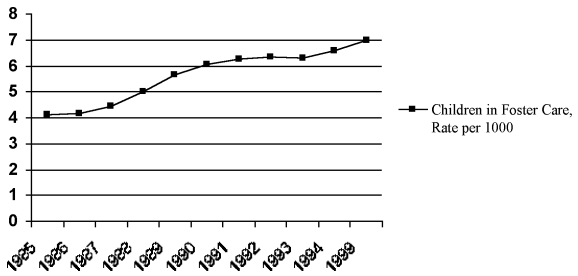

Figure 3
Rate per 1,000 of children in foster care, 1985–94 and 1999. Source: Schwarz and Fishman 1999

behavioral and emotional outcomes than children placed with non-relatives (Iglehart 1994, Berrick et al. 1994), the number of relatives fostering children is steadily increasing (Berrick 1998). As indicated in Fig. 2, 27 percent of the children in foster care in March 1999 were living with relatives. While kinship families often provide stable placements, they tend to stop short of pursuing legal guardianship or adoption of their foster children. Many kinship caregivers are reluctant to consider adoption because they consider themselves 'family' already (Berrick et al. 1994). In fact, adoption rates for children placed with relatives are lower than adoption rates for children placed with non-relatives (Berrick 1998). This is apparent in Fig. 3, which indicates that in 1998, only 15 percent of the adoptions of foster children were kinship adoptions. Policy initiatives that support permanent adoptions by kinship caregivers both culturally and economically are needed to ensure that children have the option of being raised among capable and concerned family members.

4. *Conclusion*

As this article indicates, public child welfare services in the USA are in the midst of change. Increased emphasis is being placed on insuring safety and permanency for abused and neglected children. To this end, there is a gradual movement toward kinship care and privatizing child welfare services, introducing such private sector concepts as incentives, performance indicators, and the use of technology into the system.

See also: Child Abuse; Child Care and Child Development; Childhood Health; Children and the Law; Dissolution of Family in Western Nations: Cultural Concerns; Family, Anthropology of; Family as Institution; Lone Mothers in Affluent Nations; Lone Mothers in Nations of the South; Partnership Formation and Dissolution in Western Societies; Repartnering and Stepchildren

Bibliography

Berrick J D, Barth R P, Needell B 1994 A comparison of kinship fosters homes and foster family homes: Implications for kinship foster care as family preservation. *Children and Youth Services Review* **16**: 33–63

Berrick J D 1998 When children cannot remain home: Foster family care and kinship care. *The Future of Children* **8**: 72–87

Child Welfare League of America 2000 Summary of the Adoption and Safe Families Act of 1997. www.cwla.org/cwla.hr867.html

Christian S 1999 1998 State Legislative Responses to the Adoption and Safe Families Act of 1997. *Report to the National Conference of State Legislatures*. www.ncsl.org/programs/CYF/asfaslr.htm

Duquette D N, Hardin M, Dean C P 1999 *The President's Initiative on Adoption and Foster Care. Guidelines for Public Policy and State Legislation Governing Permanence for Children, Children's Bureau*.www.acf.dhhs.gov/programs.cb/publications/adopt02/index.htm

Eggers W D 1997 There's no place like home. *Policy Review* **83**: 43–7

Golden O A 2000 Testimony on the Final Rule on Federal Monitoring of State Child Welfare Programs. *Testimony before House Ways and Means Committee* February 17th, 2000. www.hhs.gov/progorg/asl/testify/t000217b.htm

Iglehart A P 1994 Kinship foster care: Placement, service, and outcome issues. *Children and Youth Services Review* **16**: 107–22

Petr C G, Johnson I C 1999 Privatization of foster care in Kansas: A cautionary tale. *Social Work* **44**(3): 263–7

Poole P S 2000 Privatizing child welfare services. Models for Alabama. www.alabamafamily.org/pubs/privchild.html

Schwartz I M, Fishman G 1999 *Kids Raised by the Government*. Praeger Publishers, Westport, CT

I. M. Schwartz, S. Kinnevy, and T. White

Adorno, Theodor W (1903–69)

In a letter from the 1940s in which Thomas Mann asked Theodor W. Adorno to characterize his origins, Adorno wrote: 'I was born in Frankfurt in 1903. My father was a German Jew, and my mother, herself a singer, is the daughter of a French officer of a Corsican, but originally of Genoese descent and a German

singer. I grew up in an atmosphere completely steeped in theoretical (also political) and artistic, but above all musical interests'. Like many other members of the generation of Jewish intelligentsia born at the turn of the twentieth century, his mental disposition also formed in resistance to the assimilated Christian-convert parental home. The philosophically coded motifs of Jewish mysticism and theology strewn throughout his work originate from this.

Theodor Wiesengrund Adorno the individual eludes every classification into academic field or discipline. He was no ordinary run-of-the-mill citizen of the scholar republic, but as Habermas notes, 'an artist amongst civil servants.' Adorno was a composer, music theorist, literary theorist and critic, philosopher, social psychologist, and last but by no means least, a sociologist. In the early years of his academic life, it could not have been foreseen that he would become a world-famous sociologist. In 1925 he moved to Vienna to study theory of composition with Alban Berg. But instead of becoming a composer by profession, he returned to Frankfurt in 1927, in order to write his 'habilitation'-thesis. In these years his relationship with the discipline of sociology was not free from resentment. Indeed, his evaluation of the subject in a review of Karl Mannheim's works in the early 1930s seemed almost formed out of contempt. At that time he placed 'sociology' on a par with a sociology-specific lesson in ideology, that is to say, a formalistic consideration of cultural content which is not concerned with the mental substance of the analyzed person.

Adorno's vivid rhetoric about the sociologist as a 'cat burglar' is also notorious. The sociologist—so the image goes—feels his way exclusively around the surface of social architecture, whereas only the philosopher is capable of decoding its ground plan and inner structure. If one further adds Adorno's later position in the positivism debate to this biographically early perception of sociology, one might form the impression that he tackled the subject with nothing but disdain. But such stereotyping would be unfair to Adorno.

In 1933 after Hitler's rise to power, he went to England. But different from his colleagues who definitely left Germany and immigrated to the USA, Adorno kept returning to Germany various times until it got too dangerous. He joined the Institute for Social Research in New York in 1938.

For sociologists, his name is associated with three great empirical studies, each one of which was, in its own right, pioneering in its respective field of research. In the late 1930s, he worked on the 'Radio Research Project' led by Paul Lazarsfeld which was to establish the modern field of media research. In the 1940s he moved together with Max Horkheimer to California, USA. They started to write the book *Dialectics of Enlightenment*, which only appeared after the end of the war. At the same time he worked within a team of sociologists on the famous study *Authoritarian Personality*, a classic of prejudice research even today. In 1949 Theodor W. Adorno and Max Horkheimer returned to Frankfurt, in order to reopen the *Institut für Sozialforschung*.

In the 1950s, he inspired the study entitled *Group Experiment*, one of the first studies into the political consciousness of West Germans. While there were still conflicts between Lazarsfeld and Adorno over the latter's criticism of a purely quantitative orientation of research, Adorno set out in both of the aforementioned seminal socio-psychological studies his own qualitative methods of analysis which showed a close proximity to phenomenological approaches. However, Adorno's contribution to sociology did not merely amount to these empirical and methodological works.

In a work written after his return from exile, Adorno revealed an astounding knowledge of the sociology he seemingly scorned, the history of its dogma from Comte to Marx, and its contemporary German and American proponents. The latter he knew from his period of emigration, in the most part personally through his active involvement in the committee of the German Sociological Association.

He proudly accepted his nomination to the chairmanship in 1963; and in 1968, at the Sociologists' Day held in Frankfurt, and against the backdrop of the student protests, he gave the closely followed key note lecture 'Late Capitalism or Industrial Society.'

Adorno's contributions to a critical theory of society certainly had an extensive cultural-scientific effect reaching far beyond the sole discipline of sociology. The *Dialectic of Enlightenment*, written together with the philosopher Max Horkheimer that first appeared in a small number of copies in Amsterdam in 1947, enjoyed a paradigmatic status, but its effect was only really felt in the 1970s.

In the *Dialectic of Enlightenment*, Horkheimer and Adorno construe the beginning of history as a *Fall of Man*—as mankind's breaking out of its context in nature. For them, the evolution of mankind's treatment of nature is—contrary to almost the entire bourgeois and socialist thought tradition—not the road to guaranteed progress but the well-beaten track to a regression of world history. In the *Dialectic of Enlightenment*, they attempt to expose this track by means of a paradoxical figure of thought. The development of the history of mankind is, in the current meaning of the phrase, 'originating in nature'—that is, invisible and heteronomous—for as long as mankind cuts itself off from the consciousness of it own naturalness. Horkheimer and Adorno conceive the course of the history of mankind with the psychoanalytical motif of the return of the repressed. It is in the catastrophic evolution of history where irreconcilable nature seeks vengeance. The central connecting theme of this development is professed through perverted reason which, cut off from its own basis in

nature, can only get hold of itself and its object in instrumentally limited identifications. The critique of this identification principle is the system-philosophical center of Adorno's theory of society.

It is in this way that he criticizes a form of cognition which brings to the phenomena being perceived a medium which is external and conceptual to them, and which within this medium only pretends to 'identify' them. He criticizes a societal type of work which denies individuals the development of a relationship to the self exactly because it compels them to 'dispose of' their labor in the medium of exchange value. He criticizes the identity compulsion of political institutions that gain their false stability precisely by means of their intolerance of subjective differences. He criticizes a form of socialization and upbringing, which demands of individuals a biographical consistency, which is external to their naturalness.

It should be remembered when fixing Adorno's socio-theoretical reflections in philosophical terms, so as not to form the wrong impression, that they conform to a theory in the sense of a network of empirical hypotheses which can be reconstructed by discursive means.

The historicist motif of identifying the invisible power of integration of the avenging domination of nature in the economic, socio-psychological, and cultural forms of society therefore forms the basis of Adorno's fully developed theory of society. He is helped in this plan by the image—construed in terms of an ideal type—of liberal market capitalism acting as a foil. Against this backdrop, the form of social integration that became obviously totalitarian in National Socialist society came to the fore. In critical theory in general, as in Adorno's work in particular, this image of society consists of three socio-theoretical components: (a) a political economy theory of totalitarian state capitalism; (b) a socio-psychological theory of authoritarian character; (c) the theory of mass culture.

(a) Under the term 'state capitalism,' the economist Friedrich Pollock subsumes the structural characteristic of a new politico-economic order as formed in Germany in the 1930s. In 'state capitalism,' the liberal separation of the political and economic spheres is done away with. The heads of the major companies become subordinate government agencies. Through state terror, worker organizations are robbed of their rights to the representation of their interests and forcibly incorporated into the planned system. The ruling power in state capitalism is a political apparatus that emerges from the fusion of state bureaucracy and the heads of major companies. In this apparatus, plutocratic exploitation interests and political ruling interests become intermingled to form a closed cyclical system.

The concept of 'state capitalism'—from the point of view of a Marxist crisis theory—was to help explain the paradoxically emerging phenomenon of 'organized capitalism.' It was meant to show how capitalism could, through the installation of 'planned economic' elements delay its own end.

At that time (1943) Adorno even went so far as to see in the post-liberal economic structure denoted by the term 'state capitalism' a symbol for the primacy of politics over the economy—of course, under the conditions of, and at the price of, totalitarianism. This acceptance conformed to the historicist image of a prehistory of domination which was coming to an end and which would only be slowed down by the interlude of liberal capitalism.

(b) On the basis of Erich Fromm's empirical analyses and Max Horkheimer's theoretical considerations, Adorno presupposed—for Western societies in general and for German society in particular—the decline of a bourgeois social character which had, for a short historical period of time, made the formation of autonomous individuals possible. From this point of view, classic bourgeois society had allowed at least some of its male members the cultivation of an ego identity which was only prepared to accept such limitations of freedom which themselves appeared necessary in the light of rational inquiry. On the other hand—so the theory goes in short—the conditions for the formation of the individual in the late bourgeois age now only produce forms of dumb obedience to social and political claims to power.

Adorno, in his *Studies on Fascist Propaganda Material* and in his critique of psychoanalytical revisionists, once again reformulated the changed post-liberal conditions for socialization in the terms of psychoanalytical socialization theory. He showed that, as the extra-familial forms of social authority land directly—i.e., no longer mediated via the socialization achievement of the father—on the child, the socially necessary achievement of the establishment of conformity is accordingly no longer performed by the 'mediating power of the ego.' The dominating directness of social obligation now takes the place of the conscious reflexive ego-function. The domination of society over the individual—or rather within the individual—does not, for Adorno, limit itself to the weakening of the ego-function, described as the unleashing of the super-ego. The socially produced weakening of the former set dynamism in motion in the psychological apparatus of the individual which the very deprivation of power of the ego strengthens further. The ego which is overpowered by the super-ego tries to save itself by virtue of an archaic impulse of self-preservation through a libidinous self-possession which is characteristic of the pre-Oedipal phase.

This narcissistic regression manifests itself in irrational desires for fusion with the social power that is now no longer parental, but direct and anonymous. The conscious achievement of conformity of the ego is thus threatened from two unconscious poles simultaneously—from the unmediated super-ego and the identification with the aggressor. In this way, Adorno

attempted, in the terms of Freud's theory of personality, to explain how the National Socialist regime was successful in that the imposition of an unlimited reality principle could still be experienced by the subjugated subjects with such zest.

(c) Adorno also describes the peculiarity of late capitalist culture—like other critical theorists—through the stylized juxtaposition of high bourgeois and late bourgeois eras. In the former, art still offered the opportunity for productive leisure in which the bourgeois could, through the enjoyment of art, rise above the business of everyday life. The classic bourgeois works of art portrayed the utopia of human contact, which bourgeois everyday life certainly belied. Art was consequently ideological because it distracted from the realization of a truly human society; at the same time it was the form in which bourgeois society presented the utopian images of better opportunities under the appearance of beauty.

In late bourgeois culture, especially manifest in National Socialist cultural politics and in American mass culture (which Adorno was experiencing at first hand at that time), the precarious unity of utopian and ideological moments, which bourgeois culture still kept a firm hold of, was in decline. 'National' art and the consumer products of mass culture serve—albeit with different strategies—the sole purpose of ideological integration. This integrative directness of late capitalist culture is the end result of an ousting, over two centuries, of pre-capitalist remnants from artistic production. As the capitalist logic of exploitation also takes hold of culture, it reveals itself as that which it has been since the prehistoric origins of the instrumental mind—throughout all pre-modern historico-cultural epochs—that is, a medium for a dominating safeguard of conformity.

After the interlude of autonomous bourgeois art in the liberal capitalist epoch, the utopian function granted to aesthetic culture could only be kept intact in art forms which systematically elude the maelstrom of mass communication by virtue of their esoteric conception.

These are no longer utopian in the sense of a positive representation of unseized opportunities, but rather as a negative censure, as a 'wound,' which is meant to call to mind the irreconcilable condition of the social world.

These were—according to differences specific to the fields of political economy, social psychology, and culture—the main thematic strands of a theory which starts from a natural history of domination which came to an end in the Nazi system. At the Fascist end of history, the integrity of nature, violated in prehistory, takes its revenge in totalitarian social integration, that is, one in which all control of rational subjects is lost. An exchange rationality which becomes totalitarian and hermetic, a socializing structure which embeds the claim of authority of a power which has become anonymous in the ego-structure of the subjects themselves, and an industrially fabricated mass culture which serves the sole purpose of manipulative rectification, all combine together to form the terrifying image of a perfectly and systemically integrated society. This terror, set out in the theory, also had a determining influence on Adorno's socio-theoretical reflections when their immediate contemporary historical cause had lapsed due to the military defeat of National Socialism.

The inner architecture of Adorno's theory of society is, in spite of its anti-systematic claim, of a suggestive conciseness. Its consistent dominating functionalism is even in the twenty-first century capable of causing sparks to fly, the flashes from which are lighting up the problematic aspects of modern societies. It is nevertheless difficult to ascribe to the theory a direct, comprehensive, and contemporary diagnostic relevance. This comes as a result of the basic assumption—set out in the *Dialectic of Enlightenment* and popularized in Herbert Marcuse's *One Dimensional People*—that modern societies which have been shaped by scientific and technical developments have fallen into entropic systems, that is, into systems which are incapable of overcoming their own status quo.

In the light of contemporary developments, this assumption of hermetic unity, of total integration, no longer proves adequate. Present-day societies can better be described in a perspective of disintegration. This not only means the dramatic end of the post-war period with the collapse of Communist imperium. It also alludes to the foreseeability of further ecological disasters, diverse political lines of conflict and social fractures that globalized capitalism leaves in its wake, and the decline of nation state and bureaucratic organizational structures. Also worth mentioning are changes in the world of work with the expansion of the service sector, the flexible forms of industrial rationalization, and above all the unforeseeable consequences of biotechnology on the life world, as well as cultural changes such as the emergence of post-materialist, i.e., hedonistic and participation-oriented value-orientations and the erosion of traditional norms, role models, and modes of living. All of these developments point to a 'society of disintegration,' i.e., societies which—in system-theoretical terms—are not in a position to control their own environment.

Whoever did not have the luck to have experienced Adorno in person will perhaps find it paradoxical that a critic of society like Adorno lived and worked in post-war West German society and, to a large extent, was able, through public statements and radio lectures, and through the education of his students, to influence them politically. Even though this is, however, completely out of the question according to his theory of a hermetically fixed status quo.

If one looks at his *Introduction to Sociology*, the last lectures Adorno delivered, this contradiction loses much of its sharpness. One may accuse these lectures of having a lack of philosophical depth. But they make

up for this lack with a sociological depth which acts as a complement to the sociological pallor of Adorno's philosophical theory of society which actually hardly gets beyond a hermetic functionalism.

It is in fact a dialectic version of the integration of a society via exchange that paved the way to a productive development of critical theory. In the lecture *Introduction to Sociology*, the positions are already clearly marked out where a short time later Juergen Habermas, Claus Offe, and others would break up the orthodox version of the theory of an exchange society. Dialectic here means a socio-theoretical view of integration, according to which society does not merely reproduce itself as a system, that is, behind the backs of individuals, but also reproduces itself through them. Critical social research has to establish itself at the sources of friction of that which Habermas later called 'system' and 'life world.' Adorno already sets out a post-Marxist view in Marxian terms in his second lecture. Unlike in Marx's time, when the key starting-point had in fact been the forces of production, the starting-point here is, under the conditions of contemporary late capitalism, the primacy resides in the relations of production, i.e. their political mediation. It is in fact necessary to assume the concept of exchange analytically in the analysis of society, but at the same time to bear in mind that a pure implementation of the exchange principle destroys a society. In consequence, this means that an analysis of the political means by which the exchange society defers or delays its own destruction has to be an inherent part of the theory. The historical evolution of the exchange principle does not—as one might have thought with Adorno the philosopher—simply lead to a perfectly integrated society, but rather to a varied overlapping of integration and disintegration phenomena. This overlapping of integration and disintegration is actually the real sociological substance of what Adorno meant with the concept of *Dialectic of Enlightenment*:

> And if you want to reduce to a formula to learn what is meant by the *Dialectic of Enlightenment* in real social terms, this is the time. I would like to go a step further and at least broaden the problem horizon by asking whether intersecting tendencies towards disintegration oppose each other more and more, in the sense that the different social processes which have welded together arise extensively from divergent or self-contradictory interest groups and not from the increasing integration of society, rather than maintaining that moment of neutrality, of relative indifference to each other, which they once had in the earlier phases of society. (Adorno 1972, p. 79)

The considerations that Adorno sets out in the fourth lecture on the status of political reform in late capitalism to some extent go against the grain of hermetic functionalism. Were the reader to take the hermetic image of society from the philosophical writings seriously, the person who had been enlightened by the *Dialectic of Enlightenment* would be confronted with the hair-raising alternative of whether to pursue a career in closed society or to lose his mind outside its walls. Social criticism would have no place in society itself if there were no third way between the extremes of perfect integration and total disintegration. Within the term 'criticism' itself, this third way is presupposed. Adorno links the conditions for the possibility of a 'critical' influence on society with a dialectic contemplation of the status of political reforms in democracy:

> It would be a bad and idealistic abstractness if, for the sake of the structure of the whole, one were to trivialize or even negatively accentuate the possibility for improvements within the framework of the existing conditions. There would in fact be a concept of totality in this which disregards the interests of the individuals living here and now, and this calls for a kind of abstract trust in world history which I, in any case, am absolutely unable to summon up in this form. I would say that the more the present social structure has the character of a monstrously rolled-up second nature, and that as long as this is the case, the most wretched intrusions into the existing reality will also have a much greater, indeed symbolic, meaning than befits them. Therefore, I would think in the present social reality one should be much more sparing with the reproach of so-called reformism than in the past century. Where one stands in respect to reform is, to a certain degree, also a function for evaluating the structural relations within the whole, and since this change in the whole no longer seems possible in the same directness as it did around the middle of the past century, these questions pass over into a completely different perspective. That is what I wanted to tell you at this point. (Adorno 1972, p. 53)

Theodor W. Adorno died in Sils-Maria, Switzerland, on August 6, 1969.

See also: Authoritarian Personality: History of the Concept; Authoritarianism; Authority, Social Theories of; Bourgeoisie/Middle Classes, History of; Capitalism; Critical Theory: Contemporary; Critical Theory: Frankfurt School; Cultural Rights and Culture Defense: Cultural Concerns; Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Culture as Explanation: Cultural Concerns; Enlightenment; Individual/Society: History of the Concept; Integration: Social; Lazarsfeld, Paul Felix (1901–76); Marxism in Contemporary Sociology; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Mass Society: History of the Concept; Media Ethics; National Socialism and Fascism; Personality Psychology; Personality Theory and Psychopathology; Political Economy, History of; Positivism, History of; Psychoanalysis in Sociology; Socialization: Political; Socialization, Sociology of; Sociology, History of; State and Society; Theory: Sociological; Totalitarianism

Bibliography

Adorno T W 1972 *On the (with Max Horkheimer) The Dialectic of Enlightenment*. Trans. John Cumming. New York: Herder and Herder, ["The Concept of Enlightenment"; "Excursus I: Odysseus or Myth and Enlightenment"; "Excursus II: Juliette or Enlightenment and Morality"; "The Culture Industry: Enlightenment as Mass Deception"; "Elements of Anti-Semitism: Limits of Enlightenment"; "Notes and Drafts"]

Ashton E B 1983 *Introduction to the Sociology of Music*. Seabury Press, New York

Bernstein J M (ed.) *1991 The Culture Industry: Selected Essays on Mass Culture*. Routledge, London [CB 427 A3 1991]

Domingo W 1983 *Against Epistemology: A Metacritique—Studies in Husserl and the Phenomenological Antinomies*. MIT Press, Cambridge, MA

Frankel-Brunswick E, Levinson D J, Sanford R N 1950 *The Authoritarian Personality*. Harper and Row, New York

Gödde C 2000 *Introduction to Sociology*. [trans. Edmund Jephcott]. Stanford University Press, Stanford (forthcoming)

Nicholsen S W 1993 *Hegel: Three Studies*. MIT Press, Cambridge, MA

Pickford H W 1998 *Critical Models: Interventions and Catchwords*. Trans. Henry W. Columbia University Press, New York

Tarnowski K, Will F 1973 *The Jargon of Authenticity*. Northwestern University Press, Evanston, IL

Tiedemann R (ed.) 1998 *Beethoven: The Philosophy of Music*. trans. Edmund Jephcott. Stanford University Press, Stanford

H. Dubiel

Adult Cognitive Development: Post-Piagetian Perspectives

There has been an abundant history of experimental work on how cognitive processes such as memory and attention decline with age. However, a different picture may emerge if a life-span developmental approach is taken. From a traditional developmental perspective, the question to ask is whether there are adaptive qualitative changes in cognition that take place beyond adolescence. In initial attempts to address this question, the 1970s brought a proliferation of studies examining Piaget's theory of cognitive development in adulthood. However, much like the experimental cognitive aging work, cross-sectional studies indicated that many adults do not attain formal operations, Piaget's final stage of cognitive development (Kuhn 1992). Formal operational thinking is characterized by hypothetico-deductive reasoning about abstract concepts in a systematic fashion, that is, scientific thinking. It is governed by a generalized logical structure that provides solutions to hypothetical problems. In response to these findings, Piaget (1972) concluded that formal operations is probably not universal, but tends to appear only in those areas in which individuals are highly trained or specialized.

The major problem with simply examining adult cognitive development in terms of age differences in formal operational functioning in adulthood is that it may underestimate the cognitive functioning of adults. In other words, comparing age groups on formal operations uses adolescent or young adult thinking as the standard of competence. Is this a valid assumption to make when examining adaptive cognition in adulthood? Or, do more mature ways of thinking emerge during adulthood?

In response to this concern, Riegel (1976) proposed one of the first alternative models of cognitive development beyond formal operations. He argued that formal operations is limited in its applicability, in that the hypothetico-deductive mode of reasoning does not adequately represent the qualitatively different types of thinking that adults use. Other researchers also pointed out that Piaget's stage of formal operations is primarily limited to explaining how individuals arrive at one correct solution. In other words, the manner in which adults discover or generate new problems and how they consider several possible solutions are not explained. Finally, the fact that adults often restrict their thinking in response to pragmatic constraints is contradictory to the unconstrained generation of ideas characteristic of formal operations. The limitations of formal operations in explaining adult thinking set the stage for a wave of research documenting a continued cognitive growth beyond formal operations called postformal thought (Commons et al. 1989, Sinnott 1996).

1. Definition of Postformal Thought

Postformal thought is characterized by a recognition that (a) truth varies from situation to situation, (b) solutions must be realistic to be sensible, (c) ambiguity and contradiction are the rule rather than the exception, and (d) emotion and subjective factors play a critical role in thinking. These characteristics result in two types of thinking: relativistic and dialectical thinking. Relativistic thinking involves the ability to realize that there are many sides to any issue, and that the right answer depends upon the circumstances. Dialectical thinking involves the ability to consider the merits of differing viewpoints and synthesize them into a workable solution. Both of these modes of thinking accept the fact that knowledge and decisions are necessarily subjective. Thus, postformal thinkers adopt a contextual approach to problem solving in that solutions must be embedded in a pragmatic context (i.e., applying knowledge and decisions to changing circumstances of one's life).

For example, in a seminal study on cognitive growth beyond adolescence, Perry (1970) traced the developmental trajectory of relativistic and dialectical

thinking across the undergraduate years. He found that cognitive development moves from reliance on the expertise of authorities in determining what is true or false to increased levels of cognitive flexibility. The first step in this process is a shift toward relativistic thinking. This type of thinking produces a healthy dose of skepticism and lack of certainty regarding potential solutions to problems. However, Perry demonstrated that adults, in order to progress beyond skepticism, develop commitments to particular points of view. Thus, later stages allow adults to engage in dialectical thinking. They recognize that they are their own source of authority, that they must make a commitment to a position, and that others may hold different positions to which they are equally committed.

2. *Research on Postformal Thought*

Perry's research on the development of relativistic thinking opened the door to further studies documenting systematic changes in thinking beyond formal operations. King and Kitchener (1994) extended Perry's investigation of the relativistic nature of adult thinking by mapping out the development of reflective judgment. On the basis of longitudinal studies of young adults, they identified a systematic progression of thinking. The first three stages in the model represent prereflective thought. In these stages, individuals do not acknowledge that knowledge is uncertain, and maintain that there must be a clear and absolutely correct answer. In stages 4 and 5, individuals recognize that there are problems that contain an element of uncertainty. However, they are not adept at using evidence to draw a reasonable conclusion. The final stages 6 and 7 represent true reflective judgment. Individuals realize that knowledge is constructed and thus must be evaluated within the context in which it was generated. Progression through stages involves both skill acquisition, a gradual process of learning new abilities, and an optimal level of development, the highest level of cognitive capacity that a person can reach. However, because the environment does not provide the support necessary for high-level performance on a daily basis, individuals do not operate at their optimal level most of the time (King and Kitchener 1994).

Sinnott (1996) examined relativistic thinking in the area of interpersonal understanding. Sinnott investigated the degree to which individuals are guided by the fact that points of view in interpersonal relations are necessarily subjective and can be contradictory. She found that when solving problems designed to assess both formal and relativistic thinking in real-life situations, young adults tended to solve all types of problems by a formal mode of thinking, i.e., looking for one correct answer. In contrast, older adults were more likely to use relativistic thinking.

On a measure assessing paradigmatic beliefs about the social world, adults of all ages endorsed statements reflecting dialectical thinking more than statements reflecting relativistic thinking and absolutist thinking (i.e., endorsing one correct point of view) (Kramer and Kahlbaugh 1994). Furthermore, adults' scores on paradigmatic beliefs were unrelated to verbal intelligence and various personality variables such as tolerance of ambiguity. From these findings, it appears that dialectical thinking is rated higher than relativistic thinking in mature thought.

Finally, the interpersonal flavor of postformal thinking is reflected in Labouvie-Vief's theory of adult cognitive development (1992, 1997). Labouvie-Vief contends that adults, as they grow older, develop the ability to integrate emotion with logic in their thinking. From this perspective, a major goal of adult thinking is to handle everyday living effectively. Instead of generating all possible solutions to problems, adults make choices on pragmatic, emotional, and social grounds. This demands making compromises and tolerating ambiguity and contradiction. Researchers speculate that in the area of social reasoning, middle-aged and older adults have some expertise due to their respective accumulation of experience (Blanchard-Fields 1997, Labouvie-Vief 1997). For example, findings indicate that younger age groups reason at a lower developmental level (e.g., less relativistic thinking), especially when confronted with problems that are emotionally salient to them (Blanchard-Fields 1986, Blanchard-Fields and Norris 1994).

Finally, research on the pragmatics of intelligence in the form of wisdom (Baltes et al. 1998, Staudinger and Baltes 1996) also demonstrates adult reasoning that is related to postformal development. According to the Berlin Wisdom Paradigm, wisdom involves the coordination of cognition, motivation, and emotion in a combination of exceptional insight and mature character (Staudinger and Baltes 1996). Specifically, wisdom embodies five criteria closely related to the characteristics of postformal thought. These include factual knowledge, procedural knowledge, contextualism, value relativism, and the acceptance of uncertainty. Studies assessing wisdom ask participants to think aloud about difficult life problems, which are evaluated on the five wisdom-related criteria. Findings indicate that there are no negative age trends in wisdom-related performance. Second, older adults with wisdom-facilitative experiences (e.g., older clinical psychologists and wisdom nominees) disproportionately represent individuals with a large share of the higher-level responses.

Overall, there is some evidence that adults tend to reason more in a postformal manner than younger adults and adolescents, although the age differences are not strong. Indeed, postformal thinking is qualitatively different from formal operational thinking, which relies primarily on a formal logical mode of analysis. It provides a counterperspective to the view

that with increasing age comes inevitable decline. More specifically, postformal reasoning affords adults the ability to embrace the complexities of social reality and emotional involvement in problem solving. The importance of socioemotional aspects of adult reasoning is reflected in recent research trends in social cognition and aging and solving practical problems. However, although there is an emerging consensus that adulthood yields qualitatively different styles of thinking, such as relativistic or dialectic, there is no consensus on whether postformal thinking reflects a true adult developmental cognitive stage.

3. *Postformal Thought as a Stage of Development*

As indicated above, there is much debate in life-span cognitive development on the question of whether adult cognitive development progresses in stage-like fashion toward higher levels of reasoning (Baltes et al. 1998, Basseches 1984, Labouvie-Vief 1992). Although there is evidence that some adults conceptualize reality by postformal styles of reasoning, especially in socioemotional domains, there is a substantial amount of evidence that a large proportion of adults do not display all of the characteristics of postformal development (Labouvie-Vief 1997). Thus, a number of researchers have taken a functionalist approach to explain the lack of a strong positive developmental trajectory in postformal thinking during adulthood. In this case, development is characterized as adaptations to the local environment (Baltes and Graf 1996, Labouvie-Vief 1992). From this perspective, changes in experiences and demands in life determine whether different styles of adult thinking emerge across the latter half of the life span. Thus, the hallmark of adult development is interindividual variability rather than uniformity of cognitive functioning.

A second approach to explaining the variability in the maturity of adult thinking is that in adulthood knowledge becomes more specialized on the basis of experience, a development which in turn reflects a lesser role of age-related neurological development and social demands for increased specialization of knowledge and expertise (Hoyer and Rybash 1994). Thus, knowledge becomes encapsulated in that it is increasingly complex and resistant to change. Because it is experientially based, cognitive development in adulthood is directed toward mastering competency in specific domains rather than being uniform across domains, as in childhood stages of cognitive development (Hoyer and Rybash 1994).

4. *Future Directions*

In conclusion, research on cognitive development beyond adolescence highlights the positive aspects of cognitive changes (i.e., adaptive cognitive reasoning in a social context) in the aging adult from a contextual perspective. The goal of development is successful adaptation to the individual's context of living. However, it is important to acknowledge that many older adults may not achieve higher levels of postformal thinking. Future research needs to address this issue. For example, is it the case that the loss of fluid abilities well documented in the literature on psychometric intelligence and aging may influence postformal thinking styles? Or, is it the case that postformal assessment strategies do not adequately tap into the domains specific to complex reasoning in older adulthood? Future researchers may need to pay more attention to methods explicitly focused on the nature of reasoning styles in areas more relevant in advanced age.

Finally, including an individual differences approach to the study of cognitive development in adulthood promises to advance the field in important ways. The individual differences approach makes it explicit that age is only probabilistically associated with levels of cognitive functioning, and that this association can in fact be influenced and even moderated by a host of relevant variables (e.g., beliefs, attitudes, dispositional styles, and ego level). Thus, an individual differences model could make it possible to evaluate the conditions under which adults of varying ages and of different personological and developmental characteristics are likely to engage in qualitatively different strategies of cognitive functioning.

See also: Adult Development, Psychology of; Adult Education and Training: Cognitive Aspects; Adulthood: Developmental Tasks and Critical Life Events; Aging, Theories of; Cognitive Development in Childhood and Adolescence; Education in Old Age, Psychology of; Lifespan Theories of Cognitive Development; Parenthood and Adult Psychological Developments; Personality Development in Adulthood; Piaget's Theory of Human Development and Education; Social Learning, Cognition, and Personality Development; Wisdom, Psychology of

Bibliography

Baltes P B, Graf P 1996 Psychological aspects of aging: Facts and frontiers. In: Magnusson D (ed.) *The Life Span Development of Individuals: Behavioral, Neurobiological, and Psychosocial Perspectives*. Cambridge University Press, Cambridge, UK, pp. 427–59

Baltes P B, Lindenberger U, Staudinger U 1998 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology*, 5th edn. *Theoretical Models of Human Development*. Wiley, New York, Vol. 1

Basseches M 1984 *Dialectical Thinking*. Ablex, Norwood, NJ

Blanchard-Fields F 1986 Reasoning in adolescents and adults on social dilemmas varying in emotional saliency. *Psychology and Aging* **1**: 325–33

Blanchard-Fields F 1997 The role of emotion in social cognition across the adult life span. In: Schaie K W, Lawton M P (eds.)

Annual Review of Gerontology and Geriatrics. Springer, New York, Vol. 17, pp. 238–65
Blanchard-Fields F, Norris L 1994 Causal attributions from adolescence through adulthood: Age differences, ego level, and generalized response style. *Aging and Cognition* **1**: 67–86
Commons M, Sinnott J, Richards F, Armon C 1989 *Adult Development: Comparisons and Applications of Developmental Models*. Praeger, New York
Hoyer W, Rybash J 1994 Characterizing adult cognitive development. *Journal of Adult Development* **1**: 7–12
King P, Kitchener K 1994 *Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults*. Jossey-Bass, San Francisco
Kramer D A, Kahlbaugh P E 1994 Memory for a dialectical and a nondialectical prose passage in young and older adults. *Journal of Adult Development* **1**: 13–26
Kuhn D 1992 Cognitive development. In: Bornstein M, Lamb M (eds.) *Developmental Psychology: An Advanced Textbook*. Erlbaum, Hillsdale, NJ, pp. 211–72
Labouvie-Vief G 1992 A neo-Piagetian perspective on adult cognitive development. In: Sternberg R J, Berg C A (eds.) *Intellectual Development*. Cambridge University Press, New York, pp. 197–228
Labouvie-Vief G 1997 Cognitive-emotional integration in adulthood. In: Schaie K W, Lawton M P (eds.) *Annual Review of Gerontology and Geriatrics*. Springer, New York, Vol. 17, pp. 206–37
Perry W 1970 *Forms of Intellectual and Ethical Development in the College Years: A Scheme*. Holt, New York
Piaget J 1972 Intellectual evolution from adolescence to adulthood. *Human Development* **15**: 1–12
Riegel K 1976 The dialectics of human development. *American Psychologist* **31**: 689–700
Sinnott J 1996 The developmental approach: Postformal thought as adaptive intelligence. In: Blanchard-Fields F, Hess T (eds.) *Perspectives on Cognitive Change in Adulthood and Aging*. McGraw-Hill, New York
Staudinger U, Baltes P B 1996 Interactive minds: A facilitative setting for wisdom related performance. *Journal of Personality and Social Psychology* **71**: 746–62

F. Blanchard-Fields

Adult Development, Psychology of

Many of the classic developmental theories hold to the view that development takes place in childhood but not during adulthood. For example, psychoanalytic (e.g., Freud) and organismic (e.g., Piaget) approaches to development include end states (genital stage or formal operations, respectively) which occur in early adolescence. A number of subsequent theories have challenged these child-centric views of development (e.g., Buhler and Masarik 1968, Erikson 1963, Jung 1971), and many recent theories acknowledge the possibility of development and change throughout the adult years (Baltes et al. 1997). The realization that many aspects of psychological functioning show both growth and declines in the adult years has led to the study of the nature of the changes as well as their antecendents and consequences. Key theories and findings in the adult development field will be summarized below.

1. *The Adult Years*

Some researchers suggest the use of chronological age as a marker for the timing of adulthood, whereas others suggest the transition to adulthood is better characterized by events or rites of passage such as graduation from school, starting a job, or having a family (Neugarten and Hagestead 1976). Adulthood is usually divided into several periods: young or early adulthood (approximately aged 20–39), middle adulthood (40–59), and old age (60+). Old age is typically divided into the periods of young old (60–75) and old old (75 and up).

Subjective definitions of age are also important. When adults are asked how old they feel, their responses often do not correspond to their actual chronological age. Those in adolescence often feel slightly older then their age, young adults usually report feeling close to their age, whereas in midlife and old age adults feel on average 10–15 years younger than their age (Montepare and Lachman 1989). The older one is, generally, the larger the discrepancy between age and subjective age. Feeling younger than one's age is typically associated with better health and well-being.

The social clock is another important organizing framework for adulthood. Based on cultural and societal norms, there is a sense of when certain events or milestones should be achieved. Thus individuals can gauge whether or not they are on-time or off-time relative to these norms (Neugarten and Hagestead 1976). There are consequences associated with being early or late with regard to certain events (graduation, marriage, having a child, getting a job). However, the research indicates that many individuals set their own timetables which do not correspond to societal norms. For example, those with more education will be likely to get married, start a family and begin a job at later ages than the general population. It is one's own constructed timetable that appears to be most important for well-being, and the social clock may be less critical.

2. *Life-span Approach to Adult Development and Aging*

According to a life-span developmental approach (Baltes et al. 1997) there are a number of guiding principles for the study of adult development and aging. First is the assumption that development is a lifelong process, and is just as likely to occur in adulthood as at other points in the life span. De-

velopment is also assumed to take many forms, in that there are multiple paths to the same outcome and there are many outcomes that are desirable or adaptive. Changes during adulthood may be gradual or abrupt. Functioning and behaviors in most domains are modifiable through interventions or changes in environmental conditions. The nature of development is best understood by considering persons within their contexts. In addition to examining psychological functioning, it is also useful to consider the social, biological, anthropological (cross-cultural), economic, political, and historical (cohort) manifestations and intersections. Also, according to a life-span view, all developmental changes in adulthood are the result of both nature and nurture, although at different points in the life span and for different areas of functioning the influence of either heredity or environment may be more salient. Developmental changes may be influenced by normative age-graded factors (e.g., puberty or retirement), normative history graded factors (Great Depression, polio epidemic), or nonnormative events (illness, loss of spouse). Development at all points in the life span is best understood as a cumulative process of gains and losses.

3. *Theories of Adult Development*

Freud's ([1916–17] 1963) theory describes psychological development in a series of psychosexual stages completed by adolescence. Personality was determined based on the resolution of these stages in interaction with the social environment, especially the mother. Carl Jung ([1933] 1971) disagreed with Freud on a number of points, including the primary role of sexuality in development as well as the potential for change during adulthood. Jung ([1933] 1971) held that personality can develop throughout life, and called this the individuation process. According to Jung, the individual strives to become whole by integrating the unconscious undeveloped parts of the psyche with the more conscious ones.

Erikson (1963) also modified Freud's psychosexual theory and formulated a psychosocial approach to development across the life span. This epigenetic theory involves a series of eight stages from infancy through old age. At each stage there is a central crisis to resolve. The adult stages include intimacy vs. isolation, generativity vs. stagnation, and ego integrity vs. despair. The theory suggests that individuals negotiate the crises of each stage in a fixed sequence and the way earlier crises are resolved influences the outcomes of later stages. Thus, according to Erikson, before it is possible to successfully negotiate the issues surrounding intimacy in adulthood, it is critical to resolve one's identity vs. role confusion during adolescence. The stage of generativity does not imply only a focus on one's own offspring, but rather a broader interest in the succeeding generation, in the world at large, at work, among family and friends, as well as in terms of passing on one's legacy such as through works of art or literature.

Vaillant and Milofsky (1980) tested and expanded Erikson's theory. Through this research comparing well-educated college men and inner city men, they found evidence that Erikson's stages were equally applicable to both socioeconomic groups. Moreover, they found evidence that the stages were negotiated in the same sequence in both social realms. However, the natures of the issues addressed at each stage were different in content. Moreover, Vaillant and Milofsky (1980) further differentiated the adult stages by adding two substages. Between the intimacy and generativity stages they added a substage called career consolidation. In this stage they found that adults were focused on establishing their work identity and working towards accomplishing career goals. Resolution of this stage then led to the generativity stage in which workers often became mentors to the younger generation. The generativity stage often lasts for 20 years or more and Vaillant and Milofsky (1980) found evidence for a substage called 'keepers of the meaning.' In this stage, adults were concerned with transmitting their values and ideas to the next generation.

Ego integrity vs. despair is the final stage of life proposed by Erikson (1963). In this period the adult is faced with accepting that death is inevitable. To navigate this stage successfully involves coming to terms with one's life and accepting it and moving on to make the most of the remaining time. Those who go the route of despair, however, are often filled with regrets about what did not happen and often fear death because they do not have a sense of accomplishment or sense that they have lived a good life. This period is often characterized by an intensive life review (Butler 1963). This involves a therapeutic and useful process of sorting through one's life and reminiscing about the past in order to make sense of it. It is through this process that one is able to accept one's life and this allows people to focus on the present and the future so as to age successfully. Although there is evidence that people reminisce at many different points in life, it is more common in later life and should be considered a natural part of the aging process.

4. *Personality Theories*

Some aspects of personality remain relatively stable throughout life, whereas others appear to change. From a trait perspective there is longitudinal and cross-sectional evidence for long-term stability of the Big Five personality dimensions: extraversion, neuroticism, agreeableness, conscientiousness, and openness to experience (Costa and McCrae 1994). Over time, people maintain their rank orders on these key personality traits relative to others. However, there is some evidence for changes in overall level, with older

adults on average becoming less extraverted, less neurotic, and less open to experience, but more agreeable. When other dimensions of personality such as components of well-being are considered, there is evidence for greater changes in personality (Ryff 1995). Purpose in life and personal growth were found to decrease with age, whereas environmental mastery and positive relations were found to increase with age.

There is also evidence for changes in sex role characteristics in adulthood. The evidence suggests that with aging, the genders become less differentiated in terms of masculine and feminine traits. In a classic study, Neugarten and Guttman (1968) found that men adopted more communal characteristics as they aged and women adopted more agentic ones. This integration of masculine and feminine characteristics with age is consistent with what Jung called the process of individuation. Jung's theory suggested that with aging the process of individuation involves an integration of the conscious and unconscious parts of the ego. For men the feminine side is usually repressed, as is the masculine side for women. With aging, the process of bringing the undifferentiated aspects of the self into awareness is considered adaptive.

In addition to the objective data on personality over time, subjective analyses have yielded information about the processes of perceived change in personality and well-being. Adults typically expect to change in the future and see changes relative to the past (Ryff 1991). By reflecting back on the past and projecting into the future, there is evidence that adults have experienced change and expect to undergo further changes. The present functioning is in part understood in relation to what has come before and what is anticipated in the future.

5. *The Self*

The self in adulthood comprises many components such as self-esteem, self-confidence, self-concept, multiple selves, self-efficacy, and the sense of control. Some aspects of the self change, whereas other aspects remain stable in adulthood. The number of imagined or possible selves appears to decrease in later life; however, the number of undesired or dreaded selves increases. Moreover, the number of health-related selves increases in later life (Cross and Markus 1991). The sense of self is differentiated across domains. Whereas overall mastery remains relatively stable, there are some domains in which perceived control declines (Lachman 1991). Perceived control over children, physical functioning, and memory decline in adulthood, whereas perceived control over marriage relationship and work increase in adulthood (Lachman and Weaver 1998). Self-efficacy influences the choice of tasks as well as persistence and level of anxiety and stress. Those who have a greater sense of control are more likely to exert effort and choose effective strategies. There appears to be a physiological link as well. Lack of control is associated with greater stress and poorer immune system functioning. Although there appear to be fewer critical life events in late adulthood, the events that do occur typically are associated with greater stress than those experienced in young adulthood. Two strategies for handling stress are problem-focused and emotion-focused coping (Aldwin 1994). Older adults are more likely than younger ones to use emotion-focused approaches to coping. Also when in the face of difficult circumstances or challenges, it is adaptive to use primary approaches (changing the environment to meet personal goals) and secondary approaches (changing the self to accommodate to the environmental demands) to control. Although primary control strategies are used consistently throughout adulthood, the use of secondary control strategies increases with aging (Heckhausen and Schulz 1995).

6. *Cognitive Functioning*

There is evidence that some aspects of cognitive functioning decline in adulthood, whereas others increase (Baltes et al. 1997). Again, a multidimensional perspective is most useful given the different trajectories of change. For the pragmatics of intelligence, also known as crystallized intelligence, there is continued growth in adulthood. These aspects of intelligence are related to accumulated experience and knowledge and are associated with education and acculturation. Thus, the longer one lives in general the more knowledge one acquires. Presumably with aging one also can attain greater wisdom, or the ability to solve complex problems. In contrast, for the mechanics of development there is decrement starting in early adulthood. This includes mechanisms such as speed of processing, memory, and fluid intelligence. These functions rely on more biologically based processing and thus show decrements that are tied to changes in the central nervous system and brain-related changes.

7. *Social Relationships*

In adulthood, the nature of social relationships changes in both quality and quantity. The number of close friends and confidants increases in young adulthood and remains relatively stable during midlife (Antonucci and Akiyama 1997). In later life, one may begin to lose family and friends due to retirement from one's job, moving to a new residence, or death. Carstensen (1995) has found that in later life adults prefer to have a smaller number of close relationships, and they become increasingly selective in their choice of those with whom they interact. Social support is an important resource throughout adulthood, and especially in later life. It may involve material assistance

or help with tasks and chores, as well as emotional support. Social support may be provided by formal (e.g., institutions or community organizations) or informal (e.g., family or friends) sources. Those who report greater social support tend to be healthier and live longer (Antonucci and Akiyama 1997).

8. Successful Aging

People have long sought after the fountain of youth, looking for ways to live longer and to improve the quality of life. Not only are people now living longer than in the earlier part of the twentieth century, the quality of life is also improving. What factors contribute to a successful later life? Research has shown that there are many factors that involve lifestyle choices that are associated with successful aging (Baltes and Baltes 1990, Rowe and Kahn 1998). Some of the psychosocial factors and behavioral factors are exercise, healthy diet, sense of efficacy and control, mental stimulation, and social support. Those who have friends and family they can rely on report less depression and greater well-being. There is also evidence that having access to support from others is associated with better health and greater longevity. It is not yet known what the mechanisms are that link social support with health. It is likely to be a combination of factors such as reducing stress and boosting immune functioning (Rowe and Kahn 1998). It is clear that there are many factors and behaviors under one's control that affect the nature and course of adult development and aging.

9. Future Directions

We have learned a great deal about the nature of adult development and aging over the past few decades. We know that there are wide individual differences in later life and that the extent and direction of change varies across people. Researchers are conducting studies to investigate the processes that link psychosocial functioning with biomedical factors. It will be useful to understand how beliefs and attitudes such as the sense of control or personality factors contribute to health and longevity. Some of the promising biomarkers are cortisol (stress hormone) and fibrinogen (blood clotting substance). It is likely that psychosocial factors have an effect on the immune system as well as on health behaviors such as exercise and healthy diet. Ultimately, researchers are interested in developing treatments and interventions not only to cure and remediate problems in adulthood and old age, but also to prevent them by taking precautions and action during earlier age periods.

See also: Adult Cognitive Development: Post-Piagetian Perspectives; Developmental Psychology; Ego Development in Adulthood; Erikson, Erik Homburger (1902–94); Human Development, Successful: Psychological Conceptions; Jung, Carl Gustav (1875–1961); Lifespan Development, Theory of; Lifespan Theories of Cognitive Development; Midlife Psychological Development; Personality Development in Adulthood; Personality Psychology; Personality Theories

Bibliography

Aldwin C M 1994 *Stress, Coping, and Development*. Guilford Press, New York

Antonucci T C, Akiyama H 1997 Concern with others at midlife: Care, comfort, or compromise? In: Lachman M E, James J B (eds.) *Multiple Paths of Midlife Development*. University of Chicago Press, Chicago, pp. 145–170

Baltes P B, Baltes M M (eds.) 1990 *Successful Aging: Perspectives from the Behavioral Sciences*. Cambridge University Press, New York

Baltes P B, Lindenberger U, Staudinger U M 1997 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology: Vol. 1. Theoretical Models of Human Development*, 5th edn. J. Wiley, New York, pp. 1029–143

Buhler C, Masarik F (eds.) 1968 *The Course of Human Life*. Springer, New York

Butler R N 1963 The life review: An interpretation of reminiscence in the aged. *Psychiatry* **26**: 65–76

Carstensen L 1995 Evidence for a life-span theory or socioemotional selectivity. *Current Directions in Psychological Sciences* **4**: 151–6

Costa Jr. P T, McCrae R R 1994 Set like plaster? Evidence for the stability of adult personality. In: Heatherton T F, Weinberger J L (eds.) *Can Personality Change?* 1st edn APA, Washington, DC, pp. 21–40

Cross S, Markus H 1991 Possible selves across the life span. *Human Development* **34**: 230–55

Erikson E H 1963 *Childhood and Society*, 2nd edn, rev. and enlarged. Norton, New York

Freud S [1916–17] 1963 Introductory lectures on psychoanalysis. In: Strachey J (ed. and trans.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London, Vol. 16

Heckhausen J, Schulz R 1995 A life-span theory of control. *Psychological Review* **102**: 284–304

Jung C G [1933] 1971 *Modern Man in Search of a Soul*. Harcourt Press & World, New York

Lachman M E 1991 Perceived control over memory aging: developmental and intervention perspectives. *Journal of Social Issues* **47**: 159–75

Lachman M E, Weaver S L 1998 Sociodemographic variations in the sense of control by domain: Findings from the MacArthur studies of midlife. *Psychology and Aging* **13**: 553–62

Montepare J M, Lachman M E 1989 'You're only as old as you feel.' Self-perceptions of age, fears of aging, and life satisfaction from adolescence to old age. *Psychology and Aging* **4**: 73–8

Neugarten B L, Guttman D L 1968 Age–sex roles and personality in middle age: A thematic apperception study. In: Neugarten B L (ed.) *Middle Age and Aging*. University of Chicago Press, Chicago, pp. 58–76

Neugarten B L, Hagestead G O 1976 Age and the life course. In: Binstock R H, Shanas E (eds.) *Handbook of Aging and the Social Sciences*. Van Nostrand Reinhold, New York, pp. 35–55
Rowe J W, Kahn R L 1998 *Successful Aging*. Pantheon Books, New York
Ryff C D 1991 Possible selves in adulthood and old age: A tale of shifting horizons. *Psychology and Aging* **6**: 286–95
Ryff C D 1995 Psychological well-being in adult life. *Current Directions in Psychological Science* **4**: 99–104
Vaillant G E, Milofsky E 1980 The natural history of male psychological health: IX. Empirical evidence for Erikson's model of the life cycle. *American Journal of Psychiatry* **137**: 1348–59

M. E. Lachman

Adult Education and Training: Cognitive Aspects

Various cognitive functions do change with age. Decreases in memory performance, problem solving, speed and precision of perception, and concentration can be described as common age-related losses. In fact, experiencing decreases in cognitive functioning often leads people to changes in self-categorization and self-concept, i.e., the feeling of being or becoming old. However, age-related decline in basic cognitive processes does not correspond to cognitive performance in everyday context. The more experiences and knowledge are necessary to cope with everyday tasks, the less age differences in performance can be found. Moreover, decreases in basic cognitive functions show a high amount of interindividual variability; cognitive functioning in old age is highly influenced by lifelong educational processes and competencies developed in earlier phases of the life span can be used to compensate for developmental losses. As life situation in old age can be interpreted at least in part as the result of lifelong developmental processes, it is of interest whether competencies that could not be developed in earlier phases of the life span can be learned in later years as well, so that higher levels of performance, e.g., effective coping with tasks and challenges of the current life situation, become possible. The assumption that deficits in performance can be compensated for by the use of effective learning strategies is central for most adult education and training programs. However, the question whether training programs should try to educate general strategies or specific techniques that do depend very much on contextual factors has not yet been answered.

This contribution proceeds from research on older people's use of learning strategies and performance in memory tasks. In a second step the question whether adult education programs should engage in the improvement of basic cognitive processes or specific everyday activities is addressed. In the third part it is argued for a theoretical approach that challenges the traditional dichotomy of two components of human intelligence, i.e., fluid and crystallized intelligence. From this perspective the concept of practical intelligence is essential for an adequate understanding of intellectual functioning and cognitive performance in old age. In this section results from empirical studies on performance in younger and older workers and principal demands for training in the context of occupational activities are also discussed. The final part is concerned with the relationship between education and healthy aging. From a life span developmental perspective it is argued that education improves health via perceptions of control and self-effectiveness which in turn increase effective agency and healthy lifestyles.

1. *Learning Strategies of Older People*

Those who could not acquire effective learning strategies in earlier phases of development show more difficulties in organizing and restructuring new information, rarely use mnemonics and mediators, and often fail to establish associative bounds between specific contents. When people fail to organize and restructure new information effectively, they are not able to recall whole semantic clusters but have to remember single words or units instead. For this reason they need more repetition and show poorer memory performance altogether. However, memory performance can be improved through good instructions guiding the process of encoding and the training of learning strategies even in very old age. In empirical studies semantic references (i.e., expounding structural similarities between specific units) and mnemonics (e.g., imagination of pictures that summarize learning contents) proved to optimize memory performance in older people.

Difficulties in recalling stored information are also often due to a lack of effective learning strategies. Poor accessibility of learned material is seen to be caused by omission to include additional information during the encoding process which could facilitate recall. Additionally, failing to organize and restructure learning material makes recall more difficult.

Many older people had no opportunity to gain effective and differentiated strategies for encoding and decoding learning materials. Consequently, empirical studies show that older people (a) do not differ significantly from younger people in recognition tasks, (b) do not reach the performance of younger people in free recall tasks, (c) can reach higher levels of free recall performance when important features of the learned material are given as an aid. Moreover, the data suggest that the failure of older people to organize and restructure new information effectively is not due

to deficits in relevant abilities and skills, in fact they simply do not organize and restructure new information spontaneously, since they are unfamiliar with the context of a test situation.

Cognitive training primarily aims to foster acquisition and use of encoding and decoding strategies. The use of such strategies should help to compensate for deficits in training and practice among older people. Results from longitudinal intervention studies prove the effectiveness of cognitive training programs (e.g., Oswald and Rödel 1995).

Whether and to what extent the learning strategies used by older people are to be described as deficient is discussed as controversial (see Berg et al. 1994). A strategy can be defined as one of alternative methods to solve a specific cognitive task, a procedure that is optional and intentional simultaneously. Age differences in the use of strategies have been studied as possible causes for poorer performance of older people in various cognitive tasks, including memory, spatial imagination, problem solving, and solving of everyday problems. The hypothesis of deficient strategies of older people has important implications for learning from a life span developmental perspective since it is exactly the aim of numerous training programs that older people should use more efficient strategies (Willis 1987, 1990). However, from a closer look at relevant research it is obvious that the deficiency hypothesis must be rejected. Instead, empirical data support the assumption that differences between younger and older people reflect the specific adaptive functions in different ages, depending on the cognitive and pragmatic tasks in given developmental contexts. Consequently, recent research has concentrated on interindividual differences in the use of different strategies. These differences can be explained for by variables similar to those that can explain for differences in psychometric intelligence, i.e., education, experience, stimulating contexts, continuous use, etc.

2. *Training of Basic Processes vs. Training of Everyday Activities*

The distinction between two components of human intelligence, i.e., fluid intelligence as an age-related ability to solve new and unfamiliar problems and crystallized intelligence as an ability to solve familiar problems that can be preserved or even improved in old age (Horn 1982), does not mean that these components are independent of each other. Since every complex cognitive activity contains elements from fluid and crystallized intelligence and intellectual performance as a product can result from different proportions of the two components, expertise, i.e., a high level of crystallized intelligence, offers opportunities to compensate for losses in fluid intelligence.

The possibility to compensate for losses in basic cognitive processes has been proven in numerous empirical studies, especially in the field of occupational activities, but also in other meaningful everyday activities. It has been shown that performance in complex cognitive tasks does not decrease as fast as could be supposed from decreases in basic cognitive processes (Willis 1987). Strategies that allow for compensation in basic cognitive processes are, e.g., an intentional slowing of action, additional checks of solutions, restriction to a small number of activities and aims. However, as could be shown in the testing-the-limits paradigm, compensation in favor of the optimization of specific aspects generally leads to a prolongation of the time required for the task (Baltes and Baltes 1990, Kliegl et al. 1989).

The proven possibility to compensate for losses in intellectual abilities leads to the question whether everyday competence in old age can be improved by training of useful strategies and basic processes. In this context the person-centered intervention approach of Willis (1987) is instructive. According to this author complex everyday activities can be optimized by a training of basic processes. In the first step the significance of specific processes for clusters of important daily activities (e.g., reading operating instructions or an instruction leaflet) has to be determined. In a second step those processes that have an impact on performance in numerous activities can be trained. A training of basic processes would be very attractive for intervention research since participation in training programs could heighten performance in numerous contexts and activities. However, basic cognitive processes are at the very beginning of everyday performance; the relationship between the two is only poor and a satisfying prognosis of performance from basic processes is not possible. As a consequence, recent development in intervention research indicates a preference for another paradigm: the training of specific everyday activities. Since the context-independent training mnemonics failed to have the expected impact on everyday memory performance, it was proposed to offer specific courses aimed to improve memory of names or prevent people from mislaying glasses or keys instead of courses aimed to improve general memory performance. Following this approach it is necessary to create contexts of person-centered intervention that correspond very much to problematic situations in everyday living.

Consequently, from the perspective of this approach a detailed examination of individual life situations is demanded. This demand illustrates the principal dilemma of person-centered intervention programs: the expenditure of training so many people in so many specific situations is out of all proportion to the possible intervention effects. Intervention programs are often used to search for potentials for action and development, especially in the age-related component of intelligence. Numerous empirical studies have differentiated our understanding of human intelligence by demonstrating reserves of capacity for intellectual

performance. Cognitive functions can be improved through adequate training programs, especially when individual, social, and occupational aspects of the life situation are taken into account. Moreover, cognitive training can also be helpful for reaching noncognitive aims, another indication of the significance of cognition for successful management of life in our culture.

However, effects of cognitive training remain specific for concrete problems and situations. Moreover, according to Denney (1994), most training studies (naturally) focus on age-related abilities and skills where similar gains can be reached through exercise alone. Additionally, training has the greatest impact on skills that are not needed in everyday life. Therefore, Denney (1994) raises the question why people should participate in conventional training programs and whether it would not be better to create new programs that concentrate on well-developed abilities and skills, where little effects could have a great impact on possibilities to maintain an independent and self-responsible life.

3. *Practical Intelligence and Training in the Context of Occupational Activities*

The traditional research on human intelligence refers to abilities and skills that are acquired through education in adolescence and early adulthood, whereas abilities and skills required in later years, e.g., for successful occupational development, are neglected in operational definitions of this construct (see Labouvie-Vief 1985). Only since the mid-1980s did psychological research on the development of human intelligence begin to conceptualize and discover the acquisition of area-specific skills and practical intelligence (see Kruse and Rudinger 1997). The latter can be defined as the ability to solve practical everyday problems and to cope effectively with everyday tasks, i.e., as a dimension of human intelligence that may be correlated with the fluid and the crystallized component, but constitutes an own dimension that cannot be reduced to specific aspects of these two other components. Practical intelligence subsumes area-specific skills as well as more general abilities essential for effective coping with problems and tasks, e.g., overall perspective on a working field, competence in preparation and realization of decisions, development and further improvement of effective strategies.

Until recently, the development of practical intelligence has seldom been studied empirically, but the few studies are essential for understanding performance in occupational activities. In a study by Klemp and McClelland (1986), practical abilities for leadership were analyzed. For participation in this study, 150 successful senior managers were nominated; these subjects should give a detailed description of highly developed abilities and strategies for effective leadership behavior. Additionally, employees of the managers were also asked about abilities and strategies for effective leadership. According to Klemp and McClelland (1986), among the abilities and strategies that seemed to be characteristic of effective leadership in senior managers, the following eight deserve special mention: (a) planning of behavior and causal thinking, (b) synthetic and conceptual thinking, (c) active search for relevant information, (d) exerting control, (e) motivating employees, (f) ability to cooperate and teamwork, (g) serving as a model for others, (h) self-confidence and motivation. The study of Klemp and McClelland is not only illustrative of the impact of developed strategies on working performance, but also points to the demand for a continuous 'training on the job' as a precondition of higher job performance. Various studies on working expertise in older workers do show that continuous occupational activity may be associated with the development of specific skills and knowledge which can be used to compensate for age-related losses, especially in the fluid component of intelligence (e.g., Krampe 1994). In fact, meta-analyses suggest that there are no significant differences in the performance of younger and older workers (see Warr 1995). Moreover, empirical data do show that there is more variance in performance within than between age groups, i.e., age-related losses are a poor predictor of job performance (Salthouse 1984). However, age-related decline in biological and physiological processes may have an impact on performance in specific professions, e.g., occupational activities that are associated with severe and unbalanced strain of the motor system or with outside work under bad climatic conditions. Ilmarinnen et al. (1991) showed that a certain change in job profiles for older workers, e.g., reduction of physical strain (from 4,000 to 3,000 kg per shift in women packers) while handing over new experience-based duties (instruction of younger workers), can have a positive impact on health status and job performance in older workers. Results from empirical studies show that the assumption that older workers profit from a general change in job profiles from muscle work to brain work must be rejected. For example, losses in sensory functions can lead to chronic bad posture, which in turn can contribute to degenerative diseases of the motor system. Evaluation of various intervention programs showed that general conceptions and programs are less successful in maintaining functional status than specific programs for individual working conditions.

The principal demands for training in the context of occupational activities can be summarized as follows. (a) It has been shown that expertise in the field of work is to be regarded as a potential age-related gain which can be used to compensate for age-related losses; job performance of older people is not necessarily lower than job performance in younger people. Therefore, older people should have the opportunity to participate in occupational training programs to a greater

extent. (b) Bad working conditions do have a greater impact on job performance of older workers than on the job performance of younger workers. However, age-related losses in performance can be compensated for through changes in job profiles, which take into account specific job-related skills and experiences of older people. Skills and experiences of older workers can be used effectively for increasing job performance in younger workers.

4. *The Impact of Education on Healthy Aging*

Mirowsky and Ross (1998) tested three variants of the human-capital and learned-effectiveness hypothesis, i.e., the assumption that education improves health through fostering effective agency. Using data from a 1995 national telephone probability sample, these authors could show that (a) education enables people to integrate health-promoting behaviors into a coherent lifestyle, (b) education is associated with a sense of control, i.e., that outcomes in one's own life are not incidental but contingent upon intentional behavior, which in turn encourages a healthy lifestyle, (c) educated parents motivate a healthy lifestyle in their children. From the results of this research the lifelong effects of education in earlier phases of development on the possibility to maintain a personal satisfying perspective on life become apparent; older people do benefit from both education in adulthood and education in earlier phases of the life span. Consequently, from the perspective of life span development the positive effects of educational programs for younger people extend beyond jobs and earnings, i.e., they can contribute to the prevention of health problems in old age by enabling people to gain the ability to exert control over their own development. Moreover, even in old age it is not too late to foster perceptions of self-effectiveness and control; adult education can contribute to healthy aging by motivating health-producing behaviors and lifestyles.

See also: Adult Development, Psychology of; Education and Learning: Lifespan Perspectives; Education in Old Age, Psychology of; Education: Skill Training; Lifelong Learning and its Support with New Media: Cultural Concerns; Vocational Education and Training

Bibliography

Baltes P B, Baltes M M 1990 Psychological perspectives on successful aging: The model of selective optimization with compensation. In: Baltes P B, Baltes M M (eds.) *Successful Aging*. Cambridge University Press, New York

Berg C A, Klaczynski P A, Calderon K S, Strough J N 1994 Adult age differences in cognitive strategies: Adaptive or deficient? In: Sinnott J D (ed.) *Interdisciplinary Handbook of Adult Lifespan Learning*. Greenwood Press, Westport, CT

Denney N W 1994 The effects of training on basic cognitive processes: What do they tell us about the models of lifespan cognitive development? In: Sinnott J D (ed.) *Interdisciplinary Handbook of Adult Lifespan Learning*. Greenwood Press, Westport, CT

Horn J L 1982 The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. In: Craik F I M, Trehub S (eds.) *Aging and Cognitive Processes*. Plenum, New York

Ilmarinnen J, Louhevaara V, Korhonen O, Nygard H, Hakola T, Suvanto S 1991 Changes in maximal cardiorespiratory capacity among aging municipal employees. *Scandinavian Journal of Work, Environment and Health* **17**: 99–109

Klemp G O, McClelland D C 1986 What characterizes intelligent functioning among senior managers? In: Sternberg R J, Wagner R K (eds.) *Practical Intelligence in an Everyday World*. Cambridge University Press, New York

Kliegl R, Smith J, Baltes P B 1989 Testing-the-Limits and the study of adult age differences in cognitive plasticity and of mnemonic skill. *Developmental Psychology* **25**: 247–56

Krampe R T 1994 *Maintaining Excellence: Cognitive-Motor Performance in Pianists Differing in Age and Skill Level*. Springer, Berlin

Kruse A, Rudinger G 1997 Lernen und Leistung im Erwachsenenalter (Learning and performance in adulthood). In: Weinert F E, Mandl H (eds.) *Psychologie der Erwachsenenbildung*. Hogrefe, Göttingen

Labouvie-Vief G 1985 Intelligence and cognition. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*. Van Nostrand Reinhold, New York

Mirowsky J, Ross C E 1998 Education, personal control, lifestyle and health. A human capital hypothesis. *Research on Aging* **20**: 415–49

Oswald W D, Rödel G 1995 *Gedächtnistraining. Ein Programm für Seniorengruppen (Training of Memory. A Program for Seniors)*. Hogrefe, Göttingen

Salthouse T 1984 Effects of age and skill in typing. *Journal of Experimental Psychology* **113**: 345–71

Warr P 1995 Age and job performance. In: Snel J, Cremer R (eds.) *Work and Aging: a European Perspective*. Taylor and Francis, London

Willis S L 1987 Cognitive training and everyday competence. *Annual Review of Gerontology and Geriatrics* **7**: 159–89

Willis S L 1990 Cognitive training in later adulthood. *Developmental Psychology* **26**: 875–915

A. Kruse and E. Schmitt

Adult Mortality in the Less Developed World

Demographers study death as a process that reduces the size of populations and alters their characteristics. This process is mortality. The term adult mortality is sometimes used to refer to the mortality of everyone except children and sometimes to refer to the mortality of young and middle-aged adults but not the old. Either way, deaths of adults represent a growing proportion of all deaths in the less developed world. In response, health policy and research interest in adult

mortality has grown since the late 1980s. Despite this, statistics on adult mortality remain unreliable for much of the world. This article outlines what is known about adult mortality in less developed countries and places adult mortality in the context of debates about international health policy.

1. Introduction

Measures of mortality are required to calculate life expectancy. Demographers use data on current and past mortality by age to forecast mortality trends and produce population projections (see *Population Forecasts*). Epidemiologists use statistics on cause-specific mortality to investigate the etiology of diseases that kill adults. Social scientists investigate socioeconomic inequalities in adult mortality and behavior that influences adults' risk of dying. Thus, data on adult mortality are important inputs into health policy making and program evaluation, into actuarial work in the public and private sectors, and into economic and social planning more generally.

A life table provides a full description of mortality at all ages (see *Life Table*). Various measures based on the life table have been proposed as summary indices of adult mortality, including life expectancy at 15 years or some other age chosen to represent the onset of adulthood. However, to calculate life expectancy accurately one needs reliable data on old age mortality. No such data exist for most less developed countries. For this reason, and to distinguish the death of working age adults from mortality in old age, adult mortality is often measured using indices such as life table survivorship or partial life expectancy that summarize mortality between two ages representing the onset of adulthood and old age respectively. In particular, the World Bank and World Health Organization have adopted the life table probability of dying between exact ages 15 and 60 (denoted ${}_{45}q_{15}$) as their preferred measure of adult mortality (World Bank 1993, WHO 1999).

2. Data on Adult Mortality

The basic descriptive statistics on adult mortality in the less developed world are deficient. In most poor countries, few deaths are attended by physicians and no effective system exists to ensure that deaths are certified before a funeral can take place. Thus, civil registration of adult deaths is only complete in some small island states and about a dozen larger less developed countries. Even when deaths are registered, the information obtained on the cause of death is often inadequate. Moreover, inadequate and underfunded administrative systems frequently introduce further errors and delays into the production of statistical tables based on death certificates.

A few countries, such as India, have established sample systems for the registration of vital events. In addition, data on adult mortality can be collected using retrospective questions in surveys and censuses. Questions are used about deaths of household members in the year before the inquiry and about the survival of specific relatives, in particular parents and siblings (Timæus 1991). While some national statistical organizations regularly collect such data, others have never done so. Moreover, failure to report deaths or dead relatives and misreporting of ages at death or dates of death are sometimes major problems.

Data on adult mortality in the less developed world are usually analyzed using indirect methods (see *Demographic Techniques: Indirect Estimation*). The word 'indirect' originally denoted methods that estimate life table survivorship from unconventional data on the survival of relatives. In the measurement of adult mortality, however, it also describes methods that use models of demographic processes to evaluate and adjust conventional data on deaths by age.

Statistics on the mortality of old people in less developed countries are even more deficient than those on younger adults. Direct reports of deaths in old age are distorted by exaggeration of ages at death, and estimates made from data on the survival of relatives reflect the mortality of young and middle-aged adults. Moreover, so many deaths in old age are ascribed to senility and ill-defined causes that data on causes of death in elderly populations are frequently impossible to interpret.

In the absence of reliable data on adults, both international agencies such as the United Nations and national governments often publish life tables and summary indices of life expectancy that are based solely on data on children. Mortality in adulthood is imputed on the basis of infant or under-five mortality. However, adult mortality is only loosely associated with mortality in childhood and such estimates may be badly biased. Life tables produced in this way can overestimate or underestimate life expectancy at birth or at age 15 by several years. Such estimates can be very misleading and are not presented in this article.

3. Levels and Trends in Adult Mortality

In the lowest mortality countries in the developed world, only about 10 percent of men and 6 percent of women who survive to age 15 die before their 60th birthday (see *Life Expectancy and Adult Mortality in Industrialized Countries*). In 1909 in Chile, by contrast, about 63 percent of men and 58 percent of women who lived to age 15 died before age 60 (Feachem et al.1992, Chap. 2). Moreover, in India in the early twentieth century these statistics were about 77 and 71 percent, respectively (Mari Bhat 1989). In parts of francophone West Africa, adult mortality remained this high as late as the 1950s (Timæus 1993). By the 1970s though, such

Table 1
Probability of dying between ages 15 and 60 ($_{45}q_{15}$) per 1000, selected countries, 1998

Country	Women	Men
Uganda (1995)	556	547
Bangladesh	276	295
Tanzania (1996)	257	373
South Africa (1997)	237	395
Cameroon	226	275
Bolivia	219	271
Benin	198	246
Indonesia	184	236
India	182	230
Thailand	173	272
Brazil	166	292
Philippines	151	200
Pakistan	148	192
Vietnam	147	218
Uzbekistan	145	245
Algeria	120	147
Colombia	119	220
Senegal	118	127
Mexico	108	198
China	101	164
Cuba	101	141
South Korea	98	203
Chile	86	158
Sweden	63	97

Source: Africa: author's estimates; other countries: WHO (1999)

elevated mortality no longer existed except in populations afflicted by war and famine. Even in the highest mortality countries, at least half of those who lived to age 15 could expect to survive to age 60.

By the late 1990s, the level of adult mortality in a less developed country depended largely on whether it had developed a generalized epidemic of HIV. Most of the severely affected countries are in Eastern and Southern Africa (see *Mortality and the HIV/AIDS Epidemic*). For example, in Uganda, the probability of dying between ages 15 and 60 rose from about 25 percent in the early 1980s to about 55 percent in 1995, while in South Africa the rise was from around 24 percent in 1990 to about 35 percent in 1997 (see Table 1). More up-to-date statistics will only gradually become available but, by the end of the twentieth century, the probability of dying between 15 and 60 probably exceeded 50 percent across most of Eastern and Southern Africa.

Elsewhere in the less developed world, adult mortality has continued to fall. By the late 1990s, the probability of dying between ages 15 and 60 had dropped below 30 percent in all countries without generalized AIDS epidemics for which data exist. (Mortality is almost certainly higher than this in some countries that lack data, especially those with a history of war or civil war such as Afghanistan or Sierra Leone.) In some less developed countries, the probability of dying between ages 15 and 60 is now less than 10 percent for women and around 15 percent for men. These probabilities are well within the range found in the industrialized world. Indeed, adult men's mortality is much higher than this in most of Eastern Europe.

At the national level, adult mortality is associated only loosely with standards of living. It tends to be high in the least developed countries even if they have not been affected by AIDS. Yet, adult mortality is also rather high in some middle-income countries and has fallen to a fairly low level in parts of West Africa, for example, Senegal (see Table 1).

Except for a noticeable 'injuries hump' in early adulthood in the mortality schedule for men, death rates rise rapidly with age in most populations. Mortality decline disproportionately benefits young adults. Thus, in high mortality populations, the probabilities of dying between ages 15 and 45 and ages 45 and 60 are about equal but, in low mortality populations, people are about three times more likely to die in the older age range. Populations with severe HIV epidemics have very unusual age patterns of mortality. Most AIDS deaths occur at quite young ages and the risk of dying may decrease substantially in middle age before rising again in old age.

Middle-aged and elderly men always experience higher mortality than women. However, because of high rates of childbearing and associated risks, young women have higher mortality than men in parts of South Asia and Africa. Many middle-income countries, particularly in Latin America, are characterized by a large gender gap in mortality: adult men's mortality remains rather high but women's mortality is now quite low.

4. *Causes of Adult Death*

It is almost impossible to obtain useful information on causes of death in retrospect. Thus, accurate data on causes of death can be collected only in countries with an effective death registration system. Very few developing countries exist in which one can study over several decades the contribution to the mortality transition made by different causes of death. One such country is Chile.

Adult mortality was still fairly high in Chile in the 1950s (see Table 2). Nevertheless, communicable disease mortality at ages 15–60 was only slightly higher than mortality from cardiovascular disease. The non-communicable diseases, and cardiovascular disease in particular, account for a substantial proportion of adult deaths in all populations.

Between 1955 and 1986, the probability of dying between ages 15 and 60 in Chile more than halved, dropping more for women than men. Mortality from the communicable diseases fell most rapidly. By the

Table 2
Probability of dying between ages 15 and 60 ($_{45}q_{15}$) per 1000 by cause, Chile, 1955 and 1986

	Women		Men	
Cause of Death	1955	1986	1955	1986
Communicable & Maternal	74	9	85	16
Diarrhea	2	1	1	0
Tuberculosis	31	2	48	5
Sexually transmitted	1	0	2	0
Respiratory infections	24	3	31	8
Maternal	15	1	–	–
Noncommunicable	171	83	205	118
Neoplasms	46	36	38	32
Endocrine	3	3	2	3
Cardiovascular	50	19	68	31
Respiratory	2	2	3	2
Digestive	27	11	37	31
Ill-defined	26	5	29	8
Injuries	11	11	78	59
Unintentional	10	4	69	18
Suicide	1	1	5	5
Homicide	0	0	4	3
Undetermined	0	6	0	32
Total	256	103	369	193

Source: Calculated from Feachem et al. (1992, Chap. 2)

mid-1980s, they accounted for only 8 percent of adult mortality. Nevertheless, the noncommunicable diseases also made a substantial contribution to the decline in mortality. This is normal. The only diseases from which it is common for mortality to rise during the transition to low mortality are lung cancer, breast cancer, and perhaps diabetes.

In Chile, as in most countries, tuberculosis accounts for more of the decline in adult mortality than any other single disease (Timæus et al. 1996, Chap. 10). Falling mortality from respiratory infections and cardiovascular disease also had a substantial impact. In addition, a reduction in deaths associated with pregnancy and childbirth made a substantial contribution to the decline in women's mortality. Such deaths accounted for about 6 percent of women's mortality in adulthood in 1955 but less than 1 percent in 1986.

Low mortality countries like Chile in the less developed world usually have higher infectious and digestive system disease mortality than industrialized countries. However, mortality from cardiovascular disease is substantially lower (Timæus et al.1996, Chaps. 7–9, 14). In contrast, the main reason why adult mortality remains rather high in some middle-income countries is that noncommunicable disease mortality, and in particular cardiovascular mortality, has fallen by less than in Chile.

Injuries are an important cause of death in adulthood, especially for young men. The level of mortality from injuries varies markedly between countries, largely reflecting the incidence of fatal road traffic accidents, homicide, suicide, and war deaths. Injuries partly account for the relatively high mortality of men in much of Latin America (Timæus et al. 1996, Chaps. 17, 18). For example, men's mortality from injuries remains high in Chile despite the substantial contribution that the external causes made to the overall reduction in adult mortality during the three decades prior to 1986 (see Table 2).

Except in the highest mortality countries, the major noncommunicable diseases, along with influenza and pneumonia, are probably the most important causes of death in old age. For example, the most important causes of death in old age in Latin America are much the same as in North America (PAHO 1982).

In high mortality countries, most child deaths are from infections. As the risk of dying from infections has shrunk, infant and child mortality have fallen more rapidly than adult mortality (see *Infant and Child Mortality in the Less Developed World*). More children now survive to adulthood and adult mortality is emerging as a residual problem of growing *relative* importance. This process is often described as the 'epidemiological transition' (see *Mortality, Epidemiological, and Health Transitions*). Communicable disease and child health are being replaced by noncommunicable disease and adult health as the most important health issues confronting society.

The impact on the age structure of deaths of declining communicable disease mortality is being compounded by changes in the age structure of the

population. The last third of the twentieth century saw the onset of fertility decline in most of the less developed world. As a result, their populations are beginning to age (see *Population Aging: Economic and Social Consequences*). About 56 percent of the developing world's population was aged 15–59 in 1985 and about 6 percent aged 60 years or more. These proportions are projected to rise to 62 percent and 9 percent respectively by 2015.

5. *Socioeconomic Inequalities in Adult Mortality*

Data on socioeconomic inequalities in adult mortality in the less developed world are even scarcer than data on mortality trends and causes of death. The little information that is available suggests that differentials in adult mortality are large. In Peru, for example, survey data on the fathers of respondents aged 25 to 29 showed that 72 percent of educated fathers were alive, compared with only 55 percent of uneducated fathers, representing a difference of 17 percent (World Bank 1993). Lesotho is a small rural, ethnically homogenous Southern African country. Even in this context, an *absolute* difference of about 14 percent in the probability of dying between ages 15 and 60 existed in the 1970s between individuals with uneducated families and those with well-educated families (Timæus 1993).

The few data that exist on socioeconomic differentials in adult mortality by cause reveal patterns that are broadly consistent with what one might infer from trends in mortality. Data from China's Disease Surveillance Points show that women living in wealthier areas of the country have much lower mortality than women living in poor areas (Feachem et al. 1992, Chap. 2). Noncommunicable disease mortality is higher in the poorer localities but the differential is much smaller than that in mortality from the communicable diseases and injuries. This suggests that inequalities in adult mortality may shrink as overall mortality falls. Equally, interventions directed at communicable disease may be more equitable than those directed at noncommunicable disease because they benefit the poor disproportionately.

6. *Determinants of Adult Health*

The fundamental determinant underlying much adult ill-health in developing countries is poverty. For instance, poor quality and overcrowded housing and malnutrition are important risk factors for the airborne infectious diseases, including tuberculosis, that are common among adults. Similarly, use of cheap polluting fuels can cause chronic respiratory disease; inadequate storage and processing of subsistence crops and other foodstuffs can raise the incidence of cancers and other diseases of the digestive system; and living in a city in a less developed country often exposes the poor to an insecure and stressful environment and to the risk of violence.

In some developing countries as many as one third of deaths in adulthood may be linked to infections and other conditions acquired in childhood (Mosley and Gray 1993). For example, low birth weight is a risk factor for chronic respiratory disease in later life. In much of Africa and Asia, most people are infected with hepatitis B as children and 5–15 percent of them become chronic carriers. At least one quarter of adult carriers of hepatitis B die of cirrhosis or primary liver cancer.

The most worrying recent development affecting the health of adults in the less developed world is the epidemic of HIV/AIDS in Eastern and Southern Africa (see *Mortality and the HIV/AIDS Epidemic*). The less severe or more localized epidemics of HIV/AIDS found in other regions may yet spread and affect millions more people. This could produce a massive rise in adult mortality worldwide.

Development exposes adults to new risks to their health. Many of the known behavioral risk factors for premature death in adulthood are becoming more prevalent in the less developed world. In particular, the prevalence of tobacco smoking has grown rapidly. Among men, it is now more common than in the developed world. A massive epidemic of smoking-related deaths will inevitably follow during the next few decades. In countries where occupational health legislation is nonexistent or unenforced, both industrial work and the growing mechanization of agriculture are associated with a high incidence of fatal injuries and poisonings. Moreover, as the number of motor vehicles grows, mortality from road traffic accidents tends to rise.

7. *Adult Health Policy*

Until the late 1980s, few researchers and experts working on public health in the less developed world were interested in adult health and mortality. This reflects the history of international health policy. The 1970s saw a drive to redirect resources from hospital-based medicine into primary health care. The concept of primary health care was originally an integrative one, which encompassed the health of adults as well as children (WHO 1978). As a program for action, however, primary health care has concentrated on a limited number of interventions intended to reduce mortality from common communicable diseases. The United Nations Children's Fund energetically promoted the idea of a 'child survival revolution' and many agencies and governments came to implicitly or explicitly concentrate their efforts in the health sector on the reduction of child mortality (Reich 1995).

In part, the recent growth in concern about adult health and mortality is a response to the growth in the

relative importance of adult ill-health and to the epidemic of AIDS in Africa. It has also been realized that, although adult ill-health still absorbs much of the health care budget in many less developed countries, tertiary hospitals no more benefit poor adults in rural areas than they do their children. The World Bank became a major donor in the health sector by the late 1980s allowing it to promote interest in such issues. Adult health fits well with the World Bank's overall mission to promote economic development.

Despite the growing heterogeneity of the developing world, research at the World Bank (Feachem et al. 1992, Chap. 6, Jamison et al. 1993) has identified several measures that are priorities for the improvement of adult health almost everywhere. The first such measure is a negative one: withdraw public expenditure from ineffective health programs and those that benefit only a few adults. Such programs include the nonpalliative treatment of most cancers, medical management of hypertension, and antiviral therapy for adults with AIDS. Cost-effective measures to reduce adult mortality that should be available more widely include effective referral systems for obstetric emergencies, the treatment of sexually transmitted infections, tuberculosis, and leprosy, and screening for cervical cancer. Various preventive measures delivered to children, in particular hepatitis-B immunization, are priorities for the improvement of adult health. Several crucial interventions, such as fiscal measures to reduce smoking, fall outside the area of responsibility of ministries of health.

Perhaps even more than for children, effective action to reduce adult mortality is likely to meet with political opposition. For example, measures intended to reduce the prevalence of smoking will be fought by the tobacco industry and perhaps the ministry of finance; screening for cervical cancer may conflict with cultural norms about the seclusion of women; and in some countries attempts to reduce expenditure on inappropriate tertiary care have been blocked by the medical profession. Most government policy in the developed world is determined without considering its impact on mortality. Establishing the reduction of adult mortality as a major goal of governments in the less developed world is likely to be a slow process.

See also: Infant and Child Mortality in the Less Developed World; Mortality, Biodemography of; Mortality Differentials: Selection and Causation

Bibliography

Feachem R G A, Kjellstrom T, Murray C J L, Over M, Phillips M A 1992 *The Health of Adults in the Developing World*. Oxford University Press, Oxford, UK

Jamison D T, Mosley W H, Measham A R, Bobadilla J L 1993 *Disease Control Priorities in Developing Countries*. Oxford University Press, New York

Mari Bhat P N 1989 Mortality and fertility in India, 1881–1961: A reassessment. In: Dyson T (ed.) *India's Historical Demography: Studies in Famine, Disease and Society*. Curzon Press, London, pp. 73–118

Mosley W H, Gray R 1993 Childhood precursors of adult morbidity and mortality in developing countries: Implications for health programs. In: Gribble J N, Preston S H (eds.) *The Epidemiological Transition: Policy and Planning Implications for Developing Countries*. National Academy Press, Washington, DC, pp. 69–100

Pan American Health Organization (PAHO) 1982 *Health Conditions in the Americas 1977–80*. PAHO, Washington, DC

Reich M R 1995 The politics of agenda setting in international health: Child health versus adult health in developing countries. *Journal of International Development* **7**: 489–502

Timæus I M 1991 Measurement of adult mortality in less developed countries: A comparative review. *Population Index* **57**: 552–68

Timæus I M 1993 Adult mortality. In: Foote K A, Hill K H, Martin L G (eds.) *Demographic Change in Sub-Saharan Africa*. National Academy Press, Washington, DC, pp. 218–55

Timæus I M, Chackiel J, Ruzicka L (eds.) 1996 *Adult Mortality in Latin America*. Clarendon Press, Oxford, UK

World Bank 1993 *World Development Report 1993. Investing in Health*. Oxford University Press, New York

World Health Organization (WHO) 1978 *Alma Ata—Primary Health Care*. WHO, Geneva, Switzerland

World Health Organization (WHO) 1999 *The World Health Report 1999*. WHO, Geneva, Switzerland

I. Timæus

Adult Psychological Development: Attachment

In the last decade, Bowlby's attachment theory (1973) has become an important framework for understanding psychological development in adulthood. Theory and research has mainly focused on a person's style of relating to significant others ('attachment style') as well as on his/her beliefs about the self and the world ('internal-working models'). Initially, empirical efforts have been spent in examining the manifestations of attachment style in interpersonal behavior as well as in the quality of close relationships. With the progress of theory and research, more efforts have been invested in delineating the manifestations of attachment style in the process of affect regulation. In this article are presented the basic concepts of attachment theory and the psychological implications of attachment style are reviewed.

1. Attachment Style and Internal Working Models in Adulthood

According to Bowlby (1973), an attachment system evolved in humans to help maintain proximity to significant others (e.g., parents, lovers, friends) under

conditions of danger. Proximity maintenance can help the individual to manage distress with the help of other persons and to attain a sense of 'felt security' in the world. However, although all individuals need to maintain proximity to significant others in stressful situations, there are individual differences in the activation of attachment behaviors and in the extent to which people seek others' proximity and support. These individual differences reflect the pattern learned throughout the history of interactions with significant others. Whereas persons who have a history of positive interactions may develop a positive orientation towards proximity maintenance, persons who have interacted with cold and rejecting others may have serious doubts about the effectiveness of proximity maintenance as a way of obtaining comfort and security.

Attachment theory and research have conceptualized the above individual differences in terms of attachment styles—stable patterns of cognition and behaviors in close relationships. In adulthood, Hazan and Shaver (1987) proposed three prototypical attachment styles (secure, avoidant, anxious-ambivalent) that corresponded to the typical attachment typology in infancy. Brennan et al. (1998) concluded that this typology could be organized around two dimensions: avoidance and anxiety. Persons scoring low in these two dimensions correspond to the secure style and are characterized by a positive history of social interactions, comfort with proximity seeking, and confidence in others' availability in times of need. Persons scoring high in the avoidance dimension correspond to the avoidant style and are characterized by insecurity in others' goodwill and preference for social and emotional distance from others. Persons scoring high in the anxiety dimension correspond to the anxious-ambivalent style and are defined by insecurity in others' responses and an anxious and ambivalent approach to loved persons. Some recent studies have distinguished a subgroup of insecure persons who score high in both anxiety and avoidance dimensions (fearful persons) and tend to indiscriminately combine features of the avoidant and anxious-ambivalent styles.

Several self-report measures have been constructed tapping a person's attachment style (see Brennan et al. 1998 for a review). Some of these measures adopt a typological approach and ask persons to endorse the attachment style that best fits their feelings in close relationships. Other measures adopt a dimensional approach and ask persons to rate themselves along the various dimensions of attachment organization (e.g., avoidance, anxiety). Empirical efforts have also been invested in developing interview procedures, but most studies still employ self-report scales. It is important to note that all these measures assess global attachment style in adulthood rather than attachment orientation in a particular relationship or memories of childhood experiences. However, despite the tremendous development of measurement tools, more empirical work should be done in the construction of assessment techniques (e.g., observational) that could overcome the problems inherent in self-report measures.

In explaining the formation of attachment style in adulthood, attachment research has adopted Bowlby's (1973) concept of 'internal working models.' According to Bowlby, every interaction with significant others is mentally represented in terms of others' availability and responsiveness to one's attachment needs, the worthiness of the self, and the efficacy of proximity maintenance as a distress management device. In this way, people develop cognitive representations of the self and others that are generalized to new relationships and seem to be the source of continuity between past experiences and the attitude and expectations that we bring with us to current interactions. Bowlby labeled these representations as internal working models and viewed them as the building blocks of a person's attachment style. Collins and Read (1994) proposed that working models in adulthood include four components: (a) memories of attachment-related experiences, (b) beliefs and expectations about significant others and the self, (c) attachment-related goals, and (d) strategies related to the regulation of attachment needs.

According to Collins and Read (1994), persons differing in attachment style may differ in the quality of autobiographical memories of concrete episodes with significant others. Although Bowlby (1973) emphasized that these memories may be accurate reflections of a person's interactions, they may be reconstructed throughout the life span and may reflect the current organization of attachment experiences. Indeed, secure persons, compared to insecure persons, have been found to recall their parents as more available and responsive and to represent relationship histories in more positive and affectionate terms (see Shaver et al. 1996 for a review).

Attachment-style difference may also exist in a person's beliefs and expectations about significant others and the self (see Shaver et al. 1996 for a review). People who feel secure in their relationships may be prone to perceive others as loving and responsive and to feel valued by them. In contrast, people who feel insecure in their relationships may be prone to perceive others as cold and rejecting and may feel worthless in their eyes. In support of this view, secure persons, compared to insecure persons, are more likely to hold positive beliefs and expectations about their romantic partner and to explain partner's behaviors in positive and relationship-enhancing terms. Moreover, secure persons have been found to report higher self-esteem than anxious-avoidant and fearful persons. Interestingly, avoidant persons also hold positive self-views. However, whereas secure persons hold a positive self-view that is balanced by acknowledgment of negative aspects of the self, avoidant persons are reluctant to recognize these negative self-aspects.

The third component of internal working models concerns the goals people pursue in social interactions. Secure persons' positive experiences with responsive partners may teach them that attachment behaviors are rewarding and that they can continue to organize interpersonal behaviors around the basic goal of the attachment system—proximity maintenance. As a result, secure persons tend to construe their interaction goals around the search for intimacy and closeness. Insecure persons' experiences with nonresponsive others teach them that attachment experiences are painful and that other interaction goals should be developed as defenses against the insecurity caused by these experiences. In response to this insecurity, anxious-ambivalent persons hyperactivate the attachment system, construct their interaction goals around security seeking, and seek to minimize distance from others via clinging and anxious responses. In contrast, avoidant persons deactivate the attachment system and organize their interaction goals around the search for personal control and self-reliance.

The fourth component of internal working models concerns the strategies people use for achieving interaction goals and managing distress. Secure persons' interactions with supportive partners teach them that the attachment system is an effective device for attaining comfort and relief. As a result, these persons may learn to manage distress through the basic guidelines of the attachment system: acknowledgement of distress, engagement in constructive actions, and turning to others for support (Collins and Read 1994). In contrast, insecure persons learn that attachment behaviors are ineffective regulatory devices and that other defensive strategies should be developed (Bowlby 1988). Whereas anxious-ambivalent persons scoring tend to hyperactivate distress-related cues and aggrandize the experience of distress, avoidant persons tend to deactivate these cues and inhibit the acknowledgment and display of distress.

Overall, attachment research has delineated the cognitive substrate of adult attachment style. However, more research is needed examining the contribution of childhood experiences, family environment, parents' personality factors, and the person's own temperament to the development of internal working models. Accordingly, more research should be conducted on the specific ways the various components of these working models are manifested in interpersonal behavior and affect regulation.

2. *Attachment Style and Interpersonal Behavior*

According to Bowlby (1973), internal working models styles may shape the ways people interact with others and construe their close relationships. In support of this view, a growing body of research has documented attachment-style differences in the quality of close relationships and interpersonal behavior (see Shaver and Hazan 1993 for a review). However, it is important to note that most of the studies only present cross-sectional associations between self-reports of attachment style and interpersonal phenomena. Therefore, they cannot draw firm conclusions about causality (whether attachment style is an antecedent of interpersonal behaviors) as well as about the psychological mechanisms that explain these associations.

A review of research of attachment research allows us to delineate the pattern of interpersonal behaviors that characterize each attachment style. Overall, secure persons are highly committed to love relationships and tend to maintain them over long periods of time. Intimacy, supportiveness, trust, reciprocity, stability, and satisfaction characterize their romantic relationships. They tend to resolve interpersonal conflicts by discussing them with the partner and reaching integrative, relationship-enhancing solutions. They have a positive orientation toward intimate interactions and tend to disclose personal feelings to loved persons as a means of improving the quality of the relationship.

The pattern of interpersonal behaviors of avoidant persons seems to be characterized by attempts to maximize distance from partners, fear of intimacy, and difficulty depending on others. Specifically, these persons have been found to report low levels of intimacy, passion, and commitment in romantic relationships. They tend to have unstable, short-term relationships, and to grieve less than secure persons following a break-up. They feel bored during social interactions, do not like to disclose personal feelings to other persons, and do not like other persons who share intimate knowledge about themselves. They are pessimistic about romantic relationships and tend to withdraw from the partner in times of stress. Interestingly, they tend to use work and cognitive activities as an excuse for avoiding close relationships.

Obsession about partner's availability, emotional instability, worries about being abandoned, lack of satisfaction, strong physical attraction, jealousy, and a passionate desire for union characterized anxious-ambivalent persons' love relationships. They tend to construct highly conflictive relationships and to suffer from a high rate of break-up. They indiscriminately disclose their personal feelings without taking into consideration the partner's identity and responses; display argumentative and overcontrolling responses towards romantic partners; rely on strategies that aggrandize rather than reduce interpersonal conflicts; and elicit negative responses from partners. Overall, anxious-ambivalent persons' pattern of interpersonal behaviors reflects a demand of compulsive attachment from others, which may create relational tension, may result in the breaking-up of the relationship, and may exacerbate their basic insecurity and fear of rejection.

Importantly, the above patterns of interpersonal behaviors are also manifested in the nature and quality

of marital and same-sex friendship relationships. For example, husbands' and wives' attachment security seems to be related to less frequent use of destructive responses in marital conflict and to more positive marital interactions. Moreover, the marital relationship of secure spouses is characterized by more intimacy, cohesiveness, supportiveness, and flexibility than the marital relationship of insecure spouses. Secure persons also tend to have more intimate and rewarding same-sex friendships than insecure persons. Accordingly, they tend to be more committed to these relationships, to experience less conflictive friendships, and to engage in more selfless and playful interactions with same-sex friends than insecure persons.

3. *Attachment Style and Affect Regulation*

Attachment theory is highly relevant to the process of affect regulation and coping with stress. In Bowlby's (1988) terms, the attainment of a sense of felt security is an inner resource, which helps people to buffer distress. It seems to consist of expectations that stressful events are manageable, a strong sense of self-efficacy, and confidence in others' goodwill, which together evolve in an optimistic and constructive attitude toward life. As a result, secure persons tend to show better adjustment, less negative affect, and more moderate emotional reactions to stressful events than insecure persons (see Mikulincer and Florian 1998 for a review).

Attachment security is also involved in the adoption of adaptive ways of affect regulation. Studies have consistently found that secure persons attempt to manage distress by enacting effective coping responses (instrumental strategies, support seeking), coordinating attachment with other behavioral systems (exploration, affiliation), and acknowledging distress without being overwhelmed by it. There is also evidence that a sense of felt security allows people to revise erroneous beliefs and to explore strong and weak self-aspects. In this way, secure persons could develop more flexible and adjusted views of the world and the self and more reality-tuned coping plans (see Mikulincer and Florian 1998 for a review).

Along the above reasoning, insecure attachment seems to be a risk factor, which hinders well-being and leads people to adopt maladaptive ways of coping. Research has consistently shown that insecure persons tend to appraise stressful events in threatening and catastrophic terms and to have serious doubts about their abilities to deal with these events. With regard to ways of coping, avoidant persons tend to distance themselves from emotion-laden material and to show low cognitive accessibility of negative emotions. Moreover, they tend to suppress bad thoughts, to repress painful memories, and to escape from any confrontation with distress-eliciting sources of information. In contrast, anxious-ambivalent persons tend to experience an overwhelming arousal of negative emotions and an undifferentiated spreading of this arousal to irrelevant emotional themes. Moreover, they tend to mentally ruminate over negative thoughts and to approach distress in an hypervigilant way (see Mikulincer and Florian 1998 for a review).

The above strategies of affect regulation have also been found to underlie perceptions of the self and others. In dealing with stress, avoidant persons tended to inflate their positive self-view and to perceive other persons as different from themselves. Their attempt to suppress personal deficiencies favors self-inflation, whereas their attempt to maximize distance from others results in underevaluation of self–other similarity. In contrast, anxious-ambivalent persons tend to deal with stress by devaluating their self-view and perceiving other persons as similar to themselves. Their attempts to hyperactivate personal weaknesses and to elicit others' love favor self-devaluation, whereas their attempts to create an illusion of connectedness result in heightened self–other similarity. As a result, these persons tend to devaluate others. Interestingly, secure people hold more moderate and realistic views of the self and others. Their sense of felt security allows them to regulate affect without distorting mental representations.

4. *Concluding Remarks*

Attachment research clearly indicated that Bowlby's theory is a relevant framework for understanding psychological development in adulthood. Attachment style seems to be a core feature of adult personality that shapes our perception of the world and the self and guide how we interact with others, how we construe our close relationships, and how we regulate and manage distress. However, one should note that research has made only a first step in delineating attachment-style differences in interpersonal behavior and affect regulation. More research is needed tapping the involvement of attachment style in other areas of adult life as well as in the various facets of psychological development in adulthood stages. Accordingly, further research should examine the stability of attachment style, the formation, maintenance, and dissolution of particular attachments, and the protective function of attachment behaviors.

See also: Adult Development, Psychology of; Attachment Theory: Psychological; Bowlby, John (1907–90); Evolutionary Social Psychology; Love and Intimacy, Psychology of; Psychological Development: Ethological and Evolutionary Approaches; Self-esteem in Adulthood

Bibliography

Bowlby J 1973 *Attachment and Loss: Separation, Anxiety and Anger*. Basic Books, New York

Bowlby J 1988 *A Secure Base: Clinical Applications of Attachment Theory*. Routledge, London

Brennan K A, Clark C L, Shaver P R 1998 Self report measurement of adult attachment: An integrative overview. In: Simpson J A, Rholes W S (eds.) *Attachment Theory and Close Relationships*. Guilford Press, New York, pp. 46–76

Collins N L, Read S J 1994 Cognitive representations of attachment: The structure and function of working models. In: Bartholomew K, Perlman D (eds.) *Attachment Processes in Adulthood*. Jessica Kingsley, London, pp. 53–92

Hazan C, Shaver P 1987 Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology* **52**: 511–24

Mikulincer M, Florian V 1998 The relationship between adult attachment styles and emotional and cognitive reactions to stressful events. In: Simpson J A, Rholes W (eds.) *Attachment Theory and Close Relationships*. Guilford Press, New York, pp. 143–65

Shaver P R, Collins N L, Clark C L 1996 Attachment styles and internal working models of self and relationship partners. In: Fletcher G J O, Fitness J (eds.) *Knowledge Structures in Close Relationships: A Social Psychological Approach*. Erlbaum, Mahwah, NJ

Shaver P R, Hazan C 1993 Adult romantic attachment: Theory and evidence. In: Perlman D, Jones W (eds.) *Advances in Personal Relationships*. Jessica Kingsley, London, pp. 29–70

M. Mikulincer

Adulthood: Dependency and Autonomy

Dependency strictly means the ongoing need for external support (e.g., from family members, professionals, state institutions, intensive care units, or assistive devices) in order to fulfill individual or societal expectations regarding what is a 'normal' life. A less strict interpretation of dependency also encompasses human needs for affiliation, attachment, and bonding to significant others, such as to one's partner, children, grandchildren, or close friends (Baltes and Silverberg 1994). Autonomy can be defined as 'a state in which the person is, or feels, capable of pursuing life goals by the use of his or her own resources' (Parmelee and Lawton 1990, p. 465). Autonomy thus means independent and effective functioning in a variety of life domains ranging from basic activities of daily living to complex decision processes.

Historically, developmental researchers have primarily examined the dynamics between dependency and autonomy from childhood to adolescence. In old age, dependency and loss of autonomy were long viewed as direct consequences of aging worthy of systematic description (e.g., in epidemiological studies), but basically inevitable and irrevocable. It was only in the 1970s that new research findings—based on a more optimistic image of old age and stimulated by a *social learning perspective*—demonstrated the plasticity of dependency and autonomy in old age. While research in the 1980s and 1990s significantly furthered this insight, middle adulthood, and also the transitions in dependency and autonomy that are experienced as the adult individual matures, have attracted little scientific interest.

1. On the Complexities of Dependency and Autonomy in Adult Life

From a life-span perspective, childhood and adolescence are periods when striving toward autonomy and reducing dependency are among the most important *developmental tasks* (Havighurst 1972). By young adulthood, or at the very latest by middle adulthood, one is normally expected to have accomplished this successfully. Conversely, old age may be characterized, at least to some extent, as a life period that poses the risk of becoming dependent or losing one's autonomy. However, the general assumption that autonomy gradually replaces dependency and then dependency gradually replaces autonomy over the life course is clearly simplistic.

Cultural relativity becomes particularly obvious in the autonomy–dependency dynamics across the life span. For example, while the developmental goal of maintaining autonomy in a wide variety of life domains over the life span is one of the highest values in most Western cultures, one of the most 'normal' elements of many developing countries' cultures is reliance on children in the later phases of life. Second, although autonomy and dependency play their roles as individual attributes, both should be regarded predominantly as contextual constructs depending strongly on situational options and constraints. For example, the dependent self-care behavior of an 85-year-old man or woman may not reflect physical or mental frailty at all, but may primarily result from the overprotective behavior of family members and professionals (Baltes 1996). Third, autonomy and dependency should both be regarded as multidimensional, that is, gain in autonomy in one life domain does not automatically lead to reduced dependency in other life domains and vice versa. For example, being able to meet the everyday challenges of life in an independent manner does not necessarily prevent a younger individual from relying strongly on parents or significant others when making crucial life decisions (such as selecting a partner). Fourth, and finally, autonomy and dependency have strong value connotations which shape action. In Western cultures, independent behaviors are generally regarded as posi-

tive and highly adaptive, worth supporting by all means, whereas dependency has negative value connotations and should be avoided at all costs. Such global value attributions can be questioned in terms of life complexity and richness. For example, emotional dependency upon another person lies at the heart of mature intimate relationships. Conversely, striving for autonomy may become detrimental when confronted with severe chronic illness, which necessitates help, support, and the delegation of control to the external environment. These differentiations have to be kept in mind as we examine autonomy and dependency in middle and old age more closely.

2. *Dependency and Autonomy—A Challenge in Midlife?*

Answers to this question must start with the developmental challenges associated with the life phase that occurs between roughly 45 and 65 years of age. According to life-span theorist Erik Erikson (1963), one central challenge of human midlife is *generativity*, that is, support of and solidarity with the following generations not only in terms of concrete child-rearing activities, but also in the broader sense of transfer of one's own life expertise and experiences to others. Coping with the 'empty nest,' establishing bonds to grown-up children, and becoming a grandparent are other typical challenges to be met in this life phase. With respect to women, the experience of menopause is the most typically age-graded event in this life period, while one of the most challenging nonnormative events of modern middle-aged women (very much less so for men) is taking over the care-giving role for one's father or mother.

What kinds of implications do these midlife challenges have with respect to autonomy and dependency? As Baltes and Silverberg (1994) have persuasively argued, the term *interdependence* best encompasses agency and autonomy *and* attachment and bonding needs in the middle years. Interdependence typically is at work in generativity, where the older generation is dependent on the younger generation in order to feel needed (particularly after children have left home). At the same time, interdependence can strengthen feelings of agency, competence, and autonomy by providing a forum for the transfer of advice, expertise, and 'world knowledge.' The new role of being a grandfather or grandmother further adds to interdependence, particularly in women. As has been argued (Hagestad 1985), women more easily express interdependence and attachment and thus contribute more to the maintenance of family and kinship. Furthermore, interdependence finds expression in social relations in midlife by helping others and receiving help, that is, by mutuality and reciprocity, thus contributing to the establishment of a *social convoy* which guarantees social support throughout life (Kahn and Antonucci 1980). However, providing care for a chronically ill parent may also negatively impact on well-balanced interdependence: the moral obligation to care for a family member, and also the loss of autonomy one endures in order to provide such care, can lead to severe psychological distress in midlife (Pearlin et al. 1996).

3. *On the Many Faces of Dependency and Autonomy in Old Age*

With respect to later life, Margret M. Baltes (1996) has introduced distinctions among structured, physical, and behavioral dependency. *Structured dependency* implies that human worth is determined primarily by participation in the labor force. According to this understanding, society is in a sense producing dependency among certain subgroups (such as the aged) in order to provide opportunity structures for other subgroups (such as younger persons). *Physical dependency* is closely linked to the age-related emergence of chronic physical and mental impairments such as severe mobility loss or Alzheimer's disease. *Behavioral dependency* primarily reflects the effect of the interaction between the older person and his/her social environment, thus emphasizing most strongly of all three forms of dependency the role of social and behavioral processes in order to explain dependency in the later years. Owing to the understanding of dependency in social and behavioral terms, a whole empirical research program has been conducted by Margret Baltes and associates (for an overview, see Baltes 1996). The major and very robust finding was that social interactions between older adults and social partners are prone to a *dependence–support script*, that is, dependent behavior among elders is reinforced by their social partners (typically by direct helping behavior), whereas autonomous behaviors are typically ignored.

Furthermore, structured, physical, and behavioral dependency interact with each other, that is, macrosocial, biological, and psychological processes must be considered in conjunction. Three theoretical approaches have been discussed to address this dynamic interplay and its outcomes. First, dependency in old age may be due in part to *learned helplessness*, that is, an existing noncontingency between action goals and consequences (Seligman 1975). Macrosocial (see again the term structured dependency) and noncontrollable biological impairments may contribute to dependency in this vein. Second, dependency may also be interpreted as a strong instrument for *passive control*. In this regard, the dependence–support script found in the Baltes research also reflects the power of dependent behaviors of elders to provoke positive responses (typically help) from their social environment and thus

maintain environmental control as long as possible. The adaptational value of this strategy is particularly suggested by the *life-span theory of control* (Heckhausen and Schulz 1995). Third, an even more complex interpretation of the dynamics between dependency and autonomy in old age comes into the play when adaptation and the striving for a successful course of aging are viewed in terms of *selective optimization with compensation* (Baltes and Baltes 1998). In this respect, dependent behaviors in certain life areas (such as basic care needs) may be seen as a powerful compensation tool in the late phase of the human life span in order to prepare the field for further optimization and development in selected other life domains with high personal priority.

4. Conclusions and New Research Directions

The dynamic interplay between dependency and autonomy occurs in different variations across the adult human life span. While the creation and regulation of interdependence is among the major tasks of middle adulthood, the particularities of old age call for the consideration of different pathways to explain the etiology and maintenance of dependency in old age. The insight of the behavioral model of dependency, pointing to the different roles that dependency can take as an adaptive tool in old age, also means a challenge for intervention. Owing to the potential role of dependent behaviors for successful aging, intervention efforts should always be framed within a change philosophy that leaves the elderly person in control of whether external autonomy-enhancing efforts should be exerted or not. Autonomy in old age, in this sense, may also mean the autonomy to decide for dependence in some life domains, even if this might mean the risk of losing available competencies in the long run by *disuse*.

New research directions should follow at least three avenues. First, research should examine forms of interdependency in midlife, with a view to understanding how they affect dependency in old age. Second, the interaction between social and physical environmental influences and personality traits, and also their effect on fostering dependency or maintaining autonomy, deserves more consideration. Third, more knowledge is needed on the day-to-day balancing and rebalancing of dependency and autonomy and their emotional and behavioral outcomes. For this kind of empirical research, the model of selective optimization with compensation (e.g., Baltes and Baltes 1998) and the life-span theory of control (Heckhausen and Schulz 1995) are probably the best theoretical alternatives to date.

See also: Autonomy, Philosophy of; Caregiving in Old Age; Control Behavior: Psychological Perspectives; Control Beliefs: Health Perspectives; Education in Old Age, Psychology of; Learned Helplessness; Social Learning, Cognition, and Personality Development

Bibliography

Baltes M M 1996 *The Many Faces of Dependency in Old Age*. Cambridge University Press, London

Baltes P B, Baltes M M 1998 Savoir vivre in old age. *National Forum: The Phi Kappa Phi Journal* **78**: 13–18

Baltes M M, Silverberg S 1994 The dynamic between dependency and autonomy: illustrations across the life span. In: Featherman D L, Lerner R M, Perlmutter M (eds.) *Life-span Development and Behavior*. L. Erlbaum, Hillsdale, NJ, Vol. 12, pp. 41–90

Erikson E H 1963 *Childhood and Society*, 2nd edn. Norton, New York

Hagestad G O 1985 Continuity and connectedness. In: Bengtson V L, Robertson J F (eds.) *Grandparenthood*. Sage, Beverly Hills, CA, pp. 31–48

Havighurst R J 1972 *Developmental Tasks and Education*, 3rd edn. McKay, New York

Heckhausen J, Schulz R 1995 A life-span theory of control. *Psychological Review* **102**: 284–304

Kahn R L, Antonucci T C 1980 Convoys over the life course: attachment, roles, and social support. In: Baltes P B, Brim O G (eds.) *Life-span Development and Behavior*. Academic Press, New York, Vol. 3, pp. 253–86

Parmelee P A, Lawton M P 1990 Design for special environments for the elderly. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 3rd edn. Academic Press, San Diego, CA, pp. 464–87

Pearlin L I, Aneshensel C S, Mullan J T, Whitlatch C J 1996 Caregiving and its social support. In: Binstock R H, George L K (eds.) *Handbook of Aging and the Social Sciences*, 4th edn. Academic Press, San Diego, CA, pp. 283–302

Seligman M E P 1975 *Helplessness: on Depression, Development, and Death*. W. H. Freeman, San Francisco

H.-W. Wahl

Adulthood: Developmental Tasks and Critical Life Events

1. Systems of Influences on Individual Development

'Developmental tasks' and 'critical life events' represent concepts that are crucial to the framework of life-span developmental psychology, which views development as underlying, at least, three systems of influences, namely age-graded, non-normative, and history-graded (Baltes 1979). Age-grading refers to the extent to which the life span is structured and organized in time by age. Developmental psycholo-

gists have given attention to the age-relatedness of major transitions, as is represented by the concept of developmental tasks. Sociologists consider the life course as shaped by various social systems that channel people into positions and obligations according to age criteria.

In contrast, the system of non-normative influences on development is related to the concept of critical life events. These events are defined as clearly nonage-related, to occur with lower probability (as is true for many history-graded events), and to happen only to few people. Accordingly, these events are highly unpredictable, happening more or less 'by chance,' and occurring beyond the individual's control. Not surprisingly, they have been equated with 'the stress of life' threatening an individual's physical and psychological well-being, hence, were primarily the focus of clinical psychologists or epidemiologists.

Finally, history-graded events are defined as confronting large portions of the population at a given point in time, irrespective of peoples' ages or life-circumstances. Of particular interest is their differential impact depending on *when* within the life span they occur, although such a truly developmental perspective has rarely been adopted (for an exception see Elder 1998). Only recently the radical socio-historical changes associated with Germany's reunification have gained similar interest in terms of how they affect developmental trajectories in different age or birth cohorts (see Heckhausen 1999).

2. Developmental Tasks

Havighurst (1952) was one of the first to describe age-normative transitions by introducing the concept of developmental tasks, a particular class of challenges to individual development that arise at or about a certain period of time in the life span. Developmental tasks are seen to be jointly produced by the processes of biological maturation, the demands, constraints, and opportunities provided by the social environment, as well as the desires, aspirations, and strivings that characterize each individual's personality. These tasks as being age-normative has a twofold, although often not satisfactorily differentiated meaning. First, the concept refers to the statistical norm indicating that within a given age span a particular transition is (statistically) normal. Second, it also has a prescriptive connotation, indicating that at a given age individuals are expected to and have to manage certain transitions.

Evidently, due to its focus on (statistical) age-normativity, the traditional concept of developmental tasks cannot account for the overwhelmingly high variability that characterizes developmental processes particularly in adulthood. The life course is age-graded, but members of a birth cohort do not always move through it in concert according to the social roles they occupy. Some people do not experience certain transitions (e.g., parenthood), and those who do experience these transitions vary in the timing of events. In fact, the loose coupling between transitions and their age-related timing is highly reflective of individuals as producers of their own development, as is nowadays highlighted within action-orientated models of development (Brandtstädter 1998). Nevertheless, the concept has inspired the idea of age-gradedness of the life span, and developmental tasks have gained renewed interest as they are represented in individuals' normative (i.e., prescriptive) conceptions about (their own) development. These developmental conceptions set the stage for developmental prospects and goals to be attained within particular age spans ('on time') in a twofold way.

They guide social perception as age-related stereotypes (Filipp and Mayer 1999), and they inform the individual about the 'optimal' timing of investments in their development. In this latter sense, they represent nothing else than a mental image of one's own development that guides decisions and actions. Obviously, the self-regulatory power of the developing individual can be integrated into that theoretical perspective by conceiving of developmental tasks as organizers of developmental regulation (Heckhausen 1999).

Developmental tasks, moreover, share some common meaning with other traditional concepts. As is well known, Erikson (1959) proposed eight successive stages made up by a sequence of age-normative challenges. In addition, he focused on the disequilibrium associated with the normative shift of one developmental stage into another. At each of these stages, individuals are forced to manage the conflict between contradicting forces, for example, between generativity vs. stagnation (or self-absorption) in middle adulthood and between ego integrity vs. despair in old age. Similarly to the successful mastery of developmental tasks, resolution of these crises is seen to be a prerequisite of further growth; if this cannot be accomplished, various forms of psychological dysfunction may result.

Some theorists have incorporated elements of Erikson's approach into their conceptions of adult development, e.g., in addressing generativity as the developmental task of middle adulthood (e.g., Bradley 1997). At that time, individuals are seen to become 'senior members' of their worlds and to be responsible also for the development of the next generation of young adults. Yet, studies have provided mixed results on timing issues (McAdams and de St Aubin 1992, Peterson and Klohnen 1995). In addition, generativity has been conceived in terms of agency and communion, the latter representing the more mature form which is manifested through openness and union with others and in which life interest is invested in the next generation. Agentic generativity, in contrast, exists if a creation transferred to them is simply a monument to

the self, i.e., is associated with self-protection and self-absorption. Snarey (1993) has postulated that generativity is composed of three semihierarchical substages, namely biological, parental, and societal. According to his findings, having been an actively involved father at the age of 30 was linked to the expression of broad generative concerns at midlife. Marital satisfaction proved to be the strongest predictor of fathers' parental and societal generativity, underscoring that successful mastery of the preceding developmental task (intimacy vs. isolation) does in fact contribute to successful mastery of later developmental tasks.

3. *Critical Life Events*

Within the tradition of life-event research, the issue of what constitutes critical life events and of how their impact should be measured has been discussed extensively (see Filipp 1992). Various suggestions have been made, ranging from the disruptiveness or amount of change in people's lives apart from its meaning or direction up to multidimensional conceptions of what makes life experiences particularly critical ones. In general, critical refers to the fact that these types of events may be equated with turning points in the individual life span that result in one of three developmental outcomes: psychological growth, return to the precrisis level of functioning (as is stressed in homeostatic models), or psychological and/or physical dysfunctioning. Such a notion is widely acknowledged by crisis models of development, according to which transitions imply both danger and opportunity for growth. The same holds true for critical life events, as is inherent in the etymological origin of the word 'crisis' itself.

One of the most substantial contributions of a developmental perspective on 'the stress of life' was that the meaning of critical life events varies also according to their normative timetable. For instance, work place instability has different meanings across various ages, i.e., being more common before the age of 30, but being experienced rarely after the age of 50 and, thus, being more stressful in later years. Thus, life events may be considered 'critical' because they violate normative conceptions of an expected life span, e.g., death of spouse during middle rather than late adulthood. As deviations from the expected life course, off-time events can set in motion a series of off-time sequences. Due to their lack of normativity and affective quality, they evoke extremely strong reactions, provide the individual with a sense of undesirable uniqueness, and foster them to dramatically alter their conceptions of a 'good life.' In addition, critical life events are seen to interfere with the successful mastery of developmental tasks and the attainment of goals, people have set for themselves. Consequently, they bring about the necessity to disengage from commitments and to replace them with new options and goals—coping tasks that are particularly painful to accomplish (Filipp 1998). Furthermore, people normally are not 'taught' how to deal effectively with loss and crisis. Neither are they taught respective lessons at school, nor do they usually learn from models how to cope with critical life events. And even when such models were available, people rather prefer to look at the sunnier sides of life. According to the widely held belief in one's invulnerability and unrealistic optimism (Taylor and Brown 1994), people usually do not consider critical events as one of the possible realities they have to confront in their lives.

In that respect, one could borrow a term from cognitive psychology to conceive of critical life experiences as 'weakly scripted situations,' for which ways of acting (let aside behavioral routines) are not readily at hand. Some life events that accompany middle and old age are, at least partially, embedded into culturally shaped ways of responding (e.g., public rituals), often facilitating the coping process. Other events, like the initial diagnosis of cancer, represent existential plights and are seen to cause behavioral disorganization and fruitless attempts to find meaning in one's fate. In addition, almost all types of these events imply a threat to fundamental beliefs about the self (e.g., as being powerful or lovable) and the necessity to alter the self-system. Consistent with predictions of identity interruption theory (Burke 1996), one can assume that many ways of coping (like ruminative thinking, search for meaning) are ultimately related to and in the service of reconstructing the self-system. From that point of view, critical life events hardly allow for the notion of an individual that proactively regulates their development. Rather, they often put individuals in a purely reactive role for a long time, before they regain a secure basis for setting personal goals again.

4. *Conclusions*

In sum, both concepts, developmental tasks as represented in normative conceptions of development, as well as critical life events, have enriched our insight into the dynamics of life-span development. They focused our attention on road maps for human lives and regular life paths, on the one hand, and on the developmental plasticity during critical turning points within the life span, on the other. Nevertheless, a differential perspective needs to be adopted in order to account for the tremendous variability in developmental trajectories. For example, much of this work has been conducted with respect to male development in middle adulthood. Nevertheless, evidence of different developmental pathways for men and women in a variety of life domains is growing now, at least with regard to later adulthood (Smith and Baltes 1998), and this evidence definitely needs extension to the middle

years. In addition, individual strategies to cope with life challenges and demands need to be taken into consideration.

See also: Adult Development, Psychology of; Adult Psychological Development: Attachment; Coping across the Lifespan; Job Stress, Coping with; Parenthood and Adult Psychological Developments

Bibliography

Baltes P B 1979 Life-span developmental psychology: Some converging observations on history and theory. In: Baltes P B, Brim O G Jr (eds.) *Life-span Development and Behavior*. Academic Press, New York, Vol. 2, pp. 256–79

Bradley C L 1997 Generativity–stagnation: Development of a status model. *Developmental Review* **17**: 262–90

Brandtstädter J 1998 Action perspectives on human development. In: Lerner R M (ed.) *Theoretical Models of Human Development: Handbook of Child Psychology*, 5th edn. Wiley, New York, Vol. 1, pp. 807–63

Burke P J 1996 Social identity and psychological stress. In: Kaplan H B (ed.) *Psychological Stress*. Academic Press, San Diego, CA, pp. 141–74

Elder G H Jr 1998 *Children of the Great Depression: Social Change in Life Experience*, 25th anniversary edn. Westview Press, Boulder, CO

Erikson E H 1959 *Identity and the Life Cycle. Psychological Issues Monograph 1*. International University Press, New York

Filipp S-H 1992 Could it be worse? The diagnosis of cancer as a traumatic life event. In: Montada L, Filipp S-H, Lerner M J (eds.) *Life Crises and Experiences of Loss in Adulthood*. Erlbaum, Hillsdale, NJ, pp. 23–56

Filipp S-H 1998 A three stage model of coping with loss and trauma: Lessons from patients suffering from severe and chronic disease. In: Maercker A, Schützwohl M, Solomon Z (eds.) *Post-traumatic Stress Disorder: A Life-span Developmental View*. Hogrefe and Huber, Seattle, WA, pp. 43–80

Filipp S-H, Mayer A-K 1999 *Bilder des Alters*. Kohlhammer, Stuttgart, Germany

Havighurst R J 1952 *Developmental Tasks and Education*. McKay, New York

Heckhausen J 1999 *Developmental Regulation in Adulthood. Age-normative and Socio-structural Constraints as Adaptive Challenges*. Cambridge University Press, Cambridge, UK

McAdams D P, de St Aubin E 1992 A theory of generativity and its assessment through self-report, behavioral acts, and narrative themes in autobiography. *Journal of Personality and Social Psychchology* **62**: 1003–15

Peterson B E, Klohnen E C 1995 Realization of generativity in two samples of women at midlife. *Psychology and Aging* **10**: 20–9

Smith J, Baltes M M 1998 The role of gender in very old age: Profiles of functioning and everyday life patterns. *Psychology and Aging* **13**: 676–95

Snarey J 1993 *How Fathers Care for the Next Generation*. Harvard University Press, Cambridge, MA

Taylor S E, Brown J D 1994 Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin* **116**: 21–7

S.-H. Filipp

Adulthood: Emotional Development

1. Introduction and Historical Overview

Research on emotion in adult development and aging this century has been quite limited until recently and for a long time was without significant theoretical guidance. This lack of attention to emotion in human development stemmed chiefly from two phenomena: (a) the inherent complexity of emotion—which hindered definitional clarity; and (b) a long-standing Western bias against emotion, with emotion seen as an impediment to reason and rationality, rather than a psychological function in its own right.

During the 1960s and early 1970s a number of important theoretical contributions were made by Silvan Tomkins, Carroll Izard, Paul Ekman, and Robert Plutchick, building on the earlier work by Charles Darwin and William James. The latter, 'discrete emotions' theorists postulated a limited number of basic emotions—sadness, anger, fear, shame, joy, interest, contempt, disgust, and surprise—each having distinctive neurophysiological, physiognomic, motivational, and phenomenological properties. For example, there are unique motivational properties associated with each emotion: fear motivates flight, anger aggression, shame withdrawal, and so forth. It should also be mentioned that emotions are also conceptualized as having dimensional features within the discrete emotions framework; that is, emotions can vary in frequency, intensity, hedonic tone, and arousal level. These discrete emotions theories played an important role in bringing conceptual clarity to the study of emotion; they also served to challenge and undermine the view that emotions were merely disruptive and maladaptive forces in human life.

Important empirical work was conducted during the late 1970s. During this time, Ekman and Izard took issue with earlier theories which had proposed that emotions were undifferentiated. Using cross-cultural data, they were among the first to document the universality of human emotions, as well as the differential facial patterning of the fundamental emotions. Later, Robert Levenson and colleagues were able to demonstrate the differential physiologically patterning.

Although these studies led the way, a real groundswell in empirical research on the emotions did not occur until the late 1970s and early 1980s. The first inroads were made in the field of child development, and social and personality psychology. Within short order there was a rapidly expanding cadre of child, clinical, and social-personality psychologists doing very innovative research on emotion, but by and large the developmental studies were limited to the period of infancy, which was guided by a coherent body of theory. In contrast, there was little theoretical guidance for researchers interested in the adult years. For example, of the four theorists noted above, only Izard

took a developmental stance; however, almost all of his own empirical research was devoted to the study of emotional development during infancy and early childhood.

A few scattered studies on adult development that appeared during the 1970s and 1980s seemed to suggest that the course of development over the adult years was one of decline, including contradictory trends indicating affective blunting on the one hand and increased negativity on the other. However, this research was not theory driven and was based largely on institutional samples of older people. Research on emotion in adult development and aging using non-institutionalized persons only began to gain critical mass during the 1990s—largely owing to theoretical contributions that placed emotional development within a lifespan framework, including the differential emotions theory of Izard (newly expanded to actively engage issues of adult development and aging), Laura Carstensen's socio–emotional selectivity theory, and Gisela Labouvie-Vief and Powell Lawtons' regulatory models. Izard proposed that while there are changes in expressive behavior over the course of development, there is a core feeling state associated with each discrete emotion that remains constant across the lifespan. Carstensen's theory links emotions to social process, and proposes that individuals select and maintain social interactions to regulate affect and maintain self identity. In addition, Lawton's model suggests that adults engage in behavior that seeks to maintain an optimum level of arousal and that older adults become more proficient at affect regulation. Finally, Labouvie-Vief's cognitive-affective model suggests that emotion regulation is related to an individuals' cognitive-developmental level. Recent empirical research, by and large, supports these propositions. In the following, we consider this literature within the more general question of whether emotions and related socio-emotional processes undergo change over the lifecourse.

At the outset it is important to note that there are two fundamental aspects of emotions, one involving their motivational value, the other, social process. That is, emotions provide motivational cues to the self, directing the self to engage in flight, attack, approach, and so forth; because they also indicate preparedness to respond in certain ways, they provide social signals to others, inviting potential social partners to approach, to retreat, to avoid, to protect the self, and so on. Most of the basic emotions, which are part of a prewired set of behavioral propensities, depend simply on maturation for their development, and most emerge in the behavioral repertoire by the second year of life. However, they undergo modification during childhood in accordance with 'display rules,' a set of proscriptions concerning who can show what emotions, under what circumstances, to whom, and in what form. These display rules vary by culture, historical epoch, and familial patterns. Children learn to modulate and regulate their expressive behavior so as to achieve a fit with their culture and family, modifying the intensity and directness with which emotions are expressed. One may well ask whether there is anything about emotional processes that undergo change over the adult years after basic patterns have become established. Since emotions involve underlying neurophysiological patterns as well as behavioral patterns and feeling states, investigators have been interested in determining whether there are physiological changes in emotions with age, changes in expressivity, and changes in subjectivity over the adult years; they have also examined whether people's ability to regulate their emotions changes in adulthood and old age.

2. *Changes Over the Lifecourse*

2.1 *Are There Changes in Physiological Patterns?*

Given that the emotion system is grounded in basic neurophysiological processes, and given that there are widespread structural and physiological changes in the various organ systems of the human body over the adult years, including a decline in vision and hearing as early as the twenties, a slowing of metabolism with age, and neurological cell fallout in later life, we may well expect to see changes in the emotion system over the adult years. Unfortunately, there has been relatively little research in this area, with the exception of the work of Levenson and colleagues (Levenson 1992, Levenson et al. 1991). In this body of work, the autonomic nervous system responsivity of participants was monitored while they recollected and relived salient emotional events or while they assumed patterned facial expressions. Under these conditions, Levenson was able to demonstrate that there are emotion-specific patterns that distinguish between the emotions of anger, fear, sadness, and disgust. In studies of older people, he found that the latter exhibited the same emotion-specific patterns as younger people; the magnitude of the response was lower in older subjects. However, the older subjects reported the same degree of subjective emotional arousal.

2.2 *Are There Changes in the Expression of Emotion?*

The bulk of research on the nonverbal communication of emotion during this century and even during Darwin's time has been conducted on facial expressions. Research has shown that there is increasing conventionalization of facial expressions across the childhood years, which in large measure involves

adopting cultural and familial display rules and includes a general dampening of expressive behavior; there is far less research on adulthood. Although patterns of muscular activity remain basically the same, for example, oblique brows that signal sadness in children, signal sadness in younger and older adults as well, Carol Malatesta-Magai and colleagues have found several distinguishing differences in older vs. younger adult faces (see the review in Magai and Passman 1998). In one study, younger and older participants were videotaped during an emotion-induction procedure in which they relived and recounted emotionally charged episodes involving four basic emotions. Older individuals (50 years old or above) were found to be more emotionally expressive than younger subjects in terms of the frequency of expressive behavior across a range of emotions; they expressed a higher rate of anger expressions in the anger-induction condition, a higher rate of sadness during the sadness induction, greater fear under the fear-induction condition, and greater interest during the interest condition.

In another study, older adults were found to be more expressive in another sense. Malatesta-Magai and Izard videotaped and coded the facial expressions of young, middle-aged and older women while they recounted emotional experiences. Using an objective facial affect coding system, they found that while the facial expressions of the older vs. younger women were more telegraphic in that they tended to involve fewer regions of the face, they were also more complex in that they showed more instances of blended expressions where signals of one emotion were mixed with those of another. This greater complexity of older faces appears to pose a problem for those who would interpret their expressions. Young, middle-aged, and untrained 'judges' attempted to 'decode' the videotaped expressions of the women in the above study. With the objectively coded material serving as the index of accuracy, Malatesta-Magai and colleagues found that judges had the greatest difficulty with and were most inaccurate when decoding older faces; however, the accuracy with which judges decoded expressions varied with age congruence between judges and emotion expressors, suggesting a decoding advantage accruing through social contact with like-aged peers (Magai and Passman 1998).

Another aspect of facial behavior that appears to change with age has to do with what Ekman has called 'slow sign vehicle' changes—changes accruing from the wrinkle and sag of facial musculature with age. Malatesta-Magai has also noted a personality-based effect involving the 'crystallization' of emotion on the face as people get older; that is, emotion-based aspects of personality seem to become imprinted on the face and become observable as static facial characteristics in middle and old age. In one study, untrained decoders rating the facial expressions of older individuals expressing a range of emotions made a preponderance of errors; the errors were found to be associated with the emotion traits of the older expressors.

2.3 Are There Changes in the Subjectivity of Emotion?

Research suggests that older individuals are more likely to orient to the emotional content of their worlds than younger individuals, in the sense that emotion becomes a more salient experience for them, though experienced emotions may not be necessarily any more or less intense. Carstensen and colleagues tested older and younger participants for recall of narrative material they had read; they found that older individuals recalled more of the emotional vs. neutral material. In terms of changes in the intensity of emotion, the work of Levenson and colleagues indicates that younger and older persons do not differ from each other in the subjective intensity of emotions induced in the laboratory. Outside of the laboratory, there are conflicting results. Some studies find that there are no differences in frequency and intensity of reported positive and negative emotion across the adult years, whereas other studies indicate increasing positive affect in older individuals, at least up until late life. In late life (over 85 years of age), there appears to be a modest decline in positive affect, though there is not a corresponding increase in negative affect (Staudinger et al. 1999). The above pattern of results refute the earlier belief that aging is accompanied by an increase in negative affect (Carstensen et al. 1998, Levenson et al. 1991, Magai and Passman 1998).

Going beyond the issue of frequency and intensity, Labouvie-Vief and colleagues (Labouvie-Vief 1998) have examined the complexity of emotional experience. To this end they coded narrative transcripts of individuals recounting emotional experiences. They found that younger individuals rarely referred to inner subjective feelings, tended to describe their experiences in terms of norms and conventions, and controlled their emotions through such metacognitive strategies as forgetting or distracting the self. In contrast, older individuals were least bound by social convention, were more likely to refer to inner subjective states, and were more capable of discussing complex feelings and enduring states of conflict and ambivalence. Work by Lawton and colleagues (Lawton 1989) and Malatesta-Magai and colleagues (Magai and Passman 1998) has also supported the finding that older individuals are more verbally expressive of their subjective feelings. Other researchers have found that older individuals are more willing to engage in painful self-disclosure with an unfamiliar social partner than are younger individuals.

The above work suggests that people become more comfortable with their emotional selves as they age, that they are more likely to acknowledge their emo-

tional states, and that they are more capable of sustaining and reporting more complex inner affective lives. This body of work, however, does not directly address the issue of continuity of feeling states in the sense of Izard's constancy theory. While core feelings states associated with the discrete emotions may or may not remain essentially unchanged, it appears there is at least greater cognitive elaboration of inner subjectivity with age.

2.4 Are There Changes in the Ability to Regulate Emotion over the Adult Years?

It has long been noted that there is a narrowing of social networks in later life. Carstensen has proposed that this narrowing is an adaptive strategy people use to regulate emotion, a strategy that is crucial for the maintenance of well-being in later life, and which is linked to the need to conserve energy. A series of studies from her laboratory has supported this view (Carstensen et al. 1998). Older people indicate that they restrict their social contacts to those with whom they are most intimate; despite this narrowing of social networks, emotional communication is preserved, if not enhanced.

Research by Lawton and colleagues (1989) points to both an increasing ability to regulate emotion with age and the use of an optimizing emotion regulation strategy. That is, he has proposed that individuals actively create environments that permit them to achieve an optimal mix of emotionally stimulating vs. insulating features. In fact, studies have substantiated greater self-regulatory capacities in older individuals, with older people being higher in emotional control and emotional maturity through moderation. Older individuals were more likely to indicate they deliberately chose activities that would allow them to achieve just the right level of emotional stimulation.

In summary, the various strands of research that have been published since 1985 have materially advanced our understanding of emotion processes during the adult years. They take issue with earlier findings from research with institutionalized subjects and challenge stereotyped views of aging. Instead of revealing a bleak picture of diminished affective capacity and an inexorable drift towards negative affect, this latest generation of research suggests that individuals become more emotionally attuned and more emotionally complex with maturity. There are two caveats to this more comforting picture. The first is that emotional processes during advanced old age and during the chronic debilitating diseases may look quite different, as some recent studies by Lawton, Magai, and research of the Berlin Aging Study (Baltes and Mayer 1999) indicate. The second is that all of the foregoing research has been cross-sectional in nature and is thus potentially confounded by cohort effects; longitudinal research is crucial for advancing the state of knowledge in this area.

See also: Culture and Emotion; Emotion and Expression; Emotion, Neural Basis of; Emotional Inhibition and Health; Emotions, Evolution of; Emotions, Psychological Structure of; Infancy and Childhood: Emotional Development; Self-regulation in Adulthood

Bibliography

Baltes P B, Mayer K U (eds.) 1999 *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, Cambridge, UK

Carstensen L L, Gross J J, Fung H H 1998 The social context of emotional experience. *Annual Review of Gerontology and Geriatrics* **17**: 325–52

Ekman P, Davidson R J 1994 *The Nature of Emotion: Fundamental Questions*. Oxford, UK

Izard C E 1996 Differential emotions theory and emotional development in adulthood and later life. In: Magai C, McFadden S-H (eds.) *Handbook of Emotion, Adult Development, and Aging*, pp. 27–42

Labouvie-Vief G 1998 Cognitive–emotional integration in adulthood. *Annual Review of Gerontology and Geriatrics* **17**: 206–37

Lawton M P 1989 Environmental proactivity and affect in older people. In: Spacapan S, Oskamp S (eds.) *The Social Psychology of Aging*. Sage, Newbury Park, CA, pp. 135–63

Levenson R W 1992 Autonomic nervous system differences among emotions. *Psychological Science* **3**: 23–7

Levenson R W, Friesen W V, Ekman P et al. 1991 Emotion, physiology and expression in old age. *Psychology and Aging* **6**: 28–35

Magai C, Passman V 1998 The interpersonal basis of emotional behavior and emotion regulation in adulthood. *Annual Review of Gerontology and Geriatrics* **17**: 104–37

Staudinger U M, Freund A M, Linden M, Maas I 1999 Self, personality, and life regulation: facets of psychological resilience in old age. In: Baltes P B, Mayer K U (eds.) *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, Cambridge, UK, pp. 302–28

C. Magai

Adulthood: Prosocial Behavior and Empathy

Prosocial behavior represents a broad category of acts that are 'defined by society as generally beneficial to other people and to the ongoing political system' (Piliavin et al. 1981, p. 4). This category includes a range of behaviors that are intended to benefit others, such as helping, sharing, comforting, donating, or volunteering, and mutually beneficial behaviors, such as cooperation. Research on prosocial behavior has addressed not only the antecedents and consequences of these actions, but also the different motivations that may underlie these behaviors. Batson (1998), for

example, defines altruism as a motivational state with the ultimate goal of increasing another person's welfare, in contrast to egoistically motivated action, which has the ultimate goal of improving one's own welfare.

1. Research Trends

Research in this area has developed in several stages. The work of the early and mid-1960s typically focused on norms, such as social responsibility and reciprocity, that seemed to govern prosocial behavior. By the end of that decade and into the early 1970s, investigators, stimulated by public outrage at bystander apathy, investigated factors that reduced the likelihood of intervening in crisis and emergency situations. As researchers also explored the reasons why people *do* engage in prosocial activities, they began to consider more fully the role of empathy and developmental influences. In the 1980s, interest in prosocial actions declined somewhat and the questions moved primarily from *when* people help to *why* people help. Researchers frequently attempted to understand fundamental motivational processes, considering how different affective consequences of empathy could produce either egoistic or altruistic motivation. Research in the 1990s provided a clearer link between prosocial motivations and general personal, social, and intergroup orientations.

2. When Do People Help?

Research has identified a range of social and situational factors that influence helping and other prosocial actions (Schroeder et al. 1995). In terms of social factors, people are more likely to engage in prosocial activities for others with whom they are more closely related, with whom they are more similar, and with whom they share group membership; they are less likely to respond prosocially when others are seen as more responsible for their plight or otherwise undeserving of assistance. With respect to situational factors, people are more likely to help in situations that are more serious and clear. They are less likely to help when they believe that others are present and will take action, which relieves a bystander from having to assume personal responsibility for intervention.

Both affective (e.g., arousal and emotion) and cognitive (e.g., norms and perceived costs and rewards) processes are hypothesized to underlie prosocial behavior (see Dovidio and Penner in press). Cognitively, Latané and Darley's (1970) decision model of emergency intervention proposes that whether or not a person helps depends upon the outcomes of a series of prior decisions. This model has also been applied to nonemergency situations. Alternatively, a cost–reward analysis of prosocial action assumes an economic view of human behavior. In a potential helping situation, a person analyzes the circumstances, weighs the probable costs and rewards of alternative courses of action, and then arrives at a decision that will result in the best personal outcome. Current research is consistent with the central tenet of the cost–reward approach. The role of arousal in helping and other prosocial actions relates to the process of empathy.

3. Empathy

Empathy is 'an emotional response that stems from another's emotional state or condition, is congruent with the other's emotional state or condition, and involves at least a minimal degree of differentiation between the self and the other' (Eisenberg and Fabes 1990, p. 132). Empathy can be shaped by three different types of role taking (Davis 1994): (a) perceptual, relating to the capacity to imagine the visual perspective of another; (b) cognitive, involving beliefs about the thoughts, motives, and intentions of another person; and (c) affective, concerning inferences about another's emotional state. Perspective taking not only influences the degree of empathy that people experience but also, in combination with other cognitive processes, shapes the nature of the emotion that is experienced. Empathy can produce self-focused emotions, such as sadness or distress, or other-oriented emotions, such as sympathy or empathic concern.

There is substantial empirical evidence that people are fundamentally empathic and emotionally responsive to the needs of others. People are aroused physiologically and subjectively by the distress of others. This reaction appears even among very young children and occurs across cultures. In addition, people show greater empathy toward others who are closer and more similar to them. In fact, empathy is such a strong and universal phenomenon that some researchers have proposed that empathic arousal has evolutionary origins and a biological basis.

Although people appear to be generally inherently empathic, there are also systematic individual differences in dispositional empathy. Studies of identical twins, for example, have supported the heritability of empathy (see Schroeder et al. 1995). In addition, there is a general, positive association between dispositional empathy (and perceptual, cognitive, and affective perspective taking) and a broad range of prosocial actions (Davis 1994, Graziano and Eisenberg 1997).

3.1 Empathy and Emotion

Although most researchers agree that empathic arousal is important, there is much less agreement about the nature of this emotion and how it actually motivates prosocial behavior. Empathy can produce different emotions depending on the context. In severe emergency situations, bystanders may become upset

and distressed; in less critical, less intense problem situations, observers may feel sad, tense, or concerned and sympathetic (Batson 1998, Dovidio and Penner in press). How empathic arousal is interpreted, in turn, elicits either egoistic or altruistic motivation.

3.2 Empathy and Egoism

Models of egoistic motivation posit that helping is directed by the primary goal of improving one's own welfare; the anticipated benefit to others is secondary. From this perspective, empathy elicits self-oriented emotions (e.g., sadness or distress) that are experienced as unpleasant and motivate actions, such as helping, that are perceived to relieve them. Two of these approaches are the negative-state relief model and the arousal: cost–reward model.

According to the *negative-state relief model* (Cialdini et al. 1987), feelings of guilt or sadness motivate people to engage in behaviors that will improve their mood. Because through socialization and experience helping becomes self-rewarding in adults, helping represents one such behavior. Three fundamental implications of the negative-state relief model have received support (see Schroeder et al. 1995). First, a variety of negative states, including guilt from having personally harmed a person and sadness from simply observing another person's unfortunate situation, can motivate helping. Second, other events besides helping (e.g., receiving praise) may just as effectively make people feel better, and exposure to these events can thus relieve the motivation to help caused by negative states. Third, negative moods motivate helping only if people believe that their moods can be improved by helping. Negative feelings will *not* promote helping if people are led to believe that these feelings cannot be relieved or if, as with younger children, the self-rewarding properties have not yet developed.

Affective empathy can produce other emotional reactions, such as distress and upset, particularly among bystanders in emergencies. According to the Piliavin et al. (1981) *arousal: cost–reward model*, empathic arousal is generated by witnessing the distress of another person. When the bystander's empathic arousal is attributed to the other person's distress, it is emotionally experienced by the observer as unpleasant. The observer is therefore motivated to reduce it. The person then weighs various costs and rewards for helping or not helping, and chooses a course of action that relieves the unpleasant arousal while minimizing net costs and maximizing rewards. One normally efficient way of reducing this arousal is by helping to relieve the other's distress. Thus, from this perspective, arousal motivates a bystander to take action, and the cost–reward analysis shapes the direction that this action will take.

Supportive of the arousal: cost–reward model, empathic arousal attributed to the other person's situation motivates helping. Facial, gestural, and vocal indications of empathically induced arousal, as well as self-reports of empathically induced anxiety, are all positively related to helping (see Schroeder et al. 1995). Consistent with the hypothesized importance of attributing this arousal to the other's situation, people are more likely to help when arousal from extraneous sources such as exercise, erotic films, and aggressive films is attributed to the immediate need of another person. People are less likely to help when arousal generated by witnessing another person's distress is associated with a different cause (e.g., misattributed to a pill). In addition, work by Eisenberg and her associates (see Eisenberg and Fabes 1998) suggests that extreme empathic overarousal or the inability to regulate empathic arousal, which may interfere with the attribution process, can also reduce helpfulness.

3.3 Empathy and Altruism

In contrast to egoistic models of helping, Batson and his colleagues (see Batson 1998) present an *empathy–altruism hypothesis*. Although they acknowledge that egoistically motivated helping occurs, Batson and his colleagues argue that true altruism also exists. The primary mechanism in the empathy-altruism hypothesis is the emotional reaction to another person's problem. Empathy that is experienced emotionally as compassion and concern (i.e., as 'empathic concern') produces altruistic motivation. Because altruistic motivation has the primary goal of improving the other person's welfare, an altruistically motivated person will help if (a) helping is possible, (b) helping is perceived to be ultimately beneficial to the person in need, and (c) helping personally will provide greater benefit to the person in need than would assistance from another person also able to offer it.

In numerous experiments, conducted over a 20-year period, Batson and his colleagues have produced impressive empirical support for the empathy–altruism hypothesis (Batson 1998). Participants who experience relatively high levels of empathic concern (and who presumably are altruistically motivated) show high levels of helpfulness even when it is easy to avoid the other person's distress, when they can readily justify not helping, when helping is not apparently instrumental to improving the benefactor's own mood, and when mood-improving events occur prior to the helping opportunity. However, several researchers have proposed alternative explanations that challenge Batson's contention that helping can be altruistically motivated. These explanations have focused on (a) how feelings of empathic concern may be associated with special costs for not helping or rewards for helping, (b) how feelings of sadness that are aroused along with empathic concern actually are the primary determinants of helping, and (c) how manipulations used to induce empathic concern for another person

also create a greater sense of self-other overlap, or 'oneness,' so that helping may also have direct and primary benefits for the helper (see Cialdini et al. 1997). However, despite the critiques and controversies about aspects of the empathy–altruism hypothesis, the preponderance of evidence from more than 20 years of experimentation on this question strongly suggests that truly altruistic motivation may exist and that all prosocial behavior is not necessarily egoistically motivated.

4. *Sustained Helping*

In general, empathy has been viewed as much more influential for short-term, spontaneous forms of helping than for long-term, sustained volunteering, which has been hypothesized to be motivated by a range of personal (e.g., gaining knowledge, being among friends) and humanitarian goals (Clary et al. 1998). Nevertheless, developing a self-image of being empathic and helpful can produce longer-term commitments to helping others. For instance, the female gender role is often associated with the trait of 'communion,' being caring, emotionally expressive, supportive, and nurturant (Eagly and Crowley 1986). As a consequence, women may interpret empathic arousal in different ways than men and engage in different types of helping. In particular, although both men and women experience similar levels of physiological arousal when they observe distress in others, women may be more likely to interpret this arousal as a positive empathic response to the other person's needs. In accord with these findings, women are more likely than men to provide their friends with personal favors, emotional support, and informal counseling about personal or psychological problems. In contrast, consistent with their traditional gender role of being 'heroic' and effective, men are more likely than women to intervene in emergencies involving personal threat and to engage in instrumental forms of prosocial activities (Eagly and Crowley 1986). More generally, regular and public commitments to helping (such as donating blood or volunteering for charities), which may have initially been stimulated by feelings of empathy, can lead to the development of a role-identity consistent with those behaviors (Piliavin and Charng 1990, Penner and Finkelstein 1998).

5. *Conclusion*

The social context, the nature of the situation, the characteristics of the person in need, and the personality of the potential helper not only affect assessments of costs and rewards and the decisions about whether to engage in prosocial acts, but also shape empathic responses. How empathy is experienced emotionally is fundamental; empathy can produce either egoistic or altruistic motivation. Thus empathy may represent a basic mechanism for translating genetic prosocial predispositions, which may have evolutionary benefits, into action.

See also: Prosocial Behavior and Empathy: Developmental Processes

Bibliography

Batson C D 1998 Altruism and prosocial behavior. In: Gilbert D T, Fiske S T, Lindzey G. (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, Vol. 2, pp. 282–315

Cialdini R B, Brown S L, Lewis B P, Luce C, Neuberg S L 1997 Reinterpreting the empathy–altruism relationship: when one into one equals oneness. *Journal of Personality and Social Psychology* **73**: 481–94

Cialdini R B, Schaller M, Houlihan D, Arps K, Fultz J, Beaman A L 1987 Empathy-based helping: is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology* **52**: 749–58

Clary E G, Snyder M, Ridge R D, Copeland J, Haugen J, Miene P 1998 Understanding and assessing the motivations of volunteers: a functional approach. *Journal of Personality and Social Psychology* **74**: 1516–30

Davis M H 1994 *Empathy: a Social Psychological Approach*. Brown and Benchmark, Madison, WI

Dovidio J F, Penner L A in press. Helping and altruism. In: Fletcher G, Clark M (eds.) *Blackwell Handbook of Social Psychology: Interpersonal Processes*. Blackwell, Oxford, UK

Eagly A H, Crowley M 1986 Gender and helping behavior: a meta-analytic view of the social psychological literature. *Psychological Bulletin* **100**: 283–308

Eisenberg N, Fabes R A 1990 Empathy: conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion* **14**: 131–49

Eisenberg N, Fabes R 1998 Prosocial development. In: Damon W (ed.) *Handbook of Child Psychology*, 5th edn., J Wiley, New York, Vol. 3, pp. 701–98

Graziano W G, Eisenberg N 1997 Agreeableness: a dimension of personality. In: Hogan R, Johnson J A, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 795–825

Latané B, Darley J M 1970 *The Unresponsive Bystander: Why Doesn't He Help?* Appleton-Century-Crofts, New York

Penner L A, Finkelstein M A 1998 Dispositional and structural determinants of volunteerism. *Journal of Personality and Social Psychology* **74**: 525–37

Piliavin J A, Charng H W 1990 Altruism: a review of recent theory and research. *Annual Review of Sociology* **16**: 27–65

Piliavin J A, Dovidio J F, Gaertner S L, Clark III R D 1981 *Emergency Intervention*. Academic Press, New York

Schroeder D A, Penner L A, Dovidio J F, Piliavin J A 1995 *The Psychology of Helping and Altruism: Problems and Puzzles*. McGraw-Hill, New York

J. Dovidio

Adverbial Clauses

Adverbial clauses are known from traditional grammar, and basically all contemporary models of grammar, as one of three major classes of subordinate

clauses (the other two being relative and complement clauses). Their grammatical function is that of an adverbial, i.e., they provide information on the (temporal, locative, causal, conditional, etc.) circumstances depicted in the main clause. Correspondingly, the adverbial clauses in example (1) below are called temporal, locative, causal, and conditional clauses, respectively:

(1) They will meet …
  (a) before the sun rises.
  (b) where they first made love to each other.
  (c) because they need to find a solution.
  (d) if we let them.

Given the large spectrum of possible circumstances, adverbial clauses represent the most diverse semantically and (from the point of view of their interpretation) most challenging class of subordinate clauses. Given their subject–predicate structure, adverbial clauses are formally the most complex type of adverbials compared with adverbs (e.g., *soon*, *here*, *quickly*) and adverb phrases (e.g., *on Sunday*, *in the garden*, *very quickly*). Combined with a sentence frame like 'They will meet …' in (1), the latter two types of adverbial still yield a simple(x) sentence, whereas adverbial clauses yield a complex sentence. Beyond the complex sentence of which they form a part, adverbial clauses have a crucial function in the creation of a coherent discourse and are thus a prominent feature, especially of written texts. It seems that adverbial clauses can be found in all languages of the world (Thompson and Longacre 1985), even though in many languages they may look different from the prototypical adverbial clauses (with a finite verb and introduced by a subordinating conjuction such as *if* or *because*) we know from the major Indo-European languages. In the light of recent crosslinguistic research, adverbial clauses will be discussed in this article with regard to their structure (Sect. 1), the range and levels of their meanings (Sect. 2), structural properties influencing their interpretation (Sect. 3), and their functions in written and spoken discourse (Sect. 4).

1. *The Structure of Adverbial Clauses*

1.1 *Adverbial Clauses as Dependent Clauses*

Adverbial clauses are subordinate clauses in the sense that they depend for their occurrence on another, the main, clause. However, not all languages mark the distinction formally between dependent and independent clauses in the same way (e.g., German uses verb-final word order in dependent clauses and verb-second word order in independent clauses; other languages have a dependent mood: in Italian the subjunctive is used exclusively in dependent clauses). Nor do all languages make a formal distinction between these two types of clause in the first place (e.g., isolating languages Such as Chinese). In the languages of Europe nonfiniteness, i.e., the use of infinitives, participles, or related forms as predicates, is a clear indication of dependency and subordination. The same applies to the presence of certain types of lexemes serving as clause-linkers (typically) introducing, at least in languages where the verb precedes the object, a finite or nonfinite clause: relativizers such as *who*, *whose*, *whom*, *which*, *that*; complementizers such as *that*, *whether*, *if*; or adverbial conjunctions (alternatively known as adverbial subordinators) such as *where*, *when*, *after*, *before*, *because*, *if*, *although*.

Yet the presence of one of these clause-linkers by itself is no guarantee that the relevant subordinate clause can be classified as either a relative, a complement, or an adverbial clause. Just take subordinators such as *that* or *if* in examples (2) and (3), respectively, where their use as adverbial subordinators in the (a) sentences contrasts with their use as complementizers (*that* in (2b), *if* in (3b)), and the use of *that* as relativizer in (2c):

(2) (a) He talked so fast that most people couldn't follow.
  (b) He said that most people couldn't follow.
  (c) The talk that most people couldn't follow was given by a colleague of mine.

(3) (a) I'm more than glad if she's at home.
  (b) I wonder if she's at home.

What these examples show is that even in individual languages there may exist no inherent structural differences between the three major types of subordinate clause. Rather, it is the function they serve in the sentence of which they form a part that determines their classification: are they an integral part of the sentence typically serving as the argument of a verb (complement clause) or qualifying a noun (relative clause), or do they belong to the (optional) periphery of the sentence, modifying an entire state of affairs (adverbial clause)? Even then a clearcut classification may be impossible or depend on one's point of view, as the examples in (4) show:

(4) (a) She'll leave when John comes.
  (b) I forgot the bag where we met last time.

The subordinate clauses in example (4) are sometimes called *adverbial relative clauses* (with *when* and *where* analyzed as relative adverbs rather than adverbial subordinators) since they 'can be paraphrased with a relative clause with a generic and semantically relatively empty head noun' (Thompson and Longacre 1985, p. 179), such as *at the time* or *the moment* in (4a) or *at the place* in (4b). As a matter of fact, it is not difficult to find languages where in particular adverbial clauses of Time, Place, and Manner resemble and share properties with relative clauses and where, independently or in addition, the relevant adverbial subordinators 'are identical with or at least incorporate a largely desemanticized noun "place," "time," or "way, manner"' (Kortmann 1997, p. 65; on the evolution of adverbial subordinators see Kortmann 1997, 1998).

1.2 Nonfinite Adverbial Clauses

Due to their large inventories of adverbial subordinators and the pervasiveness of adverbial subordinators in adverbial clauses (obligatory in finite, optional in the much less frequent nonfinite adverbial clauses), the (mostly Indo-European) languages of Western and Central Europe, in particular, are called *conjunctional languages*. By contrast, most of the genetically rather diverse, largely non-Indo-European languages of the Eastern periphery of Europe (from Siberia down to the Caucasus) have relatively few adverbial subordinators, using rather so-called converbs, i.e., nonfinite verb forms whose main function is to mark adverbial subordination, and are thus called *converb languages* (see König 1995, Nedjalkov 1998). The division between conjunctional and converb languages and, more generally, between languages preferring either finite or nonfinite subordination strategies, correlates strongly with the basic word order found in the great majority of these two language types, namely SVO (conjunctional languages) and SOV (converb languages). Even conjunctional languages, though, make use of nonfinite (examples 5a, b) or even verbless (example 5c) adverbial clauses, albeit to differing degrees. English, for instance, stands out in this respect among the Germanic languages (see Kortmann 1991).

(5) (a) My head bursting with stories and schemes, I stumbled in next door.
(b) Inflating her lungs, Fiona screamed.
(c) Alone in his room, she switched on the light.

Interesting for the present purposes are two of their properties: even though they may have their own overt subject (so-called *absolute constructions* or *absolutes*, as in example (5a)), typically they have a subject that needs to be inferred from the main clause and is identical in reference to the subject of the main clause (known, for example, as *free adjuncts* in English, as in examples (5b, c)). More importantly, these clauses typically do not specify (for instance, by an adverbial subordinator) in which way (temporally, causally, conditionally, etc.) they modify the state of affairs expressed in the main clause, thus presenting the interpreting individual with a much more challenging problem than do most prototypical, i.e., finite, adverbial clauses (more on this in Sect. 3).

2. Semantic Types of Adverbial Clause

2.1 Circumstantial Relations

Traditionally, adverbial clauses are classified and grouped on the basis of the semantic relations that can hold between states of affairs (or propositions) depicted in different parts of a complex sentence or different chunks of discourse. The exact number and labeling of these semantic relations, variously called *adverbial*, *circumstantial*, *interclausal*, or *coherence relations*, is irrelevant. It can safely be assumed, however (and has been shown for the European languages in Kortmann 1997, that all languages use adverbial clauses and have adverbial subordinators for the expression of at least a subset of the relations in example (6) below, and perhaps for additional relations (e.g., French *faute que* expresses Negative Cause 'because no(t),' German *ohne dass* can signal both Negative Result 'as a result/consequence of which no(t)' and Negative Concession 'although no(t)'). The grouping of semantic relations suggested in example (6) corresponds largely with standard practice in many descriptive grammars when distinguishing between three major groups of adverbial clauses, each of which expresses relations which are closely related to each other: temporal clauses, modal clauses, and a third group expressing 'logical' relations, variously called causal or conditional clauses (shown here as CCC). For further discussion compare Kortmann (1997, pp. 79–89):

(6) (a) TIME — Simultaneity Overlap 'when,' Simultaneity Duration 'while,' Simultaneity Co-Extensiveness 'as long as,' Anteriority 'after,' Immediate Anteriority 'as soon as,' Terminus *a quo* 'since,' Posteriority 'before,' Terminus *ad quem* 'until,' Contingency 'whenever'
(b) MODAL — Manner 'as, how,' Similarity 'as, like,' Comment/Accord 'as,' Comparison 'as if,' Instrument/Means 'by,' Proportion 'the … the'
(c) CCC — Cause/Reason 'because,' Condition 'if,' Negative Condition 'unless,' Concessive Condition 'even if,' Concession 'although,' Contract 'whereas,' Result 'so that,' Purpose 'in order that,' Negative Purpose 'lest,' Degree/Extent 'insofar as,' Exception/Restriction 'except/only that'
(d) OTHER — Place 'where,' Substitution 'instead of,' Preference 'rather than,' Concomitance, Negative Concomitance 'without,' Addition 'in addition to'

One would assume that not all of these circumstantial relations are equally central to human reasoning. And indeed there is evidence suggesting a core of roughly a dozen cognitively most central circumstantial relations, including, above all, Simultaneity (Overlap, Duration) ('when,' 'while'), Place ('where'), Similarity ('as'), Cause, Condition, and Concession. It is for the latter three relations, for example, that all the conjunctional languages of Europe possess at least one adverbial subordinator;

similarly, it is these three relations for which the largest number of adverbial subordinators can be found in the European languages, i.e., for which the greatest need for explicit marking seems to be felt. Moreover, the adverbial subordinators marking the core relations tend to be more reduced morphologically (i.e., most lexicalized), much more frequently used, and older than those marking any of the peripheral relations (see Kortmann 1997, pp. 128–52).

Simultaneity and Cause also figure prominently in a large-scale semantic analysis of nonfinite adverbial

hardly distinguishable circumstantial relations which an equally high number of free adjuncts and absolutes can be taken to express, namely Addition/Concomitance ('and at the same time'), as in example (7), and Exemplification/Specification ('e.g., i.e., in that, more exactly,' etc.), as in example (8).

(7) (a) There he sat, wearing a white golfing cap.
(b) Sam threw himself to the ground, dragging Frodo with him.

(8) (a) Shares in Midland were worst hit, falling at one time 42p.
(b) He paid the closest attention to everything Lenny said, nodding, congratulating, making all the right expressions for him.

This shows that not all circumstantial relations are equally important in different structural types of adverbial clauses. Likewise, their relative importance as coherence relations depends on the type of discourse. For instance, Cause, Condition, and Concession play a much more important role in academic writing than they do in narrative fiction, where temporal relations as well as, for nonfinite adverbial clauses, Addition/Concomitance and Exemplification/Specification account for a much higher number of adverbial clauses (see Kortmann 1991 for statistics and a discussion of relevant literature).

2.2 *Content, Epistemic, and Speech-act Adverbial Clauses*

Adverbial clauses can be used and interpreted in different semantic domains or, alternatively, on different levels of discourse. A widely acknowledged distinction is the one suggested by Sweetser (1990), who distinguishes between three such domains: a content domain (9a), an epistemic domain (9b), and speech-act domain (9c).

(9) (a) John came back because he loved her.
(b) John loved her, because he came back.
(c) What are you doing tonight, because there's a good movie on.

In the content domain (9a), the adverbial (here causal) clause establishes a link between two objectively given, independently assertable facts; in the epistemic domain (9b) the link holds between a fact and some belief or assumption (i.e., 'John loved her' is no more than the speaker's conclusion given the fact that 'he came back'); in the speech-act domain (9c) the adverbial clause provides a motivation or justification for performing the speech act in the main clause (here asking a question). This three-level approach has been refined, modified, and complemented by other domains and notions in most recent studies (e.g., in various contributions to Couper-Kuhlen and Kortmann 2000). For example, a fourth, textual, domain is postulated, where the adverbial clause does not modify the state of affairs in the main clause, but a whole preceding text unit. For a concessive clause this is illustrated in example (10):

(10) My favourite poster is, I think, a French one for Nesquik, which shows a sophisticated-looking small boy leaning nonchalantly against something and saying that thanks to Nesquik he went back to milk. He really looks like a nice child. *Though*, there are some Adchildren that one would feel quite ashamed to have around the house …
(Greenbaum 1969 in Crevels 2000)

An important independent justification for the distinction between these domains is that there exist correlations between them and (a) the meaning and use of adverbial subordinators, and (b) the form and position of adverbial clauses. For instance, *since* has a temporal meaning only in the content domain; in the other domains it is always causal, as in 'Since you're German, how do you prepare Strudel?' Comma intonation and the causal clause following the main clause is essential for the epistemic causal in (9b); just by itself, i.e., without additional lexical or prosodic modification, a sentence like 'Because he came back John loved her' is extremely awkward, if not unacceptable. For crosslinguistic evidence of systematic correlations between the structure of adverbial clauses and the semantic domains in which they can be used, compare Hengeveld (1998) and Crevels (2000).

3. *The Interpretation of Adverbial Clauses*

There is something more than simply world knowledge or contextually grounded knowledge that is crucial for the interpretation of adverbial clauses. Formal (i.e., morphological, syntactic, and prosodic) features, too, enter into and may crucially influence the process of interpretation.

Problems of interpretation not only arise for the inherently vague nonfinite adverbial clauses (see Sect. 1.2); they also arise for finite adverbial clauses with polysemous adverbial subordinators: after all, polysemy can be observed for almost one third of the adverbial subordinators in the European languages, especially for those with a high text frequency (Kortmann 1997). Apart from that, as was shown in Sect. 2.2, formal characteristics may determine on

which discourse level(s) an adverbial subordinator may be interpreted to operate in a given adverbial clause. Here are some of the most important relevant features that may influence the interpretation (for further discussion see Kortmann 1991 and König 1995), always provided that the given language allows for a choice (i.e., has no constraints such as, for example, the obligatory use of the subjunctive mood in subordinate clauses, or subordinate clauses generally preceding their main clause):

(a) choice of tense and/or mood in the adverbial clause and, accordingly, in the main clause, as in the three semantic types of conditional (factual/real, hypothetical, counterfactual) in many languages: present tense (indicative mood) in factual/real conditional clauses (11a), past tense (or in other languages subjunctive mood) in hypothetical conditional clauses (11b), past perfect (or a conditional perfect) in counterfactual clauses (11c):

(11) (a) If she comes home, I will be very happy.
(b) If she came home, I would be very happy.
(c) If she had come home, I would have been very happy.

Many adverbial subordinators can express either Result ('so that') or Purpose ('in order to/that'). Which reading they receive in many (e.g., Romance) languages depends on the mood in the adverbial clause: indicative mood leads to a Result reading, subjunctive mood to a Purpose reading. Tense constraints hold for the two different meanings (temporal and causal) of English *since*: only when used with some past tense can *since* receive a temporal reading (but see also example (14) below).

(b) non-subordinate (in example (12) verb-second) word order is typical in spoken German for *weil*-causals and *obwohl*-concessives in the speech-act and textual domain, but can be found increasingly in the other domains, too:

(12) *Ich hab das mal in meinem ersten Buch aufgeschrieben. Weil dann glauben's die Leute ja.* ('I've written that down in my first book. Because people believe it then.')

(c) intonation, for example, the presence or absence of an intonation break (also relevant in examples (9) and (12) above): only when reading the complex sentence in example (13) as a single intonation group does the adverbial clause receive a concessive conditional reading ('even if'):

(13) /I wouldn't marry you if you were the last man on earth./
(Haiman 1986 in Kortmann 1997, p. 92)

(d) the choice of dependent vs. independent verb forms was shown to be relevant in point (a) already (indicative vs. subjunctive mood); example (14) illustrates the impact of the choice between finite and nonfinite form in the adverbial clause on the interpretation of an adverbial subordinator, here *since*. When introducing a free adjunct, *since* can only receive a temporal reading; a causal one is impossible:

(14) Since working with the new company, Frank hasn't called on us even once.

(e) constituent order, more exactly the relative order of adverbial and main clauses: in English, for example, the great majority of present-participial free adjuncts receiving a purely sequential interpretation relative to the main clause, i.e., either Anteriority ('after') or, much more rarely, Posteriority ('and then'), exhibit an iconic constituent order (see Kortmann 1991, pp. 142–57). The order of events is signaled crucially, or even solely, signaled by the relative order of adverbial and main clauses, as illustrated by the minimal pair in example (15):

(15) (a) She uncurled her legs, reaching for her shoes.
(b) Reaching for her shoes, she uncurled her legs.

The importance of iconic word order has been acknowledged for the interpretation of converbs in other languages as well (see König 1995, p. 75).

4. *Functions of Adverbial Clauses in Discourse*

Typically adverbial clauses provide background information—it is in the main clause that foreground information is given, and the plot or storyline is advanced. But they serve additional functions, including face-to-face interaction, as has been shown in a number of studies on the tasks that adverbial clauses fulfil in the organization of larger stretches of written and spoken discourse (see, for example, Thompson 1985, Ford 1993, and various contributions to Couper-Kuhlen and Kortmann 2000).

A first set of functions is concerned with the creation of a coherent discourse. Depending on whether they precede or follow their main clause, adverbial clauses create more global (or textual) coherence, or more local coherence, respectively. There is a tendency that preposed (or initial) adverbial clauses use, elaborate, and put in perspective against what follows information given in the (not necessarily immediately) preceding discourse. They serve a kind of guidepost or scene-setting function for the reader or listener by (a) filling in what has gone before, and (b) preparing the background for what is going to follow in the complex sentence, and often even a whole chunk of discourse. By contrast, postposed (or final) adverbial clauses typically have a much more local function; i.e., their scope is restricted to their immediately preceding main clause. They neither reach back into earlier parts of the discourse, nor foreshadow or prepare for what is going to follow. For instance, the subject of a postposed adverbial clause typically is identical with the main clause subject, whereas the subject of a preposed adverbial clause is often identical with that of (one of) the preceding sentence(s).

In addition to these discourse-organizing functions, adverbial clauses have been found to serve interactional functions in face-to-face conversation (Ford 1993). Thus initial adverbial clauses are often found at the beginning of relatively large speech units exactly when the speaker has maximum control of the floor. Final adverbial clauses also seem to serve a special conversational purpose, more exactly those final clauses which are separated from the main clause by an intonation break. They tend to be used preferably at those points in informal conversation where the interactants negotiate agreement. More specific interactional tasks can be identified for individual semantic types of adverbial clause.

See also: Generative Grammar; Grammar: Functional Approaches; Grammatical Relations; Semantics

Bibliography

Couper-Kuhlen E, Kortmann B (eds.) 2000 *Cause—Condition—Concession—contrast: Cognitive and Discourse Perspectives*. De Gruyter, Berlin

Crevels M 2000 Concessives on different semantic levels: A typological perspective. In: Couper-Kuhlen E, Kortmann B (eds.) *Cause—Condition—Concession—Contract: Cognitive and Discourse Perspectives*. de Gruyter, Berlin, pp. 313–39

Ford C E 1993 *Grammar in Interaction: Adverbial Clauses in American English Conversations*. Cambridge University Press, Cambridge, UK

Hengeveld K 1998 Adverbial clauses in the languages of Europe. In: van der Auwera J (ed.) *Adverbial Constructions in the Languages of Europe*. de Gruyter, Berlin, pp. 335–420

König E 1995 The meaning of converb constructions. In: Haspelmath M, König E (eds.) *Converbs in Cross-linguistic Perspective*. de Gruyter, Berlin

Kortmann B 1991 *Free Adjuncts and Absolutes in English: Problems of Control and Interpretation*. Routledge, London

Kortmann B 1997 *Adverbial Subordination: A Typology and History of Adverbial Subordinators Based on European Languages* (Empirical Approaches to Language Typology 18). de Gruyter, Berlin

Kortmann B 1998 The evolution of adverbial subordinators in Europe. In: Schmid M S, Austin J R, Stein D (eds.) *Historical Linguistics 1997, Selected Papers from the 13th International Conference on Historical Linguistics, Düsseldorf, 10–17 August 1997*. Benjamins, Amsterdam, pp. 213–428

Nedjalkov I 1998 Converbs in the languages of Europe. In: van der Auwera J (ed.) *Adverbial Constructions in the Languages of Europe*. de Gruyter, Berlin, pp. 421–56

Sweetser E E 1990 *From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press, Cambridge, UK

Thompson S A 1985 Grammar and written discourse: Initial vs. final purpose clauses in English. *Text* **5**: 55–84

Thompson S A, Longacre R E 1985 Adverbial clauses. In: Shopen T (ed.) *Language Typology and Syntactic Description*. 3 vols. Cambridge University Press, Cambridge, Vol. II, pp. 171–234

van der Auwera J (ed.) 1998 *Adverbial Constructions in the Languages of Europe*. de Gruyter, Berlin

B. Kortmann

Advertising Agencies

An advertising agency is an independent service company, composed of business, marketing, and creative people, who develop, prepare, and place advertising in advertising media for their clients, the advertisers, who are in search of customers for their goods and services. Agencies thus mediate between three different but interlocking social groups: industry, media, and consumers. The history of advertising is largely the history of the advertising agencies that have served the needs of these three groups (see *Advertising: General*). They link industry and media by creating new forms for messages about products and services; industry and consumers by developing comprehensive communications campaigns and providing information thereon; and media and consumers by conducting audience research to enable market segmentation.

1. Origins and Early Developments

Advertising agencies are the most significant organizations in the development of advertising and marketing worldwide. They came into existence in the United States in the mid-nineteenth century and, later, elsewhere because of the mutual ignorance of the needs of newspaper publishers and would-be advertisers and because of the opportunity for profit provided by both parties' desire for economic assistance. Initially, advertising agents facilitated the purchase and sale of space. In so doing they promoted the general use of advertising, helped advertisers find cheaper and more effective ways of marketing goods, and served to inform the public of the existence of those goods. They acted as a crucial bridge between the activities of selling products and of mass communication at a time when both were undeveloped and expanding rapidly (see *Journalism*; *Mass Media, Political Economy of*).

The first advertising agents were colonial postmasters in America who accepted advertisements for inclusion in newspapers. In 1843, Volney B. Palmer set up the first independent agency, soliciting orders for advertising, forwarding the copy, and collecting payment on behalf of newspapers that had difficulty in getting out-of-town advertising. This *newspaper agency* was the first of four stages through which the business of the advertising agent proceeded to pass. In the second *space-jobbing* stage, the agent became an independent middleman who sold space to advertisers and then bought space from newspapers to fill his orders, driving a hard bargain with both. In 1865, George P. Rowell initiated the third *space-wholesaling* phase when, anticipating the needs of advertisers, he bought wholesale from publishers large blocks of space that he then resold in smaller lots at retail rates. Finally, in 1867, some agents contracted annually with

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

the publications they represented to buy in advance the entire advertising space of particular newspapers, thereby acquiring the *advertising concession* in a publication. Similar processes of development can be found outside the United States, in countries such as England, France, and Japan, where advertising agencies also owe their early existence to the buying and selling of newspaper space.

Although the early agency business was often a confusing amalgam of these different arrangements, one element was common to them all and has continued to some extent to the present day. Advertising agents found themselves in an equivocal position as they attempted to serve two masters: on the one hand, the *publishers*, on whose behalf they sold advertising space and, on the other, the *advertisers*, to whom they acted as expert and 'impartial' advisors in the placement of their ads. In other words, right from the start, the advertising agency business has invited conflict between the diverse interests of the advertiser, the agent, and the media in which advertising is placed.

It was in order to overcome such a potential conflict of interests caused by space brokering that the idea of an 'open contract plus commission' was introduced. For the first time, in 1875, N. W. Ayer & Son entered into a formal arrangement that enabled a client to use the agency's services exclusively in exchange for payment of a fixed compensation based on the amount of advertising it billed each year. This so-called 'billing' system soon spread and is still widely used by advertising agencies. It is the gross billings posted by advertisers, rather than the actual payments received by agencies from their clients, which are today used to gauge the relative size of the world's advertising agencies and their industries.

There were two important social consequences of the introduction of a contract between agency and advertiser. First, under the space brokerage system, advertising agents were in constant danger of being squeezed between advertisers and publishers. The open contract, however, permitted the agent to represent the advertiser *as well as* the publisher. Not only this, but it established the agency's relationship to a *client* rather than to a customer which hitherto had given its business to various different agents.

Second, by entering into a contract with an advertiser, an agency's profits were tied to the amount that its clients billed. Its growth was thus linked to the economic well-being of the advertisers it represented, so that even today an agency's prosperity is totally determined by the prosperity of its clients. When the economy is buoyant, advertisers spend and agencies flourish; when there is a recession, advertising appropriations are cut back, and agencies struggle. Such close ties can extend to agencies' organizational expansion. When advertisers' business interests lead them to set up international offices abroad, their agencies often follow. This explains the establishment, growth, and commanding presence of American agencies in most advertising industries worldwide, although English agencies tend to have a strong presence in former British colonies. In Japan, too, a buoyant domestic market has led to the growth of major Japanese advertisers served by Japanese agencies that have resisted American dominance and followed their clients abroad, especially into East and Southeast Asia (see *International Advertising*).

2. *Full Service Agencies*

What is known as a full-service agency is an agency that handles the planning, creation, production, and placement of advertising for its clients. It may also handle sales promotion and other related services as needed by individual advertisers. In short, a full-service agency offers its clients a complete range of services beyond the preparation and placement of their advertising.

The development of the full-service agency goes back to the last quarter of the nineteenth century, when mounting competition forced agents to think of other services that they might offer (potential) customers besides the placement of their advertising. One of these was marketing. In 1879, N. W. Ayer & Son conducted its first market survey in a successful attempt to wrest a large advertising account from a competitor. This led to the implementation of advertising planning, as agencies gradually took over the functions of studying the actual and possible markets for the commodities sold by their customers and of devising plans to fit into the general strategy for tapping those markets (see *Market Research*).

It was at this time, too, that agencies first became involved in what is now known as 'creative' work. N. W. Ayer & Son, and Catkins and Holden were among the first agencies to help advertising become a profession rather than a mere business transaction when they began to write advertisements for their clients, instead of simply placing their finished ads in newspapers. As a result, as well as sending salesmen out to sell space to customers, some advertising agencies began to create, plan, write, and design the copy and artwork that went into every ad. They then produced the necessary plates, and placed them with the publishers.

There were two significant reasons underlying agencies', at first reluctant, move into the creative side of advertising. First, they recognized that the ineffectiveness of a poorly written advertisement (contracted or written by the advertiser) was sure to redound against their fledgling business. Second, they found themselves operating in a highly competitive environment in which increased skill was demanded in the writing of advertising in order to make it stand out effectively. Moreover, new forms of advertising (magazine and outdoor) were being introduced. This

necessity for distinction in the preparation of copy was made more urgent by technical advances in the printing and engraving trades. These made possible a more plentiful use of color and illustration, but they also complicated the work of advertising agencies.

Later technological innovations created further opportunities for agencies to develop into new areas of expertise and so come to be seen as essential partners in the advertising endeavor. Radio, for example, offered an entirely new medium of entertainment (and advertising) dependent on sound alone (see *Radio as Medium*). Within three years of its first transmission, and in spite of the immense technical difficulties involved, agencies had begun to take an active part, not only in the preparation of advertising commercials, but in the planning and production of radio programs. These included continuity drama or 'soaps' which, almost from the very first, were supported by advertising and publicity (see *Soap Opera/Telenovela*).

The advent of television, which became popular in most industrialized countries from the 1950s, also obliged agencies to sail further into uncharted territory—this time by learning how to combine sound and images in cinematic form (see *Television: General*). As television advertising became ever more popular, dominating print advertising in many parts of the world, and as agencies continued to be active in the planning and production of media outputs, they found themselves coming into more frequent contact with production companies, celebrities, model agencies, film studios, stylists, photographers, animators, and others in the entertainment world. The latter decades of the twentieth century, therefore, have seen the formation of increasingly close ties among the advertising, media, and entertainment industries (see *Celebrity*; *Entertainment*).

Besides 'above-the-line' advertising in the four main media of television, newspapers, magazines, and radio, full-service agencies have found themselves moving into 'below-the-line' activities involving all sorts of areas of contemporary life hitherto untouched by advertising. Some—like the ads found on the backs of Australian postage stamps during the reign of Queen Victoria and of British dog licenses in the 1920s and 1930s—have ceased to exist. Most others, however, remain. At this point it seems unlikely that the advertising currently found on airport baggage trolleys, grand prix racing cars, telephone cards, professional sports equipment, and the backs of tickets of all kinds will, short of legislation, fall into disuse.

To further improve the services that they already offer their clients, advertising agencies have in many respects acted as 'cultural intermediaries' between the economy and culture of contemporary societies. This means that they have begun to move away from the four main media that sustained the advertising industry throughout the whole of the twentieth century, and to invest in new business opportunities, including merchandizing, telemarketing, and e-commerce. The commercial sponsorship of sports, art, and other cultural events that became prominent in the 1980s and 1990s has also been driven by advertising agencies which now look for further opportunities to show their skills in matching their clients' needs for publicity and advertising to the everyday lives of consumers (see *Mass Media and Sports*). This development of 'promotional culture' has led to agencies participating in everything from the development of animation programs in consultation with television networks, supported by product placement and character merchandizing on the part of toy and other manufacturers, on the one hand, to the arrangement of marriages between celebrities, on the other. In many respects, agencies can best be described as the black boxes of cultural continuity and change.

3. *Agency Organization*

Many agencies in Europe and the United States have been founded by two entrepreneurs: one a creative person, the other a salesman or account manager. For example, the J. Walter Thompson (JWT) agency was founded by husband-and-wife team of Stanley Resor and Helen Resor who focused on management and copywriting, respectively. While, at first, they may have handled all aspects of their advertising business themselves, eventually they were obliged to take on employees with different areas of expertise and add divisions to handle the basic areas of responsibility in a full-service agency. Although agencies differ in the way they are organized, the main functional areas of a full-service agency are: (a) *account management*, (b) account planning or *research*, (c) *media planning* and buying, (d) *creative* services, and (e) *internal* services (dealing primarily with finance, personnel, and traffic). They are usually led by a chief executive officer (CEO) and one or two vice presidents, who may oversee a board of directors whose members represent different areas of their agency's responsibility.

The exact weight given to the work of each functional area has varied over time and differs also both by agency and by country. During the first half of the twentieth century, emphasis was placed by such famous copywriters as Claude Hopkins on a scientific approach to advertising—an approach supported by the founding of research organizations by A. C. Nielsen, George Gallup, and others. Later, in the 1960s, led by such creative advertising people as 'Marlboro man' Leo Burnett, David Ogilvy, and William Bernbach, art, inspiration, and intuition became the industry's buzzwords. The 1970–90s saw a return to hard-sell advertising in what was called the era of 'accountability.'

How do these different areas of responsibility work in practice? When an agency wins a new account, it

generally forms a cross-department account group, or account team, composed of an account supervisor (who is often supported by account executives), account planners and media buyers, as well as a creative team made up of copywriter and art director with, as necessary, a producer (for television and/or radio commercials). The account group proceeds along more or less predetermined lines. First, the account planning department works out and solves a marketing problem by conducting new, and making use of previous, research to find out who the targeted consumers (known as 'prime prospects') are; what their demographic characteristics might be; how the product to be advertised fits into their lifestyle; what they think of this and competing products; what the most suitable advertising medium for the product in question might be; and so on. On the basis of answers to these questions, account planners (who are sometimes simply referred to as 'marketers') formulate a strategy that positions the product in relation to targeted consumers and emphasizes the attribute(s) that will appeal to them.

Once the overall marketing strategy has been determined, the agency's creative team works on an appropriate creative strategy, writes copy, and prepares rough layouts and storyboards. At the same time, the media buyers need to work out a media plan that accords with the marketing strategy, to select an appropriate mix of media, and prepare schedules with costs. These three separate plans are then amalgamated into a single package by the account supervisor and presented to the client, which may or may not get involved in different stages of the agency's preparations along the way. The extent to which an advertiser 'interferes' in its agency's work depends very much on the company concerned. However, evidence suggests that, for reasons that will be explained below, agencies tend to work much more closely with their clients in Japan than they do elsewhere in the world.

One of the main organizational difficulties facing any account group is that it invariably finds itself having to deal with two different audiences whose interests may not be the same. On the one hand, account planners and the creative team focus their attention on how to advertise and sell a product to a particular targeted group of *consumers*. On the other, the account supervisor and his supporting account executives need to liaise between the *advertiser* and the agency, being responsible for interpreting the client's marketing needs to their colleagues, on the one hand, and selling the marketing, creative, and media plans to the client, on the other. Since it is the advertiser who pays the agency for its services, it may disagree with—and occasionally overrule—the consumer-oriented strategy proposed by its agency's account group.

A second organizational difficulty that often emerges in the work of an account group is internal and concerns the different attitudes towards advertising held by account planners, on the one hand, and creative people, on the other. The very *raison d'être* of account planning is to form marketing strategies based on objective, scientific criteria derived from in-depth qualitative and quantitative data. These data then have to be transformed into creative images by copywriters and art directors who usually claim to work according to intuitive, artistic ideas that may have little actual relationship to the expressed marketing aims. Sometimes conflict between the two types of advertising agency employee results and adversely affects the work of the account group which, ideally, needs to blend harmoniously to achieve its client's aims.

4. *Agency Compensation*

It was noted earlier that advertising agencies started out as brokers of space on behalf of newspaper publishers who paid them a commission for their services. Traditionally, this commission has amounted to 15 percent of the media cost. It may be more for some kinds of advertising (for example, outdoor), or less for certain advertisers whose account an agency, for one reason or another, prizes highly enough to accept a lower commission. In addition to media commissions, agencies may charge a production commission for production work subcontracted (type, photography, and illustration). Here the client is charged the cost of the work involved, plus a 17.65 percent commission.

As seen above in the actual creation of an advertising campaign, agencies have often found themselves dealing with the conflicting interests of two different partners. The very existence of the commission system has, historically, encouraged the advertiser to try to pay less than the sum demanded by the medium in which it sought to place its advertising. As a result, advertising agencies often found themselves caught between a desire for profit, on the one hand, and a need to retain its business, on the other. This has led, and in some parts of the world still leads, to agencies playing off one partner against the other, as they seek rebates from the media in which they place their clients' advertising, but pass on the full cost to the advertisers concerned.

It is to offset this problem of business ethics that fees have been introduced. This form of compensation is comparable to that used to pay lawyers or accountants. Agency and client agree an hourly fee for the agency's work, to which are added charges for out-of-pocket expenses, travel, and other standard items. No commission is paid on media costs which are billed net of commission. As a rule, advertiser–agency compensation arrangements are negotiated according to the specific requirements of the advertiser and the willingness and ability of the agency to fulfil those

needs at a mutually agreeable cost. Over time, the fee system has come to predominate over the commission system in the United States, although in many other countries the commission system, or a combination of fee and commission, is still the norm. It seems likely that in future the fee system will prevail, in which case the traditionally close reciprocal ties between advertising agencies and media organizations will probably weaken to some extent.

5. *The Account System*

The sum of money put aside by an advertiser and allocated to an agency for the purpose of selling a particular brand or product group, sometimes through a selected medium, is called an account. It is an advertising agency's job to persuade an advertiser that it is best suited to take on a particular account. To this end it will embark on active solicitation with the aim of being asked to participate in a competitive presentation or 'pitch' in which, together with other agencies, it will present marketing and creative strategies based on the advertiser's initial orientation of its needs. The successful agency will be asked to take on an account, usually for a fixed period, and will then enter into partnership with its new client.

The relationship between agency and client is ideally a professional one, involving a considerable exchange of confidential information which may include, for example, plans for new products or marketing strategies. As a result, when awarding an account, an advertiser will not usually permit its agency to handle companies or products with which it is in direct competition. This is known in Europe and the United States as the *competing account rule*.

The rule about an agency's not handling competing accounts means that, whenever it wins a new account that conflicts with one already being handled, it has to decide which of the two accounts it wishes to keep. Competing accounts are also often a deciding factor when, as has happened frequently since the 1980s, agencies decide to merge, although many such large agencies run their local offices independently. This they do in the hope that their clients will not regard the same type of account being handled in another office as being in competition and thus not raise an objection.

So far as the organization of the advertising industry is concerned, the competing account rule ensures two things: first, that there is a continuous circulation of accounts among agencies; and second, that no agency ever becomes excessively large. As a result, by comparison with other industries, the organizations that make up the advertising industry are quite small.

Although the competing account rule is the norm in almost all countries of the world, an alternative system of distributing advertising appropriations exists in Japan where advertisers split their accounts by product, media, or a combination of product and media, and do not object to the same agency handling a number of competing accounts. Thus large advertisers like Toyota or Kao, for example, contract more than half a dozen different agencies to handle the advertising of their numerous products. These same agencies may also be handling the accounts of rival car or toiletries manufacturers.

The *split account system* has advantages for both agencies and advertisers. Precisely because the overall advertising appropriation is divided, accounts in Japan are not nearly as large as they are elsewhere. This is to an agency's financial advantage when it loses an account, since the sheer number of accounts in circulation means that it can usually make up the financial shortfall and is not obliged, as it might be in the United States or Europe, to lay off staff. In this respect, the split account system contributes to the overall stability of the advertising agency, and of the industry of which it is a part.

Advertisers are able to use the split account system to play off one agency against another since they alone have full knowledge of their overall marketing strategy. This not only enables advertisers to minimize the problem of confidentiality that so worries those working in advertising industries elsewhere. They can use their knowledge strategically to operate a system of divide-and-rule that obliges agencies to keep in constant touch with client personnel, as well as with other agencies and media organizations, in order to gain as much information as possible about their clients' overall plans and activities. In other words, the split account system encourages much closer interaction among the three main players in the advertising industry.

There are disadvantages to the system. One result of the need for Japanese agency account executives to keep in constant touch with their clients is that advertising decisions tend to be based on personal likes and dislikes, rather than on more objective professional criteria. Another, claimed by agencies, is that they are prevented from producing effective advertising because their clients do not give them access to their *total* marketing strategy. All in all, however, the advantages appear to outweigh the disadvantages within the Japanese advertising industry. In this respect, it is significant that in the autumn of 1999, Procter & Gamble, one of the world's biggest spending advertisers, decided to divide its worldwide advertising appropriation into nine different brands and to distribute these accounts among nine different advertising agencies. It is possible, therefore, that in future the system of split accounts will become more popular with advertisers outside Japan, thereby altering the relationship hitherto developed between an agency and its clients.

See also: Advertising and Advertisements; Advertising, Control of; Advertising: Effects; Media Effects; Media, Uses of

Bibliography

Hower R 1939 *The History of an Advertising Agency: N. W. Ayer & Son at Work, 1869–1939*. Harvard University Press, Cambridge, MA

Jones P 1999 *The Advertising Business*. Sage, Thousand Oaks, CA

Leiss W, Klein S, Jhally S 1990 *Social Communication in Advertising: Persons, Products and Images of Well-being*, 2nd edn. Routledge, London

Lien M 1997 *Marketing and Modernity*. Berg, Oxford, UK

Miller D 1997 *Capitalism: An Ethnographic Approach*. Berg, Oxford, UK

Moeran B 1996 *A Japanese Advertising Agency: An Anthropology of Media and Markets*. University of Hawaii Press, Honolulu, HI

Schudson M 1984 *Advertising: The Uneasy Persuasion*. Basic Books, New York

Wells W, Burnett J, Moriarty S 2000 *Advertising: Principles and Practice*. Prentice-Hall, London

B. Moeran

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Advertising and Advertisements

1. Definition

The way the terms 'advertising' and 'advertisement' are defined depends to a great extent on the author's perspective, training, and agenda. Even so, an encyclopedia entry seems to beg for an authoritative definition. This article offers the following: an advertisement is a message paid for, but not delivered, by the sender, that (a) incorporates technologies or forms other than speech, (b) appears in a public forum, and (c) attempts to persuade receivers to behave in a way that brings direct economic benefit to the sender. This definition allows for most of what is customarily considered advertising, while excluding other acts or forms that, while sharing some characteristics with advertisements, are not generally considered such. For instance, this definition would not include a salesman's pitch, but would encompass any brochures he used. Political speeches, whether made in Parliament or during the breaks on *60 Minutes*, would not be 'advertisements' because the expected behavioral outcome is not directly economic. Corporate 'image' messages would count, even when purchase is not the object, as such advertisements ultimately have an economic goal, such as affecting the value of stock by increasing the prestige of the corporation.

2. Function

The function of advertising likewise will vary in description based upon the education, focus, and motive of a particular author. Economists traditionally have allowed 'advertising' into their models only as a form of market information. Their definition of 'information,' however, is usually limited to prices and availability. To the extent that advertisements incorporate other 'non-information' (such as prestige claims or sex appeals), advertising becomes an 'exogenous variable,' constitutive of 'tastes,' and therefore present in the system only as noise or interference. As long as advertisements serve to notify markets of supply and price, however, the institution of advertising operates neatly within the range of traditional economic models. Indeed, to the extent that advertising supports stable demand, allowing producers to become price-makers rather than price-takers, it becomes an important aid to microeconomic planning. The realization of economies of scale that accrue from such an arrangement, so the argument goes, tends to reduce prices to consumers in the long run.

Critics and consumer groups often charge that advertising merely adds to the prices paid by consumers—sometimes raising them as much as 100%—but such claims do not hold for most goods in the long run. For example, simple bar soap was sold 100 years ago in the US at a price that included advertising expenses equal to direct costs (thus representing about 50% of each unit's price). As consumption of soap grew, however, the advertising expenditures leveled off (advertisement budgets reach a plateau of effectiveness beyond which it makes no sense to spend at higher levels). Consequently, the percentage of each unit's price represented by advertising steeply declined. Since advertising expense is only one of the economies of scale created by mass marketing and production, the total unit price of each bar of soap also dropped dramatically, allowing many US citizens to wash regularly for the first time. Today, it is only the very extraordinary soap that would include an advertising expense more than 3 percent of its unit price. The same is true for most well-established, mass-produced, and heavily advertised products.

In the Marxist critique of capitalist consumer culture, advertising's most central function is to ensure capitalism continues: by stimulating ever-increasing levels of demand, advertising forces an always-growing array of goods into consumers' homes, regardless of need or utility. Here too, advertisements that inform audiences of availability and prices are not typically thought problematic. However, the 'fetishization' of objects by advertisements that name, sing, or tell stories is seen as a sinister form of industrial 'magic' through which consumers are made to value products for the wrong reasons. By adding this dubious 'exchange value' to the object's 'use value,' advertising serves to rob workers and pad the pockets of capitalists with unearned profits.

The cultural critics' picture of advertising as a guarantee of long-term profits is sharply at odds with the way advertising's function is actually viewed in practice. Though the most obvious reason for a

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

manufacturer to spend money on advertising is to stimulate purchase, advertisers are famously skeptical about the actual impact of media budgets on sales. They tend to feel vaguely robbed by their advertising agencies, forced to participate in a pointless display by social convention. And, indeed, one of the few generally accepted findings of the past 30 years of academic research on advertising effectiveness is that a direct effect on sales is virtually impossible to demonstrate. So many factors influence sales that it is hard to isolate advertising's function to stimulate demand. And, from case to case, advertising's apparent performance can vary from spectacular to abysmal with little readily discernible logic.

Accounting practices reflect the doubts of the manufacturing community: advertising cannot be treated as an investment in capital (like buying a new plant), but must be treated instead as an annual expense (like office paper). Only in the past 5 to 10 years (and only in Europe) has the possibility of recognizing the long-term benefits of advertising in financial statements been discussed. The reason for reconsidering advertising as an investment is the equity markets' valuation of companies with established brands. Consider that shares in Procter and Gamble, the largest advertiser in the US, go for more than four times book value. The price of the stock reflects the financial community's belief that some value inheres in Procter and Gamble beyond the historical cost of its equipment, patents, and real estate. Probably, the financial markets are imputing a value for Procter and Gamble's many brands, most of which are long-standing leaders in their respective product categories. Category leadership is at least partly attributable to Procter and Gamble's historical practice of consistently advertising each brand at levels that exceed all direct competitors' expenditures.

Historians have offered the view that advertising's function is to weaken, even eliminate, the 'middlemen' in goods distribution by stimulating consumer demand for products. Once consumers want a particular brand, they say, retailers are 'forced' to carry it on their shelves and brokers are 'forced' to provide it to retailers. While it is demonstrably true that the growth of advertising as an institution over, say, the past 150 years has resulted in a 'rationalizing' of distribution systems in a way that is advantageous for manufacturers (and, actually, consumers), the notion that advertising allows makers to control distributors should not be overstated. Distributors and retailers retain a great deal of power in the current system and, in fact, during the last 30 years of the twentieth century, the shift has been back toward the distributor. We see evidence of this in the production of private label brands, the growing power of stores like Wal-Mart, and the broadscale shift of advertising dollars spent by companies like Procter and Gamble into expenditures for trade allowances and the like. Further, while it is certainly plausible that advertising has had the effect of increased control over distribution, arguing that channel management is actually the primary function of advertising conflates 'effect' with 'function.'

The conflation of the terms 'effect' and 'function' is also a problem in works that purport to identify the impact of advertising upon social issues. For example, many feminists have written about the effects of advertising on the status of women, frequently falling into the habit of claiming that advertising's function is to oppress women. While advertisements may indeed contribute to the ideology of gender inequality, arguing that advertising's purpose is the oppression of women seems a misplacement of emphasis.

From the broadest cultural perspective, advertising's function may be more meta-structural than most social critics would describe—or advertisers pay for. Anthropologists argue that the primary function of consumption is cultural, not biological, because goods serve to make the categories of culture visible and, therefore, stable. Since human cognition demands a balance between total heterogeneity and total homogeneity, any system of objects will require an activity that marks items as being appropriate for certain users, places, or occasions. In pre-industrial societies, marking is accomplished through exchange rituals performed in the presence of the entire tribe. The chanting, singing, and naming in such rituals can be uncannily similar in form to modern advertising. Given the proliferation of otherwise homogeneous goods in the mass-production society, one would expect to see some mechanism arise for marking these goods as cultural artifacts. Since mass production was accompanied in history by the emergence of mass communication, it is, perhaps, not terribly surprising that the activity of marking mass-produced goods would occur in the public 'spaces' provided by the mass media. Thus emerges advertising 'as we know it.'

3. *Form*

Today, advertising is highly identified with images. Particularly in cultural studies, invoking the dangers of the visual is almost synonymous with advertising criticism. But advertising is not necessarily, nor even distinctively, visual in form. Television commercials that use only the visual dimension are as rare as silence on a sitcom—that is to say, rare indeed. Instead, television advertisements, like most other types of advertising, skillfully integrate more than one symbolic system (speech, writing, numbers, music, gesture, and so on). In radio, where the visual dimension is absent, the sound of speech is often elaborated with rhyme and rhythm or augmented with song or 'sound effects.' Furthermore, the largest category of advertisements, directory and classified advertisements, are usually composed entirely of written text. Even here, however, we see nonverbal dimensions brought to bear. The addition of color, use of stylized typefaces,

Figure 1
Visual use of speech sound similarity

Figure 2
Adaptation of familiar image

and strategic placement of written messages in all forms of print emphasize something cultural critics try to ignore: that writing is itself an image; the printed word is, by definition, speech made visible.

Writing systems, furthermore, are not always alphabetic in nature, but are sometimes based on picturing. The first shop signs operated on a kind of pictography, even in alphabetic societies, because only the elites were literate. Thus, the convention of hanging the shape of a shoe outside a cobbler's door, or painting a striped pole to designate a barber, or differentiating taverns by the visual representation of their names (the 'Bull in Boots'). Critics who insist on situating advertising in the industrial period argue that such signs were not a form of advertising because they were merely 'information,' an 'announcement' rather than an attempt to persuade. The range of early signs includes many aggressive, even poetic, pictorial statements, however. In England, the signs became so huge and outlandish that they began to cause accidents by falling or blocking the view—in 1762, they were outlawed.

Today, imagistic advertisements still reach into common linguistic forms—proverbs, colloquialisms, jargon—for the sense of their message. In the advertisement in Fig. 1, for instance, the association between 'duck' and 'duct tape' has no basis in use: one cannot buy ducks from amazon.com, nor can ducks be customers of amazon.com, nor, for that matter, would duct tape be used on a duck. The association here only 'works' because, in speech, the phrase 'duct tape' sounds like 'duck tape.'

The multiform nature of advertising is at least partly a result of the ongoing struggle to garner consumers' attention. In the early years of advertising's growth in the USA, most publishers limited advertisements to ten-point agate type. This policy, known as the 'agate-only rule,' was an attempt to equalize commercial competition and a way of saving paper. Inventive advertisers strained against the agate-only rule by printing their messages in patterns, sometimes getting notice through sheer repetition and other times by using the text to form a picture. When, eventually, publishers allowed advertisers to make use of images, the pressure only escalated. Now, advertisers wanted bigger pictures, more distinctive pictures and, especially, color pictures. The same struggle occurred in aural form during the early days of radio. The competition to stand out, be memorable, or gain credibility resulted in testimonials by leading radio stars, jokes woven into the existing program, jingles

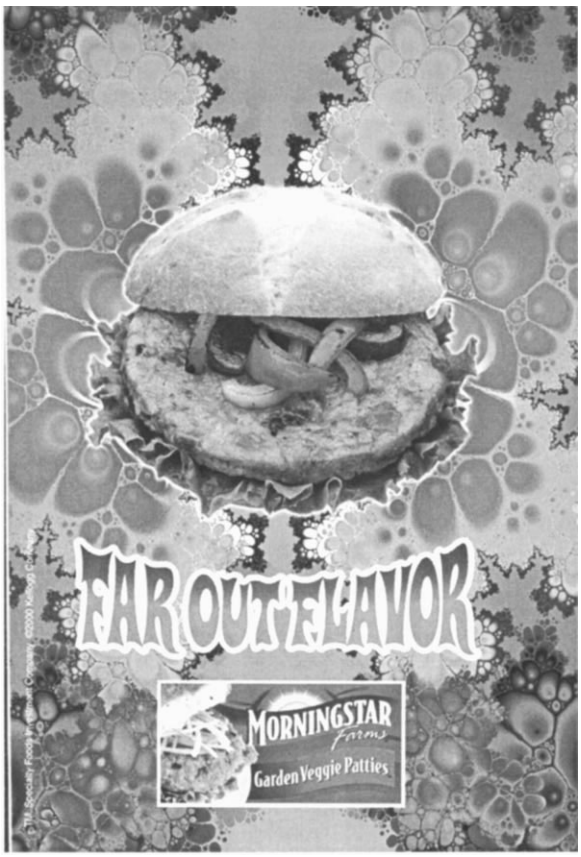

Figure 3
Use of visual style to periodize

put in with the music. The Internet advanced in less than 5 years from a text-only medium to a jumbled screen offering pictures, words, numbers, sound, animation, and even video. While the proliferation of formal features on the Internet was certainly made possible by technological innovation, it was also largely fueled by the demands of the new commercial media, the 'dot.coms,' wanting to coax their reluctant customers, national advertisers.

Advertisements have also adopted the formal features of other genres. Early in the century, for instance, the advice columns of the new national magazines in the USA were quickly imitated by advertisers, who used their own 'experts' to give advice that inevitably involved buying a particular brand. Today, advertisements like the one in Fig. 2 reach all the way to fifteenth-century Italian art for their inspiration, leading cultural critics to bemoan the 'pastiche' that seems to crop up so often in advertisements. However, advertisements like this one (and others that adapt genres like movies or comics) also testify to the growth of an elaborate visual vocabulary among the reading populace.

As this visual vocabulary builds, even the images of the past become meaningful to the readers of the present. Thus, we see frequent adaptations of older styles reused to specify a past period. For instance, the typeface, background and headline in Fig. 3 all communicate 'the late 1960s.' Yet this advertisement ran on the back of *TV Guide* in the autumn of the year 2000: many of those readers would not be old enough to remember the 1950s or 1960s, but nevertheless are clearly expected (by the advertisers) to know the style associated with that time. Achieving specific reference through subtle variation in the manner of rendering is a hallmark of advanced pictographies. Advertisements like these demand a highly developed form of reading from their viewers, as opposed to the kind of mindless absorption that most critics attribute to consumers of the 'image culture' of advertising.

Though references to the visual past appear frequently in advertising today, it is also true that in our own era—as in others before us—most advertisements struggle to look 'modern' (and here I mean 'hip,' 'new,' 'cutting edge,' etc.) The same is true in other dimensions, such as music: the first Pepsi jingle was in the swing style popular on radio in the 1930s; in the 1960s, Pepsi used Motown artists; in the 1980s, Michael Jackson and Madonna sang for Pepsi. As a result of this propensity to operate right at the horizon of a culture's formal style, advertisements are consistently among the most topical, timely, and stylish of cultural texts. Thus, the advertisements of any period tend to be, if nothing else, distinctly 'of their time,' heavily 'periodized.' As Roland Marchand noted, for instance, the advertisements of the 1930s are markedly different from those of the 1920s—more aggressive, less graceful, more pronounced in commercial intent. He attributed the difference to the declining economy. So, not only do the latest visual styles, celebrities, tunes, and dances feature heavily in advertisements, the issues and concerns of the day do as well. This influence may occur in pervasive ways, such as Marchand described, but it also makes itself felt in very specific, identifiable instances. War imagery and patriotic motifs, for instance, are much more common in US advertising between January 1942 and August 1945 than they are in the months either immediately before or immediately after. Fear of disease is the basis for many advertising appeals following the influenza epidemic of 1918; today's advertisements refer to AIDS and breast cancer. Ideology is also reflected: Thomas Frank's *Conquest of Cool* (1997) has eloquently argued that the change in advertising known as the 'Creative Revolution' had its roots in the broader political shift that rumbled through the USA between 1960 and 1970.

If one were to try to theorize an aesthetics of advertising, therefore, a key characteristic might be this tendency to focus on the present, even while evoking the past. Another, which has been noted by authors like Jennifer Wicke, is that advertising is

essentially an aesthetics of economy: saying a lot in a small space or a short time is nearly always one of the objectives. Roland Barthes' famous observation that advertising is generally characterizable by its 'frankness'—in the sense that advertisements try to seem natural—is not, in my opinion, supportable. Instead, advertising's need to 'defamiliarize' its propositions has created a corpus of texts notable for whimsy, fancy, outrageousness, and blatant artifice. A good example is a television commercial running now, which refers to the 'animal nature' of a Jeep by having the car shake mud and water from itself like a wet dog, showering its owners in the messy residue. The fundamental impulse to animate an object in the context of exchange—something that is present even in so-called 'primitive' cultures—here runs directly into the expressive capacity of early twenty-first century technology. A simple metaphorical proposition ('a Jeep is like a dog') is transformed, via computer animation, into an extraordinary fiction. The expression of a very basic material proposition in highly imaginative form would also probably be one of the key principles of an aesthetics of advertising.

For all its fictive properties, however, advertising is unlike other art forms in one very mundane way. Advertisements always have 'one foot on the ground' in the sense that they attempt to influence a concrete behavior, usually a purchase. Depending on the particular government under which it is operating, an advertisement may also have to accommodate practical concerns about the product or the status of exchange in some formal way. Commercials for drugs that stimulate weight loss, for instance, often represent attractive possibilities—followed abruptly by disclosure of some rather unattractive side-effects. Many other products, from toys to cigarettes, must include warnings or disclaimers about everything from health risks to whether batteries are included. Regulations on deceptive advertising, at least in the USA, will tolerate a range of fiction and puffery, but the inferences the average consumer makes about the product must be concretely true or the campaign is pulled. What that means, then, is that most US viewers probably do not expect a Jeep to shake like a dog, even after seeing it on TV. In fact, looking at the total range of fancy discernible in advertisements today, viewers must be making, rapidly and on a daily basis, some fairly fine distinctions between the fiction of the advertisements and the facts about the products. This, again, suggests a more skeptical, sophisticated, selective viewer than many who write about the 'manipulation' of the populace by advertisers prefer to imagine.

Raising the specter of an 'aesthetics of advertising' among those who actually make advertisements often revives a long-standing philosophical debate over what 'a good advertisement' is. On one side are those who believe that 'a good advertisement' is one that sells, even if it irritates every member of the audience. Famous admen like Rosser Reeves took this position in the 1950s. Others like Bill Bernbach, who led the Creative Revolution of the 1960s, have argued that 'a good advertisement' is one that is more 'creative' (that is, prettier, funnier, more imaginative, or in some other way a more pleasing work of art), but have also claimed that aesthetically pleasing advertisements actually sell more product anyway. Such positions are hard either to evaluate or defend because of the previously-mentioned difficulty of demonstrating the effect of advertising on sales. These positions are made even more problematic by the changes over time in (a) what is deemed to be a pleasing advertisement and (b) what is deemed to be a persuasive argument. For instance, in the 1950s, anything that smacked of science gave an aura of 'rationality' to the argument. Today, those same advertisements, with their secret ingredients, graphic cross-sections, and microscopic views, look as irrational as the sci-fi movies of the same period. Several campaigns of the Creative Revolution, like Alka-Seltzer and Volkswagen, still appear on industry 'best campaigns' lists 40 years after they were created. But others, like the Braniff campaign, are a puzzle to contemporary viewers—the Pucci costumes and 'Pop' planes aren't as fresh to today's eye.

Aesthetic notions are always closely connected to a culture's morality. Thus, outside the practical world of advertising, the distinction made between 'good' advertising and 'bad' advertising is often seated in the presumed value of 'information' versus the moral questionability of 'persuasion.' Writers trying to support such divisions among advertisements often use the presence of images as the basis for categorizing: words and numbers tend to be seen as 'informative' or 'rational,' while images are 'persuasive' and 'irrational.' Even an amateur rhetorician, however, can come up with examples in which the display of alphanumeric information (or the withholding of it) is itself an attempt at persuasion. Further, the visual presentation of information—even in the plainest style—affects its persuasiveness. Finally, information is often presented in pictorial form, as in maps and graphs. In many cultures, visual forms are used to record and transmit widely varied types of data. Today, writers like Edward Tufte have made us more aware of the intimate connection between visualization and information, whether our objective is persuasion, 'mere representation,' or facilitating our own understanding. Therefore, the distinction between the informative–written advertisement and the persuasive–imagistic advertisement seems to rest mainly on a naïve understanding of rhetoric and an outdated approach to the pictorial.

4. *How Advertisements Work*

There are almost as many theories about 'how advertisements work' as there are advertisements themselves. Several famous admen have published their

own treatises on how to make an effective advertisement. There is little agreement among them. Few, however, would subscribe to the theories that have been popular among academics. For instance, Freudian theories popular among cultural critics of the late twentieth century would inspire little more than laughter among those who actually produce advertisements. After a brief flirtation with Freudian notions in the late 1950s (an approach known in the business as 'MR,' which stands for 'motivation research'), the advertising industry abandoned attempts to sell products through subliminal cues, such as phallic symbols or 'hidden' images of genitalia. Such approaches, having proven ineffective, are now thought to be ridiculous. Shirley Polykoff (who wrote the 'Does she … or doesn't she' Clairol advertisements of the 1960s) ridiculed these attempts in her memoirs, published in 1975 (Polykoff 1975). She recalls that a TV spot introducing a women's shaving cream was changed to accommodate a psychoanalyst's suggestion: after the woman shaved, her legs should rise slowly as she leaned back in her chair, 'a position normally associated with an activity other than shaving.' The product was a bomb: 'It turned out no woman at that time, rising legs notwithstanding, was interested in paying for a shaving cream when she could use her husband's or a bar of soap that was readily at hand'(Polykoff 1975, p. 98). Cases like this one eventually put the MR people out of business.

Purchase behavior, as it turns out, is better explained by such pedestrian matters as price, practices, and preferences than by phallic imagery. So, in some ways, the subliminal advertising and motivation research fads of the fifties were displaced by a greater understanding of the reasons people consume. In *The Hidden Persuaders*, Vance Packard (1957) accused advertisers of manipulating 'hidden needs' to sell products, sparking a national controversy. Packard's 'eight hidden needs' were then thought to be irrational or even immoral as a basis for buying. Today, these needs—for self-esteem, for safety, for recognition—are no longer particularly 'hidden,' but are popularly accepted as being essential to human mental health. Further, scholars investigating consumption from a more anthropological point of view, like Mary Douglas and Mihalyi Csikszentmihalyi, have now published material emphasizing the complexity of consumption in all cultures, by all humans, and at all times. This literature richly and repeatedly illustrates the myriad of ways that people use material possessions to express themselves, to make connections with others, to protect their families, to enculturate their children, and so on. The tendency in Western post-industrial thought to insist on material life being properly separated from the spiritual or social does not hold up well under the weight of this evidence. A more sophisticated perspective now recognizes that people usually consume for social reasons even when they are satisfying biological needs: we eat dinner with our families because we need both nurture and nutrition. Indeed, the more inquiry is done on the dynamics of consumer behavior, the less currency is held even by the traditional split between needs that are biological (and therefore 'real') versus those that are social (and therefore 'false'). With this in mind, it becomes increasingly difficult to insist that advertising either plays on hidden needs or that it 'makes us buy things we don't need'—because the notion of 'need' becomes utterly redefined.

Elsewhere in academia, however, a different kind of psychology still tries to explain how advertisements work. Experimental psychologists in applied areas like advertising and marketing have tried hard to identify specific formal cues and appeals that will consistently produce sales for advertisers. These researchers design experiments that attempt to determine whether color, happy music, or big type 'work' better to sell us goods than black and white, sad music, or small type. After nearly 30 years of study, not a single generalizable finding has emerged from this research. The new psychological theories of 'how advertisements work,' therefore, are no more likely to win the respect of advertising professionals than the old, Freudian ones.

So, professionals approach the creation of advertisements in much the same way they always have: as a common-sense attempt to persuade their cultural kindred to buy a product. True, they have a great deal of information about our needs, desires, and dissatisfactions—and they do put that knowledge to use in crafting their messages. To that extent, advertisers may have some edge over the amateur persuaders we meet in everyday life: Girl Scouts who try to sell us cookies, preachers who try to get us to tithe, and offspring who try to make us buy toys. Nevertheless, advertising professionals basically work with rhetorical principles as old as Aristotle: after studying their audience, they offer information and arguments they believe will be convincing. They affect a character they think will match up with the audience's self-image. For 'sweetening' the proposition, they throw in melodies, jokes, and pictures. At the workbench level, this practice, for all the devilry attributed to it, is only a garden variety of rhetoric.

Like any other attempt at rhetoric, some of these practical efforts fail. Some sink dramatically, like Shirley Polykoff's advertisement for shaving cream. Few produce the kind of powerful propellant critics claim for this institution. The public and its pundits, though, continue to believe that advertisers have some secret power to manipulate them. Some of this suspicion is the advertising industry's own fault. In promoting their abilities to press and potential clients, advertising agencies have consistently overclaimed for their own skills. Their bravado, coupled with continued attempts to find magic formulas, makes the public justifiably wary.

Many myths thus continue about the power of advertising, lending fuel to the general discomfort with

commercial culture. For instance, the 1950s story of 'subliminal advertising,' in which unsuspecting US movie patrons were persuaded to buy cola and popcorn by flickering messages undetected in the film, is retold today as if it really happened. In actual fact, the whole event was a scam. James Vicary, the man who claimed to have accomplished this feat, was an unemployed market researcher when he reported his story to the press in 1957. While reporters and, soon, Congressmen investigated the validity of his claim, Vicary collected retainers from some of the nation's largest advertisers. But his claim quickly began to fall apart. First, the theater owner said no test had been done on his premises, as Vicary had claimed. Then, Vicary failed to produce the effect for a Congressional committee. Suddenly, less than 6 months after his name first appeared in the papers, Vicary disappeared. His closets and his checking account were left empty. He was never heard from again. Though many researchers have attempted to replicate his study, the purported effect has never been repeated. Even if Vicary was a charlatan, though, the true story does show that advertisers in the USA were willing to pay huge sums to learn how to advertise subliminally. So, however false this cultural myth may be, the suspicion of corporations that lies behind its continued viability is not without foundation.

5. *Conclusion*

Advertising seems to touch every aspect of life in the postindustrial world. As both form and institution, advertising is blamed for an array of social ills ranging from the mundane to the millennial. The poor standing advertising has in the global community at the opening of the twenty-first century is not based on the real existence of any secret formula, economic equation, or covert conspiracy. It is certainly not based on any demonstrable effect on prices, social conditions, or even sales. Rather, our attitudes toward advertising are probably a response to the rapid changes in everyday life brought about by industrial commerce, and are intensified by the knowledge that, at least some of the time, advertisers will make extreme attempts to have their way with the public. Nevertheless, the human mind, being far more subtle and sturdy than many theories would suggest, has shown remarkable resistance to all such attempts—the explosion of consumer culture notwithstanding. Our discomfort with advertising is also profoundly situated in ethnocentric prejudices against commerce, material comfort, sensual pleasure, and images. Rigorous thinking on the subject of advertising is rare—at least partly because we are blinded by these very fears and prejudices. Looking forward in a globalizing marketplace, however, should impress upon us the need to break out of such constraints.

See also: Advertising Agencies; Advertising: Effects; Advertising: General; Advertising, Psychology of; Media Effects; Media, Uses of

Bibliography

Aaker D A. (ed.) 1993 *Brand Equity and Advertising: Advertising's Role In Building Strong Brands*. L. Erlbaum Associates, Hillsdale, NJ

Barthes R 1982 *The Rhetoric of the Image, The Responsibility of Forms*. Hill and Wang, New York

Csikszentmihalyi M, Rochberg-Halton E 1981 *The Meaning of Things: Domestic Symbols and the Self*. Cambridge University Press, New York

Douglas M, Isherwood B C 1996 *The World of Goods: Towards an Anthropology of Consumption*. Routledge, London

Fox S 1984 *The Mirror Makers: A History of American Advertising and Its Creators*. Random House, New York

Frank T 1997 *The Conquest of Cool*. University of Chicago Press, Chicago

Leiss W, Kline S, Jhally S 1990 *Social Communication in Advertising: Persons, Products, and Images of Well-Being*. Nelson Canada, Scarbrough, Ontario

Marchand R 1985 *Advertising the American Dream: Making Way for Modernity, 1920–1940*. University of California Press, Berkeley, CA

Packard V O 1957 *The Hidden Persuaders*. McKay, New York

Polykoff S 1975 *Does She … or Doesn't She?: And How She Did It*. Doubleday, Garden City, NY

Presbrey F 1929 *The History and Development of Advertising*. Doubleday, Garden City, NY

Rogers S 1992/1993 How a publicity blitz created the myth of subliminal advertising. *Public Opinion Quarterly* 12–17

Schudson M 1984 *Advertising, The Uneasy Persuasion: Its Dubious Impact on American Society*. Basic Books, New York

Tufte E R 1990 *Envisioning Information*. Graphics Press, Cheshire, CT

Wicke J 1988 *Advertising Fictions: Literature, Advertisement and Social Reading*. Columbia University Press, New York

Williams R 1980 Advertising: the magic system. *Problems in Materialism and Culture* 170–95

L. Scott

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Advertising, Control of

Advertising is a form of communication between a firm and its customers, that uses independent media to communicate positive messages about a good. Firms supply it to generate sales and to counter their competitors' advertisements, but there is also a demand for advertising because consumers lack information, and much of it comes from advertisements that help lower inevitable 'search costs,' that is, consumers' expenditure of time and money to select what goods to buy.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

1. Advertising Failures

Like other institutions, advertising is prone to natural shortcomings (e.g., information is always incomplete) and artificial ones (e.g., agreements and rules not to advertise). Legal treatises and self-regulatory codes reveal that scores of advertising practices represent actual or potential failures that negatively impact the market system's dependence on adequate information and effective competition, and that impair the functioning of other institutions. Hence, control focuses on monopolistic power, consumer deception, unfairness, and social irresponsibility, although laws and codes vary in defining these concepts.

1.1 Monopolistic Power

Producers can erect 'barriers of entry' to superior and cheaper products by creating and advertising meaningless distinctions among brands. Furthermore, producers may collude to ban advertising, as in the case of legal services. Such rarely challenged restraints reduce competition and consumer welfare even when otherwise justified.

1.2 Consumer Deception

The representation of a product's features may mislead consumers acting reasonably. An advertisement cannot be as lengthy as an instruction manual or even a label, but should it include all significant information? This depends on how one defines 'information.' Do people want and need only 'objective' facts such as origin, ingredients, price, performance, and contraindications? Or should information refer to the attractiveness of a good in the context of a consumer's own buying criteria such as socioeconomic background, personality, lifestyle, aspirations, experience, and other factors—whether rational, emotional, or simply habitual? Thus, should government-sponsored lottery advertisements reveal that the odds of winning are abysmally low, such lotteries pay back the smallest share of any legal game, and the government's share is not fully or mainly used to support education or the arts, as was promised? Or should one accept the appeal of 'All you need is a dollar and a dream' because it reflects the needs of lower-income people? Both perspectives on information assume truthful communications although some puffery is tolerated (e.g., 'the King of beers'), and that advertisers should be able to substantiate their claims (e.g., 'the fastest copier').

1.3 Unfairness

'Unfair' means 'not sincere, frank, honest, loyal and right'—all subjective criteria. Unfairness is associated with particular practices such as denigrating through comparative advertising a competitive product or firm even if the latter is in fact inferior, and playing on the fears of people even when such fears are real (about sickness, death, social attractiveness, self-confidence, etc.) and could be alleviated through the advertised products and services (e.g., insurance and cosmetics). Although the legal definition of unfairness varies across countries, it usually involves: (a) malice (e.g., nobody mentions a competitor to say good things about it), or (b) calling attention to a product's benefits (e.g., comfort or relief) while minimizing or omitting mention of its costs (e.g., monetary harm or distress), or (c) the exploitation of the weaker by the stronger (e.g., large retailers out-advertising smaller ones).

1.4 Social Irresponsibility

Advertising should not undermine other institutions such as the state, the family, and the value system. Thus, advertising to children is thought to weaken parental authority and to develop consumeristic attitudes at an age where more important values should be impressed on the young. Larger and evolving questions about societal welfare and personal happiness are also involved. As a 'mirror of society,' does advertising reflect and magnify developments in ideologies and lifestyles through new symbols and habits detrimental to these goals? Is it even part of a plot to create a 'culture of consumption' essential for the capitalistic machine?

1.5 Extent of Advertising Failures

There are no counts of misleading, unfair, or irresponsible advertisements in the USA, but the self-regulatory UK Advertising Standards Authority has estimated that some 2 percent of British advertisements appear to violate some provision of its extensive Code of Advertising Practice. This is a low rate of malfeasance, compared to other institutions (e.g., the family and the educational system) and considering that much advertising misbehavior results from ignorance and carelessness rather than ill intent. Still, 2 percent of millions of advertisements add up to tens or hundreds of thousands of failures worldwide, suggesting a significant need for regulatory mechanisms to be used by competitors, consumers, governments, and concerned citizens.

2. Forms of Advertising Control

2.1 Mechanisms

Most people and organizations behave themselves because they want the esteem of other members of society, they fear losing markets, they are threatened by the law, and/or they want to lessen uncertainty

about their rivals' behavior. The community reacts to advertising's shortcomings by ignoring, discounting, and/or opposing it (e.g., defacing posters or criticizing tobacco advertisements). Individual and organizational self-discipline, in the forms of personal ethics and company codes of conduct (including media acceptance codes for advertisements), provides another vehicle for community-based control. Market responses include consumers shunning the goods of unreliable advertisers, and competitors countering advertisements with their own. State sanctions involve direct prohibition, restriction, obligation, public provision, and taxation, which reduce the reach and positive impact of advertisements, but the state can also facilitate legal action by customers and competitors. Under industry self-regulation, peers rather than outsiders establish and enforce self-imposed and voluntarily accepted rules of behavior, although such private governance may be mandated, regulated, and monitored. Since most advertisements require the use of media (the press, broadcasters, postal services, telephone companies, Internet service providers, etc.), their participation in self-regulation helps put an effective stop to objectionable advertisements.

2.2 Respective Strengths and Weaknesses

Effective control ultimately requires:

(a) developing standards;

(b) making them widely known and accepted;

(c) advising advertisers and agencies before advertisements are released;

(d) pre- and post-monitoring of compliance with the standards;

(e) handling complaints from consumers, competitors, and other parties; and

(f) penalizing bad behavior in violation of the standards, including the publicizing of wrongdoing and wrongdoers.

'Somebody' has to design, perform and fund these sizeable tasks.

Community control offers shared norms (e.g., against vulgarity) and broadly based pressure (e.g., boycotts) but lacks authoritative means to mobilize social resources beyond what can be obtained voluntarily on account of regard for others. Market competition generates more and better information but is weak on broader social responsibility issues. Government regulation benefits from universal coverage, compulsion, and legal enforceability although it may also impose high costs, stifle innovation, and be of limited effectiveness in dealing with complex and evolving issues such as bad taste and the reach of foreign media. Industry self-regulation can provide customer and competitor gains beyond the minimum standards of the law, and it benefits from the positive commitment of practitioners. However, industry support may be insufficient, unstable, and perceived as compromising the integrity of the system. Table 1 summarizes the respective strengths and weaknesses of the latter two systems—government regulation and industry self-regulation—which dominate advertising control.

2.3 Factors Affecting Use

The existence of some 200 nation-states with varying traditions, values, legal systems (common law versus civil or religious laws), economic resources, and advertising experiences precludes firm global conclusions. Still, the demand for controls reflects, first, that competitors want to protect their property rights in brands and reputations (e.g., against comparative advertisements) and to ensure fair competitive rules (e.g., against advertising allowances favoring large distributors); and, second, that customers and their associations are concerned about consumer deception, unfairness, and social irresponsibility.

On the supply side, governments are increasingly tackling such issues through regulation and (sometimes) the support of stronger market competition and self-regulation. Industry has offered self-regulation as a way of pre-empting or mitigating legislation and of improving the overall credibility of advertising. The proliferation of national and subnational 'states' as well as of supranational bodies (e.g., the European Union and the World Health Organization) has multiplied mandatory rules and advisory guidelines which have facilitated complaints, the stopping of advertisements through cease-and-desist injunctions, private suits, and even class actions by empowered competitors, consumers, and citizens.

Regulation and self-regulation are often complementary. Thus, all voluntary codes state that 'advertising must be legal,' although few self-regulatory bodies (e.g., the French Bureau de Vérification de la Publicité) pursue statutory infractions. Sometimes they specialize, as with the German ZAW industry organization that focuses on taste and decency because Germany's law against unfair competition applies to most other situations. Governments burdened by growing tasks and budget deficits increasingly invoke the principles of 'proportionality' and 'subsidiarity,' whereby higher control levels should not deal with what better informed and motivated lower levels can achieve more effectively. Even consumerist groups agree with these principles, provided they can meaningfully participate in self-regulation.

These developments reflect better knowledge about how advertising works (and how well) and about the effectiveness of controls. Thus, research has confirmed that market shares keep shifting when competitive advertising is allowed; the latter helps reduce consumer search costs; restrictions often result in higher prices and lower quality of services; and the cost of preventing and correcting harm to consumers, competitors, and citizens can be excessive. The US Supreme Court now requires that when public policy

Table 1
Strengths and weaknesses of advertising regulation and self-regulation

Advertising—control tasks	Government regulation	Industry self-regulation
Developing standards	+[a] Greater sensitivity to politicized concerns	−Greater lag in responding to such concerns
	−[b] Difficulty in elaborating standards for taste, opinion, and public decency	−More informed ability to develop and amend standards in these areas
	−Inertia in amending standards	+Faster response
Making standards widely known and accepted	+Everybody is supposed to know the law	−Difficulty in making the public aware of the industry's standards and complaint mechanisms
	−Compulsory nature of the law generates industry hostility and evasion	+Greater ability to make industry members respect both the letter and spirit of voluntarily accepted codes and guidelines
Advising advertisers about grey areas before they advertise	−Usually not provided by government	+Increasingly promoted and provided by industry—sometimes for a fee and even on a mandatory basis in the case of broadcast media and some products (toys, cosmetics, non-prescription drugs, etc.)
Monitoring compliance	±Routinely done or at the prompting of others but often with limited resources	±Increasingly done by the industry although restricted by financial resources
Handling complaints	+Impartial treatment is anticipated	−Treatment more likely to be perceived as partial
	+Potential to handle many complaints	−Limited capacity to handle many complaints in some countries
	−Slow and expensive	+Faster and less expensive
	−Cannot put the burden of proof on advertisers in criminal cases	+Usually puts the burden of proof on the challenged advertiser
Penalizing bad behavior, including the publicity of wrongdoings and wrongdoers	+Can force compliance	−Problems with non-compliers but the media usually refuse to carry ruled-out advertisements
	−Generates hostility, foot-dragging, appeals, etc.	+More likely to obtain adherence to decisions
	−Limited publicity of judgments unless picked up by the media, competitors, and activist groups	+Extensive publicity of wrongdoings and, to a lesser extent, of wrongdoers

a Strengths are indicated by '+'. b Weaknesses are indicated by '−'

justifies government controls, the latter be shown to be effective and not disproportionate to the goal to be achieved. Litigation (as in tobacco's case) is making available data about advertisers' strategies, budgets, and performance so that investigations, complaints, and suits are better informed.

Both regulation and self-regulation are being fostered by supranational governments (e.g., the European Union's directives on misleading and comparative advertising), international organizations (e.g., the World Health Organization's development of a global treaty to control tobacco sales and advertising), and industry groups (e.g., the International Chamber of Commerce's International Code of Advertising Practice, and its constant promotion of self-regulation).

2.4 *Impacts*

Self-regulatory bodies handle up to 10,000 complaints a year (with some duplications about the same advertisement) when they are actively solicited, as by the UK Advertising Standards Authority. However, the number of adjudicated cases can be relatively low

(about 100 a year by the very active US National Advertising Division of the Council of Better Business Bureaus) when these bodies choose to focus on exemplary cases that send strong signals or break new grounds. Government agencies are also very selective in their interventions (e.g., about 100 cases a year by the US Federal Trade Commission) for the same reasons and because of budgetary constraints. However, courts are increasingly handling advertising cases.

Contract law potentially protects customers everywhere against misrepresentation in advertisements. Private suing to stop advertisements or for damages has become common although usually done more by competitors than by consumers, as is also true of complaints to self-regulatory bodies. Consumers and their associations are slowly being empowered to bring class actions against companies whose products cause injury (e.g., cigarette advertisers). Under statutory law, certain practices are forbidden, circumscribed, or mandated (e.g., health warnings) and criminally punishable, but regulatory bodies can also issue civil law-based cease-and-desist injunctions as well as require corrective advertisements. Self-regulatory codes are being expanded to include more products (e.g., toys) and services (e.g., advertising by charities), as well as practices (e.g., Internet advertising).

How effective have regulation and self-regulation been in promoting truth, accuracy, fairness, social responsibility, and competition? No precise evidence exists, and their impacts vary depending on a country's level of development as well as on its priorities. There will always be crooks, so that control mechanisms can only aim at improving the overall quality of advertising at a reasonable cost through exhortation, coercion, punishment, denial of access to the media, and other mechanisms. Besides, new problems keep emerging on account of technological innovations (e.g., the Internet and satellite broadcasting, which ignore national borders), new entrepreneurial initiatives (e.g., sports sponsorship and political advertising), and citizen activism (e.g., growing concerns about children and advertising, and privacy invasion).

One can reasonably conclude that such problematic practices as unsubstantiated claims about weight reduction and other 'cures' have been significantly curbed in developed countries through regulation and self-regulation. Besides, the principle that advertisers have a 'prior reasonable basis' before making a claim has been broadly adopted and enforced in many states. More advertisements about particular goods (e.g. toys) and in certain media (particularly, broadcasting) must be approved in advance; certain claims (e.g., 'free, low-fat, and environmentally sound') are now restricted by law, while mandated information (e.g., annualized interest rates in financial advertisements) is proliferating.

However, such controls have also curbed competition and denied valuable information to consumers, as with bans on pharmaceutical, liquor, and legal-services advertising. Besides, protecting a lengthening list of 'vulnerable groups' (the young, the sick, the recently bereaved, the poorly educated, etc.) shrinks the pool of those whom advertising may target without legal or self-regulatory limitation.

3. *Ongoing Developments*

3.1 *Technology*

Means of reaching consumers keep multiplying: satellite broadcasting, international editions of newspapers, magazines and catalogs, the Internet, etc. However, technology also allows consumers to avoid advertisements (e.g., with zapping devices) and to access them only voluntarily and selectively; and it facilitates the fast sharing of information among consumers, associations, and governments as well as the filing of complaints. The Internet increasingly links advertisers, suppliers, customers, and citizens via computers, television sets, other receivers, and processors. Potentially, there will be many more millions of advertisements that can be quickly reformatted, as well as many more advertisers, including individuals. The lines between editorial content and advertising message, which have traditionally been separated, will become blurred because there will be many 'banner' advertisements and sponsors linked to programs and information. Advertising agencies that help screen advertisements may be less involved, replaced in part by less-informed and socialized Internet service providers.

3.2 *Laws and Codes*

Regarding the right to give and receive information, constitutions usually guarantee freedom of the press because democracies need independent media to challenge governments in 'the marketplace of ideas.' However, do these 'ideas' encompass non-misleading advertising messages, thereby providing for 'freedom of commercial speech?' Since the 1970s, the US Supreme Court has come to acknowledge such a right, supported also by the Council of Europe's Declaration on Freedom of Expression and Information, and by Article 10 of the European Convention on Human Rights. However, this new freedom is limited because it conflicts with growing rights to health, privacy, and sex equality—for example, pharmaceutical advertising remains restricted because it may lead to greater health expenditures.

Regulation and self-regulation may proliferate at subnational, national, and supranational levels, thereby making their harmonization less likely. Such heterogeneity will complicate the creation and delivery of advertisements although large advertisers can more

readily comply with varying rules. It also creates favorable loopholes to advertise where restrictions are lower, and to sue where litigation is most likely to succeed. Besides, more controls encourage the use of imagery ('Things Go Better with Coke') in lieu of challengeable claims.

When advertisements emanate from another sovereignty, both regulation and self-regulation face the problem of 'conflict of laws': what body, domestic or foreign, decides the case; what procedures (e.g., about required evidence) to follow; what law or code is applicable; how to recognize and enforce decisions? Governments have favored the laws of the advertisement's country of origin, and so does the self-regulatory European Advertising Standards Alliance, as well as international codes for pharmaceuticals and direct marketing. However, the European Union has recently granted consumers rights to sue locally.

Regarding the Internet, regulatory control will focus on: (a) criminalizing the sending of certain advertising contents (e.g., pornography and pharmaceuticals); (b) having advertisers and service providers notify advertisement-receivers of private-data collection practices (particularly from children) and of their right to check these data's accuracy, and be assured that such information is secure; and (c) obligating access providers (including telecommunication companies) to at least respond to complaints, since it is impossible for the latter to screen out millions of messages. Thus, the Internet will test and refine the control of advertising in coming years.

See also: Advertising and Advertisements; Advertising: Effects; Advertising: General; Broadcasting: Regulation; International Advertising; International Communication: Regulation; Mass Media, Political Economy of; Mass Media, Representations in; Media Ethics; Media, Uses of

Bibliography

Boddewyn J J 1989 Advertising self-regulation: true purpose and limits. *Journal of Advertising* **18**(2): 19–27

Boddewyn J J 1992 *Global Perspectives on Advertising Self-Regulation: Principles and Practice in Thirty-eight Countries*. Quorum, Westport, CT

Calfee J E 1997 *Fear of Persuasion: A New Perspective on Advertising and Regulation*. Agora, Monnaz, Switzerland

Petty R D 1997 Advertising laws in the United States and European Union. *Journal of Public Policy and Marketing* **16**(1): 2–13

Schudson M 1984 *Advertising, the Uneasy Persuasion: Its Dubious Impact on American Society*. Basic Books, New York

Vakratsas D, Ambler T 1999 How advertising works: what do we really know? *Journal of Marketing* **63**: 26–43

Wotruba T R 1997 Industry self-regulation: a review and extension to a global setting. *Journal of Public Policy and Marketing* **16**(1): 38–54

J. Boddewyn

Advertising: Effects

1. Introduction

The phenomenal growth of advertising in the twentieth century helped usher in and sustain a defining consumer culture. That is advertising's surest effect. Consumers come into contact with literally hundreds of ads every day. This ubiquity has its effects. People use phrases and other utterances from ads in daily discourse (Friedman 1991, Ritson and Elliot 1999), revealing more than mere mimicry, but also the ad-inspired consumer ethos of contemporary consumer culture. Advertising is now part of the interstitial tissue of daily life on this planet. Yet, the effects of advertising are often anything but clear or easily detected. In fact, the effects of advertising run the gamut from obvious to perplexing and contradictory. On the one hand, most people, most of the time, don't care much about ads. They don't pay attention to them. But every now and then, they do. And even when they don't, the effects may still accrue.

2. Context

Advertisements have been studied in one way or another for about 100 years, yet we simply don't know much about their effects. There are at least three good reasons for this lack of knowledge. First, scholars have viewed advertising as unworthy of attention. This is the same high culture (and anticommercial) bias which led libraries to cut ads out of magazines before binding them for decades. Many libraries gave space to pornography but not advertisements. It was politically difficult to propose studying something that was regarded as thoroughly unworthy of study. This bias has had a profound effect on this scant scholarly enterprise. Relatively little research has been done, and much of it is more polemic than systematic. Second, the object of study itself presents an inherent difficulty. Advertisements are by their nature textual, generally ephemeral, and encountered as a common social object, situated within layer upon layer of social discourse. They are here and then they are gone. They are commercial rhetoric, actively interpreted by various audiences in various ways. While the general meaning of an advertisement is informed and bounded by cultural forces so as to produce more likely than less likely interpretations, there is no single, objective advertisement to study (Fish 1980, Scott 1994). What an advertisement is depends on who is interpreting it and why. This seemingly simple point has thoroughly eluded most advertising researchers. Further, ads exist in their particular social–temporal microclimate and then they become something else. Even the once-legendary 'Man from Glad' is not now what he once was. Typically stripped of context, we have little idea of what advertisements meant in their own milieu.

The third reason for the paucity of work in this field is inadequate method. It is extremely difficult to demonstrate the effects of advertising, and the job is made even more difficult within the confines of a single method, a single literature, a single paradigm. Survey data suffer from the scores of variables that lie between exposure and behavior. Experiments suffer from low ecological validity; they too often have little or nothing to do with how real people experience and interpret real ads in real contexts. Ethnographic work on advertising has been rare, but suffers from its lack of generalizability, the effects of intrusion, and the privileging of informants and text. Textual analysis, and even reader response, suffers from the authority of the one constructing the interpretation. Work by economists with aggregate data suffers from a long and troubling list of assumptions, particularly those positing rational thought. Advertising is a sociotextual phenomenon, which does not travel well outside its natural environs. In their natural space advertisements are a typically trivial background to daily life. Methods sensitive to this reality have yet to be employed to any significant degree.

But advertising certainly must have effects. How could something so widespread, so much a part of contemporary existence, and something which businesses worldwide continue to invest in massively, have no effect? Beyond the reasonable and predictable retort that 'real' effects are not necessarily detectable and demonstrable effects, there must be something to which we can point and say: 'we (probably) know this.' The more robust advertising effects findings are reviewed below.

3. *Economic Effects*

Economists have studied the effects of advertising spending on various outcome variables such as GDP, aggregate and category demand, brand switching, economic concentration, and barriers to entry. Borden (1947) was one of the first to attempt a systematic economic analysis. While acknowledging that advertising might increase 'selling costs,' (p. 881) his study points to the vital need for advertising in growing 'dynamic economies' (p. 881). He sees little downside to advertising. Running counter to this view are both Packard (1960), and Galbraith (1958), who see advertising as the engine of waste and the loss of consumer agency. These are not, however, empirical examinations, but more economic–philosophical theories. They represent the political left's view sans data.

Schmalensee (1972) conducted the next major empirical analysis of the economic effects of advertising. His work showed 'that total national advertising does not affect total consumer spending for goods' (p. 85). He argues in a very real and practical sense that sales cause advertising, much more than the other way around. He goes on to say that a major implication of this is that 'advertising cannot be all-powerful if it cannot influence the decision to save.' He finds national advertising to be particularly inconsequential in driving aggregate demand. In terms of competition, he says: 'the effects of national advertising spending by the various consumer goods industries cancel each other out in the aggregate' (p. 86).

The next major exhaustive empirical examination of advertising's economic effects is that of Albion and Farris (1981). Their work revealed a much more complex and nuanced set of findings than previously reported, but were generally supportive of Borden (1947) and Schmalensee (1972). In other words, most of the criticisms of advertising were not borne out by the data. They do, however, find advertising expenditures and industry concentration to be correlated, but argue that other factors, as well as reverse causality, may be at work. Most significantly, they note the overwhelming importance of context in these effects. Schudson (1984) also makes the case for context and the limited power of national consumer goods advertising. He argues that advertising follows affluence, targeting consumers who already consume large quantities of the goods or services concerned. He gives good examples of both advertising without sales, and sales without advertising. Luik and Waterson provide one of the more recent (1996) reviews of this literature and find that advertising affects aggregate demand early in a product's life, but as time goes by, has more of an impact on brand share, and less of an impact on aggregate demand. That is, for most mature product categories, advertising seems to impact on brand share, but not total category demand.

4. *Sociocultural Effects*

The cultural effects of advertising have largely been the purview of those in the humanities, as well as those in areas of communications, sociology and anthropology. More recently, a few in marketing have contributed as well. A wide range of orientations and methods has been brought to bear on the search for advertising's effect upon societies and culture. This in fact constitutes the single largest (and most widely accessible) literature on advertising effects.

These works claim that advertising has produced several effects at the sociocultural level. For one, advertising is said to have caused a more materialistic culture (Ewen 1988, Fox and Lears 1983, Pollay 1986, Richins 1991). This argument rests on the traditional modernist critique: the real has been replaced with artifice, community with mass society, and a human orientation with one in which material things are more central to human existence. By extension, it is claimed that advertising has placed more emphasis on brand names as symbols than on the actual sum and substance of goods and services. This privileges simu-

lacra over the 'real.' Likewise, advertising is said to have worked as a hegemonic force to the detriment of women, minorities, and other marginalized peoples (Faludi 1991, Seiter 1995). These are the major effects claims at the sociocultural level. Although there are many other derivations, they are largely subsumed by those categories outlined above.

The evidence for these claims is, however, rarely as strong as the accompanying rhetoric. Still, there are a few evidence-based conclusions. First, advertising has certainly made consumer culture possible. This may be the biggest advertising effect of all. To understand this, one has to appreciate advertising's role in advancing the practice of branding. Advertising made possible national brands, without which consumer culture would have died on the vine. Advertising projected brands and brand consciousness into national consciousness and into greater social centrality (Muniz and O'Guinn 2001). Advertising made a brand-oriented society possible—perhaps inevitable. In 1850, few things were branded; by the end of the twentieth century, water and dirt were branded. The whole purpose of a brand is, in economic terms, to create greater inelasticity. A primary predictor of inelasticity is the consumer's belief that there are few acceptable substitutes. For example, when advertising made Ivory more than just soap, demand for Ivory became less elastic than demand for mere 'soap.' The same thing happened with beer, and soft drinks, and thousands of other products. So, clearly advertising lead to a proliferation of brands, which contributed to a more brand-oriented society. The brand is now a defining construct in contemporary consumer culture. This is an effect of advertising. While societies have always been involved with material culture, the massive extent to which branded culture emerged is quite clear, and not just a matter of degree. This is an advertising effect. Whether or not this has caused individuals to value things more than before is unknown, although it does seem clear that it probably produced more value for the explicitly (and highly marked) commercial object. Whether it caused individuals to value things more in relation to humans than previous generations did is a much harder case to make.

What then of advertising and hegemony? Advertising did and does provide support to modal social structures. Advertising, by its nature, is generally (but not always) a supporter of the status quo. Most researchers do not find advertising to be a great friend to progressive social movements, although it certainly has been from time to time. Unfortunately, many of the arguments regarding the hegemonic dynamics of advertising are hyperbolic and ahistorical (Scott 2002). Further, the relationship between progressive movement, resistance, consumer agency, and advertising is far more complex than typically held (Muniz and O'Guinn 2001). Consumer resistance is now played out by the rejection of one brand (perhaps an ecologically friendly, green brand) over another brand. Belief in a 'star chamber' of corporate elite, or advertisers who control consumer's minds, is being thoroughly discredited (Scott 2002) by social history research.

Advertising's effects in the areas of race and gender are far from simple, but have generally been negative. There is every reason to believe that narrow, pejorative, and stereotyped social portrayals have had detrimental effects when it comes to people of color. The same is no doubt true in the case of women, gays and lesbians. The evidence for this is more inferential than direct, but is consonant with empirical studies in other domains. We know that these groups were absent, stereotyped or under-represented in typical advertising. Exactly what the effects are on self-perception and other types of perception is less well understood. We can assume it to be anything but positive, however.

In fact, social constructions of reality in general have been shown to be influenced by mass media and advertising portrayals of social reality (O'Guinn and Shrum 1997). Recently, O'Guinn (2002) has shown that those exposed to a great deal of advertising on television believe there to be higher than actual rates of 'ad-problems' such as gingivitis, athlete's foot, bad breath, etc. There is good reason to believe that advertising has a significant impact in terms of simply making certain things more heavily represented in consumer's constructions of social reality.

5. *Individual Effects*

A considerable amount of advertising 'effects' research focuses on the individual. Most has been published in two academic fields: mass communication and marketing, typically from a psychological perspective.

In the 1940s researchers became interested in the effects of mass communication in general. Given the post-World War II consumption boom, advertising was of obvious relevance. The working model of mass communication was at that time a very simple direct effects model: message strikes audience member like a bullet. Experience with propaganda efforts in the war, the ascendancy of psychology in both the academic and larger social world, and outright naiveté led to this very simple notion of hypodermic needle-like effects (O'Guinn and Faber 1990). There was a great deal of belief in the power of communications at a critical point in the social history of advertising. Exuberant communications researchers were going to wipe out bigotry, inform the electorate, and make things better in all aspects of life, including the realm of modern consumption. This belief was a popular and self-serving one for several groups, including the advertising industry itself.

It is not coincidental that the postwar period is literally drenched in Freudian and neo-Freudian notions. With the mind-control hysteria of the cold

war, the prominence of psychology as a science and references to the subconscious mind led to a new exuberance in advertising research in both industry and academia. While the social scientists and public Freudians such as Ernest Dichter had major philosophical differences with one another, they both led to a psychological effects tradition in advertising research. To this day, discussion of advertising effects is influenced by the pervasive (and misguided) beliefs of the period.

There were, however, some exceptions. In communication, Katz (1957) proposed the 'two-step flow' model in which certain social actors and their communication networks were seen as more important than others in the spread of information and influence. This was the first time that any kind of social consciousness was included in the 'effects model' at all. This model held that other social actors mattered as much as the communication message itself, if not more. Since advertising was mass-mediated communication it was assumed to act in the same way. It was at about this time that another new field, marketing, discovered the communication literature. From the late 1950s through the mid-1970s, several marketing researchers went looking for the elusive generalized opinion leader, a person to whom advertising could be effectively targeted. This person would then pass commercial information through interpersonal channels to great effect. But alas, the search was fruitless: the generalized opinion leader could not be found, and most marketing research again eschewed the social for the more psychological 'message in a bottle' approach, at least for a while.

Another critical effects notion came along in the 1960s: the 'hierarchy of effects.' The basic idea was that effects of communication (in particular advertising) occurred along a hierarchy: from exposure to behavior (Colley 1961). Each step along the way depended upon the step before. This helped to explain why advertising was generally inefficient, why it took so much to move consumers all the way to behavior from awareness. While a vast improvement compared with an undifferentiated bullet theory, it was still bound up in the idea of a linear and fairly inflexible progression of effects. But it inspired a great deal of advertising effects research. Generally speaking, research confirmed that not all ads lead to behavior. This was progress.

A significant trend in advertising effects research began in the mid 1970s. Known as the 'information processing' tradition, its orientation has been to study ads as the bearers of 'information,' which is then 'processed' by audience members. It was assumed that people actually did something with information gained from ads. This too was progress. These researchers held (at least implicitly) that underlying psychological processes would have to be understood in order to ultimately understand advertising's effects, and make them generalizable and predictable. Most of the earliest work in this tradition was heavily influenced by then-contemporary attitude theory (Fishbein and Azjen 1975), but later became increasingly more cognitive in orientation. Of all the things advertising does, its ability to get consumers to remember the brand, its name, and something good about it, has the longest research tradition. Interestingly, recall of advertising seems important in some instances, but completely useless in others. Keller (1993) demonstrates that advertisements are much better retrieved, and retrieved in a manner most desirable to the advertiser, when there is a carefully planned congruence between the advertisement itself and the point of purchase. The point of purchase should be rich with memory retrieval cues that are entirely consistent with elements found in the ad itself. This work is very important in that it explicitly links point of encoding and point of retrieval. This helps to explain why recall of the actual ad seems important sometimes, and at other times, relatively worthless.

One model has dominated research in the psychological realm. It is the elaboration likelihood model (ELM) of Petty and Cacioppo (1986). This model says that there are two routes to persuasion: one peripheral and one direct. The direct route is marked by attention, consideration, engagement; in other words, 'elaboration.' The consumer engages arguments, or attempts at persuasion in a variety of ways, but generally takes them head-on, agrees with some, counter-argues others, etc. The peripheral route is just the opposite; there is little cognitive effort or engagement, and little motivation to process. In addition to these two routes, various cues were also assessed as either peripheral or direct. For example, source credibility, presenter attractiveness, or executional elements such as music, color, etc., are thought to be peripheral. Research emanating from the ELM has been encouraging, and generally supportive of the theory. Peripheral cues are at their best in situations where the consumer is less involved, has less at risk, and attention is low; central processing is most effective when the opposite is true. It is generally thought that long-term advertising effects are greatest when the direct route is engaged. In the long run, getting people to pay attention to ads seems to predict their greatest effect. While it is not exactly clear where emotions fit in the ELM, work by Edell and Burke (1993) demonstrates that emotions (particularly warm and upbeat ones) do influence the overall evaluation of the brand, and are recalled just as well as other types of information.

Another contribution from the effects literature is in the distinction between search attributes and experience attributes (Nelson 1974, Wright and Lynch 1995). Search attributes are objective pieces of consumer information: how many miles to the gallon, does the car come in red, how much horsepower does it have, etc. Experience attributes are those things that one can only really know via direct experience: does the engine sound good, did the seats feel comfortable, does the

stereo sound good, etc? Wright and Lynch (1995) show that search attributes are actually better learned through exposure to advertising, while experience attributes are better learned through direct contact and use of the product or service. This effect is stronger under low involvement conditions than under high involvement. This finding is important because it has long been assumed that advertising was a fairly weak force compared to direct experience, in almost all cases. This is an important qualification of advertising effects. Similar findings by Edell and Burke (1986) show that consumer's pre-existing attitudes toward the brand have a great deal to do with the attitude-to-the-ad/ attitude-to-the-brand connection, particularly as they mediate through motivation to process the ad.

Another interesting research area is the 'third person effect' (Davison 1983). This work demonstrates that individuals attribute much greater power to advertising when asked about some other (third) person than himself or herself. In other words, they believe advertising to affect others much more than themselves. The effect is particularly strong when those doing the judging are higher in socioeconomic status (Atwood 1994). So, relatively more affluent and educated people think that others, particularly poorer and less educated people, are more affected by advertising than they are. This effect has considerable implications for effects research in general, and public policy in particular. This work is typically carried out through surveys.

6. *Special Audiences*

Two audiences receive special attention in terms of advertising effects: young children (those under 12 years of age) and the elderly. The clearest findings are those reported by Roedder-John and Cole (1986). They find that significant processing difficulties exist in both children and the elderly. These are identified as 'memory-strategy' usage deficits, or strategies for the effective encoding and retrieval of stored information. These strategies are important in knowing how to effectively gain information from advertisements, and to defend oneself against some forms of deception (intentional or not). In addition, young children suffer from knowledge base deficits. They simply do not yet possess the requisite knowledge base to know how to interpret some advertisers' messages. The implication is that the effects of advertising on these populations may be significantly different than for others.

7. *Going Forward*

The arguments surrounding the relative worth of this literature are fairly typical philosophy of science ones. Experimentalists argue that the only way one is ever going to be able to understand the effects of advertising is to understand them within the tightly controlled confines of the laboratory, where all other noise is ostensibly screened out, and process is detected and discerned. More naturalistic researchers counter that advertising cannot be divorced from its social context and construction, and thus the pursuit of such effects is a comfortable conceit. Not too surprisingly, both sides make valid points. There are certain processes which would simply be hard, by their definition, to study outside a laboratory, at least where the detection of psychological process is concerned. On the other hand, things closer to actual behavior are more entangled, by their nature, in discourse, and social circumstance. Ethnographic work holds some promise. Actually watching people watch ads (at least on television) makes some degree of sense (Ritson and Elliot 1999). Yet this work is limited to what people do with advertising content in their daily lives, or the social uses of advertising. We do not typically see ethnographic research at the point of exposure, or have any idea what is going on inside the audience member's head. Furthermore, there are all the familiar criticisms of ethnography: lack of control and ungeneralizable findings. Historical and textual methods also offer promise but it is certain that no single method is going to reveal enough about advertising effects on its own. The phenomenon is simply too large, too layered, and too multifaceted.

One of the problems beyond method is that advertising is simply not very important to most people, most of the time. As Klapper argues in his 1960 limited effects model: mass communication is only effective under two conditions: (a) where the communication is completely consonant and resonates with some social theme, or (b) in a total *tabula rasa* situation. In other words, it does two things very well: delivering information and increasing knowledge. Beyond that it is an extremely weak force at the micro-individual level. The original VW Bug ads of the late 1950s and 1960s are good examples of the former. The VW advertising rode a developing wave of pop consumer counterculture. The ads provided information consonant with this social movement. In other instances, literally nothing is known about specific goods or services, and advertising provides that information. A good example is the introduction of the compact disc player. In that instance advertising supplied consumers with essential information about something that was completely unknown to them. Advertising wrote useful information on a blank slate. But this is rare. Most of the time it is the noise in the background while we are preparing dinner, or the things between the segments of our television shows. Its effects are, by advertising's nature, difficult to pin down.

See also: Advertising Agencies; Advertising and Advertisements; Advertising: General; Advertising, Psychology of; Consumer Culture; Hegemony: Cultural;

Mass Communication: Normative Frameworks; Media and Child Development; Media Effects; Media Effects on Children; Media, Uses of; Psychology and Marketing

Bibliography

Albion M S, Farris P W 1981 *The Advertising Controversy: Evidence on the Economic Effects of Advertising*. Auburn House, Boston

Atwood E L 1994 Illusions of media power: The third-person effect. *Journalism Quarterly* **71**(2): 269–81

Borden N H 1947 *The Economic Effects of Advertising*. Irwin, Chicago

Colley R H 1961 *Defining Advertising Goals for Measured Advertising Results*. Association of National Advertisers, New York

Davison B 1981 The third-person effect in communication. *Public Opinion Quarterly* **47**: 1–15

Edell J A, Burke M C 1986 The relative impact of prior brand attitude and attitude toward the ad on brand attitude after ad exposure. In: Olson J, Sentis K (eds.) *Advertising and Consumer Psychology*. Vol. 3. Praeger Publishers, Westport, CT, pp. 93–107

Edell J A, Burke M C 1993 The impact and memorability of ad-induced feelings: Implications for brand equity. In: Aaker D A, Biel A L (eds.) *Brand Equity and Advertising: Advertising's Role in Building Strong Brands*. L Erlbaum Associates, Hillsdale, NJ, pp. 195–212

Ewen S 1988 *All Consuming Images: The Politics of Style in Contemporary Culture*. Basic Books, New York

Faludi S 1991 Beauty and the backlash. In: *Backlash: The Undeclared War Against Women*. Anchor, New York

Fish S 1980 *Is There a Text in This Class?* Harvard University Press, Cambridge, MA

Fishbein Martin, Ajzen I 1975 *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA

Fox R W, Jackson Lears T J 1983 *The Culture of Consumption: Critical Essays in American History 1880–1980*. Pantheon, New York

Friedman M 1991 *A 'Brand' New Language: Commercial Influences in Literature and Culture*. Greenwood, Westport, CT

Galbraith J K 1958 *The Affluent Society*. Riverside, Cambridge, MA

Katz E 1957 The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly* **21**: 61–78

Keller K L 1993 Memory retrieval factors and advertising effectiveness. In: Mitchell A (ed.) *Advertising Exposure, Memory and Choice*. L Erlbaum Associates, Hillsdale, NJ, pp. 11–48

Klapper J T 1960 *The Effects of Mass Communications*. Free Press, New York

Lichtenstein M, Srull T K 1985 Conceptual and methodological issues in examining the relationship between consumer memory and judgement. In: Alwitt L F, Mitchell A A (eds.) *Psychological Processes and Advertising Effects: Theory, Research and Applications*. L Erlbaum Associates, Hillsdale, NJ, pp. 113–28

Luik J C, Waterson M J 1996 *Advertising & Markets: A Collection of Seminal Papers*. NTC Publications, Oxford

Muniz A Jr, O'Guinn T C 2001 Brand community. *Journal of Consumer Research*. March

Nelson P 1974 Advertising as information. *Journal of Political Economy* **83**(July/August): 729–54

O'Guinn T C 2002 'Social Anxiety Advertising and Consumers' Perceptions of Base Rates of Product Addressed Social Problems.' Unpublished manuscript.

O'Guinn T C, Faber R J 1991 Mass communication theory and research. In: Kassarjian H H, Robertson T S (eds.) *Handbook of Consumer Behavior Theory and Research*. Prentice Hall, Englewood Cliffs, NJ, pp. 349–400

O'Guinn T C, Shrum L J 1997 The role of television in the construction of consumer reality. *Journal of Consumer Research*. March: 278–94

Packard V 1960 *The Waste Makers*. David McKay, New York

Petty R E, Cacioppo J T 1986 *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer Verlag, New York

Pollay R W 1986 The distorted mirror: Reflections on the unintended consequences of advertising. *Journal of Marketing* **50**(April): 18–36

Richins M L 1991 Social comparison and the idealized images of advertising. *Journal of Consumer Research* **18**: 71–83

Ritson M, Elliott R 1999 The social uses of advertising: An ethnographic study of adolescent advertising audiences. *Journal of Consumer Research* **26**(December): 260–77

Roedder-John D, Cole C 1986 Age differences in information processing: Understanding deficits in young and elderly consumers. *Journal of Consumer Research* **13**(3): 297–315

Schmalensee R 1972 *The Economic Effects of Advertising*. North-Holland, London

Schudson M 1984 *Advertising, The Uneasy Persuasion: Its Dubious Impact on American Society*. Basic Books, New York

Scott L M 1994 The bridge from text to mind: Adapting reader response theory for consumer research. *Journal of Consumer Research* **21**(December): 461–86

Scott L M 2002 *Fresh Lipstick: Redressing Fashion and Feminism*. University of Illinois Press, Urbana, IL

Seiter E 1995 Different children, different dreams: Racial representation in advertising. In: Dines G, Humez J (eds.) *Gender, Race and Class in Media: A Text Reader*. Sage, Thousand Oaks, CA, pp. 99–108

Wright A A, Lynch J G Jr 1995 Communication effects of advertising versus direct experience when both search and experience attributes are present. *Journal of Consumer Research* **21**(March): 708–18

T. C. O'Guinn

Advertising: General

Advertising is central to the study of media and the commercial applications of the social sciences. Not only does advertising revenue provide the major source of income for print and broadcast media owners, but it gives those media their characteristic look and sound, and orients their content towards the kind of audiences which advertisers want to reach. Ad-

vertising can be defined as a kind of cultural industry which connects the producers of consumer goods and services with potential markets through the diffusion of paid messages in the media. What is loosely referred to as advertising in everyday language is really just the most visible form of marketing, in which particular images become associated with branded goods and services. It is the imaginary dimension of what may be called the manufacturing–marketing–media complex of modern societies, the whole institutional structure of production and consumption.

1. *Advertising as a Cultural Industry*

Not all media depend upon the sale of their space and time for advertising—films, for example, don't carry advertising as such—and not all advertising depends upon the media for its diffusion. Some companies spend more on 'below the line' or nonadvertising forms of promotion, such as direct mail. However, in most countries of both the developed and the developing world, the largest advertisers of consumer goods and services, many of them global corporations, have determined the direction and character of print and broadcast media development because of their demand for the means ('media') which allow them to reach potential consumers. This is true regardless of whether the media are owned by private interests or the state, as is seen in the case of the rapid commercialization of television in Europe in the 1980s and 1990s.

Advertising is cultural in that it deals in images which give public expression to selected social ideals and aspirations, but it is also an industry because of its crucial intermediary role in the manufacturing–marketing–media complex. It was in this sense that Raymond Williams called it 'the official art of modern capitalist society' (1980, p. 184).

The intermediary role is carried out by advertising agencies (see *Advertising Agencies*). These 'agents' produce and place advertising for their 'clients,' the consumer goods or service companies who are the actual advertisers. However, while the clients pay the agencies for their services, agencies also derive income directly or indirectly from the media in the form of a sales commission paid in consideration of the media space and time, such as newspaper pages or television 'spots,' which the agencies purchase on behalf of their clients.

This is the crux of the manufacturing–marketing–media complex referred to above. Indeed, this form of dealing in media space and time was the origin of advertising agencies as a form of business in the latter nineteenth century, and continues as a fundamental aspect of the industry today. The judicious purchase of media space and time is part of the range of functions routinely performed for clients by 'full-service' agencies, while there have emerged specialized agencies which deal in such media purchasing alone.

Looking at the complex of relations between advertisers, agencies, and media from the point of view of the media as an institution, it becomes evident that in most countries, the dominant print and broadcast media are 'commercial,' as distinct from government-funded or nonprofit media. This means that the measurement of readerships and audiences becomes an issue for advertisers, as they want to be sure that their advertisements are reaching the kind and range of the prospective consumers they seek (see *Audience Measurement*). For their part, the media must seek to attract those prospective consumers as readers or viewers, and this becomes the basis for the kind of content they provide.

While this relation is easiest to observe in the case of broadcast television and its constant striving to offer programs which can attract the largest audiences, mostly for the benefit of national advertisers of packaged consumer goods, to many advertisers, the kind of audience can be more important than the size. The agency's expertise in the choice of medium for the client is a key element in their mediation of the process of fitting people to products. Thus, while the broad appeal of television content might reflect 'the demographics of the supermarket' (Barnouw 1979, p. 73), other media such as prestige newspapers and special interest magazines are supported more by advertisers who are interested in reaching only affluent consumers or niche markets respectively.

All of these industrial connections are of most interest to political economists of the media, while social science and cultural studies more generally have been more concerned with the cultural output of agencies, in the form of the marketing campaigns which they devise for their clients, and the actual advertisements which they produce. Here once again, full-service agencies provide various other marketing activities, including market research, a major form in which social science methods are applied for commercial purposes (see *Market Research*). However, it is the 'creative' side of the agencies' work, including that of specialist creative 'hot shops,' which gives advertising its cultural significance.

2. *The Development of Advertising*

Advertising has its origins not just in the sale of space and time, but in its role of providing distinct identities to branded goods and services. It arises in the US and UK around the late nineteenth century when packaged household products began to replace generic goods bought in bulk: Pears rather than soap, to name one of the world's very oldest brands. By giving brands a particular character and often a logo or slogan to make them recognizable, advertising contributed to the national, and later the international, expansion of brands such as Lipton's, Gillette, Kodak, and Ford. By the 1920s, advertising was forging an alliance with what were then the still fairly new social and

behavioral sciences, drawing on sociological techniques for market research, and devising psychological appeals for advertisements themselves.

Also around this time, some of the longest-established US agencies began to expand overseas. For example, J. Walter Thompson upgraded its sales office in London, while McCann-Erickson was opening up branches in Latin America. In both cases, this expansion was initiated at the request of large clients whose brands they handled in the US, and that wanted similar advertising services provided for them on a 'common account' basis in the foreign markets they were opening up. US brands such as Coca-Cola thus became known worldwide.

However, the most important period for the internationalization of the advertising industry was the period after World War II. Apart from this being the era in which many US companies transformed themselves into multinational corporations, it was also the time when television was a new medium being adopted by one country after another around the world. Because the US agencies had their common account agreements with their clients from their home market as well as some experience with television, they were able to quickly dominate the advertising industries of those countries they entered.

Indeed, by the 1970s, this movement had provoked some resistance, and US advertising agencies became one of the targets of the rhetoric against 'cultural imperialism' in that decade. Subsequent expansion then had to proceed more through joint ventures in conjunction with national agencies. However, a more elaborate pattern emerged in the 1980s, particularly with the growth of certain British agencies, notably Saatchi & Saatchi. No longer was the worldwide advertising industry defined by a fairly simple conflict between US-based multinationals vs. the various national interests, but there were more complex tendencies emerging.

First, what the British agencies did was to keep their new US acquisitions intact and incorporate them into 'megagroup' structures, a strategy already undertaken by McCann-Erickson following its merger with SSC&B:Lintas. This was a way of dealing with 'client conflicts': for example, British Airways could be handled by Saatchi & Saatchi, while other airline accounts could be assigned to different agencies in the same group without fear of marketing secrets being leaked from one to the other. Apart from this kind of integrated concentration in the English-speaking world, there has been a trend to the interpenetration of national markets and of world-regional markets, including joint ventures by French and Japanese agencies in particular (Mattelart 1991). These trends have become more marked with the globalization of the 1990s (see *International Advertising*).

Globalization has had controversial effects upon the content of advertising. While some agencies have advocated the standardization of advertising campaigns throughout the world ('one sight, one sound, one sell'), others have argued for what the Sony Corporation calls the 'glocalization' of both products and their marketing: that is, adapting them in accordance with the cultural differences evident between the various national markets. In practice, services such as credit cards or airlines appear to benefit from global advertising, but goods such as packaged foods do not (Sinclair 1987, Mueller 1996).

3. *The Critique of Advertising*

As the advertising industry has been consolidating itself at its most intensive level of globalization ever, it is ironic that the social critique which accompanied it during its previous decades of expansion has been fading into obscurity. However, for most of the latter part of the twentieth century at least, advertising has been subject to considerable critique from various quarters of society, and on economic as well as cultural grounds, so it is worth reviewing the terms of the debate and the sources of criticism.

It has been the critics of advertising who have made the most elaborate claims about its economic importance. Marxist and liberal critics alike have argued that advertising creates and controls demand, thus attributing to it a key function in the perpetuation of consumer capitalism. Certainly, modern marketing gives producers a wide range of sophisticated strategies and techniques, but advertising is only one element among them. Furthermore, the fact that the vast majority of new products still fail when they are introduced to the market suggests that there is no one in control of demand.

Another economic criticism of advertising is that it adds to the price of goods and services, because advertising costs are real costs which are passed on to the consumer. The rebuttal is, that because 'advertising is the shortest distance between producer and consumer,' to quote J. Walter Thompson, it rationalizes the distribution process, generating only small costs which are spread easily over the large volumes of production which advertising helps to make possible.

The defenders of advertising argue that it is necessary to maintain competition, and that if producers don't advertise to maintain their market share, they will be driven out by their competitors, creating oligopolies, or markets dominated by very few producers. However, it follows that advertising is also a barrier against the article of new players, who must be able to achieve high levels of advertising from the beginning. This favours the 'market power' of the existing producers, so advertising would appear to favour oligopolistic markets.

In practice, it is evident that some of the industries which advertise the most heavily, such as packaged foods, or household and personal cleaning products, do indeed have tendencies to oligopolistic concen-

tration. Furthermore, it is worth noting that such oligopolies tend to maintain themselves not by the mass advertising of just one product line, but by product differentiation: for example, different shampoos for different types of hair. Advertising is implicated here because of its crucial role in branding, 'positioning,' and otherwise creating apparent differences between products for different types of consumer.

The connection between the ideological and the cultural dimensions of advertising is achieved most fully in the Marxist critique, where advertising is seen to perform ideological 'functions' which reproduce the capitalist system as a whole. In the US, Stuart Ewen has argued that department store owners and other ideologues of capitalism in the 1920s embarked on a 'project of ideological consumerization' (1976, p. 207) to draw working people into compliance with capitalism. More recent theory would see this as part of the 'pact' between capital and labor which stabilized the growth of capitalism in the first part of the century, the age of 'Fordism.'

In the UK, Marxist critics such as Raymond Williams (1980) were also influenced by a British tradition of social critique which denounced advertising on moral grounds, not so much because of its materialist values, but more because of its persuasive appeals and irrational associations. Something of this is also found in the critique of the Frankfurt School theorist Herbert Marcuse, who sees capitalism as sustained by 'false needs': 'The prevailing needs to relax, to have fun, to behave and consume in accordance with the advertisements' (1968, p. 22).

The Marxists have not been isolated in the critique of advertising, however. Liberal critics of capitalism in the US in the 1950s, a formative period in the history of the manufacturing–marketing–media complex, were read throughout the English-speaking world, thanks to the innovation of paperback publishing. Particularly in the context of the Cold War and popular fears about ideological 'brainwashing,' exposés of 'motivational analysis,' and the 'manipulation' of symbolic meanings in advertising, notably Vance Packard's *The Hidden Persuaders*, first published in 1957, found a ready audience. The following year, J. K. Galbraith's *The Affluent Society* came out criticizing the US economy for the way it produced the very needs which its goods were intended to satisfy (1974).

The twentieth century's last phase in the social criticism of advertising came from the women's movement, beginning in the 1970s. Advertising was a major cultural field which women cited as evidence of the social processes through which sex-roles were represented and reproduced through 'stereotyping.' In particular advertising was seen to perpetuate women in subordinate domestic roles, at the kitchen sink in detergent advertisements, for example, or alternatively, as sexual objects presenting themselves for the pleasure of the male gaze. Women's criticism of the 'sexism' of their representations in advertising was sustained into the 1980s, often with accompanying activism, even as advertising began more to represent the 'new' independent woman (van Zoonen 1994). In the 1990s, feminist critique has been absorbed into a broader academic approach to consumer culture.

In addition to the specific economic effects attributed to advertising, as outlined earlier, each of these phases of social criticism focuses on one of advertising's alleged cultural effects—that advertising has an ideological role in papering over social inequalities, that it creates false and irrational needs, and that it subordinates and degrades women. As well, there are other social concerns, with a similar perception of its power, which legitimize a public interest in the regulation of advertising—that it takes advantage of children, for example, or that it encourages young people to take up smoking (see *Advertising, Control of*).

Accordingly, advertising attracts some form of regulation in all countries, but with great variation, particularly in the ratio of self-regulation by the industry itself, as compared with government-prescribed regulation. Apart from the considerable cultural variations about what is acceptable in advertising images, advertising regulation typically restricts the type of product which may be advertised, such as alcohol and tobacco; the media which may be used, and at what times; the extent of product claims and invidious 'comparative' advertising; and in some cases, the use of foreign-produced advertisements.

No doubt the capacity of advertising to exert social influence is greatly overestimated, and its structure and processes much misunderstood, as Schudson has so thoroughly argued (1984), but even in an age of 'deregulation,' most governments are reluctant to surrender much of their control over advertising.

4. *Advertising in Social Theory*

The social criticisms of advertising referred to in the previous section, arising from liberalism, Marxism, and feminism, are grounded not just in those various social and intellectual movements as such, but in the considerable academic theorization and research which they have generated in the social sciences.

While it would be less true of the US than the UK, Canada, and Australia, Marxism provided the dominant paradigm in cultural studies and much social science, particularly sociology, from the mid-1970s until at least the end of the 1980s. This was a 'Western' Marxism in which there were two main trends: one towards 'political economy,' which emphasized the ownership, control and functioning of the economic structure of capitalism; and the other towards the cultural analysis of role of ideology in maintaining the

system as a whole. In both trends, but particularly the latter and more influential way of thinking, advertising was seen to have a crucial role to play in stabilizing a society which would otherwise be torn apart by its own contradictions. Following the French Marxist structuralist philosopher Louis Althusser, Marxist social theory in the 1980s thus shifted its analytical focus away the economic structure as the basis for capitalist society, and towards ideological reproduction—the representational and signifying practices of capitalist culture, including advertising.

This tendency in cultural Marxism found common cause with semiological structuralism (see *Semiotics*), derived from Ferdinand de Saussure, but mobilized most famously with regard to advertising by Roland Barthes (1977). In this approach, advertisements themselves became the main object of analysis, such that the emphasis was upon how the various meaningful elements, or the 'signifiers' in an advertisement, related to each other so as to produce the meaning of the advertisement as a whole (see *Advertising and Advertisements*). This is a qualitative, interpretive approach which contrasts with the quantitative method of 'content analysis.' The latter, with its roots in US behaviorism, nevertheless has been applied in complementary and productive ways in conjunction with semiological analysis, such as in Leiss et al.'s study of changing images of well-being in US advertising (1986).

Along with Marxist and semiological structuralism there has been a contribution from anthropological structuralism, stemming from Claude Lévi-Strauss. These strands are all brought together, along with feminism and the psychoanalytic development theory of Lacan, in Judith Williamson's *Decoding Advertisements*, one of the most definitive books on the analysis of advertisements (1978). She provides a coherent fusion of these theories, and applies them in the qualitative analysis of scores of magazine display advertisements, to evince such processes as 'interpellation' and the invocation of ideological 'referent systems' in the interpretation of advertisements. Apart from Williamson's application of Lévi-Strauss's theory of 'totemism' to explain how certain kinds of people become associated with particular products in advertising, such as the 'Pepsi generation,' anthropological structuralism provides a way of understanding how goods become endowed with cultural significance through their position in a total system of meaning (Douglas and Isherwood 1979, Appadurai 1986). Advertising visibly contributes to this process of giving meaning to goods, but by no means exclusively. For example, of several different international brands of sportswear advertised in similar ways, it will be the peer group that decides which of them is to be the preferred brand in any given locality.

The same relational quality of the cultural meaning of goods is also found in the poststructuralist contribution of Jean Baudrillard (1981). In his view, capitalist social structure is the source of needs as well as of the meaning of goods, and like certain of the liberal and Marxist critics of advertising cited above, Baudrillard sees the rise of consumer capitalism as a device by which the system has avoided the need to redistribute its wealth. Thus, class differences are concealed beneath an apparent democracy of consumption, a connection which is lost in the bewildering and endless display of signification.

With the advent of poststructuralism and postmodernism, the diversification of feminism, and the eclipse of Marxism, there has been much less critical attention paid to advertising as such during the 1990s. Rather, although traditional press and magazine display advertisements, television commercials, and billboards continue to provide the examples of postmodern visual culture which are cited, contemporary theory and research sees advertising in a broader and now much more theorized context. This is even wider than the manufacturing–marketing–media complex described above. Thus, for Wernick (1991), advertising is just a part of 'the promotional condition of contemporary culture,' which goes beyond the marketing of commercial goods and services to include the mode of public communication now embraced by all major social institutions, from political parties to universities, and also found in the presentation of one's own self.

In ways like this, current theory and research is moving beyond the study of advertising as such, and more towards consumer culture in general (see *Consumer Culture*). Not only has this shift encouraged attention to the role of hitherto-neglected institutions such as the department store and the supermarket, but also to transformations in work, domestic life, and cultural identities, insofar as these have become expressed and commodified in terms of consumer goods (Lury 1996). This agenda in turn gives rise to studies of how specific groups have been constructed, represented, and appealed to by marketing strategies, and of how they have responded (Nava et al. 1997).

Such a line of investigation is a welcome corrective to the preponderance of attention given to the content of advertisements themselves, without regard to their audiences, which has characterized nearly all theorization and research about advertising to date. Furthermore, it provides some insight into the 'reflexiveness' with which audiences are now seen to regard media and consumption in the era of globalization (Lash and Urry 1994). This entails a postmodern esthetic in which consumers express themselves as individual subjects by means of how they mobilize their knowledge of the codes of meaning which goods carry, codes which are partly bestowed by the images in advertising, marketing, and the media, but which become the rules of the game on the street. Clearly, this cultural relation of mediated images and their expressive use can not be understood by analysis of the images alone.

Finally, as far as studies of the manufacturing–marketing–media complex are concerned, the challenges are to keep pace with globalization, to comprehend its complexities, and to monitor the new relationships which are taking place between marketing and media with the growth of new technologies. As the much-vaunted end of the age of 'mass' media slowly becomes reality, for example, as free-to-air 'broadcast' television loses audiences to subscriber or 'narrowcast' services, advertisers are becoming more discriminating and strategic in their use of these media, and more aware of the interactive 'pointcast' access to potential consumers available over the Internet (Myers 1999). The capacity to deliver audiences to advertisers will continue to determine the development of media into the new twenty-first century as it did in the old.

Bibliography

Appadurai A (ed.) 1986 *The Social Life of Things: Commodities in Cultural Perspective*. Cambridge University Press, Cambridge, UK

Barnouw E 1979 *The Sponsor: Notes 0n a Modern Potentate*. Oxford University Press, Oxford, UK

Barthes R 1977 *Image-Music-Text*. Fontana, London, UK

Baudrillard J 1981 *For a Critique of the Political Economy of the Sign*. Telos Press, St. Louis, MO

Douglas M, Isherwood B 1979 *The World of Goods*. Allen Lane, London

Ewen S 1976 *Captains of Consciousness: Advertising and the Social Roots of the Consumer Culture*. McGraw-Hill, New York

Galbraith J 1974 *The Affluent Society*. Penguin, Harmondsworth, UK

Lash S, Urry J 1994 *Economies of Signs and Space*. Sage, London

Leiss W, Kline S, Jhally S 1986 *Social Communication in Advertising: Persons, Products, and Images of Well-being*. Methuen, London

Lury C 1996 *Consumer Culture*. Polity Press, Cambridge, UK

Marcuse H 1968 *One Dimensional Man*. Sphere, London

Mattelart A 1991 *Advertising International: The Privitization of Public Space*. Routledge, London

Mueller B 1996 *International Advertising: Communicating Across Cultures*. Wadsworth, Belmont, CA

Myers G 1999 *Ad Worlds: Brands, Media, Audiences*. Arnold, London

Nava M, Blake A, MacRury I, Richards B (eds.) 1997 *Buy this Book: Studies in Advertising and Consumption*. Routledge, London

Packard V O 1977 *The Hidden Persuaders*. Penguin, Harmondsworth, UK

Schudson M 1984 *Advertising, The Uneasy Persuasion: Its Dubious Impact on American Society*. Basic Books, New York

Sinclair J 1987 *Images Incorporated: Advertising as Industry and Ideology*. Croom Helm, London

van Zoonen L 1994 *Feminist Media Studies*. Sage, London

Wernick A 1991 *Promotional Culture: Advertising, Ideology and Symbolic Expression*. Sage, London

Williams R 1980 *Problems in Materialism and Culture: Selected Essays*. Verso Books, London

Williamson J 1978 *Decoding Advertisements: Ideology and Meaning in Advertising*. Marion Boyars, London

J. Sinclair

Advertising, Psychology of

1. Short Historical Overview

Ever since mankind has advertised for a specific goal, he has considered how the advertisement should be designed to optimalize reaching the target. As early as 1898 Lewis had formulated his well-known rule, 'AIDA' (Attention, Interest, Desire, Action). This is still used today, although it is not scientifically substantiated.

Scientific research in applied psychology had already began considering the effects of advertising by the beginning of the twentieth century (Scott 1908, Münsterberg 1912). The impact of a certain stimulus on the observer was examined. Classical advertising psychology adopted the Economic Advertising Impact Model: it assumed that advertising effects on the (purchasing) behavior of the target group could only be predicted through the characteristics of advertising—the 'Stimulus–Response Model' (SR Model). Processes within the individual were not considered.

It was not until research showed that the SR Model did not provide the required information that researchers changed the direction of their investigations. This movement was towards the cognitive, motivational, and emotional processes within the individual. A new dimension was introduced into psychology—the SR model evolved into the 'Stimulus–Organism–Response Model' (SOR Model). This continues to be used today, at times in a modified form, to explain and examine relevant processes.

2. Advertising as a Part of the (Socio-) Marketing Mix

Advertising can be perceived as a supplier-initiated, purposeful method free of obligation. It is used to influence the target individuals through specific communication methods to such an extent that they accept an offer. It is a so-called distribution-political measure, in addition to the pricing, the distribution channels, and the offer itself. The offer can be a product (for example, a car or shampoo), a service (for example, a medical service or a taxi ride), or an idea (for example, religious ideas or political programs). These distribution measures are called a marketing mix when the products have a real market price (for example, cars) and a socio-marketing mix when there is no market price (for example, a political program).

3. Definitions Used in Advertising Psychology

Different fields approach advertising from different viewpoints, for example, history, law, communication, and economics. Advertising psychology takes a special look at the experience and the behavior of the recipients of advertising. This field is very specialized and therefore it is important for advertising psychologists to cooperate and work in an interdisciplinary fashion with members from other areas when analyzing and creating advertising.

A characteristic of the psychological viewpoint is that one of its main aspects, 'the experience,' is in the phenomenal world. Other characteristics of psychology are subjective perceptions and ideas. With regard to advertising this means that the psychological view of advertising cannot be seen independently from the offer, the price, and the distribution method. An economist can differentiate easily between the price and the advertising method, but advertising psychology studies have shown that the same advertising before and after a price increase can be perceived totally differently. Moreover, the same price can be experienced differently depending on the advertising used.

On the one hand, the psychology of advertising is an applied research science. It approaches questions out of applied areas and looks for methods used in basic research to answer these questions. The psychology of advertising is, however, also put into practice. It attempts, on the basis of the current state of research, to provide helpful answers to questions that occur in practice. This approach describes the professional life of an advertising psychologist not working in research (von Rosenstiel and Neumann 2001).

4. A Psychological Model of the Effect of Advertising

Various models in advertising psychology make the assumption that advertising is interpreted by the receiver as a stimulus coming from outside creating an

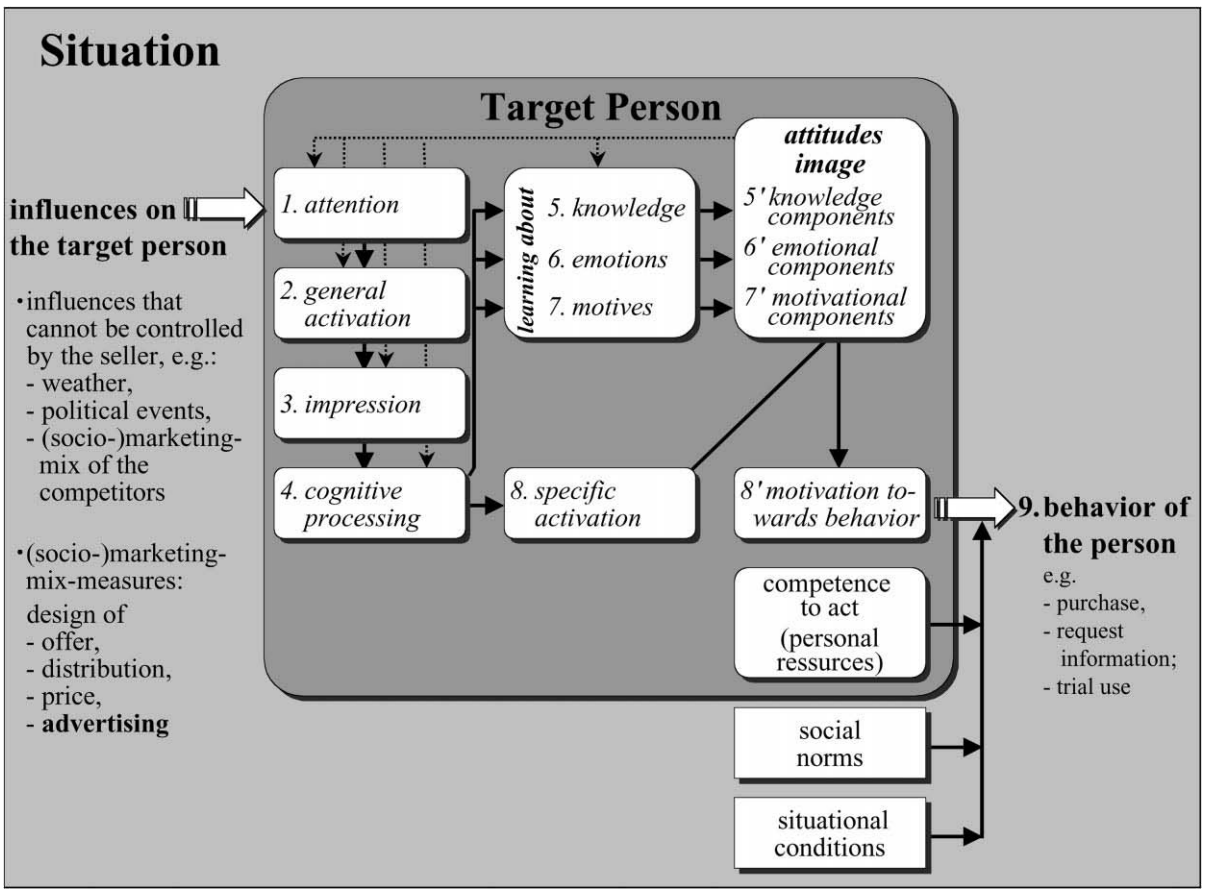

Figure 1
Psychological model of the effect of advertising (adapted from Neumann 2000, S. 18)

inner reaction. The person perceives this stimulus, is activated, and processes it cognitively, emotionally, and motivationally, which then results in a more or less extensive intention to act (for example, to ask for a brochure or to make the purchase). Whether or not this intention turns into action or not, depends on the competencies to act, the social norms involved, and the situational barriers. Figure 1 shows these relationships. Of course, this model, as all models, is a simplification and serves only to provide an understanding of the elements involved.

Figure 1 shows that advertisement, when seen as the influencing effect, is always interacting with other distribution-political measures, as well as with other factors which cannot be influenced, such as weather, economic swings, or political crises.

If advertising is to be successful and positively influence the behavior of the recipient, the psychological success criteria must be determined (as an in-between step) and indicators for their measurement must be operationalized.

The following aspects are important to analyze:

(a) Is the advertising noticed at all?

(b) Does it evoke enough general psycho-physical activation?

(c) Does it lead to optimal impressions?

(d) Is the incoming information adequately cognitively processed?

(e) Which of the following information is learned:

(i) Knowledge about the offer?

(ii) The feeling associated with the offer?

(iii) The all-encompassing motive associated with the offer?

(The learning processes (i), (ii), and (iii) can be described as image or attitude creation. Through this process the offer is positioned in the area of experience of the target person.)

(iv) Does the advertising cause a specific (motive) activation, which can then evolve into the motivation for a specific behavior?

Out of this motivation, in combination with the appropriate action competence (personal resources, such as buying power), the correct social norms as well as the beneficial or restricting conditions of the surrounding situation results the:

(v) Behavior (for example, ordering brochures, making an appointment for a test drive, or an actual purchase, which can be seen then as an indicator of economic success).

5. *Creating Advertising and the Methods for Measuring the Effects*

While the economic success of advertising can be registered by independent observers and is therefore objective in the sense of intersubjective agreement, it is only the individual perceiving the advertising introspectively. Therefore, one needs methods and procedures which make it possible to make these processes and phenomena accessible from the outside. Methods can be used which give answers to the questions about whether advertising is noticed, what feelings it evokes, or how efficient an existing image has been changed (Jones 1998). Selected methods for measuring psychological and economical advertising success can be seen in Fig. 2.

From experiences measuring future psychological advertising success (advertising success prognosis) or the actual advertising success (advertising success control), some basic principles have been determined regarding the development and use of advertising (Percy and Woodside 1983, Kroeber-Riel 1991). Examples of these are visually transported advertising such as newspapers, billboards, TV, and film.

Advertisements should be concise and clearly structured in terms of content and should correspond with the motives, attitudes, and expectations of the target group. This should make them noticeable. The written message or the picture should have topics which evoke emotions in order to create activation, but without taking the attention away from the actual product.

Regarding the overall format of the advertisement, attention must be paid to the choice of colors, shapes, and structure. Furthermore, it should not evoke negative emotions as a first impression and the written information and the picture should be easy to understand.

Since advertising often has the goal of improving the image, which can be defined as the attitudes regarding the offer, it is important to consider the following:

(a) Increase the specific knowledge about the offer by providing concise and clear information.

(b) Elicit positive emotions with pictures and words which can be applied to the offer and give it added value compared with other competitive products.

(c) Improve motivation by using written statements about potential effects of the purchase or models who show that they are successful with the offer.

In a real decision making process, these motives can turn into buying motivation if advertising is used, and this can be measured with appropriate scales which indicate the willingness to buy. Advertising is geared to access the intention to purchase and when the advertisement has hints about where, how, and under which conditions the product can be bought, the service acquired, or the idea be taken over, this can succeed. Some examples illustrating the relevance and importance of the above considerations are listed below. These show how relevant such tests before advertisement release can greatly improve the effectiveness of the advertisements on the target group.

The first example (from the authors) is of reading glasses to be used as a logo for a book club. There were two color variations for the frame, orange and red. An experiment using a tachistoscope where subjects were exposed to the glasses for 1/1,000 of a second which

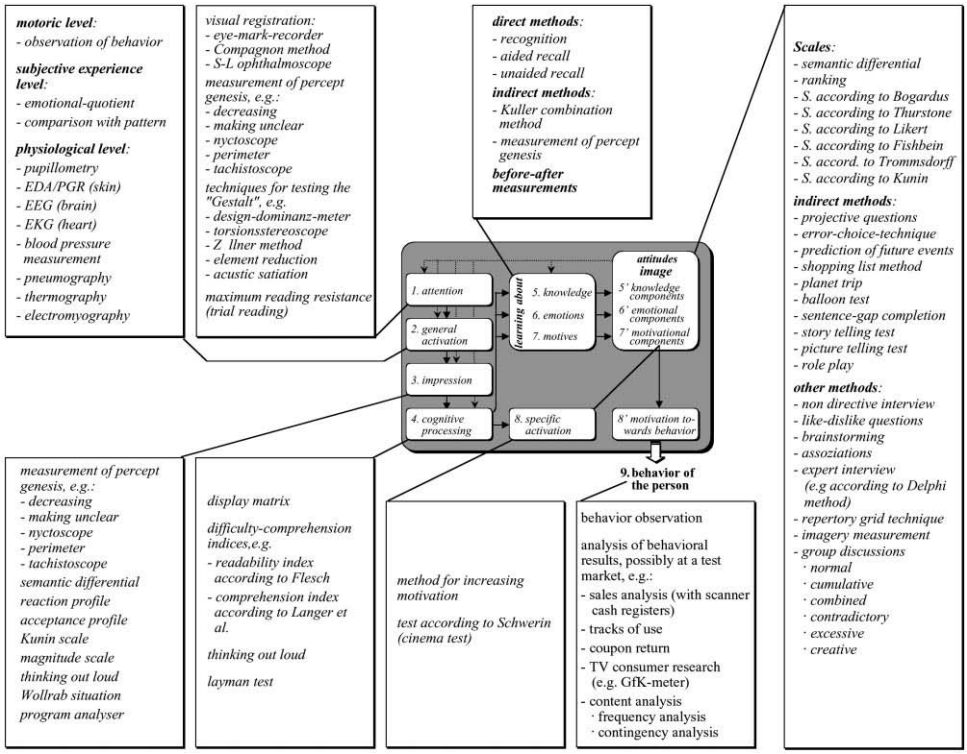

Figure 2
Methods for assessing psychological and economical advertising success (adapted from Neumann 2000, S. 268)

meant only seeing the color, not the shape, resulted in the following emotional associations:

(a) orange—mainly positive associations such as sun, warmth, and holidays, but also a few negative associations such as chemicals and danger;

(b) red—mainly negative associations such as blood, accident, hospital, and death.

On the basis of this study, the decision was made in favor of the orange glasses as the logo. Another example (Spiegel 1970, S. 61) is advertisement with a country picture with vineyards and an old train, representing the origin of a brandy. In the initial phases of perceiving the picture, associations with an industrial landscape arose. Such initial thoughts did not follow the direction of the desired advertising message. However, a systematic variation of the white steam from the train engine resulted in removing the section of steam blocking a part of the vineyards. With this alteration, the initial negative perception was removed.

In addition, one study (Neumann 2000, S. 97) illustrates that positively experienced advertisements are remembered significantly more than those experienced as less positive. In the first step, subjects were shown four advertisements for cigarettes and four for drinks and asked to rate their relative liking of each on a scale. After 30 minutes, using unaided recall, the subjects were asked to note which of the advertisements they remembered. Figure 3 shows the results, where the correlation between the two steps was very high, 0.79.

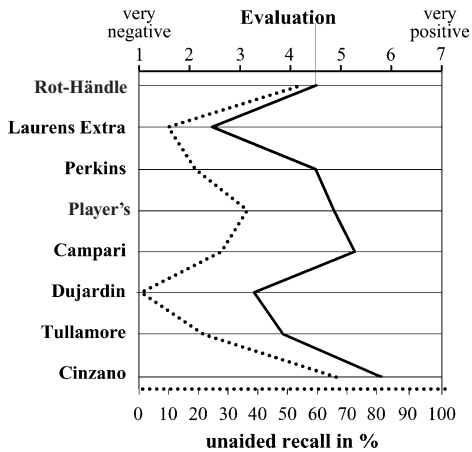

Figure 3
Evaluation and unaided recall for eight advertisements

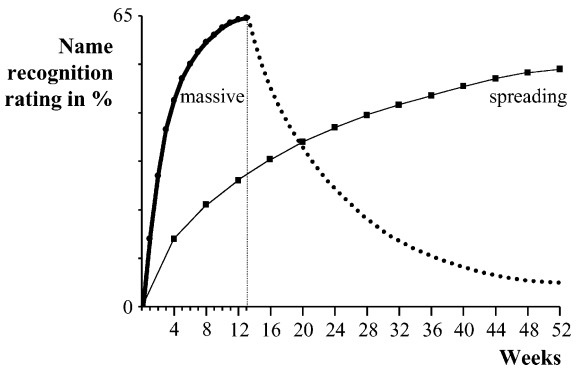

Figure 4
Name recognition rating and distribution over time (Zielske 1959)

In a classical study, Zielske (1959) tested spreading as opposed to massive printing of advertisements. A specific advertisement was shown to two parallel target groups 13 times: one group each week (massive), the other group spread over a year (spreading). The success can be read from the Fig. 4 which shows the percentage of the target group which were able to remember the advertisement.

If the target is a high level of name recognition rating massive advertisement is advisable. However the effect disappears rapidly. Typical examples for this are advertising campaigns before elections or summer/winter sales. Normally, the aim is a long-term effect where the second alternative is more appropriate, or, when the budget allows, a combination of both the massive and spreading approaches.

Finally, in a study (Neumann 2000, S. 217), four film spots were used which advertised 'alternative' targets: against torture, the killing of fur-bearing animals, nuclear power, and destruction of the ozone layer. Figure 5 shows sections of the results that were collected by using a specific characteristics profile of the optimal alternative advertising spot profile. Figure 6 shows the achieved factor analysis position of each spot based on this profile.

If the advertising method has been optimized according to psychological criteria and has been tested with the appropriate methods, then the measurement of the economic advertising success is to be seen as nearly equivalent to a test of the hypotheses for psychological advertising success. However, one has to take into account that other factors influence the economic advertising success and they must also be taken into consideration. Therefore it is hardly possible, despite optimal advertisement, to have economic success if the product is hard to find, the price is higher than what is seen as fair, or if the ownership or use of the product contradicts social norms or is perhaps even forbidden.

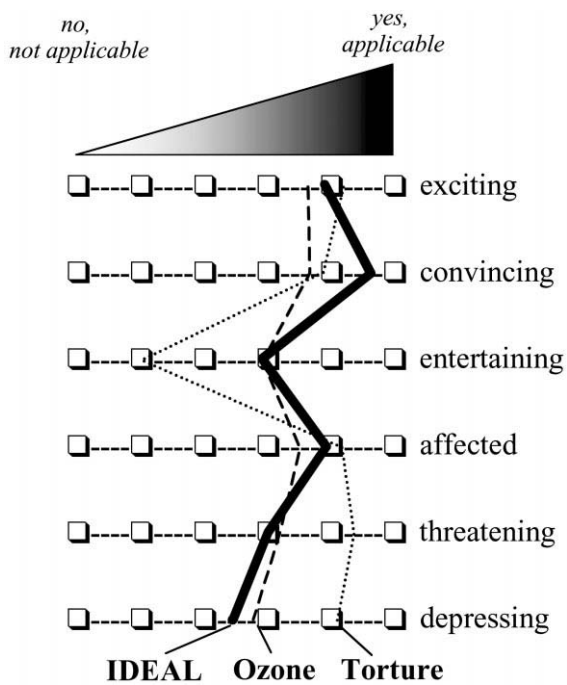

Figure 5
Characteristics profile of spots for alternative advertising (optimal conception—continuous line); spots against torture (from Amnesty International), and against the destruction of the ozone layer (from Greenpeace)

6. *Consequences and Value Problems*

Advertising is often criticized because it is created in such a way that as many people as possible notice it and because it is such an integral part of a market economy. Advertising is visible and audible for both children and adults, it influences our experience and our behavior, and it has a socializing effect. Therefore it does not come as a great surprise, but makes many upset, that for example, children know more car makes than they do tree or animal names or that the motivation to do something that costs money (for example, a cruise) is higher than to do something that is practically free of cost (for example, a hike in the nearby mountains).

There is also a great deal of criticism regarding the fact that consumer advertising shifts social norms. Advertising shows people, for example, who are astoundingly young, healthy, active, happy, and wealthy, so that an individual could perhaps see him/herself as underprivileged and then develop unrealistic expectations, consume too much, and go into great debt which could place them and their families' economic existence in danger. Especially the view of women in advertisement has been criticized

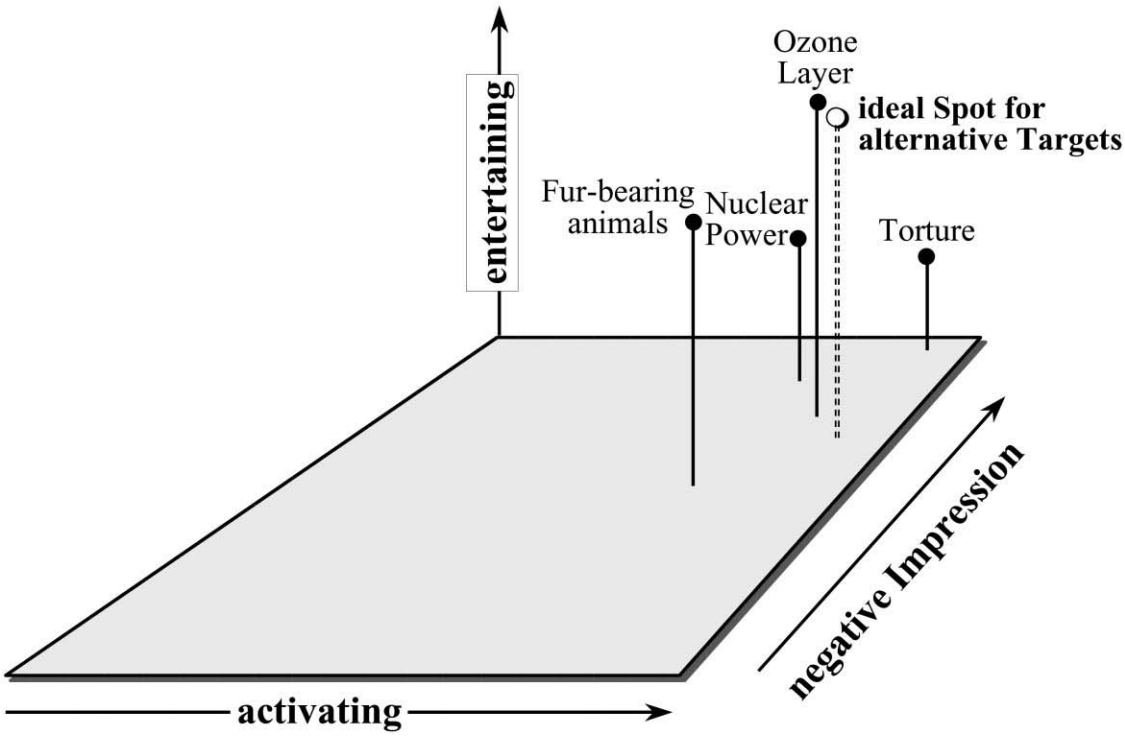

Figure 6
Factor analysis position of spots for alternative targets

greatly. By displaying extremely thin women in advertising, young girls and women could be in danger of eating disorders such as bulimia nervosa. In addition, some feel that women presented in advertising are degraded to object-status (only as objects of desire for men) and that this damages their sense of self-worth.

Critical discussions also took place regarding the assumption that advertising influences subconsciously with such weak stimulation that it is not even noticed consciously by the target individuals, but can still lead to specific consumer behavior. In the meantime, one knows through extensive research that this type of subliminal advertising has only a marginal if any behavioral effect. On the other hand, the accusation that advertising manipulates can be taken more seriously if manipulation is understood as an influencing technique which influences the target person. This would be the case if the advertising was carried out:

(a) Solely to the supplier's own advantage,

(b) Without considering the interest of the target person, and

(c) By using questionable methods which give the appearance that the target person is acting at their own free will.

Manipulative methods are typically recognized by using pictoral representations of socially unacceptable motives, whereby socially acceptable motives are displayed in writing. On the other hand, the demands made by consumer protection groups stating that advertising should only be informative are unrealistic, because they assume that humans are rational decision makers. Empirical research shows us, however, that not only rational decisions affect and influence buying behavior (Engel et al. 1992, Kroeber-Riel and Weinberg 1999).

Intense socially critical discussions have been carried out regarding the costs of advertising. In many organized industrial states with market economies the costs for advertising are higher than the costs for education. Critics say that this is a waste for society, because advertisement doesn't change the size of the cake that is to be divided up, but just changes the portions divided up by the competitors. This can be disproved, however, impulses for growth are generated by advertising.

See also: Advertising and Advertisements; Advertising: Effects; Consumer Psychology; Journalism

Bibliography

Engel J F, Blackwell R D, Miniard P W 1992 *Consumer Behaviour*. Dryden, London

Jones J P (ed.) 1998 *How Advertising Works. The Role of Research*. Sage, London
Kroeber-Riel W 1991 *Strategie und Technik der Werbung — Verhaltenswissenschaftliche Ansätze*. Kohlhammer, Stuttgart
Kroeber-Riel W, Weinberg P 1999 *Konsumentenverhalten*. Vahlen, Munich
Münsterberg H 1912 *Psychologie und Wirtschaftsleben. Ein Beitrag zur angewandten Experimental-Psychologie*. Barth, Leipzig, Germany
Neumann P 2000 *Markt- und Werbepsychologie, Band 2: Praxis*. Fachverlag Wirtschaftspsychologie, Gräfelfing
Percy L, Woodside A G 1983 *Advertising and Consumer Psychology*. Lexington Books, Lexington, MA
Rosenstiel L von, Neumann P 2001 *Einführung in die Markt- und Werbepsychologie*. Wissenschaftliche Buchgesellschaft, Darmstadt
Scott W D 1908 *The Psychology of Advertising*. Small, Maynard & Co., Boston, MA
Spiegel B 1970 *Werbepsychologische Untersuchungsmethoden*. Duncker & Humblot, Berlin
Zielske H A 1959 The remembering and forgetting of advertising. *Journal of Marketing* **23**: 239–43

L. von Rosenstiel and P. Neumann

Advocacy and Equity Planning

Until recently, most American city planners dealt solely with the physical city. They designed streets, parks, and boulevards, made plans for the way land was to be used in the community, and prepared regulations to control the use of land. Advocacy or equity planners are those professional planners who not only deal with the physical aspects of the community but who, in their day-to-day practice, also deliberately try to move resources, political power, and political participation toward the lower-income, disadvantaged populations of their cities. They are called 'advocacy' or 'equity planners' because they seek greater equity among different groups as a result of their work. Where the work of most city planners is rarely consciously redistributive, advocacy or equity planners often conceive the potential contribution of planning in broad economic and social terms and try to provide for a downward redistribution of resources and political participation in order to create a more just and democratic society.

Many observers place the birth of advocacy and equity planning in the decade of the 1960s when crowds were in the streets of American cities protesting the wholesale demolitions and displacements caused by urban renewal and the interstate highway program. These traumatic events and the anti-war and civil rights movements, which occurred at about the same time, challenged the belief in top-down planning by benign, value-free experts and created a demand for more social planning based on grassroots involvement. The events of the 1960s provided great support for advocacy and equity planning, but actually, an alternative planning practice oriented toward equity considerations had its roots in the turn of the twentieth century.

To begin this exploration into recent history, one must go back to the period between approximately 1880 and 1915, known as the Progressive Era, a time when the respectable urban bourgeois discovered the slum city festering beneath their urban world. At the time, America was rapidly changing from an agricultural to an industrial society. American farmers, forced off their homesteads, joined European immigrants flooding into industrial cities. The largest cities, as centers of manufacturing, exchange, and distribution, had grown most explosively without proper planning or the means to regulate growth. They became choked with slums built to house immigrant workers (Riis 1971). The slums, in cities like New York and Chicago, became such breeding grounds of disease, crime, and human misery that efforts at reform were introduced by civic and political leaders. It was this revulsion against the slum and the fear of revolutionary social unrest that brought housing reform and the social aspects of urbanism into modern urban planning. Progressive leaders believed that such dismal conditions could not wait on beneficent forces, but could be corrected by diligent and scientific health and housing policies.

The settlement house movement was one of their first efforts at neighborhood improvement. In the poor, immigrant neighborhoods of dozens of major cities, settlement houses like Hull House in Chicago, Henry Street Settlement in New York, and South End House in Boston, were established. These were mostly staffed by middle- or upper-class social workers who taught the English language and domestic arts to immigrants, lobbied city government for more neighborhood parks and other public facilities, and pressed for effective tenement housing reform. In the process, the settlement house workers hoped to improve housing and neighborhood conditions, and 'Americanize' the new immigrants. They were not urban planners but, then, advocacy for parks, better housing, and other improvements in the slums helped provide the reform underpinning for the nascent city planning profession.

Three typical progressive era reformers were Mary Kingsbury Simkhovitch, Benjamin C. Marsh, and Alice Constance Austin. Simkhovitch pursued a wide range of social justice issues, including women's suffrage, economic reform, and progressive politics. After establishing Greenwich House in New York City, she worked to improve housing density and building-code laws. During the 1930s, Simkhovitch performed her most important work when she helped draft the Wagner–Steagall Housing Act of 1937, providing federal participation in low-income housing for the first time in America.

Benjamin C. Marsh was the Secretary of the Committee on Congestion of Population, a prestigious committee formed in 1907 by 37 civic and philanthropic organizations to study, publicize, and promote programs to relieve the problems of excessive massing of people in New York City. Marsh was widely traveled and was strongly impressed by European city planning, especially by the ideas of the Englishman, Ebenezer Howard (Howard 1965). Howard, who loathed the industrial city with its filth and overcrowding, proposed a scheme of land development based on population dispersion into a regional pattern of small, self-contained cities. These garden cities would enjoy all the advantages of the core city including nearby jobs in industry, higher wages, and social opportunities, while also enjoying the benefits of the countryside with low rents, fresh air, agricultural gardens, and cooperative arrangements to maintain the land. A central feature of Howard's scheme included the common ownership of land so that the unearned increments in land values could be recaptured for the benefit of the entire community. Howard's sweeping planning ideas were enthusiastically adopted by Marsh, for whom city planning was a holy war against predatory forces, especially real-estate speculators. Marsh's energetic career as a skilled organizer and publicist of planning issues made the city planning movement better known nationally and more socially responsive. His small paperback book, *An Introduction to City Planning* (1909), set the stage for the First National Conference on City Planning and the Problems of Congestion convened in Washington, DC in 1909, which marks the formal birth of American city planning.

Alice Constance Austin also used Ebenezer Howard's new town ideas as examples of good city planning when she assumed the role of city planner and architect in 1915 for the partially built socialist city of Llano de1 Rio in California (Hayden 1976). Austin also learned from Patrick Geddes, a Scottish biologist, who drew up dozens of town plans in India and elsewhere, based on a cooperative model of city evolution. Austin's debt to Howard's garden city is reflected in her organization of the city of Llano del Rio, with its 'crystal palace'-like central buildings and its boulevards and street system. Her approach to building design was, however, distinctly feminist. Here, Austin proposed private gardens, but also communal kitchens and laundries to liberate women from drudgery.

Early American planners, most of whom were architects and engineers, believed that the way to bring the city under control was to reduce congestion, make physical development more attractive, and control the flow of traffic. They thought planning and zoning examples from Europe, especially from Germany and the UK, provided good direction. But, although early American planners used European models to define good planning, in one key respect the American approach diverged sharply from the European experience. In France, Germany, and the UK, the answer to the provision of low-income housing was government support; in the USA such support was decisively rejected. Lawrence Veiller, author of the New York City tenement-house legislation of 1901 who later founded the National Housing Association, maintained that only local government should concern itself with housing, but only to enforce local building regulations (Scott 1995). Otherwise, private benevolence could do the job. In Veiller's view, it was proper for local government to clear slums, but not to rehouse the displaced families.

A small group of visionaries rejected the physical determinism of most early planners. The Regional Planning Association of America (RPAA), formed in 1923, included such planning luminaries as Catherine Bauer, Stuart Chase, Benton MacKaye, Lewis Mumford, Clarence Stein, and Henry Wright. They believed in planning entire regions to achieve social objectives. Following the ideas of Howard and Geddes, RPAA members expounded their vision of small, self-sufficient communities scattered through regions in ecological balance with rich natural resources. In the distribution of electric power by regional grids and the speed of the automobile and truck, they saw new tools for rehabilitating declining urban neighborhoods and liberating large cities from congestion and waste.

The thinking and writing of RPAA's members was influential beyond their small numbers. Their ideas on regional planning and environmental conservation led in the 1930s to the creation of the Tennessee Valley Authority (TVA), the Civilian Conservation Corps, the 14-state Appalachian Trail and the regional studies of the National Resources Planning Board. RPAA member Benton MacKaye, who conceived the Appalachian Trail, was clearly one of the founders of the modern environmental movement.

RPAA's main practical experiment, however, was only a mixed success (Birch 1980). It was the garden city-inspired development of Radburn, in Fairlawn, NJ constructed in 1928 using Ebenezer Howard's ideas, but adjusted to American customs and laws. The elements of Radburn include the 'superblock,' the cul-de-sac and narrow loop lanes for residential traffic, the clustering of housing around large areas of parkland in common ownership, and the separation of vehicular and pedestrian traffic. Radburn also provided day care for working mothers and similar social services, as well as a community organization to administer commonly held land. Although Radburn has had an important influence on American planning thought and design, it failed to realize its sponsor's hopes of becoming a completely self-sustained community. It has now been swallowed up as part of northern New Jersey's amorphous sprawl.

By the time America emerged from World War I, the Progressive Era had lost much of its momentum. City planning, slum clearance and housing reform, which had been part of the upsurge of reform, now

suffered from a weakening of the liberal impulse. RPAA member Frederick L. Ackerman attacked the profession for no longer being concerned about 'the causes which give rise to the existing maladjustment' in cities. America had maladjusted communities, he said, because planners declined to interfere with 'the right of the individual to use the community as a machine for procuring individual profits and benefits, without regard to what happens to the community' (Ackerman 1919).

Urban planning and especially social or equity planning lagged during the 1920s with 'the business of America was business,' but the Great Depression and the New Deal administration of Franklin D. Roosevelt (FDR) brought both back. Roosevelt believed that the power of government should be used to restore misused land, harness wasted water, and revitalize despondent human beings. In his first term, FDR enacted legislation authorizing the widest application of planning yet proposed, including the TVA and the Resettlement Administration. By Executive Order, FDR also established the National Planning Board, later the National Resources Planning Board.

The National Resources Planning Board was America's first effort at national planning. The Board advised the Department of the Interior on regional planning, land and water conservation, and on issues of social insurance and poverty. As Secretary of the Interior, Harold L. Ickes, told city planners and members of the American Civic Association in 1933, 'long after the necessity for stimulating industry … . shall be a thing of the past, national planning will go on as a permanent Government institution.' However, the Board was resented by many old-line federal agencies like the Army Corps of Engineers who chafed under the Board's attempts to coordinate their projects, and by conservative Congressmen who feared that the agency was promoting foreign, socialist schemes. As a result, and to wide and dolorous lamentation among planners, Congress abolished the National Resources Planning Board in August, 1943.

The Resettlement Administration, established in 1935 under the leadership of Rexford G. Tugwell, suffered a similar fate. Tugwell, who once proposed that the (US) Constitution be changed to include planning as a fourth power of government, immediately embarked on an effort to improve living conditions for working people in metropolitan areas. He seized on the idea of the English suburban garden village, but adapted the concept to the American situation. The towns built by the Resettlement Administration were called Greenbelt Towns. Originally, Tugwell's staff identified eight metropolitan areas where new towns were to be built, but funding limitations narrowed the choices to three areas: Greendale, Wisconsin near Milwaukee; Greenhills, Ohio near Cincinnati; and Greenbelt, Maryland near Washington, DC. The three towns, all examples of high-quality planning, featured small homes in a well-planned garden setting and housed about 2,100 families in all. The distinctive feature of all three of the towns was the surrounding greenbelt protecting the towns from outside encroachment. The residents, largely industrial workers, were provided an opportunity to live in a low-cost, but far superior environment than they were used to. As a demonstration, the Greenbelt Towns made clear that superior alternatives existed to haphazard speculative sprawl. Unfortunately, before the Resettlement Administration could expand the three towns and build two others that were on the drafting table, Congress abolished the agency in June, 1938. It was left to post World War II British planners to build a substantial new towns program, while the USA had to content itself with the three Greenbelt towns and a few commercial initiatives including Reston, Virginia, and Columbia, Maryland.

Roosevelt's New Deal also produced the first major intervention in the field of low-rent housing. Pushed vigorously by RPAA member Catherine Bauer and others, the administration passed into law the Housing Act of 1937 which provided low-rent housing for the deserving working poor. For the first time, the federal government would provide support for the capital costs of public housing construction; tenant rents were to take care of subsequent operating and maintenance costs. Later, in 1949 and 1954, Congress returned to the public housing issue and emerged with a commitment to more low-income housing. But Congress also made a larger commitment to urban renewal, which ultimately demolished more low-income housing close to the city core than was built. The scale of demolition and forced relocation was large; a study by the National Association of Home Builders estimated that the total housing demolition by all public programs between 1950 and 1968 amounted to 2.38 million units. These units were disproportionately the homes of poor and near-poor black families.

Ultimately, a revolt broke out against these excesses. Within an astonishingly short time, this revolt caused an almost complete inversion of almost every basic value in American planning practice. Influential critics Jane Jacobs (1961) and Herbert Gans (1962) spoke for the preservation of older urban neighborhoods, as they pointed out that planners implementing the urban renewal program were thoughtlessly destroying valuable housing and irreplaceable social networks while providing opportunities for profitable real estate investments. The opposition to urban renewal coincided with the civil rights movement, and racial riots tore through the cities revealing just how little the planning process had done for the poor. Opposition to the war in Vietnam, and with it the Pentagon style of planning by top-down, bureaucratic experts, was at its peak. There was now a deep distrust of professional expertise and there was a demand for advocacy and equity planning based on grass-roots involvement.

Paul Davidoff, a lawyer, planner, and educator, made the most substantial contribution to the concept of advocacy planning (Davidoff 1965). He argued that advocacy could reinvigorate planning in three ways: by broadening public debate and participation, by sharpening the skills of planners who would have to defend their choices, and by shifting the focus of planning from the purely physical, to social and economic priorities. The essence of advocacy planning is the encouragement of alternative plans by all groups holding special values about the future of their communities. Advocacy planning would supplement the one, official city plan with alternative plans by Liberals, Conservatives, Democrats, and Republicans and other groups, but the emphasis would be on professional planning services for the poor, the black, and the underprivileged. Using legal analogies, the merits of these alternative plans were to be debated and the best plan would emerge from the debate. Davidoff's ideas were taken up by planning practitioners and educators with a strong cumulative effect.

During the 1960s advocacy planning organizations serving minorities, poor, and working-class whites were formed in many American cities (Heskin 1980). Exemplars were: Architects Renewal Committee of Harlem in New York City (1964); Urban Planning Aids in Boston (1967); and the Community Design Center in San Francisco (1967). The effectiveness of these grassroots organizations was limited given their inadequate resources and often defensive strategies, but they signaled a new diversity. A national organization called Planners for Equal Opportunity was established in the 1960s and in 1975, under the leadership of advocate planner Chester Hartman, continued as Planner's Network. By 1999, Planner's Network represented over 800 planners and academics concerned with social and economic justice. By the 1970s the American Planning Association, whose journal carried virtually no discussion of racial issues prior to 1970, had modified its Code of Ethics to reflect concern for vulnerable populations: 'A planner shall seek to expand choice and opportunity for all persons, recognizing a special responsibility to plan for the needs of disadvantaged groups and persons …'

Davidoff's writings on choice theory were important intellectual contributions to planning theory and they make him the pre-eminent symbol of progressive planning. They make perhaps the most persuasive argument on how a planner might reconcile professionalism with political engagement.

But Davidoff's substantial contributions to planning theory and practice were matched by his efforts at institutional innovation. He realized that from 1950 onward, the overwhelming preponderance of growth in population, jobs, and economic investment was not in America's cities but was in the suburbs. Because of racial discrimination and exclusionary land use controls, blacks were being shut away from the benefits of suburban growth. In response, Davidoff founded Suburban Action, a nonprofit institute for research, litigation, and advocate planning services within metropolitan regions (Davidoff et al. 1970). Suburban Action used its resources to legally challenge restrictive zoning and land use controls in the suburbs and enlarge suburban opportunities for the black and poor.

Davidoff's special concerns for the disadvantaged were also adopted in some official city planning agencies within local government. One of these was in Cleveland, Ohio, where planning director Norman Krumholz and a core staff of progressive planners developed a system of values and strategies in the 1970s which came to be called 'equity planning.' The central theme adopted by the Cleveland planners was contained in this statement: 'In the context of limited resources, first and priority attention should be given to the task of promoting wider choices for those Cleveland residents who have few if any choices' (Krumholz et al. 1975, Krumholz 1982). This goal made clear that Cleveland's poor and near-poor were to receive priority planning attention.

The Cleveland planners justified the choice of their goal with three arguments. First, they argued that a long-standing, historic moral commitment existed to seek more equity in the social, economic, and political relations among people. Second, building on the ideas of philosopher John Rawls (1971), they used reason as a means of justifying a more equitable society—the kind of society that free, equal, and rational people would establish to protect their own self-interests. Finally, they justified their goal by reality: making explicit the imbalances in income, education, health, and other social and economic variables that existed in Cleveland between city and suburb, and white and black citizens (Cleveland City Planning Commission 1975). Over the ten-year length of the equity planning experiment, and under three different mayors, the efforts of Cleveland's planners resulted in 'fair-share' low-income housing distribution plans for Cuyahoga County, progressive changes in Ohio's property law, improvements in public service delivery, enhancement of transit services for the transit-dependent population, the rescue of lakefront parklands, and many other improvements.

Another outstanding example of progressive politics and equity planning within city government was found in Chicago during the 1980s. Robert Mier, a planning professor who founded the Center for Urban Economic Development at the University of Illinois in Chicago, helped build a political coalition that in 1982 elected Harold Washington, Chicago's first black mayor. When elected, Washington took Mier and some of his associates into City Hall where they explicitly included 'redistributive and social justice goals within the governments's policy, planning and implementation frameworks' (Giloth and Moe 1999).

Mier and his fellow planners wrote the 'Chicago Economic Development Plan, 1984,' a model of equity planning. The plan proposed to use the full weight of the city's tax incentives, public financing, and infrastructure improvements to generate jobs for Chicago residents with emphasis on the unemployed. Specific hiring targets were set for minority and female employment; 60 percent of the city's purchasing was directed to Chicago businesses; 25 percent of that was to go to minority and women-owned firms.

The plan also sought to encourage a model of balanced, 'linked' growth between downtown Chicago and the city's neighborhoods. It offered public support to private developers interested in building projects in 'strong' market areas of the city, only if they would agree to contribute to a low-income housing trust fund or otherwise assist neighborhood-based community development corporations to build projects in 'weaker' areas.

Other American urban planners have adopted advocacy/equity approaches because they believe that planning along these lines holds the promise of better lives for the most troubled residents of their cities. In the 1980s and 1990s equity planning cases were documented in such cities as Denver, Jersey City, San Diego, Berkeley, and Santa Monica (Krumholz and Clave 1994). These cases make clear that there are often political and institutional barriers to equity planning practice, especially in a nation so strongly driven by market forces as the USA, and equity planners who question the status quo may face political reprisals. But obstacles to equity planning practice seem to lie primarily in the areas of the planner's personal confidence, motivation, and will.

A number of lessons seem clear. First, all forms of urban planning, especially advocacy/equity planning, prosper during periods of crisis which bring forward the reform elements of the American state. It was during the Progressive Era that housing reform and urban planning were initiated. Later, when the Great Depression brought forth FDR's New Deal, planning was introduced for the first time into the federal government, sweeping regional projects like TVA got their start and the federal government involved itself in the provision of low-income housing. During the 1960s, the ferment around civil rights, anti-war protests, and the demolitions of urban renewal and the interstate highway program, brought advocacy planners like Paul Davidoff, who combined commitment to social justice and democracy, to the fore.

Although the advocacy/equity planners often struggled against the more powerful currents of a strongly market-oriented society, they succeeded to an extent, and their work has been acknowledged and institutionalized. At the end of the twentieth century, American city planning was much more sensitive and collaborative than it had been earlier. Racial discrimination in housing and in mortgage lending has been prohibited by federal law, and citizens were encouraged to participate in the planning of programs that impacted their lives. Neighborhood-based development corporations have been successfully redeveloping parts of certain older neighborhoods, and some cities have been linking the benefits of downtown growth to their poor neighborhoods.

What remains troublesome, however, is the persistence of deep poverty in a growing number of urban neighborhoods; the continuance of patterns of racial segregation; the widening gap between rich and poor; the dismantling by conservative national administrations of the social safety net; the chronic competition among American cities that harms the vast majority of their citizens; and the growing spatial separation across metropolitan regions by race and class.

Given these challenges, a new generation of advocacy/equity planners is needed to help ease and resolve the most crucial of these social and physical issues, for what is at stake is nothing less than the future of urban life in America.

See also: Community Aesthetics; Community Economic Development; Development and Urbanization; Local Economic Development; Multi-attribute Decision Making in Urban Studies; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Public Goods: International; Real Estate Development; Strategic Planning; Urban Growth Models; Urban Life and Health

Bibliography

Ackerman F L 1919 Where goes the planning movement. *Journal of the American Institute of Architects* **1919**: 519–20

Birch E L 1980 Radburn and the American planning movement. *Journal of the American Planning Association* **46**(4): 424–39

Cleveland City Planning Commission 1975 *The Cleveland Policy Planning Report*.

Davidoff P 1965 Advocacy and pluralism in planning. *Journal of the American Institute of Planners* **31**

Davidoff P, Davidoff L, Gold N 1970 Suburban action: Advocate planning for an open society. *Journal of the American Institute of Planners* **361**: 12–21

Gans H 1962 *The Urban Villagers*. Free Press, Blencoe, New York

Giloth R, Moe K 1999 Jobs, equity and the mayoral administration of Harold Washington in Chicago. *Policy Studies Journal* **27**(1): 129–46

Hayden D 1976 *Seven American Utopias: The Architecture of Communitarian Socialism*. MIT Press, Cambridge, MA

Heskin A D 1980 Crisis and response: A historical perspective on advocacy planning. *Journal of the American Planning Association* **46**(1): 50–62

Howard E 1965 *Garden Cities of Tomorrow*. MIT Press, Cambridge, MA

Jacobs J 1961 *The Death and Life of Great American Cities*. Random House, New York

Krumholz N, Cogger J M, Linner J H 1975 The Cleveland

policy planning report. *Journal of the American Institute of Planners* **41**(5): 298–304
Krumholz N 1982 A retrospective view of equity planning: Cleveland 1969–1979. *Journal of the American Planning Association* **48**: 163–83
Krumholz N, Clavel P 1994 *Reinventing Cities: Equity Planners Tell Their Stories*. Temple University Press, Philadelphia, PA
Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA
Riis J A 1971 *How The Other Half Lives: Studies Among the Tenements of New York*. Dover, New York
Scott M 1995 *American City Planning Since 1890*. APA Press, Chicago 258

N. Krumholz

Advocacy in Anthropology

1. Definition, Scope, and Aims

Advocacy is a variety of applied anthropology advancing the interests of a community, often as a practical plea on its behalf to one or more external agencies (Paine 1985, Wright 1988). The community is usually indigenes, peasants, an ethnic minority, or refugees—those who are among the most oppressed, exploited, and abused. Advocacy is often connected to human rights, a framework internationally accepted in principle if not always followed in practice (Messer 1993). Advocacy encompasses a broad agenda for social and political activism—promoting cultural survival and identity, empowerment, self-determination, human rights, economy, and the quality of life of communities. Advocates reject the supposed neutrality of science and adopt a stance on some problem or issue to improve the situation of a community, ideally in close collaboration with it. Thus, as advocate the anthropologist is no longer just observer, recorder, and interpreter (basic research), nor consultant to an external agency (applied), but facilitator, interventionist, lobbyist, or activist for a community (advocacy).

2. Pivotal Position

Advocacy usually operates from an idealist rather than realist position, although these terms are problematic. Realists accept cultural assimilation and even ethnocide (forced cultural change or extinction) as a natural and inevitable correlate of 'civilization,' 'progress,' and 'development,' a position sometimes linked with social Darwinism. Idealists reject this, viewing ethnocide as a political decision, usually by a government violating human rights. Realists dismiss idealists as romantic, and/or trying to preserve indigenes as private laboratories, zoos, or museums. However, idealists do not argue that culture is static and should be preserved as such, only for self-determination by people to promote their cultural survival, identity, welfare, and rights. The moral and political issue is the right and power of the state to dominate indigenes or others and to implement ethnocide. Advocacy attempts to intervene in this asymmetry through demystifying the process, exposing injustices, and offering political resistance (Bodley 1999).

3. Historical Sketch

Advocacy has a long history. Bartolome de las Casas (1474–1566)—theologian, missionary, and something of a historian and anthropologist—participated in the first decades of European colonialism in the Americas. He chronicled injustice against indigenes and argued in their defense. By the mid-nineteenth century, the Anti-Slavery Society and the Aboriginal Protection Society emerged in the UK as humanitarian organizations campaigning for just policies toward indigenes and others. During this period, the early anthropological societies of London and Paris added similar concerns. In the 1960s, organizations focused on advocacy developed, most notably Cultural Survival in Cambridge, Massachusetts; the International Work Group for Indigenous Affairs in Copenhagen; and the Minority Rights Group and Survival International in London. Each publishes newsletters, journals, and/or documents to expose human rights violations and analyze issues; funds collaborative research with communities; and works to influence national and international governmental and nongovernmental agencies and the public, especially through lobbying, the media, and letter writing (e.g., Solo 1992).

Most of the history of advocacy as well as recent cases and issues can be found in the publications of these advocacy organizations (e.g., *Cultural Survival Quarterly*) and in the journals *Current Anthropology* (especially December 1968, December 1973, June 1990, June 1995, February 1996); *Human Organization* (especially January 1958, Winter 1971) and its predecessor *Applied Anthropology*; and *Practicing Anthropology*. This literature can greatly enrich anthropology courses.

4. Some Cases

Two pioneering projects in action anthropology stand out: the Fox and Vicos projects. These became models for subsequent initiatives worldwide. Action is based on the premise that if a community is adequately informed of the alternatives for change, then it will try to choose what is best. Both action and advocacy pivot on self-determination (see Castile 1975). From 1948 to 1959, Sol Tax and his students from the University of Chicago developed the Fox Project to promote the self-determination of some 600

Mesquakies in Tama, Iowa. The Mesquakies, commonly called Fox Indians, faced cultural extinction. The anthropologists facilitated the identification by the community of their needs, problems, and alternatives (Stanley 1996). From 1952 to 1957, Allan Holmberg of Cornell University led a team, including colleagues from the Indigenous Institute of Peru, in action anthropology for the Quechua community of Vicos in the Andes. About 2250 people gained significant freedom from centuries of oppression, exploitation, and abuse as the project provided technical, economic, and other assistance for community development and a cooperative (Dobyns et al. 1971). Advocacy long predates action anthropology, although the term advocacy is more recent. Action focuses on community directed cultural change and development, advocacy on pleading the case of a community to a government or other agency, especially about human rights violations. However, often advocacy and action are intermeshed.

An exemplary case from the mid-1990s is the Ye'kuana Self-Demarcation Project. Some 4000 Ye'kuana are scattered in 30 communities in the Venezuelan Amazon. Despite factions created by Catholic and Protestant missionization, the Ye'kuana united in this project. They are documenting their history, settlement pattern, resource and land use, and other aspects of their culture for legal title to ancestral land from the government. Nelly Arvelo-Jimenez (Venezuelan Institute for Scientific Investigations and Otro future) and other outsiders provided assistance including the Global Positioning System to help map Ye'kuana territory. Mapping was financed by a grant from the Canadian government through the Assembly of First Nations, the Canadian indigenous organization (Arvelo-Jimenez and Conn 1995).

In such ways advocacy contributes to indigenes, ethnic minorities, and others economic, technical, health, legal, and political assistance as well as helping raise their media savvy, cultural consciousness, and hope. Also it helps promote the global movement of pan-indigenous identity (e.g., Lurie 1999).

5. *Criticisms and Responses*

At least two criticisms contributed to the development of advocacy: first, accusations of genocide and ethnocide of indigenes by colonials and neo-colonials in frontiers like the Amazon (Bodley 1999); and second, from outside and within anthropology criticisms of basic and applied research together with calls for increasing social responsibility and relevance (Biolsi and Zimmerman 1997, Hymes 1972). In the late 1960s, Vine Deloria, Jr., a Sioux lawyer and author, launched a searing critique of anthropologists engaged in either pure research with natives as objects in their private zoo, or applied work for the colonial government, only concerned with advancing their career for status, prestige, and money; and irresponsible and unresponsive to the needs and problems of indigenes. He asserted that anthropologists should obtain informed permission from the host community for research, plan and implement it in close collaboration with them, and focus on their practical needs, problems, and concerns (Biolsi and Zimmerman 1997).

In 1971, a historically important but neglected conference of mostly Latin American anthropologists developed the Declaration of Barbados which, among other things, criticized anthropology for its scientism, hypocrisy, opportunism, and apathy in the face of the oppression, exploitation, ethnocide, and genocide of indigenes by colonials and neo-colonials. For some anthropologists this became a manifesto to join the struggles for liberation and self-determination of indigenes and ethnic minorities through advocacy (Dostal 1972).

There have also been criticisms of advocacy from within the profession, mainly for supposedly abandoning scientific objectivity and reducing or abandoning anthropology to some form of social work or political action. For instance, Elsass (1992) asserts that anthropology rests on criteria of science, objectivity through neutrality, and scholarship for the creation of knowledge; and advocacy on morality and the use of knowledge. He thinks that anthropologists are ill equipped to deal with matters such as the politics of state penetration into indigenous areas. He worries that advocacy may make things worse in the community and jeopardize the credibility and prestige of basic research. Elsass believes that fieldwork, moral commitment, a sense of justice, political observation, and anger on behalf of the community, are all part of the decision to advocate, but that anthropology as science and scholarship does not lead to advocacy (cf. Castile 1975).

Such critics fail to realize four things. First, in a general sense all anthropologists are advocates in some degree and manner. At least since Franz Boas (1858–1942), teaching anthropology advocates the value of the profession and cultural diversity; research publications advocate certain arguments, theories, and methods; and both may challenge racism and ethnocentrism in favor of valuing equality.

Second, no scientist or science is apolitical and amoral; indeed, even the decision not to act involves politics and morality. When ethnocide, genocide, or human rights violations occur, it is simply unprofessional and unconscionable for a knowledgeable anthropologist not to act. This is implicit in the code of professional ethics of organizations like the American Anthropological Association, numerous resolutions at its annual meetings over many decades, and its Committee for Human Rights.

Third, many communities believe that either the anthropologist is part of the solution or part of the problem (Biolsi and Zimmerman 1997). Increasingly host communities are excluding anthropologists unless

they demonstrate social responsibility and relevance. Indeed, if anthropology is not of some relevance to the communities from which research is derived, then some would suspect its credibility. Even when a community can readily speak for itself, it usually helps to have an outsider with some special knowledge speak as well. For instance, this transpired in the USA with numerous cases of land claims by Native Americans in which anthropologists served as expert witnesses in court.

Fourth, many anthropologists are involved in advocacy because they are sincerely concerned with applying knowledge on behalf of the communities who are indispensable for their research, as an expression of genuine reciprocity, and to avoid dehumanizing their hosts and themselves.

6. *Future*

In the future advocacy needs to more explicitly and systematically develop its foundations and operations in terms of its history, philosophy, theory, methods, ethics, practice, and politics. There are many ways to contribute to advocacy, even for those who eschew its complexities, difficulties, and risks in fieldwork (cf. Mahmood 1996). To be effective advocacy must be grounded in basic research, but it also feeds into theory. For example, advocacy should help anthropologists continue critical reflection on fundamental issues such as these problematic dichotomies: science/humanism, objectivity/subjectivity, fact/value, observer/participant, theory/practice, basic/applied, inaction/action, powerful/powerless, modernist/traditionalist, realist/idealist, and universalism/relativism.

Advocacy, action, and other varieties of applied anthropology are most likely to increase in the twenty-first century because of at least three factors: first, growing population and economic pressures on land and resources with ensuing conflicts, violence, and rights violations; second, increasing encroachment of the state, military, business, industry, and other forces into 'undeveloped' zones; and third, insistence by local communities that anthropologists be more responsible and relevant. For example, by the 1990s, most anthropologists working with the Yanomami in the Amazon between Brazil and Venezuela voluntarily shifted their emphasis from salvage ethnography (traditional culture) to advocacy, especially with epidemics and other serious problems from the mining invasion (Ramos 1999). Advocacy is likely to continue developing well into the future as a significant component of the conscience of anthropology. As Sol Tax said, 'If there is something useful I can do, then I have to do it' (Stanley 1996, p. 137).

See also: Advocacy and Equity Planning; Colonialism, Anthropology of; Conflict: Anthropological Aspects; Cultural Policy: Outsider Art; Cultural Relativism, Anthropology of; Development: Social-anthropological Aspects; Fourth World; Frontiers in History; Genocide: Anthropological Aspects; Globalization and Health; Globalization, Anthropology of; Human Rights, Anthropology of; Imperialism, History of; Land Tenure; Peace and Nonviolence: Anthropological Aspects; Refugees in Anthropology; State: Anthropological Aspects; Third World; Violence in Anthropology; War: Anthropological Aspects

Bibliography

Arvelo-Jimenez N, Conn K 1995 The Ye'kuana self-demarcation process. *Cultural Survival Quarterly* **18**(4): 40–2

Biolsi T, Zimmerman L J (eds.) 1997 *Indians and Anthropologists: Vine DeLoria, Jr. and the Critique of Anthropology*. University of Arizona Press, Tucson, AZ

Bodley J H 1999 *Victims of Progress*. Mayfield Publishing Co., Mountain View, CA

Castile G P 1975 An unethical ethic: Self-determination and the anthropological conscience. *Human Organization* **34**(1): 35–40

Dobyns H F, Doughty P L, Lasswell H D (eds.) 1971 *Peasants, Power, and Applied Social Change: Vicos as a Model*. Sage Publications, Beverly Hills, CA

Dostal W (ed.) 1972 *The Situations of the Indians of South America*. World Council of Churches, Geneva, Switzerland

Elsass P 1992 *Strategies for Survival: The Psychology of Cultural Resilience in Ethnic Minorities*. New York University Press, NY

Hymes D (ed.) 1972 *Reinventing Anthropology*. Random House, New York

Lurie N O 1999 Sol Tax and tribal sovereignty. *Human Organization* **58**(1): 108–17

Mahmood C K 1996 Asylum, violence, and the limits of advocacy. *Human Organization* **55**(4): 493–8

Messer E 1993 Anthropology and human rights. *Annual Review of Anthropology* **22**: 221–49

Paine R (ed.) 1985 *Advocacy and Anthropology*. Institute of Social and Economic Research, Memorial University, St. John's, Newfoundland, Canada

Ramos A R 1998 *Indigenism: Ethnic Politics in Brazil*. University of Wisconsin Press, Madison, WI

Solo P (ed.) 1992 At the threshold: An action guide for cultural survival. *Cultural Survival Quarterly* **16**: 1–80

Stanley S 1996 Community, action, and continuity: A narrative vita of Sol Tax. *Current Anthropology* **37**(Suppl.): 131–7

Wright R 1988 Anthropological presuppositions of indigenous affairs. *Annual Review of Anthropology* **17**: 365–90

L. E. Sponsel

Aesthetic Education

Having no fixed meaning, the phrase 'aesthetic education' may connote (a) a program of studies intended to develop dispositions to regard things from an

aesthetic point of view, (b) an emphasis on response to art in contrast to its creation, (c) concentration on the common features or the interrelatedness of the arts, (d) the cultivation of sensibility generally, not just in the arts, and (e) a special role for aesthetics as both content and method of inquiry. Given the multiple meanings of the term, definitions of aesthetic education are properly regarded as programmatic interpretations intended to convey desirable end states. Since using 'aesthetic education' as a blanket term to cover all possible relations between the arts and education would necessitate presenting a complete history of the subject, a framework must be imposed that will elicit some of the major themes of modern and contemporary thinking. The scheme discusses theorists in whose writings the cultivation of aesthetic experience plays a key role.

1. Three Generative Thinkers: Schiller, Read, and Dewey

An appropriate starting point are the writings of Friedrich Schiller, an eighteenth-century German dramatic poet and philosopher whose *On the Aesthetic Education of Man in a Series of Letters* (1793–95) (Schiller 1967) is significant for thinking about the role of a man of letters in culture, the function of art in personal and social life, and the wholeness of aesthetic experience. Problems that concerned Schiller were also addressed in different ways by such influential twentieth-century theorists as Sir Herbert Read and John Dewey. All three writers were variously preoccupied with the harmful consequences of political and social dislocation, the alienation inherent in modern productive processes and institutional arrangements, reductionism in values, and disruption of the continuity of nature and human experience. For Schiller, the Reign of Terror of the French Revolution provided the impetus for formulating in the *Letters* an idea of aesthetic education as productive of a humane and democratic society. For Read it was the advent of industrialization and the alienation of the proletariat that prompted his recommending a pedagogy capable of reuniting in human experience what modern life and production methods had sundered. For Dewey the notion of consummatory experience was a response to concerns similar to Schiller's and Read's. He further believed that art, broadly defined as experience, was important not just for the personal satisfaction it provided but for the restoration of a greater sense of community.

1.1 Schiller (1759–1805)

Schiller's career unfolded during the turbulent modern era when the power of the State and privileged classes was coming under attack in the name of Enlightenment principles of reason, freedom, and democracy. Many theorists of the time believed history was evolving in a direction that would provide individuals with greater freedom and control over their lives. Schiller admired the ideals and promise of the French Revolution, but, abhorring most forms of violence, he was dismayed at its cruelty and concluded that Man was not yet prepared for freedom. Consequently, he believed that the proper constitution of the State must be preceded by the proper constitution of individuals themselves. This transformation was to be aided by recourse to the discipline of aesthetics (newly established by Baumgarten) and the philosophical writings of Kant, as well as through energies drawn from Schiller's close association with Goethe—not to mention his own considerable strengths as a dramatic playwright. The aim of the *Letters* was to release in Man what Schiller called the living springs of human life, that is, qualities of life essentially manifested in experiences of Beauty. Such experiences would produce a healthy confluence of conflicting human impulses (the sensuous and the formal impulses) by giving free rein to a third impulse—the play impulse. The experience of Beauty, in other words, was a necessary precondition for the emergence of full humanity. Schiller found the play impulse ideally exemplified in artists' integrations of form and content in great works of art. Although he prescribed no particular curriculum or pedagogy, he was persuaded that the fostering of aesthetic culture was the required next phase in the evolution of civilization. The aesthetic path must be taken, he said, 'because it is through beauty that, man makes his way to freedom' (Schiller 1967, p. 9).

Subsequent philosophic analysis questioned Schiller's metaphysics and psychology as well as his extraordinary faith in aesthetic education's ability to advance the cause of human freedom and morality. But the inspirational force of Schiller's message was not lost on writers who have either appealed directly to his belief in the civilizing power of the arts or expressed affinity for the value he placed on art's role in the integration of the human personality. Above all, Schiller provided a significant justification for aesthetic education—the promotion of aesthetic culture—and described its potential for achieving increased social and political stability. Schiller's dated psychology notwithstanding, his discussion of the need to harmonize conflicting human drives has been of continuing interest to later theorists.

1.2 Read (1893–1968)

A poet, critic, art historian, editor, philosopher, pacifist, anarchist, and educational theorist, Read is perhaps best characterized as a humanist who was immersed in the art, culture, and politics of his time. As a humanist he was at odds with the received

cultural, intellectual, and educational traditions he believed inhibited the full realization of individuals' potentialities. And he was appalled by living and working conditions in the burgeoning factory towns of the industrial revolution. The specialization and division of labor consequent upon the triumph of technical rationality were fracturing the sense of community that had pervaded his rural upbringing. His experiences in World War I as well as his early literary training had endowed him with a poetic sensibility reminiscent of Schiller's. In his educational writings Read appeals directly to Schiller, and he evokes echoes of him when he both characterizes the kind of education that might meliorate the effects of dehumanization and prescribes a fitting instrument, model, and method for it—art and aesthetic education. But there were differences.

The materials from which Read composed his theory of education were largely of his own time. If the ideas of Marx, Morris, and Ruskin influenced his social analysis, the psychoanalytical theories of Freud and Jung, especially their notions about the structure and dynamics of the unconscious, played a significant part in his thinking about the nature of the artistic process and aesthetic education. The operations of the unconscious being essentially sensuous and sexual in character—Read called his philosophy of education a salutation to Eros—they stood in opposition to the constraints placed on human behavior by religious and moral codes. He believed that dipping into the unconscious, especially into its potent image-making powers, opened the way to greater self-realization. The crucible of the unconscious—a cauldron of memory images, feelings, and inherited attributes called archetypes—supplied source material for the creative imagination. Whatever hindered access to unconscious processes was to be discouraged, for only by utilizing them as resources could the individual benefit from their creative energies. Since Read thought that modern artists were particularly adept at exploiting the unconscious, he devoted a major portion of his career to championing their efforts.

Read's thought and career contain several complexities that cannot be dealt with here. Suffice it to say that from his social and political philosophy, conception of psychological processes, and interpretation of modern art, it was but a natural step to an educational aesthetics aimed at liberating mind and sensibility from the repressive tendencies of contemporary life and schooling. In contrast to Schiller's emphasis on studying the great works of the tradition, Read's pedagogy demoted the art object in the belief that conventional modes of art appreciation encouraged passiveness on the part of the learner and perpetuated a conception of knowledge as inert. What was wanted instead was what Dewey called learning by doing. Read also envisioned aesthetic education broadly enough to encompass the creation of a more aesthetically satisfying environment, which meant paying greater attention to the arts of everyday life. It comes as no surprise then that Read favored a pedagogy grounded in process, one that stressed the creative self-expression of the child. Given the idiosyncrasies of learners and the dispositions of teachers, the method of aesthetic education Read often referred to was less a single procedure than a collection of practices—whatever seemed to work for a particular student or situation was acceptable.

Read's impact was literally global, as witnessed by his influence on the International Society for Education through Art, which periodically confers an award in his name. But as in Schiller's case, it was the spirit of Read's message that counted more than his theoretical formulations. Few teachers were prepared to comprehend the intricacies of his major theoretical work *Education Through Art* (Read 1956). *The Redemption of the Robot* (Read 1966), in which Read recalls his encounters with education through art, is a more helpful introduction to his thought.

1.3 Dewey (1859–1952)

Dewey's roots and preoccupations resemble Read's: early childhood in a rural environment, concern about dislocations wrought by social change, impatience with educational traditions and institutions hostile to reform, compatible pedagogical ideas, and a belief in art as both a means of personal satisfaction and an instrument for reconstructing experience.

Dewey's distaste for separations and dualisms of all kinds and his almost religious feeling for the unity of human experience reflect the strong influence of Hegel, though in the course of evolving a naturalistic empiricism Dewey abandoned Hegel's metaphysics in favor of a Darwinian bio-social conception of human development. Dewey conceived experience as the interaction between organism and environment, as a doing and undergoing. Among the numerous dichotomies that worried Dewey (1958), of particular relevance to this discussion, was the separation of art from everyday life as epitomized in the museum conception of art. The theoretical task of reintegrating art into common experience that Dewey set himself required, first, defining everyday experiences in a way that revealed their inherently dramatic character and, second, claiming that whenever experience possesses certain features or qualities it may be spoken of as *an* experience—or *art*. This meant that all forms of inquiry and experience—intellectual, social, political, and practical—could under certain conditions qualify as art. Even the activities of the world of work, ordinarily characterized by a dehumanizing disjunction between means and ends, might aspire to the status of art.

Dewey tended to alternate, somewhat confusingly and not without unhappy consequences for his aesthetic theory, between two views of art and the work of

art. The first was art as the quality of *an* experience regardless of context. The second was art as commonly understood, that is, as a physical object. What is important is that in both interpretations Dewey thought art capable of endowing life with consummatory value and hence of contributing to his effort to find a philosophical justification for the reconstruction of human experience. Such reconstruction would be one of the preconditions for social reform—for Dewey was nothing if not a reformer—and was to be brought about within a framework based on liberal, democratic principles. Being the outgrowth of his theories, Dewey's educational recommendations and experiments—for example, at the University of Chicago—stressed continuity between the activities of school and society and placed reliance on learning by doing, with an emphasis on the designing of problem-solving situations.

2. *Further Developments*

Commentators have observed that subsequent theories of aesthetic education reflect attempts to escape the shadow of Dewey. This is true but requires some qualification. The educational philosophy of Arnstine (1967) is rooted in Dewey's concept of experience in that Arnstine equates learning with aesthetic experience. Dewey's analysis of qualitative thinking also influenced Eisner's (1991) notions of qualitative intelligence and educational criticism, and Dewey's pedagogy still enjoys favor among many theorists and teachers of art. His broad conception of the aesthetic is perpetuated in Howard's (1992) discussion of the role of sensibility in human life generally. Beardsley, in his several writings on the topic (Beardsley 1982), attempted to preserve what is valuable in Dewey's characterization of aesthetic experience while clearing up some of its ambiguities.

Although contemporary theorists of aesthetic education acknowledge that the compass of the field extends beyond the study of artworks, they still tend to emphasize the study of the fine arts. Another thread running through much current thinking is the idea that aesthetic experience should be cultivated for the sake of a variety of values. Broudy (1994) argues that aesthetic experiences of works of serious art serve individuals well in their quest for the good life and help them build a rich imagic store that tacitly informs their interpretations not only of artworks but of other phenomena as well. Greene's (1981) understanding of aesthetic literacy relies on the propensity of aesthetic experience to sharpen perception, expand the imagination, create a sense of freedom, and deflate stereotypes. Kaelin (1989) attributes similar benefits to art and describes how aesthetic education contributes to the effective operations of the art world, an institution that functions as guardian of aesthetic value. Since aesthetic situations are presumed to encourage open-mindedness and tolerance, aesthetic education may also help produce personality traits valued by democracies. In his interpretation of aesthetic education from a humanities point of view, Smith (1989) highlights the constitutive and revelatory values of art, the humanizing potential of aesthetic experiences, and the capacity of aesthetic studies to refine discrimination, stretch the imagination, and provide ideals for human life. Aesthetic education further prepares the young to traverse the world of art with sensitivity and percipience. Swanger (1990), who places emphasis on creative activities, thinks art's radical and destabilizing powers derives from its freshness, creativity, efficacy in the transfer of learning, and promise for transforming a materialistic consumer society into one more protective of the environment. The volume by Parsons and Blocker (1993) is noteworthy for featuring a cognitive developmental theory of aesthetic experience, a description of benefits conventionally associated with the study of art (especially the education of feeling), and a balanced account of modernism and postmodernism.

The pervasiveness of the notion of aesthetic experience in contemporary theories having been discussed, it must be mentioned that the concept has been subjected to criticism in recent aesthetic theory. Critiques tend to center on doubts about the existence of a distinctive experience or attitude called aesthetic. But the debate is hardly closed and theorists continue to argue persuasively that aesthetic encounters inspire and vitalize human experience and are therefore part of any worthwhile life. But regardless of the fate of the aesthetic in philosophical analysis, it remains indisputable that efforts by theorists to identify an aesthetic strand in human experience have contributed importantly to the depth of understanding and the quality of the experience of art and nature.

3. *The Unity of Aesthetic Education*

The diversity of aims and emphases in theories of aesthetic education raises the question of whether it has any unity as a field of study. Assuming the word 'aesthetic' sustains a more than casual relationship to 'education,' the field can be said to be unified by a solicitude for aesthetic value. No other area of study can be said to be as preoccupied with this particular value, a fact that argues for aesthetic education's occupying a singular territory within the philosophy of education. Concomitantly, the purpose of aesthetic education could be interpreted as initiating the young into a unique realm of value—the value afforded by aesthetic experience. Such an interpretation of aesthetic education's aims prompts a few further observations. First, it necessitates appropriate adjustments in teacher preparation. Second, goal achievement presupposes the mastery of relevant aesthetic concepts and the acquisition of aesthetic dispositions—which,

however, are unlikely to be attained if, as some theorists advocate, the arts should be used primarily in furtherance of the objectives of other subject areas. Third, since all the arts possess the capacity to induce aesthetic experience, it seems reasonable to organize aesthetic studies according to one of the educational schemes that recommend grouping the arts together.

Prospects for the future of aesthetic education however, would appear to be clouded. On the one hand, analytical critiques question the viability of the concept of the aesthetic, while ideology-driven theories of art and arts education often exhibit an antiaesthetic bias. On the other hand, the endurance of the American *Journal of Aesthetic Education* (1966–), evidence of increased cooperation between aestheticians and educators (Moore 1995), the founding of a committee on education within the American Society for Aesthetics, and two essays on aesthetic education in the first English-language *Encyclopedia of Aesthetics* (Vol. 2, Oxford University Press, New York, 1998) suggest continuing interest in the subject.

4. Definitions of Key Terms

Aesthetics. A branch of philosophy that inquires into the nature, meaning, and value of art; or any critical reflection about art, culture, and nature.

Aesthetic point of view. A distinctive stance taken toward phenomena, e.g., works of art and nature, for the purpose of inducing aesthetic experience.

Aesthetic experience. A type of experience that manifests the savoring of phenomena for their inherent values, in contrast to practical activities and values.

Aesthetic value. A type of value, in contrast, e.g., to economic value, etc.; also the capacity of something by virtue of its manifold of qualities to induce aesthetic experience.

Aesthetic literacy. A cluster of capacities that enables engagements of phenomena, especially works of art, with prerequisite percipience.

Aesthetic culture. A distinctive domain of society, in contrast, e.g., to its political culture, and, normatively, sensitivity in matters of art and culture, as in a person's aesthetic culture.

Interrelatedness of the arts. Implies features that different kinds of art have in common; or programs that group the arts together for purposes of study.

See also: Architecture; Art, Sociology of; Community Aesthetics; Culture, Production of; Culture-rooted Expertise: Psychological and Educational Aspects; Dewey, John (1859–1952); Fine Arts; Oral and Literate Culture

Bibliography

Arnstine D 1967 *Philosophy of Education: Learning and Schooling*. Harper and Row, New York

Beardsley M C 1982 Aesthetic experience. In: Wreen M J, Callen D M (eds.) *The Aesthetic Point of View*. Cornell University Press, Ithaca, NY, pp. 285–97

Broudy H S 1994 [1972] *Enlightened Cherishing: An Essay on Aesthetic Education*. University of Illinois Press, Urbana, IL

Dewey J 1958 [1934] *Art as Experience*. Putnam's Sons, New York

Eisner E 1991 *The Enlightened Eye: Qualitative Inquiry and the Enhancement of Educational Practice*. Macmillan, New York

Greene M 1981 Aesthetic literacy in general education. In: Soltis J F (ed.) *Philosophy and Education*. University of Chicago Press, Chicago, pp. 115–41

Howard V 1992 *Learning by All Means: Lessons from the Arts*. Peter Lang, New York

Kaelin E F 1989 *An Aesthetics for Educators*. Teachers College Press, New York

Moore R (ed.) 1995 *Aesthetics for Young People*. National Art Education Association, Reston, VA

Parsons M, Blocker H G 1993 *Aesthetics and Education*. University of Illinois Press, Urbana, IL

Read H 1956 [1943] *Education Through Art*, 3rd edn. Random House, New York

Read H 1966 *The Redemption of the Robot*. Trident Press, New York

Schiller F 1967 [1793–95] *On the Aesthetic Education of Man in a Series of Letters*. Wilkinson E M, Willoughby L A (eds., trans.). Oxford University Press, Oxford

Smith R A 1989 *The Sense of Art: A Study in Aesthetic Education*. Routledge, New York

Swanger D 1990 *Essays in Aesthetic Education*. Mellon Research University Press, San Francisco

R. A. Smith

Affirmative Action: Comparative Policies and Controversies

1. Introduction

Although the phrase 'affirmative action' apparently originated in the United States in 1961, the practice of providing benefits or preferential treatment to individuals based on their membership in a disadvantaged group can be found in a wide variety of forms in many other countries. For example, India developed affirmative programs as early as 1927, and was probably the first country in the world to create a specific constitutional provision authorizing affirmative action in government employment. Other countries with more recently developed affirmative action programs include Australia, Israel, and South Africa.

2. Comparative Issues in Designing Affirmative Action Programs

Galanter (1992) identifies several issues that are critical to a comparative study of affirmative action programs: justifications, program designers, selection of bene-

ficiary groups, distribution of benefits within a group, relations between multiple beneficiary groups, determination of individual eligibility, resources to be devoted, monitoring, and termination. This section will provide a comparative analysis of three of these issues; justifications, selection of groups, and individual eligibility.

2.1 Justifications for Affirmative Action

Affirmative action programs for racial minorities in the US typically seek to remedy harm caused to specific individuals by 'cognitive bias,' that is, harm caused by an actor who is aware of the person's race, sex, national origin, or other legally-protected status and who is motivated (consciously or unconsciously) by that awareness. Much of the current skepticism in the US about affirmative action may result from this narrow focus: many white people seem to believe themselves free of such cognitive bias and thus doubt that it is a continuing problem of sufficient magnitude to justify affirmative action. Such a focus makes affirmative action particularly vulnerable in settings like university admission, where decisions based on grades and test scores seem, to many, to be immune cognitive bias (see *Race and the Law*; *Gender and the Law*).

Although cognitive bias-type discrimination based on caste status is treated as a serious, continuing problem in India, affirmative action there is focused more on eradicating the enduring effects of centuries of oppression and segregation. There appears to be a more conscious commitment than in the US to change the basic social structure of the country. The Indian approach perhaps can be understood best using the economic theory pioneered by Glenn Loury, which distinguishes between human capital and social capital (Loury 1995). Human capital refers to an individual's own characteristics that are valued by the labor market; social capital refers to value an individual receives from membership in a community, such as access to information networks, mentoring, and reciprocal favors. Potential human capital can be augmented or stunted depending on available social capital. Economic models demonstrate how labor market discrimination, even several generations in the past, when combined with ongoing segregated social structure, can perpetuate indefinitely huge differences in social capital between ethnic communities. Since the landmark case of State of Kerala vs. Thomas (1976), decisions of the Indian Supreme Court have recognized the need for affirmative action to redress systemic inequality. Even though the constitutional provisions authorizing affirmative action are written as exceptions to guarantees of equality, the Court has characterized these provisions as providing instead a right to substantive equality rather than a simply formal equality.

Sunstein (1994) foreshadowed the potential value to the US of learning from India's differing justifications for affirmative action. The author proposed an anticaste principle in order to reconceptualize the American post-Civil War 14th Amendment (that no law may be enacted that abridges the rights of citizens of the USA), which was a source of both civil rights legislation and reverse discrimination attacks on affirmative action. Under Sunstein's anticaste principle, affirmative action would not be seen as a limited exception to the constitutional guarantee of equality, but rather as a logical, perhaps necessary, method of correcting the effects of caste, which interfere with equality. '(T)he inquiry into caste has a large empirical dimension … focus(ing) on whether one group is systematically below others along important dimensions of social welfare.' For Sunstein the key dimensions are income level, rate of employment, level of education, longevity, crime victimization, and ratio of elected political representatives to percentage of population. The range of persons who can make 14th Amendment claims would be drastically reduced from the entire population (all of whom have a race) to those who are members of a low caste. Thus, reverse discrimination claims by whites affected by affirmative action would disappear. Further, it would not be necessary to prove discrimination, either contemporaneous discrimination against an individual plaintiff or historical discrimination against that person's group, since the purpose of the 14th Amendment would no longer be interpreted as preventing or remedying discrimination but rather alleviating systemic social disadvantage. (See also Cunningham and Menon 1999, Sunstein 1999.)

India's justification of affirmative action (altering systemic inequality) can be seen as well as in several other countries' efforts to address the problems of diverse populations. Israel has developed affirmative action programs for Sephardi Jews, who typically have immigrated to Israel from Middle Eastern and North African countries, and have been socially and economically disadvantaged in comparison to Ashkenazi Jews, who typically have emigrated from Europe. These Israeli programs do not aim to combat current discrimination or to compensate for past discrimination. There is no history of Ashkenazi dominance and exploitation of the Sephardim comparable to the treatment of African-Americans in the US or the lower castes in India. Rather the programs have been justified in terms similar to the current constitutional discourse in India, recognizing that the combination of initial socioeconomic disadvantage with the continuing influence of informal networks would perpetuate a society divided along the Sephardi/Ashkenazi line, thus requiring affirmative action to counteract these social forces (see Shetreet 1987).

The new constitution of the Republic of South Africa takes the Indian approach one step further. The

very concept of equality is defined so that only unfair discrimination is prohibited. Properly designed affirmative action is thus fair discrimination. The constitution also explicitly states that 'to promote the achievement of equality, legislative and other measures designed to protect or advance persons, or categories of persons, disadvantaged by unfair discrimination may be taken.' (See Cunningham 1997, pp. 1624–28.)

Australia, in contrast, attempts to preserve principles of formal equality in its legislation designed to increase female participation throughout private-sector employment, by justifying programs as simply a 'fair go' for women and as consistent with 'best business practices.' The legislation specifically states that hiring and promotion on the basis of merit is not affected by affirmative action, which intended instead to facilitate the accurate recognition of merit among female as well as male employees (see Braithwaite and Bush 1998).

2.2 Selection of Beneficiary Groups

India appears to be unique among the countries of the world in the degree to which its affirmative action programs have wrestled with the problem of selecting beneficiary groups. The constitutional provisions authorizing affirmative action identify three general categories: (a) Scheduled Castes (descendants of the former 'untouchables'), (b) Scheduled Tribes (ethnic groups generally living in remote and hilly regions), and (c) other 'socially and educationally backward classes of citizens.' The greatest difficulty and controversy has focused on selection of groups for this third category, generally termed the OBCs (Other Backward Classes). In the first three decades after adoption of the Indian constitution, selection of groups for OBC designation was left largely to state governments within India's federal system of government. As a result, the Indian Supreme Court repeatedly struck down plans that seemed primarily to benefit politically powerful groups, or that were based on traditional assumptions of caste-based prejudice without knowing which groups were truly in greatest need.

In 1980 a Presidential Commission (known as the Mandal Commission after the name of its Chairperson) issued a comprehensive report and set of recommendations for national standards for OBC designation. Responding to the Supreme Court's concern about objective and transparent processes, the Mandal Commission conducted a national survey that started with generally recognized group categories (typically based on caste name or hereditary occupation) and tested each group using standardized criteria of 'backwardness' (such as comparing the percentage of group members who married before the age of 17, or who did not complete high school, with other groups in the same state). Eleven numerical factors, given varying weights, were assigned to each group based on the survey results and those groups with total scores below a specified cut-off point appeared in a list of OBCs. The Commission then recommended that a percentage of new hires for most central government jobs be reserved for OBC members under a quota system.

The Mandal Report generated lively debate but it was not until 1990 that the national government actually proposed implementation of the Report. This announcement, by then-Prime Minister V. P. Singh, prompted widespread civil disturbance, instances of self-immolation by high-caste Hindus in protest, and litigation leading to three months of oral argument before the Supreme Court. In 1992 the Supreme Court reached a 6–3 decision, largely approving the Report and its recommendations. A majority of the Supreme Court justices approved the following basic principles: (a) Traditional caste categories can be used as a starting point for identifying OBCs but selection criteria must include empirical factors beyond conventional assumptions that certain castes are 'backward.' (b) Identification of a group as an OBC can not be based on economic criteria alone (Indra Sawhney vs. Union of India 1993).

In contrast to India, affirmative action programs in the US have not used consistent criteria for defining group boundaries or for selecting eligible groups. For example, one US federal court struck down a law school admission program at the University of Texas, in part because only blacks and Mexican Americans were eligible for affirmative action consideration; Hispanic Americans, Asian Americans and Native Americans were excluded (Hopwood vs. State of Texas 1996). Many people who oppose affirmative action programs in the United States because they use racial categories such as black, African American, or Latino claim that equally effective and more equitable programs can be developed using only class categories, such as low-income (see Malamud 1996). Economist Glenn Loury, who is African American, has suggested that affirmative action is not needed by all African Americans but instead should be focused on a distinct group whose members share the following characteristics: (a) slave ancestry, (b) rural and Southern origins, (c) current residence in northern cities, (d) current residence in ghettos. He uses the term 'caste' to describe this group (Loury 1997).

In South Africa current affirmative action programs are haunted by the categorization systems of the apartheid regime that distinguished between black Africans, coloreds (mixed European and African ancestry), and Indians (some ancestry from the Indian subcontinent). The ruling party, the African National Congress, in its earlier role as the leading opponent to apartheid, sought political solidarity among all peoples oppressed by apartheid; it used 'black' to refer to Africans, coloreds, and Indians. The 1998 Employment Equity Act, implementing affirmative action in both the public and private sector, continued this

tradition by targeting 'Black people' (combining the three Apartheid-era categories) as well as women and people with disabilities. However, this selection system exists in tension with the recognition that coloreds and Indians were differently disadvantaged compared to those designated by apartheid as 'black Africans.' For example, a South African court has upheld a medical school admission program that gave greater preference to black African applicants than to Indian applicants (Motala vs. University of Natal 1995).

2.3 Determination of Individual Eligibility

In the US, individual eligibility for affirmative action is usually based solely on membership in one of the selected beneficiary groups. An apparent exception is the federal Disadvantaged Business Enterprise (DBE) program, an affirmative action program affecting federally-funded contracts, in which membership of one of the designated beneficiary groups only creates a presumption of eligibility (see Adarand Constructors vs. Pena 1995). However, a minority-owned business is not required to provide additional evidence of disadvantage beyond group membership to be eligible; instead the presumption is conclusive unless a third party (typically a disappointed competing bidder) asserts that the individual beneficiary is not personally disadvantaged.

In India, an individual eligibility test is being implemented pursuant to the decision of the Indian Supreme Court in Indra Sawhney vs. Union of India 1993. This 'creamy layer' approach—as it is termed in India—addresses two different but related concerns: (a) that the benefits of affirmative action are not distributed evenly throughout a backward group but instead are monopolized by persons at the socio-economic top of the group: and (b) that benefits are going to persons who do not in fact need them, because they have been raised in privileged circumstances due to parental success in overcoming the disadvantaged status of the backward group. Interestingly, the criteria proposed by the national government after the court's decision focus more on the wealth and occupation of the individual's parents than of the individual, reflecting perhaps continuing sensitivity to the role of social capital in perpetuating disadvantage (see *Class and Law*).

3. Comparative Studies of Affirmative Action

Clearly there is a need for more comparative scholarship on affirmative action, although the last years of the 1990s saw a significant increase in published work in this area. Galanter (1984), a classic in this area, points out the need to be cautious about the comparative lessons that the United States and other countries could learn from India. Thomas Sowell, a US economist critical of affirmative action policies in the United States, has frequently made use of comparative materials, most extensively in Sowell (1990) which includes sections of India, Malaysia, Nigeria, and Sri Lanka. In 1991, during the transition period that led to the abolition of apartheid and the founding of the new Republic of South Africa, the Constitutional Committee of the African National Congress convened a conference on 'Affirmative Action in the New South Africa' that included subsequently published studies of affirmative action in India and Malaysia as well as the United States Centre for Development Studies (1992). A set of conference proceedings published in 1997 includes cross-national and interdisciplinary perspectives on affirmative action by public officials and social scientists from India, South Africa, and the United States (Cunningham 1997); the same year also saw the publication of Parikh (1997). An extensive section on affirmative action in India may be found in Jackson and Tushnet's law text on comparative constitutionalism (Jackson and Tushnet 1999), and Andrews (1999) includes studies of affirmative action in Australia, India, South Africa, and the United States.

See also: Affirmative Action: Empirical Work on its Effectiveness; Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Class and Law; Critical Race Theory; Discrimination: Racial; Equality and Inequality: Legal Aspects; Equality of Opportunity; Ethnic and Racial Social Movements; Gender, Class, Race, and Ethnicity, Social Construction of; Race and the Law; Race Identity; Race Relations in the United States, Politics of; Racial Relations; Sex Differences in Pay; Sex Segregation at Work

Bibliography

Adarand Constructors vs. Pena 1995 *United States Reports* **515**: 200

Andrews P E (ed.) 1999 *Gender, Race and Comparative Advantage: A Cross-National Assessment of Programs of Compensatory Discrimination*. Federation Press, Annandale, NSW, Australia

Braithwaite V, Bush J 1998 Affirmative action in Australia: A consensus-based dialogic approach. *National Women's Studies Association Journal* **10**: 115–34

Centre for Development Studies 1992 *Affirmative Action in a New South Africa*. University of the Western Cape, Belville, SA

Cunningham C D (ed.) 1997 Rethinking equality in the global society. *Washington University Law Quarterly* **75**: 1561–676

Cunningham C D, Menon N R M 1999 Race, class, caste …? Rethinking affirmative action. *Michigan Law Review* **97**: 1296–310

Galanter M 1984 *Competing Equalities: Law and the Backward Classes in India*. University of California, Berkeley, CA

Galanter M 1992 The structure and operation of an affirmative action programme: An outline of choices and problems. In: *Affirmative Action in a New South Africa*. University of the Western Cape, Belville, SA

Hopwood vs. State of Texas 1996 *Federal Reporter* **78**(3): 932–68
Indra Sawhney vs. Union of India 1993 *All India Reports, Supreme Court* **477**
Jackson V C, Tushnet M 1999 *Comparative Constitutional Law*. Foundation Press, New York
Loury G C 1995 *One by One from the Inside Out: Essays and Reviews on Race and Responsibility in America*. Free Press, New York
Loury G C 1997 The hard questions: Double talk. *The New Republic* **23**
Malamud D C 1996 Class-based affirmative action: Lessons and caveats. *Texas Law Review* **74**: 1847–900
Motala vs. University of Natal 1995 *Butterworths Constitutional Law Reports* **3**: 374
Parikh S 1997 *The Politics of Preference: Democratic Institutions and Affirmative Action in the United States and India*. University of Michigan, Ann Arbor, MI
Shetreet S 1987 Affirmative action for promoting social equality: The Israeli experience in positive preference. *Israel Yearbook on Human Rights* **17**: 241
Sowell T 1990 *Preferential Policies: An International Perspective*. Morrow, New York
State of Kerala vs. Thomas 1976 *All India Reports, Supreme Court* **490**
Sunstein C R 1994 The Anti-caste principle. *Michigan Law Review* **92**: 2410–55
Sunstein C R 1999 Affirmative action, caste and cultural comparisons. *Michigan Law Review* **97**: 1311–20

C. D. Cunningham

Affirmative Action, Empirical Work on its Effectiveness

1. Introduction

Affirmative action generally refers to a set of public policies meant to redress the effects of past or present discrimination. Affirmative action connotes active measures to level the playing field for access to education, to jobs, and to government contracts. While a wide variety of affirmative action policies have been implemented in different countries, much of the existing research concerns experience with the affirmative action directed at expanding employment opportunities for women and minorities in the United States, the focus here.

One defining characteristic of the policy in the United States is its ambiguity. The closest the US Congress has come to explicitly requiring affirmative action in employment is in the Americans with Disabilities Act of 1990, which explicitly requires that employers make 'reasonable accommodations' to hire the disabled. This requires that employers must do more than be blind to differences between the disabled and the able, but must actively invest to overcome these differences. It is worth noting that during a period when affirmative action policies were contentiously debated, a law explicitly requiring affirmative action in employment was swiftly enacted without mention of affirmative action. Affirmative action has long been a political lightning rod. To its critics, it is symbolic of quotas unfairly and rigidly imposed. To its proponents, it is symbolic of redressing past wrongs and of leveling the current playing field. An empirical basis for policies to counterbalance current employment discrimination can be sought in (a) continuing judicial findings of systematic employment discrimination, (b) statistical evidence of wage disparities across demographic groups, and (c) audit studies of employers' hiring behavior. For reviews of this evidence, see Altonji and Blank (2000), Blau (1998) Heckman (1998), and Fix and Struyk (1993).

2. Affirmative Action in the Shadow of Title VII of the Civil Rights Act of 1964

Affirmative action in the United States has been implicitly encouraged by judicial interpretation of Title VII of the Civil Rights Act of 1964 (CRA), and explicitly required, but not defined, by Executive Order 11246 applied to federal contractors. From its inception, the CRA has embodied a tension between words that bar employment discrimination, and a Congressional intent to promote voluntary efforts to redress discrimination. The courts have struggled with making room for what they sometimes saw as the intent of Congress within the language of the CRA. In later cases, the Supreme Court read Section 703(J)'s bald statement that 'Nothing in the act shall require numerical balancing' as allowing numerical balancing as a remedy. In other cases, the court held that the act should not be read to bar voluntary acts to end discrimination. These cases left considerable if ambiguous room for affirmative action. The threat of costly disparate impact litigation under Title VII following the Griggs v. Duke Power case created considerable incentive to undertake affirmative action.

The extent of 'voluntary' (or at least non-judicially directed) affirmative action taken in response to Title VII can be roughly estimated from existing work on the overall impact of Title VII. Because firms directly subjected to litigation under Title VII represent such a small share of employment, the bulk of the black economic advance credited to Title VII must be due to its indirect effect in promoting 'voluntary' affirmative action. For example, Freeman (1973) shows the overall impact of Title VII in the time series of black–white earnings differentials. From this, the direct impact of Title VII litigation at companies sued for racial discrimination could be subtracted. While these impacts are substantial at the companies incurring litigation, employment at such companies makes up only a small fraction of employment. The impact on market wages from the outward shift in the demand curve for blacks at these companies can only be small

given their small share of the market. Most charges of discrimination never result in litigation, and much litigation never reaches a judicial decision, leaving ambiguous interpretation and little trail. It is likely that the bulk of the changes attributed to Title VII are in a broad sense the result of affirmative action taken in response to the indirect threat, rather than the direct act, of litigation. Since the impact of Title VII is at least an order of magnitude greater than that of the contract compliance program, and only a small share of this could have occurred at companies directly litigated against, the biggest impact of affirmative action, broadly defined, very likely took place in the shadow of Title VII. To date, no one has filled in the blanks in the calculation just outlined.

One difficulty in evaluating the impact of a national law such as the 1964 CRA is that a contemporaneous group unaffected by the law is usually lacking for comparison. In the case of Title VII, Chay (1998) overcomes this problem by examining the impact of the Equal Employment Act of 1972, which extended the reach of Title VII to smaller establishments. Chay finds that once smaller establishments came under Title VII's coverage, the relative employment and pay of blacks improved at these small establishments. Again, very little of this can be the direct result of litigation against small establishments.

3. *Affirmative Action Through Contract Compliance*

In the US, Federal contractors and first-tier subcontractors are required by Executive Order to pursue affirmative action. Because the Executive Order does not apply to all companies, it sets up a contrast that under some assumptions allows us to discern the impact of affirmative action under the contract compliance program by comparing the change in demographics at contractors with that at non-contractors. The interpretation of these comparisons would be complicated if each employer's demographics or anticipated change in demographics affected its selection or self-selection into contractor status. However, there is no evidence of this. To the contrary, initial demographics are generally not used as a prerequisite to becoming a contractor. And, as will be explained further below, because of peculiarities in the enforcement targeting procedures in use during much of the programs, existence, firms with unusually low proportions of minorities or females were not likely to come under greater regulatory pressure. So given the regulations as implemented, there was little reason for firms with low representation of minorities or females to avoid becoming a contractor. Since firms were not held rigidly to whatever promises they may have made to increase the employment of minorities or females, there was also no reason for firms that did not anticipate an increase in female or minority share to avoid contractor status. Given that enforcement creates a generalized pressure rather than a tight link from a firm's demographics to the targeting of enforcement, reverse causation is less of a concern. We can be more confident that the differences in changing demographics between contractor and non-contractor establishments firms are not an artifact of selection or self-selection into or out of contractor status. The research on the contract compliance program allows somewhat stronger conclusions about a weaker program because it relies not just on time-series variation but on a comparison of establishments with and without the affirmative action obligation.

Two mutually inconsistent criticisms of affirmative action have found prominent voice. The first criticism is that affirmative action goals without measurable results invite sham efforts; because these programs do not work very well, the argument goes, they should be disposed of. The second criticism is that an affirmative action 'goal' is really a polite word for a quota; in other words, that affirmative action works too well, and therefore should be disposed of. Empirical research into affirmative action programs contradicts both these criticisms. Affirmative action goals have played a statistically significant role in improving opportunities for minorities. At the same time, those goals have not resulted in 'quotas' (Leonard 1985b).

Federal Contract Compliance Program Regulations require that every contractor maintain an affirmative action plan consisting in part of a utilization analysis indicating areas of minority and female employment in which the employer is deficient, along with goals and timetables for good-faith efforts to correct deficiencies. To put this program in perspective, it is important to understand that, through most of its history, between 400 and 100 investigative officers have been responsible for enforcing compliance. The ultimate sanction available to the government is debarment, in which a firm is barred from holding federal contracts. In the history of the program, there have been fewer than 60 debarments. The other sanction typically used is a back-pay award. This sanction too is used infrequently (Leonard 1985a).

There have been several studies of the impact of the Contractor Compliance Program (Ashenfelter and Heckman 1976, Goldstein and Smith 1976, Heckman and Wolpin 1976, Smith and Welch 1984, Leonard, 1984c, Rodgers and Spriggs 1996). These studies have consistently found that employment of black males increases faster at establishments that are federal contractors than at those that are not, with the possible exception of the period during which the Reagan Administration undercut enforcement. While significant, this effect has always been found to be modest in magnitude. In general, the impact on females and on non-black minorities has been less marked than that on blacks.

Between 1974 and 1980, black male and female employment shares increased significantly faster in

contractor establishments than in noncontractor establishments. These positive results are especially noteworthy, in view of the relatively small size of the agency and its limited enforcement tools. First consider differences between contractors and noncontractors in the change in female and minority employment share without controlling for other possibly confounding variables. A summary measure, white males mean employment share, declined by 5 percent among contractors compared to 3.5 percent among noncontractors. In other words, the contractors subject to affirmative action reduced white males' employment share by an additional 1.5 percent over six years. For other groups, the comparable difference between contractors and noncontractors in the change in employment share is 0.6 percent for white females, 0.3 percent for black females, 0.0 for other minority females, 0.3 for black males, and 0.1 for other minority males (Leonard 1984c).

Controlling for establishment size, growth region, industry, occupational, and corporate structure, the annual growth rate of black male employment grew 0.62 percent faster in the contractor sector, and the annual growth rate of white males grew 0.2 percent slower. Contractor status shifted the demand for black males relative to white males by 0.82 percent per year. This is less than 1 percent per year, not a dramatic change although it can cumulate over time. Studies of earlier and later periods by Ashenfelter and Heckman (1976), Heckman and Wolpin (1976), and Rodgers and Spriggs (1996), find effects on black males of the same order of magnitude. Not surprisingly, growing establishments are better able to accommodate the regulatory pressures. Minority and female employment share increased significantly faster in growing establishments.

In addition to increasing the sheer numbers of minorities employed, the program also has had some success in helping move them up the career ladder. Affirmative action appears to increase the demand for poorly educated minority males as well as for the highly educated. Black males' share of employment increased faster in contractor establishments in every occupational group except laborers and white-collar trainees. Black females in contractor establishments increased their employment share in all occupations except technical, craft, and white-collar trainee. The positive impact of the contract program is even more marked when the position of black females is compared with that of white females (Leonard 1984b). While some part of this improvement appears to reflect title inflation by employers upgrading detailed occupational categories with a relatively high representation of minorities or females (Smith and Welch 1984), the occupational advance is accompanied by wage increases incompatible with pure title inflation (Leonard 1986). The evidence does not support the contention that this is just a program for blacks with skills. To the contrary, it helps blacks across the board. Thus, affirmative action does not appear to have contributed to the economic bifurcation of the black community.

The employment goals to which firms agree under affirmative action are not vacuous; neither are they adhered to as strictly as quotas. Affirmative action 'goals' are often considered to be a euphemism for quota. This appears to overstate what the regulatory authorities actually require: good faith efforts to meet goals that are in the first instance chosen by the employer and may subsequently be negotiated with the OFCCP. Under the Contract Compliance Program, firms agree to set goals and timetables. Firms that agree to increase minority employment by 10 percent will, on average, increase minority employment by about 1 percent. Employers are not sanctioned for failing to meet their goals. A good-faith effort toward compliance in practice means that firms make it about one-tenth of the way toward the stated goal. This falls well short of the rigidity expected of quotas (Leonard 1985b).

Both the contract compliance program and Title VII have been criticized for inducing firms simply to hire a 'safe number' of minorities and women, irrespective of qualifications. According to this argument, whatever 'quota' is imposed should be the same for firms in the same industry and region. Application of such quotas would mean that firms in the same industry hiring out of the same labor market should, over time, start to look more and more like each other in terms of the percent of women and minorities. This does not generally happen, either under Title VII or as a result of affirmative action programs, Leonard (1990).

There is some evidence that the general pressures of affirmative action have succeeded in prompting employers to search more widely, but at the same time have been flexible enough not to impinge upon the economic performance of these firms. Holzer and Neumark (1999) find that although affirmative action employers tend to hire women and minorities with lower educational qualifications, these hires do not exhibit reduced job performance.

To summarize, contractor goals do have a measurable and significant correlation with improvements in the employment of minorities and females at the reviewed establishments. At the same time, these goals are not being fulfilled with the rigidity one would expect of quotas.

4. *Future Research*

Classic economic models of discrimination recognize that discrimination is not only unfair, but also inefficient. The available evidence suggests that affirmative action has helped to promote black employment, and to a lesser degree that of females and of non-black minorities. But by themselves, these employment gains

do not tell us whether affirmative action has reduced discrimination or gone beyond that point to induce reverse discrimination. Two fundamental questions remain: has affirmative action reduced discrimination, and has it helped integrate society?

One method of answering the first question is to ask whether workers of various demographic groups subject to labor market discrimination are more likely under affirmative action to be employed and paid in proportion to their productivity without regard to their race or sex. Productivity is difficult to measure, but a model for this approach appears in recent papers by Hellerstein et al. (1999) for the US, and Hellerstein and Neumark (1999) for Israel.

Whether the employment gains associated with affirmative action and Title VII have reduced discrimination or gone beyond that to induce reverse discrimination can be addressed by asking whether industries that had come under the most pressure from Title VII or affirmative action, or industries that had increased their employment of minorities and women, had suffered in terms of productivity. The earliest paper to use this approach (of comparing relative earnings to relative wages) relied on highly aggregated data and found that minority and female employment gains under affirmative action had not reduced productivity (Leonard 1984a). Direct testing of the impact of affirmative action on productivity finds no significant evidence of productivity decline, which implies a lack of substantial reverse discrimination. However, these results at the industry-State level are too imprecise to answer this question with great confidence.

Hellerstein et al. (1999) refine this approach of comparing wage differentials to estimated productivity differentials, and develop much more persuasive results with a detailed study of establishments. They find that pay premiums for older workers are roughly matched by productivity increases, but that the female pay penalty in the US is generally not a reflection of lower productivity. The latter result suggests that far from imposing pervasive reverse discrimination against men, affirmative action for women has not yet succeeded in eliminating labor market discrimination. This direction of research holds great promise to move us beyond the formulaic debates over whether the wage differences that remain between groups after controlling for observable differences in qualifications and preferences are due to omitted human capital or discrimination.

A second approach to this question uses stock-market returns to ask whether investors consider Title VII litigation good for companies because it forces them to use labor more efficiently. Hersch (1991) finds a negative effect on investment returns. This might mean that investors do not believe that litigation will force the firms to become more efficient. Alternatively, investors might take the litigation as news that management is more inept than they thought.

The political question of whether affirmative action has served to unify or divide is of continuing concern. Sniderman and Piazza (1993) conduct survey based experiments in the US and report that just using the term affirmative action is enough to provoke more discriminatory responses.

Studies of affirmative action type policies in education and government contracting are in their infancy. Analysis of both education and public contracting face the challenge of measuring the impact of heterogeneous policies with substantial local variation that is difficult to pin down. For higher education, see Datcher-Loury and Garman (1993) in contrast to Kane (1998). Even less is known about set-aside programs at the federal, state, and local level, (see Bates and Williams 1996), although anecdotal evidence suggests that gains under these programs quickly evaporate when set-asides are suspended. The open questions include whether there is evidence of systematic discrimination in public or private contracting, as well as how discrimination and set-asides programs affect the profitability and growth of firms. Given the context of laws designed to limit discretion in public contracting, evidence of discrimination in contracting can be thought of as a miner's canary signaling loopholes in public contracting laws (see *Discrimination*).

5. *Other Mechanisms, Other Countries*

One of the most interesting mechanisms for pursuing affirmative action can be found in the German policy to promote employment of the disabled. It is a play or pay policy. Employers face an explicit quota for employment of the disabled. The German government has drawn a bright line clearly stating minimum employment share for the disabled. However, together with this clear standard is a buy out provision; firms that do not meet their quota, for whatever reason, pay into a special government fund. This mechanism is remarkably different from those used in the US. First, the quota is explicit and stated in terms of numerical standards rather than abstract rights. This reduces the uncertainty, risks, delays, and costs of enforcement through the courts. There is also little opportunity to argue over the existence of a hidden quota. At the same time, the buy-out provision defuses complaints about the costly burdens imposed by rigid quotas, because employers experiencing the greatest difficulty meeting the quota can buy their way out. The nearest example of a US policy using a similar mechanism is the creation of tradable pollution rights under US environmental law. Both mechanisms are desirable on efficiency grounds. Firms that face the greatest difficulty reaching the legal standard can buy the right to pollute (or to hire fewer disabled) from firms that can more easily meet the standard. The aggregate level of disabled employment can be adjusted by changing the

overall standard, leaving to a decentralized market mechanism the question of which firms will employ the disabled and which will pay a fine. One additional benefit is that the penalty paid by firms below quota goes into a fund used to support the training and rehabilitation of the disabled.

Whatever its desirability on efficiency grounds, this mechanism appears to be politically impractical in the context of race and sex discrimination in the US. Explicitly setting quotas and putting a price on the right to employ fewer minorities or women raises political conflicts. The current policy places only an implicit price on discrimination. But it is apparently politically advantageous to frame the discussion in terms of absolute human rights, rather than to explicitly highlight the relative price society is willing to put on these rights through enforcement budgets and employment standards.

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Disability: Sociological Aspects; Discrimination; Discrimination, Economics of; Discrimination: Racial; Employment and Labor, Regulation of; Equality and Inequality: Legal Aspects; Gender and the Law; Law and People with Disabilities; Race and the Law; Sex Differences in Pay; Sex Segregation at Work

Bibliography

Altonji J, Blank R 2000 Race and gender in the labor market. In: Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*. North-Holland, Amsterdam

Ashenfelter O, Heckman J 1976 Measuring the effect of an antidiscrimination program. In: Ashenfelter O, Blum J (eds.) *Evaluating the Labor Market Effects of Social Programs*. Princeton University Press, Princeton, NJ, pp. 46–84

Bates T, Williams D 1996 Do preferential procurement programs benefit minority business? *American Economic Review* **86**(2): 294–7

Blau F 1998 Trends in the well-being of American women, 1970–1995. *Journal of Economic Literature* **36**(1): 112–65

Bloch F 1994 *Antidiscrimination Law and Minority Employment*. University of Chicago Press, Chicago

Chay K 1998 The impact of Federal civil rights policy on Black economic progress: Evidence from the Equal Employment Opportunity Act of 1972. *Industrial and Labor Relations Review* **51**(4): 608–32

Datcher-Loury L, Garman D 1993 Affirmative action in higher education. *American Economic Review* **83**(2): 99–103

Fix M, Struyk R 1993 *Clear and Convincing Evidence*. Urban Institute Press, Washington DC

Freeman R 1973 Changes in the labor market for Black Americans, 1948–1972. *Brookings Papers on Economic Activity* **1**: 67–120

Goldstein M, Smith R 1976 The estimated impact of the Antidiscrimination Program aimed at Federal contractors. *Industrial and Labor Relations Review* **29**(3): 524–43

Heckman J 1998 Detecting discrimination. *Journal of Economic Perspectives* **12**(2): 101–16

Heckman J, Wolpin K 1976 Does the Contract Compliance Program work? An analysis of Chicago data. *Industrial and Labor Relations Review* **29**(3): 544–64

Hellerstein J K, Neumark D 1999 Sex, wages, and productivity: An empirical analysis of Israeli firm level data. *International Economic Review* **40**(1): 95–123

Hellerstein J K, Neumark D, Troske K 1999 Wages, productivity, and worker characteristics: Evidence from plant-level production functions and wage equations. *Journal of Labor Economics* **17**(3): 409–46

Hersch J 1991 Equal employment opportunity law and firm profitability. *Journal of Human Resources* **26**(1): 139–53

Holzer H, Neumark D 1999 Are affirmative action hires less qualified? Evidence from employer-employee data on new hires. *Journal of Labor Economics* **17**: 534–69

Holzer H, Neumark D 2001 Assessing affirmative action *Journal of Economic Literature* **38**: 483–56

Kane T 1998 Racial preferences and higher education. In: Jencks C, Phillips M (eds.) *The Black-White Test Score Gap*. The Brookings Institution, Washington DC, pp. 431–56

Leonard J 1984a Anti-discrimination or reverse discrimination: The impact of changing demographics, Title VII and affirmative action of productivity. *Journal of Human Resources* **19**(2): 145–74

Leonard J 1984b Employment and occupational advance under affirmative action. *Review of Economics and Statistics* **66**(3): 377–85

Leonard J 1984c The impact of affirmative action on employment. *Journal of Labor Economics* **2**(4): 439–63

Leonard J 1985a Affirmative action as earnings redistribution: The targeting of compliance reviews. *Journal of Labor Economics* **3**(3): 363–84

Leonard J 1985b What promises are worth: The impact of affirmative action goals. *Journal of Human Resources* **20**(1): 3–20

Leonard J 1986 Splitting Blacks? Affirmative action and earnings inequality within and between races. *Proceedings of the Industrial Relations Research Association* (39th annual meeting), pp. 51–7

Leonard J 1990 The impact of affirmative action regulation and equal employment law on Black employment. *Journal of Economic Perspectives* **4**: 47–63

Rodgers W M, Spriggs W E 1996 The effect of Federal Contractor Status on racial differences in establishment level employment shares: 1979–1992. *American Economic Review* **86**(2): 290–3

Smith J, Welch F 1984 Affirmative action and labor markets. *Journal of Labor Economics* **2**(2): 269–301

Sniderman P M, Piazza T 1993 *The Scar of Race*. Harvard University Press, Cambridge, MA

J. S. Leonard

Affirmative Action Programs (India): Cultural Concerns

Affirmative action in the USA has come to mean the selection of candidates from among the blacks, Hispanics and other backward communities for appoint-

ment in preference to candidates who figure higher in the merit list. In India this procedure is described as 'reservation,' that is, reserving certain posts and places in educational and professional institutions for backward communities. The objective in all such cases is to equalize opportunity in an unequal world. However, the term 'affirmative action' is being increasingly used to conform to international practice.

1. *Affirmative Action Prescribed in Constitutions*

There are two notable cases in which the idea of affirmative action is enshrined in the constitution of modern states. The first is independent India, whose constitution came into force on January 26, 1951. The second is South Africa, whose first constitution based on adult franchise was adopted in 1996.

1.1 *Indian Constitution*

In the Indian Constitution, Articles 14, 15, and 16 guarantee the fundamental rights of equality under the law: equal protection of the law and no discrimination against any person on grounds of race, religion, caste, creed, place of birth, gender, or any of these, to positions in the state or to educational institutions in the state.

Article 16 guarantees equality of opportunity, which would appear to be another way of looking at discrimination. However, clause 4 of this Article asserts that 'nothing in this Article shall prevent the state from making any provision for the reservation of appointments or posts in favor of any backward class of citizens which, in the opinion of the state, is not adequately represented in the services under the state.' The implication of this provision is that there are some groups of persons who do not in fact have equal opportunity because of financial or social deprivation: then the state is empowered to assist them by reserving posts and positions in educational institutions for them to help them catch up with the rest of society. This proviso takes note of the fact that there cannot be equal opportunity when large sections of the populace have been held down by poverty and social discrimination.

1.2 *Fundamental Rights and Directive Principles*

Prime Minister Nehru, while addressing the Constituent Assembly (First Amendment Bill, 1951), defended the notion of affirmative action, then called reverse discrimination, by distinguishing between the claims of Fundamental Rights on the one hand, and Directive Principles on the other. 'Fundamental Rights are conceived of as static and Directive Principles represent a dynamic move towards certain objectives. If in the protection of individual liberty you protect also individual or group inequality, then you come into conflict with that Directive Principle which wants … an advance, to a state where there is less and less inequality and more and more equality. Then you become static … and cannot realize the ideal of an egalitarian society which we all desire.'

1.3 *Three Classes of Deprived Citizens*

The Constitution recognizes three classes of deprived citizens.

Scheduled Tribes (S/T) and Scheduled Castes (S/C) as laid out in relevant schedules are two classes of citizens for whom reservation was to be made when the Constitution came into operation. The extent of reservation was 27 percent, and it was to last for a period of 15 years, by which time it would have fulfilled its purpose. The third class was described as Other Backward Classes.

2. *Backward Classes Commissions*

Regarding the third category, the Other Backward Classes (OBC) Commission was set up in January 1953 and submitted its report on March 31, 1955. It listed 2,399 castes as socially and educationally backward. The Commission, which consisted of 11 members under the Chairmanship of Kaka Kalelkar, was a dead letter. Five members submitted notes of dissent. The Chairman, who did not submit a note of dissent, repudiated it, saying that reservation would deprive the country of its best talent and would lower standards. It was shelved.

The Second Backward Classes Commission was set up in December 1978 under the Chairmanship of Mr. R. P. Mandal, Member of Parliament, hereafter to be known as the Mandal Commission. The Commission consisted originally of five members and a Secretary, Mr. S. S. Gill from the Indian Administrative Service. A sixth member, Mr. L. R. Naik, was added at a later date. The Commission submitted its report on the last day of 1980. A point to be noted is that all members belonged to the OBC, which led to some critical comments from the press, alleging that the Commission was a 'packed house' and therefore could not be objective. However, the Commission was appointed by the Janata government. In the last 20 years of the twentieth century the Congress Party declined in power. It was replaced by parties which defined themselves as 'Janata' or peoples' parties. An extreme right-wing Hindu fundamentalist group called themselves the Bharatiya Janata Party (BJP). The others, which comprised many splinter groups based on personal loyalties, were also called Janata parties. The common link between these parties was their alleged adherence to secularism. However, secularism in India

does not mean separation of church from state. It means equidistance and, as a corollary, equal closeness to all religions. The Commission was given extension by the Congress government under Mrs. Indira Gandhi.

2.1 Determining Criteria

The Mandal Commission faced two major issues. The first was to decide on the extent of the population which comprised Other Backward Classes and on what bases. This narrowed itself down to the question of whether the determining factor was poverty or caste. It was argued by some, including the Congress Party, that the criterion should be poverty and not caste. They based themselves on the fact that the relevant Article of the Constitution speaks of *classes*. However, noting that poverty is not necessarily an indicator of backwardness, the Commission decided on the caste indicator. After going through a complicated calculation, the Commission came to the conclusion that the OBC constitute 52 percent of the population.

2.2 Reference to Supreme Court

The question regarding the extent of reservation a society is prepared to tolerate is a matter of negotiation between political parties. At this stage, several controversial issues in the Mandal Commission's recommendations were referred to the Supreme Court by government. The Supreme Court ruled that reservation for the OBC should be pegged at 22 percent; thus the total reservation amounted to 49 percent of the whole population. The balance of 51 percent was available for open competition. The Supreme Court also ruled that reservation should only apply at the initial state of entry into service. In addition, the Court decided that conversion does not remove the stigma of caste. It had been the policy of the government to assume that if a person is converted to a religion, which does not recognize caste, such as Christianity, Islam or Sikhism, he or she no longer stands in need of reservation.

2.3 Lack of Follow-up

It is extraordinary that nearly 20 years after its publication, the Mandal Commission Report has not been discussed in parliament. Perhaps this is because it has become an election issue. On the eve of a General Election the party in power declares its intention to implement the Report as a bait for getting the vote from the lower castes. The result has been violence, rioting, arson and self-immolation on an unprecedented scale. This happened for the first time in the capital Delhi in 1981 and in Ahmedabad and Baroda in 1985. The reaction of the upper classes was immediate and extreme. Academics, professionals, and students came out on to the streets. The self-immolation deaths which were triggered off by the Delhi riots spread to all parts of India. The overall result has been that reservation quotas have been fixed for Central Government and corporate employees, and for admission to professional institutions, but other aspects of the ways and means to create a better environment for equalizing opportunity have not been examined.

3. Moral Justification

The question arises of what justification there is to penalize some persons and appoint others to certain jobs although the former have been rated superior according to the merit criteria. Kaka Kalelkar, Chairman of the First Backward Classes Commission, a noted educationist and disciple of Gandhi, argued that it is necessary for the upper castes to atone for keeping the lower castes in a state of deprivation. To this there are at least two objections: first, those guilty of discrimination are long since dead and gone. (However, discrimination is still being practiced by the upper castes to a considerable extent.) To this the reply is that dominant groups exist and persist over time, and own responsibility. The Brahmins form such a group. This argument does have loopholes: is it fair to apply the moral notions of today to events which took place centuries ago when social structures were very different? Political and religious beliefs held by the community as a whole upheld certain types of inequality which are not compatible with liberal democratic regimes. Nevertheless, the contention is not without force.

The second argument against reservation is that it is contrary to a sense of justice and fairness that a person of greater merit should be overlooked in place of one rated as less meritorious. This objection is based on the supposition that the criteria for testing merit are objective. It has been argued that criteria are only 'facially' objective; they have a hidden bias which tilts them in favor of the dominant class or classes. Thus in India it has been a long-standing practice to rate merit on a standardized written test followed by an interview. As is well known, the upper castes, especially the Brahmins, have a long tradition of rote learning, and success in such examinations depends on memory. The interview boards at the highest level are manned by members of the august Union Public Service Commission and of similar bodies at lower levels. The Commission would certainly be the embodiment of upper class and upper caste values.

One justification for rating intellectual ability as a criterion for selecting civil servants is the supposition that it can be objectively tested. However, there are other qualities which are essential for a good civil

servant, such as honesty, integrity, incorruptibility, and persistence. To test these may not be easy, but it should not be impossible to devise such tests.

3.1 Towards Greater Equality

Regarding the steps to be taken to create a better society in which there is near equality of opportunity, one has to turn back to the Mandal Report. As a result of reservation, a few of the deprived classes will get into prestigious educational and technical institutions and into Class I jobs. However, by and large, the Commission points out, these communities can only hope to compete if special schools are set up to train them. Education in the rural areas is very poor, and it is here that the bulk of the deprived classes live. Again, the multiplicity of languages in India raises serious problems. English is the first language of a very small minority. A knowledge of English is necessary for interstate communication and is the window on the world. The latest report (1998–99) of the Ministry of Human resources informs us of several steps which have been taken to improve education in the rural areas.

3.2 Reservation and Aspiration

It has been contended that a fuss is being made about reservation of posts in government because it provides at best a mere three million jobs, a minute percentage of the workforce in the open market where reservation does not apply. But against this there are two objections: first, employment in government is important because it is these officers who determine national policies and, moreover, all sections of society have a right to participate in the framing and implementation of national policies. Second, it has been pointed out that since reservation applies to a caste, the chances are that its benefits will be acquired by a limited number of powerful caste groups. The benefits will not trickle down to the weaker castes. The Mandal Report defends this probability by asserting, 'But is this not a universal phenomenon? All reformist remedies have to contend with a slow recovery along the hierarchical gradient; … human nature being what it is, a "new class" ultimately does emerge, even in classless societies.' To say this is, in effect, to negate the entire policy of reservation. Mandal was replying to a point made by L. K. Naik that the vast majority of reservations should be distributed among the lowliest and most deprived classes, and only the remainder should be kept for the more advanced. Interestingly, this move was welcomed by the current dominant classes. It would mean that for the foreseeable future the advanced OBC would be in no position to challenge them. Hence the power structure would not change unless the already emerging deprived sections are given more and more reserved positions. Then we could expect a shift from one power group to another.

4. South Africa

South Africa's constitution guaranteed equality to all citizens; equality before the law and equal protection of the law irrespective of color, race, gender, ethnic or social origin, religion, language, belief, culture, or birth. It sought to provide free education to children up to the age of 14 years, development of the several languages of the Republic, free movement within its territory, and freedom of expression. Affirmative action is prescribed.

Article 2 of its Bill of Rights reads, 'Equality includes the full and equal enjoyment of all rights and freedoms. To promote the achievement of equality, legislative and other measures designed to protect or advance persons, or categories of persons, disadvantaged by unfair discrimination, may be taken.'

Two points are worth noting. First, reservation in India is confined to government posts and educational and professional institutions. However, demands are now being made that reservation should be extended to the private sector. In South Africa it covers the whole gamut of institutions in the country. Second, no fixed percentage has been laid down for the allocation of posts or places in institutions for underprivileged persons. The number of places to be filled by the deprived will depend on the number of qualified persons available for the posts—and the degree of discrimination practiced in the past.

4.1 Measures Initiated in White Paper 1997

However, a White Paper outlining new affirmative action targets was released on November 6, 1997. It spells out practical steps that departments must follow in implementing broad affirmative action goals set out in other policy documents and statutes. Programs must focus on three main areas, namely, achieving representation, redressing disadvantages, and developing a more diverse management culture. Two of the targets, which had to be achieved by 1999, are an increase in black management by 50 percent and more than doubling of the number of women in middle and senior management from 11 percent to 30 percent. By 2005, at least 2 percent of public servants must be disabled people. Compulsory affirmative action quotas will not be introduced for the private sector, but there will be incentives for companies that promote employment equity. Employers will be required by law to implement affirmative action at their workplaces by December 1999. The government would begin implementing the Employment Equity Act in four phases from May 1999.

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on its Effectiveness; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Sex Differences in Pay

Bibliography

Chatterji P C 1984 *Secular Values for Secular India*, 2nd edn. 1995. Lola Chatterji, New Delhi

Chatterji P C 1988 Reservation: theory and practice. In: Satyamurthy T V (ed.) *Region, Religion, Caste Gender and Culture in Contemporary India*. Oxford University Press, New Delhi, Vol. 3

Cohen M, Nagel T, Scanlon T (eds.) 1977 *Equality and Preferential Treatment*. Princeton University Press, Princeton, NJ

Das V (ed.) 1990 *Mirrors of Violence*. Oxford University Press, New Delhi

Fiss O M 1977 Groups and the equal protection clause. In: Cohen M, Nagel T, and Scanlon T (eds.) *Equality and Preferential Treatment*. Princeton University Press, Princeton, NJ

Galanter M 1984 *Competing Equalities: Law and the Backward Classes in India*. Oxford University Press, New Delhi

Government of Gujarat 1990 (released in 1991) *Report of the Commission of Enquiry into the Violence in Gujarat Between February 1985 and July 1985*. Government of India, New Delhi

Government of India 1950 *Constituent Assembly Debates*. Government of India, New Delhi, Vols. 7–11

Government of India 1980 *Report of the Backward Classes Commission*. Government of India, New Delhi, Vols. 1–2

Government of India 1980a *Report of the Backward Classes Commission*. Government of India, New Delhi, Vols. 3–7

Kamath A R 1981 Education and social change amongst the scheduled castes and scheduled tribes. *Economic and Political Weekly* **16**

Mohan D 1991 Imitative suicides. *Manushi a Journal About Women and Society* **March–June**

Nozick R 1974 *Anarchy State and Utopia*. Basic Books, New York

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

P. C. Chatterji

Affirmative Action Programs (United States): Cultural Concerns

Even the term itself—'affirmative action'—is hotly contested in the United States today. Opponents generally refer to race-conscious policies designed to benefit nonwhites as 'racial preferences.' Thus, they distinguish 'affirmative action' in the form of aggressive nondiscrimination (its original meaning) from contemporary practices that involve racial double standards: admitting students, for instance, to selective institutions of higher education by different standards depending on their racial or ethnic identity. This article examines the issue from the perspective of two such critics. It reviews the history, and the arguments on both sides of what has become an ugly debate. Although preferential policies have provoked serious controversy in a number of countries, the focus here is exclusively on the United States.

1. The Original Conception

The term 'affirmative action' first entered American public discourse in executive order 10925, issued by President John F. Kennedy shortly after he took office in 1961. The president's order created a new watchdog committee to secure 'equal opportunity in employment by the government or its contractors,' and it demanded that employers engage in 'affirmative action' to secure that end.

The president's directive that companies and the government should treat their employees 'without regard to race, creed, color, or national origin' simply restated the central moral principle that had animated the civil rights movement from before the Civil War to the 1960s. The Constitution is 'color-blind,' John Marshall Harlan had declared in his famous dissent in *Plessy vs. Ferguson*, the 1896 Supreme Court decision that upheld 'separate but equal' railroad accommodations. The Rev. Martin Luther King, Jr., had dreamed of the day when Americans would be judged solely by the 'content of their character,' not 'the color of their skin.' Plaintiffs' attorneys in the 1954 landmark case, *Brown vs. Board of Education* (striking down segregated schooling in the South) had argued that the Constitution 'stripped the state of power to make race or color the basis for governmental action.' Kennedy's executive order seven years later thus reiterated a noble idea that had long been conventional civil rights wisdom.

2. An Altered Vision

Three years later, under the presidency of Lyndon B. Johnson, Congress passed the Civil Rights Act of 1964, which banned discrimination in employment, education, and public accommodations. The act did not include the phrase 'affirmative action'; it rested on the vision earlier articulated by President Kennedy who declared that 'race has no place in American life or law.'

Within a few years, however, the clarity of that moral stance was lost. Civil rights advocates adopted the view most famously stated by Supreme Court Justice Harry Blackmun in 1978. 'In order to get beyond racism we must first take account of race. There is no other way. And in order to treat some persons equally, we must treat them differently.' By the late 1970s, the vaguely Orwellian notion that some persons must be treated 'differently' in order to treat

them 'equally' had become civil rights orthodoxy, and it remains so today.

The revised view sanctioned racial double standards. If established selection procedures resulted in a statistical 'underrepresentation' of blacks, Hispanics, or American Indians in a particular business, profession, or college, they should be revised to remedy the 'imbalance.' Equal opportunity thus became synonymous with equal group results—proportionate racial and ethnic representation in selective schools, in places of employment (public and private), and in the awarding of governmental contracts.

Blackmun's idea was not new. 'You do not take a person who, for years, has been hobbled by chains and liberate him, bring him up to the starting line in a race and then say, "you are free to compete with all the others,"' President Johnson had said in 1965. Opening 'the gates of opportunity' would not suffice; racial 'equality as a fact and as a result' had to be the nation's goal. Although the president did not use the term 'affirmative action,' his image of blacks as crippled by racism laid the foundation for a generation of race-conscious measures designed to ensure 'equality as a fact.' Handicapped citizens were entitled to compete under different rules.

Perhaps this radical departure from the color-blind conception of fairness that had been advocated by liberals for many decades would have met with greater resistance if it had not been proposed on the eve of riots that erupted in the nation's cities in the summer of 1965. The looting, burning, and fighting sent tremors of fear and guilt through white America, and a subsequent official (Kerner Commission) report that purported to explain the disorders set the tone for subsequent civil rights discourse and policy. America was in grave danger of becoming 'two societies—separate and unequal'—the report concluded. It was an invitation to aggressive race-conscious action to ensure equality and thus close the divide.

Given the long and ugly history of American apartheid, the demand for equal results—blacks in the workplace and other settings in proportion to their numbers—was understandable. But it was impossible to square race-conscious measures (amounting inevitably to preferences) with the antidiscrimination language of the 1964 act and the Fourteenth Amendment to the Constitution, which guaranteed 'equal protection' to all Americans. Civil rights warriors in the 1950s and early 1960s had marched and died to rid the country of discriminatory policies; now, in revised form, they were back. Preferences involve discrimination against members of nonfavored groups.

3. *Preferences in Higher Education*

The Supreme Court played an important role in redefining the nation's commitment to civil rights. A series of decisions starting in 1973 tell a startling story of judicial creativity and confusion, culminating in *Regents of the University of California vs. Bakke* five years later. Allan Bakke had been denied admission to the medical school of the University of California at Davis, despite having had an academic record that was vastly superior to those of blacks and Hispanics accepted through a separate race-based admissions process. A deeply divided Court held for Bakke, but allowed the use of race in university admissions as only *one* factor in the equation.

It was a legal standard that was, in effect, a Rorschach test, open to a variety of interpretations. Claiming conformity to the *Bakke* rule, most schools proceeded to admit minority students whose academic records were so weak that they would never have had a chance of admission if they had been white or Asian. At the University of California at Berkeley throughout the 1980s and early 1990s, for example, the whites and Asians who were admitted fell in the top tenth of all test-takers on the SATs, the exam usually taken by high school graduates who hope to be admitted to a selective college or university. The typical black student, on the other hand, had much lower high school grades, and SAT scores at about the national average.

Berkeley was not unique. Starting in the early 1970s, all of the most selective colleges began accepting students by racial double standards. Indeed some places, like the University of Texas Law School, continued to use a separate admissions committee to read color-coded folders for members of different racial and ethnic groups. The picture was kept from public view, however. Tantalizing fragments of evidence only began to trickle out in the 1990s, as a consequence of legal and political challenges to preferential programs, inquiries under state freedom-of-information-act lawsuits, and studies by investigators with special access to the evidence.

For instance, the Center for Equal Opportunity's analysis of data from the University of Virginia reveal black applicants to the college to have had more than a hundred times better chance of admission than white candidates with the same qualifications. In three states (Texas, California, and Washington) a popular referendum or a court ruling has banned racial double standards in public institutions, but they continue to thrive across the nation in other publicly funded settings, as well as in almost all selective private colleges and universities. Thus, such preferences are now deeply embedded in the educational culture.

The fullest empirical study of racial preferences in admissions to elite colleges and universities is a volume by William G. Bowen and Derek Bok(1998), former presidents of Princeton and Harvard respectively. *The Shape of the River* was greeted with uncritical enthusiasm in the liberal press, but its arguments in favor of preferences were deeply flawed. Although the authors claimed to have shown that race was just 'one factor' in admissions decisions, their tables revealed

that affirmative action meant glaring racial double standards, not just a small bonus for belonging to an 'underrepresented' group.

The evidence in *The Shape of the River* supported rather than refuted the criticism that such programs did nothing to benefit the most disadvantaged blacks and Hispanics, whose educational development had been held back by the poor inner-city public schools they attended. The beneficiaries of preferences at the top colleges were from middle-class, often suburban, families.

Bowen and Bok's evidence also punctures the wishful argument that attending an elite school will erase the academic deficits of students admitted under lower standards. The average black student attending schools like Yale and Swarthmore performed well below average in the classroom, ranking in only the 23rd percentile in grade point averages. Even that low figure is deceptively rosy, because it is for all African-American students, although half of those admitted had credentials that merited admission without preferences. The failure to distinguish regular and preferential admits is an astonishing methodological blunder in a study purporting to assess the effects of race-conscious admissions policies.

Bowen and Bok (1998) cite the large numbers of black students accepted to the best law, medical, and other graduate schools as evidence of the success of affirmative action admissions. But a high proportion of those students were again the beneficiaries of lower standards for blacks and Hispanics. The authors ignore a wealth of evidence suggesting that preferentially-admitted black graduate students, like their counterparts in college, rank at the bottom of the class and are much more likely to leave school without a degree.

Most disturbing, those in fields that require passing external competency tests tend to perform dismally. A dismaying 43 percent of all the preferentially-admitted black students who entered an American law school in the Fall of 1991 either failed to earn their degrees or, worse yet, graduated but failed to pass the bar examination within three years of graduating. After costly years of effort, they were thus unable to engage in the practice of law. A study of 1975 medical school graduates found that less than one-third of the minority students who were given heavy preferences in admissions had earned certification in their specialty from the National Board of Medical Examiners seven years after graduation, as compared with 80 percent of Asians and whites and 83 percent of minority graduates with high college grades and good scores on the Medical College Admissions Test.

Preferential admissions to competitive schools were a failure by these objective measures. We lack careful studies of other possible negative effects, but some may be suggested. The common idea that students admitted through affirmative action will have enhanced self-esteem because they are attending an elite school seems highly dubious. Students who do poorly academically do not usually feel good about themselves. Moreover, when most black students have low grade-point averages, racial stereotypes may be reinforced, and all members of the group—whatever their academic talent—may find themselves stigmatized. Perhaps a sense of this danger explains the results of a national survey of college students conducted in 2000. While 84 percent of the respondents said that ethnic diversity on campus was important, 77 percent opposed giving admissions preferences to minority applicants.

4. *Contracting, Employment, and Voting Too*

Preferences are equally embedded in the world of contracting and employment. For instance, since the early 1970s the federal government has given a decided leg-up to minority-owned businesses who submit bids for government contracts to provide goods and services (ranging from paper clips to major construction). A 'small disadvantaged business' (which presumptively means minority-owned) will get the work even if the taxpayers end up footing a bill 10 percent higher than it would have been if a low bid from a white-owned company had been accepted. In theory, only 'small' companies qualify, but the cap on wealth is sufficiently high as to make eligible between 80 and 90 percent of business-owning families in the United States.

In recent years, the US Supreme Court has raised serious questions about such bidding preferences, but those rulings have had little impact. Bill Clinton, as well as many state and municipal governments, have worked hard to circumvent federal court decisions. In 1995 the President promised to 'mend' affirmative action, but in 1998 his administration actually extended to all federal agencies the 10 percent preference given minority firms that had previously been confined to Department of Defense contracts. The Republican-controlled Congress with which he had to work posed no obstacle; it frequently reauthorized preference programs built into legislation governing federal contracts. The rationale behind such preferences is of course a history of discrimination against minority-owned enterprises, but US government spokespersons have, by their own admission, found no discrimination by a federal procurement officer. In any case, the logical sanction is to fire such officers from their jobs.

Employers (private and public) routinely engage in race-conscious decision making, too. The 1964 Civil Rights Act (as amended in 1972) covers all private employers with at least 15 workers. With discrimination redefined to mean disparate impact, employers are vulnerable to government-initiated suits if the process of selecting employees ends up with 'too few' blacks and 'too many' whites (or perhaps Asians).

Litigation is extremely expensive; most businesses would rather create a statistically balanced workforce through the use of racial preferences (when necessary), than tangle with lawyers. Race-based hiring is a means of self-defense.

The Fourteenth Amendment also protects against employment discrimination in the public sector, but neither it nor the 1964 act allows the government to order race-conscious hiring directly. Preferences are explicitly mandated only as remedies upon a finding of discrimination. On the other hand, employers who do business with the federal government—roughly a quarter of the total American workforce—are under quite a different obligation. They are governed by an order issued by President Richard Nixon in 1970 that forthrightly demands race-conscious hiring as a condition of supplying goods and services to federal agencies.

Thus, companies like Raytheon and IBM are expected to assess the availability of black employees and file with the Department of Labor a written affirmative action plan that includes minority hiring goals and a timetable. They must show 'good faith' efforts to reach those targets. If the effort seems insufficient, the department has the power to cancel federal contracts and permanently keep a business (or educational institution) off the list of those eligible to submit bids for government work. It can also recommend further legal action.

The imposition of affirmative action policies upon these employers does not require a judicial finding of unlawful employment practices, nor does the definition of what constitutes appropriate affirmative action depend upon a court. Of course, employers can escape the coercive nature of the executive order by avoiding government contracts altogether, but that is a steep price to pay. For many companies, in fact, the federal government is their best—or only—buyer.

Almost everyone understands that racial preferences have become widespread in contracting, employment, and selective institutions of higher education. Less understood is the degree to which race-conscious policies also affect the enforcement of voting rights. In 1965 Congress passed a landmark Voting Rights Act aimed at enfranchising southern blacks, still denied their basic right to vote. Through the process of implementation, and with congressional acquiescence, that unambiguous aim of enfranchisement was soon altered. The right to vote became an entitlement to black officeholding in proportion to the black population. States became obligated to draw race-conscious districting lines to create a maximum number of safe black seats—in state legislatures, on city councils, school boards, and other elected bodies. Those race-conscious lines would protect black candidates for public office from white competition in much the same way as race-conscious admissions protected black applicants to elite colleges from white and Asian competitors.

Large employers and universities, as well as most members of Congress and other public officials, have either acquiesced or positively embraced the notion that race-conscious policies are good for business and good for America. Indeed, they have gone far beyond the demands of the law in instituting racial preferences. But a majority of the public, surveys indicate, opposes sorting and labeling Americans on the basis of skin color, and beginning in 1989 a majority of justices on the US Supreme Court began to have second thoughts about racial classifications.

Such classifications, the Court said, carry 'the danger of stigmatic harm' and 'may … promote notions of racial inferiority and lead to a politics of racial hostility.' Today, race-conscious public policies are held to a tough constitutional test. They must be narrowly tailored to serve a compelling state interest. What the Court's rulings will mean for the future of race-conscious policies is impossible to predict, however. In three states, the government can no longer discriminate—for good or invidious ends—on the basis of race; in the other 47, preferences are ubiquitous. Moreover, private companies, colleges, and other institutions in every corner of American society still favor blacks and Hispanics over Asians and whites in the interest of a more 'diverse' workforce or student body. And they are likely to continue doing so until, perhaps, demographic change—the lines of race and ethnicity blurred as a consequence of intermarriage—makes all such classifications obsolete.

5. *Arguments For and Against*

Racial preferences are one of the most polarizing issues on the American political scene, and thus few politicians discuss them. But civil rights spokespersons and numerous academics have long exchanged fire in an often ugly war.

The arguments on each side have become familiar. Supporters see race-conscious employment and other policies as essential to keeping rampant white racism in check. In fact, racism is so deeply imbedded in the nation's institutions, they assert, that individual attitudes are quite irrelevant. Thus they view affirmative action as an essential life raft—the means by which blacks and Hispanics stay afloat, the antipoverty, empowerment program that really works. In addition, they believe race-conscious programs are morally just, given the nation's history of slavery, segregation, and rampant discrimination.

Critics, on the other hand, argue that the past cannot be rectified by perpetuating color-conscious policies. There is, in fact, no way of making up for America's terrible racial history. But the past contains a lesson: judging people on the basis of the color of their skin is incompatible with racial equality. If citizens are classified by race and ethnicity, they will not view each other as equal individuals. Pasting racial and ethnic labels on everyone—assuming individuals

are defined first and foremost by the color of their skin—is no way to get beyond race. Racial classifications never worked, and never will. They are as American as apple pie, but today, as in the segregated Jim Crow South, they perpetuate terrible habits of mind. Americans are still viewed as fungible members of a group defined by race—not as unique individuals. And they are treated differently, depending on their group membership. Moreover, blacks, Hispanics, and Asians qualify for group membership on the basis of one drop of blood; the children of a black–white intermarriage are still 'black.'

Critics make a further argument. The notion that groups will be proportionately represented in all walks of life in a racially fair society is fundamentally misguided. Most doughnut shops in Los Angeles are run by Cambodians; East Indian immigrants operate a high percentage of American motels; the stereotype of the Jewish doctor has become a standing ethnic joke. The division of labor along ethnic lines is a common phenomenon in all ethnically diverse settings.

Racial inequality in America remains real, but most beneficiaries of racial preferences are already on the road to success. They are not from the black underclass, and admitting middle-class black students to Yale or Berkeley under lower standards does not address the problem of a disproportionately high percentage of black males still unemployed in a full-employment economy. Inner-city students generally lack the academic skills to compete for classroom seats in elite schools, or for well-paying middle-class jobs—only better education in the earlier years can address that problem. Indeed, poorly educated students do not make up for their lack of academic skills by being placed in a highly competitive environment; black students at elite schools, on average, end up in the bottom quarter of their class. That fact alone carries the danger of perpetuating a pernicious racial stereotype.

Racist Americans have long said to blacks, the single most important thing about you is the color of your skin. In recent years, black and white Americans of seeming good will have joined together in saying, we agree. It has been—and is—exactly the wrong foundation, these authors believe, on which to come together for a better future. Ultimately, black social and economic progress largely depend on the sense that we are one nation—that we sink or swim together, that black poverty impoverishes us all, that black alienation eats at the nation's soul, and that black isolation simply cannot work.

See also: Sex Differences in Pay

Bibliography

Belz H 1991 *Equality Transformed: A Quarter-Century of Affirmative Action. Transaction*, New Brunswick, NJ

Bloch F 1994 *Antidiscrimination Law and Minority Employment: Recruitment Practices and Regulatory Constraints*. University of Chicago Press, Chicago

Bowen W, Bok D 1998 *The Shape of the River: Long-term Consequences of Considering Race in College and University Admissions*. Princeton University Press, Princeton, NJ

Carter S 1991 *Reflections of an Affirmative Action Baby*. Basic Books, New York

Crosby F, Van De Veer C (eds.) 2000 *Sex, Race, and Merit: Debating Affirmative Action in Education and Employment*. University of Michigan Press, Ann Arbor, MI

Graham H D 1990 *The Civil Rights Era: Origins and Development of National Policy*. Oxford University Press, NY

Keith S et al. 1987 *Assessing the Outcome of Affirmative Action in Medical Schools: A Study of the Class of 1975*. The Rand Corporation, Santa Monica, CA

Kull A 1992 *The Color-Blind Constitution*. Harvard University Press, Cambridge, MA

McWhorter J H 2000 *Losing the Race: Self-Sabotage in Black America*. The Free Press, New York

Nieli R (ed.) 1991 *Racial Preferences and Racial Justice: The New Affirmative Action Controversy*. Ethics and Public Policy Center, Washington, DC

Sowell T 1990 *Preferential Policies: An International Perspective*. William Morrow, NY

Steele S 1990 *The Content of Our Character: A New Vision of Race in America*, 1st edn. St. Martin's, New York

Steele S 1998 *A Dream Deferred: The Second Betrayal of Black Freedom in America*. HarperCollins, New York

Thernstrom A 1987 *Whose Votes Count? Affirmative Action and Minority Voting Rights*. Harvard University Press, Cambridge, MA

Thernstrom A, Thernstrom S (eds.) 2000 *Beyond the Color Line: New Perspectives on Race and Ethnicity*. Hoover Institution Press, Stanford, CA

Thernstrom S, Thernstrom A 1997 *America in Black and White: One Nation, Indivisible*. Simon and Schuster, New York

Thernstrom S, Thernstrom A 1999 Reflections on *The Shape of the River*. *UCLA Law Review* **46**: 1583–1631

A. Thernstrom and S. Thernstrom

Affirmative Action, Sociology of

The phrase 'affirmative action' first became firmly associated with civil rights enforcement in 1961, the year President Kennedy directed federal contractors to take 'affirmative action' to ensure nondiscrimination in hiring, promotions, and all other areas of private employment. Over time, federal goals began to shift away from 'soft' affirmative action programs that merely required equal opportunity for members of previously excluded groups towards stronger policies mandating preferential treatment of women and minorities in order to obtain equal (or proportional) results. The shift in emphasis from 'weak' to 'strong' methods of policy enforcement emerged as many policy makers concluded that nondiscrimination alone was not sufficient to address the deep racial divisions

and inequalities that beset American society. The many different meanings and forms that affirmative action has taken on through the years create difficulties in measuring its public support and explain, in part, its tortured legal standing and questionable future.

1. *Defining the Concept of Affirmative Action*

Affirmative action involves a range of governmental and private initiatives that offer preferential treatment to members of designated racial or ethnic minority groups (or to other groups thought to be disadvantaged), usually as a means of compensating them for the effects of past and present discrimination. Justification for affirmative action programs typically rests on a compensatory rationale, i.e., members of groups previously disadvantaged are now to receive the just compensation which is their due in order to make it easier for them to get along in the world. However, other useful definitions and characterizations of affirmative action de-emphasize the retrospective, compensatory, and ameliorative nature of such programs and focus instead on the current value of such programs in enhancing diversity, particularly in educational institutions and in the workforce.

The actual programs that come under the general heading of affirmative action are a diverse lot, and can include policies affecting: (a) admissions to educational institutions; (b) public and private employment; (c) government contracting; (d) the disbursement of scholarships and grants; (e) legislative districting; and (f) jury selection. Innumerable affirmative action programs have been enacted into law at the federal, state, and local level, and many private corporations and universities have developed affirmative action programs on their own voluntary initiative. Methods of implementing affirmative action policies are similarly diverse and in the past have ranged from 'hard quotas' to softer methods of outreach, recruitment, and enforcement of anti-discrimination norms.

2. *Measuring Public Support for Affirmation Action Policy*

After years of focusing on polarization between white Americans and African-Americans, survey researchers have begun to realize that public opinion on affirmative action is highly sensitive to question wording and question context in addition to being plagued by respondent misinformation. As a result, respondents' answers to direct questions about their support or opposition to affirmative action tells us very little about the types of public policies that a given individual will endorse. In fact, one researcher has observed that respondents who say that they oppose affirmative action policies may actually support more types of affirmative action programs than a person who identifies as an affirmative action supporter. Greater awareness of the sensitivity of affirmative action questions to question context and question wording has led some researchers to conclude that validity of survey results could be greatly improved if the use of the term 'affirmative action' was abandoned and instead the content of specific policies described. When a range of survey questions is examined carefully, it is found that Americans seem to be moving towards greater on consensus on affirmative action related issues. This consensus includes a shared unease with racial preference programs coupled with a willingness to support outreach programs that benefit the economically disadvantaged regardless of race

3. *The Legal Standing of Affirmative Action*

While the shift from weak to strong methods of policy enforcement occurred in the late 1960s and was largely the result of decisions made in the executive branch of the federal government, the Supreme Court has played a crucial, if somewhat more equivocal, role in the period which followed by legitimating some of these policy changes and restricting others.

3.1 *Preferences in Hiring*

Title VII of the Civil Rights Act of 1964 is a statutory measure designed to combat racial discrimination in employment situations. Charges of 'reverse discrimination' became common during the 1970s as more and more corporations and private businesses, often under pressure from federal enforcement agencies, began more aggressive hiring of minorities and women. The Court ruled unanimously in *McDonald vs. Sante Fe Transportation Company*, 427 US 273 (1976), that whites as well as blacks are protected from racial discrimination under the antidiscrimination provisions of Title VII. Despite this ruling, a number of subsequent court decisions have held that Title VII permits the preferential treatment of minorities and women in hiring and promotion decisions (but not in decisions affecting layoffs) if such treatment is part of an affirmative action plan designed to increase the employment of previously excluded or under-represented groups (see *United Steelworkers of America vs. Weber*, 443 US 193 (1976), *Local 28 Sheet Metal Workers International Association vs. Equal Employment Opportunity Commission*, 478 US 421 (1986), *United States vs. Paradise*, 480 US 149 (1987), and *Johnson vs. Transportation Agency, Santa Clara County*, 480 US 616 (1987)). Justice Brennan writing for the majority in Weber explained 'It would be ironic indeed if a law triggered by a nation's concern over centuries of racial injustice and intended to improve

the lot of those who had been excluded from the American dream for so long constituted the first legislative prohibition of all voluntary, private, race-conscious efforts to abolish traditional patterns of racial segregation and hierarchy.' However, the Court's subsequent support for affirmative action is quite fragile as shown by the many 5:4 decisions, the restrictions placed on affirmative action programs with regard to layoffs in *Firefighters Local Union No. 1794 vs. Stotts*, 467 US 561 (1984), and *Wygant vs. Jackson Board of Education*, 476 US 267 (1986), and the changing composition of the Court.

3.2 Set-asides

One form of affirmative action preference that became very popular among state and municipal governments in the second half of the 1970s was the minority contracting set-aside. Set-aside programs usually involve the reservation of a fixed proportion of public contracting dollars that by law must be spent on the purchase of goods and services provided by minority-owned businesses. Like preferences in hiring, set-asides have been enormously controversial and cries of reverse discrimination abound. The Supreme Court first took up the issue of set-asides in the case of *Fullilove vs. Klutznick*, 448 US 448 (1980), in which it held that a federal set-aside law did not violate the equal protection provisions of the federal Constitution because it was a legitimate remedy for the present competitive disadvantages of minority firms resulting from past illegal discrimination. However, nine years later and again reflecting the influence of the Reagan-era appointees, the Court held in *Richmond City vs. J. A. Croson Co.*, 488 US 469 (1989), that racial classifications within state and local set-aside programs were inherently suspect and were to be subject to the most searching standard of constitutional review ('strict scrutiny') under the equal protection provisions of the Fourteenth Amendment. Six years after Croson, the Court extended strict scrutiny review to federal affirmative action programs that draw racial classifications (see *Adarand Contractors Inc. vs. Pena* 115 S.Ct. 2097 (1995)). These Court decisions have curtailed minority set-aside and led to greater efforts to address concerns about over-inclusiveness in the protected categories.

3.3 Preferences Higher Education

Equally important—and equally controversial—have been affirmative action policies adopted by educational institutions. Beginning in the late 1960s, many universities and professional schools began admitting minority students, particularly African-Americans and Hispanics, with substantially lower grades and lower scores on standardized tests than white students. Some nonadmitted white students charged reverse discrimination and a few brought suit in federal court claiming that affirmative action in higher education was a violation of Title VI of the 1964 Civil Rights Act, as well as of the equal protection provisions of the US Constitution. In *Regents of the University of California vs. Bakke*, 438 US 265 (1978), the Supreme Court ruled against an explicit quota system but allowed admissions officers to take race into account as one of many 'plus' factors designed to enhance the diversity of a school's student body. Affirmative action in higher education has come under increasing attack in the 1990s. In *Hopwood vs. Texas*, 78 F.3d 932 (5th Cir. 1995), the Fifth Circuit questioned the vitality of the Bakke decision and upheld a claim of racial discrimination brought by nonadmitted white students who had higher test scores and grades than admitted minority students. Two challenges to the University of Michigan's admissions policy, which takes the applicant's race into account as one of many factors bearing on admissibility, are currently working their way through the legal system and are likely to provide the vehicle for the Supreme Court to revisit the Bakke decision. In 1998, *Gratz and Hamacher vs. University of Michigan and Grutter vs. Michigan*, were combined into a class-action suit which is awaiting resolution (www.wdn.com/cir/mich1.html). These higher education cases are expected to reach the Supreme Court sometime during the twenty-first century.

4. The Outlook for the Future of Affirmative Action

Affirmative action in the US faces an uncertain future. Changes are occurring at the state and local level, as well as in the federal court system. Two states have voted to ban state-supported affirmative action programs. In November 1996, California voters approved Proposition 209, by a vote of 54 to 46 percent. This initiative provided that the state 'shall not discriminate against or grant preferences to any individual or group on the basis of race, sex, color, ethnicity or national origin in the operation of public employment, public education, or public contracting.' Similarly, in 1998 voters of Washington State, by a vote of 59 to 41 percent, passed an initiative also banning affirmative action in state-supported programs. Nevertheless, a ballot initiative in Houston, Texas failed after pro-affirmative action forces were able to control the wording of the initiative so that voters voted against dismantling affirmative action programs and for banning state supported 'discrimination.' The language used in these three initiatives seems to have been a factor in explaining the different political outcomes. Affirmative action, therefore, appears to be a vul-

nerable public policy. Changing demographic patterns could further decrease its public support, since racial and ethnic minorities are expected to reach majority status sometime during the twenty-first century. Already, a racial divide exists in the perception of the extent and nature of racial discrimination and this seems to be an important factor in explaining contemporary attitudes towards the policy. Perhaps, the perceptual gap could be narrowed by studies that identify and expose hidden racism and discrimination in housing, employment, police actions, and college admissions. Such studies could serve to heighten public awareness of the pervasiveness of discrimination and lead to greater acceptance of public policy remedies designed to help ameliorate the kinds of disparities that affirmative action was originally designed to address. It is race-based and not class-based affirmative action that seems to be most vulnerable to constitutional based challenges.

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on its Effectiveness; Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Civil Rights; Civil Rights Movement, The; Discrimination; Discrimination, Economics of; Discrimination: Racial; Gender and the Law; Gender, Economics of; Race and the Law; Racial Relations; Racism, Sociology of

Bibliography

Graham H D 1990 *The Civil Rights Era: Origins and Development of National Policy, 1960–1972*. Oxford University Press, New York

Gamson W A, Modigliani A 1987 The changing culture of affirmative Action. *Research in Political Sociology* **3**: 137–77

Skrentny J D 1996 *The Ironies of Affirmative Action: Politics, Culture, and Justice in America*. University of Chicago Press, Chicago

Steeh C, Krysan M 1996 Poll trends: Affirmative action and the public, 1971–1995. *Public Opinion Quarterly* **60**: 128–58

Swain C M 2001 Affirmative action, legislative history, judicial interpretations, public consensus . In: Smelser N, Wilson W J, Mitchel F (eds.) *Racial Trends and Their Consequences*. National Academy Press, Washington, DC

C. M. Swain

African Legal Systems

'African legal systems' means here the bodies of interrelated legal norms and accompanying institutions of norm-creation, norm-finding, and norm-enforcement which have a social existence in Africa. Legal norms are taken to be those social norms which are enforced by a relatively strong degree of coercion.

1. Customary and Religious (Non-state) Legal Systems

Customary legal systems are those systems which exist by virtue of the social observance of their norms, and not by the creation of their norms through state institutional processes such as the enactment of legislation. Customary legal norms are observed because of a continuing but usually tacit agreement among a population to accept them as obligatory. For the present purpose a practice need not have been observed for a very long period to be 'customary'. (See also: *Folk, Indigenous, and Customary Law*.)

Religious legal systems are for the present purpose a variety of customary legal systems, their distinctive feature being their derivation from a system of religious belief. In Africa the primary systems of religious law are varieties of Islamic law. There are many communities, especially in North Africa and in the Sahel region, where the predominant law is Shari'a, sometimes modified under the influence of other local customary norms.

These non-state African legal systems vary widely, and almost any generalization about them is subject to exceptions.

Perhaps the most distinctive feature of African customary legal systems is the frequency with which the parties to legal relations are communities of ascribed membership, especially membership by descent, not individuals or other corporate persons as in Western legal systems. Thus, for example, substantial interests in land are often vested in lineages or communities comprised of several lineages. Marriage is generally contracted by agreement between the lineages to which the bride and groom belong, and creates legal relations between them. A customary law community often has an individual leader, instances of which range from the 'head of family' of the relatively small lineage to the 'chief' or 'king' of a large polity such as Asante or Buganda, although there are also many acephalous communities (Middleton and Tait 1958). Leaders are determined according to customary legal norms. Although in some circumstances a charismatic leader may temporarily acquire personal authority and wield great discretionary power, generally customary law imposes strict limits on the powers of leaders: the rule of customary law prevails widely.

Another common feature of customary legal systems is the nature of the procedures and principles concerning conflicts of interests and disputes. These are directed to the achievement of social peace and harmony within the community rather than the determination of legal rights. Dispute processes are less frequently instances of adjudication than of

mediation or negotiation, in which there are social pressures on the parties to compromise.

2. *State Legal Systems*

The modern African state with its governmental and legal institutions is a product of colonization. In most parts of Africa there was relatively little immigration from the colonizing countries, and everywhere the indigenous inhabitants remained in the majority. Nevertheless the colonial powers set up systems of government which resembled those of the metropolitan states, except that they did not provide for public participation in government. The English common law was imported into the British colonies. Codes largely identical with those of France and Portugal were enacted in their colonies. (Allott 1960, 1970; see also: *Law: Imposition, Reception, and Colonial.*)

The received laws shaped the institutions of the state legal systems, such as the courts, and provided bodies of legal norms law in fields such as contract, property transactions, and personal injuries. The technical and social knowledge necessary to administer this law in such roles as those of judge, legal advisor, legislator, and police officer were initially lacking in the indigenous populations. The lacuna was filled by colonial officers brought from the metropolis, some of them members of the professional legal community and others with some training and socialization in the ways of that community. Members of indigenous communities later obtained the training and assumed the social practices which enabled them to take over these roles. Thus the laws of the colonial powers became African legal systems. At Independence (between the late 1950s and the end of the 1970s) African rulers and administrators saw their aspirations and material interests as dependent on the continued effectiveness of these legal systems.

Those customary and religious institutions and practices which the colonial powers perceived to be inconsistent with colonial domination, such as primary allegiance to chiefs, they attempted to destroy or transform. Their policy towards other components of these legal systems was to tolerate or even encourage their continuance as alternatives to those of received law. In consequence, in the social field of each colony both the state legal system and one or more customary or religious legal systems were observed. In this form of legal pluralism there was no unified hierarchy of norms. Individuals differed according to whether they gave general priority to the observance of state law rather than a non-state law, or gave priority to each on different occasions.

Within state law a further type of legal pluralism was formed. State legal systems gave recognition to African indigenous customary and religious legal systems. 'Recognition' here designates the policy whereby state legal systems treat the institutions or norms of customary and religious law as parts of the law of the state, giving effect to them and enforcing them in the same way as the institutions and norms of received law. Policies of recognition were adopted in some British colonies from the beginning, and eventually adopted everywhere.

Institutional recognition of a customary or religious legal system occurred when institutions of that system were incorporated into a state legal system, as for example when chiefs became administrative officials or judges of the state. This frequently happened in British colonies, in accordance with the policy of indirect rule which sought to govern colonies through native forms of government. With the abandonment of the policy from the 1940s this recognition became less significant, but it has continued in a limited form. In other colonies it was adopted as a concession to local opinion, as well as for practical reasons; there also it has continued since Independence.

Normative recognition of a customary or religious legal system occurred when norms of that system were incorporated into the body of norms of the state legal system. State law might provide, for example, that rights in land could be transferred by procedures specified by norms of customary law, or that the inheritance of property might be determined by customary and religious norms. The incorporated norms then became enforceable in state courts. This also occurred in most African colonial legal systems, and also has continued to the present.

Customary and religious laws were not and indeed could not be 'incorporated' into state law without radical changes to their nature and content. The state has always excluded from incorporation some portions of customary laws, such as those providing for slavery, which were contrary to the fundamental values of the state. Moreover, the acceptable elements of customary and religious legal systems could not be accommodated without reformulation. The institution of chieftaincy, for example, is transformed when its authority ceases to be derived from respect for tradition and communal identity, and becomes based upon the threat of coercion by state institutions. The forms of coercion of state institutions, such as the threat of imprisonment supporting an adjudicative decision, differ from the social pressure traditionally exerted upon disputing parties to agree to a compromise. Consequently the effect of normative recognition has generally been to compel behavior of a different type from that which would otherwise have occurred. Finally, the personnel of state institutions are frequently unfamiliar with customary or religious law and have 'recognized' norms which differed from those socially observed. Thus state institutions create new bodies of norms, 'official,' 'judicial,' or 'lawyers,' customary laws,' which differ from the 'people's,' 'folk,' 'indigenous,' or 'practiced customary laws' which have continued to be observed outside state

administrations. Some legal historians consider lawyers' customary law to be an 'invention,' bearing no significant relationship to pre-existing social norms (e.g., Chanock 1985), although others consider this extreme conclusion to be unrealistic (e.g., Woodman 1985).

3. *Past and Present Trends in Legal Development*

Legal development since the institution of the colonial state has reflected the considerable political and social change in Africa.

Changes in state public law, that is, those branches of the state legal system which regulate government, have generally been directed towards strengthening the state and elaborating the functions and powers of state executive bodies. From the late 1940s they also provided for increasing public participation in government. The Independence constitutions contained rules for democracy and constitutional government. In the immediately following decades many states instituted one-party government, or experienced the violent overthrow of their constitutions followed by military rule. More recently internal developments and external, globalizing pressures seem to have produced trends towards constitutionalism, more democracy and accountability in government, and emphasis on environmental protection. These forces, and notably the formation of the African Charter on Human and People's Rights in 1986, have produced constitutional recognition of a wide range of human rights through judicially enforceable legal provisions.

Nation-building has been a major objective of constitutional orders. The boundaries of modern African states, inherited from the colonial powers, contain different ethnic and religious groups, each with its own culture, customary law, and language. The individual citizen's loyalty and sense of belonging often refers primarily to the ethnic or religious group. Constitutional provisions, in an effort to create a sense of national identity, have prohibited discrimination on the ground of ethnicity, language, or religion in many activities, including the selection of members of state institutions and political organisations. This has been seen as necessary to economic and social development, as has the principle of indigenization of positions of economic or social power. There have been increasingly determined attempts to improve the legal position of women, who are seen as seriously disadvantaged by traditional laws and practices, and of children.

Recognition in state criminal law of customary law has never been extensive. Today the criminal law is contained exclusively in state criminal codes, together with other written laws specifying minor offences.

Generally developments in public law have increased the function of state law as a regulator of social behavior. All of the trends just noted, and the long-term tendency towards Western-style governmental systems have been accompanied by a decline in the African state's use of customary institutions such as chieftaincy, and so in the strength of these institutions. Nevertheless, the state is still not the sole, unchallenged legal authority. It remains notoriously ineffective in regulating the daily lives of its citizens outside its own bureaucratic institutions. Even within these, instances of corruption and abuse of power show that state officials are not punctilious observers of state law. One global policy may even have reduced the scope of state law. The structural adjustment policies which the World Bank and International Monetary Fund have set as conditions for economic aid have tended to reduce state involvement in economic activity in favor of private business institutions.

In the fields of private law changes are visible in the content of received law, practiced customary law, and lawyers' customary law, and in the relationships between them.

Received private law has been continuously changing both in the European countries of origin and in African states. State laws required received law to be adapted to local circumstances by the courts and legislatures. But decision makers have also asserted the value of transnational uniformity in the principles of the common law and civil law. Decisions and textbooks from the source countries of received laws have been followed in African jurisdictions. Even in the legislative reform of received law the best legislative model is often held to be the latest legislation in the country of origin, in fields as diverse as company law and divorce law. Only in some of the legislation concerned with economic development, such as that regulating foreign investment, has there been significant innovation.

Some aspects of social change have increased the extent to which received law is observed. Policies of 'modernization' have been accompanied by a general official belief that the application of Western laws gives effect to the principles of the market economy. (See *Law and Development*.) To the same effect has been the growing involvement in the global economy of private individuals and businesses through the export-oriented development of cash crops and extractive industries. There has been much expansion in the education and technology derived from the West. The same tendencies have produced and are in turn reinforced by the emergence and growth of national legal professions. Nevertheless, in activities such as the contracting of marriage, the acquisition and use of lineage assets, and land use, most people still choose to act under customary law, especially when the parties involved all observe the same customary law, or have relatively little Western education.

Non-state African legal systems have never been static. Systems of practiced customary law have changed considerably since colonization. The individual has gained more autonomy, having today the

capacity to hold extensive property and enter into weighty transactions independently of the lineage. Legal transactions are more often governed today by market forces, rather than by standard customary terms. Written records are more commonly used, and more norms have been developed to guide relations between members of communities and outsiders. New bodies of customary law have been formed, especially in the vastly expanded urban areas where religious, professional, welfare, and self-help communities with their own customary laws have been created.

Lawyers' customary law has also changed, in that norms which the state will recognize and enforce as customary law have been progressively elucidated and embodied in judicial decisions, restatements, and textbooks. To some degree, the reformulation entailed in these processes has been in the same direction as changes in practiced customary laws. However, the processes have also had a conservative effect, because a norm embodied in an authoritative pronouncement is not easily changed thereafter in response to changes in social circumstances. The gap between lawyers' customary law and practiced customary law could thus grow. Another major development in lawyers' customary law has been the extensive amendment to some areas by legislation. Land laws in particular have been amended in attempts to balance competing claims to land use while also promoting economic development and inhibiting environmental damage. Here the extent to which practiced customary law diverges from state law is a function of the extent to which the state is able to make its legislation effective.

4. *Current and Future Issues in Legal Development, Theory, and Research*

Current law-making by states is directed towards a number of practical problems. Everywhere there is a search for constitutional forms which will maintain stability against interethnic and interreligious conflict, and armed uprisings. The relationship between customary and religious legal systems and state law has become an even more acute issue since the constitutional entrenchment of human rights, since these appear incompatible with some non-state laws. Globalization intensifies certain issues. State law may be developed to protect national economic independence against the power of international businesses and institutions such as the World Trade Organisation. It may also assist economic development in the face of new threats from natural disasters such as climatic change and the AIDS epidemic.

Research into African legal systems has in the past been carried on within several disciplines. The study of customary law has used the methods of anthropology and sociology, of state law those of legal analysis, and of Islamic law those established specifically within that scholarly tradition. Today there is more interdisciplinary understanding. One consequence is renewed concern with fundamental questions about the concept of law, especially that as to whether customary and religious normative orders are truly law. These debates may also raise doubts as to whether it is appropriate to speak of 'systems' of law.

With indigenization and expansion of the African civil service and academia, most research into African legal systems will in future be conducted in Africa by African scholars. It will be closely concerned with the practical problems just mentioned. However, it is likely that the more fundamental theoretical issues will continue to inspire and to be investigated by an international community of scholars.

See also: African Studies: Culture; African Studies: History; African Studies: Politics; African Studies: Religion; Central Africa: Sociocultural Aspects; Colonization and Colonialism, History of; East Africa: Sociocultural Aspects; Folk, Indigenous, and Customary Law; Legal Pluralism; Legal Systems, Classification of; Middle East and North Africa: Sociocultural Aspects; Postcolonial Law; Southern Africa: Sociocultural Aspects; West Africa: Sociocultural Aspects

Bibliography

Allott A N 1960 *Essays in African Law: With Special Reference to the Law of Ghana*. Butterworths, London

Allott A N 1970 *New Essays in African Law*. Butterworths, London

Chanock M 1985 *Law, Custom and Social Order: The Colonial Experience in Malawi and Zambia*. Cambridge University Press, Cambridge, UK

Doucet M, Vanderlinden J (eds.) 1994 *La Réception des Systèmes Juridiques: Implantation et Destin* [The Reception of Legal Systems: Implantation and Development]. Bruylant, Brussels, Belgium

Elias T O 1956 *The Nature of African Customary Law*. Manchester University Press, Manchester, UK

Gyandoh S O (ed.) 1988 Building Constitutional Orders in Sub-Saharan Africa. Third World Legal Studies

Journal of African Law 1957. Oxford University Press for The School of Oriental and African Studies, University of London

Middleton J, Tait D (eds.) 1958 *Tribes Without Rulers: Studies in African Segmentary Systems*. Routledge & Kegan Paul, London

Reyntjens F (ed.) 1989 *Pluralism Participation and Decentralization in Sub-Saharan Africa*. Third World Legal Studies

Van Rouveroy van Nieuwaal E A B, Ray D I 1996 The new relevance of traditional authorities to Africa's future: Special Double Issue. *Journal of Legal Pluralism* 37–8

Vanderlinden J 1983 *Les Systèmes Juridiques Africains* [African Legal Systems]. Presses Universitaires de France, Paris

Woodman G R 1985 Customary law, state courts, and the notion of institutionalization of norms in Ghana and Nigeria.

In: Allott A, Woodman G R (eds.) *People's Law and State Law: The Bellagio Papers*. Foris, Dordrecht, Netherlands
Woodman G R, Obilade A O (eds.) 1995 *African Law and Legal Theory*. Dartmouth, Aldershot, UK

G. R. Woodman

African Studies: Culture

1. Introduction

1.1 Introductory Paragraph

The concept of culture is one of the most disputed in the social sciences, and nowhere more so than in African studies. The classic study of culture rested on the study of the main media of expression: music, art, material forms, and (above all) language and the written expression of thought in literature, philosophy, and theology. Texts were central. In the foundational era of African Studies, in the nineteenth century in Europe, Africa was assumed to have no texts at all. Its arts of music, performance, and oral poetics were thought to be different in kind from text-based culture. They were performative and multimedia rather than abstract and specialized. Hence their designation as 'primitive,' which survives in many museums and art galleries. The basic collections were, therefore, of sculptural art and material culture, which were relatively accessible even to amateur collectors, and which were assembled during the competitive rush for acquisitions by museums in the late nineteenth and early twentieth centuries. Even for specialists in the study of 'primitive cultures,' however, techniques for reproducing their nonwritten cultural forms—such as tape-recording, photography, and film—were still too clumsy and fragile at that time to produce a solid and reliable corpus for study. The analysis of African cultures has grown and changed, therefore, with changes in thinking about culture in general, in the technologies of scholarship, and in the place of Africa in the world. These changes are discussed in general and then their enduring contributions are summarized.

1.2 Historical Overview

Since the late nineteenth century in the systematic study of African culture, scholars and artists have struggled to rework the foundational legacy as new study techniques, approaches, and colleagues appeared on the scene. Three major changes have erased the categorical distinction between literate and oral/performative culture. First, new recording technologies have made performative culture amenable to the kinds of analysis given to texts in the past, and have made primary sources more widely available for secondary analyses. Second, and at the same time, modern Cultural Studies in Europe has moved closer to African Studies by embracing a much wider range of arts and performance than in the past, including multimedia work, and by looking at how different media can draw on the same forms and themes. Still using textual methods, they now include popular culture, film, clothing, and other expressive forms, and they link style, substance, and concepts of authorship to social context far more than was done in the classical mode. Finally, the African library of written texts has expanded greatly in the twentieth century. Far more written forms from the past are now recognized than were acknowledged in the classic period, including voluminous sources written in Arabic script and certain sculptural, geometric, and musical forms that are now realized to encode verbal content (e.g., Roberts and Roberts 1996). The other source is current artistic production. There is a new corpus of literature, music, and the arts that is not 'traditional,' but rather responds to the situation of modern life, including the global marketplace. The new expansion in Africa includes African scholars' studies of their own past and present art forms, working from their own languages, and this has contributed new expertise and a new set of ideas to the international arena.

These changes from the classic, classificatory, study of culture to the current expansive incorporation of varied and linked forms are still being worked out and they are still contentious. At the same time, the sheer enormity of the task of adequate study of African cultures is realized increasingly. Only in the European view of the confident, progressive industrial era could the whole continent and its millennia of history be grouped together into a single category as 'Africa.' Africa is the cradle of humanity; its cultures have very deep histories. Its population is also highly diverse. There are about 1,000 African languages spoken. More than was originally understood also benefit from written works. Even the popular icons of Africa that have been studied for decades—such as masks, body decoration, 'fetishes,' and drumming—are far richer as traditions than the scholarship can yet do justice to.

With these provisos about the completeness of the record, the history of cultural study can be divided schematically into four broad successive phases. The first comprises the early missionary and traveler documentation, up to the mid-nineteenth century. The first systematic efforts at documentation (a second phase, from the late nineteenth century up to World War I) were carried out in the natural history style. Nonliterate cultures were studied by collecting cultural items, and describing and classifying them according to the taxonomy of the collector. The third phase was anthropological, and lasted more or less unchallenged from about 1920 to about 1960 (the time of political independence from European colonial rule). It was devoted to the idea of cultures as coherent systems of thought and representation, where all the elements of

the life ways of a community (judged either by their common language or common membership in a political community) could be related to each other. The fourth and present phase (from about 1960 and continuing) illuminates cultural themes and techniques by drawing on the classic and modern humanities. Here culture is seen from within, as a field of imagination and debate. There have been major achievements in all four phases, each of which has posed key questions for comparative knowledge, general theory, and public debate.

2. Phases in African Cultural Studies

2.1 Missionary Understandings and Older Sources

The idea that Africa and Europe 'framed' the human experience—initiated and culminated it, expressed two ends of a spectrum of biological and cultural experience—predates the disciplinary study of the continent and has been extremely persistent in Western thought. Evidence about Africa from before the nineteenth century is both sparse and problematic. Nevertheless, these sources set agendas. Writings from that period, by Europeans in their own languages, are now seen as having 'invented' (Mudimbe 1988) an Africa that is fundamentally contrastive with the West. This conceptualization still informs the general Western lexicon about Africa, but it has been repeatedly challenged by scholars.

African philosophers and scholars outside of the European mainstream have insisted on a return to older sources, to question the basis for the contrastive model. Most prominent has been the Senegalese thinker Diop (1955), who argued for the Egyptian origins, pan-African connections and historical unity of African cultures, as against a 'primitive' and racialized conception. Painstaking scholarship in several disciplines continues, linking evidence from the old sources to new findings on, for example, the long history of urban cultures, monies, and trade, and the use of Arabic script. Archaeology and paleontology examine yet earlier traces of cultural life, not only through fossil remains (for the very distant past) but also in rock paintings, house and burial sites, the remains of iron and copper works, and so on.

It was, however, the prevailing contrastive model that was pushed in new directions by the first disciplinary scholarship about African culture.

2.2 Culture and Natural History

A passion for creating encompassing classifications of all of life gained momentum in European intellectual circles during the nineteenth century. Other cultures and their attributes were seen as part of the phenomenal world, to be treated similarly to other life forms. The actual process of collecting involved human and political encounters of a depth, complexity, and problematic nature that is still being studied (Fabian 2000). Some objects were seized during the punitive expeditions of colonial conquest from the 1880s until about 1910; some were accumulated during Christian conversion and the outlawing of the ritual cycles in which they were used; others were acquired by organized museum expeditions, where the items were bargained over and paid for. Hundreds of thousands of pieces—some of great artistic value in the West and religious value to the populations themselves, and others of technical interest only—arrived by one or other of these routes in American and European museums, where the vast majority remain. Some pieces have been missed gravely; one or two have been returned. To their lasting credit, however, some of the collectors were attuned to the breadth and completeness of their collections and to the originality of the material so that they remain a spectacular resource for scholars from all nations and fields. Some collectors also kept acquisition records, which can still be gleaned for contextual understanding of both the collecting and the objects themselves (Schildkrout and Keim 1998). And a few, such as the missionaries George Schwab (1947) with George Harley, wrote books about the cultural lives of the people who created and used the objects.

The emphasis was on material culture, with some notable photography of items in use and of daily life in general. It would be hard to exaggerate the amount and the variety of the African material culture in Western museums: everything from divination trays to fishtraps, enough spears to arm small militias, carved spoons, blacksmiths' tools, indigenous currencies, cloth, talismans, and 'fetishes.' Culture, in the natural history mode of thinking, was everything that people produced by virtue of learning, from the most intricately symbolic to the most pragmatically functional. The very large number of works that could be classified as 'art' had a further destination than the natural history museum. Some went into an art market that made its own selective choices of value. There are African works from this period that are now considered masterpieces, valued in the hundreds of thousands of dollars. The asymmetric geometry, the nonnatural figurative representations, the brilliant colors, and the general vitality of these pieces were a major influence on modernism in European art. The power of modernism, in turn, fixed these particular pieces in the mind of the public as icons of African aesthetics.

The recognizability of African icons has turned out to be a very mixed legacy because it has made innovation more of a struggle. During the museum era, the icon of African music became drumming, of art became figurative sculpture and especially masks, of architecture became what is still referred to in popular parlance as the 'hut,' of body decoration,

scarification, and so on. The following phase of cultural study involved an effort to break out of this straightjacket, and above all to place the elements back into the context of life-as-lived. Culture was not just a product of learning, but a means of articulating a world view.

2.3 Culture in Anthropology

The era between World War I and political independence, about 1920 to 1960, was the period of effective colonial rule in Africa. During this time, conditions were settled enough for scientists to stay for long phases of their careers, in some places setting up permanent research institutions. By this time, anthropology had become committed to studying societies and cultures as wholes. Where political and cultural lives were lived through oral and performative media, this meant long-term residence with them: to collect the texts, case studies, genealogies, maps, calendrics, and so on from which such an integrated model could be deduced. Political and medical conditions allowed, and intellectual aspirations promoted, the practice of field research, usually carried out for at least a year.

If cultures are conceptualized as integrated wholes then the key to their understanding must be the general concepts that animate and link the separate elements, domains, and performances. So this era of cultural study moved away from the objects, to place greater emphasis on ideas, symbols, and general principles. Particular emphasis was given to the ideas that motivate social life rather than those expressing aesthetic values: the concepts underlying kinship and political identity, religious symbols and practices, and philosophies of the composition of the life-world. In fact, cultural studies in European scholarship about Africa moved away from materiality altogether. The key works were about concepts of causation (Evans-Pritchard's eternal classic entitled *Witchcraft, Oracles and Magic Among the Azande* 1937), symbolic power (Turner 1968), kin relations (Fortes 1945), cosmology (Griaule 1948), and philosophy (Tempels 1946). Scholars of Africa such as Douglas (1966) made major contributions to general culture theory, by linking conceptions of what is pure and what is dangerous to specific ambiguities in social structures. In a sense, this period replaced a notion that African cultural 'genius' lay in the figurative arts with an 'African Genius' (Davidson 1969) that lay in the capacity to organize complex social life without the state domination characteristic of Europe. African cultural studies illuminated the study of society in general.

2.4 Culture Today

The rebellion against this approach, both within anthropology and beyond, was a rejection of the implication that any cultures could be studied as if they were 'traditional,' in the sense of closed, unchanging, and unchallenged from within. The arts resumed center-stage, not as iconic objects but as active creativity: emergent novelty, specific authorship, audience reception, and constant revision and re-creation. This was culture as a field of imagination and its students included its creators: the new novelists and artists of independent Africa, commenting on the present. The humanities became prominent again. Greenberg (1972) suggested a highly persuasive language classification, which changed the history of the modern peopling of the continent. Vansina (1961) innovated in historiography by promoting oral history, which helped to open up a new effort on orature in general. Scholars of their own languages, such as Abimbola (1976) of Yoruba and Kunene (1979) of Zulu, worked on their own language creativity. Novelists, poets, and film-makers—such as Achebe (1958), Soyinka (1964), and Sembene (1960)—used modern media to express the challenges of modernity as well as writing very influential critical works about culture and national life. New popular art forms such as juju and highlife music, the Yoruba traveling theater, textile fashion, and hair-dressing animated local engagement with the future and became subjects of study in the academy. Hountondji's (1977) attack on 'ethno-philosophy' both opened a new era of African philosophy and symbolized the break in cultural studies from the community-based traditionality of earlier anthropology. The center of gravity moved away from communities and towards the study of expertise and innovation.

In the 1990s, the cultural scene in Africa is yet more diverse, including new architecture, jazz, video, and new versions of Christian thought and expression. The vast African diasporas in the Americas and elsewhere, from both the era of the slave trade and recent global mobility, are now considered to be part of an African cultural ecumene. In turn, African history is being revisited to show how long and how widely Africa has been connected to the rest of the world, with lasting effects on many aspects of life—from cultigens to religious expression (Blier 1995)—that were once thought to express the very 'nature' of Africanity. These social geographical observations are matched by a new appreciation of the openness of cultures within Africa to novelty and originality, whether developed from within or borrowed from without. In a world era when dynamism is in full play, amongst ordinary people as well as cultural elites, this makes the study of Africa's long history, popular ebullience, and varied multiplicity the unique contribution of African Studies to the study of culture: to an understanding of cultural hybridity, memory, and power.

Diaspora Studies have been pioneered by peoples of African descent: in Brazil, the Caribbean, Europe, and undoubtedly in a growing crescendo elsewhere in the world. The 'retentions' and reconstitutions of African

culture by the continent's descendants who had been utterly stripped of every object, every community relationship, and any common African language during the era of the slave trade has remained a challenge to understanding. Herskovits's (1941) pioneering work in the 1930s stood for a long time at a tangent to cultural scholarship, which remained continent-based. In the 1990s, these themes are being reopened, and in the process African cultural studies are trailblazing in the study of global networks in artistic and intellectual life, and of memory and continuing creativity in diaspora cultures.

See also: African Studies: History; African Studies: Religion; Central Africa: Sociocultural Aspects; Colonialism, Anthropology of; Diaspora; East Africa: Sociocultural Aspects; Multiculturalism, Anthropology of; Southern Africa: Sociocultural Aspects

Bibliography

Abimbola W 1976 *Ifá: An Exposition of Ifá Literary Corpus*. Oxford University Press Nigeria, Ibadan, Nigeria

Achebe C 1958 *Things Fall Apart*. Heinemann, London

Blier S P 1995 *African Vodun. Art, Psychology and Power*. University of Chicago Press, Chicago

Davidson B 1969 *The African Genius: An Introduction to African Cultural and Social History*. Little and Brown, Boston

Diop C A 1955 *Nations nègres et culture*. Éditions Africaines, Paris

Douglas M 1966 *Purity and Danger; An Analysis of Concepts of Pollution and Taboo*. Routledge and Kegan Paul, London

Evans-Pritchard E E 1937 *Witchcraft, Oracles and Magic among the Azande*. Clarendon Press, Oxford, UK

Fabian J 2000 *Out of our Minds: Reason and Madness in the Exploration of Central Africa*. University of California Press, Berkeley, CA

Fortes M 1945 *The Dynamics of Clanship among the Tallensi; Being the First Part of an Analysis of the Social Structure of a Trans-Volta Tribe*. Oxford University Press, London

Greenberg J H 1972 Linguistic evidence regarding Bantu origins. *Journal of African History* **13** (2): 189–216

Griaule M 1948 *Dieu d'eau: Entretiens avec Ogotemmêli*. Éditions du Chêne, Paris (1965 *Conversations with Ogotemmêli: An Introduction to Dogon Religious Ideas*. Oxford University Press, London)

Herskovits M 1941 *The Myth of the Negro Past*. Harper and Brothers, New York

Hountondji P 1977 *Sur la philosophie africaine: critique de l'ethnophilosophie*. Maspero, Paris (1983 *African Philosophy: Myth and Reality*. Indiana University Press, Bloomington, IN)

Kunene M 1979 *Emperor Shaka the Great: A Zulu Epic* [trans. Kunene M]. Heinemann, London

Mudimbe V 1988 *The Invention of Africa: Gnosis, Philosophy, and the Order of Knowledge*. Indiana University Press, Bloomington, IN

Roberts M N, Roberts A F (eds.) 1996 *Memory: Luba Art and the Making of History*. Museum for African Art, New York

Schildkrout E, Keim C A (eds.) 1998 *The Scramble for Art in Central Africa*. Cambridge University Press, Cambridge, UK

Schwab G 1947 *Tribes of the Liberian Hinterland, with Additional Material by George W. Harley. Report of the Peabody Museum Expedition to Liberia*. Kraus, New York

Sembene O 1960 *Les bouts de bois de Dieu: Banty Mam Yall*. Le Livre contemporain, Paris [1976 *God's Bits of Wood*. Heinemann, London]

Soyinka W 1964 *Five Plays: A Dance of the Forests, The Lion and the Jewel, The Swamp Dwellers, The Trials of Brother Jero, The Strong Breed*. Oxford University Press, London

Tempels P 1946 *Bantoe-filosofie. Oorspronkelijke tekst*. De Sikkel, Antwerp (1959 *Bantu Philosophy*. Présence Africaine, Paris)

Turner V 1968 *The Drums of Affliction: A Study of Religious Processes among the Ndembu of Zambia*. Clarendon Press, Oxford, UK

Vansina J 1961 *Oral Tradition; a Study in Historical Methodology*. Aldine, Chicago

J. I. Guyer

African Studies: Environment

The environment, defined as biophysical landscapes and their natural resources (forests, wildlife, grasslands, water, and so on), has always figured prominently in public and scholarly images of Africa (Anderson and Grove 1987). The continent hosts a wealth of the world's biodiversity, especially its share of large mammal and bird species, and has attracted intellectual interest since at least the early nineteenth century. However, with the exception of geography, social science interest in environmental issues in Africa was minimal until the 1970s. The recent concern with environmental issues is motivated by a strong recognition that social and political forces increasingly shape the environment (Little et al. 1987). These complex relationships provide a significant empirical arena ('laboratory') for testing social science theories and methods and define a field that addresses the social and political dimensions of the physical environment. This article summarizes some of the major paradigms and topics that have influenced this field of study.

1. Theories and Approaches

Interest in the social dimensions of the African environment was provoked by a series of environmental events. The first, and perhaps most important, was the Sahelian drought of the early 1970s that instigated cries of environmental degradation and 'desertification.' Considerable doubts about whether or not desertification actually existed were voiced almost immediately, and picked up considerable momentum in the 1980s when the results of several long-term studies became available. Other events, such as the Ethiopian droughts and famines of 1971–1974 and

1984, conflicts surrounding the management of the continent's major river basins (Waterbury 1979), and West African deforestation evoked concerns of environmental catastrophes and resulted in simplistic causal statements about their origins. As early as the 1970s, social scientists were skeptical about many of the explanations for environmental degradation in Africa. Perhaps the most publicized of the social science debates concerned land tenure and the extent to which common or communal property ownership, a prevalent form of tenure in Africa, fosters natural resource abuse. Characterizing certain situations as a so-called tragedy of the commons, some researchers speculated that much of the degradation in Africa's dry regions results from the contradictions inherent when animals are owned privately, but land and other resources are held in common (see Anderson and Grove 1987). While the evidence against a tragedy of the commons is overwhelming, the position still has support among certain scholars and policymakers and still underlies current debates about land reform in Africa.

The complexity of recent environmental events challenges social scientists to rethink their methods and theories. During the 1980s and 1990s, interdisciplinary research programs on land use and environmental change in Sub-Saharan Africa continued to increase in number. Empirical findings from this collective work challenged orthodox assumptions about the relationship between population growth and environmental change, the resilience of African ecologies, and the capacity of local institutions to regulate resource use. They also pointed to concerns about the fundamental politics of natural resource use, an issue that had emerged in the early 1980s. In terms of theory, two broad bodies of work became especially appealing to social scientists and to a small number of ecologists. These are the 'ecology of disequilibrium' and the 'political ecology' approaches.

Recent theoretical advances in the 'ecology of disequilibrium (disturbance)' school are relevant to understandings of the complex relationships between human agency and African habitats. For example, the dryer portions of the African continent, including savannas, are subject to large rainfall fluctuations and sustained droughts (disequilibrium) from one year to the next. Climatic data collected since the 1980s reveal that many ecosystems of Africa are inherently unstable and, therefore, attempts to adjust conditions to some notion of stability violate the natural order and in themselves are destabilizing. Because of the higher variability of African ecosystems, many of the ecological concepts developed in the temperate zones (USA and Europe) fail to explain the dynamics of these highly variable ecosystems. Ecologists in Africa and Australia have designated these highly variable environments as disequilibrium ecosystems to distinguish them from ecosystems where climate patterns are generally reliable enough for resident plant and animal populations to reach some sort of equilibrium (Behnke et al. 1993). Most plans for preserving biodiversity in Africa (for example, national parks and biosphere reserves), however, are still based on equilibrium theory and invoke notions of carrying capacity and average stocking rates to preserve an 'undisturbed wilderness.'

Recent theoretical advances in the ecology of disequilibrium school are relevant to understandings of the complex relationships between politics, human agency, and African habitats. The approach is also more consistent with indigenous models of the African environment, which have never excluded anthropogenic disturbances nor pursued ecological equilibrium as an objective. For example, herd management strategies of East African pastoralists—which have always been the bane of conservationists—assume drought, some degree of range degradation, and fire (burning) as norms, and have never tried to pursue ideas of carrying capacity or equilibrium.

The political ecology approach, in turn, is relevant to this discussion; indeed, often the same Africanist social scientists have intellectual stakes in both schools (for example, Bassett, Behnke, Horowitz, Leach, Scoones, and Swift). Political ecology can be a useful framework for weaving together different disciplines and has contributed considerably toward understandings of the social and political processes underlying resource use in Africa and elsewhere. There are three elements that define political ecology: resource access, resource management, and ecological impacts. Political ecology starts with the political question of access to resources, but also addresses how resources are managed and the environmental effects of this. Particular areas of political ecology research in Africa include studies of: (1) forestry (Fairhead and Leach 1996); (2) wildlife conservation (Neumann 1998); and (3) pastoralism (Little 1992).

2. *Other Topics of Research*

Social science research on the African environment covers a range of significant topics, some of which have been discussed earlier: colonial history, land tenure, and pastoralism. While it is not possible to cover the breadth of research issues, there are at least four themes that require special attention.

2.1 *Gender and the Environment*

Recent studies highlight the gendered nature of the environment, whereby resources and their value are perceived differently along gender lines (Moore and Vaughan 1994). The increased dependence on wage labor markets in Africa creates additional pressures for women. They have been compelled to absorb tasks normally carried out by men, many of whom have

migrated to towns and other areas of employment. The increased workload of females in agrarian economies is at least partially a response to this loss of labor. As Sperling (1987, p. 179) shows for the Samburu pastoralists of northern Kenya, 'Male emigration intensifies the female workload … They become more directly involved in many aspects of herd care, such as fencing, watering, curative regimes, and forage collection.' These additional demands, however, are not always accepted passively. In some areas, pastoral women have refused to contribute to certain tasks (e.g., moving animals to remote pastures) that require excessive labor beyond their already heavy burdens and long absences.

Women of poor households are most seriously impacted by labor shortages and the environmental problems that ensue. Not only do they absorb additional tasks, but because of the localized degradation resulting from decreased mobility, they must search further from their homes for firewood and other natural products (e.g., wild plants) required for cooking and other domestic chores. Anthropological studies in northern Kenya estimate that because of sedentarization by pastoral groups, some women now allocate considerably more labor for collecting fuel wood than in the past (Ensminger 1987).

Very little is known about how environmental degradation in Africa is perceived by different gender groups or how their perceptions are translated into action. For example, a woman may define degradation by declining yields from dairy herds since women often control income from milk sales, or by the amount of time spent in collecting fuel wood. A male, in turn, who might control income from tree crops, may define environmental degradation in terms of how it affects coffee or tea production. Recent efforts highlight the importance of incorporating political ecology into feminist research on the African environment (Rocheleau et al. 1996). By making this linkage it is possible to show how resource access by women is shaped by power relations and policies that often disadvantage them, and force them on to marginal lands where their only option may be overexploiting the environment. The emergence of women-led environmental groups as significant political forces, such as the Green Belt Movement in Kenya, demonstrates the ways in which gender meshes with politics and environmental concerns.

2.2 Indigenous Environmental Knowledge

African farmers and herders control a wealth of sophisticated knowledge about the environment. Studies by anthropologists throughout Africa highlight the importance of understanding how local knowledge systems affect the use of environmental resources. These systems are expressed through elaborate local terminologies and classification systems, which rarely are acknowledged by governments and policy makers. Instead, outside agencies and development 'experts' often advocate new technologies and practices that may actually contradict or conflict with these local knowledge systems. Social science research in this area has emphasized the documentation of local vegetation and resources, local environmental practices, and the development potential of indigenous knowledge. As indicated earlier, these knowledge systems can have important gender components.

A few examples from Kenya demonstrate the critical role of local knowledge systems and practices in natural resource use. In the Baringo area of Kenya, for instance, many community-based irrigation schemes have achieved considerable success in a region noted for land degradation and food shortages; and where large-scale irrigation systems have failed miserably (Little 1992). Based on local councils of elders (*lamaal*), low-cost irrigation systems have been developed that conserve the fragile soils and transport water through intricate canal systems over kilometers of dry barren lands. Land and water disputes are resolved locally and, unlike neighboring villages, food-aid distribution is minimal in these areas. Because local irrigators maintain trees in their fields and do not drain or extend irrigation into local wetlands, they help to maintain some of the richest diversity of bird populations in East Africa (in excess of 700 species in an area of less than 70 square kilometers). By contrast, the mechanized, clear-cutting techniques of large-scale, government irrigation projects in the area threaten this rich diversity by transforming valuable habitats.

In the nearby Kerio Valley of northern Kenya, irrigation is managed on the basis of clans, and in some cases sub-clans. As Soper (1983, p. 91) notes, 'the water ownership unit is the clan section or, in some cases, sub-section, which is also the land-holding and the basic residential unit.' A clan or a group of clans will own a particular furrow, which they are responsible for managing and maintaining. Water from the furrows is allocated on a rotating basis to members of the clan, usually based on a 12-hour watering unit or fractions of the unit. Most agriculturalists who have surveyed the furrow system note its efficiency in conserving water and soil. They also indicate the community's commitment to maintaining it. In an area that is prone to drought and environmental problems, the system stands as a notable achievement.

2.3 Environmental 'Narratives'

Important environmental work in Africa is currently focused on how particular narratives or discourses emerge, are 'scientifically legitimized,' and are then incorporated into environmental policies. Both Fairhead's and Leach's (1996) work on deforestation in West Africa and Tiffen et al.'s (1994) treatise on soil erosion in East Africa skilfully show how flawed assumptions about population growth, local practices, and environmental degradation stem from a set of

narratives dating back to the early colonial period. These discourses, however, are still reflected in environmental policies that shape the ways in which Africans interact with their habitats.

The debate about desertification in Africa represents another environmental narrative that dates back to the colonial period but has contemporary implications. An entire book easily could be devoted to how the desertification debate was constructed, how political and institutional factors contributed to its production and reproduction, and how 'science' was invoked to justify the excessive funds and projects allocated to such an elusive issue. Concerns about this phenomenon stem from the 1930s when colonial officers pointed to the creeping deserts of West Africa and to the terrible dust storms of East Africa. The colonial soil erosion and conservation campaigns of the 1930s, a favorite theme of Africanist historians, stemmed in part from official beliefs in the desertification narrative. In the post-colonial period, elaborate sets of projects and techniques were established to measure desertification, even though climatic and aerial photos showed that the extent of the Sahara's advance had been greatly exaggerated. Similar to other narratives about environmental degradation in Africa, the 'scientific' arguments about desertification and its causes are reinforced by environmental policies, often supported by outside agencies, which blame human agency for environmental problems. Fortunately, there is a growing body of social science wisdom in Africa that challenges these truisms, as well as defines an important area of social science research.

2.4 Environment and Development

The relationship between economic development and a sustainable environment is among the most important research issues in Africa today. Questions of food security, sustainable development, and environmental welfare are imbedded in this topic. However, the tension between strategies for increasing rural incomes and development in Africa and for preserving the natural resource base remains unresolved; nor is there a good consensus on how to measure and monitor these processes. Most social scientists acknowledge the overwhelming importance that local economic incentives and benefits assume in conservation efforts in Africa, but most have not been successful in documenting the social and economic variables that facilitate effective environmental programs.

Recently, development practitioners have emphasized local participation as a possible means for achieving environmental goals, especially in villages surrounding important parks and protected areas. While it is difficult to achieve meaningful local participation in development activities, the challenges are even greater when environmental goals are pursued. Social science research shows how the recent emphases on local participation and community-based conservation are a reaction to earlier, highly centralized programs. The case of national parks, backed by restrictive legislation and heavy-handed sanctions that carved out large chunks of indigenous lands without local involvement, are classic examples of the earlier approach. This top-down approach to conservation was especially characteristic of many wildlife and forestry departments in colonial Africa.

The reality that biodiversity programs could not be limited to national parks because of migratory animal species also provoked a concern with community conservation programs. A proliferation of such activities was initiated in the 1980s and 1990s by many of the major international environmental organizations (for example, the World Wildlife Fund (WWF) and the International Union for the Conservation of Nature (IUCN)).

What has social science research in Africa told us about the relationships between development and environmental conservation? First, it has shown that local participation and conservation cannot be delinked from development concerns if environmental programs are to be sustainable. The CAMPFIRE (Communal Areas Management Programme for Indigenous Resources) program of Zimbabwe is a good example of a community conservation program with a biodiversity goal that also has a development outcome. In the CAMPFIRE effort, the local community participated in the identification of the conservation problem—in this case, better management and regulation of wildlife resources (Metcalfe 1994). Although international and national organizations were instrumental in heightening local awareness of conservation problems, the communities themselves saw the linkages between economic benefits and sustainable management of wildlife. The CAMPFIRE program was first initiated in a very poor region of Zimbabwe, and it took on the appearance of a rural development rather than a conservation project. This effort contrasts sharply with other wildlife/park schemes in Africa where wildlife conservation—as defined by external parties—has been the overriding objective, and local populations have been alienated. The Zimbabwe case, however, has proven to be a better model for promoting wildlife conservation, and has not confronted many of the social problems and conflicts that have marred wildlife conservation in Africa (Anderson and Grove 1987). In Zimbabwe, low-income producers saw the economic benefits that could accrue from tourism and hunting, while recognizing the threat that poaching posed to these activities. While there are still massive poaching problems in eastern and southern Africa, the CAMPFIRE approach demonstrates that community participation can slow down rates of resource depletion.

Other studies of environment and development in Africa point to the importance of national and

international political processes, including a renewed interest in the management of Africa's major waterways. Research shows that macro processes affecting access to natural resources vary considerably among different African countries, reflecting the varied political structures and leadership of different states. In addition, the extent to which particular states allow sufficient political space for local participation will also vary by country. With widespread political changes and turbulence in Africa, a future area for research will address the effects of collapsed states and liberation movements on natural resources and the environment (Salih 1999). In those few instances where strong states exist, however, the national political structure may prove more significant than any other variable in determining local participation. Some successful local conservation programs in Kenya and Uganda have taken place without conducive policy environments and without large amounts of external funding. In the well-documented Machakos case (Kenya), for example, local households and communities have responded to land shortages and severe soil erosion by shortening fallow periods and improving the quality and maintenance of hillside terraces (Tiffen et al. 1994). While the macro political environment has not been particularly conducive to local participation, population and land pressures motivated local communities to address environmental problems. More cases like this need to be documented by social scientists and their results disseminated to policy makers.

To conclude, environmental issues will increasingly shape social science research agendas in Africa during the twenty-first century. Scholars will be challenged even more to confront flawed and overly simplistic interpretations of ecological problems on the continent. Pressing needs to sustain adequate food production and incomes without degrading natural resources will also occupy farmers, herders, and governments, while opening space for innovative interdisciplinary programs that actively engage policy and policy makers. The success of these initiatives during the next decade will cast the critical intellectual questions about the environment, as well as contribute to improved local livelihoods.

See also: African Studies, Economics; African Studies, Gender; African Studies, History; African Studies, Politics; African Studies, Society; Environmentalism, Politics of; Environment and Development; Desertification; Land Use Regulation

Bibliography

Anderson D, Grove A (eds.) 1987 *The Scramble for Resources: Conservation Policies in Africa, 1884–1984*. Cambridge University Press, Cambridge, UK

Behnke R, Scoones I, Kerven C (eds.) 1993 *Range Ecology at Disequilibrium: New Models of Natural Variability and Pastoral Adaptation in African Savannas*. Overseas Development Institute and Russell Press, London

Ensminger J 1987 Economic and political differentiation among Galole Orma women. *Ethnos* **52**: 28–49

Fairhead J, Leach M 1996 *Misreading the African Landscape*. Cambridge University Press, Cambridge, UK

Little P D 1992 *The Elusive Granary: Herder, Farmer and State in Northern Kenya*. Cambridge University Press, Cambridge, UK

Little P D, Horowitz M M, Nyerges E A (eds.) 1987 *Lands at Risk in the Third World: Local Level Perspectives*. Westview Press, Boulder, CO

Metcalfe S 1994 The Zimbabwe Communal Areas Management Programme for Indigenous Resources (CAMPFIRE). In: Western D, Wright R M, Strum S (eds.) *Natural Connections: Perspectives in Community-Based Conservation*. Island Press, Washington, DC

Moore H, Vaughan M 1994 *Cutting Down Trees: Gender, Nutrition, and Agricultural Change in the Northern Province of Zambia, 1890–1990*. Heinemann, Portsmouth, NH

Neumann R P 1998 *Imposing Wilderness: Struggles over Livelihood and Nature Preservation in Africa*. University of California Press, Berkeley, CA

Rocheleau D, Slayter-Thomas B, Wangari E (eds.) 1996 *Feminist Political Ecology: Global Issues and Local Experience*. Routledge, London

Salih M A 1999 *Environmental Politics and Liberation in Contemporary Africa*. Kluwer Academic Publishers, Boston, MA

Schroeder R 1999 *Shady Practices: Agroforestry and Gender Politics in The Gambia*. University of California Press, Berkeley, CA

Soper R 1983 A survey of the irrigation systems of Marakwet. In: Kipkorir B, Soper R, Sssenyonga J (eds.) *Kerio Valley: Past, Present and Future*. Institute of African Studies, Nairobi, Kenya

Sperling L 1987 Wage employment among Samburu Pastoralists of north central Kenya. *Research in Economic Anthropology* **9**: 167–90

Tiffen M, Mortimore M, Gichuki F 1994 *More People, Less Erosion: Environmental Recovery in Kenya*. Chichester, New York

Waterbury J 1979 *Hydropolitics of the Nile Valley*. Syracuse University Press, Syracuse, New York

P. D. Little

African Studies: Geography

1. Geography

Geography is an integrative discipline that studies the location of phenomena on the earth's surface and the reasons for their location. Most people think that geography consists of memorizing countries and their capitals, or photographic essays of exotic places in popular magazines. However, these are only a small part of contemporary geography. Within the past 30 years the discipline of geography has undergone an unparalleled technological revolution in the spatial analysis of data. First, the personal computer has

greatly enhanced measurement technologies including remote sensing and global positioning systems, which have rapidly expanded the quantity and types of spatially referenced data about the human habitat and the physical environment. Second, the development of computer-based Geographical Information Systems (GIS) has greatly facilitated the ingestion, management, and analysis of the rapidly increasing spatially referenced data. This has allowed geographers to evaluate spatial processes and patterns, and to display the results and products of such analyses at the touch of a button. The word geography was coined by the ancient scholar Eratosthenes based on two Greek words, *geo*—meaning 'Earth,'—and *-graphy*—meaning 'to write.' Geographers ask three basic questions, where do people and environments occur on the Earth's surface? Why are they located in particular places? What are the underlying explanatory factors for the spatial patterns? This entry briefly reviews how geographers have studied the continent of Africa with reference to the past 30 years.

2. Geographic Approaches on Africa

During the 1970s and 1980s research on Africa dwelt on the many crises. While Africa has had its fair share of problems such as the collapse of the state in Sierra Leone, Liberia, Somalia, Rwanda, Zaire, Burundi and a few others, and economic shocks engendered by the continuing failure of structural adjustment programs, the 1990s witnessed momentous positive changes. For example, South Africa experienced remarkable and unprecedented social and political change and moved toward majority rule faster than previously anticipated. Dictatorial regimes in Zaire, Malawi, and Zambia as well as elsewhere in Africa crumbled and were replaced by emerging democratic systems. Gratifyingly, geographical research on Africa in the 1990s began to move away from sensationalism and over-generalization to more pragmatic and pertinent micro-level perspectives that reflected the diversity and richness of the African continent. In contrast to the gloom and doom of the 1970s and 1980s, some scholars have begun to highlight some of the positive new developments.

Most geographic work in the 1970s and 1980s concentrated on regional geography with an emphasis on country surveys, descriptions and compilation of geographic data at country or regional level. The late 1980s and 1990s saw the rise of new paradigms in the study of African geography. While the empirical subject matter may be agriculture, health, gender issues, development, etc., the theoretical paradigm guiding the geographic research during the 1990s was often about issues such as representation, discourse, resistance, and indigenous development within broader frameworks influenced by the ideas of prominent social science scholars such as Foucault (1977), Said (1978), and Sen (1981). Broadly speaking, the works fall into the three main subdisciplines of geography, namely, human geography (by far the most dominant), physical geography, now commonly referred to as earth systems science and/or global change studies, and geographic information systems (GIS). Within these three main sub-disciplines, various theoretical perspectives overlap to characterize the growing body of research by geographers on Africa.

During the 1990s geographers came to realize that the rapid and complex changes that Africa was undergoing could not be adequately explained by the conventional narrowly focused disciplinary perspectives and approaches. Geographers embraced and, in some cases, devised more complex and integrated interdisciplinary approaches. The most important of these transitions or developments in African geographical research during the 1990s include: (a) post colonial–poststructuralist–postmodern approaches; (b) political ecology championed by the works of Blaikie and Brookfield (1987); (c) Boserupian perspectives on population and environment promoted by the hotly debated book by Tiffen et al. (1994); (d) challenging environmental orthodoxies, particularly a reassessment of 'taken-for-granted' ideas about the environment championed by Fairhead and Leach (1996); (e) development from below/grassroots initiatives; (f) recognition of the importance of indigenous knowledge; (g) the impact of globalization on economic development particularly in agriculture and industrial restructuring; (h) policy-oriented studies; (i) social geographies pertaining to gender and other issues; and (j) global environmental change research involving climatologists, geomorphologists, hydrologists and biogeographers using an integrative–systems framework. The following two sections briefly highlight some of the major areas of research by geographers working on Africa.

3. Human Geography

In the subdiscipline of human geography, a number of research themes were explored by geographers. These include population, resources and the environment, population dynamics, development discourse, policy-oriented or impact-analysis studies, urban and regional development, and the geography of disease and healthcare. Theoretical orientations included the Boserupian perspective, political ecology, political economy, post-colonialism, post-structural, feminist perspectives, sustainable–green development approaches, globalization, disease ecology, and location–allocation models. For example, research on population issues moved away from neo-Malthusian approaches. Couched within the Malthusian tradition, it was commonplace for analyses of Africa over the 1980s to bemoan the combination of rapid population growth and economic and environmental decline.

Indeed, many geographers continue to be convinced that ecological degradation is a human-induced problem with a strong element of neo-Malthusian thinking. However, recent works have moved away from the neo-Malthusian trap and have concentrated on exploring the role of population growth on agrarian transformation, the role of land tenure in agrarian change, farmer–pastoral conflicts, the environment, and issues of gender and resource contestation. This literature stresses the good aspects of increasing population densities with respect to agricultural transformation. Perhaps the most influential work in the Boserupian tradition is by Tiffen et al. (1994). Their work argues that even as population densities have increased, agropastoral productivities have increased and a degraded landscape has flourished with trees, terraces, and productive farms.

The debates on indigenous land tenure systems versus privatization intensified during the 1990s. Several works questioned the traditional thought that indigenous tenure systems are an obstacle to increasing agricultural productivity. Fairhead and Leach (1996) challenge 100 years of received wisdom on the degradation of the African environment and their numerous works continue to dramatically influence 'environment' research in geography with reference to Africa. Little and Watts (1994) examine the impact of globalization on agrarian change in Africa under the rubric of contract farming. The book focuses on the genesis, growth, and form of contract farming in sub-Saharan Africa. In many geographic works that use the political economy, political ecology and liberation ecology frameworks have tackled the perplexing issues of multinational corporations versus local resources, common property rights and indigenous knowledge, land for agricultural extensification versus wildlife conservation, and afforestation–deforestation issues. The rise of post-structuralist, post-colonial, post-modern, and feminist critiques of development discourse has spawned an interesting set of case studies that examine the complex intersections among gender, agrarian change, environmental discourse, access to resources, and indigenous knowledge. Other scholars have explored the role of women in development and the changing conditions of women in rural and urban Africa.

Geographers have also been prominent in development discourse, a term which refers to the language, words, and images used by development experts in development texts to construct the world in a way that legitimates their intervention in the name of development. Geographers have contributed significantly in critiquing development discourse, particularly its characteristic language of crisis and disintegration, which gives justification for intervention. The works which critique this approach arose out of old suspicion of a hidden agenda behind the introduction and abandonment of development strategies in Africa, as well as the perpetuation of development strategies and notions that are detrimental to Africa's development, particularly structural adjustment programs. A number of influential books have fruitfully engaged and critiqued this perspective (see, for example, Corbridge 1995, Godlewska and Smith 1994). An important issue raised in some of the articles contained in these volumes and others is the persistent (mis)representation of development and of Africa itself. For example, a chapter in Godlewska and Smith's (1994) offers a critique of images in the *National Geographic* as the complicity of geography as a discipline in perpetuating colonial and post-colonial myth-making about Africa and about development.

During the 1990s, scholars from North America, Africa, and Europe collaborated to examine issues concerning urban development, industrial restructuring, the informal sector, labor, and regional development, with particular focus to the South African transition. Many geographers have been grappling with the question of what does a post-apartheid geography of this country and the southern African region look like. Thus geographers working on this issue have examined it from many perspectives, particularly the (dis)continuities between the geographies of apartheid and post-apartheid in the social and economic realms which shaped South Africa's economy, cities, and social relations under apartheid, and continue to do so in the post-apartheid era.

Medical geography research on Africa continued along the traditional lines of disease ecology and geography of health care with a clear trend toward linking health with its political and economic context. A number of scholars continued work in disease ecology focusing on specific diseases such as filariasis (elephantiasis) and dracunculiasis (guinea worm). Geographers, using a political economy or structuralist theoretical perspectives have examined the diffusion of HIV–AIDS pandemic and have typically recommended education with empowerment arguing against intervention strategies that ignore poverty.

4. *Global Change and Earth Systems Science*

This is an area in which an old disciplinary category is becoming obsolete as physical geographers increasingly refer to their subject matter as earth systems science or global change instead of physical geography. Physical geographers are examining the causes and impacts of climate change in the Sahel, rainfall patterns, and El Niño effects in Africa. Some of their work challenges conventional wisdom that the Sahara desert is expanding at a phenomenal rate. Instead they demonstrate that there has been no progressive change of either the Saharan boundary or vegetation cover in the Sahel during the last 16 years, nor has there been a systematic reduction of 'productivity' as assessed by the water-use efficiency of the vegetation cover. Some areas of study include sand transport and dune

formation, desert landscapes, soil degradation, and river morphology.

European geographers, particularly British and African geo-scientists, have been extremely active in conducting research broadly defined as physical geography. There are many hydrologists and ecologists studying wetland ecology–hydrology. The Cambridge group led by Grove and Adams is very active in examining the multifaceted physical geography processes on the African continent. Their recent book (Adams et al. 1996) is a testament to their productivity. In this book, the authors weave together biophysical and human induced processes and recognize the multitude of environmental conditions at the local, regional, and continent-wide scales. The work contains detailed discussions of the physical geography of Africa, the geomorphologic and biogeographical aspects of the continent, and the impact of human agency on African environments.

There is a tremendous amount of work on the African physical environment by geo-scientists. Geographers have had a significant role to play in this work, although this work is often 'hidden' in non-geographical journals and multidisciplinary team projects. The emerging idea of earth systems science as an integrative science has made the old-fashioned term 'physical geography' obsolete so that people who used to call themselves physical geographers are now able to cross traditional disciplinary boundaries with ease. One important development is the growing importance of remote sensing and geographic information systems (GIS) as a research tool, as well as the importance of large-scale modeling to generate global climatic models. These techniques have grown in their importance as tools for studying environmental change on the continent. Many geographic studies of Africa use these tools. Such tools have made it possible to monitor and analyze variations in, for example, grazing intensity associated with rural land practices. Other studies have used satellite imagery to study short- and long-term variability in climate within Southern Africa by determining trends in variability of vegetation greenness for evidence of climatic trends.

5. *Conclusion*

Geographic research on Africa is multifaceted and interdisciplinary in its methodological and theoretical approaches. Within the past 30 years the discipline has embraced or devised new approaches in its study of the continent. The bulk of the work has been on human geography. A number of insufficiently developed substantive areas include physical geography (geomorphology and biogeography), historical cartography, political and cultural geography, and regional geography. However, several excellent and comprehensive textbooks that address these underdeveloped areas, particularly cultural and regional geography, seem to have bridged the gap. Some of these books do an excellent job of putting together the regional geography of Africa in a systematic fashion with a coherent thematic interpretation of the regional, cultural, and development status of the continent.

One important neglected area by human geographers is the political and electoral geography of the continent. Transitions such as democratization in quite a few countries and the collapse of organized authority in others (Liberia, Somalia, Democratic Republic of the Congo, Rwanda, etc.) have received precious little in the way of attention by geographers. Another neglected area by geographers is that of population geography. Perhaps due to overreaction against the neo-Malthusian debacle, demographic factors have been completely ignored in recent research. Indeed, the major weakness of the literature that uses political ecology as its guiding framework is its muteness on dynamics of population change and demographic factors. It is apparent in reading this body of work that most of the authors, constrained by the political economy and political ecology frameworks, ignore population in their analyses. However, it should be obvious that excluding demographic factors in these important debates may result in shortsighted policy formulations.

In conclusion, the rich geographic literature on Africa reveals the adoption of newer methodological and theoretical perspectives during the late 1980s and 1990s. For example, several prominent geographers embraced and/or helped to advance the post colonial–post structuralist studies of representation and resistance and critiques of development discourse. As we move into the twenty-first century, we predict that geographical research on Africa will intensify the adoption of these newer interdisciplinary approaches, and perhaps, develop new ones in the process of discarding old and static methodologies.

See also: African Studies: History; Central Africa: Sociocultural Aspects; East Africa: Sociocultural Aspects; Postcolonial Geography; Southern Africa: Sociocultural Aspects; West Africa: Sociocultural Aspects

Bibliography

Adams W M, Goudie A S, Orme A R 1996 *Physical Geography of Africa*. Oxford University Press, London

Blaikie P, Brookfield H 1987 *Land Degradation and Society*. Methuen and Company, London

Corbridge S (ed.) 1995 *Development Studies: A Reader*. Arnold, London

Fairhead J, Leach M 1996 *Misreading the African Landscape: Society and Ecology in the Forest Savanna Mosaic*. Cambridge University Press, New York

Foucault M 1977 *The Archeology of Knowledge*. Tavistock, London

Godlewska A, Smith N (eds.) 1994 *Geography and Empire*. Blackwell, Oxford, UK
Little P D, Watts M J (eds.) 1994 *Living Under Contract: Contract Farming and Agrarian Transformation in Sub-Saharan Africa*. The University of Wisconsin Press, Madison, WI
Said E 1978 *Orientalism*. Routledge, London
Sen A 1981 *Poverty and Famine*. Clarendon Press, Oxford, UK
Tiffen M, Mortimore M, Gichuki F 1994 *More People, Less Erosion: Environmental Recovery in Kenya*. Wiley, Chichester, UK

E. Kalipeni

African Studies: Health

Health has been defined by the World Health Organization as a condition of complete physical and psychological wellbeing. In common usage the term 'good health' is employed to mean that there is present no major manifestation of ill-health, and this is also the most satisfactory way of scientifically determining the situation. The most clearly defined outcome of ill-health (or morbidity) is death (or mortality) because it can be most certainly defined and measured, and because in its irreversibility it is an index of the most extreme ill-health. This entry will focus on mainland sub-Saharan Africa plus the large island nation, Madagascar, omitting the smaller islands of the Indian and Atlantic oceans with their mixed historical, ethnic, and cultural backgrounds. All of these have lower mortality than any mainland African country and most are richer. The advances in health research since the mid-twentieth century are outlined.

1. *Researching Africa's Health Levels*

At mid-twentieth century sub-Saharan Africa was assumed to be the unhealthiest region in the world, in spite of an almost complete lack of data to confirm this view. That confirmation was to be achieved later by reconstructing the situation with data from subsequent research, which showed that as late as 1950–5 the region was characterized by a life expectancy at birth of only 37 years (United Nations Population Division 1999). The problem in making such estimates was a complete lack of vital registration, and of usable mortality and morbidity information in censuses and demographic surveys. Comprehensive national counts of deaths or illness are still not available anywhere in the region.

This problem of inadequate data has been overcome in three ways:

(a) The so-called 'indirect methods' of estimating mortality (and fertility) levels from inadequate data have been invented by William Brass and colleagues associated in later years with the London School of Hygiene and Tropical Medicine, while the stable and quasi-stable population model approach was developed by Ansley Coale and colleagues at the Office of Population Research, Princeton University. At first the 'Brass' methods provided estimates only of child mortality, but later techniques were developed for adult mortality, although the latter results were usually less secure.

(b) Censuses and national sample surveys with questions allowing indirect or even direct estimates of vital rates were developed. The first censuses with questions on births and deaths were those in British East Africa in 1948. From the 1960 census round, the United Nations assisted African censuses. From 1954, demographic sample surveys were conducted in most francophone African countries. Subsequently, many African countries participated in the great international survey programs: the World Fertility Survey (WFS) (which contained mortality questions) from 1975 and its successor, the Demographic and Health Surveys (DHS) from 1985. Nevertheless, by the year 2000 there still had been no adequate censuses or surveys of the Congo (Democratic Republic) or Angola, and only preliminary reports had been issued for the DHS surveys of South Africa and Ethiopia.

(c) High-intensity surveillance projects had been established in a range of areas. These provided demographic and health information and were often the sites for intervention projects. They included largely demographic projects such as the Sine-Saloum (1962–6) and later studies in Senegal, and those with greater attempts to investigate health: Pare-Taveta, Kenya and Tanzania (1954–6), Keneba, Gambia (1956 onward), Danfa, Ghana (1969–74), Malumfashi, northern Nigeria (1974–9), Machakos, Kenya (1974–81), Kilombera, Tanzania (1982–8) and Navrongo, Ghana (1993 onward). All programs measured health, mostly by mortality or survival measures, and all involved collaboration between African and outside institutions. African demographic estimation was the most challenging in the world, and the new techniques developed to meet that challenge were employed subsequently in other parts of the developing world and by historical demographers.

The data released from these investigations were employed by a succession of research programs:

(a) The African project of Princeton's Office of Population Research analyzed the results from predominantly francophone African demographic surveys and anglophone African censuses from 1961, and published *The Demography of Tropical Africa* in 1968. Its major contribution was its fertility estimates, and those on mortality were largely confined to infancy.

(b) The International Union for the Scientific Study of Population (IUSSP) in the mid-1980s commissioned papers on African mortality change for a 1987 conference and published *Mortality and Society in Sub-Saharan Africa* in 1992.

(c) The American National Research Council's Committee on Population established a program on the population dynamics of sub-Saharan Africa in 1989 and published *Demographic Change in Sub-Saharan Africa* and five other volumes in 1993.

(d) From the late 1980s the World Bank commissioned studies of the health of sub-Saharan Africa and in 1991 published *Disease and Mortality in Sub-Saharan Africa*. The first part of the book was constituted by studies of child and adult mortality and child malnutrition drawn from WFS and DHS data, the second part by studies of specific ailments from miscellaneous sources, and the third part by reports on morbidity and mortality from the various surveillance projects.

(e) A collaborative analytical program of persons working in African universities and institutions on West and Middle Africa resulted in 1975 in the publication of *Population Growth and Socioeconomic Change in West Africa*. In it, Cantrelle developed his thesis on tropical mortality (discussed below).

Some African research themes have a strong behavioral component. Studies of the impact of parental education, especially maternal education on child survival began in the region (Orubuloye and Caldwell 1975, Caldwell 1979, Farah and Preston 1982), and have since become numerous elsewhere. The AIDS epidemic has brought demographers and other social scientists to the study both of AIDS mortality and of sexual relations and other aspects of HIV transmission (Cleland and Way 1994, Awusabo-Asare et al. 1997).

Health in sub-Saharan Africa has not been a major feature of social science literature, probably because the data have been elusive and difficult to interpret. It has constituted 7.8 percent of all articles in *Social Science and Medicine*, 2.5 percent in *Population and Development Review* and, except for anthropological accounts of traditional healing practices, none in *Africa*. The latter point underscores the fact that the major social science input into African health research has been by demographers. Biomedical researchers have undertaken research on specific diseases. The balance of this article summarizes what the research reveals about African health.

2. The Health Situation

The first reliable estimates of sub-Saharan mortality were for the late 1940s and early 1950s, and were for infants, and then young children, made by adjusting mothers' reports of their children's deaths. These revealed infant mortality rates (deaths per 1000 births during the first year of life) ranging in most of francophone Africa between 200 and 275 (and probably equivalent to life expectancies ranging from 25 to 35 years). At the district level, Mopti in Mali and Luanda in Angola recorded infant mortality rates of 350 and 329, respectively, showing that one-third of all births resulted in deaths during the first year of life. In contrast, Kenya recorded an infant mortality rate of only 132 (a life expectancy around 45 years) and two of its districts, Central and Rift Valley Provinces, registered rates under 100. Subsequent studies increasingly involving life histories, many summarized by Althea Hill, showed infant and child survival improving nearly everywhere until at least the 1980s. West African mortality was higher than that of East and Southern Africa, but convergence was taking place. As appropriate questions were added to surveys, advances were also made in the study of adult mortality, which fell consistently until the 1980s. Thereafter, the decline became much slower in West Africa, and halted or reversed in East and Southern Africa (Timaeus 1999).

At the end of the twentieth century (1999) sub-Saharan Africa's life expectancy was 49 years, with male expectancy around 48 years and female 50 years. This compared with 61 years in South Asia, 64 years in North Africa, 65 years in Southeast Asia, 69 years in Latin America, 72 years in East Asia, 73 years in Southwest Asia, and 75 years in industrialized countries. With 9 percent of sub-Saharan African babies dying in the first year of life, and 15 percent in the first five years, the region's infant and child mortality was also the world's highest. Sub-Saharan Africa's life expectancy compares with that in Western Europe at the beginning of the twentieth century, suggesting a health lag of about 100 years. Within sub-Saharan Africa, Southern Africa recorded a life expectancy of 56 years, West Africa 52 years, Central Africa 49 years, and East Africa 44 years, the latter's poor performance being a recent product of the AIDS epidemic. National levels ranged from 55 or more years in Ghana, Liberia, South Africa, and Cameroon to 42 years or less in Malawi, Swaziland, Botswana, Zimbabwe, Niger, and Ethiopia. Such low levels as those for the last-mentioned countries are no longer found anywhere else in the world.

2.1 Past Mortality Trends

If tropical African population growth was almost stationary before the European partition of the region in the 1880s, then stable population calculations suggest that life expectancy was around 20 years. Alternatively, if, as has been suggested, the European presence on the coast had allowed the adoption of new foodstuffs permitting the denser settlement of the forest and other wetter lands, thus leading to a population growth rate as high as 0.5 percent per annum, life expectancy may have risen to around 23 years. In any case, mortality was, as traders on the west coast knew, the highest in the world, with malaria, yellow fever and other insect-borne diseases protecting the land from European settlement of the type that had occurred in Latin America. These mortality levels

Table 1
Life expectancy at birth (in years), 1950–2000 and per capita income (in US$), 1997

	Sub-Saharan Africa	North Africa	Asia	South Asia	Latin America	World
1950–55	37	42	41	39	51	46
1960–65	41	46	48	45	57	52
1970–75	45	51	56	50	61	58
1980–85	49	57	60	55	65	61
1990–95	49	62	64	60	68	64
1995–2000	49	65	66	62	69	65
Per capita income 1997	$520	$1160	$2450	$470	$3950	$5170
Adult female literacy 1995	47%	41%	50%	36%	85%	62%

Source: United Nations Population Division; Population Reference Bureau 1999; World Bank 1999.

Table 2
Average annual increases in life expectancy (in years), 1950–2000

	Sub-Saharan Africa	North Africa	Asia	South Asia	Latin America	World
1950–55 to 1960–65	0.4	0.4	0.7	0.6	0.5	0.6
1960–65 to 1970–75	0.4	0.5	0.8	0.5	0.4	0.6
1970–75 to 1980–85	0.3	0.5	0.4	0.5	0.4	0.3
1980–85 to 1990–95	0.1	0.6	0.4	0.5	0.3	0.3
1990–95 to 1995–2000	0.0	0.5	0.2	0.4	0.2	0.3

Source: As for Table 1. Increases no calculated on the rounded figures of Table 1.

are not much greater than the demographic surveys of the 1950s and early 1960s found persisting in some remote rural areas of the West African savanna.

2.2 The Second Half of the Twentieth Century

Mortality estimates spanning the second half of the twentieth century, constructed by the United Nations Population Division (1999) from the research findings described above, are shown in Table 1. Such estimates can hardly be exact but they are probably reasonably close to the truth. They are compared with estimates for the world and selected other developing regions.

What emerges from Table 1 is that sub-Saharan Africa is a healthier place than it was half a century ago. By the end of the twentieth century its life expectancy was above that of the world as a whole at mid-century, and only a little behind that of Latin America at that time. Nevertheless, it had fallen further behind every other world region. The most appropriate comparison is with south Asia, which has a somewhat lower per capita income than sub-Saharan Africa (and, taking purchasing power into account, a lower per capita parity purchasing power) and a slightly higher level of female literacy. In 45 years, Africa's life expectancy rose by 12 years, compared with 23 years in south Asia.

Table 2 explores the African health failure further. Neither sub-Saharan Africa nor North Africa participated to the same extent as Asia and Latin America in the great leap forward in reducing mortality that characterized much of the world during the two decades following World War II. Nevertheless, sub-Saharan Africa's record in health advancement was moderately successful until the 1980s, but since about 1985 there has been little health advance at all.

The relatively limited sub-Saharan African success in health up until the 1980s can be at least partly explained by the fact that ecological and other conditions meant that it was the only major world region that completely failed to reduce the level of malaria. Since the 1980s two other factors have also been important. The first was a slowdown in economic growth that necessitated the acceptance of 'structural adjustment' policies: in many countries, investment in the health sector almost ceased, and charges for government medical services were instituted. The second factor was the arrival of the AIDS epidemic with an intensity experienced nowhere else in the world.

All parts of sub-Saharan Africa have experienced a deceleration in health improvement, but this has been spread unevenly, as Table 3 shows. The health situation in the region had been described as one improving along a diagonal with the lowest levels in the northwest of West Africa and the highest levels in

Table 3
Life expectancies and average increases for the major regions of Africa, 1950–2000

	Life expectancy at birth (years)				Average annual increase in life expectancy since previous period (years)			
	West Africa	Middle Africa	East Africa	Southern Africa	West Africa	Middle Africa	East Africa	Southern Africa
1950–55	36	36	36	44	–	–	–	–
1960–65	40	40	41	49	0.4	0.4	0.4	0.5
1970–75	43	44	45	53	0.3	0.4	0.4	0.4
1980–85	46	48	47	56	0.3	0.4	0.2	0.3
1990–95	49	51	45	59	0.2	0.3	−0.2	0.3
1995–2000	50	50	45	54	0.2	−0.1	−0.2	−0.9

Source: As for Table 1.

Southern and East Africa (see Feachem and Jamison 1991, pp. 31–2). Largely as a result of the AIDS epidemic, this situation has been reversing, and by the 2010s West Africa may have the highest regional life expectancy south of the Sahara.

A very similar experience characterized trends in infant and child mortality rates. Sub-Saharan Africa's infant mortality rate in 1950–5 has been estimated as 176 deaths per 1000 births, with the south Asian rate at 186. By 1995–2000 the sub-Saharan African rate was 93 compared with a south Asian rate of 73. The child mortality rate compared even more badly; in 1995–2000 it was 152 in sub-Saharan Africa (i.e., 15.2 percent of births resulted in a death before five years of age) compared with 96, or less than two-thirds the African level, in south Asia. The explanation is what Cantrelle (1975) called the 'tropical pattern,' and others have termed the 'African pattern.' When Cantrelle was writing, not only was infant mortality in tropical Africa high, but almost as many deaths occurred to each cohort of births between their first and fifth birthdays as during the first year of life. That fraction has now fallen to 63 percent (although it is still 73 percent in West Africa) compared with 32 percent in south Asia. The reasons for very high one-to-four-year-old mortality in tropical Africa are probably the high level of infectious disease affecting that age group, poor weaning practices, and unsatisfactory foods available for weaning.

3. *Why Did Health Improve?*

It seems likely that life expectancy climbed by about 15 years between the 1880s and the early 1950s, and then by another 11 years in the next three decades. The reasons are complex, and the use of modern medicine is only part of the story.

Much of the explanation for mortality decline in the first period was probably the organization brought about by colonial governments. Inter-ethnic and individual violence probably declined. People were separated from others when plague struck, and from wild animals to curb sleeping sickness. Roads and railways helped to usher in a market economy which distributed food, and medicine, more widely. Capitalism and education increased individualism and made it likely that greater initiatives would be taken to prevent or cure illness. Immunization, led by smallpox vaccination, eventually brought the great epidemic diseases under control. Yellow fever has almost vanished, and cholera levels have declined. Digging drains and oiling stagnant water reduced malaria in towns, and mosquito nets partly protected colonial and local elites.

There was a slow spread of government and missionary hospitals. Although there is debate about the impact of modern medicine on poor, predominantly rural societies, a comparison of two areas of similar socioeconomic levels in Nigeria showed that the one which had possessed for a generation a small, adequately staffed and supplied hospital offering free services was characterized by a life expectancy 12 years greater than the area with no facilities (Orubuloye and Caldwell 1975). The treatment of water supplies, usually in urban areas and often inadequate, has improved, but is still often woefully bad. Better sanitation and hygiene practices have doubtless also reduced mortality. By the 1990s, Demographic and Health Surveys were reporting that about half of all countries had a majority of children immunized against tetanus, diphtheria, and pertussis (whooping cough), while the coverage against measles (a major killer in the region) was increasing and the incidence of poliomyelitis was declining steeply. But morbidity and mortality from diarrhea and pneumonia were still high. Malaria was almost as bad as ever, although many lives were saved by the use of drugs, and HIV/AIDS was presenting a horrific challenge.

4. *Health at the Dawn of the AIDS Era*

By the early 1980s sub-Saharan African life expectancy was 48 years, having increased by seven years in the previous decade, and seeming to promise similar gains

to come. This was not to be so, because of problems in funding the health system and the emergence of the AIDS epidemic. In contrast to most of Asia and North Africa, female child mortality was as low as that of males. There were significant mortality differentials by region (with mortality lowest in Southern Africa), ethnic group (even when neighbors), parental education (with mother's education being more important for child survival than father's education), by occupation (with farmers' death rates highest), and by residence (with urban health better). Mortality was highest in countries like Ethiopia and Mozambique where civil unrest and war had disorganized the health and other systems. It was also higher in the drought-prone savanna countries, but no higher than would be expected from their relatively low income and educational levels, which were in turn the product of impoverished agricultural resources. Drought still visited these lands regularly but evidence accumulated that, although it caused much distress and livestock loss, excess human mortality was less than might have been anticipated because of the scale of migration to better-off areas and towns.

Information on the nature of illness and the causes of death is still meager because so few people are seen by doctors or die in hospitals. By the 1980s the great epidemic diseases were largely under control. Campaigns, which mostly proved successful, aimed at eradicating onchocerciasis (river blindness), and some progress was being made against schistosomiasis (bilharziasis). Malaria, particularly its worst form, falciparum malaria, was almost universal, moderated only in the highest parts of East and Southern Africa by lower temperatures and in South Africa by a more temperate climate and successful eradication. A Gambian study revealed malaria to be the dominant cause of illness, except among children under three months of age, who were relatively free of it. Research in Tanzania showed that 31 percent of the sick were suffering from malaria, 13 percent respiratory infections and 7 percent diarrhea.

A major World Bank/Harvard University study (Murray and Lopez 1997) of global mortality estimated that 65 percent of the region's mortality was still attributed to communicable, congenital, maternal, and nutritional causes (compared with 51 percent in India, 42 percent in all developing countries, and 6 percent in developed countries), 23 percent by non-communicable disease, mostly cardiovascular and cancer (compared with 40 percent in India, 47 percent in all developing countries, and 86 percent in developed ones), and 12 percent by violence and accident (only a little higher than other areas). A World Bank survey (Feachem and Jamison 1991) found that 47 percent of deaths were from infectious and parasitic causes (malaria and measles being particularly important), which, together with perinatal problems, accounted for nearly all child deaths, and 16 percent from circulatory disease and cancer (but their category of 'other,' including indefinable, was 23 percent). A longitudinal study in Machakos, Kenya (Muller and van Ginneken 1991), an area where altitude renders malaria a minor complaint, ascribed 16 percent of deaths to respiratory infection, 13 percent to congenital factors, 11 percent to intestinal infections, 8 percent to measles, 7 percent to tuberculosis, 7 percent to other infectious and parasitic diseases, 4 percent to nutritional and metabolic causes, 4 percent to problems of the digestive system, 3 percent to malaria, 11 percent to diseases of the circulatory system and cancer, and 6 percent to violence and injuries, while 11 percent could not be determined. Hepatitis B also presents dangers. Maternal mortality has been estimated at 655 per 100,000 births, the world's highest level and 30 times that of Europe. This, together with deaths arising from miscarriages and abortion, translates, because African women still average six live births, into a lifetime chance for females of dying from maternity causes of about 5 percent. It is likely that the lifetime risk will fall substantially in the near future as the result of a fertility decline.

5. *The AIDS Epidemic*

HIV-2, perhaps the more ancient of the two human immunological retrovirus strains, is present in West Africa, Angola and Mozambique, but, because it is not a major cause of mortality, will not be discussed here. The HIV-1/AIDS epidemic began in East Africa in the early 1980s, at much the same time as it appeared elsewhere. By the mid-1990s prevalence rates in Southern Africa were higher than they have ever been in East Africa. UNAIDS/WHO (1998) estimates for the end of 1997 indicate that 9.6 million Africans have already died of AIDS (82 percent of the world total) and a further 21 million are now infected (69 percent of the world total). More startlingly, there are 15 countries, constituting the 'Main AIDS Belt' and stretching from Ethiopia through East Africa to South Africa, which contain 4 percent of the world's population and half the world's HIV/AIDS. All of them have adult HIV prevalence rates of at least 9 percent, but the rate reaches 15 percent in Malawi and Mozambique, around 20 percent in Namibia and Swaziland, and 25 percent in Zimbabwe and Botswana. The latter levels imply more than a doubling of the death rate, with a lifetime expectation of dying of AIDS over 50 percent. United Nations population projections do not anticipate Zimbabwe regaining its 1985 life expectancy without a gap of 35 years, or Botswana with a smaller gap than 50 years. Sub-Saharan Africa's epidemic is almost entirely heterosexual in primary transmission, with the result that at least as many women as men are infected. Thus, although the region contains only 59 percent of the infected men in the world, it has 81 percent of infected women, and, because of higher birthrates, probably at

least 85 percent of infected children. Thus, almost uniquely in the world, all parts of the community are affected. The AIDS epidemic is also catalyzing an increase in tuberculosis levels. There is growing evidence that HIV-positive women are likely to be rendered infertile or subfertile. Campaigns to contain the epidemic have been less intensive than the crisis demands, and largely ineffective.

6. *The African Health System*

Government hospitals are found chiefly in towns, with teaching or specialist referral hospitals in the capitals or other large cities. In most countries there are also missionary hospitals, often in rural areas. Nevertheless, most rural people depend on primary health services in the form of village health posts, clinics or dispensaries, and health centers. This system is developed to very different degrees across the region, and the health facilities often have insufficient drugs. Most have now moved to a 'user-pays' system, often characterized by declining attendances and by women losing their ability to take their children straight to the facilities without consulting male relatives (Orubuloye et al. 1991). Research has shown that most clients of health centers come from no further away than the village in which the center is located.

The system is supplemented by private pharmacies and medical stores. A Nigerian study showed that in rural areas the majority of drugs sold are malaria suppressants, worm syrups, and analgesics, although antibiotics are usually also in stock. Private doctors and nurses are becoming of increasing significance in most countries. Traditional practitioners remain important. They do not have an agreed-upon pharmacopeia as in various Asian medical systems, and their medicines are derived from animals as well as herbs. Many also identify evil forces and reveal how they can be nullified.

The great campaigns which eliminated smallpox and contained sleeping sickness, yellow fever, and cholera are largely things of the past, although watch is kept for outbreaks. Efforts continue to reduce the incidence of leprosy and tuberculosis, and to ensure that those people with malarial fever do not die. Immunization of children is now a major health weapon, and recent years have seen sustained attempts to control measles. Public health efforts continue to ensure safe drinking water and sanitation, but it is unlikely that the levels of safety are as high as respondents report to the international survey programs.

7. *Trends in the Research Literature*

More articles on the social and behavioral aspects of health are published in the journal *Social Science and Medicine* than anywhere else. In the last quarter of the twentieth century 865 of these papers were on sub-Saharan Africa. Their numbers rose rapidly at first from an annual average of nine during the late 1970s to 24 in the early 1980s and then reached a plateau at just under 50 per year. Many of the articles were derived from empirical studies of specific locations or diseases. The major topics were access to health services (132 papers), the interaction of traditional and modern health beliefs and services (75 papers), HIV/AIDS (68 papers, nearly all in the 1990s), maternal and child health (55 papers), the political economy of health (48 papers), and health economics (39 papers). Some trends were noticeable. The number of papers published on the interaction between traditional and modern health services declined after the 1980s, those on the political economy of health peaked in the late 1980s as economic structural adjustment policies were applied, and HIV/AIDS papers mostly appeared in the 1990s, with no clear trend during the decade.

8. *The Future*

The immediate future is not reassuring. In contrast to the rest of the world, the regional life expectancy is stalled at under 50 years. Most sub-Saharan African countries are unable to do more than maintain their present public health systems. The impact of such systems depends on their level of use, which is largely determined by the population's ability to pay, and by their level of education, which raises the priority given to successful treatment and the skill with which it is administered. The fee-for-service principle is limiting the expansion of use of both the health and education systems. On the other hand, the likelihood of accessing the health system is probably increasing with the transition from subsistence to market agriculture. Fertility decline has now begun in a range of African countries. Smaller families will mean fewer women dying from maternal causes, and probably greater concentration on the education and health of children.

It is now likely that no country in mainland sub-Saharan Africa will, by the year 2000, have attained any of the three criteria established by the World Health Organization to certify the attainment of the 1978 Alma Ata Declaration of 'Good Health for All' by the year 2000: a life expectancy of at least 60 years, and a survival level of 95 percent of births to one year of age and 93 percent to five years of age. Indeed, most of East and Southern Africa will probably by that date be slipping further away from these criteria.

See also: African Studies: Politics; AIDS, Geography of; Health: Anthropological Aspects; Health in Developing Countries: Cultural Concerns; Mortality and the HIV/AIDS Epidemic; Mortality, Biodemography of

Bibliography

Awusabo-Asare K, Boerma T, Zaba B (eds.) 1997 Evidence of the socio-demographic impact of AIDS in Africa. *Health Transition Review* **7** (suppl. 2)

Brass W, Coale A J, Demeny P, Heisel D F, Lorimer F, Romaniuk A, van de Walle E 1968 *The Demography of Tropical Africa*. Princeton University Press, Princeton, NJ

Caldwell J C (ed.) 1975 *Population Growth and Socioeconomic Change in West Africa*. Columbia University Press, New York

Caldwell J C 1979 Education as a factor of mortality decline: An examination of Nigerian data. *Population Studies* **33**: 395–413

Caldwell J C 1985 The social repercussions of colonial rule: Demographic aspects. In: Boahen A A (ed.) *General History of Africa, vol. 7: Africa under Colonial Domination 1880–1935*. University of California Press, Berkeley, CA and Heinemann, London for UNESCO, Paris

Cantrelle P 1975 Mortality: Levels and trends. In: Caldwell J C (ed.) *Population Growth and Socioeconomic Change in West Africa*. Columbia University Press, New York

Cleland J, Way P (eds.) 1994 AIDS impact and prevention in the developing world: Demographic and social science perspectives. *Health Transition Review* **4** (suppl.)

Coale A J, Demeny P 1966 *Regional Model Life Tables and Stable Populations*. Princeton University Press, Princeton, NJ

Committee on Population, National Research Council 1993 *Demographic Effects of Economic Reversals in Sub-Saharan Africa*. National Academy Press, Washington DC

Farah A-A, Preston S H 1982 Child mortality differentials in Sudan. *Population and Development Review* **8**: 365–83

Feachem R G, Jamison D T (eds.) 1991 *Disease and Mortality in Sub-Saharan Africa*. Oxford University Press, New York for World Bank, Washington, DC

Foote K, Hill K H, Martin L (eds.) 1993 *Demographic Change in Sub-Saharan Africa*. National Academy Press, Washington, DC

Hill A 1992 Trends in childhood mortality in sub-Saharan Mainland Africa. In: van de Walle E, Pison G, Sala-Diakanda S (eds.) *Mortality and Society in Sub-Saharan Africa*. Clarendon, Oxford

Muller A S, van Ginneken J 1991 Morbidity and mortality in Machakos, Kenya. In: Feachem R G, Jamison D T (eds.) *Disease and Mortality in Sub-Saharan Africa*. Oxford University Press, New York

Murray C J L, Lopez A D (eds.) 1996 *The Global Burden of Disease*. Harvard University Press, Cambridge, MA for World Bank, Washington, DC

Murray C J L, Lopez A D 1997 Global mortality, disability, and the contribution of risk factors: Global burden of disease study. *Lancet* **349**, 17 May: 1436–42

Orubuloye I O, Caldwell J C 1975 The impact of public health services on mortality: A study of mortality differentials in a rural area of Nigeria. *Population Studies* **29**: 259–72

Orubuloye I O, Caldwell J C, Caldwell P, Bledsoe C H 1991 The impact of family and budget structure on health treatment in Nigeria. *Health Transition Review* **1**: 189–210

Population Reference Bureau 1999 *World Population Data Sheet 1999*. Washington, DC

Timaeus I M 1999 Mortality in sub-Saharan Africa. In: Chamie J, Cliquet R L (eds.) *Health and Mortality: Issues of Global Concern*. United Nations, New York

United Nations, Population Division 1999 *World Population Prospects: The 1998 Revision*, vol. 1, *Comprehensive Tables*. United Nations, New York

UNAIDS/WHO 1998 *Report on the Global HIV/AIDS Epidemic June 1998*. Joint United Nations Program on HIV/AIDS and World Health Organization, Geneva

Van de Walle E, Pison G, Sala-Diakanda M (eds.) 1992 *Mortality and Society in Sub-Saharan Africa*. Clarendon Press, Oxford

World Bank 1999 *World Development Report 1998/99*. Oxford University Press, New York

J. C. Caldwell

African Studies: History

Central to the origins of African history as a field of inquiry is the quest to demonstrate that such a subject actually existed. Opinion leaders and intellectuals in Europe and North America had long treated Africa as the embodiment of the primitive. Nineteenth and early twentieth century science marked it a place where earlier stages of evolution could still be observed or else (as in structural–functional anthropology) as a laboratory of social specificity, where forms of social organization could be compared as if each were bounded and timeless. The decolonization of Africa shook up intellectual understandings as well as political arrangements, and from the late 1940s, some intellectuals inside and outside Africa began to argue that rethinking Africa's past was a necessary part of its future. Asserting that history could be studied scientifically was part of a new politics of intellectual inquiry. Pioneering studies stressed the dynamism of precolonial societies; resistance to conquest was a harbinger of nationalist movements. But by the 1970s, a more complicated present was leading to a more complicated past, above all to new ways of thinking about Africa's relationship to the rest of the world and the implications of this relationship for historical writing itself. Yet the time frame for considering the emergence of African history needs to be pushed back even further, and its spatial dimension—the definition of Africa, the relationship of its constituent social and political units, and the significance of the continent to Africans in the Americas—needs examination as well.

1. *Africa and the World*

Africa was in part an invention of its diaspora, a unit that became of world–historical significance because slave traders—from the sixteenth century—defined it as a place where one could legitimately develop a commerce in human beings. Over time, enslaved Africans and their descendants in the Americas began to appreciate the commonality of their fate and many looked to 'Africa' as an almost mythic symbol that they were not mere chattel who could only serve their

owners. Certain nineteenth-century African–American religious leaders looked toward 'Africa' and 'Ethiopia' (although few slaves came from that kingdom), and through such language asserted Africans' important place in a universal history—in the unfolding of Christian civilization. As some people of African descent returned to the continent in the nineteenth century—repatriated ex-slaves to the British colony of Sierra Leone, Brazilian traders in the Bight of Benin or Angola—some saw themselves as part of broad, translantic 'nations,' sharing ancestry and culture but needing Christianity to link a torn-apart past to a reintegrated future (Matory 1999).

By the late nineteenth century, Africans from coastal societies—Christian, Western-educated, but thoroughly integrated into regional social organizations—began to write about their own regions in terms that linked different historical sensibilities. Africanus Horton and Edmund Wilmot Blyden countered the primitivizing ideologies of the eras of the slave trade and colonization by writing about African societies as complex entities whose traditions of origin defined commonality, whose ideas about kingship and social hierarchy defined political order, and whose interest in commerce with the outside world, in Christianity and Islam, and in Western education marked an open, adaptable attitude to interaction. Africans, they thought, had much to learn from Muslims and Christians, but they brought something to the encounter as well.

The period of escalating European exploration and eventual conquest (from the 1870s to 1910s) brought to African contradictory regards. Some explorers encountered powerful kingdoms whose dimensions they sought to understand; others recognized in the impressive mosques of the West African desert edge or the East African coast a history of a long encounter of Africans with the outside world; but many chose to see an unchanging landscape of different peoples ensconsed in their particular cultures. The culturally mixed inhabitants of West Africa coastal region were marginalized by colonization; the worldwide connections of African Muslims were played down in favor of their 'tribal' characteristics.

After some efforts toward remaking African societies in a European image—the 'civilizing mission'—colonial regimes began to hitch their legitimacy to the chiefs whose authority toward their subjects colonial rulers needed. With that, the idea of 'tradition' as the essential quality of African life acquired a new salience in colonial ideologies. Colonial regimes—and scholars and intellectuals of that era—were interested in 'customary law,' in 'folklore,' in 'primitive art.' The growth of African ethnography in the 1920s brought to the continent foreign scholars curious about the diversity of social forms and sympathetic to victims of colonial oppression, but their emphasis on the bounded integrity of each African 'society' was ahistorical.

2. *A New Past for a New Future*

Even within the ethnic cages of colonial polities, consciousness of the past did not necessarily remain static. Some early mission converts used their literacy in French or English to record genealogies and traditions of 'their' people—using the legitimacy of 'European' writing to articulate indigenous views of the past and to emphasize the integrity of local society. Such histories were invoked to make claims for collective representation in state-sanctioned councils. Meanwhile, pan-Africanists like W. E. B. duBois countered the belittling conceptions of colonial ideology by emphasizing the importance of long history of oppression shared by Africans and African–Americans, a history which underscored the importance of the liberation movements.

The burst of interest in African history after World War II thus has a deeper context. The break, however, was fundamental. In the aftermath of a devastating war against Naziism and conquest, European states needed simultaneously to justify their rule over African peoples and to intensify their use of African resources. The dilemmas posed by challenges to legitimacy and control—from within and outside African colonies—and the increased intrusiveness of regimes into African social and economic life has made this a fascinating period for historical investigation (Marseille 1984, Cooper 1996). Even at the time, scholars and intellectuals wondered whether the conception of bounded and static units made sense of the Africa they were observing. Although a few anthropologists had seen in the 1930s that migration, for example, was redefining the nature of social connections, by the 1950s, population movement, cross-cultural interaction, and cultural adaptation demanded scholarly analysis. Meanwhile, the anthropologist Melville Herskovits (1958), had not only seen the importance of historical analysis to political anthropology but had become interested in the transmission of African culture to the new world via the slave trade, and he asked what Africans' experience of governing indigenous kingdoms had to offer to Africa's political future.

The last question was one most African political elites—and almost all social scientists—did not want to think about; the 1950s witnessed the escalation of claims from African social movements to be considered part of the 'modern' world and to enjoy the possibilities it entailed, irrespective of race or a past of colonization or enslavement. Social scientists took more interest in where history was supposed to end—'modernity'—than where people had been.

Academic historians, the professional custodians of the past, were part of the postwar political and intellectual ferment. The first African Ph.D. in history K. O. Dike of Nigeria, was trained by historians of imperial expansion, and turned their archival methodology into a means of legitimizing the telling of a

different sort of story, one of the interaction among traders, rulers, and warriors, both African and European, in the Niger Delta, and of the adaptation of African social institutions to new forms of competition in the nineteenth century. Dike (1956) insisted that oral sources could be used alongside written ones, although most of his own work was archival. At one level, his work staked itself on extending the methodological canons of history; at another his insistence that the locus of history could be found in Africa itself—that interaction was more important than transmission—was a charter for nationalist history.

The next generation of European-trained African historians went further in this direction, and many stressed explicitly that the preconquest past was a precedent for the postindependence future, showing how Africans had brought diverse populations into larger political units, how African initiative in agriculture and commerce had linked ecologically distinct regions with each other and with the outside world, how indigenous religious leaders had built networks that transcended ethnic frontiers, and how Africans had adapted Islam to particular political and cultural purposes.

The quest for a usable past—and a useable national past at that—was both African history's strength and its weakness in the 1960s. A strength, because it attracted a younger generation of Africans to believe that they could combine international scholarship with a sense of the past they had learned in their own communities and because it allowed Americans and Europeans to come to grips with the way in which the end of colonial empires had forced a reordering of intellectual categories. A weakness, for the political context privileged 'state building' over the diverse strategies of women and men, traders and religious figures to live their lives in different ways, and this both diminished a varied past and ratified an increasingly authoritarian present. The only aspect of colonial history that fitted the bill was 'resistance'; indeed, some African historians saw the 'colonial episode' as a short and not particularly important interval between an autonomous past and a promising future. A consequence of the new historiography, whatever its limitations, was a heightened interest in methodology, to solve the problem of how to reconstruct historical patterns with a paucity of written documentation, much of that from visitors and conquerors. Vansina helped to develop rigorous criteria for analyzing oral texts with the same critical eye as employed on written ones, and other scholars explored the use of linguistic and archeological material to chart the movements of people, the evolution of material culture, and the spatial configuration of state-building (Vansina 1965, Vansina et al. 1964).

African historical scholarship was vibrant in the 1960s and 1970s, and the core of the action was in Africa. Every nation had to have a university, every university a history department. Associations and journals were founded; international congresses were held; and UNESCO brought together African authors to write a comprehensive history of the continent (UNESCO 1981–93).

3. *Reconnecting Africa and the World*

Since the 1970s, African historians have become increasingly conscious of the different ways of rethinking the relationship of Africa and the rest of the world. They no longer had to prove that Africa had a history; the mounting evidence that decolonization did not free Africa from its problematic relationship to the world economy raised questions about pasts before and during colonization. Why African polities became caught up in the Atlantic slave trade, how the African presence in the Americas contributed to European wealth, and how African societies coped with unequal trade and unequal politics became topics of increasing interest (Rodney 1972). In addition, French Marxist anthropologists (Meillassoux 1975) argued that such institutions as kinship systems were not so much the product of a peculiarly African culture but of the logic of reproduction of agricultural societies, and they went from there to posit that the resulting modes of production interacted in particular ways with the expanding capitalist system. Other scholars saw that the question appeared differently if instead of asking how a 'society' responded to European traders or rulers one asked how ruling elites reacted. They showed that the strength of internal social groups made rulers look to the manipulation of external relations for power and wealth, and that this made African kingdoms and chiefdoms especially open to external relationships, most disastrously those of the slave trade and later of colonial economies (Peel 1983).

Historical research allowed not just the use of 'informants' to gain information, but the juxtaposition of different kinds of historical sensibilities and the elucidation of how thinking about the past affects and is affected by political processes in the present (Cohen 1994). Oral history revealed different lines of cleavage within African societies, including gender, age, and status. It described not the actions of fixed social groups against each other, but the flexibility of social arrangements, such as the tendency of people detached from kinship groups—through war or efforts to escape patriarchal authority—to form important but unstable groups of clients of 'big men.' Both oral sources and more nuanced reading of documents reopened the colonial era to historical investigation, not as a history of what white people did for or to Africans but as a dynamic process, in which the limited power of colonizing regimes left numerous fissures where women tried to establish autonomy even as male elders tried to contain it, where labor migrants developed networks linking distant regions, where Christians built independent churches and cults, and

where cash-crop producers used their incomes to enhance their kinship groups and chieftaincies. Such histories stand alongside those of expropriation of resources, especially in southern Africa, of segregation and discrimination against the most 'Westernized' of Africans, of arbitrary authority and daily oppression under colonial regimes, in their late, 'reformist' phases as well as brutal early ones.

Studying African history has offered the possibility of thinking about 'world' histories in a different way. Rather than seeing Africa as a peculiar place that somehow lacks what made other regions develop more rapidly, one can ask what is particular about each region of the world, what in the process of interaction produced inequalities of wealth and power. African historians have been in dialogue with historians of Latin America over economic history and with historians of India over the analysis of colonialism. World history can no longer be seen as a single narrative, but this recognition leaves in place the more difficult question of how is one then to analyze large-scale historical processes. The challenge is to chart the multiple pathways without losing sight of the development of highly unequal political and economic relations on a global scale. In the eighteenth century, European countries were perhaps five times as wealthy as African regions; now the gap is as high as 400 to 1. Such an historical change requires analysis that is about neither 'Africa,' nor 'Europe,' but their relationship in the last 200 years.

4. *The Pecularities of Disciplinary Divisions of Labor*

In the 1930s, Africa not only had no history; it held no interest for political scientists or sociologists either, and for economists only in the sense the colonial economies had spawned 'modern' sectors with certain measurable characteristics. Africa was the domain of anthropologists, the custodians of human particularity, while what was putatively 'European' was assimilated to the universal, and that was where social science found its glory. The dramatic challenges to European power in the 1940s and 1950s shook up this division of labor (Pletsch 1981). The idea of modernization was espoused within colonial bureaucracies before 'modernization theory'—in the style of the Committee on Comparative Politics of the Social Science Research Council—came into vogue. It allowed officials to think that even if African colonies became independent, they would invariably follow a road charted out to them by the early developers. However much modernization reinscribed global hierarchy, eurocentrism, and teleology, it offered African intellectuals and policy-makers a chance to position themselves as mediators between a perceived Euro–American modernity and diverse African particularities; it gave a language to leaders with which to ask industrial countries for the resources they needed to 'develop' and it allowed outsiders to see themselves as forging a world in which people of all races could advance.

American and European political scientists and sociologists flocked for a time to Africa—and their standards of empirical research deserve some of the credit for deflating modernization theory. But after that time, the interest of political scientists, sociologists, and economists in Africa has waned. Anthropologists remain the custodians of African particularity, but now this embraces the complexities of innovation and interaction more than unchanging specificities. Historians, from the 1950s, were the new element in the division of labor, and their work, despite temptations to the contrary, can offer a vision of Africa that stresses change—economic, political, social, and cultural—without confining it to a predetermined route.

Some historians have their version of the modernizing vision. The narrative of African state-building from the empire of Sundiata to the empire of Shaka was indeed a selective reading that linked a modernizing future to a modernized past. One British historian (Wrigley 1971) accused another (Fage 1969) of missing the moral and political significance of the slave trade by putting it inside a narrative of centralizing of power by successful slave-exporting West African kingdoms. Even the 'underdevelopment' school of the 1970s is as a variant of modernizing themes, for its emphasis on how over the long-term European exploitation retarded growth and transformation in African economies presumes a global narrative of capitalist development which Africa is then held to fall *under*. That Europe's oppression, not Africa's backwardness, is held culpable does not negate the inattention to the myriad innovations, struggles, and changes that have occurred within particular parts of Africa, the importance of regional, not just overseas, mechanisms of exchange and communication, and the ways in which the actions of Africans limited the actions of Europeans, whatever their intentions (Cooper et al. 1993). Much recent social and economic history bounced off the arguments of the underdevelopment school to develop more varied and more interactive views of change in various periods of African history, including those when European power was seemingly at its height (Berry 1993).

5. *Multiple Perspectives on a Varied Continent*

If Africa is neither a homogeneous space reduced into abject poverty by imperialism nor a series of autonomous societies with their own internal logics, getting a grip on units of analysis is not simple. Rather than take present-day ethnicity as a starting point, some scholars have used concepts of region, of network, and of patron–client relations to see how people in the

nineteenth and twentieth centuries constituted actual patterns of relationships, which only sometimes crystallized into groups that maintained boundaries (Ambler 1988, Glassman 1995, Barry 1998). Others have emphasized how pioneers of urban migration shaped patterns that linked particular villages with African cities (and Parisian suburbs), how religious pilgrimages and circuits of Koranic scholars created linkages across inland West Africa with further connections to Egypt and Saudi Arabia, how networks of shrines and spirit mediums shaped cross-ethnic affiliations in a large belt of Central Africa, how African American missionaries from the 1890s influenced Christianity in South Africa, and how in the 1920s African ports became nodal points in the spread of Garveyite movements across the Atlantic to large parts of Africa (Manchuelle 1997, Campbell 1995).

From there, scholars can examine the different visions of sociability held by women and men, by young and old, by elites and ordinary people. They can study not only the differential experience, say, of men and women in agriculture under precolonial and colonial regimes, but the ways in which the categories of gender and age—and the affinities and conflicts they entailed—were created and struggled over (Mandala 1990, Grosz-Ngaté and Kokole 1997). They can look at different forms of imagination and communication (White 2000).

6. Conclusion

The Senegalese novelist and film-maker Ousmanne Sembene has accused historians of being 'chronophages'—eaters of time. They impose, he insists, a one-dimensional, progressive view of time on the unruliness of people's experience. Yet the idea that other notions of past give more a more varied picture than professional history depends on aggregating different visions of the past—a quintessentially 'academic' operation. The griot or the lineage elder may also recount the past to serve narrow, presentist concerns. Historical scholarship has complicated unilinear narratives as well as imposed them.

Reflecting on historical processes confronts us with the tension between the contingency of processes and the fact of outcomes, that multiple possibilities narrowed into singular resolutions, yet led to new configurations of possibilities. The doing of history introduces a tension between an historian's imagination rooted in the present and the fragments of the past that appear in all their elusive vigor, in interviews, letters, newspaper articles, and court records. The data of history are not created neutral or equal: archives preserve certain documents but not others, and the tellers of oral tradition remember some narratives and forget others (Mudimbe 1988).

African history has offered a glorious national past that points to a national future, and it has offered an ethnicized past projected backward in time. But history can also be read as process, choice, contingency, and explanation. Since World War II, historical scholarship has provided a sense of the possibilities which mobilization can open up—and an awareness of the constraints that have made and make it so difficult for African states and societies to make their way in the world.

See also: African Studies: Culture; African Studies: Politics; African Studies: Religion; Central Africa: Sociocultural Aspects; Colonialism, Anthropology of; Colonialism: Political Aspects; Colonization and Colonialism, History of; Development: Social-anthropological Aspects; Diaspora; East Africa: Sociocultural Aspects; Historiography and Historical Thought: Sub-Saharan Africa; Southern Africa: Sociocultural Aspects; West Africa: Sociocultural Aspects

Bibliography

Ambler C 1988 *Kenyan Communities in the Age of Imperialism: The Central Region in the Late Nineteenth Century*. Yale University Press, New Haven, CT

Barry B 1998 *Senegambia and the Atlantic Slave Trade* (Trans. Armah A K). Cambridge University Press, Cambridge, UK

Berry S 1993 *No Condition is Permanent: The Social Dynamics of Agrarian Change in Sub-Saharan Africa*. University of Wisconsin Press, Madison, WI

Campbell J 1995 *Songs of Zion: The African Methodist Episcopal Church in the United States and South Africa*. Oxford University Press, New York

Cohen D W 1994 *The Combing of History*. University of Chicago Press, Chicago

Cooper F 1996 *Decolonization and African Society: The Labor Question in French and British Africa*. Cambridge University Press, Cambridge, UK

Cooper F, Isaacman A, Mallon F, Roseberry W, Stern S 1993 *Confronting Historical Paradigms: Peasants, Labor, and the Capitalist World System in Africa and Latin America*. University of Wisconsin Press, Madison, WI

Dike K O 1956 *Trade and Politics on the Niger Delta*. Clarendon Press, Oxford, UK

Fage J 1969 Slavery and the slave trade in the context of West African history. *Journal of African History* **10**: 393–404

Glassman J 1995 *Feasts and Riot: Revelry, Rebellion, and Popular Consciousness on the Swahili Coast, 1856–1888*. Heinemann, Portsmouth, NH

Grosz-Ngaté M, Kokole O (eds.) 1997 *Gendered Encounters: Challenging Cultural Boundaries and Social Hierarchies in Africa*. Routledge, New York

Herskovits M 1958 *The Myth of the Negro Past*. Beacon Press, Boston

Manchuelle F 1997 *Willing Migrants: Soninke Labor Diasporas, 1848–1960*. Ohio University Press, Athens, OH

Mandala E 1990 *Work and Control in a Peasant Economy: A History of the Lower Tchiri Valley in Malawi, 1859–1960*. University of Wisconsin Press, Madison, WI

Marseille J 1984 *Empire colonial et capitalisme français: Histoire d'un divorce*. Albin Michel, Paris

Matory J L 1999 The English professors of Brazil: On the diasporic roots of the Yoruba nation. *Comparative Studies in Society and History* **41**: 72–103

Meillassoux C 1975 *Femmes, greniers et capitaux*. Maspero, Paris
Mudimbe V Y 1988 *The Invention of Africa: Gnosis, Philosophy, and the Order of Knowledge*. Indiana University Press, Bloomington, IN
Peel J D Y 1983 *Ijeshas and Nigerians: The Incorporation of a Yoruba Kingdom, 1890s–1970s*. Cambridge University Press, Cambridge, UK
Pletsch C 1981 The three worlds, or the division of social scientific labor, circa 1950–1975. *Comparative Studies in Society and History* **23**: 565–90
Rodney W 1972 *How Europe Underdeveloped Africa*. Bogle-L'Ouverture, London
UNESCO 1981–93 *General History of Africa*. University of California Press, Berkeley, CA
Vansina J 1965 *Oral Tradition: A Study in Historical Methodology* (trans. Wright H M). Aldine, Chicago
Vansina J, Mauny R, Thomas L V (eds.) 1964 The historian in tropical Africa. Oxford University Press, London
White L 2000 *Speaking with Vampires: Rumor and History in Colonial Africa*. University of California Press, Berkeley, CA
Wrigley C C 1971 Historicism in Africa: Slavery and state formation. *African Affairs* **70**: 113–24

F. Cooper

African Studies: Politics

1. Introduction

African politics, as a distinct field of inquiry, essentially begins in the 1950s, at a moment when the rise of African nationalism and foreshortening timetables for decolonization created the prospect of an early entry into the world system of more than 50 new states. The initial focus was African nationalism and the political parties through which it found expression. The rapid succession of dominant political forms and preoccupations brought corresponding shifts in analytical focus: transitions to independence; single-party systems; military intervention; ideological radicalization; patrimonial rule; state crisis and decline; economic and political liberalization. Interactively, a series of paradigmatic orientations shaped the study of African politics: modernization, dependency and neo-Marxism, rational choice, democratic transition and consolidation.

2. Scope and Nature of Field

Officially, the African state system defines itself as coincident with the geographic continent and its offshore islands, symbolized in the membership of the Organization of African Unity (OAU). As a corpus of knowledge, however, African politics most frequently refers to the sub-Saharan states, which share a large range of cultural, sociological, and political traits. South African politics represents a distinct sphere, rendered exceptional until the 1990s by that country's apartheid regime. Africa has an extraordinary number of sovereign units (53 in 1999); however, comparative understandings of African political dynamics derive from a much smaller number of states that, by reason of their size, accessibility for research, or attractiveness as models, received disproportionate attention (for example, Nigeria, Tanzania, Kenya, Senegal, and Congo–Kinshasa).

Some particular aspects of the sociology of Africanist political knowledge merit note. There is a singular preponderance of external scholarship, North American and European, which only recently began to be balanced by African contributions. At first the paucity of African academics explained this phenomenon; subsequently, until the 1990s, most regimes had low tolerance for critical scholarship from their nationals, and by the 1970s the severe material deterioration of many African universities inhibited research from within. Methodologically, most scholarship relied upon qualitative approaches, which sharpened the debate around contending broad theoretical paradigms shaping such inquiry.

3. Decolonization and the Origins of African Politics as a Field

Until the 1950s, there were two distinct domains within African politics. At the summit lay the colonial apparatus, whose study was restricted to an administrative science conducted largely by colonial practitioners. At the base lay subordinated African societies, whose study was confined to anthropologists, missionaries, and administrators. Knowledge thus generated influenced anthropological theory and the practice of 'native administration,' but was outside the realm of comparative politics.

African nationalism emerged as a potent political force in the 1950s; its leading students (for example, Coleman 1958) are the foundational generation of the African politics field. The defining attribute of African nationalism was its autonomy from the colonial state it sought to challenge, thus constituting an authentic field of African politics. In its rendering of nationalism as a doctrine of liberation, its African form offered a dual summons to solidarity: as African, whether understood racially or continentally, and as territorial subject of the unit of colonial administration. The sources and content of nationalist thought, and the organizations through which it found expression, animated a first generation of scholars deeply sympathetic to its ends.

With the approach of independence, focus shifted to political parties. In the transitional arrangements, parties were the first institution of representative government to operate in African hands. Their in-

ternal dynamics and competitive struggle best captured the essence of African politics, well before states were 'African.' As well, their capacity to reach and mobilize urban and rural mass audiences seemed a measure of the potential for genuinely representative and hence legitimate postcolonial rule. One particularly influential school held that the prospects for stable, effective, and democratic rule were best fulfilled by mass movements successful in winning support spanning all ethnic and social groups, and thus enjoying a universal mandate as a single party (Morgenthau 1964).

4. *Modernization Theory and the Early Independence Years*

4.1 *The Single-party System*

With independence achieved, analysis shifted to the political development of new states, along with a substantial infusion of comparative political theory, greatly influenced by the Social Science Research Council Committee on Comparative Politics. The various strands of modernization theory, rooted in the premise of duality of tradition and modernity, privileged the state as indispensable central instrument of progress. Apter (1955), drawing upon Weberian theories of legitimacy, saw the key to effective modernization in the ability of the nationalist leader to transform the personal charisma achieved in this role into a routinized form of state legitimation. The nationalist parties, once in power, flowed into the state apparatus and soon lost their organizational distinctiveness; Zolberg, in a prescient study (1966), identified the party-state as dominant form, and pointed to emergent trends towards political monopolies.

With rare exceptions, the dominant parties which assumed power with independence sought to consolidate their exclusive hold on power, and to co-opt, circumscribe, or often proscribe opposition. Development, political and economic, required the undivided exercise of state power. Open opposition was likely to play upon ethnic or religious divisions, and to politicize cultural identities. Tradition resided at the periphery, the agencies of modernity at the center. Thus the widespread choice of African rulers for centralization and unification of the state, represented as nation in formation, was largely shared in the first years of independence by academic observers.

4.2 *Military Intervention*

The limitations of the first formulas for postcolonial rule stood exposed in 1965–6 when within a few months a wave of military coups occurred (Ghana, Algeria, Nigeria, Congo-Kinshasa, Benin, Central African Republic, Burkina Faso). At the time, comparative political sociology of military regimes in the developing areas supplied an unintended brief in support of such interventions. Armies, ran the argument, could serve as positive managers of modernization. The rationality and hierarchy of their internal structures, their technocratic skills, the merit basis of promotion, and their dedication to the nation equipped them for a beneficial developmental role.

Though military coups were generally justified as transitional cleansing operations, to be followed soon by restoration of civilian rule, almost invariably the new rulers concluded that national interests were best served by the permanence of their rule. To legitimate this consolidation of their rule, they adopted the political instruments fashioned by the nationalist movements in power: the single party, accompanied by co-option of prominent civilian politicians. Actual performance of such regimes proved indistinguishable from that of other party-states, deflating the military-as-modernizer theories. Closely inspected, military intervention motives bore little relation to public interest notions (Decalo 1976).

5. *The Authoritarian State Interrogated*

5.1 *The Demise of Modernization Theory*

Virtually apace, the cluster of approaches labeled 'modernization theory' lost their hold on political inquiry, and the nature of rule predominant in Africa found its credibility beginning to erode. The master concept of modernization encountered sharp criticism for its linear notions of change, its insensitivity to social cleavage and conflict, and its teleological concept of progress. A silent shift in focus took place with respect to perspectives concerning African regimes: from custodians of development to authoritarian states. In developmental perspective, the core analytic was state action in overcoming a crisis of development: integration, legitimation, penetration, distribution. Influential studies in Latin America and Asia pointed in a different direction: the nature of the state itself, and the mechanisms by which it assured the reproduction of its power. Although Latin American notions of 'bureaucratic-authoritarian' or 'national security' states did not apply, save possibly to South Africa, new currents of reflection on authoritarianism itself altered the problematic of African political inquiry.

In turn, the 1970s were a period of expanding state ambitions and further reinforcement of the centralizing and unitary state impulses. For a number, radical ideological impulses surfaced; for seven countries, these took the form of officially proclaimed Marxist–Leninist state doctrine. For others, sweeping nationalization or indigenization projects vastly enlarged the domain of state economic control. Elsewhere, the 1973 Algerian 'agrarian revolution' or the 1976 Tanzanian venture in enforced rural resettlement

in villages exemplified a mood of enlarging state aspirations for accelerated development under its control and management. But the reach far exceeded its grasp.

5.2 Dependency and Neo-Marxist Theories

The newly critical perspective towards the postcolonial state, as well as modernization theory, found potent expression in a family of conceptual approaches drawing in one way or another on Marxism. Dependency theory, which enjoyed virtual ascendancy for an extended period in Latin America, crossed the Atlantic and achieved broad influence in Africa, not only amongst scholars and intellectuals but also in ruling circles. From the dependency perspective, the issue was less the authoritarian character of the state than the class dynamics and international capitalist system which made it so. The extroverted nature of African economies, the control of the international exchanges by capital in the imperial centers, and the subordination of the domestic ruling class to the requirements of international capitalism led ineluctably to an authoritarian state to repress and control popular forces. In dependency theory's most sophisticated formulation, Leys (1974) employed a dependency perspective to question the character of development in Kenya, then still viewed as a narrative of success.

Beyond dependency theory, various currents of Marxism enjoyed intellectual influence, particularly amongst the African intellectual community in French-speaking Africa. An important revival of Western Marxism in the 1960s transcended the doctrinaire rigidities of Marxism–Leninism, and shaped an unfolding quest to resolve the riddle of the nature of class relations in Africa, a prerequisite to grasping the character of the state. Particularly interesting, though ultimately abandoned as unproductive, was the search for a 'lineage mode of production,' which could deduce class dynamics from the descent-based small-scale structures of rural society. Other hands sought to employ the ornate abstractions of the structural Marxism of Louis Althusser and Nicos Poulantzas. In the end, the schema supplied by dependency theory and neo-Marxism lost ground over the course of the 1980s. The sudden collapse of the Soviet Union in 1991 was a shattering blow in Africa as elsewhere to the credibility of Marxism, whether as regime doctrine or analytical instrument.

6. State Decline and Crisis

By the 1980s, patterns of decay became apparent in a number of states, and the notion of state crisis entered the analytical vocabulary. Corruption on a large, sometimes colossal scale became apparent; a prime example being that perpetrated by Mobutu Sese Seko, the ruler of Congo/Zaire. The rents extracted from power by ruling groups reached a magnitude which transformed citizen perceptions of states from provider to predator.

Another concept drawn from Weber, patrimonial rule, emerged as the key to understanding political practice. 'The politics of the belly,' wrote Bayart (1993), produced a 'rhizome state,' whose tangled underground root system of patron–client networks rather than its formal structure governed its operation. Office was a prebend, whose rents rewarded the occupant for service to the prince. The essence of African politics was 'big man' personal rule, or neopatrimonialism, rather than authoritanianism.

The inevitable consequence of the tentacular expansion of state ownership and control of the formal economy, and decline of government effectiveness, was a deepening economic crisis. The per capita GNP of Ghana comfortably exceeded that of South Korea at the time of independence; by 1995, that of South Korea was 25 times greater, a pattern found across much of the continent. Employing rational choice theory, Bates (1981) showed how the political logic of state operation systematically disfavored the rural sector, main source of wealth for all but a handful of oil rentier states. Hyden (1981) pointed toward the peasant response, in recourse to the exit option of the informal economy and secure reciprocities of the local kinship matrix.

7. Economic and Political Liberalization

The unmistakable economic stagnation and symptoms of state crisis drew the international financial institutions into the fray, at a moment when newly dominant political economy perspectives in the Western world called for far-reaching curtailment of the orbit of state action: privatization, deregulation, budgetary austerity, and rigor. Although a need for economic reform was acknowledged on all sides, deep divergences existed on the diagnosis of the core causes, and appropriate remedies. The 'Washington consensus' held that the explanations lay mainly in flawed domestic policies, while much African opinion, both official and scholarly, believed the root causes lay in the unjust operation of the international economy. The international financial institutions developed a standard package of 'structural adjustment programs,' holding the upper hand in the bargaining. However, these formulas were only fitfully applied, producing uneven results and strong domestic disapproval for what critics argued were their negative social consequences. The dilemmas of structural adjustment define an important fraction of African political studies of the 1980s and 1990s (Callaghy and Ravenhill 1993).

The failure of reform to reverse economic decline in the 1980s led a growing number of voices to suggest that the underlying flaw lay in the patrimonial auto-

cracies which continued to rule. Only political liberalization could empower an awakening civil society to discipline the state; accountability, transparency, and responsiveness necessitated democratization. The evaporating legitimacy of aging incumbents, and their shrinking capacity to sustain prebendal rule from declining state resources, reduced their capacity to resist political opening. The remarkable spectacle of the fall of the Berlin wall in 1989, and collapse of the Soviet Union in 1991, resonated powerfully. Leading Western donors now insisted on political reform as a condition for additional economic assistance.

The powerful interaction of internal and external pressures, and the contagious effects of the strongly interactive African regional political arena, proved irresistible. Democratization dominated the political scene in the 1990s, both on the ground and in the realm of African political study. Political opening in Africa formed part of a much larger 'third wave' of democracy, affecting Latin America, the former state socialist world, and parts of Asia as well. Initially, comparative political analysis focused upon the dynamics of transition itself, in a veritable moment of enthusiasm for the changes in course.

The initial impact was important; in at least a dozen states, long-incumbent rulers were driven from office by electoral means. However, in a larger number of other cases rulers developed the skills of managing competitive elections in a way to retain power. When attention turned to democratic consolidation in the later 1990s, the analytical mood was more somber. In the greater number of cases, only a partial political liberalization had occurred, captured in the analytical characterizations which emerged: 'illiberal democracy,' 'semi-democracy,' 'virtual democracy.' But now-dominant norms in the international system required a minimum of democratic presentability. Further, important changes had occurred in expanding political space for civil society, enlarging freedom of expression and media, and better observation of human rights. Democracy, however, was far from consolidated at the turn of the century (Bratton and van de Walle 1997, Joseph 1998).

Disconcerting new patterns appeared contemporaneous with the wave of democratization. A complete collapse of state authority occurred in Somalia and Liberia in 1991, and spread to some other countries. In a quarter of the states, significant zones of the country were in the hands of diverse militia, opening an era of 'warlord politics' (Reno 1998). In other countries such as Uganda, Ethiopia, and both Congos, insurgent bands from the periphery seized power. These events testified to a weakening of the fabric of governing, even a loss of statehood for some.

This enfeebled condition of numerous states, many if not most African scholars argued, emptied democratization of its meaning (e.g., Ake 1996). The externally imposed measures of structural adjustment had so compromised the effectiveness of states and their capacity to deliver valued services to the populace that the possibility of electoral competition had little value to the citizen. Democracy, in this view, was choiceless.

8. *African Political Study and Comparative Politics*

From the first application of modernization theories of largely extra-African derivation to African political study, a succession of conceptual perspectives drawn from comparative politics broadly defined have shaped political inquiry (dependency, neo-Marxism, rational choice, economic and political liberalism). In turn, African political study has made an important contribution to comparative politics. Instrumentalist and constructivist theories of ethnicity in their initial phases were strongly influenced by African studies (Young 1976); these modes of interpretation added important new dimensions to the comparative study of nationalism, which took on new life in the 1980s and 1990s (Rothschild 1980, Anderson 1983). The rebirth of the concept of civil society began in Africa and the former Soviet camp in the 1980s. Africa was a critical site of the late-century democratization experiments, which fed into the comparative study of democratic transitions (Diamond et al. 1995).

Africanists are prominent in the field of gender political studies. Analytical recognition of the political economy of the 'informal sector' or underground economy in good part originates in Africa-based studies. Understandings of the politics of patrimonialism rest heavily on African evidence. The collapse of some African states and the failure of others in the 1990s injected novel themes of state crisis into comparative politics (Zartman 1995). Sustained patterns of civil conflict and violence in some parts of Africa had counterparts in some regions of the former Soviet Union, suggesting the emergence of new kinds of political pathologies requiring analysis.

The singular trajectory of the African state generates multiple challenges to understanding, and divergent responses. The odyssey of independence began with high hopes and unrestrained optimism in the capacity of the state to manage rapid development and build an expanding political order from the center. Overdeveloped states and parastatalized economies ran into crisis, requiring far-reaching adjustments, economic retrenchment in the 1980s and political liberalization in the 1990s. An earlier apprehension of excessive state strength gave way to fears of state decline and weakness; analysts differed as to whether the prime cause was the inner logic of colonial autocracy embedded in the postcolonial polity (Mamdani 1996, Young 1994), or reflected a continuous pattern of underlying state weakness dating to precolonial times (Herbst 2000). The quest continues for forms of rule which could bring sustainable

development, accountable and effective governance, and also be authenticated and legitimated by a rooting in the African cultural heritage.

In sum, African politics as a field rests upon a dual dialectic. On the one hand, the rapid succession of distinctive political moments within Africa engages and defines the interpretive priorities of students of African politics. In turn, African political study remains firmly embedded in the larger field of comparative politics, whose evolving conceptual persuasions shape the orientations of its practitioners.

See also: African Legal Systems; African Studies: History; Central Africa: Sociocultural Aspects; Colonialism: Political Aspects; Dependency Theory; East Africa: Sociocultural Aspects; Nationalism, Historical Aspects of: Africa; Southern Africa: Sociocultural Aspects; West Africa: Sociocultural Aspects

Bibliography

Ake C 1996 *Democracy and Development in Africa*. Brookings Institution, Washington, DC

Anderson 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Apter D 1955 *The Gold Coast in Transition*. Princeton University Press, Princeton, NJ

Bates R 1981 *Markets and States in Rural Africa*. University of California Press, Berkeley, CA

Bayart J-F 1993 *The State in Africa: The Politics of the Belly*. Longman, London

Bratton M, van de Walle N 1997 *Democratic Transitions in Africa*. Cambridge University Press, Cambridge, UK

Callaghy T, Ravenhill J (eds.) 1993 *Hemmed In: Responses to Africa's Economic Decline*. Columbia University Press, New York

Coleman J 1958 *Nigeria: Background to Nationalism*. University of California Press, Berkeley, CA

Decalo S 1976 *Coups and Army Rule in Africa: Studies in Military Style*. Yale University Press, New Haven, CT

Diamond L, Linz J, Lipset S (eds.) 1995 *Politics in Developing Countries: Experiences with Democracy*. Lynne Rienner, Boulder, CO

Herbst J 2000 *States and Power in Africa: Comparative Lessons in Authority and Control*. Princeton University Press, Princeton, NJ

Hyden G 1980 *Beyond Ujamaa in Tanzania: Underdevelopment and the Uncaptured Peasantry*. University of California Press, Berkeley, CA

Joseph R (ed.) 1998 *State, Conflict and Democracy in Africa*. Lynne Rienner, Boulder, CO

Leys C 1974 *Underdevelopment in Kenya: The Political Economy of Neo-Colonialism*. University of California Press, Berkeley, CA

Mamdani M 1996 *Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism*. Princeton University Press, Princeton, NJ

Morgenthau R 1964 *Political Parties in French-speaking West Africa*. Clarendon Press, Oxford

Reno W 1998 *Warlord Politics and African States*. Lynne Rienner, Boulder, CO

Rothschild J 1980 *Ethnopolitics: A Conceptual Framework*. Columbia University Press, New York

Young C 1976 *The Politics of Cultural Pluralism*. University of Wisconsin Press, Madison, WI

Young C 1994 *The African Colonial State in Comparative Perspective*. Yale University Press, New Haven, CT

Zartman I (ed.) 1995 *Collapsed States: The Disintegration and Restoration of Legitimate Authority*. Lynne Rienner, Boulder, CO

Zolberg A 1966 *Creating Political Order: The Party-states of West Africa*. Rand McNally, Chicago

M. C. Young

African Studies: Religion

Most African languages lack an indigenous word for that sphere of belief and practice that is termed 'religion' in the West; their closest terms usually convey something more like 'usage' or 'custom.' So the study of African religion has tended to embrace a wide range of topics, extending to magic, witchcraft, divination, healing, cosmology and philosophy, as well as spilling over into virtually all other areas of social life and cultural endeavor. Yet still, defined as systems of belief and practice relating to the posited existence of spirits or personalized forces normally unseen by humans, African religions are not only analytically comparable in many respects to world religions such as Islam and Christianity, but have been compared in practice by the millions of Africans who over the past hundred years have converted to such religions. Scholars of African religion have thus been concerned, not just with 'traditional' religion, but with the far-reaching processes of religious change, stimulated above all by colonialism, that Africa has undergone since the late nineteenth century. African religion is thus a plural phenomenon and its study is multidisciplinary.

1. Missionary Origins

The earliest serious works on African religions were by missionary authors, many of whom also pioneered the study of African languages. Their agendas, of course, were far from disinterested: not just to explore the complexities of belief systems that many had dismissed as mere idolatry, but to find cultural leverage within them to promote the Gospel or to yield evidence for an original monotheism. Still, the best of them were serious ethnographies, grounded in long familiarity and a good command of the local language, such as Henry Callaway's study of the Zulu, Henri Junod's of the Thonga or Edwin W. Smith's of the Ila, all from Southern Africa. German scholarship was especially impressive, such as that of Diedrich Westermann on the Ewe of Togo or Bruno Gutmann on the Chagga in Tanganyika, among Protestants, or on the Catholic

side the work of several missionaries of the Society of the Divine Word (SVD), which was linked to Pater Schmidt and the journal *Anthropos* in Vienna. Sir James Frazer made use of missionary correspondents, such as John Roscoe, who dedicated his study of the Baganda to him.

After 1945, when the growing nationalist movement cast both missionaries and anthropologists under a cloud—the former for their disparagement of traditional beliefs as idolatrous, the latter as practicing a colonialist science of 'primitive' societies—the missionary tradition evolved into one of 'African theology.' A generic category of 'African traditional religion', homologous to the great scriptural religions, was first proposed by a former missionary, Geoffrey Parrinder, who was the first professor of religious studies at the University of Ibadan in Nigeria, and was taken up by a new generation of African scholars of religion, who were concerned to valorize traditional culture and to see Christianity fully 'inculturated' (to use the term that would come to be used in Catholic circles). Their theological contentions typically rested on 'ethnographic' accounts of traditional religions, whether these focused on contrasts between them and Christianity (as in the Kenyan J. S. Mbiti's comparison of eschatological concepts in the New Testament and among the Akamba) or on affinities (as in E. B. Idowu's *Olodumare: God in Yoruba Belief*). The irony of these works was that their nationalist appreciation of traditional religion often depended on their being able to write Judeo–Christian notions into it; for by now Christianity itself was fast becoming the largest religion of sub-Saharan Africa.

2. *From Administrators to Anthropologists*

The contrasting (but not wholly distinct) secular tradition of research on traditional religions had its roots in works by colonial administrators, some employed as 'government anthropologists,' such as R. S. Rattray's *Religion and Art in Ashanti* (1927) or B. Maupoil's *La Géomancie à l'ancienne Côte des Esclaves* (1944). Free from evangelistic concerns (and therefore less constrained by the category of 'religion'), they were more able to explore themes such as the mundane, practical dimensions of magical charms, or of techniques of divination, and of the cognitive or cosmological principles underlying them. The greatest figure was E. E. Evans-Pritchard, who produced two classic studies of African belief and practice, quite different from one another. The first, *Magic, Witchcraft and Oracles among the Azande* (1937) was in the mold of his teacher Bronislav Malinowski (who, though not himself an Africanist, played an important role as research director of the International African Institute). Its chief aim was to show how seemingly irrational beliefs were in their context both reasonable and effective; and its distinction between witchcraft and sorcery (made by the Azande themselves) was highly influential in subsequent studies of African witchcraft.

Most British anthropologists over the next two decades, drawing their theoretical inspiration from A. R. Radcliffe-Brown, made social structure, rather than culture, their cardinal concept. Religion (or, as preferred, 'ritual') was seen as an aspect of political and social organization, the expression of social values (like the cult of ancestors in lineage-based societies, or sacred kinship in some centralized polities, or rites of passage everywhere).

A focus on religion as culture did, however, continue elsewhere, as in the work of Melville J. Herskovits, the *doyen* of American Africanists, on the religion of Dahomey (1938); and the many essays of the French school around Marcel Griaule on the religion and cosmology of the Dogon and Bambara peoples of Mali from the late 1930s into the 1960s. But the opposition between sociological and cultural approaches was bridged in Daryll Forde's collection of essays, *African Worlds* (1954), and altogether abandoned in the fine run of studies that appeared in the ensuing decade: Evans-Pritchard's second great book, *Nuer Religion* (1956); monographs by his pupils John Middleton on the Lugbara (1960) and Godfrey Lienhardt on the Dinka (1961); Meyer Fortes's *Oedipus and Job in West African Religion* (1959), and other essays on ideas of morality and personhood among the Tallensi; and Victor Turner's studies of Ndembu rituals and symbolism (1960s and 1970s). With this body of work, the study of 'traditional' religions may be said to have reached its zenith. Essentially structural–functionalist in approach, these monographs treated systems of belief and ritual practice as distinctive wholes, expressed though the categories of a particular culture and adapted to the social and ecological setting in which they existed. Only marginally—and that mostly in the analysis of topics like witchcraft and spirit possession—did these studies address those issues of social change that were then starting to become insistent in Africa itself.

3. *Historical Perspectives*

A historical perspective—in the sense both of attention to the past, and of the analysis of change in the present—became widespread in the 1960s, with a large measure of convergence between anthropology and history. Missions, as major agents of more than just religious change, now began to attract study, first by historians for their contribution to the establishment of colonial society, as with Roland Oliver's *Missionary Factor in East Africa* (1952), or to the emergence of the educated African elite and hence, ultimately, of nationalism itself, as with the work of J. F. Ade Ajayi and E. A. Ayandele in Nigeria. T. O. Ranger probably did most to establish the historical study of African

religion, with an emphasis on the relations between religion and politics—as with the role of spirit mediums in the 1898 Rhodesia uprising—and on the specific ways on which mission Christianity became localized in East and Central Africa.

Prophetist or syncretic religious movements and independent churches attracted much attention. The seminal work was written by a Swedish missionary, Bengt Sundkler: *Bantu Prophets in South Africa* (1949/1961), which showed how independent Christian churches provided a means for the black population of South Africa to sustain alternative values to the regime of racial oppression then being imposed on them. Some Marxist-inclined analysts of nationalism in the 1950s, such as Thomas Hodgkin or Georges Balandier, saw its early precursors in religious leaders going back to just before World War I—figures such as Prophet Harris in the Ivory Coast, John Chilembwe in Nyasaland, Simon Kimbangu and his successors in the Congo. But their accounts were too reductionist, and too prone to assume that such religious movements would necessarily yield, with development, to more secular forms of politics. Most studies of contemporary movements and churches in the 1960s and 1970s—such as those by H. W. Turner and J. D. Y. Peel of the Aladura ('praying') churches among the Yoruba, Wyatt MacGaffey on Kongo prophets, M. Daneel or B. Jules Rosette on Shona independent churches, and (the richest of all in its analysis of ritual and symbolism) James W. Fernandez's *Bwiti: An Ethnography of the Religious Imagination in Africa* (1982)—laid more emphasis on the extension of African ideas of spiritual power and social renewal into Christianity.

As the tally of monographs grew, so did the demand for a theoretical synthesis. D. B. Barrett attempted in his 1968 work *Schism and Renewal in Africa* an overarching explanation for the rise of independent Christian movements, but it did not rise above the empirical identification of some rather obvious predisposing conditions, such as the depth of Christian (especially Protestant) penetration or the intensity of colonial pressure. De Craemer, Vansina and Fox (1976) proposed for Bantu Central Africa that recent religious episodes, though largely Christian in idiom, belonged to a long-established pattern, whereby phases of social malaise led communities to look for social renewal through eradicating evil (typically in the form of witches) and adopting new ritual means to ensure security and wellbeing—until things again ran down, and the cycle was repeated. The anthropologist Robin Horton, who had earlier formulated a much-debated 'intellectualist' interpretation of African cosmologies (Horton 1967), drew on it to propose the most influential general theory in a series of articles in the journal *Africa* between 1971 and 1975. This explained African conversion to the monotheist faiths as a cognitive adaptation to a basic change in social experience, from living in a 'microcosm' of confined, small-scale settings, to one in a 'macrocosm' of mobile, large-scale relations. The symbolic correlate of this was the declining relevance of local spirits and ancestors, and a new interest in the Supreme Being—relatively otiose in traditional belief, but given central position in Islam and Christianity. Horton's theory was much applied and critiqued in subsequent studies of religious change all over Africa. It had particular value in that, by placing Islam and Christianity in the same frame, it bridged the gap that had tended to develop between the study of the two world religions. Its spatial emphasis was echoed in work on regional cults and oracles, particularly by Richard Werbner on western Zimbabwe; and was combined with a Marxist perspective by Wim van Binsbergen in a bold attempt to link levels of religious development in Zambia with successive modes of production. Horton's theory was criticized on various grounds: for ignoring other kinds of religious change than the growth of monotheism; for neglecting the role of power in conversion; and (a criticism made especially by Humphrey Fisher, a historian of Islam) for laying so much emphasis on the interplay of the indigenous religious framework and Africans' experience of social change that the cultural dynamics of the world religions themselves were underplayed.

4. *Islam*

With the exception of Ethiopia, Islam's presence in sub-Saharan Africa long predated Christianity, and there is a historical and textual depth to its study not shared by the other two sectors of the African religious field. Yet the modern study of African Islam by outsiders goes back to similar missionary and administrator origins, as with traditional religion.

The French, having added a large area of Sudanic West Africa, where Islam was the hegemonic religion, to their earlier occupation of the Maghreb, were particularly concerned to gauge the political import of '*Islam noir*.' Among a notable series of scholar–administrators, the most prolific was Paul Marty, who produced no fewer than 12 volumes on Islam in different territories of French West Africa between 1913 and 1926. In English the most comparable *oeuvre* was that of an ex-missionary, J. S. Trimingham, who between 1949 and 1964 produced a series of works surveying Islam in different regions—East and West Africa, Sudan and Ethiopia—with a cultural rather than a political focus.

With the growth of a more systematic research tradition after 1950, in African universities 'Islamic Studies' was often placed with Arabic in a separate academic department from 'Religious Studies', where scholars of Islam and Muslim scholars might work together. The vital long-term project of locating and cataloging the Arabic documentation on which Islam's historical, as well as theological and legal, study

depended was begun. For centuries, the growth of Islam in Africa had been linked closely with long-distance trade and with state formation. In East Africa, the balance between external Islamic influences and internal Bantu ones in the shaping of the Afro–Islamic culture of the mercantile city-states of the Swahili coast excited debate between Islamicists, historians, archeologists and anthropologists. In West Africa a contrast was drawn between a militant Islam associated with an alliance of Fulani pastoralists and holy men, which produced, in the eighteenth and nineteenth centuries, a sequence of jihadist states, and a more accommodative Islam promoted by Dyula traders. Classic works like Murray Last's *The Sokoto Caliphate* (1967) and Yves Person's *Samori: une revolution dyula* (1969–71), still shed much light on politics in their respective modern countries, northern Nigeria and Guinea. Such institutions as Sufism, clerical lineages, and especially the religious brotherhoods that have been such a prominent feature of African Islam also received attention. Two notable studies by political scientists have examined the political and economic roles of brotherhoods in modern times: D. B. Cruise O'Brien's *The Mourides of Senegal* (1971) and John Paden's *Religion and Political Culture in Kano* (1973).

Anthropologists who worked in Muslim areas inevitably had much to say about Islam, though the initial impetus had sometimes been to marginalize it, as S. F. Nadel did in his *Nupe Religion* (1954). While historians and Islamicists tended to emphasize the long-term advance of orthodox Islam through reformist movements, the inclination of anthropology, privileging the local over the global, was to explore the substrate of indigenous practices which lie 'under' or alongside the official face of Islam. There was always *adat* ('custom') in contrast to *sharia* (Islamic law), and also a variety of less orthodox ritual practices, such as divination, charms, *sadaqa* ('alms') as sacrifice, and belief in djinns. The relation between Islam and spirit possession cults has been the subject of some fine ethnographies, such as Janice Boddy's *Wombs and Alien Spirits* (1989), on religion and gender in the northern Sudan. The label 'popular Islam,' though sometimes applied to such phenomena, misleads both because they deeply involve the religious elite and because they reach back into the past of mainline Islam. But anthropology did succeed in bringing the study of Islam and Christianity closer together, drawing parallels (for example) between patterns of conversion and the content of mundane religious practice in the two religions (Lewis 1980).

5. *African Religion at the Turn of the Millennium*

Since the mid-1980s, the study of African religion has both become more of a unified field in itself and become less compartmentalized from the rest of African studies. Two works which show this conspicuously are David Lan's *Guns and Rain* (1985), on the role of spirit mediums in the guerrilla war which won Zimbabwe's independence, and Stephen Ellis's *The Mask of Anarchy* (1999), which incorporates the 'mystical' factor into an account of the civil war in Liberia. More generally, religion—especially Islam and Christianity—came to play a larger role in public life. As the capacity of African states declined, the major churches (and the Catholic Church above all) stood out more as the most effective institutions of civil society, and by the early 1990s were playing a significant role in movements of democratization (though with fitful success), in development initiatives, and, in South Africa, in the process of post-apartheid reconciliation. At the same time, the churches had been not been able to stop the atrocities in Rwanda and religion emerged more strongly as a source of political conflict in countries such as Nigeria and the Sudan. In his *African Christianity* (1998) Paul Gifford compares Ghana, Cameroon, Uganda, and Zambia to give a nuanced picture of the public role of the churches, concluding that they reflect, at least as much as they transcend, the political values of the wider society.

New militant movements came to the fore in both Christianity and Islam, mirroring one another in their strongly global orientations. The Islamic movement might be seen as another surge of the long-term reformist trajectory, in that it is oriented to the normative standards of the Middle East, generally hostile to Sufism and the influence of the brotherhoods, and concerned to promote Islamic education, *sharia* law, and a more universal Muslim identity. However, there is still much variation from one country to another (Brenner 1993, Westerlund and Rosander 1997). In contrast, the rise of neo-Pentecostal, charismatic or 'born-again' Christianity, represents (at least on the surface) a reversal of the earlier trajectory of 'Africanization' or local inculturation, in that what attracted its youthful adherents was precisely its transnational quality, its use of electronic media, and its evocation of American modernity.

Some of the best work of the 1990s on Pentecostalism—David Maxwell on eastern Zimbabwe or Birgit Meyer on the Ghanaian Ewe, for example—shows how necessary it is to relate modern developments to earlier mission activity in its particular localized forms. The prime example of such a two-way integration of anthropology and history appears in the *oeuvre* of Jean and John Comaroff, whose massive work *Of Revelation and Revolution* (vols. 1–2, 1991, 1997) explores how Protestant missions contributed to 'the colonization of consciousness' of the southern Tswana through their many-sided impact on daily life. In contrast, Paul Landau's *The Realm of the Word* (1995), on the northern Tswana, Donald Donham's *Marxist Modern* (1999), on the contribution of missions to revolution in southern Ethiopia, and J. D. Y. Peel's

Religious Encounter and the Making of the Yoruba (2000) all give more attention to the import of the religious content of mission for new local and national identities. The forms of religion in Africa transmute with great rapidity, yet its centrality to social relations shows little sign of secular attenuation. While Christianity and Islam are now formally predominant, with 'African traditional religion' largely a thing of the past, the focus of religious concern still shows much continuity with the 'pagan': empowerment, guidance and deliverance from mundane evil are what Africans continue to ask of their gods. Moreover, many recent studies from all over Africa have drawn attention to *The Modernity of Witchcraft*, as Peter Geschiere put it in the title of his 1995 book, mainly about Cameroon. Despite its seemingly radical emphasis on renewal and its global connections, Pentecostalism maintains and even extends older discourses of witchcraft and the demonic. This social reality underscores the need for the closest interdependence between the present and the past in African religion. For just as studies of contemporary religion or those concerned with the 'advance' of the world religions must acknowledge the durability of values and ontologies grounded in the indigenous religions of Africa, so also must historical studies of religion be oriented towards those dynamics of change which have eventuated in the present complex religious disposition. A principal theoretical outcome of the study of African religion over the past century is that one of its main conceptual instruments—the distinction between the 'traditional' and the 'modern'—finally needs to be abandoned.

See also: African Studies: Culture; African Studies: History; African Studies: Politics; Christianity: Evangelical, Revivalist, and Pentecostal; Colonialism, Anthropology of; Colonization and Colonialism, History of; Evans-Pritchard, Sir Edward E (1902–73); Islam: Sub-Saharan Africa; Malinowski, Bronislaw (1884–1942); Nationalism, Historical Aspects of: Africa; Prophetism

Bibliography

Blakely T D, van Beck W E A, Thomson D L (eds.) 1994 *Religion in Africa: Experience and Expression*. Heinemann, Portsmouth, NH

Brenner L (ed.) 1993 *Muslim Identity and Social Change in Sub-Saharan Africa*. Hurst, London

De Craemer W, Vansina J, Fox R C 1976 Religious movements in Central Africa. *Comparative Studies in Society and History* **18**: 458–75

Fashole-Luke E, Gray R, Hastings A, Tasie G (eds.) 1978 *Christianity in Independent Africa*. Rex Collings, London

Forde D (ed.) 1954 *African Worlds: Studies in the Cosmological Ideas and Social Values of African Peoples*. Oxford University Press, London

Hastings A 1994 *The Church in Africa 1450–1950*. Clarendon Press, Oxford

Horton R 1967 African traditional thought and Western science (Part I and II). *Africa* **37**: 50–71, 155–87

King N Q, 1986 *African Cosmos: An Introduction to Religion in Africa*. Wadsworth, Belmont, CA

Lewis I M (ed.) 1980 *Islam in Tropical Africa*, 2nd edn. Hutchinson, London

Ray B C 1976 *African Religions: Symbol, Ritual and Community*. Prentice Hall, Englewood Cliffs, NJ

Westerlund D, Rosander E E (eds.) 1997 *African Islam and Islam in Africa: Encounters between Sufis and Islamists*. Hurst, London

J. D. Y. Peel

Age: Anthropological Aspects

Age is a product of the process of aging, which is partly determined by the social environment. Recognition of this social component has led to some initial attempts to identify features of the life course that are unique to Western civilization. However, the ethnographic literature reveals a considerable blurring of some of these stereotypes. Thus Philippe Ariès's (1962) influential argument that 'childhood' as opposed to 'adulthood' is essentially a product of the industrial revolution may be challenged with reference to the widespread practice of initiation in other cultures, which frequently marks a transitional point in the life course, distinguishing distinct stages, and is associated with a range of beliefs associated with childhood and development (La Fontaine 1985). Again, G. Stanley Hall (1904) is credited with 'discovering' adolescence as a further category that arose out of the industrial cities in America, leading to a developing interest in this topic; yet deviant subcultures associated with dispossessed youth have been noted in some traditional rural settings and extend even to studies of primate behavior (Spencer 1965, Pereira and Fairbanks 1993). Again, Leo Simmons' (1945) early survey of the role of the aged in 'primitive' societies is often cited, suggesting that this category was highly respected as compared with the West, where the family has diminishing importance in the process of urbanization (Cowgill and Holmes 1972); yet the data on this topic reveals a varied response to the problem of reconciling respect for the age and experience of older people with the frustrations of their overbearing power within the family on the one hand, or the liability of caring for them on the other, especially in impoverished situations where the family often does not survive beyond two generations.

The complexity of the problem has led to the development of more refined concepts among sociologists, focusing on particular stages of the life course as prime topics for investigation by specialized subdisciplines. In contrast to this, anthropological studies

aim to be holistic, and within a culture, any role or status associated with age has to be viewed in the context of the life course as a whole.

1. The Family and Ambivalence Towards Aging

Thus, the marginalization of certain age categories—perhaps adolescence or old age—forms part of a wider pattern involving a complex of relations between young and old. Adolescent subcultures, for instance, may be viewed as an alternative to the authoritarian structure of the family, and a milieu where the more open-ended bonds between age peers sharpen their awareness of an alternative experience involving a more creative lifestyle, preparing them for future possibilities within the wider community. Again, some studies have indicated that stress and hardship during adulthood, perhaps a midlife crisis, appears to reinforce people's ability to cope with the social discomforts of old age in due course. Conversely, a more cushioned life career appears to leave people more vulnerable to the sense of loss and isolation as they grow old.

Very broadly, ambivalent regard for old age is especially associated with the elaboration of the family as a corporate unit in preliterate agricultural societies. The evolution of the family, reaching a peak in such societies, is entwined with the evolution of age relations and the concept of 'status' coined by Henry Maine (1861), where social position is ascribed by being a member of a family. This is closely associated with power in the hands of the most senior members by age as the legitimate custodians of family tradition. In this milieu, cultural knowledge itself may be treated as a form of property, to be imparted or withheld. In a very pertinent sense, property relations within the family are age relations, creating bonds and tensions within the family (Foner 1984). A common factor underlying the urban stereotypes of old age and adolescence is the demise of the family as a dominant institution (Maine's 'status to contract').

The premise of respect and even fear for older people is especially widespread in rural Africa, where it is often associated with strong patrilineal families and polygyny. This highlights the notion that children are the property of the family in 'status-dominated' societies, giving the senior generation the power to marry off their daughters early and to delay the marriage of sons for perhaps a decade or more. As long as this regime can be maintained, extended bachelorhood facilitates widespread polygyny among older men. The array of life-stages is illustrated in Fig. 1 with reference to the Samburu of Kenya, who provide a clear-cut example of a more general phenomenon, with no dispensation for younger men to marry early or for widows to remarry. The concave shape of the age distribution is characteristic of preliterate societies, where mortality rates are especially high among the young.

The figure also indicates the contrasting life trajectories of women who are married young to much older men. The depressed status of women in such societies has to be viewed in relation to their total life-course.

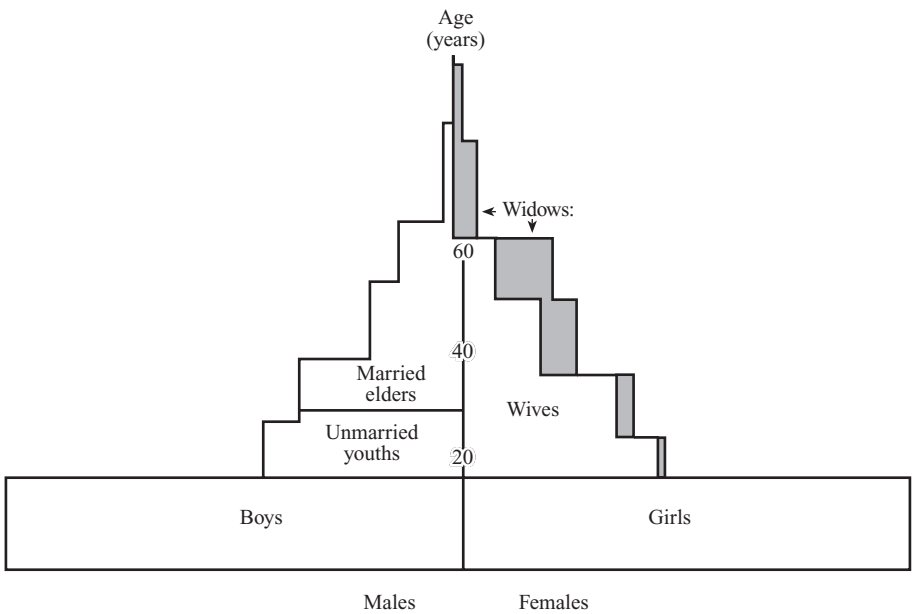

Figure 1
Age, status, and the demographic profile of a polygynous society (Samburu 1960)

Whereas a boy tends to take the first step from the obscurity of childhood towards a promising adulthood with his initiation, the marriage of a girl transforms her from the obscurity of childhood to an initially obscure role as a young wife with very restricted opportunities and a stranger in her new home. A view of women as the victims of male exploitation is particularly apt at this low point in their careers, and especially in societies where there is a sharp separation between male and female domains. An alternative viewpoint portrays women as agents who can manipulate their depressed situation to their own advantage. This becomes increasingly apt as her life course develops and notably once she surmounts the restrictions of her reproductive years and is increasingly independent of her aging husband. In the prime of middle age, the advantage still rests with men. However, this tends to be reversed beyond this point as they lose the will to assert themselves against younger, more competitive successors and are edged to the margins of activity in community affairs. Women are less hampered by aging until they are too frail to play an active role. The process of growing old for women in these circumstances is to free themselves from the domestic routine, but not from their personal networks which focus on their growing family with a freedom to choose how they wish to involve themselves (Amoss and Harrell 1981).

2. *The Experience of Maturation and Aging*

The social experience of maturation and aging shapes the perception of time and hence its meaning in a quite fundamental sense. This may be analyzed in several ways. The first involves an autobiographical approach, viewing each major event of the life course as a uniquely personal experience that may form a pattern retrospectively, but can only be anticipated within limits. The personal experience of aging coincides with the shared experience of historical change, and any analysis of one of these has to disentangle it from the other (Mannheim 1952). Thus, when older people suggest that policemen are getting younger, this may be a sign that they are getting older; but when they suggest that bank managers are getting younger, this may accurately reflect a historical trend. Correspondingly, in anthropological studies, the extent to which younger people are seen to subvert tradition may be an aspect of a continuous process of adaptation to new opportunities; and 'tradition' itself may be adapted by the new generation as they mature and take over as its custodians. Or youthful subversion may be in part a response to their subordination—an adolescent rebellion that is mounted by each successive generation. In the absence of written records, only longitudinal studies, monitoring the process of change over a period of decades, can distinguish between irreversible historical trends and the tenacity of family structures perpetuated by the recycling of vested interests and intergenerational strains. To the extent that a popular awareness of social change draws attention to recent innovations and gives an impression of the contemporary scene as a watershed between tradition and modernization, the persistence of age relations embedded in resilient family structures is less evident in short-term studies. Underpinning this resilience is the cumulative nature of privilege associated with age in 'status-dominated' societies. Those who react against traditional restrictions in their youth become the new custodians as they age.

A more interactive approach to the experience of aging, especially in tight-knit communities, focuses on the accommodation of major life transitions as a discontinuous process; a life-crisis theory of aging. The significance of rites of passage is that they involve the wider community, even beyond the family, in a shared experience of irreversible change. In the space between these events or other critical episodes, people age physically, but the configuration of social relations remains unchanged and there is a sense of timelessness as trends leading to the next critical change unfold. When this occurs or is precipitated, the configuration of roles adjusts to a new status quo and there is a distinct step in time. In dialectical terms, there is a mounting contradiction between inflexible social relations and the unstoppable process of maturation and decay, undermining the array of power relations. From this point of view, the anxieties that accompany these transitions and shifting roles are also anxieties of aging. Critical life events and transitions are key points, both in the experience of aging for the individual and with regard to adaptation and regeneration within the community.

3. *Age Systems and Hidden Knowledge*

A third approach to the experience of aging has coined the analogy of a 'cultural clock' that prescribes the appropriate ages for major life transitions within any society, facilitating adjustment, and heightening the awareness of those who forestall or lag the norm (Neugarten 1968). While this expresses an essentially conformist approach towards aging, it is a particularly appropriate model for societies with age systems. And because members of such societies are very aware of age, they provide an ideal type for examining a range of issues associated with the process of aging. In an age system, those of a similar age are grouped together as an age set (or age group), maturing and aging together. Their position at any stage may be termed as age grade, consisting of an array of expectations and privileges. In effect, the age set passes up an age ladder, rung by rung, through successive age grades in a defined progression, like children passing through school but over a more extensive period in a process

that persists into old age. The left-hand side of Fig. 1 illustrates a society with such a system, indicating the extent to which the transitions from boyhood at initiation and to elderhood with marriage are closely associated with age. Each step in the demographic profile broadly represents an age set, and the successive statuses for married elders could be elaborated to provide a more detailed set of aged grades. In general, age sets tend to involve males only, but they may also define critical aspects of men's relations with or through women, giving women a distinctive role within the age system. Thus, the position of a woman may be highlighted in relation to the age set of her husband or father or sons, but women are only rarely grouped together by age, except sometimes during the brief period leading up to their marriages.

Age systems institutionalize the cultural premise of respect for age through a form of stratification that in effect inverts stratification by caste. In contrast with the total immobility of a caste system, where status is determined by birth and persists throughout life, age systems guarantee total mobility: a young man is initiated onto the lowest rung of the age ladder as a member of the most junior age set, and is systematically promoted with his age set from one age grade to the next towards the top. The nuances of each age system relate to the process of promotion, involving certain pressures from below and resistance from above. It is these pressures, arising from the interplay between a concern for status and the physical process of aging, that provide the mainspring for the 'cultural clock,' which is perceived frequently as a recurring cycle of promotions and delays, spanning the interval between successive age sets.

Elaborate age systems are associated primarily with the pastoralist peoples of East Africa. The relevance of nomadic pastoralism appears to stem from the equality of opportunity in a setting where a mixture of acumen, commitment, and sheer luck determine the success of each stock owner. Unequal fortunes tend to even out as the more successful convert their surplus into further wives as an investment for the herding enterprise, leading to larger families and the dispersal of this wealth among the sons of the next generation. Correspondingly, the ideals underpinning age-based systems are of equality among age peers associated with the mobility of wealth rather than the inheritance of privilege based on birth and accumulated capital. This is complemented by the premise of inequality up the age ladder, regardless of family or wealth, again endowing ultimate moral authority and ritual initiative on the senior generation. Age systems were more widespread historically, but have become outmoded by the gathering complexity and inequalities in the process of urbanization.

The position of older men as the repositors of traditions has an affinity with secret societies in West Africa and Papua New Guinea, where esoteric knowledge is acquired by stages and is only fully understood in later life by those that are eventually initiated into the higher levels of the organization. It is the hiding of this secret rather than the elusive knowledge as such that displays power, impeding the progress of individuals up the career ladder towards the privileges at the top. Among East African societies with age systems, there is a similar mystique and hierarchy of power, where careers are controlled by the ritual authority ascribed to those higher up. However, it is as age sets rather than as individuals that promotion takes place; and to the extent that the pace of these is controlled by older men, manipulating the 'cultural clock' to their own advantage, they are playing for time against the inevitability of their own aging. Older women, middle-aged and even relatively young men may conspire in this, resisting the advancement of their juniors by age. For those that live long enough, the frailty of their aging does not diminish the awe for their great age.

The antithesis of the power of older men is the physical virility of youth, often associated with an alternative lifestyle in 'warrior societies.' This poses a contradiction between the moral advantage that lies with older men and the more immediate interests of younger men, who may react against traditional restrictions. Historically, situations of political turbulence offered opportunities that favored youth, overriding the constraints of the age system. However, it is also characteristic of the periodic age cycle that certain phases may be associated with greater or lesser gerontocratic control over a new age set. These are paradoxically an aspect of the system where privilege is vested heavily in the hands of older men, notably over women and marriage, and this creates a certain power vacuum on the lower rungs of the age ladder, encouraging youthful rebellion. By viewing age systems as interactive enterprises concerned with the distribution of power with age, rather than in terms of 'gerontocracies' or 'warrior societies' as such, their apparent resilience to change relates to involvement in the system at all ages. Those who can claim certain privileges of youth also have a stake in their future as elders.

A holistic approach towards age systems leads one to examine the relationship between principles of age organization and aspects of the family. As in other polygynous societies, competition for wives can give rise to rivalry between brothers, and the delayed marriage of younger men to tension between generations. Age systems can serve to diffuse these strains by imposing the restrictions on younger men from beyond the family (Samburu), by creating an alternative and prized niche for younger men (Maasai, Nyakyusa), or by maintaining a disciplined queue towards marriage (Jie, Karimojong). A variation of this general pattern occurs among the Cushitic-speaking peoples in Ethiopia and Kenya, where the age system underpins the privileges of first-born sons within the family.

A clear link between age systems and family structures is illustrated by the extent to which recruitment to an age set is often complicated by restrictions of generation: the position of the son within the system is determined in part by that of his father, giving rise to a hybrid 'age/generation' system, rather than one based solely on age (Stewart 1977). The rules can be highly elaborate, leading to speculation that they either are spurious or have been misunderstood. However, their implications for the distribution of power and authority with age are very specific, and understanding in each instance derives from a wider analysis of relations between old and young within the family and the wider community (Baxter and Almagor 1978). Younger men have the advantage of physical virility and the rapid accumulation of practical experience. Age systems, and indeed any institutions that endow older people with power, may be viewed in terms of their ability to impose a moral superstructure with a higher authority. The claim of older people to be the true custodians of tradition, and perhaps to have the closest rapport with the ancestors, places society above the brutish forces of nature and inverts the natural process of aging in a hidden display of power.

See also: Age Policy; Age, Race, and Gender in Organizations; Age, Sociology of; Age Stratification; Age Structure; Generation in Anthropology; Generations in History; Generations, Relations Between; Generations, Sociology of; Kinship in Anthropology; Life Course in History; Life Course: Sociological Aspects; Lifelong Learning and its Support with New Media: Cultural Concerns; Lifespan Development: Evolutionary Perspectives; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development; Plasticity in Human Behavior across the Lifespan; Youth Culture, Anthropology of; Youth Culture, Sociology of

Bibliography

Amoss P T, Harrell S (eds.) 1981 *Other Ways of Growing Old: Anthropological Perspectives*. Stanford University Press, Stanford, CA

Baxter P T W, Almagor U (eds.) 1978 *Age, Generation and Time: Some Features of East African Age Organizations*. Hurst, Century–Crofts, London

Cowgill D O, Holmes L D (eds.) 1972 *Aging and Modernization*. Appleton–Century–Crofts, New York

Foner N 1984 *Ages in Conflict: A Cross-cultural Perspective on Inequality Between Old and Young*. Columbia University Press, New York

Hall G S 1904 *Adolescence: Its Psychology and its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion and Education*. Appleton, New York

Kertzer D I, Keith J (eds.) 1984 *Age and Anthropological Theory*. Cornell University Press, Ithaca, NY

La Fontaine J S 1985 *Initiation*. Penguin, UK

Maine H J S 1861 *Ancient Law: Its Connection with the Early History of Society and its Relation to Modern Ideas*. Murray, London

Mannheim K 1952 The problem of generations. In: Mannheim K. (ed.) *Essays on the Sociology of Knowledge*. Routledge and Kegan Paul, London

Neugarten B I 1968 Adult personality: Toward a psychology of the life cycle. In: Neugarten B I (ed.) *Middle Age and Aging*. University of Chicago Press, Chicago

Pereira M E, Fairbanks L A (eds.) 1993 *Juvenile Primates: Life History, Development, and Behavior*. Oxford University Press, New York

Simmons L W 1945 *The Role of the Aged in the Primitive Society*. Yale University Press, New Haven, CT

Spencer P 1965 *The Samburu: A Study of Gerontocracy in a Nomadic Tribe*. University of California Press, Berkeley, CA

Spencer P (ed.) 1990 *Anthropology and the Riddle of the Sphinx: Paradoxes of Change in the Life Course Transformation*. Routledge, London

Stewart F H 1977 *Fundamentals of Age-Group Systems*. Academic Press, New York

P. Spencer

Age Policy

As government interventions in the society and economy have diversified and expanded, what can be called 'age policy' has developed for organizing and regulating phases in the life course. Age policy centers around the state for two reasons. First of all, it has come out of state interventions: constructing the welfare state cannot be separated from the task of forming and categorizing phases of life on the basis of age-based norms. Secondly, age has been a major policy tool for public authorities. Dividing the population into age groups has been the easiest way to distribute individuals to socially assigned activities (see *Age Stratification*). These two dimensions of age policy will be discussed, and a few concrete examples of the implementation of policies regarding youth and old age will be examined.

1. Age Policy, the Product of Government Interventions

Along with the building of the modern state has emerged a social and legal construction of the individual. At stake in this process is the creation of conditions 'that single out the individual' so as to enable each person to stand apart from family and community bonds. The individual has thus been taken 'as the prime holder of rights and duties and as the prime target of bureaucratic and administrative acts' (Mayer and Schoepflin 1989, p. 193). This construction of the individual has been grounded in a set of age-

based norms that have marked the chronological continuum of life with significant thresholds and organized it into successive phases.

Increasingly strict laws have been adopted about the age for working (specifically for regulating child labor and, more recently, for setting the retirement age) and the age of compulsory schooling. These laws have laid the very foundations for social constructing and institutionalizing the life course (see *Life Course: Sociological Aspects*). The life span has thus been divided into three distinct phases. Owing to its increased interventions in the economy and society, the state has regulated the ages of life. By 'policing ages' (Percheron 1991), it has become the major actor in constructing the life course. In particular, it has distributed social duties and activities by organizing the triangular relations between family, work, and school (Smelser and Halpern 1978) into an orderly model of successive phases. Each phase has thus been identified with an activity that, setting it apart from the other phases, endows it with meaning and identity. Childhood is the time for education and of dependence on the family; adulthood is defined by work; and old age is a period of rest after a life of work. This threefold organization of the life course has become an institution as the welfare state has expanded and as age norms have been enacted in law.

The invention and generalization of old-age pensions—one example of age policy—has played a decisive role in constructing and consolidating this 'tripartition' of the life course (Kohli 1987). First of all, retirement systems have been a major factor in determining the order and hierarchy between the three principal phases of the life course—with, at the center, work as the social contents of adulthood. This phase lies in between youth (devoted to education for a life of work) and old age (associated with inactivity). These systems have helped stake out a life course where the individual's contribution during adulthood to the world of work conditions the right to rest, placed at the end of life. Secondly, retirement systems, along with other social policies (such as education), have given more weight to chronological criteria for marking the transitions from one phase to another. Old-age pensions have thus chronologized the life course, marked as it is by the legal ages for starting and leaving school (the latter separating childhood from adolescence) and for going on to retirement with a full pension (an event marking the threshold of old age).

This division of the life span into three chronological phases has produced a standardized life course. At the same age, everyone moves quite predictably from one phase to the next. At an equivalent level of education, entering the world of work occurs at the same age for nearly everyone. And the retirement age sets the date when everyone will stop working. Long-term trends in the ages of exit from the labor force provide evidence of this standardization of behaviors. As retirement systems have expanded to cover more and more of the population, the moment when individuals stop working has gradually approached the age of entitlement to a full old-age pension.

Old-age pensions have fostered new expectations about the future. The individual no longer has the same prospects as in preindustrial societies, where the family and private wealth determined the timing of phases in life and where individuals had no future as such: they died young. The development of old-age pensions, along with a much longer life expectancy, has individualized the life course even as it has 'chronologized' it. Thanks to pensions, the individual is endowed with a future. As a consequence, retirement has furthered the change from a society where the person was ascribed a status through membership in a family or local group to a society of achieved statuses and, thus, of mobility. In this new society, the individual has prospects. Security is now based on the person's work and no longer on belongings, or a local or family status. To insure this security, retirement involves successive generations in forms of reciprocity and statistical, long-term solidarity. It thus contributed to erecting and developing a new social order in line with the requirements of a society undergoing industrialization. This example illustrates the state's regulatory interventions that have instituted age-based norms for 'policing ages.'

This social construction of the life course by the state, in particular through welfare policies, is an ongoing process. During the 1980s and 1990s, all sorts of early exit schemes were worked out to enable aging workers to withdraw from the labor force before the normal retirement age. This can be interpreted as a factor in deinstitutionalizing the threefold organization of the life course (Guillemard and Van Gunsteren 1991), since these schemes undermine the regulated transition from work to retirement. Early exit schemes wreak havoc in the orderly succession of the three phases in the life course. Appearances suggest that early exit is a mere event on the retirement calendar entailing no other noteworthy change. But looking beyond appearances, the impact of these new age-based measures advises interpreting early exit and preretirement schemes in terms of increased flexibility in reorganizing the end of the life course (Guillemard 1997).

The most frequently used early exit schemes have been, not old-age pension funds, but rather disability and unemployment insurance funds. These schemes have proven extremely malleable. In all countries, they have continuously evolved as a function of the employment situation. This can be interpreted as a detemporalization of the life course: the individual can no longer imagine a continuous, foreseeable life. The order of phases and activities is no longer precise, and is even contingent. The timing of definitive exit from the labor market is unforeseeable. No one working in the private sector knows when (at what age) or how (under what conditions) they will defin-

itively stop working. As a result, the end of the life course is becoming destandardized as well. Since chronological thresholds are no longer clearly set, the work inability (real or alleged) of older wage-earners is becoming a criterion more important than chronological age. Given the reforms now under way or under study, retirement tends to be timed later in life. In any case, definitive exit from the labor market has been fully separated from admission into retirement. There is now a long transition as persons move out of work and toward retirement.

The hierarchical, orderly succession of phases in the life course is coming undone. Sociologists who study youth have described this as a 'tourniquet' on young people's lives. A period of unemployment often follows education, as young people enroll in government-sponsored training programs or take up odd jobs that, instead of leading to integration in the world of work, often end in further 'mixes' of training programs with unemployment compensation. Entrance into the labor market is uncertain. A similar pattern can be detected in exit from the labor market. Arrangements in between a full-time job, full retirement, and outright unemployment now punctuate the end of careers. These changes have set off an identity crisis among economically inactive, aging persons, who do not see themselves as being retired or jobless but, instead, as being 'discouraged workers' who have given up looking for a job (Casey and Laczko 1989).

2. *Age Policy, a Major Policy Tool for Public Authorities*

As a neutral, universal criterion, age has been an especially useful tool for constructing the social security system. This system of social insurance needed to lay down universal conditions for eligibility; and age was retained as the most relevant criterion. Insurance against the risks of disability or unemployment only cover persons in the 'age of work.' And old-age funds only serve pensions after a regulatory age threshold. The welfare state has thus developed out of increasingly strict norms about the timing of phases in life. Systems of education and social (security) insurance funds have laid down clearly marked thresholds: a person is either a child in school (with life regulated by policies concerning childhood and parenthood) or else an adult at work (with risks covered by insurance funds) or else a retiree (entitled to a pension).

Entitlement under universal rules contrasts with the situation in societies before the creation of the welfare state. There, the passage from one phase to another (from childhood to work, for example) could be gradual and reversible, since families responded to needs in an occasional, particularistic way (Hareven 1986). The state has used this single criterion of age to perform its principal duties: redistributing revenue between age-groups and generations; maintaining order by assigning roles, statuses, activities and identities to each age-group and individual; and managing human resources.

As regards this management of human resources, Graebner (1980), in his history of retirement in the USA, has shown that the invention of old-age funds represented more than just a means for insuring older workers who could no longer work: firms used these funds to control the flow of labor. Companies could thus rationalize the withdrawal of older workers from the work force and replace them with young people—after all, the recently developed Taylorist Scientific Management had proven that older workers were less efficient. With respect to public policies in the UK since 1945, Phillipson (1982) has shown that the incentives offered to older workers to either keep or stop working have fluctuated depending on labor market needs. Older workers form a reserve army to be mobilized when there is a labor shortage. Or, on the contrary, they can be pushed out of the workforce during economic downturns, as unemployment rises.

As the state and and its bureaucracy have developed, age has become a major tool in government interventions in the society and economy. Dividing the population into age groups and adopting age-based programs now constitute the prevalent response to social problems. As a result, occupations, customers, and other targeted groups are segmented by age. Health care, for instance, is increasingly based on occupations specialized in handling age groups, such as childhood (Heyns 1988) and old age (Haber 1986), and as much can be said about social work.

3. *Implementing Age Policies: Childhood and Old Age*

3.1 *From Policing Ages to Working out a Fully-fledged Age Policy*

Childhood, as well as old age, has been identified as such only in modern times. During the Middle Ages, childhood did not exist as an autonomous phase in life; it emerged during the eighteenth century, when children were 'placed apart and reasoned' inside institutions specialized in educating them (Aries 1973). Likewise, the invention of old age owes much to retirement systems, which, by setting the age of entitlement to a pension, have established a threshold and thus assigned old age the socially uniform meaning of 'pensioned inactivity' (Guillemard 1983).

For each of these two phases, a variety of 'social laws' have been passed. In the case of childhood, laws regulate maternity and childcare, schooling, family

allocations, health, and child labor. Two major policies have defined old age. First of all, job and retirement policies have regulated the relation between age and work, and set the threshold for entrance into old age. Secondly, welfare policies for the elderly, which provide services to those who have physical disabilities or experience financial hardship, have assigned an identity and 'way of life' to old age (Guillemard 2000). These heterogeneous measures are more than a means of using age-based norms to police ages; they form a coherent, public policy for managing ages and the phases of life.

Once the phases of life had been clearly distinguished, broad, coherent social programs gradually emerged for managing them. In France during the 1960s, major public reports on youth and old age were published; and public interventions, coherently programmed with the clearly formulated aim of improving the management of these age-groups. French public authorities published in 1962 the first major report on age policy; its title, 'Old age policy,' clearly signaled a shift in government priorities. The USA adopted the Older American Act (Estes 1979). Meanwhile, several age policies have been programmed in Europe: for the social and vocational integration of young people, for infancy and for the frail elderly (Olson 1994).

3.2 *The Perverse Effects of Age Policies*

By implementing these various age-based policies, public authorities have invented and now regulate infancy, childhood, adolescence, old age, and advanced old age. They have assigned social contents and identities to these phases in life. These policies have not, however, always had effects in line with their initial objectives.

As studies of the old age policies implemented in the USA or Europe since the late 1960s have shown, the measures for providing social services and facilities to help the elderly continue living at home and avoid institutional custody (with the consequences of 'marginalization' and lessened autonomy) have made these persons dependent (Guillemard 1983, Walker 1980). The central argument in studies of these perverse effects is that, despite the good intentions underlying public interventions and despite tangible results for beneficiaries, these programs have, in general, not maintained or developed the autonomy of targeted age groups. The new arrangements for providing home services have perversely turned any physical or social disability into a form of 'dependence.' Thus has arisen a new definition of the senior citizen as the 'recipient of services whose extent and nature is decided by others' (Townsend 1981, p.19). Furthermore, the fragmented provision of home services have tended to define the beneficiary as a long list of needs for: health care, social ties, home helpers, cleaning services, etc. As a consequence, a category of 'professionals' has been assigned to satisfy each need.

4. *Conclusion: The Rise and Fall of Age Policy*

By imposing public education, by making decisions about health, social services, and the family, by creating retirement and then supporting preretirement, the state has, for more than a century now, regulated the relations between age groups and generations. But this 'age-management' has reached its limits, as several studies have shown.

In his pioneering book, Neugarten (1982) questioned both the pertinence of age-based policies to the lives of the elderly and the efficiency of public policies targeting age groups. In a call for an 'age-neutral society' she suggested organizing government interventions on the basis of needs instead of age. Extending this approach with the concept of 'structural lag,' Riley et al. (1994) has emphasized that using age-based criteria has considerably reduced the 'opportunity structures' that shape people's lives at every age. Despite longer life expectancies as well as improvements in health and ways of life, social structures are lagging behind. Because of this lag, there is a need for a new formula for combining work, family, and leisure so as to create an 'age-integrated society' where social activities are interwoven through all phases of life. The relevance of age policy has thus come under severe questioning, as proven by the European Commmission's call for a 'society for all ages' and by new interest shown for fighting against age barriers and age-based discrimination in employment (European Foundation 1997).

See also: Age, Sociology of; Age Structure; Life Course in History; Life Course: Sociological Aspects; Retirement and Health; Retirement, Economics of; Social Security; Welfare Programs, Economics of; Welfare State; Welfare State, History of

Bibliography

Aries P 1973 *L'Enfant et la Vie Familiale sous l'Ancien Régime*. Editions du Seuil, Paris

Casey B, Laczko F 1989 Early retirement, a long-term unemployment? The situation of non-working men 55–64 from 1979 to 1986. *Work, Employment and Society* **3**(4): 505–26

Estes C L 1979 *The Aging Enterprise*. Jossey-Bass, San Francisco

European Foundation 1997 *Combating Age Barriers in Employment. European Research Summary*. Office for Official Publications of the EC, Luxemburg

Graebner W 1980 *History of Retirement: The Meaning and Function of an American Institution (1885–1978)*. Yale University Press, New Haven, CT

Guillemard A M (ed.) 1983 *Old Age and the Welfare State*. Sage, London
Guillemard A M 1997 Rewriting social policy and changes within the life course organization: A European perspective. *Canadian Journal on Aging* **16**(3): 441–64
Guillemard A M 2000 *Aging and the Welfare State Crisis*. University of Delaware Press, Newark, NJ
Guillemard A M, Van Gunsteren H (eds.) 1991 Pathways and their prospects: A comparative interpretation of the meaning of early exit. In: Kohli M, Rein M, Guillemard A M (eds.) *Time for Retirement: Comparative Studies of Early Exit from the Labor Force*. Cambridge University Press, Cambridge, UK, pp. 362–88
Haber C 1986 Geriatric: A specialty in search of specialists. In: Van Tassel D, Stearns P N (eds.) *Old Age in Bureaucratic Society*. Greenwood, Westport, CT, pp. 66–84
Hareven T 1986 Historical change in the social construction of the life course. *Human Development* **29**(3): 171–80
Heyns B 1988 *The Mandarins of Childhood: Toward a Theory of the Organization and Delivery of Children's Services*. Basic Books, New York
Kohli M 1987 Retirement and the moral economy: An historical interpretation of the German case. *Journal of Aging Studies* **1**(2): 125–44
Mayer K U, Schoepflin U 1989 The state and the life course. *Annual Review of Sociology* **15**: 187–209
Neugarten B L (ed.) 1982 *Age or Need? Public Policies for Older People*. Sage, Beverly Hills, CA
Olson L K (ed.) 1994 *The Graying of the World. Who Will Care for the Frail Elderly?* Haworth, Binghamton, NY
Percheron A 1991 Police et gestion des âges. In: Percheron A, Remond R (eds.) *Age et Politique*. Economica, Paris, pp. 112–39
Phillipson C 1982 *Capitalism and the Construction of Old Age*. Macmillan, London
Riley M W, Kahn R L, Foner A (eds.) 1994 *Age and Structural Lag*. Wiley Interscience, New York
Smelser N, Halpern S 1978 The historical triangulation of family, economy and education. *American Journal of Sociology* **84**: 288–315
Townsend P 1981 The structured dependency of the elderly. *Ageing and Society* **1**(1): 5–28
Walker A 1980 The social creation of poverty and dependency in old age. *Journal of Social Policy* **9**(1): 49–75

A.-M. Guillemard

Age, Race, and Gender in Organizations

From the psychological perspective, an organization is *about* organization. It is the organization of people, time, resources, and activities. This article will consider the organization of people in two ways: how are they identified and brought into the organization, and once they enter the organization, how do individuals and the organization adapt to each other? In particular, it will consider the impact of the demographic characteristics of age, gender, and race on these two human resource processes.

1. *Relational Demography*

Pfeffer (1983) has characterized organizations as relational entities and introduced the concept of 'relational demography' (Mowday and Sutton 1993). The implication of this concept is that work group composition, and attempts to maintain or change that composition (formally or informally), may in turn influence recruiting, hiring, leadership, motivation, satisfaction, productivity, communication, and turnover. In a partial test of these propositions, it was found that as work groups increased in racial and gender diversity, absenteeism and turnover also increased (Tsui et al. 1991, Tsui and O'Reilly 1989).

Schneider (1987) introduced a similar concept, which he labeled the attraction-selection-attrition model (ASA) that emphasized the similarity of attitudes, values, personality characteristics, and interests rather than demographic characteristics *per se*. Like Pfeffer, Schneider proposed that individuals seek to limit work-group access to those most like them. Further, individuals will attempt to drive out those most unlike them. Both Pfeffer and Schneider hypothesize that group member similarity (demographic similarity for Pfeffer, and intrapersonal and interpersonal similarities for Schneider) creates trust and enhances communication, resulting in commitment, satisfaction, and effectiveness.

Jackson et al. (1991) studied management teams in the banking industry and found support for the models of both Pfeffer and Schneider. These models and the preliminary findings are critically important for two reasons: (a) team work is becoming the standard in many industries, requiring more worker interaction than ever before; and (b) most countries are undergoing a 'demographic revolution' either as a function of anticipated workforce population changes (e.g., the 'aging' workforce) or because of externally precipitated shifts (e.g., legislation inhibiting occupational segregation by race or gender, the creation of new sociopolitical entities such as the European Union, or the elimination of longstanding sociopolitical barriers such as the collection of 'Warsaw Pact' nations. As a result of many and interacting forces, race, gender, and age restrictions in the workplace are disappearing. If Pfeffer and Schneider are correct (as the data of Tsui et al. and Jackson et al. suggest they are), diversity becomes less a goal and more a challenge.

2. *Demographic Comparisons*

In considering differences that may be noted between any two demographic groups (e.g., men vs. women, old vs. young, ethnic minority vs. ethnic majority), it is useful to consider alternative explanatory models. Cleveland et al. (2000) have proposed three such models. The *biological model* assumes genetic, hormonal, and/or physical differences between groups

being contrasted. The *socialization model* assumes that any observed differences are learned. The *structural/cultural model* assumes that observed differences are the result of social structures and systems that work to maintain the status quo of a power hierarchy. As we consider the issue of relational demographics in organizations, elements of each of these models will become apparent.

3. Selection of Group Members

In the selection or promotion of employees, various attributes may be considered. These attributes include training and experience (e.g., educational accomplishments), abilities (cognitive and physical), personality, and skills (i.e., practiced acts). Further, these attributes may be used to predict a wide range of employee behaviors and outcomes including productivity, absenteeism, turnover, and satisfaction (Landy 1989).

3.1 Age

With respect to life-span development, it seems clear that the differences *within* any age stratum are exceeded by the differences *between* strata. Thus, while one might describe mean differences between any two age groups on tests of cognitive function, these differences are modest when considered in the context of their respective group standard deviations (Schaie 1982). Further, it is clear that job-relevant experience more than offsets any modest decline that might occur in job-related abilities (Schmidt et al. 1992). The same tradeoff between ability and experience is true, but to a somewhat lesser extent, with respect to the decline of physical abilities with age (Landy et al. 1992). For a wide range of jobs from managerial to unskilled labor positions, the age of the applicant should be largely irrelevant.

Recent meta-analyses have demonstrated that there are no differences in either the objective performance or the judged performance of older workers (Arvey and Murphy 1998, McEvoy and Casio 1989, Waldman and Avolio 1986). It does appear that older workers experience greater satisfaction and less absenteeism, but this may be more a function of increasing experience, skill development, and organizational position than age *per se* (Bedeian et al. 1992). When experience and job title are held constant, there seem, to be few differences in satisfaction between younger and older workers (Mangione and Quinn 1975). This confound is exaggerated by full-time/part-time status since part-time jobs tend to be more mundane and are most often held by younger workers. Once again, when part-time vs. full-time status is held constant, there are no differences in satisfaction between older and younger workers (Hollinger 1991).

3.2 Gender

There appears to be little difference between males and females with respect to general mental abilities. Although, on the average, females tend to do somewhat more poorly on tests of mathematical abilities (Feingold 1995), and are underrepresented in many scientific and engineering specialties, these differences may be the result of stereotypes held by employers, academic advisors, or by women themselves (Cleveland et al. 2000). With respect to personality differences, there are no clear-cut gender-based differences in either personality structure, or assessed personality dimensions (Hough 1998, Hough and Oswald 2000). Nevertheless, there are some substantial differences in physical abilities (Hogan 1991, Salvendy 1997), particularly in cardiovascular endurance and upper body strength. These differences need not be prohibitive, however, for physically demanding jobs since most jobs may be performed in a variety of ways that permit task accommodation as well as allow experience to offset lower levels of physical abilities (Landy 1989). As an example, several studies have suggested that women seek different work situations from men. Men tend to value compensation and opportunities for advancement to a greater extent than women; women, on the other hand, place a greater emphasis on hours of work and opportunity for social interaction. (Betz and O'Connell 1989, Chelte et al. 1982, Konrad and Mangel 2000, Tolbert and Moen 1998). These findings suggest that there may be male/female job satisfaction differences that result from the differential availability of rewards that each values. When job title and experience are held constant, there are no data to suggest systematic differences in the overall job satisfaction of males and females.

3.3 Race

It is not uncommon to find a mean score difference of as much as one standard deviation between whites and blacks on standardized multiple-choice cognitive ability tests (with blacks scoring lower) (Dubois et al. 1993). Hispanic test-takers usually fall midway between white and black test-takers, scoring approximately 0.5 standard deviations below whites and above blacks (Hartigan and Wigdor 1989, Jensen 1980, Sackett and Wilk 1994). But there is considerable overlap among the score distributions, suggesting the strong influence of cultural or structural issues. A debate has raged for decades with respect to the reason for these observed differences (Gottfredson 1994; Helms 1997) but there is no clear explanation at this point—just several intriguing hypotheses. The organizational reality, however, is that if standardized cognitive ability tests are used as the sole screening device for employment, blacks, and to a lesser extent Hispanic applicants, will be at a distinct disadvantage

when competing against white applicants. No such differences appear in physical abilities, personality tests, or structured interviews (Hough and Oswald 2000).

Since many, if not most, jobs depend on communication skills and personality characteristics, in addition to cognitive ability, it would seem obvious that assessment should cover a wide range of job-related attributes and not simply cognitive abilities. In expanding the comprehensiveness of the assessment process, nonwhite applicants can compete more favorably with their white counterparts, at the same time enhancing validity or job relatedness and diminishing the test score gap between applicant groups. With respect to outcomes such as job satisfaction, turnover, and absenteeism as was the case in other demographic groupings, there are no reliable differences between whites and nonwhites when job title and experience are held constant.

4. *Issues of Adaptation*

Assuming that women, older workers, and members of ethnic minority groups are employed by an organization, what are the issues of adaptation that need to be addressed? The adaptation challenges would seem to be similar for each of these demographic groups. The concept of a 'traditional' job is of value here (Cleveland et al. 2000, Sterns and Miklos 1995). Women and ethnic minorities are seeking access to occupations and organizational levels where they are historically underrepresented. Older workers are seeking entry to nontraditional occupations/job titles or to maintain an organizational position in spite of age-based stereotypes. In that sense, each group is 'dissimilar' to those already holding positions that group members seek.

If, as both Pfeffer and Schneider predict, 'dissimilar' members will be marginalized or forced out of the organization, what is the psychological mechanism by which such marginalization may occur? The most likely mechanism is a stereotype that is used as a heuristic for determining how women, older workers, and ethnic minority group members will be perceived. A stereotype is a set of beliefs and/or assumptions about a particular group of people. Stereotyping assumes that (a) the beliefs or assumptions are veridical, and (b) that all members of the group can be accurately characterized by that stereotype (Hilton and von Hippel 1996). The existence of race, gender, and age stereotypes has been well established (e.g., Cleveland et al. 2000, Sterns and Miklos 1995). Stereotypes can operate to the disadvantage of demographic subgroups in several different ways. The most obvious is by influencing the decisions of individual managers. Such decisions could involve access to training, promotion or job transfer, occupational segregation (Cleveland et al. 2000), compensation decisions or layoff decisions. In addition to the behavior of managers, stereotypes can be held by the stereotyped individuals themselves. This leads to individual decisions regarding 'appropriate' or 'expected' behavior. Thus, individuals may actually limit their own opportunities by accepting the assumptions and beliefs consistent with the stereotype. This results in a self-fulfilling prophecy that provides further support for the maintenance of the stereotype.

It is often assumed that stereotypes influence not only personnel decisions through the data used to support those decisions. One such data source is performance evaluation information provided through supervisory ratings. In spite of the intuitive appeal of this hypothesis, meta-analyses do not support it (Sackett and Dubois 1991, Pulakos 1989). Thus, it does not appear that performance ratings are being used to force 'dissimilar' members out of the organization. Recent research on stereotypes (Glick et al. 1988, Hilton and von Hippel 1996, Kunda and Thagard 1996) suggests that stereotypes are more likely to operate in the absence of individualized information about a particular individual. Thus, a manager may be opposed to the idea of 'older' engineers or female engineers in a department but feel very differently about a particular older or female engineer. This could account for the fact that even though there are no differences between the performance ratings of majority vs. minority, or male vs. female, or older vs. younger employees, there are still differences in occupational outcomes such as salary levels, progression to upper-level management ranks, or access to advanced training programs (Cleveland et al. 2000, Sterns and Miklos 1995, Borman et al. 1997).

Since it is clear that workgroup diversity will increase in the furure, the question remains with respect to what effect such diversity will have. As indicated above, prevailing theory (Pfeffer 1983, Schneider 1987) and data (Jackson et al. 1991, Tsui et al. 1991) suggests that efforts will be made by the 'in group' to exclude the 'out group.' This suggests lower levels of group satisfaction, group stability, and possibly group effectiveness. Some recent data suggest, however, that the dynamics are a good deal more complicated than they might appear. Jackson et al. (1995) conclude that group heterogeneity (broadly defined to include not only demographic characteristics but also background, experience, and personality) may actually enhance creative efforts of the group by widening the variety of approaches taken to a problem area. Watson et al. (1993) found that culturally homogeneous task groups performed better than heterogeneous groups initially, but that these differences were reversed after 15 weeks.

Taken as a whole, the following inferences might be drawn about the literature on workgroup or team diversity: (a) initially, there will be some tension and lowered effectiveness in demographically heterogeneous groups, (b) if the groups remain intact,

effectiveness will increase, and (c) the fewer the number of 'out group' members, the greater the initial tension and efforts to drive 'dissimilar' members out of the group. The general areas of group or team performance and group composition are only now receiving careful empirical attention (Guzzo and Dickson 1996, Landy et al. 1994). Given the inevitable diversification of workforces worldwide, the results of this research will prove valuable for organizational psychologists and managers alike.

See also: Affirmative Action: Empirical Work on its Effectiveness; Aging and Health in Old Age; Aging, Theories of; Cognitive Aging; Discrimination; Discrimination, Economics of; Discrimination: Racial; Education and Gender: Historical Perspectives; Equality of Opportunity; Gender and the Law; Gender, Class, Race, and Ethnicity, Social Construction of; Gender, Economics of; Job Analysis and Work Roles, Psychology of; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Law and Aging; Performance Evaluation in Work Settings; Prejudice in Society; Race and the Law; Sex Segregation at Work; Sexual Harassment: Social and Psychological Issues; Work: Anthropological Aspects

Bibliography

Arvey R D, Murphy K R 1998 Performance in work settings. *Annual Review of Psychology* **49**: 141–68

Bedeian A G, Ferris G R, Kacmar K M 1992 Age, tenure, and job satisfaction: A tale of two perspectives. *Journal of Vocational Behavior* **40**: 33–48

Betz M, O'Connell L 1989 Work orientations of males and females: Exploring the gender socialization approach. *Sociological Inquiry* **59**: 318

Borman W C, Hanson M A, Hedge J W 1997 Personne/-Selection. *Annual Review of Psychology* **48**: 299–337

Chelte A F, Wright J, Tausky C 1982 Did job satisfaction really drop during the 1970s? *Monthly Labor Review* **105**(11): 33–7

Cleveland J N, Stockdale M, Murphy K R 2000 *Women and Men in Organizations: Sex and Gender Issues at Work*. Lawrence Erlbaum, Mahwah, NJ

DuBois C L Z, Sackett P R, Zedeck S, Fogli L 1993 Further exploration of typical and maximum performance criteria: definitional issues, prediction, and White-Black differences. *Journal of Applied Psychology* **78**: 205–11

Feingold A 1988 Cognitive gender differences are disappearing. *American Psychologist* **43**: 95–103

Glick P, Zion C, Nelson C 1988 What mediates sex discrimination in hiring decisions? *Journal of Personality and Social Psychology* **55**: 178–86

Gottfredson L S 1994 The science and politics of race norming. *American Psychologist* **49**: 955–63

Guzzo R A, Dickson M W 1996 Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology* **47**: 307–39

Hartigan J A, Wigdor A K (eds.) 1989 *Fairness in Employment Testing: Validity Generalization, Minority Issues, and the General Aptitude Test Battery*. National Academy Press, Washington, DC

Helms J E 1997 The triple quandary of race, culture, and social class in standardized cognitive ability testing. In: Flanagan D P, Genshaft J, Harrison P L (eds.) *Contemporary Intellectual Assessment: Theory, Tests, Issues*. Guilford Press, New York pp. 517–32

Hilton J L, von Hippel W 1996 Stereotypes. *Annual Review of Psychology* **47**: 237–71

Hogan J C 1991 Physical abilities. In: Dunnette M D, Hough L M (eds.) *Handbook of Industrial and Organizational Psychology*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA, Vol. 2

Hollinger R 1991 Neutralizing in the workplace: An empirical analysis of property theft. Deviant Behaviour: In interdisciplining Journal **12**: 169–202

Hough L 1998 Personality issues at work: Issues and evidence. In: Hakel M D (ed.) *Beyond Multiple Choice: Evaluating Alternatives to Traditional Testing for Selection*. Erlbaum, Mahwah, NJ pp. 131–66

Hough L M, Oswald F L 2000 Personnel selection: Looking toward the future—Remembering the past. *Annual Review of Psychology* **51**: 631–64

Jackson S E 1995 Understanding human resource management in the context of organizations and their environments. *Annual Review of Psychology* **46**: 237–64

Jackson S E, Brett J F, Sessa V I, Cooper D M, Julin J A, Peyronnin K 1991 Some differences make a difference: individual dissimilarity and group heterogeneity as correlates of recruitment, promotion, and turnover. *Journal of Applied Psychology* **76**: 675–89

Jensen A R 1980 *Bias in Mental Testing*. Free Press, New York

Konrad A M, Mangel R 2000 The impact of work-life programs on firm productivity. *Strategic Management Journal* **21**: 1225–37

Kunda Z, Thagard P 1996 Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review* **103**: 284–308

Landy F J 1989 *Psychology of Work Behavior*, 4th edn. Brooks/Cole, Pacific Grove, CA

Landy F J, Bland R E, Buskirk E R, Daly R E, DeBusk R F, Donovan E J, Farr J L, Feller I, Fleishman E A, Gebhardt D L, Hodgson J L, Kenney W L, Nesselroade J R, Pryor D B, Raven P B, Schaie K W, Sothmann M S, Taylor M C, Vance R J, Zarit S H 1992 *Alternatives to chronological age in determining standards of suitability for public safety jobs. Technical Report*. The Center for Applied Behavioral Sciences, Penn State University, PA

Landy F J, Shankster L, Kohler S S 1994 Personnel selection and placement. *Annual Review of Psychology* **45**: 261–96

Mangione T W, Quinn R P 1975 Job satisfaction, counterproductive behavior, and drug use at work. *Journal of Applied Psychology* **60**: 114–16

McEvoy G M, Casio W F 1989 Cumulative evidence of the relationship between employee age and job performance. *Journal of Applied Psychology* **74**: 11–17

Mowday R T, Sutton R I 1993 Organizational behavior: Linking individuals and groups to organizational contexts. *Annual Review of Psychology* **44**: 195–229

Pfeffer J 1983 Organizational demography. *Research in Organizational Behavior* **5**: 299–357

Pulakos E D, White L A, Oppler S H, Borman W C 1989 Examination of race and sex effects on performance ratings. *Journal of Applied Psychology*. **74**: 770–80
Sackett P R, Duzois C L 1991 Rater-ratee race effects on performance evaluation: challenging meta-analytic conclusions. *Journal of Applied Psychology* **76**: 873–77
Sackett P R, Wilk S L 1994 Within-group norming and other forms of score adjustment in pre-employment testing. *American Psychologist* **49**: 929–54
Salvendy G 1997 *Handbook of Human Factors and Ergonomics*. Wiley, 2nd edn. New York
Schaie W 1982 Longitudinal data sets: Evidence for ontogenetic development or chronicles of cultural change. *Journal of Social Issues* **38**: 65–72
Schmidt F L, Ones D S, Hunter J E 1992 Personnel selection. *Annual Review of Psychology* **43**: 627–70
Schneider B 1987 The people make the place. *Personnel Psychology* **40**: 437–53
Sterns H L, Miklos S M 1995 The aging worker in a changing environment: Organizational and individual issues. *Journal of Vocational Behavior* **47**: 248–68
Tolbert P S, Moen P 1998 Men's and women's definitions of 'good' jobs: similarities and differences by age and across time. *Work and Occupations* **25**: 168–195
Tsui A S, Egan T, O'Reilly C A 1991 Being different: Relational demography and organizational attachment. *Academy of Management Best Paper Proceeding's* **37**: 183–7
Tsui A S, O'Reilly C A 1989 Beyond simple demographics: The importance of relational demography in superior–subordinate dyads. *Academy of Management Journal* **32**: 402–23
Waldman D A, Avolio B J 1986 A meta-analysis of age differences in job performance. *Journal of Applied Psychology* **71**: 33–8
Watson W E, Kumar K, Michaelsen L K 1993 Cultural diversity's impact on interaction process and performance: Comparing homogeneous and diverse task groups. *Academy of Management Journal* **36**: 590–602

F. J. Landy

Age, Sociology of

The sociology of age is currently developing as a broad, multifaceted approach to research, theory, and policy on age in society. It is concerned with: (a) people as they grow older and as cohorts succeed one another; (b) age-related social structures and institutions; and (c) the dynamic interplay between people and structures as each influences the other. These concerns reflect the hallmark of sociology itself which, unique among the sciences, emphasizes (a) people, (b) structures, and (c) the relationships among them. Thus the sociology of age, although it has so far paid more attention to (a) than to (b) and (c), provides a potential focal point for diverse multidisciplinary contributions to sociology as a whole (see *Age Stratification*).

As an emerging specialty, it will be described in this article in terms of its complex history, its consolidation of work in related fields, its developing working principles and research findings, and its promise to stand beside class, ethnicity, and gender in the sociology of the future.

1. An Emergent Field of Sociology

The sociology of age is comparatively new. It has been taking shape as a special field of sociology only since the 1970s. In 1972 Riley et al.'s *Sociology of Age Stratification* established an analytical framework for understanding the age-related dynamic interplay between people (actors) and roles (social structures). Similarly, in 1982 the Max Planck Institute for Human Development organized a center on life-course sociology and social-historical change. By 1988 the field was given separate status in a *Handbook of Sociology* (Smelser 1988). Attention to age began to crystallize at this time with the general recognition of the unprecedented increases in longevity, coupled with dawning awareness of the long-term impact of the 'baby boom cohorts.'

Though not a central focus, age—always a topic of primordial interest—had long been in the sociological air. Among the classical forerunners, Pitirim Sorokin, Talcott Parsons, S. N. Eisenstadt, and Leonard Cain wrote on age as an aspect of social structure; W. I. Thomas and Florian Znaniecki, as well as Bernice Neugarten, examined aging (growing older) as a social process; and Karl Mannheim tied the characteristics of successive generations—or 'cohorts' of people born at the same period of time—to historical and cultural change. Thus the meaning of 'age' was marked early by the distinction between the noun 'age,' as both a component of people's lives and a structural criterion for occupying and performing in roles, and the verb 'aging,' as interacting biological, psychological, and social processes from birth to death. However, this early work largely failed to explicate the irreducible reciprocity between human development and the changing society.

2. Convergences and Ramifications

As the sociology of age has been emerging as a special field, it is enriched by consolidating fragmentary work on the stages of people's lives and, at the same time, by reaching out toward multiple related disciplines.

2.1 Convergence of Life Stages

The sociology of age has gradually encompassed several subfields which have appeared sporadically, often with little awareness of each other. Paramount among these subfields is *old age*, which in the 1990s has seen a staggering rise in popular and policy attention

to the problems of older people, along with substantial scientific work supported by many agencies (notably, in the United States, by the National Institute on Aging). Forerunners include the epoch-making *Handbook of Social Gerontology* (Tibbitts 1960) that ranged over societal aspects of old age from population to values to technological and social change; it initiated a continuing series of handbooks. Back in 1950, Kingsley Davis (Davis and Combs 1950; see also Cowgill 1974) foretold that the growing numbers of older people and the rapid pace of social change would contribute to the isolation of older people and make them 'useless'—a social problem of falsely perceived uselessness which still requires correction today.

Starting even earlier, work on *childhood and adolescence* has also been proliferating, supported professionally and financially by its own organizations and agencies (notably, in the United States, by the National Institute of Child Health and Human Development). A variety of sociological studies have complemented psychological models by relating individual development to social structure and social change, for example, Glick (1947) to the 'family cycle' within which children and parents influence each other; Smelser (1968) to the movement of families into factories and mines during the Industrial Revolution; Elder (1974) to the hardships of families during the Great Depression, Vygotsky (see Cole et al. 1978) to the adult world through communal processes of sharing; and Corsaro (1997) to childhood as a structural form within which children are active agents. Meanwhile, Hernandez (1993) has analyzed in extraordinary detail the impact of several societal revolutions on the life course of children and adolescents over the past 150 years in the United States, and also in other countries. Family size plummeted. One-parent family living jumped. Family farms became rare. Formal schooling and nonparental care for children increased greatly. Unlike early longitudinal studies of aging (as by George Maddox and Gordon Streib), Hernandez, like other modern analysts, could benefit from the computer revolution and availability of data banks.

Perhaps because most work centers on adulthood as a generic category, less specific attention has been paid to the middle years *per se*, save for promised sociological reports from one of the MacArthur Foundation Networks.

Overarching these subfields is the focus on the life course as it extends from birth (or conception) to death (e.g., Marshall 1980). Intellectual progress in this area is reflected in the work of John Clausen, who corrected his limited masculine and cohort-centric 1972 analysis by his well-rounded 1986 account of both stability and change in the lives of individuals traced longitudinally over a 50 year period. Major theories of the 'institutionalization of the life course' assert that Western societies have actually been constructed to fit the 'three-box' pattern of education for the young, work and family responsibilities for the middle-aged, and leisure reserved for the old (see Meyer 1986, Kohli et al. 1991). At the same time, marked *diversity* in individual lives has been emphasized (Neugarten 1968, Dannefer 1987). Today, important as life-course studies are (see Giele and Elder 1998), the broader sociology of age awaits development of the complementary dynamics of social structures.

2.2 Multidisciplinary Ramifications

In addition to simplifying complexities in its own field, the sociology of age has been incorporating and in turn enriching relevant work in other disciplines (e.g., Bengtson and Schaie 1999). From the outset it was recognized that age-related structures cannot be understood without reference to all the social sciences—as, for example, history shows how age criteria for entering or retiring from work were institutionalized through industrialization, or anthropology shows how attitudes and feelings develop in nursing home care. Even more impressive, because aging consists of biological and psychological as well as social processes, the sociology of age has been a forerunner in the rapprochement with psychology and the life sciences (e.g. Hess 1976)—as in studies of genetic predispositions or age-related stress. Moreover, the dynamic character of aging has suggested revisions of earlier static models of age because, while people are growing older, society is recognized as changing around them, affecting both structures and the very process of aging.

3. Working Principles and Findings

Underlying such convergences and ramifications, certain common assumptions and working principles have been identified in the sociology of age, as they relate to the accumulating research findings—and also recapture the dual emphasis on structures and lives. Some principles have departed from accepted usage (for example, that intelligence declines after a peak at age 20), and some findings were demonstrably fallacious (for example, cross-sectional age differences were often erroneously interpreted as describing the process of aging). Gleaned from decades of work, the central theme of the sociology of age is that, against the backdrop of history, *changes in people's lives influence and are influenced by changes in social structures and institutions*. Linked to these reciprocal changes, just three sets of inter-related principles and findings can be illustrated in this article.

3.1 Inevitability of Change

Neither lives nor structures are entirely fixed or immutable, but vary in complex ways (Foner 1975). In the sociology of age they are conceived as 'two dynamisms'—changing lives and changing structures—that are interdependent yet distinct sets of processes. Study of the interplay between these dynamisms has succeeded in freeing the meaning of age from its early dependence on biological determinants, and has begun to suggest new meanings in a society undergoing fundamental change.

3.2 Import of Cohort Differences

Seeking to explain how lives are influenced by social as well as biological factors led sociologists to examine cohort differences (Ryder 1965, Uhlenberg and Riley 1996). The principle was formulated that: *because society changes, members of different cohorts age in different ways*. Over their lives from birth to death people move through structures that are continually altered with the course of history, thus the life patterns of those who are growing old today *cannot* be the same as the lives of those who grew old in the past or of those who will grow old in the future. Indeed, large cohort differences have been found in the aggregates of people's standard of living, educational level and technical skills, health and functioning, attitudes toward other people, and views of the world. Cohort differences characterize even the very young: newborns now weigh more than their predecessors, and children now become sexually active much earlier. Such changed characteristics of cohort members now young are bound to have predictable consequences for their later lives—their occupational trajectories, gender relationships, health and functioning.

3.3 Imbalances Between Lives and Structures

As such cohort differences were observed, it became apparent that the interplay with the dynamism of changing structures does not run smoothly. Although the two dynamisms are interdependent, *differences in timing—or asynchrony—are inherent in the interplay*. The biological lifetime of people has a definite (though variable) rhythm from birth to death. The timing of structural processes has no comparable rhythm or periodicity, but is going through entirely different historical transformations. Thus 'imbalances' arise between what people of given ages need and expect in their lives and what structures have to offer. These *imbalances exert strains on both the people and the social institutions* involved, creating pressures for further change. A current example of imbalance is 'structural lag,' as society has failed to provide opportunities in education, family, or work for the growing numbers of competent older people whose longevity is unprecedented in all history (Riley et al. 1994).

4. Future Prospects

Such intertwined principles and findings are laying the groundwork for future continuities, as the sociology of age has already begun to broaden its reach across concepts and countries.

4.1 Conceptual Reach

As one possible response to structural lag, the concept of *age integration* has been postulated as an extreme type of structure, in opposition to the extreme 'age differentiated' type of the well-known three boxes. Though originally defined as 'ideal' in Max Weber's classic sense, age integration is in some respects becoming real. Thus, the *age barriers* dividing education, work and family, and retirement are becoming more flexible (noted in the 1970s by Gusta Rehn); and incentives are sometimes advocated for interspersing these activities over people's extended lives (Riley and Loscocco 1994). Moreover, as the barriers are reduced, age integration *brings people of different ages together*. Any future shift toward age integration would challenge both the constraints placed by age on the familiar rigid structures, and also the age-related norms of 'success' and materialism now institutionalized in those structures and incorporated into people's lives.

4.2 Cross-national Reach

Future work in the sociology of age will certainly also extend to all ages its international reach. This reach dates back to the major studies of three countries by Shanas and her collaborators (1968), which showed that older people's living apart from children does not necessarily mean abandonment. The internationalism long nourished by various research committees of the International Sociological Association was updated by a 1998 discussion of age integration by scholars from seven countries. Alan Walker, speaking there about grass roots movements among older people, opined that Europeans themselves are beginning to seize the initiative in thinking that aging should not be defined as 'a simple matter of adjustment and peaceful retirement but that older people should be fully integrated citizens.' With the increasing globalization of science in the future, the multidisciplinary sociology of age should continue throughout the industrialized world as a focal point for sociology as a whole.

See also: Age Policy; Age Stratification; Age Structure; Aging, Theories of; Cohort Analysis; Generations,

Relations Between; Generations, Sociology of; Life Course in History; Life Course: Sociological Aspects; Population Aging: Economic and Social Consequences; Population Cycles and Demographic Behavior; Structure: Social

Bibliography

Bengtson V L, Schaie K W 1999 *Handbook of Theories of Sociology*. Springer, New York

Clausen J A 1972 The life course of individuals. In: Riley M W, Johnson M, Foner A (eds.) *Aging and Society: A Sociology of Age Stratification*. Russell Sage, New York

Clausen J A 1986 *The Life Course: A Sociological Perspective*. Prentice-Hall, Englewood Cliffs, NJ

Cole M, John-Steineer V, Scribner S, Sauberman E 1978 *Mind in Society: The Development of Higher Psychological Processes. L. S. Vygotsky*. Harvard University Press, Cambridge, MA

Corsaro W A 1997 *The Sociology of Childhood*. Pine Forge Press, Thousand Oaks, CA

Cowgill D O 1974 The aging of populations and societies. *Annals of the American Academy of Political and Social Science* **415**: 1–18

Dannefer D 1987 Accentuation, the Matthew effect, and the life course: Aging as intracohort variation. *Sociological Forum* **2**: 211–36

Davis K, Combs J W Jr. 1950 The sociology of an aging population. In: Armstrong D B (ed.) *The Social and Biological Challenge of our Aging Population: Proceedings*. Columbia University Press, New York

Elder G H Jr. 1974 *Children of the Great Depression: Social Change in Life Experience*. University of Chicago Press, Chicago

Foner A 1975 Age in society: Structure and change. *American Behavioral Scientist* **19**: 144–65

Giele J Z, Elder G H Jr 1998 *Methods of Life Course Research: Qualitative and Quantitative Approaches*. Sage, Thousand Oaks, CA

Glick P C 1947 The family cycle. *American Sociological Review* **12**: 164–74

Hernandez D J 1993 *America's Children: Resources from Family, Government and the Economy*. Russell Sage, New York

Hess B B 1976 *Growing Old in America*. Transaction Books, Edison, NJ

Kohli M, Rein M, Guillemard A M, van Gusteren H 1991 *Time for Retirement: Comparative Studies of Early Exit from the Labor Force*. Cambridge University Press, New York

Marshall V W 1980 *Last Chapters: A Sociology of Aging and Dying*. Brooks/Cole, Monterey, CA

Meyer J 1986 The institutionalization of the life course and its effect on the self. In: Sorenson A B, Weinert F E, Sherrod L R (eds.) *Human Development and the Life Course: Multidisciplinary Perspectives*. Erlbaum, Hillsdale, NJ

Neugarten B L 1968 *Middle Age and Aging: A Reader in Social Psychology*. University of Chicago Press, Chicago

Riley M W, Foner A, Moore M E 1968–72 *Aging and Society*. Russell Sage, New York

Riley M W, Kahn R L, Foner A, Mack K A 1994 *Age and Structural Lag: Society's Failure to Provide Meaningful Opportunities in Work, Family, and Leisure*. Wiley, New York

Riley M W, Loscocco K A 1994 The changing structure of work opportunities: Toward an age-integrated society. In: Abeles R P, Gift H C, Ory M G (eds.) *Aging and the Quality of Life*. Springer, New York

Ryder N B 1965 The cohort as a concept in the study of social change. *American Sociological Review* **30**: 843–61

Shanas E, Townsend P, Wedderburn D, Frijs H, Stehouwer J 1968 *Old People in Three Industrial Societies*. Atherton Press, New York

Smelser N J 1968 *Essays in Sociological Explanation*. Prentice-Hall, Englewood Cliffs, NJ

Smelser N J 1988 *Handbook of Sociology*. Sage, Newbury Park, CA

Tibbitts C 1960 *Handbook of Social Gerontology: Societal Aspects of Aging*. University of Chicago Press, Chicago

Uhlenberg P, Riley M W 1996 Cohort studies. In: Birren J (ed.) *Encyclopedia of Gerontology: Age, Aging, and the Aged*. Academic Press, San Diego, CA

M. W. Riley and A. Foner

Age Stratification

In the human sciences, stratification generally refers to those forms of differentiation that entail ordinal ranking along a defined dimension. Age is an inherently ordinal phenomenon, anchored in the intersection of time and the event of birth. Those individuals born within a given time period comprise a cohort. The defining cohort feature of lived time, or age, defines age difference. Age strata, however defined, are thus comprised of the continuous succession of living cohorts, glimpsed at a single point in time.

1. *Age Stratification as a Feature of Populations*

Figure 1 depicts the age composition of the US population in 1880 and the projected pattern for 2020. The ranked difference is determined by different dates of birth, and succeeding cohorts are essentially piled on top of each other at a single point in time to form a 'snapshot' of the age composition of the population. A comparison of these two figures reflects the dramatic changes that occurred in the intervening 140 years, including population growth, increased life expectancy, and the shift of the societal burden of dependency away from youth and toward age. The bulge produced by the maturing baby-boom cohorts (b. 1946–64) is also clearly visible.

Such changes in age composition are explained by fertility, mortality, and migration. Changes in these processes, in turn, can result from a diverse array of technological and other social forces; the shape of the age distribution also has an independent impact on society more generally. Scholars of age recognize immediately the consequences of such shifts in age

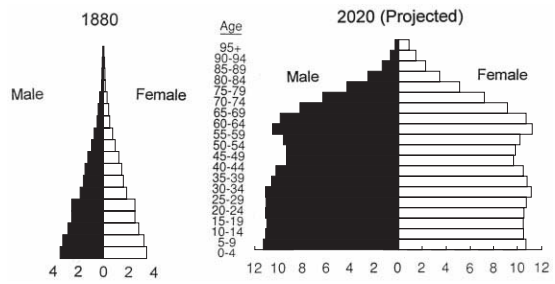

Figure 1
US population in millions by five-year age categories, 1880 and 2020 (projected) (source: Population Reference Bureau 1994, Thompson and Whelpton 1933 (adapted))

distribution of a population, which can affect aspects of society as diverse as employment prospects for graduating students, the meaning of old age, marriage markets, and one's economic life chances.

Taken alone, however, such figures reveal nothing about how age is implicated in the overall structure of society. The effects of age are always contingent on specific institutionalized regimes. For example, a shift in the aged dependency ratio might seem a very different kind of problem in modern welfare states than in present-day Russia, where economic decline has produced acute deprivation which especially affects the aged. Thus, analysis of age strata inevitably requires a consideration of stratification as a component of social structure.

2. *Age Stratification as a Feature of Social Structure*

2.1 *The Cultural Significance of Age*

As a feature of social structure, the boundaries and character of age strata are anchored in the broader social meaning and definition of age and time. The manner in which societies define age and cohort membership is, of course, not universal but variable. Indeed, the very awareness of age and cohort membership is variable and, ultimately, socially constituted as a more or less integral feature of the larger social order. Societies vary in how they count, in how they measure the passage of time, and in how they identify time of birth. In some societies, the age set—a group of individuals spanning several years—is the central social unit defining life stage and role transition processes (Foner and Kertzer 1978). In others, age is defined in remarkably descriptive terms. For example, Rwandans traditionally have not thought of individual age in metric terms, but only in relation to memorable events, such as political transitions or natural disasters. The taken-for-granted modern practices of defining age and time by precisely quantified calibrations are not essential elements of human nature, but are products of historically and socially specific systems of language and ideas.

2.2 *Age and Bureaucratic Rationality*

Within a single society, the significance of age may undergo dramatic alteration in response to changes in other aspects of social structure. An obvious prototypical case is the transformation to modernity, which generally meant the emergence of age-graded strata supported by newly identified life stages, the legal-bureaucratic reliance on age as an eligibility criterion, and an increase in age consciousness (Chudacoff 1989). The result has been the institutionalization of the life course (Kohli 1986). Viewed as a contemporaneous societal cross-section, this same phenomenon comprises the institutionalization of age strata in which the everyday experience of members of the several age strata are differentially organized by legal, governmental, and corporate policies and broader social practices that are explicitly age-graded. In age-graded societies such as modern welfare states, stratum membership is a basis for strong predictions of one's likelihood of being socially engaged or 'significant' (Uhlenberg 1988).

Standard sociological wisdom declares that modern states assign roles based on universalistic criteria that reflect individual abilities rather than sponsorship or ascription. It is thus ironic that these societies have developed criteria (both formal and informal) to govern access to and exclusion from valued social positions based on the ascribed characteristic of age. 'Children,' 'adolescents,' and 'retirees' are examples of life-stage constructs that authoritatively define capabilities based on age, even though these 'life stages' were largely unheard of 150 years ago.

Age has proven to be a useful criterion for bureaucracies faced with the task of managing and processing large populations. It has been used to regulate access to scarce roles and resources, and for attributing competence (or its lack) to individuals based on generalized assumptions of age-related capacities rather than actual abilities. Just as ageism is a social bias that often goes unrecognized by the most passionate activists for social justice in other areas, the rationality of age as an exclusionary criterion is seldom questioned by the rational-bureaucratic logic of the systems that rely upon it.

3. *Age Stratification and Social Theory*

3.1 *The Age Stratification Paradigm*

The organization of society in terms of an age-stratified social structure must be analytically dis-

tinguished from the composition of the population of individuals of different ages who occupy positions in the social structure, and who move through a sequence of age-graded roles as they age. This fundamental sociological distinction between actors and structure, persons and roles, was articulated for gerontology by Matilda Riley and associates three decades ago, but is still often overlooked (Riley et al. 1972). Confusion can arise if this analytical distinction is not recognized (e.g., when normative age-graded roles become taken for granted and assumed to be part of human nature.) The importance of this distinction was thus, from the beginning, a key premise of what Riley initially termed the age stratification perspective, a framework that crystallized a number of related theoretical principles.

One such principle is what Riley and associates term structural lag—'society's failure to provide meaningful roles...,' for people of all ages (1994). A contemporary, archetypal example is the incongruity of (a) a dramatic growth in longevity and late-life vitality among aged individuals, and (b) social institutions and policies that are premised on entrenched ageist beliefs in the incompetence and obsolescence of the aged. Riley's work thus has given a rigorous sociological foundation that supported or anticipated other scholarly (and popular and activist as well) efforts to question the prevailing social organization of age roles.

Some have criticized the age stratification approach for its reliance on terminology associated with functionalism. In fact, her approach mobilized classical concepts to highlight the crucially important distinction between structural (e.g., roles, norms) and personal (e.g., attitudes, abilities) characteristics, with the intent of challenging conventional views that prevailed within as well as beyond the social sciences. It is only by making clear, for example, that men and women in their forties are not *naturally* supposed to remain in a single career throughout their working life and old people must not *inevitably* 'retire,' that it becomes possible to have a basis for a critical analysis of the social organization of age. Allocation implies the existence of an age-stratified opportunity structure with a finite number of roles, within which individuals must find meaningful social engagement. Given the limited elasticity in the number of available age-graded roles, there is considerable potential for an 'imbalance between persons and roles' which can mean an exclusion from desired social participation that is costly for both the individual and society. Since the cohorts that populate strata are constantly aging, this is especially true when there are sharp discontinuities in the size of adjacent strata (or changes in the aspiration or abilities of stratum members). In stratification terms, such a situation may be called a disordered age structure, a cross-sectional 'snapshot' perspective on Waring's (1975) important concept of disordered cohort flow. In sum, the age stratification perspective has explicated the issue of whether the potentials of individuals can be realized within an insitutionalized role structure imposed by governmental and corporate practices and the resultant norms that regulate access to roles on the basis of age. The systems that produce allocation problems also legitimate and sustain age norms.

Thus, the age stratification framework has been used to develop a critique of the prevailing age-stratified role structure, often called the 'three boxes of life'—school, work, and retirement (Riley and Riley 1999). Using this framework, Riley proposed, alternatively, an age-integrated society, where control of the design of the individual life course shifts away from an institutional regime that has tended to stratify individual opportunities and normative possibilities according to age stratum boundaries, and toward an arrangement that affords a greater voice for self-expression.

Beyond this general perspective of social critique, the age stratification perspective has contributed to several substantive lines of research and theory, at the same time that work in related historical, demographic, economic, and sociological traditions have generated ideas that are relevant to age stratification. As examples, two general lines of theorizing will be briefly discussed: (a) the effects of cohort size and composition, and (b) the potentials for intergenerational conflict.

3.2 *Effects of Stratum Size*

Several scholars have argued that cohort size (and hence, stratum size) has a range of consequences for the lives of cohort/stratum members, including psychological (Waring 1975) and socioeconomic (Easterlin 1987) effects. The general argument is that the large cohorts (which constitute densely populated age strata) are disadvantaged relative to smaller ones, since they must compete for scarce, age-graded roles and resources. Such ideas have found support in analyses of education and work careers (Dannefer 1988). Of course, all such notions rest on the premise of a rigidly stratified age-role structure: of age as a normative if not legal qualification for occupying desirable educational and occupational positions. To the extent that this is true, the movement of unusually large cohorts through the age structure should mean that the status and resources of the various strata will change over time as the size of strata changes.

The rapid demographic change of the past century has provided a natural laboratory for exploring hypotheses about the relation of the age structure of the population and individual lives. The dramatic and still continuing increase in life expectancy has produced a population explosion in the most aged strata. Here, too, the principle of being disadvantaged by being in a large cohort has been advanced.

In preindustrial USA, the very old were often seen as the experts regarding health and longevity (cf.

Achenbaum 1979). They were few, and with no prestigious medical profession, many of these survivors carried a mystique that seemed to imply expertise. Nevertheless, the status of the aged in nineteenth century USA as in other premodern societies derived from much more than the rarity of nonagenarians. It was broadly rooted in the control of resources by the aged, and the limited options open to young people —circumstances fixed by laws of inheritance and other customary practices that organized the distribution of power in society.

3.3 Interstratum Inequality: The Intersection of Age and Other Bases of Stratification

As an ordinally ranked characteristic, age itself is an analytically distinguishable basis of stratification. Yet its significance typically involves its intersection with other dimensions of stratification, especially those involving control of economic, political, or cultural resources. However, the interpretation of interstratum age differences becomes a central concern. Stratum differences in any measurable characteristic (e.g. wealth, political attitudes; intrastratum inequality) may represent life-course, cohort, or period effects (Riley et al. 1972, Dannefer 1988). Partly because of the potential confusion that arises from issues related to discussing age and such other bases of stratification together, Riley (Riley and Riley 1999) recently renamed her analytic framework the 'Aging and Society' paradigm.

In a cross-cultural analysis of age inequality and conflict, Nancy Foner (1983) identified several factors that privilege senior age strata in many societies: control over human and material resources; knowledge, expertise and experience; prestige and positions of authority; community influence; wisdom or mystical power.

Many of these factors also appeared to operate in preindustrial Western societies. With the transformation to modernity, the aged lost both status and economic power, as the venerated characteristics of experience, skill, and wisdom were supplanted by physical strength and endurance, youthful beauty, up-to-date knowledge, and willingness to deal with rapid change.

Some scholars have argued that this structural transformation brought relief to strong intergenerational tensions that were present in many traditional agrarian families in New England and in Europe as, for example, when senior landholders survived until their children were well into or past middle age. When the family was replaced by the firm as the primary unit of economic production, the central familial relationship of economic dependency was removed (Kertzer and Laslett 1995). A hypothesized result is that the relations of adult children and aging parents were premised on volition and sentiment, rather than economic issues.

By contrast, others have proposed that the rapid pace of technological and educational change characterizing modernity tended to create cleavages between cohorts in worldview and values, dramatically increasing the likelihood of interstratum conflict.

The characterization of modernity as removing the issue of intergenerational economic strain places the entire topic of the prospect of interstratum conflict on subjective grounds. Both arguments focus on values and sentiments; in neither case is economics central to predictions about conflict. Nevertheless, the possibility of interstratum conflict based on the alignment of age and economic interest re-emerged in the late twentieth century as concern over the cost of entitlements for the age became a widespread public concern in most modern and postindustrial societies.

3.4 Economics and Generational Equity

The politicized Generational Equity controversy that has developed since 1980 in the USA returns economic issues to the foreground of potential age conflict, and illustrates the potential for age to become a basis of interest group politics. The forces underlying the debate were deeper than political opportunism. Dramatic shifts in resources among age groups, especially from children to old people, occurred, traceable to the success of policies that award special economic consideration to the senior age strata, mostly as a result of age-qualified pensions and transfer payments (Preston 1984). If being in a large cohort/stratum is disadvantageous for educational and employment opportunities, it may be the reverse for interest group politics, where the senior strata are recognized as comprising an active and growing segment of the electorate. The apparent failure of the Generational Equity campaign to polarize age strata was predicted by theoretical principles set forth by Foner. One such principle, age mobility, sees the inevitability of aging as producing an anticipation on the part of midlife adults of their own certain movement into more senior age strata (Foner 1974). A second principle involves the insulation from direct responsibility for dependent parents that Social Security and Medicare/Medicaid provide to adult offspring. Although intergenerational resource transfers may be more likely to go from parents to children than the reverse, the protection that many midlife adults are afforded by public subsidy of seniors is very real. Indeed, 'downward' resource transfers may depend, in many cases, on the public benefits afforded the senior generations.

Other sources of resistance to the so-called 'generational equity' movement may include a belief in intergenerational economic continuity within families, which implies a relatively stable intergenerational

reproduction of the structure of economic inequality along familial lines. For example, it may be the affluent neighbors or overpaid supervisor who become a focus of a sense of economic deprivation and limited life chances, and not one's comfortably situated parents.

3.5 Interstratum Variation in Intrastratum Inequality

Awareness of nonaged-related economic inequalities that are based on stratification of the general opportunity structures is made even more likely since intrastratum economic disparities appear to be greater in aged strata than in younger ones, as economic inequalities appear to cumulate across the life course of each succeeding cohort (Dannefer 1988, O'Rand 1996), creating a picture of an age structure in which inequality is higher in older age strata than in younger ones.

Despite characterizations of many traditional societies in gerontocratic terms, it appears that many aged in such societies enjoyed neither wealth nor status. Evidence for this comes from studies of the preindustrial US (e.g., Demos 1978) as well as from traditional societies (Foner 1983). If the most powerful members of traditional societies occupied the senior age strata, it does not follow that most members of those age strata were affluent or particularly respected. Just as intercohort differences in patterns of intracohort inequality has been a neglected area of life course research, interstratum differences in intrastratum inequality is a neglected question in analyses of age structure.

4. Summary

Age stratification brings together the deceptively elusive individual characteristic of age, and the complex and dynamic social phenomenon of stratification. A grasp of society as a systemic structural reality that shapes many aspects of aging has been slow to develop among those who focus on the individual-level characteristics of age and development. Conversely, a grasp that individual social actors—whether citizens, students, parents or workers—are arrayed in continuously aging and changing cohorts has been underappreciated by sociologists. The study of age stratification has contributed a clear and forceful message concerning the importance of distinguishing age and age-specific subpopulations on the one hand from normatively age-graded social practices and structures on the other. It has contributed principles helpful in understanding how population and social structure jointly impact individual life chances, and has begun to consider the conditions governing the form of the relation of age and other bases of stratification. As issues such as the impact upon individuals of age-graded structures and the question of how socioeconomic stratification and age may interact, new research agendas have begun to take shape around questions of stratification and age.

See also: Adolescent Behavior: Demographic; Age: Anthropological Aspects; Age Policy; Age, Race, and Gender in Organizations; Age, Sociology of; Age Structure; Cohort Analysis; Generations, Relations Between; Generations, Sociology of; Life Course in History; Life Course: Sociological Aspects; Life Expectancy and Adult Mortality in Industrialized Countries; Population Aging: Economic and Social Consequences; Population Cycles and Demographic Behavior; Population, Economic Development, and Poverty; Social Stratification

Bibliography

Achenbaum A 1979 *Old Age in the New Land: The American Experience Since 1790*. Johns Hopkins University Press, Baltimore, MD

Chudacoff H 1989 *How Old Are You? Age Consciousness in American Culture*. Princeton University Press, Princeton, NJ

Dannefer D 1988 Differential aging and the stratified life course. In: Maddox G L, Lawton M P (eds.) *Annual Review of Gerontology and Geriatrics*. Springer, New York, Vol. 8

Demos J 1978 Old age in early New England. *American Journal of Sociology* **84**: S248–87

Easterlin R 1987 *Birth and Fortune: The Impact of Numbers on Personal Welfare*. 2nd edn. University of Chicago Press, Chicago, IL

Foner A 1974 Age stratification and age conflict in political life. *American Sociological Review* **39**: 187–6

Foner A, Kertzer D 1978 Transitions over the life course: Lessons from age-set societies. *American Journal of Sociology* **83**: 1081–1104

Foner N 1984 *Ages in Conflict: A Cross-cultural Perspective on Inequality Between Old and Young*. Columbia University Press, New York

Kertzer D I, Laslett P (eds.) 1995 *Aging in the Past: Demography, Society and Old Age*. University of California Press, Berkeley, CA

Kett J 1977 *Rites of Passage: Adolescence in America, 1790–1920*. Free Press, New York

Kohli M 1986 Social organization and subjective construction of the life course. In: Sorensen A, Weinert F, Sherrod L (eds.) *Human Development and the Life Course*. L. Erlbaum Assoc., Hillsdale, NJ

O'Rand A 1996 The precious and the precocious: The cumulation of advantage and disadvantage over the life course. *Gerontologist* **36**: 230–8

Population Reference Bureau 1994 *The United States Population Data Sheet*, 11th edn. Population Reference Bureau, Washington, DC

Preston S 1984 Children and the elderly in the US. *Scientific American* **251**: 44–9

Riley M W, Johnson M E, Foner A 1972 *Aging and Society*. Russell Sage, New York, Vol. III

Riley M W, Kahn R, Foner A 1994 *Age and Structural Lag: Society's Failure to Provide Meaningful Opportunities in Work, Family and Leisure*. Wiley Interscience, New York

Riley M W, Riley J 1999 The aging and society paradigm. In: Bengtson V L, Schaie K W (eds.) *Handbook of Aging Theory*. Springer, New York
Thompson W S, Whelpton K P 1933 *Population Trends in the United States*. McGraw-Hill, New York
Uhlenberg P 1988 The societal significance of cohorts. In: Birren J E, Bengtson V L (eds.) *Emergent Theories of Aging*. Springer, New York
Waring J 1975 Social replenishment and social change. *American Behavioral Scientist* **19**: 237–56

D. Dannefer

Age Structure

Age is a ubiquitous and fundamental ascribed status concept within the social sciences, along with sex to which it is often linked. In spite of this distinction, age structure, the distribution of persons by age in a social unit, is often the subject of neglect as being too obvious to gain serious consideration as a theoretical concept or negative comment as a biological category without social phenomenological content. Nonetheless, consideration of age has given rise to a distinct field of study, gerontology, and remains a crucial determining variable, both distal and proximal, in empirical analyses of a wide range of phenomena and behavioral outcomes (Myers 1996a). It also serves as a defining categorization for collectivities and social groups that gives rise to specialized roles and expectations. Moreover, derived age structures, such as cohorts and generations, have achieved wide attention in studies of societal transformations and life course analyses. In short, age structure is arguably one of the most important concepts in the field of sociology.

An effective way of examining age structure is to consider it as both an outcome and an explanatory mechanism in social science investigations. At the same time, it is useful to distinguish between macro and micro levels.

1. *Macrolevel Determinants*

In the field of demography, age structure is an imbedded concept. Nonetheless, it is surprising to learn that in the eighteenth and nineteenth century there was virtually nothing written about age *per se* or age structure, although considerable attention was devoted to population size and the determinants of population change—fertility, mortality, and migration. In fact, it is not until the beginning of the twentieth century that the Swedish statistician, Gustav Sundbarg (1900), introduced a classification of countries based on the proportions of population under age 15, 15 to 49, and 50 and over. He observed that the proportions in the working ages for a number of European countries appeared to remain constant over time (roughly 50 percent), while the relative share of young and older persons shifted in magnitude from the former to the latter. Thus, he proposed that countries undergo transitions from youthful population structures (that he termed progressive) to stationary and eventually to old structures (regressive), developments largely determined by declining fertility and mortality. This remarkable insight, although subsequently found somewhat inadequate, nonetheless gave important impetus to studies of the determinants of population change, especially mortality and fertility, the possibility of transitions in these vital rates and their systematic impact on population structure, and time-series cross-national research.

Attention to the determinants of age structure benefited from the original mathematical contributions of Lotka in the early twentieth century and later at midcentury in the development of stable population and demographic accounting models by Coale, Bourgeois-Pichat, and others (Myers 1996b). The important role played by the succession of cohorts (usually determined by year/s of birth) over time has been recognized in transforming age structures. This has brought attention to the notion of disordered cohorts, in which catastrophic events (e.g., wars, famines, etc.) have produced large deficits in the numbers of persons at subsequent ages.

2. *Macrolevel Outcomes*

Although the dynamics of fertility, mortality, and migration rates determine the age composition of a population, it is important to note that the actual number of these events depend on age composition interacting with the rates. In this respect, age structure can play an important role in determining the number of births, deaths, and moves in a population. In the process of demographic transitions to lower levels of fertility and mortality, age structures have become older and overall population growth levels have declined or become negative.

Interest in the effects of declining population size emerged in the 1930s in several European countries (most notably the UK and France), but it was not until midcentury that concerted attention was drawn to potential population aging and its societal implications. A notable exception was found in the work of Maurice Halbwachs ([1938] 1960), that elaborated on the notion of social morphology, following the inspiration of his mentor Emile Durkheim. Ironically, it was his compatriot, the noted demographer Alfred Sauvy, who stressed that population aging had grave consequences on the evolution of French culture and national social structure. Scientific examination of how population aging evolves and its broad societal impact gained momentum in the 1950s, especially with the publication by the United Nations (1956) of the volume *The Aging of Populations and Its Economic and*

Social Implications. Today, aging or gerontology has become an important subdisciplinary field within sociology and the other social sciences.

Aggregate age structures have important implications for the institutions in a society—educational (schools, teachers); labor force (demand, productivity, retirement); economic (housing, savings, consumption, income); religious (attendance, volunteerism); and political (voting, government policies). This is true on both the demand and supply side in considering personnel and infrastructure to fulfill institutional functions. For example, Pampel and Stryker (1990) carried on an exchange over the relative importance of changing age structures and state corporatism effects on social welfare expenditures. Their comparative time series study demonstrated the major role that population aging played in overall welfare spending.

3. *Microlevel Perspectives*

The most ambitious treatment of age in sociological thinking can be attributed to Matilda White Riley and her associates (most notably John Riley and Anne Foner) since the early 1970s. In their view, age structure is viewed as one important component of 'age stratification.' Citing important early sociologists, such as Sorokin, Mannheim, Eisenstadt, and Parsons, the team pointed out that age is basically involved in 'group formation and intergroup relations, as a basis for social inequality, and as an intrinsic source of social change as new cohorts because of their particular historical experiences make unique contributions to social structures' (Riley et al. (1988, p. 243). Nonetheless, the basic aspects of age structures involve people and roles stratified by age, which puts a microlevel focus on the individual and normative behavioral characteristics.

This perspective owes a great debt to earlier social anthropologists who emphasized that age groups and age grading are prominent features of many traditional societies. Eisenstadt (1956), in his pioneering work on age groups, observed that age groupings also are important in more complex societies. Indeed, youth, working, and aged groups fulfill important functions in delineating roles, creating group identification, and shaping interactions with other age groups. It is interesting to note that many age groupings formed at younger ages maintain bonding throughout the life course, as is the case with so-called generational groups.

4. *Microlevel Outcomes*

In recent works, Riley and associates place great emphasis on 'structural lags' in which social institutions, such as the family, work, and so forth, have not adapted to collective individual wishes for less rigid age segmented roles (Riley et al. 1994). At the core of this framework is a social psychological perspective in which persons volitionally chose to follow certain behaviors that are in their best interest. The goal is to create a more age-integrated society. In a somewhat similar vein, O'Rand and Henretta (1999) have pointed out that recent processes of age structuring in many industrial societies have produced 'mixed patterns of uniformity and diversity in life course schedules … and decreasing importance of age for the conduct of more and more social roles' (p. 1). The increased variability in the life course, they argue, is associated with increased economic inequality.

Nonetheless, there is strong evidence that formal definitions of age continue to strongly influence behavioral outcomes. For example, expenditures on housing, utilities, and transportation in the USA are strongly affected by the varying consumption behaviors of people at different ages (Pebley 1998). Expenditures for housing rise from nearly $6,000 for persons under 25 to over $12,000 at ages 45–54 and decline to about $7,000 at ages 75 years and over. Expenditures on utilities and transportation vary less, but still follow an inverted U-shaped distribution.

5. *Measurement*

Population pyramids have been widely used to reflect the absolute (numerical) or relative (proportions) age and sex distributions in diverse population, community, and social groups. As a descriptive device, the population pyramid is unparalleled in providing a view of overall population structure at a particular point in time and the magnitude of different age and sex cohorts. Viewed in time series, these displays provide a convenient means of assessing development with regard to demographic transitions and discontinuities in cohort size and composition that can be useful in explaining emerging societal changes.

In multivariate analyses, age is frequently measured as a continuous independent variable (e.g., in ordinary least square regression) or a discrete variable defined by age groups (e.g., in logistic regression). Not uncommonly, the explanatory power of age is very strong and absorbs a considerable amount of variance. Nonetheless, it should be acknowledged that age is a variable that usually reflects other more obtuse or difficult to measure characteristics.

6. *Future Directions*

Several scholars, as we have noted, feel that age structure has or should become less salient in shaping roles and role expectations over the life course. Moreover, some have argued 'the measurement of

age, age structuring, and the life course has become more problematic as the study of human lives has moved away from global images and theoretical categories toward more detailed analyses and explanation' (Settersten and Mayer 1997, p. 234). While heterogeneity, discontinuity, and contingency always exist in considering age structures, it is important to note that age structures are but snapshots of ever-changing distributions. From a societal perspective, however, the continuing extension of life and the concomitant rectangularization and stretching of age distributions suggest that attention to age structures will persist as a major force shaping future sociological research of social structures, the life course, and individual roles and status.

See also: Age, Sociology of; Age Stratification; Aging, Theories of; Generations, Relations Between; Generations, Sociology of; Life Course in History; Life Course: Sociological Aspects; Population Aging: Economic and Social Consequences; Population Cycles and Demographic Behavior; Structure: Social

Bibliography

Eisenstadt S N 1956 *From Generation to Generation: Age Groups and Social Structure*. Free Press, Glencoe, IL

Halbwachs M [1938] 1960 *Population and Society: Introduction to Social Morphology*. Free Press, Glencoe, IL

Myers G C 1996a Aging and the social sciences: Research directions and unresolved issues. In: Binstock R H, George L K (eds.) *Handbook of Aging and the Social Sciences*. 4th edn. Academic Press, San Diego, CA, pp. 1–11

Myers G C 1996b Demography. In: Birren J E (ed.) *Encyclopedia of Gerontology: Age, Aging, and the Aged*. Academic Press, San Diego, CA, pp. 405–13

O'Rand A M, Henretta J C 1999 *Age and Inequality: Diverse Pathways Through Later Life*. Westview Press, Boulder, CL

Pampel F, Stryker R 1990 Age structure, the state, and social welfare spending: A reanalysis. *British Journal of Sociology* **41**: 16–24

Pebley A R 1998 Demography and the environment. *Demography* **35**: 377–89

Riley M W, Foner A, Waring J 1988 Sociology of Age: Society's failure to provide meaningful opportunities in work, family, and leisure. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA, pp. 243–90

Riley M W, Kahn R L, Foner A (eds.) 1994 *Age and Structural Lag*. Wiley, New York

Settersten Jr. R A, Mayer K U 1997 The measurement of age, age structuring, and the life course. *Annual Review of Sociology* **23**: 233–61

Sundbarg G 1900 Sur la repartition de la population par age et sur les taux de mortalite (On the separation of the population by age and the rates of mortality). *Bulletin of the International Institute of Statistics* **12**: 89–94, 99

United Nations 1956 *The Aging of Populations and its Economic and Social Implications*. United Nations, Department of Economic and Social Affairs New York

G. C. Myers

Agenda-setting

Agenda-setting theory develops the observations of Walter Lippmann (1922) in *Public Opinion* that the mass media act as a bridge between 'the world outside and the pictures in our heads.' The central idea is that elements emphasized by the mass media come to be regarded as important by the public. In agenda-setting research, news content is conceptualized as an agenda of items, most frequently an agenda of the major public issues of the day, and agenda-setting theory describes and explains the transfer of salience from this media agenda to the public agenda.

1. Comparing Media and Public Agendas

The media agenda is defined by the pattern of news coverage over several weeks or more, and the public agenda most often is determined by the venerable Gallup Poll question, 'What is the most important problem facing this country today?' First verified during the 1968 US Presidential election, there are now more than 300 empirical studies worldwide documenting agenda-setting effects. These studies have examined the presentation of a wide variety of public issues—and a handful of other objects—by various combinations of newspapers, television, and other mass media and the public response to these media agendas in both election and nonelection settings in Asia, Europe, Australia, and South America, as well as in the USA. Agenda-setting effects also have been produced in controlled laboratory experiments.

The seminal 1968 Chapel Hill study (McCombs and Shaw 1972), which compared the salience of five major issues defining the media agenda with the public agenda among undecided voters, found a near-perfect match in their rank-order ($+0.97$, where the maximum value of this correlation coefficient used to index the strength of agenda-setting effects is $+1.0$). The empirical correlations among general populations are somewhat lower. A year-long study during the 1976 US Presidential campaign found a peak correlation of $+0.63$ between the television agenda and the public agenda during the spring primaries (Weaver et al. 1981). In the 1995 local elections in Pamplona, Spain (McCombs in press), there were substantial matches between the public agenda and the agendas of both local newspapers ($+0.90$ and $+0.72$) and television news ($+0.66$).

2. Explaining Agenda-setting Effects

News reports are a limited portrait of our environment and create a pseudoenvironment to which the public responds. Often there is little correspondence between news coverage and underlying historical trends, in-

cluding rising trends in news coverage and public concern about situations that are unchanged or that actually have improved. These agenda-setting effects of the mass media occur worldwide wherever there are reasonably open political and media systems. Under these circumstances, the public turns to the mass media for orientation on the major issues of the day, especially those issues beyond the ken of personal experience. Even in many cases where personal experience creates high salience for an issue, people turn to the media for additional information and perspective. The concept in agenda-setting theory explaining this behavior is need for orientation, the cognitive equivalent of the physical science principle that nature abhors a vacuum. People are psychologically uncomfortable in unfamiliar situations, such as elections with a plethora of candidates and issues, and frequently turn to the media to satisfy their need for orientation. This psychological concept, which is defined in terms of relevance and uncertainty, explains, for example, the strong agenda-setting effects found in 1968 among Chapel Hill undecided voters. Obviously, both relevance and uncertainty were high for these voters, the condition defining the highest level of need for orientation.

With increased levels of media use, there also is increased agreement about the most important issues of the day among disparate demographic groups, such as men and women or those with high and low education. These patterns of social consensus have been found in Spain, Taiwan, and the USA. Consensus also is facilitated by the limited capacity of the aggregate public agenda. Typically, no more than three to five issues are able individually to garner a constituency of 10 percent or more of the public who regard that single issue as the most important issue of the day, and the public agenda is best characterized as a zero-sum game (McCombs and Bell 1996).

3. *Two Levels of Agenda-setting Effects*

Initially, agenda-setting theory focused on the objects defining the media and public agendas. However, mass media messages about public issues and other objects, such as political candidates, include descriptions of these objects. In abstract terms, objects have attributes. Just as these objects vary in salience, so do the attributes of these objects. When the mass media present an object—and when the public thinks about and talks about an object—some attributes are emphasized. Others are mentioned less frequently, some only in passing. Just as there is an agenda of objects, there is an agenda of attributes for each of these objects. The influence of the media on the relative salience of these objects among the public is the first level of agenda-setting. The influence of the media on the relative salience of these objects' attributes is the second level of agenda-setting.

Images of political leaders among the public afford examples of attribute agenda-setting (McCombs et al. 1997). In the 1994 mayoral election in Taipei, Taiwan, the median value of the comparisons between voters' images of three candidates and news coverage in two major daily newspapers was +0.68. In the 1996 Spanish general election there was substantial correspondence between the news coverage of the major candidates and their images among Pamplona voters. For six comparisons of the voters' images of the three candidates with the coverage in two local newspapers, the median correlation was +0.70. For six comparisons with two national newspapers, the median correlation was +0.81, and for six comparisons with two national TV news services it was +0.52.

Attribute agenda-setting also occurs with public issues (McCombs in press). Some aspects of issues are emphasized in the news and in how people think about and talk about issues. Other aspects are less salient. News coverage in Japanese newspapers about global environmental problems in the months prior to the 1992 United Nations Rio de Janeiro conference resulted in a steady increase in public agreement with the media agenda. By February the match was +0.68 and by April +0.78. A similar pattern was found during a three-week period prior to a local tax election in the USA. Correspondence between the voters' attribute agenda, the relative salience of various aspects of the issue, and the local newspaper's framing of the local tax increased from +0.40 to +0.65. The match with the political advertising on the issue increased from +0.80 to +0.95.

4. *Sources of the Media Agenda*

Although the majority of empirical research on agenda-setting has examined the relationship between the media agenda and the public agenda, scholars also have asked 'Who sets the media agenda?' Influences shaping the media agenda range from the external activities of major news sources to the internal dynamics of the media system (Dearing and Rogers 1996, McCombs and Bell 1996, McCombs in press). Examination of the *New York Times* and *Washington Post* across a 20-year period found that nearly half of the news stories were based substantially on press releases and other direct inputs by news sources, such as press conferences and background briefings. News coverage of Louisiana government agencies was based substantially on information provided by their public information officers to the state's major newspapers. Across an eight-week period the correspondence between the agenda originating with the press information offices and all news stories on those agencies was +0.57.

Political campaigns make a concerted effort to influence the news agenda. In the 1993 British general election, a series of comparisons between the three

major parties' agendas and seven news media, both newspapers and television, found a median correlation of +0.70. American political parties do not fare as well at the national level. A comparison of television news coverage during the 1996 New Hampshire Presidential primary, the inaugural primary in the lengthy US election year, with the candidates' speeches found only a moderate correspondence (+0.40) in their agendas. However, at the local level, in an election for Governor of Texas the combined agendas of the Democrat and Republican candidates shaped the issue agenda of both the local newspaper (+0.64) and the local television stations (+0.52) in the state capital.

The Texas election also reflected intermedia agenda-setting, the influence that one news medium has on another. In Austin, the correspondence between the local newspaper agenda and subsequent television news coverage of public issues was +0.73. A similar comparison in Pamplona, Spain, of two local newspapers with local television news found correlations of +0.66 and +0.70. In the USA, the *New York Times* is regarded as a major agenda-setter among the news media. A case study of the drug issue during the 1980s found that the *New York Times* influenced subsequent coverage by the national television networks, news magazines, and major regional newspapers.

5. *Consequences of Agenda-setting*

The agenda-setting role of the media has consequences beyond the focusing of public attention (McCombs in press). Public opinion during 1992 and 1993 about the overall performance in office by Hong Kong's last British Governor was significantly 'primed' by the pattern of news coverage on his proposals to broaden public participation in local elections. Exposure to this news coverage significantly increased the importance of these proposals in Hong Kong residents' overall approval of the Governor's performance. By calling attention to some matters while ignoring others, the news media influence the criteria by which public officials subsequently are judged, noted Iyengar and Kinder (1987). Priming represents a special case of agenda-setting in which the salience of an issue among the public becomes a significant factor in opinions about a public figure associated with that issue.

The tone of news reports as well as their content can affect subsequent attitudes and behavior. In Germany, shifts in the tone of news stories about Helmut Kohl preceded shifts in public opinion from 1975 to 1984. Daily observations during the final three months of the 1992 and 1996 US Presidential campaigns found that the positive and negative tone of television news about key campaign events influenced voters' opinions about the candidates. The pattern of negative headlines about the US economy over a 13-year period influenced both subsequent measures of consumer sentiment and major statistical measures of the actual economy.

These consequences of agenda-setting for attitudes and opinions require the revision of Bernard Cohen's (1963) seminal observation that the media may not tell us what to think, but are stunningly successful in telling us what to think about. His distinction between affective and cognitive effects of the media was an important precedent for research on first-level agenda-setting effects. At the second level, attribute agenda-setting and its consequences reinvigorate the consideration of media effects on attitudes and opinions. This expanding perspective also is a response to criticism that agenda-setting has focused narrowly on the initial stages of the mass communication and public opinion process.

Agenda-setting theory details a range of effects on the public that result from the mass media's inadvertent focus on a small number of topics and their attributes. To the extent that the news agenda is set by social forces external to the news media, the role of news institutions is important, but neutral as a transmission belt. To the extent that the news media exercise autonomy in defining the public's news diet, they are in themselves a powerful social force.

See also: Agendas: Political; Campaigning: Political; Mass Communication: Empirical Research; Mass Communication: Normative Frameworks; Mass Communication: Technology; Media Imperialism; Media, Uses of; News: General; Political Advertising; Political Communication; Political Discourse

Bibliography

Cohen B C 1963 *The Press and Foreign Policy*. Princeton University Press, Princeton, NJ

Dearing J W, Rogers E M 1996 *Agenda-setting*. Sage, Thousand Oaks, CA

Iyengar S, Kinder D R 1987 *News That Matters: Television and American Opinion*. University of Chicago Press, Chicago

Lippmann W 1922 *Public Opinion*. Harcourt Brace, New York

McCombs M in press *Setting the Agenda: Mass Media and Public Opinion*. Polity Press, Cambridge, UK

McCombs M, Bell T 1996 The agenda-setting role of mass communication. In: Salwen M B, Stacks D W (eds.) *An Integrated Approach to Communication Theory and Research*. Erlbaum, Mahwah, NJ, pp. 93–110

McCombs M E, Shaw D L 1972 The agenda-setting function of mass media. *Public Opinion Quarterly* **69**: 176–87

McCombs M L, Shaw D L, Weaver D (eds.) 1997 *Communication and Democracy: Exploring the Intellectual Frontiers in Agenda-Setting Theory*. Erlbaum, Mahwah, NJ

Shaw D L, McCombs M E (eds.) 1977 *The Emergence of American Political Issues: The Agenda Setting Function of the Press*. West, St. Paul, MN

Wanta W 1997 *The Public and the National Agenda: How People Learn About Important Issues*. Erlbaum, Mahwah, NJ

Weaver D H, Graber D, McCombs M, Eyal C 1981 *Media Agenda Setting in a Presidential Election: Issues, Images and Interest*. Praeger, New York

M. McCombs

Agendas: Political

The political agenda is the set of issues that are the subject of decision making and debate within a given political system at any one time. Significant research specifically on the topic of agenda setting, as opposed to decision making, dates mostly from the 1960s. Early studies of agenda setting were quite controversial because they were often presented as critiques of the pluralist studies of the 1950s and 1960s. Truman (1951) mostly ignored the issue of who set the agenda of political debate. Dahl (1956) discusses the matter in mentioning that ensuring that no group have control over the range of alternatives discussed within the political system is a requisite for democracy. In his study of New Haven he explicitly raises the question of agenda setting, noting that with a permeable political system virtually all significant issues would likely come to the attention of the elites. 'Because of the ease with which the political stratum can be penetrated, whenever dissatisfaction builds up in some segment of the electorate party politicians will probably learn of the discontent and calculate whether it might be converted into a political issue with an electoral pay-off' (Dahl 1961, p. 93). In Dahl's view, then, any issue with a significant potential following in the public would likely find an elite-level champion, though he also notes that issues with no large-scale electoral pay-off might never enter the agenda.

1. *Conflict Expansion*

E. E. Schattschneider (1960) focused attention on how political debates often grow from the conflict of two actors, the more disadvantaged of whom may have an incentive to 'socialize' the conflict to a broader political arena. Of course, the more advantaged disputant strives to 'privatize' the conflict. Schattschneider was one of the first to note that the composition of the political agenda was itself a fundamental part of the political process, and he was the first to give it a prominent role in his view of the political system. By around 1960, then, scholars had firmly noted the importance of the study of the political agenda as an important area of research.

After the critique of Schattschneider (1960), scholars were less willing to take the composition of the agenda for granted. Peter Bachrach and Morton Baratz (1962) provided one of the most telling critiques of pluralism when they noted that studies of decision-making, power, and influence were misleading. Their aptly titled article, 'The two faces of power,' noted that the 'first face' of power, the authority to choose between alternatives, may be less important than the 'second face' of power, the ability to control what alternatives are under discussion in the first place. Whereas Dahl and others saw this as a relatively open process, where any social group with a legitimate problem that could potentially be converted into votes in an election could gain access to the political agenda, others saw the process in a decidedly more negative light. Following Bachrach and Baratz, many scholars attempted to study not just governmental decision making, as the pluralists had done, but also non-decisions, or agenda control, as well. For example, Matthew Crenson (1971) noted that air pollution was rarely discussed in public or government in one city despite a very serious pollution problem. In another similar city with much less pollution, however, public and governmental leaders discussed it often and took steps to combat it. The reason behind the difference in the behavior between the two cities appeared to be the ability of powerful economic interests to control the agenda. John Gaventa (1980) followed this study with an analysis of poverty-stricken Appalachian towns and the 'quiescence' characterizing the demobilized populations there. These agenda theorists argued that power was most evident when objective conditions of suffering were not the subject of debate. Bachrach and Baratz (1962), Crenson (1971), and Gaventa (1980) raised important issues and directly challenged the relatively optimistic views of the pluralists but did not convince, all because of the difficulty of discerning exactly what would be a neutral political agenda. In other words, it was hard to know what findings would demonstrate elite control and what findings would demonstrate democratic openness; in this situation two scholars looking at the same findings could disagree forever (and they did; see Baumgartner and Leech 1998, chap. 3, for a discussion of these issues relating to the community power studies of the 1950s and 1960s; see also Polsby's (1980) treatment of these methodological issues).

2. *The Development of a Literature*

Roger Cobb and Charles Elder (1972), in the first book-length treatment of the political agenda, noted the difference between the systemic agenda, defined as the group of issues that were under discussion in society, and the institutional agenda, or the set of issues being discussed in a particular government institution (see also Cobb et al. 1976). Since then, scholars have variously written about the public agenda, the media agenda, the legislative agenda, and any number of other agendas as they have focused on different political institutions.

More recent studies of agenda setting have moved away from the concepts of nondecisions and power because of the difficulties inherent in designing rigorous research on the topic. Instead, scholars have

focused on the rise and fall of issues on the public or institutional agendas and how decision making during high salience periods differs from the more routine decision making that takes place when an issue is low on an agenda. Jack Walker (1977) provided one of the first statistically based studies in the area with his analysis of the US Senate's agenda. He noted that issues often rose on the Senate's agenda following heightened levels of discussion within professional communities.

John Kingdon's (1984) treatment of the public agenda set the stage for much of our current understanding of where issues come from. He emphasized the separate sources of policy problems from the solutions that may be offered to them. Government programs, he noted, come about when a given solution is attached to a particular problem, and his analysis of health care and transportation policies in the USA showed just how unpredictable these couplings can be. Political actors' search for popular issues, windows of opportunity open and close, stochastic events such as natural disasters or airplane crashes momentarily focus public attention on an issue. The confluence of many unrelated factors, often serendipitous, helps explain why a given policy is adopted, according to his study. Kingdon's (1984) was the first major book-length study on the topic since Cobb and Elder's (1972), and it was based on hundreds of interviews with government and other policymakers in the 1970s and 1980s. (Polsby 1984 also reached many of these conclusion in a book appearing in the same year as Kingdon's.)

Frank Baumgartner and Bryan Jones (1993) provided the next major treatment of political agendas in their analysis of nine different policy areas over a 40-year period. Utilizing publicly available sources such as media indices and records of congressional hearings, they noted how particular issues rose and fell on the agenda over the entire post-World War II period. They developed a punctuated equilibrium model of policy change in which episodic periods of high agenda status typically were related to dramatic and long-lasting policy changes. During these high-salience periods, institutional procedures were often created or altered. The subsequent ebbing of the issue from the public agenda enabled the newly empowered political institutions and policymakers to settle into stable routines of behavior persisting for decades at a time. Agenda setting was related to dramatic changes, often upsetting long-standing routines of behavior and power by replacing them with new ones.

3. *Issue Definition*

Studies of agenda setting have often focused on the question of issue definition. Echoing a major theme in Baumgartner and Jones (1993), David Rochefort and Roger Cobb (1994) brought together a number of essays showing how public understanding and media discussion of a given issue can change over time, often quite dramatically. Deborah Stone (1988) also discussed this in her analysis of 'causal stories.' Policy entrepreneurs frame issues by explaining the causes of a given problem with a narrative justifying a particular governmental response. Book-length studies of the issues of child abuse (Nelson 1984), pesticides (Bosso 1987), health care reform (Hacker 1997), and various natural and human-made disasters (Birkland 1997) have shown the impact of changing issue definitions and of focusing events in pushing an issue on to the public agenda. Roger Cobb and Marc Howard Ross (1997) brought together a series of essays on the rarely studied topic of 'agenda denial,' whereby political actors keep threatening issues off the agenda.

William Riker (1986, 1988, 1993, 1996) showed the importance of two related issues: the ability of strategically minded politicians to alter the terms of debate by skillfully manipulating issue definitions, and the power of formal agenda control. A voluminous literature in formal and game theory suggests that the controller of a formal agenda can affect the outcomes in a voting situation by altering the order in which alternatives are considered. Riker used game theory to illustrate how formal agenda control can affect such things as votes in a parliamentary setting, and case studies and historical illustrations to show how political leadership could be even more powerful through the means of altering issue definitions. Political leaders can utilize a combination of formal agenda control and informal debating skills to achieve their ends, according to Riker.

4. *Social Movements and the Media*

A number of scholars have noted that social movements have often successfully brought new issues onto the public agenda. Thomas Rochon's (1998) analysis of the peace movement in various Western countries fits in this tradition, as does the work of Douglas McAdam (1988), whose study of the Mississippi Freedom Summer documented the success of civil rights activists in putting the issue of racial equality on the national political agenda during the mid-1960s.

Studies of the media agenda have been legion, largely following from the early work of Max McCombs and Donald Shaw (1972); for a review of this literature, see Rogers and Dearing (1988). Bernard Cohen (1963) noted famously that while the media cannot tell the public what to think, they can have a great impact on what the public think about. Within political science, several authors have picked up on the issue of media effects on public opinion (Iyengar 1991, Iyengar and Kinder 1987). James Stimson (1991) noted the changes in a broadly measured national mood based on public opinion surveys; John Kingdon (1984) also put considerable emphasis on the national mood in his study of agenda setting in government. As

policymakers consider what issues to spend their time on, Kingdon (1984) noted they often make reference to the idea of a national mood.

Studies of the political agenda have been remarkable in political science for their integrative character: rather than focusing on any particular institution of government, scholars have traced the sources of agenda setting in the public, in the roles of interest groups and social movements, by noting the roles of policy entrepreneurs, and by looking at the government in very broad terms. Of course this does not mean that political leaders play an insignificant role. From the work of Richard Neustadt (1960) onwards students of the US Presidency have noted the need for presidents to focus their energy on a few issues (see Light 1982; for a similar study of congressional leadership see Bader 1996). Studies of the Supreme Court have noted the extremely tight control that the Court maintains over its agenda, as well as the characteristics of the cases that it is most likely to take. The Court, of course, is unusual among political institutions in that its agenda is reactive rather than proactive. Congress or the President can reach out to discuss whatever issues appeal to them; the Court can only choose from the issues that are presented for its decision (see Perry 1984, Caldeira and Wright 1988).

5. *Conclusion*

In sharp contrast to two generations ago, research on political agendas is vibrant and promising today. Though much of the work has been done within the context of US politics, comparative studies have become more common (see Hogwood 1987, Baumgartner 1989, Reich 1991, Zahariadis 1995, John 1998). New sources of quantitative data on public attitudes, government archives, and media coverage promise more systematic studies covering a greater range of issues over a longer time period than was typically possible in the past. Studies of political agendas are now firmly established as an important part of the field of political science now some 40 years after the concept was first discussed.

See also: Community Power Structure; Issue Evolution in Political Science; Power: Political; Utility and Subjective Probability: Empirical Studies

Bibliography

Bachrach P, Baratz M 1962 The two faces of power. *American Political Science Review* **56**: 947–52

Bader J B 1996 *Taking the Initiative: Leadership Agendas in Congress and the 'Contract with America'*. Georgetown University Press, Washington, DC

Baumgartner F R 1989 *Conflict and Rhetoric in French Policymaking*. University of Pittsburgh Press, Pittsburgh, PA

Baumgartner F R, Jones B D 1993 *Agendas and Instability in American Politics*. University of Chicago Press, Chicago

Baumgartner F R, Leech B L 1998 *Basic Interests: The Importance of Groups in Politics and in Political Science*. Princeton University Press, Princeton, NJ

Birkland T A 1997 *After Disaster: Agenda Setting, Public Policy, and Focusing Events*. Georgetown University Press, Washington, DC

Bosso C J 1987 *Pesticides and Politics: The Life Cycle of a Public Issue*. University of Pittsburgh Press, Pittsburgh, PA

Caldeira G A, Wright J R 1988 Organized interests and agenda-setting in the U.S. Supreme Court. *American Political Science Review* **82**: 1109–27

Cobb R W, Elder C D 1972 *Participation in American Politics: The Dynamics of Agenda Building*. Allyn and Bacon, Boston

Cobb R W, Ross J-K, Ross M H 1976 Agenda building as a comparative political process. *American Political Science Review* **70**: 126–38

Cobb R W, Ross M H (eds.) 1997 *Cultural Strategies of Agenda Denial*. University Press of Kansas, Lawrence, KS

Cohen B C 1963 *The Press and Foreign Policy*. Princeton University Press, Princeton, NJ

Crenson M A 1971 *The Un-politics of Air Pollution*. The Johns Hopkins University Press, Baltimore, MD

Dahl R A 1956 *A Preface to Democratic Theory*. University of Chicago Press, Chicago

Dahl R A 1961 *Who Governs?* Yale University Press, New Haven, CT

Gaventa J 1980 *Power and Powerlessness: Quiescence and Rebellion in an Appalachian Valley*. University of Illinois Press, Urbana, IL

Hacker J S 1997 *The Road to Nowhere: The Genesis of President Clinton's Plan for Health Security*. Princeton University Press, Princeton, NJ

Hogwood B W 1987 *From Crisis to Complacency? Shaping Public Policy in Britain*. Oxford University Press, New York

Iyengar S 1991 *Is Anyone Responsible? How Television Frames Political Issues*. University of Chicago Press, Chicago

Iyengar S, Kinder D R 1987 *News that Matters: Television and American Opinion*. University of Chicago Press, Chicago

John P 1998 *Analyzing Public Policy*. Pinter, London

Kingdon J W 1984 *Agendas, Alternatives, and Public Policies*. Little, Brown, Boston

Light P C 1982 *The President's Agenda*. The Johns Hopkins University Press, Baltimore, MD

McAdam D 1988 *Freedom Summer*. Oxford University Press, New York

McCombs M E, Shaw D L 1972 The agenda-setting function of the mass media. *Public Opinion Quarterly* **36**: 176–87

Nelson B J 1984 *Making an Issue of Child Abuse*. University of Chicago Press, Chicago

Neustadt R E 1960 *Presidential Power*. John Wiley and Sons, New York

Perry Jr H W 1984 *Deciding to Decide: Agenda-Setting on the US Supreme Court*. Harvard University Press, Cambridge, MA

Polsby N W 1980 *Community Power and Political Theory*, 2nd edn. Yale University Press, New Haven, CT

Polsby N W 1984 *Political Innovation in America: The Politics of Policy Initiation*. Yale University Press, New Haven, CT

Reich M R 1991 *Toxic Politics: Responding to Chemical Disasters*. Cornell University Press, Ithaca, NY

Riker W H 1986 *The Art of Political Manipulation*. Yale University Press, New Haven, CT

Riker W H 1988 *Liberalism Against Populism*. Waveland Press, Prospect Heights, IL

Riker W H (ed.) 1993 *Agenda Formation*. The University of Michigan Press, Ann Arbor, MI
Riker W H 1996 *The Strategy of Rhetoric*. Yale University Press, New Haven, CT
Rochefort D W, Cobb R W (eds.) 1994 *The Politics of Problem Definition: Shaping the Policy Agenda*. University Press of Kansas, Lawrence, KS
Rochon T R 1998 *Culture Moves*. Princeton University Press, Princeton, NJ
Rogers E M, Dearing J W 1988 Agenda-setting research: Where has it been, where is it going? In: Anderson J A (ed.) *Communication Yearbook 11,* Sage, Newbury Park, CA, pp. 555–94
Schattschneider E E 1960 *The Semi-Sovereign People*. Holt, Rinehart and Winston, New York
Stimson J A 1991 *Public Opinion in America: Moods, Cycles, and Swings*. Westview Press, Boulder, CO
Stone D A 1988 *Policy Paradox and Political Reason*. Scott, Foresman, Glenview, IL
Truman D B 1951 *The Governmental Process: Political Interests and Public Opinion,* 1st edn. Alfred A. Knopf, New York
Walker J 1977 Setting the agenda in the U.S. Senate. *British Journal of Political Science* **7**: 423–45
Zahariadis N 1995 *Markets, States and Public Policy: Privatization in Britain and France*. University of Michigan Press, Ann Arbor, MI

F. R. Baumgartner

Aggregation: Methodology

Aggregation is a technique that is utilized in various disciplines in the social sciences. A basic definition of aggregation is combining data from members or subordinate units of a larger, superordinate category in order to describe the superordinate category. In the social sciences aggregation typically involves obtaining data from or about individuals and combining these data into a summary statistic that would serve to characterize a larger, well-defined, socially meaningful unit that contains a large number of individuals. This summary statistic may then be used as a data point in a data set consisting of larger units for comparative purposes. Common examples of larger units with multiple members involve a social group, an organization, or geographical or administrative units—a census tract, a county, a school district, a city, or a country. Information collected from individuals is called 'individual-level' data; when these data are aggregated statistically to describe the superordinate category, the resulting data are at the 'superordinate-level' or 'aggregate-level' and called 'aggregate data' or 'aggregated data.'

When individuals are nested (i.e., located) under intact and meaningful units, a 'nested structure' is obtained. A nested structure may involve multiple levels: individuals may be nested in classrooms and classrooms may be nested under schools, schools may be nested under school districts, and so on. When there are multiple levels and the nesting is clear and hierarchical, a 'hierarchical multi-level model' is obtained.

A typical example for the process and significance of aggregation is the census where often detailed information is collected from individuals and households, and this information is used to describe census tracts, counties, zones, cities, regions, and so on. Such information is clearly important—that is why so much money and effort are put into conducting censuses all around the world—and important social policy decisions are often based on such aggregated data. More funding may be provided, for instance, for job opportunity programs in cities where unemployment rates are high.

However, census data are also a good example to illustrate the basic limitations of aggregated data. For multiple reasons, particularly for protecting privacy, census data about individuals and households are never disclosed. Instead, data about city blocks or census tracts (defined by the US Census Bureau as a group of city blocks having a total population of more than 4,000 people) are made available to the public. Statements about average household size (e.g., a household has 4.5 members on average), average number of children (e.g., an average family has 1.5 children), average number of cars, average number of jobs worked in a calendar year, etc. are clearly not about an actual household (a household cannot have 4.5 members) or an actual family (a family cannot have 1.5 children). Aggregated data describe average households or families—the typical patterns in a given census unit—but never an actual household or family. With aggregation information about actual households and the heterogeneity they may present is lost. Aggregation is often useful to characterize superordinate categories—that is, working upwards from individuals to larger social units. The reverse, however, is not true: working backwards from aggregated data to subordinate units can be very misleading.

1. *The Technique*

Aggregation is a technique that cuts across disciplines, but is most commonly utilized in disciplines that deal with collective systems, such as groups, neighborhoods, schools, markets, or organizations. Aggregation is less common in disciplines that focus on individual human beings. Aggregation is particularly important in disciplines that deal with issues where individual-level data cannot be disclosed, as in voting behavior or in household income, and data are collected or made available to the public at the aggregate level. In other cases, collecting individual-level data may be particularly difficult, time-consuming or even impossible. For instance, in criminal justice research, a researcher may have data reported by the

police department about specific crime rates in a given, well-defined area but the perpetrators of these crimes may not be known and collecting data about the motives of the perpetrators may be almost impossible.

Aggregation is not limited to the social sciences. In daily life, too, aggregation is very common. It is commonly used in various areas of life primarily for the practicality of aggregated data. For instance, households that use natural gas for cooking and heating are often provided with no information regarding daily consumption rates. Rather, monthly consumption is reported and billed. For detailed information about gas usage, information regarding daily consumption would have to be monitored and recorded. In most cases, neither the consumers nor the utility industry would be interested in such detailed information. Rather, a summary measure, total amount of gas used per month, which is based on data aggregated over time, is sufficient. Similarly most people do not monitor their calorie intake until they start a diet regimen. In daily life an aggregate perception of how much food is consumed would be sufficient. When the goal is to lose weight—that is, intervene and change the current calorie intake—each source of calorie intake may have to be monitored.

From a practical standpoint, aggregated data are very desirable. Except for those who are interested in understanding the nature of the process, such as scientists, and those who would like to intervene and change the current situation, such as policy makers and reformers, most people would be interested in what the average is or whether the average is at a desired level. For instance, when data on how much fuel it takes each car to travel a given distance is available, almost everyone would prefer to know the average fuel consumption—rather than a list of how much fuel it took each car to travel the given distance.

Basic types of aggregation or aggregate data are very easy to understand and easy to come across in daily life. Total amount of car accidents in a given geographical area per day, the average number of cars crossing a busy intersection per day, the percentange of car crashes that result in fatalities per day are such examples of aggregation. All one needs to do is to count, to take an average or to calculate the percentage of the event of interest. This is called aggregation 'across units.' For events that may show fluctuation or systematic variation over time, it is very important to specify the time period. Sometimes it may be useful to calculate an average across time to obtain more reliable estimates. If the number of cars crossing a busy intersection varies from day to day and the goal is to identify a good estimate of the average number, it would be useful to count the totals per day and then average across five days of the working week. This type of aggregation is called aggregation 'over time.' It is obvious that, depending on the question at hand, aggregation can involve aggregating data across individuals within a larger unit (e.g., a school), schools within a larger unit (e.g., a school district), districts within a large unit (e.g., a city). When time is important, it may be desirable to aggregate data across time points within a time period (e.g., a year) or to even aggregate across time points within an individual, as in monitoring health status over several months.

2. *Problems Associated with the use of Aggregated Data*

When data from individuals are obtained and aggregated to describe a socially meaningful unit, under which these individuals are nested, aggregation is often not problematic. When, however, aggregated data are used to analyze larger units, problems may arise. The most serious problem associated with aggregated data involves generalizing relationships at the aggregate level to individuals as though these relationships necessarily hold at the individual level.

Robinson (1950) is often credited with the discovery of a fundamental problem with applications of aggregation in the social sciences: the behavior of an aggregate often gives no clues to the behavior of an individual belonging to that aggregate. A relationship at the aggregate level (what Robinson called an 'ecological correlation') between two variables (e.g., crime and unemployment) does not reliably lead to an association between these two variables at the individual level (e.g., committing a crime and being unemployed). This is known as the 'ecological fallacy,' which involves deducing individual behavior from the the behavior of aggregates (Weisberg et al. 1996).

An oft-cited example comes from the political science literature, where individual-level data concerning voting behavior are rarely available. The question is simple: If the election results indicate that a voting district that consists of 90 percent ethnic minority voters and 10 percent ethnic majority voters voted 90 percent for Party A, and Party A is known to be generally supported by this ethnic minority, could we safely conclude that each minority voter actually voted for Party A—based on the aggregate data available? It may be tempting to do so, but the answer is negative. It may be the case that very few voters from the minority voted and every member of the majority voted for Party A. It is obvious from this example that a perfect match between numbers (90 percent minority—90 percent vote for Party A) can be very misleading. At the least, other data, if available, should be considered: What percentage of the minority and the majority voted? Is there any individual level data available that can shed light on this issue?

Drawing 'ecological inferences,' which involves arriving at conclusions about individual behavior using aggregate data reported for the superordinate category, is considered a very unreliable method in the social sciences. It has been noted that this problem, the

'ecological inference problem,' hinders substantive work in almost every empirical field of political science and in other disciplines where there is restricted access to individual-level data. Attempts to make ecological inferences more reliable (e.g., King 1997) have been met with strong suspicion. In general none of these attempts have yielded a satisfactory solution. Today the ecological inference problem remains a bottleneck for disciplines that have limited access to individual-level data.

In psychology and related disciplines, there is a second, historically common and much less problematic application of aggregation. This type of aggregation involves combining data across situations or time points to obtain more reliable data. This is known to reduce measurement error, which consists of random fluctuations in data. Random fluctuations often result from imperfect measurement tools, hence the name measurement error. Measurement error creates 'noise' in the data and makes it difficult to observe the true relationships. Reducing measurement error via aggregation, for instance over time, may increase the strength of the relationship (e.g., Harris 1998). In these applications, no ecological fallacy is involved. The observations are being aggregated across situations or time to describe the superordinate unit: in psychology, this unit is often the person and multiple observations are aggregated to better describe this person. This type of aggregation is suitable for situations where the process under observation is not changing, the fluctuations are not immense and the data are relatively homogenous and the fluctuations are due, not to change in the person's behavior, but to measurement error.

Aggregating data to reduce measurement error, however, should not be taken as a shortcut to reliable data and hence reliable relationships. Reliable measures are the key to capturing reliable relationships and aggregation cannot compensate for unreliable measurement. When aggregation involves aggregating observations that are very different, aggregation will yield a misleading picture: the picture will consist of an average that simply does not exist. For instance, when aggression is measured across situations and in some situations environmental factors (e.g., police presence) hinder the display of aggression, the amount of aggression will vary widely across situations. It is better to pick situations that are similar in most respects and aggregate across situations only to reduce random errors (e.g., videotape coding errors). Therefore, the long-standing notion that the use of aggregated data necessarily yields stronger relationships is not true (Ostroff 1993).

A more serious problem that arises when data are aggregated across time points, situations, or persons is the loss of information on variability or heterogeneity. When data are aggregated to obtain an average across observations, the result is a 'homogeneity bias': the average that is obtained contains no information about the degree of heterogeneity in the original data set. The homogeneity bias is a serious consideration when aggregation is done both within-persons and across-persons. When a person's behavior varies across situations, this variability may indicate underlying relationships rather than measurement error and the origins of variability across situations needs to be carefully considered.

The homogeneity bias is most serious when aggregation involves combining data across persons nested under a larger group. First, individual differences are pervasive and often quite large in almost all types of behavior that social scientists are interested in. Therefore, some degree of heterogeneity is inevitable and aggregation loses this information. When the heterogeneity is extreme, the aggregate picture (e.g., the average of the group) may be meaningless.

Second the degree of heterogeneity within a unit offers a window to social phenomena at a descriptive level. For example, neighborhoods are often assumed to be homogenous and when only aggregated data, such as census data are available, this assumption cannot be challenged. When data from individuals are obtained and within-unit heterogeneity is examined using intraclass correlations, which indicate how homogenous individuals are within a unit, neighborhoods appear to be much less homogenous than they were assumed to be (Cook et al. 1997). Finally, individual differences within units may be just what needs explaining. If, for instance, neighborhoods are theoretically powerful influences on residents and yet descriptive evidence indicates that a large degree of heterogeneity within neighborhoods, this heterogeneity needs to be studied. More explicitly, a neighborhood characteristic that is common to all neighborhood residents cannot explain why the residents are varying on a given outcome. What leads to this heterogeneity has to be investigated by focusing on variables not common to all residents. Perhaps the best example for the significance of heterogeneity comes from developmental psychology: a century of research shows that parents have powerful influences on their children and yet heterogeneity within families is always present. Siblings, even genetically identical monozygotic twins, appear to be different in multiple respects. Developmental psychologists are now considering processes that make siblings similar and different simultaneously.

3. *Current Status*

Despite the potential problems associated with the use of aggregated data and the inherent limitations of aggregated data in offering reliable insights into individual's behavior, aggregation is common in the social sciences today and there are advances in the proper utilization of the aggregation and aggregated data. These advances are due primarily to two trends:

As interdisciplinary research becomes more common in the social sciences, it becomes necessary to specify relationships and use data from more than one level. Second, advances in statistical modeling of data from multiple levels, particularly multilevel modeling, allows researchers to utilize statistical tools that match the complexity involved in multilevel research.

There is increased sensitivity in textbooks to problems associated with aggregation, often under sections dealing with the issue of the 'unit of analysis' (e.g., Singleton et al. 1988). With the wide acceptance of structural equation modeling in the social sciences, researchers are urged to specify the theoretical models and the specific relationships these models contain both in their theoretical work, at the outset of their empirical investigations and at the time of statistical analysis. This specificity facilitates specification of the unit of analysis (e.g, individuals, schools, and neighborhoods) and at which level (e.g., individual-level, school-level, and neighborhood-level) data should be collected and analyzed.

There are multiple examples in the recent literature that demonstrate how interdisciplinary research in the social sciences may necessitate working at multiple levels and the proper use of aggregation and aggregated data. When, for instance, the issue is how school reform influences change in a number of measures tapping different aspects of students' lives (e.g., Cook et al. 1999), the question is how a school-level variable (i.e., school reform) influences individual-level variables. This research brings two foci together: the school as an institution, often the focus sociology of education, and individual change, often the focus of developmental psychology. In this study, some variables are measured at the school level (e.g., school size) and outcome measures (e.g., academic performance) are measured at the student level. Several variables are measured at the individual level and then aggregated to the school level to characterize the school. For example, whether or not a student is taking algebra is determined at the individual level (yes or no), and then aggregated to the school level (proportion of students enrolled in algebra class). Such research that brings together two levels is much needed in areas where the construct of interest is collective and the individuals living in or exposed to this collective variable are expected to be influenced by it. Collective climate or organizational culture are prototypical examples of such constructs. By definition, climate or culture is collective and is expected to influence and to be influenced by, individuals living in that culture. Therefore, the issue is a multilevel issue by definition and issues of level in is a key question in organizational research (Morgeson 1999, Rousseau 1985). Often researchers would aggregate individual-level data to be able to capture and describe the collective climate (e.g., Gonzales-Roma et al. 1999) but the richer picture emerges when multilevel analysis is used and both levels are taken into account (Klein 1999).

Multilevel analysis of multilevel questions tends to produce more accurate findings (Bryk and Raudenbush 1992). Multilevel analysis avoids the heterogeneity bias aggregated data often lead to by explicitly allowing for and modeling within-unit heterogeneity. In sociology of neighborhoods, for instance, neighborhoods have often been construed as powerful influences on neighborhood residents. The evidence that supported this view was often based on aggregate data and aggregate relationships between neighborhood variables and variables measured at the individual level and aggregated to the neighborhood level. Recent research employing a two-level model (e.g., Cook et al. 1997) suggests that neighborhoods may have a much smaller influence on individual-level outcomes. When individual-level data are aggregated to the neighborhood level, as was the case before, neighborhoods appear to have stronger influences. Recent research on peer influence suggests a similar picture: when an individual is located in a network of relationships and the multiple levels involved in the network are taken into consideration, long-standing estimates of peer influence appear to be inflated (Urberg et al. 1997).

4. *Future Prospects*

There is a consensus in the social sciences that aggregated data are useful and need not be left behind as a research tool. Aggregated data are widely available through public data collection efforts (e.g., census) and easily accessible at archives, many of which are now online. The increase in the number of interdisciplinary collaborations are making it increasingly more commonplace to use individual-level and aggregate data together. The advances in multilevel modeling are enabling interdisciplinary teams to properly deal with the complexities involved in multilevel data analysis and there is a consensus that multilevel modeling is necessary to overcome the inherent limitations of aggregated data (Jones and Duncan 1998, Klein 1999, Morgeson 1999). These two trends are making it inevitable for researchers to obtain data from individuals and analyze such data at the individual level.

Thus, a third consensus is emerging around the necessity of collecting data from individuals when aggregate-level factors (e.g., economic stagnation) are expected to influence individuals (e.g., depression, job prospects, etc.). Without individual-level data, such relationships cannot be properly examined.

This consensus is bolstered by process-oriented research in the social sciences. Process-oriented research focuses on the linkages or the mediating processes that explain how an outcome may be influenced by various factors. Such research is particularly needed when interventions are necessary to influence the processes leading to negative outcomes. When an intervention is required, aggregated data

offer little or no direction, because most interventions need to target specific individuals and processes. This is particularly true for multilevel models: when global influences, such as economic stagnation, are under investigation, it is often clear from the very outset that global influences do not influence each individual in the same way and different segments of the society experience this influence in different ways. How this influence varies across individuals and social groups needs to be explained with mediating variables and moderating variables. The moderating mechanisms often involve a cross-level interaction between global factors and personal factors (e.g., depression, job prospects, etc.). Without individual-level data, such relationships cannot be properly modeled (Furstenberg et al., Sameroff et al.)

See also: Demographic Data Regimes; Ecological Fallacy, Statistics of; Ecological Inference; Statistical Systems: Censuses of Population

Bibliography

Bryk A S, Raudenbush S 1992 *Hierarchical Linear Models*. Sage, Newbury Park, CA

Cook T D, Habib F, Phillips M, Setterstein R A, Shagle S C, Değirmencioğlu S M 1999 Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal* **36**: 543–97

Cook T D, Shagle S C, Değirmencioğlu S M 1997 Capturing social process for testing mediational models of neighborhood effects. In: Brooks-Gunn J, Duncan G J, Aber J L (eds.) *Neighborhood Poverty, Volume II: Policy Implications in Studying Neighborhoods*. Russell-Sage, New York, pp. 94–119

Gonzales-Roma V et al. 1999 The validity of collective climates. *Journal of Occupational and Organizational Psychology* **72**(1): 25–41

Harris M M, Gilbreath B, Sunday J A 1998 A longitudinal examination of a merit pay system: Relationships among performance ratings, merit increases, and total pay increases. *Journal of Applied Psychology* **83**(5): 825–31

Jones K, Duncan C 1998 Modelling context and heterogeneity: Applying multilevel models. In: Scarbrough E, Tanenbaum E (eds.) *Research Strategies in the Social Sciences: A Guide To New Approaches*. Oxford University Press, Oxford, UK

King G 1997 *A Solution to the Ecological Inference Problem*. Princeton University Press, Princeton, NJ

Klein K J 1999 Multilevel theory building: Benefits, barriers, and new developments. *Academy of Management Review* **24**(2): 248–54

Morgeson F P 1999 The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review* **24**(2): 249–66

Ostroff C 1993 Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology* **78**(4): 569–82

Robinson W S 1950 Ecological correlations and the behavior of individuals. *American Sociological Review* **15**: 351–7

Rousseau D M 1985 Issues of level in organizational research: Multilevel and cross-level perspectives. In: Cummings L L, Staw B M (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, pp. 1–37

Singleton R Jr, Straits B C, Straits M M, McAllister R J 1988 *Approaches to Social Research*. Oxford University Press, New York

Urberg K A, Değirmencioğlu S M, Pilgrim C 1997 Close friend and group influence on adolescent substance use. *Developmental Psychology* **33**: 834–44

Weisberg H F, Krosnick J A, Bowen B D 1996 *An Introduction to Survey Research, Polling, and Data Analysis*, 3rd edn. Sage Publications, Thousand Oaks, CA

S. M. Değirmencioğlu

Aggression in Adulthood, Psychology of

1. *Approaches to the Study of Aggression*

Psychological analyses of adult aggression have changed over the twentieth century with the development of the behavioral, biological, and social sciences. Although nearly all of these interpretations view aggression as behavior intended to hurt or destroy some target, the early formulations, including those advanced by traditional psychoanalysts, basically attributed the action largely to endogenous motivation: an internally generated drive continuously seeking expression that supposedly had to be released directly in aggression or indirectly in substitute behavior (see Berkowitz 1993). Contemporary analyses are far more differentiated and recognize the interplay of a large variety of influences in the person's biology (including heredity), past learning, and immediate social context (see Berkowitz 1993, Reiss and Roth 1993). Even so, however, present-day psychological accounts of adult aggression typically concentrate on the psychological processes operating in the instigating situation-to-behavior sequence, although they differ in which aspects of this sequence they emphasize. Some focus on the action's goals, whereas others deal primarily with the cognitive processes promoting the behavior.

Many discussions concerned with the aggressors' goals assume the attackers are mainly motivated to achieve some end other than the victim's injury, such as achieving control or dominance over the victim (e.g., Tedeschi and Felson 1994), status attainment, the repair or enhancement of one's self-concept (e.g., Nisbett and Cohen 1996, Toch 1992), or more generally, the elimination of a noxious state of affairs (Bandura 1973). In contrast, other analyses (e.g., Berkowitz 1993) contend that sometimes the aggressors' primary goal is to harm or destroy. This latter formulation distinguishes between hostile assaults, aimed chiefly at hurting the victim (whatever other benefits might also be achieved), and instrumental aggression that is used as a means to some other, noninjurious objective. As a variation on this theme,

other investigators (e.g., Crick and Dodge 1996) differentiate between reactive and proactive aggression, with the former being a response to a perceived threat and the latter spurred by the anticipation of some gain. Reactive aggression can be regarded as hostile aggression in that both frequently occur in response to externally engendered strong negative feelings (Berkowitz 1993). And similarly, proactive aggression can be seen as instrumental aggression.

Where some discussions of hostile aggression (e.g., Berkowitz 1993) hold that quite a few assaults of this type are carried out impulsively and with little thought, the formulations emphasizing mental processes (e.g., Lindsay and Anderson 2000, Zelli et al. 1999) generally assume that the attackers' aggression-related knowledge (or belief?) structures and modes of information processing shape their decision to assault the target. According to Zelli et al. (1999), those people who believe it is proper to assault a perceived offender are also apt to make hostile attributions about ambiguous interactions with others, and these attributions then determine what behavior is enacted.

People's appraisals of an aversive situation undoubtedly can affect what they feel and do in response. Although emotion researchers do not agree in detail as to what specific interpretations produce anger and aggression, most appraisal theories insist these particular emotional reactions will not arise unless some external agent is blamed, that is, accused of having deliberately and improperly brought about the negative event. However, there is now evidence that blame appraisals are not always necessary for anger and aggression to occur, and more than this, that anger generated by irrelevant negative experiences can lead to blame being placed on innocent parties (Berkowitz 1993, 2001).

Another cognitive process promoting aggression operates through priming: Events or objects having an aggressive meaning automatically bring to mind a range of aggression-related thoughts and may even activate aggression-related motor reactions. Like other hostile acts, the primed behavior is aimed at the injury of the available target, but unlike most instances of hostile aggression, it is not spurred by intense negative affect.

2. *Internal Influences on Aggression*

2.1 *Violence-prone Personalities*

Even though every attack is not necessarily governed by the same underlying processes, those persons who are highly assaultive in some settings are apt to be also aggressive on other occasions as well, even though their actions may vary in form and target (e.g., Olweus 1979). Research has also shown that violence-prone adults are likely to have been hyperactive, impulsive, and restless as children (Reiss and Roth 1993). Consistent with these findings, Dodge and his colleagues (Zelli et al. 1999), as well as other investigators, such as Spielberger, using aggressive trait inventories (see Berkowitz 1998), indicate that many violence-prone individuals are highly reactive emotionally. These persons are also generally quick to attribute hostile intentions to others, and often react to these perceived threats with intense anger. Their aggressive urge is also apt to be facilitated by their beliefs that aggression is an appropriate and effective way to resolve their interpersonal difficulties (Zelli et al. 1999).

There apparently are also some frequent aggressors, such as the classic psychopaths (Patrick and Zempolich 1998), who are more instrumentally oriented. Sometimes termed proactive aggressors, they characteristically do not attack in the heat of anger but use their aggression as a tactic to further their ends.

Whatever its exact nature, this relatively persistent aggressive disposition is typically part of a general pattern of social deviation. Those who often depart from conventional social standards by deliberately hurting others around them are also likely to violate other traditional social norms such as by being heavy users of alcohol and drugs and engaging in crimes. This readiness to engage in antisocial conduct can continue over the years. Longitudinal investigations have repeatedly demonstrated that people who are highly aggressive as children are more likely than their less combative peers to also be convicted of a criminal offense by the time they enter adulthood (Reiss and Roth 1993).

2.2 *Cultural and Community Influences*

Community, ethnic, national, racial, and socioeconomic groupings can differ in their rates of violent crimes (Reiss and Roth 1993). Psychological accounts of adult violence usually refer to within-the-person psychological processes in explaining this variation. Some of these analyses focus on emotional reactions, proposing, for example, that the relatively high crime rates in impoverished areas (Berkowitz 1993), as well as the high incidence of homicides in the warmer regions of the globe (Anderson and Anderson 1998), stem in part from the negative affect generated by the aversive circumstances. Other interpretations have emphasized the role of widely shared values, knowledge structures, and modes of information processing, generally postulating a culture of violence in these groups and regions. Thus, according to Nisbett and Cohen (1996), among other investigators, many White (not African-American) US southerners are apt to believe they are justified in killing another person in defense of their families or property, or more generally when they are confronted by serious threats to their honor. Nisbett and Cohen (1996) also showed experimentally that southerners were typically more

likely than their northern counterparts to interpret another person's ambiguous encounter with them as an act of hostility and then become angry.

The emotional and cultural perspectives should be regarded as supplementary rather than competing accounts of group differences in the proclivity to violence. It is also clear that the differences in violence rates among a number of community, regional, and national groupings cannot be completely explained by individual-level formulations, whatever their exact nature. Any truly satisfactory analysis of the USA's high homicide rate obviously must recognize the significant contribution made by the ready availability of firearms in the USA. Then too, noting that certain urban areas continue to have high crime rates even when their ethnic or racial composition changes, some writers contend that the social disorganization and weak community controls within these areas is greatly responsible for their high levels of antisocial conduct (Reiss and Roth 1993).

2.3 *Biological Influence*

2.3.1 Heredity. Although we know that antisocial tendencies such as aggression tend to run in families, we cannot say unequivocally whether this family effect is due to the common environment, or the genetic influences shared by the family members, or both (Geen 1998). The few investigations employing behavioral measures have obtained only weak, if any, indications of a hereditary patterning in the disposition to violence. By contrast, the Miles and Carey (1997) meta-analysis found that both heritability and family environment contributed to individual differences in personality measures of aggressive inclinations. This analysis also suggested that the relative importance of genetic influences increases with entry into adulthood.

2.3.2 Gender and hormonal influences. In almost every animal species investigated, including humans, males tend to be more aggressive than females. At the human level, men have been reported to be more aggressive than women in virtually every society for which data are available, and furthermore, crime statistics around the world consistently show that far more males than females are arrested for violent crimes (Berkowitz 1993). Nevertheless, it still is not possible to make a simple, sweeping statement about gender differences in aggressiveness that holds across all situations, provocations, and modes of expression.

For one thing, men and women may differ in what kinds of situations spur them to attack a target. A meta-analysis of experimental studies examining such gender differences (Bettencourt and Miller 1996) suggests that men are more likely than women to become assaultive when their intelligence is cast in doubt or they are frustrated, whereas both genders become aggressively inclined when they are openly insulted. The genders probably also differ in what form of aggression they exhibit when they are provoked. Although men are typically more prone to attack an offender directly than are women, Lagerspetz and her colleagues (see Geen 1998) indicate that the angry males' greater propensity to direct assaults decreases as they go from childhood into late adolescence when they make greater use of indirect and verbal aggression. In sum, much of the research in this area suggests that men have a stronger biological disposition to react with direct physical aggression than do women when they are emotionally aroused, but that learning can lessen, or for that matter, even increase this gender difference in the proclivity to direct assaults.

3. *Situational Influences*

3.1 *Cognitively Primed Aggression*

Although public concern about the heavy dose of violence portrayed on TV and in the movies focuses largely on what children may learn from these frequent depictions, adults can also be affected by what they see and hear in the mass media, even if only for a relatively short time. The witnessed or reported violence can prime aggression-related thoughts and action tendencies in the audience members, especially if they already possess strong hostile dispositions (Berkowitz 1993, Geen 1998). If their restraints against aggression are weak at that time, they are apt to be hostile to others in thought, word, or deed until the priming effect subsides. Widely publicized offenses all too frequently spur 'copycat crimes' in this manner. There can even be a 'contagion' of suicides after a report that a well-known personage has taken his or her own life, as Phillips has noted (see Berkowitz 1993). People in the media audience apparently get ideas from what they have seen or read and, if they are already inclined to the same behavior, may act on these thoughts.

3.2 *Affectively Generated Hostile Aggression*

A very wide variety of aversive occurrences can also promote aggressive responses. The afflicted persons may well have a strong desire to escape from the unpleasant situation, but at the same time, their intense negative affect could also activate aggression-related feelings, ideas, and even motor impulses. Under the right circumstances (such as weak inhibitions at that time, an appropriate target, and strong aggressive dispositions), these aversively generated aggression-associated reactions can be stronger than the urge to escape so that an available target is attacked.

3.2.1 Pain and other physically unpleasant conditions. Experiments with a variety of species have now demonstrated that animals suffering from physical pain are likely to assault a nearby peer, especially when escape from the aversive stimulation is not possible and the aggressor had not previously learned to anticipate punishment for such an attack. Somewhat similarly, people in pain are often angry and even prone to hostile thoughts. Nevertheless, the pain-produced instigation to aggression may not become manifest in humans unless intervening, aggression-related cognitions are also present. As an example, the hostile ideas primed by the sight of weapons heightens the chances that great physical discomfort will lead to open aggression (Lindsay and Anderson 2000).

Other aversive conditions, such as decidedly uncomfortable temperatures, can also promote violence (see Anderson and Anderson 1998). The hottest regions of the USA and Europe typically have higher violent crime rates than the areas usually experiencing more comfortable temperatures. Further, within a specific locality (for example, in Dallas or Minneapolis), generally speaking, more violent crimes are committed on the hotter than cooler days. Still, with all of the empirical support for this temperature–aggression relationship at the area/community level, other influences can operate to mitigate the adverse effects of the unpleasant weather. The discomfort-induced instigation to aggression can at times be masked by an even stronger urge to escape from the heat, if escape is possible. And then too, the aversively generated hostility obviously will not be manifested openly if there are strong restraints against aggression in the situation and suitable targets are not available (Berkowitz 2001).

3.2.2 Frustrations and other social stressors. Frustrations can also be decidedly unpleasant and thus give rise to an aggressive urge. Although the idea that frustrations can breed aggressive inclinations has had a long and controversial history in the social sciences, it undoubtedly is best known in psychology through the monograph 'Frustration and Aggression' by Dollard et al. (1939). These writers argued that a frustration, defined essentially as an obstacle to the attainment of an expected gratification, produces an instigation to harm someone, principally but not only the agent viewed as blocking the goal attainment. This proposition has been criticized as seriously incomplete, but Berkowitz's (1989) survey of the available literature indicates that there is considerable evidence for its basic validity, and that even socially legitimate, non-ego-threatening barriers to goal attainment can produce aggressive reactions. Other, more recent, research indicates that thwartings can lead to aggression even without prior learning. In explaining why frustrations do not always have this effect, Berkowitz has proposed that frustrations will generate an instigation to aggression only to the extent that they evoke intense negative affect (see Berkowitz 1993, 2001).

This reformulation ties the frustration–aggression hypothesis together with the sociological 'social strain' conceptions of antisocial behavior. Both lines of thought basically argue that any greatly unpleasant social condition can promote antisocial conduct, including aggression. Consistent with such a statement, in one study (Catalano et al. 1993) job layoffs led to an increase in self-reported violent behavior if alternative employment was not readily available. In general, barriers to economic success can have criminogenic effects, particularly if the affected persons are not clearly threatened with punishment for any aggression they display and have not become apathetic and resigned to their privations. Social stress can also contribute to domestic violence. In Straus's (1980) survey, the greater the number of stressors the adult respondents reported experiencing, the more likely they were to say they had abused their children (Berkowitz 2001).

4. *Future Directions*

Obviously, a good deal still has to be learned about the influences, such as the mass media, peer groups, and social stressors, that contribute to adult aggression, and also about how these influences might be mitigated. Judging from contemporary trends, much of the future psychological research in these areas will focus on how mental processes operate to bring about, or lessen, the adverse effects. This relatively microanalytic approach will tell us much about aggression, but it is also apparent that complementary research by other social scientists will also be helpful if there is to be a truly comprehensive understanding of violent behavior.

See also: Agonistic Behavior; Antisocial Behavior in Childhood and Adolescence

Bibliography

Anderson C A, Anderson K B 1998 Temperature and aggression: Paradox, controversy, and a (fairly) clear picture. In: Green R, Donnerstein E (eds.) *Human Aggression: Theories, Research, and Implications for Social Policy*. Academic Press, San Diego, CA

Bandura A 1973 *Aggression: A Social Learning Analysis*. Prentice Hall, Englewood Cliffs, NJ

Berkowitz L 1989 Frustration-aggression hypothesis: Examination and reformulation. *Psychological Bulletin* **106**: 59–73

Berkowitz L 1993 *Aggression: Its Causes, Consequences, and Control*. McGraw-Hill, New York

Berkowitz L 1998 Aggressive personalities. In: Barone D F, Hersen M, Van Hasselt V B (eds.) *Advanced Personality Theory*. Plenum, New York, pp. 263–85

Berkowitz L 2001 Affect, aggression, and antisocial behavior. In: Davidson R J, Scherer K, Goldsmith H H (eds.) *Handbook of Affective Sciences*. Oxford University Press, Oxford, New York

Bettencourt B A, Miller N 1996 Sex differences in aggression as a function of provocation: A meta-analysis. *Psychological Bulletin* **119**: 422–47

Catalano R, Dooley D, Novaco R, Wilson G, Hough R 1993 Using ECA survey data to examine the effects of job layoffs on violent behavior. *Hospital and Community Psychiatry* **44**: 874–79

Crick N R, Dodge K A 1996 Social information-processing mechanisms in reactive and proactive aggression. *Child Development* **67**: 993–1002

Dollard J, Doob L, Miller N, Mowrer O, Sears R 1939 *Frustration and Aggression*. K Paul, London

Geen R G 1998 Aggression and antisocial behavior. *Handbook of Social Psychology* Vol. 2, 4th ed. McGraw-Hill, New York

Lindsay J J, Anderson C A 2000 From antecedent conditions to violent actions: A general affective aggression model. *Personality and Social Psychology Bulletin* **26**: 533–47

Miles D R, Carey G 1997 Genetic and environmental architecture of human aggression. *Journal of Personality and Social Psychology* **72**: 207–17

Nisbett R E, Cohen D 1996 *Culture of Honor: The Psychology of Violence in the South*. Westview, Boulder, CO

Olweus D 1979 Stability of aggressive reaction patterns in males: A review. *Psychological Bulletin* **86**: 852–75

Patrick C J, Zempolich K A 1998 Emotion and aggression in the psychopathic personality. *Aggression and Violent Behavior* **3**: 303–38

Reiss Jr A J, Roth J A (eds.) 1993 *Understanding and Preventing Violence*. National Academy Press, Washington, DC

Straus M 1980 Social stress and marital violence in a sample of American families. *Annals of New York Academy of Science* **347**: 229–50

Tedeschi J T, Felson R B 1994 *Violence, Aggression and Coercive Actions: A Social Interactionist Perspective*. American Psychological Association, Washington, DC

Toch H 1992 *Violent Men: An Inquiry into the Psychology of Violence*. American Psychological Association, Washington, DC

Zelli A, Dodge K A, Lochman J E, Laird R D 1999 The distinction between beliefs legitimizing aggression and deviant processing of social cues: Testing measurement validity and the hypothesis that biased processing mediates the effects of beliefs on aggression. *Journal of Personality and Social Psychology* **77**: 150–66

L. Berkowitz

Aging and Education

This article focuses on educational issues in adulthood, particularly in middle adulthood and older age. First, global demographic changes in the age structure of societies are discussed with particular emphasis on implications for education. The second section focuses on change and stability in cognitive development in adulthood, as well as the long-term effects of early education on later life and cognitive training in adulthood. Finally, current and future trends in education for adults and the aged are discussed, including efforts to promote lifespan and global learning and the potential for utilization of scientific and technological advances in adult education.

1. *Demographics of Aging and Education*

Growth in the world's population, as well as changes in the age structure of societies will impact the nature of education as well as the demographic characteristics of the learner. As the proportion of adults in middle and later part of the lifespan increases, the number of adult learners will increase, as will the diversity of this group of learners.

1.1 *The Shifting Age Structure*

During the twentieth century, the population of the world has grown substantially in both developed and developing countries. Developing countries, particularly the regions of Latin America, Asia, and Africa, are increasingly accounting for the vast majority of growth in the world population (US Census Bureau 1999). These countries face the greatest increases in population, yet have substantially fewer resources in terms of health, technology, and education. The population growth witnessed in developing countries is in contrast to many developed nations in Europe which are below population replacement (i.e., the number of deaths is greater than the number of births; US Census Bureau 1999).

In contrast to developed nations, the increase in middle age and aged adults in developing countries is occurring in just a few cohorts (i.e., generations). Developed countries, whose populations have aged more slowly, were able to adjust more gradually to demographic shifts and implement corresponding social agendas. In contrast, developing countries are aging before they have resources and social policies in place, forcing them to make major rapid social and policy changes to take into account population shifts.

1.2 *Implications of Demography Shifts for Education and Aging*

Successive cohorts of adults throughout the twentieth century have attained greater levels of formal education compared to previous cohorts. Gross enrollment ratios (i.e., percentage of the school-age population corresponding to the same level of education in a particular academic year) for participation in primary, secondary, and tertiary levels of education increased from 50 percent in 1970 to 63 percent in 1997

for the world as a whole (UNESCO 2000b). In developing countries this ratio has risen from 47 percent to 63 percent, compared to a shift of 72 percent to 85 percent for developed countries during the same time period (UNESCO 2000b). Country-specific data from the USA, mirrors the general trend for increased educational attainment for successive cohorts: 83 percent of adults over the age of 25 in 1998 had completed high school, and 24 percent had completed four years of college, compared with a rate of 25 percent and 5 percent, respectively, in 1940 (US Census Bureau 1998). Currently, American adults over the age of 64 are less likely than adults aged 35 to 64 years to possess a high school diploma; however, given cohort trends in postsecondary education, future cohorts of elderly will have increasing levels of education. While this trend is promising for future cohorts of elderly, particularly in developed countries, it also implies that current cohorts of elderly are seriously disadvantaged in educational attainment compared to current younger adult cohorts (US Census Bureau 2000).

Although the overall levels of education have risen for successive cohorts throughout the twentieth century, the gross enrollment ratios indicate that universal education is not present, even in the most developed countries. The number of expected years of formal education ranges in regions throughout the world from slightly more than one year in less developed countries to over 16 years in developed countries (UNESCO 1996). Current illiteracy rates throughout the world also indicate disparities between developing and developed countries. Developed countries in Europe and North America have very low rates of illiteracy (i.e., average = 1.4 percent) compared to developing countries (i.e., average = 26 percent); and the least developed countries in regions such as Africa and southern Asia (i.e., average = 49 percent), which are more economically challenged and have higher birth rates (UNESCO 2000a). As demonstrated by these statistics, vast differences in educational attainment exist between countries, as well as between cohorts within a country.

Groups such as older women and ethnic minorities, who are increasingly accounting for a greater proportion of the elderly, will be particularly affected, putting them at an even greater disadvantage. Despite advancements in the equity of educational opportunities, great disparities still exist for women, ethnic minorities, and the economically disadvantaged. Women's education during the twentieth century has improved tremendously, however, worldwide fewer girls attend school than boys and women comprise two-thirds of illiterate adults (UNESCO 1996). Educational equality has also been difficult for ethnic minorities within some countries.

With recent changes in population demographics, efforts to facilitate maintenance of independence and productivity in the elderly are gaining attention. It is projected that the number of elderly and their need for support will steadily increase during the next 25 years throughout the world (US Census Bureau 1999). The shifting demographics will have repercussions on numerous policy initiatives, including length of work life and retirement age. Some countries have or are considering eliminating mandatory retirement, or are raising the standard retirement age. As workers remain the in work force to later ages, maintenance of cognitive abilities and issues of educational updating will gain attention.

2. *Changes in Cognition Across Adulthood*

The changes occurring in cognitive abilities throughout adulthood have important implications for the formal education of adults in middle and older adulthood as well as self-directed learning. The variable trajectories of cognitive ability throughout adulthood, as well as the long-term beneficial effects of early formal education and the potential for cognitive training in later adulthood are discussed below.

2.1 *Age-related Changes in Cognition*

A well-established approach to the study of adult cognitive ability has been the examination of higher-order dimensions of psychometric mental abilities, particularly fluid and crystallized intelligence (Horn and Hofer 1992). Fluid intelligence refers to abilities needed for abstract reasoning and speeded performance whereas crystallized intelligence refers to knowledge acquired through one's culture including verbal ability and social knowledge (Schaie 1996).

Longitudinal research examining cognitive development has revealed that mental abilities vary in their developmental trajectories across adulthood (e.g., the Seattle Longitudinal Study: Schaie 1996, the Berlin Aging Study: Smith and Baltes 1999). A substantial body of research in the USA has demonstrated that fluid abilities, such as inductive reasoning, peak in early middle adulthood rather than in adolescence as previously thought. Fluid abilities remain stable in middle age and first show reliable decline in the mid-sixties. In contrast, crystallized abilities, such as vocabulary, do not peak until middle age and show reliable decline later in the mid-seventies (Schaie 1996). Similar developmental trajectories in abilities have been reported in Canadian and European longitudinal research (Backman 2001).

Decline in cognitive ability prior to age 60 is usually considered to be associated with ensuing pathological changes, and universal decline on all markers of intelligence in normal elderly is not evident even by the eighties (Schaie 1996). Findings from Swedish longitudinal study demonstrate that even the oldest-old (i.e., a sample of individuals aged 84 and older), who do not exhibit cognitive impairment at

baseline assessment, demonstrate relative stability over a two-year period on several markers of cognitive ability (Johansson et al. 1992).

2.2 Cohort Differences in Cognitive Ability and Education

In addition to varying individual developmental trajectories, mental abilities also show different cohort trends as well. Some abilities show positive cohort trends with successive cohorts functioning at higher levels when at the same chronological age. Other abilities exhibit curvilinear or negative cohort trends. The two abilities showing the strongest positive cohort trends are inductive reasoning and verbal memory—both representative of fluid ability. Current cohorts of the elderly are thus at double disadvantage on these abilities due to relative early age-related decline on fluid ability, combined with strong positive cohort trends on these same abilities. More modest positive cohort trends have been shown for spatial and verbal abilities. In contrast, curvilinear cohort trends have been shown for numerical abilities with birth cohorts 1918–1920s showing higher functioning compared to earlier or later cohorts when at the same chronological age. There does however appear to be a slowing of these cohort differences, and it is estimated that during the first part of the twenty-first century the differences between cohorts will become smaller (Schaie 1996).

These cohort trends in abilities are multiply determined; however, increasing levels of education across cohorts as well as medical and health advances appear to have been strong influences. The impact of increases in educational attainment as well as shifts in educational practice toward discovery learning, procedural knowledge and metacognition may have contributed in particular to the strong positive cohort trends for inductive reasoning and verbal memory. A recent reduction in the magnitude of cohort differences in abilities may be related to a plateauing of the dramatic increases in educational attainment that occurred in the later part of the twentieth century. Alternatively, the slowing of cohort trends may reflect the decline in college-entrance exam performance reported for recent cohorts of young adults; these cohorts are now in their late twenties and thirties and are represented in longitudinal studies of adult cognition.

2.3 Lifelong Benefits of Early Formal Education

Some research suggests that the benefits of early formal education extend into adulthood. Although debate exists regarding the extent, as well as the mechanisms (i.e., compensatory vs. protective) by which early educational benefits continue to be manifested in later life, numerous cross-cultural studies have found greater levels of formal education to be associated with decreased risk of cognitive impairment in later life (e.g., Kubzansky et al. 1998). Several scenarios for how early education benefits later cognitive functioning have been offered. First, greater education attainment in adolescence and young adulthood increases opportunities and access to further education through the remainder of the lifespan. Likewise, attainment of certain levels of education provides entry into particular career opportunities. A second less direct influence of early education on later cognition focuses on the increased financial and environmental resources available to those with higher educational attainment. Those with greater financial resources typically have access to better healthcare and social services, which may facilitate maintenance of cognitive functioning in late life. Finally, early educational attainment may result in a higher level of cognitive ability and thus a higher threshold of functioning from which decline occurs in later life. For example, level of education may not delay the onset of dementia; however, it appears that it may be related to a delay in its symptomatology.

2.4 Cognitive Training Research with Adults

Given that fluid abilities show age-related decline beginning in the sixties and also that positive cohort trends for some fluid abilities place current elderly cohorts at a disadvantage, the question arises of whether behavioral interventions might be effective in remediating and/or enhancing cognitive performance in later adulthood. Educational interventions have traditionally focused on the earlier part of the lifespan when children are first developing mental abilities and skills. There has been less research on educational interventions in middle adulthood (except for work-related training) and even less study on cognitive interventions in later life. Outcomes from cognitive interventions later in the lifespan may be qualitatively different from those earlier in the lifespan (Willis and Schaie 1986). For older adults suffering cognitive decline, the intervention focuses on the possibility of remediating prior loss in level of ability. In contrast, for older adults who have not declined on an ability, the question is whether interventions can boost cognitive performance above prior levels. In order to examine these questions, longitudinal data on older adults, cognitive functioning prior to the intervention are needed in order to determine whether elders have declined or not on the abilities to be trained.

Since the 1970s, there has been a growing body of cognitive intervention research in later adulthood focusing on a variety of mental abilities, including memory, reasoning, and speed of processing (Camp 1999). Much of the research has shown that non-demented, healthy older adults can improve their performance as a function of brief educational train-

ing. Researchers have focused on different questions regarding the plasticity of cognitive functioning in later adulthood. A number of researchers have compared training gains for young adults vs. older adults. Due to cohort differences, younger and older adults were performing at different levels prior to training. Significant training effects have typically been found for both young and older adults; however, the cohort differences in level of performance remained after training (Willis and Schaie 1994); that is, older adults gained significantly from the intervention, but the training did not eliminate the cohort differences in level of performance. Baltes and co-workers have focused on a form of training known as 'testing the limits' in which older and younger adults were trained on the method of loci in list learning tasks, and then recall was assessed under increasing levels of speeded performance (Kliegl et al. 1989). Although both young and old showed significant training gains, the old showed less improvement when tested under highly speeded conditions.

In the context of the Seattle Longitudinal Study, Willis and co-workers have examined whether training on fluid abilities is effective for both older adults who have declined on the target ability and for those who have remained stable (Schaie and Willis 1986). Significant training effects have been shown for both stables and decliners on two fluid abilities, inductive reasoning and spatial orientation (Willis and Schaie 1986). Seven years after training older adults trained on the ability were performing at a higher level than adults not given training on a specific ability (Willis and Schaie 1994).

While the cognitive training research appears promising, it is important to consider caveats to these findings. First, training effects have been found only with nondemented, community-dwelling elderly, not with demented patients. Second, while training effects have been demonstrated for multiple measures of the ability trained, training transfer is limited to the particular ability that was the target of training. That is training on a specific ability does not lead to significant enhancement on other primary abilities. Third, much of the training research has been conducted with young–old individuals who are White and of middle to upper socioeconomic status. Further research is needed regarding whether training effects can be demonstrated for the old–old and for minority elderly. One such study is currently being conducted by the US National Institute on Aging and National Institute of Nursing Research, which involves a multisite clinical trial examining the effects of cognitive training for more representative groups of elderly (Jobe et al. 2000).

3. *Trends in Education for Adults and the Aged*

Educational systems are continually changing in response to the political, economic, and social forces that occur in countries throughout the world (UNESCO 1996) and these forces are particularly influential in terms of adult education and vocational training. Two principal broad classes of educational trends relevant to adult learners are expected to continue during the first part of the twenty-first century, including the evolution of education into a system of lifelong learning and increasing utilization of science and technology.

3.1 *Lifespan Learning and Globalization of Education*

Education is continually affected by societal changes, particularly in the work place. These changes have promoted the emergence of lifespan learning and the globalization of education. The current transition from industrial to post-industrial economies occurring in many countries (Beare and Slaughter 1993) and a general increase in economic interdependence has fostered an interdependent, global approach to education. The shift in many countries away from industry-oriented occupations necessitates changes in occupational training, evaluation of students from all countries compared to global criteria of competence, and increased creativity and flexibility in meeting future training needs (e.g., creation of new disciplines, increased interdisciplinarity; Beare and Slaughter 1993). Furthermore, the emergence of a more global consciousness and cohesive worldview has prompted educators across the world to foster global awareness and competence in students (Beare and Slaughter 1993, Miller 2000). Systems of higher education throughout the world have begun to converge as the result of emerging national and international educational organizations and the sharing of educational information, theory, and research. This however, could come at a cost for non-English speaking countries with limited technological availability, as they may not be able to remain up-to-date. As the composition of the adult learner population changes, the challenge for educators in the twenty-first century will be the necessity to strive for universal education especially for under-served populations including ethnic minorities, women, and the economically disadvantaged.

Initial formal education, continuing education throughout adulthood, and the creation of everyday learning environments will grow increasingly intertwined. Recent increases in the length of nonworking life and amount of free time during employment, have resulted in an increasing role for education throughout the lifespan (UNESCO 1996). On-the-job training and general and vocational education are becoming more intermixed due to the growing demands to compete in the world market. Economic prosperity, company viability, and employee productivity have become increasingly interdependent. The uncertainty of the world labor market has also highlighted the need for continuing education. An unmet demand for skilled

workers and rising unemployment of unskilled workers point to the need for adequate training (Pair 1998). Education in the workplace is becoming vital as lifelong learning is required to enable job-sharing among employees with comparable training and to support and promote the growth of new occupations; workers must be prepared for present employment positions as well as positions of the future (Pair 1998). Inequalities in initial training (i.e., early formal education) have great impact on subsequent adult and lifelong learning, highlighting the importance of early education as the time for initial training with increasing amounts of subsequent training and education throughout adulthood.

3.2 Impact of Science and Technology on Adult Education

Another major trend in education has been the pervasiveness of computers and the Internet in the last decade, which have increased both older adults' formal and informal educational opportunities. For example, the use of technology throughout the later part of the twentieth century in distance learning has increased older adults' educational access and opportunity. Although distance education has been available during the past century in a variety of countries (e.g., Thailand, Pakistan, and Venezuela), it is only in the very recent years that more interactive options have been available. Early correspondence study programs relied on communication through the mail (Miller 2000), however, long distance learning has become increasingly interactive as these programs have incorporated television, videos, computers, and in the last decade, e-mail and the Internet (Miller 2000). As a result of the Internet, the classroom has become an international one (Miller 2000). The use of technology promotes flexible learning and has the advantages of decreasing cost, improving quality, and broadening access to educational materials, perhaps leading to virtual universities in the future.

The full impact of science and technology on adult education is not yet fully known. Technology has increased educational opportunities for adult learners and can be used to create a more optimal learning environment (i.e., familiar settings, accommodations such as large print and audio presentation). However, middle-aged and older adults' comfort level and ability to adapt to rapidly changing technological advances may be a challenge. The relative lack of computer experience by middle-aged and older adults and age-related changes in working memory, processing speed, and visuomotor skills can be impediments to adults' computer task performance (Czaja and Sharit 1993). However, relaxation of task pacing constraints, attentive interface design, and training are likely to increase older adults' ease and efficiency with computer-related tasks (Czaja 2001).

3.3 Implications of Trends for Adult Education

The emphases on lifelong and adult learning and the globalization of education, as well as the impact of science and technology have several implications for educators. First, education is increasingly taking place in contexts other than traditional educational institutions. This could be advantageous for adult learners as educational opportunities become more easily accessible and available in familiar environments. Second, education and access of knowledge will increasingly require competence in new technologies, which is likely to place adults and the elderly at a disadvantage compared to younger cohorts who tend to be more familiar with these technologies. Third, cohort differences and age-related change in higher order abilities such as inductive reasoning, working memory, and executive functioning may make older adults' use of new technologies particularly challenging. Finally, the rapidity of knowledge increase will require lifelong learning and adaptation, particularly in relation to work settings.

4. Dynamic Between Aging and Education

The dynamic between aging and education will continue to change as the composition of adults over the age of 60, is transformed, education takes a more global approach and encompasses learning throughout the lifespan, and technological advances continue to impact educational methods. Coming years will see an increase in the number of older women, the very oldest segment of the population (i.e., old–old: adults aged 80 and older), and greater diversity in the ethnicity and needs of the older adult population. Increased attention will be devoted to the maintenance and improvement of functioning in older adulthood, which can be aided by investment in early formal education as well as educational opportunities throughout the lifespan.

As the duration and nature of work and retirement change, the educational needs of current and future cohorts will also continue to change. Given the impact of technology in the workplace and the emergence of second careers and later retirement ages, the traditional conceptualization of the relationship between education, employment, leisure, and retirement is being reevaluated (Krain 1995, UNESCO 1996). In western cultures, individuals have typically received education and career preparation only in early childhood, worked at a career throughout early and middle adulthood, followed by retirement in older adulthood. Educational policies must increasingly address the growing number of work transitions, periods of unemployment, decreased period of transition prior to retirement, and increased part-time work after retirement. Future policies should include expansion of adult education, increased availability of lifelong

career-oriented education and training, and greater leisure-oriented education (Krain 1995).

See also: Adult Education and Training: Cognitive Aspects; Cognitive Aging; Education and Learning: Lifespan Perspectives; Education in Old Age, Psychology of; Lifespan Theories of Cognitive Development; Memory and Aging, Cognitive Psychology of

Bibliography

Backman L 2001 Learning and memory. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 5th edn. Academic Press, San Diego, CA

Beare H, Slaughter R 1993 *Education for the Twenty-first Century*. Routledge, New York

Camp C 1999 Memory interventions for normal and pathological older adults. In: Schulz R, Maddox G, Lawton M P (eds.) *Annual Review of Gerontology and Geriatrics, International Research*. Springer, New York, Vol. 18

Czaja S J 2001 Technological change and the older worker. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 5th edn. Academic Press, San Diego, CA

Czaja S J, Shant J 1993 Age differences in the performance of computer-based work. *Psychology and Aging* **8**(1): 59–67

Horn J L, Hofer S M 1992 Major abilities and development in adults. In: Sternberg R J, Berg C A (eds.) *Intellectual Development*. Cambridge University Press, Cambridge, UK, pp. 44–99

Jobe J B, Smith D M, Ball K, Tennstedt S L, Marsiske M, Rebok G W, Morris J N, Willis S L, Helmers K, Leveck M D, Kleinman K 2000 *ACTIVE: A Cognitive Intervention Trial to Promote Independence in Older Adults*. National Institute on Aging, Bethesda, MD

Johansson B, Zarit S, Berg S 1992 Changes in cognitive functioning of the oldest old. *Journal of Gerontology*: *Psychological Sciences* **47**(2): P75–80

Kliegl R, Smith J, Baltes P B 1989 Testing-the-limits and the study of adult age differences in cognitive plasticity of a mnemonic skill. *Developmental Psychology* **25**: 247–56

Krain M A 1995 Policy implications for a society aging well. *American Behavior Scientist* **39**(2): 131–51

Kubzansky L D, Berkman L F, Glass T A, Seeman T E 1998 Is educational attainment associated with shared determinants of health in the elderly? Findings from the MacArthur Studies of Successful Aging. *Psychosomatic Medicine* **60**(5): 578–85

Miller G E 2000 General education and distance education: Two channels in the new mainstream. *The Journal of General Education* **49**(1): 1–9

Pair C 1998 Vocational training yesterday, today and tomorrow. In: Delors J (ed.) *Education for the Twenty-first Century: Issues and Prospects*. UNESCO, Paris, pp. 231–51

Schaie K W 1996 *Intellectual Development in Adulthood: The Seattle Longitudinal Study*. Cambridge University Press, Cambridge, UK

Schaie K W, Willis S L 1986 Can intellectual decline in the elderly be reversed? *Developmental Psychology* **22**: 223–32

Smith J, Baltes P B 1999 Trends and profiles of psychological functioning in very old age. In: Baltes P B, Mayer K U (eds.) *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, Cambridge, UK, pp. 197–226

US Census Bureau 1998 *Higher Education Means More Money, Census Bureau Says*. US Census Bureau, on-line, *CB98-221*, http://www.census.gov/Press-Release/cb98-221.html

US Census Bureau 1999 *World Population at a Glance: 1998 and Beyond. International Brief* (*IB*) US Census Bureau, on-line, *IB/98-4*, http://www.census.gov/ipc/www/wp98.html

US Census Bureau 2000 *Aging in the United States: Past, Present, and Future*. US Census Bureau, on-line, http://www.census.gov/ipc/prod/97agewc.pdf

United Nations Educational, Scientific and Cultural Organization (UNESCO) 1996 *Learning: The Treasure Within: Report to UNESCO of the International Commission on Education for the Twenty-first Century*. UNESCO, Paris

United Nations Educational, Scientific and Cultural Organization (UNESCO): Institute for Statistics 2000a *Estimated Illiteracy Rate and Illiterate Population Aged 15 Years and Over*. UNESCO, on-line, http://unescostat.unesco.org/statsen/statistics/yearbook/tables/Table-II-S-1-Region.html

United Nations Educational, Scientific and Cultural Organization (UNESCO): Institute for Statistics 2000b *Gross Enrolment Ratios by Level of Education*. UNESCO, on-line, http://unescostat.unesco.org/statsen/statistics/yearbook/tables/Table-II-S-5-Region(Ger).html

Willis S L, Schaie K W 1986 Training the elderly on the ability factors of spatial orientation and inductive reasoning. *Psychology and Aging* **1**: 239–47

Willis S L, Schaie K W 1994 Cognitive training in the normal elderly. In: Forette F, Christen Y, Boller F (eds.) *Plasticité cérébrale et stimulation cognitive*. Foundation National de Gérontologie, Paris, pp. 91–113

S. L. Willis and J. A. Margrett

Aging and Health in Old Age

1. Introduction: Living Longer and Better or Worse?

There are three different models describing how disability may change in the US population. First, with improvements in the treatment of some chronically disabling diseases (e.g., cardiac surgery in children with Down's syndrome so they can survive past age 40, i.e., through reproductive ages), it was hypothesized that the US would enter a period of a 'pandemic' of chronic diseases and disability (Gruenberg 1977, Kramer 1980). That is, it was expected that persons with chronic diseases, and the profound disabilities they can generate, would survive many more years raising the prevalence of chronic disability and the average amount of lifetime that could be expected to be lived in an impaired state (Verbrugge 1984).

A second perspective, due to Fries (1980) and Riley and Bond (1983), was that the time (age) to the occurrence of chronic disability could be increased independently of changes in life expectancy (time to death). Life expectancy was postulated to be able to increase only to 85 years of age (Fries 1980) with the

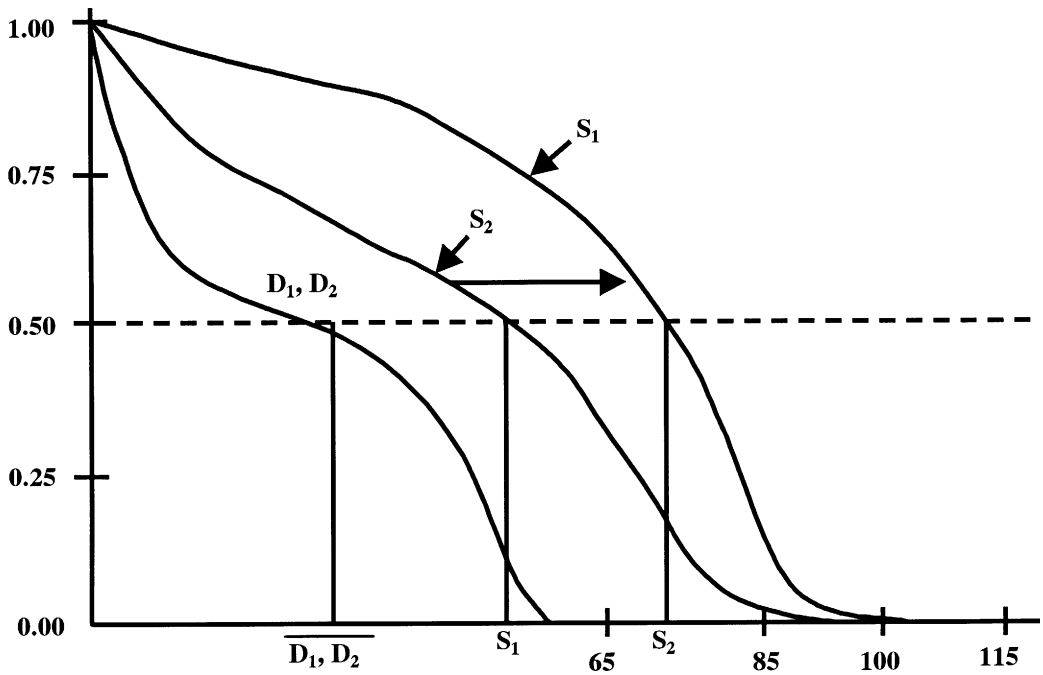

Figure 1
Pandemic of chronic disease due to prolangation of life of severely disabled persons (Kramer and Gruenberg 1977)

variance of the age at death decreasing—leading to a rectangularization of the survival curve. However, it was suggested that, as the survival curve became rectangular, so could the curve describing the age at occurrence of chronic degenerative disease so that the curves, ideally, could meet so that all life expectancy would be in an unimpaired state.

In the third model it was suggested that the times at which chronic disabilities and diseases onset could be in a dynamic equilibrium with overall survival (Manton 1982). In this case which diseases were modified by interventions, and in what ways, affect the relation of the survival and disability age-dependent onset curves over time. By appropriately selecting disease interventions, that is, by targeting for prevention those with the greatest potential for inducing chronic disability (such as Alzheimer's disease), both total life expectancy and disability-free life expectancy could be increased. This would decrease the average amount of time spent in disabled states. This perspective is referred to as 'dynamic equilibrium.' It will produce moderate decreases in the time-weighted prevalence of chronic disability (Manton 1982).

To visually compare these three theories we use concepts developed in WHO TRS 706 (1984). We define a graph (Fig. 1) where the vertical axis is the probability of surviving from birth to Age X. The horizontal axis is Age. For each of these three graphs we define four points. Two (D_1 and D_2) represent the median (50 percent) age at onset of disability at two dates separated in time (e.g., 1982 and 1999). Two (S_1 and S_2) represent the median age at death at those same two times. In graph one, we show that, although the difference between S_1 and S_2 increased (median lifetime increased), the median age of disability onset did not change ($D_1 = D_2$) so the number of years lived with disability increases. In Fig. 2, S_1 and S_2 change little because the survival curve is nearly 'rectangular.' However, since D_1 is at lower ages relative to D_2, the number of years lived with disability declines. In Fig. 3 the median of both years lived, and years lived without disability, increased over time so that, ultimately, the amount of active life expectancy increased.

The survival curves themselves contain much more information than the four median age estimates (S_1, S_2, D_1, D_2) so that comparisons can be made at any age.

2. *Empirical Evidence of Declining Disability*

Considerable scientific and policy debate has emerged regarding the validity of observations of declines in chronic disability prevalence in the US and European elderly populations (Freedman and Soldo 1994, Waidmann and Manton 1998). The US declines in functional disability were first documented in the 1982 to 1989 National Long Term Care Surveys (NLTCS). The NLTCS are sets of longitudinally related surveys (done again in 1994 and 1999) designed to assess

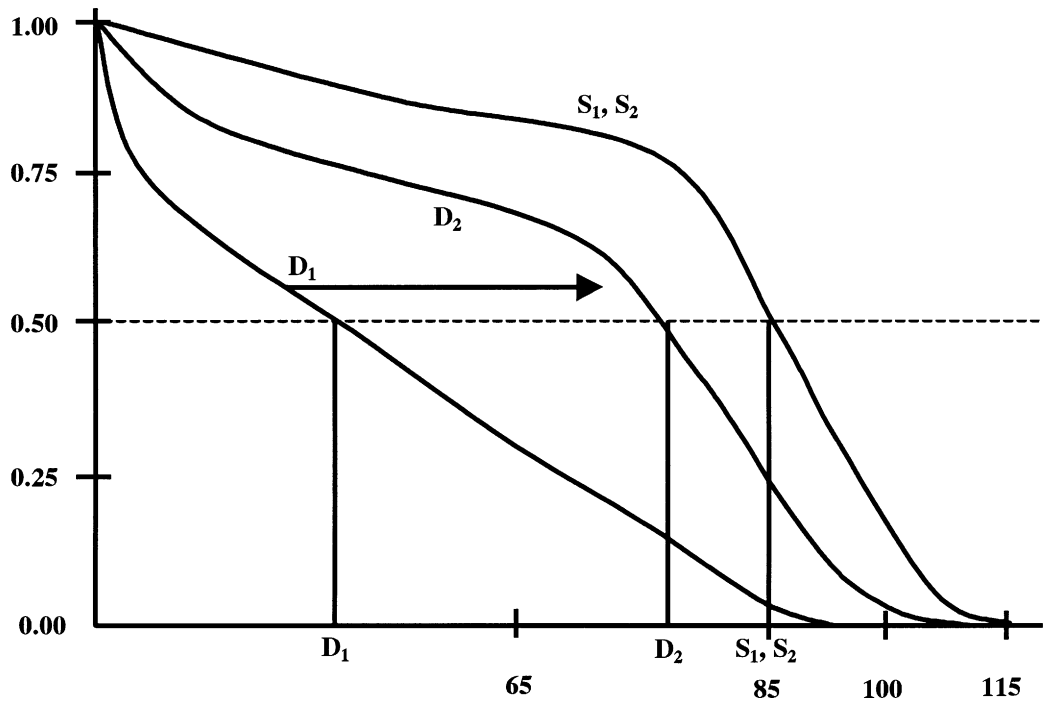

Figure 2
Compression of morbidity and morality due to the 'rectangularization' of human survival curve and delay of onset of chronic disability (Fries 1980)

changes in functional status, social conditions, and Medicare and LTC service use in the US elderly population. The NLTCS samples of individuals (not households or institutions) were drawn from Medicare enrollment lists so that nearly 100 percent of sampled persons could be followed to conduct detailed interviews, to assess functional status, to be linked to health service use and expenditures, and to document the exact date of death. Results from that longitudinal survey based on a list sample were initially thought to be at variance with estimates made from several national health surveys which were not specifically designed to longitudinally sample events generated by population processes—such as age-related health changes and disablement.

After the 1982 to 1989 results were produced, a fourth survey was done in 1994. The 1994 NLTCS confirmed the presence of the decline in chronic disability prevalence. The results from the 1982 to 1994 NLTCS are presented in Table 1.

The age-standardized rate of decline in the prevalence of chronic disability was about 0.36 percent per annum from 1989 to 1994—higher than the decline of 0.25 percent observed from 1982 to 1989 (Manton et al. 1997). This translates into a percentage point decline of 1.77 percent (25.01 to 23.24) from 1982 to 1989 (seven years) and 1.77 percent from 1989 to 1994 (five years).

Both the external and internal validity of those findings were examined in a number of ways. The declines in disability were consistent with such internal validity tests as: (a) examining whether the decline occurred after eliminating an 'age-in' sample (5,000+ new persons sampled from Medicare enrollment lists ages 65 to 69 in 1984, 1989, and 1994) so the change could be documented in only longitudinally followed population groups (i.e., persons 65+ in 1989 were age 70+ in 1994); (b) controlling for the patterns of change specific to age, race, and sex groups so that changes in demographic composition did not confound the trends; (c) determining whether comparable trends existed in the 'rate' of proxy reporting (a measure of severe disability; the proxy rate did decline in 1982 to 1994); (d) determining whether declines were consistent with changes in other covariates of disability (e.g., disability risk is lower at high levels of education; the education level of the US elderly population increased significantly); and (e) determining whether disability rate declines were consistent with declines in the prevalence of medical conditions known to cause disability (e.g., there was a large decline in the age standardized prevalence of severe cognitive impairment from 5.7 percent in 1982 to 3.8 percent in 1994, an absolute reduction from that expected (based on 1982 rates) of 610,000 cases of severe cognitive impairment in 1994).

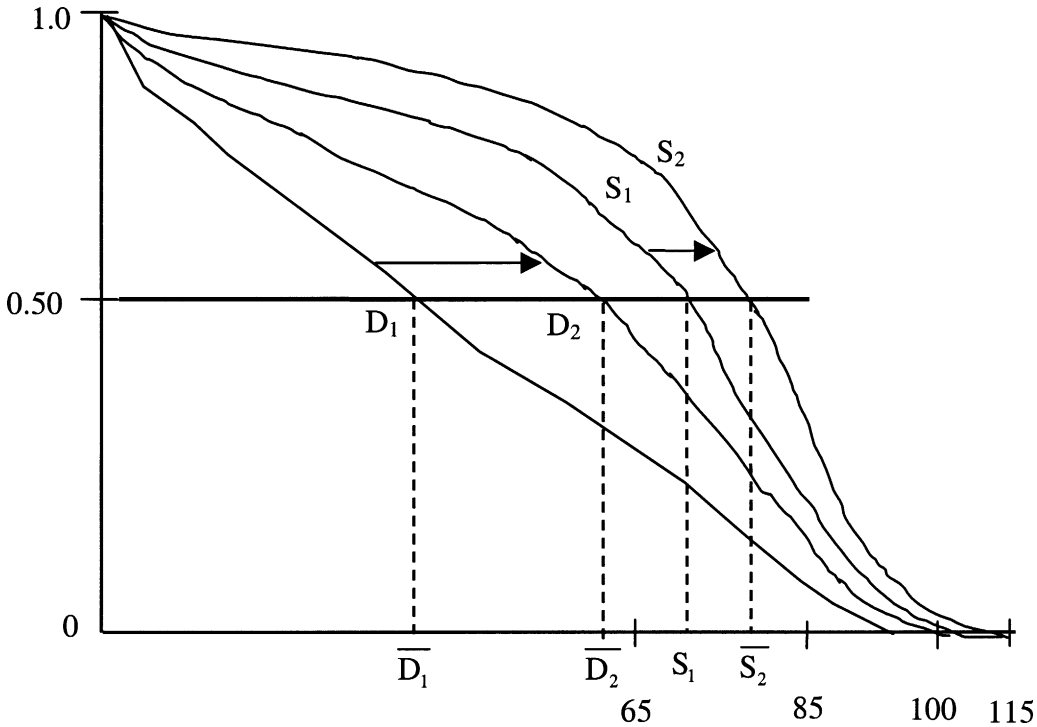

Figure 3
Dynamic equilibrium between survival and the age at onset of disability curves (Manton 1982)

Table 1
Sample weighted distribution (age standardized) of disabilities in 1982 to 1994 NLTCS

Year	1982	1984	1989	1994
Nondisabled (%)	76.28	76.28	77.31	78.53
IADL only	5.48	5.84	4.65	4.32
1 ADL	3.93	4.02	3.72	3.54
2 ADLs	2.44	2.43	2.61	2.36
3–4 ADLs	2.79	2.85	3.50	3.21
5–6 ADLs	3.39	3.08	2.75	2.81
Institutional	5.69	5.50	5.46	5.24
Housing units	92.58	93.64	94.08	94.37
Nursing home	6.30[a]	5.33	5.11	4.92
Others	1.12	1.02	0.81	0.70
Total (%) nondisabled standardized by 1994 age population distribution	74.99	75.07	76.76	78.53
Total (%) disabled standardized by 1994 age population distribution	25.01	24.93	23.24	21.47
Standardized decline rates (%) (per year)			(82–89)	(89–94)
Nonstandardized decline rates (%) (per year)			0.15	0.25

a Based on estimate that out of 1992 cases identified as in institutions, 1690 were in nursing homes and alive to potentially receive a detailed interview (only community interviews were conducted in 1982).
ADL, activities of daily living; IADL, instrumental activities of daily living.

The external validity of the disability declines was assessed by whether a decline, on the same measures of function, could be established by its replication in both European countries and in other US longitudinal surveys and historical data. Evidence of long-term (75 years) declines in chronic disease prevalence (Fogel 1994) and disability were found in studies of Civil war veterans (all male) who were aged 65+ in 1910 (birth cohorts of 1824 to 1844) when compared with World War 2 male veterans aged 65+ in the National Health Interview Survey (NHIS) in 1985–8, and to comparable groups in the National Health and Nutrition Examination Surveys. Fogel (1994) attributed the decline to improvements in nutrition. Perutz (1998) came to similar conclusions about British centenarians born after 1840 for which the population growth rate increased from 1 percent to 6 percent. The rate of decline in chronic disease was estimated by Fogel to be 6 percent per decade—or 0.6 percent per year.

Among more recent US population studies, an even larger decline than found in Manton et al. (1997) was noted by Freedman and Martin (1998) using the 1991 to 1996 Survey of Income and Program Participation—a decline which existed at the higher levels of disability and at advanced (85+) ages. Waidmann and Liu (1998) also found confirmation of declines in the 1993 to 1996 Current Medicare Beneficiary Survey. This confirmed findings in an earlier analysis which adjusted for methodological difficulties in the NHIS and which combined data from several other sources. Crimmins et al. (1997) found evidence for declines in the 1984 Supplement on Aging and the LSOA (longitudinal study of aging) from 1986 to 1990. Evidence of declines has been found in the 1985 and 1995 Supplements on Aging to the NHIS. Evidence for declines in European countries, as mentioned above, was found in Waidmann and Manton (1998).

As a consequence of this confirmatory evidence in multiple replications in the US and abroad, the focus has now shifted to questions about what caused the declines in disability, the social context of the declines, the social, economic, and health implications of the declines, and whether those declines should be included as covariates in official projections of both the size of the population by the US Census, and in forecasting the future fiscal status of Medicare and Social Security programs (Manton and Singer 2001).

3. *Consequences of Declining Disability*

A wide range of social and economic factors may be influenced by declines in disability in the elderly populations in the US and other countries. Improvements in functioning may change the ages at which retirement occurs. The trend through the 1970s and 1980s was toward a lower age at retirement. More recently the average age at retirement has tended to be static or slowly increase—at least in the US. This may increase the human capital available to the US economy and could dampen the effects of rapidly declining birth rates in a number of large (Italy) and small (Latvia) European societies.

4. *Causes of Disability Decline*

One set of observations about US disability declines questions the nature of their content and intensity. Specifically, disability is usually measured in terms of some variant of ADLs or physical performance measures. The content of these three scales are presented in Table 2.

The IADLs could have been influenced by changes in the socioeconomic environment which allow changes in socially defined gender roles (e.g., men doing more grocery shopping or laundry) or in providing more devices (e.g., improved telecommunications, better transportation systems) to aid partly impaired persons.

The functions reflected by the ADLs may be more influenced by interventions in biological and chronic disease processes (e.g., dementia). Indeed, while the IADLs were intended to reflect social and assistive device support for impairments in the elderly, the ADLs were argued to reflect a sociobiological model of disablement where functions were lost in the reverse order to which they were gained in socialization of the child. The performance measures directly reflect the ability to perform specific types of physical tasks.

A complete picture of disability requires that all of those scales be used because of their different content. Such analyses require multivariate analytic procedures to disentangle the complex inter-relation of the items and possible changes in the relation of items over time. Analyses of the ADLs, IADLs, and physical performance measures suggest that the basic nature of the underlying dimensions of disability have remained stable over time, that is, from 1982 to 1999 (the fifth and most recent of the NLTCS). From 1994 to 1999 the prevalence of chronic disability declined at an even faster rate than from 1989 to 1994. These dimensions have also shown interesting relations to a series of 29 medical conditions (e.g., severe cognitive impairment, stroke, heart attack) also surveyed in the NLTCS.

Of most interest in these relations is that the risk of Alzheimer's disease (and more generally severe cognitive impairment) declined from 1982 to at least 1994. Specifically the prevalence of severe cognitive impairment declined from 5.7 percent (age standardized to 1994) in 1982 to 3.8 percent in 1994. This change may be due to a number of biological and medical factors (e.g., effects of exogenous estrogen use in females (Tang et al. 1996); increased use of NSAIDs (non-steroidal anti-inflammatory medications) (McGeer et al. 1996)). However, it is also likely strongly linked to changes in education so that a decline in cognitive

Table 2
ADL, IADL, and physical performance measures

ADLs Needs help with	*IADLs* Difficulty doing due to health person	Levels (4) of difficulty doing
1.) Eating	1.) Heavy work	1.) Climbing stairs
2.) Getting in/out of bed	2.) Light work	2.) Bending for socks
a.) Bedfast	3.) Laundry	3.) Holding 10lb. package
b.) No inside activity	4.) Cooking	4.) Reaching over head
c.) Uses wheelchair	5.) Grocery shopping	5.) Combing hair
3.) Getting around inside	6.) Getting about outside	6.) Washing hair
4.) Dressing	7.) Traveling	7.) Grasping small objects
5.) Bathing	8.) Managing money	8.) Seeing well enough to read a newspaper
6.) Using toilet	9.) Taking medicine	
	10.) Telephoning	

impairment could be related to shifts in education in the elderly population. It is also suggested that the intensive performance of cognitive tasks could stimulate an increased complexity of neuronal connections in the brain. Preston (1992) projected that the proportion of persons aged 85 to 89 who had less than eight years of schooling would decline from 62.1 percent in 1980 to 20 percent in 2015. The prevalence of severe cognitive impairment continued to decline from 1994 to 1999 with there being one million fewer cases than expected based on the 1982 rates.

The risk of disability varies strongly with education and age; with larger declines over age 85 Manton et al. 1997). For persons aged 85+ with at least 12 years of schooling, the risk of disability was 5.8 percent lower than for persons with less than 12 years of schooling. Age standardization reduced the difference to 4.0 percent, that is, 70.2 percent of the difference was attributable to education; only 30 percent was attributable to age.

This is also consistent with major shifts in residence among the elderly where the use of nursing home facilities has declined from 6.3 percent (Table 1) in 1982 to 4.9 percent in 1994 with stays in nursing facilities also becoming shorter in duration (e.g., a median of 84 days in 1985 to a median of 63 days in 1997; Gabrel 2000) as more stays are funded by Medicare as instances of postacute rather than long-term care. More and more elderly persons, instead, are going to facilities called 'assisted living' facilities where a graded level of care is provided. A significant proportion of residents in assisted living facilities appear not to be ADL or IADL disabled in preliminary findings from the most recent (1999) NLTCS.

5. *Conclusions and Future Directions*

The nature and persistence of those declines will be evaluated and studied in several ways. First the NLTCS was repeated in 1999—a 2004 survey is being planned. The 1999 NLTCS is now being analyzed. Preliminary results suggest the rate of decline in chronic disability is accelerating further. Second, the instrumentation and temporal structure of the NLTCS is being replicated in existing (e.g., the Longitudinal Study of Danish Twins) and planned (e.g., in Sweden) European surveys. Third, the use of longitudinal data as social indicators is being exploited in measures of active life expectancy as implemented in the WHO global goals for health improvement and which is now gaining international acceptance as exploited by the efforts of the REVES (Réseau sur l'espérance de vie en sauté) groups. As a conclusion there is beginning to be more global acceptance of disability adjusted measures of demographic changes—an acceptance that has extended to such international organizations as OECD and the G7/8.

See also: Aging, Theories of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Caregiving in Old Age; Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Cognitive Aging; Dementia: Overview; Dementia: Psychiatric Aspects; Dementia, Semantic; Differential Aging; Disability, Demography of; Disability: Psychological and Social Aspects; Disability: Sociological Aspects; Ecology of Aging; Life Course in History; Old Age and Centenarians; Population Aging: Economic and Social Consequences; Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

Bibliography

Crimmins E M, Saito Y, Reynolds S L 1997 Further evidence on recent trends in the prevalence and incidence of disability among older Americans from two sources: The LSOA and the NHIS. *Journals of Gerontology Series B—Psychological Sciences and Social Sciences* **52**(2): S59–71

Fogel R 1994 Economic growth, population theory, and physiology: The bearing of long-term processes on the making of economic policy. *American Economic Review* **84**(3): 369–95
Freedman V A, Martin L G 1998 Understanding trends in functional limitations among older Americans. *American Journal of Public Health* **88**(10): 1457–62
Freedman V A, Soldo B J 1994 *Trends in Disability at Older Ages*. National Academy Press, Washington, DC
Fries J F 1980 Aging, natural death, and the compression of morbidity. *NEJM* **303**: 130–5
Gabrel C S 2000 Characteristics of elderly nursing home current residents and discharges: Data from the1997 National Nursing Home Survey. *Advance Data for Vital and Health Statistics*, No. 312. National Center for Health Statistics, Hyattsville, MD
Gruenberg R 1977 The failure of success. *Milbank Quarterly* **55**: 3–24
Jacobzone S 1999 An overview of international perspectives in the field of ageing and care for frail elderly persons. Labour Market and Social Policy Occasional Papers 38. OECD, Paris
Kramer M 1980 The rising pandemic of mental disorders and associated chronic diseases and disabilities. *Acta Psychiatrica Scandinavica* **62**(Suppl.285): 382–97
Manton K G 1982 Changing concepts of morbidity and mortality in the elderly population. *Milbank Quarterly* **60**: 183–244
Manton K G, Corder L, Stallard E 1997 Chronic disability trends in elderly United States populations 1982 to 1994. *Proceedings of the National Academy of Sciences of the USA* **94**: 2593–8
Manton K G, Singer B H 2001 Variation in disability decline and Medicare expenditures. *Proceedings of the National Academy of Sciences of the USA* in press
McGeer P L, Schulzer M, McGeer E G 1996 Arthritis and anti-inflammatory agents as possible protective factors for Alzheimer's disease: A review of 17 epidemiologic studies. *Neurology* **47**: 425–32
Perutz M 1998 And they all lived happily ever after. *The Economist* **February 7**: 82–3
Preston S 1992 Cohort succession and the future of the Oldest Old. In: Suzman R, Willis D, Manton K (eds.) *The Oldest Old*. Oxford University Press, New York, pp. 50–7
Riley M W, Bond K 1983 Beyond ageing: Postponing the onset of disability. In: Riley M W, Hess B, Bond K (eds.) *Aging and Society: Selected Reviews of Recent Research*. Lawrence Erlbaum Associates, Hillsdale, NJ
Tang M X, Jacobs D, Stern Y, Marder K, Schofield P, Gurland B, Andrews H, Mayeux R 1996 Effect of oestrogen during menopause on risk and age at onset of Alzheimer's disease. *Lancet* **348**(9025): 429–32
Verbrugge L 1984 Longer life but worsening health? Trends in health and mortality of middle-aged and older persons. *Milbank Quarterly* **62**: 475–519
Waidmann T A, Liu K 1998 *Disability Trends among the Elderly and Implications for Future Medicare Spending*. Joint Statistical Meetings, Dallas, TX
Waidmann T, Manton K G 1998 International evidence on disability trends among the elderly. Final Report for the Department of Health and Human Services
World Health Organization, Scientific Group on the Epidemiology of Aging 1984 The uses of epidemiology in the study of the elderly. Report of a WHO Scientific Group on the Epidemiology of Aging. Technical Report Series 706, WHO, Geneva, Switzerland

K. G. Manton

Aging Mind: Facets and Levels of Analysis

1. *A Zeitgeist in Search for Interdiscipline Integration*

Throughout most of the twentieth century much of the basic research on cognition has progressed in a rather segregated fashion, with differences in experimental paradigms, methodological, and theoretical orientations together with traditional discipline boundaries setting the dividing lines. Disintegrated research pursuits as such are common, as most endeavors in early stages of research development are first devoted to the discoveries of unique new phenomena and the constructions of competitive theoretical interpretations. As the field progresses with ever increasing empirical data and theories, integration is then necessary to provide a comprehensive understanding of the accumulated information.

1.1 *Proposals to Integrate the Studies of Brain, Cognition, and Behavior*

The need for developing overarching integrations across the many subfields of cognitive psychology and cognitive science became evident in the last decade of the twentieth century. Approaches for integrating the studies of brain, cognition, and behavior have been independently proposed by researchers of different specialization (Fig. 1 shows a summary diagram). For instance, researchers in the area of artificial intelligence have proposed cross-domain integration aiming at constructing comprehensive models to capture different domains of cognitive and behavioral functioning such as perception, memory, learning, decision-making, emotion, and motivation (e.g., Newell 1990). There is also the cognitive and computational neuroscience approach of cross-level integration, which aims at integrating empirical regularities and theories of cognition across the behavior, information-processing, and biological levels (see Gazzaniga 2000 for review). Others, built on Burnswik's and Gibson's earlier emphases on the embeddedness of behavior and cognition in the environmental context, have suggested a human–ecology integration stressing that functional adaptivity arising from the human–environment interaction must be considered en route to discoveries of universal principles of behavior and cognition. (e.g., Gigerenzer et al. 1999, Shepard 1995). In order to better capture dynamic exchanges between environmental support and biological resources across the lifespan, developmental psychologists (e.g., Baltes et al. 1999) have advocated a lifespan integrative approach to study behavior and cognition (see also *Lifespan Theories of Cognitive Development*). Although these approaches differ in the questions they address, they complement, rather than exclude, each other,

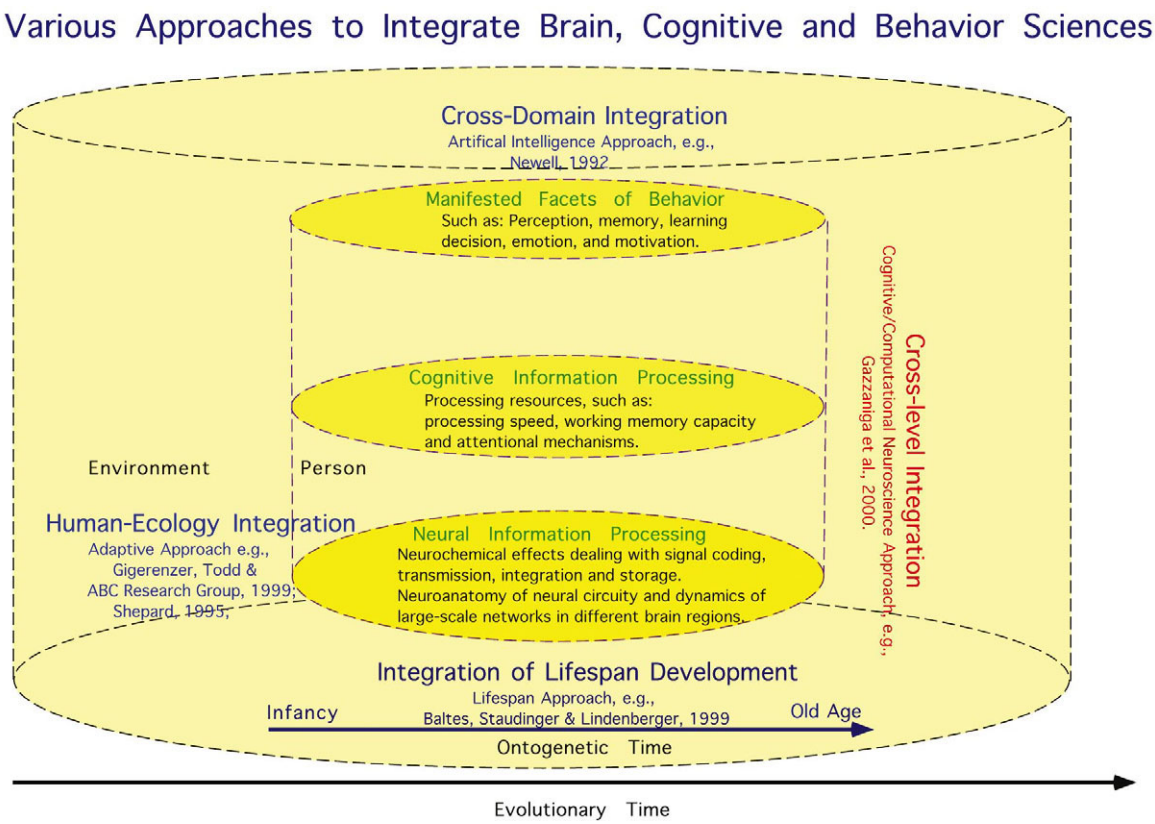

Figure 1
A summary diagram of different approaches proposed in the 1990s for integrating the various fields of brain, cognitive, and behavioral sciences

with the first two approaches focusing on different domains and levels of cognition and behavior within a person, and the last two focusing on the person–environment interaction and the evolutionary–ontogenetic dynamics.

1.2 Cognitive Aging Phenomena Studied at Various Levels

Couched within the broader research context, the field of cognitive aging had also gone through a period of disintegrated research and is now orienting towards integration. Since the 1920s when the first studies on adult age differences in mental abilities were published, studies on cognitive aging have mostly been carried out independently by individual difference and cognitive experimental psychologists and neuroscientists at the behavioral, information-processing, and biological levels (Fig. 2 gives an overview). Designs and results from animal neurobiological studies are not always readily testable in human cognitive studies, and vice versa. Therefore, until the recent advances with neuroimaging techniques (Cabeza 2001), data and theories of cognitive aging have been mostly confined within their respective levels. The goal of this article is thus to review evidence of age differences in intelligence and basic cognitive processing in ways that highlight the many facets of the aging mind and point out some recent attempts that have been undertaken since the 1990s to link previously less integrated areas of research.

2. Adult Age Differences in Intelligence

At the behavioral level, psychologists interested in how aging might affect individual differences in intelligence have taken the psychometric approach, which has a long tradition (dating back to classical works by Spearman, Galton, and Binet in the 1880s and early 1900s), and focused on the measurement of age differences in intellectual abilities. The existing psychometric data to date indicate that intellectual

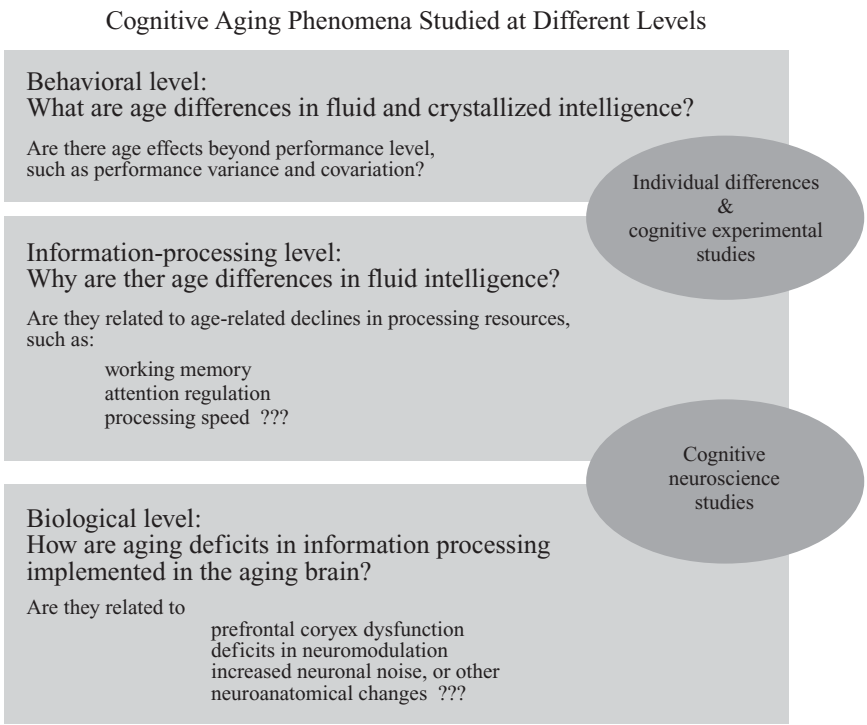

Figure 2
A summary diagram of various issues of aging mind addressed by researchers of different specializations at various levels

aging is multifaceted. Furthermore, aging effects can be observed at three aspects of the behavioral data, namely, performance level, variability, and co-variation.

2.1 Differential Age-gradients of Cognitive Mechanics and Pragmatics

Traditionally, two-component models of intelligence distinguish between fluid intelligence reflecting the operations of neurobiological 'hardware' supporting basic information-processing cognitive mechanics and crystallized intelligence reflecting the culture-based 'software' constituting the experience-dependent cognitive pragmatics (Baltes et al. 1999, Horn 1982; see also *Lifespan Theories of Cognitive Development*). Figure 3 shows that the fluid mechanics such as reasoning, spatial orientation, perceptual speed, and verbal memory show gradual age-related declines starting at about the 40s, while other abilities indicating the crystallized pragmatics such as number and verbal abilities remain relatively stable up until the 60s (e.g., Schaie and Willis, 1993). Furthermore, there have also been some recent ongoing theoretical and empirical efforts devoted towards expending the concepts of cognitive mechanics and pragmatics. In addition to the efficacy of information processing, cognitive mechanics also encompasses the optimal allocation of cognitive resources (e.g., Li et al. in press). Cognitive pragmatics has been expanded to include many other general as well as person-specific bodies of knowledge and expertise associated with the occupational, leisure, and cultural dimensions of life (e.g., Blanchard-Fields and Hess 1996). One example is wisdom, the 'expert knowledge about the world and fundamental pragmatics of life and human affair' that an individual acquires through his or her life history, that also includes an implicit orientation towards maximizing individual and collective well-being (Baltes and Staudinger 2000).

2.2 Age-related Increase in Variability and Co-variation

In addition to age differences in the performance levels of the cognitive mechanics, behavioral data also point to age-related increases of performance variations within a person (e.g., Hultsch et al. 2000) and differences between individuals (for review see Nelson and Dannefer 1992). Furthermore, much cross-sec-

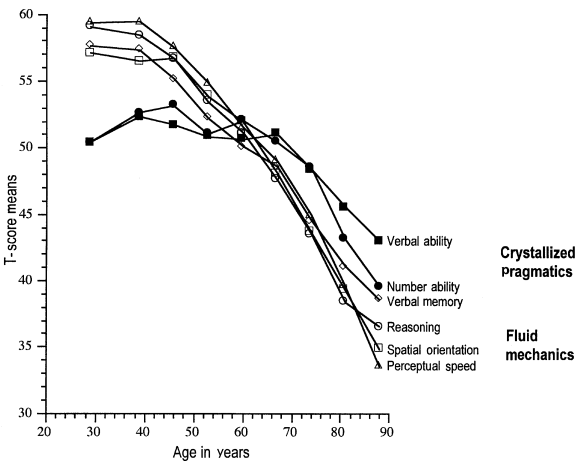

Figure 3
Differential trajectories of fluid (mechanic) and crystallized (pragmatic) intelligence. Abilities were assessed with 3–4 different tests, and were scaled in a T-score metric (data source based on Schaie and Willis 1993; figure adapted from Lindenberger and Baltes, 1994 with permission)

tional data show that as people age, performances of different subscales of intelligence tests become more correlated with each other (e.g., Babcock et al. 1997), which has been taken as indications of less differentiated ability structure in old people.

3. *Deficits in Basic Information-processing Mechanisms*

In the light of age-related declines in psychometrically measured cognitive mechanics, the information-processing approach emerged from the rise of information theory and computers in the 1940s and was advanced to explain age differences in fluid intelligence by identifying age differences in basic information-processing mechanisms. Thus far, age-related declines have been found in three main facets of information processing: people's abilities to keep information in mind, attend to relevant information, and process information promptly are compromised with age.

3.1 *Working Memory*

Working memory (WM) refers to people's ability to simultaneously hold information in immediate memory while operating on the same or other information. Age-related decline in WM capacity has been obtained on a variety of experimental tasks including backward digit span, sentence span, and several types of computational span (e.g., Park et al. 1996; see Fig. 4(A)). Age-related decline in WM capacity plays a role in many other cognitive activities where WM is implicated, ranging from long-term memory encoding and retrieval, syntactic processing, language comprehension and reasoning (for review see Zacks et al. 2000; see also *Memory and Aging, Cognitive Psychology of*).

3.2 *Attentional and Inhibitory Mechanisms*

Empirical data abound showing that old people have more problems attending to relevant information and ignoring irrelevant information. Negative age differences have been found in various selective and focus attention tasks along with the Stroop and proactive interference tasks (see Fig. 4(B)). Age-related declines in attentional and inhibitory mechanisms have functional consequences on language comprehension, memory, problem solving, and other daily activities such as driving (see McDowd and Shaw 2000 for review).

3.3 *Processing Speed*

Speed is a ubiquitous aspect of information processing. All information processing takes time, however, fast it is. There is abundant evidence showing that older people are slower in responding compared to young adults in almost every cognitive task in which processing speed is measured (see Fig. 4(C)). Many correlational analyses showed that the observed age differences in fluid intelligence are greatly reduced or eliminated after controlling for individual differences in processing speed (see Salthouse 1996 for review).

3.4 *Resource-reduction Account*

Given clear age-related declines in these fundamental aspects of information-processing, the most prominent account of cognitive aging deficits thus far has been the general conceptual framework of age-related reduction in processing resources that are indicated by working memory capacity, attention regulation, and processing speed (see Salthouse 1991 for review). However, two major difficulties limit the resource-reduction theory. First, the different aspects of processing resources are not independent of each other. Second, the account itself is circular in nature: old people's lower proficiency in cognitive performance is assumed to be caused by a reduction in processing resources, and at the same time, poor performance is taken to be the indication of reduced processing resources. One way to avoid such circularity is to establish better correspondence between the proposed processing resources and their potential neurobiological underpinnings. Lest this be viewed only as

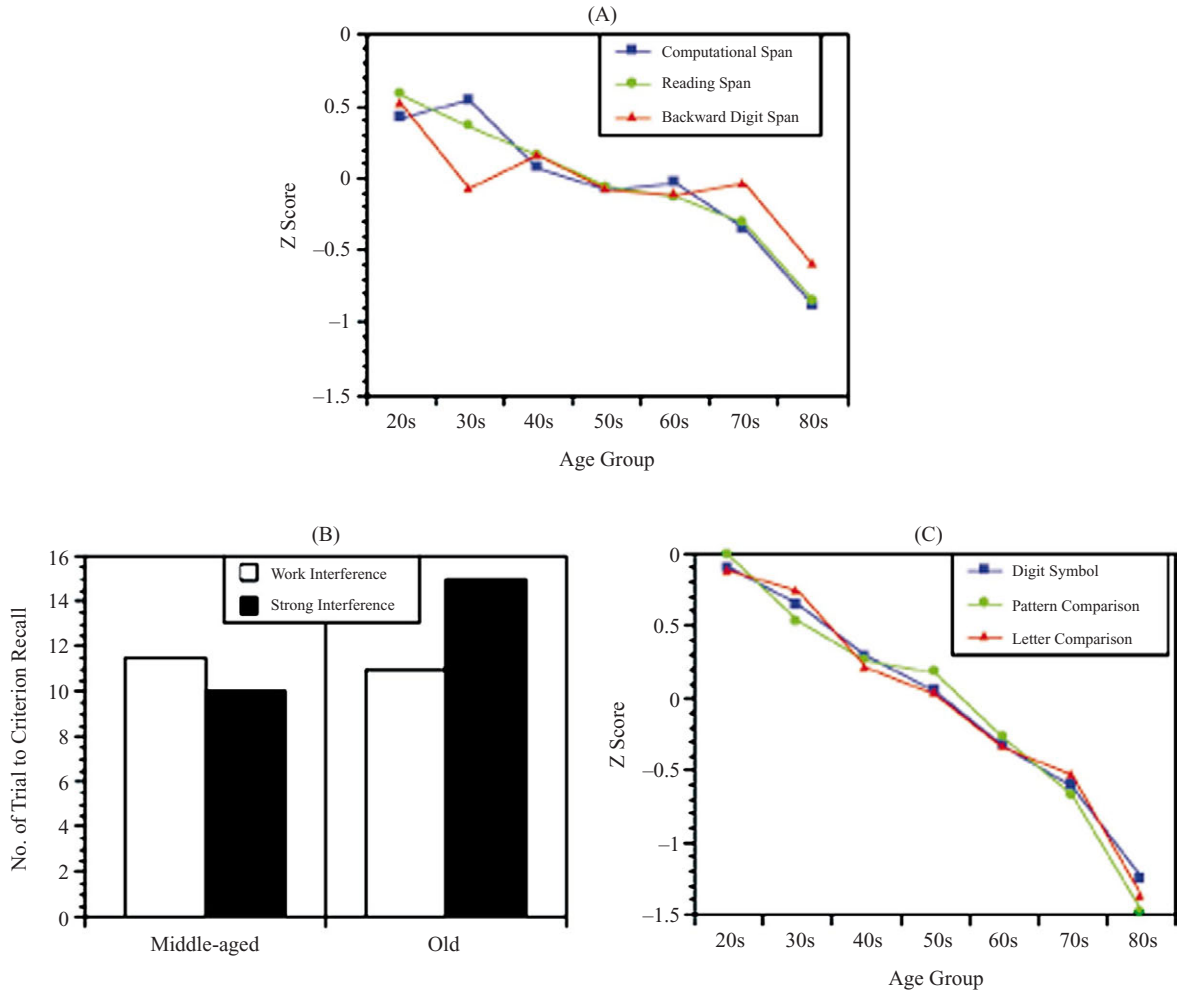

Figure 4
Negative adult age differences in working memory, proactive interference and processing speed. (A) Working memory was measured by three types span test, and were scaled in Z score metric (Data source based on Park et al., 1996. Copyright © 1996 American Psychological Association. Adapted with permission.) (B) Old adults (mean age = 64.4) required more trials to learn arbitrary word pairs than middle-aged adults (mean age = 38.8), when proactive interference was strong (data source based on Lair et al., 1969. Copyright © 1969 American Psychological Association. Adapted with permission). (C) Processing speed was measured by three perceptual speed tests (i.e., digit symbol substitution, pattern and letter comparison), and were scaled in a Z-score metric (Data source based on Park et al., 1996. Copyright © 1996 American Psychological Association. Adapted with permission.)

reductionistic, it should be mentioned that psychometric data showing stronger trends of age-related decline in biology-based fluid intelligence motivates the search for biological correlates. Experimental evidence of age-related decline in basic facets of information processing helps to focus the studies of brain aging on those aspects relevant to the affected information-processing mechanisms. Recent developments in cognitive and computational neurosciences have opened new avenues for studying the functional relationships between behavioral manifestations of the aging mind and the biology of the aging brain.

4. *The Aging Brain of the Aging Mind*

At the neurobiological level, brain aging involves both neuroanatomical and neurochemical changes. Anatomically, there are structural losses in neurons and

synaptic connections and reductions in brain atrophy (see Raz 2000 for review). Neurochemically, there is evidence for deterioration in various neurotransmitter systems (see Schneider et al. 1996 for review). However, progressive neuroanatomical degeneration resulting from cell death and reduced synaptic density is primarily characteristic of pathological aging such as Alzheimer's disease, and there is now evidence suggesting that milder cognitive problems occurring during normal aging are mostly due to neurochemical shifts in still-intact neural circuitry (Morrison and Hof 1997).

4.1 *Attenuated Neuromodulation*

Among different neurotransmitter systems, the catecholamines, including dopamine (DA) and norepinephrine (NE), are important neurochemical underpinnings of age-related cognitive impairments for several reasons. First, there is consensus for age-related decline in catecholaminergic function in the prefrontal cortex (PFC) and basal ganglia. Across the adult lifespan, dopaminergic function in the basal ganglia decreases by 5–10 percent each decade (see Schneider et al. 1996). Furthermore, many DA pathways in the basal ganglia are interconnected with the frontal cortex through the frontal–striatal circuits (Graybiel 1990), hence are in close functional association with the PFC cognitive processes. Second, research over the last two decades suggests that catecholamines modulate the PFC's utilization of briefly activated cortical representations of external stimuli to circumvent constant reliance on environmental cues and to regulate attention to focus on relevant stimuli and appropriate responses (see Arnsten 1998 for review). Third, there are many findings indicating specific functional relationships between age-related deficits in the dopaminergic system and deficits in various aspects of information processing. For instance, reduced dopamine receptor density in old rats' nigrostriatum decreases response speed and increases reaction time variability (MacRae et al. 1988). Drugs that facilitate dopaminergic modulation alleviate working memory deficits of aged monkeys who suffer from 50 percent dopamine depletion in their PFC (see Arnsten 1998 for review). In humans, age-related attenuation of dopamine D2 receptor's binding mechanism is associated with declines in processing speed and episodic memory (Bäckman et al. 2000).

4.2 *Reduced Hemispheric Asymmetry*

In addition to changes in the aging brain's neurochemical environment, recent neuroimaging evidence suggests that cortical information processing in different regions of the brain becomes less differentiated as people age, phenomena that parallel the behavioral findings of less differentiated ability structure in old people. In comparison to the more clearly lateralized cortical information processing in young adults, people in their 60s and beyond showed bilateralized (bi-hemispheric) activity during retrieval (e.g., Cabeza et al. 1997, Cabeza 2001) and during both verbal and spatial work memory tasks (Reuter-Lorenz et al. 2000).

5. *Outlooks for Integrating the Facets and Levels of the Aging Mind*

Faced with the various facets of the aging mind across the different levels, the various subfields of cognitive aging research are ever more inclined to and in need of overarching frameworks for integration (cf. Stern and Carstensen 2000). Some integrative research undertakings along the four general approaches for integrating the studies of brain, cognition, and behavior sciences are already underway.

With respect to better integrating the human–environment exchange and the evolutionary–ontogenetic dynamics, at a macrolevel some researchers embedded issues of cognitive aging within a metatheoretical framework of biological and cultural coevolution for studying lifespan human development. While the benefits of evolutionary selection and the efficacy of neurobiological implementations of the mind decrease with aging, the need for environmental–cultural support increases. In this systemic functional framework it is important for future research to investigate how declines in cognitive resources may be compensated for by the individual's more selective allocations of these resources to different task domains and by cultural–environmental supports such as cognitive training (e.g., Dixon and Bäckman 1995, Li et al. in press). At a more specific level, other researchers have also suggested an environmental support perspective for understanding age differences in episodic memory and attentional mechanisms. Better environmental stimulus and contextual supports are helpful for overcoming age-related deficits in effortful self-initiated processes implicated in various memory and attentional tasks (e.g., Craik 1986, Park and Shaw 1992).

Regarding better integrating different domains and levels of behavior and cognition within the person, some researchers have started to work towards bridging the gaps between age-related declines in basic memory and attentional processes and higher-level cognitive function such as language comprehension (e.g., Light and Burke 1988, Burke 1997). Regarding cross-level integration, there actually have been a few classical proposals trying to relate individual differences in the performance level, variance, and covariation of intellectual functioning with individual differences in general brain energy (Spearman 1927) and to link age-related cognitive aging deficits with

increased neuronal noise (e.g., Welford 1965). However, these long-range brain–behavior links could not be specified with more details in early research. It has only recently become possible to investigate these links more explicitly in cognitive and computational neurosciences. There is now some consensus for associations between PFC dysfunctions and aging-related cognitive impairments (West 1996). However, details of the functional relationships between PFC impairments, aging attenuated neuromodulation, the distribution of information processing across different neural circuitry, and various behavioral manifestations of cognitive aging deficits await further explications. Recently, one computational neuroscience approach has been undertaken to explore the links between age-related declines in neuromodulatory mechanisms innervating the PFC, noisier neural information processing, and adult age differences in episodic memory, interference susceptibility, performance variability, and co-variation (e.g., Li et al. 2000, Li in press).

These integrative research orientations have different advantages and disadvantages. While theoretical considerations about environmental and evolutionary impacts on the aging mind at the metatheoretical level have the strength in providing overarching organizations, they need to be complemented by more information-processing and neurobiologically oriented approaches to generate predictions that are more amenable for direct empirical validations. In the process of co-evolving a range of related fields, there may not be a 'right' level for integration; rather there is the task to supplement and balance the weaknesses and strengths of different approaches.

6. Conclusion

The average life expectancy in most industrialized countries has been increasing from an average of about 45 years in 1900 to about 75 years in 1995. A major task for cognitive aging research in the twenty-first century is to identify the causes of and methods to minimize or compensate for these cognitive declines, so that the blessings of improved physical health and extended life expectancy in old age could be accompanied by sound aging mind. Attempts to achieve this challenging task require collective contributions of studies from the various subfields of cognitive aging research, ranging from individual-difference based psychometric and behavioral experimental studies to cognitive and computational neurosciences in the future. Furthermore, research on the aging mind necessarily entails an applied orientation; therefore, future research also needs to include more specific focuses on identifying age-relevant knowledge, aging-friendly social and environmental contexts, and aging-rectifying training programs to help old people better allocate and compensate their declining cognitive resources.

See also: Aging and Health in Old Age; Aging, Theories of; Artificial Neural Networks: Neurocomputation; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Cognitive Aging; Computational Neuroscience; Differential Aging; Lifespan Theories of Cognitive Development; Memory and Aging, Cognitive Psychology of; Memory and Aging, Neural Basis of; Neural Networks: Biological Models and Applications; Old Age and Centenarians; Psychometrics; Recovery of Function: Dependency on Age; Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

Bibliography

Arnsten A F T 1998 Catecholamine modulation of prefrontal cortical cognitive function. *Trends in Cognitive Science* **2**: 436–47

Babcock R L, Laguna K D, Roesch S C 1997 A comparison of the factor structure of processing speed for younger and older adults: Testing the assumption of measurement equivalence across age groups. *Psychology and Aging* **12**: 268–76

Bäckman L, Ginovart N, Dixon R, Wahlin T, Wahlin A, Halldin C, Farde L 2000 Age-related cognitive deficits mediated by changes in the striatal dopamine system. *American Journal of Psychiatry* **157**: 635–7

Baltes P B, Staudinger U M 2000 A metaheuristic (pragmatic) to orchestrate mind and virtue toward excellence. *American Psychologist* **55**: 122–36

Baltes P B, Staudinger U, Lindenberger U 1999 Lifespan psychology: Theory and application to intellectual functioning. *Anuual Review of Psychology* **50**: 471–507

Blanchard-Fields F, Hess T M (eds.) 1996 *Perspectives on Cognitive Change in Adulthood and Aging*. McGraw-Hill, New York

Burke D M 1997 Language, aging, and inhibitory deficits: Evaluation of a theory. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences* **52B**: 254–64

Cabeza R 2001 Functional neuroimaging of cognitive aging. In: Cabeza R, Kingstone A (eds.) *Handbook of Functional Neuroimaging of Cognition*. MIT Press, Cambridge, MA

Cabeza R, Grady C L, Nyberg L, McIntosh A R, Tulving E, Kapur S, Jennings J M, Houle S, Craik F I M 1997 Age-related differences in effective neural connectivity. *Neuroreport* **8**: 3479–83

Craik F I M 1986 A functional account of age differences in memory. In: Klix F, Hagendorf H (eds.) *Human Memory and Cognitive Capabilities: Mechanisms and Performances*. Elsevier, Amsterdam, pp. 409–22

Dixon R A, Bäckman L (eds.) 1995 *Compensating for Psychological Deficit and Declines: Managing Losses and Promoting Gains*. LEA, Hillsdale, NJ

Gazzaniga M S (ed.) 2000 *Cognitive Neuroscience: A Reader*. Blackwell Publishers, Malden, MA

Gigerenzer G, Todd P, the ABC Research Group 1999 *Simple Heuristics that Make Us Smart*, Oxford University Press, New York

Graybiel A M 1990 Neurotransmitters and neuromodulators in the basal ganglia. *Trends in Neurosciences* **13**: 244–53

Horn J L 1982 The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in

adulthood. In: Craik F I M, Trehub S (eds.) *Aging and Cognitive Processes*. Plenum, New York, pp. 237–78
Hultsch D F, MacDonald S W S, Hunter M A, Levy-Bencheton J, Strauss E 2000 Intraindividual variability in cognitive performance in older adults: Comparison of adults with mild dementia, adults with arthritis, and healthy adults. *Neuropsychology* **14**: 588–98
Lair C V, Moon W H, Klauser D H 1969 Associative interference in the paired-associative learning of middle-aged and old subjects. *Developmental Psychology* **5**: 548–52
Li K Z H, Lindenberger U, Freud A M, Baltes P B in press Walking while memorizing: A SOC study of age-related differences in compensatory behaviour under dual-task conditions. *Psychological Science*
Li S-C in press Connecting the many levels and facets of cognitive aging. *Current Directions in Psychological Science*
Li S-C, Lindenberger U, Frensch P A 2000 Unifying cognitive aging: From neuromodulation to representation to cognition. *Neurocomputing* **32–33**: 879–90
Light L L, Burke D M 1988 Patterns of language and memory in old age. In: Light L L, Burke D M (eds.) *Language, Memory and Aging*. Cambridge University Press, New York, pp. 244–72
Lindenberger U, Baltes P B 1994 Aging and intelligence. In: Sternberg et al. (eds.) *Encyclopedia of Intelligence*. Macmillan, New York, pp. 52–66
MacRae P G, Spirduso W W, Wilcox R E 1988 Reaction time and nigrostriatal dopamine function: The effect of age and practice. *Brain Research* **451**: 139–46
McDowd J M, Shaw R J 2000 Attention and aging. A functional perspective. In: Craik F I M, Salthouse T A (eds.) *The Handbook of Aging and Cognition*. LEA, Mahwah, NJ, pp. 221–92
Morrison J H, Hof P R 1997 Life and death of neurons in the aging brain. *Science* **278**: 412–29
Nelson E A, Dannefer D 1992 Aged heterogeneity: Facts or fictions? The fate of diversity in gerontological research. *Gerontologist* **32**: 17–23
Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA
Park D C, Smith A D, Lautenschlager G, Earles J L 1996 Mediators of long-term memory performance across the lifespan. *Psychology and Aging* **4**: 621–37
Park D C, Shaw R J 1992 Effects of environmental support on implicit and explicit memory in younger and older adults. *Psychology and Aging* **7**: 632–42
Raz N 2000 Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In: Craik F I M, Salthouse T A (eds.) *The Handbook of Aging and Cognition*. LEA, Mahwah, NJ, pp. 1–90
Reuter-Lorenz P A, Jonides J, Smith E, Marshuetz C, Miller A, Hartley A, Koeppe R 2000 Age differences in the frontal lateralization of verbal and spatial working memory revealed by PET. *Journal of Cognitive Neuroscience* **12**: 174–87
Salthouse T A 1991 *Theoretical Perspectives of Cognitive Aging*. LEA, Hillsdale, NJ
Salthouse T A 1996 The processing-speed theory of adult age differences in cognition. *Psychological Review* **103**: 403–28
Schaie K W, Willis S L 1993 Age difference patterns of psychometric intelligence in adulthood: Generalizability within and across ability domains. *Psychology and Aging* **8**: 44–55
Schneider E L, Rowe J W, Johnson T E, Holbrook N J, Morrison J H 1996 *Handbook of the Biology of Aging*, 4th edn. Academic Press, New York
Shepard R N 1995 Mental universals: Toward a 21st century science of mind. In: R L, Solso Massaro D W (eds.) *The Science of the Mind: 2001 and Beyond*. Oxford University Press, New York, pp. 50–64
Spearman C E 1927 *The Abilities of Man*. MacMillan, New York
Stern P C, Carstensen L L 2000 *The Aging Mind: Opportunities in Cognitive Research*. National Academy Press, Washington, DC
Welford A T 1965 Performance, biological mechanisms and age: A theoretical sketch. In: Welford A T, Birren J E (eds.) *Behavioral, Aging and the Nervous System*. Thomas, Springfield, IL, pp. 3–20
West R L 1996 An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin* **120**: 272–92
Zacks R T, Hasher L, Li K Z H 2000 Human memory. In: Craik F I M, Salthouse T A (eds.) *The Handbook of Aging and Cognition*. LEA, Mahwah, NJ, pp. 293–357

S.-C. Li

Aging, Theories of

Because theories of aging in the behavioral and social sciences have come from a variety of disciplines it is often difficult to distinguish between formal theoretical frameworks and theoretical models that seek to systematize sets of empirical data. This article, therefore, will discuss current thought on theory building in aging, and then summarize exemplars of theoretical frameworks that inform the field origination from biology, psychology and the social sciences.

1. *Theory Building in Aging*

1.1 *Historical Development of Theories of Aging*

Early gerontologists looked for conceptual frameworks that might explain human aging by looking at popular and ancient models, including the bible, Sanskrit, medieval allegories, other ancient texts and even archaeological evidence to explain individual differences in well being and maintaining competence through the various stages of life (e.g., Hall 1922). These early models of aging typically represent broad world views, such as the biblical admonition that obedience to God's commandments would ensure a long life. New historical contexts, however, result in new explanations of aging, whether of the medieval explanation of old women as witches or the modern conception of the biological advantages of female aging. But as in Hall's writings, they may also include critiques of contemporary societal arrangements.

More modern views of the complexity of aging may be found in Cowdry's classical opus *Problems of Aging* (1939). It contains a mixture of assertions that aging

resulted from 'degenerative diseases' to contentions that social context affected the expression of aging and could lead to the difference between what Rowe and Kahn (1997) have referred to as the difference between 'normal' and 'successful' aging. As scientific insights on the aging process have accumulated during the twentieth century, a movement has occurred from broad world views on aging to more circumscribed theoretical models that are driven by disciplinary perspectives but also by the fads and explanatory frameworks that have waxed and waned in the scientific enterprise (cf. Hendricks and Achenbaum 1999).

1.2 Models and Explanation

Distinctions must be made between theories and other aspects of knowledge development. As a first stage, we find statements describing regularities detected in the process of systematic observations. A second stage is represented by prototypical models that attempt to depict how empirical generalizations are related to each other. A third stage may be characterized by the term 'paradigm' which implies a shift in scientific efforts represented by the accumulation of empirical generalizations, models, and theories. In contrast to these terms, which are of course also important for knowledge development, the focus of a theory should be upon the construction of explicit explanations that account for empirical findings (cf. Bengtson et al. 1999).

1.3 Theory Development and Research Design in Aging

Theory development in aging has been impacted markedly by advances in research design. One of the early impacts was the development of the age–period cohort model which required theory development to distinguish between age changes (measured longitudinally) and age differences (measured cross-sectionally). The distinction of within-subject maturational effects and between-subjects cohort differences has also informed theory development. In addition, the advent of restrictive factor analysis and structural equation modeling has made it possible to provide empirical tests of structural relationships in various domains that tend to change across time–age and differ across groups (cf. O'Rand and Campbell 1999, Schaie 1988).

2. Biological Theories of Aging

2.1 Biological Theories of Senescence

Theories explaining the biological basis of human aging are either stochastic theories that postulate senescence to be primarily the result of random damage to the organism, or they are programmed theories that hold that senescence is the result of genetically determined processes. Currently most popular theories include: (a) the free radical theory, which holds that various reactive oxygen metabolites can cause extensive cumulative damage; (b) caloric restriction, which argues that both lifespan and metabolic potential can be modified by caloric restriction (thus far not demonstrated in humans); (c) somatic mutation, arising from genetic damage originally caused by background radiation; (d) hormonal theories, proposing, for example, that elevated levels of steroid hormones produced by the adrenal cortex can cause rapid aging decline; and (e) immunological theories that attribute aging to decline in the immune system. Another prominent view is that the protective and repair mechanisms of cells are insufficient to deal with the cumulative damage occurring over time, limiting the replicative ability of cells (cf. Cristofalo et al. 1999, Hayflick 1994).

2.2 Stress Theories of Aging

These theories argue that excessive physiological activation have pathological consequences. Hence differences in neuroendocrine reactivity might influence patterns of aging. The focus of such theories is not on specific disease outcomes, but rather on the possibility that neuroendocrine reactivity might be related generally to increased risk of disease and disabilities. Stress mechanisms are thought to interact with age changes in the hypothalamic–pituitary–adrenal (HPA) axis, which is one of the body's two major regulatory systems for responding to stressors and maintaining internal homeostatic integrity. Individual differences in reactivity may cumulatively lead to major individual differences in neuroendocrine aging as well as age-related risks for disease. Certain psychosocial factors can influence patterns of endocrine reactivity. Perceptions of control and the so-called Type A behavior pattern may influence increased reactivity with age. Gender differences in neuroendocrine reactivity are also posited because of the known postmenopausal increase in cortisol secretion in women not treated with estrogen replacement therapy (cf. Finch and Seeman 1999).

3. Psychological Theories of Aging

As for other life stages, there do not seem to be many overarching theories of psychological aging, but emphasis in theoretical development is largely confined to a few substantive domains. A recent exception to this observation is the theory of selection, optimization and compensation (SOC) advocated by P. Baltes (1997, Baltes and Baltes 1990). This theory suggests that there are psychological gains and losses at every

life stage, but that in old age the losses far exceed the gains. Baltes suggests that evolutionary development remains incomplete for the very last stage of life, during which a societal supports no longer suffice to compensate for the decline in physiological infrastructure and losses in behavioral functionality (cf. Baltes and Smith 1999).

3.1 Theories of Cognition

A distinction is generally made between cognitive abilities that are fluid or process abilities that are thought to be genetically overdetermined and which (albeit at different rates) tend to decline across the adult lifespan, and crystallized or acculturated abilities that are thought to be learned and be culture-specific, and which tend to be maintained into advanced old age. This distinction tends to break down in advanced old age as declining sensory capacities and reduction in processing speed also leads to a decline of crystallized abilities. Nevertheless, most theories of adult cognition have focused upon explaining the decline of fluid abilities, neglecting to theorize why is it that crystallized performance often remains at high levels into late life.

Most theoretical perspectives on cognitive aging can be classified into whether the proposed primary causal influences are distal or proximal in nature. Distal theories attribute cognitive aging to influences that occurred at earlier periods in life but that contribute to concurrent levels of performance. Other distal explanations focus on social–cultural changes that might affect cognitive performance. These explanations assume cumulative cohort effects that lead to the obsolescence of the elderly. Distal theories are useful, particularly in specifying why the observed age differences have emerged, since it is generally agreed that mere passage of time can not account for these differences.

Proximal theories of aging deal with those concurrent influences that are thought to determine age-related differences in cognitive performance. These theories do not specify how the age differences originated. Major variations of these theories include strategy-based age differences, quantitative differences in the efficiency of information processing stages implicating deficits in specific stages, or the altered operation of one or more of the basic cognitive processes (cf. Salthouse 1999).

3.2 Theories of Everyday Competence

Theories of everyday competence seek to explain how an individual can function effectively on the tasks and within the situations posed by everyday experience. Such theories must incorporate underlying processes, such as the mechanics (or cognitive primitives) and pragmatics of cognitive functioning, as well as the physical and social contexts that constrain the individual's ability to function effectively. Because basic cognitive processes are typically operationalized to represent unitary trait characteristics, it is unlikely that any single process will suffice to explain individual differences in competence in any particular situation, Hence, everyday competence might be described as the phenotypic expression of combinations of basic cognitive processes that permit adaptive behavior in specific everyday situations.

Three broad theoretical approaches to the study of competence have recently been advocated. The first perspective views everyday competence as a manifestation of latent constructs that can be related to models of basic cognition (see also *Cognitive Aging*). The second approach conceptualizes everyday competence as involving domain-specific knowledge bases. In the third approach, the theoretical focus is upon the fit, or congruence, between the individual's cognitive competence and the environmental demands faced by the individual.

An important distinction must further be made of the distinction between psychological and legal competence. While the former is an important scientific construct, the latter refers to matters of jurisprudence that are involved in the imposition of guardianship or conservatorship designed to protect frail individuals as well as to limit their independent decision-making ability. Although legal theorizing incorporates aspects of virtually all psychological theories of competence, it does focus in addition the definition of cognitive functioning and competence as congruence of person and environment, upon the assignment of status or disabling condition and a concern with functional or behavioral impairment (cf. Schaie and Willis 1999).

3.3 Social–Psychological Theories

Social psychologists coming from a psychological background are concerned primarily with the behavior of individuals as a function of microsocial variables. Relying upon experimental or quasi-experimental designs, they seek to understand social phenomena using person-centered paradigms whose core is the structural and functional property of individual persons. Social–psychological approaches to aging have contributed to the understanding of numerous normal and pernicious age-related phenomena. There has been an increased interest in theoretical formulations that explain how social–psychological processes exert normative influences on life course changes. Included among theories that have received recent attention are control theories contrasting primary and secondary controls, coping theories that distinguish between accommodative and assimilative coping, and theories about age differences in attributive styles. There are also theories that blend psychological and sociological

approaches, such as the convoy theory and the support–efficacy theory.

Of particular recent interest has been the model of learned dependency (Baltes 1996). In this theory, the dependency of old age is not considered to be an automatic corollary of aging and decline, but rather is attributed in large part to be a consequence of social conditions. This theory contradicts Seligman's (1975) model of learned helplessness, which postulates dependency to be the outcome of noncontingencies and which sees dependency only as a loss. Instead it is argued that dependency in old people occurs as a result of different social contingencies, which include the reinforcement for dependency and neglect or punishment in response to pursuit of independence.

Also of currently prominent interest is socio–emotional selectivity theory. This theory seeks to provide an explanation of the well-established reduction in social interactions observed in old age. This theory is a psychological alternative to two previously influential but conflicting sociological explanations of this phenomenon. Activity theory considered inactivity to be a societally induced problem stemming from social norms, while the alternative disengagement theory suggested that impending death stimulated a mutual psychological withdrawal between the older person and society. By contrast, socio–emotional selectivity theory holds that the reduction in older persons' social networks and social participation should be seen as a motivated redistribution of resources by the elderly person. Thus older persons do not simply react to social contexts but proactively manage their social worlds (cf. Baltes and Carstensen 1999).

4. Sociological Theories of Aging

4.1 Anthropological Theories

Interest in old age came relatively late for anthropologists with an examination of ethnographic data in the *Human Relations Area Files* in 1945 that considered the role of the aged in 71 primitive societies. Early theoretical formulations propose a quasi-evolutionary theory that links the marginalization of older people to modernization. Current anthropological theorizing is informed by investigations of the contexts in which older adults are living that range from age-integrated communities to those in the inner city and in urban settings, as well as by the study of special populations that include various ethnic group and older people with disabilities. Common theoretical themes currently addressed include the complexity of the older population leading to differential experiences of aging in different cultural context, the diversity of aging within cultures, the role of context specificity, and the understanding of change over the life course across different cultural settings. Prevailing issues in anthropological theorizing on aging seem to focus first on how maturational differences are incorporated into a given social order, and second, the clarification of the variability as to how differences in maturity are modeled by human cultures in transforming maturation into ideas about age and aging.

Anthropological theories consider generational systems as fruitful ways of thinking about the life course. They argue that every human society has generational principles that organize social lives. Generations have little to do with chronological time, but rather designate position in a web of relationship; hence kinship systems are emphasized. Although age–class systems have explanatory power in primitive societies, they are not helpful as life course models in complex societies because of their variability. If anything, age–class systems are more likely to explain social structuring in males than in females. More useful for the understanding of complex societies seem to be models of staged life courses. Such models suggest that the life course in complex societies is based on combinations of generational and chronological age, and further is understood as staged or divisible into a variable number of age grades.

Anthropologists also distinguish between theories about age from those about aging or the aged. Theories about age explain cultural and social phenomena. That is, how is age used in the regulation of social life and the negotiation of daily living. Theories about aging are theories about living, the changes experienced during the life course, and the interdependencies throughout life among the different generations. Finally, theories about the age focus on late life, describing old age not only as a medical and economic problem but also as a social problem in terms of social support and care giving (cf. Fry 1999).

4.2 Life Course Theories

Life course theories represent a genuinely sociological approach to what, at the level of surface description, is a rather individual phenomenon as represented by the aging and life course patterning of human individuals. Much of this theorizing occurred subsequent to the recognition that individual aging occurred concurrently with the occurrence of social change, providing impetus to efforts trying to separate aging from cohort effects. Life course theories generally represent a set of three principles. First, the forms of aging and life course structures depend on the nature of the society in which individuals participate. Second, while social interaction is seen as having the greatest formative influence in the early part of life, such interaction retains crucial importance throughout the life course. Third, that social forces exert regular influences on individuals of all ages at any given point in time. However, such thinking also introduces three signifi-

cant intellectual problems. These are the tendency to equate the significance of social forces with social change, neglecting intracohort variability, and a problematic affirmation of choice as a determinant of the life course.

Life course phenomena can be treated at least at three levels of analysis. First, at the individual level, the structure of discrete human lives can be examined from birth to death. Second, one can examine the collective patterning of individual lives in a population. Third, it is possible to examine the societal representation of the life course in terms of the socially shared knowledge and demarkation of life events and roles. For each of these levels it is in turn possible to specify personological aspect that are thought to be part of the organism as well as the enduring contextual factors that were internalized at earlier life stages. But another crosscutting level involves the social–cultural and interactional forces that shape the life course (cf. Dannefer and Uhlenberg 1999).

4.3 Social Theories of Aging

Social theories of aging have often been devised to establish theoretical conflict and contrast. Two dimensions of contrast that have been used involve the cross-classification of normative versus interpretive theories and macro versus micro theories. But there are also intermediate theoretical perspectives that bridge these two approaches or that link different approaches. Modernization and aging theory would be an example of a normative macrotheory. Self and identity theories represent interpretive microtheories. Disengagement theory represents a normative linking theory, and the life course perspective discussed above represents a theory that is both linking and bridging (cf. Marshall 1999).

Recent generalizations that cut across most social theories seem to focus on three changes in the construction of the social phenomenon of aging. These changes suggest that life course transitions are decreasingly tied to age with a movement from age segregation to age integration. Second, that many life transitions are less disjunctive, more continuous, and not necessarily irreversible processes. Third, specific pathways in education, family, work, health, and leisure are considered to be interdependent within and across lives. Life trajectories in these domains are thought to develop simultaneously and reciprocally, rather than representing independent phenomena (O'Rand and Campbell 1999).

A prominent example of a social theory of aging is presented by the aging and society paradigm (Riley et al. 1999). The distinguishing features of this paradigm are the emphasis on both people and structures as well as the systemic relationship between them. This paradigm includes life course but it also includes the guiding principles of social structures as having greater meaning than merely providing a context for people's lives. This theory represents a cumulative paradigm. In its first phase, concerned with lives and structures, it began with the notion that in every society age organizes people's lives and social structures into strata from the youngest to the oldest, and raised questions on how age strata of people and age oriented structures arise and become interrelated. A second phase concerned with the dynamisms of age stratification defined changing lives and changing structures as interdependent but distinct sets of processes. The dynamism of changing lives began with the recognition of cohort differences and noted that because society changes, members of different cohorts will age in different ways. A second dynamism involves changing structures that redefine age criteria for successive cohorts.

In a third phase the paradigm specified the nature and implication of two connecting concepts, that of the interdependence and asynchrony of these two dynamisms, that attempt to explain imbalances in life courses as well as social homeostasis. A fourth phase deals with future transformation and impending changes of the age concepts. It introduces the notion of age integration as an extreme type of age structure as well as proposing mechanisms for cohort norm formation.

See also: Aging and Health in Old Age; Aging Mind: Facets and Levels of Analysis; Cognitive Aging; Differential Aging; Ecology of Aging; Indigenous Conceptions of Aging; Life Course in History; Old Age and Centenarians

Bibliography

Baltes M M 1996 *The Many Faces of Dependency in Old Age*. Cambridge University Press, New York

Baltes M M, Carstensen L L 1999 Social–psychological theories and their application to aging: from individual to collective. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 209–26

Baltes P B 1997 On the incomplete architecture of human ontogenesis: selection, optimization and compensation as foundations of developmental theory. *American Psychologist* **52**: 366–80

Baltes P B, Baltes M M (eds.) 1990 *Successful Aging: Perspectives from the Behavioral Sciences*. Cambridge University Press, New York

Baltes P B, Smith J 1999 Multilevel and systemic analyses of old age: theoretical and empirical evidence for a fourth age. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 153–73

Bengtson V L, Rice C J, Johnson M L 1999 Are theories of aging important? Models and explanation in gerontology at the turn of the century. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 3–20

Cowdry E V (ed.) 1939 *Problems of Aging*. Williams and Wilkins, Baltimore

Cristofalo V J, Tresini M, Francis M K, Volker C 1999 Biological theories of senescence. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 98–112
Dannefer D, Uhlenberg P 1999 Paths of the life course: a typology. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 306–26
Finch C E, Seeman T E 1999 Stress theories of aging. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 81–97
Fry C L 1999 Anthropological theories of age and aging. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 271–86
Hall G S 1922 *Senescence*. D Appleton's Sons, New York
Hayflick L 1994 *How and Why We Age*. 1st edn. Ballantine, New York
Hendricks J, Achenbaum A 1999 Historical development of theories of aging. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York pp. 21–39
Marshall V W 1999 Analyzing social theories of aging. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 434–58
O'Rand A M, Campbell R T 1999 On re-establishing the phenomenon and specifying ignorance: theory development and research design in aging. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 59–78
Riley M W, Foner A, Riley J W Jr 1999 The aging and society paradigm. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 327–43
Rowe J, Kahn R 1997 Successful aging. *The Gerontologist* **27**: 433–40
Salthouse T 1999 Theories of cognition. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 196–208
Schaie K W 1988 The impact of research methodology on theory-building in the developmental sciences. In: Birren J E, Bengtson V L (eds.) *Emergent Theories of Aging*. Springer, New York, pp. 41–58
Schaie K W, Willis S L 1999 Theories of everyday competence and aging. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 174–195
Seligman M E P 1975 *Helplessness: On Depression, Development, and Death*. Freeman, San Francisco

K. W. Schaie

Agnosia

Agnosia is a fascinating condition in which, as a consequence of acquired brain damage, patients lose the ability to recognize familiar stimuli, despite normal perception of those stimuli. For example, when encountering the faces of familiar persons such as family members or close friends, a patient with agnosia is unable to identify those persons, or even to recognize that they are familiar. A patient may look at pictures of entities such as animals or tools, and have no idea what the stimuli are. Or a patient may hear well-known sounds, such as a fire siren or a ringing phone, and not be able to identify the sounds or understand their meaning (despite being able to hear the sounds normally). Agnosia is a rare condition, and its clinical presentation borders on the bizarre; nonetheless, careful scientific study of agnosia has provided many important insights into the brain mechanisms important for learning, memory and knowledge retrieval.

1. Types of Knowledge and Levels of Knowledge Retrieval

Before discussing agnosia, it is important to explain some crucial differences in the types of knowledge that are processed by the brain, and how different task demands influence the mechanisms the brain uses to retrieve knowledge. To begin with, there is a dimension of specificity: knowledge can be retrieved at different levels of specificity, ranging from very specific to very general. Consider the following example: Knowledge about a unique horse ('Little Buck,' a sorrel roping horse) is specific and unique, and is classified at the subordinate level; less specific knowledge about horses (four-legged animals that gallop, used by cowboys; of which Little Buck is an example) is classified at the basic object level; and even less specific knowledge about living things (things that have life, of which horses and Little Buck are examples) is classified at the superordinate level.

Pragmatically, the level at which knowledge is retrieved depends on the demands of the situation, and those demands are different for different categories of entities. In everyday life, for example, it is mandatory that familiar persons be recognized at the unique level—e.g., that's 'President Clinton,' or that's 'my father Ned.' It is not sufficient, under most conditions, to recognize such entities only at more nonspecific levels—e.g., that's a 'world leader,' or that's 'an older man.' For other types of entities, recognition at the basic object level is sufficient for most purposes—e.g., that's a 'screwdriver,' or that's a 'stapler'; here, there is no need to recognize individual, unique screwdrivers and staplers in order for practical interactions with the entity to be productive.

One other critical distinction is between recognition, on the one hand, and naming, on the other. The two capacities are often confused. It is true that recognition of an entity, under normal circumstances, is frequently indicated by naming (e.g., 'stapler'; 'Little Buck'; 'siren'). However, there is a basic difference between knowing and retrieving the meaning of a concept (its functions, features, characteristics, relationships to other concepts), and knowing and retrieving the name of that concept (what it is called); moreover, this difference is honored by the brain. For example, brain damage in the left inferotemporal region can render a patient incapable of naming a wide variety of stimuli, while leaving unaffected the patient's ability to recognize those stimuli (H. Damasio et al. 1996). For the

examples of 'Little Buck' and 'siren' cited above, the patient may produce the descriptions of 'that's my sorrel roping horse that I bought two years ago and now lives on my dad's ranch,' and 'that's a loud sound that means there's an emergency; you should pull your car over to the side of the road.' Both responses indicate unequivocal recognition of the specific entities, even if their names are never produced. In short, it is important to maintain a distinction between recognition, which can be indicated by responses signifying that the patient understands the meaning of a particular stimulus, and naming, which may not, and need not, accompany accurate recognition (Caramazza and Shelton 1998, Gainotti et al. 1995, Pulvermuller 1999).

2. *The Term 'Agnosia'*

The term 'agnosia' signifies 'lack of knowledge,' and denotes an impairment of recognition. Traditionally, two types of agnosia have been described (Lissauer 1890). One, termed associative agnosia, refers to a failure of recognition that results from defective retrieval of knowledge pertinent to a given stimulus. Here, the problem is centered on memory: the patient is unable to recognize a stimulus (i.e., to know its meaning) despite being able to perceive the stimulus normally (e.g., to see shape, color, texture; to hear frequency, pitch, timbre; and so forth). The other type of agnosia is termed apperceptive, and refers to a disturbance of the integration of otherwise normally perceived components of a stimulus. Here, the problem is centered more on perception: the patient fails to recognize a stimulus because the patient cannot integrate the perceptual elements of the stimulus, even though those individual elements are perceived normally. It is important to emphasize that the nuclear feature in designating a condition as 'agnosia' is that there is a recognition defect that cannot be attributed simply or entirely to faulty perception.

The terms associative and apperceptive agnosia have remained useful, even if the two conditions do have some overlap. It is usually possible to classify a patient with a recognition impairment as having primarily a disturbance of memory (associative agnosia), or primarily a disturbance of perception (apperceptive agnosia). Not only does this classification have important implications for the management of such patients (e.g., what rehabilitation should be applied), but it also maps on to different sites of neural dysfunction. For example, in the visual modality, associative agnosia is strongly associated with bilateral damage to higher-order association cortices in the ventral and mesial occipitotemporal regions, whereas apperceptive agnosia is associated with unilateral or bilateral damage to 'earlier,' more primary visual cortices. This being said, though, the fact remains that separating associative and apperceptive agnosia can be difficult, which underscores the fact that the processes of perception and memory are not discrete. Rather, they operate on a physiological and psychological continuum, and it is simply not possible to demarcate a specific point at which perceptual processes end and memory processes begin (Damasio et al. 1990, Tranel and Damasio 1996).

In principle, agnosia can occur in any sensory modality, relative to any type of entity or event. In practice, however, some types of agnosia are considerably more frequent. Visual agnosia, especially agnosia for faces (prosopagnosia), is the most commonly encountered form of recognition disturbance. The condition of auditory agnosia is rarer, followed by the even less frequent tactile agnosia.

3. *Visual Agnosia*

3.1 *Definition*

Visual agnosia is defined as a disorder of recognition confined to the visual realm, in which a patient cannot arrive at the meaning of some or all categories of previously known nonverbal visual stimuli, despite normal or near-normal visual perception and intact alertness, attention, intelligence, and language. Typically, patients have impairments both for stimuli that they learned prior to the onset of brain injury (known as 'retrograde' memory), and for stimuli that they would normally have learned after their brain damage (known as 'anterograde' memory).

3.2 *Subtypes*

3.2.1 Prosopagnosia. The study of face processing has remained a popular topic in neuropsychology for many decades, dating back to the pioneering work of Bodamer, Hecaen, Meadows, and others (for historical reviews, see Benton 1990, De Renzi 1997). Faces are an intriguing class of stimuli (Damasio et al. 1982, Young and Bruce 1991). They are numerous and visually similar, and yet we learn to recognize individually as many as thousands of distinct faces during our lifetime; and not only can we learn many individual faces, but we can recognize them from obscure angles (e.g., from the side), attended with various artifacts (e.g., glasses, hockey helmet), after aging has radically altered the physiognomy, and under many other highly demanding conditions. Also, faces convey important and unique social and emotional information, providing clues about the emotional state of a person, or about potential courses of social behavior (e.g., approach or avoidance) (see Darwin 1872/1955, Adolphs et al. 1998). And there are a number of remarkable cross-cultural and cross-species consistencies in face processing (cf.

Ekman 1973, Fridlund 1994), which underscore the crucial and fundamental importance of this class of stimuli.

The inability to recognize familiar faces is known as prosopagnosia (face agnosia), and it is the most frequent and well established of the visual agnosias (Damasio et al. 1990, Farah 1990). The face recognition defect in prosopagnosia typically covers both the retrograde and anterograde compartments; respectively, patients can no longer recognize the faces of previously known individuals, and are unable to learn new ones. They are unable to recognize the faces of family members, close friends, and, in the most prototypical instances, even their own face in a mirror. Upon seeing those faces, the patients experience no sense of familiarity, no inkling that those faces are known to them, i.e., they fail to conjure up consciously any pertinent information that would constitute recognition. The impairment is modality-specific, however, being entirely confined to vision. For example, when a prosopagnosic patient hears the voices of persons whose faces were unrecognized, the patient will instantly be able to identify those persons accurately.

As noted above with regard to agnosia in general, prosopagnosia must be distinguished from disorders of naming, i.e., it is not an inability to name faces of persons who are otherwise recognized as familiar. There are numerous examples of face naming failure, from both brain-injured populations and from the realm of normal everyday experience, but in such instances, the unnamed face is invariably detected as familiar, and the precise identity of the possessor of the face is usually apprehended accurately. Consider, for example, the following common type of naming failure: you encounter someone whom you recently met, and cannot remember that person's name: you can remember when and where you met the person, who introduced you, and what the person does for a living—in short, you recognize the person normally. In prosopagnosia, the defect sets in at the level of recognition.

The recognition impairment in prosopagnosia occurs at the most subordinate level, i.e., at the level of specific identification of unique faces. Prosopagnosics are fully capable of recognizing faces as faces, i.e., performance is normal at more superordinate, nonspecific levels. Also, most prosopagnosics can recognize facial emotional expressions (e.g., happy, angry), and can make accurate determinations of gender and age based on face information (Humphreys et al. 1993, Tranel et al. 1988). These dissociations highlight several intriguing separations in the neural systems dedicated to processing different types of conceptual knowledge, such as knowledge about the meaning of stimuli, knowledge about emotion, and so on. In fact, these neural systems can be damaged in reverse fashion: for example, bilateral damage to the amygdala produces an impairment in recognizing facial emotional expressions, but spares the ability to recognize facial identity (Adolphs et al., 1995).

Although the problem with faces is usually the most striking, it turns out that the recognition defect in prosopagnosia is often not confined to faces. Careful assessment often reveals that the patient cannot recognize other visual entities at the normal level of specificity. The key determinants of whether other categories of stimuli are affected are (a) whether those stimuli are relatively numerous and visually similar, and (b) whether the demands of the situation call for specific identification. Whenever these conditions exist, prosopagnosics will tend to manifest deficits. For example, patients may not be able to identify a unique car, or a unique house, or a unique horse, even if they are able to recognize such entities at the basic object level, e.g., cars as cars, houses as houses, horses as horses. Similar to the problem with faces, they are unable to recognize the specific identity of a particular car, or house. These impairments underscore the notion that the core defect in prosopagnosia is the inability to disambiguate individual visual stimuli. In fact, cases have been reported in which the most troubling problem for the patient was in classes of visual stimuli other than human faces! For example, there was a farmer who lost his ability to recognize individual dairy (e.g., Holstein) cows, and a birdwatcher who became unable to tell apart various subtypes of birds (Assal et al. 1984, Bornstein et al. 1969).

Patients with face agnosia can usually recognize identity from movement. For example, upon seeing a distinctive gait of a familiar person, the patient can identify that person accurately, despite not knowing that person's face. This means not only that their perception of movement is intact, but also that they can evoke appropriate memories from the perception of unique patterns of movement. Conversely, patients with lesions in superior occipitoparietal regions (whose recognition of identity from form is normal, and hence do not have impaired face recognition) have defective motion perception and recognition. These findings underscore the separable functions of the 'dorsal' and 'ventral' visual systems, the dorsal one being specialized for spatial placement, movement, and other 'where' capacities; and the ventral one being specialized for form detection, shape recognition, and other 'what' capacities (Ungerleider and Mishkin 1982). In prosopagnosia, the dysfunction is in the 'what' system.

One of the most intriguing findings to emerge in this area of research is that despite an inability to recognize familiar faces consciously, prosopagnosic patients often have accurate nonconscious (or covert) discrimination of those faces. This phenomenon has been studied using a psychophysiological index (the skin conductance response [SCR]) to measure nonconscious discrimination (Tranel and Damasio 1985). SCRs were recorded while prosopagnosic patients

viewed a series of face stimuli. The stimulus sets included faces that were well known to the patients, mixed in random order with faces the patients had never seen before. While viewing the faces, the patients produced significantly larger SCRs to familiar faces, compared to unfamiliar ones. This occurred in several experiments, using different types of familiar faces: in one, the familiar faces were family members and friends, in another, the familiar faces were famous individuals (movie stars, politicians), and in yet another, the familiar faces were persons to whom the patients had had considerable exposure after the onset of their condition, but not before. In sum, the patients showed nonconscious discrimination of facial stimuli they could not otherwise recognize, and for which even a remote sense of familiarity was lacking. These findings suggest that some part of the physiological process of face recognition remains intact in the patients, although the results of this process are unavailable to consciousness. The fact that the patients were able to show this type of discrimination for faces to which they had been exposed only after the onset of their condition is particularly intriguing, as it suggests that the neural operations responsible for the formation and maintenance of new 'face records' can proceed independently from conscious influence.

3.2.2 Category-specific visual agnosia. Agnosia can develop for categories of stimuli other than faces, at levels above the subordinate, for example, at basic object level. For instance, patients may lose the ability to recognize animals or tools. This is generally referred to as visual object agnosia. The condition rarely affects all types of stimuli with equal magnitude (Farah and McClelland 1991, Forde and Humphreys 1999, Tranel et al. 1997, Warrington and Shallice 1984). In one common profile of visual object agnosia, there is a major defect in categories of living things, especially animals, with relative or even complete sparing of categories of artifactual entities (e.g., tools and utensils). Less commonly, the profile is reversed, in that the patient cannot recognize tools/utensils but performs normally for animals (Tranel et al. 1997, Warrington and McCarthy 1994). It has been shown that lesions in the right mesial occipital/ventral temporal region, and in the left mesial occipital region, are associated with defective recognition of animals; whereas lesions in the left occipital-temporal-parietal junction are associated with defective recognition of tools/utensils (Tranel et al. 1997).

4. Concluding Comment

Despite their relative rarity, agnosias have proved to be important 'experiments of nature,' and they have assisted with the investigation of the neural basis of human perception, learning, and memory. Careful study of agnosic patients over many decades, facilitated by the advent of modern neuroimaging techniques (computed tomography, magnetic resonance) and by the development of sophisticated experimental neuropsychological procedures, has yielded important new insights into the manner in which the human brain acquires, maintains, and uses various of knowledge.

See also: Amnesia; Face Recognition Models; Face Recognition: Psychological and Neural Aspects; Neural Representations of Objects; Object Recognition: Theories; Prosopagnosia

Bibliography

Adolphs R, Tranel D, Damasio A R 1998 The human amygdala in social judgment. *Nature* **393**: 470–4

Adolphs R, Tranel D, Damasio H, Damasio A R 1995 Fear and the human amygdala. *Journal of Neuroscience* **15**: 5879–91

Assal G, Favre C, Anderes J 1984 Nonrecognition of familiar animals by a farmer. Zooagnosia or prosopagnosia for animals. *Revue Neurologique* **140**: 580–4

Benton A 1990 Facial recognition. *Cortex* **26**: 491–9

Bornstein B, Sroka H, Munitz H 1969 Prosopagnosia with animal face agnosia. *Cortex* **5**: 164–9

Caramazza A, Shelton J R 1998 Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience* **10**: 1–34

Damasio A R, Damasio H, Van Hossen G W 1982 Prosopagnosia: Anatomic basis and behavioral mechanisms. *Neurology* **32**: 331–41

Damasio A R, Tranel D, Damasio H 1990 Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience* **13**: 89–109

Damasio H, Grabowski T J, Tranel D, Hichwa R D, Damasio A R 1996 A neural basis for lexical retrieval. *Nature* **380**: 499–505

Darwin C 1955 [1872] *The Expression of the Emotions in Man and Animals*. Philosophical Library, New York

De Renzi E 1997 Prosopagnosia. In: Feinberg T E, Farah M J (eds.) *Behavioral neurology and neuropsychology*. McGraw-Hill, New York, pp. 254–55

Ekman P 1973 *Darwin and Facial Expression: A Century of Research in Review*. Academic Press, New York

Farah M J 1990 *Visual agnosia*. The MIT Press, Cambridge, MA

Farah M J, McClelland J L 1991 A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology* **120**: 339–57

Forde E M E, Humphreys G W 1999 Category-specific recognition impairments: A review of important case studies and influential theories. *Aphasiology* **13**: 169–93

Fridlund A J 1994 *Human Facial Expression: An Evolutionary View*. Academic Press, New York

Gainotti G, Silveri M C, Daniele A, Giustolisi L 1995 Neuroanatomical correlates of category-specific semantic disorders: A critical survey. *Memory* **3**: 247–64

Humphreys G W, Donnelly N, Riddoch M J 1993 Expression is computed separately from facial identity, and it is computed separately for moving and static faces: Neuropsychological evidence. *Neuropsychologia* **31**: 173–81

Lissauer H 1890 Ein Fall von Seelenblindheit nebst einem Beitrage zur Theorie derselben. *Archiv für Psychiatrie und Nervenkrankherten* **21**: 222–70
Pulvermuller F 1999 Words in the brain's language. *Behavioral and Brain Sciences* **22**: 253–336
Tranel D, Damasio A R 1985 Knowledge without awareness: An autonomic index of facial recognition by prosopagnosics. *Science* **228**: 1453–4
Tranel D, Damasio A R 1996 The agnosias and apraxias. In: Bradley W G, Daroff R B, Fenichel G M, Marsden C D (eds.) *Neurology in Clinical Practice*, 2nd edn. Butterworth, Stoneham, MA, pp. 119–29
Tranel D, Damasio A R, Damasio H 1988 Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology* **38**: 690–6
Tranel D, Damasio H, Damasio A R 1997 A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia* **35**: 1319–27
Ungerleider L G, Mishkin M 1982 Two cortical visual systems. In: Ingle D J, Goodale M A, Mansfield R J W (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA, pp. 549–86
Young A W, Bruce V 1991 Perceptual categories and the computation of 'grandmother.' *European Journal of Cognitive Psychology* **3**: 5–49
Warrington E K, McCarthy R A 1994 Multiple meaning systems in the brain: A case for visual semantics. *Neuropsychologia* **32**: 1465–73
Warrington E K, Shallice T 1984 Category specific semantic impairments. *Brain* **107**: 829–53

D. Tranel and A. R. Damasio

Agonistic Behavior

1. Overview

Aggression and violence are serious social problems, as illustrated by acts ranging from school violence to wars. From an evolutionary viewpoint, on the other hand, aggression is often described as adaptive. From a humanitarian point of view it is difficult to imagine war among humans as being adaptive. The challenge to science is to resolve these contrasting views of aggression. Although research on aggression has been extensive, it has not led to significant progress in understanding and preventing aggressive acts. It was this lack of progress which led to the introduction of the concept of agonistic behavior in the mid-twentieth century. The definition of agonistic behavior was more inclusive of behaviors often not included under the umbrella of aggression. This provided a broader context for understanding aggression in relation to other behaviors.

The purpose of this article is to review the current status of aggression research as it relates to agonistic behaviors. The focus will be primarily on classifying and predicting human aggression. Lower animal research will be reviewed briefly in cases where the results add to the understanding of human aggression. (The term 'agonistic' has been used more frequently in research with lower animals than in human research.)

1.1 Definitions and Measurements

Although there are no universally accepted definitions of human aggression, it has generally been defined as behavior which results in physical or psychological harm to another person and/or in the destruction of property. It usually includes overt physical acts (e.g., fighting or breaking objects) or verbal abuse. Lower animals also engage in overt physical fighting. The counterpart of verbal abuse among lower animals is 'aggressive displays' in which animals vocalize and/or assume threatening postures (Kalin 1999). There are data suggesting that among lower animals size is often related to achieving dominance, and lower animals will often make themselves look larger when threatened; for example, fish will make themselves appear larger by extending their fins (Clemente and Lindsley 1967).

Agonistic behavior was defined as adaptive acts which arise out of conflicts between two members of the same species (Scott 1966, 1973). As noted, agonistic behaviors were more inclusive and provided a broader context within which to classify the more traditional concepts of aggression. In addition to overt aggressive acts or threats, agonistic behaviors included passive acts of submission, flight, and playful behaviors which involve physical contact. For example, human participation in sports or playful jostling would not generally be included as a form of aggression but would be included under the agonistic umbrella.

Since the introduction of the term 'agonistic,' the differences between agonistic and aggressive behaviors have blurred and the two labels are often used interchangeably in the literature. Its introduction did not result in more productive leads for understanding or preventing human aggression. Among humans it appears that developing techniques for killing have outstripped our knowledge of how to prevent killing. The substitution of a new term for aggression has not changed this trend. The major challenge in aggression research is to develop a model which can serve to synthesize data across a wide range of scientific disciplines (Barratt et al. 1997).

Techniques range from qualitative observations of behavior in naturalistic settings to more quantitative measures of aggressive behaviors in laboratory settings. Discipline-specific language has often produced confusion when comparing the results from cross-disciplinary research. Thus, as noted the major challenge to science is to view aggression from a more neutral context: a discipline-neutral model.

The focus here will be on classifying and measuring

both aggression and risk factors for aggression under four headings: (a) behavior, (b) biology, (c) cognitive or mental processes, and (d) environment or the setting in which psychosocial development takes place and aggression is expressed. No attempt will be made here to organize these four classes of descriptors and measurements into a model, but it should be noted that attempts to do so have been documented in the literature.

2. *Classifying and Measuring Human Aggression*

Aggression is behavior. Therefore, what is to be predicted in human aggression research are aggressive acts. These acts become the criterion measures for which risk factors or predictor measures are sought. One of the more difficult tasks in aggression research is defining these acts so they can be measured and related quantitatively to potential predictors. Unless the acts are quantitatively measured, the efficacy of various interventions for controlling aggressive acts cannot be reliably determined.

The properties of human aggressive acts which can be quantified are: (a) frequency with which the acts occur; (b) intensity of the act or degree of physical or psychological harm inflicted; (c) the target of the act; (d) the stimuli within the environmental setting which trigger the act; (e) the expressive form of the act (e.g., overt physical acts vs. verbal assaults); (f) the type of act in terms of intent. These properties of aggressive acts are often used singly or in combination as outcome or criterion measures of aggression.

There are three types of aggressive acts related to intent: (a) impulsive or reactive aggression or acting without thinking; (b) premeditated, planned or proactive aggression; (c) medically related aggression or aggressive acts which are committed secondary to a medical disorder, such as a closed head injury or psychiatric disorder. Classifying aggressive acts based on intent or effect is important because different interventions are effective with each type. If aggressive acts are a sign or symptom of a medical disorder, controlling the disorder should result in control of the aggression.

Impulsive aggression has been shown to be related in part to low levels of a neurotransmitter, serotonin, which helps selected neurons in the brain communicate with one another. Giving a medication which increases levels of serotonin has been shown to control impulsive aggression. Selected medications used to control seizures (anticonvulsants) have also been shown to control impulsive aggression.

In contrast, premeditated aggression cannot be controlled by medication but instead responds to cognitive/behavioral therapy which is based on social learning theory. This makes sense because premeditated or proactive aggression is learned in social situations. Premeditated human aggression is often compared with subhuman aggression which is related to protecting a territory for either food or reproductive purposes. These behaviors have in part a genetic basis which generally is learned in a social context.

2.1 *Human Agonistic Behaviors*

Not all agonistic behaviors among humans relate to social or clinical problems. For example, human sports activities are competitive and often result in physical harm to participants. Yet these events are condoned by society. The social value of these events is often explained in terms of the evolution of agonistic behaviors among lower animals that have become part of human biological drives. It is generally agreed that most common agonistic behaviors among lower animals relate to achieving dominance, which in turn is related to protecting a territory for purposes of food or reproduction as described above. Lower animals also engage in 'play-like' behaviors to learn to express and experience dominance in a tolerant environment. These behaviors are apparently not intended to do harm. If one observes a litter of pups as they mature, this type of 'play' behavior is obvious. At the human level, play and sports provide not only adaptive and socially acceptable outlets for aggressive impulses, but also an opportunity for non-participants to identify with a 'group,' hopefully as a 'winner.' This provides a sense of belonging.

2.2 *Techniques for Measuring Human Aggression*

As noted, one of the more difficult tasks in aggression research is quantifying the aggressive acts, especially at the human level. Opportunities to observe human aggression directly in natural settings are not common and are restricted primarily to institutions such as prisons or schools. The most common ways of measuring human aggressive acts are by structured interviews or self-report measures of aggressive acts. As emphasized earlier, aggression is behavior and should not be confused with anger or hostility, which are often precursors of aggressive acts. Self-report measures of aggression can be reliable in some instances but subjects may confuse their feelings of anger and hostility with aggression. Thus, in human-level research, reporters (e.g., spouse) who can observe an individual's behavior are also often used to document the aggressive acts of subjects. In hospital settings where aggressive patients are housed, rating scales of aggressive acts have been developed for use in quantifying patients' aggressive acts on the wards.

3. *Risk Factors for Human Aggression*

Risk factors or predictor measures of human aggression will be discussed briefly under the four headings listed in Sect. 1.1 above. Examples will be

presented in each category since lack of space precludes an indepth discussion.

3.1 Biological Predictors of Aggression

3.1.1 Neurotransmitters and hormones. The biological processes of the brain are controlled and maintained in large part by biochemicals called neurotransmitters and/or hormones. One of the most commonly quoted findings in psychopharmacology is that the serotonergic system of the brain is related to impulsive aggression, as noted above. Low levels of the neurotransmitter serotonin have been shown to be related in both lower animal and human studies to impulsive aggression, but not to other forms of aggression. Serotonin is involved primarily with brain systems which regulate behavioral inhibition (Ferris and Devil 1994). Other neurotransmitters (e.g., norepinephrine) have been shown to relate to creating the drive or impulse to be aggressive. As with most scientific findings, the results often become less clear as research progresses and it has been suggested that serotonin is not an exclusive or possibly even the best neurochemical marker for impulsive aggression. It is probable that in the long run a profile of neurochemical markers will be related to impulsive aggression rather than one or two neurotransmitters.

Hormones have also been related to aggression. For example, testosterone levels among males have been shown to be related to aggressive behaviors (Archer 1991).

3.1.2 Genetics. Although there is evidence of heritable aggressive behaviors in lower animals, especially mice and rats, there is no creditable evidence at this time for a genetic predisposition for aggression among humans. This is especially true for molecular genetic markers. There has been suggestive evidence in behavioral genetic studies for the inheritance of aggression, but these findings have been difficult to replicate.

3.1.3 Neuroanatomy. A number of brain areas have been related to aggression in lower animals but the relevance of these findings for understanding human aggression is limited because of differences in brain function and structure. One of the main problems in relating brain structures to aggression among humans is the hierarchical nature of the brain's structure, involving neurons which carry information across different parts of the brain. Implying that one area of the brain is responsible for aggressive acts ignores the interdependence of brain structures. Even parts of the same brain nucleus (e.g., the amygdala) can affect aggression differently because of their relationship with different brain systems. Neuroanatomical explanations of human aggression are limited but imaging techniques (e.g., PET scans) offer promise for the future.

3.2 Cognitive Precursors of Aggression

Research has shown that verbal skills including reading are related to impulsive aggression. It has been proposed that the reason for this relationship is that humans often covertly verbalize control of their behaviors. Among persons with verbal skill deficits this control would be diminished, hence they would more likely be aggressive if an impulse to aggress was present.

Another important cognitive process relates to conscious *feelings* of anger and hostility which are precursors of aggression. Measures of these two traits are often mistakenly used as measures of aggression. These traits are best classified as biological states which can be verbalized and cognitively experienced. One 'feels angry' but one acts aggressively.

3.3 Environmental Precursors of Aggression

It has been demonstrated among lower animals that different rearing environments can lead to changes in biological functions which are purportedly related to aggression (Kramer and Clarke 1996). For example, not having a mother in a rearing environment at critical developmental periods can lead to decreased levels of serotonin, which as noted above has been suggested as a major biological precursor of impulsive aggression. Among humans aggression is often related to living conditions (Wilson 1975). For example, persons in lower socioeconomic neighborhoods are more likely to be involved in fights than persons in higher socioeconomic neighborhoods. Again, these are complex interactions and caution is warranted in generalizing the results as 'causes' of aggression.

3.4 Behavioral Precursors and Laboratory Models of Aggression

As is generally true for most behaviors, one of the best predictors of aggression is a past history of aggressive acts. This is true for both impulsive and premeditated aggression.

Another way of studying human aggressive behavior is to generate it in laboratory situations. An example is a computer-simulated betting procedure. Individuals sit in front of a TV screen and attempt to accumulate money by pressing a button under different conditions. They think that they are competing with someone in another room for the money but they are not. Persons with tendencies toward impulsive aggression will display aggression in this well-controlled laboratory setting. This procedure can be used

to test the efficacy of 'anti-aggression' medications or for studying the effects of alcohol and other drugs on aggressive behavior.

4. Postscript

This article has focused primarily on one example of agonistic behaviors, namely aggression. The need for quantitative measures to study aggression was emphasized, as well as the problems related to predicting aggressive behaviors. It is important to realize that there are different types of aggression with different sets of precursors or risk factors for each. The greatest hindrance to advancing aggression research at this time is the lack of a discipline-neutral model which can be used to synthesize discipline-specific data in the search for precursors of aggression.

See also: Aggression in Adulthood, Psychology of; Behavior Therapy: Psychiatric Aspects; Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of; Neurotransmitters; Sex Hormones and their Brain Receptors

Bibliography

Archer J 1991 The influence of testosterone on human aggression. *British Journal of Psychology* **82**: 1–28

Barratt E S, Sanford M S, Kent T A, Felthous A 1997 Neuropsychological and cognitive psychophysiological substrates of impulsive aggression. *Biological Psychiatry* **41**: 1045–61

Clemente C D, Lindsley D B 1967 *Aggression and Defense: Neural Mechanisms and Social Patterns*. University of California Press, Los Angeles

Ferris C F, Devil Y 1994 Vasopressin and serotonin interactions in the control of agonistic behavior. *Psychoneuroendocrinology* **19**: 593–601

Kalin N H 1999 Primate models to understand human aggression. *Journal of Clinical Psychiatry* **60;** suppl. 15: 29–32

Kramer G W, Clarke A S 1996 Social attachment, brain function, and aggression. *Annals of the New York Academy of Science* **794**: 121–35

Scott J P 1966 Agonistic behavior of mice and rats: A review. *American Zoologist* **6**: 683–701

Scott J P 1973 Hostility and Aggression. In: Wolman B B (ed.) *Handbook of General Psychology*. Prentice-Hall, Englewood Cliffs, NJ, pp. 707–19

Wilson E O 1975 Aggression. In: *Sociobiology*. Belknap Press of Harvard University Press, Cambridge, MA, Chap. 11, pp. 242–55

E. S. Barratt

Agricultural Change Theory

Agricultural change refers not just to the difference between the first plantings 10,000 years ago and today's computerized, industrialized, genetically engineered production systems; agricultural change occurs on a daily basis, as farmers in every country of the world make decisions about what, where, and how to cultivate. The importance of the topic goes well beyond how much food is produced, how much money is made, and how the environment is affected: agriculture is intimately linked to many institutions in every society, and to population. This article examines the most influential theories of agricultural change in general, with particular emphasis on the role of population growth.

1. Overview

Scholarship on agricultural change has been anchored by two small books with enormous impacts, both focused on the relationship between farming and population. In 1798, British clergyman Thomas Malthus argued for an intrinsic imbalance between rates of population increase and food production, concluding that it was the fate of human numbers to be checked by 'misery and vice'—generally in the form of starvation and war. Although intended mainly as an essay on poverty, population, and Enlightenment doctrines, *An Essay on the Principle of Population* (Malthus 1798) infused popular and scientific thought with a particular model of agricultural change, in which a generally inelastic agricultural sector characteristically operated at the highest level allowed by available technology.

In 1965, Danish agricultural economist Ester Boserup claimed to upend this model of agriculture by arguing that, particularly in 'primitive' agricultural systems, farmers tended to produce well below the maximum because this allowed greater efficiency (output:input ratio). She maintained that production was intensified and additional technology adopted mainly when forced by population.

Each model is quite simple—dangerously oversimplified, many would now argue—but they provide invaluable starting points from which to address the complexities of agricultural change.

2. Malthus

Malthus's famous maxim from *Population* was that 'the power of population is indefinitely greater than the power in the earth to produce subsistence for man…Population, when unchecked, increases in a geometrical ratio. Subsistence increases only in an arithmetical ratio.' Subsequent empirical research has made this position appear dubious. He used sketchy accounts of population booms in New World colonies to show that unchecked populations double every 25 years, but such growth rates have been shown to be highly exceptional. His view of agricultural production

as relatively inelastic, with output increasable chiefly by bringing more land into tillage, has also fared poorly in subsequent comparative agricultural research.

Equally problematic has been the correlation of Malthus's 'positive checks' of starvation and warfare with populations outpacing their food supply. As Sen (1981) shows, famines result from political failures more than from inability of agriculture to keep up with population. For instance, history's greatest famine, which claimed 30–70 million Chinese peasants during Mao's Great Leap Forward in 1958–60 (Ashton et al. 1984, Becker 1996), was no Malthusian disaster, although in 1798 Malthus had opined that Chinese numbers 'must be repressed by occasional famines.' In fact, population had grown substantially since then, and has grown more since recovering from the Great Leap Forward; Chinese peasants have shown a historic capability of feeding themselves at such densities, principally through the ingenuity of highly intensive wet rice cultivation (Bray 1986). (The 1958–60 famine resulted from policies that disrupted locally-developed intensive practices as well the social institutions needed to sustain those practices (Becker 1996, Netting 1993)).

The Malthusian perspective nevertheless has proved remarkably durable in its effects on common perceptions and theories of agricultural change. Its survival is probably less related to empirical analysis than to the ways theories of agricultural change affect, and are affected by, their political context. For instance, Malthus wrote during the early stages of the Industrial Revolution in England, a time marked by a rapidly growing urban underclass and debates about obligation to feed them. Subsuming food shortages under inexorable laws of population and agricultural change was obviously appealing to prosperous segments of society, and Malthus was rewarded with a chair in political economy at the University of Haileybury. When the Irish Potato Famine hit in the late 1840s, it was widely interpreteted as a Malthusian disaster, despite Ireland's relatively low population density and the fact that food exports continued (in fact, increased) throughout the crisis (Ross 1998). The British director of relief efforts, a former student of Malthus at Haileybury, characterized the famine as 'a direct stroke of an all-wise and all-merciful Providence' (Ross 1998, p. 46).

Most recently, the perpetuation of the Malthusian perspective on agricultural change can be seen in debates on the merits of genetically modified (GM) crops. Parties in government, industry, and biological science with vested interests in GM products routinely cite famine and malnourishment in developing countries as a justification for the technology. The notion of an inelastic agriculture incapable of feeding the populace is entrenched enough that few question this claim, despite lack of evidence pointing to inadequacy of current crop plants or even the likelihood of GM plants offering higher levels of production.

3. *Boserup*

Boserup's *The Conditions of Agricultural Growth* (1965) brought an important new perspective on agricultural change. Since Malthus's time, there had been much comparative agricultural research, especially on peasant (i.e., not entirely market-oriented) systems, which Boserup used in developing a 'dynamic analysis embracing all types of primitive agriculture' (1965, p. 13). Rather than technological change determining population (via food supply), in this model population determined technological change (via the optimization of energetics). This countered Malthus's assumption that agricultural systems tended to produce at the maximal level allowed by available technology. Instead, land was shown often to be used intermittently, with heavy reliance on fire to clear fields and fallowing to restore fertility in the widespread practice of 'slash and burn' farming (Boserup 1965, p. 12). Therefore, comparisons of agricultural productivity had to be in terms of output per unit of land per unit of time—what some call 'production concentration.'

Boserup held that extensive agriculture with low overall production concentration is commonly practiced when rural population density is low enough to allow it, because it tends to be favorable in total workload and efficiency (output:input). Rising population density requires production concentration to rise and fallow times to shorten. Contending with less fertile plots, covered with grass or bushes rather than forest, mandates expanded efforts at fertilizing, field preparation, weed control, and irrigation. These changes often induce agricultural innovation but increase marginal labor cost to the farmer as well: the higher the rural population density, the more hours the farmer must work for the same amount of produce. In other words: as the benefits of fire and fallowing are sacrificed, workloads tend to rise while efficiency drops. It is because of this decreased labor efficiency that farmers rarely intensify agriculture without strong inducements, the most common inducement being population growth. Changing agricultural methods to raise production concentration at the cost of more work at lower efficiency is what Boserup describes as agricultural intensification (Fig. 1).

The model of peasant agriculture being driven by optimization of energetics, with population serving as the prime engine of change, brought a sea change in agricultural change theory. Boserup's name has become synonymous with this perspective, and indeed it was in *The Conditions of Agricultural Growth* that it was crystallized, but others have contributed significantly to this perspective. Most notable was the Russian economist Chayanov (1925), who analyzed peasant farming in terms of energy optimization, with change driven mainly by the demographic makeup of households.

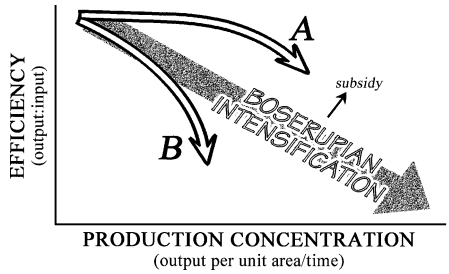

Figure 1
Schematic view of relationships between production concentration and efficiency (output:input) of agricultural methods

Agricultural change theory has now been carried far beyond the simple outlines presented in 1965. Boserup initially stressed that intensification's costs came in the field as fallows were shortened, but she (1981, p. 5) and others have also identified other modes of intensification. Capital-based intensification is characteristic of industrialized societies. The amount of human labor required to produce food generally decreases, whereas the total direct and indirect energy costs can climb to exceedingly high levels. In infrastructure-based intensification, the landscape is rebuilt to enhance, or remove constraints on, production. Land improvements used well beyond the present cropping cycle—such as terraces, ridged fields, dikes, and irrigation ditches—are termed 'landesque capital' (Blaikie and Brookfield 1987). Since landesque capital depends on long-term control (although not necessarily formal ownership and alienability), Boserup posited a general association between intensification and private land tenure, which has been supported in subsequent research (Netting 1993).

At a very general level, the Boserup model of agricultural change has been found to fit fairly well: farmers with abundant land do tend to rely heavily on methods that are land-expensive and labor-cheap; farmers under more crowded conditions do tend to adopt labor-expensive (or capital-expensive) methods; and the decline in marginal utility on inputs does offer a causal mechanism for the change. The model has an impressive record of empirical support from both cross-cultural and longitudinal studies, and it has been indispensable in explaining cross-cultural agricultural variability (Netting 1993, Turner et al. 1977, Turner et al. 1993, Wiggins 1995).

4. *Post-Boserup Research*

The Boserup model has been widely influential, but its broad-brush success comes at the cost of neglecting many important aspects of agricultural change, and researchers from various fields have fault it. Major factors shaping agricultural change beyond Boserup's simple model may be grouped into the categories of ecological, social, and political-economic.

4.1 *Ecological Variation*

Boserup depicts intensification as a universal process cross-cutting environment, but her model relies heavily on agroecological features of fire and fallow that are hardly universal. Thresholds of intensification vary with local environment (Brookfield 1972, p. 44), and the relationship between production concentration and efficiency may be quite variable among environments (Turner et al. 1977, Turner and Brush 1987).

Figure 1 schematically depicts different concentration/efficiency trajectories. The large arrow represents the global pattern emerging from the many cases where productive concentration can be raised, but only at the expense of lowered efficiency. This is the broad pattern confirmed by the empirical studies cited above: Boserupian intensification, defined as the process of raising production concentration by accepting higher labor demands and lower efficiency. In general, this trajectory fits when the labor costs of intensification are both necessary and sufficient to raise production concentration: necessary in that higher production requires proportionately more work, and sufficient in that the proportionate increase in work succeeds in raising output. Where lowered efficiency is not necessary for higher production concentration, the slope would be flatter, as indicated by non-Boserupian trajectory A. The other non-Boserupian pattern occurs where productive concentration cannot be raised, or where the cost of raising it is intolerable: trajectory B. Such a trajectory requires nonagricultural responses to rising population pressure (Stone and Downum 1999).

Although the issue is by no means settled, paddy rice production appears to exemplify trajectory A in many cases. Although it requires high labor inputs (e.g., Clark and Haswell 1967), the pattern of declining yields may be overridden by the distinctive ecology of the paddy in which fertility tends to increase rather than decrease (Bray 1986). Trajectory B is exemplified by arid areas where increasing inputs into reduced land areas cannot overcome the moisture limitations on crops, and would only serve to increase risk (Stone and Downum 1999).

4.2 *Social Factors*

Social context affects both the demands for agricultural products and the relative efficiency of different production methods. Food requirements may be affected not only by calorific needs but by what

Brookfield (1972, p. 38) calls social production, meaning 'goods produced for the use of others in prestation, ceremony and ritual, and hence having a primarily social purpose.' Among New Guinea groups, Brookfield observed production levels that were 'wildly uneconomic' in terms of energetics, but which earned a very real social dividend.

But agriculture is not only practiced partly for social ends; it is practiced by social means, which can have marked effects on how agricultural methods respond to changes in population. Nonindustrialized agriculture is run largely through social institutions for mobilizing resources. Therefore, efficiency of production strategies can vary culturally, and even a purely 'calorific' analysis must consider social institutions that affect costs and benefits.

A comparison of Kofyar and Tiv farmers in central Nigeria provides an example. Expanding out of a crowded homeland on the Jos Plateau, Kofyar farmers began to colonize a frontier near Assaikio in the 1950s. By the early 1960s, there were Kofyar living in frontier communities with population densities below 10/km^2 and agriculture was mostly extensive. By the mid-1980s population density had risen to 100/km^2 and there had been considerable agricultural intensification, with a mean yearly labor input of over 1,500 hours per person (Stone 1996). Intensification was aided by the social institutions that facilitated intensive farming in the homeland, including social mechanisms for mobilizing labor with beer, food, cash, specific reciprocity, or generalized reciprocity. The Kofyar found the main alternative to intensification—migration—expensive and risky, and tended to avoid it.

Nearby were Tiv farmers whose agricultural trajectory followed a different course. Tiv began migrating northward in the 1930s from a homeland known for settlement mobility (Bohannan 1954), and settlement was also highly mobile in the Assaikio area. Their population densities grew more slowly than the Kofyar's, and they showed a clear aversion to the intensification of agriculture. Where the Kofyar had relied on pre-existing institutions for mobilizing labor to facilitate intensification, the Tiv relied on a set of interlocking institutions to facilitate movement (Stone 1997). As long as they could maintain a relatively low population density, they could keep in place an agricultural regime that was extensive enough to allow substantial amount of free time. Much of this time went towards travel and development of social networks that lowered the costs and risks of moving.

4.3 The Role of Political Economy

Agricultural change is shaped by external economic systems, and most farmers have to contend with economic factors that affect the cost of inputs and value of output beyond local energetics. Market incentives can induce farmers to intensify in the absence of land shortage (e.g., Turner and Brush 1987). Eder's (1991, p. 246) observation that farmers 'make their production decisions in terms of pesos per hour, not kilograms per hour' is apt, although it is not so much a cash/energy dichotomy but a gradient. Few small farmers today grow crops exclusively for subsistence or sale; most do both, and they often favor crops that can be used for food or sale. Market involvement does not totally negate the Boserup model (Netting 1993), but it clearly introduces variables that can override effects of local population and energetics.

But of the factors neglected by the Boserup model, the most critical to many contemporary scholars is the variation in farmers' ability to intensify agriculture as they may wish (e.g., Bray 1986, p. 30). As Blaikie and Brookfield (1987, p. 30) put it, the Boserup model 'may be likened to a toothpaste tube—population growth applies pressure on the tube, and somehow, in an undefined way, squeezes out agricultural innovation at the other end.' Even within a single set of ecological, technological, and demographic conditions, population pressure may prompt very different patterns of agricultural change because of differences in farmers' ability to invest, withstand risk, and attract subsidy. While population pressure may stimulate technological change and creation of landesque capital, 'what appears at the other end of the tube is often not innovation but degradation' (Blaikie and Brookfield 1987, p. 30). For instance, Durham's analysis of environmental destruction in Latin America compares two separate feedback loops, both of which include population increase (Durham 1995, pp. 252–4).

The 'capital accumulation' loop leads to intensified commercial production and land concentration, while the 'impoverishment' loop leads to deforestation and ultimately reduced production; the loops feed each other.

The Boserup model is resolutely local in outlook: the cost and benefit of an agricultural operation such as plowing or tree felling is reckoned on the basis of effort required and crops produced. This holds constant the effects of external subsidy that is often available. Farmers may well achieve a higher marginal return on efforts to attract subsidy (e.g., fertilizer from a government program, irrigation ditches constructed by an NGO, or new seed stocks from a development project) than on plowing or tree felling. From the farmer's perspective, this allows new possibilities of raised production, as represented by the small arrow in Fig. 1. There may have been no absolute improvement in efficiency at all, merely a shifting of some costs to the outside by capturing subsidy. The ability to attract such subsidy is politically mediated, and it often varies sharply among segments of a farming population.

See also: Agricultural Sciences and Technology; Agriculture, Economics of; Farmland Preservation;

Internal Migration (Rural—Urban): Industrialized Countries; Population and Technological Change in Agriculture; Population Cycles and Demographic Behavior; Population Ecology; Population Pressure, Resources, and the Environment: Industrialized World; Rural Geography

Bibliography

Ashton B, Hill K, Piazza A, Zeitz R 1984 Famine in China, 1958–1961. *Population and Development Review* **10**: 613–45

Becker J 1996 *Hungry Ghosts: Mao's Secret Famine*. Free Press, New York

Blaikie P, Brookfield H 1987 *Land Degradation and Society*. Methuen, London

Bohannan P 1954 The migration and expansion of the Tiv. *Africa* **24**: 2–16

Boserup E 1965 *The Conditions of Agricultural Growth: The Economics of Agrarian Change under Population Pressure*. Aldine, New York

Boserup E 1981 *Population and Technological Change: A Study of Long Term Trends*. University of Chicago Press, Chicago

Bray F 1986 *The Rice Economies: Technology and Development in Asian Societies*. Blackwell, New York

Brookfield H 1972 Intensification and disintensification in Pacific agriculture. *Pacific Viewpoint* **13**: 30–41

Chayanov A 1925/1986 Peasant farm organization. In: Thorner D, Kerblay B, Smith R (eds.) *The Theory of Peasant Economy*. R D Irwin, Homewood, IL

Clark C, Haswell M 1967 *The Economics of Subsistence Agriculture*. Macmillan, London

Durham W H 1995 Political ecology and environmental destruction in Latin America. In: Painter M, Durham W H (eds.) *The Social Causes of Environmental Destruction in Latin America*. University of Michigan Press, Ann Arbor, MI

Eder J F 1991 Agricultural intensification and labor productivity in a Philippine vegetable garden community: A longitudinal study. *Human Organization* **50**: 245–55

Malthus T 1798 *An Essay on the Principle of Population*. J. Johnson, London

Netting R Mac 1993 *Smallholders, Householders: Farm Families and the Ecology of intensive, sustainable agriculture*. Stanford University Press, Stanford, CA

Ross E 1998 *The Malthus Factor: Population, Poverty, and Politics in Capitalist Development*. Zed, London

Sen A 1981 *Poverty and Famines: An Essay on Entitlement and Deprivation*. Clarendon Press, Oxford, UK

Stone G D 1996 *Settlement Ecology: The Social and Spatial Organization of Kofyar Agriculture*. University of Arizona Press, Tucson, AZ

Stone G D 1997 Predatory sedentism: Intimidation and intensification in the Nigerian savanna. *Human Ecology* **25**: 223–42

Stone G D, Downum C E 1999 Non-Boserupian ecology and agricultural risk: Ethnic politics and land control in the arid Southwest. *American Anthropologist* **101**: 113–28

Stone G D, Netting R M, Stone M P 1990 Seasonality, labor scheduling and agricultural intensification in the Nigerian savanna. *American Anthropologist* **92**: 7–23

Turner B L II, Brush S B (eds.) 1987 *Comparative Farming Systems*. Guilford, New York

Turner B L II, Hanham R, Portararo A 1977 Population pressure and agricultural intensity. *Annals of the Association of American Geographers* **67**: 384–96

Turner B L, Hyden G, Kates R (eds.) 1993 *Population Growth and Agricultural Change in Africa*. University of Florida Press, Gainesville, FL

Wiggins S 1995 Change in African farming systems between the mid-1970s and the mid-1980s. *Journal of International Development* **7**: 807–48

G. D. Stone

Agricultural Sciences and Technology

1. Introduction

The agricultural sciences and technology are usually seen to encompass the plant, animal and food sciences, soil science, agricultural engineering and entomology. In addition, in many research institutions related fields such as agricultural economics, rural sociology, human nutrition, forestry, fisheries, and home economics are included as well. The agricultural sciences have been studied by historians, economists, sociologists, and philosophers.

Most of the early work in Science and Technology Studies focused on physics, said to be the model for the sciences. Unlike the agricultural sciences, theoretical physics appeared disconnected from any clear social or economic interests. Indeed, one early study of the agricultural sciences described them as deviant in that they did not follow the norms found in physics (Storer 1980).

Prior to the 1970s studies of the agricultural sciences tended to be apologetic and uncritical. Then, critical historical, economic, sociological, and philosophical studies of the agricultural sciences began to emerge. These studies built on earlier work that was not within the purview of what is usually called STS. Moreover, despite attempts to incorporate perspectives from this field, it would be an exaggeration to say that studies of the agricultural sciences form an integrated body of knowledge. Indeed, fragmentation has been and remains the rule with respect to theoretical frameworks, research questions, and methods employed.

2. History

Recent historical studies have challenged the hagiographical approach of official histories, demonstrating how the organizational structure of agricultural science encouraged particular research strategies and products. Of particular import were studies of the role of the state-sponsored botanical and zoological gardens, and later the agricultural experiment stations, in the colonial project. In particular, historians began

to document the close relations between the rise of economic botany as a discipline and the creation of botanical gardens in the various European colonies in the seventeenth century (Brockway 1979, Drayton 2000). Such gardens served simultaneously to further the classification of botanical species and the colonial project by identifying plants of economic value that might serve to valorize the new colonies. Coffee, tea, cocoa, rubber, sugar, and other crops were developed as plantation crops and soon thrived in regions far from their locations of origin. In so doing, they provided revenue for colonial governments and profits for the emerging large colonial trading companies.

In the late nineteenth century, botanical gardens were superseded by agricultural experiment stations in most industrialized nations and their colonies. Until World War II, the agricultural experiment stations were *the* model and often the sole recipient of government support for non-military scientific research. The experiment stations focused on increasing yields of food crops in Europe and North America, thereby keeping industrial wages down through cheap food and avoiding feared Malthusian calamities. At the same time the experiment stations in the colonies focused their efforts on increasing yields of exports so as to provide a steady supply of raw materials to European industries. For example, the Gezira scheme in the Sudan combined science, commerce and irrigation so as to provide long staple cotton for the Lancashire mills (Barnett 1977).

In contrast to most plant and animal research, mechanical (e.g., farm equipment) and chemical technologies (e.g., fertilizers, pesticides) in agriculture were developed by private companies. Over the twentieth century, the agricultural sciences and technologies played an important role in increasing agricultural productivity per hectare and per hour of labor. However, some argue that this was only accomplished by displacing vast agrarian populations and increasing environmental degradation.

3. *Economics*

While economic studies as early as the 1930s celebrated the products of agricultural research and emphasized 'adjustment' to the new technologies by farmers, the newer literature has focused primarily on the social rates of return to agricultural research. Many such studies have been used to lobby for additional public support for research. Critics of this approach have argued that only the benefits are estimated while many of the costs are excluded as they are unmeasurable (e.g., acute pesticide poisoning, environmental damage, and research programs yielding few or no results) (Fuglie et al. 1996).

Of particular importance has been the development of the theory of 'induced innovation' as an explanation for the directions taken in agricultural research (Hayami and Ruttan 1985). Proponents of this framework argue that innovations are induced by the relative scarcity of land, labor, and capital. Thus, in Japan, where land is scarce, research has focused on increasing yields per unit of land. In contrast, in the US, where labor is scarce, research has focused on increasing yields per unit of labor. The theory further asserts that agricultural research is responsive to demands of farmers as voiced in the political sphere, since much agricultural research is publicly funded. However, critics have argued that this is only likely to be true in democratic regimes (Burmeister 1988).

Others have examined the allocation of research support across commodities, directing particular concern to what has come to be known as the problem of spillover. Since much public agricultural research has focused on the creation of products and practices that are not protectable by patents or copyrights, research completed in one nation or region can often be used with only minor adaptation in other locales. Economists have concluded that developing nations should not engage in research on commodities with high spillover, such as wheat; instead should rely on other nations and international agricultural research centers for such materials. They argue that such nations would be served better by investing in research on commodities not grown elsewhere. Others have argued successfully for the formation of international networks for research on particular commodities so as to spread the costs of research over several nations with similar agroecological conditions (Plucknett et al. 1990). Such networks have been used effectively to exchange information and materials as well as to foster collaboration. However, with the rise of stronger intellectual property rights in agriculture over the last several decades (including plant variety protection as well as utility patents for life forms), there is some evidence that spillovers may be on the decline.

Another area of interest to economists has been the division of labor between public and private financing of agricultural research. While some economists have argued that biological (as contrasted with chemical and mechanical) agricultural research is by definition a public good, others argue that stronger intellectual property rights make it possible for the private sector to shoulder most of the burden for such research. They have attempted to build the case that stronger intellectual property rights create incentives for private firms to invest in biological research (e.g., plant and animal biotechnologies, seed production), leaving only research in the social sciences and natural resource management to the public sector. Indeed, some nations (e.g., the UK) have privatized part or all of their agricultural research with varying degrees of success.

Finally, related to the division of labor is the issue of alternative public funding mechanisms. Traditionally, agricultural research has been funded institutionally,

based on annual lump-sum appropriations. However, there has been a shift toward more project-based competitive grant programs in which scientists compete to receive grants. Proponents of competitive grants argue that this approach ensures that the best research is conducted by the most competent scientists. In contrast, proponents of institutional funding argue that agriculture is fundamentally place based, necessitating that investigations be distributed across differing ecological zones. They also note that competitive grants tend to be funded over just a few years while much agricultural research requires a decade to complete.

4. *Sociology*

In sociology, adoption-diffusion theory was the dominant approach through the 1960s (Rogers 1995). Diffusion theorists accepted the products of agricultural research as wholly desirable. Hence, their work focused almost exclusively on the fate of innovations designed for farm use. They employed a communications model adopted from engineering in which messages were seen to be transmitted from sender to receiver, later adding the engineering term 'feedback' to describe receivers' responses to the messages sent to them. They argued that adoption could be best understood as based on the social psychological characteristics of adopters and non-adopters. Early adopters were found to be more cosmopolitan, better educated, less risk-averse, and more willing to invest in new technologies than late adopters, pejoratively labeled 'laggards.' This perspective fitted well with the commitments of agricultural scientists to transforming agriculture, making it more efficient and more modern. However, it ignored the characteristics of the innovations. Often they were large, costly, and required considerable skill to operate and maintain. Not surprisingly, those who rejected the innovations lacked the capital and education to use them effectively.

Later studies challenged the diffusion theorists. First, critics of the Green Revolution asked questions about the appropriateness of the research undertaken (Perkins 1997). They noted that, although inexpensive in themselves, Green Revolution varieties were often parts of packages of innovations that required considerable capital investment well beyond the means of the average farmer. While acknowledging that yields increased, they documented the considerable rural upheaval created by the Green Revolution: growing farm size, displacement of both small farmers and landless laborers to the urban slums, declining status of women, declining water tables due to increased irrigation, and contamination of ground water from agricultural chemicals.

Others asked how agricultural scientists choose their research problems (Busch and Lacy 1983). They noted that science and commerce were necessarily intimately intertwined in agriculture, in the choice of research problems, in the institutional relations between the public and private sectors, and in the value commitments of scientists (often from farm backgrounds) and wealthier farmers. They challenged the engineering model of communication, seeking to substitute for it one drawn from the hermeneutic-dialectic tradition. Drawing on philosophers such as Jürgen Habermas and Hans-Georg Gadamer, they asserted that communications between scientists and the users of the products of agricultural research had to be able to debate fundamental assumptions about what constitutes a desirable future for agriculture as well as specific technical details.

Sociologists have also studied agricultural commodity chains, i.e., the entire spectrum of activities from the production of seed through to final consumption (Friedland et al. 1978). Such studies have examined the complex interaction between scientists and engineers involved in the design of new seeds and equipment and various constituent groups. Unlike the diffusion and induced innovation theorists, proponents of this approach have engaged in detailed empirical analyses of new technologies, challenging the assumptions of the designers. For example, both the tomato harvester and the hard tomato needed to withstand mechanical harvesting were built on the initiative of scientists and engineers in the public sector rather than to meet any need articulated by growers. Together, these technologies transformed tomato production in many parts of the world by reducing the number of growers and farm workers while increasing farm size. Given the limited employment opportunities of those displaced, critics question whether this was an appropriate investment of public funds.

In recent years, sociologists have devoted considerable attention to the new agricultural biotechnologies (e.g., gene transfer, plant tissue culture), especially those involving transformations of plants (see *Biotechnology*). It is argued that these new technologies have begun to transform the creation of new plant varieties by (a) reducing the time necessary for breeding, (b) reducing the space necessary to test for the incorporation of new traits from large fields to small laboratories, and (c) making it possible in principle to incorporate any gene into any organism. However, analysts note that the vast sums of private capital invested in this sector stem as much from changes in property rights as from any advantages claimed for the new technologies. In particular, they point to the advent of plant variety protection (a form of intellectual property right), the extension of utility patents to include plants, and the imposition of Western notions of intellectual property on much of the rest of the world.

Before these institutional changes, most plant breeding was done by the public sector. Private breeding was not profitable as seeds are both means of

production and reproduction. Thus, farmers could save seed from the harvest to use the following year or even to sell to neighbors. Put differently, each farmer was potentially in competition with the seed companies (e.g., Kloppenburg 1988). In contrast, once the new intellectual property regimes began to be implemented, it became possible to prohibit the planting of purchased seed developed using the new biotechnologies. Suddenly, the once barely profitable seed industry became a potential source of profits. Agrochemical companies rapidly purchased all the seed companies capable of engaging in research in hopes of cashing in on the new opportunities. The result has been a shift of plant breeding research for major crops to the private sector and strong prohibitions on replanting saved seed.

Another line of work has examined particular agricultural scientific practices and institutions. These include approaches to irrigation and chemical pest control strategies (Dunlap 1981). Moreover, as environmental concerns have taken on greater significance for the general public, studies of agricultural science have begun to merge with environmental studies.

5. *Philosophy*

In recent years applied ethicists have begun to take an interest in the agricultural sciences, asking a variety of ethical questions about the nature of the research enterprise, its relation to larger environmental issues such as the conservation of biological diversity, and the distribution of the products generated by the agricultural sciences. Some have asked whether it is even possible to engage in applied science without considering the ethical issues raised by the research agenda. Two new interdisciplinary professional societies emerged out of that interest: The Agriculture, Food and Human Values Society and the European Society for Agricultural and Food Ethics.

In a major contribution to the field, Thompson asserts that agriculture is dominated by what he calls the 'productionist ethic', the belief that production is the sole metric for ethically evaluating agriculture (Thompson 1995). From this perspective, derived from the philosophical work of John Locke, land not in cultivation is wasted. Agricultural scientists, themselves often from farm backgrounds, have posited this as self evident. This was combined with a positivist belief in the value-free status of science and a naive utilitarianism that assumes that all new technologies adopted by farmers are ethically acceptable. From this vantage point, all distributive issues are to be resolved by making the pie bigger. Similarly, environmental problems are defined as arising from inadequate technologies. In contrast, Thompson proposes an ethic of sustainability in which agricultural production is embedded in environmental ethics. Even the quest for sustainable systems, he notes, will be filled with both ironies and tragedies.

Indeed, within the agricultural sciences, ethical concerns have a higher profile than they have had in the past. In most industrialized nations there has been greater recognition of the need for inclusion of ethics and public policy issues in agricultural scientific education and research (Thompson et al. 1994). Moreover, the challenges to the focus on production from within the agricultural sciences have increased receptivity to ethical and policy questions. For example, agronomy, once the province only of scientists concerned to enhance annual yield, has become more fragmented as those concerned with sustainable agriculture and molecular biology have entered the field. Thus, questions of the goals and practices of research, previously ignored, have moved closer to center stage.

Philosophers have also examined ethical aspects of the new agricultural biotechnologies. Among the many issues of relevance is that of informed consent. In brief, it is often argued that consumers have the right to know what is in their food and to make decisions about what to eat on the basis of that information. From this vantage point, those nations that do not label biotechnologically altered foods violate important ethical norms.

In addition, critics of biotechnology have raised questions about the ethics of the use of animals in laboratory experiments, the development of herbicide resistant crops, the use of bovine somatotropin to enhance milk production in dairy cows, the insertion of toxins from *Bacillus thuringiensis* to create insect resistance in maize and potatoes, and the establishment of intellectual property rights in plants and animals.

6. *Future Directions*

There is little evidence that the fragmentation that has plagued studies of the agricultural sciences in the past is coming to an end. Disciplinary boundaries between the relevant academic fields remain high. Moreover, there are rigid institutional boundaries that still separate academic agricultural sciences from other fields of research. In particular, agricultural research and education tend to be found in specialized institutions, partly because of the specialized activities in which they engage, and partly because of the high cost of animal herds and experimental fields. Furthermore, those who study the agricultural sciences often do so from within the confines of schools and colleges of agriculture. In some institutions, this puts restrictions on what topics are considered appropriate for research.

See also: Biotechnology; Development: Sustainable Agriculture; Food in Anthropology; Food Produc-

tion, Origins of; Food Security; Green Revolution; Rural Geography; Rural Sociology

Bibliography

Barnett T 1977 *The Gezira Scheme*. Cass, London

Brockway L 1979 *Science and Colonial Expansion: The Role of the British Royal Botanic Gardens*. Academic Press, New York

Burmeister L L 1988 *Research, Realpolitik, and Development in Korea*. Westview Press, Boulder, CO

Busch L, Lacy W B 1983 *Science, Agriculture, and the Politics of Research*. Westview Press, Boulder, CO

Drayton R H 2000 *Nature's Government: Science, Imperial Britain, and the 'Improvement' of the World*. Yale University Press, New Haven, CT

Dunlap T R 1981 *DDT: Scientists, Citizens, and Public Policy*. Princeton University Press, Princeton, NJ

Friedland W H, Barton A E, Thomas R J 1978 *Manufacturing Green Gold: Capital, Labor, and Technology in the Lettuce Industry*. Cambridge University Press, Cambridge, UK

Fuglie K O, Ballenger N, Day K, Klotz C, Ollinger M 1996 *Agricultural Research and Development: Public and Private Investments under Alternative Markets and Institutions. Report 735*. USDA Economic Research Service, Washington, DC

Hayami Y, Ruttan V W 1985 *Agricultural development: An International Perspective*. Johns Hopkins University Press, Baltimore

Kloppenburg J R, Jr 1988 *First the Seed: The Political Economy of Plant Biotechnology, 1492–2000*. Cambridge University Press, New York

Perkins J H 1997 *Geopolitics and the Green Revolution*. Oxford University Press, New York

Plucknett D L, Smith N J H, Ozgediz S 1990 *Networking in International Agricultural Research*. Cornell University Press, Ithaca, NY

Rogers E M 1995 *Diffusion of Innovations*. Free Press, New York

Storer N W 1980 *Science and Scientists in an Agricultural Research Organization: a Sociological Study*. Arno Press, New York

Thompson P B 1995 *The Spirit of the Soil: Agriculture and Environmental Ethics*. Routledge, London

Thompson P B, Matthews R J, van Ravenswaay E O 1994 *Ethics, Public Policy, and Agriculture*. Macmillan, New York

L. Busch

Agriculture, Economics of

Agriculture is distinguished from other sectors of the economy by virtue of its production processes (biological), its economic organization (on farms), and its products (food and fiber). The importance of these distinctions for economic analysis is not always evident, but they have been sufficient to make agricultural economics a separate sub-discipline of economics, with its own journals and professional organizations.

1. Agriculture's Primacy in Economic Development

In most of the world historically, and in much of the world today, the economics of agriculture is the economics of subsistence: the effort to wrest the food necessary for survival from productive but fickle resources. The essential economics concerns how individuals carry out such efforts, and how families, villages, or other social entities organize their members for doing so. Economic development begins when agriculture generates production in excess of local requirements. Until the mid-nineteenth century the majority of the labor force in most countries of Europe was employed in agriculture. By the end of the twentieth century this percentage had been reduced to less than five in the richest countries. Similar patterns have emerged since 1950 in much of Latin America and Southeast Asia. Nonetheless, the World Bank (1997) estimates that 72 percent of the world's poor live in rural areas, and the prospects for economic development in agriculture remain a matter of worldwide concern.

A long-debated issue is whether agriculture is best viewed as an engine of growth, with investment in the sector an important source of economic progress; or as an economically stagnant source of labor to be mobilized more productively elsewhere as the economy grows. 'Dual economy' models, in which agriculture is economically distinct from the nonagricultural sector, can accommodate both views, depending on how they treat mobility of labor and capital between the sectors, and the processes of technical change and investment in each. Such models can account for the observation of huge outmigration from agriculture together with wage and income levels in rural areas rising toward urban levels after falling behind in the early stages of industrialization. But they do not provide useful empirical guidance for fostering economic development in areas of the world where it still is most needed. For those purposes, attention to microeconomic and sectoral detail is necessary. For reviews of economists' work on micro-level and aggregate questions, respectively, see Strauss and Duncan (1995) and Timmer (2002).

One of the most striking, and still to some extent controversial finding about the economics of traditional agriculture is the wide extent to which farmers in the poorest circumstances in the least developed countries act consistently with basic microeconomic principles. They follow economic rationality in the sense of getting the most economic value possible with the resources at hand; but the innovation and investment that would generate economic growth are missing (Schultz 1964). What is needed is to break out of the poor but efficient equilibrium by means of 'investment in high income streams,' mainly physical capital and improved production methods embodying new knowledge, and investment in human capital that

would foster innovation in technology and the effective adoption of innovations by farmers. Events such as the 'green revolution' that boosted wheat yields in India in the 1960s showed promising trends that have been sustained in many areas, but agriculture remained moribund through the 1990s in many places, notably in Africa and the former Soviet Union. No more important task faces agricultural economics today than explaining and finding remedies for this stagnation.

2. Farms

Farms range from individuals working small plots of land with only primitive tools to huge commercial enterprises. Every operating farm embodies a solution to problems of product choice, production technique, mobilization of inputs, and marketing of output. Many of the choices to be made involve non-market, household activities. Consequently, the economic analysis of farm households has become a major area of empirical investigation, calling upon developments in population and labor economics as well as the theory of the firm. Using these tools, agricultural economists have attempted to understand alternatives that have arisen in the economic organization of agriculture: family farms, cooperatives, plantations, corporate farming, state farms (see for example Binswanger and Rosenzweig 1986).

2.1 Organization of Production

A basic decision is whether to specialize or to diversify production among a number of products. The trend is strongly toward specialization. For example, 4.2 million US farms (78 percent of all farms) had chickens in 1950, but by 1997 specialization had gone so far that only 100 thousand did (5 percent of all farms). Linked with specialization is the issue of economies of scale in farming. Throughout the developed economies there is a general tendency for farm size to increase over the last century. Data on farmers' costs indicate that a primary reason is economies of size. Yet there are many instances of very large farms failing. Collective farms in the former Soviet Union, employing thousands of workers on thousands of hectares, became paradigms of inefficiency. And in some developing country contexts there is evidence that small farms use their resources more efficiently than large ones. Optimal economic organization with respect to both specialization and scale depend on technical and institutional factors, most importantly the following.

2.2 Land Tenure

Land is a valuable asset and is necessary for farming. Yet farmers in many countries are poor. Institutional arrangements have evolved to enable farmers to cultivate and claim returns from land they do not own. The main ones are cash rental and sharecropping. Cash rental encounters several problems: the tenant may lack the means or access to credit for payment in advance, bears all the risks of crop failure or low prices, and has an incentive to use the land in ways that increase current output at the expense of future fertility of the land.

Under sharecropping, a common practice in both developing and industrial countries, the tenant pays after the harvest in the form of a fraction of the crop harvested. The share paid to the landlord varies widely, generally between one-fourth and one-half of the crop, depending on the quality of the land, the labor intensity of the crop, and the value of non-land inputs, if any, contributed by the landlord. In addition, the literature on optimal land contracting finds that shares depend on agency costs, production efficiency of tenants and landlords, and how risk averse each party is. Sharecropping divides production and price risk between landlord and tenant, and obviates the need for payment in advance. But it retains the principal-agent problem in lacking incentives to maintain future land quality, and adds a disincentive for tenant effort in that the tenant receives only part of the tenant's marginal product, and adds an incentive for the tenant to under-report output and/or price received so as to reduce the rent paid. Such problems can be dealt with through monitoring, but that is costly. For a comprehensive review of the issues, see Deininger and Feder (2001).

The problems and costs of land rental increase the attractiveness of owner-operated farms, even if they have to be smaller. However, in many countries the institutions for private land ownership are not fully developed, nor are credit markets that would enable people with few initial assets to become landowners. In developed economies, land rental functions as a mechanism through which farmers can mobilize the land resources needed to achieve the least-cost scale of production. In the case of the US in 1997, only 21 percent of cropland was on farms fully owned by their operators.

2.3 Agricultural Labor

About one-half the world's labor force works in agriculture, as either a farmer or a hired worker. (For data by country, see World Bank, *World Development Indicators*, 1998, Table 1.) Hired labor is common even on family farms. Hired farm laborers in both developing and industrial countries are among the least well paid and most economically precarious workers. Seasonal workers live under especially difficult conditions in that they often dwell in temporary quarters and are minorities or immigrants, sometimes with dubious legal status, which makes them ripe for exploitation. The plight of migrant workers in many

Table 1
Annual costs and benefits of protection of agriculture in the EU, US, and Japan

	EU	US	Japan
Consumer costs due to higher prices	35	6	7
Taxpayer costs of subsidies	12	10	0
Gains to producers	31	12	12

countries has led to legislative and regulatory attempts to limit their numbers and improve their condition, and adds to a general sense that policies should be undertaken to enable landless laborers to gain access to land of their own and become farmers themselves. Nonetheless, hired labor remains a substantial fraction of the farm labor force in both rich and poor countries.

2.4 Credit and Input Markets

As agriculture modernizes, an increased share of resources used are purchased seeds, fertilizers, chemicals, energy, and capital equipment. Farms that cannot invest become unable to compete effectively. If modern agriculture is to be undertaken by farmers other than those who already possess substantial assets, well functioning credit markets are essential. A major problem for agriculture in many countries is limited access to either purchased inputs or credit. Recent thought about credit markets has emphasized the problems that arise because of asymmetric information between lender and borrower leading to credit rationing or missing markets. If potentially productive loans do not get made, farmers and the rural economy are unable to grow to their potential. This reasoning has led many countries to provide subsidized credit to farmers, but the informational problems that cause market failure have not been overcome with government involvement. In addition the political provenance of these programs causes new problems.

2.5 Price Determination and Marketing

A recurrent complaint worldwide is farmers' lack of market power as compared to those who buy from and sell to them. Farmers typically have only a few alternative outlets for their products, and inputs they buy, but the extent of monopsony or monopoly power that results remains unclear. In many countries farmers have established marketing and purchasing cooperatives to increase their market power. In the United States, the first half of the twentieth century saw far-reaching governmental attempts to reduce the market power of meat packers, grain traders, railroads, food wholesalers and retailers, and banks through antitrust action and governmental regulatory agencies. The developed countries of the world are today replete with such efforts, and developing countries have followed suit as appeared technically and politically feasible. It is nonetheless unclear whether the economic problems of farmers have ever been principally attributable to their lack of market power, or that cooperatives or regulatory institutions have increased farm incomes appreciably.

Important recent developments in marketing involve contractual arrangements between farmers and processors that take some input provision and marketing decisions out of the hands of the farmer. Such changes have gone furthest in broiler chicken production in the United States, where the processor is an 'integrator' who supplies the baby chickens, feed, veterinary and other services, technical information, and perhaps credit. The farmer (or 'grower') receives a payment schedule, contracted for in advance, consisting of a fee per pound of chicken delivered that is adjusted for an efficiency indicator as compared to other growers (but not changes in the market for chicken) in return for the grower's effort in feeding and managing the flock and providing the properly equipped chicken house. Virtually all broilers in the country are now produced under some variant of this type of contract. Under these arrangements productivity indicators of output per unit input have grown far faster for broilers than for any other livestock product and the US price per pound (live basis) of chicken relative to beef has declined from a ratio of 1.7 in 1940 to 0.5 in 1995. Similar production arrangements are increasingly prevalent for other meat animals.

2.6 Risk Management

In subsistence agriculture, crops failing or livestock dying place the farmer at risk of starvation. In commercial agriculture, fixed costs of crops sown and interest on debt means that losing even a portion of the crop, or receiving low prices, can easily generate negative cash flow. Steps a farmer can take to manage such risk include savings, diversification of enterprises, emergency borrowing, and purchase of hazard insurance against output risk, or some form of forward pricing against price risk. It remains open to question however how risk averse farmers are. Basic evidence that risk aversion is important is farmers' willingness to pay for insurance and their interest in pricing their output in advance. Observations that give pause about the importance of risk aversion are the many farmers who do not buy even subsidized crop insurance and who do not attempt to lock in a price for their output, even when contractual means for doing so are available. Nonetheless, evidence from developing countries suggests risk aversion of a magnitude that could readily impair farmers' willingness to invest in new production methods even when innovation would pay in expected value (see Moschini and Hennessy 2001).

3. Production and Technology

The evolution of world agriculture over the long historical record is tied principally to changes in technology. Throughout the developed world a large and sustained record of growth in agricultural productivity has been achieved. In the US case, after 50 years of steady but unspectacular growth, agricultural productivity accelerated markedly after 1940 to a pace of about 2 percent annually, well above the rate of productivity growth in manufacturing. Moreover, that rate of growth has been maintained for 60 years, with little evidence of the productivity stagnation that plagued manufacturing in the 1970s and 1980s (Fig. 1).

Economists have devoted much effort to measurement and analysis of productivity changes and farmers' decisions about input use. Nerlove (1958) developed a method of estimating both short-run and long-run output response to product prices. Empirical work using many variations on his approach over the last four decades has estimated generally small short-run effects of price. But in many cases the long-run effects are substantial. Griliches (1957a) provided the first fully developed economic analyses of the adoption of technology in his study of hybrid corn in the United States. Technical change and supply response have been merged in studies of 'induced innovation.' The chief causal factors identified in both supply and productivity growth have been advances in knowledge, improved input quality, infrastructure development, improved skills of farmers, and government policies. But the relative importance of these factors, even for the most-studied countries, still remains in doubt. A comprehensive review is Sunding and Zilberman (2001).

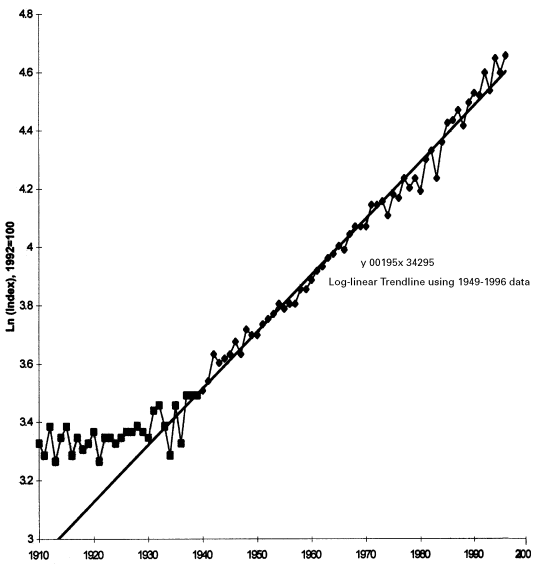

Figure 1
US Farm Total Factor Productivity Index

4. Demand and Markets

The world's population tripled in the twentieth century from the two billion of 1900. Agricultural production grew sufficiently not only to feed an additional four billion people, but also to provide the average person with a substantially improved diet. And, the incidence of famine and starvation among the world's poor has been greatly reduced. This capability was not evident 200 years ago when Malthus formulated the proposition that the earth's limited production capacity, coupled with the propensity of population to grow whenever living standards rose above the subsistence level, meant the inevitability of increasing food scarcity (and worse) over the long term. One of the most notable facts about the twentieth century is the failure of Malthusian pessimism to materialize. Nonetheless, the plausibility of elements of this view—basically the fixity of natural resources in the face of increasing population—is sufficient that the Malthusian worry resonates to the present day. It is therefore important to establish the circumstances under which food scarcity has ceased to be a salient social problem as well as the situations in which scarcity and famine remain a major cause of distress, and to understand why supply and demand have conspired to work out predominantly in the counter-Malthusian direction.

The single best indicator of food scarcity is the real price of staple commodities: cereals and other basic foods. Despite price spikes in wartime and the 1970s, the trend is for ever cheaper commodities. This trend primarily reflects lower real costs of production, a consequence of the productivity trends illustrated in Fig. 1. While all acknowledge the uncertainty of any forecast, expert participants in a recent comprehensive assessment of world food prospects were in broad agreement that the trend of lower real prices of staple food commodities is most likely to continue in the twenty-first century (Islam 1995).

An important factor in food demand is Engel's Law: the share of income spent on food decreases as consumers' incomes rise. The general rise of real incomes over the last century, coupled with the growth in agricultural productivity have meant an inexorable decline in agriculture's economic importance, and have been a source of chronic downward pressure on the economic returns of farmers. In many developing countries, especially former colonies whose economies became attuned to exports of primary products, declining commodity prices have been a key part of a bigger story of economic disappointment. Economic problems of farmers in both developing and industrial countries have kept agriculture firmly on the policy agenda almost everywhere.

5. *Government and Agriculture*

Political responses to problems of agriculture have generated a wide variety of government action. Four areas of activity warrant discussion: regulation of commodity markets; rural development policy; food policy; and resource and environmental policies.

5.1 *Commodity Programs*

Government intervention in agricultural commodity markets has been pervasive throughout recorded history. The primordial form of this intervention is taxation. With urbanization, implicit taxation of agriculture has arisen in many countries in the form of regulations intended to keep food prices from rising in times of scarcity. A sharp divide exists between the developing world, in which agricultural output is generally taxed, and the industrial world, in which agriculture is generally subsidized. This pattern of taxation and subsidy has had the unfortunate consequence of encouraging overproduction in industrial countries and discouraging investment in agriculture in developing countries, many of which have a comparative advantage in agriculture. Contrary to what one might have expected, the share of world agricultural exports accounted for by industrial countries increased from 30 percent in 1961–3 to 48 percent in 1982–4, with a corresponding decrease in developing countries (World Bank 1986, p. 10).

Not only does the protection of agriculture in industrial countries harm agriculture in developing countries, in addition each industrial country's protection makes it more costly for other industrial countries to maintain protection. The Common Agricultural Policy (CAP), created with the establishment of the European Economic Community in 1958, is notorious in this respect. The main policy instruments of the CAP go back to Britain's Corn Laws of the ninetieth century tariffs that maintain protection against imports by rising when world prices fall ('variable levies') and export subsidies to dispose of domestic surplus production (see Ritson and Harvey 1997). In the first two decades of its existence the CAP moved its members from being net importers to net exporters of wheat, rice, beef, and poultry meat. Other grain-growing countries, which also desired to maintain support prices for their producers, introduced or accelerated export promotion and subsidy programs of their own, notably the US Export Enhancement Program of the 1980s. The subsidy competition exacerbated a worldwide decline in commodity prices in the 1980s, increasing the costs of US 'deficiency payments' that made up the difference between legislated 'target' prices and market prices for grains. This in turn triggered massive acreage-idling programs; in 1985–7 about a fourth of US grain-growing land was idled.

The World Bank (1986, p. 121) assessed the annual costs and benefits of agricultural protection in the largest OECD countries as shown in Table 1 (in billions of dollars).

Note that the costs to consumers and taxpayers together far outweigh the producer (more specifically, landowner) gains, with the sum for the EU, US, and Japan being a net welfare loss of US$25 billion.

Accurate measurement of these gains and losses is difficult, but virtually all analysts estimate substantial net losses in the industrial countries and to producers in developing countries during most of the post-World War II period, and accelerating losses in the 1980s. This situation provided the stimulus for agricultural policies, after lengthy and contentious negotiations in 1986–93, to be subjected to internationally agreed disciplines that began to be implemented in 1995 under the auspices of the World Trade Organization.

Individual countries have also initiated moves towards less market distorting intervention in the commodity markets in the 1990s. In the developing world, substantive steps in deregulating commodity markets were taken in many countries of Latin America and East Asia; and in Africa many countries reformed and/or abolished marketing boards and related interventions. Most radically of all, beginning in the late 1980s (and before the breakup of the Soviet sphere in 1989), a renunciation of state control of farm enterprises occurred in China and throughout Eastern Europe and the former USSR. But the reforms have as yet achieved nothing near complete liberalization in either developed, developing, or transition economies, with the exception of New Zealand.

5.2 *Rural Development Policy*

A broader agenda of governments in promoting economic growth in agriculture and rural areas has more widespread support. Economists have generally concluded that provision of certain public goods and infrastructure investment has been crucial in the economic development of agriculture, and that the absence or deficiency of such governmental support is an obstacle, perhaps an insuperable obstacle, to economic growth in agriculture in countries where it has not yet occurred. The World Bank has taken a strong role in urging market liberalization in developing countries and at the same time proposing a broad program of public investment in pursuit of rural development (World Bank 1997).

5.2.1 Legal Institutions. The most fundamental economic service the State can provide is a system of law governing property and contracts, and protection from lawbreakers. This requirement is not of course peculiar to agriculture, but must be mentioned because legal institutions in rural areas are no-

tably weak in many transition and poor economies, and especially regarding use and control of farmland and water resources. In industrial countries, too, these institutions have to evolve in response to changes in technical and social realities, most notably in the 1990s the establishment of property rights and contractual procedures that bear on innovations in biotechnology.

5.2.2 Agricultural Research, Extension, and Education. Even with well established institutions fostering private sector research and development, research and information dissemination are likely candidates for public funding, and have long been so funded in many countries. Griliches (1957b) pioneered methods of estimating the costs and benefits of publicly supported research. Since then hundreds of studies in both developing and industrial countries have replicated his finding of extraordinarily high rates of return to public investment in research and the dissemination of knowledge through extension activities (Evenson 2001).

Since the work of Schultz (1964, Chap. 12) investment in schooling has been seen as a cornerstone of what is needed to improve the economic well-being of farm people, and of increasing agricultural productivity. Solid empirical evidence of the effects of education on farming has been hard to come by, however. Even so, there is widespread support for improved education in rural areas, recently with particular attention to the education of women. Evidence is strong that schooling improves peoples' earning capacity, so it is a promising remedy for rural poverty even if it causes its recipients to leave agriculture.

5.2.3 Rural Infrastructure. Governments in industrial countries have made major investments in roads, railways, shipping channels, and ports to provide remote areas with cheaper access to markets. Lack of such infrastructure is a major impediment to agricultural development in many parts of the world today. But we have nothing like the studies of returns to research to provide evidence on the rate of return to such investments, and the anecdotal evidence is replete with failure as well as success stories. Even more controversy swirls around investments in water projects. Irrigation was important in facilitating fertilizer response to the new grain varieties that triggered the green revolution of the 1960s, and is essential for opening up arid areas for production. At the same time, dams and irrigation works have been heavily criticized in recent years. Many cases of low or negative returns to large investments have been cited, and the environmental costs of lost habitat for endangered species and reduced water quality have been emphasized. Recent work by agricultural economists has emphasized improving institutions for water pricing and assignment of use rights more than further investment in large projects.

5.3 Food Policy

A chief source of governmental discrimination against agriculture in developing countries is a desire to keep food prices low for urban consumers. In industrial countries, too, attempts have been made via price ceilings and export restrictions to keep a lid on food prices in periods when they have risen sharply, as occurred in the commodity boom of the 1970s. More important ongoing policies address the regulation of food quality and safety, food assistance for poor people at risk of undernutrition, and famine relief. Chemical residues on food and the use of genetically modified organisms (GMOs) in agriculture were especially contentious issues in the 1990s.

An important bifurcation of countries today is between those in which food security remains a pressing national issue and those in which assurance of an adequate diet is the problem of only a small minority of the population. International food aid has become a permanent policy in industrial countries, particularly famine relief. Mobilization of funds for such efforts can be difficult except in cases of well publicized disasters, but the more salient analytical issues have involved the nature of famines and the effectiveness of alternative approaches to remedy the suffering and death they cause. It has become apparent that in most famines the problem is not so much physical unavailability of food as a lack of income with which to acquire food. This may seem a distinction without a difference but the implications are profound for the most effective administration of aid. It has been argued, for example, that it can be counterproductive simply to ship food products to be disbursed by local governments. The undesired result is distribution that too often scants those who most need the goods, and at the same time a depression of commodity prices and hence of the incomes of local farmers who produce goods that compete on the same market as goods brought in but lack income for an overall adequate diet. Generally, international donors have to be careful not to take actions harmful to local coping mechanisms, which in many poor areas are well developed from long and bitter experience. See Barrett (2002) for a comprehensive review.

5.4 Resource and Environmental Policy

The relationships between agriculture and water quality, soil and other resource depletion, wildlife habitat, and chemical contamination have become front-burner policy issues in industrial countries, and are

beginning to get attention in developing countries. A difference from the regulation of industrial polluters is that agricultural pollution sources are typically small, scattered, and difficult to monitor. Certain agricultural pesticides have been banned in industrial countries, but reasonably good substitutes have so far been available. Non-intensive uses of erodible or otherwise environmentally sensitive lands, has been fostered in the US and Europe by paying farmers to undertake recommended practices.

Nonpolluting and resource saving practices for developing countries have been promoted by international agencies as conducive to 'sustainability' of their productive resources. However, countries have resisted some of these ideas, such as restraints on opening up new land or eschewing large new dams and irrigation projects. The debate is difficult because of a lack of documentation that the loss of forests and conversion of other lands to agricultural purposes at rates now occurring is a mistake that will come to be regretted. In Europe and North America, an issue that has become prominent in recent decades is the conversion of farmland out of agriculture and into residential and commercial use in suburban areas, not so much out of concern about lost food production but rather because of the loss of open space and other community amenities that farming provides. Land use regulation and agricultural subsidies of various kinds have been introduced, most extensively in Europe.

5.5 Agricultural Politics

Why has agriculture been widely discriminated against in developing countries, and supported in developed countries? Evidence that the explanation is not country specific is that countries that have grown sufficiently rich to move from the developing to the developed category, largely in East Asia, have moved from taxing agriculture to subsidizing it. A large body of recent work has attempted to explain the strength and resilience of farmers' political clout in the richest countries, especially in Western Europe, Japan, and the North America. It is particularly notable in that this strength has been maintained even as the farm population has declined from one-fourth to one-half of the total population 50 years ago to 2 to 10 percent today. It is also instructive that some commodities are protected much more heavily than others within each country. Many reasonable hypotheses on these and related matters have been advanced, generally linked to interest group lobbying and democratic politics. Knowing more about agricultural politics is important because a governmental role is essential in many aspects of agricultural and rural development, yet governmental action in commodity support programs, trade restrictions, and other regulatory areas have imposed large social costs that are notably resistant to reform. The goal is governmental institutions that provide the services that contribute to sustainable development and that reform wasteful policies. The goal is far from being realized.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Development: Rural Development Strategies; Development: Sustainable Agriculture; Economic Geography; Food Production, Origins of; Green Revolution; Greening of Technology and Ecotechnology; Rural Geography; Rural Planning: General; Rural Sociology

Bibliography

Barrett C 2002 Food security and food assistance programs. In: Gardner B, Rausser G (eds.) *Handbook of Agricultural Economics*, Elsevier Science, Amsterdam

Binswanger H P, Deininger K 1997 Explaining agricultural and agrarian policies in developing countries. *Journal of Economic Literature* **35**: 1958–2005

Binswanger H P, Rosenzweig M 1986 Behavioral and material determinants of production relations in agriculture. *Journal of Development Studies* **22**: 503–39

Deininger K, Feder G 2001 Land institutions and land markets. In: Gardner B, Rausser G (eds.) *Handbook of Agricultural Economics*, Elsevier Science, Amsterdam, pp. 287–331

Evenson R 2001 Economic impact studies of agricultural research and extension. In: Gardner B, Rausser G (eds.) *Handbook of Agricultural Economics*, Elsevier Science, Amsterdam, pp. 573–627

Gardner B L 1992 Changing economic perspectives on the farm problem. *Journal of Economic Literature* **30**: 62–101

Griliches Z 1957a Hybrid corn: An exploration in the economics of technical change. *Econometrica* **25**: 501–22

Griliches Z 1957b Research costs and social returns: Hybrid corn and related innovations. *Journal of Political Economy* **66**: 419–31

Islam N (ed.) 1995 *Population and Food in the Early Twenty-first Century*. International Food Policy Research Institute, Washington, DC

Johnson D G 1991 *World Agriculture in Disarray*, 2nd edn. St. Martin's, New York

Moschini G, Hennessy D 2001 Uncertainty, risk aversion, and risk management by agricultural producers. In: Gardner B, Rausser G (eds.) *Handbook of Agricultural Economics*, Elsevier Science, Amsterdam, pp. 87–153

Nerlove M 1958 *The Dynamics of Supply*. Johns Hopkins University Press, Baltimore, MD

Ritson C, Harvey D 1997 *The Common Agricultural Policy and the World Economy*, 2nd edn. CAB International, New York

Schultz T W 1964 *Transforming Traditional Agriculture*. Yale University Press, New Haven, CT

Strauss J, Duncan T 1995 Human resources: Empirical modeling of household and family. In: Behrman J, Srinivasan T (eds.) *Handbook of Development Economics Vol. 3b*, Elsevier Science, Amsterdam, pp. 1883–2024

Sunding D, Zilberman D 2001 The agricultural innovation process. In: Gardner B, Rausser G (eds.) *Handbook of*

Agricultural Economics, Elsevier Science, Amsterdam, pp. 207–61
Timmer C P 2002 Agriculture and economic development. In: Gardner B, Rausser G (eds.) *Handbook of Agricultural Economics*, Elsevier Science, Amsterdam
World Bank 1986 *World Development Report*. Oxford, UK
World Bank Group 1997 *Rural Development: From Vision to Action*. (*Environmentally and Socially Sustainable Development Studies and Monographs Series 12*. World Bank, Washington, DC

B. L. Gardner

AIDS (Acquired Immune-deficiency Syndrome)

The AIDS epidemic of the 1980s shone a dark light on the assumptions of contemporary biomedicine. The world was confronted with a disease which it did not understand, could not treat, and which often attacked previously healthy young men. It had echoes of the great plagues of medieval times that decimated the cities of Europe and Asia. Theories abounded about its cause until it was discovered that a rare retrovirus instigated the destruction of human immune cells. Western medical science had recently been comfortable in the belief that infectious diseases were coming under scientific control and that the major health problems of the age lay in the understanding and treatment of chronic diseases.

What was particularly challenging about AIDS was that its impact on society extended beyond the usual concerns of biological medicine. Medical science was at the time appropriately basking in groundbreaking discoveries in the arenas of genetics and cellular and molecular biology. It was coming to better understand the mechanisms of disease and was devising innovative methods of diagnosis and treatment. What medicine did not do particularly well at the time was to attend the psychological, social, political, and ethical dimensions of illness. The onslaught of AIDS altered that, and forced the biomedical world to broaden its conception of illness and consider elements that went beyond the physical basis of pathology. The old authoritarian models of the doctor–patient relationship are gradually being replaced by a process where patients are empowered to gather more information about their condition, play a greater role in making treatment decisions, and become more self-directive in maintaining their health. All of this has taken place in an environment where new discoveries about virology, immunology, and treatment strategies are rapidly occurring.

The story of HIV infection and its culmination in the disease that is called AIDS starts in the USA in the early 1980s. Surprised physicians treating young gay men noted that they were falling ill and dying in increasing numbers. There was no clear diagonistic reason for this phenomenon, and a variety of guesses was ventured as to the cause. The number of reported incidents of this strange illness began to increase. Soon it was discovered that patients receiving transfusions for hemophilia and also intravenous drug users were being reported with this disease. Reports began coming in from Africa documenting that heterosexual men and women were likewise falling ill in alarming numbers. It became evident that the primary site of the pathology was the destruction of an important component of the immune system, the CD4 helper T cells which are responsible for mounting a critical defense against infectious agents. As a result of the loss of immune competency, an array of 'opportunistic' infections attacked the victim and instigated a wide array of illnesses which infiltrated many body organs, chiefly the lungs and central nervous system.

In the USA, the number of people contracting this disease rose over a span of two decades. By 1997 it was estimated that the prevalence of acquired immunodeficiency syndrome was approaching 900,000. AIDS became the leading cause of death among all Americans between 25 and 44 years of age. Approximately 40,000 new cases were reported yearly and the same number would die of complications of the disease. By the mid-1990s, the Centers for Disease Control estimated that the incidence of new cases had declined by six percent, but the prevalence of the disease had remained stable. (The number of new cases was approximately the same as the number of deaths by the late 1990s.) New treatments for this disease were rapidly emerging by the 1990s, but the high cost of treatment meant that many poor people and people of color did not receive the benefits of these medications.

1. The Etiology, Immunology, and Clinical Pathology of HIV/AIDS

There has been some confusion in the public mind about the use of the meaning of the concepts of HIV and AIDS. The term 'HIV' refers to a person's infection with the HIV (human immunodeficiency virus). This virus can be transported from one person to another through the exchange of certain body fluids (blood, semen, vaginal discharge). It enters certain immune cells in the body and *can* progressively destroy them over time. The presence of this virus can be detected by testing antibody titers which reveal the body's response to its presence. A person can be HIV+ and be asymptomatic for long periods of time, or show only mild flu-like symptoms after acquiring the virus. The human immunodeficiency virus belongs to a *retrovirus group* called *cytopathic lentiviruses*. This class of viruses insert their genetic material into a cell's genetic pool through a process called *reverse transcription*.

There are two types of HIV viruses which have been shown to cause AIDS in humans: HIV-1 and HIV-2.

These viruses share similar molecular structures and cause similar pathological disruptions. Currently, HIV-1 causes the majority of cases of AIDS throughout the world, while HIV-2 is found mostly in Africa. There are many different subtypes among the HIV-1 strain. These viruses can rapidly change their molecular structure within the body, making them difficult for the immune system to destroy. The mutability of these viruses also makes them difficult targets for preventive immunization strategies.

The term *AIDS* refers to a critical stage of the HIV infection when a large number of CD4 helper lymphocytes have been destroyed and the body is not able to mount an effective immune defense against secondary pathogens or the toxic effects of the HIV virus itself. The first symptomatic signs of immune breakdown may not occur for a number of years after the acquisition of the virus; latency might vary between four and ten years. The HIV virus can provoke diseases that involve the lymph nodes, lungs, brain, kidneys, and the abdominal cavity. There is no universal sequence of clinical symptoms, but the most frequent presentation can include enlarged lymph glands, pneumonia-like symptoms, decline in blood counts, fungal infections, as well as diarrhea, weight loss, bacterial infections, fatigue, and disorders of the central nervous system. The Centers for Disease Control in 1993 compiled a list of conditions which defined AIDS. This included such diseases as toxoplasmosis of the brain, tuberculosis, Kaposi's sarcoma, candidiasis infection of the esophagus, cervical cancer, and cytomegalic retinitis with loss of vision. Conditions on the list include opportunistic infections, or diseases that would otherwise not occur in the presence of a healthy immune system.

In order to best understand this devastating disease and its treatment, it is important to understand the relationship of the virus to the immune system. The HIV virus is known as a retrovirus, which means that it can alter the flow of genetic information within a cell. It uses the enzyme *reverse transcriptase* for utilizing viral RNA as a template for producing DNA. In most cells, DNA produces RNA as a genetic messenger. The sequence of infection after the entry of the virus through exchange of body fluids is the following: (a) the virus attaches to a host immune cell; (b) the virus sheds its molecular coat and begins the processes of reverse transcription; (c) viral DNA is integrated into the host cell, causing transcription and translation of the viral genetic code into viral protein, and inducing the host cell into producing more copies of the invading virus; (d) the newly formed viruses are released into the blood stream, with the death of the host cell and the re-invasion of new immune host cells.

The helper T cells have the CD4 surface receptor, which has a high affinity for the HIV surface protein gp120. However, the HIV virus can attach to other cells including macrophages and monocytes (other immune cells), as well as cells in the intestines, uterine cervix, and Langerhans cells of the skin. Researchers have found that chemicals called cytokines also play a significant role in facilitating viral invasion. Interestingly, there are certain mutant strains of cytokine genes that actually prevent the entry of the HIV virus into the target CD4 cell. The presence of these genes in certain people may explain why some individuals remain resistant to HIV infection despite exposure to virally loaded body fluids and why some people show slower rates of disease progression.

Not all infected cells immediately produce new viral copies which destroy the cell and infect many others. Some cells may remain dormant without producing new viral copies. There may be adjuvant factors that convert latent infected cells into ones that produce viral copies. It is speculated that coexisting infections with their additional antigen load may have a facilitating effect, as may certain drugs, possibly stress, fatigue, etc. Anything that reduces immune competency may become a cofactor. Many persons who are infected try to find ways to strengthen the immune system. These attempts may include meditation, exercise, herbs, dietary supplements, prayer, and support groups. Whether these have any long-lasting ameliorative effects is uncertain, but they may give the person a greater sense of self-maintenance and control over a devastating illness.

2. *Issues of Testing for HIV and Stages of the Infection*

When HIV infection was first recognized in the early 1980s, it was most prevalent among young men who had sex with other men and were living in large metropolitan cities. Prevention programs and the availability of tests for the infection have significantly decreased the rate of new AIDS cases among this population. Unfortunately, the rate of new AIDS cases reported in the USA has risen dramatically among women, African American men, and Latinos. Adolescents have shown a marked increase in AIDS during the decade of the 1990s. The major routes of infection with HIV have been shown to be through fluid-exchanging sexual behavior; use of HIV-contaminated needle exchange; from mother to infant during pregnancy, delivery, or breast feeding; and through HIV-contaminated blood products passed during transfusions. The development of accurate testing procedures for the infection became important both to assist in the diagnosis of the disease and to protect the nation's blood supply.

Various means of testing for the presence of the HIV virus or its effects have been developed since the discovery of the disease. The virus can be cultured directly from the blood, but the most efficient means of diagnosis has come from detecting antibodies to the virus. Such testing can determine whether the body

has come into contact with the virus and has mounted an immune defense. However, antibody testing can produce both false positive and negative results. Most tests take two weeks to receive results, but more rapid same-day tests are available. The usual procedure is for the patient (or blood sample) to be tested first using an enzyme-linked immunosorbent assay (ELISA), which is very reactive to HIV antibodies. False negative results are uncommon using this procedure. If the ELISA test is positive, it must be repeated. False positive results are more common using this approach. If a second ELISA test is also positive, a Western Blot procedure using an immunofluorescent assay is used as confirmation. It has a high level of specificity for detecting protein antigens of the HIV virus. New testing procedures have been developed which can also test saliva and urine, which bypasses the use of a blood draw.

Though there are reasonably accurate and fast methods of testing for the HIV virus, not all people who suspect that they have been exposed to the virus seek testing, and not all who do arrange for it early in the progress of the infection. There is a variety of reasons why this is so. Some people are simply not aware of the risks of infection and the procedures of arranging for testing. Others have been properly educated but, because of fear, denial, shame and concern about loss of privacy and public exposure, avoid learning if they have been infected.

Often when a person has received a reliable diagnosis of HIV infection, they experience a variety of negative psychological symptoms. They can become anxious, are often depressed, may feel guilt about their past behavior, and are worried about physical deterioration and eventual death. If they have observed friends experience the progress of the disease, they may have frightening images of what may happen to them. Suicidal thoughts are often present immediately after a diagnosis is made and the patient becomes exquisitely sensitive to any physical symptoms.

It is vital that such persons receive comprehensive, sensitive, and accurate medical and psychological counseling upon receiving news of a positive test. Many persons at this point in the illness feel isolated, shamed, and unable to talk about their fears and questions. Since the onset of the epidemic, many outstanding clinical and community programs providing vital information and emotional support have been established. Since significant advances in therapies have been developed, with a marked increase in survival rates, early detection and humane medical and psychological interventions is essential for all.

3. *The Clinical and Immunological Progression of the Illness*

The finding of a positive antibody test for HIV does not predict an inevitable disease course for all people. Some people exhibit only mild symptoms initially and then remain totally asymptomatic throughout their lifetime. Others demonstrate severe symptoms shortly after the virus is detected and others show no sign of illness for eight to ten years after being diagnosed as positive. HIV/AIDS is a disease of uncertain symptoms, long quiescent periods for many, and often a devastating end state.

There are many reasons to explain this variability. Some may have a genetic resistance to the virus, a robust (innate and adaptive) immune system and have been exposed to a less virulent strain of the virus. For others, diagnosis may occur long after they were exposed to the virus and their immune system may be severely compromised.

People told that they may develop AIDS face many uncertainties, complex medical and lifestyle decisions, and the need to adapt to changing medical conditions throughout their life. These decisions involve trusting communications with health care workers, friends and family, social support systems, and work associates. For some, the complexity of these decisions may prove overwhelming and they may make many poor health and lifestyle decisions. For others, the illness may reactivate old psychiatric problems or create new neuropsychological symptoms. Indeed, one of the most provocative complications of the disease is the emergence of cognitive disabilities for the patient.

As noted earlier, the harm caused by the HIV virus is instigated by its entry into the CD4 helper T lymphocyte and its subsequent capture of its genetic mechanism to produce many new viral copies. When a full-fledged infection is in progress, billions of new viral particles can be found in the human blood stream. This causes a drop in the number of CD4 cells, which usually have a level of at least 800 mg/cubic millimeter of blood. Counts below 500 cells/cm are considered serious and below 200 will set the stage for dangerous opportunistic infections. Periodically testing the concentration of CD4 cells has been an important means of measuring and predicting the course of the disease. More recently, tests have been developed to measure the viral load directly in the blood stream. These sensitive tests are considered a most reliable measure of the progress of the infection.

4. *The Progression of the Illness*

Following the entrance into the immune cells, there is a dramatic drop in CD4 counts and usually the onset of symptoms that can resemble flu or mononucleosis. These can include fever, fatigue, enlarged lymph glands, headaches, rashes, and muscle aches. These symptoms usually resolve in a few weeks, as the immune system resists the viral spread. It does so by triggering CD8 cytotoxic cell responses which destroy infected cells and by stimulating an antibody response to the virus. This reaction binds and removes many

HIV particles from the blood. These body defenses reduce the viral load but rarely eliminate the virus from the body. There tends, in many cases, to be an equilibrium established between the immune defenses and the viral level. This so-called 'set point' can be quite different from patient to patient. Whether a patient will become seriously symptomatic depends upon the balance between viral activity and immune competence.

When viral activity gains the upper hand over immune defenses, serious symptoms will occur. Depending upon when the diagnosis was first made, the set point can vary from four to ten years. Some patients, called 'non-progressors,' may never show serious symptoms and maintain good CD4 levels. Others may reveal compromised CD4 counts, have a bout of opportunistic infections, and yet be asymptomatic thereafter. Once the CD4 count drops below 200, people usually develop the many complications of AIDS. The immune system is no longer able to contain the viral spread and organisms which it usually can control begin producing dangerous infections. Most common are pneumocystis Carinii, pneumonia, and toxoplasmosis. However, many other organisms can also be activated and cause multiple organ damage. When the brain is affected in the end stages of the disease, delirium, dementia, and a variety of motor impairments can occur.

In other countries where this disease is common, there may be a different array of opportunistic infections found. In places such as Haiti and Africa, one sees more candida infections and crytococcal meningitis. Intestinal disorders with diarrhea and wasting are also common. Both abroad and in the USA, tuberculosis associated with HIV has become a major threat.

5. *The Drug Treatment of HIV Infections*

There exists at present no vaccine to prevent HIV/AIDS nor any medication to 'cure' the illness once it has been contracted. However, astonishing progress has been made since the late 1990s to more fully understand the mechanisms of viral replication and develop drugs to control proliferation of the organism. This has resulted in a sharp decline in the death rate of those taking these drugs. However, the cost of effective medication is extremely high (up to $12,000 a year). Because not all health plans will pay for such treatment, and some people are not insured, they cannot receive such pharmacological help. Many AIDS sufferers abroad, at present, have no hope of receiving these medications. This has created serious national and international concerns about creating a two-class partition concerning distribution and access to these medications. Such questions will inevitably arise as new, expensive, and sophisticated treatments are developed to treat other chronic conditions.

Since the introduction of these new treatment regimes, the death rate for AIDS in the USA between 1996 and 1997 dropped by 44 percent, and the number of HIV-caused hospitalizations has been significantly reduced. This was accomplished through a more detailed understanding of the molecular activity of the virus as it enters a human cell. Three enzymes are critical for viral replication and proliferation. These are *reverse transcriptase*, which converts viral RNA into double-strand DNA; the enzyme *integrase*, which splices the HIV DNA, thus converting it into a chromosome in the host cell, where it functions like a gene; and finally, a *protease enzyme*, which packages viral RNA into new virus particles. The new classes of drugs operate either by blocking viral replication or by inhibiting reverse transcriptase of the HIV protease. The original drug to block reverse transcription was developed in 1987 and was called *zidovudine* (*AZT*). This drug prevented the completion of the viral DNA strand in the human cell. Such drugs are called nucleoside analogs; later, *nonnucleoside reverse transcriptase inhibitors* were developed which also inhibited retroviral activity. More recently, a powerful new class of *protease inhibitors* was developed which prevents the division of newly produced HIV proteins.

The development of these powerful drugs is important because research has shown that the HIV reproduction is robust early in the disease, but remains stable because of the immune response which produces high numbers of CD4 cells. A strong initial response of CD4 cells facilitates the body's subsequent production of a CD4 subset which react selectively to HIV. Medical research has determined that the level of the viral load in the body is highly correlated with ultimate prognosis. If the viral level can drop to an almost undetectable level, the likelihood of developing opportunistic infections and other complications of AIDS declines.

Important as the development of these effective medications have been, they still further complicate the lives of HIV patients and create a myriad of complex decisions for them. As mentioned, these drugs are expensive and not always provided by insurers. They must be taken on a very rigorous schedule and in extremely large doses. Failure to maintain the timing of a dosage of the drugs may result in their being ineffective, and possibly producing viral resistance. Even when taken correctly, the drugs may not be effective, leading to disappointment. The usual prescription is for a combination of two nucleoside analogs and a protease inhibitor. New combinations are being developed and tested which promise greater potency. The cocktail approach is used because viral resistance is reduced by multiple drug assault.

Besides the financial burden and complexity (up to 16 pills daily) of the 'cocktail' regimen, other difficulties include negative side effects. These may encompass anemia, neuropathy, headache, diarrhea, rashes, and hepatitis. The complexity of this medi-

cation approach requires a close collaboration between the treating physician and the patient. It underscores the necessity of having the patient be an active participant in his or her treatment.

6. *Neuropsychiatric and Psychosocial Issues*

Although AIDS can cause a wide number of dangerous medical complications, none are more feared by persons with HIV infection than the neuropsychiatric disorders. The HIV virus can cause damage to the central nervous system itself and open the door to a myriad of opportunistic brain infections. The possible end-stage of delirium and dementia with loss of personality and body function control is a grim vision of their future.

AIDS can not only introduce a number of central nervous system (CNS) disorders, but it can also cause the reactivation of previous psychiatric illness. There is a great number of people with AIDS who also have a previous psychiatric history. Often these people may be less capable of taking reasonable precautions with regard to unsafe sex and intravenous drug use, rendering themselves vulnerable to infection.

AIDS involvement of the CNS has been found in more than 50 percent of people who are HIV+ but asymptomatic, and over 90 percent on autopsy of all AIDS patients show evidence of neuropathology. Most of the pathological findings occur in the subcortical regions of the brain. Inevitably, patients with involvement of the CNS show symptoms of cognitive impairment, movement problems, and behavioral difficulties. These difficulties may include a change in personality, withdrawal and apathy, inappropriate emotional responses, sharp mood swings, mania or suicidal impulses, and hallucinations. Often the brain involvement leads to a severe inability of the patient to carry out activities of daily living and requires institutional care or intense home assistance. There is belief among some clinicians that early diagnosis and vigorous drug treatment can delay or even prevent later AIDS dementia complex. This will depend upon the extent of CNS damage existing at the time of initiation of treatment. Lithium and neuroleptic medications have been used to systematically treat people who become manic or agitated.

The exact means by which the AIDS virus causes damage in the CNS is not certain. There is one theory that suggests that the virus damages the microglial cells in the brain which serve as connections between the neurons. Another theory points to the possibility that HIV infected cells are the source of toxins which cause the dementia.

It is often difficult to pinpoint whether mental symptoms are the direct effects of the HIV virus or produced by the multitude of opportunistic infections (particularly toxoplasmosis) that can invade the brain. The side effects of drugs used to treat the disease may cause cognitive and signs of delirium as well as the impact of frequent high fevers.

Both the organic and social stresses of HIV/AIDS are associated with the emergence of psychological distress and psychiatric symptoms. The most common psychological disorder associated with HIV infection is an adjustment disorder with features of anxiety and depression. Major depression is often observed among HIV+ patients. This is most common among those with a previous history of depression and those who are isolated and have little social support. A feeling of hopelessness and lack of personal control over the development of the disease is found among these depressed patients. Patients with HIV/AIDS experience, during the course of their illness, a variety of losses and other stresses which increase their vulnerability to psychiatric disorders. These includes loss of employment, death or illness of friends, disengagement of family members, financial losses, loss of sexual partners, abandonment of future goals, reduction of physical function, and failing cognitive abilities.

The occurrence of depression, anxiety, somatization disorders, suicidal ideation, and substance abuse can be traced to the reaction to such losses, or the fear of them. Such psychological distress may by itself compromise the immune system. Positive mental attributes acquired through therapy, support groups, prayer, etc., may modulate the psychological pain of the patient and help to improve the quality of daily living. Research has not yet determined whether developing more adequate coping mechanisms and enhanced self-esteem will effectively alter the immunological course of the illness.

Like so much else in this disorder, there is no common psychological pathway that all patients follow. Their previous psychiatric history, adaptive responses to the virus, social support, financial resources, access to good medical care, response to medication, will all play a role in helping persons with AIDS achieve positive mental equilibrium. Supportive therapy can help patients deal with fear, uncertainty, and a sense of self-recrimination. Such help can occur in a professional setting, through community groups, religious counseling and in a variety of other innovative venues. The important consideration is to help people to feel that they are not worthless, socially shunned, or without something valuable to contribute to friends and society.

7. *Policy and Ethical Issues*

This epidemic has emphatically raised the question of what is the responsibility of governments, pharmaceutical companies, and insurance plans for protecting and treating people of very limited financial means. AIDS involves populations of patients who are often out of the spotlight of public attention or who have

been morally condemned because of certain behavior characteristics. The most striking example of an ignored population is the many millions of heterosexual patients who have fallen ill with AIDS in Africa. This disease has disrupted families, created national economic disaster (particularly in agriculture), and has overtaxed the medical resources of extremely poor countries. Yet in the richest industrial and technological countries in the world, there is little concern or knowledge about the problems on this continent. The ethical responsibility for providing the benefits of modern medicine and pharmacology to people in a distant and largely unknown continent is now beginning to emerge in the public's awareness.

The USA's record in responding to the needs of minorities and underprivileged people with the infection even in its own country is not a cause for optimism. Medical care and treatment for people with AIDS is expensive and as a chronic disease, its costs mount over time. Infected people often lose their insurance, are no longer able to work, and exhaust their financial resources quickly. Yet at the beginning of the twenty-first century, there is at best a patchwork of policies to finance the development of drugs and make them available to those in need. Contemporary policy must also take into account the many women and, at times, their children who are infected with HIV. They require many additional social and educational services to deal with their medical and social problems. Attention must also be given to the special needs of adolescents who are at greater risk for contracting this disease.

Compounding the problems of money are issues of protecting both the privacy of people with the HIV infection, while also safeguarding the public's health. Generally, it has been felt that well-conceived educational programs can play a major role in both prevention and helping infected persons to make ethical decisions about disclosing their condition to others. Related to this question are issues regarding blood bank testing, protection of health care workers, and disclosure to prospective and current sexual partners.

In the USA, diseases which are sexually transmitted have become a metaphor for troubling issues about 'moral' behavior. It raises questions about the values of the society, parents' control over the behavior of their children, and the images which are conveyed by the media to the public. Among some groups, sexually transmitted diseases are seen as a punishment for immoral behavior. Much of the public's response to AIDS—even after two decades of familiarity with the disease—is shaped not by its medical and biological characteristics, but by US social and cultural attitudes towards the behaviors associated with contracting the illness. People's willingness to help those who are afflicted is molded by their social perspectives. If the public disapproves of the people who have contracted this disease, they are reluctant to provide the medical care, drugs, shelter, and the social support that they need.

The HIV/AIDS epidemic has also raised questions about the right to access to medications that have not yet met Food and Drug Administration standards for testing and release to the public. Should drugs that have not yet proven their safety be given to people who might otherwise die?

What has emerged, however, is the belief that patients and their advocates have a right to be at the table where scientific and public health decisions are being made. Patients must participate in decisions regarding the initiation and termination of treatment, rights of privacy concerning their condition, informed consent, and access to new forms of treatment.

The lessons learned in understanding AIDS are reshaping views of the roles of doctor, patient, family, and community. Hopefully, such new knowledge will provide a more humane and comprehensive attitude for the care of patients with all diseases. Illness is not an event that happens in just one person's body. Its consequences are part of the social fabric. Ethical consideration, as well as advances in biological expertise, must inform future policy and health care decisions.

See also: AIDS, Geography of; Chronic Illness, Psychosocial Coping with; Depression; HIV Risk Interventions; Homosexuality and Psychiatry; Mania; Mortality and the HIV/AIDS Epidemic; Sexual Risk Behaviors; Sexually Transmitted Diseases: Psychosocial Aspects

Bibliography

Ammann A J, Volberding P A, Wofsy C B 1997 *The AIDS Epidemic in San Francisco: The Medical Response, 1981–1984*, Vol. 3. Regents of the University of California, Berkeley, CA

Aral S O, Wasserheit J 1996 Interactions among HIV, other sexually transmitted diseases, socioeconomic status, and poverty in women. In: O'Leary A, Jemmott L S (eds.) *Women at Risk*. Plenum, New York, pp. 13–42

Aversa S L, Kimberlin D 1996 Psychosocial aspects of antiretroviral medication use among HIV patients. *Patient Education and Counseling* **29**: 207–19

Bednarik D P, Folks T M 1992 Mechanisms of HIV-1 latency. *AIDS* **6**: 3–16

Bennett R, Erin C A (eds.) 1999 *HIV and AIDS: Testing, Screening, and Confidentiality*. Oxford University Press, Oxford, UK

Cao Y, Qin L, Zhang L, Safrit J, Ho D 1995 Virologic and immunologic characterization of long-term survivors of human immunodeficiency virus type 1 infection. *New England Journal of Medicine* **332**: 201–8

Capaldini L 1997 HIV disease: Psychosocial issues and psychiatric complications. In: Sande M A, Volberding P A (eds.) *The Medical Management of AIDS*, 5th edn. Saunders, Philadelphia, pp. 217–38

Carey M P, Carey K, Kalichman S C 1997 Risk for human immunodeficiency virus (HIV) infection among persons with severe mental illnesses. *Clinical Psychology Review* **17**: 271–91

Centers for Disease Control (CDC) 1999 Guidelines for national human immunodeficiency virus case surveillance, including monitoring for human immunodeficiency virus infection and acquired immunodeficiency syndrome. *Morbidity and Mortality Weekly Report* **48** (no. RR-13): 1–31

Cohen P T, Sande M A, Volberding P A (eds.) 1990 *The AIDS Knowledge Base: A Textbook on HIV Disease from the University of California, San Francisco, and the San Francisco General Hospital*. Medical Publishing Group, Waltham, MA

Fauci A S, Bartlett J G 2000 Guidelines for the use of antiretroviral agents in HIV-infected adults and adolescents. http://www.hivatis.org/guidelines/adult/text (8/24/00)

Kalichman S C 1998 *Understanding AIDS: Advances in Research and Treatment*. American Psychological Association, Washington, DC

Lyketsos C, Federman E 1995 Psychiatric disorders and HIV infection: Impact on one another. *Epidemiologic Review* **17**: 152–64

McArthur J C, Hoover D R, Bacellar H et al. 1993 Dementia in AIDS patients: Incidence and risk factors. *Neurology* **43**: 2245–52

National Institutes of Health 2000 Summary of the principles of therapy of HIV infection. *NIH Guidelines: Report of the NIH Panel to Define Principles of Therapy of HIV Infection*. http://www.hivpositive.com/f-DrugAdvisories/NIHguidelinesJune/summary.html

Price R W, Perry S W (eds.) 1994 *HIV, AIDS, and the Brain*. Raven Press, New York

Reamer F G (ed.) 1991 *AIDS and Ethics*. Columbia University Press, New York

Shernoff M (ed.) 1999 *AIDS and Mental Health Practice: Clinical and Policy Issues*. Haworth Press, New York

Ungvarski P J, Flaskerud J H (eds.) 1999 *HIV/AIDS: A Guide to Primary Care Management*. Saunders, Philadelphia

Zegans L S, Coates T J (eds.) 1994 *Psychiatric Manifestations of HIV Disease. The Psychiatric Clinics of North America*. Saunders, Philadelphia

L. S. Zegans

AIDS, Geography of

1. Introduction

The passage of a disease agent between infectious and susceptible individuals traces a pathway in space and time along which the geography of an epidemic unfolds. To understand this diffusion requires knowledge of both those epidemiological characteristics of the agent which facilitate its transmission and societal reactions to the ensuing disease outcomes. The advent of the acquired immuno-deficiency syndrome (AIDS) and the isolation of its agent, the human immunodeficiency virus (HIV), have challenged our past experience of this inter-relationship. Unlike most other infections, the incubation period from contracting HIV to the onset of AIDS is long and allows the potential for infectious individuals to circulate freely in a community for many years unaware of their status. Similarly, the likelihood of infection has been differentiated among the population and has displayed marked variations by both risk behavior and geographical location. Given these traits, this article interprets the evolving geographical epidemiology of HIV/AIDS alongside the efforts that have been made to contain the spread of infection. In particular, the evaluation of disease prevention stresses the distinction between natural control, where the frequency of communicable events in a given population is insufficient to support sustained transmission, and direct interventions against HIV, taken either by official or voluntary agencies. Last, the implications of the current downturn in AIDS incidence in many countries are discussed.

2. Spatial Epidemiology

2.1 The Host–HIV Relationship

Following its separate and disputed isolation by French and American research teams led by Luc Montagnier and Robert Gallo, respectively, the agent of AIDS was given the agreed name the human immunodeficiency virus by the International Committee on the Taxonomy of Viruses in May, 1986. This research demonstrated that HIV is a retrovirus able to convert its own genetic materials into similar materials found in human cells. In particular, the host cells for HIV in the human body are lymphocytes known as T4 cells, which take on a surveillance role in the immune system with the capability to suppress alien infections. Initially, HIV penetrates some of these cells and then lies dormant until the body encounters some new infection. Then, this event stimulates the production of more HIV in place of the infected host T4 cells. It is believed that the recurrence of this process gradually damages the immune defenses and renders the body more vulnerable to other diseases.

This biological sequence underpins the host–agent relationship for HIV, which describes the expected timing of various transitions in human disease status from first infection to the onset of AIDS. This relationship is initiated when an infected individual's body fluids, such as blood, semen or cervical secretions are passed directly into the bloodstream of a susceptible individual. Following this occurrence, antibodies capable of suppressing HIV appear in the host within about eight weeks, during a process known as seroconversion. At this juncture, the host is thought to be maximally infectious and might develop symptoms of an illness resembling glandular fever. After seroconversion, the individual enters the symptomless chronic phase stage of the relationship when these antibodies diminish the host's power to infect others. Early estimates for the duration of this phase ranged between two and eight years but have since been revised upwards with improvements in surveillance and ther-

apy. HIV continues to destroy T4 cells throughout this phase, which is terminated when a final increase in host infectivity signals the imminent collapse of the immune system. These events initiate the patent period when the host becomes susceptible to one of a number of opportunistic infections (pneumonia, thrush, shingles) or malignancies (Kaposi's sarcoma) that characterize AIDS and are often fatal within about two years.

2.2 *The Pandemic Pathway*

This term refers to the timing displayed by an infectious disease agent as it diffuses outwards from its source area to other countries around the world. Establishing this pathway for HIV/AIDS, however, has not proved to be easy (Shannon et al. 1991, Smallman-Raynor et al. 1992). One key date in this progression is the first clinical diagnosis of AIDS made in a New York hospital in 1979. The incubation period, however, indicates that HIV must have been present long before this diagnosis and many investigations have attempted to establish the prior history of the infection. One retrospective investigation, for example, has linked the incidence Kaposi's sarcoma observed in some young males in 1882 to a prototype AIDS virus (Root-Bernstein 1989). Such early dating, however, does not necessarily support the continuous transmission of HIV in humans and might reflect sporadic outbreaks attributable to rare mutations of older and weaker strains of the virus into more virulent forms. A more reliable indicator of sustained transmission is the serological evidence obtained from infected individuals, which has implied an HIV epidemic probably began in Zaire around 1959 (Gotlieb et al. 1981). Moreover, AIDS was almost certainly present in Central Africa throughout the 1970s but was diagnosed as a wasting condition known colloquially as Slim's disease. These findings are consistent with other circumstantial evidence suggesting that strains of HIV originated in West and Central Africa owing to cross-species transmission (through eating and hunting accidents) of immunodeficiency viruses present in green monkeys and chimpanzees whose habitats are roughly coincident with the earliest identified areas of HIV endemicity.

Serological dating of viral strains has also established that, prior to 1979, HIV was transferred from Africa to the Caribbean by the early 1970s, and then entered the USA via San Francisco and New York during the mid-1970s (Li et al. 1988). In addition, contacts with infecteds in all these areas established the circulation of HIV in Western Europe before 1980, especially in France and Belgium with their strong colonial links to Central Africa (Freedman 1987). Since this era, transfers have been estimated from official records of the first diagnosis of AIDS or positive blood test for the presence of HIV. Such information has indicated the eventual entry of HIV into Asia, which was heralded by the first recorded diagnoses of AIDS in Thailand in 1984 and in India in 1986. Subsequently, HIV/AIDS has been recorded in most countries such that the World Health Organisation (WHO 1998) currently estimates the cumulative global incidence as 13.9 million for AIDS and 33.4 million for HIV.

2.3 *Localizing Elements*

A distinctive feature of the pandemic has been the variety of risk behaviors that have become associated with HIV transmission. In the USA, the early incidence of AIDS was almost exclusively among homosexual men and attracted the label 'gay plague.' The inappropriateness of this tag, however, was soon to become evident. Subsequent investigations of the African epidemic demonstrated the majority of infectious contacts were between heterosexuals while, during the mid-1980s, cases began to appear in the USA and Europe among intravenous drug users (IVDUs) who share the syringes they use for drug injection. More or less simultaneously, two further modes of transmission were recognized that do not necessarily entail direct contact between those identified to be at risk. First, the absence until 1984 of an effective screening test for the blood clotting agent Factor 8, led many hemophiliacs to become HIV positive through the receipt of contaminated blood products. Second, babies born to HIV positive mothers were observed to be at risk from transmission *in utero*.

These revelations encouraged the view that HIV/AIDS constituted a set of discrete epidemics, each characterized by a particular behavior such that the vast majority of infectious contacts are presumed to occur between those who share the same risk. An important geographical representation of this construction of the epidemic was the WHO's global typology of HIV/AIDS based upon the classification of national epidemiological profiles into one of three patterns (Piot et al. 1988). Pattern I included the countries of North America, Western Europe, and Australasia, where transmission was predominantly by homosexual men and IVDUs, and the prevalence was of median rank. Pattern II referred to most of Africa and parts of Latin America where the transmission was mainly heterosexual and the prevalence was relatively high. Last, Pattern III described most of Asia where the transmission modes were mixed and the prevalence was virtually negligible. There are dangers, however, in drawing inferences from a taxonomy based upon a single geographical snapshot taken in 1988 of an evolving and dynamic epidemic. In hindsight, it is now known that the false message of some kind of Asian immunity to HIV, which this typology appeared to convey, was due simply to the late arrival of the infection. Moreover, since first

infection, data for India has indicated an AIDS incidence curve reminiscent of the early phase in Africa and the major burden of the epidemic in the next century is expected to be in Southern Asia.

While national HIV/AIDS statistics provide essential information for the formulation of health policy, this scale of data collection obscures important local features of the transmission process. Studies of individual records often reveal highly clustered patterns of HIV/AIDS incidence, especially during the early stages of the epidemic. In the USA, infections among homosexual men were concentrated in tightly defined residential areas like Greenwich Village in New York and the Castro District in San Francisco. In comparison, the clustering exhibited by IVDUs has often been even more pronounced. Many of the heroin addicts infected early in Dublin, for example, were found to be resident in the same complex of inner city apartments (Smyth and Thomas 1996a). Moreover, a recent study of the essentially heterosexual epidemic in Rakai District, Uganda, has revealed a rural pattern where certain villages experienced in excess of 30 percent prevalence among their population, while many of their neighbors remained relatively free from infection (Low-Beer et al. 1997). This last outcome indicates the infection risk varies significantly between communities with the same behavior, in addition to the geographical variations observed between nations.

3. *Disease Prevention*

3.1 *Natural Control*

One facet of preventing the transmission of HIV, then, is to understand why only some communities seem prone to infection. In this respect, the notion of natural control describes the essentially passive protection that is conferred on communities when their collective epidemic activity is too infrequent to support the sustained transmission of HIV. This natural state may be given meaning by a statistic known as the reproduction number, which has a long history of application to understanding the control of other infectious diseases like malaria and influenza.

The basic reproduction number for HIV is formed from the values of three epidemiological parameters, each summarizing an average property of the transmission process observed in a particular community (May et al. 1989). The first is the transmission probability denoted by β. This term measures the likelihood that a single susceptible partner of an infected individual will contract HIV in a given unit of time. For homosexual men in San Francisco in the early 1980s, for example, this probability (β) has been estimated to be 0.1 per partnership per year (HIV is quite difficult to transmit). The second is the average rate of partner acquisition per unit of time, r, which for the same sample was approximately 8 partners per year. The last is the period of communicability (D) for HIV, which is thought to be about 2 years (this period is shorter than the chronic phase because the second part of the latter includes an episode when antigen levels in the host are too low to be communicable). Then, the characteristic reproduction number, R, is given by the equation $R = \beta r D$, which counts the expected number of secondary infections attributable to an initial HIV infected during the period of communicability. Moreover, a value of $R = 1$ serves as a starting threshold for an epidemic to begin. The first homosexual man with HIV in San Francisco, for example, is estimated to have acquired about $rD = 8 \times 2 = 16$ partners while communicable, of whom $R = 0.1 \times 16 = 1.6$ were expected to contract infection. Thus, this index case was more than replaced while infectious, which indicates how the virus was subsequently able to circulate in this particular homosexual community.

Alternatively, if R is less than one, then the initial infection is not reproduced and the epidemic will be expected to die out rapidly, thereby maintaining the state of natural control. It may be noticed that a partnership rate of $r = 5$ is sufficient to make $R = 0.1 \times 5 \times 2 = 1$ and, therefore, was the critical rate for the San Francisco epidemic to begin. Moreover, surveys of heterosexual activity in both the USA and UK have repeatedly reported partner acquisition rates well below this critical value, to indicate that these populations are subject to a high degree of natural protection. Such findings, however, should be treated carefully because the parameter values are subject to known geographical variations. One stark contrast is provided by an estimated transmission probability of $\beta = 0.4$ drawn from a sample of Central African heterosexuals. This easier exchange of HIV is thought to be linked to the high prevalence of genital cuts and ulcers among this sample consequent upon their prior infection with other sexually transmitted diseases (Bassett and Mhloyi 1991). This raised probability defines a much more conservative critical partnership rate of $r = 1.25$ ($R = 0.4 \times 1.25 \times 2 = 1$), which suggests a substantial proportion of the Central African heterosexual population might be subject to the prospect of continuous transmission. Moreover, the large number of people implied to be placed at risk by this interrelationship provides a plausible explanation for the high prevalence of HIV in this region.

The reproduction number, R, refers to a single risk population and, therefore, does not take account of the interactions that are known to occur both between different behaviors and geographical regions. To counter this simplification, regional reproduction numbers have been derived to count the number of secondary infections made in every locality and risk cohort that are attributable to an index case resident in a particular region (Thomas 1999). These regional numbers have been shown to possess more complex starting thresholds where a value greater than one

does not necessarily guarantee the infection will diffuse around the system of regions. For such spread to occur, a regional number must be sufficiently in excess of one to compensate for other regions and cohorts where the reproduction potential is below unity. Moreover, in conditions where spread is expected to occur, the epidemic engendered among those with low potential is often fragile. An analysis of the interchanges between those with high and low risk behaviors in the UK, for example, found the small incidence among the latter to be dominated by crossinfections rather than by contacts made between themselves (Thomas 1996). This outcome conforms with the observed incidence of HIV in most developed countries where occurrences of direct transmission between low risk heterosexual partners have been rare.

This interpretation of the epidemic in terms of reproduction numbers represents a switch in the way the infection risk is construed. Protection is now related to the frequency a particular risk activity is undertaken and not directly to behaviors like sexuality or addiction. This distinction recognizes that, although these behaviors may exhibit high incidence in certain localities, many of the individuals so categorized may not necessarily engage in frequent risk activity. The focus on activity rates, therefore, attributes the presence of high risk to specific core cohorts of individuals who make sufficiently frequent encounters to sustain a reproduction number in excess of the starting threshold. These core cohorts, therefore, are disproportionately prone to pass infection to the remainder of the population, which implies directing interventions at these networks of active individuals will be an effective strategy for reducing HIV prevalence around the regional system.

3.2 Direct Actions

In the absence of a viable vaccine, such direct action has either entailed implementing medical interventions linked to a positive blood test or social measures intended to modify high-risk behaviors. Blood testing is usually the technical responsibility of official health agencies and is intended to make those already circulating with HIV aware of their infectious status. A further option after a positive outcome is to trace and test the partners of the infectious individuals repeatedly in an effort to reconstruct the transmission pathway. This procedure of contact tracing is intended to identify all positive individuals on the local partnership network in an effort to sever this chain of infection. A more stringent response is quarantining which removes those with HIV from active circulation and so curtails their opportunities to infect susceptible individuals. In contrast, social interventions, which promote safer practices, are normally delivered voluntarily by community based organizations in an effort to reach those with a particular risk behavior. These initiatives often involve a high degree of local participation and promote changed behavior patterns through awareness and personalization of the infection risks.

The relative merits of medical and social interventions, however, have been fiercely contested, especially the use that has been made of the serological test (Smyth and Thomas 1996b). Medical opinion has often justified blood screening on the grounds that it is unethical to deny infecteds and their partners the opportunity to take precautions to prevent further passage of HIV (Knox et al. 1993). In contrast, those with high-risk behaviors have been quick to counter that these public interventions represent an extreme invasion of privacy (Krieger and Margo 1994). The strength of this resistance is indicated by the fact that Sweden is alone among the developed nations in adopting a mandatory requirement for reporting sexual partners. The established role of contact tracing in the control of other venereal diseases has been further weakened by the problem of partner recall associated with the long incubation period of HIV (Kirp and Bayer 1992). Quarantining has attracted even less support, especially after the realization that the incubation period also implies lengthy and unnecessary episodes of incarceration.

Some of the safer practices promoted by social initiatives have proved to be equally controversial. Recommendations to the US Surgeon General, for example, to provide federal funds to support the provision of needle exchange programs for IVDUs met with skeptical responses from many of the interested parties (Normand et al 1995). Some black communities were targets for these programs, yet many African-American churches teach the immorality of drug use and regard needle exchange as a facilitative venture. The response of law enforcement agencies was similarly negative, but grounded on the ambivalence of their lending support to illegal activities. Pharmacists, while often agreeing with the principle of needle exchange programs, were also concerned about the impact of a needle exchange facility on the quality of service to customers other than IVDUs. This collective mistrust, however, contrasts with the favored empirical evidence that stresses the complex etiology of individual drug abuse is unlikely to be significantly affected by a single risk factor such as the availability of sterile needles.

Nevertheless, community based initiatives and public education campaigns have gradually come to be the most frequently adopted interventions against HIV. The success of these programs in reducing AIDS incidence, however, is often difficult to gauge. The earliest and most influential prevention campaign was initiated by the homosexual community in San Francisco, and involved the establishment of both educational and legislative frameworks to support the practice of safer sex. The success of this effort was documented by a number of epidemiological studies

that recorded a decline in the rate of HIV transmission in San Francisco during the mid-1980s. Observations made on this community also found sexual activity ranging from celibacy to frequent promiscuity. In this respect, subsequent analyses of populations with varied rates of partner acquisition have shown the presence of such core cohort activity raises the prevalence of HIV early in the epidemic, but has the reverse effect later on when these individuals are the first to be removed from circulation with AIDS (Anderson and May 1991). Consequently, the early onset of AIDS among those with the most sexual partners most probably deflated the posited beneficial impacts on incidence of this campaign.

The time of implementation during the epidemic cycle of HIV infection is also crucial to the success of an intervention. In principle, the closer this time is to the date of the initial infection, the smaller is the expected cumulative AIDS incidence in the targeted community. Moreover, after the time of peak HIV prevalence, interventions are expected to become increasingly ineffective as the epidemic moves into a period of natural decline. In practice, implementation has often occurred soon after the first diagnosis of AIDS when this event raises community awareness and consciousness. The significance of these timing effects is illustrated by a comparison of the epidemics among homosexual men and IVDUs in Dublin, where the latter have been observed to progress more quickly than the former. (Smyth and Thomas 1996b). Moreover, interventions in both these communities were delayed for a number of years after the initial AIDS diagnoses by religious and political pressure. Given this background, the lower rate of transmission among homosexual men is estimated to allow 20 years of effective campaigning to the time of peak prevalence whereas, for IVDUs, this episode is likely to be just five years.

4. *Prospect*

The HIV/AIDS epidemic has been aptly named 'the slow plague' (Gould 1993). It is perhaps not surprising, therefore, that evidence for an expected downturn in the characteristic infectious disease cycle did not appear until the early 1990s. Then, recorded AIDS incidence in many developed countries, and HIV prevalence in some, began to exhibit a state of gradual decline. The outcome for AIDS is thought to be temporary and has been attributed to advances in antiretroviral drug combination therapy, which significantly delays the onset of the opportunistic infections. The risks of HIV transmission, however, are not similarly affected and the modest reductions presently observed might well signal a genuine transition in the epidemic process. Consequently, the next century might witness a gradual lessening of the devastating toll on human life taken by this tardy pandemic infection. Irrespective of this outcome, the geographical experience of AIDS to date suggests the idiosyncratic local passage of HIV will continue to pose fresh and awkward challenges for the task of disease control.

See also: AIDS (Acquired Immune-deficiency Syndrome)

Bibliography

Anderson R M, May R M 1991 *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, UK

Bassett M T, Mhloyi M 1991 Women and AIDS in Zimbabwe: the making of an epidemic. *International Journal of Health Services* **21**: 143–56

Freedman D 1987 *AIDS: The Problem in Ireland*. Townhouse, Dublin, Republic of Ireland

Gotlieb M S, Schroff R, Schanker H M, Weisman J D, Fan P T, Wolf R A, Saxon A 1981 Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *New England Journal of Medicine* **305**: 1425–31

Gould P 1993 *The Slow Plague: A Geography of the AIDS Pandemic*. Blackwell, Oxford, UK

Kirp D L, Bayer R 1992 *AIDS in the Industrialized Democracies: Passions, Politics and Policies*. Rutgers University Press, New Brunswick, NJ

Knox E G, MacArthur C, Simons K J 1993 *Sexual Behaviour and AIDS in Great Britain*. HMSO, London

Krieger N, Margo G 1994 *AIDS: The Politics of Survival*. Baywood, New York

Li W H, Tanimura M, Sharp P M 1988 Rates and dates of divergence between AIDS virus nucleotide sequences. *Molecular Biology and Evolution* **5**(4): 313–30

Low-Beer D, Stoneburner R L, Mukulu A 1997 Empirical evidence of the severe but localised impact of AIDS on population structure. *Nature Medicine* **3**: 553–7

May R M, Anderson R M, Blower S M 1989 The epidemiology and transmission dynamics of HIV/AIDS. *Daedalus* **118**: 163–201

Normand J, Vlahov D, Moses L E 1995 *Preventing HIV Transmission: The Role of Sterile Needles and Bleach*. National Academy Press, Washington, DC

Piot P, Plummer F A, Mhalu F S, Lamboray J-L, Chin J, Mann J M 1988 AIDS: an international perspective. *Science* **239**: 573–9

Root-Bernstein R S 1989 AIDS and KS pre-1979. *Lancet* **335**: 969

Shannon G W, Pyle G F, Bashkur R L 1991 *The Geography of AIDS: Origins and Course of an Epidemic*. Guilford Press, New York

Smallman-Raynor M R, Cliff A D, Haggett P 1992 *Atlas of AIDS*. Blackwell, Oxford, UK

Smyth F M, Thomas R W 1996a Controlling HIV/AIDS in Ireland: the implications for health policy of some epidemic forecasts. *Environment and Planning A* **28**: 99–118

Smyth F M, Thomas R W 1996b Preventative action and the diffusion of HIV/AIDS. *Progress in Human Geography* **20**: 1–22

Thomas R 1996 Modelling space-time HIV/AIDS dynamics: applications to disease control. *Social Science and Medicine* **43**: 353–66

Thomas R 1999 Reproduction rates in multiregion modelling systems for HIV/AIDS. *Journal of Regional Science* **39**: 359–85

WHO 1998 *AIDS Epidemic Update: December* 1998. Joint United Nations Programme on HIV/AIDS, Geneva, Switzerland

R. W. Thomas

Air Pollution

Both natural processes and human activities contribute to air pollution, with the combustion of fossil fuels being the largest anthropogenic source of air pollutants. Adverse health effects, damage to biota and materials, reduced visibility, and changed radiation balance of the atmosphere are the major consequences of high concentrations of air pollutants.

1. *Air Pollution*

Air pollution is a matter of excessive concentrations rather than a mere atmospheric presence of particular airborne elements or compounds. Air pollutants most commonly released by human activities—solid particles (dust, soot), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NO_x), and many hydrocarbons (ranging from methane to complex polycyclic molecules)—are normally present in unpolluted air in trace amounts. They are emitted by a variety of natural processes: volcanic eruptions, forest and grassland fires, soil erosion and desert storms are major sources of airborne solid particulates; wildfires release also CO and NO_x; bacteria-driven biogeochemical cycles of C, N, and S are the sources of methane, various S and N gases, and temperate and tropical forests emit large amounts of hydrocarbons.

Although air pollution is so strongly associated with modern, industrial civilization, it is actually a phenomenon with a very long history. Combustion of biomass fuels and, later (about two millennia ago in China, during the Middle Ages in Europe) of coal, and traditional color metallurgy and smelting of iron ore produced excessive concentrations of solid and gaseous pollutants. But given the relatively limited extent of these activities, as well as the fact that the pollutants were released practically at the ground level and hence they could not disperse over long distances, environmental impacts, although locally severe, were spatially quite restricted.

In contrast, the largest modern industrial sources of air pollution (power plants, iron and steel mills, smelters, refineries, chemical syntheses) often release enormous volumes of hot (more than 100 °C) mixtures of particulates and gases from tall stacks at considerable height (more than 100 m) above the ground. These emissions can rise into the mid-troposphere (about 5 km above the sea level) and can be carried hundreds of kilometers downwind before they are removed by dry deposition or precipitation. Smaller stationary (household, institutional, and manufacturing) and mobile (motor vehicles, airplanes, ships) sources of air pollutants emit particulates and gases over large urban and industrial areas and along heavily traveled routes. Combination of these large point sources, and of extensive areal pollution creates major regional, even semicontinental, environmental problems.

In traditional societies it is indoor air pollution—arising from low-efficiency combustion of solid fuels (wood, grasses, crop residues, above all cereal straws, dried dung, and coal) in unventilated, or poorly ventilated, rooms—that generally poses much greater health risks than the outdoor contamination of air (Smith 1993). Since the onset of the nineteenth century industrialization, and particularly during the latter half of the twentieth century, affluent countries have concentrated on the controls of outdoor air pollution—but more recent research has shown that even in many modern settings indoor pollutants may pose cumulatively higher risks to human health than the contaminated ambient air (Turiel 1985).

Levels of ambient air pollution are not only determined by the magnitude of emissions: atmospheric behavior and terrain combine to play a critical role. Thermal inversions reverse the normal atmospheric stratification when the warmest air is near the ground. They are produced either by intensive nocturnal cooling of the ground (more vigorously during winter months), or by the sinking of air in anticyclones (high-pressure cells) during summer. In either case, warmer air found above a cooler stratum near the ground limits the depth of the atmospheric mixing and concentrations of pollutants emitted into this restricted volume of the mixed layer can reach very high levels in a matter of days.

The inversion effect is further aggravated in places where mountain ranges or river valleys restrict horizontal air movements: Los Angeles, Vancouver, Chongqing, and Taipei, among many other places, exemplify this situation. In contrast, places where thermal inversions are less frequent, and where a relatively flat terrain allows for generally good ventilation, have much lower concentrations of air pollutants in spite of often large total emissions: New York and Boston are perhaps the two best American examples.

2. *Common Air Pollutants*

There are literally thousands of compounds whose atmospheric concentrations are now detectable by modern analytical methods, but most of them are

either present only in trace quantities (mere parts per billion or parts per trillion) or their distribution is spatially very limited (the latter case includes many occupational exposures). Masswise, large particulates and CO are the most abundant air pollutants in global terms, but the finest particulates, oxides of sulfur and nitrogen and volatile hydrocarbons, are responsible for the greatest share of undesirable impacts air pollution has on biota, human health, materials, and on the atmosphere itself (Wark et al. 1998, Heinsohn 1999).

2.1 Particulates

Particulates, or aerosols, refer to any matter—solid or liquid—with a diameter less than 500 micrometers (μm). Besides the acronyms PM (particulate matter) and SPM (suspended particulate matter), air pollution literature also uses TSP for total suspended particulates. Large, visible particulates—fly ash, metallic particles, dust, and soot (carbon particles impregnated with tar)—settle fairly rapidly close to their source of origin and are rarely inhaled. Very small particulates (diameters below 10 μm) can stay aloft for weeks and hence can be carried far downwind, and even between the continents.

Volcanic ash injected into the stratosphere, particularly from eruptions in the tropics, can actually circumnavigate the Earth. Particulates from Saharan dust storms are repeatedly deposited over the Caribbean, and detected in Scandinavia. Only 7–10 days after Iraqi troops set fire to Kuwaiti oil wells in late February 1991, soot particles from these sources were identified in Hawaii, and in subsequent months solar radiation received at the ground was reduced over an area extending from Libya to Pakistan, and from Yemen to Kazakhstan.

The US Environmental Protection Agency (EPA) estimated that in 1940 the country's combustion and industrial processes released almost 15 million tonnes (Mt) of particulates smaller than 10 μm, compared with only about 3 Mt during the late 1990s; however, field tilling, construction, mining, quarrying, road traffic, and wind erosion put aloft at least another 30–40 Mt of such particulates a year. Naturally, global estimates of particulate emissions from these sources are highly unreliable.

Particulates are sampled either by total mass (TSP) or by their size which is determined by their aerodynamic diameter: particulates smaller than 10 μm can be readily inhaled even by nose, and the smallest aerosols—those 2.5 μm and smaller—can reach alveoli, the lung's finest structures. National and international limits for particulate concentrations are the only cases of ambient air quality standards that are not chemically specific. While this makes no difference for many particulates that are inert, there is no shortage of highly toxic elements and compounds, including arsenic (emitted during the smelting of color metals and combustion of some coals), asbestos (dust from mines, brake linings, and insulation), lead (mainly from gasoline combustion), and benzo-a-pyrene (a highly carcinogenic hydrocarbon released from fuel combustion).

2.2 Carbon Monoxide

Colorless and odorless carbon monoxide is the product of incomplete combustion of carbon fuels: cars, other small mobile or stationary internal combustion engines (installed in boats, snowmobiles, lawn mowers, chain saws), and open fires (burning of garbage and crop residues after harvest) are its leading sources. Foundries, refineries, pulp mills, and smoldering fires in exposed coal seams are other major contributors. Emission controls (using catalytic converters) that began on all US vehicles in 1970 have been able to negate the effects of a rapid expansion of car ownership and of higher average use of vehicles: EPA estimates that by the late 1990s US CO emissions fell by about 25 percent compared to their peak reached in 1970.

2.3 Sulfur Dioxide

SO_2 is a colorless gas that cannot be smelled at low concentrations; at higher levels it has an unmistakably pungent and irritating odor. Oxidation of sulfur present in fossil fuels (typically 1–2 percent by mass in coals and in crude oils), and in sulfides of metals (copper, zinc, nickel) are its main sources. Petroleum refining and chemical syntheses are the other two major emitters of the gas besides fossil-fueled electricity generation and color metallurgy.

US emissions of the gas peaked in the early 1970s at nearly 30 Mt per year, and have been reduced to less than 20 Mt by the mid-1990s. Global emissions of SO_2 rose from about 20 Mt at the beginning of the twentieth century to more than 100 Mt by the late 1970s; subsequent controls in Western Europe and North America and collapse of the Communist economies cut the global flux by nearly a third—but Asian emissions (China is now the world's largest user of coal) have continued to rise (McDonald 1999).

2.4 Nitrogen Oxides and Hydrocarbons

Nitrogen oxides (NO, and to a lesser extent NO_2) are released during any high-temperature combustion which breaks the strongly bonded atmospheric N_2 and combines atomic N with oxygen: power plants are their largest stationary sources, vehicles and airplanes the most ubiquitous mobile emitters. Anthropogenic hydrocarbon emissions result from incomplete combustion of fuels, as well as from evaporation of fuels and solvents, incineration of wastes, and wear on car

tires. Processing, distribution, marketing, and combustion of petroleum products is by far the largest source of hydrocarbons in all densely populated regions. In spite of aggressive control efforts, total US emissions of NO_x and hydrocarbons have remained at roughly the same level (at just above 20 Mt per year each) since the early 1980s.

In the presence of sunlight, nitrogen oxides, hydrocarbons, and carbon monoxide take part in complex chains of chemical reactions producing photochemical smog: its major product, the tropospheric ozone, is an aggressive oxidant causing extensive damage to human and animal health as well as to forests and crops (for details see *Tropospheric Ozone: Agricultural Implications*). Ozone is also the pollutant whose generation may be most difficult to control in the coming world of megacities and intensified transportation.

Still, on the global scale a very large share of undesirable environmental and health impacts attributable to air pollution arises from classical smog and from acid deposition. Adverse health effects of classical smog—created by emissions of particulates and SO_2 from coal combustion—have been known for generations, and recent evidence suggests that fine particulates alone pose a considerable risk. Acid deposition arises from atmospheric oxidation of sulfur and nitrogen oxides: the resulting generation of sulfate and nitrate anions and hydrogen cations produces precipitation whose acidity is far below the normal pH of rain (about 5.6) acidified only by carbonic acid derived from the trace amount of CO_2 (about 360 ppm) constantly present in the atmosphere.

Only the richest economies now have fairly adequate air pollution monitoring networks whose regular measurements allow us to make reasonable judgments about the air quality and its long-term trends. Elsewhere, including the megacities of China and India, the monitoring is at best highly patchy and of questionable quality (Earthwatch 1992). Naturally, the lack of adequate knowledge of typical exposures (which must go beyond simple means or short-term maxima of major pollutants) complicates the assessment of air pollution effects on human health.

3. *Health Effects*

The recurrence of extremely high concentrations of air pollutants experienced in industrial cities of Europe and North America before the mid-1960s left no doubt about the acute harmful effects of such exposures. During the most tragic of these high air pollution episodes, in London in December 1952, about 4,000 people died prematurely within a week (Brimblecombe 1987). Similarly high levels of particulates and SO_2 are now encountered only briefly in the most polluted cities in China. Uncovering the impacts of chronic exposures to much lower levels of air pollutants has thus become a challenge for sophisticated epidemiological analyses which must eliminate, or at least minimize, the effects of numerous intervening variables ranging from socioeconomic status (a strong predictor of both morbidity and premature mortality) and diet to smoking and exposures to indoor air pollutants.

High levels of SO_2 irritate the upper respiratory tract (causing coughing and mucous secretions), and the gas adsorbed on fine particles or converted to sulfuric acid can damage lungs. Not surprisingly, chronic exposure to classical smog has been correlated with increased respiratory and cardiovascular morbidity and mortality. Those at particular risk include the elderly and small children in general, and people who are already suffering from respiratory diseases (asthma, bronchitis, emphysema) and from cardiovascular ailments. The presence of hydrocarbons in this smog has also been linked to a higher incidence of lung cancer mortality.

Epidemiological evidence assembled during the late 1980s and the early 1990s indicates that increases in human mortality and morbidity have been associated with particulate levels significantly below those previously considered harmful to human health (Dockery and Pope 1994). This effect has been attributed to particles smaller than 2.5 μm which are released mainly by motor vehicles, industrial processes, and wood stoves. For this reason the EPA introduced in 1997 new regulations to reduce concentrations of such particles. Once implemented, this new rule might prevent as many as 20,000 premature deaths a year and reduce asthma cases by 250,000—but these claims have been highly controversial, and appropriate control measures are to be phased in gradually.

Even if controls of the finest particulates are costly, studies show that—with the exception of lead, a cumulative poison which causes mental retardation in children and impairs the nervous system in adults—even greater investments are needed to lower morbidity or prevent premature mortality from exposure to many toxic air pollutants. The most dangerous organic toxins commonly encountered in polluted air are benzene (a common intermediary in chemical synthesis, and a product of burning some organic wastes), dioxin (a highly potent carcinogen released most often from solid waste incinerators), and polychlorinated biphenyls (PCBs, whose production was banned in 1977 but which continue to be volatilized from spills, landfills, and road oils).

In many low-income countries, the combined effect of indoor air pollution and smoking is almost certainly more important than the exposures to ambient air pollution. For example, in China the rural mortality due to chronic obstructive pulmonary diseases is almost twice as high in rural areas than in cities: rural ambient air is cleaner but villagers using improperly vented stoves are exposed to much higher levels of indoor air pollution (Smil 1996). The effect on children younger than five years is particularly severe: in poor

countries 2–4 million of them die every year of acute respiratory infections which are greatly aggravated by indoor pollutants.

In affluent countries indoor air pollution includes not only assorted particulates from stoves, fireplaces, carpets, and fabrics, but also volatile organic compounds from numerous household cleaners, glues, and resins, as well as from molds and feces of dust mites. High levels of radon, linked to a higher incidence of lung cancer, are common in millions of houses located on substrates containing relatively high concentrations of radium whose radioactive decay releases the gas into buildings.

4. *Other Environmental Impacts*

Reduction of visibility due to light scattering and absorption by aerosols is a ubiquitous sign of high concentration of air pollutants. High levels of aerosols can also change regional or even continental radiation balance (Hobbs 1993). Volcanic ash can be responsible for appreciable reduction of ground temperatures on a hemispheric scale and the effect can persist for months following the eruption. Sulfates in the air above Eastern North America, large parts of Europe and East Asia have been cooling the troposphere over these large regions, counteracting the effect of global warming.

These three large regions are also most affected by acid deposition: its most worrisome consequences have been the loss of biodiversity in acidified lakes and streams (including complete disappearance of the most sensitive fish and amphibian species); changes in soil chemistry (leaching of alkaline elements and mobilization of aluminum and heavy metals); and acute and chronic effects on the growth of forests, particularly conifers (Irving 1991, Godbold and Hutterman 1994). Chronic exposure to acid precipitation also increases the rates of metal corrosion, destroys paints and plastics, and wears away stone surfaces.

5. *Controlling Air Pollution*

Serious national efforts to limit air pollution date only to the 1950s (UK's Clean Air Act of 1956) and the 1960s (US Clean Air Act of 1963 and Air Quality Act of 1967). Fuel substitutions, higher combustion efficiencies, and capture of generated pollutants have been the principal strategies of effective air pollution control. Replacement of high-sulfur solid fuels by low-sulfur coals and fuel oils, and even better by natural gas, has usually been the least costly choice. These substitutions began improving the air quality in large North American cities during the 1950s and in European cities a decade later; the large-scale use of natural gas from The Netherlands, the North Sea, and Siberia has had the greatest impact.

Higher combustion efficiencies reduce the need for fuel: while traditional coal stoves were often no more than 10–15 percent efficient, modern coals stoves are commonly 40–45 percent efficient, and the best household natural gas furnaces are now rated at 96 percent efficient. Less dramatic gains resulted from the replacement of inefficient steam locomotives (less than 10 percent efficient) by diesel (more than 30 percent efficient) and electric traction. The latest gas turbines, powering commercial jet airplanes are used in stationary applications, and the internal combustion engines in cars are also more efficient.

Particulate emissions can be effectively controlled by a variety of cyclones, fabric filters, and electrostatic precipitators which can be more than 99.5 percent efficient. Lead-free gasoline is now the norm in affluent nations, but leaded fuel is still used in many low-income countries. SO_2 emissions can be reduced by desulfurization of liquid fuels and natural gases, but only to a limited extent by cleaning of coal. Flue gas desulfurization (FGD) is a costly but effective way to remove the generated SO_2; most common commercial processes use reactions with ground limestone or lime to convert the gas into calcium sulfate which must be then landfilled. Although FGD increases the capital cost of a large coal-fired power plant by at least 25 percent (operation costs are also higher), more than half of all US coal-fired power plants now desulfurize their flue gases.

Automotive air pollution controls have been achieved by a combination of redesigned internal combustion engines and by mandatory installation of three-way catalytic converters removing very large shares of CO, NO_x, and hydrocarbons. As a result, by the mid-1990s average US emissions of the three pollutants were reduced by 90–96 percent compared with the early 1970s. Continuing urbanization, including the formation of megacities with more than 20 million people, and spreading car ownership mean that new solutions will have to be adopted during the twenty-first century (Mage et al. 1996).

See also: Environmental Challenges in Organizations; Environmental Health and Safety: Social Aspects; Environmental Risk and Hazards; Transportation: Supply and Congestion

Bibliography

Brimblecombe P 1987 *The Big Smoke*. Methuen, London

Dockery D W, Pope C A 1994 Acute respiratory effects of particulate air pollution. *Annual Review of Public Health* **15**: 107–32

Earthwatch 1992 *Urban Air Pollution in Megacities of the World*. WHO and United Nations Environment Programme, Oxford, UK

Gammage R B, Berven B A (eds.) 1996 *Indoor Air and Human Health*. CRC Press, Boca Raton, FL

Godbold D L, Hutterman A 1994 *Effects of Acid Precipitation on Forest Processes*. Wiley-Liss, New York

Heinsohn R J, Kable R L 1999 *Sources and Control of Air Pollution*. Prentice-Hall, Upper Saddle River, NJ
Hobbs P V 1993 *Aerosol–Cloud–Climate Interactions*. Academic Press, San Diego, CA
Irving P M (ed.) 1991 *Acidic Deposition: State of Science and Technology*. US National Acid Precipitation Assessment Program, Washington, DC
Mage D, Ozolins G, Peterson P, Webster A, Orthofer R, Vanderweed V, Gwynne M 1996 Urban air pollution in megacities of the world. *Atmospheric Environment* **30**: 681–6
McDonald A 1999 Combating acid deposition and climate change: Priorities for Asia. *Environment* **41**: 4–11, 34–41
Smil V 1996 *Environmental Problems in China: Estimates of Economic Costs*. East-West Center, Honolulu, HI
Smith K R 1993 Fuel combustion, air pollution exposure and health: the situation in developing countries. *Annual Review of Energy* **18**: 529–66
Turiel I 1985 *Indoor Air Quality and Human Health*. Stanford University Press, Stanford, CA
Wark K, Warner C F, Davis W T 1998 *Air Pollution: Its Origin and Control*. Addison-Wesley, Menlo Park, CA

V. Smil

Alcohol-related Disorders

Alcohol (ethanol, C_2H_5OH) is a relative simple molecule which interacts with numerous transmitters and receptors in the body and brain and also changes structure and function of cells and cell membranes, among others. Virtually every organ is affected by acute or chronic alcohol intake. It is difficult to define definite cut-offs for risky alcohol consumption. The British Medical Association in 1995 considered 20 g alcohol for women and 30 g for men as the upper limit for non-risky alcohol use. Acute effects of alcohol, for example on blood pressure, circulation, or brain function must be differentiated from more chronic ones like liver dysfunction or withdrawal. Since alcohol's effects in the body are so complex only a brief overview on its basic mechanisms are given before addressing clinically relevant disorders associated with alcohol consumption.

1. *Alcohol—Metabolism and Pharmacology*

Alcohol is quite rapidly absorbed after oral ingestion in the stomach. About 95 percent of alcohol is oxidized in the liver by the enzyme alcohol dehydrogenase (ADH) to acetaldehyde, which in return is rapidly metabolized by the enzyme acetaldehydedehydrogenase (ALDH) to acetic acid which is also rapidly converted to carbon dioxide and water. Only 5 percent of alcohol is excreted unchanged in the urine, sweat, and breath. There is a genetic polymorphism for both enzymes with different isoenzymes. While most (>90 percent) of the Caucasian population have 'regular' ALDH isoenzymes, other—especially Asian populations have so-called ALDH-deficient isoenzymes (30 to 50 percent) with significant acetaldehyde levels in the blood after alcohol intake. In these individuals alcohol consumption rapidly results in aversive reactions, so-called 'flush-reaction.' Alcohol is usually metabolized at a rate of 0.1–0.15 (or 0.2) mg/liter per hour.

2. *Genetics*

There is substantial evidence from a number of family, twin, and adoption studies for a genetic transmission of alcoholism. The risk for alcoholism is increased in first-degree relatives of alcoholics. Some adoption studies have shown an up to four times increased risk for alcoholism for sons of alcoholics even if they were raised apart from their biological parents. Although the heretability of alcoholism is a topic of numerous biological and genetic studies on the genetic and molecular/biological level no vulnerability marker or gene for alcoholism is definitely identified yet. It seems most likely that alcoholism is not transmitted by a single but a number of genes. Alcoholism seems to be a polygenic disorder. Also no definite biochemical marker for alcoholism is found yet. An ambitious research project focusing on this issue is the multi centered Collaborative Study on Genetics of Alcoholism. This US study group has examined extended multigenerational families affected by alcoholism and studies the heretability by genetic linkage analysis. Genome-wide scans to identify genes mediating the risk for alcoholism have been initiated. To date the group has reported that genes affecting vulnerability for alcoholism could be located on chromosomes one and seven. There is additional modest evidence for a protective gene on chromosome seven. The latter finding has also been reported by a study in Indian Americans. This study also gave evidence for a susceptibility gene on chromosome 11. It seems of interest that alcohol dehydrogenase genes (ADH2 and 3) are located near the protective chromosome four locus. Further analysis will attempt to identify single genes mediating the risk for alcoholism (see *Mental Illness*, *Genetics of* and Zernig et al. 2000).

There are marked differences not only in alcohol metabolism but also tolerance. Experimental and follow-up studies have shown that high-risk individuals (children of alcoholic parents) usually tolerate alcohol much better than other individuals. This in part explains the increased risk for alcoholism.

3. *General Effects of Alcohol*

3.1 *Brain* (*CNS Effects*)

Different from other psychoactive substances like opioids there is no special alcohol receptor in the

brain. A number of neurotransmitters are involved in mediating alcohol's effects including GABA, glutamate, dopamine, opioids, serotonin, and noradrenalin, among others. Alcohol is a psychotropic agent that depresses the central nervous system (CNS) basically via enhancement of GABAergic neurotransmission. GABA is the most important inhibitory neurotransmitter in the brain. Acute alcohol intoxication results in enhancement of inhibitory neurotransmitters (GABA) and antagonization of excitatory neurotransmitters (glutamate, dopamine, etc.) while the neurotransmitter function in alcohol withdrawal is the opposite (increased activity and release of excitatory, inhibition of inhibitory neurotransmitters). Thus alcohol withdrawal results in an increased excitatory state in the brain, possibly leading to seizures or delirium.

The rewarding, psychotropic effects of alcohol are in part mediated by dopamine, opioids, GABA, glutamate, and serotonin. There seems to be a special addiction memory in the brain which in part involves brain structures which are of relevance for physiologic reward processes and controlling of food and fluid intake and sexuality. One of the key structures in the brain mediating these reward effects is the dopaminergil mesolimbic system including the nucleus accumbens. Activation of this system leads to positive reinforcement. Alcohol but also other psychotropic drugs are believed to act predominantly by interactions with neurons in these brain areas.

Alcohol also directly acts on neurons in the CNS. It alters the properties of lipids in the membrane of neurons but also has direct neurotoxic effects at least in higher concentrations. Chronic alcohol intake may result in cell damage and destruction of neurons in the brain but also in other regions of the body. Other alcohol-related factors such as vitamin deficiencies or malnutrition in general may contribute to the neurotoxic effects. Although alcohol-related cell loss can be found in all brain areas, the forebrain and the cerebellum are mostly effected in chronic alcoholics. To some extent cell losses in the CNS can be visualized *in vivo* by modern neuroradiological techniques such as cranial computertomography scans or NMR.

3.2 Effects on Cognitive Function and Mental Processes

Modest alcohol intake may cause a number of emotional changes such as sadness, anxiety, or irritability that predominantly occur at peak or with decreasing blood alcohol concentration (BAC). Alcohol in higher doses can cause psychiatric syndromes: (intense) sadness and anxiety, auditory hallucinations and/or paranoia without clouding of sensorium. These syndromes can be classified as organic brain syndromes or alcohol psychosis. The former is characterized by mental confusion and clouding of sensorium which can be found during alcohol intoxication usually at a BAC over 1.5 mg/liter, withdrawal, or as a consequence of alcohol-related disorders.

3.3 Behavioral Changes

They depend on age, weight, sex, prior experience with alcohol (e.g., the individual's drinking history), and age. Symptoms of alcohol intoxication are described below.

3.4 Tolerance

There are marked differences in alcohol tolerance between individuals, partially due to genetic variances in alcohol metabolism. For a number of not fully understood reasons tolerance in men is usually better than in women. Women have less water in their body so alcohol is less diluted and has greater effects in the tissue. The individual alcohol history (heavy or regular versus sporadic consumption), liver function, organic brain syndromes or other disorders have marked impact on alcohol tolerance which is usually increased in heavy drinkers and alcohol dependants, except for those late-stage drinkers with severe physical (liver!) or mental impairment. Some studies in high-risk individuals (offspring of alcoholic families) have shown that alcohol tolerance is usually better in individuals with positive family history for alcoholism and also to some extent predictive for later alcoholism.

3.5 Physical Dependence

3.5.1 Alcohol dependence. According to modern psychiatric classification systems such as ICD-10 and DSM-IV alcohol dependence is defined as a cluster of physical, psychological symptoms and social consequences of alcohol consumption (Schuckit 1995). Patients who meet ICD-10 diagnosis of alcohol dependence must display three of the following six symptoms:

(a) a strong desire or compulsion to drink;
(b) tolerance;
(c) withdrawal;
(d) loss of control;
(e) progressive neglect of alternative activities; and
(f) persistent drinking despite evidence of harm.

3.6 Physical Withdrawal

Many but not all alcoholics develop physical dependence and experience physical and psychological withdrawal symptoms after cessation of alcohol consumption. A number of physiological mechanisms are

involved in the development of the syndrome. Basically the development of withdrawal symptoms can be explained by a number of adaptive mechanisms resulting from long-term alcohol intake. While alcohol enhances the neurotransmission of inhibitory neurotransmitters (GABA) and blocks excitatory neurotransmitters (glutamate, etc.) during alcohol withdrawal, there is an increased excitability in the CNS and an autonomic nervous system dysfunction with an excess release and turnover of excitatory neurotransmitters.

Symptomatology of alcohol withdrawal covers a wide range of symptoms, which develop few hours after the last drink with a peak on day two or three which usually subsides within four or five days. While alcohol withdrawal is usually mild in some cases a severe withdrawal syndrome can develop. Key symptoms are tremor, insomnia, malaise, anxiety, inner restlessness, sweating, increase in heart and respiratory rate, mild elevations in temperature, gastrointestinal symptoms such as anorexia, nausea, and vomiting, and psychological or emotional symptoms such as anxiety or sadness. A broad number of other symptoms may also be prevalent, depending on the patient's physical condition. In more severe cases, seizures (5–10 percent or more of patients) or hallucinations may complicate the clinical course. The most severe variant of alcohol withdrawal is alcohol withdrawal delirium.

Depending on the clinical course and symptomatology, inpatient or outpatient detoxification can be necessary. Pharmacological treatment includes fluid intake, substitution of vitamins and minerals, and sedatives, predominantly benzodiazepines or clomethiazole (in Europe only).

4. *Effects on the Body and Health*

Alcohol in light to moderate doses may have a slight beneficial effect in decreasing the risk for cardiovascular disease by increasing high-density lipoproteins (HDL) although this issue is still controversial. In any case this effect is far outweighed by the health risks in individuals with heavy alcohol consumption. Mean effects of alcohol in the body are as follows:

(a) Cardiovascular and cerebrovascular system: hypertension, heart inflammation or more often myocardiopathy, arrythymia.

(b) Brain: intracerebral hemorrhage. Other data indicate that mild to moderate alcohol consumption (< 50 g/day) may have some protective effect on the cerebrovascular and cardiovascular system, possibly by effects on blood lipids (inhibition of elevated low density lipoprotein (LDL) cholesterol, increase of HDL lipoproteins) and some antiatherogenic and antithrombotic effects.

(c) Neuromuscular system: polyneuropathy, myopathy, autonomic disorders.

(d) Digestive system: increased rate of gastritis and ulcer disease, pancreatitis (possibly followed by diabetes), abnormal functioning of the esophagus including esophagitis.

(e) Liver: fatty liver, acute alcoholic hepatitis, chronic active hepatitis and finally cirrhosis. From a chronic alcohol intake of 20 g for women and 30–40 g for men the risk for liver damage is already increasing.

(f) Blood cells: the production of all types of blood cells is decreased. Red-blood-cell anemia (macrocytosis), decrease of white-cell production and function, and decreased production of platelets and clotting factors are the result. Function of thymus-derived lymphocytes, which are essential for immune function, is also impaired.

(g) Sexual functioning and hormonal changes in men testicular atrophy, hypogonadism, decreased sperm production and motility, decreased testosterone production, and sometimes impotence are typical results of chronic alcoholism. In women menstrual irregularities are of relevance, as are effects on the fetus.

(h) Other endocrine and metabolic effects of alcohol include impaired thyroid and parathyroid function with an increased risk for osteoporosis and bone fractures. The glucose and carbohydrate metabolism is effected in complex ways. A frequent complication of alcohol also is diabetes, mostly due to pancreatitis.

(i) Skin: a number of dermatological conditions can be provoked or worsened by alcohol; porphyries, psoriasis vulgaris, rosacea, cancer of the oral mucosa, pellagra, and others.

(j) Increased risk for cancer of the mouth and digestive tract (pharynx, larynx, esophagus, stomach, liver), head and neck, lungs, and breast. A number of variables contribute to this phenomenon: alcohol toxicity, comorbid nicotine dependence (smoking), malnutrition, decrease in immune function, among many others.

(k) Malnutrition, vitamin deficiency, and electrolyte changes. Typically chronic alcoholism is associated with some form of malnutrition. Typical effects of chronic alcoholism are zinc deficiency, hypokalemia, deficiencies of vitamin B ($B_{1,3,12}$), vitamin C, and folic acid, among many others.

4.1 *Morbidity and Mortality*

The mortality and morbidity of individuals with heavy alcohol consumption, harmful use or alcohol dependence are significantly increased compared to the general population. Reasons are multifactorial: there are numerous somatic and neurologic disorders related to alcohol, also the risk for accidents and suicide is much higher in alcoholics. In addition the alcoholic's lifestyle (nicotine consumption, low-protein/high caloric food, etc.) also contributes to the increased

morbidity and reduced life expectancy in alcoholism. Some studies indicate that 5 percent of all deaths are related to alcohol consumption.

4.2 *Alcohol Embryopathy* (*Fetal Alcohol Syndrome*)

A tragic and often underestimated result of drinking during pregnancy is alcohol embryopathy. Key features are intrauterine growth retardation, microcephaly, moderate to severe mental retardation, and relatively typical dysmorphologic facial malformation. Many other symptoms may also be present such as internal and genitourinary malformations, especially congenital heart defects, among many others. The degree of alcohol embryopathy is correlated to the stage of maternal alcohol illness, not to the maternal alcohol consumption.

5. *Psychiatric Complications*

A number of distinct neuropsychiatric disorders are caused by chronic alcohol intake including delirium, psychosis, anxiety, and depression and an increased risk for suicide/delinquency. There is a high comorbidity of alcoholism with depression, schizophrenia, antisocial personality, and other personality disorders, anxiety, and addiction to other substances and drugs including tobacco. Alcoholism sometimes develops prior to the psychiatric disorder but in many cases it is secondary, worsening the clinical course and symptomatology.

6. *Driving Ability and Accidents*

Alcohol has major effects on driving ability as well as the risk for accidents. These factors contribute significantly to the increased morbidity and mortality in alcoholics. Even at a BAC of 0.15 mg/liter the ability to operate a motor vehicle is significantly impaired. The risk, especially for more severe accidents, rises dramatically with increasing BAC. There are marked differences concerning permitted BAC while operating a motor vehicle in different countries. Some countries do not allow any alcohol intake, while in most states 0.5 or 0.8 mg/liter BAC are the upper limits tolerated. The risk for other accidents (home, workplace, sports) also rises with increasing BAC.

7. *Nervous System*

7.1 *Alcohol Intoxication*

Probably the most frequent alcohol-related disorder is alcohol intoxication. Although there is no strict correlation between blood alcohol level and behavioral or motor impairment the symptomatology depends on BAC, individual alcohol tolerance, and a number of confounding factors such as physical and psychiatric status, intake of other substances, sleep deprivation, among others. Alcohol intoxications can be classified as mild (BAC 0–1), moderate (BAC 1–2) and severe (BAC over 2–2.5 mg/liter). The higher the BAC, the more pronounced are the CNS depressant, sedative effects and the behavioral dysfunction. Light to moderate alcohol intoxication is usually associated with relaxation and feelings of euphoria as well as impaired coordination. Higher BAC results in severe cognitive, perceptual, and behavioral impairment. This includes blackouts, insomnia, hangover. On the physical level alcohol intoxication is associated with hypertension, cardiac arrhythmia, gastrointestinal symptoms such as vomiting, diarrhea, abdominal pain, nausea, anorexia, gastritis, and hepatitis. Neurological symptoms include ataxia, fainting, and blackouts. Trauma and accidents (traffic safety!) are of special relevance. On the psychological and behavioral level insomnia, anxiety, depression, sexual problems, inappropriate, aggressive or impulsive behavior may occur. Severe cognitive impairment, clouding of sensorium, disorientation, and amnesia can be found in higher BAC. Mortality is significant in BAC > 4 mg/liter due to respiratory paralysis, heart failure, and coma.

7.2 *Pathological Intoxication* (*Alcohol Idiosyncratic Reaction*)

In some individuals (although rare) mild to moderate alcohol intoxications can be associated with severe aggression and violence or psychotic reactions lasting for a few hours and usually followed by a more or less complete amnesia. The psychiatric and behavioral symptoms are very marked and cannot be explained by the comparatively low BAC level. This syndrome which is basically of forensic interest is very controversial among clinicians. Possible predisposing factors are hypoglycemia or other metabolic disorders, organic brain syndromes, or intake of other psychotropic drugs such as stimulants.

7.3 *Delirium*

Delirium usually starts during the first four to seven days after cessation of alcohol consumption. Key features of delirium are clouding of sensorium, disorientation, and severe confusion, fear and agitation, visual and sometimes acustic hallucinations, delusions of persecution or others. Delirium is a very serious medical disorder which is more common than other alcohol psychosis (prevalence rate about 1 percent) and has a significant mortality if untreated. Symptoms found in alcohol withdrawal can also be seen in

alcohol delirium but are usually more severe. The clinical condition is characterized by a severe overactivity of the autonomic nervous system (increased pulse rate and respiratory rate, marked elevation in blood pressure and body temperature). Frequent complications are seizures, cardiac arrhythmia, and many other medical disorders. Patients need substantial medical support and psychopharmacological treatment, usually sedatives such as benzodiazepines.

7.4 Alcohol Psychosis

Chronic alcohol consumption can result in different alcohol psychoses. In some cases a more or less chronic state with suspiciousness or more pronounced paranoid delusions can develop. This disorder is referred to as alcoholic paranoia or alcohol-induced psychotic disorder. The prototype of this psychosis is a delusional jealousy syndrome nearly exclusively found in male alcoholics who believe their spouse to have an extramarital relationship. Sometimes without the slightest evidence the alcoholic is convinced about his spouse's infidelity. Predisposing factors for the development of this syndrome are impotence or other sexual dysfunction, cognitive impairment, and a low self-esteem. The delusions often persist into abstinence. Delusional jealousy is a dangerous disorder with the patient often attacking or even killing his spouse.

The other more prevalent alcohol-induced psychosis is alcohol hallucinosis which is characterized by vivid predominantly acoustic, sometimes visual hallucinations, delusions of reference or persecution, and fear. Other psychotic symptoms may also be prevalent. Different from alcohol withdrawal delirium the sensorium is usually clear and there is no amnesic syndrome for the psychosis. The psychopathology of alcohol hallucinosis closely resembles paranoid schizophrenia but there is no evidence for a common genetic basis. Alcohol hallucinosis, like alcohol paranoia, can develop during heavy drinking or more frequently within a few days or weeks of the cessation of drinking. In abstinent patients the prognosis of alcohol hallucinosis is usually good, but in 10 to 20 percent a chronic, schizophrenia-like psychosis can develop. Psychopharmacological treatment in alcohol psychosis (neuroleptics, sedatives) is recommended.

7.5 Organic Brain Syndrome, Encephalopathy, and Dementia

While some form of cognitive impairment can be found in up to 75 percent of chronic alcoholic patients approximately 9 percent of them have clinically manifest organic brain syndrome. Alcohol itself, but also alcohol-related disorders such as malnutrition including vitamin deficiencies as well as indirect consequences of alcoholism, such as head trauma, hypoglycemia, or other metabolic disturbances can cause cognitive dysfunction, mental confusion, and clouding of sensorium. Serious confusion can be seen during alcohol intoxication and withdrawal, as a result of vitamin deficiency (e.g., thiamin), head trauma, extra- or intracranial hematoma, stroke, hypoglycemia, or simply as a result of long-term alcohol intake, or a combination of these factors.

Wernicke encephalopathy, a dramatic, very acute neurologic syndrome with high mortality, is characterized by a classical symptom trias: ataxia, opthalmoplegia, mental disorder (clouding of consciousness). Thiamin deficiency is essential for the development of the syndrome. Patients are disoriented or confused, somnolent or even in coma, show oculomotor abnormalities and gait ataxia. There are distinct symmetric punctuate hemorrhagic lesions in certain brain areas. Rapid thiamin substitution is essential for therapy.

Wernicke encephalopathy in many cases is followed by Korsakoff Syndrome (alcohol-related amnesic syndrome) which may also develop without prior Wernicke symptomatology. Key features are anterograde and retrograde amnesia, memory loss, and other cognitive impairment. Apathy, passivity, and confabulations are common symptoms. Prognosis is poor.

Other patients show symptoms of a more gradual cognitive decline and other dementia symptoms without distinct neurological symptoms. Alcohol dementia is a difficult diagnosis. A broad number of other dementia forms including Alzheimer's disease have to be excluded before diagnosis can be made. Chronic hepatic encephalopathy also goes along with cognitive impairment but other neurological symptoms can also be found: frontal release signs, hyperreflexia, pyrimidal signs, or others. Organic brain syndromes can also be found as a result of other alcohol-related disorders.

7.6 Seizures

Epileptic seizures are the most frequent neurological sequelae with prevalence estimates of 15 percent or more. The exact pathophysiological basis is unclear. Electrolyte imbalances and neurotransmitter dysfunction (GABA, glutamate) are of special relevance. This disorder is independent from duration of alcoholism and there is no evidence for a genetic risk for seizures in these patients.

Seizures usually occur within the first 24 to a maximum 48 hours of abstinence and nearly exclusively are of tonic-clonic grand mal type. The clinical and neurological status is usually normal. Other seizure types, especially focal seizures indicate a probable focal brain injury (trauma, hemorrhage, etc). Electroencephalogram and cranial computer tomography may help excluding other reasons than alcohol

for seizures but are otherwise usually normal. Prognosis for seizures in abstinent alcoholics is good, otherwise the risk for recurrent seizures is high.

7.7 Polyneuropathy

This is a frequent complication of alcoholism (prevalence 9 to 30 percent). Beside diabetes, alcohol is the most common cause for polyneuropathy. A number of peripheral nerves with sensory, motoric, or autonomic fibers are affected. The sensoric input and in more severe cases the motoric system and muscle function are impaired. Typical complaints are symmetric burning or stabbing pain in the feet and mild to more severe weakness of the limbs. Polyneuropathy usually gradually develops and the prognosis in abstinent patients is often positive.

7.8 Myopathy

Alcohol has myotoxic effects both on skeletal and cardiac muscles. The more dramatic acute myopathy which can be accompanied by sometimes extended muscle necrosis, hypokalemia, and secondary renal failure has a prevalence of 0.8–3.3 percent. Chronic myopathy, often with subclinical symptomatology is much more common (23–66 percent). Myopathy can also be secondary to polyneuropathy. While the acute form goes along with painful muscle swelling, tenderness, and muscle cramps, in the chronic form extended weakness in the muscles is reported. A rare subtype is a myopathy related to hypokalemia.

7.9 Autonomic Disorders

Alcohol can also affect autonomic nerves and cause various autonomic dysfunctions (both parasympathicus and sympathicus): dysphagia, esophageal dysfunction, abnormal pupillary reflexes, impotence, impaired thermoregulation, among many others. Autonomic disorders are seldom isolated but usually accompanied by other alcohol-related disorders.

7.10 Cerebellar Atrophy

Up to 30 percent of chronic alcoholics show some clinical or neuroradiological symptoms of cerebellar atrophy. Histologically a degeneration of Purkinje cells in the anterior and superior vermis is seen as well as in the cerebellar cortex. The disorder does not correlate with lifetime consumption of alcohol. Other factors such as vitamin deficiency seem to be of relevance. Cerebellar atrophy develops slowly. Key symptoms are dysarthria, gait and stand ataxia, tremor, and nystagmus. Lower limbs show more impairment than upper limbs. Severe forms of cerebellar atrophy cause astasia and abasia. Symptoms are often at least partially reversible in cases of abstinence and vitamin substitution.

7.11 Cerebral Vascular Diseases

There is an increased risk for intracerebral and subarachnoidal hemorrhages in chronic alcoholism with severe neuropsychiatric symptomatology depending on location and size. The association with ischemic stroke is less clear. A more frequent complication is chronic subdural hematoma, often preceded by some sometimes-minor head trauma. Symptoms can initially by very mild or even missing. Headache is the most frequent symptom.

7.12 Central Pontine and Extrapontine Myelinolysis

A rare complication of alcoholism. A very rapid substitution of hyponatremia, a common electrolyte imbalance in alcoholism, seems to be of special relevance for the development of the demyelination in the pons or some other areas of the brain. Clinical symptoms are severe with a high mortality: tetraparesis, cerebellar ataxia, bulbary symptoms, paresis of eye muscles, and central fever. The extreme form is a locked-in-syndrome with complete tetraplegia.

7.13 Marchiafava–Bignami Syndrome (Corpus Callosum Atrophy)

Another extremely rare disorder with poor prognosis and uncertain pathophysiology. In some chronic alcoholics, especially red-wine drinkers in the Mediterranean, a necrosis of the corpus callosum and sclerosis of the cerebral cortex can lead to confusion, clouding of sensorium, seizures and other neurological symptoms, coma, and death. If the patient survives dementia is the most frequent outcome.

7.14 Tobacco–Alcohol Amblyopy

A bilateral affection (demyelination) of the optic nerve, chiasma opticum, and tractus opticus can lead to blurred or loss of vision. This rare syndrome can predominantly be found in heavy-smoking alcoholics with malnutrition. Tobacco smoke contains cyanides, which cannot sufficiently be detoxified in patients with severe liver dysfunction. They are believed to affect the optic nerve by free cyanides. Prognosis is rather poor.

7.15 Alcohol-related Myelopathy

An extremely rare disorder with good prognosis. Alcohol myelotoxicity, malnutrition, and chronic liver

damage can cause a progressive myelopathy with spastic paraparesis, neurogenic bladder dysfunction, and paresthesia.

7.16 Movement Disorders

Occasionally extrapyramidal symptoms, similar to Parkinson's disease, or dyskinesias can be seen in chronic alcoholics. The prognosis in abstinent patients is usually good. The more frequent essential tremor can be suppressed by small amounts of alcohol. This syndrome is not a result of chronic alcoholism.

7.17 Sleep Disorders

Alcohol consumption has major impact on the sleep architecture. Acute intake can lead to decreased latency to sleep onset, increased slow wave sleep, and decreased REM (rapid eye movement) sleep during the first half of the night. Insomnia is frequent during alcohol withdrawal and can persist long into abstinence. Other sleep disorders, e.g., sleep apnea syndrome, is usually worsened by alcohol.

See also: Alcohol Use Among Young People; Alcoholics Anonymous; Alcoholism: Genetic Aspects; Drinking, Anthropology of; Drug Addiction; Drug Addiction: Sociological Aspects; Korsakoff's Syndrome

Bibliography

British Medical Association 1995 *Guidelines on Sensible Drinking*. British Medical Association, London

Schuckit M A 1995 *Drug and Alcohol Abuse, Fourth Edition. A Clinical Guide to Diagnosis and Treatment*. Plenum, New York

Zernig G, Saria A, Kurz M, O'Malley S S (eds.) 2000 *Handbook of Alcoholism*. CRC Press, Boca Raton, FL

M. Soyka

Alcohol Use Among Young People

Alcohol use is prevalent among people beyond childhood and shows an intriguing association with age. Consumption increases rapidly across adolescence, shows a peak in the early twenties and declines gradually thereafter, once the major developmental tasks of emerging adulthood are resolved. Whereas young children disapprove of drinking, from adolescence on alcohol consumption is most often seen as signifying one's growing social maturity.

The developmental-psychological perspective chosen departs from these observations and explains the emergence of alcohol use among the majority of young people as embedded in the normative psychosocial challenges of adolescence (Silbereisen and Eyferth 1986). This period of the life span is characterized by growing attempts to find a particular place in life, which involves dealing with new social expectations and personal aspirations. The increasing interest at this time in novel and risky activities, and the unsupervised environments associated with them, probably also has neurobiological underpinnings related to the increase in dopamine input to prefrontal cortex and limbic brain regions during early adolescence (Spear 2000). Taken together, both viewpoints justify treating alcohol use among young people as a separate issue, distinct from alcohol use in general.

Abuse of alcohol is a relatively rare form of use, characterized by consumption over extended periods of time in situations which require clarity of perception and judgment; drinking of even small amounts if educated decisions are not possible due to developmental immaturity; increasing the level of alcohol in order to compensate for declining psychoactive effects or to avoid malfunctioning; and all forms of consumption which impair health or adequate mastery of normative exchanges with the environment (Newcomb and Bentler 1989). Only a small subset of young people meets the clinical criteria for substance use disorders (see *Sexual Risk Behaviors*).

1. Consumption Prevalence and Trends Across Age

According to representative school surveys, such as the Monitoring the Future study in the USA (O'Malley et al. 1999), the lifetime prevalence of alcohol use among 12th graders is of the order of 80 percent or higher (in contrast, episodic heavy drinking [five drinks or more in a row] amounts to about 30 percent). Concerning frequency, one-third of 14- to 24-year-olds in a large German community sample reported drinking less than once per week, one-third up to twice a week, and only the remaining third reported consuming alcohol more often, including daily. With regard to quantity consumed, it has been estimated that on a drinking day about 20 percent in this age group consume up to two, but almost 50 percent more than five standard drinks (9 grams ethanol in Germany) (Holly and Wittchen 1998). In general, gender differences in consumption among the young are small among moderate drinkers.

Beginning with the teen years and their new freedoms and challenges, frequency and amount of consumption increase rapidly. According to a meta-analysis of more than 20 longitudinal studies (Fillmore et al. 1991), the increase in frequency and quantity peaks in the early twenties, followed by a similarly sharp decline, particularly for frequency, which seems

to be triggered by a general age-related trend toward conventionality (Jessor et al. 1991) and growing incompatibilities between consumption and new responsibilities as partner, parent, and worker.

Whereas countries like the US, Canada, and the UK share relatively moderate consumption, some Mediterranean and Eastern European countries rank much higher. In longer perspective, consumption in industrialized countries increased dramatically after World War II, reaching unprecedented peaks in the 1970s and 1980s, followed by stable or slightly declining figures thereafter (Silbereisen et al. 1995). Consumption in former socialist countries, however, has been increasing since the early 1990s.

2. *Immediate Negative Consequences for Well-being*

Due to the overall moderate and/or time-limited alcohol consumption among adolescents, most of the consequences for well-being are immediate. According to data from the UK (Miller and Plant 1996), between 5 percent and 30 percent of young people in mid-adolescence report problems associated with alcohol use in areas of social functioning such as personal adversities (reduced performance in school), social relationships (tensions with friends), sexuality (unwanted sexual encounters), and delinquency (trouble with police). More serious consequences are very rare.

Adverse immediate health consequences relate primarily to intoxication. Due to cultural differences in drinking habits, this experience is more commonplace in Nordic countries in Europe, in spite of higher consumption figures in the South. Very few young people develop alcohol-related conditions such as liver cirrhosis.

Among young people in mid-adolescence, 7 percent reported a buildup of tolerance, and 16 percent wanted to cut down consumption (Substance Abuse and Mental Health Services Administration 1996). In general, a substantial minority sometimes experience discomfort, including feeling dizzy, hangovers, and headaches. An estimate of the dependence potential of alcohol is the 6 percent share of those in a normal sample (ages 14–24) diagnosed with substance use disorder.

Risky sexual behavior and alcohol correlate. This is probably not so much due to being uninhibited under the influence of alcohol but is rather rooted in common situational encounters, often concentrated in small subgroups, which may also share other risk factors such as mental disorders (see *Sexual Risk Behaviors*

Alcohol use is not a gateway drug, but it is certainly true that most users and abusers of other psychoactive substances begin with (and often maintain) the use of alcohol: earlier and heavier use are associated with later drinking problems (Kandel et al. 1992), but the causal mechanism is unknown.

As far as legal consequences are concerned, in spite of public concerns about the easy accessibility to alcohol for minors, there is little attempt to prosecute. In some countries the legal age for driving is considerably lower than that for alcohol drinking and purchase, which may exacerbate the problem of young people's reckless driving under the influence of alcohol. In Germany, about 20 percent of all fatal car crashes caused by young drivers (ages 18–24) happen in a total of 12 hours dispersed across Friday and Saturday night, and occur on the way home from suburban discotheques, the car loaded with overexcited young people, and further handicapped by drivers' fatigue (Schulze and Benninghaus 1990).

3. *Role in Normative Psychosocial Development*

Following an approach forwarded by Moffitt (1993), a deeper understanding of the age trends, associations with biographical transitions, and immediate consequences of alcohol consumption can be achieved by distinguishing two sets of developmental antecedents and motives (see *Adolescent Development, Theories of*). With regard to the adolescence-limited trajectory, which is characteristic of the absolute majority, alcohol use emerges because almost all adolescents must wait for the status and privileges of adults, despite their physical maturity, for several years (due to the ever expanding schooling this gap is growing historically). Once they have resolved these issues, the frequency and intensity of problem behaviors, including alcohol, will vanish due to the influence of new environments that entail fewer opportunities and provide more deterrents concerning use. The life-course-persistent trajectory, in contrast, maintains consumption beyond the normative transitions to adulthood and is rooted in long-lasting problems of adaptation, starting in early childhood and encompassing neurological problems, attention deficit, impulsivity, and the like.

Moffitt's (1993) model matches well with more elaborate distinctions in the literature on alcohol and alcoholism where one of the subtypes is described as genetically influenced, with early behavioral maladaptations, and embedded in a long-lasting antisocial personality disorder (Tarter et al. 1999). Moreover, it also enables the remarkable co-variation among alcohol use and other, particularly externalizing problem behaviors to be understood. These problem behaviors, such as reckless driving or unprotected sexual activities, signify status for the young but are deemed inadequate by the community due to their precocity. Our general notion that it is the maturity gap which channels the alcohol use of the vast majority of adolescents sounds rather negative. Note, however, that most adolescents perceive alcohol as a means to ease social contacts and improve feelings in such contexts. Only a small minority takes alcohol with the

purpose of mood regulation when facing problems, as do many adults (Tennen et al. 2000).

Such motives turn into reality particularly with regard to the formation of peer and romantic friendships, which are major developmental tasks in the second decade of life. Moderate consumption among those on the adolescent-limited trajectory corresponds prospectively to higher status and better cohesion within one's peer group, and is associated with a higher likelihood of romantic involvement. Moreover, adolescents seem to select leisure settings that offer opportunities for friendship contacts and provide alcohol in the right quantity and environment, such as discotheques, quite deliberately (Silbereisen et al. 1992). In a nutshell, alcohol consumption also has constructive functions in healthy psychosocial development.

4. *Prevention*

Given the almost normative use of alcohol among young people in many cultures, efforts aiming at prevention typically target responsible, self-controlled, and health-conscious use rather than abstinence. The multifunctional role of alcohol in the resolution of developmental tasks during adolescence and emerging adulthood represents the major pivot for primary prevention. Appropriate measures need to be undertaken early enough, that is, in late childhood/early adolescence, parallel to the first attempts to actually utilize the possible roles of alcohol consumption in the negotiation of the adolescent challenge.

Concerning measures on the environmental level, one needs to reduce contexts which entail schedules known to provoke the habituation of drinking, such as the episodic availability of large quantities in seducing locales (like 'binge' drinking in fraternity settings). Efforts most remote to the individual concern attempts to reduce national levels of per capita consumption in general, but curbing heavy drinking seems to affect consumption among the adult population, not the young. More specific measures try to minimize the harm by enforcing controls on the drinking settings, such as the establishment of licensing hours or by training bar tenders to refuse serving alcohol to drivers (Plant et al. 1997).

The family is the proximal environment for most adolescents that represents a major source of risk factors for drinking, such as parental modeling, inconsistency in rule setting, and a lack of developmental challenge. However, very few attempts at prevention on the family level exist to date.

Concerning prevention at the individual level, targeting adolescents at school is the rule. Given the role of alcohol use in response to normative developmental difficulties, prominent programs address general life skills, such as adequate self-perception, empathy with others, critical thinking, decision-making, communication, sociability, affect regulation, and coping with stress (Botvin 1996).

In addition, as revealed by recent meta-analyses of evaluation studies (Tobler and Stratton 1997), the most successful programs are characterized by a combination of general skill development and substance-specific elements aiming at proximal risk/protective factors of use and abuse. Prominent among such programs are those offering factual information about alcohol-specific physiological and psychological states, the formation of negative attitudes (e.g., by demonstrating the partial incompatibility between alcohol and relationship goals), and the practical training of how to resist unwanted offerings of alcohol by peers. Adolescents in general are prone to conform with behavioral standards of their peers, consumption of alcohol included, but there is also a mutual selection effect among those with similar behavior patterns. It is important not to expect sustainable effects unless intervention takes places repeatedly at major milestones during adolescence and beyond.

Concerning the life-course persistent trajectory of alcohol use, prevention as described would begin too late and is inefficient (may even hurt those on the adolescence-limited trajectory due to heightened contacts with negative role models). Rather, prevention would need to start at a much earlier age, and would need to address directly the associated early childhood problems such as impulsivity.

See also: Adolescent Health and Health Behaviors; Alcohol-related Disorders; Alcoholism: Genetic Aspects; Health Education and Health Promotion; Health Promotion in Schools; Substance Abuse in Adolescents, Prevention of

Bibliography

Botvin G 1996 Substance abuse prevention through Life Skills Training. In: DeV Peters R, McMahon J (eds.) *Preventing Childhood Disorders. Substance Abuse and Delinquency*. Sage, Newbury Park, CA, pp. 215–40

Fillmore K M, Hartka E, Johnstone B, Leino E, Motoyoshi M, Temple M 1991 A meta-analysis of life course variation in drinking. *British Journal of Addiction* **86**: 1221–68

Holly A, Wittchen H-U 1998 Patterns of use and their relationship to DSM-IV abuse and dependence of alcohol among adolescents and young adults. *European Addiction Research* **4**: 50–7

Jessor R, Donovan J, Costa F 1991 *Beyond Adolescence. Problem Behavior and Young Adult Development*. Cambridge University Press, Cambridge, UK

Kandel D B, Yamaguchi K, Chen K 1992 Stages of progression in drug involvement from adolescence to adulthood: Further evidence for the gateway theory. *Journal for the Study of Alcohol* **53**: 447–57

Miller P, Plant M A 1996 Drinking, smoking and illicit drug use among 15 and 16 year olds in the United Kingdom. *British Medical Journal* **313**: 394–7

Moffitt T 1993 Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review* **100**: 674–701

Newcomb M, Bentler P 1989 Substance use and abuse among children and teenagers. *American Psychologist* **44**: 242–8

O'Malley P M, Johnson P M, Bachman J G 1999 Epidemiology of substance abuse in adolescence. In: Ott P J, Tarter R E (eds.) *Sourcebook on Substance Abuse: Etiology, Epidemiology, Assessment, and Treatment*. Allyn & Bacon, Boston, pp. 14–31

Plant M A, Single E, Stockwell T (eds.) 1997 *Alcohol: Minimising the Harm: What Works?* Free Association Books, London, New York

Schulze H, Bennighaus P 1990 *Damit sie die Kurve kriegen—Fakten und Vorschläge zur Reduzierung nächtlicher Freizeitunfälle junger Leute* [*Facts and Suggestions to Reduce Night Time Traffic Accidents Among Young People*]. Deutscher Verkehrssicherheitsrat, Bonn

Silbereisen R K, Eyferth K 1986 Development as action in context. In: Silbereisen R K, Eyferth K, Rudinger G (eds.) *Development as Action in Context: Problem Behavior and Normal Youth Development*. Springer, New York, pp. 3–16

Silbereisen R K, Noack P, von Eye A 1992 Adolescents' development of romantic friendship and change in favorite leisure contexts. *Journal of Adolescent Research* **7**: 80–93

Silbereisen R K, Robins L, Rutter M 1995 Secular trends in substance use: Concepts and data on the impact of social change on alcohol and drug abuse. In: Rutter M, Smith D (eds.) *Psychosocial Disorders in Young People: Time Trends and Their Origins*. Wiley, Chichester, UK, pp. 490–543

Spear L P 2000 Neurobehavioral changes in adolescence. *Current Directions in Psychological Science* **9**: 11–114

Substance Abuse and Mental Health Services Administration 1996 *National Household Survey on Drug Abuse: Main Findings 1994*. US Department of Health and Human Services, Rockville, MD

Tarter R, Vanyukov M, Giancola P, Dawes M, Blackson T, Mezzich A, Clark D 1999 Etiology of early onset substance use disorder: A maturational perspective. *Development and Psychopathology* **11**: 657–83

Tennen H, Affleck G, Armeli S, Carney M A 2000 A daily process approach to coping. *American Psychologist* **55**: 626–36

Tobler N, Stratton H 1997 Effectiveness of school-based drug prevention programs: A meta-analysis of the research. *The Journal of Primary Prevention* **18**: 71–127

R. K. Silbereisen

Alcoholics Anonymous

Alcoholics Anonymous (AA) is a self-help organization for persons with a desire to stop drinking (Bill 1976). AA is not a formal treatment for alcohol problems, but rather a program for living. The AA program is based on 12 steps to recovery, and is often called a '12-Step program.' Many other 12-Step programs have developed that are modeled after AA, including programs for other substance use disorders, for families of those affected by alcohol and drug use disorders, and for nonsubstance related problems that involve the experience of loss of control of certain aspects of behavior.

AA meetings are widely available. Meetings may be open meetings for any interested individual or closed meetings for alcoholics only. The format of meetings varies, with both discussion-oriented meetings and meetings with speakers. There is no charge to attend AA meetings, but voluntary contributions are accepted. The only requirement for AA membership is a desire to stop drinking. Newcomers to AA are encouraged to attend 90 meetings in 90 days. AA members work with a sponsor, a member with more experience in recovery who provides guidance and support to the member.

The organization and functioning of AA is defined by the 'Twelve Traditions,' which articulate the principles of anonymity, lack of affiliation with organizations, and the autonomy of the individual AA group (Bill 1952). AA publishes a large variety of books and pamphlets, many of which have been translated into multiple languages. The two core books are *Alcoholics Anonymous*, often call the '*Big Book*,' and *Twelve Steps and Twelve Traditions*, often called the '*Twelve and Twelve*.'

AA developed within the social context of the USA in the 1930s, and this article reviews the evolution of AA within that social context. An extensive empirical literature exists on the structure, functioning, promulgation, and effectiveness of AA, and an overview of major research findings is presented. Unique methodological issues in conducting research on AA, and future directions for such research conclude the article.

1. *Historical Origins*

1.1 Alcoholism Treatment in the Nineteenth and Twentieth Centuries

In the USA, the latter part of the nineteenth century witnessed the development of a network of facilities for the treatment of inebriety and dipsomania. Both large hospitals and smaller rehabilitative homes provided treatment that was often mandated by a judge at the request of the family of the alcoholic. After the turn of the century, however, most facilities closed, and alcoholics were assisted largely through either the social welfare system or the criminal justice system.

1.2 Founding and Early Development of Alcoholics Anonymous

Alcoholics Anonymous developed in the post-Prohibition culture of the USA in the 1930s. The country was in the midst of economic depression, and few health care professionals or scientists were focusing their attention on the problems of alcohol dependence. AA was begun in 1935 in Akron, Ohio by two alcoholic men, Bill W., a stockbroker, and Dr. Bob S., a physician. These two men developed the initial concepts and structures that continue to guide AA.

AA was developed in a culture that idealized individualism and individual achievement. The belief that individual effort would inevitably lead to progress was undermined by the experience of World War I and the Great Depression in the USA. The undermining of the assumption of the value of individualism may have contributed to the collectivist and interdependent perspective of AA.

Although AA describes alcoholism as a disease that can be arrested but not cured, the core of the AA program focuses not on drinking but on personal growth and change. The root problems underlying alcoholism were viewed as self-centeredness and loss of a spiritual center. The program of AA was designed to deflate self-centeredness and to develop a spiritually meaningful way to live (Kurtz 1979). Thus, the program emphasized powerlessness, turning over responsibility for change to a higher power, recognizing one's flaws or character defects, confessing these defects to another, making amends, maintaining a close relationship with one's higher power through prayer and meditation, and bringing the message of AA to other 'suffering alcoholics.' The traditions emphasized that no individual could be a spokesperson for AA, that AA would own no property, that each group would be autonomous, and that AA would affiliate with no other organization. All of these principles were directly contrary to the values of personal influence, autonomy, and organizational growth, and fostered reliance upon and commitment to the collective group.

1.3 Evolution and Spread of Alcoholics Anonymous

The initial development of AA was slow, with membership of only about 100 by 1940. Media attention in the early 1940s fueled a period of rapid growth. Growth of AA has continued steadily. In the 1980s and 1990s, the rate of international diffusion of AA increased dramatically. By 1990, there were an estimated 87,000 AA groups in 150 countries, and over 1.7 million members around the world. The development of AA in other countries was influenced both by people from the USA visiting other countries and by natives of other countries learning about AA when visiting the USA.

Trends in membership and meetings reflect changes in AA. Membership in AA has shifted to include a larger proportion of women, members with concurrent problems with other drugs of abuse, and younger members. Reflecting these trends, the availability of 'special interest' groups has increased, particularly in the USA. The most common special meetings are gay/lesbian meetings, young people's meetings, and women's meetings.

The basic principles of AA have remained unchanged, but the implementation of AA has varied across cultures. Research by Mäkelä et al. (1996) and others has documented both the consistency and the heterogeneity in the practice of AA across AA members, AA groups, and countries. For example, cross-talk (directly commenting on another member's statements in a meeting) and negative feedback are not accepted during AA meetings, regardless of culture, and the relative importance of different steps seems similar across countries. In other ways, the practice of AA differs across cultures. For example, interpretations of the concept of a higher power vary substantially, as does the use of sponsors. Behavior during meetings also shows considerable cultural variation, particularly in the degree of physical and personal intimacy.

2. Research on Alcoholics Anonymous

2.1 Utilization of AA

AA members enter the program by a number of routes, including self-referral; referral by family, friends, or treatment centers; or through coercion from the legal system, employers, or social welfare system (Weisner et al. 1995). Surveys of the USA population reveal that almost 6 percent of US adults have attended AA at some point in their lives. Among individuals with a history of alcohol problems, more than 20 percent of men and 15 percent of women have attended AA. Less information is available about individuals who enter AA through the criminal justice system, and the practice of judicial orders to attend AA is controversial.

Epidemiological and clinical data suggest that alcoholics attending AA average just under one meeting per week. Involvement with AA varies widely, and rates of attrition are well over 75 percent in the first year. Among those who continue their involvement with AA, the probability of remaining sober and involved with AA is about 67 percent for those with one year of sobriety, 85 percent for those with two to five years sobriety, and 90 percent for those with more than five years of sobriety.

AA members are diverse in age, gender, ethnicity, severity of alcohol dependence, and a variety of other personal characteristics. Researchers have attempted to find a profile of the type of person most likely to become involved with AA. Although no one profile characterizes AA members, a large-scale review concluded that five variables were most predictive of successful AA affiliation: a history of using external supports to cope with problems, loss of control over drinking, a greater daily quantity of alcohol consumed, greater physical dependence, and greater anxiety about drinking.

2.2 AA and Population Subgroups

Two contrasting views of AA lead to different predictions about AA and different population subgroups. One perspective suggests that AA is a program

of recovery for alcoholics, and that the common experience of alcoholism should supersede superficial individual differences. An alternative perspective states that because AA was developed by educated, middle-aged, Caucasian, Christian, heterosexual males, its relevance to the young or elderly, persons of color, non-Christians, gays and lesbians, or women, is suspect. Research data about the relevance of AA to various subgroups is limited.

Recent research has examined women and AA. Women in AA tend to be older, are more likely to be employed, and have somewhat more severe drinking problems than the women in alcoholism treatment who do not attend AA. Women attending AA see the program as crucial to their sobriety, and the fellowship, support, sharing, and spirituality in AA are all seen as important as well. Women in recovery who do not attend AA feel that they do not fit in, that AA is too punitive and focused on shame and guilt; disagree with program principles related to powerlessness, surrender, and reliance on a higher power; and perceive AA as male-dominated.

Literature on cultural, ethnic, and racial subgroups and alcoholism in general is limited, and research on AA involvement for these groups is even more limited. White, Black, and Hispanic men and women all have positive views of AA, are likely to recommend AA as a treatment for alcohol problems, and recommend AA more than any other resource. There is some variability in support for AA, with fewer Asians viewing AA as a resource than individuals from other cultural groups. Among those with drinking problems, involvement with AA of different cultural and racial groups varies depending on the study population. Overall, Hispanics are more likely to have had contact with AA (12 percent) than either whites or blacks (5 percent). Among those involved with the criminal justice or welfare system, whites are most likely to have been involved with AA, but among those in primary health care settings, blacks are most likely to have been involved with AA.

No recent research has focused specifically on the experiences of either youth or the elderly in AA. Several studies of adolescent treatment, however, suggest a strong association between AA/NA involvement and abstinence. Research on the experience of gays and lesbians in relation to AA is lacking.

2.3 *The Effectiveness of AA*

One of the most consistent research findings is that there is a positive correlation between AA attendance and good outcome. Studies of treated and untreated individuals suggest that those attending AA are about 50 percent more likely to be abstinent than those not attending AA.

Evaluation studies have followed individuals receiving treatment in 12-Step-oriented treatment programs (Ouimette et al. 1999). These treatment programs have close conceptual links to AA, but are not to be confused with AA. Evaluations of individuals receiving treatment in 12-step-oriented treatment programs have found abstinence rates of 67–75 percent six months after treatment, and 60–68 percent 12 months after treatment. However, not all individuals who received treatment are reached in follow-up evaluations. If a researcher assumes that all individuals lost to follow-up have relapsed, then abstinence rates drop considerably.

A second way to study the effectiveness of AA is to randomly assign individuals to different forms of treatment that do or do not include AA. Several studies have examined the effectiveness of AA this way. These studies have not found AA or treatments designed to facilitate the involvement of AA to be more effective than other forms of treatments. However, one study found that AA involvement led to better treatment outcomes for individuals who had many friends and family members who were heavy drinkers (Longabaugh et al. 1998).

Research also has examined what aspects of AA involvement are related to drinking outcomes. Several factors predict affiliation with AA after treatment, including perceived past and future harm from alcohol use, anticipated benefits from abstinence, degree of commitment to abstinence, and drinking problem severity. There is a significant association between participation in AA activities and drinking outcomes. Aspects of participation most strongly associated with positive outcomes include increasing involvement with AA over time, leading meetings, having a sponsor, and doing 12th step work.

A final important outcome-related question is the degree to which AA involvement is associated with positive functioning in other life areas. Popular criticism of AA asserts that, although sober, AA members are psychologically dependent on AA and therefore are poorly adjusted. The research literature contradicts this perspective, with research demonstrating that those actively involved with AA have less anxiety, cope with problems more effectively, have more social support from friends, and better overall psychological adjustment.

3. *Methodological Issues in the Conduct of Research on AA*

3.1 *Sampling Issues*

Most research on AA is hampered by difficulties in accessing AA meetings and AA members. As a voluntary, anonymous organization, AA does not keep records, and does not enter into formal collaborations with researchers. Researchers, then, are faced with the challenge of developing methods to access

representative samples of AA members, or representative samples of AA groups. A number of methodologies have been suggested, but without clear data about the overall composition of the AA membership, researchers are limited in their ability to know if their samples are indeed representative.

3.2 Definitional Issues

One complex issue related to research on AA is the question: what is AA involvement? Early research classified subjects as attending AA or not. Somewhat more sophisticated studies measured attendance quantitatively, defining greater attendance as indicative of greater affiliation. More recently, researchers have approached affiliation as a multidimensional construct (e.g., Morgenstern et al. 1997) that includes attendance, endorsement of the central beliefs of AA, use of cognitive and behavioral strategies suggested by AA, degree of organizational involvement with AA, and degree of subjective sense of affiliation with AA. Although a consensus definition of AA involvement does not yet exist, there is widespread agreement that a multidimensional model is most appropriate.

3.3 Selection of Research Questions

Much of the research on AA has asked simple and relatively static questions: is AA effective; for whom; what are the characteristics of successful AA members? Involvement with AA, however, is a rich and complex experience, which varies across individual members, and varies over time within an individual member. Development of research strategies that can capture the heterogeneous nature of AA, both cross-sectionally and longitudinally, is a further challenge.

4. Future Directions

Conducting research on AA has entered the scientific mainstream. As scientists develop more sophisticated methods for accessing individuals attending and involved with AA, a number of previously unstudied issues can be examined. Longitudinal research should be conducted to study processes of involvement in AA, as well as changes in beliefs, behavior, and interpersonal relationships. Studies of constructs core to the AA program, such as spirituality, serenity, and sobriety are also of importance. Additionally, studies examining the processes of change in AA, within the context of larger models of personal change, would be important to developing more generalizable models for understanding AA. A comprehensive listing of potential topics for research on AA was generated at a recent scientific conference on AA, and is summarized in McCrady and Miller (1993). The interested reader is referred there for a fuller listing of potential directions for future research.

See also: Alcohol-related Disorders; Alcohol Use Among Young People; Alcoholism: Genetic Aspects; Drug Addiction; Drug Addiction: Sociological Aspects; Support and Self-help Groups and Health

Bibliography

Bill W 1952 *Twelve Steps and Twelve Traditions*. Alcoholics Anonymous Publishing, New York

Bill W 1976 *Alcoholics Anonymous: The Story of how Many Thousands of Men and Women have Recovered from Alcoholism*, 3rd edn. Alcoholics Anonymous World Services, New York

Kurtz E 1979 *Not-God: A History of Alcoholics Anonymous*. Hazelden Foundation, Center City, MN

Longabaugh R, Wirtz P W, Zweben A, Stout R L 1998 Network support for drinking, Alcoholics Anonymous and long-term matching effects. *Addiction* **93**: 1313–33

Mäkelä K, Arminen I, Bloomfield K, Eisenbach-Stangl I, Bergmark K H, Kurube N, Mariolini N, Ólafsdóttir H, Peterson J H, Phillips M, Rehm J, Room R, Rosenqvist P, Rosovsky H, Stenius K, Wiatkiewicz G, Woronowicz B, Zieliski A 1996 *Alcoholics Anonymous as a Mutual-help Movement*. University of Wisconsin Press, Madison, WI

McCrady B S, Miller W R (eds.) 1993 *Research on Alcoholics Anonymous: Opportunities and Alternatives*. Rutgers Center of Alcohol Studies, New Brunswick, NJ

Morgenstern J, Labouvie E, McCrady B S, Kahler C W, Frey R M 1997 Affiliation with Alcoholics Anonymous following treatment: A study of its therapeutic effects and mechanisms of action. *Journal of Consulting and Clinical Psychology* **65**: 768–77

Ouimette P C, Finney J W, Gima K, Moos R H 1999 A comparative evaluation of substance abuse treatment. III. Examining mechanisms underlying patient–treatment matching hypotheses for 12-step and cognitive-behavioral treatments for substance abuse. *Alcoholism: Clinical and Experimental Research* **23**: 545–51

Weisner C, Greenfield T, Room R 1995 Trends in the treatment of alcohol problems in the US general population, 1979–1990. *American Journal of Public Health* **85**: 55–60

B. S. McCrady

Alcoholism: Genetic Aspects

Humans have consumed alcoholic beverages since prehistoric times. The first source of alcoholic substances most likely was accidental fermentation of fruits or grains. Mead, a fermentation product of honey, existed in the Paleolithic age, and is typically regarded as the oldest alcoholic beverage. The process of making beer and wine from the fermentation of

carbohydrates dates back to early Egyptian times. Distilling products to obtain higher alcohol concentrations can be traced to the Arab world around 800 CE (Feldman et al. 1997). Despite widespread current use of alcoholic substances in many societies, alcoholism occurs among only a small percentage of people who drink. Alcoholism is described behaviorally, and can be characterized by excessive or compulsive use, or both, of alcohol and loss of control over drinking. It also includes drinking in an amount that leads to tolerance (need for increasing amounts in order to feel its effects), and physical dependence, a condition where symptoms such as anxiety and tremulousness (or, more seriously, seizures) occur when drinking ceases.

The prevalence of alcoholism differs from country to country. In the United States, for example, approximately 10 percent of adult males are diagnosed as alcoholics, and the annual economic costs of alcohol and drug abuse are estimated at $246 billion, with alcoholism by far the most severe substance abuse problem. The reason why only a small number of those who consume alcoholic beverages become alcoholics is unknown. It is clear, however, that determinants of alcoholism include an interaction between genetic and environmental factors. Alcoholism runs in families, with one-third of alcoholics having at least one alcoholic parent. Environmental risk factors include drug availability and low economic status. While genetic inheritance probably confers heightened vulnerability to alcoholism in some individuals, environmental manipulations may prevent or further foster the development of alcoholism, underscoring the importance of studying genetic–environmental interactions.

1. *Human Findings*

Alcoholism is a complex behavioral trait mediated by factors including socioeconomic environment, individual characteristics, and pharmacological factors. It was noted in the nineteenth century that alcoholism appeared to run in families (Vaillant 1983), and it is now clear that family history of alcoholism constitutes the strongest risk factor for the development of alcoholism. Twin studies support heritability of alcohol consumption and alcohol dependence. A monozygotic co-twin of an alcoholic (who is genetically identical) is about twice as likely to become an alcoholic as a dizygotic co-twin (who shares only 50 percent of the alcoholic co-twin's genes). Children of alcoholics raised by nonalcoholic adoptive parents also show increased susceptibility to becoming alcoholics. A Danish study found that 18 percent of 133 males with a biological paternal history of alcoholism, themselves developed alcoholism compared with only 5 percent of adoptees who did not have a positive biological family history for alcoholism (Ferguson and Goldberg 1997). Furthermore, sons of alcoholic parents who were adopted away had the same increased risk of becoming alcoholics as their biological brothers raised by their alcoholic parents.

However, a number of alcoholics do not have a family history of alcoholism, suggesting that the genetic component is not inherited in a simple fashion, and indicating that there are different forms of alcoholism. Although many physicians in the nineteenth century subtyped alcoholics, it was not until 1960 that a systematic categorization of alcoholism was developed by Jellinek (Vaillant 1983). Jellinek's types ranged from people with medical and psychological complications but not physical dependence (category 'alpha') to binge drinkers (category 'epsilon'). This classification scheme was useful for categorizing behavior of an alcoholic at a point in time, but Vaillant documented the fact that across time, alcoholics manifest different symptoms of the disease, limiting the utility of Jellinek's categories. Nonetheless, Jellinek made important contributions to the field of alcoholism typology, and some concepts in his classification scheme are still apparent in typologies of alcoholism used at the beginning of the twenty-first century.

More recently, adoption studies conducted in Scandinavia led to the postulate that there are two independently transmissible forms of alcoholism: type I and type II (Cloninger et al. 1996). Type I alcoholism is characterized by anxious personality traits, rapid development of tolerance to and dependence on the anti-anxiety effects of alcohol, typically has a late onset (after 25 years of age), and genetic predisposition seems to contribute only slightly. In contrast, type II alcoholism usually has an earlier onset, tends to predominate in men, has a high genetic predisposition, and is accompanied by antisocial personality traits and low impulse control. Other investigators have argued that typologies should distinguish severe problem drinkers from those with less severe problems. One model attempting to differentiate severity of drinking separates late-onset drinkers (type A) from affiliative/impulsive alcoholics (type B) and isolative/anxious alcoholics (type C) (Morey 1996). Late-onset drinkers demonstrate signs of alcohol abuse, but develop mild alcohol-dependence symptoms. The type B and C alcoholics are at an advanced level of alcohol dependence, and differ from each other with respect to variables such as personality traits and features of alcohol use. It seems clear now that these classification schemes represent only the extremes of a continuous spectrum of manifestations of alcoholism (Cloninger et al. 1996). In other words, an individual alcoholic may appear 'type II-like' or 'type B-like,' but in reality he or she will possess a unique developmental history and collection of diagnostically relevant traits that fits no single type perfectly. This complexity creates great difficulties for genetic analyses whose goal is to resolve 'genetic risk' into the identification of specific genes

that increase or decrease risk. Like most complex traits, many genes influence alcoholism, and each such gene is likely to increase or decrease risk very modestly.

1.1 Genome-wide Screens

In 1990 the Human Genome Project was initiated to collect genetic information on a large scale, and thus provided the basis for the field of genomics. Goals of the Human Genome Project include the identification of each of the estimated 80,000–100,000 genes in human DNA, the storage of this information in useable databases, and the development of tools for data analysis. The medical industry is using and adding to the knowledge and resources created by the Human Genome Project with the goal of understanding genetic contributions to human diseases, including alcoholism. Data from the Human Genome Project should, at least in theory, enable researchers to pinpoint alterations in specific genes that contribute to alcoholism. Although gene identification is only the first step toward understanding complete genetic contributions to alcohol abuse, there are a number of ways to identify genes provisionally that might be mediating alcoholism.

One approach to finding predisposing factors to alcoholism is to study human populations with little genetic or social/environmental variability. Using a Southwestern Native American tribe, researchers found several genetic markers linked with alcoholism (Long et al. 1998). One marker was located near a gene coding for a gamma-aminobutyric acid type A ($GABA_A$) receptor, while another was located near the dopamine D4 receptor subtype gene. Alcohol and other depressant drugs act at the $GABA_A$ receptor, which modulates inhibition in the brain by decreasing nerve cell excitability. The neurotransmitter dopamine is hypothesized to be partially responsible for mediating the reinforcing and rewarding properties of drugs of abuse, including alcohol. Thus, both of these represent plausible 'candidate genes' for alcoholism risk. Results from this study require verification, and further research will need to assess whether these candidate genes have a role in determining vulnerability to alcoholism, and to what extent these results can be generalized to other human populations.

In 1989 the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health initiated the Collaborative Study on the Genetics of Alcoholism (COGA). This project is a multidisciplinary approach to investigating the genetic components of susceptibility to alcoholism. COGA performed a genetic linkage study on a large sample of the general population in the United States, selecting families affected by alcoholism. One of the genetic markers distinguishing alcoholics from nonalcoholics was provisionally mapped to a location near the gene coding for the alcohol metabolizing enzyme alcohol dehydrogenase (ADH). Other evidence suggests that possession of a variant of the ADH gene tends to protect against the development of alcoholism in Asian populations.

ADH metabolizes alcohol to acetaldehyde, and the acetaldehyde itself is rapidly converted to acetate in the human liver by aldehyde dehydrogenase (ALDH2). ALDH2 has also been implicated in protecting against the development of alcoholism. The normal allele is designated ALDH2*1, but a point mutation produces a mutant allele designated ALDH2*2. This mutant allele produces an enzyme with deficient activity, and is dominant over the normal allele (individuals who are both homo- and heterozygous for ALDH2*2 do not have detectable ALDH2 activity in the liver). Most individuals in Asian populations of Mongolian origin commonly have the inactive (ALDH2*2) variant. Such individuals show high acetaldehyde levels after alcohol consumption, due to changes in alcohol metabolism. High levels of acetaldehyde lead to a facial flushing response, nausea, and other subjective feelings of alcohol intoxication. Thus, it is hypothesized that it is the slow removal of acetaldehyde after alcohol consumption in individuals possessing ALDH2*2 that protects these individuals from the risk of alcohol abuse.

Among Asians, ADH and ALDH genotypes may be useful for predicting resistance to alcoholism. Little evidence, however, suggests that an inherited defect in alcohol metabolism among Caucasians differentiates those prone from those resistant to alcoholism. Among other factors, this has prompted researchers to investigate other possible genetic differences between alcoholics and nonalcoholics.

While human genome-wide scans have contributed to our knowledge of genes influencing alcoholism, results from different studies do not always agree. Some studies support an association of a dopamine D2 receptor subtype gene polymorphism with increased risk for alcoholism, while other studies do not support this claim (Goate and Edenberg 1998). One potential confound in the population-based studies investigating this polymorphism is the fact that allele frequencies range from 9–80 percent among populations. Thus, careful ethnic matching of alcoholics and controls is imperative for future human research (Goate and Edenberg 1998).

Clinical and epidemiological research over the past few decades, combined with historical evidence, has made it clear there is heterogeneity among those diagnosed as alcoholic. One of the current directions of the COGA study is to use narrower definitions of alcoholism prior to performing genetic analyses (Goate and Edenberg 1998). As discussed above, alcoholism is genetically heterogeneous. That is, two individuals classified as alcoholics may differ with respect to personality characteristics, age at alcohol abuse initiation, and the severity of alcohol-related

health problems. It is likely that performing genetic analyses on these symptomatic subgroups will identify different candidate genes. However, the fact that many genes contribute to alcoholism risk will make it very difficult to identify individual genes using a broad population-based association strategy.

1.2 Mapping Genes for Phenotypes

An alternative approach to genome-wide screens is to identify phenotypes correlated with alcohol dependence. Once these phenotypes of interest are identified, it may be easier to map genes that affect them. Characteristics present in those likely to become alcoholics may provide useful markers indicating potential risk for the development of alcoholism. These correlated traits are sometimes termed 'endophenotypes.' One such trait is brain electroencephalographic (EEG) activity. In resting humans, EEG activity is under strong genetic control. Resting-state EEGs in sober alcoholics contain greater activity in the 'beta-wave' category, and a deficiency in alpha, delta, and theta activity, as compared with nonalcoholics.

Event-Related Potentials (ERPs) are another measure of brain electrical activity that indicate brain responsiveness to a number of external stimuli. ERPs are significantly more similar in monozygotic twins than in dizygotic twins or unrelated controls, supporting a genetic influence. Event-related potentials can be useful in detecting differences in information processing, and so may be useful for identifying inherited vulnerability to alcoholism. When exposed to a novel stimulus, alcohol-naive sons of alcoholics have a pattern of brain waves (called P3 or P300 evoked potentials) resembling those measured in alcoholics. Abstinent alcoholics also show a significantly reduced P3 evoked potential compared to controls. These differences in brain activity are hypothesized to reflect a genetic vulnerability to alcoholism.

While studies in human populations have been useful in the preliminary identification of specific genes mediating alcoholism, a number of future directions are needed. First, we must develop descriptive epidemiology to understand alcoholism better. Second, we need to develop culture-specific models. Finally, we need to create a clear definition of typologies to describe different subcategories of drinkers. Because of the very limited statistical power to detect genes in human studies, the key to progress in genetic studies of complex traits (such as alcoholism) is better articulation of the exact phenotype for which genes of influence are sought.

2. Animal Models

Genetic animal models offer several advantages over studies using human subjects. For example, the experimenter is in control of the genotype being studied as well as environmental variables. In humans, only monozygotic twins have identical genotypes, and it is much more difficult to control environmental variables. Many human responses have been modeled successfully in animals, including sensitivity to the acute response to a drug, the development of tolerance, withdrawal symptoms, and voluntary intake. Numerous mouse and rat genotypes are readily available, making it possible to share information between laboratories and build a cumulative information database.

2.1 Selected Animal Lines

A powerful genetic animal method is selective breeding. Similar to how farmers use selective breeding to increase milk production, the technique has been used in alcohol research to manipulate genotypes toward a specific objective (in this case, a specific response to alcohol). By mating animals that are sensitive to a drug (for example, those that prefer alcohol solutions or exhibit severe withdrawal), after several generations, most of the genes leading to sensitivity will be captured in these mice. At the same time that sensitive animals are mated, animals that are insensitive to the same response are mated, fixing most of the genes leading to low responsiveness in the insensitive line. To the extent that genes contribute to the selected trait, the sensitive and insensitive selected lines will therefore come to differ greatly on the trait. If they also differ on behaviors other than those for which they were selected, this is evidence that the same genes are responsible for both traits. Studies utilizing this technique have increased our knowledge of which responses to alcohol share similar genetic influence.

In the 1940s, Mardones and colleagues at the University of Chile initiated the earliest selection study for drug sensitivity to develop rat lines with low (UChA) and high (UChB) alcohol consumption. To select for differences in drinking, rats were offered a choice of water or alcohol. Rats showing high alcohol preference were mated, and rats showing low alcohol preference were separately mated. The degree of alcohol preference in the UChA and UChB lines diverged across generations, indicating hereditary transmission of alcohol drinking. Many other studies since then have supported this notion. In the 1960s, a Finnish group led by Eriksson at Alko's physiological laboratory developed rat strains selected for voluntary alcohol consumption, the AA (Alko Alcohol preferring) and the ANA (Alko Non-Alcohol preferring). The AA and ANA lines also differ dramatically in voluntary alcohol intake, again supporting a genetic component to this behavior.

In the 1970s, Li, Lumeng, and their colleagues in Indiana also developed rat lines that either preferred (P), or did not prefer (NP), alcohol. This selection study was replicated by the same group in the 1980s, using the same protocol, to develop High (HAD) and

Low (LAD) Alcohol-Drinking rat lines. The existence of a number of different lines selected for and against alcohol preference has provided the opportunity to discover convergence in the genetic correlates of preference for alcohol. For example, one common result from these selected lines is that high-drinking genotypes appear to have lower brain serotonin, a neurotransmitter involved in mood and emotional responses. Furthermore, these results may relate to difference among humans, as at least some alcoholics also show lower serotonin activity.

Virtually all other responses to alcohol have demonstrated a heritable component as well, and selections have been performed for several responses in addition to drinking. In the early 1970s, Goldstein demonstrated that mice selectively bred for alcohol withdrawal symptoms developed progressively more severe withdrawal across three generations. In the late 1970s, Crabbe and his colleagues initiated an animal model of genetic sensitivity to severe and mild withdrawal after chronic alcohol exposure. Lines of mice were bred to exhibit either severe alcohol dependence measured by withdrawal symptoms (Withdrawal Seizure-Prone; WSP) or reduced response following dependence and withdrawal (Withdrawal Seizure-Resistant; WSR). The WSP and WSR strains differ dramatically in their withdrawal response, indicating that most of the genes leading to severe withdrawal are fixed in the WSP mice bred for this response. Conversely, most of the genes leading to withdrawal insensitivity are fixed in the WSR line. In addition to demonstrating a genetic component to dependence and withdrawal, differences between the WSP and WSR mice on other phenotypes suggest that alcohol withdrawal genes are also responsible for other responses. WSP mice have more severe withdrawal than WSR mice to diazepam (Valium), barbiturates, and nitrous oxide. These results suggest that similar brain substrates are mediating withdrawal severity to alcohol as well as a number of other drugs, and that some alcohol withdrawal risk-promoting genes also confer susceptibility to other drugs of abuse.

In addition to demonstrating genetic contributions to behavioral responses to alcohol, selected lines are useful for identifying differences in neural mechanisms mediating the response to alcohol. Long-Sleep (LS) and Short-Sleep (SS) mice have been selectively bred based on duration of loss of righting reflex (a measure of the sedative effects of alcohol). The LS and SS mice also differ in their response to other depressant drugs, again indicating similar brain substrates are mediating sedative responses to alcohol and other depressants. Administering drugs that activate or block activity of $GABA_A$ receptors affects alcohol sensitivity, and LS mice are more sensitive to these manipulations than SS mice. These findings indicate that selecting LS and SS mice for their behavioral response to alcohol has produced lines that differ in $GABA_A$ receptor activity.

Studies using selected lines have provided a wealth of information regarding how genes affect behaviors, and which alcohol responses share common genetic influence. However, despite the demonstration of genetic influences using this technique, it is difficult to identify specific genes mediating sensitivity or resistance to an effect of alcohol.

2.2 Quantitative Trait Loci (QTL) mapping strategies

The Human Genome Project has also led to genome mapping and DNA sequencing in a variety of other organisms including the laboratory mouse. Late twentieth-century developments in the physical mapping of the mouse make positional cloning of genes involved in various behaviors more likely. However, most behaviors (including responses to alcohol) are influenced by multiple genes. Behaviors, or complex traits, influenced by a number of genes are often termed quantitative traits. Within a population, a quantitative trait is not all-or-none, but differs in the degree to which individuals possess it. A section of DNA thought to harbor a gene that contributes to a quantitative trait is termed a quantitative trait locus (QTL). QTL mapping identifies the regions of the genome that contain genes affecting the quantitative trait, such as an alcohol response. Once a QTL has been located, the gene can eventually be isolated and its function studied in more detail. Thus, QTL analysis provides a means of locating and measuring the effects of a single gene on alcohol sensitivity.

In tests of sensitivity to convulsions following alcohol withdrawal, QTLs have been found on mouse chromosomes 1, 2, and 11. The QTL on chromosome 11 is near a cluster of $GABA_A$ receptor subunit genes. A number of subunits are needed to make a $GABA_A$ receptor, and the ability of a drug to act on the receptor seems to be subunit dependent. A polymorphism in the protein-coding sequence for *Gabrg2* (coding for the $\gamma 2$ subunit of the $GABA_A$ receptor) has been identified. This polymorphism is genetically correlated with duration of loss of righting reflex and a measure of motor incoordination following alcohol administration.

The use of QTL analysis has allowed us to begin the process of identifying the specific genes involved in alcohol related traits. Because each QTL initially includes dozens of genes, not all of which have yet been identified, it will require much more work before each QTL can be reduced to a single responsible gene. For the time being, one important aspect of QTL mapping in mice is that identification of a QTL in mice points directly to a specific location on a human chromosome in about 80 percent of cases. Thus, the animal mapping work can be directly linked to the human work in studies such as the COGA described in Sect. 1.1, which is in essence a human QTL mapping project. By using transgenic animal models (mice in

which there has been a deliberate modification of the genome), such as null mutants, QTLs can be further investigated.

2.3 Null Mutant and Transgenic Studies

Since about 1980, advances in embryology and genetic engineering have resulted in the creation of null mutant animals—mice that have a targeted deletion or over-expression of one of their own genes (and as a consequence, gene products). The use of null mutant mice is powerful in that it allows investigation into the role of the deleted gene by comparing the phenotype of the null mutant mice with normal mice. Until the development of this technology, the only way of studying the regulation and function of mammalian genes was through the observation of inherited characteristics, genetic defects, or spontaneous mutations, or through indirect manipulations such as selective breeding. There has been a rapid development in the use of null mutant mice in biological sciences. Since 1990, the number of procedures performed on transgenic animals in the United Kingdom, for example, has risen to more than 447,000.

Transgenic and null mutant mice are particularly useful in studying alcohol responses. As mentioned previously, there is evidence to suggest that alcohol affects the function of the $GABA_A$ receptor, although it is not clear how alcohol does this. By creating mice lacking a gene thought to mediate GABA function, it is possible to gather information regarding the function of a given neurotransmitter system, or function of a receptor. The γ isoform of the second messenger protein kinase C (PKCγ) has been implicated in one mechanism by which alcohol affects GABA function. Mice lacking the gene coding for this isoform are less sensitive to alcohol-induced loss of righting reflex and alcohol-induced hypothermia, suggesting that some biochemical processes in which PKCγ plays a role is a potential mechanism mediating responses to alcohol.

Null mutants have been created for a number of other neurotransmitter systems hypothesized to mediate responses to alcohol, including the dopamine D2 receptor. Mice lacking the D2 receptor gene consumed less alcohol in a free-choice situation, were insensitive to alcohol's locomotor depressant effects, and were less sensitive to the motor in coordinating effects of alcohol when compared to control mice. Differences between null mutants and control mice strongly support a role for this receptor in mediating several responses to alcohol.

Although studies using null mutant mice have provided much information regarding the role of specific genes in alcohol responses, alcohol responses are probably determined by the interaction of several genes. Thus, deletion of one gene will not provide information regarding the interaction of the deleted gene with other genes. If multiple genes are affecting a behavior, the elimination of one gene in an embryo will result in developmental compensation by other genes involved. The use of transgenics that have an inserted sequence allowing the experimenter to alter gene expression (commonly by using an antibiotic treatment) at any point in time allows control over when the gene is deleted. This permits the investigator to produce a null mutant, referred to as a conditionally regulated transgenic, after development has occurred, reducing developmental compensation. Genetic background may also influence results, because introduced genes may not exert similar effects when expressed on different genetic backgrounds.

3. Gene–Gene and Gene–Environment Effects on Alcohol Abuse

Gene–gene and gene–environment interactions are clearly prominent in alcoholism, although they are not often addressed. An important consideration is that alcoholism is multigenic, and might be polygenic (i.e., each individual gene might exert only a small effect on risk). However, a gene-by-gene analysis may not give a complete picture of how genes interact to mediate alcoholism. It is also clear that insights from the methods reviewed above are not sufficient to integrate a comprehensive understanding of alcoholism in the intact organism with environmental variables considered.

3.1 Epistasis (Gene–Gene Interaction)

Epistasis refers to the behavioral effect of interaction among gene alleles at multiple locations. Epistasis is observable when phenotypic differences among individuals with the same genotype at one locus depend on their genotypes at another locus. If adding the effects of each gene separately does not predict the effect of two genes, epistasis is most likely present. For example, assume that there are two genes leading to increased weight ('A' and 'B'). Each on its own induces a 1-pound increase in body weight. If an individual possessing both genes gained 2 pounds, this would imply a normal additive model of inheritance (no epistasis). If, however, an individual possessing both genes showed a 10-pound weight gain (or even weight loss) this would imply epistasis (example modified from Frankel and Schork 1996). As this example demonstrates, the effect of a gene may be detectable only within a setting that incorporates knowledge of epistatic interactions with other genes. A gene's effect may be undetectable because an interacting gene has an opposite effect, or because the gene's apparent effect is potentiated by another gene. Epistasis has been largely ignored in genome scans, but it is becoming increasingly evident that in order to understand genetic contributions to a complex trait fully, it

is an important consideration. Epistasis has been shown to be important in mouse models of epilepsy, and it is also likely to be critical to a thorough understanding of targeted gene deletion experiments.

3.2 Gene–Environmental Interaction

Several genes influencing behavioral traits in animals that may be homologous to aspects of human alcoholism are close to being isolated. In addition, human studies are identifying genes that might also be playing a role in alcoholism. While these experiments studied the important contributions of either genes, or experience, on the effects of drugs of abuse, few have provided any information on the important interaction between genes and environment. Gene–environment interaction simply refers to effects of environment that vary for different genotypes (or, effects of genes that vary for different environments). For example, take the adoption studies discussed above in Sect. 1. Type II alcoholism was highly heritable from father to son, regardless of environmental background (e.g., economic status of the adopted home). The risk for type I alcoholism, however, increased dramatically for individuals having both type I biological parents and low socioeconomic status in their adoptive home. Thus, the effect of environment can depend on the genotype, and vice versa.

A recent study in animals also illuminates the profound effect environment can have on the behavior of genetically identical animals (Crabbe et al. 1999). In this study, six commonly used mouse behaviors were simultaneously tested in three laboratories (two in the United States and one in Canada) using exactly the same genotypes. Stringent attempts were made to equate test apparatus, protocols, husbandry, age, and start time (of light cycle as well as time of year). Despite these rigorous controls, animals with the same genes performed differently on several of the behavioral tasks. It is important to note, however, that for a number of other behavioral tests, performance was very similar among the three sites. That genetically identical animals did not always respond the same depending on environment underscores the idea that, for behaviors like alcoholism, genes will define risk, not destiny.

Although the issue of gene–environment interaction seems a bit daunting, there are certainly ways of addressing the problem once researchers are aware of it. Testing animals using a battery of related tests is one way to approach the problem. For example, perhaps one test of memory relies heavily on locomotor performance. A genotypic difference in locomotion may be interpreted (incorrectly) as differences in memory. By testing genotypes on a battery of tests assessing memory, a clearer profile will emerge. Second, it is possible to evaluate genetic effects at different developmental stages, which would also provide information as to the ontological profile of a given behavior. Third, use of multiple genetic tools can help elucidate the generability of a genetic effect.

The issue of gene–environment interaction will be particularly relevant given the multiple forms of alcoholism, mediated by different sets of genes. Some forms are accompanied by antisocial personality traits, and others may be accompanied by depression. Dissociating the various genetic and environmental comorbidities affecting one's likelihood of being diagnosed with alcoholism will be a formidable task. It is clear that dissecting this disorder will require techniques that examine more than the effects of one gene at a time.

4. Conclusions

We now have some good clues from animal and human studies regarding specific genes involved in the response to alcohol. Late twentieth-century advances in QTL mapping strategies and the application of new molecular targeting techniques in genetic animal models are especially promising, because they suggest that a combined use of molecular biological techniques and animal behavioral genetic tools is beginning to occur. Understanding basic genetic mechanisms provides a context in which to interpret the findings using more recently developed molecular techniques. One important consideration is that while alcoholism may be partially mediated by genetic factors, the role of environment remains important. For example, environmental intervention including abstinence (before or after initial alcohol abuse) can prevent the expression of alcoholism. An individual at genetic risk may never develop alcoholism, for reasons not known. Genetic studies most likely identify an individual's lifetime risk for alcoholism, although at any given point in time an individual correctly identified as an alcoholic may not be manifesting symptoms of alcohol abuse (one reason why Jellinek's system was not useful longitudinally).

Ultimately, animal models further an understanding of human alcoholism. Thus, a comparison of animal and human results is an area of future importance. One of the goals of pharmacogenomic research is to develop agents to reduce drinking in alcoholics. As with other complex diseases, it is unlikely that one medication will be sufficient to treat a genetic disorder as complex as alcoholism. Understanding gene–gene and gene–environment interactions in conjunction with the action of alcohol in the central nervous system should provide new targets for the development of therapeutic agents.

See also: Alcohol-related Disorders; Alcohol Use Among Young People; Alcoholics Anonymous; Behavioral Genetics: Psychological Perspectives; Cultural Evolution: Theory and Models; Genetic Studies of Behavior: Methodology

Bibliography

Cloninger C R, Sigvardsson S, Bohman M 1996 Type I and type II alcoholism: An update. *Alcohol Health and Research World* **20**: 18–23

Crabbe J C, Wahlsten D, Dudek B C 1999 Genetics of mouse behavior: Interactions with laboratory environment. *Science* **284**: 1670–1

Feldman R S, Meyer J S, Quenzer L F 1997 *Principles of Neuropsychopharmacology*. Sinauer Associates

Ferguson R A, Goldberg D M 1997 Genetic markers of alcohol abuse. *Clinica Chimica Acta* **257**: 199–250

Frankel W N, Schork N J 1996 Who's afraid of epistasis? *Nature Genetics* **14**: 371–3

Goate A M, Edenberg J H 1998 The genetics of alcoholism. *Current Opinion in Genetics and Development* B: 282–6

Long J C, Knowler W C, Hanson R L, Robin R W, Urbanek M, Moore E, Bennett P H, Goldman D 1998 Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an American Indian population. *American Journal of Medical Genetics* **81**: 216–21

Morey L C 1996 Patient placement criteria: Linking typologies to managed care. *Alcohol Health and Research World* **20**: 36–44

Vaillant G E 1983 *The Natural History of Alcoholism*. Harvard University Press, Cambridge, MA

K. E. Browman and J. C. Crabbe

Algorithmic Complexity

1. Introduction

In the mid 1960s, in the early stage of computer science but with the general theory of Turing machines (Turing 1936) well understood, scientists needed to measure computation and information quantitatively. Kolmogorov complexity was invented by R. J. Solomonoff (1964), A. N. Kolmogorov (1965), and G. J. Chaitin (1969), independently and in this chronological order. This theory is now widely accepted as the standard approach that settled a half-century debate about the notion of randomness of an individual object—as opposed to the better understood notion of a random variable with intuitively both 'random' and 'nonrandom' individual outcomes. Kolmogorov complexity has a plethora of applications in many areas including computer science, mathematics, physics, biology, and social sciences (Li and Vitányi 1993). This article only describes some basic ideas and some appropriate sample applications.

Intuitively, the amount of information in a finite string is the size (number of bits) of the smallest program that, started with a blank memory, computes the string and then terminates. A similar definition can be given for infinite strings, but in this case the program produces element after element forever. Thus, 1^n (a string of n ones) contains little information because a program of size about $\log n$ outputs it. Likewise, the transcendental number $\pi = 3.1415\ldots$, an infinite sequence of seemingly 'random' decimal digits, contains a constant amount ($O(1)$) of information. (There is a short program that produces the consecutive digits of π forever.) Such a definition would appear to make the amount of information in an object depend on the particular programming language used. This is the case. Fortunately it can be shown that all choices of universal programming languages (such as PASCAL, C^{++}, Java, or LISP in which we can in principle program every task that can intuitively be programmed at all) lead to quantification of the amount of information that is invariant up to an additive constant. Formally, it is best formulated in terms of 'universal Turing machines,' the celebrated rigorous formulation of 'computability' by A. M. Turing (1936) that started both the theory and practice of computation.

This theory is different from Shannon information theory that deals with the expected information in a message from a probabilistic ensemble of possible messages. Kolmogorov complexity, on the other hand, measures the information in an individual string or message. The *randomness deficiency* of a binary string n bits long is the number of bits by which the complexity falls short of n—the maximum complexity—and a string is the more random the closer the complexity is to its length.

2. Theory

The Kolmogorov complexity $C(x)$ of a string x is the length of the shortest binary program (for a fixed reference universal programming language) that prints x as its only output and then halts. A string x is incompressible if $C(x)$ is at least the length $|x|$ (number of bits) of x: the shortest way to describe x is to give it literally. Similarly, a string x is 'nearly' incompressible if $C(x)$ is 'almost as large as' $|x|$.

The appropriate standard for 'almost as large' above can depend on the context, a typical choice being $C(x) \geqslant |x| - O(|\log x|)$. Similarly, the *conditional* Kolmogorov complexity of x with respect to y, denoted by $C(x|y)$, is the length of the shortest binary program that, with extra information y, prints x. And a string x is incompressible relative to y if $C(x|y)$ is large in the appropriate sense.

Intuitively, we think of such patternless sequences as being random, and we use the term 'random sequence' synonymously with 'incompressible sequence.' This is not just a matter of naming but on the contrary embodies the resolution of the fundamental question about the existence and characterization of random individual objects (strings). Following a half-century of unsuccessful approaches and acrimonious

scientific debates, in 1965 the Swedish mathematician Per Martin-Löf resolved the matter and gave a rigorous formalization of the intuitive notion of a random sequence as a sequence that passes all effective tests for randomness. He gave a similar formulation for infinite random sequences. The set of infinite random sequences has measure 1 in the set of all sequences. Martin-Löf's formulation uses constructive measure theory and has equivalent formulations in terms of being incompressible. Every Martin-Löf random sequence is *universally* random in the sense that it individually possesses all effectively testable randomness properties. (One can compare this with the notion of intuitive computability that is precisely captured by the notion of 'computable by Turing machines,' and every Turing machine computation can be performed by a universal Turing machine.) Many applications depend on the following easy facts.

Lemma 1 *Let c be a positive integer. For every fixed y, every finite set A contains at least* $(1-2^{-c})|A|+1$ *elements x with* $C(x|A, y) \geqslant [\log|A|]-c$. (*Choosing A to be the set of all strings of length n we have* $C(x/n,y) \geqslant n-c$)

Lemma 2 *Let A be a finite set. For every y, every element* $x \in A$ *has complexity* $C(x|A, y) \leqslant \log|A|+c$. (*Choosing A to be the set of all strings of length n we have* $C(x/n,y) \leqslant n+c$)

The first lemma is proved by simple counting. The second lemma holds since a fixed program that enumerates the given finite set computes x from its index in the enumeration order—and this index has $\log|A|$ bits for a set A of cardinality $|A|$.

We can now compare Kolmogorov complexity with Shannon's statistical notion of entropy—the minimal *expected* code word length of messages from random source using the most parsimonious code possible. Surprisingly, many laws that hold for Shannon entropy (that is, on average) still hold for the Kolmogorov complexity of individual strings albeit only within a logarithmic additive term. Denote by $C(x|y)$ the information in x given y (the length of the shortest program that computes x from y), and denote by $C(x, y)$ the length of the shortest program that computes the pair x, y. Here is the (deep and powerful) Kolmogorov complexity version of the classical 'symmetry of information' law. Up to an additive logarithmic term,

$$C(x, y) = C(x)+C(y|x) = C(y)+C(x|y) \qquad (1)$$

We can interpret $C(x)-C(x|y)$ as the information y has about x. It follows from the above that the amount of information y has about x is almost the same as the amount of information x has about y: information is symmetric. This is called mutual information.

Kolmogorov complexity is a wonderful measure of randomness. However, it is not computable, which obviously impedes some forms of practical use. Nevertheless, noncomputability is not really an obstacle for the wide range of applications of Kolmogorov complexity, just like noncomputability of almost all real numbers does not impede their practical ubiquitous use.

3. *Applications*

For numerous applications in computer science, combinatorics, mathematics, learning theory, philosophy, biology, and physics, see Li and Vitányi (1993). For illustrative applications in cognitive psychology see Chater (1996) (related, more informal strains of thought are the Structural Information Theory started in Leeuwenberg (1969)), in economy see Keuzenkamp and McAleer (1995), and in model selection and prediction see Vitányi and Li (2000). Here we give three applications of Kolmogorov complexity related to social sciences, explain the novel 'incompressibility method,' and conclude with an elementary proof of Gödel's celebrated result that mathematics is undecidable.

3.1 *Cognitive Distance*

For a function $f(x, y)$ to be a proper *distance measure* we want it to be a *metric*: it has non-negative real values; it is symmetrical, $f(x, y) = f(y, x)$; it satisfies the triangle inequality, $f(x, y) \leqslant f(x, z)+f(z, y)$; and $f(x, y) = 0$ iff $x = y$. Given two objects, say two pictures, how do we define an objective measure that would define their distance that is universal in the sense that it accounts for all cognitive similarities? Traditional distances do not work. For example, given a picture and its negative (i.e., exchange 0 and 1 in each pixel), Hamming distance and Euclidean distance both fail to recognize their similarity. Let us define a new distance $D(x, y)$ between two objects x and y as the length of the shortest program that converts them back and forth (Bennett et al. 1998). It turns out that, up to a logarithmic additive term,

$$D(x, y) = \max\{C(x|y), C(y|x)\} \qquad (2)$$

This distance D is a proper metric and it is *universal* in the sense that if two objects are 'close' under any distance out of a wide class of sensible and computable metrics, then they are also 'close' under d. For example, the $D(x, y)$ distance between two black-and-white pictures x and its negative y is a small constant.

3.2 *Phylogeny of Chain Letters* (*and Biological Evolution*)

Chain letters are an interesting social phenomenon that have reached billions of people. Such letters

evolve, much like biological species (rather, their genomes). Given a set of genomes we want to determine the evolutionary history (phylogeny tree). Can we use the information distance $D(x, y)$? But then the difference in length of (especially complex) genomes implies a large distance while evolutionarily the genomes concerned can be very close (part of the genome was simply erased). We can divide $D(x, y)$ by some combination of the lengths of x and y, but this can be shown to be improper as well. As we have seen, $C(y)-C(y|x) = C(x)-C(x|y)$ within a logarithmic additive constant: it is the mutual information between x and y. But mutual information itself does not satisfy the triangle inequality and hence is not a metric and therefore clearly cannot be used to determine phylogeny. The solution is to determine closeness between each pair of genomes (or pairs of chain letters) x and y by taking the ratio of the information distance to the maximal complexity of the two:

$$d(x, y) = \frac{D(x, y)}{\max\{C(x), C(y)\}} \tag{3}$$

Note that $d(x, y)$ is always a sort of normalized dissimilarity coefficient that is at most 1. Moreover, it is proper metric. Let us look a little bit closer: suppose $C(y) \geqslant C(x)$. Then, up to logarithmic additive terms in both nominator and denominator we find (using Eqn. (2))

$$d(x, y) = \frac{C(y|x)}{C(y)} = 1 - \frac{C(y)-C(y|x)}{C(y)} \tag{4}$$

It turns out that $d(x, y)$ is universal (always gives the smallest distance) in a wide class of sensible and computable normalized dissimilarity coefficient metrics. It measures the percentage of shared information, which is a convenient way to measure English text or DNA sequence similarity. We have actually applied this measure (or rather, a less perfect close relative) to English texts. Using a compression program called *GenCompress* we heuristically approximate $C(x)$ and $C(x|y)$. With the caveat 'heuristic,' that is, without mathematical closeness-of-approximation guarantees, C. H. Bennett, M. Li and B. Ma (in an article to appear in *Scientific American*) took 33 chain letters—collected by Charles Bennett from 1980 to 1997—and approximated their pairwise distance $d(x, y)$. Then, we used standard phylogeny building programs from bioinformatics research to construct a tree of these chain letters. The resulting tree gives a perfect phylogeny for all notable features, in the sense that each notable feature is grouped together in the tree (so that the tree is parsimonious). This fundamental notion can be applied in many different areas. One of these concerns a major challenge in bioinformatics: to find good methods to compare genomes. Traditional approaches of computing the phylogeny use so-called 'multiple alignment.' They would not work here since chain letters contain swapped sentences and genomes contain translocated genes and noncoding regions. Using the chain letter method, a more serious application in Li et al. (2001) automatically builds correct phylogenies from complete mitochondrial genomes of mammals. We confirmed a biological conjecture that ferungulates—placental mammals that are not primates, including cats, cows, horses, whales—are closer to the primates—monkeys, humans—than to rodents.

3.3 Inductive Reasoning

Solomonoff (1964) argues that all inference problems can be cast in the form of extrapolation from an ordered sequence of binary symbols. A principle to enable us to extrapolate from an initial segment of a sequence to its continuation will either require some hypothesis about the source of the sequence or another method to do the extrapolation. Two popular and useful metaphysical principles for extrapolation are those of simplicity (Occam's razor, attributed to the thirteenth-century scholastic philosopher William of Ockham, but emphasized about 20 years before Ockham by John Duns Scotus), and indifference. The Principle of Simplicity asserts that the 'simplest' explanation is the most reliable. The Principle of Indifference asserts that in the absence of grounds enabling us to choose between explanations we should treat them as equally reliable. Roughly, the idea is to define the universal probability, $\mathbf{M}(x)$, as the probability that a program in a fixed universal programming language outputs a sequence starting with x when its input is supplied by tosses of a fair coin (see Kirchherr et al. 1997). Using this as a sort of 'universal prior probability' we then can formally do the extrapolation by Bayes's Rule. The probability that x will be followed by a 1 rather than by a 0 turns out to be

$$\frac{\mathbf{M}(x1)}{\mathbf{M}(x0)+\mathbf{M}(x1)}$$

It can be shown that $-\log\mathbf{M}(x) = C(x)$ up to an additive logarithmic term, which establishes that the distribution $\mathbf{M}(x)$ is a mathematical version of Occam's razor: low complexity xs have high probability ($x = 11\ldots1$ of every length n has complexity $C(x) \leqslant \log n + O(1)$ and hence universal probability $\mathbf{M}(x) \geqslant 1/n^c$ for some fixed constant c), and high complexity ys have low probability (if y is the outcome of n flips of a fair coin then for example with probability 0.9999 we have $C(y) \geqslant n-10$ and therefore $\mathbf{M}(x) \leqslant 1/2^{n-10}$). This theory was further developed in Li and Vitányi (1993), Kirchherr et al. (1997) and Vitányi and Li (2000), and relates to more

informal cognitive psychology work starting with Leeuwenberg (1969) and the applied statistical 'minimum description length (MDL)' model selection and prediction methods surveyed in Barron et al. (1998).

3.4 Incompressibility Method

Analyzing the performance of computer programs is very difficult. Analyzing the average case performance of computer programs is often more difficult since one has to consider all possible inputs and take the average. However, if we could find a typical input on which the program takes an average amount of time, then all we need to do is to find the performance of the computer program on this particular input. Then the analysis is easy. A Kolmogorov random input does exactly that: providing a typical input. Using this method, we were able to solve many otherwise difficult problems. Recent examples are average case analysis of Shellsort algorithm (Jiang et al. in press) and the average case of Heilbronn's triangle problem. A popular account about how to analyze the average-case bounds on Heilbronn's triangle problem can be found in Mackenzie (1999).

3.5 Gödel's Incompleteness Result

A new elementary proof by Kolmogorov complexity of K. Gödel's famous result showing the incompleteness of mathematics (not everything that is true can be proven) is due to Ya. Barzdin's and was later popularized by G. Chaitin, see Li and Vitányi (1993). A formal system (consisting of definitions, axioms, rules of inference) is consistent if no statement that can be expressed in the system can be proved to be both true and false in the system. A formal system is sound if only true statements can be proved to be true in the system. (Hence, a sound formal system is consistent.)

Let x be a finite binary string. We write 'x is random' if the shortest binary description of x with respect to the optimal specification method D_0 has length at least $|x|$. A simple counting argument shows that there are random xs of each length.

Fix any sound formal system F in which we can express statements like 'x is random.' Suppose F can be described in f bits—assume, for example, that this is the number of bits used in the exhaustive description of F in the first chapter of the textbook *Foundations of F*. We claim that for all but finitely many random strings x, the sentence 'x is random' is not provable in F. Assume the contrary. Then given F, we can start to exhaustively search for a proof that some string of length $n \gg f$ is random, and print it when we find such a string x. This procedure to print x of length n uses only $\log n + f$ bits of data, which is much less than n. But x is random by the proof and the fact that F is sound. Hence, F is not consistent, which is a contradiction. This shows that although most strings are random, it is impossible to effectively prove them random. In a way, this explains why the incompressibility method above is so successful. We can argue about a 'typical' individual element, which is difficult or impossible by other methods.

See also: Algorithms; Computational Approaches to Model Evaluation; High Performance Computing; Information Processing Architectures: Fundamental Issues; Information Theory; Mathematical Psychology; Model Testing and Selection, Theory of

Bibliography

Barron A R, Rissanen J, Yu B 1998 The minimum description length principle in coding and modelling. *IEEE Trans. Inform. Theory* **44**(6): 2743–60

Bennett C H, Gács P, Li M, Vitányi P, Zurek W 1998 Information distance. *IEEE Trans. Inform. Theory* **44**(4): 1407–23

Bennett C H, Li M, Ma B, in press Linking chain letters, *Scientific American*

Chaitin G J 1969 On the length of programs for computing finite binary sequences: Statistical considerations. *J. Assoc. Comput. Mach.* **16**: 145–59

Chater N 1996 Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review* **103**: 566–81

Chen X, Kwong S, Li M 1999 A compression algorithm for DNA sequences and its application in genome comparison. In GIW'99, Tokyo, Japan, Dec. 1999, and in *RECOMB'00*. Tokyo, Japan, April 2000

Jiang T, Li M, Vitányi P in press A lower bound on the average-case complexity of Shellsort. *J. Assoc. Comp. Machin.* **47**: 905–11

Keuzenkamp H A, McAleer M 1995 Simplicity, scientific inference and econometric modelling. *The Economic Journal* **105**: 1–21

Kirchherr W W, Li M, Vitányi P M B 1997 The miraculous universal distribution. *Mathematical Intelligencer* **19**(4): 7–15

Kolmogorov A N 1965 Three approaches to the quantitative definition of information. *Problems Information Transmission* **1**(1): 1–7

Koonin E V 1999 The emerging paradigm and open problems in comparative genomics. *Bioinformatics* **15**: 265–6

Leeuwenberg Q J 1969 Quantitative specification of information in sequential patterns. *Psychological Review* **76**: 216–20

Li M, Badger J, Chen X, Kwong S, Kearney P, Zhang H 2001 An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* **17**(2): 149–54

Li M, Vitányi P 1993 *An Introduction to Kolmogorov Complexity and its Applications*, 1st edn. Springer Verlag, New York

Mackenzie D 1999 On a roll. *New Scientist* **164**: 44–7

Solomonoff R J 1964 A formal theory of inductive inference, Part 1 and Part 2. *Information Control* **7**: 1–22, 224–54

Turing A M 1936 On computable numbers with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* **2**(42): 230–65

Vitányi P M B, Li M 2000 Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Trans. Inform. Theory* **46**

Wooley J C 1999 Trends in computational biology: A summary based on a RECOMB plenary lecture, 1999. *Journal of Computational Biology* **6**(314): 459–74

M. Li and P. Vitányi

Algorithms

1. *Introduction*

Around the year 825 the Persian mathematician Abu Ja'far Mohammed ibn Mûsâ al-Khowârizm wrote a textbook entitled *Kitab al jabr w'al muqabala*. The term 'algorithm' is directly derived from the last part of the author's name. An algorithm is a mathematical recipe, formulated as a finite set of rules to be performed systematically, that has as outcome the solution to a well-formulated problem. In sequential algorithms these steps are ordered and should be performed one after the other. In parallel algorithms some of the rules are to be performed simultaneously. Algorithms can be graphically represented by flowcharts composed by arrows and boxes. The boxes contain the instructions and the arrows indicate transitions from one step to the next.

Algorithms were known much earlier than the eighth century. One of the most familiar, dating from ancient Greek times (c. 300 BC), is the procedure now referred to as Euclid's algorithm for finding the highest common factor of two natural numbers.

Algorithms often depend on subalgorithms or subroutines. For instance, the algorithm for obtaining the highest common factor of two numbers relies on the algorithm for finding the remainder of division for two natural numbers a and b. Dividing two numbers a and b is something we learn at school after having learnt the multiplication tables by heart. Usually we perform a division by reducing it to a sequence of multiplications and subtractions. Yet assume for the moment that to perform division we cannot rely on multiplication, nor can we express numbers in base ten. Both of these operations will require additional algorithms. We represent numbers in a very primitive form by simply writing a sequence of dots. Thus ●●●●●, for instance, represents the number 5.

The remainder algorithm works as follows (using $17 \div 5$ as an example):

(a) Write the sequence of dots for the dividend (17 in this case).
(b) Erase as many dots as correspond to the divisor (5 in this case).
(c) If the remaining number of dots is larger than or equal to the divisor go back to (a)
(d) If the remaining number of dots is smaller than the dividend print out this number
(e) STOP

The algorithm just described contains a loop. Observe that for any pair of numbers the algorithm produces the answer in a finite number of steps. Applied to our example of $17 \div 5$, the algorithm performs the following steps:

Current state	Operation
START	Write 17 dots
●●●●●●●●●●●●●●●●●	Erase 5 dots
●●●●●●●●●●●●	Erase 5 dots
●●●●●●●	Erase 5 dots
●●	Print 2 dots
STOP	

The computational algorithm for finding the remainder of a number when divided by another number can be used as a subroutine of the decision algorithm for the decidable problem 'Does b divide a?' (the answer is 'yes' if the remainder is zero). Repeated application of these algorithms produces the answer to the decidable question 'Is a a prime?' (the answer is 'no' if a is divisible by any smaller natural number besides 1).

An algorithm or machine is deterministic if at each step there is only one possible action it can perform. A nondeterministic algorithm or machine may make random choices of its next action at some steps. An algorithm is called a decision algorithm if it leads to a 'yes' or a 'no' result, whereas it is called a computational algorithm if it computes a solution to a given well-defined problem.

Despite the ancient origins of specific examples of algorithms, the precise formulation of the concept of a general algorithm dates only from the last century. The first rigorous definitions of this concept arose in the 1930s. The classical prototype algorithm is the Turing machine, defined by Alan Turing to tackle the *Entscheidungsproblem* or Decision Problem, posed by the German mathematician David Hilbert in 1900, at the Paris International Congress of Mathematicians. Hilbert's dream was to prove that the edifice of mathematics is a consistent set of propositions derived from a finite set of axioms, from which the truth of any well-formulated proposition can be established by a well-defined finite sequence of proof steps. The development and formalization of mathematics had led mathematicians to see it as the perfect, flawless science.

2. *Algorithms and the* Entscheidungsproblem

In 1931 the foundation of mathematics suffered its most crushing blow from a startling theorem proven by the Austrian logician Kurt Gödel. Gödel showed that any mathematical system powerful enough to represent arithmetic is incomplete in the sense that there must exist propositions that cannot be proven true or untrue in a finite sequence of steps. Such propositions are said to be undecidable within the given system. Turing had been motivated by Gödel's work to seek an algorithmic method of determining

whether any given proposition was undecidable, with the ultimate goal of removing undecidability as a concern for mathematics. Instead, he proved in his seminal paper 'On computable numbers, with an application to the *Entscheidungsproblem*' (1937) that there cannot exist any such universal method of determination and, hence, that mathematics will always contain undecidable propositions.

The question of establishing whether the number of steps required for a given problem is finite or infinite is called the halting problem. Turing's description of the essential features of any general-purpose algorithm, or Turing machine, became the foundation of computer science. Today the issues of decidability and computability are central to the design of a computer program—a special type of algorithm—and are investigated in theoretical computer science.

The question whether intelligent problem-solving can be described in terms of algorithms was extensively examined by Herbert Simon in the late 1940s and early 1950s. Newell and Simon proposed the first computer programs for problem-solving algorithms as well as the first programs for algorithms that prove theorems in Euclidean geometry, thus founding the new discipline of Artificial Intelligence (see *Artificial Intelligence: Genetic Programming*).

Around the same time McCulloch had developed a formal model of a neuron (McCulloch and Pitts 1943), proving that artificial neurons are capable of performing logical operations. An artificial neuron is a device that produces an output that is a function of its inputs if the sum of the inputs exceeds a threshold and otherwise produces no output. The sub-discipline of computer science known as Neural Networks deals with systems of artificial neurons firing in sequence and/or in parallel, in analogy to the operation of biological neurons.

3. *The Complexity of an Algorithm*

One important feature of an algorithm is its complexity. A number of definitions of complexity have been put forward, the most common of them being time complexity, or the length of time it takes an algorithm to be executed. Clearly, algorithms with low time complexity are to be preferred to ones with higher time complexity that solve the same problem. The question of establishing a formal definition of complexity was answered and treated formally in theoretical computer science (see *Algorithmic Complexity*).

One possibility is to count the number of operational steps in an algorithm, express this number as a function of the number of free parameters involved in the algorithm and determine the order of complexity of this function. The order of complexity of a function f is denoted by O(f), where O( ) is usually called the Landau symbol, and is defined as follows: Given two functions F(n) and G(n) defined on the set of natural numbers, we say that F is of the order of G, and write F = O(G), if there exists a constant K such that:

$$\frac{\mathrm{F}(n)}{\mathrm{G}(n)} < \mathrm{K}$$

for all natural numbers n. Thus, for instance, the function $\mathrm{F}(n) = 3n+1$ is of the order of $\mathrm{G}(n) = n$. Every polynomial in n, that is, every linear combination of powers of n, is of the order of its highest power of n. Because what is being counted is the number of steps in a sequential process it is common to view the resulting O( ) criterion as the time complexity of the algorithm, where n denotes the length of the given input.

The notion of algorithm complexity led to a fundamental classification of problems. They are classified as belonging to P or to NP, where, as we shall see, this 'or' is not exclusive. P is a shorthand for 'polynomial time' and stands for the set of all problems that can be solved by a deterministic algorithm in polynomial time, which means that the number of operations required by the algorithm can be expressed in terms of a linear combination of powers of n, where n is the number of free parameters. NP is the class of all problems that can be solved by nondeterministic algorithms in polynomial time. The strength of nondeterminism lies precisely in the machine's freedom to search through a large space of possible computations by means of a nondeterministic or stochastic process. Clearly $\mathrm{P} \subseteq \mathrm{NP}$. One of the unsolved problems in computer science is to establish whether $\mathrm{P} = \mathrm{NP}$ (Garey and Johnson 1979). Problems for which low complexity algorithms exist (i.e., problems in class P) are called tractable; other problems are called intractable.

4. *Monte Carlo Methods*

An important category of nondeterministic algorithms is the class of Monte Carlo methods. The term derives from the gambling games that are an economic mainstay of the city of that name, and originated during World War II as a code name for stochastic simulations associated with nuclear research. The most common application of Monte Carlo methods is to approximate intractable integrals. If $X_1,\ldots, X_n$ are random draws from a probability distribution with density function $f(x)$, then $S_n = \frac{1}{n}\sum_i h(X_i)$ has expected value

$$E[S_n] = \int_x h(x)f(x)\,\mathrm{d}x$$

and satisfies a central limit theorem. Thus, a difficult integral can be approximated by representing it in the

form (1) and executing the following nondeterministic algorithm:

(a) Simulate a random number X_n from the distribution $f(x)$.
(b) Compute the average $S_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$.
(c) Estimate whether sufficient accuracy has been achieved.
(d) If yes, STOP; otherwise return to (a).

A common strategy for increasing accuracy in a problem that does not admit exact solutions is to decompose it into subproblems such that exact algorithms for tractable subproblems can be combined with Monte Carlo estimates for intractable subproblems to obtain accurate approximations to the whole problem. Monte Carlo simulations are important tools in fields such as physics, statistical analysis, stochastic neural networks, and optimization.

5. *Algorithms for Learning*

Since the inception of formal algorithms, an important question has dealt with the conditions under which algorithms can be designed to learn from experience. The idea is that an algorithm that learns from experience becomes more and more adequate for solving the problem in question. The last two decades of the twentieth century witnessed an explosion of algorithms that can perform learning tasks and a sound, theoretical understanding of learning machines is beginning to emerge. The theoretical and empirical study of learning algorithms is called Machine Learning. One type of learning is typical of neural networks that modify their weights by adapting them to feedback information on their performance. Another type of learning machine is provided by what is known as a 'genetic algorithm,' where there is a natural selection between algorithmic procedures and the most effective algorithm arises by a form of survival of the fittest.

Yet another type of learning can be materialized as a search through a large set of possible algorithms for a given problem (see *Algorithmic Complexity*). Here the type of algorithm is fixed from the beginning and what is searched for is the optimal realization of this algorithm, that is, a realization that fits the given data best. Thus learning becomes a special form of model selection. For many problems, including search across models, exact solutions are intractable, yet tractable heuristics (or rules-of-thumb) work well for a large class of problem instances. As an example, consider the problem of finding an optimal decision tree algorithm to represent a given decision rule. A decision tree is a graphical representation of a rule for making a categorization decision. The graph for a decision tree consists of nodes and arcs pointing from nodes (called parents) to other nodes (called children). One node, called the root node, has no parents. Each other node has exactly one parent. At the bottom of the tree are the leaf nodes, which have no children. With each nonleaf node is associated a rule which selects a child to visit depending on information passed from the parent node.

With each leaf node is associated a rule for computing a value for the node. The decision tree algorithm proceeds as follows:

(a) Begin at root node.
(b) Execute rule to decide which arc to traverse.
(c) Proceed to child at end of chosen arc.
(d) If child is a leaf node, execute rule to compute and print value associated with node and STOP.
(e) Otherwise, go to (b).

It is frequently desired to find an optimal decision tree for a given classification problem and a given criterion of optimality. Clearly there are exponentially many possible decision trees for a set of n pieces of information and thus, searching for optimal trees is often an intractable problem. Tractable heuristics have been suggested in machine learning that help find trees good enough for given problems. One such type of classification tree has been introduced by Breiman et al. (1993) and is known as CART, which is an abbreviation of classification and regression tree.

The study of algorithms has been highly influential in psychology. The behaviorist view, which emphasizes the relationship between sensory inputs and motor outputs, has given way to the cognitive view, which emphasizes the role of internal cognitive states in producing behavior. Information processing models of cognition have become a mainstay of cognitive psychology. Hypotheses about the manner in which cognition influences behavior are encoded as algorithms and implemented in computer software which is used to simulate behavior. Simulated results are then compared to the behavior of human or animal subjects to validate or refute the hypotheses upon which they are based. One of the important issues in cognitive psychology has been to establish how the mind deals with decisions under uncertainty. While some schools have defended the approach of classical rationality, for which the mind functions by means of probabilistic algorithms, a recent development has been to view the unaided mind as relying on simple heuristics for inference. These heuristics, as proposed by Gigerenzer et al. (1999), are actually elementary, fast, and robust algorithms. In some cases they are extremely simple classification trees (see *Heuristics for Decision and Choice*).

6. *Conclusion*

Algorithms are the core element of thinking machines. The last century witnessed the birth of highly intelligent algorithms that replicate, often with great accuracy, some of the important achievements of the human mind. Yet, one of the most important discoveries concerning algorithms is that mathematics cannot be produced by an algorithmic procedure. This discovery, which was experienced as a crisis, can be viewed as liberating because it demonstrates the limitations of a

machine. Replicating intellectual, symbolic, and even basic common-sense activities by means of algorithms turned out to be daunting tasks, still mostly exceeding the grasp of computable representations.

See also: Algorithmic Complexity; Artificial Intelligence in Cognitive Science; Artificial Intelligence: Search; Mathematical Psychology, History of

Bibliography

Breiman L, Friedman J H, Olshen R A, Stone C J 1993 *Classification and Regression Trees*. Chapman and Hall, New York

Garey M R, and Johnson D S 1979 *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco

Gigerenzer G, Todd P, and the ABC Group 1999 *Simple Heuristics that Make Us Smart*. Oxford University Press, New York

McCulloch W S, Pitts W H 1943 A logical calculus of the idea immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–33 [reprinted in McCulloch W S 1965 *Embodiments of Mind*. MIT Press, MA]

Turing A 1937 On computable numbers, with an application to the *Entscheidungsproblem*. *Proceedings of the London Mathematical Society* (Ser. **2**) **42**: 230–65; a correction, **43**: 544–6

L. Martignon

Alienation: Psychosociological Tradition

The concept of alienation has a long and distinguished history (see *Alienation, Sociology of*), marked at the same time by a certain elusiveness and controversy. Its elusiveness is symbolized by the fact that the concept languished with almost no attention by social scientists until its rediscovery in the 1930s when Marx's early philosophical manuscripts of 1844 became known (Marx 1964). Thus, what we take to be a classical concept in sociological analysis has, in fact, a relatively short scientific history (leaving aside its metaphysical pre-Hegelian origins).

1. Objective and Subjective Alienation

The seminal Marxian manuscripts also established the basis for much of the subsequent controversy. Marx's concept of alienation was a complex mixture of objective and subjective elements concerning the structure of social relations, estrangement from human nature, and depersonalization, especially in relation to work. The distinction between the worker's objectively defined exploitation and lack of control, on the one hand, and the worker's subjective sense of powerlessness and self-estrangement on the other, has been a constant source of debate in the alienation literature. For some, alienation is not so much a matter of the person's awareness of various deprivations, but the fundamental deprivation of awareness (i.e., false consciousness regarding one's objective domination in capitalist society).

The unclarities that inhere in these difficulties regarding alienation have led, often enough, to calls for dismissing the concept from the lexicon of analysis. It can be argued, however, that these difficulties with the concept are no greater than those which, upon reflection, obtain for other widely used concepts in psychology and sociology (e.g., 'norms,' 'attitudes,' and, indeed, the fundamental concept of 'social structure'). In the psychosocial approach, an effort has been made to provide the requisite clarity by distinguishing the several varieties of alienation that derive from the classical tradition—for example, Marx, Durkheim, Weber; Schacht (1970)—providing at the same time the basis for empirical investigation of the sources, concomitants, and consequences of alienation.

2. Dimensions of Alienation

Six dimensions of alienation have been identified (and defined below): (a) powerlessness, (b) meaninglessness, (c) normlessness, (d) social isolation, (e) cultural disengagement, and (f) self-estrangement. Scales to measure each of these have been developed (Seeman 1991), but it is important to recognize that corollary concepts (e.g., self-efficacy, mastery, reification, sense of coherence, human agency) abound in the field; thus the definitions embodied in particular scales vary widely, and the naming of measuring instruments can be the source of considerable confusion (e.g., a measure of 'reification' can readily parallel the content of scales measuring 'powerlessness' or 'self-estrangement').

Each of the varieties of alienation is defined from the actor's point of view, but it is assumed that (a) the objective structural circumstances that generate these alienations can and should be independently identified, and (b) the person's subjective report may not coincide with these objective circumstances, yielding in that case a kind of 'false consciousness.'

2.1 Powerlessness

The sense of powerlessness is the expectancy or perception that one's own behavior cannot control the occurrence of personal and social outcomes: control is vested in external forces, powerful others, luck, or fate—as in the Marxian depiction of the domination and exploitation of the worker in capitalist society.

2.2 Meaninglessness

Meaninglessness is the sense of incomprehensibility of social affairs, events whose dynamic one does not

understand and whose future course one cannot predict—as in the Weberian depiction of the complexities of secularized and rationalized bureaucratic society.

2.3 Normlessness

Normlessness is the expectancy or perception that socially unapproved means are necessary to achieve one's goals—essentially, the personalized counterpart of Durkheim's 'anomie' or breakdown of social norms, involving as well the deterioration of trust in social relations.

2.4 Social Isolation

Social isolation refers to the person's sense of exclusion or lack of social acceptance, expressed typically in feelings of loneliness or feelings of rejection or repudiation vs. belonging—as in the concern typified by Tönnies' depiction of the historical change from 'community' to 'society.'

2.5 Cultural Disengagement

Cultural disengagement refers to a different kind of separation—namely, the person's sense of removal or distance from the dominant values in the society—as in the standard depiction of the alienation of the intellectual and the avant-garde artist.

2.6 Self-estrangement

Self-estrangement is a complex and difficult version —some would say the overarching version—of alienation, embodied, for example, in the Marxian view of alienated labor as estrangement from one's creative human nature. The complexity here is suggested by the fact that there are at least three distinctive ways of conceiving of self-estrangement; in capsule form (a) the despised self (referring to negative self-esteem); (b) the disguised self (false consciousness, being as it were 'out of touch' with oneself); and (c) the detached self (engagement in activities that are not intrinsically rewarding, a derivation from Marx's emphasis on stultifying disengagement in work).

3. The Unity of Alienation

These versions of alienation do not present a theory of the phenomenon or reflect any implicit conviction about their unity (though, indeed, they may well be correlated under given circumstances). Various proposals have been made that attempt to establish the unity of the several alienations: for example, their unity is purported to lie in the fact that (a) they appear in a typical sequence; (b) they represent the fundamental components of social action (e.g., values, norms, roles, and situational facilities, Smelser 1963); (c) they exhibit a statistical coherence—that is, a generalized first factor, as it were; and (d) they express a core theme, representing various forms of fragmentation or separation in one's experience. The latter is clearly a thin reed of unity (though perhaps the most sustainable view) in which their commonality lies only in the fact that they represent classical ways of depicting the individual's sense of separation from important commonly held values: for example, powerlessness vs. mastery, normlessness vs. order and trust, social isolation vs. community.

One of the difficulties with analysis and empirical work employing the alienation concept has been its embarrassing versatility: it has too often been used to explain particular troubles and their opposites—for example, political passivity and urban riots, conformity and deviance. It needs to be recognized that the psychosocial concepts described above (or, for that matter, similar concepts such as 'relative deprivation') cannot be expected to explain very much in themselves without adequate reference to situational circumstances or to other relevant variables that need to be taken into account.

4. Empirical Studies of Alienation

In empirical terms, most of the research on alienation has focused on the ideas of powerlessness and social isolation (though not necessarily using these two concepts). In psychology, for example, an extensive literature on 'internal vs. external control' (using a 'locus of control' measure originating in Rotter 1966) has developed, exploring the ways in which the person's sense of being in personal control of events is socialized and expressed—for example, external control as a factor in deviant behavior, family planning, depression, and alcohol use. In sociology, parallel work on powerlessness and the sense of mastery has documented the impact of such alienation on a wide range of behavior, including associations between powerlessness and (a) inferior learning and achievement (e.g., the relatively poor academic performance of minority children); (b) low political engagement; (c) participation in civil disturbances; (d) unemployment; and (e) inferior health status (including disinterest in preventive health practices and mortality consequences).

One of the best documented of these studies is the work of Kohn and Schooler (1983) showing, on the one hand, the connection between the sense of powerlessness and job conditions that have the earmarks of Marxian alienated labor (i.e., work that is not creative or self-directed), and on the other hand, the psycho-

logical consequences of such work experience (e.g., diminished intellectual flexibility). Related epidemiological work on the connection between social class and health, has explored the association between low job control and health consequences such as heart disease and mortality rates.

A similarly extensive body of research bearing on social isolation and social support has developed in recent years. In a sense, the thrust of this work—the combined effort of sociologists, psychologists, and epidemiologists—has been to undermine the earlier image of urban life as a wasteland of atomized impersonal actors. It has been shown that strong interpersonal networks persist, and more important, that engagement in such social support networks has salutary effects over a wide range of life experience. Thus, for example, social ties have been demonstrated to be important for recovery from health crises (including surgery and cancer treatment), for managing response to unemployment, for overall health status (including mortality), and for sustained high performance in old age.

These remarks concerning two of the key forms of alienation illustrate the uses that have been made of the concept in empirical work. Space permitting, a similar case could be made for the other dimensions of alienation. Thus, the idea of meaninglessness has been used to explain the genesis of ethnic hostility (prejudice and discrimination as simplified answers to societal complexity); normlessness is viewed as an element in the development and rationalization of deviant behavior and response to mass persuasion; and self-estrangement (particularly in work) is used to explain a range of behavior from alcohol problems to family troubles and mass movements. Much of this, it must be said, is not well documented, however plausible the arguments may be.

5. *Physiological Concomitants of Alienation*

It is reasonable to ask whether a number of these cited consequences of alienation have a basis in physiological mediation. There are, indeed, good grounds for predicting such consequences (particularly, the wide-ranging health effects of powerlessness), since the expected biological correlates of low control have been documented—chiefly, associations between a low sense of control and high levels of damaging stress-related hormone excretion and poorer immune function (Sapolsky 1992). Furthermore, and not surprisingly, confirming evidence has developed regarding the expected physiological concomitants of social engagement—for example, better immune system performance, less hypertension, and less damaging levels of glucocorticoids which can affect memory processes are associated with greater social integration (Seeman and McEwen 1996).

6. *Conclusion: Future Prospects*

Considerable effort has been devoted to straightforward reporting on the demographics of alienation —that is, to the social location of high alienation. It seems reasonably clear, for example, that alienation is more clearly visible in less democratic societies, and among the working class and minorities. In addition, the secular trend for measures of powerlessness or mistrust has been toward increasing alienation in Western societies.

It may well be, however, that the general looseness of argument and the difficulty of documentation commented on above has led recently to a certain avoidance of the concept for social analysis. The idea of alienation was especially popular during the 1960s and 1970s—a period of societal difficulty involving student rebellions, antiestablishment movements, Vietnam protests, and the like. But the relative calm and prosperity since the 1980s have apparently dimmed the romantic ardor associated with the idea of alienation. Nevertheless, a case could be made for the view that the dimensions of alienation described here are alive and well in contemporary analysis.

The times now being more sanguine than the post-Depression and World War era, alienation appears now under other names and in more positive guise. Thus, we now see the prevalence of work on social supports (rather than social isolation), or work on mastery and efficacy (rather than powerlessness), and analysis which emphasizes the positive theme of human agency. In a way, it hardly matters what language is used, so long as at the same time the spirit and the classical significance of the idea of alienation is not lost.

See also: Alienation, Sociology of; Anomie; Durkheim, Emile (1858–1917); Industrial Sociology; Marx, Karl (1818–89); Weber, Max (1864–1920)

Bibliography

Kohn M L, Schooler C 1983 *Work and Personality: An Inquiry into the Impact of Social Stratification*. Ablex, Norwood, NJ

Marx K 1964 (1844) *Economic and Philosophic Manuscripts of 1844*. International Publishers, New York

Rotter J B 1966 Generalized expectancies for internal vs. external control of reinforcements. *Psychological Monographs* **80**: 1–28(Whole No. 609)

Sapolsky R M 1992 *Stress, the Aging Brain, and the Mechanism of Neuron Death*. MIT Press, Cambridge, MA

Schacht R 1970 *Alienation*. Doubleday, Garden City, NY

Seeman M 1991 Alienation and anomie. In: Robinson J P, Shaver P R, Wrightsman L S (eds.) *Measures of Personality and Social Psychological Attitudes*. Academic Press, San Diego, CA, pp. 291–321

Seeman T E, McEwen B S 1996 Environment characteristics: The impact of social ties and support on neuroendocrine regulation. *Psychosomatic Medicine* **58**: 459–71
Smelser N J 1963 *Theory of Collective Behavior*. Free Press, New York

M. Seeman

Alienation, Sociology of

Rather than present an overly strict definition of this rather vague umbrella concept which many would not agree with, the italics in the following five points sum up the elements that should bring the concept in sharper focus.

(a) Alienation is an *umbrella concept* that includes, but does not necessarily or logically inter-relate, the dimensions of alienation distinguished by Seeman: *powerlessness*, *meaninglessness*, *normlessness*, *social isolation*, *cultural estrangement*, *and self-estrangement* (Seeman 1959, 1976, 1989).

(b) With the obvious exception of self-estrangement, alienation always points to a *relationship between a subject and some*—real or imaginary, concrete or abstract—*aspect of his environment*: nature, God, work, the products of work or the means of production, other people, different social structures, processes, institutions, etc. Even self-estrangement could be conceived as implying a relation between subjects and their environment: the unreachable 'real self' described by Horney (1950) and others, as the product of a society still pervaded by Cartesian dualism.

(c) Since alienation is usually employed as an *instrument of polemical criticism*, *rather than as a tool of analysis and description*, this relationship can be described as one of separation—a *separation that is considered undesirable from some point of view*. Literature about the possible positive functions of alienation is very sparse indeed, probably because desired separations do not form a serious problem for anyone.

(d) Alienation always refers to a *subjective state of an individual*, or rather to a momentary snapshot of what is usually viewed, both in psychoanalytic and Marxist theory, as a *self-reinforcing inner process*. Societies, institutions, large-scale societal processes, etc. can most certainly be alienating, but to describe them as alienated would endow them with an awareness they do not have.

(e) *Viewing alienation as a subjective individual state or process implies nothing yet about its causation*: It may either be largely brought about by another preexistent subjective, 'reified' state of the same individual, as psychoanalytic theory would hold (although admittedly, such a state would ultimately be environment-induced, e.g., by neuroticizing parents, traumatic early-life experiences, etc., but not directly environment-caused in the present), or by factors having an 'objective' existence in the individual's present environment (e.g., the Marxist and non-Marxist approaches regarding alienating work situations).

1. *A Short History of the Concept*

Alienation is a venerable concept, with its roots going back to Roman law, where *alienatio* was a legal term used to denote the act of transferring property. St. Augustine described insanity as *abalienatio mentis*; Ludz (1975) has discussed its use among the early Gnostics.

In modern times, the concept surfaced again in the nineteenth century and owes its resurgence largely to Marx and Freud, although the latter did not deal with it explicitly. After World War II, when societal complexity started its increasingly accelerated rate of change, and the first signals of postmodernity were perceived by the intellectual elite, alienation slowly became part of the intellectual scene; Srole (1956) was one of the first in the 1950s to develop an alienation scale to measure degrees and varieties of alienation. Following the 1968 student revolutions in Europe and the USA, alienation studies proliferated, at least in the Western world.

In Eastern Europe, however, even the possibility of alienation was denied; theoretically, it could not exist, since officially the laborers owned the means of production. However, the existence of alienation in the 'decadent, bourgeois' societies of the West was gleefully confirmed, as it was supposed to herald the impending demise of late capitalism.

In the Western world, and especially the USA, empirical social psychological research on alienation rapidly developed. Several alienation scales were developed and administered to college students (even national samples) and especially to different disadvantaged minority groups which, not surprisingly, tended to score high on all these scales. On the other hand, much of the theoretical work was of a Marxist persuasion and largely consisted of an exegesis of especially the young Marx's writings and their potential applicability to all kinds of negatively evaluated situations in Western society: the alienation of labor under capitalism, political alienation and apathy, suppression of ethnic or other minority groups, and so forth.

Thus, the 1970s were characterized by a great divide with, on the one hand, the empirical researchers—often, though not exclusively, non-Marxist—administering their scales and charting the degree of alienation among several subgroups, and, on the other hand, the (generally neo-Marxist) theoreticians, rarely engaging in empirical research at all.

During the 1980s, as the postwar baby boomers grew older, and perhaps more disillusioned, and willy-nilly entered the rat race, interest in alienation sub-

sided. The concept definitely became less fashionable, although a small but active international core group continued to study the subject in all its ramifications, since the problems denoted by alienation were certainly far from solved.

Maturing in relative seclusion, this core group, the Research Committee on Alienation (Geyer 1996, Geyer and Heinz 1992, Geyer and Schweitzer 1976, 1981, Kalekin-Fishman 1998, Schweitzer and Geyer 1989) of the International Sociological Association (ISA), managed to narrow the hitherto existing gap between empirical and theoretical approaches and between Marxist and non-Marxist ones. The empiricists basically knew by now who were the alienated and why, and they realized the near-tautology inherent in discovering that the (objectively or subjectively) disadvantaged are alienated. Moreover, many Marxist theoreticians had exhaustively discussed what Marx had to say on alienation, commodity fetishism, and false consciousness and were ready to engage in empirical research along Marxist lines.

It is in the work going on in alienation research since the 1990s that two developments converge: While 'classical' alienation research is still continuing, the stress is now, on the one hand, on describing new forms of alienation under the 'decisional overload' conditions of postmodernity, and on the other hand on the reduction of increasingly pervasive ethnic alienation and conflict. Summarizing, one could say that attention has shifted increasingly to *theory-driven* and *hypothesis-testing* empirical research and to attempts at discovering often very pragmatic *strategies for dealienation*, as manifested by research on Yugoslav self-management and Israeli kibbutzim.

2. *Changes in the Nature of Alienation During this Century*

To oversimplify, one might say that a new determinant of alienation has emerged, in the course of the twentieth century, which is not the result of an insufferable lack of freedom but of an overdose of 'freedom,' or rather, unmanageable environmental complexity. Of course, the freedom-inhibiting classical forms of alienation certainly have not yet been eradicated, and they are still highly relevant for the majority of the world's population. Freud and Marx will continue to be important as long as individuals are drawn into freedom-inhibiting interaction patterns with their interpersonal micro- or their societal macro-environment. However, at least for the postmodern intellectual elite, starting perhaps with Sartre's wartime development of existentialist philosophy, it is the manifold consequences of the knowledge- and technology-driven explosion of societal complexity and worldwide interdependence that need to be explained.

Perhaps this started out as a luxury problem of a few well-paid intellectuals and is totally irrelevant even now for the majority of the world's inhabitants, as it is, certainly, under the near-slavery conditions still existing in many parts of the Third World. Nevertheless, in much of the Western world, the average person is increasingly confronted, on a daily basis, with an often bewildering and overly complex environment, which promotes attitudes of political apathy, often politically dangerous oversimplification of complex political issues, and equally dysfunctional withdrawal from wider social involvements.

Postmodern philosophy has largely been an effort at explaining the effects of this increased complexity on the individual, but while it is largely a philosophy about the fragmentation of postmodern life, it often seems somewhat fragmented itself. What else can one expect perhaps, given Marx's insight that the economic and organizational substructure tends to influence the ideological superstructure? However, while postmodern philosophy certainly draws attention to a few important aspects of postmodern living, it will be argued elsewhere that modern second-order cybernetics can offer a much more holistic picture of societal development over the past few decades (see *Sociocybernetics* and Geyer 1989–98), and provides a metalevel linkage between the concepts of alienation, ethnicity, and postmodernism discussed here.

3. *Changing Emphasis Towards Problems of Ethnicity, Postmodernism, and Increasing Environmental Complexity*

Two developments converge in the work going on in alienation research since the 1990s: while 'classical' alienation research is still continuing, the stress is now, on the one hand, on describing new forms of alienation under the 'decisional overload' conditions of postmodernity and, on the other hand, on the reduction of increasingly pervasive ethnic alienation and conflict, and on alienation as caused by high joblessness rates among uneducated and disadvantaged youth in the Western world, largely as a result of the export of cheap labor to the Third World.

Since the start of the 1990s, there has again been an upsurge of interest in alienation research, caused by different developments: First of all, the fall of the Soviet empire gave a tremendous boost to alienation research in Eastern Europe, for two reasons: (a) the population as a whole was finally free to express its long-repressed ethnic and political alienation, which had accumulated under Soviet rule, while (b) the existence of alienation was no longer denied and instead became a respectable object of study. In the 1970s only a few researchers in relatively strong social positions, could permit themselves to point to the existence of alienation under communism (Schaff 1977).

Second, though processes of globalization and internationalization tended to monopolize people's

attention during the second half of the twentieth century, the hundred-odd local wars fought since the end of World War II, increasingly covered live on worldwide TV, claimed attention for the opposing trend of regionalization and brought ethnic conflicts to the fore, as demonstrated by the battle for Kosovo.

Third, postmodernism emerged as an important paradigm to explain the individual's reactions to the increasingly rapid complexification and growing interdependence of international society. Many of the phenomena labeled as characteristic for postmodernity squarely fall under the rubric of alienation; in particular, the world of simulacra and virtual reality tends to be an alienated world, for reasons that Marx and Freud could not possibly have foreseen.

Schacht (1989) argued that in modern, complex, and highly differentiated multigroup societies the struggle against alienation should be concentrated on evitable alienations. According to Schacht, one cannot possibly be involved with 'society' as one can with 'community,' but only with some of the social formations within it: i.e., specific processes and institutions, that definitely cannot be considered to stand as parts to a whole. Such involvement is necessarily selective and limited, and depends on individual preferences, character, and possibilities, sometimes even on a random and unique series of accidents. Schacht's recipe for unalienated living in what he calls post-Hegelian society departs from Nietzsche's idea of enhanced spirituality, but without its implication of a kind of extraordinary quasi-artistic development of which only the exceptionally gifted are capable. He wants to add the egalitarian spirit of Marx, but without his emphasis upon each person's cultivation of the totality of human powers, although a certain breadth in the range of one's involvements and pursuits is desirable to prevent stunted growth.

Schacht then maintains that modern, liberal society is still the best possible one for self-realization along these lines, owing to the proliferation of structured contexts into which selective entry is possible—in spite of the limited access to some of these contexts for often large parts of the population.

4. *Methodological Issues*

While agreeing with Seeman that alienation is a subjective phenomenon, one can disagree with his methodological implication, i.e., that the individual is always fully aware of his or her alienated state, and is always able to verbalize it. In that sense, both the psychoanalytic and the Marxist approach seem more realistic with their functionally roughly equivalent concepts of *repression* and *false consciousness*, but the disadvantage of these approaches is an almost inevitable authoritarianism: the external observer decides, on the basis of the subject's inputs (class position, working conditions, life history, etc.) as compared with his or her outputs (behaviors, scores on an alienation scale) whether the subject is alienated or not. The subjects themselves unfortunately have very little say in the matter, whether lying on the analyst's couch or standing on the barricades, and may or may not be or become aware of their alienation, and may even—rightly or wrongly—deny being afflicted with it.

Of course, there are many clear-cut cases where the ascription of alienation by an external observer—even if used as a critical and normative rather than as a descriptive and merely diagnostic concept—is clearly warranted, even though the persons concerned may deny their alienation because of repression or false consciousness: childhood abuse, clearly traumatizing experiences, living under conditions of extreme economic deprivation or an abject political system, exploitative working conditions, etc. But there are many not quite so appalling, but still undesirable situations in the Western world nowadays where it seems less useful to ascribe alienation to persons or groups out of a missionary drive to cure others of something they are either blissfully unaware of, or perfectly content with.

5. *Probable Future Directions of Alienation Theory and Research*

While Marxist and Freudian situations of powerlessness and other forms of alienation still abound, and the struggle against these should certainly continue, it has become evident that one is inevitably alienated from lots of things—alienation here being defined as a subjectively undesirable separation from something outside oneself (the means of production, God, money, status, power, the majority group to which one does not belong, etc.) or even inside oneself (one's 'real' inner feelings, drives or desires, as in the concept of self-alienation).

Schacht considers this indeed inevitable, and his sober appraisal contrasts with the often highly normative and evaluative character of earlier alienation studies, the Marxist ones castigating the evil effects of late capitalism on the individual, and the psychoanalytic ones deploring the effects of early-life neuroticizing influences. While admittedly Marxist and Freudian types of alienation are still prevalent in much of the world and should certainly be combated, new types of alienation have entered the scene that are caused by the increasingly accelerating complexification of modern societies. They can only be hinted at below, and have to do with phenomena like selection and scanning mechanisms, problems of information overload as well as decisional overload, and the need to engage often in counterintuitive rather than spontaneous behavior.

These modern forms of alienation have the 'disadvantage' that they are nobody's fault. No one, not even late capitalism or insensitive parents, can be

blamed for the fact that the world is becoming more complex and interdependent, that consequently causal chains stretch further geographically and timewise, and that—if one wants to reckon with their effects—one has more than ever to 'think before one acts,' and even to engage in spontaneity-reducing and therefore alienating forms of 'internal simulation.' The process of complexification is not only nobody's fault, but it is also irreversible, and cannot be turned back in spite of proclamations that 'small is beautiful.' One tends to lose a sense of mastery over one's increasingly complex environment, but it is different from the sense of mastery the alienated laborers of Marxist studies are supposed to gain if only they owned the means of production, or the psychoanalysts's clients if their neurotic tendencies would evaporate after looking at their analyst's diploma on the ceiling for half a decade while reliving early or not so early traumas.

The result of the emergence of these modern forms of alienation is that alienation studies, at least to the extent they deal with these modern forms, are becoming more value-neutral (a dirty word since the 1970s), less normative, moralistic, and value-laden. Once more: it is not implied that moral indignation and corrective action based on that indignation are not called for as long as millions of people are exploited and subjugated, or even tortured and killed, in the countless small wars that have replaced the relatively benign Cold War.

What is clear is that modern forms of alienation are emerging and will affect increasing numbers of people in the developed world, and soon also in the developing world. Several authors have hinted at this development. Lachs (1976) spoke of a *mediated world*, where the natural cycle of planning an action, executing it, and being confronted with its positive or negative consequences is broken, and where one is less and less in command of more and more of the things that impinge on one's life, without being able to impute blame on anyone or anything. Etzioni (1968) likewise saw alienation as resulting from nonresponsive social systems that do not cater to basic human needs. Toffler (1970, 1990) vividly described how change is happening not only faster around us, but even through us.

The alienated used to blame their woes on the wicked capitalist or their unsatisfactory parents, even though that was an obvious oversimplification. The common point in all these modern alienation forms is that they result from the increasing complexification of modern world society that we have brought about, reinforced by the aggregated individual and group reactions to this very complexification. One has to find ways of adapting to this irreversible process, since one cannot 'undo' the products, processes, and institutions that have emerged since the middle of the twentieth century. One cannot function adequately or participate fully in a world characterized by information overload without developing efficient selection mechanisms to select quickly what may be useful from the often unwanted information deposited at one's doorstep, and without developing effective mechanisms to scan the environment for information one needs to further one's goals. Moreover, if one tries to keep an open mind, the chance that one changes one's goals before having had the time to realize them is greater than ever before in history. Our civilization stresses the importance of learning; but has not yet sufficiently stressed the importance of *unlearning*, as Toffler (1970) has stressed; the 'halving time' of knowledge is far shorter than that of uranium.

The individual living in a world saturated with communication media is offered the possibility of thoroughly identifying with different alternative life scenarios, and at least in much of the Western world many of these scenarios can be realized if one is willing to pay the inevitable price. But a lifetime is limited, and so are the scenarios one can choose and try to realize. One of the consequences of this media-driven conscious awareness of alternative life scenarios—coupled with the freedom but also the lack of time to realize them all—is that the percentage of unrealized individual possibilities is greater than ever before, which certainly contributes to a diffuse sense of alienation: 'I'm living this life, but could have lived so many other ones.' Unlike Abraham, one cannot anymore die 'one's days fulfilled.'

Naturally, it can be maintained this is a spoilt-child syndrome, induced by the infantilizing influence of the media: fantasies are stimulated without parents telling the ever more insecure child 'this is impossible.' This accords with Schacht, who favors limited and selective involvement with the world; one cannot be involved with society as one could with community, let alone with primary group contacts.

As the development of the information society further continues, alienation towards the interpersonal environment and alienation towards the societal environment may well turn out to be inversely related. Many of those who have a high capacity for dealing with societal complexity (the educated and the academics among others), especially when they make much use of this capacity in their daily lives (the 'organization men and women,' the managers and planners), tend to generalize their 'planning attitudes,' probably due to the visible success of the associated operating procedures in the societal sphere, to encompass their interpersonal contacts. Consequently, they may become interpersonally alienated, and often see simple interpersonal relations as more complex than they actually are. They are insufficiently involved in the present, being used to internally simulate every move, to constantly think and plan ahead.

Conversely, those who have a low capacity for dealing with environmental complexity (the uneducated, especially those living in still relatively simple societies, amongst others), especially when their lowly position in complex hierarchical structures does not

require much planning regarding their wider societal environment (e.g., the unskilled), on the contrary tend to generalize their 'involvement attitudes' to include whatever societal interaction loops they are engaged in. The societally alienated tend to see complex societal relations as less complex than they actually are. They are, in direct opposition to the first group, insufficiently involved in the future, not because they cannot kick the habit of being involved in the here-and-now, but because they never developed the 'broadsight' and 'long-sight' (Elias 1939) that often characterizes the interpersonally alienated.

If it is indeed true that the interpersonally nonalienated tend to be the societally alienated, who clamor for a larger share of the societal pie, while the societally nonalienated tend to be in the power positions because they are best able to reduce societal complexity, and consequently have a fair chance of being interpersonally alienated, then the question becomes: Can a complex society ever be a nonalienating society, if it is led by those who score highest on interpersonal alienation? Or, as Mannheim asked: 'who plans the planners?' Alienation will certainly never disappear, whether in politics or in work situations, whether in interpersonal or societal interactions, but it may be considerably reduced by de-alienating strategies based on social science research.

See also: Alienation: Psychosociological Tradition; Anomie; Critical Theory: Contemporary; Critical Theory: Frankfurt School; Freud, Sigmund (1856–1939); Industrial Sociology; Marx, Karl (1818–89); Marxist Social Thought, History of; Work and Labor: History of the Concept; Work, History of; Work, Sociology of

Bibliography

Elias N 1939 *Über den Prozess der Zivilisation*. Haus zum Falken, Basel

Etzioni A 1968 *The Active Society*. Collier-Macmillan, London

Geyer F 1989–98 http://www.unizar.es/sociocybernetics/chen/felix.html

Geyer F (ed.) 1996 *Alienation, Ethnicity, and Postmodernism*. Greenwood, Westport, CT

Geyer F, Heinz W R (eds.) 1992 *Alienation, Society, and the Individual*. Transaction, New Brunswick, NJ

Geyer F, Schweitzer D (eds.) 1976 *Theories of Alienation—Critical Perspectives in Philosophy and the Social Sciences*. Martinus Nijhoff, The Hague

Geyer F, Schweitzer D (eds.) 1981 *Alienation: Problems of Meaning, Theory and Method*. Routledge and Kegan Paul, London

Horney K 1950 *Neurosis and Human Growth*. W. W. Norton, New York

Kalekin-Fishman D (ed.) 1998 *Designs for Alienation: Exploring Diverse Realities*. SoPhi Press, Jyväskylä, Finland

Lachs J 1976 Mediation and psychic distance. In: Geyer F, Schweitzer D (eds.) *Theories of Alienation—Critical Perspectives in Philosophy and the Social Sciences*. Martinus Nijhoff, The Hague, pp. 151–67

Ludz P 1975 'Alienation' als Konzept der Sozialwissenschaften. *Kölner Zeitschrift für Soziologie* **27**(1): 1–32

Schacht R 1971 *Alienation*. Allen & Unwin, London

Schacht R 1989 Social structure, social alienation, and social change. In: Schweitzer D, Geyer F (eds.) *Alienation Theories and De-alienation Strategies—Comparative Perspectives in Philosophy and the Social Sciences*. Science Reviews, Northwood, UK, pp. 35–56

Schaff A 1977 *Entfremdung als soziales Phänomen*. Europa, Vienna

Schweitzer D, Geyer F (eds.) 1989 *Alienation Theories and De-alienation Strategies—Comparative Perspectives in Philosophy and the Social Sciences*. Science Reviews, Northwood, UK

Seeman M 1959 On the meaning of alienation. *American Sociological Review* **24**(6): 783–91

Seeman M 1976 Empirical alienation studies: An overview. *Theories of Alienation—Critical Perspectives in Philosophy and the Social Sciences*. Martinus Nijhoff, The Hague, pp. 265–305

Seeman M 1989 Alienation motifs in contemporary theorizing: The hidden continuity of the classic themes. In: Schweitzer D, Geyer F (eds.) *Alienation Theories and De-alienation Strategies – Comparative Perspectives in Philosophy and the Social Sciences*. Science Reviews, Northwood, UK, pp. 33–60

Srole L 1956 Anomie, authoritarianism, and prejudice. *American Journal of Sociology* **62**(1): 63–7

Toffler A 1970 *Future Shock*. Bantam Books, New York

Toffler A 1990 *Power Shift—Knowledge, Wealth and Violence at the Edge of the 21st Century*. Bantam Books, New York

F. Geyer

Alliances and Joint Ventures: Organizational

1. Introduction

Cooperative arrangements between organizations date back to those between merchants in ancient Babylonia, Egypt, Phoenicia, and Syria who used such arrangements to conduct overseas commercial transactions. Since the 1980s, there has been a dramatic growth in the use of various forms of cooperative arrangements such as joint ventures between organizations. Several reasons have been offered for this unprecedented growth in alliances: greater internationalization of technology and of product markets, turbulence in world markets and higher economic uncertainty, more pronounced cost advantages, and shorter product life cycles. A noteworthy feature accompanying this growth in alliances has been the tremendous diversity of national origins of partners, their goals and motives for entering alliances, as well as the formal legal and governance structures utilized. The increase in cooperative arrangements has generated a renaissance in the scholarly study of alliances

(Gulati 1998). Researchers in several disciplines, including economics, sociology, social psychology, organization behavior, and strategic management, have sought answers to several basic questions: What motivates firms to enter into alliances? With whom are they likely to ally? What types of contracts and other governance structures do firms use to formalize their alliances? How do alliances themselves and firm participation patterns in alliances evolve over time? What factors influence the performance of alliances and the benefits partners receive from alliances?

We define an alliance as any voluntarily initiated and enduring relationship between two or more organizations that involves the sharing, exchange, or codevelopment of resources (e.g., capital, technology, or organizational routines). Joint ventures are a subset of alliances and typically entail the creation of a new entity by two or more partners who retain equity in the new entity. Alliances can be classified in numerous ways. They can be either horizontal or vertical, depending on the relationship of the alliance partners across the value chain. They can also be classified according to the motivations of the partners entering them (e.g., reducing costs vs. excluding potential competitors vs. developing new products) and according to the governance structure utilized (e.g., joint venture vs. minority equity position vs. licensing arrangement). The wide range of analytically useful distinctions between alliances reflects the complexity and diversity across the vast range of cooperative arrangements entered by firms.

2. *Theoretical Perspectives*

While alliances have been examined from a variety of theoretical perspectives, two disciplines have been particularly influential: economics and sociology. Within economics, there have been several approaches to the study of alliances including industrial economics, game theory, and transaction cost economics (Kogut 1988). The industrial economics perspective suggests that a firm's relative position within its industry structure is critical and that alliances can be used to increase this position relative to other rivals or consumers (Porter 1990). Industrial economists have studied how firms use alliances and other hybrid arrangements to enhance market power, maximize parent firm profits, acquire capabilities and competencies, and enhance their strategic position (Berg et al. 1982, Vickers 1985, Porter and Fuller 1986, Ghemawat et al. 1986).

Recently, scholars have extended earlier research and used insights from game theory to illuminate some of the process dynamics within alliances and their impact for firms (Gulati et al. 1994, Khanna et al. 1998). Some of the key elements have been to model the interdependence between alliance partners in terms of varying payoffs under differing scenarios. This provides a useful window into the dynamics that unfold in alliances based upon some of the ex ante conditions in the alliance.

Transaction cost economists have examined alliances and considered how transactional hazards may influence the extent to which firms use alliances as opposed to market transactions or internal production. Furthermore, scholars have also examined the formal structure of alliances and considered the extent to which transaction costs influence the governance arrangement firms use to formalize their alliances (Pisano 1990, Gulati 1995a, Gulati and Singh 1998, Oxley 1998).

A recent critique of much of the research on strategic alliances is that it has presented an undersocialized account of firm behavior (Gulati 1998). Thus, industrial economics and transaction cost economics in particular, have focused typically on the influence of structural features of firms and industries, neglecting the history and process of social relations between organizations. Sociologists have suggested that economic action and exchange operate in the context of historical structures of relationships constituting a network that informs the choices and decisions of individual actors (Granovetter 1985, Burt 1992). A network is a form of organizing economic activity that involves a set of nodes (e.g., individuals or organizations) linked by a set of relationships (e.g., contractual obligations, trade association memberships, or family ties).

The network approach to analyzing key questions about alliances builds on the idea that economic actions are influenced by the context in which they occur and that actions are influenced by the position actors occupy within a network. Organizations are treated as fully engaged and interactive with the environment rather than as an isolated atom impervious to contextual influences. Several recent scholars have followed a network approach and studied how preexisting interfirm relationships can cumulate into a network and thus influence fundamental dynamics associated with alliances (Kogut et al. 1992, Gulati 1995a, 1995b, Gulati and Gargiulo 1999, Gulati and Westphal 1999). This is based on the premise that, through their network positions, firms have differential access to information about current and potential alliance partners. Researchers have examined the role of relational embeddedness, which emphasizes the role of direct close ties as a mechanism for facilitating exchange, and also structural embeddedness, which emphasizes the structural positions of alliances partners with a network.

In the remainder of this article we shall discuss some of the new insights emerging from a network perspective on alliances for three core issues that are critical for the study of alliances (adapted from Gulati 1998): (a) formation—which firms enter alliances and with whom do they ally, (b) governance—what factors influence the governance structure firms use to for-

malize their alliances, and (c) performance—what factors influence the performance of alliances and of the firms entering alliances.

3. Formation of Alliances

Some of the key questions regarding the formation of alliances include: Which firms enter alliances and whom do firms choose as alliance partners? Instead of focusing on the factors that may motivate firms to seek alliances, we focus here on the factors that both predispose firms to enter into alliances with greater or lesser frequency and also influence their choice of alliance partners. These questions are examined at the firm and dyad level, respectively. Among the studies that have considered the factors that influence the alliance proclivity of firms, the focus has remained on some of the economic and strategic determinants of tie formation such as the customer bargaining power, market standardization, degree of asset configuration, and degree of product substitutability (Harrigan 1988). Follow-up research has remained focused on looking at the influence of firm-level attributes such as age, size, competitive positions, financial resources, and product diversity (as proxies for economic and strategic determinants) on the propensity of firms to enter alliances. This work has been extended recently to suggest that beyond the strategic imperatives, the proclivity of firms to enter alliances is also influenced by the available amount of 'network resources,' which is a function of the firm's position in its interfirm network of prior alliances (Gulati 1999).

Another line of inquiry from a transaction cost economics perspective has examined the alliance formation question as one of a choice between 'make, buy, or ally.' Since the 1980s there has been resurgence of research on the economic theory of the firm, with a particular emphasis on the 'make or buy' decision of firms (for a review, see Klein and Shelanski 1995). The key finding of this research is that the observed firm choice between these alternative governance forms (e.g., in-house production vs. outsourcing) that firms use to procure requisite goods and services is determined by the transactional hazards associated with them. The greater the transactional hazards associated with a commodity, the more likely are firms to use hierarchical governance arrangements. The logic for hierarchical controls as a response to appropriation concerns is based on the ability of such controls to assert control by fiat, provide monitoring, and align incentives. This basic premise has been considerably refined with new and detailed measures of the potential transactional hazards that could influence governance choice by firms (e.g., Masten et al. 1991). The operation of such a logic originally was examined in the classic make-or-buy decisions that were cast in Coasian terms as a question of the boundaries of the firm (e.g., Monteverde and Teece 1982, Walker and Weber 1984, Masten et al. 1991).

While there is widespread recognition that the original bipolar choice of governance arrangements examined by transaction cost theorists is no longer valid, there have been only limited inquiries into enlarging the original question and a direct consideration of the 3-way choice between make, ally, and buy (Gulati and Lawrence 1999). The same logic by which firms choose between the extremes of make or buy is also expected to operate now that firms face an enlarged choice between make, ally, and buy: the greater the transactional hazards, the more hierarchical the governance structure. While all these studies have advanced our understanding of some of the factors that influence the creation of alliances as opposed to alternative governance structures, their primary focus on transactional characteristics as determining factors has led them to incompletely examine the role of the interorganizational networks in influencing this choice. This remains an exciting arena for future research.

The two approaches to alliance formation outlined above have focused on individual firms and on individual transactions respectively. Yet another approach is to consider this issue at the dyad level and assess the factors that influence who partners with whom in an alliance. Due to the paucity of information on the reliability and competencies of potential alliance partners, there is often uncertainty over the likely behavior of the contracting parties. Network connections and third-party referrals can play an important in role in reducing such uncertainty. As a result, they may influence the availability of and access to alliance opportunities that firms may perceive. Recent empirical research corroborates these claims and suggests that the choice of partners for alliances is influenced by strategic interdependence and also the network antecedents of the partners (Gulati 1995b, Gulati and Gargiulo 1999, Gulati and Westphal 1999).

4. Governance of Alliances

A key question regarding the governance structure of alliances is: Which ex ante factors affect the choice of alliance governance structure? Cooperative interorganizational relationships such as strategic alliances and joint ventures require negotiation of property rights governing the long-term resources invested in the arrangement. The essence of cooperative strategy is achievement of an agreement and a plan to work together such that each partner essentially becomes an agent for the other(s). Firms face numerous choices in structuring their alliances. The governance structure must include a variety of components: planning and control systems, incentive systems, information systems, and partner selection systems. Property rights

cannot always be controlled fully or specified in advance, so in part, governance of joint ventures operates by firms being placed into mutual hostage positions as they commit financial capital, real assets, and other resources to the venture (Kogut 1988).

One basic governance question concerns whether the cooperative arrangement does or does not involve equity. Questions about governance structure have been examined in depth by transaction cost economists. Pisano and co-workers (Pisano et al. 1988, Pisano 1989) found that the greater the potential transaction costs, the more likely parties are to design a hierarchical contractual arrangement and to use equity alliances rather than nonequity alliances. They argue that equity alliances are an effective governance structure for mitigating transaction costs because each partner's concern for its investment reduces opportunistic behavior (i.e., there is a 'mutual hostage' situation) and because equity partners establish hierarchical supervision through their service on the board of directors. Gulati (1995a) extended this work by suggesting that such an approach erroneously treats each transaction independently and ignores the crucial role of the history of prior transactions between partners. He found that prior alliances between firms engenders interorganizational trust that reduces the likelihood of using equity arrangements in future ties. Gulati and Singh (1998) have gone a step further and found that the use of hierarchical arrangements such as equity alliances is influenced not only by appropriation concerns but also by anticipated coordination concerns that were highlighted originally by organization design theorists but have not been the focus of recent research.

5. *Performance of Alliances*

Key questions regarding the performance of alliances include: How can we measure the performance of alliances? Which factors affect the performance of alliances? Do firms receive economic and social benefits from participating in alliances? Due to onerous research obstacles related to establishing criteria for measuring performance and the logistical challenges of collecting the detailed data necessary, the performance of alliances and of alliance partners remains an exciting but underexplored area. This is especially relevant since several anecdotal accounts suggest that the failure rate of alliances may be as high as 50 to 80 percent. Scholars suggest that alliances are often difficult to manage, taxing of top management's cognitive resources, and limiting of the partners' organizational autonomy, and thus prone to failure. They have identified several factors that could influence the success rate of alliances: building interfirm trust, continuity in interface personnel, management flexibility, proactive conflict resolution mechanisms, and regular information exchange (e.g., Kanter 1989, Bleeke and Ernst 1991). The primary approach taken by empirical studies of alliance performance has been to examine factors affecting the termination of alliances such as: industry concentration and growth rates, partners' country of origin, duration of the alliance, competitive overlap between partners, and alliance governance structure (Beamish 1985, Harrigan 1986, Kogut 1989, Levinthal and Fichman 1988). However, these studies typically are limited in two respects: (a) failure to distinguish natural vs. untimely deaths of alliances and (b) an inability to distinguish gradations of alliance performance because of the consideration of all performance dichotomously as either survival or death.

A basic and particularly vexing question is how to measure alliance success (Anderson 1990). Given the multifaceted objectives of alliances, focusing on technical efficiency and financial outcomes (e.g., return on assets) often does not offer an adequate evaluation of performance and determination of antecedents. Management and strategy scholars have begun to use extensive surveys and more longitudinal data to examine performance (Harrigan 1986, Heide and Miner 1992, Parkhe 1993). One area of inquiry however that has not been pursued by such scholars has been to consider the impact of social networks and embeddedness on the relative performance of alliances. This is especially important since there is some preliminary empirical evidence that alliances embedded in social networks are less likely to terminate and are more effective in situations of high uncertainty (Kogut 1989, Levinthal and Fichman 1988, Seabright et al. 1992). A related direction for future research would be to examine the simultaneous and potentially conflicting influence of multiple social networks on alliance performance. As more and more organizations enter multiple cooperative arrangements, with some finding themselves in hundreds of alliances (e.g., General Electric, Hewlett-Packard, and IBM), a portfolio perspective is needed to examine the degree to which multiple alliance participation generates both conflicting demands and productive synergies.

Do organizations benefit from entering strategic alliances? This question is distinct from the previous issue of the alliance's overall performance. Given that many other factors besides an alliance influence an organization's performance, it can be difficult to establish a causal link between alliance participation and organizational benefit. Consequently, researchers have looked to a variety of direct and indirect measures, including stock market effects (e.g., Koh and Venkatraman 1991, Anand and Khanna 1997) and the likelihood of organizational survival (e.g., Baum and Oliver 1991, 1992). The results of these studies have been mixed but generally suggest that alliances are beneficial for organizations. Once again, there have been only few efforts to link the social structural context of alliances with the performance

benefits that are likely to ensue from them (e.g., Gulati and Wang 1999).

An important lacuna in much of the research on alliances has been insufficient attention to the dynamics of such ties. A few studies that have focused on process issues suggest that alliances are subject to dynamic evolutionary processes that cause significant transformations beyond their original designs and mandates (Hamel 1991, Larson 1992, Ring and Van de Ven 1994, Doz 1996). Change can occur at the intrafirm, interfirm, network, industry, and societal levels. At the firm level, as the pay-offs for an alliance or at least each partner's perception of those pay-offs changes, the incentives for cooperation can change (Gulati et al. 1994, Khanna et al. 1998). Key changes also takes place at the network level, as a firm's social network (the main source of its current and potential alliance partners) evolves, which changes the structural positions of organizations within the network and affects the pattern of future ties. Even more profound dynamics can operate on the industry and national levels. In this article, we have tried to outline the importance of the social network and structural embeddedness perspectives for the study of alliances. We believe that it is particularly valuable in informing scholars and managers on 'how' (as opposed to why) dyadic, network, and societal level dynamics affect the evolution and eventual performance of alliances. This perspective also opens up new critical questions for future research on the dynamics of alliances such as how firms manage a portfolio of alliances, how firms position themselves optimally within a network, and how social network membership affects alliance performance.

See also: Competitive Strategies: Organizational; Conflict: Organizational; Corporate Culture; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; Information and Knowledge: Organizational; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; International Business; International Organization; International Trade: Commercial Policy and Trade Negotiations; Learning: Organizational; Monetary Policy; Rational Choice and Organization Theory; Strategy: Organizational; Technology and Organization

Bibliography

Anand B Khanna T 1997 On the market valuation of interfirm agreements: Evidence from computers and telecommunications, 1990–1993. Working paper. Harvard Business School, Boston

Anderson E 1990 Two firms, one frontier: On assessing joint venture performance. *Sloan Management Review* **31**(2): 19–30

Baum J, Oliver C 1991 Institutional linkages and organizational mortality. *Administrative Science Quarterly* **36**: 187–218

Baum J, Oliver C 1992 Institutional embeddedness and the dynamics of organizational populations. *American Sociological Review* **57**: 540–59

Beamish P 1985 The characteristics of joint ventures in developed and developing countries. *Columbia Journal of World Business* **20**: 13–9

Berg S, Duncan J, Friedman P 1982 *Joint Venture Strategic and Corporate Innovation*. Oelgeschlager, Gunn & Hain, Cambridge, MA

Bleeke J, Ernst D 1991 The way to win in cross border alliances. *Harvard Business Review* **69**(6): 127–35

Burt R 1992 *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA

Doz Y L 1996 The evolution of cooperation in strategic alliances: Initial conditions or learning processes? *Strategic Management Journal* **17**: 55–83

Ghemawat P, Porte, Rawlinson R 1986 Patterns of international coalition activity. In: Porter M (ed.) *Competition in Global Industries*. Harvard Business School Press, Boston, MA pp. 345–66

Granovetter M 1985 Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* **91**(3): 481–510

Gulati R 1995a Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. *Academy of Management Journal* **38**: 85–112

Gulati R 1995b Social structure and alliance formation patterns: A longitudinal analysis. *Administrative Science Quarterl* **40**: 619–52

Gulati R 1998 Alliances and networks. *Strategic Management Journal* **19**: 293–317

Gulati R 1999 Network location and learning: The influence of network resources and firm capabilities on alliance formation. *Strategic Management Journal* **20**: 397–420

Gulati R, Gargiulo M 1999 Where do interorganizational networks come from? *American Journal of Sociology* **104**: 1439–93

Gulati R, Khanna T, Nohria N 1994 Unilateral commitments and the importance of process in alliances. *Sloan Management Review* **35**(3): 61–9

Gulati R, Lawrence P1999 The diversity of embedded ties. Working paper. Kellogg Graduate School of Management, Evanston, IL

Gulati R, Singh H 1998 The architecture of cooperation: Managing coordination costs and appropriation concerns in strategic alliances. *Administrative Science Quarterly* **43**: 781–814

Gulati R, Westphal J 1999 Cooperative or Controlling? The effects of CEO-board relations and the content of interlocks on the formation of joint ventures. *Administrative Science Quarterly* **44**: 473–506

Hamel G 1991 Competition for competence and inter-partner learning within international strategic alliances. *Strategic Management Journal* **12**: 83–103

Harrigan K R 1986 *Managing for Joint Ventures Success*. Lexington Books, Lexington, MA

Harrigan K R 1988 Joint ventures and competitive strategy. *Strategic Management Journal* **9**(2): 141–58

Heide J, Miner A 1992 The shadow of the future: Effects of anticipated interaction and frequency of contact of buyer-seller cooperation. *Academy of Management Journal* **35**: 265–91

Kanter R M 1989 *When Giants Learn to Dance*. Touchstone, Simon & Schuster, New York

Khanna T, Gulati R, Nohria N 1998 The dynamics of learning alliances: Competition, cooperation, and relative scope. *Strategic Management Journal* **19**: 193–210
Kogut B 1988 Joint ventures: Theoretical and empirical perspectives. *Strategic Management Journal* **9**(4): 319–32
Kogut 1989 The stability of joint ventures: Reciprocity and competitive rivalry. *Journal of Industrial Economics* **38**: 183–98
Kogut B, Shan W, Walker G 1992 The make-or-cooperate decision in the context of an industry network. In: Nohria N, Eccles R (eds.) *Networks and Organizations*. Harvard Business School Press, Boston, MA, pp. 348–65
Koh J, Venkatraman N 1991 Joint venture formations and stock market reactions: An assessment in the information technology sector. *Academy of Management Journal* **34**(4): 869–92
Larson A 1992 Network dyads in entrepreneurial settings: A study of the governance of exchange relationships. *Administrative Science Quarterly* **37**: 76–104
Levinthal D A, Fichman M 1988 Dynamics of interorganizational attachments: Auditor-client relationships. *Administrative Science Quarterly* **33**: 345–69
Masten S E, Meehan J W, Snyder E A 1991 The costs of organization. *Journal of Law, Economics and Organization* **7**: 1–25
Monteverde K Teece D 1982 Supplier switching costs and vertical integration in the auto industry. *Bell Journal of Economics* **13**: 206–13
Oxley J E 1997 Appropriability hazards and governance in strategic alliances: A transaction cost approach. *Journal of Law, Economics and Organization* **13**(2): 387–409
Parkhe A 1993 Strategic alliance structuring: A game theoretic and transaction cost examination of interfirm cooperation. *Academy of Management Journal* **36**: 794–829
Pisano G P 1989 Using equity participation to support exchange: Evidence from the biotechnology industry. *Journal of Law, Economics and Organization* **5**(1): 109–26
Pisano G P 1990 The R&D boundaries of the firm: An empirical analysis. *Administrative Science Quarterly* **35**: 153–76
Pisano G P, Russo M V, Teece D 1988 Joint ventures and collaborative arrangements in the telecommunications equipment industry. In: Mowery D (ed.) *International Collaborative Ventures in U.S. Manufacturing*. Ballinger, Cambridge, MA pp. 23–70
Porter M E 1990 *The Competitive Advantage of Nations*. Free Press, New York
Porter M E M, Fuller B 1986 Coalitions and global strategy. In: Porter M (ed.) *Competition in Global Industries*. Harvard Business School Press, Boston, MA pp. 315–43
Ring P S, Van de Ven A H 1994 Developmental processes of cooperative interorganizational relationship. *Academy of Management Review* **19**(1): 90–118
Seabright M A, Levinthal D A, Fichman M 1992 Role of individual attachment in the dissolution of interorganizational relationships. *Academy of Management Journal* **35**(1): 122–60
Shelanski H A, Klein P G 1995 Empirical research in transaction cost economics: A review and assessment. *Journal of Law, Economics, and Organization* **11**(2): 335–61
Vickers J 1985 Pre-emptive patenting, joint ventures, and the persistence of oligopoly. *International Journal of Industrial Organization* **3**: 261–73
Walker G, Weber D 1984 A transaction cost approach to the make-or-buy decisions. *Administrative Science Quarterly* **29**: 373–91

J. Gillespie and R. Gulati

Alliances: Political

Alliances are formal agreements, open or secret, between two or more nations to collaborate on national security issues. They have been variously considered as techniques of statecraft, international organizations, or regulating mechanisms in the balance of power. This article addresses several of the enduring concerns about alliances, including alliance formation, their performance and persistence; their effects on the international system and domestic politics; and their prospects in the new millennium.

1. *Alliance Formation*

Among the oldest explanations of alliances are those derived from balance-of-power theories in which the emphasis is on the external environment, including the structure, distribution of power, and state of relations among units of the system. These are often closely linked to the 'realist' approach to international politics (Morgenthau 1959, Waltz 1979, Gilpin 1981). Nations join forces as a matter of expediency in order to aggregate sufficient capabilities to achieve certain foreign policy goals, to create a geographically advantageous position, or to prevent any nation or combination of countries from achieving a dominant position. Alliance partners are thus chosen on the basis of common goals and needs, not for reasons of shared values, shared institutions, or a sense of community. According to balance-of-power theories, nations should be more likely to join the weaker coalition to prevent formation of a hegemonic one ('balancing') rather than join the dominant one in order to increase the probability of joining the winning side ('bandwagoning') (Waltz 1979). Several case studies found support for the 'balancing' hypothesis: that nations form alliances to balance threats rather than power (Walt 1987).

The 'size principle,' according to which 'coalitions will increase in size only to the minimum point of subjective certainty of winning,' is drawn deductively from 'game theory' (Riker 1962, pp. 32–3). Coalition theories share a number of characteristics with balance-of-power models, but whereas an important goal of 'ideal' balance-of-power systems is to prevent the rise of a dominant nation or group of nations, the primary motivation in game approaches is to form just such a coalition, with enough partners to ensure victory but without any additional ones who would claim a share of the spoils. Although intuitively appealing, a number of difficulties arise when this approach is applied to international politics. If war offers the prospect of winning territories or other divisible rewards, there are advantages to forming an alliance no larger than is necessary to gain victory, but even in redistribution alliances the interests of partners

may be complementary, permitting a non-competitive division of rewards. Moreover, when an alliance is formed for purposes of defense or deterrence, its success is measured by its ability to prevent conflict, not by the territorial or other gains derived from successful prosecution of a war. Another potential difficulty is the demanding requirement that leaders be able to measure capabilities with sufficient precision to define a minimum winning coalition. Any propensity to use 'worst case scenario' reasoning will likely result in a much larger alliance.

In contrast to the balance-of-power and minimum winning coalition theories are those that emphasize national attributes other than power. These approaches do not deny that calculations of national interest or power influence alliance formation, but they also emphasize that we cannot treat nations as undifferentiated units if we wish to understand either their propensity to use alliances as instruments of foreign policy, as opposed to such alternatives as neutrality, or in their choice of alliance partners. Political stability is sometimes associated with a propensity to join alliances, and instability has been seen as an impetus to go beyond non-alignment and pursue a policy of 'militant neutralism;' leaders faced with domestic instability may actively court allies in the hopes of gaining external support for a tottering regime. By utilitarian criteria, small, poor, or unstable nations are relatively unattractive alliance partners, but such nations have often been sought as allies, and they have even become the focal point of acute international crises; twentieth-century examples include Bosnia from 1908 to 1909, Serbia 1914, Cuba from 1961 to 1962, and Vietnam from 1965 to 1973.

Given a propensity to seek allies, does the choice of partners reflect a discernible pattern of preferences? Affinity theories focus on the similarities among nations as an element in their propensity to align. The premise that nations are likely to prefer partners with whom they share common institutions, cultural and ideological values, or economic interests is intuitively appealing, and NATO can be cited as an example, but it has received limited support in studies of pre-1945 alliances.

2. *Alliance Performance*

Palmerston once noted that nations have neither permanent enemies nor allies, only permanent interests. Although the longevity of most alliances can be measured better in years than decades or centuries, these observations do not explain why some alliances are cohesive and effective whereas others are not. To some, open polities are inferior allies on two counts: they experience more frequent changes in ruling elites, with the consequence that commitments to allies may also change, and the demands of domestic politics may take precedence over alliance requirements. But when democracies are under attack they are more likely to expand alliance functions and to turn alliances into communities of friendship; they are also less likely to renege on commitments by seeking a separate peace (Liska 1962, pp. 50, 52, 115).

Political instability is often associated with poor alliance performance. Unstable regimes may experience radical changes in elites, which in turn result in shifting patterns of alignment, and they may also be more willing to run high risks on behalf of their own interests but not those of allies. Differences in national bureaucratic structures and processes may be an important barrier to coordination of alliance strategies. If national security policy is the product of constant intramural conflict within complex and varied bureaucracies, even close allies may fail to perceive accurately the nuances in the bureaucratic politics 'game' as it is played abroad, and how the demands of various constituents may shape and constrain alliance policies (Neustadt 1970, Allison 1971).

Alliances differ in many ways, including the circumstances under which their provisions become operative, the type of commitment, the degree of cooperation, geographical scope, ideology, size, structure, capabilities, and quality of leadership. Presumably all nations prefer to join alliances that offer them an effective role in determining goals, strategy, and tactics; a 'fair' share of the rewards without undue costs; and the maximum probability of success in achieving their goals.

Redistribution alliances might be expected to achieve fewer successes and to break up more easily than those whose primary motives are deterrence and defense. Failure to achieve goals has a disintegrative effect and, other things being equal, it is easier for an alliance of deterrence to succeed because it is sufficient to deny the enemy a victory or to maintain the status quo. A redistribution alliance, on the other hand, must not only be able to avert defeat; it must also achieve victory if it is to be successful.

Ideology may play a role in sustaining or dissolving alliance bonds. Even those who minimize the importance of ideology in alliance formation tend to agree that a shared ideology may ensure that issues are defined similarly and facilitate intra-alliance communication. A common ideology may sustain alliances, but only as long as its tenets do not themselves become an issue.

Large alliances are usually less cohesive than small ones. The larger the alliance, the smaller the share of attention that nations can give to each ally, and as the size of the alliance increases, the number of relationships within the alliance rises even faster. Problems of coordination increase and so do opportunities for dissension. Finally, the larger the alliance, the less important the contributions of any single member, and the easier it is for any partner to become a 'free rider' by failing to meet alliance obligations.

Alliances differ widely in the kinds of political and administrative arrangements that govern their activities. Allies may undertake wide-ranging commitments to assist each other and yet fail to establish institutions and procedures for communication and policy coordination. Some suggest that hierarchical and centralized alliances are likely to be more cohesive and effective because they can mobilize their resources better and can respond more quickly to threats and opportunities. According to others, pluralistic and decentralized alliances are likely to enjoy greater solidarity and effectiveness. In pluralistic alliances, even differences on central concerns are likely to remain confined to the single issue rather than all issues (Holsti et al. 1973).

The proposition that influence within an alliance is proportional to strength may not always be valid. Under some circumstances, weakness may actually be a source of strength in intra-alliance diplomacy. The stronger nation is usually the more enthusiastic partner, it has less to gain by bargaining hard, and it can less credibly threaten to reduce its contributions. The weaker nation may also enjoy disproportionate influence within an alliance because it can commit its stronger ally, which may be unable to accept losses resulting from the smaller partner's defeat (Olson and Zeckhauser 1966).

Each state brings both assets and liabilities to the alliance. The conventional manner of assessing capabilities is to sum the assets of the members. The more realistic view is that the capabilities of an alliance rarely equal the sum of its parts. Given close coordination, similar equipment, skillful leadership, and complementary needs and resources, economies of scale may be achieved. This may sometimes be the case for alliances of deterrence, but if it becomes necessary to carry out military operations, alliance capabilities are probably less than those of the individual nations combined. Wartime operations may reveal or exacerbate problems arising from poor staff coordination, mistrust, incompatible goals, logistical difficulties, dissimilar military equipment, and organization.

The capabilities of alliance leaders are important. A powerful and wealthy bloc leader can offer side payments–private goods to supplement public ones–to smaller partners, which may in turn render the alliance more effective. According to the economic theory of 'collective action,' alliances that supplement public benefits (those that are shared by all members) with private or non-collective ones are more cohesive than alliances that provide only collective benefits (Olson and Zeckhauser 1966). Except in a pure conflict situation, there are always some tensions between the requisites of alliance cohesion and broader global concerns. If the alliance is to cope effectively with these tensions, it requires effective leadership from its leading member or members. Alliance relationships involve other kinds of tensions and fears. Members face the fears of abandonment when their own interests are at stake, and of entrapment when those of others are threatened (Snyder 1997). The ultimate test of alliance leadership comes during an international crisis when the leading partner may be forced to resolve serious tensions between alliance management (intra-alliance diplomacy) and crisis management (inter-alliance diplomacy).

Virtually every alliance faces the problem of ensuring the credibility of its commitments. The problem is most serious for alliances of deterrence. If adversaries possess serious doubts on this score, the alliance may serve as an invitation to attack. Equally important, if members have doubts about the assurances of their partners, the coalition is unlikely to be effective. 'Irrational' alliance commitments may be undertaken as part of an overall strategy of increasing the credibility of deterrence by conveying to the adversary–and to other allies–that if the alliance leader is willing to expend vast resources to protect areas of little strategic value, then it should be clear that an even greater effort will be made to defend other allies (Maxwell 1968, p. 8). Buttressing credibility by 'irrational' commitments may also entail risks, forcing the alliance leader to choose between two unpalatable alternatives: reducing the commitment under threat, thereby seriously eroding one's credibility in the future; or backing the promise to the hilt, with the possibility of becoming a prisoner of the ally's policies.

Although it is widely asserted that alliance cohesion depends upon an external threat and that it declines as the danger is reduced, this generalization must be qualified. Such a threat may give rise to divisions if only part of the alliance membership feels threatened, if the threat strikes at the basis of alliance consensus, or if it offers a solution that sets off the interests of one ally against those of another. Similarly, unless the external danger creates an equitable division of labor among alliance members, cohesion and effectiveness are likely to suffer. Although wartime alliances face considerable external threat, they may sometimes experience tensions because military operations rarely result in equal burdens. As long as the threat calls for a cooperative solution, cohesion will probably be enhanced, but should a solution which favors one ally at the expense of others appear, the alliance may lose unity or disintegrate. Finally, an alliance confronting an external threat for which there is no adequate response may also experience reduced cohesion or dissolution.

Most analysts stress the negative effects of nuclear weapons for alliances. One reason is that contemporary military technology has permitted some nations to gain preponderant power without external assistance. Another strand of the argument is that nuclear deterrence will be less than credible if it takes the form of one alliance member providing a 'nuclear umbrella' (extended deterrence) for the others. Put in its starkest form the question is: what nation will risk

its own annihilation by using its nuclear capabilities as a means of last resort to punish aggression against its allies? The other side of the argument is the fear that one may unwittingly become a nuclear target as a result of an ally's quarrels.

3. International Effects of Alliances

Alliances are intended to play a central role in maintaining the balance of power. They provide the primary means of deterring or defeating nations or coalitions that seek to destroy the existing balance by achieving a position of hegemony. Even advocates of balance-of-power diplomacy attach a number of important qualifications. Alliances among great powers contribute to instability if they are strong enough to destroy the existing balance; or if they reduce the number of nations that may act in the role of 'balancers,' which, according to the theory, are supposed to remain uncommitted until there is a threat to the balance.

Alliance critics assert that they merely breed counter-alliances, thereby leaving no nation more secure, while at the same time contributing to polarization and international tensions. A further criticism is that alliances are incompatible with collective security, which requires every nation automatically to assist the victim of aggression; alliance commitments can thus create a conflict of interest. It does not necessarily follow, however, that in the absence of alliances nations will develop an effective collective security system. Indeed, alliances may arise from disappointed hopes for collective security. Elimination of alliances may be a necessary condition for collective security, but it is not sufficient. Critics of alliances also suggest that they act as conduits to spread conflict to regions previously free of it, but Liska (1962) has come to the conclusion that alliances neither cause nor prevent conflict, nor do they expand or limit it.

Another strand of theory, not necessarily incompatible with either of the above, sees alliance as a possible step in the process of more enduring forms of integration. The premise is that effective cooperation among units in one issue area gives rise to collaboration in others and, in the long run, to institutionalization of the arrangements.

4. National Effects of Alliances

Alliance benefits may include enhanced security, reduced defense expenditures, and possible side benefits such as economic aid and prestige. A strong ally may be a necessary condition for survival. But alliances may be a net drain on national resources, they may distort calculations of national interest if allies become wedded to 'inherent good faith' models of each other, and they may lead to a loss of decision-making independence.

5. Prospects

The third quarter of the twentieth century, 'the age of alliances,' was ushered in by the formation of NATO in 1949 and the Sino-Soviet Security Treaty in 1950. Within a half-decade the Warsaw Pact, SEATO, CENTO, and a multitude of other alliances were formed, but many of them had dissolved before the end of the Cold War. Contrary to the expectations of some realists, however, NATO has outlived the threat that brought it into existence. Although many Cold War alliances failed to survive the new millennium, as long as the international system is characterized by independent political units, alliances are likely to persist as a major instrument of statecraft.

See also: Balance of Power, History of; Balance of Power: Political; Foreign Policy Analysis; Globalization: Legal Aspects; Globalization: Political Aspects; Imperialism, History of; Imperialism: Political Aspects; International and Transboundary Accords, Environmental; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; International Trade: Geographic Aspects; Peace; Peacemaking in History; War: Causes and Patterns; War, Sociology of; Warfare in History

Bibliography

Allison G T 1971 *The Essence of Decision: Explaining the Cuban Missile Crisis*. Little Brown, Boston

Bueno de Mesquita B, Singer J D 1973 Alliances, capabilities, and war: a review and synthesis. In: Cotter C P (ed.) *Political Science Annual: An International Review*. Bobbs-Merrill, Indianapolis, IN, Vol. 4

Deutsch K W, Singer J D 1964 Multipolar power systems and international stability. *World Politics* **16**: 390–406

Gilpin R 1981 *War and Change in World Politics*. Cambridge University Press, New York

Holsti O R, Hopmann P T, Sullivan J D 1973 *Unity and Disintegration in International Alliances*. Wiley, New York

Liska G 1962 *Nations in Alliance: The Limits of Interdependence*. Johns Hopkins University Press, Baltimore, MD

Maxwell S 1968 *Rationality in Deterrence*. International Institute for Strategic Studies, London

Morgenthau H J 1959 Alliances in theory and practice. In: Wolfers A (ed.) *Alliance Policy in The Cold War*, Johns Hopkins University Press, Baltimore, MD

Naidu M V 1974 Alliances and balance of power. In: *Search of Conceptual Clarity*. Macmillan Company of India, Delhi

Neustadt R E 1970 *Alliance Politics*. Columbia University Press, New York

Olson M, Zeckhauser R 1966 An economic theory of alliances. *Review of Economics and Statistics* **48**: 266–79
Riker W H 1962 *The Theory of Political Coalitions*. Yale University Press, New Haven, CT
Snyder G H 1997 *Alliance Politics*. Cornell University Press, Ithaca, NY
Walt S M 1987 *The Origins of Alliances*. Cornell University Press, Ithaca, NY
Waltz K N 1979 *Theory of International Politics*. Addison-Wesley, Reading, MA

O. R. Holsti

Allport, Gordon W (1897–1967)

Gordon W. Allport (1897–1967) was a leading American personality theorist and social psychologist throughout the mid-twentieth century. His two major works—*Personality: A Psychological Interpretation* (1937) and *Pattern and Growth in Personality* (1961)—stand as landmarks in the development of personality psychology. His *The Nature of Prejudice* (1954) remains one of the most cited applied works in social psychology.

The son of a Scottish-American medical doctor, Allport was born November 11, 1897 in Montezuma, Indiana, USA and grew up in Cleveland, Ohio. His half-century affiliation with Harvard University began in 1915 when he enrolled as an undergraduate. His student career foretold his adult interests in science and social issues. He majored in both psychology and social ethics and was impressed by his first teacher in psychology, Hugo Muensterberg. He spent much of his spare time in social service activities. This convergence of interests took institutional form later when he helped to establish both the Society for the Psychological Study of Social Issues and Harvard's Department of Social Relations.

Upon graduation in 1919, Allport taught English and Sociology at Robert College in Constantinople (then part of Greece, now Bogazici University in Istanbul, Turkey). In addition to German, he remained fluent in modern Greek throughout his life. Returning to Harvard in 1920, he completed his Ph.D. in psychology in two years. His dissertation title again reflected his dual commitment to science and social concerns: *An Experimental Study of the Traits of Personality: With Special Reference to the Problem of Social Diagnosis*. In addition, he assisted his older brother, Floyd, in editing the *Journal of Abnormal and Social Psychology*, the start of over four decades of association with the publication.

Harvard then awarded him a coveted Sheldon Travelling Fellowship. He spent the first year in Germany, where the new *Gestalt* school and its emphasis on cognition fascinated him. Indeed, he became a partial Gestaltist—partial because he could not accept the Gestaltists' assumptions about the inflexible genetic basis of cognitive processes (Pettigrew 1979). He spent his second Sheldon year at Cambridge University, where the psychologists coolly received his reports on *Gestalt* developments.

In 1924, Allport became a Harvard instructor in social ethics and taught what may have been the first personality course offered by a North American university. Two years later he temporarily severed his connection with Harvard to accept an assistant professorship in psychology at Dartmouth College. Yet, even during his brief four years away, he returned repeatedly to Harvard to teach in summer school. In 1930 he came back to Harvard to stay.

Allport's unique contributions to psychology are best described by three interwoven features of his work. First, he offered a broadly eclectic balance of the many sides of the discipline—holding to William James's contention that there were 'multiple avenues to the truth.' Second, he formulated the central future problems of the discipline and proposed original approaches to them. Finally, his entire body of scholarly work presents a consistent, seamless, and forceful perspective.

1. *Broadly Eclectic Balance*

Allport sought an eclectic balance for both methods and theory. His famous volumes on personality illustrate this dominant feature of his work (Pettigrew 1990). He urged, for example, the use of both ideographic (individual) and nomothetic (universal) methods. Since he thought the discipline relied too heavily on nomothetic approaches, he sought greater use of ideographic techniques. In an age of indirection, Allport insisted, 'If you want to know something about a person, why not first ask him?' Considered scandalously naive when he introduced it, his position helped to right the balance in assessment.

Typically, this was an expansionist, not an exclusionist, view. He simply sought a reasonable trade-off between accuracy and adequacy. He thought the two approaches together would make for 'a broadened psychology.' Indeed, his own empirical efforts ranged from personal documents, such as *Letters From Jenny* (1965), to two popular nomothetic tests on ascendance-submission and personal values. He also developed ingenious experimental procedures to study eidetic imagery, expressive movement, radio effects, rumor, the trapezoidal window illusion, and binocular rivalry.

His theoretical efforts also sought a balance. He vigorously advocated an open-system theory of personality with emphases on individuality, proaction, consciousness, maturity, and the unity of personality. As such, in Rosenzweig's (1970, p. 60) view, Allport served as the field's 'ego' in contrast to Henry Murray's

'id' and Edwin Boring's 'super-ego.' Yet Allport did not hold his emphases to be the only matters of importance for a psychology of personality. Rather, it was because he believed the discipline was granting these important ego-concerns too little attention.

In this sense, Allport was, to borrow from boxing, a counterpuncher. He opposed what he regarded as excessive trends in psychology that threatened his conception of an open, balanced discipline. In 1937, in his first personality book, he saw as the major threat the too-rigid applications of experimental psychology. By 1961, in his second personality volume, he saw as the major threat the too-loose applications of psychoanalysis. 'Although much of my writing is polemic in tone,' he confessed in his autobiography, 'I know in my bones that my opponents are partly right.' The key word is *partly*. Allport opposed excess, 'the strong aura of arrogance found in … fashionable dogmas' (1968, pp. 405–6). So he held fast to the open middle ground as he perceived it, and aimed his punches at the 'fashionable dogmas' that existed in each period. Modern readers may miss the significance of his arguments if they are unaware of which dragons he is attempting to slay.

2. *Formulating Central Problems and Offering Original Solutions*

Psychology recognized Allport throughout his career as a source for specifying the discipline's central problems. Typically his initially proposed solutions to the problems—such as functional autonomy—won limited acceptance. Nonetheless, many of Allport's initial proposals for addressing basic problems now exist in psychology with new labels and enlarged meanings. Allport only loosely sketched out his innovative ideas. Later work accepted the problem and expanded the ideas. Consider the much derided concept of functional autonomy. The notion that motives can become independent of their origins was considered heretical in 1937. Slowly, psychology came to accept the phenomenon if not the formulation. Today social psychologists reconceptualize the process in interactionist terms. Motives, established and functional in one situation, help lead individuals to new situations where the same motives persist but assume new functions.

Similarly, Allport's conception of personality traits has often been criticized. What critics attack is the mistaken notion that he held a static view of traits as pervasive, cross-situational consistencies in behavior. But Zuroff has shown that Allport advanced a far more dynamic conception of traits. In fact, he was 'an interactionist in the sense that he recognized behavior is determined by the person and situation' (Zuroff 1986, p. 993).

Indeed, Allport broke early ground for many ideas now fully developed and accepted. Thus, he provided in his 1937 volume a social constructionist interpretation of identity. His insistence on multiple indicators and methods offered an initial statement of Campbell and Fiske's (1959) multitrait–multimethod approach.

3. *A Consistent, Forceful Perspective*

Above all, Allport's contributions to psychology flowed from a consistent and forceful perspective presented in graceful prose. One reviewer of his 1937 *Personality* book wrote, 'One has all the way through it a distinct feeling that "This is Allport"' (Hollingworth 1938, p. 103). This pointed observation holds true for all his writing.

Allport's perspective remained consistent but not static throughout his career. He acquired a comprehensive knowledge of the psychological literature from his long years as a meticulous editor. Quite literally, a large proportion of North America's personality and social psychological literature crossed his editor's desk. His mastery of the literature also reflected his open-ended view of theory—a view more Popperian than the strict Vienna-circle positivism that held sway throughout most of his career.

Yet Allport held to his perspective forcefully. His writing conveyed this forcefulness. Blunt prose and forthright critiques characterize his style. As one disgruntled reviewer of an Allport book put it, 'There is something in it to irritate almost everyone' (Adelson 1956, p. 68). Hall and Lindzey (1957) suggest that Allport's many years of teaching led to his expressing his views in a uniquely salient and provocative style.

4. *Allport on Prejudice*

This directness is also apparent in his applied volume, *The Nature of Prejudice*. This influential book again brought together Allport's two sides—science and social action. He deemed it his proudest achievement because he thought it 'had done some good in the world.' Indeed, in its paperback edition, it became one of the best selling social psychological books in publishing history.

The Nature of Prejudice displays the special characteristics of Allport's contributions to psychology. The volume offers a broad, eclectic perspective with a lens model that ranges from history to the psychological effects of prejudice on its victims. Its open view of the many sides of the phenomenon again demonstrates the distinctive quality of Allport's wide-ranging thought.

The Nature of Prejudice is another seamless work that is 'Allportian' from start to finish. Deftly crafted in a simpler style than his other writings, Allport strove to make it accessible to a wider, nonacademic audience. In addition, it served to structure the entire study of prejudice in social psychology for the next

four decades. This is true for intergroup research in North America. In its many translated versions, it has proven highly influential in other parts of the world as well. Indeed, the volume's many predictions, though shaped by American intergroup data, have typically been confirmed in studies of intergroup prejudice throughout much of the world.

The Nature of Prejudice's useful definition of prejudice—'an antipathy based upon a faulty and inflexible generalization' (1954, p. 9)—stressed both affective and cognitive components, but it wisely left the complex link between prejudice and behavior an open, empirical question. This definition has served the field well, and only recently have social psychologists advanced expansions of its terms.

To attain his distinctive quality of balance, Allport again assumed the role of counterpuncher against prevailing dogmas. He gave full recognition to the importance of the psychoanalytically inspired authoritarian personality syndrome, which he also had uncovered during the 1940s. But he challenged the Freudian formulation of aggression with a rival theory. Instead of the psychoanalytic steam boiler model of aggression and catharsis, Allport proposed a feedback model with dramatically different implications for prejudice and its remediation. Aggression, he argued, feeds on itself. That is, the acting out of aggression, rather than leading to less aggression, actually increases the probability that further aggression will be expressed. Armed with this insight, *The Nature of Prejudice* proceeds to advocate policies that have indeed reduced levels of prejudice in the United States and elsewhere.

Allport also challenged the central assumption of one of his own favorite groups. The Human Relations Movement developed after World War II to improve America's intergroup relations. With Brotherhood Weeks and Dinners, the well-meaning movement hoped to combat prejudice and discrimination through formal intergroup contact. But Allport, in the book's most important theoretical contribution, questioned this assumption with his intergroup contact hypothesis. Contact alone, he argued, only set the scene for change; what mattered were the situational conditions of the intergroup interaction. The four conditions he listed—equal status in the situation, common goals, no intergroup competition, and authority sanction—have repeatedly been supported in research around the globe (Pettigrew 1998).

His treatment of prejudice also countered the then-fashionable assumption that group stereotypes were simply the aberrant cognitive distortions of prejudiced people. Advancing the view now universally accepted, Allport held that the cognitive components of prejudice were natural extensions of normal processes. Stereotypes and prejudgment, he concluded, were not aberrant at all, but unfortunately all too human.

Allport's foresight into the many advances, especially in research on group stereotypes, that this field has achieved over recent decades can be traced to his early *Gestalt* leanings (Pettigrew 1979). He devoted 10 of *The Nature of Prejudice*'s 31 chapters to cognitive factors. Psychology joined the cognitive revolution just after the volume was published. In social psychology, social cognition veered in a largely *Gestalt* direction that had molded Allport's perspective on prejudice. Thus, the same influences that shaped the study of prejudice in general and stereotypes in particular from 1960 on had earlier guided Allport's thought.

The Nature of Prejudice also presents a host of original hypotheses on specific topics that have stood the test of time. For instance, one popular theory of prejudice reduction emphasizes recategorization through identity with larger, more inclusive groups. Allport (1954) advocated precisely the same mechanism. Drawing concentric circles with family in the center and humankind at the periphery, he argued that 'concentric loyalties need not clash' and that prejudice is minimized by inclusive group membership.

The volume also devotes an entire chapter to the link between religion and prejudice. A religious man himself, Allport was disturbed that research routinely finds nonbelievers far less prejudiced on average than members of organized religions. He proposed a critical distinction between an 'institutionalized' religious outlook and an 'interiorized' one. The more numerous institutionally religious, he argued, are the highly prejudiced. Those of the interiorized type, who have deeply internalized their religious beliefs, are far less prejudiced. In his last empirical publication (Allport and Ross 1967), Allport presented additional evidence in support of his hypothesis and later research provides additional support.

Allport addressed the book primarily to his own ingroup—White, Protestant, American males. The examples of prejudice cited throughout involve anti-Black, anti-Jewish, anti-Catholic, and antifemale sentiments. He was clearly lecturing to 'his own kind.' It is safe, easy, and politically expedient to attack the prejudices of outgroups who hold negative views of our ingroup. It is quite a different matter to attack the prejudices of our own ingroup toward others. So yet another remarkable feature of *The Nature of Prejudice* is its target audience.

5. *Allport as Teacher and Mentor*

Though intellectually confident, Allport was actually a shy and modest man—though his manner was sometimes mistaken for aloofness. In his gentle and supportive way, he was a master teacher and mentor. He held to his conception of the uniqueness of personality and encouraged students to follow their own interests. Hence, though he produced dozens of well-known psychologists, he never developed a 'school' of followers.

One way Allport handled his social shyness was to prepare carefully for occasions in advance. Before giving his Hoernle Lecture at South Africa's leading Afrikaner University at Stellenbosch in 1956, he studied Afrikaans with a tutor for six months. He then skillfully gave the introduction to his lecture in Afrikaans, gracefully apologizing for not delivering his entire address in Afrikaans. The Afrikaner audience, accustomed to even other South Africans not knowing their language, reacted with surprise and delight. The audience arose as one with applause.

6. *The Lasting Contribution*

Allport's professional honors were many. He was elected president of both the American Psychological Association (1939) and the Society for the Psychological Study of Social Issues (1944). He received the Gold Medal of the American Psychological Foundation in 1963. He died October 9, 1967 in Cambridge, Massachusetts. His legacy in psychology is twofold. He helped to establish personality psychology as a science and as an integral part of the discipline of psychology. In addition, his applied work in social psychology, particularly on intergroup prejudice, furthered the practical value of the discipline's work for important social issues.

See also: Attitudes and Behavior; Cattell, Raymond Bernard (1905–98); Jung, Carl Gustav (1875–1961); Köhler, Wolfgang (1887–1967); Personality and Adaptive Behaviors; Personality and Social Behavior; Personality Psychology; Personality Structure; Personality Theories; Prejudice in Society; Psychology: Historical and Cultural Perspectives; Racial Relations; Social Psychology; Social Psychology, Theories of

Bibliography

Adelson J 1956 On man's goodness [Review of *Becoming: Basic considerations for a psychology of personality*]. *Contemporary Psychology* **1**: 67–9

Allport G W 1937 *Personality: A Psychological Interpretation*. Holt, Rinehart & Winston, New York

Allport G W 1954 *The Nature of Prejudice*. Addison-Wesley, Reading, MA

Allport G W 1961 *Pattern and Growth in Personality*. Holt, Rinehart & Winston, New York

Allport G W 1965 *Letters from Jenny*. Harcourt, Brace & World, New York

Allport G W 1968 *The Person in Psychology: Selected Essays*. Beacon Press, Boston

Allport G W, Ross J M 1967 Personal religious orientation and prejudice. *Journal of Personality and Social Psychology* **5**: 432–43

Campbell D T, Fiske D W 1959 Convergent and discriminate validation of the multitrait–multimethod matrix. *Psychological Bulletin* **56**: 81–105

Hall C S, Lindzey G 1957 *Theories of Personality*. Wiley, New York

Hollingworth H L 1938 Review of personality. *Psychological Bulletin* **35**: 103–7

Pettigrew T F 1979 The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin* **5**: 461–76

Pettigrew T F 1990 A psychological interpretation—Allport, G W. *Contemporary Psychology* **35**: 533–6

Pettigrew T F 1998 Intergroup contact theory. *Annual Review of Psychology* **49**: 65–85

Rosenzweig S 1970 E. G. Boring and the Zeitgeist: *Eruditione gesta beavit*. *Journal of Psychology* **75**: 59–71

Zuroff D C 1986 Was Gordon Allport a trait theorist? *Journal of Personality and Social Psychology* **51**: 993–1000

T. F. Pettigrew

Alternative and Complementary Healing Practices

Complementary medicine (CAM) has become increasingly popular in the Western world since about 1975 and has excited much research (Eisenberg et al. 1998). Paradoxically, this increase of interest in and the use of complementary medicine (CAM) comes at a time when the successes of the scientifically based contemporary biomedicine or orthodox medicine (OM) have never been greater.

Perhaps the different terminologies used over the years to describe complementary medicine therapies best indicate its greater acceptance both in the orthodox and medical community and also among the lay public. Thus the term 'fringe' medicine developed into 'alternative,' then 'unconventional' and finally 'complementary' medicine. Since about 1980 there has also been an enormous growth of, and interest in, all forms of CAM (Vincent and Furnham 1997, 1999). Demand for CAM has been matched by supply, and there is now a substantial list of CAM therapies available to Western metropolitan citizens (Ernst et al. 1997).

1. *Fundamental Differences Between CAM and OM*

Aakster (1986) made the following five distinctions.

(a) Health

Whereas conventional medicine sees health as an absence of disease, alternative medicine frequently mentions a balance of opposing forces (both external and internal).

(b) Disease

The conventional medicinal interpretation sees disease as a specific, locally defined deviation in organ or tissue structure. CAM practitioners stress many wider

signs, such as body language indicating disruptive forces and/or restorative processes.

(c) Diagnosis

Regular medicine stresses morphological classification based on location and etiology, while alternative interpretations often consider problems of functionality to be diagnostically useful.

(d) Therapy

Conventional medicine often aims to destroy, demolish or surpress the sickening forces, while alternative therapies often aim to strengthen the vitalizing, health-promoting forces. CAM therapies seem particularly hostile to chemical therapies and surgery.

(e) Patient

In much conventional medicine the patient is the passive recipient of external solutions, while in CAM the patient is an active participant in regaining health.

The history, philosophy, and methods of treatment of even the main forms of complementary therapy are extremely diverse. The origins of some, for example acupuncture, are ancient, while osteopathy and homeopathy date from the nineteenth century. Some (acupuncture, homeopathy) are complete systems of medicine, while others are restricted to diagnosis alone (iridology) or to a specific therapeutic technique (massage). The range of treatments is equally varied: diet, plant remedies, needles, miniscule homeopathic doses, mineral and vitamin supplements, and a variety of psychological techniques. The theoretical frameworks and underlying philosophy vary in coherence, complexity, and the degree to which they could be incorporated in current scientific medicine.

Complementary practitioners vary enormously in their attitude to orthodox medicine, the extent of their training, and their desire for professional recognition. However, there are within this diversity, some broad common themes: a vitalistic philosophy embracing the idea of an underlying energy or vital force; a belief that the body is self-healing, and so a respect for minimal interventions; general, all-encompassing theories of disease, and a strong emphasis on the prevention of disease and the attainment of positive health. While in much conventional medicine the patient is the passive recipient of external solutions, in complementary medicine the patient is more likely to be an active participant in regaining health (Vincent and Furnham 1997).

2. *The Popularity of CAM*

Surveys indicate that between 1986 and 1991 the proportion of the British population using CAM increased by 70 percent (British Medical Association 1993, Fulder and Munro 1985). A similar trend has also been noted in other parts of Europe as well as Australia and New Zealand. The same trend has been recorded in the USA. Eisenberg et al. (1998) found, by extrapolating from their survey to the population as a whole, a 47.3 percent increase in total visits to American CAM practitioners from 427 million in 1990 to 629 million in 1997, actually exceeding visits to primary-care physicians.

In Europe, surveys suggest that a third of people have seen a complementary therapist or used complementary remedies in any given year. The popularity of CAM in Europe is growing rapidly. In 1981, 6.4 percent of the Dutch population attended a therapist or doctor providing CAM, and this increased to 9.1 percent by 1985 and 15.7 percent in 1990. The use of homeopathy, the most popular form of complementary therapy in France, rose from 16 percent of the population in 1982 to 29 percent in 1987 and 36 percent in 1992.

CAM includes a wide range of therapeutic practices and diagnostic systems that stand separate from, or in some cases opposed to, conventional scientifically based modern medicine. Working definitions of CAM often rest on the fact that complementary medicines do not exert their legitimacy with reference to scientific claims and empirical regularity. Many types of CAM work within their own philosophical paradigms using alternative models to explain health and illness and thus often remain resistant to testing within the biomedical paradigm in order to prove their effectiveness (Vincent and Furnham 1997).

A British Medical Association report (1986) listed 116 different types of complementary therapy and diagnostic aids. It is possible to classify the specialties along a number of dimensions: their history; the extent to which they have been professionalized; whether or not they involve touch; and the range and type of disorders/problems they supposedly cure. The list of recognized CAM therapies continually grows.

3. *Different Perspectives on CAM*

Gray (1998, p. 55) argues that 'the topic of unconventional therapies can no longer be ignored or marginalized because, for better or worse, each seriously ill person cannot help but be confronted with choices about their possible usage.' He believes there are currently four quite different and debatable perspectives on complementary medicine. The 'Biomedicine perspective' which is concerned with the curing of disease and the control of symptoms where the physician-scientist is a technician applying high level skills to physiological problems. This approach is antagonistic toward and skeptical of CAM, believing many claims to be fraudulent and many practitioners unscrupulous. Physicians and medical scientists within this camp often believe CAM patients are naïve, anxious, and neurotic. However, the competitive health care market place has seen a shift even in the attitudes of 'hardliners' to being more interested in, and sympathetic to, unconventional therapies.

The 'Complementary perspective,' though extremely varied, those with this perspective do share certain ideas, such as: (a) rating the importance of

domains other than the physical for understanding health; (b) viewing diseases as symptomatic of underlying systemic problems; (c) a reliance on clinical experience to guide practice; and (d) a cogent critique of the limits of the biomedical approach. Interventions at the psychological, social, and spiritual level are all thought to be relevant and important, supporting the idea of a biopsychosocial model. Many advocates of this perspective believe the body has powerful natural healing mechanisms that need to be activated. They are critical of biomedicine's harsh and often unsuccessful treatments, especially with cancer. They attack biomedicine for not having most of its own interventions based on 'solid scientific evidence.'

The 'Progressive perspective' is prepared to support either of the biomedicine or complementary perspectives, mentioned previously, depending on the scientific evidence in favor. They are hardened empiricists who believe it is possible to integrate the best of biomedicine and unconventional approaches. Like all other researchers their approach is not value free—the advocates of this approach welcome the scientific testing of all sorts of unconventional therapies.

The 'Postmodern perspective' enjoys challenging those with absolute faith in science, reason, and technology, and deconstructing traditional ideas of progress. Followers are distrustful of, and cynical toward, science, medicine, the legal system, and institutionalized religion, and even parliamentary democracy. Postmodernists abandon all worldviews and see truth as a socially and politically constructed issue. Many believe orthodox practitioners to be totalitarian persecutors of unconventional medicine. They rejoice in, and welcome multiple perspectives and 'finding one's own voice.' However, because many CAM practitioners can be theoretically convinced of their position and uncompromising, they can also be subject to postmodern skepticism. Proponents of this position argue: (a) to have a complementary perspective in any debate is healthy; (b) that CAM practitioners are also connected to particular economic and theoretical interests; (c) that the variety of values and criteria for assessing success is beneficial; and (d) that the ill people themselves should be the final arbitrators of the success of the therapy.

4. *Research into CAM*

The use of alternative models, which seem incompatible with the scientific model, is one of the main sources of controversy. The effectiveness of a medicine is usually assessed using double-blind randomized-controlled trials (RCT) to prove its efficacy beyond a placebo effect (nonspecific healing mechanism) (Ernst et al. 1997). However, scientific criteria often fail to validate complementary medicines in terms of reliable and predictable outcomes, which have withstood rigorous clinical trials. Although some clinical studies do provide strong evidence for the effectiveness of some complementary therapies, the evidence is not always enough to prove their efficacy. Some have questioned the methodology of these studies, others suggest a revision of the biomedical paradigm, and most recognize this area of research is in need of more attention and resources.

Despite the lack of conclusive evidence for the efficacy of CAM, it undoubtedly remains perceived as effective, indicated by the high levels of consumer satisfaction. The efficacy of complementary medicine from a biomedical point of view, and from a lay person's point of view is, to a very large extent, two different matters. Hence research interest into the perceived efficacy of CAM and OM by members of the general public.

Furnham and Vincent (2000) have argued that there are two fundamental scientific questions concerning CAM: first, does it work? Second, why do people choose it? The first question is concerned with the quality and quantity of disinterested evidence that CAM therapies cure or 'manage' physical and psychological illnesses by the processes and mechanisms they maintain (and not by placebo effects). Second, if the evidence is limited and highly equivocal, why do so many people seek it out and pay for it?

Because of the expense and difficulty of doing scientifically satisfactory research to answer the former question, much more attention has been paid to why patients choose CAM. Various specific hypotheses have been tested, some to do with the pull of CAM (health beliefs and circumstances) and some with the push of OM (dissatisfaction with GPs). Sociologists have stressed the importance of post-modernist beliefs (reflexivity), consumer and patient rights, and holistic movements to explain the attraction of CAM. Economists have tried to account for it in terms of costs to patients, doctors, and health insurance companies. Psychologists and psychiatrists have preferred interpersonal and intrapsychic explanations while medical practitioners have focused on the flight from science beliefs as well as the benefits of longer consultations.

The growth in the USA of CAM has been paralleled by an increase in multidisciplinary research in the area. The whole question of scientific proof, as well as the pathways to care and cure, certainly merits good interdisciplinary and international research.

See also: African Studies: Health; Culture as a Determinant of Mental Health; Healing; Health: Anthropological Aspects; Health Behaviors; Health Psychology; Medicine, History of; South Asian Studies: Health; Southeast Asian Studies: Health

Bibliography

Aakster C 1986 Concepts in alternative medicine. *Social Science and Medicine* **22**: 265–73

British Medical Association 1986 *Alternative Therapy*. Oxford University Press, Oxford, UK

British Medical Association 1993 *Complementary Medicine*. Report of the Board of Science and Education. Oxford University Press, Oxford, UK
Eisenberg D, Davis R, Ettner S, App S, Wilkey S, Van Rompay A, Kessler R 1998 Trends in alternative medicine in the United States, 1990–1997: Results of a follow-up national survey. *Journal of the American Medical Association* **280**: 1569–73
Ernst E, Seivner I, Gamus D 1997 Complementary medicine—A critical review. *Israel Journal of Medical Sciences* **33**: 808–15
Fulder S J, Munro R E 1985 Complementary medicine in the United Kingdom: Patients, practitioners, and consultations. *Lancet* Sep **7**(2): 542–5
Furnham A, Vincent C 2000 Reasons for using CAM. In: Kelner M, Wellman B (eds.) *Complementary and Alternative Medicine: Challenge and Change*. Harwood, Amsterdam, pp. 61–78
Gray R 1998 Four perspectives on unconventional therapy. *Health* **2**: 55–74
Vincent C, Furnham A 1997 *Complementary Medicine: A Research Perspective*. Wiley, Chichester, UK
Vincent C, Furnham A 1999 Complementary medicine: State of the evidence. *Journal of the Royal Society of Medicine* **92**: 170–7

A. F. Furnham

Alternative Media

The designation 'alternative' is vague because it covers so many different formats and only defines these media by their difference from others. The definition used here is of small-scale, politically radical media, occupying a technological spectrum from fliers to the Web, from video to graffiti, from political song to satirical cartoons. Much of what follows will offer an historical and international survey of their roles. The review will conclude by very briefly considering the relation of these media to mainstream media, the relation between them and social movements, and the argument that they are significant in any sociological approach to media.

1. Alternative Media in History and Across the Planet

As contrasted with typical dismissal of this category of media as evanescent, petty, and ultimately irrelevant, reflection on their long and varied history in many nations and situations suggests quite otherwise. Consideration of their roles will be divided in what follows into (a) the Protestant Reformation, the English Civil War, and the American and French revolutions; (b) the US abolitionist, suffragist and labor movements; (c) Leninist political movements; (d) *samizdat* and Western shortwave radio in the later decades of the former Soviet bloc; (e) rightwing extremist alternative media; (f) 'nonmedia' alternative communication forms; and (g) the energetic attempts by state and religious authorities to suppress alternative media.

1.1 From the Reformation to the French Revolution

Martin Luther's 95 theses are only the best known example today of the 'pamphlet war' that raged in some German territories in the early 1500s. His was certainly not the most radical; one Leipzig printer, Hans Hergot, was executed for his 1527 pamphlet urging that land be held in common. But following upon the printing of the vernacular Bible, these pamphlets, many (though not all) urging religious change, played a considerable part in extending and strengthening movements of opposition to papal authority.

In the English Civil War of the 1640s, the subversive influence of the vernacular Bible and its empowerment of popular readings of the scriptures could be seen in the flood of publications by Ranters, Levelers, Diggers and other movements that claimed divine authority for their revolution against common land enclosures, despotic royal power, and Cromwell's repression. The overthrow of the monarchy and the establishment of a republic that lasted over a decade, the only one in the last 1000 years of British history, were partial echoes of these upsurges. But the media dimension of these movements was central to their impact.

In the US revolution, not only Tom Paine's famous *Common Sense* pamphlet attacking 'the Royal Brute of Great Britain,' but also a plethora of others, made their contribution to the Declaration of Independence and its armed defense. As in the German Pamphlet War, the publications took both sides in the conflict, but those that expressed the new voice of the colonists had the advantage of freshness and anger.

In the buildup to the French revolution, similarly, radical print publications played a very visible role. In order to escape the censor and the monarchy's police agents, their printing often took place just outside France's borders and then the works themselves were smuggled into the country. The content was antimonarchical and often anticlerical, sometimes expressed in ribald satirical depictions of those in power.

1.2 Media of the Abolitionist, Suffragist and Labor Movements

Focusing on the nineteenth century USA, the roles of media in these three highly influential movements are conspicuous. The earliest instances of abolitionist media were generally autobiographies of formerly enslaved Africans, either escapees or freedmen, quite often former mariners whose occupation had given them scope to travel, even in some cases to independent

Haiti. This experience both enabled them to understand the Atlantic economic and political nexus, and permitted a more wide ranging communication with fellow Africans than was the rule in the Americas. Later, escapee Frederick Douglas' autobiography and the newspaper he edited, *The Northern Star*, also played a very significant role. While many other economic and political factors played their role in tandem, these alternative media exercised considerable agency in consolidating determined opposition to the institution of slavery.

The suffragist movement, beginning from the 1848 Seneca Falls convention, made consistent use of pamphlets, tracts, cartoons, and other media to propagate its cause. While some of the cartoons and some of the propaganda romanticized the civilizing impact that women's suffrage would have upon the polity, this does not detract from their political significance. Temperance movements, for example, often connected to suffragist campaigns, gave women a sense of authority and commonality that both fitted with their supposed role as exemplars of morality and enabled them to escape the mutual segregation often entailed by domesticity. Meanwhile, as Virginia Woolf once put it, it was more and more the case that 'the middle class woman began to write,' a further mediatic step in the direction of emancipation.

The labor movement was the third great matrix of alternative media in the nineteenth century. In its earlier decades, pro-labor newspapers were part of the general hubbub of a fiercely partisan press in Philadelphia, New York and other emergent industrial centers. They were sidelined during the middle of the century by the emergence of the penny newspaper and the costs of the rotary press, but then the great labor migrations from Europe of the 1880s through to the 1920s generated a huge new alternative print sector. This partly reflected the ethnic and linguistic composition of the new labor force, but a significant part simultaneously gave voice to labor's political aspirations. For the most part the English language labor press by the end of the century represented skilled labor aristocrats, often conservative in their defense of their status, while the foreign language press expressed the perspectives of the newcomers, quite often socialist or anarchist in outlook.

1.3 Leninist Political Movements of the Twentieth Century

For much of the twentieth century, especially in nations convulsed by political strife such as Russia, China, Vietnam, India, and Indonesia, but beginning in Russia, the Leninist model of alternative media has held extraordinary sway. Effectively a model devised originally to outwit and defeat the Tsar's secret police and informer network, it became for many decades the 'scientific' procedure to organize effective oppositional media (and, in nations that experienced a Communist regime, a postrevolution media template as well).

The essential character of the Leninist media strategy was one in which alternative media were purely the transmission belt for the Communist Party leadership to diffuse its views downwards and outwards. These views might be on revolutionary tactics and strategy, on current political crises, on the global economy, on literature, in as much as Marxism always had totalizing pretensions. The notion of such media forming a general public forum in which different political viewpoints could be assigned the same moral status, was absolutely foreign to this vision. The more open the procedures, the more 'liberal' the debate, the more tactical political confusion would reign and the more vulnerable would be the newspaper's distributors and editors to police raids, exile and imprisonment.

It has to be said that this vision of alternative media as an organizational weapon was a remarkably effective survival strategy in highly polarized and repressive situations. It also helped lend Leninist political groups a hard-headed image that, faced with the menace of fascism, say, seemed tremendously attractive in contrast to endlessly debating 'academic' circles. Paradoxically, in sovietized Poland, Catholic parishes ran similarly tightly hierarchical parish organizations, for fear that more open ones would lend themselves to infiltration by the secret police. The tragic impact of Leninist media lay overwhelmingly in their perpetuation after state power had been grasped.

1.4 Samizdat and Clandestine Radio in the Collapse of Soviet Power

Samizdat literally means 'self-published' and begins to make more sense if we contrast it with the standard printed notice in any Soviet book, magazine or newspaper. This notice read '*gosizdat*,' state-published. In that tiny change of letters was concealed a radical switch of realities, the assertion of citizens' right to create their own media independently of Soviet control.

This was not an easy, or safe, process. Not easy, because even the simplest reproductive print machines such as typewriters, and later photocopiers, were licensed by the state. Typing paper was hard to come by, and large purchases, if they could be made, attracted unwanted police attention. It was also unsafe, because while legally up to nine carbon-paper copies were allowed to be made of any document, the Soviet penal code had many other clauses that gave the KGB, the political police, every rationale to repress either the author, the reader or the carrier of *samizdat*.

Russian *samizdat* publications began as a desperate response to the 1960s crackdown on Soviet dissidents, and especially after the climactic 1968 Soviet invasion of Czechoslovakia and repression of the Prague reformers. They consisted initially of typed sheets of

paper, with no margins and no blank space. Most people read blurry carbon copies. The price for being allowed to read one was to agree to retype it with another nine copies. Hardly a designer's dream, but they were generally snapped up eagerly because there was nothing else except *gosizdat*. The bulk were either nationalist—i.e., from Ukraine and the other (then) Soviet republics—or religious in character, sometimes both. The *samizdat* best known outside the USSR, though, was mostly secular and non-nationalist, dealing with other banned topics, or permitted topics in banned ways.

The second phase of *samizdat* came in the shape of whole books printed outside the USSR and smuggled back (*tamizdat*), and in forbidden cassette tapes, first audio and later video (*magnitizdat*). The audiotapes were particularly easy to hide, and the most popular in this phase reproduced the banned lyrics of Russia's dissident guitar poets.

The third phase was the expansion of *samizdat* in Eastern European Soviet bloc countries. Especially in Poland, the most populous and restive nation in the bloc, with its huge Solidarity labor movement of the 1980s, these underground media were very prevalent, despite all the government's attempts to repress them. Even when in panic in 1981 the government declared martial law, jailed many activists, and tracked down a lot of underground printing equipment, the production of *samizdat* continued regardless. By the mid-1980s, with the Gorbachev reform faction in charge in Moscow, bloody repression receded without retiring altogether, and conditions matured for Poles once again to assert themselves against Soviet domination. On the northern Polish border was Lithuania, historically united with Poland, which became the first republic to break away from the USSR. The steady spread of Polish *samizdat* media and the growth of independence expression in both nations was clearly linked, and bore fruit in Poland's election of its first postwar non-Communist government in 1989, and Lithuania's secession in 1991. The mighty Soviet empire buckled, and while once again, many economic and other factors played their part, *samizdat* media were intimately interwoven in the process.

Equally involved were the West's shortwave radio stations, the BBC World Service, Voice of America, Deutsche Welle, Radio Liberty and Radio Free Europe. Not only did 'the voices,' as they were known, supply alternative news whenever the Soviet bloc did not jam them, they also broadcast *samizdat* writings, thus making them available outside the major cities where typically their distribution was most concentrated. This amplification was extremely important.

1.5 Rightwing Extremist Alternative Media

Inclusion of these media is not for some abstract political balance. It is to underscore the sociological prevalence and power of small-scale radical media. Here the focus will be on the USA, but such media are very well known in Europe, not only in the early phases of those fascist movements that later came to power in Italy, Germany and elsewhere, but also in the second half of the twentieth century after the defeat of fascism.

In the USA, however, while Father Coughlin and Gerald Smithwere noted antisemitic and antilabor radio propagandists in the 1930s and 1940s, the second half of the century saw a large expansion in rightwing extremist alternative media. Initially in the 1950s this was a response to the slow but real advance of civil rights legislation, and was organized by the Klu Klux Klan and similar bodies. By the 1980s they were among the earliest users of the Internet for internal communication and propaganda.

As time went by, white supremacist voices found themselves not the only alternative media activists on the extremist right. The growth of the Christian Right in the 1970s, in part sharing supremacist views, but with a series of fundamentalist religious and moral priorities as well (such as total opposition to abortion rights), widely expressed itself in radio and not long afterwards on television as well. By the mid-1990s, around 10 percent of US radio stations were broadcasting Christian Right programming. Televangelism became a burgeoning profession.

Close on their heels in the 1980s were those sometimes self-described as militia or patriot groups, frequently but not universally white supremacist, antisemitic and anti-immigrant, and possessed of a pathological paranoia and hatred of government institutions beyond the county level. They frequently were convinced that a One World Government, managed by the UN, had already infiltrated the US government and was on the edge of supplanting it. They justified their obsession with arming themselves to the teeth, on occasion attacking and killing local law enforcement officers, and even going so far as to blow up a Federal building in Oklahoma City, by reference to their inalienable rights as citizens guaranteed by the miraculously donated US constitution. They often used shortwave radio for communicating their poison, afraid that Internet use would enable the authorities to track them down.

Assessing the influence of these alternative media depends very much on an evaluation of the general relation between them and the more mainstream conservative Right. Mutual interdictions and contempt do not necessarily tell us very much. Weber observed how, among religious groups, those closest to each others' sectarian beliefs are often the most ready to spy betrayal of some fundamental principle. Furthermore, with the growth of the conservative Right in general and the decline of the organized Left in the last two decades of the twentieth century, any attempt to pin down an unchanging anchor point in the political spectrum by which this relationship could

be assessed seems incognizant of this secular trend toward more conservative positions. However, the closer the practical links between various conservative vantage points, the more likely are we to see the impact of rightwing extremist media in public policy.

1.6 'Nonmedia' Alternative Communication Forms

These comprise a huge array. Graffiti, murals, posters, street theatre, dance, political song, sermons, dress, lapel-pins, festivals, street demonstrations, strikes, building-occupations, funeral parades, hunger-strikes, internet flaming, and the array of semi-covert resistance tactics that political scientist (James C. Scott 1985) once termed 'the weapons of the weak' are all part of this compendium. It is important to understand their effective interaction with more conventionally 'mediatic' alternative media. In practice, opposition expresses itself culturally by all the means it can, as the full story of the Iranian revolution of the late 1970s shows very clearly (Sreberny-Mohammadi and Mohammadi 1994).

1.7 Why Do State and Religious Authorities Seek to Suppress Alternative Media?

The question is a reasonable one, since this is indeed a very familiar pattern. The answer can only be that they fear them. This then raises a further, perhaps paradoxical question: is their fear based upon paranoia, or reason? If the former, then the conventional dismissal of alternative media as of trivial import seems to have some underpinning. If however paranoia is implausible on such a widely diffused scale, then a reverse conclusion suggests itself, namely that such media may indeed be of much more significance than researchers have often ascribed them. Perhaps, for all their combined unattractiveness, Soviet conservatives and diehard Southern US white supremacists were not stupid: they knew their respective systems rested on minute and absolute compliance, and that tinkering with them was akin to the little Dutch boy of legend removing his fist from the hole in the dyke.

2. Mainstream Media and Alternative Media

Here again, the dividing lines are blurry. Where would city culture newspapers with a critical edge, such as New York's *Village Voice*, an increasingly common genre in the 1980s and 1990s, fit? What should be made of reporters who run stories in alternative media under a pseudonym because their own editors have refused to carry them? How should a newspaper be defined that begins on the radical left and then moves over to the political center, such as France's *Libération*? How should the 'accidental midwife' role be evaluated, of radical free radio stations in Italy and France in the 1970s in forcing legislative changes that, in the 1980s, opened up broadcasting to major corporations? What of a mainstream newspaper that regularly carries material made available to it by an extreme conservative foundation?

These are some of the questions of definition that make a strictly binary categorization of alternative and mainstream media difficult, and thus complicate the analysis of their inter-relation. One reading is that alternative media of the left often receive the news first, especially in politically charged situations, and effectively do the job that mainstream journalists are theoretically supposed to do. Another is that they create a lively public sphere, in Habermas' sense (see Jürgen Habermas 1989) something that in the days of a fiercely partisan press used to be quite common but is now rare in the corporate media universe.

3. Alternative Media and Social Movements

Many alternative media cannot be understood unless related to the social movement of which they are a part, or to which they are antagonistic. Social movements of many kinds both give life to such media and are often given life by them. In the nascent phase of a movement, radical small-scale media may be very influential in maintaining focus and stimulating debate among many of the movement's future leaders. During the movement's apogee, such media are often close to being its lifeblood, updating, mobilizing, debating, exposing, and ridiculing. In the aftermath, such media—not necessarily or even usually the same ones—provide scope for reflection and regroupment.

4. Alternative Media as Key Components of the Mediascape

Returning to the question of alternative media as trivial or significant, three arguments present themselves in favor of their importance. One is quite simply their historical and current ubiquity at important junctures. Second, alternative media are widely recognized to play a vital role in organizing and consolidating social movements. Social movements often have limited access to mainstream media and, operating outside of well-established institutional frameworks, depend on alternative media for their organization and coordination. The third relates to social memory. The more ephemeral the media, the argument runs, the less their impact. However, we may conclude differently, that many of these short-

lived small-scale media make a particularly explosive dent in the political culture of the moment. In this, their mnemonic function is arguably different from mainstream media, whose power consists in sedimenting stable definitional frameworks over time within which the interpretation of society and social change takes place. Both operations, which might be likened to the brilliantly colored tropical fish and to the coral reef, are sociologically significant.

See also: Adolescents: Leisure-time Activities; Art, Sociology of; Cultural Policy; Cultural Policy: Outsider Art; Internet: Psychological Perspectives; Mass Communication: Technology; Mass Media and Cultural Identity; Media and History: Cultural Concerns; Media and Social Movements; Media Ethics; Media Imperialism; Media, Uses of; Popular Culture; Television: Genres; Television: History; Television: Industry

Bibliography

Alexeyeva L 1985 *Soviet Dissent: Contemporary Movements For National, Religious and Human Rights*. Wesleyan University Press, Middletown, CT

Armstrong D 1981 *A Trumpet To Arms: Alternative Media in America*. J. P. Tarcher, Los Angeles

Aronson J 1972 *Deadline For The Media: Today's Challenges to Press, TV and Radio*. Bobbs-Merrill, Indianapolis

Baldelli P 1977 *Informazione e Controinformazione*. G. Mazzotta, Milan

Boyle D 1997 *Subject To Change: Guerrilla Television Revisited*. Oxford University Press, New York

Darnton R 1995 *The Forbidden Best-Sellers of Pre-Revolutionary France*. W.W. Norton, New York

Dowing J 2000 *Radical Media: Rebellious communication and social movements*. Sage, Thousand Oaks, CA

Gilmont J-F (ed.) 1990 *La Réforme et le Livre: l'Europe de l'Imprimé (1517–v.1570)*. Éditions du Cerf, Paris

Habermas J 1989 *The Structural Transformation Of The Public Realm*. Beacon Press, Boston

Hill C 1975 *The World Turned Upside Down: Radical Ideas During the English Revolution*. Penguin Books, Harmondsworth, UK

Hilliard R L, Keith M C 1999 *Waves of Rancor: Tuning in the Radical Right*. M.E. Sharpe, Armonk, NY

Kahn D, Neumaier D (eds.) 1985 *Cultures in Contention*. Real Comet Press, Seattle, WA

Kintz L, Lesage J (eds.) 1998 *Media, Culture, and the Religious Right*. University of Minnesota Press, Minneapolis, MN

Lenin VI 1969 *What Is To Be Done?* International Publishers, New York

Raboy M 1984 *Movements and Messages: Media and Radical Politics in Québec*. Between The Lines, Toronto, Canada

Scott J C 1985 *Weapons of the Weak: Everyday Forms of Peasant resistance*. Yale University Press, New Haven, CT

Shanor D R 1985 *Behind the Lines: The Private War against Soviet Censorship*, 1st edn. St Martin's Press, New York

Simpson Grinberg M 1981 *Comunicación Alternativa y Cambio Social*. Universidad Nacional Autónoma de México, México

Sreberny-Mohammadi A, Mohammadi A 1994 *Small Media, Big Revolution: Communication, Culture and the Iranian Revolution*. University of Minnesota Press, Minneapolis, MN

J. D. H. Downing

Altruism and Prosocial Behavior, Sociology of

Two basic questions will be addressed in this article. One deals with the very essence of human nature, and essentially lies within the domain of philosophy: Does altruism exist? The other is an empirical social science question: How does one understand, predict, and explain positive other-oriented social action?

1. Does Altruism Exist?

Modern social science is founded mainly on the assumption that animals, including humans, are primarily motivated by egoism, that is, that each organism's basic drives involve satisfying its own needs and desires. The recognition of the importance of self-interest in human motivation goes back at least as far as the writings of Plato in the Western philosophical tradition. Various forms of this assumption are found in economic theory (e.g., classical economics), psychology (Freud's pleasure principle, Thorndike's law of effect), in social psychology (exchange theory), and in sociology (functionalism). At the dawn of the twenty-first century, the purest expression of this assumption is found in rational choice theory.

Yet many thinkers have resisted accepting the idea that all human action is selfishly motivated. Plato and Aristotle struggled to understand the source of concern for the other that is present in friendship. Even near the time of Hobbes' classic *Leviathan*, Rousseau, Hume, and Adam Smith raised doubts that egoism was the only human motivation (see Batson 1991). Smith, for example, wrote: 'How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it' ([1759] 1853, I.i.1.1).

1.1 Conceptualizing Altruism

Many different definitions have been offered for the term 'altruism.' Comte, who coined the term, defined it as an unselfish desire to 'live for others' (Comte 1875, p. 556). Social psychologists have proposed that

altruism consists of helping actions carried out without the anticipation of rewards from external sources (Macaulay and Berkowitz 1970), while others suggest that altruistic helpers must incur some cost for their actions (e.g., Krebs 1982, Wispé 1978). Note in both the focus on consequences for the helper rather than on inner motivation. Batson (1991) has proposed instead that altruism be defined by the individual's motivations: 'Altruism is a motivational state with the ultimate goal of increasing another's welfare' (p. 6).

Batson's definition makes it impossible to study altruism in nonhuman species. Sober and Wilson (1998) recognized the need to make a distinction between two types of altruism, 'evolutionary altruism' and 'psychological altruism.'

The concepts of psychological egoism and altruism concern the motives that people have for acting as they do. The act of helping others does not count as (psychologically) altruistic unless the actor thinks of the welfare of others as an ultimate goal. In contrast, the evolutionary concept concerns the effects of behavior on survival and reproduction. Individuals who increase the fitness of others at the expense of their own fitness are (evolutionary) altruists, regardless of how, or even whether, they think or feel about the action (Sober and Wilson 1998, p. 6).

They stress that, 'Even if we restrict our attention to organisms that *do* have minds, we need to see that there is no one-to-one connection between egoistic and altruistic psychological motives on the one hand and selfish and altruistic fitness consequences on the other' (p. 202). They conclude:

> The take-home message is that every motive can be assessed from two quite different angles. The fact that a motive produces a behavior that is evolutionarily selfish or altruistic does not settle whether the motive is psychologically egoistic or altruistic (p. 205).

Sober and Wilson (1998) do however propose that 'an ultimate concern for the welfare of others is among the psychological mechanisms that evolved to motivate adaptive behavior' (p. 7). They believe that both egoistic and altruistic tendencies are adaptive for survival.

1.2 Evidence for the Existence of Evolutionary Altruism

Researchers have now demonstrated mathematically (Boorman and Leavitt 1980) and by means of computer simulations (Morgan 1985) that genes for evolutionary altruism can evolve and become established in populations, through one of three mechanisms. Group selection can operate if the presence of some altruists in an isolated, endogamous group leads that entire group to survive better than groups without altruists. Kin selection operates if an altruist is more likely to save kin, whose genes are shared with the altruist and thus are more likely to survive and multiply. Finally, reciprocity selection works through a mechanism in which altruists are more likely to benefit each other, even if they are not related. Such a mechanism requires that bearers of altruistic genes be able to recognize each other—presumably through observation of past behavior. Sober and Wilson (1998, pp. 149–54) also devote considerable space to the discussion of group selection of cultural practices—such as social norms—as an alternative 'evolutionary' mechanism not involving genetic transmission by which altruistic behaviors can become established in social groups.

A strong argument has been made for empathy as the prime candidate for the inherited capacity underlying psychological altruism. At least five studies going back to 1923 have demonstrated that infants as young as a few hours old are more likely to cry at the sound of another infant's cry than at the sound of equally loud and annoying noises (Martin and Clark 1982). In addition, Matthews et al. (1981) and Rushton et al. (1986), using twin methods, found significant heritability of scores on self-report scales of altruism.

1.3 Evidence for the Existence of Psychological Altruism

Sober and Wilson (1998) discuss the difficulty both of defining what psychological altruism is and of demonstrating that it exists. If an egoistic goal is defined as 'anything one wants' one has defined altruism out of existence. If, however, altruism is defined as having irreducible preferences that the welfare of another be enhanced, it is possible to demonstrate its existence. Within social psychology, Daniel Batson has tried to demonstrate that one *can* find what Sober and Wilson (1998, pp. 245–8) call 'A over E' (other over self) pluralism: that there are some people who some of the time will choose the welfare of others over their own. Batson believes this happens when they feel empathy for the other: the 'empathy–altruism hypothesis.' For over 20 years Batson has waged academic war with several other social psychologists who espouse more egoistic models. He sets his theory against the arousal/cost-reward model of Piliavin et al. (1981), which assumes that observing another's problem arouses feelings of distress and helping alleviates those feelings, and the negative-state relief model (e.g., Cialdini et al. 1987). Cialdini assumes that children learn during socialization that helping others makes them feel good and that helping is thus motivated by hedonism. The war has been fought in a series of experimental skirmishes, summarized in Batson (1991). It appears that the rather modest goal has been achieved: some people, some of the time, do help out of altruism.

2. Determinants of Prosocial Behavior

Whether or not there are innate tendencies to care about the welfare of others, people do engage in

helpful behaviors. Two terms are commonly used in the literature. 'Prosocial behavior' means any actions, 'defined by society as generally beneficial to other people and to the ongoing political system' (Piliavin et al. 1981, p. 4). This can include paying taxes, doing volunteer work, cooperating with classmates to solve a problem, giving directions on the street, or intervening in a crime. Even acts normally deemed criminal can be prosocial in context, for example, taking medicines from a drugstore destroyed by a hurricane if needed for victims of the storm. 'Helping behavior' refers to 'an action that has the consequences of providing some benefit to or improving the well-being of another person' (Schroeder et al. 1995, p. 16). Actions from which one can also benefit (such as cooperating) are not included.

How is variation in prosocial behavior to be understood? Following the Lewinian equation $B = f(P, E)$, it is assumed that prosocial behavior is jointly determined by characteristics of the person and the environment. Thus, how individual differences in prosocial tendencies arise is first examined, and then how those tendencies combine with situational factors to influence helping behavior.

2.1 The Development of Prosocial Behavior Tendencies

Hoffman (1990) believes that, initially, infants cannot differentiate self from other, and feel only 'global empathy'; what happens to others is felt as happening to them. Starting at around a year, infants can differentiate self from others, but assume that others in distress feel exactly what they would feel. By age two or three, they not only understand that others can feel something different, but that a different response may be needed. By late childhood, children can experience empathy in response to knowledge of a person's situation, without observing the distress. Zahn-Waxler et al. (1992) provide observational evidence consistent with this theorizing regarding the early stages, and Aronfreed (1970) presents evidence regarding the learning of empathy by older children.

Cialdini et al. (1982) present a stage theory in which, in the presocialization stage (before 10), children do not know that others value helping, and will help only if asked. In the second stage, awareness, they learn that there are norms about helping, and that they can be rewarded for it, so they help in order to please. By around age 15, internalization occurs; helping becomes intrinsically satisfying. Some research supports these theories, which are consistent with research by Kohlberg (1985) regarding the development of moral reasoning. Moral reasoning, like helping, is initially based on external factors; as the child matures, decisions become based on inner motivations.

Social learning theory also informs research on the process by which children develop prosocial behavior patterns. The use of direct rewards and punishments (power assertion) and love withdrawal have been found to be less effective for the development of prosocial behavior than induction, that is, reasoning with the child. Observational learning is undoubtedly more important than direct teaching; both models who are physically present and those presented in the mass media can have significant effects. 'Preaching' altruism also has some effect, as does attributing altruistic motives to the child. Prosocial socialization can continue through adulthood, and attributional processes can be important. The 'foot-in-the-door' technique— asking a small favor and then returning to ask a larger one (e.g., Freedman and Fraser 1966)—is thought to have its effect through self-attribution. Similarly, regular participation in volunteer work leads individuals to internalize the role of helper.

The relative importance of personality and situational factors differs depending upon the kind of helping. Episodic helping—responsiveness to a request or to the perception of a sudden need—is more influenced by the situation. Sustained helping, such as volunteering, is more influenced by socialization factors and by habits, values, and personality. Interactions between personality and situational factors have also been found. Schroeder et al. (1995) present a detailed discussion of the factors influencing the various forms of helping behavior.

2.2 Determinants of Episodic Helping

Most of the empirical research on helping has used experimental methodology to study situations in which someone has a sudden need for help. Factors such as clarity and urgency of the need, the race, sex, age, or handicap of the 'victim,' how many potential helpers are present, and the relationship of victim and subject are manipulated. Latané and Darley (1970) propose that the bystander goes through a five step decision process: (a) noticing something happening, (b) deciding help is needed, (c) deciding whether one personally has a responsibility to intervene, (d) choosing a course of action, and (e) executing the plan. They consistently find that the more bystanders, the lower the likelihood that any one of them will intervene (the 'diffusion of responsibility' effect). This effect can occur at two points in the decision process. If other bystanders are visible, their actions can define whether the situation requires help (step b); if they are not visible, the knowledge that they could help can influence attribution of responsibility (step c).

Others (e.g., Piliavin et al. 1981) propose that part of the decision process involves calculating the costs to bystander and victim of intervening and not intervening. The nature of the emergency, relationship to the victim, and other situational factors enter into such calculations. The bystander's personality, background, and training can also have effects, and there can be interactions of personal and situational factors.

Wilson (1976) found that safety-oriented individuals were less likely to intervene in a perceived emergency than were esteem-oriented individuals; this difference was much greater when other bystanders were present.

2.3 Determinants of Sustained Prosocial Behaviors

Only recently have investigators seriously focused on long-term, planned helping behaviors such as blood donation, charitable giving, and volunteering for nonprofit organizations. *Voluntary Sector* has carried out surveys of charitable donation and volunteering in the USA every two years since 1988. Thus much is known descriptively about participants: They are mainly white, middle class, and express altruistic motives for their actions.

Three approaches to studying long-term helping have emerged in social psychology: attempts to find 'the altruistic personality,' explorations of the functions served by volunteering, and analyses based on the concept of role identity. After an extensive investigation of gentiles who saved Jews from the Holocaust, Oliner and Oliner (1988) proposed several personality characteristics that separate them from those who did not. Penner (e.g., Penner and Finkelstein 1998) developed a personality measure with two dimensions: other-oriented empathy (feelings of responsibility and concern about others' well-being) and helpfulness (a self-reported history of helping). Both measures distinguish volunteers from nonvolunteers, and both are related to length of service in an HIV/PWA organization and to organizational citizenship (doing optional things at work that benefit the organization).

Snyder and colleagues (e.g., Clary et al. 1998) assume that altruism is only one of many motivations behind volunteering. They measure six potential motives: enhancement (to increase self-esteem), career (to increase success in one's profession), social (to enhance friendships), values (to express who one is), protective (to escape from one's troubles), and understanding (to learn about the world). They show that these motives are relatively stable over time, people with different motives are persuaded by parallel types of appeals (e.g., people high in social motives positively evaluate social appeals), and when experiences match motives, volunteers are more satisfied.

A more sociological approach, guided by Mead's (1934) conception of roles as patterns of social acts framed by a community and recognized as distinct social objects, emphasizes helping as role behavior. A series of studies of blood donors (Piliavin and Callero 1991) demonstrate that internalization of the blood donor role is more strongly associated with a history of blood donation than are personal and social norms. More research shows similar effects for identities tied to volunteering time and giving money (Grube and Piliavin 2000, Lee et al. 1999).

Finally, there are macrosociological approaches to the study of long-term helping behaviors. For example, Wilson and Musick (1997) have presented data in support of a model using both social and cultural capital as predictors of involvement in both formal and informal volunteering. Social structure also influences the distribution of resources that may be necessary for certain helping relationships. One needs money to be able to donate to a charity and medical expertise to be able to help earthquake victims.

2.4 Cross-cultural Research

Little systematic research compares helping across cultures. Beginning in the 1970s, researchers compared helping in rural and urban areas, consistently finding that helping strangers (although not kin) is more likely in less dense areas around the world. In a real sense, then, urban and rural areas appear to have different 'cultures': Small towns are more communal or collective, while cities are more individualistic. A recent review (Ting and Piliavin 2000) examined this and many other cross-cultural studies, not only of the helping of strangers but also on the development of moral reasoning, socialization of prosocial behavior, and participation in 'civil society.' Although more collective societies generally show up as 'nicer' than individualistic societies in these comparisons, these cultures also differ in the pattern of helping. More help is provided to ingroup members than strangers in most societies, but the difference between the amount of help offered to ingroup and outgroup members is greater in communal societies.

2.5 Civil Society

Although most social scientists are still skeptical of the existence of 'pure altruism,' most serious researchers agree that some of the people some of the time consider the needs of others in decision making. Game theorists have discovered that in repeated prisoner's dilemma games and public goods problems, individuals consistently behave in more co-operative or altruistic ways than expected, and some do so more than others (Liebrand 1986). Economists and political scientists, who have long believed that all motivation is selfish, have come to grips with evidence on voting and public goods behavior which indicates that this is not true (Mansbridge 1990, Clark 1998).

See also: Adulthood: Prosocial Behavior and Empathy; Altruism and Self-interest; Attitudes and Behavior; Cooperation and Competition, Psychology of; Cooperation: Sociological Aspects; Darwinism: Social; Moral Sentiments in Society; Motivation and Actions, Psychology of; Prosocial Behavior and Empathy: Developmental Processes; Race and Gender Intersections; Sociobiology: Overview

Bibliography

Aronfreed J 1970 Socialization of altruistic and sympathetic behavior: Some theoretical and experimental analyses. In: Macauley J, Berkowitz L (eds.) *Altruism and Helping Behavior*. Academic Press, New York, pp. 103–23

Batson C D 1991 *The Altruism Question*. Erlbaum, Hillsdale, NJ

Boorman S A, Leavitt P R 1980 *The Genetics of Altruism*. Academic Press, New York

Cialdini R B, Kenrick D T, Baumann D J 1982 Effects of mood on prosocial behavior in children and adults. In: Eisenberg N (ed.) *The Development of Prosocial Behavior*. Academic Press, New York, pp. 339–59

Cialdini R B, Schaller M, Houlihan D, Arps K, Fultz J, Beamen A L 1987 Empathy-based helping: Is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology* **52**: 749–58

Clark J 1998 Fairness in public good provision: An investigation of preferences for equality and proportionality. *Canadian Journal of Economics* **31**: 708–29

Clary E G, Snyder M, Ridge R D, Copeland J, Stukas A A, Haugen J, Miene P 1998 Understanding and assessing the motivations of volunteers: A functional approach. *Journal of Personality and Social Psychology* **74**: 1516–30

Comte I A 1875 *System of Positive Polity*. Longmans, Green, London, Vol. 1

Darley J, Latané B M 1970 *The Unresponsive Bystander: Why Doesn't He Help?* Appleton-Century-Crofts, New York

Freedman J L, Fraser S C 1966 Compliance without pressure: The foot in the door technique. *Journal of Personality and Social Psychology* **4**: 195–202

Grube J, Piliavin J A 2000 Role identity and volunteer performance. *Personality and Social Psychology Bulletin* **26**: 1108–19

Hoffman M L 1990 Empathy and justice motivation. *Motivation and Emotion* **14**: 151–71

Kohlberg L 1985 *The Psychology of Moral Development*. Harper and Row, San Francisco, CA

Krebs D L 1982 Altruism—A rational approach. In: Eisenberg N (ed.) *The Development of Prosocial Behavior*. Academic Press, New York, pp. 53–76

Lee L, Piliavin J A, Call V R A 1999 Giving time, money, and blood: Similarities and differences. *Social Psychology Quarterly* **62**: 276–90

Liebrand W B G 1986 The ubiquity of social values in social dilemmas. In: Wilke H A M, Messick D M, Rutte C G (eds.) *Experimental Social Dilemmas*. Verlag Peter Lang, Frankfurt am Main, Germany

Macaulay J, Berkowitz L 1970 *Altruism and Helping Behavior*. Academic Press, New York

Mansbridge J (ed.) 1990 *Beyond Self-interest*. University of Chicago Press, Chicago

Martin G B, Clark III R D 1982 Distress crying in infants: Species and peer specificity. *Developmental Psychology* **18**: 3–9

Matthews K A, Batson C D, Horn J, Rosenman R H 1981 'Principles in his nature which interest him in the fortune of others …': The heritability of empathic concern for others. *Journal of Personality* **49**: 237–47

Mead G H 1934 *Mind, Self, and Society*. University of Chicago Press, Chicago

Morgan C J 1985 Natural selection for altruism in structured populations. *Ethology and Sociobiology* **6**: 211–18

Oliner P M, Oliner S P 1995 *Toward a Caring Society: Ideas into Action*. Praeger, Westport, CN

Oliner S P, Oliner P M 1988 *The Altruistic Personality*. Free Press, New York

Penner L A, Finkelstein M A 1998 Dispositional and structural determinants of volunteerism. *Journal of Personality and Social Psychology* **74**: 525–37

Piliavin J A, Callero P L 1991 *Giving Blood: The Development of an Altruistic Identity*. Johns Hopkins University Press, Baltimore, MD

Piliavin J A, Charng H-W 1990 Altruism: A review of recent theory and research. *Annual Review of Sociology* **16**: 27–65

Piliavin J A, Dovidio J F, Gaertner S, Clark III R D 1981 *Emergency Intervention*. Academic Press, New York

Rushton J P, Fulker D W, Neale M C, Nias D K B, Eysenck H J 1986 Altruism and aggression: The heritability of individual differences. *Journal of Personality and Social Psychology* **50**: 1192–8

Schroeder D A, Penner L A, Dovidio J F, Piliavin J A 1995 *The Psychology of Helping and Altruism: Problems and Puzzles*. McGraw-Hill, New York

Smith A [1759] 1853 *The Theory of Moral Sentiments*. Henry G. Bohn, London

Sober E, Wilson D S 1998 *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, Cambridge, MA

Ting J-C, Piliavin J A 2000 Altruism in comparative international perspective. In: Phillips J, Chapman B, Stevens D (eds.) *Between State and Market: Essays on Charities Law and Policy in Canada*. McGill-Queens University Press, Montreal, PQ, pp. 51–105

Wilson J P 1976 Motivation, modeling, and altruism: A person x situation analysis. *Journal of Personality and Social Psychology* **34**: 1078–86

Wilson J, Musick M 1997 Who cares? Toward an integrated theory of volunteer work. *American Sociological Review* **62**: 694–713

Wispé L G (ed.) 1978 *Altruism, Sympathy, and Helping: Psychological and Sociological Principles*. Academic Press, New York

Zahn-Waxler C, Radke-Yarrow M, Wagner E, Chapman M 1992 Development of concern for others. *Developmental Psychology* **28**: 126–36

J. A. Piliavin

Altruism and Self-interest

Altruism was first used circa 1853 by Auguste Comte. French *altruisme*—another; Italian *Altrui*—somebody else, what is another's; Latin *alteri huic*—to this other.

1. Definition

Altruism is behavior intended to benefit another, even when this action risks possible sacrifice to the welfare

of the actor. There are several critical aspects to altruism. (a) Altruism must entail action. Good intentions or well meaning thoughts do not constitute altruism. (b) The action is goal-directed, although this may be either conscious or reflexive. (c) The goal must be to further the welfare of another. If another's welfare is merely an unintended or secondary consequence of behavior designed primarily to further the actor's own welfare, the act is not altruistic. (d) Intentions count more than consequences. If John tries to do something nice for Barbara, and it ends up badly or with long-term negative consequences for Barbara, this does not diminish the altruism of John's initial action. Motivation and intent are critical, even though motives and intent are difficult to establish, observe, and measure objectively. (e) Altruism carries some possibility of diminution in the actor's own welfare. An act that improves both the altruist's own welfare and that of another person would be considered collective welfare, not altruism. (f) Altruism sets no conditions; its purpose is to further the welfare of another person or group, without anticipation of reward for the altruist.

Analysts often introduce various conceptual subtleties into this basic definition. We might refer to the above definition as pure altruism, and distinguish it from what could be called particularistic altruism, defined as altruism limited to particular people or groups deemed worthy because of special characteristics such a shared ethnicity, religion, or family membership (Wispé 1978).

In discussing altruism, analysts often use the term interchangeably with giving, sharing, cooperating, helping, and different forms of other-directed or prosocial behavior. The problem then becomes to recognize and allow for the subtle variations in altruism while retaining the simplicity of the single term. To solve this problem, analysts often refer to acts that exhibit some, but not all, of the defining characteristics of altruism as quasialtruistic behavior. This distinction allows us to differentiate between the many acts frequently confused with altruism (such as sharing or giving) without having to lump these significant deviations from self-interest into a catch-all category of altruism (Bar-Tal 1976, Derlaga and Brzelak 1982).

Analysts also frequently further conceptualize behavior along a continuum, with pure self-interest and pure altruism as the two poles and modal or normal behavior, including quasialtruistic acts, distributed between them. This approach avoids the problem of dichotomizing behavior into only altruistic or self-interested acts. It minimizes the confusion resulting from excessive terminological intricacies. Yet it retains the advantage of allowing us to discuss quasialtruistic acts or limited versions of altruism (such as the particularistic altruism discussed above) that would be provided for by more complex definitional terminology.

2. *What Causes Altruism?*

Analysts offer a wide range of explanations, from innate predispositions to socialization and tangible rewards. The best analyses of altruism consider more than one explanatory variable, and many of the underlying influences on altruism are frequently referred to by different names, depending on the discipline or the analyst (Kohn 1990).

Thomas Hobbes, for example, suggested an explanation for altruism that emanates not from genuine concern for the needy person but rather from the so-called altruist's personal discomfort at seeing someone else in pain (Losco 1986). Economists designate such altruism a form of 'psychic utility'; psychologists identify the same general phenomenon but refer to it as 'aversive personal distress created by arousal.' This problem presents two problems for the reader. (a) Analysts refer to the same—or to vastly similar—concepts using different terminology. These terminological differences vary, more or less systematically, from discipline to discipline; in any given analysis, they may reflect deliberate choices based on important philosophical orientations toward understanding behavior or, conversely, may merely be conventionally and uncritically adopted. (b) For purposes of analysis we need to separate predictors of altruism into distinct components in order to clarify and understand their relative influences. But in reality, these various influences often blend together and are far less distinct than our analysis suggests.

Given these caveats, it is fair to say that explanations of altruism tend to cluster into four analytical categories: sociocultural, economic, biological, and psychological.

2.1 *Sociocultural Explanations*

These focus on the individual demographic correlates of altruism. These range from religion, gender, and family background to wealth, occupation, education, or political views. The basic assumption underlying sociocultural explanations is that belonging to a particular sector of the population will predispose one toward altruism. Women are frequently said to be more altruistic than men, presumably because of genetic or socialization factors. Living in a small, close-knit community or being a member of a large, communal family is also frequently cited as encouraging altruism (Oliner et al. 1992).

2.2 *Economic Explanations*

While some economists dismiss altruism as the result of an odd utility function, most economists now tend to consider altruism a good and stress the importance

of rewards for altruism. These rewards may be material (money) or psychological (praise, honors, or simply feeling good about oneself) but are always expressed in some implicit basic economic calculus in which individual costs and benefits are entered. This leads to explanations in which altruism becomes a short-term strategy designed to obtain later goods for the altruist, through reciprocated benevolent behavior or alleviation of guilt. In reciprocal altruism, for example, John gives to Mary today in the hope that she may do something nice for him tomorrow. In participation altruism, Mary wants to be the one to give to John, as opposed to Mary's simply wanting John to benefit from anyone's altruism. Economists also explain altruism through the concept of dual utilities, in which altruism is a balancing act between John's individual self-interested utility and his other-directed utility function, or resource altruism, a variant of the dual utilities explanation in which altruism is treated as a luxury good to be indulged in once the actor has obtained his more basic self-interested needs (Margolis 1982).

The central concept underpinning all economic analyses is the idea that people think in terms of costs and benefits, and that altruism can be explained through such an economic analysis (Phelps 1975).

2.3 Evolutionary Biology

Biology is built on the Darwinian concept of individual selection and survival of the fittest. Acts in which one organism takes steps to promote the survival of another organism violates this individual selection principle. Thus, altruism presents a particular challenge to evolutionary biologists. Many biologists simply write off altruism as aberrant behavior, which eventually will disappear. Those biologists who do take altruism seriously tend to explain it through explanations that favor kin or group selection (Trivers 1971).

In kin selection, the gene, not the organism, is designated the critical unit. A selfish gene then may decide to further its likelihood of survival through foregoing the host organism's ability to propagate. Mary thus may decide not to marry since her sister has already married and Mary realizes that there will be too many offspring to survive in conditions of scarcity. Mary's sacrifice as a person willing to forego the pleasure of having children, however, actually helps further the survival of the genes Mary shares with her sister, and thus the gene is said to be selfish.

Group selectionists make a similar argument but designate the group the critical unit. They argue that groups containing a few altruists do better than groups with no altruists, since the altruists may sacrifice themselves for the good of the group. It is thus in the group's interest in the long run to protect and encourage altruism. (Tax incentives for charitable contributions are a form of this.)

Such biological analyses rely heavily on in-group/out-group distinctions and on the importance of clusters or networks of altruists who are tolerated and even protected, not for themselves but rather because their altruism benefits the group that contains them. Community size is said to encourage such altruism, as are networks or clusters of altruists, in which the mere existence and visibility of a group of altruists may influence a person to engage in similarly altruistic behavior, either through sanctions or rewards.

2.4 Psychology

Psychology offers the richest and most varied analyses of altruism. Psychologists frequently consider developmental factors such as socialization or child-rearing practices and the level of sociocognitive development. Unlike economists and biologists, psychologists allow directly for norms, usually by assuming that these are values internalized through socialization and development and are at least partially cognitive in construction (Rushton and Sorrentino 1981). Culture plays an important influence insofar as these values and norms are reinforced by the society at large. But psychologists also allow for more personal construction of values by the individual (Krebs 1982).

Psychologists frequently include some of the same factors considered by other analysts, such as reciprocity (the exchange of benefits) and often build these values into complex systems of moral judgment. Characteristics of the specific situation, including the identity of the recipient, the anonymity of the helper, and the number and identity of observers of the altruistic act, play critical roles in the emergence of altruistic behavior for psychologists (Latané and Darley 1970).

Psychological discussions, and their counterparts in philosophy, touch on the category of explanation that may be most promising (Staub et al. 1984). These works emphasize empathy, views of oneself and of the world, expectations, and identity (Batson 1991, Batson and Shaw 1991, Eisenberg and Strayer 1987). They include the cognitive and emotional bases of altruism and introduce the impact of culture via the psychological process of reasoning that leads to altruism (Oliner et al. 1992). When these factors come together in a particular way, they are said to constitute what is referred to as the altruistic personality, in which the habit of giving to others becomes so ingrained in a person that the habit becomes part of the person's personality (Oliner and Oliner 1988), or the altruistic perspective, in which the fact that the actor does, or does not, see a bond between himself and the recipient of the altruistic act, becomes critical (Monroe 1996).

This emphasis on a common humanity may provide the link with biological explanations, where there is shared genetic material, or religious explanations, in which shared membership is emphasized.

3. Importance of Altruism

Although altruism is empirically rare, its mere existence can inspire and better the world. The ordinary people who risked their lives to rescue Jews during the Holocaust help restore our hope in humanity, just as Gandhi and Mother Teresa inspire us with their acts.

Beyond this, altruism is important since its very existence challenges the widespread and dominant belief that it is natural for people to pursue individual self-interest. Indeed, much important social and political theory suggests altruism should not exist at all. It thus becomes important to consider altruism not merely to understand and explain the phenomenon itself but also to determine what its continuing existence reveals about limitations in the Western intellectual canon, limitations evident in politics and economics since Machiavelli and Hobbes, in biology since Darwin, and in psychology since Freud.

See also: Cooperation and Competition, Psychology of; Cooperation: Sociological Aspects; Identity and Identification: Philosophical Aspects; Rational Choice Theory: Cultural Concerns; Sociobiology: Overview; Sociobiology: Philosophical Aspects; Utilitarianism: Contemporary Applications

Bibliography

Bar-Tal D 1976 *Prosocial Behavior: Theory and Research*. Hemisphere, Washington, DC

Batson C D 1991 *The Altruism Question: Toward a Social Psychological Answer*. Lawrence Erlbaum Associates, Hillsdale, NJ

Batson C D, Shaw L 1991 Evidence for altruism: toward a plurality of prosocial motives. *Psychological Inquiry* **2**(2): 107–22

Eisenberg N, Strayer J (eds.) 1987 *Empathy and Its Development*. Cambridge University Press, New York

Derlega V J, Brzelak J (eds.) 1982 *Cooperation and Helping Behavior: Theories and Research*. Academic Press, New York

Kohn A 1990 *The Brighter Side of Human Nature: Altruism and Empathy in Everyday Life*. Basic Books, New York

Krebs D 1982 Psychological approaches to altruism: an evaluation. *Ethics* **92**: 447–58

Latané B, Darley J M 1970 *The Unresponsive Bystander: Why Doesn't Anybody Help?* Appeton-Century-Crofts, New York

Losco J 1986 Understanding altruism: a comparison of various models. *Political Psychology* **7**(2): 323–48

Margolis H 1982 *Selfishness, Altruism and Rationality*. Cambridge University Press, Cambridge, UK

Monroe K R 1996 *The Heart of Altruism: Perceptions of a Common Humanity*. Princeton University Press, Princeton, NJ

Oliner S P, Oliner P M 1988 *The Altruistic Personality: Rescuers of Jews in Nazi Europe*. Free Press, New York

Oliner P, Oliner S P, Baron L, Blum L A, Krebs D L, Smolenska M Z (eds.) 1992 *Embracing the Other: Philosophical, Psychological, and Historical Perspectives on Altruism*. New York University Press, New York

Phelps E S (ed.) 1975 *Altruism, Morality, and Economic Theory*. Russell Sage Foundation, New York

Rushton J P, Sorrentino R M (eds.) 1981 *Altruism and Helping Behavior: Social, Personality, and Developmental Perspectives*. Lawrence Erlbaum Associates, Hillsdale, NJ

Staub E, Bar-Tal D, Karylowski J, Reykowski J (eds.) 1984 *Development and Maintenance of Prosocial Behavior*. Plenum, New York

Trivers R 1971 The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–57

Wispé L (ed.) 1978 *Altruism, Sympathy, and Helping: Psychological and Sociological Principles*. Academic Press, New York

K. R. Monroe

Alzheimer's Disease: Antidementive Drugs

1. Introduction

Alzheimer's disease (AD) is associated with degeneration of cholinergic neurons in the basal forebrain. The resulting cholinergic deficit appears to be correlated to a certain degree with the severity of cognitive impairments, resulting in the choline-deficit hypothesis of AD (Withehouse et al. 1982, Becker and Giacobini 1988). Although it is evident that several other neurotransmitters are also involved in the pathogenesis, cholinomimetic drugs represent the only therapeutic approach with proven efficacy in mild to moderate probable AD patients (as defined by NINCDS–ADRDA (National Institute of Neurological and Communicative Disorders; Stroke/Alzheimer's Disease and Related Disorders Association) criteria (McKhann et al. 1984), Mini-Mental State Examination (MMSE, Folstein et al. 1975; usually scores of 10–26) and Clinical Dementia Rating (CDR, Hughes et al. 1982; usually 1–2). In particular, cholinesterase inhibitors (AChEI) have proven effective in stabilizing cognitive and behavioral symptoms of AD in well-designed, double-blind, placebo-controlled, clinical studies involving approximately 10,000 patients worldwide (Giacobini 2000). Tacrine, the first AChEI approved by the US Food and Drug Administration (FDA), was introduced on to the market in 1993. Its use remained limited, due to an asymptomatic elevation of liver enzymes in over a quarter of treated patients and a high frequency of cholinergic side effects. Nevertheless, tacrine prompted the development of the so-called second-generation AChEI (donepezil, rivastigmine, galantamine) that are devoid of hepatic toxicity, comparable in efficacy, and display

Table 1
Acetylcholinesterase inhibitors

Drug	Tradename	Molecule	Mechanism of action	Approval (3/2001)
Donepezil	Aricept®	Piperidine	Reversible/specific AChE > BChE	FDA, EMEA 1996
Eptastigmine	—	Carbamate	Reversible AChEI	—
Galantamine	Reminyl®	Alkaloid	Reversible/specific AChE > BChE and nicotinic receptor modulator	EMEA 2000, FDA 2001
Metrifonate	—	Dichlorvos	Pseudoirreversible AChEI	—
Physostigmine	—	Tertiary amine	Reversible/nonselective AChEI	—
Rivastigmine	Exelon®	Carbamate	Pseudoirreversible AChEI	CH 1997, EMEA 1998, FDA 2000
Tacrine	Cognex®	Acridine derivative	Reversible/nonspecific BChE > AChEI	FDA, EMEA 1993

FDA, Food and Drug Administration; EMEA, European Medical Evaluation Agency; AChE, Acetylcholinesterase; BChE, Buturylcholinesterase; AChEI, Acetylcholinesterase inhibitor.

a better risk–benefit ratio (Giacobini 1998, 2000). From eight AChEI that have been studied for use in AD, three are currently in clinical use: donepezil, rivastigmine, as well as galantamine are registered in the United States and Europe (Table 1).

2. *Efficacy*

Controlled clinical trials used standardized and validated scales as outcome measures for cognitive (Alzheimer's Disease Assessment Scale-cognitive scale, ADAS-cog, Rosen et al. 1984; Mini-Mental State Examination, MMSE, Folstein et al. 1975) and behavioral symptoms (e.g. Neuropsychiatric Inventory, NPI, Cummings et al. 1994) as well as activities of daily living or global clinical functioning (e.g. Clinician-Based Impression of Change, CIBIC; Reisberg et al. 1997). For all AChEI that are currently on the market, significant beneficial effects on various cognitive domains and behavioral function have been shown compared to placebo (Knapp et al. 1994, Ringman and Cummings 1999, Rogers et al. 1998, Wilcock 2000). The effects, however, are limited in terms of quantity and time. AD progression in mild to moderate stages of the disease (CDR stages 1–2) is assumed at a rate of approximately nine points per year on the 70 points ADAS-cog scale. According to Giacobini (2000), under current treatment conditions, AChEI may yield a maximal effect of ~ 3.6 points (ADAS-cog) compared to placebo, with an average 1.2 point difference at the end of those studies that lasted 26–30 weeks. This effect size appears to be modest and similar for all second-generation AChEI (Small et al. 1997, Giacobini 2000). Differences in clinical efficacy between the AChEI may be due to the fact that, at least for some drugs, the dose increase to reach sufficient inhibition of acetylcholinesterase of approximately 50 percent may not be achieved due to side effects.

Although most controlled clinical studies lasted only six months, open-label follow-up studies added evidence suggesting that AChEI treatment may stabilize cognitive and behavioral functioning for up to 12 months or even longer. Indeed, measures of both basic and instrumental activities of daily living tend to reach significance after 12 months of treatment, rather than after six months (Summers et al. 1996, Raskind et al. 2000). Open-label long-term studies of up to two years have shown improved activities of daily living (ADL), behavioral functioning, and social integration, suggesting that patients may profit from a long-term treatment of up to two years (Knapp et al. 1994, Morris et al. 1998, Rogers and Friedhoff 1998, Wilcock et al. 2000). Taken together, there is a growing body of evidence suggesting that the early start of treatment with a AChEI that is sustained for a longer period of time in standard dosage may yield the best results in terms of preserving functional ability. Whether tolerance to these drugs occurs during long-term administration is not known.

The percentage of patients which obviously benefit from AChEI treatment ('responders,' e.g. ADAS-cog $\geqslant 4$ points) varies from 25 percent (low-dosed rivastigmine) to 60 percent (high-dose tacrine, donepezil, galantamine) (Giocabini 2000). There are considerable interindividual variations in the degree of response: 5–10 percent of the patients show superior improvement compared to the other patients ('responder-type'), with a relevant impact on cognition, behavioral functioning, and social integration. Ten to fifteen percent of all treated patients do not improve on the ADAS-cog with any tested AChEI. The factors that influence response are not fully understood. Cholinergic brain deficit vs. other neurotransmitter imbalances, gender, APOE-4 status, or presence of other polymorphisms may play a role.

In addition to the cognitive benefit in terms of stabilization for up to 1–2 years, treatment with AChEI may have a positive socioeconomic impact, as Knopman et al. (1999) have shown that institutionalization can be delayed for up to one year. Furthermore, a reduction of caregiver burden was reported (Wimo et al. 1998; see *Gender Identity Disorders*).

3. Side Effects

Generally AChEI are well tolerated and the majority of adverse effects can be attributed to their cholinergic properties. The incidence of cholinergic side effects is dose-related and varies among the different drugs.

Adverse effects were highest in tacrine and 55 percent of the patients withdrew from the study because of adverse events vs. 11 percent of patients treated with placebo (Knapp et al. 1994). The primary reasons for withdrawal of tacrine-treated patients were asymptomatic liver transaminase elevations (28 percent) and gastrointestinal complaints (16 percent) (Knapp et al. 1994). The proportion who discontinued treatment with rivastigmine because of adverse events was significantly higher in the higher dose group (6–12 mg day^{-1}) than in the lower dose (1–4 mg day^{-1}) or placebo groups (23 percent (55/242) vs. 7 percent (18/242) and 7 percent (16/239)) (Rösler et al. 1999). Treatment discontinuation rates because of adverse events appeared to be similar between donepezil at 5 mg day^{-1} and placebo (about 4–9 percent vs. 1–10 percent) but greater with donepezil 10 mg day^{-1} (9–18 percent) compared with placebo (Dooley and Lamb 2000). Discontinuations due to adverse events occurred in 23 percent (24 mg day^{-1}) and 32 percent (32 mg day^{-1}) in galantamine treated patients (Raskind et al. 2000).

Cholinergic side effects were usually mild, transient in nature, and could be reduced by careful titration of the doses at the beginning of treatment. Most observed side effects (> 5 percent of patients) comprised the gastrointestinal system (nausea, vomiting, diarrhea, anorexia), the central nervous system (agitation, dizziness, insomnia, fatigue), and to a lesser degree peripheral symptoms like muscle cramps. The incidence of serious adverse events in the studies was not greater than in placebo (Giacobini 2000). However, some compounds derived from carbamates, such as eptastigmine, were withdrawn from clinical studies because of potential hematotoxicity. Although benefits measured by the ADAS-cog are very similar among the different drugs, there are, however, relevant differences in terms of the side effect ratio (Table 2) which may be related to the fact that the currently used compounds represent a chemically very heterogeneous group (see *Psychopharmacotherapy: Side Effects*).

4. Tacrine

Tacrine (tetrahydroaminoacridine, Cognex®), a non-selective, reversible AChEI, was FDA approved in 1993 and was the first drug on the market for the treatment of AD. With a 5.9 percent benefit vs. placebo in ADAS-cog, tacrine demonstrated the strongest acute effect on cognition compared to other AChEI during treatment of 30 weeks (Giacobini 2000). This has been attributed to additional pharmacological properties of the drug like blockade of potassium channels, inhibition of monoamine uptake, and inhibition of the monoamine oxidase (Jossan et al. 1992). Tacrine treatment appeared to delay nursing home placement for up to one year (Knopman et al. 1999). Its use, however, has been limited due to asymptomatic elevation of serum aminotransferase and a high incidence of cholinergic side effects of up to 58 percent. The most frequent observed adverse events were alanine aminotransferase elevation three times above normal in 29 percent of treated patients, nausea and vomiting in 28 percent, diarrhea in 14 percent, dyspepsia or anorexia in 9 percent, and myalgia in 7.5 percent (Wagstaff and McTavish 1994). In practice,

Table 2
Characteristics of selected acetylcholinesterase inhibitors

	Donepezil	Galantamine	Rivastigmine	Tacrine
Half-life (~ h)	70	6	1.5	2–4
Administration	oral, once/day	oral, bid	oral, bid	oral, 4 times/day
Study	Rogers et al. 1996	Tariot et al. 2000	Rösler et al. 1999	Knapp et al. 1994
Weeks	24	26	26	30
Dose (mg/d)	5–10	16–24	6–12	120–160
Treatment difference (percent) vs. Placebo (ADAS-cog, (observed cases analysis)	3.1	4.7–5.1	2.6	5.9
Patients improved (percent) (ADAS-cog ⩾ 4 points)	38–54	36–37	24	30–50
Discontinuation (percent) due to adverse events	9–18	6–10	23	55
Most common adverse events (percent)	Nausea 17	Nausea 13–17	Nausea 50	ASAT-Elevation 29/54
	Diarrhea 17	Vomiting 6–10	Vomiting 34	Nausea and/or Vomiting 35
	Vomiting 10	Diarrhea 7–9	Dizziness 20	Diarrhea 18

the unfavorable side effect profile which made liver enzyme monitoring necessary because of potential hepatotoxicity and the inconvenience of the four-times-a-day administration limited compliance and efficacy. Since second-generation AChEI have entered the market, tacrine is no longer in clinical use.

5. *Donepezil*

Donepezil (E2020, Aricept®), a reversible, CNS-selective piperidine AChEI, was the first second-generation AChEI and was FDA approved for use in 1996. Donepezil has minimal peripheral cholinesterase inhibiting activity and a long plasma half-life (70 h), allowing for once a day administration. Donepezil treatment during 24 weeks resulted in a 4.1 percent benefit vs. placebo in ADAS-cog (Giacobini 2000). Long-term efficacy data suggested that benefits in cognition and global functioning were maintained for about 21–81 weeks (Rogers and Friedhoff 1998). Donepezil was well tolerated. The most frequent adverse effects in the 10 mg day^{-1} group were nausea 22 percent, followed by diarrhea 17 percent, insomnia 18 percent, vomiting 10 percent, fatigue 8 percent, and muscle cramps 8 percent. Because of proven efficacy and a favorable side-effect profile, donepezil was considered a first-line treatment in AD (Small et al. 1997) and is widely used in patients with mild to moderate stages of the disease (Dooley and Lamb 2000) in Europe and the US.

6. *Rivastigmine*

Rivastigmine (ENA 713, Exelon®) is a relatively selective pseudoirreversible inhibitor of acetylcholinesterase with a short half-life of 1 hour and 10-hour duration of action. Studies showed a dose-dependent improvement of the ADAS-cog by 5.4 percent in a 26-week treatment trial (Giacobini 2000). The high-dose regime (6–12 mg day^{-1}) showed better effects on cognition than the 1–4 mg day^{-1} regime, but was associated with increased nausea (50 percent of treated patients) and vomiting (34 percent) (Rösler et al. 1999). Other adverse effects during treatment with 6–12 mg day^{-1} were dizziness (20 percent), headache (19 percent), and diarrhea (17 percent) (Rösler et al. 1999). Due to the short half-life it is usually administered twice a day. Rivastigmine adds to donepezil as another first-line treatment and was approved in several European countries and in the US (Spencer and Noble 1998, Rösler et al. 1999).

7. *Galantamine*

Galantamine is a natural, tertiary amarylidacea alkaloid originally isolated from the bulbs of snowdrop and Narcissus species. Recently, a synthetic version of the reversible competitive AChEI has been developed. A 'dual action' has been claimed because galantamine modulates presynaptic nicotinic receptors in addition to its AChE inhibiting properties (Fulton and Benfield 1996). Galantamin showed comparable efficacy to other AChEI (Tariot et al. 2000, Wilcock et al. 2000). The treatment effect of 24 mg day^{-1} compared to placebo was 3.9 points on the ADAS-cog/11 scale after six months (Raskind et al. 2000). Lilienfeld and Gaens (2000) reported a significant reduction of caregiver burden as evaluated by questionnaires following galantamine treatment. Reduction of time spent assisting the patient in ADL functions was calculated to be about one hour per day in comparison to placebo. Galantamine was well tolerated. As previously observed with the other AChEI, there is a dose-dependent increase of efficacy and of study dropouts due to side effects. Most of the side effects were predictably cholinergic and may be avoided by starting at a low dose and escalating the dose slowly. Among the adverse events during treatment with 24 mg day^{-1} were nausea (37.7 percent), vomiting (20.8 percent), dizziness (13.7 percent), and diarrhea (12.3 percent) (Raskind et al. 2000). Most of the side effects were predictably cholinergic and may be avoided by starting at a low dose and escalating dose slowly (in 4 weeks), as demonstrated by Tariot et al. (2000). Among the adverse effects during treatment with 24 mg day^{-1} were nausea (16.5 perceent), vomiting (9.9 percent), and diarrhea (5.5 percent). Galantamine adds to the treatment repertoire as another first-line treatment among the AchEI and was approved in European countries and in the US.

8. *Future Therapeutic Strategies*

So far, AChEI were investigated in individuals who have reached, in neurochemical and neuropathological terms, advanced stages of the disease. There is no published data regarding the efficacy of AChEI in incipient or severe AD. In the populations thus far studied, cholinomimetic treatments face obvious limitations in terms of effect size and duration of efficacy. Therefore they have to be judged as symptomatic or palliative. It is unknown, however, whether substitution of the cholinergic system is purely symptomatic or, as some authors suggest, yield certain neuroprotective properties (Giacobini 1996). The latter is supported by the fact that the stabilizing effect of AChEI is often maintained for several weeks after termination of the drug. However, significant long-term alterations of the progression of the disease have not been reported.

In summary, although of limited efficacy, the class of AChEIs represent the current treatment of choice for mild to moderate AD (Consensus statement: Small et al. 1997). Treatment effects of up to one year were shown in well-designed, controlled clinical studies. Data supporting long-term administration of chol-

inesterase inhibitors is limited to uncontrolled extension trials, but for all AChEI that have been investigated, positive effects on outcome were suggested. Thus, development of the AChEI was an important first step in the treatment of a disease that years ago appeared to be out of reach for pharmacology.

Novel treatment strategies that are currently developed address central histopathological features of AD, such as the deposition of abnormal proteins in neuritic plaques and neurofibrillary tangles. The amyloid β-peptides (Aβ) are the major constituents of brain amyloid plaques in AD. Lowering the aggregation or the production of Aβ or stimulating amyloid clearance from the brain is currently tested in model systems, e.g., transgenic mouse models or early clinical trials. Lowering Aβ production may be achieved by modulating activities of such key enzymes of amyloid precursor protein (APP) processing as α-, β-, and γ-secretases. Clearing amyloid from the brain may be achieved by means of immunization against Aβ peptides, at least in transgenic mouse models for AD. Such strategies may provide disease modifying or preventive treatments for AD in the future.

See also: Aging Mind: Facets and Levels of Analysis; Alzheimer's Disease: Behavioral and Social Aspects; Alzheimer's Disease, Neural Basis of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Dementia: Overview; Dementia: Psychiatric Aspects; Dementia, Semantic; Drugs and Behavior, Psychiatry of; Memory and Aging, Cognitive Psychology of; Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

Bibliography

Becker E, Giacobini E 1988 Mechanisms of cholinesterase inhibition in senile dementia of the Alzheimer type. *Drug Development Research* **12**: 163–95

Cummings J L, Mega M, Gray K, Rosenberg-Thompson S, Carusi D A 1994 The neuropsychiatric inventory: comprehensive assessment of psychopathology in dementia. *Neurology* **44**(12): 2308–14

Dooley M, Lamb H M 2000 Donepezil, a review of its use in Alzheimer's disease. *Drugs & Aging* **16**(3): 199–226

Folstein M F, Folstein S E, McHugh P R 1975 'Mini-Mental State'—A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* **12**: 189–98

Fulton B, Benfield P 1996 Galantamine. *Drugs & Aging* **9**(1): 60–5

Giacobini E 1996 Cholisterinase inhibitors do more that inhibit cholinesterase. In: Becker R, Giacobini E (eds.) *Alzheimer Disease: From Molecular Biology to Therapy*. Birkhäuser, Boston, pp. 187–204

Giacobini E 1998 Invited review. Cholinesterase inhibitors for Alzheimer's disease therapy: from tacrine to future applications. *Neurochemistry International* **32**: 413–19

Giacobini E 2000 Cholinesterase inhibitor therapy stabilizes symptoms of Alzheimer disease. *Alzheimer Disease and Associated Disorders* **14**(Suppl. 1): 3–10

Hughes C P, Berg L, Danzinger W L, Coben L A, Martin R L 1982 A new clinical scale for the staging of dementia. *British Journal of Psychiatry* **140**: 566–72

Jossan S S, Adem A, Winbald B, Oreland L 1992 Characterisation of dopamine and serotonin uptake inhibitory effects of tetrahydroaminoacridine in rat brain. *Pharmacology and Toxicology* **71**: 213–15

Knapp M J, Knopman D S, Solomon P R, Pendlebury W W, Davis C S, Gracon S I 1994 A 30-week randomized controlled trial of high-dose tacrine in patients with Alzheimer's disease. *Journal of the American Medical Association* **271**: 985–91

Knopman D S, Berg J D, Thomas R, Grundman M, Thal L J, Sano M 1999 Nusing home placement is related to dementia progress from a clinical trial. Alzheimer's Disease Cooperative Study. *Neurology* **52**(4): 714–18

Lilienfeld S, Gaens E 2000 Galantamine alleviates caregiver burden in Alzheimer's disease: a 12-month study (Abstract). European Federation Neurological Societies, Copenhagen, Denmark

McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan E M 1984 Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**: 939–44

Morris J C, Cyrus P A, Orazem J, Mas J, Bieber F, Ruzicka B B, Gulansk X 1998 Metrifonate benefits cognitive, behavioral, and global function in patients with Alzheimer's disease. *Neurology* **50**: 1222–30

Rösler M, Anand R, Cicin-Sain A, Gauthier S, Agid Y, Dal-Biano P, Stahelin H B, Hartman R, Gharaboni M 1999 Efficacy and safety of rivastigmine in patients with Alzheimer's disease: international randomised controlled trial. *British Medical Journal* **318**: 633–8

Rainer M 1997 Galantamine in Alzheimer's disease: a new alternative to tacrine? *CNS Drugs* **7**: 89–97

Raskind M A, Peskind E R, Wessel T, Yuan W 2000 Galantamine in AD. A 6-month randomized, placebo-controlled trial with a 6-month extension. *Neurology* **54**: 2262–8

Reisberg B, Schneider L, Doody R, Anaud R, Feldman H, Haraguch R, Lucca U, Mangone C A, Mohr E, Morris J C, Rogers S, Sawada T 1997 Clinical global measures of dementia. Position paper from Working Group on Harmonization of Dementia Drug Guide. *Alzheimer Disease Association Discord*. **11**(Suppl. 3): 8–18

Ringman J M, Cummings J L 1999 Metrifonate: update on a new antidementia drug. *Journal of Clinical Psychiatry* **60**(11): 776–82

Rogers S L, Farlow M R, Doody R S, Mohs R, Friedhoff L T, Donepezil Study Group 1998 A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. *Neurology* **50**: 136–45

Rogers S L, Friedhoff L T 1998 Long-term efficacy and safety of donepezile in the treatment of Alzheimer's diesease: an interim analysis of the results of a US multicentre open label extension study. *European Neuropsychopharmacology* **8**: 67–75

Rosen W G, Mohs R C, Davis K L 1984 A new rating scale for Alzheimer's disease. *American Journal of Psychiatry* **141**: 1356–64

Small G W, Rabins P V, Barry P P, Buchholtz N S, DeKosky S T, Ferris S H, Finkel S I, Gwyther L P, Khachaturian Z S, Lebowitz B D, McRae T D, Morris J C, Oakley F, Schneider L S, Streim J E, Sunderland T, Teri L A, Tune L E 1997

Diagnosis and treatment of Alzheimer disease and related disorders. Consensus Statement of the American Association for Geriatric Psychiatry, the Alzheimer's Association, and the American Geriatrics Society. *Journal of the American Medical Association* **278**(16): 1363–71

Spencer C M, Noble S 1998 Rivastigmine. A review of its use in Alzheimer's disease. *Drugs Aging* **13**: 391–411

Summers W K, Majovski L V, Marsh G M, Tachiki K, Kling A 1996 Oral tetrahydroaminoacridine in long-term treatment of senile dementia, Alzheimer type. *New England Journal of Medicine* **315**: 1241–5

Tariot P N, Solomon P R, Morris J C, Kershaw P, Lilienfeld S, Ding C, the Galantamine USA-10 Study Group 2000 A 5-month, randomized, placebo-controlled trial of galantamine in AD. *Neurology* **54**: 2269–76

Wagstaff A J, McTavish D 1994 Tacrine. A review of its pharmacodynamic and pharmacokinetic properties, and therapeutic efficacy in Alzheimer's Disease. *Drugs & Aging* **4**(6): 510–40

Wilcock G K, Lilienfeld S, Gaens E on behalf of the Galantamine International-1 Study Group 2000 Efficacy and safety of galantamine in patients with mild to moderate Alzheimer's disease: multicentre randomised controlled trial. *British Medical Journal* **321**(9 December): 1445–9

Wimo A, Winblad B, Grafstrom M 1998 The social consequences for families with Alzheimer disease patients: potential impact of new drug treatment. *International Journal of Geriatric Psychiatry* **14**: 338–47

Withehouse P J, Price D L, Struble R G, Clark A W, Coyle J T, DeLong M R 1982 Alzheimer's disease and senile dementia: loss of neurons in the basal forebrain. *Science* **215**: 1237–9

M. Hofmann and C. Hock

Alzheimer's Disease: Behavioral and Social Aspects

Alzheimer's disease (AD) is one of the two most common forms of *dementia* and constitutes a considerable *social problem*. It has a pathological and genetic base but produces progressive disruption of *psychological functioning* with considerable *social consequences*, especially for *carers*. Forms of intervention are being developed which so far show at least some capacity to ameliorate the consequences of the disorder.

1. Background

AD is one of the two most common dementing illnesses; the other being cerebrovascular dementia. Normally associated with *old age*, it can occur earlier although this is extremely rare. Because AD has a slow, insidious onset and can be difficult to differentiate from other dementing disorders, determining *prevalence* is not something that can be done with great reliability. One authoritative estimate indicated prevalence between the ages of 60–69 of 0.3 percent, 70–79 of 3.1 percent and 80–89 of 10.8 percent (Rocca et al. 1991). In common with other prevalence studies of dementia, this shows prevalence to be very small in the lower age ranges but to increase more-or-less exponentially as age advances. Prevalence studies on all dementia forms combined for very old age, which include a large proportion of Alzheimer's disease, found between 22.0 and 58.6 percent demented subjects in the age range of 90–94 years and between 32.0 and 54.6 percent in the age range of 95–99 years (Ritchie and Kildea 1995).

The causes of AD are, as yet, ill understood. It is associated with pathological changes in the brain (Nordberg and Winblad 1996) consisting of atrophy and the occurrence or increased manifestation of certain microscopically identifiable features such as senile plaques and neurofibrillary tangles. Biochemical analyses indicate that several neurotransmitter systems within the brain are affected and these include the cholinergic system. There are also *genetic* influences and relatives of known cases are at increased risk. For the most part, there is no clear pattern of *inheritance* but there are a few well-described families where the pattern of incidence is that of dominant inheritance (Breitner and Folstein 1984).

Certain things follow from the above. The fact that neurotransmitter systems may be involved raises the possibility of developing pharmacological treatments that may slow down or even arrest the progress of AD. A second point is that the definitive diagnosis of AD can only be done by postmortem examination of the brain. This is partly why getting accurate prevalence figures is difficult but also explains why many authorities refer to 'dementia Alzheimer type' or 'senile dementia Alzheimer type.'

AD is a slowly progressive condition that also reduces life expectancy. It has many psychological and social consequences and it is these that form the main concern of this article.

2 Psychological Manifestations

As is the case with most dementing disorders, the psychological manifestations of AD are very varied and affect almost all aspects of functioning.

2.1 Cognitive Abilities

The very term 'dementia' implies a loss of cognitive ability and it is hardly surprising that deterioration in a wide range of cognitive functions is apparent in AD. Lapses of *memory* are most commonly the first thing that alerts sufferers, or more typically their family, friends, and colleagues, to the fact that something is going wrong. Sufferers become generally forgetful and their behavior becomes disorganized as a result.

Memory disorder in dementia has been the subject of a considerable amount of research based on various models of memory developed by cognitive psychologists (Morris 1996). *Explicit memory* or the ability to acquire and remember specific items of information (e.g., the fact that a computer password is 'Victoria') has been explored in detail. A common theme running through this work is that those with AD have particular difficulty in processing incoming information and laying down new memory traces based on that information. In contrast and at least in the earlier stages, recollection of material from the distant past remains relatively well preserved.

Another aspect of memory is *implicit memory* or the ability to acquire and remember rather more general skills such as the ability to ride a bicycle. This appears relatively well preserved in AD as exemplified by the ability to learn and retain motor skills such as the pursuit rotor task. One interesting form of implicit memory task is known as '*priming*.' In priming experiments, the occurrence of a stimulus on one occasion is found to influence a second and unrelated task carried out some time later.

Participants may be presented with a list of words such as 'elephant' and 'garage' and asked to make a judgment about these words (e.g., the attractiveness of the object referred to by the word). Some time later, participants are asked to do an apparently unrelated task in which given sets of letters such as 'ELE' are presented and are asked to give a word beginning with the same letters. Normal groups are then more likely to respond with 'elephant' as compared to other possible responses such as 'element' given that the interval between the two tasks is not too great.

This phenomenon of priming is also manifested in certain groups with severe *amnesias* (such as the alcoholic *Korsakoff syndrome*) despite their extremely poor performance if asked to simply recall the words given in the first part of the task. Studies of priming in AD have given more variable results but often there is impaired priming. Whatever this means, it does indicate that there is possibly something different in the memory impairment of AD as compared to other forms of severe amnesia.

Many other aspects of cognitive functioning also deteriorate in AD (Miller and Morris 1993, Morris 1996). One aspect that merits brief comment is *language*. The first manifestation of problems with language is often a difficulty in word finding. As the condition progresses speech becomes increasingly impoverished and the ability to understand language, whether spoken or written, deteriorates.

2.2 Other Psychological Manifestations

There are also psychological changes that do not fall into the cognitive arena. A range of *psychiatric* phenomena can emerge (Miller and Morris 1993). Many people with AD are depressed and some psychotic symptoms such as hallucinations and delusions may occur. *Behavior disturbances* can also arise and it has been estimated that about a fifth will show aggression and aimless wandering. *Sexual disinhibition* is a relatively infrequent feature but raises considerable problems for carers when it does arise.

3 Social Aspects

In common with the other dementias, AD affects not only the immediate sufferer but has a wider social impact for family, friends, and the community (Miller and Morris 1993, Wilcock 1993) as well as economic implications (Wimo et al. 1998).

3.1 Impact on Carers

In most countries, the majority of people with AD and other forms of dementia are living in the community. The main burden of care, therefore, falls on relatives (usually the spouse or child). The burden of care is considerable in that carers typically report that they are unable to leave sufferers alone other than very briefly. Sufferers wander round the house, cannot hold a sensible conversation, are unsteady on their feet, and so on. Although incontinence can be a major problem, it fortunately does not usually arise until relatively late in the illness.

There is ample evidence that caring for someone with dementia results in considerable strain and distress. As far as measures of psychological distress and psychiatric symptomatology are concerned, carers typically emerge less well than comparable control groups, particularly in terms of levels of *depression*. There are also some indications that the physical health of carers can be adversely affected.

3.2 Factors Influencing Carer Well-being

It has been shown that the level of *stress or strain* in carers is not necessarily related to the degree of behavioral disturbance or physical incapacity in the sufferer although the indications are that the altered relationship with the sufferer is often a bigger source of distress than any physical burdens.

Carer distress is influenced by the quality of the past relationship with the sufferer. Where this has been high with a close, intimate relationship, carer distress and depression is lower. Another factor is gender in that it is a common finding that male carers fare better than female in terms of levels of strain and depression. Just why this should be the case is not clear. It may be the case that male carers attract more external support.

But it also may be that male carers are better able to distance themselves from the emotional aspects of the situation and adopt a practical and instrumental approach to the problems of caring. This latter form of coping style tends to be associated with lower levels of distress and depression. Finally, social support from outside the immediate household setting can be important but here it appears to be the case that the quality of support rather than its quantity is the key factor.

4 *Intervention and Management*

As indicated above, the identification of disturbed neurotransmitter systems in the brains of those with AD opens the way to pharmacological treatments that might retard or even halt the development of symptoms such as memory loss. This has been an active area of research and the development of an effective pharmacological intervention would have implications for those concerned with the psychological and social aspects of the disorder. Progress so far has been very modest but some substances such as tacrine, have been found to achieve small beneficial effects (Nordberg and Winblad 1996).

Despite marked memory impairments, those with quite advanced levels of dementia remain sensitive to environmental influences (Miller and Morris 1993, Woods 1996). In residential units, even such simple measures as grouping chairs in the dayroom round small tables rather than having them lined along the walls, has been found to increase the level of interaction between residents. Various special forms of intervention such as 'reality orientation' and 'reminiscence therapy' have been devised. These are based on discussion and the use of materials in small groups to, respectively, heighten orientation and awareness of the world around or the exploration of things from the time when sufferers were younger in order to contrast with the present. Beneficial changes have been demonstrated but again tend to be modest.

More individually based interventions have been used successfully. For example, behavior modification principles have proved of at least some value in dealing with difficult behavioral problems such as *incontinence* and aimless *wandering*. Whilst the emphasis in work on psychosocial interventions generally has been within group residential settings, some extension of interventions to the community and informal care situations has been made (Miller 1994, Miller and Morris 1993). Manuals are available offering practical advice to carers, of which one example is Hamdy et al. (1997).

See also: Aging and Health in Old Age; Alzheimer's Disease: Antidementive Drugs; Alzheimer's Disease, Neural Basis of; Caregiver Burden; Caregiving in Old Age; Chronic Illness, Psychosocial Coping with; Dementia: Overview; Dementia, Semantic; Memory and Aging, Cognitive Psychology of; Mental and Behavioral Disorders, Diagnosis and Classification of; Mental Illness, Epidemiology of; Suprachiasmatic Nucleus

Bibliography

Breitner J C S, Folstein M F 1984 Familial nature of Alzheimer's disease. *New England Journal of Medicine* **311**: 192

Hamdy R C, Turnbull J M, Edwards J, Lancaster M M 1997 *Alzheimer's Disease: A Handbook for Carers*. Mosby, St. Louis, MI

Miller E 1994 Psychological strategies. In: Copeland J R M, Abou-Saleh M, Blazer D G (eds.) *Principles and Practice of Geriatric Psychiatry*. Wiley, Chichester, UK

Miller E, Morris R G 1993 *The Psychology of Dementia*. Wiley, Chichester, UK

Morris R G 1996 *The Cognitive Neuropsychology of Alzheimer-type Dementia*. Oxford University Press, Oxford, UK

Nordberg A, Winblad B 1996 Alzheimer's disease: advances in research and clinical practice. *Acta Neurologica Scandinavia* **93**: 165

Ritchie K, Kildea D 1995 Is sevile dementia "age-related" or "ageing related"? Evidence from meta-analysis of dementia prevalence in the oldest old. *Lancet* **346**: 931–4

Rocca W A, Hofman A, Brayne C, Breteler M, Clarke M, Copeland J R M, Dartigues J F, Engedal K, Hagnell O, Heeren T J, Jonker C, Lindesay J, Lobo A, Mann A H, Morgan K, O'Connor D W, Da Silvaproux A, Sulkav R, Kay D W K, Amaducci L 1991 Frequency and distribution of Alzheimer's disease in Europe: collaborative study of 1980–1990 prevalence findings. *Annals of Neurology* **30**: 381–90

Wilcock G K 1993 *The Management of Alzheimer's Disease*. Wrightson, Petersfield, UK

Wimo A, Jonsson B, Karlsson G, Winblad B 1998 *Health Economics of Dementia*. Wiley, Chichester, UK

Woods R T 1996 *Handbook of the Clinical Psychology of Ageing*. Wiley, Chichester, UK

E. Miller

Alzheimer's Disease, Neural Basis of

Alzheimer's disease (AD), first described by the physician Alois Alzheimer in 1906, is an insidious progressive neurodegenerative disorder which can be detected clinically only in its final phase. A definitive diagnosis based on antemortem observations is complicated and often deceptive. A major histopathological criterion of AD, and one that is decisive for its post mortem diagnosis, is the assessment of distinctive alterations within the neuronal cytoskeleton, which appear in the form of neurofibrillary tangles in specific subsets of nerve cells of the human brain (see Sect. 1.1). Accompanying pathological alterations which, as a rule, appear later than the aforementioned intra-

neuronal changes, include extracellular deposits of the pathological protein beta-amyloid (see Sect. 1.2) and the formation of neuritic plaques (see Sect. 1.3).

1. The Neurodegenerative Process

1.1 The Intraneuronal Formation of Abnormal Tau Protein

An initial turning point in the degenerative process are pathological alterations of the cytoskeleton, a kind of internal moveable 'scaffolding' occurring in every living cell, which result from the formation of an abnormal (i.e., hyperphosphorylated) tau protein in a few susceptible types of neurons (Goedert et al. 1997, Goedert 1999). In healthy nerve cells the protein tau is one of several specific cellular proteins that is associated with and stabilizes components of the cytoskeleton. The abnormal but soluble material that emerges fills the entire nerve cell. The cell body and the cellular processes of such 'pretangle phase' neurons hardly deviate from their normal shape. In a second series of steps, this material aggregates to form virtually insoluble and nonbiodegradable filaments, the neurofibrillary tangles that are the hallmarks of AD. The latter, which often have a flame or comet-like appearance, gradually fill large portions of the nerve cell and appear black after staining with special silver techniques (Fig. 1a; Braak and Braak 1994, Trojanowski et al. 1995, Esiri et al. 1997). Nerve cells containing neurofibrillary tangles may survive for years despite marked cytoskeletal alterations. They forfeit many of their functional capacities, however, long before premature cell death occurs. After deterioration and disappearance of the parent cell, a cluster of the pathological material remains visible in the surrounding brain tissue as a remnant or so-called 'tombstone' tangle where it marks the site of the neuron's demise (Fig. 1a). The fact that 'tombstone' tangles are never observed in the absence of fresh neurofibrillary tangles accounts for the absence of spontaneous remission in AD patients. In the course of the illness, all of the involved nerve cells proceed through a 'pretangle phase' before developing the nonbiodegradable filaments. The potential for reversibility of the pathological process is most probably at its peak in the 'pretangle phase.'

1.2 The Extracellular Deposition of Abnormal Beta-Amyloid Protein

Among the many ingredients of the fluid filling the small space between brain cells known as the extracellular space are soluble proteins of still unknown function. Some of them probably result from processing of a normal component of nerve cell membranes

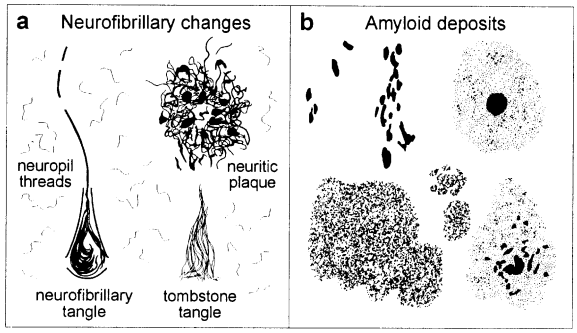

Figure 1
The hallmark of AD is precipitation of abnormal proteins in both intraneuronal and extracellular cerebral locations. **a**. The intraneuronal deposits represent the neurofibrillary alterations of the Alzheimer type and include three distinct kinds of lesions: strands of abnormal tau protein located within affected nerve cell bodies (neurofibrillary tangles) and within their dendritic processes (neuropil threads) as well as abnormal fibrous material which accumulates in swollen cellular processes of neuritic plaques. Extracellular 'tombstone' tangles mark the sites at which the parent nerve cells perished. **b.** The plaque-like extracellular deposits are composed chiefly, but not exclusively, of beta-amyloid protein. Depending on the texture of the neuropil (i.e., brain tissue consisting of nerve cell and glial cell cellular processes; see Section 1.2) they occur in different sizes and shapes. Most cortical beta-amyloid deposits evolve as globular structures with or without a condensed core.

called the amyloid precursor protein (APP). Under certain conditions and for reasons that are little understood, abnormal processing of the APP takes place, thereby resulting in the formation of the pathological beta-amyloid protein, a hydrophobic self-aggregating peptide (Selkoe 1994, Beyreuther and Masters 1997, Esiri et al. 1997). It is unclear whether all of the nerve cell types of the human central nervous system contribute to the production of the abnormal protein to the same degree. Within the extracellular space, the beta-amyloid protein eventually builds plaque-like deposits of varying sizes and shapes (Fig. 1(b)). Globular forms predominate in the cerebral cortex and occur preferentially in the cortical layers III and V. Band-like formations are seen in layer I of the cortex, whereas layers II, IV, and VI often are exempt. In the course of the disease, the number of plaques reaches a maximum; yet, even at their greatest extent, a large amount of brain tissue remains free of these deposits. At the present time, there is no clear evidence that deposits of beta-amyloid protein are capable of inducing the formation of neurofibrillary tangles within nerve cells, nor—in contrast to the neurofibrillary pathology in Sect. 1.1—do the beta-

amyloid plaques correlate with the degree of neuronal loss and/or clinical symptoms associated with AD, although intensive research has been devoted to establishing the existence of such connections (Hyman 1997).

1.3 The Appearance and Composition of Neuritic Plaques

Neuritic plaques consist of aggregations of altered glial cells and swollen cellular processes of nerve cells which, in part, contain fibrillary masses of the abnormal tau protein (Fig. 1(a)). Taken together, the various types of glial cells in the brain outnumber nerve cells by a ratio of approximately ten to one: They are the maintenance and repair 'troops,' so to speak, for the neurons. Deposits of beta-amyloid fill the extracellular space within the reaches of neuritic plaques. Neuritic plaques are more patchily distributed than simple beta-amyloid deposits and occur at much lower densities. Factors inducing the formation of neuritic plaques and the circumstances accompanying their disappearance from the tissue are unknown. A few cortical areas remain devoid of neuritic plaques, whereas others display high densities of these lesions quite early in the disease.

2. Developmental Sequence of the Lesional Distribution Pattern of the Intraneuronal Cytoskeletal Alterations

Neurofibrillary tangles develop gradually and nearly bilaterally symmetrically at predisposed sites within the cerebral cortex, thereafter overrunning other cortical sites and select subcortical nuclei. Many neuronal types, cortical areas, and subcortical nuclei remain uninvolved, while others undergo severe damage. This sequence of encroachment is predictable and remarkably consistent across cases, exhibiting little inter patient variability. By pinpointing the locations of the involved neurons and the severity of the lesions, six neuropathological stages in the evolution of the neurofibrillary changes can be differentiated (Braak and Braak 1999, Hyman and Trojanowski 1997).

2.1 Stages I and II

The first neurons to develop the pathology are specific projection cells located in small areas of the medial temporal lobe that are important cortical components of the limbic system. Bilateral structural preservation of these nerve cells is one of the prerequisites for retaining memory and learning capacities. Deposition of beta-amyloid protein is usually absent during development of these initial stages. The negligible to mild destruction still remains below the threshold required for the manifestation of clinical symptoms (see Sect. 2.2). Stages I and II represent the preclinical period of AD (Fig. 2, second line, involved areas are indicated by shading).

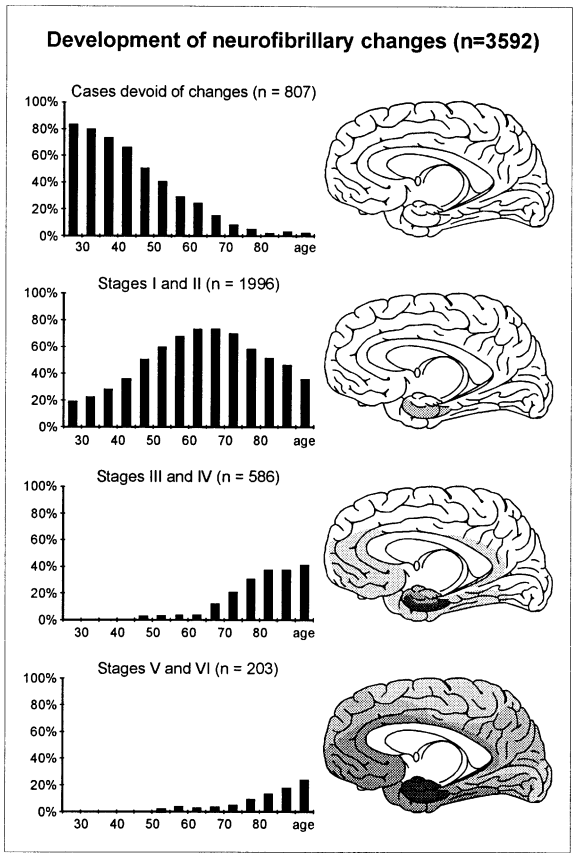

Figure 2
Development of neurofibrillary pathology in a total of 3,592 nonselected autopsy cases. In the first line, the relative prevalence of cases devoid of cytoskeletal alterations is shown for various age categories. Neurofibrillary lesions of the Alzheimer type are pathological and by no means normal concomitants of aging. The second, third, and fourth lines are similarly designed so as to show the gradual and sequential appearance of AD-related changes. Some individuals develop the initial lesions surprisingly early in life. Old age in itself is not an indispensable prerequisite for the onset of the neurodegenerative process: AD is an age-related, not an age-dependent disorder (see Braak and Braak 1994).

2.2 Stages III and IV

Subsequently, the lesions reach the hippocampal formation (stage III, and see Sect. 4) and then the

more distant neocortical destinations of the basal temporal lobe (stage IV). The clinical protocols of many individuals at stages III and IV may make reference to mild cognitive impairment (e.g., difficulties solving simple arithmetical or abstract problems), slight short-term memory or recall deficits (lapses of the so-called 'working memory'), and the presence of changes in personality ranging from suspiciousness to irascible or aggressive behavior, apathy, withdrawal, and mild or severe bouts of depression. Because such initial clinical symptoms often become manifest in stages III and IV, these cases can be referred to as clinically incipient AD (Fig. 2, third line, affected areas are marked by shading). In some patients, the appearance of symptoms still is obscured by their individual cognitive reserve capacities that are subject to influence by such factors as synaptic density, age, native acumen, education, head injury, stroke, or the co-occurrence of other neurodegenerative illnesses.

2.3 Stages V and VI

The pathology spreads superolaterally (stage V) and finally breaches the primary neocortical motor and sensory fields (stage VI). With the widespread devastation of the neocortex, patients present with severe dementia (i.e., acquired loss of memory, cognitive faculties, and judgment attended by a gradual dissolution of the personality.) The final stages V and VI correspond to clinically fully-developed AD (Fig. 2, fourth line, appropriate shading indicates the involved cortical areas). These persons become completely incontinent, are unable to dress themselves properly, to speak or recognize persons once familiar to them (spouse or children), and with the passage of time can no longer walk, sit up, or hold up their head unassisted. Increasing rigidity of major joints leads to irreversible contractures of the extremities and immobility. So-called 'primitive' reflexes normally seen only in infants also reappear in the last clinical phases of the illness (Reisberg and Franssen 1999). The period of time which elapses between the onset of clinical symptoms and death averages approximately eight and a half years (Francis et al. 1999).

3. The Relationships Between Age and the Evolutionary Stages of the Cytoskeletal Alterations

The relationships between age and the AD-associated cytoskeletal pathology can be studied by staging large numbers of nonselected brains obtained at autopsy (Braak and Braak 1997): Fig. 2 shows the percentage of cases at the various stages and according to their respective age groups. It illustrates a continuum of the cytoskeletal alterations ranging from the very first tangle-bearing nerve cell (appearing at stage I) to the extensive destruction encountered in fully-developed AD (stage VI).

The figure in the first line shows the prevalence of individuals whose brains remain entirely devoid of neurofibrillary pathology. It is important to note that a certain percentage of individuals, even at a very advanced age, refrain from developing AD-related cytoskeletal alterations (Fig. 2, first line). Neurofibrillary lesions cannot be viewed as normal concomitants of aging, even though their occurrence does in fact become more prevalent with increasing age (Hyman and Trojanowski 1997, Hyman 1998). Rather, neuronal impairment and, ultimately, destruction develop in a cell-type, layer, and area-specific fashion so that the regional pattern of nerve cell loss and atrophy (i.e., volume reduction) (see Sect. 4.1) in AD is not only quantitatively but also qualitatively different from the pattern encountered in normally aging brains.

Considerable interindividual differences exist with respect to the point at which the first neurons containing tangles actually are detectable. Many cases display a startling early onset. Often, initial lesions commence development in persons under 25 years of age, thereby implying that advanced age in itself is not a prerequisite for the development of the neurofibrillary pathology (Fig. 2, second line). The earliest cytoskeletal alterations are perfectly capable of evolving in otherwise healthy and young individuals. Accordingly, the initiation of the pathological process underlying AD is by no means an age-dependent one. Rather, it is typical of this disorder that several decades elapse between the onset of the lesions and phases of the illness in which the damage is extensive and severe enough for clinical symptoms to become apparent. Once begun, however, the destruction of the nerve cells involved progresses unyieldingly. Whether or not the neurodegenerative process underlying AD becomes clinically manifest depends solely on whether a given individual's life span permits it to attain its full expression (Braak and Braak 1997).

4. Imaging Technology and AD

4.1 Structural Imaging

Magnetic Resonance Imaging (MRI) is one of several imaging tools available to radiologists and neurologists for visualizing the blood flow and anatomical structures, including anomalies, of the human brain. Both post mortem and *in vivo* structural neuroimaging with MRI, for instance, reveal the presence of progressive atrophy in the hippocampal formation, an S-shaped or seahorse-shaped anatomical structure located in the medial temporal lobe, across a spectrum of persons suffering from age-related memory and

learning deficits associated with AD (Jack et al. 1992, Jobst et al. 1994). The hippocampal formation is responsible for, among other functions, short-term memory.

The clinical value of structural MRI brain imaging either in (1) persons considered to be at increased risk for developing AD (individuals at stage III with so-called mild cognitive impairment, the harbinger of AD) or in (2) patients with clinically diagnosed early AD, both of which groups exhibit hippocampal formation atrophy, is still purely predictive since the certain diagnosis only can be established histopathologically at autopsy (see opening paragraph). Furthermore, although the hippocampal formation becomes involved in AD relatively early (see Sect. 2.2), it is very difficult to determine *in vivo* whether progressive reductions in tissue volume that are detectable there with MRI (or those in any other given region) accurately reproduce actual nerve cell impairment or loss. This means that the use of *in vivo* MRI for early detection (e.g., in stages I–II) and diagnosis of AD is still limited and not entirely unproblematic without ongoing thorough clinical correlation. Finally, although AD brains in stages V–VI (see Sect. 2.3) generally are accompanied by macroscopically (i.e., gross) detectable ventricular enlargement, atrophy of the cerebral cortex, and a corresponding loss in brain weight, these features do not constitute specific or acknowledged criteria for AD alone.

Nevertheless, the long-range goal of *in vivo* structural MRI remains the development of a routine form of screening that will lead to the earliest possible correlation between stages of the pathology (ideally, in stages I–II while the 'prospective' patients still are just that, i.e., asymptomatic) and the clinical symptoms that emerge in the course of AD. In the meantime, when performed at regular intervals, one possible strength of *in vivo* MRI may reside in its potential not only to screen AD-candidates early but also to enhance the ability of physicians to monitor the clinical course of the disease, thereby assisting patients to preserve the quality of their lives for as long as possible and their families to plan and/or care for them more effectively.

4.2 Functional Imaging

In vivo neuroimaging with MRI makes possible visualization of activated cerebral regions during the performance of specific motor tasks or in response to external sensory stimuli. Functional MRI, however, differs from structural MRI in that the information obtained goes beyond the limits of brain morphology and topography alone. The operative principle behind the technology is that, in response to increased energy demands, blood flow in stimulated cortical regions of healthy persons may be elevated by as much as 20–40 percent and oxygen consumption by 5 percent so that activated brain tissue displays greater MRI signal intensity than nonactivated or insufficiently activated areas. Two additional *in vivo* diagnostic parameters which are measurable by means of a second functional neuroimaging technique, positron emission tomography (PET), are cerebral glucose metabolism and cholinergic neurotransmission: In individuals with mild AD, cortical hypoperfusion (decreased blood flow) can be seen in the regions involved using functional MRI. PET-scanned glucose metabolic rates are lowered in both temporoparietal lobes, and reduced cholinergic activity can be traced bilaterally as well as symmetrically in the cerebral cortex, including the hippocampus, during PET.

The practical aim of functional MRI and PET neuroimaging is the determination of the degree to which neocortical hypoperfusion and hypometabolism correlate with the severity of early AD-associated deficits. Nonetheless, even proponents of functional neuroimaging caution that, insofar as AD involves multiple neuronal systems, it is imperative for clinicians to know where in the brain and at which stages the neurotransmitter-specific cortical pathologies (e.g., cholinergic, serotonergic, GABA-ergic, noradrenergic) develop. Also, and perhaps even more importantly, clinicians need to understand mutual intersystemic implications for the patient's overall prognosis before intervening therapeutically (i.e., pharmacologically) (Francis et al. 1999). Moreover, the same dilemma mentioned in connection with structural imaging applies here as well, namely: The actual extent and severity of nerve cell damage or loss only can be surmised, not necessarily deduced or inferred, based on *in vivo* MRI or PET-detectable regional hypometabolism. Whereas the neurofibrillary pathology in AD very probably correlates with synapse loss, neuronal loss, and the clinical course of the illness, the complex causal interrelationships between the selective vulnerability of specific subsets of nerve cells, neurotransmitter-induced deficits, neurofibrillary tangle and/or beta-amyloid plaque formation, and the clinical picture of AD are only just beginning to come to light.

See also: Alzheimer's Disease: Behavioral and Social Aspects; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Dementia: Overview; Dementia: Psychiatric Aspects

Bibliography

Beyreuther K, Masters C L 1997 Serpents on the road to dementia and death. *Nature Medicine* **3**: 723–25

Braak H, Braak E 1994 Pathology of Alzheimer's disease. In: Calne D B (ed.) *Neurodegenerative Diseases*. Saunders, Philadelphia, PA, pp. 585–613

Braak H, Braak E 1997 Frequency of stages of Alzheimer-related lesions in different age categories. *Neurobiology of Aging* **18**: 351–57

Braak H, Braak E 1999 Temporal sequence of Alzheimer's disease-related pathology. In: Peters A, Morrison J H (eds.) *Cerebral Cortex*. Plenum Press, New York, Vol. 14, pp. 475–512

Esiri M M, Hyman B T, Beyreuther K, Masters C 1997 Aging and dementia. In: Graham D L, Lantos P I (eds.) *Greenfield's Neuropathology*. Arnold, London, pp. 153–234

Francis P T, Palmer A M, Snape M, Wilcock G K 1999 The cholinergic hypothesis of Alzheimer's disease: A review of the progress. *Journal of Neurology Neurosurgery and Psychiatry* **66**: 137–47

Goedert M 1999 Filamentous nerve cell inclusions in neurodegenerative diseases: Tauopathies and α-synucleinopathies. *Philosophical Transactions of the Royal Society London B* **354**: 1101–08

Goedert M, Trojanowski J Q, Lee V M Y 1997 The neurofibrillary pathology of Alzheimer's disease. In: Rosenberg R N (ed.) *The Molecular and Genetic Basis of Neurological Disease*, 2nd ed. Butterworth-Heinemann, Woburn, MA, pp. 613–27

Hyman B T 1997 The neuropathological diagnosis of Alzheimer's disease: Clinical–pathological studies. *Neurobiology of Aging* **18**(S4): S27–S32

Hyman B T 1998 New neuropathological criteria for Alzheimer's disease. *Archives of Neurology* **55**: 1174–76

Hyman B T, Trojanowski J Q 1997 Editorial on consensus recommendations for the postmortem diagnosis of Alzheimer disease from the National Institute on Aging and the Reagan Institute working group on diagnostic criteria for the neuropathological assessment of Alzheimer disease. *Journal of Neuropathology and Experimental Neurology* **56**: 1095–97

Jack C R Jr, Petersen R C, O'Brien P C, Tangalos E G 1992 MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. *Neurology* **42**: 183–88

Jobst K A, Smith A D, Szatmari M, Esiri M M, Jaskowski A, Hindley N, McDonald B, Molyneux A J 1994 Rapidly progressing atrophy of medial temporal lobe in Alzheimer's disease. *Lancet* **343**: 829–30

Reisberg B, Franssen E H 1999 Clinical stages of Alzheimer's disease. In: de Leon M J (ed.) *An Atlas of Alzheimer's disease. The Encyclopedia of Visual Medicine Series*, Parthenon, New York and London, pp. 11–20

Selkoe D J 1994 Alzheimer's disease: A central role for amyloid. *Journal of Neuropathology and Experimental Neurology* **53**: 438–47

Trojanowski J Q, Shin R W, Schmidt M L, Lee V M Y 1995 Relationship between plaques, tangles, and dystrophic processes in Alzheimer's disease. *Neurobiology of Aging* **16**: 335–40

H. Braak, K. Del Tredici, and E. Braak

American Revolution, The

Political philosopher Hannah Arendt wanted to reserve 'revolution' as an analytical term to constellations,

> 'where change occurs in the sense of a new beginning, where violence is used to constitute an altogether different form of government, to bring about the formation of a new body politic, where the liberation from oppression aims at least at the constitution of freedom, can we speak of revolution. …
> (Arendt 1963, p. 28)

Hence for Arendt, the American Revolution was the ideal revolution.

1. *From Resistance to Revolution*

Various acts of Parliament had restricted the colonists' trade since 1651. The stifling effect of this mercantilist policy for over a century is still a matter of debate: the prohibitions have to be weighed against the guaranteed market and the protection by the navy. The fact is that many a frustrated colonial merchant became a smuggler. However, the colonists did not rebel after 1764 against economic hardship because of severe exploitation. The war was fought to keep the future open for the economic development and self-government that many European settlers had experienced for several generations.

The annoying features of trade restrictions existed side-by-side with the strong self-government the English settlers had been allowed to develop, ever since the landowners of Virginia elected their first House of Burgesses in 1619. By the end of the French and Indian War in 1763, elected legislative assemblies balanced the power of the royal and proprietary governors, except in the newly acquired French-speaking colony of 'Canada.' The *de facto* free press mostly took the side of the local group of politicians critical of the governor's powers such as rewarding his supporters with patronage offices. A century and a half of experience with self-government in town meetings, county courts and representative assemblies had created a political class of plantation owners, yeoman farmers, merchants, master artisans, printers and writers, lawyers, clergy, and educators. They and the openness of their circles for 'new men' made the relatively smooth transition to complete self-rule possible, when imperial authority rule was toppled in 1775–76.

The battle cry 'no taxation without representation,' raised in 1764 against the Sugar Act, left room for negotiation. King George III, his Privy Council, ministers, and the majority Whigs in Parliament chose otherwise and enacted the Stamp Act in 1765. Only when organized mobs prevented the sale of the tax stamps by threatening the lives of sales agents appointed by the governor, and when the boycott of goods from England made London merchants suffer a two-thirds loss of their exports to the colonies in 1765–66, did Parliament repeal the Stamp Act. Coordination of this first stage of defying imperial rule was the work of 27 delegates from nine colonial assemblies meeting in

New York in October 1765. They reminded the king that 'it is inseparably essential to the freedom of a people, and the undoubted right of Englishmen, that no taxes be imposed on them but with their own consent, given personally or by their representatives.' From the beginning of the revolutionary discourse, the universalist natural-rights argument ('the freedom of a people') was used to justify rejecting direct taxation from London in addition to the rights the (uncodified) British constitution guaranteed British subjects. How to organize the expression of consent through fair representation on a large scale territories became the fundamental question of American federalism and was to remain a challenge until the ratification of the Constitution in 1788 and beyond. Its solution determined the success of the Revolution.

Repeal of the Stamp Act did not indicate a new pragmatism based on economic assessment and realistic political calculation. Together with the repeal, Parliament reasserted its constitutional power 'to make laws ... to bind the colonies and people of America, subjects of the crown of Great Britain, in all cases whatsoever.' (Declaratory Act, March 18, 1766). The colonial assemblies' assertion of their sole right to tax their voters and the crown's and Parliament's insistence on their all-inclusive power over the colonists clashed and could not exist side-by-side for long. To undermine the status of the colonial legislatures as providers of the governors' salaries, Parliament voted a new set of consumer taxes to fill the coffers of the Board of Trade in order to pay the governors and judges (Townshend duties, 1767). Once again, enough colonial merchants joined the non-importation movement to make the value of goods imported to the colonies from England drop from 2.1 to 1.3 million pounds Sterling in 1768 to 1769; in 1770 Parliament pulled back, but let one symbolic item stand: the duty on tea.

The final stage of escalation from resistance to revolution began when on December 16, 1773, Sons of Liberty dressed as Indian warriors dumped 342 chests of tea from anchored ships into the Boston harbor—the 'Boston tea party' of patriotic lore. (The governor had no constables to intervene.) The organizers behind the rioters wanted to prevent the regular landing of the tea through the customs office, because that would have meant recognizing the 'unconstitutional' duty on the tea and the East India Company's monopoly. After negotiations with Lt. Governor Thomas Hutchinson had failed, they resorted to violence, fully aware of the potential reaction in London. Slightly less violent opposition to the landing and distribution of the symbolically charged tea took place in New York, Philadelphia, and Charleston. Crown and Parliament reacted in May 1774 with the Coercive Acts, renamed by revolutionary propaganda the 'Intolerable Acts': (a) Boston harbor was closed until damages for the tea were paid; (b) crown officials accused of serious crimes could be tried in the UK in order to escape biased local juries; and (c) the position of the governor was greatly strengthened and the town meetings were drastically weakened. Plans to increase the military presence in all colonies could be seen behind the Quartering Act of June 1774 that allowed the billeting of troops in private homes in addition to public buildings and taverns that the Quartering Act of 1765 had permitted. Colonial opinion leaders also perceived the Quebec Act of May, 1774, as a threat to their future liberty, because it guaranteed the 70,000 Francophones cultural autonomy within the empire, i.e., the dominant position of the Catholic hierarchy in civil matters, the legal privileges of the *seigneurs*, and French civil law, were allowed to continue. None of the Anglo-Saxon bulwarks of the free British subject against arbitrary rule, such as an elected assembly and trial by jury were established. Worst of all, Quebec's boundaries were expanded southwestward to the Ohio and Mississippi. Parliament thus crushed with one stroke, the dreams of land speculators in the colonies with charters that left their western boundaries undefined.

Once again the authorities and politicians in London had miscalculated the effect of their measures. They wanted to punish Massachusetts and frighten the other colonies into submission; but they created solidarity expressed by the motto 'United we stand, divided we fall.' The legislatures from New Hampshire to South Carolina—some in illegal, revolutionary session—sent 56 delegates to the first Continental Congress in Philadelphia. On October 14 and 18, 1774, they declared the Coercive Acts and others to be unconstitutional, hence not binding, and called for the non-importation of British goods, the interruption of all exports to British ports, and ending the slave trade. These boycotts went beyond mere passive resistance, because only the threat of violence by self-appointed local 'committees' against disobeying merchants and other opponents made them effective. On February 9, 1775, the House of Lords voted down William Pitt's reconciliatory proposal and Massachusetts was declared to be in a state of rebellion. When Royal troops began marching into the countryside to empty the militia's armories, escalation to organized armed conflict was inevitable. On April 19, 1775, infantry and the farmer militias of Lexington and Concord near Boston clashed, 272 soldiers, and 93 militiamen died. The point of no return was passed. The political networks, that had articulated the colonial interests and led the resistance since 1765, allowed no power vacuum to develop. From May 1775 on, the Second Continental Congress meeting in Philadelphia was recognized as the sole decision-making body to speak for all 12 represented colonies (Georgia's assembly still hesitated). On June 15, 1775, it appointed George Washington general of an army in the making. Throughout the war and ever since, the army obeyed the elected civil authority; there was no room for a *caudillo* to stage his private revolution. Since Quebec and Montreal were of the utmost strategic importance, the Continental

Congress tried unsuccessfully to persuade the province to join the rebellion ('To the Oppressed Inhabitants of Canada,' May 29, 1775). Two military invasions in 1775–76 also failed.

2. *The War for Independence; the Loyalists*

The diplomatic and military situation in Europe at the end of the Seven Years War greatly favored the American cause. Having lost Quebec to Great Britain, France's absolutist ruler and his *conseil d'état* now saw a chance to weaken their adversary as a colonial and naval power, by making him lose his most precious colonies in North America. From 1775 on, a considerable amount of weaponry bought secretly with French public funds and traded for American goods like tobacco was smuggled into American ports. The two treaties of Alliance and of Amity and Commerce only followed in 1778, after the American forces had proven, in October 1777 with the battle of Saratoga in the Hudson valley, that they were worth supporting. Spanish and Dutch vessels also joined, and the war for American independence became an international naval war, as well as a guerrilla war on frontier and Indian territory, and in the more densely settled coastal areas from New York (the headquarters of the British Navy throughout the war) to Savannah. The final scene of the war illustrates the American–French military partnership and the close combination of war by land and by sea: On October 18, 1781, the last British army of 8,000 men under General Cornwallis capitulated in Yorktown on the coast of Southern Virginia. They were besieged by 9,000 Americans under General Washington and 7,800 Frenchmen under General Lafayette, but the trap only snapped shut when the British navy's rescue mission was intercepted by the French West Indian fleet under de Grasse.

The war divided the colonists into active 'Patriots' who joined Washington's troops directly or marched with their militia across state boundaries (which they were not obliged to do), 'Tories' or 'Loyalists' and, probably the largest group, the fence sitters. Exact numbers are not available. Military strategists in London miscalculated when they expected a significant swelling of the Redcoats' ranks by Loyalists; only about 50,000 men joined for more or less brief periods. Between 100,000 and 150,000 civilians fled the war to the peaceful part of British North America or Caribbean colonies or sailed 'home' to England. Over 5,000 of these refugees claimed partly substantial damages from a Royal commission (Breunig 1998, pp. 1–2, 8).

After months of tripartite peace negotiations in Paris, during which the humble republican citizens Benjamin Franklin, John Jay, and John Adams proved their skill as master diplomats in dealing with the advisors to the two most powerful monarchs of Europe, King George III recognized the victorious colonies as 'free, sovereign and independent States' (Treaty of Paris September 3, 1783).

3. *Revolutionary Ideology and Republican Constitutions*

The revolutionary moment in public debate occurred in January of 1776, when the Whig consensus of 1688 as formulated by John Locke evolved into American republicanism. Many colonists now recognized that they would never gain equality of rights within the empire. The advantages of complete independence and republican government were first openly discussed in Thomas Paine's incendiary pamphlet *Common Sense* (January 9, 1776). He broke with the ritualistic praise of the British constitution, demonstrated the absurdity of hereditary monarchy, and argued that the colonists could not be subdued by military force. Hundreds of newspaper articles and pamphlets by anonymous self-styled patriots followed. On July 2, 1776, the year-long power struggle in the assemblies and in the congress of their delegates in Philadelphia was decided with the vote of 12 colonies for independence. Two days later they laid their reasons before 'the opinions of mankind': the king of Great Britain 'has abdicated Government here' by violating the colonists' rights on 21 counts he has proven himself to be a 'Tyrant,' 'unfit to be the ruler of a free People.' The new nation's political creed was based on 'the Laws of Nature and of Nature's God' and was to apply to legitimate government worldwide: 'We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty, and the pursuit of Happiness. That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the Governed.' A collective right to revolution is expressed without use of the term: '… whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or abolish it.' That the rebelling colonists constituted 'a people' who can claim this right with impunity, is laconically asserted as a result of the King's *de facto* abdication.

The Declaration of Independence was just that, and established no legal norms for citizens to appeal to in a court of law; nor did it prescribe new institutions of government. The new internal order was agreed upon in the written state constitutions that 11 states gave themselves between 1776 and 1780. Alexander Hamilton characterized the new system as 'representative democracy.' In most states a two-chamber legislature elected by male property-owners for one to four years elected the governor from out of their midst. Institutional separation of power and functional cooperation protected the citizen from arbitrary government. So did certain rights of the individual

that had been discussed in English jurists since the Puritan Revolution and were now written into the text of the state constitutions, or in distinct Declarations of Rights. No majority of voters—the 'sovereign' of republican theory of legitimate government—could disregard them. The judges soon enhanced their role by nullifying acts of the legislators because they were 'unconstitutional.' Thus, even before sovereignty had been secured by force of arms, the first American republican or 'democratical' governments had been established at the state level. When the weak single-chamber Congress under the Articles of Confederation proved unable to cope with post-war problems—between the states and with Europe—it was replaced in 1789 by a more effective tripartite, balanced government modeled on the structure of the state governments and adapted to the needs of a federal nation. In 1791, a Declaration of Rights was added in the shape of the first 10 amendments to the Constitution. Federalism was the price for nationhood as defined in 1787, and states' rights encompassed and protected slave holding.

4. Historiography and Public Memory

Because making and preserving the nation were at stake, the Revolution and the War of Secession are the two events that have received by far the greatest attention of professional and lay historians and their different publics. Gordon (1989) assesses the major trends and works and their connections to the authors' intellectual milieu, and contemporary issues ranging from the 'Progressive' historians under the leadership of Charles Beard who emphasized economic interests and class-struggle elements among the colonists themselves to the neo-whig 'consensus' seeking intellectual historians in the1960s and 1970s (Bailyn 1992, Wood 1969) some of whom discovered a new 'republican synthesis' (Shalhope 1990). Others emphasized the influence of English and continental natural and constitutional thinking since Machiavelli (Stourzh 1970). New-left historians searched for patriotic radical roots in the founding period.

The Fourth of July as the supreme national holiday merges mythologizing the Revolution with celebrating successful nation building. Hence in times of national crisis, emotional reminders of the 'Founding Fathers' were meant to have a healing effect on the public mind. A prime example was the oratory on the occasion of the centennial in 1876, which coincided with the final phase of the post-Civil War 'reconstruction' of the union. The bicentennial of the Revolution was celebrated in a series of events from 1975–1983. President Ford and President Carter used acts of public memory from the re-enactment of the first skirmish on the village green of Lexington to the capitulation of the last British army to heal the wounds American national pride had suffered from the Vietnam War, and the constitutional crisis that had culminated in President Nixon's *de facto* impeachment in 1974. A presidential committee tried to coordinate public events and to set the patriotic tone of the public debate. The American Historical Association, the Organization of American Historians, and the Library of Congress organized special symposia and publications and reached out to the press and schools across the land and scholars abroad (*Library of Congress Symposia on the American Revolution*, published since 1972). The hoped for universal appeal of the Founders to 'the opinions of mankind' was appraised in the Library of Congress symposium on *The Impact of the American Revolution Abroad* (Library of Congress, 1976).

See also: American Studies: Politics; Arendt, Hannah (1906–75); Constitutionalism; Nations and Nation-states in History; Political Elites: Recruitment and Careers; Political Parties, History of; Revolutions, History of; State Formation

Bibliography

Adams W P 2001 *The First American Constitutions: Republican Ideology and the Making of the State Constitutions in the Revolutionary Era*. Transl. Robert and Rita Kimber, 2nd (enlarged) edn., Madison House, Madison, WI

Arendt H 1963 *On Revolution*. Viking Press, NY

Bailyn B 1992 *The Ideological Origins of the American Revolution*. Harvard UP, Cambridge, MA, enlarged ed.

Breunig M 1998 *Die Amerikanische Revolution als Bürgerkrieg*. [The American Revolution as a Civil War]. LIT, Münster, Germany

Countryman E 1985 *The American Revolution*. Hill and Wang, New York

Egnal M, Ernst J 1972 An economic interpretation of the American Revolution. *The William and Mary Quarterly* **29**: 3–32

Gordon C 1989 Crafting a usable past: Consensus, ideology, and historians of the American Revolution. *The William and Mary Quarterly* **46**: 671–95

Greene J, Pole J R (eds.) 2000 A companion to the American Revolution. Basil Blackwell, Oxford

Journal of American History 1999 Vol. 85 no. 4 with 14 articles on translating the Declaration of Independence into French, German, Italian, Spanish, Hebrew, Polish, Russian, Japanese, and Chinese

Maier P 1972 *From Resistance to Revolution: Colonial Radicals and the Development of American Opposition to Britain 1765–1776*. Norton, New York

Maier P 1997 *American Scripture: Making the Declaration of Independence*. Knopf, New York

Middlekauff R 1982 *The Glorious Cause: The American Revolution 1763–1789*. Oxford University Press, Oxford, UK

Palmer R 1959 *The Age of the Democratic Revolution: A Political History of Europe and America, 1760–1800*. Princeton University Press, Princeton, NJ

Reid J D 1978 Economic burdens: Spark to the American Revolution? *Journal of Economic History* **38**: 81–120

Shalhope R E 1990 *The Roots of Democracy: American Thought and Culture 1760–1800*. Twayne's, Boston

Stourzh G 1970 *Alexander Hamilton and the Idea of Republican Government*. Stanford University Press, Stanford, CA
Wood G 1969 *The Creation of the American Republic 1776–1787*. University of North Carolina Press, Chapel Hill, NC

W. P. Adams

American Studies: Culture

American culture has been a subject of continuous interest since the Revolution, when the question of how the newly independent states differed from England became a pressing concern. As an academic field, however, 'American studies' dates back only to the 1930s, when scholars of American history and literature began to develop—first at Yale and Harvard—an interdisciplinary framework for the study of American culture. In its early incarnation, American studies privileged literary analysis and the history of ideas as a means of understanding national character and culture. Since the 1960s, American studies has incorporated other traditions, subjects, and methods of inquiry—especially from the social sciences—which have moved it beyond its humanistic origins, although that early constellation of themes continues to exercise considerable influence. This diversity—some would say fragmentation—reflects an array of challenges to the concept of a singular and unified American culture. Increasingly, American studies serves less as a distinct field of inquiry than as an umbrella term for a range of topics and approaches that are only nominally and often contentiously organized by the concept of national culture.

1. *National Culture*

National culture is the pivotal and, in Wilfred Gallie's phrase, 'essentially contested' concept that until recently underwrote nearly all inquiry in this area. Did the United States have a national culture? If so, how could it be characterized? Answers to these questions tracked a number of important changes in the meaning of national identity and culture in the nineteenth and twentieth centuries.

Two general considerations are important in specifying how culture, especially, was understood in these circumstances and how it differed from other concepts that informed the study of national life—most prominently, society, civilization, history and politics. Since the early nineteenth century, culture has been conceived primarily as a realm of spirit, expression, character, or mind. In William Wordsworth's Romantic formulation, culture was the 'embodied spirit of a People'; in Ruth Benedict's classic anthropological definition it was a 'pattern of values.' Many scholars have commented on the shift from a humanistic definition of culture that privileged artistic expression (Wordsworth) to the anthropological view that emphasized culture as a complete and integrated 'way of life' or system of meaning (Benedict). With remarkable consistency, however, both paradigms distinguished culture from 'society,' understood as the actual practices and systemic relationships among individuals, groups, and institutions. Society, not culture, moreover, has most often been identified with processes of modernization—especially commerce, industrialization and the democratic leveling of social distinctions. This opposition has underwritten a long tradition of using culture as a basis for critiquing society. In the complicated sense introduced by early nineteenth century German Romanticism and Historicism, culture is both a bedrock of national 'character' that underlies society *and* a means of transcending society through expressions of 'high' culture—quintessentially art. Much the same relationship pertains to politics and history, for which culture sometimes served as an explanatory backdrop and sometimes as a refuge, means of escape, or vehicle for redemptive hopes. Culture has had an equally ambivalent relationship to the concept of 'civilization'—perhaps the most common interpretive framework for the study of the national experience before World War II. Studies of American civilization were generally structured around a belief in the progress of the nation as a collective historical project—one that embraced existing ways of life, history, and national achievements on the world stage. These tended to reproduce the distinction between society and culture, but could also recuperate that distinction, as in Lewis Mumford's seminal cultural history, *The Golden Day* (1926), which held the 'material fact' of civilization to be inextricable from the 'spiritual form' of culture.

The second consideration is that, since the early nineteenth century, the concept of culture has been shaped by both universalist and nationalist tendencies. From early use in reference to the cultivation of land, Enlightenment thinkers gradually appropriated the term to designate the broad, relatively unified field of human knowledge, as well as individual acquisition of that knowledge. In principle, culture was the labor and the product of all humankind, reflecting the Enlightenment's faith in the universality of knowledge—above all in the natural sciences and medicine, but with comparatively little difficulty extended to philosophy, law, art, and literature. This universalism is visible not only in the elitist view of culture as contained in the highest examples of human achievement—exemplified by Matthew Arnold's enormously influential late nineteenth century definition of culture as 'the best that has been thought and said in the world'—but also in much of the anthropological tradition's view of human culture as a singular achievement or a universal, if locally differentiated, human activity. The case for uniquely national culture, on the other hand, derives primarily from German Romanticism and

Historicism. It is this tradition that conceived the nation as the natural unit of culture, rooted in a strongly ethnic identification of culture with specific peoples and committed to forms of particularistic, nationally authentic cultural expression.

The universalist and nationalist approaches to culture coexisted and often competed, as evidenced by recurring debates between those who would judge cultural achievement in America by ostensibly universal principles, and those who viewed American culture as operating according to its own rules. Until quite recently, studies of American culture have belonged with few exceptions to the latter camp, implicitly accepting the nation as the appropriate frame of reference for culture even when they were highly critical of what they found. This perspective—organized especially around ideas of a unified national character, mind, or spirit—was not seriously inconvenienced by the transition from humanistic to anthropological priorities. Only recently has the fundamental assumption of the field—the coincidence of nation and culture—come under sustained scrutiny, leading to what some have described as a crisis or breakdown of American studies, and more generally to what Giles Gunn has called an 'indisposition to ask any longer certain questions having to do with…"the point of it all"' (1987).

1.1 The Cultural Problem in Early America

Early accounts of American culture came from a variety of sources, from European travelers seeking to explain the new country to home audiences (e.g., Francis Trollope, Charles Dickens, Alexis de Tocqueville) to the men and women of letters who formed a nascent cultural elite in New England and Virginia. Most of these accounts trafficked in the details of American manners, mores, and artistic accomplishments. Comparisons with Europe were common, and the United States was frequently found uncivilized and wanting. As James Fenimore Cooper (1824) argued, a great American literature was held back by a more general cultural failing—a 'poverty of materials…no annals for the historian; no follies (beyond the most vulgar and commonplace) for the satirist; no manners for the dramatist…' and above all no 'social capital' such as London or Paris which could provide 'a standard for opinion, manners, social maxims, or even language…'.

At the same time, the new nation did not lack for cultural entrepreneurs who saw in this absence a valuable new thing—a fresh start or a more natural, uncorrupted existence than was possible in Europe. Nor did it lack ideologues who could confidently predict the nation's glorious destiny. In both cases, writers fused nationalist sentiment with the millenarianism and redemptive rhetoric of the Puritan tradition, setting in place a quasi-mystical and extremely durable framework for interpreting the national experience and America's role in the world. The recurrent strain of 'exceptionalism' that runs through American cultural, social, and political analysis—sustained by faith in the unique mission or virtues of the United States—owes much to this tradition.

Whether as absent or as new, most partisans perceived the desirability of a distinctive basis for American life. Unlike in Germany, however, where claims for cultural identity paved the way for national aspirations and forged, through Romantic and Historicist writing, the 'natural' bond between nation and culture, American cultural identity was rarely presented as given: it was above all a problem and, for many, an opportunity.

Historians such as George Bancroft (1854) addressed this problem by writing American history as the ongoing expression of divine will—a civilizing errand into the wilderness shaped by the spirit of liberty. More consequential than this overt mythologizing, in many respects, was the emerging Romantic discourse of national cultural identity—delayed in America by some 20 years with respect to England and 30 with respect to Germany. Romanticism privileged artistic expression as a means of transcending society, which Romantics viewed as increasingly debased by the marketplace and other aspects of modernization. For most English romantics (e.g., Coleridge and Shelley), this transcendence was deeply rooted in the idealization of nature as a realm of wholeness and fulfillment. In the 1830s and 1840s, American literary elites reproduced much of this antimodernizing sentiment and idealization of nature, but also increasingly integrated the nationalistic dimension of Romanticism that associated artistic achievement with national character. The long established belief in America as 'Nature's Nation' facilitated this nationalist turn. Authentic American artistic achievement would express both nature and nation because for American Romantics, the two were much the same thing.

Where European romantics tended to present culture as a refuge from material society and art as a source of private epiphanies, the American tradition—especially after Ralph Waldo Emerson's work of the 1830s—focused on culture and its highest form, poetics, as a vehicle of social redemption. The rationale for these hopes lay in the Romantic philosophy of language, which viewed language as the medium of experience and poetics as the mastery of language. Properly conceived, therefore, poetry could unify art and life; it could reassert the unity of knowledge in the face of fragmented modern experience and redeem the promise of collective life from the tarnished realities of mass democracy. The difficulty was that this culture, by definition, did not exist. Romantic cultural criticism thus tended strongly toward prophecy: American culture was for the future.

This faith in the privileged relationship between literature and the nation, and in literature's power to crystallize an as yet nonexistent national culture, shaped a powerful, sometimes dominant, and usually highly critical American tradition of social thought that ran through much of the nineteenth and early twentieth centuries, connecting Emerson to Walt Whitman, Henry David Thoreau, later proponents of cultural invention such as Van Wyck Brooks, William Carlos Williams, Lewis Mumford, Frank Lloyd Wright, Alfred Steiglitz and Waldo Frank in the 1920s, and less directly to the cultural criticism of American Pragmatists such as John Dewey. Redemptive critique, in this context, represented one extreme of a more generalized view of culture associated with acts of creation and freedom from forms of social determination. As the nineteenth century drew to a close, this oppositional formulation grew sharper and more rarified. The defense of culture as transcendent expression shared ground with Arnold's view of culture as the best. Both implied a rejection of low or popular forms, and both became increasingly divorced from the modernizing forces that were transforming American society.

2. *Tradition and the Anthropological Concept of Culture*

Although the dominant nineteenth-century concept of national culture had enshrined ideas about national character and the myths that accompanied them, it had largely repudiated the question of tradition, which was viewed primarily as a form of constraint on individual freedom. The late nineteenth century, however, witnessed a remarkable fervor for tradition of all kinds—local history, ethnic heroes, nostalgia for the Old South, medievalism, vogues for colonial furniture, family genealogy, folklore and so forth. Nationalism underwrote a large share of this activity and increasingly came to be defined by it. Cultural practices and national history began to be understood as integral to national identity, and national identity came to be, in effect, the set of all sets of diverse local traditions and markers of identity. For the first time, this relationship began to be organized on a large scale—integrated into school curricula, promoted through Americanization campaigns, and built into the landscape through the proliferation of monuments, museums, and memorial events.

In many respects, the early anthropological tradition integrated and gave focus to this generalized passion for the past. In so doing, it challenged the humanistic notion of culture as creative expression in favor of a concept that took in the sum total of human activities and that embraced continuity rather than rupture. This new meaning initially referred to human culture in general, but increasingly acquired a discrete and pluralist sense that acknowledged the existence of separate and different cultures. The anthropologist Edward Tylor is usually credited with inaugurating this cultural turn in his celebrated 1871 description of 'Culture or civilization' as 'that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society.' Franz Boas's turn-of-the-century critique of evolutionary models such as Tylor's gave the concept its crucial pluralist inflection. Whereas for Tylor, culture referred to a singular evolutionary scale along which differen societies could be located, Boas introduced the modern sense that cultures are multiple, discrete, and derived from particular historical circumstances.

The 1920s were the watershed years for most of these intellectual and nationalist developments. The decade saw the rapid growth of social science research on American culture, especially through national character studies pursued with ethnographic methods. Boas's numerous students, including Ruth Benedict and Margaret Mead, popularized the anthropological view of culture and in many respects ensured that it would be understood in national terms. Robert and Helen Merrell Lynd's *Middletown: A Study in American Culture* (1929) showed how the new paradigm could be exploited. The 1920s also marked the high point of Progressivism as a school of historical analysis, with its signature debunking of national myths and revelations of economic interests, as well as a more general retrenchment of national identity along sharply nativist lines. It was the period in which American heritage and tradition became a veritable industry, promoted by clubs, historians, and the large-scale official sponsorship of the national past, including the construction of most of the monuments on the Washington Mall. At the same time, it was the effective end of the era in which redemptive literary ambitions in American cultural criticism could command serious attention. This was a function not only of the decline of literature as the dominant form of cultural expression, but—especially after World War Two—of the evident hegemony of American culture in the world, which brought an end to the inferiority complex that drove much of the American obsession with producing cultural achievements worthy of comparison to Europe.

2.1 *American Studies*

American studies was a product of this turmoil, but in some respects also represented a parting of the ways with developing social science approaches to culture. Strong points of commonality existed, particularly in regard to the notion of national character. Benedict's *Patterns of Culture* (1934) legitimized national character as the basis of anthropological studies of modern societies—now conceived in relation to values and psychology rather than spirit. Moreover, the critical animus of the humanist tradition remained central to

much social scientific research. Freed from the high-low dichotomy, social scientists were much more ready than their predecessors to identify American culture with the society it produced, and to see in that culture the foundation of a deeply conflicted modern life—at once materialistic and religious, individualistic and conformist, relentlessly innovative and fearful of change. Like their predecessors, nonetheless, many sought to rally the cultural resources of the nation in the name of a more genuine individualism (the Lynds), an active citizenry, or the independence of thought and action from large-scale social forces (Dewey). Reismans et al. *The Lonely Crowd* (1953) was perhaps the high point of this tradition with its indictment of modern corporate man and its nostalgia for ostensibly past values. Another symptom was the rediscovery of Tocqueville in the 1950s as a prophet of the dangers of conformity and mass society.

American studies was created in this period of popular ascendancy of the anthropological model of culture and the cult of tradition. Nonetheless, its major formulations were not greatly impacted by either. Part of the reason was professional. American literature became an acceptable topic of academic study only in the 1920s, and it quickly sought to escape the subordinate place assigned it within English departments. The new attention to American literature was directed less toward prophesizing or calling into being an American culture, however, than to confirming that search as the emblem of American culture itself. The other major force came from intellectual history—especially from the 'New History,' which focused on the role of ideas in American life and which similarly found itself marginalized in history departments.

Where earlier investments in American culture had been forward-looking, the contours of the new field were profoundly retrospective and historical. Where the cultural and historical analysis of the past decades had been largely critical in orientation—placing history in the service of reshaping American culture and debunking myths that had been honored mostly in the breach, the new scholarship tended to root American character more firmly in its faith in its myths. The monument to this new orientation was Vernon Parrington's multi-volume *Main Currents in American Thought* (1927), which went far toward establishing the operating paradigm of American studies in its first three decades. As Gene Wise (1979) has argued, this involved a loose consensual belief among scholars in the existence of a fundamental American character or 'mind' shaped by certain leading ideas and themes in American life. Individualism, Puritanism, Pragmatism, Progress, Transcendentalism, and Liberalism figured most prominently among these—though scholars could and did disagree about specifics. The humanistic ideal of high culture animated much of this work: the key national ideas, while visible in popular culture, were crystallized in the best literary and intellectual achievements. Among other virtues, this allowed intellectual historians and literary scholars to assert the privileged status of their enterprise. As an operating paradigm, this version of national culture underwrote a broad array of literary and historical work, from Perry Miller's studies of Puritan thought, to F.O. Matthiesson's account of democratic values in the American literary 'Renaissance,' to the work of the 'consensus' historians of the 1950s (Richard Hofstadter, Louis Hartz, and Arthur Schlesinger Jr., among others) who placed Lockean liberalism at the center of the American experience.

The strongest and, in many respects, field-defining version of this cultural logic belonged to the 'myth and symbol' school of the 1950s—a group that included Lionel Trilling, Henry Nash Smith, Charles Feidelson, R. W. B. Lewis, Charles Sanford, and Leo Marx. Lewis's *The American Adam: Innocence, Tragedy and Tradition in the Nineteenth-Century* (1955) and Smith's *Virgin Land: The American West as Symbol and Myth* (1950), to choose two prominent examples, worked from a widely shared but usually implicit set of assumptions about the way that forms of cultural expression—quintessentially literary texts—symbolically enacted and participated in larger national myths. The major work of the myth and symbol school fused intellectual history with the study of literature, uncovering myths of Adamic innocence, flights from civilization, the garden as a happy middle between city and wilderness, and a host of other recurrent patterns and images that allegedly defined American culture.

By the 1960s and 1970s, this work had become the establishment against which a newer generation of scholars chafed. The myth and symbol school was charged with sins of elitism and omission, with perpetuating versions of national exceptionalism and with taking an overly aestheticized view of culture that ignored power, institutions, class, gender, race and other forms of social stratification. Where myth and symbol scholars saw persistent myths, later scholars tended to see ideological formations that served particular interests, up to and including the scholarship that legitimized those ideologies as the essence of American life. But the fundamental question addressed by the myth and symbol scholars—the relationship between cultural products and larger cultural formations—did not thereby become obsolete, and their elegant if often constrained answers continued to inform a wide range of scholarship, from Sacvan Bercovitch's accounts of the Puritan intellectual legacy (1975) to Alan Trachtenburg's studies of the cultural impact of the corporation (1982).

2.2 The Pluralist Explosion

Although critics sometimes overstated the case against the myth and symbol school, there is little doubt of the magnitude of the changes in methodologies, topics

and underlying goals that reshaped the field beginning in the late 1960s. Perhaps most fundamental to this process was the contestation along multiple fronts of the idea of a singular, unified national culture. To the extent that a new consensus emerged, American culture became identified with a fundamental cultural pluralism—with a broad array of groups, traditions, and histories whose Americanness lay primarily in the acceptance of diversity as a national ideal.

The most direct of these challenges came from struggles for equality and recognition by historically marginalized groups—women, African-Americans, Native Americans, Hispanics, Asian-Americans, and later gays and other minority groups. American studies translated these social movements into as many distinct subfields, each primarily tasked with ensuring a place for its constituency in accounts of American culture. The emergence of social history or 'history from below' in the 1960s provided a methodology for many of these developments and underwrote a range of related projects of historical recovery: most prominently labor history, working-class culture, and regional studies of various kinds. They also integrated a wealth of new sources, from oral histories to folklore, popular culture, and material culture. Eugene Genovese, Sean Wilentz, and Herbert Guttman were among the pioneers in this area.

British cultural studies was another source of innovation in the 1970s. The work of Raymond Williams, E. P. Thompson, and others introduced a new style of critical and historical reflection on the concept of culture and a systematic challenge to the hierarchies of high and low culture that stratified social life. They participated in a broad groundswell of interest in how cultural products were understood and interpreted by their consumers, and helped to legitimize the study of popular and mass culture. Combined with structuralist and poststructuralist challenges to theories of interpretation (especially in regard to the primacy of authorial intention in the transmission of meaning) and the renewed influence of Marxist approaches to culture, the broad humanistic association of culture with autonomous artistic creation began to crumble.

In its place emerged a notion of culture comprised of and continuously reshaped by institutional forces. Aesthetic judgment and popular taste were now subjects with distinctive histories, and artistic creation was embedded in networks of social relations. These relations were not primarily among great artists across the generations, as in the romantic concept of tradition, but with the whole apparatus of culture: publishers, critics, buyers, technologies, opportunities for professionalization, and a range of other forces that organized cultural production into distinct fields. Although anthropological notions of culture superficially informed American studies since the 1930s—Henry Nash Smith's 1957 description of culture as 'the way in which subjective experienced is organized' was as close to a definitive statement as the field had—and although calls for greater attention to social science concepts and methods were regular features of American studies discourse, the emphasis on culture as a singular national 'pattern of values' made the specification of that culture a highly selective process and undermined the full potential of the anthropological critique. The new approaches did not dispense with the nation as a subject of analysis, but increasingly recognized national culture as a historically variable set of exclusive practices, whether built on aesthetic grounds or on compliance with certain myths, values, or exercises in national symbolism. The battles in English departments during the 1970s and 1980s over which works comprised the literary canon (and to some extent the highly politicized 'culture wars' of the late 1980s and 1990s) were symptoms of this intellectual transition.

A variety of social scientific approaches informed this transformation of the field, from the sociological accounts of culture developed by Talcott Parsons and Robert Merton, to ethnomethodological studies of everyday practices of interpretation. Thomas Berger and Peter Luckman's *The Social Construction of Reality* (1966) had a major impact on how scholars reinscribed culture within social processes and reintegrated 'high' cultural performances within the general field of cultural practice. Clifford Geertz's anthropological essays in *The Interpretation of Culture* (1973) also strongly shaped considerations of cultural practice as an inclusive but semi-autonomous system of meanings—reliant on but never reducible to social or economic life. More recently, James Clifford's textual approaches to material culture and Pierre Bourdieu's sociological histories of taste and status have proved influential. In the process, American studies became a much more self-reflective field that is aware of its history as a set of evolving interpretive paradigms and increasingly institutionalized interests.

Perhaps the most radical challenge to date has come from the recent subfield of 'border' studies, which focuses primarily on the Chicano populations of the Southwest. Although the term loosely groups together a range of perspectives and concerns, border studies scholars have consistently drawn attention to how the ideal of national culture operates as a tool of domination of the Mexican-American population. In the process, they have rejected the use of the nation as the master unit of cultural analysis. Many of the themes of border studies and related areas of transnational studies—cultural hybridity, contact zones, cultural imperialism, and the border itself as a liminal space that denies fixed identities—have entered into American studies more widely, resulting in what Carolyn Porter (1993) and others have called the 'remapping' of the field not only around cultural pluralism around cultural struggle that crosses political borders. However far this particular remapping is pursued and however useful border studies metaphors prove in

describing cultural patterns and conflict in other settings, it is clear that American studies has moved into a somewhat paradoxical post-national phase, in which most work is predicated on rejecting the institutional premises of the field. For some American studies scholars, this represents a crisis—a breakdown of the legitimating idea that gives coherence to their activity and supports their institutional identity. For others, it represents a necessary critical turn from the study of American culture to the study of culture itself as an inevitably contested field of identity and meaning.

See also: Area and International Studies: Cultural Studies; Benedict, Ruth (1887–1948); British Cultural Studies; Cultural Evolution: Overview; Cultural History; Cultural Studies: Cultural Concerns; Culture, Sociology of; Historicism; Individualism versus Collectivism: Philosophical Aspects; Intellectual History; Nationalism: General; Pragmatism: Philosophical Aspects; Progress: History of the Concept; Romanticism: Impact on Social Thought; Tocqueville, Alexis de (1805–59)

Bibliography

Arnold M [1869] 1994 *Culture and Anarchy*. Yale University Press, New Haven, CT

Bancroft G 1854 *History of the United States from the Discovery of the American Continent*. Little, Brown, Boston, Vol. 1

Benedict R [1934] 1959 *Patterns of Culture*. Houghton Mifflin, Boston

Bercovitch S 1975 *The Puritan Origins of the American Self*. Yale University Press, New Haven, CT

Berger T, Luckman P 1966 *The Social Construction of Reality*. Doubleday, New York

Cooper J F [1824] 1972 Notions of the Americans. In: Ruland R (ed.) *The Native Muse*. E. P. Dutton, New York

Gallie W B 1967 *Philosophy and Historical Explanation*. Oxford University Press, Oxford, UK

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Gunn G 1987 *The Criticism of Culture and the Criticism of Culture*. Oxford University Press, New York

Lewis R W B 1955 *The American Adam: Innocence, Tragedy and Tradition in the Nineteenth-Century*. University of Chicago Press, Chicago

Lynd R, Lynd H M 1929 *Middletown: A Study in American Culture*. Harcourt Brace, New York

Mumford L 1926 *The Golden Day: A Study in American Experience and Culture*. Boni & Liveright, New York

Parrington V [1927] 1958 *Main Currents in American Thought*. Harcourt & Brace, New York

Porter C 1994 What we know that we don't know: Remapping American literary studies. *American Literary History* **6**(3): 467–525

Reisman D, Glazer N, Denney R [1953] 1961 *The Lonely Crowd*. Doubleday, Garden City, New York

Smith H N 1950 *Virgin Land*. Harvard University Press, Cambridge, MA

Smith H N [1957] 1999 Can American studies develop a method? In: Maddox L (ed.) *Locating American Studies*. Johns Hopkins University Press, Baltimore

Trachtenberg A 1982 *The Incorporation of America*. Hill and Wang, New York

Williams R 1983 *Culture and Society: 1780–1950*. Columbia University Press, New York

Wise G [1979] 1999 'Paradigm dramas' in American studies. In: Maddox L (ed.) *Locating American Studies*. John Hopkins University Press, Baltimore, MD

J. Karaganis

American Studies: Education

1. The Structure of the American Educational System

Laws providing for public schooling were on the books in some of the American colonies as early as the mid-seventeenth century, but this public schooling was typically associated with aid to the poor. The schools were not well funded, and they fell into disuse or disrepute (Edwards and Richey 1963, Chaps. 1–3). Schooling during the colonial and early republican period was a patchwork of forms similar to that found in England. Some children learned the rudiments of literacy and mathematics in neighborhood 'dame schools.' Those who were wealthy enough might attend a private academy for a fee and later attend one of the few colleges then existing for preparation for one of the learned professions or cultural finishing before embarking on a career in commerce. Apprenticeships, rather than formal schooling, were a popular means for learning trades and professions.

The great majority of American revolutionary leaders were advocates of free and compulsory public schooling for the primary grades, but it was not until the 1830s and 1840s that the 'common school' movement mobilized under the leadership of educators from the most urbanized states. Men such as Massachusetts Commissioner of Education Horace Mann promoted free and compulsory public schooling as a support for good citizenship in a democratic republic. The common school movement struggled against citizens who were opposed to paying taxes for public education and against church leaders who wanted to exercise control over the curriculum. By the mid-1840s, free and compulsory public primary schools had become institutionalized in New England and the middle-Atlantic states. The spread of public primary schooling to the western and southern United States occurred primarily in the 1850s, but attendance remained spotty and facilities primitive in many communities (Edwards and Richey 1963, Chaps. 9–10).

Since the mid-nineteenth century, the American

system of education has been distinguished from schooling in the rest of the world by its inclusiveness and by the high average number of years students remain in school. Indeed, by the latter decades of the nineteenth century, American educators were arguing that all children had a right to secondary education, sentiments that would not become common in Europe for more than a half century. Where schooling in other industrialized countries was dominated by the idea of elite preparation, schooling in the United States developed as a means of nation building. Behind these ideological differences lie a number of social differences between nineteenth-century America and Europe. No well-entrenched aristocratic or quasi-aristocratic groups existed in the English-speaking democracies of North America to guard the universities and secondary schools as bastions of a status-linked high culture. The enfranchised groups were overwhelmingly small-property owners. The interests of this small-property-owning class, particularly when joined to the evangelical force of Protestant idealism, greatly encouraged use of state power for purposes of creating a 'virtuous citizenry.' The pragmatic spirit of the small-property-owning classes also encouraged the use of schooling as a means of teaching economically useful subjects. The expansion of the system was greatly encouraged by the efforts of Northern European Protestants to 'Americanize' the children of immigrants from poorer and Catholic regions of Europe. Compulsory schooling took root earliest in districts and states with high proportions of Catholics (Meyer et al. 1979).

Compared to schooling in other industrialized societies, the American system has also been distinctive in its decentralized organization. American federalism gave local communities the primary role in school financing and organization. Many Americans favor local control on the grounds that schools are more responsive to the particular interests and concerns of the communities in which they are located. Variation in curriculum and organization has been relatively minor, however, because educational professionals have organized schools throughout the country in remarkably similar ways. (Some variation does exist; sex education, for example, has never been popular in more conservative school districts.) Local control also encourages sharp disparities in per student spending between suburbs and inner cities, a phenomenon much less evident in more centralized systems. In recent years, many states have taken over a significant proportion of school funding, which has resulted in more equal funding between wealthy and poor school districts.

During the early and mid-twentieth century, most European states divided children after their few years into separate institutional tracks connected to differentiated adult occupational fates. Americans, by contrast, resisted tracking students into ability based institutional tracks. Students instead received a generally similar curriculum well into secondary school. Even secondary schooling has been relatively undifferentiated, with courses in the 'general' and 'college preparatory' tracks dominating. As compared to other industrial countries, only a small proportion of American secondary school students enroll in programs that can be described as primarily vocational.

Another distinguishing feature of the American education structure is the size and diversity of its tertiary sector. Until recently, most industrialized countries severely restricted enrollment in colleges and universities by requiring students to pass rigorous secondary school-leaving examinations. By contrast, the American structure was largely unplanned, unregulated, and market-driven. From the beginning of the American republic, weak state control over higher education made it relatively easy for a wide variety of groups to open colleges. Elite colleges educated children from the upper professional and business strata of the Northeast. Military colleges prepared men for the officer ranks. Denominational colleges attracted the children of coreligionists. Other private colleges served women and African Americans. In the mid and late nineteenth century, public universities, chartered with land grants from the federal government, opened their doors to children of the middle classes. These universities were required by law to combine liberal education with programs offering training in the industrial arts and agriculture. Beginning in the 1920s, state normal schools developed into teachers colleges. And junior colleges, another American innovation of the twentieth century, provided opportunities for students who lacked academic confidence or could not afford to enroll in a four-year school.

The combination of relatively undifferentiated primary and secondary schooling and competition for students among an unusually large number of colleges and universities led to comparatively high rates of college going and college graduation. For most of the twentieth century, American students were at least twice as likely as students in Europe to attend institutions of higher education. Today, two out of three American students continue their studies beyond the secondary level—still a significantly higher proportion than elsewhere in the industrialized world. The proportion of professional and managerial jobs in the economy cannot fully explain this high rate of college attendance. Countries such as England and Belgium have had a similar proportion of professional and managerial jobs in their occupational structures but much lower rates of college going. Rather than hire college graduates for white-collar jobs, employers in these countries were more likely to promote able workers from the shop floor. High rates of college-going and college graduation in the United States must be seen as reflecting not only occupational change but also expansionary tendencies in the educational structure itself—and in the ideology of opportunity that supports that structure.

Once students enter postsecondary education, they are decisively differentiated by the selectivity levels of the institutions they enter and, to a lesser degree, by their major fields of study. Highly selective institutions and scientific-technical disciplines confer a significant advantage in the market for educated labor (Bowen and Bok 1998). The combination of relative homogeneity at the primary and secondary level and extreme differentiation at the postsecondary level is the opposite of the historical pattern in Europe (Allmendinger 1989).

2. *Education as a Social Institution*

One puzzle explored by institutional analysts is why schools in such a thoroughly decentralized system became so similar to one another. Historians have shown that the organization of primary and secondary schooling became relatively standardized during the late nineteenth and early twentieth centuries, as reformers trained under the leading theorists of educational administration, George Strayer and Ellwood Cubberly, took control of big city school districts and sought to replace the waning power of religion as a moralizing force with organizationally based social control (Tyack 1974, Tyack and Hansot 1982). The triumph of these 'scientific managers' moved the schools out of the hands of people who were obsessed with personally rooting out evil and put them into the hands of people who favored creating an orderly and progressive environment through rationalized structures of administration, clearly enforced rules, scientifically tested curriculum, and regular evaluation of student progress.

Although historically important, themes of social control have played a secondary role in institutional analyses by social scientists. In the 1950s and 1960s, at the height of the Cold War, social scientists took it for granted that achievement norms were both the major socializing force in schools and the basis for the school's increasingly important role in social selection. Some of the more interesting analyses of this period showed the unintended consequences of achievement norms and how extracurricular activities and peer groups helped to blunt the potential psychological costs of the 'achievement regime.' Parsons (1959) argued that the schools' increasing emphasis on achievement, while valuable for American society, also created fertile grounds for the development of delinquency among those who were unwilling or unable to compete academically. Jackson (1968) showed how the schools' emphasis on academic success encouraged conformist students to resort to cheating and other means of manipulating the system for personal advantage. In this work, peer groups were often treated sympathetically as useful balances to the schools' emphasis on achievement. Parsons (1959) noted that many jobs and community roles in modern societies required high-level social interaction skills but only moderate academic skills. He argued that success in the informal social life surrounding schools helped to train potential leaders as well as those potentially suited to occupational and other roles in which social interaction skills were essential. Coleman (1961) observed that extracurricular activities and adolescent peer groups provided alternative avenues to status for many students who were less academically inclined and thereby contributed to the psychological health of many adolescents.

In the 1970s, as college attendance began to become expected for the majority of secondary school students, institutional analysts began to turn a skeptical eye on the idea that achievement played a central role in the American educational system. Collins (1979) criticized the view that higher levels of education were required in modern societies because these societies produced jobs requiring intellectual skills. He argued that most job skills were learned on the job and that a person's status characteristics and political skills figured more prominently in advancement than cognitive ability. As an alternative to the conventional wisdom linking high levels of education to high-technology economies, Collins developed a neo-Weberian analysis of the role of educational credentials in monopolizing access to desirable positions. Constant pressures for inflation of educational requirements exist in such a credential society, because students realize that educational credentials have come to play the stratifying role that family resources and reputation once played. Meyer and Rowan (1978) developed a similar position, arguing that many individual differences in performance are systematically obscured by the 'ritual categories' of schooling. These standardized membership categories allow unequal people to be treated more or less equally. The category 'high school graduate,' for example, is treated as a meaningful element of the American social structure, even though high school graduates include some people who know a great deal and some who can barely read and write.

Major contributions of institutional analysts to the understanding of classroom-level teaching and learning environments have been fewer in number. Lortie's (1975) occupational analysis of school teaching is a notable exception. Lortie began by distinguishing the types of people recruited into teaching. The profession has been attractive historically to people concerned less with ideas or monetary success than those interested in the psychic rewards of students' affection. In addition, as compared with other professions, Lortie argued, school teaching is distinguished by five structural characteristics: work with large, heterogeneous groups of 'immature workers'; work that requires high levels of group concentration but is marked by many interruptions; work that has multiple goals rather than a single overarching goal (e.g., socioemotional development as well as cognitive development); work that is generally performed in isolation from col-

leagues; and the absence of distinct ranks in the career. The structural characteristics, according to Lortie, lead to a dominant defensiveness in the culture of teachers. This culture is marked by a desire to protect the sanctity of the classroom from outside intrusions, a tendency to rely on trial and error as a guide to practice, an and a tendency to teach to the few students who provide teachers with the majority of their psychic rewards. Lortie recommended that teaching move in a clinical direction, modeling itself on medicine and psychotherapy, but this suggestion generated little enthusiasm. His suggestion for introducing hierarchical stages into the teaching profession has, however, been adopted by many states.

Over the course of the twentieth century, American teaching clearly became less authoritarian and more student-centered, charting the typical trajectory of classroom instruction as societies move from early to advanced forms of industrial organization (Brint 1998, Chap. 5). Cuban (1993) argued that teachers implicitly differentiate between an inner core of instructional authority and an outer periphery of social relations. The core includes lesson content, teaching techniques, and tasks to be accomplished; the periphery includes the arrangement of classroom space, the amount of student movement, the amount of ability and interest grouping, and the amount of classroom noise tolerated. Teachers have yielded considerable control over the periphery of social relations, while maintaining their control over the instructional core (Cuban 1993). Indeed, recent comparative work has emphasized the nurturing qualities of American primary school teachers. Stevenson and Stigler (1992) found, for example, that teachers they surveyed in Asia chose clarity and enthusiasm as the most important attributes of good teachers, whereas teachers in the United States more often chose sensitivity and patience. This solicitous outlook is undoubtedly connected to the expansive tendencies of a system in which two-thirds of students continue to attend educational institutions into early adulthood.

3. *Inequality and Achievement*

Though widely appreciated as a foundation, institutional analysis has not provided the dominant focus for social scientists studying American schooling. Instead, social and economic policy concerns have encouraged social scientists to focus on two topics: (a) the schools' role in perpetuating or ameliorating inequality in society; and (b) the school and classroom-level influences on students' cognitive achievement.

As in England, the sociology of education was almost synonymous in the years after World War II with concerns about the schools' role in perpetuating and transforming social inequalities. Two perspectives have clashed in these analyses: one perspective arguing that mass schooling is important primarily for legitimating the intergenerational transmission of privilege (see, e.g., Warner 1949, Bowles and Gintis 1976), and the other arguing that mass schooling is important primarily for its role in selecting intellectually able and motivated students from throughout the class structure for higher level positions in the occupational structure (see, e.g., Conant 1940, Wolfle 1954).

The empirical evidence collected and analyzed during the period failed to corroborate either the theory of social class reproduction or the theory of meritocracy. The studies showed that most of the variation in people's adult occupational and income status could not be predicted by characteristics like social background, cognitive ability, and educational credentials. Some of the unexplained variation in later life fates has to do with the vicissitudes of companies, industries, and regions and with more individual histories of good and bad fortune. Both social background and measured cognitive ability show up as important explanatory factors for that part of the variation in people's adult attainments that can be explained—primarily because they both influence the likelihood that a person will obtain the high-level educational credentials that, in turn, provide access to most good jobs. Grades and test scores are the best single predictors of educational attainment, but, even so, family background never disappears as an independent causal factor. Family background helps to predict test scores, and it also has a modest direct effect on how much schooling a person is likely to receive controlling for the person's measured cognitive ability (Jencks et al. 1972; Featherman and Hauser 1978, Jencks et al. 1979). Other factors bearing on educational attainment include having an intact two-parent household, having families and friends who value education highly, taking courses that are academically challenging (particularly courses in mathematics and science), and having strong personal aspirations to succeed (Sewell and Hauser 1975, Jencks et al. 1983).

Findings on the effects of race and gender on educational attainment show a complex pattern of declining and continuing inequalities. Although the black-white gap on achievement tests has narrowed somewhat over time, the gap continues to be sizable (up to eight-tenths of a standard deviation on mathematics achievement tests) (Jencks and Phillips 1998). Because of lower test scores, blacks and Latinos are disproportionately placed in lower tracks in elementary and secondary schools, and they experience greater difficulty gaining entrance into selective colleges and universities. By contrast, gender is a declining factor in educational stratification. Girls perform as well as boys in school and on the great majority of standardized tests. In the United States, women now outnumber men in colleges and universities (perhaps because few jobs for female high school graduates pay well). Levels of gender segregation have declined markedly in most major fields and professional scho-

ols, although gendered patterns of specialization continue to exist—for example, surgery in medicine continues to be male-dominated field. Women also remain significantly underrepresented in the physical sciences and engineering (Jacobs 1995). Most important, neither minorities nor women gain the benefits from educational credentials than white men can expect; at each level of education, their occupational and income prospects are lower than those of white men.

Because the individual characteristics associated with mobility are correlated with one another, another useful perspective is to use group-level data to compare the educational opportunities of socioeconomic strata at different points in time and in different countries. In most countries, correlations between social origins and high-level educational attainments have remained remarkably stable since the beginning of the twentieth century in spite of a rapid rise in the number of years most people stay in school (Blossfeld and Shavit 1993). Only the United States and a few Scandinavian countries have succeeded in significantly equalizing opportunities between classes. Between 1945 and 1980, the United States showed decreasing inequalities between classes (Hout et al. 1993). However, this trend has reversed since 1980 as the economic conditions of the working classes have lagged behind those of the middle and upper classes and as the unsubsidized costs of attending college have increased.

In the 1980s and 1990s, social scientists studying American education increasingly turned their attention to a second major topic: school-related factors bearing on cognitive achievement. This shift reflected the dissatisfaction of many school reformers of the period with the idea that schools are powerless to overcome the effects of social and economic disadvantages in the larger society. It also reflected policy makers' concerns about the potential implications of low educational standards for the country's ability to maintain its international economic strength. During this period, social scientists investigated relationships between a number of school-related factors and levels of student achievement. They showed that academically oriented leadership, highly qualified teachers, a disciplined and orderly school environment, and the sheer amount of time spent on learning can affect average achievement test scores at schools, controlling for the composition of the student body (Coleman et al. 1982, Ravitch 1995). At the same time, studies continued to confirm that the individual and family-related differences children bring to school account for the vast majority of variation in student performance. For primary school children, only about 20–25 percent of this variation in performance lies between schools; and at the secondary school level, only about 10–15 percent of this variation lies between schools. These figures set an upper bound on how much schools themselves can expect to affect achievement inequalities (Coleman et al. 1966; Entwistle et al. 1997).

Bibliography

Allmendinger J 1989 Educational systems and labor market outcomes. *European Sociological Review* **5**: 231–50

Blossfeld H-P, Shavit Y 1993 Persisting barriers: Changes in educational opportunities in 13 countries. In: Shavit Y, Blossfeld H-P (eds.) *Persisting Inequality: Changing Inequality in 13 Countries*. Westview Press, Boulder, CO, pp. 1–24

Bowen W G, Bok D 1998 *The Shape of the River: Long-term Consequences of Considering Race in College and University Admissions*. Princeton University Press, Princeton, NJ

Bowles S, Gintis H 1976 *Schooling in Capitalist America*. Basic Books, New York

Brint S 1998 *Schools and Societies*. Pine Forge/Sage, Thousand Oaks, CA

Coleman J S 1961 *The Adolescent Society*. Free Press, New York

Coleman J S et al. 1966 *Equality of Educational Opportunity*. Government Printing Office, Washington, DC

Coleman J S, Hoffer T, Kilgore S 1982 *High School Achievement: Public, Catholic and Private Schools Compared*. Basic Books, New York

Collins R 1979 *The Credential Society*. Academic Press, New York

Conant J B 1940 Education for a classless society: The Jeffersonian tradition. *The Atlantic Monthly* **165**: 593–602

Cuban L 1993 *How Teachers Taught: Constancy and Change in American Classrooms, 1880–1990*, 2nd edn. Teachers College Press, New York

Edwards N, Richey H G 1963 *The School and the American Social Order*, 2nd edn. Houghton-Mifflin, Boston

Entwistle D, Alexander K L, Steffel Olson L 1997 *Children, Schools, and Inequality*. Westview Press, Boulder, CO

Featherman D L, Hauser R M 1978 *Opportunity and Change*. Academic Press, New York

Hout M, Raftery A E, O'Bell E 1993 Making the grade: Educational stratification in the United States, 1925–1989. In: Shavit Y, Blossfeld H-P (eds.) *Persisting Inequality: Changing Inequality in 13 Countries*. Westview Press, Boulder, CO, pp. 25–49

Jackson P W 1968 *Life in Classrooms*. Holt, Rinehart, and Winston, Troy, MO

Jacobs J A 1995 Gender and academic specialties: Trends among college degree recipients in the 1980s. *Sociology of Education* **68**(2): 81–98

Jencks C et al. 1972 *Inequality*. Basic Books, New York

Jencks C et al. 1979 *Who Gets Ahead?* Basic Books, New York

Jencks C, Crouse J, Mueser P 1983 The Wisconsin model of status attainment: A national replication with improved measures of ability and aspiration. *Sociology of Education* **56**(1): 3–19

Jencks C, Phillips M (eds.) 1998 *The Black-White Test Gap*. Brookings Institution Press, Washington, DC

Lortie D 1975 *Schoolteacher*. University of Chicago Press, Chicago

Meyer J W, Rowan B 1978 The structure of educational organizations. In: Meyer M (ed.) *Environments and Organizations*. Jossey-Bass, San Francisco, pp. 78–109

Meyer J W, Tyack D B, Nagel J, Gordon A 1979 Public education as nation-building in the America, 1870–1930. *American Journal of Sociology* **85**: 591–613

Parsons T 1959 The school class as a social system. *Harvard Educational Review* **29**: 297–318

Ravitch D 1995 *National Standards in American Education*. Brookings Institution Press, Washington, DC

Sewell W H, Hauser R M 1975 *Occupation and Earnings: Achievement in the Early Career*. Academic Press, New York
Stevenson H W, Stigler J W 1992 *The Learning Gap*. Summit Books, New York
Tyack D B 1974 *The One Best System*. Harvard University Press, Cambridge, MA
Tyack D B, Hansot E 1982 *Managers of Virtue: Public School Leadership in America, 1820–1980*. Basic Books, New York
Warner W L 1949 *Social Class in America*. Science Research Associates, Chicago
Wolfle D 1954 *America's Resources of Specialized Talent*. Harper and Row, New York

S. Brint

American Studies: Environment

1. *The Nature of American Environmental Studies*

Environmental studies focus on interaction between ecological (ecosystems) and human societies. Ecosystems are those organized forms of natural system that integrate biotic organisms and their resource base, including air, water, and minerals. Modern scholarship notes: (a) human societies are now seen as dependent upon more features of ecosystems, and (b) ecosystems sustain human societies in ways apart from their traditional role as 'natural resources.' Natural resource views of ecosystems led to increasing pollution and depletion from societal production, exposing citizens and social institutions to new vulnerabilities from ecological scarcities. Prior to 1965, social analyses primarily traced how features of natural systems shaped the structure, location, and activities of communities, in work such as spatial ecology (Theodorsen 1961) and human ecology (Hawley 1950). Then, ecosystems were seen as autonomous from human actions.

Central problems in modern environmental studies include the complexity of relationships between social and ecological organization. It is often difficult to offer precise statements about the ecological impact of human activities, on the one hand, and the social impact of environmental disorganization, on the other. Moreover, boundaries between social scientists' and natural scientists' roles are unclear. Evaluating patterns of social impact on ecosystems and/or ecological impact on social systems does not permit the rigor of laboratory control of variables, and thus a variety of differing assessments of such impacts emerge (e.g., Dietz and Rycroft 1987, Kraus et al. 1992).

In the United States, the social science approach incorporated mainstream natural science perspectives on environmental destruction. One approach engaged in surveying American individuals, to trace how much recognition of environmental problems had emerged, and how individuals had altered their values and behaviors in the light of this recognition (Dunlap and Van Liere 1978). Another set of analyses traced the changes in twentieth-century American institutions, assessing both how they had produced negative environmental impacts (Burch 1971), and how they responded to attempts to engage in policies of managed scarcity (Schnaiberg and Gould 2000) to meliorate this impact.

2. *Individual American Responses to Environmental Challenges*

In the late 1960s and early 1970s, social surveys traced how much the public had consciousness of pollution hazards, and their attitudes towards proposed government regulation of pollution. Higher education levels predicted more concern, but later studies showed more diffused anxieties. By the late 1970s, though, individuals showed more skepticism about the implementation of new environmental protection laws through environmental agencies. More respondents were cautious about expanding government action to protect ecosystems, especially those whose livelihoods devolved around natural resource usage (Dunlap and Van Liere 1978). Better-educated Americans still expressed stronger concerns about the social impacts of pollution. More pragmatic lines of research include people's willingness to pay for hunting and fishing licenses to support conservation agencies (Heberlein 1989) and other environmental controls.

With the rise of the energy crisis in the mid and late 1970s, a variety of studies explored citizen reactions. Many citizens favored government support for broader domestic oil exploration, despite new environmental risks from drilling and transporting offshore and onshore oil. The 'problem' denoted by these segments of American society was the scarcity of petroleum products. Conversely, though, other respondents felt this was an occasion for citizens to alter their behaviors, and to engage in more energy-conserving (and resource-conserving) actions. They favored smaller cars, more public transportation, more energy-efficient appliances, and new programs of recycling social wastes (Murray et al. 1974). Middle-income respondents favored this. In contrast, low-income groups lacked the means to change their use of energy, and high-income groups resisted this decrease in their standards of living.

Towards the latter part of the 1970s, public attention was increasingly drawn to toxic-waste problems, highlighted by the problems of Love Canal in New York State, and a series of other chemical threats to human health (Levine 1982). Many social scientists began to study the emerging grass-roots social movement organizations associated with these incidents of public health (Brown and Mikkelsen 1997). Other

studies concentrated on individuals' concerns with risk (Slovic 1987), and found that individuals perceived environmental health hazards as more severe than did experts (cf. Dietz and Rycroft 1987, Kraus et al. 1992).

In the early 1980s, attention in the scientific and media circles shifted from national pollution and energy problems to new issues of global environmental change, especially ozone depletion and greenhouse gas increases that would lead to global warming. Studies of individual attitudes incorporated this new concern, adding fears about global warming and the willingness to forego some use of energy to reduce this risk. Increasingly, though, such social surveys of individuals became more problematic, as matters of environmental policy and industrial use of natural resources became more removed from individual decisions and attitudes (Buttel and Taylor 1992).

Perhaps the most recent arena in which citizen attitudes seemed to play a role was in voluntary participation in recycling their household wastes. Ironically, much less research has been addressed to this, although for some social scientists, individual recycling became a general criterion for environmentally responsible behavior (Derksen and Gartrell 1993).

3. *American Institutional Responses to Environmental Issues*

Perhaps the closest institutional analogue to the individual surveys is the studies of diverse American environmental movement organizations. Early in the modern period of environmental concern, Albrecht and Mauss (1975) reviewed the history of conservation, recreation, and environmental voluntary organizations. These citizen-based organizations sometimes predated and at other times monitored government agencies. These agencies regulated parklands, forests, and natural resources, either for purely conservationist or utilitarian purposes (Hays 1969). Most of the studies of the 1960s and 1970s focused on older and emergent national organizations, such as the Sierra Club and Environmental Defense Fund. Yet only Mitchell (1980) actually studied their members. He discovered that the bulk of these had little experience with current and previous social movements (civil rights, anti-poverty, anti-war, and feminist). Most other analysts simply traced how national organizations had brought environmental issues to the national political agenda, and to public consciousness through the use of the mass media (Mazur 1981).

In contrast, later case studies of local or grass-roots environmental organizations offered considerable insight into the motives and means of ordinary citizen-activists in these organizations (Levine 1982, Brown and Mikkelsen 1997, Szasz 1994, Gould et al. 1996). Some of these case studies treated emergent 'environmental' protest as examples of collective behavior theories, while others used the resource mobilization perspectives developed during the civil rights and other modern social movements (Turner 1981). More recent studies often focus on environmental justice movements of peoples of color, and other lower social groups. Unlike middle-class mobilization around 'protecting the environment,' these recent grass-roots organizations often dealt with members' exposure to direct health risks from local toxic waste sources, both at home (Bullard 1994) and at work (Pellow 1998).

With the political appeal of inequality-oriented movements in the US (Szasz and Meuser 1997), and the rise of concern about global inequalities (Goldman 1998), social scientists began to study the emerging web of nongovernmental organizations. In the United States, clearing-houses of local movements emerged (Brown and Mikkelsen 1997). Responding to global inequalities, multinational linkages appeared in groups protesting in Seattle and Washington, DC in the late 1990s, the World Trade Organization and the World Bank, respectively.

Beyond these studies of protest organizations, social scientists have explored the capacity of social institutions to respond to environmental problems and environmental protection. Researchers in business/management schools have stressed the motivation and capacity of economic entities to carry out environmental management (Hoffman 1997). This approach is also consonant with a Western European approach: ecological modernization (Mol and Sonnenfeld 2000). This research details how leading industrial firms have incorporated ecological criteria into their operational decision-making.

Another set of researchers has examined the prospects and limits of American and European regulatory agencies to moderate the expansionary impulses of dominant national and multinational firms (Hawkins 1984, Landy et al. 1990). These expansionary pressures have been detailed by the theory of the treadmill of production (Schnaiberg 1980, Schnaiberg and Gould 2000). Investors seeking to maximize their share values pressured managers to expand, thereby creating rising demand for natural resources, while offering fewer benefits to workers themselves from the resulting environmental exploitation.

Over the period from the mid-1960s, normative schemes have been suggested by social scientists, working with and observing social activists. These efforts focused on creating an alternative form of production and consumption. The work of E. F. Schumacher (1973) was the first approach outlining a goal of alternative or intermediate technology. Evaluating the outcomes of these programs, Schnaiberg and Gould (2000) noted how the concept had become eviscerated of its mission in most applications. Successes of these projects were only temporary, and occurred in settings which were not of interest to major investors or firms in the treadmill of production.

A second proposal was for industrial ecology (Socolow et al. 1994), in which the waste products of one firm would serve as the feedstock for another, reducing the depletion of materials, polluting wastes, and water and energy needs. Noting that examples of such projects existed in Denmark, American analysts proposed to apply these principles in the US. However, few clear examples of such systems have emerged, usually because of the limitation of capital available for the linked technologies involved, as opposed to investments in existing single-firm technologies (Weinberg et al. 2000).

More recently, much of the logic of these earlier approaches has been reconstructed in the new ideal type of sustainable development (Baker et al. 1997, PCSD 1999). Many utopian goals have been put forward by social science and activist citizen groups, despite substantial resistance from existing institutions (Daly 1996b). Paradoxically, this concept was initially proposed by natural scientists, with an aim of sustainable biodiversity. The sociopolitical reality is that most institutions seek to achieve maximal ecological protection with minimal social change (Daly 1996a). This was confirmed by recent evaluations of urban recycling as an exemplar of sustainable development policies (Weinberg et al. 2000). A related example is eco-tourism, where sustainable development is highly contingent on limiting competition for local resources (Gould 1998). Other researchers have explored existing strategies of voluntary simplicity, eco-communities, and community-based production (Shuman 1998).

Future analytic and policy studies will require far more synthetic approaches by social scientists than have been present in the distinct lines of inquiry noted above, and that is a formidable challenge for social scientists.

See also: African Studies: Environment; American Studies: Society; Area and International Studies: Economics; Area and International Studies in the United States: Institutional Arrangements; Area and International Studies: Political Economy; East Asia: Environmental Issues; East Asian Studies: Economics; Environmental Economics; Environmentalism, Politics of; Latin American Studies: Economics; Near Middle East/North African Studies: Economics; South Asian Studies: Economics; South Asian Studies: Environment; Southeast Asian Studies: Economics; Western European Studies: Environment

Bibliography

Albrecht S L, Mauss A L 1975 The environment as a social problem. In: Mauss A L (ed.) *Social Problems as Social Movements*. J P Lippincott, Philadelphia, pp. 556–605

Baker S, Kousis M, Richardson D, Young S (eds.) 1997 *The Politics of Sustainable Development: Theory, Policy and Practice Within the European Union*. Routledge, London

Brown P, Mikkelsen E 1997 [1992] *No Safe Place: Toxic Waste, Leukemia, and Community Action*, rev. edn. University of California Press, Berkeley, CA

Bullard R D (ed.) 1994 *Unequal Protection: Environmental Justice and Communities of Color*. Sierra Club Books, San Francisco

Burch W R 1971 *Daydreams and Nightmares: A Sociological Essay on the American Environment*. Harper and Row, New York

Buttel F H, Taylor P T 1992 Environmental sociology and global change: A critical assessment. *Society and Natural Resources* **5**: 211–30

Daly H E 1996a Sustainable growth? No thank you. In: Mander J, Goldsmith E (eds.) *The Case Against the Global Economy*. Sierra Club Books, San Francisco, pp. 192–6

Daly H E 1996b *Beyond Growth: The Economics of Sustainable Development*. Beacon Press, Boston

Dietz T, Rycroft R W 1987 *The Risk Professionals*. Russell Sage, New York

Derksen L, Gartrell J 1993 The social context of recycling. *American Sociological Review* **58**(3): 434–42

Dunlap R E, Van Liere K D 1978 *Environmental Concerns: A Bibliography of Empirical Studies and Brief Appraisal of the Literature*. Bibliography P-44, Public Administration Series, Vance Bibliographies, Monticello, IL

Goldman M (ed.) 1998 *Privatizing Nature: Political Struggles for the Global Commons*. Rutgers University Press, New Brunswick, NJ

Gould K A 1998 Nature-based tourism and sustainable development. *Environment, Technology and Society Newsletter*, Spring: 3–5

Gould K A, Schnaiberg A, Weinberg A S 1996 *Local Environmental Struggles: Citizen Activism in the Treadmill of Production*. Cambridge University Press, New York

Hawkins K 1984 *Environment and Enforcement: Regulation and the Social Definition of Pollution*. Clarendon Press, Oxford, UK

Hawley A H 1950 *Human Ecology: A Theory of Community Structure*. Ronald Press, New York

Hays S P 1969 *Conservation and the Gospel of Efficiency: The Progressive Conservation Movement, 1890–1920*. Atheneum, New York

Heberlein T A 1989 Attitudes and environmental management. *Journal of Social Issues* **45**(1): 37–58

Hoffman A J 1997 *From Heresy to Dogma: An Institutional History of Corporate Environmentalism*. New Lexington Press, San Francisco

Kraus N, Malmfors T, Slovic P 1992 Intuitive toxicology: Expert and lay judgments of chemical risks. *Risk Analysis* **12**(2): 215–32

Landy M K, Roberts M J, Thomas S R 1990 *The Environmental Protection Agency: Asking the Wrong Questions*. Oxford University Press, New York

Levine A G 1982 *Love Canal: Science, Politics, and People*. Lexington Books, Lexington, MA

Mazur A 1981 *The Dynamics of Technical Controversy*. Communications Press, Washington, DC

Mitchell R C 1980 How 'soft,' 'deep,' or 'left'? Present constituencies in the environmental movement. *Natural Resources Journal* **20**: 345–58

Mol A P J, Sonnenfeld D A (eds.) 2000 *Ecological Modernization Around the World: Perspectives and Critical Debates*. Frank Cass, Ilford, UK

Murray J R, Minor M J, Bradburn N M, Cotterman R F, Frankel M, Pisarski A E 1974 Evolution of public response to the energy crisis. *Science* **174**: 257–63
Pellow D N 1998 Bodies on the line: Environmental inequalities and hazardous work in the US recycling industry. *Race, Gender and Class* **6**: 124–51
President's Council on Sustainable Development (PCSD) 1999 *Towards a Sustainable America: Advancing Prosperity, Opportunity, and a Healthy Environment for the 21st Century*. Government Printing Office, Washington, DC
Schnaiberg A 1980 *The Environment: From Surplus to Scarcity*. Oxford University Press, New York
Schnaiberg A, Gould K A 2000 *Environment and Society: The Enduring Conflict*. Blackburn Press, West Caldwell, NJ
Schumacher E F 1973 *Small Is Beautiful: Economics as if People Mattered*. Harper and Row, New York
Shuman M 1998 *Going Local: Creating Self-reliant Communities in a Global Age*. The Free Press, New York
Socolow R H, Andrews C, Berkhout A F, Thomas V 1994 *Industrial Ecology and Global Change*. Cambridge University Press, Cambridge, UK
Slovic P 1987 Perception of risk. *Science* **236**: 280–5
Szasz A 1994 *Ecopopulism: Toxic Waste and the Movement for Environmental Justice*. University of Minnesota Press, Minneapolis, MN
Szasz A, Meuser M 1997 Environmental inequalities: Literature review and proposals for new directions in research and theory. *Current Sociology* **45**: 99–120
Theodorson G A (ed.) 1961 *Studies in Human Ecology*. Row, Peterson, Evanston, IL
Turner R 1981 Collective behavior and resource mobilization as approaches to social movements: Issues and continuities. In: Kriesberg L (ed.) *Research in Social Movements, Conflicts and Change*, Vol. 4. JAI Press, Greenwich, CT, pp. 1–24
Weinberg A S, Pellow D N, Schnaiberg A 2000 *Urban Recycling and the Search for Sustainable Community Development*. Princeton University Press, Princeton, NJ

A. Schnaiberg

American Studies: Politics

Observers of American politics witnessed a number of extraordinary events in the final three decades of the twentieth century. Prominent among these were the resignation of President Nixon in the wake of the Watergate scandal, the end of the war in Vietnam, a tax protest movement which spread from California across the United States, the bicentennial of both American independence and the US Constitution, a foreign policy crisis triggered by US diplomats held hostage in Iran, the end of the Cold War, the Republican party winning a majority of seats in the House of Representatives after 40 years in the minority, and the impeachment of a president for only the second time in American history. Yet, for all of these remarkable occurrences, the study of American politics in this period is marked by persistent themes: a relentless questioning of the adequacy of institutional arrangements and the unresolved nature of citizenship.

1. Institutions

The last third of the twentieth century saw no major changes to US political institutions akin to the creation of the Constitution in the 1780s, the formation of mass political parties in the nineteenth century, or the development of the welfare state in the 1930s. Indeed, while this period has been marked by policy debates over the size or merit of particular government programs, there has been remarkably little fundamental disagreement over the proper role of government. The settled nature of US institutions, however, is in marked contrast to the scholarly concerns raised with respect to the performance of contemporary American institutions as well as the evidence of public dissatisfaction.

1.1 Representative Democracy in Doubt?

The US central government is not large in comparison with other advanced industrial democracies. Yet, in relative terms, the American state grew over the course of the twentieth century to become an important provider of benefits to individuals and organizations. The expansion of the state—'big government' in the political language of the US—has raised serious questions about the performance of American political institutions. One of the most sweeping criticisms has been put forth by Lowi (1979, 1985). In Lowi's view, the institutions of American politics ceased to function as intended since the advent of the New Deal when the traditional philosophy of limited government was replaced by a philosophy of 'interest-group liberalism.' Central to this critique is the view that power had shifted from the legislative branch to the executive. By delegating authority, the legislature had made itself nearly impotent, while the powers of the presidency increased not by Constitutional amendment but through public expectations. Although few scholars fully embrace Lowi's assessment, his concerns are widely evident in the study of US institutions especially with respect to the growing informal powers of the presidency (e.g., Tulis 1987).

Another manifestation of the dissatisfaction with US institutions is the concern over 'divided government.' In the US, divided government occurs when the major parties split control of the executive and legislative branches. The potential for divided party government is rooted in both the Constitutional system and the development of political parties as a mechanism for bridging the gap between the legislative and executive branches. Although divided party government has occurred throughout American history, it became the norm in the post-1945 era. The effect of divided government on institutional performance has been the subject of some debate. While some scholars and commentators regarded divided government as leading to institutional stalemate, the careful empirical

work of Mayhew (1991) offered persuasive evidence that divided government was not dramatically different from periods of unified party control. Mayhew's analysis was not the final word, however, and a number of scholars find more substantial differences between periods of unified and divided party control (e.g., Coleman 1999).

Whatever its effect on legislative productivity, divided government can be seen as symptomatic of a broader sense of dissatisfaction with American politics. This broader discomfort encompassed several related elements of change including the erosion of public support for the Democratic party and its New Deal/Great Society agenda of activist government, the decline of party organizations and their adaptation to a 'service party' role, the increased number and prominence of interest groups in American politics, and the perception that election campaigns had become increasingly vacuous. Public manifestations of dissatisfaction with American politics are evident in lower levels of public support for the various political institutions, declining levels of trust in government, and lower turnout in presidential and congressional elections.

1.2 Remedies

In response to dissatisfaction with US institutions and party politics, some scholars have advocated measures designed to enhance deliberation and participation. Barber (1984) advocates widespread public participation and deliberation over policy issues as an antidote to the flaws of representative government, but most theorists of deliberative democracy believe widescale deliberation is impractical given the size of the American polity and the complexity of contemporary policy issues. The most empirically developed models of deliberative institutions are representative rather than direct, and are not broadly participatory. Bessette (1994) and Mansbridge (1988) advocate strengthening the deliberative norms of existing legislative institutions by developing incentives and sanctions for legislators to deliberate over the merits of public policy and to seek common ground. Fishkin's (1991) deliberative poll establishes a separate forum outside government, composed of a national sample of citizens who are brought together to discuss specific policy issues. This group is immersed in balanced information and analysis of specific issues, encouraged to discuss these issues publicly, and then surveyed about their reflective opinions.

Though more realistic than widespread deliberation, representative models like Fishkin's deliberative poll are less capable of serving the practical and ethical interests that make deliberative institutions preferable to those that are only minimally deliberative. Dahl (1997) has suggested a model of deliberation that bridges the domains of the highly involved and informed representative deliberative bodies and the less involved and informed public. For significant policy problems that normal institutions have failed to solve he proposes that a nonpartisan expert commission present alternatives to representative groups of citizens in deliberative poll settings at the national and state levels; these poll results can then be used to spark a larger public debate. Dahl's proposal is creative, yet questions persist about how more active and reflective subgroups of the citizenry can communicate their deliberations effectively to the general public.

2. Citizenship and American National Identity

2.1 American Civic Culture

Since the mid-1960s there has been a shift in focus in the study of the development of American civic culture. The influential paradigm of liberal or Lockean America stemming from Hartz (1955) remains strong but has been challenged by scholarship emphasizing the civic republican intellectual roots of the founding period. Wood (1992) and others have drawn attention to this neglected republican intellectual tradition that valued civic virtue, participation in public life, and social cohesiveness, and worried about constitutional stability and corruption. These ideas waxed in the eighteenth century but waned in the nineteenth with the growth of commercial society and the growing attractiveness of market economics. Works in constitutional theory by Michelman (1988) and in political science by Dagger (1997) attempt to use elements of the civic republican perspective, such as the imperative of civic duty and deliberative public discourse, to critique and reconstruct trends in American civic culture. In contrast to its eighteenth century forebear, contemporary civic republicanism is less skeptical about market economics, more socially inclusive, and more concerned with protecting individual rights (Dzur and Leonard 1998).

The continued relevance of nationalist political movements in the twentieth century casts new light on American civic culture. In comparative and theoretical work on the subject of nationalism the American case typifies civic, as opposed to ethnic, nationalism (Greenfeld 1992). Both forms of nationalism designate political cultures that achieve coherent and unique identities. In such cultures citizens recognize a relation to each other, to previous generations, and to their territory, that distinguishes them from members of other nations (Calhoun 1997). Civic nationalism permits universal accessibility to that national identity so that race, ethnicity, creed, or previous national affiliation are not formal barriers to becoming a fully-fledged member of the nation. The features of the American case most prominent in discussions of civic nationalism are the relative importance of constitutional patriotism and the relative unimportance of

ethnic identity as binding social forces (Habermas 1992).

2.2 Rights and Duties of Citizens

The growth of the American state and its increasing regulatory, administrative, and social welfare functions was a *de facto* rejection of the Lockean model even though Lockean language remains prominent in public discourse (Skowronek 1982). Rights of American citizenship could no longer be seen simply as negative rights—rights of protection against harm or injustice by others. Rights-bearing in the social welfare state meant, for many Progressive era intellectuals, access to food, shelter, education, and other resources needed for human development. American liberalism in the thought of someone like John Dewey became a hybrid set of values. Certainly values such as equality and freedom of speech remained central, but Dewey (1935) also asserted a positive collective commitment to human development within the context of a culturally plural, participatory, and reflective public culture.

Just how expansive welfare commitments could be justified became a central question in academic and public discourse in the 1970s—a period marked by economic recession, inflation, and tax limitation revolt. Questioning of the welfare state by libertarians like Hayek (1978) was launched from the neoclassical platform of Lockean liberalism. Taxation for support of positive rights could only be justified if it had the free agreement of those taxed, otherwise one person's positive rights were another's exploitation. Two works of normative political theory, Rawls' *A Theory of Justice* (1971) and Dworkin's *Taking Rights Seriously* (1977), developed a sophisticated defense of egalitarian social policy. Rawls and Dworkin argued that at the core of a legitimate constitutional settlement was a commitment to treating all citizens with equal concern and respect, a commitment that required close attention to the worst-off members of society. These were liberal defenses of the social welfare state since they were built on the idea of individual right and freely given constitutional arrangements. Not surprisingly, like the social movements of the 1960s and 1970s, the defense of positive rights turned to the judiciary rather than to the legislative or executive branches.

The justification of social welfare on the grounds of individual right and, more generally, the dominance of 'rights talk' in public discourse, came under scrutiny in the 1980s by academics and public intellectuals with close affinity to the civic republican tradition (Glendon 1991). 'Communitarians' such as Sandel (1982) pointed to the absence of an adequate political sociology of positive rights. Positive rights, sociologically and normatively speaking, were duties grounded in shared historical experience and common values. Such duties could not be understood as matters of reciprocity or self-interested calculation. Communitarians pointed out that civic duties in a well-functioning polity sometimes demanded sacrifice, as exemplified by military service and political participation.

2.3 Multicultural Citizenship

Recognizing the relevance of cultural pluralism as a defining feature of American civic life, Progressive-era thinkers advocated a 'federated' system of cultural preservation (Kallen 1915). Opposed to nativist renderings of a monocultural America ethnically defined by early New England settlement, and critical of an emerging mass culture that threatened homogenization, these writers applauded American cultural diversity. Their America was fundamentally an immigrant country marked by geographical regions with concentrated settlements of French, Norwegian, German, Irish, and other groups. This diversity was to be preserved, as an antidote to cultural homogenization, by state and federal acts such as bilingual education. With his concept of 'double-consciousness,' du Bois (1903) contributed to the optimistic and mostly colorblind Progressive discourse a recognition of the role African-Americans played in constructing a distinctive American identity. His work brought out the suffering, as well as the pride, bound up in the hyphenated sense of self.

Both the suffering and the pride of American cultural pluralism have been strikingly important themes politically and socially since the early 1980s. This renewed cultural pluralism was sparked in part by concerns shared by Progressive era intellectuals—homogeneity and monocultural rhetoric in public discourse—but must also be attributed to the success of liberal social movements of the 1960s and 1970s. For contemporary cultural pluralists, equal treatment under law leaves unsatisfied needs for dissimilar treatment. Politically, this view has led some to argue for group vetoes or other extrarepresentational means for preserving cultural integrity and for achieving positive affirmation of difference (Young 1990). Representation is an issue too for historians and literary critics who have pressed for a reconstruction of American history and culture that attends a multiplicity of voices. Their purpose is not simply to add neglected voices—say Mexican-Americans—to the traditional historical or literary canon, but to demonstrate how American arts, letters, and politics could not have been what they were and are without the experience of such groups (e.g., Morrison 1992). For individuals, cultural pluralism reveals new complexities in the struggle to construct personal identities. This individual struggle plays out politically when, for example, government tries to classify people using the part ethnic, part racial, part political categories of the census (Hollinger 1995).

The political impact of the multileveled category of cultural identity has come under scrutiny. Some have

argued that 'identity politics' endangers the commitment to common egalitarian ideals that marked the civil rights movement, both by emphasizing differences over commonality and by targeting group-specific political goals (Gitlin 1995). These critics note that traditional cleavages in American civic culture such as race, class, and gender still mark striking differences in individual achievements and life-plans and therefore still require civic solidarity rather than a politics of group difference. Other scholars and public intellectuals note that some differences comport poorly with others and some differences are downgraded in the discourse of cultural pluralism. How well the value of group self-determination comports with the frequently tradition-threatening value of gender equality is one concern (Okin 1999). How comfortably religious faith—a traditional difference between citizens and a difference that has historically provoked discrimination—fits into the category of 'identity' is another concern (Carter 1993).

3. *The Study of America*

The study of politics remains central to any effort to understand the American experience. Despite President Clinton's bold rhetorical claim that 'the era of big government is over,' the administrative state is well entrenched. Though there have been no major changes to US political institutions in the last three decades, scholars have raised concerns with interest group pressure, divided government, and public dissatisfaction with American politics. Creative proposals for encouraging greater public participation and deliberation are possible remedies, though much work remains to be done to gauge their effectiveness. Scholars have also sought to understand the dynamics of American civic culture during a period marked by struggles to justify substantive rights of citizenship and struggles to acknowledge cultural and other significant differences between citizens. Thorny issues persist regarding American citizenship and its intellectual heritage, with considerable disagreement about how the various components of individual and group identities ought to fit with contemporary notions of citizenship. There is, however, widespread recognition that debates over the rights, obligations, and even identities of citizenship will continue to mark this vibrant civic culture.

See also: American Studies: Society; Citizenship: Political; Civic Culture; Identity Movements; Multiculturalism; Political Parties, History of; Political Representation; Republican Party

Bibliography

Barber B 1984 *Strong Democracy*. University of California Press, Berkeley, CA

Bessette J M 1994 *The Mild Voice of Reason: Deliberative Democracy and American National Government*. University of Chicago Press, Chicago

Calhoun C 1997 *Nationalism*. University of Minnesota Press, Minneapolis, MN

Carter S L 1993 *The Culture of Disbelief: How American Law and Politics Trivialize Religious Devotion*. Basic Books, New York

Coleman J J 1999 Unified government, divided government, and party responsiveness. *American Political Science Review* **93**: 821–35

Dagger R 1997 *Civic Virtues: Rights, Citizenship, and Republican Liberalism*. Oxford University Press, Oxford, UK

Dahl R 1997 On deliberative democracy: Citizen panels and Medicare reform. *Dissent* **44**: 54–8

Dewey J 1935 *Liberalism and Social Action*. Putnam, New York

du Bois W E B 1903 *The Souls of Black Folk*. McClurg, Chicago

Dworkin R 1977 *Taking Rights Seriously*. Harvard University Press, Cambridge, MA

Dzur A, Leonard S 1998 The academic revival of Republicanism. Unpublished paper presented at the American Political Science Association annual meetings, Boston, September

Fishkin J S 1991 *Democracy and Deliberation: New Directions for Democratic Reform*. Yale University Press, New Haven, CT

Gitlin T 1995 *The Twilight of Common Dreams: Why America is Wracked by Culture Wars*. Metropolitan Books, New York

Glendon M A 1991 *Rights Talk: The Impoverishment of Political Discourse*. Free Press, New York

Greenfeld L 1992 *Nationalism: Five Roads to Modernity*. Cambridge University Press, Cambridge, UK

Habermas J 1992 Citizenship and national identity: some reflections on the future of Europe. *Praxis International* **12**: 1–19

Hartz L 1955 *The Liberal Tradition in America: An Interpretation of American Political Thought Since the Revolution*. Harcourt, Brace and World, New York

Hayek F A 1978 *Law, Legislation and Liberty, Volume 2: The Mirage of Social Justice*. University of Chicago Press, Chicago

Hollinger D 1995 *Postethnic America: Beyond Multiculturalism*. Basic Books, New York

Kallen H 1915 Democracy versus the melting pot. *Nation* **100**: 191–4, 217–20

Lowi T J 1979 *The End of Liberalism: The Second Republic of the United States*, 2nd edn. Norton, New York

Lowi T J 1985 *The Personal President: Power Invested, Promise Unfulfilled*. Cornell University Press, Ithaca, NY

Mansbridge J 1988 Motivating deliberation in Congress. In: Thurow S B (ed.) *Constitutionalism in America*. University Press of America, New York, Vol. 2

Mayhew D R 1991 *Divided We Govern: Party Control, Lawmaking, and Investigations 1946–1990*. Yale University Press, New Haven, CT

Michelman F 1988 Law's republic. *Yale Law Journal* **97**: 1493–537

Morrison T 1992 *Playing in the Dark: Whiteness and the Literary Imagination*. Harvard University Press, Cambridge, MA

Okin S M 1999 *Is Multiculturalism Bad for Women?* Princeton University Press, Princeton, NJ

Rawls J 1971 *A Theory of Justice*. Belknap Press, Cambridge, MA

Sandel M 1982 *Liberalism and the Limits of Justice*. Cambridge University Press, Cambridge, UK

Skowronek S 1982 *Building a New American State: The Expansion of National Administrative Capacities, 1877–1920*. Cambridge University Press, New York
Tulis J K 1987 *The Rhetorical Presidency*. Princeton University Press, Princeton, NJ
Wood G 1992 *The Radicalism of the American Revolution*. Knopf, New York
Young I M 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

A. W. Dzur and M. J. Burbank

American Studies: Religion

1. The Term 'Religion'

Because religion is a decisive element in much of American life, the nation's spiritual impulses, expression of faith, and religious embodiments are subjects of analysis in American Studies.

Defining religion is difficult in any circumstance, but it is notoriously so in a society that describes itself as religiously pluralistic. The word religion anywhere can refer to literature that points to the transcendent as well as to philosophy expressive of ultimate concern. It may mean dogma within formal religious bodies and also vague and individualized spirituality. Religion may refer to God but it may not. Normally it implies the awareness of a supernatural or suprahuman force or person that acts upon people who, in turn, respond. Scholars of religion in American studies are attentive to all this plus expressions in myth and symbol or rite and ceremony, in metaphysical concerns and behavioral correlates of faith. They may even study phenomena as concrete as church, synagogue, or mosque.

2. Earlier American Studies and Religion

Before American Studies became a formal complex of academic disciplines, and before these disciplines were professionalized, writers were emphasizing the role of religion in the nation. Clerics as early as Cotton Mather (1663–1728) were analyzing and, in his case, criticizing the cultural manifestations of religion in New England. Thus Mather's *Magnalia Christi Americana* (1702) concentrated on the pieties and religious declensions of New England. A century later Benjamin Trumbull (1735–1820) again concentrated on New England, arguing in 1797 that 'the settlement of New-England, purely for the purposes of religion, is an event which has no parallel in the history of modern ages.' He spoke of 'settlements' and 'sentiments' alike, and helped set American Studies on a path of New England concentrations.

Still an amateur at what became American Studies were figures like Edward Eggleston (1837–1902), a writer of fiction who saw the characters in his novels as 'forerunners of my historic studies.' A Methodist circuit rider turned self-described unbeliever after 1880, he transferred his curiosities from chronicling churches to pursuing new faith in science and criticizing the New England faith praised by people like Mather and Trumbull before him.

Most American Studies writers before the Civil War were amateurs who concentrated more on historical accounting than literary analysis. George H. Callcott tracked down the vocations of the 145 historians in *The Dictionary of American Biography* who did their main work between 1800 and 1860. Listing them by professional occupation he found that 34 were clergy, 32 lawyer-statesmen, 18 printers, 17 physicians, down to one historian. We stress this both to suggest how ill-defined were the means of disciplinary access to American Studies and to suggest why religion was programmed in to be such a major feature of these Studies in earlier times.

3. Professionalization of American Studies

The development of the modern university, marked as it is by differentiation and specialization, meant that professional historians and literary scholars pursued their ways rather independently of religious and theological scholars, particularly of those whose research focused on religious institutions. Such scholars tended to be segregated in divinity schools, often at the margins of universities, or in denominational theological seminaries out of range of graduate schools.

There were personal reasons for the marginalizing of religion in early American studies. The first generations of professional historians—Charles Beard, Carl Becker, Frederick Jackson Turner, James Harvey Robinson—all had intense childhood religion backgrounds, often in small towns. They moved from these to the urban and university scene, liberated from what they remembered as the confinements and low imagination of their churches. Most of them left behind, along with their faith, any positive curiosity about religion in American culture.

Meanwhile, when professional societies formed, they tended to downplay religion, so American Studies scholars preoccupied with religion went their own way. Thus shortly after the American Historical Association (AHA) was formed in 1884, an American Society of Church History (ASCH) under church historian Phillip Schaff developed rather independently of the secular organization. Schaff thought it to be an embodiment of 'the increase of rationalism.' Both societies attracted historians who also dealt with fields other than American. But when religious history did become the focus and American the topic, the AHA historians tended to dismiss religion. Mean-

while, the ASCH historians concentrated on the Christian churches, often seeing these apart from their cultural contexts.

4. American Religious Studies Outside America

The impression that religious studies were in a way parasitical, living off methodological developments in literature and history is reasonably accurate. Was there nothing that could be called religion standing independently in the academy? It happens that during the first half of the twentieth century, disciplines called History of Religion or Comparative Religion began to be imported form Europe. These were translations of what Germans called *Religionsgeschichte* or *Religionswissenschaft*. Giants such as Max Müller and Rudolf Otto had their disciples, importers, and adapters in America. Around the beginning of the twentieth century, six universities especially encouraged such religious studies: Boston University, the University of Chicago, Cornell University, Harvard Divinity School, New York University, and the University of Pennsylvania made commitments to pursue such 'scientific study of religion.'

Little came of these early efforts. There were few graduates and very few of these found academic employment. Most notably for our purposes, most of the pioneers chose to deal with what were then called 'primitive' or 'elementary' religions that were remote in space or time from the United States. James Freeman Clarke, Morris Jastrow, Jr., Louis H. Jordan, and George Foot Moore were notable scholars who flourished between the 1870s and the 1910s. None of them put energy into American Studies—including even Native American experience, which should have come into their purview. American Studies languished.

5. University Interest in America

Often seen as a time of turning came as a number of notable literary and historical scholars, most of them concentrating on New England sources, began to discern the revelatory character of religion in American culture. They concentrated chiefly on the long-discredited, even scorned New England Puritans. Far from dismissing these as idiosyncratic obscurantists, the new scholars took the Puritan metaphysic and piety seriously and suggested that these came to be suffusive elements far beyond the churches.

Thus Samuel Eliot Morison (1887–1976) admitted that he had once been derisive of these early Americans whose faith he did not share. But with Kenneth Murdock at his side he lifted up precisely the religious aspects of Puritan culture for positive viewing as intellectual forces.

The third of the rediscoverers of Puritan influence, usually regarded as the most significant shaper of American studies of religion in his generations, was Harvard's Perry Miller (1905–63). Miller was a formidable researcher who also did not share Puritan faith but argued that it was worthy of study, a major contributor to American cultural life. Miller was an intellectual historian who underplayed the social aspects of Puritan (and later early national) life. He did so while criticizing what he saw to be the reductionism of the social historians or the philosophy of history of the progressives. While Miller was to inspire reaction from both of these schools after his long prime (1933–63), he brought respectability of a new sort to religion in American Studies.

6. The Establishment of Religious Studies

The modern location of religious studies within secular, often tax-supported universities is usually traced to the interests of post-World War II citizens during a period of 'religious revival.' When the Soviet Union launched Sputnik in 1957, Americans responded by expanding universities. While this expansion naturally favored the sciences, the humanities also prospered and, with them, departments of religious studies. Through the 1960s and the 1970s the number of these burgeoned into the hundreds. The American Academy of Religion (AAR), the Society for the Scientific Study of Religion, the Society for Values in Higher Education (which had been the Society for Religion in Higher Education), the Religious Research Association, and any number of more specialized groups attracted and gave encouragement to thousands of professors.

While these religious studies departments included scholars who focused on everything from African Religions through Korean Religions to Womanist Approaches to Religion and Society—these being names of American Academy of Religion Groups—American Studies also found new subject matter. Thus the ARR hosted Afro-American Religious History; Asian North American Religion, Culture and Society; Church-State Studies; Hispanic American Religion, Culture, and Society; Native Traditions in the Americas; Pragmatism and Empiricism in American Religious Thought. These and many more provided windows on American Studies that went far beyond New England Puritanism and historical disciplines.

7. The Lie of the Land in American Studies: Religion

The most significant change over the century saw a move from ecclesiastically and confessionally-based studies, many of them having evangelistic or polemical intent, to phenomenological, either putatively disinterested or ideologically influenced accents that were not confined to particular religions (e.g., Freudian, Marxist, feminist, deconstructionist, and the like).

Some scholars, like those of a century earlier who had been in rebellion against their own youthful religious formation or who were seeking respectability in those parts of the academy and intellectual circles that are unfriendly to religion, adopted positivist stances and often expressed hostility to most religious expression.

Debates are waged as to whether religious studies should ever be pursued except through social scientific reductionist means. That is: if one 'appreciates' religion or writes with empathy as if from within the spiritual sphere, it is charged that this stance will skew scholarship and make the student of religion in America a biased agent. At the same time, others conversely suggest that religious participation may lead to an empathy that can inform inquiry without leading to distortion and bias. All the standard arguments on the old 'objectivity' and 'subjectivity' fronts get reworked in the service of various sides in these debates.

Since the religious theme in American studies is not pursued through a free-standing discipline—though with internal variety and argument—such as anthropology, but in a sense borrows from and contributes to existing disciplines and methods, religious studies is exposed to the various disputes in the disciplines, often with special intensity.

Thus the issues brought to the agenda by terms such as deconstruction and postmodernity are especially acute for those who study religion. This is partly the case because the final object of most religion—be it 'God' as in Judaism, Islam, and Christianity, 'Holy Emptiness' in Buddhism, or 'the Sacred' in general—eludes empirical analysis. Thus the scholars in the tradition of Morison, Murdock, and Miller, however congenial they be to the substance of Puritan thought, stand outside it. They are reduced to dealing with the human experience of the transcendental or supernatural, and not with the 'thing' itself. Most scholars of religion are content with this distancing from the subject, but some voices within religious communities argue that on such premises one cannot grasp the depth of religious commitment or meanings.

Similarly, in respect to deconstruction: the detachment of symbol from reality, or the rendering of connections between them as arbitrary, while it threatens all systems of meaning, appears to be most jeopardizing when these systems and symbols point to what Paul Tillich called 'ultimate concern.' Can one do justice to religion in American life while doing violence by scholarly inquiry into the attachments the religious citizens have to their objects? Conversely, ask others, can one do justice to religion in culture if one does *not* do such violence, wresting religious symbols and meanings from the privileged place adherents want to give them?

In the midst of such debates, some raise the question of the role of theology in all this. Most of the American people, movements, institutions, and forces that get categorized as religious, approach their ways of life with at least some broad and minimal set of theological affirmations. Theology in this context is not the same as faith or religious experience, but is a reflection on claims made by people of faith and piety. It is a second-order category, something that may be described as an interpretation of personal or public life, the life of a people (e.g., Hindus, lesbians, Asian-Americans) in the light of what they regard as a transcendent reference. What about the theologian who stands inside such believing communities?

There have been major students of American religion who are explicitly theologians. A prime example is H. Richard Niebuhr who wrote two classic works, *The Social Sources of Denominationalism* in 1929 and *The Kingdom of God in America* in 1937. Niebuhr belonged to a school of Protestant neo-orthodoxy that passed from vogue in the 1970s. But he used its critical yet faith-full vantage to trace several key ideas through an America that in 1937 was more decisively shaped by mainstream Protestantism than it could conceivably be construed as being thus influenced today.

Niebuhr unashamedly but with sophistication talked about how 'revelation means for us that part of our inner history which illuminates the rest of it and which is self intelligible,' over against 'external history.' He made no claims that 'truth' vs. 'falsehood' were here at stake so much as matters of perspective that illumine the scholars' subjects in contending ways. He saw his own earlier book as too strongly committed to 'external history' through social sciences and the latter as a corrective of it. Today both are seen as primary sources of historic approaches to religious studies.

Similarly, in 1955 Jewish theologian and sociologist Will Herberg (1955) wrote a determinative study of American religion during the Eisenhower religious revival, calling it *Protestant*, *Catholic*, *Jew*. It was a provocative analysis colored by all sorts of normative judgments based in the Hebrew prophetic tradition and modern existentialist interpretations. Works such as Niebuhr's and Herberg's are dated today, and exemplary of the work of few at the century's end. Yet they illustrate one end of a spectrum of options at the opposite pole from the social scientific reductive versions.

At the twentieth century's end, while historians in great numbers pursued a wide range of historical expressions of religion, literary studies came into their own in the American Academy [of Religion] and the American academy in general. This has meant a study of American classics—Melville, Hawthorne, Whitman, Adams, and the like; modern poets and novelists—Frost, Stevens, Eliot, and more. They had been long researched in American Studies for what they might disclose about American life.

In the subsequent deconstructionist and post-modern episodes there is much less belief that disclosure of

metanarratives or larger meanings is possible. By century's end the returns were not in on these debates and the scholarly commitments that led up to them or issued from them.

8. *The Special Issue of Pluralism*

It may create a false impression to speak in generic terms of 'American religion.' From some points of view all that religions and religious voices have in common is their issuance from American as a place. Gone are the days when a Perry Miller or Yale historian Sydney Ahlstorm could credibly and in sophisticated ways make the claim that American intellectual and cultural life was New England Puritanism writ large, or that H. Richard Niebuhr could talk of the Kingdom of God in America in Protestant terms as being disclosive of meaning in the larger whole.

Instead, historians and literary scholars tended to become more eclectic, to deal with national life in more piecemeal terms. Collections of American Studies essays are likely to focus on what the subheads suggest in an essay by historian Catherine Albanese—who operates at a pole opposite from that which Philip Schaff or Sydney Ahlstrom suggests: 'Indian Episodes,' 'Catholic Encounters,' 'Protestant Relations,' 'African American Ports of Call, Jewish Alliances, and Asian Junctures,' 'Contacts and Combinations.'

Whether this eclectic, post-modern, anti-metanarrative approach is or is not contributing to the formation of an altered canon with privileged texts seen as disclosive of larger stories is an issue that will animate and agitate American Studies: Religion for years to come.

See also: American Studies: Politics; American Studies: Society; Pluralism and Toleration; Religion and Politics: United States; Religion: Definition and Explanation; Religion, Sociology of; Religiosity, Sociology of; Secularization

Bibliography

Abrams M H 1971 *Natural Supernaturalism: Tradition and Revolution in Romantic Literature*. Norton, New York

Ahlstorm S E 1972 *A Religious History of the American People*. Yale University Press, New Haven, CT

Bowden H W 1971 *Church History in the Age of Science: Historiographical Patterns in the United States, 1876–1918*. University of North Carolina Press, Chapel Hill, NC

Brumm U 1970 *American Thought and Religious Typology*. Rutgers University Press, New Brunswick, NY

Callcott G H 1970 *History in the United States, 1800–1860; Its Practice and Purpose*. Johns Hopkins University Press, Baltimore, MA

Connolly P (ed.) 1999 *Approaches to the Study of Religion*. Cassell, New York

Conser W H Jr., Twiss S B (eds.) 1997 *Religious Diversity and American Religious History: Studies in Traditions and Cultures*. University of Georgia Press, Athens, GA

Geertz C 1973 *The Interpretation of Cultures; Selected Essays*. Basic Books, New York

Gunn, G B 1979 *The Interpretation of Otherness: Literature, Religion, and the American Imagination*. Oxford University Press, New York

Hackett D G 1995 (ed.) *Religion and American Culture: A Reader*. Routledge, New York

Herberg W 1955 *Protestant, Catholic, Jew; An Essay in American Religious Sociology*. Doubleday, New York

McCutcheon R T (ed.) 1999 *The Insider/Outsider Problem in the Study of Religion: A Reader*. Cassell, New York

McDannell C 1995 *Material Christianity: Religion and Popular Culture in America*. Yale University Press, New Haven, CT

Miller P 1939, 1953 *The New England Mind 2 vols*. Beacon, Boston

Miller P 1956 *Errand Into the Wilderness*. Belknap Press of Harvard University Press, Cambridge, MA

Moseley J G 1981 *A Cultural History of Religion in America*. Greenwood Press, Westport, CT

Murdock K S 1949 *Literature and Theology in New England*. Harvard University Press, Cambridge, MA

Niebuhr H R 1937 *The Kingdom of God in America*. Willett, Clark & Company, NY

Pals D L 1996 *Seven Theories of Religion*. Oxford University Press, NY

Ramsey P (ed.) 1965 *Religion*. [Essays by] Philip H. Ashby [and others]. Prentice-Hall, Englewood Cliffs, NJ

Scott N A Jr. 1966 *The Broken Center: Studies in the Theological Horizon of Modern Literature*. Yale University Press, New Haven, CT

Shepard R S 1991 *God's People in the Ivory Tower: Religion in the Early American University*. Carlson Pub., Brooklyn, NY

Skotheim R A 1966 *American Intellectual Histories and Historians*. Princeton University Press, Princeton, NJ

Stone J R (ed.) 1998 *The Craft of Religious Studies*. St. Martin's Press, NY

Tweed T A (ed.) 1997 *Retelling US Religious History*. University of California Press, Berkeley, CA

M. E. Marty

American Studies: Society

When the academic discipline of 'American Studies' was developed in the course of the 1930s, it was explicitly set up as an area of studies where scholars and students of, notably, history and literature could meet and join forces in an integrated, interdisciplinary context in order to redress what was felt to be an increasingly acute challenge: how to bridge the widening gap between the literary and historical approaches to the study of American society. However, the emergence of this new, holistic approach to American society immediately created its own dialectic: between those seeing American society as a whole, and those emphasizing its plurality; between those wanting to

highlight American society's unifying myths and symbols as well as its common values and aims, and those drawing attention to its centripetal forces, sometimes identifying these as a dynamics of cultural diversity, sometimes as a tendency toward social segmentation and discontinuity. Even at the beginning of the twenty-first century—as we emerge from the 'culture wars' of the late 1980s and 1990s—it remains difficult to establish any kind of consensus among analysts about the common denominators in American society, about its inner core or even its outer boundaries.

Nevertheless, a number of recurrent themes can be identified that appear to dominate past and present debates in American studies about the nature of American society. Following that society through history, distinguishing the lasting from the transient, the center from the periphery, was one of the original drives behind the emergence of American studies, and this strategy continues to inspire and energize the discipline. Although it may never have become the holistic approach originally envisaged, few in American studies today would deny that trying to understand and describe American society is in effect an attempt to fathom what sustains that society, what forces operate it and direct its purposes. Hence, even though, as Wiebe has argued in *The Segmented Society* (1975), surveying the whole of a society always tends to emphasize social patterns rather than social processes and to subordinate 'a history of themes inside American society to a history of the society incorporating them,' one may still look meaningfully at those themes, and at the history of their emergence and development. That is what American studies, from a wide variety of disciplinary angles, continues to do.

1. Terms and Periods

Apart from qualifying the unanimity among scholars in American studies regarding the object and methodology of their inquiry, a further note of caution should be sounded concerning terminology. Within the discipline of American studies—itself an umbrella term—the term 'society' is usually taken to refer to the aggregate of the politics, history, economics, education, geography, religion, environment, and culture of the United States, as well as to the relations and tensions between these elements. However, many Americanists, especially those approaching their object from the realms of literary analysis, the history of ideas, and 'cultural studies,' would refer to the same grouping of aspects of American society and the interplay between them as American 'culture.' As Potter remarked in *History and American Society* (1973), this terminological confusion is symptomatic of 'a serious general problem in the study of societies,' viz. the lack of fixed points of similarity in different societies, and between different aspects of societies, with which to measure continuity and separateness. For the sake of clarity (and following Wiebe), 'culture' will here be defined in its more narrow sense as 'those values and habits conditioning everyday choices in such areas as family governance, work, religious belief, friendship, and casual interchange,' while 'society' will be seen as combining 'these patterns with broader and more systematic realms of behavior, such as the organization of a community's life, the structure of a profession or business or religious denomination, and the formula for apportioning economic rewards.'

Traditionally, the evolution of American society during the first two hundred years since the Revolution has been divided into three phases, each associated with a distinctive social system: from the Revolution to the 1790s; from 1830 to 1890; and from 1920 to the 1970s. The intervening years—from the 1800s to the 1830s, and from the 1890s to the 1920s—are seen as crucial periods of transition from one social system to the next: from 'eighteenth-' to 'nineteenth-,' to 'twentieth-century society.' Whether the last three decades of the twentieth century constitute a fourth phase in American society (cf. for instance Lipset 1979), or a continuation of the transitional period of the 1960s and 70s, or, indeed, the 'disuniting of America' altogether as signaled by others (cf. Schlesinger 1991), is still a moot point. In the following, aspects of American society that have dominated debates and explorations in American studies will be discussed in terms of their emergence and evolutionary history.

2. Seventeenth- and Eighteenth-century Society

American society, one often reads in accounts of the settlement and early colonization of North America, began with a dream—a vision of a morally pure, socially just New Eden, itself the reflection of a long tradition of European utopianism. European observers and settlers alike represented seventeenth- and eighteenth-century American society in terms of a transatlantic opposition that is still relevant today: an opposition between Old World corruption and New World innocence, between darkness and light, between stasis and progress. For even if the early American settlers were a motley crowd internally divided by differences of ethnicity, language, ideology, religion, politics, and commercial interest, there was always a remarkable consensus among them about what America was, or was supposed to be—a society that, defining itself against the backdrop of European wars, exploitation, inequality, and persecution, was somehow unique, whole, progressive, and civilized.

More remarkable than that, perhaps, was that, despite their differences, the various stakeholders in the new nation were all prepared, if for very disparate reasons, to rally round a symbol of American society

first provided by one radical group of early settlers, the New England Puritans. As early as 1622, Puritan leader John Winthrop labeled England as 'this sinfull land,' plagued by poverty, inequitable taxation, a bureaucratic legal system, and religious intolerance; as the first elected governor of the Massachusetts Bay Company, Winthrop set off for New England in 1630, envisioning a 'city upon a Hill' as the utopian foundation for the new society that he and his fellow Puritans would be building. The New England Puritans may have had fairly specific religious reasons for regarding their community as a model for other colonists and as the first step toward establishing a kingdom of God in the New World that would lead the world into a new millennium, but the image of America as 'a beacon upon a hill' sufficiently reflected the ambitions of other, non-Puritan settlers for it to be adopted widely as one of the most resonant and sustaining symbols of American society, both in qualitative, intrinsic terms, and in its relation to the rest of the world. Thus, long after the waning of the Puritan theocracy, the founding fathers of the 1630s continued to be ritually invoked as heroes in the cause of liberty—not just in New England, but throughout other Eastern States and the South, and as much during the Great Revival as during the French and Indian War and the Revolution. The epic of the Puritan exodus became the common property of all Anglo-American settlers, and the rhetoric and ideology of the Puritan religious errand into that 'unredeemed, unchristianized, lawless region' (as Hawthorne put it) became the rhetoric and ideology of America's cultural errand into the 'wilderness.'

As a symbol reflecting the internal dynamics of the group, the idea of America as a bridgehead of civilization beleaguered by a 'vast and desolate wilderness' (Rowlandson 1994) to an important degree determined the social mechanics of what has been termed 'the ritual of consensus' (Bercovitch 1993). Faced as they were by the challenges of survival and success, consensus filled the need felt by Puritan and non-Puritan settlers alike for a certain social order. In particular, such a ritual of consensus had to regulate the rights and responsibilities of the individual versus the group. According to Bercovitch, there were three basic tenets of Puritan consensus, which in the course of the seventeenth century transformed from a tribal ritual into a national ritual of cultural and social origins. The first tenet was migration, as a function of divine mission or prophecy—which contributed significantly toward rationalizing the expansionist and acquisitive aspects of settlement (the Puritans being as much interested in material gain as in salvation). The second tenet of consensus was related to discipline: seeing that an individual's success made visible the meaning of the errand of the group, the challenge was how to endorse individualism without promoting anarchy. The third tenet of consensus was concerned with progress: constantly affirming that it was en route from a sacred past to a sacred future, the Puritan community established institutions that were geared more toward sustaining progress and growth than toward maintaining stability. Crucially, the Puritan ritual of consensus was constantly enacted—both in the sense of established in law and in the sense of publicly performed—in a series of interlocking covenants to which individuals were invited to subscribe their names if they wanted to be regarded as members of the community. The first and most famous of such civil covenants, the 'Mayflower Compact' (which was signed in 1620 by the Pilgrim Fathers on board the ship that took them to what was to become Plymouth Plantation), reflects how these covenants were aimed at establishing a 'Civil Body Politic' by preserving individual freedom while demanding submission to 'the general good.' It is this Puritan rhetoric of voluntary compacts and consensus, of community grounded in individuals committing themselves freely to their 'common providence,' that gave American society its unique slant on the universal process of introducing systematic relationships among individuals, groups, and institutions; and it is the same rhetoric that, being both universal yet culturally specific, facilitated the transition from the Puritan to the Yankee, and from errand to manifest destiny and the American Dream. It is the legacy of the Puritan errand that lies at the basis of America's social awareness vis-à-vis the world at large: America as the 'redeemer nation,' which, according to Tuveson (1968), can be reduced to the following key elements: 'chosen race; chosen nation; millennial-utopian destiny; fighting God's war between good (progress) and evil (regression), in which the United States is to play a starring role as world redeemer.' Thus, when Woodrow Wilson said that 'America had the infinite privilege of fulfilling her destiny and saving the world,' he was not saying anything startlingly new.

In the course of the second half of the eighteenth century, the spiritual and the more secular beginnings of American society (symbolized by Plymouth Plantation and Jamestown, respectively, by the 'beacon on the hill' and the Horn of Plenty) had become sufficiently blended for Crèvecoeur to be able to pose the notorious question in his *Letters from an American Farmer* (1782), 'What, then, is the American, this new man?' and to assume—or suggest—that this was actually more than a rhetorical question: that there really was such a social subject as 'an American.' What in hindsight is perhaps even more perplexing than Crèvecoeur's question as such, is his answer to that question: Crèvecoeur's 'American' and the society he is said to have created are still by and large what we would recognize as American national identity and society today. An immigrant newly arrived in America, Crèvecoeur informs us, would find himself 'on a new continent; a *modern society* offers itself to his contemplation, different from what he had hitherto seen.' He would see a society that is egalitarian and

classless, in which individuals are 'animated with the spirit of an industry which is unfettered and unrestrained, because each person works for himself.' The world's poor, oppressed and persecuted will find safety and plenty of opportunity in 'this great American asylum.' To underline the radical modernity and uniqueness of this new society, Crèvecoeur assures us that in America 'individuals of all nations are melted into a new race of men, whose labours and posterity will one day cause great changes in the world.' The millennial dreams of the New England Puritans have in Crèvecoeur's *Letters* become America's millennial dreams—now all Americans are 'pilgrims': 'Americans are the Western pilgrims who are carrying along with them that great mass of arts, sciences, vigour and industry which began long since in the East; they will finish the great cycle.'

3. *Nineteenth-century Society*

The individual's 'self-interest' and his concomitant voluntary submission to his nation's mission form the nucleus of Crèvecoeur's America—and they still do. The American Revolution, which ended the colonial stage in the evolution of American society, did not therefore constitute that sharp, radical break it has traditionally been made out to be. Since the end of the French and Indian War, self-interest had been synonymous with (Protestant) patriotism, and it was through their appeal to this patriotic self-interest that the Revolutionary leaders managed to steer the colonies victoriously through the conflict with Britain and toward nationhood. Thus, while Crèvecoeur's distressed frontiersman announces in his final letter (burdened by his creator's nostalgia for utopian pastoralism) that in 'these calamitous times' he feels forced to flee his farm and to resettle his family 'under the wigwam' in the western wilderness, the Whig leaders succeeded in bringing the violence and disruption of the Revolution under control by turning it into a redemptive, controlling ritual of national identity. The push for independence, which elsewhere (most notoriously in subsequent years in France) often triggered a complete collapse of the social fabric as the forces of change spun out of control, in America engendered a spirit of intense nationalism and enforced the model of consensus. What the Puritan Fathers had begun—overcoming discord in society by turning anxiety into a vehicle of social control—was now finished by the Founding Fathers.

With the 'family quarrel'—so unfortunate but necessary—finally over, America was ready to enter the final phase of its evolution toward full nationhood; the Revolution having given America independence from the Old World, it was now time to complete the errand into the wilderness by going west, not to hide under a wigwam, but to conquer the continent. Space and diversity, as Madison had argued in *Federalist Number Ten*, would be the best guarantors of liberty, because a dispersed and heterogeneous nation would be able to contain more factiousness than a compact and more homogenous one. And the dangers of faction were increasing all the time in the wake of the Revolution, as the system of graded and interlinking social relationships of eighteenth-century society—marked by paternalism, the moral economy, and virtuous republicanism (or 'civic virtue')—began to give way to an increasing emphasis on personal mobility—social mobility as well as geographical.

Uncurbed social ambition and a vigorous spirit of enterprise led to a rapid crumbling of deferential restraints, and the meaning of American wholeness came under significant pressure during these years of transition. The old household economy of colonial America had gone into a complete nosedive in the 1790s; fuelled by the stirrings of industrialism, the opening up of the west, the Louisiana Purchase of 1803, and the Napoleonic Wars in Europe (which virtually eliminated mercantile competition), liberal capitalism triggered a socioeconomic revolution that transformed the United States in the course of two decades into a market economy and market society. The end of the War of 1812 gave a further boost to America's economic revolution, with a rapidly growing immigrant population, and the building of the canals and the early railways permitting the republic to enter an era of unprecedented expansion, especially toward the vast western regions.

American geopolitics in the early decades of the nineteenth-century to a large extent shaped the growing awareness of nationhood and state. This resurgence of nationalism peaked with the so-called 'Monroe Doctrine,' which basically meant that the United States declared they would not accept any future colonization by European powers of 'the American continents.' But in social terms a price had to be paid for America achieving hegemony in the Western Hemisphere, and a controversy erupted over the question of not whether, but how the nation should be allowed to expand. With the election of Jackson as President of the United States in 1829 a powerful democratic movement swept the country, paving the way for mass politics and what was called at the time 'the age of the common man.' The fight against social and economical inequality and privilege dominated the national political agenda. And although there is still considerable disagreement among historians as to how much of this drive toward democratic reform and anti-elitism can actually be attributed to Jackson and his followers, the reaction against laissez-faire capitalism from among the working classes and other exploited groups in the new industrial society was considerable. Liberty and equality became the cornerstones of the new American society—even though the veneer of egalitarianism did not affect a host of distinctions underneath (notably those of race and gender).

A tourist like the French aristocrat Alexis de Tocqueville, who came to America in the 1830s to discover why the efforts at establishing democracy in France, starting with the French Revolution, had failed while the American Revolution had produced a stable democratic republic, may not have seen far beyond the surface respectability but he did make some seminal comments on American egalitarianism. According to Tocqueville, America was 'the first new nation' in the world. In fact, it was Tocqueville who first referred to the United States as 'exceptional,' that is, as a nation qualitatively different from other nations. In his extraordinarily influential book *Democracy in America* (1835–9), Tocqueville identified five elements in the American Creed as it had emerged from the revolutionary ideology—liberty, egalitarianism, individualism, populism, and laissez-faire—but, as Tocqueville pointed out, egalitarianism in American involves an equality of opportunity and respect, not of result or condition. Therefore there is no emphasis on social hierarchy in America—no monarchies or aristocracies (which distinguishes the United States from post-feudal nations as such)—although there is social segmentation. It is this, Tocqueville noticed, and many others since him, that was central to American exceptionalism: 'What held Americans together was their ability to live apart .... From this elementary principle emerged a pattern of beliefs and behavior that was recognizably American' (Wiebe 1975).

The idea of space as a sine qua non of liberty and equality was popularized at the end of the nineteenth century by the historian Frederick Jackson Turner. At a point in America's history when, he claimed, the vast, unsettled lands were gone, Turner identified the westward expansion of the United States as the crucial formative force behind American individualism, nationalism, and democracy. Because he had to constantly reinvent himself and reconfirm his essential Americanness as he pushed further west, the frontiersman, Turner argued, had defined American civilization. With the waning of the western wilderness, a crucial period in American history had closed. In fact, Turner's concerns turned out to be rather premature as in the decades following his bemoaning of the 'passing of the frontier' the US government gave away more land from the public domain than it had done before—but the 'frontier myth' kept a romantic image of the West alive for long after.

4. *Twentieth-century Society*

If, as historian Richard Hofstadter 1948 has observed, it has been America's 'fate as a nation not to have ideologies, but to be one,' then that ideology was very much in place at the beginning of the twentieth century. But although the ideology did not change materially from that of the late nineteenth century, the face of society did. One of the most conspicuous developments was the emergence of a new social elite—the occupational elite. The end of the nineteenth century saw an explosion of the number of Americans in administrative and professional jobs. Industry and the new technologies needed managers, technicians, and accountants; cities needed professionals in the medical, commercial, legal, and educational services. By the turn of the century, these new professionals had come to constitute a distinct social group—the new middle class.

Although the new professional class enabled individuals to climb the social ladder more rapidly than in the nineteenth century, the new class was by no means any more homogeneous—the nineteenth-century segmentation of society on the basis of region, ethnic background, and access to capital was replaced by a segmentation according to professional occupation—each segment (clerks, technicians, lawyers, teachers, engineers) having its distinct identity and social awareness, and the vitality of each depending on the vitality of the whole.

While diversity remained the common denominator in society, the need for cohesion in the national society never disappeared—the ideology of consensus still being as strong as in the days of the Puritans. If in the nineteenth century the outrage of national conscience had directed itself against sloth, drunkenness, and profligacy, in the twentieth century the social hegemony aimed its arrows at the leveling left, communists, and other 'un-American' elements (descendants of Crèvecoeur's 'off-casts') that threatened to undermine the 'volutary order' (Berthoff 1971).

But the most powerful cement of the new social system was consumption. Starting hesitantly in the latter decades of the nineteenth century, accelerating in the 1920s, and exploding in the 1950s, consumption became the foundation for the modern America. By the 1960s, consumption was no longer a privilege but a right—to some extent even a civil duty. Consumption became everybody's stake in society—the ultimate equalizer. In the course of the twentieth century, Americans began to regard the meaning of work as the power to spend, and increasingly began to discover their identity and social position in their power to consume. Mass consumption became an inalienable right. Women derived their freedom from being primary consumers; consumption was an effective weapon against communism; and the urban African-Americans who looted shops during race riots thereby underlined the fact that they were still being denied their basic civic right: the right to consume.

Consumption has been regarded as the ideological twentieth-century answer to the challenges of differences of class and civil rights. The nineteenth century had failed to manage the problems of class; the economic elite of the twentieth century quite early on accepted class as a social fact. By quickly and consistently demolishing labor organizations and any

manifestations of socialism wherever they appeared, and simultaneously offering occupational segmentation as a new model of social mobility, the threat of class conflict was contained. The importance of regular employment was paramount to egalitarianism for the credit it generated: for it was the credit that made the professional or worker a consumer, and the consumer a citizen. In a similar way, twentieth-century society put paid to the issue of race conflict: the economic elite offered members of minorities as well all other Americans a passport to consumerism and a place somewhere in the great interdependency of the system of segmented wholeness. Race, like class, simply was no longer relevant from a socioeconomic point of view. Racial or other minorities can either aspire to assimilate to the segmented system, even to belong to the social elite, or they can decide to rely on such benefits as are provided within the context of social security.

The much-touted revolutionary drive of the 1960s' 'counterculture' did not materially change the system of segmented wholeness. For a while it looked as if there might be more emphasis on the common denominators in American society, as if there were an America beyond the segmented system (classless, gender-neutral, race-neutral). Indeed, changes in the overarching social fabric have occasionally been affected by pressure from within one or more of the segments (such as racial desegregation, and a more equitable job market). But, by and large, America has remained essentially a nation of segments. However, if America still subscribes to this concept of nationality, it does so voluntarily. In contrast to the rest of the world, as Gorer, 1948 has observed, Americans believe that nationality is 'an act of will, rather than of chance of destiny.'

That is what Europeans have never understood about America, according to Baudrillard (1986). He argues that the great lesson of the success of America's social formation is that 'freedom and equality, like ease and grace, only exist where they are present from the outset. This is the surprise democracy ha[s] in store for us: equality is at the beginning, not the end. That is the difference between egalitarianism and democracy: democracy presupposes equality at the outset, egalitarianism presupposes it at the end. "Democracy demands that all of its citizens *begin* even. Egalitarianism insists that they all *finish* even."' Whereas Europe has remained stuck in the old rut of social difference and is constantly dragged back into the history of its bourgeois culture, America, by contrast, has achieved a state of radical modernity, which in temporal terms can be described as a perpetual present. Having internalized democracy and technological progress, America ducks the question of originality, descent, and mythical authenticity. Everything is exactly what it appears to be in America: the real and the imaginary have been collapsed into the 'hyperreal,'—a constant state of simulation, of signs having escaped their referentials in the real. The concept of history as 'the transcending of a social and political rationality, as a dialectical, conflictual vision of societies,' is alien to Americans, Baudrillard observes: the United States is, in fact, a utopia achieved—a society which has behaved from the beginning as though it were already achieved.

5. *American Society and American Studies*

Baudrillard's incisive observations about American society are not unchallenged within the social sciences, but despite the provocative rhetoric, it cannot be denied that his concepts of the 'hyperreal' and the 'utopia achieved' are valuable tools in trying to account for some of the more paradoxical complexities of American society—its fragmented wholeness; its social stability in the face of an unequal distribution of wealth; its belief in 'a pleasing uniformity of descent' (Crèvecoeur 1782) and a 'general happy Mediocrity' (Franklin 1986) when the racial divide appears to be as wide as ever. Baudrillard's agenda—to bring the paradoxical nature of American society into a universal model while retaining the apparent incompatibility of its constituent parts—is certainly unique in the evolution of American studies. From its inception in the 1920s, the discipline of American studies has almost inescapably been concerned with the search for the common denominator in American culture and society: the very success of its mission—to establish and institutionalize a scholarly discipline whose object of study was the society of the United States—depended on the new discipline being able to achieve some sort of consensus about the nature of that object of study.

This mission first took off seriously with the publication of Vernon Parrington's *Main Currents in American Thought* (1927), and was carried further with enthusiasm by the 'consensus historians' of the 1950s (including Richard Hofstadter, Louis Hartz, Arthur Schlesinger, Jr., Daniel Boorstin) and, most importantly (in terms of its impact on the field), by the 'myth and symbol' school of the 1950s (which included intellectual historians and literary critics such as Lionel Trilling, Henry Nash Smith, R. W. B. Lewis, and Leo Marx). The 'myth and symbol' scholars had indirectly been influenced by a host of literary critics in the 1930s and 1940s who were known collectively as the 'New Critics.' Approaching the literary text as an autonomous whole, the New Critics (which included such people as John Crowe Ranson, Kenneth Burke, Yvor Winters, and R. P. Blackmur) emphasized the transcendental qualities of the text: its moral content; its symbolic meaning as a reflection of national culture and identity; and its place and contribution to a national literary 'tradition.' F. O. Matthiessen's extraordinarily influential *American Renaissance: Art and Expression in the Age of Emerson and Whitman* (1941) established a 'great tradition' of nineteenth-century

American authors that represented the 'quintessential' characteristics, concerns, and values of 'mainstream' American society. Inspired by the humanistic ideal of high culture, Matthiessen's canon of nineteenth-century American literature not only introduced a firm hierarchy of 'high' over 'low' artistic expression but also provided the raw cultural materials for the authors of the myth and symbol school. In this way Smith's thesis of the wilderness as the basic ingredient of American civilization, Lewis's insistence on tragic innocence in the myth of the 'American Adam,' and Marx's symbol of the 'machine in the garden' to account for the 'American' dichotomy of pastoralism and technological progress could become the unifying myths and symbols that dominated the study of American society during the 1950s and 1960s.

When, in the course of the 1960s, mainstream American society and culture increasingly came under attack from historically marginalized groups as part of the struggle for emancipation and recognition (women, African-Americans, Native Americans, Hispanics, gays, and other minorities), the theory and methods of American studies were also subjected to critical analysis and reappraisal. If pluralism replaced universalism as the new norm in society, so it did in American studies. Thus the new social history that emerged in the late 1960s and early 1970s began to demolish American exceptionalism, imperialism, and the consensus society, while it tried to recover or create the history of America's 'forgotten' minorities, as well as America's popular culture, and its labor and regional history. Drawing from the techniques and methods of sociology, anthropology, and other human sciences, gender, class, and race became the new 'holy trinity' of American studies.

Increasingly, also new criticism and the myth and symbol school came under attack for being ahistorical and for their philosophical idealism. With the emergence of the 'New Historicism' (ushered in by Stephen Greenblatt and the journal *Representations*) interdisciplinarity and historical contextualization were put firmly on the methodological agenda of American studies. This paradigm shift is illustrated by a seminal collection of essays edited by Sacvan Bercovitch and Myra Jehlen, *Ideology and Classic American Literature* (1988), which contains reassessments of their work by important pioneers of American Studies theory and method including Henry Nash Smith, Leo Marx, and Alan Trachtenberg, as well as essays by a younger generation of scholars, including Houston Baker, Carolyn Porter, Donald Pease, Michael Gilmore, Jane Tompkins, Jonathan Arac and Myra Jehlen—work that exemplified the scholarly output of the 1980s, blending feminist, neo-Marxist, and poststructuralist theories and methods.

However, whereas among the newly emancipated groups this pluralist approach to American society was experienced as liberating (both intellectually and politically), a growing number of scholars began to interpret the new trend toward plurality as undesirable. Christopher Lasch was one of the first to sound the alarm bell. In *The Culture of Narcissism* (1979) he argued that the 'devaluation of the past' was symptomatic of a deep 'cultural crisis,' and he accused the 'demoralization of the humanities' of having reduced American 'individualism' to 'a narcissistic preoccupation with the self.' With the publication of Allan Bloom's *The Closing of the American Mind* (1987) the 'battle of the books' erupted into an all-out 'culture war.' According to Bloom, the social/political crisis of twentieth-century America was really an intellectual crisis. Despite the use of such edifying labels as 'individual responsibility, experience, growth, development, self-expression, liberation and concern,' Bloom argued, the 1960s had 'bankrupted' American universities and destroyed 'the grand American liberal traditions of education.' The battle over the canon of American literature was now openly a battle about the canon of American culture and society, about nationality and multiculturalism, about consensus and political correctness. The 1990s saw a whole spate of books calling for a truce and reconciliation, offering a wide array of solutions to the increasingly vicious national debate and the resulting ideological stalemate. In *Beyond the Culture Wars* (1992) Gerald Graff showed how 'teaching the conflicts can revitalize American education,' while Henry Louis Gates, Jr. advocated in *Loose Canons: Notes on the Culture Wars* (1992) a new civic culture that would transcend the sharp divisions of nationalism, racism, and sexism. Schlesinger rather desperately proposed in *The Disuniting of America* (1991) that Americans should attempt 'to vindicate cherished cultures and traditions without breaking the bonds of cohesion—common ideals, common political institutions, common language, common culture, common fate—that hold the republic together'; this appeal was picked up by David Hollinger, who argued in his *Postethnic America* (1995) that Americans should form 'voluntary' rather than 'prescribed affiliations' to bridge the ethnic and multicultural gaps in society.

There are no clear signs that the latest models of reconciliation have effected any kind of movement on the battleground of nationality and multiculturalism—which perhaps is not surprising since the majority of them propose as the solution to the threat of the radical implosion of the consensus model of American society a return to the very neo-humanist, essentialist ideals that are so vehemently being contested in the multiculturalism/pluralism debate. In the meantime, the discipline of American studies is as divided as the various participants in this debate. However, although there has been some degree of disciplinary backlash (with some academic programs having been reduced or cut from university calendars), most working in the field would concede that the questions about the future of American culture and society have never been as open as they are at the

beginning of the twenty-first century, and that this in itself more than legitimizes the continuing study of American society within the interdisciplinary framework of American studies.

See also: American Studies: Politics; American Studies: Culture; Civil Society, Concept and History of; Egalitarianism: Political; Egalitarianism, Sociology of; Frontiers in History; Multiculturalism; Multiculturalism and Identity Politics: Cultural Concerns; Multiculturalism: Sociological Aspects; Pluralism; Tocqueville, Alexis de (1805–59)

Bibliography

Baudrillard J 1986 [1988] *America*. Verso, London

Bercovitch S 1993 *The Rites of Assent: Transformations in the Symbolic Construction of America*. Routledge, New York

Bercovitch S, Jehlen M 1988 *Ideology and Classic American Literature*. Cambridge University Press, Cambridge, UK

Berthoff R 1971 *An Unsettled People: Social Order and Disorder in American History*. Harper & Row, New York

Bloom A 1987 [1988] *The Closing of the American Mind: How Higher Education Has Failed Democracy and Impoverished the Soul of Today's Students*. Simon & Schuster, New York

Crèvecoeur J H St J de 1782 [1981] *Letters from an American Farmer*. Penguin, New York

Franklin B 1791–1798 [1986] *Autobiography*. Norton, New York

Gates H L Jr 1992 *Loose Canons: Notes on the Culture Wars*. Oxford University Press, New York

Graff G 1992 [1993] *Beyond the Culture Wars: How Teaching the Conflicts Can Revitalize American Education*. Norton, New York

Gover G E 1948 *The American People: A Study in National Character*. Norton, New York

Hofstadter R 1948 *The American Political Tradition and the Men Who Made It*. Knopf, New York

Hollinger D A 1995 *Postethnic America: Beyond Multiculturalism*. Basic Books, New York

Lasch C 1979 [1991] *The Culture of Narcissism: American Life in an Age of Diminishing Expectations*. Norton, New York

Lipset S M 1963 *The First New Nation: The United States in Historical and Comparative Perspective*. Basic Books, New York

Lipset S M 1979 [1980] *The Third Century: America as a Post-industrial Society*. University of Chicago Press, Chicago

Lipset S M 1996 *American Exceptionalism: A Double-edged Sword*. Norton, New York

Matthiessen F O 1941 *American Renaissance: Art and Expression in the Age of Emerson and Whitman*. Oxford University Press, New York

Parrington V 1927 [1958] *Main Currents in American Thought*. Harcourt & Brace, New York

Potter D M 1973 *History and American Society*. [ed. Fehrenbacher D E]. Oxford University Press, London

Rowlandson Mary 1682 [1994] *A True History of the Captivity and Restoration of Mrs. Mary Rowlandson*. Penguin, New York

Schlesinger A M Jr 1991 [1993] *The Disuniting of America: Reflections on a Multicultural Society*. Norton, New York

Tocqueville A de 1835–9 [1945] *Democracy in America*. Vintage, New York

Tuveson E L 1968 *Redeemer Nation: The Idea of America's Millennial Role*. University of Chicago Press, Chicago

Wiebe R H 1975 *The Segmented Society: An Introduction to the Meaning of America*. Oxford University Press, New York

W. M. Verhoeven

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Amnesia

1. *Introduction*

The term *amnesia*, as a description of a clinical disorder, refers to a loss of memory for personal experiences, public events, or information, despite otherwise normal cognitive function. The cause of amnesia can be either primarily *organic*, resulting from neurological conditions such as stroke, tumor, infection, anoxia, and degenerative diseases that affect brain structures implicated in memory; or it can be primarily *functional or psychogenic*, resulting from some traumatic psychological experience (see *Amnesia: Transient and Psychogenic*). This article will focus on organic amnesia. The following questions are addressed: What are the characteristics of amnesia? What structures are involved in forming memories (and whose damage causes amnesia) and what function does each serve in the process? Does amnesia affect recent and remote memories equally and, by implication, are memory structures involved only in memory formation and shortly thereafter, or are they also implicated in retention and retrieval over long intervals? Are all types of memory impaired in amnesia or is amnesia selective, affecting only some types of memory and not others? What implication does research on amnesia have on research and theory on normal memory?

2. *Characteristics of Organic Amnesia*

The typical symptoms of organic amnesia are the opposite of those of functional amnesia: old memories and the sense of self or identity are preserved but the ability to acquire new memories is severely impaired. Though capturing an essential truth about organic amnesia, this statement needs to be qualified in important ways in light of new research.

The scientific investigation of organic amnesia effectively began with Korsakoff's (1889) description of its symptoms at the turn of the century, during what Rozin (1976) called the 'Golden Age of Memory Research.' Likewise, it can be said that the modern era of neuropsychological research on memory and amnesia was ushered in by Scoville and Milner's (1957) publication of the effects of bilateral medial temporal

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Amnesia

1. Introduction

The term *amnesia*, as a description of a clinical disorder, refers to a loss of memory for personal experiences, public events, or information, despite otherwise normal cognitive function. The cause of amnesia can be either primarily *organic*, resulting from neurological conditions such as stroke, tumor, infection, anoxia, and degenerative diseases that affect brain structures implicated in memory; or it can be primarily *functional or psychogenic*, resulting from some traumatic psychological experience (see *Amnesia: Transient and Psychogenic*). This article will focus on organic amnesia. The following questions are addressed: What are the characteristics of amnesia? What structures are involved in forming memories (and whose damage causes amnesia) and what function does each serve in the process? Does amnesia affect recent and remote memories equally and, by implication, are memory structures involved only in memory formation and shortly thereafter, or are they also implicated in retention and retrieval over long intervals? Are all types of memory impaired in amnesia or is amnesia selective, affecting only some types of memory and not others? What implication does research on amnesia have on research and theory on normal memory?

2. Characteristics of Organic Amnesia

The typical symptoms of organic amnesia are the opposite of those of functional amnesia: old memories and the sense of self or identity are preserved but the ability to acquire new memories is severely impaired. Though capturing an essential truth about organic amnesia, this statement needs to be qualified in important ways in light of new research.

The scientific investigation of organic amnesia effectively began with Korsakoff's (1889) description of its symptoms at the turn of the century, during what Rozin (1976) called the 'Golden Age of Memory Research.' Likewise, it can be said that the modern era of neuropsychological research on memory and amnesia was ushered in by Scoville and Milner's (1957) publication of the effects of bilateral medial temporal lobectomy to control intractable epilepsy in a single patient, H.M. (see Fig. 1.

Some aspects of the disorder Korsakoff described are peculiar to a kind of amnesia that now bears his name (amnesia related to vitamin (thiamine) deficiency typically associated with alcoholism), while others are common to all forms of amnesia, including the one described by Scoville and Milner. The symptoms are best described by contrasting impaired abilities with preserved ones. They are as follows:

(a) Memory is impaired for various types of material, which is why the amnesia is often referred to as global, though as we shall see not all memories are affected equally. Perception, intelligence, and other cognitive functions are relatively preserved. Thus, amnesic people perform normally on standard tests of intelligence, but are impaired by comparison on standard tests of memory. They can play chess, solve crossword and jigsaw puzzles, comprehend complex instructions, and reason logically.

(b) The memory that is impaired is only *long-term* or *secondary memory*; that memory which lasts long after the information has been received and registered, and not *short-term or primary memory* which is used to hold information briefly in mind. Unaffected also is *working memory* (Baddeley 1986) which is used to operate on the information held in mind. As an example, amnesic people have a normal digit span which refers to the number of digits one can repeat immediately in sequence (typically seven, plus or minus two) and even a normal backward digit span in which the digits must be recalled in reverse order. They are impaired, however, at retaining and recalling even a sub-span list of words after a short interval that is filled with distracting activity, even if given ample opportunity to rehearse the material before. In a more ordinary life example, amnesic people can remember sentences well enough to respond to them, but cannot follow a conversation to the end if reference is made to utterances that occurred at the beginning. The same dissociation between long- and short-term memory holds for words, stories, complex visual patterns, faces, melodies, and some esthetic stimuli (Kolb and Whishaw 1996).

(c) Long-term memory loss in amnesia is most noticeable for events and information acquired after the onset of the disorder and into the future as well as in the period immediately preceding it, but not for information acquired long before that. That is, they have an *anterograde amnesia* that extends into the future but a *retrograde amnesia* limited to the time preceding the onset of the disorder. Thus, amnesic people have difficulty remembering what they learned or experienced since the onset of the disorder, even their current address and neighborhood if they had moved to a new place, but they can remember their old home and neighborhood, and some old experiences and events (Milner 1966, Squire and Alvarez 1995). Retention of such old memories, however, may be more selective and not as well-preserved as had once been believed (see below).

(d) Anterograde long-term memory loss in amnesia applies only to information that can be recollected consciously or *explicitly*. Acquisition, retention, and recovery of memory without awareness or *implicitly*

is normal. For example, having studied a set of words or pictures, amnesic patients perform poorly when their memory is tested explicitly with recognition (Which of the following items do you recognize as having been presented to you?) or recall (Tell me the items you had studied). Their memory for studied items is normal if they are tested implicitly by seeing how performance is altered by the study experience

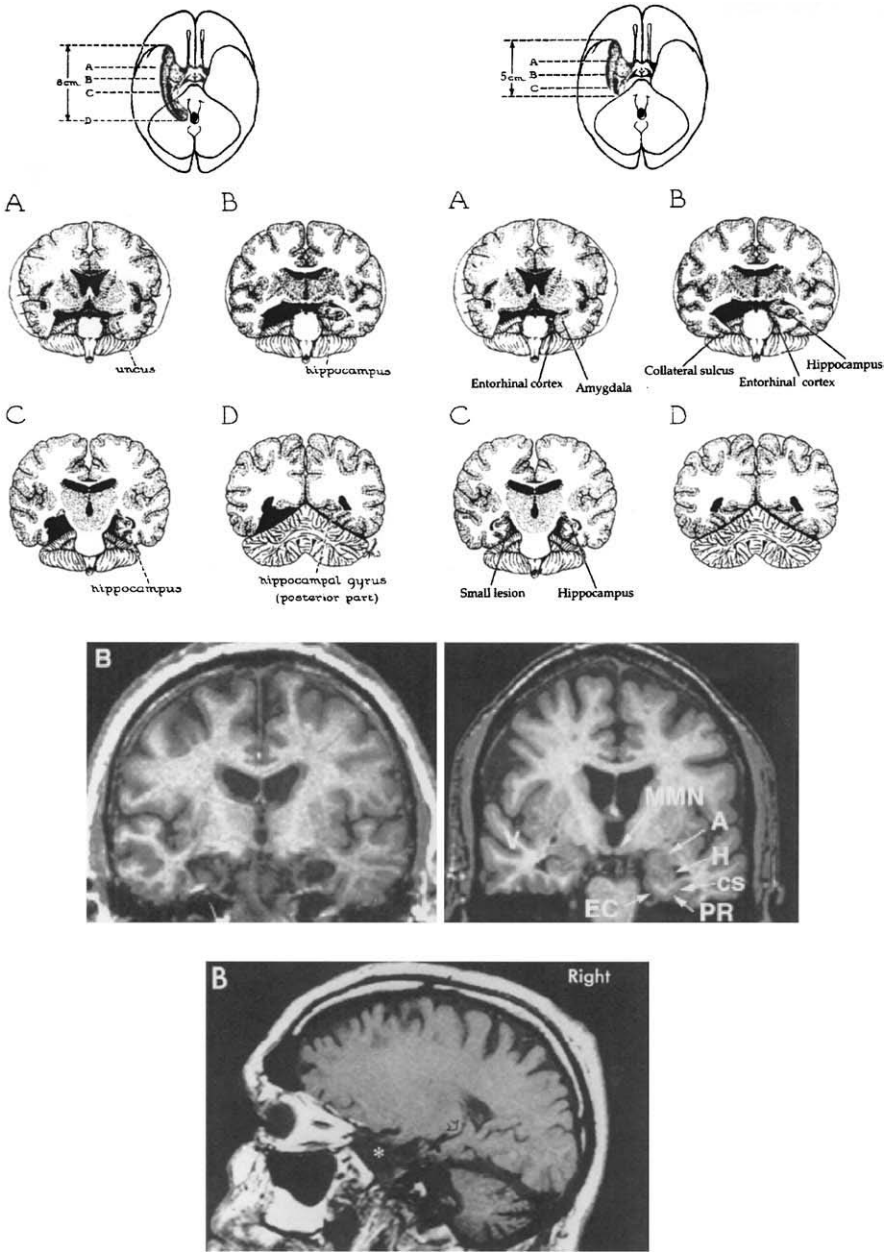

Figure 1
A recreation of H.M.'s lesion from the surgeon's report (left drawings A, B, C, D) and from a recent MRI scan (see bottom panels) of the lesion (right drawings, A, B, C, D). The surgeon overestimated the posterior extent of the lesion. The right side of each drawing is intact for comparison purposes. The two MRI scans labeled B are of H.M.'s brain, and the remaining one is of a control's brain. MMN = Mammilary bodies; A = Amygdala; H = Hippocampus; CS = collateral sulcus; PR = Perirhinal cortex; EC = Entorhinal cortex

without making any explicit reference to the study episode. Thus, though they cannot recall or recognize the items they studied, amnesic people will perceive them more quickly and accurately than items they did not study, or complete them better when they are degraded, such as by filling in the missing letters in a word or completing the lines in a picture. Performance on these implicit tests of memory indicates that information about the studied items is held in memory though they are not aware of that (Tulving and Schacter 1990, Schacter and Buckner 1998).

These characteristics will serve as the foundation for later discussion of empirical and theoretical investigations of amnesia, normal memory, and brain function. Indeed, research on amnesia, fascinating in its own right, has had a powerful impact on memory research and theory, especially since the landmark discovery of Scoville and Milner. This research can be divided into two interacting streams: a *functional neuroanatomical* one, concerned with identifying the neuroanatomical substrates of memory and determining their precise function; and a *(neuro)psychological* one, concerned with the implication that amnesia has for understanding normal memory at a functional level. Each is dealt with in turn.

3. Neuroanatomy of Amnesia

Amnesia is caused by bilateral damage to structures in the limbic system and to adjacent cortex in the medial temporal lobes (see Figs. 2 and 3). The hippocampal formation, in the medial temporal lobes, is the most prominent of the memory structures. It consists of the hippocampus proper with its various subfields and regions, plus the dentate gyrus and the subilculum. Communication between the hippocampus and neocortex occurs thorough a series of relays. The hippocampus is connected directly to the entorhinal cortex which in turn is connected to the parahippocampal gyrus and

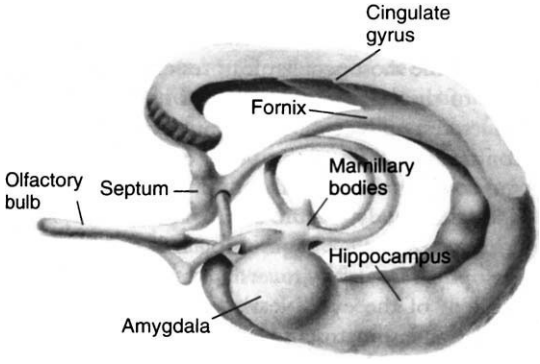

Figure 2
The limbic system and memory circuit (from Hamilton, 1976)

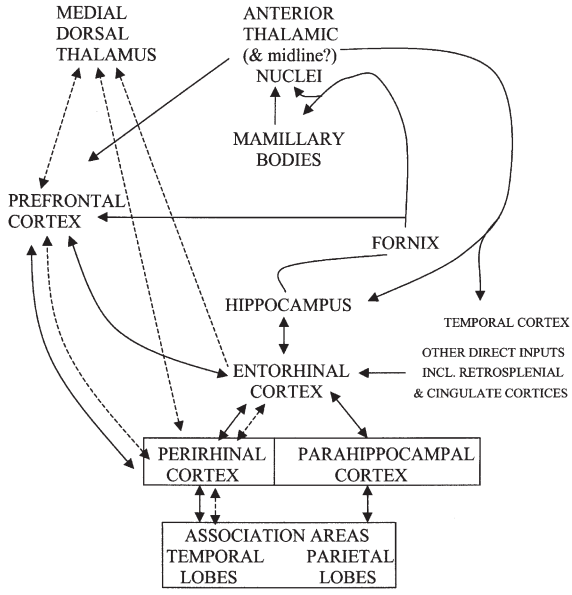

Figure 3
The hippocampal-diencephalic systems (modified from Aggleton and Brown 1999). There are two inter-related systems: The *hippocampal-fornix-anterior thalamic system* indicated by solid lines and the *perirhinal-medial dorsal thalamic system* indicated by dashed lines

perirhinal cortex which project bidirectionally primarily to the temporal and parietal lobes of neocortex, respectively. There are also projections from the hippocampus and perirhinal cortex via the fornix and anterior cingulate to the mammillary bodies and the dorsomedial nucleus of the thalamus, which are in the diencephalon. The loop of medial temporal and diencephalic structures constitute the limbic system. The hippocampus is thus ideally situated to collate information both about the cognitive (neocortex) and emotional (limbic) state of the organism and bind that information into a memory trace that codes for all aspects of a consciously experienced event (Moscovitch 1995).

3.1 Functions of the Neuroanatomical Substrates of Memory

There is considerable debate about the role that each of the components of the memory system have and how they interact with one another. Korsakoff's amnesia is associated with damage to the diencephalon whereas amnesia caused by anoxia, encephalitis, and some degenerative disorders such as Alzheimer's Disease is associated most often with medial temporal lobe damage. Although close inspection reveals differences in memory loss among the various conditions, some have not been substantiated reliably

and others could not be attributed with certainty to differences associated with diencephalic and medial temporal lobe damage, as compared to damage to other structures that often accompanies the various disorders. For example, it had been proposed that the medial temporal lobes were necessary for consolidation and retention of new information whereas the diencephalon was needed for encoding. Investigators claimed that people with diencephalic amnesia had difficulty acquiring information, but once they did so they retained it normally. People with medial temporal damage, however, showed abnormally rapid forgetting. Other investigators, however, found no difference in forgetting rates between amnesic groups, or even between them and controls (see review in Aggleton and Brown 1999).

Others noted that people with Korsakoff's, but not medial temporal, amnesia showed a number of memory deficits in addition to loss of memory for the content of an experienced event. These included confabulation, poor memory for the source and temporal order of events, susceptibility to interference, and poor meta-memory (knowledge about memory) (Moscovitch and Wincour 1992). These deficits, all of which involve strategic aspects of memory, were shown to be related more to frontal dysfunction that often accompanies Korsakoff's amnesia, rather than to diencepahlic damage *per se*.

Influenced by research on animal models of memory, investigators focused on differences among the structures in the medial temporal lobe itself, and its projections to the diencephalon. O'Keefe and Nadel (1978) proposed that the hippocampus is needed for representing allocentric spatial information, or cognitive maps but not for representing egocentric representations, routes and landmarks. While acknowledging the role of the hippocampus in allocentric spatial memory, other investigators disputed that its only function is spatial. Instead, they claimed that the hippocampus is needed for memory for all types of relational information, whether it be among spatial elements, among objects, or among words (Cohen and Eichenbaum 1993). In support of the latter idea, investigators working with animal models showed that memory for single objects is affected little, if at all, by hippocampal lesions, but is disrupted severely by lesions to the perirhinal cortex, whereas memory for the association between objects and locations is affected by parahippocampal lesions (Murray 1996, Aggleton and Brown 1999). This division of labor is consistent with the connection these regions have to neocortex. The perirhinal cortex is connected primarily to the temporal lobes which are concerned with processing objects, whereas the parahippocampus is connected to the parietal lobes which specialize in processing spatial information.

Aggleton and Brown (1999) proposed that the functional differentiation among medial temporal lobe structures extends to the diencephalon, thus forming two integrated medial–temporal–diencephalic memory systems. One system, consisting of the hippocampus and its connections to the mamillary bodies and anterior thalamic nuclei, mediates memory for recall which relies on relational information; the other, consisting of the perirhinal cortex and its connections to the dorsomedial nucleus of the thalamus, is necessary for item recognition which depends on familiarity judgments.

The lesion evidence in humans is roughly consistent with these proposals based on animal models where dissociations of function are observed along the lines predicted by the models (but see commentary in Aggleton and Brown 1999). Because amnesic people with circumscribed lesions are rare, investigators have turned to neuroimaging studies in normal people to test the hypothesis that different aspects of memory are mediated by different regions of the medial temporal lobe. In general, the evidence has been supportive of the hypotheses that have been advanced here, with greater activation in the right hippocampus on tests of spatial (Maguire et al. 1998) and relational memory (Henke et al. 1999) and in the parahippocampal gyrus (possibly entorhinal cortex) on tests of object-location memory (Milner et al. 1997) and navigation (Agguire and D'Esposito 1999).

Whatever the final verdict is regarding the role of the various regions of the medial temporal lobe and diencephalon, the evidence indicates that the type, extent, and severity of anterograde amnesia is a function of the size, side, and location of the lesion. This rule applies as well to the deficits indicative of retrograde amnesia.

4. *Retrograde Amnesia and Memory Consolidation: Where and When Are Memories Stored?*

Whereas studies of anterograde amnesia tell us about memory acquisition, studies of retrograde amnesia provide clues about the time course involved in consolidating long-term memories and the physiological processes and neural substrates which contribute to consolidation and storage. Until recently, it was widely believed that retrograde amnesia associated with medial temporal and diencephalic damage was short-lasting and temporally graded, such that memory loss was more severe for information acquired near the time of amnesia onset than for that which was acquired long before (see Sect. 2.2 above). Accordingly, the medial temporal lobes, particularly the hippocampus, and possibly the diencephalon, were considered to be temporary memory structures, needed only for memory retention and retrieval until memories were consolidated in neocortex and other structures. They were then permanently stored there and could be retrieved directly from those regions.

Nadel and Moscovitch (1997) and Nadel et al. (2000) noted a number of problems with the accepted view. Though the duration of retrograde amnesia sometimes is short, more often retrograde amnesia for details of autobiographical events after large medial temporal (or diencepahlic) lesions can extend for decades, or even a lifetime (Warrington and Sanders 1971), far longer than it would be biologically plausible for the consolidation process to last. Retrograde amnesia for public events and personalities, however, is less extensive and often is temporally graded; this is truer still of semantic memory which includes knowledge of new vocabulary and facts about the world and ourselves (our address, the names of our friends, our job), what some have called *personal semantics* to distinguish them from autobiographical episodes (see Fig. 4). The distinction between temporally extensive and temporally limited retrograde amnesia also applies to spatial memory. Schematic cognitive maps of old neighborhoods that are adequate for navigation are retained but they lack topographical details and local environmental features, such as the appearance and location of particular homes, that would allow the person to have detailed cognitive maps of their locale (see Nadel et al. 2000).

Based on this evidence, Nadel and Moscovitch concluded, contrary to the traditional consolidation model, that the function of the medial temporal system is not temporally limited but that it is needed to represent even old memories in rich, multifaceted detail, be it autobiographical or spatial, for as long as the memory exists. Neocortical structures, on the other hand, are sufficient to form domain-specific and semantic representations based on regularities extracted from repeated experiences with words, objects, people, environments, and even of autobiographical episodes that one recollects repeatedly, creating a gist of each episode. The medial temporal lobe system may aid in the initial formation of these neocortical representations, but once formed they can exist on their own. Recent evidence from studies of children whose hippocampus was damaged at birth or shortly thereafter supports this view. Vargha-Khadem et al. (1997) found that they acquired sufficient general knowledge (semantic memories) to complete high school even though their memory for autobiographical episodes was impaired. Corroborating evidence is also provided by neuroimaging studies of recent and remote autobiographical and semantic memory. These studies found that the hippocampus is activated equally during retrieval of recent and remote autobiographical memories, but not of memory for public events or personal semantics (Ryan et al. in press, Maguire 2001) (see Fig. 5).

To account for this evidence, Nadel and Moscovitch (1997) and Nadel et al. (2000) proposed a Multiple Trace Theory (MTT) according to which a

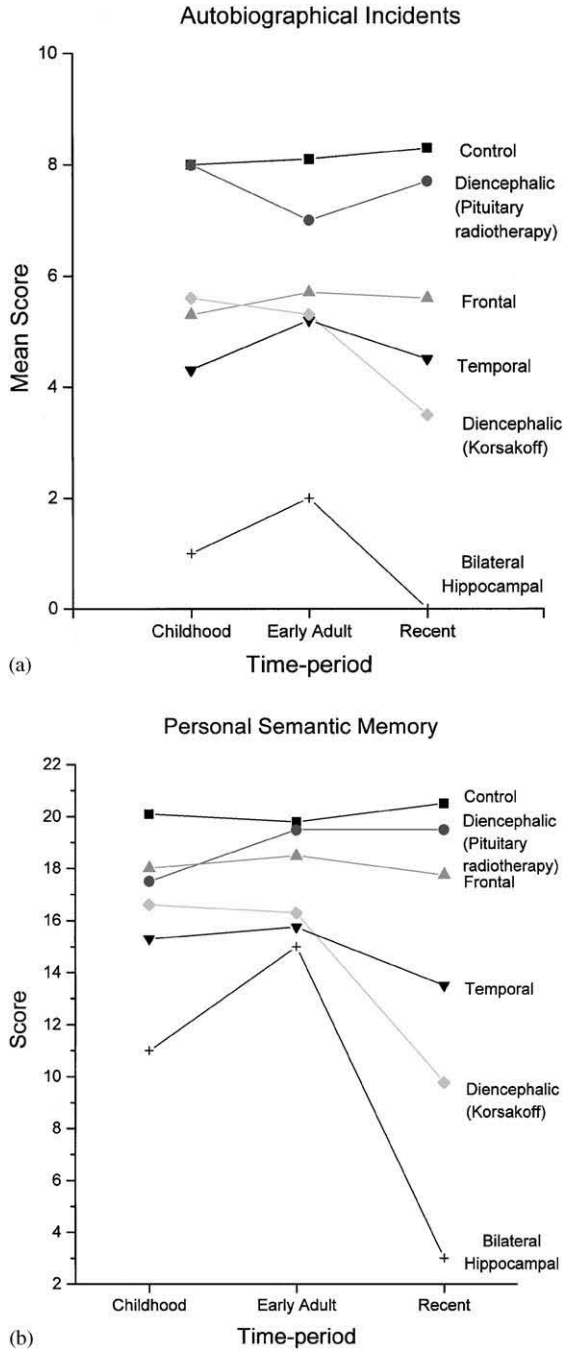

Figure 4
Example of (a) temporally-graded retrograde amnesia and (b) temporally-extended retrograde amnesia for autobiographical incidents and personal semantics in patients with bilateral medical temporal-lobe, hippocampal, and other lesions (modified from Kopelman et al. 1999 and Cipilotti et al. 2001)

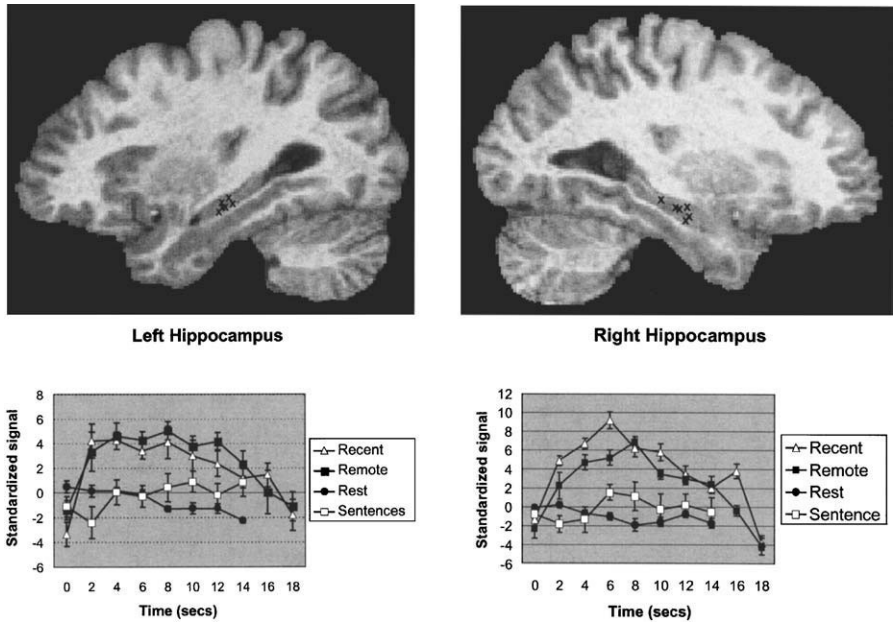

Figure 5
Hemodynamic response of the hippocampus during recall of recent and remote memories, and two baseline conditions (rest and sentence completion) (Ryan et al. in press)

memory trace of an episode consists of a bound ensemble of neocortical and hippocampal/medial temporal lobe (and possibly diencephalic) neurons which represent a memory of the consciously experienced event. Formation and consolidation of these traces, or cohesion (Moscovitch 1995), is relatively rapid, lasting on the order of seconds or at most days. Each time an old memory is retrieved, a new hippocampally mediated trace is created so that old memories are represented by more or stronger traces than are new ones, and therefore are less susceptible to disruption from brain damage than more recent ones. With respect to autobiographical episodes, the extent and severity of retrograde amnesia and perhaps the slope of the gradient are related to the amount and location of damage to the extended hippocampal complex. Remote memories for the gist of events, and for personal and public semantics, are not similarly dependent on the continuing function of the hippocampal complex (see McClelland et al. 1995 for a computational account for the usefulness of having complementary hippocampal and neocortical learning and memory systems).

Proponents of the standard consolidation model, however, argue that severe and temporally extensive retrograde amnesia is observed only when the lesion extends beyond the hippocampus to include neocortical structures where remote memories, both autobiographical and semantic, are represented (Squire and Zola 1998, but see Cipilotti et al. 2001). It remains to be determined what specific contributions the different regions of the medial temporal lobes and diencephalon make, and how they act in concert with the neocortex and other brain areas, to form and retain both detailed, contextually rich representations and context-independent knowledge (McDonald et al. 1999, Rosenbaum et al. in press).

5. *Amnesia and Neuropsychological Theories of Normal Memory: Which Types of Memory Are Affected?*

Research on amnesia and other memory disorders has influenced theories of normal memory at least since the end of the nineteenth century (Rozin 1976), but at no time has this been more apparent than in the last quarter of the twentieth century. Because amnesia is selective, affecting some memories and not others, research on amnesia has been used to promote the view that memory is not unitary, but rather consists of dissociable components, each governed by its own principles and mediated by different structures. For example, evidence showing that retrograde amnesia affects detailed autobiographical memory more than semantic memory supports the idea that episodic and semantic memory are distinct both

functionally and neurologically (Tulving 1983, Kinsbourne and Wood 1975, but see Squire and Zola 1998).

One of the characteristics of amnesia is that it affects long-term or secondary memory more than short-term or primary memory. This observation was one of the crucial pieces of evidence used in the 1960s and early 1970s to argue for a functional separation of memory into at least these two major components. The idea has since become almost universally accepted and opened the field to investigation of the functional differences between the two major components, to development of the concept of working memory (Baddeley 1986), and to identification of the mechanisms supporting working and primary memory in frontal and posterior neocortex (Smith and Jonides 1999, Petrides 2000).

Beginning in the late 1960s, research on animal models and humans indicated that the formation and retention of some types of long-term memory are spared in amnesia though there is continuing debate on how best to characterize them. Generally it is accepted that only conscious recollection is affected by amnesia. Memory retrieval without awareness seems to be intact. For example, it was noted that people with bilateral medial temporal lobe lesions, such as the patient H.M., could learn and retain motor skills for months or years, though he had no memory of the learning episode minutes after it was over. The same was true of perceptual skills involved in reading mirror-reversed words or in identifying degraded or fragmented pictures and words. That more than just a skill was involved became apparent when patients could complete or identify items to which they had been exposed more accurately and more quickly than new items, suggesting that they stored information peculiar to that item even though at a conscious level they could not recall or recognize that they had studied it (Warrington and Weiskrantz 1970) (Fig. 6). At the same time, similar phenomena were reported in normal people for material they could not consciously recollect, suggesting that the dissociation between memory with and without awareness was not peculiar to amnesia but was indicative of something fundamental about the organization of memory (see reviews in Cermak 1982, Cohen and Eichenbaum 1993, Kolb and Whishaw 1996, Tulving and Schacter 1990, Schacter and Buckner 1998, Squire and Knowlton 1999). These observations not only had a major impact on our understanding of normal memory, but also were instrumental in revitalizing research on unconscious processes in cognition and emotion, an area of investigation that had been abjured by mainstream experimental psychology for almost a century.

A number of terms have been used to refer to memory with and without awareness, including declarative and nondeclarative (or procedural memory), direct and indirect memory, memory and habit,

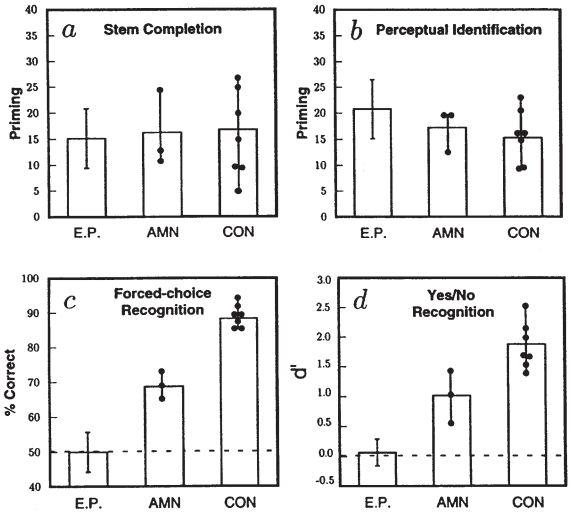

Figure 6
Example of amnesic and control performance on two implicit (stem completion and perceptual identification) and two explicit (forced choice and yes–no recognition) tests of memory (from Squire and Knowlton 1999)

controlled and automatic. We prefer the terms explicit and implicit memory because they are descriptively accurate, refer to the types of tests used to assess memory, and are close to theoretically neutral as to the processes and mechanisms involved.

Since the 1980s, research on normal people and on people with amnesia has identified a number of characteristics of implicit memory. The structures implicated in amnesia, the medial temporal lobes and diencephalon, are not needed for normal implicit memory. Instead, performance is mediated by a variety of structures, depending on the type of implicit memory that is being tested. Like explicit memory, implicit memory is not unitary and consists of a variety of different subtypes. Although a detailed description of all them is not possible, we mention three types that have been identified and about which we know a great deal: *perceptual implicit memory or priming, conceptual implicit memory, and procedural memory* (Knowlton and Squire 1999, Moscovitch 1992, Moscovitch et al. 1993, Schacter and Buckner 1998, Tulving and Schacter 1990).

On tests of perceptual implicit memory, the test stimulus resembles the studied target perceptually (e.g., a perceptually degraded version of it, or even an identical repetition of it) whereas on tests of conceptual implicit memory, the test stimulus resembles the target semantically (e.g., having studied the word 'horse,' the participant may be asked to make a semantic decision to the word at test, or asked to produce it in response to the word 'animal'). Performance is measured by speed or accuracy of

the response to the test stimulus, without explicit reference to the studied items. The implicit nature of the test is corroborated if the person is not aware that the response or test item refers to a studied stimulus. Increases in accuracy or decreases in response latency to old studied items as compared to new ones are indicative of implicit memory for the studied items.

Research on perceptual implicit memory suggests that performance is mediated by the same perceptual modules or representation systems in posterior neocortex that are involved in perceiving the stimuli. These modules are modified by the act of processing some given material, leaving behind a record of that process. As a result, processing is faster and more accurate when the system is re-exposed to that material as compared to new material (Wiggs and Martin 1998). The modules are domain specific, in that separate ones exist for objects, faces, words, and possibly places. They are also presemantic in that they do not represent the meaning of the item but only its structure. Thus, performance on perceptual implicit tests is sensitive to changes in perceptual or structural aspects of the stimulus but not to changes in semantic aspects.

The opposite holds for performance on conceptual implicit tests which is mediated by conceptual systems in the lateral temporal lobe and inferior frontal cortex (for reviews on implicit memory see Moscovitch et al. 1993, Gabrieli 1998, Schacter and Buckner 1998). As with perceptual modules, improvement in performance results from modifications in the conceptual systems themselves.

Tests of procedural implicit memory involve learning rules, motor sequences, conditioned responses, and subjective probabilities of stimulus–response associations without explicit memory for any of them. The structures that have been identified as crucial for procedural learning are the cerebellum for classical conditioning, and the basal ganglia for the others, with some indication of prefrontal involvement if learning rules or motor sequences involve strategic, sequential, or inhibitory components (Moscovitch et al. 1993, Squire and Knowlton 1999). Changes in these structures during execution of procedures underlies the changes in performance observed on implicit tests of procedural memory.

In studying implicit memory, great care must be taken to insure that performance on tests of memory that are ostensibly implicit are not contaminated by explicit components, such as might occur when asking participants to identify degraded stimuli they had studied earlier. It is for this reason that studies of amnesic patients is so useful. Because their explicit memory is so poor, equivalent performance between amnesic and normal people on an implicit test is taken as evidence that the test in question was not contaminated by explicit memory.

6. *Amnesia and Beyond: A Neuropsychological Component Process Model of Memory*

Studies of memory in amnesia have indicated that memory is not unitary but rather consists of a variety of different forms, each mediated by different component processes which in turn are subserved by different neural mechanisms. Because neither short-term or working memory, nor implicit memory is impaired in amnesia, it can be concluded that these types of memory are not dependent on the medial temporal lobes and diencephalic structures which are damaged in amnesia. The latter structures, however, are necessary for conscious recollection of long-term, episodic memories. It has been proposed that any information that is consciously experienced is picked up obligatorily and automatically by the hippocampus and related structures in the medial temporal lobes and diencephalon. These structures bind into a memory trace those neural elements in neocortex (and elsewhere) that mediate the conscious experience of an event. The episodic memory trace thus consists of an ensemble of hippocampal and neocortical neurons. The hippocampal component of the trace acts as an index or file entry pointing to the neural elements in neocortex that represent both the content of the event and the conscious experience of it. 'Consciousness' is, therefore, part of the memory trace. Retrieval occurs when an external or internally generated cue triggers the hippocampal index which in turn activates the entire neocortical ensemble associated with it. In this way we recover not only the content of an event but the consciousness that accompanied our experience of it. In short, when we recover episodic memories, we recover conscious experiences (Moscovitch 1995).

According to this model both encoding and retrieval of consciously apprehended information via the hippocampus and related structures is obligatory and automatic, yet we know from experience and from experimental investigation that we have a measure of control over what we encode and what we retrieve from memory. Moreover, if encoding is automatic and obligatory, the information cannot be organized, yet memory appears to have some temporal and thematic organization. How can we reconcile this model of memory with other facts we know about how memory works? One solution is that other structures, particularly those in the frontal lobes, control the information delivered to the medial temporal and diencephalic system at encoding, initiate and guide retrieval, and monitor, and help interpret and organize the information that is retrieved. By operating on the medial temporal and diencephalic system, the frontal lobes act as *working-with-memory* structures that control the more reflexive medial temporal and diencephalic system and confer a measure of intelligence and direction to it (Fig. 7). Such a complementary system is needed if

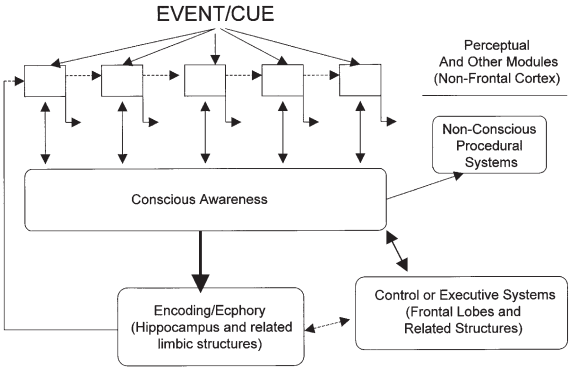

Figure 7
A schematic model of hippocampal complex-neocortical-frontal interaction during encoding and retrieval (from Moscovitch 1989, See text)

memory is to serve functions other than mere retention and retrieval of past experiences (Moscovitch 1992). As invaluable as studies of amnesia have been to our understanding of memory, those studies need to be supplemented by investigations of memory in people with other types of disorders, particularly those implicating the frontal lobes, to have a full appreciation of how memory works.

See also: Amnesia: Transient and Psychogenic; Declarative Memory, Neural Basis of; Dementia: Overview; Dementia, Semantic; Implicit Learning and Memory: Psychological and Neural Aspects; Learning and Memory, Neural Basis of; Memory, Consolidation of; Technology-supported Learning Environments

Bibliography

Aggleton J P, Brown M W 1999 Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral Brain Science* **22**: 425–89

Agguire G K, D'Esposito M 1999 Topographical disorientation: A synthesis and taxonomy. *Brain* **122**: 1613–28

Baddeley A 1986 *Working Memory*. Oxford University Press, Oxford, UK

Cermak L S (ed.) 1982 *Human Memory and Amnesia*, Erlbaum, Hillsdale, NJ

Cipilotti L, Shallice T, Chan D, Fox N, Scahill R, Harrison G, Stevens J, Rudge P 2001 Long-term retrograde amnesia … the crucial role of the hippocampus. *Neuropsychologia* **39**: 151–72

Cohen N J, Eichenbaum H 1993 *Memory, Amnesia and the Hippocampal System*. MIT Press, Cambridge, MA

Corkin S, Amaral D G, Gonzalez R G, Johnson K A, Hyman B T H.M.'s medial temporal lobe lesion: Findings from magnetic resonance imaging. *The Journal of Neuroscience* **17**: 3964–79

Eichenbaum H 1999 The hippocampus and mechanisms of declarative memory. *Behavioral Brain Research* **103**: 123–33

Gabrieli J D E 1998 Cognitive neuroscience of human memory. *Annual Review of Psychology* **49**: 87–115

Hamilton C W 1976 *Basic Limbic Anatomy of the Rat*. Plenum, New York and London

Henke K, Weber B, Kneifel S, Wieser H G, Buck A 1999 The hippocampus associates information in memory. *Proceedings of the National Academy of Science USA* **96**: 5884–9

Kapur N 1999 Syndromes of retrograde amnesia: A conceptual and empirical analysis. *Psychological Bulletin* **125**: 800–25

Kinsbourne M, Wood F 1975 Short-term memory processes and the amnesic syndrome. In: Deutsch D, Deutsch A J (eds.) *Short-term Memory*, Academic Press, New York

Kolb B, Whishaw I Q 1996 *Fundamentals of Human Neuropsychology*, 4th edn. Freeman, New York

Kopelman M D, Stanhope N, Kingsley D 1999 Retrograde amnesia in patients with diencephalic, temporal lobe or frontal lesions. *Neuropsychologia* **37**: 939–58

Korsakoff S S 1889 Etudes medico psychologique sur une forme du maladi de la mémoire. *Review Philosoph* **28**: 501–30 (trans. and republished by Victor M, Yakovlev P I 1955 *Neurology* **5**: 394–406)

Maguire E A (in press) Neuroimaging studies of autobiographical event memory. *Philosophical Transactions of the Royal Society of London Series B–Biological Sciences*

Maguire E A, Burgess N, Donnett J G, Frackowiack R S J, Frith C D, O'Keefe J 1998 Knowing where and getting there: A human navigation network. *Science* **280**: 921–4

McClelland J L, McNaughton B L, O'Reilly R C 1995 Why are there complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–57

McDonald R M, Ergis A-M, Winocur G 1999 Functional dissociation of brain regions in learning and memory: Evidence for multiple systems. In: Foster J K, Jelicic M (eds.) *Memory: Systems, Process, or Function*. Oxford University Press, Oxford

Milner B 1966 Amnesia following operation on the temporal lobe. In: Whitty C W M, Zangwill O L (eds.) *Amnesia*, Butterworth, London

Milner B, Johnsrude I, Crane J 1997 Right temporal-lobe contribution to object-location memory. *Philosophical Transactions of the Royal Society of London Series B* **352**: 1469–74

Moscovitch M 1992 Memory and working with memory: A component process model based on modules and central systems. *Journal of Cognitive Neuroscience* **4**: 257–67

Moscovitch M 1995 Recovered consciousness: A hypothesis concerning modularity and episodic memory. *Journal of Clinical Experimental Neuropsychology* **17**: 276–91

Moscovitch M, Vriezen E, Goshen-Gottstein Y 1993 Implicit tests of memory in patients with focal lesions or degenerative brain disorders. In: Boller F, Spinnler H (eds.) *The Handbook of Neuropsychology*, Vol. 8. Elsevier, Amsterdam, The Netherlands pp. 133–73

Moscovitch M, Winocur G 1992 The neuropsychology of memory and aging. In: Craik F I M, Salthouse T A (eds.) *The Handbook of Aging and Cognition*. Erlbaum, Hillsdale, NJ, pp. 315–72

Murray E A 1996 What have ablation studies told us about the neural substrates of stimulus memory? *Seminars in Neuroscience* **8**: 13–22

Nadel L, Moscovitch M 1997 Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinions in Neurobiology* **7**: 217–27

Nadel L, Samsonovich A, Ryan L, Moscovitch M 2000 Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results. *Hippocampus* **10**: 352–68

O'Keefe J, Nadel L 1978 *The Hippocampus as a Cognitive Map*. Oxford University Press, Oxford, UK

Petricles M 2000 The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research* **133**: 44–54

Rosenbaum R S, Winocur G, Moscovitch M (in press) New views on old memories: Reevaluating the role of the hippocampal complex. *Experimental Brain Research*

Rozin 1976 The psychobiological approach to human memory. In: Rozenzweig R Bennett E L (eds.) *Neural Mechanisms of Learning and Memory*. MIT Press, Cambridge, MA

Ryan L, Nadel L, Keil K, Putnam K, Schnyer D, Trouard D, Moscovitch M (in press) The hippocampal complex and retrieval of recent and very remote autobiographical memories: Evidence from functional magnetic resonance imaging in neurologically intact people. *Hippocampus*

Schacter D L, Buckner R L 1998 Priming and the brain. *Neuron* **20**: 185–95

Scoville W B, Milner B 1957 Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry* **20**: 11–21

Smith E E, Jonides J 1999 Storage and executive processes of the frontal lobes. *Science* **283**: 1657–61

Squire L R, Alvarez P 1995 Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinions in Neurobiology* **5**: 169–77

Squire L R, Knowlton B J 1999 The medial temporal lobe, the hippocampus and the memory systems of the brain. In: Gazzuniga M S (ed.) *The Cognitive Neurosciences*, 2nd edn., MIT Press, Cambridge, MA, pp. 765–80

Squire L R, Zola S M 1998 Episodic memory, semantic memory, and amnesia. *Hippocampus* **8**: 205–11

Tulving E 1983 *Elements of Episodic Memory*. Clarendon Press, Oxford, UK

Tulving E, Schacter D L 1990 Priming and human memory systems. *Science* **247**: 301–6

Vargha-Khadem F, Gadian D G, Watkins K E, Conneley A, Van Paesschen W, Mishkin M 1997 Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**: 376–80

Warrington E K, Sanders H I 1971 The fate of old memories. *Quarterly Journal of Experimental Psychology* **23**: 432–42

Warrington E K, Weiskrantz L 1970 Amnesic syndrome: Consolidation or retrieval? *Nature* **228**: 628–30

Wiggs C L, Martin A 1998 Properties and mechanisms of perceptual priming. *Current Opinions in Neurobiology* **8**: 227–33

M. Moscovitch

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

and Memory: Psychological and Neural Aspects; Learning and Memory, Neural Basis of; Memory, Consolidation of; Technology-supported Learning Environments

Bibliography

Aggleton J P, Brown M W 1999 Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral Brain Science* **22**: 425–89

Agguire G K, D'Esposito M 1999 Topographical disorientation: A synthesis and taxonomy. *Brain* **122**: 1613–28

Baddeley A 1986 *Working Memory*. Oxford University Press, Oxford, UK

Cermak L S (ed.) 1982 *Human Memory and Amnesia*. Erlbaum, Hillsdale, NJ

Cipilotti L, Shallice T, Chan D, Fox N, Scahill R, Harrison G, Stevens J, Rudge P 2001 Long-term retrograde amnesia … the crucial role of the hippocampus. *Neuropsychologia* **39**: 151–72

Cohen N J, Eichenbaum H 1993 *Memory, Amnesia and the Hippocampal System*. MIT Press, Cambridge, MA

Eichenbaum H 1999 The hippocampus and mechanisms of declarative memory. *Behavioral Brain Research* **103**: 123–33

Henke K, Weber B, Kneifel S, Wieser H G, Buck A 1999 The hippocampus associates information in memory. *Proceedings of the National Academy of Science USA* **96**: 5884–9

Kapur N 1999 Syndromes of retrograde amnesia: A conceptual and empirical analysis. *Psychological Bulletin* **125**: 800–25

Kinsbourne M, Wood F 1975 Short-term memory processes and the amnesic syndrome. In: Deutsch D, Deutsch A J (eds.) *Short-term Memory*. Academic Press, New York

Kolb B, Whishaw I Q 1996 *Fundamentals of Human Neuropscyhology*, 4th edn. Freeman, New York

Kopelman M D, Stanhope N, Kingsley D 1999 Retrograde amnesia in patients with diencephalic, temporal lobe or frontal lesions. *Neuropsychologia* **37**: 939–58

Korsakoff S S 1889 Etudes medico psychologique sur une forme du maladi de la mémoire. *Review Philosoph.* **28**: 501–30 (trans. and republished by Victor M, Yakovlev P I 1955 *Neurology* **5**: 394–406)

Maguire E A in press Neuroimaging studies of autobiographical event memory. *Philosophical Transactions of the Royal Society of London Series B–Biological Sciences*

Maguire E A, Burgess N, Donnett J G, Frackowiack R S J, Frith C D, O'Keefe J 1998 Knowing where and getting there: A human navigation network. *Science* **280**: 921–4

McClelland J L, McNaughton B L, O'Reilly R C 1995 Why are there complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–57

Milner B 1966 Amnesia following operation on the temporal lobe. In: Whitty C W M, Zangwill O L (eds.) *Amnesia*. Butterworth, London

Milner B, Johnsrude I, Crane J 1997 Right temporal-lobe contribution to object-location memory. *Philosophical Transactions of the Royal Society of London Series B* **352**: 1469–74

Moscovitch M 1992 Memory and working with memory: A component process model based on modules and central systems. *Journal of Cognitive Neuroscience* **4**: 257–67

Moscovitch M 1995 Recovered consciousness: A hypothesis concerning modularity and episodic memory. *Journal of Clinical Experimental Neuropsychology* **17**: 276–91

Moscovitch M, Vriezen E, Goshen-Gottstein Y 1993 Implicit tests of memory in patients with focal lesions or degenerative brain disorders. In: Boller F, Spinnler H (eds.) *The Handbook of Neuropsychology*, Vol. 8. Elsevier, Amsterdam, The Netherlands, pp. 133–73

Moscovitch M, Winocur G 1992 The neuropsychology of memory and aging. In: Craik F I M, Salthouse T A (eds.) *The Handbook of Aging and Cognition*. Erlbaum, Hillsdale, NJ, pp. 315–72

Murray E A 1996 What have ablation studies told us about the neural substrates of stimulus memory? *Seminars in Neuroscience* **8**: 13–22

Nadel L, Moscovitch M 1997 Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinions in Neurobiology* **7**: 217–27

Nadel L, Samsonovich A, Ryan L, Moscovitch M 2000 Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results. *Hippocampus* **10**: 352–68

O'Keefe J, Nadel L 1978 *The Hippocampus as a Cognitive Map*. Oxford University Press, Oxford, UK

Rozin P 1976 The psychobiological approach to human memory. In: Rozenzweig R, Bennett E L (eds.) *Neural Mechanisms of Learning and Memory*. MIT Press, Cambridge, MA

Ryan L, Nadel L, Keil K, Putnam K, Schnyer D, Trouard, Moscovitch M in press The hippocampal complex and retrieval of recent and very remote autobiographical memories: Evidence from functional magnetic resonance imaging in neurologically intact people. *Hippocampus*

Schacter D L, Buckner R L 1998 Priming and the brain. *Neuron* **20**: 185–95

Scoville W B, Milner B 1957 Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry* **20**: 11–21

Squire L R, Alvarez P 1995 Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinions in Neurobiology* **5**: 169–77

Squire L R, Zola S M 1998 Episodic memory, semantic memory, and amnesia. *Hippocampus* **8**: 205–11

Tulving E 1983 *Elements of Episodic Memory*. Clarendon Press, Oxford, UK

Tulving E, Schacter D L 1990 Priming and human memory systems. *Science* **247**: 301–6

Vargha-Khadem F, Gadian D G, Watkins K E, Conneley A, Van Paesschen W, Mishkin M 1997 Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**: 376–80

Warrington E K, Sanders H I 1971 The fate of old memories. *Quarterly Journal of Experimental Psychology* **23**: 432–42

Warrington E K, Weiskrantz L 1970 Amnesic syndrome: Consolidation or retrieval? *Nature* **228**: 628–30

Wiggs C L, Martin A 1998 Properties and mechanisms of perceptual priming. *Current Opinions in Neurobiology* **8**: 227–33

M. Moscovitch

Amnesia: Transient and Psychogenic

Memory is an essential feature for an integrated personality and the ability to lead a normal life. The most complete form of memory disturbance, amnesia

(see *Amnesia*), frequently is regarded as a permanent state arising from focal brain damage in bottleneck structures of the brain. This form of so-called global amnesia (Table 1) is usually caused by bilateral damage of the regions in the limbic system (see *Limbic System*), divided into the medial temporal lobe (see *Hippocampus and Related Structures*; *Temporal Lobe*), the medial diencephalon (see *Hypothalamus*), and the basal forebrain.

On the other hand, already in the nineteenth century various reports on patients with transient forms of amnesia had appeared and frequently were subsumed under the heading of hysteria. Hysteria was seen in principle as a treatable or even curable state of an individual with a personality demonstrating a deviance or discrepancy between their nonconscious presentation and their true character. In line with this view is the described selectivity of the amnesic condition which is confined to person-relevant or individual-specific episodes, implying that only the episodic memory system (see *Episodic and Autobiographical Memory: Psychological and Neural Aspects*) is affected while world or general knowledge (the semantic memory system) (see *Semantic Knowledge: Neural Basis of*) is preserved (Table 1). When the personal past is relearned, this relearning occurs usually with a reduced affect (*la belle indifference*) (Markowitsch 1999). (Episodic memory refers to the possibility of traveling back in time; that is, it allows the reinstatement of past episodes with the context of time and place; semantic memory, on the other hand, refers to a context-free instatement of general facts.)

Psychic conditions, in particular environmental stress factors (see *Stress, Neural Basis of*), may lead to a number of amnesic conditions, some of which are of a more general and others of a more specific nature. The general conditions refer to an inability to retrieve the personal past such as in psychogenic amnesia (Markowitsch et al. 1997a), or psychogenic fugues (Markowitsch et al. 1997b), conditions nowadays subsumed under the dissociative states. Furthermore, in rare cases there may be an inability to acquire new (episodic) information long-term (Markowitsch 1999). In more specific instances, the ability to retrieve episodes of a certain nature or of a limited time period may be impaired and, consequently, the respective episodes during which a person may have been sexually abused, may remain inaccessible over years and decades (Markowitsch 1999). Sometimes, somatic injuries (e.g., whiplash injury) may provoke amnesic conditions (Markowitsch 1999). From this introduction it can be inferred correctly that a strict division into brain-damage-caused ('somatic') and psychically generated amnesias is frequently not possible. In fact, there is increasing evidence for a mixture, justifying the use of the expression 'functional amnesia' which leaves the origin or etiology open (Markowitsch 1999). Nevertheless, the more homogenous form of transient global amnesia, which is considered to be a neurological disease, will first be described, and then functional amnesias will be discussed.

1. *Transient Global Amnesia*

The dimension of affect is of importance as well in a number of cases with so-called transient global amnesia (TGA) (Hodges 1991, Markowitsch 1990). TGA is a still not very well understood condition, largely because of its time-limited nature and its many

Table 1
General forms of amnesic states and their behavioral characteristics

Feature	Global amnesia	Transient global amnesia	Functional (psychogenic) amnesia
Occurrence	Usually suddenly as a consequence of focal brain damage	Suddenly, frequently after significant physical or psychic events	Suddenly, either after chronic stress or after a single major event
Duration	Frequently permanent	$\leqslant$ 24 h	Variable
Affected forms of memory	Episodic and semantic	Mainly episodic	Episodic–autobiographic
Affected time range	Anterograde and to a minor degree retrograde	Anterograde and limited retrograde	Retrograde and/or (more rarely) anterograde
Patient's behavior	Abnormal, sometimes confuse, neurological signs	Confuse, but no neurological signs	Intentional, reality-oriented
Remission/recovery	Various degrees, depending on the lesioned locus	Complete, except for the episode of the attack itself	Sometimes complete, sometimes persistent

possible etiologies. In principle, TGA refers to a sudden severe memory loss without concomitant brain damage (or epilepsy); intelligence and consciousness (see *Consciousness, Neural Basis of* and *Conscious and Unconscious Processes in Cognition*) are preserved, the duration lasts less than one day. Transient amnesias caused by specifically induced external interventions, such as electroconvulsive therapy (see, e.g., Squire and Slater 1975), also do not fit the definition of TGA.

Essential criteria for diagnosing TGA differ somewhat between authors (Caplan 1990, Hodges 1991, Frederiks 1993). In essence, the short duration ($\leqslant 24$ h) is emphasized and the predominance of its occurrence late in adulthood (usually $\geqslant 60$ yr) (Hodges 1991; Markowitsch 1990). It is characterized by a major amnesia which goes mainly in the anterograde, but also in retrograde direction (Härting and Markowitsch 1996). Though initially it was assumed that attacks rarely recur, a more thorough screening of relevant data indicates that many patients experience more than one event (Caplan 1990, Frederiks 1993). Risk factors include high blood pressure, coronary heart disease, previous strokes or transient ischaemic attacks, migraine, hyperlipidemia, smoking, diabetes, and peripheral vascular diseases (Caplan 1990, Frederiks 1993, Hodges 1991). Imaging results (see *Functional Brain Imaging*) indicated cellular edema in the anterior temporal lobe during TGA or hypometabolic zones in the memory processing regions of the limbic system.

The precipitants of TGA include physical and psychic factors; among them are emotional stress, pain, sexual intercourse, physical activity, exposure to cold or hot water (Caplan 1990, Markowitsch 1990). This list shows that various stressors (see *Stress, Neural Basis of*) and hemodynamic challenges can lead to TGA.

The TGA episodes are neuropsychologically quite uniform. There are anterograde and retrograde components to the amnesia. As the patient recovers, the retrograde amnesia shrinks and then totally disappears just as the patients gain their ability to retain new information. Following the ictus, the patient is left with a permanent island of memory loss for the period encompassed by the TGA episode. During the ictus, the patient repetitively asks questions about his or her plight, identity and location in ways that indicate an acute awareness of the amnesia.

2. *Functional Amnesias*

The spectrum of antecedents of 'functional amnesias' is quite variable; among other features minor head trauma, psychiatric disorders, depression, stress (see *Stress, Neural Basis of*), and chronic fatigue syndrome can be listed. That amnesia may follow both psychic and somatic shock conditions has been known for a long time. Already at the turn of the nineteenth century prominent scientists distinguished between four forms of shock and emphasized the existence of a psychic shock which may result in memory disturbances. The phenomenon of post-traumatic stress disorder (PTSD) can be viewed as a present-day description of such early observations. Functional amnesias may interfere with the recall of all autobiographical information (Markowitsch et al. 1997b), or may even lead to both retrograde and anterograde amnesia of a more general nature (Markowitsch et al. 2000). In the following, common forms of functional amnesias will be described.

Minor head injury (without identifiable structural brain damage) is sometimes accompanied by lasting and major retrograde amnesia for autobiographic memory. Barbarotto et al. (1996) described a woman who slipped and fell in her office. Though no brain damage was detected, she nevertheless remained retrogradely amnesic even when tested six months after the event. The authors describe a personality pattern compatible with conversion hysteria (see Table 2). Persistent anterograde amnesia with preserved retrograde memories after a whiplash injury without measurable brain damage was described for a young woman (Markowitsch 1999).

Cases with pure psychogenic amnesia (i.e., without evidence for somatic [brain] injuries) show some common features: a weak, underdeveloped personality, a problematic childhood or youth and the appearance of emotionally negative events such as sexual abuse (Markowitsch 1999) during early life. The amnesia may be interpreted as a mechanism for blocking awareness of previous traumatic events (Markowitsch 1998).

Functional brain imaging such as positron emission tomography (PET) (see *Functional Brain Imaging*) may help to obtain evidence for altered neural processing in patients with functional amnesias. Markowitsch et al. (1997b) investigated a 37-year-old man with a persistent fugue. He remained unsure of his relationship to family members and changed a number of personality traits. While having been an avid car driver prior to the fugue, he became quite hesitant to enter a car thereafter, lost his asthma, gained substantially in body weight and changed his profession. Regional cerebral blood flow was measured with PET in an autobiographic memory paradigm and revealed a mainly left-hemispheric activation instead of the usual right-hemispheric frontotemporal one found with the same paradigm by Fink et al. (1996). This result suggests that he might indeed have been unable to recall his own past and processed this information as new and unrelated to his person. Related to this finding, another functional imaging technique (single photon emitted computer tomography; SPECT) revealed in another patient with psychogenic amnesia a hypometabolic zone in exactly that frontotemporal junction area which is necessary for normal retrieval of autobiographical memories (Markowitsch et al.

1997a). A similar hypometabolic zone in the same brain region was detected after organic brain damage and retrograde amnesia (Calabrese et al. 1996). These findings suggest that there may be a block or disconnection of memory-processing neural nets, leading to both 'psychogenic' or 'organic' amnesia (Markowitsch 1999).

Related to these findings, Markowitsch et al. (2000) described a patient who at the age of four witnessed a man burn to death and then at the age of 23 had an open fire in his house. Immediately following this second exposure to a fire, the patient developed anterograde amnesia and a retrograde amnesia covering a period of six years. PET results demonstrated a severe reduction of glucose metabolism, especially in the memory-related temporal and diencephalic regions of the brain. His amnesic condition persisted for months but improved thereafter as his brain metabolism returned to normal levels after one year. Even then his long-term memory was quite poor so that he remained unable to return to his former job.

Findings of this kind indicate that psychological stress can conceivably alter the structure and function of memory-processing brain areas, perhaps through the mediation of stress-related hormones which are active through the pituitary-adrenal axis (see *Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of*) (Markowitsch 1999). For example, patients with PTSD (see Table 2) are vulnerable to becoming depressed and to manifesting memory disturbances; the release of endogenous stress-related hormones is dysregulated in PTSD and this abnormality may result in altered neural function (Markowitsch 1999).

Some studies even point to hippocampal volume reductions in association with stress (Gurvits et al. 1996, Stein et al. 1997). Massive and prolonged stress might induce changes in regions with a high density of glucocorticoid receptors such as the anterior and medial temporal lobe. In animals, it was shown that stress enhances hippocampal long-term depression (see *Long-term Potentiation (Hippocampus)*) and blocks long-term potentiation (see *Long-term Depression (Hippocampus)*). Teicher et al. (1993) found that early physical or sexual abuse hinders the development of the limbic system and may therefore induce a predisposition for the outbreak of stress-related amnesia syndromes.

3. Conclusions

The possibility of a co-existence of neurological and psychiatric or psychogenic forms of amnesia in one individual is evident from the above descriptions and had been emphasized before with Mai's (1995) statement that 'the presence of a neurological abnormality does not necessarily rule out psychogenic amnesia' (p. 105). Mai (1995), who discussed the use of the term 'hysteria' in clinical neurology, pointed out that the behavioral conditions described above make it necessary to distinguish between a *disease* and an *illness*. 'Disease' being the condition associated with pathological disturbance of structure or function, and 'illness' being the experience associated with ill-health, symptoms, and suffering. Consequently, individuals may have an illness without a disease or a disease without an illness. The independence between these two states is most clearly evident in the amnesias described above and makes them particularly interesting examples for neuroscientific investigations on the representation of memory in the brain. In the search for uncovering the neural mechanisms of information processing, particular emphasis should be laid on dynamic biochemical alterations in the brain—such as the availability or block of transmitters, neuromodulators, and hormones (including stress hormones).

See also: Amnesia; Memory, Consolidation of; Memory Retrieval

Bibliography

Barbarotto R, Laiacona M, Cocchini G 1996 A case of simulated, psychogenic or focal pure retrograde amnesia: Did an entire life become unconscious? *Neuropsychologia* **34**: 575–85

Calabrese P, Markowitsch H J, Durwen H F, Widlitzek B, Haupts M, Holinka B, Gehlen W 1996 Right temperofrontal cortex as critical locus for the ecphory of old episodic memories. *Journal of Neurology, Neurosurgery, and Psychiatry* **61**: 304–10

Caplan L R 1990 Transient global amnesia: Characteristic features and overview. In: Markowitsch H J (ed.) *Transient Global Amnesia and Related Disorders*. Hogrefe, Toronto, pp. 15–27

Fink G R, Markowitsch H J, Reinkemeier M, Bruckbauer T, Kessler J, Heiss W-D 1996 A PET-study of autobiographical memory recognition. *Journal of Neuroscience* **16**: 4275–82

Frederiks J A M 1993 Transient global amnesia. *Clinical Neurology and Neurosurgery* **95**: 265–83

Gurvits T V, Shenton M E, Hokama H, Ohta H, Lasko N B, Gilbertson M W, Orr S P, Knis R, Jolesz F A, McCarley R W, Pitman R K 1996 Magnetic resonance imaging study of hippocampal volume in chronic, combat-related post-traumatic stress disorder. *Biological Psychiatry* **40**: 1091–9

Härting C, Markowitsch H J 1996 Different degrees of impairment in recall/recognition and anterograde/retrograde memory performance in a transient global amnesic case. *Neurocase* **2**: 45–9

Hodges J R 1991 *Transient Amnesia: Clinical and Neuropsychological Aspects*. Saunders, London

Mai F M 1995 Hysteria in clinical neurology. *Canadian Journal of the Neurological Sciences* **22**: 101–10

Markowitsch H J (ed.) 1990 *Transient Global Amnesia and Related Disorders*. Hogrefe & Huber, Toronto

Markowitsch H J 1998 The mnestic block syndrome: Environmentally induced amnesia. *Neurology Psychiatry and Brain Research* **6**: 73–80

Markowitsch H J 1999 Functional neuroimaging correlates of functional amnesia. *Memory* **7**: 561–83

Markowitsch H J, Calabrese P, Fink G R, Durwen H F, Kessler J, Härting C, König M, Mirzaian E B, Heiss W-D, Heuser L, Gehlen W 1997a Impaired episodic memory retrieval in a case of probable psychogenic amnesia. *Psychiatry Research: Neuroimaging Section* **74**: 119–26

Markowitsch H J, Fink G R, Thöne A I M, Kessler J, Heiss W-D 1977b Persistent psychogenic amnesia with a PET-proven organic basis. *Cognition and Neuropsychiatry* **2**: 135–58

Markowitsch H J, Kessler J, Weber-Luxenburger G, Van der Ven C, Heiss W-D 2000 Neuroimaging and behavioural correlates of recovery from 'mnestic block syndrome' and other cognitive deteriorations. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology* **13**: 60–6

Squire L R, Slater P C 1975 Electroconvulsive therapy and complaints of memory dysfunction: A prospective three-year follow-up study. *British Journal of Psychology* **142**: 1–8

Stein M B, Koverola C, Hanna C, Torchia M G, McClarty B 1997 Hippocampal volume in women victimized by childhood sexual abuse. *Psychological Medicine* **27**: 951–9

Teicher M H, Glod C A, Surrey J, Swett C 1993 Early childhood abuse and limbic system ratings in adult psychiatric outpatients. *Journal of Neuropsychiatry and Clinical Neuroscience* **5**: 301–6

H. J. Markowitsch

Amygdala (Amygdaloid Complex)

'Amygdala' and 'amygdaloid complex' are interchangeable names used today for this temporal lobe structure which has been linked to some of the most complex functions of the brain, including emotion, mnemonic processes, and ingestive, sexual, and social behavior. Not surprisingly, the underlying neuronal circuits through which the amygdala expresses such a variety of functions are exceedingly complex and highly differentiated.

1. *Location and Basic Amygdalar Divisions*

The amygdala occupies a large portion of the inferior temporal lobe, beginning just caudal to the nucleus of the diagonal band in the front, and extending almost to the end of the cerebral hemisphere. It was first identified in human brain and named by Burdach at the beginning of the nineteenth century. Starting about 70 years later, other early descriptions of the amygdala in various species followed, and the microscopic examination of histological tissue sections began to reveal structural differentiation in the amygdala. Today, more than a dozen distinct cell groups (nuclei) can be identified within the amygdala on the cytoarchitectonic, histochemical, connectional, and functional basis. Even though there are variations in the size and position of some nuclei in different species, the basic pattern of amygdalar nuclei appears to be the same in all mammals.

Johnston (1923) introduced the fundamental description of amygdalar structure in widest use today, based on detailed analysis of comparative vertebrate material. He proposed that the amygdala be divided into a primitive group of nuclei associated with the olfactory system (central, medial, and cortical nuclei, and the nucleus of the olfactory tract), and a phylogenetically new group of nuclei (basal and lateral). Many physiologists have adopted this parcellation, resulting in a large body of evidence suggesting functional differences between these two groups. For instance, the corticomedial group, which receives direct olfactory input from both the main and accessory olfactory bulbs, is involved in the expression of agonistic, and various aspects of sexual and ingestive behavior. The basolateral group, which receives most other classes of sensory information, and has widespread isocortical and hippocampal connections, is important for emotional responses as well as mnemonic processes such as learning and memory. Although this general division works reasonably well, rapidly accumulating neuroanatomical, physiological, and behavioral results have made it clear that the amygdala is divided into more than two functional groups, and its complex intrinsic connections suggest that different functional systems are highly interactive.

2. *Amygdalar Function*

At the beginning of the twentieth century it was still widely believed that the amygdala was involved primarily in olfactory functions, and that its main connections were with olfactory areas and the hypothalamus. Theories about the function of the amygdala underwent radical changes after Klüver and Bucy (1939) examined the effects of the temporal lobe lesions in monkeys. These lesions, which included the amygdala, produced severe behavioral impairments including tameness, lack of emotional responses, hypersexuality, and excessive oral tendencies. The Klüver–Bucy syndrome can be summarized as an overall inability to identify the biological significance of stimuli. A little over a decade later, Weiskrantz performed specific lesions of the amygdala alone in monkeys, and produced similar effects lending to the hypothesis that the amygdala is necessary for establishment of positive or negative reinforcing stimuli.

Since the 1950s further evidence has been provided for amygdalar involvement in circuits necessary for stimulus-reinforcement learning, and a number of theoretical models of emotion included the amygdala as a central component. The exact role of the amygdala in emotion and motivation, however, is not clear yet. The best understood amygdalar function is, arguably, its role in conditioned fear, the main research focus for the last two decades. Rapidly accumulating neuroanatomical and functional evidence strongly suggest

that the amygdala may be crucial for the acquisition and expression of conditioned fear (see *Fear Conditioning*; *Anxiety and Fear, Neural Basis of*).

In addition, the amygdala is important for other types of aversive learning, such as avoidance learning (see *Avoidance Learning and Escape Learning*). In this type of learning the amygdala is believed to be modulating learning and memory circuits elsewhere in the brain, through connections via the stria terminalis pathway. Most recently, a role in other aspects of cognition, such as attention, has been suggested for the amygdala, although considerably less is known about the exact circuitry and mechanisms of this function.

Pioneering work of Kaada (1972), together with a large body of physiological and behavioral evidence accumulated over the last 50 years, provide powerful evidence for amygdalar influence on arousal, motivation, and the expression of reproductive, ingestive, and species-specific defensive and aggressive behavior. In more recent years, amygdalar influence on circuitry mediating reward-related behaviors has been shown as well. Generally it is believed that the amygdalar role in these behaviors is modulatory, and is exercised through an extensive network of projections to other brain regions more directly involved with these functions.

Despite the literature there is no real consensus about a unifying role for the amygdala. The most widely accepted hypothesis is that the amygdala attaches motivational and emotional significance to incoming sensory stimuli through associations with rewarding or aversive events. Based on the outcome of this pairing process the amygdala modulates other functional systems including reflex circuits, as well as systems involved in learning, memory, reward, arousal, and attention. Within this context it is important to emphasize again that the amygdala is not a unified region, and that its different subsystems are likely performing distinct functions. Thus, the challenge for future research will be in identifying amygdalar functional subsystems, and the larger circuitry they form within the cerebral hemisphere.

Our understanding of amygdalar functions is based primarily on evidence provided by animal studies. Nevertheless, in recent years a number of human studies using functional magnetic resonance imaging or positron emission tomography have confirmed the findings from the animal work, and in addition, provided evidence for amygdalar role in development of various pathologies including anxiety, depression, and schizophrenia.

3. *Connections of the Amygdala*

A detailed review of the literature on amygdalar connection—which is vast, complex, and contradictory—is beyond the scope of this article. Instead, a brief summary of major amygdalar connections, with several broad generalizations, will be presented.

The basic outline of amygdalar connections was demonstrated in early studies using axonal degeneration methods, however, most of our current knowledge is based on experiments using modern neuroanatomical pathway tracing techniques. The amygdala shares connections with all parts of the forebrain including areas in the cortex and basal ganglia in the telencephalon; thalamus, and hypothalamus in the diencepahlon; and with a number of sensory, autonomic, and behavioral areas in the midbrain and brainstem.

Most amygdalar connections are bi-directional and relay information concerning virtually all sensory modalities. Amygdalar nuclei receive inputs from polymodal cortical, thalamic, and brainstem sensory areas, as well as direct input from the olfactory bulbs. Thus, sensory information arrives to the amygdala after vastly different amounts of processing. These range from relatively unprocessed sensory information arriving from the olfactory bulb, brainstem, and thalamus, to highly processed (cognitive) information from the temporal, insular, and perirhinal cortical areas, and the hippocampal formation. A complex pattern of the intrinsic connections suggest that sensory information undergoes extensive processing within the amygdala before it is relayed further to a number of distinct functional systems, which are described below.

First, there is a well-known topographically organized projection directly to the medial and lateral zone of the hypothalamus, and via the bed nuclei of the stria terminalis to the periventricular zone. This output provides a route for the amygdala to influence circuits within the hypothalamus critical for expression of goal-oriented (ingestive, sexual, aggressive, and defensive) behaviors.

Second, is the projection to the basal ganglia, which includes topographically organized inputs to the dorsal and ventral striatum, and parts of the pallidum. Within the striatum, the amygdala innervates parts of the nucleus accumbens, fundus striatum, olfactory tubercle, and the entire caudoputamen. Interestingly, most amygdalar nuclei projecting to the caudoputamen innervate specifically its medial ('limbic') region, while only one amygdalar cell group (anterior basolateral nucleus) reaches its dorsolateral (somatomotor) part. Via its striatal projections, the amygdala can influence both limbic and somato-motor processing within the caudoputamen and reward-related processing within the nucleus accumbens, as well as learning associated with both of these. In contrast, its pallidal projections end specifically in distinct regions (substantia innominata and bed nuclei of the stria terminalis) which are associated with autonomic responses.

Third, is the projection to the medial and lateral prefrontal cortical areas and to the mediodorsal

thalamic nucleus. These projections, mainly from the basolateral nuclei, enable the amygdala to influence prefrontal cortex processing, most commonly linked to learning, working memory, and decision-making.

Fourth, are projections to the brainstem sensory and motor areas, originating primarily in the central and medial amygdalar nuclei. The central nucleus is also the recipient of major sensory information arriving from the brainstem. These efferents provide a 'feedback' to the sensory areas in the brainstem, and allow the amygdala direct access to the autonomic areas necessary for the expression of emotional and motivational responses.

Fifth, the amygdala projects back to the cortical sensory areas from which it receives inputs, thus providing a route for a feedback on sensory processing. In addition, it sends substantial, topographic inputs to various regions of the hippocampal formation, which are critical for learning and memory.

Clearly, the amygdala is well placed for a critical role in the stimulus-reinforcement type of associative learning, it receives sensory information from fundamentally all sensory modalities, and in turn, is in a position to influence a variety of motor systems via its wide ranging efferents. Furthermore, it is increasingly clear that distinct functional subsystems exist within the amygdala, and they each have their unique set of projections. An understanding of the underlying principles of the functional and anatomical organization of the amygdala is one of the most challenging questions that the field is facing today.

4. *Models of Amygdalar Organization*

In the 1980s and 1990s two models of amygdalar organization have been proposed. The first model suggests that part of the amygdala (central and medial nuclei) together with the substantia innominata and the bed nuclei of the stria terminalis, form a structural and functional unit—'the extended amygdala.' This hypothesis is based on cytoarchitectural, histochemical, and connectional similarities between the three parts of the continuum (De Olmos and Heimer 1999). Although this model fosters better understanding of the organization of central and medial parts, it does not provide a place for the rest of the amygdala.

Another more comprehensive model of amygdalar organization was proposed based on current embryological, neurotransmitter, connectional and functional data. This model argues that the amygdala is neither a structural nor a functional unit, but rather an arbitrarily defined collection of cell groups in the cerebral hemispheres, originally based on cytoarchitectonics. This suggests that it is more useful to place the various amygdalar cell groups within the context of the major divisions of the cerebral hemispheres—cortex and basal ganglia—and then to define the topographical organization of functionally defined systems within these divisions. Thus, various parts of the amygdala can be classified as belonging to one of the three distinct telencephalic groups: caudal olfactory cortex, specialized ventromedial extension of the striatum, or ventromedial expanse of the claustrum. Functionally they belong to the olfactory, autonomic, or fronto-temporal cortical systems, respectively (Swanson and Petrovich 1998).

Future research hopefully will delineate how the dynamics of information flow through the different amygdalar subsystems contribute to its different functions.

See also: Emotion, Neural Basis of; Fear: Potentiation of Startle; Fear: Psychological and Neural Aspects; Motivation, Neural Basis of; Reinforcement: Neurochemical Substrates

Bibliography

Aggleton J P (ed.) 1992 *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*. Wiley-Liss, New York

De Olmos J S, Heimer L 1999 The concepts of the ventral striatopallidal system and extended amygdala. *Annals of New York Academy of Science* **877**: 1–32

Kaada B R 1972 Stimulation and regional ablation of the amygdaloid complex with reference to functional representations. In: Eleftheriou B E (ed.) *The Neurobiology of the Amygdala*. Plenum Press, New York

Johnston J B 1923 Further contributions to the study of the evolution of the forebrain. *Journal of Comparative Neurology* **35**: 337–482

Klüver H, Bucy P C 1939 Preliminary analysis of functions of the temporal lobes in monkeys. *Archives of Neurological Psychiatry* **42**: 979–1000

Swanson L W, Petrovich G D 1998 What is the amygdala? *Trends in Neurosciences* **21**: 323–31

Weiskrantz L 1956 Behavioral changes associated with the ablation of the amygdaloid complex in monkeys. *Journal of Comparative Physiological Psychology* **49**: 381–91

G. D. Petrovich

Analysis of Variance and Generalized Linear Models

Analysis of variance (ANOVA) models are models that exploit the grouping structure in a set of data and lend themselves to the examination of main effects and interactions. These models are often referred to as being hierarchical in that it makes no sense to test whether main effects can be dropped from a model that includes that factor in an interaction, and it makes no sense to test whether a lower order interaction can be dropped from a model that includes any higher order interaction involving all of the same factors. The

modeling ideas also extend to generalized linear models (GLIMs).

1. One-way Analysis of Variance

One-way ANOVA models identify groups of observations using a separate mean for each group. For example, the observations can be the age of suicide categorized by mutually exclusive ethnic groups. Suppose there are a groups with N_i observations in each group. The jth observation in the ith group is written y_{ij} with expected value $E(y_{ij}) = \mu_i$. The one-way analysis of variance model is

$$y_{ij} = \mu_i + \varepsilon_{ij} \tag{1}$$

$i = 1, \ldots, a$, $j = 1, \ldots, N_i$ where the ε_{ij}s are unobservable random errors with mean 0, often assumed to be independent, normally distributed with variance σ^2, written $\varepsilon_{ij} \sim N(0, \sigma^2)$.

The ANOVA is a procedure for testing whether the groups have the same mean values μ_i. It involves obtaining two statistics. First, the mean squared error (MSE) is an estimate of the variance σ^2 of the individual observations. Second, the mean squared groups ($MSGrps$) is an estimate of σ^2 when the means μ_i are all the same and an estimate of σ^2 plus a positive number when the μ_i are different. The ratio of these two statistics, $MSGrps/MSE$ should be approximately 1 if the μ_is are all the same and tends to be larger than 1 if the μ_is are not all the same. Under the normality assumptions for the ε_{ij}s, when the μ_is are all the same the exact value of $MSGrps/MSE$ is random and follows an F distribution with $a-1$ degrees of freedom in the numerator and $n-a$ degrees of freedom in the denominator. Here $n = N_1 + \cdots + N_a$ is the total number of observations in the data. If the observed value of $MSGrps/MSE$ is so much larger than 1 as to be relatively inconsistent with it coming from the F distribution, one concludes that the μ_is must not all be the same.

In the special balanced case where $N_i = N$ for all i, the computations are particularly intuitive as analyzing variances. To find the MSE, simply find the sample variance within each group, i.e., compute the sample variance from $y_{i1}, \ldots, y_{iN}$, then average these a numbers to get MSE. This provides an estimate of σ^2. To get $MSGrps$, first compute the sample mean for each group, say $\bar{y}_i$. Then compute the sample variance of the $\bar{y}_i$s and multiply by N to get $MSGrps$. $MSGrps$ estimates $\sigma^2 + Ns_\mu^2$ where s_μ^2 is a new parameter consisting of the sample variance of the unknown parameters $\mu_1, \ldots, \mu_a$. The μ_is are all equal if and only if $s_\mu^2 = 0$. $MSGrps/MSE$ estimates $1 + Ns_\mu^2/\sigma^2$.

Terminology varies considerably for these concepts. The mean squared error is also called the mean squared residual and the mean squared within (groups). The mean squared groups is also called the mean squared treatments (because often the groups are identified as different treatments in an experiment) and the mean squared between (groups).

An alternative but equivalent model used with one-way ANOVA is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{2}$$

$i = 1, \ldots, a$, $j = 1, \ldots, N_i$. Here μ is a grand mean and the α_is are differential effects for the groups. The problem with this model is that it is overparameterized, i.e., the μ and α_i parameters are not identifiable. Even if you know the μ_is, there is an infinite number of ways to define the μ and α_i parameters. In fact, you can arbitrarily pick a value for any one of the μ or α_i parameters and still make them agree with any set of μ_is. Without including extraneous side conditions that have nothing to do with the model, it is impossible to estimate any of the μ and α_i parameters. It is, however, possible to estimate some functions of them, like values $\mu + \alpha_i$, or contrasts among the α_is like $\alpha_1 - \alpha_2$. Linear functions of the μ and α_i parameters for which linear unbiased estimates exist are called estimable functions. See also *Statistical Identification and Estimability*.

The F test can also be viewed as testing the full one-way ANOVA model against the reduced model

$$y_{ij} = \mu + \varepsilon_{ij}.$$

The reduced model can be viewed as either dropping the subscript i from μ_i in model (1) or as dropping the α_is from model (2). In either case, the reduced model does not allow for separate group effects. The F statistic comes from the error terms of the two models

$$F = \frac{SSE(\text{Red.}) - SSE(\text{Full})}{dfE(\text{Red.}) - dfE(\text{Full})} \Big/ MSE(\text{Full}) = \frac{MSGrps}{MSE}.$$

Extensions of ANOVA to more general situations depend crucially on the idea of testing full models against reduced models; see *Linear Hypothesis*.

2. Two-way Analysis of Variance

Two-way ANOVA can be thought of as a special case of one-way ANOVA in which the groups have a two-factor structure that we want to exploit in the analysis. For example, Everitt (1977) discusses 97 ten-year-old school children who were cross-classified by two factors, first, the risk of their home environment: not at risk (N) or at risk (R), and then the adversity of their school conditions: low, medium, or high. This defines six groups, each a combination of a home environment and a school condition. To illustrate the modeling concepts, suppose the dependent variable y

is the score on a test of verbal abilities. In general, we would write a model

$$y_{gk} = \mu_g + \varepsilon_{gk} \tag{3}$$

where $g = 1, \ldots, G$ indicates the different groups and $k = 1, \ldots, N_g$ indicates the observations within the group. In the specific example, $G = 6$ and $(N_1, N_2, N_3, N_4, N_5, N_6) = (17, 8, 18, 42, 6, 6)$.

To use the two-factor structure, we begin by identifying the factors mumerically. Let $i = 1,2$ indicate home environment: not at risk ($i = 1$), at risk ($i = 2$). Let $j = 1, 2, 3$ identify school adversity: low ($j = 1$), medium ($j = 2$), high ($j = 3$). In general, let $i = 1, \ldots, a$ denote the levels of the first factor and $j = 1, \ldots, b$ denote levels of the second factor. Without changing anything of substance in the model, we can rewrite the ANOVA model (3) as

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \tag{4}$$

$i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, N_{ij}$. All we have done is replace the single index for the groups g with two numbers ij that are used to identify the groups. This is still a one-way ANOVA model and can be used to generate an F test of whether all the μ_{ij}s are equal. In our example, $a = 2$, $b = 3$, and $(N_{11}, N_{21}, N_{12}, N_{22}, N_{13}, N_{23}) = (17, 8, 18, 42, 6, 6)$. Note that with a two factor model the number of groups is $G = ab$ and $E(y_{ijk}) = \mu_{ij}$.

The one-way ANOVA model (4) is often rewritten in an overparameterized version that includes numerous unidentifiable parameters. The overparameterized model is called the two-way ANOVA with interaction model, and is written

$$y_{ijk} = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij} + \varepsilon_{ijk}. \tag{5}$$

Here μ is a grand mean, the α_is are differential effects for the first factor, the η_js are differential effects for the second factor, and the $(\alpha\eta)_{ij}$s are called interaction effects. In reality, all of the μ, α_i, and η_j parameters are extraneous. If we drop all of them, we get the model $y_{ijk} = (\alpha\eta)_{ij} + \varepsilon_{ijk}$, which is just the one-way ANOVA model (4) with the μ_{ij}s relabeled as $(\alpha\eta)_{ij}$s. In model (5), without the introduction of side conditions that have nothing to do with the model, it is impossible to estimate or conduct any test on any function of the parameters that does not involve the $(\alpha\eta)_{ij}$ parameters. Since main effect parameters are extraneous when interactions are in the model, computer programs that purport to give tests for main effects after fitting interactions are really testing some arcane function of the interaction parameters that is determined by a choice of side conditions used by the program.

The interesting aspect of the two-way ANOVA with interaction model is that it suggests looking at the two-way ANOVA without interaction model

Table 1
Mean verbal test scores, additive model

		j (School)		
	μ_{ij}	1 (Low)	2 (Medium)	3 (High)
i	1 (N)	110	105	95
(Home)	2 (R)	100	95	85

Table 2
Mean verbal test scores, interaction model

		j (School)		
	μ_{ij}	1 (Low)	2 (Medium)	3 (High)
i	1 (N)	110	105	95
(Home)	2 (R)	100	95	80

$$y_{ijk} = \mu + \alpha_i + \eta_j + \varepsilon_{ijk} \tag{6}$$

This model is not equivalent to the one-way ANOVA model, but it includes nontrivial group effects. It amounts to imposing a restriction on the μ_{ij}s that

$$\mu_{ij} = \mu + \alpha_i + \eta_j \tag{7}$$

for some μ, α_is, and η_js. This indicates that the group effects μ_{ij} have a special structure in which the group effect is the sum of a grand mean and differential effects for each factor, hence model (6) is referred to as an additive model. The two-way ANOVA without interaction model is still overparameterized, so none of the individual parameters are estimable, but typically contrasts in the α_is and η_js are estimable as well as values $\mu + \alpha_i + \eta_j$.

To illustrate the modeling concepts, suppose that in our example the group means for the verbal test scores take the values shown in Table 1. These μ_{ij}s satisfy the additive model (7) so they do not display interaction. For example, take $\mu = 0$, $\alpha_1 = 110$, $\alpha_2 = 100$, $\eta_1 = 0$, $\eta_2 = -5$, $\eta_3 = -15$. Note that μ, the α_is, and the η_js are not identifiable (estimable) because there is more than one way to define them that is consistent with the μ_{ij}s. For example, we can alternatively take $\mu = 100$, $\alpha_1 = 5$, $\alpha_2 = -5$, $\eta_1 = 5$, $\eta_2 = 0$, $\eta_3 = -10$. However, identifiable functions of the parameters include $\mu + \alpha_i + \eta_j$, $\alpha_1 - \alpha_2$, $\eta_1 - \eta_3$, and $\eta_1 + \eta_2 - 2\eta_3$. Identifiable functions take on the same values for any valid choices of μ, the α_is, and the η_js. For example, the effect of changing from a not at risk home situation to an at risk home situation is a decrease of $\alpha_1 - \alpha_2 = 10$ points in mean verbal test score. Similarly, the effect of changing from a low school adversity situation to a high school adversity situation is a decrease of $\eta_1 - \eta_3 = 15$ points in mean verbal test score.

The beauty of the additive model (6) is that there is one number that describes the mean difference between students having not at risk home status and students having at risk home status. The difference is a drop of 10 points *regardless* of the school adversity status. This number can legitimately be described as the effect of going from not at risk to at risk. (Note that unless the students were assigned randomly to their home conditions, this does not imply that changing a student's status from at risk to not at risk will *cause*, on average, a 10 point gain.) Similarly, the effect of school adversity does not change with the home status. For example, going from low to high school adversity induces a 15 point drop regardless of whether the students are at risk or not at risk. Statistically, tests for main effects are tests of whether any of the α_is are different from each other and whether any of the η_js are different from each other, in other words whether the home statuses are actually associated with different mean verbal test scores and similarly for the school adversities.

The existence of interaction is simply any structure to the μ_{ij}s that cannot be written as $\mu_{ij} = \mu + \alpha_i + \eta_j$ for some μ, α_is, and η_js. For example, consider Table 2. In this case, the relative effect of having at risk home status for low or medium adversity schools is a drop of 10 points in mean test score, however for highly adverse schools, the effect of an at risk home status is a drop of 15 points. Unlike cases where the additive model (7) holds, the effect of at risk home status depends on the level of school adversity, so there is no one number that can characterize the effect of at risk home status. It makes no sense to consider the effect of home conditions without specifying the school adversity. Similarly, the effects of school adversity change depending on home status. For example, going from low to high school adversity induces a 15 point drop for students whose homes are not at risk, but a 20 point drop for at risk students. Again, there is no one number than can characterize the change from low to high school adversity, so there is no point in considering this change without specifying the home risk status. One moral of this discussion is that when interaction exists, there is no point in looking at tests of main effects, because main effects are essentially meaningless. If there is no one number that describes the effect of changing from not at risk to at risk, what could $\alpha_1 - \alpha_2$ possibly mean?

In practice, analyses such as these are conducted on estimates of the μ_{ij}s, and the estimates are subject to variability.

A primary use of model (6) is to test whether this special group effects structure fits the data. Model (6) is used as a reduced model and is tested against the full model (5) that includes interaction. This test is referred to as a test of interaction. In particular, the interaction test is not a test of whether the $(\alpha\eta)_{ij}$s are all zero, it is a test of whether every possible definition of the $(\alpha\eta)_{ij}$s must be consistent with the two-way without interaction model. The easiest way to think about this correctly is to think about testing whether one can simply drop the $(\alpha\eta)_{ij}$s from model (5).

If the two-way without interaction model (6) fits the data, it is interesting to see if any special case (reduced model) also fits the data. Two obvious choices are fitting a model that drops the second factor (school) effects

$$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \tag{8}$$

and a model that drops the first factor (home) effects

$$y_{ijk} = \mu + \eta_j + \varepsilon_{ijk} \tag{9}$$

If model (8) fits the data, there is no evidence that the second factor helps explain the data over and above what the first factor explains. Similarly, if (9) fits the data, there is no evidence that the first factor helps explain the data over and above what the second factor explains.

Given model (8) with only the first factor effect, we can evaluate whether the data fit a reduced model without that effect,

$$y_{ijk} = \mu + \varepsilon_{ijk} \tag{10}$$

This model comparison provides a test of whether the first factor is important in explaining the data when the second factor is ignored. Similarly we can start with model (9) and compare it to model (10). If model (10) fits, neither factor helps explain the data.

Through fitting a series of models, we can test whether there is evidence for the interaction effects, test whether there is evidence for the second factor effects η_j when the first factor α_i effects are included in the model, test whether there is evidence for the second factor effects η_j when the first factor α_i effects are not included in the model, and perform two similar tests for the importance of the α_i effects. In the special case when the N_{ij}s are all the same, the test for α effects including ηs, that is, the test of model (6) vs. model (9), and the test for α effects ignoring ηs, that is, the test of model (8) vs. model (10), are identical. Similarly, the two tests for η effects are identical. While this identity greatly simplifies the analysis, it only occurs in special cases and is not generally applicable. Moreover, it does not extend to generalized linear models such as log-linear models and logistic regression. The appropriate way to think about these issues is in terms of model comparisons.

3. *Higher-order Analysis of Variance*

The groups in a one-way ANOVA can also result from combining the levels of three or more factors. Consider a three-way cross-classification of Everitt's 97 students by further classifying them into students displaying or not displaying deviant classroom behavior. This deter-

mines $G = 12$ groups. The one-way ANOVA model $y_{gm} = \mu_g + \varepsilon_{gm}$, $g = 1, \ldots, G$, $m = 1, \ldots, N_g$ can be rewritten for three factors as $y_{ijkm} = \mu_{ijk} + \varepsilon_{ijkm}$ where the three factors are identified by $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$ so that $G = abc$, the group sample sizes are N_{ijk}, and $\mathrm{E}(y_{ijkm}) = \mu_{ijk}$. In the example, we continue to use i and j to indicate home and school conditions, respectively, and now use $k = 1, 2$ to indicate classroom behavior with $k = 1$ being nondeviant.

In its most overparameterized form, a three-way ANOVA model is

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + (\eta\gamma)_{jk} + (\alpha\eta\gamma)_{ijk} + \varepsilon_{ijkm} \tag{11}$$

This includes a grand mean μ, main effects for each factor α_is, η_js, γ_ks, two-factor interactions $(\alpha\eta)_{ij}$s, $(\alpha\gamma)_{ik}$s, $(\eta\gamma)_{jk}$s, and a three-factor interaction $(\alpha\eta\gamma)_{ijk}$s. In model (11), dropping the redundant terms, that is, everything except the three-factor interaction terms, gives a one-way ANOVA model $y_{ijkm} = (\alpha\eta\gamma)_{ijk} + \varepsilon_{ijkm}$. The primary interest in model (11) is that it suggests a wealth of reduced models to consider.

The first order of business is to test whether the three factor interaction is necessary, i.e., test the full model (11) against the reduced model without the three-factor interaction

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + (\eta\gamma)_{jk} + \varepsilon_{ijkm} \tag{12}$$

Focusing on only the important parameters, model (12) is equivalent to $y_{ijkm} = (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + (\eta\gamma)_{jk} + \varepsilon_{ijkm}$. In fact, even this version of the model is overparameterized. Model (11) is equivalent to the one-way ANOVA model, so if the three-factor interaction test is significant, there is no simplifying structure to the treatments. Probably the best one can do is to think of the problem as a one-way ANOVA and draw whatever conclusions are possible from the groups.

If model (12) fits the data, we can seek simpler models or try to interpret model (12). Conditioning on the levels of one factor, simplifies interpretations. For example, if $i = 1$, $y_{1jkm} = \mu + \alpha_1 + \eta_j + \gamma_k + (\alpha\eta)_{1j} + (\alpha\gamma)_{1k} + (\eta\gamma)_{jk} + \varepsilon_{1jkm}$ or $y_{1jkm} = [\mu + \alpha_1] + [\eta_j + (\alpha\eta)_{1j}] + [\gamma_k + (\alpha\gamma)_{1k}] + (\eta\gamma)_{jk} + \varepsilon_{1jkm}$. This is simply a two-factor model with interaction. If we change to $i = 2$, we again get a two-factor interaction model, one that has different main effects, but has the same interaction terms as when $i = 1$. For example, in Table 3 μ_{ijk}s satisfy model (12), that is, $\mu_{ijk} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + (\eta\gamma)_{jk}$ for some definitions of the parameters.

The key point in Table 3 is that, regardless of school adversity, the relative effect of not being at risk is always one point higher for nondeviants than for deviants. Thus, for high adversity the nondeviant home difference is 101–99 which is one point higher than the deviant home difference 99–98.

Table 3
No three-factor interacton

μ_{ijk}		Adversity of School (j)					
		Low		Medium		High	
	Home (i)	N	R	N	R	N	R
Classroom	Nondeviant	101	100	101	99	101	99
Behavior (k)	Deviant	100	100	100	99	99	98

Table 4
Two two-factor interactions

μ_{ijk}		Adversity of School (j)					
		Low		Medium		High	
	Home (i)	N	R	N	R	N	R
Classroom	Nondeviant	101	100	101	99	100	97
Behavior (k)	Deviant	100	98	100	97	99	95

Table 5
Rearrangement of Table 4

μ_{ijk}		Home (i)					
		N			R		
	School (j)	L	M	H	L	M	H
Classroom	Nondeviant	101	101	100	100	99	97
Behavior (k)	Deviant	100	100	99	98	97	95

If model (12) is adequate, the next step is to identify which of the two factor interaction terms are important. For example, we can drop out the $(\eta\gamma)_{jk}$ term to get

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + \varepsilon_{ijkm} \tag{13}$$

Dropping unimportant parameters, this is equivalent to $y_{ijkm} = (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + \varepsilon_{ijkm}$. To interpret model (13), condition on the factor that exists in both interaction terms. In our example, this is home conditions. The resulting model for $i = 1$ is a no interaction model $y_{1jkm} = [\mu + \alpha_1] + [\eta_j + (\alpha\eta)_{1j}] + [\gamma_k + (\alpha\gamma)_{1k}] + \varepsilon_{1jkm}$ with a similar no interaction model for $i = 2$ but with different main effects. For example, suppose the m_{ij}s are given in Table 4. Rearranging Table 4 to group together categories with i fixed gives Table 5. For fixed home conditions, there is no interaction. For not at risk students, deviant behavior is associated with a one point drop and high school adversity is associated with a one point drop. For at risk students, deviant behavior is associated with a 2 point drop and, relative to low school adversity, medium and high adversities are associated with a one point and a three point drop, respectively.

Table 6
One two-factor interaction

μ_{ijk}		Adversity of School (j)					
		Low		Medium		High	
	Home (i)	N	R	N	R	N	R
Classroom	Nondeviant	101	101	101	100	101	100
Behavior (k)	Deviant	100	100	100	99	100	99

Alternatively, we could eliminate both the $(\alpha\gamma)_{ik}$s and $(\eta\gamma)_{jk}$s from model (12) to get

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + \varepsilon_{ijkm} \tag{14}$$

which is equivalent to $y_{ijkm} = \gamma_k + (\alpha\eta)_{ij} + \varepsilon_{ijkm}$. We can think of this as a model for no interaction in a two-factor analysis in which one factor is indicated by the pair ij and the other factor is indicated by k. In the example, this means there is a main effect for classroom behavior plus an effect for each of the six combinations of home and school conditions. In particular, consider Table 6. Here, deviant behavior is always associated with a one point drop but there is interaction between home and school. For low school adversity, there is no effect of home conditions but for medium or high adversity, being at risk is associated with a one point drop.

From model (14L50), we could then fit a model that involves only the main effects

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + \varepsilon_{ijkm} \tag{15}$$

Model (15) is equivalent to $y_{ijkm} = \alpha_i + \eta_j + \gamma_k + \varepsilon_{ijkm}$. If model (15) is adequate, we could consider models that successively drop out the α_is, the η_js, and the γ_ks.

Fitting this sequence of successively smaller models, (11) then (12) then (13) then (14) then (15), then three models with the main effects successively dropped out, provides one tool for analyzing the data. There are 36 different orders possible for sequentially dropping out the two-factor interactions and then the main effects. In the balanced case of $N_{ijk} = N$ for all i, j, k, the order of dropping the effects does not matter, so one can construct an ANOVA table to examine each of the individual sets of effects, that is, main effects, two-factor interactions, three-factor interaction. For unbalanced cases, one would need 36 different ANOVA tables, one for each sequence, so ANOVA tables are almost never examined except in balanced cases. (Two-factor models only generate two distinct ANOVA tables). For unbalanced cases, instead of looking at the ANOVA tables, it is more convenient to simply report the *SSE* and *dfE* for all of the relevant models using a notation that identifies models by only their important parameters. For four or more factors, the number of potentially interesting models becomes too large to evaluate all of them.

After deciding on one or more appropriate models, the μ_{ijk}s can be estimated subject to the model and the models interpreted subject to the variability of the estimates.

To interpret a three-factor interaction, recall that just as a two-factor interaction exists when the effect of one factor changes depending on the level of the other factor, one can think of a three-factor interaction as a two-factor interaction that changes depending on the level of the third factor. However, it may be more productive to think of three-factor interaction in a three-factor model as simply specifying a one-way ANOVA. As with two-factor models, it makes little sense to test the main effects of a factor when that factor is involved in an important interaction. Similarly, it makes little sense to test that a lower-order interaction can be dropped from a model that includes a higher-order interaction involving all of the same factors.

A commonly used generalization of the ANOVA models is to allow some of the parameters to be unobservable random variables (random effects) rather than fixed unknown parameters. See *Hierarchical Models: Random and Fixed Effects*.

4. *Generalized Linear Models (GLIM)*

ANOVA is best suited for analyzing normally distributed data. These are measurement data for which the random observations are symmetrically distributed about their mean values. Generalized linear models (GLIMs) use similar linear structures to analyze other kinds of data, such as count data and time to event data.

We can rethink two-way ANOVA models as independent observations y_{ijk} normally distributed with mean m_{ij} and variance σ^2, write $y_{ijk} \sim N(m_{ij}, \sigma^2)$. The interaction model (5) has

$$m_{ij} = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij}$$

The no interaction model (6) has

$$m_{ij} = \mu + \alpha_i + \eta_j$$

Now consider data that are counts y_{ij} of some event. Assume the counts are independent Poisson random variables with means $\mathrm{E}(y_{ij}) = m_{ij}$. The use of two subscripts indicates that the data are classified by two factors. For Poisson data the m_{ij}s must be nonnegative, so $\log(m_{ij})$ can be both positive and negative. Linear models naturally allow both positive and negative mean values, so it is natural to model the $\log(m_{ij})$s with linear models. For example, we might use the log-linear model

$$\log(m_{ij}) = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij}$$

or, without interaction,

$$\log(m_{ij}) = \mu + \alpha_i + \eta_j$$

These log-linear models are also appropriate for multinomial data and independent groups of multinomial data.

Similarly, if the data y_{ij} are independent binomials with N_{ij} trials and probability of success p_{ij}, then we can analyze the proportions $\hat{p}_{ij} \equiv y_{ij}/N_{ij}$ with $\mathrm{E}(\hat{p}_{ij}) = m_{ij} \equiv p_{ij}$. Probabilities are defined to be between 0 and 1, the odds $p_{ij}/(1-p_{ij})$ take positive values, and the log-odds can take any positive or negative value. It is natural to write linear models for the log-odds such as

$$\log[p_{ij}/(1-p_{ij})] = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij}.$$

or, without interaction,

$$\log[p_{ij}/(1-p_{ij})] = \mu + \alpha_i + \eta_j$$

More generally, in GLIMs we create linear models for $g(m_{ij})$ using any strictly monotone function $g(\cdot)$, called a link function. The linear structures used for one-way ANOVA, two-way ANOVA, and higher-order ANOVA models all apply to GLIMs, and appropriate models can be found by comparing the fit of full and reduced models similar to methods for unbalanced ANOVA. In particular, cross-classified tables of counts often involve many factors.

We now examine in more depth ANOVA-type models for binomial data. Reconsider Everitt's 97 children cross-classified by the risk of their home environment and adversity of their school conditions. Rather than studying their verbal test score performance conditional on their classroom behavior as we did when discussing three-factor ANO we now examine models for whether the children display deviant or nondeviant behavior. In particular, we model the probability of a student falling into the deviant behavior category given their membership in the six home–school groups, write the probability of deviant behavior when in the *ij* group as p_{ij} and model the log-odds with analysis of variance type models,

$$\log[p_{ij}/(1-p_{ij})] = \mu_{ij}$$

The overparameterized model

$$\log[p_{ij}/(1-p_{ij})] = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij}$$

is equivalent to the original model. Neither of these models really accomplishes anything because they fit a separate parameter to every home–school category, so the models place no restrictions on the observations.

Table 7
Log-odds : additive model

		j (School)		
	μ_{ij}	1 (Low)	2 (Medium)	3 (High)
i	1 (N)	−2	−1	1
(Home)	2 (R)	0	1	3

Table 8
Log-odds : nonadditive model

		j (School)		
	μ_{ij}	1 (Low)	2 (Medium)	3 (High)
i	1 (N)	−2	−1	1
(Home)	2 (R)	−1	1	4

The additive model

$$\log[p_{ij}/(1-p_{ij})] = \mu + \alpha_i + \eta_j$$

constitutes a real restriction on the parameters.

Suppose the log-odds, the μ_{ij}s, satisfy Table 7. These have the structure of an additive model. In this case the log-odds for deviant behavior are two larger for at risk homes than for not at risk homes, e.g., $2 = 0-(-2) = 1-(-1) = 3-1$. This means that the odds are $e^2 = 7.4$ times larger for at risk homes. In particular, the log-odds of deviant behavior for low adversity schools and not at risk homes is -2 so the odds are $O_{11} = e^{-2} = 0.135$ and the probability if $p_{11} = O_{11}/[1+O_{11}] = 0.12$. Similarly, the odds and probability of deviant behavior for low adversity schools and at risk homes are $O_{21} = e^0 = 1$ and $p_{21} = 0.5$. The change in odds is $0.135 \times 7.4 = 1$. The 7.4 fold increase in odds is the same for all school conditions. Similarly, comparing low to high school adversities, the difference in log-odds is $3 = 1-(-2) = 3-0$, so the odds of deviant behavior are $e^3 = 20$ times greater in highly adverse schools, regardless of the home situation. The odds of deviant behavior in highly adverse schools and not at risk homes is $O_{13} = e^1 = 2.7$ which is 20 times greater than $O_{11} = 0.135$, the odds for low adversity schools and not at risk homes.

An example of a nonadditive model has log-odds as shown in Table 8. For low adversity schools the log-odds only differ by 1, for medium adversity schools they differ by 2, and for high adversity schools they differ by 3. Thus the effect of home conditions on the log-odds depends on the level of school adversity.

5. *Conclusions*

The fundamental ANOVA model is the one-way model. It specifies different mean values for different groups. When the groups are identified as combi-

nations of two or more factors, models incorporating main effects and interactions become a useful device for examining the underlying structure of the data. Appropriate models can be identified by fitting sequences of successively smaller models and using general testing procedures to identify the models within each sequence that fit well. Having identified a model, it is important to interpret what that model suggests about the underlying data structure. ANOVA models, the sequential model fitting procedures, and the interpretations apply to both balanced and unbalanced data and to generalized linear models. The sequential model fitting simplifies in balanced ANOVA allowing for it to be summarized in an ANOVA table.

See also: Multivariate Analysis: Discrete Variables (Logistic Regression); Simultaneous Equation Estimates (Exact and Approximate), Distribution of

Bibliography

Agresti A 1990 *Categorical Data Analysis*. Wiley, New York

Christensen R 1996a *Plane Answers to Complex Questions: The Theory of Linear Models*, 2nd edn. Springer-Verlag, New York

Christensen R 1996b *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Chapman and Hall, London

Christensen R 1997 *Log-linear Models and Logistic Regression*, 2nd edn. Springer-Verlag, New York

Everitt B S 1977 *The Analysis of Contingency Tables*. Chapman and Hall, London

Fienberg S E 1980 *The Analysis of Cross-classified Categorical Data*, 2nd edn. MIT Press, Cambridge, MA

Graybill F A 1976 *Theory and Application of the Linear Model*. Duxbury Press, North Scituate, MA

Hosmer D W, Lemeshow S 1989 *Applied Logistic Regression*. Wiley, New York

Lee Y, Nelder J A 1996 Hierarchical generalized linear models, with discussion. *Journal of the Royal Statistical Society, Series B* **58**: 619–56

McCullagh P, Nelder J A 1989 *Generalized Linear Models*, 2nd edn. Chapman and Hall, London

Nelder J A 1977 A reformation of linear models. *Journal of the Royal Statistical Society, Series A* **140**: 48–63

Scheffé H 1959 *The Analysis of Variance*. Wiley, New York

Searle S R 1971 *Linear Models*. Wiley, New York

Seber G A F 1977 *Linear Regression Analysis*. Wiley, New York

R. Christensen

Analytic Induction

Analytic induction (AI) is a research logic used to collect data, to develop analysis, and to organize the presentation of research findings. Its formal objective is causal explanation, a specification of the individually necessary and jointly sufficient conditions for the emergence of some part of social life. AI calls for the progressive redefinition of the phenomenon to be explained (the explanandum) and of explanatory factors (the explanans), such that a perfect (sometimes called 'universal') relationship is maintained. Initial cases are inspected to locate common factors and provisional explanations. As new cases are examined and initial hypotheses are contradicted, the explanation is reworked in one or both of two ways. The definition of the explanandum may be redefined so that troublesome cases either become consistent with the explanans or are placed outside the scope of the inquiry; or the explanans may be revised so that all cases of the target phenomenon display the explanatory conditions. There is no methodological value in piling up confirming cases; the strategy is exclusively qualitative, seeking encounters with new varieties of data in order to force revisions that will make the analysis valid when applied to an increasingly diverse range of cases. The investigation continues until the researcher can no longer practically pursue negative cases.

1. The Methodology Applied

Originally understood as an alternative to statistical sampling methodologies, 'analytic induction' was coined by Znaniecki (1934), who, through analogies to methods in chemistry and physics, touted AI as a more 'scientific' approach to causal explanation than 'enumerative induction' that produces probabilistic statements about relationships. After a strong but sympathetic critique by Turner (1953), AI shed the promise of producing laws of causal determinism that would permit prediction. The methodology subsequently became diffused as a common strategy for analyzing qualitative data in ethnographic research. AI is now practiced in accordance with Znaniecki's earlier (1928), less famous call for a phenomenologically grounded sociology. It primarily continues as a way to develop explanations of the interactional processes through which people develop homogeneously experienced, distinctive forms of social action.

The pioneering AI studies centered on turning points in personal biographies, most often the phase of commitment to behavior patterns socially defined as deviant, such as opiate addiction (Lindesmith 1968), embezzlement (Cressey 1953), marijuana use (Becker 1953), conversion to a millenarian religious sect (Lofland and Stark 1965), abortion seeking (Manning 1971), and youthful theft (West 1978, a rare study focusing more on desistance than onset). Studies at the end of the twentieth century have addressed more situationally specific and morally neutral phenomena. These include occupational perspectives exercised in

particular work settings (Katz 1982 on lawyers, Strong 1988 on doctors, and Johnson 1998 on union representatives), and distinctive moments in the course of everyday life (Flaherty 1999 the experience of time as passing slowly, and Katz 1999 laughter in a funhouse).

There is no particular analytical scale to the phenomena that may be addressed with AI. The research problem may be macro social events such as revolutionary social movements, mid-scale phenomena such as ongoing ways of being a student in a given type of educational institution, or everyday microsocial phenomena such as expressive gestures that can only be seen clearly when videotape is repeatedly reviewed.

1.1 How AI Transforms Theory

AI transforms and produces a sociological appreciation of phenomena along recurrent lines. The explanandum is often initially defined as a discrete act or event, for example the ingestion of a drug, exceeding a specific tenure on a job, or the commission of a fatal blow. The target phenomenon is progressively redefined to address a process: a persistent commitment, for example, being addicted to a drug; the maintenance of a perspective, such as a way of being involved with the challenges of a job; or a phase of personal change, as in the emotional transformation experienced when becoming enraged. Explanatory conditions, often originally defined from the outside as biographical and ecological background factors, are redefined to specify the interactions through which people, by learning, recognizing, or becoming aware of features of their pasts and circumstances, in effect set up the motivational dynamics of their own conduct. The methodology of AI thus dovetails with the theoretical perspective of symbolic interaction (SI) (Manning 1982; see *Symbolic Interaction: Methodology*), which stipulates that a person's actions are built up and evolve over time through processes of learning, trial-and-error, and adjustment to responses by others.

Although authors do not necessarily present their findings in these categories, a common theoretical result of AI is to highlight each of three types of explanatory mechanisms. One points to the practicalities of action (e.g., learning distinctive techniques for smoking marihuana). A second relates to matters of self-awareness and self-regard (e.g., attributing physical discomfort to withdrawal from opiates). The third refers to the sensual base of motivation in desires, emotions, or a sense of compulsion to act (e.g., embezzling when feeling pressured to resolve a secret financial problem). Perhaps the most ambitious long-term objective of AI is to develop the most economical set of inquiries capable of unveiling the distinctive processes that constitute any experienced moment in social life.

1.2 As Used in Ethnography

In the 1950s, reports of AI studies often took the form of tracing how negative cases led, step-by-step, to the final state of the theory. By 1980 an AI study was more likely to be presented in the style of an ethnographic text. Ethnographers find the principles of AI useful for guiding data gathering and shaping analysis. Often starting as outsiders and typically concerned to document social reality as lived by members, ethnographers redefine categories toward homogeneity 'from the inside.' When a sense of redundancy develops in interviews and observations, they commonly seek unusual data that will implicitly serve as negative cases when explicit analysis begins. Ethnographers are better positioned to trace how social life develops than to control rival variables as a means to argue why particular types of actions occur. In turn, they gravitate away from predictive theory and toward documenting regularities in the evolution of significant forms of behavior.

The methodological strategy of AI also dovetails with ethnography's narrative style. Because the logic of proof in AI relies solely on the richness or variety of cases that have been shown to be consistent with the final explanation, not on counting confirming cases, the researcher demonstrates the evidentiary strength of the theory by showing how variations of the explanans, A, B and C (A^{1-n}, etc.) fit with instances of the explanandum, X (X^{1-n}). Similarly, a common format for ethnographic writing is to entitle an article or chapter 'the career of …,' or 'doing …,' or 'becoming a …,' and then, in separate subsections, to describe the various ways that each of the explanatory conditions and the resulting phenomenon take shape. The author implicitly lays out the 'coding' or interpretive procedures that have been applied to the data set. A separate section may be devoted to cases of desistance, or transitions to non-X. For example, in a study of drivers becoming angry, a variety of cases describes how experiences of being 'cut off' emerge, how drivers come to see themselves in asymmetrical relations with other drivers, how they mobilize for revenge; and how each of these explanatory conditions is negated in cases showing anger subsiding (Katz 1999).

2. Limitations and Advantages

The logic of AI implies ideal conditions for data gathering that are rarely satisfied. The researcher should not be committed to a preset or conventional definition of the explanandum. Funding sources, however, are usually motivated by problems as defined by popular culture and as documented by official statistics. The researcher should constantly alter the data search as analysis develops. The practicalities of ethnographic projects, however, often press toward a

less flexible involvement in the field. Data should track the emergence and decline of the explanandum, the data should remain constant through repeated inspections, and there should be an inexhaustible series of instances against which to test hypotheses. Such data may be created through unobtrusive video-recordings of situated action, but the range of phenomena that can be described contemporaneously through phases of emergence and decline, in situ, in massive number, and without reactivity either during original recording or infinite reinspection, is severely limited.

2.1 Critiques and Rejoinders

Even so, an appreciation of how AI could exploit ideal evidence helps in assessing the central criticisms that have been addressed to it. One frequently cited weakness is that AI only specifies necessary but not sufficient conditions. Another is that it produces tautological explanations. If indeed the researcher only looks for factors common in the etiology of X, narrowing definitions and shedding cases when encountering negative cases, the explanation may only specify preconditions that are necessary but not particularly distinctive to X, much less sufficient to cause it. Cressey's claim that a 'nonshareable problem' and 'rationalization' explained a specified form of embezzlement was especially vulnerable on these grounds, especially since he never described the situated action of embezzlement. But if, as has often been the case, the researcher finds data describing transition points from non-X to X, as well as data describing progressions from X to non-X (e.g., desistance studies), claims of sufficiency may be precisely tested.

As to tautology, when the explanatory conditions are social-psychological matters of interactive behavior, as opposed to psychological and internal matters of thought and outlook, they can be coded independently of X. Note that any true causal explanation of behavior should turn up some potentially tautological cases. The very idea of causal sufficiency is that, with no existential break, the simultaneous presence of A, B, and C instantly produces X; in some cases there should not be any evidence that permits the coding of A, B, and C independent of data describing X. It should be expected that in some cases the development of the explanatory factors would be depicted as continuous with the emergence of the target phenomenon. But AI also leads the researcher to hunt for case histories in which, alternatively, each of A, B, or C had been absent and then came into existence, leading to X; as well as cases in which X had been present and then, alternatively, A, B, or C declined or ended, leading to non-X. AI is especially attuned to exploit contrasting states in temporal as opposed to cross-sectional data. The best examples of AI present evidence in just this sequential form, for example showing the development of addiction after an explicit and abrupt recognition that a long-standing pattern of distress has been due to repeated opiate withdrawal (Lindesmith 1968).

2.2 Generalization, Prediction and Retrodiction

Although AI studies, in order to allow the definition of explanatory elements to develop, cannot proceed from probabilistic samples and produce meaningful statistics attesting to representativeness, they are fundamentally geared toward generalization. By seeking negative cases, the researcher tests the explanation against claims that in times, places, and social circumstances other than those defining the initial collection, the explanation will not hold. As the explanation is redefined, it becomes both more nuanced and more wide-ranging in demonstrated validity. External validity depends on internal variety, not on the quantity and logically prederived uniformity of the data set. For this reason, an AI study reporting data that are monotonous, abstracted, and static will be methodologically weak.

AI cannot produce predictions in the sense of specifying the conditions at time 1 that will result in particular behaviors at time 2. The causal homogeneity that the method demands depends on subjects' defining their situation in common ways, and no study has ever found 'objective' conditions that will perfectly predict people's understandings of their biographical backgrounds and ecological contexts. The methodology does, however, support what might be called 'retrodiction': assertions that if a given behavior is observed to have occurred at time 2, specific phenomena will have occurred at time 1. While AI has never attempted to produce natural histories that specify the order of sequencing through which given social forms emerge (first condition A, then condition B, then condition C), it always makes claims of one or more, individually necessary and jointly sufficient preconditions of the explanandum. Thus, for example, AI will not attempt to predict who will become a murderer, but if one finds a case of enraged homicidal assault AI can support assertions as to what must have happened on the way to the assault: an emotional transformation from humiliation to rage, a recognition of being at a last stand for defending self-respect, and optimism about practical success (Katz 1988). Much of what is said to be valuable about prediction, such as the potential for intervention and control, remains available when results support retrodiction. Indeed, perhaps the most useful focus for policies of intervention is the identification of a narrowly defined precondition that is distinctive to a troublesome behavior, even if that single condition alone is not sufficient to cause the problem.

2.3 Unique Contributions

By redefining phenomena from the actor's perspective, and by discovering and testing an analysis of how given forms of social life come into existence, AI makes unique contributions that may be appreciated without gainsaying the contributions of statistical research. As it redefines the explanandum from a definition initially taken from conventional culture, AI typically reveals the social distance between insiders and outsiders, and the realities of culture conflict. Although rarely touted as 'policy research,' the upshot is a documented portrayal of some segment of social life that is systematically misrepresented by the culture that supports power. To the extent that social control, as influenced by populist voting and as implemented through officials' quotidian actions, is based on stereotypes about problematic behavior, AI can play a significant role in policy reform over the long term, especially if its ethnographic texts become widely used in university education. For scholars of cultural history and cultural differentiation, AI can document the changing variety of experiences in some area of social life (e.g., the variety of experiences in illicit drug use, professional acumen, or the behavior of laughter). Perhaps most generally, AI can specify the 'essence' of sociological phenomena in the sense of documenting what is entailed in a given line of action and form of social experience.

3. Prospects

Over its 75-year history, AI has shed a rhetorical claim to priority as the logic that should guide sociological data collection, metamorphosing into a pervasive, if typically implicit, strategy for analyzing qualitative data. Within the philosophy of science, the methodology's ill-considered claim of 'induction' has been replaced by a concept of 'retroduction,' or a double fitting of analysis and data collection (Ragin 1994). Similarly, it has been recognized that 'retrodiction' but not 'prediction' captures the thrust of AI's explanatory power.

The prospects for AI rest on three grounds. First, there is broad consensus that explanations, whether probabilistic or 'universal,' are likely to work better, the better they fit with subjects' perspectives. AI focuses most centrally on the foreground of social life, reaching into subjects' backgrounds to varying lengths, but always requiring careful examination of the contents of the targeted experience. It thus balances a relative indifference in much statistical research to the specific content of the explanandum, which is often left in such gross forms as 'serious' (FBI 'Part One') crime, or self-characterized 'violence.' Second, the findings of AI indicate that if social research imposes definitions on subjects regardless of the meaning that their conduct has to them, it will risk perpetuating artificial stereotypes and supporting power relations ill-suited to effective policy making. Finally, and most broadly, the utility of AI depends on two persistent features of sociological thinking. One is a fascination with the endless variety of distinctive forms into which people shape their social lives. The other is the observation that each subjectively distinctive stretch of personal experience comes with a tail of some biographical length. If one may never predict behavior with perfect confidence, still the forms that characterize small and large segments of lives are not superficial matters that emerge wholly made and with random spontaneity. Even as people constantly bootstrap the foundations for their conduct, they ground the objectives of their action in rich depths of temporal perspective; they act with detailed, often hard-won practical competence; and they consider the matter of how their conduct will appear to others with a care that is seasoned, even if it is routinely exercised in a split second of consequential behavior. AI's quest for systematic knowledge is no less secure than is the understanding that social life takes shape as people crystallize long-evolved perspectives, elaborate familiar behavioral techniques, and weave cultured interpersonal sensibilities into situationally responsive, experientially distinctive patterns of conduct.

See also: Symbolic Interaction: Methodology

Bibliography

Becker H S 1953 Becoming a marihuana user. *American Journal of Sociology* **59**: 235–42

Cressey D R 1953 *Other People's Money*. Free Press, Glencoe, IL

Flaherty M G 1999 *A Watched Pot: How We Experience Time*. New York University Press, New York

Johnson P 1998 Analytic induction. In: Symon G, Cassell C (eds.) *Qualitative Methods and Analysis in Organizational Research*. Sage, London

Katz J 1982 *Poor People's Lawyers in Transition*. Rutgers University Press, New Brunswick, NJ

Katz J 1988 *Seductions of Crime: Moral and Sensual Attractions in Doing Evil*. Basic, New York

Katz J 1999 *How Emotions Work*. University of Chicago Press, Chicago

Lindesmith A R 1968 *Addiction and Opiates*. Aldine, Chicago

Lofland J, Stark R 1965 Becoming a world-saver: A theory of conversion to a deviant perspective. *American Sociological Review* **30**: 862–75

Manning P K 1971 Fixing what you feared: Notes on the campus abortion search. In: Henslin J (ed.) *The Sociology of Sex*. Appleton-Century-Crofts, New York

Manning P K 1982 Analytic induction. In: Smith R B, Manning P K (eds.) *Handbook of Social Science Methods: Qualitative Methods*. Ballinger, Cambridge, MA

Ragin C C 1994 *Constructing Social Research*. Pine Forge Press, Thousand Oaks, CA

Strong P M 1988 Minor courtesies and macro structures. In: Drew P, Wootton A (eds.) *Erving Goffman: Exploring the Interaction Order*. Polity Press, Cambridge, UK

Turner R 1953 The quest for universals in sociological research. *American Sociological Review* **18**: 604–11
West W G 1978 The short term careers of serious thieves. *Canadian Journal of Criminology* **20**: 169–90
Znaniecki F 1928 Social research in criminology. *Sociology and Social Research* **12**: 302–22
Znaniecki F 1934 *The Method of Sociology*. Farrar and Rinehart, New York

J. Katz

Analytical Marxism

Analytical Marxism is a cross-disciplinary school of thought which attempts to creatively combine a keen interest in some of the central themes of the Marxist tradition and the resolute use of analytical tools more commonly associated with 'bourgeois' social science and philosophy.

Marx is often interpreted as having used a 'dialectical' mode of thought which he took over, with modifications, from Hegel, and which can be contrasted with conventional, 'analytical' thinking. Marx himself may have said and thought so. But according to analytical Marxists, one can make sense of his work, or at least of the main or best part of it, while complying with the strictest standards of analytical thought. This need not mean that all of what is usually called 'dialectics' should be abruptly dismissed. For example, the study of the typically dialectic contradictions which drive historical change constitutes a challenging area for the subtle use of analytical thought (see Elster 1978).

The resolute option for an analytical mode of thought does, however, imply that the research program stemming from Marx should by no means be conceived as the development of an alternative 'logic,' or of a fundamentally different way of thinking about capitalist development, or about social reality in general. It rather consists of practicing the most appropriate forms of standard analytical thought—using conventional conceptual analysis, formal logic and mathematics, econometric methods, and the other tools of statistical and historical research—in order to tackle the broad range of positive and normative issues broached in Marx's work. Analytical tools may have been developed and extensively used by 'bourgeois' social science and philosophy. This does not make them unfit, analytical Marxists believe, to rigorously rephrase and fruitfully develop some of the central tenets of the Marxian tradition.

Thus, the techniques characteristic of Anglo-American analytical philosophy can be used to clarify the meaning of key Marxian concepts as well as the epistemic status of the central propositions of the Marxian corpus and their logical relations with each other (see e.g., Cohen 1978). Formal models resting on the assumption of individually rational behavior, as instantiated by neoclassical economic theory and the theory of strategic games, can be used to understand the economic and political dynamics of capitalist societies (see e.g., Przeworski 1985, Carling 1991). The careful checking of theoretical conjectures against carefully collected and interpreted historical data can be used to test Marx's grand claims about transitions from one mode of production to another (see Aston & Philpin 1985). Making use of these tools does not amount to taking on board all the statements they have allegedly helped establish in the hands of conservative philosophers and social scientists. Nor is it meant to serve dogmatic defense of any particular claim Marx may have made. Rather, analytical Marxists view a competent, inventive, and critical use of these tools as an essential ingredient of any effective strategy both for refining and correcting Marx's claims and for challenging the political status quo.

The areas in which analytical Marxists have been active range from medieval history to socialist economics and from philosophical anthropology to empirical class analysis. The issues about which they have been involved in the liveliest controversies include:

(a) Are the central propositions of historical materialism to be construed as functional explanations, i.e., as explanations of institutions by references to the functions they perform? If so, are such explanations legitimate in the social as well as in the biological realm? (See Cohen 1978; Van Parijs 1981; Elster 1983.)

(b) Is it possible, indeed is it necessary, for a Marxist to be committed to methodological individualism, i.e. to the view that all social-scientific explanations should ultimately be phrased in terms of actions and thoughts by individual human beings? Or are there some admissible 'structuralist' Marxian explanations which are radically irreducible to an individualistic perspective? (See Elster 1982, Elster 1985; Roemer 1985.)

(c) Is there any way of salvaging from the ferocious criticisms to which it has been subjected the so-called theory of the falling rate of profit, i.e., Marx's celebrated claim that capitalist economies are doomed to be crisis-ridden, owing to a systematic tendency for the rate of profit to fall as a result of the very process of profit-driven capital accumulation? If the criticisms are compelling, what are the consequences both for the methodology of Marxian economics and for Marxian crisis theory? (See Roemer 1981, Elster 1985, Van Parijs 1993.)

(d) Can one vindicate the labor theory of value, i.e., the claim that the labor time required to produce a commodity is the ultimate determinant of its price, against the numerous objections raised against. If not, does this have any serious consequence for positive Marxian theory, bearing in mind that Marx himself considered this theory essential to the explanation of capitalist profits. Does it have any serious consequence for Marxian normative theory (if any), bearing in

mind that Marx's concept of exploitation is usually defined in terms of labor value? (See Roemer 1981, Cohen 1989.)

(e) Does Marx leave any room for ethical statements, i.e., statements about what a good or just or truly free society would be like, as distinct from statements about how a society could be more rationally organized or about what it will turn out to be by virtue of some inexorable laws of history? Or should one instead ascribe to him a consistently immoralist position and, if the latter, can such a position be defended? (See Wood 1981, Elster 1985, Cohen 1989.)

(f) Can the concept of exploitation—commonly defined as the extraction of surplus labor, or as the unequal exchange of labor value—be made independent of the shaky labor theory of value? Can it be extended to deal with late-capitalist or postcapitalist societies, in which the possession of a scarce skill, or the incumbency of some valued job, or the control over some organizational asset may be at least as consequential as the ownership of material means of production? Can such a more or less generalized concept of exploitation provide the basis for an empirically fruitful concept of social class? By providing a precise characterization of what counts as an injustice, can it supply the core of an ethically sensible conception of justice? (See esp. Roemer 1982, Wright 1985, Wright et al. 1989.)

(g) How can the Marxian commitment to equality (if any) be rigorously and defensibly formulated? Could the egalitarian imperative that motivates the demand for the socialization of the means of production also be satisfied by an equal distribution of the latter, or by a neutralization of the impact of their unequal distribution on the distribution of welfare? To what extent is this imperative compatible with every individual owning (in some sense) herself, as taken for granted, it would seem, in the typically Marxian idea that workers are entitled to the full fruits of their labor? (See Roemer 1994c, Cohen 1995.)

(h) After the collapse of East-European socialism, is it possible to reshape the socialist project in a way that takes full account of the many theoretical and practical objections that have been raised against it? Can a system be designed in which the social ownership of the means of production can be combined with the sort of allocative and dynamic efficiency commonly ascribed to capitalist labor and capital markets? (See Roemer 1994b.) Or should the radical alternative to capitalism as we know it rather be found in a 'capitalist transition towards communism' through the introduction and gradual increase of an unconditional basic income or in a highly egalitarian redistribution of privately owned assets? (See van der Veen and van Parijs 1986, Bowles and Gintis 1998.)

The boundaries of analytical Marxism are unavoidably fuzzy. Defined by the combination of a firm interest in some of the central themes of the Marxist tradition and the unhibited use of rigorous analytical tools, it extends far beyond, but is strongly associated with, the so-called 'September Group' founded in 1979 by the Canadian philosopher G. A. Cohen (Oxford) and the Norwegian social scientist Jon Elster (Columbia) The group has crystallized an attitude into a movement, by endowing it with a name, a focus, and a target for criticisms (see Roemer 1985, 1994a, Ware and Nielsen 1989). Having started with a critical inventory of Marx's heritage, it gradually took a more prospective turn, with a growing emphasis on the explicit elaboration and thorough defense of a radically egalitarian conception of social justice (see Cohen 1995, 1999, Van Parijs 1995, Roemer 1996) and a detailed multidisciplinary discussion of specific reforms (see Van Parijs 1992, Roemer 1998, and the volumes in the *Real Utopias Project* directed by Erik O. Wright). This development has arguably brought analytical Marxism considerably closer to left liberal social thought than to the bulk of explicitly Marxist thought.

See also: Communism; Equality: Philosophical Aspects; Market and Nonmarket Allocation; Marxism/Leninism; Socialism

Bibliography

Bowles S, Gintis H 1998 *Recasting Egalitarianism. New Rules for Communities, States and Markets*. Verso, New York

Aston T H, Philpin C H (eds.) 1985 *The Brenner Debate: Agrarian Class Structure and Economic Development in Preindustrial Europe*. Cambridge University Press, Cambridge, UK

Carling A H 1991 *Social Division*. Verso, London

Cohen G A 1978 *Karl Marx's Theory of History. A Defence*. Oxford University Press, Oxford, UK

Cohen G A 1989 *History, Justice and Freedom. Themes from Marx*. Oxford University Press, Oxford, UK

Cohen G A 1995 *Self-Ownership, Freedom and Equality*. Cambridge University Press, Cambridge, UK

Cohen G A 1999 *If You Are an Egalitarian, How Come You Are So Rich?* Harvard University Press, Cambridge, MA

Elster J 1978 *Logic and Society. Contradictions and Possible Worlds*. Wiley, Chichester, UK

Elster J 1982 Symposium on "Marxism, functionalism and game theory." *Theory and Society*. **11**: 453

Elster J 1983 *Explaining Technical Change*. Cambridge University Press, Cambridge, UK

Elster J 1985 *Making Sense of Marx*. Cambridge University Press, Cambridge, UK

Przeworski A 1985 *Capitalism and Social Democracy*. Cambridge University Press, Cambridge, UK

Roemer J E 1981 *Analytical Foundations of Marxian Economic Theory*. Cambridge University Press, Cambridge, UK

Roemer J E 1982 *A General Theory of Exploitation and Class*. Harvard University Press, Cambridge, MA

Roemer J E (ed.) 1985 *Analytical Marxism*. University of Calgary Press, Calgary, Canada

Roemer J E (ed.) 1994a *Foundations of Analytical Marxism*. E. Elgar, Aldershot, UK

Roemer J E 1994b *A Future for Socialism*. Harvard University Press, Cambridge, MA

Roemer J E 1994c *Egalitarian Perspectives. Essays in Philosophical Economics*. Cambridge University Press, Cambridge, UK
Roemer J E 1996 *Theories of Distributive Justice*. Harvard University Press, Cambridge, MA
Roemer J E 1998 *Equality of Opportunity*. Harvard University Press, Cambridge, MA
van der Veen R J, Van Parijs P 1986 Symposium on 'A Capitalist Road to Communism.' *Theory and Society* **15**(2): 635–655
Van Parijs P 1981 *Evolutionary Explanation in the Social Sciences. An Emerging Paradigm*. Rowman & Littlefield, Totowa, NJ
Van Parijs P (ed.) 1992 *Arguing for Basic Income. Ethical Foundations for a Radical Reform*. Verso, London
Van Parijs P 1993 *Marxism Recycled*. Cambridge University Press, Cambridge, UK
Van Parijs P 1995 *Real Freedom for All. What (if Anything) Can Justify Capitalism?* Oxford University Press, Oxford, UK
Ware R, Nielsen K (eds.) 1989 *Analyzing Marxism*. University of Calgary Press, Calgary, Canada
Wood A 1981 *Karl Marx*. Routledge & Kegan Paul, London
Wright E O 1979 *Class Structure and Income Determination*. Academic Press, New York
Wright E O 1985 *Classes: Methodological. Theoretical and Empirical Problems of Class Analysis*. New Left Books, London
Wright E O 1997 *Class Counts: Comparative Studies in Class Analysis*. Cambridge University Press, Cambridge, UK
Wright E O et al. 1989 *Debates on Classes*. Verso, London

P. van Parijs

Anaphora

The term 'anaphora' comes from the Greek word ἀναφόρα, which means 'carrying back.' In contemporary linguistics, there are three distinct senses of 'anaphora/anaphor/anaphoric': (a) as a relation between two linguistic elements, in which the interpretation of one (anaphor) is in some way determined by the interpretation of the other (antecedent), (b) as a noun phrase (NP) with the features [+anaphor, −pronominal] versus pronominal as an NP with the features [−anaphor, +pronominal] in the Chomskian tradition, and (c) as 'backward' versus cataphora/cataphor/cataphoric as 'forward,' both of which are endophora/endophor/endophoric, as opposed to exophora/exophor/exphoric.

Anaphora is at the center of early twenty-first century research on the interface between syntax, semantics, and pragmatics in theoretical linguistics. It is also a subject of key interest in psycho- and computational linguistics, and to work on the philosophy of language, and language in cognitive science. It has aroused this interest for a number of reasons. In the first place, anaphora represents one of the most complex phenomena of natural language, which, in itself, is the source of fascinating problems. Second, anaphora has long been regarded as one of the few 'extremely good probes' in furthering our understanding of the nature of the human mind, and thus in facilitating an answer to what Chomsky considers to be the fundamental problem of linguistics, namely, the logical problem of language acquisition—a special case of Plato's problem. In particular, certain aspects of anaphora have repeatedly been claimed by Chomsky to furnish evidence for the argument that human beings are born equipped with some internal, unconscious knowledge of language, known as the language faculty. Third, anaphora has been shown to interact with syntactic, semantic, and pragmatic factors. Consequently, it has provided a test bed for competing hypotheses concerning the relationship between syntax, semantics, and pragmatics in linguistic theory.

1. *Typologies of Anaphora*

Anaphora can be classified on the basis of (a) syntactic categories, (b) truth-conditions, and (c) discourse reference-tracking systems.

1.1 *Anaphora and Syntactic Categories*

In terms of syntactic category, anaphora falls into two main groups: (a) NP- (noun phrase-), including N- (noun), anaphora, and (b) VP- (verb phrase-) anaphora. In an NP-anaphoric relation, both the anaphor and its antecedent are in general NPs, and both are potentially referring expressions (1). NP-anaphora corresponds roughly to the semantically defined type of 'identity of reference' anaphora. NP-anaphora can be encoded by gaps (or empty categories), pronouns, reflexives, names, and descriptions. By contrast, in an N-anaphoric relation, both the anaphor and its antecedent are an $\bar{N}$ rather than an NP, and neither is a potentially referring expression (2). N-anaphora corresponds roughly to the semantically defined type of 'identity of sense' anaphora. Linguistic elements that can be used as an N-anaphor include gaps, pronouns and nouns.

(1) John said that he didn't know how to telnet.

(2) John's favorite painter is Gauguin, but Mary's Ø is van Gogh.

The other main category is VP-anaphora. Under this rubric, five types may be isolated: (a) VP-ellipsis, in which the VP of the second and subsequent clauses is reduced (3); (b) gapping, in which some element (typically a repeated, finite verb) of the second and subsequent conjuncts of a coordinate construction is dropped (4); (c) sluicing, which involves an elliptical construction consisting only of a wh-phrase (5); (d) stripping, an elliptical construction in which the ellipsis clause usually contains only one constituent (6); and

(e) null complement anaphora, an elliptical construction in which a VP complement of a verb is omitted (7).

(3) John adores Goya's paintings, and Steve does, too.

(4) Reading maketh a full man; conference a ready man; and writing an exact man.

(5) John donated something to Médecins Sans Frontières, but I don't know what.

(6) Pavarotti will sing 'Nessun dorma' again, but not in Hyde Park.

(7) Mary wanted to pilot a gondola, but his father didn't approve.

1.2 Anaphora and Truth-conditions

From a truth-conditional, semantic point of view, anaphora can be divided into five types: (a) referential anaphora, one that refers to some entity in the external world either directly or via its co-reference with its antecedent in the same sentence/discourse (1); (b) bound-variable anaphora, which is interpretable by virtue of its dependency on some quantificational expression in the same sentence/discourse, thus seeming to be the natural language counterpart of a bound variable in first-order logic (8); (c) E[vans]-type anaphora, one which, for technical reasons, is neither a pure referential anaphor nor a pure bound-variable anaphor, but which nevertheless constitutes a unified semantic type of its own (9); (d) anaphora of 'laziness,' so-called because it is neither a referential anaphor nor a bound-variable anaphor, but functions as a shorthand for a repetition of its antecedent (10); and (e) bridging cross-reference anaphora, one that is used to establish a link of association with some preceding expression in the same sentence/discourse via the addition of background assumption (11) (see Huang 2000a and references therein).

(8) Every little girl wishes that she could visit the land of Lilliput.

(9) Most people who bought a donkey have treated it well.

(10) The man who gave his paycheck to his wife was wiser than the man who gave it to his mistress.

(11) John walked into a library. The music reading room had just been refurbished.

1.3 Anaphora and Discourse: Reference-tracking Systems

Reference-tracking systems are mechanisms employed in individual languages to keep track of the various entities referred to in an ongoing discourse. In general, there are four major types of reference-tracking system in the world's languages: (a) gender/class, (b) switch-reference, (c) switch-function, and (d) inference.

In a gender system, an NP or (less frequently) a VP is morphologically classified for gender/class according to its inherent features, and is tracked through a discourse via its association with the gender/class assigned. This reference-tracking device is found to be present in a large variety of languages such as Archi, Swahili, and Yimas. Next, in a switch-reference system, the verb of a dependent clause is morphologically marked to indicate whether or not the subject of that clause is the same as the subject of its linearly adjacent, structurally related independent clause. Switch-reference is found in many of the indigenous languages spoken in North America, of the non-Austronesian languages spoken in Papua New Guinea, and of the Aboriginal languages spoken in Australia. It has also been found in languages spoken in North Asia and Africa. Somewhat related to the switch-reference system is the switch-function system. By switch-function is meant the mechanism that tracks the reference of an NP across clauses in a discourse by means of verbal morphology indicating the semantic function of that NP in each clause. This system is found in a wide range of languages including English, Dyirbal, and Jacaltec. Finally, there is the inference system. In this system, reference-tracking in discourse is characterized by (a) the heavy use of zero anaphors, (b) the frequent appeal to sociolinguistic conventions, and (c) the resorting to pragmatic inference. This mechanism is particularly common in East and Southeast Asian languages like Chinese, Javanese, and Tamil. Of the four reference-tracking devices mentioned above, the first is considered to be lexical in nature, the second and third, grammatical in nature, and the fourth, pragmatic in nature (Foley and Van Valin 1984, Comrie 1989, Huang 2000a).

2. Intrasentential Anaphora: Three Approaches

Anaphora can be intrasentential, that is, when the anaphor and its antecedent occur within a single sentence. It can also be discoursal, that is, when the anaphor and its antecedent cross sentence boundaries. In addition, there are anaphoric devices that lie in between pure intrasentential anaphora and pure discourse anaphora.

With regard to intrasentential anaphora, three main approaches can be identified: (a) syntactically oriented, (b) semantically oriented, and (c) pragmatically oriented. There has also been substantial research on the acquisition of intrasentential anaphora, especially within the generative framework, but I will not review it in this article.

2.1 Syntactic

Central to the syntactically oriented analyses is the belief that anaphora is largely a syntactic phenomenon, and as such references must be made to conditions and constraints that are essentially syn-

Table 1
Chomsky's typology of NPs

	Overt	Empty
a. [+anaphor, −pronominal]	reflexive/reciprocal	NP-trace
b. [−anaphor, +pronominal]	pronoun	pro
c. [+anaphor, +pronominal]	—	PRO
d. [−anaphor, −pronominal]	name	wh-trace

Table 2
Chomsky's binding conditions

A. An anaphor is bound in a local domain.
B. A pronominal is free in a local domain.
C. An r-expression is free.

Table 3
Anaphoric expressions

a. Handel_1 admired himself_1.
b. Handel_1 admired him_2.
c. Handel_1 admired Handel_2.

Table 4
Reinhart and Reuland's binding conditions

A. A reflexive-marked syntactic predicate is reflexive.
B. A reflexive semantic predicate is reflexive-marked.

Table 5
Reinhart and Reuland's typology of overt NPs

	SELF	SE	pronoun
Reflexivizing function	+	−	−
Referential independence	−	−	+

tactic in nature. This approach is best represented by Chomsky's (1981, 1995) binding theory within the principles-and-parameters theory and its minimalist descendant.

Chomsky distinguishes two types of abstract feature for NPs: anaphors and pronominals. An anaphor is a feature representation of an NP which must be referentially dependent and which must be bound within an appropriately defined minimal syntactic domain; a pronominal is a feature representation of an NP which may be referentially dependent but which must be free within such a domain. Interpreting anaphors and pronominals as two independent binary features, Chomsky hypothesizes that one ideally expects to find four types of NP in a language—both overt and non-overt—see Table 1. Of the four types of NP listed in Table 1, anaphors, pronominals, and r[eferential]-expressions are subject to binding conditions A, B, and C respectively (see Table 2). Binding is defined on anaphoric expressions in configurational terms, appealing to purely structural concepts like c-command, government, and locality. The binding theory accounts for the distribution of anaphoric expressions in Table 3. It is also applied to empty categories: NP-trace, pro, and wh-trace. It has even been extended to analyzing VP-ellipsis and switch-reference, but that work will not be surveyed here.

There are, however, considerable problems with the binding theory cross-linguistically. The distribution of reflexives violates binding condition A in both directions: on the one hand, a reflexive can be bound outside its local domain (so-called long-distance reflexive), as in Chinese, Icelandic and Tuki; on the other, it may not be bound within its local domain, as in Dutch and Norwegian. Binding condition B is also frustrated, for in many of the world's languages (such as Danish, Gumbaynggir, and Piedmontese), a pronominal can be happily bound in its local domain. Next, given Chomsky's formulation of binding conditions A and B, it is predicted that anaphors and pronominals be in strict complementary distribution, but this predicted complementarity is a generative syntactician's dream world. Even in a 'syntactic' language like English, it is not difficult to find syntactic environments where the complementarity breaks down. Finally, even a cursory inspection of languages like English, Japanese, and Vietnamese indicates that binding condition C cannot be taken as a primitive of grammar.

2.2 *Semantic*

In contrast to the 'geometric,' syntactic approach, the semantically based approach maintains that anaphora is essentially a semantic phenomenon. Consequently, it can be accounted for in semantic terms. Under this approach, binding is frequently defined in argument-structure terms. Reinhart and Reuland's (1993) theory of reflexivity, for example, belongs to this camp (see Tables 4 and 5).

On Reinhart and Reuland's view, reflexivity is not a property of NPs, but a property of predicates. The binding theory is designed not to capture the mirror-image distribution of anaphors and pronominals, but to regulate the domain of reflexivity for a predicate. Putting it more specifically, what the theory predicts is that if a predicate is lexically reflexive, it may not be reflexive-marked by a morphologically complex SELF anaphor in the overt syntax. And if a predicate is not

Table 6
Huang's revised neo-Gricean pragmatic apparatus (simplified)

(i) The use of an anaphoric expression x I-implicates a local coreferential interpretation, unless (ii) or (iii).
(ii) There is an anaphoric Q-scale ⟨ x, y ⟩, in which case, the use of y Q-implicates the complement of the I-implicature associated with the use of x, in terms of reference.
(iii) There is an anaphoric M-scale {x, y}, in which case, the use of y M-implicates the complement of the I-implicature associated with the use of x, in terms of either reference or expectedness.

lexically reflexive, it may become reflexive only via the marking of one of its co-arguments by the use of such an anaphor.

While Reinhart and Reuland's theory constitutes an important step forward in our understanding of binding, it is not without problems of its own. First, cross-linguistic evidence has been presented that marking of reflexivity is not limited to the two ways identified by Reinhart and Reuland. In addition to being marked lexically and syntactically, reflexivity can also be indicated morphologically. Second, more worrisome is that the central empirical prediction of the reflexivity analysis, namely, only a reflexive predicate can and must be reflexive-marked, is falsified in both directions. On the one hand, a predicate that is both syntactically and semantically reflexive can be non-reflexive marked; on the other, a non-reflexive predicate can be reflexive-marked.

2.3 Pragmatic

One of the encouraging progresses in the study of anaphora in the last decade has been the development of pragmatic approaches, the most influential of which is the neo-Gricean pragmatic theory of anaphora constructed by Levinson (1987, 2000) and Huang (1991, 1994, 2000a, 2000b). The central idea underlying this theory is that anaphora is largely pragmatic in nature, though the extent of anaphora being pragmatic varies typologically. Therefore, anaphora can largely be determined by the systematic interaction of some general neo-Gricean pragmatic principles such as Levinson's Q- (Don't say more than is required.), I- (Don't say less than is required.), and M- (Don't use a marked expression without reason.) principles, depending on the language user's knowledge of the range of options available in the grammar, and of the systematic use or avoidance of particular anaphoric expressions or structures on particular occasions. Table 6 is a revised neo-Gricean pragmatic apparatus for anaphora. Needless to say, any interpretation generated by Table 6 is subject to the general consistency constraints applicable to Gricean conversational implicatures. These constraints include world knowledge, contextual information, and semantic entailments.

There is substantial cross-linguistic evidence to show that empirically, the neo-Gricean pragmatic theory of anaphora is more adequate than both a syntactic and a semantic approach. Not only can it provide a satisfactory account of the classical binding patterns (Table 2), it can also accommodate elegantly some of the anaphoric patterns that have always embarrassed a generative analysis. Conceptually, this theory also has important implications for current thinking about universals, innateness and learnability.

3. Discourse Anaphora: Four Models

One of the central issues in discourse anaphora is concerned with the problem of anaphoric distribution, namely, how to account for the choice of a particular referential/anaphoric form at a particular point in discourse. For any entity to which reference is to be made, there is a (potentially large) set of possible anaphoric expressions each of which, by a correspondence test, is 'correct' and therefore could in principle be used to designate that entity. On any actual occasion of use, however, it is not the case that just any member of that set is 'right.' Therefore, an 'appropriate' anaphoric form from that set has to be selected from time to time during the dynamic course of discourse production, but what contributes to the speaker's choice of that form? Surely anaphoric distribution in discourse is a very complex phenomenon, involving, among other things, structural, cognitive, and pragmatic factors that interact with each other. Nevertheless, currently there are four main models of discourse anaphora: (a) topic continuity or distance-interference (Givón 1983), (b) hierarchy (Fox 1987), (c) cognitive (Tomlin 1987, Gundel et al. 1993), and (d) pragmatic (Huang 2000a, 2000b). In addition, there is a formal, Discourse Representation Theory model (Kamp and Reyle 1993), but how it can be applied to real discourse data is unclear.

3.1 Topic Continuity

The main premise of this model is that anaphoric encoding in discourse is essentially determined by topic continuity. The continuity of topic in discourse is measured primarily by factors such as linear distance, referential interference, and thematic information. Roughly, what the model predicts is this: the shorter the linear distance, the fewer the competing referents, and the more stable the thematic status of the protagonist, the more continuous a topic; the more continuous a topic, the more likely that it will be encoded in terms of a reduced anaphoric expression

like pronouns and zero anaphors. Some of these ideas have recently been further developed in centering theory (Walker et al. 1997).

3.2 Hierarchy

On this model, it is assumed that the most important factor that influences anaphoric selection is the hierarchical structure of discourse. From this assumption follows the central empirical prediction of the theory, namely, mentions (initial or non-initial) at the beginning or peak of a new discourse structural unit tend to be done by an NP, whereas subsequent mentions within the same discourse structural unit tend to be achieved by a reduced anaphoric expression.

3.3 Cognitive

The basic tenet of this model is that anaphoric choice in discourse is largely dictated by cognitive processes such as activation and attention. Its central empirical claim is that NPs are predicted to be used when the targeted referent is currently not addressee-activated, whereas reduced anaphoric expressions are predicted to be selected when such a referent is estimated to be currently both speaker- and addressee-activated. A number of hierarchies have been put forward in the literature to capture this cognitive status/anaphoric form correlation.

3.4 Pragmatic

The kernel idea behind the neo-Gricean pragmatic model is that anaphoric distribution in discourse, as in anaphoric distribution in the sentence, can also be largely determined by the systematic interaction of the Q-, M-, and I-principles, mentioned above. In Huang (2000a, 2000b), it has been demonstrated that by utilizing these principles and the resolution mechanism organizing their interaction, many patterns of discourse anaphora can be given a sound explanation. Furthermore, a careful consideration of anaphoric repair systems in conversational discourse in Chinese and English shows that the neo-Gricean pragmatic analysis is consistent with what interlocutors in conversational discourse are actually oriented to.

Clearly, there are at least three interacting factors that are at work in predicting anaphoric distribution in discourse: structural, cognitive, and pragmatic. Of these factors, the structural constraint (both linear and hierarchical) seems largely to be a secondary correlate of the more fundamental cognitive and/or pragmatic constraints. However, the interaction and division of labor between the cognitive and pragmatic constraints are not well understood and need to be further studied.

See also: Anaphora Resolution; Linguistics: Overview; Pragmatics: Linguistic; Semantic Knowledge: Neural Basis of; Semantics; Syntactic Aspects of Language, Neural Basis of; Syntax; Syntax—Semantics Interface

Bibliography

Chomsky N 1981 *Lectures on Government and Binding*. Foris, Dordrecht, The Netherlands

Chomsky N 1995 *The Minimalist Program*. MIT Press, Cambridge, MA

Comrie B 1989 Some general properties of reference-tracking systems. In: Arnold D et al. (eds.) *Essays on Grammatical Theory and Universal Grammar*. Oxford University Press, Oxford, UK, pp. 37–51

Foley W A, Van Valin R D 1984 *Functional Syntax and Universal Grammar*. Cambridge University Press, Cambridge, UK

Fox B A 1987 *Discourse Structure and Anaphora*. Cambridge University Press, Cambridge, UK

Givón T (ed.) 1983 *Topic Continuity in Discourse*. John Benjamins, Amsterdam

Gundel J, Hedberg N, Zacharski R 1993 Cognitive status and the form of referring expressions in discourse. *Language* **69**: 274–307

Huang Y 1991 A neo-Gricean pragmatic theory of anaphora. *Journal of Linguistics* **27**: 301–35

Huang Y 1994 *The Syntax and Pragmatics of Anaphora*. Cambridge University Press, Cambridge, UK

Huang Y 2000a *Anaphora: A Cross-linguistic Study*. Oxford University Press, Oxford, UK

Huang Y 2000b Discourse anaphora: Four theoretical models. *Journal of Pragmatics* **32**: 151–76

Kamp H, Reyle U 1993 *From Discourse to Logic*. Kluwer, Dordrecht, The Netherlands

Levinson S C 1987 Pragmatics and the grammar of anaphora. *Journal of Linguistics* **23**: 379–434

Levinson S C 2000 *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge, MA

Reinhart T, Reuland E 1993 Reflexivity. *Linguistic Inquiry* **24**: 657–720

Tomlin R 1987 Linguistic reflections of cognitive events. In: Tomlin R (ed.) *Coherence and Grounding in Discourse*. John Benjamins, Amsterdam, pp. 455–80

Walker M, Joshi A K, Prince E (eds.) 1997 *Centering Theory in Discourse*. Oxford University Press, Oxford, UK

Y. Huang

Anaphora Resolution

The term *anaphora* (which comes from a Greek root meaning 'to carry back') is used to describe situations in which there is repeated reference to the same thing in a text. Sentence (2) below contains three instances of anaphora.

(1) John noticed that a window had been left open.

(2) *He* walked over to *the window* and closed *it* firmly.

He, *the window*, and *it* mentioned in (2) refer back to the previous mentions of *John* and *a window* in (1). In

general, anaphors, like those in sentence (2), refer back to previously mentioned entities. However, anaphora can also occur with temporal or spatial reference. Temporal expressions, such as *then*, *the next day*, or *the week before*, often refer back to previously established times and spatial expressions, such as *there*, often refer back to previously mentioned locations. Thus, anaphora is an important linguistic device for establishing the coherence of an extended piece of discourse.

Anaphora resolution is the process of interpreting the link between the anaphor and the previous reference, its *antecedent*. It is especially interesting because it frequently involves interpretation across a sentence boundary. Hence, despite playing a crucial role in discourse comprehension, it falls outside the scope of traditional psycholinguistic accounts of sentence processing (Kintsch and van Dijk 1978). There are two main issues that have driven the psychological research on anaphora resolution. First, there is the issue of how the resolution process relates to more basic sentence comprehension processes, such as syntactic parsing and semantic interpretation. This issue has been addressed principally by investigating the time course of anaphora resolution as compared with other sentence interpretation processes. The second issue concerns the nature of the mental representation of the discourse context that is required to support anaphora resolution. In this article these two issues are examined in turn.

1. *The Time Course of Anaphora Resolution*

Anaphora enables readers and listeners to infer links or bridges between the different sentences in a text. Thus, one of the earliest accounts, developed by Haviland and Clark (1974), characterized it as a *bridging-inference* process. It was assumed that the anaphoric expression (i.e., pronoun or definite noun phrase) triggered a search through the previous text to find something that could link anaphor to antecedent. For the anaphors shown in (2) this can be done on the basis of simple matching, but it is often more complicated. Consider for instance, the slightly different text below.

(3) John noticed it was chilly in his bedroom.

(4) He walked over to *the window* and closed it firmly.

Here the reader needs to infer that *the window* in (4) is intended as an indirect anaphoric reference to part of the bedroom mentioned in (3). Thus, he or she will have to draw an inference, based on prior knowledge of windows and rooms, to link the two together. Self-paced reading experiments clearly demonstrate that readers spend time on such bridging operations during normal reading (e.g., see *Inferences in Discourse, Psychology of*).

The special nature of this anaphoric bridging process raises questions about how it relates to other more basic sentence interpretation processes (e.g., see *Sentence Comprehension, Psychology of*). One way of addressing the issue is by investigating the time course of anaphora resolution in relation to the other processes. Are anaphors resolved as soon as they are encountered or is resolution delayed until after a local analysis of the sentence has been established? Several methods have been developed to address this question.

1.1 *Methods for Measuring Anaphora Resolution*

Methods for investigating the time course of anaphora resolution tap into the process in two ways. One approach is to record precisely when an antecedent is reactivated after encountering an anaphor. The other approach sets out to determine how soon after encountering an anaphor the information associated with the antecedent becomes incorporated into the interpretation of the rest of the sentence. The two techniques produce rather different results and this has led to the suggestion that resolution occurs in two distinct stages. First, there is a process of antecedent identification and recovery, and, then at some later point there is a process of integration of the antecedent information into the interpretation of the current sentence.

The principal technique for establishing precisely when an antecedent is reactivated during processing uses *antecedent probe recognition*. A passage containing an antecedent is presented on a computer screen one word at a time. At predetermined points in the passage a probe word is then shown and readers have to judge whether or not it matches a previous word in the text. The difference in recognition times for probes presented before and after an anaphor give a measure of the degree to which the antecedent is reactivated on encountering the anaphor. By judicious placement of the probes it is therefore possible to measure the time course of anaphora resolution. In a classic study of this kind, Dell et al. (1983) used texts like the following:

A burglar surveyed the garage set back from the street.

Several milk bottles were piled at the curb.

The banker and her husband were on vacation.

The criminal/a cat slipped away from the street lamp.

At the critical point following either the anaphor *the criminal* or the nonanaphor *a cat*, they presented the probe word *burglar* for recognition. They found that recognition was enhanced immediately following *criminal* as compared with *cat*. They also obtained a similar effect for words drawn from the sentence in which the antecedent had appeared (e.g., *garage*). This finding, together with related findings from experiments using proper-name anaphors, indicate that antecedents are reactivated almost immediately after encountering the anaphor.

However, the antecedent probe recognition technique has not been so successful in demonstrating immediate reactivation of antecedents after reading a pronoun anaphor (Gernsbacher 1989). This led to the conclusion that the antecedent recovery process depends on the degree to which the anaphor specifies its antecedent in the discourse. Repeated proper names and repeated nouns directly match their antecedents whereas pronouns do not.

The second way of investigating the time course of resolution is to examine when antecedent information is incorporated into the interpretation of the sentence containing the anaphor. This approach has been used to investigate the interpretation of both spoken and written language. A recent example with written language is an experiment by Garrod et al. (1994) which used eye movement recording to tap into the comprehension process. They presented readers with passages such as the following:

A dangerous incident in the pool
Elizabeth$_1$ was an inexperienced swimmer and would not have gone in if *the male lifeguard*$_2$ had not been standing by the pool. But as soon as she got out of her depth she started to panic and wave her hands about in a frenzy.

(a) Within seconds (she$_1$/Elizabeth$_1$) sank into the pool.

(b) Within seconds (she$_1$/Elizabeth$_1$) jumped into the pool.

(c) Within seconds (he$_2$/the lifeguard$_2$) sank into the pool.

(d) Within seconds (he$_2$/the lifeguard$_2$) jumped into the pool.

The passages introduce two characters of different gender (i.e., *Elizabeth* and *the male lifeguard*), who are subsequently referred in the target sentences (a) to (d) using either a pronoun or a repeated noun. The crucial manipulation was with the verbs in the target sentences. Each verb was chosen to make sense only with respect to one of the antecedent characters. For example, whereas *Elizabeth* might be expected to *sink* at that point in the story, she could not *jump* because she is out of her depth, whereas the *lifeguard* might be expected to *jump*, he could not *sink* because he is standing on a solid surface. Thus sentences (a) and (d) describe contextually consistent events, but sentences (b) and (c) describe contextually anomalous events. The eye-tracking procedure makes it possible to measure the point in the sentence when the reader first detects these contextual anomalies by comparing the pattern of eye movements on matched contrasting materials (e.g., (a) vs. (b) or (d) vs. (c)).

The results from this and similar studies with spoken language produce a different pattern from the earlier reactivation experiments. They show that pronouns can be resolved at the earliest point in processing just so long as they are unambiguous and refer to the principal or topic character in the discourse at that point (see Sect. 2.2). Hence, readers immediately respond to the anomaly with sentence (b) as compared with sentence (a) above. In contrast, noun anaphors or pronouns that refer to other nontopic characters lead to delayed resolution. Hence readers only detect the anomaly in (c) as compared with (d) after they have read the whole clause. This discrepancy in the findings from the two techniques leads to the conclusion that anaphora resolution is a two-stage process. In the first stage, the anaphor triggers a search for a matching antecedent, as suggested originally by Clark, and this antecedent is recovered. However, it is only at a second later stage that the anaphoric information is fully incorporated into the interpretation of the sentence. The first recovery stage is affected by the accessibility of the antecedent and the degree to which it matches the anaphor, whereas the later integration stage is affected by the topicality of the antecedent at that point in the text (for a fuller discussion, see Garrod and Sanford 1994).

In summary, evidence on the time course of anaphora resolution, both from antecedent reactivation studies and from studies that tap into the consequences of resolution, indicate that it occurs in tandem with other more basic sentence comprehension processes. However, the time course experiments also indicate that different kinds of anaphor affect resolution in different ways. This leads to the other main issue associated with anaphora resolution, which concerns the nature of the discourse representation required to support the process.

2. *Anaphora Resolution and Discourse Representation*

Although anaphors relate to antecedent mentions, their interpretation is usually based on the antecedent's reference rather than its wording or its meaning. This means that anaphora resolution depends on access to a representation of the prior discourse that records these discourse referents: the people and other entities that the discourse has referred to up to that point.

The structure of the representation has to account for how referents become more or less accessible as the text unfolds. Discourse representations are also important for explaining indirect anaphora of the kind illustrated in sentences (3) and (4) above. These issues are considered in more detail below. First, there is a discussion of referential discourse models, then a discussion of accessibility of referents in terms of their degree of *focus*, and finally a discussion of what other kinds of information might be represented in the discourse model.

2.1 *Referential Discourse Models*

Typically anaphors are coreferential with their antecedents. That is, they are interpreted in terms of what the

antecedent refers to rather than its wording or meaning. For example consider the interpretation of *it* in sentence (6).

(5) Mary painted her front door.

(6) Jill painted *it* too.

If the wording of the antecedent *her front door* were substituted for the pronoun in (6), it would be Jill's door that Jill had painted. However, that is not the interpretation readers make. Rather the pronoun is taken to refer to the original referent (i.e., Mary's door). Furthermore, not all noun phrases introduce such referents. For example, the mention of *a teacher* in sentence (7) does not license an anaphoric reference to *the teacher* in (8), because the noun phrase does not introduce a new discourse referent.

(7) Harry was *a teacher*.

(8) *The teacher* loved history.

Thus, readers construct a representation that contains a record of the referents that have been introduced into the discourse up to that point. These discourse referents then serve as antecedents for anaphora. Such representations have been described as models of the discourse world or *discourse models* (e.g., see *Mental Models, Psychology of*).

Discourse models have been discussed in detail by various semantic theorists. For example, Kamp and Reyle (1993) developed a general framework to account for the circumstances under which sentences introduce discourse referents and the degree to which the referents become accessible for subsequent anaphora. It is called *discourse representation theory* and has had a strong influence on computational modeling of anaphora. Among other things the theory explains why *her front door* in (5) introduces a discourse referent, whereas *a teacher* in (7) does not. An important aspect of such models is how they reflect the degree to which antecedents shift in and out of *focus* as a text unfolds. To illustrate this point consider a slight variant of sentences (5) and (6):

(9) Mary painted *her front door*.

(10) She planned to paint the windows next and then the walls.

(11) Jill painted *it* too.

The use of the pronoun *it* in sentence (11) is infelicitous, despite the fact that the front door is the only antecedent discourse referent that matches in number and gender. Thus, discourse referents vary in their accessibility as antecedents for pronouns as the text unfolds. This phenomenon has been called *discourse focus* and has attracted interest both in psychology and computational linguistics.

2.2 Discourse Models and Focus

There are a variety of factors that affect the degree to which an antecedent is focused at any point in a piece of discourse. In general, the more text between the antecedent and the pronoun, the less focused, and hence less accessible, the antecedent becomes. However, several other factors play an important role in discourse focus, such as the number of alternative discourse referents that have been introduced since encountering the antecedent and whether the antecedent was originally introduced as a named principal protagonist in a narrative. There is also the issue of how the antecedent was referred to in the previous sentence. This last factor has been the subject of considerable theorizing in computational linguistics in the form of *centering theory* (Walker et al. 1998). According to the theory, a sentence projects candidate antecedents for pronominal reference in the following sentence and the candidates are ranked in terms of accessibility. High-ranking antecedents are appropriately referred to with pronouns whereas low-ranking antecedents require more explicit anaphors. The theory has been used to predict when a pronoun is more acceptable than a fuller anaphor and is consistent with the findings on the time course of resolution of pronouns as compared with fuller forms of anaphora discussed in Sect. 1.

For pronouns, anaphora resolution is influenced by discourse focus. However, with noun anaphors it is more strongly influenced by other aspects of the discourse representation. This is particular true with indirect anaphora, as when the writer refers to *the window* following mention of *a room* (e.g., see sentences (3) and (4) above).

2.3 Discourse Representation and Indirect Anaphora

The prototypical noun anaphor is a definite noun phrase, such as *the window*. However, studies of text corpora indicate that the majority of such references occur in contexts that do not contain explicitly introduced antecedents, as in sentences (3) and (4) above. Thus, they only serve as indirect anaphors.

In general, indirect anaphora takes longer to resolve than direct anaphora. However, there are certain circumstances under which it does not. For instance, indirect references to *the lawyer* in the context of a court case or to *the car* in the context of driving can be resolved as quickly as direct anaphora (Garrod and Sanford 1990). This leads to the question of whether there is more in the discourse representation than discourse referents alone. One suggestion is that the discourse model is extended to also include *discourse roles*, and that they can serve as antecedents in a similar fashion to discourse referents. Roles, such as *lawyer* or *vehicle*, are represented in certain accounts of knowledge representation as essential components of knowledge of court cases or driving (e.g., see *Mental Models, Psychology of*; *Schemas, Frames, and Scripts in Cognitive Psychology*). It is an open research question as to whether they also play a key role in indirect anaphora resolution.

See also: Inferences in Discourse, Psychology of; Mental Models, Psychology of; Sentence Comprehension, Psychology of

Bibliography

Dell G S, McKoon G, Ratcliffe R 1983 The activation of antecedent information during the processing of anaphoric reference in reading. *Journal of Verbal Learning and Verbal Behaviour* **22**: 121–32

Garrod S, Freudenthal D, Boyle E 1994 The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language* **33**: 39–68

Garrod S, Sanford A J 1990 Referential processes in reading: focusing on roles and individuals. In: Balota D A, Flores d'Arcais G B, Rayner K (eds.) *Comprehension Processes in Reading*. Erlbaum, Hillsdale, NJ, pp. 465–84

Garrod S, Sanford A J 1994 Resolving sentences in a discourse context: how discourse representation affects language understanding. In: Gernsbacher M (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA, pp. 675–98

Gernsbacher M A 1989 Mechanisms that improve referential access. *Cognition* **32**: 99–156

Haviland S E, Clark H H 1974 What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior* **13**: 512–21

Kamp H, Reyle U 1993 *From Discourse to Logic*. Kluwer, Dordrecht, The Netherlands

Kintsch W, van Dijk T A 1978 Toward a model of text comprehension and production. *Psychological Review* **85**: 363–94

Walker M A, Joshi A K, Prince E F (eds.) 1998 *Centering Theory in Discourse*. Oxford University Press, Oxford, UK

S. Garrod

Ancestors, Anthropology of

The study of beliefs and rites involving ancestors has long been a central anthropological concern. Many early anthropologists suggested that ancestral worship was the forerunner of all religions and devoted much time to building conjectural religious histories. Twentieth century anthropologists have devoted attention to the working of ancestral cults as aspects of wider religious systems and to their links to rights in property, succession to office, and cohesion of descent groups. The main questions asked include how the living define ancestors, transform the dead into them, and use them as guardians of morality and agents of social continuity.

The word 'ancestor' is often used as a synonym for 'the dead': but they should be distinguished—not all the dead are remembered and defined as ancestors. Those selected as ancestors provide a history memorializing certain events of the past that give significance to claims of the present. Defined ancestors may either actually have lived at a remembered time, may be given a place in a genealogy irrespective of the date of death, or may be invented so as to provide their believed descendants with history and identity.

The status of ancestor is given by a rite of transition performed by the living descendants of the deceased. Robert Hertz (1960) listed the three components of mortuary rites: the corpse, the mourners, and the soul. The functions of the rites are to dispose of the corpse; to remove the deceased from the immediate domain of the living and to transform him or her into an ancestor; to re-form the descent group of the deceased's living kin; and to ensure the continuity of the essential element of the deceased's human identity, the soul or some similar concept. The forms the rites take reflect the importance of the deceased: those for a king may be highly elaborate and last many months or years, for a respected elder be less so and take a few days, for a child be minimal and take only a few minutes (Huntington and Metcalf 1979, Bloch and Parry 1982).

The living decide which of the dead shall be transformed into ancestors. The key criterion is the link of kinship between a living group or person and an ancestor. Without that link the deceased is not an ancestor but merely a dead person and soon forgotten. To possess recognized ancestors gives the living descendants full identity and status; to lack them (as in the case of slaves in many societies) is to lack full status.

Several ancestral categories may be defined. For example, the Lugbara of Uganda (Middleton 1960) distinguish the *ori*, a descent group's direct and renowned male patrilineal forebears, to whom are made ritual offerings; *oku-ori*, female *ori*, daughters of the patriline who, although they bore sons to other descent groups (by the rule of exogamy), are remembered as powerful women; and *a'bi*, the general mass of other dead members of a given clan. Only relatively few of the dead become *ori* and *oku-ori*, whereas the number of *a'bi* is beyond reckoning. The ancestors of other clans are of no concern and are merely *a'bi* without definition.

Ancestral categories may also depend upon the manner of death. The Akan of Ghana, for example (Fortes 1959), distinguish three kinds of death: good death, the natural death of man or woman followed by a future in the land of the ancestors; bad death, due to evil behavior during life and a double death, first that in the living world and later within the land of the ancestors, whose inhabitants refuse to accept the deceased who is returned to the land of the living to linger as a bad and dangerous specter; and sudden death, as in war, childbirth, accident, or suicide—such a person is not buried properly in a clan cemetery but in the bush land and is then 'lost' as an ancestor.

Among the Chinese, as a third example, Confucianism makes a general distinction between the ances-

tors and the non-ancestral dead (Hsu 1948, Ahern 1973, Wolf 1974). The former are men of strictly patrilineal descent groups, senior kinsmen with clearly recognized rights due from their descendants and obligations towards them. Women are not defined as ancestors, although they have been daughters, wives, and mothers to the ancestors of a particular descent group.

These cases show that whether individuals are defined as ancestors depends upon their kinship ties with their descendants, linked by mortuary rites and by inheritance of social position and property. Ancestors possess no independent status in themselves but are given it by their living kin.

Once made into ancestors, where they dwell and the nature of their relationships to their descendants vary from one society to another. Christian cultures hold that claimed ancestors and other dead dwell either with the Deity in a heaven, a place of peace without further passing of time, and/or a hell, a place of punishment and atonement. The Akan peoples of Ghana hold that ancestors dwell in a land of the dead across a river, which has the same patterns of stratification and authority as that of the living. The Lugbara of Uganda consider that they dwell underground near their descendants' settlements, where they themselves have been buried. In virtually all religions ancestors are considered to be senior kin living apart from their descendants yet still linked by spatial propinquity and kinship interdependence. They are typically believed to see and hear their living kin, who as their juniors are unable to see or hear them except in unusual circumstances, such as their appearance as specters or when people are possessed by them.

Associated with the concept of ancestor is that of soul, linked to individual morality rather than to descent and kinship. The concept of soul was of great concern to early anthropologists such as Frazer (1911–15), Lévy-Bruhl (1927) and Tylor (1871), who coined the term 'animism' for the belief in souls attributed to all living things. Their views were essentially conjecture: more recently ethnographic research, including that on religion, has become based on detailed study of the beliefs and practices of particular cultures. For example, the Lugbara word *ori* is related to the word *orindi* (literally 'the essence of the *ori*') which might be translated as 'soul,' an element of the living that is not exterminated by death but which persists into the afterlife. Only men of full adult status have what might be called full souls; women who are the first-born of a set of siblings may be attributed souls if they acquire respected status before death, and it is they who become *oku-ori*. Other men and women lack souls or, more accurately, do not realize the potentiality of developing them, and it is they who are placed in the loose category of *a'bi*. Different societies order these concepts and elements in differing ways, but the basic factor is that of an individual's moral passage through life, the soul being that element that bestows him or her with morally responsible behavior.

Ancestors have to be remembered, usually by being listed in genealogies. Evans-Pritchard (1940) showed for the Nuer of the Sudan that the generational depth of genealogies reflects the width and composition of descent groups on the ground: the wider the group the more the number of generations needed to include all its members in a single genealogy. The identity of effective ancestors usually depends upon the recognition of a particular mode of descent (patrilineal, matrilineal, or cognatic) as the basis for local social organization. Genealogies are not, or only extremely rarely, historically accurate statements. There is typically a telescoping of generations so as to maintain the same depth of a genealogy over time; and in patrilineal systems, certainly, the exclusion of most women is usual as they do not exercise the same degree of authority as do men. The situation is different in those societies in which ancestors are memorialized in writing or on tablets (e.g. China and Mormon America); here the purpose may be to include all forebears by name irrespective of their immediate and active links to their descendants.

Terms commonly used in discussions of ancestors include 'ancestral spirits,' 'ancestral worship,' and 'ancestral cults,' that refer to the relationships between the living and their own ancestors. The beliefs and practices concerning ancestors nowhere form a distinct religion but rather a religious practice within a wider societal religion. This practice typically includes lineage and family rites, periodic rites on ancestral death days, and often annual rites for collectivities of a group's ancestors. Rites linked to the non-ancestral dead in general should not be included. Ancestors are virtually everywhere held to care for and protect their descendants; they may teach, criticize, and punish them for unfitting behavior. They may send sickness or other affliction to those who disobey their rules and wishes, and they are in general held to be conservative in their views and to represent an idealized stability and moral order. They may act on their own initiative or may be evoked by their descendants. The living may consult oracles or diviners, or interpret omens, to discover the identities and motives of specific ancestors concerned.

The living may communicate with their ancestors both by making verbal requests and by offering them food (often flesh and blood) and drink, commensality being a central act of kinship bonding. Offerings may be regular or made only when the ancestors show that they wish to receive them, at shrines of many kinds, from ancestral graves to special buildings, and may be simple and cursory or elaborate and long-lasting, depending on the ancestors' remembered status.

The term 'ancestral spirit,' used for those aspects of ancestors thought to come into direct contact with the living, is clumsy: ancestors and spirits are in their

natures quite distinct. Ancestors have been among the living, who remember the more recent of them as individuals and understand their wishes and demands. Even if their individual personalities are in time forgotten, they represent human experience, skills, and sentiments, and are senior kin. They are used by the living both to maintain conservative authority and moral order and to help bring about orderly settlement of disputes. They provide and maintain the long-lasting stability of otherwise ever-changing and fluid descent groups. On the other hand, spirits are refractions of divinity, have never been living and so have no descendants: their natures and demands are different and may be beyond human understanding. The living can exercise a more immediate control over their ancestors by stressing their link as kin, than they can over spirits. Ancestors may act as unvarying custodians of order, unlike spirits which are by contrast individualistic, capricious, and unpredictable.

An important point is that of the persistence of beliefs in ancestors in the 'modern' industrialized world. The rites of invocation and appeasement of ancestors may lose importance for small-scale descent groups precisely as these lose importance in radically changing societies; ancestors are those who, when they lived, were important in the lives and continuity of their descent groups of the time. By being remembered and commemorated in ritual, they are objects of the group's memory of the past and its identity of the present. Ancestors used as mnemonic figures remain important, as do their material representations such as tombs and shrines. This is especially so in the case of those who in their lives were kings, sacred personages who embodied the identity and continuity of their kingdoms. Such royal ancestors are not only used as signs of memory by their genealogies: they are typically given royal tombs that materially represent the past in the present and so act as centers of societal identity. In societies that accept Christianity, Islam, or other world religions, greater emphasis may be given to spirits: ancestors' powers are limited to their own descendants whereas those of spirits are typically held to affect all members of a society irrespective of their descent groups.

See also: Death, Anthropology of; Kinship in Anthropology; Religion: Evolution and Development; Religion: Family and Kinship; Ritual

Bibliography

Ahern E 1973 *The Cult of the Dead in a Chinese Village*. Stanford University Press, Stanford, CA

Bloch M, Parry J (eds.) 1982 *Death and the Regeneration of Life*. Cambridge University Press, Cambridge, UK

Evans-Pritchard E E 1940 *The Nuer*. Clarendon Press, Oxford, UK

Fortes M 1959 *Oedipus and Job in West African Religion*. Cambridge University Press, Cambridge, UK

Frazer J G 1911–15 *The Golden Bough*. Macmillan, London

Hertz R 1960 (1st edn. 1909) *Death, and the Right Hand*. Cohen and West, London

Hsu F L K 1948 *Under the Ancestors' Shadow*. Norton, New York

Huntington R, Metcalf R 1979 *Celebrations of Death*. Cambridge University Press, Cambridge, UK

Lévy-Bruhl L 1927 *L'Ame Primitive*. (English edn., *The Soul of the Primitive* 1928) Alcan, Paris

Middleton J 1960 *Lugbara Religion*. Oxford University Press, London: new edition 1999 James Currey, Oxford, UK

Tylor E B 1871 *Primitive Culture*. Murray, London

Wolf A 1974 *Religion and Ritual in Chinese Society*. Stanford University Press, Stanford, CA

J. Middleton

Androgyny

Androgyny is most simply defined as the combination of masculine and feminine characteristics within a single person. It achieved widespread popularity within gender psychology beginning in the early 1970s as part of a model for conceptualizing gender that uses familiar constructs of masculinity and femininity while avoiding prescriptive, sex-specific values inherent in earlier studies. In this article the conceptualization of androgyny within gender psychology, major research findings, and current status will be discussed.

1. The Concept of Androgyny

The idea of androgyny is an ancient one, expressed in mythology and literature centuries ago. Broadly speaking, androgyny denotes any blurring of distinctions between the sexes. In this sense there can be 'androgynous' persons in physical sex characteristics (hermaphroditism), sexual preference (bisexuality), unisex dress styles, or societies providing equal economic and political rights for the sexes. Social scientists have used the term more restrictively to describe an individual who manifests in either personality or behavior a balanced combination of characteristics typically labeled as masculine (associated with men) or feminine (with women) in our society.

The characteristics traditionally associated with each sex are diverse yet recognized easily by members of a particular society. A combination of Parsons and Bales' (1953) description of familial roles and Bakan's (1966) philosophical treatise on fundamental modalities of all living organisms has been adopted widely to distill the common themes inherent in feminine/masculine trait distinctions. The core of those psychological characteristics stereotypically associated with

women concerns sensitivity, selflessness, emotionality, and relationships with others (expressive/communal). In contrast, psychological characteristics stereotypically associated with men reflect goal orientation, self-development, assertiveness, and individuation (instrumental/agentic). Androgyny essentially represents a combination of these two themes within one person.

1.1 Origins in Masculinity–Femininity Research

The modernized concept of androgyny is founded in earlier twentieth-century ideas about the nature of psychological differences between the sexes, but endorses a much broader range of gender-related behavior for individuals regardless of biological sex. Traditional models of gender psychology viewed a clear differentiation between the sexes in a wide range of characteristics to be natural, typical, and desirable. Manifestation of 'masculine' attributes by men and 'feminine' attributes by women signified fulfillment of a basic genetic destiny. Preandrogyny studies reflected researchers' conviction about the existence of a single innate psychological trait differentiating the sexes, and focused on specifying the content universe of this masculinity–femininity trait. As Constantinople (1973) discussed in her landmark review, this trait was posited to be a continuum with femininity and masculinity serving as bipolar endpoints and logical reversals of one another. Researchers considered a given masculinity–femininity measure to be valid if it reliably clustered women's and men's responses into two distinct groups, regardless of the items' content. The final item pool was frequently a melange of content sharing only an ability to distinguish between the sexes. Endorsement of the 'masculine' pole by men and the 'feminine' pole by women was considered typical, and indicative of psychological health. Individuals who failed to exhibit attributes associated with their biological sex or who endorsed attributes typical of the other sex were suspected of being sexually confused, homosexual, and/or psychologically maladjusted.

This traditional model of gender differentiation did not explain the common occurrence of similarities across the sexes; items on which the sexes gave similar responses typically were deleted from these scales. Variability within each sex in endorsement of masculine and feminine items was also ignored in favor of highlighting mean differences between the sexes. Such inconsistencies became increasingly troublesome to researchers. Also with the mid-century resurgence of the feminist movement, the gender-related values codified within the traditional model increasingly appeared to be restrictive, outdated, and harmful to individuals. The alternative of androgyny quickly became what Mednick (1989) referred to as one of feminist psychology's conceptual bandwagons of the 1970s and early 1980s.

1.2 Assumptions Underlying Androgyny Theory and Research

Proponents of androgyny typically continued the earlier focus on masculinity and femininity as trait dimensions. In androgyny theory and research, however, a unique series of assumptions are operative. Femininity and masculinity are no longer seen as opposite ends of a single dimension in which being less feminine automatically meant being more masculine. In the conceptualization of androgyny, masculinity and femininity are portrayed as independent but not mutually exclusive groups of characteristics. Individuals can be meaningfully described by the degree to which they endorse each group as self-descriptive. One can be high on both, low on both, or high on only one. Both masculinity and femininity are also hypothesized to have a unique and positive impact on a person's psychological functioning. That is, both sexes presumably benefit from being 'feminine' and 'masculine' to some degree. Possession of high levels of both sets of characteristics, or androgyny, should thus represent the most desirable gender-relevant alternative. This model ingeniously provided researchers with a virtually limitless array of researchable hypotheses, a methodology consistent with psychology's positivist tradition, and an explicit values statement comfortably fitting with the era's Zeitgeist of expanded human rights and roles.

2. Research on Androgyny

Research on androgyny can be classified loosely according to purpose: (a) to develop androgyny measures fitting these newly formulated assumptions about femininity and masculinity; (b) to determine the meaningfulness of the masculinity and femininity dimensions as represented on the androgyny measures; and (c) to explore the implications of various combinations of these femininity and masculinity dimensions within an individual.

2.1 Development of Androgyny Measures

Development of psychometrically sound masculinity and femininity scales based on the revised assumptions was an important first step for androgyny researchers. The favored scale format was paper-and-pencil self-descriptions using Likert scales. Criteria for item selection were somewhat variable. Although a (small) number of measures were eventually developed, only two achieved prominence: The Bem Sex Role Inventory (BSRI) (Bem 1974) and the Personal Attributes Questionnaire (PAQ) (Spence and Helmreich 1978). The items on the BSRI and the PAQ reflected judges' ratings of personality characteristics utilizing criteria of sex-based social desirability or of sex

typicality, respectively. The PAQ incorporated only characteristics generally seen as desirable. The BSRI included some femininity items with less positive connotations (e.g., 'childlike,' 'gullible'), a decision that complicated subsequent analyses considerably (Pedhazur and Tetenbaum 1979).

Correlations between the masculinity and femininity scales of a single androgyny measure tended to be small in magnitude as was desired, and the content of corresponding scales across androgyny measures was overlapping but not identical. Factor analyses (e.g., Wilson and Cook 1984) indicated that the content of the femininity and masculinity scales corresponded generally to theoretical definitions of femininity as representing empathy, nurturance, and interpersonal sensitivity and masculinity as representing autonomy, dominance, and assertiveness. The emergence of this factor structure is interesting in that item selection procedures did not specifically select items to be congruent with the expressive/communal and instrumental/agentic distinctions. These content distinctions appear to be central to the broad-based perceptions of the sexes' personalities and behaviors elicited by the androgyny measures (Cook 1985).

2.2 Research on Masculinity and Femininity Scales within the Androgyny Measures

Research generally has shown that each scale is related as expected to variables linked to the instrumental/agentic and expressive/communal distinctions. However, masculinity scale correlations with measures of self-esteem and psychological adjustment have typically been stronger than femininity scale correlations for both men and women, depending on measures and samples used (Taylor and Hall 1982, Whitley 1983). This pattern of findings does not support the basic hypothesis about the equal value of both dimensions for both sexes. Explanations for this pattern have variously implicated the specific content included on the androgyny measures; the adequacy or appropriateness of criterion measures; a failure of researchers to assess negative aspects of each dimension that may offset its benefits; the greater valuing of masculine attributes in society; or the multiple determined nature of gendered phenomena. Each of these explanations is likely to have some merit. Researchers did generally agree that each dimension as operationalized on the androgyny measures may be beneficial for individuals in some respects.

2.3 Androgyny as Type: Theoretical Considerations

The most provocative studies within the androgyny literature addressed the implications of various levels of masculine and feminine characteristics within an individual. These studies were based on the assumption that the self-descriptions elicited by the androgyny measures are indicative of an enduring typological differentiation of individuals. Preandrogyny gender research generally classified two types of individuals, feminine or masculine, with pervasive consequences predicted from their internalization of either own sex-typed (masculine men and feminine women) or cross sex-typed characteristics (feminine men and masculine women). Sex typing was considered normative and good, and cross-sex typing as deviant and harmful. Individuals whose responses placed them in neither group were given little attention. Within the androgyny literature, it was hypothesized that both masculinity and femininity dimensions conferred certain benefits on men and women alike. An expanded typology was needed to acknowledge a portion of the population overlooked in preandrogyny masculinity–femininity studies, and to operationalize a new gender ideal.

Androgyny researchers exploring androgyny as type accepted that the pattern of self-descriptions elicited by androgyny measures corresponded to a meaningful typology of individuals composed of different blendings of feminine and masculine attributes. The manner in which masculinity and femininity might work together to produce androgyny was variously explained. Androgyny was proposed alternatively to mean the balancing or moderating of femininity and masculinity's excesses or deficits by the other dimension; a beneficial summation of the positive qualities of each dimension; the emergence of unique, albeit vaguely defined qualities from the synergy of the femininity and masculinity dimensions; or the elimination of sex-stereotypic standards for behavior in an individual's perceptions and decisions, thus making traditional, prescriptive masculine vs. feminine distinctions irrelevant to her or him.

Bem provided the most influential and elegant renderings of androgyny as a type of individual. Her original theory (1974) contrasted sex-typed and nonsex-typed persons, a focus that was reminiscent of the preandrogyny literature's bipolar classification of individuals. According to Bem, sex-typed individuals have internalized society's sex-appropriate standards for desirable behavior to the relative exclusion of the other sex's typical characteristics. This one-sided internalization has a negative impact on the sex-typed person's view of self and others, expectations and attitudes, and behaviors. In contrast, nonsex-typed individuals are free from the need to evaluate themselves and others consistent with prescriptive sex-linked standards, and thus are able to behave more adaptively and flexibly. In her later gender schema theory, Bem (1981) proposed that sex-typed individuals cognitively process incoming information in terms of culturally-based definitions of masculinity and femininity. These definitions are effectively internalized to function as cognitive schema shaping perceptions and subsequent behavior. Gender-related

connotations are not similarly salient for nonsex-typed persons, who are able to utilize other schema as appropriate. The implication is that freedom from reliance on gender schema is likely to be preferable in many situations.

2.4 Androgyny as Type: Scoring Methods

Conflicting views about the manner in which masculinity and femininity were believed to influence each other gave rise to controversy about the best way to derive the typological classification from the femininity and masculinity scale scores. Bem (1974) first proposed *t*-ratio scoring consistent with her primary distinction between sex-typed vs. nonsex-typed persons. In *t*-ratio scoring, androgyny is defined operationally as the lack of a statistically significant difference between femininity and masculinity scale scores. Sex-typed individuals do have statistically significant scale score differences. Researchers interested in the ramifications of a balance between masculine and feminine characteristics have tended to favor *t*-ratio scoring or some variant of it.

The second dominant view of androgyny is the additive view, in which androgyny is defined as the summation of the positive and basically independent influences of the femininity and masculinity dimensions. Whereas the balance view considers as androgynous those individuals endorsing roughly comparable levels of both dimensions at any level (low–low to high–high), in the additive view only high–high scorers are considered androgynous. Early research using the androgyny measures suggested the heuristic value of distinguishing between high–high and low–low scoring individuals. Spence and Helmreich (1978) proposed a median split of each masculinity and femininity scale score distribution to yield a four-way classification: androgynous, undifferentiated (low–low), and two sex-typed groups reporting a predominance of one set of characteristics. This scoring method was adopted quickly by researchers; Bem herself used some variation of median split scoring on occasion. Variations and alternatives to these two types of androgyny scoring procedures have been proposed but not adopted widely.

The issue of what scoring procedure is preferable has generated considerable controversy but little resolution. The best method to portray the conjoint influence of masculinity and femininity may depend on the specific nature of the hypotheses used. Unfortunately, the proliferation of scoring variations has contributed to confusion in the field. Rarely have researchers explicitly justified their choice of scoring method based on theoretical grounds. Choice of scoring method affects classification of individuals using the same androgyny measure, and may thereby influence what conclusions are derived from data analyses.

2.5 Androgyny as Type: Associated Characteristics

Numerous studies were conducted to demonstrate the unique characteristics associated with each typological category. In particular, because androgyny was seen as an ideal of human functioning, androgynous persons commonly were hypothesized to demonstrate superior adaptability, flexibility, and psychological health compared with sex-typed individuals, or to undifferentiated individuals who see neither set of characteristics as particularly self-descriptive. Unfortunately, too many studies appeared to be fishing expeditions to find any conceivable pattern of relationship between the types and other measures linked loosely to gender.

The evidence delineating the types has not been as compelling as many proponents would have liked. Generally, distinctions between the categories have been irregular and modest in size, in directions predictable from simple correlations with the masculinity and femininity scales. Androgynous persons tend to be favored, although not always, and the significant effects were frequently attributed to the power of one dimension shared with sex-typed counterparts (e.g., androgynous and masculine persons both scoring high on self-esteem because of a positive correlation with masculinity scores). Individuals classified as undifferentiated (low on both masculinity and femininity) appeared typically disadvantaged to some extent.

The most serious questions about the adequacy of the androgyny model have arisen partly because the multidimensionality of gender-related variables was underestimated. The presence of numerous sex differences reported in data analyses suggested that the process, likelihood, and implications of self-descriptions into the same category may well differ substantially for men and women. For example, Spence and Helmreich (1978) documented differential relationships by sex between femininity and masculinity and respondents' perceptions of relationships with their parents. Such interactions by sex were not difficult to acknowledge in the case of the sex-typed categories (e.g., feminine women and feminine men), but contradicted the notion that androgynous persons somehow transcend sex-based distinctions. Relationships between feminine and masculine self-descriptions and other gender-relevant variables such as attitudes and stereotypes have also not been strong. For example, an individual who is androgynous in self-description could appear to be quite traditional in feminist attitudes, feminine in appearance and informal social interaction, and masculine in behaviors on the job. Research generally has contradicted assumptions about the unitary nature of gender phenomena; a single set of traits (masculinity and femininity) or process (e.g., Bem's gender schema theory) simply cannot account for the complexity of gender-related differences within and between women.

Spence, an eminent researcher in gender psychology, has repeatedly cautioned researchers to expect such apparently disconfirming findings. From the earliest androgyny studies she emphasized that traits are, at best, behavioral predispositions that may be overridden by a multitude of individually idiosyncratic or situational factors. Trait-behavior connections are likely to be further strained when the dependent measure is peripherally related to the instrumental and expressive content embodied in the androgyny measures. Gender-related attributes, beliefs, and behaviors are likely to be substantially independent from one another, yet each related to other factors that may or may not be interconnected themselves. Spence has recommended abandoning the conceptualization of pervasive gender-related typologies in favor of assessing independently the diverse factors and processes describing and maintaining sex-based differentiation.

3. *Current Status*

Androgyny research appeared to lose momentum in the late 1980s, probably because the complexities revealed in empirical research contradicted the appealingly straightforward hypotheses characteristic of its heyday. Publications in the 1990s tended to focus on applying concepts and measures to an expanded range of cultural and age groups; continued testing of hypothesized typological differences and scoring variations; and most pertinent to recent developments within gender psychology, exploration of multifactorial models (e.g., Twenge 1999) and contextual differences (e.g., Swann et al. 1999), for example, in gender-related determinants of career behavior (Eccles et al. 1999). The conceptualization of androgyny retains its appeal as an alternative to restrictive, gender-based ideologies and roles. In a society that maintains gender distinctions at every level from personal self-concepts to sociopolitical institutions, actualizing this ideal appears more elusive than once thought. The masculinity and femininity scales developed for the androgyny measures continue to be most useful to study variables conceptually linked to the instrumental/agentic and expressive/communal dimensions represented by them.

See also: Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gay, Lesbian, and Bisexual Youth; Gender and Feminist Studies; Gender and Feminist Studies in Anthropology; Gender and Feminist Studies in Psychology; Gender Differences in Personality and Social Behavior; Gender History; Gender Identity Disorders; Gender Ideology: Cross-cultural Aspects; Gender-related Development; Masculinities and Femininities; Sex-role Development and Education; Sexual Behavior: Sociological Perspective; Sexual Orientation: Historical and Social Construction; Sexuality and Gender

Bibliography

Bakan D 1966 *The Duality of Human Existence*. Rand McNally, Chicago

Bem S L 1974 The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology* **42**: 155–62

Bem S L 1981 Gender schema theory: A cognitive account of sex typing. *Psychological Review* **88**: 354–64

Constantinople A 1973 Masculinity-femininity: An exception to a famous dictum? *Psychological Bulletin* **80**: 389–407

Cook E P 1985 *Psychological Androgyny*. Pergamon, Elmsford, NY

Cook E P 1987 Psychological androgyny: A review of the research. *The Counseling Psychologist* **15**: 471–513

Eccles J S, Barber B, Jozefowicz D 1999 Linking gender to educational, occupational, and recreational choices: Applying the Eccles et al. model of achievement-related choices. In: Swann W B Jr, Langlois J H, Gilbert L A (eds.) *Sexism and Stereotypes in Modern Society: The Gender Science of Janet Taylor Spence*. American Psychological Association, Washington, DC pp. 153–92

Mednick M T 1989 On the politics of psychological constructs: Stop the bandwagon, I want to get off. *American Psychologist* **44**: 1118–23

Parsons T, Bales R F 1953 *Family, Socialization, and Interaction Process*. Free Press, New York

Pedhazur E J, Tetenbaum T J 1979 Bem Sex Role Inventory: A theoretical and methodological critique. *Journal of Personality and Social Psychology* **37**: 996–1016

Spence J T 1993 Gender-related traits and gender ideology: Evidence for a multifactorial theory. *Journal of Personality and Social Psychology* **64**: 624–35

Spence J T, Helmreich R L 1978 *Masculinity and Femininity: Their Psychological Dimensions, Correlates, and Antecedents*. University of Texas Press, Austin, TX

Swann Jr W B, Langlois J H, Gilbert L A (eds.) 1999 *Sexism and Stereotypes in Modern Society: The Gender Science of Janet Taylor Spence*. American Psychological Association, Washington, DC

Taylor M C, Hall J A 1982 Psychological androgyny: Theories, methods, and conclusions. *Psychological Bulletin* **92**: 347–66

Twenge J M 1999 Mapping gender: The multifactorial approach and the organization of gender-related attributes. *Psychology of Women Quarterly* **23**: 485–502

Whitley B E 1983 Sex role orientation and self-esteem: A critical meta-analytic review. *Journal of Personality and Social Psychology* **44**: 765–85

Wilson F R, Cook E P 1984 Concurrent validity of four androgyny instruments. *Sex Roles* **11**: 813–37

E. P. Cook

Animal Cognition

Animal cognition is that branch of the biological sciences that studies how animals perceive, learn about, remember, understand, and respond to the world in which they live. The mechanisms it investi-

gates are evolved ways of processing information that promote the fitness or survival of the organisms that possess them. A synonym for animal cognition is comparative cognition, because researchers are often interested in the different ways in which evolution has equipped different species to process information. Comparisons between the human mind and those of animals have always held a fascination for both laymen and scientists. Careful experimental research in animal cognition shows how people and different species of animals are similar and different in the ways they comprehend their environments. Interesting new findings about animal cognition have been revealed as part of the cognitive revolution in psychology of the last 30 years. Most animals show cognitive competence in three major dimensions of the natural world—time, number, and space. New discoveries and theories offer important insights into the ways animals keep track of time, count the frequency of events, and learn to navigate successfully through space.

1. *Keeping Track of Time*

Animals, like people, keep track of time using two types of clocks, a relatively coarse 'time-of-day clock' and a fine-grain 'interval-timing clock.'

1.1 *The Time-of-Day Clock*

Time of day is given by *zeitgebers* or 'time givers' provided by internal biological circadian rythms that cycle through the same stages each day. Proof that animals keep track of time of day was provided in a classic experiment carried out by Biebach et al. (1989). Garden warblers were kept in a chamber that contained a living room and four feeding rooms. Each feeding room contained a feeder that provided a bird with food at different times of day: a bird could gain access to food in Room 1 from 6.00 to 9.00 a. m.; in Room 2 from 9.00 to 12.00 noon; in Room 3 from 12.00 to 3.00 p.m.; and in Room 4 from 3.00 to 6.00 p.m. Within only a few days of experience with this schedule, warblers went to each room at the time of day when it contained food. Even when food was made available in all of the rooms throughout the day, birds still went to Room 1 first, switched to Room 2 at about 9.00 a.m., switched to Room 3 at about 12.00 noon, and switched to Room 4 at about 3.00 p.m. Thus, even in the absence of external cues for time of day, such as the sun, birds knew the time of day when food would be in different locations.

1.2 *The Interval-timing Clock*

By contrast to the time-of-day clock, the interval-timing clock tracks intervals of a few seconds to a few minutes. Suppose a rat placed in a chamber is occasionally rewarded with a food pellet for pressing a bar. Each time a tone is presented, the first bar press made after 30 seconds causes a pellet to be delivered. Notice that presses before the end of the 30-second interval are futile. After some training, the rat learns to begin pressing the bar only a few seconds before the 30-second interval elapses. In this way, it earns the reward with a minimum of effort. The rat has learned to respond only near the time when it will be rewarded. If a curve is plotted showing the rate of bar pressing against time since the tone began, the peak of the curve appears at just about 30 seconds. This experiment has been performed with many species of animals, and they all show an ability to precisely time short intervals.

1.3 *Scalar Timing Theory*

A highly successful theory of interval timing, called *scalar timing theory*, was devised by Gibbon and Church (1990). A flow of information diagram depicting this model is shown in Fig. 1. The model is based on the assumption that a pacemaker in the brain is constantly emitting pulses at a fixed rate. When a signal is presented, it activates a switch which closes and allows pulses to flow from the pacemaker to an accumulator. In the example of the rat pressing a bar for food, the tone would close the switch, and pulses would flow into the accumulator until the rat earned its reward and the tone stopped, reopening the switch 30 seconds later. Pulses in the accumulator enter a temporary working memory, and the total in the working memory at the moment of reward is sent for permanent storage to a reference memory. Over a number of trials, a distribution of pulse totals representing intervals centered around 30 seconds will be stored in reference memory. Once a set of criterion times has been established in reference memory, an animal can track time by using a comparator that compares the pulses accumulating in the working memory with a criterion value retrieved from reference memory. Importantly, this comparison is made as a ratio; the difference between the criterion value and the working memory value is divided by the criterion value. Thus if the criterion retrieved from reference memory represented 30 seconds, and only 15 seconds had elapsed since the signal began, the comparator would yield a ratio of $(30-15)/30 = 0.50$. The model assumes that once this ratio drops below a threshold, say 0.25, the organism makes the decision to begin responding. The fact that the decision to respond is based on a ratio gives the model its scalar property. The scalar property predicts that error in responding will be proportional to the length of the interval timed. Thus, if the threshold ratio is 0.25, and the interval timed is 30 seconds, an animal should start responding 7.5 seconds before the end of the interval. If the interval timed is 60 seconds, however, the animal

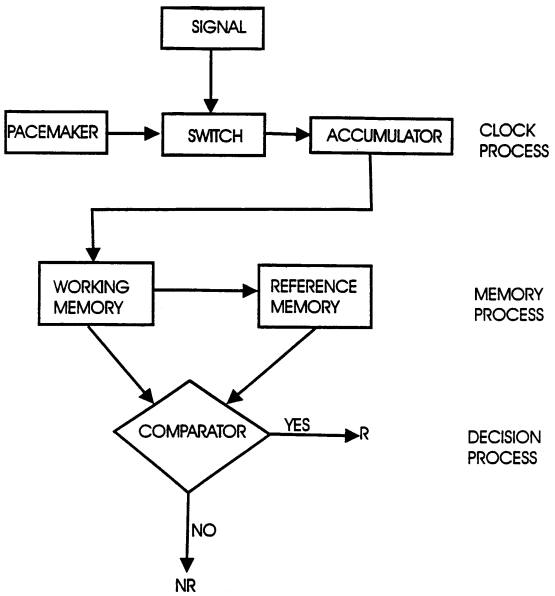

Figure 1
This cognitive model of timing shows how pulses produced by a pacemaker flow into an accumulator and are sent to working memory. Criterion values stored in reference memory are retrieved into a comparator that compares the criterion number with the accumulating total in working memory and makes a decision to respond (R) or not to respond (NR)

should start responding at 15 seconds before the end of the interval. Numerous experiments show that this is exactly what happens.

2. *Numerical Operations in Animals*

Can animals count? This is a highly controversial question. Although animals may not be able to perform all of the numerical operations humans can, there is good evidence that they can keep track of the frequency of events.

2.1 *Eating a Fixed Number of Food Pellets*

Based on experiments with birds first performed by the German ethologist Otto Koehler (1951), Davis and Bradford (1991) allowed rats to walk along a narrow elevated pathway from a platform to a chair in order to eat food pellets placed in a pile. There were three groups of rats, designated '3 eaters,' '4 eaters,' and '5 eaters.' The rats in each group were verbally praised if they ate their designated number of pellets and then left the chair. If a rat ate more than its group assignment, however, it was scolded and shooed back to the platform. After considerable training (200 trials), the rats in each group ate predominately the correct number of pellets assigned to their group.

2.2 *The Clever Hans Problem*

The Davis and Bradford experiment is reminiscent of one performed at the beginning of the twentieth century. A retired German school teacher named von Osten owned a horse named Clever Hans, who he claimed could not only count but could perform a number of numerical operations. When asked to add, subtract, multiply, or divide two numbers, Hans would tap his hoof on the ground and reliably stop tapping when he reached the correct number. Hans convinced many doubters by passing tests even when problems were given to him by new testers he had not worked with before. Eventually a young scientist named Oskar Pfungst (1965) tested Hans by showing him problems that could not be seen by the tester standing in front of him. Now Hans failed miserably, and it was concluded that prior successes had arisen not from the ability to perform mathematics but from cues provided unwittingly by testers, who slightly bowed their heads when Hans had reached the correct number of taps.

In order to avoid the problems raised by the Clever Hans effect, Davis and Bradford allowed their rats to be tested with no-one in the testing room and observed the rats' behavior over closed-circuit television. In the complete absence of a human tester, and thus any possible Clever Hans cues, the rats still ate the appropriate number of pellets they had been trained to eat. The rats apparently had learned a restricted eating rule based on number of pellets taken from the pile.

2.3 *Addition of Numbers*

Beyond a sensitivity to number, is any animal capable of performing numerical operations that combine numbers? Some surprising research carried out with chimpanzees suggests that they are able to add numbers. Boysen and Berntson (1989) trained a chimpanzee named Sheba to associate Arabic numbers with numbers of objects. Eventually, Sheba could be shown a number of objects, such as two oranges or five bananas, and Sheba would correctly choose a card containing the number '2' or '5' that correctly indicated the number of items shown. Different numbers of oranges then were hidden in three different locations in the laboratory, and Sheba was allowed to inspect each location. She was then allowed to select among cards containing the numbers 1–4. Sheba reliably chose the card that equalled the total number of oranges she had found. An even more impressive finding was revealed when Sheba visited the hiding places and found cards containing number symbols. Now when offered a choice among the numbers 1–4, she chose the card that represented the sum of the

numbers just encountered. The fact that Sheba chose the total of both objects and number symbols without summation training suggests that the ability to add may have been a quite natural and automatic process for her.

3. Navigating through Space

One of the most impressive cognitive abilities of animals is that of accurately navigating through their spatial environment. It is critical for an animal's survival that it knows the locations of such things as its home, water, foods of different types, mates, competitors, and other species that might prey on it. In addition, an animal must keep track of its own travel through its environment and be able to plot a course readily back to its home base. Research has shown that animals use a number of cues and mechanisms to navigate through space. The cues used are divided into two general classes, egocentric cues and allocentric cues. Egocentric cues are cues generated by an organism as it moves through space and that allow it to keep track of its own position relative to its starting location through the mechanism of deal reckoning. Allocentric cues are structures or objects external to the organism that are used as landmarks to guide travel through space. Two important allocentric cues are the geometric frame and landmarks.

3.1 Dead Reckoning

Both insects and mammals can travel long and twisting paths from their home base in total darkness and still return home along a straight-line route. A person wearing a blindfold can be led some distance from his/her starting position through several turns; when asked to return to the start, the person will walk directly to an area near the start. In all of these cases, the return route is computed by dead reckoning, also known as path integration. The return route has two components, direction and distance. As an organism moves through space, it has several sources of internal cues that are recorded and indicate position in space. In mammals, vestibular organs in the semicircular canals measure angular acceleration created by inertial forces when turns are made. In addition, distance information is provided by proprioceptive feedback from the muscles involved in locomotion and by efference copies from motor commands. These cues are integrated to plot a return vector to home that contains information about both direction and distance. Although dead reckoning is highly effective, errors arise as the number of turns in a trip increases, and the angle of the return vector may deviate progressively from the correct one. Thus, animals often use dead reckoning to return to an area near their home and then use local landmarks to find their nest or hole.

3.2 The Geometric Frame

Cheng (1986) placed hungry rats in a walled-in rectangular arena which contained loose bedding on the floor. Food was always buried under the bedding in one corner of the arena. Rats learned to dig in this location for food on repeated trials. However, they often made an interesting error; they dug in the corner diagonally opposite the correct one. The rats' mistake told Cheng something very important about the way these animals had encoded the location of food in space. They had used the entire rectangular framework of the arena and thus coded food as being in the corner with a long wall on the left and a short wall on the right. Unfortunately for the rats, this code also fit perfectly the diagonally opposite corner and thus led to frequent error. People too frequently use the geometric framework of their home or office to find important locations but are largely unaware of this cue.

3.3 Landmarks

In the open world, animals frequently use isolated landmarks, such as trees, rocks, or man-made structures, to guide their travel. Many animals use a sun compass by navigating relative to the azimuth of the sun and use celestial cues provided by the stars at night. A forager honeybee having found a source of nectar returns to the hive and performs a waggle dance that informs its hive mates about the location of the food relative to the position of the sun. If the bees are not released from the hive for some time during which the position of the sun has changed, they still accurately fly to the location of the food. They have a built in clock that allows them to correct for the change in the position of the sun.

The movement of landmarks is frequently performed by researchers to examine how animals use landmarks. If an animal learns to search for food at a fixed location relative to some rocks and logs, an experimenter may arrange a test on a given day by moving each landmark a fixed distance to the west. The animal's search location then also will be shifted this distance to the west, showing its use of landmarks. Some interesting recent experiments carried out in the laboratory with pigeons show that these birds do not always use landmarks in the same way people do. Spetch et al. (1996) trained pigeons and people to locate a hidden goal on a computer monitor touchscreen. The touchscreen automatically detects and records any place on the monitor that the pigeon pecks or a human touches. One pattern of landmarks is shown in the Control panel of Fig. 2. In order to

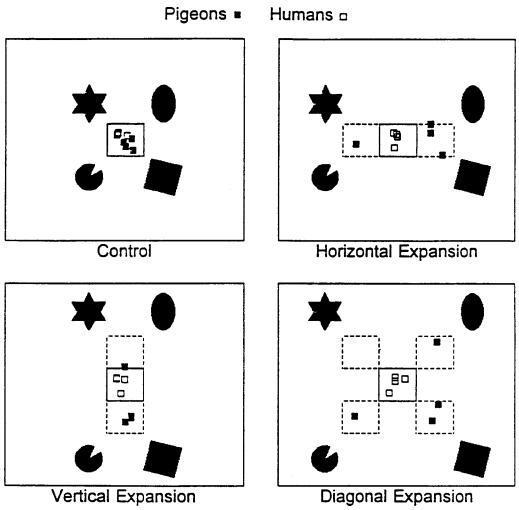

Figure 2
The black geometric patterns formed a landmark array on a touchscreen within which the invisible goal was located. After pigeons and humans learned to touch the screen at the goal location (Control), both species were tested with expansions of the landmark array in horizontal, vertical, and diagonal directions. The small squares show the locations where individual pigeons and humans made most of their contacts with the screen (reproduced by permission of the American Psychological Association from the *Journal of Comparative Psychology*1996, **110**: 22–8)

receive a reward (food for the pigeon and points for the human), the subject had to contact the touchscreen in an area equidistant from each of the geometric pattern landmarks. The configuration of landmarks was moved about the screen from trial to trial, so that the subjects had to learn to use the landmarks and could not always respond to a fixed location on the screen. The Control panel shows the average locations where four pigeons and four people learned to respond relative to the landmarks.

Tests then were carried out on which the landmarks were expanded, horizontally, vertically, or diagonally in both directions. Notice that all of the people continued to search in the central area, an equal distance from each landmark. When asked, each person said 'the goal was in the middle of the landmarks.' Pigeons did something quite different. The locations where pigeons searched indicated that each pigeon had coded the goal as being a fixed distance and direction from one landmark. Different pigeons used different landmarks. This experiment provides an example of a species difference in cognitive processing. Humans used all of the landmarks and coded the goal as being in the middle, but pigeons chose only one landmark and coded the position of the goal relative to that single landmark. Later research showed the same species difference when pigeons and people had to locomote across a field to find a hidden goal (Spetch et al. 1997).

4. *Advanced Processes*

Advanced processes refer to abilities once considered unique to humans. These processes typically seem to require higher level reasoning and abstraction of information beyond that which is perceptually available. Their investigation in animals sometimes provides surprising insights into animal cognition and sometimes confirms the hypothesis that an ability is unique to humans. Examples from two areas of research will be described.

4.1 *Concept Learning*

A concept refers to a common label we give to a number of objects, actions, or ideas that are not exactly alike but have some properties in common. We use concepts all the time when we give verbal labels to things. For example, we call all cats 'cat' and all flowers 'flower,' even though all cats and all flowers do not look alike. Somehow, we have learned that these things share sufficient properties in common to be given a common label. Concepts provide us with considerable cognitive economy, as it would be laborious to distinguish each cat and each flower and give it a separate name.

Can animals form concepts? The answer appears to be yes, at least for the pictorial concepts just described. Experiments originally performed by Richard Herrnstein (Herrnstein et al. 1976) at Harvard University and more recently by Edward Wasserman (1995) at the University of Iowa show that pigeons can readily distinguish between the same categories of objects that people do. A pigeon is tested in a chamber that contains a square-shaped screen on one wall, with keys placed at the four corners of the screen. The pigeon is presented with a visual slide show that consists of the successive presententation of pictures from the categories of people, flowers, cars, and chairs, with each picture containing a different example of each category. The pigeon is rewarded with food for pecking only one of the keys when a picture from a particular category is shown. For example, a peck on the key in the upper left corner might yield reward whenever a flower picture appears, but only a peck on the key in the lower right corner would yield reward when a picture of a chair is shown. Repeated experiments have shown that pigeons learn to peck the correct key with about 80% accuracy whenever a picture in any of the categories is shown. One might worry that pigeons would memorize the correct response to these pictures, since they are shown repeatedly over sessions. Memorization does not appear to be the basis for the pigeon's

ability to sort out these pictures, because the pigeons continue to respond accurately the first time they encounter new pictures from each of these categories. This demonstration is striking because it shows that in the absence of language, which humans use to label categories, a pigeon is able to sort two-dimensional pictures into the same semantic categories that are meaningful to people.

4.2 Theory of Mind

The term 'theory of mind' refers to the fact that people know about minds. Each of us knows that we have a mind of our own. Thus, you know that your mind has a memory and that that memory contains certain information you cannot retrieve at this moment but may be able to retrieve at some later time. You also know that other people have minds and that those minds know some things that your mind knows and other things your mind does not know. The inferences you make about others minds may often guide your behavior. For example, if you are a teacher, you may teach a particular lesson if you believe your students do not know it, but you will skip it if you believe it is already contained in their minds.

Researchers working with non-human primates have raised the question, 'Do apes or monkeys have a theory of mind?' (Premack and Woodruff 1978). Although this question is highly controversial, some observations of primates suggest that it is possible. Menzel (1978) kept a group of chimpanzees in an outdoor enclosure near an open field. One chimpanzee was selected and taken about the field to be shown various locations where food was hidden. This chimp then was returned to its companions, and they were all released to search for food. The chimpanzees soon learned to track the path of the one that had been shown the locations of food, often searching ahead of the knowledgeable chimp. Some of the chimps that were shown food seemed to learn to throw off their pursuers by running in a false direction and then suddenly doubling back to get the food before the others could catch up to them. One interpretation of these behaviors is that the chimps pursued the informed individual because they knew only its mind contained valid information about food locations. Similarly, the knowledgeable chimp ran in the wrong direction in order to deceive the minds of its companions.

5. Conclusion

The examples discussed are a sampler of the many problems studied by scientists interested in animal cognition. They should convey the impression that among the many species of animals that inhabit the earth, each processes information in ways that are unique to it and in ways that are common to other species. The purpose of the field of animal cognition is to understand these similarities and differences and how they were selected by the process of evolution.

Bibliography

Biebach H, Gordijn M, Krebs J R 1989 Time-and-place learning by garden warblers (*Sylvia borin*). *Animal Behaviour* **37**: 353–60

Boysen S T, Berntson G G 1989 Numerical competence in a chimpanzee (*Pan troglodytes*). *Journal of Comparative Psychology* **103**: 23–31

Cheng K 1986 A purely geometric module in the rat's spatial representation. *Cognition* **23**: 149–78

Davis H, Bradford S A 1991 Numerically restricted food intake in the rat in a free-feeding situation. *Animal Learning & Behavior* **19**: 215–22

Gibbon J, Church R M 1990 Representation of time. *Cognition* **37**: 23–54

Healy S 1998 *Spatial Representation in Animals*. Oxford University Press, New York

Herrnstein R J, Loveland D H, Cable C 1976 Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes* **2**: 285–302

Koehler O 1951 The ability of birds to count. *Bulletin of Animal Behavior* **9**: 41–5

Menzel E W 1978 Cognitive mapping in chimpanzees. In: Hulse S H, Flowler H, Honig W K (eds.) *Cognitive Processes in Animal Behavior*. Erlbaum, Hillsdale, NJ

Pfungst O 1965 *Clever Hans: The Horse of Mr. von Osten*. Holt, Rinehart and Winston, New York

Premack D, Woodruff G 1978 Does the chimpanzee have a theory of mind? *Behavioral Brain Sciences* **1**(4): 515–26

Roberts W A 1998 *Principles of Animal Cognition*. McGraw-Hill, Boston

Shettleworth S J 1998 *Cognition, Evolution, and Behavior*. Oxford University Press, New York

Spetch M L, Cheng K, MacDonald S E 1996 Learning the configuration of a landmark array: I. Touch-screen studies with pigeons and humans. *Journal of Comparative Psychology* **110**: 55–68

Spetch M L, Cheng K, MacDonald S E, Linkenhoker B A, Kelly D M, Doerkson S R 1997 Use of landmark configuration in pigeons and humans: II. Generality across search tasks. *Journal of Comparative Psychology* **111**: 14–24

Wasserman E A 1995 The conceptual abilities of pigeons. *American Scientist* **83**: 246–55

W. A. Roberts

Animal Rights in Research and Research Application

1. The Concept of Animal Rights

1.1 Animal Rights and Animal Welfare

In analogy to human rights, *animal rights* can be described as rights that are attributed to animals as consequence of being animal without any further condition. The term animal rights has two different

meanings: in a narrower sense it refers to a point of view that animals have an inherent right to live according to their nature, free from harm, abuse, and exploitation by the human species. This is contrasted by an *animal welfare* position, which holds to minimize animal suffering without rejecting the human practice to use animals for e.g., nutritional or scientific purposes. Animal rights in a wider sense include both positions.

Scope as well as legitimacy of animal rights is highly controversial. Among the issues under discussion is beyond the question whether animals can be attributed rights at all, the problem of drawing a boundary: do only mammals have rights, all vertebrates or even bacteria? There are also different foundations of these rights: are they to be derived from utilitarian principles or are they to be assigned on a deontological basis?

The concept of animal rights extends well beyond the field of research which is the focus of this article. Most ethical considerations discussed below apply equally to the issues of raising and killing animals for food, using animals for sport or as companion animals (e.g., pets) as well as for fighting pests (i.e., poisoning rats or pigeons in cities) most of which affect larger numbers of animals than research.

1.2 Moral Agents and Moral Patients

Ethics distinguishes between *ethical subjects* or *moral agents* on the one hand as the set of subjects who have ethical obligations, and *objects of ethics* or *moral patients* on the other hand designating the set of individuals or objects ethical obligations are owed to. Since it requires certain preconditions to act ethically, among them rationality, most ethical theories agree to equalizing mankind and the set of moral agents. Exceptions of this are slightly more controversial but usually include individuals to whom rationality cannot be attributed, for instance young children or senile and mentally ill people.

1.3 Different Realms of Moral Patients

Classical anthropocentrism is the ethical position held for centuries that takes for granted that only human beings can be the objects of ethics. This equals the set of moral agents and moral patients, except for the few cases mentioned above. In contrast to this the approach taken by many animal rightists can be described as *pathocentric*, since it defines the set of moral patients by the definition 'all that can suffer.' Thus a pathocentrist approach fits the traditional utilitarian position that aims to minimize the total amount of suffering in the world.

1.4 Deontological and Teleological Foundations of Ethics

Philosophical foundations of ethical theories can roughly be divided in two lines of reasoning: Deontological principles imply that certain actions are intrinsically bad (or good), regardless of their consequences. Philosophers or theologists arguing this way require an instance which determines the set of bad actions or at least a method or law to determine this set, like Kant proposed with his categorical imperative. Problems with this reasoning are on the one hand the establishment of the ethical instance (since there will be competing models, for example other religions) and the deduction or interpretation of ethical principles from this instance (be it religious amendments or be it 'the will of the nature').

A strict animal right view, which ascribes certain rights to all animals and the classical anthropocentric view, which disregards animals as objects of ethics are both deontological positions that may differ in the derivation of their justification.

The other way of grounding ethical principles is teleological reasoning, which judges actions by their outcome (usually disregarding the intentions). The problem here is how to determine on which dimensions which outcome of an action is to be considered as morally good (hedon) vs. bad (dolor). Different variants of utilitarianism provide different answers to what has *intrinsic value*: Jeremy Bentham named pleasure as hedon and pain as dolor, John Stuart Mill sets out for the more general happiness as primary good.

2. History

Welfarist and animal right ideas developed against the background of the Western anthropocentric ethics. Before sketching this development it should be noted, however, that there are other cultures and religions that emphasize the sanctity of all life and thus are a realization of animal rights in the narrower sense. The concept of *ahimsa* (Sanskrit for noninjury) as the standard by which all actions are to be judged for example is rooted in Hinduism as well as in Buddhism and finds its strictest interpretation by the Jains in India, where cloth mouth covers are worn to prevent the death of small organisms by inhalation.

From the beginning of philosophy there were differences among philosophers with respect to the treatment of animals. While Pythagoras was vegetarian, Aristotle performed vivisections. The right of men to use animals for their purposes was part of Christian tradition from the very beginning. Mankind, made after God's own image and superior to all other beings, should have dominion over every little thing that moveth upon the earth (see also *Research Conduct: Ethical Codes*).

A continuous development of the moral status of animals can be reconstructed since René Descartes. When he established the dualism of mind (attributed

only to men) and matter as mere spatial extension, animals were seen as devoid of *res cogitans* and hence as pure mechanical automatons without subjective experience. This position lead to the abandonment of welfarist positions especially in research. Vivisections of alive, conscious animals were common since anesthetics were not invented yet and contemporary scientists argued that an animal yelling is like a clock that also makes predictable noises when handled appropriately. These experiments however lead to the discovery of remarkable similarities between animals and humans and it might be partly their merit that the assumption of a fundamental difference between humans and animals was undermined.

During the Enlightenment it became common sense (again) that animals are sentient beings. The rising welfarist attitude might be summed up in the term 'gentle usage' used by David Hume in his *Enquiry Concerning the Principles of Morals* (1751, Chap. 3). Most philosophers did not go as far as to include animals in the set of ethical objects for their own sake. Their indirect duty positions state that we have certain obligations to animals for the sake of their owner or for the sake of 'compassion.' This aptitude was useful in respect to our behavior towards neighbors but blunted by cruelties against animals, Kant wrote in §17 of his *Metaphysik der Sitten* in 1797.

Jeremy Bentham's *An Introduction to the Principles of Morals and Legislation* (1789), which aims to derive a law system from the principles of social eudaimonism, the utilitarian principle of 'the greatest happiness for the greatest number,' does contain the footnote most often quoted in the animal right discussion. Here Bentham criticizes that the interests of animals are not included in the existent laws, compares this with juridical systems neglecting the rights of slaves, and suggests that sentience is the 'insuperable line' beyond which the interest of any being has to be recognized: 'the question is not, Can they reason? Nor, Can they talk? But, Can they suffer?' (Chap. 17).

While the first example of an anticruelty law dates back to 1641 and is to be found in the legal code of the Massachusetts Bay Colony (US), only in the nineteenth century were efforts made to implement in laws some of the new ethical intuitions regarding animals. This juridical development, e.g., the *Martin Act* in the UK (1822) and the *Law Grammont* (1840) in France, was paralleled by the foundations of societies against cruelties to animals, which often had the purpose of enforcing the newly passed laws. Darwin's theories continued to narrow the gap between humans and animals during the late nineteenth century.

3. *Contemporary Positions*

The time since the enlightenment can be reconstructed as a transition from indirect ('anthropocentric-esthetic') to direct duty positions, with welfarist positions gaining popularity. It nowadays is common sense that animals—at least starting with a given phylogenetic complexity—are sentient creatures and thus to be included in some way in the set of moral patients.

A landmark in establishing a welfarist position in animal experimentation was made with *The Principles of Human Experimental Technique*, published first 1959 by William M. S. Russell and Rex Burch which postulated that humane experimental techniques had to conform to the *3Rs: replacement* of animal experiments whenever possible, *reduction* of the numbers of animals used where animal experiments cannot be avoided (i.e., by sophisticated experimental plans that lead to significant results using less animals), and *refinements* of methods to minimize pain and distress experienced by the animals in a given experiment.

One of the most influential works with respect to the moral status of animals is *Animal Liberation* (1975) by the Australian philosopher Peter Singer. This book was written to persuade as many readers as possible to abandon the use of animals regardless of their suffering less than being a waterproof piece of academic work. Singer was extraordinarily successful in propagating his position, but at the price of the quality of argumentation.

For a critical examination of his chapter on animal experimentation see Russell and Nicoll (1996). Taking a utilitarian view, Singer claims sentience as the only dimension that defines the set of moral patients and uses the ability to suffer as a standard example. Two conclusions follow from this line of argumentation.

First, the common use of animals for human purposes, which is more or less inevitably associated with their suffering, is to be condemned, since the animal suffering outweighs by far the human pleasure in the examples given. Even in the absence of a list of hedons and dolors which would allow one to estimate quantitatively whether an act is moral or not by computing the consequences for the animal and the 'general welfare' (as Porter tried for animal experiments in 1992), most utilitarians including Singer agree on a special status of beings which are able to plan and anticipate the future. For those beings one assumes *preference utilitarianism*, which takes the satisfaction of preferences as good, whereas *hedonistic utilitarianism* is assumed for the rest of sentient beings. Because fulfillment of preferences is usually thought to have a bigger moral weight than the intrinsic values of the hedonistic utilitarianism this rules out the possibility to use these subjects against their will in ways that will harm them but will be a benefit for the rest of the moral community.

But on the other hand putting sentience as the only yardstick by which to determine moral patients has a flipside that even preference utilitarianism fails to avoid. It does not condemn sacrificing so-called *human marginal cases* like anencephalic newborns or coma-

tose patients to whom no preferences can be attributed because they lack consciousness or even sentience for purposes of the general welfare, given no other individuals like the parents would suffer from this. This obviously deviates from the ethical intuitions held by a large part of the population in the Western hemisphere. Using human marginal cases for research is thought to be immoral, while the use of animals who might suffer more is instead thought to be morally less problematic, if not sound under certain circumstances.

Well aware of this, Singer uses the notion *speciesism* in analogy to racism for this, quoting Bentham's comparison with slavery among others as examples that exclusion of a given set of (in the examples always human) subjects from the set of moral patients in a certain society is by no means a guarantee that this restriction is not recognized as inappropriate by the following generations. He pleads for using neither higher mammals nor humans in research but his position logically includes that there are circumstances that justify sacrificing men as well as other mammals.

Utilitarian positions furthermore allow each moral agent to kill moral patients in a fast and painless way, provided nobody else suffers from this ('secret killing argument'), unless an additional assumption of being alive as hedon is made, for instance grounded on possible positive experiences in the future. Besides that this is against intuition in the case of humans, it also allows so-called terminal experiments in the life sciences which put the (humanely raised) animals under anesthesia and sacrifices them after the experiment is finished, since no suffering would be induced.

There are many attempts to avoid the problems regarding the human marginal cases, among them to use a deontological corollary that human life is protected even when it fails the utilitarian criteria for being the status of a moral patient. This may mirror the ethical intuitions of at least the Western population but it is exactly the attitude criticized as speciist. At this point there is a conflict between two tasks of philosophical ethics: on the one hand it should organize the set of moral intuitions into a coherent system, while on the other hand it should be based on and reflect the grown and sometimes incoherent moral intuitions of an epoch.

In *The Case for Animal Rights* Tom Regan (1983) rejects utilitarian foundations of animal rights beyond others for the above reasons and offers a deontological alternative. Starting point is a careful derivation of criteria for moral patients. All those beings that have a psychophysical identity over time, beliefs, desires, and a sense of the future, including their own future, are defined as 'subjects-of-a-life.' Regan grounds his deontological position by trying to fulfill certain formal demands, among them conceptual clarity, rationality, impartiality, and compliance to reflected moral principles (considered beliefs). On these grounds he considers it as the best ethical position to attribute an *inherent value* to all subjects-of-a-life. This inherent value is held by the subject, whereas the utilitarian intrinsic value is held by the subject's *experiences* (hedons and dolors). Regan criticizes this as treating the subject as 'mere receptacle' of experiences.

Regan goes on to derive several principles, first of which is the *respect principle*: moral agents are obliged to treat individuals with inherent value in a way that respects this value. From this the *harm principle* is derived: each moral agent has a direct duty not to harm any moral patients, the flipside of which is the inherent right of the moral patient not to be harmed.

Harm is given by inflictions or deprivations, and since to kill a 'subject-of-a-life' is to deprive a moral patient of its future experience, all killings are explicitly condemned by Regan's view, which rules out one of the weaknesses of utilitarianism.

Regan derives more principles, showing that according to the rights view it is immoral to negate the inherent value of subjects-of-a-life by using them in medical research even in case the general welfare of future generations would be promoted for sure. Intrinsic values like the general welfare and inherent values are incommensurate and cannot be set off against each other.

While utilitarian opponents of animal research are obliged to discuss the potential benefit of the research, not only including potential results but also the benefit of the scientists, the suppliers of scientific equipment etc. Regan can avoid this discussion in the first place: 'research should take the direction away from the use of any moral agent or patient.'

While being consequent in the 'consequences aside' notion in demanding respect for the inherent value of moral patients in the above examples, the rights view has considerable weaknesses when it comes to exceptions. Regan interprets the rights view as compatible with the violation of other's values in self-defense or by punishing other moral agents besides other cases. Justification is basically drawn from our (reflected) 'moral intuitions.' These stand against declaring pacifism as moral obligatory by defining self-defense as wrong, 'consequences aside.' Since our moral intuitions, however, *do* also allow animal experiments (at least) in some exceptional cases, it is hard to see why these should not be moral on the same grounds. This turns the uncompromising rights view quite a bit towards a welfarist position. This is avoidable if no exceptions are allowed, like in the eastern philosophy of ahimsa.

4. *The Animal Rights Movement*

In addition to the established animal welfare institutions, since the 1970s a number of more radical animal rights organizations have been founded. While the latter usually define themselves as an essential comp-

lement or even replacement of the former, there is in fact a continuum in the movement's positions under which circumstances animals may be used for scientific purposes. The heterogeneity of the movement may explain the wealth of umbrella organizations like the International Fund for Animal Welfare (1969) or worldanimal.net (1997).

While at educational institutions animal rights were originally rooted in the philosophical departments, recently also law schools deal with the topic offering classes in animal rights law.

In some aspects the history of 'animal liberation' so far resembles that of other liberation movements, but one should remember the essential difference that in this case the liberated population does not participate itself. More comparable to other social movements is the following development: the less radical parts, like the animal welfarists, got integrated in society at large, which adopted some of their demands. This in turn weakens the more extreme parts of the movement and is fertile ground for radicalization.

Underground organizations like the Animal Liberation Front (ALF) use violence and terror as means to impose their views of a humane society, be it by arson to animal laboratories or by torture to their opponents. In 1999 the journalist Graham Hall was kidnapped and branded the letters 'ALF' to his back because of his prize-winning undercover documentary on the criminal violent methods of the ALF.

It should be noted, however, that there are also many groups which are radical in refusing the welfarist 'be nice to them until you kill them' approach while sticking to nonviolent (albeit not always legal) methods of fighting for their goals. The methods they use are the same as in other domains of resistance and include concerted actions to solicit publicity against special cases of animal usage, political lobbying, and fundraising. These organizations are no small 'Robin Hoods': people for the ethical treatment of animals (PETA) as the biggest group which is totally opposed to animal experimentation reported annual net assets of over $7 million in the late 1990s.

Scientists entered discussion only after having been attacked by the animal liberation movement and having been presented to the public as conscienceless monsters for quite a while. After activities of the movement finally reached the ivory tower stopping some research programs due to lobbying and when more radical activists raided laboratories and threatened scientists, eventually societies like 'Society for health and research' (GGF, Germany, 1985) or 'Americans for Medical Progress Educational Foundation' (1991) were founded. Their goals are to 'bolster public understanding and support of the humane use of animals in medical research' (AMPEF). Scientists now offensively promote the results of research by setting up websites, producing brochures aimed at special target groups, and watch and comment on the activities of the animal right groups.

Both sides seem to be aware of the fact that the most important battlefield might be education since moral intuitions are a heritage that is based more on cultural than on biological aspects. This is one of the central distinctions between human and animal behavior codes, which for example allow carnivores to kill prey, but prevents them in general from killing members of their own species. The development of human ethical intuitions has become the subject of scientific investigation in developmental psychology. Although such a complicated process can probably never be fully explained, it is obvious that it can be influenced. Thus Peter Singer demands 'animal rights' children books, and scientists strive for an intensification of education in the live sciences.

Analyses of the Animal Right movement can be found in Francione (1996) from a rightist and in Guither (1998) from a welfarist position.

5. *Juridical Implementation*

In case animal rights in a narrower sense are implemented in laws there are inevitable collisions with basic human rights such as the freedom of religion or the freedom of science. This may be among the reasons that most countries do not implement animal rights at the level of the constitution, where other basic rights are set but in special laws that do not override the constitutional rights.

The ethical intuition that animals should not suffer unnecessarily while it is morally justifiable to use animals for human purposes like research is reflected by the fact that the regulating systems for animal experiments are similar in most Western countries and consist of three columns. First, each experimenter has to have a *personal license* to ensure a quality standard of operation (for an example of educational requirements, see the guidelines of the Federation of European Laboratory Science Organizations). Second, animal experiments are only allowed within research projects, the goals of which have been reviewed and approved to guarantee that no animal experiments are conducted for trivial purposes. Third, research institutions which perform animal experiments have to be licensed and are bound to tight regulations considering animal care during, but also outside the experiments, which is important if one considers that the average laboratory animal spends much more time in the cage than in experiments.

In addition to nationwide requirements, most research institutions, especially universities, also have their own ethical guidelines, which can be more restrictive than the law and which are usually enforced by ethics committees (see also *Ethics Committees in Science: European Perspectives* and *Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives*).

While the structure of the law regarding animal experiments is very similar in most countries, the actual requirements diverge widely, as does the enforcement of laws. While the term animal as used in the US Animal Welfare Act (1966, see Animal Information Center resources) specifically excludes rats, mice, and birds, the German animal protection law protects all vertebrates and, while other laws just aim to prevent unnecessary suffering, aims also to protect the *life* of animals.

Despite the strict German law there may be a general tendency to use the smallest common denominator when putting ethical intuitions into national laws in the age of global competition, slowing down the trend to further implement animal welfare in national laws.

See also: Bioethics: Examples from the Life Sciences; Bioethics: Philosophical Aspects; Clinical Psychology: Animal Models; Comparative Neuroscience; Consequentialism Including Utilitarianism; Ethical Dilemmas: Research and Treatment Priorities; Ethics and Values; Experiment, in Science and Technology Studies; Human Rights, Anthropology of; Medical Experiments: Ethical Aspects; Objectivity of Research: Ethical Aspects; Research Conduct: Ethical Codes; Research Ethics: Research

Bibliography

Americans for Medical Progress: www.ampef.org

Animal Rights FAQ: http://www.hedweb.com/arfaq/

Animal Rights Law at Rutgers University: www.animal-law.org

Animal Rights Net (critical update on AR movement): www.animalrights.net

Animal Rights vs. Animal Welfare (welfarist, excellent resources): http://www.dandy-lions.com/animal_rights.html

Animal Welfare Information Center (US): http://www.nal.usda.gov/awic/

Federation of European Laboratory Animal Societies (FELASA): www.Felasa.org

Francione G L 1996 *Rain Without Thunder: The Ideology of the Animal Rights Movement*. Temple University Press, Philadelphia, PA

Guither H D 1998 *Animal Rights: History and Scope of a Radical Social Movement*. Southern Illinois University Press, Carbondale, IL

LaFollette H, Shanks N 1996 *Brute Science: Dilemmas of Animal Experimentation*. Routledge, London, New York

Porter D G 1992 Ethical scores for animal experiments. *Nature* **356**: 101–2

Regan T 1983 *The Case for Animal Rights*. University of California Press, Berkeley, CA

Russell S M, Nicoll C S 1996 A dissection of the chapter 'Tools for research' in Peter Singer's 'Animal Liberation'. *Proceedings of the Society for Experimental Biology and Medicine* **211**: 109–54

Russell W M S, Burch R L 1992/1959 *The Principles of Human Experimental Technique*. University Federation for Animal Welfare, Herts, UK

Singer P 1991/1975 *Animal Liberation*. Random House, New York

Singer W 1993 The significance of alternative methods for the reduction of animal experiments in the neurosciences. *Neuroscience* **57**: 191–200

F. Borchard

Anomie

'Anomie' comes from the Ancient Greek *anomia*, meaning absence of rules, norms, or laws (the adjective form *anomos* was more common). The word is still sometimes used in this highly general sense. In Greek thought it had the negative connotations of disorder, inequity, and impiety; and when it reappeared briefly as 'anomy' in English theological literature of the seventeenth century (Orru 1987), its meaning was somewhat similar to the Ancient Greek one. The term came back at the end of the nineteenth century in writings of French philosophers and sociologists, and the French spelling imposed itself in English-language sociology despite attempts to Anglicize it to 'anomy.' 'Anomie' was particularly popular among American sociologists in the 1960s. It is one of the few lexical inventions in sociology, perhaps the only one to have been specific to the sociological community. But sociologists used it with differing and even contradictory meanings—sometimes without any meaning at all—and its career eventually ended in total confusion.

1. Durkheim's Concept

'Anomie' was reinvented by Jean-Marie Guyau, a French philosopher with a sociological bent, in two books: *Esquisse d'une morale sans obligation ni sanction*, published in 1885, and *L'Irréligion de l'avenir*, published in 1887. Guyau opposed *anomie* to *autonomie* (Kantian antonomy). Because of progressive individualization of beliefs and criteria for ethical conduct, the morality of the future would be not only autonomous but anomic—determined by no universal law. For Guyau this process was both ineluctable and desirable.

After Guyau, the word appeared occasionally in late nineteenth-century writings of French philosophers and sociologists such as Gabriel Tarde. But it was Emile Durkheim who would inscribe 'anomie' in the vocabulary of sociology, the new discipline he intended to found.

Durkheim could not accept Guyau's anarchizing individualism since for him every moral fact consisted of a sanctioned rule of behavior. He therefore considered anomie a pathological phenomenon. In *De la division du travail social* (*The Division of Labor in Society*), his doctoral thesis, defended and published in 1893, Durkheim called anomie an abnormal form of

the division of labor, defining it as the absence or insufficiency of the regulation necessary to ensure cooperation between different specialized social functions. As examples of anomie he cited economic crises, antagonism between capitalists and workers, and science's loss of unity due to its increasing specialization. In all three of these cases, what was missing were continuous contacts between individuals performing different social roles, who did not perceive that they were participating in a common undertaking. At the end of his chapter on anomic form, Durkheim discussed another kind of pathology characteristic of industrial societies: the alienation of the worker performing an overspecialized task. This explains why a number of Durkheim's readers understood job meaninglessness to be an integral part of anomie. In fact, they introduced a 'parasite' connotation into his concept. Such alienation is not only different from anomie, it is in many ways its opposite.

Durkheim further developed the theme of anomie in *Le Suicide* (*Suicide*) (1897), defining anomic suicide as that resulting from insufficient social regulation of individual aspirations. In this connection he discussed economic anomie, caused by a period of economic expansion, and conjugal or sexual anomie, due to the instituting and spread of divorce. In both these cases, an opening up of the horizon of the possible brings about the indetermination of the object of desire—or unlimited desire, which amounts to the same thing. This leads to frustration. The anomie that follows a passing crisis may be acute, but Durkheim was especially interested in chronic, even institutionalized anomie. This type of anomie, the 'morbid desire for the infinite,' was an unavoidable counterpart of modern industrial society, residing simultaneously in its value system, e.g., the doctrine of constant progress; its institutions, e.g., divorce law; and its functioning, e.g., competition in an ever-expanding market. However, along with 'progressive' anomie, Durkheim also evoked a 'regressive' type, which actually pertained to 'fatalism.' This last notion, which Durkheim did not really develop, refers to the impossibility of internalizing rules considered unacceptable because unjust or excessively repressive. Here, desires and aspirations run up against new norms, deemed illegitimate; there is a closing down of possibilities. This situation is in fact the opposite of progressive—true—anomie, where the wide, open horizon of the possible leads to unlimited desires. Fatalism, then, is the opposite of anomie, just as altruism is the opposite of egoism (Durkheim's terms for the other types of suicide). In fact, this bipolar theory of social regulation of aspirations was not clearly laid out by Durkheim, and this no doubt explains the concept's peculiar later career: after Durkheim, 'anomie' underwent a complete semantic revolution.

Anomie was not a permanent theme in Durkheim's thought; he was more concerned with the question of social integration. Indeed, both word and concept disappeared from his work after 1901, and they are not to be found anywhere in that of his disciples or collaborators in the 'French school of sociology' and in *L'Année sociologique*. While in *Les Causes du suicide*, a purported continuation of Durkheim's study of suicide published in 1930, Maurice Halbwachs confirmed Durkheim's results on many points he either neglected or rejected everything that could possibly pertain to Durkheim's theory of anomie and he made no reference to conjugal anomie and refuted Durkheim's hypothesis of a progressive economic anomie.

Yet it was in the sociology of suicide, particularly American studies of the subject, that the Durkheimian notion of anomie would endure for a time, following the 1951 English translation of *Le Suicide*. In this connection mention should be made of Henry and Short's 1954 study, in which they sought to apply Durkheim's concept of anomie in their analysis of the effect of the economic cycle on suicide frequency. The largely negative results of this and other tests of Durkheim's progressive anomie hypothesis did not encourage sociologists to explore this research avenue further. Conjugal anomie, meanwhile, was left almost entirely aside in suicide studies, which gave primacy to Durkheim's egoistic suicide, reflecting a preference for his theory of integration over the less developed one of regulation.

2. *The Americanization of Anomie*

As with sociological studies of suicide, the future of 'anomie' was linked to American sociology's reception of Durkheim's work, particularly *Suicide*. Interest in Durkheim began to manifest itself only in the 1930s. With the notable exception of Parsons (1937), however, little attention was paid at first to 'anomie.' The term made no headway at the University of Chicago, where a tradition of research on social disorganization had developed in the 1920s. It was at Harvard University that 'anomie' underwent its American naturalization. Used by Elton Mayo (1933), Talcott Parsons (1937), and Robert K. Merton (1938), it appeared as a new label—and possibly a new theoretical framework—for the study of social disorganization and deviant behavior, and enabled the new department of sociology at Harvard to position itself in relation to the old Columbia–Chicago struggle for intellectual supremacy in the discipline. This strategic aspect may be seen in the young Harvard theoreticians' use of the French spelling (sociologists at Columbia, like Durkheim's translators, spelled it 'anomy'). Distinguishing themselves from the American tradition and appropriating the continental European one, that of Durkheim, Weber, and Pareto, they wrote 'anomie.'

Merton's successful *Social structure and anomie* in its various versions (1938, 1949, 1957, 1964) both

made the term familiar and metamorphosed the concept. In fact, Merton never clearly defined what he meant by anomie. We can nonetheless say that Merton's theory ultimately identified anomie as the contradiction between the culturally defined goal of success and the individual's lack of access to legitimate means of achieving this goal. To resolve this contradiction, the individual may engage in deviant behavior, one form of which is using illegitimate means. Merton called this mode of adaptation 'innovation.'

Merton's theory of anomie is fairly equivocal, but it was understood and used as a theory of differential access to the means to success, one which predicted that delinquent behavior was most likely among the least privileged social strata. This theory, which reached its apogee in the mid-1960s, has been considered the core of 'strain theory,' a major sociological approach to deviance. It should be noted that Merton's theory was only rarely tested in empirical research on delinquency and that when it was, the results were generally negative.

The Mertonian concept of anomie—to the extent that its content can be identified and such as it has been generally understood—is in many ways not only different from Durkheim's but opposed to it. The opposition was not perceived at the time because Merton never related his own use of the term to Durkheim's. Durkheim's anomie is about the lack of restriction on goals; Merton's anomie refers to restriction on means. In Merton's theory, goals are given, defined, even prescribed by the cultural system, whereas for Durkheim the central feature of anomie is indeterminacy of goals. Durkheim's anomic individual is uncertain of what he should do because the horizon of possibilities is so open, whereas for Merton that individual, with the clearest of ideas of the objective to be reached, finds the possibilities for success closed down.

This semantic revolution attained its full effect in American sociology between 1955 and 1970, in numerous sample surveys using attitude scales designed to measure the psychological counterpart of social anomie. This research in turn was based on theoretical essays in which anomie had appeared for the first time outside Harvard: at the end of the 1940s sociologists and political scientists such as David Riesman, Robert MacIver, and Sebastian De Grazia had psychologized the concept, using the word 'anomie' to designate individual feelings of anxiety, insecurity, and distrust. These feelings figured in 'anomie scales,' meant to measure degree of pessimism in world views, feelings of having no control over one's situation, and renunciation of hope. (The most famous of these was developed by Leo Srole in 1956 and called the 'anomia scale.') In Durkheim's thought, these phenomena were characteristic of fatalism—the opposite of anomie. Anomie had become interchangeable with the notion of alienation, which sociologists were trying to apply in attitude scales; normlessness had been associated with powerlessness. All these empirical studies using attitude scales were based on the (mistaken) postulate of a single unified concept of anomie, the myth of a continuous tradition of research, with Durkheim as precursor and Merton as prophet.

Anomie scales were contemporaneous with the other main usages of the term. As we have seen, it was between the mid-1950s and the late 1960s that 'anomie' was most heavily used, with its Durkheimian meaning in the sociology of suicide and its Mertonian meaning (mainly) in the sociology of deviant behavior. That contradictory meanings of one and the same term should have been successful at the same time shows that anomie's use function in sociology was more decorative than cognitive.

In the 1960s, anomie came to be considered the sociological concept par excellence, despite (or because of) the semantic confusion surrounding it. In the 1968 edition of the *International Encyclopedia of the Social Sciences*. Talcott Parsons explained that anomie was one of the few concepts truly central of contemporary social science. At that time the term had reached the peak of its glory. Because 'anomie' made possible a kind of bridge between references to classic theoretical sociology works and sample surveys using attitudes scales purporting to test hypotheses drawn from those theories, it served as the emblem of the dominant research practice of the time (opposed to the ecological analysis and fieldwork favored by the Chicago School).

Though the word 'anomie' has lost much of his attraction for sociologists, some still use it rather vaguely to mean deregulation or disorder. It would be preferable to restrict usage of the term to Durkheim's meaning of an excessive growth in aspirations or expectations. This concept refers to a paradox present in Tocqueville and in the notion of relative deprivation (and for which Raymond Boudon (1977) has constructed a model): when individual objective conditions improve, collective subjective dissatisfaction may increase.

See also: Alienation: Psychosociological Tradition; Alienation, Sociology of; Control: Social; Delinquency, Sociology of; Deprivation: Relative; Durkheim, Emile (1858–1917); Integration: Social; Labor, Division of; Norms; Parsons, Talcott (1902–79); Solidarity, Sociology of; Suicide, Sociology of

Bibliography

Besnard P 1986 The Americanization of anomie at Harvard. *Knowledge and Society* **6**: 41–53

Besnard P 1987 *L'anomie, ses usages et ses fonctions dans la discipline sociologique depuis Durkheim*. Presses Universitaires de France, Paris

Besnard P 1990 Merton in search of anomie. In: Clark J, Modgil C, Modgil S (eds.) *Robert K. Merton: Consensus and Controversy*. Falmer Press, London
Besnard P 1993 Anomie and fatalism in Durkheim's theory of regulation. In: Turner S P (ed.) *Emile Durkheim: Sociologist and Moralist*. Routledge, London
Boudon R 1977 *Effets pervers et ordre social*, 1st edn. Presses Univesitaires de France, Paris
Boudon R 1982 *The Unintended Consequences of Social Action*. St. Martin's Press, New York
Durkheim E 1893 *De la division du travail social*. Alcan, Paris
Durkheim E 1897 *Le Suicide. Etude de sociologie*. Alcan, Paris
Durkheim E 1933 *The Division of Labor in Society*. Free Press of Glencoe, New York
Durkheim E 1951 *Suicide. A Study in Sociology*. Free Press, Glencoe, IL
Guyau J M 1885 *Esquisse d'une morale sans obligation ni sanction*. F. Alcan, Paris
Henry A F, Short Jr J F 1954 *Suicide and Homicide: Some Economic, Sociological, and Psychological Aspects of Aggression*. Free Press, Glencoe, IL
Mayo E 1933 *The Human Problems of an Industrial Civilization*. Macmillan, New York
Merton R K 1938 Social structure and anomie. *American Sociological Review* **3**: 672–82
Merton R K 1949 Social structure and anomie. In: Merton R K (ed.) *Social Theory and Social Structure*. Free Press, Glencoe, IL
Merton R K 1957 Continuities in the theory of social structure and anomie. In: Merton R K (ed.) *Social Theory and Social Structure*. Free Press, Glencoe, IL
Merton R K 1964 Anomie, anomia and social interaction. In: Clinard M B (ed.) *Anomie and Deviant Behavior*. Free Press of Glencoe, New York
Orru M 1987 *Anomie: History and Meanings*. Allen & Unwin, Boston
Parsons T 1937 *The Structure of Social Action: A Study in Social Theory with Special Reference to a Group of Recent European Writers*. McGraw-Hill, New York
Parsons T 1968 Durkheim, Emile. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. MacMillan, New York
Srole L 1956 Social integration and its corollaries: an explanatory study. *American Sociological Review* **21**: 709–16

P. Besnard

Anthropology

Anthropology as a discipline is concerned with human diversity. In its most inclusive conception, this is what brings together the four fields of sociocultural anthropology, archaeology, biological (or physical) anthropology, and linguistics. With its formative period in the historical era when Europeans, and people of European descent, were exploring other parts of the world, and establishing their dominance over them, and when evolutionary thought was strong, it also came to focus its attention especially on what was, from the western point of view, distant in time or space—early humans or hominoids, and non-European peoples. In that early period, there was indeed a strong tendency to conflate distances in time with those in space: some living non-Europeans could be taken to be 'contemporary ancestors,' dwelling in 'primitive societies.' Understandings of the discipline have changed over time, however, and they are not now entirely unitary across the world. What is held together under one academic umbrella in one place may be divided among half a dozen disciplines somewhere else. A mapping of the contemporary state of the discipline, consequently, may usefully begin by taking some note of international variations.

1. Terminologies and Boundaries

It is particularly in North America that academic anthropology has retained what has come to be known as 'the four-field approach,' although in recent times its continued viability has come under debate. Among the founders of the discipline, some were perhaps able to work (or at least dabble) in all the main branches, but with time, in American anthropology as well, it has certainly been recognized that most scholars reach specialist skills in only one of them—even as it may be acknowledged that a broad intellectual sweep across humanity has its uses, and at the same time as it is recognized that here as elsewhere, research in the border zones between established disciplines or subdisciplines often brings its own rewards. On the whole, in any case, scholars in archaeology, biological anthropology, linguistics, and sociocultural anthropology now mostly work quite autonomously of one another, and while terminologies vary, in many parts of the world, they are understood as separate disciplines.

In Europe, varying uses of the terms 'anthropology,' 'ethnology,' and 'ethnography,' between countries and regions as well as over time, often reflect significant historical and current intellectual divides (Vermeulen and Alvarez Roldán 1995). In parts of the continent, in an earlier period, the term 'anthropology' (in whatever shape it appeared in different languages) tended to be used mostly for physical anthropology, but in the later decades of the twentieth century, it was largely taken over by what we here term sociocultural anthropology—itself a hybrid designation for what is usually referred to either as social anthropology or as cultural anthropology. (Especially in German usage, however, 'anthropology' also frequently refers to a branch of philosophy.) Physical or biological anthropology, meanwhile, was absorbed in many places by other disciplines concerned with human biology, while archaeology and linguistics maintained their positions as separate disciplines.

In some European countries, now or in the past, the term 'ethnography' has been used, unlike in present-day usage in Anglophone countries, to refer to

sociocultural anthropology as a discipline. Matters of discipline boundaries are further complicated, however, by the fact that sociocultural anthropology especially in northern, central, and eastern Europe is often itself divided into two separate disciplines, with separate origins in the nineteenth and early twentieth centuries. One, which was often originally designated something like 'folk life studies,' had its historical links with cultural nationalism, and concerned itself with local and national traditions, especially with regard to folklore and material culture. This discipline mostly did not acquire a strong academic foothold in those western European countries which were most involved in exploration and colonialism outside Europe, particularly Great Britain and France, where on the other hand, that sociocultural anthropology which focused on non-European forms of life was earliest and most securely established. The more Europe-oriented, or nationally inclined, 'folk life studies' discipline has tended, in recent decades, to redesignate itself as 'European ethnology'—or in some contexts, simply as 'ethnology,' in contrast to a rather more non-Western oriented, or at least internationalist, 'anthropology.'

In another usage, 'ethnology' has been taken to refer to a more historical and museological orientation (in contrast with what was for a time a more presentist social anthropology), while in other contexts again, it is more or less synonymous with 'sociocultural anthropology.' Yet further national variations in terminology continue to make direct transpositions of terms between languages treacherous. Meanwile, outside the heartlands of western academia, the idea that the study of non-European societies and cultures must be a separate discipline has not always been obvious and uncontroversial. In African universities, founded in the late colonial or the postcolonial period, there has often been no distinction made, at least organizationally, between anthropology and sociology—and not least in Africa, anthropology sometimes has to carry the stigma of being associated with the evils of colonialism and racism (Mafeje 1998). In India, when anthropology is recognized as a distinct discipline, it is frequently taken to be particularly preoccupied with 'tribal' populations, and perhaps with physical anthropology—while some of the scholars recognized internationally as leading Indian anthropologists, concerned with the mainstream of Indian society, may be seen as sociologists in their own country.

Back in North America and Europe, again, the framework of academic life is not altogether stable over time. 'Cultural studies,' putting itself more conspicuously on the intellectual map from the 1970s onwards, has been most successful as a cross-disciplinary movement in the Anglophone countries but has had an impact elsewhere as well. Its center of gravity may have been mostly in literary and media studies, but insofar as it has engaged with methods of qualitative field research and with issues of cultural diversity, not least in the areas of multiculturalism and diaspora studies, it has sometimes come close to sociocultural anthropology and—in certain European countries—to ethnology, as discussed above. While some anthropologists may see this as an undesirable intrusion into their disciplinary domain, others see it as a source of new stimuli (Dominguez 1996, Nugent and Shore 1997).

2. *Conceptions of the Core*

If a concern with human diversity may define anthropology as a whole, it is perhaps too airy a notion to offer a strong sense of the distinctive working realities of the anthropological practitioner. In discussing the changing assumptions of what is the core of anthropological work and thought, we will concentrate on sociocultural anthropology—by now the segment of the wider field which remains, internationally, most clearly identified as anthropology.

Obviously the tradition of studying more exotic, non-Western forms of life has remained strong, and some would still maintain that this is what a 'real anthropologist' does. Yet it is a conception of anthropology which is now, in its own way, susceptible to charges of 'Eurocentrism,' with uncomfortable and controversial links to a past in which the divide between a dominant West and a subjugated 'Rest' was more likely taken for granted, at political, intellectual, and moral levels. By the latter half of the twentieth century, past vocabulary such as 'primitive societies' had become mostly an embarrassment, and in a more egalitarian mood, some anthropologists had taken to arguing that it was quite acceptable for North American and European anthropologists to go to Africa, Asia, or Oceania, as long as anthropologists from these regions were also welcomed to do research in the Occident, in a generous worldwide intellectual exchange effort. Such a symmetrical exchange, however, has hardly occurred. The relatively small number of professional anthropologists originating in non-Western, non-Northern parts of the world have seldom had the inclination, or the funding, to do their research abroad. They have been much more likely to conduct their studies in their home countries—where they may thus sometimes share their fields with expatriate anthropologists from Europe or North America (Fahim 1982).

In the latter part of the twentieth century, moreover, the emphasis on studying exotic others—in what had now become the Third World, or the Fourth World—was weakening in American and European sociocultural anthropologies as well. In some parts of Europe where the discipline was relatively late in getting established, most anthropologists from the very beginning did their research in their own countries, even as they drew intellectual inspiration from

the classic studies of exotics imported from older national anthropologies. The growing interest in, and increasing legitimacy of, 'anthropology at home' has been based partly on a sense that humanity is indivisible, and an academic segregation of the West from the Rest indefensible. The preoccupation with exploring human diversity is to a certain extent retained by emphasizing the internal diversity of forms of life (lifestyles, subcultures, etc.) of contemporary societies. The exotic may be around the corner. Moreover, the study of one's own society is often motivated by a sense of relevance: one may be better placed to engage actively with whatever are perceived to be the wrongs of the place where one has the rights and the duties of a citizen. Of course, such a repatriation of sociocultural anthropology may blur certain disciplinary boundaries–perhaps with sociology, perhaps in some contexts with an 'ethnology' of the European type, which could also identify itself as an 'anthropology at home.'

If anthropology is no longer quite so committed to identifying itself in terms of exotic fields, its methodology appears, in some views, to offer another sense of distinctiveness. Anthropologists, since at least the early twentieth century, have typically done 'field work.' They emphasize 'participant observation,' or 'qualitative research'—in contrast with the handling of more or less impersonal statistical data—or 'ethnography,' in the sense of integrated descriptions of ways of life. In this discipline, the pure theorist is an anomaly, if not entirely nonexistent. The direct involvement with another way of life, whether it is in a village on the other side of the world, or a neighborhood across town, or another occupation, tends to become more than a methodological choice. It becomes a central personal experience, a surrender with strong moral and esthetic overtones and a potential for rich satisfaction and life-long memories.

Even so, placing field work at the core of the distinctiveness of anthropology is also problematic. In part, the problem is that 'doing ethnography' has become increasingly common in various other disciplines as well, even if anthropologists often complain that it is then done badly, superficially. It is also true that field work of the most orthodox variety does not fit readily into every setting. In all field studies, anthropologists tend to do not only participant observation, but other kinds of work as well. They talk to informants, elicit life histories, collect texts, do surveys, and engage in some variety of activities to acquire new knowledge. To what extent their observational work can really be actively participatory must depend a great deal on the contexts. But furthermore, in contemporary life observational work is sometimes not very rewarding. Following the daily round of an agricultural village is one thing; observing an office worker, or even a creative writer, at his desk, in front of his computer screen is another. On the other hand, fields in the present-day world may involve other kinds of data than those of classic anthropology—more media use, for example. Notions of 'field' and 'field work' are increasingly coming to be debated in anthropology, as the world changes and as anthropologists try to fit their pursuits into it, and there are clearly also anthropologists who prefer less distinctively anthropological methodological repertoires (cf. Gupta and Ferguson 1997, Bernard 1998).

The idea of gaining knowledge and understanding from field work is certainly not entirely separate from the issues of where, and among whom, anthropologists work either. While anthropologists have been inclined to assume, or argue explicitly, that there are special insights to be gained from combining an outsider view with immersion in another way of life, the rise of 'anthropology at home,' the increasingly frequent copresence of indigenous and expatriate anthropologists in many fields, and the growth elsewhere in academic life of more emphatically insider-oriented fields of study, such as some varieties of ethnic studies, leave assumptions and arguments of this kind more clearly open to debate. Some would question as a matter of principle what kind of validity an outsider's perspective can have. Others could argue that one can never be an anthropologist and an insider at the same time, as these are distinct intellectual stances. And then, surely, such issues are further complicated by the fact that insider and outsider are not necessarily either-or categories, and hardly altogether stable either. An outsider can perhaps become an insider—and on the other hand, one can probably start as an insider and drift away, and become an outsider. Here again is a complex of problems which remains close to the heart of anthropological work (Merton 1972, Narayan 1993).

One may wish to object to the somewhat unreflected inclination to put field work at the center of the discipline because it would tend to make anthropology primarily a methodological specialty—perhaps, because of its qualitative emphasis, the counterpart of statistics. Undoubtedly many anthropologists would prefer to identify their discipline with particular ideas, theories, and intellectual perspectives, and the history of anthropology is most often written in such terms. It may still turn out to be difficult to point to any single, uncontested, enduring central concept, or structure of concepts.

Again, however, there is the exploration of human diversity. A key idea here has obviously been that of 'culture,' not least in the plural form of 'cultures.' Most fundamentally, in its conventional but multifaceted anthropological form, 'culture' stands for whatever human beings learn in social life, as contrasted with whatever is inborn, genetically given. But with the idea that human beings are learning animals also goes the understanding that they can learn different things, so that what is cultural tends to exhibit variation within humanity. Moreover, the tendency has been to conceptualize culture at a level of

human collectivities (societies, nations, groups), so that members are held to be alike, sharing a culture, while on the other hand there are cultural differences between such collectivities. Culture, that is, has been taken to come in distinct packages—culture*s*. In the classic view of culture, there has also been a tendency to try and see each such culture 'as a whole'—to describe entire ways of life and thought, and to throw light on the varied interconnections among ideas and practices.

Such a basic notion of culture has allowed anthropologists to proceed with the task of drawing a panoramic map of human diversity. There has certainly also been a widespread preoccupation with showing how much of human thought and behavior is actually cultural (that is, learned) and thus variable, rather than altogether biologically based, and uniform. This emphasis has been visible, for example, in anthropological contributions to gender studies.

Yet culture has also been a contested concept in anthropology, not least in recent times. It is true that it was never ally central to anthropological thought in all varieties of anthropology. American sociocultural anthropology, which has more often been identified as 'cultural anthropology,' has focused rather more on it, for example, than its British counterpart, more inclined over the years to describe itself as 'social anthropology' (and perhaps now offering 'sociality' as an emergent alternative or complementary core concept). By the late twentieth century, however, critics in the United States as well as elsewhere argued that the established style of cultural thought tended to exaggerate differences between human collectivities, to underplay variations within them, to disregard issues of power and inequality and their material bases, and to offer overly static, ahistorical portrayals of human ways of life. While some anthropologists have gone as far as to argue for the abolition of the culture concept, others would be inclined to accept many of the criticisms, and yet take a more reformist line. Whichever view is taken, it seems that debates over the understanding of culture offer the discipline one of its lively intellectual foci (Hannerz 1993, Kuper 1999).

The notion of cultural translation has also had a part in defining anthropological work, and inevitably it is drawn into the arguments about the culture concept generally. Indeed, bridging cultural divides by making the ideas and expressions of one culture understandable in terms of another has been a major anthropological activity, although this linguistic analogy is certainly too limiting to describe all anthropology (Asad 1986, Pálsson 1993). In a somewhat related manner, it has gone with the interest in human diversity to describe anthropology as a discipline centrally involved with comparison. In a general way, this is clearly valid. Much of anthropology is at least implicitly comparative, in its inclination to emphasize what is somehow notably different about the ideas, habits, and relationships of some particular population. To the extent that the origins of the discipline have been European and North American, the baseline of such comparisons and contrasts has no doubt been in a general way Western. At the same time, there has been a strong tendency to use the knowledge of diversity to scrutinize Western social arrangements and habits of thought to destabilize 'common sense.' In such a way, anthropological comparison can serve the purpose of cultural critique, and this possibility has again been drawing increasing attention in recent times (Marcus and Fischer 1986). On the other hand, the form of large-scale comparative studies, focusing on correlations between particular sociocultural elements across a large number of social units, which was prominent in the discipline in some earlier periods, has not been a central feature for some time, and the intellectual and methodological assumptions underlying comparison are also a more or less continuous topic of argument (Köbben 1970, Holy 1987).

3. *Specialties and Subdisciplines*

If culture in the singular and plural forms, field work, translation, and comparison may count as ideas and practices which hold anthropology together, not in consensus but often rather more through engagement in debate, it is often suggested that the discipline also tends to be systematically segmented along various lines—not only in terms of what is understood to make up the discipline, as noted above, but also with regard to what draws together closer communities of specialists. One of the three major dimensions here is area knowledge. Anthropologists tend to be Africanists, Europeanists, Melanesianists, or members of other region-based categories. Relatively few of them ever commit themselves to acquiring specialized knowledge of more than one major region to the extent of doing field research there. It is true that if field work is often highly localized, it does not necessarily lead to a wider regional knowledge either, but as a matter of convention (perhaps at least to fit into normal job descriptions in the discipline), the tendency has been to achieve specialist status by reaching toward an overview of the accumulated anthropological knowledge of some such unit, and perhaps seeking opportunities to familiarize oneself with more of it through travel. In countries where 'area studies' have been academically institutionalized, such regional anthropologies have had a significant role in the resulting interdisciplinary structures, and generally, shared regional specialization has often been important in scholarly exchanges across discipline boundaries.

The second major dimension of specialization can be described as topical. Although anthropology has a traditional orientation toward social and cultural 'wholes,' there is yet a tendency among many anthropologists to focus their interest on particular kinds

of ideas, practices, or institutions. Frequently, such specializations have tended to follow the dominant dividing lines between other academic disciplines—there have been a political anthropology, an economic anthropology, a psychological anthropology, an anthropology of law, an anthropology of art, and an ecological anthropology, for example. Drawing on knowledge of social and cultural diversity, one aspect of such specialization has been to scrutinize and criticize concepts and assumptions of the counterpart disciplines with respect to their tendencies toward Western-based intellectual ethnocentrism; but obviously there is also a continuous absorption of ideas from these other disciplines. Here, then, are other interdisciplinary connections. It may be added that such subdisciplines often have their own histories of growth in some periods, and stagnation in others (Collier 1997).

Third, anthropologists sometimes specialize in the study of broad societal types—hunter-gatherers, peasants, pastoralists, fishermen, and so forth. In some ways urban anthropology may be seen as a similar kind of specialty. In fields like these, too, intensities of collective engagement and intellectual progress have tended to vary in time.

4. *The Practical Uses of Anthropology*

It has occasionally been argued that anthropology is less engaged with practical application than many other academic disciplines, and more concerned with achieving a somewhat lofty overview of the human condition, in all its variations. One factor underlying such a tendency may be a sort of basic cultural relativism—it may go with the acceptance, and even celebration, of human diversity to be somewhat sceptical of any attempt to impose particular arrangements of life on other people. This stance may well be supported by the fact that anthropologists have often been outsiders, working in other societies or groups than their own, and feeling that they have no mandate for meddling, or for that matter any actual realistic opportunity.

Nonetheless, there have always been a number of varieties of 'applied anthropology.' In an early period, when the study of non-Western societies was carried out in contexts of Western colonialism, it was held that anthropological knowledge could be useful in colonial administration. While some administrators had some anthropological course work before taking up their overseas assignments, and while some anthropologists did research oriented toward such goals, it seems that such connections remained rather limited, and administrators were often impatient with the anthropologists' preference for a long-term buildup of basic knowledge. Particularly toward the end of the colonial period, anthropologists also often saw such involvements as morally and politically questionable.

In the postcolonial era, anthropological knowledge has been engaged in a larger scale, and in many different ways, in development work in what had now turned into 'the Third World.' Many governments, international agencies, and international nongovernmental organizations have thus drawn on anthropological advice, and have offered arenas for professional anthropological activity outside academia. At the same time, with 'development' turning into a global key concept, around which an enormous range of activities and organizations revolve, this again has become another focus for critical theoretical scrutiny among anthropologists (Dahl and Rabo 1992, Escobar 1995).

When anthropologists work in their own countries, the range of applications may be wider, and as already noted, practical and political relevance is frequently one reason for doing anthropology 'at home.' Not least as transnational migration has made more societies increasingly ethnically and culturally heterogeneous, anthropologists have been among the social scientists engaged, in one way or other, in the handling of minority affairs and multiculturalism. Educational and medical anthropology are not always involved in practical application, and research in such fields is also carried out in other societies than the researcher's own, but often work in these subdisciplines similarly centers on the encounter between major institutional complexes and culturally heterogeneous populations. Some number of anthropologists in Europe and North America now also make use of their disciplinary perspectives as specialists on organizational culture, and as marketing analysts. A further practical field which emerged in the late twentieth century as a profession in itself is that of 'intercultural communication,' which offers training and consultancy in the handling of concrete situations of cultural difference, and, putting it somewhat dramatically, of 'culture shock.' This field has links to several academic disciplines, but to a considerable degree it draws on a more or less anthropological conception of culture—of a type, it may be added, that many anthropologists might now find rather old-fashioned and clumsy.

5. *The Future of Anthropology*

In all its varying shapes, in space and over time, anthropology has tended to straddle conventional academic classifications of disciplines. In its scope of subject matter—family and kinship, politics, market and exchange, for example, on the one side; art, music, and dance on the other—it extends across the social sciences and the humanities. Insofar as it has to take into account what are the biological givens of human thought and action, and inquires into the interactions of humankind with its natural environment, it reaches into the natural sciences as well. But its multiple affiliations are not only a consequence of its varied

subject matter. They are also implied in the variety of intellectual approaches: in field research, in theoretical work, and in styles of presentation.

In many ways the enduring characteristics of anthropology, throughout this range of forms, continue to be expressions of the concern with diversity, with the highly varied manners of being human. To the global public stock of ideas it brings such notions as taboo, witchcraft, cargo cults, totemism, or the potlatch exchange feasts of Northwest Coast American Indians. Concepts such as 'Big man' (out of Melanesia), or 'patron–client relationships' (not least out of the Mediterranean area), or 'caste' (out of India), allowed to travel out of their areas of origin, can enrich our thinking about power, politics, and inequality in many contexts. There is a rich intellectual universe here, to be drawn on within the discipline and from outside it. And anthropology has its classic preoccupations, such as ritual or kinship, concerning which new materials about yet more variations are continuously gathered worldwide, and around which theoretical debates never seem to cease.

At the same time, anthropology goes on reconfiguring itself (cf. Wolf 1964, Hymes 1972, Fox 1991, Ingold 1996). One might perhaps have thought that a discipline revolving around the diversity of human forms of life and thought would find itself in difficulty at a time when increasing global interconnectedness may lead to cultural homogenization and a loss of local or regional cultural forms. Indeed, it is true that many people in the world no longer stick to the beliefs they used to hold, and are discarding some of their past practices, ranging from spirit possession to headhunting. In part, the responsibility of anthropology here becomes one of preserving a record of the ways of being humans, past and present, and keeping that record alive by continuing to scrutinize it, interpret it, and bringing it to bear on new developments.

It is also true, however, that diversity is not diminishing as much as some, perhaps fairly superficial, views of globalization might suggest. Traditions may be remarkably resilient, adapting to new influences and absorbing them in various ways; there may be more than one 'modernity.' Moreover, new interconnections may generate their own social and cultural forms, so that there may be cultural gain as well as cultural loss. A growing interest in anthropology with such notions as hybridity and creolization bears witness to this. In addition, however, anthropology has continuously expanded the field of social and cultural variations with which it has actively engaged. As the discipline was practiced in the earlier decades of the twentieth century, one might have discerned a tension between on the one hand the ambitious claims of offering a view of humanity, and on the other hand the actual concentration on villages of horticulturalists, bands of hunter-gatherers, or other exotic and small-scale sociocultural arrangements. But then by midcentury, there was an increasing involvement with peasant societies, and the civilizations of which they were a part, and later yet urban anthropology appeared as another subdiscipline, practiced in field settings in every region. In recent times, the anthropology of science has been another growing specialization, as the practices of scientists are scrutinized as yet another kind of cultural construct, and as laboratories are added to the range of possible sites of field work (Downey and Dumit 1997). It appears, consequently, that the enduring wider claims of the discipline to an overall engagement with human modes of thought and action are increasingly being realized.

Obviously, in many of its fields of interest, anthropologists now mingle with scholars from a variety of other backgrounds, and the division of labor between disciplines here may not seem obvious. Undoubtedly there are blurred boundaries as well as cross-fertilization, but anthropologists would emphasize the intellectual potential of a conceptual apparatus trained on, and informed by a knowledge of, a great variety of cultural assumptions and institutional arrangements. There is also the commitment to close-up observation of the relationships between human words and deeds, and the strain toward understanding 'wholes' which, despite its vagueness, may usefully involve a particular commitment to contextualization as well as skills of synthesis.

Anthropologists now find themselves at work in a global ecumene of increasing, and increasingly polymorphous, interconnectedness. This is a time in which that diversity of social and cultural forms with which they are preoccupied is constantly as well as rapidly changing, and where new social, political, economic, and legal frameworks encompassing and rearranging that diversity are emerging. More people may have access to a larger part of the combined human cultural inventory than ever before; conversely, whether one likes it or not, more of that cultural diversity can also come in one's way. This is also a time when debates over the limits of diversity are coming to new prominence, for different reasons. Evolutionary biologists are setting forth new views of human nature which need to be carefully confronted with understandings of cultural variation. As people sense that they live together in one world, questions also arise over what is, or should be, shared humanity, for example in the area of human rights (Wilson 1997). There would seem to be a place in the public life of this era for a cosmopolitan imagination which both recognizes diversity and seeks for the ground rules of a viable and humane world society. For such a cosmopolitan imagination, one would hope, anthropology could continue to offer materials and tools.

See also: Anthropology and History; Anthropology, History of; Cultural Critique: Anthropological; Cultural Relativism, Anthropology of; Culture: Contemporary Views; Ethnography; Ethnology; Field

Observational Research in Anthropology and Sociology; Fieldwork in Social and Cultural Anthropology; Hunting and Gathering Societies in Anthropology; Modernity: Anthropological Aspects; Primitive Society; Psychological Anthropology; Qualitative Methods, History of; Thick Description: Methodology; Tradition, Anthropology of; Tribe

Bibliography

Asad T 1986 The concept of cultural translation in British social anthropology. In: Clifford J, Marcus G E (eds.) *Writing Culture: The Poetics and Politics of Ethnography: A School of American Research Advanced Seminar*. University of California Press, Berkeley, CA

Bernard H R (ed.) 1998 *Handbook of Methods in Cultural Anthropology*. AltaMira Press, Walnut Creek, CA

Collier J F 1997 The waxing and waning of 'subfields' in North American sociocultural anthropology. In: Gupta A, Ferguson J (eds.) *Anthropological Locations: Boundaries and Grounds of a Field Science*. University of California Press, Berkeley, CA

Dahl G, Rabo A (eds.) 1992 *Kam-Ap or Take-Off: Local Notions of Development*. Stockholm Studies in Social Anthropology, 29. Almqvist and Wiksell International, Stockholm

Dominguez V R 1996 Disciplining anthropology. In: Nelson C, Gaonkar D P (eds.) *Disciplinarity and Dissent in Cultural Studies*. Routledge, New York

Downey G L, Dumit J (eds.) 1997 *Cyborgs & Citadels: Anthropological Interventions in Emerging Science and Technologies*. School of American Research Press, Santa Fe, NM

Escobar A 1995 *Encountering Development: The Making and Unmaking of the Third World*. Princeton University Press, Princeton, NJ

Fahim H (ed.) 1982 *Indigenous Anthropology in Non-Western Countries*. Carolina Academic Press, Durham, NC

Fox R G (ed.) 1991 *Recapturing Anthropology: Working in the Present*. School of American Research Press, Santa Fe, NM

Gupta A, Ferguson J (eds.) 1997 *Anthropological Locations: Boundaries and Grounds of a Field Science*. University of California Press, Berkeley, CA

Hannerz U 1993 When culture is everywhere: reflections on a favorite concept. *Ethnos* **58**: 95–111

Holy L (ed.) 1987 *Comparative Anthropology*. Blackwell, Oxford, UK

Hymes D (ed.) 1972 *Reinventing Anthropology*. Pantheon, New York

Ingold T (ed.) 1996 *Key Debates in Anthropology*. Routledge, London

Köbben A J F 1970 Comparativists and non-comparativists in anthropology. In: Naroll R, Cohen R (eds.) *A Handbook of Method in Cultural Anthropology*. Natural History Press, Garden City, NY

Kuper A 1999 *Culture: The Anthropologist's Account*. Harvard University Press, Cambridge, MA

Mafeje A 1998 Anthropology and independent Africans: suicide or end of an era? *African Sociological Review* **2**(1): 1–43

Marcus G E, Fischer M M J 1986 *Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences*. University of Chicago Press, Chicago, IL

Merton R K 1972 Insiders and outsiders: a chapter in the sociology of knowledge. *American Journal of Sociology* **78**: 9–47

Narayan K 1993 How native is a 'native' anthropologist? *American Anhropologist* **95**: 671–86

Nugent S, Shore C (eds.) 1997 *Anthropology and Cultural Studies*. Pluto Press, London

Pálsson G (ed.) 1993 *Beyond Boundaries: Understanding, Translation, and Anthropological Discourse*. Berg, Oxford, UK

Vermeulen H F, Alvarez Roldán A (eds.) 1995 *Fieldwork and Footnotes: Studies in the History of European Anthropology*. Routledge, London

Wilson R A (ed.) 1997 *Human Rights, Culture and Context: Anthropological Perspectives*. Pluto Press, London

Wolf E 1964 *Anthropology*. Prentice-Hall, Englewood Cliffs, NJ

U. Hannerz

Anthropology and History

For anthropology, history is not one but many things. First, it is the past, and especially the past that survives in archives and other written or oral records; 'prehistory' is its more remote counterpart. Second, it is change, diachronic as opposed to synchronic process. Third, history is a domain of events and artifacts that make manifest systems of signification, purpose, and value, the domain of human action. Fourth, it is the domain of all the diverse modes of the human experience and consciousness of being in time. Last, it is a domain encompassing all those practices, methods, symbologies, and theories that human beings—professional academic historians among them—have applied to the collection, recollection, and comprehension of the past, the present, and the relations between the two.

1. Anthropology and Natural History

At the end of the sixteenth century, anthropology emerged in Europe not in contrast to history but rather within it. Thence, and for some two and a half centuries forward, it would be understood broadly as that branch of 'natural history' that investigated the psychophysical origins and diversification of the human race—or races, as was very often the case. Demurring throughout the period to the theological calculus of the creation of Adam, it confined itself to treating developments presumed to have transpired over only a few millennia. In 1858, miners at England's Brixham cave unearthed tools and other human remains that stratigraphers could prove to be at least 70,000 years old. Theological authority suffered a blow from which it would not recover; anthropological time suddenly acquired much greater archaeological depth. Meanwhile, a growing scholarly coalition was coming together to support the doctrine that 'culture' was that common human possession which made manifest the basic psychic unity of all the putative races of mankind. In 1854, James Pritchard would accordingly launch 'ethnology,' and send the racialists

off on their separate ways (Stocking 1987). 'Cultural' and 'physical' anthropologists would never again keep their earlier company. Yet, even through the turn of the twentieth century, natural history remained the largely uncontested source of the methods and aims of both. Questions of origin and development continued to have pride of place. The 'savage' or the 'primitive' became all the more entrenched as a disciplinary preserve, but also as the rudimentary pole of any number of ambitious reconstructions of the probable steps or stages that had marked the human passage to 'civilization' or 'modernity.' Tylor's *Primitive Culture* (1871), Morgan's *Ancient Society* (1877), and the several volumes of Frazer's *Golden Bough* (1990) are classic examples of the genre.

Such far-reaching treatises would strike the more meticulous of the subsequent generation of their readers as undisciplined, even whimsical. By the 1920s, and despite all their other differences, Franz Boas, Bronislaw Malinowski, and A. R. Radcliffe-Brown were inaugurating a turn away from 'speculative history' in favor of meticulous observational attention to the 'ethnographic present.' Not even these sober empiricists were, however, opposed in principle to reconstructive or evolutionary analysis.

With due regard for rigor, the more adventurous of their colleagues would continue to pursue it; and virtually every subfield of anthropology has contributed to it ever since. Physical anthropology is now a 'genetic' science in both the larger and the stricter sense of the term. Morgan's project survives most explicitly in the US, from Leslie White to Marvin Harris, as 'cultural materialism' (see Harris 1968), but also endures implicitly in the technological determinism that informs Goody's argument in *The Domestication of the Savage Mind* (1977). Tylor and Frazer are the precursors not simply of Claude Lévi-Strauss's structuralism (see infra) but also of the burgeoning interdisciplinary and neo-Darwinist vocation known as 'evolutionary psychology.' Boas himself is among the bridges between an older 'comparative philology' and ongoing efforts to trace the family tree of all the world's languages (see Kroeber 1935).

2. *Anthropology and Ethnohistory*

A more modest anthropological tradition of diachronic research has borrowed its methods less from natural history than from empirical historiography. It has a partial foreshadowing in the particularistic study of the drift and dissemination of traits and artifacts from the centers to the peripheries of cultural production with which diffusionists in both Germany and the US were occupied between the 1890s and the 1930s. It has its more definitive commencement in the immediate aftermath of the Second World War. Its most familiar designation is still 'ethnohistory,' however misleading the suggestion of parallels with 'ethnoscience' or 'ethnomethodology' may be. In any event, the signature task of ethnohistory has always been the investigation and documentation of the pasts of those native or 'first' peoples whom anthropologists had, until recently, proprietarily or conventionally claimed as 'their own.'

In the US, its more concrete initial impetus came with the 1946 ratification of the Indian Claims Act, which soon led to anthropologists serving as expert witnesses—sometimes for the plaintiffs, sometimes for the defense—in the readjudication of the treaties of the pioneer era. In the US and elsewhere, it had a crucial catalyst in the granting of public access to the administrative archives of the pioneers, the missionaries, and the bureaucrats of European colonization. Hence, its characteristic focus: the dynamics of contact and conflict between the subaltern and their would-be overlords: pioneering, missionary, colonizing, or enslaving (Cohn 1968, 1980).

Narrowly delimited, ethnohistory remains a specialist's craft. Since the 1970s, however, its methods and its themes have met with an ever widening embrace, and if 'historical ethnography' and 'historical anthropology' are not yet synonymous with standard disciplinary practice, they are certainly of a piece with it. The anthropological gaze is now less often 'from afar' than it is longitudinal. Such a vantage has been pivotal in the renovation of political economy. It has also brought fresh and stimulating perspectives to the address of kinship, race, national and personal identity, and gender. Its deployment and its impact may or may not be indications of greater disciplinary enlightenment, but they are by no means indications of passing intellectual fashion alone. The lengthening of the anthropological gaze has rather gone hand in hand with the ascendance of a postcolonial order in which social and cultural boundaries have become increasingly permeable and structures increasingly indistinguishable from processes. It has gone hand in hand as well with the ascendance of the postcolonial demand that anthropology offer a reckoning, not simply of its relation to the colonial past but also of the status of the knowledge that it claims to produce.

3. *Anthropology and Hermeneutics*

In our postcolonial order, anthropology itself needs interpreting; so, too, do many other sociocultural phenomena. Anthropology's current troupe of interpreters espouses, however, an even stronger postulate: that sociocultural phenomena, as historical phenomena, permit only of interpretation; that they permit of contextual understanding, but not of general explanation. 'Interpretive anthropology,' thus, stands starkly at odds with the loftier versions of the natural history of culture, but also with any empirical historiography that has the inductive abstraction of general types or causal relations as its end. It now comes in many versions of its own. The most venerable

of them commences with the Boasians. In 1935, Alfred Kroeber would accordingly remark that the ethnographies that Ruth Benedict and his other colleagues were busy producing were 'historical' in type. True enough, their temporality was synchronic, not diachronic. Their mode, however, was particularistic; their method contextualist; their analyses rarely if ever causal; and their model consequently that of what Wilhelm Dilthey had defined as the *Geisteswissenschaften*—'sciences of mind,' literally, but better glossed as 'sciences of meaning' or simply 'hermeneutics.' In fact, Kroeber was quite correct; Benedict, Margaret Mead, and the other cultural anthropologists of Boas's circle were indeed hermeneuticians. They were hence establishing the methodological legacy to which Geertz is the most celebrated heir (see Geertz 1973).

The more proximate wellspring of contemporary interpretation flows out of the tumult of the later 1960s. Elaborating in 1972 upon the call for a 'critical and reflexive anthropology,' Scholte was among its earliest manifestologists, though he borrowed many of his own philosophical and methodological tenets from Johannes Fabian's slightly earlier exposition of Hans-Georg Gadamer's revision of Dilthey's thought (Scholte 1972, Fabian 1971, 1983). The channels thus opened have remained critical and reflexive, if perhaps not always so politically committed as Scholte might have hoped. Gadamer's heremeneutics has, moreover, served as only one of many subsequent grounds on which to establish interpretive license. Bypassing Gadamer, Bourdieu (1991) has found a cardinal inspiration for his own program of reflexive critique in Martin Heidegger. Walter Benjamin (see Taussig 1993) and Giambattista Vico (see Herzfeld 1987) have won admirers of their own. Among the French poststructuralists, Bourdieu himself has attracted a substantial following, especially in his own country; Michel Foucault has had the greater impact on both reflexion and critique in the US and elsewhere.

Such a well populated census might suggest that, as much now as at its beginnings, anthropology belongs to history (as a discipline) no less than in it (as contingent process). Yet, even many of those who label themselves 'historical anthropologists' or 'interpreters' of one or another stripe would object to such a subsumption. No doubt, professional territorialism plays a part in their resistance. Often very palpable differences of professional sensibility play another part. Matters more strictly intellectual, however, play a part of their own, and their stakes are no more evident than within the anthropology of history itself.

4. *The Anthropology of History*

Durkheim's *Elementary Forms of the Religious Life* (1965) opens the arena of the anthropology of history with its argument for the social causation of the experience and conceptualization of time. Halbwachs's *Collective Memory* (1980) expands it, and after many decades of neglect, has come to be among the foundational texts of a recent surge of ethnographic and comparative inquiry into the techniques, the media, and the politics of remembering—and forgetting (see, e.g., Boyarin 1994, Shryock 1997). The most imposing precedent of the broader anthropology of history is, however, Lévi-Strauss's *Savage Mind* (1966). A treatise devoted to the analysis of the analogical—and ahistorical—matrices of mythic and totemic thought, *The Savage Mind* concludes with an extended polemic against the 'historical, structural anthropology' that Jean-Paul Sartre had advocated in the introduction to his *Critique of Dialectical Reason*. Sartre had no time for totemists.

His anthropology would instead rest with the charting, and the heightening, of 'historical consciousness.' Against it, Lévi-Strauss had two general rejoinders. The first was that 'historical consciousness' was the expression not of dawning wisdom but instead of a collective devotion to 'development,' and its absence not the expression of error but instead of a collective devotion to homeostasis. Some societies ran historically 'hot'; others ran 'cold.' All were equally human. The second was that history—as narrative, as diachronic interpretation—was always 'history-for,' always biased. Stripped of its bias, it amounted to nothing more than the methodical application of temporal scales of measure to the flow of human and nonhuman events alike. If not that, then it amounted simply to the preliminary 'cataloging' with which any 'quest for intelligibility'—the anthropological quest included—had to begin. But it could be a beginning only: 'as we say of certain careers, history may lead to anything, provided you get out of it' (Lévi-Strauss 1966).

In the face of much criticism, Lévi-Strauss has granted that his division between 'hot' and 'cold' societies was in need of much refinement. He has not, however, retreated from his division between historical and properly anthropological knowledge. On the contrary: from the outset of his career, he has counted history consistently among those disciplines limited to the merely statistical representation and analysis of their objects. His anthropology is for its part a model-theoretic discipline, an axiomatic and deductive science. Its object remains what it was for Tylor: the psyche. Its quest, however, culminates not in the hypothetical reconstruction of the path toward enlightened modernity but instead in the formal exegesis of the universal 'grammar,' the structural and structuring properties, of the mind itself. 'Structuralism,' thus construed, is far less influential than it was in the 1960s, but by no means bereft. In cognitive anthropology, it has its most secure contemporary home; and there, history (as intellectual or epistemological paragon) continues to meet with a cool reception.

For Lévi-Strauss as for other philosophical and scientific rationalists, there is an insuperable gap between 'the facts' and their theoretical intelligibility. For positivists and empiricists, the relation between facts and theories is putatively more seamless. Unabashedly positivist anthropologists are a rather rare breed at present, at least in the sociocultural field, though many cultural materialists and evolutionary psychologists might quietly reckon themselves as such. The positivist anthropology of historical consciousness or the 'historical sensibility' is rare indeed, but has a singularly unabashed spokesperson in Donald Brown. In *History, Hierarchy, and Human Nature* (1988), Brown offers a cross-cultural survey of those traits—from divination to record-keeping—most suggestive of a preoccupation with the meaning and significance of events. Among literate peoples, he discerns a relatively stable correlation: between the presence and prominence of such a preoccupation and the absence of caste or other fixed hierarchies. He concludes that history (as sensibility, as worldview, and as mode of inquiry) takes its most regular nourishment from an ideology of social mobility. Whether or not correct, the conclusion is compatible with Lévi-Strauss's own considered judgments. Here, too, the anthropologist takes history as his analytical object.

Sahlins offers an alternative, which also takes history as its analytical object—but further as its analytical mode. It preserves the structuralist principle that systems of signification are never mere derivatives of their social or material environment. Yet it casts them less as revelations of the grammar of the psyche than as interfaces or differentials between the past and the future of given social practice. Their effect is threefold. First, they determine the internal historical 'temperature' of practice, which is relatively colder when governed by prescriptions, warmer when not. Second, they vary in their capacity to accommodate potentially anomalous or disruptive events. They are in other words more or less historically 'sensitive,' and the greater their sensitivity, the more the continuity of practice is itself at risk. Finally, they influence the symbolic 'weight' or import of actors and their acts. Where some men are 'kings,' and their authority unchallenged, the historian is right to monitor them with especial care; where democracy holds sway, they would do better to monitor status groups or classes or parties. Hence, the best historian should be a good anthropologist, and the best anthropologist always prepared to be a good historian as well (Sahlins 1985).

Sahlins's standard of goodness is still the standard of objective accuracy. It thus stands in partial contrast to the standard that has at least implicitly guided interpretive anthropology since the Boasian 'golden age.' Though subject to diverse formulations—some more vividly critical than others—the latter standard is practical or pragmatic, a matter of consequences. Perhaps for a majority of interpretive anthropologists, it has been nothing short of 'liberation'—whether from psycho-sexual repression, as for Benedict and Margaret Mead, or from political domination and economic exploitation, as for Scholte. Many of its prominent inflections remain reformist, though of more qualified scope. For the Geertzian, however, the pragmatic proof of an interpretation lies in its facilitating conversation, its translational efficacy.

For others, it lies in a broadening or enrichment of our imagination of the ways of being human. For a few others still, it lies in therapeutic release—perhaps from prejudice, perhaps from alienated isolation. Interpretation must in every case be informed factually; but at its best, is always also 'informative.' Its analytical mode is historical; but however much Lévi-Strauss or Sahlins might disapprove, it is always also 'history-for' and 'anthropology-for' alike. The interpretive history of anthropology and the interpretive anthropology of history must, moreover, occupy the same epistemological plane. For the interpreter, historical and anthropological knowledge are of precisely the same kind.

History, then, is not simply a thing of many refractions. It is also a thing of plural and incompatible anthropological estimations. These in turn are among the most telling indices of plural and incompatible visions of the basic enterprise of anthropology itself. One can hence condemn the discipline for its intellectual indecisiveness or incoherence. Or one can applaud it for its perspectival diversity. One can in any event note that its many byways have a common point of departure—in the question of whether human nature itself is transhistorically fixed, or instead historically variable. This is anthropology's first question, and if the past is any indication of the future, it is likely to remain so—at least until either Man, or history, comes to an end.

See also: Anthropology; Anthropology, History of; Civilization, Concept and History of; Civilizations; Cultural Critique: Anthropological; Fieldwork in Social and Cultural Anthropology; Historical Archaeology; Historical Explanation, Theories of: Philosophical Aspects; Historicism; Historiography and Historical Thought: Current Trends; History and the Social Sciences; History: Forms of Presentation, Discourses, and Functions; Knowledge, Anthropology of; Modernity; Modernity: Anthropological Aspects; Postmodernism: Philosophical Aspects; Primitive Society; Psychological Anthropology; Qualitative Methods, History of; Tradition, Anthropology of

Bibliography

Bourdieu P 1991 *The Political Ontology of Martin Heidegger* (trans. Collier P). Stanford University Press, Stanford, CA

Boyarin J (ed.) 1994 *Remapping Memory: The Politics of Timespace*. University of Minnesota Press, Minneapolis, MN

Brown D 1988 *History, Hierarchy, and Human Nature*. University of Arizona Press, Tucson, AZ
Cohn B S 1968 Ethnohistory. In: Sills D (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, pp. 440–8
Cohn B 1980 History and anthropology: The state of play. *Comparative Studies in Society and History* **22**(2): 198–221
Durkheim E 1965/1915 *The Elementary Forms of the Religious Life* (trans. Fields K E). Free Press, New York
Fabian J 1971 Language, history, and anthropology. *Journal of the Philosophy of the Social Sciences* **1**(1): 19–47
Fabian J 1983 *Time and the Other: How Anthropology Makes its Object*. Columbia University Press, New York
Frazer J G 1990 *The Golden Bough: A Study in Magic and Religion*. Macmillan, London
Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York
Goody J 1977 *The Domestication of the Savage Mind*. Cambridge University Press, Cambridge, UK
Halbwachs M 1980/1950 *The Collective Memory* (trans. Ditter F J Jr, Ditter V Y). Harper and Row Books, New York
Harris M 1968 *The Rise of Anthropological Theory: A History of Theories of Culture*. Crowell, New York
Herzfeld M 1987 *Anthropology through the Looking-Glass: Critical Ethnography in the Margins of Europe*. Cambridge University Press, New York
Kroeber A 1935 History and science in anthropology. *American Anthropologist* **37**(4): 539–69
Lévi-Strauss C 1966/1962 *The Savage Mind* (translation of *La pensée sauvage*). University of Chicago Press, Chicago
Morgan L H 1877 *Ancient Society*. Holt, New York
Rabinow P 1989 *French Modern: Norms and Forms of the Social Environment*. MIT Press, Cambridge, MA
Sahlins M 1985 *Islands of History*. University of Chicago Press, Chicago
Scholte B 1972 Toward a critical and reflexive anthropology. In: Hymes D (ed.) *Reinventing Anthropology*. Vintage Books, New York, pp. 430–57
Shryock A 1997 *Nationalism and the Genealogical Imagination: Oral History and Textual Authority in Tribal Jordan*. University of California Press, Berkeley, CA
Stocking G W Jr 1987 *Victorian Anthropology*. Free Press, New York
Taussig M 1993 *Mimesis and Alterity: A Particular History of the Senses*. Routledge, New York
Tylor E B 1871 *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art, and Custom*, 2 Vols. J. Murray, London

J. D. Faubion

Anthropology, History of

Social or cultural anthropology can be defined, loosely and broadly, as the comparative science of culture and society, and it is the only major discipline in the social sciences that has concentrated most of its attention on non-Western people. Although many of the classic problems investigated by anthropologists are familiar to the European history of ideas, the subject as it is known today emerged only in the early twentieth century, became institutionalized at universities in the Western world in mid-century, and underwent a phenomenal growth and diversification in the latter half of the century.

1. Foundations and Early Schools

1.1 Proto-anthropology

Interest in cultural variation and human universals can be found as far back in history as the Greek city-state. The historian Herodotos (fifth century BC) wrote accounts of 'barbarian' peoples to the east and north of the peninsula, comparing their customs and beliefs to those of Athens, and the group of philosophers known as the Sophists were perhaps the first philosophical relativists, arguing (as many twentieth century anthropologists later did) that there can be no absolute truth because, as one would put it today, truth is contextual. Yet their interest in cultural variation fell short of being scientific, chiefly because Herodotos lacked theory while the Sophists lacked empirical material.

Centuries later, scholarly interest in cultural variation and human nature re-emerged in Europe because of the new intellectual freedom of the Renaissance and questions arising from European overseas exploits. Michel de Montaigne (sixteenth century), Thomas Hobbes (seventeenth century), and Giambattista Vico (eighteenth century) were among the thinkers of the early modern era who tried to account for cultural variability and global cultural history as well as dealing with the challenge from relativism. Eighteenth century philosophers such as Locke, Hume, Kant, Montesquieu, and Rousseau developed theories of human nature, moral philosophies, and social theories, taking into account an awareness of cultural differences. The early German romantic Herder challenged Voltaire's universalistic vision by arguing that each people (*Volk*) had a right to retaining its own, unique values and customs—in a manner reminiscent of later cultural relativism romanticism. Indeed, by the end of the eighteenth century, several of the general questions still raised by anthropologists had already been raised: Universalism versus relativism (what is common to humanity; what is culturally specific), ethnocentrism versus cultural relativism (moral judgments versus neutral descriptions of other peoples), and humanity versus the rest of the animal kingdom (culture versus nature). Twentieth century anthropology has taught that these and other essentially philosophical problems are best investigated through the detailed study of living people in existing societies through ethnographic fieldwork, and by applying carefully devised methods of comparison to the bewildering variety of 'customs and beliefs.' It would

take several generations after Montesquieu's comparative musings about Persia and France, in his *Lettres persanes*, before anthropology achieved this mark of scientific endeavor.

1.2 Victorian Anthropology

The first general theories of cultural variation to enjoy a lasting influence were arguably those of two men trained as lawyers: Henry Maine (1822–88) in Britain and Lewis Henry Morgan (1818–82) in the USA. Both presented evolutionist models of variation and change, where Western European societies were seen as the pinnacle of human development. In his *Ancient Law* (1861), Maine distinguished between *status* and *contract* societies, a divide which corresponds roughly to later dichotomies between traditional and modern societies, or, in the late nineteenth-century German sociologist Ferdinand Tönnies' terminology, *Gemeinschaft* (community) and *Gesellschaft* (society); status societies are assumed to operate on the basis of kinship and myth, while individual merit and achievement are decisive in contract societies. Although simple contrasts of this kind have regularly been severely criticized, they continue to exert a certain influence on anthropological thinking (see *Evolutionism, Including Social Darwinism*).

Morgan's contributions to anthropology were wide-ranging and, among other things, he wrote a detailed ethnography of the Iroquois. His evolutionary scheme, presented in *Ancient Society* (1877), which influenced Marx and Engels, distinguished between seven stages (from lower savagery to civilization). His materialist account of cultural change influenced Marx and Engels. His pioneering work on kinship divided kinship systems into a limited number of types, and saw kinship terminology as a key to understanding society. Writing in the same period, the historian of religion Robertson Smith and the lawyer J.J. Bachofen offered, respectively, theories of monotheistic religion and of the (wrongly assumed) historical transition from matriliny to patriliny.

An untypical scholar in the otherwise evolutionist Victorian era, the German ethnologist Adolf Bastian (1826–1905) reacted against simplistic typological schemata. Drawing inspiration from Herderian romanticism and the humanistic tradition in German academia, Bastian wrote prolifically on cultural history, avoiding unwarranted generalizations, yet he held that all humans have the same pattern of thinking, thus anticipating structuralism.

The leading British anthropologist of the late Victorian era was Edward Tylor (1832–1917), whose writings include a famous definition of culture (dating from 1871), seeing it as the sum total of collective human achievements (thus contrasting it to nature). Tylor's student James Frazer (1854–1941) published the massive and very influential *Golden Bough* (1890, rev. ed. 1911–15), an ambitious comparative study of myth and religion.

Intellectual developments outside anthropology in the second half of the nineteenth century also had a powerful impact. Charles Darwin's theory of natural selection, first presented in his *Origin of Species* from 1859, would both be seen as a condition for anthropology (positing that all humans are closely related) and, later, as a threat to it (arguing the primacy of the biological over the cultural). The emergence of classic sociological theory in the works of Comte, Marx, and Tönnies, and later Durkheim, Weber, Pareto, and Simmel, provided anthropologists with general theories of society, although their applicability to non-European societies continues to be disputed (see *Sociology, History of*).

The quality of the data used by the early anthropologists was variable. Most of them relied on written sources, ranging from missionaries' accounts to travelogues of varying accuracy. The need for more reliable data began to make itself felt. Expeditions and systematic surveys now provided researchers around the turn of the twentieth century with improved knowledge of cultural variation, which eventually led to the downfall of the ambitious theories of unilineal evolution characteristic of nineteenth-century anthropology.

An Austro-German specialty proposed both as an alternative and a complement to evolutionist thinking, was diffusionism, the doctrine of the historical diffusion of cultural traits. Never a part of the mainstream outside of the German-speaking world, elaborate theories of cultural diffusion continued to thrive, particularly in Berlin and Vienna, until after World War II. As there were serious problems of verification associated with the theory, it was condemned as speculative by anthropologists committed to fieldwork and, furthermore, research priorities were to shift from general cultural history to intensive studies of particular societies.

In spite of theoretical developments and methodological refinements, the emergence of anthropology, as the discipline is known today, is rightly associated with four scholars working in three countries in the early decades of the twentiethth century: Franz Boas in the USA, A.R. Radcliffe-Brown and Bronislaw Malinowski in the UK, and Marcel Mauss in France.

1.3 Boas and Cultural Relativism

Boas (1858–1942), a German migrant to the USA who had briefly studied anthropology with Bastian, carried out research among Eskimos and Kwakiutl Indians in the 1890s. In his teaching and professional leadership, he strengthened the 'four-field approach' in American anthropology, which still sets it apart from European anthropology, including both cultural and social anthropology, physical anthropology, archaeology,

and linguistics. Although cultural relativism had been introduced more than a century before, it was Boas who made it a central premise for anthropological research. Against the evolutionists, he argued that each culture had to be understood in its own terms and that it would be scientifically misleading to rank other cultures according to a Western, ethnocentric typology gauging 'levels of development.' Boas also promoted *historical particularism*, the view that all societies or cultures had a unique history that could not be reduced to a category in some universalistic scheme. On related grounds, Boas argued incessantly against the claims of *racist* pseudoscience (see *Race: History of the Concept*).

Perhaps because of his particularism, Boas never systematized his ideas in a theoretical treatise. Several of his students and associates nevertheless did develop general theories of culture, notably Ruth Benedict, Alfred Kroeber, and Robert Lowie. His most famous student was Margaret Mead (1901–78). Although her best-selling books on Pacific societies have been criticized for being superficial, she used material from non-Western societies to raise questions about gender relations, socialization, and politics in the West, and Mead's work indicates the potential of cultural criticism inherent in the discipline.

One of Boas' most remarkable associates, the linguist Edward Sapir (1884–1939), formulated, with his student Benjamin Lee Whorf, the Sapir–Whorf hypothesis, which posits that language determines cognition. Consistent with a radical cultural relativism, the hypothesis implies that, for example, the Hopi perceive the world in a fundamentally different way from Westerners, due to differences in the structure of their respective languages.

1.4 The Two British Schools

While modern American anthropology had been shaped by the Boasians and their relativist concerns, as well as the perceived need to record native cultures before their anticipated disappearance, the situation in the major colonial power, Great Britain, was different. The degree of complicity between colonial agencies and anthropologists is debatable, but the very fact of imperialism was an inescapable premise for British anthropology until after World War II.

The man who is often hailed as the founder of modern British social anthropology was a Polish immigrant, Bronislaw Malinowski (1884–1942), whose over two years of fieldwork in the Trobriand Islands (between 1914 and 1918) set a standard for ethnographic data collection that is still largely unchallenged. Malinowski argued the need to learn the local language properly and to engage in everyday life, in order to see the world from the actor's point of view and to understand the interconnections between social institutions and cultural notions. Malinowski placed an unusual emphasis on the acting individual, seeing social structure not as a determinant of but as a framework for action, and he wrote about a wide range of topics, from garden magic, economics, and sex to the puzzling *kula* trade. Although he dealt with issues of general concern, he nearly always took his point of departure in his Trobriand ethnography, demonstrating a method of generalization very different from that of the previous generation with its scant local knowledge.

The other major British anthropologist of the time was A.R. Radcliffe-Brown (1881–1955). An admirer of Durkheim's sociology, Radcliffe-Brown did relatively little fieldwork himself, but aimed at the development of a 'natural science of society' where universal laws of social life could be formulated. His theory, known as structural-functionalism, saw the individual as unimportant, emphasizing instead the social institutions (kinship, norms, politics, etc.). Most social and cultural phenomena were seen as functional in the sense that they contributed to the maintenance of the overall social structure (see *Functionalism in Anthropology*).

Despite their differences in emphasis, both British schools had a sociological concern in common (which they did not share with most Americans), and tended to see social institutions as functionally integrative. Both rejected the wide-ranging claims of diffusionism and evolutionism, and yet, the tension between structural explanations and actor-centered accounts remains strong in British anthropology even today.

Malinowski's students included Raymond Firth, Audrey Richards, and Isaac Schapera, while Radcliffe-Brown, in addition to enlisting E.E. Evans-Pritchard and Meyer Fortes—arguably the most powerful British anthropologists in the 1950s—on his side, taught widely, and introduced structural-functionalism to several foreign universities. British interwar anthropology was characteristically oriented towards kinship, politics, and economics, with Evans-Pritchard's masterpiece *The Nuer* (1940) demonstrating the intellectual power of a discipline combining detailed ethnography, comparison, and elegant models. Later, his models would be criticized for being too elegant to fit the facts on the ground—a very Malinowskian objection.

1.5 Mauss

No fieldwork-based anthropology developed in the German-speaking region, and German anthropology was marginalized after World War II. In France, the situation was different. Already in 1902, Durkheim had published, with his nephew Marcel Mauss (1872–1950), an important treatise on primitive classification; in 1908, Arnold van Gennep published *Les Rites de Passage*, an important analysis of initiation rites, and Lucien Lévy-Bruhl elucidated a theory, later refuted by both Evans-Pritchard, Mauss and others, on the

'primitive mind,' which he held to be 'pre-logical.' New empirical material of high quality was being produced by thorough observers such as Maurice Leenhardt in New Caledonia and Marcel Griaule in West Africa.

Less methodologically purist than the emerging British traditions and more philosophically adventurous than the Americans, interwar French anthropology, under the leadership of Mauss, developed a distinct flavor, witnessed in the influential journal *L'Année Sociologique*, founded by Durkheim and edited by Mauss after Durkheim's death in 1917. Drawing on his vast knowledge of languages, cultural history, and ethnography, Mauss, who never did fieldwork himself, wrote several learned, original, compact essays ranging from gift exchange (*Essai sur le Don*, 1924) to the nation, the body, and the concept of the person. Mauss' theoretical position was complex. He believed in systematic comparison and the existence of recurrent patterns in social life at all times and in all places, yet he often defended relativist views in his reasoning about similarities and differences between societies.

Not a prolific writer, Mauss exerted an enormous influence on later French anthropology through his teaching. Among his students and associates were most of the major French anthropologists at the time, and the three leading postwar scholars in the field—Louis Dumont, Claude Lévi-Strauss, and Georges Balandier—were all deeply indebted to Mauss.

1.6 Some General Points

The transition from evolutionist theory and grand syntheses to more specific, detailed, and empirically founded work, in reality amounted to an intellectual revolution. The work of Tylor and Morgan had been relegated to the mists of history, and the discipline had been taken over by small groups of scholars who saw intensive fieldwork, cultural relativism, the study of single societies, and rigorous comparison as its essence. Today, the academic institutions, the conferences, and the learned journals all build on the anthropology of Boas, Malinowski, Radcliffe-Brown, and Mauss. This is to a great extent also true of the anthropological traditions of other countries (Vermeulen and Roldán 1995), including India, Australia, Mexico, Argentina, The Netherlands, Spain, and Scandinavia. Soviet/Russian and East European anthropologies have followed different itineraries, retaining a connection with the German *Volkskunde* tradition.

2. Anthropology in the Second Half of the Twentieth Century

The numbers of anthropologists and institutions devoted to teaching and research in the field grew rapidly after World War II. The discipline diversified. New specializations such as psychological anthropology, political anthropology, and the anthropology of ritual emerged, and the geographical foci of the discipline multiplied: whereas the Pacific had been the most fertile area for theoretical developments in the 1920s and Africa had played a similar part in the 1930s and 1940s, while the preoccupation with North American Indians had been stable throughout, the 1950s saw a growing interest in the 'hybrid' societies of Latin America and the Caribbean as well as the anthropology of India and South-East Asia, and the New Guinean highlands became similarly important in the 1960s. Such shifts in spatial emphasis had consequences for theoretical developments, as each region posed its own peculiar problems.

From the 1950s, the end of colonialism also affected anthropology, both in a banal sense—it became more difficult to obtain research permits—and more profoundly, as the subject–object relationship between the observer and the observed became problematic as the traditionally 'observed' peoples increasingly had their own intellectuals and spokespersons who frequently objected to Western interpretations of their way of life.

2.1 Structuralism

The first major theory to emerge after World War II was Claude Lévi-Strauss' structuralism. Lévi-Strauss (1908–) developed an original theory of the human mind, drawing on structural linguistics, Mauss' theory of exchange and Lévy-Bruhl's theory of the primitive mind (which Lévi-Strauss opposed). His first major work, *Les Structures Elémentaires de la Parenté* (*The Elementary Structures of Kinship*, 1949), introduced a grammatical, formal way of thinking about kinship, with particular reference to systems of marriage (the exchange of women between groups). Lévi-Strauss later expanded his theory to cover totemism, myth, and art. Never uncontroversial, structuralism had an enormous impact on French intellectual life far beyond the confines of anthropology. In the English-speaking world, the reception of structuralism was delayed, as Lévi-Strauss' major works were not translated until the 1960s, but he had his admirers and detractors from the beginning. Structuralism was criticized for being untestable, positing as it did certain unprovable and unfalsifiable properties of the human mind (most famously the propensity to think in terms of contrasts or binary oppositions), but many saw Lévi-Strauss' work, ultimately committed to human universals, as an immense source of inspiration in the study of symbolic systems such as knowledge and myth (see *Structuralism, History of*).

A different, and for a long time less influential, brand of structuralism was developed by Louis Dumont (1911–99), an Indianist and Sanskrit scholar who did fieldwork both in the Aryan north and the Dravidian south. Dumont, closer to Durkheim than

Lévi-Strauss, argued in his major work on the Indian caste system, *Homo Hierarchicus* (1969), for a holistic perspective (as opposed to an individualistic one), claiming that Indians (and by extension, many non-modern peoples) saw themselves not as 'free individuals' but as actors irretrievably enmeshed in a web of commitments and social relations, which in the Indian case was clearly hierarchical.

Most later, major French anthropologists have been associated with Lévi-Strauss, Dumont, or Balandier, the Africanist whose work in political anthropology simultaneously bridged gaps between France and the Anglo-Saxon world and inspired both neo-Marxist research and applied anthropology devoted to development.

2.2 Reactions to Structural-functionalism

In Britain and her colonies, the structural-functionalism now associated chiefly with Evans-Pritchard and Fortes was under pressure after the war. Indeed, Evans-Pritchard himself repudiated his former views in 1949, arguing that the search for 'natural laws of society' had been shown to be futile and that anthropology should fashion itself as a humanities discipline rather than a natural science. Retrospectively, this statement has often been quoted as marking a shift 'from function to meaning' in the discipline's priorities; Kroeber expressed similar views in the USA. Others found other paths away from what was increasingly seen as a conceptual straitjacket. Edmund R. Leach, whose *Political Systems of Highland Burma* (1954) suggested a departure from functionalist orthodoxies, notably Radcliffe-Brown's dictum that social systems tend to be in equilibrium and Malinowski's view of myths as integrating 'social charters', would later be a promoter and critic of structuralism in Britain. Leach's contemporary Raymond Firth proposed a distinction between social structure (the statuses in society) and social organization, which he saw as the actual process of social life, whereby choice and individual whims were related to structural constraints. Later in the 1950s and 1960s, several younger social anthropologists, notably F.G. Bailey and Fredrik Barth, followed Firth's lead as well as the theory of games (a recent development in economics) in refining an actor-centered perspective on social life. Max Gluckman, a former student of Radcliffe-Brown and a close associate of Evans-Pritchard, also abandoned the strong holist program of the structural-functionalists, reconceptualizing social structure as a loose set of constraints, while emphasizing the importance of individual actors. Gluckman's colleagues included several important Africanists, such as A.L. Epstein, J. Clyde Mitchell, and Elizabeth Colson. Working in southern Africa, this group pioneered urban anthropology and ethnicity studies in the 1950s and 1960s.

2.3 Neo-evolutionism, Cultural Ecology and Neo-Marxism

The number of anthropologists has always been larger in the USA than anywhere else, and the discipline was always diverse there. Although the influence from the Boasian cultural relativist school remains strong, other groups of scholars have also made their mark. From the late 1940s onwards, a resurgent interest in Morgan's evolutionism led to the formulation of neoevolutionist and materialist research programs. Julian Steward, a student of Robert Redfield (who had been a student of Radcliffe-Brown), proposed a theory of cultural dynamics distinguishing between 'the cultural core' (basic institutions such as the division of labor) and 'the rest of culture' in a way strongly reminiscent of Marx. Steward led research projects among Latin American peasants and North American Indians, encouraging a focus on the relationship between culture, technology, and the environment. Leslie White held more deterministic materialist views, but also—perhaps oddly—saw symbolic culture as a largely autonomous realm. Among the major scholars influenced by White, Marvin Harris has strengthened his materialist determinism, while Marshall D. Sahlins in the 1960s made the move from neoevolutionism to a symbolic anthropology influenced by structuralism.

Cultural ecology sprang from the teachings of Steward and White, and represented a rare collaboration between anthropology and biology. Especially in the 1960s, many such studies were carried out, including, notably, Roy Rappaport's *Pigs for the Ancestors* (1968), an attempt to account for a recurrent ritual in the New Guinean highlands in ecological terms. The upsurge of Marxist peasant research, especially in Latin America, in the 1970s, was also clearly indebted to Steward.

The appearance of radical student politics in the late 1960s, which had an impact on academia until the early 1980s, had a strong, if passing, influence on anthropology. Of the more lasting contributions, the peasant studies initiated by Steward and furthered by Eric Wolf, Sidney Mintz, and others must be mentioned, along with French attempts, represented in the very sophisticated work of Maurice Godelier and Claude Meillassoux, at synthesizing Lévi-Straussian structuralism, Althusserian Marxism, and anthropological comparison. Although Marxism and structuralism eventually became unfashionable, scholars—particularly those engaged in applied work—continue to draw inspiration from Marxist thought (see *Marxist Social Thought, History of*).

2.4 Symbolic and Cognitive Anthropology

More true to the Boasian legacy than the materialist approaches, studies of cognition and symbolic systems developed and diversified in the USA. A leading theorist is Clifford Geertz, who wrote a string of

influential essays advocating hermeneutics (interpretive method) in the 1960s and 1970s (see *Hermeneutics, History of*). While his originality as a theorist can be questioned, his originality as a writer is obvious, and Geertz ranks as perhaps the finest writer of contemporary anthropology. Marshall Sahlins is, with Geertz, the foremost proponent of cultural relativism today, and has published a number of important books on various subjects (from Mauss' theory of exchange to sociobiology and the death of Captain Cook), consistently stressing the autonomy of the symbolic realm, thus arguing that cultural variation cannot be explained by recourse to ecology, technology, or biology.

In Britain, too, interest in meaning, symbols, and cognition grew after the war, especially from the 1960s (partly due to the belated reception of Lévi-Strauss). British anthropology had hitherto been strongly sociological, and two scholars who fused the legacy from structural-functionalism with the study of symbols and meaning were Mary Douglas and Victor Turner. Taking his cue from van Gennep, Turner, a former associate of Gluckman, developed a complex analysis of rituals among the Ndembu of Zambia, showing their functionally integrating aspects, their meaningful aspects for the participants, and their deeper symbolic significance. Douglas, a student of Evans-Pritchard, famous for her *Purity and Danger* (1966), analyzed the human preoccupation with dirt and impurities as a way of thinking about the boundaries of society and the nature/culture divide. Prolific and original, Douglas is a major defender of a reformed structural-functionalism.

Against all these (and other) perspectives regarding how 'cultures' or 'societies' perceive the world, anthropologists emphasizing the actor's point of view have argued that no two individuals see the world in the same way and that it is preposterous to generalize about societies. The impact of feminism has been decisive here. Since the 1970s, feminist anthropologists have identified often profound differences between male and female world-views, indicating how classic accounts of 'societies' really refer to male perspectives on them as both the anthropologist and the main informants tended to be male. For example, in a restudy of Trobriand society undertaken in the 1970s and 1980s, Annette Weiner showed that Malinowski's famous work was ultimately misleading as he had failed to observe important social processes confined to females.

2.5 Anthropology and the Contemporary World

Since the pillars of modern anthropology were erected around the First World War, the former colonies became independent, 'natives' got their own educated elites (including social scientists), economic and cultural globalization led to the spread of capitalism and consumer culture, and transcontinental migration blurred the boundaries between the traditional 'us' and 'them.' This situation entailed new challenges for anthropology, which were met in various ways—revealing continuities as well as breaks with the past.

A late field to be incorporated into anthropology, but one that became the largest single area of interest from the 1970s, was the study of identity politics, notably ethnicity and nationalism. Since the publication of several important texts around 1970 (by, *inter alia*, Barth and Abner Cohen in Europe, and George DeVos in the USA), anthropological ethnicity studies investigated the interrelationship between ethnic identity and ethnic politics, and explored how notions of cultural differences contribute to group identification. Since the publication of several important texts on nationalism in the early 1980s (by Ernest Gellner, Benedict Anderson and others), this also became an important area for anthropologists. Ethnicity and nationalism are partly or wholly modern phenomena associated with the state, and thus denote a departure from the former mainstay of anthropology, the study of nonmodern small-scale societies. While ethnicity and (especially) nationalism could not be studied through participant-observation only (other kinds of data are required), it was evident that anthropologists who engaged in this field remained committed to the classic tenets of the discipline; ethnographic fieldwork, comparison, and a systemic view of social reality. Also, the study of identity politics emerged as an interdisciplinary field where anthropologists, sociologists, historians, and political scientists profited from each other's expertise.

Other modern phenomena also received increased attention in anthropology from the 1970s, including consumption, 'subcultures,' wagework, and migration. The boundary between the 'Western self' and the 'non-Western other' became blurred. Anthropological studies of Western societies became common, and Europe was established as an ethnographic region along with West Africa, South Asia, and so on. Even anthropologists working in traditional settings with classic topics increasingly had to see their field as enmeshed, to a greater or lesser extent, in a global system of communication and exchange.

Because of the increased penetration of the formerly tribal world by capitalism and the state, and accompanying processes of cultural change, there was a growing demand for a reconceptualization of culture in the 1980s and 1990s, and scholars such as Ulf Hannerz and Marilyn Strathern developed fluid concepts of culture seeing it as less coherent, less bounded, and less integrated than the Boasian and Malinowskian traditions implied.

Some scholars saw the postcolonial situation as sounding the death knell of anthropology: Since the 'primitive' was gone, and the former informants were now able to identify and describe themselves (they no longer needed anthropologists to do it), the science of

cultural variation seemed to have lost its *raison d'être*. Following the lead of Edward Said's *Orientalism* (1978), an influential critique of Western depictions of the 'East', and often inspired by Michel Foucault, they saw anthropology as a colonial and imperialist enterprise refusing non-Western peoples a voice of their own and magnifying the distance between 'us' and 'them.' Especially in the late 1980s and early 1990s, this view had many followers, some of whom abandoned empirical research, while others tried to incorporate the autocriticism into their work. Yet others saw these pessimistic views as largely irrelevant, since anthropology had always been fraught with similar tensions, to which each new generation found its solutions. In this regard, it must be pointed out that the earlier feminist critique of anthropology, far from repudiating the subject, led to its enriching by adding new implements to its toolbox and new dimensions to its worldview. The same could be said of the reconceptualizations of culture, which arguably offered an improved accuracy of description (see *Postcoloniality*).

3. The Situation at the Turn of the Millennium

Over the course of the twentieth century, anthropology became a varied discipline with a strong academic foothold in all continents, although its centers remained in the English- and French-speaking countries. It was still possible to discern differences between American cultural anthropology, British social anthropology, and French *ethnologie*, but the discipline was more unified than ever before—not in its views, but in its approaches. Hardly a part of the world had not now been studied intensively by scholars engaging in ethnographic fieldwork, but since the world changes, new research is always called for.

Specializations proliferated, ranging from studies of ethnomedicine and the body to urban consumer culture, advertising, and cyberspace. Although the grand theories of the nineteenth and twentieth centuries—from unilinear evolutionism to structuralism—had been abandoned, new theories claiming to provide a unified view of humanity were being proposed; for example, new advances in evolutionary psychology and cognitive science offered ambitious general accounts of social life and the human mind, respectively. The problems confronting earlier generations of anthropologists, regarding, for example, the nature of social organization, of knowledge, of kinship, and of myth and ritual, remained central to the discipline although they were explored in new empirical settings by scholars who were more specialized than their predecessors.

Anthropology has thrived on the tension between the particular and the universal; between the intensive study of local life and the quest for general accounts of the human condition. Is it chiefly a generalizing science or a discipline devoted to the elucidation of the unique? The general answer is that anthropologists ultimately do study Society, Culture, and Humanity, but that in order to do so, they must devote most of their energies to the study of societies, cultures, and humans. As long as their mutual differences and similarities are not fully understood, there will be an intellectual space in the world for anthropology or, at least, a discipline like it (see *Anthropology*; *Spatial Thinking in the Social Sciences, History of*).

See also: Anthropology; Boas, Franz (1858–1942); Cognitive Anthropology; Colonialism, Anthropology of; Community/Society: History of the Concept; Darwin, Charles Robert (1809–82); Human Behavioral Ecology; Psychiatry, History of; Structuralism; Symbolism in Anthropology

Bibliography

Clifford J 1988 *The Predicament of Culture: Twentieth-Century Ethnography, Literature and Art*. Harvard University Press, Cambridge, MA

Kuper A 1996 *Anthropology and Anthropologists: The Modern British School*, 3rd edn. Routledge, London

Lévi-Strauss C 1987 *Introduction to the Work of Marcel Mauss* [transl. Baker F 1987]. Routledge, London

Moore J D 1997 *Visions of Culture: An Introduction to Anthropological Theories and Theorists*. AltaMira Press, Walnut Creek, CA

Stocking G W, Jr 1987 *Victorian Anthropology*. Free Press, New York

Stocking G W, Jr (ed.) 1996 *Volksgeist as Method and Ethic: Essays on Boasian Ethnography and the German Anthropological Tradition*. University of Wisconsin Press, Madison, WI

Vermeulen H F, Roldán A A (eds.) 1995 *Fieldwork and Footnotes: Studies in the History of European Anthropology*. Routledge, London

T. H. Eriksen

Antidepressant Drugs

Because of historical developments, a stringent classification for all available agents with antidepressant properties has not yet been agreed upon. Initial classificatory attempts focused on the chemical structure of the drugs. Recent classifications emphasize the pharmacological profile of the drug. This latter approach is of greater clinical relevance, because it depicts the primary target sites in the central nervous system (e.g. transporter proteins or various receptor subtypes), allowing a better prediction of the potential clinical application and the side-effect profile of the drug. As a result of both approaches, many clinicians today distinguish the following groups of antidepressants: (a) monoaminooxidase inhibitors (MAOI), (b)

tricyclic antidepressants (TCA), (c) selective serotonin reuptake inhibitors (SSRI), and (d) antidepressants with different pharmacological profiles.

1. History

1.1 MAOI

Iproniazid and its precursor isoniazid had been developed in the laboratories of the company Hoffmann-La Roche primarily as antituberculosis agents. As a consequence of casuistic observations on mood-elevating properties of iproniazid, this effect was first confirmed by Loomer and colleagues (1957) in a clinical trial. However, significant hepatotoxicity was observed during iproniazid treatment. Subsequent development of drugs inhibiting MAO systems led to the introduction of isocarboxazid, phenelzine and tranylcypromine as irreversible inhibitors of MAO. Their use has been limited because of possible hypertensive crises after ingestion of food containing large amounts of tyramine, forcing patients to follow dietary restrictions during pharmacotherapy with these drugs. This led to the development of reversible MAOI such as moclobemide, which no longer required dietary restrictions.

1.2 TCA

After the discovery of the antipsychotic effects of chlorpromazine in 1952, the compound G22355, later called imipramine, was developed in the laboratories of the company Geigy. Imipramine was a derivative of chlorpromazine and was first tested in clinical trials in schizophrenia. While no relevant antipsychotic effects were noted, mood-elevating effects in patients with a co-occurring depressive syndrome were observed. At this point, both the company and the clinical investigator Kuhn decided to test imipramine in patients with depression as a potential antidepressant. The results of the clinical trials were first published in 1957, and are still a remarkable document of skillful clinical observation, as the report contains many details about the antidepressant effects of imipramine which still hold true today (Kuhn 1957).

A variety of compounds that were chemically related to imipramine have been developed. They all shared the basic chemical structure of a tricyclus. Subsequent research revealed that in most instances, while the tricyclic basal structure was preserved, minor alterations of its side chains resulted in markedly different pharmacological profiles of the drugs. The chemically related compound maprotiline, often termed a tetracyclic, was developed as a consequence of systematic variation of the tricyclic structure.

From the 1960s to the 1980s, the TCA were the dominating group of antidepressants in clinical practice throughout the world. As many of the TCA compounds induce substantial side effects, they gradually lost their predominance with the availability of the SSRI during the 1990s.

1.3 SSRI

In the 1970s, on the basis of increasing knowledge about the mechanism of action of MAOI and TCA, pathophysiological models of depression were proposed. Consequently, potential antidepressants with more specific neurotransmitter receptor effects were developed. According to the norepinephrine depletion hypothesis of depression, drugs to selectively potentiate central noradrenergic neurotransmission were developed. On the other hand, the various available TCA were scrutinized for their potency of serotonin reuptake inhibition. Clomipramine was shown to have the strongest effects on serotonin reuptake inhibition; however, the substance does not selectively inhibit the serotonin transporter but also shows effects on norepinephrine.

Subsequently, in a collaboration between a pharmaceutical company (Astra) and the research team of Carlson in Sweden, the first SSRI zimelidine was developed. The first results from a clinical trial on zimelidine were published in 1977 (Benkert et al. 1977). However, the antidepressant had to be dropped from the market because of the emergence of neurological side effects.

In the 1980s, fluoxetine ('Prozac'), an SSRI developed by Eli-Lilly Company, was made available in the USA. The introduction of this drug represents the breakthrough of SSRIs in the treatment of depression. Successful marketing strategies and the low toxicity of fluoxetine may explain the broad spectrum of indications for this drug. The prescription of fluoxetine in patients with personality disorders and, finally, the use of Prozac by healthy people to improve their general wellbeing, led to an ongoing discussion regarding the acceptability of so-called 'life-style' drugs.

In the following years, four additional SSRIs were approved: fluvoxamine, paroxetine, sertraline, and citalopram. The latter two substances have favorable pharmacokinetics and a low risk for drug–drug interactions—an issue that became increasingly important in the 1990s.

1.4 Antidepressants With Different Pharmacological Profiles

The introduction of selective antidepressants reduced the rate of unwanted side effects. The higher selectivity of these drugs was, however, not accompanied by higher efficacy. Therefore, research and drug development focused on compounds which exerted their effects on different and interrelated neurotransmitter

receptors in the brain to enhance their benefit: risk ratio based on theoretical reasoning about the anticipated profile of effects and side effects.

The push to develop highly selective antidepressants targeting a single type of neurotransmitter receptor or transporter had ended. Highly effective and well-tolerated drugs like mirtazapine and venlafaxine are examples of the new strategy of drug design.

Rather independent from this line of progress, the herbal St. John's Wort has achieved great popularity throughout the world, based on the drug's low toxicity, its high acceptance, and a proven efficacy in the treatment of mild to moderate depression. The exact mechanism of action is not completely known.

2. *Mechanism of Action of Antidepressants*

There is considerable evidence that many antidepressants facilitate neurotransmission of one or more monoamines—serotonin, norepinephrine, or dopamine. This effect appears to be mediated by different mechanisms:

(a) blockade of one or more types of synaptic monoamine transporters;

(b) blockade of the enzymes that inactivate monoamines;

(c) stimulation and/or blockade of receptor subtypes.

2.1 *Blockade of One or More Types of Synaptic Monoamine Transporters*

Synaptic monoamine transporters are protein structures which rapidly retrieve released neurotransmitter molecules from the synaptic cleft into the originating neuron. By specifically blocking the transporter, the reuptake of the neurotransmitter is hampered, thus increasing its availability in the synapse. Examples of compounds that predominantly or selectively inhibit the reuptake of serotonin are the SSRI listed above, or the tricyclic compound clomipramine. Drugs predominantly or selectively inhibiting the norepinephrine reuptake are reboxetine, and the tricyclic compounds desipramine and nortriptyline. Many antidepressants block both the serotonin and norepinephrine transporter, more or less with the same potency, e.g. imipramine, amitriptyline, doxepine, milnacipran, or venlafaxine. Examples of dopamine reuptake inhibitors are bupropion (also affecting norepinephrine uptake), or amineptine.

2.2 *Blockade of the Enzymes that Inactivate Monoamines*

The main enzyme relevant in this context is the MAO. Its isoform MAO-A degrades mainly serotonin and norepinephrine, while MAO-B predominantly metabolizes dopamine. MAO can be blocked irreversibly and unselectively by compounds such as phenelzine or tranylcypromine. Reversible MAO-A inhibitors such as moclobemide selectively block the metabolism of serotonin and norepinephrine within therapeutic ranges. Selective MAO-B inhibitors have not consistently proved to be antidepressants.

2.3 *Stimulation and/or Blockade of Receptor Subtypes*

In this category, various compounds with different modes of action are mentioned. From the tricyclic compounds, trimipramine blocks dopamine D2, alpha-1-adrenergic H1-histaminergic and muscarinic receptors without any substantial effect on monoamine transporters. Mirtazapine blocks alpha-1 and 2-adrenergic receptors, as well as 5-HT_2 and 5-HT_3 and H1 histaminergic receptors. Nefazodone mainly blocks 5-HT_2 receptors, with weak inhibition of serotonin and norepinephrine uptake.

The pharmacological profile of most antidepressants has been well characterized, providing an explanation for some of the acute and prolonged effects of these agents. In particular, the side-effect profile of antidepressants can be readily derived from their pharmacological profile. However, the postulated enhancement of monoaminergic function may not be the only or even the crucial factor explaining why antidepressants work. The pharmacological profile of a drug does not sufficiently explain its antidepressant properties. This is borne out by the following findings:

(a) The pharmacological target sites within the CNS are affected within minutes after administration of the drug. Many side effects of these agents occur early in the course of treatment, in good time correlation with the administration of the drug. However, virtually all antidepressants need several days to weeks of regular application until their full therapeutic effect becomes clinically visible.

(b) Other pharmacological agents acting as monoamine reuptake inhibitors comparable to the above-mentioned antidepressant drugs have not been found to possess convincing antidepressant properties (e.g., cocaine, amphetamine).

(c) Drugs without primary effect on monoamine sites may also possess antidepressant effects, such as substance P antagonists or CRF antagonists.

Several hypotheses have been developed to explain the delay of full antidepressant effects. It has been postulated that the effects of antidepressants at their primary target sites (receptors or transporters as first messengers) induce secondary effects, e.g., adaptatory changes in intracellular signal transduction at the level of second messengers (such as the adenylyl cyclase system) or of gene expression. This may lead to altered receptor sensitivity (as described for beta-adrenergic receptor down-regulation or decreased sensitivity of serotonergic receptors, such as 5-HT_2 receptors).

Changes of the synthesis of brain-derived neurotrophic factor and/or alterations of the hypothalamic–pituitary axis may represent pathways by which antidepressants finally exert their therapeutic effects. Changes in synthesis of protein structures might explain the time lag needed for antidepressants to develop their full therapeutic effect on mood and affect.

3. Indications

By now, many drugs with antidepressant effects have also been evaluated as effective agents in several indications, which are briefly outlined below.

3.1 Depressive Disorders

The classical indication for TCA, MAOI, SSRI and other antidepressants is depressive episodes (e.g., major depression). Their use is the treatment of choice in moderate or severe depression. Antidepressants are effective both in acute treatment and in preventing relapse of unipolar depression during long-term treatment (continuation or maintenance phase, after the acute symptoms have been treated effectively). Several authors have found that depressive episodes with melancholic features are more likely to respond to drug treatment than to psychotherapy. So-called atypical depressive episodes are reportedly more likely to respond to MAOI (and perhaps to SSRI) than to TCA such as imipramine. In the treatment of psychotic depression, the combination of an antidepressant with an antipsychotic is warranted. Contrary to former belief, long-lasting mild to moderate depressive states, currently labeled as dysthymia, also respond well to antidepressant drug treatment in a majority of cases. In bipolar disorder, the use of antidepressants during a depressive episode may trigger a switch into a manic state; therefore their use in this indication is currently a matter of debate.

3.2 Panic Disorder

Starting with TCA such as imipramine or clomipramine, a variety of antidepressant drugs have been demonstrated to be an effective treatment in panic disorder with or without agoraphobia. Currently, SSRI are the treatment of choice in this indication, if a drug treatment with an antidepressant is established. SSRI have shown convincing anxiolytic properties in combination with a favorable side-effect profile for a majority of patients.

3.3 Obsessive–compulsive Disorder (OCD)

This often chronic and disabling disorder, characterized by intruding thoughts or repetitive compulsions that often are time-consuming and may interfere considerably with an individual's functioning, seems to respond selectively to compounds such as clomipramine or SSRI that act as predominant or selective serotonin reuptake inhibitors, because norepinephrine reuptake inhibitors or MAOIs have not consistently shown therapeutic effects in this indication.

3.4 Generalized Anxiety Disorder

In the treatment of this chronic anxiety disorder characterized by waxing and waning anxiety symptoms and excessive worrying, several antidepressants such as SSRI or venlafaxine have demonstrated beneficial effects in reducing cognitive anxiety symptoms in particular.

3.5 Social Phobia

MAOI and SSRI are effective treatments for this disorder, characterized by marked anxiety in social situations, where the individual is exposed to unfamiliar people or to possible scrutiny by others.

3.6 Bulimia Nervosa

This disorder, with recurrent episodes of binge eating, loss of control, overeating impulses, and inappropriate compensatory behavior to prevent weight gain, can also be beneficially influenced by antidepressants. Binge frequency and purging behavior in particular can be significantly reduced by regular treatment.

3.7 Premenstrual Syndrome

Symptoms of tension, dysphoria, and irritability, accompanied by various somatic complaints and occurring during the late luteal phase of the menstrual cycle, can be effectively treated by continuous or intermittent administration of SSRI or clomipramine. As in the case of OCD, there is evidence that only predominant or selective serotonin reuptake inhibitors are effective from the spectrum of antidepressant drugs.

3.8 Further Indications

Antidepressants have also been demonstrated to be of therapeutic value in the treatment of pain syndromes of various etiologies, in the prophylaxis of migraine, in the treatment of post-traumatic stress disorder, the treatment of children with enuresis or attention-deficit and hyperactivity syndrome, or in patients with narcolepsy (especially clomipramine). SSRI may also be of therapeutic value in treating premature ejacu-

lation, obesity, or anger attacks associated with mood or personality disorders. In impulsive and affectively unstable personality disorders (e.g., borderline personality disorder) SSRI may be beneficial, whereas MAOI may be helpful in avoidant personality disorder.

4. *Therapeutic Principles*

If treatment with antidepressant drugs is established, the prescribing clinician and the treated patient need to consider some general guidelines, which apply to most indications for which these agents have been evaluated:

(a) Treatment with antidepressants generally requires continuous treatment over several weeks on a regular basis; intake on an as-needed base generally is not sensible.

(b) Therapeutic effects generally develop gradually during continuous treatment with antidepressants. In a majority of cases, 2–3 weeks are needed until more than 50 percent symptom reduction is achieved.

(c) During the course of treatment, side effects usually appear before the desired therapeutic effect reaches its maximum. This particular issue needs to be clarified with the patient in order to ensure compliance with treatment up to the point of clinical remission.

(d) Antidepressants need to be given in a sufficiently high dose to achieve optimal treatment effects. Efficient doses for TCA usually lie between 100 and 300 mg/day. SSRI such as citalopram, fluoxetine, or paroxetine need 20 mg/day, sertraline or fluvoxamine at least 50 mg/day. Depending on the side-effect profile of a drug, initial treatment requires gradual dose increase until effective doses are achieved (for TCA recommended starting doses are 25–50 mg/day; SSRI can usually be started with the minimal effective dose).

After remission is reached, it is generally recommended to maintain the effective dose for at least several weeks or months in order to prevent recurrence of the symptoms or relapse. In cases of known high risk of recurrence, continued treatment over years is needed. Combining antidepressant drug treatment with psychotherapy (especially cognitive-behavioral or interpersonal therapy) usually provides additive benefit for the patient; both approaches are not contradictory.

Antidepressants do not have addictive properties.

5. *Disputable Issues: Ethical, Methodological, Clinical*

Despite the widely accepted beneficial effects of antidepressants, there is ongoing debate on how these effects and benefits can be assessed. The scientific gold standard established over a period of about 50 years, and representing the basis of approval of new antidepressant drugs by legal authorities, is to assume *efficacy* of antidepressants if they are clearly superior to placebo in at least two well-conducted randomized controlled trials. Ethical issues are primarily related to the use of placebo in such pivotal trials. A recent study has shown that suicide rates and mortality in patients treated with placebo in randomized controlled trials were not discernible from those of patients treated with an antidepressant verum (Khan et al. 2000).

The demonstration of efficacy of antidepressants in controlled clinical trials is necessary but not sufficient for evaluating the therapy of major depression and related disorders. The focus of interest is shifting from pure efficacy to *effectiveness* studies of antidepressants. This implies asking the question whether a treatment still works when used by the average clinician with the average patient in an average therapeutic setting. Additionally, *efficiency* studies are essential to assess the required level of resources to produce reasonable benefit. Research at the interface of clinical trials and effectiveness studies, including cost-utility and econometric studies, is indispensable to estimate the benefit and the shortcomings of available antidepressant drug therapy (Wells 1999).

Evidence from double-blind, placebo-controlled trials of antidepressant drugs from all classes has shown them to be effective in alleviating symptoms in approximately 60 to 75 percent of patients with major depression. In contrast, the placebo response rate averages 35 percent after one month of treatment. However, the very high variability of verum and placebo responses makes it nearly impossible to extrapolate from one study to another or from the results of a clinical trial to real clinical situations. This variability of study outcomes can be reduced by patients' selection based on DSM-IV criteria, and by clinicians trained on standard depression assessment scales. A symptom reduction of at least 50 percent, e.g., in the Hamilton Depression Rating Scale score, is widely used as an arbitrarily chosen but operational measure of treatment response. However, a 50 percent improvement may leave many patients with significant symptoms, and, paradoxically, it is more difficult to achieve a 50 percent response if the baseline severity of depression is high. As a consequence, antidepressant drug efficacy in less severely depressed patients is not easily demonstrable, although most of these patients require effective treatment.

The effects of antidepressive drugs are seen in all of the symptoms of major depression; however, vegetative symptoms, e.g., sleep disturbances and loss of appetite, are often the most rapidly improved. When only a single global measure is used to assess depression severity or improvement, differential drug effects and differential response patterns in subgroups of patients are overlooked. Thus, the methodology used for the assessment of symptoms and symptom changes, as well as the distribution of prevailing symptoms in different patients' samples, can influence

study results. There is still a lack of studies considering these issues and other factors which can potentially influence treatment response, e.g., patients' gender, age, or personality.

Furthermore, the efficacy of antidepressants is often confounded with the time course of improvement. There is still controversy over whether or not some antidepressants exert a faster onset of action. Survival–analytical approaches (Stassen et al. 1993), exponential recovery models (Priest et al. 1996), and pattern analysis (Quitkin 1999) are statistical methods to assess differential dynamics of improvement under antidepressants. There are additional important clinical and methodological issues relevant to treatment with antidepressant drugs.

One of these issues is the question whether patients with mild or moderate depression should be pharmacologically treated. This issue is linked to the issue of reliable screening and diagnosis.

There is a lack of convincing trials on the comparative and combined effectiveness of biological and psychological treatments in depression. Despite the long experience in the use of antidepressants and psychotherapy, there is no evaluation standard for these issues.

Another controversy directly related to mental health policy is the recommendation of antidepressants to prevent manifestation of major depressive disorder at an early stage, when only prodromal or single symptoms are present. This issue also includes the questionable prescription of antidepressants for children and adolescents—a generally underinvestigated field.

Secondary prevention, i.e. reducing the risk for recurrence and chronicity of depressive disorders, also raises the question of how long antidepressants should be prescribed after remission of the target symptoms. The potentially advantageous prophylactic use of antidepressants is challenged by the risk of triggering switches into manic states, particularly in bipolar disorder. Decisive studies are required to demonstrate which antidepressants have the lowest risk for mood switches and phase acceleration.

6. *Conclusion and Future Prospects*

Today, a respectable armamentarium for the treatment of depressive disorders is at hand, and antidepressants are standing in the front line. When used adequately, the available antidepressant drugs are effective in reducing an enormous burden of costs, distress, suffering, and death.

Notwithstanding the tremendous progress in the field of antidepressant drugs the ultimate goals have not yet been reached. The key objectives of antidepressive drug treatment remain as follows:

(a) to reduce and ultimately to remove all signs and symptoms of the depressive syndrome in the largest possible proportion of patients in the shortest possible time;

(b) to restore occupational and psychosocial function to that of the asymptomatic state;

(c) to attenuate and ultimately to eliminate the risk of relapse, recurrence, and chronicity.

There is still an urgent need to develop new and better treatments for depression, including drugs and other biological treatments, psychotherapeutic strategies, and combinations of some or all of these. The benefit of having a wide array of antidepressant drugs and other treatment options is based on evidence that there may be different forms of the illness, which respond to different mechanisms of action. More research is needed to develop predictors of differential responsiveness in order to tailor antidepressant treatment to the individual patient's requirement (Preskorn 1994).

Depression is—among other neurological and psychiatric disorders—at present a treatable, but still little understood, disease. From the very beginning, the history of antidepressive drug treatment was accompanied by scientists' expectation that the knowledge about the mechanism of action of antidepressants would lead to an extended insight into the biological nature of the treated illnesses. In fact, the progress in research on antidepressants had an immense impact on biological psychiatry. However, there is still a lack of knowledge concerning the molecular understanding of the etiology and pathophysiology of depression, and the complex mechanisms of action of antidepressant drugs are still enigmatic.

Despite the chemical and pharmacological diversity of the presently available antidepressant drugs, they have some properties in common. The differences between antidepressant drugs lie mainly in side effects, tolerability, and the potential for pharmacokinetic interactions. Regarding the mechanism of action of the diversity of available antidepressants, a final common pathway is likely to be involved. A modulation of a number of different types of neurotransmitter receptors seems to contribute to the long-term effects of antidepressants, in most cases irrespective of the acute effects of the diversity of available drugs on a particular system. On a higher level of action, shared effects of different antidepressants on second-messenger systems and gene expression can be assumed but specific data are not available. For instance, an active cross-talk between serotonergic and noradrenergic receptors which can be modulated by a protein kinase has been demonstrated. Furthermore, a link between central glucocorticoid receptors and neurotransmitter receptor systems is likely.

Promising new candidates of alternative mechanisms of action for antidepressive drugs include substance P (neurokinin I) receptor antagonists (Kramer et al. 1998), neuropeptide Y agonists (Stogner and Holmes 2000), and corticotropin-releasing hormone receptor (CRH-R) antagonists (Holsboer 1999). Despite encouraging initial results from animal and

human studies, a firm conclusion regarding safety and efficacy of these new therapeutic principles cannot yet be drawn.

Rational antidepressive drug development in the future will presumably be guided by the tools of modern drug discovery, which comprise:

(a) the genetic understanding of depression and environmentally influential factors;

(b) animal models that closely reflect the pathophysiology of depressive signs and symptoms;

(c) complex cellular models with cloned transmitter and neuropeptide receptors;

(d) a detailed atomic-level understanding of the target molecules where the drugs are intended to act (Cooper et al. 1996).

Brain imaging techniques and molecular genetic methodology will not only expand our knowledge about biological manifestations of depressive disorders. These techniques will also help increasingly in monitoring and understanding the biological effects of antidepressants *in vivo*.

It does not seem overoptimistic to expect that with the rapid progress of genetic and molecular research, the core pathophysiology of depressive disorders could be elucidated in the near future. This attainment alone could provide a rationale for a causal therapy of depression and related disorders, and for logically designed antidepressant drugs with a clear therapeutic superiority over those available today.

Superiority should be in future understood not only in terms of higher efficacy, but also with regard to lower toxicity, better subjective tolerability and treatment adherence, improved functional outcome, and quality of life. The increasingly available evidence should serve as the basis for the prescription of a particular drug to a given patient. Long-term mental health issues for the development of better antidepressants should be the cost-effective reduction of patients' disability, morbidity, and mortality caused by depression and related disorders.

See also: Childhood Depression; Depression; Depression, Clinical Psychology of; Psychological Treatment, Effectiveness of; Psychotherapy and Pharmacotherapy, Combined

Bibliography

Benkert O, Laakmann G, Ott L, Strauss A, Zimmer R 1977 Effect of Zimelidine (H 102/09) in depressive patients. *Arzneimittel-Forschung* **27**: 2421–3

Benkert O, Hippius H 1996 *Psychiatrische Pharmakotherapie*. Springer, Berlin

Bloom F E, Kupfer D 1996 *Psychopharmacology: The Fourth Generation of Progress*. Raven Press, New York

Carlson A, Wong D T 1997 A note on the discovery of selective serotonin reuptake inhibitors. *Life Sciences* **61**: 1203

Cooper J R, Bloom F, Roth R H 1996 *The Biochemical Basis of Neuropharmacology*. Oxford Academy Press, New York

Feldman R S, Meyer J S, Quenzer L F 1997 *Principles of Neuropsychopharmacology*. Sinauer Associates, Sunderland, MA

Holsboer F 1999 The rationale for corticotropin-releasing hormone receptor (CRH-R) antagonists to treat depression and anxiety. *Journal of Psychiatric Research* **33**: 181–214

Khan A, Warner H A, rown W 2000 Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials. *Archives of General Psychiatry* **57**: 311–17

Kramer M S, Cutler N, Feighner J, Shrivastava R, Carman J, Sramek J J, Reines S A, Liu G, Snavely D, Wyatt-Knowles E, Hale J J, Mills S G, MacCoss M, Swain C J, Harrison T, Hill R G, Hefti F, Scolnick E M, Cascieri M A, Chicchi G G, Sadowski S, Williams A R, Hewson L, Smith D, Rupniak N M et al. 1998 Distinct mechanism for antidepressant activity by blockade of central substance P receptors. *Science* **281**: 1640–45

Kuhn R 1957 Über die Behandlung depressiver Zustände mit einem Iminodibenzylderivat G 22355. *Schweizerische Medizinische Wochenschrift* **87**: 1135–40

Loomer H P, Saunders I C, Kline N S 1957 A clinical and pharmacodynamic evaluation of iproniazid as a psychic energizer. *Psychiatric Research Publications of the American Psychiatric Association* **8**: 129

Nathan P E, Gorman J M 1998 *A Guide to Treatments That Work*. Oxford University Press, New York

Preskorn S H 1994 Antidepressant drug selection: criteria and options. *Journal of Clinical Psychiatry* **55** (9 Suppl A): 6–22

Priest R G, Hawley C J, Kibel D, Kurian T, Montgomery S A, Patel A G, Smeyatsky N, Steinert J 1996 Recovery from depressive illness does fit an exponential model. *Journal of Clinical Psychopharmacology* **16**: 420–24

Quitkin F M 1999 Placebos, drug effects, and study design: a clinician's guide. *American Journal of Psychiatry* **156**: 829–36

Schatzberg A F, Nemeroff C B 1998 *Textbook of Psychopharmacology*, 2nd edn. American Psychiatric Press, Washington, DC

Stassen H H, Delini-Stula A, Angst J 1993 Time course of improvement under antidepressant treatment: a survival-analytical approach. *European Neuropsychopharmacology* **3**: 127–35

Stogner K A, Holmes P V 2000 Neuropeptide-Y exerts antidepressant-like effects in the forced swim test in rats. *European Journal of Pharmacology* **387**: R9–R10

Wells K B 1999 Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *American Journal of Psychiatry* **156**: 5–10

O. Benkert, A. Szegedi, and M. J. Müller

Antiquity, History of

Antiquity is commonly understood as the history of the Mediterranean world and its neighboring regions before the Middle Ages. The concept is a legacy of the tripartite division of history which has developed since the Renaissance. There was always a tension between mere concentration on Greek and Roman history as representing 'classical' antiquity, and an approach including the high cultures of Egypt and the Near East. On the one hand, the Assyrian and Persian

empires were due to the theory of the four world monarchies (adopted from oriental sources in the Book of Daniel) part of a Christian view of history. On the other hand, there was a strong tradition (going back to Aristotle) of contrasting Western freedom with oriental despotism where subjects lived under slave-like conditions. Until the nineteenth century oriental history could only be viewed through classical and biblical sources. Since then, decipherment of hieroglyphs and cuneiform as well as excavations have led to a reconstruction based on monumental sources. Their interpretation, however, required a linguistic and archaeological specialization which led to a new departmentalization between classical and oriental studies.

1. *Early Greece*

Until the later nineteenth century, the beginnings of Greek history were equated with the first literary sources, the Homeric epics (eighth century BC). Since then excavations have enabled detection of the main structures of the Bronze Age civilizations—the 'Minoan' in Crete and the 'Mycenaean' in parts of mainland Greece, where great palaces served as centers of a redistributive economy. Those cultures were finally destroyed during the twelfth century BC by waves of invaders whose identity is far from clear, although they were probably not identical with the Dorian Greeks who afterwards settled in parts of mid-Greece, the Peloponnese, and Crete. The structures of the 'Dark Ages' up to the eighth century can only be reconstructed conjecturally. There was no need to build up greater units, since there were no attempts by foreign powers to establish control. Social life took place in village communities with some collective organization for warfare and regulation of internal conflicts. The leading part was played by well-to-do landowners, who had to rely on the support of the peasants and permanently competed for a precarious leadership within the community.

The so-called archaic age from ca. 750 BC to ca. 500 BC is the formative period of Greek culture. The adaptation of the Phoenician alphabet was a cultural breakthrough, the epics of Homer and Hesiod (ca. 700 BC) canonized the ensemble of gods and heroes, and Panhellenic festivals such as the Olympic Games presented places of contact for aristocrats from all over the Greek world. Cultural unity was also fostered by the 'colonization' which led to numerous Greek settlements on the coasts of Asia Minor, Sicily, southern Italy, North Africa, and in the Black Sea area; contacts with indigenous populations reinforced the consciousness of an ethnic homogeneity based on language, religion, and customs. Colonization reacted to population growth, aiming at occupation of arable land. Newly created settlements became independent units. Constructing a new community implied a process of rationalization by which the order at home was also understood to be subject to change by deliberative decisions.

2. *The City-state*

The city-state (*polis*), first developed in colonization areas, became the typical form of organization in most (yet not all) parts of mainland Greece. It consisted of a fortified center on a hill, a residential town at the foot of a hill, and the agricultural hinterland. Its political unity was symbolized by a meeting-place and communal temples. Collective decision making was formalized by the establishment of magistrates with defined competences and terms of office, councils with administrative functions, and assemblies which took decisions about war and peace, and the administration of justice. For a long time, public functions could only be fulfilled by well-to-do people, but they had to take into account the opinion of the peasants who made out the bulk of the military.

The development toward formalized political structures overlapped with increasing tensions between rising and declining families of the social elite on the one hand, and between the aristocracy and the peasantry on the other. Considerable parts of the peasantry were in danger of indebtedness and finally enslavement. There were demands for cancellation of debts and redistribution of land. In several cities internal tensions led to the seizure of power by a tyrant, an aristocrat who controlled the city through a mixture of military force and patronage. Tyrants improved the condition of the peasantry, developed the infrastructures of the cities, and fostered communal identity by establishing central cults. Only later were tyrants considered to be exercising arbitrary rule incompatible with the freedom of citizens. In certain cities, 'legislators' were entrusted with the formulation of rules which should stabilize the community.

At the end of the archaic age, a firm political structure was achieved in most cities, with considerable differences as to the degree of participation open to the bulk of the citizenry. Priests were responsible for the administration of cults and temples but did not dispose of political power. There were hundreds of autonomous *poleis*, most of them with a territory of 50–100 km^2 and some hundred or a few thousand (male, adult) citizens. There were, however, some exceptions, in particular Sparta and Athens.

2.1 *Sparta*

Sparta had embarked on an unparalleled expansion. During the second half of the eighth century BC, it conquered Laconia and Messenia, the southern part of the Peloponnese (a territory of ca. 8,500 km^2). The subjugated populations were treated as helots, state

serfs, or as *perioikoi* ('those who dwell about'), who lived in communities enjoying local autonomy but bound to follow Sparta's military leadership.

Uprisings among the Messenians in the late seventh century BC led to a total reorganization of Spartan society, which was achieved by a series of reforms probably completed about 550. The citizens (Spartiates, ca. 9,000) received almost equal plots of land, which were tilled by the helots, whose products enabled the Spartiates to serve as full-time hoplites. Every Spartiate underwent a public education and afterwards lived in a warrior community. Society had a uniquely militaristic character, distinguished by an austere life-style, which has provoked admiration as well as abhorrence (from antiquity until the present day). The political system was unique again: there was a hereditary kingship in the form of a dual monarchy (of unknown origin), the kings being the leaders of the army; home affairs were administered by magistrates (*ephors*) appointed every five years and a council of men over 60 elected for life (*gerousia*); decisions on foreign policy were taken by the assembly of the Spartiates. In spite of the ostentatious stress on equality within the Spartiates, there was probably an informally ruling elite, although internal politics remain obscure. During the sixth and fifth centuries BC, Sparta enjoyed remarkable stability, and the strength of its army was unequalled in the Greek world.

2.2 Athens

Athens was an exceptional case due to the fact that the whole territory of Attica (ca. 2,500 km^2) had been formed into one political unit during the Dark Ages. Solon's legislation in 594 BC ensured social peace by establishing that the indigenous population should not be enslaved (which in turn created the demand for chattel slaves from outside), and by codifying large parts of Athenian law. The tyranny of the Peisitratids (561–510 BC) had further unifying effects. In 508/7 BC Cleisthenes laid the foundations for a political system based on the participation of the citizenry. A reorganization of the subdivisions of the citizenry made newly constituted local communities (demes) the backbone of the system. A council of 500, consisting of members delegated by every deme, took on the preparation of assembly meetings and a number of administrative functions. The council's importance grew, especially when the older council, the Areopagus, consisting of life members, lost political influence in 462/1 BC. During the fifth century, appointment by lot was introduced for councilors and the majority of (several hundreds of) magistrates, as well as daily allowances for them and the hundreds of laymen who served as jurors in courts that decided in criminal and civil cases. Every honest citizen (above the age of 30) was considered as being able to undertake public functions, and, due to the great number of positions that had to be filled each year, a great part, if not the majority, of the 30,000–40,000 citizens must have acquired this experience.

All important decisions, especially on foreign affairs, were taken in the assembly, which met regularly. Leadership in the assembly was assumed by generals (strategoi) whose office (a board of 10) constituted an exception, since they were elected (not appointed by lot) and re-election was permitted. But only a few generals acquired a reputation through military achievement and rhetorical ability such that the assembly was likely to follow their lead. The trend toward democracy (a term coined in the later fifth century BC) was not accompanied by disputes of an ideological nature. The central position of the assembly was the result of a dramatic increase in military and diplomatic activities, as tremendous successes in this field fostered the legitimacy of the system.

3. The Greek World during the Fifth and Fourth Centuries BC

At the Battle of Marathon (490 BC), Athens achieved a sensational victory over Persian troops. In 480 BC, Persia launched an attack to conquer Greece and was beaten off by an alliance of most Greek states, Sparta taking the lead in the land battles, Athens in the maritime operations (with a great victory in the Salamis naval battle). Due to the strength of its fleet, Athens became the leading power in the Aegean Sea and build up a collective military system from which Sparta kept away. To continue fighting Persia an alliance (the Delian League) was established, with a great number of *poleis* financially contributing to the maintenance of the Athenian fleet. Athens enjoyed a period of prosperity and became a cultural center that in the era of Pericles (the leading statesman since the 440s BC) attracted intellectuals, artisans, and craftsmen from all over the Greek world. Athens' political and cultural hegemony materalized in the great temple buildings on the Acropolis and in the great festivals including the theatre performances.

Athens' claim for leadership was based on its role in the Persian Wars but operations against Persia ceased after ca. 450 BC. Nevertheless, Athens did not allow any city to drop out of the alliance and showed an increasing tendency to interfere in the affairs of member states. Suspicion of Athenian ambitions to dominate all Greece arose in Sparta and its allies (especially Corinth) with the result that minor conflicts culminated in the 'Peloponnesian War' between Athens and Sparta and their respective allies. Both sides probably considered a final confrontation as inevitable and thought to fight a preventive war. The first period of warfare (431–421 BC) led to a confirmation of the *status quo ante*; the second one opened

with Athens' attack on Sicily (415–413 BC), a disaster for the Athenians, and ended finally in Sparta's total victory (404 BC), which was facilitated by Persian financial aid.

Sparta was not able to keep its hegemonic position for much longer. There were quarrels within the ruling group and, due to military losses and a concentration of landed property, the number of full citizens dwindled continually (down to 2,500 Spartiates by 370 BC). Athens partly regained its naval strength, although a new alliance, founded in 377 BC, lasted for only two decades. For a short time, 371–362 BC, Thebes maintained the position of a leading military power. The instability of the power system was increased by Persia's paying subsidies to varying allies, which led to a permanent reversal of coalitions. In the mid-fourth century BC, Macedonia, under King Philip II, entered the scene and tried to acquire a hegemonic position. Athens organized an anti-Macedonian alliance, but this was decisively beaten in 338 BC.

Athenian democracy had survived the defeat in the Peloponnesian War. The terror regime ('Thirty Tyrants') during the final phase of the war had definitively discredited any constitutional alternative. Democracy was restored in 403 BC. Fundamental criticism of political equality as articulated by Plato and Aristotle did not reflect even the opinion of the upper classes. Present scholarship is agreed that there was no decline in political participation. The growing tendency toward a separation of political and military leadership had certain negative effects on military efficiency, but the final defeat by Macedonia was due to the lack of sufficient resources. That Athens (and the other Greeks) should better have accepted a 'national' unity under Macedonian leadership (as was often said in nineteenth-century historiography) is an anachronistic idea, incompatible with the tradition of the autonomous city-state.

4. *The Hellenistic Age*

Macedonia did not formally abolish the autonomy of the *poleis*, but they were put under firm control and lost their capacity to pursue an independent foreign policy. Macedonia wanted to legitimize its leadership by embarking on retaliation against Persia. Philip's plan was realized by his successor Alexander ('the Great'). Heading an army of Macedonian and Greek troops, Alexander conquered Asia Minor, Syria, Egypt, Mesopotamia, and Iran with incredible speed (334–330 BC). Having declared the Panhellenic revenge war terminated, he dismissed the Greek contingents and started conquering the eastern parts of the Persian empire with Macedonian and Iranian troops. In 326 BC he reached the Indus Valley, but his intention to march to the 'edges of the world' failed as his exhausted troops refused. In his lifetime Alexander won a reputation for invincibility; his death in 323 BC, at the age of 33, made him a figure that was reflected upon by philosophers, historians, and rhetoricians. He became the prototype of a world conqueror later rulers would imitate, and a subject of popular romances which would receive adaptations in the major languages of the Middle Ages. Modern historians often cannot resist following their own imaginations with respect to Alexander's 'final plans.'

Alexander took over the provincial system of the Persian empire and considered himself a successor to the Persian Great King and the Pharaoh. His adoption of Persian court ceremonial and attempts to integrate Iranians into his elite units provoked alienation with his Macedonian troops. At the time of Alexander's death there was no successor. Control was taken over by several generals, who acted as governors in various parts of the empire. In a series of wars between them the unity of the empire was broken. Finally, around 280 BC, a new power structure consisting of Macedonia, Egypt under the Ptolemies, and Asia Minor, Syria, Mesopotamia, and parts of Iran under the Seleucids was established. Macedonian generals had succeeded in founding dynasties. In the Ptolemaic and Seleucid kingdoms leading positions in the army and bureaucracy were occupied by Macedonians and Greeks; Alexander's attempts to include native elites were not repeated. Greek-Macedonian and native populations lived under different legal orders. The system was kept together by the king, who was considered absolute ruler and overlord of the territory.

In spite of these limits of integration, there was a tendency toward assimilation of Greco-Macedonian and indigenous cultures, the spread of Greek as an international language, and the spread of the Greek life-style to cities and garrisons with Greek populations. Mercenaries, tradesmen, intellectuals, and artists enjoyed mobility throughout the Mediterranean and Near Eastern world. Alexandria in Egypt became a cultural center where the literary heritage of the fifth and fourth centuries BC was cultivated. These achievements of the 'Hellenistic age' (a nineteenth-century coinage) were still persisting when the Romans established their domination over the eastern parts of the Mediterranean world.

5. *Early Rome and the Middle Republic*

According to late Republican Roman tradition, the city of Rome was founded in 753 BC. Archaeological evidence suggests that settlements on the hills of the future city were established from the tenth to the eighth centuries BC. Formation of a city-state took place considerably later, and was due to strong Etruscan influences, although Rome was never dependant on an Etruscan city. The primordial kingdom, ever in need of cooperation with a landowner patriciate, was finally abolished around 500 BC. After that Rome was dominated by an aristocracy which,

however, changed its character. There were tensions between the patriciate and rising families, as well as peasant demands for land distributions and debt relief. The plebeians, i.e., all nonpatricians irrespective of their social status, combined in a separate organization with magistracy (tribunes of the people) and assembly to achieve access to political functions hitherto monopolized by the patriciate, and to secure the fulfillment of social demands. The 'Struggle of the Orders' is a somewhat dramatic term for a long process in which phases of conflict and reconciliation alternated. Piecemeal changes from the fifth to the early third centuries BC led to a political system with an equilibrium between magistrates, senate (a council of former magistrates sitting for life), and assemblies, which decided on war and peace, legislation, and the election of magistrates. Politics was determined by a nobility formed of the old patriciate and the plebeian elite who now had access to the magistracy. Office tenure for one year, collegiality in all magistracies, a hierarchy of magistracies to be occupied in successive order, and the senate's overall control ensured a certain equality within the ruling class. Ordinary citizens participated through the assemblies. Voting, however, took place in units structured according to wealth and place of residence, which implied a blatant weight in favor of the propertied classes; initiative for legislation was the monopoly of magistrates. Roman assemblies were far from being democratic in the Greek sense. But their very existence and their function as a forum for communicating with the people ensured consensus within the citizenry. The nobility presented itself as a ruling class not indulging in conspicuous consumption, but absorbed by their duties in public office.

Constitutional development and territorial expansion mutually enforced each other; the peasantry serving in the army put pressure on the nobility, who in turn would fulfil their demands by land distributions in newly acquired territories. During the fourth and early third centuries BC, Rome won control first over the neighboring region of Latium, then over mid- and south Italy. This tremendous success was not only due to the strength of a highly disciplined military but also the result of a prudent policy of integration. Whole communities were admitted to Roman citizenship (with or without the right to vote); others gained a special status by which their members could opt for Roman citizenship. All others were treated as allies enjoying autonomy, their obligation to military support being rewarded by a share in the profits of war. Thus Rome could rely on indirect domination based on the cooperation of local elites.

Roman involvement in Southern Italy led to confrontation with the Carthaginians who occupied parts of Sicily. As a result of the First Punic War (264–241 BC) Rome established domination over Sicily, Sardinia, and Corsica. The second war with Carthage (218–201 BC) opened with catastrophic Roman defeats against Hannibal, who had invaded Italy from Spain. But finally the Romans were able to make Hannibal retreat to North Africa, where he was decisively beaten. In the end, Rome extended its domination over Spain. During the second century BC, Rome established its hegemony in wider parts of the Mediterranean area. The Macedonian state was destroyed in 168 BC, Carthage in 146 BC. The Seleucid position in Asia Minor was undermined in favor of the rise of Pergamum, which in 133 BC would fall to Rome.

Beginning with Sicily, newly acquired overseas territories were made provinces under permanent administration by Roman magistrates and 'pro-magistrates' (whose authority was prolonged for that purpose after the regular one-year turn). Governors relied on the cooperation of local elites and, as a rule, did not change internal administration structures, especially the systems of tax collection. There were, however, always complaints about exploitation by governors who impudently enriched themselves.

Rome's overseas expansion did not follow a blueprint, and actual decisions were not taken primarily with respect to the acquisition of resources. But a tendency toward 'imperialism' was inherent in a system in which magistrates had just a short period to gain prestige and wealth, and soldiers were accustomed to participate in booty. The more conquest, the more chances to 'defend' Roman interests; Rome's determination to impose its will upon other states was compatible with its self-representation as fighting only 'just wars.'

6. *The Late Republic*

The establishment of a large empire had unforeseen repercussions at home. Part of the vast resources available because of the conquests was invested in Italian land. Aristocrats acquired great landed estates which were worked by slaves for the production of cash crops. Peasants were displaced and no longer met the property qualification for army service. This picture, going back to ancient narrative sources, represents a general trend, although it certainly needs qualification to take account of regional differences. The land reforms by Tiberius and Gaius Gracchus (133 BC and 123/2 BC, respectively) had only limited effects. Later the recruitment system was changed: volunteers without property qualification were enrolled; these soldiers showed special loyalty to their commanders, who in turn felt obliged to secure landed property for veterans. Sulla, who came to power (82 BC) after a series of civil wars (starting with the Italian allies' forcing their access to Roman citizenship) and the war with Mithridates of Pontus (who had attacked the Romans in Asia Minor and Greece), was the first to satisfy the soldiers' demands by large-scale confiscations in Italy. Sulla's attempt as *dictator* (tra-

ditionally only a short-time supreme commander) to strengthen the senate's authority was not really successful. It became necessary to entrust the great general Pompey with long-term command, but after his complete reorganization of Roman domination in Asia Minor and Syria (66–62 BC), Pompey could not arrange for the land allotments he had promised his soldiers. He engaged in a coalition with the ambitious Caesar, who used his provincial command in Gallia for a large-scale conquering campaign (58–51 BC). At last Pompey realigned with the senate to crush Caesar. Caesar won the ensuing 'civil war,' which was fought all over the Roman empire (49–45 BC). Being in sole power, Caesar was finally appointed dictator for life and the senate indulged in conferring extreme honors upon him.

Caesar's ostentatious disgust with aristocratic equality caused his murder by a group of senators (44 BC). His supposed 'final plans' have bedeviled generations of historians, but it seems doubtful whether he had had a clear-cut conception of a lasting reorganization of the political system. That a monarchical head of the empire was a historical necessity (as nineteenth-century scholarship put it in Hegelian terms) was an idea apparently not familiar to contemporaries, since military success over centuries had been achieved by an aristocracy. But the traditional republic could not be restored either, and after a further series of civil wars a new system emerged.

7. *The Early Empire*

The new order was established by Octavian, Caesar's personal heir, on whom the senate conferred the title 'Augustus' in 27 BC. Through cautious moves he constituted a monarchy in disguise, proclaimed as a 'restoration of the republic.' The emperor's formal authority consisted of a bundle of competencies derived from traditional magistracies, which allowed him to control all affairs. Continuance of magistracies and senate, and employment of senators as military commanders and governors ensured the loyalty of the senatorial elite, although the subsenatorial 'equestrian order' got a share in administration. The emperor played the part of the patron to the capital's population by grain and cash distributions as well as public games. The problem of veteran settlement was solved by the purchase of land in Italy and overseas colonization. Provincial governors still relied on the cooperation of local elites, who were rewarded with legal privileges and offered chances for ascendance into the equestrian and senatorial ranks.

Such a summary of the 'principate' inaugurated by Augustus should not result in ignoring internal tensions. Keeping a republican facade demanded the self-restraints of rulers, but the lack of constitutional controls and, for example, the tendency of Eastern provinces to confer god-like honors on them made emperors show their true faces as autocratic rulers again and again. Provincial government was improved, but charges of corruption and maladministration against governors did not cease (and led to a great uprising in case of the Jews in AD 66–70). The absence of a law of succession reflected the tension between meritocratic and dynastic elements within the emperorship. Designation of a successor was not necessarily decisive—at the crucial moment, the senate, the imperial elite troops (praetorians), even the capital's population might be influential, or a provincial army might promote a pretender.

All in all, during the first and second centuries AD the empire enjoyed stability and prosperity which was manifest in the sophistication of urban culture in all parts of the Roman world. Expansionist policy was given up in favor of territorial consolidation (although Britain was subjected to Roman rule throughout the first century). The majority of troops were stationed in fortified positions near the frontiers, especially the Rhine-Danube and Euphrates borders. During the third century AD, Rome came under increasing pressure by the Goths (who were invading the Danube area) and the new Sassanid dynasty in Persia. The situation was aggravated by instability on the throne: again and again army units proclaimed officers as emperors; most of them met violent deaths after a short time.

8. *The Later Empire*

After half a century of turmoil, Diocletian (AD 284–305) embarked on a comprehensive reform. It was based on devolving the imperial office on two senior and two junior emperors (as prospective successors), who would each have authority in a part of the empire but would govern jointly. The emperorship was symbolically enhanced by attributing a 'sacral' character to it and adopting oriental court ceremonial. However, regulated succession failed; from AD 305 onwards various pretenders fought each other, until Constantine and Licinius could consolidate their positions in the West and East respectively. Finally Constantine became sole ruler (AD 324).

The administrative reforms started by Diocletian were continued by Constantine. They included a reorganization of the army (now divided into field army and border troops) and of the provincial system (now including Italy), a separation of military and civil administration, and a new taxation system to meet the demands of the army and the enlarged governmental apparatus. These led to considerable consolidation although the intention to regulate more and more features of social life was partly defeated by corruption and maladministration, as well as the attempts of various social groups to escape the pressure put on them from above—local grandees sheltered their dependants from the state's demands and exploited

them at the same time. Citizenship (since AD 212 extended to all inhabitants of the Empire) lost its importance, as can be seen in the harsh penal system with distinctions according to the social status of citizens. Peasants, who had to take the burden of taxation, were reduced to a serf-like status; craftsmen were bound to their occupation and place of residence. These points signify tendencies; they should, however, not be understood as covering the multifariousness of social reality. There were considerable differences between the various regions of the empire, and discrepancies between the increasing legislation and its enforcement. The political and societal order of the post-Diocletian era had manifestly changed in comparison with the early empire, but there were also remarkable elements of continuity, for example with respect to higher education, local administration, and civil law. The traditional label of 'despotism' is not helpful.

A most important change happened with respect to the Christians. From the first century AD onwards Christianity had spread through all parts of the empire, and had taken on firm organization with the episcopal constitution. Christians who did not participate in the ruler cult were considered disloyal citizens. During the first two centuries AD, there were sporadic persecutions of Christians, mostly inspired by local communities. The Christians reacted, on the one hand, by praising martyrdom; on the other, by stressing the coincidence of Christ's coming and the Augustan reign of peace and (answering the charge of being a lower-class religion) by adopting 'pagan' literary culture. Since the mid-third century AD, rulers had considered the restoration of state cults necessary for the preservation of the empire. That led to systematic persecutions under the reigns of Decius (AD 250), Valerian (AD 257), and Diocletian (AD 303); indictments of clerics, prohibition of church meetings, confiscation of holy scripts and church property were aimed at destroying the church's very existence. As a whole the measures failed and were stopped in AD 311. After AD 312, Constantine pursued a policy not only of toleration but also of actively promoting the church by donations, granting privileges to clerics, conceding a public role to bishops, promoting Christians in public offices, and finally giving the new capital Constantinople, founded at the site of Byzantium (AD 324), a Christian outlook. Constantine's religiosity and the motives of his policy remain obscure. The strengthening of imperial authority by a theology that declared the emperor God's vicar on earth had to be paid for with involvement in inner-church conflicts, especially the Arian schism, which was not to be solved by the theological compromise found at the Council of Nicaea (AD 325) under Constantine's personal lead. Finally, Theodosius declared Christianity the state religion (AD 380) and pagans, heretics, and Jews became subject to persecution. Bishops played an important role as the heads of cities, and some, especially in the West, in imperial politics as well. The increasing interweaving of state and church, however, fostered ascetic monasticism.

The division of the empire into Western and Eastern parts after the death of Theodosius (AD 395) would last, although the fiction of unity was never given up. For a number of reasons the Eastern Empire did far better with the pressures from migrating Germanic tribes, especially the Goths. The weakness of the Western empire is symbolized by the Visigoths' sack of Rome (AD 410). The end of Western emperorship in AD 476 is traditionally taken as signifying the end of Antiquity. Whereas in the West, Germanic states were established, the East took up the Roman heritage (as symbolized by Justinian's codification of Roman Law, AD 534) and preserved its continuity through the Middle Ages.

9. *The Legacy of Antiquity*

In spite of manifold contacts with the oriental world, the political and cultural achievements of Greece and Rome were products of indigenous development. The notions of citizenship and individual rights, the ideas of republicanism and democracy, as well as universal emperorship, the continuity of Roman law and of the church (including the use of the Latin language), Christian preservation of 'pagan' literature and other traditions became formative for European culture through structural continuities and conscious renaissance. In this sense, it is still appropriate to treat Greco-Roman antiquity as a historical epoch of its own.

See also: Christianity Origins: Primitive and 'Western' History; Democracy, History of; Democratic Theory; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Imperialism, History of; Republicanism: Impact on Social Thought; Warfare in History

Bibliography

Brown P R L 1978 *The Making of Late Antiquity*. Harvard University Press, Cambridge, MA

Brown P R L 1992 *Power and Persuasion in Late Antiquity. Towards a Christian Empire*. University of Wisconsin Press, Madison, WI

Cambridge University Press 1970 *The Cambridge Ancient History*, new edn. Cambridge University Press, Cambridge, UK

Gruen E S 1984 *The Hellenistic World and the Coming of Rome*. University of California Press, Berkeley, CA

Meier C 1980 *Res publica Amissa. Eine Studie zu Verfassung und Geschichte der späten römischen Republik*. Suhrkamp, Frankfurt am Main, Germany

Millar F G B 1977 *The Emperor in the Roman World (31 BC–AD 337)*. Duckworth, London

Momigliano A D, Schiavone A (eds.) 1988–93 *Storia di Roma*. Einaudi, Torino
Nicolet C 1977 *Rome et la conquête du monde méditerranéen 264–27 avant J.-C. I: Les structures de l'Italie romaine*. Presses Universitaires de France, Paris
Nicolet C (ed.) 1978 *Rome et la conquête du monde méditerranéen 264–27 avant J.-C. II: Genèse d'un empire*. Presses Universitaires de France, Paris
Osborne R 1996 *Greece in the Making 1200–479 BC*. Routledge, London
Will E 1972 *Le monde grec et l'orient. I: Le Ve siècle*. Presses Universitaires de France, Paris
Will E, Mossé C, Goukowsky P 1975 *Le monde grec et l'orient. II: Le IVe siècle et l'époque héllenistique*. Presses Universitaires de France, Paris

W. Nippel

Anti-Semitism

The term 'Antisemitism' was first introduced into public discourse in Germany in the 1870s and thereafter quickly replaced all previous words denoting hostility to Jews, both in and out of the German *Kaiserreich*. Within months it was also applied to past cases of Jew-hating and to the historiography of anti-Jewish attitudes and policies, written by Jews and non-Jews alike. Scholarly attempts to restrict the meaning of Antisemitism either to the modern era or to that kind of anti-Jewish sentiment relying on racism, have usually failed. Thus, despite its limitations, the term continues to dominate the discussion of all anti-Jewish attitudes and measures in every period, in all cultures and in every geographical region.

1. The Term and its Early Applications

During the late eighteenth century, a group of mostly ancient Middle Eastern languages was first named 'Semitic' and soon afterwards the term 'Arian' was coined to denote another vaguely defined group of languages, later also known as Indo-Germanic. By the mid-nineteenth century, both terms began to be applied, as for instance by Ernst Renan, to peoples and to ethnic groups, too. The adjective 'Antisemitic' appeared during the 1860s, in at least two major German reference works, but it was only in Berlin, during the late 1870s, in the wake of a violent anti-Jewish campaign, that the term really entered public discourse. It was apparently the *Allgemeine Zeitung des Judentums*, in an article published on September 2, 1879, that first used the term in print. It reported plans to publish an Antisemitic weekly by Wilhelm Marr—by then one of the more outspoken anti-Jewish journalists in the German–Prussian capital. The circle around Marr may have indeed applied the term, though it was only somewhat later that he himself began to make use of it. By the end of the same month, the *Antisemiten-Liga* has been established, and despite its meager performance aroused some interests in liberal and Jewish circles. The debate that ensued, especially after the publication of Heinrich von Treitschke's article, *unsere Aussichten*, in the *Preussische Jahrbücher* of November 15, 1879, was known as the *Antisemitismusstreit*, and the petition against the legal and social position of Jews in Germany, circulating a year later—as the *Antisemiten-Petition*.

2. The Novelty of Modern Antisemitism

The term apparently served a needed function. It seemed to have indicated a new anti-Jewish attitude, unhinged from the old traditional Jew-hatred and directed against the modern Jewish community, now in possession of full civil rights and on the way—so it seemed to many—to full social integration. The new word made it possible to regard Jew-hating as a full-fledged ideology, presumably like Liberalism or Conservatism. Those who made it the focus of their overall social thought tried to explain it by both their own particular misfortunes and all evil in the world at large. By the late-nineteenth century, a full-blown conspiracy theory, identifying the Jews as an imminent danger to civilization and as enemies of all culture was added to this *Weltanschauung*; salvation finally meant freeing the world from this particular threat. Friedländer (1997) argues that such an encompassing kind of 'redemptive Antisemitism' was first elaborated by the members of Richard Wagner's circle of friends and admirers in Bayreuth, inspired by Houston Stewart Chamberlain's *Foundations of the Nineteenth Century*. An equally ambitious ideological system was more or less simultaneously advanced by Eugen Dühring's *Die Judenfrage als Rassen- Sitten- und Kulturfrage* (1881), while another version was formulated in France by Eduard Drumont in his widely read and much appreciated *La France Juive* (1886).

It has often been argued that the new term for what seemed only another version of age-old hostility to Jews did in fact signal the emergence of a new phenomenon. Its novelty was two-fold: first, it was based on racism instead of the old religious grounds. Second, it indicated the growth of a new political movement, whose purpose was to reverse the legal equality of Jews, sometimes even to rid Germany or France of them. To be sure, racial theories did reach a new level of presumed scientific precision at that time, and by the 1870s they were fairly well known in wide circles of the European educated public. Nevertheless, none of the Antisemites before World War I relied exclusively upon racial arguments. Jews continued to be attacked for their economic role as speculators, blamed for destroying the livelihood of small artisans and shopkeepers, and above all accused of destroying

the unique culture of the people amongst whom they dwelled. Moreover, the religious impulse behind the Antisemitism of these days should not be underestimated (Tal 1975). Even the overtly anti-Christian Antisemites displayed an eschatological fervor that was far from being wholly secular in tone or in character. Racism did not in fact replace previous Antisemitic views, but was crafted upon them.

The political organizations of that period, too, were not entirely new, nor were they of great importance at this stage. Toury (1968) has clearly shown the *political* function of Antisemitism in the revolution of 1848/1849 and in fact, even in Antiquity or in medieval Europe, many incidents of extreme Antisemitism were politically motivated and/or manipulated. Furthermore, the Antisemitic parties in Germany before Nazism, both those of the social-conservative persuasion and the more radical, oppositional ones (such as Theodor Fritsch's *Antisemitische deutsch-soziale Partei* or Otto Boeckel's *Antisemitische Volkspartei* later known as the *Deutsche Reformpartei*) had a very uncertain existence. Despite their fiery rhetoric they rarely managed to present a united front and even during their heyday, with 16 representatives in the *Reichstag*, they remained entirely ineffective. In France, too, public Antisemitism had a short blooming during the Dreyfus affair, but was of no real consequence as a parliamentary force. The strength of Antisemitism is better measured by the degree of its infiltration into the established parties, such as the *Deutsch-konservative Partei* since 1892, and the various associations and interest groups, such as the *Bund der Landwirte*, the *Deutschnationale Handlungsgehilfenverband*, and some of the students organizations (Jochmann 1976). The powerful *Pan-German League* too, added Antisemitism to its aggressive, expansionary nationalism, especially during the later years of World War I.

Despite the strength of the Antisemitic tradition, many historians insist upon the novelty of its 'post-emancipatory' stage (Rürup 1975). Arendt (1951) indicated the changing position of Jews within the emerging modern national state as the main prerequisite for Antisemitism in modern times. A strong historiographical trend has insisted upon the particular circumstances, socioeconomic or political, that brought about its rise. What may likewise be considered unique about this so-called 'Modern Antisemitism' was its link to a particular cluster of other, mainly cultural tenets and beliefs. Even the small, openly Antisemitic parties in Germany were never devoted to achieving anti-Jewish measures only. They all clung to monarchical and nationalist tenets and worked for a variety of social policies, usually in support of one of the pre-modern economic sectors. They urged control of what they deemed 'unworthy capitalist competition,' and propagated other explicitly anti-modern social and cultural views. By the end of the century, accepting Antisemitism became a code for the belief in all of these (Volkov 1990). In the liberal to mildly conservative atmosphere of pre-World War I Europe, supporting Antisemitism meant opposing the status quo, rejecting democratization, and allying oneself with nationalism and an imperialist foreign policy. In France, too, Antisemitism meant more than hostility to Jews. *Dreyfusards* were identified with the republic and its dominant values, while their opponents basically represented the antirepublican, catholic front.

3. *The Functions of Jew-hating*

3.1 *The Middle Ages*

With the possible exception of the ancient, pre-Christian world, anti-Jewish positions were always symbolic for more comprehensive views and associated with issues well beyond and outside the so-called 'Jewish question.' In early Christianity, the need to distinguish true believers from Jews, who rejected the messianic message of the new religion, was naturally very keenly felt. A complete disregard of Judaism was unthinkable at the time, since the new faith accepted the holiness of the Old Testament. By the twelfth century, Jews were considered an integral part of Christian society, fulfilling two essential functions within it, both in the present life of the church and in its sacred history. First, they were regarded as witnesses to the antiquity and truthfulness of the Bible, 'living letters of the law' (Cohen 1999). Second, their low status was seen as a proof of their anachronism, a consequence of their theological stubbornness.

Furthermore, the symbolic meaning of the Jew has had a dynamism of its own. By the time of the early Crusades, minor fluctuations turned into a major change, considered by some the most crucial break in the history of Antisemitism (Langmuir 1990). From that time onward, Jews became targets of direct attacks, both in theory, that is, within a new theological discourse, and, more significantly no doubt, in practice. A series of bloody assaults on their communities sought to achieve mass conversion, and, as this goal was repeatedly frustrated new accusations against them, concerning ritual murder and later also Host-desecration and well-poisoning, justified violent acts of revenge against them. In an atmosphere of radical struggle against the infidel—Moslems in the outside world and new heretics at home—the fight against Jews and Judaism received a new impetus. Stress was put on Jewish post-Biblical literature, especially on the Talmud, perceived now as an esoteric Jewish document, distorting the original message of the Old Testament and actively scheming against Christianity. This sometimes led to the burning of the Talmud, as in Paris, 1240, but actual, physical attacks

on Jews, carried out in numerous northern German and French towns at about the same time, are more often explained by socioeconomic rather than theological factors. As moneylenders and especially as pawnbrokers Jews were all too often hated by the peasantry and exploited—politically and economically—by landlords. In the aftermath of the Black Death, during the mid-fourteenth century, new massacres were clearly a part of an uncontrolled social upheaval. At the same time, the popular demonization of Jews continued unabated. They were seen as truly Satanic, infecting the healthy body of Christian society, arousing that dangerous mixture of hatred and fear that has since been typical of Antisemitism.

3.2 *The Early Modern Period*

The growing stress on rationalism in medieval thought, as shown by Funkenstein (1993), was no defense against Antisemitism. Neither were Humanism and the Protestant Reformation reliable bearers of toleration. Most humanists were either indifferent to Jews or fell under the influence of ancient anti-Jewish writers. While the growing interest in Hebrew and in ancient Hebrew texts led some to reconsider their negative attitudes towards Jews, others reaffirmed them. In contrast, the position of Lutheranism was more consistent. Luther himself had at first hoped to Christianize the Jews as part of reforming corrupt Catholicism, but he soon gave up that project and turned against them with unparalleled vehemence. Historians tend to see in his anti-Jewish writings the culmination of medieval hostility towards Jews rather than the opening of a new era (Oberman 1984). The demands to shake off the constraints of a previous faith legitimized previous doubts within Christianity and under the circumstances reformers could not afford to relax the boundaries that had for so long separated it from Judaism. Once again, they urgently needed to distinguish the true from the false faith, God from Satan. Jews continued to serve as a tool in building up their new identity.

At the same time, the Reformation and the ensuing religious wars deeply changed the nature of the prevailing political system in Europe. The independent secular state now became the seat of a single religion, in which—ideally at least—the crown ruled over a homogeneous society of subjects/believers. Exceptions could at best be tolerated, as under Calvinist governments, or at worst repressed and expelled—under Lutheranism and the Counter Reformation. To be sure, the needs of homogeneity were felt within the old Catholic world too, though exact motivations were locally varied, of course. England expelled its Jews in 1290. The kingdom of France ordered their exit first in 1306 and then again in 1394. Most outstanding were developments in the Iberian Peninsula after the *reconquista*, leading to expulsion and/or forced conversions in Spain and Portugal during the last decade of the fifteenth century. In the days of Moslem rule, Christian and Jews did occasionally suffer as adherents of minority religions, but under the ambitious government of the conquering Christian kings their position became practically unbearable. A thriving culture, relatively open to the 'other,' was replaced by a demand for uniformity and the merciless elimination of all unbelievers. The Inquisition, indeed, acted primarily against the 'New-Christians,' suspected of secretly upholding their old faith. But even true *conversos* were frowned upon in a Spain, obsessed with notions of blood and an early version of racism. In some reformed German states, too, rulers could not resist public pressure to expel Jews, as in Saxony (1536), and Bohemia (1541 and 1557), while Pope Paul IV ordered the establishment of the first Ghetto in Rome (1655) and tightened up all restrictions designed to prevent Jewish integration in Counter-Reformation Italian society. The worst physical attack against Jews in this period, however, took place in Eastern Europe. Jews were caught in the Cossacks' revolt of 1648/1649, made responsible for the brutal exploitation of the local peasantry, and fell victims to the fight over the nature of the Polish State. It was only during the next century that toleration began to be acknowledged as a necessary principle in a Europe torn by endless inner strife.

4. *Nationalism and the Fear of Modernity*

Even during the Age of Reason, rationality alone did not suffice for combating Antisemitism. While the English Hebraists of the late seventeenth century positively re-evaluated the contribution of Jewish writings, and a few of them even drew a defense of contemporary Jewry from their scholarly projects, the majority of the so-called Deists concluded that Judaism, past and present, was too particularistic and too unnatural to deserve their respect. Ettinger (1978) suggested, that although theirs was a negligible influence in eighteenth century England, they did effect some of the French *Philosophes*, especially Voltaire, who despised the old Jewish religion as much as the living Jews, with whom he apparently had unpleasant personal encounters. Nevertheless, the significance of the 'turn to rationalism' (Katz 1980) should not be underestimated. In its wake, that all too limited secularism of the Enlightenment, joined to a growing emphasis upon human equality, gradually made possible the acceptance of Jews as subjects/citizens in England, France, and the American colonies of the late-eighteenth century. In Germany, too, despite the twisted course of legal emancipation, Jews entered bourgeois society and were accepted within it to a degree never known before. Economic growth and the increasing influence of the bourgeoisie made for an uncongenial atmosphere to Antisemitism. Interest in

the actual life of Jews, their customs, and their habits, worked against xenophobic generalizations and the exploitation of old hatred for other purposes seemed outdated in the age of progress and liberalism.

Nevertheless, old-style, religiously motivated anti-Jewish positions had not entirely disappeared by then and Judaism continued to be seen as a legalistic, cold, and fundamentally inhuman religion. In addition, the process of making Jews stand for other negative aspects of contemporary European life received a new momentum. As early as 1781, in response to Christian Wilhelm Dohm's book on the *Bürgerliche Verbesserung der Juden*, Johann David Michaelis, an expert on ancient Judaism not previously known to hold Antisemitic views, insisted that Jews could not be equal members in a modern, nationally defined Germany. Precisely at the time when Jews were beginning to accept the option of turning their religion into yet another confession within a more tolerant and efficiently run state, their 'otherness' was being redefined so as to exclude them once again. Just as they were abandoning their separate group-identity in favor of joining the 'imagined' national communities, energetically formed everywhere around them, a new nationalism was making their integration all the more problematic, giving old-style Antisemitism yet another meaning.

As in the past, processes of identity formation and the drawing of new group boundaries required an apparent figure of the 'other' for their completion. Especially in Germany, Jews were made to play the role of the enemy from within. Early nationalism used open antisemitic slogans in trying to draw clear boundaries: Jews were to be excluded for all times from the emerging German nation. Throughout the first two-thirds of the nineteenth century, while on the one hand conservatives were only slightly modifying their traditional negative attitudes to Jews and while on the other hand many liberals were clamoring for their emancipation, nationalists, sometimes even of the liberal persuasion, sought to stress their status as outsiders. The fixed place of Jews in society, obvious in the prenational era was now becoming a major 'problem,' the so-called *Judenfrage*. During the revolution of 1848/1849, while the Frankfurt *Parlament* decided to grant full equality to Jews, opposition was rampant everywhere in the country. Peasants still considered Jews as allies of the oppressive local landlord and the tax-collecting, centralizing bureaucracy and in towns they were seen as representatives of a feared and hated new order. As early as the summer of 1819, while the 'Hep-Hep Riots' spread from Bavaria to the Rhine and then eastward and northward across Germany, Jews were identified as the forerunners of social and economic change. It was in this context too that they were attacked in the ensuing decades (Rohrbacher 1993). No less a thinker than Karl Marx, son of a converted Jew, accepted the symbolic role of Jews within the Capitalist system. In his *Zur Judenfrage* (1844) he envisioned their emancipation as liberation from their own nature, to be achieved only with the final collapse of the bourgeois world-order. By the late-nineteenth century, this was still the position of European Social Democracy, rejecting antisemitism, while objecting to any sign of 'Philosemitism' and negating every notion of a separate Jewish group-identity. At this time, both hostility to Jews and a defense of their emancipation were readily combined with more comprehensive world-views, used as symbols of their overall ethos.

5. *The 'Functionalist' vs. the 'Essentialist' Approach*

This 'functional' approach to the study of Antisemitism, however, has often been criticized, since it describes the dynamics of Antisemitism, not its sources. In applying it, one tends to stress the changing needs of the non-Jewish environment and neglect the search for a single explanation, based on Jewish uniqueness and history. In order to do that it is necessary to go back to a point in time, in which the forces of tradition may be, at least partially, neutralized and begin 'in the beginning.' According to Marcel Simon (1948), for instance, two aspects of Jewish life were the fundamental causes of ancient Antisemitism: their 'separatism' and their unique religion. The two, in fact, are inseparable. From the outset, Judaism made Jews different and adhering to it was the cause of their social isolation and the resentment they aroused among their neighbors. Beyond strict monotheism and the peculiarity of their God, three manifestations of Jewish life were the focus of ancient Judophobia (Schäfer 1997): the abstinence from pork, the upholding of the Sabbath, and the habit of circumcision. As early as the third century BC, one encounters accusations against Jews as misanthropic, cruel, and dangerous. The pagan counter-history to the story of Jewish exodus from Egypt describes them as lepers and unwanted foreigners. Apion of Alexandria has apparently reported stories of Jews worshipping a donkey-head and practicing human sacrifice. Popular outbursts against Jews are known as early as the destruction of their temple in Elephantine in 411 BC, and as late as the riots in Alexandria in 38 AD. While some historians argue that Jews were basically treated no worse than other barbarians, others convincingly show that they were treated, in fact, as 'more barbarian than others' (Yavetz 1997). After all, Jews were known to be an ancient, civilized people and their transgressions could not be as easily pardoned. They further aroused special resentment by their insistence on upholding their special way of life even in Greek and later in Roman exile. Indeed, Theodor Mommsen believed that Antisemitism was 'as old as the Diaspora' and historians who regard Antisemitism as no more than the age-old

'dislike of the unlike,' feel this was surely magnified by the fact of Jewish life among the nations. As in later periods, in Antiquity too, Jewish 'otherness' had been more or less tolerated in a culturally mixed pagan world. But it was 'the idea of a world-wide Greco–-Roman Civilization that made it possible for Antisemitism to appear (Schäfer 1997).

The controversy over the role of the Jews themselves in the history of Antisemitism has not been limited to the study of Antiquity. Medievalists, too, sometimes point out that Jews had some part in inciting hatred against them. It is generally accepted that their particular economic role was the cause of much resentment. Yuval (1993) also argues that close relationships existed between the hope for a 'revenging salvation,' often expressed by medieval Ashkenazic Jewry, and the Antisemitism of that time. He further links Jewish suicide in sanctification of God (*Kidush-ha'Shem*) with the practically synchronous emergence of the ritual murder accusations. While causal relationships are difficult to establish, such interpretations offer a more complex view of both Jewish life and Antisemitism during the Middle Ages.

Within a more modern context, too, the problem of Jewish 'responsibility,' if not out-right 'guilt,' has always accompanied the study of Antisemitism. Jews are known to have often taken the blame upon themselves. Orthodox made the unorthodox responsible for the evil that befell all Jews. Some historians regard Jewish capitalists as the main culprits; others blame Jewish socialists. Zionists considered Jewish life in the *Galuth* (exile) corrupt and reprehensible, and at the height of Jewish integration in German society, before World War I, Jews occasionally reproached themselves for being too successful, too culturally prominent, and too obtrusive. By then, the concept of Jewish self-hate also began to gain currency. In *fin de siecle* Vienna, days before he committed suicide, the young Otto Weininger published his *Geschlecht und Charakter* (1902), explaining that Antisemites were only fighting the feminine-Jewish side within themselves and that this fight had to be joined by every self-respecting Jew.

6. *National–Socialism*

Meantime, Antisemitism was gradually taking on an altogether different guise. In the wake of the 'Great War' most European countries were plunged into a prolonged crisis—economic, social, cultural, and political. Antisemitism indeed seems to thrive under such circumstances. While the new Europe was striving to regulate the treatment of minorities and the Weimar Republic repealed all remaining discriminations against Jews, popular Antisemites everywhere continued to blame them for causing and then profiting from military defeat, the hyperinflation, Bolshevism, and 'international' Capitalism. In *Mein Kampf*, recalling his early encounters with Jews and local Antisemites in prewar Vienna, Hitler finally called for action—in all matters, to be sure—but especially in respect of the 'Jewish Question.' Indeed, historians are still divided as to the exact role of Antisemitism in the Nazi project of exterminating European Jewry. Those known as 'intentionalists' see stressing Antisemitism as the only way of 'explaining the unexplainable' (Kulka 1985. The 'functionalists' introduce into their narrative other elements too, and see the 'Final Solution' as a cumulative response to more immediate problems, devoid of prior definition of aims or methods (e.g., Schleunes 1970). All agree, however, that Antisemitism was rampant among the Nazi 'true believers,' and was a necessary if not a sufficient precondition for the Holocaust. Recently, a middle-of-the-way position, shared by many, seem to have robbed the controversy of its pertinence. Instead, arguments about the relationships between pre-Nazi and Nazi Antisemitism have been renewed. Obviously, no history of the 'Final Solution' can be complete without joining the tradition of European Antisemitism to the unique exterminatory rage of the Nazis. Goldhagen (1996) presented a seamless continuity between the two, while most historians prefer a more differentiated approach. They insist on comparing Germans Antisemites to others, in and out of Europe, and on taking into account the century of essentially successful Jewish life in Germany prior to Nazism. It is also useful to follow Hilberg's (1992) distinction between perpetrators and bystanders. Perhaps the activists were spurred to action by racism, while passive participants, by no means only Germans, still relied upon older, more traditional forms of Antisemitism. In the end, all aspects of Jew-hating were necessary preconditions for the execution of the Nazis' murderous project, though none seems entirely sufficient for explaining it.

7. *Social Scientists Confront the Problem*

Facing the enormity of the problem, scholars outside the discipline of history also attempted to deal with it. As early as 1882, Leon Pinsker, in his path breaking *Autoemanzipation* (1881), diagnosed 'Judophobia' as a psychosis, an inherited pathological reaction of non-Jews to the ghost-like Jewish existence in the Diaspora. At the same time, Nahum Sokolov, another early Zionist, preferred to see in it a 'normal' phenomenon, to be explained in social-psychological terms. These two approaches characterize later efforts by social scientists. Especially in the USA, even before World War II, interest in racism, raised in conjunction with attitudes to blacks as well as to Jews, motivated social scientists to analyze what they then named 'prejudice.' Antisemitism was thus considered a particular instance of a more general phenomenon. Some psychoanalytical concepts, such as the Oedipal complex, seemed readily applicable. On the basis of

Freud's *Moses and Monotheism* (1939) one attempted to associate Jews with the distant punishing father of the Old Testament and Christians with the suffering punished son (Loewenstein 1951). Displaced aggression and anger, or the projection of guilt could also be applied here. Adorno and co-workers published *The Authoritarian Personality* in 1951. The force of their theory was the linkage they presented between a specific personal pathology and a particular form of political and social structure. Later on, however, emphasis shifted to the 'normalcy' of prejudice, adding the figure of the social conformist to that of the pathologically prejudiced, and stressing the prominence of social training and education in the fight against Antisemitism. In his *Towards a Definition of Antisemitism*, Langmuir (1990), summarizing existing theories in the field, suggests a division into 'realistic,' 'xenophobic,' and 'chimerical' Antisemitism. According to him, these represent both historical stages, from antiquity through the early Middle Ages until the irrationalism of the twelfth century, and classification of the different types of Antisemitism arranged by their distance from reality and the depth of anti-Jewish fantasies associated with them. Usually, it seems, all three kinds appear together and it is perhaps their special mix or relative importance that define the intensity of prejudice and the threat inherent in it.

8. *Antisemitism in the Post-Holocaust World*

In any event, in the post-Holocaust world, though Antisemitism has by no means disappeared, there is little evidence for the expansion of its 'chimerical,' virulent variety. But there are a few significant exceptions. These are the antisemitism in Soviet and post-Soviet Russia, as well as in some of the Eastern European countries; the case of the Middle Eastern Arab states, and the sporadic outbursts of anti-Jewish rhetoric and occasional violence characteristic of extreme right-wing parties and groups in the West. Thus, examples of the multiple manifestations of Antisemitism are available today too. Czarist Russia, to take the first case, was known for its repressive anti-Jewish policies during the nineteenth century (Wistrich 1991). While civil equality became the rule in the West, most Jews in Russia were restricted to the 'Pale of Settlement' and subjected to numerous humiliating decrees. The pogroms of 1881, partially condoned by the Czarist government, spread to more than 160 cities and villages and claimed thousands of lives. The later Kishiniev pogrom of 1903 aroused a great deal of indignation, especially outside of Russia, but this too could not stop further attacks upon the Jews. During the Civil War, following the 1917 Revolution, some 100,000 Jews were massacred by the Whites. The Revolution, indeed, put an end to all previous discriminations against Jews. Their religious and institutional life suffered of course from the atheist campaign against all religions, but no open Antisemitism was allowed in Soviet Russia until the late 1930s. After a short respite during the war, in which Jews often felt protected by the government, an official anti-Jewish campaign was pursued vigorously, reaching a peak during Stalin's last years and becoming particularly vicious after the 1967 Arab–Israeli war. Few expected a renewed wave of Antisemitism to characterize the collapse of the Communist regime. But under conditions of economic crisis and political chaos, mainly verbal attacks against Jews became common again. After all, exploiting the Jew as a scapegoat for all misfortunes is a well-known pattern and although it seems rather marginal in most parts of the world today, it has not completely faded away.

Under Islam, to take the second case, while Jews were never free of discrimination, they were only rarely subject to actual persecution (Lewis 1986). Like other non-Moslems, they enjoyed limited rights while their inferiority was formally established and considered a permanent fact of life. Ideological Antisemitism was imported into the Moslem Middle East. Its intensity there is a result of the actual political conflict with the State of Israel. In most countries involved in it, a distinction between Judaism and Zionism is usually preserved, but explaining defeat by reference to 'Jewish power' and some kind of 'Jewish conspiracy' has often proven irresistible. As in the previous Soviet Union, Antisemitism here too usually is directed from above, but unlike the Russian case, it is not based on traditional, popular enmity but on a real, ongoing struggle. It was, in fact, the combination of Soviet and Arab anti-Israeli position that produced the 3379 UN resolution of 1975, declaring that 'Zionism is a form of racism and racial discrimination.' Significantly, the resolution was supported by many developing countries, too, expressing what they saw as solidarity with the Arab world, while adopting Antisemitism as a code for their anti-Colonial, anti-Western attitude (Volkov 1990).

The third focus of Antisemitism in today's world is the right-wing organizations in Europe and in the USA. While America seemed at first an unlikely place for the growth of antisemitism, it has experienced a truly racist wave as early as the aftermath of World War I. Christian conservatives and revivalists of all sorts espoused Antisemitism, Ku-Klux-Klan activists incited against 'aliens,' and the *Protocols* were disseminated by such antisemitic advocates as Henry Ford (Wistrich 1991). Finally, the Immigration Law of 1924 legitimized the racist atmosphere, ever more noticeable during the 1930s. In the post-World War II years, Jews were sometimes associated with the danger of Communism, but soon the favorable economic circumstances of later years helped reduce tension and American Jews were able to improve their status considerably. In Western Europe too, openly neo-Nazi and neo-Fascist parties proved of little political

consequence and the small Jewish communities enjoyed relative security and prosperity. Lately, however, anti-immigration and xenophobic sentiments seem to have given rise to parties with a measure of wider mass appeal, propagating, though often only implicitly and always among other things, an Antisemitic message. Despite the fact that nowadays immigrants are only rarely Jews, and that in most countries they constitute only a very small minority, hostility towards them accompanies old fears of foreigners and the new panic in the face of rapid change. Antisemitism among Blacks in America, for instance, seemed for a while to be a real menace, while in Europe, the declining power of the nation–state, the new world of communication, and, above all, the specter of globalization produce sporadic manifestations of Antisemitism, too. Ultra radical, terrorist groups—all too often agents of Antisemitism—plague many countries. Side by side with the presumably intellectual brand of Holocaust denial, occasionally infiltrating even respectable university campuses, pseudo-Nazi groupings insist upon reviving Antisemitism in its crudest forms. A flood of Internet sites appeal to a new kinds of young audiences. Jews are once again symbolic for everything they hate and fear. Even in places such as Japan, where there is no Jewish minority to speak of, the presumed Jewish power and evil influence is a source of concern for some.

In most of the democratic countries, however, both right-wing parties and racist activists of the more militant type face a political system, determined to limit their activities. In countries where such countervailing forces are weak, outbursts of Antisemitism cannot always be controlled, but elsewhere, despite minor incidents—though sometimes numerous and occasionally violent—it does not present a real danger. Antisemitism continues to exist, and in view of past experience must be regarded as a serious potential threat, but clearly, the spreading of democratic education and the strengthening of democratic institutions have proven capable of curbing its propaganda and checking its power and influence.

See also: Authoritarian Personality: History of the Concept; Ethnic Cleansing, History of; Ethnic Groups/Ethnicity: Historical Aspects; Ethnocentrism; Historiography and Historical Thought: Christian Tradition; Holocaust, The; Judaism; Judaism and Gender; National Socialism and Fascism; Parties/Movements: Extreme Right; Prejudice in Society; Totalitarianism: Impact on Social Thought; Western European Studies: Geography

Bibliography

Adorno T W 1951 *The Authoritarian Personality*, 1st edn. Harper, New York

Allport G W 1958 *The Nature of Prejudice*. Doubleday, Garden City, New York

Arendt H 1951 *The Origins of Totalitarianism*, 1st edn. Harcourt Brace, New York

Baron S W 1964 *The Russian Jew under Tzars and Soviets*. Macmillan, New York

Berger D (ed.) 1986 *History and Hate: The Dimensions of Anti-Semitism*, 1st edn. Jewish Publication Society, Philadelphia, PA

Cohen J 1999 *Living Letters of the Law. Ideas of the Jews in Medieval Christianity*. University of California Press, Berkeley/Los Angeles/London

Cohn N 1967 *Warrant for Genocide: The Myth of the Jewish World Conspiracy and the Protocols of the Elders of Zion*, 1st US edn. Harper, New York

Dinnerstein L 1994 *Antisemitism in America*. Oxford University Press, New York

Ettinger S 1978 *Modern Antisemitism. Studies and Essays*. Moreshet, Tel Aviv, Israel [Hebrew]

Friedländer S 1997 *Nazi Germany and the Jews. Vol 1: The Years of Persecution, 1933–1939*. Harper Collins, New York

Funkenstein A 1993 Changes in anti-Jewish polemics in the twelfth century. In: Funkenstein A (ed.) *Perceptions of Jewish History*. University of California Press, Berkeley/Los Angeles/Oxford

Gager J 1983 *The Origins of Anti-semitism: Attitudes towards Judaism in Pagan and Christian Antiquity*. Oxford University Press, New York

Goldhagen D J 1966 *Hitler's Willing Executioners. Ordinary Germans and the Holocaust*, 1st edn. Knopf, New York

Hilberg R 1992 *Perpetrators, Victims, Bystanders: The Jewish Catastrophe 1933–1945*, 1st edn. Aaron Asher Books, New York

Jochmann W 1976 Struktur und Funktion der deutschen Antisemitismus. In: Mosse W (ed.) *Juden in Wilhelminischen Deutschland 1890–1914*. Mohr, Tübingen, Germany

Katz J 1980 *From Prejudice to Destruction: Anti-Semitism 1700–1933*. Harvard University Press, Cambridge, MA

Kulka O D 1985 Major trends and tendencies in German historiography on National Socialism and the Jewish question (1924–1984). *Leo Baeck Institute Yearbook* **30**: 215–42

Langmuir G I 1990 *Towards a Definition of Antisemitism*. University of California Press, Berkeley/Los Angeles/Oxford

Lewis B 1986 *Semites and Anti-Semites. An Inquiry into Conflict and Prejudice*. Weidenfeld and Nicolson, London

Loewenstein R M 1951 *Christians and Jews. A Psychoanalytical Study*. International Universities Press, New York

Oberman H A 1984 *The Roots of Anti-semitism in the Age of Renaissance and Reformation*. Fortress Press, Philadelphia, PA

Pulzer P G J 1964 *The Rise of Political Anti-Semitism in Germany and Austria*. Wiley, New York

Rohrbacher S 1993 *Gewalt im Biedermeier. Antijüdische Ausschreitungen in Vormärz und Revolution (1815–1848/49)*. Campus, Frankfurt/New York

Rürup R 1975 *Emanzipation und Antisemitismus*. Vandenhoeck & Ruprecht, Göttingen, Germany

Sartre J P 1948 *Anti-Semite and Jew*. Schoken, New York

Schäfer P 1997 *Judeophobia. Attitudes towards Jews in the Ancient World*. Harvard University Press, Cambridge, MA

Schleunes K A 1970 *The Twisted Road to Auschwitz*. University of Illinois Press, Urbana, IL

Simon M 1986 *Verus Israel. A study of the Relations between Christians and Jews in the Roman Empire*. Oxford University Press, Oxford, UK, pp. 135–425

Tal U 1975 *Christians and Jews in Germany: Religion, Ideology and Politics in the Second Reich, 1870–1914*. Cornell University Press, Ithaca/London
Toury J 1968 *Turmoil and Confusion in the Revolution of 1848*. Moreshet, Tel Aviv, Israel
Volkov S 1990 *Jüdisches Leben und Antisemitismus im 19. und 20. Jahrhundert*. Verlag C. H. Beck, Munich, Germany
Wistrich R A 1991 *Antisemitism. The Longest Hatred*. Thames Metheuen, London
Yavetz Z 1997 *Judenfeindschaft in der Antike*. C H Beck, Munich, Germany
Yuval I J 1993 Vengeance and damnation, blood, and defamation: From Jewish martyrdom to blood libel accusations. *Zion* **58**: 33–90
Zimmermann M 1986 *Wilhelm Marr: The Patriarch of Antisemitism*. Oxford University Press, New York

S. Volkov

Antisocial Behavior in Childhood and Adolescence

Antisocial behavior is a broad construct which encompasses not only delinquency and crime that imply conviction or a possible prosecution, but also disruptive behavior of children, such as aggression, below the age of criminal responsibility (Rutter et al. 1998). The age of criminal responsibility varies from 7 years of age in Ireland and Switzerland to 18 years of age in Belgium, Romania, and Peru. In the United States, several states do not have a specific age. Legal, clinical, and developmental definitions of antisocial behavior have different foci.

1. *Definitions of Antisocial and Aggressive Behavior*

Legal definitions of criminal offences committed by *young people* cover: (a) noncriminal but risky behavior (e.g., truancy) which is beyond the control of authorities; (b) status offences where the age at which an act was committed determines whether it is considered damaging (e.g., gambling); (c) crimes to protect the offender from being affected (e.g., possession of drugs); and (d) crimes with a victim (e.g., robbery) broadly defined (Rutter et al. 1998). The most common crimes among young people are thefts.

Only some forms of delinquency involve aggression which is a narrower construct than antisocial behavior. A meta-analysis of factor analytic studies of antisocial behavior (Frick et al. 1993) revealed four major categories of antisocial behavior defined by two dimensions (overt to covert behavior, and destructive to less destructive) as follows: (a) aggression, such as assault and cruelty (destructive and overt); (b) property violations, such as stealing and vandalism (destructive and covert); (c) oppositional behavior, such as angry and stubborn (nondestructive and overt); and (d) status violations, such as substance use and truancy (nondestructive and covert). Aggression and violence are related but not synonymous concepts. Violence usually refers to physical aggression in its extreme forms.

Clinical definitions of antisocial behavior are focussed on psychopathological patterns in individuals. Oppositional defiant disorder, which includes temper tantrums and irritable behavior, becomes clinically less problematic by age eight, but some children, more often boys than girls, are unable to outgrow these problems.

Conduct disorder is diagnosed on the basis of a persistent pattern of behavior which violates the rights of others or age-appropriate societal norms. To individuals who must be at least 18 years of age, a third diagnosis, antisocial personality disorder, can be applied. These psychopathological patterns may involve delinquent behavior, but the criteria of their diagnosis are broader in terms of psychological dysfunction.

Developmental approaches to antisocial behavior are focussed on its developmental antecedents, such as hyperactive and aggressive behavior in childhood, and maladjustment to school in early adolescence. The younger the children are, the more their 'antisocial' behavior extends beyond acts that break the law. Different delinquency-related acts may be indicators of the same underlying construct such as low self-control, or they may indicate developmental sequences across different but correlated constructs.

Development of antisocial behavior is studied using a longitudinal design which means repeated investigations of the same individuals over a longer period of time. The increasing number of longitudinal studies indicates a high continuity of behavior problems from childhood to adulthood. There is continuity between disobedience and defiance of adults, aggression towards peers, and hyperactivity at age three, and similar or more serious behavior problems in later childhood. Hyperactivity during the preschool years associated with aggressive behavior has the most robust links to later antisocial behavior.

Common definitions of *aggression* emphasize an intent to harm another person (Coie and Dodge 1998). References to the emotional component of aggression are not typically made in these definitions. *Anger*, the emotional component of aggression, and *hostility*, a negative attitude, motivate a person for aggressive acts, but aggressive behavior may also be displayed instrumentally. Hostile aggressive responding is characterized by intense autonomic arousal and strong responses to perceived threat. In contrast, instrumental aggression is characterized by little autonomic activation and an orientation toward what the ag-

gressor sees as a reward or expected outcome of behavior.

Each aggressive act has a mode of expression, direction, and motive. An aggressive act may be expressed physically, verbally, or non-verbally, and targeted, in each case, more directly or indirectly. It also varies in its harmfulness or intensity. The motive of the aggressive act may be defensive (reactive) or offensive (proactive). Among school children, proa ctive aggression is often displayed in bullying behavior, which means purposefully harmful actions repeatedly targeted at one and the same individual. From four percent to 12 percent of children—boys slightly more often than girls—can be designated as bullies and as many as victims depending on the method of identification, age of children, and culture. Both bullying others and being victimized tend to endure from one year to another, and they are related to relatively stable personality patterns. Besides Bullies and Victims, the participant roles include Assistants who are more or less passive followers of the bully, Reinforcers who provide Bullies with positive feedback, Defenders who take sides with the victim, and Outsiders who tend to withdraw from bullying situations (Salmivalli 1998).

Self-defense and defense of others are often culturally accepted, and many children limit their aggressive behavior to defensive aggression. Longitudinal findings show that 'defense-limited' aggression in early adolescence predicts more successful social adjustment in adulthood than 'multiple' aggression, which also includes proactive aggression. Only multiple aggression predicts criminal offences at a later age (Pulkkinen 1996). The distinction between hostile and instrumental aggression is not parallel to defense-limited and multiple aggression or to reactive and proactive aggression. Although proactive aggression often is instrumental, reactive (or defensive) aggression may be either instrumental or hostile.

2. *Development of Aggression and Antisocial Behavior*

2.1 *Development of Aggression*

Anger expression cannot be differentiated from other negative affects in newborns, but by four months of age angry facial displays—the eyebrows lower and draw together, the eyelids narrow and squint, and the cheeks elevated—are present and they are directed to the source of frustration (Stenberg and Campos 1990). The most frequent elicitors of aggression in infancy are physical discomfort and the need for attention. Peer-directed aggression, seen in responding to peer provocations with protest and aggressive retaliation, can be found at the end of the first year of life. At this age children become increasingly interested in their own possessions and control over their own activities.

During the second year of life, oppositional behavior and physical aggression increase. Most children learn to inhibit physical aggression during preschool age, but other children continue displaying it (Tremblay et al. 1999). Verbal aggression sharply increases between two and four years of age and then stabilizes. It is a time of fast language development which helps children to communicate their needs symbolically. Delays in language development are often related to aggressive behavior problems.

Between six to nine years of age, the rate of aggression declines, but at the same time its form and function change from the relatively instrumental nature of aggression in the preschool period to increasingly person-oriented and hostile (Coie and Dodge 1998). Children become aware of hostile intents of other people and they, particularly aggressive children, perceive threats and derogations to their ego and self-esteem which elicit aggression. Most longitudinal studies show a decrease in the ratings of aggression, that is, in the perceived frequency of aggressive acts as children enter adolescence. Nevertheless, serious acts of violence increase. Individual differences in aggressive behavior become increasingly pronounced.

2.2 *Individual Differences in Aggression and Antisocial Behavior*

Individual differences in anger expression emerge early in life. At the age of two years, consistency of anger responses across time is already significant. Individual differences in aggression remain rather stable during childhood and adolescence. Correlations vary slightly depending on the measures used, the length of interval, and the age of children, but they are generally between 0.40 and 0.70. The stability is comparable for males and females. Individual differences in responding to a conflict lie both in the frequency of aggressive behavior and in prosocial attempts to solve conflicts. The latter are facilitated by language development. Language may, however, provide children with verbal means of aggression. Additional factors, such as the development of self-regulation, perspective taking, empathy, and social skills, are needed for the explanation of individual differences in aggression (Coie and Dodge 1998).

Gender differences in aggression appear in preschool age, boys engaging in more forceful acts both physically and verbally. This sex difference widens in middle childhood and peaks at age 11 when gender differences in aggressive strategies emerge: girls display relational aggression (e.g., attempts to exclude peers from group participation) more than boys, and boys engage in fighting more than girls (Lagerspetz and Björkqvist 1994). Both fighting and relational aggression may aim at structuring one's social status in a peer group, but by different means.

Antisocial and other externalizing behavior is more common, and the offending career is longer among males than among females. There have been, however, changes in the ratio between male and female offenders during the 1980s and 90s in several western countries. Adolescent girls are increasingly involved in antisocial behavior. In the United Kingdom, the sex ratio was about 10:1 in the 1950s, and 4:1 in the 1990s. The peak age of offending among girls has remained at around age 14 or 15, but the peak age for male offenders has risen in thirty years from 14 to 18 (Rutter et al. 1998). The peak age of registered offences is related to police and prosecution policy, and varies by offense. For instance, peak age is later for violent crimes than for thefts.

2.3 Continuity in Antisocial Behavior

A multiproblem pattern is a stronger predictor of delinquency than a single problem behavior. For instance, aggression in childhood and adolescence predicts delinquency when associated with other problem behaviors, such as hyperactivity, lack of concentration, and low school motivation and achievement, and poor peer relations (Stattin and Magnusson 1995). Peer rejection in preadolescence, which indicates social incompetence rather than social isolation, predicts delinquency even independently of the level of aggression. Continuity from early behavioral problems to delinquency and other externalizing behavior is higher among males than among females, whereas girls' behavioral problems predict internalizing behavior (depression and anxiety) more often than boys' behavioral problems (Zoccolillo 1993).

Several studies show that a small group of chronic offenders accounts for half of the offences of the whole group. They tend to display the pattern of antisocial behavior called 'life-course-persistent.' It is characterized at an early age by lack of self-control, reflecting an inability to modulate impulsive expression, difficult temperament features, hyperactivity, attentional problems, emotional lability, behavioral impulsivity, aggressiveness, cognitive, language and motor deficits, reading difficulties, lower IQ, and deficits in neuropsychological functioning. Thirteen percent of the boys in the study by Moffitt et al. (1996) met criteria for early onset, but only half of them persisted into adolescence. Therefore, several assessments are needed for the identification of life-course-persistent offenders.

An adolescence-limited pattern of offending is more common than the life-course-persistent pattern. It reflects the increasing prevalence of delinquent activities during adolescence. Both overt (starting from bullying) and covert (starting from shoplifting) pathways toward serious juvenile offending have been discerned (Loeber et al. 1998). Compared to the life-course-persistent pattern, the adolescence-limited pattern is less strongly associated with difficult temperament, hyperactivity and other early behavioral problems, neuropsychological deficits, and poor peer relationships. Many problem behaviors are very common in adolescence. Self-reports show that half of males and from 20 percent to 35 percent of females have been involved in delinquency. Rutter et al. (1998) conclude that antisocial behavior 'operates on a continuum as a dimensional feature that most people show to a greater or lesser degree' (p. 11).

3. Determinants of Antisocial Behavior

With the increase of empirical findings, theories of crime which try to explain crime with a single set of causal factors have been increasingly criticized. The climate has changed also in regard to the possible role of individual characteristics as determinants of antisocial behavior. In the 1970s, theories emphasized social causes of crime, and paid little attention to individual factors. The situation is now different. Empirical studies have revealed that determinants of antisocial behavior are diverse ranging from genetic to cultural factors.

Studies on genetic factors in antisocial behavior have shown that the estimates for the genetic component of hyperactivity are about 60 to 70 percent. Antisocial behavior linked to hyperactivity, which is generally associated with poor social functioning, is strongly genetically influenced. In contrast, antisocial behavior which is not associated with hyperactivity is largely environmental in origin (Silberg et al. 1996). The genetic component for violent crime is low compared to the heritability of aggression (about 50 percent), but this difference may also be due to differences in prevalence of these behaviors, and its effects on statistical analyses.

There is no gene for antisocial behavior; it is multifactorially determined. Genetic effects increase a liability for antisocial behavior, but they operate probabilistically, which means that the effects increase the likelihood of antisocial behavior, if environmental and experiential factors affect in the same direction. This conclusion also concerns the XYY chromosomal anomaly.

The importance of experiential factors, particularly early family socialization, in the development of aggression has been shown in many studies (Coie and Dodge 1998). Aggressive individuals generally hold positive views about aggression and believe it is normative. Child-rearing strategies are related to subsequent aggression in the child, for instance, insecure and disorganized attachment with the caregiver, parental coldness and permissiveness, inconsistent parenting, and power-assertive discipline (Hinde et al. 1993). Low monitoring is particularly important to adolescent involvement with antisocial behavior. Parenting affects children's behavior in interaction with their temperament resulting in differ-

ences in self-control that are related to adult outcomes, such as criminality (Pulkkinen 1998).

An adverse immediate environment which increases a risk for antisocial behavior in interaction with genetic factors includes parental criminality, family discord, ineffective parenting such as poor supervision, coercive parenting and harsh physical discipline, abuse, neglect, and rejection, delinquent peer groups, unsupervised after school activities, and youth unemployment. These risk factors also increase the use of alcohol and drugs, which is often related to crime, and are very similar in different countries, although there are also some differences (Farrington and Loeber 1999).

There are also several sociocultural factors which may serve to raise the level of crime in the community, such as income differentials, antisocial behavior in neighborhood, the availability of guns, media violence, the quality of school and its norms, unemployment rate, and involvement in a drug market. Poverty is strongly related to aggression and possibly operates through disruption of parenting. Violent virtual reality is available for children of the present generation via electronic games playing. TV programs and video films are passive in nature, whereas electronic games involve the player's active participation and often violent winning strategies (Anderson and Ford 1986). Some minority ethnic groups are overrepresented in crime statistics, but causal factors are complex. Cultural traditions cause, however, vast differences in crime rates between different countries. In Asian countries, particularly in Japan, crime rates are lower than in western countries.

4. Conclusions

Official statistics in developed countries show that crime rates among young people have been rising since the 1970s. The results of recent longitudinal studies have increased our understanding of factors contributing to the incidence of antisocial behavior. A greater understanding of how causal mechanisms operate is, however, needed for the development of effective means of preventing crime. Since persistent antisocial behavior starts from conduct problems in early childhood, support for families and work with parents and teachers to improve their management skills are extremely important. An affectionate parent, nonpunitive discipline, and consistent supervision are protective factors against antisocial behavior. Parent management training is a neglected area in western educational systems. Socialization of children and youth might also be supported by, for instance, legislation against gun availability and sociopolitical improvements of family conditions.

See also: Adolescent Development, Theories of; Adolescent Vulnerability and Psychological Interventions; Aggression in Adulthood, Psychology of; Behavior Therapy with Children; Children and the Law; Crime and Delinquency, Prevention of; Developmental Psychopathology: Child Psychology Aspects; Early Childhood: Socioemotional Risks; Personality and Crime; Personality Theory and Psychopathology; Poverty and Child Development; Social Competence: Childhood and Adolescence; Socialization in Adolescence; Socialization in Infancy and Childhood; Violence and Effects on Children

Bibliography

Anderson C A, Ford C M 1986 Affect of the game player. Short-term effects of highly and mildly aggressive video games. *Personality and Social Psychology Bulletin* **12**: 390–402

Coie J D, Dodge K A 1998 Aggression and antisocial behavior. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology. Vol. 3. Social, Emotional and Personality Development*, Wiley, New York, pp. 779–862

Farrington D P, Loeber R 1999 Transatlantic replicability of risk factors in the development of delinquency. In: Cohen I, Slomkowski C, Robins L N (eds.) *Historical and Geographical Influences on Psychopathology*, Erlbaum, Mahwah, NJ, pp. 299–329

Frick P J, Lahey B B, Loeber R, Tannenbaum L, Van Horn Y, Chirst M A G 1993 Oppositional defiant disorder and conduct disorder: a meta-analytic review of factor analyses and cross-validation in a clinic sample. *Clinical Psychology Review* **13**: 319–40

Hinde R A, Tamplin A, Barrett J 1993 Home correlates of aggression in preschool. *Aggressive Behavior* **19**: 85–105

Lagerspetz K M, Björkqvist K 1994 Indirect aggression in girls and boys. In: Huesmann L R (eds.) *Aggressive Behavior: Current Perspectives*, Plenum Press, New York, pp. 131–50

Loeber R, Farrington D P, Stouthamer-Loeber M, Moffitt T E, Caspi A 1998 The development of male offending: Key findings from the first decade of the Pittsburgh Youth Study. *Studies on Crime and Crime Prevention* **7**: 141–71

Moffitt T E, Caspi A, Dickson N, Silva P, Stanton W 1996 Childhood-onset versus adolescent-onset antisocial conduct problems in males: Natural history from ages to 18 years. *Development and Psychopathology* **9**: 399–424

Pulkkinen L 1996 Proactive and reactive aggression in early adolescence as precursors to anti- and prosocial behavior in young adults. *Aggressive Behavior* **22**: 241–57

Pulkkinen L 1998 Levels of longitudinal data differing in complexity and the study of continuity in personality characteristics. In: Cairns R B, Bergman L R, Kagan J (eds.) *Methods and Models for Studying the Individual* Sage, Thousand Oaks, CA, pp. 161–184

Rutter M, Giller H, Hagell A 1998 *Antisocial behavior by young people*. Cambridge University Press, New York

Salmivalli C 1998 *Not only bullies and victims. Participation in harassment in school classes: some social and personality factors*. Annales Universitatis Turkuensis, ser. B/225, University of Turku, Finland

Silberg J, Meyer J, Pickles A, Simonoff E, Eaves L, Hewitt J, Maes H, Rutter M 1996 Heterogeneity among juvenile antisocial behaviours: Findings from the Virginia Twin Study of Adolescent Behavioral Development. In: Bock G R, Goode J A (eds.) *Genetics of Criminal and Antisocial Behaviour* (Ciba

Foundation Symposium no. 194), Wiley, Chichester, UK, pp. 76–85
Stattin H, Magnusson D 1995 Onset of official delinquency: Its co-occurrence in time with educational, behavioural, and interpersonal problems. *British Journal of Criminology*, **35**: 417–49
Sternberg C R, Campos J J 1990 The development of anger expressions in infancy. In: Stein N L, Leventhal B, Trabasso T (eds.) *Psychological and Biological Approaches to Emotion*, Erlbaum, Hillsdale, NJ, pp. 247–82
Tremblay R E, Japel C, Pérusse D, McDuff P, Boivin M, Zoccolillo M, Montplaisir J 1999 The search for the age of 'onset' of physical aggression: Rousseau and Bandura revisited. *Criminal Behavior and Mental Health* **9**: 8–23
Zoccolillo M 1993 Gender and the development of conduct disorder. *Development and Psychopathology* **5**: 65–78

L. Pulkkinen

Antitrust Policy

The term antitrust, which grew out of the US trust-busting policies of the late nineteenth century, developed over the twentieth century to connote a broad array of policies that affect competition. Whether applied through US, European, or other national competition laws, antitrust has come to represent an important competition policy instrument that underlies many countries' public policies toward business. As a set of instruments whose goal is to make markets operate more competitively, antitrust often comes into direct conflict with regulatory policies, including forms of price and output controls, anti-dumping laws, access limitations, and protectionist industrial policies.

Because its primary normative goal has been seen by most to be economic efficiency, it should not be surprising that antitrust analysis relies heavily on the economics of industrial organization. But, other social sciences also contribute significantly to our understanding of antitrust. Analyses of the development of antitrust policy are in part historical in nature, and positive studies of the evolution of antitrust law (including analyses of lobbying and bureaucracy) often rely heavily on rational choice models of the politics of antitrust enforcement.

The relevance of other disciplines notwithstanding, there is widespread agreement about many of the important antitrust tradeoffs. Indeed, courts in the US have widely adopted economic analysis as the theoretical foundation for evaluating antitrust concerns. Interestingly, however, antitrust statutes in the European Union also place heavy emphasis on the role of economics. Indeed, a hypothetical conversation with a lawyer or economist at a US competition authority (the Antitrust Division of the Department of Justice or the Federal Trade Commission) or the European Union (The Competition Directorate) would be indistinguishable at first sight.

While this article provides a view of antitrust primarily from the perspective of US policy, the review that follows illustrates a theme that has worldwide applicability. As our understanding of antitrust economics has grown throughout the past century, antitrust enforcement policies have also improved, albeit sometimes with a significant lag. In this survey the following are highlighted: (a) the early anti-big business period in the US, in which the structure of industry was paramount; (b) the period in which performance as well as structure was given significant weight, and there was a systematic attempt to balance the efficiency gains from concentration with the inefficiencies associated with possible anti-competitive behavior; (c) the most recent period, which includes the growth of high technology and network industries, in which behavior theories have been given particular emphasis.

1. *The Antitrust Laws of the US*

In the USA, as in most other countries, antitrust policies are codified in law and enforced by the judicial branch. Public cases may be brought under Federal law by the Antitrust Division of the Department of Justice, by the Federal Trade Commission, and/or by each of the 50-state attorneys-general. (The state attorneys-general may also bring cases under state law.) Further, there is a broad range of possibilities for private enforcement of the antitrust laws, which plays a particularly significant role in the USA.

1.1 *The Sherman Act*

Antitrust first became effective in the US near the end of the nineteenth century. Underlying the antitrust movement was the significant consolidation of industry that followed the Civil War. Following the war, large trusts emerged in industries such as railroads, petroleum, sugar, steel, and cotton. Concerns about the growth and abusive conduct of these combinations generated support for legislation that would restrict their power. The first antitrust law in the USA—the Sherman Act—was promulgated in 1890. Section 1 of the Act prohibits: 'Every contract, combination in the form of trust of otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations.'

Section 2 of the Sherman Act states that it is illegal for any person to '…monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce among the several States, or with foreign nations….' These two sections of the Act contain the

two central key principles of modern antitrust policy throughout the world—conduct that restrains trade and conduct that creates or maintains a monopoly—is deemed to be anticompetitive.

1.2 The Clayton Act

Early in the twentieth century it became apparent that the Sherman Act did not adequately address combinations, such as mergers, that were likely to create unacceptably high levels of market power. In 1914 Congress passed the Clayton Act, which identified specific types of conduct that were believed to threaten competition. The Clayton Act also made illegal conduct whose effect 'may be to substantially lessen competition or tend to create a monopoly in any line of commerce.' Section 7 of the Clayton Act is the principal statute for governing merger activity—in principle Sect. 7 asks whether the increased concentration will harm actual and/or potential competition.

Other sections of the Clayton Act address particular types of conduct. Section 2, which was amended and replaced by Sect. 1 of the Robinson-Patman Act in 1936, prohibits price discrimination between different purchasers of the same type and quality of a commodity, except when such price differences are cost-justified, or if the lower price is necessary to meet competition.

Section 3 of the Clayton Act specifically prohibits certain agreements in which a product is sold only under the condition that the purchaser will not deal in the goods of a competitor. This section has been used to challenge exclusive dealing arrangements (e.g., a distributor that is obligated to sell only the products of a particular manufacturer) and 'tied' sales (e.g., the sale of one product is conditioned on the buyer's purchase of another product from the same supplier). The Act does not prohibit all such arrangements—only those whose effect would be likely to substantially lessen competition in a particular line of commerce.

1.3 The Federal Trade Commission Act

The US is nearly unique among competition-law countries in having two enforcement agencies. In part to counter the power granted to the Executive branch under the Sherman Act, Congress created the Federal Trade Commission (FTC) in 1914. Section 5 of the Federal Trade Commission Act, which enables the FTC to challenge 'unfair' competition, can be applied to consumer protection as well as mergers. In addition, the FTC has the power to enforce the Clayton Act and the Robinson-Patman Act. While the agencies act independently of one another, the FTC and the DOJ's enforcement activities have generally been consistent with one another during most of enforcement history.

What then are the differences between the two enforcement agencies? A simple answer is that the FTC is responsible for consumer protection issues, whereas criminal violations of Sect. 1 of the Sherman Act (e.g., price fixing and market division) are the responsibility of the Antitrust Division. It is also important to note that most enforcement activities, when successful, lead to injunctive remedies, where the party that has violated the law is required to cease the harmful activity. Exceptions are the criminal fines that are assessed when Sect. 1 of the Sherman Act is criminally violated, and fines that are assessed in certain consumer protection cases. Damages are rarely assessed by the federal agencies, otherwise, although in principle there can be exceptions (e.g., when the US represents the class of government employees).

2. The Goals of Antitrust

Despite considerable economic change in the economy over the twentieth century, the federal antitrust laws have continued to have wide applicability in part because their language was quite vague and flexible. This has led, quite naturally, to extensive historical study of the legislative history and substantial debate about congressional intent, especially where the Sherman Act is concerned. Some scholars have argued that the Sherman Act was directed almost entirely towards the achievement of allocative efficiency (e.g., Bork 1966). Others have taken the view that Congress and the courts have expressed other values, ranging from a broad concern for fairness, to the protection of specific interest groups, or more simply to the welfare of consumers writ large (e.g., Schwartz 1979, Stigler 1985).

The debate concerning the goals of the antitrust laws continues today. The enforcement agencies, for example, evaluate mergers from both consumer welfare and total welfare (consumer plus business) points of view, in part because the courts are not clear as to which is the appropriate standard under the Clayton Act. While it seems clear that the Sherman Act was intended in part to protect consumers against the inefficiencies of monopolies and cartels, it is significant that at the time of the passage of the Act many economists were in opposition because they believed that large business entities would be more efficient. Whatever the goals of the Sherman Act, it is notable that a merger wave (1895–1905) soon followed the passage of the Act. It may be that outlawing various types of coordinated behavior (Sect. 1) may have encouraged legal coordination through merger.

3. Historical Developments

Because antitrust focuses on the protection of competitive markets, it was natural to suspect other nonstandard organizational forms as potentially anticompetitive. During the early part of the 1900s, most antitrust enforcement was public, and it was directed against cartels and trusts. Early successes included

convictions against the Standard Oil Company (*Standard Oil vs. US*) and the tobacco trust for monopolization in 1911.

It is important to note, however, that the Sherman Act does not prohibit all restraints of trade—only those restraints that are unreasonable. The distinction between reasonable and unreasonable restraints remains a subject of debate today. It is significant, however, that the distinction between *per se* analysis (in which a practice is deemed illegal on its face) and rule of reason (in which one trades off the procompetitive and anticompetitive aspects of a practice) was made during this early period, and it remains important today.

By its nature, a *per se* rule creates a rebuttal presumption of illegality once an appropriate set of facts is found. *Per se* rules have the advantage that they provide clear signals and involve minimal enforcement costs. Yet actual firm behavior in varying market contexts generates exceptions that call for in-depth analyses. It is not surprising, therefore, that the *per se* rule is the exception to the rule-of-reason norm.

An example of the application of rule of reason is price discrimination. Viewed as an exercise of monopoly power, the practice was seen as suspect. Indeed, the Robinson-Patman Act presumes that with barriers to entry, price discrimination marks a deviation from the competitive ideal that was presumed to have monopoly purpose and effect. Efficiency justifications for price discrimination and other 'restrictive practices'—promoting investment, better allocating scarce resources, and economizing on transaction costs—were overlooked during the early enforcement period. Confusion also arose between the goal of protecting competition and the practice of protecting competitors. At the beginning of the twenty-first century, however, efficiency arguments are clearly pertinent to the evaluation of Robinson-Patman claims.

There is no doubt that antitrust enforcement is difficult, even with the more sophisticated tools of industrial organization that are available today. The Sherman Act does not offer precise guidance to the courts in identifying illegal conduct. In both the Clayton Act and the Sherman Act, Congress has chosen not to enumerate the particular types of conduct that would violate the antitrust laws. Instead, Congress chose to state general principles, such as attempts to monopolize, and contracts or combinations that restrain trade, without elaborating on what actually qualifies for the illegal behavior. It was left to the courts to ascertain the intent of the antitrust statutes and to distinguish conduct that harms competition from conduct that does not.

In a significant early case, *Board of Trade of the City of Chicago vs. USA* (1918), the Supreme Court reiterated the reasonableness standard for evaluating restraints of' trade. The Court concluded that 'The true test of legality is whether the restraint imposed is such as merely regulates and perhaps thereby promotes competition or whether it is such as may suppress or even destroy competition.' In this and in subsequent cases, the court failed to provide a clear statement of when and how a rule of reason analysis should be applied.

Antitrust enforcement agencies and the courts continue to debate the mode of analysis that is most appropriate in particular market contexts and when particular practices are at issue. Both courts and agencies deem it appropriate to undertake some form of 'quick-look analysis,' one which goes beyond the application of a per se rule, but which falls short of a full rule of reason inquiry (see *California Dental Association vs. FTC* 1999, Melamed 1998). *Per se* rules are applied to certain horizontal restraints, involving two or more firms operating in the same line of business. Quick-look and rule of reason analyses are more prevalent when the restraints at issue are vertical (involving two or more related lines of business, e.g., a manufacturer and a supplier).

Quick-look and more complete rule-of-reason analyses have relied on a number of basic principles. First and foremost is the market power screen. Courts have appreciated that antitrust injury necessitates the exercise of market power. Antitrust analysis typically begins with the measurement of market share possessed by firms alleged to engage in anticompetitive conduct. As many scholars have noted (e.g., Stigler 1964), in the absence of collusion, the exercise of market power in unconcentrated markets is unlikely. Market concentration is often seen as a necessary but not sufficient condition for the exercise of' market power, since ease of entry, and demand and supply substitutability can limit the ability of firms to raise prices even in highly concentrated markets (e.g., Baumol et al. 1982). But, market concentration —market power screens are not essential; witness the Supreme Court's opinion in *FTC vs. Indiana Federal of Dentists* (1986).

Although courts have long recognized the importance of market power to conclusions about antitrust injury, the standards by which they adjudicate antitrust cases, and their willingness to apply sophisticated economic analysis, has varied significantly over time. Antitrust policy once treated any deviation from the competitive ideal as having anticompetitive purpose and effect. Vertical arrangements were particularly suspicious if the parties agreed to restraints that limited reliance on market prices. Over time, however, the critique of such arrangements diminished as economists began to appreciate the importance of efficiencies associated with a range of contractual practices.

3.1 The Structure-conduct Interventionist Period

During the 1950s and the 1960s the normative analysis of antitrust was dominated by the work of Joe Bain on

barriers to entry, industry structure, and oligopoly (Bain 1968). Bain's relatively interventionist philosophy was based on the view that scale economies were not important in many markets, that barriers to entry are often high and can be manipulated by dominant incumbent firms, and that supracompetitive monopolistic pricing is relatively prevalent. This tradition held, not surprisingly, that nonstandard and unfamiliar practices should be approached with skepticism. During this era, antitrust enforcers often concluded that restraints which limited the number of competitors in a market necessarily raised prices, and that the courts could protect anticompetitive conduct by appropriate rule making.

In this early structure-conduct era, industrial organization economists tended to see firms as shaped by their technology. Practices, such as joint ventures, that reshaped the boundaries of the firm (e.g., joint ventures) were often seen as suspect. Significant emphasis was placed on the presence or absence of barriers to entry, which provided the impetus for a very tight market power screen.

Because the government was often seen as benign, it was not surprising that antitrust looked critically at mergers and acquisitions. Moreover, absent a broader theory that encompassed a variety of business relationships among firms, it is not surprising that the Supreme Court often supported government arguments without seriously evaluating the tradeoffs involved (e.g., *USA vs. Von's Grocery Co.* 1966).

Economies of scale are illustrative. That they can create a barrier to entry was emphasized under Bain's point of view, whereas the clear benefits that scale economies provide were given little recognition. For example, the Supreme Court argued that 'possible economies cannot be used as a defense to illegality' (*Federal Trade Commission vs. Procter & Gamble Co.* 1967).

Government hostility during this period also applied to markets with differentiated products. In *USA vs. Arnold, Schwinn & Co.* (1907), for example, the government was critical of franchise restrictions that supported product differentiation because the restrictions were perceived to foster the purchase of inferior products at higher than competitive prices. The possibility that exclusive dealing was procompetitive was not given serious consideration during the 1950s and 1960s.

The failure of antitrust enforcers to appreciate the benefits of discriminatory contractual practices was also evident in the early enforcement of the Robinson-Patman Act. Through the 1960s, the Act was vigorously enforced. Legitimate reasons for discriminatory pricing were very narrow (e.g., economies of scale in production or distribution). Since the mid-1960s, however, there has been a significant shift. Fears that the Act might discourage procompetitive price discrimination have led to less aggressive enforcement by the FTC.

The 1960s and 1970s marked a period of substantial empirical analysis in antitrust, motivated by structural considerations. The literature on the correlation between profit rates and industry structure initially showed a weak positive correlation, suggesting that high concentration was likely to be the source of anticompetitive firm behavior. However, this interpretation was hotly disputed. If behavior that is described as a barrier to entry also served legitimate purposes, what can one conclude even if there is a positive correlation between profitability and concentration? The challenge to this line of empirical work is typified by the debate between Demsetz (1974) and Weiss (1974). Demsetz argued that concentration was a consequence of economies of scale and the growth of more efficient firms—in effect the early empirical work suffered from problems of simultaneity. If concentrated markets led to higher industry profits, these profits were the consequence and not the cause of the superior efficiency of large firms and they consequently are consistent with competitive behavior. Today, these early studies are viewed critically, as having omitted variables that account for research and development, advertising, and economies of scale. It is noteworthy that the observed positive correlation reflected both market power and efficiency, with the balance varying on a case-by-case basis.

3.2 *The Influence of the Chicago School*

All of the assumptions that underlie the tradition of the 1950s and 1960s came under increasing criticism during the 1970s, led in part by the influence of the Chicago School. While the views of its proponents (e.g., Posner 1979, Easterbrook 1984) are themselves both rich and varied, they have come to be typified as including the following: (a) A belief that the allocative efficiencies associated with economies of scale and scope are of paramount importance; (b) A belief that most markets are competitive, even if they contain a relatively few number of firms. Accordingly, even if price competition is reduced, other nonprice forms of competition will fill the gap; (c) A view that monopolies will not last forever. Accordingly, the high profits earned by dominant monopolistic firms will attract new entry that in most cases will replace the monopolist or at least erode its position of dominance; (d) A view that most barriers to entry, except perhaps those created by government, are not nearly as significant as once thought; (e) A belief that monopolistic firms have no incentive to facilitate or leverage their monopoly power in vertically related markets (the 'single monopoly rent' theory); (f) A view that most business organizations maximize profits; firms that do not will not survive over time; (g) A belief that even when markets generate anticompetitive outcomes, government intervention, which is itself less

than perfect, is appropriate only when it improves economic efficiency.

The period of the late 1960s and 1970s not only marked a significant influence by the Chicago School; it also saw an upgrading of the role of economists in the antitrust enforcement agencies. The increased role continued through the end of the twentieth century in the US, as it did in the European Union. For example, two economists play decision-making roles at the Department of Justice—one a Deputy for Economics and the other a Director in charge of Economic Enforcement.

Because of the influence of the Chicago School, nonstandard contractual practices that were once denounced as anticompetitive and without a valid purpose were often seen as serving legitimate economic purposes. Not surprisingly, for example, the *per se* rule limiting exclusive distribution arrangements in *Schwinn* was struck down by the US Supreme Court in *GTE Sylvania, Inc. vs. Continental TV, Inc.* (1977). Although the US DOJ–FTC Merger Guidelines does not explicitly spell out a tradeoff analysis of efficiencies and competitive effects, accounting for the economic benefits of a merger is now standard practice. Antitrust enforcement is alert to tradeoffs, and less ready to condemn conditions for which there is no obvious superior alternative.

3.3 The Post-Chicago School Analyses of Strategic Economic Behavior

In the 1970s and continuing through the 1990s new industrial organization tools, especially those using game-theoretic reasoning came into prominence. These tools allow economists to examine the ways in which established firms behave strategically in relation to their actual and potential rivals. The distinction between credible and noncredible threats, which was absent from the early entry barrier literature, has been important to an assessment of the ability of established firms to exclude competitors and to the implications of exclusionary conduct for economic welfare (e.g., Dixit 1979). These theories also illuminated a broader scope for predatory pricing and predatory behavior. Previously scholars such as Robert Bork (1978) had argued that predatory pricing imposes high costs on the alleged predator and is unlikely to be profitable in all but extraordinary situations. Developments in the analysis of strategic behavior provide a richer perspective on the scope for such conduct. Thus, nonprice competitive strategies that 'raise rivals' costs' (Krattenmaker and Salop 1986, Ordover et al. 1990) are now thought to be quite prevalent. Indeed, models of dynamic strategic behavior highlight the ability and opportunity for firms to engage in coordinated actions and for firms to profit from conduct that excludes equally efficient rivals (e.g., Kreps and Wilson 1982, Milgrom and Roberts 1982).

The implications of the various models of strategic behavior remain hotly debated. On the more skeptical side is Franklin Fisher, who stresses that while these new developments offer insights, they are insufficiently complete to provide firm conclusions or to allow us to measure what will happen in particular cases (Fisher 1989). A more optimistic view is given by Shapiro (1989) who believes that we can now analyze a much broader range of business competitive strategies than before. We also know much more about what to look for when we are studying areas such as investment in physical and intangible assets, the strategic control of information, network competition and standardization, and the competitive effects of mergers.

The analysis of strategic behavior has emphasized the potential for exercising market power, often to the detriment of consumers. At the same time, other analyses have stressed that there are gains to consumers from coordinated behavior among firms with market power. Thus, Coase (1937) argued that market forces were only one means to organize market activities, and that nonmarket organizations could provide a viable alternative. Further, Williamson (1985) developed the theory of nonmarket organization to show that contractual restraints can provide improved incentives for investments in human and physical assets that enhance the gains from trade. This transaction cost approach helps to provide a foundation for understanding the efficiency benefits of contractual restraints in vertical relationships.

Coupled with the game-theoretic analyses of firm behavior in imperfect markets has been the application of modern empirical methods for the analysis of firm practices. These newer methods allow for the more precise measurement of market power, and they indicate generally that market power is relatively common, even in markets without dominant firms (Baker and Rubinfeld 1999).

3.4 The Public Choice Approach

Today, a balanced normative approach to antitrust would involve reflections on the broad set of efficiencies associated with various organizational forms and contractual relations, as well as the possibilities for anticompetitive strategic behavior in markets with or without dominant firms. A positive (descriptive) approach focuses on the relationship between market structure and the politics of the interest groups that are affected by that structure.

The theory of public choice has been used as the basis for limiting competitor standing to bring antitrust suits. In support is the argument that competitors' interests deviate from the public interest because competitors view efficiencies associated with alleged exclusionary practices to be harmful, and therefore are not in a position to distinguish efficient from inefficient practices. On the other hand,

competitors are often the most knowledgeable advocates, and are likely to know more about the socially harmful effects of exclusionary practices long before consumers do. The debate over standing to sue has surfaced with respect to indirect purchaser suits (e.g., *Illinois Brick Co. vs. Illinois* 1977) and with respect to predatory pricing (*Brooke Group Ltd. vs. Brown & Williamson Corp.* 1993).

4. *High Technology and Dynamic Network Industries*

In the latter part of the twentieth century rapid changes in technology have altered the nature of competition in many markets in ways that would most likely have surprised the antitrust reformers of the late nineteenth century. The early debates centered on scale—did the benefits of economies of scale in production outweigh the associated increase in market power? In many dynamic high technology industries today, demand, rather than supply is often the source of substantial consumer benefits and significant market power. Economies of scale on the demand side arise in industries such as computers and telecommunications because of the presence of network effects, whereby each individual's demand for a product is positively related to the usage of the product (and complementary products) by other individuals. Network effects apply to communications networks (where consumers value a large network of users with whom to communicate, such as compatible telephone systems and compatible fax machines), and they apply to virtual networks or hardware–software networks (where there is not necessarily any communication between users).

In industries in which network effects are significant, a host of issues challenge our traditional views of antitrust. Because network industries are often characterized by large sunk costs and very low marginal costs, there is a substantial likelihood that a successful firm will come to dominate a market and to persist in that dominance for a significant period of time. Indeed, while there is no assurance that a single standard will arise in network industries, it is nevertheless often the case that users will gravitate toward using compatible products. This combination of economic factors makes it possible for firms to adopt price and nonprice policies that exclude competition and effectively raise prices significantly above what they would be were there more competition in the market (Rubinfeld 1998).

A host of hotly debated antitrust policy issues are raised by the increasing importance of dynamic network industries. Whatever view one holds, there is little doubt that the antitrust enforcement stakes are raised. On one hand, because the path of innovation today will significantly affect future product quality and price, the potential benefits of enforcement are huge. This perspective clearly motivated the Department of Justice and twenty states when they chose to sue Microsoft for a variety of antitrust violations in 1998 (*US vs. Microsoft* 1998). On the other hand, because the path of innovation is highly uncertain and technology is rapidly changing, barriers to entry that seem great today could disappear tomorrow, and the potential costs of enforcement are large as well. The threat of potential entry by innovative firms has been a significant part of Microsoft's defense in the Department of Justice case.

5. *Standard Setting*

In network industries competitive problems may arise in the competition to develop market standards. Dominant firms may have an incentive to adopt competitive strategies that support a single standard by preventing the products of rivals from achieving compatibility. Indeed, when the dominant firm's product becomes the standard for the industry, firms that are developing alternative standards may find it difficult to compete effectively (Farrell and Katz 1998). Alternatively, firms might collude to affect the outcome of the standard-setting (price-setting) process. Such collusion can be difficult to detect because firms often have pro-competitive reasons for cooperating in the race to develop new technology. Indeed, firms often possess assets and skills that can make collaboration in developing a market standard an efficient arrangement. An example is the pooling of technology related to video and audio streaming on the Internet. In its business review letter of the MPEG-II patent pooling agreement, the Antitrust Division of the Department of Justice spelled out a set of conditions under which the pooling of assets would be deemed to be pro-competitive. Where and how to draw the line between procompetitive sharing of assets of skills and other anticompetitive activities that will discourage competitors in the battle for the next generation standard remains a highly significant antitrust issue.

As a general rule, one should expect that sufficient competition will exist to develop a new product or market standard whenever more than a few independent entities exist that can compete to develop the product or standard. This rule of thumb is consistent with case law and conclusions in guidelines published by the US antitrust authorities. For example, the DOJ/FTC *Antitrust Guidelines for the Licensing of Intellectual Property* conclude that mergers or other arrangements among actual or potential competitors are unlikely to have an adverse effect on competition in research and development if more than four independent entities have the capability and incentive to engage in similar R&D activity (DOJ/FTC 1995). Similarly, the 1984 DOJ/FTC *Merger Guidelines*

(revised in 1992) conclude that mergers among potential competitors are unlikely to raise antitrust concerns if there are more than a few other potential entrants that are similarly situated.

5.1 Leveraging

Leveraging occurs when a firm uses its advantage from operating in one market to gain an advantage in selling into one or more other, generally related markets. Leveraging by dominant firms in network industries may take place for a variety of reasons that can be procompetitive or anticompetitive, depending on the circumstances. The challenge for antitrust policy is to distinguish between the two. On one hand, leveraging can be seen as a form of vertical integration, in which a firm may improve its distribution system, economize on information, and/or improve the quality of its products. Leveraging, however, can be anticompetitive if its serves as a mechanism by which a dominant firm is able to raise its rivals' costs of competing in the marketplace.

Leveraging can be accomplished by a variety of practices (e.g., tying, bundling, exclusive dealing), each of which may have anticompetitive or procompetitive aspects, or a combination of the two. For example, with tying, a firm conditions the purchase (or license) of one product—the tied product—on the purchase (or license) of another product—the tying product. A firm might choose a tying arrangement for procompetitive reasons, including cost savings and quality control. Suppose, for example, that a dominant firm offers to license its dominant technology only to those firms that agree to also license that firm's complementary product, and suppose that the complementary product builds on the firm's next generation technology. Such a tying arrangement could allow the dominant firm to create a new installed base of users of its next generation technology in a manner that would effectively foreclose the opportunities of competing firms to offer their products in the battle for the next generation technology (Farrell and Saloner 1986).

6. International Antitrust

The emergence of a global marketplace raises significant antitrust policy issues. On one hand, free trade and the opening of markets makes it less likely that mergers will significantly decrease competition, and indeed, creates greater opportunities for merging or cooperating firms to generate substantial efficiencies. On the other hand, international price-fixing agreements are more difficult for US or other domestic enforcement agencies to police, since success often requires some degree of cooperation with foreign governments or international organizations. In the 1990s we saw a significant increase in the prosecution of international price-fixing conspiracies. Indeed, in 1998, the Department of Justice collected over $1 billion in criminal fines, the bulk of which arose from the investigation of a conspiracy involving the vitamin industry.

The internationalization of antitrust raises a host of difficult jurisdictional and implementation problems for all antitrust enforcement countries. One can get a sense of the enormous difficulties involved by looking at the problems from the perspective of the US. The Sherman Act clearly applies to all conduct, which has a substantial effect within the US. However, crucial evidence or culpable individuals or firms may be located outside the US. As a result, it is imperative for US antitrust authorities to coordinate their activities with authorities abroad. This is accomplished in part through mutual assistance agreements, as, for example, in the agreement between the US and Australia under the International Enforcement Assistance Act. It is also accomplished through informal 'positive comity' arrangements, whereby if one country believes that its firms are being excluded from another's markets, it will conduct a preliminary analysis and then refer the matter to the foreign antitrust authority for further investigation, and, if appropriate, prosecution. (Such an agreement was reached with the EU in 1991.) Further, in 1998, the Organization for Economic Cooperation and Development formally recommended that its member countries cooperate in enforcing laws against international cartels. What role, if any, the World Trade Organization will play in encouraging cooperation or resolving antitrust disputes remains an open question today.

7. Conclusions

Antitrust policy has undergone incredible change over the twentieth century. As the views about the nature of markets and arrangements among firms held by economists and others have changed, so has antitrust. The early focus was on the possible anticompetitive effects of mergers and other horizontal arrangements among firms. The primary source of market power was the presence of scale economies in production. However, vertical arrangements receive significant critical treatment, and the sources of market power are seen as coming from the demand side as well as the supply side. Further, a wide arrange of contractual practices are now judged as creating substantial efficiencies, albeit with the risk that market power will be used for exclusionary purposes. Finally, significant changes in international competition, resulting from free trade and the communication and information revolution, have reinforced the importance of reshaping antitrust to meet the needs of the twenty-first century. Significant new antitrust challenges lie ahead. Whatever those challenges may be, we can

expect that industrial organization economics and antitrust policy will be sufficiently flexible and creatively to respond appropriately.

8. Statutes

Clayton Act, 38 Stat. 730 (1914), as amended, 15 USCA §§12-27 (1977).

Federal Trade Commission Act, 38 Stat. 717 (1914), as amended, 15 USCA §§41-58 (1977).

Robinson-Patman Act, 49 Stat. 1526 (1936), 15 USCA §13 (1977).

Sherman Act, 26 Stat. 209 (1890), as amended, 15 USCA §§1-7 (1977).

9. Cases

Board of Trade of the City of Chicago vs. USA, 246 US 231 (1918) (see Sect. 3).

Brooke Group Ltd. vs. Brown & Williamson Tobacco Corp., 509 US 209, 113 S. Ct. 2578 (1993) (see Sect 3.4).

California Dental Association vs. FTC, 119 S.Ct. 1604 (1999) (see Sect. 3).

Federal Trade Commission vs. Indiana Association of Dentists, 476 US 447 (1986) (see Sect. 3).

Federal Trade Commission vs. Procter & Gamble Co, Inc., 386 US 568, 574 (1967) (see Sect. 3.1).

GTE Sylvania, Inc. Continental TV, Inc., 433 US 36, 45 (1977) (see Sect. 3.2).

Illinois Brick Co. vs. Illinois, 431 US 720, 97 St. Ct. 2061 (1977) (see Sect. 3.4).

Standard Oil Co. vs. United States, 221 US 1 (1911) (see Sect. 3).

USA vs. Arnold, Schwinn & Co., 388 US 365 (1907) (see Sect. 3.1).

USA vs. Microsoft, Civil Action, 98-1232 (1998) (see Sect. 4).

USA vs. Von's Grocery Co., 384 US 270, 301 (1966) (see Sect. 3.1).

See also: Business History; Business Law; Firm Behavior; Policy Process: Business Participation; Regulation and Administration; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulation Theory in Geography; Regulatory Agencies

Bibliography

Bain J 1968 *Industrial Organization*. Wiley, New York

Baker J B, Rubinfeld D L 1999 Empirical methods in antitrust: Review and critique. *American Law and Economics Review* **1**: 386–435

Baumol W, Panzar J, Willig R 1982 *Contestable Markets and the Theory of Industry Structure*. Harcourt Brace, New York

Bork R 1966 Legislative intent and the policy of the Sherman Act. *Journal of Law & Economics* **7**: 7–48

Bork R 1978 *The Antitrust Paradox: A Policy at War with Itself*. Basic Books, New York

Coase R 1937 The nature of the firm. *Economica* **4**: 380–405

Demsetz H 1974 Two systems of belief about monopoly. In: Goldschmid I, Mann H M, Weston J F (eds.) *Industrial Concentration: The New Learning*. Little Brown, Boston, pp. 164–83

Dixit A 1979 A model of duopoly suggesting a theory of entry barriers. *Bell Journal of Economics* **10**: 20–32

Easterbrook F 1984 The limits of antitrust. 63. *Texas Law Review*. **1**: 1–40

Farrell J, Katz M L 1998 The effects of antitrust and intellectual property law on compatibility and innovation. *The Antitrust Bulletin* **43**: 609–50

Farrell J, Saloner G 1986 Installed base and compatibility: Innovation, product preannouncements, and predation. *American Economic Review* **76**: 940–55

Fisher F M 1989 Games economists play: a noncooperative view. *RAND Journal of Economics* **20**: 113–24

Krattenmaker T G, Salop S C 1986 Anticompetitive exclusion: Raising rivals' costs to achieve power over price. *Yale Law Journal* **96**: 209–93

Kreps D M, Wilson R 1982 Reputation and imperfect information. *Journal of Economic Theory* **27**: 253–79

Melamed A D 1998 Exclusionary Vertical Agreements, Speech before the ABA Antitrust Section, April 2

Milgrom P, Roberts J 1982 Predation, reputation and entry deterrence. *Journal of Economic Theory* **27**: 280–312

Ordover J A, Saloner G, Salop S C 1990 Equilibrium vertical foreclosure. *American Economic Review* **80**: 127–42

Posner R 1979 The Chicago School of antitrust analysis. *University of Pennsylvania Law Review* **127**: 925–48

Rubinfeld D L 1998 Antitrust Enforcement in Dynamic Network Industries. *The Antitrust Bulletin* **Fall-Winter**: 859–82

Schwartz L B 1979 Justice and other non-economic goals of antitrust. *University of Pennsylvania Law Review* **127**: 1076–81

Shapiro C 1989 The theory of business strategy. *RAND Journal of Economics* **20**: 125–37

Stigler G 1964 A theory of oligopoly. *Journal of Political Economy* **72**: 44–59

Stigler G J 1985 The origin of the Sherman Act, 14 *Journal of Legal Studies* **14**: 1–12

US Department of Justice and Federal Trade Commission 1984 *Merger Guidelines*. Reprinted in 4 Trade Reg. Rep. (CCH) paras. 3, 103

US Department of Justice and Federal Trade Commission 1992 *Horizontal Merger Guidelines*. Reprinted in 4 Trade Reg. Rep. (CCH) paras. 13, 104

US Department of Justice and Federal Trade Commission 1995 *Antitrust Guidelines for the Licensing of Intellectual Property*, April 6, 1995. Reprinted in 4 Trade Reg. Rep. (CCH) 113, 132

Weiss L 1974 The concentration-profits relationship and antitrust law. In: Goldschmid H, Mann H M, Weston J F (eds.) Industrial Concentration: The New Learning. Little Brown, Boston pp. 184–232

Williamson O 1985 *The Economic Institutions of Capitalism*. Free Press, New York

D. L. Rubinfeld

Anxiety and Anxiety Disorders

In addition to happiness, sadness, anger, and desire, anxiety is one of the five important normally and regularly occurring emotions which can be observed

throughout all human cultures and in several animal species (Ekman 1972). Anxiety *per se* is a complicated concept since several difficulties arise in defining this emotion and, in addition, it has to be differentiated from fear and stress (see below). Moreover, anxiety refers to a variety of other emotional experiences, for example, apprehensiveness, nervousness, tension, and agitation, which are partially directly related to anxiety, but occur also in other emotional states. Anxiety is defined by subjective, behavioral, and physiological characteristics. Anxiety involves the experience of dread and apprehensiveness and the physiological reactions of anxiety usually include trembling, sweating, elevated heart rate and blood pressure, and increases in muscle tone and skin conductance. Anxiety is defined as pathological when occurring inadequately or with much more pronounced severity and debilitating features. An additional defining criterion in standardized diagnostic manuals is the concomitant occurrence of anxiety and avoidance. Representatives for these diagnostic entities, in which anxiety is the leading symptom, are panic disorder and generalized anxiety disorder. Anxiety is experienced in phobic disorders when the subject is confronted with the feared stimulus, which results in its avoidance. One of the reasons for the declaratory confusion of the term *anxiety* is its psychological similarity to fear and its vegetative similarity to stress. Similar to anxiety, *fear* also includes the experience of dread, and fear seems to be largely included into the concept of anxiety. Moreover, anxiety and fear induce bodily reactions, which represent the so-called stress responses. Stress responses again can be divided into two large entities, the excitatory fight or flight response postulated by Cannon (1929) and the endocrine stress concept raised by Selye (1956). In contrast, the results of several studies indicate that the brain function during stress involves structures which mediate the perception of anxiety such as amygdala, hippocampus, and other limbic structures (see below).

1. *Differentiation of Anxiety*

Both anxiety and fear are regularly experienced within a range of normal emotional responses of everyday life. Specifically, fear is necessary to achieve personal growth and individual freedom during ontogeny.

1.1 *Anxiety*

Anxiety represents one of the five basic emotional states and can be defined by affective (basic emotional feelings), perceptive (realization of bodily or psychomotor sensations), and cognitive components. Besides these subjective components, behavioral and physiological characteristics can be used to define anxiety phenomenologically. In contrast to experiencing anxiety during everyday life, anxiety as psychopathologic disturbance or anxiety disorder includes specific diagnostic criteria, neurobiological dysfunctions, and a specific genetic background and leads to social and occupational disabilities.

1.2 *Fear*

Fear is the normal reaction to threatening stimuli and is common in everyday life. When fear is greater than warranted by the situation or starts to occur in inappropriate situations, a specific phobia arises which belongs to the diagnostic entity of anxiety disorders. One distinction between fear and anxiety is based upon the presence of commonly defined stimuli, a realistic relation between dangerousness and elicited fear, and the potential to adapt to the stimuli. Specific phobias are defined as persistent, irrational, exaggerated, and pathological dreads of a stimulus or situation combined with a compelling desire to avoid this feared challenge.

1.3 *Stress*

Stress is also regularly experienced by all organisms, and refers generally to physical or psychological stimuli or alterations that are capable of disrupting the homeostasis of one individual or animal. With regard to psychological aspects of stress, predictability, control, and coping skills are important determinants, which, however, are also threatened during anxiety or fear. Hence, anxiety and fear also represent important psychological stressors with their physiological sequelae being similar to stress reactions. The differentiation between stress and anxiety is difficult, since psychological and biological aspects of stress are linked to each other and are mutually interdependent.

2. *Anxiety Disorders*

As defined by means of the diagnostic and statistical manual of mental disorders (*DSM-IV*; American Psychiatric Association 1994), anxiety disorders comprise a heterogeneous group which share *anxiety* as a symptom. However, each of these disturbances has a different etiology and outcome and different physiological characteristics.

2.1 *Panic Disorder*

Panic disorder is characterized by recurrent paroxysmal anxiety which can even surmount the fear of death during an acute myocardial infarction. These attacks are regularly combined with bodily sensations such as tachycardia, suffocation, shaking, trembling,

sweating, abdominal distress, and dizziness. They typically have a sudden onset and are either unpredictable or occur before or during specific situations. If these attacks are affiliated to specific situations they can lead to avoidance of these specific events and an agoraphobia develops (see *Panic Disorder*).

2.2 Phobias

Phobias are usually differentiated into three specific subtypes: agoraphobia, as frequent sequels of panic disorder, social phobias, and simple phobias. Agoraphobia is the fear of being in situations from which escape is not immediately possible. The symptoms regularly include depersonalization, derealization, dizziness, and cardiac symptoms. Agoraphobia may occur without preceding a panic attack, but remain consolidated between attacks. Social phobias are characterized by the fear that someone may be exposed to a situation where this person is inappropriately scrutinized by others or where this person may behave inadequately. Exposure leads to prominent symptoms of anxiety including bodily alterations, and anticipatory anxiety leads to the avoidance of these situations. Simple phobias are characterized by a persistent fear of a defined object or situation such as fear of spiders or fear of height. The anticipatory anxiety is common and these stimuli are largely avoided, which can impair daily life routines.

2.3 Generalized Anxiety Disorders

This disorder is characterized by an unspecific, unrealistic, and excessive apprehension about a large variety of future events, which are difficult to control for the person. It has been classified as a chronic disorder lasting longer than six months. Specific physiological symptoms include motor tension, autonomic hyperreactivity, and sleep disturbances.

2.4 Post-traumatic Stress Disorders (PTSD)

In contrast to the above-mentioned disorders, PTSD does not present anxiety as the leading symptom. Typically this disturbance follows a psychologically and also physically distressing incident which is outside the realm of normal human experience and which is frequently life threatening. Incidents may be experienced alone or in groups and include natural or human-made emergencies. The disorder is characterized by re-experiencing the traumatic event in different ways: by recurrent intrusive thoughts, dreams, and flashbacks accompanied by intensive feeling or reliving the trauma in a dissociative state. Reminders of the incident can cause intense psychological distress. In addition, loss of interest, depressive mood, and feelings of detachment from family members or friends frequently occur. PTSD patients have a persistently increased level of arousal, concentration problems, sleep disturbances, and an increased sympathetic arousal, but surprisingly a decreased secretion of cortisol and related stress hormones (see *Post-traumatic Stress Disorder*).

2.5 Obsessive Compulsive Disorders

As mentioned for PTSD, this disorder also does not present anxiety as a leading symptom: essential features of this disorder are recurrent obsessions and compulsions that are strong and persistent enough to cause distress or that are disabling in daily life. Obsessions are defined as prominent and persistent thoughts about the same person, object, or action and compulsions are repetitive intentional behaviors that accompany these obsessive thoughts. All these behavioral manifestations are performed stereotypically and excessively. The actions are a function of a conflict between wanting to pursue and to resist this behavior. Whereas the patient realizes the irrationality, the compulsive behaviors typically provide a release of tension and fear (see *Obsessive–Compulsive Disorder*).

3. Sources of Anxiety in Humans

Anxiety has to be derived from complex origins and an interplay of genetic, biological, social, and psychological events and influences. Among the most important factors are the genetic or biological disposition, the developmental and environmental impact upon one individual, and acute stressors and experiences which challenge one person and lead to a variety of adaptational changes.

3.1 Genetic and Biological Disposition

Hints for a genetic background of anxiety disorders and indicators for their heritability have been considered as long as for mood disorders, despite the change of diagnostic criteria and labels for different anxiety disorders over the years. Among the anxiety disorders the genetics of panic disorder and generalized anxiety disorder have been most studied (Burrows et al. 1992). From a methodological point of view, family studies, twin studies, and linkage studies have to be differentiated. Regarding panic disorder it has been shown that relatives of patients have an increased risk of a similar disturbance. Among relatives a risk of up to 30 percent is reported, which is significantly different from a lifetime prevalence of about 2 percent in the general population. Also, twin studies support a heritable component since several

studies indicate that the concordance rates for panic disorder are higher in mono- than in dicygotic twins. Linkage studies have been attempted several times, but no single gene locus could be identified. Considering the great complexity of this disorder, it is to be expected that a single gene locus is unlikely to be responsible for the diagnostic entity *panic disorder*. However, association studies might lead to the detection of genes responsible for an enhanced vulnerability to anxiety disorders. Support for a genetic basis of anxiety stems also from preclinical studies. By selective breeding, different lines of rats can be established which differ markedly in their innate anxiety behavior. This might be also of importance in the initiation of alcohol consumption by these rats: animals with higher innate levels of anxiety show a greater preference for alcohol. It has been demonstrated by genetic knockout strategies in mice for specific receptors that deficiency of receptors, which are considered to be involved in anxiety and stress reactions, is correlated with a lower innate anxiety behavior. These receptors include, among others, the corticotropin-releasing factor (CRF) receptor (Steckler and Holsboer 1999; see *Endocrinology and Psychiatry*).

3.2 Social and Environmental Influences

Although the strong impact of child rearing or untoward events during childhood is evident, it is worth remembering that simple relationships cannot be constructed. Some child-rearing conditions such as family conflict situations have been correlated with anxiety during adolescence. Also other factors such as parental support, child rearing style, and personality traits have been linked to anxiety in adolescence (Spielberger and Sarason 1977). Especially spanking in childhood carries an association and increased lifetime prevalence of anxiety disorders as well as depression and alcohol abuse. Despite the great interest in this aspect of anxiety, the literature on the relationship between parent–child interactions and anxiety disorder is inconsistent. Besides these factors age also contributes to the expression of anxiety. Anxiety and anxiety disorders have a higher incidence in adolescence that cannot be reduced only to the use of different diagnostic tools, but seems to be related to other factors (*Psychiatry and Sociology*).

3.3 Life Events

In context with environmental and developmental influences, traumatic events, which are regularly out of the realm of normal human experience, are of special importance. A traumatic event can lead to increased anxiety after the event, but may have also long-term effects that emerge with future traumas. This has several implications: PTSD and increased arousal in response to the experience of a traumatic event have been related to adversities during childhood (e.g., childhood abuse). Childhood abuse, however, appears to increase an individual's risk to develop PTSD in response to extreme stressors in adulthood. Besides abuse, other adversities such as parental loss have also been related to the development of anxiety including PTSD (Friedman et al. 1995, Heim and Nemeroff 1999).

4. Neuronal Basis of Anxiety

Both the increasingly differentiated analysis of anatomical structures and biochemical and neurophysiological pathways have led to a more detailed concept about the neurobiology of anxiety and especially of panic attacks.

4.1 Neuroanatomical Structures

Whereas fear is one of best investigated emotions in terms of brain mechanisms, a direct comparison of animal models of fear is limited with respect to the spectrum of human anxiety disorders. It has been proposed that panic disorder involves the same pathways that support conditioned fear in animals. These findings support the theory that panic attacks arise from loci in the brain stem that control serotonergic and noradrenergic neurotransmission and respiratory control. Further, it was postulated that anticipatory anxiety arises from kindling of limbic areas and phobic avoidance from precortical activation. Sensory inputs for conditioned stimuli are mediated through the connection of the anterior thalamus to the lateral and then to the central nucleus of the amygdala. The latter coordinates physiological and behavioral responses related to anxiety. Efferents of this nucleus have several targets, for example, the parabrachial nucleus producing an increase in respiratory rate, the lateral nucleus of the hypothalamus activating the sympathetic system, the locus coeruleus resulting in an increase in noradrenaline release with its sequelae of increased blood pressure and heart rate and behavioral fear responses, and the nucleus paraventricularis of the hypothalamus causing an increase in corticosteroids. As outlined by LeDoux (1998), the overlap between effects of brain-stem activation by the central nucleus of the amygdala in animals with physiological effects in humans during panic attacks is striking. Besides these connections, mutual interactions between the amygdala and the thalamus, the prefrontal and the somatosensory cortex are obvious. An impairment of the cortical processing could lead to a misinterpretation of visceroafferent cognitions, leading to the activation of the above-mentioned systems. Because of these complex interactions with the autonomic and

endocrine regulation, panic attacks apparently result in equivocal physiological and behavioral sequelae (see below) (Gorman et al. 2000).

4.2 Transmitter Systems

Considering a large body of clinical and preclinical findings, the monoamine transmitters serotonin and noradrenaline and the neuropeptide corticotropin-releasing factor are most important in the regulation of the neuroanatomical structures involved in anxiety and fear. Regarding serotonergic neurotransmission, several findings support its involvement in mediating anxiety: serotonin neurons in the raphe nuclei have an inhibitory effect on noradrenergic neurons at the locus coeruleus. In addition, these neurons act at the periaquaeductal gray modifying the escape responses and are also thought to inhibit the hypothalamic release of CRF. From a clinical point of view, these findings are supported by the effects of serotonin reuptake inhibitors, that is, pharmaceuticals which inhibit the uptake of serotonin back into the presynaptic neuron and increase the amount of serotonin in the synapse to bind both to pre- and postsynaptic sites where more than 13 subtypes of serotonin receptors coupled to different membranous and intracellular effector systems have been identified (Kent et al. 1998). Overall, a long-term increase of serotonergic transmission by these compounds exerts antipanic and anxiolytic effects.

The other important neurotransmitter system involved in anxiety disorders is the noradrenergic system (Sullivan et al. 1999). Noradrenalin neurons largely originate in the locus coeruleus and some other nuclei in the medulla and pons. Projection sites include the prefrontal cortex, the amygdala, the hippocampus, the hypothalamus, the thalamus, and the nucleus tractus solitarius. Conversely, the locus coeruleus is innervated by the amygdala. Therefore, the locus coeruleus seems to integrate external sensory and visceral afferents influencing a wide range of neuroanatomical structures related to fear and stress. Clinically it has been proven that noradrenergic alpha-2 receptor antagonists such as yohimbine can be used to provoke panic attacks acting via an increase in synaptic availability of noradrenaline. In contrast, clonidine, an alpha-2 adrenergic agonist, exerts anxiolytic-like effects in experimentally induced panic attacks such as lactate infusions.

Both transmitter systems, the serotonergic and the noradrenergic, interact with the release of CRF, a 41 amino acid neuropeptide (Arborelius et al. 1999, Koob 1999). Neurons containing CRF and its receptors have been shown to be distributed throughout the brain and CRF has emerged as a neurotransmitter that plays a central role not only in stress regulation but also in anxiety and depression. CRF neurons are found in the amygdala, the hypothalamus, and the locus coeruleus. Their activity is regulated by adaptive responses. CRF neurons also project from the amygdala to the locus coeruleus. Hence CRF could act as a modulator of cognitive and physiological symptoms of anxiety. This factor initiates on the one hand a humoral cascade which enhances via the secretion of corticotropin the release of glucocorticoids, which in turn act at central gluco- and mineralocorticoid receptors. CRF is on the other hand involved in the modulation of anxiety and depression. Stress results in increased CRF concentrations in the locus coeruleus and CRF increases the firing rate noradrenergic neurons. In contrast, noradrenaline also potently stimulates the release of CRF. The involvement of CRF is interesting also with respect to respiratory alterations during panic attacks which have led to the 'suffocation false alarm theory' (Klein 1993), since CRF seems to be an important modulator of respiratory centers in the brain stem. Several studies support the contention that antagonists and inhibitors of the synthesis of CRF exert anxiolytic-like effects. A CRF-1 receptor-deficient mouse showed a significantly lowered anxiety behavior in comparison with controls. Antagonists of CRF receptors have also been examined in clinical trials for their anxiolytic and antidepressant potency.

Since serotonin reuptake inhibitors are involved in the inhibitory regulation of noradrenergic neurons of the locus coeruleus and since serotonin reuptake inhibitors are thought to reduce the hypothalamic release of CRF, these complex interactions suggest that noradrenergic, serotonergic, and CRF-regulated neurotransmission are linked together mediating the responses to anxiety, fear, and stress (see *Peptides and Psychiatry*).

5. Models of Anxiety

Anxiety is not merely one of the most important emotions throughout phylogeny and ontogeny, but is also provocapable by different means and can then be readily observed under experimental conditions. Both in humans and in animals a variety of investigations have been conducted which allow thorough insights into the pathophysiological conditions and the cognitive and neurobiological processes involved in these specific emotional states.

5.1 Animal Models

Animal studies in anxiety can be used both for investigating the physiological and anatomical substrates of anxiety and for studying pharmacological strategies for their potential anxiolytic or anxiogenic effects (Westenberg et al. 1996). Basically there are two types of animal behavioral models to detect anxiolytic-like effects. One is based upon conditioned behavior

and detects responses controlled by operant conditioning procedures. The other type of model involves unconditioned behavior which is mainly based upon naturally occurring behavior and is called ethologically based models. A different type are separation models, which mainly investigate the behavior of an offspring during separation from its mother and involves the investigation of developmental disturbances.

5.1.1 Conditioned emotional responses. The most important conditioned models comprise conflict models where behavior is suppressed by aversive stimulation. The release of the suppressed behavior without altering the levels of punished responding following pharmacological intervention is estimated as the anxiolytic-like effect. Using these models in rodents, for example, benzodiazepines are consistently effective, whereas for other compounds such as serotonin reuptake inhibitors anxiolytic-like effects are difficult to find. Other important models are the fear-potentiated startle response where the startle response of rats is augmented by fear conditioning. During the conditioning phase another stimulus is presented signaling the presence of, for example, a shock stimulus. During the startle response presentation of the stimulus enhances the startle amplitude. Also in this paradigm benzodiazepines exert anxiolytic effects. In addition to these models a variety of other conditioned responses and active and passive avoidance reactions can be determined.

5.1.2 Ethological models. In contrast to the conditioned responses, the ethological models are based upon naturally occurring behavior. The most important and frequently used models are the elevated plus-maze, the open-field, and the dark–light-box models. The elevated plus-maze uses the conflict between exploration and aversion to elevated open places. In this test, the anxiety is generated by placing the animal on an elevated open arm, where height and openness rather than light are responsible for anxiogenic effects. The device is shaped as a plus sign with two open arms and two arms enclosed by high walls. The time that rodents spend on the open arms and the number of entries are related to the anxiolytic effects. The open-field test investigates the distance traveled by rodents in a locomotor box within a given time interval. Usually, rodents avoid open areas and try to remain at the edge of the locomotor box. The overall distance traveled and the transitions of the central area of the box are related to the anxiolytic potency by a treatment. The dark–light box uses the number of transitions between a light and a dark, closed compartment as measure of anxiety, since rodents prefer the dark compartment. Peptide receptor ligands such as CRF (Arborelius et al. 1999) and cholecystokinin tetrapeptide (CCK-4) (Bradwejn and Vasar 1995) show anxiogenic effects in all three paradigms, whereas other substances such as atrial natriuretic peptide (ANP) (Wiedemann et al. 2000) and neuropeptide Y (Heilig and Widerlöv 1995) indicate anxiolytic-like effects.

5.2 Human Models

The interest in human models of anxiety has been catalyzed to a large extent by findings that panic attacks can be stimulated by a variety of different psychological, physiological, and pharmacological paradigms. Attempts to alter basic anxiety levels and especially those via induction of psychological stress have led to equivocal findings. This might indicate that within the above-mentioned neuroanatomical and physiological systems strong interfering factors exist, which modulate the responses to anxiety and stress.

5.2.1 State and trait anxiety. When investigating human anxiety, the distinction between state and trait anxiety is most important. State anxiety can be defined as a transitory emotional state consisting of feelings of apprehension, nervousness, and physiological sequelae such as an increased heart rate or respiration (Spielberger 1979). Whereas everyone can experience state anxiety occasionally, there are large differences among individuals in the frequency, duration, and severity. State anxiety can be determined by several rating instruments developed in the past. Trait anxiety represents a fairly stable characteristic related to personality. Experiencing more frequently state anxiety combined with a general view of the world as being threatening and dangerous is used as marker of trait anxiety. The initiation and maintenance of trait anxiety have been related to several factors as outlined above.

5.2.2 Challenge studies. The profound interest in state anxiety and especially panic attacks stems from a large variety of investigations provoking anxiety and panic attacks experimentally (Nutt and Lawson 1992). Panic attacks are unique in the spectrum of psychiatric disorders since their core psychopathology is temporally limited and can be provoked and investigated under laboratory conditions. Information provided by these studies has led to new cognitive and physiological theories about the basis of panic anxiety and anxiety diseases. Moreover, owing to the experimental character of these investigations, closer comparisons with experiments in animals can be drawn in contrast to other psychiatric animal models. Panic attacks can be elicited by various

Table 1
Experimentally induced panic attacks

Panicogen	Heart rate stimulation	Respiratory stimulation	HPA stimulation	NE stimulation
Cognitive stimuli	+	+	−	−
Metabotropic agents				
L-Lactate	+	+	−	−
D-Lactate	+	+	−	−
Bicarbonate	−	+	−	−
CO_2	+/−	+	−	−
Receptor ligands				
Yohimbine	+	−	+	+
Fenfluramine	+	−	+	−
β-Carboline	+	−	+	+
Caffeine	+	+	+	−
CCK-4	+	+	+	(+)
CRF	+	+	+	(+)

HPA, hypothalamo–pituitary–adrenocortical system; NE, noradrenergic system; CCK-4, cholecystokinin tetrapeptide; CRF, corticotropin-releasing factor.

means which are listed in Table 1. As indicated, the different paradigms can be differentiated into cognitive, metabotropic and direct receptor interactions. Especially naturally occurring, cognitive, and metabotropic panic attacks share many features. One of the most amazing findings is that, despite the dramatic anxiety felt, a uniform stress response either of the hypothalamo–pituitary–adrenocortical (HPA) or the sympathetic system is largely missing. These findings led to the hypothesis that in addition to a variety of stimulating agents, strong inhibitors also exist which physiologically antagonize the altered transmitter and modulator systems involved in panic anxiety. Considering the hypothesis that CRF is one important modulator of anxiety in humans and rodents, it is astonishing that no activation of the HPA system occurs in naturally occurring or metabotropic panic attacks. In contrast, compounds interfering with monoamine and peptide receptors stimulate the HPA activity and the noradrenaline system. Of the latter group, one of the most potent panicogens is CCK-4 (Bradwejn and Vasar 1995), which seems to exert its effect via CRF. Up to now only a few modulators have been identified, which are able to inhibit the exaggerated HPA system activity and, in addition, exert anxiolytic-like effects. One of these inhibitors might be ANP which is secreted in the atria of the heart and in various brain regions involved in anxiety. Hence it may be speculated that peptides such as ANP might help to explain the so far unknown mechanisms of terminating panic anxiety (Wiedemann et al. 2001). Despite a tremendously increased knowledge about the induction of anxiety, fear, and stress, the mechanisms of coping and terminating these emotional alterations need further investigations.

See also: Anxiety and Fear, Neural Basis of; Anxiety Disorder in Children; Bowlby, John (1907–90); Culture and Emotion; Emotion: History of the Concept; Emotions, Evolution of; Emotions, Psychological Structure of; Emotions, Sociology of; Freud, Sigmund (1856–1939); Harlow, Harry Frederick (1905–81); Stress and Health Research; Test Anxiety and Academic Achievement

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Diseases*, 4th edn. American Psychiatric Association, Washington, DC

Arborelius L, Owens M J, Plotsky P M, Nemeroff C B 1999 The role of corticotropin-releasing factor in depression and anxiety disorders. *Journal of Endocrinology* **160**: 1–12

Bradwejn J, Vasar E (eds.) 1995 *Cholecystokinin and Anxiety: From Neuron to Behavior*. Springer, New York

Burrows G D, Roth M, Noyes R 1992 *Contemporary Issues and Prospects for Research in Anxiety*. Elsevier, Amsterdam

Cannon W B 1929 *Bodily Changes in Pain, Hunger, Fear and Rage*. Appleton, New York

Ekman P (ed.) 1982 *Emotion in the Human Face*, 2nd edn. Cambridge University Press, Cambridge, UK, New York; Editions de la Maison des Sciences de l'Homme, Paris

Friedman M J, Charney D S, Deutch A Y (eds.) 1995 *Neurobiological and Clinical Consequences of Stress: From Normal Adaptation to Post-traumatic Stress Disorder*. Lippincott-Raven, Philadelphia, PA

Gorman J M, Kent J M, Sullivan G M, Coplan J D 2000 Neuroanatomical hypothesis of panic disorder, revised. *American Journal of Psychiatry* **157**: 493–505

Heilig M, Widerlöv E 1995 Neurobiology and clinical aspects of neuropeptide Y. *Critical Reviews in Neurobiology* **9**: 115–36

Heim C, Nemeroff C B 1999 The impact of early adverse experiences on brain systems involved in the pathophysiology

of anxiety and affective disorders. *Biological Psychiatry* **46**: 1509–22
Kent J M, Gorman J M, Coplan J D 1998 Clinical utility of the selective serotonin reuptake inhibitors in the spectrum of anxiety. *Biological Psychiatry* **44**: 812–24
Klein D F 1993 False suffocation alarms, spontaneous panics, and related conditions. An integrative hypothesis. *Archives of General Psychiatry* **50**: 306–17
Koob G F 1999 Corticotropin-releasing factor, norepinephrine, and stress. *Biological Psychiatry* **46**: 1167–80
LeDoux J 1998 Fear and the brain: Where have we been, and where are we going? *Biological Psychiatry* **44**: 1229–38
Nutt D, Lawson C 1992 Panic attacks: A neurochemical overview of models and mechanisms. *British Journal of Psychiatry* **160**: 165–78
Selye H 1956 *The Stress of Life*. McGraw-Hill, New York
Spielberger C D 1979 *Understanding Stress and Anxiety*. Harper and Row, New York
Spielberger C D, Sarason I G 1977 *Stress and Anxiety*. Hemisphere, Washington, DC
Steckler T, Holsboer F 1999 Corticotropin-releasing hormone receptor subtypes and emotion. *Biological Psychiatry* **46**: 1480–508
Sullivan G M, Coplan J D, Kent J M, Gorman J M 1999 The noradrenergic system in pathological anxiety: A focus on panic with relevance to generalized anxiety and phobias. *Biological Psychiatry* **46**: 1205–18
Westenberg H G M, DenBoer J A, Murphy D L (eds.) 1996 *Advances in the Neurobiology of Anxiety Disorders*. Wiley, Chichester, UK
Wiedemann K, Jahn H, Yassouridis A, Kellner M 2001 Anxiolytic-like effects of atrial, natriuretic peptide on cholecystokinin tetrapeptide-induced panic attacks. *Archives of General Psychiatry* **58**: 371–7

K. Wiedemann

Anxiety and Fear, Neural Basis of

1. *Stress, Anxiety, and Fear*

The terms stress, anxiety, and fear are commonly used in the daily language, as well as psychological, psychiatric, and neuroscientific literature. Consequently, these terms have been defined from a variety of descriptive, phenomenological, psychodynamic, and biological standpoints. Behavioral, electrophysiological, pharmacological, and genetic methods increasingly facilitate research on the underlying cell biological mechanisms and neuronal circuitry of stress, anxiety, and fear.

Stress represents an integrated neuroendocrine response of humans and animals to stimuli perceived as novel or threatening. Although a variety of physical, psychological, and physiological stimuli may be perceived as stressful, at a biological level, the stress response is highly conserved, and consists of a set of emotional, cognitive, autonomic, and metabolic reactions that are elicited to enable rapid adaptation to the environment. One of the earliest chemical signals released after a stimulus is perceived as stressful is corticotropin-releasing factor (CRF), a hypothalamic peptide (Spiess et al. 1981) acting via the portal bloodstream at the pituitary thereby stimulating the secretion of the peptide corticotropin (ACTH). ACTH enters the bloodstream and elicits the release of glucocorticoid hormones from the adrenal gland. In addition to the activation of the hypothalamo-pituitary–adrenal (HPA) axis, the organism responds with a strong activation of the adrenergic system. Thus, a stress response can be reliably evaluated by measuring the blood levels of ACTH, glucocorticoids, and adrenaline at a very early stage. In addition, a stress response elicits numerous other molecular processes in the brain, which are not detectable in the systemic circulation, and valuable data about these processes have been generated from laboratory animals, in particular rats and mice. These processes consist of changes of protein phosphorylation, gene expression, and alteration of growth factor levels leading to plastic changes in the neuronal synapses. Upon re-exposure to the same or other stressful stimuli, the pattern of behavioral and molecular processes is altered. For example, in the limbic forebrain, the transcription factor FOS is not produced after exposure to a previously encountered stimulus independently from its aversive properties, and thus, in this context FOS may be regarded as a molecular marker for novelty (Radulovic et al. 1998a).

Rodents, as well as humans, respond to a stressful situation with alerting emotional responses such as anxiety or fear. An attempt to discriminate between fear and anxiety on a biological basis is undertaken below. For the presentation of several animal models of fear and anxiety (see Rodgers 1997, for review), it may be sufficient to point to the rather focused expression of fear when compared to the diffuse expression of anxiety.

2. *Animal Models: Anxiety and Fear*

2.1 *Animal Models of Anxiety*

In rodents, anxiety is measured by a number of paradigms that evaluate different sets of behaviors of rodents under defined environmental conditions. The most commonly applied tests measure the preference for dark over light environments (elevated plus-maze test, dark–light emergence task), the intensity of muscle contraction in response to sensory stimuli (startle), or contacts during social interactions. In the shock-probe-burying test, an animal encounters an electrified probe, and copes with it by burying or avoiding the probe. This test has been regarded as a model for fear as well as anxiety. Although the findings that lateral septal lesions affect burying whereas amygdala lesions affect probe avoidance may indicate that burying and probe avoidance reflect anxiety and fear, respectively, conclusive biological evidence is still lacking.

Most of the anxiety tests also measure additional behaviors, such as exploration, locomotion, and risk assessment that may or may not be directly linked to anxiety. As the interdependence of these behaviors and anxiety is difficult to evaluate, anxiety is optimally identified and quantified when the other behaviors are not affected (Weiss et al. 2000). With the exception of the startle assay, the rodent tests for anxiety evaluate acute anxiety responses and are optimally carried out once. Multiple exposures to the test are associated with strong interference of habituation, which adds learning components to anxious behavior.

2.2 Animal Models of Fear

With the exception of inborn fears such as fears of predators, evaluation of fear responses requires a two-step procedure. Classical fear conditioning occurs when the animal learns that an originally neutral stimulus (conditioned stimulus, CS) is predictive of danger. For training, the animal is placed in a novel environment (context) and at the end of an exploratory period a foot shock (unconditioned stimulus, US) is delivered. If a tone or light is presented as a CS before the shock, the animal learns to associate the CS with the US. The fear response is behaviorally evaluated after re-exposure to the used context, tone, or light by measuring the freezing behavior reflecting conditioned fear. Alternatively, conditioned light or tone may be presented when the animals are in a startle box, and in response to the CS exhibit fear-potentiated startle.

2.3 Potentiation of Anxiety

Animal models of potentiated anxiety apply the same tests that measure anxiety; however, before the anxiety test the animal experiences a stressful event that is unrelated to the stimuli of the test situation. Potentiated anxiety is commonly observed following exposure to uncontrollable stressors, such as immobilization, social defeat, ethanol withdrawal, or classical fear conditioning. Generalization, observed in response to stimuli other than the CS used for fear conditioning, may reflect potentiated anxiety. However, animals pre-exposed to a stressor they can control or escape from, as during active avoidance (when for example the animal's transition from one context to another stops the delivery of shock), do not develop potentiated anxiety.

Potentiated anxiety can also be induced pharmacologically by compounds that increase anxious behavior. For example, injection of the peptide CRF into the bed nucleus of the stria terminalis elicits a long-lasting facilitation of the startle response, whereas its injection into the lateral septum mimics stress-induced anxiety on the elevated plus-maze.

3. Similarities and Differences Between Fear and Anxiety

Anxiety and fear have been clearly distinguished in psychiatry. However, neurobiologists commonly treat these emotional responses as the same process, possibly on the basis of the close molecular, neuroanatomical, and functional relationship between fear and anxiety. This close relationship is demonstrated by the findings that: (a) both fear and anxiety elicit similar behavioral, somatic, motor, and visceral responses mediated by common pathways within discrete hypothalamic, midbrain, and brain stem nuclei; (b) acquisition of fear responses can be significantly prevented by anxiolytic drugs such as benzodiazepines and agonists of the serotonin receptor 5-HT1A; and (c) acquisition of conditioned fear responses is closely paralleled by increased anxiety which is maximal after memory consolidation (Radulovic et al. 1998b).

The differences between anxiety and fear are significant at the level of cognitive processing of the stimuli that elicit the emotional response, so that the stimulus recognition at the level of the cortical–limbic system significantly determines whether fear or anxiety is expressed.

3.1 Neuroanatomical Circuitry of Anxiety and Fear

As presented in Table 1, it is notable that lesions of distinct brain regions differentially affect anxiety, potentiated anxiety, and conditioned fear. These data suggest that the fear and anxiety systems are diffusely distributed throughout the limbic forebrain and that their expression recruits distinct sets of behaviors under different anxiety- or fear-provoking environmental conditions. A detailed analysis of anxiety and fear responses at the neuroanatomical level has been performed in a series of experiments using the model of startle response (Davis et al. 1997). These authors delineated the limbic structures differentiating fear-potentiated from light- or CRF-potentiated startle. These structures, however, do not affect anxious behavior on the elevated-plus maze, which is in turn highly susceptible to lesions of the lateral septum (Figure 1). Another differentiation can also be made with regard to the thalamic and cortical input providing these limbic structures with sensory information. For example, whereas the basolateral amygdala receives auditory and visual projections from the thalamus and perirhinal cortex, the major telencephalic input of the lateral septum is provided by the hippocampus, which is essential for processing of spatial and contextual stimuli and receives cortical fibers from the entorhinal cortex. Accordingly, either fimbria–fornix or lateral septal lesions can disrupt elevated plus-maze anxiety.

Taking into account the remarkable dissociation between anxiety and fear among the structures of the

Table 1
Dissociation of anxiety and fear by regional brain lesions of the rat

Test	Lesioned brain area	Effect	Reference
Elevated plus-maze test	Lateral septum	Increased time spent in the open arms	Treit et al. 1993
Shock-probe burying test	Lateral septum	Decreased burying	
	Amygdala	Increased contact with probe	
CRF-enhanced startle	Bed nucleus of the stria terminalis	Impaired	Davis et al. 1997
Light-enhanced startle	Amygdala	Impaired	
	Bed nucleus of the stria terminalis	Impaired	
Fear-potentiated startle	Amygdala	Impaired	
Contextual fear conditioning	Hippocampus	Impaired	Fendt and Fanselow 1999
	Basolateral amygdala	Impaired	
	Lateral septum	Enhanced	Vouimba et al. 1998

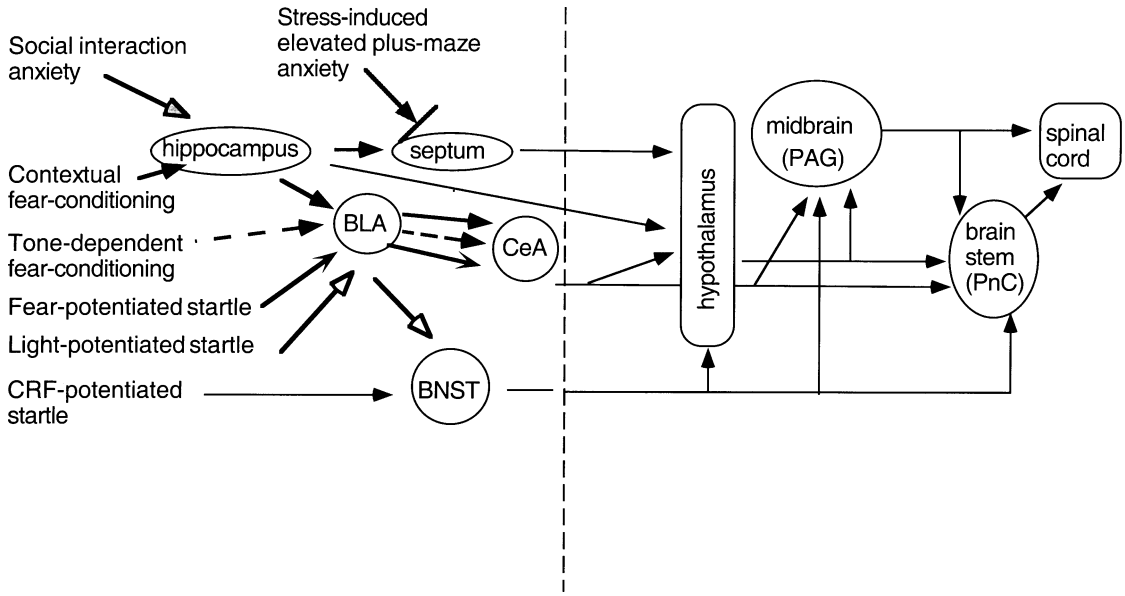

Figure 1
Brain regions and pathways mediating fear and anxiety. BLA, basolateral amygdala; BNST, bed nucleus of the stria terminalis; CeA, central amygdala; PAG, periaquaeductal gray matter; PnC, caudal pontine nucleus

limbic system, it remains unclear why anxiety develops along memory consolidation of conditioned fear. It is hypothesized that anxiety and fear potentiate each other by facilitation of the common pathways at the level of hypothalamus, midbrain, and brainstem.

3.2 Molecular Mechanisms of Anxiety and Fear

The balance between the activity of excitatory, glutamatergic, and inhibitory, GABA-ergic neurons is of utmost importance for the expression of anxiety (Clement and Chapouthier 1998). The blockade of NMDA or AMPA/kainate receptors prevents both anxiety and fear responses. In contrast, blockade of GABA receptors, in particular GABA-A receptors, increases anxiety whereas activation of the benzodiazepine site of the GABA-A receptor decreases anxiety. To date, the GABA-A receptor agonists, benzodiazepines, are widely employed in the treatment of anxiety disorders.

In general, the activation of glutamate receptors involved in anxiety is also essential for fear conditioning, whereas activation of GABA-A and sero-

tonergic 5-HT 1A receptors that prevent anxiety also prevent fear conditioning. These results are commonly observed not only after systemic drug application, but also after local injections into brain regions that differentially process anxiety and fear. Thus, generation of anxiety and conditioned fear appears to depend indistinguishably on excitatory and inhibitory amino acid transmission.

However, we have recently demonstrated that within a defined brain area, peptidergic action may differentially affect anxiety and learning measured as conditioned fear. Shortly after a stressful experience, the animal responds with general arousal and anxiety, whereas at a later time it responds with increased associative learning of specific aversive stimuli (Radulovic et al. 1999). Thus, the peptidergic action within the brain responds with a different regional time pattern after stress. Furthermore, without affecting anxiety, CRF enhances fear conditioning through hippocampal CRF receptor 1 whereas it impairs fear conditioning through lateral septal CRF receptor 2. Even within the CRF receptor 2 system, remarkable differences are observed. Thus, septal CRF receptor 2 mediates stress-induced anxiety (Radulovic et al. 1999), whereas non-septal CRF receptor 2 decreases baseline anxiety. Increased anxiety in male mice lacking CRF receptor 2 is accompanied in several brain areas by a significant reduction of phosphorylated CREB (cAMP responsive element binding protein), serving as a transcriptional activator. Thus, it appears on the basis of this observation that successful coping with a stressful stimulus is linked—at least in the mouse under investigation—to enhanced CREB phosphorylation (Kishimoto et al. 2000).

4. Perspective

Significant research efforts are targeted at the elucidation of the biological correlates of behavior with the objectives to fundamentally understand the basic principles of higher brain function and eventually emotional and cognitive disorders. In particular, anxiety disorders show increasing incidence. Free-floating anxiety or phobias occur in the absence of specific or appropriate association, and these and other clinical forms of anxiety represent chronic states in humans. The chronicity of these disorders largely impairs the delineation of the molecular mechanisms causally linked to anxiety from compensatory/secondary molecular changes. Therefore, animal experiments dealing with acute anxiety may provide unambivalent insight into the genetic, cellular, and biochemical mechanisms underlying the induction and termination of anxious behavior in the nervous system. Elucidation of these mechanisms could facilitate approaches to chronic anxiety states in humans. Interesting anxiolytic drug developments may result from the cell biological research on the transductional mechanisms assigning roles to CRF and CREB phosphorylation in coping with anxiety.

See also; Anxiety and Anxiety Disorders; Fear Conditioning; Fear: Potentiation of Startle; Fear: Psychological and Neural Aspects

Bibliography

Clement Y, Chapouthier G 1998 Biological bases of anxiety. *Neuroscience and Biobehavioral Reviews* **22**: 623–33

Davis M, Walker D L, Lee Y L 1997 Roles of the amygdala and bed nucleus of the stria terminalis in fear and anxiety measured with the acoustic startle reflex—possible relevance to PTSD. *Annals of the New York Academy of Science* **821**: 305–31

Fendt M, Fanselow M S 1999 The neuroanatomical and neurochemical basis of conditioned fear. *Neuroscience and Biobehavioral Reviews* **23**: 743–60

Kishimoto T, Radulovic J, Radulovic M, Lin C R, Schrick C, Hooshmand F, Hermanson O, Rosenfeld M G, Spiess J 2000 Gene deletion reveals an anxiolytic role for corticotropin-releasing factor receptor 2. *Nature Genetics* **24**: 415–19

Radulovic J, Kammermeier J, Spiess J 1998a Generalization of fear responses in C57BL/6N mice subjected to one-trial foreground contextual fear conditioning. *Behavioural Brain Research* **95**: 179–89

Radulovic J, Kammermeier J, Spiess J 1998b Relationship between FOS production and classical fear conditioning: Effects of novelty, latent inhibition, and unconditioned stimulus preexposure. *Journal of Neuroscience* **18**: 7452–61

Radulovic J, Rühmann A, Liepold T, Spiess J 1999 Modulation of learning and anxiety by cotricotropin-releasing factor (CRF) and stress: Differential roles of CRF receptors 1 and 2. *Journal of Neuroscience* **19**: 5016–25

Rodgers R J 1997 Animal models of 'anxiety': Where next? *Behavioural Pharmacolology* **8**: 477–96

Spiess J, Rivier J, Rivier C, Vale W 1981 Primary structure of corticotropin-releasing factor from ovine hypothalamus. *Proceedings of the National Academy of Sciences of the United States of America* **78**: 6517–21

Treit D, Pesold C, Rotzinger S 1993 Dissociating the anti-fear effects of septal and amygdaloid lesions using 2 pharmacologically validated models of rat anxiety. *Behavioural Neuroscience* **107**: 770–85

Vouimba R M, Garcia R, Jaffard R 1998 Opposite effects of lateral septal LTP and lateral septal lesions on contextual fear conditioning in mice. *Behavioural Neuroscience* **112**: 875–84

Weiss S M, Lightowler S, Stanhope K J, Kennett G A, Dourish C T 2000 Measurement of anxiety in transgenic mice. *Reviews in the Neurosciences* **11**: 59–74

J. Radulovic and J. Spiess

Anxiety Disorder in Children

Anxiety disorders relate to recurrent, excessive, and intense fears and anxiety relating to one or more situations resulting in disruption and interference to

1. Separation anxiety disorder (SAD)
   - excessive anxiety about separation from significant others or from home

2. Social phobia
   - excessive anxiety about social or performance situations in which the individual fears scrutiny or exposure to unfamiliar people, and in which the individual feels embarrassment may occur (APA 1994)

3. Agoraphobia
   - anxiety and/or avoidance of situations in which the individual feels that escape may be difficult or that help may not be available

4. Panic disorder
   - primarily defined by the occurrence of a panic attack, i.e., a short period of intense fear that is accompanied by a variety of physical or cognitive symptoms

5. Specific phobias
   - may be of animal, natural environment, blood-injection or situational type
   - defined as a marked and persistent fear of the object or situation

6. Generalized anxiety disorder (GAD)
   - excessive anxiety and worry about a number of things
   - accompanied by at least three physical symptoms

7. Obsessive compulsive disorder (OCD)
   - may experience obsessions (persistent ideas, thoughts, impulses or images, e.g., fear of contamination), compulsions (behaviors that are repeated in order to prevent or reduce anxiety, e.g., repetitive hand-washing) or both

8. Post-traumatic stress disorder (PTSD)
   - may occur after an individual experiences a traumatic event
   - includes persistent re-experiencing of the event and avoidance of stimuli related with the event
   - duration of at least one month

9. Acute stress disorder
   - may occur after an individual experiences a traumatic event
   - includes persistent re-experiencing of the event and avoidance of stimuli related with the event
   - duration must be less than one month

10. Anxiety due to a medical condition
    - occurs as a direct consequence of a general medical condition

11. Substance-induced anxiety disorder
    - panic attacks, obsessions, or compulsions as a consequence of substance intoxication or withdrawal.

Figure 1
Anxiety disorders that may present during childhood

daily living and personal competence. There are several different types of anxiety disorder that children may experience with each disorder being characterized by a pattern of presenting symptoms. While the *Diagnostic and Statistical Manual of Mental Disorders, 4th edition* (*DSM-IV*; American Psychiatric Association 1994) outlines only one anxiety disorder specific to children and adolescents, namely separation anxiety disorder, children may also experience several types of anxiety disorders that are also found in adulthood. Indeed, although there are a few developmental differences in the way that children manifest anxiety disorders compared to adults, to a large extent anxiety problems in childhood closely resemble those experienced by adults. Figure 1 lists the various anxiety disorders that may present during childhood and the primary features of these conditions.

Contrary to widely held public beliefs, childhood anxiety disorders are not simply a normal, transient part of childhood development. Rather, these problems are associated with a range of negative consequences relating to personal, social, and academic adjustment. Furthermore, such problems are likely to persist if left untreated, with many adults reporting that the onset of their difficulties commenced in childhood or adolescence.

1. Prevalence of Anxiety Disorders in Children

Anxiety disorders represent one of the most common and debilitating forms of psychopathology in children. Anxiety disorders have been variously estimated to affect between 8 and 17 percent of the child population at a given point in time depending upon the definition and assessment measures used to determine the presence of an anxiety disorder (e.g., Kashani and Orvaschel 1990). Clearly, anxiety disorders represent one of the most commonly presenting problems of childhood. Generally, anxiety problems have been found to be more common among girls than boys, although it is interesting to note that gender differences are not typically found for the prevalence of obsessive compulsive disorder (March et al. 1995). The age of the child also affects the frequency with which the problem is found. For example, separation anxiety disorder is more common among younger children and tends to decrease in prevalence in later childhood and adolescence, whereas social phobia tends to become more prevalent in later childhood and adolescence (Kashani and Orvaschel 1990, Last et al. 1992). The findings of epidemiological studies therefore rarely concur with respect to prevalence rates for specific child anxiety disorders, as the rates tend to vary depending on the age of the children involved in the study and the criteria used to determine the presence of a problem. Generally speaking, generalized anxiety disorder, social phobia, simple phobias, and separation anxiety present most frequently, with obsessive compulsive disorder and post-traumatic stress disorder being less common. The picture is also complicated by a high level of comorbidity, in that children who experience one anxiety disorder are highly likely to present with another anxiety problem. Indeed, more than 50 percent of children who manifest an anxiety disorder are also likely to meet diagnostic criteria for some other anxiety problem. Clinically anxious children are also more likely than other children to report other forms of psychopathology such as depression and attention deficit disorder.

2. Etiology

Empirical research in the area of childhood anxiety has identified a number of risk factors that, when present, increase the likelihood of the development of such problems. More recently, evidence is also emerging regarding protective factors that reduce the negative impact of risk factors.

2.1 Genetic Transmission

Anxious children are more likely to have anxious parents and anxious parents are more likely to have anxious children. These familial relationships could indicate either genetic or family environment influences. Evidence confirms that genetic factors do play a part in determining the development of childhood anxiety disorders, but clearly this explanation does not account for many cases of childhood anxiety. The research has found heritability estimates of around 40–50 percent (Thapar and McGuffin 1995), meaning that other factors play an important role, in addition to genetic determination. Although genetic factors are clearly involved in the development of anxiety for some children, it remains to be shown exactly what is inherited. What appears to be inherited is an increased propensity to develop anxiety related problems, rather than a specific anxiety disorder. This propensity may relate to some temperament pattern in the child that increases the risk of developing anxiety disorders.

2.2 Child Temperament

Temperament theorists believe that early child temperament is of etiological significance to the later development of childhood anxiety. 'Behavioral inhibition' is the term used to describe one particular pattern of childhood temperament that has been most frequently linked with childhood anxiety problems. It can be defined as a relatively stable temperament style characterized by initial timidity, shyness, and emotional restraint when exposed to unfamiliar people, places, or contexts. This temperament pattern is associated with elevated physiological indices of arousal and has been shown to have a strong genetic component. Most importantly, children exhibiting a temperament style of behavioral inhibition demonstrate an increased likelihood of developing child anxiety (see Kagan 1997 for a review of this area). Other temperament theorists argue for the existence of three stable factors: positive affectivity/surgency (PA/S), negative affectivity/neuroticism (NA/N), and effortful control (EC) (Lonigan and Phillips in press). According to this theory, high NA/N combined with low EC places children at risk for the development of anxiety problems, and there is some tentative evidence to support this proposition. However, as not all children exhibiting an early temperament style of behavioral inhibition, or high NA/N combined with low EC, go on to develop an anxiety disorder, the presence of moderating or mediating variables appears to be likely. In particular, attachment style and parenting characteristics (see Sects. 2.3 and 2.4) are likely to interact with early childhood temperament to determine the development of anxiety problems.

Although the literature regarding childhood temperament is interesting, it tells us little about the exact mechanism of action. It remains to be determined whether temperament impacts upon anxiety through greater susceptibility to conditioning processes, greater emotional and/or physiological arousability

to stressful events, or through cognitive processes. For example, it is feasible that 'at risk' temperaments have their impact though greater tendencies to detect and attend to threatening stimuli in the environment, or expectations regarding the occurrence of negative outcomes. It has been shown in several studies that anxious children are more likely than others to think about negative events and to expect negative outcomes from situations.

2.3 Parenting Characteristics

The strong family links found in childhood anxiety could also be explained to some degree by parental behavior and the family environments in which the children are brought up. Parenting behavior has been suggested to impact upon child anxiety in a number of ways. From a learning theory perspective, certain forms of parenting behavior may increase the probability that children learn to respond in an anxious manner and fail to acquire the skills needed to cope with the inevitable stressful events that occur during children's lives. Observational studies have demonstrated that parents of anxious children are more likely to model, prompt, and reinforce anxious behavior, such as avoidance and distress in stressful situations. Furthermore, parents of anxious children are more likely to draw their children's attention to the threatening aspects of situations and less likely to encourage 'brave' solutions (Rapee in press).

The parents of anxious children are also more likely to engage in behaviors that make it less likely that children will learn how to solve stressful problems themselves. Empirical enquiry has found that parents of anxious children demonstrate higher levels of overcontrolling and overprotective behaviors that disrupt coping skills development. As a group, they are also more likely to be critical of their child's coping attempts, thereby reducing children's confidence in their abilities to solve their own life problems (Dumas, La Freniere and Serketich 1995, Krohne and Hock 1991). These parenting styles may interact with childhood temperament in explaining why some behaviorally inhibited children develop anxiety problems and some do not. For example, parental overprotection and overcontrol appears to be influential in determining the stability of behavioral inhibition in children (Hirshfeld et al. 1997a, 1997b). Parental behavior has also been found to be important in determining the impact of traumatic life events upon childhood psychopathology. Following trauma, children are more likely to develop emotional and behavioral difficulties if their parents react in an overprotective manner after the event (e.g., McFarlane 1987).

It is also important to consider that children have an influence upon parents, and anxious child behavior may cause parents to behave in particular ways. In much of the literature to date, it is not clear whether the overprotective behaviors of parents are definitely a cause of childhood anxiety or whether they could be a consequence of living with an anxious child. Future research needs to clarify these relationships.

2.4 Attachment Style

Recognizing the reciprocal effects of parents and their children, researchers have started to examine the quality of the attachment relationship between children and their caregivers. For example, Warren et al. (1997) found that anxious-resistant attachment at 12 months predicted anxiety disorders in adolescence, even after the effects of maternal anxiety and infant temperament were removed. There is also some evidence that attachment style may interact with infant temperament in the prediction of early markers of anxiety problems (e.g., Fox and Calkins 1993, Nachmias et al. 1996). It appears likely that certain patterns of behavior characteristic of particular forms of early childhood temperament make it difficult for parents and children to form secure attachments. Although this research is in its early stages, it appears to be an area that warrants further investigation. The quality of the parent–child attachment relationship may represent one mechanism through which familial transmission could occur. It is well recognized that parental psychopathology, particularly depression, disrupts parenting skills and interferes in attachment relationships. It may be that high levels of parental anxiety also disrupt effective parenting and attachment relationships, thereby contributing to intergenerational transmission of anxiety.

2.5 Traumatic, Negative and Stressful Life Events

The effect of traumatic, negative, and stressful life events on the development of anxiety in children is another area of etiological investigation. Perhaps not surprisingly, higher rates of anxiety disorders are associated with a range of natural disasters and traumatic life events (Benjamin et al. 1990). However, as not all children experiencing traumatic, negative, or stressful life events go on to develop anxiety disorders, the moderating or mediating influence of parenting behavior has been suggested. Indeed, what emerges from the literature relating to the etiology of childhood anxiety is a complex picture of interacting determinants and multiple pathways through which such problems may develop.

2.6 Protective Factors

Protective factors refer to variables that increase resilience to psychological disorder by reducing the

impact of risk factors. Positive social support, particularly from a significant adult, is one such protective factor that has been suggested to provide a buffer against the development of anxiety problems, and indeed against the development of psychopathology in general. For example, a strong negative relationship has been found between child anxiety level and family social support (White et al. 1998). Child coping style is another protective factor suggested to play a role in child anxiety. Coping style is a generic term that relates to the way in which individuals attempt to cope with negative or aversive situations. There is some tentative evidence to suggest that children employing problem-focused strategies are less likely to experience psychopathology, whereas emotion-focused and avoidant coping styles are associated with higher levels of anxiety and depression (Compas et al. 1988).

3. *Assessment of Childhood Anxiety*

Research conducted on childhood anxiety is reliant upon methods of identifying and quantifying anxiety and different forms of anxiety disorder. Professionally, anxiety measures may also assist in the guidance of treatment. Various types of assessment measures are used, such as interviews, questionnaires, and direct observation. Methods also vary according to whether the informant is the child, a parent, teacher, or an independent observer. Several diagnostic interviews exist for the identification of childhood anxiety disorders, such as the Anxiety Disorders Interview Schedule for Children (Parent and Child Versions; Silverman and Albano 1996). Diagnostic interviews are extremely useful for obtaining a clinical diagnosis but are time consuming and require adequate interviewer training in order to obtain a reliable assessment. For large-scale screening purposes it may be necessary to use child and parent questionnaires for measures of childhood anxiety. Questionnaire data also provides valuable information to supplement the diagnostic interview. Various forms of questionnaires exist. Some focus on the more general aspects of the subjective, physiological, and behavioral aspects of anxiety, such as the State Trait Anxiety Inventory for Children (Spielberger 1973) or the Revised Manifest Anxiety Scale (Reynolds and Richmond 1978). There are also several fear survey schedules that examine children's fear of a wide range of trigger situations or outcomes. In the late 1990s researchers started to develop anxiety questionnaires for children and parents that examine the specific symptoms of anxiety associated with particular anxiety disorders, such as the Screen for Child Anxiety Related Emotional Disorders (SCARED; Birmaher et al. 1997) and the Spence Child Anxiety Scale (SCAS; Spence 1997). There is also an increasing number of instruments that focus in depth upon one specific anxiety disorder.

An important consideration in the assessment of child anxiety is the notoriously low reliability between child and parent sources. Given this difficulty, assessment information is generally obtained from a range of sources in order to obtain a fuller picture of the child's presenting problems.

4. *Treatment Strategies*

There is convincing evidence to demonstrate that childhood anxiety disorders can be treated effectively. Since the early 1990s, the majority of treatment outcome studies in this area have focused upon the evaluation of cognitive behavioral treatments (CBT). Most studies have examined the efficacy of a combination of treatment approaches, including the training of coping skills (e.g., positive self-talk and relaxation), graded exposure to a hierarchy of feared situations, and identification and challenging of irrational thoughts and beliefs relating to the feared events. The majority of programs have also included some form of modeling, prompting, and reinforcement of 'brave' and approach behavior to the feared situation. Generally, parents are also instructed to ignore and not to reinforce fearful and avoidance behavior.

Several studies have now demonstrated the effectiveness of this combined approach to the reduction of childhood anxiety disorders. The most frequently evaluated program has been the 'Coping Cat' approach (see Kendall 1994). Generally, the evidence suggests that around 60–70 percent of children are no longer regarded as experiencing clinically significant anxiety problems one year after participating in treatment. The challenge for researchers in the future will be to develop treatments that are effective with the 30 to 40 percent of children who either do not respond to treatment or relapse afterwards.

5. *Prevention*

Given the high financial cost of treatment and the personal cost in terms of emotional suffering and disruption to daily living for anxious children and their families, there is a strong case for developing methods to prevent the development of childhood anxiety disorders. In keeping with the recognition of the importance of prevention of mental health problems generally, there has been a recent increase in efforts to develop effective methods of preventing anxiety disorders in children. To date, most universal strategies that target entire populations have focused upon the enhancement of mental health generally, rather than focusing specifically upon the prevention of anxiety problems. However, several programs have been investigated that can be described as 'selective' preventive interventions. These aim to target sub-

groups or individuals who are assumed to have a high lifetime or imminent risk of developing a problem as the result of exposure to some biological, psychological, or social risk factor(s). Examples of selective prevention strategies include those aimed at children whose parents have been divorced, those who are making the transition to high school, and children undergoing painful medical and dental procedures. Researchers have started to examine the possibility of intervening with children who manifest early childhood temperaments of behavioral inhibition or disrupted attachment relationships, in order to determine whether it is possible to reduce the probability of the development of anxiety problems. However, these studies are in their early stages and there is no clear indication as to their efficacy.

Some tentative data does exist to suggest that 'indicated' prevention strategies offers promise in the prevention of childhood anxiety disorders. The Queensland Early Intervention and Prevention of Anxiety Project (Dadds et al. 1997) represents an 'indicated' prevention program that targeted high-risk children demonstrating minimal but detectable symptoms of anxiety. This intervention made use of a one-term program that taught anxiety management and coping skills to elementary school children and their parents. At two-year follow-up, significantly fewer children who participated in the preventive intervention met diagnostic criteria for an anxiety disorder compared to those who did not take part. Importantly, the study also demonstrated that children who showed mild, nonclinical symptoms were at particular risk of developing a full-blown clinical anxiety disorder over the following two-year period if they did not receive the intervention.

While prevention research and practice remains in its infancy, a number of issues concerning prevention research warrant discussion. First, methods of childhood anxiety prevention may be child, parent, or environmentally based, and should be derived from the plethora of information regarding etiological factors and effective treatment strategies. Second, preventative efforts must be aimed in accordance with the developmental level of the child, as different risk factors may impact upon a child at different developmental stages. Third, the importance of multilevel intervention must be recognized, as effective prevention must go beyond the acquisition of personal skills and must include environmental and community change.

See also: Anxiety and Anxiety Disorders; Attachment Theory: Psychological; Genetic Studies of Personality; Parenting: Attitudes and Beliefs; Personality Development and Temperament; Shyness and Behavioral Inhibition; Temperament and Human Development; Temperament: Familial Analysis and Genetic Aspects

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders-IV*. American Psychiatric Association, Washington, DC

Benjamin R S, Costello E J, Warren M 1990 Anxiety disorders in a pediatric sample. *Journal of Anxiety Disorders* **4**: 293–316

Birmaher B, Khetarpal S, Brent D, Cully M, Balach L, Kaufman J, Neer S M 1997 The screen for child anxiety related emotional disorders (SCARED): Scale construction and psychometric characteristics. *Journal of the American Academy of Child and Adolescent Psychiatry* **36**: 545–53

Compas B E, Malcarne V L, Fondacaro K M 1988 Coping with stressful events in older children and young adolescents. *Journal of Consulting and Clinical Psychology* **56**: 405–11

Dadds M R, Spence S H, Holland D E, Barrett P M, Laurens K R 1997 Prevention and early intervention for anxiety disorders: A controlled trial. *Journal of Consulting and Clinical Psychology* **65**: 627–35

Dumas J E, La Freniere P, Serketich W J 1995 'Balance of power': A transactional analysis of control in mother–child dyads involving socially competent, aggressive, and anxious children. *Journal of Abnormal Psychology* **104**: 104–13

Fox N A, Calkins S D 1993 Pathways to aggression and social withdrawal: Interactions among temperament, attachment and regulation. In: Rubin K H, Asendorpf J B (eds.) *Social Withdrawal, Inhibition and Shyness in Childhood*. Lawrence Erlbaum, Hillsdale, NJ, pp. 81–100

Hirshfeld D R, Biederman J, Brody L, Faraone S V, Rosenbaum J F 1997a Associations between expressed emotion and child behavioral inhibition and psychopathology: A pilot study. *Journal of the Academy of Child and Adolescent Psychiatry* **36**: 205–13

Hirshfeld D R, Biederman J, Brody L, Faraone S V, Rosenbaum J F 1997b Expressed emotion towards children with behavioral inhibition: Association with maternal anxiety disorder. *Journal of the Academy of Child and Adolescent Psychiatry* **36**: 910–17

Kagan J 1997 Temperament and the reactions to unfamiliarity. *Child Development* **68**: 139–43

Kashani J H, Orvaschel H 1990 A community study of anxiety in children and adolescents. *American Journal of Psychiatry* **147**: 313–18

Kendall P C 1994 Treating anxiety disorders in children: Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology* **62**: 100–10

Krohne H W, Hock M 1991 Relationships between restrictive mother–child interactions and anxiety of the child. *Anxiety Research* **4**: 109–24

Last C G, Perrin S, Hersen M, Kazdin A E 1992 *DSM-III-R* anxiety disorders in children: Sociodemographic and clinical characteristics. *Journal of the American Academy of Adolescent Psychiatry* **31**: 1070–6

Lonigan C J, Phillips B M in press Temperamental influences on the development of anxiety disorders. In: Vasey M W, Dadds M R (eds.) *The Developmental Psychopathology of Anxiety*. Oxford University Press, New York

March J S, Leonard H L, Swedo W E 1995 Obsessive–compulsive disorder. In: March J S (ed.) *Anxiety Disorders in Children and Adolescents*. Guilford Press, New York, pp. 251–78

McFarlane A C 1987 Posttraumatic phenomena in a longitudinal study of children following a natural disaster. *Journal of the American Academy of Child and Adolescent Psychiatry* **26**: 764–9

Nachmias M, Gunmar M, Mangelsdorf S, Parritz R H, Buss K 1996 Behavioral inhibition and stress reactivity: The moderating role of attachment security. *Child Development* **67**: 508–22

Rapee R M in press The develoment of generalised anxiety. In: Vasey M W, Dadds M R (eds.) *The Developmental Psychopathology of Anxiety*. Oxford University Press, New York

Reynolds C R, Richmond B O 1978 What I think and feel—a revised measure of children's manifest anxiety. *Journal of Abnormal Child Psychology* **6**: 271–80

Silverman W K, Albano A 1996 *Anxiety Disorders Interview Schedule for DSM-IV*. The Psychological Corporation/Harcourt Brace & Company/Graywind Publications, San Antonio, TX

Spence S H 1997 Structure of anxiety symptoms among children: A confirmatory factor-analytic study. *Journal of Abnormal Psychology* **106**: 280–97

Spielberger C D 1973 *Manual for the State-Trait Anxiety Inventory for Children*. Consulting Psychologists Press, Palo Alto, CA

Thapar A, McGuffin P 1995 Are anxiety symptoms in childhood heritable? *Journal of Child Psychology and Psychiatry* **36**: 439–47

Warren S L, Huston L, Egeland B, Sroufe L A 1997 Child and adolescent anxiety disorders and early attachment. *Journal of the American Academy of Child and Adolescent Psychiatry* **36**: 637–44

White K S, Bruce S E, Farrell A D, Kliewer W 1998 Impact of exposure to community violence on anxiety: A longitudinal study of family social support as a protective factor for urban children. *Journal of Child and Family Studies* **7**: 187–203

C. L. Donovan

Apartheid

An Afrikaans word meaning 'separateness,' apartheid was the name given to the legislative program enacted by the National Party that ruled South Africa from 1948 to 1994. Created to entrench domination by the white minority, apartheid represented a refinement of race-based policies implemented by successive colonial governments for three centuries. Apartheid was developed in the 1960s as Grand Apartheid, the never-realized vision of creating separate, independent states for different ethnic groups. Increasingly disruptive internal opposition to apartheid led, in 1960, to the banning of liberation movements that continued their campaigns to overthrow the apartheid state from exile and underground. The failure of apartheid's Total Strategy to stem opposition, the resurgence of popular protest in South Africa in the 1980s, and the ending of the Cold War led to the dismantling of legal apartheid and the nation's first democratic elections in 1994. Beyond its impact in Southern Africa, the struggle against apartheid fostered crucial developments in the evolution of international human rights, uniting the fractious United Nations Security Council to intervene, for the first time, in the domestic race relations of a sovereign state.

1. Apartheid's Antecedents

Formal apartheid originated in the legal and social structures that followed the settlement of southern Africa by Europeans in the mid-seventeenth century. After agents of the Dutch East India Company created the first permanent European settlement in modern-day Cape Town in 1652, they soon transformed their supply station into a base for European expansion. Over the next 250 years, the settlers consolidated their control of land and livestock, conquering the indigenous Khoisan and Bantu peoples through war and disease. The British colonial authorities who administered the Cape Colony from 1806 perpetuated the Dutch policies of segregation and discriminatory legal standards. While the British abolished slavery and granted putative equality of political rights to the Khoisan in the Cape Colony in 1828, they denied meaningful political representation to the indigenous population who came increasingly under colonial control.

The discovery of vast diamond and gold deposits in the interior of South Africa in the late 1800s spurred colonial settlement and British interest in areas that had been declared independent republics by Afrikaans-speaking settlers of Dutch, French, and German descent. It would take the devastating South African War (Anglo–Boer War) of 1899–1902, during which a British force of 200 000 troops stamped its authority on the independent republics, to settle the territorial claims of the Boer, British, and local people. While the 1902 peace treaty entrenched the property rights of white settlers, the Union of South Africa, established in 1910, reasserted the race-based policies of the formerly independent colonies, denying the franchise to non-whites in all but the Cape Colony. The 1913 Natives Land Act delineated reserves for indigenous people, eventually barring them from owning land in 87% of the country, and forcing thousands into the labor market by banning share-cropping on white land.

During the 1930s, political leaders became increasingly successful in stimulating the nationalist ambitions of the Afrikaans-speaking population. Heavily influenced by the strict, Calvinist, Dutch Reform Church, the Afrikaners largely constituted a white underclass of small farmers and workers. The Reunited National Party's promise to overturn British domination, campaign against the South African government's military alliance with Great Britain, and an economic and social platform that focused on preserving white privilege and segregation earned them a surprise victory in the 1948 general election. Some ideologues within the National Party (NP), many influenced by Nazi ideology, sought complete separation between whites and all other people. Complete separation, however, would have removed the source of abundant cheap labor on which white wealth relied, undermining the National Party's plans

for uplifting Afrikaners economically. Apartheid evolved over the next half-century as an attempt to resolve the dilemma of how the outnumbered whites could exploit black labor while maintaining political control and racial separation.

2. *Evolution of Apartheid Policy and Protest*

Upon assumption of power in 1948, the NP quickly passed a series of laws aimed at reversing and preventing even the smallest steps toward integration. The Prohibition of Mixed Marriages Act and Immorality Act of 1949 banned miscegenation, while the Group Areas Act of 1950 extended the powers created by previous land acts to segregate urban residential and commercial areas. The Population Registration Act of 1950 underpinned apartheid legislation by recording the racial category of all South Africans from birth. The NP also ensured their continued legislative success by removing the descendents of Khoisan and mixed-race citizens from the voters roll in the Western Cape in 1956. Emphasizing the Afrikaners' independence, the NP led South Africa out of the British Commonwealth in 1961. On the economic front, successive NP governments channeled state resources into Afrikaner empowerment schemes and set aside jobs in the burgeoning bureaucracy for their supporters. The *Broederbond*, a shadowy network of Afrikaner intellectuals, politicians, clergymen, and other powerful figures, took control of most of the important positions in the public sector.

2.1 *Grand Apartheid*

After the initial flurry of apartheid legislation failed to halt the influx of Africans into cities, the government sought to implement Grand Apartheid. Advocates of this extreme form of separation envisioned transforming the least productive 13% of the country's land into homelands—independent nation states in which blacks would undertake 'separate development.' Black South Africans would be granted citizenship in one of the homelands, each representing a supposed tribal group. Stripped of their South African citizenship, they would only be allowed to enter white South African areas to work. Despite the rhetoric that described the homelands as part of a constellation of independent states in Southern Africa, the South African government would maintain political control through largesse and the right to appoint political representatives.

Four homelands eventually did opt for 'independence' between 1976 and 1980, but the concept of separate development neither mollified internal dissent nor acquired international legitimacy. With consolidation plans for the homelands scuttled by white farmers unwilling to part with fertile land, in reality the homelands became a patchwork of overgrazed lands that could not sustain the populations imposed on them. The homelands policy, and the enforcement of other segregation legislation, eventually led to the forced removal of 2 million people.

2.2 *Opposition to Apartheid*

The imposition of formal apartheid from 1948 reinvigorated internal resistance to the government and domestic and international support for what was increasingly seen as part of the broader anticolonial struggle sweeping Africa. A coalition led by the African National Congress (ANC) waged a campaign of civil disobedience in 1952 and proposed a dramatic alternative to apartheid in the 1955 Freedom Charter. When the state responded with arrests and repression, banning liberation movements in 1960, the ANC and other groups organized armed resistance from exile and underground. Brutal and effective police action, and the economic boom that followed the increase in commodity prices in the 1960s and the sharp rise in the price of gold in the early 1970s, allowed the state to crush internal opposition. At the same time, by portraying the liberation struggle as linked to a global communist conspiracy that might threaten access to South Africa's strategically important minerals and sea lanes, the apartheid state was able to win important, if wary, political support and security assistance from the West and neighboring countries under colonial domination or heavy Western influence.

In the mid-1970s, rising repression encountered a resurgence of activism, especially among young proponents of the black consciousness philosophy. Challenged by a wave of worker action in 1973, the state provoked international outrage with the 1976 killing of student protestors in Soweto and an ensuing clampdown. With the overthrow of colonial governments in Mozambique in 1975 and Zimbabwe in 1980, the South African government also faced a renewed threat from external guerilla armies who could use neighboring countries as rear bases. These international developments, coupled with growing worldwide isolation of the apartheid government, forced a further evolution of apartheid policy.

2.3 *Total Strategy*

Known as the Total Strategy, this new policy attempted to preserve white control in South Africa by making cosmetic changes in apartheid policies, co-opting other racial minority groups, and winning political co-operation from neighboring countries, while increasing the repressive might of the state. The

government implemented a tricameral parliamentary system that created token parliamentary houses for Indian and mixed-race representation in 1984, repealed the Mixed Marriages Act in 1985, and abolished the hated pass laws in 1986. Building on the outward policy of the 1970s, the government also signed a non-aggression pact with Mozambique in 1984. At the same time, the Total Strategy channeled more resources into covert operations to undermine neighboring black governments and unleashed the armed forces to engage in cross-border raids and political repression.

Like earlier apartheid policies, the Total Strategy failed to stem opposition or win international favor. Instead of deflating protest, the tricameral parliament proposals galvanized internal opposition under the banner of the United Democratic Front, an umbrella of hundreds of diverse opposition groups. The structure of grass-roots leadership, with subterranean ties to the banned ANC, coordinated national protests with techniques that confounded the apartheid state. From 1984, an increasingly devastating spiral of unrest, repression, and international condemnation took hold.

2.4 International Intervention

Personified in the campaign to release Nelson Mandela, the rising tide of international protest against apartheid became the first major international movement that asserted the prerogative of all people and international organizations to protest against human rights violations in a sovereign state. By imposing mandatory sanctions in 1979, the United Nations Security Council for the first time overrode traditional international legal norms that held problems of national integration to be an exclusively domestic matter. Spurred by a grass-roots campaign, the US Congress imposed sanctions against South Africa in 1986, overturning a presidential veto. These actions, coupled with the decisions by private banks to deny access to new capital in the face of rising political uncertainty in the mid-1980s, weakened the apartheid government's resolve, though to what extent their impact was material, not just symbolic, remains disputed.

3. Apartheid's Negotiated Demise

The winding down of the Cold War in the late 1980s undercut both the opposition's financial support base and the ability of the South African government to defend its repressive actions as an anti-communist crusade. The resulting stalemate between an increasingly powerful, but militarily diminished, opposition movement and the highly militarized, but isolated, state ended with the unbanning of opposition parties in 1990 and the initiation of political negotiations. The cornerstones of apartheid legislation, the Population Registration Act and Group Areas Act, were repealed the following year. In April 1994, after a series of complex negotiations, virtually the entire South African adult population voted in a peaceful election, choosing the ANC to become the ruling party.

Apartheid's demise, by political rather than military means at the end of history's most violent century, confounded many experts who believed that the divisions were too deep and the economy too weak to support a compromise. Yet South Africans, with very little help from outsiders, brought to an end one of humanity's longest-running dramas; the struggle against colonialism, segregation, and other forms of institutionalized racism in Africa. Even after the abolition of legal apartheid, the huge economic and social disparities that it created will persist for generations in South Africa. Post-apartheid South Africa represents one of the world's most prominent attempts to meet what the British philosopher, Sir Isaiah Berlin, termed the greatest challenge facing humanity: building the political frameworks to manage cultural diversity. Few nations have had to deal with as profound cultural, racial, and religious differences, compounded by a history of race-based political oppression and economic deprivation. If South Africans can continue to govern themselves peacefully, their example could potentially have a greater political impact on the global spread of democratic values in the twenty-first century than the fight against apartheid had on the spread of human rights in the twentieth century.

See also: African Legal Systems; African Studies: History; African Studies: Politics; Ethnic Conflict, Geography of; Ethnic Conflicts; Race and the Law; Race: History of the Concept; Racial Relations; Racism, History of; Racism, Sociology of; Residential Concentration/Segregation, Demographic Effects of; Social Mobility, History of; Southern Africa: Socio-cultural Aspects

Bibliography

Adam H, Giliomee H 1979 *Ethnic Power Mobilized: can South Africa change?* Yale University Press, New Haven, CT

De Villiers M 1970 *White Tribe Dreaming: Apartheid's Bitter Roots as Witnessed by Eight Generations of an Afrikaner Family*. Penguin, New York

Karis T, Carter G M 1972 *From Protest to Challenge: A Documentary History of African Politics in South Africa 1882–1990*. Hoover Institution Press, Stanford, CA

Karis T, Gerhart G M 1991 *From Protest to Challenge: A Documentary History of African Politics in South Africa 1882–1990*. Hoover Institution Press, Stanford, CA
Mandela N 1994 *Long Walk to Freedom: The Autobiography of Nelson Mandela*. 1st edn. Little Brown, Boston
O'Meara D 1996 *Forty Lost Years: The Apartheid State and Politics of the National Party, 1948–94*. Ravan Press, Randburg, South Africa
Posel D 1991 *The Making of Apartheid 1948–61: Conflict and Compromise*. Oxford University Press, Oxford, UK
Reader's Digest 1988 *Illustrated History of South Africa: The Real Story*. Reader's Digest Association South Africa, Cape Town, South Africa
Thompson L 1995 *A History of South Africa*. rev. edn. Yale University Press, New Haven, CT
Waldmeir P 1997 *Anatomy of a Miracle*. 1st edn. Norton, New York

A. Levine and J. J. Stremlau

Apathy

Apathy stems from the ancient Greek *apathies*, which means 'lack of feeling.' Apathy plays an important role in theories of democracy that stress citizens' involvement in public affairs (see *Democratic Theory*; also see *Democracy*). Ancient Athenians' praise for attentive citizens and condemnation of apathetic ones established a tradition in democratic theory. Apathy saps public spiritedness, which is why it is thought to be one indicator of waning 'social capital' in modern societies (see *Social Capital*). Apathy also inhibits citizens' 'cognitive mobilization,' which is an important political resource (Inglehart 1997). Small wonder that apathy among citizens of democratic countries worries politicians, pundits, and professors.

Apathy means political indifference; its opposite is political interest. Apathy/interest entails the expression of 'curiosity' about public affairs (Gabriel and van Deth 1995). Apathy/interest is an attitude, not an absence of activity. Apathy does not mean 'nonvoting,' for example, for people do not vote for many reasons (see *Voting: Turnout*), most having nothing to do with lack of interest in public affairs. Before passage of the 1965 Voting Rights Act in the US, for example, African-Americans living in the South were prevented from voting by several means, including intimidation and violence (see *Race Relations in the United States, Politics of*). It would be wrong to equate southern blacks' absence from the voting booth as indicating apathy.

Although apathy was once equated with alleged 'pathologies' such as alienation, hostility, isolation, and suspicion (Campbell 1962), that is no longer true. Psychological involvement in public affairs is one of the most important political dispositions a person has. Citizens who pay attention to public affairs are different political actors than those who are indifferent (Almond and Verba 1963, Bennett 1986, Converse 1972, Gabriel and van Deth 1995, Inglehart 1997, van Deth 1990, Verba et al. 1995).

How should apathy/interest be measured? Prior to the advent of scientific public opinion surveys in the 1930s (see *Polling*), generalizing about mass publics was risky, although it could be done well (Lippmann 1925). Even after public opinion polling emerged, estimates of mass publics' apathy/interest were not problem free. Although some researchers think it is possible to use a single item to measure apathy/interest, it is best to use multiple items to tap this attitude. It is preferable to avoid combining measures of 'subjective political interest'—the topic of this article—with indicators of political behavior, such as talking about politics with family and friends.

In their study of political attitudes in five western democracies, Almond and Verba (1963; see also *Civic Culture*) developed a multi-item indicator of subjective political interest. They combined a measure of general political interest with another, tapping attention to election campaigns to form 'the civic cognition.' Following Almond and Verba, Bennett (1986) constructs 'the Political Apathy Index,' which is a combination of general political interest and attentiveness to election campaigns. Since both items have had the same wording since 1968, and had virtually the same location on the University of Michigan's biennial National Election Studies since 1978, the Political Apathy Index provides an excellent vehicle for exploring Americans' interest in public affairs over more than 20 years. Multiple item measures of Europeans' political interest do not exist over a very long period (Gabriel and van Deth 1995, van Deth 1990).

What do researchers know about democratic citizens' interest in public affairs? Except for short-term emergencies or catastrophes, most people are not very interested in public affairs (Bennett 1986). Most Americans normally express only a 'lukewarm' interest in public affairs. Nevertheless, Americans are more politically interested than most West Europeans, probably because educational attainment is higher in the US (Powell 1986). There is little evidence of growing political interest in most European nations in recent years (Gabriel and van Deth 1995).

Several factors affect psychological involvement in public affairs. Education strongly shapes interest in the US and elsewhere (Bennett 1986, Converse 1972, Gabriel and van Deth 1995, Nie et al. 1996, van Deth 1990). Formal schooling imparts the intellectual skills and motivation needed to pay heed to public affairs, and exposure to higher education often ensconces people in social niches that encourage and reward political attentiveness.

Location in a social structure affects apathy/interest. It is easier to pay attention to public affairs if one's occupation and lifestyle place the individual at or near the center of a society. Some professions

encourage political interest. Members of the legal profession, for example, tend to be very attentive to government and public affairs. Those living on the periphery of society—by virtue of their job, race, religion, or ethnicity—are less inclined to be politically interested.

Age also affects political attentiveness. Young people tend to be less politically attentive than their elders, mostly because they are distracted by the 'start-up phenomenon,' which involves completing school, getting started in a job or career, searching for a life-mate, and even being socially mobile. Political interest requires the capacity to concentrate on issues and events outside one's immediate concern, and most young people tend to focus on personal matters. In this view, the passage of time and the assumption of mature adult roles produced a steady increase in political interest over the life cycle, an increase that would be partly reversed as people reached extreme old age.

Some evidence questions the life-cycle explanation for the relation between youth and apathy. American men who came of age during World War II were especially politicized, and they remained atypically interested in politics over the next four decades (Bennett 1986). Similar evidence occurs among Early Baby Boomers in the US, whose male members were exposed to the Vietnam-era draft, and Late Baby Boomers, who were born too late for the draft (Bennett and Bennett 1990). On the other hand, younger citizens of West Germany expressed more political interest than older persons in 1994, perhaps because the latter were still haunted by the Nazi past (Bennett et al. 1996). (When looking at the relation between most social factors and a disposition such as apathy, one should take note of a nation's history, culture, and institutions.)

Another contradiction to the life-cycle thesis is the emergence in the US of 'Generation X,' or persons born between 1965 and 1978, who have been particularly apathetic (see *Generations: Political*). Young Americans today (Bennett 1997), and to some degree young Europeans (Gabriel and van Deth 1995), are less politically attentive than youth were a generation ago, and American young people show little inclination to become more politically interested as time goes by.

Certain political dispositions also affect apathy/interest. People who believe they have a moral obligation to be politically active are more likely to be attentive than those who lack a sense of 'civic duty' (Bennett 1986). In addition, strong adherents of a political party are more politically attentive than independents and 'apoliticals' (see *Party Identification*). Therefore, Generation Xers' tendency to avoid attachment to a political party has worrisome consequences for their political attentiveness. Other political attitudes, such as the belief that one is a competent citizen and that political activity is worthwhile, or political efficacy (see *Efficacy: Political*), are tied to political interest, but scholars cannot tell which causes which.

What difference does it make if citizens are politically indifferent? Apathy violates the assumption that a healthy democracy requires attentive citizens. There are demonstrable consequences of apathy that trouble many (e.g., DeLuca 1995), but not everyone (Berelson et al. 1954). Not only is psychological involvement in public affairs a powerful goad to political participation (see *Participation: Political*), interest also affects exposure to the mass media and political information (Bennett 1986). Finally, although the relationship is complicated, interest also affects political sophistication (Converse 1972). If one is interested in the grassroots foundations of democracy, there are ample grounds for concern with apathy/interest.

Scholars stand on the threshold of several discoveries about apathy/interest. It now appears that apathy/interest has at least two related components: an 'ego involvement' dimension and a 'general subjective interest' dimension. If true, the disposition is more complex than researchers previously assumed. The future may also witness the passing of the well-documented fact that women have been less politically attentive than men (Bennett and Bennett 1989, Inglehart 1981). As older women who were raised to believe that 'politics is a man's business' pass from the electorate, and especially if new birth cohorts do not subscribe to traditional norms, the next century may see an end of gender differences in political interest.

Additional research will also improve people's understanding of apathy's causes and consequences. Previous scholarship was limited by both the way in which apathy/interest was conceptualized and measured and by the research tools used to study the disposition's causes and effects. As new means to measure the phenomenon emerge, and as more sophisticated data analysis procedures are utilized, future scholarship will sharpen and refine what is known about apathy/interest in the US and elsewhere. It will help greatly, for example, to understand better the complex relationships between apathy/interest and other political dispositions such as party identification, efficacy, and sense of civic duty. The nexus between apathy/interest and exposure to the mass media also needs to be better understood. Critics allege that the American news media's style of political coverage saps people's interest in public affairs. Researchers need to see if this is true in other polities, as well as in the US. New, more sophisticated comparative studies can shed useful light on the association between a nation's political culture and its citizens' attention to public affairs (see *Political Culture*).

Finally, Western nations are experiencing a renewal of civic education. Scholars do not understand very well how to motivate more young people to become politically attentive, but the will to accomplish that goal seems to be emerging. The struggle to educate the

young to the norms of citizenship will call for blending theoretical and applied research (see *Socialization: Political*). As ancient Greek democrats appreciated, encouraging political interest among democratic citizens is a worthy enterprise.

See also: Attitudes and Behavior; Participation: Political; Party Systems; Public Opinion: Political Aspects; Voting: Class; Voting: Compulsory; Voting: Issue; Voting, Sociology of; Voting: Turnout

Bibliography

Almond G A, Verba S 1963 *The Civic Culture*. Princeton University Press, Princeton, NJ

Bennett L L M, Bennett S E 1989 Enduring gender differences in political interest. *American Politics Quarterly* **17**: 105–22

Bennett L L M, Bennett S E 1990 *Living With Leviathan: Americans Coming to Terms With Big Government*. University Press of Kansas, Lawrence, KS

Bennett S E 1986 *Apathy in America, 1960–1984*. Transnational, Dobbs Ferry, NY

Bennett S E 1997 Why young Americans hate politics, and what we should do about it. *PS: Political Science and Politics* **30**: 47–53

Bennett S E, Flickinger R S, Baker J R, Rhine S L, Bennett L L M 1996 Citizens' knowledge of foreign affairs. *Harvard International Journal of Press/Politics* **1**(2): 10–29

Berelson B, Lazarsfeld P, McPhee W N 1954 *Voting*. University of Chicago Press, Chicago

Campbell A 1962 The passive citizen. *Acta Sociologica* **6**(1–2): 9–21

Converse P E 1972 Change in the American electorate. In: Campbell A, Converse P E (eds.) *The Human Meaning of Social Change*. Russell Sage, New York

DeLuca T 1995 *The Two Faces of Apathy*. Temple University Press, Philadelphia, PA

Gabriel O W, van Deth J W 1995 Political interest. In: van Deth J W, Scarbrough E (eds.) *The Impact of Values*. Oxford University Press, New York

Inglehart M L 1981 Political interest in West European women. *Comparative Political Studies* **14**: 299–326

Inglehart R 1997 *Modernization and Postmodernization*. Princeton University Press, Princeton, NJ

Lippmann W 1925 *The Phantom Public*. Harcourt Brace, New York

Neuman W R 1986 *The Paradox of Mass Politics*. Harvard University Press, Cambridge, MA

Nie N H, Junn J, Stehlik-Barry K 1996 *Education and Democratic Citizenship in America*. University of Chicago Press, Chicago

Powell G B 1986 American voter turnout in comparative perspective. *American Political Science Review* **80**: 17–43

van Deth J W 1990 Interest in politics. In: Jennings M K, van Deth J W et al. (eds.) *Continuities in Political Action*. de Gruyter, Berlin

Verba S, Lehman Schlozman K, Brady H E 1995 *Voice and Equality*. Harvard University Press, Cambridge, MA

S. E. Bennett

Aphasia

1. *Definition*

The term 'aphasia' refers to disorders of language following diseases of the brain. As is discussed in other articles in this encyclopedia language is a distinctly human symbol system that relates a number of different types of forms (words, words formed from other words, sentences, discourse, etc.) to various aspects of meaning (objects, properties of objects, actions, events, causes of events, temporal order of events, etc.). The forms of language and their associated meanings are activated in the processes of speaking, understanding speech, reading, and writing. The processes whereby these forms are activated are largely unconscious, obligatory once initiated, fast, and usually quite accurate. Disturbances of the forms of the language code and their connections to their associated meanings, and of the processes that activate these representations in these ordinary tasks of language use, constitute aphasic disturbances. By convention, the term 'aphasia' does not refer to disturbances that affect the functions to which language processing is put. Lying (even transparent, ineffectual lying) is not considered a form of aphasia, nor is the garrulousness of old age or the incoherence of schizophrenia.

Language consists of a complicated system of representations, and its processing is equally complicated, as described in other entries in this encyclopedia. Just the representation of the minimal linguistically relevant elements of sound—phonemes—and the processing involved in recognizing and producing these units constitute a highly complex domain of functioning. When all the levels of language and their interactions are considered, language processing is seen to be enormously complex. Aphasic disturbances would therefore be expected to be equally complex. Researchers are slowly describing the very considerable range of these disorders.

2. *History of the Field: The Classic Aphasic Syndromes, and Alternative Views*

However, the first modern scientific descriptions of aphasia were quite modest with respect to the descriptions of language processing that they contained. These descriptions were made by neurologists in the second half of the nineteenth century. Though modest with respect to the sophistication of the descriptions of language, these studies laid important foundations for the scope of work on aphasia and for the neural basis for language processing, which has always been a closely associated topic. The first of these late nineteenth century descriptions was that by Broca (1861), who described a patient, Leborgne, with a severe speech output disturbance. Leborgne's speech

Table 1
Classical aphasic syndromes

Syndrome	Clinical manifestations	Postulated deficit	Classical lesion location
Broca's aphasia	Major disturbance in speech production with sparse, halting speech, often misarticulated, frequently missing function words and bound morphemes	Disturbances in the speech planning and production mechanisms	Posterior aspects of the 3rd frontal convolution (Broca's area)
Wernicke's aphasia	Major disturbance in auditory comprehension; fluent speech with disturbances of the sounds and structures of words (phonemic, morphological, and semantic paraphasias)	Disturbances of the permanent representations of the sound structures of words	Posterior half of the first temporal gyrus and possibly adjacent cortex (Wernicke's area)
Pure motor speech disorder	Disturbance of articulation Apraxia of speech, dysarthria, anarthria, aphemia	Disturbance of articulatory mechanisms	Outflow tracts from motor cortex
Pure word deafness	Disturbance of spoken word comprehension	Failure to access spoken words	Input tracts from auditory system to Wernicke's area
Transcortical motor aphasia	Disturbance of spontaneous speech similar to Broca's aphasia with relatively preserved repetition	Disconnection between conceptual representations of words and sentences and the motor speech production system	White matter tracts deep to Broca's area connecting it to parietal lobe
Transcortical sensory aphasia	Disturbance in single word comprehension with relatively intact repetition	Disturbance in activation of word meanings despite normal recognition of auditorily presented words	White matter tracts connecting parietal lobe to temporal lobe or portions of inferior parietal lobe
Conduction aphasia	Disturbance of repetition and spontaneous speech (phonemic paraphasias)	Disconnection between the sound patterns of words and the speech production mechanism	Lesion in the arcuate fasciculus and/or corticocortical connections between Wernicke's and Broca's areas
Anomic aphasia	Disturbance in the production of single words, most marked for common nouns with variable comprehension problems	Disturbances of concepts and/or the sound patterns of words	Inferior parietal lobe or connections between parietal lobe and temporal lobe; can follow many lesions
Global aphasia	Major disturbance in all language functions	Disruption of all language processing components	Large portion of the perisylvian association cortex
Isolation of the language zone	Disturbance of both spontaneous speech (similar to Broca's aphasia) and comprehension, with some preservation of repetition	Disconnection between concepts and both representations of word sounds and the speech production mechanism	Cortex just outside the perisylvian association cortex

was limited to the monosyllable 'tan.' Broca described Leborgne's ability to understand spoken language and to express himself through gestures and facial expressions, as well as his understanding of non-verbal communication, as being normal. Broca claimed that Leborgne had lost 'the faculty of articulate speech.' Leborgne's brain contained a lesion whose center was in the posterior portion of the inferior frontal convolution of the left hemisphere, an area of advanced cortex just adjacent to the motor cortex. Broca related the most severe part of the lesion to the expressive language impairment. This area became known as 'Broca's area.' Broca argued that it was the neural site of the mechanism involved in speech production.

In a second very influential paper, Wernicke (1874) described a patient with a speech disturbance that was very different from that seen in Leborgne. Wernicke's patient was fluent; however, her speech contained words with sound errors, other errors of word forms, and words that were semantically inappropriate. Also unlike Leborgne, Wernicke's patient did not understand spoken language. Wernicke related the two impairments—the one of speech production and the one of comprehension—by arguing that the patient had sustained damage to 'the storehouse of auditory word forms.' The lesion in Wernicke's case was unknown, but a lesion in a similar case was the area of the brain next to the primary auditory receptive area, which came to be known as Wernicke's area.

These pioneering descriptions of aphasic patients set the tone for much subsequent work. First, they focused the field on impairments of the usual modalities of language—producing and producing speech and later on, reading and writing. This seems like an obvious area for aphasiology to be concerned with, but not all researchers of the period agreed with this focus. In another famous paper, the influential British neurologist John Hughlings Jackson (1878) described a patient, a carpenter, who was mute but who mustered up the capacity to say 'Master's' in response to his son's question about where his tools were. Jackson's poignant comments convey his emphasis on the conditions that provoke speech, rather than on the form of the speech itself:

> 'The father had left work; would never return to it; was away from home; his son was on a visit, and the question was directly put to the patient. Anyone who saw the abject poverty the poor man's family lived in would admit that these tools were of immense value to them. Hence we have to consider as regards this and other occasional utterances the strength of the accompanying emotional state' (Jackson 1878, p. 181)

Jackson sought a description of language use as a function of motivational and intellectual states, and tried to describe aphasic disturbances of language in relationship to the factors that drive language production and make for depth of comprehension. Broca, Wernicke, and the researchers who followed, focused aphasiology on patients' everyday language use under what was thought to be normal emotional and motivational circumstances.

These and related subsequent papers tended to describe language impairments in terms of the entirety of language-related tasks—speaking, comprehending, etc.—with only passing regard for the details of the language forms that were impaired within a task. A patient's deficit was typically described in terms of whether such an entire function was normal or not, and in terms of whether one such function was more impaired than another was. Here, for instance, is the description of Broca's aphasia by two twentieth-century neurologists whose work follows in this tradition:

> 'The language output of Broca's aphasia can be described as nonfluent … Comprehension of spoken language is much better than speech but varies, being completely normal in some cases and moderately disturbed in others. (Benson and Geschwind 1971, p. 7).'

The one level of language that descriptions did tend to concentrate on was the level of words. For instance, many patients with language disturbances that are classified as Wernicke's aphasia make many errors in word formation, substituting one type of word ending for another, but Wernicke's description of the impairment in his patient dealt only with the storehouse for individual words, not the locus of the word formation process. Put in other words, though the early work on aphasia emphasized the usual tasks of language use, this work, and research that followed in this tradition, did not describe these impairments systematically in either linguistic or psycholinguistic terms.

This approach to aphasia led to the recognition of some 10 aphasic 'syndromes.' These are listed, along with their proposed neural bases, in Table 1.

3. *Psycholinguistic Approaches to Aphasia*

As noted in Sect. 2, these classical syndromes do not give a complete account of the range and specificity of aphasic impairments. More recent descriptions of aphasia add many details to the linguistic and psycholinguistic descriptions of these disorders. It is impossible to review all these impairments in the space of a short article, but a few examples will illustrate these results. For instance, at the first step of speech processing—converting the sound waveform into linguistically relevant units of sound—researchers have described specific disturbances affecting the ability to recognize subsets of phonemes, such as vowels, consonants, stop consonants, fricatives, nasals, etc. (Saffran et al. 1976). In the area of word production, patients have been described with selective impairments of the ability to produce the words for items in particular semantic categories, such as fruits and vegetables, but sparing animals and man-made

tools (Hart et al. 1985), selective impairments affecting the ability to produce nouns and verbs (Damasio and Tranel 1993), and other highly restricted deficits. In the area of reading, patients have been found to have impairments of the ability to sound out novel written stimuli using letter-sound correspondences but retained abilities to read familiar words, and vice versa (Marshall et al. 1980, Patterson et al. 1985).

Linguistic theory provides a basis for exploring the nature of aphasic disorders, by providing evidence for different types of linguistic representations. Models of the psychological processes involved in activating linguistic representations suggest other possible loci of impairment. Many researchers have worked backwards from clinically observed phenomena, developing or modifying theories of language structure and processing on the basis of the disorders seen in aphasic patients. For instance, Shallice and Warrington (1977) challenged the then-popular view that verbal short-term memory fed verbal long-term memory by documenting a patient with a severely reduced verbal short-term memory capacity whose performance on tests of verbal long-term memory was normal. Ullman and his colleagues (Ullman et al. 1997a, 1997b) have argued that the impairments seen in patients with Alzheimer's and Huntington's Disease provide support for a view of language that distinguishes between regular, rule-based, word formation processes, and irregularly formed complex words that are listed in a mental dictionary. (For a discussion of this distinction and its broader implications for language and the mind, see Pinker 1999.) Characterizing aphasic disorders is an interactive, interdisciplinary, bootstrapping process that is presently in active evolution.

The psycholinguistic approach to aphasia is based upon a model of language structure and processing. Experts disagree about these models. The greatest disagreements center on the issue of the extent to which linguistic representations are highly abstract structures that are produced and computed in comprehension tasks by rules (Chomsky 1995), as opposed to far less abstract representations that are processed largely by highly developed pattern associations (Rumelhart and McClelland 1986). If language is seen in the former perspective, many aphasic impairments are considered to be the result of damage to specific representations and/or processing operations. If language is seen in the second perspective, aphasic disturbances are largely conceptualized as resulting from reductions in the power of the associative system, due to loss of units, increases in noise, etc. Empirical study suggests that both specific impairments and loss of processing power are sources of aphasic disturbances. This can be illustrated in one area—disorders affecting syntactic processing in sentence comprehension.

Disorders of syntactically based comprehension affect the ability to extract the relationships between the meanings of words in a sentence that are determined by the syntactic structure of a sentence. For instance, in the sentence 'The dog that scratched the cat killed the mouse,' there is a sequence of words—the cat killed the mouse—which, in isolation, would mean that the cat killed the mouse. However, this is not what the sentence means, because of its syntactic structure. 'The cat' is the object of the verb 'scratched;' 'the dog' is the subject of the verb 'killed' and is the agent of that verb. Caplan and his colleagues have explored the nature of these disturbances (Caplan et al. 1985, 1996). They found that, in many hundred of aphasic patients, mean group performance deteriorated on sentences that were more syntactically complex and that more impaired groups of patients increasingly performed more poorly on sentences that were harder for the group overall. These patterns suggest that the availability of a processing resource that is used in syntactic comprehension is reduced to varying degrees in different patients. A second finding in their studies has been that individual patients can have selective impairments of syntactic comprehension, just as is the case in the other areas of language processing previously mentioned. Published cases have had difficulty constructing hierarchical syntactic structures, disturbances affecting reflexives or pronouns but not both, and other more subtle impairments of syntactic processing (Caplan and Hildebrandt 1988).

Overall, these studies suggest that a patient's aphasic impairment can be described in terms of a reduction in the processing resources needed for this function and disruption to specific operations. An unresolved question is whether the entire pattern of performance seen in these disorders can be attributed to just one of these types of impairments, as the two types of models previously outlined maintain. This may be possible, but the challenges in explaining all these aspects of these (and other) aphasic disorders within a model that either does not incorporate the idea of a processing resource limitation or one that does not recognize specific operations are considerable.

4. *Functional Consequences of Aphasic Impairments*

The focus of this article has thus far been on aphasic disturbances as impairments of the largely unconscious processes that activate the elements of language in the usual tasks of language use. The functional consequences of these disorders deserve a brief comment. Functional communication involving the language code occurs when people use language to accomplish specific goals—to inform others, to ask for information, to get things done, etc. There is no simple, one-to-one relationship between impairments of elements of the language code or of psycholinguistic processors, on the one hand, and abnormalities in performing language-related tasks and accomplishing the goals of language use, on the other. Patients adapt

to their language impairments in many ways, and some of these adaptations are remarkably effective at maintaining at least some aspects of functional communication. Conversely, patients with intact language processing mechanisms may fail to communicate effectively. Nevertheless, most patients who have disturbances of elements of the language code or psycholinguistic processors experience limitations in their functional communicative abilities. In general, as the intentions and motivations of the language user become more complex, functional communication is more and more affected by disturbances of the language code and its processors. Thus, though 'high-level' language-impaired patients may be able to function well in many settings, their language impairments can cause substantial functional limitations. The language code is a remarkably powerful code with respect to the semantic meanings it can encode and convey, and psycholinguistic processors are astonishingly fast and accurate. Without this code and the ability to use it quickly and accurately, one's functional communicative powers are limited, no matter how elaborate one's intentions and motives. This is the situation in which many patients who have disorders affecting the language code and the processors dedicated to its use find themselves.

5. *Concluding Comments*

This essay should not end on this negative note. Rather, it is important to appreciate that many aphasic patients make excellent recoveries, for a variety of reasons. The natural history of many aphasic impairments is for considerable improvement, especially those due to smaller or subcortical lesions. Though still in their infancy, modern approaches to rehabilitation for aphasia are developing a sounder scientific basis. Technological advances allow for more professionally guided home training using computers, improved augmentative communication devices, and other useful support mechanisms. Support groups for patients and their families and friends are increasing in number; these help patients adjust to the changes in their lives and remain socially active. Though aphasia deprives a person of an important function to a greater or lesser degree, reactions to aphasia are as important as the aphasia itself in determining functional outcome and many aphasic patients function in vital ways after their loss. Lecours et al. (1983) cite a patient described by the Soviet psychologist A. R. Luria, who continued to compose music after a stroke that left him very aphasic; some critics thought his work improved after his illness. Time, rehabilitation, support, and a positive attitude can allow many aphasic patients to be productive and happy.

See also: Speech Production, Neural Basis of; Speech Production, Psychology of; Syntactic Aspects of Language, Neural Basis of

Bibliography

Benson D F, Geschwind N 1971 Aphasia and related cortical disturbances. In: Baker A B, Baker L H (eds.) *Clinical Neurology*. Harper and Row, New York

Broca P 1861 Remarques sur le siège de la faculté de la parole articulée, suives d'une observation d'aphemie (perte de parole). *Bulletin de la Société d'Anatomie* **36**: 330–57

Caplan D, Hildebrandt N 1988 *Disorders of Syntactic Comprehension*. MIT Press (Bradford Books), Cambridge, MA

Caplan D, Baker C, Dehaut F 1985 Syntactic determinants of sentence comprehension in aphasia. *Cognition* **21**: 117–75

Caplan D, Hildebrandt N, Makris N 1996 Location of lesions in stroke patients with deficits in syntactic processing in sentence comprehension. *Brain* **119**: 933–49

Chomsky N 1995 *The Minimalist Program*. MIT Press, Cambridge, MA

Damasio A, Tranel D 1993 Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Science* **90**: 4957–60

Hart J, Berndt R S, Caramazza A 1985 Category-specific naming deficit following cerebral infarction. *Nature* **316**: 439–40

Jackson H H 1878 On affections of speech from disease of the brain. In: Taylor J (ed.) *Selected Writings of John Hughlings Jackson 1958*. Basic Books, New York

Lecours A R, Lhermitte F, Bryans B 1983 *Aphasiology*. Bailliere Tindal, Paris, Chap. 19

Marshall J C, Patterson K, Coltheart M 1980 *Deep Dyslexia*. Routledge, London

Patterson K, Coltheart M, Marshall J C 1985 *Surface Dyslexia*. Lawrence Elrbaum, London

Pinker S 1999 *Words and Rules*. Basic Books, New York

Rumelhart D, McClelland J C 1986 *Parallel Distributed Processing*. MIT Press, Cambridge, MA

Saffran E M, Marin O, Yeni-Komshian G 1976 An analysis of speech perception and word deafness. *Brain and Language* **3**: 209–28

Shallice T, Warrington E K 1977 Auditory-verbal short-term memory impairment and conduction aphasia. *Brain and Language* **4**: 479–91

Ullman M T, Corkin S, Coppola M, Hickok G, Growdon J, Koroshetz W, Pinker S 1997a A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory and grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience* **9**: 289–99

Ullman M T, Bergida R, O'Craven K M 1997b Distinct fMRI activation patterns for regular and irregular past tense. *NeuroImage* **5**: S549

Wernicke C 1874 *Der aphasische Symptomenkomplex*. Cohn and Weigart, Breslau, Germany, Reprinted in translation in, *Boston Studies in Philosophy of Science* **4**: 34–97

D. Caplan

Appeals: Legal

An *appeal* is a proceeding in a higher court of law initiated by a party contending that a decision of a subordinate court is erroneous. Appeal is to be distinguished from other proceedings sometimes initiated in high courts for purposes other than the

correction of error by a subordinate court. It is also to be distinguished from review proceedings conducted without regard for any previous disposition by the subordinate court, as where a case is subjected to *trial de novo* in the higher court. The latter form of proceeding is, for example, standard practice at the first level of review in Germany (Meador et al. 1994, pp. 893–979). It is not unknown in the USA, where its use is most common when a higher court is reviewing a small claims court where professional lawyers seldom represent the parties and where the judge is sometimes nonprofessional. Appeal is also to be distinguished from the *discretionary review* that may be observed in the Supreme Court of the United States and most other highest courts of states or of large nations. A court performing discretionary review may select the issues or rulings that it chooses to review; its primary purpose in making its selection is to explain the correct resolution of issues that are likely to recur in the lower courts or are otherwise important to persons other than those who are parties. In the USA and in many other countries, most appeals are taken to intermediate courts that are subject to discretionary review in courts of last resort. This article is an account only of the principles governing such appeals, primarily in the intermediate courts of the USA. Some variations in other systems will be noted, but those variations are so numerous that they defy synthesis.

1. Not Universal

The appeal is not a universal feature of legal systems. Some tribal courts, for example, bring to bear in the first instance most of the wisdom and authority of the community (see, e.g., Gluckman 1955). There is then no body of higher persons to whom an appeal might appropriately be addressed. Also, there are autocratic systems, generally those having strong religious roots, in which an individual chief, priest, or judge is accorded the power of decision without possibility of review. Islamic law is noted as an example; the *kadi* administers an elaborate code of conduct contained in holy writ that no mere lawyer can presume to interpret, and the *kadi's* decision is not subject to review (Shapiro 1980).

The appeal as it is known in Western legal systems appears to have been devised by the early Byzantine Empire and rested on the idea that all power was a delegation from the emperor (Pound 1941). It was reinvented in the twelfth century as a method of centralizing power and was used for that purpose by French monarchs as early as the thirteenth century. For similar reasons, the appeal was used intensively in socialist legal systems patterned on the Soviet model (Damaska 1986, pp. 48–52). Its use in the USA reflects the different reality that trial courts there enjoy a measure of political autonomy and are supported by the politically important institution of the local jury, so that the appeal is needed to correct for the tendency of American law to be diffuse.

Typically, in those systems using an appeal, the review is conducted by a larger bench of judges than the court whose judgment is under review. In the USA, there is almost without exception a single judge presiding over a trial court from whose decisions an appeal may be taken to a three-judge court. Judges sitting on an appellate court are generally designated as persons of higher rank, and are likely to receive marginally higher salaries than judges sitting on the trial courts they review.

Even in the USA, the right of appeal in a civil case is generally a matter of legislative grace. Until 1888, there was no appeal from a criminal conviction in a federal court, even if the convicted person was subject to capital punishment (Frankfurter and Landis 1928). However, the constitutions of some states have long guaranteed the right of appeal in criminal cases. It is still the law in all American jurisdictions that a state may not appeal a judgment of acquittal in a criminal case because a second trial would place the defendant in double jeopardy of conviction. Where there is no right to appeal, trial judges were seen to accumulate excessive discretionary power over the individuals engaged in disputes brought before them and on occasion to engage in seemingly lawless behavior. The Congress of the United States and state legislatures have in this century been attentive to the importance of constraining that discretionary power and have provided for appellate review in civil cases and in criminal cases not resulting in judgments of acquittal.

2. Principles of Restraint

The development of the appeal in American courts has resulted in the formulation of principles of restraint that are also in use in various forms in other legal systems. The institution described here is analogous to the *writ of error* familiar to ancient practice of English common law courts as an instrument for bringing a local one-judge decision to the attention of the larger court sitting in Westminster, but departs from that usage in important respects. Contemporary American practice took its present form in the federal and in the courts of the states over the course of the nineteenth century. Six principles of restraint emerged.

2.1 Reversible Error: The Adversary Tradition

The first of these is the concept of *error*, a principle rooted in the Anglo-American adversary tradition placing primary responsibility for the conduct of litigation on the parties and their counsel. Because the burden is on the parties to present the evidence and inform the court of their claims or defenses, it is

not generally error for a court to fail to identify a fact not proven or to fail to enforce a legal principle not invoked. The political function of this principle is to reduce the role of the court: disappointed parties often share responsibility for their own defeats. The administrative purpose of the principle is to encourage thorough preparation and presentation of the case by counsel and to protect the court from being trapped by deceitful or negligent counsel. It follows from this principle of reversible error that an appellant, to be successful on appeal, must generally point to a ruling made by the lower court to which the appellant made timely *objection* (Tigar and Tigar 1999). In times past, it was often required not only that there be an objection to the erroneous ruling, but that an appellant have taken *exception* to the adverse ruling, thus putting the trial judge on notice that the ruling may be challenged on appeal. The requirement of an exception has been eliminated from the practice of federal and most state courts.

Even in American courts, the requirement of a timely objection is sometimes disregarded for erroneous rulings so egregious that the judgment under review deeply offended the appellate court's sense of propriety. Such errors are denoted as *plain errors*. Rarely is the plain error doctrine invoked because it allows counsel to proceed in a no-lose situation, knowing that if the error is not corrected by the trial judge there will be a successful appeal affording a non-objecting party with a fresh start. On the other hand, appellate courts may be reluctant to punish a party having a clearly meritorious claim or defense for no reason other than a lapse on the part of counsel.

2.2 Harmless Error

A second and more universal requirement is that no error however egregious is an occasion for reversal unless it was consequential. The familiar expression is that a *harmless error* is not reversible. Thus, an appellant who should and would have lost on other grounds cannot secure relief from an adverse judgment even if that party can identify blatant errors of fact or of law that were committed by the court below and to which timely objection was made.

2.3 Standing to Appeal

A third requirement is *standing* to appeal. This principle is related to that just stated. Generally, only a party who has participated in the proceeding is bound by the decision below. Hence, one who was not a party is not directly harmed by a judgment even though he or she may strongly disapprove of the outcome and assert that a grave injustice has been done. The requirement of standing may be extended to bar appeal by nonparties who are only remotely affected by a judgment. Thus an investor or employee lacks standing to appeal a judgment rendered against a corporation in which he or she owns shares or by which he or she is employed. The decision to appeal a judgment rendered against a corporation resides with the corporate directors acting on the advice of the officers, and to no one else.

2.4 Ripeness for Review

A fourth common requirement is that of ripeness for review. In general, an appeal is premature until a *final decision* has been reached in the court below. This is a principle of economy. An error of the lower court may turn out to be harmless; the higher court cannot know until a final decision has been reached. Moreover, there would be serious diseconomies in allowing every litigant adversely affected by a provisional ruling of a trial court to take an appeal at once. Not only would this afford parties a means of imposing needless financial costs on adversaries, but it would force the trial court or perhaps even the appellate court to decide in each instance whether the lower court proceedings should be stayed pending the outcome of the appeal. If a stay is granted, the appellant has been empowered to delay the proceeding; and the delay itself may often result in injustice. If a stay is denied, there is the risk that further proceedings will be set aside as a result of the interlocutory decision of the higher court.

There are in most jurisdictions numerous exceptions to the ripeness requirement. For example, in the federal practice, an *interlocutory appeal* may be taken from the grant or denial of a preliminary injunction for the reason that an error in such a ruling can have grave consequences for the party adversely affected (Steinman 1998, pp. 1388–476). Immediate appeal may also be allowed with respect to other rulings involving substantial procedural rights that are very likely to affect the outcome of a case. With respect to important issues of judicial administration, it may be said that an otherwise unripe appeal should be entertained in order to sustain the supervisory power of the appellate court, i.e., to prevent lawless behavior by a trial judge. An appellate court in the USA will also generally possess the power to issue an *extraordinary writ* such as a writ of *mandamus* (a tool of Roman origins) to forestall an abuse of discretion by a subordinate court. In addition, a lower court may be permitted to certify to a higher court a question of law to which the lower court has no answer and which is central to a lengthy trial. And in New York state courts, a ripeness requirement is itself an exception to the more general rule that a party aggrieved by a ruling of the trial division can seek prompt review in the appellate division of the same court.

2.5 The Record on Appeal

A fifth concept shared by all American and most other jurisdictions is that of the *record on appeal*. That is an

official account of the proceedings below, or at least that part of the official account that the lawyers deem pertinent to the issues raised on appeal and therefore worth the cost of reproduction. The record is generally produced by the clerk of the court below and will contain documents filed with the court and a transcript of any oral proceedings prepared by a professional court reporter. In general, an appellate court will not consider information that is not contained in the record (Marvell 1978, pp. 160–66). There is an exception to this principle generally known as the *judicial notice* doctrine. A court may without proof take notice of common knowledge not contained in the record, or knowledge that is readily available to all, such as the coincidence of days of the week with days of the month. In a civil case, the appellant must generally advance the cost of preparing such a record, but a state may be required to bear the cost in a criminal case if the convicted person is indigent.

2.6 *Deferential Review of Factual Determinations*

Finally, there is a principle of deference to trial courts. Among the aims of this principle are to dignify the proceedings below and to discourage appeals challenging the guesswork inevitably done in the trial court to resolve issues on which there is conflicting evidence. The principle is expressed in an elusive distinction between *law and fact*. It is generally agreed that trial courts are entitled to no deference in their rulings on questions of law. Such rulings may be embodied in the *instructions* on the law given to a jury, or if there is no jury demanded, in the *conclusions of law* stated by the court to explain its disposition. An erroneous statement in either of those utterances of the trial judge is a sufficient ground for reversal. On the other hand, trial courts having dealt directly with the evidence submitted by the adversaries are extended the benefit of some doubt with respect to the determinations of fact. The appellate court hears no witnesses and sees only a transcript of the testimony. If the trial court judgment rests on a *jury verdict*, the factual determination in a civil case can be reversed on appeal only if the appellate court finds that there is no *substantial evidence* in the record to support it. In the absence of a jury, a trial judge's decision on evidence will in a civil case be expressed in the judge's *finding of facts*; such findings may be reversed in American courts only if the reviewing court finds '*clear error*.' It is generally assumed that the latter is a lower standard and that fact finding by a judge is more closely scrutinized on appeal than is fact finding by a jury.

The distinction between fact and law is subtle and even sometimes circular; a much oversimplified summary is that issues of fact are those involving specific past events about which there is doubt, whose resolution has little or no bearing on future cases. Issues of law are typically those involving an interpretation of legal texts, but it is often said that whether there is evidence sufficient to support a jury verdict is itself a question of law. All this really means is that sufficiency of the evidence is a question in the first instance for the trial judge but that the ruling on the issue will be reviewed without deference. It is therefore not wrong to say that an issue of fact is simply one that the courts leave for decision by a trier of fact while an issue of law is any issue that the appellate court chooses to decide on its own.

A homely example may assist understanding of this professional jargon. A farmer seeks compensation from the railroad adjoining his farm for a cow hit and killed by the train. The controlling law is that the railroad has a duty to fence livestock out of its right-of-way. But it allows the farmer to keep a gate in the fence. If the cow went through a hole in the fence, the railroad is responsible; if she went through the gate, it is not. If the accident happened at a spot equidistant to the gate and a hole in the fence, convention would say that there was substantial evidence from which a trial court might infer that the railroad's negligence probably caused the misfortune. If, however, the accident occurred near the gate and a long distance from the hole in the fence, it would be unreasonable to infer that the cow probably came through the hole and not the gate. Convention would then say that there is no substantial evidence to support a jury verdict for the farmer; that such a decision by a judge sitting without a jury would be clear error; and that *as a matter of law* the railroad has no liability. How close to the gate and how far from the hole the accident must be in order for the case to present an issue of fact for a jury to decide is itself a question of law.

This distinction is not made in some other legal systems. It is unknown to Japanese practice modeled on the French. It is generally less useful in systems placing heavy reliance on written submissions of evidence. In such systems, the appellate court has access to the same information as the trial court, and hence there is less reason for deference to the judicial officer who saw and heard the adversary presentation of the parties.

2.7 *Abuse of Discretion*

Notwithstanding these settled principles of restraint, an appellate court is also empowered to correct actions of a trial judge that it deems to be an abuse of discretion. This principle is most frequently invoked to challenge and correct procedural rulings that are seen to be idiosyncratic or manifestly unjust (Friendly 1982).

3. *Appellate Procedure*

In addition to these principles of appellate jurisdiction, American courts share a traditional procedure that is replicated in many legal systems. Until about 1960,

this process was universal in the USA and included (1) the submission of written *briefs* prepared by counsel on both sides that present their legal arguments and provide citations to pertinent legal texts and authorities; (2) an *oral argument* at which counsel might engage the appellate judges in dialogue and answer questions they might pose; (3) a conference of the judges responsible for the appellate decision, and (4) a published *opinion of the court* explaining the legal principles underlying the disposition on appeal. These procedural amenities have since been foreshortened in most American appellate courts, especially since the number of criminal appeals has increased precipitously in recent decades, many raising no serious issue worthy of the effort to conduct oral argument, confer, and write an opinion. It has also become common, especially in federal appellate courts, for much of the responsibility to be delegated to law clerks serving as members of the judges' staffs. In recent years, it has been argued that the right of appeal should be abolished in federal practice in recognition of the reality that many appeals are never seriously studied by the judges commissioned as members of the court (e.g., Parker and Chapman 1997). Others have resisted this trend (Arnold 1995). It has even been contended that the erosion of appellate procedure has diminished the *raison d'etre* of the federal appellate courts (Carrington 2000). This is an aspect of judicial administration that is likely to be profoundly affected by electronic communications (Carrington 1998).

4. *The Opinion of the Court*

English appellate courts have since ancient times favored oral opinions delivered separately by each appellate judge in immediate response to the oral argument (Meador et al. 1994, pp. 751–892). This method has the virtue of accelerating the decision, but it is less instructive to lower courts and citizens expected to obey the utterances of the court. The American concept of the opinion of the court is an invention of John Marshall as Chief Justice of the Supreme Court of the United States. Before his time, American appellate judges like their English forebears rendered their decisions orally from the bench after oral argument and without conferring among themselves. The opinion of the court is most useful for courts of last resort having the duty of explaining legal texts not only to subordinate judges and other officials, but also to citizens expected to obey the law and conform their behavior to its requirements. Out of the practice of publishing opinions of the court comes the understanding that American courts 'make law.' In many other legal systems, such opinions of the court are not prepared and published, or if prepared are presented in such summary and didactic form that they shed little illumination on the meaning of the legal texts cited. In countries adhering to that practice it is seldom said that courts make law.

See also: Courts and Adjudication; Judicial Review in Law; Procedure: Legal Aspects; Supreme Courts

Bibliography

Arnold R S 1995 The future of the Federal Courts. *Mo. L. Rev.* **60**: 540

Carrington P D 1998 Virtual civil litigation: A visit to John Bunyan's Celestial City. *Columbia Law Review* **98**: 501

Carrington P D 2000 The obsolescence of the United States Courts of Appeals. *Journal of Law and Politics* **20**: 266

Civil Appellate Jurisdiction, Part 2, 47-3 L. and Contemp. Probs. (P D Carrington (ed)., 1984).

Damaska M R 1986 *The Faces of Justice and Authority*. Yale, New Haven, CT

Frankfurter F, Landis J 1928 *The Business of the Supreme Court: A Study in the Federal Judicial System*. Macmillan, New York

Friendly H J 1982 Indiscretion about discretion. *Emory Law Journal* **31**: 747

Gluckman M 1955 *The Judicial Process Among the Barotse of Northern Rhodesia*. Manchester University Press, Manchester, UK

Jolowicz J A 2000 *Civil Procedure*. Cambridge University Press, Cambridge, UK

Marvell T 1978 *Appellate Courts and Lawyers: Information Gathering in the Adversary System*. Greenwood, Westport, CT

Meador D J, Rosenberg M, Carrington P D 1994 *Appellate Courts: Structures, Functions, Processes and Personnel*. Michie, Charlottesville, VA

Meador D J, Bernstein J S 1994 *Appellate Courts in the United States*. West, St. Paul, MN

Parker R M, Chapman R 1997 Accepting reality: The time for accepting discretionary review in the Courts of Appeals has arrived. *SMU Law Review* **50**: 573

Phillips J D 1984 The appellate review function: Scope of review. *Law and Contemporary Problems* **47**: 1

Pound R 1941 *Appellate Procedure in Civil Cases*. Little Brown, Boston

Shapiro M L 1980 Islam and Appeal. *California Law Review* **68**: 350

Steinman J 1998 The scope of appellate jurisdiction. *Hastings Law Journal* **49**: 1337

Tigar M E, Tigar J B 1999 *Federal Appeals, Jurisdiction and Practice*. West, St. Paul, MN

P. Carrington

Appeasement: Political

Appeasement is a policy of settling international disputes by admitting and satisfying grievances through rational negotiation and compromise, thereby avoiding war. Because of British and French concessions to Hitler at Munich in 1938, appeasement has

acquired an invidious, immoral connotation. In the classic balance-of-power system, however, appeasement was an honorable policy reflecting the principle that the international system needed some means of peaceful adjustment to accommodate changing national power and aspirations. As a small insular power with a far-flung empire, the British successfully pursued appeasement throughout the nineteenth century. Even the conventional wisdom about the 'lessons of Munich' has been questioned by revisionist historical scholarship. Scholars and policy analysts now view appeasement as a useful strategy for maintaining international stability under certain conditions.

1. *Definition*

As late as 1944, Webster's dictionary defined appease as 'to pacify (often by satisfying), to quiet, soothe, allay.' The 1956 edition of Webster's dictionary adds another clause: 'to pacify, conciliate by political, economic, or other considerations; now usually signifying a sacrifice of moral principle in order to avert aggression' (Herz 1964).

Appeasement, however, may have little to do with moral principles but instead reflect national interests. In the classical era of diplomacy, appeasement was a method of adjusting the balance of power to preserve an equilibrium between the relative power of states and the distribution of benefits. A declining state might accommodate a rising power with colonies or spheres of influence to dissuade it from engaging in a costly bloody war to overturn the international system. The British followed such a policy in the nineteenth century as their industrial, military, and economic strength declined relative to the United States and Germany.

Appeasement was also used to satisfy a revisionist state's legitimate grievances so that it would not go to war. Diplomatic theorists believed that it was futile and self-defeating to try to prevent all change because the revisionist state would eventually try to achieve its objectives by force unless some effort was made at negotiation and conciliation.

2. *British Tradition of Appeasement*

Historians Paul Schroeder and Paul Kennedy have shown that Munich was not a departure from the traditional British policy of maintaining a balance of power, but a continuation of a tradition of appeasement that was a prudent response to entangling obligations and commitments. For the British in the nineteenth century, appeasement referred to the attempt to stabilize Europe and preserve peace by satisfying a revisionist powers' justified grievances.

The British tradition of appeasement was a response to ideological, strategic, economic, and domestic political considerations. The British adhered to internationalist principles favoring arbitration and negotiation of differences between states, disarmament, and abhorrence of war except in self-defense. By the middle of the nineteenth century, Britain's Royal Navy could not defend the far-flung global empire, and British commitments outreached its military capabilities. The British suffered from strategic overextension. With her military forces stretched thinly over the world, Britain had an incentive to establish priorities among global interests, settle disputes peacefully where possible, and to reduce the number of its enemies. The center of a global economy, Britain imported raw materials and foodstuffs, and exported manufactured goods and coal. Britain provided insurance and overseas investment. The British economy would have been severely disrupted by war, as imports would have exceeded exports and Britain's income from 'invisible' services would have been cut off. Peace for Britain was a vital national interest. As the franchise expanded after 1867, British governments increasingly had to take into consideration public opinion. The British public disliked wars, especially expensive ones, preferring expenditures on social programs and economic reforms.

The British often used appeasement, and it was a successful policy for them, particularly toward the United States. The United States was a rising power. By the 1820s the United States had a larger population than Britain; by the 1850s it had a larger gross national product (GNP); and in the 1890s, the United States was expanding its navy. Yet, Britain and the United States did not fight a hegemonic war for world power. Much of the credit for avoiding war should be given to Britain for its numerous concessions to the United States.

The British had commercial and strategic reasons for appeasing the United States. Britain bought American cotton and wheat. The British were also aware of the vulnerability of Canada; in any war with Britain, the United States would invade Canada. The British did not have ground troops in Canada. Finally, Britain was already involved in quarrels with France, Russia, and Germany, and needed to reduce the number of its enemies.

In the 1842 Webster–Ashburton Treaty, Britain handed over most of northern Maine and the head of Lake Superior to the United States. The 1846 Oregon Treaty extended the border of Canada to the Pacific coast along the 49th parallel, giving the United States most of what is now called Washington State and Oregon. The Oregon territory, which included what is now Washington, Oregon, Idaho, parts of Montana and Wyoming, and half of British Columbia, had been held jointly by Great Britain and the United States. President James Polk had campaigned on the slogan 54–40 or fight, meaning that the United States should have the entire Oregon territory. After being elected, Polk announced that the United States was with-

drawing from the treaty sharing Oregon with Britain, and that the USA would put forts and settlers in the territory. Using deterrence, the British promptly sent thirty warships to Canada. Polk backed down, and submitted a compromise to Congress at the 49th parallel. Canadians viewed these treaties as sellouts to the United States, but Britain did not believe that the territory was worth fighting for, and if there was a fight, Britain might lose.

In 1895, Britain accepted US arbitration in a territorial dispute between Venezuela and British Guiana that almost led to war between the United States and Britain. Britain claimed part of Venezuela for British Guiana. In 1895, the USA demanded that it be allowed to arbitrate the dispute under the Monroe Doctrine. After four months, the British sent a note declining arbitration and the legitimacy of the Monroe Doctrine. President Grover Cleveland was enraged by the condescending tone of the note and by denial of the American right to settle the dispute. In December 1895, in a belligerent speech to Congress, President Cleveland threatened to go to war. The British were beginning to fear Germany, and they had no desire to get into a war with the United States. In 1896, the British accepted the US proposal that the dispute be arbitrated, and the crisis was over. Afterwards, American–British relations improved dramatically. The USA and Britain were never again close to war.

In 1902, Britain gave up its right under the Clayton–Bulwar treaty to have a share in any canal constructed in Panama. At the turn of the century, Britain withdrew most of her navy from the Western Hemisphere. Through appeasement, Britain satisfied American demands and altered American attitudes toward Great Britain, which led to a transformation of their relationship. The United States did not increase its demands as a result of British concessions, and America fought with Britain in World War I.

3. *Munich*

British and French attempts in the 1930s to buy off aggressors at the expense of weaker states have made a permanent stain on the policy of appeasement. The same strategic, economic, and domestic political conditions that encouraged the British to pursue appeasement in the nineteenth century, though, were even more pressing in the interwar period. The worldwide depression caused a contraction in Britain's invisible earnings as well as its overseas exports. By 1935, when the need for rearmament was apparent, defense expenditures were limited by the need to avoid jeopardizing Britain's economic recovery and financial position. In 1937, the Chiefs of Staff warned that British defenses were not powerful enough to safeguard British trade, territory, and vital interests against Germany, Italy, and Japan. Diplomacy must be used to reduce the number of potential enemies and to gain the support of allies. The British public had horrific memories of World War I (Kennedy 1983). Britain could not go to war without the support of the Dominions and before completing its rearmament program. In short, the British policy of appeasement was overdetermined by the popular fear of another devastating war, military unpreparedness, concern about the British economy and the Empire, and isolationism in the Dominions and the United States (Schroeder 1976).

A widely drawn lesson of Munich is that appeasement merely increases the appetite of aggressors, avoiding war now in return for a worse war later. The Munich agreement signed by Italy, France, Britain, and Germany on September 29, 1938 ceded to Hitler the Sudetenland of Czechoslovakia. On March 15, 1939, the Germans invaded Prague and annexed the rest of Czechoslovakia, provoking an outcry among British public opinion, and permanently discrediting the policy of appeasement.

Recent scholarship, though, brings into question some aspects of the conventional wisdom concerning 'the lessons of Munich.' British Prime Minister Neville Chamberlain suspected and mistrusted Hitler, but did not believe that the fate of the Sudetenland Germans was a cause justifying war, given the apparent reasonableness of the demand for self-determination. British appeasement did not whet Hitler's appetite for territory, because he had already formulated his foreign policy aims. In 1939, Hitler decided to attack Poland and was willing to risk war with Britain to do so. Hitler could not have been deterred from his plans by British threats to go to war; he *wanted* war (Richardson 1988).

4. *Uses and Limitations of Appeasement*

Rather than using 'appeasement' as a term of opprobrium, we need to identify the conditions under which it is a viable strategy for avoiding war, and those under which it will increase the likelihood of war. Appeasement of a revisionist state's grievances may be necessary if the status quo power has other competing geopolitical interests and domestic public opinion does not favor armed resistance to any changes in the international system. Yet, almost no systematic work has been done to compare successful and unsuccessful uses of appeasement.

Preliminary analysis, though, suggests that in order for appeasement to work, the revisionist power's demands should be limited. Ideally, the revisionist power's claims should have inherent limits—bring members of their ethnic group into their territorial boundaries, more defensible frontiers, or historic claims to some land. For example, the United States was expanding throughout the continent, but did not have any aspirations for overseas colonies or for world

domination. Because American aims were limited, British concessions were not likely to provoke additional demands. If concessions are made gradually, then the status quo power can assess the other state's intentions and provide incentives for good behavior.

Appeasement is usefully combined with deterrence toward any additional changes to a settlement. This means that the appeasing state should retain the option of using force—by developing adequate military capabilities to defend a settlement. Concessions made from a position of strength are more likely to be viewed by the opponent as an attempt to conciliate rather than as evidence of weakness. In the nineteenth century, Britain was more powerful than the United States. Neither deterrence nor appeasement is likely to be useful by itself.

See also: Balance of Power: Political; Conflict/Consensus; Deterrence; Deterrence: Legal Perspectives; Diplomacy; Dispute Resolution in Economics; Foreign Policy Analysis; International Law and Treaties; National Security Studies and War Potential of Nations; Second World War, The

Bibliography

Craig G A, George A L 1995 *Force and Statecraft: Diplomatic Problems of Our Time*. Oxford University Press, New York

Herz J H 1964 The relevancy and irrelevancy of appeasement. *Social Research* **31**: 296–320

Kennedy P M 1983 The tradition of appeasement in British foreign policy, 1865–1939. In: Kennedy P M (ed.) *Strategy and Diplomacy 1870–1945*. Allen and Unwin, London

Richardson J L 1988 New perspectives on appeasement: Some implications for international relations. *World Politics* **40**: 289–316

Schroeder P W 1976 Munich and the British tradition. *The Historical Journal* **19**: 223–43

D. W. Larson

Appetite Regulation

As it is commonly used, the concept of appetite is identified with the sensation of hunger or the subjective 'urge to eat.' Accordingly, much research on appetite regulation is devoted to specifying the processes or mechanisms that are involved with meal initiation, termination, frequency, duration, and food selection. Although most of this work attempts to describe physiological (e.g., neural, hormonal, metabolic) control mechanisms, the role of psychological factors such as learning about foods and the consequences of eating have also received much recent attention. The purpose of this article is to summarize briefly what is known about the physiological and psychological bases of appetite regulation.

1. Physiological Controls of Appetite

1.1 Central Control Mechanisms

Early studies showed that electrical stimulation of the lateral hypothalamus (LH) evokes robust eating, whereas destruction of this area produces a dramatic suppression of intake. In contrast, respective stimulation and lesion of the ventromedial hypothalamus (VMH) has the opposite effects on food intake. Based on such findings, the LH and VMH were conceptualized as forebrain hunger and satiety centers, respectively. More recently, emphasis on hypothalamic control of intake has been reduced based on evidence that brainstem areas also have a significant role in appetite regulation. Decerebration is a procedure that can be used to surgically disconnect the hypothalamus and other forebrain structures from neural input originating at sites in and below the level of the brainstem. Although decerebrate rats do not search for or initiate contact with food, when liquid food is infused directly into the mouth, these rats consume discrete meals. In addition, the taste reactivity (appetitive orofacial responses) of these rats also appears to be sensitive to food deprivation and other manipulations that influence the intake of intact animals. As noted by Berthoud (2000), these and other findings suggest that hypothalamic and brain stem nuclei are important components of a larger neural system for appetite control. This system also includes brain regions that are involved with the rewarding aftereffects of eating (e.g., prefrontal cortex, nucleus accumbens), that mediate learning about internal signals (e.g., hippocampus), and that are potential processing sites for peripheral signals related to energy balance (e.g., central nucleus of the amygdala).

1.2 Peripheral Controls and Signals

Appetite regulation by the brain undoubtedly depends on signals received from the periphery. The stimulus to eat, at first claimed to originate in the stomach, is now proposed to be a consequence of departures from homeostatic levels of glucose, lipids, or amino acids. Debate continues about whether changes in the availability of each of these metabolic fuels give rise to separate feedback signals or whether this information is integrated to produce a single, common, stimulus to eat. Receptors for such metabolic signals have been found in the liver (see Langhans 2000). The liver is centrally located with respect to metabolic traffic and is capable of detecting even slight changes in circulat-

ing metabolic fuels, including those that are denied access to the brain by the blood–brain barrier. Infusing metabolic antagonists directly into the liver promotes food intake and induces electrophysiological changes in vagal afferent pathways that project from the liver to brainstem loci known to be involved with feeding.

The termination of meals also seems to depend on the detection of specific peripheral signals. For example, when nutrients are first absorbed in the gut, the rate at which food is emptied from the stomach slows resulting in stomach distension as eating continues. Vagal afferent fibers carry signals produced by stomach distension to the brain. Food intake also produces short-term hormonal changes that appear to be involved with meal termination. Early findings that systemic injection of the gut peptide cholecystokinin (CCK) reduces meal size in a variety of species, including humans, led to the suggestion that CCK might play a role in normal meal termination (see Smith 1998). Findings that CCK is released by the presence of preabsorptive nutrients in the small intestine, and that blocking the effects of endogenous CCK release by administration of CCK receptor antagonists increases meal size support this hypothesis. Furthermore, the presence of CCK increases vagal afferent activity, an effect that is also blocked by CCK antagonists. Thus, CCK release may lead to meal termination as part of a mechanism that informs the brain when food has been detected in the gut.

Neuropeptides may also play an important role in the long-term control of appetite. Recent evidence indicates that the brain receives information about the status of bodily fat stores via neuropeptide signaling systems. One such signal may be provided by leptin, a circulating neuropeptide that is the product of the adipose tissue-specific *ob* gene. Secretion of leptin is correlated with body fat mass. Furthermore, extreme fasting decreases, whereas overfeeding increases, leptin concentrations in the blood. In addition, genetic mutations that produce defects in either leptin production or in its hypothalamic receptor lead to the development of obesity in several rodent models. Thus, leptin may be part of a signaling system that enables information about body fat mass to be communicated to the brain. The brain may use the information provided by the leptin signaling system to defend a 'set-point' or homeostatic level of body fat. Many other neuropeptides have been shown to alter food intake. For example, neuropeptide Y (NPY), and several opioid peptides are known to increase food intake. Along with CCK, serotonin, estradiol, and glucagon are prominent on the list of peptides that suppress feeding. In addition, the central melanocortin system has endogenous agonists and antagonists that are known to suppress and promote intake, respectively. Furthermore, some neuropeptides may influence the selection of specific macronutrients. For example, NPY and serotonin have been reported to selectively alter carbohydrate intake, whereas as galanin and enterostatin modulate selectively the intake of fat. This suggests that neuropeptides may influence not only energy balance but also the types of foods that are selected.

2. *Psychological Controls of Appetite*

2.1 *Social/Cultural Factors*

The traditions and standards of one's social group or culture to which we belong are important determinants of meal initiation, termination, and food selection. For example, the relative tendency to eat chilli peppers, sushi, and cheeseburgers is influenced by which type of food is most common or available in one's culture. Furthermore, even though a young child may at first reject, for example, very spicy or piquant foods, preference for such foods often develops after repeated exposure. Because one's culture determines, in large part, the types of food to which one is exposed (e.g., spicy foods, fatty foods, highly caloric foods), including what foods are considered acceptable at all (e.g., insects, dog meat, pork), it follows that food preferences and, to some extent, level of caloric intake are subject to strong cultural influences.

2.2 *Learned Controls of Eating*

One way to develop a preference for a particular food is to associate that food with a stimulus that is already preferred. Conversely, a preferred food will be liked less to the extent that it is associated with an aversive or unpleasant stimulus. A form of learning that produces changes in food preferences is flavor–flavor learning. For example, animals will come to prefer flavor A relative to flavor B (two hedonically neutral flavors) to the extent that flavor A has been combined previously with the sweet taste of saccharin. This outcome is obtained when both flavors are presented without saccharin during testing. Because saccharin contains no calories, increased preference for flavor A must have been based on its association with saccharin's sweet taste rather than its caloric consequences. Conversely, presenting a neutral flavor in solution with quinine (a bitter substance that is normally rejected by rats) reduces the preference shown by rats for that flavor when it is presented without quinine.

The flavor of food can also be associated with the postingestive consequences of intake. Conditioned taste aversion, which is demonstrated when animals avoid eating a normally acceptable food that has been associated previously with intragastric malaise, provides a robust example of this type of learning. Conversely, food preferences result when flavors are associated with the caloric or nutritive postingestive aftereffects of eating. This flavor–nutrient learning has

been demonstrated with rats in studies where the consumption of one noncaloric, flavored cue solution (CS+) is paired with the infusion of carbohydrate or fat solutions directly into the stomach, whereas intake of a different noncaloric solution (CS–) is paired with gastric infusion of water (see Sclafani 1997). After several of these training trials, strong preferences develop for the flavor that was followed by intragastric infusion of nutrients. Thus, the fact that conditioned food preferences are observed even when the nutritive consequences of intake completely bypass the oral cavity confirms that flavor–nutrient and flavor–flavor learning can occur independently.

Learning also contributes to meal initiation. Conditioned meal initiation occurs when environmental cues that are associated with eating when animals are hungry acquire the capacity to initiate large meals when the animals are tested when food sated. For example, repeatedly consuming meals at the dining room table when hungry may endow that place with the ability to initiate meals even when hunger is absent. The phenomenon of conditioned meal initiation has been demonstrated for both human and nonhuman animals, with punctate cues (e.g., discrete tones or lights) as well as contextual stimuli. Thus, the initiation of feeding appears under the control of learning mechanisms.

Experience with the sensory aspects of food can also contribute to meal termination independent of any reduction of physiological need produced by that experience. For example, hungry rats will eat a substantial meal of Food A before they voluntarily stop feeding. If the rats are offered the same food again a short time later they refrain from eating. However, if they are offered a new food, Food B, they ingest a second meal that can be calorically equal to or greater than the first. This is not simply a consequence of a difference in palatability between the original and novel foods. A large second meal is consumed even when the order of the foods eaten is reversed. Thus, rats do not become satiated for calories *per se* but become 'satiated' for specific tastes, textures, odors, or other sensory properties of food. This type of specific satiety has also been found in studies of human eating.

2.3 Appetite and Incentive Value

Animals are said to be attracted to pleasant or rewarding stimuli. Incentive motivational accounts propose that increased hunger promotes eating and appetitive behavior by enhancing the reward value of food (e.g., its taste or positive postingestive aftereffects). In addition, the attractiveness of environmental stimuli that are associated with the presentation of food are also enhanced by hunger. Enhancing the attractiveness or value of a stimulus improves its ability to compete with other environmental cues for behavioral control. Thus, modulation of the value of the orosensory and postingestive consequences of food and of stimuli associated with food may be an important psychological basis for appetite regulation.

See also: Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Eating Disorders, Determinants of: Genetic Aspects; Food in Anthropology; Food Preference; Food Production, Origins of; Hunger and Eating, Neural Basis of; Obesity and Eating Disorders: Psychiatric; Obesity, Behavioral Treatment of; Obesity, Determinants of: Genetic Aspects

Bibliography

Berthoud H R 2000 An overview of neural pathways and networks involved in the control of food intake and selection. In: Berthoud H-R, Seeley R J (eds.) *Neural and Metabolic Control of Macronutrient Intake*. CRC, Boca Raton, FL, pp. 361–87

Capaldi E D 1993 Conditioned food preferences. In: Medin D (ed.) *The Psychology of Learning and Motivation*. Academic Press, New York, Vol. 28, pp. 1–33

Langhans W 2000 Portal-hepatic sensors for glucose, amino acids, fatty acids, and availability of oxidative products. In: Berthoud H-R, Seeley R J (eds.) *Neural and Metabolic Control of Macronutrient Intake*. CRC, Boca Raton, FL, pp. 309–23

Legg C R, Booth D A 1995 *Appetite: Neural and Behavioural Bases*. Oxford University Press, Oxford, UK

Sclafani A 1997 Learned controls of ingestive behavior. *Appetite* **29**: 153–8

Smith G P 1998 Cholecystokinin—the first twenty-five years. In: Bray G A, Ryan D H (eds.) *The Pennington Center Nutrition Series: Nutrition, Genetics, and Obesity*. Louisiana State University Press, Baton Rouge, LA, pp. 227–45

Thorburn A W, Proietto J 1998 Neuropeptides, the hypothalamus and obesity: insights into the central control of body weight. *Pathology* **30**: 229–36

T. L. Davidson

Applied Geography

Applied geography emphasizes the social relevance of geographic research, which focuses on human-environment interactions, area study, and spatial-location problems. In general, the spatial distributions and patterns of physical and human landscapes are examined, as well as the processes that create them. Applied geography takes place within and outside of university settings and often bridges the gap between academic and nonuniversity perspectives.

Wellar (1998a) has referred to the difference between purely academic research and applied research as client-driven vs. curiosity-driven research. Whether university or nonuniversity based, geography becomes applied when the researcher undertakes a problem for

a client. While academic geographers' research is driven by their curiosity to understand patterns and processes, applied geographers perform research for a client with a 'real-world' problem. This results in a very different approach than an academic endeavor. The client can be a retailer in a capitalist economy, a group that receives unequal treatment from a capitalist or some other society and, therefore, would benefit from the empowerment that applied research findings may provide, or any other person or agency seeking or needing relief from a solution to a geographic problem.

The 'applied' approach provides a user-orientation and an effectuation plan quite different from those of curiosity-driven academic research. The research results are also presented in different formats and used differently. Applied research results may be input into the design of a new product, may result in a strategy for delivering a new product, or may be a set of recommendations that inform decisions. Academic research pursued for curiosity, typically results in the production of a book, monograph, or research article in a journal publication for the purpose of informing colleagues of findings, especially in terms of the contribution to existing theory. Applied geographers in vs. out of the university often differ in their views of what applied geography is. Despite these differences, applied geography is united by emphasis on useful knowledge.

University-based applied geography is narrower in scope than nonuniversity applied geography. The university presumes a theoretical and an empirical basis for applied research. Nonuniversity-based applied geography occurs within business and government cultures that often results in highly specialized roles for individuals applying geographic knowledge and skills in the solution of problems. The differences between the two workplaces shape research perspectives, define broader nonresearch obligations, and result in different perspectives. For example, in the university context the researcher controls and participates in all aspects of research, leading most academics to define 'applied' as the research performed by the researcher for a client. Most academics also emphasize research over effectuation planning.

By contrast, in a nonuniversity setting an applied geographer has a job description that indicates specific tasks and places the employee in a reporting hierarchy. Individual responsibilities are tied to a 'client relationship' that is inherent to a production environment. The employment specializations of nonacademic geographers range from planner-technician to analyst, from a research scientist to director of operations, and from project director to executive (Frazier 1994). In the nonacademic world every cog in the wheel is crucial. In short, the very nature of the nonacademic world requires multiple stages in the creation of a product. An applied geographer may be performing research or implementing the research of others, doing none herself, but contributing greatly to the process. Being directly involved in all parts of the applied research process is not a prerequisite to doing applied geography outside academia.

What holds applied geography together is the value context placed on knowledge. The applied geography approach seeks useful knowledge that helps solve a specific problem for a specific client. Pacione (1999) describes examples of practical recommendations for policy that address the issue of ethical standards in applying science. Specifically, he used British and American examples that illustrate applied geographers taking anti-establishment, unpopular positions to effectuate research findings that would change existing urban policies.

The goals of applied geographic research differ from those of basic research but the boundary between the two can be fuzzy. The value of 'useful knowledge' helps clarify the differences between the two approaches.

1. *Applied Geography: An Historical Context*

1.1 *Early Leaders by Example*

A dialectic between academic (pure) and nonacademic (applied) geography is reflected in the work of many famous geographers in the early twentieth century. Perhaps the most frequently cited case of geography in action is the work of L. Dudley Stamp in the Land Utilization Survey of Great Britain. Stamp called for geography not only to use its methods and concepts to interpret the world, but also to help forward the solution of some of the great world problems (as quoted in Frazier 1981, p. 3). Stamp's work in the 'Survey' put geographic methods and concepts into action and provided a touchstone for their utility. The 'Survey' methods and findings were replicated by others and informed geographic theory and approach for decades, providing not only planning inputs but research topics in academic geography.

In the US, several geographers of national reputation converted their academic interest in human-environmental topics into government employment. Among them were Harlan Barrows, who between 1933 and 1941 worked on a variety of water resource projects, including the design of comprehensive development schemes on a regional basis (Colby and White 1961). Even earlier, Carl O. Sauer, an architect of academic cultural-historical geography, played a leadership role in the Michigan Land Economic Survey. The frequently employed theme of 'human as agent' was applied in this land assessment and classification effort. Sauer applied well-established academic geographic concepts and methods in an applied research mode. He dealt directly with specific forms of 'destructive exploitation' and their specific impacts on the land. Further, his scheme included a model of potential future uses, including modifications

that would realize better results. Sauer's efforts were utilized in land management planning as well as being published in the academic literature (Sauer 1919); they informed future action and future academic research. In Sauer's case he followed up on applied research by being involved in the implementation stages as an activist.

Perhaps the most internationally respected expert on natural hazards is Gilbert F. White. He is another example of an academic geographer that has served the nonacademic community for half a century. White's early concern for victims of flooding and other natural hazards led to applied work for the administration of Franklin D. Roosevelt and new policy. White established academic and nonacademic followings and his leadership resulted in the creation of The Center for Natural Hazards Studies at the University of Colorado, Boulder. White and his colleagues have used geographic concepts and methods to understand different types of natural hazards and find ways to mitigate their impacts on people. Stressing physical and human variability, White has focused on improving the plight of human kind (White 1974).

While this is not a comprehensive review of applied geography, we would be remiss to exclude the founder of business geography in the nonacademic world. In the 1930s William Applebaum accepted a position with The Kroger Company. His introduction of very basic geographic concepts and methods for analyzing market regions was highly successful and created a career path for others to follow. Initially, applied geographers focused on measurement techniques, market delimitation and competitor analyses. However, Applebaum maintained a close relationship with a small group of academic geographers and hired new geography graduates. Slowly, applied business geographers expanded their roles to include location analyses, site selection, and the application of models. As a result of this dialectic relationship between Applebaum and, later, others from the business world with academics, a new subspecialty, business/marketing geography, emerged in the US and new concepts and methods were exchanged between the two environments. In Europe, Ross L. Davies is an example of an academically based applied geographer who has spent a career conducting research that guides understanding of retail structure and behavior and that has informed business decisions.

1.2 Institutionalizing Applied Geography: Two American Examples

Geography has been incorporated into government and business at various levels and in various ways. At one level, geography is inherent in business decisions and has resulted in many firms hiring geographers as location and/or site analysts. Some have risen to managers and executives. At another level, however, agencies formalize geography through their missions and/or internal organizational structure. Two American examples suffice.

1.2.1 The American Geographical Society: useful knowledge and geography in action. When writing the history of the American Geographical Society, J. K. Wright noted its earliest public purposes: '... The advancement of geographical science and the promotion of business interest of a "great maritime and commercial city" were the Society's two leading purposes' (Wright 1952, p. 69). These were most often expressed in the society's early years through published research reports, including those resulting from Society-backed field expeditions. An early example, which also reflects another Society purpose, supporting religious causes, was the appointment of the 'Committee on Syrian Exploration,' which was championed by a rabbi and, according to the Committee, would 'lead to important discoveries and human development by reclaiming the desert, creating self-sufficiency and linking the region to world trade' (Wright 1952, pp. 39–40). This was in keeping with the Society's purposes and exemplifies its efforts to put geography into action that would promote human welfare.

In attempting to institutionalize applied geography in society, the AGS sought to provide important data useful to government and business decision making. This is obvious in a number of the Society's efforts. In its early years statistics had been hailed as knowledge capable of eliminating 'the fogs of human ignorance and suffering' (Wright 1952, p. 47). The Society attempted to influence census taking and even proposed a new federal tax system as its 'most ambitious attempt to influence government policy' (Wright 1952, p. 50). The intended applied thrust of the AGS is perhaps best stated in its inaugural issue of the *Geographical Review*: 'It is the essence of the modern ideal that knowledge is of value only when transformed into action that tends to realize the aspirations of humanity. It is precisely this view that the Society has always taken' (*Geographical Review*, 1916, pp. 1–20, as quoted by Wright 1952, p. 195).

Other national and international roles by the Society and its members are well documented. Among the most prominent are Society activities associated with the Congressional Act that established the International Meridian Conference in Washington, DC and the leadership role of Society Director, Isaiah Bowman, in 'The Inquiry' after World War I. This enterprise was headquartered at The Society. At the Paris Peace Conference that followed, Bowman served in an executive capacity at the request of President Wilson.

1.2.2 Applied geography in government: The US Bureau of Census Geography Division. Perhaps no federal agency anywhere has been more influenced by

the geographic perspective than the US Bureau of the Census. For more than a century, the value of geographic knowledge has been applied in the acquisition and reporting of data for the establishment and monitoring of national and state programs in Commerce, Agriculture, Housing and Urban Development, Health and Human Resources, Indian Affairs, Transportation, Energy, Defense and CIA, and others. Applied geographers have played important roles throughout the federal government in such activities. However, the Geography Division of the US Census has long been charged with very specific duties, including the delineation of meaningful statistical areas, the creation and/or implementation of methods for the capture, organization, and reporting of useful geographic knowledge for problem solving. Torrieri and Radcliffe (in press) noted the early work of Census Chief Geographer, Henry Gannett, who by 1890 led the development of area definitions and mapping techniques, and was largely responsible for a variety of population reports, including the 'Report of the Nation' (Torrieri and Radcliffe in press).

Over the next century Geography's role expanded and by 1978 Chief Geographer, Jacob Silver, provided a summation of responsibilities that included the development of national geographic reference files and digital mapping techniques (Silver 1978). By the close of the twentieth century the Geography Division supports the most widely used digital base file in the world (TIGER) and is planning new census techniques, area definitions and portrayal methods for census data in the twenty-first century.

1.3 Themes in Applied Geography

Most subdisciplines of geography contain examples of applied geography. Examples include a concern for environmental processes and patterns that influence human development and well-being, such as the physical processes of erosion, sedimentation, and desertification. Human–environmental relationships take on many forms but prominent among them are natural hazards behaviors and environmental degradation of all types. The applied regional approach, or area study, involves any effort to create or analyze geographic zones for the benefit of human welfare. We already reported the leadership role of census geography, but others include the creation of planning regions by applied geographers to plan, administer, and monitor urban and environmental areas. Location principles are applied to a wide range of analyses but are probably best known in applications for site selection in business and government and for retail market analyses, including the internationally known Huff model.

Recently, Torrieri and Radcliffe identified seven general classes into which they believe most applied geography falls: (a) market/location analysis; (b) medical geography; (c) settlement classification and statistical geographic areas; (d) land use, environmental issues and policy; (e) transportation planning and routing; (f) geography of crime; and (g) developmental tools and techniques for aiding geographic analysis (Torrieri and Radcliffe in press). They admit that these classes are not exhaustive. It is unlikely that any nomenclature will satisfy all applied geographers. However, such classifications are useful in understanding the wide range of applied geography that occurs internationally.

One contemporary example illustrates geography at work in the transportation and routing area. Barry Wellar, who worked for the Provincial government of Ontario before becoming a professor at the University of Ottawa, has led a research team under contract with the government to develop a 'Walking Security Index.' This index will be used to assist Canadian officials and citizens in assessing the safety of signalized intersections for pedestrians (Wellar 1998b). After developing the necessary indices and examples for the government, Wellar and his colleagues have presented their results in public meetings and through the media. The final report contained 17 intersection recommendations based on the research and interactions with the three clients, government officials, staff professionals and citizens. It also contains implementation issues and strategies.

1.4 Future of Applied Geography

Several trends indicate a bright future for applied geography. As noted above, applied geography was institutionalized both inside and outside academia in important ways. One trend is its increasingly formal treatment as part of academia, where it is considered a valid approach challenged in the philosophical debates of the discipline (Pacione 1999). Further, the British journal, *Applied Geography*, created a lasting publication forum for those interested in particular aspects of this approach. Finally, in America, the annual 'Applied Geography Conferences' can be considered an institution that has brought thousands of applied geographers together from business, government, and the university.

The second trend in Geography is the recent improvements in geographic education at all levels. Professional European Geography was founded in the nineteenth century at least in large part due to educational concerns. At the close of the twentieth century, concerns for geographical ignorance spawned a revolution in geographic education. In America, the generosity of the National Geographic Society and the labor of educators of the National Council for Geographic Education have contributed significantly to improving the quality of geographic teaching through geographic alliances in all 50 states. If applied geographers can be viewed as the 'outputs' of geographic education, then children are the 'inputs' that are shaped and mentored at all levels of the edu-

cational process. This trend of improved geographic education to rid the world of geographical ignorance has contributed positively to a third trend, that of greater public awareness of the importance of geography and its value, not only in better appreciating and understanding our local and global environments, but in contributing to the solution of significant global (e.g., global warming), national (e.g., poverty and inequality), and local (e.g., environmental health) problems.

Another trend that has contributed to our ability to recognize and solve geographic problems and to institutionalizing geography in the public and private sectors is the creation of geographic-based automated technology. Geography is now inseparably linked to information systems via GIS (geographic information systems). Many employers seek technical and analytical employees who can use this technology in problem solving. Among the other digital technologies often linked to GIS and, therefore, extended to applied geography, are remote sensing and global positioning systems (GPS). Both technologies frequently find their academic homes in geography departments and are requirements for undergraduate and graduate degrees. In short, technical skills are now often associated with applied geography and its practitioners because they acquire, portray, and analyze useful geographic knowledge for problem solving.

Finally, and perhaps most important, there is no shortage of pressing global, regional and local problems that require an applied geography approach. Many have been mentioned here. As long as there are geographers willing to contribute to the solution of problems that are inherently geographic or have geographic dimensions, there will be applied geography. Where geography is put into action, applied geography occurs.

See also: Cultural Geography; Environmental Policy; Geography; Human—Environment Relationships; Social Geography

Bibliography

Colby C C, White G F 1961 Harlan Barrows 1877–1960. *Annals of the Association of American Geographers* **51**: 395–400

Frazier J W (ed.) 1982 *Applied Geography Selected Perspectives*. Prentice-Hall, Englewood Cliffs, NJ

Frazier J W 1994 Geography in the Workplace: A Personal Assessment with a look to the future. *Journal of Geography* **93**(1): 29–35

Geographical Review 1916 **1**: 1–2

Harris C D 1997 Geographers in the U.S. government in Washington, D.C. during World War II. *Professional Geographer* **49**(2): 245–56

James P E 1972 *All Possible Worlds. A History of Geographical Ideas*. Odyssey Press, Indianapolis, IN

Mayer H M 1982 Geography in city and regional planning. In: Frazier J W (ed.) *Applied Geography Selected Perspectives*. Prentice-Hall, Englewood Cliffs, NJ, pp. 25–57

Pacione M 1999 Applied geography: in pursuit of useful knowledge. *Applied Geography* **19**: 1–2

Sant M 1982 *Applied Geography: Practice, Problems and Prospects*. Longman, London

Sauer C O 1919 Mapping the utilization of the land. *Geographical Review* **8**: 47–54

Silver J 1978 Bureau of the Census—applied geography. *Applied Geography Conference SUNY Binghamton* **1**: 80–92

Taylor P 1985 The value of a geographic perspective. In: Johnston R J (ed.) *The Future of Geography*. Methuen, London, pp. 92–110

Torrieri N K, Radcliffe M R in press Applied geography. In: Gaile G L, Willmott C J (eds.) *Geography in America at the Dawn of the Twenty-First Century*. Oxford University Press, Oxford, UK

Wellar B 1998a Combining client-driven and curiosity-driven research in graduate programs in geography: some lessons learned and suggestions for making connections. *Papers and Proceedings of Applied Geography Conferences* **21**: 213–20

Wellar B 1998b *Walking Security Index*. DOT Regional Municipality of Ottawa-Carleton (RMOC), Canada

White G F (ed.) 1974 *Natural Hazards*. Oxford University Press, New York

Wright J K 1952 *Geography in the Making. The American Geographical Society*. American Geographical Society, New York, pp. 2851–1951

J. Frazier

Apportionment: Political

Most contemporary students of democratic theory take for granted that the basis of political representation will be geographic. There are two key components of any geographic system of representation: apportionment and districting. While the two terms are often used synonymously, formally, *apportionment* refers to the determination of the number of representatives to be allocated to pre-existing political or geographic units, while *districting* refers to how lines are drawn on a map within those units to demarcate the geographic boundaries of individual constituencies. *Malapportionment* refers to differences in the ratio of number of voters/electors to number of representatives across different constituencies. *Gerrymandering* refers to the drawing of districting lines for purposes of political (e.g., partisan or ideological or ethnic) advantage/disadvantage.

Use of geographical districting leaves open many key questions: How many and how large are the districts to be? Will seats be allocated to whole political units, such as provinces or towns, or will district lines be permitted to cut across existing political sub-unit boundaries? Will district lines be required to satisfy standards of compactness or contiguity? To what

extent will apportionment and districting lines be based (entirely or almost entirely) on total population? Or on population of (eligible) voters?

The USA has been a leader in defining standards of apportionment and districting to implement the principle of popular sovereignty. The US House of Representatives was intended by its founders to be the representative chamber of a bicameral legislature and its apportionment rules were set up to require a purely population-based allocation of House districts to the states, with changes in seat allocations made after each decennial census. Indeed, the various apportionment methods that have been used for the House over the past several centuries are mathematically identical to proportional representation methods such as N'Hondt and Ste. Laguë (Balinski and Young 1982; also see *Electoral Systems*). In Baker vs. Carr, 369 US 186 (1962), the US Supreme Court held that failure to redraw district lines when new census data was available was unconstitutional and that courts could fashion appropriate remedies. In subsequent landmark districting decisions, such as Reynolds vs. Sims 377 US 533 (1964), the US Supreme Court went much further, proclaiming '*one person*, *one vote*' as the only appropriate standard for both districting and apportionment.

While one person, one vote notions of representation have had a profound influence throughout the world, by and large, the USA remains extreme among nations in its insistence on strict adherence to one person, one vote standards. For state legislative and local redistricting plans, where the one person, one vote standard is derived primarily from the equal protection clause of the Fourteenth Amendment, Supreme Court cases in the USA have established a 10 percent total deviation as prima facie evidence of constitutionality. (*Total deviation* is the sum of the absolute values of the differences between actual district size and ideal district size of the largest district and the smallest districts, normalized by dividing through by ideal district size.) For congressional districting, where standards are based directly on the supposed meaning of language in Article I of the Constitution, the Supreme Court has held that districts must be as close to zero deviation as is practicable. For example, in Karcher vs. Daggett, 462 US 725 (1983) a congressional plan with a total deviation of only 0.698 percent was invalidated. In reaction to this ruling, in some 1990s congressional plans, districts were drawn that were equal in population to within a handful of persons. In contrast, in other countries, especially those using plurality elections, no such strict population requirements exist. Many countries require (or even only just suggest) that differences should be no greater than plus or minus 25 percent or plus or minus 50 percent of ideal (Butler and Cain 1992).

However, the notion that near perfect equality of political representation has been achieved in the USA is misleading. The grossly malapportioned US Senate tends to be omitted from international comparisons despite the fact that it is a co-equal chamber. The US House requires that each state have at least one representative—a rule that usually gives representation to several small states who would not otherwise be entitled to seats. Also, since states are the units, 'rounding rules' create variation in average House district population across states. For example, based on 1990 census figures, the largest House district in the 1990s apportionment was 1.7 times the size of the smallest House district, and the House had a total deviation of 61 percent (based on absolute deviations from the ideal size of 572,465 of 231,289 (Montana, too many) plus 118,465 (Wyoming, too few)). The discrepancies have been even greater in earlier apportion ments. Moreover, even districts that are equal in population need not be equal in terms of (eligible) voters. Perhaps most importantly, unless we somehow regard voters as completely interchangeable units, neither population nor voter equality across constituencies, however perfect, guarantees equality of effective representation of the disparate groups and interests within a society.

The degree and geographic locus of malapportionment and differential turnout across groups interact with how a group's voting strength is distributed across districts to affect the translation of a group's voting strength into actual electoral impact (Grofman *et al.* 1997). Indeed, malapportionment is sometimes referred to as a form of '*silent gerrymander*,' since malapportionment can easily translate into the political disadvantage of groups whose influence has been diminished because their members are disproportionally concentrated in constituencies whose voters have been underrepresented relative to their numbers. Even without conscious gerrymandering, the way in which districting lines are drawn will necessarily have an impact on the representation of different parties or groups (Dixon 1968).

The term gerrymandering comes from word play on the last name of Elbridge Gerry, Governor of Massachusetts. In 1812, Gerry signed into law a districting plan for the Massachusetts Senate, allegedly designed to maximize the electoral successes of Republican-Democrat candidates and minimize the electoral successes of Federalist candidates, which included some rather strangely shaped districts. In a map in the Boston Gazette of March 26 1812, the strangest of these districts was shown as a salamander, given tongue and teeth (Fig. 1). Perhaps the most pernicious aspect of this figure is that it has led to a potentially misleading association of gerrymanders with oddly shaped districts. The defining aspect of a gerrymander is the political consequences it entails, not its shape. Political disadvantage can come about even when districts look like squares or hexagons (Grofman 1990). In fact, however, the 1812 Senate plan did achieve partisan advantage for the Republican–Democrats; in the next election they won

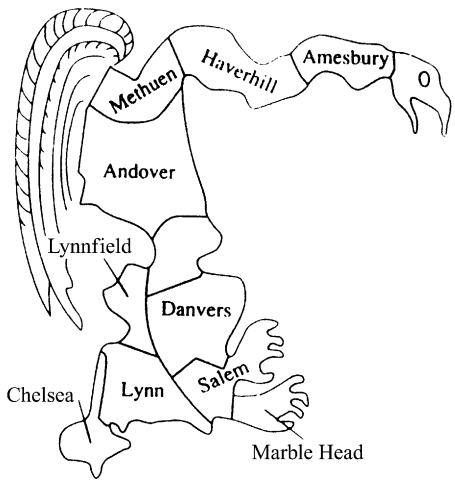

Figure 1

29 of 40 seats even though they received less than half of the votes (Hardy 1990).

Gerrymanders can be classified as partisan, bipartisan (often called 'incumbent gerrymanders'), racial, and personal, depending on who can be expected to be harmed or helped. In the USA, for example, the debate about gerrymandering has been fought largely over racial rather than partisan issues, e.g., over the extent to which plans should seek to place members of historically disadvantaged groups such as African-Americans into districts where they comprise the majority of the population even if doing so meant drawing districts that were irregular in appearance or cut across municipal and other political unit boundaries (Grofman 1998).

There are two basic techniques of gerrymandering: (a) 'packing' members of the group that is to be disfavored into districts that are won by very large majorities, thus 'wasting' many of that group's votes; and (b) 'cracking' the voting strength of members of the group by dispersing the group's population across a number of districts in such a fashion that the group's preferred candidates will command a majority of the votes in as few districts as possible. In addition, if elections are held under plurality, a group's voting strength may be submerged in multimember districts that use bloc voting—a technique sometimes called 'stacking.' The terms 'affirmative action gerrymander,' and 'benign gerrymander' have been used to denote districting done to advantage members of a historically disadvantaged group. However, it is important to distinguish between plans that are drawn with an aim to create a level playing field by avoiding unnecessary fragmenting of minority population concentrations, but that otherwise generally take into account the usual districting criteria such as respect for natural geographic boundaries and historical communities of interest, and plans that seek to specially privilege particular groups by totally disregarding features other than race in drawing lines.

Because the way in which lines are drawn can be expected to matter, an important issue has to do with who draws the lines. In most democracies, especially those electing under plurality, non-partisan boundary commissions are responsible for drawing district lines (Butler and Cain 1992). In the USA, the preponderant pattern is for a legislature to be responsible for its own redistricting, and for each state legislature to be responsible for the drawing of congressional district lines for its state. However, in most US legislatures, no plan can be passed without gubernatorial agreement. Because of divided party rule and other factors, states may be unable to reach agreement on plans, thus throwing decision-making into the courts. One way in which districting practices in the USA are distinct from those in other countries is the extent to which courts play a critical role as arbiter. In recent decades, all but a handful of states have had a legislative or congressional plan challenged in court, and many plans have been rejected—in the 1960s and 1970s mostly for reasons having to do with population inequalities across districts, in the 1980s and 1990s for reasons having to do with racial representation (Grofman 1998). Indeed, throughout these decades, courts themselves were responsible for drawing some of the legislative or congressional districting plans that were actually used. Another peculiarity of US districting practices is the role of the US Department of Justice under Section 2 and Section 5 of the Voting Rights Act of 1965 as amended in 1982 (Grofman 1998).

From a comparative perspective, we may say that, generally speaking, gerrymandering is more important in plurality elections than in elections under proportional or semi-proportional rules. In particular, when there are more than two candidates or political parties competing, districting can have a dramatic impact on outcomes in plurality elections Taylor et al. 1986). *Ceteris paribus*, for elections under proportional or semi-proportional methods, the larger the average *district magnitude* (the number of representatives to be elected from the constituency), the less the probable impact of districting choices on outcomes; in contrast, for elections under plurality voting, the greater the average *district magnitude* (the number of representatives to be elected from the constituency), the greater is the expected impact on outcomes, since *plurality bloc voting* (the extreme case of which is an *at-large election*) can result in the virtual submergence of the views of those in the minority. However, even under proportional representation, expected outcomes can still be manipulated by districting choices, especially choices as to district magnitude (Mair 1986).

See also: Electoral Geography; Electoral Systems; Latin American Studies: Politics; Political Geography;

Political Parties, History of; Political Representation; Proportional Representation

Bibliography

Balinski M L, Peyton Young H 1982 *Fair Representation: Meeting the Ideal of One Man. One Vote*. Yale University Press, New Haven, CT

Butler D, Cain B 1992 *Congressional Redistricting*. Macmillan, New York

Dixon R G 1968 *Democratic Representation: Reapportionment in Law and Politics*. Oxford University Press, New York

Grofman B 1990 Toward a coherent theory of gerrymandering: Bandemer and Thornburg. In: Grofman B (ed.) *Political Gerrymandering and the Courts*. Agathon, New York, pp. 29–63

Grofman B (ed.) 1998*Race and Redistricting*. Agathon, New York

Grofman B, Koetzle W, Brunell T 1997 An integrated perspective on the three potential sources of partisan bias: malapportionment, turnout differences, and the geographic distribution of party vote shares. *Electoral Studies* **16**(4): 457–70

Hardy L 1990 *The Gerrymander: Origin, Conception and Re-Emergence*. Rose Institute of State and Local Government, Claremont McKenna College, Claremont, CA

Mair P 1986 Districting choices under the single-transferable vote. In: Grofman G, Lijphart A (eds.) *Electoral Laws and Their Political Consequences*. Agathon, New York, pp. 289–308

Taylor P, Gudgin G, Johnston R J 1986 The geography of representation: A review of recent findings. In: Grofman B, Lijphart A (eds.) *Electoral Laws and Their Political Consequences*. Agathon, New York, pp. 183–92

B. Grofman

Apprenticeship and School Learning

Apprenticeship models provide a view of school learning processes that is quite different from traditional models. In particular, apprenticeship learning is less prone to the inert knowledge phenomenon. In this article, the main characteristics of apprenticeship learning and its theoretical background are described, and pros and cons of apprenticeship learning vs. traditional school learning are discussed.

1. Apprenticeship Models as Solutions for Problems with School Learning

Qualification and integration-enculturation are amongst the most important functions of schools for society as well as for individuals. School learning therefore should teach students facts and skills that are necessary for later life. However, evidence exists that school learning is often far removed from application situations out of school; critics have argued that it frequently leads to knowledge that is inert (Bransford et al. 1991) and thus cannot be used for solving real-world problems. A plausible reform idea is to make school learning resemble learning out of school. Resnick (1987) identified four principles in which learning in school and out differ:

(a) Individual cognition in school vs. shared cognition outside. Learning in school focuses on individual performance, students have to work in isolation; in contrast, learning out of school usually focuses on shared knowledge and cooperative problem solving.

(b) Pure mentation in school vs. tool manipulation outside. School learning focuses on abstract mental activities that have to be done without the use of any external support like one's own notes or Internet research systems; in contrast, learning out of school heavily relies on individuals' competence to use adequate tools in an adequate way.

(c) Symbol manipulation in school vs. contextualized reasoning outside school. With its focus on symbol-based reasoning, school learning often lacks close connections to events and objects in the daily world that are characteristic for learning out of school.

(d) Generalized learning in school vs. situation-specific competencies outside. School learning aims at the acquisition of general, widely usable principles, whereas learning out of school focuses on solving problems that actually arise at places and contexts the individual is situated in.

The analysis of discrepancies between learning in school and out includes criticisms of present school instruction concerning both the qualification function of school and its integration-enculturation function. Similar arguments were brought forward by the German *Reformpädagogik* (educational reform) at the beginning of the twentieth century. Alternative instructional models developed in these years are closely related to the proposals nowadays made by situated learning theorists. One common principle is that students participate in more advanced individuals' activities and thus increasingly become part of the community of practice. In such apprenticeship models of learning, besides knowledge students have to acquire the ways of thinking in communities of practice.

Since the early 1990s, a number of situated learning models have been developed in order to decrease the discrepancy between learning in and out of school and to avoid the acquisition of inert knowledge (Gruber et al. 2000). In each, students learn within complex contexts like apprentices by solving authentic problems in a community of practice. The approaches are based on the idea that knowledge is socially shared so that plain teaching of 'objective' knowledge does not suffice because each working situation includes de-

mands for adaptive problem solving (Derry and Lesgold 1996). The view that knowledge in principle is situated in the environment within which it was acquired is the perspective of situated learning.

2. Theoretical Background: 'Reformpädagogik,' Situated Learning

Many aspects of the mentioned situated learning approaches resemble concepts of the *Reformpädagogik*. When the agents of this German educational movement expressed their ideas a century ago, they were well aware of developments in the Deweyan Chicago school (e.g., the project plan; Kilpatrick 1922), and vice versa. *Reformpädagogik* criticized the 'book school' in a similar way school learning nowadays is criticized. The central counter-idea then was the 'work school,' most prominently proposed by Kerschensteiner (1854–1932). Kerschensteiner broadly reorganized the elementary and vocational school system; the German dual system in vocational learning emerged from his work. Kerschensteiner (1912) emphasized the importance of manual work which should be closely tied to mental work; among his school reforms was the introduction of kitchen and garden instruction in schools. Based on authentic activities, students learned chemistry, physics, physiology, and mathematics. Instructional questions usually were not posed by the teacher, but rather by the students when working on problems. Such self-regulated activity reduced the discrepancy between school and everyday life: work school no longer was an *Ersatzwelt* (artificial world) in which ersatz activities had to be carried out. Since the problems posed were of relevance to the students' life outside school, besides skills and knowledge, learners acquired socially conscious attitudes and a sense of responsibility as members of society.

Later, Kerschensteiner became extremely concerned about the question of whether school prepared for later professional life. He thus initiated the German vocational training system, the dual system, which is a form of training tied to the workplace with supplementary teaching in the compulsory, part-time vocational school. In this system, not only manual work is stressed: the notion of work can also be denoted to cognitive learning processes or, in Gaudig's term, 'free mental work.'

Many of the *Reformpädagogik* ideas were theoretically reconsidered in the situated learning movement under cognitive psychological and constructivist perspectives. Five major instructional principles can be identified:

Learning by solving authentic problems. In problem-oriented learning, an authentic, rather complex problem marks the starting point. Its function is not only motivating students, but also, by embedding the learning process in meaningful contexts, knowledge acquisition through application instead of abstract teaching. Knowledge is conditionalized to application conditions from the beginning.

Multiple perspectives. Domains are analyzed from multiple perspectives and multiple contexts in order to foster multiple conditionalization. Preventing the emergence of dysfunctional oversimplifications and fostering active abstraction processes are the main instructional functions, thus supporting flexible use of knowledge and skills.

Articulation and reflection. Externalizing mental processes enables students to compare their own strategies with those of experts and other students. Articulation facilitates reflection and thus fosters the decontextualization of acquired knowledge.

Cooperative format. Cooperative learning facilitates the comparison with experts and peer learners. In addition, competencies for cooperation are useful in later professional life.

Learning as enculturation. A particular function of learning is to become enculturated into communities of practice. To achieve this, students have to acquire a sense of authentic practice. They have to acquire functional belief systems, ethical norms, and the tricks of the trade.

These principles of situated learning have been realized in several instructional models, most convincingly in the model of cognitive apprenticeship.

3. Microanalytic Use of Apprenticeship in School Learning: The Model of 'Cognitive Apprenticeship'

The apprenticeship metaphor for designing situated learning arrangements has its origin in skilled trade domains, e.g., tailoring or midwifery. Collins et al. (1989) adapted the concept for cognitive domains. In 'cognitive apprenticeship' as in trade apprenticeship, students are introduced into an expert culture by authentic activities and social interactions. Important is the sharing of problem-solving experiences between students and mentors (experts, teachers, or advanced students), thus negotiating their understandings and actions through dialogue. Since in cognitive domains mental activities prevail, the explication of cognitive processes is extremely important in order to publicly expose knowledge and thinking processes involved in the cooperative problem solving (Derry and Lesgold 1996).

The core of cognitive apprenticeship is a particular instructional sequence. Learning takes place in sequenced learning environments of increasing complexity and diversity. In the early stages of learning, mentors provide overall direction and encouragement; as students improve, mentors gradually withdraw support, encouraging students to work and think more independently. At all phases of apprenticeship, the mentor is assigned an important role as a model and as a coach providing scaffolding. However, the

student has to increasingly take over an active role, as the mentor gradually fades out. Articulation and reflection are promoted by the mentor in order to persist on externalization of cognitive processes.

This instructional sequence supports students in increasingly working on their own (exploration) and taking over the role initially held by the mentor. Lave and Wenger (1991) similarly described the development of competence in nonexplicit learning contexts as development from legitimate peripheral participation to full participation. Learning is thus not confined to the acquisition of knowledge or skills, but is a social process of enculturation by becoming a full participant in a community of practice that can cope with the problems typical for the domain in a flexible manner. Considering learners as apprentices ascribes them a novel role: From the beginning, they are active participants in authentic practices instead of passive recipients (Gruber et al. 1995).

Empirical evidence exists that cognitive apprenticeship learning supports the acquisition of applicable knowledge in a variety of domains (e.g., programming, medicine). However, successful learning is not guaranteed. High demands are made on learners, so that careful instructional support is necessary (Stark et al. 1999).

4. *Macroanalytic Use of Apprenticeship in School Learning: Communities of Practice—Implementation of Apprenticeship in School Systems*

Cognitive apprenticeship stresses that learning performance, like all authentic cognitive activity, is socially constituted. Learning environments inevitably are part of a complex social system, and the structure and meaning of learning processes are substantially influenced by that system. It is unlikely that an adequate understanding of learning processes can be developed without taking the social context into consideration in which learning processes are situated. Conceiving learning as becoming enculturated into communities of practice thus concerns not only the individual processes during learning, but also the implementation of learning environments in large social contexts, frequently denoted as the prevailing 'learning culture.' Taking the apprenticeship idea seriously, such macroanalytic concepts of learning are inseparably connected with microanalytic analyses of learning processes. In a study of insurance claims processors at work, Wenger (1990) showed that there are large discrepancies between the official agenda of insurance management and what was actually learned and practiced at workplaces. Wenger concluded that knowledge and expertise cannot be understood separately from the social environment in which they are observed. A consequence for instruction is that learning has to be bound to application situations and has to be integrated into a large system with adequate learning culture.

The principle of the dual system of vocational training in Germany tries to fulfill this requirement. It includes simultaneous qualification at two learning locations: vocational school and enterprise. Thus school learning and professional practice are closely tied. Winkelmann (1996) compared the experience of apprenticeship graduates to that of graduates from universities, full-time vocational schools, and secondary schools when entering the job. He showed that apprentices experienced fewer unemployment spells in the transition to their first full-time employment than did nonapprentices. One reason is the interaction of different social components in the training process.

Efforts to implement the dual system in other countries (Harthoff and Kane 1997, Heikkinen and Sultana 1997) showed that many organizational preconditions are required, such as flexibility of enterprises in the adoption of the system, an elaborated widespread school system, adequate teacher education, etc. The dual system cannot serve as a panacea if the learning culture within a society is not prepared. Additionally, modern developments in workplaces—globalization of enterprises, new principles of working organization, etc.—yield difficulties for the dual system. New approaches in educational research are being developed as a response to emerging imbalances between the learning locations school and enterprise. For example, 'discipline and location-crossed training' aims to implement new ways of cooperation between learning locations.

5. *Outlook: What Remains of 'Traditional' School Learning?*

The view that knowledge is essentially situated has major educational implications. Learning is regarded as a process of enculturation in order to take part in a particular community of practice, to get acquainted with the culture of the community, the jargon used, the beliefs held, the problems raised, and the solving methods used. This entails the notion of apprenticeship to learning.

Traditional school learning is accused of producing inert knowledge as students fail to transfer their knowledge to tasks embedded in contexts other than school settings. Lave (1988) argued that traditionally school is treated as if it were a neutral place, with no social and cultural features of its own, in which competencies are acquired that easily can be transferred later to any other situation.

As mentioned above, there is evidence for the inert knowledge phenomenon. However, this evidence does not necessarily require the importance of traditional school learning to be denied. Even if there exist resources other than teaching through which apprentices acquire competences, apprenticeship is not the

only effective means of learning. The academic context of school learning proves useful for a variety of learning situations. It is very likely that an interaction exists between learning purposes and preferable learning contexts. Educational psychology has to decide carefully which kind of learning context is appropriate for which kind of learning, and should use the results of respective analyses for the design of school learning as well as for the design of teacher education.

Thus, the notion that 'traditional' school learning should be entirely abolished cannot be maintained. However, it is worth analyzing under what conditions apprenticeship models are an attractive alternative to traditional school learning. For which kinds of competence and knowledge, in terms of their complexity (elementary, intermediate, sophisticated) and nature (declarative, procedural, strategic), is apprenticeship deemed necessary? By what means can students obtain maximal benefits? These questions must be addressed in future research in order to determine the role apprenticeship can and should play in future so that an adequate balance between traditional school learning and situated learning is found.

See also: Apprenticeship: Anthropological Aspects; Cooperative Learning in Schools; Education: Skill Training; Educational Research and School Reform; Environments for Learning; Pedagogical Reform Movement, History of; Progressive Education Internationally; School (Alternative Models): Ideas and Institutions; School Learning for Transfer; Simulation and Training in Work Settings; Situated Learning: Out of School and in the Classroom

Bibliography

Bransford J D, Goldman S R, Vye N J 1991 Making a difference in people's ability to think: reflections on a decade of work and some hopes for the future. In: Sternberg R J, Okagaki L (eds.) *Influences on Children*. Erlbaum, Hillsdale, NJ, pp. 147–80

Collins A, Brown J S, Newman S E 1989 Cognitive apprenticeship: teaching the craft of reading, writing and mathematics. In: Resnick L B (ed.) *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*. Erlbaum, Hillsdale, NJ, pp. 453–94

Derry S, Lesgold A 1996 Toward a situated social practice model for instructional design. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Macmillan, New York, pp. 787–806

Gruber H, Law L-C, Mandl H, Renkl A 1995 Situated learning and transfer. In: Reimann P, Spada H (eds.) *Learning in Humans and Machines: Towards an Interdisciplinary Learning Science*. Pergamon, Oxford, UK, pp. 168–88

Gruber H, Mandl H, Renkl A 2000 Was lernen wir in Schule und Hochschule: Träges Wissen. In: Mandl H, Gerstenmaier J (eds.) *Die Kluft zwischen Wissen und Handeln: Empirische und theoretische Lösungsansätze*. Hogrefe, Göttingen, Germany, pp. 139–56

Harthoff D, Kane T J 1997 Is the German apprenticeship system a panacea for the US labor market? *Journal of Population Economics* **10**: 171–96

Heikkinen A, Sultana R G (eds.) 1997 *Vocational Education and Apprenticeships in Europe. Challenges for Practice and Research*. Tampereen Ylioposto, Tampere, Finland

Kerschensteiner G 1912 *Begriff der Arbeitsschule*. Teubner, Leipzig, Germany

Kilpatrick W H 1922 *The Project Method: The Use of the Purposeful Act in the Educative Process*. Teachers College Press, New York

Lave J 1988 *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life*. Cambridge University Press, Cambridge, UK

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK

Resnick L B 1987 Learning in school and out. *Educational Researcher* **16**: 13–20

Stark R, Mandl H, Gruber H, Renkl A 1999 Instructional means to overcome transfer problems in the domain of economics: empirical studies. *International Journal of Educational Research* **31**: 591–609

Wenger E 1990 Toward a theory of cultural transparency. Unpublished Ph.D. dissertation, University of California, Irvine, CA

Winkelmann R 1996 Employment prospects and skill acquisition of apprenticeship-trained workers in Germany. *Industrial & Labor Relations Review* **49**: 658–72

H. Gruber and H. Mandl

Apprenticeship: Anthropological Aspects

Anthropological treatments of apprenticeship have until relatively recently been sporadic and piecemeal. Lately, however, there has been a surge of interest in the topic, and since the mid-1980s it has received much more sustained attention. Three themes emerge from the literature as predominant anthropological approaches to apprenticeship: first, as a form of social organisation occurring in a variety of social and cultural contexts; second, as an anthropological method of field research; third, as a domain for the analysis of social, cognitive, and bodily processes of learning. These aspects will be dealt with in turn below, but first some remarks to indicate how the concept has been formulated in the West.

The subject of apprenticeship comes with a good deal of conceptual baggage derived from European and North American history. An institution known in ancient Greece and Rome, as well as from the early history of the Middle East, apprenticeship in Europe in the Middle Ages was regulated by craft guilds; specifically in Britain under sixteenth-century Elizabethan statutes which were later repealed in the early nineteenth century. As a nineteenth-century institution, it has received a particularly bad press in Britain and the United States, where it was used as means of poor relief for pauper children who were bound and

indentured into trades so as to relieve local communities of the burdens placed on them by the poor. The lives of such children have been described as 'at best a monotonous toil,' 'at worst a hell of human cruelty.'

Anthropological perspectives on apprenticeship and the use of child labor in non-Western cultures are confronted today by a not dissimilar problem: what relationship should be drawn between the condemnation of the use of child labor and of the frequently accompanying economic poverty, on the one hand, and the recognition, on the other hand, that in some cultural contexts children are not excluded practically or ideologically from relations of production (see Nieuwenhuys 1996). The dilemma over how to approach this problem is informed in part by the history of the institution in the West, and by our knowledge of the abuses to which it has been put in the past. In order for apprenticeship to be viewed as a cross-cultural concept, therefore, it has to be disentangled from our own conceptual baggage. This is not a simple or easy task.

Apprenticeship is a key institution, then, in the West's history of craft production, of capitalist industrialized manufacture, and of the transition between the two. Whilst Adam Smith regarded it as an archaic institution whose abolition should be welcomed, Marx saw it as a form of organisation that protected a skilled workforce from the logic of machine manufacture and from the subsequent deskilling of labour (Marx 1887/1947).

Apprenticeship, viewed in the context of craft production or of industrial manufacture is a mode of organization for acquiring trade skills and for the supply of able practitioners in a trade. It also involves a relationship between a novice, who is usually a minor, and a master (usually not the novice's parent) who teaches a trade and who is recompensed for the instruction given in the form of the novice's products. Furthermore, the master stands *in loco parentis* to the novice, for they are responsible for the minor's moral development and general conduct, as well as for furnishing their board and lodgings.

1. *Apprenticeship as a Form of Social Organization*

Coy's edited collection on the topic (Coy 1989) was one of the first works by a group of anthropologists to treat apprenticeship explicitly as a social institution in comparative cross-cultural perspective. His definition of it echoes that given in the preceeding paragraph, but he suggests also that it is 'rite of passage' involving 'specialized' and 'implicit' knowledge; he emphasises too observation and participation as a mode of learning about a craft, about social relations, and about the social self and forms of cultural identity (1989 pp. 11–12) (see *Craft Production, Anthropology of*). Ethnographic examples in the volume range from north to south America, from Africa to the Far East. One of the major conclusions of the volume is that apprenticeship as a mode of social organization 'displays more similarity cross-culturally and historically than any of us realised' (Coy 1989 p. 15).

Goody (1989) goes on to theorise about these commonalities, arguing that apprenticeship is a characteristic of social systems undergoing increasing differentiation in a division of labor that has breached the limits of domestic, kin-based production; systems that are undergoing structural changes by virtue of entry 'into the market.' Whilst there is debate about the factors that determine the similarity of forms that apprenticeship takes across the world, some anthropologists have highlighted as well the unique features it takes in particular contexts (see Singleton in Coy 1989, Kondo 1990).

Coy's volume sets a contemporary benchmark for developments in the study of apprenticeship: first, it provided a set of comparable ethnographic descriptions of apprenticeship systems in a range of cultures; second, it examined in depth the idea of apprenticeship as an anthropological field method. More recent work on the topic has moved debates along in a number of directions. For example: Is apprenticeship a means of imparting knowledge or of restricting it? (cf. Singleton in Coy 1989 on 'stealing the master's secrets') What are the relationships of dominance operating within, and indeed outwith, apprenticeship systems? (Herzfeld 1995). Is there an objective body of knowledge that is passed on to novices or is knowledge negotiated and situational? (Lave and Wenger 1991, Keller and Keller 1996).

The first two questions introduce the issue of power into the organization of apprenticeship, viewed as a means of the transmission of knowledge and skills to future generations. Training apprentices produces potential future competition in the form of skilled practitioners who might one day, once they are economically independent, take trade away from their former master. These issues must be seen also within the purview of the wider political economy in which the reproduction of a trade does or does not take place.

The third question about the form that knowledge takes relates to whether specialists possess a body of knowledge that exists in the conventional terms in which we often think about it. Even if knowledge does not exist in the form of books, databases and so on, but in the form of oral traditions, does the achieving of knowledgeable practice rest on the acquisition of the latter? In many cases of apprenticeship, very little seems to be passed verbally from master to novice. So in which case what is being passed on? The development of knowledgeable practice might then be viewed in the framework of interaction and negotiation between two parties, and might too be connected with bodily praxis: a topic dealt with in Sect. 3.

The significance of the master–apprentice relationship is another topic of debate, especially whether it should be seen as a central characteristic of apprenticeship. Whether apprenticeship is a formal or an informal arrangement, many studies emphasise the relation between teacher and pupil as crucial to the processes and the experience of learning (e.g., Stoller and Oikes 1987 and Singleton in Coy 1989). Cooper (1989) describes the constant criticism by his master of his own practice as an apprentice woodcarver, and it seems that the theme of discomfort between master and pupil is one which recurs in many other studies, too (e.g., Marchand 2002). However, following Lave and Wenger's suggestions (1991) about a 'decentred' approach that focuses on 'communities of practice' rather than single dyadic relations, some analyses have now shifted emphasis away from the master–apprentice dyad to focus on whole communities of actors (e.g., Palsson 1994). The question raised here concerns how much apprentices actually learn directly from a master and how much is simply absorbed through participation with a group of skilled practitioners. There seems to be a degree of variation in the emphasis placed on master–novice relations depending on the kind of specialism being learned.

2. *Apprenticeship as an Anthropological Field Method*

The second aspect of apprenticeship is as a method of anthropological field research (see *Fieldwork in Social and Cultural Anthropology*). An increasing number of field researchers have apprenticed themselves to practitioners of a variety of crafts, trades, and other specialisms in order to acquire particular social and practical skills, and to gain a deeper insight to cultural practice and processes; in short, to learn about cultural learning (see e.g., Coy 1989, Keller and Keller 1996, Marchand 2002, Stoller and Oikes 1987). The intense interaction entailed by such participation provides a point of entry into a community and a way of learning through practice. Types of knowledge, secrets, and specific skills are also accessed, and these might otherwise remain hidden to investigations by other methods. Tedlock (1992) describes how she and her husband used the method of apprenticeship successfully to learn the art of Mayan divination.

Apprenticeship as a field method is not without its critics, who point out the methodological dangers of compromising 'objectivity' and succumbing to partiality. Few anthropologists, however, rely totally on this method and often use other techniques to investigate wider, macrosocial processes beyond the workshop. Cooper discusses the 'fiction' of apprenticeship: the requirement for fieldworker and fellow practitioners to enter into a pretence and to know when the pretence ends. A kind of 'enforced schizophrenia prevails' (Cooper 1989, p. 138). The 'enforced schizophrenia' inherent in the method allows for a continual methodological movement in and out of roles, and indeed sets up a distance from as well as an engagement with the subjective and experiential aspects of being apprenticed. Apprenticeship has of course much in common with traditional participant-observation, in which the fieldworker is as much a kind of 'cultural apprentice' as he or she is a detached observer (see for e.g., Coy 1989, Goody 1989, Keller and Keller 1996, Palsson 1994 on this comparison). 'Apprentice participation' is, thus, an extension of traditional anthropological methods of participation and observation, and is pertinent to the investigation of particular kinds of activities that may be closed, specialized or in some way inaccessible.

3. *Apprenticeship and Forms of Learning*

The third aspect of apprenticeship is the way it has been linked to the anthropology of education (see Pelissier 1991), and to the analysis of cultural theories of learning (see *Education: Anthropological Aspects*). Lave and Wenger (1991) have reworked this area to propose a shift away from the study of apprenticeship *per se* to the more general idea of 'situated learning' and 'legitimate peripheral participation,' in which the learning process is embedded in 'communities of practice.' The analysis of cognitive activity is firmly placed back in the domain of everyday life. Their work reviews a range of learning contexts from Yucatec midwifery to the US navy, and it raises questions about the relationship between participation and knowledge.

This relationship can be seen as problematic in that much of the knowledge learnt in such contexts is not verbalized and indeed is nonpropositional in nature. What is learnt is gained through mimesis, practice, and repetition of tasks performed routinely by skilled practitioners. Bodily knowledge, bodily techniques, and aesthetics now become the focus of study (see *Body: Anthropological Aspects*). Examples of this type of approach are Marchand's study of Yemeni master builders (Marchand 2002), and especially Wacquant's examination (1998) of his own apprenticeship as a boxer among prize-fighters in Chicago. Novices learn through their bodies as well as their minds, and the bodily discipline of a trade creates moral and aesthetic sensibilities. Moreover, these learning processes are not just about practical skills or producing objects, but are also concerned with the creation of cultural identities and selves. As Kondo points out with regard to Japanese artisans, work involves self-realization, a 'polishing of self through hardship'; for 'a mature artisan is a man who, in crafting fine objects, crafts a finer self' (Kondo 1990, p. 241). These studies suggest, therefore, that the rigid Cartesian distinction between mind and body, thought and action needs to be critically re-examined.

The concept of apprenticeship is a matrix connecting a set of keystone issues within the contemporary frame of social and cultural anthropology. When employed in cross-cultural comparison, it demands explanations for its apparent commonalities across a wide range of social settings as well as for its cultural particularities specific to place or time. The practitioners of apprentice field methods have raised questions about the delicate balance between subjective participation and distanced observation that lie at the heart of empirical fieldwork procedures: indeed, it suggests a way of bridging this distinction. As a domain of enquiry, apprenticeship highlights cultural conceptions of knowledge and the processes of learning, of embodiment and bodily praxis, of self and identity formation. It highlights too the need to examine the connections between knowledge and power in dynamic interactive contexts embracing both individual practitioners situated in different social roles and communities of skilled actors whose practices create wider webs of relations. These concerns about knowledge and power extend, moreover, beyond the parochial boundaries of the workshop or place of learning to the broad social world in which apprenticeship takes place.

See also: Apprenticeship and School Learning

Bibliography

Cooper E 1989 Apprenticeship as field method: Lessons from Hong Kong. In: Coy M W (ed.) *Apprenticeship: From Theory to Method and Back Again*. SUNY, Albany, NY, pp. 137–48

Coy M W (ed.) 1989 *Apprenticeship: From Theory to Method and Back Again*. SUNY, Albany, NY

Goody E N 1989 Learning, apprenticeship and the division of labour. In: Coy M W (ed.) *Apprenticeship: From Theory to Method and Back Again*. SUNY, Albany, NY, pp. 233–56

Herzfeld M 1995 It takes one to know one. In: Fardon R (ed.) *Counterworks: Managing the Diversity of Knowledge*. Routledge, London, pp. 124–42

Keller C M, Keller J D 1996 *Cognition and Tool Use: The Blacksmith at Work*. Cambridge University Press, Cambridge, UK

Kondo D 1990 *Crafting Selves: Power, Gender and Discourses of Identity in a Japanese Workplace*. University of Chicago Press, Chicago

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK

Marchand T 2002 *Minaret Building and Apprenticeship in Yemen*. Curzon, London

Marx K 1947[1887] *Capital: A Critical Analysis of Capitalist Production*, Vol. I. International Publishers, New York

Nieuwenhuys O 1996 The paradox of child labor and anthropology. *Annual Review of Anthropology* **25**: 237–51

Palsson G 1994 Enskilment at sea. *Man* **29**(4): 901–27

Pelissier C 1991 The anthropology of teaching and learning. *Annual Review of Anthropology* **20**: 75–95

Stoller P, Olkes C 1987 *In Sorcery's Shadow: A Memoir of Apprenticeship among the Songhay of Niger*. The University of Chicago Press, Chicago

Tedlock B 1992[1982] *Time and the Highland Maya*. University of New Mexico Press, Albuquerque, NM

Wacquant L 1998 The prizefighter's three bodies. *Ethnos* **63**(3): 325–52

R. M. Dilley

Apraxia

The concept of apraxia was elaborated by Hugo Liepmann (Liepmann 1908) in the early twentieth century. Liepmann noted that patients with left sided brain damage (LBD) committed errors when performing motor actions with either hand. This obvious deviation from the rule that each hemisphere controls the motor action of only the contralateral hand led him to conclude upon a general dominance of the left hemisphere for the control of motor actions. At that time it had already been established that the left hemisphere is dominant for comprehension and production of speech, and indeed most of Liepmann's apraxic patients were also aphasic, but he found some apraxic patients without aphasia and argued convincingly that faulty motor actions could not be referred to being a sequel of language impairment.

The nature of left hemisphere motor dominance and its relationship to language gave rise to various conflicting interpretations and remains an unsettled question after 100 years of reasearch. Any valid interpretation must take account of the fact that apraxia following LBD does not affect all kinds of motor actions. There is a striking discrepancy between fast and accurate performance of some motor action and hesitant and grossly erroneous performance of other actions which do not pose higher demands on the coordination of muscular innervations. Three kinds of actions are traditionally investigated for a clinical diagnosis of apraxia, because they yield clear manifestations of apraxic errors: imitation of gestures, demonstration of meaningful gestures, and use of tools and objects. This article will examine each of them on its own and then return to their implications for understanding hemisphere specialization of action control.

1. Imitation of Gestures

Faulty imitation of gestures has been said to prove that apraxia is a disorder of motor control and not a sequel of language disturbance or general asymbolia, that is, inability to comprehend and produce any signs or significations. The conclusion is strongest for faulty imitation of meaningless gestures. As these gestures have neither a verbal label nor a conventional signification, their imitation should be immune against disturbances of language or symbolic thought. More specifically it has been proposed that errors in

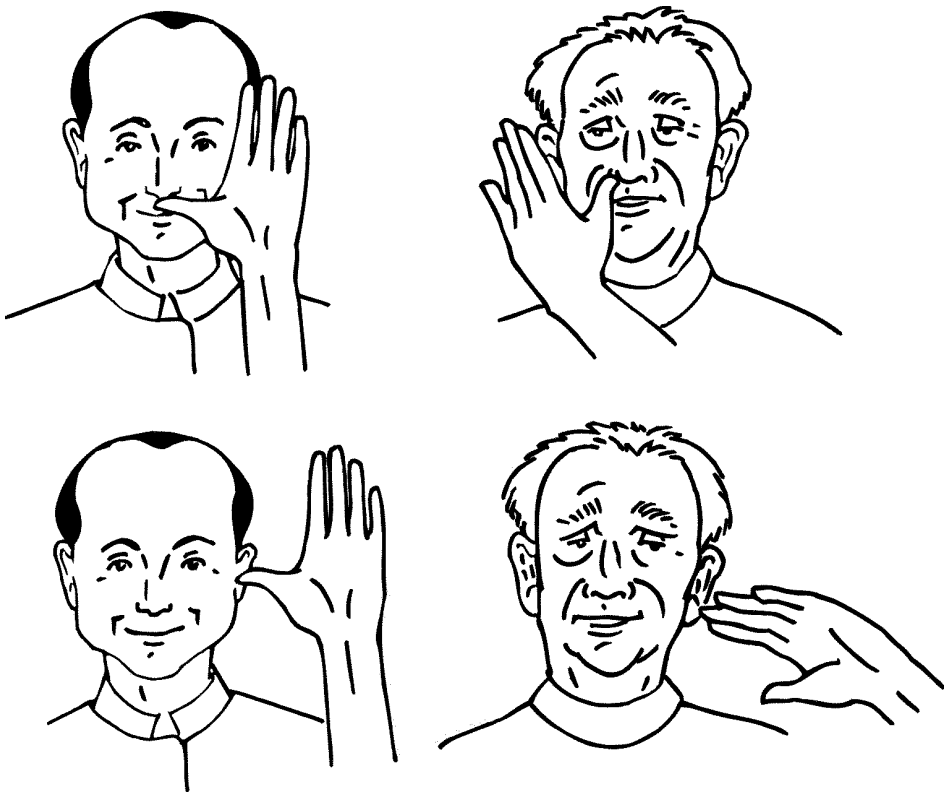

Figure 1
Imitation of hand postures by a patient with visuo-imitative apraxia. Left: model, right: imitation (reprinted from *Neuropsychologia*, **35**, Goldenberg G and Hagmann S 1997, The meaning of meaningless gestures: A study of visuo-imitative apraxia, pp. 333–341 Copyright 1997, with permission from Elsevier Science)

imitation testify disturbance of an executional or 'ideo-motor' stage of motor control succeeding to a conceptual or 'ideational' stage in which a plan of the intended action is formed (Barbieri and De Renzi 1988, Roy and Hall 1992). This proposal rests on the assumptions that the demonstration of the gesture for imitation leaves motor execution as the only possible source of errors. Defective imitation thus appears as evidence endorsing motor theories of hemisphere dominance which assume that there is a left hemisphere motor dominance which preceded and laid the ground for its language dominance (Liepmann 1908, Kimura and Archibald 1974). There are, however, several lines of evidence against this seducingly limpid interpretation of faulty imitation in apraxia.

The idea that imitation disorders arise at an executional stage of motor control predicts that patients who commit errors on imitation of meaningless gestures should encounter similar difficulties when performing meaningful gestures in response to a command specifying the meaning that is to be expressed by the gesture. There is no reason to assume that the motor implementation of gestures varies depending on whether the shape of the intended gesture is given by direct demonstration or by its meaning. This prediction is falsified by the observation of patients with 'visuo-imitative apraxia' who commit errors on imitation of meaningless gestures but not when demonstrating meaningful gestures. They may even achieve a correct imitation of meaningful gestures by first understanding their meaning and then reproducing them from long-term memory (Goldenberg and Hagmann 1997).

Kinematic studies of imitation show deviations from the normal profile of ballistic movements in patients with left brain damage and apraxia, but no correlation between the severity of these abnormalities and spatial errors of the finally achieved position (Hermsdörfer et al. 1996). There are even single patients who arrive at wrong final positions by kinematically perfect movements. This has led to the proposal that hesitancy, searching, and blocking of normal joint coordination are a reaction to ignorance of the exact shape of the intended gesture. This reaction may be absent in single apraxic patients who do not even note that they have not been able to build

up a correct representation of the demonstrated gesture and hence reach an incorrect target position with normal movements.

Further evidence that difficulties with imitation of meaningless gestures arise at a conceptual level preceding motor execution comes from the observation that patients who commit errors when imitating meaningless gestures commit errors also when asked to replicate these gestures on a mannikin or to match photographs of meaningless gestures demonstrated by different persons and seen under different angles of view (Goldenberg 1999).

The interpretation of left hemisphere dominance for imitation is further complicated by findings of defective imitation by patients with right brain damage (RBD). Whereas imitation of hand positions like those shown in Fig. 1 is affected exclusively by LBD, imitation of finger configurations, like those used for finger spelling in sign language, is affected by RBD even more than by LBD (Goldenberg 1999). RBD patients may also have difficulties when imitating sequences of gestures rather than single postures (Kolb and Milner 1981). Apparently, a left hemisphere contribution is necessary but not always sufficient for imitation.

The inconsistencies of a motor interpretation of LBD patients' difficulties with imitation motivated the revival and elaboration of an idea which had been put forward by Morlaas in 1928, held some popularity until the 1960s, but was then abandoned in favor of a return to Liepmann's original ideas (compare the articles on apraxia in the 1969 and 1985 edition of the *Handbook of Clinical Neurology* (De Ajuriaguerra and Tissot 1969, Geschwind and Damasio 1985). It was proposed that the left hemisphere contributes to imitation by coding meaningless gestures with reference to a classification of body parts (Goldenberg and Hagmann 1997, Goldenberg 1999). This classification reduces the multiple visual features of the demonstrated gesture to simple relationships between a limited number of body parts and accommodates novel and meaningless gestures to combinations of familiar elements. Furthermore, translating the gesture's visual appearance into relationships between body parts produces an equivalence between demonstration and imitation which is independent of the different modalities and perspectives of perceiving one's own and other persons' bodies. Absence of body part coding renders imitation an error prone to 'trial and error' matching between multiple visual details of perceived gestures, motor actions, and feedback about the own body's configuration (Goldenberg 1999).

Additional right brain involvement for some types of gestures can be referred to as demands on visuospatial analysis which may be lower for hand postures than for finger configurations. Hand postures are determined by relationships of the whole hand to perceptually salient body parts with very different shapes like the lips, the cheeks, or the ears. It is likely that demands on visuospatial analysis increase when, for example, finger postures require a distinction between extensions of index, middle, or ring finger.

2. *Meaningful Gestures*

Meaningful gestures serve communication. Their dependence on simultaneous verbal communication varies from gestures which emphasize or modulate the meaning of simultaneous oral speech to sign languages which are independent of, and can completely substitute for, oral language (McNeill 1992). The meaningful gestures which are usually examined for a diagnosis of apraxia lie in the middle of this continuum: they carry a meaning of their own and can be understood without accompanying speech, but their range of expressions is very limited, and there are no syntactic rules for combining them to a full language. Such gestures may either have a conventionally agreed, more or less arbitrary, meaning like 'somebody is nuts,' 'military salute,' or 'okay,' or they may indicate objects by miming their use. Usually, diagnosis and research on apraxia concentrate on miming of object use, because aphasic patients may not understand the verbal label of gestures with conventional meaning, whereas comprehension of the object name can be facilitated by showing either the object or a picture of it.

Examination of meaningful gestures requires that they are demonstrated outside their appropriate behavioral context. The instruction is 'show me how you *would* show to somebody that they are nuts' or 'show me how you *would* use a hammer.' This is significantly different from the instructions given for imitation: 'Do as I do' or, for actual object use, 'use this object.' To follow such instructions requires symbolic thought or, respectively, an 'abstract attitude.' Miming the use of an object without tactual contact poses additional demands on imagination and inventiveness. The motor actions of actual object use are partly determined by mechanical constraints and properties of the used objects. The patients must compensate for the absence of this information by conjuring up a mental image of actual object use and extracting from this image the shape and motion path of the hand holding the object. Indeed, it has been observed that provision of an object whose tactual properties resemble those of the pretended object (e.g., a stick for a hammer) may induce a significant improvement of miming. Of course, miming object use would be impossible without any knowledge about how the actual object is to be handled. This knowledge—the nature of which will be the subject of the following section—is needed in addition to the ability to demonstrate it without touching the object.

Defective demonstration of meaningful gestures is exclusively linked to LBD (Barbieri and De Renzi 1988, Goldenberg and Hagmann 1998). Although

some deviations from normal performance have been documented in RBD patients by applying sophisticated measurement of all spatial and temporal details of gestures, these deviations never approach the gross errors or total failure which apraxic patients with LBD encounter when asked to demonstrate meaningful gestures. Many LBD patients who cannot mime object use can demonstrate the use of the same objects when allowed to take them in their hands (De Renzi et al. 1982, Goldenberg and Hagmann 1998), whereas the possibility of a reverse dissociation of impaired actual object use with intact miming is questionable. By contrast, there are patients who cannot demonstrate meaningful gestures but can imitate meaningless gestures (Barbieri and De Renzi 1988) and—as already discussed—patients who can demonstrate meaningful gestures but cannot imitate. Such a double dissociation suggests that demonstration of meaningful gestures and imitation involve nonoverlapping components of left hemisphere competence.

The multifaceted nature of meaningful gestures makes it difficult to draw any firm conclusions as to which aspect of them is exclusively bound to the left hemisphere. Possibly, the left hemisphere contribution is essential for the elaboration of comprehensive communicative signs and and for the ability to demonstrate them in the absence of their habitual behavioral context.

3. *Tool and Object Use*

The inability to use real tools and objects is the least frequent but most dramatic manifestation of apraxia. For example, patients may try to cut bread with the reverse edge of the knife or even with a spoon, may press the head of the hammer upon the nail rather than hitting, or may try to press toothpaste out of a firmly closed tube. There is general agreement that such errors arise at a conceptual level of motor control. 'Agnosia of utilization' (Morlaas 1928) renders patients unable to recognize how objects should be used.

Knowledge about how to use tools and objects can have several sources: it may be specified by 'instructions of use' stored in semantic memory and retrieved as one of multiple semantic features when the object has been identified. Such instructions can exist only for familiar objects and are likely to specify their prototypical use, like inserting nails for a hammer and extruding nails for pincers. One can, however, use pincers for hammering. Possible nonprototypical uses of familiar objects as well as possible uses of unfamiliar objects could be detected by a direct matching between structural properties of objects and the affordances posed by actions (Vaina and Jaulent 1991), that is, by a direct inference of function from structure.

When the task transgresses the use of single tools and objects to require the coordination of multiple actions with several objects, like, for example, preparing a meal or fixing household repairs, additional cognitive resources are invoked. The task must be parsed into its component actions and their adequate sequence must be determined considering hierarchical relations between goals and subgoals. During the course of actions the running of the sequence must be checked, updated, and possibly revised. It is likely that these demands pose loads on memory and on general reasoning abilities.

There is evidence that the first two of these sources, retrieval of instructions of use from semantic memory and inference of function from structure, are exclusively bound to left hemisphere function. Patients with LBD make errors when requested to match objects according to similiarities of function rather than to perceptual similarities (Vignolo 1990). As already noted, retrieval of knowledge about object use is a component of pantomiming object use which is deficient exclusively in LBD patients, and although real object use is usually less affected than miming, the severities of their disturbances are correlated (Goldenberg and Hagmann 1998). It thus seems very likely that LBD patients have difficulties with retrieval of instructions of use from semantic memory.

Evidence for an inability to directly infer function from structure comes from experiments in which patients are requested either to find alternative uses of familiar objects for accomplishing a given task (e.g., selecting a coin for screwing when there is no screwdriver), or to find out the possible applications of unfamiliar tools (Heilman et al. 1997, Goldenberg and Hagmann 1998). With both types of tests only patients with LBD encounter difficulties.

The role of the left hemisphere is much less clear for the additional cognitive components that come into play when the task affords a chain of actions with several tools and objects. Whereas disturbances of simple object use (e.g., hammering a nail, opening a bottle) are found exclusively in patients with LBD, complex actions sequences (e.g., preparing a lunch, wrapping a gift) also pose difficulties for patients with RBD or with diffuse brain damage (Schwartz et al. 1999).

4. *Conclusions*

Apraxia was crucial to Liepmann's proposal that the left hemisphere is dominant for the control of motor actions. This proposal was attractive because it promised to explain many, if not all, clinical symptoms of left brain damage within the framework of one coherent theory of hemisphere specialization. One hundred years of research have falsified the hypothesis by demonstrating that apraxia embraces a collection of heterogeneous symptoms and cognitive deficits which can be brought forward by examining motor actions but cannot plausibly be reduced to insufficient

motor control. These symptoms are, however, worthy of being studied in their own right. They refer to central domains of human competence. Learning novel skills through imitation, the use of symbols to denote absent objects and events, and creation and use of tools, have all been proposed as being unique to man and crucial for the development of human culture. Abandoning motor dominance as their common denominator leaves open the question why decisive components of these aptitudes are bound to left hemisphere function. Apraxia continues to promise a key to understanding hemisphere specialization and its importance for the development of specifically human aptitudes.

See also: Brain Asymmetry; Classical Mechanics and Motor Control; Motor Control; Motor Control Models: Learning and Performance; Motor Skills, Psychology of; Neural Representations of Intended Movement in Motor Cortex

Bibliography

Barbieri C, De Renzi E 1988 The executive and ideational components of apraxia. *Cortex* **24**: 535–44

De Ajuriaguerra J, Tissot R 1969 The apraxias. In: Vinken P J, Bruyn G W (eds.) *Handbook of Clinical Neurology*. North-Holland, Amsterdam, Vol. 4, pp. 48–66

De Renzi E, Faglioni P, Sorgato P 1982 Modality-specific and supramodal mechanisms of apraxia. *Brain* **105**: 301–12

Geschwind N, Damasio A R 1985 Apraxia. In: Frederiks J A M (ed.) *Handbook of Clinical Neurology*. Elsevier, Amsterdam, New York, Vol. 1 (49), pp. 423–32

Goldenberg G 1999 Matching and imitation of hand and finger postures in patients with damage in the left or right hemisphere. *Neuropsychologia* **37**: 559–66

Goldenberg G, Hagmann S 1997 The meaning of meaningless gestures: A study of visuo-imitative apraxia. *Neuropsychologia* **35**: 333–41

Goldenberg G, Hagmann S 1998 Tool use and mechanical problem solving in apraxia. *Neuropsychologia* **36**: 581–9

Heilman K M, Maher L M, Greenwald M L, Rothi L J G 1997 Conceptual apraxia from lateralized lesions. *Neurology* **49**: 457–64

Hermsdörfer J, Mai N, Spatt J, Marquardt C, Veltkamp R, Goldenberg G 1996 Kinematic analysis of movement imitation in apraxia. *Brain* **119**: 1575–86

Kimura D, Archibald Y 1974 Motor functions of the left hemisphere. *Brain* **97**: 337–50

Kolb B, Milner B 1981 Performance of complex arm and facial movements after focal brain lesions. *Neuropsychologia* **19**: 491–503

Liepmann H 1908 *Drei Aufsätze aus dem Apraxiegebiet*. Karger, Berlin

McNeill D 1992 *Hand and Mind*. University of Chicago Press, Chicago

Morlaas J 1928 *Contribution à l'étude de l'apraxie*. Amédée Legrand, Paris

Roy E A, Hall C 1992 Limb apraxia: a process approach. In: Proteau L, Elliott D (eds.) *Vision and Motor Control*. Elsevier, Amsterdam, pp. 261–82

Schwartz M F, Buxbaum L J, Montgomery M W, Fitzpatrick-DeSalme E J, Hart T, Ferraro M, Lee S S, Coslett H B 1999 Naturalistic action production following right hemisphere stroke. *Neuropsychologia* **37**: 51–66

Vaina L M, Jaulent M C 1991 Object structure and action requirements: A compatibility model for functional recognition. *International Journal of Intelligent Systems* **6**: 313–36

Vignolo L A 1990 Non-verbal conceptual impairment in aphasia. In: Boller F, Grafman J (eds.) *Handbook of Clinical Neuropsychology*. Elsevier, Amsterdam, pp. 185–206

G. Goldenberg

Archaeology and Cultural/National Memory

Individual humans can remember occurrences of their own lifetimes as well as of a collective past, whether near or far removed in time from the moment of remembering. While the former is restricted to events and processes at which the particular person was physically present as a conscious human being, the latter includes accounts of occurrences happening both before and during an individual's life. Although records of personal memories may be an important source for some areas of archaeological research (e.g., twentieth-century archaeology, San rock art), most archaeologists are dealing with collective memories.

This article discusses several models of how collective memory functions. An overview will be given of the different implications for both the archaeological study of ancient sites and objects, and archaeology as an academic discipline in the context of a given society or nation of the present.

1. Archaeology and Theoretical Approaches to Memory

For the longest time in intellectual history, human memory has been seen as a huge archive in every human being from which, at any point in time, specific items can be retrieved in the process of remembering. This was, for example, the view of Saint Augustine (*Confessions*, X.viii(12)). In this perspective, individuals may create mnemonics in order not to forget on particular occasions (Yates 1966) and they are able, in principle, to remember past events accurately. Collective memories may work accordingly. More recently, however, it has been argued that the issues involved in remembering and forgetting are not that simple.

1.1 Memory as a Social Construction

Remembering, and indeed forgetting, are strongly influenced by circumstances of the time in which they take place. Moreover, memory can fail completely or

'make up' stories or events that allegedly took place in the past. Maurice Halbwachs argued that some key factors affecting memory derive from the social arena which people always inhabit when they remember. Halbwachs introduced the concept of '*mémoire collective*' (collective memory), and emphasized how strongly social processes influence not only people's personal memories of their own lifetimes, but also a community's shared memories of the past. Collective memories are crucial for the identity of groups such as families, believers of a religion, or social classes (Halbwachs 1992). Although memories vary between individuals and no two persons share identical 'collective' memories, anyone who has ever lived for some time in a foreign country knows what it means not to share the same collective memories as colleagues and friends. Collective memories can be evoked at particular sites (see below), and they have, in turn, the potential to shape such sites. As an example of the latter, Halbwachs discussed the Medieval 'creation' of the Holy Land by the superimposition of an imaginary landscape on Palestine (Halbwachs 1971).

The collective memory of archaeology provides a topical example. Although little research has been done on that topic, it is clear that some historical figures such as J. J. Winckelmann, C. J. Thomsen, General A. Pitt-Rivers, and H. Schliemann are considered widely to be the 'fathers' of the discipline. Similarly, sites such as the Valley of the Kings, Troia (Troy), Teotihuacan, and Stonehenge, are remembered collectively as key sites of archaeology and are now major tourist sites. While the significance of any of these persons and sites is not disputed here, it is clear that they function *de facto* as parts of the collective memories of archaeology and its successes. The way they are remembered tells us as much, if not more, about specific people and their identities in the present as about the history of archaeology.

Going beyond Halbwachs's argument, social scientists have argued recently that memory of the past is not only influenced but also constituted by social contexts of the present (Middleton and Edwards 1990, Fentress and Wickham 1992). Such reasoning questions the separation of past and present in a fundamental way. It becomes pointless to discuss whether or not a particular remembered event or process corresponds to what 'actually' happened: all that matters are the specific conditions under which such memory is constructed, and its full personal and social implications at a given point in time and space. The distinction between personal and collective memory is thus not necessarily a sharp one. Both reflect, first and foremost, the conditions of the present in which they originate. Individual persons learn collective memories through socialization, but they retain the freedom to break out of it and offer alternative views of the past which may themselves later inform the collective memory. It may eventually become impossible to tell which memory is indeed an accurate remembrance of an occurrence at which a person was present, and which is merely a remembrance of an earlier remembrance that was created by a story read in a book or watched on television.

1.2 'Les lieux des mémoire'

Pierre Nora edited a monumental work of seven volumes about the loci memoriae of France, entitled *Les lieux des mémoire* (Nora 1984–92). A '*lieu de mémoire*' (site of memory) is any significant entity, whether material or nonmaterial in nature, which by dint of human will or the work of time has become a symbolic element of the memorial heritage of any community (see *History and Memory*). Nora deals specifically with sites where modern France constructs its national identity by constructing its past. Crucially, these sites of memory do not only include places, e.g., museums, cathedrals, cemeteries, and memorials; but also concepts and practices, e.g., generations, mottos, commemorative ceremonies; and objects, e.g., inherited property, monuments, classic texts, and symbols. Archaeological sites of memory discussed in individual contributions include Palaeolithic cave-paintings, megaliths and the Gauls.

According to Nora, all sites of memory are artificial and deliberately fabricated. Their purpose is to stop time and the work of forgetting, and what they all share is 'a will to remember.' Such sites of memory are not common in all cultures, but a phenomenon of our time: they replace a 'real' and 'true' living memory which was with us for millennia but has now ceased to exist. Nora thus argues that a constructed history replaces true memory.

1.3 Cultural Memory

Jan Assmann employed the term '*kulturelles Gedächtnis*' (cultural memory) in a study of the past in ancient Egypt (Assmann 1992). Cultural memory can be defined as the cultural expressions of a collective memory, including written texts, symbols, monuments, memorials, etc. It is distinct from the 'communicative memory' of individuals' lifetimes, which is expressed by means of direct communication (see *Oral History*). Cultural and collective memories are 'retrospective.' Correspondingly, the hopes and expectations of people for what will be remembered in the future can be termed 'prospective memory.' Such hopes find their expression not only in the design and 'content' of a site of memory, but also in its monumentality and the durability of the materials employed. If prospective memory is the desired memory when building a site of memory, retrospective memory is how this site is in fact later remembered.

According to Assmann, cultural memory embraces both '*Erinnerungskultur*' (memory culture) and '*Ver-*

gangenheitsbezug' (reference to the past). Memory culture is the way a society ensures cultural continuity by preserving, with the help of cultural mnemonics, its collective knowledge from one generation to the next, rendering it possible for later generations to reconstruct their cultural identity and maintain their traditions. Memory culture relies on later references to (mnemonics of) the past. Such references reassure the members of a society of a shared past and collective identity, and supply them with an awareness of their unity and singularity in time and space. Cultural mnemonics, such as preserved ancient monuments or artifacts, trigger recollections of past times, although these can be fabrications rather than accurate representations of the past. Arguably, this second aspect of cultural memory has the wider implications. The fact that collective memories of the past occur without being able to ensure cultural continuity, suggests that the past can be given specific meanings for reasons other than accurate reconstruction. In effect, as interpretations of the past have changed over time, ancient sites and artifacts have been interpreted quite differently during their long 'life-histories (Holtorf 2000).'

For archaeologists, both actual continuities and various archaistic reinventions or other symbolic references to a rediscovered or recreated past are interesting phenomena to study. Because both already occur widely in antiquity, Nora is arguably misguided in limiting the existence of sites of memory to the modern era. There is evidence for deliberate attempts to stop time and make contemporaries remember by 'artificial' means, from the ancient Near East (e.g., Assmann 1992, Jonker 1995) as well as from prehistoric Europe (Holtorf 2000).

2. *The Relationship between Archaeology and Collective Memories*

2.1 *Archaeology vs. Memory*

It has often been assumed that the academic study of the past is superior epistemologically to popular notions of the past, as they are reflected in folklore, myths or in other expressions of collective memory. Maurice Halbwachs contrasted memory and history as two oppositional ways of dealing with the past. Whereas historians aim at writing a single objective and impartial universal history, collective memories are numerous, limited in their validity to members of a particular community, subject to manifold social influences, and restricted to the very recent past in scope.

Likewise, Pierre Nora argued that memory and history are two very different phenomena, but his preference was the opposite to that of Halbwachs. Nora distinguished true memory, borne by living societies maintaining their traditions, from artificial history, which is always problematic and incomplete, and represents something that no longer exists. For Nora, history holds nothing desirable. In each case, the term 'history' might be replaced with 'archaeology.'

2.2 *Archaeology as Memory*

Recently the split between history/archaeology and memory was challenged fundamentally, and a more fluid transition proposed instead (e.g., by Burke 1989). In this view, history and archaeology are seen as special cases of social and cultural memory. Doing archaeology simply means (a) to recognize, and treat, certain things as 'evidence' for the past; and (b) to describe and analyze it in a way that is valuable according to the standards of archaeologists. These practices are not categorically different from how people engage with the past in their everyday lives. History and archaeology, like all forms of memory, give particular meanings to ancient sites and artifacts. Some of these 'academic' meanings eventually influence collective memories of the past in the present. But many never leave the sphere of academia, and are short-lived there too. What is interesting, therefore, in studying collective memories of the past, is not only how accurately they might fit (some) part of a past reality, but also why particular memories are created and adopted (or not) in particular contexts, and what falls into oblivion.

2.3 *Archaeology and Nationalist Agendas*

The history of archaeology provides a fitting example of the great importance of the past for collective identities. Archaeology originated as an academic discipline in the nineteenth century—the great age of the European nation states. This is no coincidence: archaeology has from its beginnings been directly bound up with politics (see *Archaeology, Politics of*). By establishing the origins of their respective national people in the distant past, the young nations reassured their citizens of a shared past, and thus a shared identity and legitimacy in the present. A case in point are the defences of Masada in Palestine, which fell to the Romans after several years of Jewish resistance in 73 AD, but not until the last defenders had committed collective suicide. They were not rediscovered as a prime object of Jewish collective memory until the 1920s coinciding with the rise of Zionism (Schwartz et al. 1986). At times of fundamental social change, people turn to their 'origins' and seek reassurance either in a 'better' past, or in historical traditions and collective identities which legitimate a political movement or new social order. Not surprisingly, much of early archaeological research focused on 'culture-historical' approaches, and invested much effort in the ethnic interpretation of material evidence (see *Ethnic Identity/Ethnicity and Archaeology*). This connection between archaeology and the modern nation is still

visible at the beginning of the twenty-first century in many 'national museums' of archaeology, national academic journals with titles such as *Germania* and *Gallia*, and in the key role of the national pasts in both teaching curricula and popular culture. During the last decade of the twentieth century, and with the emergence of new nation states in many parts of Eastern Europe and the former Soviet Union, similar processes were repeated (Kohl and Fawcett 1995, Diaz-Andreu and Champion 1996).

In some cases, collective memories support nationalist ideologies effectively, while contradicting results gained from archaeological research. Some archaeologists too may lend their authority to support such ideologies. At first this appears to raise a dilemma for those archaeologists who are prepared to accept the equal legitimacy and value of collective memories of different groups, whether academic or not: either they have to accept all constructions of the past as legitimate alternatives to their own accounts, or they would contradict themselves by dismissing some as false. A realistic solution is to allow many alternative collective memories in principle, but be prepared to fight some of them on political or other grounds, and publicize widely any dangerous consequences or implications.

Such issues have recently come to the fore in several regions of the world. They demonstrate that archaeology and collective/national memory are deeply intertwined and ultimately interdependent. Instead of arguments about theoretical principles, future archaeologists will be challenged increasingly to find pragmatic guidelines as to how to behave and act in politically highly charged situations.

See also: Collective Beliefs: Sociological Explanation; Collective Identity and Expressive Forms; Collective Memory, Anthropology of; Collective Memory, Psychology of; Cultural History; Nationalism and Expressive Forms; Nationalism: General; Nationalism, Sociology of

Bibliography

Assmann J 1992 *Das kulturelle Gedächtnis. Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. Beck, Munich, Germany

Burke P 1989 History as social memory. In: Butler T (ed.) *Memory: History, Culture and the Mind*. Blackwell, Oxford, UK, pp. 97–113

Díaz-Andreu M, Champion T (eds.) 1996 *Nationalism and Archaeology in Europe*. UCL Press, London

Fentress J, Wickham C 1992 *Social Memory*. Blackwell, Oxford, UK

Halbwachs M 1971 [1941] *La topographie légendaire des évangiles en terre sainte. Etude de mémoire collective*. Presses Universitaires de France, Paris (Conclusion translated in Halbwachs 1992.)

Halbwachs M 1992 *On Collective Memory*. [Ed. trans., and intro. by L A Coser]. University of Chicago Press, Chicago, IL

Holtorf C 2000 *Monumental Past: The Life-histories of Megalithic Monuments in Mecklenburg-Vorpommern (Germany)*. CITD Press, Toronto, Canada (electronic monograph: http://citd.scar.utoronto.ca/CITD Press/Holtorf/)

Jonker G 1995 *The Topography of Remembrance. The Dead, Tradition and Collective Memory in Mesopotamia*. Brill, Leiden, The Netherlands

Kohl P, Fawcett C (eds.) 1995 *Nationalism, Politics, and the Practice of Archaeology*. Cambridge University Press, Cambridge, UK

Middleton D, Edwards D (eds.) 1990 *Collective Remembering*. Sage, London

Nora P (ed.) 1984–92 *Les lieux de mémoire*. 7 Vols. Edition Gallimard, Paris [Abridged English translation, Nora P and Kritzman L D (eds.) 1996–8 *Realms of Memory*. 3 Vols. Columbia University Press, New York]

Schwartz B, Zerubavel Y, Barnett B M 1986 The recovery of Masada: A study in collective memory. *The Sociological Quarterly* **27**: 147–64

Yates F 1966 *The Art of Memory*. Routledge & Kegan Paul, London

C. Holtorf

Archaeology and Philosophy of Science

Archaeologists have long been concerned with questions about the scientific status of their discipline, periodically engaging in debate about goals and standards of practice that raises issues central to the philosophy of science. In the 1960s and 1970s, with the advent of the New Archaeology in North America, the philosophical content of these debates became explicit. The New Archaeologists advocated a scientific program of research modeled on logical positivist theories of science, and their critics have since drawn on a range of post-positivist philosophies of science for alternative guidelines and ideals. Philosophers of science have been direct participants in some of these debates. There is now growing interest, among both philosophers and archaeologists, in a range of philosophical questions that extend well beyond debate about whether archeology does (or should) conform to models of practice derived from the natural sciences: the naturalist commitments (in philosophical terms) that initially motivated archaeologists to turn to the philosophy of science.

1. Interactions

The influence of naturalist ideals are evident at many junctures in internal archaeological debate. In the early twentieth century, when archaeology was taking shape as a museum and university-based profession, advocates of a 'new archaeology' defined the difference between archaeological and antiquarian practice in

terms of a commitment to anthropological goals and scientific methods. They insisted that the value of archaeological material lies, not in its intrinsic or artistic merits, but in its capacity to serve as evidence in a program of systematically testing 'multiple working hypotheses' about the cultural past; in this they invoked an influential account of scientific method that appeared in *Science* (Chamberlin 1890).

1.1 Critiques of 'Narrow Empiricism'

When these themes reemerged in the 1930s and 1940s it was in opposition to forms of archaeological practice that were described, in explicitly philosophical terms, as 'narrowly empiricist.' Internal critics objected that the discipline had become mired in empirical detail. Although anthropological goals were widely endorsed, rarely were they directly addressed. The process of recovering and systematizing archaeological data had become an end in itself; it was assumed that interpretive or explanatory theorizing must be deferred until an exhaustive archaeological data base had been secured and systematized. These forms of practice were impugned as 'narrowly empiricist' because they were said to presuppose, in a particularly stringent form, the central presuppositions of an empiricist theory of knowledge: that empirical or sensory experience is the source and ground of all legitimate knowledge and, more specifically, that it constitutes a foundation for knowledge that is independent of the theoretical or interpretive claims it may be used to support or refute.

One prominent anthropological critic, Kluckhohn, published a critique of such practice in *Philosophy of Science* (1939) in which he challenged the assumption, based on these principles, that anthropological goals can be realized only inductively, that is, by first collecting archaeological data and then gradually building up a body of interpretive or explanatory theory. Kluckhohn and aligned archaeological critics argued that the contents of the archaeological record have no significance except in light of a theoretical framework; and if interpretive or explanatory questions about the cultural past are to be answered, empirical investigation must be theoretically informed and problem-oriented. These critics drew on a number of philosophical sources including Whitehead, philosophers of history such as Teggart and Mandelbaum, and pragmatists such as Dewey. A decade later these questions about epistemic foundations emerged again in the context of debate over the status of typological constructs: are typological categories discovered in or derived from empirical data, or are they interpretive constructs, and, if they are constructs, are they irreducibly subjective? Those who took the former position appealed to the 'liberal positivism' of such philosophers of science as Bergman, Brodbeck, Feigl, and Hempel.

It was primarily to Hempel that the New Archaeologists turned in the 1960s and 1970s when they argued the case for a self-consciously scientific research program predicated on the principles of logical positivism (Watson et al., 1971). What they drew from Hempelian positivism were models of scientific explanation and confirmation which they saw as an antidote to 'traditional' (empiricist and inductivist) practice. Hempel's models are representative of what has since been described as 'received view' philosophy of science, the product of 50 years of careful reconstruction of the logic of scientific reasoning that presupposes the central tenets of the empiricist tradition. On his covering-law (C-L) model, explanation is accomplished when a particular event or property can be shown to fit (by deductive subsumption) the patterns of conjunction or succession captured by law-like generalizations. Explanation and prediction are thus symmetrical: the laws support deductive inferences from established regularities to instances which show that they can be, or could have been expected. This symmetry is central to Hempel's hypothetico-deductive (H-D) account of confirmation in which hypotheses about prospective law-like regularities are tested by determining what empirical implications follow if they are true and then systematically searching for evidence that fits or violates these expectations.

The New Archaeologists were confident that, if they made it their central objective to develop and test law-governed explanatory hypotheses, following the guidelines suggested by these deductive models, they could transcend the limitations of 'narrow empiricism' without indulging in inductive speculation. On an H-D model of confirmation, interpretive and explanatory hypotheses become the departure point for empirical inquiry, rather than its deferred goal; and if laws connect antecedent cultural processes to archaeological outcomes, explanatory inference is secured. What the New Archaeologists did not appreciate is that Hempelian positivism presupposes the central tenets of empiricism; logical positivism is an empiricist theory of science. On such an account, empirical evidence is not only the final court of appeal for adjudicating theoretical claims, but the exhaustive source of their content. If claims about unobservables such as the cultural past are to have cognitive significance, they must be reducible to, or derivable from, the observations they subsume.

1.2 Post-positivism

The positivism of the New Archaeology drew immediate critical attention, both from fellow archaeologists and from philosophers of science. Many objected that the 'received view' philosophy of science had met its demise by the time archaeologists invoked it as a model for their practice. Post-positivist philosophers and historians of science, most famously Kuhn, had

decisively challenged its foundationalist assumptions, arguing that theory and evidence are interdependent (evidence is theory-laden). Moreover and the enthusiasm for 'theory demolition' and deductive certainty had been called into question by critics who showed that the most interesting theoretical claims overreach all available evidence (theory is underdetermined by evidence). But beyond this critical consensus, responses to the demise of positivism diverged sharply.

Many philosophical commentators were sympathetic to the scientific (naturalist) ideals of the New Archaeology but, together with internal archaeological critics, they made the case for alternative models of science that better fitted the conditions of practice and ambitions of archaeology. Those proposed by Salmon (1982), and the Popperian models advocated by Bell (1994), fall within the ambit of a liberal empiricism even if they are not positivist in conception, while others represent a more fundamental reconception of scientific practice: the scientific realism endorsed by Gibbon (1989) and the coherentism elaborated by Kelley and Hanen (1988). More radical departures are to be seen in post-processual critiques of New Archaeology. The advocates of broadly interpretivist, humanistic approaches (e.g., contributors to Tilley 1993) reject the naturalist assumption that archaeology is best conceived as a scientific enterprise. They draw inspiration from a range of philosophical traditions outside analytic philosophy of science (e.g., critical theory, phenomenology, and philosophical hermeneutics).

2. *Issues*

While debate over naturalist ideals has most often been the catalyst for interaction between archaeologists and philosophers of science, several more narrowly defined philosophical issues have become a recurrent focus of attention.

2.1 *Explanation*

The explanatory goals of archaeology are a perennial concern. Critics of the Hempelian C-L model argue for a range of alternatives, chiefly accounts which recognize forms of explanation that do not depend on laws. Systems models of explanation were initially popular among archaeological critics. Salmon proposed a causally enriched statistical-relevance (S-R) model on which explanation is accomplished, not by invoking a covering-law but by enumerating the factors that can be shown (statistically) to make a difference to the occurrence of the events or conditions that require explanation (Salmon 1982, Chap. 3, 5, 6, pp. 84–139). Gibbon elaborates the causalist elements of Salmon's model in a realist account. According to this approach, explanation is a matter of building models of the antecedent causal processes and conditions that were responsible for the surviving archaeological record (1989, Chap. 7, pp. 142–72).

This makes sense of the central role, in archaeological explanations, of claims about unobservables, both material and intentional, a feature of scientific explanation with which logical positivists have always had difficulty (Hempel 1965, pp. 177–87). By contrast, Kelley and Hanen advocate an anti-realist, pragmatist position. They consider that explanatory power is a key consideration in evaluating competing hypotheses but not the central aim of scientific inquiry. Explanations are answers to 'why-questions' that deploy whatever scientifically credible information will satisfy a specific inquirer (1988, Chap. 5, pp. 165–224).

2.2 *Theory and Evidence*

A second focus of jointly philosophical and archaeological interest is a family of questions about the nature of archaeological evidence and its use in formulating and evaluating claims about the cultural past. In early programmatic statements, the New Archaeologists repudiated any dependence on inductive forms of inference in favor of an H-D testing methodology. They quickly realized, however, that archaeological data stand as test evidence only under interpretation and, in most cases, this requires the use of background or collateral knowledge ('middle range theory') in reconstructive inferences that rarely realize deductive security. By the late 1970s many had turned their attention to development of the necessary linking principles and, since the mid-1980s, one of the most pressing philosophical problems at the interface between philosophy and archaeology has been to explain how evidence can be theory-laden without risking vicious, self-justifying circularity when used to test interpretive theory. One strategy of response, developed by Kosso (1992) and Wylie (2000) is to identify conditions of epistemic independence between test hypotheses and linking principles which ensure that the 'middle range theory' used to interpret archaeological data does not necessarily ensure a favorable outcome for a particular test hypothesis.

At a broader level of analysis, a number of alternatives to the H-D model confirmation have been proposed. Salmon advocates a modified Bayesianism model which captures the complex reasoning by which archaeologists weigh the significance of evidence for a particular hypothesis against its prior probability of being true (1982, Chap. 3, pp. 31–57). Many of these considerations figure in Kelley and Hanen's account of archaeological practice as a comparative process of inference to the best explanation, but they also emphasize several other epistemic virtues such as explanatory power and consistency with established 'core beliefs.' Popper's falsificationist account of theory testing has been advocated by Bell (1994) who

argues that archaeologists should proceed, not by building support for bold conjectures confirmationally, but by searching for the evidence that is most likely to refute them.

3. Metaarchaeology

Although archaeologists have long drawn on the philosophy of science to articulate their programmatic goals and guidelines, philosophers took little systematic interest in archaeology before the 1970s. There are, however, some notable exceptions. One is Collingwood, a philosopher of science and of history who was also an active archaeologist and historian of Roman Britain in the inter-war period (Collingwood 1978, [1939]). He made frequent and subtle use of archaeological examples to develop models of historical inquiry that do not fit neatly on either side of the conventional naturalist–anti-naturalist divide. His influence is evident in the work of Clarke, a British contemporary of the New Archaeologists who was a staunch naturalist but insisted that archaeologists should not assume that any existing models of scientific practice will serve them well as a guide to more systematic empirical practice (1973). At the same time, Collingwood's philosophy of history has been an important inspiration for interpretivist critics of the New Archaeology.

Since the 1970s sustained interaction has developed between archaeologists and philosophers of science. In the process, attention has shifted away from questions about whether archaeological practice does, or should, fit existing models of scientific practice. Increasingly the focus is on distinctive problems of archaeological practice, or on the use of archaeological examples as a basis for reframing and extending philosophical models developed in other contexts. The result has been the formation of a vigorous field of 'metaarchaeological' inquiry (Embree 1992) at the intersection between philosophy and archaeology.

See also: Empiricism, History of; Explanation: Conceptions in the Social Sciences; History of Science; Meta-analysis: Overview; Positivism, History of

Bibliography

Bell J A 1994 *Reconstructing Prehistory: Scientific Method in Archaeology*. Temple University Press, Philadelphia, PA

Clarke D L 1973 Archaeology: the loss of innocence. *Antiquity* **47**: 6–18

Collingwood R G 1978 [1939] *An Autobiography*. Oxford University Press, Oxford, UK

Chamberlin T C 1890 The method of multiple working hypotheses. *Science* **15**: 92

Embree L (ed.) 1992 *Metaarchaeology: Reflections by Archaeologists and Philosophers*. Kluwer, Boston

Gibbon G 1989 *Explanation in Archaeology*. Basil Blackwell, London, UK

Hempel, C G 1965 *Aspects of Scientific Explanation and Other Essays in Philosophy of Science*. Free Press, New York

Kelley J H, Hanen M P 1988 *Archaeology and the Methodology of Science*. University of New Mexico Press, Albuquerque, NM

Kluckhohn C 1939 The place of theory in anthropological studies. *Philosophy of Science* **6**: 328–344

Kosso P 1992 Observation of the past. *History and Theory* **31**: 21–36

Salmon M H 1982 *Philosophy and Archaeology*. Academic Press, New York

Tilley C (ed.) 1993 *Interpretative Archaeology*. Berg Publishers, Oxford, UK

Watson P J, LeBlanc S A, Redman C L 1971 *Explanation in Archaeology: An Explicitly Scientific Approach*. Columbia University Press, New York

Wylie A 2000 Rethinking unity as a working hypothesis for philosophy of science: How archaeologists exploit the disunity of science. *Perspectives on Science* **7**(3): 293–317

A. Wylie

Archaeology and the History of Languages

Possible correlations between the histories of the major language families and major traditions within the archaeological record have exercised the minds of scholars since Gustav Kossinna, and Gordon Childe attempted early in the twentieth century to trace the archaeological record of the Indo-European languages. But long before the rise of archaeology as a research discipline, some of the major language families had already come into historical perspective through comparative linguistic research. This perspective is often claimed to have emerged when Sir William Jones in 1786 suggested that Greek, Sanskrit, Latin, Gothic, Celtic, and Old Persian were 'sprung from some common source.' (Jones 1993).

In the twenty-first century, language history and the archaeological record can be studied in combination to recover history at two major (but clearly overlapping) levels:

(a) at the level of the individual language, *ethnolinguistic group*, or historical community; and,

(b) at the level of the *language family* or major subgroup.

It is also possible to seek linguistic correlations for some archaeological complexes, particularly those which are sharply bounded and defined by consistent stylistic features, although this tends to become more difficult as the complex in question extends further back into prehistory and becomes more diffusely defined. Such correlations, for obvious reasons, also benefit from the assistance of written and translatable texts.

In general, it is a very difficult task to trace the identity into deep levels of prehistory of a specific ethnolinguistic or historical population (e.g. Celts, Greeks, Etruscans), unless one is dealing with a very isolated region or an island where one can assume there has been no substantial population replacement

during the period in question. A good example of the latter would be certain Pacific islands, for example Easter Island or New Zealand, both fairly isolated since their first human settlements by Polynesians (Kirch and Green 1987). However, this entry is not primarily concerned with such society- or culture-specific correlations amongst language, history, and archaeology, but focuses instead on the study of languages as members of genetically constituted and evolving families, combined with the study of large-scale archaeological traditions as they spread, evolve, and interact through time and space.

Historical reconstruction at this level tends to be organized such that language families (e.g. Indo-European, Austronesian) are foregrounded as the major foci of enquiry, rather than archaeological complexes. This is because language families are usually more sharply defined and reveal much clearer patterns of genetic inheritance than do archaeological complexes. In such situations, archaeology tends to be used to support or refute historical linguistic questions (for instance, where was the Indo-European homeland located, what lifestyle did its inhabitants enjoy, and when?). However, some archaeological complexes of particularly wide distribution, internal homogeneity, and short time span (e.g., the Linearbandkeramik (LBK) early Neolithic of Central Europe, the Lapita cultural complex of the western Pacific) are also sometimes foregrounded as requiring a paleolinguistic identity. For instance, does the LBK correlate with Indo-European dispersal into Central Europe; does Lapita correlate with Austronesian dispersal through Melanesia into Polynesia?

In order to understand how the data of historical linguistics and archaeology might be compared against each other in order to improve understanding of the human past, it is first necessary to state clearly the abilities and limitations of the two disciplines.

1. *Language as a Source of Information on Human Prehistory*

The branch of linguistics which is of most interest to prehistoric archaeologists is that known as comparative historical linguistics, in which the structures and vocabularies of present-day or historically recorded languages are compared in order to identify families, and subgroups within these families. The methodology of comparative linguistic reconstruction is precise. Like the methodology of cladistics, as applied in biology, its main goal is to identify shared innovations which can identify language subgroups. Such subgroups comprise languages which have shared a common ancestry, apart from other languages with which they are more distantly related. Languages which comprise a subgroup share descent from a common 'protolanguage,' this being in many cases a chain of related dialects. The protolanguages of subgroups within a family can sometimes be organized into a family tree of successive linguistic differentiations (not always sharp splits, unlike real tree branches), and for some families it is possible to postulate a relative chronological order of subgroup formation. For instance, many linguists believe that the separation between the Anatolian languages (including Hittite) and the rest of Indo-European represents the first identifiable differentiation in the history of that family. Likewise the separation between the Formosan (Taiwan) languages and the rest of Austronesian (Malayo-Polynesian) represents the first identifiable differentiation within Austronesian.

The vocabularies of reconstructed protolanguages (e.g., Proto-Indo-European, Proto-Austronesian) can sometimes provide remarkable details on the locations and lifestyles of ancient ancestral communities, with many hundreds of ancestral terms and their associated meanings reconstructible in some instances. There is also a linguistic technique known as glottochronology which attempts to date protolanguages by comparing recorded languages in terms of shared cognate (commonly inherited) vocabulary, applying a rate of change calculated from the histories of Latin and the Romance languages. But the rate of change varies with sociolinguistic situation, often a complete unknown in prehistoric situations. Glottochronology can be used only for recent millennia and for those languages which have not undergone intense borrowing from languages in other unrelated families. It is not a guaranteed route to chronological accuracy.

The other major source of linguistic variation, apart from modification through descent, is that termed by linguists 'borrowing' or 'contact-induced change.' This operates between different languages, and often between languages in completely unrelated families. Borrowing, if identified at the protolanguage level, can be as much an indicator of the homeland of a language family or subgroup as can genetic structure. It can also reflect important contact events in language history. (See *Phylogeny and Systematics*.)

2. *Archaeology as a Source of Information on Human Prehistory*

Archaeology is concerned mainly with the recovery and interpretation of the material remains of the human past, and the environmental contexts in which those remains were originally deposited. Such remains can be dated, and grouped into regional complexes of related components. Such complexes can then be compared with other complexes, and the natures of the boundaries between such complexes can be studied carefully. Some are sharply bounded, hence possible candidates for correlation with an ethnolinguistic group, others are simply nodes of relative homogeneity in a kaleidoscope of ever-shifting patterning. Archae-

ology alone cannot pinpoint ethnicity, unless of course it operates in an environment associated with literacy and the availability of written records (and even then ambiguity can plague interpretation, as in the modern debate in the UK archaeological literature about the definition and archaeological history of the Celts). Any correlations between the archaeological and linguistic records will always require care—prehistoric artifacts cannot talk!

3. *How Can Language Family History and Archaeological Prehistory be Correlated?*

Because languages change constantly through time, and because relationships between languages become ever fainter as we go back in time, it is assumed by most linguists that language family histories apply only to the past 8,000 to 10,000 years. At a greater timescale we enter the arena of 'macrofamilies' such as Nostratic and Amerind, concepts which cause rather vituperous debate amongst linguists because of their very ambiguity and elusiveness. Most of the examples discussed below relate to historical trends which have occurred since the beginnings of agriculture and which do not extend back as far as the macrofamily level.

Correlation of the archaeological and linguistic records is not always a simple matter because the two classes of data are conceptually quite discrete. However, correlations can be made when language family distributions correspond with the distributions of delineated archaeological complexes, particularly when the material culture and environmental vocabulary reconstructed at the protolanguage level for a given family correspond with material culture and its environmental correlates as derived from the archaeological record. Many reconstructed protolanguages, for instance, have vocabularies which cover crucial categories such as agriculture, domestic animals, pottery, and metallurgy, these all being identifiable in the archaeological record.

The concept of the language family is more sturdy than that of the archaeological culture. This is important, because linguists have come to a remarkable level of agreement on the classification of the world's language families. Apart from a small number of Creoles, mostly a result of European colonization and population translocation, the vast majority of the world's language families are clearly bounded in a classificatory sense and not beset with huge numbers of 'mixed' languages. As an example, the Indian subcontinent has been a region of interaction between the speakers of Indo-European and Dravidian languages for at least three millennia. Languages within these two families have borrowed extensively from each other, to the extent that the subcontinent is often referred to by linguists as a 'linguistic area' (a zone of widespread areal diffusion). Nevertheless, the subcontinent is not covered by languages which are half Indo-European, half Dravidian, chaotically mixed. This means that language families are coherent entities, capable of maintaining coherence and independence through long periods of time. As such, they are believed to carry traceable records of history and to be associated, in their origins, with homeland regions and processes of population dispersal. Where different language families meet, we can infer that different populations have met as well. Although some populations have changed their languages in the past, it is unlikely that language shift, as opposed to an actual dispersal of ancestral speakers of a protolanguage, can be the main mode of dispersal of a major language family. All of this means that the origins and dispersals of the protolanguages from which language families are created should correlate with major population movements, frequently on a scale which should be visible in the archaeological record.

4. *Some Examples of Language Family Origin and Dispersal Histories with Claimed Archaeological Correlations*

Some of the major language families of the Old World are shown in Fig. 1 (the American families are much more mosaic-like in distribution and cannot be mapped so easily). Also shown in Fig. 2 are regions where agriculture developed independently. Many archaeologists and linguists today recognize that many language families could owe their initial creations to population dispersal as a result of population growth following on from the development of agriculture. If this is so, then the homelands of these families can be expected to overlap with the regions of early agriculture, as indeed seems to be the case for the Middle East, China, and Mesoamerica. However, it is important to remember that many language families are associated totally with hunting and gathering populations, and presumably always have been, so their histories obviously will not involve this factor. Such hunter-gatherer families include Khoisan in southern Africa, the Australian languages (probably several families), Athabaskan and Eskimo-Aleut, and the languages of western North America and southern South America. Other families, such as Uralic, Algonkian, and Uto-Aztecan, have both agricultural and hunter-gatherer populations. In some of these cases it is possible that former agricultural peoples have actually become hunters and gatherers in difficult environments (e.g., the Great Basin Uto-Aztecans).

It is also apparent that some languages and subgroups (but not whole language families) have been recorded as spreading over large distances in historical times, under conditions of statehood, religious evangelism, and colonialism. Thai, the Chinese languages, Arabic, and of course English and Spanish all come to

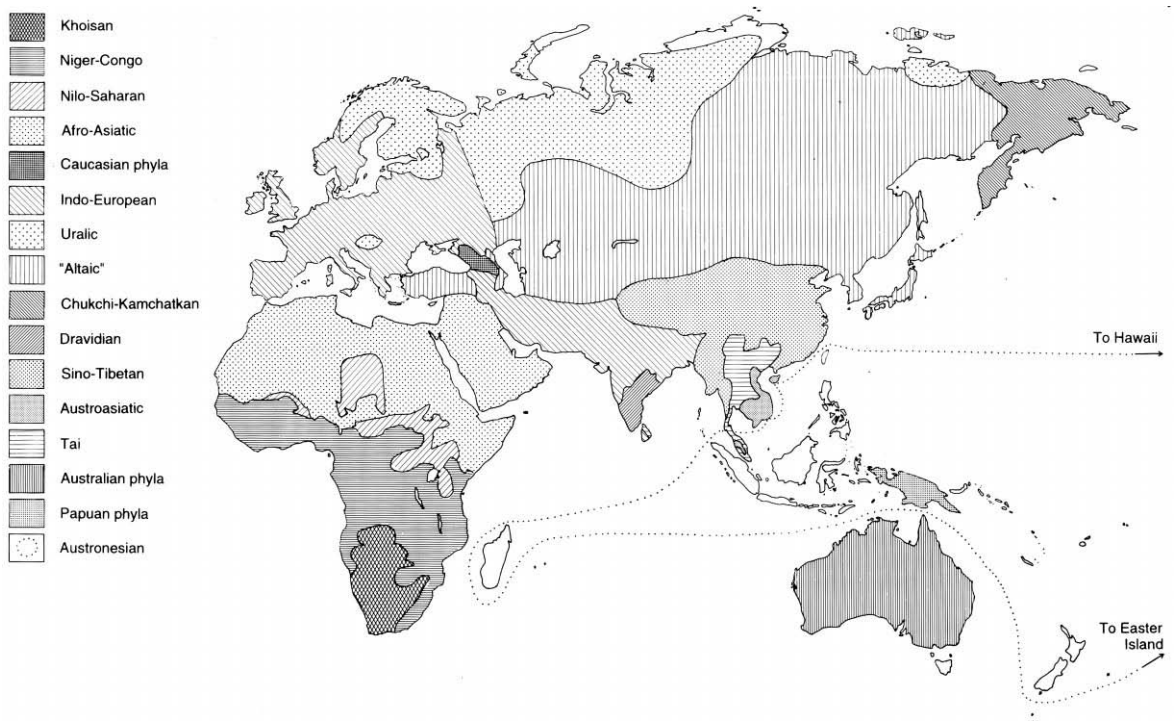

Figure 1
The major language families of the Old World

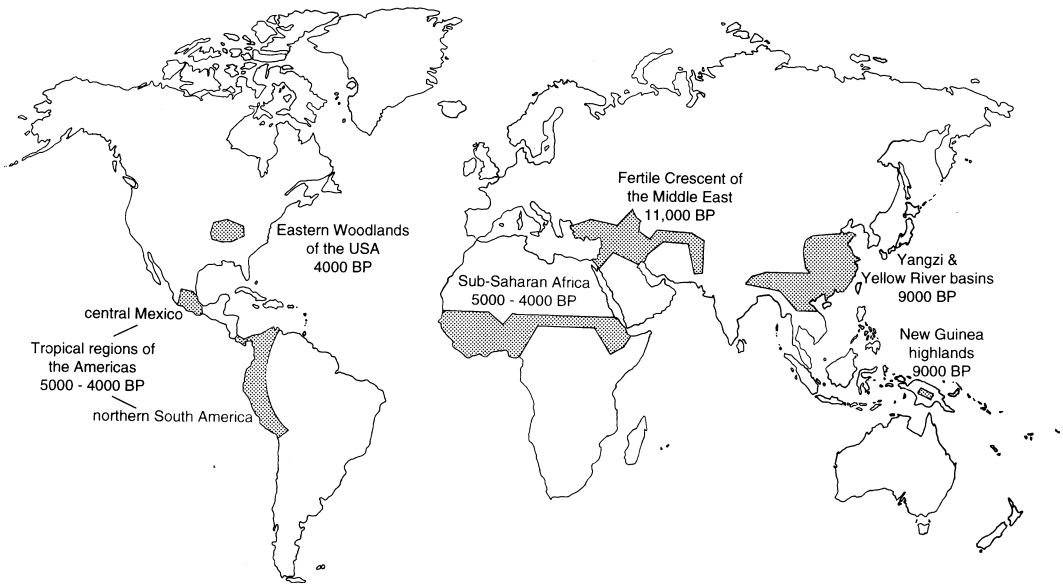

Figure 2
Regions of the world where agriculture is believed to have developed independently

mind here. In the twenty-first century, it is also apparent that lingua francas and national languages can spread rapidly as a result of educational policy, literacy, mass media, and sociolinguistic status; but it is more difficult to imagine such processes of language adoption as being of great significance amongst the small-scale societies of preurban prehistory. Nevertheless, many prehistorians have suggested that language replacement processes of this type, whereby people adopt a language deemed to be of high social status and abandon their original language, have been instrumental in the spread of some families. One such family is Indo-European, for which some linguists and archaeologists have long agreed on a homeland in the steppes north of the Black Sea, followed by a spread into Europe by Late Neolithic and Bronze Age pastoral peoples with domesticated horses and wheeled transport. According to the archaeologist Marija Gimbutas, these people undertook their migrations into Europe between 4500 and 2500 BC, dominating and absorbing the older Neolithic societies in the process.

This view of Bronze Age conquest and language replacement for Indo-European dispersal has been challenged by the archaeologist Colin Renfrew (1987, 1996), who opts instead for an association of early Indo-European with early Neolithic farming dispersal into Europe from Turkey. Nowadays, the idea that many of the major agriculturalist language families spread as a result of the early development of agriculture is taking firmer hold. Farmers typically have larger populations than hunter-gatherers, and if farmers are not enclosed by other farming populations i.e., if they live in a region surrounded essentially by low-density hunter-gatherers, then expansion is a likely outcome, exactly as in the European frontiers in Australia and western North America. So, while the agricultural dispersal hypothesis is no more 'provable' than any other hypotheses to explain language family origins, it does at least have the strong supporting factor of a historically proven mechanism which can allow and encourage population expansion to occur. Such expansion need not mean extinction of all hunter-gatherers. In many ethnographic situations, hunter-gatherers have survived in the interstices of agricultural or pastoralist landscapes, perhaps for millennia.

Put simply, the farming dispersal hypothesis would see the protolanguages for Indo-European, Semitic, Turkic, Sumerian, Elamite, and possibly Dravidian located in the wheat, barley, cattle, and caprine zone in the Middle East, with dispersals occurring mainly in the period between 6500 and 3000 BC. During this time mixed farming became widely established and the archaeological record tells us unambiguously that population was increasing in an overall sense quite rapidly (despite periodic environmental setbacks and short-term population retractions). Sino-Tibetan, Austroasiatic, Austronesian, Tai, and Hmong-Mien would all have begun their dispersal from the region of rice and millet cultivation in China, focused in the middle and lower Yellow and Yangzi valleys, between 5000 and 2000 BC (with Austronesians eventually colonizing the greater part of the Pacific). Niger-Congo (including Bantu) resulted from the development of agriculture in West Africa and the Sahel zone, mainly after 3000 BC, and perhaps following earlier pastoralist dispersals in northeastern Africa by Afroasiatic (Berber, Chadic, Cushitic) and Nilo-Saharan speakers. In the Americas, the Mayan, Otomanguean, Mixe-Zoque, Uto-Aztecan, and Chibchan language families probably spread as a result of agricultural developments in Greater Mesoamerica after 3500 BC. In South America the picture is a little more diffuse, but some of the major Andean and Amazonian families might have spread as a result of the establishment of maize and manioc agriculture after about 2500 BC—examples here would include Quechua and Aymara, and lowland Amazonian families such as Arawak, Carib, and Tupi.

Archaeologically, these suggested language radiations associated with early agricultural societies should be reflected in the distributions of some very widespread archaeological complexes. In particular, it has been noted in many regions that the archaeological complexes of early agricultural phases are much more widespread and homogeneous in content than the highly regionalized complexes of later periods. This appears to be the case in early Neolithic Europe, East Asia, and the Pacific, and amongst the Early Formative cultures of the Americas.

5. *Language Contact and Cultural Contact*

Language and archaeology correlations can be sought not only for the origins and dispersal histories of language families, but can also reflect the contacts which take place from time to time between languages in different families and subgroups. For instance, the Austronesian speakers of New Guinea have been in intense contact for upwards of 2,000 years with the speakers of Papuan languages in several unrelated families. This has led to a great deal of contact-induced change and even language shift, and it is therefore not surprising to discover that distributions of material culture often cross-cut language boundaries. The archaeological record, however, suggests that quite strong differences in material culture would have distinguished Papuan and Austronesian societies 3,000 years ago, at the time of the Lapita archaeological spread through much of the western Pacific. The Lapita spread was probably associated with the initial Austronesian colonization of many of the western Pacific Islands, but it is significant that it appears to have avoided the island of New Guinea itself, where Austronesian speakers even today are found only in a

few pockets of coastal distribution (Kirch and Green 2001).

Indeed, the linguist Robert Dixon (1997) has suggested that the overall history of the major language families can be separated into short periods of widespread dispersal, when the families are actually founded, interspersed with long periods like that described for New Guinea when populations of quite different linguistic and cultural origin interact, whether peacefully or belligerently. This hypothesis resembles the theory of punctuated equilibrium as applied to the biological evolution of species.

A final point to note is that, whereas archaeology and language history can often come together to throw independent light to support a plausible historical reconstruction, we often find that genetic data are not in full agreement. It is not the intention to discuss human genetics here, but it is perfectly obvious that not all of the speakers of some of the major language families are of tightly defined and geographically restricted genetic origin. For instance, the speakers of Austronesian languages range from Southeast Asians to Melanesians and Polynesians. The speakers of Indo-European languages range from northern Europeans to northern Indians. It is possible, but rather unlikely, that these differences represent no more than natural selection operating since the initial population dispersal which founded the language family in question. But it is far more likely that these differences reflect population mixing not always paralleled by an equivalent amount of language mixing. In other words, language families can have a life of their own, as can nodes of biological variation. Many population dispersals must have incorporated large numbers of the existing inhabitants of the newly settled regions, with consequent genetic effects stamped on later generations. This does not mean that there are no correlations between variations in language and biology in the human species, but we must be aware that any correlations will not always be clear-cut and obvious. They must be teased apart with care.

This field of archaeolinguistic research is not one in which we can expect absolute proofs for suggested correlations, particularly when dealing with prehistoric societies, but firm hypotheses are worthy of the research effort. The goals of archaeolinguistic research are laudable ones since they help us to interpret and understand so many fundamental developments and transitions in human prehistory.

Bibliography

Anthony D 1995 Horse, wagon and chariot: Indo-European languages and archaeology. *Antiquity* **69**: 554–65

Bellwood P 1991 The Austronesian dispersal and the origin of languages. *Scientific American* **265**: 88–93

Bellwood P 1995 Language families and human dispersal. *Cambridge Archaeological Journal* **5**: 271–4

Bellwood P 1997 *Prehistory of the Indo-Malaysian Archipelago*, rev. edn. University of Hawaii Press, Honolulu, HI

Bellwood P, Fox J J, Tryon D (eds.) 1995 *The Austronesians*. Department of Anthropology, Research School of Pacific and Asian Studies, Australian National University, Canberra, Australia

Blust R A 1995 The prehistory of the Austronesian-speaking peoples. *Journal of World Prehistory* **9**: 453–510

Childe V G 1926 *The Aryans: A Study of Indo-European Origins*. Kegan Paul, Trench, Trubner, London

Dixon R M W 1997 *The Rise and Fall of Language*. Cambridge University Press, Cambridge, UK

Ehret C, Posnansky M (eds.) 1982 *The Archaeological and Linguistic Reconstruction of African History*. University of California Press, Berkeley, CA

Fiedel S 1991 Correlating archaeology and linguistics: The Algonquian case. *Man in the Northeast* **41**: 9–32

Gimbutas M 1991 *Civilization of the Goddess*. Harper, San Francisco

Higham C 1996 Archaeology and linguistics in Southeast Asia. *Bulletin of the Indo-Pacific Prehistory Asssociation* **14**: 110–8

Hill J 2001 Proto-Uto-Azlecan: a community of cultivators in Central Mexico? *American Anthropologist* in press

Jones W 1993 The third anniversary discourse. In: Pachori S S (ed.) *Sir William Jones: A Reader*. Oxford University Press, Delhi, pp. 172–8

Kaufman T 1976 Archaeological and linguistic correlations in Mayaland and associated areas of Meso-america. *World Archaeology* **8**: 101–18

Kirch P V, Green R C 1987 History, phylogeny and evolution in Polynesia. *Current Anthropology* **28**: 431–56

Kirch P V, Green R C 2001 *Hawaiki, Ancestral Polynesia*. Cambridge University Press, Cambridge, UK

Mallory J P 1989 *In Search of the Indo-Europeans*. Thames and Hudson, London

Mallory J P 1996 The Indo-European phenomenon: Linguistics and archaeology. In: Dani A H, Mohen J P (eds.) *History of Humanity*. UNESCO, Paris, Vol. 3, pp. 80–91

McConvell P, Evans N (eds.) 1997 *Archaeology and Linguistics: Aboriginal Australia in Global Perspective*. Oxford University Press, Melbourne, Australia

Noelli F S 1998 The Tupi: Explaining origin and expansions in terms of archaeology and historical linguistics. *Antiquity* **72**: 648–63

Pawley A K, Ross M 1993 Austronesian historical linguistics and culture history. *Annual Review of Anthropology* **22**: 425–59

Renfrew C 1987 *Archaeology and Language – the Puzzle of Indo-European Origins*. Jonathan Cape, London

Renfrew C 1989 Models of change in language and archaeology. *Transactions of the Philological Society* **87**: 103–55

Renfrew C 1994 World linguistic diversity. *Scientific American* **270**: 116

Renfrew C 1996 Language families and the spread of farming. In: Harris D (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London, pp. 70–92

Ruhlen M 1987 *A Guide to the World's Languages*. Stanford University Press, Stanford, CA, Vol. 1

Thomason S, Kaufman T 1988 *Language Contact, Creolization and Genetic Linguistics*. University of California Press, Berkeley, CA

Zvelebil M 1995 At the interface of archaeology, linguistics and genetics. *Journal of European Archaeology* **3**: 33–70

P. Bellwood

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Archaeology, Politics of

The results of archaeological research have long been used by individuals, groups, and nations for political purposes such as nationalism, colonialism, tourism, and the development of group identities, but archaeologists have generally not seen themselves conducting work which is political in nature. The traditional view of archaeology has been that of a group of scholars working systematically and scientifically to discover and excavate information (usually in the form of objects and monuments) about past societies around the world. What people did with that information once it was published and described was seen as being distinct from what archaeologists did; archaeologists saw themselves as objective in their interpretations and descriptions. This view of archaeology began to change in the 1980s, but became dramatically different by the 1990s, with the development of new attitudes both within and outside archaeology. In the United States, the passage of the Native American Graves Protection and Repatriation Act (NAGPRA) in 1990 was a particularly important milestone.

1. Archaeology and Politics: A Brief Historical View

In discussing nationalism as one political force that has affected archaeology, Trigger (1995) outlines the history of archaeology from this perspective, and the changing views he summarizes represent one example of how politics can affect archaeology.

Trigger notes (1995, p. 266) that throughout antiquity, royal families and ethnic groups strengthened their positions by linking themselves to particular figures or events of the past. During the Renaissance, scholarship was often used more broadly to support political changes by providing precedents from antiquity.

There were several shifts in the way archaeology was used during the Enlightenment. Trigger (1995, p. 267) notes that revolutionary leaders in France supported Napoleon's 1798 invasion of Egypt because they saw ancient Egypt as a source of wisdom. A more scientific archaeology replaced the object-focused archaeology of the previous periods, and the attention of archaeologists shifted to cultural evolution, and to archaeology as a means of documenting the progress of human development. As colonialism expanded, however, the application of theories of cultural evolution, as a benefit to everyone, began to change. Europeans could benefit from cultural progress, but indigenous peoples were viewed as less developed and less capable of development. This racist view helped support the expansion of colonialism, and the then limited archaeological record was used to support this perspective (Trigger 1995, p. 268).

In the 1860s, nationalism took a more prominent role in shaping archaeological research, although its influence depended upon the impact of colonialism, class struggles, and ethnic nationalism (Trigger 1995, p. 269). Archaeology responded by shifting its focus to reconstructing the history of specific peoples or, more formally, the development of culture historical archaeology. Archaeological cultures were identified and defined as early representations of historically known groups. This allowed groups to add to their own history, and glorify themselves in relation to others. In the United States, where there was no bond to the indigenous people being studied, culture history also became popular, primarily because it could account for geographic variation and change that could not be explained by cultural evolution (Trigger 1995, p. 269).

The early twentieth century practice of tying interpretations of past groups to specific indigenous peoples and tribes became less common as archaeology became more complex, and as archaeologists realized that it was difficult to associate an ancient group with a particular modern tribe. Because of intervening years of movement (often dramatic because of forced movements of tribes by the US government) and change, the match was inexact and difficult to support with the kind of scientific certainty that archaeologists would like. Further, archaeologists had begun to ask other kinds of questions beyond whether or not a particular archaeological culture was associated with a particular living one. As the discipline moved in new directions, linking past cultures to modern ones became less and less common. Archaeologists now understand that race, language, and culture are independent variables that can change for different reasons and in many different ways.

During the 1980s and 1990s, another shift in archaeology can be subsumed under the heading of postmodernism. For this discussion, the most important aspect of this development is the rise of a self-reflective archaeology which questions the perspective and political nature of everything that an archaeologist produces, and stresses the importance of cultural relativism, which in its most extreme form, suggests that all archaeological interpretations are subjective, and that any one interpretation is as valid as any other. A major focus of this approach has been to give voice to indigenous groups and their interpretations, as well as to the views of the general public. A recent statement (Shack 1994, p. 115) provides an example:

> Archaeology and history share common features of malleability, continually recreating the past, a principal function of which past is to socially construct the present. … The impulse to preserve the past is part of the impulse to preserve the self, an impulse that is given 'legitimacy' when grounded in objects from the past.

Shack concludes with the observation that constructions and representations will be misconstrued until

indigenous voices are heard unfiltered and in the first person.

2. Changing Attitudes and Times

During the last two decades, with the increasing demand by indigenous groups to be heard and given a say over their past, present, and future, archaeology increasingly found itself in an awkward position. Groups that have been disenfranchised and ignored have moved to empower themselves and develop political strength and attention. Archaeologists, who often see themselves as champions of such groups and as people whose work focuses on the histories of these groups, are cast as enemies. It is often difficult, unfortunately, for disenfranchised groups to get people's attention on issues such as education, food, and healthcare, but it is easier to draw the attention of the media and the public by focusing on human bones. Archaeologists, a not particularly powerful political group, are portrayed as grave robbers and people who ignore the beliefs and desires of indigenous peoples. When combined with the cavalier treatment of native sensibilities and concerns that archaeologists and physical anthropologists exhibited in the late nineteenth and early twentieth centuries, the discipline appears in a bad light. This framing of the problem gets immediate public attention. Bones represent powerful cultural symbols, especially in the United States, and this approach by indigenous peoples is understandable, if uncomfortable for archaeologists and physical anthropologists.

2.1 Who Pays for Archaeological Research?

Although archaeologists have often been uncomfortable with the idea that their work is used for political purposes, because archaeology is important for political and cultural reasons, it is seen as important for the State to fund. As this funding has increased, however, there has been a growing awareness that archaeologists have a direct obligation to the public and to the people being studied. These obligations can perhaps best be represented by two of the principles of archaeological ethics crafted by the Society for American Archaeology (Lynott and Wylie 1995): the principle of accountability, and the principle of public education and outreach. The principle of accountability notes that archaeologists must acknowledge public accountability and must 'make every reasonable effort, in good faith, to consult actively with affected group(s), with the goal of establishing a working relationship that can be beneficial to the discipline and to all parties involved' (Watkins et al. 1995, p. 33). In public outreach and education, archaeologists are told to engage the public in stewardship for the archaeological record, to explain how archaeology is used in understanding human behavior, and also to explain interpretations of the past. There is also a recognition that a variety of different kinds of audiences exist for these efforts (Herscher and McManamon 1995, p. 43).

2.2 Power and Control

Developments and changes in archaeology and in the larger society have resulted in a real shift in power and control. The question now posed is who owns or controls the past, and the shifts in archaeology are similar to those found in other disciplines. Changes in medicine provide a useful analogy.

2.3 An Analogy with Medicine

For generations, physicians used authoritative knowledge in healing patients; they knew best, and our job as patients was to follow directions. Patients were often not even told the details of their condition. Many, however, grew up in communities which had their own medicines and ways of treating illness. The success of such approaches was uniformly rejected by physicians as representative of 'old wives tales' or as being anecdotal in nature. Science held the correct and proven ways to treat illness. Over time, people began to questions physicians more, demanding more say in their treatment and care, and wanting empowerment. In addition, many returned to their traditional approaches to medicine, in part because they worked or worked as well as 'scientific' remedies, and they were often cheaper and easier to use. Eventually, physicians began to accept some of these treatments, indicating that they were not harmful and might have some placebo effect. As more evidence supported the value of traditional approaches, and as patients gained more say in their own care, the views of the medical establishment shifted to the point that alternative medicine is now taught at many medical schools. A number of traditional approaches are covered by insurance, and foundations and federal agencies are funding research in alternative medicine. The trajectory in archaeology is similar, with an equal proportion of professionals who see such changes as wrong and dangerous, those who see them as long overdue, and those who see them as necessary but problematic.

3. Repatriation as an Example

The call for the repatriation or return of Native American human remains, funerary objects, and sacred objects in the United States represents the best

example of change in the political nature of archaeology and the resulting changes in the way that archaeology is conducted.

3.1 The History of Repatriation in the United States

As in medicine, archaeology for many years presented authoritative knowledge about the past. While Native American tribes who lived throughout the US were seen as distant descendants of the people who lived there prior to the 1600s, their perspectives were either ignored or used only as analogy. An important exception was the use of the culture-historical approach, in which the history of modern tribes was traced backward in time to link to earlier groups. Nonetheless, the data used by archaeologists in creating culture histories was generally authoritative knowledge provided by ethnographers, not information directly from tribes.

For many members of native communities, the past, and especially human remains and sacred objects, represent symbols of political and spiritual power, and that power does not diminish over time. The importance attributed to time differs across tribes and cultures, and the distinctions archaeologists see between historic (written records) and prehistoric (prior to written records) are not understood by many who make no distinction between the present and the past. Similarly, the distinctions we make between the written and oral record vary in their relevance to traditional groups: many see the two as equally valid. How can such dramatic differences be reconciled?

Price (1991) outlines a history of repatriation law and policies, and provides a summary of federal laws and policies and of state laws. He notes that for many years tribes focused on their diversity and differences, but legal procedures such as the Indian Claims Commission, established in 1946, educated Indians to the effects and advantages of legal representation (Price 1991, p. 10). Later, an event that many believe began a national Indian movement, the American Indian Chicago Conference, was organized in 1961 by anthropologist Sol Tax (Price 1991, p. 10). The conference brought together over 90 tribes and bands for the first time. Subsequent advances in technology and communication made sharing of knowledge between tribes much faster and easier, including increased communication internationally. New laws were the eventual result.

The federal laws which provide the clearest recognition of aboriginal rights and interests in human remains and sacred objects are the National Museum of the American Indian Act of 1989 (which applies only to the Smithsonian Insitution) and the Native American Graves Protection and Repatriation Act (NAGPRA) of 1990 (which excludes the Smithsonian). More recent revision of the Museum of the American Indian Act has made the two laws more comparable in scope, but the Smithsonian has a separate review panel. Both panels are composed of representatives of native communities and scientific organizations, with native representatives holding the majority.

Some archaeologists viewed NAGPRA as designed to empty museums of all of their collections, but ignored the second part of the law which may have more far-reaching implications. Under NAGPRA, Native American cultural items and remains discovered or excavated on federal or tribal lands are under the control of Native American groups, listed in a priority order. Many states have followed suit, writing or revising laws to mirror this procedure. Ultimately, there may be very few places that archaeologists can excavate without the direct involvement of native groups.

4. The Future of Archaeological Research in a Political World

The museums of the US have not been emptied by the introduction of NAGPRA, but this does not mean that the law has not had a permanent effect on the conduct of archaeology. Archaeologists have been forced to change, and to acknowledge that 'The past belongs to everyone' (Lowenthal 1981, p. 236). Sharing control of the past is not easy, and archaeologists have had to learn to change their approaches and their methods of communication in order to level the playing field (see Leone and Preucel 1992, Goldstein 1992 for examples). Equally significantly, archaeologists have had to expand the lines of evidence they use to develop interpretations of the past, and the potential of oral traditions is one of the most exciting and difficult areas to incorporate. The shifts in archaeology do not mean that any view is as valid as any other view, but rather that archaeologists must realize that their work will be used for political purposes, they must take a more active role in directly involving Native American communities in their work, they must acknowledge and address the many publics interested in the past, they must continually expand the lines of evidence they employ, and they must remember that a static view of the past does not allow anyone to learn.

See also: Aboriginal Rights; Anthropology, History of; Cultural Relativism, Anthropology of; Cultural Resource Management (CRM): Conservation of Cultural Heritage; Environmentalism: Preservation and Conservation

Bibliography

Goldstein L 1992 The potential for future relationships between archaeologists and Native Americans. In: Wandsnider L (ed.) *Quandaries and Quests: Visions of Archaeology's Future*. Center

for Archaeological Investigations, Southern Illinois University at Carbondale, IL
Herscher E, McManamon F P 1995 Public education and outreach: the obligation to educate. In: *Ethics in American Archaeology: Challenges for the 1990s*. Society for American Archaeology, Washington, DC
Leone M P, Preucel R W 1992 Archaeology in a democratic society: a critical theory perspective. In: Wandsnider L (ed.) *Quandaries and Quests: Visions of Archaeology's Future*. Center for Archaeological Investigations, Southern Illinois University at Carbondale, IL
Lowenthal D 1981 Conclusion: dilemmas of preservation. In: Lowenthal D, Binney M (eds.) *Our Past Before Us: Why Do We Save It?* Temple Smith, London, pp. 213–37
Lynott M J, Wylie A (eds.) 1995 *Ethics in American Archaeology: Challenges for the 1990s*. Society for American Archaeology, Washington, DC
Price H M 1991 *Disputing the Dead: US Law on Aboriginal Remains and Grave Goods*. University of Missouri Press, Columbia, MI
Shack W A 1994 The construction of antiquity and the egalitarian principle: social constructions of the past and present. In: Bond G C, Gilliam A (eds.) *Social Construction of the Past: Representation as Power*. Routledge, London, pp. 113–8
Trigger B G 1995 Romanticism, nationalism, and archaeology. In: Kohl P L, Fawcett C (eds.) *Nationalism, Politics, and The Practice of Archaeology*. Cambridge University Press, New York, pp. 263–79
Watkin J, Goldstein L, Vitelli K, Jenkins L 1995 Accountability: responsibilities of archaeologists to other interest groups. In: Lynott M J, Wylie A (eds.) *Ethics in American Archaeology: Challenges for the 1990s*. Society for American Archaeology, Washington, DC

L. Goldstein

Archaeometry

1. Archaeometry

'Archaeometry' is a specialized discipline within archaeology in which various scientific methods of chemical and physical analysis are applied to archaeologically derived materials. Archaeometry therefore centers on research whose aim is to explain and test archaeological questions about ancient things or phenomena related to human cultural activities. The research measures or quantifies parameters using analytical techniques borrowed from earth science, chemical, biological, and other scientific disciplines. The field of archaeometry includes, for example, determining the ages of sites and artifacts, sourcing objects to the original localities of raw materials, identifying the components and processes involved in converting earth materials into metals and ceramics, and determining patterns of dietary exploitation. Many analytical techniques have been applied in these investigations and some common ones are briefly described below to illustrate the diversity of archaeometric techniques.

1.1 Age Determinations

Probably the first example of the use of archaeometrical methods was the realization that the annual growth rings of trees could be used to determine the age of construction of prehistoric pit houses in the southwest of the USA using cores taken from wooden beams (Douglass 1936). The tree rings were also shown to indicate variations in climate within the time span of the life of the tree (Judd 1954). This method of tree-ring dating, or dendrochronology, is still widely used to date habitation sites in the Americas and in Europe.

The dating of carbon-bearing substances associated with archaeological deposits has revolutionized the determination of absolute ages in archaeology. Radiocarbon dating (Libby et al. 1949) has been used to place time markers on important periods of human activities and climatic change, and to date the extinctions of animals, for example the woolly mammoth and sabre-toothed tiger (Ho et al. 1969). Refinements to the method and development of the accelerator mass spectrometer (AMS) have allowed for the dating of microgram quantities of carbon (Nelson et al. 1977). For example, a famous case involves the AMS radiocarbon dating of the Shroud of Turin (Damon et al. 1989). While many people believe the shroud was wrapped around Christ's body, fragments of its linen fibers were dated independently by three radiocarbon laboratories to the late thirteenth century.

Archaeologists have also been able to date individual seed grains, beeswax resins (Nelson et al. 1995), charcoal scrapings from Paleolithic cave paintings (Valladas et al. 1992), and buried Australian rock art (Watchman 1993). Pigment painted on rock in South Africa was the first application of the AMS radiocarbon technique for dating rock art (Van der Merwe et al. 1987). Innovative methods have since been developed for extracting carbon from paintings. Oxidation of carbon compounds associated with pigments or rock surface mineral deposits using either an oxygen plasma (Russ et al. 1990), laser energy (Watchman et al. 1993), or permanganate chemistry (Gillespie 1997) can be used to prepare samples prior to AMS radiocarbon dating. Rock art in Texas has been dated to between 4,100 to 3,200 years ago (Russ et al. 1990). Rock paintings in northern Australia have been dated using carbon in rock surface encrusted 'canvasses' and mineral coatings (Watchman 1993) and in plant fibers used as paint binders (Watchman and Cole 1993).

Analytical methods employing light (luminescence) are also being used to date sediments in occupation shelters and in pottery. The basis of the luminescence methods is that natural radiation provides energy to the electronic structure of some crystals, particularly quartz and feldspar minerals (Aitken 1985). Grains that are heated (thermoluminescence) or illuminated by green light (optical luminescence) emit small amounts of light that reflect the level of radiation and

length of time since they were incorporated in a sediment or pot. In a controversial case the floor sediments containing stone artifacts at Jinmium in northern Australia were dated at 116,000 years old using thermoluminescence (TL) and quartz grains (Fullagar et al. 1997). The age of those sediments were disputed by proponents of an optical stimulated luminescence method (OSL) who found that the maximum age of the deposits was only 10,000 years (Roberts et al. 1998). Arguments about the reliability of the age determinations for the sediments are concerned essentially with bulk-samples versus single-grain analyses, incomplete bleaching by sunlight of the luminescence centers in quartz grains generated by *in situ* disintegration of rock fragments compared with well-bleached sand grains. These controversial situations and measurements highlight the complex nature of many archaeometric techniques, and indicate the potential problematical results obtained from applications where the experimental conditions are not well known.

2. *Identifying the Sources of Artifacts*

Determining the characteristics of raw materials obtained from quarries and recognizing the same attributes in archaeological objects allows an archaeologist to describe trading routes and interactions between groups of people. The simplest archaeological technique is to look at earth materials using a microscope. Such petrological analyses can be used, for example, to identify the rock types selected for production of prehistoric polished edge-ground axes. The texture and assemblage of minerals in stone tools can be compared with similar features in rocks at known quarry sites. This method was used to indicate complex prehistoric trading patterns in southeastern Australia. Hard, fine-grained hornfels from the Mt. William quarry was traded across hundreds of kilometers (McBryde 1984). Similarly, Neolithic communication networks and boundaries were defined in England based on the petrological analysis of edge-ground hand axes (Cummins 1980).

Subatomic particles can also be used to provide geochemical information about rocks and minerals to substantiate petrological and stylistic information about artifacts. For example, in the proton induced X-ray emission (PIXE) analysis of artifacts, high-energy protons are used to induce X-ray excitations from a range of elements. The measurement of multiple trace element abundances in artifacts, waste materials, and known sources can differentiate between quarries and can allocate artifacts to individual deposits. An example of this archaeometric method is the characterization of obsidian artifacts and sources for the investigation of the production, trade, and patterns of consumption of that natural resource in Melanesia. The systematic analysis of obsidian deposits and excavated flakes on New Britain have demonstrated that two exposures and possibly a third site supplied most of the obsidian for use as stone tools for more than 11,000 years (Summerhayes et al. 1993). Another analytical technique employing subatomic particles is neutron activation. Energetic neutrons, like the protons in PIXE, allow the measurement of elemental abundances in artifacts. Neutron activation analysis (NAA) has been used to source Late Neolithic and Early Bronze Age obsidian in Macedonia (Kilikoglou et al. 1996). Sources more than 300 kilometers north and south of the archaeological site at Mandalo were shown to have provided raw obsidian for use as artifacts.

Neutron activation analysis has also been used to characterize steatite (soapstone) sources in eastern North America (Truncer et al. 1998). Relatively small quantities of rock can be analyzed using the NAA method to produce gamma rays from many elements, which are measured over four weeks. A wide range of trace and major elements, as well as rare earth elements can be measured after a single long irradiation in a flux of neutrons from a reactor. Statistical analysis of the large numbers of analytical measurements eases the burden of discriminating between quarry materials, and permits the identification or 'fingerprinting' of elements that are characteristic of each quarry.

The chemical and mineralogical characteristics of Australia ochers have been determined to discriminate between various sources and to confirm ethnographic accounts of long-distance trade. PIXE analyses have been used to characterize ocher sources in central Australia, with implications for defining trading networks and, delineating boundaries between Aboriginal populations (Smith et al. 1998). X-ray fluorescence has also provided major and trace element analyses, and Rietveld X-ray diffraction has identified the principal mineral phases for each known ocher source in southern Australia (Jercher et al. 1998). Ochers smeared on bones and objects have been traced to potential sources within geologically defined areas. The high degree of natural variability in ocher compositions makes such sourcing studies extremely challenging, so much so that it may be easier to exclude certain locations rather than identify specific sources.

Chemical analyses of other materials can also be used by the archaeometrist to provide useful information for the archaeologist. For instance, the minor ingredients in glass, especially the presence of lead, can be used to indicate the likely sources of glass production and the existence of trading networks. The ecclesiastical glass found at Koroinen, Finland illustrates how the tools of archaeometry provide insights into medieval manufacturing and trading processes (Kuisma-Kursula and Raisanen 1999). Using X-rays generated from focused electrons in a scanning microscope and a beam of protons directed onto small glass fragments, the average chemical composition of the Finnish glasses were found to be remarkably

similar to northwestern European glasses. The abundances of minor elements, particularly lead, sodium, and calcium, in medieval Finnish glasses are inconsistent with likely Russian sources, but match German and southern European glasses. The indication is that glass-making was not practiced in Finland at that time, but that supplies of colored glass for Finnish monasteries depended on trade links with Western and Central Europe.

3. *Metallurgy*

Unlike other geochemical sourcing studies that rely on finding significant archaeological variations in the intrinsic trace element compositions between different sources, the study of metal processing is far more challenging for an archaeometrist. Concentrations of trace elements vary between the ores and the processed metal, and complications arise because of the introduction of fluxes and refractory components. Lead isotope analyses, on the other hand, provide an alternative means for sourcing alloyed or leaded artifacts because the relative proportions of isotopes from ores to artifacts are not measurably affected by chemical and pyrometallurgical processes (Srinivasan 1999). For example, thermal ionization mass spectrometry produces lead isotope ratios ($Pb^{208/207}$ and $Pb^{207/206}$) which were used to discriminate between known deposits of Indian lead ores and artifacts made from them. Matching artifacts to specific ore deposits using the isotopic ratios resolved contentious chronological issues based on art-historical criteria. Western Indian lead ores rather than local sources were found to have been used for brass in northern India during the premedieval period, and also in lead and brass in southern India during the medieval period. Recycling of materials was also observed in Indian coins where lead isotope analyses of later bronzes fitted the trends established for earlier groupings.

4. *Ceramics*

Geochemical studies of pottery and porcelain have focused mainly on grouping complete objects or sherds into products that were either made locally or imported (see Ceramics in Archaeology). This has been done to establish or confirm suspected cultural links and trading associations between groups of people. Basic analyses of pottery sherds usually include determining the mineralogical compositions of the clay and temper to find out what specific materials were used and where a pot was made. This is done using petrographic and scanning electron microscopy. More detailed geochemical analyses may include the use of inductively coupled plasma emission spectroscopy, NAA, PIXE, or gamma-ray emissions. An example of one of these techniques is the confirmation of the extent of local production of fineware between the seventh and second centuries BC at a village on the Calabrian coast, Italy (Mirti et al. 1995). Often, distinctive patterns of elemental abundances are not readily evident in the large amount of geochemical data that is collected. Multivariate statistics are needed to separate disparate groups of pots, as in the case of Roman Samian pottery (Argyropoulos 1995).

5. *Paleodiets*

Adaptation by people to changing environments and the transition in prehistoric societies from hunting and gathering to agricultural subsistence are topical themes in archaeology. These changes can be better understood through the measurements of stable carbon ($*^{13}C$) and nitrogen ($*^{15}N$) isotopes obtained from skeletal materials, and edible plants and animals (DeNiro and Epstein 1978, 1981). Stable carbon isotopes were first used to study the introduction of maize into northeastern North America (Vogel and Van der Merwe 1977). Their use is based on the observation that different groups of plants differ markedly in their isotopic compositions, and that therefore animals that live mainly on particular plants will have bones of matching composition. For example, studies of the paleodiets of people occupying southern Ontario, Canada (Katzenberg et al. 1995) and coastal New England, USA (Little and Schoeninger 1995) have shown reliance on animal proteins rather than on maize or legumes during the Late Woodland period.

These examples of the applications of various analytical techniques reveal how archaeometry plays an important role in archaeology. Archaeometric methods are not used in isolation, but complement a range of other archaeological observations that build a more complete picture of the past. Archaeometry therefore provides a variety of tools that allow archaeologists to understand better human relics of the past.

See also: Bioarchaeology; Ceramics in Archaeology; Environmental Archaeology; Geoarchaeology

Bibliography

Aitken M J 1985 *Thermoluminescence Dating*. Academic Press, London

Argyropoulos A 1995 A characterization of the compositional variations of Roman Samian pottery manufactured at the Lezoux production center. *Archaeometry* **37**: 271–85

Cummins W A 1980 Stone axes as a guide to Neolithic communications and boundaries in England and Wales. *Proceedings of the Prehistoric Society* **46**: 45–60

Damon P E, Donahue D J, Gore B H, Hatheway A L, Jull A J T, Linick T W, Sercel P J, Toolin L J, Bronk C R, Hall E T, Hedges R E M, Housley R, Law I A, Perry C, Bonani G,

Trumbore S, Woelfli W, Ambers J C, Bowman S G E, Leese M N, Tite M S 1989 Radiocarbon dating of the Shroud of Turin. *Nature* **337**: 611–5
DeNiro M J, Epstein S 1978 Influence of diet on the distribution of carbon isotopes in animals. *Geochimica et Cosmochimica Acta* **42**: 495–506
DeNiro M J, Epstein S 1981 Influence of diet on the distribution of nitrogen isotopes in animals. *Geochimica et Cosmochimica Acta* **45**: 341–51
Douglass A E 1936 The Central Pueblo Chronology. *Tree Ring Bulletin* **2**: 29–34
Fullagar R L K, Price D M, Head L M 1996 Early human occupation of northern Australia: Archaeology and thermoluminescence dating of Jinmium rock-shelter, Northern Territory. *Antiquity* **70**: 751–73
Gillespie R 1997 On human blood, rock art and calcium oxalate: Further studies on organic carbon content and radiocarbon age of materials relating to Australian rock art. *Antiquity* **71**: 430–7
Ho J Y, Marcus L F, Berger B 1969 Radiocarbon dating of petroleum-impregnated bone from tar pits at Rancho La Brea, California. *Science* **164**: 1051–2
Huntley D J, Godfrey-Smith D I, Thewalt M L W 1985 Optical dating of sediments. *Nature* **313**: 105–7
Jercher M, Pring A, Jones P G, Raven M D 1998 Rietveld X-ray diffraction and X-ray fluorescence analysis of Australian Aboriginal ochres. *Archaeometry* **40**: 383–401
Judd N M 1954 *The Material Culture of Pueblo Bonito*. Smithsonian Miscellaneous Collections 124, Washington, DC
Katzenberg M A, Scwarcz H P, Knyf M, Melbye F J 1995 Stable isotope evidence for maize horticulture and paleodiet in southern Ontario, Canada. *American Antiquity* **60**: 335–50
Kilikoglou V, Bassiaskos Y, Grimanis A P, Souvatzis K, Pilali-Papasteriou A, Papanthimou-Papaefthimou A 1996 Carpathian obsidian in Macedonia, Greece. *Journal of Archaeological Science* **23**: 343–49
Kuisma-Kursula P, R≅is≅nen J 1999 Scanning electron microscopy-energy dispersive spectrometry and proton induced X-ray emission analyses of medieval glass from Koroinen (Finland). *Archaeometry* **41**: 71–9
Libby W F, Anderson E C, Arnold J R 1949 Age determination by radiocarbon content: world-wide assay of natural radiocarbon. *Science* **109**: 949–52
Little E A, Schoeninger M J 1995 The Late Woodland diet on Nantucket Island and the problem of maize in coastal New England. *American Antiquity* **60**: 351–68
McBryde I M 1984 Kulin greenstone quarries: The social contexts of production and distribution for the Mt William site. *World Archaeology* **16**: 267–85
Mirti P, Casoli A, Barra Bagnasco M, Preacco Ancona M C 1995 Fine ware from Locri Epizephiri: A provenance study by inductively coupled plasma emission spectroscopy. *Archaeometry* **37**: 41–51
Nelson D E, Korteling R G, Stott W R 1977 Carbon-14: Direct detection at natural concentration. *Science* **198**: 507–8
Nelson D E, Chaloupka G, Chippindale C, Alderson M S, Southon J R 1995 Radiocarbon dates for beeswax figures in the prehistoric rock art of northern Australia. *Archaeometry* **37**: 151–156
Roberts R, Bird M, Olley J, Galbraith R, Lawson E, Laslett G, Yoshida H, Jones R, Fullagar R, Jacobsen G, Hua Q 1998 Optical and radiocarbon dating at Jinmium rock shelter in northern Australia. *Nature* **393**: 358–62
Russ J, Hyman M, Shafer H J, Rowe M W 1990 Radiocarbon dating of prehistoric rock paintings by selective oxidation of organic carbon. *Nature* **348**: 710–11
Smith M A, Fankhauser B, Jercher M 1998 The changing provenance of red ochre at Puritjarra rock shelter, central Australia: Late Pleistocene to Present. *Proceedings of the Prehistoric Society* **64**: 275–92
Srinivasan S 1999 Lead isotope and trace element analysis in the study of over a hundred south Indian metal coins. *Archaeometry* **41**: 91–116.
Summerhayes G R, Gosden C, Fullagar R, Specht J, Torrence R, Bird J R, Shahgholi N, Katsaros A 1993 West New Britain obsidian: Production and consumption patterns. In: Fankhauser B L, Bird J R (eds.) *Archaeometry: Current Australasian Research*. Department of Prehistory Research School of Pacific Studies, Australian National University, Canberra, pp. 57–68
Truncer J, Glascock M D, Neff H 1998 Steatite source characterization in eastern North America: New results using instrumental neutron activation analysis. *Archaeometry* **40**: 23–44
Valladas H, Cachier H, Maurice P, Bernaldo de Quiros F, Clottes J, Cabrera Valdes V, Uzquiano P, Arnold M 1992 Direct radiocarbon dates for prehistoric paintings at the Altamira, El Castillo and Niaux caves. *Nature* **357**: 68–70
Van der Merwe N J, Seely J, Yates R 1987 First accelerator carbon-14 date for pigment from a rock painting. *South African Journal of Science* **83**: 56–7
Vogel J C, Van der Merwe N J 1977 Isotopic evidence for early maize cultivation in New York State. *American Antiquity* **42**: 238–42
Watchman A L 1993 Evidence of a 25,000-year-old pictograph in Northern Australia. *Geoarchaeology* **8**: 465–73
Watchman A, Cole N 1993 Accelerator radiocarbon dating of plant-fibre binders in rock paintings from northeastern Australia. *Antiquity* **67**: 355–8
Watchman A L, Lessard A R, Jull A J T, Toolin L J, Blake W Jr 1993 Dating of laser-oxidized organics. *Radiocarbon* **35**: 331–3

A. Watchman

Architectural Psychology

1. *Definition*

Architectural psychology may be defined as that field within the discipline of applied psychology which deals directly with the response of people to designed environments. In this way, architectural psychology is differentiated from environmental psychology (see *Environmental Psychology: Overview*). The latter may be found under the appropriate headings in this encyclopedia. Generally, the primary focus of architectural psychology has been on cognitive and affective responses to conditions which are, at least partly, under the control of building designers. Responses to attributes of enclosure (shapes, colors, sounds, temperature, lighting, degree of complexity, and so on),

control of interaction with others, and ability to find one's way around buildings (both with signposting and without) are three broad categories which have enticed many researchers. Responses of specific groups to buildings commonly used by them—children and schools, university students and dormitories, the sick and hospitals, workers and offices—have often established areas of study.

2. History

The identification of a subfield within psychology to be called environmental psychology, and its own subfield, architectural psychology, dates from the last half of the twentieth century only. The latter subfield seems to have been inspired by interests among architects, and by the potential for practical application of psychological theorizing, at a period when applied psychology was flourishing.

The first architectural psychology program was established in the USA in the early 1960s at the University of Utah. It was funded as a focus for investigation of the relationship between psychiatric disorder and physical environments. In the UK, recognition of a 'new' approach to architectural issues was given in 1965 by the 'journeyman's' architectural journal, *The Architects' Journal*, in an article by a psychologist, B. W. P. Wells. Wells had been involved in an interdisciplinary study of a workplace—an office building (Manning 1965). Two members of that research team were involved in the establishment of the Building Performance Research Unit at the University of Strathclyde in Glasgow in 1967, and members of that unit organized the first British architectural psychology conference in 1969. Examination of the proceedings shows that participants shared a common orientation to person–environment interaction: an orientation which assumed a close causal relationship between the physical environment and individual behavior.

Also in 1969, the first conference of the Environmental Design Research Association (EDRA) was held in Raleigh, North Carolina. Assumptions of the value of empirical (and preferably experimental) psychological methods are evident in the reports. The results of findings would result in 'better' architectural decision making. This was the hope. There was also a hope that the interaction between many physical, social, and psychological variables would be increasingly understood. To quote one widely used textbook in the discipline whose authors were pioneers in the field (Ittelson et al. 1974, p. 9):

> We are dealing with a theory of environment that removes the individual from the physical isolation in which he is usually studied. This in itself is a significant advance in our understanding of human behavior.

Perhaps unfortunately, this humanistic and/or ecological orientation has not been evident in all, or even the majority, of the published work in the area. Analysis of the content of the early editions of the first specialized journal in the field (of environmental psychology) *Environment and Behavior*, from 1969, indicates that the majority of published articles presuppose the possibility of manipulation of people, in one way or another. The underlying assumption seems to be that human behavior in relation to the environment is capable of being understood in terms of a discrete number of variables (then-developing methods of statistical manipulation of multivariate data, such as factor analysis, had an influence on research design, too). If the assumption were true, then control of behavior by manipulation of these variables becomes a possibility. With control comes power; a valued commodity within professions, and something which could have market value (however ethically dubious its use value).

In the early 1970s, the first reports appeared from Australia and Scandinavia. David Canter (a student of Wells' and subsequently influential in the field) ran a course in architectural psychology in Sydney, Australia in 1971. A conference was organized in Lund, Sweden in 1973, and there the influential reported work on color and measurement techniques derived from a cognitive orientation. Generally, the subdiscipline developed later in continental Europe, though by the late 1970s, *psicologia ambientale* was a recognized subfield in Italy, for example. The International Association for the Study of People and their Physical Surroundings (IAPS), with significant European membership, was founded in 1981.

Some ideas that were not seen by their authors as 'environmental' were influential in shaping the subdiscipline. George Kelly is one example. His personal construct theory of the 1950s focuses on the individual's understanding of his or her world, but was adapted to architectural psychology by several researchers. Roger Barker's behavior-setting theory was not well developed in respect of physical environmental components (a deficiency recognized by Barker), but the theory has powerful explanatory potential for understanding physical environments. In methodology, too, the interviewing techniques developed by Carl Rogers have highlighted the importance of listening skills, and have been used to focus attention on users' interpretations of their life world, for example. These three investigators share a humanistic orientation. Their central concern is not prediction, but understanding. Researchers using this approach use methods which are almost always qualitative, impossible to reduce to general hypotheses about the relation between people and their physical world (indeed, the very act of distinguishing the two is an error, seen from the standpoint of some). Prediction is not understood to be the major objective. The majority of published research, however, is not of this kind, but falls within the rubric of cognitive psychology, with its empirical/experimental biases.

3. Specific Areas of Significance in Architectural Psychology

As indicated above, a number of specific fields of psychological study have been seen to have potential value for architects. The most obvious are:

(a) Perception/cognition;
(b) Color;
(c) Proxemics (the study of people spacing, and including studies of crowding and privacy);
(d) Wayfinding;
(e) Affect (the relation of emotion and/or mood to variations in physical environment).

These fields are not autonomous. Color is widely believed to influence emotional response (and there is some experimental support for this view). Perception is involved in establishing whether one feels crowded. Some studies attempt to consider the complexity of environments in social terms as well as physical. The variations in the literature are significant. Nonetheless, the five fields listed provide a simple taxonomy which includes a very large number of reported studies and texts.

3.1 Perception/Cognition

The variety of approaches to perception reflects differing ideas of what is important. No approach is specific to architectural psychology. Kaplan and Kaplan (1981) focus on functioning in the world. They ask the question: 'What does the process of perception/cognition help us to do?' The answer given is that what is important in real life is functioning—and with minimum stress, if possible. People search to make the world familiar, so that they may move smoothly and confidently through it. Experience is organized, so that we learn what constitutes salient information. Thus, the way the world seems is directly related to the way we process received environmental information.

By an understanding of these processes, designers might be enabled more confidently to solve problems with people in mind.

3.2 Color

Few subjects have so great a potential for claiming the interest of both designers and architectural psychologists as does the topic of subjective responses to various colors. There has been a long history of attempts to 'prove' (show scientifically) certain beliefs about the effect of color on people, generally with little success, or with little relation to everyday environments.

Taking an experiential perspective, it might be claimed that people do know about color in their everyday lives, that it is part of their lived experience. Given this way of thinking about color, it might be expected that there would be some phenomenological accounts of color that might lead to greater understanding, if not to prediction, of emotional response to colored environments. Yet these are not easy to find.

It is 'common knowledge' that perception of red is 'arousing' and of green is 'calming' or 'soothing.' It seems that 'common knowledge' is not necessarily correct. Working with strict experimental controls and with a strong physiological bias, Mikellides (1990) found that it is not hue that directly influences responses but chromatic strength. Pale colors are less arousing, whatever the hue.

When actual interiors are considered, the limitations of experimental studies become obvious, however. 'Real' interiors inevitably display a range of colors, in which a single dominant color is extremely rare. Psychologists are still a long way from being able to say very much of value to designers about color perception or response to colored interiors. Thus, several commonly cited texts in the field of architectural psychology make no reference to color in their index.

3.3 Proxemics

The study of social spacing in humans (proxemics) was an early focus of study. Studies of crowding, for example, illustrated that the phenomenon was not simply a function of the density of occupancy. It is now recognized that there are aspects of crowding which were often ignored in early research—situational, affective, and behavioral—and that even knowledge of the response of others in similar situations can modify behavior.

Generally, the extent to which individuals feel in control in relation to their spatial environment has much to do with their satisfaction, but studies of privacy have illustrated its complexity in terms of interacting cultural, personal, and physical conditions, and no simple hypotheses cover all cases, except perhaps the following: The critical characteristic of human behavior in relation to the physical environment is control—the ability to maximize freedom of choice. Clearly, this is not very helpful to designers, or even to, say, office managers, who are almost certainly going to want to limit the freedom of subordinate workers, for example.

Recognizing the application problem, Robert Sommer, who completed the original research, observing the way that students occupied seats at library tables—a kind of territorial behavior—wrote (in Lang et al. 1974, pp. 205–6):

> When I did this research originally, I believed it would be of use to architects. Since architects were concerned with designing spaces and this research was concerned with space, there must be something useful in it for architects. Looking back I think this assumption was, if not unwarranted, at least overoptimistic.

Nonetheless, Sommer still believes that architectural psychology offers architects the means to avoid making false assumptions about the ways in which their decisions will influence behavior.

Irwin Altman deserves special mention in this context. His 1975 book on the material was broadly influential, as was a 1977 edition of the *Journal of Social Issues* (33, 3) devoted entirely to privacy theory and research (and to which Altman contributed).

3.4 Wayfinding

The task of finding one's way to a desired destination in complex environments has interested researchers because it has clearly defined preferred behavioral outcomes and a relatively limited number of physical variables which affect the outcome. These variables are mainly maps and signs of various kinds, but 'building legibility' is rather more complex, and concerns the organizing principles people use to make sense of the buildings which they occupy. Factors such as spatial landmarks and spatial distinctiveness and cognitive mapping (or the way in which individuals order, store, and retrieve their cognitions) have been central to numerous studies.

In solving wayfinding design problems, information-processing theories of perception/cognition can be applied. As hinted at in Sect. 3.1 above, it seems that there are no features of the environment which are essential to cognition. An adequate sample of characteristic features is enough. Taken together these allow for recognition. The environment is diverse, complex, and uncertain. Situations do not repeat exactly. Despite this, adequate theory can lead to the making of choices within the physical environment which reduce ambiguity and make for greater legibility within large buildings.

3.5 Affective Response to Environments

For some time, in the 1970s especially, there were high hopes held by some psychologists that physical environments could be 'measured' using various polar verbal scales (semantic differentiation scales). J. A. Russell was a prominent researcher in this area, and developed the view that responses based on two key dimensions—arousal and pleasure—accounted for most of the variability in human responses. For Russell, the two are independent bipolar variables (orthogonal if graphically represented). Other researchers in the area (without apparent interest in architectural psychology) lend support to the idea that affective response to building environments could be 'measured' by measurement of internal arousal states. Further, there seems some support for the idea that self-report methods (say the adjectival scales) correlate pretty well with physiological measures. Despite this, in recent times, there have been very few reported attempts to investigate responses to interior spaces using such techniques. Perhaps, because mood is not usually related to a particular stimulus in field situations, and the range of responses within a given setting is likely to be large, there has been a loss of confidence that any variation is attributable to the physical environment.

4. Phenomenological Approaches

Although much of the theoretical work and methods adopted in research designs derive from a cognitive orientation, mention must be made of the contribution of phenomenological approaches. Many explicitly qualitative studies have been reported by people whose first interest is not psychology, but David Seamon (1982) has explained how a humanistic psychology which emphasizes intersubjectively shared understanding has contributed to understanding of place. This is especially the case when considering responses to the concept 'home,' for example.

5. Postoccupancy Evaluation

One area where architectural psychology has changed the behavior of architects, to a degree, is in the evaluation of buildings after completion and a period of use. The argument in favor of undertaking post-occupancy evaluation (POE) studies goes as follows. It is assumed that all building designers have intentions in respect of their designs in relation to human behavior (which, of course, for the psychologist includes a wide gamut, including thinking). There is little evidence that designers ever follow up on these assumed behavioral responses to check whether their intentions are reflected in actual outcomes. POE is intended as a method of rectifying that perceived lack of feedback. Without feedback, each new design is a set of untested behavioral speculations. If there is to be cumulative knowledge in design, there have to be some supportable hypotheses about the relation between designed environments and behavioral outcomes.

All of the above assumes there is a direct relationship between physical environments and behavior which is causal. No determinism is implied here. The best way to understand the relationships involved is to consider the idea of 'affordance': does the physical environment help or hinder people in achieving their preferred behavioral outcomes (whether action, cognition, or affect)?

In research, a variety of methods have been adopted. However, in many instances, in the practice of architecture, feedback is limited to forced choice pencil-and-paper responses and the findings have more to do with meeting requirements of quality assurance procedures of bureaucracies than they have with ensuring a closer fit between agreed design criteria and building outcomes.

6. Future Directions

A recent textbook on environmental psychology (Bonnes and Secciaroli 1995) suggests, by its content, that the impetus for new research in the subdiscipline has lessened in recent years. While the book covers much more than just architectural psychology, it is of interest that, of close to 600 citations, fewer than 20 percent are dated 1985 or later, while a third of all of the citations date from the 1970s. This hardly suggests a growing and expanding subdiscipline.

The potential for failure in the dialog between the social and behavioral science and architecture was summarized by Jameson (1970). A breakdown in communication was evident in 1970, as it still is. In the late 1960s and 1970s, it would seem that schools of architecture had high hopes for the potential of architectural psychology. Courses in architectural psychology were incorporated in the curriculums of many schools, either as core material or as significant option streams. Nearly all have disappeared. While the work of some psychologists might have led to greater understanding of, and empathy for, users of buildings, such understanding seems to have done little to inform the practices of architects (although there may be greater concern for users in some practices). Without the impetus of potential applications for research findings, it is not clear what directions architectural psychology will take.

See also: Architecture; Community Environmental Psychology; Environmental Psychology: Overview; Residential Environmental Psychology

Bibliography

Altman I 1975 *The Environment and Social Behavior: Privacy, Personal Space, Territoriality and Crowding*. Brooks/Cole, Monterey, CA

Bonnes M, Secciaroli G 1995 *Environmental Psychology: A Psycho-social Introduction*. Sage, London

Fisher J D, Bell P A, Baum A 1984 *Environmental Psychology*, 2nd edn. Holt, Rinehart and Winston, New York

Ittelson W H, Proshansky H M, Rivlin L G, Winkel G H 1974 *An Introduction to Environmental Psychology*. Holt, Rinehart and Winston, New York

Jameson C 1970 The human specification in architecture: A manifesto for a new research approach. *The Architects' Journal* **154**: 919–54

Kaplan S, Kaplan R 1981 *Cognition and Environment*. Praeger, New York

Lang J, Burnette C, Moleski W, Vachon D (eds.) 1974 *Designing for Human Behavior: Architecture and the Behavioral Sciences*. Dowden, Hutchinson & Ross, Stroudsburg, PA

Margulis S T (ed.) 1977 Privacy as a behavioral phenomenon. *Journal of Social Issues* **33** (3) (Special issue)

Manning P (ed.) 1965 *Office Design: A Study of Environment*. University of Liverpool Press, Liverpool, UK

Mikellides B 1990 Color and physiological arousal. *Journal of Architectural and Planning Research* **7**(1): 13–20

Passini R 1984 *Wayfinding in Architecture*. Van Nostrand Reinhold, New York

Seamon D 1982 The phenomenological contribution to environmental psychology. *Journal of Environmental Psychology* **2**: 119–40

Sommer R 1969 *Personal Space: The Behavioral Basis of Design*. Prentice Hall, Englewood Cliffs NJ

Wells B W P 1965 Towards a definition of environmental studies: A psychologist's contribution. *The Architect's Journal* **142**: 677–83

D. Philip

Architecture

Architecture is a complex area of social life to which the social sciences have paid relatively scant attention. The term refers to an art with a long history and a theoretical tradition and also to its products, which are one constituent of the built environment. This article examines how the architects' claim to be the foremost producers of architecture has been grounded historically in their relations with power, and considers the main shifts in Western societies' conceptions of architecture. It concludes with the implications of architecture for the social sciences.

1. Architecture and Building

For the philosopher Suzanne Langer, architecture 'is a total environment made visible.' Architecture creates its own specific illusion, not the ingredients and fragments of a culture, but a total image of it. Langer calls it the image of 'an ethnic domain,' implicitly stressing its sacred, collective origins and its public nature. She follows in this the dictum of Le Corbusier, one of the twentieth century's greatest architects, that 'Architecture is the masterly, correct and magnificent play of masses brought together in light.' But for a philosopher like Roger Scruton, architecture is an art of everyday life, the application of a sense of what is culturally appropriate. Any person with a sense of 'visual decorum' can pursue architecture, and the visual order it creates. If these acceptations may seem incompatible, it is because 'architects' architecture,' solely recognized as art by patrons and intellectuals, stands in opposition to 'mere' building and the untutored search for what looks right.

Architecture *is* building. It serves as shelter and as the stage of social life, but it rises above utility and transforms it. In ancient civilizations, the *telos* of architecture, its function and its traditional forms,

were sacred and transcendent, part of the ritual and theological knowledge monopolized by the priesthood. In Greece and Rome architects were secular, and their access to elite positions depended on mastery of both *telos* and *techne*, the organizing form and the technology of construction. Historians of architecture have noted that Roman buildings endow with meaning and dignity forms such as the barrel vault and the arch, and materials such as concrete that had only had utilitarian purposes. Roman civilization expands and diversifies the register of architectural types, while its domestic architecture marks the emergence of the architect as interpreter of the patron's private needs. But not until the Italian Renaissance do we find a new social definition of design and construction, harbinger of the modern architect's role.

The kind of building that we call architecture is defined in its essence by the relationship of *telos* and *techne*, conception and execution, symbolic intentions and materialization, but also by the fact of patronage. Before the Renaissance, special groups of builders or exceptional individuals had appeared as mediators of the elites' desire to express their piety or their hope for immortality in beautiful buildings. In Greek cities, the pursuit of beauty and excellence in building had already given rise to theorizing addressed by architects to other builders, artists, and intellectuals. We may call these mediators 'architects,' insofar as they inserted their practice between the *telos* and the *techne* of construction, but they *had* to be in control of technology. The complexity and the large scale of a project obtained for them a better chance to claim a measure of authorship from their patrons and enter their name in the historical record.

By 1400, the historical repertory already contained three potential strategies of collective social upgrading for the 'architects': one pre-eminently based on the development and mastery of technology, of which large-scale projects were the base; another dependent on the service of the state and official religion; and finally, an 'intellectual' strategy aimed directly at the symbolic and aesthetic dimensions of buildings. These three ways overlap, each historical situation offering or closing some alternatives. In the Italian Renaissance architects strove to appropriate the *telos* of architecture by intellectual and almost *purely stylistic* means; architecture would henceforth be distinguished from mere building by scale and a new set of stylistic conventions. In this crucial phase, Western architecture severed its intimate relationship with the technology of construction; *design* was instituted as the new discipline's foundation, while monumental projects remained the architects' *voie royale* for making their mark upon the city. The relationship of architects' architecture to the city's anonymous fabric was marked by latent conflict, both visual and social. It was transformed by change in the nature of cities, and no change was more revolutionary than urbanization in the industrial age.

2. *The Social Construction of Architecture in Western History*

The development of Gothic buildings provided remarkable constructional means to the fifteenth century, and a labor force with a tradition of technical competence that has not been equaled since. Renaissance architects could therefore concentrate on designing buildings, only needing to know what was technically possible in order to command the construction crews.

2.1 *Architecture and Power*

In the emblematic case of Florence after the Black Death, a revived economy gave the merchant elites capital, and the desire to spend it in celebration of themselves and their city. They looked for design talent in the ranks of decorative artists much rather than among the building trades (of the major architects only Sangallo the Younger and, later, Palladio came from the trades). Their keen interest in building sustained the designers' social ascension, as also did every city's need for civil engineering projects and complex fortifications. Architecture emerged thus in early modern Europe as a special medium for the needs of power and the expression of status.

Regular collaboration with their patrons gave architects, among other artists, a social position inaccessible to mere craftsmen. The distance they established between themselves and their technical base was never to be closed in the subsequent evolution of their role, but patronage continued to fetter their autonomy. Individual patrons and building committees not only supervised their considerable investments, but they involved themselves frequently in design, claiming authorship of the building (and they still do today!). At this social and ideological turning point, architects used the possibilities offered by the new forms of patronage to separate form from function in the *telos* of building, claiming conception, which translates function into form, for their own. The organization of construction proper was therefore left open to *other* delegates of the patrons, or other mediators. Architects turned to humanist intellectuals to explain the new style and its significance, while natural scientists began to study the nature of materials, and physical builders concentrated on the machinery of building. The rivalry with engineers, characteristic of the industrial age, was thus prefigured at birth for the specialized occupation of architecture.

Renaissance architects used the theory produced on behalf of their art to advance and legitimize their position. By the late sixteenth century, with the first academies, they began once again to write about practice for other architects. Theoretical foundations and treatises meant that architecture had to be studied, and that the title of architect could thus be denied to

the uneducated. In seventeenth century France, where the monarchy had been developing the administration of buildings for three centuries, the occupation of architect-urbanist entered an academic and official phase. As Louis XIV's phenomenal building programs were emulated by other monarchs, so was the Royal Academy of Architecture established in 1671. It integrated the Renaissance conception of the architect-artist with an *ancien regime* version of the civil servant's role, introducing a premodern notion of corporate professional power. The elite academicians were the expert judges of the beautiful, and monumental royal projects their preserve. The engineer also had made his appearance and asserted domain upon utility, his presence increasingly blocking any attempt by architects to recapture the control of *techne*.

Academic teaching diffused the conception of architects as specialists in stylistic codes, carrying it into the industrial age, where challenges against the architects' role in the system of construction multiplied. Architects defended themselves in a state of stylistic disunity, brought about by the exhaustion of the classic orders. Where an academy existed, officially charged with defending taste and elaborating doctrine as in France, they still enjoyed considerable advantages over other building designers. The French Academy, dissolved in 1793, reappeared in 1819 as the celebrated Ecole des Beaux-Arts. The classic coherence that it managed to preserve made the Ecole into a professional model, closely emulated in the United States practically until World War II.

Thus, at the beginning of the industrial age, architects retain in principle the control of aesthetics through the discourse that defines good 'architecture.' Among their patrons, state and church commissions remain important, although new monumental types have appeared often at the behest of business: universities, post offices, railroad stations, hospitals, but also stores, hotels, and the American 'tall office buildings.' Architects' architecture, propagated on paper by its own institutions, depends for realization on finding and pleasing patrons or clients. It works, as it always has, in the service of power and social status, which is now pursued by culturally divided and heterogeneous elites. As a learned discipline, architecture faces the challenges of industrial capitalism and professionalization in its own ranks with a traditional identity ill-suited to its very modern ambitions.

2.2 Architecture and the City

By the seventeenth century, the military engineer had become a separate role, but architects retained control over the monumental design of cities and over the complex, dynamic form of individual buildings. Neither the relentless axiality of monumental plans, nor the palatial style of life and conspicuous consumption of the age of the baroque could defeat, however, the cities' unruly growth. The anonymous urban fabric had always been a reminder that architects cannot monopolize architecture for, in a sense, the more beautiful and coherent a city is, the better it sets out the architects' designs, and the more it denies the uniqueness of their function.

Since the sixteenth century, capital cities held in a straight jacket by their walls had known overcrowding and rising land values, but urban disorder could still be contained physically. In the seventeenth and eighteenth centuries, the regimented holistic environments built for the state contrasted with mansions built for new classes of patrons—a physical counterpart of the stratification produced among architects by private and public patronage. But architects had nothing to say, except in planned utopias, about the need for housing of working populations. Before industrial capitalism, three major ways of addressing the city's problems had emerged: the gridiron plan of real estate speculators; improved lines of circulation and transportation (the basic approach of engineers, which Lewis Mumford called the 'bull-dozing habit of mind' and Baron Haussmann applied unforgivingly in Paris); and utopian schemes, proposed mainly by architects. The latter juggled uneasily with total monumental conceptions, with the integration of nature into the city, or with escape and reconstruction.

The industrial revolution's utilitarian buildings put old and new materials—iron, glass, concrete—to bold new uses, while the unplanned growth of the modern metropolis seemed to confirm architects' architecture as a superfluity of the rich and powerful. Yet new needs and new technologies meant that it could become directly relevant to larger sectors of society than ever before in history. Architects had to redefine their role in the politics of construction to confront the task. As in the crucial phase of the Renaissance, they used doctrine, which they controlled, to secure their social foothold.

3. Doctrinal Shifts in Architecture

A specialized occupation, which produces discourse about its activity and objectives, will use it to legitimize, confirm, and also transform its conditions of practice. The audience for whom discourse is intended will inflect and influence its content and manner.

Thus, in the Renaissance, intellectuals educated potential patrons in the *stil novo*, providing explanations, and exhortations to spend money. From the treatises, a more autonomous discourse also emerged. Leon-Battista Alberti integrated the rediscovery of the classic Orders with a theory of harmonic proportions, pronouncing architecture the spatial materialization of mathematical truth. The rationalism of an emancipated minority of intellectuals echoed the buoyant sense of power of the new patron class, contributing to the divinization of art and conferring charisma upon

architecture. The ideology and the new architects' masterpieces made architects into artists who henceforth competed with their patrons, not in the political economy of construction, but in a symbolic dialectic of charisma. Intellectual work and publics allowed architects to appropriate the pure *telos* of architecture, but the environment was changed only by the decisions to build which belonged to the patrons.

Five centuries later, the modernist *avant gardes* also launched a powerful symbolic and ideological movement. Renaissance architects had advanced their own collective social status and that of their discipline relying on their own shared origins and locality, as well as on the worldview of the elite whom they served. Modernist architects worked in different nations and in different circumstances, but they too led a transnational movement with a coherent doctrine. These minorities within architecture shared a long-standing discontent with what architecture had become; their opposition to academicism, which ignored the problems and the possibilities of the modern era, unified them. Four key factors describe modernism's conditions of birth: the existence of artistic *avant gardes* in the European capitals; the devastating experience of World War I and the massive need for housing it exacerbated; the response to socialism and the revolutionary movements of the brief interwar period; and the demonstration of enormous productivity provided by large-scale industry during the war effort. At all the levels of their aesthetic and ideological battle, the modernists sought an integration of architecture within the mainstream of industrial production. As Reyner Banham has shown, the aesthetics were derived from an *image* of machine-made objects, with the ideological premise that art would conquer mass production by fusing with it.

In the first part of the twentieth century, architectural modernism represented as total a departure from the immediate past as the Renaissance was from the Gothic. In both cases, ideology and aesthetic theory were the guiding principles, just as in both cases the movement leaders attempted a social redefinition of the architect's role. Modernism, however, depended closely on new technologies and new materials for the realization of its radically antihistorical aesthetics. Claiming to speak for the largest number and to the largest public, modernists presented their ideas as a world architecture for the masses. Their doctrine called for the abandonment of regionalism, nationality, traditionalism, and local particularities, for the ahistorical domain of function, never a guiding principle of design but an image wedded to that of the machine. Abolishing the worn-out *signifiers* of the past, and leaving only pure, efficient Form, the modernists attempted to bridge the opposition between buildings where people work and buildings where they live. Freed from load-bearing walls by the new technology, they could design continuous spaces and 'dematerialized' glass walls that went beyond the antinomy between inside and outside inscribed in heavy ornamented facades.

Modernism did not redefine the position of architects in the political economy of construction. Driven to America by Nazism and war, the German modernists, in particular, conquered for a time an almost hegemonic academic base. After World War II, the steel and glass aesthetic identified with Ludwig Mies van der Rohe fused with the large architectural office (an American late nineteenth century invention) to furnish the corporate reconstruction of the world with its towering glass boxes. The next doctrinal shift, starting in the 1960s, attacked once again architects' architecture for having become a mere instrument of profit and power.

The revisionist attacks gathered under the label of *postmodernism* had different origins and a different thrust in America and in Europe. European architects, dependent on public funds for their most important commissions, never abandoned the idea that architects have an important social role to play (Champy 1998). Yet on both sides of the Atlantic, the attack started within the specialized *discourse* of architecture, in which architects reserve for themselves the authority to participate. The passage from the specialized discourse (situated mainly in universities, museums, journals, and intellectual circles, and interesting primarily for architects and cognoscenti) to the streets required as always that new, defiant buildings be realized. In the United States, postmodernism was marked by a manner rather than a coherent conception of architectural design; the manner admitted ornament and thrived on eclectic allusions, ranging from reinvented history to regional vocabularies and populist gestures toward a mostly commercial 'vernacular.' The movement attacked the modernist aesthetic concretely embodied in its archetypal buildings. The modernism of skyscrapers and 'growth machines' was indicted in the name of another architecture, which thrived on small projects, relatively modest housing, new kinds of clients, and new kinds of needs, both frequently subsidized by the War on Poverty. The battle for the control of architecture's discourse marked the arrival of a new generation, but also the affirmation of a different kind of practice (Larson 1993).

During the 1970s, however, the professional elites, buffeted with a crisis of construction that reached depression levels, gradually conferred legitimacy upon almost any definition of what constituted good architecture. In the ensuing confusion, a brand of revisionism that rejected both the social mission of architecture and the historicist vocabulary asserted once again the supremacy of design and form as the primary competence and concern of architects. The ideological restoration of the architect-as-artist role opened the 1980s, a decade of ferocious real estate speculation and chronic overbuilding, which coincided with the Reagan era and incorporated all the forms of

postmodern revisionism into the establishment's architecture. The era saw the preponderance of architecture as provider of eclectic *images*, prime assets in the ever-faster cycles of upscale consumption, and the never-ending search for product differentiation. Architecture became more glamorous than ever at the service of postindustrial and global capitalism, but it was neither more secure nor autonomous vis-à-vis clients and competitors than before. At the end of the twentieth century, some leading architects and critics saw as deeply problematic the relations between scenography and construction, between image and reality in architecture, agonizing once more about the role of architecture and architects in the now global economy.

4. Concluding Remarks

Architecture has interested the social sciences as a profession, whose weakness is in part explained by its base in aesthetics. It has not succeeded in establishing jurisdiction against either professional competitors or lay resistance, and it is marked by apparently unsurmountable lines of internal stratification, based on the form and the volume of practice. Typical of architecture is the deep cleavage that distinguishes the elite designers who control discourse from everyone else. Logically, architecture has also been studied as a form of production of culture closely interdependent with the economy and inserted into a very complex division of labor (Blau 1984, Gutman 1988, Larson 1993, Moulin 1973).

The architects' expertise is challenged by cultural plurality, permissible and encouraged in the arts as it is not in the sciences, or even in the law. In the global economy, both practitioners and theorists, seeing architecture subjected to the realities of 'transnational' construction and real estate promotion, have become aware of endangered cultures, always a concern of lay people (Saunders 1996). The behavioral sciences may have been closer than sociological research to the interests of architects, although the former have tended to privilege individual reactions to very general characteristics of the built environment. Recently, social scientists have approached the problem of *reception* of architecture (Larson 1997) and included it in the development of a sociology of place (Gieryn 2000). The emphasis is on the scripts, conventional and otherwise, that may be inscribed in the architectural object, and read, or followed, by the users. The importance given to meaning and agency seeks to bridge the gap between architecture and building, and between lay people and their culture, returning thus to the point where this account began, and to the contradictions inscribed in architects' architecture.

See also: Art History; Human–Environment Relationships

Bibliography

Banham R 1960 *Theory and Design in the First Machine Age*. MIT Press, Cambridge, MA

Blau J 1984 *Architects and Firms*. MIT Press, Cambridge, MA

Champy F 1998 *Les Architectes et la Commande Publique*. Presses universitaires de France, Paris

Gieryn T 2000 A space for place in sociology. *Annual Review of Sociology*. **26**: 453–96

Gutman R 1988 *Architectural Practice*. Princeton Architectural Press, New York

Kostof S (ed.) 1977 *The Architect*. Oxford University Press, New York

Langer S K 1953 *Feeling and Form*. Scribner, New York

Larson M S 1993 *Behind the Postmodern Façade*. University of California Press, Berkeley, CA

Larson M S 1997 Reading architecture in the Holocaust Memorial Museum. In: Long E (ed.) *Sociology and Cultural Studies*. Blackwell, London

Lane B M 1968 *Architecture and Politics in Germany*. Harvard University Press, Cambridge, MA

Le Corbusier 1970 *Toward a New Architecture*. Praeger, New York

Moulin R 1973 *Les Architectes*. Calmann-Levy, Paris

Saunders W (ed.) 1996 *Reflections on Architectural Practice in the Nineties*. Princeton Architectural Press, New York

Scruton R 1979 *The Aesthetics of Architecture*. Princeton University Press, Princeton, NJ

Stieber N 1998 *Housing Design and Society in Amsterdam: Reconfiguring Urban Order and Identity, 1900–1920*. University of Chicago Press, Chicago

M. S. Larson

Archival Methods

'[N]o documents, no history' declare Charles-Victor Langlois and Charles Seignobos (1898, p. 2) at the end of the nineteenth century as they begin their guide to historical research. They continue: 'To hunt for and to gather the documents is therefore a part, logically the first part, and one of the most important parts, of the historian's craft' (1898, p. 2). They thus state a methodological principle as well as a defining marker for a social science discipline. A standard for progress: if more recent historians are in any way superior to those of the past it is because they have 'the means to be better informed' (1898, p. 3); these means are the modern public archive, with its large and effectively catalogued collection, and the careful training of fledgling historians in the critical use of documents. A century later, historians were still far more likely than other social scientists to exchange tales of archival discomforts and delights, and to guide their graduate students in crafting essays built around properly referenced primary sources. But new information technologies were somewhat blurring the distinction between archival and published sources, between primary and secondary sources, and between the

methods of history and the methods of other social sciences.

1. Archives as Research Sites

Archives are the records of an organization's activities. The term is used both for the place those records are housed and for the body of records themselves. These records may have been produced by that organization or gathered by it. Such organizations include businesses, religious bodies, and government agencies, and we may stretch the notion to include the records of individual persons. Since the records were produced for the purposes of that person or organization, they tend to be organized for those purposes and not those of researchers. This means that archives are often difficult and sometimes impossible to use effectively without detailed knowledge of the organization in question. Researchers could hardly make much of the documents in some governmental archive without understanding the agencies that produced or collected the documents.

Archives tend, therefore, to be organized idiosyncratically. One archive may well differ from another not only in its classificatory categories, but also in the extensiveness of inventories, catalogues, and other aids to research. There may be no list of individual documents, but just indications of more or less broad categories. The records of some country's regional or local governments may not be organized in a uniform manner. As a result of such challenges to a user's ingenuity and patience, a researcher may derive considerable pleasure from the sense of mastering an archive. By contrast a researcher can get very far in most libraries the world over by expecting to find an alphabetical catalogue of authors and titles.

2. The Modern Public Archive

The linking of the modern practice of historical research and the modern public archive is generally dated from the French Revolution, but archives of some form are considerably older. Since states have often needed records of laws, tax assessments, and interstate treaties, state archives existed in antiquity. *Arkheia* is the Greek for an authority's records. It is even probable that the invention of writing was spurred by state need for documentation. So archives were part of the technology of state power, but they could also play an important role in the power claims of other social actors. In the European middle ages, lords and ecclesiastical institutions maintained records spelling out their privileges (etymologically: private laws). The possession of documents attesting to immunities from royal claims or the antiquity of claims over others (for seigneurial dues, for example) was an important instrumentality of such privilege.

Early modern European states had their document collections, too, but these were often semiprivate, with individual agencies and even individual officials having their own archives. Steps towards centrally organized collections were part of the history of European statemaking; early centralizers included the Spanish archives of Charles V in 1545, the English State Paper Office in 1578, and the Archivium Secretum Vaticanum in 1611 (Favier 1958, p. 24). But these were not publicly accessible, secrecy being as significant an attribute of the monarch's documentary collection as of the lord's.

The French Revolution launched the modern public archive. In France's villages, peasant attacks on the documentary collections of lords and monasteries testified to their continuing significance. Although the new revolutionary authorities promoted some archival destruction, the overwhelming thrust of their efforts ran in a very different direction. The separate collections of royal agencies, old regime corporate institutions, and local powerholders were made public in two senses:

(a) They were brought under state management and thereby opened up to systematic classification, not only in the National Archives housed in Paris, but in the uniformly organized departmental archives as well. This development greatly facilitated measures for preservation, as well as cataloguing. A modern profession of archivists formed around these tasks.

(b) They were made accessible to the public. With exceptions for state secrets (a far more limited notion than before the Revolution) and the privacy of individual citizens whose identities might figure in state records, documents were to be made available for the scrutiny, and research, of French citizens. Archival documents were redefined from being primarily instruments of power, often private power, to aspects of the national cultural heritage.

In establishing France as the center of a reconstructed Europe, Napoleon had vast archival holdings transferred from Spain, Vienna, and much of Italy to Paris. In the wake of the French defeat, the recovery of documents from France further strengthened the notion that a respectable modern state needed its own respectable modern archives; as the nineteenth century went on this increasingly meant that, like France's, other archives were going to be open to a far broader public than in the past.

3. Professional History

The historical profession that emerged in the nineteenth century combined a distinctive sense of evidence, a distinctive form of professional apprenticeship, and a distinctively organized intellectual product. This was being developed in many places but the model for the new professional history was provided

in the writings and seminars of German historian Leopold von Ranke, beginning in the late 1820s (Novick 1988, Smith 1998).

(a) Historical claims were supposed to be grounded in the best possible primary sources, generally those located in searching relevant archives. Historical knowledge would be disciplined, a notion that acquired important additional baggage as the German *wissenschaft* was transmuted into the English *science* as a description of the kind of learning produced by historical research.

(b) Apprenticeship included practice in making something of primary sources in graduate seminars. To the extent the seminar was seen as a sort of 'laboratory' (Novick 1988, p. 33), the sense of history as a sort of science was augmented. Mastery would be further demonstrated by the completion of an original work dependent for its primary sources on unpublished, generally archival, materials. The budding historian would acquire not only the cognitive skills but also the character needed for archival labors. Historians needed to be 'calm, reserved, circumspect; in the midst of the torrent of contemporary life which swirls about him, he is never in a rush.' Those 'always in a hurry to get to the end of something…may manage to find honorable employment in other careers' (Langlois and Seignobos 1898, p. 103).

(c) Professionally respectable intellectual products consisted of books and articles whose factual claims were, in principle, verifiable from the primary documents to which the text would explicitly refer. The footnote to sources was a vital part of the prose of professional history (Grafton 1997). Denoting the structure of reference—footnotes plus bibliography—as the scholarly 'apparatus' fostered rhetorical associations with science.

Important and diverse consequences followed from its characteristic methods for such a professionalized history:

(a) *National specialization*. Historical specialties tended to be organized thematically around national histories because the archives were organized by states. Beyond linguistic familiarity, practitioners who had invested their energies in mastering one archive or set of archives found it advantageous to continue future work that would reap the benefit of those investments.

(b) *Building national identities*. Partly as a consequence of this technology of history, and partly because states were supportive of such work, the writing of history itself became an important component of the very forging of national identities in the nineteenth and twentieth centuries.

(c) *Focus on elites*. Since what was in these state archives was, by definition, documents that had been of interest to state managers, professionalized history had a tendency to privilege the doings of states and other human practices of interest to administrators while paying lesser attention to other ways of exploring human experience.

(d) *Focus on Western history*. Since the standards of professional history could best be attained where there were rich archival collections, cared for by skilled practitioners of preservation and cataloguing, those parts of the world relatively well endowed in those regards were attractive to researchers who hoped to achieve professional advance. Such conditions meant governments: (i) with rich documentary collections; (ii) with the resources to support the preservation and cataloguing of these collections; and (iii) for which the French revolutionary model of a public national archive as an essential component of a modern national identity had some resonance. Nowhere were these conditions better met than in France itself. Part of what has made France such an important center of innovation in professional history in the twentieth century has been its marvelously organized archives. Countries (i) lacking extensive documentation, (ii) having documents but lacking the resources or the interest in preserving them or organizing them coherently, or (iii) with strong traditions of archival secrecy were all, in their different ways, more difficult terrain for professional historians.

(e) *Marginalization*.The inverse of the previous points: social activities, social strata, and entire continents that have either been of little interest to states' producers of documents, or that have not developed such extensive and usable documentary collections, have tended to be marginal areas for professional history. Historians interested in such subject matters have had to be methodologically innovative and have sometimes faced an enormous challenge for professional recognition.

(f) *Eschewing explicit methodological discussion*. A professional socialization oriented to the careful scrutiny of the provenance of particular documents and the opportunities for knowledge and the motivations for falsification on the part of that document's author encouraged historians to delight in the particular. This was reinforced by the particularities of archives, knowledge of whose quirks was not readily transferable to other archives. Historians understood their own methods in large part as the location of useful texts in idiosyncratically organized archives in conjunction with the careful criticism of documentary sources. Such skills were acquired by experience and the patient application of a few readily assimilable critical rules of thumb. The capstone of training involved immersion in some idiosyncratic archive. Thus Langlois and Seignobos: '[t]o learn to distinguish in this enormous confused literature of printed inventories (to confine ourselves to such), that which merits confidence from that which does not, in a word to be able to make use of them, is a complete apprenticeship' (1898, p. 12). Historians have therefore directed far less of their intellectual efforts at methodological discussion and have therefore also had fewer methods courses as part of graduate training than the other social sciences. In 1971, one study of

graduate education commented that 'methodology is the orphan of the history curriculum' (Landes and Tilly 1971, p. 82); there is no reason to think that three decades later things had significantly changed.

(g) *Eschewing explicit theory construction*. Dependence on the fortunes of document survival, as Murray Murphey observes (1973, p. 148), is a major reason historians have been less prone than other social scientists to pursue a model of inquiry in which abstract hypotheses are formulated, then tested against available evidence. The unpredictabilities of archival research suggest instead the development of hypotheses in tandem with the exploration of data. One often finds wholly unexpected documents sitting on a shelf somewhere. Theory building as an explicit and valued agenda is therefore less characteristic of people in departments of history than it is of those in sociology or political science and very much less than in economics. Nor does a broad knowledge of history's big themes without a deep reserve of pertinent facts earn much esteem from professional peers. By way of contrast, graduate students in top US economics programs do not regard a deep knowledge of the economy as nearly as important to their professional futures as they do skill in mathematics (Klamer and Colander 1990). One could not imagine graduate students in first-rate departments of history similarly downplaying the significance of a rich and concrete knowledge of some specific time and place in favor of mastering methodological or theoretical tools.

4. *Working with Documents*

The notion of 'archival methods' comprehends two rather different sorts of activity. First, there are the methods of collection, preservation, and cataloguing on which documentary research in archives depends (Schellenberg 1956, Brooks 1969). Second, there is the work of historians (or others) with those documents. Essential to nineteenth century notions of historical method was the 'criticism of sources.' Classic works of method, like that of Langlois and Seignobos (1898) provided a broad taxonomy of ways in which one document might be superior to another. For example, historians became skilled at considering:

(a) possible errors of transcription in the reproduction of or quotation from ancient texts

(b) whether a document's author was in a position to reliably know that which was claimed

(c) what motivations that author might have for slanting a story this way and that

(d) how to select from among multiple documents the most credible account.

According to Langlois and Seignobos (1898, p. 131), historians need to cultivate an attitude they call 'methodical distrust': an author's 'every claim' must be suspected of being 'mendacious or mistaken' (p. 132). (There follows a taxonomy of possible guises assumed by documentary deceit and error as a guide for avoiding the snares and pitfalls that might lead a researcher too easily to accept some statement as historical truth.)

More recent reflection has pondered the forms of distortion that might exist in whole collections of documents, and recognizes with Murray Murphey (1973, p. 146) that the survival of documents has often depended on fires, floods and 'the concern of loving daughters.' If there is still much to be said for the formulation of Langlois and Seignobos that there is no history without documents, we may consider how it is that documents come to be available to the historian's scrutiny. We should consider (Shapiro et al. 1987):

(a) *Recording*: what social processes bring documents into existence? This includes taking into consideration what sorts of things are or are not of interest to states, economic enterprises, religious bodies, and other recorders of words. An increase in counted murders may indicate a greater interest of states in whether or not people are killing each other, for example. No doubt in many times and places social struggles in and near centers of government generate far more paper than those far away.

(b) *Preservation*: what social processes destroy or preserve documents? We must consider not only the destructiveness of war and fire which often are random destroyers of texts and the wishes of loving (or hostile) relatives for preservation (or elimination) but also the existence of organized agencies for preserving or eliminating the records of the past. The existence of well-trained and dedicated archival professionals is more probable in some countries than others, more (we may speculate) in large urban centers, more in relation to some subject matters (but which?) than others. (The study of the professional cultures of archivists would seem an important agenda for a self-critical history.)

When economists, sociologists, or political scientists turn to archival sources to construct statistical series about social processes that extend over time and space they are likely to encounter several problems. Some of these problems may be amenable to amelioration by statistical manipulation but others demand the kinds of analysis of sources that are characteristic of historians—and still others call for both in conjunction.

(a) *Missing data*. Random document loss from moisture, fire, and the gnawing criticism of the mice; deliberate destruction of troublesome records by those who might be embarrassed by some datum or data; and errors in filing by clerks all generate considerable missing cases in many potentially valuable data series. Statistical judgments about the appropriateness of completing the series by extrapolation and interpolation from surviving data may be complemented by the institutional knowledge that sometimes permits a researcher to find an alternative source for the same data.

(b) *Changing definition*. Much data of interest to states is subject to changing definition and therefore many series of great interest to social scientists cannot be used intelligently without a study of those definitions. Crime, ethnicity, conflict, and poverty, for example, are all subject to considerable redefinition, both formally and informally. To make use of any long data series on crime, one would have to have acquired considerable knowledge of the changing interests of states in accurate data collection for different kinds of offenses. One would also have to know something of the changing mores that redefine actions as criminal or not (and which kind of crime they are). In addition, the boundaries of such administrative subdivisions as counties and municipalities often change, and a study of changing administrative geography is often an essential step on the way to some form of statistical correction.

5. *Challenges to Professional Traditions*

The last quarter or so of the twentieth century saw challenges to such notions of research methods from several converging sources. Since what was easy to examine with prevailing methodologies were those facets of human experience that had left traces in the documentary collection of states, the subject matters of professional history gave great weight to adults, men of weapon-bearing and income-earning age, dominant ethnicities, state politics, and dominant powers on the world scene. Those curious about other realms of human experience, including members of some of the under-represented groups and places, were led to develop new methodologies.

To some extent these involved innovative use of state archives, as in the mobilization of records of births and deaths to develop the new historical demography, a vital window on the past lives of ordinary people. To some extent these involved the use of new kinds of sources, as in the discovery of the value of visual materials for clues to the history of childhood or aging. To some extent these involved extensive interaction with other social science disciplines, as in important interchanges with anthropology in the hope for getting a handle on people whose activities have left less easily pursued traces in the written records of states or in the importation of quantitative methods from economics, sociology, and political science (Rabb and Rotberg 1982, Revel and Hunt 1995).

Fueled by these trends, and fueling them, professional historians became inclined to doubt the clarity and coherence of the entire project of a scientific history. This further eroded the rationale for a methodologically distinctive enterprise, thus opening the way for even further cross-disciplinary fertilization. But it also opened the way for a radical skepticism about any form of historical knowledge or inquiry whatsoever. In the last quarter of the twentieth century, increased sensitivity to the ways in which the subject matter and methods of past historical research had been intertwined with social power was leading some to abandon the notion of any coherent historical truth altogether (Appleby et al. 1994).

6. *Challenge and Opportunity of Electronic Information Technology*

The new information technology becoming increasingly widely available at the end of the twentieth century seemed likely to reshape the ways in which archives would be used and perhaps to reshape what an archive was.

(a) *Cataloguing* could now be enhanced radically. Particular items could be cross-classified in innumerable ways since electronic data files are not subject to the same sorts of space limits as a catalogue's page or a card file, making it easier for researchers to find all that is relevant. User-manipulated databases meant that users could sort through files according to the categories important to them, relieving them, to at least some limited extent, of the challenging burden of mastering the intricacies of the classificatory systems of particular archives. Since electronic information could be transmitted around the world at high speed through the Internet, researchers were becoming less dependent on visits to distant archives to have some notion of their contents.

(b) *Reproduction* of documents was immeasurably enhanced. Age-old errors of transcription, inherent in hand copying of archival materials, were alleviated by the development of microfilming technologies, and further reduced by photocopying. Both microfilming and photocopying technologies enhanced the labors of researchers from afar. But they also meant that archives and research libraries could duplicate their own scarce holdings and distribute them either to individual scholars or other research institutions. The potential of high-quality electronic scanners as input devices to computers was only at the very beginning of use in archival work at the beginning of the twenty-first century, but no one could doubt that the potential impact on research activities was going to be enormous. Not only could a visiting scholar store large collections of source materials, but such materials could also be readily transmitted to far-off scholars. If, as seems likely, there come to be improvements in the quality of software for optical character recognition, an electronic archive including scanned documents opens up the possibility of high-speed searches according to multiple criteria, further freeing researchers from the idiosyncrasies of particular catalogues.

(c) *Publishing*. New modes of electronic publishing, barely launched at the beginning of the twenty-first century, looked likely to be similarly revolutionary, and with similarly important implications for

redefinition of professional scholarly standards (Darnton 1999). A conventionally published book, for example, might omit its apparatus of learned footnotes and bibliography, but make them available to other interested scholars on a disk or an Internet site. One might speculate that this would lead to new forms of scholarly precision and richness (as scholars supplemented a basic text with footnoted references to sources, critiques of learned predecessors, methodological commentaries on sources or predecessors, annotated rather than bare-bones bibliographies, and even scanned copies of vital original documents). At the beginning of the twenty-first century, for example, it was technologically feasible for the learned footnote at the bottom of the printed page to be replaced by the electronic note in some online text in the form of a hyperlink to a scanned copy of the referenced text itself. Such super-references would actually send the primary sources to the reader's screen or printer, rather than tell the reader in which place in which archive the source could be consulted (if that reader had the time and the good fortune to win a fellowship and to have a dean kind enough to grant leave). The capacity of historians to check each other's research against the primary sources would be greatly facilitated. Graduate students with no research funds to travel to document collections could nonetheless hone their critical faculties by studying the sources of prominent works without mastering archival idiosyncrasies.

7. *The Future of Archival Methods*

The new technologies, moreover, were opening up what could be called *the electronic archive* (Shapiro and Markoff 1998). Databases, possibly accompanied by programs to facilitate use, once created, could be readily distributed, and even augmented by new scholars.

Possible consequences for the future culture of historical research might include:

(a) Blurring the distinction between research on primary and secondary materials as researchers grew comfortable with electronic archives built of many components that might range from scanned primary sources, quantitative data extracted from such sources, quantified data created from such sources by a researcher, and new variables or indexes added by a community of users.

(b) Blurring the distinction between archival and published primary sources as information became available in electronic form, regardless of whether it was extracted from a book or found in an archival dossier.

(c) Increasing attention to issues of representativeness of large bodies of data (and diminished concern for the traditional criticism of sources?).

(d) More research using primary sources by social scientists with the appropriate technical skill who were not employed by departments of history.

(e) Increasing use of archival materials by scholars, whether professional historians or not, who are not primarily specialists in the national history of the country from which the data are derived as it becomes easy for someone on the other side of the planet to use an electronic archive without having to master the particularities of traditional repositories of documents.

(f) Increasing accessibility of materials vital to historical problems that cross national frontiers, as teams of scholars assemble relevant electronic archives.

In such ways the boundaries between history and the other social sciences may become less sharp than they have been since the emergence of the social science disciplines.

See also: Archives and Historical Databases; Archiving: Ethical Aspects; Data Archives: International

Bibliography

Appleby J, Hunt L, Jacob M 1994 *Telling the Truth About History*. Norton, New York and London

Brooks P C 1969 *Research in Archives. The Use of Unpublished Primary Sources*. University of Chicago Press, Chicago and London

Darnton R 1999 The new age of the book. *New York Review of Books* **46**: 5–7

Favier J 1958 *Les Archives*. Presses Universitaires de France, Paris

Grafton A 1997 *The Footnote: A Curious History*. Harvard University Press, Cambridge, MA

Klamer A, Colander D 1990 *The Making of an Economist*. Westview Press, Boulder, CO

Landes D S, Tilly C (eds.) 1971 *History as Social Science*. Prentice-Hall, Englewood Cliffs, NJ

Langlois C V, Seignobos C 1898 *Introduction aux études historiques*. Hachette, Paris

Murphey M G 1973 *Our Knowledge of the Historical Past*. Bobbs-Merrill, Indianapolis, IN and New York

Novick P 1988 *That Noble Dream. The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK

Rabb T K, Rotberg R I (eds.) 1982 *The New History. The 1980s and Beyond*. Princeton University Press, Princeton, NJ

Revel J, Hunt L 1995 *Histories. French Constructions of the Past*. New Press, New York

Schellenberg T R 1956 *Modern Archives. Principles and Techniques*. University of Chicago Press, Chicago

Shapiro G, Markoff J 1998 *Revolutionary Demands. A Content Analysis of the Cahiers de Doléances of 1789*. Stanford University Press, Stanford, CA

Shapiro G, Markoff J, Baretta S 1987 The selective transmission of historical documents: the case of the parish cahiers of 1789. *Histoire et Mesure* **2**: 115–72

Smith B G 1998 *The Gender of History. Men, Women, and Historical Practice*. Harvard University Press, Cambridge, MA

J. Markoff

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Archives and Historical Databases

This article discusses the history of archives from antiquity to the present day. It makes clear that archives follow a legal tradition that reaches back to the beginnings of occidental civilization. The use of archives for research is a relatively recent development that basically began with the French archives law of 1794. A broad organization of archives, both public and private, has developed since that period and now encompasses all the countries in the world. The scope of this material (i.e., charters, records, registers, maps, plans, audiovisual media, electronic documents, private collections, and archives) and its appraisal pose serious problems for archivists today. Those problems include all questions dealing with preservation and information technologies.

1. *Terminology and History of Archives*

The term 'archive' comes from *arché*, the Greek word for beginning. In Herodotus' *History*, *Archeion* is the town hall or government office. The word was not used to refer to preservation of the written word in ancient Greece, but was used in that manner by Josephus and later by Eusebius during the Hellenistic period. Roman jurists later adapted the Greek *Archeion*, using it in the Latin *Archium* and later *Archivum*. By the end of the fourth century, *Archivum* refers to the place where public records are preserved (*locus quo acta publica asservantur*). The archival tradition, therefore, is not based on the concept of preservation for research, but it was preservation for public administration purposes. Elements of this idea that archives only preserve public documents can still be found in present-day British and Dutch archives. It follows that the English language makes a difference between records that possess legal character and papers that possess no such qualities. The official character of such archival material, as well as the fact that its preservation is a responsibility of the state, both explain the understanding of public faith and the concept of unbroken custody in the Anglo-Saxon world.

Research done by Johannes Papritz, Ernst Posner, and others indicates that clay tablets from 3000 BC found in the Near East were preserved to facilitate commercial and public administration. Strictly speaking, these tablets are registries, but were not intended for permanent preservation as an archive. Literature that uses the term 'clay tablet archives,' often referring to the royal palaces of Minoan civilization on Crete and the Palace of Nestor at Pylos (Peloponnese), is therefore misleading. While clay tablets are exceptionally durable, even fire-resistant, other material used to write on, such as leather, wooden tablets, or papyrus, proved to be less durable. Finds of such materials are therefore quite fragmentary. In Athens, laws and decrees of councils and assemblies, documents concerning acts of state, and copies of state-commissioned plays were stored at the town hall and, after the fourth century, at the Temple of Kybele. In Rome, senate resolutions were kept in the Temple of Saturn until 78 BC, and after that time senate minutes, state payment demands, census records, and so forth were placed in the 'Tabularium.'

Some remains of this ancient tradition can be found in the registers of the Middle Ages, that is, in the secure documentation of resolutions and records through continuous notation on archival tablets, which were later replaced by papyrus rolls. In 538 AD, emperor Justinian ordered all cities to preserve the *commentarii* and *gesta municipalia*, to facilitate jurisdiction, in a separate location, the *Archeion* or *Archivum*. Nevertheless, political upheavals in late antiquity and the beginning of the medieval period led to breaks in archival traditions. Roman registers can still be found in the cities of northern Italy and Gallia Cisalpina. Notary registers as compiled by Johannes Scriba of Genoa came into use in the middle of the twelfth century. The Papal Curia reintroduced register-keeping in the eleventh century (on parchment since 1088, though only records from 1198 survived due to a fire in the Lateran). The papal example may have had influence on the creation of a royal archives in Aix-la-Chapelle under Charlemagne. The lack of a permanent residence among his heirs prevented the development of an archival organization. The influence of Persian–Arabian legal practice under Norman rule in Sicily could still be felt in France and England in the twelfth century.

Unlike the archives of antiquity, those developed in the medieval period did not serve the public interest in the first place, but rather helped to secure legal positions of individual institutions and groups, i.e., church and monastic repositories, municipal archives, and archives of the nobility. The refinement of paper in the fourteenth century led to a reorganization of office and chancellery management. Registers, minutes, bills, and so forth became an administrative memory. Older archives were rearranged and concentrated in the fifteenth and sixteenth centuries: the Spanish court archives at Simancas (1548) is one notable example.

2. *Modern Archives since the French Revolution*

The French Revolution was an important turning point in the advancement of the archives. It was the revolution that introduced the developments we associate with archives today. The concentration of archival materials in central locations, the organization of an archives administration, the civil administration's jurisdiction over registries, public access,

and historical research are all characteristics of this new world of archives.

The National Archives in Paris, established in 1789–90, was initially intended only for the preservation of records produced by the national assembly. Records from *ancien régime* administration boards were not incorporated until 1793. The law passed on June 25, 1794 declared the National Archives to be a *Centre Commun*, giving citizens the right of access to the documents preserved. Two years later, the departmental archives were created and put under the jurisdiction of the National Archives. Much literature refers to 1794 as the 'constitution' of modern archives. Napoleon's 'Universal Archive' of 1810–11, which integrated important artifacts from the conquered German empire, Austria, Italy, Spain, and the Vatican, lasted only briefly. What made the National Archives revolutionary was the fact that the arrangement of the archives followed the librarians' principle of pertinence. In 1839, initial calls were made, however, to introduce in the departmental archives the principle of provenance, which was codified in a circular on April 24, 1841. This concept of *respect de fonds* stipulates that all records created by one institution remain together. The paleographer Natalis de Wailly said of it, '*La methode est fondée sur la nature des choses*.' The same principles were introduced on a municipal level in 1857.

In 1861, the principle of provenance was realized in Denmark's ministerial archives. In Germany, it was Prussia that prescribed the principle of provenance with the regulation of July 1, 1881: 'Every government board will, once it generates records, receive its own repository.' The regulation initially applied only to the Prussian Secret State Archives. Not until 1896 were the provincial archives advised to follow suit. Primary credit for the international success of the principle of provenance in the following decades must be given to the Dutch: in the Netherlands the principle of provenance was introduced in 1897. The archivists Samuel Muller, Johan Adriaan Feith, and Robert Fruin described the principle of provenance in their handbook of archives, achieving worldwide recognition for Dutch archival theory. Translations into German (1905), Italian (1908), French (1910), and English (1920) followed. The conclusion of this discussion, which dominated the late nineteenth and early twentieth centuries, came at an international conference in Brussels in 1910. Archivists claimed the principle of provenance as fundamental for their profession. The principle of provenance had been accepted by Denmark in 1903, and by Sweden and the USA in 1909. Vienna's Haus-, Hof- und Staatsarchiv accepted it in 1914. In 1920 Hilary Jenkinson achieved implementation in England, where the law had controlled archives since 1838. A decree by Lenin issued on June 1, 1918, dealing with the reorganizations of archives, gave Russian archives the chance to respect the principle of provenance whenever possible. An international conference in Stockholm in 1993 proved that the principle of provenance continues to play a role in international archival theory and discussion.

3. *Records as Historical Sources: Appraisal and Use*

Historical research and the development of critical methods, as exemplified by the German editions of the *Monumenta Germaniae Historica* since 1826, promoted archival description of older documents. Inventories have been published in England, France, and Belgium since the third decade of the nineteenth century. Editions of charters and records from Prussian archives have appeared since 1878.

While historians in Germany still discussed Karl Lamprecht and his cultural and sociohistorical theory from 1893 to 1898, archivists were already discussing topics that were very modern at the time, such as whether or not to preserve judicial records, which contained material on the administration of justice as well as the moral, financial, and economic circumstances of different classes of society. There were demands for the preservation of records relating to the labor movement and social questions, for census records and, from 1904 onward, there were tendencies to preserve archives from private firms. It is no coincidence that during the first decade of the twentieth century in Germany the first private business archives were established. These were quite different from Europe's preindustrial trade and merchant registries that survived from the late Middle Ages (e.g., the Datini archives in Prato).

The disintegration of states and society as well as the massive amount of documentation in the twentieth century has led to an international discussion on archival appraisal and preservation. All efforts to come up with a generally applicable model of appraisal have led to dead ends. Obviously, every model of appraisal is in a certain way subjective in that historians' initial questions are formed by their generation: 'Every political system considers different documents worthy of preservation' (Papritz 1983). A free society cannot dictate binding rules for appraisal. The result of this insight was a closer look at Theodore R. Schellenberg's (1903–70) theories. Schellenberg, former administrative head of the US National Archives, first published *Modern Archives. Principles and Techniques* (Schellenberg 1964) in 1956. In the Anglo-American world at least, Schellenberg's ideas changed appraisal theory fundamentally. Schellenberg called for the appraisal of an entire institution's records. Besides the information in the documents about people, places, and events that were documented (the 'content' of records), the principle of evidence is focused on. Evidence in public records involves the archivist finding out how the originating government bodies functioned and what competence they had. For

appraising purposes it is indispensable that the archivist knows how the records in question were produced. The decision of whether or not to preserve records is then made on both formal and content-based criteria.

Papers issued by the Paris-based International Council of Archivists (ICA) indicate that networks of archives exist worldwide, even in developing nations. Because of its rich cultural tradition, China has the greatest number of archivists. Most countries have laws regulating users' access to archives. Conditions on access vary greatly. The Vatican only allows access to documents originating before 1922, while Scandinavians enjoy freedom of information rights that include access to current registries. In the Anglo-Saxon countries and in Germany there is a limitation on access of 30 years. Records on individuals and private data are regulated separately.

4. *Current Problems: Preservation and Information Technology*

Archives deal with two major problems today: the conservation/preservation of archival material, and the effects of information technology. The material that is to be preserved does not age well, compared to medieval parchment. It requires the employment of complicated preservation techniques, among them restoration, microfilm, and digitization.

As far as information technology is concerned, archives are presented with two major challenges: they must develop models for appraising and taking over of electronic media, and they wish to provide universal access for researchers as well as for an interested public. The NARA (National Archives and Record Administration) in Washington, DC are the most experienced in this field. As of May 1999, some 100,000 files comprising more than 500 gigabytes had been taken over and stored there. Some 10,000 new files are expected to be added every year. Data that is transferred must first be copied onto separate tapes or cassettes. This procedure then has to be repeated every 10 years (concept of migration). A provisional list of the data available can be found on the Internet. Since data loss can occur in the copying process, Jeff Rothenberg has come up with a concept for emulating software that allows archivists to make copies using the original or related programs. The concept is currently undergoing tests as part of a project in the Netherlands. The goal of the project is the long-term preservation of digitized data in their original, authentic form (concept of emulation). In Germany, the German federal archives, the *Bundesarchiv*, is the most experienced in the appraisal and storage of digitized data. The *Bundesarchiv* is involved in a pilot project sponsored by the federal administration office called DOMEA (Document Management and Electronic Archiving in IT Supported Operations), which is developing methods of saving, appraising, and preserving electronic files permanently.

For research purposes, the new possibilities which allow the searching of archives both nationally and internationally via the Internet are especially interesting. In Germany, most archives are accessible via the Internet. The state of North-Rhine Westphalia alone has made some 450 guides to state, municipal, church, and business archives accessible for Internet users. Marburg Archive School has presented the prototype of an online search aid especially suited to German descriptive methods, but also with an interface to an international standard of Internet presentation of search aids.

A major undertaking being discussed internationally is called EAD (Encoded Archival Description). In the USA the Berkeley Art Museum, California State University, and the Library of Congress were instrumental in developing EAD, and have been joined by the Public Record Office in London. EAD has found an ever growing following in the Anglo-Saxon world since the mid-1990s. It has meanwhile been made compatible with the international ISAD standard issued by ICA. Actually implementing these standards on a large scale requires a retroconversion of handwritten and typewritten search aids to a digital format. In Britain, home to some 2000 archives, the National Council on Archives introduced an archival network in 1998, which, after retroconversion of existing search aids, will provide a virtual guide to the archives and an incentive to make increased use of the archives over the Internet. The San Francisco-based Research Library Group (RLG) knows no national boundaries. Founded in 1974, the group comprises some 160 member universities, libraries, archives, museums, and scientific organizations. Databases of libraries and digitized finding aids of archives are made accessible by RLG.

In the 1990s, technological progress has changed archives in a way experts never thought possible. Following the digitization of archival guides, and the retroconversion of existing search aids and the organization of this metadata into databases, the digitization of entire fonds seems at least possible. Most progress has been made in digitizing collections of individual items like photographs and posters from various institutions such as museums, libraries, and archives. The digitization of files, that is, documents with several pages, has not yet begun. Information technology has made archives still more accessible and interesting for the public and has blurred the lines between archives, libraries, and documentary institutions.

See also: Archival Methods; Databases, Core: Anthropology and Museums; Databases, Core: Demography and Genealogies; Databases, Core: Demography and Registers; Digital Computer: Impact on the Social Sciences; Historical Archaeology; Historical Demography

Bibliography

Abukhanfusa K, Sydbeck J (eds.) 1994 *The Principle of Provenance. Report from the First Stockholm Conference on Archival Theory and the Principles of Provenance, 2–3 September 1993*. Swedish National Archives, Stockholm

Archival Legislation 1981–1994, 2 Vols., 1995, 1996. Saur, Munich, Germany

Bischoff F M, Reininghaus W (eds.) 1999 *Die Rolle der Archive in Online-Informationssystemen. Beiträge zum Workshop im Staatsarchiv Münster 8.–9. Juli 1998*. Staatsarchiv, Münster, Germany

Black-Veldtrup M, Dascher O (eds.) 2001 *Archive vor der Globalisierung? Symposion des Hauptstaatsarchivs 11.–13. September 2000*. Hauptstaatsarchiv, Düsseldorf, Germany

Brennecke A, Leesch W 1953 *Archivkunde. Ein Beitrag zur Theorie und Geschichte des europäischen Archivwesens*. Köhler and Amelang, Leipzig, Germany

De Lusenet Y 2000 *Preservation Management. Between Policy and Practice*. European Commission on Preservation and Access, Amsterdam

Eckelmann S, Kreikamp H-D, Menne-Haritz A, Reininghaus W 2000 *Neue Medien im Archiv: Onlinezugang und elektronische Unterlagen. Bericht über eine Studienreise nach Nordamerika, 10.–21. Mai 1999* (Veröffentlichungen der Archivschule Marburg No. 32). Archivschule, Marburg, Germany

Franz E G 1999 *Einführung in die Archivkunde*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany

International Directory of Archives 1992. Saur, Munich, Germany

Jenkinson H 1937 *A Manual of Archive Administration*. Lund, Humphries, London

Ketelaar E 1997 *The Archival Image. Collected Essays*. Verloren, Hilversum, The Netherlands

Metzing A (ed.) 2000 *Digitale Archive—Ein neues Paradigma? Beiträge des 4. Archivwissenschaftlichen Kolloquiums der Archivschule Marburg* (Veröffentlichungen der Archivschule Marburg No. 31). Archivschule, Marburg, Germany

Muller S, Feith J A, Fruin R 1898 *Handleiding voor het ordenen en beschrijven van archieven*. Van der Kamp, Groningen, The Netherlands

Papritz J 1983 *Archivwissenschaft*, 4 Vols. Archivschule, Marburg, Germany

Rothenberg J 1999 Avoiding technological quicksand: Finding a viable technical foundation for digital preservation. A report to the Council on Library and Information Resources (http://www.clir.org/pubs/reports/rothenberg/pub77.pdf)

Schellenberg T R 1964 *Modern Archives. Principles and Techniques*. Angus and Robertson, London. (Menne-Haritz A (trans. and ed.) 1990 *Die Bewertung modernen Verwaltungsschriftguts*)

O. Dascher

Archiving: Ethical Aspects

1. Archiving Social and Behavioral Research By-products

Archiving refers to the process of appraising, cataloging, organizing, and preserving documentary material—of any type and in any medium—for open use by specific (e.g., scholarly) or general audiences. The social and behavioral sciences produce intellectual by-products at various stages of the research process that, if preserved and organized, could further basic and applied research, aid policy making, and facilitate the development and replication of effective social intervention programs. A variety of institutions preserve such materials. They include government archives, academic data archives and libraries, and specialized organizations in both the public and private sectors. Professional organizations of social science archivists and librarians have been formed to further the field.

The National Archives and Records Administration (NARA) is the US federal agency that preserves and ensures access to those official records which have been determined by the Archivist of the United States to have sufficient historical or other value to warrant their continued preservation by the Federal Government, and which have been accepted by the Archivist for deposit in his custody (44 U.S.C. 2901). Information about NARA's electronic records holdings (most of which are data files) can be obtained from the Internet site http://www.nara.gov/nara/electronic.

The Inter-university Consortium for Political and Social Research (ICPSR) is the largest academic-based social science data archive. Founded in 1962 at the University of Michigan, ICPSR is a membership-based organization which provides access to a large archive of computer-based research and instructional data in political science, sociology, demography, economics, history, education, gerontology, and criminal justice. More information about ICPSR and its holdings is available from http://www.icpsr.umich.edu.

Sociometrics Corporation was established in 1983. The company's primary mission is the development and dissemination of social science research-based resources for a variety of audiences, including researchers, students, policymakers, practitioners, and community-based organizations. Sociometrics has pioneered in the establishment and operation of topically-focused data, instrument, publication, and (since the mid-1990s) program archives (http://www.socio.com): (a) *Data Archives*: collections of original machine-readable data from over 300 exemplary studies, many of them longitudinal, on the American family, teen sexuality and pregnancy, social gerontology, disability, AIDS and STDs, maternal drug abuse, and geographic indicators; (b) *Instrument Archives*: the questionnaires, interview protocols, and other research instruments that were used to collect the data in the data archives; (c) *Bibliographic Archives*: collections of abstracts of research papers, books, and other publications dealing with topics covered by the data archives; and (d) *Program Archives*: collections of program and evaluation materials from several dozen intervention programs that have proven effective in preventing risky behaviors such as unprotected sex and drug use. These topically-focused archives synthesize research in the field in one

place; facilitate further research with the best existing data and accompanying instruments; promote data-based policymaking; and help service providers and practitioners use the insights gained from research.

Two professional organizations of social science data archivists and librarians are the Association of Population Libraries and Information Centers (APLIC) and the International Association for Social Science Information Service & Technology (IASSIST). APLIC's membership, consisting of both individuals and organizations, represents some of the oldest population and family planning agencies and institutions in the US. IASSIST is an international organization dedicated to the issues and concerns of social science data librarians, data archivists, data producers, and data users. This unique professional association assists members in their support of social science research. The APLIC and IASSIST membership lists provide pointers to the various social science data collections housed all over the world. (http://www.med.jhu.edu/ccp/aplic/APLIC.html; http://datalib.library.ualberta.ca/iassist/index.html).

2. *Ethical Aspects in Archiving*

The development of collections such as those contained in the above archives involves a series of decisions with ethical considerations and implications.

2.1 *Protecting the Integrity of the Selection Process*

Given the limited nature of resources allocable to archiving, how should the contents of the collection be selected? Some archives sidestep this challenge by merely cataloging and warehousing archival material (e.g., data sets) donated to them by the field. While this procedure undoubtedly results in the lowest per-capita archiving cost, the quality of the resultant archival collection is uncertain at best. A better procedure is to set objective technical and substantive standards for inclusion in the archival collection and then actively recruit material meeting or surpassing such standards. The previously described data and program archives at Sociometrics have worked with Scientist Expert Panels in establishing criteria for inclusion in the various collections. For the data archives the selection criteria are scientific merit, substantive utility, and program and policy relevance of the data sets comprising the collection. For the program archives the selection criterion is documented effectiveness in preventing the social problem or disease (e.g., drug use, teen pregnancy, sexually transmitted disease, HIV/AIDS) or in changing these problems' risky-behavior antecedents (e.g., delaying age at first intercourse, increasing the use of contraception and/or an STD-prophylactic at first and every act of sexual intercourse, abstaining from or reducing the frequency of drug use). Having established these objective inclusion criteria, archive staff then work with their respective Scientist Expert Panels to identify and prioritize available data sets and intervention programs for inclusion in the collections. The end result is an archival collection with integrity and credibility.

2.2 *Protecting Respondents' Confidentiality*

Data archives often contain responses to sensitive questions, some of which, for example, ask respondents to admit to illegal, immoral, or 'private' behavior such as abortion, premarital or extramarital sexual activity, mental illness, alcohol abuse, and drug use. How can researchers' need to know (the incidence, prevalence, antecedents, and consequences of these social problems) be balanced against respondents' rights to privacy? This ethical consideration is most often addressed by stripping all archival material of information that could be used to identify individual subjects, for example, name, address, social security number, exact date of birth (often only month and year of birth are included in a public use database). A problem arises when data holders want to strip the data set of key variables such as those measuring the sensitive behaviors listed above, prior to placing a data set in a public use archive. This desire is motivated by the fear that such information could be linked to particular respondents by malicious, hardworking sleuths, even without the help of individual identifiers such as name, address, and so forth. Such censorship restricts the range of uses to which the data set can be put by future researchers, and archives typically make an active attempt to find alternate solutions. For example, users could be asked to sign a confidentiality agreement prior to being allowed access to the data, pledging to use the data for legitimate research purposes only.

2.3 *Censoring Potentially Controversial or Offensive Material in the Collection*

Intervention program archives occasionally encounter an analogous censorship-related challenge. For example, several of the effective programs selected for archiving by the Scientist Expert Panel for the Program Archive on Sexuality, Health and Adolescence (PASHA) contain sexually explicit material that could be viewed as offensive and inappropriate by some individuals and communities. However, because these prevention programs are targeted at high-risk, already sexually active youth, the material could also be seen as appropriate, even necessary, to drive home relevant points. In addition, these programs, like other PASHA programs, meet the collection's inclusion criterion of demonstrated effectiveness in changing sexual risk-related behavior in at least one subgroup of teens. The

decision was made to include the material without alteration or censorship, but to publicize and disseminate the collection as an eclectic one, with different schools and communities encouraged to replicate those programs consistent with their own values, norms, and target populations. A complimentary program abstract was developed so that both the approach and the content of the program packages could be perused prior to requesting the program from the archive.

2.4 Timing of Release of Information to an Archive

Holders of data sets and developers of effective programs often, and understandably, want to reap some payoff from their professional investments by keeping the data or programs to themselves until they have published what they wish from the data (or tweaked the intervention program to their satisfaction). The ethical issue arises when this 'private' or 'proprietary' period of time stretches to what the field would view as abnormally long. This is especially true when the data were collected, or the intervention program developed, with government funds. Several US federal agencies are trying to forestall the problem by building resource-sharing ground rules into the original funding award document. Thus, the grant or contract may specify that the data to be generated from a research study will be placed in a public archive two years after the expiration of the project. This solution gives the original developer the fair 'head start' their efforts have earned, while ensuring that the data collected with government funds will be shared with the field before it gets stale.

2.5 Assignment of Due Credit to Both Original Producer and Archivist

Data sets and program materials typically are received by an archive in a format that data developers and their colleagues found workable, but one not yet suitable for public use. The archivist contributes significant additional value in preparing the database for public use. For example, with the approval of the data donor, inconsistencies in the database are eliminated, or at least documented. The documentation is augmented, both at the study level (describing study goals, sampling and data collection procedures) and at the variable level (assigning names and labels for each variable; documenting algorithms for constructed scale variables). Occasionally, the variable and scale documentation is done using the syntax of a popular statistical analysis package such as SPSS or SAS, facilitating future data analysis. Archivists who prepare intervention program packages for public use make analogous contributions. Program materials are edited and 'prettified' for public use. User's Guides and Facilitator's Manuals are created so that the package is replication-ready in the absence of the original developer. In short, the archiving process is best viewed and executed as collaboration between original developer and archivist. Care must be taken to give due credit for the final product to both individuals, teams, and institutions.

2.6 Tension between Fidelity and Usability in the Archiving Process

The collaborative model is productive not only for assignment of due credit but also for joint resolution of the fidelity vs. usability issues that occasionally arise during the archiving process. Should obvious errors in the data-base be corrected or only documented? Should original program materials that were found effective in the developmental site be altered when replication sites find them unclear or when the curriculum they present is based on out of date data? Issues such as these are best resolved on a case-by-case basis by the archivist and original developer working side by side in collaborative fashion.

2.7 Ownership of the Research and Development By-products

The purposes and procedures of the archive accepting the donation should be made clear to the donor at the outset. It should be communicated to data donors that the research by-product they are donating is being put in a public archive whose main goal is the preservation of the resource. Some archives also actively publicize and disseminate their holdings. In addition, as seen above, archives vary in the extent to which they work with the donor in 'upgrading' the material for public use. The donor should be informed in advance of what to expect along these lines. Issues of credit and ownership should also be agreed to before archiving work begins. How will professional credit for the collaborative product be allocated? Will the resultant product be sold to the end user (at cost or for profit) or given away? If the product will be sold for profit, will royalties be given to the original developer? If the product will be sold at cost, will free or discounted copies be made available to the original developer?

3. Conclusion

Social science research yields many by-products that, if properly archived, could be used to further future research, aid policymaking, and foster the development and replication of effective prevention and treatment programs. Several challenges, some with ethical implications, arise in the archiving process. All

are resolvable with good will and commitment to the public good on the part of both original developers and archivists.

See also: Archival Methods; Archives and Historical Databases; Confidentiality and Statistical Disclosure Limitation; Data Archives: International; Databases, Core: Sociology; Deceptive Methods: Ethical Aspects; Ethics Committees in Science: European Perspectives; Privacy: Legal Aspects; Privacy of Individuals in Social Research: Confidentiality; Research Subjects, Informed and Implied Consent of

J. J. Card

Arctic: Sociocultural Aspects

Western curiosity about the peoples of the North can be traced back to the ancient Greeks. Still, the anthropological study of the Arctic as an area of shared cultural traits and environmental conditions hardly predates the year 1900. The early twentieth century paradigms of diffusionism and environmental determinism were instrumental in creating the simplistic notion of a unified Arctic or circumpolar culture (e.g., Bogoras 1929, Hatt 1914). Detailed ethnographic research conducted since that time has demonstrated that variation is an intrinsic characteristic of Arctic sociocultural systems (some of this research is summarized in Berg 1973, Graburn and Strong 1973, Irimoto and Yamada 1994). Nevertheless, this article approaches the subject by focusing on sociocultural similarities, without denying the existence of considerable differences.

1. *The Arctic and Its Indigenous Inhabitants*

Culturally, the Arctic can be subdivided into a North American and a northern Eurasian part. The North American Arctic is primarily inhabited by speakers of Eskimo-Aleut languages. The old collective ethnonym Eskimo is little used today, and commonly replaced by Inuit and Yupik. Aleut, Yupik, and Inuit societies inhabit the coastal areas of northern North America, stretching from southern Alaska to eastern Greenland. Here the geographical boundary between Arctic and Subarctic coincides more or less with the cultural boundary between Inuit/Yupik/Aleut and Athapaskan and Algonquian groups of North American Indians (see *North America and Native Americans: Sociocultural Aspects*).

In the northern part of Eurasia, the physical boundary between Arctic and Subarctic does not coincide as neatly with cultural boundaries. In Siberia, the cultural realm of the Arctic extends into the Subarctic ecological zone, and reaches its limits only in the steppes of southern Siberia. Samoyedic Tungusic, and Paleoasiatic languages are spoken by the culturally 'most typical' Siberians. Speakers of Turkic languages inhabit large parts of the eastern Siberian tundra and boreal forest, but their historical and cultural background points to Central Asia. In northern Europe, where the vegetational and climatic zone of the Arctic is narrow, conventionally only the Saami (speakers of Finno-Ugric languages) are considered an indigenous Arctic people. Other peoples who also inhabit the northern margins of Europe but are organized into large-scale agricultural societies, are not considered in this overview (see *Europe: Sociocultural Aspects*).

2. *Indigenous and Colonial Histories*

Human habitation of the circumpolar North extends over several thousand years. The direct presence of European colonial powers in the Arctic is a relatively recent phenomenon, but findings of iron and other items that were not produced locally attest to long-standing connections with trade centers to the South. The territory of the Saami has a history of almost 2,000 years of outside intervention, as Vikings pushed north along the western coast of contemporary Norway to extract natural resources and to acquire items of Saami production through trade and tribute. Similarly, European expansion into other parts of the North was fueled by the quest for marketable resources. From the seventeenth century onwards, the rich boreal forests of Siberia and Canada became staging areas for the fur trade. The areas north of the tree line were little affected by fur trapping prior to the twentieth century, but the coastal areas of the Arctic close to the Atlantic and Pacific Oceans became important destinations for the Euro-American whaling industry in the eighteenth and nineteenth centuries.

The early days of colonial rule had little direct administrative impact in many areas of the North. Often it amounted to little more than the purely nominal claim to 'owning' the land, and to sporadic resource extraction. In certain areas, however, the state took a more active role in regulating the lives of the indigenous population: For example, in Greenland a particular form of 'enlightened paternalism' was the guiding principle of Danish rule from the eighteenth century onwards. The second half of the nineteenth century was in this respect far more disruptive than earlier periods: The newly emerging ideology of nationalism brought indigenous peoples from northern Scandinavia to Alaska under increasing pressure to adopt the language and culture of the respective dominant societies.

3. Domains of Arctic Sociocultural Systems

The following section will provide a discussion of various aspects of 'traditional' Arctic cultures. Using the turn of the twentieth century as the ethnographic present, the material is organized along topical lines.

3.1 Ecology and Economy

Characteristic of the Arctic fauna is the relative small number of available species, while the number of individual animals at particular times can be high. Similarly, vegetation north of the tree line is sparse, enjoying only a brief but dramatic growing season. Thus, periods of abundance at a particular locale are followed by periods of shortage, and human groups must schedule their movements accordingly. Due to the wide distribution of permafrost, in combination with climatic and vegetational factors, the practice of agriculture was impossible until recently. Thus, all indigenous Arctic peoples practiced foraging forms of subsistence, such as hunting, fishing, and gathering (see *Hunting and Gathering Societies in Anthropology*).

In the tundra zone, collective hunts of wild reindeer (caribou) during spring and fall migrations constituted extremely important subsistence activities. Along the coasts of the Arctic Ocean the pursuit of sea mammals was life-sustaining. Seals formed the staple, and areas visited by migrating whales and walrus had the opportunity to acquire large quantities of meat through one successful hunt. The lower reaches of major rivers provided excellent opportunities for seasonal fishing. The inland areas of boreal forests could generally rely less on one or two major resources, but had to combine hunting (wild reindeer, moose, etc.), fishing, and gathering. The changes brought about by the fur trade triggered dramatic shifts in the seasonal rounds of the peoples inhabiting the Subarctic. The hunt for fur bearers, which previously had been of little significance, became the most important economic activity.

The major form of subsistence outside of hunting/gathering/fishing is reindeer herding, a form of pastoralism (see *Pastoralism in Anthropology*). Reindeer domestication is found in northern Eurasia, from northern Scandinavia to the Bering Strait, but did not penetrate into northern North America until the late nineteenth century. Among the various local forms of reindeer herding, two major types emerge. Small herds of domesticated reindeer are found throughout the boreal forest regions of Siberia; their primary use is for transportation during hunting. Large-scale reindeer herding of the tundra, on the other hand, is geared toward the maximization of animal numbers, and provides the group with meat and other reindeer products. The only pan-Arctic domesticated animal is the dog, which is used for transportation (and sometimes as a sacrificial animal).

Most economic production was geared toward household and community consumption. Sharing of resources within these limits was a consistent feature of Arctic and Subarctic societies. Exchange relations with other communities were often facilitated through ritual partnerships which extended the limits of reciprocity. With the advent of colonialism, production for outside markets commenced: Furs, meat of domesticated reindeer, and other products were individually appropriated; they fueled the emergence of economic stratification.

3.2 Social and Political Organization

It has long been suggested that Arctic societies are characterized by a bilateral type of social organization (Gjessing 1960). Traditional Saami societies, the Paleoasiatic groups of northeastern Siberia, as well as the majority of Inuit and Yupik societies clearly fit this pattern. The societies inhabiting the western, central, and eastern parts of Arctic Siberia, however, display patrilineal types of organization. Given the fact that vectors of cultural influence in Siberia run generally from south to north, it is possible to suggest that unilineal social organization among those groups was stimulated by southern (Central Asian) influences. The unilineal tendencies of certain Yupik and Aleut societies are less easy to comprehend. There are indications of patrilineal 'clans,' sometimes with endogamous tendencies; however, the basic bilateral framework of Eskimo-Aleut societies (privileging horizontal links over lineal ties) is also present in these cases.

It is noticeable that none of the Arctic or Subarctic societies of Eurasia show any traces of matrilineal organization. In Subarctic North America, on the other hand, the majority of northern Athapaskans (including the linguistically related Eyak and Tlingit) are textbook examples of descent reckoned in the female line. Generally speaking, the Subarctic forests are dominated by unilineal kinship systems. However, it is unclear whether there is a causal relationship between unilineality and Subarctic lifeways and between bilaterality and tundra lifestyles. While bilateral systems are more flexible and seem more appropriate for small-scale groups scattered over large territories than more rigid unilineal systems, population densities in the boreal forests were hardly larger than in the tundra. On the contrary, certain Arctic coastal communities (e.g., near the Bering Strait) sustained quasi-sedentary settlements of several hundred inhabitants; other than along the mouths of salmon-rich rivers, no comparable population concentrations are known from the aboriginal Subarctic.

Throughout the circumpolar North, political organization was local-group-oriented and did not entail hierarchies based on hereditary status. Leadership continues to be situational: It is directed toward the solution of group problems, and the decision to

comply is voluntary. In addition to seniority, individual achievements are most relevant in gaining leadership positions. The major exception are the Sakha/Yakut, whose political organization has been characterized as a chiefdom (Graburn and Strong 1973). For the known cases of ranking in the southern parts of Alaska (e.g., Aleut, Alutiiq), influence from the northern northwest coast of North America (especially from the Tlingit) can be conjectured.

Men have often been the only political actors visible to outside observers. Although it is true that many Siberian societies with patrilineal descent have pronounced male-centered ideologies, not all Arctic societies were dominated by 'man, the hunter.' For example, Inuit males, who seemingly provided the large majority of food resources, were entirely dependent on women to process the meat and skin of the slaughtered animals, to make them into usable food and clothing. Thus, the roles of men and women were strongly complementary and did not sustain rigid gender hierarchies.

3.3 *Religion and Worldview*

Shamans and shamanism are probably the most evocative symbols of circumpolar religion and worldview (see *Shamanism*). There is no doubt that—until recently—most Arctic communities had religious functionaries who were able to communicate with and to 'master' spirits. These 'shamans' were engaged in healing and other activities aimed at improving communal and individual well-being. In the small-scale societies under consideration here, these functionaries held extremely important social positions, which sometimes led to an abuse of power. However, the notion of 'shamanism' can easily be misconstrued as a unified system of beliefs, which it never was in the Arctic. Instead, in addition to a limited number of common elements, circumpolar shamanisms show profound differences in the belief systems with which they are associated. Especially in northern Eurasia, elements of worldviews associated with highly organized religions (such as Buddhism or Christianity) found their way into localized forms of shamanism long before the direct impacts of colonialism.

Animism—the belief that all natural phenomena, including human beings, animals, and plants, but also rocks, lakes, mountains, weather, and so on, share one vital quality—the soul or spirit that energizes them—is at the core of most Arctic belief systems. This means that humans are not the only ones capable of independent action; an innocuous-looking pond, for example, is just as capable of rising up to kill an unsuspecting person as is a human enemy. Another fundamental principle of Arctic religious life is the concept of humans being endowed with multiple souls. The notion that at least one soul must be 'free' to leave the human body is basic to the shaman's ability to communicate with the spirits.

Since the killing and consumption of animals provides the basic sustenance of circumpolar communities, ritual care-taking of animal souls is of utmost importance. Throughout the North, rituals in which animal souls are 'returned' to their spirit masters are widespread, thus ensuring the spiritual cycle of life. While most prey animals receive some form of ritual attention, there is significant variation in the elaboration of these ceremonies. One animal particularly revered throughout the North is the bear (both brown and black), as has been demonstrated by Hallowell (1926) in his classic comparative study of 'bear ceremonialism.'

By the twentieth century, hardly any Arctic community had not yet felt the impact of Christian missionary activity. However, there is considerable variation as to when these activities commenced: Christianity reached the Arctic areas of Europe almost 1,000 years ago, while the indigenous inhabitants of the Chukchi Peninsula (Russia) had little first-hand experience of Christianity before the 1990s. Generally speaking, the eighteenth and nineteenth centuries mark the major periods of religious conversion in the Arctic. Although no other major world religion has significantly impacted the North, the spectrum of Christian denominations represented in the Arctic is considerable. There is also considerable variation in how 'nativized' the individual churches have become.

4. *Contemporary Developments*

Throughout the circumpolar North, World War II triggered developments that made the once-distant Arctic frontiers into strategically important areas. The resulting infrastructural and demographic developments altered the social fabric of the North and put an end to official policies of isolationism. The first two decades after 1945 were generally characterized by state policies geared toward 'modernization' and assimilation. It is frightening how similar state policies from Siberia to Greenland were: Native people were relocated from ancient villages to faceless new towns, a new emphasis on 'productivity' favored newly introduced economic activities over traditional subsistence pursuits, and educational systems were reoriented toward non-native knowledge and skills. It may seem ironic that the very same policies provided the educational infrastructure for those native elites who would challenge the political and cultural hegemony of the colonial powers in the years and decades to come.

Between the 1960s and the 1980s, radical political change affected most parts of the circumpolar North. With the exception of the Soviet North, the general tendency was to repeal colonial status and chart a course toward self-determination. In many cases, opposition to large-scale development projects served as a rallying point for newly developing indigenous

movements. For example, the conflict around the construction of the Alta hydroelectric dam in Norway in the late 1970s brought international attention to indigenous causes, and marks a sea change in the political history of the Saami. In Alaska, it was the discovery of oil in 1968 which put the long-neglected topic of native title to land on the agenda. The resulting 'Alaska Native Claims Settlement Act' (1971), which awarded title to 12 percent of Alaska's land, and cash compensation to newly created native corporations, was then considered a spectacular success.

One of the most impressive results of the era was the 1979 passing of the 'Home Rule Act' in Greenland, which provided for far-reaching autonomy within Denmark (control over most public affairs except defense and foreign relations). Between 1973 and 1993, Finland, Norway, and Sweden installed so-called Saami parliaments, which—although only advisory—provide Europe's northernmost indigenous peoples with powerful symbols of sovereignty. The most recent change on Canada's political map is the April 1999 creation of Nunavut, the first and only Canadian territory in which Inuit make up the majority population.

Developments in the Soviet North followed a different pace. The first 'ethnic' administrative units were formed in Siberia in the 1920s, when such events would have been unthinkable in most other countries, but the political conditions of the subsequent decades reduced notions of 'national autonomy' to mere propaganda instruments. Thus, the reforms of the late 1980s, which led eventually to the demise of the Soviet Union, gave hope for advances in the realm of native rights. Although many changes were undoubtedly positive, the social and economic situation of most native communities in the Russian North deteriorated throughout the 1990s. Many Siberian natives have entered the twenty-first century with nostalgic memories of the Soviet period.

The fact that most indigenous Arctic communities have come a long way since the days of colonial rule and outright discrimination should not blind us to persistent problems. Indeed, most communities face a plethora of cultural and social ills, which are enhanced by precarious economic and ecological conditions (see Smith and McCarter 1997). While many of these problems can be attributed directly to the impact of colonization, the option of isolation from global forces has long since disappeared (if it ever existed). Contemporary Arctic indigenous peoples have understood that the challenge is not to choose between 'Western' modernity and 'unchanging tradition,' but to find a livable combination of the two. Given the political sophistication of local communities, to work in the North has become a tremendously rewarding learning experience for anthropologists and other social scientists. No longer content with being mere objects of study but, at the same time, realizing the potential benefits of social science, Arctic communities are engaged in an ongoing process of defining collaborative and mutually beneficial research.

Bibliography

Beach H 1990 Comparative systems of reindeer herding. In: Galaty J G, Johnson D L (eds.) *The World of Pastoralism: Herding Systems in Comparative Perspective*. Guildford Press, New York

Berg G (ed.) 1973 *Circumpolar Problems: Habitat, Economy, and Social Relations in the Arctic. A Symposium for Anthropological Research in the North, September 1969*. Pergamon Press, Oxford, UK

Bogoras W G 1929 Elements of the culture of the circumpolar zone. *American Anthropologist (n.s.)* **31**: 579–601

Gjessing G 1960 Circumpolar social systems. In: Larsen H (ed.) *The Circumpolar Conference in Copenhagen 1958*. Ejnar Munksgaard, Copenhagen, Denmark

Graburn N H H, Strong B S 1973 *Circumpolar Peoples: An Anthropological Perspective*. Goodyear Publishing, Pacific Palisades, CA

Hallowell A I 1926 Bear ceremonialism in the northern hemisphere. *American Anthropologist (n.s.)* **28**: 1–175

Hatt G 1914 *Arktiske skinddragter i Eurasien og Amerika. En etnografisk studie*. J H Schultz, Copenhagen, Denmark [Trans. 1969 Arctic skin clothing in Eurasia and America: An ethnographic study. *Arctic Anthropology* **5**: 3–132]

Hoppál M, Pentikäinen J (eds.) 1992 *Northern Religions and Shamanism*. Akadémiai Kiadó and Finnish Literature Society, Budapest, Hungary and Helsinki, Finland

Irimoto T, Yamada T (eds.) 1994 *Circumpolar Religion and Ecology: An Anthropology of the North*. University of Tokyo Press, Tokyo

Minority Rights Group (ed.) 1994 *Polar Peoples: Self-Determination and Development*. Minority Rights Publications, London

Paulson I, Hultkrantz Å, Jettmar K 1962 *Die Religionen Nordeurasiens und der amerikanischen Arktis* [*The Religions of Northern Eurasia and the American Arctic*]. W. Kohlhammer Verlag, Stuttgart, Germany

Shephard R J, Rode A 1996 *The Health Consequences of 'Modernization': Evidence from Circumpolar Peoples*. Cambridge University Press, Cambridge, UK

Smith E A, McCarter J (eds.) 1997 *Contested Arctic: Indigenous Peoples, Industrial States, and the Circumpolar Environment*. University of Washington Press, Seattle, WA

P. P. Schweitzer

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Are

Area and International Studies: Archaeology

Within the realm of formally constituted academic disciplines, archaeology is of relatively recent vintage, although interest in the past is a human universal. Every living society has developed means by which to explain its origin and past, and these conceptions of the past are inevitably used to explain, validate, or challenge current conditions. This interest in the past is an archetypal expression of the human mind and represents far more than idle curiosity about things that are either only dimly perceived or altogether unknown; it is a search for roots that are seen as endowing us with an earned and secure place within our social and physical universe. It is perceived and stated histories that locate individuals and groups within larger social networks and assign to them rights, privileges, and obligations vis-à-vis each other, and the resources they extract from it.

Given the universality of human interest in the past, and the importance concepts of the past have in constituting and validating social frameworks of existence, it is not surprising that the academic elaboration of this interest in the form of archaeology is among those scholarly disciplines that attract great popular interest. By the same token, the pursuit of archaeological studies and the protection of archaeological resources tend to be of interest to governments. Most governments regulate access to archaeological resources (sites, museum collections, etc.) and either directly or indirectly support archaeological research.

Since archaeology deals largely with physical evidence in the form of artifacts, geological sediments, remains of plants and animals, chemical and physical traces in the soil, etc., it is sometimes considered to be closely allied with the 'hard' sciences. In the popular view, the analysis of archaeological data results in establishing factual frameworks of what happened in antiquity with a high level of reliability. To some degree, this view is valid with regard to the most basic level of archaeological information and analysis such as dates based on physical and chemical techniques, descriptive studies of artifacts or features, statistical measures of distribution, etc.

Yet, while archaeology shares many of its primary analytical tools with the physical sciences, its ultimate goal is the interpretation of physical evidence in terms of cultural traditions and social behavior, organization, and structures and the processes and causes of their change. Thus, it is truly a social science and shares with the social sciences an important characteristic: archaeological explanations are formulated within, and are dependent on, conceptual frameworks which, in turn, are derived from, or articulated with, broader social philosophies.

In practice, this conditionality of archaeological explanations means that the results of archaeological research tend to be more or less strongly influenced, and in some cases controlled, by prevailing social and political philosophies. It also means that the regulation of archaeological research by governments provides opportunities for the manipulation of archaeological findings and interpretations for political ends and that archaeology may be, and is, used as an arena for cultural and political contests.

1. *Archaeology and Area Studies*

Since the pursuit of area studies aims at an understanding of the cultural and social conditions in contiguous geographic regions, and since historical constructs play such a central role in the construction of culture and social systems, one would expect that archaeology should play a relatively prominent role in the area studies enterprise. An appreciation of the long-term historical background and context under which artistic, intellectual, and social traditions have emerged, and of the conditions within which they have changed over short and long periods of time, as well as an understanding of autochthonous conceptions of history, should be of profound interest. Also, within the comparative methodological framework of the social sciences, the expansion of the heuristic perspective into ancient history and prehistory should be of great interest inasmuch as it vastly increases the range of variability of the human condition that is the subject matter of social science scholarship.

Among the forerunner disciplines of area studies, research into the ancient history and prehistory of colonial dependencies was a prominent and integral part of the activities of Orientalist and Africanist studies as practiced in the nineteenth and early twentieth centuries by academicians and amateurs tied to the great colonial powers of England, France, and Holland. In Asia, much of this work was carried out under the sponsorship of learned societies such as the Royal Asiatic Society of England, the *Ecole Francaise*

d'Extreme Orient, or the *Koninklijk Nederlandsch Aardrijkskundig Genootschap*, or by researchers and administrators associated with colonial government agencies such as the Archaeological Survey of India or the *Service Geologique de l'Indochine*. As a rule, economic, administrative, and scholarly interests were intermingled in the constitution and operation of these organizations.

The archaeological research carried out in those early years continues to provide much of our basal knowledge of the antiquity of those regions, and its findings and interpretations continue to linger in our contemporary perceptions of the pre-modern civilizations and societies of places like Egypt, the Levant, Mesopotamia, and the Indus Valley. At the same time, research in these places had great impact on the emerging discipline of archaeology itself and helped shape some of its fundamental techniques and methodologies. Further to the east, beyond India, the scope and quality of early archaeological work was somewhat more modest, but even there, it was far from negligible or inconsequential. Scholars of that tradition gave us our first detailed accounts and analyses of the 'classical' Buddhist and Hindu monuments of Burma, Cambodia, Indonesia, and Vietnam. During that period, archaeological research in East and Southeast Asia also penetrated deep into prehistory and provided us, among others, with such enduring concepts as the Dongson civilization of Vietnam, the Yangshao culture of China, the Hoabinhian culture of Indochina, and the Homo erectus-grade hominid fossils of Java, Indonesia, and Zhoukoudian, China.

Curiously, within the context of area studies as it emerged after World War II, archaeology has played a relatively minor role. This is surprising for a number of reasons. For one, the field of archaeology has, over the same period of time, seen exponential growth in terms of developments in method and theory as well as in terms of the number of practitioners. Moreover, the growing wealth of Western nations, together with extensive government support for scientific research, has made it increasingly possible for Western scholars to pursue research in far-flung corners of the globe. Finally, most postcolonial nations of Asia, Africa, and Latin America have developed indigenous cadres of archaeologists who are engaged in research programs within their countries, and sometimes beyond, often supported by their governments. Surprisingly, although our knowledge of the ancient history and prehistory of all parts of the world has increased dramatically over the past five decades or so, neither the pursuit of archaeological inquiry nor its results have attracted much interest in the context of area studies.

The relative neglect of archaeology in contemporary area studies probably has several causes. Among them is the strong social science orientation of archaeologists following the American archaeological tradition, which makes it difficult for humanists in the area studies field to appreciate either the processes or products of archaeological research. By contrast, Orientalists tended to practice archaeology from a humanist perspective, and many of them freely moved among a number of subject areas that we now consider distinct disciplines, including linguistics, philology, archaeology, art history, history, literature, and religious studies. Another likely cause of this neglect is that social scientists among area studies scholars tend to be strongly oriented toward contemporary issues and affairs, thus they see archaeology as of marginal relevance. The questions that emerge, then, are: can archaeology make a significant contribution to area studies? What is the nature of this contribution? And, is this contribution critical to our understanding of the social and cultural dynamics of diverse areas and regions of the world?

2. *Historic and Geographical Continuities and Discontinuities*

On a very basic level, archaeology allows us to reveal the broad historical and geographical continuities and discontinuities in social and cultural histories that underlie the emergence of ancient as well as modern cultures and societies and the political structures within which they are embedded. I mean here, for instance, important historical and prehistoric events that literally remade the human landscape, such as the advance of Neolithic farming populations into Mesolithic Europe beginning during the seventh millennium BC or the great migrations of the early first millennium AD; the Bantu expansion in Africa; the Inca conquests of the first millennium BC in South America; the great Han expansion in China during the early first millennium AD; the emergence of new technologies; the evolution of early cities and empires; the entry of world religions such as Buddhism, Islam, and Christianity into East and Southeast Asia. In all cases, these ancient events laid the foundations for subsequent developments in pre-modern and modern history so that their impact still reverberates today. It would be impossible fully to understand the contemporary face of any world region without being cognizant of this historical and archaeological background.

To some degree, discovery and elucidation of these events is based on sets of empirical data, the reliability of which depends chiefly on the amount and quality of archaeological fieldwork conducted. The interpretation of these events in terms of social process and causation, however, is governed by conceptual paradigms constructed within broader intellectual conventions. A good example of this is the issue of the formation of the early, protohistoric states of Southeast Asia. The evidence of these early states and civilizations is found in monumental architecture, sculptural and other arts, inscriptions, and settlement

patterns. All evidence attests clearly to the presence of cultural, artistic, and intellectual elements derived from India and China. The question is: through which processes, and why, did these elements come into Southeast Asia, and what role did they play in the formation of complex local polities?

Earlier generations of Orientalist scholars, influenced by nineteenth-century notions of the cultural capacity of tropical populations, had variously surmised that the process was one of military conquest, deliberate missionization, or colonization through expansionary trade and involved the imposition of colonial dependencies in Southeast Asia by Indian rulers (cf. Coedes 1968). That view was, at least in part, supported by indigenous historical traditions within the region. Expanding empirical archaeological data uncovered since the 1960s and paradigm shifts within the archaeological community have led to extensive revisions. The view at the beginning of the twenty-first century is that the process of state formation in the region was internally driven, and the acquisition of foreign ideologies and material goods, as well as even the construction of historical traditions claiming foreign descent for local rulers, were the result of competition between emerging indigenous elites (Higham 1996).

In a similar vein, when the Bronze Age antiquities of northern China were first discovered in the 1920s, Western scholars assessed them, correctly, as representing a very sophisticated technology and as being part of a highly developed civilization. Yet, they saw no evidence of gradual local development of either technology or the culture associated with it, and were disinclined on *a priori* grounds to credit early societies in China with such advanced developments. Instead, they sought a putative source in Western Asia, which they thought was the font of Bronze Age civilizations in both Europe and East Asia. More recent archaeological research has generated ample evidence for long-term continuous technological and social developments preceding the Shang state, and a change in archaeological thought has led to proposals that see continuities in Chinese tradition reaching back to prehistoric communities of the Yangshao level (e.g., Chang 1986, Keightley 1983).

It is evident that the broad historical and geographical continuities and discontinuities archaeology can reveal bear on our understanding of the growth and nature of contemporary cultural and social conditions. For that reason alone, archaeology should be an integral part of an area studies framework. Perceived continuities and discontinuities are also invoked in constructing historical consciousness and identity by individuals, groups, and nations, and, because archaeological interpretations are conditioned by given intellectual milieus, become malleable tools in ongoing social and political processes. With this, the conduct of archaeological studies as well as the use of its results enters the arena of contemporary political affairs, which is such an important part of international and area studies.

3. *Interpretive Frameworks and Political Agendas*

The conditionality of archaeological interpretations has become painfully clear in the context of decolonization. Since the 1950s, the conduct of archaeological research outside Europe and North America has increasingly passed from Western scholars to indigenous researchers, and the control of archaeological resources and research has passed from colonial powers to newly established, or re-established, nation states in Asia, Africa, and Latin America. In the process, archaeologists of the new generation have increasingly begun to reject older archaeological reconstructions as rooted in a colonial ideology and are engaging in vigorous revisions of interpretations that had become textbook wisdom in previous decades. A number of non-Western scholars have also argued, explicitly or implicitly, that the pursuit of archaeology should ultimately be restricted to indigenous researchers, not only because it engages their heritage, but, more importantly, it is felt that only indigenous researchers have the cultural knowledge necessary to interpret archaeological findings within their domain properly. The most extreme restriction on foreign archaeological research was probably found in the People's Republic of China, where Communist authorities after 1949 completely outlawed any involvement by foreign scholars in fieldwork. This ban was lifted only in 1990, and archaeological research by non-nationals in China is now permitted under highly regulated and supervised conditions.

While it may be difficult to find many examples of archaeological research projects under colonial rule that were explicitly promoted or conducted to advance colonial political interests, it would be even more difficult to deny that the conduct of research as well as its interpretive results were heavily impregnated by the colonial *zeitgeist* and its views of non-European cultures and societies. Thus, the postcolonial critique is justified in principle. By the same token, however, it must be recognized that the epistemological critique cannot be restricted to colonial contexts only. That is, the conduct of archaeological research (e.g., problem orientation, selection of evidence, etc.) and the interpretation of field data, are subject to influences of dominant social and political ideologies anywhere, including those of postcolonial societies.

In rare circumstances, state ideologies exercise overt control of archaeology, particularly in centralized, autocratic political systems. Egregious examples are found in Germany during the 'Third Reich' (Arnold 1990) and countries under Communist control like Russia, China, or Vietnam (Kohl and Fawcett 1996). In China after 1949, archaeologists were explicitly required to apply Marxist-Leninist theory as well as

Maoist thought in archaeological interpretation (Falkenhausen 1993). They were also told that archaeology, like all science, had to serve the goals of class struggle. Such ideological compulsion had relatively little effect on the reporting of empirical archaeological evidence, but it did force broader interpretive attempts into the procrustean bed of the Marxist-Leninist-Morganian framework of patriarchal, matriarchal, slave, and feudal social formations.

Archaeology is commonly exploited for political purposes in the context of defining and fostering local and national identities, advancing national unity in multi-ethnic states, and promoting national prestige, and in the struggle for supremacy among competing ethnic forces. Examples abound from around the world, including Europe (Diaz-Andreu and Champion 1996), the Middle East (Meskell 1998), and eastern Asia.

Trying to forge a new, postcolonial Filipino identity, President Ferdinand Marcos promoted the notion of an early Malay *barangay* society as the basis for Philippine social organization and projected himself as a personification of an ancient Malay *rajah*. In promoting Islam as the focus of Malay identity underlying the young state, Malaysian political leaders have viewed archaeological research dealing with pre-Islamic antiquities with great suspicion and skepticism. Similarly, the nature of the connection between Kofun and Nara-period Japan and Korea is the subject of heated debate (e.g., Ledyard 1975), and the Japanese government is said to have restricted research access to certain monuments for fear of internal political repercussions over potential findings. Both Vietnam and Korea have had difficulties coming to terms with the role of Chinese elements at the emergence of their civilizations (Pai 1992), and Thai archaeologists argue about the nature of the Khmer foundations of ancient Thai history (Vallibhotama 1996).

Once again, China presents a particularly interesting case in point. Even though Chinese Communist leaders were dedicated to destroying all vestiges of the country's feudal history, they did not hesitate to use the spectacular archaeological treasures discovered in the Yellow River valley for promotional purposes on a global scale and to construct and promote a North China-centered model for the origin and unity of Chinese history and civilization. Mao Zedong is said to have seen himself in line with the famous Qin emperor Shihuangdi, whose stunning tomb had been unearthed at Xian.

In the interest of national identity, unity, and pride, archaeologists in China were also under pressure to find evidence of Chinese chronological primacy in the appearance of important prehistoric technological innovations, and suggestions of external influences in the development of Chinese civilization were officially frowned upon. On the other hand, under the influence of liberalization since about 1985, some archaeologists working in southern China have recently begun to use important Bronze Age discoveries in Sichuan and Yunnan not only to advance a multilinear model of development of Chinese civilization (Tong 1987) but also to promote greater decentralization in contemporary Chinese politics—a reflection of an age-old political contest between center and periphery.

Archaeology, then, relates to international and area studies on two planes. In one sense, it is an extension of history and provides important information on long-term historical relationships that predicate the constitution of contemporary cultural and social systems. It would be impossible fully to understand any culture area without this perspective. On the other hand, archaeological understandings, like historical ones, are constructed in the context of political relationships and contests and frequently are invoked as powerful tools in political processes. Indeed, since archaeology deals with tangible, visible evidence, it often supplies even more powerful symbols than historical characters or events. In this sense, archaeology enters the very core of political processes the study of which is an essential part of international and area studies.

See also: Aboriginal Rights; Archaeology, Politics of; Cultural Resource Management (CRM): Conservation of Cultural Heritage; Theory in Archaeology

Bibliography

Arnold B 1990 The past as propaganda: Totalitarian archaeology in Nazi Germany. *Antiquity* **64**: 464–78

Chang K C 1986 *The Archaeology of Ancient China*. Yale University Press, New Haven, CT

Coedes G 1968 *The Indianized States of Southeast Asia*. East-West Center Press, Honolulu, HI

Diaz-Andreu M, Champion T 1996 *Nationalism and Archaeology in Europe*. University College London Press, London

Falkenhausen L v 1993 On the historiographic orientation of Chinese archaeology. *Antiquity* **67**: 839–49

Higham C 1996 *The Bronze Age of Southeast Asia*. Cambridge University Press, Cambridge, UK

Keightley D N 1983 *The Origins of Chinese Civilization*. University of California Press, Berkeley, CA

Kohl P L, Fawcett C 1996 *Nationalism, Politics, and the Practice of Archaeology*. Cambridge University Press, Cambridge, UK

Ledyard G 1975 Galloping along with the horseriders: Looking for the founders of Japan. *Journal of Japanese Studies* **1**: 217–54

Meskell L 1998 *Archaeology Under Fire: Nationalism, Politics and Heritage in the Eastern Mediterranean and Middle East*. Routledge, London

Pai H I 1992 Culture contact and culture change: The Korean peninsula and its relations with the Han Dynasty commandery of Lelang. *World Archaeology* **23**: 306–19

Tong E 1987 The South—a source of the long river of Chinese civilization. *Southern Ethnology and Archaeology* **1**: 1–3

Vallibhotama S 1996 *Syam Prathet. Background of Thailand from Primeval Times to Ayutthaya.* Matichon, Bangkok, Thailand

K. L. Hutterer

Area and International Studies: Cultural Studies

Area studies has been the site in the United States academy where an in-depth understanding of the languages and cultures of other societies has been nurtured. Its institutional development since World War II has been the product of financial pressures in funding from the government and foundations; its intellectual development has been the product of the changing views of language, discourse, and culture that have now come to be known as the 'linguistic turn' in the humanities and social sciences.

In the new millennium, 'area studies' finds itself at a familiar crossroads between funding pressures and new intellectual issues. The development of area studies has always relied upon a mix of foundation and federal support. In the early part of the twentieth century, the establishment of the great philanthropic institutions such as the Carnegie Foundation for the Advancement of Teaching (1905), the Rockefeller Foundation (1913), and then the Ford Foundation (1936), provided the initial financial base for supporting research and teaching about other parts of the world, often as an adjunct to the foundations' own activities abroad. Particularly important in these efforts was the Rockefeller Foundation whose experiences in China led to its establishment of a humanities program that focused on modern languages and area studies, including the co-funding of an East Asian studies program with the Carnegie Foundation and the American Council of Learned Societies. These activities paralleled the growing public interest in China and Japan; the first area studies organization was the Far Eastern Association (now the Association for Asian Studies), founded in 1943.

World War II would bring government support of area studies for explicitly strategic and intelligence gathering purposes. The Far Eastern Association was created under the specter of World War II, and the prospects of a communist China. The containment of communism required understanding the enemy, and it was in this light that the centers of Russian studies at Columbia and Harvard were established with a mix of government (including 'laundered' CIA funding) and foundation support (Cummings 1997). The Ford Foundation soon became the major private support for area studies. From 1951 to 1966, its International Training and Research Program granted US$270 million to US universities and other institutions, to build research and training programs in area studies. In addition, from 1950 to 1996, it also provided US$87 million to the Social Science Research Council (SSRC) and the American Council for Learned Societies (ACLS) for joint area studies programs.

The key event that changed the public and government perception of area studies was the Soviet launch of the unmanned Sputnik satellite in 1957. The United States was immediately seen as having fallen behind the Soviet Union at every level of education, and by the fall of 1958, Congress was set to pass the National Defense Education Act (NDEA). Under Title VI of this act, funds were set aside 'to teach modern foreign languages if such instruction is not readily available'; in addition to language instruction, the government would provide half of the cost of 'National Resource Centers' devoted to the study of specific languages and areas. From an initial annual appropriation of US$500,000, Title VI would grow to US$14 million in 1966. Despite the failure of Congress in 1966 to pass an International Education Act that would have provided additional support for area studies, Title VI and Fulbright-Hays funding continued to support area research and training, including academic infrastructure and library resources. By 1999, there were 119 Title VI centers divided across 11 areas of the world. These form the backbone for present-day area studies.

By the 1990s, 'area studies' was in crisis. Foundation funding had leveled off, enrollments were declining, and the end of the Cold War eliminated one of the motivations that had played so important a role in federal funding: the containment of communism. At the same time, the intellectual configurations that nurtured area studies were changing. In the early 1990s, the fall of the Soviet Union and a newly emancipated Eastern Europe revitalized interest in civil society and even international civil society, but the initial optimism soon faded with the resurgence of embittered nationalisms. Soon, another term began to dominate discussions of area studies: globalization. The challenge of globalization was soon reflected in the proposed reorganization in 1996 of the joint ACLS and SSRC international program that funded much area research. The following statement, written by the president of SSRC, Kenneth Prewitt, would soon launch a thousand proposals, not only to SSRC, but also the foundations involved in supporting area-based research, such as Ford, MacArthur, Rockefeller, and Mellon.

> Now free from the bipolar perspective of the Cold War and increasingly aware of the multiple migrations and intersections of people, ideas, institution, technologies and commodities, scholars are confronting the inadequacy of conventional notions of world 'areas' as bounded systems of social relations and cultural categories. (Prewitt 1996, p. 1)

Almost at the same time, the Ford Foundation initiated its 'Crossing Borders: Revitalizing Area Studies' initiative, in which it described area studies as being 'at a significant and somewhat tumultuous turning point in its history as it attempts to respond to and illuminate dramatic changes in the world in recent decades and to understand complex relationships between the "local" and the "global".' The program would provide US$25 million dollars over a six-year period in a major effort to respond to the changing conditions that area studies found itself in as it approached the new millennium. By reaching out beyond established area research centers to include colleges and smaller universities, the pool of over 200 proposals provided a snapshot of what were the leading edges of area and international studies in the United States, as they confronted an age of globalization. Summarizing the proposals, the introduction to the project website states:

> The projects described here reveal a multiplicity of approaches to rethinking area studies in the twenty-first century. Some work to reframe the very notion of 'area' by exploring an array of new geographies: diaspora and diasporic communities, links between areas (Africa and South Asia, for example), maritime rather than terrestrial perspectives, challenges to American exceptionalism, or a focus on 'globalized sites.' Other projects begin with transnational themes of compelling interest across regions. The legacies of authoritarianism, human rights, social movements, alternative modernities, the rise of new media technologies, and performance and politics are some of the themes addressed.

The intellectual motivation behind many of the proposals was a rethinking of the very notion of culture and its ties to explicitly national frameworks, as the emphasis on diasporic, ethnic, and transnational phenomena indicates. These interests contrasted sharply with those developing in international relations. There were no proposals that dealt with the two themes highlighted in a controversial collection edited by the international relations experts Lawrence Harrison and Samuel Huntington (2000), *Culture Matters: How Values Shape Human Progress*—civilizational perspectives on culture and globalization as economic development. Just as area and cultural studies have begun to 'deconstruct' notions of shared, national cultures in their attempts to deal with globalization, international relations and development economics have combined theses taken from Max Weber's analysis of the Protestant ethic, added a comparative civilizational framework taken from anthropologists such as Robert Redfield, and made civilization-based cultural differences the sources of development (mainly in the West) and underdevelopment (everywhere else).

The origins of these very different approaches to globalization lie in the historical development of the term 'culture.' One important input into area studies has always been the study of languages, especially those of cultures that had their own written histories and literatures. These language-based approaches focused on the historical examination of texts, and required both philological and hermeneutic methodologies taken primarily from the humanities; the result has been a close tie both in theory and practice between notions of language and those of culture. These concerns often clashed with social sciences-based interests in contemporary societies that focused on strategic issues and economic development, in which language played basically a utilitarian role and was never the source of theoretical insight. An uneasy truce emerged in which area departments were placed in the humanities, while area centers were interdisciplinary and more social science oriented. History departments often acted as a bridge, since they could be located either in humanities or social science divisions.

One result is that department-based language and civilization introductions to world cultures have often been part of the undergraduate curriculum in many colleges and universities as a part of 'general education.' With the support of federal (NDEA Title VI) and foundation funding, area programs, such as those at Columbia, Chicago, Berkeley, and Harvard universities, were typically built around a large library collection, and tried to create 'vertical' depth in language instruction, history, and literary studies. The area committees of ACLS and SSRC were staffed by area experts specializing in particular parts of the world (Eastern Europe, Southeast Asia, East Asia, etc.) and funded research on these regions. The emphasis was not on international relations issues, but more area-specific work that required in-depth language, culture, and historical training. Area research thus tended to draw its theoretical and methodological frameworks from the various disciplines they were historically connected to; theoretical advances made first in the disciplines were then applied to area research. The distinctiveness and autonomy of area studies lay not in its theoretical advances, but in language teaching and the translation and preservation of specialized texts.

On the other hand, international relations programs, especially those affiliated with professional schools such as Woodrow Wilson at Princeton, SAIS at Johns Hopkins, the Kennedy School at Harvard, and Fletcher at Tufts, focus on the economics and politics of nation-states. Professional schools usually exist independently of the intellectual agenda of the university, especially the undergraduate curriculum, and therefore have had little influence on the debates about general education. However, their impact on the undergraduate curriculum is accelerating, as many universities have made 'internationalizing' their institutions a top priority for both research and teaching. International studies majors are proliferating, and often require some mix of economics and area-specific

courses, leading some institutions to see international studies as a way of catalyzing multidisciplinary work.

The close links between traditional area studies and its place within the humanities explains the influence of what has been termed the 'linguistic turn,' in which problems of meaning, interpretation, and discourse became central to much of the humanities and social sciences. The linguistic turn was triggered by the development of structuralism after World War II. In 1969, Claude Levi-Strauss published *The Elementary Structures of Kinship*, which applied the structural linguistics of Ferdinand de Saussure and the Prague School to the analysis of kinship and exchange relationships. The result was what might be called a 'paradigm' shift that would have a profound influence on French humanities and social sciences, influencing several generations of thinkers that included Roland Barthes, Michel Foucault, Jacques Lacan, Louis Althusser, Jacques Derrida, and Pierre Bourdieu. It would serve as a lightning rod for the development of feminist, ethnic, and cultural studies in the 1970s and 1980s.

Structuralism arrived in the United States in the mid-1960s, and in 1966, there was a famous colloquium at Johns Hopkins on 'The Languages of Criticism and the Sciences of Man,' in which Derrida, Barthes, Lacan, Gerard Genette, Jean-Pierre Vernant, Lucien Goldmann, Tzvetan Todorov, and Nicholas Ruwet participated. What was impressive was not only the caliber of thought, but its range, including literary criticism, psychoanalysis, history, philosophy, semiotics, and linguistics. Although united underneath a structuralist banner, the conference also introduced post-structuralist thought to US audiences. Post-structuralism would make its initial beach-head at Johns Hopkin's University and would soon conquer comparative literature departments such as Cornell and Yale where the so-called 'gang of four' of Paul DeMan, James Miller, Geoffrey Hartmann, and Harold Bloom would consolidate what came to be known as 'deconstructionism.'

While structuralism had little direct effect on area studies, the focus on language and meaning would have a huge impact across many disciplines. By the late 1960s, the triumvirate of Noam Chomsky, Jean Piaget, and Levi-Strauss promised to uncover the 'deep structures' of the human mind and society. Analytic philosophy focused on the 'ordinary language philosophy' of John Austin, Ludwig Wittgenstein, and the more formalist work of Quine, Kripke, and Davidson. Besides French-based structuralism and post-structuralism, the more German hermeneutic tradition represented by Hans-Georg Gadamer combined with the Weberian tradition of 'verstehen' analysis to influence thinkers as diverse as Clifford Geertz and Jurgen Habermas. The result was that at some of the leading 'area studies' research universities, such as the universities of Chicago and Columbia, language and discourse became models for cultural analysis. The rise of symbolic anthropology at the University of Chicago quickly set off an 'interpretive turn' associated with such figures as Marshall Sahlins, Victor Turner, David Schneider, Stanley Tambiah, and Clifford Geertz; all except Sahlins would move elsewhere, thereby transforming the field's core from social to cultural anthropology.

At Columbia, the presence of Edward Said and Gayatri Spivak would pave the way for what would become known as post-colonial studies. At Berkeley, Stephen Greenblatt would develop what became known as the 'new historicism,' itself heavily influenced by the mix of history and anthropology developed by Geertz. The interest in meaning and discourse would quickly spread from anthropology and literary studies into area studies as part of what became more generally known as the 'linguistic turn,' a term coined by the philosopher Richard Rorty (1967) and applied first to analytic philosophy. The term was quickly applied to all language-based inquiries, as indicated by Rorty's own work on deconstructionism and Derrida. At the same time, there was a dramatic increase in the number of foreign graduate students, especially from South Asia, East Asia, and Latin America whose interests in contemporary cultural developments in their home countries contributed to the growing interest in popular culture triggered by cultural studies. The relaxation of immigration restrictions in the 1960s also increased their presence among the US faculty.

The introduction of these language and discourse themes in the late 1960s and early 1970s would intersect with the development of new social movements inspired by the civil rights movement and the events of 1968. The rise of feminism and ethnic studies (especially on the West Coast where there were degree-granting programs in ethnic studies) catalyzed research into alternative, nonmainstream histories. Post-structuralist formulations of subjectivity and identity, along with neo-Marxist analyses of culture provided not only new theoretical frameworks, but also new areas of study that challenged traditional area studies. British cultural studies introduced a theoretically sophisticated discussion of the mass media and popular culture, including Althusser's reworking of Marx and Lacanian film analysis. Cultural studies entered the United States through communications programs, most notably at the University of Illinois at Champaign-Urbana. Because of its origins in England (the so-called Birmingham school) and its location in communications studies in the United States, it initially had no impact on area studies or anthropology. However, as it became increasingly apparent (most dramatically by the events of 1989) that popular culture and the mass media were playing a crucial role in the development of societies that were the traditional object of area studies, its influence quickly spread to the point where anthropologists began complaining about how cultural studies was

invading their domain and control of the concept of culture.

Even approaches that were not structuralist or post-structuralist contributed to the linguistic turn. The Frankfurt School critiques of mass culture were updated by Jurgen Habermas' work on the public sphere (Habermas 1989, Calhoun 1992) and then his explicitly linguistically based theory of communicative action (1984, 1987) and his critiques of post-structuralism. His ideas would introduce a distinctly communications-oriented dimension to the civil society discourses that became popular after the downfall of communism. Benedict Anderson (1983) would draw upon Walter Benjamin's work for his provocative theory of the communicative origins of nationalism, while Charles Taylor's hermeneutic explorations into the origins of Western concepts of the self (1989) invited comparative work.

The development of these trends paralleled the expansion of area, ethnic, and cultural studies in the 1970s and 1980s. Yet if the events of 1968 could be said to roughly herald in the linguistic turn in the United States, the events of 1989 would prove to be the beginning of its demise. The downfall of communism began the undoing of not only Marxist discourses, but also the relevance of the more politicized domestic versions of multiculturalism and identity politics for global comparisons. The civil society discourse that seemed to hold such promise at the end of the 1980s would quickly unravel in the 1990s, as the uneven development of Eastern Europe and Russia, the clampdowns in China, and the rise of new nationalisms and fundamentalisms, questioned the viability of the civil society concept in an age of globalization.

Whereas in the 1970s and 1980s, theory was very much the at cutting edge of the linguistic turn, by the mid-1990s, 'globalization' had replaced it, not so much as a theoretical category, but rather as an empirical challenge to the standard disciplines and area studies. The transition from the linguistic turn to globalization formed the intellectual backdrop to the Ford Foundation's Crossing Borders Initiative. The over 200 proposals represented the state-of-the-art in area studies, as it attempted to deal with contemporary global transformations. While there was a large variety in theoretical approaches, none emerged as predominant. Instead, most striking was the range of phenomena that the proposals tried to deal with, from opening up ethnic studies and globalizing identity politics, to reconfiguring notions of areas and regions (e.g., thinking of oceans as areas). One lacuna was particularly noticeable and undoubtedly fed into the crisis feeling about area and cultural studies. Despite the calls for interdisciplinarity, there were no proposals that dealt substantively with the relation between economics and culture.

This problem was not unrelated to the linguistic turn in cultural analysis. Despite some innovative works by the literary scholar Mary Poovey (1998) and the philosopher Ian Hacking (1970, 1975), interpretive approaches developed for and applied to literary and historical texts have not been easily transferred to the analysis of mathematical and statistical data. The inability to treat these discourses as cultural phenomena has heightened the split between economic and cultural analyses, just at the moment when historical economics was being eliminated in many economics departments. This gap was exacerbated by the collapse of Marxism, which in the hands of the Frankfurt School and contemporary thinkers such as David Harvey (1982, 1989) and Frederic Jameson (1998), had at least tried to keep the analysis of the relations between culture and capital alive.

The textual focus also overlooked the dynamics of the circulatory processes in which texts are embedded and transmitted. Text-oriented approaches often assumed the fixity of the text artifact and such categories as the author, reader, audience, and the act of reading itself; even the destabilizing moves of deconstructionism presupposed elitist conceptions of reading and re-reading. The linguistic turn's focus on the interpretation of texts, and discourses and the analytic frameworks derived therefrom, produced a paradox: the more approaches treated texts or discourses as intrinsic objects of analysis, the more difficult it was to understand the norms of the interpretive communities that made such approaches possible, i.e., the context that makes interpretation possible. A focus on texts and discourses *per se* overlooks how they serve as ways of connecting individuals and groups and how these connections create various kinds of interpretive communities. Yet, if contemporary globalization involves the creation of new forms of connectedness through a globalized market, mass media, and new technologies, then it is not surprising that the linguistic turn has become increasingly anachronistic as it confronts these phenomena.

Globalization presents a unique challenge to area and cultural studies. A focus on language and identity leaves economic processes unanalyzed. If the seeming triumph of capitalism indicates that the leading edge of contemporary changes is the spread of homogenizing economic processes, then analyses that focus on the cultural dimensions of identity formation seem doomed to play catch-up with forces beyond their control. The civilizational approaches developing in international relations invoke cultural values to explain the failure to develop, thereby presupposing that economic success depends on adopting what are basically Western cultural values; a holistic conception of shared culture as civilization-based values provides the diacritics for developmental differences. Given the historical trajectory of area and cultural studies, it seems highly unlikely that they can go back to a conception of culture that they have been abandoning over the last several decades.

Instead, there is a growing realization in area and cultural studies that while globalization has produced

many convergences in institutions and practices across societies—market economies, bureaucratic states, parliaments, elections—these differ significantly from society. Unlike the rhetoric of economic development, globalization is not a single process spreading across the world. Instead, it spreads through diffusion, influence, imitation, negative reaction, borrowing, etc. that work across complex circulations of goods, technologies, ideas, images, and forms of collective imagination. At the same time, the velocity, scale, and form of these processes and circulations challenge virtually all existing narratives of culture, place, and identity and the intellectual and academic frameworks used to study them.

The complexity of these global processes suggests that if circulation is to be a useful analytic construct that replaces more traditional notions of shared culture, it must be more than simply the movement of people, ideas, and commodities from one culture to another. Instead, recent work suggests that circulation is a cultural process with its own forms of abstraction, evaluation, and constraints that are created by the interactions between specific types of circulating forms and the interpretive communities built around them; the circulation and interpretation of specific cultural forms creates new ways of connecting individuals and groups that can be the bases for social movements, identity formations, and communities. A much cited example would be Benedict Anderson's claim that novels and newspapers help people imagine forms of connectedness among strangers that are at the heart of the idea of nationalism. However, similar themes run through Habermas' work on the public sphere, the Chartier group on print mediation (Chartier 1989, 1994, Martin 1994), and Arjun Appadurai on global flows and 'scapes' (1996).

The linking of circulation with the construction of different types of imagined and interpretive communities also provides a way of overcoming some of the barriers between cultural and economic approaches to globalization. In much of the contemporary work on globalization, culture is seen as that which is subjectively shared by a community, and therefore local. Economic processes are seen as objective, universal, and global. Often overlooked, however, is the interpretive dimension of economic processes. Global capital flows presuppose the intertranslatability of financial instruments and information technologies while at the same time demanding that 'local' economic activities be translated into cross-culturally comparable statistical categories. The circulation of financial instruments rests upon this work of translation and depends upon an interpretive community of institutions (exchanges, banks, clearing houses, etc.) that understand and use these forms.

What these phenomena share is an internal connection between circulating forms and the forms of connection and community built around them. Communities are built around the ways they connect individuals and groups. The ways they are connected depend upon the circulation of specific cultural forms that both enable and constrain how they are to be used and interpreted. Whether it is the novels, newspapers, magazines, publishing houses and coffee houses and salons of Anderson's and Habermas' accounts of nationalism and the public sphere, financial derivatives or junk bonds, or the architecture and codes of the Internet, interpretation and circulation interact to create new forms of collective subjectivity. Instead of simply being the movement of people, ideas, and commodities from one culture to another, circulation is a cultural process with its own forms of abstraction, reification, and constraint. The circulation of a 'community' of forms creates new forms of community with hierarchies of evaluation, contrast, and difference; what might be called a 'culture of circulation.' A culture of circulation would be a 'translocal' form of connectedness that might have the geographical reach of a civilization, but not its fixity or sharedness. It is from these cultures of circulation that new forms of subjectivity and consciousness arise that are the bases for the political cultures of modernity, including the public sphere and nationalism.

Arjun Appadurai has suggested (1999) that research itself should be thought of as an imagined community that spreads through the global circulations of people and ideas as well as institutions and practices. The peculiarity of area studies is not that it is an interpretive community within the larger cultural circulations that characterize modern research; that would be true of any research community. Instead, its potential as a form of knowledge lies in its self-reflexivity. Its objects of study are the cultures of circulation of which it is a part. Both the objects of study and those who study them are increasingly in motion. This 'double circulation' is already creating new communities of research that challenge traditional models of area study by de-centering the Euro-American Academy as the main source of innovative research; post-colonial studies may be the welcome harbinger of things to come. These confrontations are not simply producing competing hypotheses that can be tested for their validity, but are rather the indices of new research imaginaries in the making. A self-reflexive area studies, instead of just providing the data to test theories generated in other disciplines, would be at the cutting edge of understanding the conditions for the production of knowledge in an age of globalization.

See also: Area and International Studies in the United States: Intellectual Trends; British Cultural Studies; Critical Theory: Contemporary; Cultural Studies: Cultural Concerns; Deconstruction: Cultural Concerns; Globalization and World Culture; Language and Society: Cultural Concerns; Linguistic Turn; Postmodernism: Philosophical Aspects; Structuralism; Structuralism, Theories of

Bibliography

Anderson B 1983 *Imagined Communities*. Verso Press, London
Appadurai A 1996 *Modernity at Large*. University of Minnesota Press, Minneapolis, MN
Appadurai A 1999 Globalization and the research imagination. *International Social Science Journal*. June: 229–38
Calhoun C (ed.) 1992 *Habermas and the Public Sphere*. MIT Press, Cambridge, MA
Chartier R 1989 *The Cultural Uses of Print in Early Modern France*. Princeton University Press, Princeton, NJ
Chartier R 1994 *The Order of Books*. Stanford University Press, Stanford, CA
Cummings B 1997 Boundary displacement: Area studies and international studies during and after the Cold War. *Bulletin of Concerned Asian Scholars*. Jan–March: 6–26
Habermas J 1984, 1987 *Theorie des kommunikativen Handelns* [*The Theory of Communicative Action*]. Polity Press, Cambridge, UK
Habermas J 1989 *Strukturwandel der Öffentlichkeit: Untersuchungen zu Einer Kategorie de Bügerlichen Gesellschaft* [*The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*]. MIT Press, Cambridge, MA
Harvey D 1982 *Limits to Capital*. Blackwell, London
Harvey D 1989 *The Condition of Post-modernity*. Blackwell, Oxford, UK
Hacking I 1970 *The Taming of Chance*. Cambridge University Press, Cambridge, UK
Hacking I 1975 *The Emergence of Probability*. Cambridge University Press, Cambridge, UK
Harrison L, Huntington S (eds.) 2000 *Culture Matters: How Values Shape Human Progress*. Basic Books, New York
Jameson F 1991 *Postmodernism, or, The Cultural Logic of Late Capitalism*. Verso Press, London
Jameson F 1998 *The Cultural Turn*. Verso Press, London
Levi-Strauss C 1969 *Les Structures Élémentaires de la Parenté*. [*The Elementary Structures of Kinship*] Beacon Press, Boston
Martin H-J 1994 *The History and Power of Writing*. University of Chicago Press, Chicago
Poovey M 1998 *The Invention of the Modern Fact*. University of Chicago Press, Chicago
Prewitt K 1996 Presidential items. *ITEMS* **50**(2–3): 1–9
Rorty R 1967 *The Linguistic Turn: Recent Essays in Philosophical Method*. University of Chicago Press, Chicago
Taylor C 1989 *Sources of the Self: The Making of the Modern Identity*. Harvard University Press, Cambridge, MA

B. Lee

Area and International Studies: Development in Eastern Europe

The very designation 'Eastern' Europe is itself controversial, and closely connected to the broader issue of the under-development of the area. Uneven economic growth on the European continent produced a neatly regressive pattern of political, social, and cultural development running from the north-west to the south and east, and that was already apparent to observers in the late eighteenth and early nineteenth centuries (Dobrogeanu-Gherea 1910) By the end of the nineteenth century, 'East' generally referred to those areas lying east of the River Elbe and within the Danubian basin. Such a description, of course, tended to be a self-fulfilling prophesy in that the newly independent countries that emerged from the collapse of the German, Russian, Ottoman, and Habsburg monarchies after World War I were treated within the international community as being backward. The collapse of the successor states into dictatorship in the 1930s, and their absorption into the communist orbit after World War II tended to reinforce the image of the region as somehow a world apart from the rest of Europe. After World War II, the study of the region in the English-speaking world took place under the aegis of area studies centers funded for the purpose of understanding communist countries, and therefore requiring specialized knowledge and methods that were not easily transferable to other areas of the world. These factors, historical and political, all conspired to create the region that by the 1960s came to be known in common parlance as Eastern Europe. The countries that fell under this rubric at the height of the Cold War were Albania, Bulgaria, Czechoslovakia, East Germany, Hungary, Poland, Romania, and Yugoslavia, as well as the European areas of the Soviet Union.

1. Debating the Term 'Eastern Europe'

As the communist world slowly unraveled in the 1980s, some students of the region began deconstructing the notion of a unified 'Eastern Europe.' The idea of Mitteleuropa or Central Europe was revived, and analysts debated which countries exactly belonged in this intermediate category. The revival of Mitteleuropa was a project of both East European dissidents, who wanted Western assistance in challenging Soviet hegemony in the region, and British and American social scientists, who genuinely believed in the existence of an alternative political geography of the European continent. The most influential arguments identified the lands of the former Habsburg and German empires—Czechoslovakia, Hungary, and Poland—as genuinely Central European, having more in common with Germany and Austria than with Bulgaria or Russia. Others maintained that even Soviet Ukraine and the Baltic Republics were properly Central European, leaving eventually Russia as the self-affirming 'other,' the sole East European nation. For both political and analytical reasons, after the Cold War many scholars wanted to eliminate the term 'Eastern Europe' altogether.

Eliminating the designation 'Eastern Europe' would indeed make sense if there were no common problems specific to the area. This is not the case, however. Over the twentieth century, scholars both within and outside

Eastern Europe have identified consistently two interrelated features of the region that define it as an object of analysis. The first is the problem of creating a stable institutional order in economically backward countries (a problem of much of the developing world). The second is the problem of importing institutional models developed in different social settings (Gerschenkron 1962). During the twentieth century, Eastern Europe was an ideological and institutional laboratory for every major ideology and institutional order. In roughly chronological order, the countries of the region experienced liberal democracy (1900–30), right-wing or fascist dictatorship (1930–45), Soviet-style communism (1945–89), and once again liberal democracy (1989–the present). Of course, such a periodization is problematic, and misses some important variations within the region. It nevertheless captures much of the reality. With the exception of the present period, the outcome of which is not yet known, scholars have maintained that each of these orders failed due to the relative backwardness of the region, and the corruption of the original institutional design in the face of local resistance or circumstance.

2. *The First Liberal Period*

Even before the collapse of the imperial orders, liberal institutions had been adopted throughout much of Eastern Europe. After World War I, the new post-imperial states (except those in the Soviet Union after 1922) implemented broad constitutional guarantees of freedom of speech and assembly, parliamentary government, near universal male suffrage, and judicial independence. Economic backwardness would be overcome through integrating the economies of these new countries into the broader markets of Western Europe and North America. The rights of ethnic minorities, arguably the most thorny issue in the region, would be guaranteed through a series of treaties and documents drawn up by the League of Nations.

The problem with this institutional design was that liberalism was not home grown. Instead, it had been adopted in order to emulate the West rather than as a response to industrialization and the growth of capitalism. National bureaucracies, for example, developed in anticipation of rather, than in reaction to, increased economic complexity and industrialization. They tended to be overstaffed, inefficient, and corrupt. Although in the Czech lands, due to the particularities of Habsburg economic policy, a native middle class had developed, in most of the other countries of the region, entrepreneurial activity was dominated by ethnic minorities. Even as this situation began to change during the 1920s, as native entrepreneurs grew in numbers, an ethnic division of labor remained in place, and political careers and state employment remained the preserve of the dominant national groups (Janos 1982).

Even more important than the differences in the composition of the middle classes between East and West, however, was the differences in the lower classes, in particular the peasantry. Bloated East European states could only be sustained, and development policies pursued, by extracting resources from an already poor peasantry. Slow industrial development, caused by the economic chaos in Germany after World War I, and disrupted trade after the collapse of the imperial orders, meant that cities could not possibly absorb the huge numbers of landless peasants. Land reform, the initial answer to rural poverty, was largely abandoned, or watered down, in Poland and Hungary in the 1920s, and even where implemented, as in Romania, it tended to create unproductive subsistence holdings whose inhabitants continued to live in squalor (Berend 1996).

Such difficult economic and social circumstances conspired to make liberal democracy an extraordinarily precarious project throughout Eastern Europe. Impoverished peasants, marginalized ethnic minorities, and industrial workers could easily be mobilized into radical, antiliberal politics. Political control and 'democracy' could therefore only be maintained through electoral corruption, 'managed' elections in which certain parties were not allowed to compete, or quasi-military dictatorships (as in Poland after 1926). Liberal political and economic institutions in Eastern Europe were thus corrupted by the circumstances in which they developed.

Even if corrupted, however, liberalism was not abandoned altogether. Elections were fixed in some districts, but they continued to be competitive in others. The police and other state officials sometimes violated rights to free speech and even property, but courts frequently reversed such acts of arbitrariness. Public discourse was often impolite but it continued to exist, and the press remained lively. Not until the rise of the Nazi dictatorship and the presentation of an ideological and institutional alternative did the elites of the region become completely unhitched from their liberal moorings. Even here, however, there are crucial differences among the countries. Czechs, and to a lesser degree Poles, resisted right-wing radicalism because their territory was the immediate object of German revisionist claims. Hungary, Romania, and the Slovak lands, on the other hand, all succumbed to fascist dictatorships, hoping to benefit from Nazi power or at least be spared the more unbearable forms of discrimination that were starting to take shape in the 'new European order' (Polonsky 1975).

3. *Fascism*

There is very little agreement among scholars about the causes or social roots of fascism. Some argue that

it is a form of psychological escape into irrationalism that is inherent in modernity. Others argue that it is the revenge of the middle and lower middle classes on the radical left, a sort of Marxism for idiots (Lipset 1963). Still others maintain that it is in fact a radical form of developmental dictatorship which appears quite regularly in late industrializing societies. Neither East European scholars nor British and North American specialists have been able to resolve the core disagreements on this question. Similarly, regarding fascism's impact on Eastern Europe, economic historians continue to debate whether or not the German economic and trade offensive in Eastern Europe in the 1930s was a net gain or loss for Eastern Europe. In the short run, it appeared to have a positive effect, or at least was perceived to have had one among the East European elites. The design was a simple one and in some ways resembled the arrangements that the Soviets later instituted in the region. East European agricultural goods and raw materials would supply the German military industrial buildup. In return, the countries of the region received credits against which they could buy German industrial goods. Of course, in the long run these credits were all but worthless, and the onset of the war in the east ensured that the Nazis would never repay the debts they incurred in the 1930s. Yet the modest recovery of the East European economies during the early Nazi years could not help but to draw these countries more firmly into the Nazi sphere of influence.

The hope of many regional elites was that Germany and Italy would accept their Eastern neighbors as junior partners as long as their institutional and legal orders mirrored those of the masters. Thus, in Hungary, Poland, Romania, and Estonia, local Hitlers and Mussolinis came to power in the 1930s, and even where they did not come to power they waited in the wings for the day when German or Italian armies would install them at the top of the political pyramid. Of course, the ideological design of the right and its subsequent institutional expressions were far less elaborate or well articulated than the liberal one. This was so not only because it was much newer, but also because most ideologies of the right were explicitly antiprocedural and antiorganizational in nature. The 'little dictators' of Eastern Europe did not, for the most part, share Hitler's racial fantasies, since Nazi ideology had little good to say about the non-Germanic peoples of the area, but they did use the opportunity to free themselves from parliamentary and other liberal restraints in pursuit of economic development and regional power. As in the earlier liberal era, political elites corrupted the pure German or Italian model in an attempt to turn it to their own purposes. Nevertheless, the conflict between the fascist right, which favored a party dictatorship, and the technocratic right which favored a nonpolitical dictatorship was never fully resolved in any country of the region until the onset of the war in the east in 1939.

With the onset of World War II, the scales tilted in favor of the fascist right. An important indicator of these differences can be seen in minorities policies, especially with regard to the Jews. Anti-Jewish laws had been on the books in several countries of the region since the mid-1930s, and in some cases even earlier. By the late 1930s, often under German pressure, but sometimes voluntarily and with a good deal of enthusiasm, they were implemented in full force. It is nevertheless important to distinguish between the institutionalized discrimination of the 1930s and the historical 'revenge' against the Jews that was exacted in horrific form by the Germans and their East European helpers on the fascist right during World War II. From the standpoint of 'development,' the organized massacre of Jews that took place in the region during World War II marks clearly the difference between the corrupted developmentalist model of the East European right during the 1930s, and the antidevelopmentalist bacchanalia of the 1940s. Whereas interwar liberalism had failed in Eastern Europe because it could neither overcome the backwardness of the region nor adapt its institutional order to the problems of scarcity, the fascist order failed because it did not really have an institutional response to backwardness at all. Instead, it retreated into the psychological appeal of glory inherent in war, the pleasure of feeling superior to one's 'inferiors,' or the negative empathy inherent in exacting revenge on one's historical enemies. Although it probably did not appear as such to most East European elites in the early 1930s, by the end of the war it must have been clear that the fascist order ultimately had little to do with development at all.

4. *The Communist Experience*

Although Western scholarly debates on the nature of communism were often influenced by the seminal dissident works of Djilas, Havel, Konrad and Szelenyi, and Solzhenytsin, the political restrictions of Soviet rule in the region meant that the study of Eastern Europe was done mostly from abroad. Two schools of thought dominated the analysis of communism: Totalitarianism and modernization theory.

The totalitarian school was inspired by the writings of Hannah Arendt and Carl Friedrich (Arendt 1966, Friedrich and Brzezinski 1956). Its adherents argued that despite the doctrinal differences between Nazi Germany and communist Russia, they had so much in common that it made sense to group the two dictatorships together as essentially the same. For one thing, both professed an ideology of earthly salvation and were prepared to cast aside conventional moral restraints in order to attain their goals. For another, both destroyed existing civic and personal attachments for the purposes of creating a single locus of devotion. And while the Nazis stressed the importance of the

leader, and the communists the importance of the party, in practice both devolved into personal dictatorships. The Nazis believed in hierarchy, and the communists in equality, but in practice these ideological differences exercised a small impact on political, or even social, organization. Most important, however, these theorists told us, the unprecedented capacity for social control inherent in modern political technologies and bureaucratic organizations render totalitarian orders exceedingly difficult to change. Students of Eastern Europe during the 1950s had little difficulty finding proof of totalitarian parties with instrumental views of their own societies. Private property was expropriated and liberal freedoms were either never restored after the liberation from Nazi Germany or were abolished in steps that culminated with the onset of the Cold War in 1948. Under careful Soviet tutelage, East European secret police forces thoroughly intimidated entire societies. Similar to the Soviet Union of the 1930s, the 'little Stalins' of Hungary, Bulgaria, and Czechoslovakia staged trials of 'traitors' from among the highest ranks of the party, all of whom confessed under duress to having worked for Western intelligence agencies throughout their long careers as revolutionaries.

Although not questioning the characterization of communist politics as essentially antiliberal, scholars inspired by modernization theory in the 1960s began to challenge the totalitarian school's interpretation of the *dynamics* of communism, that is, how it would change over time. The essence of modernization theory is its assertion that, even accounting for broad ideological differences, all societies that industrialize, urbanize, and educate their populations face the same kinds of pressures and will most likely have similar kinds of politics. Furthermore, over time, the functional prerequisites of modern societies produce a convergence of cognitive orientations toward power, politics, and justice. Again, studying the Soviet Union and Eastern Europe after Stalin's death in 1953, and the subsequent critique leveled against Stalin by Khrushchev in 1956, scholars had little trouble finding proof of what they were seeking. The Soviet leadership and its East European counterparts appeared to espouse a more pragmatic, less 'ideological' approach to the problems of their own societies (Hough 1977). No longer were shortcomings the result of 'wreckers' and 'saboteurs,' but rather problems to be dealt with and overcome through the 'scientific technical revolution.' Marxism–Leninism, the official ideology of communist East Europe would not be cast aside completely, but in such highly industrialized countries as East Germany or Czechoslovakia the clash between a mobilizational dictatorship and the prerequisites of industrial modernity would most likely be resolved in favor of the latter.

As it turns out, both schools were wrong. Contrary to the expectation of the totalitarian school, the communist world did change, but it did not change in a direction predicted by modernization theory. Rather than a leadership increasingly infused by rational–technical and pragmatic orientations that would yield policies that worked regardless of ideology, Soviet-style institutions throughout the region produced economic stagnation and widespread corruption. Concerning economic dynamism, the key error of the modernization theorists was to confuse Soviet-style industrialization with capitalist economic development. No Soviet or East European economic theorist was ever able to articulate a nonmarket and post-mobilizational model of economic growth. In fact, the post-Stalinist economists in both Poland, and especially Hungary, articulated quite convincing work which demonstrated just why it was impossible to generate growth based on greater allocative efficiency in a Soviet-type economy. On the question of corruption, in the absence of some mechanism for ensuring the circulation of elites, the end of Stalinist police terror simply turned public offices into private sinecures. Such a possibility was laid out in the early work of Barrington Moore on the Soviet Union and developed into a full-blown Weberian model of 'neotraditionalism' by Ken Jowitt at the beginning of the 1980s (Moore 1954, Jowitt 1992). Jowitt explained the decay of communist rule by the inability of communist leaders to articulate a new, postmobilizational 'combat task' that would have provided a yardstick against which to judge bureaucratic rectitude. Others, mainly Western but also East European economists, began to build related models based on the organizational, as opposed to the ideological, features of the Soviet political economy, which they argued was in essence one giant rent-seeking machine (Kornai 1992).

The Soviet Empire in Eastern Europe underwent significant changes in its 40-year history. As in the interwar liberal and then the fascist periods, local conditions conspired to alter the original institutional design. Communism in the region after World War II began essentially as a classical colonial operation in which local elites were controlled by Soviet supervisors, political direction stemming from of the Soviet embassy, and resources extracted through trade agreements that favored the Soviets. After 1953, however, the various states began to move in their own directions. The essential dilemma for both Soviet and East European elites was that, in order to gain some measure of local legitimacy, policy had to be dictated by local circumstances. These local variations of communism, however, always threatened to go beyond the bounds of what the Soviets wanted in order to maintain a cohesive empire. In 1953 in East Germany, 1956 in Hungary, 1968 in Czechoslovakia, and 1980 in Poland, local communist leaders made concessions to local sentiment by making significant institutional changes. Each time, the logic of these changes led to a weakening of party control, a threat to Soviet hegemony in the country, and, ultimately, a

military crackdown and restoration of communist party rule. During the 1970s and 1980s, the relationship of exploitation between the Soviet Union and Eastern Europe was reversed, with the former subsidizing the latter and shielding it from the full effect of the dramatic increase in world oil prices after 1973.

The growth within Eastern Europe of dissident and antipolitical groupings throughout the 1970s and 1980s, especially the emergence of the revolutionary trade union Solidarity in Poland in 1980–1, unleashed a plethora of interpretations. Some modernization theorists maintained the rise of civil society was the fruit of Soviet-type modernization. After decades of repression, an educated, urbanized, industrialized society had emerged to demand more say in how its affairs were being run (Lewin 1988). Others argued that this really had nothing to do with modernization, but with poor economic performance caused by high level of military spending and dysfunctional economic policy making.

The disagreement was never really settled among academics before the entire system in Eastern Europe began to collapse in 1989. The causes of 1989 continue to be debated. Most scholars point to the importance of Soviet leader Mikhail Gorbachev and his attempt to salvage the Soviet empire by remobilizing society through ersatz democratic structures. Others point to deeper causes: changes in military technologies in the 1980s that effectively bankrupted the Soviet state, the drop in oil prices at the end of the 1980s that depleted hard currency revenues, and the rising cost of empire. Whatever the ultimate reason (or combination of reasons), between 1989 and 1991 European communism disintegrated completely and the states of the region found themselves once again trying to adapt institutions imported from abroad—this time, once again, liberal democracy—to the particular postcommunist conditions of their countries.

5. *Postcommunist Democracy*

Between 1989 and 1991, 27 independent states emerged from the collapse of the Soviet empire and Yugoslavia. A decade later, some of these states had established capitalist economies and meaningful institutions of democratic representation. Others made little progress or quickly slid back into a form of semi-authoritarian democracy. What accounts for the huge differences in outcomes? Some have argued that initial institutional choices shape outcomes in decisive ways. In particular, the choice of strong presidentialism appears to undermine the development of representative programmatic parties, parliamentary responsibility, and civic organization (Fish 1998). Others argue that long-term cultural and bureaucratic legacies affect how willing states are to defend economic and political rights (Kitschelt et al. 1998). Still others maintain that geopolitical position is the main driving factor, especially the capacity of selected countries of the postcommunist world to join the economic and security structures of the West embodied in the institutions of NATO and the EU (Kopstein and Reilly 2000). These highly intrusive institutions have permitted Hungary, Poland, and the Czech Republic to suppress internal disagreements in pursuit of the larger goal of entry to the West. The prospect of being admitted to the West also helped Slovakia and Croatian democrats to overthrow dictatorial postcommunist regimes. East European scholars, now free to contribute to the scientific debates about their own countries, tended to point to a combination of internal and external conditions that determined the initial variation in postcommunist outcomes.

Is history repeating itself? In some ways it is. Once again, the countries of Eastern Europe are attempting to plant the institutions of liberal democracy in unfamiliar soil. Once again, the countries of the region are bit players in the game of international capitalism, trying to 'catch up' with the already developed countries in the West. Yet there are important differences, both internal and external, to Eastern Europe. For the first time in history a select group of East European countries really is now thought of as Western and is being admitted to the economic and security structures of the West. Contrary to the rhetoric of the 1980s, even Poland and Hungary in the pre-World War II era were not really considered fully Western. This has now changed, not only within Europe as a whole but, perhaps more importantly, within these countries themselves. No one doubts any more where these countries are "located". Furthermore, no matter how difficult the transition has been, even for countries such Romania and Bulgaria, there does not appear to be any viable ideological or institutional alternative to liberal democracy. Of course, both of these conditions are subject to change. East and West are not only objective categories but also social constructs. The EU and NATO may close their doors to further membership, or new antiliberal ideological challengers might appear on the new eastern periphery of Europe. If this occurs, one can expect the subversion of liberalism in the region once again.

See also: Communism; Democratic Transitions; East Asian Studies: Politics; East Asian Studies: Society; Eastern European Studies: Culture; Eastern European Studies: Economics; Eastern European Studies: History; National Socialism and Fascism; Revolutions of 1989–90 in Eastern Central Europe; Social Evolution, Sociology of; Socialist Societies: Anthropological Aspects

Bibliography

Arendt H 1966 *The Origins of Totalitarianism*. Harcourt, Brace, and World, New York

Berend T I 1996 *Central and Eastern Europe 1944–1993*. Harvard University Press, Cambridge, MA
Dobrogeanu-Gherea C 1910 *Neoiobagia*. Edetura Libariei, Bucharest, Romania
Fish M S 1998 Democratization's requisites. *Post-Soviet Affairs* **14**: 212–38
Friedrich C J, Brzezinski Z K 1956 *Totalitarian Dictatorship and Autocracy*. Praeger, New York
Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Harvard University Press, Cambridge, MA
Hough J F 1977 *The Soviet Union and Social Science Theory*. Harvard University Press, Cambridge, MA
Janos A C 1982 *The Politics of Backwardness in Hungary 1825–1945*. Princeton University Press, Princeton, NJ
Jowitt K 1992 *New World Disorder*. University of California Press, Berkeley, CA
Kitschelt H, Mansfeldoua Z, Markowski R, Toka G 1998 *Post-communist Party Systems*. Cambridge University Press, New York
Kopstein J S, Reilly D A 2000 Geographic diffusion and the transformation of the postcommunist world. *World Politics* **53**: 1–37
Kornai J 1992 *The Socialist System*. Princeton University Press, Princeton, NJ
Lewin M 1988 *The Gorbachev Phenomenon*. University of California Press, Berkeley, CA
Lipset S M 1963 *Political Man: The Social Bases of Politics*. Doubleday, Garden City, NY
Moore B 1954 *USSR: Terror and Progress*. Harvard University Press, Cambridge, MA
Polonsky A 1975 *The Little Dictators*. Routledge, London

J. Kopstein

Area and International Studies: Development in Europe

The field of International studies in Europe today raises the question of boundaries: geographical, territorial boundaries, and disciplinary boundaries. As a matter of fact, the construction of Europe as a new political space introduces *de facto* a fluidity of frontiers, and a multiplicity of approaches which have led to the reconstruction of the social sciences in which all sorts of boundaries are blurred (Bigo 1996).

International studies has a broader scope than international relations since it refers not only to relationships that a state maintains with other states, but also to relationships between states and other societies, other groups and communities that have emerged and have been organized in other political and cultural contexts than its own. It includes studies on migration, on minorities and ethnicity, and the emergence of transnational communities. It refers also to interactions among actors—individuals and/or institutions—each carrying different national identities to be negotiated on a transnational level (Kastoryano 1997, 1998).

In this perspective, Europe constitutes a specific historical and political setting for the analysis of international studies and its development. It is, indeed, in the eighteenth century that the nation-state, defined as a cultural, territorial, and political unity was born (Rokkan, Tilly 1976). The same nation-state is questioned today as a universal political structure and major actor in international studies. It is again in Europe that, because of the project of a new political unit, called the European Union, concepts such as citizenship, nationality, public space (Habermas 1996, 2000), and cosmopolitanism (Held et al. 1999, Linklater 1998), need to be redefined. Furthermore, cultural, sociological, and political plurality within Europe provides empirical evidence for the development of international studies and more specifically for the analysis of the switch from a realist perspective based on the rationality of the state (Weber) to a liberal one where increasing interdependence among states leads to an analysis in terms of integration, both regional and European.

Such a development leads to a methodological confusion and to an obvious interdisciplinarity. History, sociology, anthropology, political science, and juridical studies contribute altogether to the knowledge and understanding of the political structure, the institutions, and the social organization in a comparative perspective, and of course, of Europe as a new political space. Moreover, the increasing complexity of social and political reality and an inevitable interdependence of internal and external political decisions require a combination of various theories and intellectual frameworks of interpretation, methods, and approaches, as well as conceptual tools of analysis.

1. *Europe of Nation-states*

War and peace during the twentieth century have not only changed the political geography within Europe but have also stimulated an interest in international studies. Born as a reaction to World War I, international studies has focused on the relationship among states on the line of the treaties of Westphalia (1648), declaring a territorial sovereignty of all states of the Empire and their right of concluding alliances with one another and with foreign powers. This perspective developed by the 'realist' theory of international relations relies on concepts such as sovereignty, territoriality, and security. In addition to the Weberian definition of the state—a collectivity that within the limits of a given geographical space claims for its own interest the monopoly of a legitimate violence—the 'realists' have considered the state as a homogeneous unit on the international scene. Its action is qualified as 'rational.'

Following the path of positivism in social sciences, the 'realist' approach expressed for the first time by E. H. Carr in 1939, and formalized by H. J.

Morgenthau after World War II aims at 'objectivity.' The analysis of the international scene relies on a sociological and political knowledge that is 'real' and amoral. It brings to light states' interests and a scheme of rational actions that characterize them. Such an approach reminds one indeed of Auguste Comte's formula: 'to know in order to predict.' The sentence illustrates best the link between international studies and foreign policy as well as security issues: to know about other societies, other political systems, other administrative structures in order to protect the nation and define an appropriate foreign policy.

This tradition based on the logic of expertise has guided the area studies in Europe. More than any theoretical considerations, knowledge about other 'places' and other 'customs' have been produced by the military, by missionaries, and by diplomats. Based on their imperial tradition, France and Great Britain have given priority to the study of their colonies in order to understand the functioning of the society and obviously to exercise their power.

National characteristics appear also in their method in connection with the tradition of social sciences in each country. Whereas France has privileged a juridical and administrative approach in the description and analysis of other countries, Great Britain looked for grand strategy through international history of diplomacy, based on the descriptions of British diplomats recalling the methods of social anthropology, and developed theories on international studies along the line of the International Society tradition of the English school.

Enriched by the missionaries and useful for diplomats, the realist vision on international studies meant to fight, during the interwar period, against the idealist approach according to which ideas are more important than states' interests. Away from a Machiavellian logic, their approach, qualified as 'utopian,' was based on juridical analysis and aimed at finding new theoretical models and solutions to avoid conflict by introducing a moral argument in interstate relations (Kant). The confrontation of realists and idealists nevertheless brought a dynamic perspective in the perception of the state, where moral values can generate social change and affect relations among states.

According to the liberal vision developed in the 1960s in the United States, 'the good of individuals has moral weight against the good of the state or the nation' (Doyle 1997). The state is not a homogeneous unit but is split into various interest groups and individuals; it is considered as an actor influenced by rational individuals acting and shaping the institutions and the political decisions (Keohane and Nye 1971, Waever), and the 'competition among states' takes into consideration the relationship within and across societies (Raymond Aron 1962). Liberal economic and political models have been transposed in international studies in Europe with the objective of reducing the risk of war and establishing a permanent 'democratic peace,' a concept that has gained more legitimacy after the end of the Cold War with a new perspective called liberal internationalism.

2. *The Plurality of Civil Societies*

The new dynamic thus emphasizes plurality within civil society. It takes into consideration multiple voices and movements in decision making. It incorporates into its problematic the variable of identity formed in relation to various institutions.

Along the same lines, research on comparative politics has switched from a focus on state and power, social and political organizations, institutions and law, to processes of decolonization, to theories of modernization, to 'models of democracy' (Held et al. 1999), in short from structures to issues following the historical events that caused 'great transformations' (Polanyi 1944). To compare the state–society relationship implied measuring the implications of such an interaction for the definition of a political culture, elite formation, the understanding of civic virtues, the nature and the scope of social movements. Since there is an increasing interdependence of the states on the international scene, the ways in which each of these issues is treated are supposed to affect the power relationship within society and between states.

Disciplinary approaches and theory followed the evolution and the practices. With the development of social sciences in the 1970s and 1980s history, law, philosophy, sociology, and political sciences have completed the 'traditional' diplomatic history and fostered international studies in Europe (Aron 1962). German tradition, more interested in the study of peace and geopolitics, incorporated international studies in its intellectual tradition, developing a university research basis and a highly theoretical perspective in the domain of sociology, law, philosophy, and economy.

Theories, in order to reflect the reality, have to take into consideration the complex interdependence between national and international institutions, between domestic and foreign politics, between local and transnational actors. The evolution led in the 1980s and 1990s to a neo-realist approach combining scientific theories and empirical research to show the 'anarchical nature of the international system' (Waltz 1979), confronted by a neo-liberal one that emphasizes the 'institutionalization of world politics' and its effect on the ways in which states cooperate (Keohane and Nye 1971). Such a dynamic confirms the understanding of an international society, defined by Hedley Bull as a 'group of states, conscious of certain common interests and common values that contribute to the formation of a society in the sense that they conceive themselves to be bound by a common set of rules, in their relations with one another, and share in the working of common institutions' (Bull 1977).

The international society implies therefore the establishment of common norms and conventions, rules of interaction. These principles constitute the basis for 'new institutionalism', an approach developed in studies of the European Union. They rely on the existence and importance of international institutions as an instance of socialization for individuals interacting beyond boundaries, sharing the same norms and values and changing their conception of interest and identity; therefore the interaction between actors and institutions has shaped each other and national society as well as international society. The switch from a state-oriented to an individual-oriented action in international studies, from interstate relations to 'transnational' relations (Keohane and Nye 1971), has been considered more determinant of world politics.

The French version of the analysis of transnational relations introduces the network built by nonstate actors and political parties and unions and their actions (Merle 1974). In any case transnationality refers to the interaction of multiple actors and strategies, leading to a political action beyond boundaries and even to transnational social movements (Tarrow 2001). The tendency in the 1990s had become, however, the intersubjectivity (Habermas 1996, 2000) and the role of agents—in interaction with states—in the definition of political norms beyond institutional frameworks and their diffusion. Categorized as constructivist, this approach privileges the identity issues and claims for recognition in the public sphere and establishes the society as the locus of political change.

All these competing theories (realist and liberal and their neo's) and their variations (institutionalism, neo and liberal) do not remove the state away from international studies. Some include the state into the transnational space and movement of interaction (Bigo 1996, Kastoryano 1997, 1998, Tarrow 2001). What is at stake is the 'absolute' state as a homogeneous actor, its limits, and its capacity to shape the political community, and the future of the nation state as a universal and legitimate political structure.

3. *Transnational Europe?*

This debate is at the core of the construction of Europe. Europe as a geographical setting has switched to an economic community (EEC) and since 1992, with the treaty of Maastricht, to a political unit called the European Union. The transformation had an effect on area and comparative studies within and among member states. Area studies within European countries have become internal to Europe, and comparative analysis emphasizes the convergence among member states on issues such as immigration, demography, family structure, as well as security, environment, welfare, even citizenship and nationality. (Each of these themes constitutes a broad research program stimulated by different European Institutions and mainly from the Commission.) France, Germany, Great Britain, Italy, or Spain related through multilateral conventions constitute all 'parts' of one expected political setting called the European Union.

The European Union as a new political unit has changed the paradigms in social sciences, raising the question of frontiers, territory, identity, migrations, sovereignty, all together linked to the future of the nation state. New paradigms partly inspired by the classical approaches in international studies take into consideration economic, sociological, and political integration where the local is confronted by the global, where territory is replaced by space, where citizenship is detached from nationality. The realist approach based on interstate relations has been replaced by the study of 'intergovernmental' relations (Hoffmann 1994) focusing on the power of each nation state, whereas a neofunctional approach has drawn attention to the emergence of a European space as a new space for political action or mobilization for states, as well as for groups organized in transnational networks throughout Europe. The liberal intergovernmental theory maintains the state as a rational actor, and argues that the power within Europe is the result of bargaining among governments of member states (Moravcsik 1993). On the other hand, the importance of institutions as conceived in the 1990s was re-evaluated by the neo-institutionalists and 'historical institutionalists' who include in the definition of institutions formal rules, and political norms (Pierson 1996). The main issues were public policies, and their harmonization, the definition of common political agendas (Muller), multilevel governance focusing on the effect of various transnational networks built by economic, social, and political actors, and of 'network governance' according to which national, subnational, and supranational institutions. They all constitute a penetrated system for a transnational policy process (Kohler-Kock), and transnational networks built by interest groups and immigrants.

To reduce the construction of the European Union to theories is to neglect all the complexity of its integration. Europe constitutes *de facto* a space of complex interaction among various actors, states, national and European institutions, and of their interpenetration. The outcome is paradoxical: there is on the one hand an emphasis on the national specificity of each nation state expressed in terms of 'models' and projected onto the European level: French model of citizenship, British model of liberalism, German model of democracy, Scandinavian model of a welfare-state … , each of these being projected onto a European level (Schmidt 1997). On the other hand, the European Union stands for the idea of open-minded conciliation and negotiations of all identities—national, regional, ethnic, religious—for an alternative conception of universality.

In any case the construction of Europe as a political unity is a challenge to nation states, leading them to reformulate the founding principles, to revise national rhetoric, and restructure the institutions. As a matter of fact, migrations from within Europe from without have generated scenarios announcing the end of the nation state or at least its weakening (Habermas 1996, 2000) and the 'transformation of the political community' (Linklater 1998), and questions on ways to go beyond national understanding of democracy, citizenship, on the emergence of a European civil society not limited to the market, on the construction of a new political space and on new 'model' of democratic society and political structure, a European state (Ferry 2000)—plural (Mouffe), multicultural (Kastoryano 1997, 1998), or cosmopolitan (Archibugi and Held 1995).

All these questions are related to the question of European identity and citizenship. The answers are normative and come mainly from political philosophers. They have developed in the 1990s concepts such as 'postnational' to underline the limits and the difficulties of nation states facing the changing political context, and suggest a membership beyond the nation state and their nationalist definition of citizenship (Ferry 2000). Habermas on the other hand sees in the 'constitutional patriotism' a way to unify all the cultural diversity in one common political culture, and projects it onto the European context (Habermas 1996, 2000). This model implies the separation between nationality and citizenship linked in the context of the nation state, therefore a separation between feelings of membership carried by national citizenship and its juridical practice that is extended beyond the nation state. These normative views of citizenship nourish discourses and stimulate debate and research for a new model of citizenship, including the nationals and non-nationals—immigrants. The question remains the emergence of a European public space denationalized, integrated into one political culture carried out by voluntary associations representing a multiplicity of interests in a public arena (Habermas 1996, 2000), on the search for a 'political community' where all the internal diversity of Europe coexist in order to produce an European identity, to define the European citizen and insure his (or her) identification with the new political entity. Studies on the European Union, privileging the political philosophy, sociology, and anthropology, converge implicitly or explicitly towards this direction. They question the identification of citizens organized in interest groups on the European level, or of immigrants resident in one of the member states with a European society. Sociological research brings some evidence on transnational networks from professionals, corporations, and voluntary associations, unions that cover the European space like a spider's web, and have introduced a new way of political participation on a national as well as European level. They show that some of these networks stem from local initiatives, but most of the time are encouraged by supranational institutions (Kastoryano 1997, 1998, Tarrow 2001), which mobilize resources for voluntary associations or groups to help them to consolidate their organization based on identity or interest or both throughout Europe, and contribute to the formation of a transnational European civil society. A transnational political participation is a sign of Europeanization of actions and can produce a European political culture shaped by interaction with supranational institutions.

Studies of the political construction of Europe, like the process of globalization at large, reveals paradoxes. Transnational networks contribute to the formation of external communities. At the same time transnational networks are now imposed on the states as unavoidable structures for the negotiation of collective identities and interests.They aim to influence the state from outside and within. Clearly, the objective of transnational networks is to reinforce their representation at the European level, but their practical goal is its recognition at the national level. In other words, the ultimate goal is to reach a political representation that can only be defined at the national level (Kastoryano, 1998).

Such an argument contradicts recent claims of the end of the nation state. Of course, an organization which transcends national borders such as transnational ones brings to the fore the multiple identifications deriving from the logic of a political Europe, and runs against the principle of nation states. Others argue that the relevance of the nation state in a political Europe does not necessarily imply the erosion of the nation state. In reality the state remains the 'driving force' of the European Union. Even submitted to supranational norms, the state keeps its autonomy in internal and international decisions, and remains the framework for negotiations of recognition. Therefore the permanence of the nation state as a model for a political unit in the construction of Europe relies very much on its capacity 'to negotiate' within and without, that is, its capacity to adopt structural and institutional changes to the new reality (Kastoryano, 1998).

This debate is closely linked to globalization, broader but raising the same conceptual questions of multiple loyalties and citizenship that derives from intercultural interaction in a common political space. Liberal universalism sees in this evolution a new commitment of individuals to a project that is cosmopolitan, and an open membership to a global political community (Held et al. 1999).

The evolution of social reality and the dynamics of the international scene have had a great impact on 'area and international studies' in general and more specifically in Europe. The prospect of European integration raises questions about the relevance of the national boundaries and makes interdisciplinary approaches necessary. The insights at 'realism,' which

have been so dominant in international studies, have not been invalidated. However, new perspectives and paradigms emerge in which the state is one political actor among others (e.g., individuals, groups, institutions) and in which interdependence and interpenetration become keys to international studies.

See also: European Union Law; Globalization: Political Aspects; Nationalism, Historical Aspects of: The West; Nations and Nation-states in History; Western European Studies: Culture; Western European Studies: Society

Bibliography

Arbuchi D, Held D 1995 *Cosmopolitan Democracy. An Agenda for the New World Order*. Polity Press, Oxford, UK

Aron R 1962 *Paix et guerres entre les nations*. Calmann Levy, Paris

Bigo D 1996 *Polices en réseaux. L'expérience européenne*. Presses de Sciences, Paris

Bull H 1977 *The Anarchical Society*. Macmillan, London

Carr E H 1939 *The Twenty Years' Crisis 1919–1939*. Macmillan, London

Doyle M 1997 *New Thinking in International Relations Theory*. Westview Press, Boulder, CO

Ferry J-M 2000 *L'Etat européen*. Gallimard, Paris

Foucher M 2000 *La République Européenne*. Belin, Paris

Groom A J R, Light M 1994 *Contemporary International Relations. A Guide to Theory*. Pinter, London

Habermas J 1996 *Between Facts and Norms*. MIT Press, Cambridge, MA

Habermas J, Rochlitz R 2000 *Après l'Etat-nation. Une nouvelle constellation politique*. Fayard, Paris

Hassner P 1995 *La violence et la paix*. Esprit, Paris Seuil

Held D, Mcgrew, Goldblatt, Perraton 1999 *Global Transformations*. Polity Press, Cambridge, UK

Hoffmann S 1994 *The European Sysiphus. Essays on Europe 1964–1994*. Westview Press, Boulder, CO

Julien E, Fabre D (eds.) 1996 *L'Europe entre cultures et nations*. Editions de la Maison des Sciences de l'Homme, Paris

Kastoryano R 1997 *La France, l'Allemagne et leurs immigrés. Négocier l'identité*. Armand Colin, Paris

Kastoryano R (ed.) 1998 *Quelle identité pour l'Europe? Le multiculturalisme à l'épreuve*. Presses de Sciences-Po, Paris

Katzenstein P J, Keohane R O, Krasner S D 1998 International organization and the study of world politics. *International Organization* **4**(52): 645–85

Keohane R, Nye J (eds.) 1972 *Transnational Relations and World Politics*. Harvard University Press, Cambridge, MA

Lenoble J, Dewandre N (eds.) 1992 *L'Europe au soir des siècles*. Seuil, Paris

Linklater A 1998 *The Transformation of the Political Community*. Polity Press, Oxford, UK

Merle M 1974 *Sociologie des Relations Internationales*. Dalloz, Paris

Moravcsik A 1993 Preferences and power in the European Community: a liberal intergovernmental approach. *Journal of Common Market Studies* **31**(4): 473–524

Morgenthau H J 1960 *Politics Among Nations*, rev. edn. Knopf, New York

Neumann I B, Waever O (eds.) 1997 *The Future of International Relations Masters in the Making*. Routledge, London

Pierson P 1996 The path to European integration. A historical institutional analysis. *Comparative Political Studies* **29**(2): 123–63

Polanyi K 1944 *The Great Transformation*. Farrar & Rinehart, New York

Risse T (ed.) 1995 *Bringing Transnational Relations Back In*. Cornell University Press, Ithaca, NY

Roseneau J 1990 *Turbulence in World Politics: A Theory of Change and Continuity*. Princeton University Press, Princeton, NJ

Schmidt V A 1997 European integration and democracy: the differences among member states. *Journal of European Public Policy* **4**(1): 128–45

Smouts M-C (ed.) 1998 *Les nouvelles relations internationales. Pratiques et théories*. Presses de Sciences-Po, Paris

Tarrow S 2001 La contestation transnationale. *Cultures et Conflits* **38/39**:

Tilly Ch 1976 *The Formation of the Nation-State in Western Europe*. Princeton University Press, Princeton, NJ

Waever O 1998 The sociology of not so international discipline: American and European developments in international relations. *International Organization* **4**(52): 687–727

Waltz K 1979 *Theory of International Politics*. Addison-Wesley,

Weiler J H H 1999 *The Constitution of Europe. Do the New Clothes Have an Emperor? And Other Essays on European Integration*. Cambridge University Press, Cambridge, UK

R. Kastoryano

Area and International Studies: Development in South Asia

Development can be understood as an activity, a condition, an event, or a process. In social science, development is most often studied as a complex set of institutional activities that employ both public and private assets for public benefit. It takes many forms according to ideas and environments that guide its conduct and condition its results. Policies, institutions, outcomes, and analysis together constitute development as a process that is distinct from the related processes of economic growth and social progress, because development explicitly includes the activities of state authorities who establish public priorities and implement policy, includes official relationships among people inside and outside the state, includes public assessments of policy, and includes political efforts to change policy.

Objects and trajectories of development are defined and measured variously. There is thus a vast literature on economic, political, social, cultural, industrial, agricultural, technological, moral, and human development. Even economic development can be assessed by different yardsticks: aggregate increases in

national wealth and productivity are common measures; but national autonomy, food security, and social stability are often important priorities; and a particular state regime's stability, revenue, military might, and cultural legitimacy often preoccupy policy makers. Primary, secondary, explicit, and implicit priorities typically jostle in policy making and various measures of success are typically used by various participants in development debates.

Economic development is the subject of this article. Though 'the economy' as studied by economics consists primarily of markets, 'an economy' is a more complex environment that includes natural endowments, social power relations, and political history. Economic development embraces all the institutional and material conditions that constitute specific economies. Because development requires the self-conscious use of power by particular groups in specific contexts, development regimes represent formations of organized power that define the history of development.

In South Asia, premodern regimes developed regional economies before 1800. A modern development regime emerged under British rule after 1800. National regimes took over development after 1945. Since 1970, the leadership capacities of national regimes have declined as international trends have favored global investors and struggles to represent the poor and previously marginal peoples have favored local movements and nongovernmental organizations.

In 1929, an erudite British agricultural officer, William Moreland, concluded from his research that the 'idea of agricultural development was already present in the fourteenth century.' His conclusion can now be extended much further back in time, because now we know that ancient and medieval rulers in South Asia invested to increase productivity, most prominently by organizing irrigation. By the fourteenth century, royal finance and protection were also expanding markets and manufacturing by building transportation infrastructure. By the eighteenth century, state activities that developed agriculture, commerce, and manufacturing flourished around capital cities in Bengal, Gujarat, Punjab, the Indo-Gangetic plains, and the peninsular river basins.

Premodern regimes increased state revenue and enriched bankers, farmers, and manufacturers. But they worked in what Moreland called a 'political and social environment … unfavourable to [modern goals of development],' because, he said, military and political struggles undermined investments in farming, manufacturing, and banking, as pillage and plunder fed destructive armies and rapacious taxation fattened unproductive ruling elites (Moreland 1929).

The British imperial development regime was built upon a pre-modern legacy but introduced new ideas, institutions, and priorities. In 1776, Adam Smith's *Wealth of Nations* became Britain's first modern treatise on economic development. Smith attacked Crown support for monopolies like the East India Company and promoted the expansion of commerce as the nation's top development priority. British conquest in South Asia proceeded from the mid-eighteenth to the late nineteenth century as Britain became the world's foremost industrial nation. Industrialization helped to sustain the imperial enterprise and vice versa. Modern imperialism defined the first institutional framework for modern economic development in the United Kingdom, British India, Ceylon, and other colonial territories.

Until the 1840s, Indian tax revenues were assigned primarily to meet the cost of conquest, administration, and imperial finance. Policy priorities shifted over decades onto *laissez faire* lines to open India for Britain's commercial interests. In 1813, Parliament renewed the Company charter but ended its trading monopoly and allowed private merchants freer access to British territories overseas. In 1833, Parliament made the Company an administrative institution and made English the official language of state law, administration, and education. British India became a territory for imperial development inside a world empire; and in 1833, the abolition of slavery triggered petitions from Caribbean planters that spurred the Indian government to send shiploads of indentured workers from Calcutta to keep English sugar plantations running in the West Indies.

British industrial interests were prominent in imperial development policy. As early as 1793, public debates ensued on how to best manage of 'Asiatic possessions' in the national interest. Increasingly prohibitive tariffs against Indian cloth protected Lancashire, and after 1815, Lancashire sent cloth virtually free of tariffs to India. As Smith predicted, British consumers benefited from commercial imperialism. English merchants sold Bengal opium in China to buy tea and porcelain for English households; sugar from Caribbean plantations sweetened English tea. Monetary policy kept relative prices of the Indian rupee and English pound favorable for English investors, importers, exporters, and consumers. In 1818, James Mill's *History of British India* composed a British national history, justification, and ideology for British governance in India; and English businessmen were soon cutting Indians out of commercial partnerships to garner national benefits from the imperial trades.

The real value of taxes in India rose rapidly as prices dropped from 1823 to 1854. During this long price depression, it became more cost effective to invest Indian taxes in India. At the same time, outlets for British industrial capital were being sought in London and supply systems for industrial raw materials were being developed. In the 1840s, London launched plans for building infrastructure in India to cheapen English supplies of commodities and raw materials, to expand military operations, to increase revenue, and to extend British capital investments in plantations, railways,

cities, roads, ports, shipping, irrigation, and other ventures.

In the 1840s, an irrigation engineer, Arthur Cotton, argued forcefully that Indian crop production and security could be advanced by state irrigation investments that would pay for themselves with higher taxes on more productive land. At the same time, a commission of Parliament met to consider ways to improve supplies of raw cotton to Lancashire mills. Bombay Presidency attracted special attention, along with Egypt. Measures were sought to expand cotton exports from these regions to counterbalance England's dependence on cotton supplies from the American South. When the US Civil War broke out, Egypt and India filled a void in cotton supplies created by the Union blockade of Confederate ports.

A transition to a modern development regime consumed the decades 1840–1880. In 1853, Governor General Dalhousie announced a plan to build an Indian railway with state contracts that guaranteed English companies a minimum 5 percent return; and to secure that return, government kept control of railway construction and management. In 1871, the Government of India obtained authority to raise loans for productive purposes, and large irrigation projects began, following earlier success raising revenues from smaller projects. Development projects were all government endeavors that employed many native contractors and their benefits also filtered down to native owners of land receiving new irrigation and producing commodity crops.

By 1880, regions of specialized production for world markets had been developed in South Asia. Ceylon was a plantation economy. Coffee plantations expanded from 50,000 to 80,000 acres between 1847 and 1857, and peasants devoted another 48,000 acres to coffee for export. Coffee acreage expanded another 35,000 acres in the 1860s. In the 1880s, leaf disease killed coffee cultivation, which was rapidly replaced by tea, rubber, coconut, and cinchona plantations. Ceylon and India replaced China as the major suppliers of English tea. British plantation investors drove out peasant producers and controlled export markets. Labor supplies posed the major constraint for tea planters, and the solution was found in the institution of (eventually permanent) labor migration from southern Tamil districts in British India. British plantations in Malay colonies also depended on migrating Tamil workers.

British Burma and East Africa also developed in circuits of capital accumulation anchored in India. In Burma, Tamil Chettiyar bankers became prime financiers for agricultural expansion in the Irrawaddy River delta, which generated huge exports of rice for world markets, including India, where urbanization increased demand for imported rice. In East and South Africa, merchants from Gujarat and emigrant workers from Bombay, Calcutta, and Madras provided both labor and capital for railway construction and formed urban nuclei for the colonial economy. Between 1896 and 1928, 75 percent of emigrants from Indian ports went to Ceylon and Malaya; 10 percent to Africa; 9 percent to the Caribbean; and the remaining 6 percent to Fiji and Mauritius, which also became island plantation economies.

The Deccan plateau in India's peninsula became cotton country. In 1876, cotton duties were abolished in England to further cheapen supplies from India, and a year later, the biggest famine ever recorded struck Deccan cotton-growing districts. Under *laissez faire* economic policy and imperial bureaucracy, little was done to alleviate famine suffering, but famine sharpened government attention to investments in protective irrigation. Famine commissions and policies were implemented.

By 1914, most goods arriving at South Asia ports were destined for export: cotton, wheat, rice, coal, coke, jute, gunny bags, hides and skins, tea, ores, and wool. Most cotton came to Bombay from Maharashtra. All tea came to Calcutta and Colombo from British-owned plantations in Assam, Darjeeling, and hills around Kandy. Most export rice came to Rangoon. Wheat came primarily from fields under state irrigation in Punjab (60 percent) and the western United Provinces (Uttar Pradesh) (26 percent). Oilseeds came to Bombay from Hyderabad territory (Andhra Pradesh), the Central Provinces (Madhya Pradesh), and Bombay Presidency (Maharashtra). Coal, coke, and ores came from mines around Jharkhand into Calcutta and Bombay, where they stoked local industry as well as exports. Eastern Bengal (Bangladesh) produced almost all the world's jute, which went to Scotland and then to Calcutta, where jute cloth output surpassed Dundee by 1908.

Between 1880 and 1914, industrial development in India took off during decades of low prices in Europe and America when rising prices in South Asia encouraged investments in India by firms producing for Indian markets and for diversified world markets. Commodity prices in India rose with export commodity production until 1929. Imported industrial machinery was domesticated in new Indian factory towns. It 1853, the first Indian cotton mill appeared in Bombay, and the Factory Act (1881) imposed rules on Indian factories to reduce their comparative advantage in virtue of low labor costs and cheap access to raw materials in India. In 1887, J. N. Tata's Empress Mill arose at Nagpur, in the heart of cotton country, in 1887. The Tatas became India's industrial dynasty. Tata Iron and Steel Works at Jamshedpur consumed increasing supplies of ore and coal, which by the 1920s rivaled exports from Calcutta. In 1914, India was the world's fourth largest industrial cotton textile producer: cotton mills numbered 271 and employed 260,000 people, 42 percent in Bombay city, 26 percent elsewhere in Bombay Presidency (mostly Nagpur), and 32 percent elsewhere in British India, at major railway junctures. Coal, iron, steel, jute, and other

industries were developed at the same time, producing specialized regional concentrations of heavy industrial production around Bombay, Ahmedabad, Nagpur, Kanpur, Calcutta, Jamshedpur, and Madras. Jute mills around Calcutta multiplied from 1 to 64 between 1854 and 1914; the number of looms and scale of employment increased twice as fast. In 1913, manufactured goods comprised 20 percent of Indian exports that were valued at 10 percent of national income, figures never since surpassed.

World War One stimulated policies to enhance India's industrialization to make India less dependent on imports, and the Great Depression, 1929–33, again boosted incentives for industrial growth by reducing prices for farm output compared to manufactures. As a result, industrial output in British India grew steadily from 1913 to 1938 and was 58 percent higher at the end of the Depression than at the start of World War One, compared to slower and more uneven rates of growth in the UK and Germany. By contrast, plantations languished from the early 1900s to the 1940s, the major exception being rubber, which benefited from war booms. Native States and non-British firms participated in the industrial trend. In 1902, Mysore government installed an electric generator built by General Electric with techniques and equipment pioneered at Niagara Falls. Bangalore was the first South Asian city lighted with electricity. In 1921, a third of India's industrial production was driven by electricity and Mysore had a higher proportion of electrified industry (33 percent) than British Indian provinces of Madras (13 percent) or Bengal (22 percent).

By 1920, South Asia contained national economies dominated by agriculture but also including large public sectors and major industries. Indian investors and nationalist politicians were by this time vocal advocates for increasing state development efforts. By 1920, British India was also a land of opportunity for global investors. In 1914, the US Consul at Bombay, Henry Baker, had called it 'one of the few large countries of the world where there is an "open door" for the trade of all countries.' England was still India's dominant trading partner, but was losing ground. In 1914, the UK sent 63 percent of British India's imports and received 25 percent of its exports; and by 1926, these figures stood at 51 percent and 21 percent, respectively. By 1926, total trade with the UK averaged 32 percent for the five major ports (Calcutta, Bombay, Madras, Karachi, and Rangoon). Bombay and Rangoon did 43 percent of their overseas business with Asia and the Middle East. Calcutta did a quarter of its business with America.

South Asia's early globalization also appears in migration data. In 1911, the British in British India numbered only 62 percent of all resident Europeans (54 percent in Native States and Agencies). Four times more immigrants came into India from Asia than from Europe and 7 of 10 came overland from Nepal (54 percent) and Afghanistan (16 percent). In 1911, Nepalis entering British India (280,248) exceeded the resident British population by 50 percent; Asian immigrants were three times as many. By 1921, emigration far exceeded immigration. Between 1896 and 1928, 83 percent of 1,206,000 emigrants left British India from Madras (which accounted for only 10 percent of total overseas trade), and they went mostly to work in Ceylon (54 percent) and Malaya (39 percent). Bombay emigrants went mostly to East and South Africa; Calcutta emigrants, to Fiji and the West Indies.

In 1920, Britain still controlled the highest echelons of South Asia's political economy, but process of capital accumulation inside South Asia had escaped British control. Before the war, London's political position in South Asia seemed secure. After the war, London declined visibly in relation to other metropolitan powers and also to cosmopolitan powers in South Asia that were mobilizing for national control of development. A national development regime emerged inside the British Empire. In 1920, the Indian government obtained financial autonomy from Britain.

Indian nationalists focused sharply on economic issues. The Indian National Congress first met in Bombay, in 1885, and then met every year in late December in a different city of British India. Following the Deccan famines, in1879, Dadabhai Naoroji published his influential *The Poverty of India* to document the negative economic impact of imperial policies on India, and he presided at Congress meetings in 1886, 1893, and 1906, where delegates from all the provinces discussed government policy and argued for lower taxes and increased state development expenditure. In 1905, the Congress launched a Swadeshi Movement to induce Indian consumers to buy Indian made cloth rather than British imports.

The Great Depression dramatized the social cost of India's open imperial economy: it sparked peasant and worker's movements demanding economic security and spurred nationalist efforts to make government more responsible for public well-being in India. By this time, government had long experience as an economic manager and investor in infrastructure. Government owned and managed most mineral and forest resources. Government agricultural departments, colleges, and experiment stations supported scientists and engineers who worked on state-funded development projects. The vast state sector of the imperial economy was managed, however, within a *laissez faire* policy framework that favored foreign investors.

In 1930, the new Congress president, Jawaharlal Nehru, announced new ambitions for national development. He took nationalist economic thought in a new direction when he said 'the great poverty and misery of the Indian People are due, not only to foreign exploitation in India but also to the economic structure of society, which the alien rulers support so that their exploitation may continue,' and he went on

to proclaim that 'In order therefore to remove this poverty and misery and to ameliorate the condition of the masses, it is essential to make revolutionary changes in the present economic and social structure of society and to remove the gross inequalities.'

Bitter experience of state failure, social disruption, and mass death during the Great Depression, the Great Bengal Famine (1943–4), and the Partition of British India (1947) laid the groundwork for national planning that stressed national autonomy, security, and economic integration under strong state leadership. In 1951, Prime Minister Nehru chaired India's Planning Commission, and in the 1950s, all South Asian countries wrote national plans stressing self-sufficiency and addressing problems of national economic growth, poverty, and inequality.

Three decades from the start of India's first Five Year Plan in 1952 to the end of its Sixth Plan in 1985 where the heyday of nationally planned development in South Asia. This was also the most creative period for development theory, a practical strain of economic thought devoted to increasing the productivity and well-being in nations emerging from European imperial control.

National planning required the institutional enclosure of national economies. Around the world, national economies were more self-contained in the 1950s and 1960s than they had been in the heyday of European imperialism. Foreign direct investment declined globally from roughly 10 percent of world output in 1913 to less than 5 percent in the 1960s, when the rate of increase in world merchandise exports was well below the 1.7 percent that pertained from 1870 to 1914.

South Asia's national plans focused on national markets. National planners formulated priorities for allocating state resources acquired both internally and externally. External funds came in grants and loans from countries involved in the Cold War as well as from the Bretton Woods institutions sponsored by the richest capitalist countries. National plan allocations for agriculture and industry were intended to enhance private investment. Planning instituted a new public-and-private apparatus for monitoring national economies. Planning agencies organized regional and local initiatives like cooperative societies and community development programs. National governments set up public food procurement and distribution systems to establish a ceiling on food costs for the poor. National health and education expanded. State ownership expanded to basic industries, public utilities, banks, and insurance.

Economic progress became a central feature of national discourse. Public intellectuals and organizations representing farmer, worker, business, and other interests became intensely involved in planning debates as the national public was mobilized politically under the universal adult franchise. In order to address national needs, however, deficit spending increased demands for external funding; and national economic growth depended on private capital rather than poor voters. Finance and politics pushed nationally planning in opposing directions. Popular participation favored citizen groups while financial pressures favored major investors.

Plans initially focused on industrial import substitution and on producing basic goods in public sector enterprises. Even so, eighty percent of India's industrial production remained in the private sector, where public sector output lowered input prices. Protective controls on imports, exports, and operations inside national markets were stricter than ever before and spawned a regulatory bureaucracy as well as black and gray markets. Plan allocations were in practice mixed in with political patronage. By the late 1960s, foreign exchange shortages began to put private and public sector companies into direct competition for funding. Nehru died in 1964. Drought and famine struck India in 1965–7 and the food distribution system relied on US foreign aid. As a result, planners thrust new energy into the Green Revolution, which combined irrigation, pesticides, and high-yielding hybrid wheat and rice seeds. Plans concentrated on extending the Green Revolution by investing in sites of intensive cultivation where well-endowed landowners controlled local labor, finance, and political institutions. Critics called this strategy 'betting on the rich.' Defenders saw the Green Revolution as the foundation of national food security.

During the 1970s, state planning began to lose its grip on development. Policy makers in Sri Lanka, Bangladesh, and Nepal were first to shift priorities away from national autonomy as they sought to meet demands from the urban middle classes and rural landowners by using massive external assistance for large development projects, epitomized by the Mahaveli scheme in Sri Lanka, then the largest irrigation project in the world. New external debt came with new conditions. From 1973, rising oil prices brought recession along with inflation to rich countries in Europe and North America, as it drove up the cost of South Asia's industrial growth, middle class consumption, and the Green Revolution. The effects were most drastic in smaller South Asian countries, which began borrowing on a much higher scale and soon came under structural adjustment policies introduced by the World Bank and International Monetary Fund, where development theory had shifted to focus on government policy reform in borrowing countries. Among economists, a critique of national state control of development became more insistent.

In 1981, India also began to rely more heavily on foreign debt. By 1990, internal pressure from middle class consumers and major industrial concerns combined with conditions imposed by external lenders to force the liberalization of economic policies in favor of freer market operations. State policy shifted away from regulated planning toward institutional reforms,

which have preoccupied government since 1991 and opened India's economy substantially.

The 1980s and 1990s witnessed a profound shift in relationships among participants in development. Dismantling central government controls, opening governments to public scrutiny and popular participation, and making the state more accountable and transparent for private citizens became the order of the day. In India, private capital and state governments gained more independence from a central government that is today composed of shifting coalitions of regionally based parties rather than being dominated by a single national party. State governments gained powers to make contracts with foreign countries and businesses. State Chief Ministers now compete to attract investors. In Nepal, electoral democracy was established in 1991, opening development to wide public debate at the same time as foreign investments grew. In Pakistan, a national government threatened by struggles for regional autonomy also had to absorb disruptions from two decades of war in Afghanistan, leading to more stringent authoritarianism. Since 1981, Sri Lanka has been wracked by civil war over Tamil regional autonomy; and like Bangladesh, it depends on foreign investors while it struggles to resist reforms that undermine national sovereignty.

The problem of governance became increasingly central in debates about development. Since 1980, active institutional participants in development have multiplied in all the countries as global investors have increased their power inside national economies. Together, these two trends have weakened the capacities of national governments to maintain strong leadership. At the same time, government reform and privatization of public enterprise has become a policy priority for international funding agencies. Effective governance in development has scattered and fragmented, while the responsibility of the national state for macro-economic management, property and human rights protection, and political stability has become more demanding.

Development today has no one guiding vision or dominant logic and several contradictory trends are prominent. National economies are more global as are the cultural communications that shape national politics. In the 1990s, television media owned by multinational corporations flooded public information systems. The growth of exports from South Asian countries measured 13.5 percent annually in the 1990s, almost four times the rate of the 1970s. Foreign direct investment (FDI) grew, though it remains a small proportion of India's GDP at 0.1 percent before 1991 and 0.5 percent in 1992–6. In 1990–6, FDI increased (in millions of US dollars) from under 100 to over 5,000 in India, from under 250 to over 650 in Pakistan, from under 60 to over 600 in Bangladesh, and from under 60 to over 2,400 in Sri Lanka. In the first six months of 1996 alone, Korean companies made nine technical and 25 financial agreements in India.

Forging alliances between national and international business preoccupies national policy and linkages between FDI and national investors have increased the pool of investment capital inside national economies. A repeat of the nineteenth century trend that created specialized economic regions is underway. In Nepal, tourism and hydroelectricity attract foreign and domestic partnerships. In Bangladesh, the garment industry has been the fastest growing employer relying on imports for all material inputs and exporting all its output. Sri Lanka is a free-trade zone. The South Indian cities of Bangalore and Hyderabad are growth nodes for high technology and business collaboration. The Sylhet region of Bangladesh specializes in labor exports to Britain.

Globalization has not fostered much cooperation among South Asian countries; rather, national governments and business compete in world markets. More market activity crossing national borders escapes regulation and monitoring. External labor migration has reached staggering proportions but is impossible to assess empirically. The largest overseas flow is to the Persian Gulf, where Bangladesh alone sent 1,600,000 workers in 1995. Available data indicate that only a fraction of remittances are recorded and that most flow through informal channels to finance domestic consumption, investment, and foreign trade in the migrants' home country. Illegal trades also flourish in drugs and arms; organized crime has gone beyond its old interest in black market radios and videos to trafficking in women and child sex-workers.

More citizen groups have become vocal critics of state leadership, priorities, and administrative practice in development. Popular movements against the Narmada Dam in India and against the Arun Three hydroelectric project in Nepal represent many mobilizations to make development more respectful of the environment and responsible to people marginalized and displaced by state development projects.

Countless grassroots movements now seek to wrest control of development from national states. These include regional and local democratic movements fighting for the interests of farmers, workers, industrialists, women, and the poor; they also include the Maoist insurgency that has spread like wildfire in Nepal, Tamil separatists in Sri Lanka, and militant struggles for autonomy among tribal peoples in several mountain regions.

Nongovernmental organizations (NGOs) have become prominent development institutions. NGOs number in the hundreds of thousands. Most are small and locally financed but some have grown huge by combining local initiatives with government funding, international finance, and business activity. In 1976, the Grameen Bank was established by Muhammad Yunnus in Bangladesh to make small loans to poor women and today it counts its clients in the millions

and values its loans in billions of dollars. Despite its size, however, Grameen still reaches a tiny proportion of the rural poor in Bangladesh.

Contemporary development includes contradictory tendencies that do not form one dominant trend. Globalization, regionalism, and localization are progressing at the same time. The conventional use of national statistics to study development has become inadequate as economic conditions have become more disparately local, regional, national, and global in their form and content. Overall economic growth accelerated in the 1990s, but there were also a series of good monsoons, poverty did not decline significantly, and inequalities as well as instability and conflict over development increased.

Who is leading development, who is benefiting, and where today's trends are moving remain debatable. Some analysts have said that development itself is dead. It is more accurate to say that development entered a new phase in the last decades of the twentieth century when increasingly numerous, vocal, and contentious participants organized effectively to pursue disparate, perhaps contradictory goals, including free market globalization, economic growth, ending poverty, and empowering the poor majority of citizens in South Asia who have never had their own effective institutional voice.

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Development, Economics of; Development: Social; Economic Growth: Theory; South Asia, Archaeology of; South Asian Studies: Culture; South Asian Studies: Economics; South Asian Studies: Politics; Western European Studies: Religion

Bibliography

Agarwal A, Narain S 1989 *Towards Green Villages; A Strategy for Environmentally-sound and Participatory Rural Development*. Centre for Science and Environment, New Delhi

Bagchi A K 1987 Development planning. In: Milgate M, Newman P, Eatwell P (eds.) *The New Palgrave: A Dictionary of Economics*. Macmillan, London, Stockton Press, New York, Marnzen, Tokyo

Barber W 1975 *British Economic Thought and India 1600–1858: A Study in the History of Development Economics*. Oxford University Press, Oxford, UK

Bardhan P 1984 *The Political Economy of Development in India*. Oxford University Press, Delhi

Chandra B 1966 *The Rise and Growth of Economic Nationalism in India: Economic Policies of Indian National Leadership*. People's Publishing House, New Delhi

Chaudhuri P 1979 *Indian Economy: Poverty and Development*. St Martin's Press, New York

Drēze J, Sen A (eds.) 1998 *Indian Development: Selected Regional Perspectives*. Oxford University Press, Delhi

Frankel F R 1978 *India's Political Economy, 1947–1977: The Gradual Revolution*. Princeton University Press, Princeton, NJ

Habib I 1982 *An Atlas of Mughal Empire: Political and Economic Maps With Notes, Bibliography and Index*. Oxford University Press, Delhi

Hossain M, Islam I, Kibria R 1999 *South Asian Economic Development: Transformations, Opportunities and Challenges*. Routledge, London

Johnson B L C 1983 *Development in South Asia*. Penguin, Harmondsworth, UK

Kabeer N 1994 *Reversed Realities: Gender Hierarchies in Development Thought*. Verso, London

Kothari R 1971 *The Political Economy of Development*. Orient Longman, Bombay

Kumar D (ed.) 1970 *The Cambridge Economic History of India, Vol. 2, circa 1757–1970*. Cambridge University Press, Cambridge, UK

Leys C 1996 *The Rise & Fall of Development Theory*. EAEP, Nairobi and Indiana University Press, Bloomington, IN

Mahbub ul Haq Human Development Centre *1999 Human Development in South Asia 1999*. The University Press Limited, Dhaka

Moreland W 1929 *The Agrarian System of Moslem India*. Cambridge University Press, Cambridge, UK, pp. 205–6 [reprint 1968 Delhi]

Myrdal G 1968 *Asian Drama: An Inquiry into the Poverty of Nations*. Random House, New York

Majid Bawtree, Rahnema V (eds.) 1997 *The Post-Development Reader*. University Press Limited, Dhaka

Raychaudhuri T, Habib I (eds.) 1982 *The Cambridge Economic History of India, Vol. 1*. Cambridge University Press, Cambridge, UK

Tomlinson B R 1993 *The New Cambridge History of India: The Economy of Modern India, 1860–1970*. Cambridge University Press, Cambridge

D. Ludden

Area and International Studies: Development in Southeast Asia

Southeast Asia as a political/geographical entity came into 'existence' during World War II when the Allied Chiefs of Staff divided the world into specific 'war' commands. The Southeast Asia Command covered all the present countries of the Association of Southeast Asian Nations (ASEAN) with the exception of the Philippines. After independence, the Philippines was also included in the 'entity.' Thus, Southeast Asians themselves did not know that they were 'Southeast Asians' until the Europeans and Americans informed them. Consequently, they made little use of this connection, and Southeast Asian scholarship showed no interest in the larger unit other than their individual countries. Nevertheless, in universities set up by the colonial powers, the emphasis was on colonial histories, colonial possessions, and their common heritage.

The defining moment for awareness of being Southeast Asian came with the formation of small regional associations (under the aegis of the former colonial powers) and the larger ASEAN entity in 1967. ASEAN

originally comprised the five most economically advanced countries of Singapore, Malaysia, Thailand, the Philippines, and Indonesia and by 1999 had expanded to include all 10 countries in Southeast Asia. This expansion as a political association or unit also led to an increasing use of the term the 'ASEAN region,' rather than the Southeast Asian region. Indeed, although the former term is widely used in the United States, the United Kingdom, Europe, and Australia, there is an increasing trend to use the latter in the region. The term has been widely adopted by the ASEAN Committee on Culture and Information which has requested that researchers undertaking research programs under its umbrella to concentrate on 'ASEAN studies.' Thus, although ASEAN studies can be considered a specialist area within Southeast Asian studies in general, much like Thai studies, for example, nevertheless it encompasses the entire region. In this article, the term ASEAN is used to denote all countries in the region but the preferred usage is Southeast Asian Studies.

Most of the original members of ASEAN have achieved remarkable economic growth over the last 30 years. Behind this lies high levels of investment, open economies, and 'outstanding' export performance. This stimulated a rethinking of Southeast Asia's history and the meaning and context of the growth, leading to diverse studies of the region. In the 1990s, political change and internal transformations related to 'new' political visions, democratization and civil society, and recently the financial crisis and a slow return to previous levels of prosperity have again focused international attention on the region. While in the past emphasis was placed on the need to synthesize and compare variety and variation in Southeast Asia, the current trend is to see the region as an essential player in the global economy. This growth and change in focus have by no means been confined to 'Western'-based scholarship only. Southeast Asianists themselves have actively participated in the studies of their region in the wider context of the world around them and in the process made efforts to internationalize Southeast Asian studies.

1. *Early Approaches and Frameworks by Specialists Mainly Outside the Region*

There is an important heritage of scholarship relating to Southeast Asia within the Asian region, principally by Chinese and Indian scholars. One focus of Indian scholarship was the Indian Ocean area in the early modern period, the resilience of Asian economic forms and activities, and the relations between the Asians and early Europeans in the region. Chinese scholarship focused on China's tributary state and trading relationships with native rulers, and observations on European colonialism in Southeast Asia, especially in the nineteenth century. This approach essentially viewed the Southeast Asian (and Asian) region in terms of networks of maritime trade and trading ports connected within and across 'national' boundaries.

The internationalization of Asian studies, which commenced with the first International Congress of Orientalists convened in 1873 in Paris, revolved around the needs of textual scholars whose enquiry concentrated on linguistic, religious, cultural, and political pluralisms. Asian diversity was included in all congresses, which were held in rotation in all continents. Contemporary political issues, which had been excluded from the earlier meetings of the congress, were included from 1954. The name of the congress was also changed to CISHAAN (*Congres International des Sciences Humaines en Asia et Afrique du Nord*) in 1973, and in 1983, the English form ICANAS (International Congress of Asian and North African Studies) became official.

In the twentieth century, the development of Southeast Asian area studies was closely related to the European powers' imperial interests in Asia and the need for expertise in the relevant vernacular languages. The work of scholar-administrators in the then Indochina, the Netherlands East Indies, Malaya, Burma, and the Philippines falls within this genre of policy oriented research. After World War II, decolonization, the Cold War politics, and development optimism fostered further policy oriented research in order to contribute to the strategic interests of government and business circles. This research challenged the previous, more conservative work of colonial scholars. At the same time a 'new' generation of Western scholars representing the humanities, particularly anthropology, sociology, history, and the social sciences—political science, economics—ventured into cultural contexts, interactions, power relations, and the state, thus challenging the applicability of overarching grand theories. These scholars fostered interdisciplinary programs and trained a generation of Southeast Asianists. With their legacy of regional, comparative, and disciplinary specialisms they not only sustain interest in the area but also are training a new generation of Southeast Asianists. Their focus has also changed with the redefinition of the region through the expansion of ASEAN and an increasing proliferation of meetings and organizations.

2. *Research by Southeast Asianists: Early Directions*

In the 1960s and 1970s, most Southeast Asianists focused on country studies or sub-regional unit studies. The majority of scholars concentrated on historical research/empirical work to constitute knowledge. History, anthropology, sociology, and

political science represented most of the intellectual streams in the field. The principal institutional setting for these disciplines was the universities that trained students in the humanities and the social sciences, and were the sites of research. Politicians and university administrators, encouraged by nationalist sentiment, fostered awareness of national/Southeast Asian studies, and encouraged the development of research interests on the region. The stronger the identity of the Western 'other' in the mind of the Southeast Asianists, the more inclusive became the notion of Southeast Asia. In turn, as these area studies programs attracted students who believed that their employment prospects were enhanced, so were further research interests in Southeast Asian studies developed. A wide range of professional associations was also set up to sustain the humanities and social sciences.

Historians were in the forefront of change. There was a two-stage move away from colonial and political emphasis to economic and then to social and cultural structures. This in turn brought about a discernible realignment from the social sciences towards the humanities. Historical scholarship was not confined to institutional settings and the discipline was open to contributions from other fields. Apart from national historical associations, state historical societies were also formed and there were cultural organizations like museums, which provided vitality to the discipline. Nevertheless, the encouragement of local history in some cases eroded any systematic attention to historical skills.

3. *Principal Institutional Settings and International Networks*

While university departments offered undergraduate teaching, postgraduate training was principally undertaken in the United Kingdom, the United States, or Australia. This was contemporaneous with two major developments. The first was the emergence of Southeast Asian area studies as the most dynamic branch of area studies in the 1960s and 1970s. The region's increasingly important role in view of the actual and potential Cold War conflicts in the area not only focused media attention on the region but also was assisted by support from two main sources. The first was support from Southeast Asian governments, which established Southeast Asian programs or institutes for policy related research pertaining to self-determination struggles, governance, and state policies that promoted political stability in diverse and plural societies. Support also came from the United States, the United Kingdom, and Australia. These countries not only promoted and established Southeast Asian studies centers as a field of study in their own right but also provided postgraduate training for Southeast Asianists through fellowship schemes. They also served as a source of Western academics that came to the region for teaching and research stints. It must be stressed however, that for the latter, Southeast Asia was still largely viewed from the periphery of Europe, the US, and Australia.

In the US, the Rockefeller Foundation made a grant to Cornell University to establish a center in 1950. Between the late 1950s and early 1970s, there was increased US funding for Southeast Asian study programs at Yale, Michigan, Northern Illinois, and Wisconsin-Madison. Columbia expanded its Southern Asian Institute to include Southeast Asian Studies. In the UK, the School of Oriental and African Studies (established 1938), was followed by a Southeast Asian studies center at Hull and Kent. In Australia, Monash University under the direction of John Legge, became a leading center of Southeast Asian studies, followed later by the Australian National University and others. These centers complemented and networked with the centers set up in Southeast Asia (see section 4 below).

4. *The 'Indigensization' of Southeast Asian Studies*

The formation of ASEAN in 1967, the impact of the Vietnam War, and the withdrawal of the US from Vietnam bolstered new interdisciplinary studies. Subsequently, at ASEAN's first summit meeting in 1976 the promotion of Southeast Asian Studies and the initiation of centers reflected a growing awareness of a wide variety of cross-border commonalities and the need to make greater efforts to understand neighboring countries and societies. The more challenging fields of study were cultural studies, gender studies, and indigenous and postcolonial studies, all of which had their own problems while at the same time, relying on the older disciplinary bodies of knowledge. The courses taught were discipline-based and the scholars regarded as discipline-based social scientists or humanists. They were only regarded as Southeast Asianists outside their countries. Moreover, although many or most of the Southeast Asian universities established courses on Southeast Asia studies, only in Malaysia and the Philippines did these institutions award both undergraduate and graduate degrees in Southeast Asian Studies [during the 1990s in Thailand as well]. Interestingly too, the only 'other' languages widely studied were those of Europe, especially English. Only in Malaysia and Singapore were other Southeast Asian and Asian languages offered to students. In the department of Southeast Asian Studies at the University of Malaya, students are required to take a Southeast Asian language, other than Malay.

The rationale for the promotion and funding of Southeast Asian area studies by most countries in the

region stemmed from the need to contribute to better understanding within the region; generate materials on the region which emphasize cross-border commonalities and shared interests, and establish specializations and the teaching of Southeast Asian languages. Funding for these activities came mainly from the state with additional international funding.

5. *Changing the Balance: The 1980s*

The economic ascendancy of Japan, the rise of the East Asian 'Tigers,' and the greater integration of Southeast Asia into the East Asian region led to changes in the direction of research and emphasis. The 'new 'concerns—industrialization, trade, investment, the sociology of production, and the disciplines of economics and political economy—became the new and international 'focus' areas. In terms of area studies, Japan and East Asia became the new paradigms. The commonality of East Asia's economic success, particularly for the countries that benefited most from it, promoted an awareness of shared values (Asian values), as distinct from 'Western values' and the construction of an Asian identity. Coincidentally, Japan (and to a lesser extent South Korea), became the new 'funding ' players, promoting not only Southeast Asian Studies but also Japanese and Korean studies.

At the same time, social science paradigms developed principally in the West were being modified or even rejected within the region. By the end of the 1980s, most Southeast Asians, in common with other Asians, began to perceive themselves and their countries not as objects but as subjects of study.

Three other concerns stand out during this period: labor, women, and the environment. Their study on a regional basis and through regional projects was promoted by international agencies such as the World Bank, the United Nations Environmental Program, and the United Nations Development Program. Large amounts of money were poured into the region, principally to established centers like the Institute of Southeast Asian Studies, Singapore. This institute has become a major postgraduate studies center for Southeast Asian Studies in the region. It has traditionally had a very active publishing program, its own publishing house for books, and publishes journals, of which Contemporary Southeast Asia and Southeast Asian Affairs is very well known. The new programs of the Institute include the regional economic studies program, which focuses on economic and related issues of the Asia-Pacific Economic Cooperation (APEC) forum, with a special focus on ASEAN, and the East Asian Development Network. Other regional projects included the East Asian Caucus, the expansion of ASEAN membership, and the coordination of Asian positions at the Asia-Europe meetings.

Southeast Asian Studies as area studies programs also advanced with the establishment of the Southeast Asian program at the National University of Singapore and the enlargement of the Southeast Asian Studies program at the University of Malaya into a Department that awarded degrees. In the meantime, in Thailand the Institute of Asian Studies at Chulalongkorn University (established in 1967) also began to focus on Southeast Asian Studies. Thammasart University established an Institute of Southeast Asian Studies in 1986, while the Arts faculty established a Bachelors degree in Southeast Asian studies in the late 1990s. The latter also initiated an important Foundation for the Promotion of Social Sciences and Humanities Textbooks Project. This foundation has produced a large number of books, in Thai, on Southeast Asia. The Institute of East Asian Studies (IEAS) established in 1998 at University Malaysia Sarawak, claims to be the first of its kind in the ASEAN region. Its role 'is to promote a range of interdisciplinary programs and activities to advance a better understanding of the East Asian region.'

The pride of being 'Asian,' sometimes expressed in anti-Western policies such as the Look East Policy, also led to a greater focus on Islamic identity and Islamic values, especially in Malaysia, Indonesia, the Philippines, and Thailand. Since the late 1970s, the region has experienced an unprecedented religious resurgence. The expansion of religious schools, the growth of a market in Islamic books, magazines and newspapers, and the rise of a well-educated Muslim middle class have played a role in the development of Islamic studies and Muslim discourse. Consequently, although Islamic studies in the region had earlier been regarded as at the intellectual periphery of the Islamic world, in the 1980s a systematic understanding of Islam and Islamic civilization has made Islamic Studies an ascendant field of area studies. This has been contemporaneous with increased relations with the Middle East and the Muslim bloc. In Malaysia, where Islam is the state religion, but where Muslims form a small majority only, the state has actively promoted Islamic area studies. In common with other countries in the region, pluralism, intellectualism, and openness to dialog with other faiths and institutions mark the study of Islam. The Malaysian government has also promoted Islamic studies through the establishment of an International Islamic University, an international Institute of Islamic Thought and Civilization (ISTAC), and a Center for Civilization Dialogue. The Islamic University, established in 1983, attracts scholars from the Asian region and elsewhere. ISTAC, which was established in 1991, is a research and postgraduate institution, affiliated to the International Islamic University and offers courses in Islamic thought, civilization, and science. The Center for Civilization Dialogue, which is based at the University of Malaya, was formed to encourage cross-cultural dialog and promote harmonious relations in Malaysia.

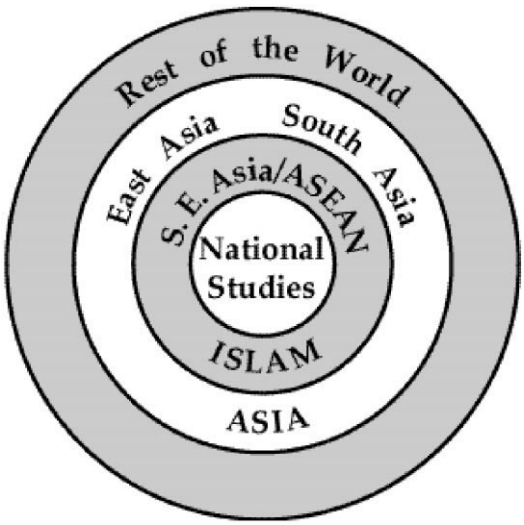

Figure 1
Southeast Asia: area and international

Islamic studies therefore form a core research activity in Malaysia and Indonesia with participation from key neighboring countries.

6. *The 1990s, Diversification, and New Research Directions*

The 1990s started with a continuation of the discourse on 'Asianess,' 'Asian values,' and an 'Asian century,' all of which lost overall relevance when the Asian miracle turned to meltdown. National values and national ideologies began to be voiced again, as some countries sought to distance themselves from others. In regional terms, especially over East Timor and Myanmar, there was also a lack of unanimity. Nevertheless, the Southeast/Asian identity in the wider world continued, especially in relations with East Asia, the US, and Europe.

Many of the new directions were in Women's and Gender Studies, Migration Studies, and Environmental Studies and there is not a single discipline in the humanities and social sciences that has remained untouched by these topics. There are many new conjunctions, especially with regard to Women's and Gender Studies, encompassing critical race theory, postcolonial theory, multiculturalism, and cultural, political, and social theory. The tremendous growth in intra-Asian labor migration associated with labor shortages/labor surpluses in the region also provided a new focus of area studies: migration studies. Not surprisingly, the Philippines, which relies heavily on short-term contract labor and remittances, took the lead in promoting migration studies. The Scalabrini Migration Center in the Philippines, which is a branch of the Federation of Centers for Migration Studies, became the focus of migration studies and publishes two quarterlies, *Asian Migrant* and the *Asian and Pacific Migration Journal*. Migration studies programs were also initiated in Malaysia, Thailand, and Indonesia.

After the middle of 1997, the region witnessed momentous and tragic events that led to a substantial decline in the living standards of its inhabitants. Consequently, just as in the 1980s, there was a great emphasis on explaining why the second tier newly-industrializing countries grew so fast, the major preoccupation of economists and political economists was on understanding how and why these economies fell off their high growth trajectories. Since this is not the kind of crisis to which governments and international agencies were accustomed, there was a substantial shift to understanding the crisis. The ASEAN Inter-University Seminar Series, which was launched in 1993, had as its theme Social Development: Post Crisis Southeast Asia for the 2001 Seminar. Two of the key panels included Southeast Asian Families: Surviving the Crisis and The Political Economy of Crisis and Response. The Series focused on common pursuits in the exploration of social issues, with an emphasis on collaboration, mutual understanding, and regional cooperation.

While national and regional studies continue to expand, ASEAN has now found it imperative to gain a knowledge and appreciation of other countries. The changing emphasis is reflected in Fig. 1 below National studies are placed at the 'core' of the Sotheast Asian world and other area studies are positioned relative to their importance to Southeast Asia.This is in keeping with the region's growing ties with the European Community and ASEM initiatives. At ASEM I in 1996, Malaysia took the lead to establish an Asia-Europe University (AEI) for European studies. The AEI is aimed at 'studying the diversities of both Asia and Europe and the forces of integration at work in both regions.' New research directions include an enhancement of Asia-Europe relations and partnerships through dialog and cooperation in the field of higher education, and an advancement of European studies. The Malaysian government funds the AEI, with support from other ASEM partners, international institutions in Europe, and the corporate sector in Asia and Europe. It is sited at the University of Malaya in Kuala Lumpur, and the key topics in this 'new' area studies are policy-related research on Asia, Asian business, European studies, European business, and e-commerce.

Thus while Southeast Asian studies continue to be emphasized, the region is finding it imperative to retain a knowledge and appreciation of other countries in Asia, Europe, the US, and the rest of the world, in that order. This will make Southeast Asian area studies remain competitive and holistic and overcome regional parochialism.

7. Future Directions

The interconnections between area studies and global concerns is shown in the growth of various studies–gender relations, labor and environmental standards and regionalism and trading blocs. This focus on global issues is set to continue in the research agendas of Southeast Asian studies in the third millenium. Consequently, Southeast Asian studies will continue to involve an understanding of the comparative nature of change in the global economy, and it is this comparative aspect that gives the field its intellectual breadth and vigor.

See also: Southeast Asia: Sociocultural Aspects; Southeast Asian Studies: Economics; Southeast Asian Studies: Geography; Southeast Asian Studies: Politics; Southeast Asian Studies: Society

Bibliography

Abraham I (ed.) 1999 *Weighing the Balance: Southeast Asian studies. Ten years after*. Social Science Research Council, New York

Asian Studies Association of Australia 1999 Symposium on Asian Studies in Asia. *Asian Studies Review* **23**(2): 141–203

Halib M, Huxley T (eds.) 1996 *An Introduction to Southeast Asian Studies*. Taurus Academic Studies, London

Hirschman C, Keyes C F, Hutterer K (eds.) 1992 *Southeast Asian Studies in the Balance: Reflections from America*. The Association for Asian Studies, Ann Arbor, MI

Morris-Suzuki T 2000 Approaching Asia from Asia. *Southeast Asian Studies Bulletin* 1/00: 19–21, 28

Shin Y H, Oh M S 1998 Southeast Asian Studies Overseas: A survey of recent trends. In: Kwon T H, Oh M S (eds.) *Asian Studies in the Age of Globalisation*, Seoul National University Press, Seoul, South Korea, pp. 173–92

Withaya Sucharithanrugse 1998 An insider's view of Southeast Asian Studies. In: Kwon T H, Oh M S (eds.) *Asian Studies in the Age of Globalisation*, Seoul National University Press, Seoul, South Korea, pp. 192–208

A. Kaur

Area and International Studies: Economics

At least since Adam Smith's *Wealth of Nations* (1776) the economics discipline has been concerned mainly with explaining the production and distribution of goods and services through the spontaneous interaction of self-interested demanders and suppliers. Economic 'efficiency' in a market system came to be understood as the condition in which land, labor and capital were allocated so that the largest possible output was obtained of those goods and services desired by consumers. Improvement in efficiency is achieved by extending the market geographically and obtaining thereby both greater competition among buyers and sellers, and reductions in costs of production through specialization and economies of scale. All impediments to extensions of the market, such as barriers to mobility of goods and services and of productive inputs, came to be viewed with suspicion or regret by economists. Nation states with distinct cultures, languages, and bodies of law were clearly sources of potential economic immobility, and economists often arrayed themselves against nationalism and in favor of a cosmopolitan world.

In the nineteenth century the 'marginal revolution' in economic theory formalized, and in some respects simplified, economic thinking; it presumed that all economic agents, regardless of their location on the globe, employed a utilitarian calculus. This undermined the case still more for economists to engage in the sympathetic study of the distinctive or unique characteristics of individual nations and peoples. By the start of the twentieth century, then, economists for the most part held to a universalistic ideology that saw in too much attention to what would become known later as international and area studies the danger that a case would be constructed for intervention in the world economy and an excessive role for the state. Added to this danger was the threat that too sharp a focus on the details of national behavior might lead to disturbing questions about the behavioral postulates of economic agents upon which the theoretical system of economic science now depended.

Nevertheless several special circumstances from time to time drew the attention of economists to area and international studies. These circumstances may be grouped roughly into three categories: first, the existence of empires and the challenge of achieving economic development in poor countries; second, international conflict and the need to understand peculiarities of the 'economic systems' of friend and foe; and finally, the structure of the international economic order, taking into account not only the microeconomic efficiencies that could be achieved therein, but also an increasing degree of macroeconomic interdependence.

1. Empire and Economic Development

In contrast to the Mercantilist Writers who preceded them, Smith and some of the Classical economists suggested that colonies, including those in North America, were likely to cost more than they were worth, and it was better to let them become independent trading partners, in which case there was no particular reason to study them more than any other nations. Regardless of this advice, the major nations of Europe, and even to a degree the USA, 'scrambled' for colonies during the nineteenth and early twentieth centuries. Many prominent economists such as John Stuart Mill and John Maynard Keynes became directly involved in colonial administration and neces-

sarily deepened their knowledge of the dependencies in the process. Some later classical economists, such as Edward Gibbon Wakefield, extolled the economic benefits to both colonizers and colonists of the imperial links; others such as Karl Marx and John R. Hobson saw empire as part of a predatory process into which industrialized countries were inevitably drawn by the unstable nature of their own economic systems, but on both sides of this debate over the value of empire there seemed little reason to examine the peculiarities of the colonies themselves, and arguments were conducted mainly at the level of high theory.

During the nineteenth century, the one notable exception to the general inattention by economists to cultural and societal distinctions was the German Historical School which held, among other things, that the postulation of theory should follow inductive inquiry, and that economic development typically proceeded through successive stages, for each of which different public policies were likely to be appropriate. For example, they asserted that there should be no unqualified endorsement of free trade and prohibition against an active state in the economy. Each case required its own analysis, and in certain circumstances even tariff protection—anathema to most free-market economists—might be appropriate. Friedrich List (1827) described the success of protectionist American development policies after residence in the country. Other Germans—graduate students as well as senior scholars—fanned out around the world to gather information about foreign countries in a way that was never fashionable among British classical political economists.

When empires largely collapsed after World War II, attention among economists shifted toward the developmental challenges of the newly decolonized nations. This attention was especially intense in the USA because of the strong sense of an important new world role for the country. To a degree never attempted before, economists in those years reached out to other disciplines to help explain reasons for stagnation, and to help design policies for the future. To some extent the new field of 'Economic Development' in economics, as it was called, reached back to the interdisciplinarity and multinational orientation of the German Historical School, but now filtered through the distinctive features of American Institutionalist Economics, found in the works of Thorstein Veblen, Wesley Clair Mitchell, and John R. Commons. In teaching and research units devoted to development studies such as those at Yale, Stanford, the University of Wisconsin, and the University of Sussex, and through such periodicals as the *Journal of Development Studies* and *Economic Development and Cultural Change*, economists joined with other disciplines, and especially the other social sciences, to explore developmental strategies and especially innovative roles for the state. Enthusiasm for 'economic planning' in this literature was reminiscent of the excitement felt for the Tennessee Valley Authority in the 1930s. Often leaders in the new development economics, although much honored as individuals, moved outside the mainstream of the discipline. Gunner Myrdal and John Kenneth Galbraith are prominent examples. In some parts of national governments and international organizations, such as the Economic Commission for Latin America under the leadership of Raul Prebisch, there was even wholesale rejection of the free-market model wherein, it was charged, each nation was compelled to accept the economic activities meted out to it by the principle of comparative advantage. This model was pictured as no more than a rhetorical device used by the industrialized nations to keep poorer nations as hewers of wood and drawers of water. Instead, it was said, each nation should select its own development path and seek to limit external dependence by producing its own substitutes for imports (Prebisch 1950, Singer 1950). In that it was predominantly Latin American, and later African, economies that followed this advice—as opposed to Asian economies that focused on expanding manufactured exports—the Prebisch–Singer dependency theory fostered a regionally-defined character of inquiry into the particulars of developing economies.

For the first few decades after World War II the subdiscipline of Economic Development prospered and interacted in complex ways with other disciplines and with international and area studies programs on many campuses and in many countries. Within the discipline, a handful of scholars such as Hirschman (1958) emphasized the importance of understanding the different structural issues of the developing world and advocated country-specific inquiry. Economists became engaged with the details and particularities of the societies of other countries as never before, and often they were involved with the planning and implementation of the policies they held under review. Beginning in the 1970s, however, the subdiscipline came under stress. The tension between its operating assumptions and those of the disciplinary core of economics became obvious and intolerable to some disciplinary leaders. The development economists accepted complex behavioral postulates that were culturally determined, at least in part, in place of profit and utility maximization, and they proposed policy norms that substituted values like independence and cultural survival for consumer sovereignty and market freedom. Perhaps nowhere was the debate as sharp as in research on rural and agricultural development, where the 'efficient peasant' school (Schultz 1964) stood in stark contrast to the view that poor farmers were driven by cultural constraints and 'survival algorithms' (Lipton 1968).

The research style of the subdiscipline of development was also discovered to be very different from that of the 'core,' demanding extensive field work and personal identification with the subjects under study, as well as resort to other disciplines such as anthro-

pology and history. The development economists did not depend upon abstract modeling and empirical testing to the extent that this was prescribed in the canonical methodological essays of the time, such as those by Lionel Robbins (1932) and Milton Friedman (1953).

The methodological debate was rendered moot for many observers of the world economy—within and outside the economics discipline—in the 1970s and 1980s, as the evidence mounted that free markets and free trade had demonstrated conclusively their virtues as the best path to successful economic development; central planning and other forms of public intervention lay discredited. Old-style development studies therefore became associated with the losers in this appraisal of the development race. With the evident failure of the import-substitution strategies of Latin America, as against the success of the export-led growth of east and south-east Asia, it seemed to many observers that the special field of economic development had been a pied piper, and that those poorer countries which had followed the mainstream of the economics discipline, where country-specific knowledge was not required, were better off. With the decline in fashion of old-style development studies in economics there followed a reduction in support for the subdiscipline from charitable foundations, governments, and international organizations, and a decline also of interest among students.

By the 1990s, economic development as a distinct subfield of economics was in many places on the wane, or was moving away from its former distinctive commitment to an area studies approach toward simply the application of the conventional tools of applied mainstream economics to problems of poorer countries. Developing-area data sets were widely supplied by international organizations, but there remained little sense of a need for analytical tools uniquely appropriate to conditions of the 'developing world.'

As consensus built during the 1980s and 1990s on the strength of free-market policies, more and more less-developed countries tried to follow in the path of the Asian economies (some under the express direction of the World Bank and International Monetary Fund) and liberalize their markets. Most of these efforts were profoundly disappointing in terms of economic growth. Evidently, the strict neoclassical paradigm was not a panacea, but neither were the interventionist models it replaced. The search went on for a new explanation, and many focused their attention on political and social institutions and their role in facilitating or hindering economic growth. If now-discredited development studies had an ideological connection to the original American school of Institutionalist Economics, the New Institutional Economics offered a neoclassically credible approach with the study of region- and country-specific structures. Douglas North (1990) and Oliver Williamson (1996), among others, developed an evolutionary approach to the origins of legal and commercial institutions, an approach that resonated with economic historians and those in the field of economic development who remained committed to the differences in, rather than universality of, the countries they studied.

2. *International Conflict and Comparative Economic Systems*

After the problems of economic development in poor countries, international conflict has been the second major stimulus for economists to attend to international and area studies. From the earliest days of the discipline, economists joined their countrymen in speculating about the behavior and motives of their traditional opponents: the British about the Spanish and then the Germans, the Americans about the British, the Canadians about the Americans, and so on. Such speculation required some familiarity with the cultural peculiarities of the antagonists. By the 1930s, traditionally free-market economists concluded that at least two alternative types of economic system had emerged that were a continuing political and economic threat to the very idea of a free society and a competitive market economy. These were represented by the fascist dictatorships of Hitler, Mussolini, and Franco, and by the Stalinist planned economy of the Soviet Union. Attention by economists to these alternative economic systems was strengthened, first by the sense that the world economy had become so interrelated that difficulties in one part, like a recession, or selfish actions in another part, like imposition of a tariff, could injure all. This is partly the reason why John Maynard Keynes paid so much attention to economic development in Germany and the USSR. The second reason for attention to these authoritarian experiments was because they were perceived, correctly, as direct challenges to the general applicability of the competitive market model. Advocates of totalitarian economies claimed that authoritative decision makers could do jobs better than could the free market. The systemic extremes were in a sense locked in a battle for the minds of the modern economic man.

As soon as World War II began, the new economics subdiscipline of 'comparative economic systems' emerged formally as a way for governments of the allied nations to better understand the behavior of both enemies (Germany, Italy and Japan) and allies (the Soviet Union). At the start the 'systems' were seen as divided roughly into two broad categories: authoritarian planned economies on one side and liberal democratic economies on the other, but in the decades after World War II a more fine-grained categorization was attempted between the two extremes, to include across a political spectrum from left to right such

categories as non-communist planned (India and Tanzania), liberal welfare-state (Sweden and the UK) and modified communist (Yugoslavia). All these variants received some attention from economists specializing in economic systems and using an approach that involved area and international studies.

The decline of comparative systems as a subfield within economics parallels that of economic development and is tied up with the same historical and ideological trends. Over the course of the Cold War many economists came increasingly to conclude that all deviations from the competitive market norm were simply short-term aberrations and unworthy of serious scholarly attention. These, they believed, were the work mainly of selfish rent-seekers and bureaucrats and could not long survive in the face of the liberal alternative. Events of the 1980s, and especially the *annus mirabilis* 1989 that saw the collapse of the Soviet system, seemed to bear out this prophesy, but if 1989 brought an end to the study of comparative systems, it launched the (presumably temporary) subdiscipline of 'transition economics,' with its own necessarily regional focus. Economic systems specialists turned their attention to the evolution of post-planned economies toward the final steady state of democracy and free market capitalism (Kornai 1990, Boycko et al. 1995). Free-market transitions, like economic development, have proven both slow and difficult, suggesting that economists will remain a voice in post-Soviet studies for the foreseeable future.

With the declining political threat from the communist world and rising concern over the Middle East, it is perhaps not surprising that economists' interest in regional study has also been focused increasingly on the economics of the Muslim world. Islamic economics, with a focus on the Koran-prescribed rules of financial transactions, has emerged as the most recent comparative system for analysis (Khan and Mirakhor 1987, Kuran 1995). Reflecting perhaps the new interest in social institutions, much of the literature has approached the religious constraints on the free market not as an obvious inefficiency but, in the light of new attention to group-based lending and social capital, as potential corrections to the failures of free-market financial institutions.

3. *The International Economic Order*

A third stimulus for economists to pursue area and international studies was a growing appreciation that since the existence of nation states rendered impossible a truly competitive and unrestrained international economy, it was necessary to think about 'second best' alternatives. Up until World War I the informal Gold Standard seemed to impose a salutary degree of monetary and fiscal discipline on nations engaged in international trade and finance. After the war, however, attempts to recapture the benefits of the Gold Standard through the largely uncoordinated efforts of individual nations were mainly unsuccessful. By the later stages of World War II it was widely agreed, even among prominent free-market economists, that new institutional structures were needed to facilitate international trade and finance and, above all, to prevent the return of the Great Depression. Lord Keynes, Harry Dexter White, and other prominent economists who had risen to high levels in government, were at the heart of the planning at Bretton Woods that led to creation of the International Monetary Fund, the World Bank, and ultimately the General Agreement on Tariffs and Trade. Economists were given prominent roles in all of these, and in many other international organizations that came into being with some jurisdiction over the world economy.

It is hazardous to generalize about as complex phenomena as the place of the economist in the design, construction, and operation of the architecture of the international economic order that has emerged since World War II. But dominant characteristics that were present also in the approaches to economic development and economic systems discussed above, are easily visible here as well: an abiding faith in markets as a solution to economic problems, suspicion of special pleading as the basis for policymaking, skepticism of vigorous governmental action of any kind, and, wherever practicable, a preference for rules over authorities as a fundamental policy posture.

Despite near consensus among the economists on the optimal global architecture, progress toward these goals has been incremental and an ongoing area for research. Efforts to remove barriers to trade and financial integration accelerated in the 1980s and 1990s once developing economies lost whatever special status they could claim for strategies such as import substitution, which required tariff protection. Although much of the literature has been polarized between country-specific perspectives and global ones, recently two phenomena have emerged to refocus attention to regional problems.

First, there have been increasing efforts of governments to reduce trade barriers within regions, if not yet worldwide. Regional trade agreements such as NAFTA have raised concern over trade diversion while resuscitating the debate over regional integration. The prospective benefits of regional integration have been a boon, not just for trade theorists, but also for area specialists with an interest in the historical context as well as the implications for economic development (Collier and Gunning 1995).

Second, a series of regional financial crises, in Latin America in the mid-1990s and in Asia in 1997–8, coupled with plans for currency integration such as the EU have renewed interest in regional currency management and optimal currency zones. As with the with debate on trade, many of these issues have deep historical precursors, and economic historians—whose subdiscipline remains a haven for regional

expertise—have added contextual nuance to the monetary theorist's musings (Eichengreen 1996).

4. Conclusion

Economists have tended to focus their attention on increases in efficiency and competition that arise from the development of markets, and they view with suspicion arguments for intervention that privilege national borders or cultures. They have come to view economic agents as relatively homogeneous in their behavior across the globe. In their view, these agents respond differently mostly because they are faced with different institutions and different incentive structures. The New Institutional Economics has been instrumental in using the standard economist's toolkit to model these differences, allowing behaviors previously thought to be in the cultural domain to be formalized within the neoclassical paradigm.

These new domains may revive interest among economists in studies of areas, but it is unclear whether such interests will fall within the realm of area studies. The regional expertise required to legitimize regionally-oriented work within the economics discipline is of a different character than the expertise required in other disciplines. One notable feature of the debate on regional integration is how many economists are publishing on NAFTA, MERCOSUR, APEC, and the EU simultaneously. There is little incentive to invest in context-specific expertise at the expense of the universally-applicable tools, especially since the empirical record would argue against economists going outside the neoclassical paradigm, especially when offering policy advice. There is a fundamental rift between this view and the original vision of area studies, which has stressed nonuniversality and resisted the 'monoeconomics' standard. This ideological rift threatens to leave area studies without a voice from the economics discipline, while simultaneously removing the voice of regional expertise from the economic arena

See also: Dependency Theory; Development, Economics of; Development: Rural Development Strategies; Development: Social; Development: Socioeconomic Aspects; Economic Anthropology; Economic Sociology; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Financial Institutions in Economic Development; Imperialism: Political Aspects

Bibliography

Boycko M, Shleifer A, Vishny R 1995 *Privatizing Russia*. MIT Press, Cambridge, MA

Collier P, Gunning J W 1995 Trade policy and regional integration: Implications for the relations between Europe and Africa. *The World Economy* **18**(3): 387–410

Eichengreen B 1996 Deja vu all over again: Lessons from the Gold Standard for European monetary unification. In: Bayoumi T, Eichengreen B, Taylor M P (eds.) *Modern Perspectives on the Gold Standard*. Cambridge University Press, Cambridge, UK, pp. 365–87

Friedman M 1953 *Essays in Positive Economics*. University of Chicago Press, Chicago, IL

Hirschman A O 1958 *The Strategy of Economic Development*. Yale University Press, New Haven, CT

Khan M S, Mirakhor A (eds.) 1987 *Theoretical Studies in Islamic Banking and Finance*. Institute for Research and Islamic Studies, Houston, TX

Kornai J 1990 *The Road to a Free Economy, Shifting from a Socialist System: The Example of Hungary*. W. W. Norton, New York

Kuran T 1995 Islamic economics and the Islamic subeconomy. *Journal of Economic Perspectives* **9**(4): Fall: pp. 155–73

Lipton M 1968 The theory of the optimizing peasant. *Journal of Development Studies* 327–51

List F 1827 *Outlines of American Political Economy*. Samuel Parker, Philadelphia, PA

North P C 1990 *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, New York

Prebisch R 1950 *The Economic Development of Latin America and its Principal Problem*. United Nations, New York

Robbins L 1932 *An Essay on the Nature and Significance of Economic Science*. Macmillan, London

Schultz T W 1964 *Transforming Traditional Agriculture*. Yale University Press, New Haven, CT

Singer H 1950 The distribution of trade between investing and borrowing countries. *American Economic Review* **40**(2): 473–85

Smith A 1776 *An Inquiry into the Nature and Causes of the Wealth of Nations*. Whiteston, Dublin

Williamson O E 1996 *The Mechanisms of Government*. Oxford University Press, New York

C. D. Goodwin

Area and International Studies in the United States: Institutional Arrangements

Area and international studies refers to research and teaching about other countries. In the United States, until the 1940s, international studies were simply parts of academic disciplines. Over the past fifty years they have become more distinct, self-aware, and organized. Their growth has paralleled a more general trend toward the internationalization of American universities: the expansion of opportunities for study abroad by students and faculty; bringing more foreign students to the campus; staffing and managing technical assistance programs; developing formal overseas linkages including the establishment of satellite campuses abroad; and durable transnational links between faculty and students in the US and abroad. Here, only those aspects dealing with research and teaching about countries other than the US will be discussed.

As international studies have grown, they have split into a number of specialties which for purposes of

analysis will be divided into two segments; one, those that deal primarily with a single country or region, and the other, those that comprise studies of transnational phenomena. Two examples of studies focused on single countries or regions will be discussed: area studies and risk analysis. With respect to transnational research several overlapping specialties will be discussed: comparative studies, international relations and foreign policy analysis, security studies, peace studies and conflict resolution, and international political economy studies.

1. Studies of a Single Country or Region

1.1 Area Studies

1.1.1 Origins. Of all of the international studies specialties the developmental path of language and area studies is the easiest to identify. The seeds of area studies lay in the 'tiny band' of area-focused scholars who, before World War II, were scattered throughout the faculty of a few universities, primarily historians or scholars who studied the literature of great civilizations. During World War II, the anticipated manpower needs for intelligence and possible military government led to the establishment of a number of prototype area studies programs, Army Specialized Training Programs (ASTP), on 55 American college and university campuses, to train specialists on particular countries or areas. This early emphasis on (a) a perceived national need which was unmet by normal academic processes and (b) remedying a manpower shortage—other programs were established under ASTP for perceived shortages in mathematics, physics, electricity, and engineering—remain until today the core rationales for federal government support of language and area studies programs.

Since the ASTP programs were focused on preparing people to deal with current affairs, the social sciences and the currently spoken languages of the region were added to the pre-existing humanities base on the campuses to create a new academic model. In the initial programs the students were all enlisted in the military. After World War II area studies were gradually incorporated into the general educational mission of universities and colleges. By the 1990s almost half of the universities and colleges in the US offered curricular concentrations focused on one or another world area. The area studies model was also used in various government agencies for training foreign service officers and other personnel preparing for service abroad.

1.1.2 Funding. It was clear from the outset that special financing was needed to sustain and expand these programs. The normal organizational style of universities was inhospitable. The normal priorities of disciplinary departments made it difficult to assemble the requisite specialized faculty in a variety of disciplines. Student demand had to be developed from scratch, and funds were needed to support the added time that language learning and overseas doctoral research demanded. Moreover, expanding and cosmopolitanizing library collections to provide the basis for research and teaching about other countries was costly.

Over the course of the next several decades grants from a number of private foundations—primarily the Rockefeller, Ford, and Mellon Foundations, and the Carnegie Corporation—provided the special support that area studies programs needed. In 1958, private foundation funds were supplemented by funding from the federal government. Following the minor panic that resulted from the Russians' unanticipated launching of the satellite Sputnik, the US government established an annual support program for university-based language and area studies centers under Title VI of the National Defense Education Act (NDEA), later the Higher Education Act (HEA). This governmental funding program has continued uninterrupted for more than fifty years and has played a critical role in the maintenance of language and area programs. However, universities have provided the bulk of the support for area studies programs out of their own funds. Outside moneys rarely cover as much as ten percent of the full cost of programs.

Since a primary goal of area studies has been the expansion of a national cadre of experts it was essential that a steady flow of graduate students be recruited into the field. An early fellowship program—Foreign Area Fellowships—was introduced by the Ford Foundation to provide support for domestic and overseas education of area specialists. Over the years, this program has come to emphasize support for dissertation research overseas. Supplemented by support from other donors, it is now administered jointly by the Social Science Research Council and the American Council of Learned Societies, the primary overarching research organizations in the social sciences and humanities respectively. Subsequently, annual government support for study at area centers was provided as Foreign Language and Area Studies Fellowships under NDEA, Title VI. Support for dissertation research abroad was funded by the US Department of Education under a specifically earmarked section of the Fulbright–Hays Act, and, since 1991, through fellowships made available under the National Security Education Act, administered by the Department of Defense. Over the years, language and area studies centers have received a series of general support and endowment grants from the Ford and Mellon Foundations, as well as a variable flow of project money for individual research projects. In addition to general funding for area studies as a whole, supplementary

support has been made available for specific world areas.

1.1.3 Area Specialists. The term 'area specialist' refers to an individual, most often a scholar, who dedicates most of his or her research and teaching to a particular country or region, possesses a substantial amount of multidisciplinary erudition about that country or region, and is competent in one or more of the languages of the area. While the term can include individuals concentrating on any portion of the world, in practice it tends to be used primarily for those whose area of specialization is in the non-Western world (including Russia and East Europe), plus Latin America. Scandinavian studies is usually included as area studies, but European studies is not.

In the period immediately after World War II, before area studies centers were established, most area specialists were self-recruited and self-trained. Their tended to be recruited through prior overseas residence; in the case of developing societies this was often the Peace Corps. In recent decades, most new area experts have been the products of language and area centers—in universities in the case of academics, in government area training centers for the diplomatic service, the army, the navy, the marine corps, or the various intelligence services.

1.1.4 Organization of area studies education. Growing out of the model established by ASTP, and institutionalized in part by the terms of the annual competition for support under HEA, Title VI, the organizational and curricular style of language and area studies is relatively standardized. Campus-based area centers comprise a set of faculty members whose home affiliations lie in a variety of disciplinary departments. These centers do educate undergraduates who take a major or minor in the study of a particular world area. The area studies undergraduate major or minor usually requires that the student takes a spread of disciplinary courses focused on the area, plus a modern language of the area. However, while centers do provide for undergraduate education, their principal concern is with graduate students training to be specialists on the area. While such students almost always take their degrees in a particular discipline, for certification as an area specialist graduate students are required to spread their course work across a number of disciplines and acquire a high level of competence in a regional language. Moreover, there is an almost universal requirement that dissertation research must be carried out in the area. This supplemental layer of area-specific courses, the time required to gain an advanced competence in a language, and dissertation research abroad normally adds several years of course work to earn a PhD.

Centers vary greatly in their organizational form and degree of cohesion. A few centers are full academic departments, complete with their own faculty, staff, and students. More often, the faculty is scattered throughout the various disciplinary departments, but the center sets the curriculum for area studies majors and certifies degrees or certificates. They also maintain centrally held resources such as area-specific library collections, access to fellowships for domestic and overseas study, auxiliary teaching and support staff, and external programmatic and research funding.

Area studies are now spread throughout higher education in the USA. About half of American colleges and universities provide at least one identified concentration of courses on a country or world region. The major research universities have a number of large, well-organized centers, in some cases as many as six or seven, each one dealing with a different world area. Individual area centers vary in size from a handful of faculty members to very large centers which are staffed by as many as 50 faculty members spread over more than a dozen disciplines. The largest and most developed of the centers are awarded annual program support and a quota of graduate student fellowships under HEA, Title VI.

1.1.5 Organization of area research. While the centers play a key role in the organization of area studies, they tend not to be units of research collaboration. Research is carried out by individual scholars. Collaborative research, where it occurs, tends to link scholars across universities rather than within centers. Facilitating these transinstitutional links and serving as research facilitators and accumulators are national area studies membership organizations: the African Studies Association, the American Association for the Advancement of Slavic Studies, the Association for Asian Studies, the Latin American Studies Association, and the Middle East Studies Association. Individual scholars and students are also served by a series of more focused organizations that represent sub-sections of the constituency—e.g., the American Oriental Society representing scholars pursuing the textualist tradition in the Near and Far East. There are also a series of organizations promoting overseas research, some based in the United States, such as the International Research and Exchange Board (IREX) which awards fellowships for study in Russia, and some based abroad, such as the American Institute of Indian Studies whose headquarters are in New Delhi. Standing committees of the Social Science Research Council and the American Council of Learned Societies, supported largely by the Ford Foundation, annually allocate dissertation-level fellowships and sponsor centralized planning and assessment for each world area.

1.1.6 Geographic coverage. Over the years, the geographic domain covered by each world area, whose boundaries represent 'a residue of colonial cartography and European ideas of civilization,' has remained fairly constant. The established geographic units are Africa, Central Asia, East Asia, East Europe, Latin America, Middle East, South Asia, and Southeast Asia. Programs on smaller campuses may define particular areas more broadly. Scandinavian studies, Central Asian, and Oceanic studies constitute smaller area study groups. Scholars studying one or more Western European countries generally do not consider themselves part of the area studies communities, although West European studies have been added to the list of federally supported area groups.

Within each world area particular countries tend to receive the bulk of scholarly attention: China and Japan in East Asian studies, Mexico and Brazil in Latin American studies, India in South Asian studies, Egypt in Middle Eastern studies, Thailand and Indonesia in Southeast Asian studies, and the former Soviet Union, now Russia, in East European studies. More recently, there has been a tendency among students to direct their studies to other countries within their world area.

World area study groups tend to be almost totally discrete both on the campus and nationally. While there may be several area programs on particular campuses and some faculty members may belong to more than one area studies group, each group has its own organizational style, intellectual tradition, professional association, and scholarly journals. There are recurrent attempts to create new geographic units, e.g., the development of the Pacific Rim studies, diaspora studies, and research on the Muslim world as a whole. More recently, part of the rationale for the Ford Foundation's 'Crossing Borders' funding project is to encourage the linking together of several world area studies groups.

1.1.7 Disciplinary coverage. While area studies conceptually covers all the academic disciplines, they differ in their hospitality to area studies as an intellectual approach. The degree of hospitality is reflected in the representation of members of the various disciplines in area studies. History and language and literature, with their emphasis on substantive erudition, are the most fully represented in area studies, followed by political science and anthropology. Least hospitable to area studies are the 'hard' social sciences, those that emphasize strong methodology, quantitative methods, and abstract conceptualization and theorizing: economics, sociology, psychology, and a major portion of political science. The Ford Foundation, in a program begun in 1990, attempted to remedy these disparities by providing fellowships to allow students specializing in one of these disciplines to add area and language training to their education. It also allowed area studies students who were majoring in one of these four disciplines to add to their area course work advanced theoretical and methodological training in their discipline.

1.1.8 Languages. From the start, the federal government's interest in language and area studies has been primarily in the less commonly taught languages, including the training of specialists and the maintenance of a capacity to teach these languages on a variety of campuses. Of special interest are the least commonly taught languages, for instance, Albanian, Armenian, Azeri, Cebuano, Kaqchikel, Latvian, Somali, etc., for which enrollments are so small that no university is likely to offer instruction in them without outside support. The cost of teaching such a language may be as high as $14,000 per student per year, and if the language is taught at all, an institution will have to offer at least two and probably more years of instruction. Partly as a result of the federal government support program, many major research universities currently teach as many as forty languages other than English. Most of the non-European languages are taught in area studies programs.

The degree of language competence required of area specialists varies by world area. It is greatest among Latin American, East Asian, and East European specialists, and least among students of regions with a large number of languages and/or a widely used colonial language, such as South Asia or Africa. At the same time, the amount of time a student must devote to language training varies considerably by world area from very little for Latin American specialists who tend to bring into the program almost all of the language skills they will need to five years or more of language study for students who must master a non-cognate language like Japanese, Chinese, or Arabic. Several of the world area studies groups maintain a facility overseas for advanced training in languages.

1.2 Risk Analysis

Area studies is not the only intellectual format for the study of single countries and world regions. Much research and teaching still takes place within the established disciplines outside of the area studies community. Several streams of internationally-focused, intra-disciplinary research and teaching have coalesced into distinct specialties. These research traditions tend to develop a special theoretical and analytic approach and to use that approach to describe a particular geographic area. An example of such a stream is risk analysis, a branch of applied economics. It analyzes a specific set of quantifiable demo-

graphic, economic, and political variables to estimate the suitability of a particular country or region for investment. While the origins of much of the theory and analytic methodology of risk analysis was developed within the scholarly community—e.g., probabilistic analysis, game theory, information economics—the developers and users of risk analysis are primarily international practitioners such as commercial banks and multinational corporations.

2. *Transnational Studies*

Area studies and other analyses of a single country or region are not the only styles of internationally-focused research and teaching. A number of international studies specialties, some of them predating area studies, have developed that focus not just on one country or region, but are concerned with a number of countries or regions, or, in some cases, all geographic areas at the same time. As in the case of area studies, these intellectual specialties have tended to separate themselves from the general internationalization of disciplines. Moreover, while there is some overlap with area studies and among the transnational specialties themselves, they have developed into quite separate traditions of research and teaching. Here we can deal with only four styles of cross-country and cross-regional research and teaching: comparative analysis, studies of international relations and foreign policy, security studies, and conflict resolution, peace studies, and international political economy studies.

2.1 *Comparative Studies*

Unlike area studies and risk analysis which focus on a single country or region, in comparative studies a deliberate search is made for uniformities and differences across a number of geographic areas. Sometimes the United States is matched with another country. More commonly similarities and differences among a variety of other countries are examined. Early examples were the numerous essays on differences between East and West. Another common style is the tracing of a single phenomenon across a number of different areas, as in comparative studies of entrepreneurship, or Islam. In a third style of comparative studies a common conceptual and analytic framework is developed, then relevant data are collected in a variety of countries. An example of this approach is the study of comparative political development in which information on a substantial number of political systems, primarily in the third world, was assembled using a common descriptive and analytic format. All of these styles of comparative analysis dramatize one of the major tensions in international studies: the degree to which emphasis should be put on the particularities of individual cases or on bending country idiosyncracies to fit a common conceptual and analytic framework.

2.2 *International Relations and Foreign Policy Studies*

While comparative studies can analyze any or all features of countries or regions, other forms of transnational analysis tend to concentrate on a more limited range of topics. The oldest of such international studies specialities is the study of international relations. It is primarily concerned with relations among nation states, their foreign policies, and the operation of international organizations. On most campuses the study of international relations was initially a sub-division of political science. Gradually, free-standing academic departments of international relations developed on a few campuses. In a number of universities, the study of international relations expanded and crystallized into separate schools of international affairs whose primary mission was the education of MA level students who sought internationally oriented careers. There are now 13 members of the Association of Professional Schools of International Affairs (APSIA) which specialize in international relations. Several of them were established before World War II. In earlier years, these schools served as academic trainers for future foreign service employees. As the government developed its own training facilities the schools broadened their mandate to train for service with international organizations and multinational corporations. At the same time, their curricula expanded from a narrow focus on international relations to a broader range of international studies approaches, with a predominantly academic rather than applied professional orientation. Organizationally, the schools differ in the extent to which they are free-standing within the university with separate faculty and degrees. Several of them serve as administrative homes for the area studies centers on their campuses.

2.3 *Security Studies*

Within the study of international relations, a separate field of research and teaching is primarily concerned with security relations. It covers such topics as military policy, the cold war, and the management and use of nuclear and other high technology. In the early years, security studies analysts were primarily political scientists and physicists, both inside and outside of the government. Analyses tended to concentrate on US foreign policy, or on the international system as a whole. Its purpose was preeminently a practical one, to influence foreign policy decisions particularly with respect to the use of military power. A large part of this

research enterprise was carried out within the government or in academic centers with strong government ties.

In 1984 the Catherine T. McArthur Foundation funded a program through the Social Science Research Council whose goal was to transform the field of security studies. Through the awarding of grants for fellowships, conferences, and workshops, the Foundation and the Council's intention was to engage a broader academic community in security studies, including the diversification of research participants to include more young scholars, more women, more scholars from abroad, and to expand the range of disciplines among the participants. Similarly, the scope of security studies shifted from an exclusive concern with military security and technology to include such topics as environmental issues, nationalism and ethnicity, and the changing nature and role of the state in violent conflicts. Its academic home is in the International Studies Association rather than area studies associations. In fact, most area studies specialists deal with internal affairs in the countries they study rather than their external relations. Moreover, most security analysts prefer to work with entire regions, or with international systems as a whole rather than individual countries.

A substantial portion of security studies was carried out within governmental organizations or in research-oriented centers on campuses with strong links to government.

2.4 *Peace Studies and Conflict Resolution*

In part as a reaction to the perceived militarization of the outlook of security studies, two interrelated research specialties developed which were concerned not with strategies for winning conflicts but with their peaceful resolution or with the avoidance of conflict entirely. Peace studies, like security studies, are primarily concerned with a segment of international relations. War and peace are seen as a continuum in the relations of nation states. The peaceful avoidance or settlement of international disputes, including the role of international organization and international law, are of special interest. More recently, the scope of peace studies has broadened to include factors that are presumed to enhance what is referred to as 'positive peace': human rights, ecology, economic well-being, non-violence, peace movements.

The field of conflict resolution overlaps in part with peace studies in that it is also concerned with the settlement of international conflicts, primarily as case studies. However, its emphasis is on conflict as a more general process, and much of the analysis deals with conflicts that are not specifically international. The study of conflict resolution will also deal with such topics as marital conflict, intergroup relations, race and ethnic conflict, and labor relations.

Organizationally, the study of conflict resolution is carried out by individual scholars who sometimes organize an on-campus center which links together a set of on-going group research projects. It may also provide a curricular concentration leading to a degree or a certificate. Nationally, the field is served by a specialized journal and membership association.

2.5 *International Political Economy Studies*

In recent years another more amorphous specialty has emerged from the combined efforts of economics, political science, economics, and sociology. Referred to as international political studies (IPE) it includes research on the international political economy as a whole. It covers such topics as global capitalism, trade regimes, international commodity flows, transnational crime and policing, intergovernmental mechanisms, and the like.

2.6 *Funding*

Unlike area studies, there is no centralized source of federal government funding for transnational research and teaching. There have, in fact, been two attempts to create such a source. In 1967, the International Education Act which would have provided a general purpose fund for international education was enacted by Congress, but no money was ever appropriated for it. Similarly, under the aegis of the Association of American Universities, the Social Science Research Council and in 1989, 165 academic associations concerned with international education formed a Coalition for the Advancement of Foreign Languages and International Studies (CAFLIS) in an attempt to create a free-standing federally funded foundation to provide broad support for international education. Lack of consensus among the various international studies organizations and the unwillingness of the foreign language scholarly community to support the effort defeated this attempt. In the absence of such an overarching source of funds for international education, the strategy has been to expand the mandate of HEA, Title VI to cover a variety of enterprises beyond language and area studies. It now provides funds for international business programs, undergraduate international education, the cosmopolitanization of university library collections, the introduction of international studies into historically black institutions, research and development in foreign language instruction, and overseas research centers.

Non-governmental support for non-area-oriented international studies tends to be more narrowly targeted by topic and purpose, although on a number of campuses the Hewlett Foundation provided en-

dowment funds in support of international studies broadly defined. Most private foundation funding, however, has been for substantively defined, time-limited projects. In recent years, the leading private donors for international studies have been the Ford, Kellogg, MacArthur, Mellon and Rockefeller Foundations and the Pew Charitable Trusts.

3. Looking Ahead

The forces that will shape future developments within international studies are already apparent. Internally, the various sub-specialties are likely to have different trajectories. International relations will broaden its perspective beyond interstate relations to include international business and other aspects of the global society. Peace and conflict studies, security analysis and risk analysis will come to resemble the other temporary coalitions of scholarly interest in International Political Economy Studies. Comparative studies will lose its distinctiveness.

With respect to area studies, during most of its history area specialists have played an important role in linking the American academic world with events and scholarship in other countries. What made this role possible was the area specialists' combination of language competence, area knowledge, familiarity with forefront disciplinary scholarship both here and abroad, and access to the American academic world. It will be interesting to see whether this role becomes less important as the spread of English becomes more pervasive, as strong academic communities develop within the countries being studied, and as more members of those communities themselves integrate with the American general academic community and with the increasingly interlinked global world of scholarship. These changes are occurring at a different pace with respect to different world areas. European studies never developed an area studies perspective in part because this intermingling of American and European scholarship was already well advanced. Latin American studies is already well along in this loss of the traditional role for area specialists. At the other extreme, in East Asian studies the difficulty of mastering Asian languages and the importance of a knowledge of their cultures will inhibit the transnational homogenization of scholarship. African studies and Central Asian studies are at the very beginning of this cycle. Hanging over all of the components of international studies is uncertainty about the continuation of the external financial support that in the past has underwritten the special costs of international scholarship and, above all, the overseas sojourns of PhD students. It is clear that, in some form, international studies will continue to be a strong component of American scholarship in the behavioral and social sciences.

See also: Area and International Studies in the United States: Intellectual Trends; Area and International Studies in the United States: Stakeholders; Foreign Language Teaching and Learning; Foreign Policy Analysis; Human–Environment Relationship: Comparative Case Studies; International Research: Programs and Databases

Bibliography

Barash D P 1991 *Introduction to Peace Studies*. Wadsworth, Belmont, CA

Chandler A 1999 *Paying the Bill for International Education: Programs, Partners and Possibilities at the Millennium*. NAFSA, Association of International Educators, Washington, DC

Goheen R F 1987 *Education in US Schools of International Affairs*. Princeton University, Princeton, NJ

Lambert D 1984 *Beyond Growth: The Next Stage in Language and Area Studies*. Association of American Universities, Washington, DC

Lambert D 1989 *International Studies and the Undergraduate*. American Council on Education, Washington, DC

Perkins J A 1979 *Strength Through Wisdom, A Critique of US Capability: A Report to the President from the President's Commission on Foreign Languages and International Studies*. United States Government Printing Office, Washington, DC

Sjoberg R L (ed.) *Country and Risk Analysis*. Routledge, London

Worschel S, Simpson J A (eds.) *Conflict Between Peoples and Groups*. Nelson-Hall, Chicago, IL

R. D. Lambert

Area and International Studies in the United States: Intellectual Trends

The fundamental role of area studies in the United States has been to deparochialize US- and Eurocentric visions of the world in the social sciences and humanities, among policy makers, and the public at large. Within the university, area studies scholarship attempts to document the nature, logic, and theoretical implications of the distinctive social and cultural forms, values, expressions, structures, and dynamics that shape the societies and nations beyond Europe and the United States. The broad goals are (a) to generate new knowledge for both its intrinsic and practical value, and (b) to contextualize and denaturalize the universalizing formulations of the social sciences and humanities which continue to draw largely on US and European experience. When successful, area studies research and teaching demonstrates the limitations of analyses of other societies,

based largely on the contingent histories, structures, and selective and often idealized narratives of 'the West.' More ambitiously, area studies can provide understandings of other societies in their own terms, and thus materials and ideas to construct more inclusive and effective tools for social and cultural analysis.

Area studies communities have not always succeeded in this; there have been false starts, dead ends, and other agendas as well. And area studies has evolved over time, and thus must itself be historicized and contextualized. Nevertheless, research and teaching on Africa, Asia, Latin America, the Middle East, and the Soviet Union has become a powerful social and intellectual invention. By generating new kinds of data, questions, and insights into social formations and cultural constructions that undermine received wisdom and established theories, by creating new interdisciplinary academic programs, and developing close collaborations with overseas colleagues rooted in different national and intellectual cultures, area studies scholars have challenged the social science and humanities disciplines to look beyond, and even to question and reconstruct, their initial origins and formulations.

These challenges have often involved sharp intellectual, institutional, and political struggles. Tensions and debates between area studies and the disciplines continue over intellectual issues, economic resources, and the structure of academic programs. To complicate matters, the various area studies fields are not at all homogeneous; there are striking differences in their political, institutional, and intellectual histories, and their relationships with the disciplines.

Area studies can be seen as a family of academic fields and activities with a common commitment to: intensive language study; in-depth field research in the local language(s); close attention to local histories, viewpoints, materials, and interpretations; testing, elaborating, critiquing, or developing grounded theory through detailed observation; and multidisciplinary approaches. Most area studies scholars concentrate their research and teaching on one or a small number of related countries, but try generally to contextualize their work in larger regions of the world (e.g., Africa, Latin America, Southeast Asia), beyond the USA and Western Europe. Those working on the three one-country area studies fields (China, Japan, and Korea), often engage in at least implicit comparisons among them, and often command literatures in languages from two or more of these countries. (Scholars with historical interests in Japan and Korea need to read Chinese. Likewise, serious scholars of China need to read the vast literature in Japanese.) The geopolitical boundaries of all the area studies fields—especially East Europe, Soviet, and Southeast Asia—are historically contingent, pragmatic, and highly contestable. The conventional boundaries have been intellectually generative, but they also clearly have limits.

1. *The Growth of Area Studies*

Prior to World War II, internationally-oriented teaching and research in US colleges and universities rarely went beyond European History and Literature, Classics, and Comparative Religion. At the start of the twenty-first century, thousands of college and university faculty regularly teach on the histories, cultures, contemporary affairs, and international relations of Africa, Asia, Latin America, the Middle East, and the former Soviet Union. Topical courses in the social sciences, humanities, and professional schools now use examples, readings, ideas, and cases from across the world.

Area studies has been institutionalized in US universities in (a) area studies departments, and (b) area studies centers, institutes, or programs. Area studies departments usually offer undergraduate degrees combining course work in the language, literature, history, religion, and sometimes the politics, of the particular region. In general, these departments are multidisciplinary but tilt heavily to the humanities. At the graduate level, area studies departments tend to concentrate on literature and history. In the 1940s and 1950s, these departments were regarded as crucial to training area specialists. However, by the 1960s the overwhelming majority of doctoral students specializing in the non-Western world were being trained and hired to teach in standard social science and humanities departments; anthropology, art history, geography, history, language and literature, music, political science, sociology. At the University of California, Berkeley, for example, since 1946, over 90 percent of the advanced degrees dealing with Southeast Asia were granted from the core disciplines or professional schools. Across the country, however, Psychology was always absent, and Economics has now essentially stopped producing area specialists. Through the early 1970s, small numbers of economists working on Third World development issues counted themselves area specialists. But as Third World development problems turned out to be more intractable than imagined, and as Economics has moved towards quantitative analyses and formal modeling, the subfield of development economics lost status and very few US economists now claim to be, train others to be, or indeed, engage intellectually with area studies scholars. Nevertheless, nearly all area studies faculty have at least double identities; for example, as a historian but also as a China scholar, as a sociologist but also as a Latin Americanist.

While area studies departments generally have declined in their centrality, area studies centers, institutes, and programs have grown dramatically. US universities now house over 500 such units focused on every region and all the major countries of the world. Only rarely are they formal teaching units, and they do not usually grant degrees. Instead, they draw in and on faculty and graduate students from across

the social sciences, humanities, and professional schools by organizing multidisciplinary lecture series, workshops, conferences, research and curriculum development projects, advanced language instruction, publication and library collection programs, and a wide variety of public outreach activities. By these various means, they often become active intellectual and programmatic focal points for both new and established scholars concerned with their particular area of the world.

Despite this dramatic growth, debates continue regarding the adequacy of international content in the curriculum; how it should relate to the undergraduate majors or advanced degree programs; the interests of diasporic populations in the student body; the relationships to 'multiculturalism'; and what, and how, foreign languages should be taught. Equally debated are the most valued topics, theories, intellectual perspectives, and methods for faculty and graduate student research.

The growth and worldwide coverage of area studies scholarship and teaching in the USA has no equivalent elsewhere. Many European universities have, or have had, centers or programs focused on their own colonial or ex-colonial possessions. Japanese and Australian universities have active centers concerned with their neighbors in East and Southeast Asia, but support relatively little scholarship on more distant regions such as the Middle East, Africa, or Latin America. Elsewhere, only a few universities have programs that go beyond their own world region—and the study of the USA or North America. While area studies is growing slowly in parts of Asia and the Middle East, only the USA has numerous universities with multiple area studies programs dealing with several regions of the world. Variously overlapping and competitive, jointly they provide global coverage.

Area studies in the USA began shortly before World War II, when small bands of scholars of Latin America and the Soviet Union joined forces to encourage increased research on those regions. During the war, a large percentage of the few US specialists on other regions became involved in intelligence work and helped to train officers for overseas commands and postwar occupation forces. After the war, some continued with the government but most returned to university life.

In the late 1940s, the Ford and Rockefeller Foundations, and the Carnegie Endowment convened a series of meetings among scholars and government officials in the belief that the Cold War and the prospects of decolonization in Africa and Asia would require the USA to play a vastly expanded role in world affairs. It was felt that the USA would need much more expertise than was currently available, and it would have to cover every region of the world. Such expertise, it was argued, would be needed by policy analysts, diplomats, and development workers, but also by society at large—in business, banking, the media, in primary and secondary education, in the foundation world, and by US personnel in international agencies. It was needed especially in higher education, where such expertise could be generated, mobilized, and directed toward overseas projects, but also be taught and disseminated broadly.

For most Americans at the time, the only familiar area of the world beyond the USA was Western Europe. (In 1951, an SSRC survey was able to identify only 55 people in US universities with expertise on any country in all of South and Southeast Asia. (Bennett 1951)). Most Americans had studied something of Europe in secondary school, some had traveled there, and some had recently fought there. Likewise, the vast majority of faculty and students in US universities came from families of European background. European institutions, politics, economies, cultures, and social formations were at least somewhat familiar from the media, and they were often similar to and sources for their US counterparts. Thus increasing US expertise on Europe did not seem to be of the highest priority. In contrast, US ignorance about the rest of the world was overwhelming. Furthermore, perceived challenges and threats from the Soviet Union, China, and communism generally, suggested that the USA needed internationally oriented economists and political scientists capable of constructing programs to encourage capitalist economic development, 'modernization,' and democracy in order to achieve social and political stability.

At the same time, at least some academic and foundation leaders were aware that the American and Eurocentric knowledge and experience of most US economists and political scientists might not be adequate for understanding the non-Western world. At least some felt the direct application of Western models and examples, and techniques in societies of very different character and history might not work at all. Economist, George Rosen (1985), provides a compelling account of the failures of MIT and Harvard economists in attempting to suggest or impose decontextualized economic development strategies in India and Pakistan during the 1950s and early 1960s. The USA seemed to need expertise in other fields to understand the structures and dynamics of other societies; their social organization, demography, social psychology, cultural and moral values, religious, philosophical and political orientations, economic potentials, international relations, etc.

The lead was taken on the campuses; indeed, the vast majority of support for area studies has always come from the universities through long-term investments in faculty, foreign language facilities, fellowships, libraries, research funding, etc. Nevertheless, external support has been crucial. A small Fulbright overseas teaching and exchange program had begun in 1946. But in 1950, The Ford Foundation established the large-scale Foreign Area Fellowship Program (FAFP), designed to create a much more sophisticated

and knowledgeable cadre of international scholars. FAFP awards provided a year of interdisciplinary and language training on a country or region of the world, plus two years' support for overseas dissertation research and write-up. By 1972, the FAFP had supported the training and research of some 2,050 doctoral students. That year, the FAFP was transferred to the interdisciplinary Area Studies Committees jointly sponsored by Social Science Research Council (SSRC) and the American Council of Learned Societies (ACLS). By the year 2000, with continuing Ford and other funding, the two Councils had provided over 5,000 more area studies dissertation fellowships and postdoctoral research grants. The foundations also provided several million dollars for area studies workshops, conferences, and publication programs at the two Councils and other similar institutions. Between 1951 and 1966, the Ford Foundation also provided $120 million to some 15 US research universities to establish interdisciplinary area studies centers. By 1999, the Foundation had invested on the order of $400 million in area studies training, research, and related programs (Beresford in Volkman 1999).

Although The Ford Foundation was the single most important source of private extra-university funding for the institutionalization of multidisciplinary area studies, other important funding programs followed. The post-sputnik National Defense Education Act of 1957 established the Department of Education's program that helps to support the administrative, language teaching, and public service (outreach) costs of some 125 university-based area studies centers, and the Fulbright Programs were much expanded in 1961. Likewise, the National Science Foundation and the National Endowment for the Humanities fund international research, workshops, conferences, exchanges, and related activities. Private foundations (e.g., Mellon, Henry Luce, Tinker), have also provided major support for area studies programs on particular countries or regions of the world. Still others (the Rockefeller Foundation, Carnegie Endowment for International Peace, the John D. and Katherine T. MacArthur Foundation), have both funded and drawn on area studies scholars for their own topically focused international programs. But it was the long-term, massive, and continuing support by The Ford Foundation at key research universities and through the SSRC/ACLS joint committees that established area studies as a powerful and academically legitimate approach to generating knowledge about the non-Western world. The Ford Foundation's 1997 $25 million 'Crossing Borders' initiative is the latest manifestation of this long-term commitment.

Although Cold War concerns were central to founding US area studies in the late 1940s and early 1950s, the scholars in the universities and SSRC/ACLS area studies committees quickly captured the initiative with broader academic agendas including the humanities, history, and other fields far from immediate political concerns. Indeed, from the 1960s, many area studies scholars criticized publicly the US government's definition of 'the national interest' and its policies and activities in the region of the world they were studying. This was most obvious in the Southeast Asia and other Asian fields during the Vietnam War. But numerous Latin Americanists had long been deeply critical of US policies towards Cuba, the Caribbean, and Latin America. And many South Asia scholars vigorously protested the US government's 'tilt towards Pakistan' during the Indo-Pakistan conflict of 1965. Numerous area studies scholars and organizations criticised US government-sponsored Third World 'development' and 'modernization' programs variously as ill-conceived, unworkable, counter-productive (if not simply counter-insurgency), self-serving, elite oriented, and of limited value to the poor of the countries they were claiming to aid. In effect, in varying but substantial degrees, all of the area studies fields quickly came to include a much wider range of political views and research agendas than their origins might have suggested.

Aside from expanding beyond initial political concerns, each of the area studies fields rapidly took on their own distinctive intellectual and research agendas, debates, and characteristics. US research on Southeast Asia began with heavy emphasis on political issues and the social sciences but then became heavily 'cultural' in orientation. In Latin American studies a variety of political economy frameworks have spiraled through the field influencing heavily political science, sociology, and theories of development generally, far beyond Latin America. From the 1960s to the end of the 1980s the key debates in Soviet Studies turned around whether the USSR could evolve towards more rational sociopolitical forms, or would necessarily degenerate.

African studies is marked by conflicting visions of Africa and divergent research agendas among mainstream (white) Africanist scholars, African-American scholars, and their counterparts in African universities. In contrast, while the South Asia field in the USA was built on nineteenth-century European humanistic studies of Sanskrit religion and philosophy, since the mid-1970s the intellectual life of the field has been redirected by the Subaltern movement and epistemological debate over the position of the scholar, and appropriate categories and subjects for the study of post-colonial societies.

But area studies in the USA has meant more than simply the addition of new research agendas or distinctive scholarly communities in US universities. By generating new data, new concepts, new approaches, and new units of analysis; by legitimating the intrinsic and analytic value of culturally rooted interpretations, and by creating new types of multidisciplinary academic units, area studies scholars have challenged intellectually and structurally, and to some

degree transformed, US universities and the established disciplines.

As Immanuel Wallerstein, as chair of the Gulbenkian Commission on the Restructuring of the Social Sciences, has pointed out (Open the Social Sciences, Stanford 1996) the current disciplinary division of labor in the social sciences was established in the late nineteenth century. A domain in the world, an intellectual discipline, and an academic department were seen to be mutually defining; the market required a discipline and department of economics, politics required a discipline and department of political science, society called for a discipline and department of sociology, etc. Legitimating each other, these hierarchically structured departments became the fundamental building blocks of US universities developing their own agendas, concepts, curricula, jargon, research methods, internal debates, specializations and subfields, journals, national organizations, and intellectual and interdepartmental hierarchies.

In this context, cross-disciplinary or multidisciplinary training and research were always difficult, and often denigrated. At the same time, recognition has grown that this nineteenth-century compartmentalization of the world reflected in twentieth- or twenty-first-century department does not fit current understandings of how societies and cultures actually operate. Not only are specializations internal to departments reducing their coherence, but it become obvious that the market, polity, society, culture, etc.—the domains that once justified current disciplinary boundaries—all penetrate, interact, and shape each other, and cannot be studied in isolation. Scholars now often seek out intellectual colleagues in other departments, and there are frequent calls for greater interdisciplinarity. Institutionally, even if internally riven, most departments remain sharply bounded based on the power to hire and recommend or deny tenure, buttressed by exclusionary discourses or jargons, and in competition with each other for university resources. The resulting tensions and contradictions, and the critiques they engender, have created a 'crisis in the disciplines' (Timothy Mitchell in Szanton in press) at least as problematic as the debates surrounding area studies.

Nevertheless, area studies units are not about to replace the disciplines, or even attain institutional equivalency. Short of an unlikely intellectual revolution and the reconstruction of the social science and humanities departments, area studies units and the disciplinary departments will continue to stand in productive tension with each other. At the same time, by demonstrating there are intellectually, politically, and socially important forms of knowledge, and legitimate modes of generating knowledge that require interdisciplinary collaboration which the traditional disciplines are unlikely to produce on their own, area studies has paved the way for the subsequent creation of women's studies, gender studies, African-American studies, ethnic studies, Asian-American studies, cultural studies, agrarian studies, and numerous other interdisciplinary centers and programs since the 1970s. In effect, area studies has legitimated a series of venues for cross-disciplinary research and debate recognized increasingly as essential for understanding the mutually constitutive elements of any society.

2. *The Critiques of Area Studies*

Despite the relative success of area studies centers in legitimating intellectual and organizational changes in US universities, area studies continue to be critiqued by scholars who define themselves solely, or largely, in terms of a disciplinary affiliation.

First, an unlikely combination of positivist and critical left scholars have charged that area studies was a politically motivated Cold War effort to 'know the enemy,' and with the collapse of the Soviet Union and the end of the Cold War, it is now obsolete. This critique is most frequent from within political science, a discipline currently taken with rational choice theories and also most directly affected by the political rivalries of the Cold War, its termination, and the transitions that have followed. But another version is also heard from the academic left, long opposed to US foreign policy and international activities, which has claimed that area studies has been largely a component of, and aid to, US hegemony, and opposed to progressive change elsewhere in the world. However, as noted above, while Cold War issues provided the major impetus for the development of area studies in the 1940s and 1950s, since that time all of the fields have in fact expanded intellectually and politically far beyond those initial concerns.

Second, others in the positivist tradition have charged that area studies has always been largely ideographic, merely concerned with description, as opposed to the nomothetic, or the theory building and generalizing character of the core social science disciplines. At its worst, this view sees area studies simply as generating exotica, which, however interesting, cannot add up to useful theories. At best, this view sees area studies as a source of data and information, fodder for more universal theorizing by scholars in the disciplines with broader vision, more sophisticated techniques, and greater intellectual skills.

In fact, there is little evidence that area studies research has been any less theory-driven than social science and humanistic research on the USA and Western Europe. Few social scientists or humanists ever propose grand new theoretical statements and proofs. Most, more modestly, see themselves as analyzing an interesting, or to them important, subject or object, in the process testing, critiquing, confirming, elaborating, or refining some presumed understandings or theories. This is equally true of scholars writing

on the politics of Bangkok or the politics of Washington DC, on Russian novels or US novels.

The issue is not the presence or absence of theory, but the kinds of theory being used, and how explicit or implicit, ambitious or modest, scholars are in articulating their theoretical assumptions and concerns. Here there is vast room for variation and debate as theories come and go, attract attention, are tried out against diverse data, materials, and concerns, and are then rejected, refined, celebrated, or absorbed into disciplinary (or common) knowledge. Not only have the area studies fields been thick with theory and theoretical debates, but frequently they have generated theoretical developments and debates within the disciplines. Nor should this be surprising, for, as previously noted, the vast majority of area studies scholars are institutionally located in the core social science and humanities departments. Privileging (or worse, universalizing) theory derived from narratives or analysis of US experience or phenomena alone overlooks the fact that the US is, although 'unmarked' by Americanists, as much a contingent, historically shaped and particular, if not peculiar, 'area' as China, or India, or Latin America. Indeed, on many dimensions, the USA is one of the more unusual and least 'representative' societies in the world—and thus a particularly poor case from which to build generalizing theory. In addition, Area Studies scholars working outside the USA usually recognize, at least implicitly, both the comparative value and the limits of their research arenas. In contrast, Americanists working on similar issues at home often seem to treat the USA as the 'natural' society, theorize, universalize, and advise others freely, and see no bounds to their findings.

A third and more subtle set of critiques of area studies scholars argues that they have absorbed and continue to use uncritically the politically biased categories, perspectives, and theories of their colonialist scholar-administrator predecessors—or indeed, of contemporary US or Western leaders attempting to maintain or expand hegemonic control over the rest of the world. The claim, dramatically put forward by Edward Said (1979), and echoed across subaltern and cultural studies generally, is that despite area studies scholars' evident personal interest in and specialized knowledge of the area of the world they are studying, the conceptualization of their projects, their research agendas, and what they have taken as relevant models of society and social change, remain fundamentally US- or Eurocentric.

In effect, there are two different charges here. One is that area studies scholars have sometimes or often failed to study other societies in their own terms, as social and cultural life and processes are experienced and might be construed, constructed, analyzed, and critiqued from the inside. The second is that they have failed to extract themselves from their conscious or unconscious political biases, and therefore have not framed their analyses adequately in some purportedly more universal theory, whether neoliberal, neo Marxist, postmodern, etc. Instead, area studies scholars are accused of—at best, naively, at worst, intentionally—imposing their own personal and/or national agendas and variously idealized formulations of the historical experience of 'the West,' both to explain, and often in the process, to denigrate, other societies that have almost always been, in one way of another, politically and economically subordinated.

These charges carry weight; political power and position and the generation of knowledge are inevitably entwined. But this hardly limited to area studies scholarship. All social scientists and humanists—both insiders and outsiders to a society—are influenced by their political context and commitments. Implicitly or explicitly, politically freighted categories and theories always shape how issues are framed, what kinds of question are raised, what equally valid questions are ignored, and who benefits from the research.

But the issue is more complex, because the current international economic and political hegemony of the US subtly, or not so subtly, encourages the notion that the research questions, assumptions, concepts, and procedures of US scholars are incontrovertible or irresistible. This can provoke deep resentments in other parts of the world. But it can also generate powerful alternative analytic approaches, for example, subaltern studies. Nevertheless, area studies scholars have one advantage in dealing with this problem. Intensive language- and history-based research conducted outside one's home country, in at least partially unfamiliar settings, is more conducive to a self-conscious recognition of these power issues than research carried out in the familiar USA.

A fourth critique of area studies derives from current fascination with 'globalization.' Although there is huge debate on how new it really is, how to define it, and how to study it, globalization (as financial, population, media, or cultural flows, as networks, 'deterritorialization,' etc.), is seen broadly as erasing boundaries, forcing the homogenization of localities, cultures, and social and economic practices. From this viewpoint, an area studies focus on the specificities or unique dynamics of particular localities is seen as being beside the point—an outdated concern for a world that at best is fading rapidly away.

In fact, globalization, however defined, when examined in particular places is rarely a homogenizing force, or erasing all other social or cultural forms and processes. Not only is globalization producing increased disparities in power and wealth—both nodes of rapid accumulation, and zones of exploitation and poverty—but its particular manifestations are always shaped by local histories, structures, and dynamics. Likewise, the recent growth and virulence of divisive ethnic movements and identity politics often seem both a consequence of, and a reaction to, elements of globalization. In this context the intensive multidisci-

plinary analysis of particular locations and areas—the hallmark of area studies—is even more essential.

In contrast, transnationalism is leading to more significant changes in the conceptualization and procedures of area studies. The geographic regions into which the area studies world was divided in the 1950s—South Asia, Sub-Saharan Africa, Latin America, etc.—were politically defined, and in cultural–historical terms often arbitrary and debatable. At the time of writing, these conventional categorizations are being questioned, and boundaries are being redrawn. Furthermore, recent attention to transnational diasporas is emphasizing the importance of new social and cultural formations cross-cutting previous nation-state and area boundaries. Likewise, Britain and France, as past centers of empires, have been—and continue to be—deeply shaped by their (ex-) colonial activities and subjects—and indeed are now becoming subjects of study by scholars in their ex-colonies. The analytic value of the older geopolitical area studies units has not disappeared completely, but the geographies of power are changing. Many boundaries have become more permeable, and the importance of sometimes new, sometimes longstanding, transnational social, economic, and cultural formations are increasingly being recognized, studied, and becoming the basis for new institutional support and organizational arrangements.

3. *Some Future Directions*

The new geopolitics and softening of area or national boundaries is being reflected in new area studies attention to population diasporas. Once one could study comfortably Southeast Asia, or, for example, the Philippines, Vietnam, or Laos as relatively bounded units. Today, there is growing recognition of the necessity of studying the flows of people from such areas or countries as they spread around the world. The intellectual reasons are several; diasporized populations have numerous feedback effects on the dynamics of their homelands. They drain educational investments, alter the age structure, and reduce population and sometimes political pressures. They also send back remittances, economic intelligence, political ideas, and entrepreneurial skills; and reshape the world views, opportunities, and networks of those remaining at home. Diasporas also often affect the political and diplomatic relations between their host country and original homeland. German relations with Turkey, US relations with Cuba, Chinese relations with Indonesia, etc., are all affected by the immigrant populations from those countries. In addition, in the new setting of a host country, immigrant communities may reveal previously unremarked elements of their homeland—or of the host society and culture. Thus new African and Middle Eastern populations in Sweden have brought out previously unnoticed degrees of racism in that country (Pred 2000). In the USA, the children of historic and current diasporas constitute increasingly large proportions of college and university students. As such, they are demanding new courses on the language, culture, and history of their ex-homelands, on their own diaspora, and critical courses on their relationship to the USA. And the growing numbers of scholars from other regions of the world now in US universities are generating new intellectual approaches, theories, and understandings.

The expanding attention of area studies scholars to diasporic populations is recontextualizing the prior focus on the nation state as the primary actor and natural unit of international analysis. The nation state was a great social invention of the nineteenth century, and since then it has spread across the globe as the seemingly necessary macropolitical unit for organizing societies and interstate relations. Yet public, political, and scholarly interest in nation states has drawn attention away from other powerful world-shaping macro institutions and processes, including the diasporas now coming into focus, the multiple forms of capitalism, and world-girdling institutions and movements, from the World Bank, IMF, and the United Nations and its international conventions to the environmental or feminist movements. These alternative macro-foci to the nation state vary in salience and manifestations in different world areas, but all are variously overriding or circumventing traditional nation state and area boundaries. Their significance, however, requires close analysis of particular manifestations and processes in diverse parts of the world, the classic role of area studies scholarship.

Currently, US area studies is also changing with a growing recognition of the necessity of serious collaboration with scholars in other parts of the world. While well trained area studies scholars, as outsiders, may discern elements of a society or culture that insiders tend to take for granted, as outsiders they inevitably also miss key local understandings and dynamics. Scholars and intellectuals inside those societies have different perspectives, experiences, agendas, and priorities than their US counterparts, and can answer questions, redirect straying analyses, and illuminate unimagined domains. Russian scholars are now working with US, European, and other counterparts to clarify the multiple transitions their society is experiencing. The theoretical generativity of Latin American studies, a field long marked by high levels of collaboration, only underscores this point. And scholars in other regions of the world who command the local historical dynamics, languages, literatures, philosophies, and cosmologies are powerfully challenging Western formulations and presumptions (Smith 1999). Indeed, a major source of new social theory seems likely to derive from efforts to integrate analyses of social experience in a much wider variety of societies than has been the case in the past.

More fundamentally, collaborators abroad assist US scholars in seeing the particularities and limitations of their own US-based agendas, perspectives, and theories. Collaboration and complementarity, engaging the insider's and outsider's views, should provide fuller and more analytically rich and useful accounts of both US and other societies, than either view alone. While this is easy to assert in principle, it is often difficult to achieve in the current global context. European social theorists (Bourdieu, Foucault, Giddens, Gramsci, Habermas, Hall, etc.) and South Asian 'Subalterns' are providing important new perspectives and intellectual frameworks. However, US scholars and universities still shape much world-wide academic (and public-policy) discourse. US views tend to define the key questions, approaches, and methods, and US universities train large numbers of scholars from all over the world, socializing them into the particular assumptions and perspectives of the US disciplines. In this context, genuinely collaborative relationships—most likely led by area studies scholars—drawing on multiple national perspectives, will increasingly be important to avoid reading into other societies the presumptions of one's own.

Although a political impetus initiated area studies in the USA, the character and intellectual agendas of the individual area studies fields have diverged dramatically since the 1950s. They have varied with the shifting mix of the disciplines involved most centrally in particular fields, and the fashions within them; with the difficulty of access and learning local languages and histories; with events in the countries being studied; with US foreign policy, and domestic politics and demography; with funding sources and funders' interests; and with the accumulated prior scholarship of, access to, and collaborative relationships with, scholars in the area being studied. Individually and collectively, however, the area studies fields have played a deeply innovative and generative intellectual and institutional role in US universities. At the start of the twenty-first century, area studies continues to produce intellectual challenges to the humanities and social sciences, structural challenges to the organization of the US university, innovative approaches to the generation of social theory, at least rough translations and greater knowledge of other societies and cultures, and greater comparative understandings of US society and culture as well.

See also: Area and International Studies in the United States: Stakeholders; Comparative History; Comparative Studies: Method and Design; Regional Geography; Regional Science

Bibliography

Bennett W C 1951 *Area Studies in American Universities*. Social Science Research Council, New York

Mitchell T (in press) The Middle East in the past and future of social science. In: Szanton D (ed.) *The Politics of Knowledge: Area Studies and the Disciplines*, University of California Press, Berkeley, CA

Pred A 2000 *Even in Sweden: Racisms, Racialized Spaces, and the Popular Geographical Imagination*. University of California Press, Berkeley, CA

Rosen G 1985 *Western Economists and Eastern Societies: Agents of Change in South Asia, 1950–1970*. Johns Hopkins University Press, Baltimore, MD

Said E 1979 *Orientals*. Vintage Books, New York

Smith L T 1999 *Decolonizing Methodologies: Research and Indigenous Peoples*. Zed Books, London

Volkman T 1999 *Crossing Borders: Revitalizing Area Studies*. The Ford Foundation, New York

Wallerstein I et al. (eds.) 1966 *Open the Social Sciences: Report of the Gulbenkian Commission on the Restructuring of the Social Sciences*. Stanford University Press, Stanford, CA

D. L. Szanton

Area and International Studies in the United States: Stakeholders

Area and international studies (A&IS) represent a crossing point between academic disciplines, research and policy institutions, numerous ethnic foreign-language communities, various international exchange organizations, foreign and domestic corporate interests, national security agencies, and issue-oriented oppositional movements. As a result, the A&IS enterprise is contested by shifting coalitions defined by such factors as professional self-interest, intellectual paradigms, national security threats, and ideological agendas.

The various A&IS stakeholder communities can be divided into three major categories: the producers of knowledge (individuals and institutions, as well as the associations that represent them), the consumers of knowledge (in the form information or in the form of A&IS-trained personnel), and the investors (in both production and consumption).

1. *The Producers of A&IS*

The institutional locus of area and international studies varies across nations. This locus is a key determinant of the complexity of the stakeholder community. In many countries, including France, the former Soviet Union, and China, A&IS research and training is located in national academies, or government-funded think tanks (Lambert 1990, pp. 712–32). These are self-contained nonuniversity institutions

whose staff and students are devoted to language training, empirical research, and policy analysis. A variant, most notably in Great Britain, is the free-standing research institute devoted to philology and historical studies of a foreign civilization. Such self-contained institutes are separate, and often distant, from institutions of higher education in their countries. As a consequence, A&IS expertise in these countries is generally not conducted with reference to the standards of the academic disciplines, but represents a nondisciplinary or multidisciplinary tradition that serves as its own point of reference. The stakeholders are basically limited to the area specialists in the national academies and their client government agencies with foreign diplomatic, economic, military, or national security concerns.

In contrast, A&IS in the United States are found in institutions of higher education, particularly in the research universities. The standard organizational model consists of a coordinating center or institute that offers the less commonly taught languages (LCTLs), while the more commonly taught languages and relevant courses in the various academic fields are offered by the disciplinary departments. The area and international studies faculty are employed by these departments and judged by them for tenure and promotion. As a result, research and teaching in A&IS are judged primarily by disciplinary standards. Faculty in A&IS programs have primary appointments in disciplinary departments, teach courses in those departments, and conduct research in those disciplines. Thus, while the A&IS programs are interdisciplinary inasmuch as they list faculty and courses from different disciplines, A&IS teaching and research are almost exclusively disciplinary in nature.

A&IS graduate degrees, especially PhDs, are therefore normally conferred in a discipline. Department of Education data for all graduate degrees produced by federally-funded comprehensive A&IS centers in the 1991–4 period show that 91.5 percent received disciplinary or professional degrees, while only 8.5 percent received areas studies degrees,and these were largely at the MA level (Schneider 1995a, p. 9, Table C). Even those programs that award a graduate degree in study of a foreign area do so through a curriculum based on disciplinary courses offered by departments, usually with a disciplinary major.

Because of this integral relationship to higher education, A&IS studies are far more developed in the United States than in other countries. In the year 2000, federal funding through Title VI of the Higher Education Act supported 113 comprehensive (graduate and undergraduate) A&IS centers at colleges and universities, another 57 undergraduate centers, 26 undergraduate international business programs, 25 graduate centers for international business education and research, 7 foreign language resource centers, 11 overseas research centers, and one program to recruit A&IS students at minority-serving institutions. A&IS programs that do not receive federal funding are far more numerous, at least four times the number receiving aid. The total number of formally organized A&IS programs in the USA is probably in excess of 500.

The A&IS stakeholders in the US system are far more varied than in countries with the national academy model. The faculty members involved in the academic programs are usually members of disciplinary associations (such as the American Historical Association), and members of A&IS professional associations (such as the Latin American Studies Association and the International Studies Association), which means that these associations are stakeholders.

The membership of the A&IS professional associations are a guide to the size of the knowledge producer population. Although only about two-thirds of the area studies association's membership are faculty, not all foreign area studies faculty belong to these associations. Past studies have therefore estimated that total membership in the area studies association provide a good approximation of faculty employment in foreign area studies (Lambert et al. 1984, p. 13). The total membership of the five major area studies associations, which cover Africa, Asia, Latin America, the Middle East, and the former Soviet Union and Eastern Europe, was approximately 16,000 in 1990 (NCASA 1991). This compares with a 1979 membership in all area studies associations of about 18,000. If the 1990 memberships of the smaller area studies associations such as Brazilian, Canadian, European, and Caribbean studies are added to those of the five major associations, the total remains approximately 18,000, about the same as in 1979 (Barber and Ilchman 1977, p. 15). Membership in the International Studies Association adds another 3,000 persons, for a combined total of 21,000. Another faculty stakeholder group is foreign-language and literature teachers, for whom two sources of information exist, membership in professional organizations, and surveys of the National Center for Education Statistics of the US Department of Education. Using these two sources, a recent study estimated a total 36,000 post-secondary language and literature faculty (Merkx 2000, pp. 93–100). Added to the previous groups, the estimated total of faculty stakeholders is 47,000 persons.

Campus units that support A&IS, such as the library and the language departments, have an investment, which means that their professional organizations (such as the Association of Research Libraries or the National Council for the Less Commonly Taught Languages) are also involved. The numerous academic disciplines that contribute to A&IS programs, including foreign languages and literature, have an interest. Likewise, those universities whose prestige and enrollments reflect success in A&IS programs have a stake, as do the presidential associations that

represent their interests in Washington (such as the Association of American Universities (AAU)).

Campus-based components of A&IS face internal rivals on campus, however, and these rivalries may be reflected in division within their respective associations. Within key departments, such as history or political sciences, there may be struggles for resources and faculty lines with non-A&IS factions. Within the library, the A&IS collection effort competes with other collection priorities. The language department may be unwilling to offer foreign languages that are needed by A&IS programs but attract low enrollments. The entire A&IS community of a university must compete for funding against student aid, faculty compensation, science and technology programs, and other priorities. Even at the national level, the associations and organizations representing the A&IS campus-based community may face internal conflicts over competing priorities, or find themselves lobbying at cross-purposes. In turn, collaboration or conflict at the national level has a significant influence on levels of investment in A&lS.

Additional A&IS knowledge is produced in the USA by government intelligence agencies, military institutions, government-sponsored think tanks, and risk analysts employed by corporations. This information is generally not accessible to the public, and hence is a relatively minor component of A&IS knowledge. These agencies and institutions are important stakeholders, however, as consumers of A&IS information and training, and to a lesser extent, as funders.

2. *The Consumers of A&IS*

A second category of stakeholders is constituted by those institutions that need A&IS information or need A&IS-trained personnel. These can be divided in turn into three broad sectors: government, business, and education.

There is a surprising consistency over time in estimates of US government manpower needs for foreign-language and area-trained personnel. The most cited and thorough study is that of James R. Ruchti of the US Department of State, prepared in 1979 for the Perkins Commission, which surveyed more than 25 agencies and concluded that the federal government employed between 30,000 and 40,000 individuals whose jobs required competence in a foreign language, and that, of these persons, between 14,000 and 19,000 were in positions that required skills in the analysis of foreign countries and international issues (Ruchti 1979, cited at length in Berryman 1979, pp. 75–114). Although estimates of declining government employment have been assumed to reduce the need for foreign language skills in the federal government, this was not evident in the mid-1990s. The most recent survey of foreign language needs at 33 federal agencies, undertaken by Stuart P. Lay in 1995, concludes that these agencies have over 34,000 positions that require foreign language proficiency, of which an estimated 60 percent are found in the defense and intelligence community (Lay 1995). Anecdotal evidence suggests that reductions in force since 1995 may have lowered these figures in the nondefense government sectors, but the preponderance of defense and intelligence employment would reduce the effect of such reductions. If 30,000 positions are taken as a possible lower end estimate to account for reductions, and it is assumed that because of the relatively high turnover of military personnel 20 percent of the federal positions will require replacement in any given year, a replacement need of 6,000 government positions per year can be estimated.

Business demand is hard to estimate. Anecdotal evidence suggests that the US business climate has an increased need for A&IS information and skills. In the 1970s, US business was widely seen as uncompetitive on world markets and lagging in productivity. A survey from this period indicates that less than 1 percent of jobs at 1266 US firms, which accounted for the great majority of industrial exports, required foreign language skills. Nevertheless, there were 57,000 jobs at these firms that required or benefited from foreign language skills (Wilkins and Arnett 1976). Given the predominance of employment in small firms as opposed to large industrial firms, this was clearly an underestimation of private-sector demand at that time.

Adding an equal-sized small-firm component leads to an estimate of about 100,000 positions, which with a 10 percent turnover would have required replacement of 10,000 positions per year. Since the 1970s, the share of the US gross national product resulting from international trade has quadrupled. It is therefore reasonable to assume that employment in the private sector of personnel with foreign language and area skills has increased by several magnitudes. Even if such employment has only doubled since the 1970s, the business sector would need to fill 20,000 positions per year involving foreign language and area competence.

Educational demand for A&IS has been better studied, but not without controversy. The 1970s recession in higher education employment, combined with the Nixon-inspired drop in Title VI funding, led to gloomy projections about the future academic demand for language, international, and area-trained personnel (see, for example, the extended discussion in Berryman et al. 1979, pp. 30–74). There was also a concomitant reduction in the production of international and area studies PhDs, compared with the 1960s. However, the optimistic projections of the Barber and Ilchman study of 1977 proved more accurate: they noted that tenured area studies faculty were significantly older than the general tenured faculty population, and predicted a surge of retire-

ments over the following ten years (Barber and Ilchman 1977). The academic employment market for language and foreign area specialists was indeed strong during the 1980s.

Another dimension of change in higher education created additional need for foreign language, international, and area studies faculty, namely the expansion of public-sector undergraduate teaching institutions, most notably community colleges and branch campuses. The growth of this sector included a substantial and largely unforeseen growth in international education activities, including the teaching of language, international, and to a lesser extent, foreign-area content courses. By the end of the 1980s a sizable proportion of members of the foreign area studies association were located at undergraduate teaching institutions.

During the 1990s the demand for post-secondary faculty strengthened, and this has contributed to a sense of crisis in the Title VI community. Unemployment among PhDs dropped by more than one-third in the mid-1990s, and all major professional associations, including those in foreign language and area studies, reported increased postings of job announcements. In part this reflected the retirement of faculty hired during the higher education boom of the 1960s. Because area and international studies grew even more rapidly than higher education as a whole in the 1960s and 1970s, the job market for these specialists will be strong through the first decade of the twenty-first century.

Detailed projections of retirement patterns based on the age cohorts of the area studies association's membership were prepared in 1991 by the National Council of Area Studies Associations, which represents the five major area studies associations. Exit rates of present humanities and social science faculty were based on respondents' plans to retire, estimated at 16.9 percent for 1997 through 2001, and at 16.8 percent for 2002 through 2007, for a total of 33.7 percent, or one-third of current faculty. These estimations do not include projections of exits due to morbidity or mortality based on the age structure of the cohorts, which would, if included, lead to an overall exit rate of approximately 40 percent. Using the latter figure and assuming that exiting faculty are replaced but there is no growth in academic demand, 40 percent, or 7,000, of the current 18,000 area studies faculty will need to be replaced in the first 10 years of the twenty-first century, or 700 positions per year. An additional 1,200 positions will be opening in international studies, or 120 faculty positions a year.

Another faculty stakeholder group consists of foreign-language and literature teachers, for whom two sources of information exist, membership in professional organizations, and surveys by the National Center for Education Statistics of the US Department of Education. Using these two sources, a recent study estimates the combined foreign language and literature faculty in higher education at the start of the twenty-first century at 36,000 persons. If the exit projection of 40 percent for area studies faculty over the first decade of the new century is applied to the estimated total of 36,000 post-secondary language and literature faculty, an additional 14,000 positions would need to be filled.

3. *Demand and Supply*

The demand in federal government, business, and higher education for personnel who have foreign language skills, and international and foreign area knowledge will be substantial in the coming decades. Additional demand from state and local government and from secondary education may also be expected. Overseas employment offers another, as yet unexplored, source of demand. Annual demand for the first decade of the twenty-first century is estimated at 20,000 business jobs, 6,000 government jobs, and 10,000 education jobs, for a total of 36,000 foreign-area, international, and language-trained personnel.

The supply side of the equation is far simpler to estimate, as are the implications for Title VI legislation. The number of federal Foreign Language and Area Studies (FLAS) Fellowships awarded annually through Title VI of the Higher Education, which was 600 in 2000, does not meet even the academic demand, although it is a stimulus for recruiting superior students. The overall annual production of PhDs by Title VI centers was about 1,400 per year in the early 1990s. Beginning in 1993 there was a substantial increase in the number of universities receiving NRC or FLAS funding, leading to a jump of PhD production to about 1,900 language and area trained personnel (Schneider 1995b). These numbers, which have since been stable, are less than the estimated annual higher education demand for 2,100 foreign language or area studies faulty over the first decade of the twenty-first century.

The production of MA degrees by Title VI centers is higher, averaging about 6,000 per year in the 1990s (Schneider 1995b). That number is far below the annual combined demand for about 14,000 persons coming from the K-12 education sector and the federal government. Production of BAs by Title VI centers approximated 27,000 students annually by the early 1990s, of which an estimated 6,000 enter graduate school, and 21,000 of the BA graduates enter the job market. Of the 6,000 MA recipients, about 2,000 continue graduate study and 4,000 enter the market. Thus a combined total of 25,000 BA and MA graduates with foreign language or area training enter the job market, compared with a demand from business, government, and K-12 education estimated above at 34,000 positions. These positions must be filled by persons trained in other programs or on the job.

As a consequence of the shortfall of trained personnel, the federal government spends considerable sums of money on in-house training programs, such as the Department of Defense's Defense Language Institute (DLI), the National Security Agency's National Cryptologic School (NCS), and the Department of State's Foreign Service Institute (FSI). The DLI and the NCS together train about 4,600 students annually. The Department of Defense alone spent over $78 million to train linguists to meet its need, considerably more than the cost of all Title VI programs (Lay 1995, p. 1). In-house foreign language training at the FSI cost an additional $10 million. These figures do not include the salaries of the personnel who are being trained.

It should be noted in passing, however, that at least two Department of Defense programs draw on institutions of higher education to meet future needs for foreign language and area competence. The US Army Foreign Area Officer (FAO) Program annually sends approximately 100 mid-career officers to Title VI centers to obtain graduate degrees in preparation for overseas assignments in embassies, foreign war colleges, or military aid missions. The National Security Education Program provides on a competitive basis portable scholarships and fellowships to students undertaking foreign language and area training, as well as grants to enhance the institutional capacity for such training.

4. *The Investors in A&IS*

Investment in area and international studies in most of the world is supported almost exclusively by national governments. In the United States, funding for A&IS has come from the federal government, but also from two other sources—private foundations, and from colleges and universities themselves. Title VI of the Higher Education Act (HEA), formerly the National Defense Education Act (NDEA), administered by the Department of Education, has been the primary instrument through which the federal government supports A&IS (cited in McDonnell et al. 1981, p. 11). Beginning in the mid-1990s, the National Security Education Act provided additional support through the Department of Defense.

The history of A&IS been something of a roller coaster. Federal funding was critical to the development of A&IS in the United States, but has been threatened or reduced a various times. Foundation funding was highly important at one time, but was later reduced to modest levels. Despite this variation of support through time, the commitment of higher education institutions support for A&IS, once established, has proven relatively constant.

The original rationale for NDEA as a whole was narrow and clearly articulated: 'To insure trained manpower of sufficient quality and quantity to meet the national defense needs of the United States' (cited in McDonnell et al. 1981, p. v). In the post-Sputnik atmosphere of concern about Soviet achievements in science and technology, the NDEA legislation focused primarily on training in the physical sciences and engineering. However, *prior* to Sputnik the US Office of Education had prepared draft legislation on foreign language and area training. This legislation, according to one contemporary official, had been drafted because, 'By the mid-1950s responsible people in the Government were beginning to realize that university resources in non-Western studies were wholly inadequate to meet present and anticipated national needs. Some measure of Government assistance to language and area studies seemed essential' (Mildenberger 1966, pp. 26–9). Kenneth W. Mildenberger organized and headed Title VI programs following passage of NDEA.

The Office of Education's foreign language and area studies draft was incorporated in the NDEA bill as the result of negotiations between Assistant Secretary of Health, Education and Welfare, Eliot L. Richardson, on behalf of the Eisenhower Administration and the sponsors of the legislation, Senator Lister Hill and Representative Carl Elliott (Clowse 1981). Like all sections of NDEA, the original Title VI emphasized training, in this case of individuals in modern foreign languages 'needed by the Federal Government or by business, industry, or education' and 'not readily available in the United States.' Such individuals were also to be trained 'in other fields needed to provide a full understanding of the areas, regions, and countries in which such language is commonly used,' including 'fields such as history, political science, linguistics, economics, sociology, geography, and anthropology' (National Defense Education Act of 1958, as Amended, reproduced in Bigelow and Legters 1964).

The original NDEA Title VI legislation envisioned a 50–50 partnership in which the costs of foreign area centers would be divided equally between the universities and the federal government (a munificent arrangement in comparison to the 20–1 ratio of university to federal support for Title VI centers at the time of writing) (McDonnell et al. 1981, p. 38). Federal support for foreign area studies was augmented by sizable investments from the philanthropic community, led by foundations such as Ford, Rockefeller, Mellon, Carnegie, and Tinker. The Ford Foundation alone contributed about $27 million annually between 1960 and 1967 for advanced training and research in international affairs and foreign area studies, more than federal appropriations for Title VI in the same period.

The establishment of foreign-area programs was a costly venture for universities even with federal and foundation subsidies. Yet universities responded to NDEA Title VI with enthusiasm. This reflected a consensus between academia and government on

national needs in the field of international education, forged by World War II and reinforced during the early stages of the Cold War. Many, if not all, foreign-area scholars, university administrators, and foundation executives had served in State Department, intelligence, or military positions during World War II or the Korean War. The boundaries between government and academic institutions were permeable and amicable. The relatively small size of the community of foreign area specialists inside and outside government meant that people knew one another. Perhaps most important of all, they shared a similar perspective on the US role in world affairs. The alliance between academia and government for foreign area training and research was a natural outgrowth of these affinities.

The Title VI experiment proved highly successful. Within a decade the United States had established a network of centers covering most foreign areas, generating unprecedented quantities of research, and producing substantial numbers of new foreign language and area specialists. Ancillary institutions such as foreign-area studies associations and research journals multiplied quickly. The prestige and the funding conveyed by Title VI designation continued to be rewarded by university administrations (Merkx 1990, pp. 1, 18–23).

Following the promising early start of Title VI came the Vietnam War, which by the late 1960s led to funding freezes and then to declines. Moreover, the Vietnam War itself produced a significant deterioration in the relationship between government and academia. Academic dissent from US foreign policy in South-West Asia began to grow during the Johnson Administration and reached high levels during the Nixon Administration. The events of Watergate did little to increase academic confidence in government. Resentment by government officials of criticisms from academia, including from some of the very foreign area specialists trained under Title VI, grew as well. The gradual retirement of the World War II generation of leaders further contributed to the growing gulf between academic and government cultures.

The Carter Administration made some effort to improve relations with the academic community. Carter appointed a Presidential Commission on Foreign Language and International Studies (known as the Perkins Commission), which in 1979 issued a strong call for increasing federal support of international education to more than three times existing levels (US Department of Health, Education, and Welfare 1979). Among its recommendations were the establishment of undergraduate and regional foreign area centers as complements to the Title VI national resource centers, and the provision of annual federal grants of $50,000 to each of the national resources centers for library costs. However, the timing of the Perkins Commission report was not propitious. Double-digit inflation and high interest rates were leading to reductions, not increases, in federal expenditures.

A more successful initiative by the Carter Administration was to repeal NDEA and incorporate its Title VI functions in a new Higher Education Act (HEA). This step ratified the separation of university-based foreign area studies from the national defense needs that had been the original justification for federal funding. While pleasing many campus-based area specialists, the change of name made Title VI a more vulnerable target for those in the next administration who were to argue that federal subsidies were no longer in the national interest. HEA was also accompanied by the controversial establishment of the Department of Education (ED) as a cabinet-level institution, despite strong objections from the Republican minority in the Congress.

Within months after the establishment of the Department of Education, President Carter lost the 1980 elections to Ronald Reagan. The Department of Education and Title VI were among the targets singled out by the incoming Republican administration. Reagan's Office of Management and the Budget (OMB) recommended elimination of all Title VI appropriations and continued to do so for all of the first seven Reagan budgets, even after the White House had conceded defeat on its goal of eliminating the Department of Education. University–government relations were not enhanced by faculty criticisms of the Reagan Administration's increased emphasis on defense expenditures and its aggressively interventionist stance in Third World zones of conflict, nor by administration insinuations that foreign-area specialists were unpatriotic partisans of the countries they studied.

Thus the early years of the Reagan Administration represent the nadir of the partnership between academia and government in foreign area studies. Due to the high inflation of the late 1970s and early 1980s, and essentially level funding in current dollars, Title VI appropriations in real terms were at their lowest level since the inception of the program, falling to well below half their 1967 high point. The Administration was formally opposed to the program. Higher education in turn was suffering from the consequences of inflation, the growth of university enrollments had taken a sharp downturn for demographic reasons, and foundation support for international education had evaporated.

Nonetheless, the partnership between government and academia survived. Support for Title VI came from three directions. A few well-placed government officials recognized the value of Title VI research and training, and were moved to intervene on its behalf. Perhaps the most famous example is the letter of 11 March 1983 sent by Secretary of Defense, Caspar Weinberger, to Secretary of Education, Ted Bell, with a copy to OMB Director, David Stockman, requesting reconsideration of the zero-funding of Title VI.

Weinberger noted, 'My concern is shared by other officials within the Department of Defense, and members of the academic community on whom we depend for both a solid research base in area studies, as well as for production of foreign language specialists.' The Deputy Director of Central Intelligence, Rear Admiral Bobby Inman, was another outspoken defender of Title VI programs.

A second source of support came from the directors of Title VI-funded area centers, who were galvanized by the realization that the survival of Title VI programs could no longer be taken for granted. The second-generation center directors of the 1980s lacked the Washington contacts and political knowledge of the founding generation. They were forced, however, to increased activism by the disastrous implications of a total cut-off of federal support for their programs. After the failure of their efforts to reverse the Reagan Administration zero-funding of Title VI in the early 1980s, the center directors focused on the Congress. (The author, for example, met with Vice President Bush's chief of staff, Boyden Gray, in 1981 to submit several proposals concerning Title VI funding and administration, which were then transmitted to the Vice Present in writing, eliciting from him a pleasantly noncommittal response.) These early attempts at congressional relations were for the most part uncoordinated individual initiatives, but were not without effect. Congress proved more responsive to the universities than the administration had been, and Title VI funding survived, albeit at the relatively low levels of the Carter period.

The third source of support came from individuals outside of government and academia that were affiliated with institutions involved directly or indirectly with international education, such as the major foundations, higher education associations, and professional organizations. Some of these persons had served on the Perkins Commission. Others were involved with the successful effort to obtaining funding for Soviet Studies that led to the passage of the Soviet–Eastern European Research and Training Act of 1983 (Title VIII of the Department of State Authorization Act). From this sector came two important studies calling attention to inadequacies in the US international studies, the *Report of the Task for on National Manpower Targets for Advanced Research on Foreign Areas* (NCFLIS 1981) and *Beyond Growth: The Next State in Language and Area Studies* (Lambert et al. 1984).

The growing sense of shared concern about the chronic underfunding of international education in general, and the threat to Title VI in particular, coalesced in an initiative that began with a dinner at the Smithsonian Institution in late 1984 'to discuss what might be done to stabilize long-term federal support for international studies and foreign language training' (Prewitt 1986, p. v). Several foundations funded, under the aegis of the Association of American Universities (AAU), a study which resulted in the monograph *Points of Leverage: An Agenda for a National Foundation for International Studies* (Prewitt 1986, p. v), accompanied by the recommendation of an AAU advisory committee chaired by Kenneth Prewitt of the Rockefeller Foundation (AAU 1986). The Prewitt committee offered draft legislation for the establishment of a National Foundation for Foreign Languages and International Studies. The intent of the proposal was to centralize federal support for international education in a single entity, as opposed to the existing 196 international studies and exchange programs in 35 departments of the federal government. It was thought that this would allow supporters of international education to focus their advocacy on a single, high-profile institution that would be analogous to the National Science Foundation or the National Endowment for the Humanities.

The AAU Legislative proposal was intended to serve as a rallying point for the various international education constituencies. It had the opposite effect. The proposal was viewed with suspicion as being the product of the elite institutions associated with the AAU. It was widely misinterpreted as constituting a threat to existing programs. The process by which the proposal was developed was criticized for having too narrow a base and not including a sufficiently broad spectrum of international education interests. Perhaps the most positive thing that can be said of the reaction was that its vigor reflected the growth of concern about the future of the US international education effort.

In a constructive response to these criticisms, the AAU obtained additional foundation support to broaden the dialogue about international education needs, establishing in 1987 a two-year effort known as the Coalition for the Advancement of Foreign Languages and International Studies (CAFLIS). An open invitation to participate was sent to virtually every type of group that might be interested. Ultimately, some 150 organizations participated in one or more stages of CAFLIS discussion, including most higher education associations, area studies associations, language groups, organizations engaging in foreign exchanges, peak organizations in the social sciences, and professional associations.

Despite, or perhaps because of, the inclusive nature of the process, CAFLIS ultimately ended in failure. The various constituencies represented in the CAFLIS process had different agendas and found difficulty coming to an overall agreement. A fundamental cleavage existed between those groups that emphasized the need for increased investment to meet the original Title VI goals of advanced training and research in foreign language and area studies (a focused and less costly agenda), and those who were advocating federal subsidies for language training in primary and secondary schools, for internationalization of undergraduate education, or for adding

international dimensions to professional training in fields such as business, medicine, and engineering (a diffuse and more costly agenda). Even after the recommendations of the three CAFLIS working groups were watered down and widened to satisfy groups dissatisfied with the original Title VI agenda, some of these groups refused to ratify the final recommendations, which appeared in late 1989 (CAFLIS 1989). The result was a proposal for a new federal foundation that would not incorporate existing federal programs but merely add to them, an agenda that failed to generate a congressional response.

The disappointing outcomes of both the AAU legislative proposal and the CAFLIS process led the core A&IS groups to refocus attention on Title VI. The Higher Education Act was to expire in 1990. The Association of American Universities, the National Association of State Universities and Land Grant Colleges, and the other higher education presidential associations organized a series of meetings on Title VI reauthorization that culminated in the appointment of an Interassociation Task Force on HEA–Title VI Reauthorization. The directors of foreign area centers receiving federal funding formed a Council of Title VI National Resource Center Directors (CNRC). The foreign-area studies associations organized a coordinating group of executive directors and presidents known as CASA (Council of Area Studies Associations).

This resurgence led to a strategy of building support for Title VI programs in particular, and A&IS in general, by mobilizing the campus-based core constituencies, their professional associations, and national organizations such as the AAU. At the invitation of CNRC a meeting of all these groups took place at the Library of Congress in 1991, for the purpose of defining a common agenda with respect to Title VI reauthorization. When the final law reauthorizing HEA was approved by the Congress in 1992, Title VI incorporated, virtually intact, almost all of the recommendations of the Interassociation Task Force which had been supported by the core constituencies.

This success of the 1991–2 reauthorization effort resulted in the establishment of the Coalition for International Education, which includes 26 national organizations representing the various A&IS stakeholder communities. This membership includes six higher education associations representing college and university presidents, two associations of international education administrators, four associations of Title VI-funded program directors (representing area centers, business centers, language centers, and schools of international administration), two library associations, two international exchange associations, one council of overseas research centers, one area studies association, a social science association, a humanities alliance, an overseas university, and an association of graduate schools.

Federal funding for A&IS reached a low point in constant dollars during the mid-1980s. Since the establishment of the Coalition for International Education, it has grown steadily, if modestly. Adjusted for inflation, funding for Title VI grew 6.5 percent between the 1994 and 2000 federal budgets. The original Title VI programs, such as the area centers, are funded at about one-half their constant dollar figures of the late 1960s. The overall International Education and Foreign Language Studies account in the year 2000 federal budget was approximately $70 million.

5. *Overview*

Area and international studies constitute a diverse and often fractious group of stakeholders in academia and government. In the country where they are most developed, the United States, area and international studies are divided by disciplinary, geographic, linguistic, professional, and institutional fault lines, and face a constant struggle for resources with competing academic fields. Until the 1990s, the different A&IS stakeholders failed to cooperate effectively. Since that time a working alliance has been effective in lobbying for continued government support. Demand for graduates with A&IS training is strong and growing, which appears to account in part for healthy enrollments and continued support by colleges and universities. The post-Cold War increases in international trade and other forms of globalization suggest that area and international studies will remain a growing component of both the humanities and social sciences, despite internal rivalries and external competition for scarce academic resources.

See also: Cold War, The; Policy Knowledge: Universities

Bibliography

AAU 1986 To Strengthen the Nation's Investment in Foreign Languages and Internatioal Studies: A Legislative Proposal to Create a National Foundation for Foreign Languaes and International Studies. Association of American Universities, Washington DC, October 3

Barber E G, Ilchman W 1977 *International Studies Review*. The Ford Foundation, New York

Berryman S E et al. 1979 *Foreign Language and International Studies Specialists: The Marketplace and National Policy*. The Rand Corporation, Santa Monica, CA, September, pp. 30–74

Bigelow D N, Legters L H 1964 *NDEA Language and Area Centers: A Report on the First Five Years*. US Department of Health, Education, and Welfare, Office of Education. US Government Printing Office, Washington DC

CAFLIS 1989 *The Federal Government: Leader and Partner*. Report on the recommendations and findings of the Coalition

for the Advancement of Foreign Languages and International Studies: Working Group on Federal Support for International Competence, December. CAFLIS, Washington DC
Clowse B B 1981 *Brainpower for the Cold War: The Sputnik Crisis and the National Defense Education Act of 1958*. Greenwood Press, Westport, CT
Lambert R D 1990 Blurring the disciplinary boundaries: Area studies in the United States. *American Behavioral Scientist* **33**(6) July/August: 712–32
Lambert R R et al. 1984 *Beyond Growth: The Next Stage in Language and Area Studies*. Association of American Universities, Washington DC, p. 13
Lay S P 1995 Foreign language and the Federal Government: Interagency coordination and policy, MA thesis, University of Maryland.
McDonnell L M et al. 1981 *Federal Support for International Studies: The Role of the NDEA Title VI*. The Rand Corporation, Santa Monica, CA, p. 11
Merkx G W 1990 Title VI accomplishments, problems, and new directions. *LASA Forum* **21**(2) Summer: 1: 18–23
Merkx G W 2000 Foreign language and area studies through Title VI: Assessing supply and demand. In: Lambert R D, Shoharny E (eds.) *Language Policy and Pedagogy. Essays in Honor of A. Ronald Walton*. John Benjamins, Amsterdam, pp. 93–110
Mildenberger K W 1966 The federal government and the universities. *International Education: Past, Present, Problems and Prospects*. Task Force on International Education, John Brademas, Chairman, Committee on Education and Labor, House of Representatives. US Government Printing Office, Washington DC, pp. 26–9
NCASA 1991 *Prospects for Faculty in Area Studies*. Report of the National Council of Area Studies Associations. NCASA, Stanford, CA
Prewitt K 1986 Preface. In: Lambert R D (ed.) *Points of Leverage: An Agenda for a National Foundation for International Studies*. Social Science Research Council, New York, p. v
Ruchti J R 1979 The U.S. government employment of foreign area and international specialists. Paper prepared for the President's Commission on Foreign Language and International Studies, July 12
Schneider A I 1995a Title VI FLAS Fellowship awards, 1991–1994. Memorandum to Directors of Title VI Centers and Fellowships Programs, Center for International Education, US Department of Education, September 15, p. 9, Table C
Schneider A I 1995b 1991–94 Center graduates: Their disciplines and career choices. Memorandum to Directors of Title VI Centers and Fellowships Programs, Center for International Education, US Department of Education, September 26
NCFLIS 1981 *Report of the Task Force on National Manpower Targets for Advanced Research on Foreign Areas*. Mimeo. National Council on Foreign Languages and International Studies, New York
US Department of Health, Education, and Welfare 1979 *Wisdom: A Critique of U.S. Capability*. Report to the President from the President's Commission on Foreign Language and International Studies. US Government Printing Office, Washington DC, November
Wilkins E J, Arnett M R 1976 *Languages for the World of Work*. Olympus Research Corporation, Salt Lake City, June

G. W. Merkx

Area and International Studies: International Relations

1. *Origins of Area Studies*

Area studies originated from American frustration over the inadequacies of European-originated social sciences, which focused primarily on European and North Atlantic societies for a long time. In other words, the discipline arose because of the frustration of many Americans in coming up with relevant knowledge of and insights into non-Western societies in the twentieth century (Hall 1948). First of all, empirical data were pitifully scarce. Second, European-originating social sciences did not seem to give credence to their own propositions about economic development, democratization, and rule of law. Third, for their war effort, at first, and then later for their own vision of global governance, the Americans needed broader coverage of social sciences over the globe. Area studies was born of such factors in the USA.

At first, area studies focused on local societies, local dynamics and local logics, and tried to accumulate 'thick description' as described above. The Human Relations Area Studies File compiled at Yale University is one of the best examples of such efforts. It is an overwhelming, detailed ethnographic work covering the beliefs, customs, and social practices of many societies in the world, carried out mostly by anthropologists and geographers (Murdoch 1981). Another, far lesser known, example is the file of food patterns compiled under the auspices of the United Nations University in Tokyo. It details the range and nature of food intake in a vast number of places in the world in relation to the anticipated needs for food assistance, training of personnel for the preservation of health and sanitary conditions, and technical assistance in canned food. This type of area study has long been carried out by anthropologists and geographers alike.

But in the 1950s and 1960s a number of leading economists, sociologists, and political scientists felt compelled to come up with empirically testable propositions about such general subjects as economic development and democratization. This ushered in the heyday of the American-originating modernization theory which tried to give guidance to developing countries as well as to the US government on to how to proceed with the task of economic development and democratization, with US experiences portrayed as the best example. The most notable authors in this area are Walt W. Rostow and Seymour Martin Lipset (Rostow 1991, Lipset 1981). Those social scientists thirsty for generalizable propositions about economic development and democratization did not dismiss area studies at all. For them, area studies are good data basis for social scientific endeavor. Moreover, they were eager to collect evidence that accorded with their own generalizations about economic develop-

ment and democratization, i.e., their various versions of modernization theory. Therefore area studies expanded dramatically in terms of staff appointment and student enrollment in the 1960s.

As area studies became more systematized and theorized under the influence of American-centric modernization theory, many of the themes associated area studies were incorporated into the ordinary social sciences. Thus, textbooks of comparative economic systems included chapters on market economies, centrally planned economies, and developing economies, the last of which was a generalized treatment of those economies which were normally covered by area studies. Likewise, textbooks of comparative politics consisted of chapters on industrial democracies, communist dictatorships, and developing authoritarianism, the last of which was a general treatment of those political systems normally covered by area studies (Eckstein A 1971, Eckstein H and Apter 1963). At one point in the 1960s it seemed as if area studies had merged happily with ordinary social science with the injection of American-centric modernization theory into area studies. However, the picture changed fairly soon as the American experience in Vietnam in the late 1960s through mid-1970s failed to give credence to the American-centric modernization theory (Packenham 1973, Latham 2000). Furthermore, the end of the Cold War in the early 1990s drastically altered the pictures of the world those textbooks portrayed. More fundamentally, the three trends of digitalization, globalization, and democratization have started to alter the whole framework and tenets of comparative economic systems and comparative politics (Inoguchi 2001).

With these three trends steadily intensifying, the framework and tenets of ordinary social science, which focused on the nation-state, the national economy, and the national culture, has begun to look slightly too narrow to deal with increasingly globalized, localized, and trans-nationalized economic transactions and political interactions (Inoguchi 1999, Katzenstein et al. 1998). Therefore, from the late 1990s onwards textbooks of comparative economic systems were replaced by those of the more open or less open economies under the globalizing market system. Likewise, textbooks of comparative politics categorizing regimes into industrial democracies, communist dictatorships, and developing authoritarianism were replaced by those categorizing regimes into established democracies, newly emerging or transitional democracies, and non-democracies and failed states (Sachs 1993, Kesselman et al. 1999). It was clear that by the start of the twenty-first century the two benchmarks—well-functioning market economies and well-functioning democracies—have prevailed as the organizing principles of mainstream social sciences. Thus the subject of comparative market systems such as the Anglo-American model, the Continental European model, and the Japanese model has become fashionable. Likewise, the subject of comparative democratization has become a commonplace (Thurow 1997, Rose and Shin 2001, Inoguchi 2000, APSA 2000).

What, then, has become of area studies? Area studies looks now as if it has been submerged by ordinary social sciences focusing on market and/or democracy. However it seems that, with these terribly simplified frameworks, some have been left uncertain and uneasy and have tried to come up with tighter and more readily fathomable concepts on which to focus. Their solutions are the rule of law and 'high trust.' Such authors as David Landes, Eric Jones, Francis Fukuyama, and Robert Putnam all seem to be saying that how ready people are to observe the rule of law and maintain high trust in human interactions and common endeavors makes a difference, and these things are fathomable only through historically and culturally sensitive understanding of the collective activities of human beings (Landes 1998, Jones 1979, Fukuyama 1996, Putnam 1994, 2000). To sum up, area studies has been largely incorporated into ordinary social sciences under the rubrics of market and democracy while the remaining huge residues are to be understood by the resort to history (Putnam 1994, Inoguchi 2000). So, ironically, area studies is now played down institutionally and in terms of flows of money. However, the tasks assigned to area studies as envisaged by Americans in the 1940s and 1950s remain to be solved.

2. *Origins of International Relations*

International relations as a discipline was born after World War I in Europe. All the major European intellectuals pondered on the causes and consequences of the most disastrous war ever experienced. Most of them were historically oriented, and yet such authors as F. H. Hinsley, David Wight, and Edward H. Carr all argued that the notion of international society that had been long held by major European powers had been disrupted in the course of the twentieth century and that may be regarded as the basic cause of such a war (Hinsley 1986, Wight 1991, Carr 1939, Bull 1977). Yet the discipline of international relations was brought into its own by the Americans. Overcoming the idealism and isolationism that characterized the USA through the early part of the twentieth century, the country started producing works that became the classics of the international relations genre in the 1940s and 1950s. Realism and internationalism became the dominant modes of American thinking about international affairs by the 1940s. Such authors as Hans J. Morgenthau, George Liska, and Arnold Wolfers were representative authors (Morgenthau 1978, Liska 1977, Wolfers 1962). At the same time behavioral, i.e., systematic and empirical, examinations of international relations were produced the 1940s by, among others, Quincy Wright and Harold

Lasswell and his associates (Wright 1942, Lasswell et al. 1980). Furthermore, area studies occupied an important place in the study of international relations most broadly defined. The study of international relations covered not only inter-state relations but also anything that took place outside the USA. All in all, Americans dominated the field of international relations studies by the 1950s (Hoffmann 1977). Most intellectual currents were represented by their writings, while the new mode of analyzing international relations in the framework of behavioral science and the new area of study called 'area studies' that was attached to the study of 'international relations' prospered in the USA. American dominance was reflected by the salience throughout the world of the US publications *Foreign Affairs* and *World Politics*; journals representing the policy-oriented establishment and the academic establishment respectively.

3. *Intersection of Area Studies and International Relations*

The place of area studies in the study of international relations has always been ambiguous. Initially it was simple and clear. From the dominant US perspective, anything that took place outside the USA was defined as international affairs, and that presumably included all subjects that also came under 'area studies.' However, as the study of international relations became more diversified and as the study of comparative politics encompassed much of area studies as it applied to developing countries, at one point in the 1960s it seemed as if international relations and area studies were to be isolated from each other. International relations focused on realism of various kinds and area studies focused on politics in developing countries, which was then referred to as political development in the Third World—Asia, Africa, and Latin America. Yet the isolation of the disciplines turned out to be short-lived. The key concepts giving coherence to international relations and political development respectively, i.e., state sovereignty and modernization, came to be more easily compromised by such forces as economic interdependence and the nonlinearity of the move from economic development to political development, which became more pronounced in the 1970s through the 1980s. From the 1970s onward those forces that undermine the system of international relations under the guidance of state sovereignty—such as economic interdependence and transnational relations—became so pronounced that they came to be widely regarded as concepts that shape international relations no less than the traditional concepts of state sovereignty, popular sovereignty, and the loss of sovereignty (Inoguchi 1999). This way, international relations have come to be called global politics, encompassing international relations, domestic politics, local politics, and all transnational relations. Here again, the relationship between international relations and area studies has been blurred (Baylis and Smith 1997, Held et al. 1999, George 1994).

In terms of institutional arrangements, area studies and international relations have a similarly ambiguous relationship. When 'area studies' meant studies of developing countries including colonies and when international relations meant foreign affairs taking place outside the USA, their relationship was not an issue. That was the case indeed in those years before 1945. In Japan their relationship is not a big issue. International relations means more or less everything that is taking place outside Japan, and area studies means studies of foreign countries including developed countries (Inoguchi and Bacon 2001). Area studies occupies an important place in the Japan Association of International Relations. Area studies and diplomatic history are two of the three major categories of academic genre along with international relations theory in terms of the number of members. In terms of academic training of graduate students, path dependence of a sort is easily discernible. First, those graduate students with a social science background focus on international relations theory. Second, those graduate students with a history background focus on diplomatic history. Third, those graduate students with the background of foreign languages focus on area studies. Since these three major kinds of training all produce graduate students of international relations, all the three major genres are quite evenly represented in the Japan Association of International Relations. In the USA, area studies and international relations are both fully under the department of political science. Many area studies programs may have disappeared, but language and history departments continue to give training to those political science graduate students needing some area-specific knowledge and training. International relations has been studied mostly within the department of political science, and therefore, given the theory-driven nature of American international relations, a vast majority of students are exposed to theories of international relations, whether they are structural realism, constructivism, critical theories, or behavioralism (Weaver 1998).

Reference to journals and their subject areas helps provide an understanding of the ever-self-differentiating drive of American journals. *Foreign Affairs* and *World Politics*, mentioned above, both have very large circulations, in part because of their wide coverage of everything pertaining to US foreign affairs. Besides these two, *International Organization* is the most highly ranked journal in the field. It focuses on international political economy, and is intensely theory-driven. It is quintessentially 'American.' *International Security* is another highly regarded journal

focused on international security. It is both theory-oriented and policy-oriented. *International Studies Quarterly* is the most behaviorally oriented of major American international relations journals. It is sustained by a large number of professionally trained academics in behavioral science. *Journal of Conflict Resolution* is the highest-ranking journal with a peace research orientation and behavioral approach. It is multidisciplinary with psychology, formal theory, sociology, political science, and social psychology all well represented. *International Studies Review* focuses on critical review essays on international relations, with emphasis on multidisciplinary contributions, which its editor believes are not well represented in either *International Organization* or *International Studies Quarterly*.

Outside the USA, the following journals stand out: *Review of International Studies*, *European Journal of International Relations*, *Journal of Peace Research*, and *International Relations of the Asia-Pacific*. *Review of International Studies* focuses on theories and historical events of international relations. It presents a good mix of theory, history, and philosophy. *European Journal of International Relations* focuses on theoretical and philosophical analysis and reflections on European and international affairs. *Journal of Peace Research* is a highly respected peace-oriented journal with a predominantly behavioral orientation. *International Relations of the Asia-Pacific* is a new journal focusing on the Asia-Pacific area. It covers contemporary events and actors in the region from the points of view of theory, history, and policy.

The above comparison of major journals conveys a sense of the various components of international relations studies. On the other hand, those products of area studies (whether it is single-country-focused or comparative) are accommodated in a very different set of journals. They are either flagship journals of national political science associations (and their equivalents) or area-studies-focused journals. The former includes *American Political Science Review*, *American Journal of Political Science*, *Political Studies*, *British Journal of Political Science*, *European Journal of Political Research*, *Comparative Politics*, *Comparative Political Studies*, *Government and Opposition*, *Asian Journal of Political Science*, and *Japanese Journal of Political Science*.

A quick survey of major journals of international relations and political science seems to give a sense of separation between area studies and international relations mentioned earlier, but the major trends are in fact 'comparative' and 'global' (McDonnell 2000). 'Comparative' is understood to mean that they provide in-depth comparisons of two or more political systems rather than focusing individually on single countries. This is, incidentally, a trend that the United States Social Science Research Council wishes to promote at a time when area studies programs and funding have been steadily disappearing. 'Global' studies aim to treat all politics within a larger framework of co-existence on the planet. This is an inevitable and irreversible trend that the neat distinction between domestic politics and international relations cannot properly tackle. One should gain consolation from the meaning behind the word 'area studies.' Clearly for Americans, Japanese politics is the right subject of area studies, whereas for the Japanese, American politics is the right subject of area studies. American funding on area studies may be receding, whereas some other countries' funding on their respective national politics may be on the rise.

Area studies is an important area of social science research. Its relationship with the field of international relations varies from one country to another. In some countries such as Japan, area studies and international relations more or less go together or even sometimes merge with one another. In other countries such as the USA, area studies and international relations are not regarded as overlapping subject areas.

See also: Area and International Studies in the United States: Institutional Arrangements; Area and International Studies in the United States: Intellectual Trends; Area and International Studies in the United States: Stakeholders; Globalization: Political Aspects; International Relations: Theories

Bibliography

American Political Science Association (APSA). APSA welcomes new organized sections http://www.apsanet.org/new/sections.cfm, October 30, 2000

Baylis J, Smith S 1997 *The Globalization of World Politics*. Oxford University Press, Oxford, UK

Bull H 1977 *The Anarchical Society: A Study of Order in World Politics*. Macmillan, London

Carr E 1939 *The Twenty-Years' Crisis*. Macmillan, London

Eckstein A 1971 *Comparison of Economic Systems: Theoretical and Methodological Approaches*. University of California Press, Berkeley, CA

Eckstein H, Apter D E (eds.) 1963 *Comparative Politics: A Reader*. Free Press, New York

Fukuyama F 1996 *Trust: The Social Virtues and the Creation of Prosperity*. Free Press, New York

George J 1994 *Discourses of Global Politics*. Lynne Riener, Boulder, CO

Hall R 1948 *Area Studies: With Special Reference to Their Implications for Research in the Social Sciences*. Committee on World Area Research Program, Social Science Research Council, New York

Held D et al. 1999 *Global Transformations*. Polity Press, Cambridge, UK

Hinsley F H 1986 *Power and the Pursuit of Peace: Theory and Practice in the History of Relations among States*. Cambridge University Press, Cambridge, UK

Hoffmann S 1977 An American Social Science: International Relations. *Daedalus* **196**: 41–60
Inoguchi T 1999 Peering into the future by looking back: the Westphalian, Philadelphian and anti-utopian paradigms. *International Studies Review* **1**: 173–91
Inoguchi T 2001 *Global Change: A Japanese Perspective*. Palgrave, New York
Inoguchi T 2000 Social capital in Japan. *Japanese Journal of Political Science* **1**: 73–112
Inoguchi T, Bacon P 2001 The study of international relations in Japan: toward a more international discipline. *International Relations of the Asia-Pacific* **1**: 1–20
Jones E 1979 *The European Miracle*. Cambridge University Press, Cambridge, UK
Katzenstein P, Keohane R, Krasner S (eds.) 1998 *Exploration and Contestation in the Study of World Politics*. MIT Press, Cambridge, MA
Kesselman M, Krieger J, Joseph W 1999 *Introduction to Comparative Politics: Political Challenges and Changing Agendas*, 2nd edn. Houghton Mifflin, Boston
Landes D 1998 *The Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor*. Norton, New York
Lasswell H et al. (ed.) 1980 *World Revolutionary Elites: Studies in Coercive Ideological Movements*. Greenwood, New Haven, CT
Latham M 2000 *Modernization as Ideology: American Social Science and 'Nation Building' in the Kennedy Era*. University of North Carolina Press, Chapel Hill, NC
Lipset S M 1981 *Political Man: The Social Basis of Politics*. Johns Hopkins University Press, Baltimore, MD
Liska G 1977 *Quest for Equilibrium*. Johns Hopkins University Press, Baltimore, MD
McDonnell M 2000 Critical forces shaping social science research in the 21st century. Paper presented at the seminar on Collaboration and Comparison: Implementing Social Science Research Enterprise. Tokyo, Keio University, March 30, 2000, jointly sponsored by the Social Science Research Council and the Center for Global Partnership
Morgenthau H 1978 *Politics among Nations*. Knopf, New York
Murdoch G P 1981 *Atlas of World Cultures*. University of Pittsburgh Press, Pittsburgh, PA
Packenham R 1973 *Liberal America and the Third World Political Development: Ideas in Foreign Aid and Social Science*. Princeton University Press, Princeton, NJ
Putnam R 1994 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ
Putnam R 2000 *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster, New York
Rose R, Shin D C 2001 Democratization backwards: the problem of third-wave democracies. *British Journal of Political Science* **30**: 331–54
Rostow W W 1991 *The Stages of Economic Growth: A Non-Communist Manifesto*. Cambridge University Press, Cambridge, UK
Sachs J 1993 *Macroeconomics in the Global Economy*. Prentice Hall, New York
Thurow L 1997 *The Future of Capitalism: How Today's Economic Forces Shape Tomorrow's World*. Penguin, Harmondsworth, UK
Weaver O 1998 The sociology of a no so international discipline: American and European developments in international relations. *International Organization* **52**: 687–727
Wight M 1991 *International Theory: The Three Traditions*. Leicester University Press, Leicester, UK
Wolfers A 1962 *Discord and Collaboration: Essays on International Politics*. Johns Hopkins University Press, Baltimore, MD
Wright Q 1942 *A Study of War*. University of Chicago Press, Chicago

T. Inoguchi

Area and International Studies: Law

Soon after the beginning of twentieth century, the Carnegie Endowment for International Peace concluded a report on the teaching of international law in the United States by suggesting that although data on student enrollment in such offerings 'probably convey a favorable impression with respect to the extent to which International Law is taught in the … United States … [a] closer examination … will, however, make it clear that a relatively small number of students actually take the courses offered' (Carnegie 1913). Shortly before the twentieth century's end, a study based on a survey conducted by the American Bar Association subtitled 'Plenty of offerings, but too few students,' concluded that '[w]hile international law offerings have exploded since the 1960s, the percentage of students taking these courses has remained relatively constant' (Barrett 1997). Although US law school offerings in international law are just one facet of the broader question of the relationship between area and international studies and law, these two studies, conducted, respectively, prior to World War I and following the Cold War, suggest that the relationship has been a complex one, surprising at times even to those engaged professionally in the intersection of these different fields.

After a brief discussion of definitions, suggesting the capaciousness of the relevant terms, this article examines the relationship of area and international studies to law, with particular attention being paid to the serious questions that efforts at integrating them raise about the nature of each field of inquiry. It argues that one cannot appreciate fully the history of such efforts, nor the directions in which they might proceed, without taking full account of the tensions that exist between the particularity in which area studies is grounded, and the purported universality of law.

1. *The Relevant Terms*

There is no single agreed-upon definition of what constitutes either of the terms central to this article, be it from a purely scholarly or a more practical perspective. This article treats area and international studies as the intensive cross-disciplinary study of a

region or of the global system that is centered typically in history, language, literature, and/or the social sciences, and regards law as that field of inquiry concerned with the rules, formal and informal, that societies create at a variety of levels (such as the national, subnational, international, and transnational). Its principal emphasis will be on ways in which, within the legal academy, the former subject has—and has not—informed the latter, as manifested most concretely through foreign, comparative, and international legal studies, although it will also discuss briefly the treatment within area studies of law.

2. *A Selective History*

At least in terms of concrete indices, the period since the conclusion of World War II, and especially the last three decades of the twentieth century, witnessed a marked increase in attention to area and international studies in the legal academy. In the United States, the extent and range of legal scholarship that might be said to incorporate some dimension of area and international studies has grown enormously at both elite and other institutions. So, too, have the number of specialized law journals, dedicated research centers, course offerings, foreign visiting faculty, exchange programs and other opportunities for study abroad, and the like. For example, there are today scores of journals focused on international, comparative, and foreign law, and associated issues, few of which can trace their history as far back as the 1960s (Crespi 1997).

Western European legal education has taken on a decidedly more comparative and international, flavor, especially as concerns issues raised by European integration. In its most pronounced form, this is leading to academic programs (such as those of the European University) and scholarship intended to create 'European,' rather than national, law and lawyers.

In East Asia, the idea of university legal education from its outset approximately a century ago drew heavily on foreign and particularly German and French civil law models. This arguably has imparted a significant, if not always fully acknowledged, area and international studies tinge to legal studies and scholarship. More recently, the growing global concern with foreign and international legal studies has been accompanied in many parts of East Asia by a conscious attempt by some to recast the legal academy (and profession) along so-called American lines. This has been evidenced for instance, by an emphasis in pedagogy on what are said to be practical, problem-solving skills and enhanced student participation, as opposed to more abstract, doctrinally focused lectures. And throughout much of the developing world, particularly beyond universities with an Islamic mission, there has been an upswing in scholarly and curricular concern with international issues, if not with area studies more generally.

Potential explanatory factors abound. Perhaps most significantly, global economic integration, as manifested in the expansion of international trade (at a postwar rate four times that of overall economic growth), of foreign direct investment, and of the transborder flow of capital, technology, information, and personnel has generated changes in legal institutions, law, and the legal profession warranting greater attention in the legal academy to area and international studies, broadly defined. With respect to institutions, for example, the European Coal and Steel Community and the General Agreement on Tariffs and Trade have become the European Communities and the World Trade Organization. In the process, each has expanded substantially in scope, membership, and global importance, building an increasingly elaborate jurisprudence and institutional structure worthy of serious study by legal scholars that many would agree is abetted by familiarity with fields of international studies such as international relations and international institutions, and of studies of at least some areas. As regards the law itself, the growing fragmentation of production across national borders and the concomitant tendency to see the corporate form as plastic, rather than fixed (in the sense, suggested by the new institutional economics, of readily adding and shedding functions) has, for example, heightened scholarly, as well as more practical, interest in contract law in both transnational and various foreign national settings. And the legal profession itself also reflects this phenomenon. This is evident not only in the much publicized growth of so-called mega law firms with hundreds of attorneys scattered in dozens of satellite offices circling the globe, but increasingly, even for those lawyers anchored firmly in their domestic legal setting (still the vast majority worldwide)—all 'justifying' greater concern with the foreign and international in legal studies.

There is much more involved, however, than global economic integration. Politically, the second half of the twentieth century witnessed a marked expansion of bodies of law—such as international human rights and international environmental law (with close to 200 international agreements in the latter discipline)—that barely existed as formal doctrine prior to World War II. And the end of the Cold War, the collapse of apartheid, and the changes under way in China, together the associated rise of legal development as an instrument of broader developmental work (of the type sponsored by institutions such as the United Nations Development Programme) and of foreign policy more generally, have all drawn academics from granting, recipient, and other countries into law reform work that highlights the significance of area and international studies (Alford 2000). Demo-

graphically, the professoriate in both the developed and developing world has changed in ways that would seem to make it potentially more receptive to incorporating area and international studies in its work. The percentage of legal academics in nations such as the US with serious training in a social science has expanded in recent years, as has the number of their developing-nation counterparts educated abroad (with, for instance, the master's and doctoral programs at most American law schools comprised predominantly of foreign students). And, finally, there is a need to take account of broader academic institutional considerations, as the funding that governments, foundations, businesses, and others have provided for area and international studies has not gone unnoticed by law schools, much as the desire of universities to accentuate their international character via exchange programs and the like has increasingly reached professional education.

Much of the preceding discussion has focused on the impact of area and international studies on law (perhaps a cost of the author's principal disciplinary affiliation), but we ought not to neglect the converse. Growing legalization requires that at least some area and international specialists be attentive to law in ways that may not have previously been the case. As both the substantive doctrinal rules and the dispute resolution processes of regional and multilateral bodies become more elaborate and have greater effect (as, e.g., the WTO's more adjudicatory mode of settling disputes replaces the GATT's more negotiation oriented format), scholars whose principal concerns lie in fields such as area studies, diplomacy, or international economics increasingly need to pay the law greater heed. One would be hard put, for instance, to appreciate changes in the Mexican polity, or to assess the economic implications of intellectual property protection without some understanding, respectively, of the North American Free Trade Agreement and the Trade Related Intellectual Property agreement of the WTO's Uruguay Round. Somewhat analogously, the increasing transformation of what were once moral or political claims into rights with at least some potential of legal enforcement is leading not only students of international affairs, but also growing numbers of philosophers, historians, and others having an area focus into the language and argumentation of law, as evidenced quite graphically by the international dialogue between scholars of Confucianism and human rights launched by the distinguished sinologists William Theodore deBary of Columbia University and Wei-ming Tu of Harvard University (deBary and Tu 1998). And law has been critical to the blossoming since the 1980s of fields such as social history and cultural studies, among others, given how valuable a source legal materials have proven to be for primary data on the lives of ordinary citizens and efforts at state control (leading, for instance, to a recasting of the conventional wisdom that major East Asian societies had an aversion to formal legality) (Huang 1996, Macauley 1998).

3. *Still at the Margins*

As suggested by the quotations in the introduction to this article, for all the impressive growth in foreign, comparative, and international legal studies, and notwithstanding the rationale for law taking area and international studies seriously, at the dawn of the twenty-first century, the notion of 'plenty of courses, but too few students,' and the more general marginality it implies remain apt, at least as concerns American legal academe. Each of the major schools of legal thought prevalent in the US—law and economics, critical legal studies, and more traditional doctrinalism—in its own way, offers a universalist paradigm (of, to put it crudely, economic analysis, critical thought in the manner of deconstructionism, and what, for lack of a better term, is often described as 'thinking like a lawyer,' respectively) in which there is little room for, or for the most part, even interest in the particularism of area and international studies. Accordingly, only rarely does what is seen as cutting-edge mainstream scholarship touch upon foreign, comparative, or international law, or life beyond the US more broadly. Even when it does, typically it provides an application of, or otherwise confirms, established signature theoretical positions. Moreover, the focus of most legal scholarship on formal rules may lead even comparative legal scholars to lose sight of one of the key lessons to be gleaned from area studies—namely, that the extent to which public, positive law is relied upon to address certain concerns may vary enormously between societies.

Much the same pattern is replicated in other dimensions of US legal academe. The most prestigious and widely read of law journals (such as the *Harvard Law Review* or *Yale Law Journal*, and other 'flagship' general reviews edited by students at the leading law schools) address only very sporadically foreign or international subject matter. Such concerns are relegated to more specialized journals (the advent of which, arguably, has had the unintended consequence of isolating work in these fields from a general audience). The first-year law school curriculum, which remains substantially what it was in the late nineteenth century at most American schools, incorporates foreign, comparative, or international law marginally, if at all, and is not accompanied by any requirement that students in their final two years do coursework in these areas. As a result, most of the enrollment in such offerings consists of 'repeat players' who, arguably, are those who are already most open to perspectives other than that of their own nation, while at least two-thirds and perhaps as many as three-quarters of all law students, even at institutions that extol their inter-

national programs, graduate with no curricular exposure to these areas (Barrett 1997). Faculty members specializing in foreign or comparative law (who often form one-person 'departments' responsible for whole continents, if not most foreign legal systems) are, by their own statement, on the periphery, rather than at the center, of their institution's intellectual life (Reimann 1996). Increasingly, they are called upon to generate a fair degree of their own financial support, typically (and unfortunately in terms of its potential for at least the appearance of conflicts of interest) from the very nations that are the subjects of their scholarship. And in a similar vein, in a manner anomalous in American graduate education, the overwhelming majority of foreign students are directed into a one-year master's program that is neither their institution's principal or most prestigious degree, nor intended primarily to lead to further academic or professional study.

Comparative law scholars in particular have tended to respond to this marginality with intimations that those who set intellectual trends in legal academe are less worldly than they might be (Berman 1989), expressions of concern about isolation (Merryman 1999), and even calls for the end of comparative law as a distinct field (Reimann 1996). Yet, arguably, area and international specialists bear some responsibility for their plight. However strong it might be on its own terms, comparative legal scholarship is rarely viewed as making major contributions of a broader theoretical nature while all too frequently it is seen as inaccessible to nonspecialists, and more than occasionally so obscure as to deter the type of engagement that over time builds an ongoing scholarly dialogue. Moreover, recent critiques from the vantage points of both law and economics, and critical legal studies suggest that some comparative law scholarship suffers from a disingenuousness or inattention to issues of power, invoking cultural difference in an ill-defined manner that may obscure, rather than clarify (Ramseyer and Nakazato 1998, Kennedy 1997), and, according to some, paying insufficient heed to the political implications of its treatment of the 'other' (Riles 1999). International law scholarship may speak to a broader range of readers, focused as it has done increasingly on more readily comprehended projects of international governance and regulation. But in the minds of some observers, it may be insular in its own way, containing, as noted with respect to American work, 'very few references to non-American writings, even in English, let alone in French or German' (not to mention other languages) (Gross 1989). Curricular offerings in comparative and foreign law face a tension between striving to be comprehensive, and endeavoring to avoid being superficial. And notwithstanding the heightened attention to legal issues in area studies, many historians and social scientists continue to portray the law in excessively formalistic, overly literal terms, evidencing little appreciation of scholarly debates about the nature of law that would suggest the malleability with which law might be interpreted (Alford 1997).

There is no shortage of proposals in this 'age of globalization' as to how better to accommodate the foreign, comparative, and international with legal studies. And yet, even the most sophisticated of these may not grasp the heart of the problem. Although law might seem a grounded discipline, given how anchored it would appear to be in the soil of a single nation, ultimately it has universalist presumptions and aspirations (at least with respect to methodology), which do not accommodate comfortably the particularism of area and international studies as they (and especially the former) have for the most part been practiced. The effort, in a sense, is akin to that of trying to bring disciplines such as philosophy and anthropology together, leading some to suggest that before legal academe can engage the history and sociology of law in other societies effectively, it needs first to do so more thoroughly for its own society (law, having as a discipline, been surprisingly inattentive to the ways in which the rules of which it is comprised actually work in practice).

Of course, the universalist presumptions and aspirations of American law, by way of example, are at some level grounded in the experience of a particular society, even if not always fully appreciated as such by their proponents. Particularly in the current era of triumphalism, with its suggestions of an inevitable convergence of law and other institutions along what are said to be American lines, this blurs the distinction between what might genuinely be termed universal and what might be more specific to the United States, thereby complicating serious efforts to incorporate more fully other area perspectives into thinking about law in general (Ackerman 2000). These difficulties are, arguably, further exacerbated by the nature of law as both an academic and a practical discipline, housed in the United States (and increasingly elsewhere) in a professional school. This means that most faculty members do not have formal academic training beyond their professional degrees, may be participants in that about which they are writing, and, principally, are teaching practically-oriented students who will work chiefly within a single jurisdiction.

To acknowledge the foregoing challenges is not necessarily a counsel of despair. Spurred by the globalizing considerations discussed above, legal academe, particularly beyond the United States, is taking fuller account of foreign, comparative, and international law, even as newer work in area and international studies, in the manner examined in the *Area and International Studies in the United States: Intellectual Trends*, is imbuing what has long been a field rich in description with a keener appreciation of the importance of theoretical engagement.

See also: Area and International Studies: Political Economy; International Law and Treaties; Law:

History of its Relation to the Social Sciences; Legal Education

Bibliography

Ackerman B 2000 The new separation of powers. *Harvard Law Review* **113**: 633–729

Alford W 1997 Law, law, what law? Why Western scholars of Chinese history and society have not had more to say about its law. *Modern China* **23**: 398–419

Alford W 2000 Exporting the pursuit of happiness. *Harvard Law Review* **113**: 1677–715

Anon 1989 The state of international legal education in the United States. Special Feature. *Harvard International Law Journal* **29**: 239–316

Barrett J A 1997 International legal education in US law schools: Plenty of offerings, but too few students. *The International Lawyer* **31**: 845–67

Berman H 1989 Interview. *Harvard International Law Journal* **29**: 240–5

Carnegie Endowment for International Peace 1913 *Report on the Teaching of International Law in the Educational Institutions of the United States*. Carnegie Endowment for International Peace, Washington DC

Crespi G S 1997 Ranking international and comparative law journals: A survey of expert opinion. *The International Lawyer* **31**: 867–85

deBary W T, Tu W 1998 *Confucianism and Human Rights*. Columbia University Press, New York

Gross L 1989 Interview. *Harvard International Law Journal* **29**: 246–51

Huang P 1996 *Civil Justice in China: Representation and Practice in the Qing*. Stanford University Press, Stanford, CA

Kennedy D 1997 New approaches to comparative law: Comparativism and international governance. *Utah Law Review* : 545–637

Macauley M 1998 *Social Power and Legal Culture: Litigation Masters in Late Imperial China*. Stanford University Press, Stanford, CA

Merryman J 1999 *The Loneliness of the Comparative Law Scholar and Other Essays in Foreign and Comparative Law*. Kluwer Law International, The Hague

Ramseyer J M, Nakazato M 1998 *Japanese Law: An Economic Approach*. University of Chicago Press, Chicago

Reimann M 1996 The end of comparative law as an autonomous subject. *Tulane European and Civil Law Forum* **11**: 49–72

Riles A 1999 Wigmore's treasure box: Comparative law in the era of information. *Harvard International Law Journal* **40**: 221–83

W. P. Alford

Area and International Studies: Linguistics

In most of the world, 'you are what you speak,' because national identity is often aligned with linguistic identity. Geopolitical regions are partially defined in terms of language, and the subject matter of area and international studies is embedded in local languages. Despite the importance of linguistic expertise for understanding the peoples of a region and accessing primary material, linguistics is typically regarded as a peripheral discipline for area and international studies, relative to 'core' disciplines such as political science, history, economics, anthropology, sociology, and geography. This peripheral status results from (largely correct) perceptions that linguistics is highly technical and impenetrable, that linguistics is theoretically fractured, and that most linguists in the US are not interested in topics relevant to area and international studies. However there is evidence of renewed linguistic interest in issues of language in the contexts of geography, politics, history, and culture, as well as a commitment to be accessible to other disciplines and language learners.

1. Linguistics and Area and International Studies

Linguistics is directly relevant and beneficial to area and international studies:

(a) when it contributes to understanding the geographical distribution of peoples (by means of typology, dialect geography, historical linguistics, fieldwork, and language planning and intervention);

(b) when it contributes to understanding the different world views of peoples (by means of linguistic anthropology, discourse analysis, literary analysis, and poetics); and

(c) when it contributes to language learning (through the development of pedagogical and reference materials).

Linguistics can also achieve an area or international studies dimension in other endeavors (for example, the development of formal theories) when there is sustained focus on a given language group.

2. A Brief History of Relevant Linguistic Developments

In the early part of the twentieth century (approximately 1900–40), linguistics was dominated by Sapir and Whorf, whose objective was to explore how languages reveal people's worldviews and explain cultural behaviors. This view of language as a direct artifact of the collective philosophy and psychology of a given society was inherently friendly to the goals of understanding nations and their interactions. The Sapir-Whorf emphasis on the relationship between language and its socio-geographical context (later retooled as 'functional linguistics') might have engendered significant cross-disciplinary efforts, but unfortunately, its heyday was largely over before area and international studies became firmly established as academic disciplines.

By the time the US government made its first Title VI appropriations in the late 1950s, a landmark event in the founding and building of area and international studies as known at the end of the 1990s, linguistics had moved on to a fascination with mathematical models that would predominate (at least in the US) well into the 1980s. The theoretical purpose of an algebraic approach to the explanation of grammatical phenomena is to provide a formal analysis of the universal features of language. This theoretical perspective of 'formal linguistics' marginalizes or excludes issues relevant to area and international studies since language context is not considered a primary factor in language form. The relationship of pure math as opposed to applied mathematical sciences (economics, statistics, etc.) is analogous to the relationship between formal and functional linguistics and their relative sensitivity to contextual factors: the objective of both pure math and formal linguistics is analysis independent of context, whereas functional linguistics and applied math make reference to concrete domains (extra-linguistic or extra-mathematical).

The popularity of mathematical models was widespread in the social sciences in the late twentieth century, creating tension between the so-called 'number-crunchers' and area and international studies scholars, and disadvantaging the latter in hiring and promotion. Formal linguistics has played a similar role in the broader discipline of linguistics and yielded a framework that does not focus on language pedagogy or the geographic distribution and differing worldviews of peoples. Formal linguistics has been primarily inspired by the work of Noam Chomsky, whose framework has been successively known as generative grammar, the government binding theory, and the minimalist program. Other important formalist theories include relational grammar and head-driven phrase structure grammar.

Since the 1980s there has been renewed interest in the relationship between language function and language form, known as 'functional linguistics.' Though functionalist approaches are not a retreat into the past, they comport well with pre-Chomskyan theories, enabling linguists to build on previous achievements. Functional linguistics is also more compatible with many linguistic traditions outside the US, especially in areas where Chomsky is not well known (for example the former Soviet Bloc countries, where Chomsky's linguistic work was banned in reaction to his political writings), or in areas where there has been sustained focus on mapping and codifying indigenous languages (such as Australia, Latin America, and the former Soviet Union). The most significant functionalist movement is known as cognitive linguistics, and has George Lakoff and Ronald Langacker as its primary proponents. Cognitive linguistics has rapidly gained popularity in Western and Eastern Europe, in the countries of the former Soviet Union, Japan, and Australia. In addition to cognitive linguistics, many traditional sub-disciplines of linguistics continue their commitment to functionalist principles, among them dialectology, discourse analysis, historical linguistics, and typology. These traditional endeavors and cognitive linguistics bear a mutual affinity since both focus on language-specific data (as opposed to language universals). Because the context of language and its role in meaning are central to the functionalist view of linguistics, the potential contribution of functional linguistics to area and international studies is great. And because functionalist linguistics tends to avoid intricate formal models, it is more accessible to specialists in other disciplines, and its results are transferable to language pedagogy.

At the time of writing, formalist and functionalist linguistics are engaged in an often-antagonistic competition. (For further information on the history and present state of formal vs. functional linguistics, see *Generative Grammar*; *Functional Approaches to Grammar*; *Cognitive Linguistics*; *Sapir-Whorf Hypothesis*; and Newmeyer 1998, Lakoff 1991, Croft 1998).

3. Linguistic Contributions to Area and International Studies

Many time-honored endeavors of linguists (investigation of unknown languages, research on the relations among languages, preparation of descriptive and pedagogical materials) yield valuable results for area and international studies. Relevant methods and results are discussed under three broad headings below.

3.1 Contributions to Understanding the Geographic Distribution of Peoples

Linguists use the empirical methods of fieldwork to discover the facts of existing languages, recording features of phonology (language sounds), morphology (shapes of words), syntax (grammatical constructions), and lexicon (meanings of words). Investigation of how these features vary through space is known as dialectology, and each line on a map corresponding to one of these features is known as an isogloss. Isoglosses usually correspond to geographic (mountains and rivers), ethnic (often religious), or political (more often historical than current) boundaries. Despite the use of scientific discovery procedures, linguists do not have an operational definition for language as opposed to dialect.

Language is often closely tied to national identity, and the cohesiveness of a given speech community is often more dependent upon the sociopolitical imagination of speakers than on the number of features they share or the number of isoglosses that divide them. Chinese, for example, is a remarkably diverse linguistic entity that elsewhere in the world would probably be considered a family of related languages.

There is only a gradual cline rather than a bundle of isoglosses between Macedonian and Bulgarian, and the speakers do not agree on the status of their distinction: Bulgarians believe Macedonians are speaking a 'Western Bulgarian dialect,' whereas Macedonians assert they are speaking a distinct language. Minor dialectal differences are sometimes amplified for political gain. The various ethnic groups in the former Yugoslavia that speak the language historically known as Serbo-Croatian have used relatively minor distinctions as flags of national identity, claiming distinct languages in order to fracture the country and justify seizure of territory.

The aim of historical linguistics is to discover relationships among languages. Historical linguistics uses two methods to arrive at a description of historical changes and their relative chronology. The first method is internal reconstruction, which compares linguistic forms within a single language in an attempt to reconstruct their historical relationships. The second is the comparative method, which compares cognate forms across related languages in an attempt to arrive at how modern forms developed from a shared proto-language. Any given language change usually spreads gradually across the territory of a language. Over time this yields isoglosses, the primary material of dialect geography, and these isoglosses reflect the relative chronology of historical changes. Thanks to historical linguistics, we know a lot about how languages are related to one another, information valuable for understanding the history, migrations, and ethnic backgrounds of peoples.

Despite considerable removal in both time and space, linguistic relationships continue to inspire political and other behavior. During the Cold War Ceaucescu's communist regime raised money by selling babies for adoption to infertile French couples; this plan played upon a desire to procure genetically related offspring, since both Romanian and French are Romance languages. The notion of Slavic unity was used to justify much of the Warsaw Pact, and after the break-up of the Soviet Union Solzhenitsyn suggested that the Belarussians and Ukrainians join Russia to form a country based upon the relation of their languages (since Belarussian, Ukrainian, and Russian constitute the East Slavic language subfamily).

Languages in contact can influence one another regardless of any genetic relation. As a result, groups of contiguous languages tend to develop shared features, known as areal phenomena. The languages of the Balkans include a variety of South Slavic and other very distantly related Indo-European languages, among them Serbo-Croatian, Albanian, Macedonian, Romani, Greek, and Bulgarian. Together they share certain features, pointing to a greater unity of the Balkans that transcends their diverse heritage. Sustained or intensive language contact can result in the creation of new types of languages.

This takes two forms: one is 'creolization,' two or more languages melded into a new language; and the other is 'pidginization,' a simplified version of a language (often borrowing words from another language). An example of a creole is Papiamentu, a mixture of Spanish, Portuguese, Dutch, and indigenous languages, spoken in the Dutch Antilles; pidgin English is a language of trade created in Asia and the South Pacific for communication between indigenous peoples and outsiders. A further type of linguistic coexistence is 'diglossia,' the use of one language for spontaneous oral communication, but another language for formal and literary expression. For example, after two centuries of German domination removed Czech from the public arena, the Czech National Revival resurrected a literary language from an archaic Bible translation. As a result, there is a significant gap between spoken Czech and the Czech literary language.

Typology compares the structure of both related and unrelated languages. Typology suggests a positive correlation between the severity of geographic terrain and the density of linguistic diversity (Nichols 1990). Perhaps the best example is the Caucasus mountain region, arguably a part of the world with more languages per unit of inhabitable surface area than any other, predictably matched by a high level of ethnic and political tension. Global linguistic diversity is threatened by the phenomenon of language death, and it is predicted that 90 percent of the world's languages will disappear by the end of the twenty-first century (Krauss 1992, p. 7). Endangered languages are those of minorities who must acquire another language (of a politically dominant group) in order to survive. Protection of minority rights requires protection of minority languages, and can entail fieldwork and the preparation of pedagogical materials. Another significant language-planning issue involves the status of languages in the Central Asian republics of the former Soviet Union. After decades of Russian domination, the majority languages of these new countries are being elevated to the status of official literary languages.

3.2 Contributions to Understanding Behaviors and Worldviews

People use language to describe their experiences of reality and to make hypothetical projections from those experiences. Human experience is mediated by both perceptual mechanisms and conceptual systems. Though much of human perceptual ability is universal, input can be both ambiguous and overly detailed. Perception provides much more opportunity for distinction than any one language can codify in its grammar or any human being can meaningfully attend to. The highly textured world of perception does not suggest any unique strategy for carving nature at its

joints. Thus, perception is inseparably joined with conceptual decisions concerning what to ignore and what goes with what (Talmy 1996 has coined the term 'ception' to describe the concurrent operation of perception and conception). If, as functional linguists believe, linguistic categories are conceptual categories, and can be specific to a given language, then linguistic categories should reveal important facts about how people understand and interact with their world. Time, for example, is a phenomenon that human beings do not have direct experience of, because humans perceive only time's effects on objects and events. It is therefore possible to conceive of time in different ways, and a plethora of tense and aspect systems present artifacts of varying conceptions of time. The relationships that exist between beings, objects, and events can likewise be understood in many different ways; a testament to this is the variety of case systems and other means that languages use to express relationships.

Language is the essential vehicle of a number of cultural phenomena, ranging from the daily rituals of oral communication, the subject of discourse analysis, through the artistic use of language that is the subject of literary analysis and poetics. Linguistic analysis of use of metaphor and poetic structure can be valuable in interpreting literary culture.

3.3 Contributions to Language Pedagogy and Reference

Linguistic expertise is essential for the production of effective language textbooks, reference grammars, and dictionaries, tools that enable area and international studies scholars to gain language proficiency. Academic promotion procedures fail to recognize the exacting scholarship and creative thinking that pedagogical authorship and lexicography require. In the US, there is not enough of a market for publications in languages other than French, Spanish, and German to provide financial incentive to take on these tasks. As a result, linguists are reluctant to author textbooks and reference works, and materials for lesser-taught languages are usually inadequate or absent. Faced with financial crises in the 1990s, some colleges and universities acted on the popular myth that native ability is the only qualification needed to teach language, and replaced language professionals with part-time and/or adjunct native speakers. Although it now has competition from functional linguistics, formal linguistics continues to dominate the field, and its findings are not generally relevant or transferable to pedagogy and lexicography (since this is not the aim of formal linguistics). Collectively, academic bias, small market share, de-professionalization of language teaching, and theoretical focus greatly reduce linguists' impact on language pedagogy and reference materials.

For detailed treatment of the above topics and for further references, see *Linguistic Fieldwork*; *Dialectology*; *Historical Linguistics*; *Internal Reconstruction*; *Comparative Method*; *Areal Linguistics*; *Pidgin and Creole Languages*; *Diglossia*; *Linguistic Typology*; *Language Endangerment*; *Language Policy*; *Language and Literature*; *Language and Poetic Structure*.

4. Probable Future Directions of Theory and Research

Internet technology provides instantaneous access to vast quantities of language data, an unprecedented resource that linguists are only beginning to use. A large number of national language corpora, even for lesser-taught languages, are now available on the Web. There are also search tools, such as google.com, that are extremely useful to linguists researching the use of forms and constructions (at least in languages with Latin alphabets; despite the advent of Unicode, fonts continue to pose some of the most intractable technological problems linguists face). The sheer quantity and availability of language-specific data seems guaranteed to facilitate research relevant to area and international studies. Perhaps the best example of how corpora and technology can be integrated into linguistic research is Charles Fillmore's FrameNet, a digital dictionary of the grammatical constructions of a language, based on a language corpus. Originally developed for English, FrameNet is now being expanded to other languages, and promises to be a valuable tool for linguistics and language pedagogy. Perhaps projects like these will raise awareness of the need for lexicographical and other reference materials, and enhance the prestige of such endeavors.

Funding always plays a crucial role in guiding research trends. The US Department of Education and the National Science Foundation are the greatest sources of support for linguistic research, and both agencies fund projects relevant to area and international studies. While linguistics plays merely a supportive role in US Department of Education Title VI National Resource Center grants, it is a central player in Title VI Language Resource Center (LRC) grants. There is a new trend for LRC grants to focus on a region of the world. In 1999 three LRC grants were awarded for projects with areal focus: the National East Asian Languages Resource Center at Ohio State University, the National African Languages Resource Center at the University of Wisconsin, Madison, and the Slavic and East European Language Resource Center at Duke University-University of North Carolina, facilitating the creation of technologically enhanced pedagogical materials and area-specific linguistic research. The launching of LRCs focused on world regions is a major step forward in fostering linguistic projects that are responsive and responsible to area and international studies. Continued attention and funding may enable the relationship between area and international studies and

linguistics to realize its potential, much of which today remains untapped.

See also: Areal Linguistics; Cognitive Linguistics; Comparative Method; Diglossia; Functional Approaches to Grammar; Generative Grammar; Historical Linguistics; Internal Reconstruction; Language and Literature; Language and Poetic Structure; Language Endangerment; Language Policy; Linguistic Fieldwork; Dialectology; Linguistic Typology; Pidgin and Creole Languages; Sapir–Whorf Hypothesis; Language and Gender; Linguistics: Overview

Bibliography

Croft W 1998 What (some) functionalists can learn from (some) formalists. In: Darnell M, Moravcsik E (eds.) *Functionalism and Formalism in Linguistics*. J Benjamins, Amsterdam, Vol. 1, pp. 85–108

Krauss M 1992 The world's languages in crisis. *Language* **68**: 4–10

Lakoff G 1991 Cognitive versus generative linguistics: How commitments influence results. *Language and Communication* **11**: 53–62

Newmeyer F J 1998 *Language Form and Language Function*. MIT Press, Cambridge, MA

Nichols J 1990 Linguistic diversity and the first settlement of the new world. *Language* **66**: 475–521

Talmy L 1996 Fictive motion in language and 'ception'. In: Bloom P, Garret M F, Peterson M A (eds.) *Language and Space*. MIT Press, London, pp. 211–76

L. A. Janda

Area and International Studies: Political Economy

Modern 'political economy' explores relationships among economic and political organizations (e.g., states, corporations, unions), institutions (e.g., laws and practices regulating trade and competition), policies (e.g., restrictions on international capital mobility), and outcomes (e.g., rates of economic growth, political regime stability). Political economists differ along several important dimensions, as discussed below. This essay offers an overview of the evolution of competing versions of political economy, and their relationship to Area and International Studies, since World War Two.

1. The Fordist Moment

The classical political economists of the late eighteenth and nineteenth centuries—Smith, Ricardo, Malthus, Marx, and J. S. Mill—addressed fundamental questions such as the appropriate economic role of the state, the implications of trade liberalization for the economic fortunes of different economic classes, the ecological constraints on continuous economic growth, and the economic and political contradictions of different modes of production, including capitalism.

In the first quarter-century after World War Two, some of these questions remained at the center of debates between modernization theorists (Rostow 1960) and dependency theorists (Cardoso and Faletto 1979) who argued about the logics of, and possibilities for, economic and political development in what the Cold War framed as the Third World. Marxist versions of political economy became the new orthodoxy in the Second World, which soon encompassed most of Eastern Europe and much of Asia. In the First World, however, most students of market dynamics within Economics departments began to abandon political economy approaches. Prior to World War Two, when institutional economics remained the dominant tendency in the economics departments of the United States, this break with the core assumptions and research agendas of political economy had gone furthest in the United Kingdom. However, after the war, British neo-classical microeconomics and Keynesian macroeconomics gained ground rapidly in the United States. In the Cold War context, the dominant theoretical orientations of the US hegemon exerted a powerful gravitational pull on social science in the academies of the First World.

The new mainstream economists had a much narrower intellectual agenda. At the micro level, they drew on the pioneering work of Marshall and Pareto in an effort to demonstrate by formal, mathematical means the superiority of competitive markets as efficient allocators of resources. Arguments for trade liberalization and critiques of most forms of state intervention in market allocation processes were developed in this spirit. At the macro level, in First World economies, the new mainstream drew on Keynes in an effort to theorize how best to employ fiscal and monetary policies to reduce the amplitude of business cycle fluctuations. There were significant tensions between these micro-and macro-economic agendas, but they only became salient toward the end of this period.

Most of the new mainstream economists—whether micro or macro in focus—sought quasi-natural laws governing market dynamics regardless of time and place. They paid little attention to the political and institutional parameters within which markets existed or to the balance of power among social forces that shaped these parameters. There was an irony here. The divorce between politics and economics in First World academic economics was possible because a new kind of political economy—sometimes called 'Fordism' (Lipietz 1987)—was developed after World War Two.

The institutional and political details of Fordist regulation varied from country to country but everywhere it represented a fundamental shift away from wage regulation through competitive labor markets maintained by the repression of worker rights, to form democratic unions and engage in collective bargaining. Fordist institutions linked national real wage growth to the expansion of national labor productivity through some combination of collective bargaining (e.g., pattern bargaining in the United States) and state regulation (e.g., minimum wages) (Piore and Sabel 1984). Fordist regulation generated higher rates of economic growth and distributed the gains from that growth more broadly among the workforce than any form of economic regulation before or since (Marglin and Schor 1990). These successes contributed to the depoliticization of economic policy, which could more easily be seen as an administrative matter to which there were technical answers. This facilitated the shift away from political economy's focus on the reciprocal relationship between political power and economic outcomes.

Non-Marxist versions of political economy remained a significant current within the comparative politics subdiscipline of political science in this period. In the Cold War struggle, a great deal of public research funding was made available to those pursuing such studies (Gendzier 1985). Area Studies encompassed the countries of the Second and Third Worlds, where the influence of Marxist political economic analysis was strong among state elites, academics, and organizations such as unions and co-operatives. So First World academic analysts found it necessary to engage questions of class power, organization, and institutions in studying these countries. To this task most brought Weber and Durkheim, as interpreted and synthesized by Talcott Parsons, under the rubric of modernization theory (Leys 1986).

Area Studies thus helped to preserve political economy when it was marginalized in the Economics departments of the First World countries in which it had originated. In turn, political economy offered a coherent basis for distinguishing among different areas. Latin America, for example, made sense as a region to be contrasted with others because most of its states shared a particular kind of political economy. In the nineteenth century, the countries in this region were characterized by primary commodity production for export, and republican regimes that secured their independence from Iberian empires. In the crisis of the 1930s, most of these countries embraced a particular kind of economic development strategy—import-substitution industrialization—that gave rise to parallel economic and political dynamics, including the rise of a significant industrial working class, and the formation of corporatist political systems.

There was no parallel symbiosis between political economy and the International Relations (IR) subdiscipline as it existed in US political science in this period. IR was dominated by international security debates between those who supported a *realpolitik* firmly rooted in a narrow account of national self-interest and those who asserted that international cooperation rooted in shared liberal values was a surer guide to national security and world peace. IR paradigms, particularly as they were formalized by Waltz and his followers, took it as axiomatic that states were highly autonomous from the domestic societies in which they were embedded, at least as regards the formation of foreign policy (Waltz 1959). On this view, the international distribution of state power resources (e.g., concentrated in two rival superpowers versus dispersed more evenly among leading countries organized into alliances), and differences in state elite strategies for realizing their power maximizing objectives, were the main explanations for variations in state behavior and resulting international dynamics (Krasner 1976). States might liberalize trade as part of their grand strategies for enhancing their power relative to their rivals, but there was little reciprocal causality in this model. That is, neither international economic dynamics, nor classes defined in economic terms, had much impact on state goals or strategies.

Beyond IR as practiced in the United States, International Studies in these years was roughly equivalent to diplomatic history. Greater methodological and theoretical eclecticism created more space for recognizing the significance of domestic factors in international relations. A few approached these matters from standpoints that paid greater attention to the kinds of factors highlighted by political economy. Still, most studies of international diplomacy remained in the realm of 'high politics,' and so had only passing contact with the methods and concerns of political economy.

2. *The Neo-liberal Moment*

As the Fordist economic order began to disintegrate in the late 1960s, many argued that both the causes of the crisis and the remedies for it lay in changes in the balance of economic and political power among nations, between labor and capital within nations, or both (e.g., Gourevitch 1986). Economics and economic policy were thus repoliticized as an intense political struggle over how to understand and respond to the crisis of Fordism under way. In this context, rival strands of political economy emerged, each associated with advocates of a different response to the crisis. Critical political economists—a diverse group influenced in varying degrees by Marx, Weber, and Polanyi—were concentrated in political science and sociology departments. This strand of political economy was strongest among area specialists who focused on Latin American, Asia, and Western Europe and among students of peasant rebellion and social rev-

olution. Their diagnoses of the crisis tended to support policies that would reinforce and extend the basic principles of Fordist regulation, or a move beyond capitalism to some form of democratic socialism.

Neoclassical political economists began from the premise that neoclassical accounts of economic dynamics were basically sound, as were capitalist economies. The problem as they saw it was how to develop an equally sound science of political dynamics, a science that would explain why state intervention could seldom improve on market outcomes, even when market failures were acknowledged. Some sought to build a new political science on the 'rational choice' premise that all individuals and organizations are instrumentally rational, self-interested actors (Alt and Shepsle 1990). Others were less programmatic, turning traditional analytic tools to the service of policy goals deriving from neoclassical economics. The first strategy generated theories of 'rentier' states and 'political failure' paralleling the theory of market failure that justified Fordist regulation on efficiency grounds (Kruger 1974, Bates 1988). These analyses, together with neoclassical micro-economic doctrines, provided the intellectual rationale for 'neo-liberal' policy prescriptions—that is, the redefinition as the primary economic role of the state as the creation and maintenance of competitive markets. The second strategy generated (among other things) assessments of strategies for implementing structural adjustment policies successfully in democracies where popular opposition to such policies was widespread. The neoclassical strand of political economy was concentrated in economics and political science departments (particularly in the United States).

The neoclassical approach to political economy was very much in tune with the neo-liberal response to the crisis of Fordism championed by the United States under Ronald Reagan and the United Kingdom under Margaret Thatcher. The third world debt crisis soon facilitated the export of neo-liberal policies, via conditions imposed on debtor nations in return for assistance in restructuring their loans. In this context, neoclassical political economy became the more prominent of the two approaches, particularly among policy élites and in the United States. However, there were important intellectual innovations within both tendencies. The result was a renaissance of political economy analysis in the neo-liberal era, with both the form and the implications of that analysis intensely contested.

World systems theory was an important strand of the critical political economy analysis that emerged in this period (Wallerstein 1976). While influenced by earlier dependency theories, world systems theory was distinctive in two ways: the degree to which it treated the international economy as a system governed by its own systems-level logic, and the degree to which that logic was seen to determine the development possibilities of the nations whose system functions marked them as peripheral or semi-peripheral. World systems theory took hold primarily in sociology departments, where Marx enjoyed more equal status with Durkheim and Weber.

A second strand of critical political economy, with institutional roots in sociology and political science, emerged under the banner of 'bringing the state back in' (Evans et al. 1985). This strand explored the significance of national differences in state characteristics and relations between the state and societal actors, factors that were treated as secondary in most world systems analyses. Within the Third World, critical political economists began to explore the social and political consequences of structural adjustment policies in Africa and Latin America in the wake of the debt crisis (e.g., Bierstecker 1995). There was also increased interest in the role of 'developmental states' in enabling a small number of countries—concentrated in Asia's Newly Industrializing Countries—to escape the travails of the debt crisis and successfully reorient national economies to an export-driven model of industrialization (e.g., Evans 1995). Finally, a literature emerged on the export processing zones created in many countries subject to neo-liberal restructuring, and on the international supply chains that linked them to First World corporate producers and retailers (Gereffi and Korzeniewicz 1994).

A strand of neoclassical political economy also focused on the world system. Most important here was the 'hegemonic stability' theory advanced by the self-styled neo-realists and their neo-liberal interlocutors. These analysts, located within international relations sections of political science departments, mainly in the United States, debated the extent and significance of the decline of US economic hegemony evident by the late 1960s. Hegemonic stability theories asserted that, without disproportionate US economic power, the international monetary system created at Bretton Woods and multilateral trade liberalization under the auspices of GATT would not have been possible (Gilpin 1987). They interpreted the collapse of fixed exchange rates and an alleged shift toward protectionism in the form of 'nontariff barriers,' as evidence that the postwar international economic regimes constructed by the United States were indeed unraveling.

Neo-liberals such as Keohane drew on game theory to argue that states *qua* rational actors might choose to support and extend trade liberalization and other aspects of international regulation out of an enlightened sense of self-interest, even in the absence of a hegemon, under certain conditions (Keohane 1984). Other neoclassical political economists took a different tack, identifying a variety of societal, institutional, and ideological factors that might explain why the US state maintained its trade-liberalizing trajectory in the 1970s and 1980s, despite declining US economic hegemony and rising social costs (e.g., Goldstein 1993). Among neoclassical political economists fo-

cusing on the global South, attention was devoted to explaining the wave of democratization that began in the late 1970s, particularly to possible links between economic liberalization and democratization.

In the 1990s, the end of the Cold War and the acceleration of the economic globalization—by which analysts generally meant increased international trade and capital mobility, and sometimes also neo-liberal policies such as privatization and deregulation—shifted the focus of research and the terms of debate. The concepts of the Second and Third Worlds were rendered obsolete; many analysts began dividing the world into the 'global North' (i.e., rich capitalist democracies) and the 'global South' (i.e., all others). An important debate developed concerning whether there was anything sufficiently novel about the international economy of the 1990s to warrant the use of the term globalization (Held et al. 1999). There was also great interest in the causes and consequences of economic globalization. Neoclassical political economists tended to take a positive view of economic globalization and often treated that the process as natural and/or inevitable. Critical political economists typically saw the shift as the product of power politics within and among nations and its negative affects as more substantial (Cox 1994). There was great interest in whether this new economic order was significantly narrowing state policy autonomy, forcing governments toward a more *laissez-faire* model of economic organization regardless of their political stripe and the preferences of voters (e.g., Rodrik 1997). There was also interest in the implications of economic globalization for the power of organized labor (e.g., Kitschelt et al. 1999). Finally, there was growing interest in the origins and character of organized resistance to the neo-liberal model of globalization, a discussion leavened with the insights of social movement theory (e.g., Castells 1997), as well as more traditional political economy approaches (Arrighi et al. 1989).

These developments had important implications for the evolution of area and international studies. The resurgence of political economy strongly legitimated international studies' supra-national and interdisciplinary character and added another important approach to how such work might be organized. As to area studies, political economy may afford new grounds for drawing area boundaries. For example, since the passage of the North American Free Trade Agreement (NAFTA) in 1993, Canada and Mexico have become much more integrated with the US economy. For many of the questions of interest to political economists, it now makes sense to treat North America as a region to be studied as a unit. This contrasts with the old area studies practice of studying the United States in splendid isolation, Mexico as part of Latin America, and largely ignoring Canada. Similarly, as more East European countries join the European Union (EU), it will become sensible to frame many political economy questions in terms of a new EU region that straddles what were once regions in the First and Second Worlds.

See also: Area and International Studies: Economics; Area and International Studies: International Relations; Area and International Studies: Sociology; Dependency Theory; Development and the State; Development: Socioeconomic Aspects; Globalization: Political Aspects; Nations and Nation-states in History; Political Economy, History of; Political Economy in Anthropology; Political Science: Overview; State Formation; World Systems Theory

Bibliography

Alt J E, Shepsle K A (eds.) 1990 *Perspectives in Positive Political Economy*. Cambridge University Press, New York

Arrighi G, Hopkins T, Wallerstein I 1989 *Antisystemic Movements*. Verso, London

Bates R H (ed.) 1988 *Toward a Political Economy of Development: A Rational Choice Perspective*. University of California Press, Berkeley, CA

Bierstecker T J 1995 The 'triumph' of liberal economic ideas in the developing world. In: Stallings B (ed.) *Global Change, Regional Response: The New International Context of Development*. Cambridge University Press, New York

Cardoso F H, Faletto E 1979 *Dependency and Development in Latin America*. University of California Press, Berkeley, CA

Castells M 1997 *The Power of Identity, Vol. 2. The Information Age: Economy, Society and Culture*. Blackwell, Cambridge, MA

Cox R W 1994 Global restructuring: Making sense of the changing international political economy. In: Stubbs R, Underhill G R D (eds.) *Political Economy and the Changing Global Order*. St. Martins Press, New York, pp. 45–59

Evans P 1995 *Embedded Autonomy: States, Firms, and Industrial Transformation*. Princeton University Press, Princeton, NJ

Evans P, Rueschemeyer D, Skocpol T 1985 *Bringing the State Back In*. Cambridge University Press, New York

Gendzier I 1985 *Managing Political Change: Social Scientists and the Third World*. Westview Press, Boulder, CO

Gereffi G, Korzeniewicz M (eds.) 1994 *Commodity Chains and Global Capitalism*. Greenwood Press, Westport, CT

Gilpin R 1987 *The Political Economy of International Relations*. Princeton University Press, Princeton, NJ

Goldstein J 1993 *Ideas, Interests and American Trade Policy*. Cornell University Press, Ithaca, NY

Gourevitch P 1986 *Politics in Hard Times: Comparative Respon ses to International Economic Crises*. Cornell University Press, Ithaca, NY

Held D, McGrew A G, Goldblatt D, Perraton J 1999 *Global Transformations*. Stanford University Press, Stanford, CA

Keohane R 1984 *After Hegemony*. Princeton University Press, Princeton, NJ

Kitschelt H, Lange P, Marks G, Stephens J D (eds.) 1999 *Continuity and Change in Contemporary Capitalism*. Cambridge University Press, New York

Krasner S D 1976 State power and the structure of international trade. *World Politics* **28**(3): 317–47

Kruger A 1974 The political economy of the rent-seeking society. *American Economic Review* **64**: 291–303

Leys C 1996 *The Rise and Fall of Development Theory*. Indiana University Press, Bloomington, IN
Lipietz A 1987 *Mirages and Miracles: The Crises of Global Fordism*. Verso, London
Marglin S, Schor J (eds.) 1990 *The Golden Age of Capitalism: Reinterpreting the Post-War Experience*. Clarendon Press, Oxford, UK
Piore M, Sabel C 1984 *The Second Industrial Divide: Possibilities for Prosperity*. Basic Books, New York
Rodrik D 1997 *Has Globalization Gone Too Far?* Institute for International Economics, Washington, DC
Rostow W W 1960 *Stages of Economic Growth: A Non-Communist Manifesto*. Cambridge University Press, New York
Wallerstein I 1976 *The Modern World-System*. Academic Press, New York
Waltz K N 1959 *Man, the State, and War: A Theoretical Analysis*. Columbia University Press, New York

I. Robinson

Area and International Studies: Sociology

Area studies brings many disciplines to bear on the study of one geographic or cultural area, such as Latin America, the Middle East, East Asia, or Japan. International studies is a collective term for area studies, but also refers to the study of processes, institutions, and interactions that transcend national boundaries. Sociology is one of several disciplines that may be incorporated into area and international studies. Sociology encompasses the general study of society, including large-scale processes of social change, the organization and functioning of whole societies, social institutions, processes, and groups within societies, and social interaction. There is both synergy and potential for conflict in the relations between area and international studies and sociology.

1. Differences of Perspective and Points of Intersection

Areas studies and sociology constitute two different academic communities with their own sets of assumptions and criteria for evaluating scholarship. These criteria in turn affect the training of graduate students, availability of research support, issues of intellectual interest, infrastructure for research cooperation, venues for presenting research papers, and outlets for publication. Understanding these differing professional perspectives provides a foundation for examining how the two communities relate to each other and how their interaction may stimulate new intellectual contributions.

1.1 How Area Studies Fields View Sociological Research and Contributions to Knowledge

From the perspective of area and international studies, sociology contributes certain ways of analyzing a society or interpreting social phenomena. Sociological research is useful to the extent that it reveals interesting things about the area, which in turn may clarify or extend the existing multidisciplinary body of area knowledge.

Since the aim of area studies research is to contribute to knowledge of the area, scholars are expected to be familiar with the current state of that knowledge in order to identify appropriate research questions. In areas with a strong indigenous research community, the current state of knowledge may encompass both the research literature produced by scholars inside the area and published in their own languages, and the research literature published outside the area in other languages.

The questions that build on this body of area knowledge may be pursued using whatever research materials, opportunities, and strategies are available in the area's research context. Some sophisticated sociological research methods may not work well within particular area studies communities. Rather than collecting new data systematically for quantitative analysis, in some research environments it is more feasible and more appropriate to use observational field methods and interviews, or available documentary sources. These approaches often receive a more favorable reception within the community of area scholars, as well as from local gatekeepers of research access.

Observation, interviewing, and documentary research methods place a premium on language facility rather than on the skills of formal quantitative analysis. In many areas of the world such research requires speaking or reading ability in local vernacular languages, the language of a former colonial power, or some other commercial or regional *lingua franca*. Consequently, area studies scholars place considerable value on appropriate language competence as a basic qualification for scholars and a fundamental tool of scholarly research. If the linguistically competent scholar has mined the available resources appropriately, the resulting research contribution will be evaluated on the basis of its analytical power and the degree to which the findings resonate with what is already known about the subject.

The audience for area studies research is broadly interdisciplinary and may also be quite international. The lines between academic disciplines are much less significant than the period, geographic area, or specific topic of study. Consequently, the most knowledgeable

audience emphasizes the contribution of the research toward understanding of the particular issue or phenomenon within its natural social and historical context.

Theory is relevant in this research environment to the extent that it elucidates the particular case, or conversely, when the evidence from the area refutes a prevailing theory developed elsewhere. However, an area-based case study may generate theory that can then be applied in other settings. Similarities and differences with other cases in other geographic areas are of relatively lesser interest to area studies scholars, although implicit comparisons between the observer's home country and the area of study often underlie (and may distort) the analysis. Hence the theoretical contributions made by area specialists often need to be noticed and utilized by scholars who are not specialists in the original area in order for their general relevance in the social sciences to be recognized.

Area studies scholars evaluate sociological research about their area in terms of its contribution to substantive knowledge of the area, and its ability to provide interpretive frameworks that clarify the social patterns and processes they encounter. Formal training in area studies emphasizes the application of the findings from many disciplines to knowledge of the area, but pays less attention to the theoretical and methodological underpinnings of those disciplines. Hence area specialists without training in a specific discipline may have ample empirical knowledge but lack the tools to conduct empirical research or to draw analytical conclusions. Among area studies scholars who do have disciplinary training in sociology, participation in the interdisciplinary area studies community broadens perspectives and provides additional tools and resources for research, as well as offering an audience that can appreciate and evaluate new research findings about the area.

Sociologists have made major intellectual contributions to area studies in virtually every area of the world. In East Asia, for example, these include the work of Ronald Dore (1958, 1959), Ezra Vogel (1963, 1969), and William Parish and Martin White (1984). Most of these studies are better known among area specialists than among sociologists.

1.2 How Sociology Views Area Studies Research and Contributions to Knowledge

Theory and methods hold pride of place in the discipline of sociology. The aim of sociological research is to contribute to the developing body of sociological theory, rather than to empirical knowledge of a particular place. Sociology tends to view area studies as a collection of available knowledge that can be mined as a resource by scholars who want to pursue theoretical ideas through comparative analysis. This produces a different set of criteria for the conduct and evaluation of research, and can create difficulties for the scholar who wishes to be both sociologist and area studies scholar.

Sociological theory is supposed to be general, though not necessarily universal. That is, it should specify the conditions under which certain results ought to occur (prediction) or explain the processes that operate in a particular case (explanation), by reference to more general concepts that presumably would apply in other similar cases. The theories themselves concern the relationships between such general concepts, which are subject to testing to find out if they continue to hold true or can be rejected on the basis of empirical evidence. New theories can be proposed or old ones elaborated through empirical research, but the findings of empirical research must be couched in theoretical terms.

The habits of thought that are cultivated in the study of sociology thus emphasize extracting from the particular case those properties that can be compared or generalized. Such properties are conceptualized as belonging to limited sets of alternatives, or constituting points on a continuum. The logic of research is then to identify circumstances in which the crucial properties vary, either through internal variation within a large sample or by the selection of cases for systematic comparison, in order to test the validity and limits of the theory. It is also common practice to undertake a single case study, either to apply an existing theory and assess its explanatory power, or to generate new theoretical ideas out of the intriguing properties and dynamics of the case.

Research questions derive from the current state of sociological theory and substantive knowledge about some social phenomenon, abstracted from its geographic location. For the audience of professional sociologists, empirical research contributions are valued to the extent that they are methodologically rigorous and contribute to the advancement of sociological theory. However, there is lively debate within the discipline of sociology about the relative merits of different styles of theory and consequently about the most appropriate research approaches. These methodological orientations also reflect different levels of sociological interest in how well research represents and illuminates the actual cases under study.

The development of sophisticated multivariate methods for analyzing quantitative data has encouraged sociological research to move toward internal comparison of subgroups that cluster or diverge on certain variables within a single dataset. This approach shifts attention toward the proper execution of methodological procedures, and away from the assumptions, operational definitions, and methodological decisions that connect the quantitative research findings to the underlying social reality they claim to measure. Moreover, in international research, the

common framework for data collection may distort findings in favor of the theoretical assumptions of the dominant party, regardless of their relevance in other cultural contexts.

Although qualitative research methods have a long history in sociology and at the beginning of the twenty-first century are enjoying a resurgence in popularity, the predominance of quantitative methods raises standards for qualitative sociological research as well. Qualitative researchers may feel obliged to build internal comparisons into their research design with multiple research sites and subgroups, or to include some systematic quantitative analysis to bolster their qualitative arguments. These demands increase the methodological rigor of qualitative research and the discipline's receptivity to it. However, they may also greatly extend the time required to conduct the research, which is usually a solo undertaking, and may distract the researcher's attention from the contextual analysis that is the hallmark of good qualitative research.

Since in many research contexts serious qualitative research requires strong facility in a vernacular language, the potential range of applicability of the researcher's skills for comparative work becomes a function of the geographic range of the language he or she commands. Hence a specialist with language competence in Spanish, Russian, Chinese, or Arabic may have a potentially wider range of comparative possibilities than someone whose language competence is in Japanese or Hungarian.

The sociologist who is linguistically qualified to do independent area studies research represents only one of several strategies for utilizing area knowledge in sociology. A sociologist who does not have an area studies background may decide that the properties of some society fit the conditions needed for testing an idea in a comparative study. Depending upon the study design and methods to be employed, this might require one or more of the following research strategies: working with sociologists from the area as collaborators in a joint project involving systematic data collection in two or more countries, using common instruments; using the research literature on the area as a secondary resource, to the extent that it exists in a language the sociologist can read; hiring research assistants from the area to gain access to vernacular resources; or going to the area to conduct research, either independently or with the assistance of translators and interpreters.

Such research might result in a multinational comparative study that uses specific countries to represent particular structural conditions, a study using intra-area or intra-country comparisons to highlight variations on a particular theme, a study that compares a phenomenon found in one country with a similar phenomenon found previously in another setting, or a study that explores in detail an institution or phenomenon that appears to produce different results from those found elsewhere. These forms of research are most likely to be presented to a sociological audience, and to be couched in theoretical language as general contributions to the discipline of sociology rather than as contributions to understanding of the area.

Despite the strong demand for theory-driven research, in fact many sociologists initially become intrigued by some social situation or research opportunity, and then develop a theoretical rationale for pursuing it. The key lies in linking the situation to a sociological question of current interest to the discipline. However, the most intriguing issues or research opportunities in an area studies context may not mesh well with the current theoretical concerns of the discipline. The relevant subfield may espouse a theory that asks different research questions, or the area case may contradict the dominant theory, whose advocates may be more inclined to dismiss the troublesome case than to reject their theory.

Conversely, an awareness of current issues in the discipline of sociology may lead area studies researchers to explore new questions that have previously been ignored or even deliberately avoided because of their sensitivity in the cultural context. For example, American sociologists have raised research questions concerning the status of women and minorities in many world areas where local scholars had previously ignored them. Greater interaction across the area studies–sociology divide can thus challenge received understandings and lead to new intellectual developments on both sides. In addition, as sociological theories fall in and out of favor, or conditions in the area change, different area studies concerns may gain new sociological relevance and vice versa.

2. *A Brief History of Area and International Studies in Sociology*

While some of the tendencies discussed above may be found in other social science disciplines, particularly in recent years, the deep concern with both theory and methodology is characteristic of sociology as a discipline and has left its mark on the relations between area and international studies and sociology. The history of area studies within sociology reveals how these potential synergies and conflicts have fluctuated over time and in different contexts.

2.1 *The Global Perspectives of Sociology's Nineteenth-century Founders*

The European founders of the discipline of sociology viewed the world as a laboratory in which similarities between societies offered evidence of general laws, while differences between societies provided clues to the large-scale processes of social change that were sweeping nineteenth-century Europe and America. In

their search for general laws and processes, they made extensive use of available research materials about other societies, even if they did not venture into the field themselves. Area and international studies as known today did not yet exist, but there was considerable information available even on very remote societies. Created partly as a result of colonial relations, these materials included scholarly research by historians, philologists, geographers, and anthropologists, translations of major cultural texts, and reports from colonial administrators, missionaries, and adventurers.

Emile Durkheim made extended use of anthropological research on totemism among American Indian tribes and Australian aborigines to propound the theory that the social order itself was the symbolic focus of religious rituals (Durkheim 1912). Karl Marx and Friedrich Engels used historical and anthropological materials, as well as participant observation of contemporary social life and political events in several countries, to analyze the development of capitalism and its transformation of social relations (Engels 1884, Marx 1852).

Max Weber made even more extensive and systematic use of the historical, anthropological, and cultural materials of particular world areas in his ambitious comparative study of religion and society in India, China, the ancient Middle East, and early modern Europe (Weber 1922a, 1922b). Although Weber's focus was on the link between religious beliefs and economic behavior, his comparative studies constituted a thorough examination of the legal institutions, political order, economic institutions, social structure, and social stratification of each society and how it had changed over time. Some of Weber's interpretations have been superseded by new scholarship, but his comprehensive approach to the understanding of particular societies remains a strong model for both sociology and area studies today.

2.2 Isolation and Internationalism in Twentieth-century American Sociology

While this broad international and historical perspective continued in twentieth-century European sociology, a new generation of American sociologists turned their attention to the social institutions and social processes developing within American society. In a young society whose cities were absorbing millions of immigrants even as the population pushed westward to settle a still-open frontier, American sociologists relied increasingly on observation, interviews, and surveys to analyze the social dynamics swirling around them. They theorized about people who created their own social rules and meaning through social interaction, rather than living with centuries of custom and inherited position. There was little inclination to look to history or other countries for evidence when comparative cases for analysis could be found in subcultures right at home.

As American sociology came into full flower in the 1950s, its insular tendency was further reinforced, despite the influence of many European émigré scholars and the continuing tradition of comparative international research as exemplified by the work of S. N. Eisenstadt (1963) and Seymour M. Lipset (1959). The field was soon dominated by functionalist theory, which emphasized analysis of how the internal components of a society work together to produce a smoothly functioning and stable whole, based on a common set of values. This perspective generated a flood of research on various aspects of American society, couched as general contributions to sociological knowledge. Sociological methodology flourished as well, with increasing emphasis on quantitative analysis of survey data on attitudes and reported behavior, which fit neatly into the assumptions of functionalist theory.

Yet during the same postwar years, new programs of interdisciplinary language and area studies were being developed at large American universities, with strong financial support from the federal government and private foundations. The impetus for the development of area studies came from a combination of America's experience of World War II, and the subsequent Cold War. Acutely aware of the nation's lack of citizens with foreign-language skills and useful knowledge of particular world areas, the US government sought to create such a reserve for future national defense needs. The resulting government-funded program provided general infrastructure support to develop interdisciplinary area studies programs with course offerings in language and various academic disciplines, plus fellowships that the institution could award to graduate students willing to undertake the new courses of study, which required intensive foreign-language study (Lambert 1973).

The funding encouraged the development of area studies programs at the master's level, but the intent was also to encourage and support students who continued with doctoral-level studies in a discipline, in order to staff the continued expansion of international studies in colleges and universities. Major foundations contributed to the government-led effort with additional programs of infrastructure and research support for international studies, including fellowship programs to fund dissertation field research in foreign areas. Among the research support institutions that provided infrastructure for the area studies initiative were the American Council of Learned Societies (ACLS), which provides national-level leadership and coordination in the humanities, and the Social Science Research Council (SSRC), which plays a similar role for the social sciences. These two institutions provided the infrastructure for a series of national-level research planning committees for specific world areas, which in turn served as re-granting bodies for large block grants

of research and research training funds for area studies research provided by major private foundations.

Under these public and private initiatives, academic institutions were encouraged not only to utilize existing staff, but to hire new faculty to broaden their interdisciplinary offerings. The willingness of academic departments to accommodate area specialists depended to a considerable extent on whether the discipline's internal intellectual organization recognized geographic or cultural areas as natural subdivisions. Sociology was particularly resistant to the notion of area specialists, because its internal intellectual organization was oriented to specialization in particular social institutions and processes, and the discipline as a whole was heavily oriented to universalistic theories.

However, the growing dominance of modernization theory in American social science during the 1950s and 1960s lent indirect support to area studies research. Modernization theory was an American elaboration of evolutionary and Durkheimian ideas about how societies could make the transition from traditional to modern, based on a model of full modernity epitomized by the contemporary United States. In sociology, the functionalist model of internally differentiated, modern American society was projected backward to a theoretical model of an undifferentiated traditional society based largely on kinship and holistic communities bound together by shared religious beliefs (Lerner 1958). Indicators were then developed to measure the progress of societies along the road from traditional to modern (Inkeles and Smith 1974), which in many cases also corresponded to prescriptive programs for American assistance to less-developed countries.

Area specialists could apply modernization theory to the area they studied, and in some cases they could also find employment designing and evaluating modernization programs for the area. For areas that were already defined as modern or nearly so, the task was to show how well the theory predicted their trajectory. Unfortunately, the assumption was that the theory must be correct; any misfit between case and theory was either forced to fit, or dismissed as an irrelevant case for sociological study because of its exceptionalism.

In addition to the small cohort of American sociologists trained as area specialists, American postwar affluence and various development programs for other parts of the world were also drawing foreign nationals into American graduate programs in sociology in growing numbers. These young scholars came specifically to learn the theories and methods of American sociology, and were eager to participate in large, multinational studies directed by their American mentors. If they had reservations about the relevance of the survey questions or the applicability of the theory, modernization theory could also subtly imply that this reflected a deficiency in the foreign student's understanding or the backwardness of the native country, rather than a flaw in the theory.

By the 1970s, as funding for area studies programs was drying up, disillusionment with modernization theory was growing among younger sociologists. A cluster of new approaches to development issues emerged, prompted in large measure by scholars with area studies interests. Scholars of Latin America embraced the alternative of dependency theory, which argued that the lagging development of Latin American countries was the result of colonial relations and economic dependency on northern hemisphere countries, rather than on the traditional values and internal backwardness posited by modernization theory (Frank 1967, Cardoso and Faletto 1979). This resonated well with the broader world systems theory propounded by Immanuel Wallerstein, who theorized that the development of capitalism was not a phenomenon that took place within individual nations, but rather was an international set of processes that profoundly altered relationships within the world system of states. In a multivolume study informed by historical area studies research, he traced the decline of Africa and Eastern Europe to dislocations in agricultural markets and trade relations as capitalism expanded around the globe (Wallerstein 1974, 1980). World systems theory offered a new set of transnational variables for studies of development within countries, and focused attention on the international ties of different segments within a society. The theory attracted both scholars with particular area studies interests, and those who wanted to study transnational processes on a more general level.

Attempts to apply world systems theory in Asia were not particularly fruitful, but by the 1980s Japan had risen to a position of economic rivalry with the United States, which prompted new interest in Japanese methods of business organization and industrial production. Although area specialist sociologists of social organization and industrial sociology analyzed the Japanese methods, it was schools of business administration rather than sociology departments that were most eager for their research. Sociological interest deepened as it became more apparent that other Asian countries were following a Japanese model of state-directed, export-led development. This did not fit any of the earlier development theories, but resonated with new theoretical interest in relations between state and society (Skocpol 1979).

In the post-Cold War 1990s, language and area studies programs came under heavy attack from social science disciplines as a relic of the Cold War (see Samuels and Weiner 1992). In sociology this did not signal a return to American isolation so much as a general internationalization of the discipline, in which the many alternate ways of accessing area knowledge made the trained area studies specialist less significant as a gatekeeper for that knowledge. Ironically, in

Japanese studies, by the 1990s the postwar investment in language and area studies scholars had produced a large enough body of area studies research in English to support secondary research by sociologists without Japanese language skills.

Sociology was also becoming more internationalized through a new wave of European postmodern and poststructural theories that called attention to nuances of culture and symbolic language even as they emphasized the internationalization of popular culture and the breakdown of stable cultural systems of meaning (Bourdieu 1977, Foucault 1975). In drawing closer to new European theories, American sociology was also reconnecting to the relatively unbroken European tradition of sociology as a discipline interested in area-specific knowledge about the whole world.

3. *Institutional Relationships between Sociology and Area and International Studies*

Sociologists are organized internationally through the International Sociological Association, and nationally through national or regional disciplinary associations, while area specialists are organized in North America, Europe, and elsewhere through interdisciplinary professional associations that are area- or region-specific, such as the Association for Asian Studies, the African Studies Association, the Latin American Studies Association, the European Association of Asian Studies, and the American Studies Association in Japan.

The International Sociological Association is structured around a series of international committees representing subfields of the discipline, which facilitate the interaction of scholars from different countries who work on similar issues. The implicit assumption is that scholars do research in their own areas to contribute to collective sociological knowledge, and area-specific research or knowledge has no independent significance. Conversely, sociology as a discipline has little visibility in area studies associations, which tend to have strong representation from history and literature along with many social science disciplines.

In the mid-1990s the International Sociological Association organized a series of regional conferences to consider the state of sociology from the perspective of each region. The resulting publications reflect the substantive issues most relevant to each region, including in some cases a strong desire to develop new theories out of regional experience or a regional and linguistic community of sociologists. The base, however, remains sociologists *from* the region as opposed to sociologists *of* the region (see Wallerstein et al. 1998).

As the ISA has taken up regional concerns, the large American Sociological Association is becoming steadily more international in the scope of its membership and its concerns. The ASA has long had an institutional committee devoted to international ties, which in the early 1990s voiced concern over the widening gap between area studies and the social sciences. Rather than leading toward greater integration of area studies scholars into sociology and other social science disciplines in the United States, this concern combined with serious financial exigencies fed into the decision of the Social Science Research Council to dismantle its long-standing infrastructure support for specific area studies committees in favor of very broad regional committees composed of American scholars and scholars from the region. The change was intended to encourage research across regions and on transnational processes, but it also reflected the growing hostility of the social science disciplines to area specialists and the enterprise of interdisciplinary area studies.

Internally, the ASA has made international or comparative sociological studies the thematic focus of several recent annual meetings, and its journals publish international research regularly. Since the 1970s several ASA sections have formed with explicitly international interests, including sections on Asia and Asian America; Comparative and Historical Sociology; International Migration; Latina/o Sociology; Peace, War, and Social Conflict; and Political Economy of the World System. Although their themes are not as explicitly international, some other sections, such as Collective Behavior and Social Movements, have also become thoroughly international in scope. The two explicitly area-focused sections of the ASA combine an interest in a world area with an identity community of sociologists who have ethnic roots in the area but do not necessarily study the area itself. Many smaller interest groups for particular areas and countries also sponsor gatherings at the ASA annual meetings, which help to link area studies, area nationals, and sociology, but also may blur the intellectual focus on the sociological study of each area.

While tensions remain between area and international studies and sociology, the discipline is international in scope and is getting better at accommodating area and international research. Still, the gap remains wide enough that scholars who hope to combine area specialization with sociology must be prepared to navigate between two academic communities with different standards and expectations. The potential rewards of doing so promise to strength en area studies with the theoretical and methodological rigor of sociology, and to challenge sociology with the theoretical insights of research that is deeply contextualized by interdisciplinary area studies knowledge.

See also: Area and International Studies in the United States: Intellectual Trends; Comparative Studies:

Method and Design; Dependency Theory; Durkheim, Emile (1858–1917); Human–Environment Relationship: Comparative Case Studies; Marx, Karl (1818–89); Modernization, Sociological Theories of; Sociology, History of; Sociology: Overview; Weber, Max (1864–1920); World Systems Theory

Bibliography

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK

Cardoso F H, Faletto E 1979 *Dependency and Development in Latin America*. University of California Press, Berkeley, CA

Dore R P 1958 *City Life in Japan: A Study of a Tokyo Ward*. Routledge & Kegan Paul, London

Dore R P 1959 *Land Reform in Japan*. Oxford University Press, London

Durkheim E 1912 *Les Formes Élementaires de la Vie Religieuse* [1961 *The Elementary Forms of the Religious Life*. The Free Press, New York]

Eisenstadt S N 1963 *The Political Systems of Empires*. The Free Press, New York

Engels F 1884 *Ursprung der Familie, des Privateigentums und des Staats* [1972 *The Origin of the Family, Private Property, and the State*. International Publishers, New York]

Foucault M 1975 *Surveiller et Punir: Naissance de la Prison*. Editions Gallimard, Paris [1979 *Discipline and Punish*. Vintage Books, New York]

Frank A G 1967 *Capitalism and Underdevelopment in Latin America*. Monthly Review Press, New York

Inkeles A, Smith D H 1974 *Becoming Modern: Individual Change in Six Developing Countries*. Harvard University Press, Cambridge, MA

Lambert R D 1973 *Language and Area Studies Review*. American Academy of Political and Social Science, Philadelphia

Lerner D 1958 *The Passing of Traditional Society*. Free Press, Glencoe, IL

Lipset S M 1959 *Political Man: The Social Bases of Politics*. Doubleday, Garden City, NY

Marx K 1852 *Klassenkämpfe in Frankreich 1848 bis 1850* [1964 *The Class Struggles in France 1948–1850*. International Publishers, New York]

Parish W L, Whyte M K 1984 *Urban Life in Contemporary China*. University of Chicago Press, Chicago

Samuels R J, Weiner M 1992 *The Political Culture of Foreign Area and International Studies: Essays in Honor of Lucian Pye*. Brassey, Washington, DC

Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge University Press, Cambridge, UK

Vogel E F 1963 *Japan's New Middle Class: The Salary Man and His Family in a Tokyo Suburb*, 2nd edn. University of California Press, Berkeley, CA

Vogel E 1969 *Canton Under Communism: Programs and Politics in a Provincial Capital, 1949–1968*. Harvard University Press, Cambridge, MA

Wallerstein I 1974 *The Modern World-System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

Wallerstein I 1980 *The Modern World-System II: Mercantilism and the Consolidation of the European World-Economy, 1600–1750*. Academic Press, New York

Wallerstein I, Walby S, Main-Ahmed S, Fujita K, Robert P, Catano G, Fortuna C, Swedberg R, Patel S, Webster E, Moran M-L 1998 Spanning the globe: Flavors of sociology. *Contemporary Sociology: A Journal of Reviews* **27**(4): 325–42

Weber M 1922a *'Konfuzianismus und Taoismus' Gesammelte Aufsätze sur Religionssoziologie*. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany [1968 *The Religion of China: Confucianism and Taoism*. Free Press, New York], *'Antike Judentum' Gesammelte Aufsätze sur Religionssoziologie* [1967 *Ancient Judaism*. Free Press, Glencoe, IL] *'Hinduismus und Buddhismus' Gesammelte Aufsätze sur Religionssoziologie* [1967 *The Religion of India: The Sociology of Hinduism and Buddhism*. Free Press, New York]

Weber M 1922b *Wirtschaft und Gesellschaft: Grundriss der verstehenden Soziologie*. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany [1978 *Economy and Society: An Outline of Interpretive Sociology*. University of California Press, Berkeley, CA]

P. G. Steinhoff

Areal Linguistics

Areal linguistics is concerned with the diffusion of structural features of language among the languages of a geographical region. Various linguistics areas are exemplified in this article and the importance of areal linguistics to the study of linguistic change in explained.

1. *Linguistic Areas*

A linguistic area is a geographical area in which, due to language contact and borrowing, languages of a region come to share certain structural features—not only borrowed words, but also shared elements of sound and grammar. Other names sometimes used to refer to linguistic areas are Sprachbund, diffusion area, adstratum, and convergence area. Areal linguistics is concerned with linguistic areas, with the diffusion of structural traits across language boundaries.

2. *Defining Linguistic Areas*

Central to a linguistic area is the experience of structural similarities shared among languages of a geographical area (where usually some of the languages are unrelated or at least are not all close relatives). It is assumed that the reason the languages of the area share these traits is because they have borrowed from one another.

Areal linguistics is important in historical linguistics, whose goal is to find the full history of languages. A full history includes understanding of both inherited traits (shared in related languages because they come from a common parent language, for example features shared by English and German because both inherited

traits from Proto-Germanic, their parent) and diffused features (shared through borrowing and convergence among neighboring languages; examples below).

While some linguistic areas are reasonably well established, based on a number of shared areal traits, all linguistic areas could benefit from additional investigation.

Some proposed linguistic areas amount to barely more than preliminary hypotheses, while in general linguistic areas have been defined, surprisingly, on the basis of a rather small number of shared traits.

3. *Examples of Linguistic Areas*

For understanding of areal linguistics, it will be helpful to consider the better known linguistic areas together with some of the traits shared by the languages in each area (for more details, see Campbell 1998, pp. 299–310).

3.1 *The Balkans*

The Balkans is the best known linguistic area. The languages of this area are: Greek, Albanian, Serbo-Croatian, Bulgarian, Macedonian, and Rumanian; some scholars also add Romani (the language of the Gypsies) and Turkish. Some salient traits of the Balkans linguistic area are: (a) A central vowel (somewhat like the vowel in English 'but') (not in Greek or Macedonian). (b) Syncretism of the dative and genitive cases (merged in form and function); this is illustrated by Rumanian *fetei* 'to the girl' or 'girl's' (compare *fată* '*girl*'), as in *am dat o carte fetei* 'I gave a letter to the girl' and *frate fetei* 'the girl's brother.' (c) Postposed articles (not in Greek), e.g., Bulgarian *mâž-ât* 'the man'/*mâž* 'man,'-*ât* 'the.' (d) Futures signalled by an auxiliary verb corresponding to 'want' or 'have' (not in Bulgarian or Macedonian), e.g., Rumania *voi fuma* 'I will smoke' (literally, 'I want to smoke') and *am să cânt* 'I will sing' (literally 'I have sing'). (e) Perfect with an auxiliary verb corresponding to 'have.' (f) Absence of infinitives (instead, the languages have constructions such as 'I want that I go' for 'I want to go'); for example, 'give me something to drink' has the form corresponding to 'give me that I drink,' e.g., Rumanian *dă -mi să beau*, Bulgarian *daj mi da pija*, and Greek *dós mu na pjó*. (g) Double marking of objects which refer to humans or animals by use of a personal pronoun together with the object, e.g., Rumanian *i-am scris lui Ion* 'I wrote to John,' literally 'to.him-I wrote him John,' and Greek *ton vlépo ton jáni* 'I see John,' literally 'him.Acc I see the/him.Acc John' (see Joseph 1992).

3.2 *South Asia* (*Indian Subcontinent*)

This area is also well known. It is composed of languages belong to the Indo-Aryan, Dravidian, Munda, and Tibeto-Burman families. A few traits from the list of several shared among languages of the area are: (a) retroflex consonants (pronounced with the tip of the tongue pulled back towards the hard palate); (b) absence of prefixes (except in Munda); (c) Subject–Object–Verb (SOV) basic word order, including postpositions rather than prepositions (i.e., the equivalent of 'people with' instead of, as in English, 'with people'), (d) absence of a verb 'to have'; and (e) the 'conjunctive or absolutive participle' (meaning that subordinates clauses tend to have particles, rather than fully conjugated verbs, which are placed before the thing they modify, for example, the equivalent of 'the having eaten jackal ran away' where English has 'the jackal which had eaten ran way').

Some of the proposed areal features are not limited to the Indian subcontinent (e.g., SOV basic word order, found throughout much of Eurasia, and in other parts of the world). Some traits are not necessarily independent of one another (for example, languages with SOV basic word order tend also to have subordinate clauses with participles, not fully conjugated verbs, and tend not to have prefixes) (see Emeneau 1980).

3.3 *Mesoamerica*

The language families and isolates (languages with no known relatives) which make up the Mesoamerican linguistic area are: Nahua (branch of UtoAztecan), Mayan, MixeZoquean, Otomanguean, Totonacan, Xincan, Tarascan, Cuitlatec, Tequistlatecan, and Huave. Five diagnostic areal traits are shared by nearly all Mesoamerican languages, but not by neighboring languages outside this area. They are: (a) Possessive construction of the type *his-dog the man* 'the man's dog,' as in Pipil (UtoAztecan): *ipe:lu ne ta:kat*, literally 'hisdog the man.' (b) Relational nouns (locational expressions composed of noun and possessive pronominal prefixes, which function as prepositions in English), of the form, for example, *my-head* for 'on me,' as in Tz'utujil (Mayan): (*č*)*ri:x* 'behind it, in back of it,' composed of *č* 'at, in,' *r* 'his/her/its' and *i:x* 'back,' contrasted with *čwi:x* 'behind me,' literally 'atmyback.' (c) Vigesimal numeral systems based on twenty, such as that of Chol (Mayan): *hun-k'al* '20' (1×20), *ča?-k'al* '40' (2×20), *uS-k'al* '60' (3×20), *ho? k'al* '100' (5×20), *hun-bahk'* '400' (1-*bahk'*), *ča?-bahk*' '800' (2×400). (d) Non-verbfinal basic word order (no SOV languages) although Mesoamerica is surrounded by languages both to the north and south which have SOV word order, all languages within the linguistic area have VOS, VSO, or SVO basic order, not SOV. (e) Many loan translation compounds (calques) are shared by Mesoamerican languages, e.g., 'boa' = 'deersnake,' 'egg' = 'birdstone/bone,' 'lime' = 'stone(ash),' 'knee' = 'leghead,' and 'wrist' = 'hand-neck.' Since these five traits are shared almost unan-

imously throughout the languages of Mesoamerica but are found extremely rarely in languages outside Mesoamerica, they are considered strong evidence of the validity of Mesoamerica as a linguistic area. Additionally, a large number of other features are shared among several Mesoamerican languages, but are not found in all of the languages of the area, while some other traits shared among the Mesoamerican languages are found also in languages beyond the borders of the area. To cite just one example found in several but not all Mesoamerican languages, sentences which in English have a pronoun subject such as 'you are carpenter,' are formed in several Mesoamerican languages with a prefix or suffix for the pronoun attached directly to the noun, as in Q' eqchi' (Mayan) *iSq-at* (women-you) 'you are a women,' *kwinq-in* (man-I) 'I am a man': Pipil *ni-siwa:t* (you-woman) 'I am a woman,' *ti-ta:kat* 'you are a man' (see Campbell et al. 1986).

3.4 The Northwest Coast of North America

The best known linguistic area in North America is the Northwest Coast. It includes Tlingit, Eyak, the Athapaskan languages of the region, Haida, Tsimshian, Wakashan, Chimakuan, Salishan, Alsea, Coosan, Kalapuyan, Takelma, and Lower Chinook. The languages of this area are characterized by elaborate systems of consonants, which include series of phonetically very complex sounds. In contrast, the labial consonant series (which includes in English, for example, *p*, *b*, and *m*) is typically either lacking or contains few consonants: labials are completely lacking in Tlingit and Tillamook, and are quite limited in Eyak and most Athabaskan languages. The vowel systems are limited, with only three vowels (*i*, *a*, *o* or *i*, *a*, *u*) in several of the languages. Some shared grammatical and word-formation traits, from a list of many, include: extensive use of suffixes; nearly complete absence of prefixes; reduplication processes (where the first part of the word is repeated) of several sorts, signaling various grammatical functions, e.g., iteration, continuative, progressive, plural, collective, etc.; evidential markers in the verb (a suffix indicating, for example, whether the speaker has first-hand knowledge of the event, has only hearsay knowledge, or doubts it); and directional suffixes in the verb (telling whether the action is towards the speaker, away from the speaker, and so on); a masculine/feminine gender distinction in demonstratives and articles; a visibility/invisibility opposition in demonstratives (that is, for example, one word corresponding to English 'that' used for things the speaker can see, 'that boy' (visible), and an entirely different word for 'that' for things not visible to the speaker, 'that boy' (not visible, known about, but not present). Northwest Coast languages have distinct verb words for singular and plural (that is, an entirely different root may be required with a plural subject, for example, 'the children sat on the ground,' different from the root used with a singular subject, such as 'the child sat on the ground'—where the word for 'sat' would be distinct in the two instances).

Some other traits are shared by a smaller number of Northwest Coast languages, not by all. One example is the so-called 'lexical suffixes,' found in a number of the languages (Wakashan and Salishan). Lexical suffixes are grammatical endings which designate familiar objects (which are ordinarily signaled with full independent words in most other languages) such as body parts, geographical features, cultural artifacts, and some abstract notions. Wakashan, for example, has 300 of these. Another example is the very limited role for a contrast between nouns and verbs as distinct categories in several of the languages (see Campbell 1997, pp. 330–4).

3.5 The Baltic

The Baltic area is defined somewhat differently by different scholars, but includes at least Balto-Finnic languages (especially Estonian and Livonian), Latvian, Latgalian, Lithuanian, and Baltic German. Some would include Swedish, Danish, and dialects of Russian, as well. Some of the shared features which define the Baltic area are: (a) first-syllable stress, palatalization of consonants (a'y'-like release of these consonant sounds, as in Russian $p^{y}at^{y}$ 'five'), (b) a tonal contrast; (c) partitive case (to signal partially affected objects, equivalent to, for example, 'I ate (some) apple' found in Balto-Finnic, Lithuanian, Latvian, some dialects of Russian); (d) evidential voice ('John works hard (it is said)'; Estonian, Livonian, Latvian, Lithuanian); (e) prepositional verbs (German *ausgehen* (out-to-go) 'to go out'; Livonian, German, Karelian dialects); (f) SVO basic word order; and (g) adjectives agree in case and number with the Nouns they modify) (see Zeps 1962).

3.6 Ethiopia

Languages of the Ethiopian linguistic area include: Cushitic, Ethiopian Semitic, Omotic, Anyuak, Gumuz, and others. Among the traits they share are: (a) SOV basic word order, including postpositions; (b) subordinate clause preceding main clause; (c) gerund (non-conjugated verbs in subordinate clauses, often marked for person and gender); (d) a 'quoting' construction (a direct quotation followed by some form of 'to say'); (e) compound verbs (consisting of a noun-like 'preverb' and a semantically empty auxiliary verb); (f) negative verb 'to be'; (g) plurals of nouns are not used after numbers (equivalent to 'three apple' for

'three apples'); (h) gender distinction in second and third person pronouns (English has the 'he'/'she' gender distinction for third person, but nothing like 'you' (masculine)/'you' (feminine), found here); and (i) the form equivalent to the feminine singular used for plural agreement (feminine singular adjective, verb, or pronoun is used to agree with a plural noun) (see Ferguson 1976).

4. *How are Linguistic Areas Determined?*

On what basis is it decided that some region constitutes a linguistic area? Scholars have at times utilized the following considerations as criteria: (a) the number of traits shared by languages in a geographical area, (b) bundling of the traits in some significant way (for example, clustering at roughly the same geographical boundaries), and (c) the weight of different areal traits (some are counted differently from others on the assumption that some provide stronger evidence than others of areal affiliation) (see Campbell et al. 1986).

With respect to the number of areal traits necessary to justify a linguistic area, in general, the linguistic areas in which many diffused traits are shared among the languages are considered more strongly established; however, some argue that even one shared trait is enough to define a weak linguistic area. Without worries over some arbitrary minimum number of defining traits, it is safe to say that some areas are more securely established because they contain many shared traits, whereas other areas may be more weakly defined because their languages share fewer areal traits. In the linguistic areas mentioned above, there is considerable variation in the number and kind of traits they share which define them.

With respect to the relatively greater weight or importance attributed to some traits than to others for defining linguistic areas, the borrowed word-order patterns in the Ethiopian linguistic area provide an instructive example. Ethiopian Semitic languages exhibit a number of areal traits diffused from neighboring Cushitic languages. Several of these individual traits, however, are interconnected due to the borrowing of the SOV basic word order patterns of Cushitic languages into the formerly VSO Ethiopian Semitic languages. The orders Noun–Postposition, Verb–Auxiliary, Relative Clause–Head Noun, and Adjective–Noun are all correlated and thus they tend to co-occur with SOV order cross-linguistically (see *Word Order*). If the expected correlations among these constructions are not taken into account, one might be tempted to count each one of these word orders in different constructions as a separate shared areal trait and their presence in Ethiopian Semitic languages might seem to reflect several different diffused traits (SOV counted as one, Noun–Postposition as another, and so on), and they could be taken as several independent pieces of evidence defining a linguistic area. However, from the perspective of expected word-order co-occurrences, these word-order arrangements may not be independent traits, but may be viewed as the result of the diffusion of a single complex feature, the overall SOV word-order type with its various correlated orderings in interrelated constructions. However, even though the borrowing of SOV basic word order type may count only as a single diffused areal trait, many scholars would still rank it as counting for far more than some other individual traits based on the knowledge of how difficult it is for a language to change so much of its basic word order by diffusion.

With respect to the criterion of the bundling of areal traits, some scholars had thought that such clustering at the boundaries of a linguistic area might be necessary or at least helpful for defining linguistic areas properly. However, this is not so. Often one trait may spread out and extend across a greater territory than another trait, whose territory may be more limited, so that their boundaries do not coincide ('bundle'). This is the most typical pattern, where languages within the core of an area may share many features, but the geographical extent of the individual traits may vary considerably one from another. However, in a situation where the traits do coincide at a clear boundary, rare though this may be, the definition of a linguistic area to match their boundaries is relatively secure. As seen earlier, several of the traits in the Mesoamerican linguistic area do have the same boundary, but in many other areas, the core areal traits do not have the same boundaries, offering no bundling and no clearly identifiable outer border of the linguistic area in question.

5. *Areal Linguistics and Language Classification*

Unfortunately, it is not uncommon to find cases of similarities among languages which are in reality due to areal diffusion but which are mistakenly taken to be evidence of a possible remote family relationship among the languages in question. One example will be sufficient to illustrate this: the 'Altaic' hypothesis.

The core Altaic hypothesis holds that Turkic, Mongolian, and Manchu-Tungusic are related in a larger language family, though versions of the hypothesis have been proposed which would include also Korean, Japanese, and sometimes also Ainu. While Altaic is repeated in encyclopedias, most specialists find that the evidence at hand does not support the conclusion of a family relationship among these language groups. The most serious problems for the hypothesis has to do with the fact that much of the original motivation for joining these languages seems to have been based on traits which are shared areally, for example, vowel harmony (where within a word there is a restriction on which vowels can cococur with each other, for example, only combinations of back

vowels (*a*, *o*, *u*) or only vowels from the front vowel set (*i*, *e*, *æ*) but not some from one set and others from the other set in the same word), relatively simple inventories of sounds, agglutination, suffixing, SOV word order, and subordinate clauses whose verbs are participles, not fully conjugated verbs. These are also areal traits, shared by a number of languages in surrounding regions whose structural properties were not well known when the hypothesis was first framed. Because these traits may be shared among these languages due to areal linguistic contact and borrowing, such traits are not compelling evidence of a family relationship (with its assumption that the traits were inherited from an earlier common ancestor).

From this example, it is easy to see one reason why the identification of areal traits is so important in historical linguistics. In this case, failure to recognize the areal traits led to a questionable proposal of genetic relationship among neighboring language families.

See also: Historical Linguistics: Overview; Languages: Genetic Classification; Linguistic Typology; Linguistics: Overview

Bibliography

Campbell L 1997 *American Indian Languages: The Historical Linguistics of North America*. Oxford University Press, New York

Campbell L 1998 *Historical Linguistics: An Introduction*. MIT Press, Cambridge, MA

Campbell L, Kaufman T, Smith-Stark T 1986 Mesoamerica as a linguistic area. *Language* **62**: 530–70

Emeneau M B 1980 *Language and Linguistic Area: Essays by Murray B. Emeneau* (selected and introduced by Dil A S). Stanford University Press, Stanford, CA

Ferguson C 1976 The Ethiopian language era. In: Bender M L et al. (eds.) *Language in Ethiopia*. Oxford University Press, Oxford, UK, pp. 63–76

Joseph B 1992 Balkan languages. *International Encyclopedia of Linguistics*, Oxford University Press, Oxford, UK, Vol. 1, pp. 153–5

Zeps V 1962 *Latvian and Finnic Linguistic Convergence* (Uralic and Altaic Series, Vol. 9). University of Indiana Press, Bloomington, IN

L. Campbell

Arendt, Hannah (1906–75)

Born on October 14, 1906 in Hannover, Arendt grew up in a liberal Jewish family in Koenigsberg. Later, she studied philosophy, protestant theology, and Greek philology in Marburg, Heidelberg, and Freiburg. During this time, she got to know and was greatly influenced by two philosophers: Martin Heidegger and Karl Jaspers. In 1928 she received her Ph.D. in Heidelberg for her thesis on 'The concept of love in the work of St. Augustin.' Understanding the implications of Hitler's accession to power, she left Germany in 1933. She temporarily settled in Paris where she volunteered for a Jewish refugee organization. In 1941 she emigrated to the USA, and became an American citizen 10 years later. She worked as a lecturer at various American universities and colleges, and was a well-known freelance writer who contributed to the American intellectual culture as well as to the postwar European political discourse. Arendt died on December 4, 1975 in New York.

1. Arendt's Methodological Approach to Theory and Thinking

One cannot comprehend Arendt's thought without taking into account her personal experience of National Socialism and the genocide of the European Jewry. Throughout her life, Arendt always pointed to the existence of National Socialist and Stalinist terror and to concentration camps, which she considered the utmost challenges to political thinking.

It appears that she interlinked all her concepts and categories as well as the history of thought on which she based her notions with the experience of totalitarianism. She considered totalitarianism not only to be a breach in modern civilization but was even convinced that totalitarian rule challenges modern political thinking. This view gave rise to her seemingly unscientific methodological approach to thinking: How to create a world to which mankind feels attached and in which the individual does not lose their ability to judge? How to protect the 'body politic' from self-destruction? In this context, the concept of 'understanding the world' seems to become the overarching hermeneutic idea which interrelates all other categories: judging, acting, being in the public, etc.

Arendt focuses on reconsidering the instruments applied in social sciences. She believes that these were reduced to an absurdity by totalitarianism (see also Lefort 1988, p. 48). Her considerations encompass a wide range of historical, philosophical, and political references. Her concept of 'the world,' and her radicalization of hermeneutics by questioning the self as a subject have political and philosophical connotations.

There are some specific methodological aspects that need to be considered to understand Arendt's way of thinking:

(a) Historically Arendt focuses on the disintegration of the public sphere in the European societies (e.g., France and Germany but also Russia) at the end of the nineteenth century. Her point of reference is the emergence of a new form of total domination in the 1930s, which basically differs from former types of

tyranny or dictatorship but refers to the concept of natural law and/or of dialectical law. For Arendt, the need to understand totalitarianism epistemologically arises from the fact that totalitarianism cannot be sufficiently explained by historical and social sciences (Arendt 1994, p. 317). She argues that totalitarian rule has had a major impact on the categories applied in social sciences (Arendt 1994, p. 318). She criticizes social sciences for assuming that political action is basically a rational process (in the sense of Max Weber, see *Weber, Max (1864–1920)*). But totalitarian terror is, in Arendt's view, not at all rational but meaningless and contingent. Therefore, it is not possible to grasp the absolute meaninglessness of destruction, evil, and mass murder by adopting social-science approaches. This is the reason why Arendt's hermeneutic concept is closely interlinked with her criticism of social sciences.

(b) Arendt aims to 're-open' the political dimension of modern political thinking by applying her concept of understanding. In this context, 'reopening' means that a dimension, which has been lost, has to be rediscovered: the public sphere as a sphere of human interaction. It is the public sphere that keeps the body politic (the political system) alive. In Arendt's view, the removal of institutions and legal systems as well as the abolition of political freedom under totalitarian rule are direct consequences of the dissolution of the public sphere in modern times.

(c) In terms of philosophical categories, understanding is taken to be different from cognition. Understanding is not simply an intellectual concept of the world but is a hermeneutic process, and thus also a philosophical and emotional approach to the world. Understanding is interlinked with experience and judgement.

(d) Understanding is also supposed to be different from having 'correct information and knowledge,' although the two are interrelated. Understanding also differs from scientific knowledge. In Arendt's critical view knowledge is meaningless without preliminary understanding. Here, she breaks with the tradition of natural and social sciences according to which the knowledge about facts and details of a phenomenon and putting these pieces together in systematical order means to recognize these facts and details or to realize their truth.

Furthermore, 'understanding' does not mean 'causal and historicist explanation': 'The necessity which all causal historiography consciously or unconsciously presupposes does not exist in history. What really exists is the irrevocability of the events themselves, whose poignant effectiveness in the field of political action does not mean that certain elements of the past have received their final, definite form, but that something inescapably new was born' (Arendt 1994, p. 326).

In this way, Arendt again draws a line between political thinking and the world of scientific analysis. Her approach denies all teleological and deterministic as well as deductive thinking. However, it includes a fundamental criticism of the concept of progress and rationality in modern history.

Methodologically, her work is challenging insofar as she opposes the traditional (usual) meaning of concepts. In all of her greater works she critically reviews concepts and conceptions as facets of the history of thinking; she eventually tries to re-establish an authentic relation between a given concept, reality and the acting individual; furthermore she seeks to reveal what is new about the object of recognition. Thus, her work is also a major contribution to hermeneutics in social sciences.

2. *Basic Concepts in Arendt's Thinking*

Arendt did not evolve a systematic political theory nor did she elaborate a philosophical system. However, her approach to political thinking remains a challenge for both political theory and philosophy. It is based on a different understanding of the major concepts of political theory. Her critical view of the techniques applied in social sciences induced Arendt not only to do research on etymological changes in social science concepts (e.g., freedom, morality) but also to relate her own concept to the overarching idea how to reveal the political dimension in the community of citizens.

2.1 *The Concept of Society*

Society is considered a never-ending process in which citizens act and judge towards an open future within a public sphere; it is not perceived as a set of institutions and ways of living. Arendt opposes traditional views of theory on society by denying that society is to be seen as a (national) entity.

In her view institutions have a protective function. However, they can also pose a threat to the 'humanness of the world.' In this respect, Arendt follows Max Weber in his criticism of bureaucracy. She agrees that the public realm cannot do without institutions, but here again her concept of institution is different, for example, the act of setting up a 'political body' during a revolution is taken to be an institution by itself.

The 'humanness' is supposed to be maintained only by the citizens. Thus, it may not be surprising that Arendt relates her basic concepts to certain abilities of citizens: thinking, judging, and acting in the public realm.

2.2 *The Political*

Arendt's concept of 'the political,' which is basically different from 'politics,' is based on her distinction

between the public sphere and the needs and desires dominating the private sphere.

The political is not a substance of humankind but arises in the midst of citizens in the public realm. This is another example of the way in which Arendt relates her concepts to the sphere of interaction. In her view, the political gives rise to a new beginning in the sphere of interaction.

Arendt's way of thinking can be closely related to the tradition of political thinking which emerged in Greek and Roman antiquity. She acknowledged the Greek 'polis' and the Roman republic as historical forms of political sphere, to which die American founding fathers referred to perceiving the republic as a permanent founding of itself (self-instituting) within a system of checks and balances.

2.3 Political Freedom

The term 'freedom' (like 'society') refers to a process of interaction in the political sphere and not primarily to institutions. Freedom is not restricted to the individual or to the private sphere. Moreover, 'political freedom is distinct from philosophic freedom in being clearly a quality of the I-can and not of the I-will. Since it is possessed by the citizen rather than by man in general, it can manifest itself only in communities, where the many who live together have their intercourse both in word and in deed regulated by a great number of *rapports*—laws, customs, habits, and the like. In other words, political freedom is possible only in the sphere of human plurality, and on the premise that this sphere is not simply an extension of the dual I-and-myself to a plural We' (Arendt 1981, p. 200).

Basically, political freedom appears to have a double meaning: on the one hand there is the act of setting up a 'political body' (e.g., by means of a constitution) and on the other hand there are the *rapports* of citizens in favor of their political body.

2.4 Power

As with the concept of freedom to which it is linked, power is interrelated to the public realm. Arendt does not interlink power with domination, and violence. She disagrees with the definition of power given by major social scientists and especially by Max Weber (Weber 1978, p. 53).

In Arendt's thinking power, that is, political power, cannot be 'possessed' by individuals nor—as this was assumed during the French Revolution—by 'the people.' On this point Arendt differs from the liberal as well as from the revolutionary concept of power. 'Power corresponds to the human ability, not just to act in concert. Power is never the property of an individual; it belongs to a group and remains in existence only so long as the group keeps together' (Arendt 1972, p. 143).

This means that power belongs to the public sphere. It can be incorporated by means of a revolution, a political uprising or an act of founding. However, if power is institutionalized it tends to become bureaucratic. In this respect Arendt agrees with the skepticism of Max Weber about modern bureaucracy.

Thus, power comes up and disappears again depending on the contingency of history.

2.5 'Alienation from the World' and 'Being in the World'

Arendt experienced emigration, exile and the deplorable status of a stateless person over many years. It is no accident that the issue of alienation plays a major role in her works. But again, alienation is not related to the private realm but to the public sphere. Alienation refers to the estrangement of individuals from their political community. According to Arendt's theory of humanness, individuals have to 'appear' in the world, join their fellow men and commit themselves to the preservation of the world they live in through their actions. This is what 'citizen' means.

In this context, 'world' (and opposed to 'being without a world'/worldlessness) refers to the fundamental fact of being born into an already existing world.

However, Arendt's concept of alienation differs from the one of Marx or other philosophical authors insofar as it focuses not on the individual but on the relationship between the citizens and the world which they have created (see *Marx, Karl (1818–89)*).

This is the starting point from which Arendt pursues two objectives: to reflect upon how to make the world worldly and to humanize thinking. This is a constant theme throughout her work.

2.6 Totalitarianism

All of Arendt's works on political thinking refer to the political events of the time. Her research on totalitarianism during the 1940s focuses on various aspects.

(a) She identifies totalitarianism as a new type of total domination. It bases on violent antidemocratic movements which later become part of a political system which engages in the control of institutions and the personal lives of the people, sets up a specific ideology, and establishes systematic terror.

(b) She focuses on the 'The origins of totalitarianism,' that is, imperialism and anti-Semitism as the major sources of destroying the political body of the European democracies since the nineteenth century.

(c) She considers totalitarianism as representing a breach in modern civilization which began a long time ago. Here again she argues that the categories of thinking are deeply affected by this process.

Arendt considers totalitarian rule as a possible outcome of the weaknesses of modern age. It is bewildering that modern democracies themselves may generate totalitarian elements. Hence, in Arendt's view totalitarian rule cannot be dismissed as an 'accident in history'; there is always the possibility that this form of rule emerges.

This is why it is problematic to discuss moral behavior in a traditional sense after the emergence of totalitarian rule has occurred. Wherever human nature was reduced to the level of 'material' it seems inappropriate to simply re-establish rules that were violated or manipulated. For Arendt, it is important to reflect on the dimensions of a world which gave rise to such a development and which continues to exist after the break-down of totalitarianism.

2.6.1 Radical and Banal Evil. One word within Arendt's concept of totalitarianism generated major criticism: 'evil.' Arendt distinguishes between radical and banal evil. 'Radical evil' makes human beings superfluous. Arendt uses this concept to find an explanation for Nazi death camps. 'Radical evil' stands for both an absolute and a contingent negation of humanness. The other side of evil is its banality. Arendt takes the Nazi functionary Adolf Eichmann as an example. Her characterization of Eichmann as 'banal' has been widely misinterpreted as a diminution of his personal responsibility (Arendt 1963). But what Arendt really meant has more to do with another dimension of evil: the absence of thinking (that means, of self-reflection and of conscience). Thus, Arendt presents Eichmann as a specific type of modern person who has lost their relationship to the world as a 'common good.'

3. *Arendt as a Political Thinker of Morality*

At first sight it appears strange to describe Arendt's way of thinking as moral (ethical) because Arendt does not evolve a theory of 'acting correctly.' Neither does she construct a system of values. She does not apply moral standards to judge political thinking and acting. Moreover, she deconstructs the popular meaning of morality by uncovering its etymological roots. It turns out that 'morality' has more to do with being committed to the world than with internalizing values.

Seen from her perspective, acting within the political sphere means to establish civil (civic) manners; this can only be achieved if political action is related to 'home' which may also be called 'the world people share.'

Arendt's principal work 'The origins of totalitarianism' concludes by referring to Heidegger's philosophy. There is a quasi-existential fixed point which is inaccessible to totalitarian rule: the fact of natality. The simple fact that people are born cannot be removed; the birth of a child will continue to represent a new beginning for the world. 'Beginning, before it becomes a historical event, is the supreme capacity of man; politically it is identical with man's freedom… This beginning is guaranteed by each new birth; it is indeed every man' (Arendt 1981, p. 479).

Hence the basic prerequisite for human action is there; it is one of the fundamentals of human existence. The question that arises is: How can this existential potentiality of political action be given shape?

Acting politically is only possible in a public sphere. This sphere provides the common world people share in which thinking in public and acting take place. Citizens have to see to it that this public sphere can be renewed. At this point Arendt differs from Kant's theory of morally acting. Unlike Kant, Arendt does not perceive 'polical community' as a limited place, that is, a city (polis) but as a never-ending process of taking action in a sphere that can only be protected by the citizens themselves. The citizens' concerns arise from a joint interest in the world they live in, a world, which keeps on giving new beginning to them.

Arendt considers a renewal of the public sphere to be the only possibility to counteract the (self-)destructive potential of modern times. The conclusion that can be drawn after experiencing totalitarian rule is that evil in the world can only be faced if people commit themselves to making the world they share fit to live in for their fellow citizens and if they permanently renew this 'habitability.'

The difficulty which has arisen is that the distinction between public and private sphere becomes increasingly blurred: it was intentionally destroyed by totalitarian rule, and similarly, it is jeopardized by the increasing importance of social interests and needs in modern society. Nevertheless Arendt argues, a world that is shared by people can only emerge by creating a public sphere. However, this can only be achieved to the extent to which citizens succeed in agreeing on making freedom the overall objective of their actions.

See also: Anti-Semitism; Balance of Power, History of; Citizenship and Public Policy; Citizenship: Political; Civil Liberties and Human Rights; Civil Society, Concept and History of; Civil Society/Public Sphere, History of the Concept; Democracy; Democracy, History of; Democratic Theory; Dictatorship; Ethics and Values; Freedom: Political; Genocide: Historical Aspects; Marx, Karl (1818–89); National Socialism and Fascism; Nazi Law; Political Thought, History of; Power in Society; Power: Political; Public Interest; Public Sphere: Eighteenth-century History; Totalitarianism; Totalitarianism: Impact on Social Thought; Weber, Max (1864–1920)

Bibliography

Arendt H 1958 *The Human Condition*. University of Chicago Press Chicago (1960 *Vita Activa oder Vom tätigen Leben*. Kohlhanno, Stuttgart, Germany)

Arendt H 1963 *Eichmann in Jerusalem. A Report on the Banality of Evil*. Viking, New York (1964 *Eichmann in Jerusalem. Ein Report von der Banalität des Bösen*. Piper, München, Germany)
Arendt H 1973 *The Origins of Totalitarianism*. Harcourt Brace Jovanovic, New York (1980 *Elemente und Ursprünge totaler Herrschaft*. Piper, München, Germany)
Arendt H 1981 *The Life of the Mind*. Harcourt Brace, New York (1979 *Vom Leben des Geistes*. Piper, München, Germany)
Arendt H 1981 On violence. In: *The Crisis of the Republic*. Harcourt Brace, San Diego, CA (1990 *Macht und Gewalt*. Piper, München, Germany)
Arendt H 1994 *Essays in Understanding*. In: Kohn J (ed.) Harcourt Brace, New York
Bernstein R 1996 *Hannah Arendt and the Jewish Question*. Cambridge, MIT Press, MA
Canovan M 1992 *Hannah Arendt. A Reinterpretation of Her Political Thought*. Cambridge University Press, New York
Forti S 1994 *Vita della mente e tempo della polis. Hannah Arendt tra filosofia e politica*. Angeli, Milan
Lefort C 1988 *Democracy and Political Theory* [trans. Macey D] University of Minnesota Press, Minneapolis, MN
Villa D R 1996 *Arendt and Heidegger. The Fate of the Political*. Princeton University Press Princeton, NJ
Young-Bruehl E 1982 *Hannah Arendt. For Love of the World*. Yale Unversity Press, New Haven, CT (1986 *Hannah Arendt. Leben und Werk*. Fisher, Frankfurt-am-Main, Germany)
Weber M 1978 *Economy and Society*. In: Roth G, Wittich C (eds.) University of California Press, Berkeley, CA (1980 *Wirtschaft und Gesellschaft*. Mohr, Tübingen, Germany)

A. Grunenberg

Aristocracy/Nobility/Gentry, History of

1. Troublesome Terminology

This article concentrates on Europe. No room can be spared for explicative forays to other parts of the world but one should mention the striking similarities between medieval nobility in Europe and the Japanese *samurai*, as well as in modern times export of European aristocratic structures to other parts of the world.

Even within the European core, diverse terminologies reflected different social structures of particular countries. In the British usage, 'nobility' equals 'aristocracy,' the term 'gentry' being used for the landowners without hereditary titles. English terms do not necessarily fit the Continent. In *The New Cambridge Modern History* (1970) J. P. Cooper wrote: 'Though the word noble was usually reserved for the peerage in England, … in France, Poland and other countries it included those without titles who in England were called the gentry.' In this article the term 'nobility' will be used in this comprehensive meaning, corresponding with the German *Adel*, Italian *nobiltà*, or French *noblesse*, all terms originally connected with land ownership. In particular languages there were parallel terms that laid stress on the origins of the nobility. In German *Ritter*, as well as in Swedish *riddare* and French '*chevalier*' or Spanish '*caballero*,' originally signified 'horseman.' In Poland the word *szlachta* stressed upon its hereditary character (from German *Geschlecht* over Czech *šlechta*), while in Sweden medieval *frälse* underlined their freedom from taxes. All these terms have been value-charged. They gave birth to various modern descriptive terms like 'workers' aristocracy' or 'trade-unions' aristocracy. The derivatives of 'gentry' have in English a particular value: e.g., this quotation from a London real estate newsletter: 'Cleared sites (in Battersea) ready for development ... herald for the next area to be *gentrified*.'

2. Ancient-world Origins

The word 'aristocracy' is of ancient Greek origin and signifies the 'rule of the best.' In Homeric times 'the best' signified chiefs of the noble families who pretended to share with the king a descent from the gods, and were also prominent by their wealth and personal prowess. They formed a class of 'horsemen' or 'knights' (*hippeis*) connected by blood and by various community institutions. They governed the state by means of the council of the *gerontes* (the elder). In the eighth and early seventh century BC, social position of the aristocrats was based on their land ownership but also upon commerce, robbery, and piracy. They dominated the communities (*poleis*) and organized colonization. Many factors contributed to destruction of aristocratic rule, like change of military tactics (riders in single combat were replaced by *phalanx* of heavily armed foot soldiers) and ascent of nonagrarian social groups striving for power.

The fate of the aristocracy in ancient Greece had shown the track that many other ruling groups would follow: from undisputed moral and political domination to the rise of rival groups, to loss of oligopoly of power. However, prestige related to ancient roots (real or fictitious) survived and would become a constituent of all aristocracies. Aristocracy as the 'rule of the best' was a moral ideal; if birth was replaced by wealth as the decisive qualification it became oligarchy.

In republican Rome several groups consecutively enjoyed oligopoly of prestige and power. The earliest hereditary estate was the patricians (the *patriciate*) who in the late fifth century BC reached almost complete oligopoly of offices. In the later fourth century, their competitors, the plebeians (*plebs*), also got access to power. The outcome was a sort of convergence of top strata of both estates through matches and family alliances. In the late fourth and the third centuries a new aristocracy was emerging, the *nobilitas*. Its base were great landed estates run by the slaves and by the peasants, more and more dependent as clients. About 30 houses had access to power both

civil and military. Electoral system and honorary unpaid offices secured their domination. Territorial expansion offered them benefits (fruits of power in the provinces) and created new dangers. 'The mighty few' (*pauci potentes*, so wrote Cicero) were being pressed by the *equites*, originally a moneyed group which in stormy times of civil wars strove for power and virtually assimilated to the *nobiles*.

Augustus and his imperial successors changed the role of *nobilitas* by the very introduction of the Imperial Court. Later on, along with spatial expansion of the Roman Empire emerged provincial aristocracies whose role increased when the center was losing its grip on more distant provinces invaded by the 'Barbarians.' Roman traditions influenced medieval and modern vocabularies of elites.

3. *The Early Middle Ages and la Féodalité*

The dissolution of the Roman Empire in the West (fifth century AD) brought about new nobilities from among the retinues of Germanic chieftains and local Roman governors. Their structures became somehow stabilized with emergence of new states. The milieu of princely (royal) households offered opportunities of advancement for diverse social types like the prince's companions, allied tribal chiefs, valiant warriors, and also bondsmen or slaves who had a chance to serve the prince's person directly. A parallel but closely interwoven network was created by the Church. The bishops exercised secular control over large territories, and the Church offered a convenient upward path to able and ambitious persons of modest origin. This is true for later periods as well. When after the twelfth century nunneries multiplied, they constituted cache for daughters of good stock who easily became abbesses or prioresses of great prestige and considerable power. Individual success signified family advancement as well. Close relation to the Church has been for many centuries an important factor of aristocratic way of life. In Germany, even Protestant noble families claimed right to particular benefices of the Catholic Church they had traditionally enjoyed before the Reformation.

Medieval and modern aristocracies with their hereditary titles of honor had their roots in the royal household and in the administration of the early state. Office holders strove at permanent and hereditary positions. Instability of the early medieval states helped them. So in Carolingian times 'count' (Latin: *comes*) was a judge who accompanied the Emperor. In the twelfth century emerged other *comites*, provincial governors, and 'viscounts' as their representatives. Border provinces of the Empire were governed by the 'marquesses' (German *Markgraf*), 'Duke,' or (German) *Herzog* which meant 'leader of hosts.' 'Baron' (German *Freiherr*) was either a great landowner or just a substantial 'free man,' i.e., primarily the prince's companion, not his simple subject; it was also an equivalent of 'nobleman.' So the recipients of the *Magna Carta* (1215) were *barones*.

More rigid rules of interpersonal relationships were conceived originally in France in the twelfth, and developed into sophisticated structures in the subsequent centuries. In that system, now often called 'feudal,' whoever counted was somebody's vassal and had his own vassals as well. Everybody knew his specific rung on a long ladder on top of which was the ruler who had a special relationship with God. This formed a power system based on personal relationships, a skeleton of the medieval 'state,' called by the German constitutional historians *Personenverbandstaat*. Diverse alterations and finesses of these rules cannot be analyzed here albeit they determined the vicissitudes of particular aristocracies. Enough to mention the notion of 'pairs' (the king's direct vassals equal to each other and also his companions like the legendary 'Knights of the Round Table') making the council of prince's advisors with juridical privileges, and a pressure group conscious of their common interests. In Norman-dominated countries (most important were England and the Kingdom of Naples) the ruler was recognized as direct *seigneur* of all vassals; in other countries the principle reigned that 'the vassal of my vassal is not my vassal.' The Norman principle became a legal tool whenever the king wished to curb his vassals. In the thirteenth century it helped to create in Naples a state of rather bureaucratic character. The other system tended to become a complicated mixture of conflicting loyalties. It troubled France and Germany.

From the late Middle Ages the nobilities in particular countries were being shaped by factors as different as intensity of money economy and customs or laws of inheritance. Money had a destructive influence upon feudal relationships: debts and mortgage haunted landed estates. A system, later called 'bastard feudalism,' replaced the fiefs by money payments. In the Hundred Years War feudal hosts of France proved helpless time and again against English foot militia and the longbows. Eventually mercenary troops raised by entrepreneur-commanders replaced them. Most successful commanders aspired to aristocratic titles. In Italy, mercenary commanders (*condottieri*) were nobles, even sovereign princes, and they often recruited their workforce among lesser nobles from Romagna, Friuli, or Marche. The nobles were looking for their place in constantly growing state machine.

4. *Early Modern State: Crisis and Adjustment*

Not everybody was, however. The fifteenth and early sixteenth centuries witnessed multiform critical phenomena: rents were falling, so robber knights emerged in Bohemia (they were active among the Hussites) and in East Germany. A revolt of lesser nobles broke out in

many parts of the Empire: princes and bishops eventually won and razed many knights' castles. In the Renaissance, assemblies of estates became all over Latin Europe a power factor parallel to the monarch. The nobles participated in them chiefly through their elected members (along with the clergy and the burghers) but the pairs and titled nobles usually had their personal seats secured. The Assemblies became the forum of open competition between the nobility and the burghers. The nobles of a given province usually aimed at an oligopoly of posts and offices, against the burghers and any outsiders.

Recent research has concentrated attention on the 'crises' of nobility and its particular strata. Precious little remained of the original theses. In most countries the nobles (in the broadest sense) rather successfully adapted themselves to new economic and political conditions but this created major shifts within the estate; thence the 'crisis.' The ideal of an unspoiled and bucolic noble life competed with that of a life of public service. A precondition for state service was legal training, and the gentry largely disdained from formal education. Therefore some rulers encouraged foreign travel or founded special lay colleges (like the Knights' Academy in Siegen or nobles' colleges in Denmark and Sweden). In Elizabethan and Stuart England, young gentlemen crowded Inns of Court, and in Scotland the legal training of young noblemen was encouraged by law. In the seventeenth century German nobles became more prone to acquiring legal training as they needed it to service their princes. Not so for the French *noblesse*.

Internal stratification of the estate was of utmost importance. It differed according to size. Here are some estimates: in Denmark, around 1600, the nobles made 0.25 percent of population; in Bavaria, Saxony, and Holland slightly more of the population were nobles; in Bohemia and in Prussia, in 1700, 1 percent; in Poland 7 percent, and in Castile about 10 percent (see Labatut 1978, Meyer 1973). In some regions (Vizcaya in Spain, Masovia in Poland) membership of the noble estate was massive.

In the sixteenth century, ideas of unspoiled bucolic life competed with that of public service. However, if economic and social differentiation of the estate was great enough, the more mighty organized their lesser neighbors into clienteles. 'Liveries' and noble retinues were forbidden in Tudor England and virtually suppressed in Spain in the mid-sixteenth century. In France, a century later, they failed as a weapon of the princes during the *Fronde*.

The seventeenth century may be regarded as a century of opposition between the 'Court' and the 'Country.' The former became a principal source of prestige and privilege, a center of culture and power. It was never that important in England as it was in France, Spain, and Austria.

In the East, where demesne farming prevailed, the state was 'coercion-intensive' (as Charles Tilly 1990 defines it). In Prussia and Livonia the landed gentry (*Junker*) identified with state and its army. In Russia they were the Tsar's dependent servants and did not become an estate before the mid-eighteenth century. In Poland a sequence of fifteenth-century statutes secured independence from the royal administration. The structure of politics in Poland-Lithuania was shaped by the relationship between the gentry and the magnates. Neither of them were interested in a strong executive, after an elective monarchy was introduced, and the political patronage of the magnates became the dominant factor.

In Sweden, nobility, having virtually no medieval traditions, developed into an estate of civil and military officers through ennoblement of commoners. Absolutist reforms of 1680 got their full support against the aristocracy. The system of rangs (known also in Denmark and later introduced in Russia) served their interests.

In France the *noblesse d'epée* felt endangered by the robins (*noblesse de robe*), or commoners, who were massively buying land and titles of nobility on a large scale. Strict *dérogeance* (ban on improper, disqualifying, sources of income) created economic problems. The high aristocracy had enjoyed military command and governorships of provinces as their *chasse gardée*. But their influence was rather restricted to the Court intrigues, masterfully described in the memoirs of Duke de Saint-Simon (1699–1752).

5. *Pride and Prejudice*

One may argue that noble culture has been based on these two vices. The whole idea of nobility drew from tradition, and much imagination was employed to deepen the roots of each family tree. The Renaissance discovery of Roman literature enriched noble imagination with hosts of ancient heroes. Most seminal were the escapees from burning Troy. When in 1624 a scion of the distinguished Lithuanian family Pac visited Florence, both he and his hosts, the Pazzi, were happy to get 'proof' of their common heroic, Homeric origins. The nobilities favored imagined glorious ancestry: the Germanic Francs (as opposed to the commoners descending from the Celtic Gauls) in France, and in England the Norman knights of William the Conqueror. The Lithuanian gentry descended from a Palemon (of Troy, for that matter) and the Polish one from Japhet son of Noah (as contrasted with Cham for the peasantry and Sem for the Jews). In Italy, France, and Spain noblemen and scholars discussed the relative importance of long, impeccable pedigree, lifestyle, and personal valor.

Of more practical importance were property rights and the rules of inheritance. They determined the economic continuity of families. Chiefly in the south (Naples, the islands, Spain) feudal property was dominant, which in cases of conflict, became a tool of

the monarch. In the England of the early Stuarts, the Court of Wards was an instrument of fiscal oppression of the nobles and the Long Parliament (1640) abolished it instantly. In Naples and Sicily the nobles continuously pressed their Spanish sovereigns for more liberal rules of feudal inheritance.

The preferred rule of inheritance was the entail which treated landed estates as family property and neutralized the tendency to division. However, it limited opportunities to mortgage the real estate and left open the fate of younger sons and daughters. In the English case, entail served landed property well: it kept property intact and the younger sons easily found jobs in business, professions, royal service, and eventually in the colonies. On the Continent, ambitious, less substantial nobles as 'noblemen-errant' searched for service abroad. Count Achatius zu Dohna (from Prussia) and Prince Eugene (of Savoy) were among the best known. In Poland, shared inheritance contributed to the impoverishment of the gentry and the growth of the magnate. The latter secured entail, for themselves, as a legal exemption.

6. *Urban Aristocracies*

The cities had aristocracies of their own. North of the Alps, they were usually called 'patriciate' and in Italy simply *nobiltà*. In the High Middle Ages, before town guilds took over the government, most cities were run by royal officers of noble extraction. The most impressive relics of that early stage and its lifestyle are thirteen towers that dominate the cityscape of San Gimignano and the incredible two constructions still towering over Bologna. In most cities, however (Florence, Basle for example) victorious town guilds destroyed these bastions of noble power, only to create the elites of their own. The large, wealthy town offered great benefits to the people of influence, and led to strong oligarchic trends in urban 'communes.' The top families were often more aristocratic than the aristocrats *sensu strictiori*. In Italy and later also in south Germany, ruling elites introduced the *serrata*, or closed lists of families with exclusive right to city offices in a given city. The most famous was the Golden Book of Venice (1297) but it was by no means unique. The patricians cultivated their special ways and observed *dérogeance* taboos. In Germany, the symbol of urban independence was 'Roland,' the legendary companion of Charlemagne and personification of knightly valor. King Arthur too became the eponym of exclusive clubs (*Artushöfe*). The patricians built country residences and acquired landed property chiefly as a status symbol. It was often apparent that they treated this as a collective enterprise. In the fifteenth and sixteenth centuries, in the Swiss city of Luzern, a small cluster of patrician families raised urban troops and leased them to neighboring powers. The patricians commanded them and collected most of the money paid for their services. Profits from power and politics were closely interwoven.

The *connubium* or marriage market strictly limited to elite families created problems, and between the sixteenth and eighteenth centuries many urban aristocracies from Venice to Lübeck experienced a shortage of skilled candidates for urban offices. The medieval register of noble families centuries later did not secure wealth compatible with their social standing. Therefore the Venetian *nobiltà* created, at the city's expense, a system of social welfare for the noble poor.

In general, the urban patriciate in many respects was similar to the aristocracy. They were closely related, too, because landed aristocrats did not scorn urban money. In Germany, patricians from many cities of the Empire (*Reichsstädte*) enjoyed the status of the *Reichsadel*. The Habsburgs granted aristocratic titles to wealthy burghers who supported them with loans and banking services. Names like Fugger (Augsburg), Doria, and Spinola (Genoa) are only the most famous among many. In some cases the patrician was automatically regarded as a noble: in the tiny imperial city (*Reichsstadt*) of Wels, owners of important local salt (brine) wells were, by imperial patent, regarded as nobles (incidentally, this status was enjoyed by Franz von Papen, who preceded Hitler as German Chancellor). In France some city posts were ennobled almost automatically. Even the most exclusive Maltese Order excepted candidates from Florence, Genoa, Lucca, and Pisa from its very strict *dérogeance* rules.

Members of the urban elite merged with the nobility *tout court* in many ways. They met at Royal Councils, acquired landed estates and prestigious titles, and profited from the inflation of honors in the sixteenth and seventeenth centuries. They defended their exclusive status against bureaucratic encroachments of enlightened absolutism. The elites of some larger cities made a particularly closed, self-conscious caste. The twilight of the patrician civilization was best described by Thomas Mann in *The Buddenbrooks* (1901).

7. *Triumph and Twilight of Aristocracy*

The French Revolution closed a chapter, but not the whole story, of European aristocracy. Although aristocrats were given preference onto the scaffold (even a heiress of Tchernobyl, who visited Paris at a wrong time, lost her head), the First Consul and later Emperor created his own Court and constituted a new stratum of *notables* based on loyalty to the ruler and on merit. Political changes largely destroyed the prestige of French élites whose membership reflected, in the nineteenth century, the vicissitudes of the nation. But still, in 1839, the Marquess Astolphe de Custine argued that the true nobility cannot be bestowed on or purchased. The ruler, he continued, can appoint dukes; education, opportunity, and genius or virtue

can create heroes; but nobody and nothing can make a noble. We had been twelve ducs and pairs, he quoted a Napoleon's appointee, but I was the only nobleman among them. However, in Custine's opinion, such restrictive concept of nobility became a mirage and possibly has never been anything more. Everywhere commoners were being ennobled: bankers, industrial entrepreneurs, statesmen and politicians, army and navy officers. Monarchies, even restricted by the constitutions of the second half of the nineteenth century, needed traditional elites as political support and as an ornament. In Berlin, the last emperor Wilhelm II recreated a very formal and lavish court ceremonial. In Britain, in 1876 Queen Victoria became Empress of India and made Benjamin Disraeli, a statesman of Jewish extraction, Earl of Beaconsfield. Immemorial traditions were being cheerfully invented. In the 'first industrial state' the high nobility had a greater than ever share of land and substantial political influence.

For a contrast, in the East, after the dissolution of the Polish Republic (1795) the three partitioning powers inherited numerous lesser nobles but almost no titled ones. This unusual gap had to be filled and eventually anybody who was able to prove members of the Senate among his ancestors who pursued a noble lifestyle, could become baron or count. 'Austrian Count' became, in Poland, an ironic comment. In general, however, in the monarchies, ennoblement and aristocratic titles were becoming a mere ornament and final prize for distinguished service and/or business performance.

Constitutional and political changes in Europe in 1917–18 and after were destructive for the nobility, and not only those in Europe. For instance, neither republican nor Frankist Spain were solving the problems that were vital for the nobles in former Spanish colonies in Latin America. There was great demand for a sovereign overlord.

After 1918 radical projects of agrarian reforms threatened the foundation of noble interest in many countries but it was destroyed only in the Soviet Russia. The former Russian aristocrat becoming taxi driver in Paris became a literary *topos*. Another totalitarian system, Nazi Germany, finally won support of the Prussian Junkers. In the beginning, Adolf Hitler's party seemed to them too radical, too plebeian, but the preceding Weimar Republic had also been disappointing for them. Nevertheless the Junkers shared nationalistic commitments and a fear of the left, and loyalty to the sovereign persisted as the foundation of their ethos. But in later years, when Germany was losing the war, many of them had to reconsider their attitudes vis-à-vis Hitler's régime.

Perhaps the *Credo* of the late aristocrats eventually challenging the Nazi dictator could have been outlined that way: the aristocracy never collaborates with the tyrant in defense of order; its tradition is on the one hand defense of the people against despotism, on the other defense of civilization against the revolution, that most dreadful of tyrants (or so Astolphe de Custine had had it). The failed coup against Hitler on 20 July, 1944 brought them to a bloody end, and the postwar loss to Poland of the territories beyond Oder-Neisse frontier cut off their economic base.

The conspiracy against Hitler was perhaps the last conscious political action of an aristocracy. The political influence of that class seems lost for good, notwithstanding the distinguished intellectual, scientific, or political performance of their individual members. What remains is the value of their collective tradition and the sentiment or flavor so well presented to the modern world by a *Principe* Giuseppe Tomaso di Lampedusa or a Jean Marquis d'Ormesson.

See also: Bourgeoisie/Middle Classes, History of; Class: Social; Elites: Sociological Aspects; Family and Kinship, History of; Feudalism; French Revolution, The; Inequality; Inequality: Comparative Aspects; Middle Ages, The; Peasants and Rural Societies in History (Agricultural History); Revolutions, History of; Social History; Social Inequality in History (Stratification and Classes)

Bibliography

Adonis A 1993 *Making Aristocracy Work. The Peerage and the Political System in Britain 1884–1914*. Oxford University Press, Oxford, UK

Asch R G, Birke A M 1991 *Princes, Patronage and the Nobility: The Court at the Beginning of Modern Age c. 1450–1650*. Oxford University Press, Oxford, UK

Bush M L 1983 *Noble Privilege*. Manchester University Press, Manchester, UK

Bush M L 1988 *Rich Noble, Poor Noble*. Manchester University Press, Manchester, UK

Cannadine D 1990 *The Decline and Fall of the British Aristocracy*. Longman, London

Cooper J P (ed.) 1970 *The New Cambridge Modern History*, Cambridge University Press, Cambridge, UK, Vol. 4

Dewald J 1996 *The European Nobility, 1400–1800*. Cambridge University Press, Cambridge, UK

Donati C 1995 *L'Idea di nobiltà in Italia. Secoli XIV–XVIII*. Laterza, Roma-Bari, Italy

Goody J, Thirsk J, Thompson E P (eds.) 1976 *Family and Inheritance. Rural Society in Western Europe, 1200–1800*. Cambridge University Press, Cambridge, UK

Labatut J-P 1978 *Les Noblesses européennes de la fin du XVe siècle à la fin du XVIIIe siècle*. 1st edn. Presses Universitaires de France, Paris

Mączak A 1989 *Der Staat als Unternehmen. Adel und Amtsträger in Polen und Europa in der Frühen Neuzeit*. Oldenbourg, Munich, Germany

Meyer J 1973 *Noblesses et pouvoirs dans l'Europe d'Ancien Régime*. Hachette, Paris

Mingay G E 1977 *The Gentry. The Rise and Fall of A Ruling Class*. London

Powis J 1984 *Aristocracy*. Oxford University Press, Oxford, UK

Scott H M (ed.) 1995 *The European Nobilities in the Seventeenth and Eighteenth Centuries*. Longman, London

Stone L, Fawtier Stone J 1964 *An Open Elite? England 1540–1880*. Clarendon Press, New York
Tilly C 1990 *Coercion, Capital and European States, AD 990–1990*. Blackwell, Oxford, UK
Wehler H-U (ed.) 1990 *Europäischer Adel, 1750–1950*. Vanderhoek and Ruprecht, Göttingen, Germany

A. Mączak

Aristotelian Social Thought

Aristotelian social thought refers both to Aristotle's own thinking about society and politics, and to the body of ideas inspired by his works. Since the latter has taken many different forms and often departs significantly from the former, this article will devote a separate section to each.

1. *Aristotle's Social and Political Thought*

There are good reasons for accepting the common characterization of Aristotle as the first social or political scientist. Certainly, none of his predecessors engaged in anything like his systematic study of political life as it is actually lived. Nevertheless, in order to appreciate both Aristotle's own ideas as well as their appeal to later students of society and politics, it is important to recognize the ways in which his understanding of social science differs from its contemporary counterparts.

First, Aristotle integrates rather than separates empirical and normative analysis. As a result, the contemporary distinction between political science and political philosophy would have made little sense to him. Aristotle distinguishes, instead, between what he calls practical and theoretical knowledge. The former, which includes both the study of politics and the study of ethics, takes its bearing from human purpose and choice. It therefore can never reach anything like the precision and certainty that the theoretical sciences, which begin with necessary causes and certain first principles, seek to attain.

Aristotelian social science also diverges from its contemporary counterparts in its reliance on teleological principles. In particular, Aristotle believes that the natural needs and ends of human beings draw them into political life. It is this belief that allows Aristotle to integrate the empirical and normative analysis of social relations, for it leads him to argue that we cannot understand the good life that human beings seek without understanding the dynamics of the ordinary political interactions which make it possible, and that we cannot understand the dynamics of ordinary political life without understanding the good life of virtuous activity that makes political life desirable.

Aristotle develops these ideas in a number of texts, the most important of which are the *Politics* (Aristotle 1956b) and the *Nicomachean Ethics* (Aristotle 1956a). Aristotle and his students also assembled a collection of accounts of the constitutions of hundreds of Greek city-states; but only one of these accounts, *The Constitution of Athens* (Aristotle 1950), has survived.

The key concepts in Aristotle's social thought are community, friendship, justice, and regime (or constitution). Community (*koinonia*) is the term that Aristotle uses to characterize the whole range of human association, from business partnerships and groups of travelers, to families and polities. Wherever people share something, be it interest, pleasure, a sense of the good, or common origins, there is community, according to Aristotle. And wherever community develops, there also develop distinctive forms of friendship and justice, the bonds of mutual concern and mutual obligation that, Aristotle argues, we find in even in the most ephemeral or self-serving forms of human association.

The form of sharing that distinguishes political community, according to Aristotle, is taking turns in the process of ruling and being ruled. When Aristotle argues that human beings are by nature political animals, he is arguing that fully mature individuals will eventually form such communities in their pursuit of the good, unless prevented by external coercion or unfortunate circumstances. The fact that most human beings, apart from the Greeks and a few other Mediterranean peoples, do not organize their lives in this way poses a problem that Aristotle deals with only in a very cursory way; for example, by suggesting that although Asians are intelligent enough to form political communities, they lack the spiritedness or self-assertion needed to replace the despotic empires in which they live.

Aristotle distinguishes among the different forms of political community with reference to two things: the number of people who actually share in the process of ruling, and the claims to rule that they make. His famous sixfold classification of regimes—three correct regimes: monarchy, aristocracy, and polity; and three deviations from the correct regimes: tyranny, oligarchy, and democracy—reflects these principles of differentiation. Behind this relatively simplistic system of classification, however, there lies a much richer and more insightful political sociology. In almost all actual cases that Aristotle discusses, political community takes the form of oligarchy, democracy or some mixture of the two. Democracy, as Aristotle understands it, is not just the regime in which the majority rule; it is the regime in which the egalitarian principles and shared freedom on which the many base their claim to rule shapes everyday social life. Similarly, oligarchy is not just the regime in which the few distinguished people rule; it is the regime in which the inegalitarian principles, and especially the unequal wealth, on which the few base their claim to rule,

shapes social relations. As a result, the regime (*politeia*) represents for Aristotle both an ordering of offices or constitution and the particular way of life promoted by the principles of justice that support this ordering of offices. (The closest modern parallel is probably the way in which Alexis de Toqueville 1988 characterizes democracy and aristocracy in *Democracy in America*.)

Thus when Aristotle argues that a mixed regime, a blending of oligarchic, aristocratic, and democratic political institutions, offers the greatest hope for improving ordinary political life, he is not just seeking to lessen political conflict by giving competing groups a stake in the regime. He is also seeking to improve the standards of political morality within communities by mixing the egalitarian and inegalitarian principles of justice that sustain each group's claims to power. Furthermore, Aristotle suggests that the middle class, where it exists, has an interest in supporting these improved standards of justice associated with the mixed regime. Overall, the mix of class, regime, and moral analysis in the middle books of the *Politics* contain a variety of insights that deserve greater attention from contemporary social and political scientists.

2. *Aristotelian Social Thought After Aristotle*

Although Aristotle founded one of the leading ancient schools of philosophy, known as the Lyceum or the Peripatetics, his social and political ideas were not very influential in the ancient world. His analysis of communal life was so single-mindedly focused on the Greek city-state (*polis*) that it must have seemed irrelevant in a world increasingly dominated by the vast empires created by the Macedonians and the Romans. The afterlife of Aristotelian social thought thus takes the form of a series of revivals and rediscoveries, rather than a continuous tradition of commentary and interpretation.

These revivals and rediscoveries of Aristotelian social thought have played an important role in the intellectual life in at least three periods: the high Middle Ages, the years immediately following the French Revolution, and the decades since World War II. In each of these periods students of contemporary society and politics came to believe that Aristotle's texts preserved fundamental insights that had been lost or obscured by their contemporaries.

2.1 *The High Middle Ages*

Among the Muslim, Jewish, and Christian philosophers of the high Middle Ages, Aristotle's name was accorded a degree of authority that few, if any, secular thinkers have ever received. To most of them Aristotle was 'the Philosopher', and his texts were a compendium of the knowledge of the natural and social world that Holy Scripture could not provide.directly As a result, the most famous thinkers of the Middle Ages, such as al-Farabi, Maimonides, and Thomas Aquinas, all looked at the social world from the perspective of Aristotelian concepts and categories, even when that perspective seemed out of line with the world of feudal privilege, sanctified monarchy, and universal scripture-based religions in which they lived.

There was, however, at least one important political problem facing medieval philosophers that Aristotle had not addressed: the tension between secular and religious authority or, put more broadly, between reason and revelation. Some of the most important and interesting transformations of Aristotelian social thought developed out of the need to address this problem. Thomas Aquinas (1959), for example, used Aristotle's brief and cryptic remarks about the difference between natural and conventional right (in Book 5 of the *Nicomachean Ethics* 1956a) to reconcile reason and revelation. He argued that natural right was that part of the eternal law by which God governs the universe that human beings can know without the aid of scriptural guidance. In doing so, he founded the tradition of natural law thinking that continues to shape Catholic social doctrines, a tradition that is often, rather misleadingly, identified with Aristotle's own thinking about natural right. (When, for example, sixteenth-century Catholic thinkers debated whether slavery violated natural law or not, it was Aristotle's arguments in favor of slavery that proved decisive for many of them.) Marsilius of Padua (1956), in contrast, used Aristotle's theory of political community to justify the independence of politics from religious authority. He argued that it was consent of those who take turns in ruling, rather than any superior knowledge of God's will or the natural world, that sustains political authority. In doing so, he recruited Aristotle as an ally in the struggle for republican government and political autonomy.

2.2 *After the Revolution*

Aristotle's extraordinary authority for medieval philosophers made him a conspicuous target for many Renaissance and early modern critics of scholasticism, an approach to philosophy that Thomas Hobbes often ridiculed as 'Aristotelity.' Although his works continued to be read during these periods, it was not until the end of the eighteenth century and the reaction to the French Revolution that his social thought again played a prominent role in Western intellectual life, at least outside of the tradition of Catholic natural doctrines.

In the wake of the social upheavals of the French Revolution, many social critics came to view Aristotle's claims about the priority of the community to the individual as a healthy corrective to the

individualism and social contract thinking that had been popular among the Revolution's supporters. Echoes of Aristotle's claim that human beings are by nature political animals can be heard repeatedly in the writings of the Revolution's critics—even if they were at the same time careful to keep his defense of republican forms of government at arm's length. For them, Aristotelian social thought served as an alternative starting point for reflection on society and politics, rather than as, for medieval philosophers, an exemplary system of analysis.

2.3 After World War II

The postwar period has seen a variety of attempts to revive Aristotelian approaches to social and political analysis. All of these efforts present themselves as alternatives to what is perceived as the mainstream academic approach to the study of politics, society, and morality. Many are inspired by a sense that modern social science cannot explain—and may have even contributed to—the catastrophes of twentieth-century totalitarianism. For these latter-day Aristotelians, Aristotle's emphasis on the integration of normative and empirical analysis provides a better way of both explaining and resisting the seductions of nihilism and totalitarianism. For some, like Leo Strauss (1953), the key to this alternative social science is Aristotle's understanding of natural right, his sense that there is ultimately a natural basis for our judgments about political morality. For others, like Hannah Arendt (1958), it is Aristotle's understanding of citizenship as direct political engagement that is inspiring. Aristotelian social thought has, accordingly, been very influential among both conservative moralists and radical democrats in the postwar intellectual world.

In more recent years, it has been Aristotle's emphasis on community and the moral virtues that has drawn the most attention from students of social and political life. Aristotelian social thought, in this instance, provides a corrective to the limitations of liberalism rather than to the menace of totalitarianism. The most influential example of this contemporary reinterpretation of Aristotle is Alisadair MacIntyre's *After Virtue* (1984).

Any return to something like the authoritative status that Aristotelian social thought had in the Middle Ages seems extremely unlikely. But as a rallying point for critics of both liberal political philosophies and purely empirical forms of social science, the Aristotelian approach to the study of politics and society is likely to maintain its vitality for some considerable time.

See also: Arendt, Hannah (1906–75); Aristotle (384–322 BC); Community Sociology; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Justice and its Many Faces: Cultural Concerns; Marxist Social Thought, History of; Nostalgic Social Thought; Pragmatist Social Thought, History of; Social Justice; Utilitarian Social Thought, History of; Weberian Social Thought, History Of

Bibliography

Aquinas T Saint 1959 *Selected Political Writings* [trans. Dawson J G]. Blackwell, Oxford, UK

Arendt H 1958 *The Human Condition*. University of Chicago Press, Chicago

Aristotle 1950 *The Constitution of Athens and Related Texts*. Hafner, New York

Aristotle 1956a *Nicomachean Ethics*. Loeb Classical Library, Cambridge, MA

Aristotle 1956b *Politics*. Loeb Classical Library, Cambridge, MA

Bien G 1973 *Die Grundlagen der politischen Philosophie bei Aristoteles*. Alber, Munich, Germany

Jaffa H 1952 *Thomism and Aristotelianism: A Study of the Commentary by Thomas Aquries on the Nicomachean Ethics*. University of Chicago Press, Chicago

MacIntyre A 1984 *After Virtue: A Study in Moral Theory*. University of Notre Dame Press, Notre Dame, IN

Marsilius of Padua 1956 *Defender of the Peace*. Columbia University Press, New York

Newman W L 1887 *The Politics of Aristotle*. Clarendon Press, Oxford, UK, 4 Vols

Rybicki P 1984 *Aristote et la Pensée Sociale Moderne*. Ossilineum, Wroclaw, Poland

Salkever S G 1990 *Finding the Mean: Theory and Practice in Aristotelian Political Philosophy*. Princeton University Press, Princeton, NJ

Strauss L 1953 *Natural Right and History*. University of Chicago Press, Chicago

Tocqueville A de 1988 *Democracy in America*. Harper Collins, New York

Yack B 1993 *The Problems of a Political Animal: Community, Conflict and Justice in Aristotelian Political Thought*. University of California Press, Berkeley, CA

B. Yack

Aristotle (384–322 BC)

1. Life

Aristotle remains the most influential philosopher of antiquity apart from Plato. Reactions against his philosophy marked the rise of modern science and political theory; some contemporary developments in the social and political sciences now involve a revival of aspects central to his thought.

Aristotle was born in Stagira (Chalcidice), the son of Nicomachus, physician to the Macedonian king, in

384 BC. At the age of 17 he moved to Athens and, for 20 years, studied and taught in Plato's Academy, the international centre of science and philosophy at the time. On the basis of his encyclopedic knowledge, Aristotle discussed problems posed by Plato, raising objections against his teacher's theory of Forms. At this stage, Aristotle developed his own positions in areas of the philosophy of nature, of ethics, metaphysics and rhetoric. He also developed the first known system of formal logic (syllogistics) and collected rules of dialectical and rhetorical argumentation (*topoi*).

After Plato's death, Aristotle left Athens for the first time, travelling for 12 years in the eastern Aegean. He stayed briefly (343–2 BC) at the Macedonian court of Philip II as tutor to his son Alexander (the Great).

Political changes allowed Aristotle to return to Athens, where he set up his own school in the Lyceum. Probably all his major treatises were written or finished during this period (335–4 to 323–2 BC), but of Aristotle's extensive oeuvre only a fraction survives. After Alexander's death and surges in political opinion against Macedonia, Aristotle was accused of impiety. To save the Athenians from committing a second crime against philosophy (referring to the death of Socrates), Aristotle left Athens for Chalcis where he died (322) at the age of 63.

2. *General Contribution to Knowledge*

Aristotle's philosophy represents a unity between systematically assembled and compared arguments, empirical data concerning language, nature and society, and a search for causes and first principles in terms of which these data can be understood. This effort is guided by an unprecedented quest for conceptual precision without loss of systematic and terminological flexibility. He establishes criteria and methods of inquiry that for centuries defined scientific knowledge and procedure. Aristotle's inquiries are the first to compartmentalise knowledge into separate theoretical, practical, and productive disciplines such as mathematics, the philosophy of nature, metaphysics, ethics, and politics, rhetoric and poetics. Aristotle is also the first to lay the foundations for psychology, biology, botany, anatomy, zoology, and geology.

Aristotle's ethical and political investigations and methods of inquiry in the area of human action have now become disputed paradigms of practical philosophy, but are of particular relevance for the social sciences. In contrast to an understanding of social/political sciences as independent of ethics, Aristotle argues for their mutual interdependence. He also elucidates the sense in which they deal with areas of knowledge that make specific claims to conceptual and methodological precision and truth. Theoretical accounts in these areas tend not to be exact in the same sense in which a science can be understood as 'perfect knowledge,' but nor are they imprecise. This special mode of reasoning does not render them inherently deficient as forms of knowledge. On the contrary, it makes them particularly suited for completing their proper task, for their goal is not theory pursued for its own sake but, as Aristotle argues, the alteration of praxis.

2.1 *Theory of Scientific Knowledge*

Aristotle regards it as a fact that 'all human beings naturally desire to know.' The resulting types of knowledge attain their ultimate end in an understanding of truth which implies universality and necessity, since it is based on knowledge of first principles. Although these are evident *per se*, they are frequently not well known to us. Scientific procedure, as described in the *Analytica*, involves two intrinsically related activities: the inductive-abstractive process (*epagoge*) leading from data commonly known to first principles which are frequently unknown, and the deductive process (*apodeixis*) of inference from self-evident first principles to subsequent statements. The first procedure originates in perception, imagination, memory, and experience, when the active intellect grasps the universal by collecting, comparing, and abstracting it from the manifold data presented to it by the passive intellect. Conversely, the deductive procedure of science that, for the most part, follows rules of formal logic, begins with first principles grasped by the intellect and deduces further sets of valid statements from them.

Although the program of presenting areas of scientific knowledge as deductive systems gave rise to the notion of science as knowledge *more geometrico*, only small portions of Aristotle's own scientific and ethicopolitical writings fit this pattern.

For Aristotle, the search for first principles, the clarification of difficulties involved in problem-solving, the elimination of unacceptable premises, and consensus are achieved most efficiently by disciplined discourse between researchers. Aristotle's method of dialectics, designed specifically to serve this purpose, involves continuous training in the use of *topoi*, mainly listed his *Topica*. They are especially suited for logically coherent reasoning based on reputable opinions, which may form the starting points of scientific inquiry or may represent essential elements of analysis within disciplines related to social sciences, such as ethics, politics, or rhetoric.

2.2 *Theoretical Sciences: Physics and Metaphysics*

Aristotle's treatise *Physics* attempts to explicate causes and principles of nature (*physis*). This involves the analysis of essential features of the process of natural change and of its sources which include teleology. In

the Aristotelian tradition, distinctions made in this context were also used for describing sociopolitical change. Whatever is in motion, changes from an actual state in which it is not yet whatever it is potentially, to another actual state in which it has become what it was potentially, provided that no obstacle occurs. In this context, 'matter' is the reality underlying all physical change, although matter is never actually found unformed. Whenever a new form is acquired, this also means the privation of the previous form. Matter (such as the bronze of a statue: material cause) and form (its shape: formal cause) are sources of change, just like that which initiates it (the artist's action: efficient cause) and the end for the sake of which the forming of the item was begun (final cause). Finally, Aristotle discusses the notion of an ultimate source of motion in the universe, which leads him to a concept of a prime unmoved mover different from nature.

The treatises collected as *Metaphysics* describe differing but related approaches to a 'first science' which underlies any particular area of knowledge: a science of principles and causes attributable to every being qua being (ontology), of basic constituents of reality (theory of substance), and of its prime constituent (theology). Elements of metaphysics comprise a theory of principles considered to be universal in thought and reality, such as the principle of non-contradiction. Aristotle's theory of substance distinguishes between substance as primary reality and its properties (accidents). 'Substance' denotes a concrete individual of a natural kind or the species that it exemplifies. Real in the strict sense are individual substances which exist actually and not only potentially. The highest rank among them is taken up by an entity characterized by immaterial, immutable, separate, and intelligent actual existence, called God.

Although Aristotle's theory of substance was replaced by later ontologies, his theoretical terminology of matter, form, potentiality, actuality, and causality had a lasting impact on the language of science and philosophy. The substantialist notion of individuals of a natural kind persists until today, and is implied by aspects of contemporary debates on the respective roles of social structure and human agency.

2.3 Practical Sciences: Ethics and Politics

The science presented in the *Politics* is based on a study of the history and development of 158 constitutions of Greek states, the *Constitution of the Athenians* being the only one that survived. It deals with the city-state (*polis*) which exists for the sake of the good life of its citizens, the culmination of a natural process of congregating. People can only flourish as human beings within this context, Aristotle considers. Flourishing as a human being (*eudaimonia*) is the ultimate end of human action. It is the primary concern of a type of ethical reflection that finds its most celebrated form in Aristotle's *Nicomachean Ethics*. Flourishing is based on intellectual and moral virtues. They are neither innate—as if teleology involved deterministic development—nor against nature. They must be acquired by instruction, by deliberate choice, and practice within a given ethos of a community. They include bravery, magnanimity, justice, or friendship. For the most part, Aristotle understands moral virtues as habits of manifesting, relative to our own capabilities, the optimum of a mean between too much and too little in any situation involving action or emotional response (this does not imply a quantifiable average, valid for everyone).

Whereas ethics studies forms of excellence as a human being, excellences of character and related intellectual habits such as practical wisdom (phronesis), political science examines societal institutions, constitutions, forms of government, and virtues as a citizen which are necessary for the good life and for the well-being of the whole community. Both disciplines attempt to explain how under ideal and less than ideal conditions these ends may or may not be achieved. They are intrinsically connected. Insofar as constitutions and legislation are committed to envisaging the city-state as 'a community in a good life in houses and villages with the aim of a life perfected in itself,' ethics is the discipline which also provides norms of human excellence for the household (economics) and for politics. Insofar as the notion of political science frequently takes on the wider meaning of practical philosophy it covers not only strategy, economics, and rhetoric but also ethics.

Since it is a characteristic feature of a human being not only to be a living being capable of reason (*zoon logon echon*) but also a being which dwells in city-states (*zoon politikon*), Aristotle distinguishes between two basic forms of life through which a person's ultimate end may be achieved: the theoretical life of contemplating unchangeable truth and the political life. Although the theoretical life promotes the highest form of activity, each form has a specific dignity of its own.

Not all forms of state, distinguished by differences between types of constitution or their distribution of offices, are equally conducive to a good life. A city-state may be governed by a single ruler, by small, or by large groups. These may govern in the interest of all or in their own interest. Consequently, Aristotle distinguishes between three basic features of correct government existing for the common good: kingship, aristocracy, polity, and three of deviant, self-interested character: tyranny, oligarchy, and democracy. His analysis of subtypes reveals their differing socio-economic bases and inherent tensions between rich and poor. Because of their effects on the polis, he pays special attention to sources of revolution and socio-political stability. These inquiries lead him to favor a 'mixed' constitution (polity) providing for peace.

Those who are neither rich nor poor should hold the balance of power, which is based on collective decision-making and the public promotion of human excellence. Since it is in conditions of peace that men live well, any city-state should provide for self-sufficiency and self-defence and hence avoid both economic or cultural dependence, as well as expansionist policies which risk enlarging the state, thus destroying it.

2.4 Productive Sciences or Arts: Rhetoric and Poetics

For Aristotle, rhetoric is the medium of political reasoning and hence a necessary element of political life, which depends on joint deliberations and decisions on the basis of public speeches for and against what should be done or avoided, praised or blamed, accused or defended. Consequently, the art of public discourse described in his *Rhetoric* is rooted in political science and ethics and the art of dialectical argumentation. Rhetoric consists in a method of discovering and convincingly presenting deliberations or presenting a case in the Assembly, on festive occasions of the state or at Court, and these concern matters that were, are or will be conducive or inimical, to the good life in a city-state. Knowledge of the rhetorical method also lays bare the structure of discourse that leads to error. This art is necessary for a culture of political oratory, in turn necessary for optimising decision-making, particularly in areas of human action where undisputed certainties are unavailable and reasonable arguments based on reputable opinions or probability are as much required as character and emotional commitment. Thus, Aristotle's paradigmatic type of rhetoric is the deliberative genus of the Assembly, rather than doctrinal teaching or forensic oratory.

The art described in the *Poetics*, dealing centrally with a theory of tragedy, has an equally practical and political function. Poets, in Aristotle's view (in opposition to Plato's), count among the best educators of a people. The specific task of a tragedy consists in leading members of the audience—while they empathize with the rise and fall of the hero of the drama—through woeful compassion and fearful alarm to a cleansing of their own most profound emotive dispositions to the way they live their lives as human beings in their polis and their cosmos. The dispositional balance reached by this process helps to reorganize citizens' relations to human flourishing. It also creates a bond of unanimity, necessary for the good life of a community.

Although rhetoric and poetics are arts (*technai*), they are not like mechanical arts for producing objects for use or consumption. their products are characteristic of human beings *qua* human: they are speeches or actions which relate to human praxis either thematically or representationally (*mimesis*), by revealing basic tensions and truths related to living a life as a human being. They are indispensable aids for the sociopolitical invention of a good life in finite time and the formation of cultural identity. Social science disregards their constitutive function for sociability only at the risk of substantial theoretical loss.

3. Aristotle and the Contemporary Social Sciences

No one denies the historically contingent character of Aristotle's theories or his practical concern with analyzing conditions of living well in a polis-community rather than in a modern nation-state. Yet neither are his thoughts entirely moulded by his historical situation, nor does the qualitative difference between the polis and modern society render them obsolete. Aristotle's descriptive and analytical work, particularly in connection with his theory of human action and his typology of aspects of constitutions, made the *Nicomachean Ethics* and the *Politics* classics of ethics and political theory. Basic concepts of Aristotle's practical philosophy are discussed in contemporary philosophy and political science.

Analytical as well as narrative theories of action, of human agency and the self (Ricoeur 1992) are based on Aristotle's ethics and poetics, which highlight the fact that a discussion of narrative and of its constituent structures not only concerns the nature and scope of social and political theory but also raises crucial questions about personal and cultural identity. Moreover, the focus on the practical impact of the relation between thought, emotions and character apparent in the *Rhetoric* is now reappearing in feminist and other contemporary approaches to sociology.

Insofar as a hermeneutical consciousness of human interaction is a necessary element of discovery and explanation in the social and historical sciences, they share family resemblances with practical rationality (*phronesis*) as described by Aristotle (Gadamer 1975). They involve research that strictly cannot be separated from the persons who possess and share this knowledge. Consequently, scientific reasoning about social interaction and institutions involves rhetorical forms of argument or reasoning with a practical purpose (Edmondson 1984).

The fundamental distinction introduced in the *Nicomachean Ethics* between praxis and productive activity, besides supplying reasons for esteeming praxis highly, functions as a critical tool for understanding the genesis of contemporary Western societies and their potential of providing a good life under conditions of labor (Arendt 1958).

Reference to an Aristotelian concept of praxis is also prominent in reconceptions of the political in communitarianism (MacIntyre 1985). They highlight the constitutive role of practices for a community, insofar as they are shared activities that are undertaken

not as means to an end but as choice-worthy in themselves. Their perceived importance for human well-being involves a reconsideration of central issues of Aristotelian virtue ethics and its guiding concept of the good life, as well as an analysis of features essential for functioning as a human being. For the most part, such reconceptions attempt to integrate sociocultural pluralism, involving a multiplicity of competing ideas of the good life. Such pluralism requires a conception of politics as basically deliberative, since reasoned social and political decision-making depends on creating joint convictions in areas of the contingent. This process depends on a rhetorical culture as a constitutive element of political culture. Since Aristotle's *Rhetoric* represents a synthesis of conceptual frameworks necessary for understanding the functioning of rhetorical culture, he provides a paradigm for developing a theory of political deliberative argumentation under pluralist sociopolitical conditions. Under the title of '*topoi*,' Aristotle's heuristic for systematically exploring and presenting whatever is potentially convincing has entered theories of (legal) argumentation (Perelman 1969) and social science/political science research (Hennis 1977). Even authors with a critical distance from neo Aristotelianism nonetheless adopt Aristotle's rhetorical heritage by discussing notions of deliberative democracy or politics. The scope and rigor of Aristotle's thought will no doubt continue to inspire generations of social scientists.

See also: Aristotelian Social Thought; Causation: Physical, Mental, and Social; Citizenship: Political; Counterfactual Reasoning: Public Policy Aspects; Counterfactual Reasoning, Qualitative: Philosophical Aspects; Democracy; Democracy: Normative Theory; Ethics and Values; Idealization, Abstraction, and Ideal Types; Identity and Identification: Philosophical Aspects; Individualism versus Collectivism: Philosophical Aspects; Knowledge (Explicit and Implicit): Philosophical Aspects; Knowledge Representation; Meaning and Rule-following: Philosophical Aspects; Models, Metaphors, Narrative, and Rhetoric: Philosophical Aspects; Person and Self: Philosophical Aspects; Personal Identity: Philosophical Aspects; Policy History: State and Economy; Power: Political; Practical Reasoning: Philosophical Aspects; Responsibility: Philosophical Aspects; Rhetoric; Rhetorical Analysis; State and Society; State: Anthropological Aspects; State Formation; States and Civilizations, Archaeology of; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects; Virtue Ethics

Bibliography

Arendt H 1958 *The Human Condition*. University of Chicago Press, Chicago
Aristotle 1926/1965 *The Loeb Classical Library*, *Greek Authors*, 17 Vols. Harvard University Press, Cambridge, MA
Barnes J, Schofield M, Sorabji R (eds.) 1975/79 *Articles on Aristotle I–III*. Duckworth, London
Edmondson R 1984 *Rhetoric in Sociology*. Macmillan, London
Flashar E 1983 *Äeltere Akademie*, *Aristoteles—Peripatos*. In: Flaschar E (ed.) *Grundriss der Geschichte der Philosophie*. *Die Philosophie der Antike*. Schwabe, Basel
Gadamer H G 1975 *Truth and Method*, 2nd edn. Sheed and Ward, London
Guthrie W K C 1990 *A History of Greek Philosophy, Vol. VI: Aristotle. An Encounter*. Cambridge University Press, Cambridge, UK
Hennis W 1977 *Politik und Praktische Philosophie*. Klett-Cotta, Stuttgart
Keyt D, Miller F D (eds.) 1991 *A Companion to Aristotle's Politics*. Blackwell, Oxford, UK
MacIntyre A 1985 *After Virtue*. Duckworth, London
Perelman Ch 1969 *The New Rhetoric. A Treatise in Argumentation*. University of Notre Dame Press, Notre Dame, IN
Ricoeur P 1992 *Oneself as Another*. University of Chicago Press, London
Totok W 1997 *Handbuch der Geschichte der Philosophie*. Klostermann, Frankfurt, Germany, Vol. 1, pp. 359–466
Wőrner M 1990 *Das Ethische in der Rhetorik des Aristoteles*. Alber, Freiburg/Munich, Germany

M. Wörner

Arms Control

Arms control, a term popularized in the early 1960s, may be defined as the effort, between and among countries, to negotiate binding limitations on the number and types of armaments or armed forces, on their deployment and disposition, or on the use of particular types of armaments. It also includes measures designed to reduce the danger of accidental nuclear war and to alleviate concerns about surprise attack. Although the two terms are often used interchangeably, it is distinct from disarmament, which has the more ambitious objective of seeking to eliminate, also by international agreement, the means by which countries wage war (Blacker and Duffy 1984). The goal of eliminating war extends far back in history, but in modern times disarmament came into focus with the Hague Peace Conferences in 1899 and 1907. These and subsequent efforts met with what can only be termed limited success.

Schelling and Halperin, in their seminal work, *Strategy and Arms Control*, first published in 1961, listed three objectives for arms control: to reduce the likelihood of war; to limit the extent of damage should war occur; and to reduce expenditures on military forces. As US and Soviet weapons arsenals mushroomed during the Cold War and each country spent liberally to keep pace with the other militarily, later analysts tended to offer less sweeping goals. These

included reducing the number of nuclear weapons and redirecting the arms race into areas less likely to threaten the stability of the international system.

The end of the Cold War and the collapse of the Soviet Union gave rise to expectations, both at the expert level and among publics more broadly, that the sudden cessation of the superpower arms race would create opportunities for radical reductions in the nuclear and conventional weapons arsenals of the major powers.

Although the leading industrialized countries have taken some steps in that direction—US armed forces, measured in terms of total numbers, have declined by one-third since 1990—progress in eliminating the nuclear-weapons stockpiles of the US and Russia has proven to be an elusive goal. To compound the problem, the number of nuclear-armed states has actually grown in recent years as India and Pakistan, in a series of highly publicized weapons tests in 1998, announced their arrival as full-fledged nuclear powers. The means to deliver these weapons, as well as chemical and biological agents, across hundreds (and even thousands) of miles has also spread as countries such as North Korea, Iraq, and Iran continue to invest heavily in programs to develop and deploy long-range ballistic missiles.

Arms control, particularly between the US and the Soviet Union, aroused controversy from the outset. The Limited Test Ban Treaty, signed by representatives of the US, the United Kingdom, and the Soviet Union in 1963, provoked sharp debate when US president John F. Kennedy submitted this arms control 'first' for Senate approval. Some critics saw the treaty, which prohibits the testing of nuclear weapons above ground, under water, and in space, as unenforceable. These and other concerns notwithstanding, the treaty eventually was ratified and fears of Soviet noncompliance proved to be unfounded.

More favorably received in the US was the United Nations-sponsored Nuclear Non-Proliferation Treaty (NPT), which sought to restrict the size of the 'nuclear club' by inducing non-nuclear weapons states to renounce the acquisition of such weapons in exchange for a commitment (among other pledges) on the part of the nuclear-weapons countries to reduce their own arsenals. The NPT, which entered into force in 1970, was renegotiated in 1995, at which time the signatory states agreed to extend the treaty's provisions for an unlimited period of time. Conspicuous by their absence as states-parties to the NPT are the two newest declared nuclear powers, India and Pakistan, as well as Israel, which is believed to possess a small but sophisticated arsenal of nuclear weapons.

The NPT was critically important in facilitating the start of bilateral US–Soviet negotiations in 1969 to limit central strategic forces. Known by the acronym SALT, for Strategic Arms Limitation Talks, the negotiations resulted in two arms-control agreements in 1972. The first and more important was the Antiballistic Missile (ABM) Treaty, by which the two countries agreed to limit the number of ABM sites and thus not deploy nationwide defensive systems to protect their homelands against nuclear-missile attack. The second accord was a 5-year freeze on the construction of long-range land- and sea-based ballistic missile 'launchers' (underground silos and submarine missile tubes, respectively). The so-called Interim Agreement on Offensive Weapons was a temporary measure to slow the competition in offensive weaponry, pending negotiation of a more permanent and restrictive treaty (Newhouse 1973).

The second phase of the negotiations, lasting from 1972 to 1979, led to the signing of several agreements, including an accord further limiting US and Soviet ABM deployments and the 1974 Threshold Test Ban Treaty, which restricts the yields of underground nuclear weapons tests. The most important agreement concluded during this period was the 1979 SALT II treaty, a lengthy and complex document that attempted to extend and refine many of the provisions of the 1972 Interim Agreement (Talbott 1979). The US Senate never ratified the treaty, however, largely because of the dramatic deterioration in superpower relations that began during the abbreviated administration of President Gerald Ford and accelerated during the term of his White House successor, Jimmy Carter. Despite the ambiguous legal status of the treaty following the failure to obtain its ratification, both the US and the Soviet Union abided by most of its provisions well into the 1980s.

Negotiations on US and Soviet strategic nuclear forces, renamed the Strategic Arms Reduction Talks (START) by President Ronald Reagan, resumed in the summer of 1982. Through the remainder of Reagan's presidency, the two sides reached consensus on many key points, including the desirability of 50 percent reductions in long-range nuclear forces and the need for intrusive, on-site inspections to prevent cheating.

Among the issues not resolved was how to construct the preferred relationship between strategic offensive and defensive forces. In a dramatic reversal of its earlier negotiating position, the US now favored the rapid development and deployment of nationwide defensive systems, as shown by its sponsorship of the Strategic Defense Initiative (SDI), first outlined by Reagan in March 1983 in a nationally televised address (Fitzgerald 2000, Nolan 1989). The Soviet Union was sharply critical of SDI and resisted the conclusion of any new agreement to reduce strategic offensive weapons, pending a commitment by the US to abide by the terms of the ABM treaty, narrowly interpreted, through at least the end of the century.

The rise to power of the Soviet leader Mikhail Gorbachev in 1985 signaled a fundamentally new phase in arms control. One of Gorbachev's most important foreign policy objectives was to curtail sharply the political rivalry between the USSR and the

West by, among other steps, delimiting their military competition. The first major result of this policy was the Intermediate-Range Nuclear Forces (INF) Treaty, concluded in 1987, which eliminated all US and Soviet land-based nuclear missiles with ranges between 500 and 5,500 kilometers. This was followed in 1990 by a treaty to reduce conventional forces in Europe (CFE), signed by 22 member-states of the North Atlantic Treaty Organization (NATO) and the now-defunct Warsaw Pact. At about the same time, US and Soviet negotiators completed work on the long-awaited START treaty, and in July 1991, President George Bush traveled to Moscow to join Gorbachev in signing the accord. Less than 6 months later, Gorbachev—the architect of the most ambitious reform program in the 74 year history of the Soviet Union—resigned as president and the USSR ceased to exist.

At a hastily called summit meeting in June 1992, Bush and Russian Federation president Boris Yeltsin agreed to press for early ratification of the START treaty. They also pledged to conclude a second and more ambitious agreement to reduce US and Russian strategic nuclear arsenals by up to two-thirds within a decade and to eliminate all multiple-warhead land-based missiles. Some 2 weeks before the inauguration of Bill Clinton as US president in January 1993, the two sides made good on their promise and concluded the START II treaty. As relations between the US and Russia deteriorated during the second half of the 1990s, the treaty itself languished; although finally ratified by both sides, nearly a decade after its signing most of the agreement's provisions still had not been implemented.

The US–Russian arms-control agenda has also been complicated by the confused, and confusing, nuclear legacy of the Soviet Union. In the wake of the Soviet collapse, four of the country's now independent republics, including Russia, found themselves in possession of thousands of nuclear weapons and hundreds of long-range delivery systems. Under pressure from Russia and the West, in May 1992 the governments of Belarus, Kazakhstan, and Ukraine promised to abide by the terms of the 1991 START I agreement and to join the NPT as non-nuclear weapons states. With financial and technical assistance from the US, all nuclear warheads and their associated missile systems deployed on Belarusian, Kazakh, and Ukrainian territory eventually were disarmed and dismantled. The denuclearization of the three former Soviet republics constitutes the most important—and perhaps the only unambiguous—arms control success story of the 1990s.

The abrupt end of Soviet rule gave rise to a second kind of security problem for which policymakers everywhere were ill prepared. The frequency and intensity of ethnically and religiously inspired conflicts in such heretofore remote parts of the former USSR as Tajikistan and the southern Caucasus increased dramatically following the collapse of central authority in Moscow. The breakdown of political control in these and proximate regions, coupled with the extreme poverty that afflicted many of the people caught in the fighting, served both to prolong these struggles and to frustrate diplomatic efforts to contain them.

In addition, the unraveling of the Soviet Union's alliance relationships left a number of countries politically orphaned and therefore less secure. This too served to increase regional instability. It also encouraged some states—North Korea being a case in point—to seek to develop nuclear weapons of their own despite the determined opposition of the major powers, especially the US.

At the start of the twenty-first century the most important challenge to arms control, however, is the prospective change in the relationship between long-range, offensive nuclear forces and weapons designed to defend against them. For the last 40 years, what Philip Green characterized in the mid-1960s as the deadly logic of nuclear deterrence helped preserve the uneasy truce between Washington and Moscow. In its simplest form, deterrence held that no rational leadership in possession of nuclear weapons would ever intentionally authorize their use against a nuclear-armed adversary because of the near-certain knowledge that the victim of such an attack would retaliate in kind. With effectively no ability to ward off or deflect such a retaliatory strike, the would-be aggressor would thus be deterred from initiating a nuclear exchange in the first place (Brodie 1965, Schelling 1960).

As the capacity to build nuclear weapons spreads to countries other than the so-called great powers, interest in acquiring the means to defend against their possible use has grown, particularly in the US. Even a comparatively modest system of active defenses—one designed to defeat a handful of incoming ballistic missile warheads launched from whatever quarter—arouses concern, however, because of the latent ability of such a system to expand and improve over time. The larger and more robust a system of national missile defense becomes, the better it will be at defending against more complex threats. The more able it becomes, in other words, the more threatening it will seem to the more established nuclear powers, including Russia and China, for whom the unchallenged ability to retaliate with overwhelming force constitutes the bedrock of their security (Wilkening 2000). According to the classic tenets of arms control, the large-scale deployment of strategic defensive systems could therefore erode stability and increase the likelihood of war.

It is probably not beyond human ingenuity to design and construct a 'mixed' strategic environment that allows for a modicum of defense while preserving nuclear deterrence in its essentials. Given the understandable urge to escape the persistent threat of nuclear annihilation, it seems safe to assume that governments will persist in their efforts to square this

strategic circle. Policymakers everywhere would do well to remember, however, that nuclear deterrence, for all its flaws, has kept the peace for half a century, and that any attempt to replace it with something else is likely to entail serious risks and potentially enormous costs.

See also: Conflict and War, Archaeology of; Conflict/Consensus; Geopolitics; Military and Politics; Military Geography; Military History; War: Causes and Patterns

Bibliography

Blacker C D, Duffy G (eds.) 1984 *International Arms Control: Issues and Agreements*, 2nd edn. Stanford University Press, Stanford, CA

Brodie B 1965 *Strategy in the Missile Age*. Princeton University Press, Princeton, NJ

Fitzgerald F 2000 *Way Out There in the Blue: Reagan, Star Wars, and the End of the Cold War*. Simon & Schuster, New York

Green P 1966 *Deadly Logic: The Theory of Nuclear Deterrence*. Ohio State University Press, Columbus, OH

Newhouse J 1973 *Cold Dawn: The Story of SALT*, 1st edn. Harper & Row, New York

Nolan J E 1989 *Guardians of the Arsenal: The Politics of Nuclear Strategy*. Basic Books, New York

Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA

Schelling T C, Halperin M H 1961 *Strategy and Arms Control*. Twentieth Century Fund, New York

Talbott S 1979 *Endgame: The Inside Story of SALT II*, 1st edn. Harper & Row, New York

Wilkening D 2000 *Ballistic Missile Defense and Strategic Stability, Adelphi Paper 334*. International Institute for Strategic Studies, London

C. D. Blacker

Aron, Raymond (1905–83)

The life and works of Raymond Aron coincide with the period of conflicts generated by ideologies. Born in 1905, twelve years before the Bolshevik Revolution, he died in 1983, in the middle of the European missile crisis, the last event of the Cold War before the fall of the Berlin Wall in 1989.

1. Chronology

Raymond Aron was born on March 14, 1905 in Paris into a family of Jewish origin, completely integrated into patriotic and republican society. A brilliant student, he went to the Ecole Normale Supérieure, where he became friends with Sartre and Nizan, and then went on to the *Agrégation de Philosophie*. Germany in the 1930s revealed to him the violence of history, and drew him into a critical attitude and a personal approach which made him unique among the French intellectuals of the twentieth century. In the tradition of Montesquieu, Constant, Tocqueville, and Elie Halévy, he is the most eminent representative of liberal thinkers in France in the twentieth century.

From 1930 to 1933, Aron stayed in Cologne and in Berlin. The rise of Nazism led him to break with the socialism and pacifism of his youth. Reading Max Weber and the phenomenologists—particularly Husserl and Heidegger whom he introduced in France—took him away from the idealism and positivism which at that time dominated the French academic philosophy. The doctoral thesis he defended in 1938 dealt with the philosophy of history; it created a scandal in the French university by using epistemological doubt to criticise positivism in the field of social sciences. Called up in 1939, Aron answered General de Gaulle's call in June 1940 and reached London, where he edited the review *La France Libre* until the Liberation. The Second World War represented a major upheaval, with the quadruple shock of defeat, exile, dismissal from the university following the status of Jews as applied by Vichy, and finally of genocide. During the Cold War, Aron identified himself, with André Malraux as one of the few well known intellectuals to oppose the attraction of communism in France and to participate in the Congress for Cultural Freedom, created to contain Soviet influence throughout the world. As a result this isolated him completely. He pursued a double career up to his death in 1983, first as a university professor, at the Sorbonne and later at the Collège de France, after an interim period at the Ecole des Hautes Études, and second as a journalist, at *Combat*, then at *Figaro* (1947–77) and finally at *L'Express* (1977–83). He was faithful to the choice of vocation he had made in the 1930s, to be a committed witness trying to reflect on history and politics as they happened. He was fully recognized outside France where he was held by academics and politicians to be an interlocutor of the first order. At the end of his life, Aron became reconciled with French intellectuals, at the same time as they were converted to the fight against totalitarianism following the disclosures on the Gulag by Solzhenitsyn as well as with the general public which welcomed enthusiastically his *Mémoires* (Aron 1983). He died on October 17, 1983 in Paris.

2. Works and Beliefs

Aron defines his works as:

> a thought on the twentieth century, in the light of Marxism, and an attempt to throw light on all areas of modern society: economics, social relations, class relationships, political systems, relationships between nations, and ideological discussions.

Freeing himself from the traditional separations between disciplines, his thoughts have covered many

areas of knowledge, mainly philosophy, sociology, international relations, ideological controversy, and commentaries on current events. They find their unity, however, in the idea of the condition of man as presented in his thesis, *Introduction à la Philosophie de l'Histoire* (1938), whose meaning can be summarized in the formula. 'Man is in history; man is historical; man is a history.' Human existence is tragic, as each of us is forced to make decisions about one's destiny from partial knowledge and with limited powers of reasoning. For all that, it is not condemned to nihilism and despair, as one's commitment allows one to overcome the relativism of history and knowledge to reach a portion of freedom and truth. According to Aron, freedom is first, this primacy being historical and not metaphysical. The idea of modern freedom became clear in Europe in the Age of Enlightenment, then gained strength in the presence of the industrial society of the nineteenth century, and later with the resistance against totalitarianism.

Aron stands out by his early understanding of ideologies whose hostility to democracy give structure to the twentieth century. From the end of the 1930s, he showed, with Elie Halévy, the novelty and the common features which united Fascism, Nazism, and Communism, while proving that before all they were fighting democracies. Contrary to Hannah Arendt, Aron does not think of either democracy or totalitarianism in terms of essence, but as historical constructions mixing a general design, institutions, and the action of men such as Lenin and Stalin, Mussolini and Hitler, Churchill and de Gaulle. After the war, he was also the first French intellectual to give a correct interpretation of Cold War—which he defined with the formula 'Peace impossible, war unlikely'—and of nuclear deterrence. Through his critical commentary of Marx—where he separates the sociologist of the industrial civilization from the prophet of the revolution—and of the French Marxists—primarily among them Sartre, Merleau-Ponty, and Althusser,—Aron proved the impossibility of reconciling historical determinism with human freedom, and contrasted the development of Western economies with the prediction of an unavoidable crisis of Capitalism.

The comparison between liberal and socialist systems also nourished the sociology of industrial societies—*Dix-huit Leçons sur la Société Industrielle* (Aron 1962a), *La Lutte de Classes* (Aron 1964), *Démocratie et Totalitarisme* (Aron 1965a). For Aron, 'an industrial society is one in which big companies constitute the typical form of the organization of labor,' which accompanies the accumulation of capital and the generalization of economic calculus. The common features of the capitalist and communist systems do not mean their convergence, as their political structures remain completely opposed. Pluralism is in conflict with the one-party system, fundamental liberties with a state truth, the autonomy of organized forces in society with their control, the constitutional state with an oversized machinery of repression, market economy with centralized planning. The primacy of political variables excludes any symmetry between the two blocks. Aron is against pluralism or market economy being made into values. They are of means and not ends. Political liberalism separates itself from the utilitarian tradition, of which the most complete version in the twentieth century is to be found in Hayek's *The Constitution of Liberty*. He assigns a most important role to the state, whose task is to establish civil rights within a society and to defend the sovereignty of a country in the field of international relationships.

The study of international relations is indispensable to counterbalance the analysis of industrial society: on the one hand, the rise of violence with the alternation of war and peace, and the fighting of nations and empires; on the other hand, the working of commercial society, based on peaceful competition, and conveying an individualism which attempts to free itself from state tutelage. He was introduced to strategy, while staying in London, through the analysis of the theater of operations in World War II, and he participated in the thinking on the use of nuclear weapons very early on and was a regular commentator on international events. Aron put forward in *Paix et Guerre* (Aron 1962b) a theoretical interpretation of the world diplomatic and strategic system, based on the key role of the states, the only arbitrators in case of armed conflicts. Recognizing the pre-eminence of sovereign states led Aron to be seen outside France as the thinker behind the foreign policy of General de Gaulle, while he was considered in France the most severe critic of de Gaulle's grand design. *Penser la Guerre*, *Clausewitz* (Aron 1976) continues the exploration of conflicting relationships between violence and reason, sovereignty and empires. From the ambivalence of Clausewitz, who was the theoretician of total war and limited conflicts, the rise to extremes and the control of force, Aron shows how the different patterns of international relations in the twentieth century, a European inheritance from the nineteenth century, the interwar period, and the Cold War, combine the passions of peoples and the interest of states, a strategic global view, and an unstable balance between rival powers.

At the same time as his university work, Aron exercised an increasing intellectual and moral authority on French opinion through his columns in *Figaro* and *L'Express* as well as through his papers which threw light on the political crises of the country. From 1957, he pronounced himself in favor of Algerian independence, explaining its inevitable nature. Even a military victory could not prevent a political defeat, as France was fighting in Algeria the ideals which she had claimed for herself from the Revolution, primarily the right of nations to self-determination. In 1968, he analyzed the events of May as a pseudo-revolution in which the extravagance of ideological speeches con-

cealed the lack of a political plan, leading to a nihilism destroying the Republic as well as the University *La Révolution Introuvable* (Aron 1968b). *Les Désillusions du Progrès* (Aron 1969a) develops a meditation on the disenchantment of democratic societies, while the *Plaidoyer Pour l'Europe Décadente* urges rich and vulnerable Europe to find again a major role in politics avoiding the alternatives of integration within the US sphere of influence or subservience to the Soviet empire.

Aron's thought combines a philosophy of history and a moral doctrine in action which rests on the wisdom of statesmen and the commitment of citizens. He rejects the traditional separation between liberalism, which underestimates the weight of history, the strength of violent passions, and the clash of ambitions, and politics, ready to be exonerated from any link with truth and reason. The structure of the multiple aspects of modern societies thus allows one to find out the complex interactions between the organizational transformations, the play of political forces and rival interests, and the ultimate freedom of people to finally '…create their own history, even if they do not know the history they are creating.' Hence, a method at the same time realistic, probabilistic, and dialectical. Realistic, because it refuses any transcendental principle and continuously claims for itself a moral doctrine of responsibility; probabilistic, because it tries to throw light on the complexity of decisions in history by studying the range of possibilities; dialectical because it refuses any determinism and any Manichaeism to take on the complexity and uncertainty which only can allow the resolution of the conflict between the unpredictable side of history with the search for knowledge and the hope of a common vocation for humanity.

This approach breaks with that of most twentieth century French intellectuals who have tried to think about power, whether Alain, Malraux, or Sartre. Alain's ambition was to draw up permanent rules which should govern the relationships between citizens and the governments, from an equilibrium between the liberating principle of universal suffrage and the resistance to alienation. Hence there was a disregard for history and a pacifist commitment which tragically ignored the menace of totalitarianism. Malraux meant to give sense and dignity to man by his revolt against destiny, by a commitment to a cause, and by his participation in an epic, in this case that of General de Gaulle from 1945. Between indifference and adhesion, Sartre, as Aron, claims to represent critical commitment, but this remains abstract, in a state of weightlessness in relation to history. Sartre postulates a radical freedom of conscience, liberated from time and space, composed by the exaltation of violence. Conscience finds itself enslaved and solitary and frees itself through collective revolt, confirmed by terror. This individualistic ontology which attributes an almost mystical function to violence espouses closely the prophetic dimension of Marxism, and is the starting point for Sartre's long association with communism. Aron opposes him by asserting the reinstatement of history in the field of politics and the role of institutions, which open up a possibility for men to act on their future as well as a field for reform of free societies. In the face of these pacifist, epic, or revolutionary visions, Aron defends relentlessly the existence and the rights of a government both fully liberal and fully responsible.

Raymond Aron's liberal political science finds its ultimate horizon in the wager laid on the idea of reason, in the meaning of Kant. At the point of maximum tension between the universal and the particular, it gives its meaning to the commitment of man in history. Aron respected religious faith for which he reserved a space to which he did not enter, the idea of a revelation or of a sacred history remaining fundamentally foreign to him. He considered, however, reason as a 'hidden universal,' which, by freeing man from natural condition and historicity, opens the possibility of a reconciliation between power and liberty. After the death of God and the end of ideologies, in the midst of the fight against the barbarity of genocide and the mass terror of totalitarian regimes, Aron elaborates and sketches the outline of moderate and intelligent policies which mobilize the forces of freedom and man's reasoning to contain the explosion of passions and violence. He reminds statesmen that there is something above politics, Truth. He reminds men of science and of faith, that there is only incomplete knowledge and there is no true commitment if not free. Between the reasoning of power and the power of reason, Aron's moderate realism bases democracy on the fragile equilibrium of tensions which it generates and from which freedom is nourished, on the appreciation of compromise which allows liberty to take roots for a long period. He was a patriot and a cosmopolitan, a relentless adversary of totalitarianism and spokesman of universal history, and he is one of the modern thinkers of freedom and its contradictions in the twentieth century. His views on the power of the citizen are still current, in a world where the disillusionment of free nations is in contrast with the aggressive renewal of feelings of national identity, in a postwar period which must reconcile at the same time the utopia of a world based on the coming of market democracy and the warlike fatalism of a clash of civilizations which would place the twenty-first century in a confrontation of cultures, following those of nationalisms in the nineteenth century and of ideologies in the twentieth century.

3. *Summary*

Raymond Aron, as with other great liberal thinkers, was too respectful of each person's freedom and too

influenced by the principles governing methodological individualism to found a school of thought. But his mark in French political history and thinking was just as profound due to the fact that, far from relying on a network of disciples and institutions, it was anchored in an intellectual heritage which transcends partisan differences or academic studies. Aron succeeded in continuing and revitalizing, in the midst of the century of ideologies, the tradition of French political liberalism, in particular through the creation and development of the reviews *Preuves*, *Contrepoint*, and *Commentaire*. In France, his attitude and his writings—particularly the publication of *L'Opium des Intellectuels* Aron (1955)—contributed decisively in progressively detaching the intellectuals from Communism and, from there, to the resistance of French society to Communism. At the same time, Aron's action and thinking were of tremendous importance in the conversion of a large part of governmental and administrative elite to the Atlantic Alliance and the building of the European Community. Abroad, he acquired a huge moral notoriety and had a privileged relationship with numerous scholars, philosophers, and statesmen, such as Henry Kissinger with whom he maintained a frequent dialogue. Aron thus was a unique example of a twentieth-century French intellectual who was both patriot and European, republican and liberal, antitotalitarian and cosmopolitan. In the face of collective passions and demagogy which overtake regularly democracies in general and French political life in particular, Aron played the major part, according to Claude Lévi-Strauss of a 'teacher of intellectual hygiene.'

See also: Ideology, Sociology of; Liberalism: Historical Aspects; Sociology, History of; Theory: Sociological

Bibliography

Aron R 1935 *La Sociologie Allemande Contemporaine*. Alcan, Paris [new edn. 1981, PUF, Paris]

Aron R 1938 *Introduction to the Philosophy of History. An Essay on the Limits of Historical Objectivity*. Gallimard, Paris. [1961 Weidenfeld and Nicolson, London]

Aron R 1951 *The Century of Total War*. Gallimard, Paris. [1954 Verschoyle, London]

Aron R 1955 *L'Opium des Intellectuels*. Calmann-Lévy, Paris [new edn. 1991, Hachette, Paris]

Aron R 1958 *War and Industrial Society*. Oxford University Press, London

Aron R 1962a *Dix-huit Leçons sur la Société Industrielle*. Gallimard, Paris [new edn. 1988, Gallimard, Paris]

Aron R 1962b *Paix et Guerre Entre les Nations*. Calmann-Lévy, Paris [new edn. 1992, Calmann-Lévy, Paris]

Aron R 1964 *La Lutte de Classes. Nouvelles Leçons sur les Sociétés Industrielles*. Gallimard, Paris [new edn. 1981, Gallimard, Paris]

Aron R 1965a *Démocratie et Totalitarisme*. Gallimard, Paris [new edn. 1992, Gallimard, Paris]

Aron R 1965b *Main Currents of Sociological Thought*. Weidenfeld and Nicolson, London

Aron R 1968a *De Gaulle, Israel and the Jews*. Plon, Paris [1969 Deutsch, London/Praeger, New York]

Aron R 1968b *La Révolution Introuvable, Réflexions sur la Revolution de Mai*. Fayard, Paris

Aron R 1969a *Les Désillusions du Progrès. Essai sur la Dialectique de la Modernité*. Calmann-Lévy, Paris [new edn. 1987, Julliard, Paris]

Aron R 1969b *Marxism and the Existentialists*. Gallimard, Paris [1970 Harper and Row, New York]

Aron R 1973a *History and the Dialectic of Violence: An Analysis of Sartre's 'Critique de la Raison Dialectique'*. Gallimard, Paris [1975 Blackwell, Oxford, UK]

Aron R 1973b *The Imperial Republic. The United States and the World 1945–1973*. Calmann-Lévy, Paris [1973 Prentice Hall, New York]

Aron R 1976 *Penser la Guerre. Clausewitz*. Vol. I, *L'Age Européen*, Vol. II, *L'Age Planétaire*. Gallimard, Paris [new edn. 1989 (Vol. I) and 1984 (Vol. II), Gallimard, Paris]

Aron R 1977 *In Defense of Decadent Europe*. R. Laffont, Paris

Aron R 1978 *Politics and History*. Free Press, New York

Aron R 1983 *Mémoires. Cinquante Ans de Réflexion Politique*. Julliard, Paris [new edn. 1990, Presses Pocket, Paris]

Aron R 1985 *History, Truth, Liberty: Selected Writings of Raymond Aron*. University of Chicago Press, Chicago

Colquhoun R 1986 *Raymond Aron: The Philosopher in History 1905–1955, The Sociologist in Society 1955–1983*, 2 vols. Sage, London

Mahoney D J 1992 *The Liberal Political Science of Raymond Aron*. Rowman and Littlefield, Lanham, UK

N. Baverez

Arousal, Neural Basis of

One of the most important discoveries in the early years of brain research was that by Moruzzi and Magoun (1949). They reported that electrical stimulation of the brainstem of the anesthetized cat produced a pattern of electrical activity recorded from the cerebral hemispheres that was identical to that observed in an awake, alert cat. This pattern was elicited by stimulation of the reticular formation, a diffuse area of neurons and axons, and was characterized by low voltage fast activity, or LVFA, in the electroencephalogram (EEG), which is the record of the electrical activity of the outermost region, or neocortex, of the cerebral hemispheres. LVFA is commonly referred to as neocortical arousal or EEG desynchronization. It replaced the high amplitude slow wave activity pattern that characterized the anesthetized or sleeping cat, and it is considered to be the most prominent index of arousal, alertness or vigilance. With this discovery, the concept of the ascending reticular activating system (ARAS) was established, a system that was believed to maintain the neocortex in an energized, alert state required for the most efficient processing of incoming

information. Soon thereafter the search began for the exact brain substrates in terms of the neurochemistry and brain pathways that contributed to maintaining the neocortex in this energized state, as described below.

1. Forebrain Contributions to Arousal: The Nucleus Basalis of Meynert

Research following the discovery of the ARAS suggested that brainstem neurons within this system contained the neurotransmitter, acetylcholine (ACh), and that they sent projections to widespread neocortical areas where they released ACh to maintain neocortical arousal. Support for this hypothesis was based on several observations. First, greater amounts of ACh were released in the neocortex during LVFA than during high amplitude slow activity. Second, electrical stimulation of the reticular formation elicited ACh release in neocortex. Third, pharmacological agents that blocked the receptors for ACh reduced LVFA in the awake, behaving animal (see Steriade and Biesold 1990 and Dringenberg and Vanderwolf 1998 for reviews of this early research).

As research on the ARAS progressed, several important observations suggested that this early hypothesis needed modification. First, complete transection of the presumed connections from the reticular formation ARAS to the neocortex did not result in a permanent loss of LVFA. Second, ACh containing neurons of the traditionally defined brainstem ARAS were not found to project directly to the neocortex, suggesting the existence of structures located more rostrally in the brain that were capable of sustaining neocortical arousal via the release of ACh. With the development of more sophisticated neuroanatomical techniques, it became clear that major groups of ACh-containing neurons exist in the forebrain and that one group, the nucleus basalis of Meynert (NB), sent widespread projections to the neocortex (Wainer and Mesulam 1990).

More recent research has demonstrated that the NB projection to the neocortex is a major contributor to neocortical arousal. Several lines of evidence support this conclusion (see Steriade and Buzsaki 1990). First, stimulation of the NB elicits LVFA and results in an increase in the extracellular concentration of ACh that depolarizes neocortical sensory neurons, thereby enhancing their response to sensory stimuli. This depolarization is the basis for the observed LVFA and creates a stand-by mode for the most efficient processing of incoming information by neocortical neurons. Second, lesions of the NB that markedly decrease neocortical ACh result in an increase of high-amplitude slow wave activity in the EEG similar to that observed following administration of drugs that block ACh receptors and that are commonly referred to as cholinergic receptor antagonists. Third, the activity of neurons in the NB that project to the neocortex demonstrate significant, positive correlations with the amount of neocortical arousal; that is, the greater their activity, the greater the amount of LVFA. Fourth, stimulation of the NB enhances the response of sensory cortex neurons to sensory stimuli. These accumulated results strongly suggest that NB cholinergic neurons function in neocortical arousal and create a neocortical state for the optimal processing of information; in essence, a physiological correlate of arousal.

2. Brainstem Contributions to Arousal: Cholinergic, Noradrenergic, Serotonergic, and Histaminergic Neuronal Groups

2.1 The Ch-5 Cholinergic Contribution to Arousal

Recall that Moruzzi and Magoun (1949) reported that electrical stimulation of the reticular formation elicited neocortical arousal in the cat. Recent research has demonstrated that ACh containing neurons are located in this region. This group of neurons has been designated as the Ch-5 cholinergic cell group (Wainer and Mesulam 1990). Neurons within the region where the Ch-5 group is located project to the region of the NB, and electrical stimulation of the Ch-5 region activates neurons in the NB that demonstrate positive correlations with neocortical arousal (Detari et al. 1997). This projection, therefore, offers a pathway by which stimulation within the traditionally defined ARAS elicits neocortical arousal via an influence on the NB. Importantly, however, the majority of the neurons that project from the Ch-5 cell group to the NB are not acetylcholine containing (Jones and Cuello 1989). This suggests that the stimulation-induced neocortical activation from the Ch-5 region is not due to activation of Ch-5 cholinergic neurons but perhaps to activation of neurons containing the excitatory neurotransmitter, glutamate.

While the cholinergic neurons of the Ch-5 group do not appear to contribute to neocortical arousal via an influence on the NB, they nevertheless make a significant contribution to arousal via their release of ACh on neurons of the thalamus that receive sensory information directly from the sense organs. For example, Ch-5 neurons project directly onto neurons of the dorsal lateral geniculate nucleus (dLGN) (Wainer and Mesulam 1990). The latter receive information directly from the retina and transmit that information to the visual neocortex for further processing. When active, Ch-5 neurons release ACh on to dLGN neurons which influences these neurons in a manner identical to the influence of ACh on neocortical sensory neurons. It depolarizes them, making them more sensitive to incoming sensory information (Steriade and Busaki 1990, McCormick and Bal 1997).

Thus, Ch-5 ACh neurons, similar to the influence of NB ACh neurons on the neocortex, enhance the processing of incoming information in the sensory thalamus.

2.2 The Serotonergic Contributions to Arousal

While antagonists of ACh receptors have been demonstrated to block neocortical arousal, these antagonists do not block neocortical arousal under all conditions. It has been convincingly demonstrated in the rat that the neocortical arousal that accompanies certain types of behaviors such as walking, stepping, head movements, rearing, postural adjustments or spontaneous limb movements occurs in the presence of cholinergic receptor antagonists. However, neocortical arousal that occurs during grooming, licking, chewing and immobility behaviors is lost (Dringenberg and Vanderwolf 1998). Additional research has demonstrated that antagonists of the neurotransmitter, serotonin (5-HT), block neocortical arousal accompanying the former group of behaviors (Dringenberg and Vanderwolf 1998). The raphe nuclei, a series of serotonin containing cell groups within the midbrain, provide a rich serotonergic projection to the neocortex. Electrical stimulation of the raphe nuclei produces neocortical arousal that is blocked by serotonergic receptor antagonists, whereas selective destruction of serotonergic cells in the raphe nuclei abolishes the neocortical arousal that is resistant to cholinergic receptor antagonists. Additional research has demonstrated that injections of serotonergic antagonists directly into the neocortex can block neocortical arousal produced by noxious stimulation, thereby suggesting that 5-HT, similar to the actions of ACh, works directly at the level of the neocortex to elicit neocortical arousal (Dringenberg and Vanderwolf 1998).

The fact that 5-HT has been reported to depolarize neocortical cells and enhance their excitability is consistent with its ability to produce neocortical arousal (McCormick 1992). It, therefore, appears that both the serotonergic and cholinergic systems contribute to neocortical arousal by direct actions on the neocortex. Indeed, it has been demonstrated that the combined application of cholinergic and serotonergic receptor antagonists completely blocks neocortical arousal accompanying all behaviors, suggesting that other areas which contribute to neocortical arousal may exert their effects via an action on either the serotonergic and/or cholinergic systems (Dringenberg and Vanderwolf 1998). One such area that appears to exert such an indirect effect is the locus coeruleus (LC), as described below.

2.3 The LC Noradrenergic Contribution to Arousal

The LC, more than any other brain structure, possesses the most extensive connections with other brain areas. It comprises a small group of neurons in the dorsal brainstem that contain the neurotransmitter, norepinephrine (NE). The widespread projections of the LC to the neocortex makes it a prime candidate for a function in arousal. Indeed, much research over the years has strongly implicated the LC in neocortical arousal. Perhaps the most compelling evidence for this function derives from the demonstration in rats that infusions into the LC of a small volume of an agent that excites LC neurons elicit a shift from neocortical high amplitude slow wave activity to LVFA (Berridge and Foote 1994). There also appeared a shift in the EEG recorded from the hippocampus to one of intense theta wave activity, an activity pattern that is observed in concert with neocortical arousal. These effects were blocked by intraperitoneal injections of a NE receptor antagonist. Contrariwise, infusions into the LC of a small volume of an agent that inhibits the activity of LC neurons produced a shift in the neocortical EEG from LVFA to high amplitude slow wave activity and abolished hippocampal theta wave activity. Infusion sites within 0.5mm of the LC were without effect. Additional research is consistent with these observations. For example, recordings taken from LC neurons in the monkey demonstrated that their activity correlated positively with neocortical arousal; that is, they showed increased activity during LVFA during waking and decreased activity during high amplitude slow wave activity during drowsiness (Aston-Jones et al. 1996). Finally, the LC is the sole source of NE in the neocortex, and NE applied to neocortical sensory neurons decreases their spontaneous firing rate while enhancing their response to sensory input; in essence, increasing the signal-to-noise-ratio of the neuron. A similar action for NE on thalamic sensory neurons has been observed (McCormick 1992).

The accumulated results suggest that the LC, when active, releases NE in the neocortex which in turn elicits neocortical arousal. Recent research, however, suggests that the LC exerts its effect indirectly since antagonists of cholinergic receptors block the effects of LC electrical stimulation on neocortical arousal (Dringenberg and Vanderwolf 1998). Consistent with these findings is recent research demonstrating that the effects of NE on the neocortical EEG are exerted by an excitatory action on the cholinergic cells of the medial septum which influences neocortical arousal in as yet an unknown manner (Berridge et al. 1996). Nevertheless, the noradrenergic neurons of the LC appear to play an important role in neocortical arousal, although an indirect one.

2.4 Histaminergic Contributions to Arousal

It has long been recognized that antihistamines produce drowsiness accompanied by high amplitude slow wave activity in the EEG. Research has demonstrated that the sole source of histamine-containing neurons in the brain originates in a structure called the

tuberomamillary nucleus (TM) located in the basal hypothalamus and that these neurons send widespread projections to many brain areas including the neocortex (Inagaki et al. 1988). Antihistamines are antagonists of histamine receptors, and their effects on behavioral state and the EEG strongly implicate the histamine-containing neurons of the TM in neocortical arousal. Numerous observations support this hypothesis. For example, selective activation of histaminergic receptors by intracranial injections of histamine elicits neocortical arousal whereas chemically induced inactivation of TM neurons produces high amplitude slow wave activity in the EEG and sleep (Lin et al. 1989, Tasaka et al. 1989). Finally, TM neurons in the cat have been observed to be most active in the aroused, awake, state but demonstrate decreased activity during high amplitude slow wave activity indicative of slow wave sleep (Vanni-Mercier et al. 1984).

Histamine has been shown to depolarize neurons, rendering them more likely to fire in response to incoming information (McCormick 1992). This action is similar to that observed for ACh and 5-HT and could result in neocortical arousal. Some recent evidence, however, suggests that histamine's effect on neocortical arousal, like the effect of NE, may be indirect. For example, neocortical arousal is still present after large depletions of brain histamine, suggesting that histamine is not essential for neocortical arousal (Dringenberg and Vanderwolf 1998). This observation has led to the suggestion that histamine may exert its effects on neocortical arousal via modulation of the cholinergic or seronergic arousal systems (Dringenberg and Vanderwolf 1998).

3. *Modulation of Neocortical Arousal Systems*

The research described above suggests that the cholinergic, serotonergic, noradrenergic, and histaminergic systems contribute either directly or indirectly to neocortical arousal. Given these contributions, an important question arises concerning the environmental stimuli that activate these different systems and that would, in turn, enhance neocortical arousal. Insight into the answer to this question derives from research describing the neuronal responses of these systems to environmental stimuli. Recall from the previous discussion that the spontaneous activity of NB, LC, and TM neurons correlates with the state of neocortical arousal. They demonstrate increased frequency during LVFA during waking, and decreased frequency during high amplitude slow wave activity indicative of drowsiness and sleep. It is important to emphasize, however, that the spontaneous activity of NB and LC neurons is markedly increased by the presentation of stimuli that elicit neocortical arousal and that are relevant to the survival of the organism. For example, novel stimuli that signal a change in the environment, or conditioned, emotionally-arousing stimuli that predict the imminent occurrence of either an important pleasant or unpleasant event, elicit significant increases in activity in NB and LC neurons (Aston-Jones et al. 1996, Richardson and DeLong 1991, Whalen et al. 1994). Presumably, these increases in activity would serve to create a brain state for the most efficient processing of environmental information. This would contribute to the organism's welfare in the face of a changing environment or in the presence of stimuli which are predictive of events of import to the organism's welfare. Research is now underway to determine the brain pathways and structures that serve to convey these stimuli to these arousal systems. For example, neurons located in the amygdala project to the NB and respond to novel as well as conditioned, emotionally arousing stimuli (Kapp et al. 1992).

Furthermore, electrical stimulation of the amygdala has been reported to elicit neocortical arousal and to excite neurons in the NB region that project to neocortex (Dringenberg and Vanderwolf 1998). These observations suggest that stimuli relevant to the organism's welfare may be conveyed to the NB arousal system from the amygdala.

An important focus for future research will be to determine if the amygdala has a similar influence on the brain's other arousal systems. An equally important focus will be to determine the extent to which these systems play redundant and/or uniquely different roles in modulating the level of neocortical arousal. Finally, it is important to note that brain areas that promote sleep actively inhibit many of these arousal systems. An important current research focus is on the delineation of the exact mechanisms by which these arousal systems become inhibited during sleep, a focus described in more detail in the article describing the neural systems contributing to sleep (see *Autonomic Classical and Operant Conditioning*).

See also: Attention-deficit/Hyperactivity Disorder, Neural Basis of; Attention, Neural Basis of; Autonomic Classical and Operant Conditioning; Cardiovascular Conditioning: Neural Substrates; Conscious and Unconscious Processes in Cognition; Consciousness, Cognitive Psychology of; Consciousness, Neural Basis of; Electrical Stimulation of the Brain; Neurotransmitters; Orienting Response; Sleep: Neural Systems

Bibliography

Aston-Jones G, Rajkowski J, Kubiak P, Valentino R J, Shipley M T 1996 Role of locus coeruleus in emotional activation. *Progress in Brain Research* **107**: 379–402

Berridge C W, Bolen S J, Manley M M, Foote S L 1996 Modulation of electro-encephalographic activity in halothane-anesthetized rat via actions of noradrenergic B-receptors within the medial septal region. *Journal of Neuroscience* **16**: 7010–20

Berridge C W, Foote S L 1994 Locus coeruleus-induced modulation of forebrain electroencephalographic (EEG) state in halothane-anesthetized rat. *Brain Research Bulletin* **35**: 597–605

Detari L, Semba K, Rasmussen D D 1997 Responses of cortical EEG-related basal forebrain neurons to brainstem and sensory stimulation in urethane-anaesthetized rats. *European Journal of Neuroscience* **9**: 1153–61

Dringenberg H C, Vanderwolf C H 1998 Involvement of direct and indirect pathways in electrocorticographic activation. *Neuroscience and Biobehavioral Review* **22**: 243–57

Jones B E, Cuello A C 1989 Afferents to the basal forebrain cholinergic cell area from pontomesencephalic-catecholamine, serotonin, and acetylcholine-neurons. *Neuroscience* **31**: 37–61

Kapp B S, Whalen P J, Supple W F, Pascoe J P 1992 Amygdaloid contributions to conditioned arousal and sensory information processing. In: Aggleton J P (ed.) *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*. Whiley-Liss, New York

Lin J S, Sakai K, Vanni-Mercier G, Jouvet M 1989 A critical role of the posterior hypothalamus in the mechanisms of wakefulness determined by microinjection of muscimol in freely moving cats. *Brain Research* **13**: 225–40

McCormick D A 1992 Neurotransmitter actions in the thalamus and cerebral cortex and their role in neuromodulation of thalamocortical activity. *Progress in Neurobiology* **39**: 337–88

McCormick D A, Bal T 1997 Sleep and arousal: Thalamocortical mechanisms. *Annual Review of Neuroscience* **20**: 185–215

Moruzzi G, Magoun H W 1949 Brain stem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology* **1**: 455–73

Richardson R T, DeLong M R 1991 Electrophysiological studies of the functions of the nucleus basalis in primates. In: Napier T C, Kalivas P W, Hanin I (eds.) *The Basal Forebrain: Anatomy to Function*. Plenum Press, New York

Steriade M, Biesold D (eds.) 1990 *Brain Cholinergic Systems*. Oxford University Press, New York

Steriade M, Buzsaki G 1990 Parallel activation of thalamic and basal forebrain cholinergic systems. In: Steriade M, Biesold D (eds.) *Brain Cholinergic Systems*. Oxford University Press, New York

Tasaka K, Chung Y H, Sawada K, Mio M 1989 Excetatory effect of histamine on the arousal system and its inhibition by H1 blockers. *Brain Research Bulletin* **22**: 271–5

Vanni-Mercier G, Sakai K, Jouvet M 1984 Waking-state specific neurons in the caudal hypothalamus of the cat. *C R Academy of Science III* **298**: 195–200

Wainer B H, Mesulam M-M 1990 Ascending cholinergic pathways in the rat brain. In: Steriade M, Biesold D (eds.) *Brain Cholinergic Systems*. Oxford University Press, New York

Whalen P J, Kapp B S, Pascoe J P 1994 Neuronal activity within the nucleus basalis and conditioned neocortical electroencephalographic activation. *Journal of Neuroscience* **14**: 1623–33

B. S. Kapp and M. E. Cain

Art and Culture, Economics of

1. Introduction

Though Adam Smith, William Jevons, Alfred Marshall, David Ricardo, and John Maynard Keynes (who also collected paintings) were intrigued and had some views on the issues which will be discussed, research on the economics of art and culture is said really to have started after the essay by Baumol and Bowen (1966) on why, unless they receive financial support, the performing arts may disappear. The field is now relatively well established, has an association, the Association of Cultural Economics International, which publishes a quarterly journal, the *Journal of Cultural Economics*, organizes, every two years, a conference that is attended by 200 to 300 scholars, and has been able to generate teaching positions in a few economics departments in Europe and in the United States. The field is, nevertheless, still in its infancy

The topic is not well-defined, since it is located at the crossroads of several disciplines: art history, art philosophy, sociology, law, management, and economics, and tries (or should try) to tackle questions such as why Van Gogh's paintings are expensive, and why copies of his works are cheap; why pre-Raphaelite painters came back into vogue in the 1960s, after having been completely forgotten during almost a century (with an obvious effect on their prices); why European public or national museums are not allowed to sell parts of their collections; how the performance of museums should be evaluated; which (and given the budget constraint, how many) buildings should be saved from demolition, and kept for future generations; why the arts should be supported by the state; why there are superstars who make so much money; and whether works that have been sold should nevertheless be subject to copyright laws. From this enumeration, it should appear that all the fields are interacting, but that this does not often make an art historian understand mathematical economics, nor an economist show interest in pre-Raphaelite painters.

The list also raises the issue of whether economists interested in cultural economics should simply apply their usual tools to questions related to and data coming from the arts, or whether they should take culture as an opportunity to add new issues to the existing economic literature. Should economists take as granted that prices and consumers' incomes are the main determinants of the demand for theater plays, or should they also understand what quality means in such heterogeneous markets, define it, measure and enter it as a variable in their regressions?

To describe the main problems which arise in cultural economics, it is useful to make the distinction between the performing arts (music, theatre, opera, dance), the visual arts (paintings, sculpture, art objects), and cultural heritage (museums, historical buildings, monuments and sites), though there are some unavoidable intersections (museums accumulate paintings, some artists in the visual arts produce unique performances in galleries). There is a tendency, nowadays, to add what came to be called the cultural industries, which cover books, movies, popular music, records, and the media (radio, television, newspapers

and, of course, the ever present and invasive Internet). The discussion will be devoted mainly to the first three groups, which are often referred to as 'high culture,' and will try to point out what cultural economists have found interesting, discuss the tools which they have imported from economics to analyze culture, which new insights the analysis of art markets has added to economics, and what the questions are that have been left open or untouched. However, to start with: a topic that applies to all groups, has kept many cultural economists busy, and has led to a large literature in the field. Should the state support the arts and, if so, how?

2. *Public Support of the Arts. Why?*

Many arguments have been invoked to justify public support, and to correct for assumed market failures. Some apply to all forms of artistic activities (including the cultural industries), and some are more specific. Here is a nonexhaustive list of the most important economic rationalizations for such support.

(a) The oldest and most often invoked argument is that art, whatever its form, is a public good. It benefits not only those who attend or see it, and who pay for it, but also all other consumers, who do not necessarily wish to contribute voluntarily to its production (performing or visual arts) or to its preservation (museums) and free ride, or who cannot contribute, since they are not yet born (heritage). If the arts are left to the market they will not be priced correctly, and will thus be underproduced or not saved for future generations. Artistic activities are also said to produce externalities that cannot be sold on the marketplace, such as civilizing effects, national pride, prestige, and identity (justifying, for example, the French position in their fight against the free trade of movies and TV programs), social cohesion, etc. (such arguments were suggested earlier by Stanley Jeavons, Adam Smith, and Arthur Pigou) which benefit all consumers.

(b) But the arts are also said to yield economic externalities. Old castles, well-known opera houses or orchestras, and art festivals attract visitors and tourists. So do museums with good collections, while newly constructed museums are claimed to contribute to city renewal (an argument used, for instance, to attract public support for the recent Guggenheim Museum in Bilbao). This is supposed to have spillover effects on hotels, nearby restaurants and shops, generate new activities, and this effect came to be called the 'arts multiplier.' Grampp (1989, p. 247), ironically, gripes that the arts multiplier takes 'its place besides the investment multiplier, the foreign trade multiplier, and the balanced budget multiplier in the kitchen midden of Keynesianism.' (See also the paper by Seaman 1987.)

(c) Art is a 'merit' good. It 'is a means of educating the public's taste and the public would benefit from a more educated taste' (Scitovsky 1972). Since consumers are not fully informed, they are unable to evaluate all its benefits without public intervention. Moreover, even if there is addiction to the arts, it develops only slowly, and consumers have to be exposed as much as possible.

(d) For equity reasons, art should be made available also to low-income consumers who cannot afford to pay. Poor artists should also be supported. Schemes for doing this will be discussed below.

(e) Culture is transmitted by education but also from parents to children. Since parents can hardly be considered as purely altruistic, an additional externality is generated, which needs support for efficiency (and equity) reasons.

(f) The last argument on the difficulty or impossibility of achieving productivity gains is more specific to the performing arts. It was put forward, more than 30 years ago, by Baumol and Bowen (1966), and came to be known as the Baumol cost disease. It can be stated briefly as follows. Since wages escalate in sectors other than culture, they must also do so in the performing arts to make these attractive enough for artists to enter, but since no productivity gains are possible, wage increases have to be passed fully to prices. Therefore, the relative price of the performing arts increases and, unless subsidized—or supported by donors and private funds—the sector will shrink and eventually disappear. (It is worth quoting the (now) standard argument by Baumol and Bowen: 'The output per man-hour of the violinist playing a Schubert quartet ... is relatively fixed, and it is fairly difficult to reduce the number of actors necessary for a performance of Henry IV, Part II.')

The issues just discussed have, of course, their supporters and their detractors. The main detractors base their case on the public choice theoretic claim that some people wish to use the state's resources for their own benefit, and on the regressive nature of supporting high-income groups who are more likely than others to attend, and who can afford to do so. In 1982, in the United States, for example, 38.5 percent of the high-income population (earning more than $50,000 per year) attend classical music live performances, while this percentage drops to 8.1 percent for low income groups (earning less than $10,000 per year) (see Heilbrun and Grey 1993, p. 43).

Several additional arguments can be set against the Baumol and Bowen reasoning. Empirical studies on the performing arts point to price elasticities which are smaller than one, so that there is still some room for price increases; the performing arts have not exhausted the possibilities of price discrimination (popular shows do not seem to charge more, most venues could be scaled in more sections than observed, intertemporal price discrimination and peak load pricing during certain days are seldom implemented, bundling—

bookstores at opera houses, museums, drinks during intermissions—can also be used more systematically); technical progress (sound and lighting, for instance) can make for larger audiences; broadcasts and records—though these forms may also be subject to the cost disease—could be used to cross-subsidize live performances, and make these less expensive; synthetic music is being composed, which needs no, or fewer, performers, and is even sometimes used in ballet or opera performances; opera music is performed in the form of concerts; contemporary theatre plays save on the number of characters, etc. In the visual arts, Marcel Duchamp and later, Andy Warhol, created ready mades and multiples by the hundreds. Reproductions of paintings in books pay royalties to artists, which is also a form of cross-subsidization.

The cost disease has led to a host of quantitative studies, which show that the curse is not as severe as Baumol and Bowen originally suggested. But there is still more empirical work needed to prove (or disprove) the reality of the other arguments which call for public support.

3. *Public Support of the Arts. How?*

In most developed countries, the arts are supported, though the volume of public intervention can be very different. It is very low in the United States ($3 to $4 per head), reaches $16 in the United Kingdom, $19 in Australia, and $30 to $40 in continental Europe, where private support is, at the present time, negligible (see Throsby 1994, p. 21 for more detailed data). No convincing argument has yet been put forward to explain these large differences in behavior between the Anglo-Saxon and other cultures.

Government intervention takes several forms. It can be direct: some art institutions or companies are state-owned (museums in most European countries, but also opera companies, theaters, symphony orchestras) and are, therefore, supported, at least in part, by public funds. Their efficiency has been questioned on many occasions. Arts education programs are also often supported by the state. In the 1960s, Peacock (1993, p. 122) suggested the distribution to poor consumers of art vouchers, exchangeable for seats or visits to museums.

In most cases, public funds, including lottery funds, are channeled to private companies, or directly to artists, through subsidies, or tax relief (reduced VAT rates, no tax or lower rates of taxation on profits or on income). In the United States, most of the effort is concentrated on tax deductions or exemptions to donors (individuals or corporations) for charitable contributions; gifts or bequests of artworks are also encouraged by tax concessions, and more and more often, inheritance taxes can be paid for by artworks. This indirect subsidization system is used much less in European countries, though the incentive schemes are also present.

The less direct (and, sometimes, most insidious) form of intervention is regulation, which is mainly concerned with (a) guaranteeing artistic freedom, though this right has often been breached—for example, to 'protect' consumers against pornography; (b) ensuring moral rights—that is, protecting artists against being plagiarized, or against alterations or destruction of their works; and (c) ensuring financial rewards to living artists and to their heirs during as much as 70 years after their death, in the European Union, in the form of copyright. Copyright has led to a very large literature on the confines of law and economics. Its economic justification is based on the idea that it provides incentives to create, and that the absence of protection would lead to underproduction—the same idea is used to justify protection of industrial patents.

Copyright applies mainly to plays, and literary or musical works. Authors usually do not 'sell' their creation, but get compensated through authors' rights for each copy (book or record) sold, or for each performance of a play or of a score. However, the idea has been extended to all other forms of art, including the performance itself (the staging of a play can be copyrighted), and the visual arts. Reproduction of a work pays rights to the artist even after it has been sold; resale rights (*droit de suite*) pay to the artist a percentage on the resale of a work on the secondary market, for instance between two dealers (note that since such transactions are difficult to follow, *droit de suite* is usually only enforced for sales at auctions). The photographer of a work by Rembrandt is, in some (European) countries, entitled to rights for his 'original' photography. One can wonder whether copyright laws still provide incentives for art creation, or whether they generate incentives to use Mozart's music in movies, and Schiele or Klimt paintings (no longer covered by copyright) to illustrate book covers, in order to avoid the intricacies of the law on contemporary scores and illustrations. Obviously, regulation has led to overregulation, perhaps driven by the rent-seeking behavior of author societies, which collect transaction fees, and not by artists themselves.

The international trade of cultural goods has also been the concern of regulators. Some countries prohibit completely the export of ancient artifacts—although illegal exports from Mexico, Italy, or Greece continue. Some countries (France and the United Kingdom) are more lenient and require export licenses for 'national treasures;' this procedure is often used to delay the export and give time to the potential exporting country to raise funds and buy the object at the price at which it could be sold abroad. More recently, during the 1994 GATT negotiations, France insisted on maintaining import quotas on American TV programs to 'protect' local culture and national identity. This is the so-called cultural exception argu-

ment, which is very reminiscent of its forefather, the infant industry argument.

4. *Rich Man, Poor Man, Beggar…*

A folk theorem states that most artists are poor—or have lower income than the average consumer—and die poor, while salerooms, their dealers, and editors become rich.

Using the 1980 US Census data, Filer (1986) concludes that, on average, artists do not earn any less than they would in a nonartistic employment (after standardizing for personal characteristics, productive attributes, and age). It has later been shown, however, that this is due to multiple job holdings, including in nonartistic activities (see, e.g., Wassal and Alper 1992). If these extra jobs are not accounted for, artists' average incomes are smaller than those of workers with comparable education. It is also said that artists devote less time to their work on art than do others to their main occupation, that they may have to go to lower paid jobs to supplement their income, and that their earnings are more variable and less directly influenced by education than those of other workers. To explain then why artists accept this income differential, one has to appeal to differences in motivation; this is summarized by Throsby (1994, p. 17) who suggests that 'the primary desire to create art as a principal occupation must be recognized as the essential driving force behind an artist's labor supply decisions.'

But the income distribution of artists is also markedly skewed, and there exist superstars whose earnings are impressive. In a beautiful paper, Rosen (1981) shows that talent alone does not explain these differences. He models income as a convex function of talent, so that small differences in this endowment may result in large differences in income. This is compounded by the fact that there is little if any substitution between various degrees of talent, or between quality and quantity. As Rosen points out, 'hearing a succession of mediocre singers does not add up to a single outstanding performance.'

5. *Issues in the Market for Visual Arts*

Art works are, with some exceptions, heterogeneous. Each creation is unique and markets can hardly be analyzed by the usual supply–demand mechanism. The very high prices reached by some works has led economists to study (a) what can be said about the characteristics that determine values, (b) whether the returns obtained on artworks outperform those of financial markets, and (c) why tastes change. None of these issues has been given a definitive or even a satisfactory answer.

Values are obviously determined mainly by the name (or the reputation) of the artist, but this explains little, and one should understand why reputations differ, and what makes a painting 'good' or 'bad.' Decomposing 'quality' in terms of (hopefully) objective characteristics should be possible. After all, this is what, long before Kevin Lancaster, the French art historian de Piles (1708) had in mind when he wrote that (the English translation is borrowed from Rosenberg (1969, p. 33) 'the true understanding of painting consists in knowing whether a picture is good or bad, in distinguishing between that which is good in a certain work and that which is bad, and in giving reasons for one's judgment.' This led him to grade, on a scale of 20, characteristics such as composition, drawing, color, and expression, for a group of painters, and rank them accordingly (note that quality concerns also the performing arts, movies, TV programs, etc.). Art historians and philosophers have strong objections to such views. Economists are used to working on horizontal and vertical differentiation, but consider quality in art to be the playground of art historians and philosophers. Some collaboration between the two sciences should therefore be possible!

Financial returns on works of art have been analyzed very often (see Frey and Eichenberger 1995) for a survey). Baumol's (1986) paper has probably been one of the most influential studies, since the conclusions are based on prices obtained at auction over 300 years (1650–1960). Baumol finds that the average real return rate is equal to 0.55 percent per year, some 2 percent lower than the return on bonds. The difference is understandable and, according to Baumol, attributable to the return provided by aesthetic pleasure. Buelens and Ginsburgh (1993) qualify the results and show that this conclusion does not preclude the existence of 20 to 40 years-long time intervals during which returns are much higher; since tastes change slowly, there may thus be opportunities to do better than bonds. In a recent paper, Landes (2000) shows that skillful collectors can even beat the stock market, as long as they spend time and effort and follow a buy and hold policy. Buelens and Ginsburgh show that when the 1914–49 period of turmoil is excluded, the rate of return is 2.5 percent, thus equal to the return on bonds. Art is thus attractive even if aesthetic returns are ignored.

Tastes change, and prices commanded by works of art often follow these movements. There are many examples of painters who have been 'rediscovered,' including Botticelli, Vermeer, Goya, or the pre-Raphaelites (see Haskell 1980 for an account of these changes in France and in the United Kingdom between the late eighteenth century and the early 1920s). It is unclear whether such changes are due to fads or fashion, or to more fundamental reasons. Grampp (1989, pp. 66–8) suggests that the rapid succession of styles in contemporary art is due to the rapid increase in consumers' income, 'who seek novelty more often

than they would if their income was lower.' The changing tastes issue has, to our knowledge, never been investigated seriously by economists. (At least empirically: Stigler and Becker 1977 have developed a theory of changing tastes in the neoclassical framework of unchanging preferences—see below.)

The status and the prices of copies (to be distinguished from fakes) have also varied over time. While copies were close substitutes for originals during the seventeenth century, and priced accordingly at about half the price of originals (see De Marchi and Van Miegroet 1996) they are considered to have hardly any value nowadays. Again, little is known about the reasons and the period at which the break occurred. Photography is sometimes said to have been at the root of the change.

The issues discussed above have often been analyzed on the basis of works sold and prices obtained at auctions. This has led to several observations on the possible inefficiency of auction mechanisms, which theory considers to be efficient—this is at least the case for the second-price auction mechanism often used in art auctions. Among these observations, let us single out the declining price anomaly (prices of homogeneous lots sold in sequence decrease over time, but this is also the case for the ratio between hammer prices and pre-sale estimates for heterogeneous works of art (Ashenfelter 1989, Beggs and Graddy 1997)), the existence of systematic price differences across salerooms and/or countries (Pesando and Shum 1996)—though this seems to make arbitrage possible, saleroom commissions and other transaction costs are usually high enough to reduce such possibilities, at least in the short run—the biasedness of pre-sale estimates by salerooms, and the fact that pre-sale estimates do not reflect all the information that is available (though theory shows that it is efficient for the auctioneer to reveal the truth).

6. *Cultural Heritage, Where Do We Go?*

Forty-five thousand monuments are listed in France, and this number is increasing every year, without any legal possibility of dropping a monument once it has been listed. In the United Kingdom, there are 40,000 scheduled monuments and sites, and 500,000 listed buildings. Maintaining this heritage obviously implies costs, and these costs are increasing. The number of specialists who are able to reconstruct and restore these sites is dwindling, the materials needed are becoming rare and often expensive to extract, and monuments deteriorate at an increasing rate, since they are visited more often, and are subject to damage by pollution (Benhamou 2000, p. 58). A curse similar to the one said to prevail in the performing arts (the previously mentioned cost disease) is ahead of us. The decision to list heritage sites less will have to be made, and perhaps also the decision to drop monuments and allow their demolition. Property rights may have to be redefined to induce private owners to care for their properties, and tools, such as contingent valuation, polls, or even referenda may be needed to elicit preferences and help decision makers.

The number of museums has also increased dramatically since the 1980s, and so have collections, generating serious constraints on exhibition space (the forms used by contemporary art installations, for example, make this constraint even more binding). It happens that as much as 80 to 90 percent of the works are not on show (97 percent in the case of the Art Institute in Chicago). Though the International Council of Museums (ICOM) recommends that each musuem should have a statement describing its collecting policy, many have no such written documentation. To the layman, and sometimes even to the expert, some transactions look bizarre, to say the least. In 1999, though it possesses several works by Breughel the Elder, the largest Belgian museum bought a copy by one of his descendents, Breughel the Younger, for the stunning amount of 3 million dollars. This amount should be compared to the museum's $150,000 annual public subsidy for new acquisitions! Why can't museums stop increasing in numbers, and dimensions. Why is it that if curators have the power to buy works of art, it is made difficult (in the United States) if not impossible or unlawful (in Europe) for them to sell works ('deaccession,' in museum language), and certainly those which are not and will probably never be shown. This would not only provide financial leverage to museums, but also reduce the incentive for illegal excavations and sales.

Should museums charge entrance fees? Curiously enough, the French, whose successive governments claim to have the largest budgets to support the arts, insist that museums should charge visitors—and they do so—while free admission at the six British national museums—some of which did make charges—was introduced in April 2001. Some (natural) experiments are being run in French museums on the effect of charging, and not charging entrance fees on Sundays once every four weeks, but the evidence that this increases the number of visitors is scant. More serious econometric work is obviously needed in this area.

How would the management of museums be affected if the value of their collections were considered as capital? How should museums (as well as many other cultural institutions) be evaluated? Their output can obviously not be measured by profits, nor by the sole number of visitors, as has often been suggested, and done. According to the ICOM definition, a museum 'acquires, conserves, researches and communicates, and exhibits for the purpose of study, education and enjoyment, material evidence of people and their environment.' This makes for an activity with multiple outputs; measuring its performance calls for the use of appropriate efficiency frontier methods.

7. *The Cultural Industries: How Will Newspapers and TV Programs Look in the Future?*

Here we turn to more conventional commodities, since the cultural industries mainly produce 'copies,' such as books (some books can be viewed as works of art; this is obviously the case with old manuscripts, but also with some contemporary books, often illustrated by painters, printed in limited numbers, and which become collectors' items), records, movies, newspapers, and TV programs, and the value of the work is not conferred by its uniqueness or its scarcity. Note that this is also partly true for the performing arts, though each performance can be considered as unique. As in other sectors, these industries are faced with globalization and follow the move to more concentration, both in terms of production and of distribution. This may open the road to more uniformity and is sometimes considered as presenting a threat for democracy, especially in the case of newspapers and TV or radio news programs.

At issue are also (a) the public support for (and even ownership by the state of) television and radio stations and, in some countries (e.g., France), newspapers, and the complex problems of (dis)information this may generate, (b) the import quotas on cultural programs in a world which moves to free trade in all other industries, and (c) the complex economic and legal copyright problems generated by the Internet.

8. *Other Issues and Why Not Culture for the Enjoyment of Life?*

We are left with many unresolved questions. The economics of certification or authentification ('authentic' works vs. copies or fakes) is one of these. While many art philosophers maintain that this does not change the aesthetic qualities of a work, it definitely changes its price! The formation of values also raises the interesting issue of who are the agents endowed with the power to discriminate between good and bad art (the 'gurus'). What is the expertise of those who grant prizes and awards? Who still reads Sully Prudhomme (Nobel literature winner in 1901), Paul von Heyse (Nobel 1910), Verner von Heidenstam (Nobel 1916), or of more recent vintage, Ivo Andric (Nobel 1961). But even worse, who still remembers their names? Who would today award the 1959 Oscar for best movie to *Ben Hur* and forget about *Some Like it Hot*, produced during the same year? Time obviously does a far better job than the experts.

Why did blockbuster exhibitions or music, opera, and theatre festivals become so numerous? Don't such events simply displace the public from more traditional venues? Or do they, as is sometimes claimed, attract visitors who also have a look at the permanent collections, and familiarize with music those tourists who would otherwise never attend a concert in their own city. Much has been said on the rent-seeking behavior of those who organize such events, but little is known about the long-term effects on the behavior of those who attend.

The last question, which is also one of the most important ones, is how culture affects the preferences of consumers. This is the subject of the important paper by Stigler and Becker (1977), who were among the first to discuss changing tastes and addiction to culture (as well as to other goods or bads) in terms of a household production function and the stock of past consumption (of culture) and not consumption itself, appearing as argument in the utility function.

Addiction (if positive) should enhance the role of education to culture. More education to culture would certainly lead to more participation in cultural activities (see Bergonzi and Smith 1996) and, therefore, to less need for public support. It would also increase efficiency, since we would have to accept that consumers possess the knowledge and have the right to choose in a sovereign way, and therefore also accept the consequences to which this could lead.

9. *Bibliography Notes*

For obvious reasons of space, it is impossible to do justice to and mention the many original contributions to cultural economics. The two volumes edited by Towse (1997) contain many of the important papers that have made cultural economics into a field. The book by Peacock (1993) is a delicious and highly informative account by an economist *cum* musician and composer, of the years he spent as Chairman of the Scottish Arts Council, and member of the Arts Council of Great Britain. It also contains references to his many and highly regarded other contributions to cultural economics. The iconoclastic views in Grampp (1989) make for very entertaining and witty reading that is also very serious economics. So is Frey's and Pommerehne's (1989) contribution, which contains a very complete and useful list of references written before 1989. O'Hagan (1998) covers most of what should be known on the relations between the art sector and the state. So does Schuster (1997) for the cultural heritage, Throsby and Withers (1979) for the performing arts, and Caves (2000) and Hoskins et al. (1997) for the main problems faced by the cultural industries. Throsby (2001) bridges the gap between economics and culture. The situation in the United States is described by Netzer (1978), Heilbrun and Gray (1997), and by Feldstein (1991) for museums. A very good overview of the French situation (which is almost polar to the American one) is given in Benhamou (2000). The two volumes by Merryman and Elsen (1998) are invaluable as a first approach to the understanding of the intricacies of law and the visual arts.

See also: Art: Anthropological Aspects; Art, Sociology of; Arts Funding; Cultural Resource Management (CRM): Conservation of Cultural Heritage; Film and Video Industry; Fine Arts; Markets: Artistic and Cultural; Television: History; Television: Industry

Bibliography

Ashenfelter O 1989 How auctions work for wine and art. *Journal of Economic Perspectives* **3**: 23–36

Baumol W J 1986 Unnatural value: Or art investment as floating crap game. *American Economic Review* **76**: 10–14

Baumol W J, Bowen W G 1966 *Performing Arts. The Economic Dilemma*. Twentieth Century Fund, New York

Beggs A, Graddy K 1997 Declining values and the afternoon effect: Evidence from art auctions. *Rand Journal of Economics* **28**: 544–65

Benhamou F 2000 *L'économie de la culture*. La Découverte, Paris

Bergonzi L, Smith J 1996 *Effects of Arts Education on the Participation in the Arts*. Seven Locks Press, Santa Ana, CA

Buelens N, Ginsburgh V 1993 Revisiting Baumol's 'Art investment as floating crap game.' *European Economic Review* **37**: 1351–71

Caves R E 2000 *Creative Industries*. Harvard University Press, Cambridge, MA

De Marchi N, Van Miegroet H J 1996 Pricing invention: Originals, copies and their relative value in seventeenth century art markets. In: Ginsburgh V, Menger P-M (eds.) *Economics of the Arts. Selected Essays*. North-Holland, Amsterdam, The Netherlands

de Piles R 1708/1992 *Cours de peinture par principes*. Gallimard, Paris

Feldstein M (ed.) 1991 *The Economics of Art Museums*. The University of Chicago Press, Chicago

Filer R K 1986 The starving artist—Myth or reality? Earning of artists in the United States. *Journal of Political Economy* **94**: 56–75

Frey B S, Eichenberger R 1995 On the rate of return in the art market: Survey and evaluation. *European Economic Review* **39**: 528–37

Frey B S, Pommerehne W W 1989 *Muses and Markets: Explorations in the Economics of the Arts*. Blackwell, Oxford, UK

Grampp W D 1989 *Pricing the Priceless. Art, Artists and Economics*. Basic Books, New York

Haskell F 1980 *Rediscoveries in Art. Some Aspects of Taste, Fashion and Collecting in England and in France*. Phaidon Press, London

Heilbrun J, Gray C M 1993 *The Economics of Art and Culture. An American Perspective*. Cambridge University Press, Cambridge, UK

Hoskins C, McFadyen S, Finn A 1997 *Global Television and Film*. Clarendon Press, Oxford, UK

Landes W M 2000 Winning the art lottery: The economic returns to the Ganz collection. *Louvain Economic Review* **66**: 111–30

Merryman J H, Elsen A E 1998 *Law, Ethics and the Visual Arts*, 3rd edn. Kluwer Law International, London

Netzer D 1978 *The Subsidized Muse: Public Support for the Arts in the United States*. Cambridge University Press, Cambridge, UK

O'Hagan J W 1998 *The State and the Arts. An Analysis of Key Economic Policy Issues in Europe and the United States*. Edward Elgar, Cheltenham, UK

Peacock A 1993 *Paying the Piper. Culture, Music and Money*. Edinburgh University Press, UK

Pesando J E, Shum P S 1996 Price anomalies at auction: Evidence from the market for modern prints. In: Ginsburgh V, Menger P-M (eds.) *Economics of the Arts. Selected Essays*. Elsevier, Amsterdam, The Netherlands

Rosen S 1981 The economics of superstars. *American Economic Review* **71**: 845–58

Rosenberg J 1969 *On Quality in Art*. Phaidon, London

Schuster J M (ed.) 1997 *Preserving the Built Heritage. Tools for Implementation*. University Press of New England, Hanover, NE

Scitovsky T 1972 Arts in Affluent Society—What's wrong with the arts is what's wrong with society. *American Economic Review* **62**: 62–9

Seaman B A 1987 Art impact studies: A fashionable excess. In: Radich A J (ed.) *Economic Impact of the Arts: A Sourcebook*. National Conference of State Legislatures, Washington, DC

Stigler G J, Becker G S 1977 De gustibus non est disputandum. *American Economic Review* **67**: 76–90

Throsby D 1994 The production and consumption of the arts: A view of cultural economics. *Journal of Economic Literature* **XXXII**: 1–29

Throsby D 2001 *Economics and Culture*. Cambridge University Press, Cambridge, UK

Throsby D, Withers G A 1979 *The Economics of the Performing Arts*. St. Martin's Press, New York

Towse R (ed.) 1997 *Cultural Economics: The Arts, the Heritage and the Media Industries*. Edward Elgar Cheltenham, UK, 2 Vols

Wassal G H, Alper N O 1992 Toward a unified theory of the determinants of the earnings of artists. In: Towse R, Kakhee A (ed.) *Cultural Economics*. Springer Verlag, Berlin

V. A. Ginsburgh

Art: Anthropological Aspects

The term 'art' has several meanings, all having to do with the performance and products of human skills, associated with aesthetic (rather than functional, sacred, or scientific) qualities appealing to the human senses. Art can appeal to a single sense (e.g., easel painting), or to multiple senses (e.g., ritual dances—vision, kinesthesia, sound). This essay focuses on the visual arts rather than on song, music, or dance.

1. The Term 'Art'

The meaning of 'art,' from the Latin *ars*, *artis* (skill), has changed over time, reflecting technological and social contexts. In English, 'art' is often twinned with 'craft' from the Germanic *kraft* ('skill, cunning' in English, since 850 AD). After the Norman Conquest and the Renaissance, 'craft' became associated with manual skills and lower-class occupations, and 'art,'

from the French *art*, with the conquerors and the Church, and hence with learned practices such as mathematics, rhetoric, logic, and grammar (as in Master of Arts).

In Enlightenment Europe, the meaning of 'art' changed again, becoming a synonym for '*beaux arts*,' a set of primarily expressive skills entertaining to the senses but with few practical aims. These arts were the enjoyment of the wealthy, educated classes, whose tastes brought into being nonecclesiastical forms such as opera, symphonic music, figurative painting, and choreographed dance (and later, ballet). Once detached from 'bread-and-butter' issues, art was justified as a superior expression of the human soul, whose practitioners embodied inner genius and were owed freedom of expression.

Many aesthetic practices of the unschooled members of European populations: folk song, dance, and handicrafts; and of women: quilting, lullabies, and sewing, were not considered to be arts. Others, having both aesthetic and practical aspects: architecture (vs. building), landscape architecture (vs. gardening), or cuisine (vs. cooking), bestowed 'artist' status only on trained practitioners, usually men. These distinctions are fading with the postmodernist revaluations of many folk and vernacular practices as 'arts.'

2. *The Challenge for Anthropologists*

All these arts are embedded in the historical, social and cultural structures of the Western world, and anthropologists find it difficult to identify 'art' in other societies. They have tackled the problem in two ways.

First, to identify practices and products resembling Western arts, in form (appeal to certain senses), and in relation to context, e.g., religion, hierarchy, ritual. This was easy in the West's encounter with other complex civilizations (often archaeological), but difficult in more egalitarian, nonliterate societies studied by early anthropology. In his *Primitive Art* (1955) Franz Boas emphasized the universality of arts as skill-based productions of aesthetically valued forms:

> When the technical treatment has attained a certain standard of excellence … such that certain typical forms [later identified as 'styles'] are produced, we call that process art … perfection is essentially an aesthetic judgement … [but] we cannot tell just where the aesthetic attitude sets in (1955, p. 10).

He added, 'the twofold source of artistic effect, the one based on form alone, the other on ideas associated with form' (1955, p. 13). Presumably, the former would be aesthetic universals and the latter culturally specific, but Boas admits that other authors claim that, iconically, all arts represent something.

For Maquet (1979) all societies have aesthetic foci, where excellence is judged by the senses. But 'Art' comprises only those productions which circulate within the metropolitan 'art world' (Becker 1982), and 'primitive art' only exists in the galleries and homes of Western societies. Arts are created two ways (a) *by intention*, made by self-conscious 'artists' for the art world; and (b) *by metamorphosis*, produced either in the past or elsewhere, outside the art world. For instance, eighteenth-century quilts were not considered art until late in the twentieth century, and contemporary Dogon granary doors made for use in West Africa, only become 'art' when imported to the West. The current globalization of commodities makes Maquet's suggestion cogent for almost everything that would be 'art' by Boas' definition.

3. *Meaning, the Content of Art*

Twentieth-century anthropological approaches to art have been concerned with the mechanisms and nature of the messages carried by art. They have drawn upon either psychology or linguistics, i.e., semiotics, reflecting functionalist and structuralist modes of anthropology.

3.1 *Psychological Approaches*

Psychoanalytic approaches treated art, like dreams and myths, as expressions of the subconscious, translated into forms acceptable to the conscious Ego. Most assume the universality of the Oedipus complex and the sublimation of aggression. Jungian approaches posit a universal set of archetypes, whose meanings are personally or locally adapted to solving problems.

Devereux (1961) said that art expresses taboos in palatable, aestheticized forms. Taboos may be universal (e.g., incest); culturally specific (e.g., sex in a puritanical society); or idiosyncratic (neurotic). Arts are admired locally when the taboos are cultural, and more widely when they relate to universal human problems. This 'aesthetics as alibi' approach also underlies the works of Margaret Mead, Bateson, Schieffelin, and to some extent Kardiner.

Fischer (1961), following Wallace and Barry, proposed that artists sense how their patrons receive satisfactions to needs generated by local sociopsychological contexts: the latent (nonrepresentational) content of art, the 'style,' presents aestheticized versions of social fantasies giving security or pleasure. For instance, symmetrical designs with simple, repetitive elements characterize the arts of egalitarian societies, whereas the arts of complex hierarchical societies present enclosed asymmetrical designs integrating disparate elements. He also found that preferences for

the curvilinear reflected male security and fantasizing about females. Teilhet (1978) suggested that the equivocal role of women artists in nonliterate societies resulted from men's jealousy of women's reproductive powers, compensating by control of artistic creativity and the hoarding of valuable materials.

3.2 Linguistic Approaches

Art, like language, is a communicative form, using a medium that carries a coded message understood by both creator and audience. Forge claimed that, for the Sepik peoples, visual art could be read directly, but not reduced to words. In contrast, Munn showed how Australian Walbiri painting consists of a number of iconic but polyvalent elements, combined in sentence-like structures, to make symbolic statements about dream-time ancestors, often suggesting speech.

Lévi-Strauss (1966) posits that 'art lies halfway between scientific knowledge and mythical and magical thought … midway between design [the scientist builds structures from first principles] and anecdote [the *bricoleur* makes things from the 'debris of past events'] … The aesthetic emotion is the result of this union between the structural order and the order of events.' Other structuralist analyses, such as Adams' on Sumbanese textiles, reveal underlying systems of meanings interpreted locally.

For Bloch (1974) relations between aesthetics and emotion may be summarized as 'You cannot argue with a song!' He makes the analogy: Phatic/illocutionary speech is to ritual/art as discursive/propositional speech is to secular behavior and mundane expression. Formal mechanisms, for example, repetition, symmetries, limited range of elements, fixed sequences, and archaic images, give the authorities who control ritual, rhetoric, and the arts, power over those enculturated to these patterns.

3.3 Sociopolitical Approaches

Social anthropologists have stressed the functions of art in maintaining the social. Turner showed how Ndembu political rites use normative and naturalistic symbols and colors to structure the event and foster *communitas* among the participants, and Biebuyck showed how human and animal figurines recall proverbs, having the function of training initiates in the Lega *bwami* cult. Gell (1998) asserts that anthropologists focused wrongly on aesthetics, and on relationships between formal style and other cultural characteristics, because (a) 'art' is not always about aesthetics; and (b) the expression of a culturally specific aesthetic system is *not* the function of art objects. Art is created to function in social relations to do something, that is, it has agency. In the West, art serves class and power interests, in gender exclusion (Nochlin) and the maintenance of hierarchy (Bourdieu).

3.4 Aesthetics, Ethno-aesthetics and Universals

Little is understood yet about aesthetic universals. A number of anthropologists have explicated the local, i.e., ethno-aesthetic systems of, for example, the Yoruba (Thompson), the Kilenge (Dark), and the Inuit (Graburn). Some have tried inconclusively to test the cross-cultural nature of aesthetic taste by, for instance, showing photographs of art objects both to their non-Western makers and to Western students of art (Child and Siroto). However, Graburn (1978), drawing on the work of Salvador among the Kuna, was among the first to use a test exhibit of the actual art objects in comparing Native (Inuit) and Euro-American preferences. Other approaches, drawing on Wittgenstein's philosophical work and works on Paleolithic and ethnographic arts (Conkey, Marshak) show promise; there are possibilities that underlying aesthetic universals are related to neuro-ocular phosphenes.

4. Art in the Contemporary World

Early anthropological analyses focused on embedded 'authentic, traditional' arts, known as 'primitive' or 'non-Western' arts (Gerbrands 1957). In *Ethnic and Tourist Arts* (1976) Graburn examined their entry into the 'World Art System,' looking at the primary audiences (locals or the external market) and the sources of aesthetic forms (traditional, external models, or novelties). Others have focused on such arts, in Africa and the Americas where traditional 'primitive' arts were found, and in literate civilizations such as Thailand which have fallen under Western influences (Phillips 1992). Questioning the historically contingent distinctions of hybrid/purebred, Morphy (1992) revealed identical structural and aesthetic patterns underlying functional (locally directed) and commercial (market) Yirrkala bark paintings. Neich examines a colonial era (1870–1910) form of Maori figurative painting influenced stylistically by *pakeha* (white) culture, but expressing Maori cosmology.

Cole, Appadurai, Clifford, and Thomas have recently shown the historical depth and sociocultural complexity of art production in colonial or post-colonial, often touristic, contexts. Dealers, curators, and tour directors may present historical artifacts as 'purebred' symbols of rich heritages, but practically everything in our ethnographic collections was acquired in colonial situations. They may be trade

replicas or, like the carved arts of the Northwest Coast Indians of North America, they may have enlarged and proliferated with new tools and economic stimuli. Often colonists both consumed and imitated contact arts, attempting to assume a new 'ethnic' identity by appropriating the culture of the conquered.

Recent works have focused less on the objects and more on the agency of the personnel of the art world. Ethnohistorical analyses (Cohodas, 1999, Batkin 1999) and ethnographic research (Steiner 1995) reveal the complex activities of artists, intermediaries, and consumers. Artists may work consciously in a variety of styles, aiming at a number of consumer markets, or they may themselves play the role of intermediaries.

When artists and consumers are culturally or geographically separated, mediating agents assume greater importance. They not only transmit the physical art object from the producer to the consumer, but they also control the flow of information about its origin, age, meaning, and status as a commodity or a treasure (Appadurai 1986). They also transmit back the demands and the ideology of the collectors, in terms of price, repeat orders, or information about form, content, color, and materials.

Artists dependent on the cross-cultural market are rarely socialized into its symbolic and aesthetic system, knowing little about the art world, or even about 'Art' itself. They live in a minefield of rules which they overstep at their peril, unless they assimilate into the 'art world' (Graburn 1993). The Australian Aborigine watercolorist, Namatjira, painted successfully in the style of his white mentor, engendering both praise and jealous racism among the Australian public, leading to his downfall. Today's Third and Fourth World peoples have increasing exposure to mainstream arts and art materials. Many have attended art schools and assumed the burdens of (Western) art history and the ideological values of originality, artistic freedom, and individual creativity.

The separation of the art circuits for Native arts and mainstream (Euro-American) arts is eroding. Many Native artists work in genres of the metropolitan world art. The contents of their art may not be recognizable as 'ethnic'—possibly because white artists have appropriated so many ethnic motifs! When competing in the world market, Native artists signal their ethnicity through the presentation of traditional content, inclusion of stereotypical motifs and color schemes; choice of title or use of a Native name (or 'tribal' designation with a nonethnic name); or the inclusion of written statements expressing a Native point of view (Sturtevant 1986). Inclusions of non-Native contents, motifs, names and titles may draw criticism or rejection by the market.

Ethnic artists are everywhere pressured to create arts conforming to the image desired by the mainstream market. Even 'Magiciens de la Terre' which exhibited 'world artists' in Paris in 1987, revealed by juxtaposition how the 'ethnic' artists were expected to create something stereotypically traditional, but the metropolitan (white) artists drew their motifs or materials from natural or cultural alterity (Otherness) without limitation.

Very few anthropologists have tackled the core institutions and activities of the metropolitan art world itself. De Berlo, Gerbrands, Graburn, Krech (1999), and Sally Price have all examined how this art world has classified and examined non-Western arts. Myers, in his work on contemporary Australian aborigine dual purpose (sacred and commercial) arts has best analyzed the Western art discourse about such arts (Marcus and Myers 1995). Plattner (1997) has showed how hierarchies of prestige, stemming from New York, control economic rewards which dictate the life choices of a range of professional artists in St. Louis. This is the field in which the disciplinary boundaries between anthropology, art history, and cultural studies dissolve in synthetic collections in books and in journals such as *Res*, *Art in America*, or *Visual Anthropology Review*, reflecting the final breakdown of the boundaries between art, craft, material culture, and perhaps of the singularity of 'Art' itself.

See also: Art and Culture, Economics of; Art, Sociology of; Arts Funding; Authenticity, Anthropology of; Craft Production, Anthropology of; Dance, Anthropology of; Music: Anthropological Aspects; Popular Culture; Visual Anthropology

Bibliography

Appadurai A 1986 *Social Life of Things: Commodities in Cross-Cultural Perspective*. Cambridge University Press

Batkin J 1999 Tourism is overrated: Pueblo pottery and the early curio trade. In: Phillips R, Steiner C (eds.) 1999 *Unpacking Culture: Art and Commodity in Colonial and Post-colonial Worlds*. University of California Press, Berkeley, CA, pp. 282–97

Becker H 1982 *Art Worlds*. University of California Press, Berkeley, CA

Bloch M 1974 Symbols, song, dance and features of articulation: Is religion an extreme form of traditional authority. *European Journal of Sociology* **15**: 55–98

Boas F 1955[1927] *Primitive Art*. Dover, New York

Cohodas M 1999 Elizabeth Hickox and Karuk basketry: a case study in debates on innovation and authenticity. In: Phillips R, Steiner C (eds.) 1999 *Unpacking Culture: Art and Commodity in Colonial and Post-colonial Worlds*. University of California Press, Berkeley, CA, pp. 143–61

Coote J, Sheldon A (eds.) 1992 *Anthropology, Art and Aesthetics*. Clarendon Press, Oxford

Devereux G 1961 Art and mythology: A general theory. In: Kaplan B (ed.) *Studying Personality Cross-culturally*. Harper & Row, New York

Fischer J 1961 Art styles as cultural cognitive maps. *American Anthropologist* **63**: 79–93

Gell A 1996 *Art and Agency: An Anthropological Theory*. Clarendon Press, Oxford, UK

Gerbrands A 1957 *Art as an Element of Culture, Especially in Negro Africa*. Brill, Leiden, The Netherlands
Graburn N (ed.) 1976 *Ethnic and Tourist Arts: Cultural Expressions from the Fourth World*. University of California Press, Berkeley, CA
Graburn N 1978 'I like things to look more different than that stuff did': An experiment in cross-cultural art appreciation. In: Greenhalgh M, Megaw J (eds.) *Art in Society: Studies in Style, Culture and Aesthetics*. Duckworth, London
Graburn N 1993 Arts of the Fourth World: The view from Canada. In: Whitten D, Whitten N (eds.) *Imagery and Creativity; Ethnoaesthetics and Arts Worlds in the Americas*. University of Arizona Press, Tucson, AZ
Krech S, Hail B (eds.) 1999 *Collecting Native America*. Smithsonian Press, Washington DC
Lévi-Strauss C 1966 The science of the concrete. In: Levi-Strauss C (ed.) 1996 *The Savage Mind*. University of Chicago Press, Chicago, IL, pp. 1–34
Maquet J 1979 *Introduction to Aesthetic Anthropology*, 2nd edn. Undena, Malibu CA
Marcus G, Myers F (eds.) 1995 *The Traffic in Culture: Refiguring Art and Anthropology*. University of California Press, Berkeley, CA
Morphy H 1992 From dull to brilliant: the aesthetics of spiritual power among the Yolngu. In: Coote J, Shelton A (eds.) 1992 *Anthropology, Art and Aesthetics*. Clarendon Press, Oxford, pp. 181–208
Phillips H 1992 *The Integrative Art of Modern Thailand*. University of Washington Press, Seattle, WA
Phillips R, Steiner C (eds.) 1999 *Unpacking Culture: Art and Commodity in Colonial and Post-colonial Worlds*. University of California Press, Berkeley, CA
Plattner S 1997 *High Art Down Home: An Economic Ethnography of a Local Art Market*. University of Chicago Press, Chicago, IL
Steiner C 1995 The art of trade: on the creation of value and authenticity in the African art market. In: Marcus G, Myers F (eds.) *The Traffic in Culture: Refiguring Art and Anthropology*. University of California Press, Berkeley, CA, pp. 151–65
Sturtevant W 1986 The meanings of Native American art. In: Wade E L (ed.) *The Arts of the North American Indians: Native Traditions in Evolution*. Hudson Hills, New York, pp. 23–244
Teilhet J 1978 The equivocal role of women artists in non-literate societies. *Heresies* **1**: 96–102
Wade E (ed.) 1986 *Native American Arts in Transition*. Hudson Hills, New York

N. Graburn

Art History

The task of art history is to collect, preserve, order, present, and interpret works of visual art. Its activity is particularly visible in the field of historic preservation, aimed at the conservation of historically significant objects as well as regional and urban ensembles, and in the establishment and maintenance of art museums. As invulnerable as the museum buildings may seem and as long-term as the effect of historic preservation can be, the foundations of art history nonetheless lie in a cultural consensus that is not axiomatic, but rather subject to continual fluctuation. Art history is bound to the value placed on its object. In the interest of its own legitimation, it is structurally conservative, yet must also untiringly evaluate whether current artistic developments run counter to the legitimating consensus. Since its invention, therefore, the discipline of art history has had simultaneously to adopt both an open and a defensive posture. Its current possibilities and problems are a product of this conflict.

1. *The Beginnings of Art History*

The Renaissance consciousness of the break between its own art and the stylistic forms of earlier epochs marks the birth of modern art history. These early attempts found their first canonical form in the artist biographies of the artist and theorist Giorgio Vasari, published in 1568 (Rubin 1995). Vasari's *Lives* number among the great achievements of historiography since they, like no other historical work, nourished a conception of the modern age as emerging victoriously from the shadow of the Gothic and Byzantine periods. Since Vasari's time, art historical stylistic designations such as Romanesque, Gothic, Renaissance, Mannerism, Baroque, and Classicism have defined the outline and physiognomy of historical epochs.

With their fixation on the Florentine Renaissance, Vasari's *Lives* imbued early art history with a particular stylistic, geographical, and political slant. The generations that followed saw a considerable expansion of the temporal and geographical context (Bickendorf 1998). With Bernard de Montfaucon's *Monuments de la monarchie françoise* (1729–33) the field of inquiry was expanded systematically to include the Middle Ages, while Comte de Caylus's *Recueil d'antiquités* (1752–67) sought to develop an almost physically precise stylistic history of the art of antiquity (Bazin 1986). More successful, however, was Johann Joachim Winckelmann's stylistic history in his *Geschichte der Kunst des Alterthums* (1764). By deriving the Greek artistic ideal from the political climate of freedom, the work became an element of prerevolutionary discourse (Bäumer 1986).

2. *The Art History of the French Revolution*

Accordingly, the French Revolution saw itself as proof of Winckelmann's conviction that only political freedom could provide a suitable forum for great art. On the basis of this premise, works of art confiscated from the church and the emigrated nobility became a

Figure 1
Benjamin Zix, allegorical portrait of Dominique-Vivant Denon, drawing, Paris, Louvre, Cabinet des Dessins

national legacy. Selected pieces on view in the *Musée des Monuments Français*, opened in 1795, constituted a first ensemble of native artistic development no longer bound to the norms of Vasari and Winckelmann. Art history had become a national affair (Wescher 1976).

No one person embodied this shift more completely than Dominique-Vivant Denon, the 'eye of Napoleon.' Denon was responsible for selecting and overseeing the vast quantity of works from all epochs and regions brought together in the Louvre as the booty of the French army. Benjamin Zix, who accompanied Napoleon's victorious advances with his draftsman's pen, drew an allegorical portrait of Denon surrounded by masses of antique and postantique works and books, an image that aptly characterized the situation (Fig. 1). The sheer range and quantity of works in question made a stylistically, dynastically, and religiously neutral art history, equipped with transepochal methods, an objective necessity. Since in the French view the works had been liberated from foreign captivity into the Louvre, this European perspective contained a national element as well (Denon 1999).

3. *Establishment in the Universities*

One effect of the Napoleonic plundering of works of art was an emphatic association of art history with national consciousness. The art museum erected by Karl Friedrich Schinkel between 1825 and 1830, for example, was intended to help give Prussia a new identity not only through the founding of the University of Berlin, but also through an art museum (Fig. 2). Among the curators of the museum, Gustav Hotho gave lectures at the university in Berlin, while Gustav Friedrich Waagen taught as an adjunct professor in order to build a bridge between the cultivation of taste as represented by the museum and the academic training in the university. This constituted the first step toward the academization of art history (Dilly 1979).

Elsewhere in German-speaking Europe, attempts were likewise made to cultivate art history as an academic discipline, but institutional establishment succeeded only with the appointment of Jacob Burckhardt as professor of art history at the *Eidgenössische Hochschule* in Zurich (1855). Burckhardt therewith became one of the founding fathers of academic art history. In the years that followed, professorships were likewise established at polytechnic universities (Stuttgart, Karlsruhe, and Berlin–Charlottenburg, 1865–1868), since the training of architects in the period of historicism required an art historical foundation (Beyrodt 1991).

The first professorships for art history at classical universities were established in Bonn and Vienna (1862 and 1868), followed by Strasbourg, Leipzig, and Berlin after the founding of the Empire (1871–1873). By the turn of the century, numerous other universities had followed suit, viewing art history as necessary not only for the specific training of museum curators, but for aesthetic education in general. A century after revolutionary France had appropriated for itself the status of homeland for all masterpieces, the cultural policies of imperial Germany aimed at cultivating the general public taste through art history had become the driving force behind the academization of the field.

In France, where art history had been taught in Lyon since 1876 and in Toulouse since 1893, the academic tradition began only in 1893 with the establishment of a professorship for Henri Lemonnier at the Sorbonne and Emile Mâle's appointment to Toulouse (Therrien 1998, pp. 314ff.). In the United States, efforts had been made since the 1920s to disengage art history from archaeology or general artistic education and to make it a separate academic discipline; the breakthrough, however, occurred with the arrival of such German emigrants as Erwin Panofsky and Max J. Friedländer (Michels 1999, Wendland 1999).

In England, individual researchers such as Herbert Horne had overcome the tradition of the erudite writer as represented by John Ruskin. With the foundation

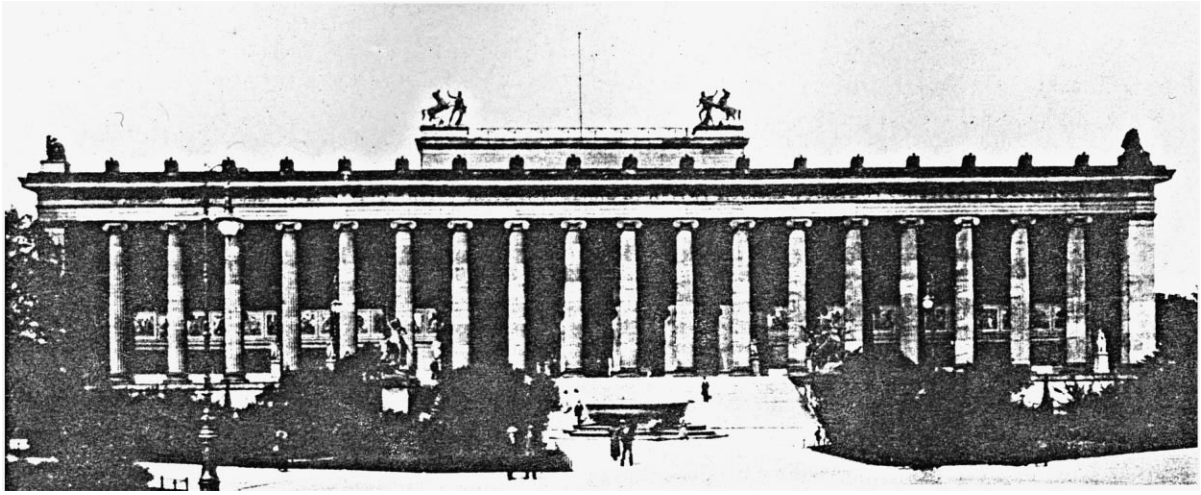

Figure 2
Karl Friedrich Schinkel, Museum, Berlin

of the Courtauld Institute in 1932, art history was officially established as a scholarly discipline, a development furthered by the transfer of the Warburg Library from Hamburg to London in 1933 (Mann 1993).

4. *Methods*

Art history as an academic discipline was mainly the product of German-speaking Europe, a fact that had consequences for the development of its specific methods. These ranged from methods strictly oriented to the object to philosophical and cultural–historical approaches.

In analogy to the critical methods of the historians, Friedrich von Rumohr and Gustau Friedrich Waagen, director of the Gemäldegalerie in Berlin, developed a method of source criticism and a technique of formal observation that permitted the assessment and continuation of Vasari's *Lives* (Bickendorf 1991). The achievements of this school were embodied in the *Allgemeines Künstlerlexikon*, planned by the two private scholars Ulrich Thieme and Felix Becker and published in 37 volumes from 1907 to 1950, with entries for ca. 150,000 artists and craftsmen from 92 countries.

Post-Hegelian Berlin saw the emergence of Gustav Hotho's and Carl Schnaase's developmental histories, oriented to the philosophy of the world spirit, and Franz Kugler's history of world art encompassing all times and peoples, derived from the Kosmos lectures of Alexander von Humboldt. Burckhardt's cultural-historical orientation likewise drew essential inspiration from this climate. Burckhardt strove for an art history 'according to problems,' in other words in accord with external aims, in order, on the basis of this kind of investigation, to be able to define the autonomy of art even more precisely (Waetzold 1924, Vol. 2, pp. 47–92, 141–210).

The two poles of dependence on external conditions and the unmistakable uniqueness of art determine art historical methods to this day. A series of attempts undertaken around 1900 on the basis of Burckhardt's work, aimed at responding to modern experience, still maintain their essential validity (Podro 1982). Representatives of the Vienna School such as Alois Riegl operated with the concept of the *Kunstwollen* in order to develop an art history that transcended individual artistic personalities. Riegl's *Spätrömische Kunstindustrie* (1901), which sought to structurally analyze a hitherto unconsidered epoch, used late antiquity as a mirror to reflect the art industry coming to expression in the mass culture of Riegl's own time. With a view to an 'art history without names' in accord with the abstractions of modern art, Burckhardt's Swiss student Heinrich Wölfflin developed criteria of a stylistic history which, as a strict school of vision, rejected everything oriented to other disciplines.

With sensitivity to contemporary tensions between art and politics, the Hamburg art historian Aby Warburg developed a method diametrically opposed to that of Wölfflin: iconology. Already in his dissertation on Sandro Botticelli (1893), Warburg dealt with problems of the psychology of form connecting historical forces of action and image-making. His dictum, formulated in 1912, that art history should be written as cultural history in a struggle against the boundaries of the academic disciplines (Warburg 1999, p. 585) lives on in present-day efforts to understand artistic form in its liturgical, political, and social

function and its relation to literature and documentary sources. Warburg's foundation of the *Kulturwissenschaftliche Bibliothek Warburg* and the development of the iconological method constitute probably the most profound impulse art history has contributed to the history of scholarship in the twentieth century. The continuing publication and translation of his collected works into English will probably further increase his influence.

Another response to the problematization of the ego and the artistic personality in modern mass culture was the sociology of art, a field that moved between the poles of Burckhardt—who, although he wanted to write art history on the basis of defining 'problems,' nonetheless saw it as an equal partner alongside the political and social realm—and Karl Marx, who derived artistic activity from social conditions (Warnke 1970). The sociology of art has researched the changing social status of the artist, examined relations between patrons and artists, defined the determinative role of function for artistic form, explored the role of patronage, and reconstructed historical manifestations of the art market. In addition, stylistic forms have been associated repeatedly with socially determined ways of seeing and thinking.

Landmarks of the sociology of art were written on the Italian Renaissance. Fundamental works dedicated to the status of art and artists (Wackernagel 1938, Kempers 1992, Goldthwaite 1993) and to the sociology of style (Antal 1947, Meiss 1951). Meiss's work, inasmuch as it mediates between style and class structure far more flexibly than does Antal, numbers among the most important works of art sociology of all time. Hauser's handbook (1953) constitutes the most comprehensive work on the social history of art and is notable for having also included film as a genuine area of art history. In contrast to Hauser's social history, oriented to the bourgeoisie and the cities, Warnke's work on the court artist (1996) investigates the princely court as the motor of artistic development, while Bätschmann (1997) focuses his analysis on the modern market- and exhibition-oriented artist.

5. *Methodological Pluralism*

Postmodernism has not failed to leave its mark in art history, especially since the discipline played a determinative role in the establishment of the term through the medium of architectural criticism. The attacks on methodological certainties, historical teleologies, and one-dimensional perspectives have above all been directed at those areas of art history that focused on high art and studied it using the criteria of stylistic development. While the history of style and connoisseurship have asserted themselves as the basis for all art historical work, the overall picture of the discipline is defined by a plurality of methods such as previously existed only between ca. 1900 and 1933 (Preziosi 1989, Belting 1983, 1995).

The field of gender studies has been particularly active in recent decades, pioneering not only the reconstruction of works by women artists from the Middle Ages to the modern period, but also uncovering gender-determined ways of looking and representing (Pollock 1999).

The history of relations between semiotics and art history belongs in the broader field of methodological expansion. To what extent exponents such as Roland Barthes are the secret heirs of this interchange remains an open question. Although some representatives of art history have used semiotic approaches since around 1970, whether these methods can do better justice to the differentiated nature of complex works of art than traditional hermeneutics is still controversial. The same holds true for structuralist ethnology as represented by Claude Lévi-Strauss. He oriented himself to iconology, and the latter profited from his work as well (Bredekamp 1995b). Finally, an expanding field is the connection between art history and the history of science, extending from the sixteenth century to the present (Stafford 1991, Bredekamp 1995a, Crary 1990).

6. *Universalization*

From the fifteenth to the eighteenth century, the dominant language of art history was Italian, around 1800 it was French, from 1820 to 1933 German, and thereafter English. The Anglicization of art history, in which the work of German emigrants to England and the United States played a decisive role, also ensured that after World War II its field extended beyond Europe to all of world art. From the originally narrowly defined field of Italian art of the late Middle Ages and the Renaissance, art history has expanded to include everything from late antiquity to the present. Geographically, art history deals with the art of all continents, although for other continents besides Europe—where late antiquity constitutes a relatively clear line of demarcation to the field of archaeology—the temporal boundaries are less sharply defined. The universal aspirations of the discipline have enabled it to discover that other cultures such as China have a history of art as well, in which elaborate standards have been developed over centuries (Ledderose 2000). Today, the study of art history is established in all countries that have achieved an initial consolidation after the wave of industrialization.

This worldwide orientation has been embodied in encyclopedic form since 1983 in the publication of the *Allgemeines Künstlerlexikon*, a work that will include entries for several hundred thousand artists from all periods and countries. The symbol of art historical

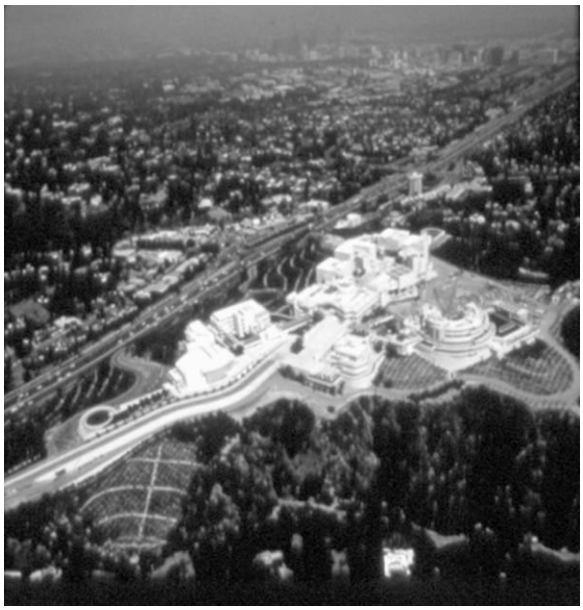

Figure 3
Richard Meyer, Getty Center, Los Angeles

universalism is the research institute of the Getty Center in Los Angeles (Fig. 3), whose mere location makes it clear that European art history does not stand alone. This does not mean, however, that its methods are useless. The objection that the Western conception of 'art' as an elevated sphere is unsuitable for Eastern cultures and that art history can gain no suitable access to this hemisphere presupposes that the discipline is concerned only with high art. This, however, is a prejudice that serves to resist all new fields and all questions of medium.

Within architecture, sculpture, and painting, art history by no means deals only with high art. These artistic media have been joined by photography, video, computer art, and, in analogy to the historic interest in arts and crafts, the history of design and advertising. Despite promising earlier approaches, however, film and television have not been adopted into the canon of art historical objects with the same consistency, although more recent attempts are striving for a historical foundation in this field as well.

In this context, the question of the internal consistency of art history arises with unprecedented clarity. The art forms that have appeared in the twentieth century have provoked the question of whether art history should also devote itself to the images of the so-called New Media or leave them to other disciplines such as film studies or cultural studies (Belting 1995). Through a narrower definition of art history and its confinement to the traditional media of architecture, sculpture, painting, and perhaps photography the discipline would thus be a kind of art history of the predigital age, a 'second archaeology.'

Yet when we consider the recently published *Dictionary of Art* (1996)—a total of 36 volumes with 41,000 articles dealing with persons, artists, art historians, collectors, and all relevant geographical and object areas as well as photography, performance, and multimedia from the entire world, from which film has been excluded only because it is such a comprehensive area in itself—the conflict seems to have been decided in favor of an open conception of art history. Here the discipline manifests itself as the historical science of images defined by the early Warburg School already at the beginning of the twentieth century.

The same process of integration of the New Media is occurring in museums of fine art, applied arts, and related areas. In the last decades of the twentieth century even elaborate museums were built as in all preceding epochs put together, a trend that continues unbroken to the present. In many places in the world, for example, in Bilbao with its Guggenheim Museum (1997), museums form urbanistic landmarks that shape the character of entire regions and serve as points of attraction for cultural and mass tourism. They thereby come to represent a considerable cultural and economic force (Lampugnani and Sachs 1999). The media of video and digital art are included in many of these museums, and the *Zentrum für Kunst and Medientechnologie* in Karlsruhe (1996) has for the first time devoted an entire complex to these new media, consisting of both an art museum and a laboratory. As the response to a media revolution encompassing the world but remaining ephemeral in its production, museums as the primary seats of art history remain the fixed points of a locally accessible but materially permanent realm of experience, one that also incorporates the products of the art–technological revolution.

See also: Art and Culture, Economics of; Art: Anthropological Aspects; Art, Sociology of; Fine Arts; French Revolution, The; Postmodernism: Philosophical Aspects; Renaissance

Bibliography

Antal F 1947 *Florentine Painting and its Social Background: The Bourgeois Republic before Cosimo de Medici's Advent to Power, XIV and Early XV Centuries*. Paul, London

Bäumer M L 1986 *Klassizität und republikanische Freiheit in der außerdeutschen Winckelmann–Rezeption des späten 18. Jahrhunderts*. In: Gaehtgens Th W (ed.) *Johann Joachim Winckelmann 1717–1786*. Gaehtgens, Hamburg, Germany, pp. 195–219

Bazin G 1986 *Histoire de l'histoire de l'art de Vasari à nos jours*. Albin Michel, Paris

Bätschmann O 1997 *Ausstellungskünstler: Kult und Karriere im modernen Kunstsystem*. Dumont, Cologne, Germany

Belting H 1983 *Das Ende der Kunstgeschichte? (The End of History of Art?)*. Deutsche Kunstverlag, Munich, Germany
Belting H 1995 *Das Ende der Kunstgeschichte. Eine Revision nach zehn Jahren*. Beck, Munich, Germany
Beyrodt W 1991 *Kunstgeschichte als Universitätsfach. Kunst und Kunsttheorie 1400–1900*. Harrosowitz, Wiesbaden, Germany, pp. 313–33
Bickendorf G 1991 *Die Anfänge der historisch–kritischen Kunstgeschichtsschreibung. Kunst und Kunsttheorie 1400–1900*. Harrosowitz, Wiesbaden, Germany, pp. 359–74
Bickendorf G 1998 *Die Historisierung der italienischen Kunstbetrachtung im 17, und 18. Jahrhundert*. Gebrüder Mann, Berlin
Bredekamp H 1995a *The Lure of Antiquity and the Cult of the Machine. The Kunstkammer and the Evolution of Nature Art and Technology*. Markus Wiener Publishers, Princeton, NJ
Bredekamp H 1995b *Words, Images, Ellipses. Meaning in the Visual Arts: Views from the Outside. A Centennial Commemoration of Erwin Panofsky (1892–1968)*. Princeton, NJ, pp. 363–71
Crary J 1990 *Techniques of the Observer: On Vision and Modernity in the Nineteenth Century*. MIT Press, Cambridge, MA
Denon D-V 1999 *l'oeil de Napoléon*. Exhibition catalogue. Editions de la Réunion des Muśees Nationaux, Paris
Dilly H 1979 *Kunstgeschichte als Institution*. Suhrkamp, Frankfurt, Germany
Goldthwaite R A 1993 *Wealth and the Demand for Art in Italy 1300–1600*. Johns Hopkins University Press, Baltimore, MD and London
Haskell F 1993 *History and its Images: Art and the Interpretation of the Past*. Yale University Press, New Haven, CT
Hauser A 1953 *Sozialgeschichte der Kunst und Literatur. (The Social History of Art)*. Beck, Munich, Germany
Kempers B 1992 *Painting, Power and Patronage*. Penguin, London
Kultermann U 1993 *The History of Art History*. Aboris Books, New York
Lampugnani V M, Sachs A 1999 *Museen für ein neues Jahrtausend. Ideen Projekte Bauten*. Prestel, Munich, London, and New York
Ledderose L 2000 *Ten Thousand Things. Module and Mass Production in Chinese Art*. Princeton, NJ
Mann N 1993 *The Warburg Institute—Past, Present and Future. Porträt aus Büchern. Bibliothek Warburg und Warburg Institute*. Dölling und Galitz, Hamburg, Germany, pp. 133–43
Meiss M 1951 *Painting in Florence and Sienna after the Black Death*. Princeton University Press, Princeton, NJ
Michels K 1999 *Transplantierte Kunstwissenschaft: deutschsprachige Kunstgeschichte im amerikanischen Exil*. Academie Verlag, Berlin
Podro M 1982 *The Critical Historians of Art*. Yale University Press, New Haven
Pollock G 1999 *Differencing the Canon: Feminist Desire and the Writing of Art's Histories*. Routledge, London and New York
Preziosi D 1989 *Rethinking Art History. Meditations on a Coy Science*. Yale University Press, New Haven, CT and London
Riegl A 1901 *Spätrömische Kunstindustrie. (Late Roman Art Industry)*. Druck und Verlag de Kaiserlich-Koeniglichen Hof- und Staatsdrukerei, Vienna
Rubin P 1995 *Giorgio Vasary, Art and History*. Yale University Press, New Haven, CT
Stafford B 1991 *Body Criticism. Imaging the Unseen in Enlightenment Art and Medicine*. MIT Press, Cambridge, MA and London
Therrien L 1998 *L'histoire de l'art en France. Genèse d'une discipline universitaire*. Editions du C.T.H.S. Paris
Vasari G 1906 *Vite de' più eccelenti architetti, pittori et scultori italiani (The Lives of the Painters, Sculpturs, and Architects)*. Sansoni, Florence, Italy
Wackernagel M 1938 *Lebensraum des Künstlers in der Florentinischen Renaissance. Aufgaben und Auftraggeber, Werkstatt und Kunstmarkt (The World of the Florentine Renaissance Artist: Projects and Patrons, Workshop and Art Market)*. E A Seemann, Leipzig
Waetzold W 1924 *Deutsche Kunsthistoriker*, 2 Vols. E A Seeman, Leipzig
Warburg A 1999 *The Renewal of Pagan Antiquity*. Getty Research Institute for the History of Art and the Humanities, Los Angeles
Warnke M 1970 Jakob Burckhardt und Karl Marx. *Neue Rundschau* **81**: 702–23
Warnke M 1996 *Hofkünstler: zur Vorgeschichte des modernen Künstlers (The Court Artist: On the Ancestry of the Modern Artist)*. Dumont, Cologne, Germany
Wendland U 1999 *Biographisches Handbuch deutschsprachiger Kunsthistoriker im Exil: Leben und Werk der unter dem Nationalsozialismus verfolgten und vertriebenen Wissenschaftler*. Saur, Munich, Germany
Wescher P 1976 *Kunstraub unter Napoleon*. Mann, Berlin
Wölfflin H 1915 *Kunstgeschichtliche Grundbegriffe: das Problem der Stilentwicklung in der neueren Kunst (Principles of Art History: The Problem of the Development of Style in Later Art)*. Bruckmann, Munich, Germany

H. Bredekamp

Art, Sociology of

Evidently, the sociology of art experiences, to constitute itself as a specific domain, the same difficulties as those encountered at the end of the nineteenth century by sociology. Despite the numerous attempts made since the 1960s to bring about progression, one wonders to what extent the bitter statement established in 1960 by P. Francastel is appropriate today. In fact, despite the abundance of publications on the subject, the attempt to constitute a reasoned bibliography of the question leads us once again to discover that the best ways to constitute a sociology of art are found under different bibliographical headings. Having revealed this initial reserve, we rapidly need to present in this article the main work already carried out in order to constitute a sociology of art, then to indicate the methods which seem to be the most fruitful to achieve this.

Essentially three methods have already been widely explored. The first, which claims to be scrupulously loyal to a classical conception of sociology, situates its object, art, within the epistemological framework of

its discipline. It shows that art is a collective phenomenon, crossed over and invested through and through by the society and history. One can therefore link, to give an idea of this enterprise, the works of H. Becker, P. Bourdieu, R. Moulin and so many others with those of S. Alpers, M. Baxandall, and F. Haskell. The privilege which is given here to collective phenomena and denied of individual creations can result in two forms of reduction, that which Marxist sociology operated, that of Lukacs, Hauser, Goldmann, and that cultivated by P. Bourdieu, which criticizes both the Marxist sociologists and those who maintain according to them the fancy of a creating subject, freed from institutional constraints and history.

The second proceeds in the opposite way to the first and recommends the singularity of the artist and the work, singularity which tends towards the solipsism of a creator out of text, nonaltered by his belonging to a society and a history. Cultivated in this method is the biography and study of style, following which the artist, considered more as a creator than as a producer, appears to be the carrier of a singularity which could be entered into by nothing.

The third, more recent, will delve into the resources of analytical philosophy, and of ethnomethodological interactionism, pragmatism. What is of prime importance is no longer in general, neither the artist nor the work, in a word the world of art, but the method, the way in which the world is thought of and told about. The practice of works separates itself from what it considers to be hindrances to research. The descriptive concern incites us to refute the normative approach. One is no longer trying to define, legitimize, or invalidate values, but rather to take into account what actors have lived, how they define, legitimately or not, these values. It is therefore necessary for the sociologist, not or no longer to say what art is worth or what it means, but what it does, no longer what sociology does to art, but 'what art does to sociology.' There again, the results are unequal. A. Danto, N. Goodman, and N. Heinich, despite their differences, enter into these research perspectives.

The first two methods being classical, one must note here some difficulties and inherent aporias with this third method as presented by N. Heinich. The priority given to the description and questioned by the normal approach causes a series of epistemological prejudices which can arouse reservations. Indeed, the statement is expressed in a description based on an evaluation, which itself is based on values. The methodological concern which N. Heinich describes seems to reveal useless precaution: 'To be as attached as possible to description, while abstaining from all normativeness (…) implies for the sociologist an important reduction in his space for intervening.' The 'pragmatical' sociology which it claims to be chooses a method which claims to be a-critical, pluralistic, relativistic, also claiming the epistemological neutrality of Weber. This project, in the very terms of its wording, can seduce: nothing is more consensual as all these terms which rhetorically oppose, even in an implicit way, their solar face with severe shadows of criticism and univocity which are supposedly reducing. However, this sociological concern which gives way to the fetishism of methods and which deludes itself despite its pretensions about the nature of axiological neutrality, brings about a major drawback well conceptualized by analytical philosophy. From the moment we are interested in art, we must recognize that a work of art is not an object which is given and which can be observed and described in a neutral way, but a semiotic entity which is constructed within and by interpretation. In order to avoid an incorrect conception of the work of art, we must admit that it supposes intentionality and representation. The two together suppose an interpretation, which is not hermeneutic but constituting. 'The work and the interpretation are born together in aesthetic conscience' writes A. Danto whose important theses must be remembered. The path followed in the transfiguration of the ordinary is in two stages. The first stage, negative, claims that the ontological specificity of the work of art does not reveal a material or perceptual property: the work of art, as is stated above, is not given to us in the same way as the objects of the world. The second stage, positive, tries to establish the criteria which are no longer perceptual but conceptual, categorical, thanks to which works of art come about for us as works. Posture which, by disciplinary provincialism no doubt, claiming to be very descriptive, lacks an essential dimension of the art with which sociology must count if it wants to avoid the trap of sterility, be it or not disguised in methodological precaution or in ethical concern.

N. Goodman is quite right to notice that the real question 'what is art?' too often has as its implicit equivalent this less correct question 'what is good art?' (It is useless to evaluate here to what extent this question can fuel arguments or abuse of those who consider contemporary art as trickery.) If it is possible to not share these viewpoints on the answers he gives, it is difficult to not join him in the validity of this correct question 'when is there art?' One can connect to this interrogation another question which stems from what we will call a cognitive sociology of art: in what and how can art be a source of knowledge for the sociologist?

Very simply but very usefully, art can function as a sign or symptom for the sociologist, and therefore serves as a political and social revealer. Tocqueville already underlined this cognitive dimension: 'When I arrive in a country and I see the arts giving some admirable products, this teaches me nothing about the social state and the political constitution of the country. However, if I notice that the products of the arts are generally imperfect there, in large quantities and at low prices, I am certain that, for the people where this occurs, privileges are weakening and the

classes are starting to mix together and will soon be confused.' This cognitive dimension exists in all the areas where art is enrolled at least partially in institutional dimensions. Adorno gives an interesting demonstration of this in the sociology of music; it can also be observed for example in urbanism and in architecture: monuments, the choices made in terms of conservation and restoration, translate a political willingness whose monument as a work of art is the vector. A. Riegl established this more than a century ago.

The cognitive contribution of art to the sociology of the same name is demonstrated in the most diverse ways. Whatever the noticeable differences between A. Warburg and E. Panofsky are and even if iconology conjugates no doubt excessively the verb 'to see' with the verb 'to know,' it is true that it paves the way for a fertile route of sociological research, as it gives a distinguished place to the faculty of knowing and ensures the possibility of an objective reading of works, given that this knowledge in no way exhausts the richness of the perception of the work. According to the hypothesis of E. Panofsky, gothic architecture, full of rigor and harmony, would be the pendant, materialized in stone, the rules of academic education recommending severity, rationality, and conceptual hierarchy. The market study of art offers an undeniable interest: its evolution allows for example the measurement of liberties and constraints linked for the artist and the work to the substitution of the tradesman for the guardian. In very contemporary studies, R. Moulin and N. Heinich insist upon the essential cognitive resources for the evolution of the 'world of art.' It is an undeniably rich perspective for the art sociologist to study the factors of contextualization which function like parameters for the actors. In this way, the elements of cooperation and conflict which weave the threads of the 'world of contemporary art' are correctly interrogated: among the art critics, conservationists, experts, buyers, gallery agents, guardians, auctioneers, and the numerous institutions, galleries, museums, sales by auction, local councils, public ministries, the interactions are complex and as it is ultimately up to either the representative of the citizen, which means the political figure, or the cultural intermediary, the representative of the artist, to engage plots, jousts, whose most obviously materialistic stakes are of financial order. We can appreciate the passion which stirs simple citizens and a part of their income, whatever their taste in terms of art, will be allocated to the subvention of works which they appreciate or despise.

The sociologist can also be inspired by the original path cleared by N. Goodman to enrich the sociological theory. For the latter, two types of existence characterize works of art, essentiality and transcendence. Essentiality corresponds to the type of object in which the work is constituted and is distributed in these two regimes which make up the 'autographics' and the 'allographics.' In accordance with the first, the object of essentiality, be it a picture or a sculpture, is material and shows itself. In the second, this object, a text, a musical composition, an architectural plan, is ideal, it is conceived by reduction beginning with its physical demonstrations, books, scores, and performances. Transcendence is defined by the diverse ways according to which a work projects its essentiality, mainly for what is important to the sociologist when the work acts in a different way according to places, periods, individuals, contexts, to use the reformulation which G. Genette gives of it. It is in relation to this last field of investigation that the work does not limit itself to its object of essentiality, because its being is not separable from its action, which leads to answering the question 'when is there art?' Among other contributions, the cognitive virtues to which the art sociologist can be attached are considerable. The properties of works of art act upon the emotions which perceive them; by touching the receiver of the work, they intervene with their cognitive categories by renewing and questioning certain intellectual categories. N. Goodman is quite right to maintain that mental categories grow rich when, concerning an area of red created by Rothko, it becomes possible, in a strange way, but not easily disputable, to recall the 'resonance' of a color. The sociologist is capable of giving precision to the last question by asking for example, when is there modern art and what would a sociology of modern art be made up of. The answer to this question summons up the contributions of the most diverse authors, sociologists, philosophers, and art historians from G. Simmel to L. Steinberg, E. Gombrich and R. Krauss among others. If, to be brief, modernism is distinguished by the emergence of the individual and individuality, how can we characterize certain works as realizing this modernism and how can artistic creation, questioned by the sociologist, increase the sociological knowledge surrounding modernism? Is there, and if so which one, a register of expressiveness which is particular to an art which could be qualified as being modern by the sociologist?

This register is that of simultaneity which art entrusts in all its forms. Simultaneity is declined in different ways. Below are the main ones.

(a) In the species of time, simultaneity is shown in the triumph of the instant compared to time as succession, length, and flow.

(b) In terms of style, simultaneity adopts a non-narrative character which is combined with different effects of oxymore.

(c) Simultaneity in modern art is strongly connected to the notion of living metaphors (P. Ricoeur).

Simultaneity and instance are strongly linked. In 1917, G. Simmel gives a conceptual compulsion in sociology whose impact on the sociology of art is worth underlining, by saying that 'the coexistence of individuals having shared relationships amongst

themselves, causes within each of them what cannot be explained about only one of them.' It appears here that it is not or is no longer subjectivity which leads and precedes intersubjectivity, but the opposite: it is mutual action which generates within all protagonistical effects which inform, model, determine, and constitute their own subjectivity. One must therefore, and this is what Rodin suggests, 'read' by Simmel, give up the classical conception of an ego founder: the me, modern, is crossed, shaped by the encounter with others, it is not constituted in a previous way to the reciprocity of action, it is for a large part the result of it. With Rodin, interpreted by Simmel, but also by Rilke, it is the individual who loses his substantial stability to be, in the light of modernism, thought of as a process, as mobility. This radicalism of movement, displayed by the instant, is expressed no longer in terms of succession but in terms of simultaneity: man is no longer alternatively, following the course of time, happy then alternatively worried, frightened then reassured, gay then gloomy, he lives the harmony of these contrasts within the very moment.

Another perceptible demonstration of the instant in the actual style of works lies within the physical and technical means which, by the impression of incompletion, give priority to the instant, to a time whose length slips away. The importance can be measured not only with Simmel but with others also who dedicate, on this subject, their analyses to Rembrandt and to Rodin. S. Alpers notes that Rembrandt, and this is an element of rupture with his contemporaries, initiates a 'rugged' way of painting. This absence of the 'completed,' contrasted with his contemporaries who adopted as from 1650 a 'smooth' style, at the same time a security for the guardian or the seller, who read within this plainness an impression of toil almost convertible into hours of work, in a word, a guarantee for them. L. Steinberg, in the field of sculpture, notes with Rodin, not the indifference of the artist, but his or her choice to lead perception in the direction of the instant, materialized there again by the impression of incompletion. The sculptor, by deliberate choice, uses all the artifacts intended to show the work as an accidental product of a process. For example, the balls of wet clay, in principal temporary, that the sculptor or his assistants add by precaution in order to be able to give more dimension to a surface, remain within the bronze, even if Rodin is determined that he does not use this. In this perspective, many works of Rodin owe a part of their modern style to this welcoming of the accidental which marks the rise of the instant.

Modernism finishes with narrative, and what surrounds the narrative, the coherence of a univocal story. With Rodin (R. Krauss), the part devolving to time is retracted to that of the instant. The strategy used to do justice to this instantaneous character appears in the choice of the non-narrative procedure by excellence; on the contrary, the *Marseillaise de Rude*, low relief which ornates the *Arc de Triomphe*, chooses, compared with Rodin, to represent the fertile moment, the one which in some ways, in a narrative way, tells and reveals the historical sense of a moment freed from the pure contingency of the event. The 'reading' aroused by the architecture of the low relief invites the spectator to a deciphering regulated by an order of succession, leading the eye from right to left in a movement of elevation to demonstrate progressive awakening of the conscience in the sense of freedom. What the sociologist teaches Rodin is that the latter, in the classical sense of the term, teaches us nothing. He tells no story, teaches us nothing about the flow of an event, does not inform us. From him to us, nothing of any sort of educational ambition, but the distilling of a concern, of an uncertainty which nothing could remove, nor the knowledge of a master, nor the univocal interpretation of the interpreter. In his work *La Porte de l'enfer*, there is no trace of the univocity of a speech skilfully declined by the spatial organization of scenes which would follow each other. On the contrary, all the methods are summoned to erase the apparent narrative pretext which is the reference to Dante; this reference is totally subverted by numerous devices. One will be retained due to the richness of its effects, oxymore.

Oxymore, the fact, despite the logic and its requirement of the excluded third (a traffic light is red or green, it cannot be one and the other at the same time) that in the area of art, more or less and with modern art precisely, the polarity of logical contradictions is displayed and maintained, creates a new language of art. Referring to the paintings of Böcklin, Simmel speaks in an oxymorical way about a happy melancholy of living. This polarity of apparent opposites of which none will ultimately give way to another, is the most effective and concerning vector of modern individuality, conceptually inexpressible but perceptible in modern works. Simmel and E. Gombrich made no mistake. This uncertainty of the effect of oxymore, compared to the univocity of narration, is constant with Rodin. In a simultaneous way, opposed axes of interpretation are imposed upon us, without the receiver of the work having to solve an enigma in order to retain the 'right' interpretation: it is their maintained polarity which is the whole of the modern individual. Rodin thus chose the ambiguity of contradictory meanings. In this way, the *Le Fils prodigue*, with his enormous arms stretched upwards, expresses an indissociable unity of gratitude and despair, of entreaty and actions of grace. With Rembrandt, this procedure of oxymore can be found clearly, via which modern individuality becomes specific. The analyses of Simmel are congruous with those of E. Gombrich, without the latter having read Simmel. E. Gombrich, like Simmel, notices this ability that Rembrandt possesses to surpass and neglect conceptual expression, generalizing and therefore not individualizing passions embodied within such a singular person. The forger who tries to represent the disciples of *Emmaüs*

does not manage to represent the charge of emotion of this scene; he or she exaggerates the line in an expressionistic way. Instead of seizing the singularity of the emotion composed of contradictory and simultaneous emotions, he or she chooses one of them, terror, which he or she freezes in an academical general nature. Instead of this, the work of the hand of Rembrandt creates an effect of oxymore which translates the complexity of life into one of these instants: the sketch of a disciple allows us to see, in a simultaneous way, how the joy of the recognition of Christ is composed of an impression of fear.

Simultaneity in art is partly connected to the notion of metaphor as P. Ricoeur understands it, and with the notion of model as it is conceptualized by R. Boudon and M. Black. The model presumes, compared to a narrative text, a tabular reading, instantly right. Compared to R. Jakobson who stands for a substitutive conception of metaphor, and therefore allows no cognitive content, the metaphor is a method of interaction, in working within modern art, which produced in a cognitive way this sudden appearance, this new aspect, this supplement to knowing that one can hope for without having the means to anticipate it. The metaphor, as P. Ricoeur correctly states, maintains two thoughts of different things simultaneously active within an expression whose signification is the result of interaction. In this sense, the metaphor, towards which art sociologists should direct their thoughts as its effectiveness does not confine itself to this part of art which is modern art, is of the nature of creation of invention. A partial exploration of this is found with A. Danto, and more significantly with N. Goodman.

The consensual relativism currently in fashion in the 'art world' would be eased by an in-depth study of the metaphorical processes which are at the origins of art. Even the relativism of N. Goodman is moderated by this epistemological idea according to which the 'ways of making the worlds' emerge into a diversity without limits but are corrected by the possibility of distinguishing between correct and incorrect versions. Without intending to force the thoughts of N. Goodman, we can usefully come back to Aristotle who offers a definition of the metaphor which is capable of wavering ambient relativism: to metaphorize well is to notice the similar. The 'immeasurable' works of relativism, revisited by the riddle of the metaphor, could therefore be arranged, at least partially, according to a line of intelligibility which distinguishes the similar despite the difference.

See also: Art and Culture, Economics of; Art: Anthropological Aspects; Arts Funding; Censorship and Transgressive Art; Cultural Policy: Outsider Art; Culture, Production of; Culture, Sociology of; Fine Arts; Prehistoric Art; Simmel, Georg (1858–1918); Values, Sociology of

Bibliography

Alpers S 1991 *L'atelier de Rembrandt* [trans. Sené JF]. Gallimard, Paris
Boudon R 1995 *Le juste et le vrai, études sur l'objectivité des valeurs et de la connaissance*. Fayard, Paris
Boudon R 1995 De l'objectivité des valeurs artistiques, ou, les valeurs artistiques entre le platonisme et le conventionnalisme. *Archives de philosophie du droit* **40**: 76–95
Boudon R, Clavelin M 1994 *Le relativisme est-il résistible?* PUF, Paris
Bourdieu P, Darbel A 1969 *L'Amour de l'art*. Minuit, Paris
Danto A 1989 *La transfiguration du banal*. Seuil, Paris
Deroche-Gurcel L 1997 *Simmel et la modernité*. PUF, Paris
Genette G 1994 *L'œuvre de l'art*. Seuil, Paris
Gombrich E 1987 *L'art et l'illusion*. Gallimard, Paris
Goodman N 1990 *Langages de l'art*. Jacqueline Chambon, Paris
Heinich N 1998 *Ce que l'art fait à la sociologie*. Les Editions de minuit, Paris
Krauss R 1997 *Passages*. Macula, Paris
Mesure S 1998 *La rationalité des valeurs*. PUF, Paris
Michaud Y 1997 *La crise de l'art contemporain*. PUF, Paris
Moulin R 1992 *L'artiste, l'institution, le marché*. Flammarion, Paris
Ricoeur P 1975 *La métaphore vive*. Seuil, Paris
Riegl 1984 *Le culte moderne des monuments*. Seuil, Paris
Simmel G 1994 *Rembrandt* Circé, Paris
Steinberg L 1991 *Le retour de Rodin*. Macula, Paris

L. Deroche-Gurcel

Artificial and Natural Computation

1. Introduction

1.1 Survey

This article discusses the notion of computation in natural, as well as in artificial systems. Particular emphasis lies on artificial computation models which are derived from natural systems that perform computation. The term 'computation' here refers to all processes that systematically evaluate a certain (mathematical) function. That is, computation refers to all processes that determine a value or magnitude depending on one or more factors (parameters). The article sketches the main direction of research in the past decades on modeling computational processes in natural systems. This concerns on the one hand computation in nervous systems, such as the human brain and their artificial models. On the other hand, there is the process of biological evolution, of which artificial computer models proved useful as a general approach to solving difficult design and optimization problems.

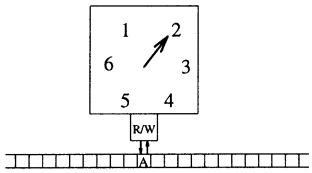

State	Symbol		State	Symbol	Head
1	A	→	2	A	L
1	B	→	3	C	R
1	C	→	6	C	L
...	...		...	...	...
6	D	→	3	C	R

Figure 1
The schema of a Turing machine (left) and the corresponding Turing table defining the operations of the Turing machine (right)

1.2 Historical Background

In the 1930s a precise mathematical notion of computation was developed, which also captures everything that can be computed in modern computers. What exactly natural computation is, and how it is conducted in natural systems, such as the human brain, is less well understood. However, the mathematical notion of computation has proved a convenient and very useful form for 'describing' information processing in natural systems.

Wiener (1948) introduced the discipline of cybernetics, which was concerned with control systems, including natural systems, such as the human body, in which feedback loops control the secretion of glands, etc. In the following decades, computational models of biological systems were developed and used to address a variety of difficult problems for which conventional solutions were not readily available. The research on models of natural computation has mainly focused on models of the human brain and its neurons on the one hand and on biological evolutionary processes on the other hand.

Another type of computation in 'natural systems' was shown by Adleman (1994), when he theoretically demonstrated the use of DNA molecules as a means of solving computational problems. The work showed the potential for building computers on the basis of molecular biology.

This article focuses on the main research efforts in the past decades and discusses artificial models of natural computation in neural systems. Furthermore, the use of models of biological evolution for solving computational problems will be sketched. The article is organized as follows: first, the mathematical notion of computation is discussed. This is followed by a description of computation in those natural systems, for which most of the artificial models of natural computation have been developed: the human brain and evolutionary processes in biology. The relationship between the natural computing processes and the artificial models of them are thereafter discussed.

2. The Mathematical Notion of Computation

In the 1930s it was an important problem in mathematics to delineate those functions for which the function value can be 'computed' for all possible arguments, from those functions for which the function value may not be effectively computable, i.e., for which the value cannot be determined by executing a finite number of 'calculations' or evaluation steps. (For functions, which contain themselves either directly or indirectly in their definition, it is not straightforward to decide whether a value can effectively be determined for all possible arguments.)

Turing (1937) developed a model for what a mathematician can possibly do, when computing the value of a function according to clearly defined rules, and he argued that no human mathematician could possibly follow other clearly defined rules, which are not covered by his model. This model became known as the 'Turing machine.' Other mathematicians developed alternative models for how to effectively compute mathematical functions. However, it was proved that with all these different models, exactly the same class of functions was effectively computable.

2.1 The Turing Machine

A Turing machine has a finite number of internal states. A tape on which the Turing machine can read and write symbols from a finite alphabet via its read/write head is unlimited on both sides. The Turing machine starts its operation in a particular internal state and terminates its operation in another particular state. The read/write head will focus on the first square on the tape, where the input string (in consecutive squares) to its computation is provided. The Turing machine will execute a sequence of operations. Each of these operations depends on the current symbol under the read/write head and its internal state, and is specified in the corresponding Turing table. The operation may erase the old symbol on the tape in the square under the read/write head and replace it by another symbol. Furthermore, the Turing machine may change its internal state and move (optionally) its read/write head one square to the left or to the right. The entire computation terminates when the Turing machine reaches the special internal termination state. The result of the computation is found at a pre-specified location on the tape relative to the position of the read/write head. An example is shown in Fig. 1.

An important concept is the 'universal Turing machine,' which is a special Turing machine that is capable of simulating any other Turing machine, if a description of that machine is provided as input on the tape. The existence of the universal Turing machine demonstrates the relativity of the actual device on which the computation is performed. Generally speaking, a computing device can behave in quite different ways depending on what data are provided to it. In the case of a computationally universal computing device, all other computing devices can be simulated, if appropriate data, i.e., an appropriate program, is provided.

Today's computers are generally said to be computationally universal, although they are not really as a physical computer has limited memory (a tape of finite length). Nevertheless, it is commonly assumed that all natural computation can be simulated on an artificial computation device. Strictly speaking, however, there is not enough known about what 'programs' are actually 'executed' in natural computation processes. While today's scientists commonly assume that the human mind works deterministically, it should be noted that this is really just a working hypothesis for which there is no proof whatsoever. Opposed to that are computations in Turing's sense deterministic by definition.

3. *Artificial and Natural Computation in Neural Systems*

3.1 *Artificial Computation in Neural Systems*

Artificial computation in artificial neural networks is a metaphorical model of natural computation in biological neural systems, such as the human brain. Usually, large sets of computing units, connected in a network like manner, are used to simulate the way natural neural networks in humans and animals perform information processing. Often, the computing units are all of the same type and compute some sort of linear threshold function, i.e., they sum up the weighted input signals and compare the sum against a threshold value. Today, output functions of a sigmoidal shape or a simple step function are common. The simple model of a neuron's functionality using a step function is due to McCulloch and Pitts (1943). The unit computes internally the function:

$$f(i_1,\ldots, i_n) = \sum_{j=1}^{n} w_j i_j$$

where i_j is the jth input and w_j is the weight associated to the jth input. This internal function is used to calculate the output value by applying, either a step function or a differentiable function, e.g., a sigmoid function to it, such as:

$$out(i_1,\ldots, i_n) = \frac{1}{1+e^{-af(i_1,\ldots, i_n)}}$$

with a scaling parameter a. This maps the internal function value in the range of $[-\infty, \infty]$ (often actually a rather small range instead) to a value in the range [0, 1] strictly monotonically as shown in Fig. 2 (right). The weights for each individual unit usually vary and are often determined by an automatic adjustment process, also called 'training' or 'learning.' The first well-known learning algorithm for such neuronal models was developed by Rosenblatt (1962), who used a step function as opposed to the mentioned sigmoid output function.

Artificial neural networks are usually composed of a large number of such units, e.g., in engineering applications, the number of units typically ranges between as few as ten and as many as several thousand. The units are normally arranged in a layered fashion, similar to the structure in Fig. 2 (left), where the outputs of the units on one layer are fed as inputs to the units on the next layer. Such networks are also known as Multi-Layer Perceptrons (usually abbreviated as MLP since the hallmark publication of Rumelhart et al. (1986)). In combination with learning techniques which find appropriate weights for each unit, these networks have been successfully applied to

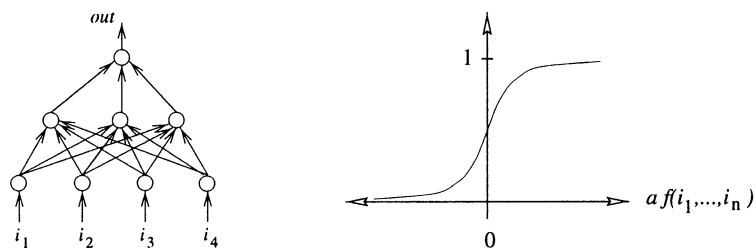

Figure 2
Left: A feed-forward neural network (multi-layer perceptron). Right: A sigmoid output function

a large number of engineering problems, such as general classification problems, recognition of visual, acoustic and other patterns, and adaptive control of nonlinear systems.

It must be noted that the computational models of cognitive systems are not at all accurate replications of what is known about their biological functionality. To contrast the artificial neural networks, the following presents some of the current knowledge about the physiological basis of computation in the human brain.

3.2 Natural Computation in the Human Brain

The human brain is made up of a vast network of neurons coupled with receptors and effectors. The input to the network is provided by millions of receptors while the output of the brain is transformed into effects on the body and the external world via effectors. The receptors continually monitor changes in the internal and external environment. Hundreds of thousands of cells control the effectors, such as muscles and the secretion of glands.

The human brain is made up of thousands of millions of neurons. A typical estimate is about 10^{10}–10^{11}, i.e., 10–100 billion neurons. Many neurons have between 1,000 and 10,000 connections to other neurons in the brain. There are thousands of different types of neurons in the brain. This includes neurons whose purpose has been identified as being specialized to the processing of sensory information. Those specialized neurons which deal directly with the signals coming from sensors, which are receptors for light, sound, touch, etc., are called 'sensor neurons.' Other neurons are responsible for controlling muscles and are, therefore, called 'motor neurons.'

The brain can hardly be considered as a simple stimulus-response chain from receptors to effectors. The vast network of billions and billions of neurons is rather interconnected in extremely complex and diverse ways, including simple loops as well as complex feedback cycles. Thus, signals coming from the receptors are mixed with signals already traveling in this vast network resulting in a change of the signals traveling inside the network as well as generating signals for controlling the effectors.

In this way, the brain manages to generate actions which depend on both the current stimulation as well as on some residue of past experiences. Past experiences may be reflected in current activities of the network, i.e., certain signals traveling inside the network, as well as in the current pattern and strength of connections.

It is estimated that there are more than 1,000 different types of neurons in the human brain. As a consequence of this large variety of neurons, it is impossible to speak of 'the typical neuron.' However, a model of a neuron is sketched in Fig. 3, which captures some of the basic features and functionality found in many different types of neurons.

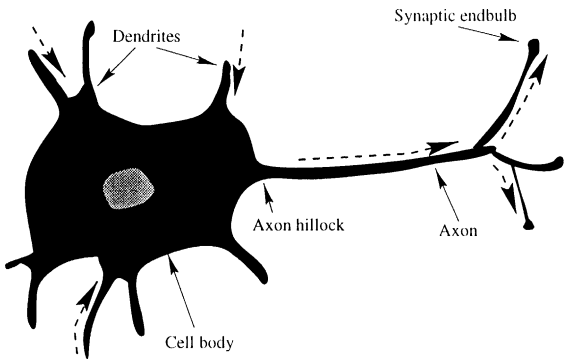

Figure 3
A schematic view of a neuron. The activity coming from receptors or from other neurons changes the membrane potentials on the dendrites and cell body (soma). The effects of these stimulations accumulate at the axon hillock where, for appropriate spatio-temporal patterns of stimulation, a threshold is reached which triggers a pulse of membrane potential to be propagated along the axon. This pulse is carried via the branching of the axon to the synaptic endbulbs, which in turn stimulate the membrane potential of the neurons or muscle fibers to which the synaptic endbulbs are connected. The dashed arrows indicate the direction of activity propagation through the axons

More than 40 properties of biological neurons are known to influence a neuron's information processing. One distinguishes three different parts in biological neurons: soma, axons, and dendrites (see Fig. 3). Biological neurons are connected with each other. Their connections are called 'synapses.' The axon may be considered as the output channel of the neuron which conveys its activation level to the synaptic connections with other neurons. The ends of the axons are called nerve terminals or 'endbulbs.' The dendrites on the other hand act as the neuron's input receptors for signals coming from other neurons. They convey the (electrical) input potential to the neuron's soma, i.e., to the neuron's body. The input potential in that case is also called 'post-synaptic potential.' The neuron's soma, in particular where it connects to the axon, i.e., in the 'axon hillock,' acts as an accumulator of input potentials and/or as an amplifier of these signals.

The tips of the branches of the axon impinge upon other neurons or upon other effectors. One says that the cell with the endbulb 'synapses upon' the cell with which the connection is made. In fact, the axonal branches of some neurons can have many 'endbulbs' that allow them to make synapses along their lengths as well as their ends.

3.3 Dynamics of Biological Neurons

Neural activations are mostly stimulated circularly. A neuron is activated by other neurons to which it is connected. In turn, its own activation stimulates other connected neurons to activation.

If an impulse is started at any one place on the axon, it propagates in both directions. However, if the pulse is started at one end of the axon (normally the axon hillock), it can only travel away form that end, since once a section has been triggered it becomes refractory until well after the impulse has passed out of range. An impulse traveling along the axon triggers new impulses in each of its branches, which in turn trigger impulses in their even finer branches. Axons come in two kinds: myelinated and unmyelinated. A myelinated fiber is wrapped in a sheath of myeline, a sequence of Schwann cells wrapped tightly around an axon with small gaps, called nodes of Ranvier, between adjacent segments. Instead of the somewhat slow active propagation down an unmyelinated fiber, the nerve impulse in a myelinated fiber jumps from node to node, thus speeding passage and reducing energy requirements. When an impulse arrives at one of the endbulbs, after a slight delay it yields a change in potential difference across the membrane of the cell upon which it impinges. The membrane of the endbulb is called the 'pre-synaptic membrane,' and the membrane of the surface upon which the endbulb impinges is called the 'post-synaptic membrane.' At most synapses the direct cause of the change in (electrical) potential of the post-synaptic membrane is not electrical but chemical. There are some exceptional cell appositions that are so large, or have so tight coupling, that the impulse affects the polarization of the post-synaptic membrane without chemical mediation. However, the normal process is that the electrical pulse that reaches the endbulb causes the release of 'transmitter' molecules from little packets called 'vesicles' through the pre-synaptic membrane. The transmitter then diffuses across the very small 'synaptic cleft' to the other side. When the transmitter reaches the post-synaptic membrane, it causes a change in the polarization of this membrane. The effect of the transmitter may be of two basic kinds: either 'excitatory,' tending to move the potential difference across the post-synaptic membrane in the direction of the threshold (depolarizing the membrane), or 'inhibitory', tending to move the polarization away from threshold (hyperpolarizing the membrane).

The many small changes in potential differences across the membrane of a neuron, due to the activity of many or all the synapses which impinge upon it, propagate—in most cases—passively through the membrane of its dendrites and its cell body. They decay as they propagate. At the axon hillock, many different and often small potential changes converge, so that the total contribution by which excitation exceeds inhibition may be quite large. In other words, the transmitter released when the impulse of a single synapse arrives at the endbulb, and generally causes a change in the postsynaptic membrane, is not usually large enough to trigger the neuron. Nonetheless, the cooperative effect of many such subthreshold changes may yield a potential change at the axon hillock that exceeds threshold, and if this occurs at a time when the axon has passed the refractory period of its previous firing, then a new impulse will be fired down the axon. See e.g., Arbib 1989, Arbib et al. 1997, Bower 2000, or the Journals *Biological Cybernetics*, *Journal of Neurobiology*, and others (see also *Connectionist Approaches*; *Perceptrons*)

From the foregoing it is clear that the current artificial neural networks are only very rough approximations of the natural neural systems. Furthermore, it has been proved that certain types of artificial neural networks (cellular automata) are even computationally universal. This, however, also implies that nothing has been gained from copying aspects of the human brain with respect to the function the cellular automata will compute. The function depends completely on the 'activity levels of the individual cells.' That is, with suitable initial activity patterns the system can compute any function one chooses.

4. Evolutionary Computation

Evolutionary computation is another field, that is strongly inspired by nature (see *Artificial Intelligence: Genetic Programming*). This field was pioneered independently in the 1960s by Fogel et al. 1966, Holland 1975, Rechenberg 1973. The latter two authors published their work in a widely accessible form only in the 1970s. Rechenberg used evolutionary strategies to develop highly optimized devices, such as irregularly shaped reduction pieces for pipes, e.g., for an air conditioning system, which proved to have a lower air flow resistance than ordinary reduction pieces.

In evolutionary computation, the process of natural evolution is used as a role model for a strategy for finding optimal or near optimal solutions for a given problem. In genetic algorithms, an important class of evolutionary computing techniques, candidates for a solution are encoded in a string, often a binary string only containing '0's and '1's. Evolution takes place by modifying the genetic code of a candidate.

Evolutionary techniques are generally applied to optimization problems. Many of those problems are combinatorial optimization problems, which are computationally hard (NP-hard) (see also *Algorithmic Complexity*).

This roughly means that programs are expected to require a computing time that grows exponentially with the size of the problem. One such problem is the traveling salesperson problem, where a number of

cities are given along with a distance matrix providing the traveling distance or cost for each pair of cities. The problem asks for the shortest possible round trip, i.e., the traveling salesperson departs from a given city and should visit all other cities before returning to the first city. For n cities, apart from the city of origin, there are $n!$ different round trips, provided there is a direct connection between every pair of cities. A substantial proportion of them has to be evaluated before one can be sure that the optimal round trip has been found. Hence, the number of computing steps grows exponentially with the number of cities to be visited. That is, the required computing time to find the optimal solution becomes impractical even for moderate numbers of cities on today's fast computers and those computers projected for the near future. As a consequence, one has to be content with a good solution as opposed to an optimal solution.

The traveling salesperson problem can be handled by a genetic algorithm according to the following sketch: a population of potential solutions (more or less randomly generated orderings of the cities) is created initially as the first generation in the evolutionary process. The fitness of each individual for a solution to the posed problem is assessed, i.e., the traveling distance is calculated. The next generation is created by retaining the better suited individuals in the population, i.e., those with smaller traveling distances. The less fit individuals are discarded. Furthermore, new individuals are created as 'offsprings' from the current population and added to the new generation. Offspring individuals are usually created by random mutations of the 'parents.' This may also include the combination of the genetic code of multiple parents, such as 'cross over' operators on the genetic code of two parent individuals.

From a more abstract point of view, such evolutionary approaches fall into the category of probabilistic algorithms. Other probabilistic algorithms, such as simulated annealing (which constitute another computer model of a natural, physical process), share the following common feature with the sketched genetic algorithms: a solution to the given optimization problem is developed by iteratively modifying one or more candidate solutions until a satisfactory solution is found. The iterative modification is based on some random decisions. Usually, a fitness measure is used to decide whether a new candidate solution is deemed any closer to an acceptable solution than previously generated candidates. Consequently, if a new candidate appears closer to an acceptable solution, the new candidate will normally be used to derive another modification. This is also called 'hill-climbing' in a search space. However, since no exact measure for the closeness of a candidate to an acceptable solution can generally be expected, a candidate that appears less close to an acceptable solution may still be selected for modification, but with less probability.

While the evolutionary strategies appear to be universally applicable, experience showed that in most cases problem-specific adaptations of the evolutionary scheme need to be made in order to allow the evolutionary process to generate a good solution in reasonable computing time. Such adaptations may consist of finding a suitable genetic encoding of a solution candidate, determining an effective size of the population, suitable probability parameters for generating certain mutations, etc. Another application of the sketched genetic techniques is the development of programs, as is done in the field of 'genetic programming' (see Koza et al. 1999 for more details). (See also *Artificial Intelligence: Genetic Programming*.) For more details on evolutionary computation, see e.g., Bäck et al. 1997.

5. *Summary and Prospects*

The presented artificial computing models of natural processes are the most studied models in recent years. They are models of biological neural systems and models of evolutionary processes in biology.

While nature inspired substantial research on these models, they are rather metaphorical models of the natural processes. Many known details of the natural processes are missing in these models. The models are not accurate and they are not meant to be accurate. Once simple models were developed, they were studied for their use in particular applications, such as solving certain combinatorial optimization problems or learning to classify images of characters. Hence, from an engineering perspective it is less important whether the models represent the natural processes authentically, as long as the models are useful for solving practical problems.

In regard to artificial neural networks, it is certainly desirable to enhance their capabilities, so that they come closer to the performance of the human brain. Given the substantial gap between the artificial models and what is known about the brain's physiology, an accurate computational model of the brain's physiology is not in sight. While it is unknown to what extent the brain's physiology needs to be modeled in order to produce a behavior, which would be comparable to a human, it seems likely that not all aspects of the physiology need to be modeled. Which aspects are necessary and sufficient is unknown at this stage.

The artificial evolutionary computation models are not accurate models of the biological processes either. However, their application in engineering and elsewhere are clearly directed towards artificial domains, i.e., engineering problems. Hence, one would not expect that the evolutionary computation models need to be accurate models of biological evolution.

However, it remains an open question how far the study of such metaphorical models of nature will lead? Perhaps there exist more suitable approaches which have no resemblance with nature. After all, there are

plenty of examples where excellent engineering solutions are very different from what is found in nature. For example, the invention of the wheel or the steam engine. On the other hand, there are also examples where a partial, metaphorical model of nature's solution to a problem can be supplemented with an artificial engineering approach. For example, the development of airplanes, which do not move their wings, but have rather 'artificial' propulsion mechanisms. See also Hoffmann (1998) for a discussion of the requirements of a computing model for being used as an engineering approach to the design of complex systems.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence: Genetic Programming; Artificial Intelligence in Cognitive Science; Artificial Neural Networks: Neurocomputation; Brain, Evolution of; Computational Neuroscience; Deductive Reasoning Systems; Intelligence: History of the Concept; Linear Algebra for Neural Networks; Neural Networks and Related Statistical Latent Variable Models; Neural Networks, Statistical Physics of; Scientific Discovery, Computational Models of

Bibliography

Adleman L 1994 Molecular computation of solutions to combinatorial problems. *Science* **266**: 1021–4

Arbib M A 1989 *The Metaphorical Brain II*. Wiley, New York

Arbib M A, Erdi P, Szentagothai J (eds.) 1997 *Neural Organization: Structure, Function, and Dynamics*. MIT Press, Cambridge, MA

Bäck T, Fogel D B, Michalewicz Z (eds.) 1997 *Handbook of Evolutionary Computation*. Oxford University Press, Oxford, UK

Bower J M (ed.) 2000 *Computational Neuroscience: Trends in Research, 1997*. Plenum, New York

Fogel L J, Owens A J, Walsh M J (eds.) 1966 *Artificial Intelligence Through Simulated Adaptation*. Wiley, New York

Hoffmann A 1998 *Paradigms of Artificial Intelligence: A Methodological and Computational Analysis*. Springer-Verlag, Heidelberg, Germany

Holland J H 1975 *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI

Koza J R, Bennett F H, Andre D, Keane M A (eds.) 1999 *Genetic Programming iii: Darwinian Invention and Problem Solving*. Morgan Kaufmann Publishers, San Francisco, CA

McCulloch W S, Pitts W 1943 A logical calculus for the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–43

Rechenberg I 1973 *Evolutionsstrategie: Optimierung technischer Systeme und Prinzipien der biologischen Evolution*. Frommann-Holzboog, Stuttgart, Germany

Rosenblatt F 1962 *Principles of Neurodynamics*. Spartan, Chicago

Rumelhart D E, McClelland J L, PDP Research Group 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, i & ii*. MIT Press, Cambridge, MA

Turing A M 1937 On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* **2**(42): 230–65, (43): 544–6

Wiener N 1948 *Cybernetics*. Wiley, New York

A. Hoffmann

Artificial Intelligence: Connectionist and Symbolic Approaches

Perhaps the most significant feature of current artificial intelligence (AI) research is the co-existence of a number of vastly different and often seriously conflicting paradigms, competing for the attention of the research community (as well as research funding).

In this article, two competing paradigms of artificial intelligence, the connectionist and the symbolic approach, will lie described. Brief analysis and articism of each paradigm will be provided, and possible integration of the two will their respective shortcomings.

1. The Two Paradigms

The two main competing paradigms in artificial intelligence can be summarized as follows:

(a) The traditional symbolic paradigm (Newell and Simon 1976). The field of AI, since its inception, has been conceived mainly as the development of models using symbol manipulation. The computation in such models is based on explicit representations that contain symbols organized in some specific ways and aggregate information is explicitly represented with aggregate structures that are constructed from constituent symbols and syntactic combinations of these symbols.

(b) The more recently established connectionist paradigm (Rumelhart and McClelland 1986, Smolensky 1988). The emergence of the connectionist paradigm largely resulted from various dissatisfactions with symbol manipulation models, especially in their inability to handle flexible and robust processing in an efficient manner. The connectionist paradigm aims at massively parallel models that consist of a large number of simple and uniform processing elements interconnected with extensive links, that is, artificial neural networks and their various generalizations. In many connectionist models, representations are distributed throughout a large number of processing elements (in correspondence with the structure of such models). Sometimes the constituent

symbolic structures of aggregate information are embedded in a network and difficult to identify. Due to their massively parallel nature, such models are good at flexible and robust processing, and show promise at dealing with some tasks that have been difficult for the symbolic paradigm.

It should be emphasized that up until now, no single existing paradigm can fully address all the major AI problems. Each paradigm has its strengths and weaknesses, and excels at certain tasks while falls short at some others. This situation, in a way, indicates the need to integrate these existing paradigms somehow (see Sect. 4).

2. *Symbolic AI*

The *physical symbol system hypothesis* introduced by Newell and Simon (1976) clearly articulated the tenets of symbolic AI. They defined a *physical symbol system* as follows:

> A physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (symbol structure). Thus a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another).

They further claimed that symbols can designate arbitrarily: 'a symbol may be used to designate any expression whatsoever'; 'it is not prescribed *a priori* what expressions it can designate.' 'There exist processes for creating any expression and for modifying any expression in arbitrary ways.' Based on that, they concluded: 'A physical symbol system has the necessary and sufficient means for general intelligent action,' which is the (famed) *physical symbol system hypothesis*.

The physical symbol system hypothesis has spawned (and was used to justify) enormous research effort in traditional AI (and in cognitive science). This approach (classical symbolism) typically uses discrete symbols as primitives and performs symbol manipulation in a sequential and deliberative manner.

2.1 *Representation and Search*

Two fundamental ideas that originated in the earliest days of AI are search and representation,which have played central roles in symbolic AI.

Let us discuss the idea of search space first. In any problem (to be tackled by AI), there is supposed to be a space of *states* each of which describes a step in problem solving (inference). *Operators* can be applied to reach a new state from a current state. Search techniques adopted in the early days of AI include depth-first search and breadth-first search. In depth-first search, from the current state, the system examines one alternative (a 'path') at a time by applying one of the operators to the current state, which leads to a new state. Then from the new state, the same process is repeated, until it reaches a goal state or hits a dead-end where there is no operator that can be applied. In the latter case, the system backs up to a previous state and tries a different alternative. In breadth-first search, however, the system examines all alternatives ('paths') at once by applying each of all the applicable operators. Such exhaustive search techniques are inefficient. To speed up the search, many heuristic search algorithms have been proposed. The idea of search space has been applied in all the areas of AI, including in problem solving, natural language processing, robotics and vision, knowledge representation/reasoning, and machine learning (see Russell and Norvig 1995 for more details regarding search).

Another important idea is representation. It embodies the belief that knowledge should be expressed in an internal form that facilitates its use, corresponding to the requirement of the task to be handled and mirroring the external world (in some way). A variety of symbolic representational forms have been developed over the years in AI, and most of them are used in conjunction with search algorithms for inference. One of the earliest representational forms involves rule-based reasoning, in which discrete rules are used to direct search (inference). Rules are composed of both conditions, which specify the applicability of a rule, and conclusions, which specify actions or outcomes. Rules are modular: ideally each rule can be added to or deleted from a system, without affecting the other parts of the system (modularity may, however, inadvertently hamper computational flexibility and dynamic interaction in reasoning).

A popular form of rule-based reasoning is the production system, which evolved from some psychological theories that emerged in the 1960s and 1970s. A production system consists of (a) a production rule base (for storing all the rules), (b) a working memory (for storing initial, intermediate, and final results), and (c) a control structure for coordinating the running of the system. The inference process in a production system can be either forward chaining or backward chaining.

Formal logics constitute an alternative approach in rule-based reasoning, as advocated chiefly by John McCarthy (McCarthy 1968). They are relatively simple, formally defined languages capable of expressing rules in a rigorous way. Logic inference is performed in formally defined ways that guarantee the completeness and soundness of the conclusions, which can be carried out using a variety of algorithms beyond simple forward and backward chaining. Formal logics (and most production systems) are restrictive: one needs to have all conditions precisely specified in order to perform one step of inference. Thus, they are unable to deal with partial, incomplete, or approximate

information. There is also no intrinsic way for handling reasoning involving similarity-based processes (Sun 1994).

Another type of representation aims to capture the aggregate *structures* of knowledge (instead of dispersing such structures as in production systems). Knowledge can be organized in structured chunks, each of which is centered around a particular entity, and each can contain, or be contained in, other chunks. Each chunk contains all the pieces of information regarding a certain entity as well as their interrelations. For instance, a *frame* (as proposed by Marvin Minsky) represents a concept in terms of its various attributes (slots), each of which has a name (label) and a value. By organizing knowledge in frames, all the relevant pieces of information can be accessed in a structured way. A *semantic network* (as proposed by Ross Quillian) consists of a set of nodes, each of which represents a particular (primitive) concept, and labeled links, each of which represents a particular relation between nodes. Semantic networks also allow efficient and effective access of knowledge, albeit in a different way, by following links that go from one concept to all the others that have a particular relation to the original one, through, for example, 'spreading activation.' *Scripts* (proposed by Roger Schank and associates) are used for representing prototypical event sequences in stereotypical situations, for example, eating in a restaurant. In such a situation, there is an almost fixed sequence. Scripts help with the efficient recognition and handling of these sequences. The shortcomings of the above types of representations include: (a) structures often need to be determined *a priori* and hand coded, (b) they are usually fixed and cannot be changed dynamically, and (c) they can become too costly and unwieldy to capture the full extent of complex real-world situations.

Some more recent developments in the symbolic paradigm aim to remedy some of the problems with traditional symbolic representation, especially those of logic-based approaches, as discussed earlier. These extensions include Default Logic and Circumscription, among many others. For example, *Default Logic* is proposed to model default reasoning: it deals with beliefs based on incomplete information, which might be modified or rejected later on based on subsequent observations. These logics have some shortcomings, including the lack of capabilities for dealing with (a) approximate information, (b) inconsistency, and most of all, (c) reasoning as a complex, interacting process (Sun 1994). Thus, they are somewhat deficient from the standpoint of capturing human reasoning. For an overview of these models, see Davis (1990).

In a totally different vein, Zadeh proposed *fuzzy logic* (see Zadeh 1988), primarily to capture vagueness or approximate information in linguistic expressions, such as 'tall' or 'warm,' which have no clear-cut boundaries. The basic idea is as follows: for each concept, a set of objects satisfying that concept to a certain degree form a subset, namely a fuzzy subset. This fuzzy (sub)set contains as its elements pairs consisting of an object and its grade of membership, which represents the degree to which it satisfies the concept associated with the set. A fuzzy logic can therefore be constructed with which one can reason about the fuzzy truth values of concepts.

The *probabilistic approach* (especially the Bayesian network approach) treats beliefs as probabilistic events and utilizes probabilistic laws for belief combinations (Pearl 1988). Note, however, that human reasoning may not always conform to the assumptions and the laws of probability theory, probably in part because of the complexity of the formal models. In addition, it is not always possible to obtain adequate probability measures in practice.

2.2 Symbolic Learning

Learning is a major issue in AI, in Cognitive Science, and in other related areas. While very sophisticated representations have been developed in symbolic AI, learning is, comparatively speaking, difficult for symbolic AI. This is in part due to the fact that, from its inception, symbolic AI has been centered on representation, not learning.

Nevertheless, there have been some developments of symbolic learning, especially since the late 1980s (see Shavlik and Dietterich 1990). The majority of work in symbolic machine learning focuses on 'batch learning.' Typically, the learner is given all the exemplars/instances, positive and/or negative, before learning starts. Most of these algorithms handle the learning of concepts with simple rules or decision trees (Quinlan 1986). They typically involve (a) a division of all the instances/exemplars into mutually exclusive or overlapping classes (i.e., clustering), and/or (b) the induction of a set of classification rules describing the makeup of a concept (Michalski 1983, Quinlan 1986).

More recently, some symbolic learning algorithms have been extended to deal with noisy or inconsistent data and to deal with incremental learning. In addition, inductive logic programming, tries to induce more powerful first-order rules from data. Recently, there have also been some reinforcement learning algorithms that handle dynamic sequences, which necessarily involve temporal credit assignment (i.e., attributing properly a success/failure to some preceding steps). (See Shavlik and Dietterich 1990.)

2.3 Criticisms

Although the symbolic paradigm dominated AI and Cognitive Science for a long while, it has been steadily receiving criticisms from various sources (e.g., from Rodney Brooks, John Searle, and Hubert Dreyfus; see Dreyfus and Dreyfus 1986). The critics focused largely

on the disembodied abstractness of this approach, especially in relation to interacting with the world (Dreyfus and Dreyfus 1986). Some, such as John Searle, emphasized the importance of biological substrates of intelligence. Some other more specific criticisms have been identified earlier.

3. Connectionist AI

In the 1980s, the publication of the PDP book (Rumelhart and McClelland 1986) started the so-called 'connectionist revolution' in AI and cognitive science. The basic idea of using a large network of extremely simple units for tackling complex computation seemed completely antithetical to the tenets of symbolic AI and has met both enthusiastic support (from those disenchanted by traditional symbolic AI) and acrimonious attacks (from those who firmly believed in the symbolic AI agenda). Even today, we can still feel, to some extent, the divide between connectionist AI and symbolic AI, although hybrids of the two paradigms and other alternatives have flourished. However, much of the controversy was the result of misunderstanding, overstatement, and terminological differences.

Connectionist models are believed to be a step in the direction toward capturing the intrinsic properties of the biological substrate of intelligence, in that they have been inspired by biological neural networks and seem to be closer in form to biological processes. They are capable of dealing with incomplete, approximate, and inconsistent information as well as generalization.

3.1 Connectionist Learning

Connectionist models excel at learning: unlike the formulation of symbolic AI which focused on representation, the very foundation of connectionist models has always been learning. Learning in connectionist models generally involve the tuning of weights or other parameters in a large network of units, so that complex computations can be accomplished through activation propagation through these weights (although there have been other types of learning algorithms, such as constructive learning and weightless learning). The tuning usually is based on gradient descent or its approximations. The best known of such learning algorithms is the backpropagation algorithm (Rumelhart and McClelland 1986).

In terms of task types tackled, connectionist learning algorithms have been devised for (a) supervised learning, similar in scope to aforementioned symbolic learning algorithms for classification rules but resulting in a trained network instead of a set of classification rules; (b) unsupervised learning, similar in scope to symbolic clustering algorithms, but without the use of explicit rules; (c) reinforcement learning, either implementing symbolic methods or adopting uniquely connectionist ones.

Connectionist learning has been applied to learning some limited forms of symbolic knowledge. For example, Pollack (1990) used the standard backpropagation algorithm to learn tree structures, through repeated applications of backpropagation at different branching points of a tree, in an auto-associative manner (named which was auto-associative memory, or RAAM). Since trees are a common symbolic form, this approach is widely applicable in learning symbolic structures. Similarly, Giles and co-workers (see, e.g., Giles and Gori 1998) used backpropagation for learning finite-state automata, another common symbolic structure. Connectionist learning algorithms combine the advantages of their symbolic counterparts with the connectionist characteristics of being noise/fault tolerant and being capable of generalization. For an overview of both symbolic and connectionist learning, see Shavlik and Dietterich (1990).

3.2 Connectionist Knowledge Representation

Although it is relatively difficult to devise sophisticated representations in connectionist models (compared with symbolic models), there have been significant developments of connectionist knowledge representation. Many so-called 'high-level' connectionist models have been proposed that employ representation methods that are comparable with, and sometimes even surpass, symbolic representations, and they remedy some problems of traditional representation methods as mentioned earlier. For an overview of connectionist knowledge representation, see Sun and Bookman (1995).

Let us look into some of these developments in detail. First of all, logics and rules can be implemented in connectionist models in a variety of ways. For example, in one type of connectionist system, inference is carried out by constraint satisfaction through minimizing an error function. The process is extremely slow though. Another type of system, as proposed by Shastri and many others in the early 1990s, uses more direct means by representing rules with links that directly connect nodes representing conditions and conclusions, respectively, and inference in these models amounts to activation propagation. They are thus more efficient. They also deal with the so-called variable binding problem in connectionist networks. Those advanced logics as mentioned earlier that go beyond classical logic can also be incorporated into connectionist models (see, e.g., Sun 1994).

Aggregate information can also be incorporated into connectionist models. A system developed by Miikkulainen and Dyer (1991) encodes scripts through dividing input units of a backpropagation network

into segments each of which encodes an aspect of a script in a distributed fashion. The system is capable of dealing with incomplete (missing) information, inconsistent information, and uncertainty. There are also localist alternatives (such as those proposed by Lange and Dyer in 1989 and by Sun in 1992), in which a separate unit is allocated to encode an aspect of a frame.

Search, the main means of utilizing knowledge in a representation, is employed or embedded in connectionist models. Either an explicit search can be conducted through a settling or energy minimization process (as discussed earlier), or an implicit search can be conducted in a massively parallel and local fashion. Symbolic search requires global data retrieval and is thus very costly in terms of time. Global energy minimization (as in some connectionist models) is also time consuming. Local computation in connectionist models is a viable alternative. Knowledge is stored in a network connected by links that capture search steps (inferences) directly. Search amounts to activation propagation (by following links, similar to semantic networks in a way), without global control, monitoring, or storage.

The advantage of connectionist knowledge representation is that such representation can not only handle symbolic structures but goes beyond them by dealing with incompleteness, inconsistency, uncertainty, approximate information, and partial match (similarity) and by treating reasoning as a complex dynamic process. However, developing representation in highly structured media such as connectionist networks is inherently difficult.

4. *Hybrid Models*

Given the different emphases and strengths of connectionist and symbolic systems, it seems plausible that combining them would be a promising avenue for developing more robust, more powerful, and more versatile systems. The need for such systems has been slowly but steadily growing. There has been a great deal of research recently, leading to so-called hybrid systems.

The relative advantages of connectionist and symbolic models have been amply argued—see, e.g., Waltz and Feldman (1988), Smolensky (1988), and Sun (1994). The computational advantage of the combination thereof is thus relatively easy to justify. Naturally, we want to take advantages of both types of models and their synergy (see, e.g., Dreyfus and Dreyfus (1986), Sun and Bookman (1995), and Sun and Alexandre (1997) for further justifications).

There are many important issues to be addressed in developing hybrid connectionist-symbolic systems. These issues concern architectures, learning, and various other aspects. First, hybrid models likely involve a variety of different types of processes and representations. Multiple heterogeneous mechanisms interact in complex ways. We need to consider ways of structuring these different components; in other words, we need to consider *architectures*, which thus occupy a more prominent place in this line of research. Second, although purely connectionist models, which constitute part of any hybrid system, are known to excel in their learning abilities, hybridization makes it more difficult to perform learning. In a way, hybrid systems inherit the difficulty with learning from the symbolic side and mitigate to some extent the advantage that (purely) connectionist models have in terms of learning.

4.1 *Architectures and Representations*

We divide systems into two broad categories: *single-module* and *multimodule* architectures. Among single-module systems, along the representation dimension, there can be the following types (Sun and Bookman 1995): symbolic, localist (with one distinct node for representing each concept; see, e.g., Lange and Dyer 1989, Shastri and Ajjanagadde 1993) and distributed (with a set of nonexclusive, overlapping nodes for representing each concept; see, e.g., Pollack 1990). Among multimodule systems we can distinguish between *homogeneous* and *heterogeneous* systems. Homogeneous systems are similar to single-module systems, except they contains several replicated copies of the same structure, each of which can be used for processing the same set of inputs, to provide redundancy for various reasons.

Heterogeneous multimodule systems are more interesting. This category constitutes the true hybrid systems. As an example, CONSYDERR (Sun 1994) belongs to this category. It consists of two levels: the top level is a network with localist (symbolic) representation, and the bottom level is a network with distributed representation; concepts and rules are represented diffusely in the bottom level by sets of feature units overlapping each other. This is a similarity-based representation, in which concepts are 'defined' in terms of their similarity to other concepts in these representations. The localist network is linked with the distributed network by connecting each node in the top level representing one concept to all the feature nodes in the bottom level representing the same concept. Through a 3-phase interaction between the two levels, the model is capable of both rule-based and similarity-based reasoning with incomplete, inconsistent, and approximate information, and accounts for a large variety of seemingly disparate patterns in human reasoning data.

A variety of distinctions can be made here. First of all, a distinction can be made in terms of representations of constituent modules. In heterogeneous

multimodule systems, there can be different combinations of different types of constituent modules, for example, a system can be a combination of localist modules and distributed modules (as in CONSYDERR discussed above) or it can be a combination of symbolic modules and connectionist modules (either localist or distributed, e.g., as in SCRUFFY described by Hendler in Barnden and Pollack (1991)). Some of these combinations can be traced to the ideas of Smolensky (1988), who argued for the dichotomy of conceptual and subconceptual processing.

Another distinction which can be made among heterogeneous multimodule systems is in terms of the *coupling* of modules: a set of modules can be either loosely coupled or tightly coupled. In loosely coupled situations, modules communicate with each other, primarily through message passing, shared memory locations, or shared files, as in, for example, SCRUFFY (see Hendler's chapter in Barnden and Pollack 1991). Such loose coupling enables some loose forms of cooperation among modules. One form of cooperation is in terms of pre/postprocessing vs. main processing: while one or more modules take care of pre/postprocessing, such as transforming input data or rectifying output data, a main module focuses on the main part of the task. Another form of cooperation is through a master-salve relationship: while one module maintains control of the task at hand, it can signal other modules to handle some specific aspects of the task. For example, a symbolic expert system, as part of a rule, may invoke a neural network to perform a specific classification or decision making. Yet another form of cooperation is the equal partnership of multiple modules. In this form, the modules (the equal partners) can consist of (a) complementary processes, such as in SOAR/ECHO (Johnson et al. in Sun and Alexandre 1997), or (b) multiple functionally equivalent but structurally and representationally different processes, such as in CLARION (Sun and Peterson 1998), or (c) they may consist of multiple differentially specialized and heterogeneously represented experts each of which constitutes an equal partner in accomplishing a task.

In tightly coupled systems, on the other hand, the constituent modules interact through multiple channels (e.g., various possible function calls), or may even have node-to-node connections across two modules, such as CONSYDERR (Sun 1994) in which each node in one module is connected to a corresponding node in the other module. There are a variety of forms of cooperation among modules, in ways quite similar to loosely coupled systems.

4.2 Learning

Learning, which can include (a) learning the content (knowledge) in a hybrid model or (b) learning and developing the model architecture itself, is a fundamental issue that is clearly difficult. However, learning is indispensable if hybrid systems are ever to be scaled up. Over the years, some progress on learning has been made. While some researches have tried to extend connectionist learning algorithms to learn complex symbolic representations, others have instead incorporated symbolic learning methods. For example, Sun and Peterson (1998) presented a two-module model CLARION for learning sequential decision tasks, in which symbolic knowledge is extracted on-line from a reinforcement learning connectionist network and is used, in turn, to speed up connectionist learning and to facilitate transfer. The work showed not only the synergy between connectionist and symbolic learning, but also that symbolic knowledge can be learned autonomously on-line, from subsymbolic knowledge, which is very useful in developing autonomous agents.

Learning methods that may be applied to hybrid systems include gradient descent and its many variations (extending typical connectionist learning algorithms), Expectation-Maximization and its many instantiations (including hidden Markov model algorithms), search algorithms, evolutionary algorithms, and heuristic methods (such as decision trees or rule induction; see Shavlik and Dietterich 1990). Some of these methods may be combined with others (as in Sun and Peterson 1998), which likely results in improved learning. There are a variety of other learning approaches being proposed also, including many rule extraction or insertion algorithms.

There is a sense that future advance in this area is dependent on progress in the development of new learning methods for hybrid systems and the integration of learning and complex symbolic representations. As mentioned above (see Sun and Peterson 1998), symbolic representation and reasoning may well emerge from subsymbolic processes through learning, and thus an intimate and synergistic combination of symbolic and subsymbolic learning processes should be pursued.

4.3. Discussion

Despite the diversity of the approaches discussed thus far, there is an underlying common theme: bringing together symbolic and connectionist models to achieve a synthesis of the two seemingly radically different paradigms.

There is no doubt that we need to invest substantial efforts in comparing and analyzing various paradigms, especially the symbolic paradigm and the connectionist paradigm, to reveal their respective strengths and limitations, as well as underlying assumptions, which can lead to better understanding and more rapid advancement of this field. Although significant advances have been achieved so far, this process is still on-going today, and it appears that it will not be concluded until a better understanding is achieved of the nature of intelligence in computational terms.

In the meantime, it is an appealing idea that representation and learning techniques from both symbolic processing models and connectionist network models shall be brought together to tackle problems that neither type of model alone can apparently handle very well.

One such problem is the modeling of human cognition, which requires dealing with a variety of cognitive capacities. Several researchers (e.g., Smolensky 1988, Dreyfus and Dreyfus 1986) have consistently argued that cognition is multifaceted and better captured with a combination of symbolic and connectionist processes. Many methods and frameworks reviewed above share the belief that connectionist and symbolic methods can be usefully integrated, and such integration may lead to significant advances in our understanding of cognition. It thus appears that hybridization is a theoretically sound approach, in addition to being a practically expedient approach.

See also: Artificial and Natural Computation; Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence in Cognitive Science; Artificial Neural Networks: Neurocomputation; Connectionist Approaches; Connectionist Models of Concept Learning; Connectionist Models of Language Processing; Intelligence: History of the Concept; Knowledge Representation; Scientific Discovery, Computational Models of

Bibliography

Barnden J A, Pollack J B (eds.) 1991 *Advances in Connectionist and Neural Computation Theory*. Ablex Publishing Co., Norwood, NJ

Davis E 1990 *Representations of Commonsense Knowledge*. Morgan Kaufmann, San Mateo, CA

Dreyfus H L, Dreyfus S E 1986 *Mind Over Machine*. Free Press, New York

Giles C L, Gori M 1998 *Adaptive Processing of Sequences and Data Structures*. Springer, New York

Lange T, Dyer M 1989 High-level inferencing in a connectionist network. *Connection Science* **1**: 181–217

McCarthy J L 1968 Programs with common sense. In: Minsky M (ed.) *Semantic Information Processing*. MIT Press, Cambridge, MA

Rumelhart D E, McClelland J L, PDP Research Group 1986 *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*. MIT Press, Cambridge, MA

Michalski R S 1983 A theory and methodology of inductive learning. *Artificial Intelligence* **20**: 111–61

Miikkulainen R, Dyer M 1991 Natural language processing with modular PDP networks and distributed lexicons. *Cognitive Science* **15**(3): 343–99

Newell A, Simon H 1976 Computer science as empirical inquiry: Symbols and search. *Communication of ACM* **19**: 113–26

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA

Pollack J B 1990 Recursive distributed representations. *Artificial Intelligence* **46**(1–2): 77–106

Quillian M 1968 Semantic memory. In: Minsky M (ed.) *Semantic Information Processing*. MIT Press, Cambridge, MA, pp. 216–60

Quinlan R 1986 Inductive learning of decision trees. *Machine Learning* **1**: 81–106

Russell S J, Norvig P 1995 *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ

Shavlik J W, Dietterich T G 1990 *Reading in Machine Learning*. Morgan Kaufmann, San Mateo, CA

Shastri L, Ajjanagadde V 1993 From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences* **16**(3): 417–51

Smolensky P 1988 On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**(1): 1–23

Sun R 1994 *Integrating Rules and Connectionism for Robust Commonsense Reasoning*. Wiley, New York

Sun R, Alexandre F (eds.) 1997 *Connectionist Symbolic Integration*. Erlbaum, Mahwah, NJ

Sun R, Bookman L A (eds.) 1995 *Computational Architectures Integrating Neural and Symbolic Processes*. Kluwer, Boston

Sun R, Peterson T 1998 Autonomous learning of sequential tasks: Experiments and analyzes. *IEEE Transactions on Neural Networks* **9**(6): 1217–34

Turing A 1950 Computing machinery and intelligence. *Mind* **59**: 433–60

Waltz D, Feldman J A (eds.) 1988 *Connectionist Models and Their Implications*. Ablex, Norwood, NJ

Zadeh L A 1988 Fuzzy logic. *Computer* **21**(4): 83–93

R. Sun

Artificial Intelligence: Genetic Programming

1. Introduction

The term genetic programming (GP) describes a research area within artificial intelligence (AI) that deals with the evolution of computer code. The term evolution refers to an artificial process analogous to natural evolution of living organisms, but which has been abstracted and stripped of most of its intricate details. The resultant algorithms then yield approximate solutions to problems in machine learning or induce precise solutions in the form of grammatically correct (language) structures for the automatic programming of computers.

Genetic programming is part of the growing set of evolutionary algorithms which apply search principles analogous to those of natural evolution in a variety of different problem domains, notably parameter optimization. Evolutionary programming, evolutionary strategies, and genetic algorithms are three other

branches of the area of evolutionary algorithms which mostly find applications as optimization techniques.

All evolutionary algorithms follow Darwin's principle of differential natural selection. This principle states that the following preconditions must be fulfilled for evolution to occur via (natural) selection:

There are entities called individuals which form a population. These entities can reproduce or can be reproduced.

There is heredity in reproduction, that is to say that individuals produce similar offspring.

In the course of reproduction there is variety which affects the likelihood of survival and therefore of reproducibility of individuals.

There are finite resources which cause the individuals to compete. Due to overreproduction of individuals not all can survive the struggle for existence. Differential natural selection exerts a continuous pressure towards improved individuals.

The representation of programs, or generally structures, in GP has a strong influence on the behavior and efficiency of the resulting algorithm. As a consequence, many different approaches toward choosing representations have been adopted in GP. The resulting principles have been applied even to other problem domains such as design of electronic circuits or art and musical composition.

2. *The Mechanisms Behind Genetic Programming*

Genetic programming works with a population of programs that are executed or interpreted in order to judge their behavior. Usually, a scoring operation called fitness measurement is applied to the outcome of the behavior. For instance, the deviation between the quantitative output of a program and its target value (defined through an error function) could be used to judge the behavior of the program. This is straightforward if the function of the target program can be clearly defined. Results may also be defined as side-effects of a program, such as consequences of the physical behavior of a robot controlled by a genetically developed program. Sometimes, an explicit fitness measure is missing, for instance in a game situation, and the results of the game (winning or losing) are taken to be sufficient scoring for the program's strategy. The general approach is to apply a variety of programs to the same problem and to compare their performance relative to each other (see Fig. 1).

The outcomes of fitness measurement are used to select programs. There are a number of different

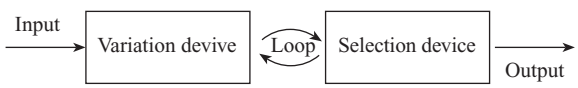

Figure 1
The variation selection loop of GP

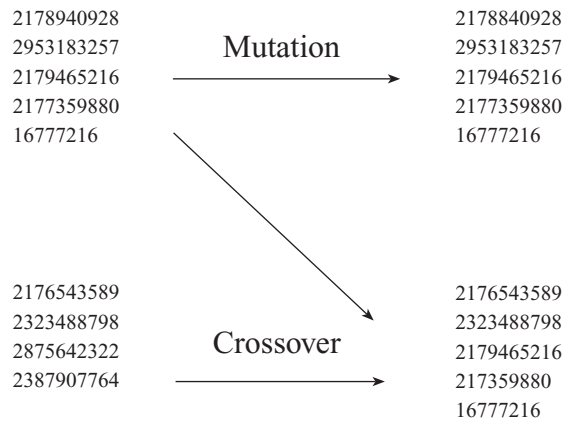

Figure 2
The primary operations of GP, mutation, and crossover are applied here to programs represented by the sequences of instructions. The instructions are coded as integer numbers which allows easy manipulation by access to these numbers

methods for selection, both deterministic and stochastic. Selection determines (a) which programs are allowed to survive (overproduction selection), and (b) which programs are allowed to reproduce (mating selection). Once a set of programs has been selected for further reproduction, the following operators are applied: reproduction, mutation, and crossover. Reproduction simply copies an individual, mutation varies the structure of an individual under control of a random number generator, and crossover mixes the structure of two (or more) programs to generate one or more new programs (see Fig. 2). Additional variation operators are applied in different applications. Most of these contain problem knowledge in the form of heuristic search recipes which are adapted to the problem domain.

In this way, fitness advantages of individual programs are exploited in a population to lead to better solutions. A key effort in GP is the definition of the fitness measure. Sometimes the fitness measure has to be iteratively improved in order for the evolved solutions actually to perform the function they were intended for. The entire process can be seen in close analogy to breeding animals. The breeder has to select those individuals from the population which carry the targeted traits to a higher degree than others.

Researchers (Turing 1950, Friedberg 1958, Samuel 1959) suggested ideas similar to GP already in the early days of AI, but did not get very far. So it was only after other techniques in evolutionary algorithms had been successfully developed that GP emerged. Earlier work concerned genetic algorithms (Holland 1975), evolution strategies (Schwefel 1981), and evolutionary programming (Fogel et al. 1966). These methods have been applied successfully to a wide spectrum of

problem domains, especially in optimization. However, it was unclear for a long time whether the principles of evolution could be applied to computer code, with all its dependencies and structural brittleness. Negative results from early experiments seemed to indicate that evolution of computer code was not possible. Successes were all in the area of constraint optimization (Michalewicz 1996), where methods were made available for how to deal with structural brittleness. These methods found their way into programming and gave rise to the new field of GP (Koza 1992).

3. *Progress and State-of-the-art*

In his seminal work of 1992, Koza established the field of GP by arguing convincingly that manipulation of symbolic tree structures is possible with evolutionary algorithms and that the resulting technique would have a wide variety of applications. In subsequent years, the field experienced both broadening and deepening (Banzhaf et al. 1998). Many different representations for GP were studied, among them other generic data structures such as sequences of instructions or directed graphs, as well as more exotic data structures such as stacks or neural networks. Today, different approaches are considered as GP, from the evolution of parse trees to the evolution of arbitrary structures. The overarching principle is to subject structures with variable complexity to forces of evolution by applying mutation, crossover, and fitness-based selection. The results must not necessarily be programs.

An ever-present difficulty with GP is that the evolution of structures of variable complexity leads often to large structures with considerable redundancy. Notably, variable complexity often leads to inefficient and space-consuming code. It was subsequently recognized that the evolutionary forces exerted a pressure toward more complex solutions, most of which could be removed after evolution without doing any harm to the evolved solution. By drawing an analogy from biological evolution of genomes, this phenomenon was called 'intron growth,' or growth of ineffective code. Thought the explanation for this phenomenon is not fully understood yet, it was found that at least two different influences were at work promoting the growth of complexity during evolution. The most important one has to do with the protection effect of redundant code if subjected to the action of crossover or mutation. Redundant code was resistant to crossover and mutation and allowed its carrier solution to survive better, compared to other individuals which did not possess the redundancy.

Currently, many researchers are working to transfer results from research in genetic algorithms to genetic programming. To achieve results in GP is generally more difficult since it works with variable complexity and multiple fitness cases for fitness scoring. The schema theory of genetic algorithms (Goldberg 1989, Vose 1999) has been a primary target of knowledge transfer. In the meantime, a number of different schema theorems have been formulated for GP.

When analyzing search spaces of programs it was realized that their size is many orders of magnitude larger than search spaces of combinatorial optimization problems. A typical size for a program search space might be $10^{100,000}$, as opposed to a typical search space for a combinatorial optimization problem of the order of 10^{100}. Although this might be interpreted as discouraging for search mechanisms, it was also realized that the solution density in program spaces is, above a certain threshold, constant with changing complexity (Langdon 1999). In other words, there are proportionally many more valid solutions in program spaces than in the spaces of combinatiorial optimization problems.

4. *Applications*

The main application areas of GP are (from narrow to wide) (Banzhaf et al. 1998): computer science, science, engineering, and art and entertainment. In computer science, the development of algorithms has been a focus of attention. By being able to manipulate symbolic structures, genetic programming is one of the few heuristic search methods for algorithms. Sorting algorithms, caching algorithms, random number generators, and algorithms for automatic parallelization of code (Ryan 2000), to name a few, have been studied. The spectrum of applications in computer science spans from the generation of proofs for predicate calculus to the evolution of machine code for accelerating function evaluation. The general tendency is to try to automate the design process for algorithms of different kinds.

Typical applications in science are to modeling and pattern recognition. Modeling certain processes in physics and chemistry with the unconventional help of evolutionary creativity supports research and understanding of the systems under study. Pattern recognition is a key ability in molecular biology and other branches of biology, as well as in science in general. Here, GP has delivered first results that are competitive if not better than human-generated results.

In engineering, GP is used in competition or cooperation with other heuristic methods such as neural networks or fuzzy systems. The general goal is again to model processes such as production plants, or to classify results of production. Control of man-made apparatus is another area where GP has been used successfully, with process control and robot control the primary applications.

In art and entertainment, GP is used to evolve realistic animation scenes and appealing visual graphics. It also has been used to extract structural in-

formation from musical composition in order to model the process so that automatic composition of music pieces becomes possible.

Many of these problems require a huge amount of computational power on the part of the GP systems. Parallel evolution has hence been a key engineering aspect of developments in GP. As a paradigm, genetic programming is very well suited for a natural way of parallelization. With the advent of inexpensive parallel hard- and software (Sterling et al. 1999), a considerable proliferation of results is expected from GP systems.

5. *Methodological Issues and Future Directions*

In recent years, some researchers have claimed human-competitive results in the application of genetic programming to certain problems (Koza et al. 1999). These claims are based on a comparison between the best known at present human solution to a problem and its GP counterpart. Usually, a large amount of computational power had to be invested in order to gain human-competitive results from genetic programming runs. Due to the ability of the human mind quickly to grasp the recipes of problem solution an artificial system has applied, the question remains open whether GP solutions will stay better than human solutions.

Theory of genetic programming is presently greatly underdeveloped and will need to progress quickly in order to catch up with other evolutionary algorithm paradigms. Most of the obstacles stem from the fact of variable complexity of solutions evolved in GP. Implementation of GP will benefit in the coming years from new approaches which include research from developmental biology. Also, it will be necessary to learn to handle the redundancy forming pressures in the evolution of code.

Application of genetic programming will continue to broaden. Many applications focus on controlling behavior of real or virtual agents. In this role, genetic programming may contribute considerably to the growing field of social and behavioral simulations. Genetic algorithms have already been found beneficial in optimizing strategies of social agents (Gilbert and Troitzsch 1999). With its ability to adjust the complexity of a strategy to the environment and to allow competition between agents, GP is well positioned to play an important role in the study and simulation of societies and their evolution.

See also: Adaptation, Fitness, and Evolution; Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence: Search; Evolution: Optimization; Evolution: Self-organization Theory; Evolutionary Theory, Structure of; Intelligence: History of the Concept; Natural Selection

Bibliography

Banzhaf W, Nordin P, Keller R, Francone F 1998 *Genetic Programming—An Introduction on the Automatic Evolution of Computer Programs and its Applications*. Morgan Kaufmann, San Francisco, CA

Fogel L J, Owens A, Walsh M 1966 *Artificial Intelligence Through Simulated Evolution*. Wiley, New York

Friedberg R M 1958 A learning machine: part I. *IBM Journal of Research and Development* **2**: 2–13

Gilbert N G, Troitzsch K G 1999 *Simulation for the Social Scientist*. Open University Press, Buckingham, UK

Goldberg D E 1989 *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, Reading, MA

Holland J H 1975 *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan Press, Ann Arbor, MI

Koza J R 1992 *Genetic Programming: on the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA

Koza J R, Andre D, Bennett F, Keane M 1999 *Genetic Programming III: Darwinian Invention and Problem Solving*. Morgan Kaufmann, San Francisco, CA

Langdon W 1999 Boolean function fitness spaces. In: Poli R, Nordin P, Langdon W, Fogarty T (eds.) *Proceedings EuroGP'99*. Springer, Berlin, pp. 1–14

Michalewicz Z 1996 *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, Berlin

Ryan C 2000 *Automatic Re-engineering of Software Using Genetic Programming*. Kluwer, Boston, MA

Samuel A 1959 Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **3**: 210–29

Schwefel H P 1981 *Numerical Optimization of Computer Models*. Wiley, Chichester, UK

Sterling T L, Salmon J, Becker D, Savarese D 1999 *How to Build a Beowulf: A Guide to the Implementation and Application of PC Clusters*. MIT Press, Cambridge, MA

Turing A M 1950 Computing machinery and intelligence. *Mind* **59**: 433–60

Vose M D 1999 *The Simple Genetic Algorithm: Foundations and Theory*. MIT Press, Cambridge, MA

W. Banzhaf

Artificial Intelligence in Cognitive Science

Artificial intelligence (AI) and cognitive science are, at the turn of the twenty-first century, two distinct disciplines, with overlapping methodologies, but rather different goals. AI is a branch of computer science and is concerned with the construction and deployment of intelligent agents as computer programs, and also with understanding the behavior of these artifacts. The core scientific goal of AI is to understand basic principles of intelligent behavior that apply equally to animal and artificial systems (Russell and Norvig 2000). Almost all of the work is math-

ematical or computational in character and much of the literature is technique oriented.

Cognitive science (see *Cognitive Science: History*; *Cognitive Science: Overview*) is an explicitly interdisciplinary field that has participation from AI, but also from linguistics, philosophy, psychology, and subfields of other social and biological sciences. The unifying goal of cognitive science is to understand and model human intelligence, using the full range of findings and methodologies of the complementary disciplines. As one would expect, a wide range of techniques from the mathematical, behavioral, social, and biological sciences are employed. Cognitive science, in contrast with AI, is defined more by phenomena than by methodology (Posner 1985). There are research groups that are active in both AI and cognitive science, but they tend to produce different papers for journals and conferences in the two areas.

1. *Shared Origins in the Postwar Cognitive Revolution*

Both AI and cognitive science evolved after 1950 and, in their early development were more tightly integrated than at present (see *Cognitive Science: History*). For much of the first half of the twentieth century, the Anglo-American study of cognition was dominated by the behaviorist paradigm, which rejected any investigation of internal mechanisms of mind. The emergence of both AI and cognitive science was part of a general postwar movement beyond behaviorist theories, which also included new approaches in linguistics and the social sciences. The idea of computational models of mind was a central theme of what is sometimes called the postwar cognitive revolution. One of the leading early AI groups, under the leadership of Allen Newell and Herbert Simon at Carnegie Mellon (Newell 1990), was explicitly concerned with cognitive modeling using the symbolic processes of AI. The current version of this continuing effort is described in the article on SOAR (see *Cognitive Theory: SOAR*). Another traditional symbolic approach to modeling intelligence is presented in the article on ACT (see *Cognitive Theory: ACT*). But there is very little work in contemporary AI that is explicitly focused on modeling human behavior as opposed to intelligent systems in general. There is some continuing work on human and machine game playing (see *Chess Expertise, Cognitive Psychology of*), but it is not integrated into either field.

The central idea of AI is computational modeling of intelligent behavior and this is its main contribution to cognitive science. The basic notion of a computational model is now commonplace in all scientific fields and many other aspects of contemporary life. One builds a detailed software model of some phenomenon and studies the behavior of the model, hoping to gain understanding of the original system. Much of the work in AI has the engineering goal of producing practical systems, and there is no sharp boundary between AI and other applied fields of computer science and engineering. AI techniques are now commonplace in the full range of business, scientific, and public applications. While all fields *use* computational models, researchers, in computer science in general and AI in particular, *invent* and study computational techniques for constructing models, presenting the results of simulations, and understanding the limitations of the simulation. AI has traditionally studied the modeling of the most complex phenomena—those relating to intelligence. Because of the technical challenges arising in the construction of these complex simulations, many innovations in computing have arisen in AI and then been more widely applied.

2. *Domain-focused Research that Cuts Across AI and Cognitive Science*

The relationship between AI and cognitive science is further complicated by the fact that there are currently several distinct research fields that cut across both disciplines, but have separate journals, meetings, etc. The most prominent of these research areas are speech, language, vision, and neural networks. Each of these fields has thousands of practitioners, many of whom are interested in AI, cognitive science, or both. Appropriately, each of these areas is represented by multiple articles in this encyclopedia. As AI and cognitive science have grown, specialized areas such as language and vision modeling have become largely independent, but they do continue to share the development of underlying methodologies. The main areas that have remained as core AI include knowledge representation and reasoning, planning and problem solving. The study of learning has evolved somewhat differently and this will be discussed below. There are some common scientific paradigms that cut across all of these fields and this article will try to outline them as they relate to the social and behavioral sciences. The common thread linking AI to cognitive science is reliance on computational models of different kinds.

3. *The Role of Formal Logic in AI and Cognitive Science*

To a great extent, the early development of AI was based on symbolic, as opposed to numerical, modeling. This led to the introduction of some novel representations such as those of SOAR (see *Cognitive Theory: SOAR*), but the main effect of this was to align AI with formal logic for much of its early history. In fact, much of the driving force for the creation of the

field called cognitive science came from people who saw mathematical logic as its unifying theme. This remains a fruitful approach to AI and constitutes one major area of overlap with cognitive science. Mathematical logic is elegant and well developed and can be shown to be, in some sense, general enough to represent anything that can be described formally. There remains a significant community of linguists, philosophers, and computer scientists for whom logic is the only scientific way to study intelligence (Posner 1985).

But the twentieth century was not kind to categorical, deterministic theories in any field and cognitive science was no exception (see *Imagery versus Propositional Reasoning*). A central issue in the study of cognition has been how to describe the meaning of words and concepts. In formal logic, a concept is defined by a set of necessary and sufficient conditions. For example, a bachelor might be defined as a male who never married. The limitations of classical, all-or-none, categories were already recognized by Wittgenstein, who famously showed that concepts such as 'game' could not be characterized by necessary and sufficient conditions, and were better described as family resemblances. Even the definition of bachelor becomes graded when we consider cohabitation or how old a male needs to be before being considered a bachelor (see *Word Meaning: Psychological Aspects*).

Starting in the 1960s, a wide range of cognitive science studies by Rosch and others showed the depth and complexity of human conceptual systems and their relation to language. This helped give rise to the subfield called cognitive linguistics (Lakoff 1987), which overlaps with cognitive science, but has its own paradigms, journals, and conferences. The graded and relational nature of human categories undermined the attempt to create a unified science of mind based on formal logic. Another attack on the formalist program arose from the growing understanding of the neural basis of intelligence, also discussed in several articles (see *Cognitive Control (Executive Functions): Role of Prefrontal Cortex*; *Declarative Memory, Neural Basis of*; *Dreaming, Neural Basis of*; *Episodic and Autobiographical Memory: Psychological and Neural Aspects*; *Emotion, Neural Basis of*; *Implicit Learning and Memory: Psychological and Neural Aspects*). A crucial insight was that basic human concepts are grounded in direct experience and that more abstract concepts are mapped metaphorically to more embodied ones. This undercuts the formalist program, and also leads to a separation of AI, which studies intelligence in the abstract, from cognitive science, which is explicitly concerned with human minds. While much excellent work continues to be done using logic, it is now generally recognized that there are a wide range of phenomena that are better handled by biologically based models and/or some form of numerical, often probabilistic, modeling. For a variety of reasons, the movement to quantitative numerical models followed somewhat different paths in AI and cognitive science, but there are some recent signs of reconvergence.

4. *Learning and the Connectionist Approach to Cognitive Science*

As it happens, while the formalists were trying to establish a cognitive science based on formal logic, an antithetical neural network movement was also developing, and this approach has become a major force in cognitive science. The two contrasting approaches to cognitive modeling, neural modeling and logic, were mirrored in the two different methods by which early computer scientists sought to achieve AI. From the time of the first electronic computers around 1950, people dreamed of making them 'intelligent' by two quite distinct routes. The first, which was described above as conventional AI, is to build standard computer programs as models of intelligence. This remains the dominant paradigm in AI and has had considerable success. The other approach was to try to build hardware that was as brain-like as possible and have it *learn* the required behavior. The history of this 'neural modeling' approach is described in several articles (see *Artificial Intelligence: Connectionist and Symbolic Approaches*; *Connectionist Approaches*; *Connectionist Models of Concept Learning*; *Connectionist Models of Development*; *Connectionist Models of Language Processing*). After some promising early mathematical results on learning in simple networks, the neural learning approach to modeling intelligence fared much less well for three decades and had little scientific or applied success. Around 1980, a variety of ideas from biology, physics, psychology, and computer science and engineering coalesced to yield a 'new connectionist' approach to modeling intelligence that has become a core field of cognitive science, and also the basis for a wide range of practical applications (McClelland and Rumelhart 1986). Among the key advances was a mathematical technique (back-propagation) that extended the early results on learning to a much richer set of network structures.

Connectionist computational models are almost always computer programs, but programs of a different kind than those used in, for example, word processing or symbolic AI. Connectionist models are specified as a network of simple computing units, which are abstract models of neurons. Typically, a model unit calculates the weighted sum of its inputs from upstream units and sends to its downstream neighbors an output signal that is a nonlinear function of its inputs. Learning in such systems is modeled by experience-based changes in the weights of the connections between units. The basic connectionist style of modeling is now being used in three quite different ways in neurobiology, in applications, and in cognitive science. Neurobiologists who study networks of neurons employ a wide range of computational mod-

els, from very detailed descriptions of the internal chemistry of the neuron to the abstract units described above. The use of connectionist neural models in practical applications is part of the reconvergence with AI and will be discussed in the final section.

In cognitive science, connectionist techniques have been used for modeling all aspects of language, perception, motor control, memory, and reasoning. This universal coverage represents a potential breakthrough; previously the computational models of, for example, early vision and problem solving used entirely different mathematical and computational techniques. Since the brain is known to use the same neural computation throughout, it is not surprising that neurally inspired models can be applied to all behavior. Unfortunately, the existing models are neither broad nor deep enough to ensure that the current set of mechanisms will suffice to bridge the gap between structure and behavior, but the work remains productive.

Connectionist models in cognitive science fall into two general categories, often called structured and layered (also called PDP) networks. Most modelers are primarily interested in learning, which is modeled as experience-driven change in connection weights. There is a great deal of research studying different models of learning with and without supervision, different rules for changing weights, etc. Because of the focus on what the network can learn, any prewired structure will weaken the results of the experiment. The standard approach is to use networks with unidirectional connections arranged in completely connected layers, sometimes with a very restricted additional set of feedback links. This kind of network contains a minimum of presupposed structure and is also amenable to efficient learning techniques such as the back-propagation method described above. Most researchers using totally connected layered models do not believe that the brain shares this architecture, but there is an ongoing controversy about the implications of PDP learning models for theories of mind, discussed in Sect. 4.1.

Structured connectionist models are usually less focused on learning than on the representation and processing of information. Essentially all the modeling done by neurobiologists involves specific architectures, which are known from experiment. For structured connectionist models of cognitive phenomena, the underlying brain architecture is rarely known in detailand sometimes not at all at the level of neurons and connections. The methodology employed is to experiment with computational models of the behavior under study which are consistent with the known biological and psychological data and are also plausible in the resources (neurons, computing time, etc.) required. This methodology is very similar to what are called spreading activation models, widely used in psycholinguistics. Some studies combine structured and layered networks (Regier 1996), or investigate learning in networks with an initial structure that is tuned to the problem area or the known neural architecture.

4.1 Nature and Nurture: Rules vs. Connections

Perhaps the most visible contribution to date of connectionist computational models in cognitive science has been to provide a new jousting ground for contesting some age-old issues on the nature of intelligence. Much of the current debate is being published in *Science* magazine, which suggests that it is considered to be of major importance by the US scientific establishment. The nature/nurture question concerns how much of some trait, usually intelligence, can be accounted for by genetic factors, and how much depends on postnatal environment and training. Some PDP connectionists have taken very strong positions, suggesting that learning can account for everything interesting (Elman et al. 1996). In the particular case of grammar, an important group of linguists and other cognitive scientists take an equally extreme nativist position, suggesting that humans only need to choose a few parameters to learn grammar. A related issue is whether human grammatical knowledge is represented as general rules or just appears as the rule-like consequences of PDP learning in the neural network of the brain. There is ample evidence against both extreme positions, but the debate continues to motivate a great deal of thought and experiment.

5. Current and Future Trends

Although the fundamental split between the AI focus on general methods and the cognitive science emphasis on human intelligence remains, there are a growing number of areas of overlapping interest. As was discussed above, quantitative neural models are playing a major role in cognitive science. It turns out that the mathematical and computational ideas underlying learning in neural networks have found application in a wide range of practical problems, from speech recognition to financial prediction. The basic idea is that, given current computing power, back-propagation and similar techniques allow large systems of nonlinear units to learn quite complex probabilistic relationships using labeled data. This general methodology overlaps not only with AI but also with mathematical statistics, and is part of a unifying area called computational learning theory. There is also a large community of scientists and engineers who identify themselves as working on neural networks and related statistical techniques for various scientific and applied tasks, along with conferences and journals to support this effort.

While probability was entering cognitive science from the bottom up through neural models, AI experienced the introduction of probabilistic methods from general theoretical considerations, which only later led to practical application. As was discussed in Sect. 3, the limitations of formal logic became well recognized in the 1960s. Over the subsequent decades, AI researchers, led by Judea Pearl of UCLA (Pearl 1988), developed methods for specifying and solving large systems of conditional probabilities. These belief networks are now widely used in applications ranging from medical diagnosis to business planning. A growing field that involves both symbolic and statistical techniques is 'data mining,' processing large historical databases to search for relationships of commercial or social importance. Recent efforts to learn or refine belief networks from labeled data are another area of convergence of AI and cognitive science in computational learning theory.

Of course, the explosion of Internet activity is affecting AI along with the rest of the computing field. Two AI application areas that seem particularly important to cognitive science are intelligent Web agents and spoken-language interaction.

As the range of users and activities on the Internet continues to expand, there is increasing demand for systems that are both more powerful and easier to use. This is leading to increasing efforts on the human–computer interface, including the modeling of user plans and intentions—clearly overlapping with traditional concerns of cognitive science. One particularly active area is interaction with systems using ordinary language. While machine recognition of individual words is relatively successful, dealing with the full richness of language is one of the most exciting challenges at the interface between AI and cognitive science, and a problem of great commercial and social importance.

Looking ahead, we can be confident that the increasing emphasis on intelligent systems will continue. From the scientific perspective it is very likely that most of the interdisciplinary research in cognitive science will remain focused on specialized domains such as language, speech, and vision. General issues including representation, inference and learning will continue to be of interest, and will constitute the core of the direct interaction between AI and cognitive science. With the rapid advances in neurobiology, both fields will increasingly articulate with the life sciences, with great mutual benefits.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Chess Expertise, Cognitive Psychology of; Cognitive Science: History; Cognitive Science: Overview; Cognitive Theory: ACT; Cognitive Theory: SOAR; Imagery versus Propositional Reasoning; Neural Networks: Biological Models and Applications; Word Meaning: Psychological Aspects

Bibliography

Elman J, Bates E, Johnson M 1996 *Rethinking Innateness: A Connectionist Perspective on Development (Neural Network Modeling and Connectionism)*. MIT Press, Cambridge, MA

Lakoff G 1987 *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, Chicago

McClelland J L, Rumelhart D E 1986 *Parallel Distributed Processing*. MIT Press, Cambridge, MA

Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA

Posner M I (ed.) 1985 *Foundations of Cognitive Science*. MIT Press, Cambridge, MA

Regier T 1996 *The Human Semantic Potential*. MIT Press, Cambridge, MA

Russell S J, Norvig P 2000 *Artificial Intelligence*. Prentice-Hall, Upper Saddle River, NJ

J. Feldman

Artificial Intelligence: Search

Search is the study of computer algorithms designed to solve problems by systematic trial-and-error exploration of possible solutions. Problem-solving tasks include single-agent path-finding problems, such as Rubik's Cube, two-player games such as chess, and constraint-satisfaction problems, such as the Eight Queens problem. We will consider each in turn.

1. Single-agent Path-finding Problems

A classic example of a single-agent path-finding algorithm is the *Fifteen Puzzle* (see Fig. 1). It consists of a 4×4 frame containing 15 numbered tiles, and an empty or blank position. Any tile horizontally or vertically adjacent to the blank can be slid into the blank position. The goal is to rearrange the tiles from some random initial configuration to a particular goal configuration. The *states* of the problem are the

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

Figure 1
Fifteen Puzzle

different permutations of the tiles. The *operators* are the legal moves. The Fifteen Puzzle task is to find a sequence of operators that map the initial state to the goal state, and ideally a shortest such sequence.

A search algorithm may be systematic or nonsystematic. A systematic search algorithm is guaranteed to find a solution if one exists, and may in fact guarantee a lowest-cost or optimal solution. If no solution exists, a systematic algorithm may or may not detect this fact. Nonsystematic algorithms, such as simulated annealing, genetic algorithms, heuristic repair, and other local search or stochastic approaches, are either not guaranteed to find a solution at all, or are not guaranteed to find an optimal solution, and are not considered here.

The simplest systematic search algorithms are called *brute-force search* algorithms, and do not use any knowledge about the problem other than the states, operators, initial state, and description of the goal state. For example, *breadth-first search* starts with the initial state, then considers all states that can be reached from the initial state with one operator application, then all states two operator applications away, etc., until a goal is reached. *Uniform-cost search*, or *Dijkstra's* single-source shortest-path algorithm (Dijhstra 1959) allows operators to have different costs, and visits states in increasing order of their total cost from the start, where the cost of a state is the sum of the operator costs incurred in reaching that state. A drawback of each of these algorithms is that they require enough memory to hold all the states considered so far, which is prohibitive in large problems.

Depth-first search more closely approximates how one would search using a single physical copy of a puzzle, such as the Fifteen Puzzle. From the current state, depth-first search applies a sequence of operators, until some depth limit is reached. It then backtracks to the last state that hasn't been completely explored, and applies a new sequence of operators from there. The advantage of depth-first search is that it only requires enough memory to hold the current operator sequence from the initial state.

A drawback of all brute-force search algorithms is that they are uninformed about the location of the goal. Heuristic search, however, is directed toward the goal. A heuristic search, such as the *A* algorithm* (Hart et al. 1968), makes use of a *heuristic evaluation function*, which estimates the cost to reach the goal from any given state. For example, the *Manhattan distance* is a common heuristic estimate for the Fifteen Puzzle. It is computed by counting the number of grid units each individual tile is away from its goal position, and summing these values over all tiles. For every state n visited by A*, it adds the number of moves made from the initial state to reach the current state, $g(n)$, to the heuristic estimate of the number of remaining moves to the goal, $h(n)$, to estimate the total number of moves on a path from the initial state to the goal that passes through the current state, $f(n) = g(n) + h(n)$. A* starts with the initial state, and generates all the states immediately adjacent to it. It then evaluates those states n using the cost function $f(n) = g(n) + h(n)$. At each step, it generates and evaluates the neighbors of the unexplored state n with the lowest total cost estimate, $f(n)$. It stops when it chooses a goal state to explore next.

A* is guaranteed to find a solution if one exists. Furthermore, if the heuristic function $h(n)$ never overestimates actual cost, A* is guaranteed to find a shortest solution. For example, since each move only moves one tile one grid unit, the Manhattan distance is a lower bound on the actual number of moves required to solve a Fifteen Puzzle state, and A* using the Manhattan distance heuristic function is guaranteed to find a shortest solution to the Fifteen Puzzle, given sufficient memory.

The bane of all search algorithms is *combinatorial explosion*. For example, the Fifteen Puzzle contains $16!/2 \approx 10^{13}$ different reachable states. Even at ten million states per second, a brute-force search examining all these possibilities would take over 12 days. The 5×5 *Twenty-Four Puzzle* contains $25!/2 \approx 10^{25}$ states. At ten million states per second, brute-force search would take over 24 billion years, longer than the age of the universe! With heuristic search techniques, optimal solutions have been found for both the Fifteen and Twenty-Four Puzzles, but for no larger versions.

2. *Two-player Games*

The classic two-player game is chess and has been an object of research in artificial intelligence for over fifty years. The current best programs are comparable to the best human players. The standard algorithm for these problems is *minimax search* with static evaluation, described below, and *alpha-beta pruning*, a technique that makes minimax search much more efficient. The states of the problem are the legal board positions, and the operators are the legal moves of the game. From the current board position, we consider all the legal moves we can make, then all the possible responses by our opponent, then all our possible countermoves, etc. In small games, like tic-tac-toe, there is sufficient time for the machine to explore the entire game tree, or all possible plays by both players, allowing the machine to play perfectly.

In more complex games, brute-force search and hence perfect play are impossible. For example, the chess game tree is estimated to contain over 10^{40} different board positions. In that case, the machine searches ahead as many moves as it can, given the time available to make each move. The limit of how far it can look ahead is called the *search horizon*. Each of the game states at the search horizon is then evaluated by a heuristic function that returns a number, which estimates the relative merit of that board position for

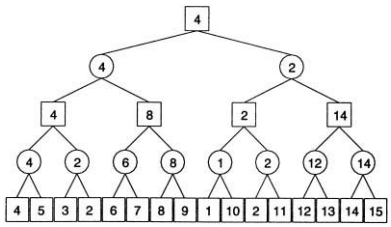

Figure 2
Minimax search tree example

each player. A large positive value is advantageous for one player, called MAX, and a negative value of large magnitude is beneficial to the player called MIN. For example, the simplest evaluation function for chess would sum all the pieces on the board belonging to MAX, weighted by their relative values, and subtract the weighted sum of all of MIN's pieces on the board, to get the relative material advantage of MAX over MIN. For example, a weighting function for chess might count the queen as worth nine points, the rooks as five points each, the bishops and knights as three points, and one point for each pawn.

Given the numeric evaluation of the board positions at the search horizon, the move to make is computed by the minimax algorithm. We compute the value of each position in the game tree recursively from the value of the positions that result from each different legal move. If the MAX player is to move in some position, the value of that position is the maximum of the values of the positions that result from each legal move. Similarly, if the MIN player is to move in some position, the value of that position is the minimum of the values of the positions that result from each legal move. By backing these values up the tree, we can compute the value of the current game state, and make a move to the position that gave rise to this value.

Figure 2 shows an idealized four-level game tree, where the square nodes represent the game states where MAX is to move, and the circle nodes represent states where MIN is to move. The numbers at the bottom level of the tree from the heuristic evaluation function, and the numbers at the interior nodes are computed from the nodes immediately below them by the minimax rule. The move recommended by this procedure is to the left child of the root node. A much more efficient version of this algorithm, known as alpha-beta pruning (Knuthand Moore 1975), is used in practice.

3. *Constraint-satisfaction Problems*

A classic constraint-satisfaction problem is the Eight Queens Problem. The task is to place eight queens on a standard 8×8 chessboard so that no two queens are attacking along the same row, column, or diagonal. Figure 3 shows one solution to the Eight Queens problem. In a constraint-satisfaction problem, the task is not to find a sequence of moves, but to identify a state that satisfies a set of constraints. The standard algorithm for these problems is known as *backtracking*. We systematically place queens on the board one at a time, one per row, so that no two queens on the board are in the same column or diagonal. If we get to a state in which no additional queens can be placed without attacking an existing queen, and all the queens have not yet been placed, we backtrack by moving the last queen that was placed to the next column where it doesn't attack any of the previously placed queens. The algorithm continues until either one solution is found, all solutions are found, or it is determined that no solutions exist.

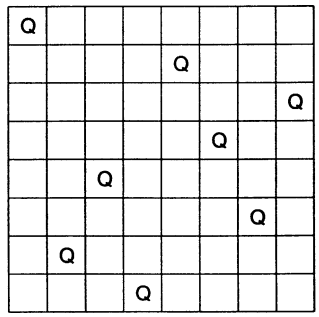

Figure 3
A solution to the Eight Queens Problem

See also: Artificial Intelligence: Uncertainty; Artificial Intelligence in Cognitive Science; Decision Support Systems; Deductive Reasoning Systems; Intelligence: History of the Concept; Scientific Discovery, Computational Models of;

Bibliography

Bolc L, Cytowski J 1992 *Search Methods for Artificial Intelligence*. Academic Press, London

Dijkstra E W 1959 A note on two problems in connexion with graphs, *Numerische Mathematik* **1**: 269–71

Hart P E, Nilsson N J, Raphael B 1968 A formal basis for the heuristic determination of minimum cost paths, *IEEE Transactions on Systems Science and Cybernetics* **4**(2): 100–7

Korf R E 1998 Artificial intelligence search algorithms. In: Attalah M J (ed.) *CRC Handbook of Algorithms and Theory of Computation*. CRC Press, Boca Raton, FL, pp. 36–20

Knuth D E, Moore R E 1975 An analysis of alpha-beta pruning, *Artificial Intelligence* **6**(4): 293–326

Pearl J 1984 *Heuristics*. Addison-Wesley, Reading MA

R. E. Korf

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Artificial Intelligence: Uncertainty

Artificial intelligence (AI) is the discipline of computer perception, reasoning, and action. Dealing with uncertainty is a central challenge for artificial intelligence. Uncertainty management capabilities are required to combine evidence about a new situation with knowledge about similar situations, to draw inferences, and predict the effects of actions. Numerical computing has traditionally focused on problems where measurement imprecision is the sole source of reasoning uncertainty. AI researchers aim to develop software systems for applications such as automated learning, perception, natural language, and speech understanding. Such systems must deal with many sources of uncertainty: including equally plausible alternative explanations; missing information; incorrect object and event typing; diffuse evidence; ambiguous references; prediction of future events; and deliberate deception.

AI researchers in the late 1950s worked with computing environments inferior to a wristwatch calculator in 2000. As computing power increased, emphasis shifted from optimizing speed and memory to increasing expressiveness in knowledge representation and reasoning strategies. Automating intellectual, symbolic, and common-sense activities turned out to be daunting tasks, still mostly exceeding the grasp of computable representations. Over decades several common and interdependent themes emerged that explained key and repeating sources of failure in AI programs.

One such failure is incorrect categorization of object types. An example of incorrect categorization is to call a cat a dog because it is a housepet with four legs. Another problem is inadequate treatment of uncertainty or ambiguity with respect to explanations for a given pattern of evidence. Examples include pronouns with multiple possible antecedents, or multiple chains of inference to explain the same event, e.g., the ground is wet because it rained, because the sprinkler was on, or both. A third source of failure is incomplete, inadequate, or nonexistent models for explaining events, often due to unstated common-sense assumptions. An example is failure of the computer to appreciate that a truck at a loading dock will not wait indefinitely for someone to load it. All of these failures can be viewed as issues in automated handling of uncertainty in reasoning.

It has become clear that effective treatment of such uncertainty requires considering multiple alternatives throughout the incremental accrual of evidence. Although this ability appears to be innate in humans, naïve computations of alternatives in the face of ambiguous evidence quickly become a combinatorial nightmare, exploding memory and computational requirements at exponential rates.

Remarkable innovations in computing have arisen out of the search for efficient methodologies for representing multiple alternatives in the face of conflicting, ambiguous, and multiple sources of evidence, for weighing evidence in favor of different alternatives, and for deciding among multiple options. These innovations include logic-based theorem proving, expert systems, neural networks, fuzzy set theory, belief function theory, and Bayesian networks. The following sections present a roughly chronological view of key developments in the history of managing uncertainty in AI.

1. Symbolic Computing, Confidence and Class Membership

In the 1960s the AI community established the importance of symbolic computing as distinct from traditional numerical computing. Mainstream AI research focused on purely qualitative, symbolic approaches to reasoning with uncertainty such as default or non-monotonic logic, and the application of heuristic strategies for selecting among alternative world models and resolving conflicts (McCarthy 1960).

With algebraic and other equation-oriented representations, the focus tends to be on the result of the computation. Logic representations, being sentence-like with typed categorical terms, lend themselves naturally to being stored as 'declarative' knowledge. Declarative knowledge is expressed in the form of data structures that can be manipulated, decomposed, and analyzed by a reasoner independent of content. Expert systems apply theorem provers, or algorithms capable of deriving true propositions from other true propositions, to declarative knowledge in the form of facts and rules of inference specific to a domain of application. Successful expert systems have been deployed with great success in a variety of application areas, and are regarded as one of the great engineering successes of artificial intelligence.

Rule sets proved to be a natural way to represent computable systems that did not have complete knowledge of the problem to be addressed. Such a system begins with what is initially known and infers additional facts, as evidence becomes available. As richer rule-based representations were developed and larger-scale problems were tackled, technological limitations became apparent. Key issues included confidence in applicability of rules to particular situations as well as in their derived consequences. A related issue was uncertain typing of logical variables. The third was combinatorial explosion of alternatives as uncertain evidence activated multiple and often conflicting rules.

Medical diagnostic expert systems provided some of AI's great initial successes, but the complexity of applying medical knowledge required estimates of the confidence that rules and their conclusions applied in specific situations. Buchanan and Shortliffe (1974) developed the first calculus for confidence in expert

systems, called 'certainty factors.' Heralded as a breakthrough at the time of their invention, certainty factors were later proven by one of Shortliffe's students (Heckerman 1986) to be mathematically inconsistent and formally superseded by Bayesian probabilistic reasoning.

At the same time that certainty factors were in development, others pursued different formalisms for numerical representations of uncertainty. Shafer (1976) was concerned with representing situations in which evidence for estimating probability values is incomplete. He adapted the work of Dempster to develop a formal calculus for inductive reasoning with incomplete evidence.

Zadeh (1975) addressed the issue of uncertain typing of variables with the invention of fuzzy set theory. Fuzzy set theory extends traditional set theory by allowing varying degrees of membership in a set. For example, a man 2.3 meters tall would be assigned a high degree of membership in the fuzzy set 'very tall', a lower degree of membership in the fuzzy set 'tall', and a zero degree of membership in the fuzzy set 'short.' By contrast, a man 1.5 meters tall would be assigned a high degree of membership in the fuzzy set 'short', a low degree of membership in the fuzzy set 'tall', and a zero degree of membership in the fuzzy set 'very tall.' Zadeh's fuzzy calculi were later extended over logical representations to yield fuzzy logic and fuzzy expert systems.

By the late 1970s, a growing segment of AI researchers recognized the combinatorial explosion inherent in rule-based systems, as well as in straightforward combinations of logic with other theories such as fuzzy sets. Proposals to incorporate new structures for numerical uncertainty measures were sometimes dismissed by arguing that, with the exception of highly trained individuals such as statisticians and weather forecasters, humans rarely reason explicitly with numerical likelihoods. Critics argued that if humans can behave intelligently using heuristic processes that appear to be primarily based on qualitative symbolic rules, then qualitative processes should be the solution of first, and last, resort for AI. Somewhat paradoxically, it was also argued that computational complexity precluded the application of probability theory and related numerical uncertainty management approaches.

2. *Graphical Networks of Alternatives*

Several related trends coalesced into a shift in AI community consensus in the 1980s. One trend was the resurgence of interest in connectionist models (e.g., Rumelhart and McClelland 1985). The connectionist movement, which includes the development of neural networks (see *Neural Networks and Related Statistical Latent Variable Models*; *Neural Networks: Biological Models and Applications*), lent strong support to the thesis that fundamentally numerical approaches could give rise to computational systems that exhibited intelligent behavior. Connectionism also sparked interest in symbol-level representations that integrated smoothly with numerical sub-symbolic representations, especially for reasoning from perceptual signals to higher level abstractions.

The development of this research direction culminated in a series of breakthroughs in automated inference and the development of graphical models and associated algorithms for automated probabilistic decision making (Pearl 1988, D'Ambrosio 1999 and *Bayesian Graphical Models and Networks* and *Latent Structure and Casual Variables*). Graphical models combine qualitative rule-like and object-like knowledge structures with quantitative measures of the uncertainty associated with inferences. Alternative inferences are represented in all the possible chains of reasoning implicit in the graphical structure, and need not be explicitly enumerated. The loosely coupled, modular architecture of graphical models enables the creation of knowledge representations and tractable algorithms for inference, planning, and learning for realistically complex problems. Directed graphical probability models are called 'Bayesian networks' and undirected graphical probability models are called 'Markov graphs' (Pearl 1988, Jensen 1996). Shafer and Shenoy combined Dempster-Shafer calculus and Bayesian network concepts to build even more general knowledge structures out of graphs encoding dependencies among variables, and proved the existence of a universal representation for automating inductive inference (Shafer and Shenoy 1990). Graphical models became increasingly popular as a common framework, independent of uncertainty calculus, for representing the loosely coupled dependency relationships that give rise to the modular representations that are basic to AI. Graphical models are also useful for expressing the causal relationships that underlie the ability to predict the effects of manipulations and form effective plans (Pearl 2000, Spirtes et al. 2000).

Many uncertain attributes of knowledge, including belief, credibility and completeness, can be expressed using graphical models and their related computational calculus. The most prominent issue in the field of uncertainty in AI has been the representation and reasoning about belief in alternatives given uncertain evidence. More recently there has been increased focus on planning and action, as well as approaches integrating perception to symbolic-level reasoning, planning, and action.

3. *Incremental Model Construction: An Example*

A limited example, originally due to Max Henrion, serves to illustrate key concepts such as the interplay between structural assumptions and numerical beliefs, the distinction between causal and evidential reasoning, the benefits of loosely coupled modular representations, the integration of uncertainty management

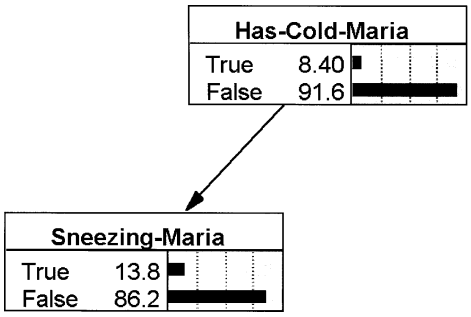

Figure 1
Initial model for Maria's sneezing episode

into knowledge representation and reasoning architectures, and the role of learning. The example uses a Bayesian network to represent uncertainty; nevertheless, the issues illustrated by this example are common to all belief calculi.

Maria is visiting her friend Jose when she suddenly begins sneezing. 'I must be getting a cold,' she thinks. Fig. 1 illustrates how such a human reasoning pattern can be encoded in a directed graphical model, or Bayesian network. This Bayesian network represents both the deductive knowledge that colds cause sneezing, and the inductive knowledge that sneezing constitutes evidence for a cold. The probability bars in Fig. 1a represent the prior probabilities assigned to whether or not Maria has a cold and whether or not she will be sneezing. Fig. 1b shows how these probabilities change after incorporating the evidence that Maria is sneezing.

After concluding that she probably has a cold, Maria notices scratches on the furniture. 'Oh, it's nothing but my cat allergy acting up,' she sighs in relief. Fig. 2 shows how the Bayesian network can be extended to incorporate this additional knowledge. Each time a new item of evidence is incorporated, an inference algorithm is applied to update the probabilities of the propositions for which truth-values remain unknown. When Maria began sneezing, the probability she had a cold increased from about 8 percent to about 53 percent. When she saw the scratches, however, the probability of having a cold dropped back to about 15 percent.

This non-monatomic change in probabilities that occurs with the incorporation of evidence for a competing explanation is known as explaining away. The evidence for Maria's having an allergic reaction explains the symptom for which the cold hypothesis had been nominated as an explanation. The cold hypothesis, which has no corroborating evidence and is no longer needed to explain the sneezing, consequently drops in probability as evidence mounts for the allergy hypothesis. Only the final probability of Maria's having an allergy is shown in the figure. In fact, this hypothesis increases monotonically in probability from about 3 percent prior to the onset of sneezing, to about 20 percent after sneezing is observed, to 88 percent after seeing the scratches.

The phenomenon of explaining away is a well-known structural feature of causal argumentation. It is an instance of a more general phenomenon known as intercausal dependence, in which two, heretofore unrelated, causes become relevant to each other when information about a common effect becomes known. The ability of Bayesian networks to model intercausal dependence comes about as a natural consequence of the independence assumptions encoded by directed graphs as representations of dependence among pro-

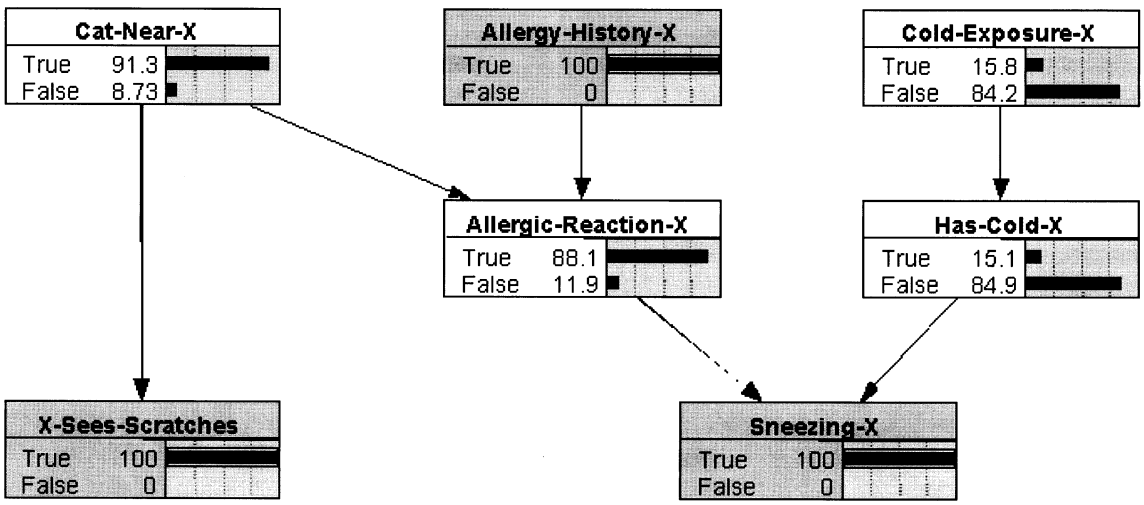

Figure 2
Revised model for Maria's sneezing episode

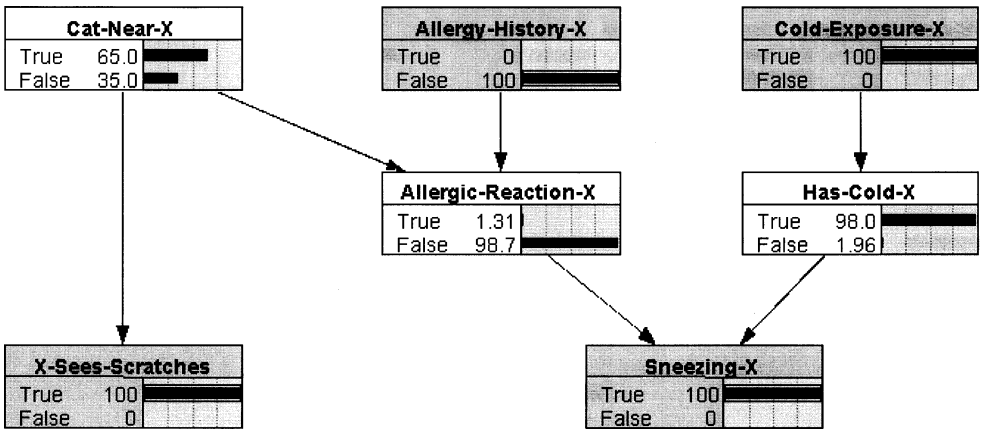

Figure 3
Generic model for sneezing episodes

positions. The two causes of sneezing are independent of each other prior to observation of the sneezing episode. They become negatively related due to their roles as competing explanations for an observed episode of sneezing. Bayesian networks exhibit intercausal dependence as a consequence of the mathematics of probability theory, and the manner in which directed graphs are applied to represent conditional independence assumptions. Observing the value of a node at the head of converging links in a directed graphical model opens the flow of information through the link. This enables information favoring one cause of sneezing to flow through the converging link and decreases the probability of the other cause of sneezing (Jensen 1996).

A different kind of dependence relation is exhibited by diverging links, such as the links connecting the node Cat-Near-Maria to its two children, and serial links, such as the links entering and leaving the node Allergic-Reaction-Maria. Observing a node at a diverging or serial link cuts off the flow of evidence through the link. Thus, learning about the scratches increases the probability that a cat is present, which in turn increases the probability of an allergic reaction. However, if we already know whether or not a cat is present, the additional information that there are scratches is no longer relevant to the question of whether Maria is having an allergic reaction.

These different types of relevance relationships exhibit the kinds of qualitative behavior one expects of common causes of an effect (converging links), independent effects of a common cause (diverging links), and nonsynergistic causal chains (serial links). For this reason, it has been argued that directed graphs are natural structures for representing knowledge about causal relationships in a domain (Pearl 2000, Spirtes et al. 2000). This makes Bayesian networks useful for applications such as planning and diagnosis that depend on the ability of models to reason appropriately about cause-and-effect relationships. However, Bayesian network arcs can represent relationships other than causality, including correlation, imputation, and credibility. The only required condition for a directed graph to encode a domain's dependence relationships is that each node (i.e., variable or logical expression of variables) is probabilistically independent of its nondescendants given its parents (e.g., Jensen 1996).

Fig. 3 shows how the Bayesian network model of our example can be extended to form a generic knowledge structure that can be applied to other individuals in other situations. The name 'Maria' has been replaced by the variable 'X', indicating that this Bayesian network applies to any individual in the population it represents. Fig. 3a applies this Bayesian network model to the available evidence for Maria's case. The probabilities of Fig. 3a apply to Maria, as well as to any individual in the domain who has a history of allergies, experiences a sneezing episode, sees scratches on the furniture, and does not know whether s/he has been exposed to a cold or whether or not there is a cat or other cause of allergic reaction nearby. Fig. 3b represents the same situation for an individual with no history of allergies, but who is known to have been exposed to a cold. Using the same Bayesian network with different evidence, it is concluded that Maria is probably suffering from allergies, whereas the second individual probably has a cold. The simpler Bayesian network of Fig. 2 can be obtained from the more complex network of Fig. 3 by conditioning, on the assumption that the person in question has a history of allergies. The Bayesian network of Fig. 1 can be obtained by a process called marginalization. A major research advance of the past decade has been the development of methods for ensuring incrementally constructed models remain

Variable	Parent Variable Values		True	False
Cold_Exposure_X			10%	90%
Has_Cold_X	Cold_Exposure_X=True		75%	25%
	Cold_Exposure_X=False		5%	95%
Allergy_History_X			1%	99%
Allergic_Reaction_X	Allergy_History_X=True	Cat_Near_X=True	75%	25%
		Cat_Near_X=False	1%	99%
	Allergy_History_X=False	Cat_Near_X=True	1%	99%
		Cat_Near_X=False	1%	99%
Sneezing_X	Allergic_Reaction_X=True	Has_Cold_X=True	99%	1%
		Has_Cold_X=False	99%	1%
	Allergic_Reaction_X=False	Has_Cold_X=True	99%	1%
		Has_Cold_X=False	5%	95%
Cat_Near_X			5%	95%
X_Sees_Scratches	Cat_Near_X=True		60%	40%
	Cat_Near_X=False		5%	95%

Figure 4
Local probability tables for sneezing model

consistent and are no more complex than necessary (Wellman et al. 1992, Goldman and Charniak 1993).

Fig. 4 shows the probability information for the Bayesian network of Fig. 3. The probability information in a Bayesian network is encoded as one or more local probability distributions at each node. There is a single probability distribution for each root node and a distribution for each combination of values of the parents of each non-root node. Additional structural assumptions can be applied within these local probability distributions to further simplify both knowledge acquisition and inference. Several authors have noted that a Bayesian network such as Fig. 3 can be defined by augmenting standard AI knowledge structures such as frames or objects to include information about probabilistic relationships. A probability model involving many variables can be represented as a collection of frames or objects, each encoding a small portion, or fragment, of the overall network. A mathematical foundation for such systems has been established by extending symbolic logic to incorporate probabilities, and then establishing conditions under which a knowledge base encodes a coherent probability model for problem instances in a given domain (e.g., Goldman and Charniak 1993).

Many common knowledge-based system architectures use a generic inference engine to apply a knowledge base of rules to facts stored in a database or in working memory. Facts about an entity in a domain can be represented as logical statements in existential/conjunctive form (Sowa 2000). For example, the facts in the cat allergy problem can be represented as the existentially quantified conjunction, viz., that there exists an individual in the domain, to whom the label 'Maria' is given, about whom the propositions Sneezing-Maria, Allergy-History-Maria, and Maria-Sees-Scratches, all take on the value 'True.' A standard knowledge-based system might have a Sneezing-Episode object type to organize information about sneezing episodes for individuals in the domain. Entering the fact Sneezing-Maria = True into the system triggers creation of an object instance Sneezing-Episode-Maria. The object would have attributes for each of the variables in the Bayesian network of Fig. 3. Traditional knowledge-based system architectures such as CLIPS or Prolog apply rules from the knowledge base to draw inferences about the causes of sneezing from the available evidence about Maria's case. By contrast, the Bayesian network described here calculates probabilities for the different causes by constructing the relevant portion of a globally consistent probability model of the application domain, and then applying a belief propagation algorithm.

4. *Belief Propagation*

Although rules are represented as universally quantified generalizations, most applications require the ability to qualify rules to allow for the possibility of exceptions. The Bayesian network of Fig. 3 can be thought of as a collection of rules with probabilistic qualifiers organized in a structure that ensures the existence of a consistent joint probability distribution over logical propositions related to the sneezing episode. The existence of such a consistent joint probability distribution guarantees that, at least in theory, any new information about entities in the domain can be incorporated and evidence can be propagated consistently to compute updated probabilities for the rest of the propositions of interest.

The propagation of evidence in a Bayesian network is bi-directional. Information can flow in the same direction as the arcs, as when evidence of exposure to a cold increases the probability of a cold, or in the opposite direction from the arcs, as when evidence of scratches increases the probability that there is a cat nearby. Furthermore, the direction of evidential flow is automatically performed by the solution algorithm.

A knowledge-based system whose generalizations consist of a complete and consistent set of Bayesian network fragments has 'possible worlds' semantics with an associated probability measure. That is, each assignment of values to the variables in the Bayesian network corresponds to a 'possible world', and the collection of local probability distributions implicitly specifies likelihood for each world. From this information, a probability can be computed for any logical proposition involving features of the world included in the Bayesian network.

Observation of evidence amounts to restricting the set of possible worlds to those consistent with the observed evidence. This results in a revised probability distribution obtained by applying the probability updating formula known as Bayes' Rule:

$$\frac{P(H_1|E,K)}{P(H_2|E,K)} = \frac{P(E|H_1,K)}{P(E|H_2,K)} \frac{P(H_1|K)}{P(H_2|K)}.$$

The left-hand side of this expression is the 'posterior odds ratio', which measures the relative likelihood of hypotheses H_1 and H_2 after observing evidence E, where K represents the information assumed in the reasoning context. The second factor on the right-hand side is the 'prior odds ratio', which measures the relative likelihood of the hypotheses prior to observing the evidence. The first factor on the right-hand side is the 'likelihood ratio', which measures how likely the evidence E is under each of the two hypotheses. This expression demonstrates that evidence E increases (decreases) the probability of H_1 relative to H_2 if it is more (less) likely to occur under H_1 than under H_2.

A knowledge-based system based on graphical models can implicitly encode an infinite number of probability statements. Research in knowledge-based model construction (Wellman et al. 1992) focuses on incremental construction of finite models to respond to specific queries. A typical query is to compute the probability of one or more target propositions given evidence in the form of facts about domain entities. The example above illustrates model construction in response to receipt of a sequence of facts about Maria's sneezing episode.

The probability measure over possible worlds can be used to express modal propositions about beliefs and knowledge of an agent. Beliefs of different agents having consistent ontologies (i.e., knowledge vocabularies and structures) can be represented by encoding different agent-specific probability distributions over the same set of possible worlds.

Incomplete probability distributions can be represented using interval probabilities or belief functions. Fuzzy logic enables the representation of ambiguous, as distinct from uncertain, truth-values (Zadeh 1975). Counterfactual or causal logic augments the semantics to include information about causality and the effects of manipulations. Temporal reasoning is represented by dynamic Bayesian networks. Planning problems can be treated by incorporating actions into graphical models. Partially observable Markov decision processes (POMDPs) are graphical probabilistic models that also represent actions available to agents and permit the computation of optimal actions given available information.

5. *Discovery, Learning, and Numerical Estimation*

Although the semantics of their associated numerical measures vary, all certainty calculi require prior numerical valuations. For example, Bayesian networks require prior conditional probabilities; fuzzy logic requires a priori fuzzy set membership estimates; Dempster-Shafer calculus requires estimates of prior values for basic probability masses. The need to acquire and calibrate prior numerical estimates is both a challenge and a standard argument against the use of belief calculus. There are several answers to this challenge.

First, any source of exemplar cases provides a source of prior numerical estimates. Second, a history of successful knowledge engineering efforts has demonstrated that, at least for certain sorts of variables and problems, humans can provide the necessary judgments to build models that perform well on a variety of tasks (e.g., Heckerman 1991). Third, techniques for data mining and for learning graphical models from data, or from a combination of expert judgment and data (Jordan 1999; See *Bayesian Statistics*), can be applied to discover numerical estimates 'on the fly.' In particular, Bayesian learning methods have been developed that exhibit both accuracy and simplicity suggesting, via an Occam's razor argument, that researchers are successfully meeting this technical challenge (see *Hierarchical Models: Random and Fixed Effects*).

6. *Conclusion*

Since 1980 there have been major advances in automating inductive reasoning in AI systems. Modular, tractable, and cognitively plausible representations for uncertainty in AI systems have been developed and implemented in widely available software tools. Uncertainty management is widely recognized as a science

and corresponding engineering technology is required for effective knowledge representation and reasoning architectures for all AI systems. The capabilities achieved so far point the way to a much broader set of issues and objectives, many of which concern the possibilities of automated acquisition of inductive knowledge relationships. Graphical network representations combined with data mining, automated learning, and other recent developments are a natural computational environment in which to develop scientific hypotheses and model experimental and other evidential reasoning in support of complex hypotheses. Uncertainty management-driven advances in automated inductive reasoning raise the hope that AI systems might soon accelerate the rate of scientific discovery.

The web site of the Association for Uncertainty in Artificial Intelligence, is a good source of up-to-date information on topics related to uncertainty in artificial intelligence. This site contains links to the online *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, which contains abstracts for all papers since the conference started in 1985, as well as soft-copy versions of many papers. The web site also contains links to a number of tutorials on subjects related to uncertainty in AI. Good introductions to Bayesian networks can be found in the seminal work by Pearl (1988), the more recent books by Jensen (1996) and Cowell (1999), and a good nontechnical introduction by Charniak (1991). A review of recent work in decision theoretic methods in artificial intelligence is provided by Haddawy (1999). A recent overview of fuzzy set methodology is presented in Dubois and Prade (2000). Smets and Gabbay (1998) describes different approaches to representing uncertainty. The anthology by Pearl and Shafer (1990) compiles a number of seminal articles in the field of uncertainty in artificial intelligence.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence in Cognitive Science; Bayesian Graphical Models and Networks; Graphical Models: Overview; Scientific Discovery, Computational Models of

Bibliography

Buchanan B, Shortliffe E 1974 *Rule-Based Expert Programs: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA

Charniak E 1991 Bayesian networks without tears. *AI Magazine* **12**(4): 50–63

Cheeseman P 1985 In defense of probability. In: *Proceedings of the International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers, San Mateo, CA, pp. 1002–9

Cowell R (ed.) 1999 *Probabilistic Networks and Expert Systems*. Springer-Verlag, Berlin

Dubois D, Prade H 2000 *Fundamentals of Fuzzy Sets*. Kluwer Academic, Dordrecht, The Netherlands

D'Ambrosio B 1999 Inference in Bayesian networks. *AI Magazine* **20**(2): 21–36

Goldman R P, Charniak E 1993 A language for construction of belief networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**(3): 196–207

Haddawy P 1999 An overview of some recent developments in Bayesian problem solving techniques. *AI Magazine* **20**(2): 11–9

Heckerman D 1986 A *Probabilistic Interpretation for MYCIN's Certainty Factors. Uncertainty in Artificial Intelligence 1*. North-Holland, Amsterdam, pp. 167–96

Heckerman D 1991 *Probabilistic Similarity Networks*. MIT Press, Boston, MA

Henrion M 1987 Uncertainty in artifical intelligence: Is probability epistemologically and heuristically adequate? In: Mumpower J L, Philipps L D, Renn O, Uppuluri V R R (eds.) Expert Judgment and Expert Systems, Nato ASI Series F, Vol. 35. Springer-Verlag, Berlin, pp. 105–30

Jensen F 1996 *An Introduction to Bayesian Networks*. Springer-Verlag, New York

Jordan M I (ed.) 1999 *Learning in Graphical Models*. MIT Press, Cambridge, MA

McCarthy J 1960 Recursive functions of symbolic expressions and their computation by machine. *Communications of the ACM* **7**: 184–95

Ngo L, Haddawy P, Helwig J 1995 A theoretical framework for context-sensitive temporal probability model construction with application to plan projection. In: *Uncertainty in AI, Proceedings of the Twelfth Conference*. Morgan Kaufmann Publishers, San Mateo, CA, pp. 419–26

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA

Pearl J 2000 *Causality: Models, Inference and Reasoning*. Cambridge University Press, Cambridge, UK

Pearl J, Shafer G (eds.) 1990 *Readings in Uncertain Reasoning*. Morgan-Kaufmann, San Mateo, CA

Rumelhart D, McClelland J, PDP Group 1987 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA

Shafer G 1976 *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ

Shafer G, Shenoy P 1990 probability propagation. *Annals of Mathematics and AI* **2**: 327–52

Smets P, Gabbay D (eds.) 1998 *Handbook of Defeasible Reasoning and Uncertainty Management Systems: Quantified Representation of Uncertainty and Imprecision* (Vol. 1), Kluwer Academic, Dordrecht, The Netherlands

Sowa J 2000 *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks-Cole, London

Spirtes P, Glymour C, Scheines P 2000 *Causation, Prediction, and Search*. MIT Press, Boston, MA

Wellman M P, Breese J S, Goldman R P 1992 From knowledge bases to decision models. *The Knowledge Engineering Review* **7**(1): 35–53

Zadeh L 1975 Fuzzy logic and approximate reasoning. *Synthese* **30**: 407–28

K. B. Laskey and T. S. Levitt

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Artificial Neural Networks: Neurocomputation

A neurocomputer is an abstract mathematical neural network model capable of computing solutions to specific information processing problems. Neurocomputers are usually referred to as 'neural networks,' 'artificial neural networks (ANN),' 'connectionist systems,' or 'parallel distributed processing (PDP) systems.' Neurocomputers are important to neuroscientists since they may implement specific theories of information processing in brain subsystems. The behavior of a neurocomputer may also be systematically compared with animal and human behavior within restricted behavioral task domains in order to explore the implications of a particular psychological theory. Finally, neurocomputers are important because they may suggest novel approaches to solving data analysis and computational problems which occur in an engineering context. This article introduces neurocomputation from a historical perspective.

The typical neurocomputer consists of a collection of *units* where each unit is assumed to be a highly abstract model of either a neuron, a portion of a neuron, or a group of neurons. Each unit has a state represented by a real number which is referred to as the unit's *activity level* which is typically assumed to be abstractly related to neuronal firing frequency (seeFig. 1). The set of activity levels of all units in the network is called an *activation vector* whose magnitude corresponds to 'signal strength' and whose direction corresponds to 'signal identity.' Learning in a neurocomputer usually involves some algorithm for adjusting connections among the units as a function of the neurocomputer's past experiences. Figures 2 and 3 illustrate two typical neurocomputer architectures.

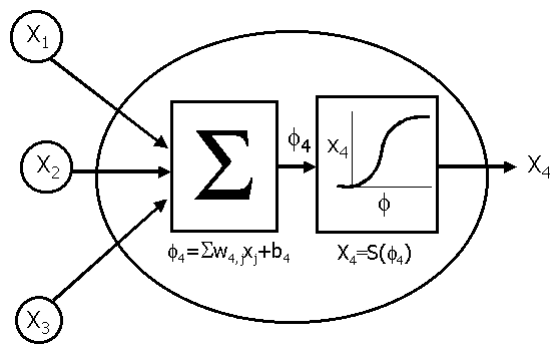

Figure 1
A formal neuron common to many neurocomputer architectures. The *net input* to the ith unit, ϕ_i, is given by the formula: $\phi_i = \Sigma w_{ij}\, x_j + b_i$ where w_{ij} is the synaptic connection weight associated with the jth input to unit i and x_i is the activation level of the ith unit in the network. The quantity b_i is often referred to as the *bias* for the ith unit. Choosing S such that $S(\phi_i) = 1$ if $\phi_i > 0$ and $S(\phi_i) = 0$ otherwise gives a popular version of the McCulloch–Pitts formal neuron. Choosing S to be a differentiable monotonically nondecreasing (i.e., 'sigmoidal') function yields the 'quasilinear' formal neuron

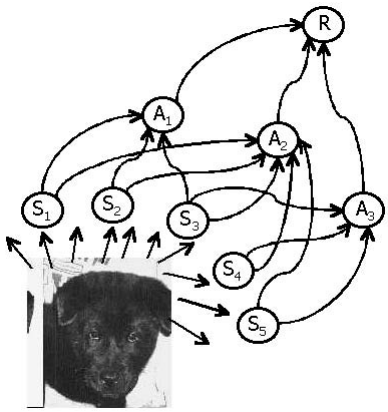

Figure 2
An example 'feed-forward' neurocomputer architecture. In Rosenblatt's (1962) generic 'feed-forward' perceptron neurocomputer, incoming information (e.g., light intensities generated from a photographic image) is detected using an array of input units called 'sensor units' or 'S-units.' The S-units are then randomly connected through unmodifiable connection weights to a set of hidden units called 'association units' or 'A-units.' The A-units are then connected to the output unit called the 'response unit' or 'R-unit.' All units in the perceptron were McCulloch–Pitts formal neurons. The 1986 Rumelhart, Hinton, and Williams generic 'feed-forward' back-propagation network architecture is essentially identical to the perceptron architecture except: (a) the units are differentiable sigmoidal functions (see Fig. 1), and (b) the input to hidden unit connection weights are modifiable through a learning process as well

1. *The Origins of Neurocomputation (1890–1965)*

1.1 *Hebbian Learning*

In 1890, the psychologist William James argued that it is more fruitful to consider associations between 'brain-processes' (and not 'ideas') as the basis for human associative memory. Information retrieval, according to James, was proposed to arise by a process where:

> The amount of activity at any given point in the brain-cortex is the sum of the tendencies of all other points to discharge

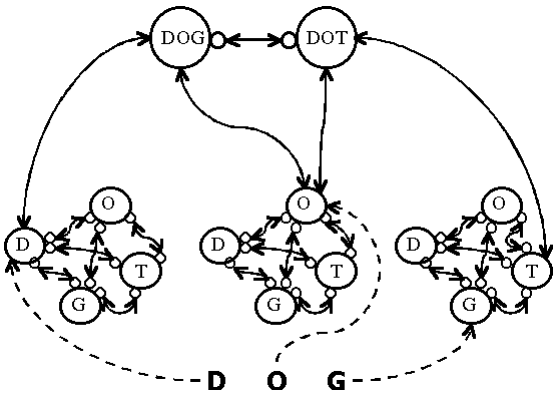

Figure 3
An example 'recurrent' neurocomputer architecture. In contrast to the 'feed-forward' architecture shown in Fig. 2, feedback connections among units are permitted as well. Activation is passed back and forth among the units until the activation pattern over all units 'stabilizes.' This figure depicts the architecture used in the 1986 Rumelhart and McClelland Interactive Activation Model (see Anderson and Rosenfeld 1988a for further details). Notice how the recurrent connections permit letter string orthographic structure to influence perception of the component individual letters

> into it…(James 1890 cited in Anderson and Rosenfeld 1988a, p. 5).

Moreover, William James also suggested that information such as the past history of intensity and co-occurrence of such brain processes influenced actual physical changes in the brain such that:

> When two elementary brain-processes have been active together or in immediate succession, one of them, on re-occurring, tends to propagate its excitement into the other (James, 1890 cited in Anderson and Rosenfeld 1988a, p. 5).

In many respects, these ideas can be viewed as marking the origins of the concept of neurocomputation since James was providing a sketch of how mechanisms for storing and retrieving information in the brain might operate.

An elaboration of these ideas was developed a half century later by the psychologist Donald Hebb in 1949 using a 'neuron model.' A *neuron* consists of incoming or 'afferent' fibers and an outgoing fiber called the *axon*. Hebb knew that a voltage pulse could be propagated along the neuron's axon in situations where the neuron's afferent fibers were connected to axons of other neurons that were propagating voltage pulses. The term *firing* was used to refer to the process of voltage pulse propagation in the neuron.

These neuroscience concepts appeared in Donald Hebb's famous Hebbian learning rule which expressed the now common neurocomputer learning principle.

> When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes place in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased (Hebb 1949 cited in Anderson and Rosenfeld 1988a, p. 50).

1.2 McCulloch–Pitts Formal Neurons

McCulloch and Pitts in 1943 (see Anderson and Rosenfeld 1988a) proposed an abstract mathematical model of a neuron which today is referred to as a 'McCulloch–Pitts neuron.' A McCulloch–Pitts neuron (MP-neuron) has two states: 'active' and 'inactive.' Physiologically, the activation level of an MP-neuron measured the likelihood the MP-neuron would generate at least a certain number of voltage impulses within some fixed time frame.

As illustrated in Fig. 1, the MP-neuron summed its active excitatory synaptic inputs and compared that sum to some internal threshold value. If the summed activity of its active excitatory synaptic inputs exceeded the MP-neuron's threshold value and no inhibitory synaptic input to the MP-neuron was active within some given timeframe, then the MP-neuron would become active and could consequently activate excitatory and inhibitory synaptic inputs of other MP-neurons. One key theoretical result obtained by McCulloch and Pitts was that deductive logic could be implemented by networks of MP-neurons. This theoretical result had an important impact on the field of neurocomputation and strongly influenced the conceptual development of the modern digital computer concept.

1.3 Learning Networks of McCulloch–Pitts Neurons

By the mid-1960s several researchers such as Frank Rosenblatt, Bernard Widrow, Oliver Selfridge, Nils Nilsson, and E. R. Caianiello were focusing on the problem of building networks of MP-neurons which could learn (see Anderson and Rosenfeld 1988a; Anderson et al. 1990). Since McCulloch and Pitt had pointed out that a network of MP-neurons could realize any arbitrary logical function, it would seem useful to postulate learning mechanisms that could create appropriate networks of MP-neurons for different statistical environments.

Rosenblatt was a psychologist who analyzed and simulated multilayer networks of MP-neurons capable of 'learning.' He called such networks *perceptrons*. Although Rosenblatt discussed many different types of perceptrons, the 'generic perceptron' (as shown in Fig. 2) consisted of a set of *S-units* (i.e., 'sensor units'

or 'input units') whose outputs were randomly connected to the inputs of a set of *A-units* (i.e., 'association units' or 'hidden units'). The outputs of the A-units, in turn, were connected to an *R-unit* (i.e., the 'response unit' or 'output unit').

The connections from the S-units to the A-units were nonmodifiable and chosen according to a particular probabilistic law. The connections from the A-units to the R-unit could be modified by a learning process called the *perceptron learning rule*. The learning rule worked as follows. First, the perceptron would generate an activation level for the R-unit based on the current stimulus. The R-unit activation level, R, would be equal to either zero or one. The *desired response D* for the R-unit (generated by the 'training data') is also equal to zero or one. Let A_i be the activation level of the ith A-unit. The connection weight from the ith A-unit to the R-unit would then be incremented by an amount $\gamma(D–R)A_i$. The positive number γ is called the *learning rate* and the quantity $(D–R)$ is often called the *error signal*. Note that no weight adjustments are made when the perceptron makes the correct classification (i.e., $D = R$).

An important consequence of the perceptron learning rule was the famous *perceptron convergence theorem* (see Anderson and Rosenfeld 1988a and Minsky and Papert 1969 for reviews) which essentially stated that if at least one network architecture exists to solve a particular classification problem, then the perceptron learning rule was guaranteed to find a solution after a sufficiently large (but finite) number of training samples have been presented! This theoretical result was strengthened further by the work of McCulloch and Pitts which basically had shown that the multilayer perceptron architecture consisting of S-units, A-units, and an R-unit was sufficiently rich (given enough correctly wired A-units) to represent the solution to any possible classification problem! Such theoretical results in conjunction with extensive empirical research provided an exciting and stimulating environment for research in neurocomputation (Anderson and Rosenfeld 1988a, Anderson et al. 1990).

2. Neurocomputation: The Dark Ages (Late 1960s–Early 1980s)

By the late 1960s, research in the field of neurocomputation was less popular. The early (and important) successes of the field had been excessively popularized, and promises had been made and broken. Rosenblatt's research, in particular, had been oversold by the popular press. By the late 1960s, many researchers were more interested in attacking (rather than developing) a methodology for neurocomputation. For example, in their well-known book titled *Perceptrons*, Minsky and Papert (1969) formally explored the computational limitations of a single generalized MP-neuron but their chosen book title implied that the generic multilayer perceptron network architecture was being discussed! The time period beginning in the late 1960s might be viewed as the 'dark ages' of neurocomputing in the sense that a climate of skepticism regarding neural computation prevailed. However, this same time period was an amazingly productive period of research as well, where many of the key fundamental ideas of the field were initially developed.

2.1 Quasilinear Formal Neurons and Learning

By the late 1960s, the biological assumption underlying generalized MP-neurons began to be reconsidered. First, neuroscientists such as Vernon Mountcastle had shown that the *firing rate* of a sensory neuron (i.e., the number of voltage pulses generated per second) in response to a stimulus could sometimes be realistically modeled as a linear function of the stimulus magnitude. Experiments of this type suggested that the 'relevant informational state' of a neuron might be better modeled as a continuous numeric quantity (i.e., a real number) as opposed to a categorical binary variable. Second, work by Ratliff, Hartline, and others had shown that the peripheral visual system of the horseshoe crab could be effectively modeled using linear systems theory. Research of this nature encouraged modelers to consider a 'quasilinear' continuous-state neuron model (as opposed to the 'two-state' MP-neuron model) for the purposes of neurocomputation. Figure 1 illustrates the key idea underlying the quasilinear neuron model which is essentially that the change in the model neuron's activation level can be approximately modeled as a linear transformation of the activation levels of the other model neurons in the system. In the early 1970s the quasilinear neuron model became relatively popular and researchers such as James Anderson, S. I. Amari, Stephen Grossberg, and Teuvo Kohonen independently published papers concerning Hebbian learning in networks of abstract quasilinear neurons (see Anderson and Rosenfeld 1988a, Grossberg 1982, Levine 1991, and Lau 1992, for relevant reviews).

2.2 Recurrent Networks of Quasilinear Neurons

In the late 1960s, neuroscientists had obtained experimental evidence suggesting that some biological neural networks were organized in a *recurrent on-center off-surround anatomy*. In this configuration, an excited neuron activates a few selected neurons while simultaneously 'inhibiting' a large group of neighboring neurons (see Fig. 3). Thus, in some cases, positive as well as negative feedback is present. This combination of positive and negative feedback implements

a general type of neural categorization mechanism where the network is constructed such that only a small number of neurons are permitted to respond to a stimulus, while the responses of the remaining neurons tend to be diminished in magnitude. Following the research in the early 1970s concerned with associative linear networks involving Hebbian-like learning mechanisms, a number of researchers began to consider problems of neurocomputation using recurrent networks (such as on-center off-surround) of quasilinear neurons. This research led to both theoretical and simulation work which showed how recurrent networks could function as 'feature detectors,' 'pattern amplifiers,' and 'categorization mechanisms.'

2.3 Self-organizing Maps

It had been known since the early 1960s that: (a) feature detection units existed in visual cortex, and (b) feature detection units with similar response properties tended to physically cluster together. Malsburg in 1973 (see Anderson and Rosenfeld 1988a) hypothesized that the development and organization of such feature detection cortical cells might be understood in terms of a learning process. To examine this hypothesis, Malsburg simulated a highly simplified mathematical model of the visual cortex that shared important similarities with concurrent theoretical work by Grossberg (Anderson and Rosenfeld 1988a, Grossberg 1982).

The key ideas of the Malsburg computer simulation model were as follows. First, the initial pattern of neuronal connectivity was random. This initial random connectivity pattern is important because it implicitly provides a set of 'biases' encouraging different cells in the simplified formal model of the visual cortex to have different response characteristics. Second, a Hebbian-type learning rule was instantiated which increased the strength of a connection between two cells when both cells became simultaneously active within some timeframe. Constraints on the weight vector for each neuron were introduced in order to prevent the weights from growing without bound.

From a computational perspective, the initial random connectivity principle and the Hebbian learning principle are sufficient to account for the development of neurons which respond to specific features in their environment. These two principles by themselves, however, cannot account for the qualitative finding that feature detection neurons with similar response properties tend to cluster together in the visual cortex. The third essential property of Malsburg's model was that his model incorporated an on-center, off-surround architecture which had the effect of forcing physically close feature detection neurons to have similar response properties.

3. The Renaissance of Neurocomputation (1980–90)

By the early 1980s, enough progress in the field of artificial intelligence had been made for many researchers to appreciate the strengths as well as the limitations of methodologies entrenched in deductive logic. Second, the cost of computing resources was dropping rapidly. This meant that the implications of complex mathematical neural models could be systematically explored through simulation studies. Thus, obscure mathematical neural models suddenly became more accessible and understandable to a much larger scientific community through the use of computer simulation examples. Moreover, using the new computer simulation technology, new ideas regarding the capabilities and limitations of a potential mathematical model could be explored long before complete or even partial mathematical analyses had been accomplished.

3.1 Recurrent Networks of Sigmoidal Neurons

In 1981, the experimental psychologists McClelland and Rumelhart proposed a neurocomputer for letter perception which had the behavioral property, common to human behavior, that letters in words were perceived more efficiently than letters by themselves (Anderson and Rosenfeld 1988a). The model was designed using an on-center, off-surround anatomy (similar to Grossberg's late-1970s work) such that activation of a concept resulted in decreasing the activation of its competing concepts (see Fig. 3).

McClelland and Rumelhart demonstrated, through a series of careful and systematic computer simulation studies, that the patterns of errors generated by the mechanism were both qualitatively and quantitatively similar to both old and novel experimental data. The McClelland and Rumelhart model eventually became a highly influential example of how a computer simulation methodology could be exploited to empirically evaluate the performance of a neurocomputer from a purely behavioral perspective.

In 1982, Hopfield (see Anderson and Rosenfeld 1988a) showed how symmetrically connected recurrent networks composed of MP-neurons could be mathematically analyzed using Lyapunov function techniques from nonlinear dynamical systems theory (see Golden 1996 for a relevant discussion of Lyapunov function methods; also see *Optimal Control Theory*; *Self-organizing Dynamical Systems*). The Lyapunov function technique was shown to provide a methodology for expressing the computational goal of a recurrent network as minimizing some nonlinear objective function. Shortly thereafter, similar analyses of closely related networks of quasilinear neurons (e.g., Cohen–Grossberg, Cohen–Grossberg–Kosko,

BSB, Hopfield, Harmony Theory, and Boltzmann neural networks) were analyzed using Lyapunov function methods (see Anderson and Rosenfeld 1988a, Anderson et al. 1990, Golden 1996 for relevant reviews).

3.2 Self-organizing Maps

Kohonen (1984) discussed a highly abstract version of Malsburg's (1973) self-organizing map model whose computational performance was comparable to the original Malsburg neural model. From a neuroscience perspective, this was informative because it emphasized those aspects of the neuroscience model which were especially relevant for realizing a useful information processing model. From an engineering perspective, this was interesting because the resulting mathematical model was much simpler and more amenable to computer implementation and mathematical analysis. Kohonen applied his model to a variety of engineering problems concerned with the problem of extracting features for speech recognition systems.

In 1987, Carpenter and Grossberg introduced into the literature what is now called 'fuzzy' Adaptive Resonance Theory (ART). In the ART model, a set of input units is initially randomly connected to a set of hidden units. When a stimulus is presented to the input units, the hidden units compete among themselves until the maximally active hidden unit inhibits the activation levels of all other hidden units. The winning unit then generates a feedback signal to the set of input units. If the feedback signal agrees with the incoming stimulus pattern, then 'resonance' occurs and the connections from the input units to the winning hidden unit are adjusted so the stimulus is more likely to activate the winning hidden unit in the future. If no agreement occurs, then a 'reset operation' occurs and the winning unit is temporarily disabled and the process repeats itself. A useful feature of the ART methodology is that new units can be recruited on an as-needed basis to solve a given pattern recognition problem. In addition, unlike many other neurocomputing architectures, the ART methodology incorporates an 'attentional component' which plays a role in determining whether or not resonance in ART has occurred.

3.3 Multilayer Network Architectures

An important difficulty with the original generic perceptron architecture was that the connections from the input units to the hidden units (i.e., the S-unit to A-unit connections) were randomly chosen. Thus, in many important situations, the chances of obtaining a useful network architecture were relatively small. Rumelhart, Hinton, and Williams in 1986, Le Cun in 1985, Parker in 1985, Werbos in 1974, and Amari in 1967 independently contributed key components to the solution of this problem which is now referred to generically as 'back-propagation' or 'multilayer perceptron' learning methods (for relevant reviews see Chauvin and Rumelhart 1995, Golden 1996, Werbos 1994). The essential idea of such approaches involved 'smoothing' the sigmoidal nonlinearity of the generalized MP-neuron by substituting a differentiable sigmoidal function (see Fig. 1).

Using this device, a standard nonlinear optimization method for differentiable objective functions known as 'gradient descent' could be used to simultaneously estimate the input to hidden unit and the hidden unit to output unit connections. In order for this approach to work, it was also necessary that the initial pattern of connections among the units be randomly chosen so that (like the Malsburg's 1973 model) different hidden units would be biased to detect different features in the input pattern set.

4. Modern Neurocomputation

Today, neurocomputer architectures have proven to be useful tools for developing, evaluating, and modifying theories in the fields of both neuroscience and psychology (for relevant reviews of both neuroscience and psychology research see Anderson 1995, Anderson and Rosenfeld 1988a, Anderson et al. 1990, Durbin et al. 1989, Ellis and Humphreys 1999, Golden 1996, Grossberg 1982, Levine 1991, McClelland and Rumelhart 1986). For example, Kleinfeld and Sompolinsky in 1988 proposed a biologically plausible neurocomputer model of the swimming escape response in the sea mollusk *Tritonia* (see Golden 1996 for a review). Grajski and Merzenich in 1990 (in an extension of the previous work by von der Malsburg in 1973) have described a self-organizing biologically plausible neurocomputer for modeling the development of somatosensory topographic maps in the adult monkey (see Golden 1996 for a review).

Kruschke in 1992 developed a neurocomputer for modeling qualitative aspects of human category learning (see Ellis and Humphreys 1999; see *Connectionist Models of Concept Learning*). In 1991, Plunkett and Marchman developed a neurocomputer type architecture for explaining patterns of 'undergeneralization' and 'overgeneralization' in child language acquisition (see Ellis and Humphreys 1999; see *Connectionist Models of Development*).

Neurocomputer architectures have also made important contributions towards the development of new approaches to solving complex optimization and pattern recognition problems in real-world engineering contexts (for relevant reviews see Anderson 1995, Anderson and Rosenfeld 1988a, Anderson et al. 1990,

Golden 1996). Neurocomputer architectures have been successfully applied to solving real-world engineering problems in areas such as image analysis, data analysis, robot control, adaptive signal processing, financial prediction, medical prediction, inductive reasoning, and speech recognition (for additional examples see Anderson and Rosenfeld 1988a, Anderson et al. 1990, Dagli et al. 1991, Golden 1996). In addition, appropriate techniques for proving theorems regarding the behavior of neurocomputers are becoming more widely recognized. Such techniques are crucial for not only guaranteeing 'correctness of design' in engineering applications but also for advancing our theoretical understanding of neurocomputer-based theories of the mind and brain (see Golden 1996, for an introduction to such methods).

Finally, it is important to emphasize that the field of neurocomputation is still in its infancy. Indeed, it is likely that the diversity of neurocomputer architectures will continue to increase as our understanding of the neural basis of computation evolves in future years. Highly abstract models of neurocomputation will continue to be developed and will continue to play important roles in theory development in psychology and neuroscience. The brain is extremely complex and it is clear that multiple neurocomputer models associated with multiple overlapping levels of abstraction will play a critical role in theory development in psychology, neuroscience, and neuropsychology (see *Connectionist Approaches*).

Moreover, as our understanding of the mathematical basis of neurocomputation increases, the formal implications of specific neurocomputer modeling assumptions will become more apparent. Golden (1996; also see Smolensky et al. 1996) reviews mathematical methods for neurocomputer analysis and design showing how such methods are closely related to previous and current work in theoretical mathematical psychology (see *Mathematical Psychology*; *Mathematical Psychology, History of*). Extremely detailed models of neurocomputation (Bower and Beeman 1998) which incorporate detailed knowledge of the neurophysiology will also play an increasingly larger role in theory development and evaluation in neuroscience and psychology. Another new but promising research trend will be the role of genetic algorithms for formally exploring how evolutionary pressures might influence the development of mechanisms for neurocomputation (Husbands and Meyer 1998).

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Connectionist Approaches; Connectionist Models of Concept Learning; Connectionist Models of Development; Neural Networks and Related Statistical Latent Variable Models; Neural Networks: Biological Models and Applications

Bibliography

Anderson J A 1995 *An Introduction to Neural Networks*. MIT Press, Cambridge, MA

Anderson J A, Pellionisz A, Rosenfeld E 1990 *Neurocomputing 2: Directions for Research*. MIT Press, Cambridge, MA

Anderson J A, Rosenfeld E 1988a *Neurocomputing*. MIT Press, Cambridge, MA

Anderson J A, Rosenfeld E 1998b *Talking Nets: An Oral History of Neural Networks*. MIT Press, Cambridge, MA

Bower J M, Beeman D 1998 *The Book of Genesis: Exploring Realistic Neural Models with the General Neural Simulation System*, 2nd Edn. TELOS, Santa Clara, GA

Chauvin Y, Rumelhart D E 1995 *Backpropagation: Theory, Architectures, and Applications*. Erlbaum, Hillsdale, NJ

Dagli C H, Kumara S R, Shin Y C 1991 *Intelligent Engineering Systems Through Artificial Neural Networks*. ASME Press, New York

Durbin R, Miall C, Mitchison G 1989 *The Computing Neuron*. Addison-Wesley, Reading, MA

Ellis R, Humphreys G 1999 *Connectionist Psychology: A Text with Readings*. Psychology Press, Basingstoke, UK

Golden R M 1996 *Mathematical Methods for Neural Network Analysis and Design*. MIT Press, Cambridge, MA

Grossberg S 1982 *Studies of Mind and Brain*. Reidel, Boston, MA

Hassoun M H 1995 *Fundamentals of Artificial Neural Networks*. MIT Press, Cambridge, MA

Husbands P, Meyer J A 1998 *Evolutionary Robotics*. Springer-Verlag, New York

Kohonen T 1984 *Self-Organization and Associative Memory*. Springer-Verlag, New York

Lau C 1992 *Neural Networks: Theoretical Foundations and Analysis*. IEEE Press, New York

Levine D S 1991 *Introduction to Neural and Cognitive Modeling*. Erlbaum, Hillsdale, NJ

Malsburg C 1973 Self-organization of orientation sensitive cells in the striata cortex. *Kybernetik* **14**: 85–100

McClelland J L, Rumelhart D E 1986 *Parallel Distributed Processing, Volume 2: Psychological and Biological Models*. MIT Press, Cambridge MA

Minsky M L, Papert S A 1969 *Perceptrons*. MIT Press, Cambridge, MA

Rosenblatt F 1962 *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington DC

Rumelhart D E, McClelland J L 1986 *Parallel Distributed Processing, Volume 1: Foundations*. MIT Press, Cambridge MA

Smolensky P, Mozer M C, Rumelhart D E 1996 *Mathematical Perspectives on Neural Networks*. Erlbaum, Mahway, NJ

Werbos P J 1994 *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. Wiley, New York

R. M. Golden

Artificial Social Agents

A fundamental question in the social sciences is what makes human beings social? What is the root of sociality? Today, computational models are being used to address this issue. Computational social theorists

are not simply providing descriptions of society at a particular point in time. Rather, they are in the business of providing insight into the fundamental basis of socialness, understanding of how societies change, and how individuals negotiate, constrain, enable, and are constrained by the social world in which they live. They are in the business of explaining the dynamics by which social agents, groups, teams, organizations, societies, cultures evolve and co-evolve.

Computational analysis is playing an increasingly important role in the development of our understanding of what it means to be a social agent (Weiss 1999, Bainbridge et al. 1994). Computational analysis using social agents has been used to theorize about a large number of social behaviors. These include but are not limited to: organizational exploration and exploitation, cooperation, coordination, diffusion and social evolution, organizational adaptation, change in social networks and exchange networks, collective action, emotions, markets, and transactive memory. This exploration provides insight into the fundamental nature of the social agent (Carley and Newell 1994) and the laws by which artificial societies operate. The conceptualizations coming out of computational social science are building on, yet moving us far beyond, traditional conceptualizations of the social agent. Such traditional conceptualizations include Rousseau's noble savage, *homo economicus*, Skinner's contingently reinforced human, Simon's boundedly rational human (see *Bounded and Costly Rationality*), the imperfect statistician, Mead's social symbolic human, Blau's bundle of sociodemographic parameters human, the structuralist's human as a social position, *homo faber* (human, the toolmaker), and Huzinga's *homo ludens* (playful human).

The work on artificial agents asks—what is the fundamental basis of socialness? You might assume that a computational approach to socialness will become trapped in the simplicity of machine-like systems. However, exactly the opposite is the case. Current theories of human cognition, based on information-processing notions, are in many ways quite rich. Such theories include SOAR, ACT, and neural networks. Moreover, these formalisms provide an analytical tool to ask critically what more is needed to attain social behavior? Computational models allow us to: peel away what is understood about cognition, and consequently reveal the basis of socialness; theorize about the possible joint effects on social behavior of cognitive architectures and complex social environments; and examine dynamic social systems.

Artificial social agents are being used in all branches of the social and organizational sciences. These agents range from simple cellular automata-type agents to detailed and highly cognitively realistic agents (such as SOAR agents). Artificial social agents range from: whether or not they can learn and adapt; the level of knowledge with which they are endowed; and the type of social activities in which they can engage. In a sense, this work can be divided into three categories—A-life, organizational, and cognitive. In the A-life models (artificial life) there are large numbers (hundreds to thousands) of simple agents (few rules) interacting in a two-dimensional space and collectively generating socially realistic behaviors (such as aggression and segregation) (Epstein and Axtell 1997). The Sugarscape and Schelling agents are of this type. In the organizational models, there are a moderate number (tens to hundreds) of moderately cognitively complex (some rules) agents. Each of these agents has a specific organizational position, which constrains and enables access to information. The agents collectively perform organizational tasks (such as resource assignment, classification, or search) theory (Carley and Prietula 1994). The VDT, Orgahead, and Construct agents are of this type. In the cognitive models, there are a few (two to 10) highly cognitively realistic agents (many rules) each with specific detailed expertise and often a specific task-based role. Examples are the SOAR-IFOR agents and Plural-SOAR agents.

1. *The Model Social Agent*

Carley and Newell (1994) construct a definition of the social agent—a Model Social Agent. They developed their definition by asking the question, 'What is necessary in order to build an adequate artificial social agent?' They describe the social agent using two dimensions: processing capabilities and differentiated knowledge of the self, task domain, and environment. The processing dimension reflects limitations to the agent's capabilities that enable it to be social. The knowledge dimension reflects the richness of the agent's perceived environment that evokes and supports social behavior. As the agent's information-processing capabilities become more limited and the agent's knowledge becomes more complex (both in type and in quantity) the Model Social Agent is achieved. As the agent's information processing capabilities decrease, task performance may degrade, but different, and often more complex, behaviors emerge. Basically, agents with more information processing capabilities do not need certain behaviors, for example, the omnipotent agent does not need to gather information. Regardless of an agent's capabilities, if its situational knowledge is impoverished, it will not behave as a social agent. For knowledge, there is a cumulative sequence of information types. The agent can engage in distinct actions as it gains more types of information. In Table 1, the Newell–Carley Scheme is displayed and common agent types placed in the relevant location.

2. *Social Agents in the Loop*

Computational agents have served as 'conspirators' and agents in human laboratory experiments. For

Table 1
Model social agent and artificial social agent models

	Knowledge (increasingly rich situation)					
Information processing capability	Nontask specific	Multiple agents	Real interaction	Social–structural	Multiple–goals	Cultural–historical
Omniscient agent		Game theory				
Rational agent		Economics rational choice	Tit for tat	Schelling segregation NK, NKC Exchange theory VDT		Sugarscape
Boundedly rational agent	Expert systems Chess programs Behavioral decision theory	Exploration and exploitation		Garbage can construct	AAAIS Orgahead	
Cognitive agent	SOAR SDML ACT-R Neural networks	Multiagent SOAR				
Emotional cognitive agent		Affect control theory				Model social agent

example, in the artificial markets project at MIT (http://web.mit.edu/lfe/www/Research/artificial-mkts.html), researchers are running experiments with both humans and artificial social agents in order to examine the rich dynamics arising from interactions between human and artificial agents in a stochastic market environment. This work is at the very heart of electronic commerce.

Whether or not artificial agents are used in the laboratory, social agent models can be used to fine tune human experiments. For example, Markovsky (1987) used a model of triads of simple social agents to explore the power associated with position in a three-person social network, where person A could interact with persons B and C, who however could not interact with each other. In a typical round in this virtual experiment, each agent makes an offer of dividing 24 points between itself and another agent. Agent A compares the offers of B and C, and selects one of them. The points are then divided between A and the selected partner using the average of the B and C's offers. After the first round (when the offers are randomly determined), each agent adjusts its offer as follows. If the previous offer was accepted, the new offer will be more demanding. If the previous offer was rejected, the new offer will be less demanding. Given these simplistic agents, Markovsky was able to build a series of 19 experiments that varied the strategies employed by bargaining agents. These results then informed laboratory experiments, and directed future research in exchange theory (see also Markovsky et al. 1993).

Computational agents have served as players in computer tournaments. An example is the 'prisoner's dilemma' computer tournament organized by Robert Axelrod. The prisoner's dilemma is a classic game-theory problem that explores the conditions under which cooperation may arise between self-interested agents who have the potential to garner short-term gains if they violate agreements to cooperate (Rapoport and Chammah 1965). Researchers submit artificial social agents—computational models—that do the 'task' of the tournament. Tournaments enable theories, realized as social agents, to 'compete' with each other. This leads to a better understanding of their relative efficacy, common features, and differences. For Axelrod's prisoner's dilemma tournament, the submitted programs were agent models following strategies for playing the iterated game. Rapoport's winning social agent employed the 'tit for tat' strategy in which it cooperated on the first move and then imitated the previous action of its partner.

3. *Social Agents and Virtual Worlds*

At a high level of abstraction, there are two generic approaches to doing social science with artificial social

agents: bottom up and top down. In the top down approach, the computational social scientist utilizes social agents to search for underlying fundamental principles of sociality by endowing those agents and the worlds they inhabit with known constraints on social and cognitive behavior. The virtual social worlds are populated with the constraints known to exist in the real world, and a small number of highly realistic agents. The expectation is that the interaction among these agents and the world will lead to fundamental insights about the nature of social life and predictions about how to affect social behavior. In the bottom up approach, virtual social worlds are populated with large numbers of minimally intelligent agents. The expectation is that through interactions among these social agents, social realities will emerge in the virtual world.

Virtual worlds are computer simulated worlds of particular social and/or physical environments. Within a virtual world, there is a particular set of physical, temporal, social and cultural laws that must be obeyed. These worlds are populated by artificial agents. Examples range from game systems such as SimCity to research worlds for understanding the complexities of biological and social life such as Sugarscape. Virtual worlds become scientific test-beds in which researchers can grow artificial life and social theory from the ground up. Virtual worlds enable the researcher and student to reason about social behavior in a controlled setting, like a human experiment. However, unlike a human experiment, these worlds can be quite large. Thus, issues of scale, emergence, and time varying behavior can be addressed.

One of the earliest examples of a virtual world in sociology is Bainbridge's (1987) sociological laboratory. Bainbridge created worlds in which the student could explore classic theories of social behavior. Within these worlds, the student could run virtual experiments to see the impact of changing norms of social behavior. Another classic example of a virtual world is Schelling's (1969) segregation game. In this world, agents on a grid choose to move or stay in neighborhoods, based on a preference to be in a neighborhood with at least a minimum number of agents of their own kind. Using this model, he demonstrates that the optimization of local choices leads to globally destructive consequences—complete segregation and ghettos.

Much of the early work used cellular automata. Today, this same branch of research is being reborn in the work on artificial life, or A-Life. A typical A-Life model is Sugarscape (Epstein and Axtell 1997). Epstein and Axtell argue that they are using computers to do 'bottom up' social science. Given a collection of very simple agents with very simple rules, they try to grow complex social behavior in the Sugarscape world. In Sugarscape, multiple simple agents engage in social interactions in the processes of consuming sugar, moving across planes, giving birth, and dying. Sugarscape is a virtual world where agents can live, eat, die, and engage in social interactions. In its simplest form, the Sugarscape world is a torus made by wrapping a grid of 50 by 50 squares, such that sugar grows in some of the squares and not others. This world is populated by a number of agents, each of whom can consume sugar or move one square on the von Neuman grid (NSEW) each time period, as long as they do not occupy a square occupied by another agent. Consumed sugar translates into energy, or can be stored for later usage. Energy is needed for movement, reproduction, trade, engagement in conflict, etc. Without energy, the agents die. Agents placed randomly on the grid will, over time, develop a collective intelligence that will move them towards the 'fields of sugar.' Additional rules, for example, the inheritance rule, produce more social behavior. The more factors and rules added, the more social are the behaviors, which can be explored. For example, introducing a second resource, such as spice, leads to the emergence of trading and an economic market. Epstein and Axtell demonstrate that complex social outcomes need not have complex causes. As noted by Schelling, simple behaviors on the part of individual agents can have major social consequences.

4. *Social Agents and the Generally Intelligent Agent*

Research on individual cognition has generated a number of models of general intelligence and learning that extend beyond task-specific models of performance. Examples include SOAR (Laird et al. 1987), ACT (Anderson 1996), and neural networks (Rumelhart and McClelland 1987). These models take a more system level approach to cognition. If these general models of individual cognition are sufficient models of social behavior, then a group composed of such agents, each realized using one of these cognitive models, should account for all forms of social and organizational goal-oriented deliberation, which underlie social behavior (Carley and Newell 1994). Multiagent models can be constructed by weaving into organizational design sophisticated computational models of individual cognitive agents, tasks, and social situations.

According to ACTs, organizations are composed of agents who are cognitively-restricted, task-oriented, and socially-situated. The actions of these agents are constrained by the human cognitive architecture, the characteristics of the task, and the nontask characteristics of the environment in which the agents are situated (the social situation). An agent's actions are a function of the agent's cognitive architecture and knowledge. Consequently, manipulating the organizational design or social structure manipulates the

constraints on agent action imposed by the agent's knowledge, task and social situation.

ACTs theory is consistent with Simon's (1981) comment that the seeming complexity of human behavior is due, not to the reasoning mechanisms, but to the task environment. That is, the agent's cognitive architecture is seen as immutable and affects the process by which the agent gathers and communicates information, makes decisions, and learns. Each agent's knowledge changes over time through learning and forgetting, as a function of the agent's position within the organization or society, the task in which the agent is engaged, and various stresses on behavior. Knowledge mediates the constraints on agent actions, dictated by social and cultural situations.

5. *Theoretical Breakthroughs: The Nature of the Social Agent*

Models employing artificial social agents have a long history in the development of social theory and methodology. Results derived from computational models have led to a number of important theoretical breakthroughs that collectively generate a more complete understanding of the social agent.

One of the earliest and most profound uses of computational models was to develop the theory of bounded rationality (see *Bounded and Costly Rationality*). Prior to the 1960s, most formal theories of social and organizational behavior assumed rational actors with complete information and total insight. Arising out of the Carnegie School, Herb Simon, and others, argued that humans were boundedly rational, that is, social structure limits their access to data, and human cognition limits their ability to process that information. As a result, decisions are made by 'satisficing' rather than 'optimizing.' In *A Behavioral Theory of the Firm* (Cyert and March 1962) and in the *Garbage Can Model of Organizational Choice* (Cohen et al. 1972), the authors demonstrate, through computational theorizing, that human limitations affect what choices are made when and how. Further, by taking such bounds into account, theoretical propositions better match actual observations. This work was instrumental in revolutionizing theoretical and empirical work on group and organizational behavior. The resulting information processing view is now an integral part of many social theories.

Another theoretical breakthrough is in the area of chaos (see *Chaos Theory*). The notion of deterministic chaos has captured the imagination of scholars and the public. Social behavior can become chaotic. This in itself is not a novel theoretical proposition. However, they ran a series of social agent simulations that demonstrated that when the agents are social, that is have the capacity to base their actions on beliefs about others' strategies and on the observed behavior of the collection of agents, meaningful order emerges. This suggests that, for recognizable social behavior to emerge, the content of individual cognition needs to contain mental models of others and their actions.

A related breakthrough is in the area of transactive memory. Wegner (1995), using a computer system as a metaphor for human memory, developed the powerful idea of transactive memory. Transactive memory refers to the ability of a group to have a memory system exceeding that of the individuals in the group. The basic idea is that knowledge is stored as much in the connections among individuals as in the individuals. Wegner argues that factors relevant in linking computers together such as directory updating, information allocation, and coordination of retrieval, are also relevant in linking the individuals' memories together into a group memory. Empirical evidence provides some confirmation and suggests that for a group, knowledge of 'who knows what' is as important as the task knowledge. Transactive knowledge can improve group performance. Thus, for recognizable social behavior to emerge, the content of individual's mental models needs to include knowledge of who knows what and who knows who.

Today, important advances in the social sciences are made possible by the use of artificial social agents. This work has led to a new paradigm, the neo-information processing paradigm, in which social and cognitive behavior is jointly determined by both cognition and interaction. Social behavior realizes social outcomes as the result of changes in interaction among agents and agents whose mental models take into account what others know, who they know, and what they are likely to do. This is the interaction-knowledge perspective.

See also: Artificial and Natural Computation; Artificial Intelligence in Cognitive Science; Artificial Intelligence: Uncertainty; Cognitive Theory: ACT; Cognitive Theory: SOAR; Computational Approaches to Model Evaluation; Decision Support Systems; Intelligence: History of the Concept; Game Theory; Game Theory and its Relation to Bayesian Theory; Game Theory: Noncooperative Games

Bibliography

Anderson J R 1996 ACT: A simple theory of complex cognition. *American Psychologist* **51**: 355–65

Bainbridge W S 1987 *Sociology Laboratory*. Wadsworth, Belmont, CA

Bainbridge W, Brent E, Carley K, Heise D, Macy M, Markovsky B, Skvoretz J 1994 Artificial social intelligence. *Annual Review of Sociology* **20**: 407–36

Carley K, Newell A 1994 The nature of the social agent. *Journal of Mathematical Sociology* **19**(4): 221–62

Carley K M, Prietula M (eds.) 1994 *Computational Organization Theory*. Lawrence Earlbaum, Hillsdale, NJ
Cohen M D, March J G, Olsen J P 1972 A garbage can model of organizational choice. *Administrative Sciences Quarterly* **17**(1): 1–25
Cyert R, March J G 1992 *A Behavioral Theory of the Firm*, 2nd edn. Blackwell, Cambridge, MA
Epstein J, Axtell R 1997 *Growing Artificial Societies*. MIT Press, Boston
Laird J, Newell A, Rosenbloom P 1987 SOAR: An architecture for general intelligence. *Artificial Intelligence* **33**: 1–64
Markovsky B, Skvoretz J, Willer D, Lovaglia M J, Erger J 1993 The seeds of weak power: An extension of network exchange theory. *American Sociological Review* **58**: 197–209
Rapoport A, Chammah A M 1965 *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press, Ann Arbor, MI
Rumelhart D E, McClelland J L 1987 *Parallel Distributed Processing*. MIT Press, Cambridge, MA
Schelling T 1969 Models of segregation. *American Economic Review* **59**: 488–93
Simon H 1981 *The Sciences of the Artificial*, 2nd edn. MIT Press, Cambridge, MA
Wegner D M 1995 A computer network model of human transactive memory. *Social Cognition* **13**(3): 319–39
Weiss G (ed.) 1999 *Distributed Artificial Intelligence*. MIT Press, Cambridge, MA

K. M. Carley

Artisans and Guilds, History of

1. Definitions

Artisanal labor is a particular form of labor that has existed in most—if not all—human societies. Historically, it first appeared as non-agricultural production, and during the last 200 years primarily as small scale or small commodity production, in contrast to mass production and mechanized factory industry. The focus of the definition is on production, thus also distinguishing artisanal labor from economic activities such as commerce. Indeed, one cannot draw strict boundaries here. The English term 'trade,' for instance, has been in use for artisanal production as well as for commerce and exchange. Furthermore, artisanal labor is not necessarily limited to production, but also includes some activities that in modern economic thought are assigned to the tertiary or service sector. Artisan is a narrower concept than artisanal labor. While artisanal labor may be performed by men and women as one activity among many others, artisans can be defined as persons who specialize in artisanal labor and make their living mainly, if not exclusively, from this kind of work. There are some almost universal elements in the concept of the artisan: manual labor and the use of tools; specialization of labor, which connects the concept of the artisan with the concept of profession or occupation; skill and some kind of training as a precondition of skill; and work not for subsistence but for consumers or markets. Besides these universal elements, there are also some historically variable ones that fluctuate in accordance with broader social and economic change.

2. Artisans: Concepts and History

The origins of artisans are connected with social changes at the end of the Neolithic Revolution (see *Work, History of*) in approximately the fourth millennium BC. Foremost among these was the development of technology, above all in the fields of metallurgy, ceramics, and woodworking. This promoted specialization in the production of metals as well as in the fashioning of tools, weapons, and a growing number of practical implements. Second, the emergence of social stratification brought about increased demand for luxury goods by the upper classes, the production of which called for special training and particular skills. Third, there was the formation of urban cultures, especially in the Near East and the Mediterranean. Artisans were concentrated in cities, and it was here that the division of labor within artisanal production accelerated and specialized crafts and trades came into being. During the first millennium BC, dozens of different artisanal occupations emerged in the urban cultures of India, the Near East, and the Mediterranean.

The value assigned to labor in the crafts and trades and the social position of the artisan has been characterized by a high degree of ambivalence. In all societies in which work was held in low esteem, this applied to artisanal labor too. In European history, though, there is also a long tradition of valuing the work of artisans more highly than other forms of manual labor. Whereas in some non-European cultures, such as ancient China, artisans were assigned a lower social status than peasants, in the European tradition they enjoyed a relatively high level of prestige. In early ancient Greece, this was expressed in the myth of Prometheus the god of fire, pottery, and metallurgy. 'In the eighth century BC the crafts of blacksmith, pottery, and weaving were considered in the same light as that of priests, bards and doctors (…). All these activities were called *technai*, that is, activities implying the use of secret processes, specialized knowledge based on long apprenticeship and an initiation ritual.' (Godelier 1980) Some crafts appeared as sacred in Gallic and Germanic mythology as well, and in the European Middle Ages, for instance, in the scale of social values in early Germanic law, artisans clearly had a higher rank than persons engaged in agricultural activities. 'The craftsmen who

forged the weapons for the aristocratic warriors (…); and the goldsmiths who decorated these weapons and created jewels for the women of these warriors (…) were important personages who upheld the prestige of technical skill' (Le Goff 1980).

2.1 Artisans in Corporate Society

The urbanization process in twelfth- to fourteenth-century Europe strongly influenced the concept and the social standing of the artisan. Most importantly—and not at all usual for people working with their hands—artisans, as urban citizens, were freemen. In many English towns, for instance, completion of an apprenticeship entitled an individual to the freemanship of the respective town (which meant permission to take up residence, claim to social relief if the need arose, and often some form of political participation). The increasing division of labor in towns created a wide range of artisanal occupations, and the expanding guild system (cf. guilds, See Sect. 3) served as an institutional framework for training and work in these occupations. Both of these developments had consequences, which became basic to the concept of the artisan in the corporate societies of Europe from the late Middle Ages until the eighteenth century.

First, artisanal work—even if it was manual labor—was not simply regarded as labor. The negative connotations of labor (travail, hard work) as toil, pain, burden, or penitence were replaced by an understanding of artisanal labor as art. *Arts et métiers* (arts and trades) was the common French term used to group together all artisans (a generalizing term which appeared in the sixteenth century). In eighteenth-century France, artisanal crafts were regarded as honorable because they were based not only on the 'effort of the body' but also on the 'subtlety of the mind' (Sewell 1980, p. 23). Second, by organizing in guilds or other associations, artisans had a recognized standing in society and the state. Whether they enjoyed royal privileges, municipal statutes, or the permission of the church to form a confraternity, all these defined them as a corps or as an estate (French: *état*; German: *Stand*). In both respects, inherent in the concept of the artisan, was a sharp destinction vis-à-vis unskilled laborers.

2.2 The Ambiguous Journeyman

Traditionally, the concept of the artisan referred primarily to the master artisan—an independent craftsmen who ran his own workshop and casually or regularly employed a few hands. As head of the household, the master was part of the corporate structure and exercised domestic authority over his family and his workforce. On the other hand, the conception of the artisan also included different categories of employees. First, it referred to apprenticeship, which was understood as a period (and a social position) of training—generally, although not necessarily, connected with adolescence. In the late Middle Ages, there appeared a further group: artisans who had finished their apprenticeship but had not—or not yet—become masters. In English, the term journeyman (which stems not from the journey of the tramping artisan, but from the French word *jour* meaning day, and designating one who works and is paid by the day) became most widespread to describe this group. While the concepts of apprentice and master are relatively clear, the concept of the journeymen is fluid and ambiguous. On one hand, it refers to a life-cycle stage, as the completion of training, mainly by tramping a couple of years. From this perspective, journeymen were young, single men subject to the domestic authority of the master (when at work and not on the road), until they themselves became masters and householders. On the other hand, the concept of journeymen refers to wage laborers—often on a life-long basis—having their own households and families. In practice, it is almost impossible to make a clear distinction between these two concepts and modes of life. Very roughly, the first type prevailed in small-scale production (bakers, tailors, and the like), the second in textiles (particularly where the craft was transferred into outwork or domestic industry) and in the building trades. Furthermore, the first seems to be more common in Central Europe, whereas the second is more familiar in the West, particularly in England. Many historians also believe that over the course of the late medieval and early modern history, the first type was gradually replaced by the second, but this is probably a far too general and too linear model (cf. Reith 1999). Generally, the relations between masters and journeymen included solidarity and moral community as well as conflict.

2.3 Redefinition of the Artisan in Class Society

In the emerging class societies of eighteenth- and nineteenth-century Europe, the concept of the artisan became redefined both in *petit bourgeois* and in proletarian terms. On one hand, the concept still focused on the independent master craftsman working in his own workshop with his own tools. Apprentices and journeymen were included in this concept as future or anticipated masters. On the other hand, the concept of the artisan increasingly referred to life-long employed workers working in a wide range of economic settings—even in factories—but being distinguished from ordinary laborers by their common 'property of skill.' Artisans were 'those who through apprenticeship or its equivalent had come to possess a

skill in a particular craft and the right to exercise it' (Rule 1987, p. 102).

In social theory, the first model became predominant. In Enlightenment philosophy as well as in political economy, the artisan began to be regarded as a counter-model to capitalism. Karl Marx conceptualized the artisan as a personification of a non- (or pre-) capitalist mode of production, which he called small commodity production (*kleine Warenproduktion*). At the turn of the century, the German social scientist Werner Sombart regarded the artisan as the 'ideal type' of the pre-capitalist economic system, which, in his view, was characterized by empiricism and traditionalism (as opposed to rationality) and a lack of interest in accumulation and profit. This perspective became extremely powerful and shaped more or less completely the historical interpretation of the artisan.

In practice, however, in the economically advanced regions of Western Europe, most of the workmen who identified themselves as artisans were in wage employment. Adam Smith in his 'Wealth of Nations' (1776) estimated that in England there were 20 men working for wages for every one who was his own master (Rule 1987, p. 103). Here, the artisan as a skilled workman became central to the self-image, identity, and rhetoric of the emerging working class. On the continent, and particularly in Germany, even in the mid-nineteenth century, the number of masters almost equaled the number of journeymen in the crafts and trades. Here, the concept of the artisan became synonymous with independent master craftsmen and their respective *petit bourgeois* political movements.

3. Guilds: Concepts and History

In Western history, both the concept and the practice of artisans are closely connected with guilds. Associations of artisans existed in many historic societies and cultures. For example, there is evidence of their existence in classical India, ancient Rome, Tokugawa Japan, and Colonial America. Nevertheless, the term guild or craft guild refers to a particular type of artisanal association, which had its place in the corporate societies of Europe from the late Middle Ages to the nineteenth century. Despite enormous variations from city to city, region to region, and one historical period to another, the concept of the guild refers to a combination of several characteristics. First, a major organizational principle was occupation. Usually, although not necessarily, guilds organized artisans of the same occupation or of related/similar occupational groups, and occupation formed the major line of differentiation separating one guild from another. Changes in the division of labor and in occupational structures are therefore the major dynamic forces in the development of the guild system. Second, guilds were strongly concerned with economic matters in their respective sector of the economy. Market regulation—both in respect to commodity markets and labor markets—was an essential part of their economic activities. Third, guilds were primarily organizations of master craftsmen. Occasionally they also included apprentices and journeymen in minor positions; on other occasions, they definitely excluded these groups, thus providing the impetus for journeymen to found their own clandestine or legal associations. Fourth, guilds were formally constituted and officially recognized bodies. Statutes or privileges granted by municipal or state authorities made them a solid part of the corporative structures of society. In this context, guilds also performed political functions, casually participating in the political system, occasionally even dominating urban politics. Lastly, guilds were not only economic and political bodies, but also multifunctional institutions engaged in religious, cultural, and social activities.

The first guilds—in the sense of this definition—were established in France, England, Germany, and Italy in the twelfth century. In the thirteenth and fourteenth centuries, guilds existed almost everywhere in urban Europe with a wide variety of titles and designations. Latin sources use terms such as *officium*, *ministerium*, *magisterium*, *artificium*, *collegium*, and many others. The French language speaks of *métiers*, *corps des métiers*, or *corporations*; in English they are called guilds, crafts, or mysteries; in German *Zunft*, *Zeche*, *Innung*, or *Amt*; in Italian *arti*, *gremi*, and so on.

3.1 Guilds in Corporate Society

Despite these obviously medieval roots, European guilds became a mass phenomenon and a nearly universal institution in the Early Modern Era. In the Netherlands and France, Spain and the various Italian states, German-speaking Central Europe, and the Nordic countries, most guilds were founded in the sixteenth, seventeenth, and eighteenth centuries. Even in England with its highly developed market economy, many guilds were re-established in the Restoration, and in cities like Bristol and Bath, there were even new guilds being founded as late as the mid-eighteenth century. A guild system also developed in the Ottoman Empire, particularly in the Balkans, the Near East, and Northern Africa. In the Russian Empire, on the other hand, guilds were unknown. In Colonial America during the early eighteenth century, many artisans took an interest in setting up craft guilds. Even though some associations were actually founded (such as the Carpenter's Company of the City and County of Philadelphia in 1724), no general guild system came into being.

Even when guilds were at the pinnacle of their importance in Europe, not all artisans were incor-

porated. In Western Europe and particularly in France and England, guilds were limited to cities, and did not even exist in all cities. In many Central European states—for instance, Austria—guilds transcended municipal boundaries and became regional or nationwide institutions organizing urban artisans as well as craftsmen from villages and the countryside. In those places where guilds did exist, there were few instances in which all artisans belonged to them. A large proportion of artisanal labor always took place outside of the guild system.

Mercantilist theorists and political leaders of the seventeenth and eighteenth centuries regarded guilds as instruments of unified state regulation of trades, labor, and taxes. During this period, many European states indeed tried to make guilds universal and to establish the guild system as a framework for all economic activities; one such example is the effort of French Prime Minister Colbert in his trade laws of 1673.

3.2 The End of the Guild System

A hundred years later, the opinion of guilds held by political and intellectual elites had changed completely. During the second half of the eighteenth century, an antagonistic discourse emerged. Mainstream economic and political thought—expressed in the fields of political economy, physiocracy, or in Enlightenment thought in general—came to regard guilds as obsolete relics of the past that sought to retard or prevent economic development. Proponents of this approach called for complete freedom of trades and the abolition of guilds. Conclusions diametrically opposed to these were drawn by conservative thinkers who saw guilds as the embodiment of a corporative ideal and as a means of combating the assumed decay of authority and order.

When the leading physiocrat and Enlightenment thinker Turgot served briefly (1774–76) as French prime minister, he issued a decree in 1776 completely abolishing all guild privileges as well as expressly forbidding 'all masters, journeymen, workers, and apprentices (…) to form any association or assembly among themselves (…), (Sewell 1980, p. 73). Despite the fact these measures were repealed only a few months after Turgot's fall from power, they nevertheless initiated the process of the elimination of guilds from Europe over the course of the next 100 years. Revolutionary France took a radical step as early as 1791 and banned guilds along with all other labor organizations. England's Combination Acts of 1799/1800 also mandated the 'prohibition of all combinations'—those of masters as well as of journeymen (Chase 2000, p. 84). Most Central European states oscillated between anti-guild and pro-guild principles. They indeed forbade the organization of artisanal workers, but permitted the guilds—as organizations of masters—to remain in existence. It was not until the 1860s that the guilds were dissolved once and for all in the German states and in the Habsburg Monarchy. This process reflected the guilds' loss of their social function. In most countries, dissolving the guilds met no opposition, but it was precisely in Central Europe that guilds remained alive as the object of the projections engendered by a corporative social ideal and conservative nostalgia.

3.3 Guild Traditions in Petit Bourgeois and Labor Movements

The end of the guild system did not also mean the end of artisanal organizations. These developed over the course of the nineteenth century in two separate directions. On one hand, there was the organizational continuity of master artisans, independent craftsmen, and employers. This line of tradition was significant above all in Central Europe. In the German and the Habsburg Empires, in Denmark, and is some other countries, cooperative organizations in the crafts and trades continued to exist even after the dissolution of the guilds. From the 1880s on, these organizations were granted wide-ranging authority in their respective fields, such as control over the training of apprentices. Politically, these trade associations were closely associated with the 'middle class movements,' the political movements of the *petits bourgeois*. They demanded the restoration of guilds and developed corporative ideals in the framework of conservative and right-wing politics (cf. Haupt 1996).

The tradition of guild crafts and trades, however, also played an important role in the emerging labor movement. Although journeymen's associations had been banned in many European countries at around 1800, confraternities, *compagnonnages*, houses of call, and similar institutions continued to exist on an informal or illegal basis. Although there was no direct institutional continuity between guilds and trade unions, there was nevertheless a close connection on the conceptual, experiential, and personnel levels. Artisanal workmen who came out of guild traditions were used to organizing in such associations, which they considered their political right. They were likewise accustomed to settling labor disputes collectively, be that in the form of negotiations or job actions. Furthermore, they also carried on the tradition accentuating and defending their social status, which was based upon the 'property of skill.' All these traditions were important incentives for the founding of the first trade unions in the first half of the nineteenth century.

In a broad sense, artisans were also the driving force behind the early political labor movement. Of course the concept of class contradicted the tradition of craft consciousness, since it stressed the common experiences of all working men, skilled or unskilled, in whichever craft or trade they engaged. Nevertheless,

there is almost universal agreement 'that skilled artisans, not workers in the new factory industries, dominated labor movements during the first decades of industrialization. Whether in France, England, Germany, or the US, whether in strikes, political movements, or incidents of collective violence, one finds over and over the same familiar trades: carpenters, tailors, bakers, cabinetmakers, shoemakers, stonemasons, printers, locksmiths, joiners, and the like. The nineteenth-century labor movement was born in the craft workshop, not in the dark, satanic mill' (Sewell 1980, p. 1).

4. *Revisionism in Guild and Artisan Historiography*

The negative assessment of guilds in eighteenth and nineteenth century thought also characterized historiography. Until recently, most historians had assumed that guilds were technologically regressive, that they restricted economic growth and straitjacketed the development of free market capitalism.

All of these assumptions have been subjected to critique since around 1980. A growing number of historians is now engaged in a 'revisionist version of guild history' (cf. Epstein et al. 1998, Ehmer 1998). With respect to early modern Europe, artisans and guilds have become reintegrated into the mainstream of the development toward capitalism. The Dutch case shows most clearly what can be seen in many other European regions as well: 'guilds were part and parcel of commercial capitalism,' showing extensive growth particularly in the 'Golden Age' (Lucassen 1995). The increasing dependency of Europeans on markets, the growing variety of available goods and of artisans who made them is now seen as evidence of a 'dynamic, flexible and creative craft economy' (Farr 1977, p. 25). The guilds' emphasis on skill, their system of formalized training, and the technological transfer through tramping artisans, is now said to explain, 'the undisputed technological leadership of guild-based production,' particularly in comparison to the rural putting-out system (proto-industry) and early centralized factories (Epstein 1998, p. 705).

Furthermore, guilds are no longer regarded as stable institutions of equals, but as highly stratified and dynamic fields of social relations. Merchant capitalists used some guilds to control and discipline economically dependent masters and journeymen, whereas other guilds were instruments of resistance to merchant capital. There are also new perspectives on artisanal cultures, conceptualizing them as historically variable and adaptable sets of practices. In this new light, customs that were regarded as 'ancient' by the artisans themselves as well as by the historians who came after them are now often being shown to have a rather brief history. For instance, French *compagnonnages*, journeymen's associations with highly elaborate rituals and an aura of mystery, experienced their 'golden age' from the mid-eighteenth century to the 1830s/1840s—the very period of redefinition of the position of artisans and guilds in society. (cf. Sonenscher 1987, pp. 31–63).

Another area in which revision is taking place is the role of the artisans in the Industrial Revolution. In eighteenth- and nineteenth-century Britain, small-scale and labor-intensive production, skill and 'hand technology' drawn from a huge reservoir of urban and rural artisans played a key role, not as a mere remnant of the past, but as a central and dynamic component of industrial capitalism (cf. Berg 1985). Generally, artisanal small-scale production displayed remarkable persistence in industrializing societies, adapting to the needs of industry, and benefiting from the growing demand of consumer society. The generally successful process of adaptation was not harmonious and free of conflict. Some artisanal branches survived, some new ones were born, but others—including all textile production and most metal manufacturing—declined and vanished. Artisans in these trades experienced the devaluation of their skills. The same happened to the journeymen of the metropolitan garment industries, which left the artisans' workshops and became reorganized as an urban domestic industry based on the cheap labor of women and children instead of that of skilled workmen.

The role of women in the world of the artisans is a major point of recent discussions (Hafter 1995, Hanawalt 1986). It is generally agreed that women in most human societies regularly performed artisanal labor, but rather seldom became artisans in the sense of professional, organized, and honorable workers. In European history, the concept of the artisan and, to an even greater extent, that of the guild were male concepts. Artisans who practiced crafts and trades in European cities during the late Middle Ages produced commodities that had previously been produced in households by women as well as by men. The difference was that the professional artisans then produced for the market and their labor became regulated by a set of formal and informal rules. Even if 'the early guild ordinances (…) did not regard skilled guild labour as a male preserve,' in the long run the switch from household production to market production, from unregulated artisanal labor to a 'craft,' excluded women (Wiesner 1989).

Recent research clearly shows that women participated in artisanal labor in many respects, as masters' wives and daughters, as domestic servants in artisans' households, as widows, and, last but not least, as autonomous independent producers. Nevertheless, in all these roles they were not officially recognized as skilled workers. This shows that the very concept of skill, which is central to the conception of the artisan, is highly ambiguous. On one hand, skill indeed requires a certain body of knowledge, experience, and training; on the other hand, skill is a historically

variable social construct. Not all labor activities that require some expertise are regarded as skilled work. In European history, skill was defined by male artisans and guildsmen not only with respect to actual work practices, but also as property, status, and a means of social distinction.

See also: Capitalism; Industrialization; Industrialization, Typologies and History of; Labor History; Labor Unions; Labor, Division of; Middle Ages, The; Social Stratification; Work and Labor: History of the Concept; Work, History of

Bibliography

Berg M 1985 *The Age of Manufactures Industry, Innovation and Work in Britian 1700–1820*. Fontana, London

Chase M 2000 *Early Trade Unionism. Fraternity, Skill and the Politics of Labour*. Ashgate, Aldershot, Brookfield, Singapore, Sidney

Crossick G, Haupt H-G 1995 *The Petite Bourgeoisie in Europe 1780–1914*. Routledge, London and New York

Epstein S A 1991 *Wage Labor and Guilds in Medieval Europe*. The University of North Carolina Press, Chapel Hill and London

Epstein S R 1998 Craft Guilds, Apprenticeship, and Technological Change in Preindustrial Europe. *The Journal of Economy History* **58**: 684–713

Epstein S R, Haupt H G, Poni C, Soly H (eds.) 1998 *Guilds, Economy and Society*. Publicacioines de la Universidad de Sevilla, Seville, Spain

Ehmer J 1998 Traditionelles Denken und neue Fragestellungen zur Geschichte von Handwerk und Zunft. *Lenger* 19–77

Farr J R 1997 On the shop floor: guilds, artisans, and the European market economy, 1350–1750. *Journal of European Modern History* **1**: 24–54

Farr J R 2000 *Artisans in Europe, 1300–1914*. Cambridge University Press, Cambridge, UK

Godelier M 1980 Work and its representations: a research proposal. *History Workship Journal* **10**: 164–74

Guenzi A, Massa P, Piola Caselli F (eds.) 1998 *Guilds, Markets and Work Regulations in Italy, 16th–19th Centuries*. Ashgate, Aldershot, Brookfield, Singapore, Sidney

Hafter D M (ed.) 1995 *European Women and Preindustrial Craft*. Indiana University Press, Bloomington and Indianapolis, IN

Hanawalt B (ed.) 1986 *Women and Work in Preindustrial Europe*. Indiana University Press, Bloomington and Indianapolis, IN

Haupt H-G 1996 Zum Fortbestand des Ancien Régime im Europa des 19. Jahrhunderts: Zünfte und Zunftideale. In: Hettling M, Nolte P (eds.) *Nation and Gessellschaft in Deutschland. Historische Essays*. Verlag C. H. Beck, Munich, Germany

Haupt H-G 2000 *Das Ende des Ancien Régime. Zünfte in Europa um 1800*. Vandenhoek und Ruprecht, Göttingen, Germany

Hugger P (ed.) 1991 *Handwerk zwischen Idealbild and Wirklichkeit. Kultur- und sozialgeschichtliche Beiträge*. Verlag Paul Haupt, Bern and Stuttgart, Germany

Le Goff J 1980 *Time, Work and Culture in the Middle Ages*. The University of Chicago Press, Chicago

Lenger F 1998 *Sozialgeschichte der deutschen Handwerker seit 1800*. Frankfurt am Main, Suhrkamp, Germany

Lenger F (ed.) 1998 *Handwerk, Hausindustrie und die historische Schule der Nationalökonomie. Wissenschafts- und gewerbegeschichtliche Perspektiven*. Verlag für Regionalgeschichte, Bielefeld, Germany

Lis C, Soly H (eds.) 1997 *Werelden van Verschil. Ambachtsgildern in de Lage Landen*. VVB Press, Brussels

Lucassen J 1995 Labour and early modern economic development. In: Davids K, Lucassen J (eds.) *Miracle Mirrored. The Dutch Republic in European Perspective*. Cambridge University Press, Cambridge, UK, pp. 367–409

Reith R 1999 *Lohn und Leistung. Lohnformen im Gewerbe 1450–1900*. Franz Steiner Verlag, Stuttgart, Germany

Rule J 1987 The property of skill in the period of manufacture. In: Joyce P (ed.) *The Historical Meanings of Work*. Cambridge University Press, Cambridge, UK, pp. 99–118

Schulz K (ed.) 1999 *Handwerk in Europa. Vom Spätmittelalter bis zur Frühen Neuzeit*. R. Oldenbourg Verlag, Munich, Germany

Sewell W 1980 *Work and Revolution in France. The Language of Labor from the old Regime to 1848*. Cambridge University Press, Cambridge, UK

Sonenscher M 1987 Mythical work: workship production and the compagnonnages of eighteenth-century France. In: Joyce P (ed.) *The Historical Meaning of Work*. Cambridge University Press, Cambridge, UK, pp. 31–63

Wiesner M E 1989 Guilds, male bonding and women, work in early modern Germany. *Gender and History* **2**: 125–37

J. Ehmer

Arts Funding

1. The Term 'Arts Funding'

The term 'arts funding' comprises those processes that are aimed at providing the frameworks for the production, dissemination, and consumption of art objects. Furthermore, a description of arts funding focuses particularly on the economic aspects of these processes. That is, how individuals, groups, institutions, and agencies, over time have been willing to pay for art objects to be produced, disseminated, and consumed. Thus, arts funding research has been exploring economic preconditions for creative and performing artistic work, the various social settings in which this has taken place, *who* has paid for such activities, and *why* they have done so. 'Patron' is the term for various funding sources. 'The arts' is often associated with the visual arts. When describing arts funding, other art forms such as music and literature are equally important. This means that studies of arts funding can be concerned with creative artists such as painters, sculptors, composers, and writers, and with performing artists such as musicians and actors, all being funded directly by patrons and by institutions where they have worked, or indirectly through various regulatory means.

Arts funding is a social phenomenon in the sense that its existence and organizational form depend on persons and institutions constituting certain configu-

rations which differ over time. These configurations may also differ across societies and political systems. In this article, these configurations will be examined in turn. The last section describes some issues in current arts funding research.

2. Why the Arts Have Been Funded

Arts funding research is interested in those individuals or institutions willing to put money into artistic activities, and their intentions in doing so. The reasons vary from wanting art for art's sake, to viewing the arts as fulfilling different instrumental purposes. Some functions are more closely connected with some specific patrons than with others. Furthermore, the connections between patrons and functions have changed throughout history. The presentation below also shows that the term 'public' has manifested itself differently throughout history.

Art as a phenomenon can be traced back to prehistoric times. In *The Story of Art*, Gombrich (1995) alludes repeatedly to 'art patrons' and 'patronage,' beginning around 700 AD. This article's historic overview of arts funding begins here, indicating the most important art patrons over time. It is mostly about Europe and North America, but examples from Asia, Africa, and Latin America are also included:

(a) ca. 700–1300: the church,
(b) ca. 1300–1660: the aristocracy,
(c) ca. 1600–1770: royal courts,
(d) late eighteenth century: the bourgeoisie in private clubs and associations,
(e) primarily after World War II: the state, government,
(f) at all times, but particularly since the 1960s: private patrons,
(g) particularly after the 1960s: corporations and other private patrons.

2.1 Church and Aristocratic Patronage

In 326 AD, a Vatican church in Rome was consecrated in memory of the Apostle Peter. In 1506, Pope Julius II (1443–1513) tore down this basilica, commissioning the architect Donato Bramantes to build St. Peter's Church on the same spot. The purpose was twofold—to create a room for religious worship, and fulfilling a genuine artistic interest, including a desire to support individual artists.

The Medici family, ruling from 1434, made Florence the center for renaissance culture, supporting artists like Michelangelo and Leonardo da Vinci. They were funded for their artistic excellence, but also as tutors and coordinators of apprentices and skilled workers who decorated church walls and ceilings. Thus, they were artists as well as experts in a certain handcraft, serving representative as well as aesthetic functions. The Medici family also established Europe's first public libraries in Florence, thus enhancing access to literary art. However, the idea of autonomous art was not yet relevant.

Religious and aristocratic patronage also explains the construction of magnificent temples between the tenth and fifteenth centuries in, example, India, Korea, Afghanistan, and Nepal. For instance, in Sri Lanka (Bandara 1972): 'Buddhism [...] determined the aims of artistic creation and functioned as the common link among the ruler (sponsor), the artist (creator), and the public (recipient).'

2.2 The Royal Court as Patron

During the fifteenth century the royal courts increasingly became the primary arenas for artistic endeavors. Royal arts patronage reached its peak during Louis XIV's reign from 1643. Art should illuminate the splendor of the king and his court, which left the artists in the role as servants. Thus the representative function was more important than artistic content. This posed a dilemma for artists who focused primarily on their creative abilities, like Wolfgang Amadeus Mozart (1765–91). Mozart began composing when art was still regarded as a useful handicraft, at the same time that the idea of art for art's sake began to emerge. He was torn between an artistic role in which he could compose 'pure' music, and a representative role, being obliged to serve patrons commissioning and paying for familiar, and popular, types of music.

The courts offered a direct relationship between the artists, the art objects, and the art consumers, making art presentation a somewhat personal, but rather exclusive affair, being isolated from the rest of the population and its plebeian culture. The artists' prestige was closely connected with the king's personal interest in the arts. King Christian IV (1588–1648) in Denmark, the Swedish King Gustav III (1746–92), as well as several kings in France and Great Britain were active art patrons, commissioning murals and portraits, and giving grants to artists for travel abroad. A major part of Nepal's cultural heritage consists of wonderfully decorated temples that were built during the Malla dynasty (1350–1768) to honor and appease certain gods, and to impress their subjects and other royalty. In Nigeria, tribal kings bestowed prestige upon artists in return for masks, music, clothing, etc., for use in everyday life and during ceremonies.

With the rise of liberal political ideas, culminating in the French Revolution of 1789, the royal courts lost their role as centers for art and culture. Instead, artistic activity and presentation were moved outside the walls of the castles and into the rooms of private clubs and salons where the bourgeoisie reigned, establishing them as meeting places for a variety of people wanting to participate in the arts world, each having particular interests at stake. The owners of the salons obtained

prestige as the patrons of such meeting-places, and art dealers and collectors became the dominant art patrons, either as private individuals or as representatives of galleries and museums.

This new configuration had consequences for the functions of art, for the position of artists, and for arts funding. Art objects continued to serve representative purposes, but the content came in focus. For instance, when the kings lost their political power—particularly during the reign of Louis XV (1715–74)—writers were approached by political parties and asked to write propaganda pamphlets. With new interpretations of art and its functions followed an interest in, and concern for, the individual artist as a creative and independent personality. Art works were no longer seen as handcraft only, but had become autonomous objects with intrinsic value, that is, art for art's sake.

2.3 Diversification of Art Patrons

The expansion of arts audiences gave the artists a more prominent position in society. The number of literate people increased, and the opening of concert halls and art galleries for the general public made the arts more accessible. Artists obtained a variety of funding sources: dealers and buyers of visual arts, publishing companies for books and music sheets, and the general public as cultural consumers. These changes within the arts world indicate the beginning of a democratization process that made the arts available to an expanding 'public.' While jeopardizing a fairly secure income at court, the diversification also freed the artists from restricting themselves to the changing fashions among the aristocracy. Arts funding had become a matter of artists producing goods that were offered on a market. For instance, the expanding literary audience demanded a variety of content: mysteries, romantic fiction, sensations, drama, and social novels. Daniel Defoe (1660–1731), Jonathan Swift (1667–1745), Victor Hugo (1802–85), and Charles Dickens (1812–70) were among authors meeting this demand.

Thus, the diversified art market at the beginning of the nineteenth century comprised a diversity of art forms, offering a variety of options for artists as well as the audience. The significant change was the expansion of popular art forms, like vaudeville within theater, and the serial realistic novel about issues in contemporary society. Classical art forms were still created, but had to compete with the more popular art forms. The art objects had become goods for sale on a market, the artist had become an independent individual who had to deal with the market, and the audience had developed into consumers who had the privilege to choose. This also meant more differentiated arts funding structures than before with a variety of patrons: members of private clubs and associations, audiences at public exhibitions and performances, and consumers in bookseller's shops. Artists were faced with different options. They could cater to popular demands from the new audiences. It was a good way of earning money, but was looked down upon by those who regarded themselves as genuine creative artists. Their patrons had to be found among a niche audience.

The period from the late eighteenth to the early twentieth century is characterized by the dominant position of the bourgeoisie, particularly the new middle class that came with industrialization and increased business and commerce. Similarly to the church and the royal courts, people within the bourgeoisie also wanted to distinguish themselves from peasants and industrial workers. Showing interest in the arts was one convenient option, either by being seen at public places such as concert halls, theaters, and art exhibitions, or by commissioning and buying visual art. The meaning of the term *public* also expanded during this period. First, the public comprised an invited audience, modeled on the eighteenth century events at the European courts. Then came a period when such events, for instance classical concerts, were open to subscribers and members of music associations. Finally, around 1880, everyone could buy tickets to arts performances, making the event truly public. Thus, arts funding involved more patrons than before, although people from the middle class were the dominant contributors. In some countries, its social, political, and cultural dominance was so profound that it excluded government arts funding. This was particularly the case in Britain, where the landed aristocracy put money into portrait paintings to decorate their estates, thus: 'culture that required a public rather than a private setting, such as theater, opera, and ballet, had only a limited place in this pattern' (Ridley 1987, p. 225).

2.4 Corporations as Patrons

The expanding industrial production brought business corporations into arts funding, particularly in the United States. In 1903, the Atchison, Topeka, and Santa Fe Railway commissioned artists to paint pictures of the western parts of the United States. Not in order to decorate train compartments and waiting rooms, but to develop visuals for advertising, menus, posters, travel brochures, etc. Such material was ordered directly from the artists. At the end of the 1930s, artists were funded indirectly when corporations began assembling their own art collections, IBM being among the first to do so. This type of arts funding became more pronounced during the 1970s when most corporations began purchasing art for their own collections. Individual business owners and company presidents had already supported the arts from personal interest, while corporate giving to art institutions must be viewed in connection with the Federal Revenue Act from 1935. This allowed a certain percentage of corporate deductions of net income for contributions to charitable and educational organ-

izations that had been designated as not-for-profit according to Section 501(c)(3) in the Federal tax code. European countries and Australia also have a tax incentive system for arts funding, but it does not play as crucial a role as in the United States.

In 1967, the Business Committee for the Arts (BCA) was created, confirming the vital role of corporate arts funding in the United States. At a meeting for business leaders, David Rockefeller (1966) stated that traditional art patrons—primarily wealthy individuals and foundations— were no longer able 'to cope with the growing needs.' Neither did Rockefeller think that the government could meet the need, despite the establishment of the National Endowment for the Arts in 1965. The BCA should encourage corporate arts funding by providing expert counseling about opportunities in the arts, and assist arts organizations in becoming reliable and accountable partners. *Nonprofit Enterprise in the Arts* (DiMaggio 1986) describes how and why corporations get involved in arts funding.

3. *The State and Arts Funding*

Several arts funding partners have been described in the section above. The crucial issue is that different people and institutions have filled the role as public patron. *The Patron State* (Cummings and Katz 1987) documents that government funding of the arts in most countries dates back to the middle of the nineteenth century. State involvement in the arts is related closely to each country's political system, particularly to whether the state is active or passive. Since the end of the nineteenth century, most European countries have had an active state whose primary concern has been to develop social welfare programs. State arts funding began as support to individual artists of 'high quality' and in order to establish national cultural institutions as part of the nation-building process during the nineteenth century. In Italy, government arts funding was aimed—and still is—primarily towards preserving the national heritage, leaving creative arts to private funding sources. This has also been the case in Asian and African countries. The strong decentralized tradition in Germany has given the *Länder* most of the responsibility for support of the arts as well as other policy areas. The opposite is the case in France, with its strong, centralized system. Public arts funding in Great Britain has been characterized as a 'night watchman' type, the state being responsible for arts education and cultural heritage, leaving most of the performing and creative culture and arts activities to the market place. Canada represents an example of how arts funding systems can emerge from an existing social structure. Canada lacked a strong tradition of royal and aristocratic patronage of the arts. There were also almost no financially endowed individuals or estates comparable to those of Western Europe who could or would support the arts. Thus arts funding became the responsibility of the state.

Some of the most spectacular examples of indirect government arts funding were the New Deal arts programs in the USA during the 1930s and early 1940s (Mankin 1995). In 1933 and 1934, the Public Works Art project employed painters and sculptors to create arts works for public buildings. Despite its short existence, it was nevertheless 'the first time the government had subsidized an art project of national dimension' (Mankin 1995, p. 77). The following Section of Fine Arts administered several programs, the most famous being the Works Progress Administration (WPA) Art Projects, running from 1935 until 1943. It was the most comprehensive program, including writing, music, theater, and visual arts. All three projects were not aimed at arts funding as such, but reflected the Roosevelt administration's general worry about the high unemployment rate, also among artists. Thus, Relief Administrator Olin Dows stated (quoted in Mankin 1995):

> Human economic relief was the motive behind all the New Deal's art programs. This is why they were so easily accepted by the public and the politicians. If it had not been for the great depression, it is unlikely that our government would have sponsored more art than it had in the past.

These economic based programs nevertheless had long-lasting cultural effects. For instance, many of the symphony orchestras established in the United States with money from the New Deal arts programs continued after World War II and are still active.

The American government has been called the 'reluctant patron.' This is appropriate also for the British government. Before 1940 there was virtually no government support of the arts in Britain. Early in World War II, a group of private citizens established the Council for the Encouragement of Music and the Arts (CEMA) aimed at 'brightening the lives of factory workers, military personnel, and isolated civilian groups, thus fortifying national morale' (Ridley 1987, p. 228). Government became a funding partner only after the organization had been established. However, in 1945, the work done by CEMA was continued and expanded by the Arts Council, a government agency. Despite the government's reluctant stance, the British Arts Council was the first among all councils, agencies, and endowments for the arts that, during the following decades, particularly around 1964–5, were established in the United States, in Canada, and in European, Asian, African, and Latin American countries (see the UNESCO Series 1970s/1980). One recent type of indirect government arts funding is the money granted to arts programs at single events such as the World's Fair (e.g., Expo 1967 in Canada) and the Olympic Games (e.g., Albertville, France in 1984, and Lillehammer, Norway in 1994).

There are two theoretical arguments for state funding of the arts. The first is to compensate for

market failure, and the other is that the arts serve collective needs, having what cultural economists call 'external benefits.' The arts' external benefits are (Heilbrun and Gray 1993, pp. 205–9): (a) availability, that the arts are available regardless of regular use, (b) legacy for future generations—cultural heritage, (c) national identity and prestige, (d) supplementing liberal education, and (e) artistic innovation. Securing the cultural heritage and promoting national identity are the most frequently mentioned cultural policy goals throughout the world (UNESCO Series 1970s/1980s).

4. *Issues in Arts Funding Research*

Systematic arts funding research began during the 1960s, often within a broader cultural policy perspective. It can aim at explaining certain relationships in arts funding, or at finding the most effective means to implement policy goals. Schuster (1996) emphasizes the importance of distinguishing between 'pure' social science research, searching for explanations, and the more applied-oriented policy research. They are not, however, mutually exclusive. For instance, knowing that one country's tax exemption regulations enhance corporate donations to the arts may encourage new policy initiatives in other countries.

Arts funding research has been concerned with conditions within one country, it has been cross-national, or comparative. Section 2 has described how, throughout history, the dominant art patrons in different countries have changed over time, depending upon their social and political position in society, and the functions that the arts had for them. Cross-country descriptions and analyses (e.g., Cummings and Katz 1987) shed light on similarities and differences between countries. This can, in turn, lead to comparative studies, the major issues being governments' cultural policy goals and organizational structure for funding of arts and culture. The crucial question has been: does it make any difference whether the arts are funded by a government based or market based system? Heeding the problems and pitfalls involved (Schuster 1996), comparative research has revealed some tendencies.

Most European nations have a strong welfare state, being based on the principle of equal citizen rights. Thus, the arts are regarded as collective goods, the use of which can be facilitated by state appropriations directly to cultural institutions such as public libraries, museums, theaters, and orchestras. The government compensates for market failure. A market-based system, such as in the United States, emphasizes the individual's freedom as consumer. The state has no obligations towards its citizens, leaving funding of the arts to the private sector. Despite this difference in the arts funding structure, comparative research indicates that both systems favor cultural institutions that potentially can reach large audiences. The state is interested in justifying spending the tax payers' money on the arts, and potential private art patrons, corporations in particular, are looking for arenas offering large exposure. Looking at the art forms, however, research indicates that a market based funding system to a higher degree steers, for example, concert repertoires and book publishing toward popular content than a government arts funding system (Bakke 1996).

Entering the new millennium, arts funding research is faced with four crucial tendencies (Boorsma et al. 1998, Lewis 1995, Pankratz and Morris 1990). First, market elements in traditionally strong government-based arts funding countries are increasing. Will this imply more commercial art forms? Second, the fact that government comments to appropriations to the arts have changed from alluding to general arts policy goals to specifying result goals indicates increased emphasis on efficiency in government spending. What are the relevant quantitative and qualitative result indicators within the arts? This also shows that efficiency demands are not the result of privatization above. Third, will arts patronage become more diversified in the age of electronic media, bringing, for instance, art consumers in more direct contact with the artists? Also, will artists' copyrights be at stake? Fourth, these issues are linked to the ongoing globalization process, inviting researchers to look for converging tendencies in arts funding patterns as well as diversification of cultural tastes and art forms (Johnson 1993). Globalization also challenges one of the traditional purposes of state patronage, particularly in small countries and in former colonial nations, to preserve national culture identity (Chen 1998).

See also: Art and Culture, Economics of; Art: Anthropological Aspects; Art, Sociology of; Cultural Policy: Outsider Art; Culture, Sociology of; Film: History; Fine Arts

Bibliography

Bakke M 1996 Does it make any difference? Cultural policy implications: Classical music concert repertoires in Norway and the United States. *Studia Musicologica Norvegica* **22**: 7–23

Balfe J H (ed.) 1993 *Paying the Piper. Causes and Consequences of Art Patronage*. University of Illinois Press, Urbana, IL

Bandara H H 1972 *Cultural Policy in Sri Lanka*. UNESCO, Paris

Boorsma P B, Hemel A van, Wielen N van der (eds.) 1998 *Privatization and Culture. Experiences in the Arts, Heritage and Cultural Industries in Europe*. Kluwer Academic Publishers, Boston

Chen K-H (ed.) 1998 *Trajectories. Inter-Asia Cultural Studies*. Routledge, London

Cummings M C Jr, Katz R S (eds.) 1987 *The Patron State. Government and the Arts in Europe, North America and Japan*. Oxford University Press, New York

Cummings M C Jr, Schuster M D (eds.) 1989 *Who's to Pay for the Arts*? American Council for the Arts, New York

DiMaggio P 1986 *Nonprofit Enterprise in the Arts: Studies in Mission and Constraint*. Oxford University Press, New York

Gombrich E H 1995 [1950] *The Story of Art*. Phaidon, London
Heilbrun J, Gray C M 1993 *The Economics of Art and Culture. An American Perspective*. Cambridge University Press, New York
Johnson R (ed.) 1993 *Pierre Bourdieau. The Field of Cultural Production*. Polity Press, Cambridge, UK
Lewis J 1995 [1990] *Art, Culture, and Enterprise. The Politics of Art and the Cultural Industries*. Routledge, London
Mankin L D 1995 Federal arts patronage in the New Deal. In: Mulcahy K V, Wyszomirski M J (eds.) *America's Commitment to Culture. Government and the Arts*. Westview Press, Boulder, CO
Pankratz D B, Morris V B (eds.) 1990 *The Future of the Arts. Public Policy and Arts Research*. Praeger, New York
Ridley F F 1987 Tradition, change, and crisis in Great Britain. In: Cummings C C Jr, Katz R (eds.) 1987 *The Patron State. Government and the Arts in Europe, North America and Japan*. Oxford University Press, London
Rockefeller D 1966 *Culture and the Corporation*. Founding Address, Business Committee for the Arts, Inc. Business Committee for the Arts, New York
Rueschemeyer M 1993 State patronage in the German Democratic Republic: Artistic and political change in a state socialist society. In: Balfe J H (ed.) 1993 *Paying the Piper. Causes and Consequences of Art Patronage*. University of Illinois Press, Urbana, IL
Schuster J M D 1987 Making compromises to make comparisons in cross-national arts policy research. *Journal of Cultural Economics* **11**(2): 1–36
Schuster J M D 1996 Thoughts on the art and practice of comparative cultural research. In: Hamersveld I van, Wielen N van der (eds.) *Cultural Research in Europe*. Boekman Foundation & Circle, Amsterdam
UNESCO 1970s/1980s Series on studies and documents on cultural policies. UNESCO, Paris
Zolberg V 1990 *Constructing a Sociology of the Arts*. Cambridge University Press, Cambridge, UK

M. Bakke

Asceticism

The term 'asceticism' basically means a type of behavior that tends to submit an individual's impulses and desires to a systematic way of life in order to permanently gain specific virtues (athletic-heroic, ethical-moral, or religious) which are considered important by these individuals to reach ethical self-fulfillment. Such behavior, when conducted systematically and consciously, is inseparable from a real ethical doctrine that renders it desirable and even dutiful. Asceticism, often also defined by the term 'ascetism' (from Greek *ásketés*), thus means the extreme will of inner transformation by means of a strict practice to achieve what individuals consider their elective nature and which they feel called to. Giving up one's desires and restraining one's impulses become more important the more the physical and moral nature of individuals consider themselves unable to achieve the virtues they pursue throughout the course towards asceticism. From this point of view, asceticism is based, at least implicitly, on a conflict between nature and spirit, and between passion and desire on the one side and ethical-religious imperatives on the other.

1.1 The Rise of Asceticism as a Problematic Concept: A Religious Behavior

In Greek culture, the term 'ascetism' implies the methodical application to an art just as much as constant training of an athlete's or warrior's physical strength. In classical philosophy (in particular in the Sophists, Stoics, and Pythagorean schools), ascetism denotes a systematic effort of the spirit towards wisdom. In the work of Filone d'Alessandria, ascetism became a systematic effort, both moral and religious, to perfect the soul and prepare it for God's contemplation. It was, however, within a religious environment that the concept of asceticism paved the way towards its first and most widespread practice. Ascetism, considered as an attempt at a radical allegiance to the divine law by fighting against the temptations of human nature, can be found in all salvation religions based on the constant practice of virtues which are pleasing to God. Examples of this can be found in the Old Testament of Judaism, where the virtues of wisdom imply an arduous practice of moral virtues, in the first Christian communities, and in the Patristics philosophy.

It is in the context of third century Christianity that the weaving together of theological considerations and Greek philosophy came about. Ascetism became a methodology of perfection which, through self-reflection, positively coincides with meditation, prayer, and charitable work; or more negatively viewed it implies the estrangement from any alternative bond with respect to absolute unity with God. In the fourth century monastic ascetism initially arose in the form of retreat from the world so as to fight against the inner enemy which every monk sees within himself. The changeover from a hermit's life to the cenobitic community reveals the importance of practicing charity towards neighbors, as well as using everyday life as an essential means to achieve such. The Rule defines ascetism as a means of perfection, imposing, through the needs of the spiritual path, virtues of obedience (alongside those of poverty and chastity), and stresses the inner character of the search for virtues.

Following the move from Egypt and Syria towards central Europe, after the Arabian conquest, monastic asceticism had to temper the need for prayer and meditation alongside that of demanding agricultural work. With the beginning of town life and the search for unity with God in the world, asceticism continued to develop its characteristics of inner silent discipline, very different from the penitential meditation of oriental asceticism, but also completely integrated in

the social world, where it established itself with the sacramental hierarchy, life in the world, and urban ethics. In *devotio moderna*, closely related to the artisan town life in the Dark Ages, the ascetic was no longer isolated by the desert nor protected by the walls of the monastery, and had to achieve a real discipline of inner life. With the Reformation, the blessing, seen as a gift of God, became no longer a result of a subject's personal effort. A completely inner-worldly asceticism emerged, aimed at self-renunciation in career and professional service for the benefit of everyone. In this way, ascetism tended, ever more clearly, actually to integrate into the coextensive secularization process establishing itself in the modern world. In some cases, as with Puritanism, the ascetic behavior did not give rise to any specific institution; in others, like monasticism and subsequent religious orders, it gave rise to statutes, rules, orders, and institutions.

2. *Asceticism in the Secular City*

For Emile Durkheim, the detachment from worldly profane things (objects, passion, physical needs) is the root of every religious learning. All religions introduce restrictions to be observed and thus sacrifices to be made, and asceticism is therefore the actual element of each of these. In light of this, it is not a prerogative of the virtuous class only and neither does it coincide with penitential hardships, actually hypertrophies of the negative cult. Asceticism, the core of every religion, can also be found in the 'positive cult' by worshiping God with prayers, offerings, and actions pleasing to him. Not considering oneself, sacrificing one's time, renouncing, and other acts are just as much elements of separation from the profane world. Moreover, the religious rites become more sound and effective the more one disengages from activities which are in some way loved and useful.

In addition, ascetism is not only a prerogative of the religious world, but in so far as every community claims a superior identity over individuals, they also tend to demand the same methodology of renunciation that permeates religious practice from the people who are part of it. In this perspective, ascetism loses every voluntary characteristic to stand as a required substantive methodology of deprivation (negative ascetism) in order to properly approach the community where sacred means expression and an identifying point of reference. The link which Durkheim constantly outlines between religion and society shows how ascetism works in the first as well as the second. Society, therefore, similar to religion, also needs to appeal to asceticism which thus entails deprivation, renunciation, and above all, subordination of personal interests and passions to public interest: duty towards the city prevails over every other duty.

Max Weber, differently from Durkheim, does not see a structural characteristic of the sacred in asceticism, but a method to obtain the state of grace needed for the salvation of one's soul. Ascetic procedures are therefore found, at least in principle, in salvation religions. Salvation may be placed between two opposite extreme poles: on one side it is apparently a result of a subject's completely personal behavior while on the other it may be regarded as a completely free, inexplicable gift of grace from God himself. Consequently, asceticism ends up by giving rise to two strategies that are completely different in social behaviors.

Whereas awareness prevails within the determined role of an individual, ascetism becomes a search to constantly receive grace obtained through rationalization of a salvation method: from ancient Buddhism to the Jesuit Order, through the anchorites of the third and fourth century, the problem involves the constant and completely vigilant rule over one's thoughts as well as actions. All those who reach this objective develop a truly religious level within the believers' community. The ascetics end up being given privileged positions: this also occurs in Indian salvation religions as well as the Christian faith (where such religious elites paved the way to religious orders), including the Protestantism of the ascetic sects and pietistic conventicles. Ascetic religiousness takes the shape of an actively ethical behavior in this case: a religious methodology of salvation combined with the awareness that God guides such behavior. The world, along with the social relations of the average person, are not only a danger because they show signs of ethically irrational receptive pleasures, but are also profoundly conventional and utilitarian. It is this concept of banality in ordinary social relations that confers to the ascetic, at least initially, along with the rejection of the world, the undertaking to witness and transform the world as God's instrument (inner-worldly acestism). If aascetics, in rejecting the world, turn their practice toward a sensitive discipline and look towards God as a constant path of existence, in the case of inner-worldly commitment, they may become reformers of the world: examples range from Cromwell's parliament of 'saints' to the Quaker's state and different forms of radical communism practices in the pietistic conventicles.

If the whole world cannot be raised to the demanding ethics of the virtuous elite, it however remains the only possible place where such people can prove to themselves that they have full possession of their state. The worldly orders thus become a vocation that the ascetic has to achieve in a rational way. Rationality of social relations free from worldly passions (power, wealth, erotic deification of the being, violence as irrational behavior…) become a way of life that ascetics can make their own in a world which is, however, desired by God.

The methodology of salvation, in order to seek constant possession of grace, can also find expression in mystical union. The contemplative mystic actually

breaks away from the world and lives estranged from it to 'rest' in God's contemplation. Only in this extreme case does ascetism become a true *fuga mundi* and the mystic ascetic, considering it unlikely that God needs 'instruments' to carry out his will, is pleased to be a simple receptacle for the divine spirit.

Asceticism takes a different form if salvation is regarded as a completely free gift of grace given by God, who is absolutely unsearchable as to his decisions. We, therefore, face a predestined grace in which believers are sure of their salvation only when they are certain that they belong to the group of the chosen ones. Since God expresses certain rules on behaviors pleasing to him, the calling to a vocation is not only an invitation to cooperate in order to accomplish God's will but also true evidence of being among the chosen ones. Acting in compliance with God's will becomes much more than a duty: it is already a sign of salvation.

If a person devoted to good (or evil), regardless of his or her personal merits, is a result of the predestination dogma, then the world is nothing else but the outcome desired by the divine power to increase its glory. The believer may be nothing else but a precious and tireless collaborator in such project. The Weber thesis is extended and refined by Ernst Troeltsch. He distinguishes between Jesus's ethics (closer to heroism than to true ascetism) and Paolo's ethics, which is linked to depreciation of worldly life. Monarchism brings such asceticism towards rationalization of the Rule while Lutheranism completely opposes monastic ascetism to favor ascetism in the world, a self-renunciation that takes place through actively and systematically performing one's profession. Troeltsch developed various ascetic experiences in the Protestant sects by combining ascetic needs with social protest, wherein asceticism is no longer a free choice of the elite few but a dutiful behavior for all those who are part of the new church.

3. *Contemporary Developments: Asceticism as Element of Rational Social Behavior*

Durkheim, Weber, and Troeltsch all question the importance of asceticism in the edification of the modern world. For Durkheim, asceticism, once confirmed as an essential expression of the sacred, experiences notable changes but becomes inevitable in the behavioral ethos of the secular citizen at the moment religion and society end up by amalgamating, thus becoming indistinguishable since they are reciprocally functional. For Weber, asceticism, taken in its historical evolution, is the transposition, in the forum of inner spirituality, of the same principle of systemization and rationalization that forms the essence of Western civilization. For Weber and Troeltsch, the ascetic dimension is not separated by a sort of implicit protest (and also explicit in some cases) toward the surrounding society. The sectarian dimension, studied in particular by Troeltsch, is often given body by abandoning activities, behaviors, and principles of political and social rule considered to conflict with Old Testament ideals. Ascetism of entire communities which reject behaviors and acts considered unacceptable for those who, by the Grace of God, regard themselves as already redeemed, stands against the monastic ascetism of those who abandon the world to reach perfection that constantly evades them.

In contemporary sociology, the term asceticism has often been used with its general meaning relating to deferment of pleasures and unnecessary consumption. Ascetic behavior, understood as voluntary deprivation from seeking unnecessary possessions, was used in the study of consumption behaviors of social groups. In the sociology of religion it still defines the spiritual methodology of ascetics of every creed, group, or religious denomination. In particular, links with social protest forms were studied and, through these, the messianic and millenarian utopias which such protests feed on.

As far as epistemology is concerned, asceticism is quite comprehensible not only as a historical concrete form of perfection methodologies aimed at salvation but also, and above all, as the ideal Weberian type. The 'unilateral exaggeration' of one or more elements characterizing asceticism are historically traceable both in outer-worldly ascetism of the religious elite of the fourth century as well as in the sectarian inner-worldly ascetism of modern times. When traced back to its inner-worldly setting, whether religiously or laically oriented, asceticism shows its heuristic potentiality, so accounting for the wide range of religious behaviors. It not only explains the professional ascetism of the Calvinistic entrepreneur but all the lay variations—that is, the actual combinations of the ideal-typical model—of the ethics of the city.

Being the latter turned to legitimize the approval of the social contract as well as to oppose it whereas some principles diverge from the religious doctrines on which the single methodology of perfection is based.

If asceticism is destined to disappear as a historical feature, since the world rests on mechanical foundations, it continues to determine the driving principle of every constantly practiced axiological rationality. Only by systematic and constant practice of a completely vigilant conscience does axiological rationality cease to be occasional or temporary, to rise to a permanent passage of an individual's social existence. From the cells of the monks to the austerity of the revolutionary militants in every age, to the sacrifice for the city or country, through all different levels of public duties, asceticism—understood from an ideal-typical viewpoint—remains one of the elements that structures ethical behaviors raised to a voluntary methodology of social behaviors, considered as building behaviors of a new community or of rebuilding of

an already existing community. As a methodology of life, asceticism can be regarded as the crowning achievement of the utopian projects for the transformation of the world as well as a pattern of systematic practice of given historical values (more or less explicitly religious, however perceived so in their inviolability) with the object of constantly maintaining these as informing and ruling principles of the entire social existence of individuals. The principle of rationalization, implicit in inner-worldly asceticism as a methodology of life oriented towards constant preservation of virtues, makes it a structural component of social behavior able to continuously invite participants towards a strict and not just occasional coherence of conformity with values.

See also: Altruism and Self-interest; Buddhism; Catholicism; Christianity Origins: Primitive and 'Western' History; Durkheim, Emile (1858–1917); Ethics and Values; Historiography and Historical Thought: Christian Tradition; Historiography and Historical Thought: Jewish Tradition; Judaism; Norms; Religion: Definition and Explanation; Religion: Evolution and Development; Religion: Morality and Social Control; Religion, Sociology of; Weber, Max (1864–1920)

Bibliography

Abbruzzese S 1995 *La vita religiosa. Per una sociologia della vita consacrata*. 1st edn. Guaraldi, Rimini, Italy

Boudon R 1999 *Le sens des valeurs*. Presses Universitaires de France, Paris

Bourdieu P 1979 *La distinction. Critique sociale du jugement*. Editions de Minuit, Paris

Brown P 1978 *The Making of Late Antiquity*. Harvard University Press, Cambridge, MA

Brown P 1988 *The Body and Society. Men, Women and Sexual Renunciation in Early Christianity*. Columbia University Press, New York

Durkheim E 1963 *L'Éducation morale*. Presses Universitaires de France, Paris

Durkheim E 1968 *Les formes élémentaires de la vie religieuse. Le système totémique en Australie*. Presses Universitaires de France, Paris

Guillaumont A 1979 *Aux origines du monachisme chrétien. Pour une phénoménologie du monachisme*. Abbaye de Bellefontaine, Bégrolles en Mauges, France

Séguy J 1964 L'ascèse dans les sectes d'origine protestante. *Archives de Sociologie des Religions* **IX**(18): 55–70

Séguy J 1980 *Christianisme et société, Introduction à la sociologie de Ernst Troeltsch*. Cerf, Paris

Séguy J 1999 *Conflit et utopie, ou réformer l'Église. Parcours wébérien en douze essais*. Cerf, Paris

Troeltsch E [1912] 1931 *The Social Teaching of the Christian Churches* [trans. O. Wyon]. George Allen and Unwin, London

Weber M [1922] 1963 *The Sociology of Religion* [translated E. Fischoff]. Beacon Press, Boston

Weber Max, [1922] 1965 *The Protestant Ethic and the Spirit of Capitalism* [trans. T. Parsons]. Allen and Unwin, London

S. Abbruzzese

Assessment and Application of Therapeutic Effectiveness, Ethical Implications of

1. Definitions

Medical therapy is an intervention which aims at removal or at least attenuation (and in a broader sense also at prevention) of diseases or disorders, as well as at subjective suffering due to such illnesses of the individuals concerned. In general, understanding the first aspect is focused in the term cure (and prophylaxis), the second one in the term care. Medical cure comprises somatic (by drugs, surgery, ray, or physical) and psychotherapeutic (by a broad range from counseling to psychodynamic to behavioral) interventions. Medical care comprises the embedding of the cure in a sufficient (in number) and satisfying (by personal quality and qualification) setting of personnel, facilities (beds or treatment places), and supply. Removal is related more to a causally, attenuation more to a symptomatically effective therapy.

Intervention has effects, wanted and unwanted. Wanted effects with regard to the aim of therapy are called therapeutic effectiveness (or efficacy) and unwanted effects are side-effects. Occasionally effectiveness is limited to application in everyday practice, and differentiated from efficacy in controlled clinical trials. The relation of therapeutic effectiveness to its costs, both medically in terms of side-effects and risks, and particularly economically in terms of financial burdens, is labeled as efficiency. The assessment of these different outcome aspects of therapeutic interventions has some implications.

Effects, both physiological and psychic ones, can be observed directly. The assessment of effectiveness, however, implies presuppositions with regard to more or less complex theoretical constructs such as illness or needs or therapeutic aims (Winterer and Herrmann 1996). A behavioral model of mental illness will give reason for aiming therapeutically at symptom relief whereas an intervention based on a psychodynamic model may be oriented towards a maturing development of personality, i.e., to follow either an adaptational or an emancipatory model of psychotherapy (Helmchen 1998); or, a sociogenic or even an antipsychiatric model of mental illness may lead rather to social interventions with the aim of acceptance of the disorder or the disordered subject respectively, by

both the individual and society, whereas the medical model aims more likely at symptom relief and recovery of functional capacity by drug treatment. Furthermore, the aims of therapy may be different from the perspectives of the patient (e.g., well-being), of the therapist (e.g., maturation of personality), and of the insurance company (e.g., fitness for work) (Kottje-Birnbacher and Birnbacher 1998). Finally, but not the least, these constructs depend on contemporary context.

Hence, assessment of effectiveness of a therapeutic intervention means to evaluate the fitness of achieving these construct-related aims. This will be done either in general by research in order to establish the strength and specificity of the relatedness of the intervention to the aim, i.e., to find an evidence-based indication, or in the individual case, by judging whether the indicated intervention actually did achieve the aim. The strength of the specific relatedness of an intervention to a defined aim may be a continuous variable, but will be categorized by cut-offs for being usable as an indicator of therapeutic effectiveness, e.g., the reduction of a symptom score below a defined threshold. Here, again, different standards are in use: some accept a symptom reduction of already 50 percent, others admit only a complete relief of symptoms as therapeutic effectiveness. Therefore, universally accepted aims for therapeutic interventions do not exist as well as 'there are no universally accepted standards of efficacy and safety' (Kane and Borenstein 1997).

This seems to be even more valid for the assessment of efficiency because it has to deal with much more complexity: a divorce as a side-effect of strengthening the autonomy by psychotherapy may be acceptable for one patient, but not for the other (notwithstanding the problems of informed consent with regard to such a potential side-effect). Tardive dyskinesia as an irreversible side-effect of a neuroleptic long-term medication urgently needed for suppression of agonizing schizophrenic symptoms will be put up with by one patient but not by another. Although from the medical point of view in this case the application of a drug with the same antipsychotic effectiveness, but without the risk of this unwanted effect, undoubtedly indicated the doctor to be efficient with regard to his limited budget, nevertheless he or she is urged to decide in favor of a less expensive drug. If he prescribes the most effective but also the most expensive drug, he perhaps cannot prescribe drugs urgently needed by other patients. With regard to subsequent burdens to society, it could be much more efficient to choose a cost-intensive treatment for help-seeking people with subthreshold morbidity (morbid states below the threshold of officially used operationalized diagnostic systems such as *DSM-IV* (*Diagnostic and Statistical Manual* of the American Psychiatric Association) or *ICD-10* (*International Classification of Diseases* of WHO)) than to withhold therapeutic interventions from them (Helmchen 1999).

The assessment of effects, effectiveness, and efficiency of a medical intervention is most developed in drug therapy. Therefore, the following considerations will be related to drug therapy, although the criteria of its evaluation cannot simply be applied to other types of medical interventions. Both the assessment of effectiveness and safety of a potentially therapeutic drug by research, and the efficient application of a marketed drug to ill persons in practice, imply ethical problems.

2. *Assessment*

In therapeutic research the basic ethical problem on the one hand is to expose the participating subject to a potentially ineffective intervention or to unknown risks. However, without such research on the other hand, the risk is to expose people to the uncontrolled risks of an intervention with a drug, the effectiveness and safety of which are not really known. This had been called the ethical-paradox of therapeutic research: it may be unethical to subject a person to a controlled clinical trial, and it may be even more unethical to introduce an unproven drug into the market (Helmchen and Müller-Oerlinghausen 1975) because in practice the application of an ineffective intervention will not only produce unnecessary costs for the patient but, much worse, may withhold him from an effective therapy. Thus, the scientific evaluation of therapeutic effectiveness is a moral imperative (Wing 1981). Also the refunding of medical interventions by the insurance companies is made increasingly conditional on the scientifically proven efficacy of new treatments. The need and the risks of therapeutic research have produced highly specified regulations (laws and guidelines) which among other things define standards of efficacy and safety as well as of participation of fully informed and voluntarily consenting subjects in such therapy research. They oblige the researcher to minimize the risks and inconveniences for patients (Helmchen 1990), particularly:

(a) to formulate the research question as specifically as possible as the basis of a biometric calculation of the needed sample size in order to avoid the inclusion of more persons than needed and, thus, to worry patients unnecessarily, as well as of fewer people than needed for a valid result and thus to burden all participants for nothing.

(b) to choose adequate selection criteria in order to minimize the drop-out rate and to achieve at least some generalizability of the results. Otherwise the results may be invalid or not applicable to other patients than those of the study sample and in that case patients may be burdened unnecessarily. Often therapy-resistant patients will be excluded from clinical trials. However, the need to overcome therapy-resistance is much more urgent than, for example, to test a further antidepressant drug of an assumed

known therapeutic mechanism for a questionable quantitative improvement of existing standard therapies.

(c) to not withhold effective therapies from patients, e.g., by use of placebo-controls. However, in order to control the effect of subjective influences and spontaneous remissions on the evaluation of effectiveness placebo-controls may be unavoidable. Thus, e.g., in the case of depression, a placebo-controlled trial may be ethically justifiable only in mild depressions in which the effectiveness of available antidepressant drugs is questionable (Paykel et al. 1988) or in therapy-resistant depressions. Otherwise there remains no alternative but to test the experimental drug against a standard drug. However, in order to avoid both the beta-error and a large sample size, one should test for falsification of the zero-hypothesis, i.e., that the investigational drug is more effective than the standard. Since, as in many countries regulatory agencies do not demand evidence of superiority of the investigational drug, an investigator can decide for himself to choose criteria of superiority, e.g., more than 20 percent of effectiveness.

All such details of testing therapeutic interventions have to be considered with regard to ethical principles, the primary essential ones of which in medicine are (a) respect for autonomy and dignity of the patient, (b) beneficence, (c) nonmaleficence, and (d) justice (Beauchamp and Childress 1994). Respect for autonomy requires improving the patient's understanding of the disorder (its symptoms, course, outcome, and perhaps causes) and of the benefits and risks of both, the recommended and alternative therapeutic interventions as well as the elucidation of their meaning for the patient's personal values, especially his expectations towards the therapy with the consequence of an informed consent. Most therapeutic interventions do not fulfil the therapeutic ideal (Ehrlich's 'therapia magna sterilisans'): to achieve immediate and complete elimination of the disorder in all patients with the disorder for which the therapy is indicated (specific efficacy) with no unwanted effects (safety). Usually some benefit can be obtained only at the cost of some more or less serious unwanted effects. Therefore, the relationship between beneficence and nonmaleficence must be evaluated and this risk-benefit-estimation should be imparted to the patient—but only in the case the physician comes to the conclusion that this estimation ethically justifies the intervention, e.g., to expect better as existing effectiveness and/or lesser or only minimal risks and burdens for the patient.

Furthermore, all scientifically based therapeutic trials do not aim only at the benefit of the study patient but also at a better knowledge for the benefit of other patients too. The problem becomes more difficult if such broadening of the perspective beyond the individual patient leads to a benefit-risk-estimation which may be less favorable for the study patient than for other patients in other stages of the disorder or in the future. The ethical consequence is the obligation to minimize both the burden and the number of patients needed for a valid result. In this context the understanding of the principle of justice means predominantly an equal distribution of burdens and benefits, i.e., among all members of a group, e.g., all patients with a defined disease, or among all participants of an insurance company. Patients receiving an effective therapy are beneficiaries of the trial burden of former patients and, therefore, could be asked to participate in a new trial for improving existing therapies.

The steadily increasing specification (and red tape) of controlled clinical trials of new drugs does not only reflect these problems as well as further relevant difficulties such as those of full information of the patient, or of assessment of his competence to consent, or of subjecting him to strong standardization procedures; it also may point to corresponding difficulties in much less elaborated fields of psychotherapeutic or social interventions. These latter mentioned difficulties may be both greater and less considered as a still valid quotation from Wing may illustrate:

> There is nowadays a fairly general acceptance of the view that new drugs should not be introduced without being tested. There is no similar consensus in respect of social treatments, most of which become firmly adopted before they have been thoroughly examined. The harm that may come from the application of misguided social theories, or the misapplication of sensible theories, is at least as great that can follow the prescription of a harmful drug or an unnecessary course of psychotherapy. In fact, it can be much greater, since harmful social practices can become institutionalized into the structure of a complete psychiatric service. The "custodial era" in psychiatry, although it was not as black as it has sometimes been painted, nevertheless illustrates how the practices inherent in the concept of the "total institution" can become generally and uncritically adopted, even though many of them were quite unnecessary and demonstrably harmful' (Wing 1981, p. 278).

3. *Application*

The efficacy of a drug developed in the highly artificial setting of a controlled clinical trial may be considerably altered in the natural setting of therapeutic practice due to uncontrollable influences such as co-morbidity, co-medication, compliance, lifestyle, etc. The evaluation of these effects of natural settings upon effectiveness as well as upon safety of a drug in the individual patient is the task of the treating physician, with regard to general knowledge it is the purpose of so-called phase IV research (Linden 1999).

The therapist has to consider the consequences to treat against the consequences not to treat, not only from a factual point of view but also in an ethical perspective. To treat predominantly requires the evalu-

ation of the risk of unwanted effects against a possible worsening of the patient's condition, e.g., the risk to apply lithium, a drug with proven effectiveness for relapse prevention in recurrent affective disorders, in a patient with unreliable compliance may be worse with regard to—principally controllable but in this case uncontrolled—side-effects than a relapse of the disease. Not to treat evidently has much more ethical implications. Nowadays, in times of managed care, the physician has to keep strongly (under penalty of recourse) to his limited budget and this means that he has to ration principally available therapies not only according to their evidenced effectiveness but also to the severity and acuity of need of patients and to their costs (APA 1997). Costs nowadays are predominantly paid from (third party) collective resources. Therefore, the therapist has obligations not only towards his individual patient but also towards the according collective. Physicians from a traditional and well reasoned point of view argue against this double-agent role of physicians as unethical (Simon 1987). However, Physicians Associations at least in some countries such as Canada, Germany, and the UK, tell their members that 'Finite resources can never match potentially infinite demands or expectations…' (consequently)… 'it is doctor's ethical duty to use the most economic and efficacious treatment available' (British Medical Association, cit. Sabin 1996). But besides the efficient application of effective therapies in terms of costs (both side-effects and financial burdens) the therapists as well as society are asked to justify ethically, with regard to justice, to apply or not to apply an effective therapy. This question arises particularly in cases such as antidementive drugs which have a significant but only small effectiveness on the progression of dementia making it possible for a demented person to stay longer in his familiar home before a transfer to institutional care becomes unavoidable (Max 1996) or, lipid reducing drugs which are effective in the prevention of arteriosclerotic diseases and which are so expensive that a treatment of all persons who are in need of it cannot be paid for (Kübler 1999). Furthermore, the physician has to consider the scope of informing the patient, e.g., whether he or she has to give information of a better alternative of which the patient cannot get hold.

4. Conclusion

Therapeutic effectiveness can be assessed only by relating the effects of a therapeutic intervention to a therapeutic aim. Its elucidation with and by the patient is one objective of the informed consent process which is required by respect for autonomy and dignity of the patient. Relating beneficence to nonmaleficence is the medical part of the assessment of efficiency of a therapy, relating this to the financial costs and limitations is its economical part. Particularly the latter provokes the question of justice. It belongs to the obligations of therapists and society to reflect the ethical implications of the medical, as well as the financial aspects, of assessment and application of therapeutic effectiveness and efficiency.

See also: Ethical Dilemmas: Research and Treatment Priorities; Ethics for Biomedical Research Involving Humans: International Codes; Medical Experiments: Ethical Aspects; Pecuniary Issues in Medical Services: Ethical Aspects; Placebos, Use of: Ethical Aspects; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Clinical Trials; Psychotherapy: Ethical Issues

Bibliography

American Psychiatric Association 1997 *The Psychiatrist's Managed Care Primer*, 1st edn. American Psychiatric Press, Washington, DC

Beauchamp T L, Childress J F 1994 *Principles of Biomedical Ethics*. Oxford University Press, New York

Helmchen H 1990 Ethical problems and design of controlled clinical trials. In: Benkert O, Maier W, Rickels K (eds.) *Methodology of the Evaluation of Psychotropic Drugs*. Springer-Verlag, Berlin, pp. 82–8

Helmchen H 1998 Ethische Implikationen von Psychotherapie (Ethical Implications of Psychotherapy). *Nervenarzt* **69**: 78–80

Helmchen H 1999 Care of people with psychiatric subthreshold disorders: Ethical challenges. In: Guimón J, Sartorius N (eds.) *Manage or Perish? The Challenges of Managed Mental Health Care in Europe*. Kluwer Academic/Plenum Publishers, New York, pp. 429–40

Helmchen H, Müller-Oerlinghausen B 1975 The inherent paradox of clinical trials in psychiatry. *Journal of Medical Ethics* **1**: 68–73

Kane J M, Borenstein M 1997 The use of placebo controls in psychiatric research. In: Shamoo A E (ed.) *Ethics in Neurobiological Research with Human Subjects. The Baltimore Conference on Ethics*. Gordon and Breach Publishers, Amsterdam, pp. 207–14

Kottje-Birnbacher L and Birnbacher D 1998 Ethische Aspekte bei der Setzung von Therapiezielen. In: Ambühl H, Strauß B (eds.) *Therapieziele*. Göttingen, Germany, pp. 15–31

Kübler 1999 Diskussionsbeitrag. In: Häfner H (ed.) *Gesundheit—unser höchstes Gut?* Springer, Heidelberg, Germany, pp. 198–9

Linden M 1999 Die Phase IV der Therapie-Evaluation. *Nervenarzt* **60**: 453–61

Max W 1996 The cost of Alzheimer's Disease. Will drug treatment ease the burden? *PharmacoEconomics* **9**(1): 5–10

Paykel E S, Hollyman J A, Freeling P, Sedgwick P 1988 Predictors of therapeutic benefit from amitriptyline in mild depression: a general practice placebo-controlled trial. *Journal of Affective Disorders* **14**: 83–95

Sabin J E 1996 Is managed care ethical care? In: Lazarus A (ed.) *Controversies in Managed Mental Health Care*. American Psychiatric Press, Washington, pp. 115–26

Simon R I 1987 The psychiatrist as a fiduciary: avoiding the double agent role. *Psychiatric Annals* **17**: 622–6

Wing J 1981 Ethics and psychiatric research. In: Bloch S, Chodoff P (eds.) *Psychiatric Ethics*. Oxford University Press, Oxford, pp. 277–94

Winterer G, Herrmann W M 1996 Effect and Efficacy—on the function of models in controlled phase III trials and the need for prospective pharmacoepidemiological studies. *Pharmacopsychiatry* **29**: 135–41

H. Helmchen

Asset Pricing: Derivative Assets

1. Introduction

A derivative instrument is a financial contract in which the *cash flow* received by the holder is linked to the *price* of another asset or to another financial variable (such as an interest rate). Typical examples of such contracts include forward contracts, swaps, and, of particular importance here, *options*.

This article is about the theory of derivative pricing, that is, the valuation of derivative contracts. Although there are a number of important early contributions, these tended to founder on the problem of determining a discount rate that would account correctly for the risk of the cash flows (see, for example, Sprenkle 1961). The modern theory essentially begins with the seminal contributions of Black and Scholes (1973) and also of Merton (1973). (Scholes and Merton shared the Nobel Prize for economics in 1997. If Black had not died (in August 1995) it seems certain that he would have shared the prize. For an account of the development of the Black–Scholes model see Black 1989; for further discussion see Duffie 1998, Schaefer 1998, and Jarrow 1999.) This theory shows that, under certain conditions, it is possible to determine the price of an option, *relative to the price of the asset on which the option is written*, using only the condition that the prices do not offer investors an opportunity for pure arbitrage (a free lunch). Beyond that, the theory makes no assumptions regarding the preferences of agents. In particular, it shows that the relation between the price of an option and that of the underlying asset does not depend on the attitude of agents towards risk.

Although options, forward contracts, and, perhaps, other forms of derivatives have probably existed since the earliest times (Ingersoll 1989 gives the example of the code of the Hammurabi (c. 1780 BC) which contains a provision for contingent debt contracts), the last 30 years of the twentieth century have witnessed a vast increase in the number and variety of contracts and also in the volume of trading. A precise estimate of the scale of derivatives trading is difficult to obtain, but a recent report by the Group of Thirty (1993) estimated the total notional amount of over-the-counter (OTC) derivatives on interest rates and currencies worldwide at around $30 trillion. For the 50 largest US banks the replacement cost (roughly, the net present value) of their derivatives portfolios was $140 billion, or 11 percent of their total asset values. Derivative instruments, by any measure, are now of central importance to the world financial system and it is interesting to speculate whether this revolution would, or indeed could, have occurred without the development of derivative pricing theory.

Interestingly, at the time of Black, Scholes, and Merton's [BSM] seminal contribution, financial options (a *financial option* is one in which the underlying asset is a financial asset, for example, a stock or a bond, as distinct from, say, a commodity) themselves were, in Merton's words, '… specialized and relatively unimportant financial securities …' (Merton 1973, p. 141). Then, the most promising applications of the new theory appeared to be those that arose from the isomorphism between financial options and corporate liabilities. An important example, noted in the original paper by Black and Scholes, is corporate debt where the equity holders in a limited liability firm that has debt outstanding have an option to default if the amount owed to lenders is greater than the value of the firm. As things have turned out, however, the unpredicted and dramatic growth in the market for financial options markets has made this, rather than corporate liabilities, the principal area of application.

2. What is a Derivative?

Consider a security, for example, a stock, with current (time t) price S_t and another security, for example, an option on the stock, which, at some future date T, pays a single cash flow, Y_T, that depends on the price of the stock at that time. Thus

$$Y_T = g(S_T) \tag{1}$$

where $g(.)$ is some function. The security defined by Eqn. (1) is called a *derivative* security because its cash flow, Y_T, and therefore its value, derives from the price another security, the *underlying* security. The function $g(.)$ defines the type of derivative and is, in effect, specified in the contract. Some common examples are given below

$$\begin{aligned}
&\text{Forward (buy) contract} && g(S_T) = S_T - X \\
&\text{Forward (sell) contract} && g(S_T) = X - S_T \\
&\text{Call Option} && g(S_T) = \mathrm{Max}(S_T - X, 0) \\
&\text{Put Otion} && g(S_T) = \mathrm{Max}(X - S_T, 0)
\end{aligned} \tag{2}$$

In Eqn. (2) X is the *contract* (or *strike*) price at which the underlying asset is bought, in the case of a call option and a forward contract to buy the asset, or

sold, in the case of the put option and a forward contract to sell. The examples given in Eqn. (2) are the simplest and commonest forms of derivatives but there are many others. In some the cash flow on the derivative depends on the *history* of the price of the underlying asset up to the payment date, T, rather than on S_T alone; in others it depends on the prices of several underlying assets rather than one. In yet others the cash flow depends on some quantity that may be neither an asset price (e.g., the yield-to-maturity on a bond) nor even a function of an asset price (e.g., changes in the credit rating of a corporate bond).

Finally, the word derivative, although accurately suggestive of one important feature of these instruments, is unfortunate in that it may suggest that they are in some sense 'less important' than the underlying assets to which they are linked. Often this is quite untrue. Many derivative markets are large and highly liquid and, in some cases, for example, the interest rate swap markets in countries such as the US and the UK, some traditional contracts, for example, loans, may be priced with reference to the derivative rather than the other way around. In such cases it becomes difficult to tell the tail from the dog.

3. Derivative Pricing vs. General Asset Pricing

The price of every risky asset, including derivatives, depends on agents' expectations of the future cash flows and on their willingness to bear risk; in other words, on their preferences. (For a description of modern asset pricing theory see Duffie 1996: Campbell 2000 reviews both the theory and empirical evidence.) What, then, is special about the problem of pricing derivatives as distinct from other risky assets? The key difference is that in a number of important cases, including the Black–Scholes model, while the prices of *both* the derivative and the underlying security depend on expectations and preferences, the *relation* between the two does not. In this case, given the price of the underlying security, the price of a derivative may be determined independently of agents' expectations of future cash flows and their preferences.

The economic principle that is used to establish the link between the price of a derivative and the price of its underlying asset is *no-arbitrage*. An arbitrage opportunity is defined as one that requires no capital, has a positive probability of a profit, and a zero probability of a loss. In other words, it provides 'something for nothing' and any rational investor, who prefers more to less, irrespective of their attitude towards risk, would choose to exploit such opportunities. In a well-functioning market, therefore, arbitrage opportunities should not exist and, in the economy studied by Black and Scholes, it turns out that the conditions which eliminate the possibility of arbitrage between a derivative and its underlying asset are also sufficient to determine the relation between their prices.

4. Pricing Nonlinear Contracts Where Prices Follow a Binomial Process

The most important and distinctive aspect of derivative pricing theory is its application to contracts where the function $g(.)$—the relation between the cash flow on the derivative and the price of the underlying asset—is nonlinear. This is the case for options while for forward contracts and many swap contracts the relation is linear (see Eqn. (2)). As a first step we analyze a linear contract; the treatment of options, while more complex, is conceptually very similar.

4.1 Pricing a Forward Contract

Consider a forward contract to buy one unit of an asset at a price of X. Using the notation introduced above, the value of the contract at maturity is given in Eqn. (2) and is a linear function of the price of the underlying asset:

$$g(S_T) = S_T - X$$

The key concept in computing the value of this and all other derivative contracts is that of the *self-financing replicating portfolio*: a position containing the underlying asset, combined with borrowing or lending, which, after it is initially established, (a) neither requires additional investment nor generates cash outflows ('*self-financing*') and (b) is guaranteed to have exactly the same value at time T as the forward contract ('*replicating*'). If the cost of the replicating portfolio at some earlier time $t < T$ is V_t, say, then this must also be the *price* of the forward contract, H_t. If this were not the case, arbitrage profits could be made either by buying the forward contract and selling the replicating portfolio (if $H_t < V_t$) or doing the reverse (if $H_t > V_t$). In other words, to prevent arbitrage, the price of a derivative security must be equal to the cost of its replicating portfolio.

In the case of the forward contract in the example, the replicating portfolio consists simply of holding one unit of the underlying asset and borrowing an amount equal to the present value of the contract price, X. Thus, the cost of this portfolio at time t, V_t, is given by:

$$V_t = S_t - e^{-r(T-t)}X \tag{3}$$

where r is the (continuously compounded) riskless interest rate. Clearly, at time T the value of this portfolio is equal to S_T–X, the cash flow on the forward contract. Notice that the *price* of the forward contract is to be distinguished from the so-called *forward price* of an asset. The latter is simply the contract price, X, that sets the price of the contract to zero. Solving Eqn. (3) for the value of X at which V_t is zero, the forward price, F_t, is obtained as

$$F_t = e^{r(T-t)}S_t \tag{4}$$

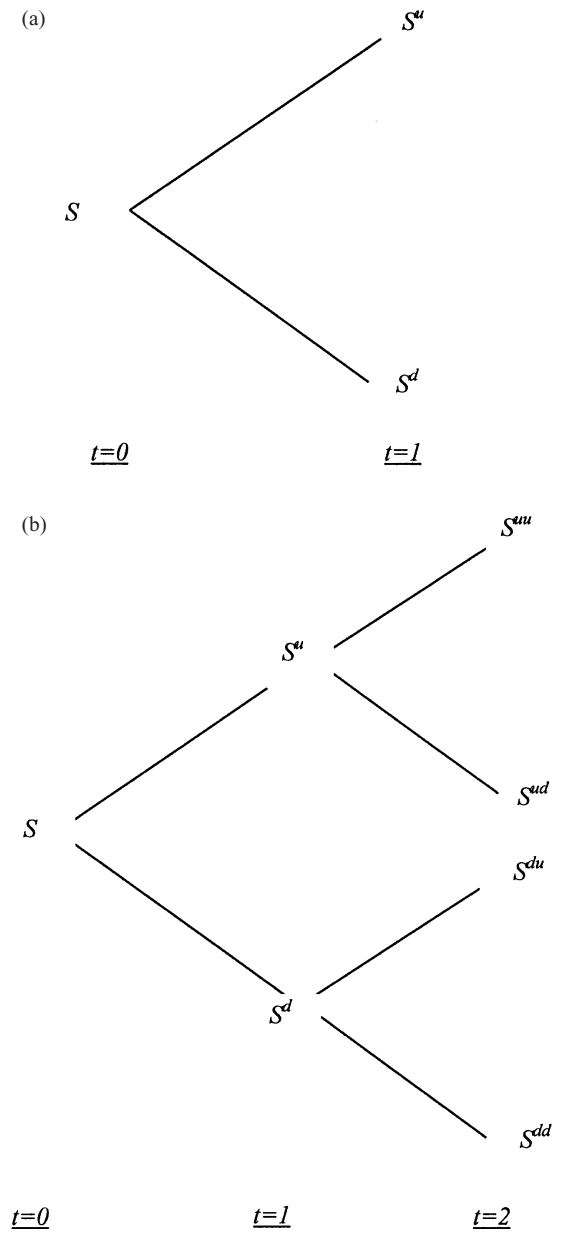

Figure 1
(a) Single period binomial model. (b) Two-period binomial model

that is, the current spot price compounded up to the maturity date at the riskless interest rate.

Although straightforward, this example illustrates accurately the logic of valuation by replication used in the BSM approach. However, it is by no means clear that, in the case of nonlinear contracts, replication is actually *feasible*. Whether it does depends mainly on the process followed by the price of the underlying asset—and, to some extent, on the nature of the payoff function, $g(.)$. In some cases, notably (a) the Brownian motion case considered by BSM and (b) the binomial case that we discuss next, replication remains feasible. In others, for example, where price of the underlying asset experiences 'jumps' of a random size, replication is not possible and, in these cases the relation between the price of a derivative and the underlying asset cannot be established from no-arbitrage conditions alone.

4.2 Pricing an Option

We consider first the problem of pricing an option where the underlying asset price follows a binomial process in discrete time (Fig. 1(a)). Thus, if the current price of the underlying asset is S, then, at time $t+1$, the price will take one of two values. S^u (the 'up' price) and S^d (the 'down' price). In this example, the timing of prices is clear and, to simplify notation, the time-subscript has been dropped. Interestingly, the probability of these two states is not important; what is critical is that these are the only two states with a nonzero probability.

The cash flows to the option buyer in the two states are denoted C^u and C^d. If the option were a call, for example, then C^u and C^d would be given by Eqn. (5) below but the argument below does not, in fact, depend on the type of option

$$\begin{array}{ll} \text{'up'} & C^u = Max(S^u - X, 0) \\ \text{'down'} & C^d = Max(S^u - X, 0) \end{array} \tag{5}$$

Now, exactly as in the case of the forward contract, we calculate the price of the derivative as the cost of its replicating portfolio. The composition of this portfolio is simple to compute: suppose that we hold Δ units of the underlying asset and purchase B units of a riskless 'bond' which pays one unit at time $t+1$. Denoting the riskless interest rate, as before, as r per unit time, the price per unit of this bond is e^{-r}. Equating the time $t+1$ value of this portfolio and the value of the option in the two possible states, we have

$$\begin{array}{ll} \text{'up'} & \Delta S^u + B = C^u \\ \text{'down'} & \Delta S^d + B = C^d \end{array} \tag{6}$$

From Eqn. (6) the composition of the replicating portfolio is easily calculated as

$$\Delta = \frac{C^u - C^d}{S^u - S^d} \quad \text{and} \quad \mathrm{B} = \left[\frac{C^d S^u - C^u S^d}{S^u - S^d} \right] \tag{7}$$

As before, because this portfolio replicates exactly the payoff on the option *at maturity*, the *current* cost of the portfolio and the *current* price of the option must be

equal. If not, pure arbitrage is possible. Thus, the option price, C, is given by

$$C = \Delta S + Be^{-r} \tag{8}$$

where Δ and B are given in Eqn. (7).

4.3 State Prices and Risk Neutral Probabilities

A little rearrangement of Eqns. (7) and (8) allows us to write the pricing relation in way that provides a different, and very useful, perspective on the pricing formula. First, defining

$$\phi^u = \frac{S - e^{-r}S^d}{S^u - S^d} \quad \text{and} \quad \phi^d = \frac{e^{-r}S^u - S}{S^u - S^d} \tag{9}$$

we may write the option price as

$$C = C^u\phi^u + C^d\phi^d \tag{10}$$

In Eqn. (10) ϕ^u and ϕ^d may be interpreted as 'state prices,' that is, the prices of 'pure securities,' analogous to Arrow–Debreu prices, which pay one unit in a particular state and zero in all other states. The option price is then simply the sum of the cash flow in each state times the corresponding state price. The fact that the state prices are uniquely identified here shows that the market is *complete*, that is, that for each state it is possible to construct a portfolio that replicates the cash flow on the corresponding pure security. If the market is complete then it possible to replicate the cash flows on *any* security, irrespective of the pattern of cash flows across states. In particular, we can replicate the cash flows on nonlinear contracts.

Market completeness is thus a critical element in the theory of derivative pricing. In the binomial case the market is complete both in the single period case just discussed and in the multiperiod case (see Fig. 1b). As we discuss below, completeness also holds in the continuous time framework of the BSM model.

Eqn. (10) may be further restated in a different and highly intuitive form. First, using Eqn. (10) to value a riskless bond paying one unit at time $t+1$ we have that

$$e^{-r} = \phi^u + \phi^d \tag{11}$$

Second, since they represent the prices of claims on positive (unit) cash flows in particular states, the ϕ's are strictly positive and may therefore be normalized to add to one. Defining

$$\pi^u \equiv \frac{\phi^u}{\phi^u + \phi^d} \quad \text{and} \quad \pi^d \equiv \frac{\phi^d}{\phi^u + \phi^d} \tag{12}$$

and using Eqn. (11) we have that

$$\pi^u = e^r\phi^u \quad \text{and} \quad \pi^d = e^r\phi^d \tag{13}$$

Substituting for ϕ^u and ϕ^d in Eqn. (10), the option price may be written as

$$C = e^{-r}(\pi^u C^u + \pi^d C^d) \equiv e^{-r}\hat{E}(\tilde{C}) \tag{14}$$

where, for the second equality, we interpret the π's as 'probabilities'—since the π's are positive and add to one, we are free to do this—and $\hat{E}(.)$ means 'expectation under the probability distribution defined by the π's.' In Eqn. (14) the expected cash flow under the π's is discounted at the riskless interest rate, as if agents were risk-neutral, and for this reason the π's are termed *risk-neutral probabilities*. We emphasize, however, the economic content of this representation comes from the fact that the ϕ's are *prices*. In fact, as Eqn. (13) shows, each π is simply equal to the corresponding ϕ compounded up at the riskless interest rate. From Eqn. (4) we see that this implies that the risk neutral probabilities are simply the *forward prices* of Arrow–Debreu securities.

4.4 The No-arbitrage Condition in Terms of Risk-neutral Probabilities and State Prices

The no-arbitrage argument presented above shows that if a self-financing portfolio exists which replicates exactly the cash flows on another asset (e.g., an option) then the price of the option must equal the cost of this portfolio or an arbitrage opportunity exists. This is what Eqn. (8) represents: the option price is equal to the sum of the price times the quantity of each asset in the replicating portfolio. This Eqn. is easily recognizable as a no-arbitrage condition. However, Eqns. (10) and (14) are merely reformulations of this same condition—in terms of state prices and risk-neutral probabilities, respectively—and are, therefore, equivalent expressions of the no-arbitrage condition.

Equivalence between the absence of arbitrage and the existence of state prices or risk-neutral probabilities holds very broadly and is by no means limited to the binomial setting. We now derive this result in a more general, discrete state setting.

We first show that there is no arbitrage if and only if there exists of a set of ϕ's under which the value of each asset is equal to its price. To see this, let $\boldsymbol{A}$ $(N \times M)$ denote a matrix of cash flows to N securities in M states and $\boldsymbol{p}$ $(N \times 1)$ denote the vector of prices of the N securities. (In our example, the number of assets is three—the stock, the bond, and the option—and the number of states is two—'up' and 'down'). An *arbitrage portfolio*, $\boldsymbol{x}$ $(N \times 1)$ is a vector of portfolio holdings which satisfies:

$$\begin{aligned} x'p &= 0 \\ x'A &\geqslant 0 \end{aligned} \tag{15}$$

where $\geqslant$ denotes semipositive. (A 'semipositive' relation, $a \geqslant 0$, is one in which all elements are non-

negative and at least one is strictly positive). Thus, a portfolio, $\boldsymbol{x}$, which satisfies Eqn. (15) has a zero cost and provides nonnegative cash flows in all states and a strictly positive cash flow in at least one state; in other words it promises something for nothing.

If there is no solution to (15)—i.e., if there is no arbitrage—then, from Farkas's lemma (see, e.g., Gale 1960), there exists an $(M \times 1)$ vector $\phi > 0$ such that:

$$A\phi = p \tag{16}$$

The ϕ's in Eqn. (16) correspond to the state prices introduced in Eqns. (9) and (10) above. Thus, Eqn. (16) states that there is no arbitrage if and only if there exists a strictly positive vector of state prices such that the value of each security is equal to its price. If such a vector, ϕ, exists then, as before, we may define a corresponding vector, π, of risk-neutral probabilities by normalizing the ϕ's (and noting that $\sum_i \phi_i = e^{-r}$)

$$\pi_j \equiv \frac{\phi_j}{\sum_{i=1}^{M} \phi_i} = e^r \phi_j \quad j = 1, \ldots, M \tag{17}$$

which allows Eqn. (16) to be rewritten as

$$p = A\phi = e^{-r} A\pi = e^{-r} \hat{E}[CF] \tag{18}$$

In Eqn. (18) $\hat{E}[CF]$ denotes the expected value of the cash flows computed using the π's, defined in Eqn. (17), as 'probabilities,' exactly as in Eqn. (14). Eqn. (18) is exactly equivalent to Eqn. (16); in other words there is no arbitrage if and only if there exists a vector of (risk-neutral) *probabilities* such that the price of each asset is equal to its expected payoff under these probabilities, discounted at the riskless interest rate.

4.5 Market Completeness

Notice that if Eqns. (16) and (18) hold for *some* ϕ's and, therefore, for the equivalent π's, then there is no arbitrage irrespective of the uniqueness of the ϕ's and the π's. Without uniqueness, however, Eqns. (16) and (18) cannot be used constructively to value derivatives, or other assets, with arbitrary cash flows. When unique ϕ's and π's exist and the market is 'complete' and any pattern of cash flows across states may be replicated and valued using existing securities.

5. The Black–Scholes Theory

The B–S model is developed in continuous time using a model in which security prices follow a process related to Brownian motion. The continuous-time framework is a critical element of the theory but space unfortunately does not permit us fully to develop this analysis here. For a good general introduction to option pricing (as well as option markets and derivatives in general) see Hull (2000). For a collection of key papers in the development of the field see Constantinides and Malliaris (2001).

We assume that the value of the derivative at time t, $V_t(.)$, is a function of the current asset price, S_t, and time alone and that $V_t(.)$ is twice continuously differentiable in S_t and once t; this will turn out to be correct but a proof lies outside the scope of this article. (This assumption is made in the original Black and Scholes 1973 paper. The first proof that dispensed with this assumption was Merton 1977; also, see Harrison et al. 1979). Thus we assume

$$V(.) = V(S_t, t) \tag{19}$$

In the binomial model described above, the market is complete, which means that it is possible to replicate the payoff on a derivative using just the underlying asset and riskless borrowing and lending. The surprising—and deep—result of the Black–Scholes theory, is that exactly the same holds good in their model, even though it is clear that completeness does not derive, as it does in the binomial case, from the number of available assets being at least as great as the number of states. To understand the intuition behind this remarkable result it is useful to go back to the example discussed in Sect. 4 where we discussed valuing a forward contract via replication. In this example replication was feasible, even though the distribution of the stock price was continuous—and the number of states infinite—because the payoff on forward contract is linear in the price of the underlying asset. Derivatives with payoffs which are nonlinear in the underlying asset cannot, in general, be replicated with a portfolio containing only the underlying asset and borrowing or lending because the payoff on the latter is linear. The size of hedging 'error,' that is, the difference between the value of the derivative and the price of the portfolio, depends on two variables. The first is the degree of nonlinearity in the derivative—the curvature of the relation between the value of the derivative and the underlying asset—and the second is the variability of the underlying asset price. The greater the curvature or the variability, the larger, in general, is the error. In the case of a linear contract, such as the forward contract discussed earlier, the curvature is zero and so, therefore, is the hedging error.

In the Black–Scholes model, the properties of Brownian motion combined with the fact that the composition of the hedge portfolio is revised continuously through time, means that the hedge error is (surprisingly) completely predictable. The hedge portfolio composition can therefore take this effect into account initially and a perfect hedge can be constructed.

This means that, in effect, the price of the derivative and the underlying asset are perfectly correlated over

short intervals and, in this case, a simple relation exists between their risk premia and the volatility of returns:

$$\frac{\mu_S - r}{\sigma_S} = \frac{\mu_V - r}{\sigma_V} \tag{20}$$

where μ_S is the instantaneous or 'local' mean rate of return on the underlying stock, r is the instantaneous riskless interest rate, σ_S is the local volatility of the rate of return on the stock, and μ_V and σ_V are the corresponding quantities for the derivative. Eqn. (20) has a very simply interpretation: it says that if two assets are perfectly correlated, and therefore exposed to exactly the same *type* of risk, the risk premium (the numerator) per unit of risk (the denominator) must be the same for two assets. The common risk in this case is the risk of price fluctuations in the underlying asset.

Using a result known as Îto's lemma, the values of μ_V and σ_V may be calculated from μ_S, σ_S and the derivatives of V with respect to the underlying asset price and time. Substituting these values into Eqn. (20) gives the famous partial differential equation obtained by Black and Scholes

$$\tfrac{1}{2}\sigma_S^2 S^2 V_{SS} + rSV_S + V_t - rV = 0 \tag{21}$$

This equation applies to all derivatives written on the same underlying asset. Solving this equation, subject to boundary conditions specific to the particular contract under analysis, gives the price. For example, the price at time t of a European call option maturing at time T with strike price X is given by

$$V_t = S_t N(d_1) - e^{-r(T-t)} X N(d_2) \tag{22}$$

where:

$$d_1 = \frac{\ln[S_t/(e^{-r(T-t)}X)]}{\sigma_S\sqrt{T-t}} + \tfrac{1}{2}\sigma_S\sqrt{T-t}$$
$$d_2 = d_1 - \sigma_S\sqrt{T-t} \tag{23}$$

and the function $N(.)$ is the cumulative normal distribution.

Just as in the binomial case, the 'risk neutral' probability distribution can also be identified. Under this distribution, instantaneous rates of return on the underlying asset have the same variance as under the natural distribution but, again as before, an expected rate of return equal to the riskless interest rate. The value of a derivative may, therefore, also be written as the expected value of its payoff under this distribution, discounted at the riskless interest rate, exactly as in Eqn. (14) for the binomial case.

6. *Generalizations of Black and Scholes*

While the papers by Black and Scholes (1973) and Merton (1973) provided formulae for the values of some of the simpler derivatives, the methodology they introduced—pricing by replication—has been applied successfully to a vast and still increasing number of derivative instruments (see a recent monograph on the mathematical theory of derivative pricing by Musiela and Rutkowski 1997 that lists over 1,000 references). Although many of the recent applications are much more complicated than the problem of pricing European call and put options considered by Black and Scholes, and Merton, in terms of economic principles, there is little, if anything, that is fundamentally new in what has followed.

Some of the generalizations of the original theory have been developed in response to the increasing complexity of contracts traded in actual markets. One early problem was the valuation of American options which carry the right to exercise at any point up to maturity, as distinct from European options which may be exercised only at maturity. A canonical problem is the valuation of the American put option, for which no closed form solution is available, and this stimulated the application of numerical methods to the valuation of derivatives.

Among the techniques that have been applied are finite difference schemes, a variety of other lattice methods, including the binomial model, and Monte Carlo. Recent treatments of numerical methods applied to derivatives pricing include Rogers and Talay (1997) and Clewlow and Strickland (1998).

A large body of literature has developed devoted to so-called 'exotic options'; contracts that may retain some flavor of simple call and put options but are in other respects much more complicated. An important early contribution is Margrabe (1978) who analyzed options to exchange one risky asset for another. Other examples include 'lookback' options, which are similar to conventional options except that the exercise price is set at the minimum or maximum price of the stock over a given period prior to maturity; 'Asian options,' where the asset price used to compute the payoff is the average over a period prior to maturity and 'Quanto' options where the contract pays a number of units of one currency, for example, US dollars, equal to the price of an asset expressed in another currency (e.g., pounds sterling). Rubinstein (1991) gives an excellent discussion of many exotic options.

A second important area of application is to assets which are not in themselves derivatives but which have characteristics in common with derivatives.

One important example is the valuation of credit risky debt, discussed by Black and Scholes in their original paper. This was also the subject of an early paper by Merton who considered the case of a firm whose liabilities consist of a single zero-coupon loan (see Merton 1974). The relation with derivatives arises because, at maturity, the cash flows received by the lenders in this case may easily be shown to equal those on an equivalent default-free loan minus a put option on the assets of the firm. Black and Cox (1976) generalized Merton's work in a model where the firm

may default prior to the maturity of the debt. Recently, several authors have tackled the same problem in a rather different way and have modeled directly the firm's probability of default. These 'reduced form' models contrast with the traditional approach where the event of default is modeled as the first time the value of the firm's assets reach a given threshold. Lando (1997) provides an excellent survey of recent work on the valuation of credit risky debt.

A third major area of application for the theory has been the term structure of (default-free) interest rates. In some early work, for example, Vasicek (1977), both the dynamics of the short-term interest rate and the functional form of the price of risk are specified exogenously and the BSM no-arbitrage approach is then used to derive the prices of bonds and related derivatives. The same approach, and the same partial differential equation can be used to obtain the prices of interest rate contingent claims more generally, for example, options on bonds and interest rate options as well as bonds themselves. In an important paper, Cox et al. (1985) use a general equilibrium model to derive both the dynamics of the short rate and the price of risk. More recently, many authors have addressed the problem of constructing term structure models that are 'calibrated' to market data such as the term structure of zero coupon rates and/or the prices of particular options. Much of this literature is excellently reviewed in Sundaresan (2000). Ross (2000) contains a collection of many important contributions.

The real options often faced by firms undertaking real investments represents a further important application of the concepts of derivative pricing. Examples of the options in this case include the opportunity to abandon a project (and avoid continuation costs) or to make follow-on investments (see Dixit and Pindyck 1994).

7. *Future Directions*

Measured either by the volume of research it has generated or the extent of its application in practice, the theory of derivative pricing has been an astonishing success and, despite what has already been achieved, the intensity of research activity shows little sign of slackening. Some of this work is devoted to problems in application and much of the work devoted to numerical solutions to derivative pricing problems falls in this category. However, much activity is currently directed towards those aspects of the theory, or methods of application, which are viewed, to a greater or lesser extent, as unsatisfactory.

One example is the theory of the term structure of interest rates where no one formulation has yet been developed which appears to fit all the features of the data. A second concerns the assumptions on the properties of price paths in the standard model and whether it may be important to include the possibility of jumps (see Bates 1996). A third has to do with differences in liquidity between assets and the inclusion of transaction costs and a fourth with the fact that most derivatives pricing theory is inapplicable when the market is incomplete. Finally, there is still much debate about the best way to carry out empirical tests of derivative pricing models.

As mentioned earlier, when Merton wrote his seminal paper on derivative pricing he thought that they were 'unimportant financial securities (Merton 1973). Although a sound judgment at the time it was a poor prediction. The market in derivatives has grown out of all recognition and the development of derivative pricing research has followed a similar path. It would be a brave man who would predict the imminent demise of either.

See also: Asset Pricing: Emerging Markets; Stock Market Predictability

Bibliography

Bates D 1996 Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *Review of Financial Studies* **9**: 69–107

Black F 1989 How we came up with the option formula. *Journal of Portfolio Management* **15**(2): 4–8

Black F, Cox J C 1976 Valuing corporate securities: Some effects of bond indenture provisions. *Journal of Finance* **31**: 351–67

Black F, Scholes M 1973 The pricing of options and corporate liabilities. *Journal of Political Economy* **81**: 637–54

Campbell J 2000 Asset pricing at the millennium. *Journal of Finance* **55**(4): 1515–67

Clewlow L, Strickland C 1998 *Implementing Derivatives Models: Numerical Methods*. Wiley Series in Financial Engineering

Constantinides G M, Malliaris A G 2001 *Options Markets. The International Library of Critical Writings in Financial Economics*. Edward Elgar, Cheltenham, UK and Brookfield, VT

Cox J C, Ingersoll J E, Ross S A 1985 A theory of the term structure of interest rates. *Econometrica* **53**: 385–407

Dixit A, Pindyck R 1994 *Investment under Uncertainty*. Princeton University Press, Princeton, NJ

Duffie D 1996 *Dynamic Asset Pricing Theory*, 2nd edn. Princeton University Press, Princeton, NJ

Duffie D 1998 Black, Merton and Scholes: Their central contribution to economics. *Scandinavian Journal of Economics* **100**(2): 411–24

Gale D 1960 *The Theory of Linear Economic Models*. McGraw-Hill, New York

Group of Thirty 1993 *Derivatives: Practices and Principles*. Group of Thirty, Washington, DC

Hammurabi, Code of (trans. King LW) http://www.fordham.edu/halsall/ancient/hamcode.html#text

Harrison J, Kreps M, Kreps D M 1979 Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* **20**(3): 381–408

Hull J C 2000 *Options Futures and Other Derivatives*, 4th edn. Prentice-Hall, Upper Saddle River, NJ

Ingersoll J 1989 Option pricing theory. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: Finance*. Macmillan, London

Jarrow R A 1999 In Honor of the Nobel Laureates Robert C. Merton and Myron S. Scholes: A partial differential Eqn. that changed the world. *The Journal of Economic Perspectives* **13**(4): 229–48
Lando D 1997 Modeling bonds and derivatives with credit risk. In: Dempster M, Pliska S (eds.) *Mathematics of Derivative Securities*. Cambridge University Press, Cambridge, UK, pp. 363–93
Margrabe W 1978 The value of an option to exchange one asset for another. *Journal of Finance* **33**: 177–86
Merton R C 1973 Theory of rational option pricing. *Bell Journal of Economics* **4**: 141–83
Merton R C 1974 On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* **29**: 449–69
Merton R C 1977 On the pricing of contingent claims and the Modigliani–Miller theorem. *Journal of Financial Economics* **5**(2): 241–9
Musiela M, Rutkowski M 1997 *Martingale Methods in Financial Modeling*. Springer-Verlag, Berlin
Rogers L C G, Talay D (eds.) 1997 *Numerical Methods in Finance* (Publications of the Newton Institute). Cambridge University Press
Ross S A 2000 *The Debt Market, The International Library of Critical Writings in Financial Economics*. Edward Elgar, Cheltenam, UK and Brookfield, VT
Rubinstein M 1991 Exotic options (working paper). University of California
Schaefer S M 1998 Robert Merton, Myron Scholes and the development of derivative pricing. *Scandinavian Journal of Economics* **100**(2): 425–45
Sprenkle C M 1961 Warrant prices as indicators of expectations and preferences. *Yale Economic Essays* **1**: 178–231 (reprinted in Cootner P H (ed.) 1967 *The Random Character of Stock Market Prices*. MIT Press, Cambridge, MA)
Sundaresan S M 2000 Continuous-time methods in finance: A review and an assessment. *Journal of Finance* **55**(4): 1569–1622
Vasicek O 1977 An equilibrium characterization of the term structure. *Journal of Financial Economics* **5**: 177–88

S. M. Schaefer

Asset Pricing: Emerging Markets

Asset pricing theory is a framework designed to identify and measure risk as well as assign rewards for risk bearing. This theory helps us understand why the expected return on a short-term government bond is a lot less than the expected return on a stock. Similarly, it helps us understand why two different stocks have different expected returns. The theory also helps us understand why expected returns change through time. The asset pricing framework usually begins with a number of premises such as: investors like higher rather than lower expected returns, investors dislike risk, and investors hold well-diversified portfolios.

These insights help us assess the 'fair' rate of return for a particular asset. Such information is critical for the investment decision facing both corporations evaluating projects, and investors forming portfolios. In the corporate setting, the theory helps us characterize the risk of a particular project or acquisition, and assign a discount rate that reflects the risk. In choosing projects that have a higher promised rate of return than what would be assumed by the risk theory, corporations create value. In the portfolio investment setting, the theory helps us identify overvalued and undervalued assets. The theory is also integral to establishing a framework to help the investor understand the risk to be faced with a particular portfolio.

The foundation work in asset pricing originated with Nobel Laureate William Sharpe (1964) and the late John Lintner (1965). While there have been many advances in asset pricing since 1964, to understand the issues that we face with asset pricing in emerging markets, it is useful to follow the framework of the first asset pricing theory, the Capital Asset Pricing Model (CAPM) of Sharpe and Lintner. The key to understanding the complexities of emerging market asset pricing lies with the assumptions of the asset pricing theory.

The CAPM suggests that investors hold well-diversified portfolios. Investors like higher expected returns and dislike variance. This is the framework pioneered by Nobel Laureate Harry Markowitz (1959).

There is an important implication of the notion of well-diversified portfolios. The risk of the well-diversified portfolio is its variance—and the risk of a particular asset is not its own variance. The logic works as follows; you care about the variance of the portfolio—not of individual assets. A particular asset might have greater or lower variance than the portfolio. However, it does not make any sense to reward the asset based on its own variance. Correlation is the missing ingredient. It is possible that a very high variance asset can reduce the overall portfolio variance because it has low or negative correlation with the portfolio returns. Indeed, one can think of this high volatility asset with low correlation as providing insurance or hedging for the overall portfolio.

Let's follow this example further. Investor's don't like variance in their portfolio returns. The particular asset with high variance and low correlation is not judged on its own variance. It is judged on how it contributes to the variance of the well-diversified portfolio. In the example, it reduces the variance of the portfolio. As a result, this asset is valuable (investors like variance reducing assets) and the expected return is low as a result. In other words, because investors value the variance reducing properties of this asset in the context of their portfolio, the price is bid up to the point that the future expected returns are low.

So, to be clear here, it is possible that a high volatility asset has a low expected return and it is also possible that the high volatility asset has a high-expected return. It is not the variance of the asset that matters—it is the contribution to the variance of the portfolio. This contribution is the covariance.

The model of Sharpe (1964) and Lintner (1965) formalize this. Expected returns on assets are different

because the assets have different covariances with a well-diversified portfolio. The model also includes a reward for covariance risk. That is, in order to translate the covariance into expected return, we need the price of covariance risk—how this risk is treated in the marketplace.

There are numerous ways to derive the CAPM and we will not go into the different ways in this article. However, some of the most important assumptions are: investors only care about mean and variance; asset returns are multivariate normally distributed (or equivalent assumptions on investor utility could be made to replace this assumption); capital markets are perfect (all information is correctly reflected in prices as in Fama (1970); there are no transaction costs, no taxes, etc.); and there are no disagreements about the returns distributions. All of these assumptions are counterfactual. However, they provide a framework to derive a simple model that has rich implications.

One serious problem in applying this model to international finance is the assumption of perfect capital markets. In an international setting, this assumption also means that markets are perfectly integrated. This means the following: the same risk asset commands the same expected return regardless of location (country). A sufficient condition for this to work is that there are no effective barriers to portfolio investment across borders. That is, local investors are free to add any stock in the world to their portfolio and international investors are free to choose any stock within a particular country.

With capital market integration, we get a world version of the Capital Asset Pricing Model. That is, assets within a particular country are rewarded in terms of their contribution to a well-diversified world portfolio. What matters is the covariance with the world portfolio. There is also a world price of covariance risk that translates the contribution into expected returns. The world price is directly linked to the weighted average risk aversion in the world. Higher risk aversion implies a higher world price of covariance risk.

The world CAPM is a powerful model and has met with some success in being applied to developed market returns (see Harvey 1991). However, the same model fails when applied to emerging market returns (see Harvey 1995). There are many reasons why the model fails in emerging markets—but a leading and logical candidate is the lack of market integration in some emerging markets.

To understand the impact of market integration, consider a completely segmented (non-integrated) country. In this country, local investors are not allowed to own foreign securities. Foreign investors are not allowed to own local securities.

If the CAPM is held in this segmented country, then the relevant risk that investors face is the asset's contribution to variance of a diversified portfolio within the particular country—not the world. The risk that investors face is the variance of the country portfolio.

Let us make the distinction clear. In the integrated world, a country portfolio's risk is its covariance with world returns. This covariance is rewarded with a common world price, which is linked to weighted average risk aversion in the world. In the segmented world, a country portfolio's risk is it variance. The variance is rewarded with a country specific price, which is linked to a weighted average risk aversion within the particular country. These scenarios, integrated/segmented, are polar extremes. Early capital markets research on emerging markets recognized the importance of market integration and realized that it was likely that many markets were not completely integrated into world capital markets but they were not completely segmented either. For example, Errunza and Losq (1985) proposed a model of partial integration. Roughly speaking, one could think of expected returns in a partially segmented emerging market as reflecting some reward for the covariance with world returns as well as some reward for the market's own variance. One gets a hybrid CAPM that includes both variance as well as covariance with the world.

Bekaert and Harvey (1995) critique the usual implementation of the partial integration/segmentation model. The traditional model assumes that the degree of integration/segmentation is fixed over time. However, this flies in the face of substantial liberalizations of equity markets in many emerging countries in the late 1980s. That is, the traditional framework does not handle the dynamics of capital market integration. Bekaert and Harvey (1995) present an alternative framework for the valuation of emerging market assets. This framework explicitly recognizes that the integration process is gradual. Bekaert and Harvey parameterize and estimate a model that allows for a time-varying market integration. In the polar case of market integration, their model reduces to the world capital asset pricing model. In the case of market segmentation, their model reduces to a local CAPM. In the partial integration world, the expected rate of return is a weighting of world covariance times world price of covariance risk and the local volatility and the reward for the local volatility. To make the model dynamic, this weighting changes through time. The model assumes that the weighting is a function of two variables that proxy for the openness of the market: the size of the trade sector and the capitalization of the local equity market.

What happens when an emerging market liberalizes and becomes more integrated into world capital markets? This is a growing area of research that sheds much light on asset pricing in emerging markets. We will first consider what the theory suggests and second we will consider the evidence.

As previously mentioned, in the segmented capital market, variance counts. In addition, the variance of

the country portfolio is high (because it is not a truly diversified portfolio in a world context). To make matters worse, many emerging markets do not have the breadth of industrial sectors that Developed Countries have. That is, the firms come from very few industries. Further, most of the local firms' prospects are tied to the local economy. As a result, the returns of these firms tend to move in the same direction on any given day. This is another reason why variance is so high in segmented markets. Expected returns are also high. The local investors do not want to bear this extreme volatility. However, local corporations need to raise funds for investment projects. In order to get local investors to purchase local equity, the price must be low (expected rewards must be high). A subsidiary, but important point is that local corporations decline to pursue a number of seemingly profitable capital projects because the cost of equity capital is so high. For example, a project might promise a 25 percent rate of return, which is extraordinary in the context of developed markets. However, this project might be rejected because the cost of equity capital is 30 percent.

Now consider the integration process. Suppose regulations are changed such that local investors can purchase stocks outside their country and foreigners are allowed into the local market.

The most important impact will come from foreigners. They will be attracted to the emerging market for two reasons. First, at current values, the prices are cheap and the expected returns are high compared to what could be earned in developed markets. Second, because of the different industrial compositions of emerging markets relative to developed markets, the correlation of these markets' returns with world returns is lower than the correlation of developed markets' returns with world returns. This second point is important. Even though the volatility of the individual emerging market is high, the correlations are low or negative, which implies that the addition of emerging markets to a well-diversified world portfolio could reduce portfolio volatility.

When the foreign investors pour money into the market to take advantage of the low correlation/high expected return opportunities, prices rise and expected returns decrease. So, one immediate implication of capital market liberalization is that expected returns should decrease. Remember that the expected returns started out from quite a high level. There is plenty of room for the expected returns to decrease.

Bekaert and Harvey (2000) and Henry (2000) document this phenomenon. Bekaert and Harvey propose a set of dates that reflect capital market liberalization. They find that the expected returns decrease after liberalizations.

There are powerful implications to the decrease in expected returns. An immediate implication is that the cost of equity capital decreases. For local firms, this means that the investment projects that in the past were rejected because of expensive equity financing—are now profitable. We would expect to see an increase in capital investment as a proportion to GDP with a decrease in the cost of capital. We would also expect to see high GDP growth.

The evidence in Bekeart et al. (2000) provides convincing evidence in favor of these implications. In a cross-country study of the determinants of GDP growth, Bekaert et al. find that liberalization of capital market leads to a 1.5 percent increase in real GDP growth.

There are other potential implications of capital market liberalization. Policy makers, in particular, are often concerned with the volatility of capital market returns. Is it the case that liberalization or the move to more integrated capital markets increases local equity volatility?

In the case of volatility, there are many possible theories. It might be that capital market integration brings new trading volume and many new analysts watching emerging market stocks thereby increasing the informational efficiency. As a result, these stocks react faster to relevant information and we may see a natural increase in volatility. Volatility might also increase as so-called 'hot' foreign portfolio investment is withdrawn from a particular country on the hint of economic/financial/political unrest. Volatility may also increase as local firms specialize their product line to focus on the goods that they have a demonstrable competitive advantage in producing.

There are reasons for volatility to decrease too. Financial integration is often accompanied or preceded by economic integration. As firms trade their goods and services in world markets, they become less susceptible to shocks or economic fluctuations from the local economy. That is, world trade provides a natural economic hedge. Less sensitivity to the local economy can reasonably be translated into lower volatility of equity returns.

The evidence presented in Bekaert and Harvey (1997, 2000) suggests that there is no significant impact on volatility. Market integration sometimes leads to higher volatility and sometimes leads to lower volatility—it depends on the country. This ambiguous result is consistent with the different theories of volatility.

What about correlation with world returns? Historically, many researchers have looked at increases in correlation as evidence of capital market integration. However, this approach is potentially flawed. Two countries may be completely integrated and, at the same time, their equity returns are uncorrelated. The low correlation could simply reflect the different industrial mixes in the two countries.

The argument articulated in the previous paragraph that local companies might find more of their business in world product markets, is a powerful reason why correlations may increase. In an integrated world, the same type of world shocks that companies in developed markets experience will influence local com-

panies. For example, if the US falls into recession, this is bad news for many stocks in Developed Countries because the US is a major consumer of their goods. While segmented, the local economy may have been largely shielded from such fluctuations. However, as both the economic and financial integration processes are initiated, local companies could be very much affected by, for example, a US recession. Correlations should increase as a result.

The evidence in Bekaert and Harvey (2000) suggests that correlations increase after financial liberalizations. Note that one of the important reasons for foreign investors to enter these emerging markets and purchase securities is the low correlation. Importantly, the evidence in Bekaert and Harvey does not suggest that the diversification potential be eliminated. Even the new higher level of correlation provides substantial benefits to international diversification and is well below the threshold level set by developed markets.

For asset pricing theory, it is critical to know when and if an emerging market has effectively liberalized. Notice the use of the word 'effective.' Bekaert and Harvey (1995) argue that a country might be liberalized to the letter of the law but effectively segmented. That is, new laws might be passed that make it easier for foreigners to access the local capital markets—but it is possible that no foreigners bother to enter the market. This might happen if the liberalization was not credible—or if there was a threat of future policy makers reversing the liberalizations.

Conversely, it is possible, by the letter of the law that a country appears completely segmented but it effectively integrated Indeed, the growth of country closed-end funds as well as American Depository Receipts (ADRs) has made it possible for foreign investors to access local securities without directly purchasing the securities on the local market (where they might not be allowed to transact).

So, it is potentially problematic to judge the degree of integration by looking at particular regulations. The analysis in Bekaert and Harvey (2000) considers a number of different dating schemes: regulatory, introduction of first ADR, introduction of country fund, and a composite indicator of the first sign of liberalization (first date of regulatory, ADR, and country fund). Bekaert and Harvey also examine US portfolio flows. Indeed, if the market were truly open, then supporting evidence of integration would be a dramatic pickup in foreign portfolio flows to the emerging market.

The idea of portfolio flows as a proxy for market integration is analyzed in Bekaert et al. (2000a). Using the econometrics of break point analysis, they examine the interrelationship between capital flows and expected returns in emerging markets.

Indeed, it is possible that the market integration process fundamentally impacts a number of financial and economic variables. Bekaert et al. (2000b) presents the first multivariate tests that try to date the integration of world capital markets. The idea of this paper is to let the data speak for itself. They simultaneously examine a number of important aggregates and use endogenous break point analysis to reveal the common break point. They link this break point to capital market liberalization. For example, we expect to see a decrease in the cost of capital and an increase in foreign portfolio flows. The technology in this paper uses the information in many different economic aggregate series to come up with a composite break point.

The analysis in Bekaert and Harvey (2000) as well as Bekaert et al. (2000a, 2000b) suggests that many emerging markets have successfully liberalized their capital markets. There was a clustering of liberalizations in the late 1980s and early 1990s. This suggests that it is more likely that some version of the world capital asset pricing model holds more today than 10 years ago in 1990 for emerging markets. This is consistent with the results presented in Harvey (2000) who studies the ability of a world version of the CAPM to explain expected returns in 20 developed markets and 27 emerging markets.

There is one additional important issue. Traditional asset pricing theory operates in a world of mean and variance. The assumption of multivariate normality is often invoked. However, the evidence in Bekaert and Harvey (1997) strongly suggests that emerging market returns are highly non-normal. There is also evidence that the returns in many Developed Countries are non-normal. In addition, there is considerable discussion of the 'downside risks' of investing in emerging markets. That is, emerging markets are known to suffer from sharp crises. Two recent memories are the Mexican crisis of December 1994 and the Asian crisis of July 1997.

The analysis of downside can be handled within the traditional asset pricing frameworks. For example, Rubinstein (1973) and Kraus and Litzenberger (1976) developed an asset-pricing framework that explicitly considers the skewness of portfolio returns. Harvey and Siddique (2000) present tests of an asset-pricing model with a dynamic measure of skewness.

The logic of these models is much like that of the CAPM—indeed, it is a straightforward extension of the CAPM. Investors like expected returns, dislike variance, and dislike negative skewness (but like positive skewness). Investors' preference for positive skewness is evidenced by the unusually high price that they will pay for a lottery ticket (which has an expected return of –50 percent). We also see evidence in the options markets where 'put' options on the S and P indexes have very high prices. Investors are using these 'put' options to protect against extreme downside moves in the equity market. These options reduce negative skewness in an investor's portfolio. As a result, they are valuable (high price).

Like variance, it is not the skewness of the particular asset that matters—it is the contribution of that asset to the portfolio's skewness. This is called the coskew-

ness. If the coskewness is negative, it means that the asset is contributing negative skewness to the portfolio. In order to get people to purchase an asset with such an undesirable property, the price must be low (expected returns high). If the asset offers to increase the portfolio skewness (which is desirable), the price will be bid upwards (expected returns will be low).

Harvey and Siddique (2000) present tests of an asset-pricing model that incorporates skewness. They find some success in explaining the cross-section of US equity returns with this framework.

Harvey (2000) looks at a model with skewness for 47 international stock markets. His analysis is broken into three groups: Developed Countries, emerging markets, and all countries. He only considers data from 1988 which marks the beginning of an intense period of capital market reforms in emerging markets.

If markets were completely segmented, then what counts is the country's variance and total skewness. If markets are completely integrated, what counts is the covariance and coskewness. Harvey presents evidence that developed markets are not impacted by variance and total skewness, which is consistent with them being integrated. Evidence is also presented that covariance and coskewness also do a reasonable job in emerging markets—suggesting that many of these markets have successfully integrated into world capital markets. However, his evidence suggests that these emerging markets be not completely integrated. In some of his tests, both variance and total skewness provide an incremental ability to explain the cross-section of expected returns in emerging markets. The evidence suggests that it is unwise to assume that emerging markets are fully integrated into world capital markets.

There are a host of additional issues that we face in investing in emerging markets. Transaction costs (execution fees, bid-ask spread, and market impact) are higher than developed markets. Acute information asymmetries can exist (local investors may have better information than foreign investors may). Regulations, such as insider trading laws, may not exist or even if they exist, are not enforced. Many securities suffer from chronic infrequent trading making the price data unreliable. There is also the general issue as to how to handle foreign currency risk within asset pricing theory (see Dumas and Solnik 1995 as well as the early work of Solnik 1974 and Stulz 1981).

Each of these additional issues increases the probability that standard asset pricing models will fail when applied to emerging markets.

Finally, there is one overriding issue that impacts the study of asset pricing in emerging markets—as well as a more general study of asset pricing. It is not clear that the standard models are adequate to capture the complexities of security valuation in developed markets, let alone in emerging markets. Recently, in a number of studies of US equity returns, the traditional CAPM has come under attack. Indeed, most view the US market as one of the most efficient markets in the world, which would maximize the chance that the CAPM would work.

However, the international evidence is more generous to the traditional framework. It is paramount to explicitly allow for the role of capital market integration. After doing so, the traditional frameworks, perhaps augmented with measures of coskewness, are able to do a reasonable job in explaining the cross-section of expected returns.

Bibliography

Bekaert G 1995 Market integration and investment barriers in emerging equity markets. *World Bank Economic Review* **9**: 75–107

Bekaert G, Harvey C R 1995 Time-varying world market integration. *Journal of Finance* **50**: 403–44

Bekaert G, Harvey C R 1997 Emerging equity market volatility. *Journal of Financial Economics* **43**: 29–78

Bekaert G, Harvey C R 2000 Foreign speculators and emerging equity markets. *Journal of Finance* **55**: 565–614

Bekaert G, Harvey C R, Lunblad C 2000 Emerging equity markets and economic growth, Unpublished manuscript, Columbia University and Duke University

Bekaert G, Harvey C R, Lumsdaine R 2000a The dynamics of emerging market equity flows, Unpublished manuscript, Columbia University, Duke University and Brown University

Bekaert G, Harvey C R, Lumsdaine R 2000b Dating the integration of world capital markets, Unpublished manuscript, Columbia University, Duke University and Brown University

Black F, Jensen M, Scholes M 1972 The capital asset pricing model: some empirical tests. In: Jensen M (ed.) *Studies in the Theory of Capital Markets*. Praeger, New York

Dumas B, Solnik B 1995 The world price of foreign exchange rate risk. *Journal of Finance* **50**: 445–80

Errunza V R, Losq E 1985 International asset pricing under mild segmentation: Theory and test. *Journal of Finance* **40**: 105–24

Fama E F 1970 Efficient capital markets: a review of theory and empirical work. *Journal of Finance* **25**: 383–417

Harvey C R 1991 The world price of covariance risk. *Journal of Finance* **46**: 111–57

Harvey C R 1995 Predictable risk and returns in emerging markets. *Review of Financial Studies* **8**: 773–816

Harvey C R 2000 The drivers of expected returns in international markets. *Emerging Markets Quarterly* forthcoming

Harvey C R, Siddique A 2000 Conditional skewness in asset pricing tests. *Journal of Finance* **55**: June 2000, 1263–95

Henry P B 2000 Stock market liberalization, economic reform, and emerging market equity prices. *Journal of Finance* **55**: 529–64

Kraus A, Litzenberger R 1976 Skewness preference and the valuation of risk assets. *Journal of Finance* **31**: 1085–100

Lintner J 1965 The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* **47**: 13–37

Markowitz H 1959 *Portfolio Selection: Efficient Diversification of Investments*. John Wiley, New York

Rubinstein M The fundamental theory of parameter-preference security valuation. *Journal of Financial and Quantitative Analysis* **8**: 61–9

Sharpe W 1964 Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* **19**: 425–42
Solnik B 1974 An equilibrium model of the international capital market. *Journal of Economic Theory* **8**: 500–24
Stulz R 1981 A model of international asset pricing. *Journal of Financial Economics* **9**: 383–406

C. R. Harvey

Assimilation of Immigrants

Immigration is a transformative force, producing profound and unanticipated social changes in both sending and receiving societies, in intergroup relations within receiving societies, and among the immigrants themselves and their descendants. Immigration begets ethnicity—collectivities that perceive themselves and are perceived by others to differ in language, religion, race, national origin or ancestral homeland, cultural heritage, and memories of a shared historical past. Their modes of incorporation across generations may take a variety of forms—some leading to greater homogenization and cultural solidarity within the society (or within segments of the society), others to greater ethnic differentiation and heterogeneity. Assimilation is a multidimensional process of boundary reduction that blurs an ethnic or racial distinction and the social and cultural differences and identities associated with it. At its end point, formerly distinguishable ethnocultural groups become effectively blended into one. At the group level, assimilation may involve the absorption of one or more minority groups into the majority, or the merging of minority groups. At the individual level, assimilation denotes the cumulative changes that make individuals of one ethnic group more acculturated, integrated, and identified with the members of another. Ideologically, the term has been used to justify selective state-imposed policies aimed at the eradication of minority cultures. But in the social scientific study of immigration and intergroup relations, it remains an indispensable concept (Alba and Nee 1997, Yinger 1981).

1. Evolution of the Concept of Social Assimilation in American Social Science

A May 15, 1880, editorial in *The New York Times* reflected the popular usage of the concept on the eve of a new era of mass immigration:

> There is a limit to our powers of assimilation and when it is exceeded the country suffers from something very like indigestion … We know how stubbornly conservative of his dirt and ignorance is the average immigrant who settles in New York … these wretched beings change their abode, but not their habits in coming to New York.

But in Chicago a generation later, with immigration unabated and the large majority of the city's residents already consisting of first- and second-generation immigrants, Robert Park, the leading sociologist on the subject, could write (in Park and Burgess 1924, pp. 757–8):

> In America it has become proverbial that a Pole, Lithuanian, or Norwegian cannot be distinguished, in the second generation, from an American born of native parents … As a matter of fact, the ease and rapidity with which aliens, under existing conditions in the United States, have been able to assimilate themselves to the customs and manners of American life have enabled this country to swallow and digest every sort of normal human difference, except the purely external ones, like color of the skin.

What had become proverbial in America would become canonical in American sociology.

Ironically, Park is best known for the formulaic notion that assimilation is the final stage of a natural, progressive, inevitable and irreversible 'race relations cycle.' But in a prolific career, Park wrote about a 'race relations cycle' only twice: first in a sentence near the end of a 1926 article, 'Our racial frontier in the Pacific;' then a decade later (Park 1937) in a brief introduction to a book on interracial marriage in Hawaii written by one of his former students. In the first instance he was arguing against the likelihood that a 'racial barrier'—which the passage of exclusionary laws sought to establish by barring Asian migration to the USA—could be much of a match against global economic, political, and cultural forces that have brought about 'an existing interpenetration of peoples … so vast and irresistible that the resulting changes assume the character of a cosmic process' (Park 1926, p. 141). And in his 1937 introduction, he explicitly rebutted any notion of a predictable assimilative outcome to race conflict and change ('what are popularly referred to as race relations'), arguing instead that when stabilization is finally achieved, race relations would assume one of three configurations (Park 1937, p. xii):

> They will take the form of a caste system, as in India; they will terminate in complete assimilation, as in China; or the unassimilated race will constitute a permanent racial minority within the limits of a national state, as in the case of the Jews in Europe … All three types of change are involved … in what we may describe as the 'race relations cycle.

Park and Burgess gave the concept of assimilation its classic formulation: 'a process of interpenetration and fusion in which persons and groups acquire the memories, sentiments, and attitudes of other persons and groups, and, by sharing their experience and

history, are incorporated with them in a common cultural life' (Park and Burgess 1924, p. 735). They distinguished systematically between 'four great types of interaction'—competition, conflict, accommodation, and assimilation—which they related respectively to economic, political, social, and cultural institutions. *Competition* and *conflict* sharpen ethnic boundaries and the consciousness of intergroup difference. An *accommodation* (of a conflict, or to a new situation) may take place quickly, and the person or group is typically a highly conscious protagonist of the process of accommodating those circumstances. In *assimilation*, by contrast, the changes are more subtle and the process is typically unconscious, so that the person is incorporated into the common life of the group largely unaware of how it happened. Assimilation is unlikely to occur among immigrants who arrive as adults. Instead, accommodation most closely reflects the modal adaptation of first-generation adult immigrants, while assimilation can become a modal outcome ultimately only for the malleable young and for the second generation, and then only if and when permitted by structural conditions of inclusion at the primary group level. Indeed, the research literature on the adaptation of twentieth-century European immigrant groups in the USA suggests that evidence of assimilation was not manifestly observed at the group level until the third or even fourth generation.

Assimilation takes place most rapidly and completely in primary—intimate and intense—social contacts, including intermarriage; accommodation may be facilitated through secondary contacts, but they are too distant and remote to promote assimilation. Because the nature (especially the interpersonal intimacy, 'the great moral solvent') of the social contacts is what is decisive, it follows that 'a common language is indispensable for the most intimate associations of the members of the group,' and its absence is 'an insurmountable barrier to assimilation,' since it is through communication that gradual and unconscious changes of the attitudes and sentiments of the members of the group are produced. But language and acculturation alone cannot ensure assimilation if a group is categorically segregated, racially classified, and 'regarded as in some sense a stranger, a representative of an alien race'—which is why the English-speaking Protestant 'Negro, during his three hundred years in this country, has not been assimilated … not because he has preserved in America a foreign culture and an alien tradition, for with the exception of the Indian … no man in America is so entirely native to the soil' (Park 1930, p. 282). Race and place (or more precisely, racial discrimination and residential segregation) become critical structural determinants of the degree of assimilation precisely insofar as they delimit possible forms of primary social contact; social relations are inevitably embedded and bounded in space, which is why social distance is typically indexed by physical distance.

In *The Social Systems of American Ethnic Groups*, perhaps the definitive statement on the subject near mid-century, Warner and Srole (1945) described the progressive advance of eight European-origin immigrant groups in the major status hierarchies of 'Yankee City' in Massachusetts, explicitly linking upward social mobility to assimilation, which they saw as determined largely by the degree of ethnocultural (religion and language) and above all racial difference from the dominant group. While racial groups were subordinated and excluded through caste restrictions on residential, occupational, associational, and marital choice, the clash of ethnic groups with the dominant institutions of the host society was not much of a contest, particularly among the young. The industrial economy, the polity, the public school, popular culture, and the American family system all undercut and absorbed ethnicity in various ways, so that even when 'the ethnic parent tries to orient the child to an ethnic past … the child often insists on being more American than Americans' (Warner and Srole 1945, p. 284). And for the upwardly mobile, with socioeconomic success came intermarriage and the further dilution of ethnicity. Or at least, as qualified by a contemporary analysis (Kennedy 1944), the assimilation process was segmented in a 'triple melting pot,' with ethnic intermarriage and blending occurring within Protestant, Catholic, and Jewish religious groupings and not across religious lines.

That general view of assimilation as linear progress, with sociocultural similarity and socioeconomic success marching in lock-step, was not so much challenged as refined by Milton Gordon in *Assimilation in American Life* (1964), published ironically on the eve of the beginning of the latest era of mass immigration to the USA—and of the denouement of the concept itself in the wake of the 1960s. He broke down the assimilation sequence into seven stages, of which 'identificational assimilation'—a self-image as an unhyphenated American—was the end point of a process that began with cultural assimilation, proceeded through structural assimilation and intermarriage, and was accompanied by an *absence of prejudice and discrimination* in the 'core society.' Once structural assimilation had occurred (that is, extensive primary-level interaction with members of the 'core group'), either in tandem with or subsequent to acculturation, 'the remaining types of assimilation have all taken place like a row of tenpins bowled over in rapid succession by a well-placed strike' (Gordon 1964, p. 81). For the children of white European immigrants, in fact, the acculturation process was so 'overwhelmingly triumphant' that the greater risk consisted in alienation from family ties and in role reversals of the generations that could subvert parent–child relationships. Still, what it was that one was assimilating to remained largely taken for granted.

Gordon was aware of the ways in which the ideal and the ideological get wrapped up in the idea of

assimilation, and saw 'Anglo-conformity' as the most prevalent ideology of assimilation in American history, but he did not focus on the historical contexts that have shaped the ideas and ideals embodied in the notion of assimilation. Kazal (1995) sees the apogee of the concept in the 1950s and early 1960s as reflecting the need generated by World War II for national unity and the postwar tendency to see American history as a narrative of consensus rather than conflict; and the political and social upheavals of the 1960s as shattering the 'consensus school' and the rationale for studying assimilation, bringing back instead a focus on the ethnic group and ethnic resilience, and more inclusive conceptions of American society. 'To know how immigrants came to fit in, one had to understand what it was they were fitting into … When the notion of an Anglo-American core collapsed amid the turmoil of the 1960s, assimilation lost its allure' (Kazal 1995, p. 437). It took until the 1990s, once more in a new era of mass immigration, for a systematic reevaluation of the concept of assimilation to emerge, and its application in contemporary scholarship seeking to contrast differences and similarities between the old and the new immigration.

2. *Dimensions and Determinants of Assimilation*

Assimilation, as shown by this brief overview of the evolution of the concept, involves a series of interrelated cultural (acculturation), structural (integration), and psychological (identification) processes. These may be elaborated further, with some additional consideration of contextual and group factors shaping each of these dimensions—either by promoting or precluding assimilative outcomes (Cornell and Hartmann 1998, Yinger 1981).

Acculturation, involving complex processes of cultural diffusion and changes producing greater cultural similarity between two or more groups, is generally more extensive among members of smaller and weaker groups, and particularly immigrant groups. Nonetheless, acculturation is never exclusively one-sided; dominant groups, too, are culturally influenced by their contacts with other ethnocultural groups in the society, from cuisine and music to language and religion. In the American experience, language shifts have been overwhelmingly one-sided, with the switch to monolingual English typically being accomplished by the third generation. At the individual level, a key distinction to make is that between *subtractive* (or *substitutive*) acculturation and *additive* acculturation. The first is essentially a zero-sum game that involves giving up some elements of a cultural repertoire (such as language) while replacing them from another; the second does not involve losing but gaining to form and sustain a more complex repertoire (bilingualism and biculturalism). Available research has yet to examine systematically the multiplicity of conditions and contexts yielding subtractive vs. additive acculturative outcomes.

The degree of acculturation, as noted previously, is by itself not a sufficient condition for assimilation. Structural integration was the crux of the matter for Gordon (1964), although he focused on the entrance of the minority group into 'the social cliques, clubs, and institutions of the core society at the primary group level.' Given the many different institutions involved here—and the fact that integration in the economy, the polity, and the community at the secondary group level tends to be ignored by that formulation—a conceptual distinction can be made between secondary structural assimilation and primary structural assimilation. The former refers to a wide range of key integrative processes, including socioeconomic and spatial (residential) assimilation, and the acquisition of legal citizenship. The latter—extensive interaction within personal networks, and intermarriage—is unlikely to take place under conditions of status inequality. While it is clear that all of these dimensions are interdependent—acculturation, integration, intermarriage—the linkages between them are historically contingent and will vary depending on a number of factors, particularly the context of reception within which different immigrant groups are incorporated (Portes and Rumbaut 1996).

Conventional accounts of ethnic identity shifts among the descendants of European immigrants, conceived as part of a linear process of assimilation, have pointed to the 'thinning' of their ethnic self-identities in the USA. For their descendants, at least, one outcome of widespread acculturation, social mobility, and intermarriage with the native population is that ethnic identity became an optional, leisure-time form of 'symbolic' ethnicity (Gans 1979, Waters 1990). As the boundaries of those identities become fuzzier and less salient, less relevant to everyday social life, the sense of belonging and connection to an ancestral past faded. This mode of ethnic identity formation, however, was never solely a simple linear function of socioeconomic status and the degree of acculturation—that is, of the development of linguistic and other cultural similarities with the dominant group—but hinged also on the context of reception and the degree of discrimination experienced by the subordinate group. Identity shifts, like acculturative changes, tend to be from lower to higher status groups, *mutatis mutandis*. But where social mobility is blocked and hindered by prejudice and discrimination, members of lower status groups may react by reaffirming their shared identity. This process of forging a *reactive ethnicity* in the face of perceived threats, persecution, discrimination, and exclusion is not uncommon. On the contrary, it is another mode of ethnic identity formation, accounting for the 'thickening' rather than the dilution of ethnicity.

However, compared to language loyalty and language shift, generational shifts in ethnic self-

identification are far more conflictual and complex. Paradoxically, despite the rapid acculturation of European immigrants in the USA, as reflected in the abandonment of the parental language and other ethnic patterns of behavior, the second generation remained *more* conscious of their ethnic identity than were their immigrant parents. The parents' ethnic identity was so much taken for granted that they were scarcely explicitly aware of it, but the marginality of their children made them acutely self-conscious and sensitive to their ethnicity, especially when passing through adolescence. Moreover, as parents and children acculturated at different rates, a generational gap grew so that by the time the children reached adolescence 'the immigrant family had become transformed into two linguistic subgroups segregated along generational lines.' Finally, by the third generation 'the grandsons became literally outsiders to their ancestral heritage,' and their ethnic past an object of symbolic curiosity more than anything else (Nahirny and Fishman 1965).

By the end of the twentieth century a new era of mass immigration—and hence of ethnogenesis—now overwhelmingly non-European in composition, again raised familiar doubts about the assimilability of the newcomers and alarms that they might become consigned to a vast multiethnic underclass, on the other side of a new color line. It also raised questions about the applicability of explanatory models developed in connection with the experience of European ethnics, despite the fact that contemporary immigrants are being incorporated in a more complex, post-civil-rights context characterized more by ethnic revivals and identity politics than by forced Americanization campaigns. Relative to the first generation, the process of ethnic self-identification of children of immigrants is more complex, and often entails the juggling of competing allegiances and attachments. Situated within two cultural worlds, they must define themselves in relation to multiple reference groups (sometimes in two countries and in two languages) and to the classifications into which they are placed by their native peers, schools, the ethnic community, and the larger society.

While assimilation may still represent the master process in the study of today's immigrants, it is a process subject to too many contingencies and affected by too many variables to render the image of a relatively uniform and straightforward path convincing. Instead, the present second generation of children of immigrants is better defined as undergoing a process of 'segmented assimilation' where outcomes vary across immigrant minorities, and where rapid integration and acceptance into the American mainstream represent just one possible alternative. Why this is so—and how it is that different groups may come to assimilate to different sectors of American society—is a complex story that hinges on a number of factors, among which the following can be considered decisive: the history of the immigrant first generation, including the human capital brought by immigrant parents and the context of their reception; the differential pace of acculturation among parents and children, including the development of language gaps between them, and its bearing on normative integration and family cohesiveness; the cultural and economic barriers confronted by second-generation youth in their quest for successful adaptation; and the family and community resources for confronting these.

These factors combine in ways that magnify the relative social advantages or disadvantages of particular immigrant groups as they make their way in America, and mold both the character of parent–child relations within immigrant families and the adaptive experiences and trajectories of the children. Internal characteristics, including the structure and cohesiveness of their families, interact in complex but patterned ways with external contexts of reception—government policies and programs, the state of the economy in the areas where they settle, employer preferences in local labor markets, the extent of racial discrimination and nativist hostility, the strength of existing ethnic communities—to form the conditions within which immigrant children adapt, react, and assimilate to different segments of American society.

Bibliography

Alba R, Nee V 1997 Rethinking assimilation theory for a new era of immigration. *International Migration Review* **31**(4): 826–74

Cornell S E, Hartmann D 1998 *Ethnicity and Race: Making Identities in a Changing World*. Pine Forge Press, Thousand Oaks, CA

Gans H J 1979 Symbolic ethnicity: the future of ethnic groups and cultures in America. *Ethnic and Racial Studies* **2**(1): 1–20

Gordon M M 1964 *Assimilation in American Life: The Role of Race, Religion, and National Origins*. Oxford University Press, New York

Kazal R A 1995 Revisiting assimilation: the rise, fall, and reappraisal of a concept in American ethnic history. *American Historical Review* **100**(2): 437–71

Kennedy R J R 1944 Single or triple melting-pot? Intermarriage trends in New Haven, 1870–1940. *American Journal of Sociology* **49**: 331–9

Nahirny V C, Fishman J A 1965 American immigrant groups: Ethnic identification and the problem of generations. *Sociological Review*, NS **13**: 311–26

Park R E 1926 Our racial frontier in the Pacific. *Survey Graphic* IX (May): 192–6 [Reprinted in Park's collected papers, *Race and Culture* 1950, The Free Press, Glencoe, IL, Vol. 1, pp. 138–51]

Park R E 1930 Assimilation, social. In: Seligman E R A, Johnson A (eds.) *Encyclopedia of the Social Sciences*, Vol. 2. Macmillan, New York

Park R E 1937 Introduction. In: Adams R C (ed.) *Interracial Marriage in Hawaii; A study of the mutually conditioned processes of acculturation and amalgamation*. Macmillan, New York, pp. vii–xiv

Park R E, Burgess E W [1921] 1924 *Introduction to the Science of Sociology*. University of Chicago Press, Chicago
Portes A, Rumbaut R G 1996 *Immigrant America: A Portrait*. University of California Press, Berkeley, CA
Warner W L, Srole L 1945 *The Social Systems of American Ethnic Groups*. Yale University Press, New Haven, CT
Waters M C 1990 *Ethnic Options: Choosing Identities in America*. University of California Press, Berkeley, CA
Yinger J M 1981 Toward a theory of assimilation and dissimilation. *Ethnic and Racial Studies* **4**(3): 249–64

R. G. Rumbaut

Associative Modifications of Individual Neurons

What is known about the learning of associations at the cellular level? Can associative learning be attributed to the changes in the membrane properties of a single cell, or a change in the efficacy of a specific synapse? The answers to these questions are beginning to be understood for the two major forms of associative learning: operant conditioning and classical conditioning.

1. *Operant Conditioning*

Operant conditioning is a process by which an animal learns to associate consequences to its own behavior. In an operant conditioning paradigm, the delivery of a reinforcing stimulus is contingent upon the expression of a designated behavior. The probability of expression of this behavior will then be altered. In order to understand the mechanisms underlying conditioning, it is often helpful to develop cellular analogues of the behavior. These cellular analogues attempt to capture the essence of the behavior, while also being amenable to experimental manipulations to uncover the processes mediating learning.

It is commonly believed that reinforcement is mediated by certain endogenous neurotransmitters. Self-administration experiments support this view. First, consider drug self-administration, where an animal will consistently make responses that lead to drug injection into the brain. Animals will ardently self-administer only a few substances. One of these substances is dopamine. Next, consider electrical self-stimulation experiments. These experiments demonstrate that the electrical stimulation of certain brain areas can reinforce the response leading to the stimulation. Often the brain areas that give positive reinforcement are coincident with areas rich in dopaminergic tracts or cell bodies and dopamine levels increase after self-stimulation of these areas (a). These results suggest that dopamine can act to reinforce a behavioral response. The reinforcing role of dopamine has been further characterized in rat hippocampal slices (b). Since the CA1 region is known to be a target for endogenous dopaminergic tracts, it was hypothesized that the electrical bursting activity in a pyramidal cell might be strengthened when there was a contingent application of dopamine. Indeed, when contingent reinforcement was applied, the bursting was strengthened, whereas a noncontingent application of dopamine produced no burst strengthening. This preparation demonstrated successful conditioning, indicating that dopamine was capable of mediating the reinforcement.

Feeding behavior in *Aplysia* has been used to gain insights into the modification of a behavioral response by reinforcement. This system is highly tractable. *Aplysia* have a relatively simple nervous system and the neurons are large and accessible, making it possible to consistently record from a single identified neuron. These advantages have led to identification of much of the cellular circuitry controlling feeding behavior. Thus, it is possible to monitor and experimentally manipulate neurons with known behavioral significance. One such neuron is denoted B51. This cell is implicated in the expression of ingestive behavior.

Aplysia ingest food by protracting a toothed grasping structure called the radula. The radula contacts seaweed and closes, grasping the food. The radula then retracts, completing the cycle. The rejection of food (known as egestion) starts by protracting the radula, which is grasping the soon to be rejected food. The radula then opens while it retracts, which releases the food. The timing of radula closure determines whether the behavior is ingestive or egestive. When radula closure occurs only during the protraction phase, the behavior is egestive. However, if radula closure occurs primarily during the retraction phase, then the behavior is ingestive. The cellular network responsible for generating feeding behavior has been largely identified. Both extracellular recordings from nerves and intracellular recordings from individual neurons have shown patterns of activity that correspond to feeding behavior (e.g., protraction, retraction, and closure). Neuron B51 is active only during the retraction phase. Furthermore, when B51 is recruited into a pattern, it both lengthens the retraction phase and recruits radula closure motor neurons (see Fig. 1). Thus, B51 recruitment contributes importantly to the expression of ingestive behavior.

An *in vitro* analogue of operant conditioning has been developed utilizing the ganglion responsible for generating feeding motor patterns (c). The ganglion expresses both ingestion-like and rejection-like patterns. The type of pattern expressed is not predictable. Ingestion-like patterns were selectively reinforced with a contingent shock to the esophageal nerve, which contains dopaminergic tracts. Following conditioning, ingestion-like patterns were more likely to be expres-

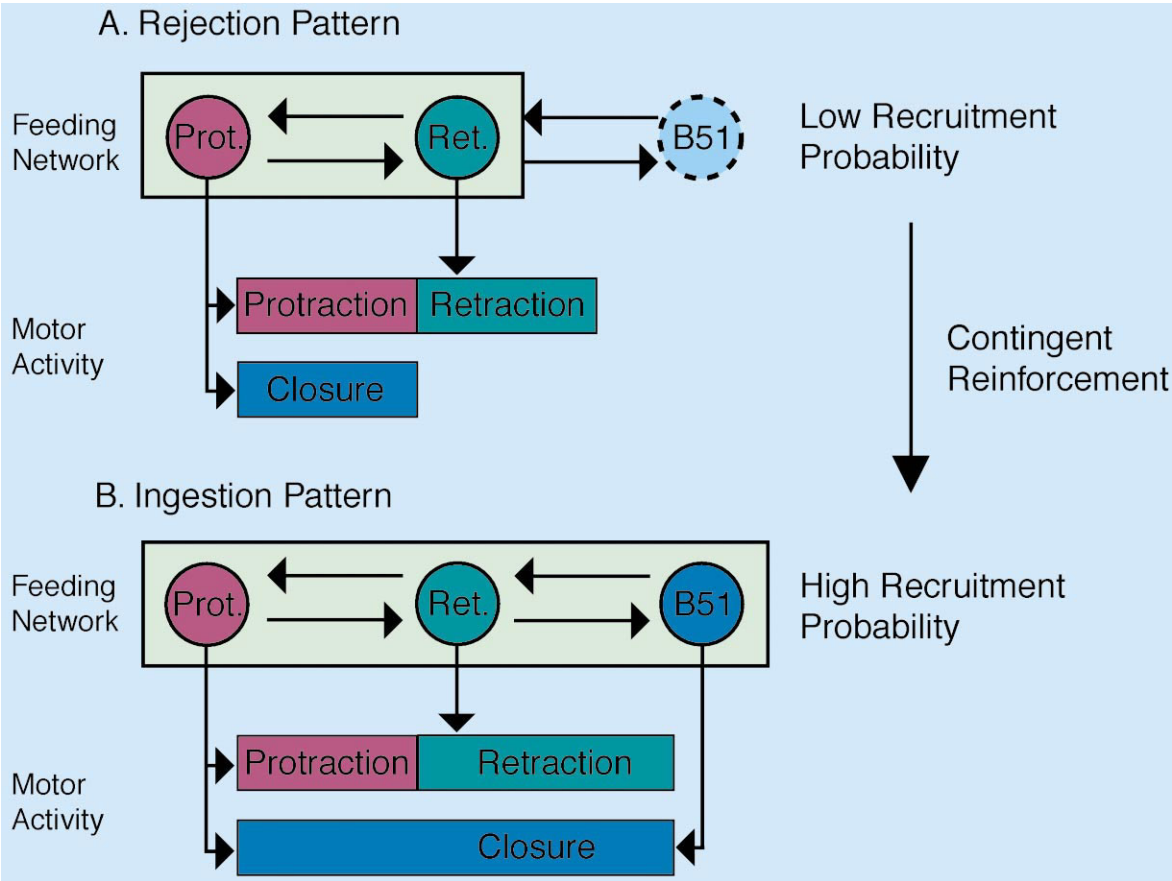

Figure 1
Model of operant conditioning of feeding in *Aplysia*. The circles represent the cellular network that mediates feeding motor programs. The rectangles below represent motor activity comprising a feeding pattern. (A) The radula protraction generating neuron (Prot.) is initially active followed by the radula retraction generating neuron (Ret.). In this naïve state, neuron B51 needs a relatively large magnitude stimulus to become active. Consequently, B51 is not recruited into the feeding motor pattern. Radula closure occurs only during the protraction phase, so this pattern is rejection-like. (B) Following contingent reinforcement, B51 now has a lower threshold for recruitment. Thus, B51 is recruited into the motor pattern. B51's activity in the pattern leads to radula closure occurring primarily during the retraction phase, so this pattern is ingestion-like

sed compared to noncontingent controls. Thus, the conditioning was successful. Contingent reinforcement also modified the membrane properties of neuron B51. The magnitude of the stimulus required to elicit electrical activity in the cell and the resting conductance were both decreased. Both of these changes in the membrane properties acted in the same direction and made the cell more likely to be active in the future. In another set of experiments utilizing the isolated ganglion, induced electrical activity in B51 was used as the analogue of ingestion-like behavior (d). Shocks to the esophageal nerve remained as the analogue of reinforcement. After conditioning, both the resting conductance and the amount of stimulus needed to elicit activity were decreased similarly to the previous experiment. Again, cell B51 was modified such that it would be more likely to be recruited into the motor pattern. This change in the recruitment probability can account for the shift in the likelihood of future ingestion-like patterns. Furthermore, the conditioning of B51 was blocked by the dopamine antagonist methylergonivine. Dopamine appears to act as an agent of reinforcement in invertebrates also.

The analogue has been further reduced to the level of the single cell (e). B51 was removed from the ganglia and placed in culture. Induced electrical activity in this isolated B51 was the analogue of behavior and a direct, temporally discrete, application of dopamine

was the analogue of reinforcement. After conditioning, B51 was more likely to be active in the future. Thus, it appears that the conditioned expression of a behavior can be attributed to associative modifications made at the level of the single cell.

2. *Classical Conditioning*

The mechanisms underlying classical conditioning are also being revealed at the cellular level. Classical conditioning occurs when an animal learns to associate a typically neutral stimulus with a later salient event. The animal learns that the stimulus can serve as a predictor. For example, a conditioned stimulus (CS) is paired with an unconditioned stimulus (US). Typically, the US produces an unconditioned response (UR). After successful paired training, the CS can invoke a conditioned response (CR) in anticipation of the US.

In *Aplysia*, analyses of simple withdrawal reflexes have provided the substrate for mechanistic studies of classical conditioning (f, g, h, i). A diagram of the general scheme is presented in Fig. 2. The US pathway is activated by a shock to the animal, which elicits a withdrawal response (the UR). When a CS is consistently paired with the US, a CR of withdrawal will develop. Activity-dependent neuromodulation is proposed as the mechanism for this pairing specific effect. The US activates both a motor neuron (UR) and a modulatory system. The modulatory system delivers the neurotransmitter serotonin (5-HT) to all the sensory neurons (parts of the various CS pathways). Sensory neurons whose activity is temporally con-

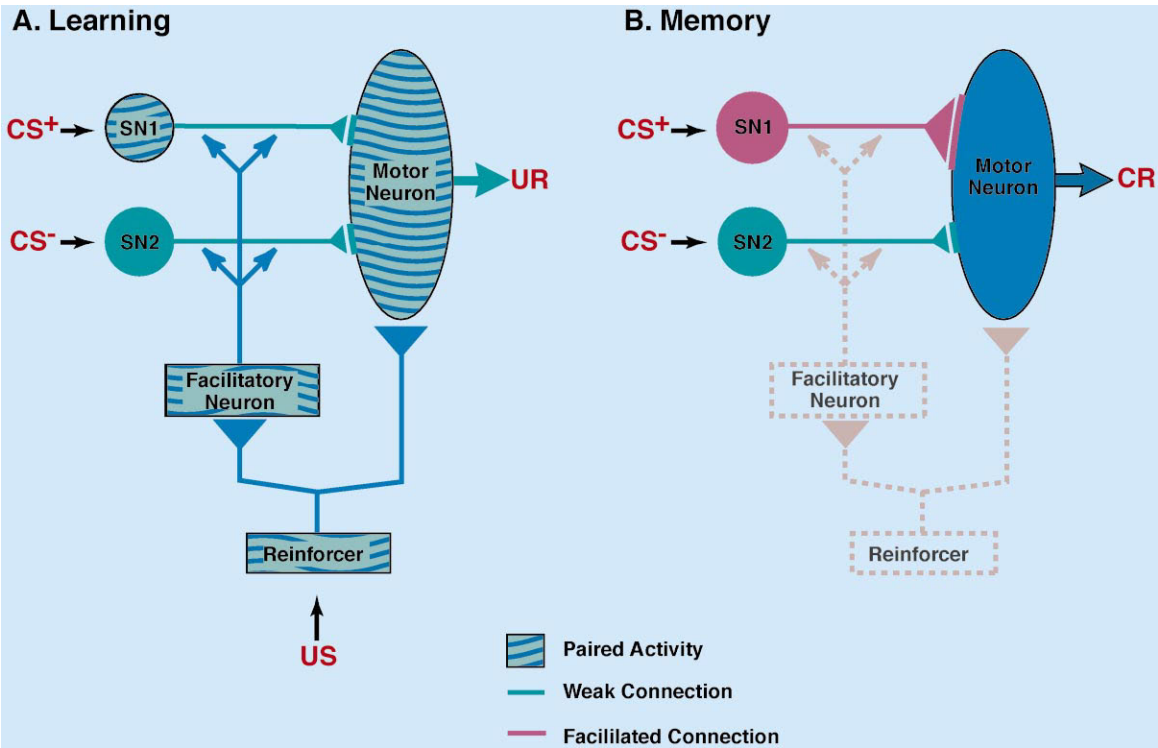

Figure 2
Model of classical conditioning of a withdrawal reflex in *Aplysia*. (A) Activity in a sensory neuron (SN1) along the CS+ (paired) pathway is coincident with activity in neurons along the reinforcement pathway (US). However, activity in the sensory neuron (SN2) along the CS– (unpaired) pathway is not coincident with activity in neurons along the US pathway. The US directly activates the motor neuron, producing the UR. The US also activates a modulatory system in the form of the facilitatory neuron, resulting in the delivery of a neuromodulatory transmitter to the sensory neurons. The pairing of activity in SN1 with the delivery of the neuromodulator yields the associative modifications. (B) After the paired activity in A, the synapse from SN1 to the motor neuron is selectively enhanced. Thus, it is more likely to activate the motor neuron and produce the conditioned response (CR). The synapse between SN2 (whose activity is not paired with the US) and the motor neuron is much less enhanced compared to the synapse between SN1 and the motor neuron. (Modified from Lechner and Byrne 1998)

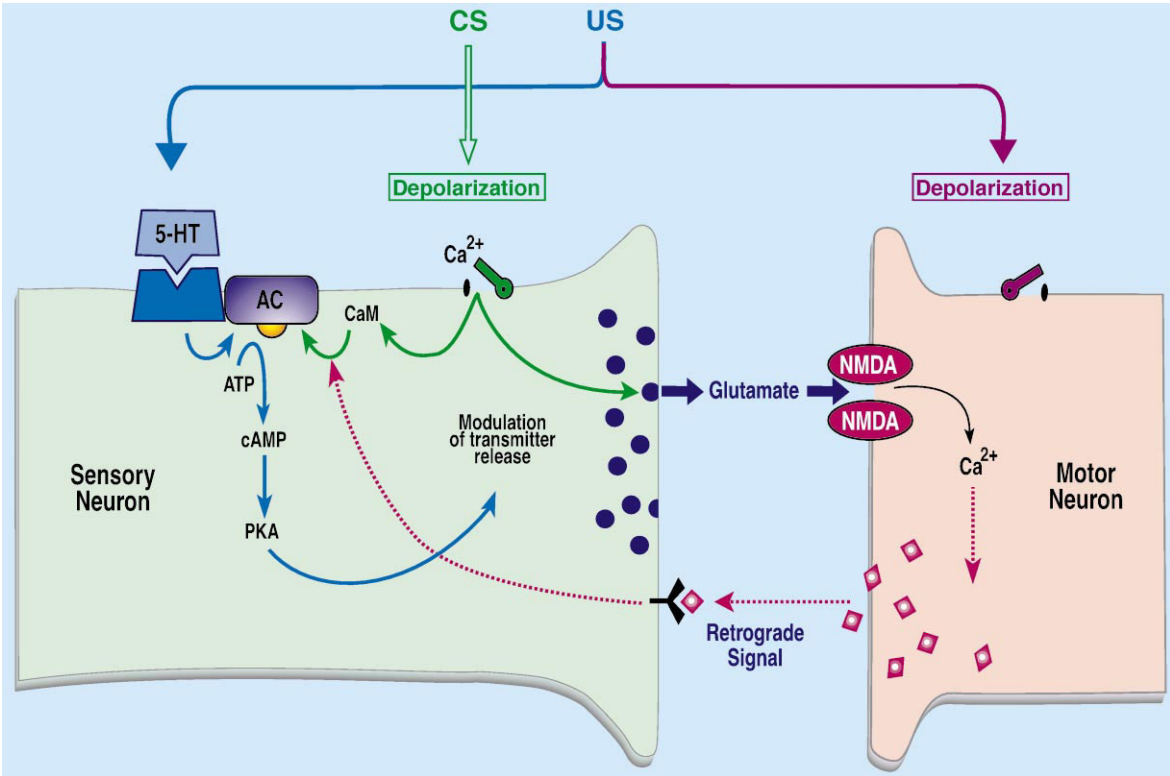

Figure 3
Model of associative facilitation at the *Aplysia* sensorimotor synapse. This model has both a presynaptic and postsynaptic detector for the coincidence of the CS and the US. Furthermore, a putative retrograde signal allows for the integration of these two detection systems at the presynaptic level. The CS leads to activity in the sensory neuron, yielding presynaptic calcium influx, which enhances the US-induced cAMP cascade. The CS also induces glutamate release which results in postsynaptic calcium influx through NMDA receptors if paired with the US induced depolarization of the postsynaptic neuron. The postsynaptic calcium influx putatively induces a retrograde signal, which further enhances the presynaptic cAMP cascade. The end result of the cAMP cascade is to modulate transmitter release and enhance the strength of the synapse. (Modified from Lechner and Byrne 1998)

tiguous with the US mediated reinforcement are selectively modulated. Spiking in sensory neurons during the presence of 5-HT leads to changes in that cell relative to other sensory neurons whose activity was not paired with reinforcement. Thus, a subsequent CS will lead to an enhanced activation of the reflex (Fig. 2(B)).

Figure 3 illustrates a model of the proposed cellular mechanisms responsible for this example of classical conditioning. The modulator (US) acts by increasing the activity of adenylyl cyclase (AC), which in turn increases the levels of cAMP. Spiking in the sensory neurons (CS) leads to increased levels of intracellular calcium. The increased calcium in a spiking sensory neuron enhances the action of the modulator to increase the cAMP cascade. This system determines CS–US contiguity by a method of coincidence detection at the presynaptic terminal. Now, consider the postsynaptic side of the synapse. The postsynaptic region contains NMDA (N-methyl-D-aspartate) receptors. These receptors need concurrent delivery of glutamate and depolarization in order to allow the entry of calcium. The glutamate is delivered by the sensory neuron (CS) and the depolarization is provided by the US (h, i). Thus, the postsynaptic neuron provides another example of coincidence detection. The increase in intracellular calcium putatively causes a retrograde signal to be released from the postsynaptic to the presynaptic terminal, ultimately acting to further enhance the cAMP cascade in the sensory neuron. The overall amplification of the cAMP cascade acts to raise the level of protein kinase A (PKA). The increase in PKA leads to the modulation of transmitter release. These activity-dependent changes enhance the synaptic efficacy between the specific sensory neuron of the CS pathway and the motor

neuron. Thus, the sensory neuron along the CS pathway will be better able to activate the motor neuron and produce the CR.

Another example of associative modifications in an individual neuron can be found in the marine mollusc *Hermissenda* (j). Light is paired with 5-HT application in an analogue of classical conditioning. As with classical conditioning of the withdrawal reflex in *Aplysia*, 5-HT can also act to mediate reinforcement in *Hermissenda*. The conditioning results in both an increase in the excitability of a single cell and an enhancement in the strength of a specific type of synapse.

3. Conclusion

With classical conditioning of the *Aplysia* withdrawal reflex, the paired CS and US form an association by converging on a second messenger cascade within a single cell. This convergence results in the enhancement of a specific synapse. With operant conditioning of *Aplysia* feeding behavior, the association is made through contingent reinforcement. Contingent reinforcement of the response results in the alteration of a cell that mediates the expression of that response. Conditioning occurs through a modulation of the membrane properties of this single cell. Thus, modifications made to individual neurons (via intrinsic membrane properties and synapses) can account for both types of associative learning phenomena.

See also: Classical Conditioning, Neural Basis of; Ion Channels and Molecular Events in Neuronal Activity; Long-term Depression (Cerebellum); Long-term Depression (Hippocampus); Long-term Potentiation and Depression (Cortex); Long-term Potentiation (Hippocampus); Memory: Synaptic Mechanisms; Neural Plasticity; Neurotransmitters; Synapse Formation; Synaptic Efficacy, Regulation of; Synaptic Transmission

Bibliography

Bao J X, Kandel E R, Hawkins R D 1998 Involvement of presynaptic and postsynaptic mechanisms in a cellular analog of classical conditioning at *Aplysia* sensory-motor neuron synapses in isolated cell culture. *Journal of Neuroscience* **18**: 458–66

Frysztak R J, Crow T 1997 Synaptic enhancement and enhanced excitability in presynaptic and postsynaptic neurons in the conditioned stimulus pathway of *Hermissenda*. *Journal of Neuroscience* **17**: 4426–33

Hawkins R D, Abrams T W, Carew T J, Kandel E R 1983 A cellular mechanism of classical conditioning in *Aplysia*: Activity-dependent amplification of presynaptic facilitation. *Science* **219**: 400–5

Lechner N A, Byrne J H 1998 New perspectives on classical conditioning: A synthesis of Hebbian and non-Hebbian mechanisms. *Neuron* **20**: 355–8

Lorenzetti F D, Baxter D A, Byrne J H 2000 Contingent reinforcement with dopamine modifies the properties of an individual neuron in *Aplysia*. *Society for Neuroscience Abstracts* **26**: 1524

Murphy G G, Glanzman D L 1997 Mediation of classical conditioning in *Aplysia californica* by long-term potentiation of sensorimotor synapses. *Science* **278**: 467–71

Nargeot R, Baxter D A, Byrne J H 1999a *In vitro* analog of operant conditioning in *Aplysia*. I. Contingent reinforcement modifies the functional dynamics of an identified neuron. *Journal of Neuroscience* **19**: 2247–60

Nargeot R, Baxter D A, Byrne J H 1999b *In vitro* analog of operant conditioning in *Aplysia*. II. Modifications of the functional dynamics of an identified neuron contribute to motor pattern selection. *Journal of Neuroscience* **19**: 2261–72

Phillips A G, Blaha C D, Fibiger H C 1989 Neurochemical correlates of brain-stimulation reward measured by ex vivo and in vivo analyses. *Neuroscience Biobehavioral Reviews* **13**: 99–104

Stein L, Xue B G, Beluzzi J D 1993 Cellular targets of brain reinforcement systems. *Annals of the New York Academy of Sciences* **702**: 41–60

Walters E T, Byrne J H 1983 Associative conditioning of single sensory neurons suggests a cellular mechanism for learning. *Science* **219**: 405–8

F. D. Lorenzetti and J. H. Byrne

Asymmetry of Body and Brain: Embryological and Twin Studies

An individual's behavior is generated and modulated by the structure and function of the brain and body. The nervous system, which becomes the physical substratum of human cognitive processes, is formed during embryogenesis. During life, environmental stimuli are perceived, processed, and eventually manifest as behavior through the action of the brain and body. Thus, the structure of the body in general, and the brain in particular, are crucial factors in behavior and psychological processes. Therefore, an understanding of the embryonic processes which pattern the human organism is crucial to a complete understanding of cognition and behavior. Because of the ease of experimentation on animal embryos, and the high evolutionary conservation of molecular embryonic mechanisms from invertebrate animals through man, studies in animal embryos have shed much light on how the complex structure of the organism is formed during embryonic development.

1. Concepts in Molecular Embryology

The complex biochemical events which create an organism from a single fertilized egg produce intricate pattern on many scales, from the molecular to the organismic. This is accomplished through complex

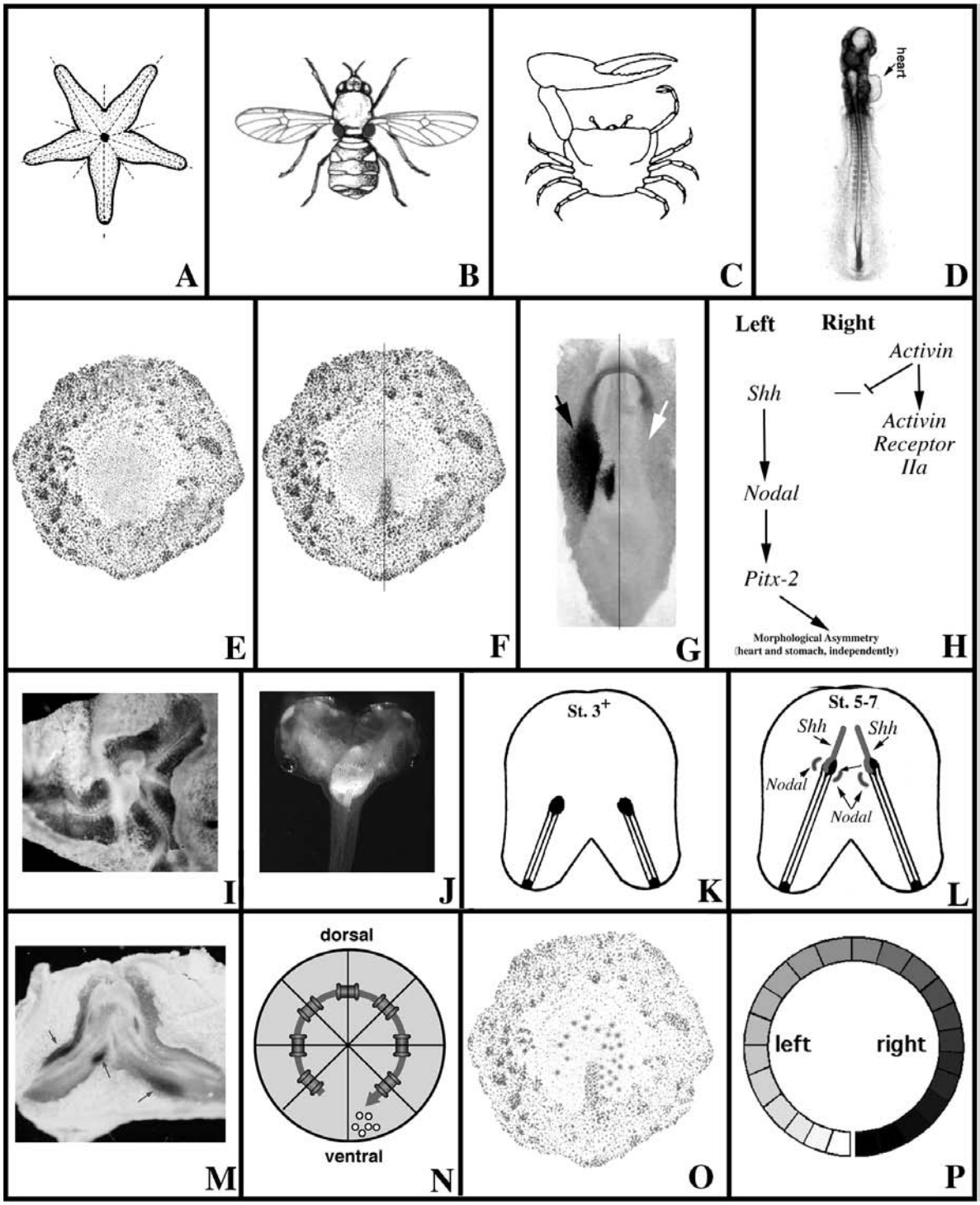

Figure 1
Caption on next page

programs of gene expression which are turned on and off as the egg begins to divide and grow. The cells which are formed as the result of this proliferation have different fates in the embryo, in that they take up different positions and biochemical functions. The genes and maternal products stored in each of the embryo's cells are transcribed into mRNA and translated into proteins which can perform physiological functions or regulate the expression of other genes. Thus, the embryo is a complex dynamical system where a precisely orchestrated interplay of information and materials takes place to generate an organism with a well-defined shape.

2. *Left–Right Asymmetry*

The most basic aspect of any organism's structure is the large-scale symmetry of the body plan (see Fig. 1(A)–(D)). While some invertebrates possess chirality (snail shells), radial symmetry (sea anemonies), or no symmetry at all (sponges), the vertebrate body plan is bilaterally symmetric (defined as having two sides, each of which is identical to its mirror reflection across a midline axis of symmetry). While the general plan of the human body and nervous system is bilaterally symmetric, it contains important and consistent asymmetries of the brain as well as visceral organs such as the heart. The normal human anatomy features lungs which are trilobed on the right and bilobed on the left, a left-pointing cardiac apex and aortic arch, a right-sided liver and gall-bladder, and a left-sided stomach and spleen (Casey and Hackett 2000, p. 32). In particular, lateralization of the brain is involved in many aspects of higher mental function such as speech (Corballis 1983).

The left–right (LR) axis itself follows automatically from the definition of the anterior–posterior (AP) and dorsal–ventral (DV) axes, as it is perpendicular to both; however, consistently imposed asymmetry across it is fundamentally different from patterning along the other two axes. First, while the AP and DV axes can be set by exogenous cues such as gravity, or sperm entry point, there is no independent way to pick out the left (or right) direction, since no obvious macroscopic aspect of nature differentiates left from right. This problem can be most acutely felt by imagining trying to explain to someone using only words which hand is the 'Left' hand: unless an asymmetric object exists to which one can refer, it is an extremely difficult task.

Second, all normal members of a given species are asymmetrical in the same direction (e.g., the heart is always on the same side in normal individuals). However, animals with complete mirror reversal of internal organs can arise (*situs inversus*) and are otherwise phenotypically unimpaired. Thus, while it is possible to come up with plausible evolutionary reasons for why organisms might be asymmetric in the first place (optimal packing of viscera, etc.), there is no obvious reason for why they should all be asymmetric in the same direction. It is, after all, much easier to

(A)–(D) *Symmetry types in the animal kingdom*: starfish have a fivefold radial symmetry (A); some insects such as Drosophila are bilaterally symmetric (B); Interestingly, some invertebrates such as crabs have a bilaterally symmetric bodyplan with conspicuous asymmetries such as claws (C). The vertebrate body plan is also bilaterally symmetrical, but contains conserved asymmetries of the viscera and heart (D). (E)–(I) *Left–right asymmetry in chick embryos*: the early chick embryo is a flat, radially symmetric two-layer disk of cells (E). The radial symmetry is broken into a bilateral symmetry by the appearance of the primitive streak at one point on the periphery. As the streak grows away from the periphery, its base becomes the posterior end of the embryo, while its tip becomes the anterior, and will induce the formation of the head and other nervous structures (F). Several genes have asymmetric expression patterns in early embryos; for example, the gene *Nodal* is expressed in two domains on the left side of the midline, shown here in the chick embryo (G). The asymmetric genes function in a pathway (H): activin on the right side inhibits the expression of *Shh*, which is then present only on the left. *Shh* induces expression of *Nodal* (only on the left side), which induces *Pitx-2*, and then transduces this information to the forming heart (I) and viscera. (J)–(L) *Conjoined twinning and asymmetry*: conjoined twins occur in embryos of chick (such as the triplets in panel I), frog (J), and other animal models. Human twins which are joined at the midline (J) are known to often exhibit disturbances of laterality. In the chick, conjoined twins sometimes occur by the appearance of two primitive streaks which begin far apart in the embryonic field (K) but then grow together with time. As the tips of the streak near each other, the expression of *Shh* on the left side of the tip of the primitive streak can induce an ectopic nodal domain on the right side of the left twin (L). This is observed in spontaneous twins which are examined for expression of nodal (M); such twins show laterality disturbances in the left twin because of the extra nodal domain on the right side. (N)–(P) *Gap junctions and asymmetry*: gap junctions are tunnels formed by connexin proteins on either side of the membrane of two opposing cells. In early frog embryos, radial patterns of open gap junctions allow the traversal of small molecular signals (morphogens) across embryonic fields; these molecules then accumulate on one side of a junctional barrier on the ventral midline (N). A very similar process occurs in the chick (O), where the junctional paths allow morphogens to flow across the blastoderm to accumulate on one side of the primitive streak. This process is schematized in (P): junctional paths allow signals to traverse embryonic fields to preferentially accumulate on one side of the midline, thus converting a cell-level asymmetry (which drives one-way transfer through gap junctions) to an embryo-wide asymmetry

imagine a developmental mechanism for generating asymmetry (such as positive feedback and amplification of stochastic biochemical differences) than for biasing it to a given direction. The left–right axis thus presents several unique and deeply interesting theoretical issues.

Normal asymmetry (*situs solitus*) can be perturbed in one of several ways, and various human syndromes recapitulate all of these (Burn 1991, Winer-Muram 1995). *Situs inversus* is a complete mirror image inversion of asymmetry, and does not usually cause complications because all interconnections are made properly. However, isomerism (a loss of asymmetry, where an embryo has two left or two right sides, which can result in polysplenia or asplenia, a midline heart, etc.) and heterotaxia (a loss of concordance, where each organ makes an independent decision as to which side it will go towards) are associated with serious medical complications.

3. *The Molecular Basis of Asymmetry*

Molecular embryology has made great strides in understanding the patterning of the antero-posterior and dorsoventral axes. However, prior to 1995, very little was known about the molecular mechanisms which give asymmetric cues to the left and right sides of the body. A large catalog of asymmetries in the animal kingdom was available (Neville 1976). Besides the human syndromes which exhibited alterations in laterality, a variety of mouse mutants existed which displayed various laterality phenotypes. However, the nature of the mutations responsible for these effects was unknown. It had been observed that a variety of drugs was able to induce asymmetry phenotypes in embryos (Levin 1997), and timing experiments using these drugs indicated that processes which lead to asymmetry function very early in embryonic development (prior to neural tube formation).

The identification of a chicken gene, *Sonic Hedgehog* (*Shh*), cloned in 1993 because of its role in limb and neural tube patterning (Riddle et al. 1993) turned out to be pivotal in understanding left–right patterning. The chick embryo begins as a flat disk only three cell layers thick and is radially symmetric (Fig. 1(E)). However, soon afterwards a structure appears which is called the primitive streak (see Fig. 1(F)). This groove of special cells which are undergoing gastrulation is common to mammals and many other types of embryos, and marks the breakage of symmetry from the radial to the bilateral, since it defines the midline of the embryo, and sets up the antero-posterior axis. When the chick embryo at this stage is examined for the presence of mRNA of the gene *Shh* by a procedure which deposits blue stain wherever the gene is active, it is seen that this gene is only transcribed on the *left* side of the tip of the primitive streak, and not the right; this is even more obvious for a gene called *Nodal*, which is expressed in a wide domain on the left flank. The identification of several other asymmetric genes followed, and now a variety of genes is known which are active on the left or right sides (Levin 1998b). More importantly, by a variety of experiments involving artificial retroviruses, gene cloning and misexpression, and transgenic mouse technology, it was shown that these genes function in a pathway: they turn each other's expression on and off in a spatially and temporally specific manner to give each side of the body an identity. The presence of the *Shh* gene defines the left side as left; its absence defines the right side. These gene cascades provide information which is interpreted by asymmetric organs such as the heart to enable them to know which direction is left and which is right (summarized in Fig. 1(H)). For example, it was shown that if the normally left-sided *Shh* expression is artificially induced on the right side as well (thus producing an embryo both of whose sides think they are 'left'), the heart forms properly, and loops correctly in 50 percent of the cases, but loops to the opposite side in the other 50 percent. In effect, the heart (and other viscera) must choose randomly when presented with conflicting information. The details of the molecular cascade of asymmetric genes can be seen in reviews such as Burdine and Schier (2000), Fujinaga (1996), and Wood (1997).

Several of these genes are of particular interest to behavioral science, as they presage signals for the determination of brain laterality. For example, the chick gene *Cerberus* is expressed on the left side of the embryonic head, shortly after neural tube closure (Zhu et al. 1999). The same is true of genes such as *Pitx-2* in zebrafish embryos (Essner et al. 2000). It should be kept in mind that though most of this data comes from animal models, there is a high degree of conservation of embryonic mechanisms, and these gene cascades are very likely to be crucial in patterning all vertebrates including man (Casey and Hackett 2000).

4. *Conjoined Twins and Laterality Defects*

The identification of asymmetrically expressed genes whose protein products specify the spatial asymmetry of the viscera and brain made it possible to explain, on a molecular level, an observation which had been found in conjoined human twins. Twins and the twinning process have had a relevance for laterality research ever since the tantalizing experiments of Spemann (Spemann and Falkenberg 1919). It has been noted (Burn 1991, Morrill 1919, Newman 1916, Schwind 1934, Winer-Muram 1995) that conjoined twins in human beings, mice, and frogs tend to exhibit laterality defects. Animal models of twinning have been instrumental in understanding this defect in embryogenesis (Fig. 1(I), (J)).

In reviewing the human literature on conjoined twins, it was observed that parapagus and thoracopagus twins (twins joined at the chest or side-to-side) tend to exhibit *situs* abnormalities (Levin et al. 1996); these are twins thought to originate from two adjacent embryonic streaks developing side by side, either in parallel or obliquely (schematized in Fig. 1(K)). Guided by the LR pathway, Levin et al. (1996) examined the expression of the asymmetric genes in analogously positioned chick twins and proposed two models explaining laterality defects found in conjoined twins. These are both based on molecules in the LR pathway crossing some distance in the blastoderm and affecting the conjoined embryo. The precise details of geometric arrangement and timing determine which members of the LR cascade affect the twin, and thus control which twin exhibits the *situs* anomaly.

For example, when primitive streaks arise far apart, but grow towards each other during gastrulation (Fig. 1(K), (L)), *Shh* expression proceeds normally in both twins on the left side of the tip of the primitive streak. However, during head-fold stages, the *Shh* expression of the right twin induces not only the normal left-sided *Nodal* expression, but also induces aberrant *Nodal* expression on the right side of the left twin (Fig. 1(L)). When head-fold stage spontaneous chick twins with oblique streaks are examined for *Nodal* expression (Fig. 1(M)), as predicted it was seen that the right twin has *nodal* expression only on the left side (i.e., normal expression), while the left-most twin has expression (black arrows) on both the left and right sides (which leads to laterality defects). Thus, studies of twins in model systems such as the chick have provided a crucial arena in which to test molecular models which explain clinical findings in human teratology.

5. *Nonconjoined Twins and Asymmetry*

The models discussed above present plausible explanations of laterality defects in conjoined twins. There is, however, an interesting set of observations which suggest that they do not tell the whole story, and that even in mammals, chirality is determined as early as the first few cell divisions, and certainly before the streak appears. Nonconjoined monozygotic twins, while not exhibiting the kinds of visceral laterality defects that occur in conjoined twins, do manifest many subtler kinds of mirror-image asymmetry ('bookend' or enantiomer twin pairs), where some characteristic is left-sided (or counterclockwise) in one twin but right-sided (or clockwise) in the other.

Pairs of such twins have been noted to present such mirror asymmetries in hand preference, hair whorl direction, tooth patterns, unilateral eye and ear defects, and even tumor locations and undescended testicles (Beere et al. 1990, Carton and Rees 1987, Cidis et al. 1997, Gedda et al. 1981, Morison et al. 1994, Newman et al. 1937, Townsend and Richards 1990, Yager 1984). Most healthy, nonconjoined twins presumably result from separation of cleavage, morula, or early blastocyst stage embryos (James 1983). Thus, some chiral information may be present in the very early mammalian embryo, manifesting itself in hair whorls etc. if the cells are separated at an early stage. In contrast, the asymmetry of the major body organs seems to be unspecified (or at least plastic enough to be respecified) at those stages, and is developed correctly for both monozygotic twins. This may be related to the fact that heterotaxic reversals in hair whorls and tooth patterns would not be expected to be disadvantageous, while discordant *situs* for internal organs clearly is subject to negative evolutionary pressure.

6. *Upstream of Asymmetric Genes*

Pathways of asymmetric gene expression beg the question: what determines the sidedness of the first asymmetric gene? The very first step of asymmetry at the cellular level is presently unknown, but is thought to perhaps involve a chiral (handed) molecule which is tethered with respect to the other two axes, and has an activity which distinguishes left from right (Brown and Wolpert 1990). Interestingly, the identification of such a mechanism would still leave an important question unanswered, because for a cell, knowing which direction is left vs. right is a very different problem than knowing which side of the midline it is on. Knowing its position relative to the midline is crucial for proper expression of asymmetric genes. Thus, what is needed is a mechanism for transducing left–right directional information at the level of a single cell into embryo-wide fields of asymmetric gene expression.

Recent experiments have shown that a system of cell–cell communication exists by means of which the two sides of very early embryos decide which side is to be the right side and which is to be the left (Levin and Mercola 1999). Such a system involves cell–cell channels known as gap junctions, through which small molecular signals can pass (Fig. 1(N)). Once crucial LR decisions are made in a small group of cells, this information can be distributed to the whole organism through circumferential paths of open gap junctions (Fig. 1(O), (P)). Twin embryos will be crucial in testing these models because the gap junction models make specific predictions about what should happen to circumferential paths when more than one primary axis exists in an embryonic field (as in conjoined twins growing side by side, in parallel).

7. *Twins and the Origin of the Primary Axis*

Bilateral asymmetry presupposes the existence of a midline axis, across which symmetry or asymmetry is observed. In the chick this is accomplished by the

primitive streak growing out of a specific point on the periphery of the blastoderm, and thus breaking radial symmetry into bilateral symmetry. However, it is known that when suitably stimulated, any point on the periphery is able to serve as the site of streak initiation. Yet, in the great majority of embryos, only one primary axis is observed (singleton embryos). What is responsible for the appearance of one and only one streak, when many points could potentially initiate streaks?

Research has showed that at very early stages, many chick embryos in fact consist of more than one primary axis, and are conjoined twins. Interestingly, at this crucial stage of embryogenesis, the streaks secrete a streak-inhibitory factor in an attempt to suppress the appearance of other streaks (Levin 1998a). In most embryos, this competition results in the primary dominant streak inhibiting other streaks and in fact causing the extra streaks to dissolve. Thus, twinning illustrates the mechanisms underlying the control of the number of primary axes in the embryonic field—a key component of embryogenesis.

8. *Prospects for the Future*

The phenomenon of twinning provides crucial insight into the earliest steps of embryonic pattern formation; luckily, twins can be experimentally induced in various systems with ease (Fig. 1(I), (J)). Conjoined animal twins allow researchers to answer questions regarding mechanisms by which primary axes arise from a competent embryonic field. Moreover, conjoined twins are the perfect system in which to characterize paths which facilitate, and barriers which block, the flow of left–right information during embryogenesis. This information is of vital importance in understanding normal mechanisms of embryonic development.

Nonconjoined twins provide a special challenge to molecular embryology. The bookending phenomenon, which displays a conservation of chirality in nonconjoined twins, provides tantalizing clues about the very first steps in the generation of embryonic asymmetry, and has implications for evolution of the vertebrate body plan. The data suggesting that the head and brain have a system of asymmetry determination distinct from the rest of the body have far-reaching implications for cognitive science. Progress on this issue will be a key part of a molecular understanding of the embryonic development of systems which generate and shape human behavior.

See also: Behavioral Genetics: Psychological Perspectives; Brain Asymmetry; Comparative Method, in Evolutionary Studies; Genetic Studies of Behavior: Methodology; Genetic Studies of Personality; Intelligence, Genetics of: Cognitive Abilities

Bibliography

Beere D, Hargreaves J, Sperber G, Cleaton-Jones P 1990 Mirror image supplemental primary incisor teeth in twins: Case report and review. *Pediatric Dentistry* **12**: 390–2

Brown N, Wolpert L 1990 The development of handedness in left/right asymmetry. *Development* **109**: 1–9

Burdine R, Schier A 2000 Conserved and divergent mechanisms in left-right axis formation. *Genes Development* **14**: 763–76

Burn J 1991 Disturbance of morphological laterality in humans. *CIBA Foundation Symposium* **162**: 282–96

Carton A, Rees R 1987 Mirror image dental anomalies in identical twins. *British Dentistry Journal* **162**: 193–4

Casey B, Hackett B 2000 Left-right axis malformations in man and mouse. *Current Opinion in Genetic Development* **10**: 257–61

Cidis M, Warshowsky J, Goldrich S, Meltzer C 1997 Mirror-image optic nerve dysplasia with associated anisometropia in identical twins. *Journal of the American Optometry Association* **68**: 325–9

Corballis M C 1983 *Human Laterality*. Academic Press, New York

Essner J J, Branford W W, Zhang J, Yost H J 2000 Mesendoderm and left-right brain, heart and gut development are differentially regulated by pitx2 isoforms. *Development* **127**: 1081–93

Fujinaga M 1996 Development of sidedness of asymmetric body structures in vertebrates. *International Journal of Developmental Biology* **41**: 153–86

Gedda L, Brenci G, Franceschetti A, Talone C, Ziparo R 1981 Study of mirror imaging in twins. *Progress in Clinical Biological Research* **69A**: 167–8

James W 1983 Twinning, handedness, and embryology. *Perceptual and Motor Skills* **6**: 721–2

Levin M 1997 Left–right asymmetry in vertebrate embryogenesis. *BioEssays* **19**: 287–96

Levin M 1998a Follistatin mimics the endogenous streak inhibitory activity in early chick embryos. *International Journal of Developmental Biology* **42**: 553–9

Levin M 1998b Left–right asymmetry and the chick embryo. *Seminars in Cell Developmental Biology* **9**: 67–76

Levin M, Mercola M 1999 Gap junction-mediated transfer of left-right patterning signals in the early chick blastoderm is upstream of Shh asymmetry in the node. *Development* **126**: 4703–14

Levin M, Roberts D, Holmes L, Tabin C 1996 Laterality defects in conjoined twins. *Nature* **384**: 321

Morison D, Reyes C, Skorodin M 1994 Mirror-image tumors in mirror-image twins. *Chest* **106**: 608–10

Morrill C 1919 Symmetry reversal and mirror imaging in monstrous trout and a comparison with similar conditions in human double monsters. *Anatomical Record* **16**: 265–92

Neville A 1976 *Animal Asymmetry*. Edward Arnold, London

Newman H 1916 Heredity and organic symmetry in armadillo quadruplets. *Biological Bulletin* **XXX**: 173–203

Newman H, Freeman F, Holzinger K 1937 *Twins: a Study of Heredity and Environment*. University of Chicago Press, Chicago

Riddle R, Johnson R, Laufer E, Tabin C 1993 Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell* **75**: 1401–16

Schwind J 1934 Symmetry in spontaneous twinning in *Rana sylvatica*. *Anatomical Record*. **58**: 37

Spemann H, Falkenberg H 1919 Über Asymmetrische Entwicklung und Situs inversus viscerum bei Zwillingen und Doppelbildungen. *Roux Archives* **45**: 371–422

Townsend G, Richards L 1990 Twins and twinning, dentists and dentistry. *Australian Dentistry Journal* **35**: 317–27
Winer-Muram H 1995 Adult presentation of heterotaxic syndromes and related complexes. *Journal of Thoracic Imaging* **10**: 43–57
Wood W 1997 Left-right asymmetry in animal development. *Annual Reviews in Cell Development Biology* **13**: 53–82
Yager J 1984 Asymmetry in monozygotic twins. *American Journal of Psychiatry* **141**: 719–20
Zhu L, Marvin M, Gardiner A, Lassar A, Mercola M, Stern C, Levin M 1999 Cerberus regulates left-right asymmetry of the embryonic head and heart. *Current Biology* **9**: 931–8

M. Levin

Atomism and Holism: Philosophical Aspects

In the philosophy of the social sciences, atomism is the view that human beings can be thinking, rational beings independently of social relations. Holism, by contrast, is the view that social relations are essential to human beings insofar as they are thinking, rational beings. This article first provides an overview of different sorts of atomism and holism (see Sect. 1). It then briefly sketches the historical background of these notions in modern philosophy (Sect. 2). The main part is a systematic characterization of atomism and holism (Sect. 3) and a summary of the most important arguments for both these positions (Sect. 4).

1. *Types of Atomism and Holism*

If one speaks of atomism or holism, one has to specify with respect to what atomism or holism is claimed. Atomism with respect to one thing or aspect can go with holism with respect to another thing or aspect, and *vice versa*. In the social sciences, the discussion on atomism vs. holism focuses on the question whether social relations to other humans are essential to human nature. The question thus is whether society in the sense of a social community of human, thinking beings is an atomistic or a holistic system. The question is neither whether a human being can exist in the biological sense without social relations nor whether a human being is dependent on social relations for maintaining a fully satisfactory human life. The point at issue is whether a human being is dependent on social relations to other humans insofar as he or she is a thinking, rational being. Since, on most accounts, thought in the full sense of the term as exhibited by adult human beings is tied to rationality constraints and since rational agency, including moral agency, presupposes thought, the point at issue can be formulated thus: are social relations necessary for a human being to have thoughts with a determinate conceptual content? Social atomism is the thesis that an individual considered in isolation can have thoughts with a determinate conceptual content. Social holism, by contrast, is the thesis that social relations are essential for a human being in order to have thoughts with a determinate conceptual content. Social holism thereby implies that having thoughts is tied to speaking a public language; social atomism does not have such an implication.

The discussion on atomism vs. holism concerns not only the conditions under which humans can have thoughts; it extends also to aspects of thoughts. The most important issue in this respect is the question whether meaning is atomistic or holistic. Semantic atomism is the thesis that each thought or sentence has a meaning independently of other thoughts or sentences. Semantic holism, in reverse, is the thesis that the meaning of a thought or a sentence consists in its inferential relations to other thoughts or sentences in a system of thoughts, a theory, etc. There is a middle position, namely semantic molecularism. This is the thesis that only some particular inferential relations to other thoughts or sentences, most prominently the analytic ones, constitute the meaning of a thought or a sentence.

Atomism and holism are put forward with respect to further aspects of thoughts. Confirmation atomism is the thesis that thoughts or sentences can be empirically confirmed or disconfirmed one by one. Confirmation holism, by contrast, is the thesis that only a whole system of thoughts or a whole theory can be confirmed or disconfirmed by experience. Consequently, if a conflict between theory and experience occurs, there are several possibilities for adapting the theory to experience. Justification atomism is the thesis that there are some thoughts or sentences which can be justified without referring to other thoughts or sentences. These thoughts or sentences are the foundation of knowledge. Justification atomism is, therefore, known as foundationalism. Justification holism, by contrast, is the thesis that the justification of any thought or sentence consists in relations to other thoughts or sentences. Justification holism is, therefore, known as coherentism: the justification of a thought or a sentence is its being integrated into a coherent system of thoughts or sentences. Although there are arguments which link all these sorts of atomism or holism together, endorsing any one sort of atomism or holism does not logically commit one to endorse any other sort of atomism or holism as well.

2. *The Historical Background of Atomism and Holism*

The issue of atomism vs. holism in the social sciences is linked closely with philosophical reactions to what is perceived as the methodology of modern physics.

Thus, social atomism starts when Hobbes applies the method of physical science to social science. Hobbes—in contrast to, for instance, Gassendi—is not a physical atomist: he does not commit himself to the thesis that there are smallest, indivisible bodies which have a few basic properties such as position and motion and of which every complex body is composed (*De corpore/Of Body*, Chap. 15, Sect. 2, Chap. 27, Sect. 1). However, he transfers to the social sciences the method of (a) dividing a complex system up into its constituent parts, (b) considering the parts and their properties in isolation from each other, and (c) understanding the nature and function of a complex system on the basis of the properties of its constituent parts considered in isolation (*De cive/Of the Citizen*, Preface). Hobbes is famous for applying this method to political science. He thereby commits himself to social atomism: in his political philosophy, he presupposes individuals each of whom is a thinking, rational being independently of the other ones (*De cive/Of the Citizen*, Chap. 8, Sect. 1). He reconstructs social and political relations on that basis.

Despite the widespread acceptance of the sketched methodology of the natural sciences as a paradigm for philosophy, holism is proposed as an ontology in early modern philosophy. Most prominently, Spinoza maintains in his *Ethics* that there is only one substance and that everything is a way of being (a mode) of that substance. However, it is only when the supposition that the methodology of physical science is a paradigm for philosophy is challenged in romanticism and its predecessors that social holism emerges as an option. This development is linked with a focus on language and the thesis that language is essential to thought. Thus, for instance, Herder claims (a) that thought is tied to language and (b) that language is tied to social relations. Consequently, he comes to the conclusion that a human can be a thinking being only in virtue of social relations to other humans (see, in particular, *Abhandlung über den Ursprung der Sprache*; partial English translation in Herder 1993 and see Spencer 1996). Hegel systematizes this conception and integrates it into his holistic ontology according to which the world is the dialectical development of a spiritual substance. He goes as far as considering individual thinking beings as accidents of a collective substance (for instance, *Grundlinien der Philosophie des Rechts*, Sect. 156; English translation Hegel 1991). He can be seen as thereby transforming social holism into collectivism.

Social holism in the romanticists is part of their reaction against conceiving philosophy on the model of the methodology of the physical sciences. But the romanticists do not go as far as developing a proper methodology for the social sciences. It is only from the beginning of the twentieth century on that social holism is proposed as a proper methodology for the social sciences. Most prominent in that respect are the works of Durkheim (1927) and Mead (1934). This development again links up with a development in the natural sciences, namely the emergence of systems theory. Systems theory seeks to understand a system on the basis of its constituent parts—but not on the basis of properties which these parts have independently of each other; on the contrary, systems theory sets out to understand a complex system on the basis of the relations into which its constituent parts enter (von Bertalanffy 1968). Applied to social systems, systems theory can therefore lead to social holism. Moreover, as regards physics, it is claimed that atomism is replaced with a comprehensive holism in our description of the microphysical level consequent upon quantum theory (for instance, Teller 1986). However, this holism has—as yet?—no repercussions for the social sciences.

3. *A Characterization of Atomism and Holism*

Both atomism and holism, taken in a general sense, are about complex systems in which constituent parts can be distinguished. Having constituent parts is not sufficient for something to be an atomistic system. In reverse, there is no reason to recognize something which does not admit of parts as a holistic system; such a thing would rather be an atom in the literal sense. Both atomism and holism therefore presuppose the analysis of a complex system in terms of parts. For a characterization of atomism in opposition to holism, it is crucial to consider (a) the sort of independence or dependence that holds among the parts and (b) the properties of the parts with respect to which that sort of independence or dependence holds.

Relations of causal dependence among the parts of a complex system are not sufficient to settle the issue of atomism vs. holism, even if they are complicated relations of mutual or circular causal dependence. No one denies that the development of thought and rationality in a human being causally depends on the social environment and that these social relations can be complex ones of mutual and circular causal dependence. But this commonplace does not amount to a refutation of social atomism and an argument for social holism, on pain of trivializing social holism and turning social atomism into an absurd position.

It is, therefore, appropriate to focus on whether or not there are relations of ontological dependence among the parts of a complex system. Ontological dependence is about something depending for its very existence on the existence of some other thing. Two sorts can be distinguished. (a) If an individual *a* cannot exist unless there is a particular individual *b*, then we have a case of rigid ontological dependence. (b) If there can be no individual of a certain kind unless there is some other individual of a certain kind, then we have a case of generic ontological dependence (Simons 1987, Chap. 8.3). The issue of atomism vs. holism is about generic ontological dependence among

the parts of a complex system. The holist claims that some social environment is necessary for thought, but not that it has to be, say, one's family, or one's tribe.

According to the standard account of modal sentences in today's philosophy, the necessity which ontological dependence implies is to be analyzed in terms of possible worlds. One can claim that the necessity is metaphysical and thus concerns all possible worlds. But one does not have to go as far as that: one can restrict the scope of ontological dependence so that it applies, for instance, only to all physically possible worlds, i.e., all those worlds in which the same physical laws hold as in our world. Consequently, one can take the followig position: social atomism or social holism applies only to human thinking beings, i.e., to those who share a human nature and who are subject to the physical and biological laws of our world. One thereby refuses to commit oneself as to whether or not in the imagined case of Maritians, who are thinking beings, but who do not have a human nature, social relations are necessary for thought. Hence, one can conceive social atomism or social holism in such a way that these positions formulate a conceptual necessity about all possible thinking beings; or one can restrict these positions in such a way that they apply only to us humans.

When it comes to atomism vs. holism, we can further specify generic ontological dependence: (a) we are interested in generic ontological dependence as a symmetric relation; for we are enquiring into the sort of dependence that captures the way in which the parts of a holistic system depend on each other. (b) In cases such as social atomism vs. social holism and atomism vs. holism with respect to meaning, confirmation, or justification of thoughts, generic ontological dependence concerns the dependence of things of one sort on there being other things of the same sort (Fodor and Lepore 1992, pp. 1–2). In social atomism vs. social holism, the question is whether there have to be other human thinking beings if there is one human thinking being. This point does, however, not apply to all cases of holism: if an organism is a holistic system, then something which is a heart is not ontologically dependent on there being other hearts, but on there being blood, kidneys, etc.

Most importantly, (c) the issue of atomism vs. holism can be formulated in a more fine-grained way than being about generic ontological dependence that concerns the existence as such of the parts of the kind of systems in which we are interested. The issue is about generic ontological dependence among the parts of a complex system with respect to certain of their properties. The claim of social holism is not that a human being cannot exist unless there are other humans. The claim is that a human being is ontologically dependent on social relations to other humans with respect to the property of thought and rationality, i.e., insofar as it is a thinking, rational being. Otherwise, social holism would be an absurd position: it would have to claim either that a baby already has thoughts in the full sense of the term, or that a baby and an adult cannot be the same individual. Thus, for a characterization of atomism vs. holism, it is necessary to filter out properties for which it is a substantial issue whether or not the parts of a complex system are ontologically dependent on each other with respect to these properties.

For every system of a qualitative kind, there is a family of nondisjunctive, qualitative properties which make something a system of the kind in question. Such a family of properties can include both nonrelational and relational properties. Something is a system of the kind *S* if and only if it has all—or by far most of all—the properties which make something an *S*. For instance, if something is a grain of sand, a certain molecular structure as well as a shape, seize, and mass within a certain margin count among the family of properties which make something a grain of sand. Applied to the parts of an *S*, this notion is to say: There is a family of qualitative, nondisjunctive properties which make something a part of an *S*—in case the thing is arranged with other things in a suitable way. Having all or nearly all the properties that belong to such a family of properties is a necessary condition for something to be a part of an *S*; it is a sufficient condition in conjunction with the condition of a suitable arrangement with other things.

The condition of a suitable arrangement has to be imposed in order to exclude those properties for which it is trivial that something can have them only by being part of a complex system. For instance, a human being can exercize a social role—such as being a judge, or being a salesman—only by being part of a social system. Social roles are arrangement properties. Pointing out social roles is not sufficient to make a case for social holism, again on pain of trivializing social holism and turning social atomism into an absurd position. The question is whether properties that are the prerequisite for exercising a social role—such as the property of thought and rationality in the sense of having thoughts with a determinate conceptual content—require social relations. By excluding from the mentioned family of properties those relational properties in which the arrangement with other things consists, we pick out the properties which underlie the arrangement. There is a substantial case of holism if and only if these properties are relational as well—in the sense that one thing can have these properties only if there are other things together with which this thing is arranged in such a way that there is an *S*. Consequently, there is a substantial case of atomism if and only if something can have these properties independently of whether or not there are other things with which this thing is arranged in such a way that there is an *S*.

The issue of atomism vs. holism can now be characterized as follows: Consider a system of the kind *S* and its constituent parts. For every constituent part

of an *S*, there is a family of qualitative, nondisjunctive properties which make something a constituent part of an *S* provided that there is a suitable arrangement. The issue whether or not an *S*is holistic depends on whether or not the following condition is satisfied by all the things which are its constituent parts: with respect to some of the properties that belong to such a family of properties, a thing is ontologically dependent in a generic way on there being other things together with which it is arranged in such a way that there is an *S* (Esfeld 1998). If there are some properties which satisfy this condition, the system is holistic with respect to these properties. With respect to all those properties within the mentioned family of properties that do not satisfy this condition, the system is atomistic.

This is a fine-grained characterization of atomism and holism: one and the same system can be atomistic in regard to some properties, but holistic in regard to other properties. Apart from the above mentioned variable of the scope of ontological dependence, there are two variables with respect to the properties that atomism or holism concerns: (a) the number of these properties: for instance, in the case of a system of thoughts, one can maintain that holism concerns both confirmation and meaning, but also that it concerns only confirmation; one can hence claim that a system of thoughts is a holistic system with respect to confirmation, but an atomistic system with respect to meaning (Fodor and Lepore 1992, Chap. 2); (b) the importance of these properties: all the properties which atomism or holism concerns are important in that they belong to the family of properties which make something a constituent part of a system *S* provided that there is a suitable arrangement. Nonetheless, the properties which are touched by atomism or holism can be more or less important as far as the identity of the system and its constituents is concerned. For instance, in the case of a thinking human being who is a constituent part of a social community, the property of having thoughts is not necessary for the identity of the human being concerned. If a human being develops from a baby into a normal adult and then goes insane, she nevertheless remains the same human being throughout this development. By contrast, consider the case of atomism or holism of thought with respect to meaning: if a constituent of a system of thoughts changes its meaning, it is no longer the same thought.

The outlined characterization makes clear that the issue of atomism vs. holism is distinct from the issue of individualism vs. collectivism. The former issue is about the way in which the constituent parts of a complex system are interrelated with each other. It is noncommittal on the issue of whether or not there are forces or regularities by means of which a whole influences the behaviour of its parts. Collectivism is the claim that there are such social forces; this claim is not implied by social holism. Social holism can go with individualism: it can be conceived as being about the requirements under which a human being can develop into an individual personality (Pettit 1993, Chaps. 3, 4); see *Individualism versus Collectivism: Philosophical Aspects*.

4. *Arguments for Atomism and Holism*

The main argument for social holism and the main challenge which social atomism has to meet stem from the work of the later Wittgenstein, in particular the *Philosophical Investigations* and the interpretation of that work by Kripke (1982). Kripke sets out what is known as the problem of rule-following, i.e., the problem how the thoughts of finite thinking beings can have a determinate conceptual content. He raises the following issue: any finite sequence of examples or of actions including linguistic actions of persons can in principle be described in terms of infinitely many different rules. Consequently, (a) there are infinitely many possible ways of continuing any such sequence in any new situation, and (b) for any new situation, it is not determined what is the correct way to go on. The problem of rule-following then is this one: What can select for us one of infinitely many rules so that for us one way of going on is singled out as the correct one? See *Meaning and Rule-following: Philosophical Aspects*.

According to Kripke's interpretation of Wittgenstein, to the extent that there is rule-following at all, social interactions are necessary in order to (a) select for a person one of infinitely many meanings in a finite sequence of examples or of actions and (b) enable a person to have a distinction between correct and incorrect rule-following at their disposal. This is the main argument for social holism. This argument speaks against the programme to reduce the description of thoughts to a description in terms of the natural sciences.

There are two main ways in which social holism can be spelled out. One possibility is to conceive social holism in terms of an asymmetric I–we relation: an individual becomes a thinking being by being integrated into a social community. The members of the community assess the actions of the individual according to what they take to be fit and proper. In this case, norms are reduced to social facts: the fact of community approval or disapproval is identified with correctness or incorrectness. The actions on which the members of a community agree are not themselves liable to an assessment as being correct or incorrect. This position consequently is a social relativism. Its most extreme version is a radical meaning finitism that considers the extension of a concept only to be determined insofar as factual communal agreement reaches at a time (Bloor 1997, Chap. 2).

The other possibility is to conceive social holism in terms of symmetric I–thou relations between individuals assessing each other's actions as correct or incorrect. This conception does not have to identify

norms with social facts. Instead, it can regard these assessments as an open-ended process. Each normative attitude of taking something to be correct or incorrect, including those attitudes which are shared by all the members of a community at a time, can be challenged as not being correct. This position makes it possible to conceive the normative attitudes of taking something to be correct or incorrect as being responsive and responsible to the way the world is. It thus combines social holism with a realism (Pettit 1993, Chap. 4, Brandom 1994, Chaps. 1, 8).

If one intends to defend social atomism, one has to meet the challenge which Kripke poses. Two strategies can be distinguished. One strategy is to make a case for meanings as entities beyond the physical including persons and their interactions (Katz 1990, Chaps. 2–4). Having thoughts with a determinate conceptual content then does not depend on social relations, but on standing in an appropriate relation to such entities, i.e., grasping meanings. The other, most widespread strategy is to argue for some sort of a naturalization of semantics, i.e., an account of meaning in the vocabulary of the natural sciences. One then has to show how an account in terms of dispositions to behavior can solve both the infinity and the normativity problem which Kripke raises. Prominent proposals are an account in terms of computation (Miscevic 1996) and a teleofunctional account (Millikan 1990). According to these accounts, certain dispositions which each human being has considered in isolation from other human beings are sufficient for a human being to have thoughts with a determinate conceptual content—although, of course, social relations influence the way in which the thoughts of a person develop.

Social holism is a social externalism: the thoughts of a person are externally individuated, namely by the person's relations to her social environment. Social atomism, by contrast, is an internalism: the thoughts of a person are individuated by factors which are internal to that person such as her dispositions to behavior. Independently of Kripke's argument, Burge (1979) makes a case for social externalism by claiming that the meaning of the thoughts of a person depends on the way in which the terms in which that person expresses her thoughts are used in her social community. Burge's method is to develop thought experiments in which he keeps all the factors that are internal to a person fixed and varies the social environment. He thereby intends to show that the social environment enters into fixing the meaning of the thoughts of a person. Burge's argument fits into the method which dominates the discussion on externalism in general. A common reply to the externalist arguments is to distinguish between narrow and wide content of thoughts: the externalist argument is granted as far as wide content is concerned. But it is claimed that there is a narrow content of thoughts which is internally individuated and which is all that is needed to explain behavior. Thus, social atomism is vindicated as far as narrow content is concerned.

Social holism can be combined with semantic holism if one maintains that social relations determine a content for a thought of one type only by determining a content for other types of thoughts as well which constitute an inferential context for the thought in question (Brandom 1994). However, one does not have to combine semantic holism in form of an inferential semantics with social holism. The most widespread version of semantic holism is an inferential semantics applied to narrow content (Block 1986). Hence, a semantic holism can go with a social atomism. Moreover, subsequent to the influential criticism of semantic holism by Fodor and Lepore (1992), arguments for semantic atomism as well as semantic molecularism have gathered new momentum (Devitt 1995).

By contrast, holism with respect to the confirmation of thoughts or sentences is widely accepted owing to Quine (1951). Confirmation holism is one of the backgrounds of the view of science of Kuhn (1962), which is also influential in the social sciences. It is in dispute to what extent confirmation holism and Kuhn's view of science lend support to social relativism. Because of the presumed relativistic consequences, Quine (1969) rejects Kuhn's view of science and excludes observation sentences from both confirmation and semantic holism, defending atomism with respect to these sentences.

See also: Autonomy, Philosophy of; Collective Beliefs: Sociological Explanation; Collective Identity and Expressive Forms; Collectivism: Cultural Concerns; Colonialism, Anthropology of; Hegel and the Social Sciences; Hegel, Georg Wilhelm Friedrich (1770–1831); Hobbes, Thomas (1588–1679); Individualism versus Collectivism: Philosophical Aspects

Bibliography

Block N 1986 Advertisement for a semantics for psychology. In: French P A, Uehling T E. Jr., Wettstein H K (eds.) *Studies in the Philosophy of Mind. Midwest Studies in Philosophy*. University of Minnesota Press, Minneapolis, MN, Vol. 10, pp. 615–78

Bloor D 1997 *Wittgenstein, Rules and Institutions*. Routledge, London

Brandom R B 1994 *Making It Explicit. Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Cambridge, MA

Burge T 1979 Individualism and the mental. In: French P A, Uehling T E. Jr., Wettstein H K (eds.) *Studies in Metaphysics. Midwest Studies in Philosophy*. University of Minnesota Press, Minneapolis, MN, Vol. 4, pp. 73–121

Devitt M 1995 *Coming to Our Senses. A Naturalistic Program for Semantic Localism*. Cambridge University Press, Cambridge, UK

Durkheim E 1927 *Les règles de la méthode sociologique*. Alcan, Paris 8 [1938 The Rules of Sociological Method. Trans.

Solovay S A, Mueller J H]. University of Chicago Press, Chicago
Esfeld M 1998 Holism and analytic philosophy. *Mind* **107**: 365–80
Fodor J A, Lepore E 1992 *Holism. A Shopper's Guide*. Blackwell, Oxford, UK
Hegel G W F 1991 *Elements of the Philosophy of Right* (Trans. Nisbet H B). Cambridge University Press, Cambridge, UK
Herder J G 1993 *Against Pure Reason. Writings on Religion, Language, and History* (Trans. Bunge M). Fortress Press, Minneapolis, MN
Katz J J 1990 *The Metaphysics of Meaning*. MIT Press, Cambridge, MA
Kripke S A 1982 *Wittgenstein on Rules and Private Language: an elementary exposition*. Blackwell, Oxford, UK
Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
Mead G H 1934 *Mind, Self, and Society. From the Standpoint of a Social Behaviorist*. University of Chicago Press, Chicago
Millikan R G 1990 Truth rules, hoverflies, and the Kripke–Wittgenstein paradox. *Philosophical Review* **99**: 323–53
Miscevic N 1996 Computationalism and rule following. *Proceedings of the Aristotelian Society* **96**: 215–29
Pettit P 1993 *The Common Mind. An Essay on Psychology, Society, and Politics*. Oxford University Press, Oxford, UK
Philosophical Explorations 1 1998 Special Issue on Social Atomism and Holism
Quine W V O 1951 Two dogmas of empiricism. *Philosophical Review* **60**: 20–43
Quine W V O 1969 Epistemology naturalized. In: Quine W V O (ed.) *Ontological Relativity and Other Essays*. Columbia University Press, New York, pp. 69–90
Simons P M 1987 *Parts. A Study in Ontology*. Clarendon Press, Oxford, UK
Spencer V 1996 Towards an ontology of holistic individualism: Herder's theory of identity, culture and community. *History of European Ideas* **22**: 245–60
Teller P 1986 Relational holism and quantum mechanics. *British Journal for the Philosophy of Science* **37**: 71–81
von Bertalanffy L 1968 *General System Theory: Foundations, Development, Applications*. Braziller, New York

M. Esfeld

Attachment Theory: Psychological

Cross-cultural attachment research in Africa, China, Japan, Indonesia, and Israeli kibbutzim showed that attachments emerge inevitably between helpless offspring and protective parents. Similar antecedents of attachment security arise, and similar perceptions of the value of a secure attachment relationship seem to be prevalent. As attachment has also been observed in a wide variety of nonhuman primates and in other species, the bond between protective parents/caregivers and offspring is considered to be a universal phenomenon.

Although attachment is universal, three different organized attachment patterns have been detected in several hundreds of studies across the world. Infants who actively seek proximity to their caregivers upon reunion, and then readily return to exploration are classified as secure (B; 65 percent). Infants who ignore or avoid the caregiver following reunion are classified as insecure-avoidant (A; 20 percent). And infants who combine strong proximity with contact resistance and who are not able to return to play are classified insecure-ambivalent (C; 10 percent). Caregivers of secure infants are found to respond sensitively to their infants' attachment signals. Caregivers of insecure-avoidant infants have been found to reject their infants' negative emotions. Caregivers of insecure-ambivalent infants have been found to be inconsistently sensitive to their infants' attachment signals. There is growing evidence for the continuity as well as lawful discontinuity of these basic attachment patterns across the life span.

1. *The Concept of Attachment*

In a strange environment, infants become distressed when they are left alone by their parents for a brief period of time. Some infants will start crying, whereas other infants will move around to search for the absent parent. These responses to a brief separation are attachment behaviors and they indicate the attachment relationship that infants and parents usually establish during the first year after birth. The concept of attachment describes the enduring affective bond between children and their primary caregivers; from this bond children derive comfort in times of stress and distress. The attachment relationship enables them to explore new objects, persons, or environments and the attachment figure serves as a safe haven to which the child can return when exploration becomes too dangerous or stressful. Children can become attached to their (biological) mother and father, but also to other caregivers such as step-parents, babysitters, or professional caregivers in daycare centers. John Bowlby, British child psychiatrist, developed attachment theory to account for striking similarities in the ways in which children respond to major separations: with protest, despair, and detachment. The Second World War disrupted the life of many families, and many children became orphans. He sought to explain why children depend on their parents and caregivers, not only for food and shelter, but more importantly, for feelings of security and confidence.

1.1 *Historical, Cultural, and Biological Context*

Attachment is not an invention of modern times; attachment behavior is not restricted to modern, industrialized countries and cultures, and it is not an exclusively human characteristic. About 3,000 years ago, Homer had already described an example of infant attachment behavior to mother, father, and professional caregiver. In the famous *Iliad*, the Trojan

hero Hector returns from the battlefield to his family, smeared with mud and blood of the previous fights. Achilles is waiting for him to revenge earlier assaults on the Greek warriors, and he will defeat Hector in an atrocious fight. Hector wants to spend the remaining time with his wife, Andromache and his son, Skamander who is in the arms of his 'nanny.' Hector scares Skamander with his imposing helmet and with the traces of the earlier battle. In response to the reunion with Hector, Skamander seeks close physical contact with his nanny because he does not recognize his father. Hector notices the signs of fear that his son displays, takes off his helmet, and starts to play. His sensitive interactions stimulate Skamander to smile and play with his father; Andromache looks at this peaceful and moving scene in tears as she foresees the death of her husband in the fight with Achilles. After a while Hector carries his child to Andromache who cuddles him, and takes leave. Mother and son will never meet with their husband and father again.

The empirical study of attachment did not start in a Western culture. In fact, Ainsworth's first investigation of attachment was conducted in Uganda, a former British protectorate in East Africa. Here, she discovered the now-famous tripartite classification of insecure-avoidant, secure, and insecure-ambivalent attachment relationships (see later). It was in this African culture that she also studied, for the first time, the antecedents of attachment, in particular parental sensitive responsiveness. After this path-finding field-study of attachment, the development of attachment between parents and infants in several other African cultures has been studied, for example in the Gusii, the Dogon, the Hausa, the Efe, and the !Kung San. Furthermore, cross-cultural attachment research has been carried out in China, Japan, and Indonesia, and in the unique setting of the Israeli kibbutzim. In these diverging cultures attachments seem to emerge inevitably between helpless offspring and protective parents, similar patterns of associations between antecedents and sequelae of attachment security seem to arise, and similar perceptions of the value of a secure attachment relationship for the individual, the family, and for society seem to be prevalent. Cultural diversity in the expression of attachment is, however, acknowledged. In the Gusii culture, for example, infants are accustomed to being greeted by their returning mothers with a handshake instead of a hug, and accepting or refusing the handshake can be considered indications for the security of the attachment relationship in the Gusii dyad. The patterning of attachment behaviors is independent of the specific attachment behaviors that express the children's emotions, and these patterns of attachment have been found in all cultures studied thus far. Attachment, therefore, appears to be a universal phenomenon, and attachment theory claims universal validity (Van IJzendoorn and Sagi 1999, see also *Cultural Variations in Interpersonal Relationships*).

Attachment has been observed in many species, and John Bowlby drew heavily on the results of ethological investigations of attachment in nonhuman primates for the construction of attachment theory. Harlow's experiments with rhesus monkey infants showed that the development of attachments is not dependent on the provision of food and, thus, the result of reinforcement schedules, but seemed to be 'instinctive' behavior directed at objects providing warmth and protection. Field studies of rhesus monkeys showed that in the first few months after birth the infants develop enduring and unique bonds with the mother who protects the infants against predators and other environmental dangers. Rhesus monkey infants use their mother figure as a secure base to explore the environment, and to regulate their negative emotions and stresses. This role of a secure base is fulfilled uniquely by the mother and not by other members of the group or even by relatives. Bowlby derived from these ethological findings the idea that human infants were born with an innate bias to become attached to a protective adult because the affective bond would have evolutionary advantages (see also *Psychological Development: Ethological and Evolutionary Approaches*). Infants may differ in the expressions of their attachment 'needs,' and the patterns of their attachment behavior may diverge, but, basically, every representative of the human species would show attachment behaviors. In his later work he introduced the concept of 'inclusive fitness' into attachment theory to take into account the parental side of the first attachment relationship. It should be noted that nonhuman primates differ widely in the emergence and display of attachment relationships between infants and mothers. For example, capuchin monkey infants do not only rely on their mothers, but also on their peers and relatives for contact comfort and protection, and their bond with the mother seems much weaker than in the case of the rhesus monkey infants. The evolution of attachment theory might have been completely different if, in the 1950s and 1960s, ethologists had studied the social relationships of capuchin instead of rhesus monkey infants (Suomi 1999).

2. *Secure and Insecure Attachments*

In the development of attachment in young children four phases can be discriminated. In the first phase the baby is indiscriminately orienting and signaling to the environment, with a preference for human stimuli that seems to be most salient. Fairly soon the baby is able to fixate the eyes of caregivers, and begins to cry, smile, and grasp. It is unclear whether the baby is able to differentiate between people before 8–12 weeks of age. In the second phase the baby starts to orient and signal to a few specific caregivers for whom a preference is developed, presumably first through smell and sight. In this phase, attachment behaviors, such as crying,

can also be more easily terminated by caregivers instead of strangers. With the emergence of 'person-permanence' the infants enter into the third phase of attachment proper, in which they are able to discriminate the attachment figures from other persons, and address their attachment behaviors exclusively to these selected and preferred caregivers. They are now more actively seeking proximity and maintaining contact in stressful circumstances, and explore their environment from the safe base of the attachment figure. At around 6–8 months of age their attachment behaviors become organized in a goal-directed system, in which the set goal of proximity or 'felt security' is served by several different means of more distal or proximal behavior. In this phase, infants react to major separations in the manner that Bowlby described succinctly, with protest and anger, followed by despair and apathy, and finally detachment and the formation of new bonds. In the fourth phase the children learn to take the perspective of their attachment figures into account, and around three years of age they develop a goal-corrected partnership with their caregivers. In this phase, children construct a mental representation of their caregivers, the interactions and relationships with their attachment figures, and their expectations of caregivers' responses to their attachment signals.

Individual differences in quality of attachment relationships have been studied most intensively in the third phase of goal-directed development. In the Ainsworth Strange Situation Procedure, also termed Strange Situation (Ainsworth et al. 1987), infants in their second year of life are confronted with three stressful components, a strange environment, interaction with a stranger, and two brief separations from the caregiver. The first episode begins when the experimenter leads caregiver and child in the unfamiliar room and gives some last instructions.

The observations start when the caregiver brings the infant towards the pile of toys. The second episode is spent with the caregiver, together with the child, in the playroom. In the third episode an unfamiliar adult enters the room and, after a while, starts to play with the infant. In the fourth episode the caregiver leaves, and the infant is left with the stranger. In the fifth episode the caregiver returns after about three minutes of separation. Episode 6 starts when the caregiver leaves again and the infant is alone in the room. In the seventh episode the stranger returns, and in the eighth episode the caregiver also returns. The compound stress of this procedure has been proven to elicit attachment behaviors, and the patterning of this behavior, in particular in the reunion episodes, is taken as evidence for the quality of the attachment relationship with the caregiver.

Three organized attachment patterns have been found in several thousands of Strange Situation Procedures across the world. Infants who actively seek proximity to their caregivers upon reunion, communicate their feelings of stress and distress openly, and then readily return to exploration, are classified as secure (B) in their attachment to that caregiver. Infants who seem not to be distressed and ignore or avoid the caregiver following reunion (although physiological research shows that their arousal during separation is similar to other infants), are classified as insecure-avoidant (A). Infants who combine strong proximity seeking and contact maintaining with contact resistance, or remain inconsolable, without being able to return to play and explore the environment, are classified insecure-ambivalent (C). About two thirds of the infant–caregiver dyads are classified secure, one fifth are classified as insecure-avoidant, and about one-tenth are classified insecure-ambivalent. Caregivers of secure infants are found to respond sensitively to their infants' attachment signals, and the infants trust their availability in times of stress. Caregivers of insecure-avoidant infants have been found to reject or ignore their infants' communications of negative emotions, and the avoidant infants seem to minimize the expression of distress.

Caregivers of insecure-ambivalent infants have been found to be inconsistently sensitive to their infants' attachment signals, and the ambivalent infants appear to maximize the expression of negative emotions in order to draw their caregiver's attention. In numerous correlational and experimental studies on normal, nonclinical dyads this causal association between caregiver sensitivity and infant attachment security has been established.

During the past few decades attachment research has been broadened to include clinical groups, that is, samples with physically or psychologically impaired and disturbed children or caregivers. In particular, in abusive families, the traditional tripartite attachment classification system appeared to reach its limits, and many abused infants had to be classified secure although their parents were known to have abused the specific child. Main and her co-workers examined cases that were difficult to classify within the traditional Ainsworth system, and they developed the concept of 'disorganized' attachment, defined as the momentary breakdown of an otherwise organized attachment strategy to deal with stress. Examples of disorganized attachment behavior are contradictory behavior (e.g., at the same time seeking proximity to as well as avoiding the attachment figure), misdirected attachment behavior (e.g., seeking proximity to the stranger instead of the parent on reunion), stereotypical behavior (e.g., hair pulling with a dazed expression during stress and in presence of the caregiver), freezing and stilling (e.g., stopping all behavior as if in a trance or unable to choose between behavioral alternatives), and direct apprehension or even fear of the caregiver upon reunion. Using the disorganized attachment classification in abusive families, almost 80 percent of the dyads had to be classified as disorganized. In several studies disorganized children

have been shown to manage or regulate their physiological stress (measured by cortisol levels and heart rate) less well than children with organized (secure or insecure) attachments, and, in the long run, these infants seem to be at risk for psychopathology, in particular externalizing problems. Disorganized attachment is somewhat similar to the behavior described in the Reactive Attachment Disorders of the *DSM-IV* of the American Psychiatric Association, and similar pathogenic care is supposed to be the cause of disorganized attachment: abuse and neglect, as well as frequent changes of caregivers and, thus, major separations from attachment figures (see also *Early Childhood: Socioemotional Risks*).

Abusive parents create a paradox that their young children cannot resolve. Abusive caregivers are, at the same time, a source of fright and the only potential haven of safety in threatening situations. The combination of frightening parental behavior and the parental figure as a focus of attachment constitutes a paradox that the children respond to with contradictory, misdirected, or otherwise disorganized behaviors that bear some resemblance to dissociative states in older children and adults. In fact, some studies have found associations between disorganized attachment in infancy and elevated dissociative tendencies in adolescence. Main and her colleagues argued that frightening parental behavior does not only occur in abusive parents, but also in caregivers who suffer from unresolved loss or other trauma and cannot avoid entering into somewhat dissociative states themselves, even in presence of their children. The children would perceive this parenting as frightening and incomprehensible. In several studies with the, so-called, Adult Attachment Interview of Main and her co-workers, a rather strong association between unresolved loss or trauma and disorganized infant attachment has been found (Van IJzendoorn 1995). Furthermore, in at least three studies the hypothesis of the mechanism of frightening parental behavior linking unresolved loss to disorganized attachment has been confirmed.

3. *Continuity of Attachment Across the Life Span?*

The Adult Attachment Interview not only contributed to the investigation of antecedents of disorganized attachment, but also represented a 'move to the level of representations' and a broadening of attachment research into the domain of adolescence and adulthood (see also *Adult Psychological Development: Attachment*). Whereas the Strange Situation dominated attachment research in the 1970s and 1980s, the Adult Attachment Interview emerged as one of the most important attachment measures in the 1990s, and it facilitated the investigation of continuities and discontinuities in the development of attachment from infancy to adulthood. The Adult Attachment Interview is a semistructured, hour-long interview with open-ended questions that revolve around issues of attachment, separation, and loss during childhood and adulthood. In fact, like the Strange Situation procedure, the interview constitutes a dual task. On the one hand the participants must focus on their attachment experiences—which, in the case of bad experiences, may sometimes be rather uncomfortable. On the other hand, while contemplating their past the participants should keep focused on and follow the rules of discourse. Insecure adults, in particular, are not able to complete this dual task successfully. They remain too much focused either on the discourse context ('dismissing' subjects) or on the past experiences ('preoccupied' participants). Only the secure subjects are able to keep a balance between focus on the present discourse and the past experiences, even when they were treated badly in their childhood. The Adult Attachment Interview can be considered a stressful situation for adults in the way that the Strange Situation is for young children, and the struggle for balance between discourse context and autobiographical content runs parallel to the balance children have to strike between attachment and exploration in the Strange Situation. In fact, Main and her coworkers developed the Adult Attachment Interview coding system with the assumption that adult attachment classifications should map on corresponding Strange Situation classifications. In subsequent independent replication studies, this assumption has been confirmed (Van IJzendoorn 1995).

In several longitudinal studies continuities as well as discontinuities in the development of attachment from infancy to adulthood have been established. In stable (negative or positive) child rearing arrangements the likelihood of continuity is enhanced, and attachment security in adolescence or early adulthood appears to be almost perfectly predictable on the basis of assessment of attachment security in infancy (see, for example, Beckwith et al. 1999, Waters et al. 1995). If major life events, such as parental divorce, occur during the first seven years of life this will not only disrupt family life but also the continuity of attachment in the children. Lawful discontinuity has been established in several longitudinal studies in which disruptive life events were linked to discontinuities in attachment security. In unstable child rearing circumstances Sroufe and his colleagues were not able to detect any continuity of attachment security across the first 16 years of life, although attachment disorganisation in infancy was highly predictive of psychopathology and dissociative tendencies in adolescence (Ogawa et al. 1997). More longitudinal studies will come of age in the next few years (Grossmann and Grossmann 1991), but the available evidence suggests that attachment experiences in infancy do not leave permanent scars or give robust protection against adverse events later in life. The concept of a critical

attachment period in the first year or two has become obsolete. Attachment security seems to be environmentally labile until about the age of five. Thereafter, the development of attachment security becomes more and more stable because mental representations of earlier attachment experiences influence the perception and evaluation of later experiences and tend to become self-fulfilling prophecies. Nevertheless, environmental discontinuities may leave their traces on attachment at any major transition during the life cycle—for the better or the worse. Future, longitudinal-attachment studies should focus more carefully on repeated age-appropriate assessments of attachment, as well as on detailed descriptions of relevant environmental developments. Continuity and lawful discontinuity of attachment can only be found in the transactions between individual development and the changing environment (Sameroff 1997).

Bibliography

Ainsworth M D S, Blehar M C, Waters E, Wall S 1987 *Patterns of Attachment: a Bychological Study of the Strange Situation. Lawrence*. Erlbaum, Associates, Hillsdale, NJ

Beckwith L, Cohen S E, Hamilton C E 1999 Maternal sensitivity during infancy and subsequent life events relate to attachment representation at early adulthood. *Developmental Psychology* **35**: 693–700

Bowlby J 1988 *A Secure Base: Parent–Child Attachment and Healthy Human Development*. Basic Books, Inc., New York

Grossmann K E, Grossmann K 1991 Attachment quality as an organizer of emotional and behavioral responses in a longitudinal perspective. In: Parkes C M, Stevenson-Hinde J, Marris P (eds.) *Attachment Across the Life Cycle*. Tavistock, Routledge, London, pp. 93–114

Main M 1990 Cross-cultural studies of attachment organization: Recent studies, changing methodologies, and the concept of conditional strategies. *Human Development* **33**: 48–61

Main M, Kaplan N, Cassidy J 1985 Security in infancy, childhood, and adulthood: a move to the level of representation. In: Bretherton I, Waters E (eds.) *Growing Points of Attachment Theory and Research*. Society for Research in Child Development, pp. 66–104

Sameroff A J 1997 Developmental contributions to the study of psychopathology. Master lecture presented at the biennial meeting of the Society for Research in Child Development, Washington, DC, April, 1997

Ogawa J R, Sroufe L A, Weinfield N S, Carlson E A, Egeland B 1997 Development and the fragmented self: Longitudinal study of dissociative symptomatology in a nonclinical sample. *Development and Psychopathology* **9**: 855–79

Suomi S J 1999 Attachment in Rhesus monkeys. In: Cassidy J, Shaver P R (eds.) *Handbook of Attachment. Theory, Research, and Clinical Applications*. Guilford Press, New York, pp. 181–97

Van IJzendoorn M H 1995 Adult attachment representations, parental responsiveness, and infant attachment: A meta-analysis on the predictive validity of the Adult Attachment Interview. *Psychological Bulletin* **117**: 387–403

Van IJzendoorn M H, Sagi A 1999 Cross-cultural patterns of attachment: universal and contextual dimensions. In: Cassidy J, Shaver P R (eds.) *Handbook of Attachment Theory, Research and Clinical Applications*. Guilford Press, New York, pp. 713–35

Waters E, Merrick S K, Albersheim L J, Treboux D 1995 Attachment security from infancy to early adulthood: A 20-year longitudinal study. Paper presented at the biennial meeting of the Society for Research in Child Development, Indianapolis, IN

M. H. van IJzendoorn

Attention and Action

1. Attention

Within information processing psychology, the term 'attention' refers to a mechanism that selects a spatially coherent subset of sensory information from among all information available. With regard to the function of attention two different theoretical camps can be distinguished: a major perception camp and a minor action camp.

2. Perception

Theorists in the perception camp, for instance Broadbent (1958) and Treisman (1988), answer the question 'What is (visual) perception for?' with the answer: 'For (visual) perception!' From this answer it immediately follows that the study of 'Visual cognition, then, deals with the processes by which a perceived, remembered, and thought-about world is brought into being from as unpromising a beginning as the retinal patterns' (Neisser 1967, p. 4). Just because perception is for perception, these are the processes the experimental research has to be concerned with.

After Broadbent (1958), the basic assumption in this line of theorizing is that the human information processing system has severe central capacity limitations. Perception is construed as a two-stage process. In the first, 'preattentive,' stage all information receives a preliminary, superficial evaluation. In the second, 'attentive,' stage, only a part of that preprocessed information is selected and subjected to a definitive, complete interrogation. Limited central capacity is regarded as the basic functional characteristic of the human information processor and attentional selection as a secondary functional consequence.

There are two basic reasons to doubt the limited capacity assumption, one theoretical and one empirical. The theoretical reason is that in this theorizing an observed phenomenon (people show overt performance limitations) is simply translated into an explanatory construct (people suffer internal capacity

limitations). The empirical reason is that behavioral experiments have not convincingly demonstrated these central capacity limitations. Moreover, also neurophysiology and neuroanatomy do not support the assumption of insufficient computational power in the brain (van der Heijden and Bem 1997).

3. Action

Theorists in the action camp, Allport (1987, 1989) and Neumann (1987, 1990) for instance, answer the question 'What is visual perception for?' with the answer: 'For (control of) action!' In their view, visual perceptual systems, just as all other '… perceptual systems have evolved in all species of animals solely as a means of guiding and controlling action … ' (Allport 1987 p. 395). Two control problems are recognized: the problem of effector recruitment (effectors can be used for different actions and one of the actions has to be selected) and the problem of parameter specification (a selected action can be executed in a number of ways and one of these has to be selected).

In correspondence with these two control problems two selection mechanisms are distinguished. The first is in charge of determining which *action* out of the total repertoire of actions is given priority at a certain moment in time (Allport 1987, p. 395). It determines which skill is allowed to recruit what effectors now (Neumann 1987, p. 376). The second, attention, is in charge of determining which *object* is acted upon (Allport 1987, p. 395). It determines from what region in space the parameters are taken that are allowed to specify a selected action in detail (Neumann 1987, p. 376). Selection for action is regarded as the basic functional characteristic of the human information processor and overt behavioral limitations as a functional consequence.

The perception for action approach presents a convincing functional analysis. Especially important is the idea that the control structures that serve to guide the selected actions, e.g., walking, throwing, or catching, have built-in selective capacities. These control structures determine and extract the parameters that specify the selected actions in detail. Therefore, besides 'action selection' and 'object selection,' no additional 'parameter selection' is required.

4. Impact

The perception for action approach has not really influenced the theorizing in information processing psychology. The prime reason for this lack of impact has to be found in the fact that the mainstream perception for perception approach is not very much helped by the *theorizing* within the perception for action approach. The perception for perception approach is not concerned with the behavior that animals and humans have in common; with how they forage for food, avoid predators, find mates, and move efficiently around. The experiments performed within this camp capitalize nearly always upon a uniquely human capability: the use of language. Subjects have to name colors or positions or to read letters or words.

5. Language

For the language behavior, used in the perception for perception camp, the analysis in terms of 'action selection' and 'object selection' of the perception for action camp still needs to be further developed. For the actions that humans share with animals the action control structures might always determine and extract the required set of action parameters from a selected object. This, however, is generally not the case with language behavior, i.e., with 'the act of speaking.' This is even not the case when in 'the act of speaking' a distinction is made between 'the act of naming' and 'the act of reading.' Two different problems have to be considered.

First, 'the act of naming' cannot select the required action parameters. When the selected object is, for instance, a small, red, triangle at the left, 'the act of naming' insufficiently constrains the verbal behavior that has to occur. The size, color, form, position and many more properties can be named. The introduction of 'size naming,' 'color naming,' 'form naming,' and 'position naming' as independent actions with devoted control structures is not a solution. That deprives the perception for action theory from its parsimony and elegance.

Second, 'the act of reading' cannot select all required action parameters. When the selected object is, for instance, the capital letter A, 'the act of reading' insufficiently constrains the verbal behavior that has to occur. Subjects need not respond with 'a.' They can refrain from responding; an adequate reaction under the instruction: say 'o' when there is an O and refrain from responding when there is an A. They can also respond with 'yes' and 'no'; adequate reactions under the instructions: say 'yes' ('no') when there is a vowel (consonant) and 'no' ('yes') when there is a consonant (vowel). Subjects can also respond with 'one'; an adequate reaction under the instruction: name the serial position of the letter in the alphabet. They can also respond with 'ape' or 'cat'; adequate reactions under the instruction: give an animal name that begins (not) with an A. Besides the selected object more than just 'the act of reading' is required for explaining a subject's task performance.

So, perception for action theorists have failed to generalize their view in such a way that also 'the act of speaking,' as encountered in the experiments of the perception for perception camp, can be dealt with.

6. Two Factors

From the perception for action approach nevertheless an important lesson follows. That lesson is that the perception for perception camp is not performing the *experiments* they think they perform and are not providing the *explanations* they have to provide. For recognizing this, it is important to see that in all experiments two experimental 'factors' are operative.

The first experimental 'factor,' that is generally acknowledged, consists of the stimuli that are presented in an experiment. The second experimental 'factor,' that is not generally acknowledged, consists of the instruction to act in a specified way. With their emphasis on 'action selection' and 'object selection,' the perception for action theorists make clear that in the laboratory experiments the instruction, for covert and overt 'action selection,' and the stimuli, for 'object selection,' are of equal importance. The instruction specifies the task to be performed and the action to be executed. The stimuli provide the objects to act upon and to specify the required action in detail. Of course, no experiment will work and produce interpretable data without one or another kind of stimuli and without one or another kind of instruction.

7. The Instruction

Within the perception for perception camp the importance of the instruction is not recognized. Reasons for this state of affairs are not too difficult to find. The perception investigators are almost exclusively interested in the bottom-up effect of the stimulus, in how it is processed and perceived despite the limited capacity. Moreover, in line with this preoccupation, in nearly all experiments, the stimulus is varied from trial to trial and produces an interesting pattern of results that invites a detailed explanation, but the instruction remains the same and can play no role in that explanation (see van der Heijden 1992).

But, recognized or not, in all experiments there is always an instruction to perform a task and to act in a specified way. So, not perception as such is investigated, but task performance involving or using perception. By failing to recognize this point and by overseeing the importance of the instruction as the factor that induces the covert and overt task performance, a discrepancy arises between the issues *explicitly addressed* (the processing of the stimulus information) and the processes *actually involved* (the instruction-induced task performance) (see Neumann 1990, p. 232).

8. The Action Plan

The important idea of the perception for action approach is that a selected action has the causal power to select the subset of parameters that is required to specify that action in detail. It is not difficult to see how that important idea can be generalized to account for the observed (language or symbolic) behavior in standard information processing experiments. Whether fixed or varied, in each information processing experiment there *is* always an instruction. The internal representation of that instruction has the causal powers that make that the subjects perform according to the instruction. That internal representation can be called an 'action plan' (Neumann 1987, p. 381). This instruction-induced 'action plan' selects its parameters in the experiments of the perception for perception camp just as the selected 'action' selects its parameters in the tasks considered in the perception for action camp.

In current theorizing in the information processing approach, the important theoretical issues with regard to 'action plans'—how they are derived from instructions, how they are structured, how they are implemented in the information processing system, and how they can exert their causal effects—are generally neglected. Therefore, the theoretical accounts are incomplete. Indeed, the internal representation of a fixed instruction cannot contribute to the explanation of the interesting *details* of the results obtained within an experiment. Without the instruction-induced action plan with its causal powers, however, the results that exhibit that interesting pattern had never been produced in the first place.

See also: Attention: Multiple Resources; Attention, Neural Basis of; Motivation and Actions, Psychology of; Perception and Action

Bibliography

Allport D A 1987 Selection for action: Some behavioral and neurophysiological considerations of attention and action. In: Heuer H, Sanders A F (eds.) *Perspectives on Perception and Action*. Erlbaum Associates, Hillsdale, NJ

Allport D A 1989 Visual attention. In: Posner M I (ed.) *Foundations of Cognitive Science*. MIT Press, Cambridge, MA

Broadbent D E 1958 *Perception and Communication*. Pergamon, London

Neisser U 1967 *Cognitive Psychology*. Appleton-Century-Crofts, New York

Neumann O 1987 Beyond capacity: A functional view of attention. In: Heuer H, Sanders A F (eds.) *Perspectives on Perception and Action*. Erlbaum Associates, Hillsdale, NJ

Neumann O 1990 Visual attention and action. In: Neumann O, Prinz W (eds.) *Relationships between Perception and Action*. Springer Verlag, Berlin

Treisman A M 1988 Features and objects: The fourteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology* **40-A**: 201–37

van der Heijden A H C 1992 *Selective Attention in Vision*. Routledge, London; New York

van der Heijden A H C, Bem S 1997 Successive approximations to an adequate model of attention. *Consciousness and Cognition* **6**: 413–28

A. H. C. van der Heijden

Attention-deficit/Hyperactivity Disorder (ADHD)

Attention-deficit/hyperactivity disorder (ADHD) brings to mind high energy levels, rapid reactions, and eagerness to explore novel environments. One also thinks of unfinished tasks, distractibility and disorganization, rule-breaking and risk-taking, disruptions and intrusions. Whether ADHD is considered a clinical syndrome, a cluster of behavior problems, or a dubious social construction, the characteristics are well known. Behaviors typical of ADHD are exhibited by most people on occasion, and by most children on many occasions, suggesting that people with ADHD differ more in degree than in kind from their nondiagnosed counterparts. Children are the most likely candidates for a diagnosis of ADHD, but increasing numbers of adolescents and adults are being referred for similar problems.

1. The Changing Faces of ADHD: Characterizing the Disorder

ADHD can be construed as a disorder of context, timing, and modulation. One of the biggest puzzles about children with ADHD is the number of things they can do well. They often appear to have mastered the requisite competencies, be they academic or social. Their problems emerge not as knowledge or skill deficits, but rather as application failures or mismatches between acts and settings. The appropriate time to ask the teacher what you're supposed to do is not immediately after she has given complete instructions. The optimal opportunity for practicing karate moves is not during a spelling test. The curiosity that leads a child to unwind rolls of toilet paper down the staircase is better expressed when the mother is not hosting a dinner for her employer.

1.1 Analyzing Performance: 'Can't Do' vs. 'Won't Do'

Ongoing research is identifying two focal problem areas: integrating information over time and context, and sustaining effortful problem solving. These difficulties are exacerbated when the task requires prolonged inhibitory control or the withholding of preferred actions. Children with ADHD spend less time than their peers observing task-relevant features, especially when engaging distractors are present. Tasks that require shifting from one dimension or response to another and back again on cue pose particular challenges. Also problematic are tasks that demand cognitive transformations such as integrating new data with knowledge stored in memory, or inferring causal connections between sequences of events. The metaphor presented by Milch-Reich et al. (1999) is particularly fitting. If the active processing of incoming information by most children can be likened to a cohesive, unfolding film, the mental transformations made by children with ADHD resemble a discrete and disorganized slide presentation.

These propensities have been well documented in task contexts and are also surfacing in the social arena. It seems reasonable to infer that youngsters who have difficulties organizing information over time, drawing causal connections, and shifting response sets on cue may also have problems forming accurate social impressions, understanding people's motivations, and engaging in reciprocal interactions. Children with ADHD show relatively low rates of effortful attention in social contexts; they spend little time observing how other people interact. They often seem oblivious to their impact on others and mystified when their actions provoke irritation or avoidance. Sadly, it is often the children with the strongest social interests who engender the most conflict and rejection.

Further research is required to evaluate the extent to which these information-processing differences are attributable to an impulsive response style, problems allocating attention efficiently, failures to sustain mental effort, or the pursuit of task-irrelevant goals. Distinguishing the 'can't do' from the 'won't do' facets is a continuing challenge.

1.2 Three Myths About Performance Patterns

1.2.1 Myth 1: Unpredictability. The putative unpredictability of children with ADHD is often predictable. In many situations, consistency rather than variability distinguishes them from their peers. When the task or role requirements change, when they need to stop one behavior and start another, children with ADHD often have difficulties. On cognitive tasks they tend not to slow down after an error or speed up when instructed to do so, and they have problems adjusting to a new rule or dimension (Douglas 1999). This difficulty modulating behavior has also been documented in social contexts that involve shifting from the role of leader to follower or from talk show host to guest (Whalen and Henker 1998). Thus despite the fact that spontaneous change seems to be a hallmark of ADHD, change-on-cue is problematic. These findings illustrate the hazards of interpreting

behaviors without considering contexts and task demands.

1.2.2 Myth 2: Rapid reactivity. Another counter-intuitive aspect of ADHD concerns response speed. The reputation for impulsivity is based on quick reactions that occur without reflecting on options and consequences. But on some cognitive tasks, especially those requiring sustained attention, reaction times are sluggish and variable, resulting in impaired performance. The timing problems appear to be another example of difficulties sustaining effortful control and adapting to shifting task demands rather than a global tendency to respond with undue speed.

1.2.3 Myth 3: Insensitivity to consequences. The repeated failures of children with ADHD have led many to infer a certain imperviousness to consequences. Recent studies do indeed implicate disturbed reward sensitivity, but they suggest hypersensitivity rather than a cavalier attitude toward reinforcement. It may be that children with ADHD actually care too much, becoming distracted by reinforcers and frustrated when they are delayed. Thus, the partial reinforcement schedules that so effectively maintain performance in people without ADHD may disrupt performance in those with the disorder.

1.3 A Core Deficit?

Despite extensive efforts and continuing progress in identifying a core deficit (e.g., Barkley 1998, Douglas 1999, Sergeant et al. 1999), consensus has not yet been achieved. Early conceptualizations focused on motor activity, and children with hyperactivity were described as perpetual motion machines, always on the go, acting as if driven by a motor. In the 1970s and 1980s, the focus shifted to attentional mechanisms, the abilities to focus selectively and continuously while resisting distraction. More recent formulations involve deficits in self-regulation and inhibitory competence, executive functions that involve the ability to modulate one's own responding in accord with varying cues and conventions.

1.4 Diagnostic Nomenclature, Criteria, and Subtypes

Variations in diagnostic labels and criteria have paralleled these changes in conceptualization and research emphasis. Over the second half of the twentieth century, four modifications of diagnostic labels and criteria appeared in successive versions of the *Diagnostic and Statistical Manual* (*DSM*) of the American Psychiatric Association (see *Mental and Behavioral Disorders, Diagnosis and Classification of*). In the current version (*DSM-IV*, American Psychiatric Association 1994), an ADHD diagnosis requires problems with either inattention (ADHD-I), hyperactivity and impulsivity (ADHD-HI), or the two combined (ADHD-C). ADHD-C is the most common and ADHD-HI the least common subtype. Children who meet the criteria for ADHD-I have higher rates of cognitive-academic and lower rates of social-behavioral problems and are also more likely to be girls than are those in the other two subgroups. Some specialists even question whether ADHD-I is a legitimate subtype or a distinctive disorder. The face of ADHD will probably change in subsequent revisions of *DSM-IV*. The frequent changes in diagnostic schemas attest to the continuing uncertainties about the nature and boundaries of ADHD as well as the heterogeneity of people given this label.

2. ADHD: Some Knowns and Unknowns

Of all childhood psychological dysfunctions, ADHD has been studied the most extensively and rigorously. A selection of fundamental findings that have withstood the tests of time and replication follows, laced with some remaining questions.

2.1 Nature and Nurture

Although there are no known biological markers for ADHD, multiple strands of evidence converge on the conclusion that most cases have a genetic origin. ADHD runs in families, and the behavioral characteristics as well as the clinical conditions show strong heritabilities ranging from 60 to 90 percent. Neuroimaging studies reveal subtle reductions in size and activity of several brain regions, including the prefrontal cortex, basal ganglia, and cerebellum. Neurochemical studies suggest a dysregulation of the catecholamines, especially dopamine, and genetic studies are identifying aberrant candidate genes within the dopamine system. Consistent with these findings is the fact that the most effective drugs for treating ADHD, the stimulants methylphenidate (Ritalin) and dextroamphetamine (Dexedrine), alter catecholamine functioning. Other biological systems have also been implicated, however, and inconsistencies across studies abound. Despite the complexities, rapid developments in neurobiological models and methods promise continuing elucidation of the pathophysiologies of ADHD (Castellanos 1999, Solanto 1998, Swanson et al. 1998, Tannock 1998).

The strong biological evidence does not negate the role of psychosocial and environmental contributions. Psychosocial factors influence the severity, course, and outcome of ADHD and may also modulate treatment response. In some cases, psychosocial adversity or insults such as environmental lead or maternal smoking during fetal development may also play an etiological role.

2.2 *Prevalence, Gender, and Culture*

ADHD is a common childhood disorder. The best prevalence estimate is 3–5 percent of school-age children, although rates range from less than 2 percent to over 20 percent across studies, depending on definitions and sampling. During middle childhood, boys outnumber girls at ratios of four or five to one, but the gender gap decreases with increasing age and may even reverse in adulthood. The changes in diagnostic criteria over successive versions of the DSM have been accompanied by increasing prevalence rates, especially in girls. ADHD has been found in all countries and cultures studied to date, with wide variations across geographical regions and diagnostic practices.

2.3 *Heterogeneity and Comorbidity*

People with ADHD are highly heterogeneous, showing markedly different profiles of strengths and deficiencies. Childhood ADHD typically coexists with one or more other disorders, most frequently oppositional defiant or conduct disorder. Anxiety disorders, depression, and learning problems may also accompany ADHD. In fact, a 'pure' diagnosis of ADHD tends to be the exception rather than the rule. Defining the boundaries of ADHD and identifying the subtypes continue to be controversial enterprises. Some specialists assert that there should be additional subtypes of ADHD, whereas others contend that the high rates of comorbidity raise doubts about the very validity of the syndrome.

2.4 *Pervasiveness and Severity*

Problems associated with ADHD surface across settings and affect multiple domains of everyday life. There are difficulties completing schoolwork, responding to directions, interacting harmoniously, adjusting behaviors in accord with changing settings and activities, solving the minor problems that surface during any social exchange, and managing affect, motivation, and self-esteem. The current version of the *DSM* recognizes the extensity of the problems by requiring that difficulties be observed in at least two settings and be sufficiently severe to impair functioning. Important areas for future study include a functional taxonomy of situations and delineation of settings that provoke, and those that diminish, the problems associated with ADHD.

2.5 *Developmental Course and Outcome*

ADHD is a chronic disorder. Although many children improve over the years, at least half continue to have difficulties in everyday activities and relationships, with problems changing in form if not in severity. In the transition to adolescence and adulthood, restlessness may replace hyperactivity, frequent job changes may replace frenetic shifts of activities, and argumentativeness or irritability may replace physical aggression and social disruption. Even when these problems no longer reach diagnostic thresholds, they continue to compromise overall functioning. Compared to peers and siblings, many people diagnosed with ADHD do not progress as far in school, obtain lower-ranked jobs and poorer employee evaluations, are involved in more accidents and receive more traffic citations, have more encounters with law enforcement, and show less engagement in constructive activities. Supportive and mutually gratifying social relationships continue to elude them.

The perennial search for sturdy predictors of long-term adjustment continues to yield disappointing results. Early psychosocial adversity and aggressive behavior portend a rocky developmental course, but specific prognostic indicators have not emerged. In need of systematic study is the role of intervening events in influencing outcomes. Does some of the continuity of ADHD result from avoidable consequences of problem behaviors, for example, responding to social rejection by selecting a deviant partner or peers who in turn reinforce dysfunctional responding (Taylor 1998)?

2.6 *Assessment*

There is no definitive diagnostic test of ADHD, and problem behaviors may not be observable in the doctor's office. The most useful information is obtained from parents and teachers, using standard behavioral ratings. Systematic interviews and classroom observations are also informative but often infeasible. Medical tests cannot confirm ADHD but may be useful in ruling out medical disorders that may mimic ADHD symptoms. Self-report is of limited help during the school-age years but adds an important dimension during adolescence and adulthood.

2.7 *Treatment*

The therapeutic armamentarium for ADHD is extensive, ranging from traditional behavioral, cognitive, and psychosocial therapies to unproven ap-

proaches such as biofeedback, megavitamins, exercise, sensorimotor integration, and low-sugar or additive-free diets. Parenting programs, social skills training, and educational interventions are often helpful. The most effective single treatment continues to be one that has been used since the middle of the twentieth century: stimulant pharmacotherapy (Whalen and Henker 1997). The immediate and often dramatic short-term improvements seen with stimulant medications have been documented repeatedly, and newer medications such as Adderall are dealing with some of the limitations of standard stimulants (e.g., short span of effectiveness). There is still very little evidence of enduring effects, however, either in academic/occupational or social arenas. Research is needed on modes of transforming the welcome short-term gains into long-term improvements.

2.8 The Social Ecology of ADHD and Stimulant Pharmacotherapy

ADHD affects not only individuals but also the surrounding social environment. In the presence of a child with ADHD, peers, parents, and teachers become more negative and controlling. When a child with ADHD improves on medication, other people become more positive and interactions more harmonious. These interpersonal ripple effects have significant intervention implications.

2.9 Adults with ADHD

Recent years have witnessed a surge of interest in adults with ADHD, many of whom are first diagnosed when they take their children for evaluation of ADHD. The diagnosis in adults is especially difficult to substantiate, however. Not enough is known about appropriate measures and cut-off scores, and many of the behavioral symptoms consistent with ADHD are found in other syndromes such as antisocial and substance use disorders. Concerns have been raised that some adults may actively seek a diagnosis of ADHD, perhaps because this diagnosis is more benign than other labels, or perhaps because it yields special considerations (e.g., extra time to complete certification exams; accommodations by employers) or prescriptions for controlled drugs.

2.10 Links Between ADHD and Cigarette Smoking

Mothers who smoke during pregnancy are at elevated risk for having a child who develops ADHD. Adolescents with ADHD are more likely than their peers to smoke cigarettes; they initiate these behaviors at younger ages and, as adults, have more difficulty quitting. These findings, along with indications that nicotine enhances attention and cognitive performance and reduces anger, raise important questions about whether people with ADHD may smoke as a form of self-medication. Also intriguing are converging lines of evidence implicating genes involved in dopamine regulation in both ADHD and cigarette smoking.

3. Continuing Controversies and Future Directions

Ironically, ADHD is one of the most carefully studied childhood disorders yet remains one of the most controversial. Ongoing and at times heated debates have an impact not only on how society deals with ADHD, but also on the research agenda for childhood disorders and treatments more generally. A few of the most salient controversies are:

(a) Is ADHD a socially constructed pathology imposed upon healthy children who are required to function in unsuitable environments?

(b) Are the increasing rates of ADHD diagnoses and stimulant prescriptions cause for serious concern?

(c) Are there distinctive subtypes of ADHD, with differing causes, concomitants, and courses?

(d) Does ADHD take different forms in girls than in boys? Are girls with ADHD understudied and underdiagnosed?

(e) To what extent can behavioral and psychosocial treatments obviate the need for stimulant medications, decrease the required dosage or duration, or enhance long-term outcomes?

(f) Is normalization a feasible treatment goal?

(g) Even if feasible, is normalization desirable? Or, are there contexts in which some of the characteristics of people with ADHD may prove valuable or adaptive?

There is no question that ADHD is a pervasive, perplexing, and persistent syndrome. This family of disorders has served a catalytic function, generating sophisticated procedures and instruments for studying fundamental psychobiological processes, including attentional mechanisms, self-regulation, peer interaction, developmental continuity, and risk and protective factors. Stimulant drugs have been used effectively as pharmacological probes that have yielded increased understanding of brain–behavior relationships in areas that extend far beyond ADHD. Future years should see clearer delineation of alternative developmental pathways that lead to and away from ADHD, increased understanding of how environmental factors modulate biological vulnerabilities, and systematic testing of multimodal treatment packages tailored to specific ADHD subtypes.

See also: Behavior Therapy with Children

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV* 4th edn. American Psychiatric Association, Washington, DC

Barkley R A 1998 Attention-deficit hyperactivity disorder. *Scientific American* **279**: 66–71

Castellanos F X 1999 The psychobiology of attention-deficit/hyperactivity disorder. In: Quay H C, Hogan A E (eds.) *Handbook of Disruptive Behavior Disorders*. Plenum, New York, pp. 179–98

Douglas V I 1999 Cognitive control processes in attention-deficit/hyperactivity disorder. In: Quay H C, Hogan A E (eds.) *Handbook of Disruptive Behavior Disorders*. Plenum, New York, pp. 105–38

Milch-Reich S, Campbell S B, Pelham W E, Connelly L M, Geva D 1999 Developmental and individual differences in children's on-line representations of dynamic social events. *Child Development* **70**: 413–31

Sergeant J A, Oosterlaan J, van der Meere J 1999 Information processing and energetic factors in attention-deficit/hyperactivity disorder. In: Quay H C, Hogan A E (eds.) *Handbook of Disruptive Behavior Disorders*. Plenum, New York, pp. 75–104

Solanto M V 1998 Neuropsychopharmacological mechanisms of stimulant drug action in attention-deficit/hyperactivity disorder: A review and integration. *Behavioural Brain Research* **94**: 127–52

Swanson J, Castellanos F X, Murias M, LaHoste G, Kennedy J 1998 Cognitive neuroscience of attention deficit hyperactivity disorder and hyperkinetic disorder. *Current Opinion in Neurobiology* **8**: 263–71

Tannock R 1998 Attention deficit hyperactivity disorder: advances in cognitive, neurobiological, and genetic research. *Journal of Child Psychology and Psychiatry* **39**: 65–99

Taylor E 1998 Clinical foundations of hyperactivity research. *Behavioural Brain Research* **94**: 11–24

Whalen C K, Henker B 1997 Stimulant pharmacotherapy for attention-deficit/hyperactivity disorders: an analysis of progress, problems, and prospects. In: Fisher S, Greenberg R P (eds.) *From Placebo to Panacea: Putting Psychiatric Drugs to the Test*. J. Wiley, New York, pp. 323–55

Whalen C K, Henker B 1998 Attention-deficit/hyperactivity disorders. In: Ollendick T H, Hersen M (eds.) *Handbook of Child Psychopathology*, 3rd edn. Plenum, New York, pp. 181–211

C. K. Whalen

Attention-deficit/Hyperactivity Disorder, Neural Basis of

Attention-deficit hyperactivity disorder (ADHD) is a clinical syndrome that was initially described in the early twentieth century in the aftermath of an encephalitis epidemic. Originally called Minimal Brain Disorder and then Minimal Brain Dysfunction, the assumption was that the disorder had a neurological etiology. More recently, the terms Hyperkinetic Syndrome, Attention Deficit Disorder (ADD) and, most recently, Attention-Deficit Hyperactivity Disorder (ADHD), replaced earlier nomenclature to capture the primary descriptive features of the disorders and avoid any commitment about etiology (Biederman 1998, Shaywitz et al. 1997). Detailed descriptions of the disorder, controversies in description and classification, prevalence and developmental course are included in the article on Attention-deficit/hyperactivity disorder (ADHD) by Carol K. Whalen (*Attention-deficit/Hyperactivity Disorder (ADHD)*). The clinical features as a prelude to a discussion of the putative psychological, neurological and biological causes of this disorder and the relationship of these causes to treatment strategies are reviewed.

1. Clinical Features

Attention-deficit/hyperactivity disorder is characterized by three core behavioral symptoms—inattention, hyperactivity, and impulsivity. The behavioral indications of inattention do not correspond to a single psychological construct, but rather combine features of sustained attention, selective attention, and active working memory (Shaywitz et al. 1997). By current definitions, inattention with or without hyperactivity and impulsivity must have persisted for a considerable amount of time—operationally since before age 7 years and at least 6 months in duration—and must characterize the child's behavior in multiple settings, operationally usually home and school (American Psychiatric Association 1994). Finally, the behaviors must lead to clinically significant functional impairments in social, academic, or occupational functioning (American Psychiatric Association 1994).

ADHD typically presents during the preschool or early school aged years. The majority of children with ADHD continue to show core symptoms in adolescence and some retain these characteristics into adulthood (Biederman 1998), although the clinical manifestations may change over time. Co-existing conditions, such as anxiety and conduct disturbance, also persist into adolescence and may predispose the teenager with ADHD to additional problems such as substance abuse and juvenile delinquency. These features—long-standing and pervasive symptoms—are suggestive but not at all conclusive that the disorder has a strong biological basis.

A debate persists as to whether ADHD should be conceptualized as a single disorder or whether inattention and hyperactivity are separate, but frequently co-occurring, deficits. Current definitions represent the middle ground: the name of the disorder includes behaviors from two dimensions and definitions allow for three different variants: ADHD-predominantly inattentive type, ADHD-predominantly hyperactive-impulsive type, and ADHD-

combined type (American Psychiatric Association 1994). The implication of this debate is that combining two overlapping syndromes into a single category may have obscured the underlying etiology. Future research will be important to determine the nosology of the condition.

ADHD frequently co-occurs with other neuropsychiatric and neurobehavioral disorders. Aggressive disorders, including oppositional defiant disorder and conduct disorder, have been reported in 15.6 to 48.3 percent of children with ADHD (Green et al. 1999). Anxiety disorder was found in 24.0 to 35.7 percent of children with ADHD (Green et al. 1999). Depression disorder was found in 14.3 percent and 18.8 percent of children with ADHD (Green et al. 1999). Learning disorders are also common in children with ADHD, though precise estimates are difficult to obtain because of the variations in the definition of learning disorders. ADHD frequently co-occurs with problems of movement and coordination. ADHD has been found to be associated with other medical and neurological disorders, including Tourette syndrome. Thus, any discussion about underlying causes must take into account the high prevalence of co-existing conditions.

2. *Underlying Psychological and Biological Mechanisms*

Some investigators conceptualize ADHD as a disorder of inhibition rather than a problem of perception or information processing (Barkley 1997, Casey et al. 1997). In this model, inattention is explained as a problem with inhibiting automatic, potent, and/or interfering responses. A major appeal of this model is that the deficits of hyperactivity/impulsivity and inattention can be unified through this single construct. Moreover, it links the core symptoms with cognitive and metacognitive deficits associated with the disorder. According to Barkley (1997), behavioral inhibition makes a fundamental contribution to other executive functions, including working memory; self-regulation of affect, motivation and arousal; and reconstitution, or the ability to analyze information and re-synthesize it into novel structures. Executive functions support the inhibition of rapid and automatic responses to immediate environmental inputs, thereby resulting in delays in responding and in control of thinking and action by internal representations rather than environmental stimuli. Together, executive functions allow prioritization, organization, and goal-directed activity. According to this model, some cases of ADHD-predominantly inattentive type may be reconceptualized as a distinct disorder. This other disorder would be characterized by slow information processing and difficulties with selective attention, rather than as impairments of inhibition, goal-directed persistence, and self-regulation. Further research is necessary to determine if the clinical condition should be reformulated into two, possibly correlated disorders.

Executive functions are the neuropsychological functions subserved by the prefrontal cortex. Children with ADHD bear a weak resemblance to patients with deficits of executive function from known causes such as frontal lobe brain injury. However, children with ADHD are typically less impaired than these patients.

A broad range of studies using multiple different methodologies support the concept that ADHD has a neurological basis. Increasing evidence from structural imaging studies comparing children with ADHD to unaffected controls find anatomic differences in prefrontal cortex and connecting networks, including in the basal ganglia, particularly in the right hemisphere and in the corpus callosum (Swanson et al. 1998, Zametkin and Liotta 1998). Functional imaging studies are also suggestive of altered functioning of the fronto-striatal circuits (Casey et al. 1997). These findings fit with the construct of ADHD as a deficit in executive functioning since these pathways originating in the prefrontal cortex are presumed to be involved in inhibition of automatic responses, planning, and organizing. Other anatomical differences have been reported, including differences in posterior regions including the parietal lobe and cerebellum, the hippocampus, and the thalamus (Zametkin and Liotta 1998).

The magnitudes of the neuroanatomical differences in all of these studies are relatively small and the distributions of the sizes of anatomic regions in the different groups overlap. Therefore, these measurements, interesting for the insights they provide regarding the neurological basis of ADHD, are currently not useful for clinical diagnostic purposes. Moreover, the association of size differences and ADHD should not be used to infer that the anatomical differences are the primary cause of the disorder. These structural and functional differences might be a reflection of a history of different neurological function rather than the cause of the disorder.

Genetic causes of ADHD are suspected because of the finding that siblings of affected children and children of affected adults are at increased risks of the disorder. Twin studies also report a higher concordance between monozygotic than dizygotic twins. Evidence is mounting that the genetic causes may relate to either abnormal dopamine receptors or dopamine transporter molecules (Biederman 1998). The role of dopamine in the etiology of ADHD has long been implicated because stimulant medications that increase dopamine concentrations in synaptic spaces have been shown in hundreds of studies to decrease the core features of ADHD. Some family studies implicate a single autosomal dominant gene (Zametkin and Liotta 1998). However, genetic heterogeneity is likely. For example, another genetic mutation in a thyroid receptor gene on chromosome 3 has been associated

with a rare disorder of ADHD plus generalized resistance to the hormone.

ADHD is also associated with a variety of prenatal and acquired medical conditions. Indeed, the disorder was initially described in the early twentieth century to characterize the behavioral problems of patients in the aftermath of encephalitis and similar neurological insults. ADHD is also associated with complications of pregnancy, labor, delivery, or infancy and also with maternal smoking or alcohol use *in utero* (Milberger et al. 1997). These data suggest that many biological conditions may give rise to ADHD, but that despite differences in etiology, all share a similar phenotypic expression. Whether all of the predisposing conditions impact the same brain regions, circuits, or neurotransmitters as those implicated in cases of familial ADHD, has yet to be determined.

It is intriguing to consider how the high prevalence of co-existing conditions with ADHD may further elucidate the underlying mechanisms. Neuropsychiatric disorders such as anxiety, depression, and learning disabilities are also thought to have a neurological basis. It is not at all clear whether co-existing conditions are based on the same underlying neural mechanisms, different but correlated disorders or secondary behavioral problems related to life difficulties with ADHD.

A dramatically different but not entirely inconsistent view of the causes of ADHD considers whether the core traits of the condition could be conceptualized as adaptive responses to certain environmental conditions over the evolution of the species or the development of the individual. The evolutionary viewpoint (Jensen et al. 1997) rests on the assumption that a disorder this prevalent in the population would not be maintained through natural selection unless it conferred some relative advantage to at least some affected individuals. High motor activity, rapid shifts of attention, and quick response might have been desirable traits in some harsh and dangerous ancestral environments. If so, the genetic basis of this response style may have been the object of natural selection, thus sustaining the high prevalence of the condition in the population. However, current conditions in the industrialized world, which emphasize universal education, literacy, and logical sequential problem solving, render these same traits as maladaptive.

The analogous argument on the developmental timetable begins with the assumption that certain conditions in the early life of the child, including the prenatal environment, infancy, and toddler years, might favor the development of a neural system that makes rapid shifts of attention and quick responses. Such environments might include excessive television watching, limited opportunities for sustained conversation, or adverse conditions in rearing, such as inconsistent and unpredictable social contact (Biederman et al. 1998). After years of such exposure, a neural system of rapid attention shifting may dominate neural functioning, even when the environment calls for different characteristics such as sustained attention and careful planning. These personal traits may interact with cultural norms and expectations, which could serve to motivate change or to encourage persistence. In such cases, the neural circuits predisposing to this condition may be present even though genetic causes might not be. This area is in need of vigorous research.

3. *Treatments*

Medications classified as stimulants, including methylphenidate hydrochloride and dextroamphetamine sulphate, are the most widely used forms of medical management for ADHD. These medications have been shown to have dramatic short-term ameliorative effects on the core systems of ADHD. A 14-month multi-center study found that about 67 percent of children treated with stimulant medications showed behavioral ratings on medication within the normal range (The MTA Cooperative Group 1999). The stimulant medications also improve some of the coexisting conditions such as anxiety and oppositional behavior.

Stimulant medications work by increasing the levels of neurotransmitters (dopamine, norepinephrine, and serotonin) at the presynaptic cleft. Thus, the utility of this treatment supports the concept of ADHD as a neurological disorder primarily in catecholamine pathways. The effects of stimulant medications on cognition, social and academic functioning and other features of the co-existing conditions remain controversial.

The other mainstay of treatment is behavior management. The goal of behavior management is to create specialized environments that provide the child with ADHD with appropriate opportunities to use motor activity and rapid attention shifts, and that offer clear contingencies for sustained on-task effort, self-paced learning, and self-monitoring. In this tailored environment, the child receives frequent feedback and reinforcement, both positive and negative, contingent on his or her behavior. Treatments are geared toward parents and teachers who are coached on how to provide frequent, consistent, appropriate, and motivating feedback to the children about desirable and unacceptable behavior. In a 14-month multicenter study, medication alone and combinations of medical and behavior management did not differ significantly in the treatment of core symptoms of ADHD, suggesting that the behavior management did not add substantially to the improvements associated with medication. However, in the treatment of oppositional/aggressive symptoms, internalizing symptoms such as anxiety, teacher-rated social skills, parent-child relations, and reading achievement, a combination of medication and behavior management

proved superior to intensive behavioral treatment alone while medication management alone was not (The MTA Cooperative Group 1999). Most clinicians recommend a combination of treatments. The use of behavior management in treatment may not have a direct effect on the underlying biological basis of ADHD. It is possible that the relearning that a child experiences through behavior management allows for modification of underlying neural networks. This topic is an important one for future research.

4. Conclusion

ADHD is a complex and prevalent neurobehavioral disorder that presents in childhood and may persist into adulthood. The condition frequently co-exists with other behavioral and cognitive disorders. Evidence of a biological basis for the disorder is mounting through the use of neural imaging studies and genetic analysis. Conceptualizing the disorder as a mismatch of the child's underlying personal characteristics and the environment does not necessarily negate the possibility of a biological basis for ADHD. Management of the disorder includes stimulant medications that operate on neural pathways putatively involved in the etiology of the disorder. Behavior management may operate at the psychosocial level but have an influence on neurological organization.

See also: Anxiety Disorder in Children; Attention-deficit/Hyperactivity Disorder (ADHD); Attention, Neural Basis of; Child and Adolescent Psychiatry, Principles of; Developmental Psychopathology: Child Psychology Aspects; Ergonomics, Cognitive Psychology of; Neurotransmitters; Prefrontal Cortex Development and Development of Cognitive Function

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental disorders*, 4th edn. American Psychiatric Association, Washington, DC, pp. 78–85

Barkley R A 1997 Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin* **121**: 65–94

Biederman J 1998 Attention-Deficit/Hyperactivity Disorder: A lifespan perspective. *Journal of Clinical Psychiatry* **59** (Supplement 7): 4–16

Casey B J, Castellanos F X, Jiedd J N, Marsh W L, Hamburger S D, Schubert A B, Vauss Y C, Vaituzis A C, Dickstein D P, Sarfatti, S E, Rapoport J L 1997 Implication of right fronto-striatal circuitry in response inhibition and attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry* **36**: 374–83

Green M, Wong M, Atkins D, et al. 1999 Diagnosis of attention deficit/hyperactivity disorder. Technical review No. 3. *AHCPR Publication No. 99-0036*. Agency for Health Care Policy and Research, Rockville, MD

Jensen P S, Mrasek D, Knapp P K, Steinberg L, Pfeffer C, Scholalter J, Shapiro T 1997 Evolution and revolution in child psychiatry: ADHD as a disorder of adaptation. *Journal of the American Academy of Child and Adolescent Psychiatry* **36**: 1672–9

Milberger S, Biederman J, Faraone S V, Guite J, Tsuang M T 1997 Pregnancy, delivery and infancy complications and attention deficit hyperactivity disorder: issues of gene-environment interaction. *Biological Psychiatry* **41**: 65–75

Shaywitz B A, Fletcher J M, Shaywitz S E 1997 Attention-deficit Hyperactivity Disorder. *Advances in Pediatrics* **44**: 331–67

The MTA Cooperative Group 1999 A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry* **56**: 1073–86

Zametkin A J, Liotta W 1998 The neurobiology of attention-deficit/hyperactivity disorder. *Journal of Clinical Psychiatry* **59** (Supplement 7): 17–23

H. M. Feldman

Attention: Models

'Attention' is a general term for selectivity in perception. The selectivity implies that at any instant a perceiving organism focuses on certain aspects of the stimulus situation to the exclusion of other aspects. In this article, theoretical models of attention are reviewed and their development is described.

1. Theoretical Beginnings

The first modern theory of attention was the *filter* theory of Broadbent (1958) (see Figure 1). In this theory, information flows from the senses through many parallel input channels into a short-term mem-

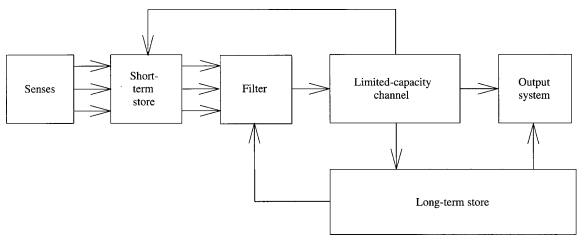

Figure 1
Flow diagram illustrating the filter theory of Broadbent (1958)

ory store. The short-term store can hold the information for a period of the order of seconds. Later in the system there is a *limited-capacity channel*, whose capacity for transmitting information is much smaller than the total capacity of the parallel input channels. Between the short-term memory and the limited-capacity channel is a selective filter which acts as an all-or-none switch that selects information from just one of the parallel input channels at a time.

Broadbent (1958) defined an input channel as a class of sensory events that share a simple physical feature (e.g., a position in auditory space). Except for analysis of such features, stimuli on unattended channels should not be perceived. This conjecture accounted for results of studies in the early 1950s by E. C. Cherry on the ability to attend to one speaker in the presence of others (the *cocktail party problem*). Cherry asked his subjects to repeat a prose message while they heard it, rather than waiting until it finished. When the message to be repeated (*shadowed*) was presented to one ear while a message to be ignored was presented to the other ear (*dichotic* presentation), subjects typically were unable to recall any words from the unattended message.

Later studies by Moray (1969) and others showed that subjectively important words (e.g., the subject's own name) tended to be recognized even if presented on the nonshadowed channel. To accommodate such findings, A. M. Treisman developed a variation of filter theory in which the filter operates in a graded rather than an all-or-none fashion. In Treisman's *attenuation* theory, unattended messages are weakened rather than blocked from further analyses. Both selected and attenuated messages are transmitted to a pattern recognition system with word recognition units. Because thresholds of recognition units for important words are lowered, important words tend to be recognized even if appearing in attenuated messages.

In the filter theories of Broadbent and Treisman, attentional selection occurs at an earlier stage of processing than pattern recognition. Such theories are called *early-selection* theories. In *late-selection* theories, attentional selection occurs later in processing than pattern recognition. The first late-selection theory was outlined by J. A. Deutsch and D. Deutsch in 1963. Deutsch and Deutsch argued that the early filtering mechanism in Treisman's theory was redundant. They proposed that attended and unattended messages receive the same amount of analysis by the pattern recognition system. However, after a stimulus has been recognized, the importance of the stimulus is retrieved and the stimulus with the greatest importance is selected for further processing, including conscious awareness.

The theories of Broadbent, Treisman, and Deutsch and Deutsch set the stage for the development of more specific, quantitative models of attention. Most of these models were based on experimental findings on visual processing: data on our ability to *divide* attention between multiple, simultaneous targets and data on our ability to *focus* attention on targets rather than distractors.

2. Serial Models

In serial models of attention, only one stimulus is attended at a time. This section examines the development from simple serial models to selective serial models.

2.1 Simple Serial Models

In visual *whole-report* experiments by G. Sperling in the early 1960s, subjects were instructed to report as many letters as possible from a briefly exposed array of unrelated letters followed by a pattern mask. The number of correctly reported letters depended on the stimulus-onset asynchrony (SOA) between the letter array and the mask. Corrected for guessing, the score appeared to be zero when the SOA was below a certain threshold. As the SOA exceeded the threshold, the mean score initially increased at a high rate of about one letter per 10–15 ms. The mean score leveled off as it approached a value of about four letters or the number of letters in the stimulus, whichever was smaller.

Sperling proposed a simple serial model to account for the initial strong and approximately linear increase in mean score as SOA exceeded threshold. By this model, the subject encodes one letter at a time, requiring 10–15 ms to encode a letter. The serial encoding is interrupted when the stimulus is terminated by the mask or when the number of encoded letters reaches the immediate memory span of the subject.

Simple serial models for *visual search* were developed in the 1960s by W. K. Estes, S. Sternberg, and others. In most experiments on visual search, the subject is instructed to indicate 'as quickly as possible' whether or not a certain type of target is present in a display. Positive (target present) and negative (target absent) reaction times are analyzed as functions of the number of items in the display (*display set size*).

By a simple serial model, items are scanned one by one. As each item is scanned, it is classified as a target or as a distractor. A negative response is initiated if and when all items have been scanned and classified as distractors. Thus, the number of items processed before a negative response is initiated equals the display set size, N. Accordingly, the rate of increase in mean negative reaction time as a function of display set size equals the mean time taken to process one item, Δt.

In a *self-terminating* serial search process, a positive response is initiated as soon as a target is found. As the order in which items are scanned is independent of

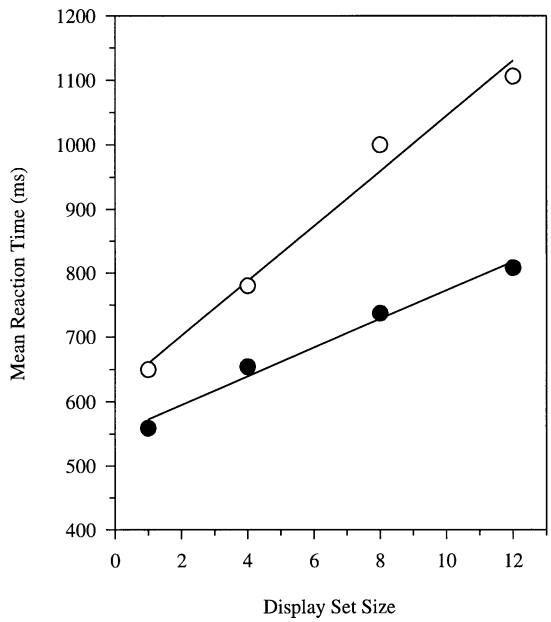

Figure 2
Positive and negative mean reaction times as functions of display set size in visual search for a T in one of four possible orientations among similarly rotated Ls (The observed reaction times were read from Wolfe 1994, Figure 1b. The slopes of the fitted least squares lines are 22 ms/item and 43 ms/item, respectively)

their status as targets vs. distractors, the number of items processed before a positive response is initiated varies randomly between 1 and N and averages $(1+N)/2$. Thus, the rate of increase in mean positive reaction time as a function of display set size equals one half of the mean time taken to process one time, $\Delta t/2$. A representative pair of search reaction time functions with a positive-to-negative slope ratio of about 1:2 is illustrated in Fig. 2.

A. M. Treisman has introduced a distinction between feature and conjunction search. In *feature* search, the target possesses a simple physical feature (e.g., a particular color, shape, or size) not shared by any of the distractors. For example, the target can be a red T among black Ts. In *conjunction search*, the target differs from the distractors by possessing a predesignated conjunction of physical features (e.g., both a particular color and a particular shape), but the target is not unique in any of the component features of the conjunction (i.e., in color or in shape). For example, the target can be a red T among black Ts and red Xs.

Many experiments on conjunction search have yielded positive and negative mean reaction times that are approximately linear functions of display set size with substantial slopes and positive-to-negative slope ratios of about 1:2. This pattern conforms to predictions from simple self-terminating serial models, and Treisman and co-workers have concluded that conjunction search is performed by scanning items one at a time. Experiments on feature search with low target-distractor discriminability have yielded a similar pattern of results.

2.2 Selective Serial Models

In selective serial models, items in the stimulus display are attended one at a time, but the sequential order in which items are attended depends on their status as targets vs. distractors: When a target and a distractor compete for attention, the target is more likely to win.

The first selective serial model of visual search was published by J. E. Hoffman in 1978. It was motivated by findings from C. W. Eriksen's laboratory on the time taken to shift attention in response to a visual cue. The findings cast doubt on the notion that attention can be shifted from item to item at the high rates presumed in simple (nonselective) serial models of processing.

In Hoffman's model, visual search is a two-stage process in which a parallel evaluation of the entire stimulus display guides a slow serial processor. The parallel evaluation is preattentive and quick, but error prone. For each item in the display, the outcome is an overall measure of the similarity between this item and the prespecified targets. Items are transferred one by one to the second stage of processing. The serial transfer mechanism is slow (about one item per 100 ms), but it makes search efficient by transferring items in order of decreasing overall similarity to the prespecified targets. Thus, if there is a target in the display, the target is likely to be among the first items that are transferred to the second stage of processing.

J. M. Wolfe, K. R. Cave, and S. L. Franzel have proposed a selective serial model, which they called *Guided Search* (cf. Wolfe 1994). The model combines elements of the two-stage model of Hoffman with elements of the *feature integration* theory of Treisman and co-workers (cf. Treisman 1988). As in feature integration theory, simple stimulus features such as color, size, and orientation are registered automatically, without attention, and in parallel across the visual field. Registration of objects (items defined by conjunctions of features) requires a further stage of processing at which attention is directed serially to each object. As in Hoffman's model, the outcome of the first, parallel stage of processing guides the serial processing at the second stage. The guidance works as follows.

For each feature dimension (e.g., color, size, or orientation), the parallel stage generates an array of activation values (attention values). The array forms a

map of the visual field. Each activation value is a sum of a bottom-up and a top-down component. For a particular location within a map for a given feature dimension, the bottom-up component is a measure of differences between the value of the feature at that location and values of the same feature at other locations. The top-down component for a feature dimension at a particular location is a measure of the difference between the value of the feature at that location and the target value for the feature dimension. After activations have been calculated in separate maps for each feature dimension, they are summed across feature dimensions to produce a single overall activation map. In simulations, a certain level of gaussian noise is also added at each location. The final overall activation values represent the evaluation given by the parallel stage of how likely the stimulus at each location is to be the target.

The serial stage processes the items one by one in order of decreasing activation in the overall activation map. Each item it processes gets classified as a target or as a distractor. The serial processing continues until a target is found or until all items with activations above a certain value have been processed.

The guided search model accounts for many findings from experiments on visual search. It was motivated, in particular, by demonstrations of fast conjunction search. Some demonstrations of fast conjunction search are accommodated by assuming that for some feature dimensions, top-down control is very effective. Other demonstrations are accommodated by assuming that in some subjects, the level of gaussian noise is very low.

3. *Parallel Models*

In parallel models of attention, several stimuli can be attended at the same time. This section examines the development from simple parallel (independent-channels) models to limited-capacity parallel models and race-based models of selection.

3.1 *Independent Channels Model*

The first detailed parallel model of visual processing of multi-item displays was the Independent Channels model developed by C. W. Eriksen and co-workers in the 1960s. It was based on the assumption that display items presented to separated foveal areas are processed in parallel and independently up to and including the stage of pattern recognition. The assumption implies that the way in which a display item is processed is independent of random variations in the way in which other display items are processed. It also implies that the way in which an item is processed is independent of display set size.

The notion of independent channels (*unlimited-capacity parallel processing*) is used mainly to account for cases in which visual search is highly efficient (small effects of display set size). This includes cases of feature search with high target-distractor discriminability. In a theory proposed by R. M. Shiffrin and W. Schneider in 1977, it also includes cases of search for more complex targets such as particular alphanumeric characters when subjects have been trained consistently in detecting those particular targets.

In the theory of Shiffrin and Schneider, slow, serial, *controlled search* for particular items can develop into fast, parallel *automatic detection* of the same items. Automatic detection occurs without subject control and without stressing the capacity limitations of the system. The development of automatic detection presupposes that the mapping of stimuli to responses is consistent rather than varied over trials.

3.2 *Evidence from Automatic Interference*

Empirical support for the assumption of parallel processing has come from demonstrations by Eriksen and others of *Stroop-like interference* in processing of multi-item displays. In the original task developed by J. R. Stroop in the 1930s, subjects are asked to name the color of the ink used to print a word. The *Stroop effect* denotes the fact that the task is more difficult when the word itself is the name of a different color (e.g., *red* printed in blue) than when the word does not refer to a color or refers to the color shown by the ink (*blue* printed in blue).

The original Stroop task concerns selective attention to a particular feature of an item (*featural attention*). In the *flankers task* of Eriksen, subjects are asked to focus attention on a target presented in a known spatial location (*spatial attention*). Usually, the target is a letter presented at fixation, and the subject is required to make a speeded binary classification of the letter (e.g., move a lever in one direction if the letter is a T or a K but in the opposite direction if the letter is an S or a C). The target is flanked by letters that should be ignored. However, the task is more difficult when the flankers are *response-incompatible* with the target (Ss or Cs flanking a T) than when the flankers are *neutral* (e.g., Xs flanking a T) or *response-compatible* with the target (Ts or Ks flanking a T). Eriksen suggested that the flankers are processed in parallel with the target up to and including the stage of pattern recognition.

3.3 *Limited-capacity Models*

The linear relations between mean reaction time and display set size predicted by simple serial models are hard to explain by parallel models with independent

channels (unlimited-capacity parallel models). However, the linear relations can be explained by parallel models with *limited processing capacity*. In 1969, the following example was published independently by J. T. Townsend and R. C. Atkinson, and his co-workers.

Consider a display of N items that are processed in parallel. Let the processing speed for an item (technically, the 'hazard function' for the processing time of the item; cf. Townsend and Ashby 1983) equal the amount of processing capacity devoted to that item, and suppose the total processing capacity spread across items in the display is a constant C. Then the time taken to complete processing of the first item is exponentially distributed with rate parameter C. Suppose that when the first completion occurs, the processing capacity is redistributed among the remaining N–1 items. Then the time from the first to the second completion is exponentially distributed with the same rate parameter C. Let the process repeat until all N items have been completed. If so, then the mean time taken to complete processing of the N items increases linearly with display set size N.

In 1977, M. L. Shaw and P. Shaw proposed a model for *Optimal Allocation of Cognitive Resources to Spatial Locations* that showed how limited-capacity models could be extended to situations in which the probability that a target occurs at a given location varies across display locations. In such situations, capacity was assumed to be allocated and reallocated among display items so that performance is optimized.

In the model of Shaw and Shaw, attention can be split among noncontiguous locations in the visual field. Processing capacity can be allocated to several separated locations at the same time. C. W. Eriksen and co-workers have proposed an alternative conception, a *zoom lens* model of the visual attentional field. In this conception, the attentional field can vary in size from an area subtending less than 1° of visual angle to the full size of the visual field. Because total processing capacity is limited, the amount of processing capacity allocated to a given attended location decreases as the size of the attentional field increases. However, the attentional field cannot to be split among noncontiguous locations. Direct tests of this hypothesis have been attempted, but the issue is still open.

3.4 Race Models of Selection

In race models of selection from multi-item displays, display items are processed in parallel and attentional selection is made of those items that first finish processing (the winners of the race). Thus, selection of targets rather than distractors is based on processing of targets being faster than processing of distractors.

In 1988, H. Shibuya and C. Bundesen proposed a fixed-capacity independent race model (*FIRM*). The model describes the processing of a stimulus display as follows. First, an *attentional weight* is computed for each item in the display. The weight is a measure of the strength of the sensory evidence that the item is a target. Then the available processing capacity (a total amount of C items/s) is distributed across the items in proportion to their weights. The amount of processing capacity that is allocated to an item determines how fast the item can be encoded into visual short-term memory (VSTM). Finally, the encoding race between the items takes place. The time taken to encode an item is assumed to be exponentially distributed with a rate parameter equal to the amount of processing capacity that is allocated to the item. The items that are selected (i.e., stored in VSTM) are those items whose encoding processes complete before the stimulus presentation terminates and before VSTM has been filled up.

Detailed tests of FIRM have been made in *partial-report* experiments in which subjects report as many targets as possible from a briefly exposed display showing a mixture of targets (e.g., red letters) and distractors (e.g., black letters). FIRM has provided accurate accounts of effects of selection criterion, exposure duration, and numbers of targets and distractors in the display (see Fig. 3).

3.5 Theory of Visual Attention (*TVA*)

The theory of visual attention (*TVA*) proposed by Bundesen (1990) is a generalization of FIRM. In TVA, both visual recognition and selection of items in the visual field consist in making *perceptual categorizations* (i.e., encoding categorizations into VSTM). When one makes the perceptual categorization that a given item belongs to a certain category, the item is said (a) to be selected, and (b) to be recognized as a member of the category.

Recognition and selection depend on the outcome of a biased race between possible perceptual categorizations. The rate at which a possible categorization ('item x belongs to category i') is processed increases with (a) the strength of the sensory evidence that supports the categorization, (b) the subject's bias for assigning objects to category i, and (c) the attentional weight of item x. When a possible categorization completes processing, the categorization enters VSTM if memory space is available there. The span of VSTM is limited to about four items. Competition between mutually incompatible categorizations of the same item are resolved in favor of the first-completing categorization.

TVA accounts for many findings on single-stimulus recognition, whole report, partial report, search, and detection. The theory has been extended by G. D. Logan to encompass aspects of perception and memory as well as attention.

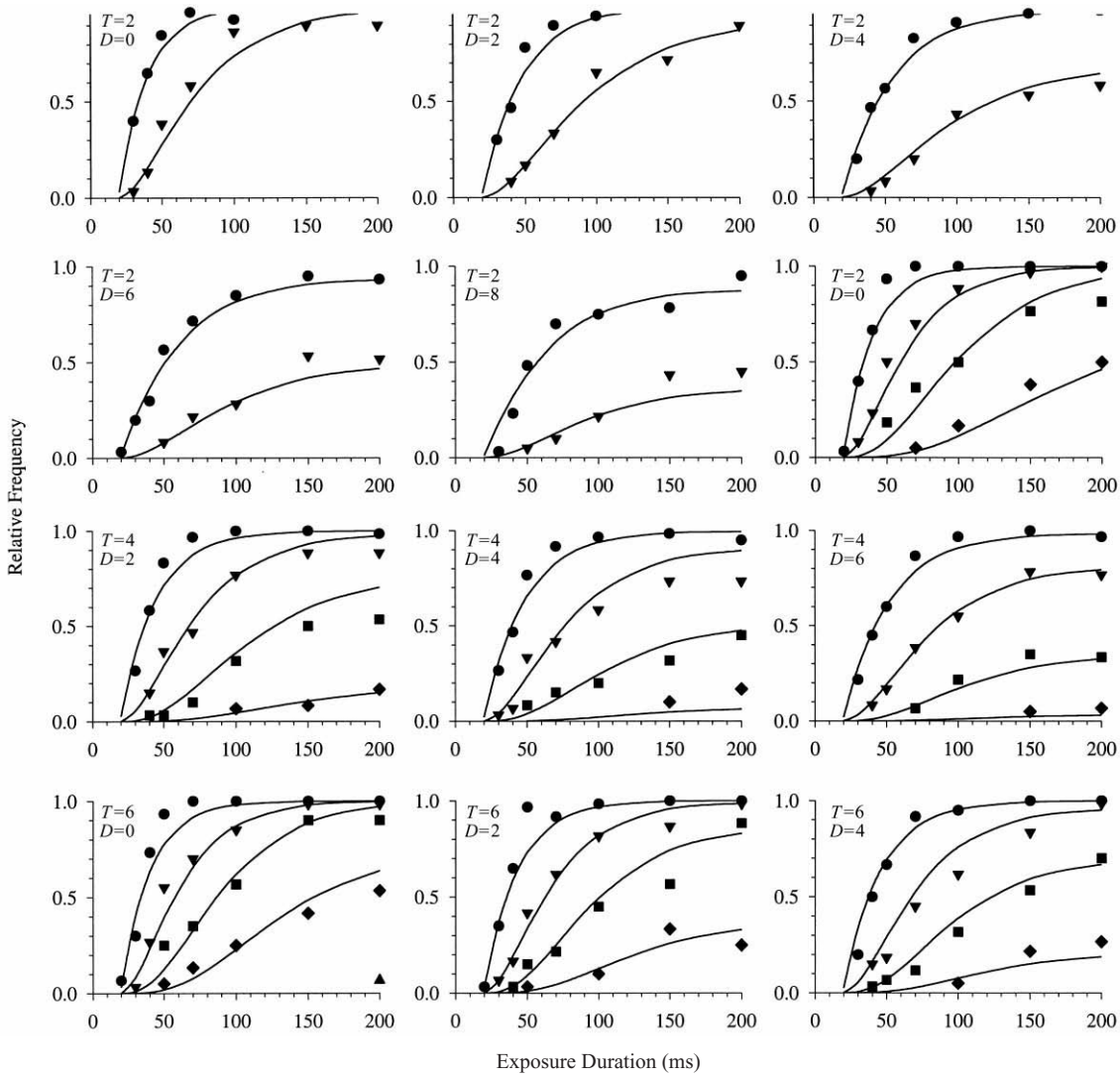

Figure 3
Relative frequency of scores of *j* or more (correctly reported targets) as a function of exposure duration with *j*, number of targets *T*, and number of distractors *D* as parameters in partial report of digits among letters (Parameters *T* and *D* vary between panels. Parameter *j* is l (circles), 2 (downward pointing triangles), 3 (squares), 4 (diamonds), or 5 (upward pointing triangle). Smooth curves represent a theoretical fit to the data by the fixed-capacity independent race model (FIRM). For clarity, observed frequencies less than 0.02 were omitted from the figure. Source: Visual selection from multi-element displays: Measuring and modeling effects of exposure duration by Shibuya and Bundesen 1988, *Journal of Experimental Psychology: Human Perception and Performance*, p. 595. ©American Psychological Association)

4. *Neural Network Models*

Formal models like TVA are highly abstract. Neural network modeling is an attempt to theorize at a level that is closer to neurobiology. In neural network models, information processing consists in a flow of activation through a network of neuronlike units that are linked together by facilitatory and inhibitory connections.

A simple way of implementing attentional selection in a neural network is by arranging the connections so that (a) units representing mutually compatible categorizations of the same item facilitate each other but (b) units representing incompatible categorizations in-

hibit each other, and (c) units representing categorizations of different items also inhibit each other. Search for a red target, for example, can then be done by preactivating units representing redness. If a red target is present, the preactivation will directly facilitate the correct categorization of the target with respect to color. Indirectly the preactivation will facilitate categorizations of the target with respect to other properties than color but inhibit categorizations of any other items than the target (*integrated competition*; Duncan 1996).

Many neural network models of attention have appeared since the mid-1980s. Good examples are the model for selective attention and pattern recognition by K. Fukushima, the multiple object recognition and attentional selection model (MORSEL) of Mozer (1991), the selective attention model (SLAM) of R. H. Phaf, A. H. C. van der Heijden, and P. T. W. Hudson, and the search via recursive rejection model (SERR) of G. W. Humphreys and H. J. Müller.

5. *Conclusion*

Current models of attention have sprung from the theoretical framework developed by Broadbent in the 1950s. The first detailed models of visual attention were the serial scanning models of Sperling, Estes, and others and the independent channels model developed by Eriksen and co-workers. Early tests of the models brought important discoveries but no simple resolution of the serial vs. parallel processing issue. Attempts to integrate the empirical findings led to selective serial models and to parallel models with differential attentional weighting, including race models of selection. No extant model has accounted for the full range of empirical findings but substantial progress has been made.

See also: Attention and Action; Attention: Models; Attention: Multiple Resources; Broadbent, Donald Eric (1926–93); Dual Task Performance; Interference and Inhibition, Psychology of; Working Memory, Neural Basis of; Working Memory, Psychology of

Bibliography

Allport A 1989 Visual attention. In: Posne M I (eds.) *Foundations of Cognitive Science*. MIT Press, Cambridge, MA

Broadbent D E 1958 *Perception and Communication*. Pergamon Press, London

Bundesen C 1990 A theory of visual attention. *Psychological Review* **97**: 523–47

Bundesen C 1996 Formal models of visual attention: A tutorial review. In: Kramer A F, Coles M G H, Logan G D (eds.) *Converging Operations in the Study of Visual Selective Attention*. American Psychological Association, Washington, DC

Cowan N 1995 *Attention and Memory: An Integrated Framework*. Oxford University Press, Oxford, UK

Desimone R, Duncan J 1995 Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* **18**: 193–222

Duncan J 1996 Cooperating brain systems in selective perception and action. In: Inui T, McClelland J L (eds.) *Attention and Performance XVI: Information Integration in Perception and Communication*. MIT Press, Cambridge, MA

Duncan J, Humphreys G W 1989 Visual search and stimulus similarity. *Psychological Review* **96**: 433–58

Kinchla R A 1992 Attention. *Annual Review of Psychology* **43**: 711–42

van der Heijden A H C 1992 *Selective Attention in Vision*. Routledge and Kegan Paul, London

van der Heijden A H C 1993 The role of position in object selection in vision. *Psychological Research* **56**: 44–58

LaBerge D 1995 *Attentional Processing: The Brain's Art of Mindfulness*. Harvard University Press, Cambridge, MA

Logan G D 1996 The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review* **103**: 603–49

Moray N 1969 *Attention: Selective Processes in Vision and Hearing*. Hutchinson, London

Mozer M C 1991 *The Perception of Multiple Objects: A Connectionist Approach*. MIT Press, Cambridge, MA

Norman D A 1976 *Memory and Attention: An Introduction to Human Information Processing*, 2nd edn. Wiley, New York

Pashler H E (ed.) 1998a *Attention*. Psychology Press, Hove, UK

Pashler H E 1998b *The Psychology of Attention*. MIT Press, Cambridge, MA

Posner M I, Petersen S E 1990 The attention system of the human brain. *Annual Review of Neuroscience* **13**: 25–42

Schneider W X, Maasen S (eds.) 1998 *Mechanisms of Visual Attention: A Cognitive Neuroscience Perspective*. Psychology Press, Hove, UK

Shiffrin R M 1988 Attention. In: Atkinson R C, Herrnstein R J, Lindzey G, Luce R D (eds.) *Stevens' Handbook of Experimental Psychology. Vol. 2: Learning and Cognition*. Wiley, New York

Sperling G, Dosher B A 1986 Strategy and optimization in human information processing. In: Boff K, Kaufman L, Thomas J (eds.) *Handbook of Perception and Performance*, Vol 1. Wiley, New York

Styles E A 1997 *The Psychology of Attention*. Psychology Press, Hove, UK

Swets J A 1984 Mathematical models of attention. In: Parasuraman R, Davies D R (eds.) *Varieties of Attention*. Academic Press, Orlando, FL

Townsend J T, Ashby F G 1983 *The Stochastic Modeling of Elementary Psychological Processes*. Cambridge University Press, Cambridge, UK

Treisman A 1988 Features and objects: The 14th Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology* **40A**: 201–37

Wolfe J M 1994 Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review* **1**: 202–38

C. Bundesen

Attention: Multiple Resources

Among the diverse conceptions of attention, a major one considers that selective attention is a consequence of the limited-capacity resources of our mental proces-

ses. Attention must be able to restrict the processing of the stream of information available from sensory and memory sources. Several questions have arisen over 50 years of research, and this article will present some related to the organization of the resources, and to the mechanisms involved when attention is focused on a piece of information.

1. Organization of Processing Resources

1.1 Limits of Processing Resources

Everyday life abounds in examples of situations in which two tasks are performed simultaneously (e.g., walking while talking). However, examples also abound of inefficient time-sharing, specifically when tasks are complex. Driving a car and talking with passengers can be relatively easy, but it becomes difficult or impossible if the demands of the driving task increase, or if the conversation requires reflection and arguments. Tasks are assumed to demand processing resources, and these resources are limited in their availability.

Many experiments have been developed to demonstrate, evaluate, and quantify the limits of mental or processing resources. In the dual-task method, subjects are required to perform two concurrent tasks simultaneously (see *Dual Task Performance*). Performance is degraded if attentional demands brought about by the tasks are high. For example, writing down a list of words heard through headphones while reading a text for meaning is extremely difficult. Subjects can be trained to carry out these two tasks simultaneously, but writing, in this case, becomes quite automatic: When subjects are asked after the experiment if there was a dominant semantic category in the list of words they had to write, results show that they did not attend to the meaning of the written words.

Other paradigms are used to explore how resources are shared when several stimuli are processed. The RSVP (rapid serial visual presentation) method consists of the fast presentation of several successive stimuli. This method shows that the identification of a first visual target disrupts the detection of a second target when the delay between both targets is too short (below 300 ms). The second target is merely not seen (a phenomenon known as the 'attentional blink') although it would have been detected if no task occurred on the first target. The identification (and/or the access to working memory) of the first target requires processing resources, which are therefore not available for the detection of another stimulus. These results are interpreted as a difficulty in sharing resources between the processing of both targets, or as the consequence of the time needed to shift attention from processing one target to processing the other.

Another type of experiment, the search task, demonstrates that several stimuli might not be processed simultaneously. In a typical search task, only one predefined target has to be detected among distracters (e.g., the target T among several Ls). Search time increases proportionally with the number of distracters (about 60 ms per additional distracter). Resources cannot easily be allocated to several stimuli in a visual display, and the search time is related to the size of the display.

1.2 Level of Selection

Because several stimuli cannot be processed simultaneously, or several concurrent tasks cannot be performed efficiently, selection has to occur. Selection allows the allocation of more processing resources to a privileged target or action. The act of attending selectively to a piece of information gives priority to more elaborate operations. The information processing flow could be considered as beginning with some parallel system followed by a serial system in which only attended information enters. The determination of the location of the bottleneck between both systems has been debated for a long time. At which level in the information processing flow is the selection performed, or which level has limited resources?

External information must undergo several successive and more or less discrete operations before an appropriate response is performed. The initial operations correspond to sensory recording: after the transduction, the different primary features are computed (e.g., color, texture, orientation, movement, etc.). A second series of operations ends with the perception of an object or event, the access to a semantic code. Then, according to the individual's goal, a strategy for action takes place (immediate motor response of different kinds, naming, access to working memory when a delayed response is required, etc.). Of course, things are more complicated, but two broad theories can be described concerning the selection level. According to the early theory of attention, the resources allowing access to the semantics are limited. In other words, only stimuli located in the focus of attention can be identified. According to the late theory of attention, the limited resources are located at the stage when a decision of response to a piece of information is taken. All the stimuli in a display can access the semantic code (provided that each stimulus is in an area of sufficient acuity), and the selection role is to choose which stimulus will provoke the response and/or which response will be performed.

In the auditory modality, the dichotic listening method (different messages delivered in each ear) has shown that, when a subject is attending to a continuous message in one ear, the identification of occasional messages in the 'ignored' ear is very poor, but not completely eliminated. This type of result is known as

the 'cocktail party' effect. At a cocktail party, people usually do not hear what is being said nearby when they are absorbed in a conversation. We are able to select a channel and filter out other channels, as predicted by the early theory. However, it might happen that a person would react, for example by turning the head, if a someone close beside them spoke their name. Consequently, an unattended piece of information can access the semantic code, as predicted by the late theory (the person's own name was identified). However, the defenders of the early theory have developed counter-arguments. First, attentional selection would be necessary to identify a piece of information unless this information was extremely familiar to the subject. Second, in the 'cocktail party' effect, as in several experiments using dichotic listening, subjects could sometimes shift their attention rapidly from the attended channel towards another channel of information. At the time of writing, the debate is still continuing, but most authors seem to agree that attentional selection can occur early *and* late in the processing of information, depending of the task, the familiarity of the information, etc.

1.3 Multiple Resources

A question which has aroused the interest of researchers working in the domain of attention concerns the unicity of the processing resources. The possibility for attention to occur at two (or more) different levels in the information flow, early and late, would indicate that the 'pool' of the mental resources may not be unique, but be composed of several 'reservoirs.' Other researches have also provided much evidence against the unicity of resources, independently of the early/late debate. For example, it is rather difficult to perform a linguistic task (repeating a text, counting in reverse order, etc.) and a spatial task (detect a target on a video screen, etc.) simultaneously. And it is still more difficult to perform two linguistic tasks simultaneously (listening and reading), or two spatial tasks, especially if these tasks concern information presented in the same modality.

Neuropsychology—the study of the performance of patients suffering from cerebral lesions—seems also in favor of such a modular organization. Lesions in different parts of the brain yield different deficits. Usually, lesions in the left hemisphere produce language disturbances, while lesions in the right hemisphere produce difficulties in visuo-spatial functions. Also, different types of language and perceptual deficits may occur, according to the exact location of the lesion. Many years ago, John Hughlings Jackson observed that cerebral lesions do not usually disrupt functioning completely, but only the most complex and propositional aspects of the function: more automatic behavior may be unimpaired. Alajouanine described the case of an anomic patient who, while not remembering the name of her daughter when asked, turned the head toward her and said 'See, my poor Jacqueline, I cannot remember even your own name.' Prosopagnosic patients cannot identify people's faces, but show different electrodermal response to familiar and unfamiliar faces.

Hemineglect patients tend to ignore the half of the visual space opposite to their lesion, but can sometimes identify information on this side when it is very familiar. Amnesic patients can demonstrate implicit learning. There are many other examples of patients who present some residual ability in their impaired function, showing that they can still process information as long as they do not need to attend to it specifically. The fact that different functional operations are linked to specific areas favors a multiple resource organization of the brain (however, no function is linked to a single area of the brain, and networks of connected areas are considered as the basis for any behavior). Neurophysiological studies (electrical cell recording, functional metabolism) also tend to show that attention is expressed 'everywhere' in the brain, and that each cortical area can have an automatic or an attentional state according to the task.

Processing resources are multiple and hierarchically organized. For example, auditory and visual resources are separated to some degree, and each of these resources can be shared by two tasks within the modality. There also exist more general resources such as verbal and spatial perception, each one able to be shared between modalities. Above all, resources can be allocated to any activity, or stage of processing, by a higher level executive program, a central processor, itself being divided into perceptual and response components. Whatever the exact organization of the hierarchy is, the role of the central processor is the command of behavior, the elaboration of strategies in allocating resources. In neuropsychology, patients suffering from frontal lesion usually have no difficulty in performing simple tasks, but are impaired considerably in complex tasks in which strategies have to be elaborated. For example, although they classify cards according to the features of the drawings on the cards quite well (i.e., color, form, number of shapes), they have great difficulty when asked to change the classification (i.e., shifting from color to form). The frontal lobe is crucial to the decision process of organizing attentional strategies.

2. Mechanisms of Attention

Research in the attention domain is not restricted to the overall organization of processing resources. Several lines of research have been developed in order to explore the mechanisms of attention (or allocation of

resources). Here, only a few examples of fundamental questions will be presented regarding (a) selection and its role in information processing, (b) preparatory attention and its components, and (c) the opposition between activation and inhibition.

2.1 Integration and Selective Attention

Selective attention enables a piece of information access to a more elaborated level of processing, in order to achieve a more accurate perceptual judgment, action, etc. In a search task, attention seems to be required when the discrimination between the target and the distracters is complex. A vertical bar can be detected easily among horizontal bars: The target will 'pop out.' Things are more complicated in the example cited above of a T among Ls, however, presumably because selection is necessary to combine or integrate specific primary features (vertical and horizontal bars), each of which being present in the distracter as well as the target. The need for a precise localization of each primary feature could explain why selection is necessary (see the work of Treisman and colleagues). Selection could be required to process attributes of objects and events as unified entities, which are the basis for the representation of knowledge in the mind.

2.2 Preparatory Attention

Another role of attention is to prepare for the processing of a stimulus or an action prior to the time when that stimulus or action is expected to occur. Attention can be maintained over a variable period of time in which some 'preprocessing' occurs, reducing the amount of processing needed when the stimulus itself appears, thus resulting in a more accurate and faster response.

A simple, noninformative signal announcing the stimulus onset can reduce the response time if the delay between the signal and the target stimulus is about 500 ms. Other types of preparation effect are even faster. Most studies in preparatory attention have used informative signals (usually called primes or cues), that is, signals sharing codes with the target. In spatial cueing experiments, a target stimulus (for example, a simple asterisk the subject has to detect by pressing a key, a letter to identify, etc.) is preceded by a spatial cue. The cue can be an arrow in the location where the subject's gaze is fixated, pointing towards one of the possible locations of the target in peripheral vision. It could also be a small signal at one of the possible locations of the target. Responses are faster and more accurate when the location indicated by the cue is congruent with the target location (valid condition) compared to when the indication is wrong (invalid condition), because preparatory attention allows the allocation of resources at the precise location of the following target. This effect occurs even if subjects do not move their eyes towards the location indicated by the cue. Preparatory attention in the visual space is usually followed by a saccade, placing the information on the fovea (area of best acuity), but both mechanisms can be dissociated, and one can attend to a location of the space without moving the eyes toward this location (covert attention).

The spatial cueing paradigm helps to dissociate between exogenous and endogenous attention. Typically, in cueing experiments, valid trials are frequent, and invalid trials rare, so the subjects' strategy is to use the information given by the cue. Endogenous attention, an active process the subject develops in order to orient voluntarily toward a location, requires a delay of 300 ms or more between the cue and the target. However, even if the subject does not elaborate strategies—for example, when the proportion of valid and invalid trials is 50/50, an effect of a peripheral spatial cue can occur with delays of 60 ms to 150 ms between the cue and the target. The cue attracts the exogenous attention automatically toward its location.

2.3 Inhibition/Activation

Another fundamental question concerns the respective contribution of inhibition and activation in the attentional mechanisms. When attention is directed toward a target stimulus in a visual display, a possibility is that inhibition would first attenuate the impact of distracters, followed by the activation of the target (although other models exist). The consequence is that the selected target stimulus can access a higher level of processing (enhancement) than nonselected stimuli.

Several components have been depicted in the cueing procedure by Posner and colleagues. In a nonvalid condition, subjects would first have to 'disengage' from the cue (inhibition) before moving and 'engaging' attention on the location of the target (activation). Behaviorally, the demonstration of inhibition or activation is not so easy, mainly because it must rest on the comparison between an attentional condition and a neutral condition, and neutral conditions are difficult to define. Nevertheless, this description of the cueing effect had the immense advantage of beginning, in the 1980s, a new era in characterizing performance of patients, as well as in understanding the cerebral mechanisms of attention. Patients who present with the syndrome of hemineglect (ignoring the part of the space controlateral to their lesion, usually in the parietal lobe) almost always have a deficit in disengaging spatial attention from the 'good' hemispace (ipsilateral to the lesion), without showing a systematic deficit in engaging attention toward the 'neglected' hemispace. According to

neuropsychological and neurophysiological data, disengagement and engagement seem related to different parts of the brain.

Finally, another important attentional operation is the 'inhibition of return.' Using the spatial cueing paradigm, it has been demonstrated that a location in the space is inhibited 500 ms or more after being activated by orienting attention to it. This operation has been generalized to other contexts and could represent some useful and adaptive device (imagine the case of an animal attending exclusively to eating, while being stalked by a predator).

3. Conclusion

Research in attention has led to the description of a hierarchical organization of processing resources allocated to information, with a central processor elaborating attentional strategies. Many experiments have been conducted to describe the operations of attention, and models developed in cognitive psychology are also helpful in explaining brain mechanisms, and in understanding the performance of patients suffering from brain lesions.

Nonetheless, the 'resource' approach has been criticized because it does not resume the concept of attention. In other words, the management of the processing resources in the brain may not be the only way for attentional operations to provide adaptive advantages within a complex environment. Attention can also be viewed as the cause of the coherence as well as the flexibility of an individual's behavior. Coherent and flexible behavior supposes the integration, comparison, and evaluation between an intention, determined by a person's motivated goals (represented in our working memory), and the actual modifications occurring in the environment. By inhibiting an immediate response to the environment, or by activating a piece of information to a higher level of processing, attention provides the time necessary for reflection. Consequently, nonroutinized strategies can be activated when someone faces a challenging problem. Also, sustaining attention may represent the condition for pleasurable activities like tasting food, listening to music, and so forth.

See also: Attention and Action; Attention: Models; Attention, Neural Basis of; Dual Task Performance; Modularity versus Interactive Processing, Psychology of

Bibliography

Baddeley A, Weiskrantz L (eds.) 1993 *Attention: Selection, Awareness and Control. A Tribute to Donald Broadbent*. Oxford University Press, Oxford, UK

Broadbent D 1958 *Perception and Communication*. Pergamon, London

Inui T, McClelland J L (eds.) 1996 *Attention and Performance, XVI: Information Integration in Perception and Communication*. MIT Press, Cambridge, MA

Kahneman D 1973 *Attention and Effort*. Prentice-Hall, Englewood Cliffs, NJ

LaBerge D 1995 *Attentional Processing: The Brain's Art of Mindfulness*. Harvard University Press, Cambridge, MA

Neisser U 1976 *Cognition and Reality: Principles and Implications of Cognitive Psychology*. Freeman, New York

Parasuraman R (ed.) 1998 *The Attentive Brain*. MIT Press, Cambridge, MA

Parasuraman R, Davies D R (eds.) 1984 *Varieties of Attention*. Academic Press, New York

Pashler H E (ed.) 1998 *Attention*. Psychology Press

Pashler H E 1998 *The Psychology of Attention*. MIT Press, Cambridge, MA

Posner M I 1978 *Chronometric Explorations of Mind*. Oxford University Press, Oxford, UK

Posner M I, Raichle M E 1994 *Images of Mind*. Freeman, New York

E. Siéroff

Attention, Neural Basis of

The flow of stimuli that is constantly available to our sense organs is far greater than our brains can handle. Studies in cognitive psychology have clearly established that our perceptual system is severely limited in its capacity to fully process and identify more than one object at a time. Moreover, our response system is limited inherently in its ability to perform concurrent actions on a large number of objects: we can only gaze at one object at a time, and reach at most to two objects. To cope with the flood of sensory inputs and achieve coordinated, purposeful behavior a powerful mechanism is available to the brain: selective attention. This term defines the mental ability to select the most important aspects of the sensory scene and filter out those that are irrelevant or potentially interfering. Perceptual experience of the sensory world critically depends on attention. Selected objects can enter awareness, guide our actions, and undergo memory storage, whereas ignored objects are often left at the margins of our consciousness.

For several decades, selective attention has been studied at the psychological level. More recently, however, behavioral research has been strengthened by a variety of neuroscience approaches as diverse as single-neuron recordings in monkeys, functional neuroimaging and event-related potentials in humans, and lesion studies in neurological patients. This interdisciplinary and convergent effort is beginning to reveal how the brain activity is regulated through attention. This article will focus mainly on selective visual attention and review some recent advances in understanding the neural underpinnings and mechanisms of this cognitive function.

1. Multiple Visual Areas

Contemporary neuroscience has revealed that the neural representations of visual objects are highly distributed. Visual inputs activate almost simultaneously a variety of cortical and subcortical regions within the brain. Among visually responsive structures are 30 or more separate cortical areas, which appear to be organised within two major parallel pathways, or streams (Ungerleider and Mishkin 1982). Each visual stream begins with the primary or striate visual cortex (area V1) and involves multiple areas beyond V1. One of the pathways, the ventral stream, leads toward the inferior temporal (IT) cortex and is crucial for object perception and recognition (the 'what' pathway). The other pathway, the dorsal stream, is directed into the posterior parietal cortex and is mostly concerned with spatial localization and action (the 'where' or 'how' pathway). Physiologically, the processing of visual information appears largely hierarchical within each cortical pathway. The latency of visual responses, the size of the receptive field, and the complexity of neuronal tuning properties increase steadily as one proceeds from one area to the next in the processing hierarchy. Anatomically, two types of connections between areas can be identified: (a) feed-forward connections that relay information from lower to higher visual areas in each pathway and (b) feed-back connections that provide recurrent signals from higher to lower levels of processing. As described below, feedback, re-entrant connections from parietal and frontal areas may form the anatomical basis of top-down attentional signals that modulate activity elsewhere in the visual system.

2. How Does Visual Selective Attention Work?

Traditionally, visual attention has been viewed as a mental spotlight or zoom lens that illuminates stimuli in selected regions of space, thus easing their access to higher processing centers (Posner 1980). According to this classical conception, visual attention is inherently spatial in nature; objects are found by rapidly shifting the attentional spotlight from one location to another. The canonical view also poses that attention is mainly controlled by a unified, topographical 'master' map located in the parietal cortex and associated spatial processing structures. However, several recent findings suggest that any effect of attention is best explained in terms of competitive interactions among the neural representations of objects present in the visual field, as outlined by the 'biased competition' model of visual attention (Desimone and Duncan 1995). According to this model, inputs compete with one another at all levels of the visual system, and different mechanisms can tip the competitive balance in favor of one stimulus' representation over others.

At the behavioral level, competition is reflected in the loss of accuracy or slowing in reaction time that result when subjects are asked to identify multiple stimuli at once, rather than a single object. At the physiological level, single-unit recording studies in monkeys have shown that responses to two stimuli in the receptive field of visual neurons are smaller than responses to either stimulus alone (Reynolds et al. 1999). This indicates that multiple stimuli in the visual field are not processed separately from each other but rather they interact in an inhibitory, suppressive fashion. In agreement with this result, a recent study using functional magnetic resonance imaging (fMRI) in humans (Kastner et al. 1998) has shown that neural responses from visual areas are lower when stimuli are presented simultaneously than when identical stimuli are presented one after the other in the same locations. The reduced activation in the simultaneous condition may reflect sensory suppression among multiple concurrent visual stimuli.

Competitive interference between stimuli has been especially observed in ventral stream areas. However, there is also evidence for sensory competition in structure associated with the dorsal stream, for example, in the frontal eye field and in areas MT and MST. Here, cells respond best to stimuli that move in a particular 'preferred' direction across the receptive field, but respond weakly or not at all to stimuli moving in the opposite or 'null' direction. Treue and Maunsell (1996) found that the response to an otherwise optimal stimulus (i.e., moving in the cell's preferred direction), placed within the receptive field of a MT neuron, was critically reduced by the addition of a second, nonpreferred stimulus at a different location in the same receptive field.

These competitive interactions between neuronal populations encoding different visual objects can be modulated by virtue of bottom-up and top-down biases, such that neurons encoding one object win out over the competition and that object receives further processing in the brain. Bottom-up factors refer to largely automatic and unconscious processes that operate early in vision and presumably contribute to perceptual pop-out phenomenon—the quick and effortless detection of a unique item on a homogeneous background. Bottom-up biases are thought to be mediated by hard-wired neural mechanisms, such as the center–surround structure of receptive fields found in many areas of the visual system. However, competition among visual neurons can also be biased by top-down cognitive influences, in which objects that meet the requirements of the current task are voluntarily selected for focal attention. For example, top-down biases allow you on one occasion to search efficiently for your mug on a desk cluttered with many objects, while on another occasion you might search the same complex visual scene for your pencil. Top-down biases, as the name implies, are imposed on low-level sensory processes by higher-level control mechanisms, and are probably mediated by extensive feedback connections from higher areas to low-level

sensory areas. It is to these processes and their contribution to resolve competition between stimuli that the article will now turn.

3. What is Selected by Attention?

In contrast to the spotlight model of visual attention, a central tenet of the biased-competition model is that the feedback bias from higher areas is not purely spatial. Depending on the context, different visual properties—spatial location, color, shape, or motion—can be used to direct attention and assign limited processing resources to a particular object. Some psychologists have arrived at the notion of attentional template, a flexible, short-term description, formed to suits the task at hand, that can be used to give a competitive advantage to visual inputs matching that description. The attentional template can specify any property of the relevant object. Spatial location is, therefore, only one among many features that can contribute to the attentional template. For example, consider a subject that is asked to search for and report numbers colored in red and to disregard numbers in other colors. Following this instruction, neurons selective for red inputs are primed or preactivated in visual areas coding colors, such that a stimulus that possesses the behaviorally relevant color is processed preferentially over others.

3.1 Space-based Selection

In the past, much research on the control of visual attention has focused on selection in space. In most behavioral studies, subjects covertly (i.e., without executing eye movement) direct their attention onto stimuli shown at a specific spatial location in the visual field, whereas distracting information is present at the other locations. Performance typically is enhanced at attended location, and impaired at the distractor locations, suggesting that competition for focal capacity is biased towards objects at the attended region of space. Single-cell recording work in monkeys has made a substantial contribution to our understanding of space-based selection mechanisms. A host of studies have reported that attention can increase the neural response to a visual stimulus when it appears at an attended location compared with when the animal attends to a different stimulus location. This selective enhancement effect, which increases with task difficulty and with the amount of visual clutter outside of the receptive field, has been demonstrated in numerous visual structures. These include primary visual cortex, areas V2 and V4, but especially in dorsal stream areas implicated in visuomotor control, such as the lateral intraparietal area, the frontal eye field, and the dorsolateral prefrontal cortex (Colby and Goldberg 1998). In humans, brain imaging and event-related potential (ERP) studies have shown similar results. In these experiments, identical stimuli were presented on either side of fixation, while normal subjects focused attention to the left or right visual field. Results showed enhanced activity in extrastriate visual areas, generally lateralized to the hemisphere contralateral to the attended stimulus, in agreement with the fact that many visual areas have a predominantly crossed representation of the visual field (Kastner and Ungerleider 2000). Overall, these findings suggest that top-down biases operate by increasing the input strength for stimuli at the attended locations. Most notably, however, these studies generally failed to find any effect of spatial attention on activity in human V1 (see below).

An alternative, not exclusive, possibility, however, is that spatial attention filters out distracting information when multiple stimuli appear simultaneously within the cell's receptive field. Studies in extrastriate ventral areas V2 and V4, but also in dorsal areas MT and MST, have shown that when monkeys attend to a stimulus at one location within the RF, competitive interference from stimuli at ignored locations within the RF is greatly attenuated or suppressed (Moran and Desimone 1985). Apparently, the cell receptive field 'shrinks' around the attended location, with unattended stimuli being filtered out. In addition, the effects of attention are much weaker when either the target or the distractor is placed outside the RF. Indeed, in such a configuration stimuli are no longer competing for the cell's response, thus reducing the need of attentional biases. The suppressive mechanism of attention may explain why we have little awareness of unattended stimuli, and is highly consistent with the idea that spatial attention biases the sensory interactions among neurons, causing them to respond primarily to the attended stimulus.

In accordance with monkey neurophysiology, one fMRI study in humans showed that directing attention to one object in a cluttered visual scene reduces the sensory interference exerted by distracting stimuli at nearby locations (Kastner et al. 1998). Although attention can suppress irrelevant stimuli, it does not imply that it always does. According to one theory of attention, the extent to which spatial attention inhibits irrelevant information depends on the processing load imposed by the primary task, with stronger inhibition occurring when processing load is high. Consistent with this, an fMRI study by Rees et al. (1997) showed that neural responses to irrelevant moving stimuli, as registered in the motion-processing visual areas MT and MST, are much weaker when the primary task is difficult than when it is easy.

3.2 Object-based Selection

Although spatially directed attention has proven to be an excellent model system for investigating attentional

mechanism, objects can be also selected by attention based on nonspatial properties. For example, color can be used efficiently to control attention, particularly when there is a large hue difference between target and distractors. Neurophysiological studies in monkeys have considerably extended our understanding of feature-based processes for directing visual attention. In a single-neuron study in IT cortex, monkeys were instructed to perform a simple visual search task (Chelazzi et al. 1993). On each trial, a central cue was presented, followed by a blank delay period. At the end of the delay, an array of stimuli was presented and the monkey was rewarded to make a saccadic eye movement toward the stimulus matching the cue (target). The location of the target was not known in advance and thus selection could not be based on spatial information. Just like attention to location, attention to an object (non-spatial) feature was found to inhibit neuronal responses to irrelevant stimuli. Neurons in IT cortex showed an initial response if the choice array contained their preferred stimulus, without regard of the target identity. When the preferred stimulus was the eye movement target, the initial response remained high. However, when the preferred stimulus was an irrelevant, distracting stimulus in the array, the initial response was rapidly inhibited, long in advance of the onset of the saccade. Interestingly, neurons selective for the cue showed higher sustained activity during the delay. It has been argued that this tonic activity during the interval represents the neural correlate of the attentional template, which can be used to bias competition in favor of cells representing the target. Other neurophysiological studies also reported attentional modulation based on object features. Interestingly, studies that have used a single stimulus within the cell receptive field found facilitation of attended stimuli as well as inhibition of unattended stimuli.

Functional neuroimaging studies provided additional insight into feature selection. In a classic PET study (Corbetta et al. 1991), subjects compared two successive visual displays, with reference to the color, shape, or speed of motion. Depending on the relevant visual feature, different extrastriate areas were active: attention to color led to enhanced visual response in area MT, whereas attention to shape and motion color induced response enhancement in area V4.

Similar results have been reported using ERPs and fMRI. Together, these studies argue that selective attention to a particular visual feature enhances neural activity in areas selectively specialised for that attribute. Attention selection can also occur from whole object representations. In a recent fMRI study (O'Craven et al. 1999), two stimuli, a face and a house, were superimposed transparently in the same location. In each trial, either the face or the house oscillated back and forth. Subjects' attention was directed in different conditions to the face, the house, the direction of motion, or the spatial location of the stationary stimulus. As expected, different brain regions were activated depending on whether the relevant feature was the face (fusiform face area), the house (parahippocampal place area), or the direction of motion (areas MT and MST). Importantly, however, results showed that attention to one feature of an object enhanced the representation of other features of the same object, even when these features were completely irrelevant for the task. These findings strongly suggest that attention selects entire objects, rather than selecting simple visual attributes.

4. *The Anatomical Locus of Attentional Selection*

Historically, the parietal lobe has been viewed as central in visual selection. Damage to this region commonly produces unilateral neglect, a deficit in perceiving and interacting with events on the contralesional side of space, which has often been explained as a difficulty in directing attention to that side. Patients with left neglect, for example, fail to complete the left half drawings, or to read the left side of words or passages of text. The attentional nature of neglect is well shown by the observation that the deficit is greatly exaggerated by competing inputs on the intact, ipsilesional side, a phenomenon termed extinction. In fact, recent evidence based on large series studies indicates that neglect and extinction may follow a wide variety of unilateral lesions, both cortical and subcortical. This indicates a multiplicity of competitive systems modulated by attention and behavioral relevance, rather than a single anatomical level at which attention affects processing.

Data from lesion studies in monkeys are consistent with this conclusion. Indeed, different methodologies, such as single-unit recording, ERPs, functional imaging, and lesion studies provide converging evidence that attention can affect neuronal activity in most of the extrastriate visual areas where such modulation has been investigated. Paradoxically, many electrophysiological reports suggested that the processing of attended and unattended items does not differ in primary visual cortex. Monkey physiology in the past yielded conflicting results in this regard. In some studies approximately 30 percent of V1 neurons revealed some form of attentional modulation. More recently, however, Roelfsema et al. (1998) trained monkeys on an ingenious task in which animals had to trace one of two curves—the attended curve—while maintaining fixation. They found that the firing rate of V1 neurons was significantly higher when the receptive field was on the curve that needed to be traced, than on the distracting curve. The long latency of the enhancement (about 200 ms) suggested that this effect might be mediated by feedback processing.

Recent functional magnetic resonance imaging studies in humans have also shown that spatial attention is indeed capable of modulating neuronal

activity in V1. One study compared the magnetic resonance signals in central and peripheral parts of V1 while subjects were either attending to a rapidly presented series of letters in the center of the visual field or performing a direction-of-rotation discrimination task on a peripheral rotating disk (Somers et al. 1999). In striking contrast with previous imaging studies, substantial increases in activity were found in V1 regions corresponding to the attended stimulus and decreases in regions representing the nonattended stimulus. Together these studies demonstrated that, depending on the visual task, attentional modulation of responses does indeed occur in the primary visual cortex. For the region of V1 representing the attended part of the visual field, neuronal responsiveness increases; conversely, when attention is directed away from a particular location there appears to be a consequent decrease in responsiveness in the retinotopically corresponding V1 region.

How can these recent findings be reconciled with previous failure to find attentional modulation in V1? One explanation is that, in the earlier studies, the tasks were not sufficiently demanding. In the new studies, stimulus parameters were chosen so that tasks were difficult and only 70–90 percent correct responses were achieved. Alternatively, it is possible that some techniques, like ERP recordings, may reflect only initial, feed-forward processing in V1. Apparently, the early part of the visual response of V1 neurons is unaffected by attention, as single-unit studies have shown. In contrast, the relatively poor time resolution of fMRI measurements would capture both early feed-forward processing and later top-down attentional influences from higher visual areas.

5. *The Neural Mechanisms of Selection*

Do attentional signals affect visual processing by facilitation of attended stimuli or by inhibition of irrelevant distractors, or both? One view suggests that attention exerts its influence by inhibiting unwanted stimuli early in visual processing, and by enhancing relevant information later on. The assumption underlying this view is that neurons in early visual areas respond at close to maximal rates to visual stimuli even during passive viewing, and therefore there is little room for response enhancement in early stages. In contrast, in higher visual areas, cells become less stimulus-driven and more sensitive to behavioral relevance. Overall, recent single-unit and neuroimaging data do not support such a suggestion, and argue instead for a close interplay between suppressive and enhancing mechanisms at all levels of the visual system. For example, one recent fMRI study reported that attention to a particular spatial location, in addition to an enhanced response to the attended visual stimulus, results in a widespread suppression of activity at all other locations (Smith et al. 2000). However, ERP studies have shown that the 'costs' of spatial cueing are associated with suppression of unattended inputs at an early stage, whereas the 'benefits' of cueing are associated with enhancement of attended stimuli at a later stage (Hillyard and Anllo-Vento 1998). Thus, there is at least some suggestive evidence that suppression and enhancement may entail qualitatively different mechanisms, which operate at different levels of the visual pathways.

Top-down attentional signals may also affect visual processing by increasing the spontaneous (baseline) level of neural activity in the absence of any visual stimulation. Single-cell recording studies have clearly shown a substantial elevation of baseline activity for neurons in area V2 and V4, when the monkey was instructed to attend covertly to a location within the receptive field, even before the stimulus was presented there (Luck et al. 1997). The shift in baseline activity during spatial attention closely resembles the sustained, tonic activation found in IT neurons during visual search experiments. This increase in spontaneous firing rate is consistent with the idea of top-down feedback signals priming neurons coding the attended location of space. In a similar vein, several fMRI studies in humans have shown that during the expectation period preceding the arrival of a stimulus at an attended location, baseline activity increased in numerous extrastriate visual areas. The effect is spatially selective, as it occurs only in subregions of visual areas with a representation of the attended location; thus, it cannot be due to nonspecific arousal. Remarkably, an increase in baseline activity has been found even in the primary visual cortex during a demanding threshold detection task, and its magnitude was highly predictive of the subject's behavioral performance (Ress et al. 2000).

6. *The Control of Visual Attention*

A central issue in the current debate on selective visual attention concerns the brain sources of the top-down attentional signals to extrastriate and striate visual areas. Studies of brain-damaged patients, single-neuron recordings in primates and more recently blood-flow neuroimaging methods have identified a network of interconnected cortical and subcortical areas that play a key role in the control of selective visual attention. These areas include the dorsolateral prefrontal and posterior parietal cortex, the anterior cingulated gyrus, and the pulvinar nucleus of the thalamus. It has been proposed that in this circuitry, the prefrontal cortex would initiate and maintain the selective sensory bias in working memory, whereas the parietal cortex would direct and shift attention to specific objects and locations in space. In turn, this control network would enhance or suppress stimulus

representations in the extrastriate visual areas according to current concern. Parietal and prefrontal cortex have anatomical and functional properties ideal for providing attentional templates. Both regions are extensively and reciprocally interconnected with virtually all visual cortical areas, and are therefore in a position to influence visual processing at multiple levels of the visual system. Neurons in the prefrontal and parietal cortex are capable of encoding the behavioral relevance or saliency of a visual stimulus. Furthermore, they display a pattern of activity that is well suited to form and maintain on line a representation of an object or of a location for as long as that representation is relevant to behavior. In the past years, several functional neuroimaging studies in humans have shown frontal and parietal activations during visual attention tasks (see Kastner and Ungerleider 2000, for a recent review). However, only recently top-down control signals for allocating attention have been distinguished from the modulatory effects that these signals have on the processing of incoming visual stimuli. For example, Hopfinger et al. (2000) asked which brain regions were activated in response to cues instructing shift or attention, and which were activated by the subsequent visual target presentation. A number of different visual areas were active; the surprising finding was that there was relatively little overlap between the two sets of visual responses. Importantly, posterior parietal and prefrontal areas were selectively engaged by attention-directing cues, thus suggesting that these two structures play a fundamental role in top-down attentional control. It should be noted, however, that the fronto-parietal network is not exclusively related to visual attention, but may coincide or overlap with cortical regions involved in oculomotor processing. This overlap supports the hypothesis that the neuronal signals used to direct attention correspond to the premotor signals that drive the eyes. One way to disentangle attentional and oculomotor processes would be an attentional task that does not require eye movements. In one study, for instance, subjects performed rapid attention shifting between the shape and color of stimuli presented at fixation, so that the need for eye movements was eliminated (Le et al. 1998). Nevertheless, significant activations were observed in superior parietal lobule and other occipital visual areas, as well as the lateral cerebellum. Thus, it seems that at least some activity in the parietal and frontal cortex reflects attentional signals that are independent of eye movement. The frontoparietal circuit may therefore play a very general role in visual attention, governing the access of competing visual representations to higher brain centers and ultimately to perceptual consciousness.

See also: Arousal, Neural Basis of; Attention and Action; Attention-deficit/Hyperactivity Disorder (ADHD); Attention: Models; Attention: Multiple Resources; Brain, Evolution of; Conscious and Unconscious Processes in Cognition; Consciousness, Cognitive Psychology of; Consciousness, Neural Basis of

Bibliography

Chelazzi L, Miller E K, Duncan J, Desimone R 1993 A neural basis for visual search in inferior temporal cortex. *Nature* **363**: 345–47

Colby C L, Goldberg M E 1998 Space and attention in parietal cortex. *Annual Review of Neuroscience* **22**: 319–49

Corbetta M, Miezin F M, Dobmeyer S, Shulman G L, Petersen S E 1991 Attentional modulation of neural processing of shape, color and velocity in humans. *Science* **248**: 1556–9

Desimone R, Duncan J 1995 Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* **18**: 193–222

Hillyard S A, Anllo-Vento L 1998 Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Science USA* **95**: 781–7

Hopfinger J B, Buonocore M H, Mangun G R 2000 The neural mechanism of top-down attentional control. *Nature Neuroscience* **3**: 284–91

Kastner S, De Weerd P, Desimone R, Ungerleider L G 1998 Mechanisms of directed attention in human extrastriate cortex as revealed by functional MRI. *Science* **282**: 108–11

Kastner S, Ungerleider L G 2000 Mechanisms of visual attention in human cortex. *Annual Review of Neuroscience* **23**: 315–41

Le T H, Pardo J V, Hu X 1998 4 T-fMRI study of nonspatial shifting of selective attention: cerebellar and parietal contributions. *Journal of Neurophysiology* **79**: 24–42

Luck S J, Chelazzi L, Hillyard S A, Desimone R 1997 Neural mechanism of spatial selective attention in areas V1, V2 and V4 of macaque visual cortex. *Journal of Neurophysiology* **77**: 24–42

Moran J, Desimone R 1985 Selective attention gates visual processing in extrastriate cortex. *Science* **229**: 782–4

O'Craven K M, Downing P, Kanwisher N 1999 fMRI evidence for objects as the units of attentional selection. *Nature* **401**: 584–7

Posner M 1980 Orienting of attention, *Quarterly Journal of Experimental Psychology* **32**: 3–25

Rees G, Frith C D, Lavie N 1997 Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science* **278**: 1616–9

Ress D, Benjamin T, Backus B T, Heeger D J 2000 Activity in primary visual cortex predicts performance in a visual detection task. *Nature Neuroscience* **3**: 940–5

Reynolds J H, Chelazzi L, Desimone R 1999 Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience* **19**: 1736–53

Roelfsema P R, Lamme V A, Spekreijse H 1998 Object-based attention in the primary visual cortex of the macaque monkey. *Nature* **395**: 376–81

Smith A T, Singh K D, Greenlee M W 2000 Attentional suppression of activity in the human visual cortex. *NeuroReport* **7**: 271–7

Somers D C, Dale A M, Seifert A E, Tootell R B H 1999 Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Science USA* **96**: 1663–8

Treue S, Maunsell J H R 1996 Attentional modulation of visual motion processing gain in macaque visual cortex. *Nature* **399**: 575–79

Ungerleider L G, Mishkin M 1982 Two cortical visual systems. In: Ingle D J, Goodale M A, Mansfield R J W (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA, pp. S49–S86

G. di Pellegrino

Attitude Change: Psychological

Attitudes refer to people's global and relatively enduring (i.e., stored in long-term memory) evaluations of objects, issues, or persons (e.g., I dislike chocolate; I'm opposed to the governor's tax policy). Numerous procedures have been developed to modify these evaluations with some change techniques involving considerable thinking about the attitude object and some requiring little. Attitudes are one of the most studied and important constructs in psychology because of the critical role of attitudes in guiding behavior (see *Attitudes and Behavior*).

1. Overview of Attitudes and Attitude Change

Attitudes are based on some combination of cognitive, behavioral, and affective influences, and are typically measured by self-report scales such as the 'semantic differential,' where a person rates the target on bipolar evaluative dimensions such as how good/bad or favorable/unfavorable it is. Increasingly, researchers have appreciated that it is also useful to assess attitudes on dimensions other than their valence, such as their accessibility (how quickly the attitude comes to mind) and ambivalence (how consistent the basis of the attitude is). These indicators of attitude 'strength' are useful in determining which attitudes are consequential and which are not. Strong attitudes are those that persist over time, are resistant to change, and predict other judgments and actions (Petty and Krosnick 1995). At any given moment, one's expressed evaluation can be influenced by a variety of contextual factors, but the common assumption is that one's core 'attitude' is the underlying evaluation that is capable of guiding behavior (one's actions), cognition (one's thoughts and memories), and affect (emotional reactions).

Attitude change occurs when one's core evaluation shifts from one meaningful value to another, and is typically inferred from a change in a person's scale rating, although behavioral and other indirect or implicit procedures for assessing change are sometimes used. Most studies of attitude change involve exposing individuals to a persuasive communication of some sort but, as noted below, some attitude change techniques do not involve exposure to any message. The earliest work on attitude change attempted to examine which variables and procedures increased and which decreased the likelihood of change (e.g., did more change occur when the message source was described as an expert than when the source lacked expertise even though the message was the same?).

2. Two Routes to Attitude Change

After numerous studies, the accumulated evidence suggests that even the simplest variables (e.g., being in a positive mood) sometimes increase, sometimes decrease, and sometimes have no impact on the likelihood that a person's attitude will change. Numerous theories and psychological processes have been proposed to account for these divergent results. Even though the many different theories of attitude change have different names, postulates, and particular effects and variables that they specialize in explaining, the many different theories of attitude change can be thought of as emphasizing just two relatively distinct 'routes to persuasion' (Petty and Cacioppo 1981). The first change technique, persuasion via the central route, focuses on the information that a person has about the central merits of the object under consideration. Some of the central route approaches postulate that comprehending and learning the information presented is critical for attitude change, whereas others focus more on the evaluation, elaboration, and integration of this information. In contrast, the peripheral route approaches emphasize attitude changes that are brought about without much thinking about information central to the merits of the attitude issue. Thus, the peripheral approaches deal with changes resulting from rewards, punishments, and affective experiences that are associated directly with the attitude object, or simple inferences that people draw about the appropriate attitude to adopt based on their own behavior or other simple cues in the persuasion environment. For example, a person might be more persuaded by a message containing nine rather than three arguments because each of the arguments is evaluated and determined to be compelling (central route), or because the person simply counts the arguments and reasons, 'the more the better' (peripheral route). Before integrating these approaches, it is useful to describe the major central and peripheral processes responsible for persuasion.

3. Central Route Approaches to Attitude Change

3.1 Message Learning Approach

One of the most influential programs of research on attitude change was that undertaken by Carl Hovland

and his colleagues at Yale University in the years following World War II (e.g., Hovland et al. 1953). The Yale group studied how source, message, recipient, and channel factors affected the comprehension, acceptance, and retention of the arguments in a persuasive communication. Although no formal theory tied together the many experiments conducted by this group, they often attempted to explain the results obtained in terms of general learning principles such as the more message content you learned, the more your attitudes should change (see McGuire 1985). Contemporary research shows that people can be persuaded without learning or remembering any of the message content. That is, people are sometimes persuaded solely by the 'cues' associated with the message (e.g., the source is expert; Petty and Cacioppo 1986). Or, the message might elicit a favorable thought that persists in the absence of memory for the information that provoked it (Greenwald 1968). Message learning appears to be most important when people are not engaged in an on-line evaluation of the information presented to them, such as when they do not think they have to form an opinion at the time of information exposure. In such cases, subsequent attitudes may be dependent on the valence of information they have learned and can recall (Hastie and Park 1986).

3.2 Self-persuasion Approach

Self-persuasion theories hold that people's attitudes can change in the absence of any new external information. This is because people can self-generate reasons to favor or disfavor any position. The powerful and persisting effects of completely self-generated messages were shown in early research on 'role-playing' where people were asked to generate messages on certain topics (e.g., the dangers of smoking). The subsequent attitudes of these people were compared to those who had either passively listened to the communication or who had received no message. A consistent result was that active generation of a message was a successful strategy for producing attitude change, and these changes persisted longer than changes based on passive exposure to a communication. Finally, merely asking someone to think about an issue, object, or person can lead to attitude change as a result of the evaluative thoughts generated. Cognitive response theorists hold that just as one's thoughts can produce change in the absence of a message, so too are one's own thoughts responsible for attitude change even when a persuasive message is presented. That is, to the extent that a person's thoughts in response to the message are favorable, persuasion should result, but to the extent that they are unfavorable (e.g., counterarguments), resistance or even boomerang are more likely. These theorists hold that persistence of persuasion depends upon the decay function for cognitive responses rather than message arguments *per se* (see Petty et al. 1981, for review).

3.3 Expectancy-value Approach

The message learning and self-persuasion approaches focus on the information (either externally or internally generated) that is responsible for persuasion. Neither approach has much to say about the particular features of the information that are critical for influencing attitudes. In contrast, expectancy-value theorists analyze attitudes by focusing on the extent to which people expect the attitude issue to be related to important values or produce positive and negative consequences. In one influential expectancy-value model, Fishbein and Ajzen's theory of reasoned action (1975) holds that the attributes (or consequences) associated with an attitude object are evaluated along two dimensions—the likelihood that an attribute or consequence is associated with the object, and the desirability of that attribute or consequence. If a persuasive message says that raising taxes will lead to reduced crime, the effectiveness of this argument should depend on how likely people think it is that crime will be reduced if taxes are increased (likelihood), and how favorably they view the outcome of reducing crime (desirability). Although some questions have been raised about the necessity of one or the other of these components, a large body of research supports the idea that attitudes are more favorable the more that likely–desirable consequences (or attributes) and unlikely–undesirable consequences are associated with them. The major implication of this theory for persuasion is that a message will produce attitude change to the extent that it introduces new attributes of an object, or produces a change in the likelihood and/or the desirability components of an already accepted attribute. Another proposition of this theory is that the items of information constituting an attitude are combined in an additive fashion. Other theorists, however, have contended that an averaging mechanism is more appropriate (see Anderson 1971).

3.4 Functional Approach

In their expectancy-value theory, Fishbein and Ajzen speculate that five to seven attributes or consequences are critical in determining a person's overall attitude. It is not clear, however, which particular attributes will be the most important (i.e., how the attributes are weighted). Functional theories of persuasion focus on the specific needs or functions that attitudes serve for a person and are therefore relevant for understanding the underlying dimensions of the attitude that are

most important to influence (e.g., Smith et al. 1956). For example, some attitudes are postulated to protect people from threatening truths about themselves or to enhance their own self-image ('ego-defensive function'), others give expression to important values ('value-expressive function'), or help people to understand the world around them ('knowledge function') or facilitate achieving rewards and avoiding punishments ('utilitarian function'). According to these theories, change depends on challenging the underlying functional basis of the attitude. Thus, if a person dislikes lowering taxes because of concern about social inequality (value expressive function), an argument about the amount of money the taxpayer will save (utilitarian function) will be ineffective.

3.5 Consistency Approach

Just as functional theories hold that attitudes serve important needs for individuals, dissonance and related theories hold that attitudes are often in the service of maintaining a need for consistency among the elements in a cognitive system (Festinger 1957). In Festinger's original formulation of dissonance theory, two elements in a cognitive system (e.g., a belief and an attitude; an attitude and a behavior) were said to be consonant if one followed from the other, and dissonant if one implied the opposite of the other. Two elements could also be irrelevant to each other. One of the more interesting dissonance situations occurs when a person's behavior is in conflict with his or her attitudes or beliefs because behavior is usually difficult to undo. According to the theory, dissonance, experienced as an aversive tension, may be reduced by changing beliefs and attitudes to bring them into line with the behavior. Thus, if you were opposed to the election of Candidate Smith, it would be inconsistent to sign a petition in favor of this candidate. According to dissonance theory, signing such a petition would produce discomfort that could result in a more favorable evaluation of the candidate in an effort to restore consistency.

Although early dissonance research was generally supportive of the theory, several competing formulations were proposed. Although it is now clear that many of the behaviors described by Festinger induce in people an 'unpleasant tension,' just as the theory predicts, current research has begun to focus more on understanding the precise cause of that tension. For example, some have questioned Festinger's view that inconsistency *per se* produces tension in many people. Rather, some argue that people must believe that by their behavior they have freely chosen to bring about some foreseeable negative consequence, or that the inconsistency involves a critical aspect of oneself or a threat to one's positive self-concept (see Harmon-Jones and Mills 1999).

4. Peripheral Route Approaches

Each of the central route approaches described above assumes that attitude change results from people actively considering the merits of some position either in a fairly objective manner, or in a biased way (such as when seeking to restore consistency). The next group of theories does not share this assumption. Instead, these theories suggest that people often prefer to conserve their cognitive resources and form or change attitudes with relatively little cognitive effort.

4.1 Inference Approaches

Rather than effortfully examining all of the issue-relevant information available, people can make an evaluative inference based on some meaningful subset of information. One popular inference approach is based on 'attribution theory' and holds that people come to infer underlying characteristics about themselves and others from the behaviors that they observe and the situational constraints imposed on these behaviors (e.g., Kelley 1967). Bem (1965) suggested that people sometimes have no special knowledge of their own internal states and simply infer their attitudes in a manner similar to that by which they infer the attitudes of others. In his self-perception theory, Bem reasoned that just as people assume that the behavior of others and the context in which it occurs provides information about the presumed attitudes of these people, so too would a person's own behavior provide information about the person's own attitude. Thus, a person might reason, 'since I signed Candidate Smith's petition, I must be in favor of her election.'

The attribution approach has also been useful in understanding the persuasion consequences of making inferences about relatively simple cues. For example, when external incentives (e.g., money) provide a salient explanation for a speaker's advocacy ('he was paid to say it'), the message is less effective than when a discounting external attribution is not possible. Research indicates that these simple attribution processes are most likely to influence attitudes when people are relatively unmotivated or unable to think carefully about the issue, such as when they have relatively little knowledge on the topic and the issue has few anticipated personal consequences.

Like the attributional framework, the heuristic–systematic model of persuasion postulates that when people are not motivated or able to process all of the relevant information available, attitude change can result from the use of certain heuristics or rules of thumb that people have learned on the basis of past experience and observation (see Eagly and Chaiken 1993). To the extent that various persuasion heuristics are available in memory, they may be invoked to evaluate persuasive communications. For example,

either because of prior personal experience or explicit training, people may evaluate a message with many arguments by invoking the heuristic 'the more arguments, the more valid it is.' If so, no effortful learning or evaluation of the actual arguments presented is necessary for influence to occur.

4.2 Approaches Emphasizing Affect

The attribution and heuristic models focus on simple cognitive inferences that can modify attitudes. Other peripheral route theories emphasize the role of affective processes in attitude change. One of the most direct means of associating 'affect' with objects, issues, or people is through classical conditioning (e.g., see Staats and Staats 1958). In brief, conditioning occurs when an initially neutral stimulus (the conditioned stimulus; CS) is associated with another stimulus (the unconditioned stimulus; UCS) that is connected directly or through prior conditioning to some response (the unconditioned response; UCR). By pairing the UCS with the CS, the CS becomes able to elicit a conditioned response (CR) that is similar to the UCR (Pavlov 1927). So, when food is paired over and over again with a bell, eventually the bell elicits salivation in the absence of food.

Considerable research has shown that attitudes can be influenced by pairing initially neutral objects with stimuli about which people already feel positively or negatively. For example, peoples' evaluations of words, other people, political slogans, consumer products, and persuasive communications have been modified by pairing them with such affect-producing stimuli as unpleasant odors and temperatures, the onset and offset of electric shock, harsh sounds, and elating and depressing films. People are especially susceptible to the simple transfer of affect from one stimulus to another when the likelihood of object-relevant thinking is rather low.

Another procedure for modifying attitudes through simple affective means was identified by Zajonc (1968) in his work on mere exposure. In this research, Zajonc and his colleagues showed consistently that when objects are presented to an individual on repeated occasions, the mere exposure is capable of making the individuals' attitudes toward these objects more positive. Recent work on this phenomenon indicates that simple repetition of objects can lead to more positive evaluations even when people do not recognize that the objects are familiar. Mere exposure effects have been shown in a number of studies using a variety of stimuli such as polygons, tones, nonsense syllables, Chinese ideograms, photographs of faces, and foreign words. Interestingly, what these stimuli have in common is that they tend to be meaningless and are relatively unlikely to elicit spontaneous thought. In fact, the simple affective process induced by mere exposure appears to be more successful in influencing attitudes when processing of the repeated stimuli is minimal. When more meaningful stimuli have been repeated such as words or sentences, mere exposure effects have been less common. Instead, when processing occurs with repetition, the increased exposures enhance the dominant cognitive response to the stimulus. Thus, repeating strong arguments tends to lead to more persuasion (at least to the point of tedium), and repeating weak arguments tends to lead to less persuasion (Petty and Cacioppo 1986).

5. A Dual Process Approach to Understanding Attitude Change

Although the theories just described continue to be useful in accounting for a variety of persuasion phenomena, much of the contemporary literature on attitude change is guided by one of the available 'dual process' models of judgment. For example, one of the earliest approaches of this type, the elaboration likelihood model (ELM) represents an attempt to integrate the many seemingly conflicting findings in the persuasion literature under one conceptual umbrella by specifying a finite number of ways in which source, message, recipient, and contextual variables have an impact on attitude change (see Petty and Cacioppo 1986, Petty and Wegener 1998; for reviews of the ELM, the related heuristic–systematic model, and other dual process approaches, see Chaiken and Trope 1999). The ELM is based on the notion that people want to form correct attitudes (i.e., those that will prove useful in functioning in the environment) as a result of exposure to a persuasive communication, but there are a variety of ways in which a reasonable position can be adopted.

The most effortful procedure for evaluating an advocacy involves drawing upon prior experience and knowledge to scrutinize carefully and think about all of the issue-relevant information available in the current environment, along the dimensions that are perceived central to the merits of the attitude object. According to the ELM, attitudes formed or changed by this central route are postulated to be relatively persistent, predictive of behavior, and resistant to change until they are challenged by cogent contrary information along the dimension or dimensions perceived central to the merits of the object. However, it is neither adaptive nor possible for people to exert considerable mental effort in processing all of the persuasive information to which they are exposed. This does not mean that people never form attitudes when motivation and/or ability to think are low, but rather that attitudes are more likely to be changed as a result of relatively simple associations, on-line inferences, and well-learned heuristics in these situations. Attitudes formed or changed by these peripheral

route processes are postulated to be relatively less persistent, resistant, and predictive of long-term behavior than those based on central route processes. Thus, the ELM holds that both central and peripheral processes are important for understanding attitude change, but their influence varies depending on the likelihood of thinking.

The ELM holds that there are many variables capable of affecting elaboration and influencing the route to persuasion. Some variables affect a person's motivation to process issue-relevant information (e.g., the personal relevance of the issue; personal accountability for a decision), whereas others affect their ability or opportunity to think about a message (e.g., the extent of distraction present; the number of times the information is repeated). Some variables affect processing in a relatively objective manner (e.g., distraction disrupts both favorable and unfavorable thinking) whereas others influence elaboration in a biased fashion (e.g., a positive mood makes positive thoughts more likely than negative thoughts when people are motivated and able to think). Biases can stem from both ability factors (e.g., a biased knowledge store), or motivational factors (e.g., when a desire to maintain one's current attitude is more salient than one's desire to consider new information objectively).

Research on the ELM has shown that when the elaboration likelihood is high (e.g., high personal relevance, high knowledge of topic, simple message in print, no distractions, etc.), people typically know that they want and are able to evaluate the merits of the information presented, and they do so. Sometimes this effortful evaluation is relatively objective, but sometimes it is biased. On the other hand, when the elaboration likelihood is low, people know that they do not want and/or are not able to carefully evaluate the merits of the information presented (or they do not even consider exerting effort). Thus, if any evaluation is formed, it is likely to be the result of relatively simple associations or inferences (e.g., agreement with an expert source; counting the number of arguments presented). When the elaboration likelihood is moderate (e.g., uncertain personal relevance, moderate knowledge, moderate complexity, etc.), however, people may be unsure as to whether the message warrants or needs scrutiny, and whether or not they are capable of providing this analysis. In these situations, they may examine the persuasion context for indications (e.g., is the source credible?) as to whether or not they should attempt to process the message.

There are at least two important implications of the ELM. First, the model holds that any one variable can produce persuasion by different processes in different situations. For example, putting people in a positive mood can influence attitudes because of a simple inference process when the likelihood of thinking is low (e.g., 'I feel good so I must like it'), bias thinking when the likelihood of thinking is high (i.e., making positive interpretations more likely than negative ones), and influence the extent of thinking when it is not already constrained to be high or low (e.g., thinking about an unpleasant message less when happy than when sad). Second, as explained next, the model holds that not all attitude changes of the same magnitude are equal. Specifically, thoughtful attitude changes (central route) tend to be more consequential than nonthoughtful changes (peripheral route).

6. *Consequences of Attitude Changes Produced by Different Processes*

It is now clear that there are a variety of processes by which attitudes can be changed, and that the different processes dominate in different situations. That is, some change processes dominate when motivation and ability to think are high, but other change processes dominate when motivation and ability to think are low. Research suggests that attitudes formed by different processes often have different characteristics (Petty and Cacioppo 1986). For example, persistence of persuasion refers to the extent to which attitude changes endure over time. When attitude change is based on extensive issue-relevant thinking, it tends to persist longer than when it is not. However, multiple exposures to positive cues can also produce relatively persistent attitudes.

Resistance refers to the extent to which attitude change is capable of surviving an attack from contrary information. Attitudes are more resistant the stronger the attack they can withstand. Although attitude persistence and resistance tend to co-occur, their potential independence is shown conclusively in work on cultural truisms (McGuire 1964). Truisms such as 'you should brush your teeth after every meal,' tend to be highly persistent in a vacuum, but very susceptible to influence when challenged. People have very little practice in defending truisms because they have never been attacked. These beliefs were likely formed with little issue-relevant thinking at a time during childhood when extensive thinking was relatively unlikely. Instead, the truisms were probably presented repeatedly by powerful, likable, and expert sources. As noted above, the continual pairing of an attitude with positive cues may produce a relatively persistent opinion, but it may not prove resistant when attacked. The resistance of cultural truisms and other attitudes can be improved by motivating and enabling people to defend their positions in advance of a challenging communication. One such 'inoculation' treatment involves exposing people to a few pieces of counter-attitudinal information prior to the threatening communication and showing them how to refute it. The inoculation procedures do not change the valence of a person's initial attitude, but it makes it stronger. Other persuasion treatments that seem ineffective in changing the valence of attitudes might nonetheless be

effective in modifying the strength of the attitude—making it more or less enduring, resistant, or predictive of behavior than it was initially.

7. Summary

In sum, contemporary persuasion theories hold that changes in attitudes can come about through a variety of processes which imbue them with a multiplicity of characteristics and render them capable of inducing a diversity of consequences. According to the popular dual process logic, the processes emphasized by the central route theories should be largely responsible for attitude change when a person's motivation and ability to scrutinize issue-relevant information is high. In contrast, the peripheral route processes should become more dominant as either motivation or ability to think is attenuated. This framework allows understanding and prediction of what variables affect attitudes and in what general situations. It also permits understanding and prediction of the consequences of attitude change. It is now accepted that all attitudes can be based on cognitive, affective, and behavioral information, and that any one variable can have an impact on persuasion by invoking different processes in different situations. Finally, attitudes that appear identical when measured can be quite different in their underlying basis or structure and thus can be quite different in their temporal persistence, resistance, or in their ability to predict behavior. Work on attitude change to the present has focused on the intrapsychic processes responsible for change in adult populations mostly in Western cultures. Future research is needed on the interpersonal processes responsible for attitude change, and the potentially different mechanisms that produce change in different population groups (e.g., children versus elderly individuals, those in individualistic versus collectivist cultures). In addition, as the field matures, current theories are ripe for exportation to important applied domains such as health promotion (e.g., AIDS education), political participation (e.g., determinants of voter choice), and others.

See also: Attitude Formation: Function and Structure; Attitude Measurement; Attitudes and Behavior

Bibliography

Anderson N H 1971 Integration theory and attitude change. *Psychological Review* **78**: 171–206

Bem D J 1965 Self-perception theory. *Advances in Experimental Social Psychology* **6**

Chaiken S, Trope Y (eds.) 1999 *Dual Process Models in Social Psychology*. Guilford Press, New York

Eagly A H, Chaiken S 1993 *The Psychology of Attitudes*. Jovanovich College Publishers, Fort Worth, TX

Festinger L 1957 *A Theory of Cognitive Dissonance*. Row, Peterson, Evanston, IL

Fishbein M, Ajzen I 1975 Belief, attitude, intention, and behavior: an introduction to theory and research. Addison-Wesley, Reading, MA

Greenwald A G 1968 Cognitive learning, cognitive response to persuasion, and attitude change. In: Greenwald A G, Brock T C, Ostrom T M (eds.) *Psychological Foundations of Attitudes*. Academic Press, New York, pp. 147–70

Harmon-Jones E, Mills J (eds.) 1999 *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology*. 1st edn. American Psychological Association, Washington, DC

Hastie R, Park B 1986 The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review* **93**: 258–68

Hovland C I, Janis I L, Kelley H H 1953 *Communication and Persuasion*. Yale University Press, New Haven, CT

Kelley H H 1967 Attribution theory in social psychology. In: Levine D (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, Vol. 15, pp. 192–241

McGuire W J 1964 Inducing resistance to persuasion: some contemporary approaches. *Advances in Experimental Social Psychology* **1**: 191–229

McGuire W J 1985 Attitudes and attitude change. In: Lindzey G, Aronson E (eds.) *Handbook of Social Psychology*, 3rd edn. Random House, New York, pp. 233–46

Pavlov I P 1927 *Conditioned Reflexes*. Oxford University Press, New York

Petty R E, Cacioppo J T 1981 *Attitudes and Persuasion: Classic and Contemporary Approaches*. W. C. Brown, Dubuque, IA

Petty R E, Cacioppo J T 1986 The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology* **19**: 123–205

Petty R E, Krosnick J A (eds.) 1995 *Attitude Strength: Antecedents and Consequences*. Erlbaum, Mahwah, NJ

Petty R E, Ostrom T M, Brock T C (eds.) 1981 *Cognitive Responses in Persuasion*. Erlbaum, Hillsdale, NJ

Petty R E, Wegener D T 1998 Attitude change: multiple roles for persuasion variables. In: Gilbert D, Fiske S, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, Vol. 1, pp. 323–90

Smith M B, Bruner J, White R W 1956 *Opinions and Personality*. Wiley, New York

Staats A W, Staats C K 1958 Attitudes established by classical conditioning. *Journal of Abnormal and Social Psychology* **57**: 37–40

Zajonc R B 1968 Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monograph*. **(Suppl.) 9**: 1–27

R. E. Petty

Attitude Formation: Function and Structure

Attitude is a psychological tendency that is expressed by evaluating a particular object or entity with some degree of favor or disfavor. In this definition, an object or entity can be virtually any 'thing' in a person's

internal or external social environment. Thus, people hold attitudes, or evaluations, toward an endless variety of objects, including the self, another person or groups of persons, social and political policies, goals, and behaviors. Specific terms are sometimes applied to certain classes of attitudes. For example, attitudes toward minority groups are often called *prejudice*, and an attitude toward one's self is often called *self-esteem*. Attitudes have two basic components at the level of abstract evaluation, direction (positive vs. negative) and intensity—one may, for example, be extremely positive or only moderately positive toward 'Democrats.' This definition of attitude is consensual within social psychology, the discipline that has generated most basic theory and research on this concept. From a very early point in social psychology, attitude has been regarded as a fundamental construct (Allport 1935). This high status in social psychology, and throughout the social sciences, has endured and shows no signs of abating.

1. Attitude and Other Constructs

For many years attitude was viewed solely as a hypothetical construct. Like intelligence and other hypothetical constructs, attitude was considered a theoretically and empirically useful construct capable of explaining certain regularities in social thought and behavior. In this S-O-R (stimulus-organism-response) approach, attitude could be inferred from overt cognitive, affective, or behavioral responding. But attitude itself was considered neither directly observable nor likely to have a genuine physical or neural reality. At the beginning of the twenty-first century, attitude remains largely a hypothetical construct. Nonetheless, there is widespread agreement that attitudes are represented in memory, and increased understanding that as measurement technology in psychological science progresses, attitudes will eventually be shown to have more directly observable physiological, especially neural, underpinnings.

Before considering the formation, functions, and structure of attitudes, it is important to differentiate attitudes from other, seemingly related individual difference constructs. While attitudes do differ from person to person, they are always held in relation to a specific object, unlike broader personality dimensions such as extraversion–introversion, and unlike the relatively temporary evaluative states known as moods. Also, while personality differences like extraversion are usually thought to be chronic and thus relatively stable—and despite early treatments of attitudes as *enduring* dispositions—attitudes may in fact be either stable or unstable. Stability has become a theoretical and empirical issue, with stability, or persistence, dependent on factors such as attitudinal structure and people's exposure to new direct or indirect experience with attitude objects (Eagly and Chaiken 1993).

Habits, opinions, beliefs, and values have often been confused with attitudes. Habits refer more to regularities in behavior, are considered relatively automatic, and do not necessarily imply evaluation. Opinions and beliefs are interchangeable in contemporary usage. While beliefs do often imply evaluation, the hallmark of attitude, they are best viewed as likelihoods, or subjective probabilities; for example, the belief that attitude object X (e.g., African-Americans) possesses or is associated with certain attributes (e.g., athleticism). Beliefs may contribute to the formation of attitudes and are best viewed as part of an attitude's structure (see Sect. 3). Finally, attitudes and values are typically distinguished, though they need not be. Strictly speaking, values are 'just' attitudes in the sense that they convey people's evaluations of 'objects' (e.g., one values freedom). Yet researchers continue to use both terms because the objects toward which we hold values are broader than the objects toward which we hold attitudes. We hold values toward very broad and, at least within cultures, consensually held personal or social goals (e.g., being responsible) and desired end-states of existence (e.g., equality). The objects toward which we hold attitudes are typically narrower (e.g., health care reform).

2. Attitude Formation

How are attitudes, our abstract evaluations of objects in our environment, formed? Traditional accounts of attitude formation have emphasized that attitudes are *learned*. Indeed, early definitions of attitudes emphasized that attitudes are 'learned' predispositions to respond favorably or unfavorably toward objects in the environment (e.g., Allport 1935). That many, if not most, attitudes are learned is theoretically and empirically uncontroversial. Nonetheless, it should be noted that some scholars have argued that at least some attitudes are unlearned in the sense that people may have predispositions to react positively or negatively to novel attitude objects. Such predispositions may reflect evolutionary adaptations and/or biological substrates that are inherited (e.g., Tesser 1993). Predispositions to evaluate, however, are not the same thing as attitudes because attitudes arise from *evaluative responding* to an attitude object. Thus, people do not have an attitude until they first encounter the attitude object and respond evaluatively to it on an affective, cognitive, or behavioral basis. For example, although humans may be predisposed to react negatively to sour tastes, a person would not have a negative attitude toward a sour-tasting fruit until he or she had some first- or second-hand experience with it (e.g., tasting it, hearing another describe its taste).

Cognitive learning theories have been especially prominent in explaining how attitudes are formed (and modified). In the most widely accepted cognitive learning model of attitude formation, Fishbein and Ajzen (e.g., 1975) proposed that people's evaluations

of objects are based on the beliefs that they form about these objects, whether through direct or indirect experience. In this expectancy-value approach, attitudes are viewed as the sum of the evaluative implications of a person's beliefs about the attitude object. Beliefs about an object (e.g., Democrats care about the poor) represent the expectancy (or subjective probability) component, and evaluations of the attributes of beliefs (e.g., 'caring for the poor') represent the value component. This general theory has obvious implications for how attitudes are formed and modified (see also *Attitudes and Behavior*; *Attitude Change: Psychological*). Beliefs may be acquired directly, through firsthand experience with attitude objects (e.g., playing with a dog), and also indirectly, via socialization agents such as parents, teachers, peers, and the mass media. The large literature on designing persuasive messages attests to the popularity of attempting to change attitudes by instilling new beliefs (or refuting old beliefs) about attitude objects (see also *Sexual Attitudes and Behavior*). By these various means we form beliefs about, and therefore attitudes toward, a vast array of socially significant (and insignificant) objects in our environment.

Fishbein and Ajzen's approach stressed beliefs, or cognitions, as the foundation of attitudes. In contrast, other models of attitude formation have emphasized *affective* experience. Traditional affective approaches included both operant and classical conditioning explanations of attitude formation. Most notably, classical conditioning studies paired neutral ('conditioned') attitude–object stimuli (e.g., 'George') with evaluative ('unconditioned') stimuli (e.g., 'friendly,' or 'naughty') guaranteed to elicit positive or negative ('unconditioned') responses. These studies showed that after repeated pairings, previously neutral stimuli came to elicit positive or negative evaluations, depending upon the positive/negative nature of the unconditioned stimuli they were paired with.

Empirical evidence for classical (as well as operant) conditioning of attitudes was criticized during the 'demand characteristics' crisis in social psychology, in the 1960s. Critics argued that the simplicity of these experimental paradigms made clear to participants the purposes of the research enterprise, and argued that experimental outcomes were simply the result of 'good subjects' providing 'respected' researchers their hoped-for results. Yet, more recent analysis of the conditioning literature and these critiques suggest that the fundamental proposition that attitudes can be formed through simple affective conditioning mechanisms remains plausible (Eagly and Chaiken 1993).

Research on mere exposure also gives importance to affective experience in attitude formation. The mere exposure effect, demonstrated initially by Zajonc (see Zajonc 1980) and subsequently by numerous others, is that the more frequently people are minimally, or 'merely,' exposed to stimuli—ranging from other persons to simple geometric figures—the more positive their attitudes toward such stimuli become. Some commentators dismissed this work as the product of demand characteristics, or as demonstrating little more than cognitive learning (e.g., Fishbein and Ajzen 1975). Yet more recent research has bolstered Zajonc's (e.g., 1980) contention that mere exposure is the result of a simple, and *automatic*, affective conditioning process.

In current usage, an automatic process is fast, consumes little if any cognitive capacity, and is largely unconscious in the sense that it occurs without awareness and is neither intentional nor controllable (e.g., Bargh 1997). Automatic attitude formation is therefore of great importance insofar as it challenges traditional cognitive theorizing that attitudes (and numerous other types of judgments) are the product of deliberative cognitive processing of information in one's environment. Deliberative, or controlled, processing is everything that automatic processing is not: It is relatively slow, demands cognitive resources, and proceeds with intentionality and awareness. A theoretical possibility in the 1980s, automatic attitude formation has now been demonstrated in a variety of experimental paradigms. Most notably, mere exposure effects have been found even when research participants cannot consciously recognize the stimuli to which they are (more or less frequently) exposed. In much of this work, participants have been exposed to stimuli subliminally. Despite the phenomenological fact that participants have no idea which stimuli are 'old' (i.e., those that have been presented subliminally) and which stimuli are 'new' (i.e., those that have never been presented) they nevertheless exhibit a strong evaluative preference for 'old' stimuli (see Zajonc 1980). In related work, researchers have shown affective priming effects (e.g., Murphy et al. 1995). In such studies, participants' attitudes toward novel stimuli (e.g., geometric figures) are more positive (or more negative) when these stimuli are presented just after (a single) subliminal exposure to an emotionally evocative positive (or negative) picture.

Complementing this work has been research demonstrating that people may automatically evaluate virtually everything in their environment (see Bargh 1997). In this research paradigm, participants are briefly exposed to an 'attitude object prime' of positive or negative valence (e.g., 'summer' or 'snake') and, a fraction of a second later, to a target adjective of positive or negative valence (e.g., 'beautiful' or 'ugly'). Participants' task is to respond in some manner to the target word (e.g., to pronounce it) while ignoring the prime word. Results across a variety of instructional sets and stimuli (including familiar as well as novel stimuli) indicate that participants are engaged in automatic evaluation of the prime word. This is demonstrated by the fact that participants are, for example, faster to pronounce a target word (e.g., 'ugly') if it has been preceded by a prime of like valence (e.g., 'snake' vs. 'summer').

In addition to attitudes arising out of cognitive and affective responding to attitude objects, attitudes can of course also arise from behavioral responding. In most such direct experience situations, behavior's impact on attitude is probably mediated by a combination of affective and cognitive responding; in other words, behavior vis-à-vis an attitude object elicits both beliefs and feelings in perceivers which, in turn, give rise to an overall attitude, or evaluation. There are, however, other mechanisms by which behavior creates attitudes. In classic dissonance research, participants induced to engage in a behavior assumed to be at odds with their (often unmeasured) initial attitudes subsequently report attitudes that are congruent with their induced behavior, presumably out of a desire to reduce the dissonance, or discomfort, caused by an attitude–behavior discrepancy. A more purely cognitive account of the same phenomenon is provided in self-perception theory. In this account, people are assumed to have little access to their interior feelings and beliefs. Hence, when induced to perform a certain behavior (e.g., signing a petition) they simply infer that they possess an attitude congruent with that behavior, perhaps using the simple heuristic that 'behavior x implies disposition or attitude y' (see Chaiken et al. 1996).

This section has emphasized how attitudes are formed in an experiential way based on direct or indirect cognitive, affective, or behavioral responding to an attitude object. Contemporary scholarship holds that this responding is represented in memory as attitude object–response associations. As evaluative meaning is abstracted from these associations, an attitude is formed as a generalization from more elementary associations. In contrast to this intra-attitudinal mode of attitude formation, people may also form attitudes by forging linkages between an attitude object and other attitude objects. Such linkages, or associations, are represented in memory along with the target attitude itself. Often this interattitudinal mode of attitude formation entails an inference by which a new attitude (e.g., toward recycling) is deduced from a more abstract or general attitude that has already been formed (e.g., the value of 'a world of beauty').

3. *Attitude Structure*

Attitudes have most often been measured by self-report. Paralleling the automatic-controlled processing distinction, contemporary researchers have begun to label such measures *explicit*, because respondents are usually well aware that their attitudes are the target of measurement. For decades researchers have explored various 'unobtrusive' attitude measures, including physiological assessments such as heart rate, pupil size and, more recently, facial muscle patterns. These have been newly labeled *implicit*, insofar as respondents are usually unaware of the purpose of these measurements. Promising new implicit measures rely on response-time methods such as those featured in the automatic evaluation paradigm described earlier, and dissociations between explicit and implicit measures are attracting a great deal of empirical attention (e.g., Fazio 1995; see also *Attitude Measurement*).

Regardless of their explicit or implicit nature, measures of attitude have traditionally been used to locate individuals along a bipolar dimension running from highly positive to highly negative toward the attitude object. If theorists assumed that attitude structure reflected this operational definition (though generally they do not), attitudes would have, at best, rather simple structural properties. Although some such properties have been ascribed to attitudes, most researchers have focused on (a) attitude's relation to other attitudes, as in Heider's classic balance theory and some ideological conceptions of structure; and even more so on (b) attitude's associations with attitude-relevant beliefs, affects, and behaviors (see Eagly and Chaiken 1993, 1998). This discussion focuses on the latter, more extensively researched issue, namely intra-attitudinal structure. Critical to work on this topic is the assumption that attitudinal responding is better understood when attitudes' internal structure is taken into account along with the sheer positive vs. negative nature of overall attitude.

Considering that attitudes are formed through affective as well as cognitive (and behavioral) responding, and that attitudes and associated beliefs and affects are represented in memory, it is unsurprising that questions regarding the nature of such associations would be of interest to investigators. Most work has concentrated on attitude-relevant beliefs, and structural properties such as the amount, complexity, and consistency of beliefs have been examined. For example, how many beliefs a person holds about an attitude object is important for understanding the processing of new attitude-relevant information; more knowledgeable people tend to process information deeply, or systematically, whereas less knowledgeable people tend to process information superficially, or heuristically (e.g., Wood et al. 1995). As a second example, belief complexity is predictive of either attitudinal extremity or attitudinal moderation, depending on other conditions such as the extent to which beliefs are evaluatively redundant (e.g., Judd and Lusk 1984). Finally, the extent to which beliefs are themselves more or less evaluatively congruent has also yielded interesting insights. Inconsistency among beliefs, often termed *ambivalence*, is associated with attitude instability and heightened susceptibility to social influence. Similar structural properties no doubt exist with respect to affective associations to attitude objects but have to date not been well researched.

A particularly interesting structural property is evaluative-cognitive consistency—the extent to which

the evaluative implications of a person's stored (or accessible) beliefs match the evaluation implied by overall attitude. Research on this property has, for example, shown that high consistency attitudes are better predictors of behavior and are also more resistant to social influence attempts (see also *Attitudes and Behavior*; *Attitude Change*: *Psychological*). More recent work has also demonstrated the importance of evaluative-affective consistency—the extent to which the evaluative implications of affective associations match the evaluation implied by overall attitude. For example, research on selective memory has shown that higher evaluative-affective consistency produces a memory bias favoring attitude-congruent information; in contrast, higher evaluative-cognitive consistency produces a memory bias favoring attitude-incongruent information (see Eagly and Chaiken 1998). This research reflects another recent structural distinction, whether affect-based attitudes behave differently from cognitively based attitudes. In these terms, the selective memory research just described could be restated as showing that affective attitudes produce a congeniality bias, whereas cognitive attitudes produce an incongruency bias. Additional issues regarding attitude basis are sure to receive increased empirical attention in the future.

One of the most important reasons for examining attitude structure is its relation to the concept of attitude *strength*. Although traditional treatments of attitude held that they are important to study because attitudes guide thought, judgment, and behavior, contemporary investigators know well that not all attitudes are equally predictive of such outcomes. Hence the concept of attitude strength; 'strong' attitudes are highly predictive of thought, judgment, and behavior while 'weaker' attitudes are not. But what, exactly, constitutes a strong attitude? There have been numerous conceptions of attitude strength; for example, that such attitudes are 'personally important,' or 'ego-involving,' or 'intense,' or 'come easily to mind,' and so on. Yet there is not yet consensus on strength's conceptual definition or its specific properties or, for that matter, whether there are multiple aspects of strength. Research on attitude structure, however, presents a promising and relatively integrative way of examining the nature of attitude strength (see Eagly and Chaiken 1995). For example, some strength research suggests that affective attitudes may be stronger than cognitive attitudes; such attitudes come to mind more easily (i.e., are more accessible) and may well be more predictive of behavior (e.g., Fazio 1995). Other work suggests that regardless of whether attitudinal associations in memory are mostly affective, mostly cognitive, or some combination, attitudes are stronger to the extent that their internal structure consists of many (vs. few) affective and/or cognitive associations, especially when such associations possess greater *evaluative consistency* (see Eagly and Chaiken 1995, 1998).

Although issues of attitude structure have been addressed throughout the history of attitude theorizing and research, there has been a renaissance of sorts on this topic since the early 1990s. Because of its importance to a large number of attitudinal phenomena (e.g., attitude strength, selective information processing, attitudinal persistence and resistance to change), it is likely to be a prominent area of research for many years to come.

4. *Attitude Functions*

Why do people hold attitudes? Like other broad questions, this one can be answered on several levels. Yet the assumption underlying all such answers is the usual functionalist premise that attitudes enable individuals to adapt to their environment. The most profound sense in which attitudes help people adapt emerges readily from their definition. To evaluate—to assess the degree of goodness vs. badness inherent in features of one's world—is absolutely fundamental to the task of survival. If people were unable to identify 'good' stimuli, those that enable survival and enhance well-being, and to identify 'bad' stimuli, those that threaten survival and lessen well-being, they would fail to thrive.

Consistent with this view, the seminal work of Osgood et al. (1957) argued that evaluation is primary, and offered as their main evidence numerous factor analyses of meaning in which an evaluative factor accounted for more variability than any other factor (in particular, potency and activity). Although always influential in attitude theory, the strongest evidence for the primacy of evaluation has come only recently, in research on evaluative primacy and the automatic nature of attitude formation. As cited earlier, for example, Bargh and colleagues (see Bargh 1997) have shown that people immediately (i.e., within a fraction of a second) evaluate as good or as bad any stimulus object presented to them—regardless of whether the object is a word or a picture, regardless of the strength of one's stored evaluation of the object, and regardless of whether the object is familiar or novel.

Smith and colleagues (e.g., Smith et al. 1956) captured this essential function of attitudes in their conception of *object appraisal*, a function by which attitudes allow people to appraise stimuli in terms of their goals and concerns. Shortly thereafter, Katz (1960) essentially partitioned the object appraisal function into two components. Whereas Katz's *knowledge* function emphasized the cognitive function of facilitating people's ability to interpret and make sense of otherwise disorganized perceptions, his *instrumental* function emphasized the motivational function of facilitating people's ability to maximize rewards and minimize punishments. Object appraisal, which joins the knowledge and instrumental aspects of attitudes, should be regarded as attitudes' universal function.

Indeed, many contemporary discussions emphasize that attitudes allow people to categorize objects, persons, and events and to differentiate between things they should approach and things they should avoid (Eagly and Chaiken 1998).

The universal, object appraisal function of attitudes implies a very general view of the kinds of rewards and punishments with which attitudes are associated. These positive and negative outcomes may reflect both narrow self-interest and more abstract benefits such as the affirmation of one's values or self-concept. To stabilize terminology with respect to attitudes' facilitation of rewarding outcomes, Eagly and Chaiken (1998) have proposed that object appraisal should indeed encompass all classes of positive and negative outcomes. In contrast, a narrower *utilitarian* function should encompass only those concrete rewards and punishments (e.g., money, opportunities) that the attitude object provides directly. This narrower function is useful for contrast with an *expressive* function of attitudes wherein psychological gains and losses follow from holding or expressing an attitude. For example, in discussions of *symbolic politics*, researchers have tried to determine whether people's attitudes toward policies such as affirmative action are better predicted by narrow self-interest or by internalized social values such as fairness and equality (see Sears and Funk 1991).

Three additional functions of attitudes have frequently been identified. The *value-expressive* function assumes that people find it inherently rewarding to hold and express attitudes that affirm their core values and self-concepts. The *social adjustive* function assumes that there are inherent rewards in holding and expressing attitudes that are pleasing to others or that coincide with the norms and values of important reference groups. Finally, the *ego-defensive* function assumes that holding or expressing attitudes can also serve to defend the self (and one's important reference groups) from potentially threatening events. Because these functions are not inherent in the nature of all attitudes, they, along with the utilitarian function, coexist with the universal object appraisal function. Moreover, these more specific functions are assumed to be more important for some individuals, for some attitude objects, and under some circumstances. Identifying when attitudes serve these particular functions, and how attitude functions map on to structural issues has been and will remain a key issue for researchers (see Eagly and Chaiken 1998).

5. *Future Directions*

The purpose of this article has been to inform readers outside the area of attitudes about classic and contemporary theoretical and empirical issues. As noted earlier, attitude theory and research has been central to the discipline of social psychology since its inception. While this area of inquiry has been highly programmatic and remains vigorous and rigorous in its theorizing and methodologies, there are ample issues for future research. Some have already been mentioned. For example, future directions should include achieving a better understanding of distinctions such as automatic vs. controlled processing in attitude formation (and change), implicit vs. explicit measures of attitude, and the relation between attitude structure and attitude strength and between attitude structure and attitude function. In addition, future research would benefit greatly from greater attention to other psychological disciplines as well as other social science disciplines. For example, little social psychological work on attitudes has considered developmental issues such as how attitudes develop over time or across the lifespan. Finally, although attitude is an important construct throughout the social sciences, there has been too little scholarly contact between social psychological and other social science treatments of attitudes. Greater attention to integrating knowledge across relevant disciplines should ensure the continuing fundamental importance of the attitude construct.

See also: Attitude Change: Psychological; Attitude Measurement; Attitudes and Behavior; Belief, Anthropology of; Collective Beliefs: Sociological Explanation; Environmental Cognition, Perception, and Attitudes; Opinion Formation; Public Opinion: Microsociological Aspects; Sexual Attitudes and Behavior; Shared Belief; Social Categorization, Psychology of; Social Influence, Psychology of; Social Psychology: Research Methods; Stereotypes, Social Psychology of

Bibliography

Allport G W 1935 Attitudes. In: Murchison C (ed.) *Handbook of Social Psychology*. Clark University Press, Worcester, MA, pp. 798–844

Bargh J A 1997 The automaticity of everyday life. In: Wyer R S Jr. (ed.) *Advances in Social Cognition*. Erlbaum, Mahwah, NJ, Vol. 10, pp. 1–61

Chaiken S, Wood W, Eagly A H 1996 Principles of persuasion. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology: Handbook of Basic Principles*. Guilford Press, New York, pp. 702–44

Eagly A H, Chaiken S 1993 *The Psychology of Attitudes*. Harcourt Brace Jovanovich, Fort Worth, TX

Eagly A H, Chaiken S 1995 Attitude strength, attitude structure, and resistance to change. In: Petty R E, Krosnick J A (eds.) *Attitude Strength: Antecedents and Consequences*. Erlbaum, Mahwah, NJ, pp. 413–32

Eagly A H, Chaiken S 1998 Attitude structure and function. In: Gilbert D, Fiske S F, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, Vol. 2, pp. 269–322

Fazio R H 1995 Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In: Petty R E, Krosnick J A (eds.) *Attitude Strength: Antecedents and Consequences*. Erlbaum, Mahwah, NJ, pp. 247–83
Fishbein M, Ajzen I 1975 *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA
Judd C M, Lusk C M 1984 Knowledge structures and evaluative judgments: Effects of structural variables on judgmental extremity. *Journal of Personality and Social Psychology* **46**: 1193–1207
Katz D 1960 The functional approach to the study of attitudes. *Public Opinion Quarterly* **24**: 163–204
Murphy S T, Monahan J L, Zajonc R B 1995 Additivity of nonconscious affect: Combined effects of priming and exposure. *Journal of Personality and Social Psychology* **69**: 589–602
Osgood C E, Suci G S, Tannenbaum P H 1957 *The Measurement of Meaning*. University of Illinois Press, Urbana, IL
Sears D O, Funk C L 1991 The role of self-interest in social and political attitudes. In: Zanna M P (ed.) *Advances in Experimental Social Psychology*. Academic Press, San Diego, CA, Vol. 24, pp. 2–91
Smith M B, Bruner J S, White R W 1956 *Opinions and Personality*. Wiley, New York
Tesser A 1993 The importance of heritability in psychological research: The case of attitudes. *Psychological Review* **100**: 129–42
Wood W L, Rhodes N, Biek M 1995 Working knowledge and attitude strength: An information-processing analysis. In: Petty R E, Krosnick J A (eds.) *Attitude Strength: Antecedents and Consequences*. Erlbaum, Mahwah, NJ, pp. 283–313
Zajonc R B 1980 Feeling and thinking: Preferences need no inferences. *American Psychologist* **35**: 151–70

S. Chaiken

Attitude Measurement

The term *attitude* refers to a hypothetical construct, namely a predisposition to evaluate some object in a favorable or unfavorable manner. This predisposition cannot be directly observed and is inferred from individuals' responses to the attitude object, which can run from overt behavior (such as approaching or avoiding the object) and explicit verbal statements to covert responses, which may be outside of the individual's awareness (such as minute facial expressions). This article reviews diverse measurement procedures that draw on these different manifestations of attitudes (see Oskamp 1991, for a history).

1. *Explicit Self-reports*

Most attitude researchers rely on respondents' answers to questions, such as 'Do you approve or disapprove of how President Clinton is handling his job?' Direct questions are the most feasible procedure for assessing the attitudes of the population at large, e.g., in representative sample surveys. Unfortunately, self-reports of attitudes are highly context dependent and minor changes in question wording, question format, and question order can dramatically alter the obtained results (for reviews see Schuman and Presser 1981, Schwarz and Sudman 1992, Tourangeau et al. 2000). This section addresses the underlying processes and their implications for the measurement and conceptualization of attitudes.

1.1 *The Question Answering Process*

To answer an attitude question, respondents need to: (a) understand the meaning of the question; (b) retrieve relevant information from memory to; and (c) form an attitude judgment. In most cases, they also need to: (e) map their judgment onto the response alternatives provided by the researcher; and they (f) may want to edit their response before they report it, due to reasons of social desirability and self-presentation. Respondents' performance at each of these steps is highly context dependent (see Sudman et al. 1996 for a comprehensive treatment). The cognitive and communicative processes involved in these tasks are reviewed in *Data Collection: Interviewing*, *Questionnaires: Assessing Subjective Expectations*, and this material applies directly to attitude questions. The following discussion is restricted to specific aspects of the judgment stage and the use of different response formats.

1.2 *The Construction of Attitude Judgments*

Once respondents determined what the question refers to, they need to recall relevant information from memory. In some cases, they may have direct access to a previously formed relevant judgment that they can offer as an answer. In most cases, however, they will not find an appropriate answer readily stored in memory and will need to compute a judgment on the spot. To do so, they need to form a mental representation of the attitude object and of a standard, against which the object is evaluated. The resulting judgment depends on which information happens to come to mind at that point in time and on how this information is used.

As a large body of research demonstrated, individuals rarely retrieve all information that may bear on an attitude object but truncate the search process as soon as enough information has been retrieved to form a judgment with sufficient subjective certainty (Wyer and Srull 1989). Hence, the judgment depends on the first few pieces of information that come to mind. Whereas some information may always come to mind when the person thinks of this object (and is therefore called *chronically accessible*), other information may be only *temporarily accessible*, for example, because it

has been brought to mind by preceding questions. Changes in what is temporarily accessible are at the heart of many context effects in attitude measurement, whereas chronically accessible information contributes some stability to attitude judgments.

How accessible information influences the attitude judgment depends on how it is used. Information that is included in the mental representation formed of the attitude object results in assimilation effects, i.e., the judgment becomes more positive (negative) when information with positive (negative) implications comes to mind. In contrast, information that is used in forming a mental representation of the standard against which the attitude object is evaluated results in contrast effect. In this case, the judgment becomes more negative (positive) when information with positive (negative) implications comes to mind and is used as a positive (negative) standard of comparison. Suppose, for example, that a preceding question brings a politician to mind who was involved in a scandal (say, Richard Nixon). When a subsequent question pertains to the trustworthiness of American politicians in general (a superordinate category), Richard Nixon is likely to be included in the representation formed of 'politicians' as a group, resulting in judgments of decreased trustworthiness (an *assimilation effect*). Suppose, however, that the question pertains to the trustworthiness of Bill Clinton (a lateral category), rather than to politicians in general. In this case, Richard Nixon cannot be included in the representation formed of the attitude object 'Bill Clinton'; instead, he is likely to serve as a (rather negative) standard of comparison, relative to which Bill Clinton will be evaluated as more trustworthy than would otherwise be the case (a *contrast effect*). As this example illustrates, the same context question can have profoundly different effects on superficially similar subsequent questions, resulting in many apparent inconsistencies: in the present case, a researcher might conclude that political scandals decrease trust in politicians when judgments of the group are used as the dependent variable, but increase trust in politicians when judgments of individual politicians are used as the dependent variable.

If a given piece of information is used in constructing a representation of the attitude object (resulting in assimilation effects), or of a standard of comparison (resulting in contrast effects), depends on a host of different variables. Sudman et al. (1996, Chap. 5) review these variables and summarize a theoretical model that predicts the direction, size, and generalization of context effects in attitude measurement.

1.3 Some Attitude Scale Formats

Textbooks consistently advise researchers to use multi-item scales. Yet, the classic examples of multi-item attitude scales—namely, *Thurstone scales, Guttman scales*, and *Likert scales* (see Eagly and Chaiken 1993, Chap. 2)—are rarely used in practice, primarily because they require extensive, topic-specific development work. An exception is Osgood and colleagues' *Semantic Differential Scale*(see Eagly and Chaiken 1993, Chap. 2), which is a ready-to-use scale that can be applied to any topic, making it considerably more popular. Respondents are asked to rate the attitude object (e.g., 'welfare') on a set of seven-point bipolar adjective scales. The adjectives used as endpoint labels reflect three general factors, namely evaluation (e.g., good–bad; pleasant–unpleasant), potency (e.g., strong–weak; small–large), and activity (e.g., active–passive; fast–slow). Of these factors, evaluation is considered the primary indicator of respondents' attitude towards the object, as reflected in the objects' (relatively global) connotative meaning.

However, in most studies respondents' attitudes are assessed by only one or two questions, presented either with multiple-choice response alternatives or a rating scale. When the question offers several distinct opinions in a *multiple-choice format*, it is important to ensure that the set of response alternatives offered covers the whole range of plausible positions. Any option omitted from the set is unlikely to be reported at all, even when respondents are offered a general 'other' response option, which they rarely endorse. Similarly, few respondents report not having an opinion on an issue when this option is not explicitly provided—yet, they may be happy to report so when 'don't know' is offered as an alternative (see Schuman and Presser 1981, Schwarz and Hippler 1991, for reviews).

When *rating scales* are used, respondents are presented with a numerical scale with verbally labeled endpoints (e.g., -3 = 'strongly disagree,' $+3$ = 'strongly agree') and asked to check the number that best represents their opinion. Alternatively, each point of the rating scale may be labeled, a format that is common in telephone interviews. In general, the retest-reliability of fully labeled scales is somewhat higher than that of partially labeled scales and retest reliability decreases as the number of scale points increases beyond seven, reflecting the difficulty of making many fine-grained distinctions (Krosnick and Fabrigar in press). What is often overlooked is that the numeric values of the scale may influence respondents' substantive interpretation of the question. For example, when asked to report how highly they think of a politician along a scale from 'not so highly' to 'very highly,' respondents need to determine if 'not so highly' pertains to the absence of positive thoughts or to the presence of negative thoughts. When the numeric values run from 0 to 10, respondents interpret 0 = 'not so highly' as reflecting the absence of positive thoughts; but when the scale runs from -5 to $+5$, they interpret -5 = 'not so highly' as pertaining to the presence of negative thoughts. In general, scales that

range from negative to positive numbers suggest that the researcher has a bipolar dimension in mind, implying that one endpoint label refers to the *opposite* of the other; in contrast, scales that present only positive numbers suggest a unipolar dimension, implying that different scale values pertain to different degrees of the *same* attribute. Such shifts in question meaning result in markedly different responses (see Sudman et al. 1996, Chap. 3, for what respondents learn from scales).

1.4 Implications

The emergence of context effects in attitude measurement raises important theoretical issues. Traditionally, attitudes have been conceptualized as enduring dispositions, implying that context effects represent undesirable noise. In fact, some researchers suggested that context effects may only emerge when respondents have not yet formed an opinion about the object, but not when respondents have formed strong, or crystallized, attitudes. Empirically, however, this *attitude strength* hypothesis has received little support (see Krosnick and Abelson 1992, for a review). Alternatively, we may conceptualize attitudes as evaluative judgments that are formed on the spot, when needed, based on whatever information is accessible at the time. These diverging perspectives have differential implications for the conditions under which we can expect attitude stability over time, or close attitude-behavior relationships, and suggest different research and measurement strategies (see Schwarz and Bohner 2001).

2. Implicit Attitude Measures

Complementing direct attitude questions, researchers have developed a number of indirect measures, often intended to assess attitudes that respondents may not be willing to report or of which they may themselves be unaware. Such measures include projective procedures, such as the *Thematic Apperception Test*, as well as more recent developments that rely on response time measurement, such as lexical decision or pronunciation tasks. These measures are commonly referred to as *implicit measures* and their hallmark is that respondents are not aware that their attitude is being assessed.

The use of *lexical decision tasks* is based on the observation that a letter string (e.g., 'doctor') is more quickly recognized as a meaningful word when preceded by a closely related concept (e.g., 'nurse') rather than an unrelated one (e.g., 'butter'). In applications of this method, meaningful words as well as meaningless letter strings are presented for a very short time and respondents are asked to say, as quickly as possible, if what they see is a word or a nonword; alternatively, they may be asked to *pronounce* the word as fast as possible. Of interest is how fast they can recognize a given evaluatively laden word when it is, or is not, preceded by the attitude object (for reviews see Bassili 2001, Dovidio and Fazio 1992). Using this procedure to assess racial attitudes in the United States, researchers observed, for example, that words describing positive personality traits were recognized faster than words describing negative personality traits when preceded by 'Whites,' whereas the reverse held true when the same traits were preceded by 'Blacks.' These findings indicate that positive traits are more closely associated with the attitude object 'Whites' than negative traits, whereas the reverse holds for the attitude object 'Blacks.' More complex variants of this general approach include the *Implicit Associations Test*(*IAT*), which draws on the observation that a given response, like pressing a key with one's left index finger, is easier (and hence faster) when all objects associated with this response share the same valence (see Bassili 2001, Banaji et al. 2001, for reviews).

In most studies, the results obtained with these implicit procedures are only weakly related to explicit self-reports of attitudes or overt behavior. Some researchers concluded that this observation demonstrates the power of implicit measures: presumably, the weak relationships indicate that these procedures assess attitudes that respondents do not want to admit or may not even be aware of. In contrast, others wondered what weak relationships imply for the validity of implicit measures: do they really assess evaluative dispositions or do the results merely reflect the structure of semantic knowledge with limited behavioral implications? Recent research further suggests that implicit measures are as likely to show context effects as explicit measures, potentially thwarting the hope that they may provide an unbiased window on individuals' attitudes (see Banaji et al. 2001). Future work in this very active area will need to settle these issues.

3. Observational and Psychophysiological Measures

In principle, an individual's attitude towards some object may be inferred from his or her behavior towards it. Unfortunately, however, individuals' behavior is influenced by many variables other than their attitudes and the attitude-behavior relationship is typically weak. Hence, direct *behavioral observation* is rarely used in attitude measurement.

As an alternative observational approach, researchers may assess individuals' physiological responses to the attitude object (for reviews see Cacioppo et al. 1994, Winkielman et al. 2001). If an attitude object evokes strong feelings, exposure to it should be

associated with increased activation of the sympathetic nervous system. Increased sympathetic activation results in increased sweat glands activity, which can be measured by assessing the resistance of the skin to low level electric currents. This *electrodermal measurement*, however, has the disadvantage that sweat glands activity only reflects the intensity but not the direction (favorable or unfavorable) of the evaluative response, thus limiting its usefulness.

More promising are attempts to assess changes in individuals' facial expression in response to an attitude object. Overt facial expressions (like smiling or frowning) may often be observed in response to attitude objects that elicit strong reactions. Yet, these expressions may be intentionally concealed and many evaluative reactions may be too subtle to evoke overt expressive behaviors. Even subtle evaluative reactions, however, are associated with low-level activation of facial muscles that can be detected by *electromyography* (EMG). These muscle reactions reflect the direction (favorable vs. unfavorable) as well as the intensity of the evaluative response. Unfortunately, the obtained measures can be distorted by facial movements that are unrelated to the evaluative reaction, rendering their interpretation ambiguous in the absence of additional evidence.

Another potentially promising development involves the measurement of brain activity through *electroencephalography* (EEG), the assessment of small electric signals recorded from the scalp. This procedure, however, does not lend itself to a direct assessment of positive or negative responses. Instead, it capitalizes on the observation that unexpected stimuli evoke brain wave activity that differs from the activity evoked by expected stimuli. Hence, one may detect if a target object is evaluated positively or negatively by embedding its presentation in a long series of other objects with a known evaluation. The brain activity evoked by the target object will then indicate if its evaluation is consistent or inconsistent with the evaluation of the context objects.

4. Future Directions

Different attitude measures may suggest different conclusions, giving rise to debates over which one best reflects respondents' 'true' attitude. Subliminal exposure to 'cod liver oil,' for example, may facilitate your identification of 'disgusting' as a 'bad' word and may elicit a negative physiological response, yet you report that you think cod liver oil is good for you and you take it regularly. Which of these diverging responses should we consider an unbiased indicator of your 'true' attitude? Each measure reflects a different aspect of a complex phenomenon, rendering the designation of any single one as the only 'true' window on individuals' attitudes counterproductive. Understanding the interplay between cognitive, affective, physiological, and behavioral responses presents the most challenging issue for future research.

See also: Attitude Change: Psychological; Attitude Formation: Function and Structure; Attitudes and Behavior

Bibliography

Bassili J 2001 Cognitive indices of social information processing. In: Tesser A, Schwarz N (eds.) *Intraindividual Processes* (*Blackwell Handbook of Social Psychology, Vol. 1*). Blackwell, Oxford, UK, pp. 68–88

Banaji M R, Lemm K I, Carpenter S J 2001 The social unconscious. In: Tesser A, Schwarz N (eds.) *Intraindividual Processes* (*Blackwell Handbook Of Social Psychology, Vol. 1*). Blackwell, Oxford, UK, pp. 134–58

Cacioppo J, Petty R, Losch M, Crites S 1994 Psychophysiological approaches to attitudes: Detecting affective dispositions when people won't say, can't say, or don't even know. In: Shavitt S, Brock T C (eds.) *Persuasion*. Allyn and Bacon, Boston, pp. 43–72

Dovidio J F, Fazio R H 1992 New technologies for the direct and indirect assessment of attitudes. In: Tanur J M (ed.) *Questions About Questions*. Russell-Sage, New York, pp. 204–37

Krosnick J A, Abelson R P 1992 The case for measuring attitude strength. In: Tanur J M (ed.) *Questions About Questions*. Russell-Sage, New York, pp. 177–203

Krosnick J A, Fabrigar L in press *Handbook of Attitude Questionnaires*. Oxford University Press, Oxford, UK

Oskamp S 1991 *Attitudes and Opinions*. Prentice-Hall, Englewood Cliffs, NJ

Schuman H, Presser S 1981 *Questions and Answers in Attitude Surveys*. Academic, New York

Schwarz N, Bohner G 2001 The construction of attitudes. In: Tesser A, Schwarz N (eds.) *Intraindividual Processes* (*Blackwell Handbook of Social Psychology, Vol. 1*). Blackwell, Oxford, UK, pp. 436–57

Schwarz N, Hippler H J 1991 Response alternatives: The impact of their choice and ordering. In: Biemer P, Groves R, Mathiowetz N, Sudman S (eds.) *Measurement Error in Surveys*. Wiley, Chichester, UK, pp. 41–56

Schwarz N, Sudman S (eds.) 1992 *Context Effects in Social and Psychological Research*. Springer-Verlag, New York

Sudman S, Bradburn N M, Schwarz N 1996 *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco

Tourangeau R, Rips L J, Rasinski K 2001 *The Psychology of Survey Responses*. Cambridge University Press, Cambridge, UK

Winkielman P, Berntson G G, Cacioppo J T 2001 The psychophysiological perspective on the social mind. In: Tesser A, Schwarz N (eds.) *Intraindividual Processes* (*Blackwell Handbook of Social Psychology, Vol. 1*). Blackwell, Oxford, UK, pp. 89–108

Wyer R S, Srull T K 1989 *Memory and Cognition in its Social Context*. Erlbaum, Hillsdale, NJ

N. Schwarz

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Attitudes and Behavior

Although there is no single, universally accepted definition of 'attitude', social psychologists use the term to refer to a relatively enduring tendency to respond to someone or something in a way that reflects a positive or a negative evaluation of that person or thing. Eagly and Chaiken (1993, p. 155) define attitudes as 'tendencies to evaluate an entity with some degree of favor or disfavor, ordinarily expressed in cognitive, affective, and behavioral responses.' This immediately raises the issue of the relationship between attitude and behavior, since 'behavioral responses' are given as one of the ways in which an individual can express evaluation of the attitude object. If cognitive, affective, and behavioral measures of evaluative responses to an object are all indices of the same underlying construct, attitude, there should be some consistency between them. If measures of cognitive and affective responses to an attitude object fail to correlate with measures of behavioral responses, the widespread assumption that attitudes consist of these three components is called into question.

This 'three-component' view of attitudes is one theoretical basis for expecting attitudes to be correlated with behavior. Another is Festinger's (1957) theory of cognitive dissonance. Festinger argued that humans are characterized by a need to maintain consistency between their cognitions. These cognitions can be about one's behavior ('I told somebody else that I really enjoyed the meal') and about one's attitudes ('At the time I ate the food I did not like it very much'). If individuals become aware of inconsistencies between cognitions, they are said to experience 'cognitive dissonance,' an unpleasant state that they will be motivated to reduce or eliminate. Thus where an individual has had some behavioral experience with the attitude object, there should be a positive correlation between attitudes and behavior.

1. *Early Empirical Evidence*

There are various reasons, then, for expecting attitudes and behavior to be correlated. However, early empirical studies yielded evidence of a weak relation between these constructs. In 1969 Wicker published an influential review of 45 studies on the attitude–behavior relation. He found that the correlation between measures of attitude and measures of behavior rarely exceeded 0.30 and was often closer to zero. The mean correlation was 0.15. Wicker (1969, p. 65) concluded that 'it is considerably more likely that attitudes will be unrelated or only slightly related to overt behaviors than that attitudes will be strongly related to actions.' This led Wicker and others to suggest that the attitude concept should be abandoned. Although there was a crisis in confidence in the attitude–behavior relationship in the early 1970s, the attitude concept has not been abandoned; indeed, research on the attitude–behavior relationship has flourished.

2. *The Principles of Aggregation and Compatibility*

The field was transformed by Ajzen and Fishbein's (1977) paper, in which the authors developed two principles—the principle of aggregation, and the principle of compatibility. The principle of aggregation is a straightforward application to the attitude–behavior relation of ideas well known to psychometricians. A standard notion in psychological measurement is that one should not try to measure anything but the simplest of constructs using a single item. Precisely the same point applies to the measurement of attitudes and of behavior. In attitude measurement the logic of this argument is generally recognized, and present-day researchers would not attempt to assess attitudes using a single item. However, researchers have not always been as careful to apply these elementary principles to the assessment of behavior. Take DeFleur and Westie's (1958) study. They used a multiple-item measure of white persons' attitudes to black persons, but a single-item measure of behavior, namely whether or not a white participant was willing to be photographed with a black person of the opposite sex. How validly does this measure reflect a dispositional tendency to behave favorably or unfavorably towards black persons? Someone with nonracist attitudes might refuse to have his or her photograph taken because of a reluctance about being photographed. The solution is to use more than one item to assess the behavioral disposition to act positively or negatively towards black persons. This is the principle of aggregation: by aggregating across different measures of behavior toward the entity in question, one arrives at a more valid and reliable index of an evaluative tendency expressed in behavior. Thus it should be easier to find strong attitude–behavior correlations where both attitudes and behavior are measured using multiple items, and there is empirical evidence to support this (see Eagly and Chaiken 1993).

The principle of aggregation leads to the conclusion that general measures of attitude should be better predictors of general measures of behavior than of specific measures of behavior. The principle of compatibility is a logical extension of this argument, being more precise about what is meant by 'general' and 'specific.' Ajzen and Fishbein (1977) noted that a behavior varies not only with respect to the action performed; other elements are the target at which the action is directed, the context in which the action takes place, and the time at which it takes place. The principle of compatibility states that measures of attitudes and behavior are more likely to be correlated with each other if they are compatible with respect to

action, target, context, and time. Thus if your attitudinal measure is a general one, not tied to any specific action, target, context, or time, it should not be surprising to find that behavioral measures that are tied to specific actions, objects, contexts, and times correlate poorly with the attitudinal measure. The corollary is that if the behavior one seeks to predict is specific with respect to action, object, context, and time, then the attitudinal measure employed should be compatible with the behavioral measure in these respects. There is also empirical evidence to support this (see Eagly and Chaiken 1993).

Armed with these principles, we can account for the early failures to find substantial correlations between attitudes and behavior. DeFleur and Westie (1958), along with many others, used general measures of attitude to predict quite specific measures of behavior, usually involving a specific action toward designated objects in a certain context and at a particular time. Ajzen and Fishbein (1977) reviewed previous studies of the attitude–behavior relation in terms of whether or not the attitudinal and behavioral measures had been compatible. Where the measures were not compatible, the correlation was typically nonsignificant; but in all 26 studies in which the measures were compatible, the correlation was significant.

3. *The Theory of Reasoned Action*

These methodological analyses of the attitude–behavior relation were not the only responses to Wicker's (1969) conclusions. Social psychologists also began developing theoretical frameworks in which the attitude–behavior relation was nested within a broader causal model. The central point of these models is that thinking of behaviors as being determined exclusively or even principally by attitudes is an oversimplification. Moreover, this sits rather strangely in a discipline—social psychology—whose primary focus of concern can be regarded as the impact of social factors on behavior. Although attitudes are implicitly regarded as 'social products' by most attitude researches, in practice they are treated as attributes of the individual. Thus, much of the research effort on the attitude–behavior relation can be seen as focusing on the way in which an individual's behavior is shaped by personal attributes. Yet many of the classic studies in social psychology (e.g., Asch 1951, Milgram 1963) testify to the powerful effects that the implicit and explicit expectations of others can have on behavior. The best known of the theoretical frameworks in which the attitude–behavior relation was set in a broader causal model, incorporating the impact of social factors, is what came to be known as the theory of reasoned action (TRA) (Fishbein and Ajzen 1975).

The TRA has two key attributes. First, consistent with the principle of compatibility, it holds that strong relations between attitudes and behavior will only be found where attitudinal measures and behavioral measures are compatible with each other. Second, attitude is construed as just one determinant of behavior. The model is represented diagrammatically in Fig. 1.

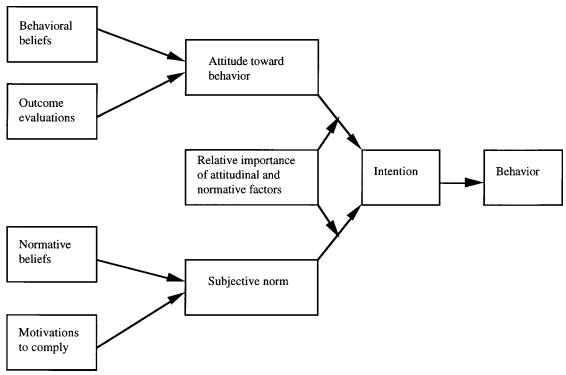

Figure 1
The theory of reasoned action (after Fishbein and Ajzen 1975)

The immediate determinant of behavior in this model is behavioral intention, or how the individual intends to act. This reflects an important assumption of the TRA, namely that most of the behaviors that social psychologists are interested in studying are 'volitional' in nature, in the sense that they can be performed (or not performed) at will.

Intention, in turn, is determined by the individual's attitude to the behavior in question, and by his or her subjective norm. Attitude to behavior is, of course, the individual's evaluation of the behavior. Subjective norm refers to the individual's belief that important others expect him or her to perform (or not to perform) the behavior. The relative contributions of attitude to behavior and subjective norm in determining behavioral intentions is left open; which of these factors is more important in shaping intentions is thought to depend on the behavior and the individual in question.

The TRA unfolds further by specifying the determinants of attitudes to behavior and of subjective norms. In the case of attitudes, the determinants are said to be behavioral beliefs, that is, beliefs about the consequences of performing the behavior; and outcome evaluations, or the individual's evaluation of each of those consequences. Thus attitude to behavior is treated as an evaluation of performing the behavior, based on the perceived likelihood that carrying out the behavior leads to outcomes of which one approves or disapproves. In the case of subjective norms, the determinants are said to be normative beliefs, or the individual's beliefs that each of a number of significant others expects him or her to act in a certain way; and motivations to comply, or the individual's inclination to conform to these other people's expectations. Thus

subjective norm is based on the perceived expectations of others, weighted by the extent to which one is inclined to fulfil those others' expectations.

The TRA has in many respects been highly successful. It has been widely used by social psychologists and other social scientists, and it has generally performed quite well in terms of the ability of the constructs in the model to predict behavior and behavioral intentions. The performance of the TRA has been assessed in various meta-analytic studies. In a widely cited meta-analysis of 87 separate studies, Sheppard et al. (1988) found a weighted average multiple correlation of 0.66 for the relationship between attitude and subjective norm, on the one hand, and behavioral intention, on the other. The average correlation between behavioral intention and behavior was 0.53. It is clear, then, that behavioral prediction can be achieved with reasonable accuracy if one assesses behavioral intentions, rather than attitudes; and that behavioral intentions, in turn, can be predicted with acceptable accuracy if one assesses attitude to behavior and subjective norm.

However, the TRA is concerned with more than simply the prediction of intentions and behaviors. It also attempts to provide an explanation for behavior, by specifying the determinants of intentions (i.e., attitude to behavior and subjective norm), as well as the factors that underlie these determinants (i.e., behavioral beliefs, outcome evaluations, normative beliefs, and motivations to comply). How does this help to explain behavior? Fishbein and Ajzen (1975) assume that the basic building blocks of volitional behavior are beliefs, values, and motives. Attitudes to behavior are built up through experience (direct and indirect) of the behavior in question, in the course of which individuals form and change beliefs about the consequences of the behavior. They may also form or change their evaluations of these consequences. If behavioral beliefs and outcome evaluations together determine an individual's attitude to a given behavior, and attitude to behavior in turn shapes behavioral intentions, identifying which behavioral beliefs and outcome evaluations distinguish people with positive intentions from those with negative intentions helps us to understand and explain behavior.

Exactly the same reasoning applies to the role of normative beliefs and motivations to comply in shaping subjective norms. To the extent that those who engage in a behavior (or who intend to engage in that behavior) differ significantly on these variables from those who do not, it can be argued that key determinants of the behavior in question have been identified. From an applied perspective, the importance of this process of identifying at the level of behavioral beliefs, outcome evaluations, normative beliefs, and motivations to comply which factors discriminate 'intenders' from 'nonintenders', or those who do perform the behavior from those who do not, resides in the enhanced potential for designing effective interventions. In devising such interventions (e.g., to quit smoking, engage in regular exercise, drive within the speed limit, use condoms), it clearly makes sense to focus the communication on those factors that are known to determine the behavior you are trying to change.

Successful though the TRA has undoubtedly been, it has certain limitations. One objection is that it is better suited to the prediction and understanding of challenging behavioral decisions that require the individual to deliberate the pros and cons of alternative courses of action than to the prediction and understanding of everyday, routine behaviors, of behaviors that have a strong habitual component. Much of our everyday behavior, according to this argument, is engaged in rather spontaneously and unthinkingly: It is more a matter of habit than of careful consideration. Consistent with this, various investigators have reported a significant direct link between previous behavior and current behavior, unmediated by the constructs of the TRA. This is inconsistent with the theory of reasoned action, which holds that the influence of past behavior on present behavior is mediated by constructs within the model. One explanation for this sort of finding is that both past and present behaviors are influenced by essentially noncognitive factors, such as habit, conditioning, or addiction (see Ouellette and Wood 1998).

One way in which Fishbein and Ajzen (1975) could respond to the challenge presented by the notions of habit, conditioning, and addiction would be to point out that the TRA was developed in order to predict and understand behaviors that are engaged in voluntarily by the individual. Habitual, conditioned, and addictive behaviors therefore lie outside the model's explanatory scope. However, the restriction to volitional behaviors also excludes any behaviors that require skills, resources, or the cooperation of others, a point made by Liska (1984). In other words, only the simplest and most trivial of behaviors are entirely under volitional control, in the sense that performance of such behaviors is dependent only on the individual's motivation to perform the behavior. This issue of the role played by control over behavior was a major reason for the development of a successor to the TRA, in the form of the theory of planned behavior.

4. *The Theory of Planned Behavior*

Ajzen (1985) proposed the theory of planned behavior (TPB) with the explicit goal of extending the explanatory scope of the TRA. Recognizing that at least some of the behaviors that social psychologists want to explain and understand are not under complete volitional control, Ajzen added a new construct, which he called perceived behavioral control. This construct represents the individual's perception of how easy or difficult it is to perform the behavior in question. A

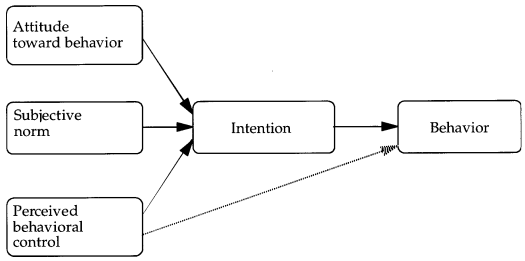

Figure 2
Theory of planned behavior (after Ajzen 1991)

behavior that is seen as easy to perform is high in perceived behavioral control; one that is seen as difficult to perform is low in perceived behavioral control. The TPB argues that an individual who has high perceived behavioral control with respect to a particular behavior is more likely to form the intention to perform that behavior, and is more likely to act on that intention in the face of obstacles and setbacks, than someone who is low in perceived behavioral control.

Just as attitudes to behavior and subjective norms are seen within the TRA as being founded on beliefs, so perceived control is regarded within the TPB as being founded on control beliefs. These are expected to reflect direct, observed, and related experiences of the behavior and 'other factors that may increase or reduce the perceived difficulty of performing the behavior in question' (Ajzen 1988, p. 135).

The TPB is shown in Fig. 2. It can be seen that perceived behavioral control determines both intentions (together with attitudes to behavior and subjective norms) and behavior (together with intentions). The joint determination of intentions is straightforward: it is assumed that when individuals form intentions they take into account how much control they have over the behavior. The joint determination of behavior (together with intention) can be understood in two ways. The first relates to motivation: an individual who has high perceived behavioral control and who has formed the intention to do something will simply try harder to carry out that action than someone with an equally strong intention but who has lower perceived behavioral control. The second explanation assumes that when someone has the intention to perform a behavior and fails to act on that intention, this failure is attributable to his or her lack of control over the behavior. The role of perceived behavioral control here is 'nonpsychological' in the sense that it is not the perception of control that causes the failure to act in accordance with intentions; rather, it is a lack of actual control. However, to the extent that perceived control is accurate thereby reflecting lack of actual control, a measure of perceived behavioral control should help to predict behavior. This is why the direct link between perceived behavioral control and behavior is depicted in Fig. 2 as a broken line, rather than a solid one: perceived behavioral control only helps to predict behavior if the individual has sufficient experience with the behavior to be able to make a reasonably accurate estimate of his or her control over the behavior.

Ajzen (1991) reviewed the findings of more than a dozen empirical tests of the TPB. In most of these studies the addition of perceived behavioral control to the TRA resulted in a significant improvement in the prediction of intentions and/or behavior. More recently, Godin and Kok (1996) reviewed the results of 54 empirical tests of the TPB within the domain of health behavior, and came to broadly similar conclusions. It is safe to assume that for most behaviors that are likely to be of interest to social scientists, it is worth using the TPB rather than the TRA.

5. *Direct Experience of the Attitude Object*

As we have just seen, direct experience of the attitude object can have an impact on the extent to which a dispositional construct (such as perceived behavioral control) is predictive of behavior. Fazio (1990) has developed a theoretical model in which the influence of direct experience on attitudes to an object is of central importance. His argument is that attitudes formed on the basis of direct behavioral experience with an object are more predictive of future behavior towards that object than are attitudes based on indirect experience. The rationale for this argument is that attitudes based on direct experience are stronger (they have greater clarity and they are held with greater certainty and confidence) and more accessible.

Attitude strength is seen as the strength of the association between an attitude object and the perceiver's evaluation of that object. The more experience one has with the object, the stronger will be this associative link. One's attitude will also be more accessible, in the sense that it will be more easily and quickly retrieved from memory. This is why attitudes based on direct behavioral experience are more predictive of future behavior, according to Fazio. Attitudes that are retrieved from memory easily and quickly are more likely to be activated whenever the attitude object is present. Frequent interaction with another person whom you like should lead to a highly accessible positive attitude to this person. The result is that whenever one is responding to the actual or symbolic presence of that person, whether this is in the context of a questionnaire assessing attitudes or in the context of an assessment of behavior toward that person, the same positive attitude should be elicited rapidly and automatically, and this attitude should guide both questionnaire responses and behavior. Attitudes and behavior should therefore be strongly correlated when the attitude is highly accessible.

Fazio operationalizes attitude accessibility as the time taken to respond to a question about that attitude. If an attitude is highly accessible it should take less time for people to retrieve their evaluation of the object in question from memory, and they should therefore be quicker to answer questions about attitudes to that object. Thus attitudes based on direct behavioral experience with an object should be (a) more accessible, in the sense that reaction times to questions about these attitudes should be faster; and (b) more predictive of behavior toward that object. There is empirical support for both these predictions (see Fazio 1990).

The significance of Fazio's work is that it shows that there are conditions under which reasonably strong attitude–behavior relations can be found even when one measures attitudes to the targets of behavior, rather than attitudes to behavior towards those targets. Thus the principle of compatibility can be violated without necessarily weakening the relationship between attitudinal and behavioral measures. For this to apply the respondent has to have a highly accessible attitude towards the target of the behavior.

6. Conclusion

Social psychologists' understanding of the attitude–behavior relation has developed considerably since early studies such as the one by DeFleur and Westie (1958). Wicker (1969) and others regarded these studies as evidence of a worrying inconsistency between what people say and what they do. The crisis that ensued has to a large extent been resolved, partly by a careful conceptual and methodological analysis of the problem, and partly by locating the attitude–behavior relation within more general theoretical frameworks.

See also: Action Planning, Psychology of; Attitude Change: Psychological; Attitude Formation: Function and Structure; Attitude Measurement; Cognitive Dissonance; Sexual Attitudes and Behavior

Bibliography

Ajzen I 1985 From intentions to actions: A theory of planned behavior. In: Kuhl J, Beckmann J (eds.) *Action Control: From Cognition to Behavior*. Springer-Verlag, Berlin, pp. 11–39

Ajzen I 1988 *Attitudes, Personality, and Behavior*. Open University Press, Milton Keynes, UK

Ajzen I 1991 The theory of planned behavior. *Organizational Behavior and Human Decision Processes* **50**: 179–211

Ajzen I, Fishbein M 1977 Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin* **84**: 888–918

Asch S 1951 Effects of group pressure on the modification and distortion of judgments. In: Guetzkow H (ed.) *Groups, Leadership, and Men*. Carnegie Press, Pittsburgh, PA, pp. 177–90

DeFleur M L, Westie F R 1958 Verbal attitudes and overt acts: An experiment on the salience of attitudes. *American Sociological Review* **23**: 667–73

Eagly A H, Chaiken S 1993 *The Psychology of Attitudes*. Harcourt Brace Jovanovich, Fort Worth, TX

Fazio R H 1990 Multiple processes by which attitudes guide behavior. The MODE model as an integrative framework. *Advances in Experimental Social Psychology* **23**: 75–109

Festinger L 1957 *A Theory of Cognitive Dissonance*. Row Peterson, Evanston, IL

Fishbein M, Ajzen I 1975 *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA

Godin G, Kok G 1996 The theory of planned behavior: A review of its applications to health-related behaviors. *American Journal of Health Promotion* **11**: 87–98

Liska A E 1984 A critical examination of the causal structure of the Fishbein–Ajzen attitude–behavior model. *Social Psychology Quarterly* **47**: 61–74

Milgram S 1963 Behavioral study of obedience. *Journal of Abnormal and Social Psychology* **67**: 371–8

Ouellette J A, Wood W 1998 Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin* **124**: 54–74

Sheppard B H, Hartwick J, Warshaw P R 1988 The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research* **15**: 325–43

Wicker A W 1969 Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues* **25**: 41–78

A. S. R. Manstead

Attributional Processes: Psychological

1. Background

Psychological research on attribution was inspired by Heider (1958), who developed models of attribution for object perception as well as person perception. Heider assumed, following the philosophical tradition of phenomenology, that a perceiver faced with observational data tries to attribute these data to their underlying cause—to an object in the case of sensory data, and to characteristics of the agent or the environment in the case of human behavioral data.

Heider recognized, however, that person perception is more complex than object perception—due to the manifold observational data available and the manifold causes (e.g., beliefs, desires, emotions, traits) to which these data can be attributed. Heider believed that people master this complexity by means of a sophisticated conceptual framework that he labeled 'commonsense psychology'—nowadays also called 'folk psychology' or 'theory of mind.' Using this conceptual framework, perceivers bring order to the massive stream of behavioral data by classifying them and attributing them to mental categories—including

motives, obligations, sentiments, and personality traits. For example, certain behaviors are classified as 'intentional actions', and the perceiver looks for the specific intention underlying each action (what the agent was trying to do).

Heider's research method was philosophical. He relied on conceptual analysis complemented by linguistic examples and thought experiments, but he provided no empirical experiments to test his model. The impact of Heider's work, then, was largely mediated by other researchers, who turned some of his ideas into empirical research programs. The evolution of attribution research can therefore be seen as a selective exegesis of Heider—with different scholars picking up different ideas and developing them in more detail.

This survey begins with a brief sketch of Heider's charismatic work, then turns to the two main strands of post-Heiderian attribution research: trait attribution work and causal explanation work. Finally, recent theoretical developments are addressed that return to some of Heider's original concerns and link attribution research to developmental research on children's theory of mind (see *Theory of Mind*) and philosophical work on folk psychology.

2. Heider

Despite his early focus on object perception, Heider's deepest concern was always with the topics of person perception and social interaction. Just like Lewin and Asch before him, Heider recognized that a psychology of social interaction must chart out people's *subjective* conceptions of the social world. At first Heider tried to develop a general theoretical framework to describe these conceptions, but then he realized that such a powerful framework already existed—it was commonsense psychology, the system of concepts ordinary people use to describe and understand human behavior. In his 1958 book, Heider thoroughly explored this framework, analyzing such commonsense concepts as *can*, *want*, *try*, and *ought*. Heider's blend of empirical observation with conceptual and linguistic analysis was visionary and unlike anything psychology had seen before—resembling more the tradition of ordinary language philosophy or contemporary cognitive linguistics.

Two major insights stand out from Heider's analysis of social perception. The first is that social perception is a process of reconstructing invariance out of variance. According to Heider, the social perceiver is presented with streams of behavior, and the cognitive system conceptualizes these streams in terms of a network of fundamental concepts, trying to reconstruct the invariances underlying behavior. Heider's notion of invariances encompassed perceptions, intentions, motives, abilities, and sentiments, which are all relatively invariant against the stream of ongoing behavior. Among these invariances, the agent's motives occupied a special role: 'The underlying causes of events, especially the motives of other persons, are the invariances of the environment that are relevant to [the perceiver]; they give meaning to what he experiences' (Heider 1958, p. 81). Heider's second major insight was that the reconstructive process of identifying invariances is a form of causal (i.e., attributional) analysis, rendering judgments of causality one of the primary elements in social perception. According to Heider, the commonsense conception of causality breaks down into two separate models. For one, there is *impersonal causality*, applied to unintentional human behaviors (such as sneezing or feeling pain) as well as physical events (such as stones rolling or leaves falling). For another, there is *personal causality*, applied only to human agents who perform an intentional action (such as cleaning the kitchen or inviting someone to dinner): 'Personal causality refers to instances in which p causes x intentionally. That is to say, the action is purposive' (Heider 1958, p. 100). For Heider, this distinction between unintentional and intentional behavior was central to commonsense psychology because it guides how people predict, evaluate, and influence social behavior.

In the literature, Heider's distinction between personal (intentional) and impersonal (unintentional) causality is typically portrayed as a distinction between internal (or person) causes and external (or situation) causes of behavior. But Heider did not subsume all internal attributions under the label 'personal causality'; only the attribution of purposive behavior to intentions and motives was characterized that way. The remaining internal causes fall under the 'impersonal' label, because they bring about behavior in a mechanical, nonintentional way. Thus, Heider drew a line between intentional and unintentional events, whereby intentional events are based on a special class of person causes (intentions and motives) and unintentional events can be based on person or situation causes.

3. Attribution Research After Heider

The rise of empirical attribution research in the 1960s and 1970s was sparked by a focusing and narrowing of Heider's broad model of social perception. Two main strands of research can be identified, each of which used Heider's ideas as a starting point and then developed them in more detail.

3.1 Attribution as Trait Inference

Heider theorized about the conditions under which people infer intentions from behavior. In their classic

paper, Jones and Davis (1965) took this idea a step further by developing a model of 'correspondent inferences' from a given intentional behavior ('he acted friendly') not just to a corresponding intention ('he tried to be friendly') but to a corresponding trait ('he is friendly'). The model identified conditions that increase the likelihood of such correspondent trait inferences. For example, the social perceiver will be more likely to infer a trait the more the action in question deviates from what is normally seen as desirable, and the more unique (positive) effects it has compared to alternative courses of action. If none of these conditions hold, the perceiver will not infer a correspondent trait but instead attribute the agent's behavior to forces of the situation.

The initial model was limited to intentional behaviors and predicted that people do not infer traits from unintentional behaviors or from behaviors performed under strong constraints (e.g., upon request by an authority). However, subsequent research found that people infer traits even from unintentional behaviors such as mistakes and accidents (e.g., 'he's clumsy') and even when the agent acted under strong constraints. This tendency to make correspondent trait inferences, even from single behaviors, and the parallel reluctance to attribute behavior to situational forces, was later dubbed the 'correspondence bias' or fundamental attribution error' (Ross 1977).

Researchers became increasingly interested in the question as to why people are so ready to infer traits from behavior and what conditions would moderate this bias. The currently dominant models of trait inference postulate a multistage process in which a behavior is identified as a certain type of action, a corresponding trait is then inferred by default, and only subsequent corrections take situational influences into account (Gilbert and Malone 1995). However, this default trait inference does not occur, for example, when the situational influence is salient or when the perceiver suspects that the agent acted out of ulterior motives. A potential limitation of trait attribution work is the scarcity of studies that examine the frequency and conditions of trait inferences in naturalistic social contexts.

3.2 Attribution as Causal Judgment

The second major strand of attribution research was inspired by Kelley's (1967) model of causal attribution. Focusing on Heider's insight that causal judgments are pivotal in social perception, Kelley proposed that such judgments are based on a simple information-processing rule: people infer those causes that covary with the event in question. Specifically, when an agent *A* behaves toward object *O*, the cause of the behavior is perceived to lie in *A* (internal attribution) if few other people behave as *A* does (low consensus), if *A* behaves the same way toward *O* over time (high consistency), and if *A* behaves the same way toward other objects (low distinctiveness). By contrast, the cause of the behavior is perceived to lie in *O* (external attribution) if most other people behave as *A* does (high consensus), if *A* behaves the same way toward *O* over time (high consistency), and if *A* behaves differently toward other objects (high distinctiveness). For Kelley, the principles of covariation applied equally to behaviors (whether intentional or unintentional) and physical events. His model was therefore embraced by causal reasoning researchers in both social and cognitive psychology.

Empirical tests of the covariation model tended to support it, although this support was limited to experimental settings in which explicitly presented covariation information had an effect on judgments. At the turn of the twenty-first century, no studies had demonstrated that people spontaneously, in everyday situations, seek out covariation information before answering a why question. Moreover, the covariation model makes no predictions about situations in which people lack covariation information (e.g., single observations) or in which they are not motivated to make use of the information (e.g., under time pressure). For those instances, Kelley later proposed additional rules of causal reasoning ('causal schemata'). One of them is the discounting principle, which states that, under certain circumstances, a second cause weakens the plausibility of a first cause. For example, if a student aces a difficult exam, we might explain it by assuming that he studied hard; but upon hearing that he cheated we may no longer believe that he studied hard, thereby discounting the previous explanation.

The problem that Kelley's causal attribution models tried to address was the 'causal selection problem'—how perceivers select particular causes for explaining a given behavior or event. Even though Kelley's models themselves did not completely solve this problem (alternative models were developed later), Kelley's work profoundly influenced attribution research by assuming that (a) people break down causes into internal ones (something about the agent) and external ones (something about the situation) and (b) that the internal–external dichotomy generally applies to all behaviors and events alike. This dichotomy proved to be a compellingly simple dependent variable that allowed researchers to explore a variety of interesting phenomena. One of these phenomena is the self-serving bias in explanations (see Miller and Ross 1975)—the tendency for people to explain their own positive and negative outcomes so as to maintain favorable self-perceptions or public impressions. For example, students would be expected to explain a good grade by citing internal causes (e.g., ability or hard work) and a bad grade by citing external causes (e.g., bad luck or an unreasonable teacher). Another important phenomenon is the actor–observer asymmetry, which is the tendency for people to explain their

own behaviors and other people's behaviors in systematically different ways. Specifically, Jones and Nisbett (1972) argued, and later studies confirmed, that people tend to explain their own behaviors by reference to external factors (e.g., 'I chose psychology as my major because *it's interesting*') but explain other people's behavior by reference to internal attributes (e.g., 'He chose psychology as his major because *he wants to help people*').

Some researchers proposed alternatives to the internal–external dichotomy of causes. For example, Weiner introduced two additional distinctions—one between stable and unstable causes and one between controllable and uncontrollable causes—and thereby improved predictions for people's emotions and motivations in the wake of explaining achievement outcomes, health outcomes, and deviant social conduct (see Weiner 1995). Abramson et al. (1978) introduced another distinction, between global and specific causes, to describe a hopeless explanatory style in depression (see *Depression*). However, most attribution work at the beginning of the twenty-first century still uses a single internal–external (or sometimes trait–situation) distinction.

The focus on internal vs. external causal attributions to all behaviors and events alike can be contrasted with Heider's claim that ordinary people sharply distinguish between (a) attributing *intentional actions* to the actor's motivation and (b) attributing *unintentional behaviors* (e.g., failure, depression) to causal factors internal or external to the agent. Heider felt, as an interview in 1976 reveals, that people's attributions of events were adequately depicted in Weiner's model, but that attributions of actions to motives were inadequately treated by contemporary research (Ickes 1976). A second limitation of contemporary research is that it developed primarily cognitive models of the attribution process, whereas Heider demanded careful attention to the communicative functions of attributions and the language of causality. Theoretical developments of the 1990s tried to address both of these potential limitations.

4. *Theoretical Developments of the 1990s*

4.1 *Attribution as Folk Theory of Mind*

Over the decades, attribution research has offered increasingly sophisticated information-processing models for the basic internal–external distinction, but it has not examined the distinctions that actually guide people's explanations within folk psychology—an omission that Kelley and Michela (1980) called the central irony of attribution research. The study of people's own concepts and distinctions has recently gained force among researchers who endorse what may be called a folk-theoretical approach to explanations. This approach considers behavior explanations an integral part of people's folk theory of mind and assigns the concept of *intentionality* a central role in these explanations, because people reliably distinguish between intentional and unintentional human behaviors and use distinct modes of explanation for each type of behavior (Buss 1978, Malle 1999).

In particular, unintentional behavior is explained by *causes*, perceived as simply bringing about the behavior in a mechanical way (e.g., 'She cried because *she was sad*'; 'He fell because *the floor was slippery*'). By contrast, intentional behavior is primarily explained by *reasons*, conceptualized as the beliefs and desires an agent considers when forming an intention to act (e.g., 'She quit her job because *she felt the pay was too low*'; 'They worked extra hours because *they needed to finish the project*'). Thus, when people offer reason explanations, they presume the agent's awareness of the reason content and a more or less rational process of forming an intention to act in light of those reasons. According to the folk-theoretical approach, the traditional internal–external dichotomy applies well to people's cause explanations of unintentional behavior but not to reason explanations of intentional behavior. For example, the above reason explanation, 'she felt the pay was too low,' refers to the agent's own belief ('she felt'), but the belief itself has a content that refers to the situation ('the pay was too low'), so it is unclear how this reason should be classified within an internal–external scheme. To analyze adequately people's reason explanations of intentional behavior, new theoretical categories have been developed that take into account the complex conceptual and linguistic nature of reasons (Malle 1999).

One advantage of the folk-theoretical approach lies in its compatibility with work on responsibility attribution. For example, Shaver (1985) argued that assignments of responsibility and blame are critically dependent on the intentionality of the behavior in question, and so the choice of explanations (e.g., causes vs. reasons) often manages assignments of blame. Another advantage of this approach lies in the integration of social-psychological work on explanations with developmental research on children's emerging theory of mind and philosophical analysis of the properties and evolution of folk psychology (Malle et al. 2001). These disciplines have highlighted people's unique conception and explanation of intentional action, and the next generation of attribution research will have to flesh out in detail the conceptual, linguistic, and social nature of these unique explanations.

4.2 *Attribution as Communication*

Heider examined the concepts of commonsense psychology in their everyday conversational context because he was convinced that tools of social perception help people to accomplish their goals in social communication and interaction. Along these lines,

innovative work by Hilton (1990) and Antaki (1994) suggests that attributions are not just cognitive processes but rather communicative acts that obey the rules of conversation; that is, behavior explanations are altered for impression-management purposes (e.g., to appear rational or fend off blame). Explanations are also responsive to an interlocutor's concerns and questions (e.g., 'How was this possible?' elicits different modes of explanation than 'For what reason?'). Finally, explanations have an immediate impact on others' perceptions and evaluations of the explainer and the agent whose behavior is being explained.

5. *Future Directions*

Promising new directions of attribution research may arise from a merging of the folk-theoretical and communicative approaches. By recognizing people's fine-grained theory of mind and their distinctions between various modes of explanation, researchers can more effectively trace the cognitive as well as communicative functions that explanations serve in social interaction. Classic phenomena such as the self-serving bias, the fundamental attribution error, and the actor–observer asymmetry may then be fruitfully re-examined, not just under laboratory constraints, but within their natural contexts of occurrence. Greater attention to the social context and functional roles of attribution processes may perhaps be the most important desideratum for future research. As a result, attribution theory might lead to powerful applications in the domain of social relations (e.g., conflict resolution, relationship dynamics) and would contribute to the scientific knowledge of both social perception and social interaction, reverberating Heider's (1958) timeless goal of providing a psychology of interpersonal relations.

See also: Control Behavior: Psychological Perspectives; Control Beliefs: Health Perspectives; Explanatory Style and Health; Small-group Interaction and Gender; Social Psychology, Theories of; Theory of Mind

Bibliography

Abramson L Y, Seligman M E P, Teasdale J D 1978 Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology* **87**: 49–74

Antaki C 1994 *Explaining and Arguing: The Social Organization of Accounts*. Sage, Thousand Oaks, CA

Buss A R 1978 Causes and reasons in attribution theory: A conceptual critique. *Journal of Personality and Social Psychology* **36**: 1311–21

Gilbert D T, Malone P S 1995 The correspondence bias. *Psychological Bulletin* **117**: 21–38

Heider F 1958 *The Psychology of Interpersonal Relations*. Wiley, New York

Hilton D J 1990 Conversational processes and causal explanation. *Psychological Bulletin* **107**: 65–81

Ickes W 1976 A conversation with Fritz Heider. In: Harvey J H, Ickes W J, Kidd R F (eds.) *New Directions in Attribution Research*. L. Erlbaum Associates, Hillsdale, NJ, Vol. 1, pp. 3–18

Jones E E, Davis K E 1965 From acts to dispositions: The attribution process in person perception. *Advances in Experimental Social Psychology* **2**: 219–66

Jones E E, Nisbett R E 1972 The actor and observer: Perceptions of the causes of behavior. In: Jones E E, Kanouse D, Kelley H H, Nisbett R E, Valins S, Weiner B (eds.) *Attribution: Perceiving the Causes of Behavior*. General Learning Press, Morristown, NJ, pp. 79–94

Kelley H H 1967 Attribution theory in social psychology. In: Levine D (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, Vol. 15, pp. 192–240

Kelley H H, Michela J L 1980 Attribution theory and research. *Annual Review of Psychology* **31**: 457–501

Malle B F 1999 How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review* **3**: 21–43

Malle B F, Moses L J, Baldwin D A (eds.) 2001 *Intentions and Intentionality: Foundations of Social Cognition*. MIT Press, Cambridge, MA

Miller D T, Ross M 1975 Self-serving biases in the attribution of causality: Fact or fiction?*Psychological Bulletin* **82**: 213–25

Ross L 1977 The intuitive psychologist and his shortcomings: Distortions in the attribution process. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 10, pp. 174–221

Shaver K G 1985 *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Springer-Verlag, New York

Weiner B 1995 *Judgments of Responsibility*. Guilford, New York

B. F. Malle

Auctions

1. *Introduction*

The use of auctions to allocate goods is quite common and has been known for centuries. Numerous and quite different goods are sold by auctions, that is, antiques, fish, flowers, real estate, gold, securities, and oil exploration rights. Cassady (1967) gives a complete list of goods sold by auction. The literature on auctions is enormous, and it is increasing rapidly. It is concerned mainly with revenue and efficiency considerations, namely to find the auction that yields the greatest expected revenues to the auctioneer and that allocates the goods to the bidders which value them most highly. Further issues, like the robustness of auctions, transactions costs, bid preparation costs, and the vulnerability to cheating are addressed rarely.

2. Definitions

According to Cassady (1967, p. 8), auctioning is

> a unique system of allocating … property based on price making by competition of buyers for the right to purchase.

More formally, Myerson (1981) defines an auction mechanism by (a) a pair of outcome functions, which gives the allocation of the goods and the bidders' fees in dependence of the bidders' strategies, and (b) a description of bidders' strategic plans. According to those definitions, there is a plethora of different auctions. To fix ideas, assume that there are N units of a homogenous good for sale and that each bidder is restricted to one unit of the goods. First of all, open and sealed-bid auctions can be distinguished. An auction is open, if the bids become known to all participants of the auction when submitted. In contrast to this, an auction is sealed-bid, if bidders submit their bids privately. Thus, bidders have no knowledge of the bids of their competitors.

The best known open auctions are the English and the Dutch auction (cf. Milgrom and Weber (1982)). In the English auction, the auctioneer raises the price up to the point, where N bidders remain, with the understanding that the N units of the good are sold to the remaining bidders at the last called-out price. In the Dutch auction the auctioneer lowers the price interrupted only by bidders accepting the last called-out price for one unit of the good. The auction ends when all units are sold or a lower price limit is reached.

The best known sealed-bid auctions are the discriminatory and the uniform-price (competitive) auction (cf. Smith 1966, Harris and Raviv 1981). In the discriminatory auction the N units are allocated to the bidders with the highest price bids in descending order. Successful bidders are charged an amount equal to their bids. Thus, for more than one unit for sale, different bidders pay different unit prices in a discriminatory auction. If there is only one unit for sale, the discriminatory auction is called first-price auction. The rules of the uniform-price auction are the same as those of the discriminatory auction, except for the prices paid by successful bidders. In the uniform-price auction all successful bidders are charged the same price, which equals the highest unsuccessful bid. If there is only one unit for sale, the uniform-price auction is called second-price auction or Vickrey auction. Vickrey (1961) introduced this type of auction.

Allowing for more than one seller, a double auction results. In a double auction sellers submit bids to sell and buyers submit bids to buy. The sellers' (buyers') bids are ranked in ascending (descending) order. The price in a nondiscriminatory double auction is determined by the intersection of the obtained supply and demand schedule. This auction as well as other types may be performed as a discrete-time auction at predetermined times during a trading period, or as a continuous-time auction at any moment during a trading period (cf. Friedman 1991).

The goods sold by auction can be classified with respect to the monetary value that bidders attach to them. Milgrom and Weber (1982) distinguish between goods with private value, goods with common value, and goods with general value. Goods are of the private value type (or of preference uncertainty (cf. Myerson (1981)), if its value to a bidder is a purely personal matter. A painting bought for purely personal reasons provides an example. Goods are of the common value type (or of quality uncertainty); (cf. Myerson 1981), if the monetary value bidders attach to it is the same for all bidders. This value, however, generally will be unknown at the time the auction takes place. Typically, securities have a future uncertain monetary value. Thus, they are of common value. A good is called to be of the general value type, if its value to a bidder is a combination of private and common values. A painting bought for both personal reasons and later resale could be considered being of the general value type.

3. Analyzing Auctions

The auction literature can be divided into three groups: *decision-theoretic auction models*, *game-theoretic auction models*, and *empirical studies*. Decision-theoretic auction models deal with the bidder's decision problem in a way which ignores her strategic response to the bidding behavior of her competitor. The game-theoretic approach, on the contrary, is based on this strategic aspect of the bidding process. Empirical studies make use of either the decision-theoretic or the game-theoretic approach when examing data sets resulting from real-life auctions or controlled laboratory experiments.

The earliest auction models are of the decision-theoretic type. They are motivated by real-world auction design problems concerning the sale of oil exploration rights (Hughart 1975) and US Treasury bills (Smith 1966, Scott and Wolf 1979). Formal in nature, they assume that bidders are expected utility maximizers, where the expectation is with respect to a subjective probability density function for the lowest accepted bid (cp., e.g., Smith 1966). In particular, it is assumed that a bidder's action does not affect the realization of the lowest successful bid (cp., e.g., Scott and Wolf 1979). Despite obvious modeling deficiencies, those early models already succeed in identifying certain effects, which later are proved rigorously within the game-theoretic approach to auction theory.

The earliest game-theoretic auction studies primarily are concerned with the determination of Bayes–Nash equilibrium bidding strategies. The models are quite simple in that they assume symmetric bidders with independent value estimates. Those results are generalized in different directions in order to account

for bidders with correlated value estimates, for goods of the general value type (Milgrom and Weber 1982), and for goods with a resale market (Bikhchandani and Huang 1989).

Empirical studies of auctions are concerned with the analysis of real-life auction data sets or controlled laboratory experiments. Examples are auctions of Federal leases on the Outer Continental Shelf (Hendricks and Porter 1988), US Treasury bills (Cammack 1991), wine (Ashenfelter 1989), and collectible trading cards over the Internet (Lucking-Reiley 1999).

4. *A Game-theoretic Auction Model*

For ease of exposition there is only one unit of goods to be sold in an auction. The *auctioneer*, who for simplicity is assumed to be the owner of the goods, sets the auction rules and makes them known to all participants. The auction rules are binding even though it might be in the auctioneer's interest to change them after the bids have been submitted. The auctioneer is risk-neutral, that is, interested in maximizing the expected revenues resulting from the sale of the goods. There are n risk-neutral bidders competing with each other in the auction. When preparing their bids, they do not cooperate. Every bidder i has some private information x_i, a realization of her information variable. Bidders are distinguished solely by their private information; otherwise they are identical. It is assumed that the number of bidders and the joint distribution of all information variables is common knowledge.

In an auction a rational bidder takes the bidding behavior of the competitors into account and behaves strategically. However, a bidder has only incomplete information about the relevant features of the bidding situation. Thus, an auction is to be considered a non-cooperative game with incomplete information among the bidders. To analyze this game the Bayes–Nash equilibrium concept (Harsanyi (1967/68) is employed. For that purpose, a bidder's strategy is defined as a real-valued function mapping possible realizations of her information variable into bids. A Bayes–Nash equilibrium $(B_1, \ldots, B_n)$ of bidders' strategies is described by the following properties. B_i maximizes bidder i's expected gains assuming bidding strategies $B_1, \ldots, B_{i-1}, B_{i+1}, \ldots, B_n$ of her competitors and the assumed bidding behavior is correct.

5. *Theoretical Results*

Fundamental to the theory of auctions is the bidding behavior under varying assumptions. This naturally includes alternative auction rules. In what follows the Dutch, English, first-price, and second-price auctions will be considered and compared.

The simplest set-up is the one where the good to be auctioned is of private value and the values that different bidders attach to it are independent. This assumption is well suited for goods without resale, but it excludes cases like the painting of an artist whose work has some prestige among collectors.

5.1 *The Private Value Model*

5.1.1 Independent private values. Although the rules of the standard open and sealed-bid auctions are quite different, Vickrey (1961) found an astonishing similarity. When comparing the Dutch and the first-price auction, he noted that a bidder would follow the same strategy in both auctions. In either auction, a bidder has to determine a price at which she is willing to accept the good. In the first-price auction this would correspond to her bid; in the Dutch auction this would be the lowest called-out price at which she would claim the good. Thus, the Dutch and the first-price auction are strategically equivalent. A similar argument applies to the English and the second-price auction. In either auction, a bidder would bid her private value. This ensures that she will accept all prices below and no price above her private value of the good.

To be more specific in the first-price auction, consider the decision problem of bidder i. Her expected gains are given by

$$x_i \times \text{probability of winning} - \text{expected payments} \quad (1)$$

where x_i denotes her private value. Note that lowering the bid has two opposing effects on expected gains. The probability of winning decreases and the gains from the bid increase. Intuitively, the optimal bid will be such that it balances those two effects, thus maximizing Eqn. (1).

Since bidders differ only by their private information, bidding behavior can be described by a common bidding strategy B. The resulting bidding strategies in the standard sealed-bid auctions are as follows, where F is the distribution function of X_i:

(a) *Bidding strategies* (Vickrey 1961, Harris and Raviv 1981).

(i) In the second-price auction the symmetric bidding strategy is given by

$$B(x) = x$$

(ii) In the first-price auction the symmetric bidding strategy is given by

$$B(x) = x - \frac{\int_0^x (F(y))^{n-1}\,\mathrm{d}y}{(F(x))^{n-1}} \quad (2)$$

From this a number of conclusions follow that are subject to empirical testing. Both in the first- and second-price auction, bidding strategies are increasing

in bidders' private values. In a first-price auction, bidding becomes more aggressive with an increasing number of bidders. A comparison of first- and second-price auctions yields that a bidder submits a lower bid in the first than in the later auction. This difference in bids is represented by the second term of Eqn. (2).

Uniqueness of symmetric equilibria can be established (cp. Milgrom and Weber 1982). Furthermore, efficiency is guaranteed in both types of auction, that is the bidder with the highest private value is the one who wins the good.

From the described bidding behavior, bidders' expected gains and the auctioneer's expected revenues can be derived. Both coincide in the case of first-price and second-price auction. This is a special case of the revenue equivalence theorem (Vickrey 1961, Harris and Raviv 1981, Myerson 1981, Riley and Samuelson 1981).

(b) *Revenue equivalence*. The auctioneer's expected revenues are the same in the single-unit English, single-unit Dutch, first-price, and second-price auction.

Similarly, a bidder's expected gains coincide in the above mentioned auction types.

5.1.2 Strictly affiliated private values. Suppose that bidders' private values are strictly affiliated, that is, large values of one variable are more likely, when the other variables are large as well. Milgrom and Weber (1982) introduced this generalization of nonnegative correlation. This concept allows for goods of private value, where bidders have some common liking.

Although Dutch and first-price auction are still strategically equivalent, this is no longer true for English and second-price auction. In particular, Milgrom and Weber (1982) established the following ranking of auctions.

(a) *Auction ranking*. The auctioneer's expected revenues in the first-price auction are never higher than those in the second-price auction. The auctioneer's expected revenues in the second-price auction are never higher than those in the single-unit English auction. The single-unit Dutch auction yields the same expected revenues as the first-price auction.

Similarly, in the second-price auction a bidder's expected gains are smaller or equal to those in the first-price auction.

5.2 The Common Value Model

In the common value model the unknown value of the good and bidders' information (about the unknown value) are affiliated. Auctions of financial securities like US Treasury bills and of Federal leases on the Outer Continental Shelf can be analyzed within this framework. The informational effect, however, makes bidding quite complicated and gives rise to the following phenomenon.

5.2.1 Winner's curse. Common value models suffer from an effect called the winner's curse.

> learning that others have bid less than [his own bid]… is bad news about the value of the item being acquired (Milgrom 1987, p. 18).

Unsophisticated bidding leads to the effect that the successful bidder tends to be the one who overvalued the goods. Substantial losses in common value auctions are quite often attributed to this effect. Thus rational bidders take account of the winner's curse effect, when preparing their bids. The 'winner's curse' was the subject of a lot of controversy (cf., e.g., Kagel and Levin (1986). Milgrom (1987) formalizes the winner's curse effect for a sealed-bid common value auction for one unit of the good, where the highest bid wins the auction. A bidder's expected unit value of the goods, when she has won with her bid, is smaller than her expected unit value. Thus, the winner's curse effect can be measured by this difference.

Any information on the unknown common value will weaken the winner's curse effect. In particular, linking the price in an auction to affiliated information of the common value reduces the winner's curse. This is the essence of the linkage principle (Milgrom and Weber 1982). Reducing the winner's curse effect leads to more aggressive bidding and higher auction prices. Also competition matters, that is, the winner's curse effect gets stronger with an increasing number of bidders.

5.2.2 Independent information variables. If the bidders' information variables are independent, the results of the common value model resemble those of the independent private value model. In particular, the auctioneer's expected revenues in the first-price and second-price auction coincide. Similarly, the same is true for a bidder's expected gains in both auctions.

5.2.3 Strictly affiliated information variables. The bidding behavior in the common value set up can be described in the same way as in the private value model. Again, Dutch and first-price auction are strategically equivalent. This, however, is not the case for the English and second-price auction. Fundamental to the analysis is a bidder's expected value of the good given her information and the fact of winning the auction with her bid. This way she takes care of the winner's curse effect. Taking this into account, existence and uniqueness of symmetric equilibria can be established (cf. Milgrom and Weber 1982).

The central results of the common value auction literature concern the ranking of auctions, the auctioneer's information policy and the information aggregation property of auctions. All those results rely on

the comparison of symmetric equilibria of the respective auctions. The basic results are as follows:

(a) *Auction ranking* (Milgrom and Weber 1982). The auctioneer's expected revenues in the first-price auction are never higher than those in the second-price auction. The auctioneer's expected revenues in the second-price auction are never higher than those in the single-unit English auction.

If the auctioneer has private information about the unknown common value, she may use this fact to influence the outcome of the auction. She may commit to supplying information publicly in order to increase revenues. There is a whole range of possible information policies. Milgrom and Weber (1982) mention complete revelation (fully reporting of all information), censoring (reporting of favorable information only), randomizing (reporting after having added noise to the information), and concealment (no reporting of private information).

(b) *Information revelation* (Milgrom and Weber 1982). The auctioneer's policy of revealing any affiliated information publicly and completely raises her expected revenues in the first-price auction, the second-price auction and the single-unit English auction.

There has always been the question of whether the price of a good of unknown value reflects (aggregates) the dispersed information of market participants. Under certain assumptions this could be answered in the affirmative in the case of auctions.

(c) *Information aggregation* (Wilson 1977). The winning bid in a first-price auction converges almost surely to the common value V as the number of bidders n goes to infinity.

5.3 The General Value Model

The general value model gives results in much the same way as the common value model. In particular, the auction ranking and the information revelation results hold true. However, the information aggregation property relies on the simple structure of the pure common value model and thus does not generalize.

6. Empirical Results

There is an increasing number of studies which provide tests of auction market behavior by means of experimental methods or real-life auction data sets (field data). Whereas the experimental approach allows for controlling a specific auction model including informational assumptions, the field approach provides a test of auction theory in situations where most theoretical assumptions are unobservable.

Quite often, experimental studies are concerned with the independent private values model, which predictions are being tested. The results obtained are mixed (cf. Smith 1991, Kagel 1995, and their citations for the following). Bidding behavior is characterized by deviations from the strategic equivalence of first-price and Dutch auctions and second-price and English auctions, respectively. In particular, in first-price (second-price) auctions bidders submit higher bids than in the corresponding Dutch (English) auctions. Consequently, revenue equivalence does not hold. However, as predicted by theory, individual bids are strictly increasing in bidders' private values. As concerns comparative statics, increasing the number of bidders leads to more aggressive bidding on the bidders' part in first-price auctions. When introducing affiliated private values in experiments, the resulting bids could be better explained by Bayes–Nash bidding behavior than by alternative ad hoc bidding models.

Fundamental to experimental work in the common value set-up is the study of Smith (1967), who examines individual bidding behavior in discriminatory and uniform-price auctions. His results are consistent with the predictions of auction theory. He establishes that bids in the discriminatory auction are lower than those in the uniform-price auction and that the same is true for the auctioneer's revenues. Other results are less supportive of auction theory (cf. Kagel and Levin 1986, Kagel 1995, and his citations for the following). As predicted by theory, individual bids are strictly increasing in bidders' information. However, comparative statics might fail. This is the case in second-price auctions, where increasing the number of bidders leaves bids unaffected.

In general, Bayes–Nash bidding behavior in sealed-bid auctions is questioned by strong winner's curse effects especially for inexperienced bidders. Furthermore, the common value auction rankings could not be confirmed in experiments. Due to the winner's curse effect, sealed-bid auctions yield higher revenues than English auctions. Even information revelation can be affected by the winner's curse effect. Kagel and Levin (1986) find that information revelation reduces average revenues, when the number of bidders is high. Only for smaller numbers of bidders, the theoretical predictions of information revelation could be confirmed. This effect applies to sealed-bid as well as English auctions. Thus, in first- and second-price auctions theoretical predictions are valid only when the winner's curse effect is weak.

Early results of field data support auction theory. For example, Tenorio (1993) establishes that in the Zambian foreign exchange uniform-price auction average revenues are higher than in the corresponding discriminatory auction. Lucking-Reiley (1999), on the contrary, reports a failure of revenue equivalence for auctions of collectible trading cards over the Internet. He finds that revenues in the Dutch auction are higher than those in the first-price auction. With respect to English and second-price auctions, however, the data

support revenue equivalence. This contrasts sharply with the results of independent private values experiments.

7. Double Auctions

Since the nineteenth century, trade at organized exchanges to a large extent was conducted by means of double auctions. Nowadays, double auctions are important tools for computerized trade. They also serve well for the organization of markets for natural gas as well as for electric power networks. The theoretical analysis concerning double auctions is much more involved than that of one-sided auctions. Determination of Bayes–Nash equilibrium is complicated, if at all feasible. Often the analysis gives rise to a multiplicity problem. For that reason, the theoretical analysis centers around the independent private values model for a single indivisible unit. Satterthwaite and Williams (1993) give results for particular sealed-bid nondiscriminatory double auctions. For m risk-neutral sellers and m risk-neutral buyers symmetric Bayes–Nash equilibria are determined. Convergence of equilibria to ex post efficiency as both the number of sellers and buyers increase is shown to hold.

However, due to the abovementioned technical problems, the analysis of double auctions is done almost always by means of empirical or experimental studies (cf., e.g. Friedman and Rust 1993, Sect. 3).

8. Application to Financial Markets

Interest in the market institutions that govern the trade of financial securities grew rapidly during the last decades of the twentieth century. Along with the successful introduction of new exchanges, the trading rules at existing exchanges were examined more closely. Common to all those activities is an interest in the price formation process under alternative trading rules. An important example of this approach is the discussion of the design of the US Treasury bill auction, one of the most important auctions of financial securities. Every Monday the US Treasury auctions a previously announced quantity of Treasury bills maturing in 91 and 182 days, respectively.

Since the late 1950s there has been quite a controversy concerning the design of this auction. This controversy started with a proposal from Milton Friedman to change the Treasury bill auction from discriminatory to uniform-pricing. The reasons given in favor of this proposal were quite different in nature. They ranged from revelation of bidders' true demand curves and an increase in the number of bidders to discouragement of collusion under uniform-pricing. They did not, however, include an increase in the Treasury's average revenues from a sale under uniform-pricing. In fact, the early opponents to this proposal argued by a reduction in revenues through a change from discriminatory to uniform-pricing. Several years later, models of the decision- and the game-theoretic type lent additional support to the Friedman proposal.

In that respect, the results of Smith (1966) and Milgrom and Weber (1982) on the superiority of uniform over discriminatory pricing in terms of auctioneer's expected revenues were especially successful. Despite the fact that a completely satisfactory theoretical model of the Treasury bill auction is not available yet, those results made the Treasury reconsider the Friedman proposal. Those considerations led to experimentation with a uniform-price auction in the 1970s. However, the Treasury retained the discriminatory rules. Only after the Salomon Brothers Inc admission of violating the US Treasury auction rules, the experiments were taken up again in September 1992 for the two-year and five-year note auctions (cf. Bikhchandani and Huang (1993)). Experimentation ended in a decision effective with the February 11, 1998 auction to auctioning all bills using the uniform-pricing rules.

9. Conclusion

The basic game-theoretic auction model gives rise to a number of strong results which have been subject to empirical and/or experimental tests. Naturally, the question arises as to whether those results are robust with respect to a relaxation of assumptions. Unfortunately, there are several assumptions which, when relaxed, give rise to different results. There are a number of factors which have to be dealt with. First, risk neutrality matters. When allowing for risk aversion on the bidders' part, the expected revenue comparison will yield different results. For example, in models with both risk-aversion and strictly affiliated information variables, the established ranking of first- and second-price auctions with respect to expected revenues fails to hold. Only in special cases (e.g., independently and identically distributed (i.i.d) information variables) general results can be established (cf. Milgrom and Weber 1982).

Second, multiunit demand by bidders changes results dramatically. Endogenizing quantity, for example, as in the case of a share auction, might result in a multiplicity of equilibria and a substantial lower expected sale price than in comparable unit auctions (cp. Wilson 1979). Third, in auctions of financial securities bidders usually meet more than once. The introduction of this phenomenon produces all effects that apply to repeated noncooperative games. Finally, a number of features cannot be solved in the outlined theoretical auction model due to mathematical complexities. It is in those cases that one has to rely on controlled laboratory experiments in order to draw conclusions.

See also: Decision-making Systems: Personal and Collective; Game Theory; Game Theory and its Relation to Bayesian Theory

Bibliography

Ashenfelter O 1989 How auctions work for wine and art. *The Journal of Economic Perspectives* **3**: 23–36

Bikhchandani S, Huang C 1989 Auctions with resale markets: an exploratory model of Treasury bills markets. *The Review of Financial Studies* **2**: 311–39

Bikhchandani S, Huang C 1993 The economics of Treasury securities markets. *Journal of Economic Perspectives* **7**: 117–34

Cammack E B 1991 Evidence on bidding strategies and the information contained in Treasury bill auctions. *Journal of Political Economy* **99**: 100–30

Cassady R Jr 1967 *Auctions and Auctioneering*. University of California Press, Berkeley, CA

Friedman D 1991 The double auction market institution: A survey. In: Friedman D, Rust J (eds.) *The Double Auction Market. Institutions, Theories and Evidence*. Addison-Wesley, Reading. MA, pp. 3–25

Friedman D, Rust J (eds.) 1993 *The Double Auction Market. Institutions, Theories and Evidence*. Addison-Wesley, Reading, MA

Harris M, Raviv A 1981 Allocation mechanism and the design of auctions. *Econometrica* **49**: 1477–99

Harsanyi J C 1967, 1968 Games with incomplete information played by Bayesian players, Part I. *Management Science* **14**: 159–82, Part II. *Management Science* **14**: 320–34, Part III. *Management Science* **4**: 486–502

Hendricks K, Porter R H 1988 An empirical study of an auction with asymmetric information. *American Economic Review* **78**: 865–83

Hughart D 1975 Informational asymmetry, bidding strategies, and the marketing of offshore petroleum leases. *Journal of Political Economy* **83**: 969–85

Kagel J H 1995 A survey of experimental research. In: Kagel J H, Roth A E (eds.) *The Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ, pp. 501–85

Kagel J H, Levin D 1986 The winner's curse and public information in common value auctions. *American Economic Review* **76**: 894–920

Lucking-Reiley D 1999 Using field experiments to test equivalence between auction formats: magic on the Internet. *American Economic Review* **89**: 1063–80

Milgrom P R 1987 Auction theory. In: Bewley T (ed.) *Advances in Economic Theory*. Cambridge University Press, Cambridge, UK, pp. 1–32

Milgrom P R, Weber R J 1982 A theory of auctions and competitive bidding. *Econometrica* **50**: 1089–1122

Myerson R B 1981 Optimal auction design. *Mathematics of Operations Research* **6**: 58–73

Riley J G, Samuelson W F 1981 Optimal auctions. *American Economic Review* **71**: 381–92

Satterthwaite M A, Williams S R 1993 The Bayesian theory of the k-double auction. In: Friedman D, Rust J (eds.) *The Double Auction Market. Institutions, Theories and Evidence*. Addison-Wesley, Reading, MA, pp. 99–123

Scott J H, Wolf C 1979 The efficient diversification of bids in Treasury bill auctions. *Review of Economics and Statistics* **61**: 280–7

Smith V L 1966 Bidding theory and the Treasury bill auction: does price discrimination increase bill prices? *Review of Economics and Statistics* **48**: 141–6

Smith V L 1967 Experimental studies of discrimination versus competition in sealed-bid auction markets. *Journal of Business* **40**: 56–84

Smith V L 1991 Auctions. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics*. Macmillan, London, pp. 138–44

Tenorio R 1993 Revenue equivalence and bidding behavior in a multi-unit auction market: an empirical analysis. *Review of Economics and Statistics* **75**: 302–14

Vickrey W 1961 Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* **16**: 8–37

Wilson R B 1977 A bidding model of perfect competition. *Review of Economic Studies* **44**: 511–18

Wilson R B 1979 Auctions of shares. *Quarterly Journal of Economics* **93**: 675–89

S. Müller

Audience Measurement

Audience measurement, as the term is commonly used, refers to regular assessments of the size and composition of media audiences. It is based on survey research wherein individual attributes (e.g., age, gender, etc.) are quantified to produce statistical summaries of various audience aggregates. These numbers, which include broadcast audience ratings, are sold in the form of syndicated reports that are indispensable to advertiser-supported media.

1. Uses of Audience Measurement

The major impetus for audience measurement is advertising. By the late nineteenth century most newspapers and mass circulation magazines had begun to sell space to advertisers. The value of space was determined largely by the number of readers who would see each publication, and by extension each ad. Paid circulation served as a reasonable surrogate for the size of the readership. Many publishers, however, made inflated claims about their circulation, undermining the confidence of advertisers. To make print media credible, in 1914 the industry created the Audit Bureau of Circulations, an independent organization that verified circulation (Beniger 1986). To this day, such audits continue in the USA and many other countries around the world.

The growth of radio in the 1920s presented new problems for those who would use it as an advertising medium. Unlike print, there were few traces of the radio audience save for fan mail or the sheer number of receivers being sold. For all intents and purposes,

the radio audience was invisible. To provide reliable estimates of the number of people listening to specific programs, the industry sponsored audience surveys. By the 1930s independent research firms had begun to produce what were called 'ratings' (i.e., percentages of the population tuned to a particular program or station). These made the audience visible and allowed advertisers to buy spot announcements in radio with a degree of certainty. When television and other forms of electronic media emerged in the latter half of the century, similar mechanisms for measuring the size and composition of audiences were quickly put in place (Beville 1988).

Today, advertising is a multi-billion dollar business. Audiences are bought and sold like commodities. Advertising expenditures are typically guided by audience measurement and the cost of reaching various audience segments. In a business world increasingly interested in target marketing, research firms have been called upon to produce ever finer demographic distinctions, as well as data on lifestyles and product purchases. Curiously, all published audience measurements describe something that has already happened, while all advertising expenditures are made in anticipation of audiences that have yet to materialize. This odd system works because, in the aggregate, media audiences are quite predictable. For example, advertisers and networks know the total number of people who are likely to watch television at any time during the year, and negotiate ad prices on that basis. The stability of audience behavior has also allowed researchers to identify various 'law-like' features of the mass audience (Comstock and Scharrer 1999, Goodhardt et al. 1987, Webster and Phalen1997).

The fact that advertising is the major source of revenue for several forms of media (including broadcasting, newspapers, and magazines) has embedded audience measurement in the operation of these industries. Obviously, the system places a premium on audiences that will be attractive to advertisers, either by virtue of their sheer size or desirable composition. Content, therefore, is evaluated with an eye toward its audience-making potential. Moreover, broadcasters pay careful attention to how audiences 'flow' across an entire schedule of programming, relying on established patterns of media use to deliver people to their program offerings. In all of these endeavors, audience measurement is crucial. In fact, a good case can be made that audience measurement is a necessary condition for the emergence of any form of advertiser-supported media (Ettema and Whitney 1994).

Even in social systems not fully committed to advertising, audience measurement can become a fixture. Public service broadcasting around the world must often justify its existence by demonstrating that it has an audience (Ang 1991). Here the techniques of audience measurement are similar to those in commercial systems, though they may involve the use of more 'qualitative' measures, such as the appeal or usefulness of media materials (Beville 1988, Lindlof 1995). Similarly, government officials concerned with good or bad media effects often relate these to the audience and the magnitude of exposure to good or bad content (Webster and Phalen 1997).

2. *Methods of Audience Measurement*

As a type of survey research, audience measurement confronts issues common in sampling and statistical inference. Its distinctiveness comes from the methods it uses to measure the behavior of audience members. One of three techniques is typically employed: interviews, diaries, and meters.

The first regular surveys of radio audiences were conducted using telephones. Respondents were called at home and asked either what they were currently listening to (i.e., telephone coincidental method) or what they had listened to in the recent past (i.e., telephone recall method). Both techniques are used today. In fact, telephone coincidentals are generally regarded as the most accurate way to measure broadcast audiences and are, therefore, the standard against which other methods are judged (Beville 1988).

The advent of radio audience measurement forced print media to go beyond circulation as the sole indication of readership. Since a single copy of a magazine or paper might well be read by more that one person, circulation alone seemed to understate the audience for print. The industry eventually adopted survey research to better estimate readership. Personal interviews were conducted in which respondents were shown stripped down copies of various magazines and asked if they recalled reading them (i.e., through-the-book method). Alternatively, to economize on time and increase the number of publications that could be assessed, respondents were asked to sort through cards containing magazine logos and indicate which of these were read (i.e., recent reading method). Today, the recent reading method remains the standard of magazine audience measurement (Bogart 1996). Between the 1940s and 1970s local radio audiences were also measured using personal interview techniques. Respondents were presented with a list of radio stations and asked which they listened to (i.e., the roster recall method). Eventually the roster recall technique gave way to the use of diaries (Beville 1988).

Diaries are small booklets in which respondents keep a log of their radio or television usage throughout the day. Typically, diaries are kept for a week, though time periods of different durations are possible. Diaries are a relatively inexpensive data collection technique and, properly filled out, contain a wealth of information. However, many people are reluctant to accept a diary, or are unwilling or unable to make accurate entries. In a world of push-button radios, multichannel television, and remote control devices, completing a diary can be a burdensome task. Nonetheless,

diaries are a common measurement technique for producing local radio and television audience reports (Webster and Wakshlag 1985).

Many of the problems associated with diaries can be solved by the use of meters. These devices attach to a receiver and automatically record the source to which the set is tuned. By the early 1940s Arthur Nielsen had developed such a device and had placed it in a panel of American households to produce national radio network ratings. Household meters were eventually adapted to the new medium of television. While household meters eliminate various problems of response, they are expensive and offer no information about who is actually using the set. With advertisers increasingly concerned about demographics, the latter is a serious shortcoming. To address it, meters were adapted so that people could identify themselves by pressing a button. This so-called 'peoplemeter' is the technique now used to produce network television ratings in the United States and several countries around the world (Webster et al. 2000).

The fact that respondents need to actively signal their presence before the set by pressing a button reintroduced response errors. Children, for example, are not particularly diligent button pushers. From the industry's point of view, the ideal measurement device would be a passive peoplemeter, one that would somehow register who was watching the set without requiring any action on the part of viewers. Experiments have been conducted using infrared sensing and image recognition technology, but at the time of writing no such system has been put into commercial operation. Pilot studies have also been conducted on wireless devices that sense an inaudible code placed in a radio or television signal. In one manifestation, a pager-like device worn by a respondent picks up and identifies all radio programming within earshot of the listener.

The rapid growth of the Internet in the 1990s and the prospect of using the World Wide Web as an advertising medium motivated a number of companies to measure Internet use. While individual web sites could track people who visited their pages, what was needed was a 'user-centric' form of measurement to track people from site to site. Firms providing this type of measurement typically create a panel of computer users who agree to have monitoring software loaded on their machines. These programs record the various web pages the user visits. Data are then aggregated to produce web site ratings not unlike television ratings (Webster et al. 2000).

3. *Sources of Error in Audience Measurement*

There are four sources of error in audience measurement: sampling, nonresponse, response, and processing. The first three are, again, problems commonly associated with survey research. The last includes a variety of issues that have to do with bringing a saleable research product to market.

Most syndicated audience measurement is based on some form of probability sampling. Compounding considerations of sampling error is the tension between sample size and the growing complexity of the media environment. Virtually all forms of mass media are becoming more diversified: there are dozens of television networks available on broadband distribution systems, there are hundreds of special interest magazines, and there are an untold number of web sites. This, in combination with the need to estimate ever narrower audience segments, constantly presses sample sizes to the limit. In fact, samples for estimating magazine readership and Internet use already number in the tens of thousands. Unfortunately, the high cost of some measurement devices (e.g., meters) and the diminishing marginal returns associated with increases in sample size have forced difficult compromises on sample size and allowable levels of sampling error. Moreover, individual subscribers who manipulate databases to suit themselves are under no restrictions as to how they 'crunch the numbers' or the sampling error attendant to the estimates they produce.

There are also potential problems of response. Written forms of feedback (e.g., self-administered questionnaires, diaries, etc.) often suffer from low rates of response. Further, as people become increasingly leery of telemarketers, telephone survey completion rates are threatened. All these lead to concerns about non-response (i.e., the people being studied are systematically different from those who don't respond). Even among those who agree to cooperate, there may be problems with the quality of their responses (e.g., response error). There is anecdotal evidence that people, increasingly aware of the impact of audience ratings, can deliberately misrepresent their behaviors (Beville 1988). More likely, the flood of modern media simply overwhelms the reporting capabilities of even the most conscientious respondent.

The information collected by interviews, diaries and meters can be thought of as raw material that must be processed into a saleable product. Data must be coded and checked for obvious errors. Missing values are occasionally imputed using a process called ascription. New information, like program schedules, must be added. Responses from under-represented or over-represented groups are sometimes assigned a mathematical weight in a process called sample balancing. Responses generated by one method may be adjusted to parameters set by another method in a controversial process called calibration. Finally, data must be aggregated and projected to form the population estimates in published reports. In all of these matters, errors of fact or judgment may creep into the process (Bogart 1996, Webster et al. 2000).

4. Institutional Aspects of Audience Measurement

In most developed countries, audience measurement is a business that sells research reports or data to multiple subscribers. Rarely do more than two firms compete for any given market. Sometimes only one firm provides the measurements for an entire medium. As such, their reports take on the air of quasi-public, official documents. Though various errors in the data mean the numbers are not nearly as 'hard' as they appear, the organizations that use them have little alternative but to treat them as such. And an enormous amount rides on those numbers. The process of allocating billions of dollars in advertising and programming resources is informed, some would say tyrannized, by audience measurement (Ang 1991, Bogart 1995).

The institutional significance of audience measurement has not gone unnoticed. In the 1960s the United States Congress held hearings into the quality and consequences of the American broadcast ratings industry. Public scrutiny induced the industry to create an ongoing council to audit and accredit ratings research methods, as well as sponsor industry-wide methodological research (Beville 1988). For these reasons, and because audience measurements are used by powerful institutions with competing interests, it is probably the case that no research firm can drift too far from truthful reporting without sowing the seeds of its own destruction.

Even so, much about audience measurement is negotiable and left to industry consensus. Like any self-interested business, research firms try to adapt their products to client needs. This has led to charges that traditional audience measurement fails to address factors that others find important (i.e., people's attentiveness, preferences, needs, or understandings). It has also given rise to a more cynical view that research companies simply adopt whatever methods and produce whatever results serve dominant institutional interests (Meehan 1990).

The matter of turning the audience into a commodity has also drawn critics. Audience measurement essentially lumps people into categories, with one person being the functional equivalent of all others in the group. These groups are then sold to advertisers at some 'cost-per-thousand.' At best, the business is dehumanizing. At worst, it has been argued, the entire enterprise serves to colonize, manipulate, and ultimately victimize audiences (Ang 1991).

Alternatively, audience measurement can be seen as exercising a democratizing influence on popular culture. Material that is successful in the marketplace proliferates, while content that is unpopular fades away. While this notion of cultural democracy is oversimplified, a case can be made that measurement empowers audiences by casting them in a form to which institutions can and do respond (Webster and Phalen 1997). One way or another, it can certainly be said that audience measurement helps shape mass culture.

See also: Advertising: Effects; Advertising: General; Audiences; Broadcasting: General; Entertainment; Market Research; Mass Media, Political Economy of; Media Effects; Media, Uses of; Radio as Medium; Television: Industry

Bibliography

Ang I 1991 *Desperately Seeking the Audience*. Routledge, London

Beniger J R 1986 *The Control Revolution: Technological and Economic Origins of the Information Society*. Harvard University Press, Cambridge, MA

Beville H M 1988 *Audience Ratings: Radio, Television, Cable*. Rev. edn. Erlbaum, Hillsdale, NJ

Bogart L 1995 *Commercial Culture: The Media System and the Public Interest*. Oxford University Press, New York

Bogart L 1996 *Strategy in Advertising: Matching Media and Messages to Markets and Motivations*. 3rd edn. NTC Business Books, Lincolnwood, IL

Comstock G, Scharrer E 1999 *Television: What's On, Who's Watching, and What It Means*. Academic Press, San Diego, CA

Ettema J S, Whitney D C 1994 *Audiencemaking: How the Media Create the Audience*. Sage, Thousand Oaks, CA

Goodhardt G J, Ehrenberg A S C, Collins M A 1987 *The Television Audience: Patterns of Viewing*. 2nd edn. Gower, Aldershot, UK

Lindlof T R 1995 *Qualitative Communication Research Methods*. Sage, Thousand Oaks, CA

McQuail D 1997 *Audience Analysis*. Sage, Thousand Oaks, CA

Meehan E R 1990 Why we don't count. The commodity audience. In: Mellencamp P (ed.) *Logistics of Television: Cultural Criticism*. Indiana University Press, Bloomington, IN

Webster J G, Phalen P F 1997 *The Mass Audience: Rediscovering the Dominant Model*. Erlbaum, Mahwah, NJ

Webster J G, Phalen P F, Lichty L W 2000 *Ratings Analysis: The Theory and Practice of Audience Research*. Erlbaum, Mahwah, NJ

Webster J G, Wakshlag J 1985 Measuring exposure to television. In: Zillmann D, Bryant J (eds.) *Selective Exposure to Communication*. Erlbaum, Hillsdale, NJ

J. G. Webster

Audiences

The notion of audience in social research has been largely derived from the image media producers have of the actual, or the intended, people or groups of people that they imagine as the main recipients of their products. Media products—newspapers, television shows, films, radio broadcasts—are for the most part

manufactured with the aim of capturing the attention of certain of these people, the audience members. Often this interest is commercial: When media can capture the attention of particular audiences, their producers can sell this attention to various advertisers who may profit from this exposure.

While a long tradition of scholarship has accepted this idea of the audience, other thinkers have been rancorous in their disapproval of the commercial nature of the very concept. According to Ang, the '"television audience" only exists as an imaginary entity, an abstraction constructed from the vantage point of the institutions, in the interest of the institutions' (Ang 1991, p. 2). Nevertheless, scholars as well as market researchers study these groups of people that media products assemble. Audience research has been partitioned according to the type of media consumed. Different media have inspired different methods of primary research, as media differ in how they are received, and in the type of access scholars have to studying their modes of reception. The film audience has been analyzed most often from a historical or a psychoanalytic perspective, for example, while the television audience has been studied from a wide variety of perspectives, humanistic as well as social-scientific. Recent studies of the audiences for television, popular fiction, and popular music have used ethnographic methods. The trend toward ethnography attempts to recognize that the media 'audience' cannot be entirely separated from the role that media play in the everyday lives of the people who are medias' users, spectators, and viewers.

1. *Changes in Focus Over Time*

Audience research grew out of theoretical debates about the nature of public opinion and knowledge, and concern about the influence of the mass media on it. Early theories posited a 'hypodermic needle' model of media influence on its audience (sometimes called the 'magic bullet' theory), whereby audience members were powerfully influenced directly by media exposure ('injected' with ideas by the media, so the metaphor goes). This model of powerful media influence over a passive audience was supplanted by the work of Paul Lazarsfeld in the 1940s and 1950s. Directing a research program at the Bureau of Applied Social Research, affiliated loosely with Columbia University, Lazarsfeld had strong ties to industry and private money to finance targeted studies of the media audience.

Along with Elihu Katz, Lazarsfeld elaborated a 'limited effects' model in the work *Personal Influence* (Katz and Lazarsfeld 1955). This became famous for elaborating the 'two-step' flow model of the influence of the media on the audience. Instead of directly and powerfully influencing people, Katz and Lazarsfeld argued, media influence was mediated by influential people whom they called 'opinion leaders,' to whom audience members attended more strongly than they did directly to media itself. In this work, Katz and Lazarsfeld integrated the insights of social-psychological small-group theory into their approach to media influence. In general, the study confirmed the idea that media effects, rather than being powerful and evil, were actually for the most part rather minor and noninvasive, and that the media audience, as a social entity, was more powerful and resistant to them than previously thought. The limited effects model was influential in audience research throughout the 1960s. While limited effects research does not necessarily argue that mass media have no impact, this body of research generally asserts that its effect is primarily to reinforce existing opinions.

One critical response to the limited effects model was 'agenda setting' research (see *Agenda-setting*). This work argues that the mass media don't tell people what to think, but they do tell people what to think about. In particular, agenda-setting research has focused on the impact and meaning of news, and often argues that news has a structuring impact on the way both viewers and social policy makers prioritize society's social and political issues. Many of the insights of agenda-setting research have informed the cultural approaches to audience analysis discussed below.

Another response came in a direct, vitriolic critique of limited effects theory by Todd Gitlin (1978), widely read and cited, and having an important influence on audience research since the early 1980s. Gitlin accuses Lazarsfeld and his followers of misconceiving the problem of media effects. By conceptualizing an 'effect' too narrowly and concretely, they miss the many less measurable ways the mass media influence their audience. Much of this influence, Gitlin asserts, is not particularly amenable to the kind of 'scientific' study that Lazarsfeld championed. When media effects on their audience are too broad and diffuse, they can be theorized, described, and analyzed, but not necessarily measured, and this, Gitlin argued, was the nature of the modern mass medias' impact on their audience.

Other audience scholars working simultaneously to Gitlin developed what has come to be known as the 'uses and gratifications' tradition of audience research. Uses and gratifications research emerged from the functionalist perspective in US sociology, which was ascendant in the 1950s and 1960s. It focused on the active processes by which audiences 'used' the media they consumed, in an instrumental fashion, to gratify a variety of needs researchers identified to be present in audience members. This shift in emphasis from a discussion of 'effects' to 'uses' highlighted and elaborated the nature of active audience involvement in media reception. However, the emphasis upon an active audience was limited by its rather instrumental concept of audience members' orientation toward media, as one of purely satisfying particular 'needs' that individuals had (for instance, a need for in-

formation, for escape, for status enhancement), the nature of which remained merely taken for granted and unquestioned by researchers. It remained for scholars in the critical tradition to push the active audience notion further, while questioning the 'uses'–'needs' paradigm of this perspective.

2. *Emphases in Current Theory and Research*

While earlier studies tended to build audience theory as though the experiences of all types of media audiences could be analyzed and understood with one set of concepts, audience theory since the 1970s has tended to be more segmented. Partly this is a result of scholars questioning the uniform, aggregate vision of the audience which emerges from broad theories of media 'effects' or the lack thereof. Current approaches to media audience research tend to draw in part from the uses and gratifications tradition to emphasize the active engagement of the audience with the mass media, rather than focusing on the passivity of audiences in the face of media exposure. Yet active audience research tends to eschew the cognitive dimension of this research, instead combining the focus on activity with a critical perspective on media and their uses often deriving from the Marxist tradition and from theories of democracy, which both criticize and herald the role of media in a democratic society. Particularly important for current audience research has been the cultural studies tradition, both the British (see *British Cultural Studies*) and the US versions.

British cultural studies has been associated with the work of those at the Centre for Contemporary Cultural Studies in Birmingham. Stuart Hall, David Morley, Angela McRobbie, Paul Willis, Dick Hebdige, and Roger Silverstone have all done work that looks at media audiences not as isolated phenomena, but as individuals and groups of individuals who must be studied in the context of the rest of their lives, and whose nature as a part of the media audience is only one segment of an overall set of cultural practices that characterize their identities.

Also emphasized in cultural audience work has been the multilayered nature of texts as well as the complexity of how they are received. Cultural audience researchers have drawn from semiotic and hermeneutic approaches to text analysis to make clear the multitextured nature of media 'messages.' Formerly treated as fixed and transparent, media messages themselves are now understood as complex and multitextured.

Perhaps the first work to bring together textual and audience analysis from these new, critical perspectives was Stuart Hall's (1973) essay about television reception, entitled 'Encoding/Decoding in the Media Discourse.' In this work, Hall theorizes both the complex nature of the meanings 'encoded' primarily in the television text, and the necessarily separate, but equally complex nature of the process by which viewers decode these messages. This seminal article paved the way for the more empirical audience studies which were produced in its wake by cultural audience researchers. Audience research carried out in this tradition emphasizes overall the ways in which the media are part of culture, and concomitantly, the way in which audience reception of media is only one facet of the way an entire cultural system influences those who live within it.

David Morley's work has been particularly influential in current audience study. He has authored two famous studies, *The Nationwide Audience* and *Family Television* (Morley 1980, 1986). In *Nationwide Audience*, he looked at the way people of different occupations, social class statuses, and ethnicities interpreted a television news program differently. Even a supposedly 'objective' news show, he found, was open to different interpretations, made by different types of audiences. In *Family Television*, Morley took the family locale of the television audience seriously. He and his research team went into the homes of families and both interviewed them about television, and observed them watching it. Through this research Morley was able to talk about the gender dynamics of the family audience for television. He observed how husbands rather than wives often commandeered the remote control, enabling them to make many of the family viewing decisions.

Morley's work, and the work of others in cultural studies, helped to shift research from a scientific paradigm which attempted to measure audience exposure and effects, to a more holistic one which looks at audiences in the context of their everyday lives. Rather than searching for more narrow, measurable influences of media, cultural researchers examine the narratives embedded within media, and look at how these are interpreted and indeed adopted by audience members in the course of many activities in their lives. Audience research is less focused only on 'audiences' *per se*, and has expanded to include the many uses people make of culture and media in postmodern society.

Many studies now emphasize active audience interpretation in a broad cultural context. Among the most important of these are Liebes and Katz' (1990) work, *The Export of Meaning*. In their research, Liebes and Katz looked cross-culturally at the ways audiences in different countries, and members of different ethnic and religious groups within each country, interpreted the same episodes of the prime-time soap opera *Dallas*, which had become a global phenomenon, being broadcast worldwide and achieving an avid following in many different national and cultural contexts. They found striking ethnic and religious group differences in how these audiences read the very same episodes of this popular television show. For example, Israeli Arabs and Russian immigrants were defensive about the US way of life pictured in the show, and attempted

to shield their children from it, while others in cultures closer to that pictured in the show read it more as nonthreatening, simple entertainment.

Other cultural studies work often chooses locales rather than particular media as the frame for audience study. Willis (1978), for example, studied bars and looked at the types of audiences for different sorts of music assembled there. More contemporary work focuses on theme parks like Disneyland, and charts the activities of audiences there, or focuses on formal as well as informal types of museums or other audience gatherings.

Simultaneous to the cultural studies productions of Morley, Willis, and other members of the Birmingham School, scholars primarily located in literature, humanities, and film departments have produced an enormous body of work theorizing the film audience. This work has drawn on different theoretical traditions from most of the television-focused work discussed so far, which was produced primarily in social science and communication departments. Drawing on psychoanalytic theory, film scholars have theorized the way filmic texts structure and position what they call the 'spectator' of film. They theorize what is in many ways an asocial type of audience for this medium, best illustrated by early scholars' discussions of how the situation of the film audience closely approximates the dream state, a resemblance underscored by the darkened movie theater, which makes acknowledgment of audience members, and discussion with them, difficult. Much of the early film theorizing was published in the British journal *Screen* (originally named *Screen Education*). Screen theory elaborated the contours of what was termed the 'filmic spectator.' This universalized 'spectator' was constructed with the aid of Freudian concepts of the family, which were applied to the way Hollywood cinema inserts—or 'sutures'—the viewer as an individual into the family setting and other family-like relationships set up by films. Such theory was based largely on psychoanalytically informed text analyses, rather than the study of actual viewers.

While retaining some of the insights of each of these traditions, current audience work has introduced new emphases: the increasing globalization of the media (Ang 1991); the historical specificity of the audience; the segmentation and diversity of the media audience; and recent and historical changes in media such as television, film, radio, newspapers, and magazines since the early 1990s, and the implications this has had for the media audience.

For example, Levine (1988) looks historically at the development of the distinction between 'high' and 'low' culture by studying theater audiences in the eighteenth and nineteenth centuries, emphasizing their often lower-class composition and their appreciation of authors like Shakespeare, whom modern audiences consider only for intellectuals interested in high culture. Film historians like Hansen (1991) continue this tradition by studying the informal ethnic and class organization among film audiences in the early twentieth century. She shows how film was received in a social and communal way by women in New York's immigrant communities in the early 1900s, and highlights the role film played in the development of ethnic identities. Scholars like Dayan and Katz (1992) analyze the audience for particular national media events, like the coronation of Queen Elizabeth in the UK, or Rabin's assassination in Israel, and find a political meaning and potential for mobilization in this type of event-oriented audience which may not exist in more day-to-day media audiences. Their study raises questions about many conventional approaches to audience study, which often perhaps falsely individualize reception and ignore its social functions and political impact.

2.1 Feminist Audience Study

Feminist study of the audience has grown auspiciously over the past several decades. It now constitutes a sizeable body of work with its own history, subfields, criticism, and problems. In film, Mulvey's (1975) seminal article put forth the idea that the female spectator must assume what is essentially a male identity in order to derive pleasure from classical Hollywood cinema. Others responded to this claim with extensive discussions of the melodrama and other allegedly 'female' genres of the classical cinema, asserting that these genres opened a space for an essentially female spectatorial identity. Perhaps, too, because gender was already so central a concept for psychoanalytic theory generally, the entire field of spectator theory was reframed by these feminist debates, which now often revolve around gender-related issues.

Many feminist audience researchers have employed what is often called 'feminist methodology' to study the media audience. Generally, feminist methodology employs interpretive, intersubjective techniques to probe the creative responses to oppression that women and members of other oppressed groups make. Topically, feminist audience researchers seek to look at the particular uses women and others make of our culture in response to their oppression.

Janice Radway (1984), for example, in *Reading the Romance*, interviewed female members of a Midwestern community who were fans of romance novel reading. Feminist communication researchers of the time had deemed romance literature sexist, oppressive to women, trashy, and of little literary value. Radway illustrated how, contrary to these dicta, women fans found great value in this literature, and admired it for its portrayal of strong and independent heroines. In addition, many used their reading time aggressively, to claim time away from family duties and household chores. Her interpretive methodology, which consisted

of speaking to the women on their own terms, and in their own language and spaces, helped her to elicit these findings, which were surprising at the time. This book was innovative and influential in its focused, qualitative study of readers rather than texts, and its focus on a popular 'woman's' genre, the romance, which had never before been taken seriously enough to be the object of scholarly study.

Press (1991) continued in Radway's tradition with the study *Women Watching Television*, an interviewing study of working-class and middle-class women of different ages in which she discusses the role of television in their lives. In this book, Press documents the role television plays for women of different social classes and age groups in reinforcing the culture's hegemonic values. She found that television seemed to reinforce different values and beliefs for different groups of women. For middle-class women, television's hegemony operated around issues concerning gender, often reinforcing stereotypical notions of proper feminine and masculine behavior. For working-class women, television hegemony operated along a class dimension, reinforcing the normality of a middle-class or often an upper-middle-class way of life in the USA, and underscoring the notion that poor or working-class people are abnormal and somehow individually inadequate.

Television soap opera in particular, traditionally considered a 'woman's genre' (although this is changing, particularly with an increasingly flexible workforce), has inspired researchers worldwide to study its audience. Often these studies assess the impact of US television soaps, which have been exported to many cultures around the globe, and are heavily consumed in a variety of contexts. Gripsrud (1995) studied the notable success of the prime-time soap opera *Dynasty* in Norway, and documented the enormous public debate over its broadcast in a television system that was tightly state-controlled and concerned that US shows would erode traditional culture. In the end, the show's widespread popularity, and public expression of this accolade in the mass media, swayed officials to permit its continued broadcast. Liebes and Katz (1990) document the failure of another prime-time soap opera, *Dallas*, phenomenally popular worldwide, to capture the popular audience in Japan. The unconventional family and gender relationships on the show were incomprehensible to the Japanese audience, they conclude. Other scholars have documented the continuing widespread popularity of telenovelas on Brazilian television (see *Soap Opera/Telenovela*).

3. *Probable Future Directions of Theory and Research*

Possible future trends of audience research will undoubtedly continue current insights into the nature of the media audience as increasingly fragmented into specific cultural 'segments' separated by race, nationality, ethnicity, sexuality, religion, and 'lifestyle.' Audience researchers are being challenged to refine their notion of the 'audience' to accommodate increasing recognition on the part of media, new media, and advertisers of its specialized and segmented nature. For example, the decline of viewership for nightly television news and of readership for the major national newspapers, and the concomitant rise of unconventional news sources such as MTV news and on-line news forms, have led to a level of fragmentation and specialization among news audiences inconceivable even at the start of the 1990s. These developments have spawned new types of local audience research which focus on the processes of reception and appropriation of media specific to particular groups of people, in particular locations, at particular moments in time. New media such as computers and the Internet pose particular challenges in this regard for current audience researchers. Often this research looks at the social, interactive, and consequently the often political media audience in a way ignored by earlier researchers, as in the work of Dayan and Katz (1992) discussed above.

In particular, the proliferation of new media, and the concomitant changing face of the old media since the 1980s have pushed audience research in new and ever-changing directions. Many researchers are interested in studying the impact of electronic media, both on children and adults. Some are discussing the way in which virtual communities have changed identities. Turkle (1995), for example, discusses the intriguing issue of how writing on computers, and communicating with others through computers, changes the sense of self (see *Virtual Reality, Psychology of*). She argues that a new sense of identity, decentered and multiple, is emerging as more and more people spend more and more of their daily lives communicating by computer. In a sense, Turkle asserts, computers bring the concepts of postmodernism 'down to earth,' concretizing the decentered sense of identity that scholars have been theorizing since the 1980s. Others discuss how digital technologies have transformed the very notion of interaction and dialog, or the particular implications of these changes for women and members of various minority groups in contemporary societies.

Certainly such observations transform the concept of 'audience.' As electronic media have become more participatory and dialogical, our notion of audience has transformed to become a more participatory concepts itself. Similarly, the notion of community that undergirded the conventional concept of audience has also been radically challenged. Virtual communities have some of the characteristics of communities in the traditional sense, but not others. The global nature of new communication technologies is fundamentally transforming the sense of space, and of what collectivity means. As notions of community develop along-

side rapidly changing electronic media, so will new dimensions of the term audience continue to unfold.

See also: British Cultural Studies; Broadcasting: Regulation; Mass Media: Introduction and Schools of Thought; Media Effects; Media Effects on Children; Media, Uses of; Soap Opera/Telenovela; Television: History

Bibliography

Ang I 1991 *Desperately Seeking the Audience*. Routledge, London

Dayan D, Katz E 1992 *Media Events: The Live Broadcasting of History*. Harvard University Press, Cambridge, MA

Gitlin T 1978 Media sociology: The dominant paradigm. *Theory and Society* **6**(1): 205–54

Gripsrud J 1995 *The DYNASTY Years: Hollywood Television and Critical Media Studies*. Routledge, London

Hall S 1973 Encoding/decoding in the media discourse, Stencilled Paper 7. Centre for Contemporary Cultural Studies, Birmingham, UK

Hansen M 1991 *Babel and Babylon: Spectatorship in American Silent Film*. Harvard University Press, Cambridge, MA

Katz E, Lazarsfeld P 1955 *Personal Influence: The Part Played by People in the Flow of Mass Communication*. The Free Press, Glencoe, IL

Levine L 1988 *Highbrow/Lowbrow: The Emergence of Cultural Hierarchy in America*. Harvard University Press, Cambridge, MA

Liebes T, Katz E 1990 *The Export of Meaning: Cross-Cultural Readings of DALLAS*. Oxford University Press, New York

Morley D 1980 *The Nationwide Audience: Structure and Decoding*. British Film Institute, London

Morley D 1986 *Family Television: Cultural Power and Domestic Leisure*. Comedia, London

Mulvey L 1975 Visual pleasure and narrative cinema. *Screen* **16**(3): 6–18

Press A 1991 *Women Watching Television: Gender, Class and Generation in the American Television Experience*. University of Pennsylvania Press, Philadelphia, PA

Radway J 1984 *Reading the Romance*. University of North Carolina Press, Chapel Hill, NC

Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon and Schuster, New York

Willis P 1978 *Profane Culture*. Routledge and Kegan Paul, London

A. L. Press

Audition and Hearing

Hearing, like most senses, allows an organism to determine objects in its environment. Objects can vibrate, producing sound that provides information about the object. However, sounds from the many sound sources in one's environment are combined into one complex sound field, and therefore sounds from individual sources do not reach the listener as separate events. This complex sound field is the frequency components of the sound's spectrum, the level of the spectral components, and the timing relationships among the components. The auditory periphery codes for the frequency, level, and timing parameters of the components of the entire complex sound field and not for sound sources *per se*. The spectral–temporal code generated by the auditory periphery contains information that is processed by brainstem and cortical auditory centers to help determine the originating sound sources (Yost et al. 1993).

1. Outer and Middle Ears

The head and torso transform sound spectrally as it travels from its source to the outer ear canal. Resonances of the outer ear canal provide further modifications (increases in pressure in the 2,000–6,000 Hz frequency region) before sound vibrates the tympanic membrane and ossicular bones (see *Auditory System* for a description of auditory anatomy). The vibrations of the ossicular chain cause the fluids and membranes of the inner ear to move, which begins the biomechanical and neural transduction of vibration into neural discharges. The fluids and membranes of the inner ear offer a significant impedance to the vibratory motion established by the incoming pressure wave. The resonance of the outer ear canal and, most important, the lever action and pressure transformations provided by the ossicular chain, compensate for this impedance mismatch, providing an efficient transfer of pressure from air to the fluids of the inner ear (Dallos 1973).

2. The Inner Ear

The coiled cochlea constitutes the auditory part of the inner ear with the cochlear partition, lying within the middle segment (scala media) of the three-part cochlea, housing the major structures (e.g., inner and outer haircells) responsible for the biomechanical and neural sound transduction. As shown in Fig. 1, the stapes vibrates the fluids of the cochlea in a piston-like manner (Dallos et al. 1996). The resulting pressure wave develops a pressure gradient across the cochlear partition, setting it in motion such that it moves from the base of the cochlea toward its apex. The stiffness gradient of the cochlear partition; the widening of the cochlear partition toward the apex; and the opening (the habenula perforata) between the cochlear partition and the cochlear end lead to a traveling motion of the cochlear partition. For high-frequency sounds, the traveling wave produces maximal displacement at the base and the wave does not propagate to the apex. For low-frequency sounds, the traveling wave propagates toward the apex where maximal displacement occurs. Cochlear partition vibration follows the vibratory pattern of the incoming sound, but maximal

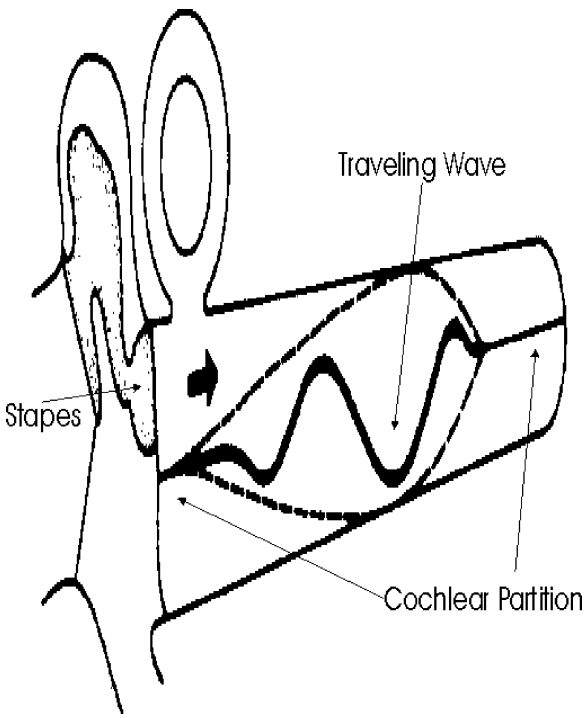

Figure 1
A schematic diagram of the uncoiled cochlea and the middle ear ossicles with the moving stapes initiating a pressure gradient across the cochlear partition which vibrates with a traveling wave motion from base (near the stapes) to apex. The envelope of the traveling wave (the dotted curve) shows that the point of maximal cochlear partition displacement is frequency dependent (reproduced by permission from Yost, 2000)

displacement varies from base to apex depending on the frequency of the originating sound: maximal displacement toward the base for high frequencies and toward the apex for low frequencies. Cochlear partition motion is in phase with stapes movement near the cochlear base but, due largely to traveling wave travel times, there is a phase lag between cochlear partition and stapes vibrations toward the apex.

The vibratory pattern of cochlear partition displacement can be estimated using laser interferometry or the Mossbauer technique to measure basilar membrane displacement (the basilar membrane forms one boundary of the cochlear partition). Figure 2 shows an iso-displacement plot for one location along the basilar membrane for tones of different frequencies, indicating the sound level required to displace the basilar membrane by a threshold amount. For this location near the base of the cochlea, sounds with frequency components near 8,000 Hz require the lowest level for threshold displacement (8.3 kHz is the best frequency (BF) for this basilar membrane location).

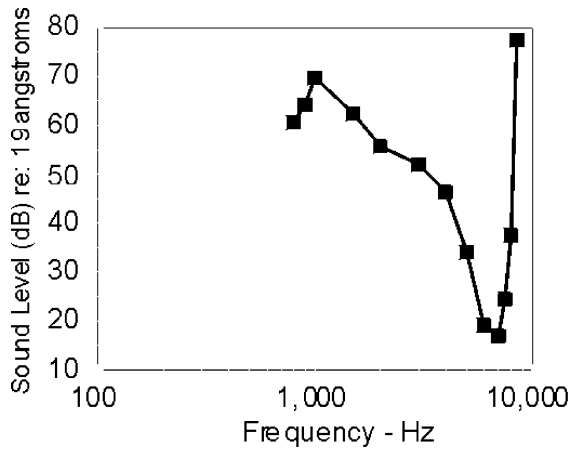

Figure 2
The level of tones of different frequencies required to displace the basilar membrane 19 angstroms using the Mossbauer technique. The BF for this function is 8.3 kHz (reproduced by permission of from Yost (2000), adapted from Ruggero and Rich, 1986)

Each point along the basilar membrane produces a function similar to that shown in Fig. 2, but with a different BF for each point: maximal displacement at the base for high frequencies, and toward the apex for low frequencies. The amount of basilar membrane displacement is proportional to stimulus level, but for tones with frequencies near BF displacement, it is a nonlinear compressive function of sound level. At all points along the membrane, the temporal pattern of displacement follows the sound's pressure waveform, except that displacement magnitude depends on interactions between basilar membrane location and frequency, and there is a time delay for displacement at apical locations.

Cochlear partition motion causes the stereocilla of the inner and outer haircells to bend (shear), triggering an ion exchange across the haircell membrane wall, most likely near the base of the stereocilla, that produces a neural generator potential propagating to auditory nerve fibers innervating the haircells (Dallos et al. 1996). Shearing of stereocillia is due to differential motion of the basilar membrane on which the haircells sit, and the tectorial membrane that lies at the top of the haircells' stereocilla. The generator potential creates neural discharges within the auditory nerve, which carries this neural information about the incoming sound to the cochlear nucleus, the first auditory brainstem nucleus.

Inner and outer haircells serve two different functions in the transduction process (Dallos et al. 1996). Inner haircells are the actual biological transducers that convert stereocillia shearing into neural generator potentials. Approximately 90 percent of the afferent auditory nerve fibers innervate inner haircells, with

only a few auditory nerve fibers innervating any one inner haircell. The outer haircells most likely modify the mechanical coupling within the cochlear partition that provides the high sensitivity and frequency selectivity of the biomechanical traveling wave. The outer haircells are motile in that they change length in response to electrical stimulation in vitro, and most likely also change length in response to cochlear partition vibrations in vivo (Dallos et al. 1996). Since the outer haircells are connected to basilar and tectorial membranes, the change in outer haircell confirmation can alter the coupling between these two membranes that is crucial for stereocilla shearing. Outer haircell motility is also most likely under some efferent control from the olivocochlea bundle coming from the cochlear nucleus. Direct outer haircell motility operates on a cycle by cycle basis as the cochlear partition vibrates (as fast as fractions of a millisecond), whereas efferent control of outer haircell function operates on a longer timescale (many milliseconds). Outer haircell changes are consistent with a system that feeds energy back into the cochlea as it is stimulated. This feedback system is referred to as the cochlear amplifier, in that the function of the outer haircells may be to 'amplify' the sensitivity of the biomechanical system. When outer haircells are damaged and inner haircells remain functional, the sensitivity of the inner ear is significantly reduced, frequency selectivity is poor, and the system becomes nearly linear—no longer displaying a compressive nonlinearity. These alterations in cochlear function demonstrate the significant role of outer hair cells in auditory processing.

Hearing loss due to overexposure to sound, exposure to some ototoxic drugs, and aging results from deterioration of haircells, both inner and outer. The haircells are either destroyed or their ciliary structures become malformed. In mammals, haircells do not regenerate once they are destroyed (Rubel et al. 1999). However, haircells do appear to regenerate in many birds and fish. In birds, there is strong anatomical evidence for haircell regeneration, and mounting evidence that the physiological properties of regenerated haircells are the same as for normal haircells. In addition, it does not appear that birds suffer permanent hearing loss following intense sound exposure as mammals do. It is not clear what causes haircells to regenerate in birds and fish and not in mammals. Several hypotheses concerning genetic triggers are currently being investigated.

3. *Auditory Nerve*

Activity within the inner haircells triggers neural responses in auditory nerve fibers lying in the VIIIth cranial nerve bundle. Figure 3 displays tuning curves for different auditory nerve fibers in the VIIIth nerve. Each tuning curve displays the sound level required to increase neural firing rate a threshold amount above spontaneous rate as a function of frequency. Each auditory nerve fiber has a different BF; fibers with high-frequency BFs innervate haircells near the base of the cochlear partition, and fibers with low-frequency BFs innervate apical locations (note the similarity between Fig. 1 displaying tuning at the level of basilar membrane displacement and Fig. 2 displaying neural tuning). Due to the refractory properties of neuronal firing, auditory nerve fibers discharge in synchrony with the temporal pattern of the sound waveform up to approximately 5 kHz. Any one auditory neuron responds over a 30–50 decibel (dB) range of sound level. Thus, the auditory nerve carries a code for sound to the auditory brainstem. Frequency is coded primarily by the differential sensitivity of each nerve to frequency (tonotopic organization); sound level is coded by overall neural discharge rate; and timing information is coded by the temporal pattern of neural discharges (Geisler 1998). This spectral–temporal code has limitations in its ability to represent faithfully the originating sound. While each auditory nerve fiber is most sensitive to one frequency, it does respond to a range of frequencies. Any one auditory neuron only responds with a change in discharge rate to a stimulus change of 30 to 50 dB, which is a small part of the 120 dB dynamic range of hearing. Each nerve can only code for temporal structure up to approximately a rate of 5 kHz (or a timing difference of 0.2 ms), whereas the auditory system is sensitive to frequencies as high as 20,000 Hz. However, this peripheral code accounts for much of auditory perception. And the peripheral code is refined in many ways as neural information ascends through the auditory brainstem toward the cortex.

4. *Auditory Brainstem and Cortex*

The auditory pathway from the auditory nerve to the auditory cortex is a complex system of afferent and efferent subpathways (see *Auditory System*). The varying morphological cell types within the various brainstem nuclei and auditory cortex appear to have corresponding physiological function (Fay and Popper 1992). For instance, fusiform cells in the dorsal cochlear nucleus produce neural firing patterns with an excitatory tuning curve as well as inhibitory areas. These inhibitory areas reveal cases in which tones of one frequency inhibit the firing rate of the cell caused by tones of a different frequency (two-tone inhibition; Geisler 1998). Such two-tone inhibition exists at the level of the auditory nerve and throughout the auditory brainstem. Some of these inhibitory cells may be part of lateral inhibitory networks that could sharpen the neural response to sounds with complex spectral patterns (Shamma 1985). The behavioral significance of the function of central neural circuits in several species of bats and song birds is better understood than those of most other animals (see *Birdsong and Vocal Learning during Development*).

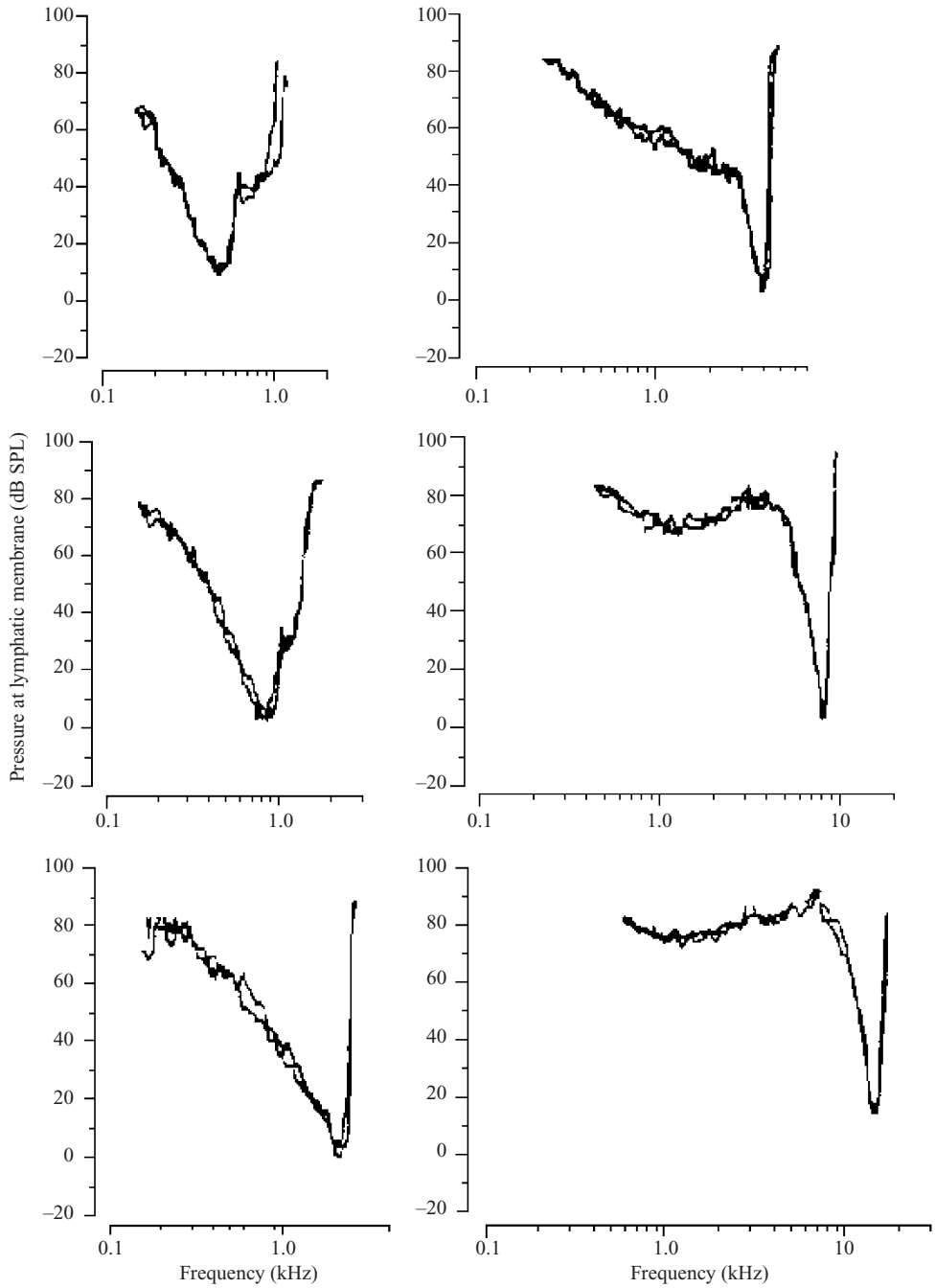

Figure 3
Neural tuning curves from six auditory nerve fibers showing the level of tones of different frequencies required to increase the firing rate a threshold amount above spontaneous activity (reproduced by permission from Yost (2000), adapted from Liberman and Kiang, 1978)

The peripheral neural code contains information about the individual sound sources of a complex sound field that can be processed by higher auditory centers. One form of processing is binaural processing, which makes possible sound source location in the horizontal or azimuth plane (Yost and Gourevitch 1987). Sound from a horizontally located source reaches the ear closest to the source sooner and with a greater level than that reaching the other ear. Differences in interaural (between the ears) time and level are cues used for azimuthal sound localization. There are only a few bilateral neural connections between the two ears within the cochlear nucleus, but there are strong bilateral connections at the next brainstem nucleus, the olivary complex. Neural units within the medial superior olive (MSO) respond to interaural time differences, while the lateral superior olive (LSO) processes interaural level differences. MSO and LSO outputs are sent to the inferior colliculus and then to the auditory cortex, where additional binaural processing takes place. While there is some evidence for auditory neural spatial maps (Yin et al. 1997), determining azimuthal spatial location depends on neural computations of interaural time and level differences.

Animal models such as the barn owl, which may be adapted specially to use sound to locate prey, have provided valuable information about binaural hearing. In several animal species, interaural time differences are likely processed by a coincidence network of bilateral cells (Yin et al. 1997). In the barn owl these cells are located in the nucleus laminaris, which is a homolog to the MSO. Such coincidence networks perform a form of cross-correlation on the signals reaching the two ears, providing a spatial map for processing interaural time as a cue for azimuthal sound localization. Bats (Popper and Fay 1995) and dolphins (Au 1993), which use echolocation for navigation and feeding, are also valuable animal models for studying sound localization.

Location in the vertical plane most likely depends on spectral information contained in the Head Related Transfer Function (HRTF), which describes the spectral changes sound undergoes as it passes over the torso and head on its way from the source to the outer ear canal (Wightman and Kistler 1993). Parts of the torso and head (e.g., pinna) alter the spectral information reaching the outer ear canals, especially in the high frequencies where the spectral locus of deep spectral valleys in the amplitude spectrum of the HRTF covaries with vertical stimulus location. Thus, the spectral location of these valleys is the likely cue for vertical sound localization. The complex neural circuitry of the cochlear nucleus (e.g., the dorsal cochlear nucleus) is a candidate for initially processing complex spectral structure, such as that associated with the HRTF (Young et al. 1997).

Segregating one sound source from others in a complex sound field may be aided by the fact that an independent, slow temporal amplitude modulation of the sound can be imparted by each source (Yost et al. 1993). Neurons at several brainstem levels (cochlear nucleus and inferior colliculus) and in the auditory cortex are differentially sensitive to sounds with different rates of amplitude modulation. These neural units display a modulation transfer function suggesting that they may be tuned to a particular rate or pattern of amplitude modulation. There is evidence that inferior colliculus units in the central nucleus are organized according to their preferred amplitude modulation rate (Langner et al. 1998). Thus, these circuits may subserve modulation processing as one means of segregating sound sources in complex, multisource environments.

The periodic envelope of amplitude-modulated sounds often provides a complex periodicity pitch cue that can be used for sound source determination and segregation. The same central auditory system circuits described above that process amplitude modulation have also been suggested as circuits for complex periodicity pitch processing (Langner et al. 1998). Other sounds have a similar complex pitch. Some of these sounds do not have the periodic envelopes of amplitude modulated sounds, but they do have a temporal regularity in their waveform fine structure (Yost et al. 1996). One such sound, iterated rippled noise, is formed by a sound and its added delayed reflections. The amplitude envelope of such sounds is flat, but there is a temporal regularity in the fine structure equal to the delay, the reciprocal of which is the perceived pitch. Studies of the anteroventral cochlear nucleus have shown that the neuronal discharge pattern of these units preserves this temporal regularity (Shofner 1991). Autocorrelation of the neural spike trains reveals a strong temporal regularity associated with iterated ripple noise stimuli that might be the basis for the perceived complex pitch. While autocorrelation of neural spike trains has been a useful analysis tool (Cariani and Delgutte 1996), there is little evidence for neural coincidence networks that might form an autocorrelator. Auditory science is learning more about the functional properties of the auditory brainstem nuclei and cortex as a means of understanding the neural basis for sound source determination and segregation.

See also: Auditory Models; Auditory Scene Analysis: Computational Models; Auditory System; Neural Plasticity in Auditory Cortex

Bibliography

Au W L 1993 *The Sonar of Dolphins*. Springer, New York

Cariani P A, Delgutte B 1996 Neural correlates of the pitch of complex tones I. Pitch and pitch salience. *Journal of Neurophysiology* **76**: 1698–716

Dallos P 1973 *The Auditory Periphery: Biophysics and Physiology*. Academic Press, New York

Dallos P, Popper A N, Fay R R (eds.) 1996 *Springer Handbook of Auditory Research, Vol. 8. The Cochlea*. Springer, New York
Fay R R, Popper A N (eds.) 1992 *Springer Handbook of Auditory Research, Vol. 2. The Auditory Pathway: Neurophysiology*. Springer, New York
Geisler C D 1998 *From Sound to Synapse*. Oxford University Press, New York
Langner C L, Schreiner C E, Biebel U W 1998 Function implications of frequency and periodicity pitch in the auditory system. In: Palmer A R, Rees A, Summerfield A Q, Meddis R (eds.) *Psychophysical and Physiological Advances in Hearing*. Whurr, London
Liberman M C, Kiang N Y S 1978 Acoustic trauma in cats. *Acta Otolaryngology*. Supplement, **358**: 1–63
Popper A N, Fay R R (eds.) 1995 *Springer Handbook of Auditory Research, Vol. 5. Hearing by Bats*. Springer, New York
Rubel E, Popper A N, Fay R R (eds.) 1999 *Springer Handbook of Auditory Research, Vol. 13. Plasticity of the Auditory System*. Springer, New York
Ruggero M A, Robles L R, Rich N C, Reico A 1992 Basilar membrane response to two-tone and broadband stimuli. *Philosophical Transactions of the Royal Society of London* **336**: 307–15
Shamma S A 1985 Speech processing in the auditory system II: Lateral inhibition and the central processing of evoked activity in the auditory nerve. *Journal of the Acoustical Society of America* **78**: 1622–32
Shofner W P 1991 Temporal representation of rippled noise in the anteroventral cochlear nucleus of the chinchilla. *Journal of the Acoustical Society of America* **90**: 2450–66
Wightman F L, Kistler D J 1993 Sound localization. In: Yost W A, Popper A N, Fay R R (eds.) *Springer Handbook of Auditory Research, Vol. 3. Psychoacoustics*. Springer, New York
Yin T C T, Joris P X, Smith P H, Chan J C K 1997 Neuronal processing for coding interaural time disparities. In: Gilkey R, Anderson T (eds.) *Binaural and Spatial Hearing in Real and Virtual Environments*. Erlbaum Press, New Jersey
Yost W A 2000 *Fundamentals of Hearing: An Introduction*, 4th edn. Academic Press, New York
Yost W A, Gourevitch G (eds.) 1987 *Directional Hearing*. Springer-Verlag, New York
Yost W A, Patterson R D, Sheft S 1996 A time domain description for the pitch strength of Iterated Ripple Noise. *Journal of the Acoustical Society of America* **99**: 1066–78
Yost W A, Popper A N, Fay R R (eds.) 1993 *Springer Handbook of Auditory Research, Vol. 3. Psychoacoustics*. Springer, New York
Young E D, Rice J J, Spirou, G A, Nelken I, Conley R 1997 Head related transfer functions in cat: Neural representation and the effects of pinna movement. In: Gilkey R, Anderson T (eds.) *Binaural and Spatial Hearing in Real and Virtual Environments*. Erlbaum Press, New Jersey

W. A. Yost

Auditory Models

For more than one hundred years models of hearing have played a significant role in the development of auditory theories. Modeling efforts in hearing can be divided into three categories: models of the inner ear (cochlear models), models of auditory detection and discrimination, and auditory computational models.

In the last part of the nineteenth century several 'theories of hearing' were proposed (Boring 1942). In reality these were theories of frequency coding by the cochlea. The bases for these theories were actual physical models, descriptions of such models, or mathematical realizations of the biomechanical vibratory patterns that might exist within the cochlea and how those patterns code for the frequency content (the spectrum) of sound. Cochlear modeling is still an extremely active and important effort. Following World War II a group of psychologists and engineers at the University of Michigan developed Signal Detection Theory (SDT, e.g., see *Signal Detection Theory, History of*) and adapted it as a model of hearing for predicting the detection and discrimination of tonal signals in noisy backgrounds. Models based on SDT still provide a fundamental base for many models of auditory perception. With the advent of fast computers, computational models of hearing have been developed for both neural and perceptual data including those describing auditory neurons and neural circuits, spatial localization, pitch perception, speech perception, amplitude modulation processing, and, recently, auditory scene analysis (e.g., see *Auditory Scene Analysis: Computational Models*).

1. Cochlear Modeling

Sound enters the ear canal and vibrates the eardrum (tympanic membrane) and the middle ear bones (ossicular chain). The stapes, the ossicular bone nearest the inner ear, transmits this vibration to the fluids of the cochlea (part of the inner ear). The moving fluids vibrate a set of structures within the cochlea called the cochlear partition, such that cochlear partition vibrations are sorted into a frequency-selective pattern. The inner haircells lying within the cochlear partition, which serve as the biological transducers, produce neural activity in proportion to cochlear partition vibratory displacement. The auditory nerve relays the neural activity to the central nervous system such that each auditory nerve fiber has a preferred (or best) frequency to which it discharges due to the location of the inner haircells along the cochlear partition. Thus, frequency is place coded in that each nerve fiber carries information about a particular frequency region of the auditory spectrum. This place code is the basis for all theories of hearing (see Yost 2000 for an introduction to hearing).

Due to the difficulty of measuring cochlear-partition vibrations, cochlear models have played a major role in understanding cochlear-partition vibrations. George von Bekesy, who won the Nobel prize in 1959, built real models made of wood, metal, rubber, and fluids scaled to the actual dimensions of the cochlea to understand the frequency selectivity of cochlear-

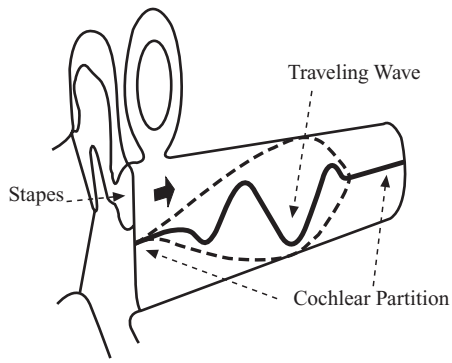

Figure 1
A schematic depiction of the uncoiled cochlea and the middle ear ossicles with the stapes shown. A predicted instantaneous traveling wave motion of the cochlear partition is shown indicating that the area of maximal displacement (dotted curve) is about two-thirds of the way toward the cochlear apex for this low-frequency stimulus

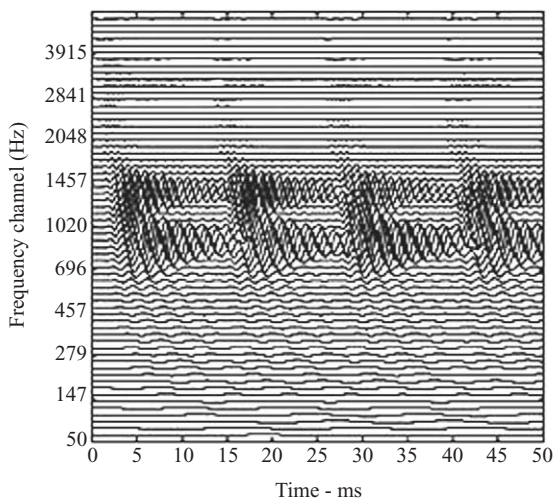

Figure 2
Simulation of the vibration pattern in the cochlea based on the output of a bank of 64 gammatone filters to the vowel /a/

partition wave motion (von Bekesy 1960). Modern models are based on hydrodynamic or equivalent electrical circuits (Hawkins et al. 1996). The key prediction of these models is that cochlear-partition wave motion produces a traveling wave that travels from the cochlear base toward its apical end. Figure 1 is a schematic depiction of the cochlea and an instantaneous traveling-wave. The place of the traveling wave's maximal displacement is frequency dependent, with high frequencies generating a traveling wave that travels partway toward the cochlear apex and having its maximal place of displacement near the cochlear base. Low frequencies generate a traveling wave that travels to the apex, with its maximal place of displacement near the apex. These wave-propagation models are based on longwave and shortwave hydromechanical assumptions (Dallos et al. 1996). They do a good job of describing classic measures of cochlear vibration and indirect measures of these vibrations, and thus they account for frequency coding by the auditory periphery.

The classic measures of cochlear vibrations were based on cochleas that were severely compromised by the surgery required to make the measurements, and, thus, do not represent the true biological state of the cochlea. The earlier wave-propagation models cannot account for the high sensitivity and acute frequency selectivity measured for biologically vital cochleas. In addition, the healthy cochlea produces vibration patterns that are nonlinear and the earlier models are linear. The role of the outer haircells must be included in an improved cochlear model. The function of outer haircells is minimized in damaged cochleas. It is now known that outer haircells are motile, acting like motors that probably modify the cochlear-partition vibrations. Outer haircells may feed energy back into the cochlear partition vibration process providing an active mechanism that increases the partition's vibratory sensitivity and improves its frequency selectivity. Such feedback mechanisms are nonlinear. The difficulty of measuring outer haircell function, especially in vivo, has made cochlear models critically important (Dallos et al. 1996). The active process is often modeled as a form of 'negative damping' that counteracts the inherent damping (friction) provided by the cochlear partition membranes and fluids (Dallos et al. 1996). Modern laser-interferometry techniques allow accurate measures of cochlear-partition vibrations, and several current models can account for the sensitivity, frequency selectivity, and nonlinear properties of these vibrations. However, knowledge about the details of cochlear biomechanics is changing rapidly, and the modeling efforts are, therefore, dynamic ones.

As described above the biomechanics of the cochlea generate a spectral-temporal pattern of neural discharges within the auditory nerve bundle providing the neural code for the stimulus' spectrum. This spectral-temporal pattern can also be modeled as if the stimulus were filtered by a bank of narrow band filters. A filterbank can also account for the perception of the interactions that takes place when sounds of different frequencies coexist in an acoustic environment. The sound of one frequency (masking sound) can interfere with the detection of a sound of a different frequency (signal frequency) only if the masker frequency is in the spectral region of the signal frequency. This result can be explained by assuming that the masker is filtered by a bandpass filter whose center frequency is

that of the signal, and the total power of the masker at the output of this hypothetical filter determines signal thresholds (Moore 1986).

Thus, it is not surprising that several models of these 'internal filters' have been proposed to model physiological and psychoacoustical data. One such filter is based on a gammatone impulse function (an impulse function with a gamma-function envelope of a cosinusoidal fine structure) which is derived from direct neural measurements of auditory nerve tuning (Patterson et al. 1995). Figure 2 depicts the filtered output of the vowel /a/ provided by a gammatone filterbank. Each line is the filtered output of a gammatone filter with center frequencies ranging from low frequencies at the bottom to high frequencies at the top. By looking across frequency at any one point in time, an estimate of an instantaneous measure of the traveling wave can be established. The display also reveals the spectral-temporal code provided by the cochlea. The time delay seen in low-frequency channels result from the impulse time of low-frequency filters and corresponds to the traveling wave delay toward the cochlear apex (where low frequencies are coded). The gammatone filter bank has provided a good fit to many psychoacoustical and auditory nerve data.

2. *Models of Auditory Detection and Discrimination*

Auditory models based on Signal Detection Theory have played a major role in psychoacoustically based theories of hearing (e.g., see *Signal Detection Theory*). The classic auditory stimulus condition for these models involves the detection of a tonal signal in a background of Gaussian noise. The energy-detection model, as developed by Dave Green and colleagues in the 1960s, was derived from SDT (Green and Swets 1974). It assumes that the listener uses the distribution of instantaneous amplitudes of the two sounds (signal-plus-noise and noise alone) for a decision as to whether or not a tone was added to the noise. The instantaneous amplitudes of Gaussian noise are normally distributed with zero mean and a standard deviation equal to the root-mean-square (rms) amplitude of the noise. When a sinusoidal tone is added to the noise, the distribution remains normal with a mean equal to the energy of the signal and the same standard deviation. The decision is assumed to be based on the likelihood that a sampled instantaneous amplitude would come from the signal-plus-noise distribution as opposed to the noise distribution. The distributions of the likelihood ratios remain normal and the distance (d') between the two distribution means normalized to the common standard deviation takes the form: $d' = \sqrt{(2E/N_o)}$, where E is signal energy and N_o is noise spectrum level. Thus, as the energy of the signal increases relative to the level of the noise, d' increases, suggesting higher signal detectability.

A d' measure of performance can be derived from a masking task. In the single-interval psychophysical task the signal-plus-noise is presented randomly on half of the trials and the noise on the other half. The listener indicates either 'Yes' the tone-plus-noise was presented or 'No' only the noise was presented. The hits (proportion of Yes responses given signal-plus-noise) and false alarms (proportion of No responses given signal-plus-noise) are converted to normal standard scores (z-scores) and $d' = z_{hit} - z_{fa}$, where z_{hit} is the hit z-score and z_{fa} is the false alarm z-score.

According to the model's assumptions the listener establishes a criterion likelihood ratio for making the Yes/No decision. If the obtained likelihood ratio is greater than the criterion value, the decision is 'Yes' the signal-plus-noise was presented, otherwise the decision is 'No.' Because of the normal form of the underlying distributions as explained above, z_{hit} and z_{fa} represent the appropriate response outcomes given the model's assumptions. Thus, the obtained d' values can be compared with the predicted values. The predictions are close to the obtained data, but miss in both slope and intercept. Better fits can be obtained, but the model adjustments usually miss fitting the proper slope (Green and Swets 1974). The general energy-detection model has been applied to a wide range of stimulus conditions (Swets 1964) such as different stimulus durations and different frequency content for the signal and masker, and to tasks involving discrimination stimulus changes (e.g., in signal level). The energy-detection models are most accurate for limited data sets.

The energy-detection model is used less today, but several aspects of the SDT approach remain as powerful components of many current auditory models. The energy model is often termed an 'ideal detector' in that the model's prediction represents the maximal use of all available information, and given this information no other decision process is more accurate in determining the presence or absence of the signal. Many current models use such an 'ideal detector' approach. The result of $d' = \sqrt{(2E/N_o)}$ can be equivalently derived by assuming that the observer cross correlates a sample of the instantaneous amplitude to a known template of the signal-plus-noise distribution. As such this template matching calculation is an ideal detector. Current auditory models often include such template matching operations (Dau et al. 1996). Many current models contain a decision-processing stage. The concepts of SDT, in which the decision process is divided into two classes of variables, those controlling the sensitivity of the observer to the stimulus conditions and those that control the observer's response bias, have been used extensively in many current auditory models. Other models have assumed other divisions of the variables that might

control a decision, such as the stimulus context in which a signal exists (Durlach and Braida 1969).

3. Auditory Computational Models

Most computational models of hearing have a modular approach, with each module representing a different stage of auditory processing. The 'front-end' to many models contain an outer and middle ear module, a cochlear module, and a haircell module (Patterson et al. 1995). The outer and middle ear module is usually based on a simple bandpass-filter model to represent the filtering of the middle ear and its relationship to the thresholds of hearing (Patterson et al. 1995). Recent work has shown that the spectral transformations that sounds undergo as they travel from the source across the body and head (especially the pinna) before reaching the outer ear are extremely important in providing sound localization cues, especially in the vertical plane. These head related transfer functions (HRTFs) are crucial characterizations of the acoustic input to the outer ear for which modeling efforts have been important for understanding sound localization and providing realistic virtual auditory environments (Gilkey and Anderson 1996).

The gammatone filter bank (see Fig. 2) is a common cochlear model. An often used neural transduction model is the Meddis haircell, which is a computational model of inner haircell function (Meddis 1986). The parameters governing neurotransmitter release in the Meddis haircell have been tailored to the unique physiology of auditory haircells. Figure 3 displays the output of the Meddis haircell for the same /a/ vowel used in Fig. 2. The display reveals the spectral-temporal code provided by the auditory periphery.

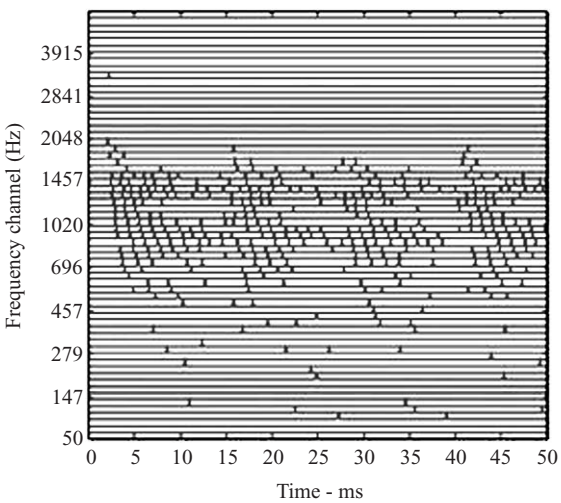

Figure 3
Simulation of the neural spectral-temporal pattern in the auditory nerve based on the neural output of the Meddis haircell model for the 64 tuned channels shown in Fig. 2

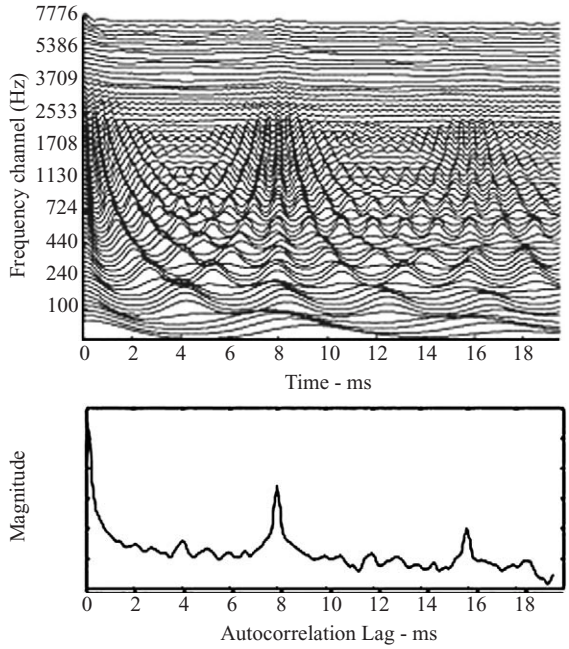

Figure 4
Top: The correlogram for 64 neural channels (as in Fig. 3) formed by autocorrelation. The stimulus was an iterated rippled noise that produces a 125-Hz pitch. Bottom: The summary correlogram formed by adding the correlogram across channels, showing the clear peak at 8 ms

In most computational models additional modules process the information provided by these 'front-end' modules. Often these higher modules are based on pattern recognizing models derived from neural network technology. These models attempt to find appropriate patterns of information in displays such as those shown in Figs. 2 and 3. Sometimes, especially for models of auditory scene analysis (e.g., see *Auditory Scene Analysis*), the neural network requires some prior knowledge about the acoustic or listening situation in order for accurate predictions to emerge. In other models a decision module is provided (Dau et al. 1996), and if so, they use models following those from SDT as described above.

Figure 4 displays the output of a model used to predict the pitch of complex sounds (Yost et al. 1996). The front end to this model is a middle-ear filter, the gammatone filterbank, and Meddis haircell. Following the earlier suggestions of Licklider (Licklider 1951), the final module is based on autocorrelation performed on the neural spike trains of each frequency channel (such as shown in Fig. 3). Such a display is

called a correlogram. At the bottom of the figure is a summary correlogram generated by summing across frequency channels for each autocorrelation lag. The stimulus for this display is a regular-interval noise stimulus generated by delaying a noise by 8 ms and adding it back to itself several times (iterated rippled noise, Yost et al. 1996). This stimulus has a flat amplitude envelope and a spectrum with noisy spectral peaks and valleys, and it produces a salient pitch of 125 Hz. The summary correlogram shows a clear peak at a lag of 8 ms suggesting an 8-ms temporal regularity. The model predicts that this iterated rippled noise stimulus should have a pitch of 125 Hz (the reciprocal of 8 ms), and that the pitch should be relatively strong given the prominence of the summary correlogram peak. Both predictions are in excellent agreement with psychoacoustical data.

A similar approach has been proposed to account for interaural time differences as a major cue for horizontal-plane sound localization. Jeffress (1948) proposed a model for processing interaural time differences in which the neural activity at the output of each cochlea was cross correlated. Several versions of such cross-correlation models have been successful in accounting for a wide range of binaural and spatial hearing data (Colburn and Durlach 1978, Hawkins et al. 1996). Such a cross-correlation process could be realized by a coincidence network of bilateral cells receiving inputs from each cochlea (Jeffress 1948).

The modeling efforts described in this chapter and many other models have provided valuable links between empirical research on hearing and the development of theories of auditory processing. Occasionally, as with cochlear models, the models provide a close approximation to nature. The recent advances in computational power and in our understanding of hearing suggest that many more models will be soon be available which will provide useful descriptions of the hearing process.

See also: Auditory Scene Analysis: Computational Models; Auditory System; Fechnerian Psychophysics; Neural Plasticity in Auditory Cortex; Psychophysics

Bibliography

Boring E G 1942 *Sensation and Perception in the History of Experiential Psychology*. Appleton Century, New York

Colburn H S, Durlach N I 1978 Models of binaural interaction. In: Carterette E C, Friedman M P (eds.) *Handbook of Perception* (*Hearing*). Academic Press, New York

Dallos P, Popper A N, Fay R R (eds.) 1996 *The Cochlea*. Springer, New York

Dau T, Kollmeier B E, Kohlrausch A 1996 A quantitative prediction of modulation masking with an optimal-detector model. *Journal of the Acoustical Society of America* **99**: 2565–73

Durlach N I, Braida L D 1969 Intensity perception I: Preliminary theory of intensity perception. *Journal of the Acoustical Society of America* **46**: 372–83

Gilkey R H, Anderson T (eds.) 1996 *Localization and Spatial Hearing in Real and Virtual Environments*. Erlbaum Press, Hillsdale, NJ

Green D M, Swets J A 1974 *Signal Detection Theory and Psychophysics*. Krieger, Huntington, New York

Hawkins H L, McMillian T A, Popper A N, Fay R N (eds.) 1996 *Auditory Computation*. Springer-Verlag, New York

Jeffress L A 1948 A place theory of sound localization. *J. Comparative Physiology and Psychology* **41**: 35–9

Licklider J C 1951 A duplex theory of pitch perception. *Experientia* **7**: 128–34

Meddis R 1986 Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America* **79**: 702–11

Moore B C 1986 *Frequency Selectivity in Hearing*. Academic Press, London

Patterson R D, Allerhand M, Giguere C 1995 Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *Journal of the Acoustical Society of America* **98**: 1890–4

Swets J (ed.) 1964 *Signal Detection and Recognition by Human Observers*. Wiley, New York

von Bekesy G 1960 *Experiments in Hearing*. McGraw-Hill, New York

Yost W A 2000 *Fundamentals of Hearing: An Introduction*, 4th edn. Academic Press, New York

Yost W A, Patterson R A, Sheft S 1996 A time domain description for the pitch strength of iterated rippled noise. *Journal of the Acoustical Society of America* **99**: 1066–78

W. A. Yost

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Auditory Scene Analysis

Sounds are created by acoustic sources (sound-producing activities) such as a horse galloping or a person talking. The typical source generates complex sounds, having many frequency components. [Its *spectrum* (pl. *spectra*) consists of the frequency and amplitude of every pure-tone component in it.] In a typical listening situation, different acoustic sources are active at the same time. Therefore, only the sum of their spectra will reach the listener's ear. For individual sound patterns to be recognized—such as those arriving from the human voice in a mixture—the incoming auditory information has to be partitioned, and the correct subset allocated to individual sounds, so that an accurate description may be formed for each. This process of grouping and segregating sensory data into separate mental representations, called *auditory streams*, has been named 'auditory scene analysis' (ASA) by Bregman (1990).

The formation of auditory streams is the result of processes of sequential and simultaneous grouping. Sequential grouping connects sense data over time,

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

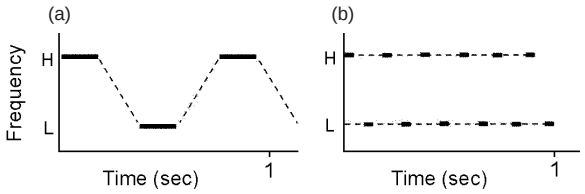

Figure 1

ff

ff

ff

ff

ff

ff

ff

Auditory Scene Analysis: Computational Models

1. Introduction

Human listeners have a remarkable ability to separate a complex mixture of sounds into discrete sources. The processes underlying this ability have been termed 'auditory scene analysis' (Bregman 1990) (see also *Auditory Scene Analysis*). Recently, an interdisciplinary field known as 'computational auditory scene analysis' (CASA) has emerged which aims to develop computer systems that mimic this aspect of hearing (Rosenthal and Okuno 1998).

Work in CASA is motivated both by a desire to understand the mechanisms of auditory perceptual organization and by a demand for practical sound separation devices. Currently, automatic speech recognizers perform badly in noisy acoustic environments. It is likely that their performance could be improved by integrating CASA with speech recognition technology. Other applications of CASA include hearing prostheses and music analysis.

This article considers three general classes of CASA system and discusses their relative merits. Evaluation techniques for CASA are described and outstanding challenges in the field are identified.

2. Theoretical Influences

CASA has been motivated principally by Bregman's (1990) account of auditory organization. However, it has also drawn inspiration from Marr's (1982) work in the field of machine vision. Marr regarded visual processing as a series of representational transforms, each of which makes explicit some aspect of the preceding representation. Starting with Cooke (1993), several workers in the field of CASA have acknowledged Marr's influence and have developed systems in which representational transforms of the acoustic signal play a key role.

3. Three Classes of CASA System

The stages of processing in a CASA system mirror the conceptual stages of auditory scene analysis (Fig. 1). The input to a CASA system is a digitally recorded mixture of sound sources; following feature extraction, this acoustic mixture is decomposed into parts (segmentation). Subsequently, a grouping mechanism identifies segments that are likely to have arisen from the same sound source, and combines them to form a structure that corresponds to a perceptual stream. A final evaluation stage appraises the quality of sound separation.

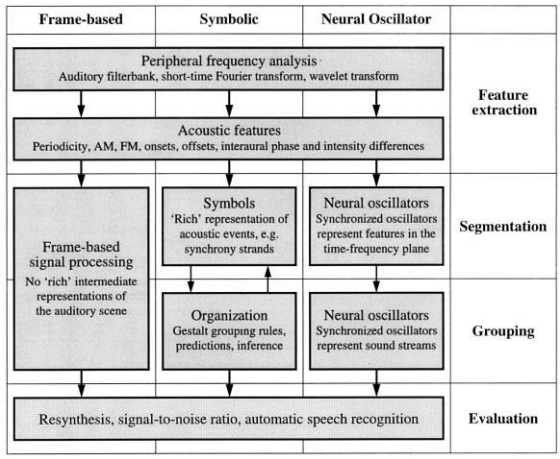

Figure 1
Characteristics of frame-based, symbolic, and neural oscillator approaches to computational auditory scene analysis

Approaches to CASA can be separated broadly into three categories, which differ both in their motivation and in their implementation of the four processing stages identified above. Symbolic systems (e.g., Cooke 1993) adopt a Marrian approach: the segmentation stage derives intermediate representations that characterize properties of the acoustic input. In particular, symbolic systems tend to employ representations of the auditory scene in which temporal continuity has been made explicit. In contrast, frame-based systems (e.g., Denbigh and Zhao 1992) do not use such 'rich' representations; they make grouping decisions within short overlapping time windows ('frames'). Neural oscillator systems (e.g., Wang and Brown 1999) are based more firmly on neurobiological principles than the other types of CASA system. Motivated by reports of neural oscillations in sensory cortex, they achieve segmentation and grouping of acoustic components by forming blocks of synchronized activity in networks of neurons with oscillatory firing responses.

The following sections review the four processing stages of CASA from the perspective of symbolic, frame-based, and neural oscillator systems.

4. Feature Extraction

4.1 Peripheral Auditory Function

The frequency analysis performed by the peripheral auditory system can be modeled by a bank of bandpass filters with overlapping passbands. Each 'auditory filter' channel simulates the frequency response associated with a particular point on the basilar membrane.

Neuromechanical transduction in the cochlea can be modeled by a detailed simulation of inner hair cell function (Brown 1992) or may be approximated by half-wave rectifying the auditory filter outputs and applying a static nonlinearity to compress the dynamic range. CASA approaches which are motivated less strongly by mechanisms of auditory function employ conventional methods of spectral analysis, such as the short-time Fourier transform or discrete wavelet transform (Nakatani and Okuno 1999).

4.2 Auditory Representations of Acoustic Features

Most CASA systems process the peripheral auditory representation in order to derive further representations that make aspects of the acoustic signal explicit. Symbolic and neural oscillator approaches place greater emphasis on these intermediate representations than do frame-based systems.

Human listeners tend to group acoustic components that have the same fundamental frequency (F0). The periodicity of auditory nerve firings provides a cue to F0 and this can be extracted by a running autocorrelation of the activity arising from each auditory filter channel (Brown 1992). Amplitude modulation (AM) in high frequency channels of the auditory filterbank provides another cue to F0 and is used by some CASA systems (Cooke 1993).

Many CASA systems identify the times at which appreciable onsets and offsets of energy occur in the auditory spectrum, since acoustic components which start and stop at the same time are likely to have arisen from the same sound source (Brown 1992). The movement of resonances in the time-frequency plane may also provide information that can be used to track harmonics and formants over time or to identify acoustic components that have a common pattern of frequency modulation (FM) (Mellinger 1992).

Human listeners are also able to separate sounds on the basis of their spatial locations. Phase and intensity differences between the two ears are important cues for spatial location; accordingly, some CASA systems estimate these parameters from a binaural peripheral model (Denbigh and Zhao 1992).

5. Segmentation

The segmentation stage of CASA aims to represent the auditory scene in a manner that is amenable to grouping. Frame-based systems omit this stage of processing: they operate directly on the acoustic features described above.

In many CASA systems, the segmentation stage makes temporal continuity explicit. Typical is the approach of Cooke (1993), which tracks changes in instantaneous frequency and instantaneous amplitude in each channel of an auditory filterbank to create 'synchrony strands' (Fig. 2). Each strand traces the evolution of an acoustic component (such as an harmonic or formant) in the time-frequency plane. This approach offers advantages over frame-based processing: because frame-based schemes make

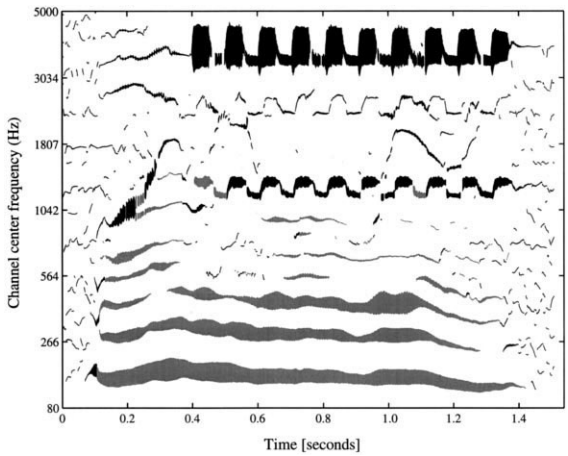

Figure 2
Segregation of speech from an interfering telephone ring in the 'symbolic' CASA system described by Cooke (1993). The utterance is 'why were you all weary?' spoken by a male speaker. A group of harmonically related synchrony strands belonging to the speech source are highlighted in gray. The remaining strands (shown in black) predominantly belong to the telephone sound

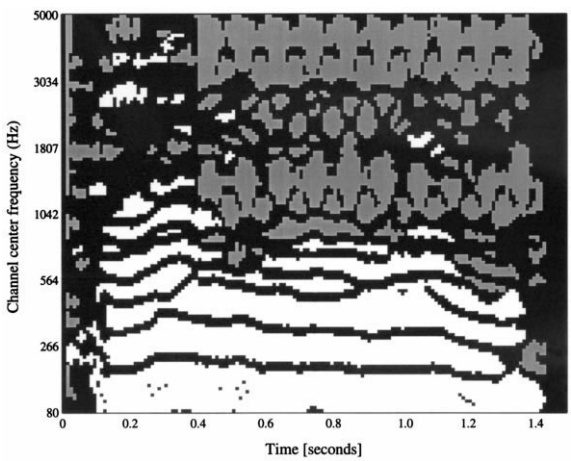

Figure 3
Behavior of the Wang and Brown (1999) system for the sound mixture shown in Fig. 2. Two groups of synchronized oscillators are shown, corresponding to the speech (white) and telephone (gray)

grouping decisions locally in time, they must resolve ambiguities which would have an obvious solution if temporal continuity were taken into account.

Neural oscillator models of CASA also exploit temporal continuity in the time-frequency plane. In this approach, groups of features that belong to the same acoustic source are represented by a population of neural oscillators whose firing is synchronized. Other groups of features are also represented by synchronized populations but oscillators coding different sound sources are desynchronized. The model of Wang and Brown (1999) employs an architecture consisting of a 'segmentation layer' and a 'grouping layer.' Each layer is a two-dimensional network of oscillators with respect to time and frequency. In the segmentation layer, lateral connections are formed between oscillators on the basis of local similarities in energy and periodicity. Synchronized populations of oscillators emerge that represent contiguous regions in the time-frequency plane (Fig. 3).

6. *Grouping*

The grouping stage of CASA identifies acoustic components that are likely to have arisen from the same sound source. This is achieved by implementing Gestalt-like grouping principles that underlie auditory organisation in human listeners (Bregman 1990).

A number of CASA systems apply grouping principles using a simple search engine. For example, Cooke's (1993) system initiates search from a 'seed' synchrony strand, and identifies other overlapping strands which match the seed according to some principle of organization. Figure 2 shows a group of synchrony strands that have been identified as harmonically related in this way. A higher-level grouping stage then integrates groups that have been formed by different organizational principles (such as common AM, common FM, onset and offset synchrony).

This approach to grouping is entirely data-driven (i.e., information flows strictly from top to bottom in Fig. 1). However, it is clear that human listeners are able to predict the behavior of acoustic sources (such as speech) in order to disambiguate an acoustic mixture. A more flexible computational framework is needed to combine this top-down information flow with data-driven grouping mechanisms.

Ellis (1999) describes such a 'prediction-driven' architecture. In his system, a world model is updated in response to errors between observed and predicted signals. Context-sensitive behavior emerges as a result of the interaction between model predictions for (possibly overlapping) segments in the auditory scene. For instance, consider a tone that is interrupted by a burst of noise. The observed acoustic evidence may be compatible with both the predictions made for the noise burst and the predictions made for the tone; since the tone is assumed to be continuous rather than intermittent, it is 'perceived' as continuing through the noise. Hence, Ellis's architecture is able to account for the occurrence of perceptual restoration phenomena (Bregman 1990). A similar 'residue-driven' system is described by Nakatani and Okuno (1999).

In the neural oscillator architecture of Wang and Brown (1999), grouping arises from the dynamics of their oscillator network. In the second layer of the network, excitatory connections are made between segments if the segments are related to the same fundamental frequency and inhibitory connections are made otherwise. As a result, synchronized groups of oscillators emerge which define sound streams (Fig. 3).

7. *Evaluation*

Most CASA systems allow a time-domain waveform to be resynthesized for each separated 'stream.' Hence, sound separation performance can be judged by listening tests or can be quantified as a change in signal-to-noise ratio (Brown 1992).

A number of workers have evaluated CASA systems as components of automatic speech recognition systems. Early approaches treated CASA as a separate preprocessor; resynthesized speech was presented to an unmodified speech recognition algorithm (Weintraub 1985). However, the performance of this approach was disappointing: although interfering sounds were rejected, the speech was distorted and hence little improvement in recognition rate was obtained. Recent efforts to integrate CASA and speech recognition more tightly have produced encouraging results (Ellis 1999).

8. *Future Challenges*

Although the CASA systems described here show promise as useful sound separation devices, their performance is still far below that of human listeners. Some workers claim that improved performance can be obtained by an approach which is based more closely on the neurobiology of hearing (Wang and Brown 1999). Equally, however, 'blind' statistical techniques are able to achieve near-perfect sound separation when certain constraints are placed on the statistical properties of the sound sources and the number of sound sensors (Bell and Sejnowski 1995). It remains to be shown, therefore, whether sound separation systems which are strongly motivated by an auditory account have an advantage over engineering approaches.

Another challenge for CASA systems is real-time performance; currently, they tend to be computationally expensive and nonreal-time. The neural oscillator approach may offer some advantage here, since parallel and distributed architectures are easier to implement in hardware than are the complex algorithms associated with symbolic CASA approaches.

See also: Fechnerian Psychophysics; Psychophysics; Psychophysical Theory and Laws, History of; Signal Detection Theory; Signal Detection Theory, History of

Bibliography

Bell A J, Sejnowski T J 1995 An information maximization approach to blind separation and blind deconvolution. *Neural Computation* **7**: 1129–59

Bregman A S 1990 *Auditory Scene Analysis*. MIT Press, Cambridge, MA

Brown G J 1992 *Computational auditory scene analysis: A representational approach*. Ph.D. thesis, University of Sheffield

Cooke M 1993 *Modelling Auditory Processing and Organisation*. Cambridge University Press, Cambridge, UK

Denbigh P N, Zhao J 1992 Pitch extraction and the separation of overlapping speech. *Speech Communication* **11**: 119–25

Ellis D P 1999 Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Communication* **27**: 281–98

Marr D 1982 *Vision*. W H Freeman, San Francisco

Mellinger D K 1992 Event formation and separation in musical sound. Ph.D. thesis, Stanford University, CA

Nakatani T, Okuno H G 1999 Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication* **27**: 209–22

Rosenthal D F, Okuno H G (eds.) 1998 *Computational Auditory Scene Analysis*. Erlbaum, Mahwah, NJ

Wang D L, Brown G J 1999 Separation of speech from interfering sounds using oscillatory correlation. *IEEE Transactions on Neural Networks* **10**(3): 684–97

Weintraub M 1985 A theory and computational model of auditory monaural sound separation. Ph.D. thesis, Stanford University, CA

G. J. Brown

Auditory System

We perceive an acoustic landscape through a system of exquisite microarchitectures along the auditory pathway to the brain. These structures range from the biomechanical at the periphery to topographical organizations more centrally. Rarely stationary, the acoustic landscape is transformed ultimately into our familiar objects of auditory perception such as speech and music. The ocean of sound waves surrounding us encounters two small points of sensory transduction on both sides of the head—leading remarkably to our perception of a full three-dimensional world of sound. The acoustic medium is often viewed as a complementary source of information; after all, higher order primates for the most part are highly visual beings. However, the physical characteristics of sound offer certain advantages in the detection of an object regardless of its location relative to the animal, such as the ability to detect the location and movement of a sound source in darkness or when the line of sight is occluded. In this article the serial and parallel processing pathways of the auditory system are described briefly, together with observations on the neural coding of auditory information. The term 'neural code' refers to a precise language—inferred by experimental observation—that reliably conveys information along neural pathways. Whether the code is actually used by the animal can only be evaluated with respect to the code's causal impact on the behavior of the animal.

The auditory system, like other sensory systems, consists of a series of connected nuclei that process acoustic information via a complex neural code that necessarily must exist in time and anatomical space (or place). Neural activity in the auditory system is capable of preserving the temporal fine structure of acoustic waveforms, up to a certain degree. Consequently, the timing of individual action potentials likely serves as a major channel of auditory information, particularly at the early levels of the auditory pathway. Centuries ago Jean Baptiste Joseph Fourier (1768–1830) defined the reciprocal relationship between time and frequency of sinusoidal waveforms that form a composite known as the spectrum. A single sinusoidal waveform can be represented as a periodic function in the time domain, but also represented as a single component among others in the composite spectrum of frequencies. This dual relationship plays a prominent role in the representation of acoustic information in the auditory system, beginning with the anatomical frequency mapping on the basilar membrane of the cochlea. The majority of nuclei in the auditory system reflect this frequency mapping, an observation that is commonly referred to as cochleotopic or tonotopic organization.

Acoustic vibrations are initially transferred to the tympanic membrane through the external ear, which causes an increase in sound pressure due to resonance formed in the ear canal. The familiar fleshy appendages we commonly refer to as 'ears' are called pinnae. Pinnae play an important role in our ability to locate sounds in space. Next, the middle ear couples acoustic energy from the external ear via three small bones called the malleus, incus, and stapes to the oval window of the cochlea in the inner ear.

1. Cochlea

The cochlea forms the foundation for signal processing in the auditory periphery (see Geisler 1998). It is embedded deep in the temporal bone and contains a coiled basilar membrane that is shown unwrapped in Fig. 1. Three fluid-filled ducts can be observed (scala vestibuli, scala media, and scala tympani), divided by a stiff cochlear partition composed of the basilar membrane and associated structures. The organ of Corti sits atop the basilar membrane and contains the

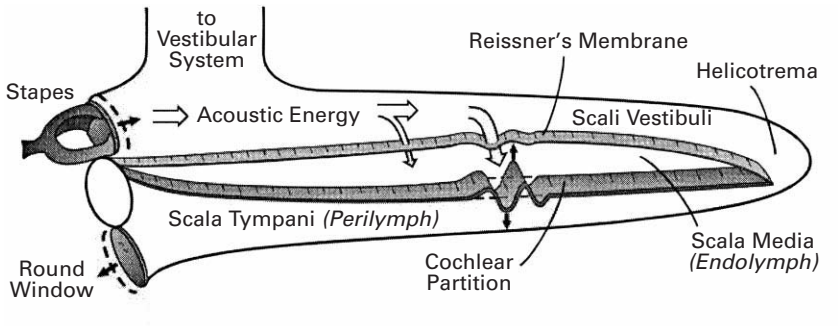

Figure 1
Schematic of the uncoiled mammalian cochlea showing a resonant vibration in the cochlear partition tuned to a particular frequency. The oscillation shows a broadness of tuning that accounts for a broad representation of a single pure tone; this coarse tuning in turn is reflected at subsequent stages in the auditory pathway (source: Geisler, 1998. *From Sound to Synapse: Physiology of the Mammalian Ear*, © Oxford University Press. Reproduced with permission).

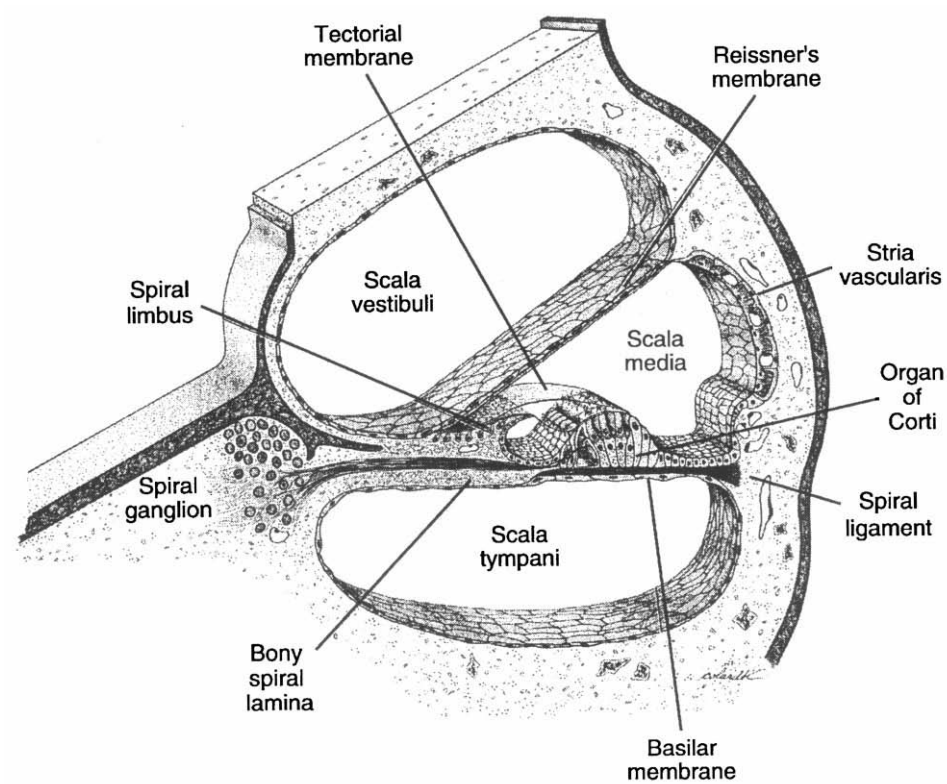

Figure 2
A more detailed cross-section of the cochlea showing the mechanical structures of the cochlear partition (source: Fawcett, 1994. *A Textbook of Histology*, © Chapman and Hall. Reproduced with permission).

transducing hair cells, together with various supporting structures shown in Fig. 2. The uncoiled human cochlea is approximately 35 mm in length and innervated by some 30,000 nerve fibers (Pickles 1988). The basilar membrane partition becomes wider and less stiff with distance from the base. A traveling wave

initiated by the incoming sound travels rapidly from the base towards the apex. The amplitude of the wave increases at first and then decreases quite abruptly, due to the dispersive nature of the basilar membrane as shown in Fig. 1. As a result of these mechanical properties, complex sounds composed of multiple frequencies produce maximum displacement at different places along the basilar membrane, resonating to higher frequencies at the base and lower frequencies at the apex. In effect, the cochlea acts as a spectrum or frequency analyzer, although with limited resolving power.

2. *Auditory Nerve* (*AN*)

Auditory nerve fibers are ideally suited to take advantage of the cochlea's spatial frequency analysis due to nearly uniform innervations along the basilar membrane partition. Between the basilar membrane and the tectorial membrane are hair cells that together form the organ of Corti. Hair cells on the inner side of the organ of Corti are known as inner hair cells. Hair cells nearer the opposite side of the organ of Corti are known as outer hair cells that may actively influence the mechanics of the cochlea. Most auditory nerve fibers connect to inner hair cells that transduce displacements of the basilar membrane partition through shearing motion. A small minority of auditory nerve fibers innervates the outer hair cells.

As a consequence of the mechanical properties of the organ of Corti, complex sounds are decomposed into a spectral series of signals distributed along the cochlear partition. At low sound pressure levels the basilar membrane responds most vigorously to low frequencies at its apex and to high frequencies at the base. This place coding of frequency is translated into discrete populations of AN fibers responding to individual frequencies shown in Fig. 3. A distinctive property of AN fibers is their ability to temporally synchronize (phase-lock) their discharge to a restricted portion of the stimulus cycle. Fibers are capable of encoding information concerning stimulus frequency in terms of the temporal properties of the discharge, in addition to the cochlear place identity of the fiber. In response to a complex signal, with a broad range of frequencies the temporal information could, in principle, provide better resolution of the stimulus spectrum than possible through place information alone.

AN fibers vary in terms of their spontaneous discharge activity. Approximately 60 percent are highly active in the absence of sound. About a quarter show moderate rates of background activity, while the remainder (15 percent) are spontaneously inactive or nearly so.

3. *Brainstem Auditory Pathways*

3.1 *Cochlear Nuclei* (*CN*)

The auditory nerve fibers project to the cochlear nuclear complex made up of the posteroventral (PVCN) and dorsal (DCN) cochlear nuclei and the anteroventral cochlear nucleus (AVCN) (refer to Fig. 4). A complete tonotopic organization is main-

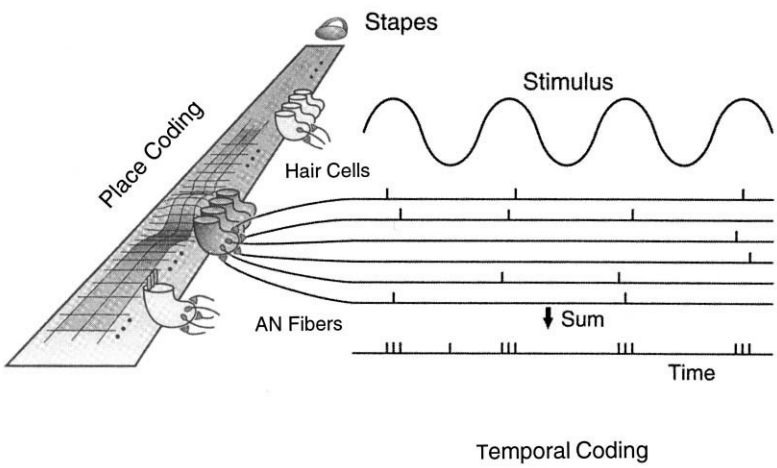

Figure 3
Schematic showing the transduction of deflections of the cochlear partition by hair cells, leading to auditory nerve fiber spikes. This figure illustrates the ensemble temporal code for a periodic waveform and the cochlear place code for its frequency (source: adapted from Geisler, 1998. *From Sound to Synapse: Physiology of the Mammalian Ear*, © Oxford University Press. Reproduced with permission).

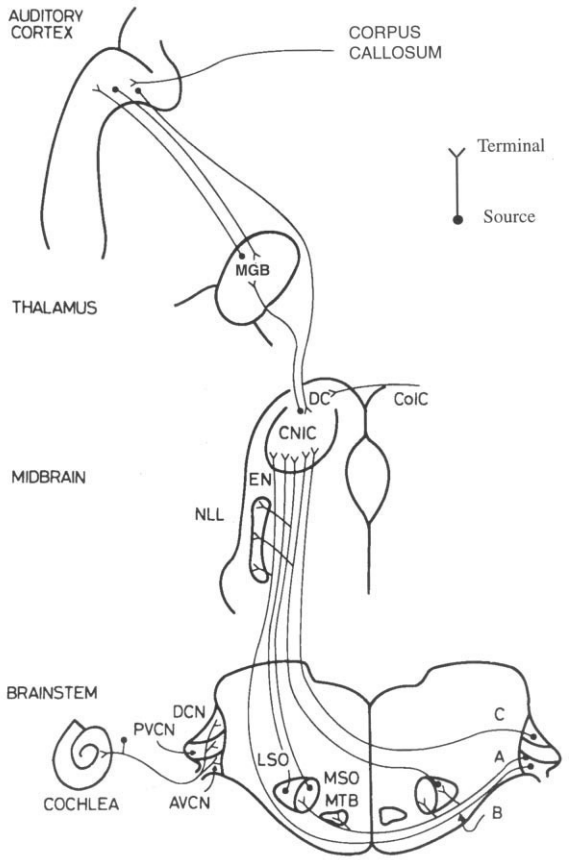

Figure 4
Schematic of the auditory pathway. See text for definitions. Parallel processing streams are denoted A, B, and C (source: adapted from Aitkin, 1990. *The Auditory Cortex*, © Chapman & Hall. Reproduced with permission).

tained in each subdivision. There is a varied taxonomy of cell types in the cochlear nuclei. The branch of the auditory nerve fiber that innervates the AVCN usually has a large, specialized calyceal ending, also known as an end bulb that more or less engulfs the cell body of bushy cells (Rhode and Greenberg 1992). Because of this secure connection, bushy cells faithfully convey timing information from the auditory nerve and therefore the cochlea. Stellate cells, found in both the AVCN and PVCN, receive their input from auditory nerve fibers spatially distributed on their dendrites. Octopus cells, found in the PVCN, possess long dendrites that serve to integrate information across neighboring frequencies—perhaps serving to bring neighboring spectral information into a common temporal register (Oertel 1999). The form and structure of these different cells is reflected in the individual firing patterns—bushy cells evidence firing patterns that are similar to that of auditory nerve fibers, stellate cells typically show a 'chopping' pattern, and octopus cells respond primarily to the onset of a tone whose frequency is within the response region of the cell.

3.2 Superior Olivary Complex (SOC)

The SOC is made up of many nuclei, but the main nuclei associated with the ascending auditory pathway are the lateral superior olive (LSO), the medial superior olive (MSO), and the medial nucleus of the trapezoid body (MTB). Other nuclei in the SOC are descending efferents involved in controlling the stiffness of the basilar membrane and the middle ear reflex. The primary afferents projecting to the LSO, MSO, and MTB are from the AVCN and the PVCN (Aitkin 1986). Neural computation in the SOC nuclei is inherently based on the physics of the relative acoustic delays resulting from the spatial separation of the two ears.

The difference in the interaural time-of-arrival of sound pressure waves at the two ears is an important source of information for the degree of 'leftness' or 'rightness' of a sound source, or in common spherical (astronomical) coordinates, its azimuthal position. Based on early anatomical evidence, Lloyd Jeffress speculated on a coincidence detection mechanism for operating on interaural time differences that has become known as the Jeffress model (1948). Interestingly, there is now considerable anatomical and physiological evidence for such a mechanism in the auditory brainstem, as well as psychophysical support. This is perhaps one of the rare examples where an abstract mathematical model predicted future anatomical and physiological observations so well. Electrophysiological observations have implicated the medial superior olive (MSO) in mammals (nucleus laminaris in birds) as the anatomical site of cross-correlation of sound synchronized responses from the two ears (see Oertel 1999). However, the inference that the MSO serves only as a simple coincidence detector of excitatory inputs from the two ears underestimates the full functionality of the circuit. Neurons in the MSO also respond to sounds presented in just one ear, so coincidence from the two sides is not exclusively required for responding. Also, neurons in the MSO receive inhibitory inputs from the MTB, which may serve to improve the temporal resolution of the circuit. The MSO receives tonotopically organized input from both ears, directly and indirectly from the AVCN, where the frequencies are systematically mapped on the MSO. The map is biased toward the lower frequencies, likely due to the low-frequency restriction of useful interaural time and phase differences resulting from the acoustic delay between the two ears. Electrophysiological evidence suggests that the LSO is the site where sound level differences between the two ears are computed. Neurons of the LSO also detect the

timing of excitation from one ear and inhibition from the other ear, so LSO neurons are sensitive not only to interaural level but also interaural time differences (Oertel 1999). The ability to detect rapid timing changes in interaural level differences may have implications for the perception of moving sounds.

3.3 *Inferior Colliculus* (*IC*)

The central nucleus of the inferior colliculus (CNIC) is the target for the vast majority of terminals from the CN and SOC and therefore represents a common point of convergence serving as a relay to higher cortical processing (Aitkin 1990). The ascending fibers run in a tract called the lateral lemniscus, and send collaterals to the nuclei of the lateral lemniscus (NLL). A tonotopic representation can again be found in the CNIC. There is a precise arrangement of neurons in the CNIC, with a continuum of frequency selectivity, with isofrequencies represented as planes in a three-dimensional structure. The LSO projects to the CNIC on both sides of the brainstem, and the MSO projects primarily to the ipsilateral (or same) side of the CNIC. There are also projections from the contralateral (or opposite) side DCN that terminate in the CNIC. Hence, the CNIC serves as an important site of convergence for several parallel paths of acoustic information. Interaural phase sensitive neurons in the SOC that converge on the lateral lemniscus and the CNIC likely account for the sound localization properties of IC neurons. Neurons sensitive to higher frequencies are less able to follow the temporal fine structure of complex sounds such as speech at the level of the CNIC compared to that of neurons lower in the pathway (AN and CN). Consequently, it is likely that periodic sounds such as speech, that were previously coded in time, may be recoded at some level into a place code for periodicity in the CNIC. Neurons in the external nucleus (EN) and dorsal cortex (DC) tend to respond to the stimulus onset and have lost their frequency selectivity. This absence of frequency selectivity is sometimes referred to as being 'diffuse' in contrast to tonotopic. The EN is not modality specific and represents a possible motor pathway in the auditory pathway.

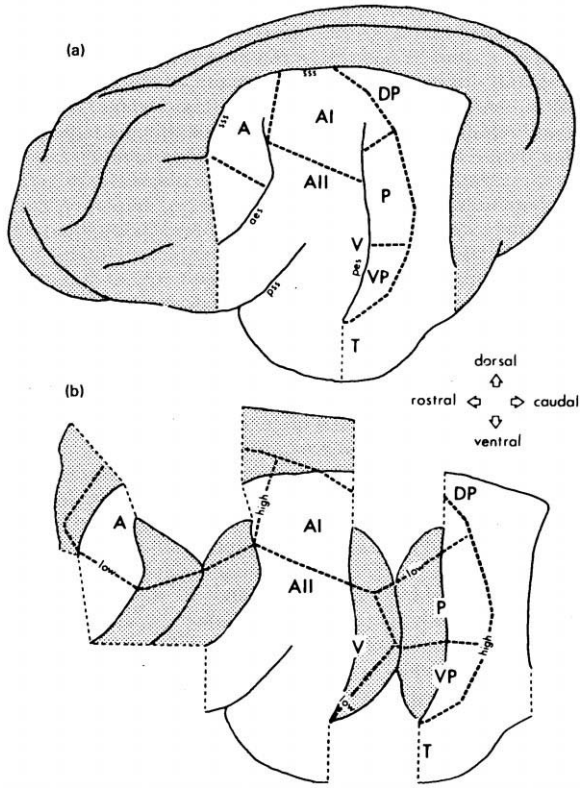

Figure 5
Left cerebral hemisphere of the cat showing the locations and boundaries of auditory cortical fields. The lower diagram shows the cortex unfolded revealing the auditory cortex hidden within the fissures. Four of the fields have a known tonotopic organization—primary auditory (AI) cortex, anterior auditory field (A), posterior field (P), ventral posterior field (VP). *High* and *Low* denote the highest and lowest frequencies represented in each field. Surrounding fields include secondary auditory (AII) cortex, temporal field (T), dorsal posterior field (DP) are likely to be more diffuse (source: Imig and Reale 1980. *Journal of Comparative Neurology* © Reproduced with permission).

4. *Thalamocortical System*

The auditory thalamus and cortex can best be thought of as an integrated system with feedback (see Aitkin 1990, Clarey et al. 1992, de Ribaupierre 1997, Rouiller 1997). Projections from the thalamus to cortex are reciprocated to afford feedback control of the cortex, producing a powerful information processing system.

4.1 *Medial Geniculate Body* (*MGB*)

The MGB is the major auditory nucleus of the thalamus and can be divided into ventral, dorsal, and medial divisions. The major ascending auditory pathway from the CNIC is through a bulge in the lateral surface of the brainstem of the IC to the ventral division of the MGB. The ventral division projects to the primary auditory cortex (AI), as well as the adjacent cortical fields, anterior (A), posterior (P), and ventral posterior (VP). The medial division is less specifically auditory, projecting diffusely to auditory cortex, and the dorsal division projects to the more diffuse secondary cortex (AII) and temporal field (T) (see Fig. 5).

4.2 Auditory Cortex

The primary auditory cortex, as in other sensory cortices, has systematic columnal organization that exists within a three-dimensional volume. The organization of the auditory cortex has largely been explored in the cat, but other species, including primates, have been investigated as well. A faithful tonotopic map can be observed along one surface dimension of AI in nearly all mammalian species. In some cases, such as the mustache bat, the organization, although systematic, can be quite complex (Suga 1982). The dimension orthogonal to the tonotopic axis shows no change in frequency and hence reflects bands of isofrequency. Several functional organizations other than frequency have been shown. First, there is evidence of periodic patches of binaural interactions along the isofrequency bands. In addition, systematic gradients of frequency tuning, response latency, and best intensity have also been observed (Schreiner 1998). Although the tonotopic dimension is fairly well structured, the organizations orthogonal to the frequency axis are less so, and highly individualized across different animals.

Auditory cortex is generally consistent with the classical six-layer architecture of other cortical areas (Winer 1992). There are specific relationships among the interconnections between the six layers (see Fig. 6). Within a cortical column, which descends through the cortical layers, there is a distribution of processing that begins with input from the thalamus (MGB) to layer IV. Information is then sent to other layers of the same column where it is integrated with other sources of information and then broadcast to other columns, subdivisions, and brainstem nuclei. Cortical projections back to the thalamus originate in the deep cortical layers V and VI. These descending projections return to the same tonotopic regions of the MGB that projected originally to the cortex. The functional role of the corticothalamic feedback is likely that of adaptive filtering of the acoustic information specific to each of the subdivisions, since the parallel streams remain segregated in the feedback pathways as well. Layers II and III are involved in local cortical interconnections. Various areas of auditory cortex also have connections between hemispheres via the corpus callosum.

The overlay of several dimensions of functionality (i.e., frequency, intensity, broadness of tuning, and so on), onto essentially two surface dimensions, means that the acoustic landscape must be represented by a multivariate pattern of action potential firings across the cortical expanse. Decoding these patterns into particular speech sounds or sound locations in space is a classical example of the perceptual inverse problem in neural computation. The precise and orderly thalamocortical feedback circuits, although computationally powerful, make interpretation of cortical firing patterns very difficult. However, advances in computational neuroscience (Rieke et al. 1997), particularly the analysis of timing patterns across ensembles of cortical neurons, are just beginning to shed light onto the possible neural codes for our objects of auditory perception.

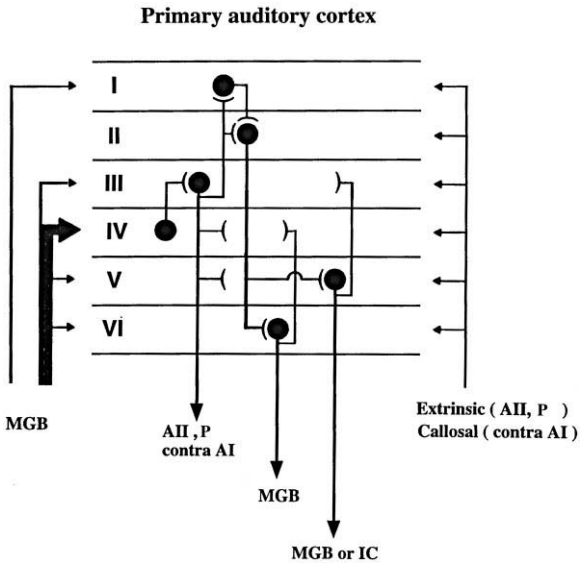

Figure 6
Schematic of neuronal connections in primary auditory cortex (source: adapted from Mitani et al. 1985, *Journal of Comparative Neurology* © reproduced with permission).

See also: Auditory Models; Audition and Hearing; Cross-modal (Multi-sensory) Integration; Dyslexia, Developmental; Neural Plasticity in Auditory Cortex; Sensation and Perception: Direct Scaling; Signal Detection Theory: Multidimensional

Bibliography

Aitkin L 1986 *The Auditory Midbrain: Structure and Function in the Central Auditory Pathway*. Humana Press, Clifton, NJ

Aitkin L 1990 *The Auditory Cortex: Structural and Functional Bases of Auditory Perception*. Chapman and Hall, London

Brugge J F 1992 An overview of central auditory processing. In: Fay R R, Popper A N (eds.) *The Mammalian Auditory Pathways: Neurophysiology*. Springer-Verlag, New York, pp. 1–33

Clarey J C, Barone P, Imig T J 1992 Physiology of thalamus and cortex. In: Fay R R, Popper A N (eds.) *The Mammalian Auditory Pathways: Neurophysiology*. Springer-Verlag, New York, pp. 232–334

Fawcett D W 1994 *A Textbook of Histology*. Chapman and Hall, London

Geisler C D 1998 *From Sound to Synapse: Physiology of the Mammalian Ear*. Oxford University Press, New York
Jeffress L A 1948 A place theory of sound localization. *Journal of Comparative Physiology and Psychology* **41**: 35–9
Mitani A, Shimokouchi M, Itoh K, Nomura S, Kudo M, Mizuno N 1985 Morphology and laminar organization of electrophysiologically identified neurons in the primary auditory cortex in the cat. *Journal of Comparative Neurology* **235**: 430–47
Oertel D 1999 The role of timing in the brain stem auditory nuclei of vertebrates. *Annual Review of Physiology* **61**: 497–519
Pickles J O 1988 *An Introduction to the Physiology of Hearing*. Academic Press, London
Reale R A, Imig T J 1980 Tonotopic organization in auditory cortex of the cat. *Journal of Comparative Neurology* **192**: 265–91
Rhode W S, Greenberg S 1992 Physiology of the cochlear nuclei. In: Fay R R, Popper A N (eds.) *The Mammalian Auditory Pathways: Neurophysiology*. Springer-Verlag, New York, pp. 94–152
Ribaupierre F de 1997 Acoustical information processing in the auditory thalamus and cerebral cortex. In: Ehret G, Romand R (eds.) *The Central Auditory System*. Oxford University Press, New York, pp. 317–97
Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W 1997 *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA
Rouiller E M 1997 Function organization of the auditory pathways. In: Ehret G, Romand R (eds.) *The Central Auditory System*. Oxford University Press, New York, pp. 3–96
Schreiner C E 1998 Spatial distribution of responses to simple and complex sounds in the primary auditory cortex. *Audiology and Neuro-otology* **3**: 104–22
Suga N 1982 Functional organization of the auditory cortex: Representation beyond tonotopy in the bat. In: Woolsey (eds.) *Cortical Sensory Organization. Vol. 3, Multiple Auditory Areas*. Humana Press, Clifton, NJ, pp. 152–218
Winer J A 1992 The functional architecture of the medial geniculate body and the primary auditory cortex. In: Webster D B, Popper A N, Fay R R (eds.) *The Mammalian Auditory Pathways: Neuroanatomy*. Springer-Verlag, New York, pp. 222–409

R. L. Jenison

Australia and New Guinea, Archaeology of

1. Dating the Arrival of Humans

Although today Australia and New Guinea comprise three nations on two landmasses, during times of lowered sea levels in the late Pleistocene they formed a single continent, commonly called *Sahul* or *Greater Australia*. There is general consensus that the first human colonizers were both anatomically and behaviorally modern humans; indeed the necessary crossing of a minimum water barrier of c. 90 km to reach New Guinea or Australia in the late Pleistocene constitutes one measure of that modernity.

Beyond that, there is little consensus for the date of earliest arrival. Oldest claims, in the 100–200 ka (thousands of years) range, all depend upon ecological evidence—usually charcoal counts in pollen cores interpreted as anthropogenic rather than natural fires—and are claimed in both the north and south of the Pleistocene continent, as well as Eastern Indonesia. The second group of claims consists of analyses of mtDNA diversity which put humans in New Guinea and Australia c. 70–100 ka. A third group of claims associate direct archaeological evidence with uranium series, luminescence, and electron spin resonance dates in the 50–70 ka range. The fourth category includes those sites which form the beginning of a continuous sequence of occupation in the Pleistocene continent which may extend to 43–45 ka sidereal years ago. Certainly by c. 35 ka all the major ecological zones of Australia and New Guinea had been colonized by humans.

Claims in the first group have little support because they lack direct artefactual evidence and because it is difficult to accept that humans in Greater Australia before 100 ka could have been anatomically and behaviorally modern. *Homo erectus* probably reached Java one million years ago but was apparently prevented from moving further east by the water barrier known as Wallace's Line.

Three specifically archaeological claims put humans in Australia before or at 50 ka. The western New South Wales burial known as Mungo 3 has recently been redated to c. 56–68 ka using uranium series and ESR (electron spin resonance) techniques on the skeletal elements and OSL to date the surrounding deposits. This is particularly dramatic because the Mungo 3 human was ritually buried, covered in red ocher. Such burial is seen as a marker of modern behavior, and if as old as this, would be the oldest unambiguous ritual burial of *H. sapiens* presently known. The second claim is for two rockshelter sites in the Northern Territory, Nauwalabila and Malakunanja, where the lowest stone artifacts are associated with TL and OSL dates having central tendencies between 50 ka and 60 ka. Here critics question the *in situ* nature of the lowest artifacts being dated rather than the dating techniques themselves (see *Chronology, Stratigraphy, and Dating Methods in Archaeology*). Similar criticisms have been leveled at the third example, the recent AMS dating of the Devil's Lair cave in southwest Western Australia to c. 48–50 ka, where the lowest levels of the site which produced this date are water-scoured and contain only a few stone flakes, some not clearly of human manufacture.

Defenders of the longer chronologies criticize the shorter chronology on the grounds that the great majority of dates are based on older radiocarbon dating which was unreliable for ages much over 40 ka, so that this age for earliest arrival might easily be a product of the technique. However, modern radiocarbon techniques now produce accurate dates in the

range 35–45 ka and have done so regularly in Australia in the 1990s for nonarchaeological samples, but only on a very few occasions for archaeological sites.

The resolution of this debate is needed to understand the rate at which Greater Australia was colonized, and the adaptive strategies used to occupy a wide range of new environments. Equally important are the implications of this debate for understanding the spread of modern humans elsewhere in the world, since behaviorally modern humans are unknown before c. 50 ka anywhere in Eurasia.

1.1 Routes of Entry

The Sunda shelf of Southeast Asia is separated from the Sahul shelf of Australia and New Guinea, and this break forms a significant biogeographical boundary first recognized by Alfred Russell Wallace. Wallace's Line separates an Asian fauna which includes primates, carnivores, elephants, and ungulates from the marsupial fauna of Australia and New Guinea. However, the widespread elements of Asian flora in Indonesia, New Guinea, and into northern Australia indicate that this boundary still permitted the natural transfer of this flora.

Two routes of entry from Southeast Asia seem possible, the first and more northern route through the Indonesian Moluccan archipelago to New Guinea and a second more southern route along the Flores chain to Timor and thence to northern Australia. Any of the subroutes which might form the northern route had two sailing advantages which the southern route lacked—intervisibility between departure point and target landfall on all legs and the ability to return to the departure point on existing winds and currents. Sea craft could journey blind from Timor to Australia on the summer monsoon, but not return. Interestingly, altered heights of islands because of sea level fluctuations had no appreciable effect on either intervisibility or distance between landfalls.

The earliest coastal site in New Guinea is at its extreme eastern end, emphasizing the absence of systematic archaeological exploration of the western half of the island. This site, an open site on uplifted coralline terraces on the Huon Peninsula, has yielded dates of c. 40,000 or a little older for large tanged and waisted flaked axes which must have been hafted.

1.2 New Guinea Highlands

New Guinea is dominated by a steep central cordillera rising to 4,000 m in places. Because of rainfall patterns and cloud cover, Pleistocene colonists likely settled the series of intermontane valleys ranging between 1,200 m and 2,400 m. Currently the oldest dates for humans in the highlands are c. 30,000 ka, the time lag from the coast perhaps indicating the significant adaptations required to settle this region. The subsistence patterns reflected in excavated sites like the Nombe rockshelter reflect a strategy of hunting and collecting across these altitudinal zones. While forests and grasslands contained mainly small game, Pleistocene New Guinea was also home to a small number of large mammals, all now extinct. At Nombe, four species—all forest browsers—occur between c. 25,000 and c. 14,000 ka, but it is uncertain whether these constitute human prey remains. The general paucity of game animals in New Guinea is, however, offset by the huge array of plants, of which more than 200 are used for food. While little direct evidence exists for the development of plant processing and plant storage, it is likely that these processes developed from the beginning. The Huon axes also occur at Nombe, Kosipe and other early sites, where they have been identified as tools used to thin forest patches by trimming and ring-barking so as to promote the growth of useful plants. The burning of forests at this time is reflected in the pollen records and supports this interpretation.

If accurate, these data suggest a long evolution of plant management which culminates in a distinctive highlands agriculture at 9–6000 BP at Kuk in the Wahgi Valley. Here evolutionary stages of garden technology have been identified, which include hydraulic practices involving both drainage and irrigation. The most recent phases reflect the familiar root crop/arboriculture/pig-husbandry systems present in the highlands today. Pigs, not indigenous to New Guinea, appear in the archaeological record 5–6,000 ka with some claims back to 10,000 ka.

1.3 Island Melanesia Before Lapita

Moving into the nearer islands of Melanesia required adapting to the simplified terrestrial ecology of an oceanic world. The initial colonists adopted strategies which compensated for this reduced range of edible plants and terrestrial animals by high mobility across large territories. This, in turn, depended on safe, if simple, watercraft. Initially it appears such craft took people to scattered resources, but over time, as populations and specialized resource knowledge increased, moving resources to people increasingly replaced the first strategy.

Because of tectonic uplift and deep submarine contours along the north coast of New Ireland, Pleistocene sites now normally flooded by later rising seas have been preserved. The oldest known of these, Buang Merebak, recently redated to 40–41,000 ka, and Matenkupkum (c. 35,000 ka) are about 140 km apart, and both contain marine shell middens which are very similar. Both contain marine fish bones indicating some form of fishing at this age, and both reflect strategies of shellfish collecting and processing the

shells for food and tools. People also moved inland, quarrying chert outcrops near Yombon in New Britain at c. 36,000 ka. At Kilu Cave in the northern Solomons, stone tools dated between 29,000 ka and 20,000 ka carry residues of *Colocasia* and *Alocasia* taro, in quantities sufficient to suggest that successful selection had already taken place for starch-rich tubers.

Getting to Kilu required a sea journey about twice that of any prior necessary crossing between mainland Southeast Asia and the Bismarcks, a voyage without the aid of two-way intervisibility. But by 21,000 ka people had reached the Admiralties Group, a voyage of over 200 km which required that 60–90 km be traversed totally out of the sight of any land; this is the earliest known example of true seafaring anywhere.

By 20,000 ka, moving resources to people is reflected in New Ireland in the appearance of obsidian from a New Britain source 300–350 km distant and the introduction of *Phalanger orientalis*, a marsupial which became an important food source. It is the first of a series of animal translocations in the region and flags the possibility that useful plants and nut bearing trees were also moved about well before the end of the Pleistocene.

A reduction in cave use in the region during the early Holocene raises the question of whether these islanders were more dependent on cultivated plants than previously. While the evidence is piecemeal, gradual economic intensification is reflected in aspects of the archaeology such as intensive and extensive trade in obsidian, plant residues on tools, domesticated and transported nut remains in sites, elaborate stemmed and hafted obsidian and chert tools, stone and shell ground axes, open site settlement patterns and expansion onto smaller offshore islands, all aspects which seemingly demand more substantial food productive systems than those provided solely by hunting and gathering.

Maritime trading systems with their genesis in the Pleistocene elaborated during the later Holocene, as horticulture underwrote village development both in the near islands and around the New Guinea coast. The growth of these systems, particularly after the spread of Lapita-introduced pottery technology, has been a major research concern of archaeologists in the region (see *Pacific Islands, Archaeology of*).

2. *Australia*

Regardless of where first landfall occurred, the move south was into environments more foreign in their floras and faunas and in increasing contrast to presumed tropical humid homelands. Although more than 160 Pleistocene age sites are known in the region, they occupy a diverse range of environments and are frequently hundreds of km apart, such that no neat synthesis of past human behaviors is yet possible. A few short generalizations will provide some background. (a) With some glaring exceptions, such as edge-ground axes in Arnhem Land at c. 20 ka, Pleistocene flaked stone tools across Australia are mostly amorphous, showing little of the development which characterizes the Upper Palaeolithic in Europe or the Later Stone Age in Africa. (b) A host of larger marsupials, collectively dubbed megafauna, appear to have become extinct around the time of human arrival, although the dating of the extinction event(s) is poor. While it is difficult not to implicate humans in this extinction, the processes are unclear. Direct extinction by hunting appears unlikely, given the range of species, the size of the continent and its varieties of environments, the presumed low numbers of humans, and the conspicuous absence of kill sites. Many commentators appeal to less direct effects, particularly habitat modification by the human use of fire, but not even this effect is universally endorsed. Arguments of climate change are still advanced to counter human involvement, but these are also widely criticized. (c) A tradition of small blade tools and points appears c. 4 ka throughout Australia, but not in Tasmania, with clearer stylistic differences between them identified at the regional level. (d) Despite the example of New Guinea to the north, Aborigines did not adopt or develop agriculture, remaining a continent of hunter–gatherers until European arrival. Even so, the successful colonization of even marginal areas of the continent reflects the adaptability of these humans, while the widely documented development of art and engraving on rocks and cave walls, now demonstrated from earliest arrival, reflects an intellectuality and social awareness among Aborigines previously only documented ethnographically.

2.1 *The Arid Zone*

The arid zone occupies an area of c. five million square km—about 70 percent of Australia's land surface—and during the Pleistocene this area was even larger. In this region occupation and movement has always been constrained by the distribution of food and water. Areas of relatively high relief, the ranges of the central desert and the Kimberly and the Pilbara in particular, concentrate run-off in stream channels and waterholes; similarly, riverine floodplains contain perennial and seasonal streams which contain surface water in some seasons and store accessible subsurface reserves. The riverine floodplains, particularly those along the Darling system in New South Wales, offer a range of aquatic resources, including fish, shellfish, hydrophytic plants, and water birds.

Before 25 ka, conditions were perhaps wetter but cool to cold, but between 25 ka and 12 ka conditions

were much colder and drier, especially at the last glacial maximum. Much of this zone would have been inimical to human settlement at this time. Although people were on its fringes by 40 ka and occupying the 'dead heart' before 30 ka, many sites were abandoned at the height of the ice age.

People were at Mandu rockshelter on the arid Western Australian coast c. 34 ka and left behind shell beads apparently used for personal adornment, as well as ochre carried there from the Pilbara area. As the sea dropped, offshore islands were incorporated into coastal territories as people followed the receding coastline. Today the bare rocks of the Montebello islands contain small rockshelter sites which reveal human presence at c. 27 ka, with both terrestrial and coastal resources represented in the deposits, and again between 7 ka and 10 ka, when the prograding sea was again near. These sites were not used between times and were abandoned when they again became offshore islands.

Thus models of changing land-use patterns put forward by arid-region scholars all emphasize mobility and adaptation to changing environmental conditions and suggest small populations thinly spread across the arid zone. Diets may not have been greatly different from those recorded ethnographically in these regions. They are generally similar and broadly based and included, in addition to marsupials, reptiles, birds, larvae, tubers, fruits, seeds, and aquatic resources where available. Recent claims that grass and tree seeds might have been gathered and processed in the Pleistocene and particularly before the last glacial maximum is notable because although seeds are nutritious, their gathering and processing is laborious. For this reason, seed processing elsewhere in the world is seen as a key indicator of changes in human behavior connected to the appearance of new technologies, larger and more settled populations and in places the establishment of agriculture. Its presence in an Australian form in preglacial arid Australia has significant theoretical relevance. Traces of seeds and other plant food remains from Pleistocene levels in Carpenters Gap in the northwest is complemented by pieces of grinding stones dating to c. 30 ka at Cuddie Springs some thousands of kilometers to the east. These carry usewear and residue traces suggesting the processing of starchy and siliceous plants, evidence coincident with a local change to grasslands seen in the pollen record at this time.

Data from the Mungo region of western New South Wales indicate how favorable these lake systems were when lacustral conditions ensued. A range of resources—fish, shellfish, crustaceans, and birds—were taken. Fish may have been taken using nets and spears and possibly traps. A range of mainly small terrestrial game was also captured and again in this region unspecialized grinding tools suggest at least generalized processing of plant foods and seeds during the Pleistocene.

2.2 *Tasmania*

Today, and for the last 12 ka, humans have occupied the north and east of Tasmania. In the southwest, temperate rainforest continues to exclude settlement. During the late Pleistocene, however, this situation was reversed. At times glacial deserts made the east inhospitable, while colder and drier conditions forced the rainforests in the southwest into refugia at lower altitudes, opening up the montane valleys to open forest with patches of grassland on limestone substrates in the valley bottoms. Since 1980 c. 30 sites, predominantly limestone caves have been located in the rainforest and c. 12 tested by excavation. The oldest, Warreen, was occupied c. 35 ka and all were abandoned by 12 ka, when warmer and wetter conditions encouraged the return of the forests, which drove out game animals and their human hunters. At the height of the last glacial maximum the mountains had glaciers extending as low as perhaps 700 m above sea level, only a few hundred meters above some of the occupied caves. With temperature reductions of c. 7 °C and winds reducing temperatures sometimes tens of degrees it was indeed an extreme climatic place, some 40 degrees of latitude south of their ancestors' equatorial point of departure.

The cluster of known sites provides a rare glimpse at a regional pattern of Pleistocene behavior extremely different to that depicted for arid Australia. Here the subsistence strategy was the opposite of broad spectrum. There are no obvious vegetable staples which might have supported even small transient groups and the plentiful faunal remains repeat the same pattern from site to site. Up to 80 percent of bones from humanly consumed animals were from the single species *Macropus rufogriseus*—the Bennetts wallaby, supplemented in small amounts by wombat and platypus. In more than 630,000 animal bones processed from eight sites, there is not a single extinct megafaunal species represented, although some are present in nonarchaeological deposits in caves in the same region, suggesting that those species known from Tasmania had become locally extinct before human arrival, or that they were not hunted or scavenged.

3. *Holocene Continuities and Changes*

As the forests drove Tasmanians from the southwest, rising sea levels at the end of the Pleistocene cut off Tasmanian access to the Australian continent and for 8 ka the Tasmanians were deprived of human contact beyond their island. When reunited with the world at the end of the eighteenth century, they no longer made bone points, caught fish, or made fire, and were obliged to keep fire burning or rekindle flame from neighboring groups. Their technology comprised no composite or hafted tools, and consisted only of single piece wooden spears, clubs, digging sticks, kangaroo

skin capes, skin and kelp water containers, string, and small baskets. A simple stone tool technology was compared by Europeans to Mousterian industries. While the loss of fish in the diet c. 4 ka may reflect some established island-wide taboo, this dietary loss was offset by the continued capture of seals and terrestrial mammals and an expanded range of shellfish species. In the recent Holocene more use seems to have been made of the less hospitable south and west coasts and people also used nearby offshore islands. Bundles of reeds tied in the shape of small canoes are known ethnographically where they were used as flotation devices.

In contrast, the end point of Holocene changes on the mainland differentiated Tasmanians and mainland Australians so distinctly that early explorers and anthropologists considered them different races. But while the Tasmanian devil and marsupial tiger or thylacine—both extant in Tasmania when Europeans arrived—had disappeared from the mainland when the dingo was introduced there from Southeast Asia or New Guinea c. 4 ka, few other Holocene changes can clearly be attributed to external influences. In particular, the temporal and distributional evidence for the appearances of small points, geometric microliths, and other tools of the 'Australian Small Tool Phase' during the mid and late Holocene suggest different and regional origins within the country. While some of these tool types disappeared perhaps 1–2 ka in some regions, in the north larger bifacially points and other macroblade tools appeared more recently.

This proliferation of regional tool types, increases in site numbers, site types, and the range of organic and inorganic artifacts to be found in them, together with the appearance of cemeteries, evidence for exchange networks and labor-intensive specialized food procurement activities like constructing artificial channels to capture migrating eels has been taken as a package by some Australian archaeologists to promote a model of socioeconomic intensification in the late Holocene. Even though stationary sea level and the relatively young ages of these sites clearly promotes better preservation and higher visibility, this package is contrasted with the Pleistocene and early Holocene, where population densities were lower, territories larger, people were less sedentary, there was less social ranking and less complex food procurement strategies. Critics of this model question its testability in the archaeology and some prefer to continue seeking environmental explanations for the changes seen in the archaeological record.

The unsurpassed heritage of Australian rock art has been recorded, classified, and interpreted for more than a century. The ability to now date many of the varied styles directly is a recent revolution which is enabling the integration of art with other aspects of the archaeological record. Many new and ingenious techniques, such as luminescence dating of sand grains in mudwasp nests overlying painted surfaces, or dates that use carbon in oxalate salts to indicate the time when such encrustations covered the art, are now demonstrating a long Pleistocene history for many paintings, providing a further window on the nature of hunter–gatherer intellectuality.

See also: Archaeometry; Chronology, Stratigraphy, and Dating Methods in Archaeology; Hunter–Gatherer Societies, Archaeology of; Migrations, Colonizations, and Diasporas in Archaeology; Pacific Islands, Archaeology of

J. Allen

Australian Aborigines: Sociocultural Aspects

Australian Aborigines are surely the most famous of indigenous peoples, due to their place in western evolutionary thought. In the nineteenth and early twentieth centuries they were studied as living evidence of the earliest, most primitive form of a universal human past. All of us, it was asserted, were once hunter-gatherers like these. The nineteenth-century anthropologist Lewis Henry Morgan exemplified primitive forms of kinship from descriptions of the Kamilaroi and Kurnai of South-eastern Australia (Fison and Howitt 1880/1991). Aborigines were central characters in the origin myths of Freud and Durkheim. Such erudite studies occupied elite scholarship, but theories of primitivity and evolution also formed the foundation of public understanding and government policies.

Controversies about Australian Aborigines in the first half of the twentieth century concerned their antiquity and the proper way to describe their social traditions. Today, there is intense public dissension about the category Aboriginal itself, as well as about political and symbolic representation, rights to land, and rights to compensation for past wrongs. Since about 1970 many Aboriginal people have participated vigorously in these controversies, marking a major shift in the place of indigenous peoples in the West's academic canon. At the same time radical changes have occurred in Aborigines' conditions in relation to the Australian nation and state in law, scholarship, political activism, and public perception.

Archaeologists and physical anthropologists have found evidence of human habitation in Australia from 50,000 years ago, though prehistorical anomalies and mysteries are still being unraveled. The search for the 'racial origins' of Aborigines began to be muted in the 1950s because its language was that of scientific racism. More recently, such research has had to face Aborigines' assertions that the skeletons in Australian soil

are the ancestors of contemporary Aborigines and should be left undisturbed. Yet recoil at the racial connotations of past scholarship has been ameliorated by the value Aboriginal people, poets as well as political activists, have recognized in the legitimacy which expert evidence of great antiquity confers on them. Fifty thousand years of undisputed possession has immense symbolic power. Further, the kind of evidence archaeologists produce is valuable when land claims are mounted. Thus, Aborigines have often inverted the connotations of being the most primitive peoples in the world by emphasizing the spiritual wisdom and rights to land conferred by antiquity.

1. *The Reign of Social Anthropology*

Aborigines are now also questioning social anthropology's representations. When Australian anthropology was established at the University of Sydney in 1928, it claimed expertise in managing indigenous populations. As 'culture' became privileged over 'race,' social anthropology displaced prehistory as the pre-eminent science in relation to Aborigines. These 'hunter-gatherers' were defined by their artefacts and rituals, languages, religious beliefs, and kinship systems. Those who had 'retained their culture' were in the northern and central parts of Australia where the invading British had only lately and sparsely occupied the land.

From the 1930s, systematic field workers such as Lloyd Warner, Phyllis Kaberry, Donald Thomson, and T. G. H. Strehlow and in the 1950s and 1960s W. E. H. Stanner, Catherine and Ronald Berndt, Mervyn Meggitt, and Lester Hiatt traveled north to conduct careful studies of remote Aboriginal communities. All assumed that the authentic culture of Aborigines was disappearing. Writings about social change were premised mostly on a dichotomy between 'full-bloods' who were 'traditional' and those who had 'lost the essential elements of that heritage, that living link with their cultural past' (Elkin 1938/1974).

'Their cultural past' was depicted as static by structural-functional social anthropology established by Radcliffe-Brown's seminal study of *The Social Organisation of Australian Tribes* (1931). In fact, these traditions were under extreme pressure from the colonisers. Anthropologists were far behind the frontier where the first and most fundamental social disruption, the removal from land, took place gradually from the late eighteenth to the twentieth century. The invaders' access to the land was assured as Aborigines were variously massacred, removed to reserves and missions, and re/formed through a vast plethora of legislation and regulation which reached its height in the 1950s (Rowley 1970). By the time Australia began to recognize Aboriginal traditional ownership of land in the Northern Territory in the 1970s there was a solid body of ethnographic documentation of Aboriginal traditions in relationship to the social significance of country. Since the High Court (1991) affirmed that native title exists in Australia, research in this field has burgeoned.

2. *'Country' or 'Native Title'*

The seminomadic groups of 'hunter-gatherers' were not understood to be deeply attached to particular areas of land until proto-anthropologists such as Spencer and Gillen (1899) began to document totemic affiliations and dreaming tracks. There emerged detailed ethnographic accounts of the classic Aboriginal traditions, kinship and genealogical ties, clan organization, conception and birth sites, and historical associations which give individuals authority and responsibility in relation to various sites. These are now understood to involve flexible principles rather than fixed rules, and debate over rights and obligations under ancestral law are understood as the stuff of everyday life.

Ceremonies are arranged and authorized by mutually dependent clans and lineages involving large networks of intricately related individuals who negotiate endlessly about their authority. For instance in the Arnhem Land area, the senior 'boss' (djungkayi) and 'owner' (gidjan) share responsibility as manager and performer of the Yappaduruwa ceremony; these responsibilities are reversed for the Gunappipi ceremony (Maddock 1982). Sophisticated negotiations take place constantly as democracy is practiced; no rights are taken for granted but certain underlying principles are called upon to settle disputes in ever changing circumstances (Myers 1991, Sansom 1988).

Such complex, flexible sets of rights and responsibilities to country are incommensurable with the Australian law which typically recognizes particular individuals as exclusive owners of specific, bounded areas. Further, the commercial exploitation which renders land the passive object of human agency is incommensurable with Aboriginal understanding of the productive interdependence between people and land. Sitting in country and walking through country are actions which generate meanings and events as the sentient country responds to people and people respond back (Povinelli 1993). Aboriginal people assert endlessly to government officials that the interactive, ceremonial, and productive relationship with country is necessary for the continuation of social life.

3. *Kinship and Religion*

Aboriginal kinship was long a major focus of anthropologists' research. They recorded the system of moieties, clans and lineages, sections and subsections

which identified people as nodes in a web of intricate affiliations. The meaning and corporate power of the clan and its reproduction over time were the subject of many ethnographic and theoretical studies (Hiatt 1965, Meggitt 1962).

Aboriginal religion has always attracted both scholarly and popular interest. Variously designated as 'the law' and 'the dreaming' (Stanner 1979), the body of beliefs saturates all facets of the classic traditions. Kinship, totemism, rights to land, dance, song and body art, and hunting and gathering practices, are all ways to 'follow up the dreaming.' Each of these terms is but a crude reference to a precisely integrated cosmology and ontology based in concepts that are at odds with Western traditions of thought. It was probably inevitable that attempts by ethnographers to grasp and convey the significant themes of this system with its specific epistemology have tended to represent it as given, rather than as an arena in which its participants are engaged creatively with their own reproduction over time.

Perhaps the contemporary irrelevance of much of the body of ethnography stems from the fact that, while the centrality of kinship and land to social relations among Aborigines was being documented, governments were systematically breaking up family groups, taking half-cast children, removing 'tribal remnants' from their country, and trying to break down the system of exchange partly by instigating a system of rationing (Rowse 1998). This is well known, yet only lately have researchers begun to be examine the social conditions in which the body of ethnography was produced.

4. *Anthropology's Legacy*

In the middle of the twentieth century three comprehensive anthropological textbooks summarized existing knowledge. Elkin's *The Australian Aborigines* was published in 1938/1974, with several later editions and reprints. In 1964, Ronald and Catherine Berndt published *The World of the First Australians*, and this was superseded in 1972 by a shorter, more sophisticated work by Maddock which reverted to Elkin's title, *The Australian Aborigines*. Each of these texts defined the classical Aboriginal traditions as the 'real object' of Australian anthropology. The provision of accurate and profound knowledge of Aboriginal cultures accrued intellectual and moral worth because it would both provide the basis of appropriate policies and lead to sympathetic understanding from other Australians. However, the many specialized ethnographic studies not only failed to protect the fundamental institutions of Aboriginal society, but also failed to make much public impression. Knowledge of even basic elements of traditional Aboriginal cultural life is absent from the intellectual baggage of educated nonindigenous Australians, as has become clear during recent indigenous struggles concerning sacred sites, native title, and colonial history. Also, these studies largely left out what are now the majority of people who are known and who know themselves as Aborigines, those whose social lives were disrupted to the extent that they no longer recognize moieties, practice ceremonies, or speak traditional languages and whose ancestors include white or other immigrant Australians.

5. *'The Aboriginal Race' or 'Traditional Culture'*

In much of the older ethnographic work, Aboriginal cultural characteristics appeared as fixed as those of race, though in fact both racial and cultural characteristics were changing. Though racial classifications and racial explanations were eschewed, a submerged or implied definition of Aborigines as a race was retained. That race was identified with an unsullied tradition known as 'Aboriginal culture' which remained a discrete *a priori* category. Australian anthropology continued to be associated with the study of 'traditional Aborigines' whose definitive characteristics were formed before colonization. Such studies reinforce the popular judgment that only remote 'full-bloods' are real Aborigines. The metaphor of destruction became entrenched, fixing the complex, ongoing events of colonization into a one way process of collapse. No explanation replaced that of race for the social inequality which separated Aborigines from Europeans throughout Australia. The boundaries of racial categories had been smudged, but they were not re-examined or redrawn, and until the 1970s no serious cultural analysis was undertaken of the relationship between the two populations.

The few studies of what were known as 'mixed bloods' or 'detribalized remnants' usually sought 'traditional' features rather than an understanding of the forces that had transformed them. The actual social lives of these groups were judged depressing and deprived, rather than sites where transformations of cultural forms were taking place. While Aboriginal activists in the 1970s acquired the power to give authoritative expression to a counterdiscourse and to modern Aboriginal culture, there is still a heavy burden of inauthenticity associated with those who display no elements or echoes of an autochthonous past.

With the emergence of Aboriginal voices in the public arena, the ethnographic legacy came under widespread and intense though contradictory criticism, both for complicity in the colonial management of Aborigines and for ignoring their plight. In the 1980s, however, ethnographies began to branch out in several directions. The politics of representation is now firmly on the academic agenda, with articles and

books deconstructing the 'construction of Aboriginality' (Beckett 1988) and an insistence that the study of 'Us' is part of the study of 'Them' (Collmann 1988). Detailed cultural analysis of colonial relations in Australia has showed the complex, creative responses to the distinctive forms of governance Aborigines were subjected to (Morris 1989). Anthropological writings now vie with cultural studies, history, sociology, and law faculties for authority in this domain. Aboriginal intellectuals and academics provide powerful and varied challenges to the writings of the old 'experts.' Some social scientists fear the inevitable political struggles that constitute the realm of Aboriginal studies. However, as the intellectual and ethnographic boundaries have been breached, different kinds of research, such as ethnography on the racial borderlands, can flower (Cowlishaw and Morris 1997).

While critiques of anthropology's essentialism and primitivism have become popular, the moral authority of the old ethnographies has also been revived, as they provide evidence concerning native title claims, both in those few areas where Aborigines have unbroken connections with their traditional land and for groups who were displaced in an earlier era.

6. *Cultural Conundrums*

Aboriginality today is varied and complex. There are fringe dwellers and intellectuals, remote speakers of Aboriginal languages and urban sophisticates, jail birds and film makers, politicians and lawyers. This 2 percent of Australia's population (386,000 people) is the focus of much dispute and debate. A dominant image of distress is supported by statistical evidence that Aborigines are vastly over-represented among those who die young or are in jail.

The Australian constitution of 1901 decreed that Aborigines, meaning those known as 'full-bloods,' would not be counted in the population census, and gave the federal government no jurisdiction over Aborigines residing in the various states. In 1967, a referendum removed these two items from the constitution. The official definition of an Aboriginal person now has three elements, descent, self-identification, and acceptance by an Aboriginal community. But two centuries of categorization as 'full-bloods,' 'mixed bloods,' 'half-castes,' and 'quadroons' have left a complicating racial legacy, for instance, for those who, by force or under legal or social pressure, were removed from family and community, and ceased to identify as Aboriginal. Now individuals can be denigrated for belatedly claiming inclusion in the category 'Aboriginal' when not sufficiently dark, or for trying to remain outside it when a shade too dark. In remote rural communities family histories are well known, but for an increasingly urban, anonymous population the necessity to have one's Aboriginality authorized by a community can be a source of conflict. Another source of conflict is created by the recognition of native title and official protection of sacred or significant Aboriginal sites.

7. *Contrasting Epistemologies*

The problem of legitimizing Aboriginal culture was typified in the events surrounding one much publicized disputed claim for the protection of an Aboriginal sacred site (Bell 1998). Aborigines and anthropologists were accused of conniving in fabricating cultural material. 'Secret women's business' became a widely used term of derision. The dispute itself and the publicity it received revealed the threat posed by contrasting epistemologies. In the classic Aboriginal tradition, knowledge does not circulate freely. Elements of cosmology are the responsibility of particular people (men or women, certain clans, those related to particular places). That is, knowledge is both localized and 'owned.' But if knowledge is secret, how can indigenous people prove their claims, and how can sites be recognized and protected by the state? In fact, a series of answers to this problem have emerged. Land tribunals have accepted the performance of a dance, a song, or a design as supporting a claim and have agreed to restricting the circulation of evidence. But in some cases the insensitivity of authorities to secrecy and sacredness has led to the abandoning of claims (McRae et al. 1997).

Attempts to legitimize the knowledge of dispossessed Aborigines have given rise to sophisticated arguments about the legitimacy of cultural transformation and the interweaving of historical experience with cultural identities. There is a bitter irony here, for it is commonly the fixed tradition, the given culture, the ancestor's true knowledge that is appealed to by Aborigines as the basis of claim. Scholars who want to assert the fictional basis of all culture can appear to undermine those claiming authority over their own cultural knowledge and history. The notion of culture has powerful public meanings. The policy of indigenous self-determination has registered 'Aboriginal culture' with formal respect, but has given it no systematic place in the state funded structure of organizations or enterprises. 'Culture' is the icing on the cake of 'social' matters, with 'viable projects' to be run by Aborigines. The myth of a fragile Aboriginal tradition has been replaced with its opposite; now Aborigines can become part of capitalist culture without tarnishing their own cultural identity which is rendered as an indigenous version of elite culture, with ritual and ceremony severed from social conditions and everyday meanings.

Aboriginal culture is also called upon to play a redemptive role for the white nation. Its symbols are used routinely to construct a unique imagery for

Australia. The fear of being complicit in this appropriation and the constricting traditionalism leads to ambivalence about, for instance, hanging bark paintings on the wall. Continuing controversy over the status of modern Aboriginal art (Michaels 1994), and continuing celebrations of the primitive, confuse the interpretations and the intentions of scholarship in this field.

8. History

Historians began, in the 1970s, to document the horrific atrocities and the accompanying savage rhetoric of British settlers in the eighteenth and nineteenth centuries. Some conservative scholars and politicians deplore the emergence of what they call the 'black armband' view of history, making Australians ashamed of their forebears. But the opposite problem is that the racism of the past is not seen to be an organic and ongoing part of colonialism. As Australian readers recoil in horror from the bloody history, the obsession with color and race is attributed to some deranged, long-gone strangers rather than to our own grandparents. But these violent and racist people certainly left us something, if not the land they took or the wealth they made from it, then the culture they were developing. As we position ourselves against the invaders, we obscure the way a racialized society has been reproduced over time and the cultural source of contemporary events is lost. Where before anthropologists' writings offered no access to the contemporary experiences of the timeless Aborigines, now historical discourse immerses readers in an awful past and distracts them from contemporary forms of violence and racism.

The destructive effects of well intentioned government policies are now receiving attention. From the early twentieth century successive governments have attempted to ameliorate the conditions of Aborigines through varied policies of protection, assimilation, integration, and self-determination. In each case there have been assimilative forces which have led to a degree of empowerment of a minority of people, and communities have had to adjust to new relationships with the state's institutions (Cowlishaw 1999).

See also: Aboriginal Rights; Australian Settler Society: Sociocultural Aspects; Clan; Colonialism, Anthropology of; Ethnic Cleansing, History of; Ethnic Conflict, Geography of; Hunter–Gatherer Societies, Archaeology of; Hunting and Gathering Societies in Anthropology; Minorities; Tribe

Bibliography

Beckett J (ed.) 1988 *Past and Present: The Construction of Aboriginality*. Aboriginal Studies Press, Canberra, Australia

Bell D 1998 *Ngarrindjeri Wurruwarrin*. Spinifex, Melbourne, Australia

Berndt R, Berndt C 1964 *The World of the First Austrlians*. Ure Smith, Sydney

Collmann J 1988 *Fringe-Dwellers and Welfare*. University of Queensland Press, St Lucia, Australia

Cowlishaw G 1999 *Red-necks, Egg-heads and Blackfellas*. University of Michigan Press, Ann Arbor, MI

Cowlishaw G, Morris B (eds.) 1997 *Race Matters*. Aboriginal Studies Press, Canberra, Australia

Elkin A P 1938/1974 *The Australian Aborigines*. Angus and Robertson, Sydney, Australia

Fison L, Howitt A W 1880/1991 *Kamilaroi and Kurnai*. Australian institute of Aboriginal and Torres Strait Islander Studies Press, Canberra, Australia

Hiatt L 1965 *Kinship and Conflict*. Australian National University, Canberra, Australia

Maddock K 1972 The *Australian Aborigines*. Penguin, Ringwood, Australia

McRae H, Nettheim G, Beacroft L (eds.) 1997 *Indigenous Legal Issues: Commentary and Materials*. LBC Information Service, Sydney, Australia

Meggitt M 1962 *Desert People*. Angus and Robertson, Sydney, Australia

Michaels E 1994 *Bad Aboriginal Art*. University of Minesota Press, Minneapolis, MN

Morris B 1989 *Domesticating Resistance*. Berg, Oxford, UK

Myers F 1991 *Pintupi Country, Pintupi Self*. University of California Press, Berkeley, CA

Povinelli E 1993 *Labor's Lot*. University of Chicago Press, Chicago

Radcliffe-Brown A R 1931 *The Social Organisation of Australian Tribes*. Macmillan, Melbourne, Australia

Rowley C D 1970 *The Destruction of Aboriginal Society*. Penguin, Ringwood, Australia

Rowse T 1998 *White Flour, White Power*. Cambridge University Press, Cambridge, UK

Sansom B 1988 A grammar of exchange. In: Keen I (ed.) *Being Black*. Aboriginal Studies Press, Canberra, Australia

Spencer B, Gillen F 1899 *The Native Tribes of Central Australia*. Macmillan, London

Stanner W E H 1979 *White Man Got No Dreaming*. Australian National University Press, Canberra, Australia

G. Cowlishaw

Australian Settler Society: Sociocultural Aspects

Australia ranks among the more affluent and stable of Western societies. It has also experienced major changes since the 1970s. The more significant material and cultural transformations have been accompanied by a marked uncertainty over the country's cultural standing and its collective identity.

1. Prosperity and Paradox

Two properties of contemporary Australia distinguish it from other nation-states with a comparable pattern of settlement. One is that a clear majority shares a high standard of living, a predominantly middle-class life

style. Australians enjoy one of the highest rates of private property ownership in the Western world: household ownership of a second or third car is commonplace: a wide range of costly consumer goods is customary; even the possession of a second dwelling for vacation purposes is nothing untoward. There is a strong emphasis too on costly hedonistic activities, notably multiple forms of gambling.

The other distinguishing feature is political stability. Unlike South Africa or Israel, for example, the political process is strikingly free from outright conflict. Constitutional crises are rare, elections are routine, and changes in government are uncomplicated. This it not to suggest the political arena is free from major lines of cleavage, nor that public confrontations are unknown: the divide between Aboriginal and non-Aboriginal interests is profound and open dispute is recurrent. But the absorption of major conflict is undoubtedly among the more pronounced qualities of the political system.

Yet despite affluence and order, a distinct disquiet prevails over Australia's cultural and national identity. It repeatedly surfaces in debate about how 'We' differ from significant 'Others,' whether 'Our' culture is the equal of 'Theirs.' Australia's most renowned author writes: 'Identity can be experienced in two ways. Either as a confident being-in-the-world or as anxiety about our-place-in-the-world; as something we live for ourselves, or as something that demands for its confirmation the approval of others.' Malouf (1998) argues that the latter condition prevails in Australia, so he suggests: 'Perhaps it is time to stop asking what our Asian neighbours will think of us, or the Americans, or the British, and try living free of all the watchers but our own better and freer and more adventurous selves.'

This, then, is the paradox at the core of Australian society: a prosperity and stability that are the envy of others, coupled with a profound concern about cultural worth and collective identity. The roots of the paradox lie in recent economic and political developments.

2. *Rural Change*

Australia's small, peripheral economy (the population is 18.5 million) has long been integrated into the global marketplace, mostly to distinct economic advantage. Throughout the first half of the twentieth century, the export of agricultural produce and minerals to Britain and mainland Europe generated a prosperity from which only particular minorities were categorically excluded.

This affluence was sustained after World War II as new international markets were added to traditional ones. By the early 1970s, however, this export dependent economy proved vulnerable to changed global conditions (Bell 1997), and the major impact fell on the small-to-moderate sized family farms. Declining commodity prices, increased interest rates, price hikes in key inputs, and rising transport costs, impacted most severely on the bottom 25 per cent of the rural hierarchy; and their problems have been compounded by the loss of wage-earning opportunities, as well as cutbacks to the infrastructure and welfare services. The result has been the loss of young people to the cities, the annual closure of between 2,000 and 3,000 production units, and increased dependence on minimal welfare by the already marginal, along with a deepening of the divide between prosperous and poor.

Lawrence (1987) has made it clear that these changes to the rural class structure are inseparable from transnational corporations' increased hegemony over agriculture, especially through control over the supply of critical inputs to the rural sector (mechanical equipment, seeds, building materials) as well as its major outputs (canning, packaging, shipping). The larger units of rural production that are contractually integrated into transnational structures realize higher rates of profit than ever before, but smaller ones find it increasingly difficult to survive.

In these circumstances, the independence of farm family producers, so integral in the past to the culture of rural Australia, becomes illusory. But the significance of these material changes is more wide-ranging. The bush has always provided the more powerful symbols from which cultural specialists forged narratives about the distinctiveness of the Australian character (Ward 1966). Self-reliance, egalitarianism, and mateship are all idealizations drawn from outback experience, and all have been part of distinguishing Australian culture and the nation from their European origins. Where the realities of rural experience are characterized by increasing inequalities of class, destruction of community life, and transnational domination, the bush inevitably loses its symbolic worth in the process of constructing Australia as a distinct cultural space.

3. *Urban Transformation*

These changes have also compounded rural–urban differentials: by the 1990s, the average urban income was half as much again as the rural one. But it has been class inequalities that have become most conspicuous in the major metropolises, where well over two-thirds of Australians reside. The most powerful members of the country's dominant class are from the transnational capitalist class: their major allies are drawn from the Anglo-Celtic power bloc of business, political, and media institutions that are all similarly wedded to a narrow ideology of late capitalist development. The life style of this class is cosmopolitan and ostentatious; since it attracts extensive media coverage, it contrasts all the more objectionably with the circumstances of urban residents who are 'doing it

hard,' especially the one-third of the population reliant on some form of welfare. The long-hallowed ethos of Australia as an open, 'fair go' society, rooted in individual achievement rather than ascribed status, thus has increasingly limited relevance.

It is, however, changes in ethnic composition that have generated the most telling questions as to how Australians imagine themselves to be as a nation. With the exception of Israel, no other modern society has relied so heavily on immigration to build up its settled population. Between 1947 (when the population stood at 7.5 million) and 1987, some 4 million immigrants entered the country. Since the UK and Ireland were no longer able to meet Australia's demand for labor, substantial numbers were recruited from southern Europe and the Levantine by the 1960s. Regardless of place of origin, these new immigrants were offered citizenship to encourage permanent residence.

The result was an exponential increase in cultural heterogeneity. Inner city areas were transformed from predominantly Anglo-Celtic working-class neighborhoods into multiethnic lower-class suburbs. These were not destined to become ethnic ghettoes, however, because a pattern of modest clustering was sufficient to establish residential stability, gain employment, and bring other family members from Italy, Greece, Yugoslavia, and elsewhere. Then again, as these migrants accumulated capital and became small business owners, they moved into more prestigious suburbs. Finally, second and third generations have displayed even less inclination to being restricted by ethnic considerations in their choice of spouses, personal friendships, and business associates (Price 1988).

It would be more than appropriate to utilize such concepts as creolization and hybridization (Hannerz 1996) to detail the innovative cultural relationships and associations that have so transformed Australia's urban areas. The monochromatic mould of Anglo-Celtic dominance is, at this level, a thing of the past. But not only have urban social analysts generally opted for more conventional analytic categories, successive governments have worked hard to obscure the rich character of these creolizing, hybrid processes, with a stultifying rhetoric of multiculturalism.

4. *Multiculturalism*

Multiculturalism was adopted as state policy in the early 1970s, by which time it was more than evident that immigrants from southern Europe had no intention of assimilating into the Anglo-Celtic ethnic majority as the previous policy had (unsuccessfully) dictated (Jakubowicz 1984). The following decade saw the establishment of numerous governmental organizations whose function was to implement different facets of multiculturalism, even though what was meant by the term proved difficult to establish (Jupp 1991). Two considerations, however, were returned to time and again in policy statements. First, multiculturalism entailed unqualified respect for those cultures imported into Australia by successive immigrant waves. Second, certain core values had to be honoured by all in order that the stability of society could be maintained.

Notwithstanding the liberal rhetoric of tolerance and coexistence woven around the ideal of a multicultural society, this proved not only a remorselessly conservative set of ideas but also a discourse of control from which many came to feel alienated. For in political practice, multiculturalism entailed the privileging of the Anglo-Celtic core culture already enshrined in the dominant institutions of Australian society, and this monocultural core was to be surrounded by the diverse range of ethnic cultures to which 'Non-English-Speaking-Migrants' were assumed to be still committed. The Anglo-Celtic core was not acknowledged as the preserve of a particular ethnic category; it was represented instead as a naturalized Australian culture. Those not fortunate enough to have been born into this cultural mainstream could now couple it with the culture which was their inheritance—and thus become 'Greek–Australian,' 'Chinese–Australian,' and so on.

It was this arbitrary privileging of Anglo-Celtic culture that alienated many people. Multiculturalism proved to be a state policy that discriminated against cultural difference rather than celebrating it. More problematic still, the policy subsequently endorsed a new, emergent form of racism which also equated an Anglo-Celtic heritage with the supposedly national culture, while stridently claiming also that other particular cultures (those of Asian immigrants, for example) were quite incompatible with the latter (Stratton 1998). All things considered, it was hardly surprising that for many immigrants multicultural policy raised contentious questions about how those in power evaluated their position and their worth in Australian society. Notwithstanding their formal citizenship, at least in cultural terms they were being considered less than full and equal members of a rapidly evolving and increasingly complex nation-state.

5. *Globalization*

The final development that transformed the social and political landscape has been the globalization of culture, especially in its popular mass forms. The expansion of a middle-class prosperity and life style, including extended leisure time, has coincided with the ready availability of popular cultural commodities, and at least the superficial impact has been wide-ranging. Especially among the young, but by no means exclusive to them, the appeal of American fashion is pronounced: the major US corporations dominate the lucrative fast food market; and films, videos, CDs, and other forms of entertainment are predominantly im-

ported, along with spectacular mega-events. As with the exceptional growth of tourism into Australia from south east and east Asia, not only has the expansion of popular culture compounded levels of economic growth; it too has been constitutive of newly hybrid, cultural developments. Imported commodities spawn not only imitations but also equally lucrative innovations. There is also a wide-ranging reverse flow, including the export of Aboriginal art, soap operas, and films.

In the judgment of some social analysts, however, the impact has been much more than superficial: mass produced commodities from the US are considered a substantial threat to the already fragile and uncertain status of Australian culture (see Bell and Bell 1998). There are always difficult questions to be asked about the multiple meanings which consumers glean from, or invest in, these predominantly ephemeral cultural forms. But whether these justify such newspaper headlines as 'Australia—the 51st State' and turns of phrase like 'the Trojan horse of American culture' remains a moot point, not least because it is rarely made clear which aspects of Australian culture are most at hazard, and why this culture is considered especially fragile and threatened in the first place. At the very least, analysis of the impact of global culture needs to be coupled with, rather than separated from, exploration of the national fervour so conspicuously generated by sporting achievements.

6. *Conclusion*

In sum, the paradox of prosperity and stability on the one hand, and concern about cultural worth and collective identity on the other, is in substantial part indicative of the varied ways in which Australia is articulated within the global system. Increased affluence and social order have been sustained by such processes as the continued export of agricultural and mineral commodities, the recruitment of skilled and educated immigrant labor, and engagement in the extensive (albeit uneven) exchange of global culture in its multiple forms, as well as a general commitment to important principles of conduct such as the search for consensus through sustained negotiation.

But the course of rapid and fragmenting social change has seen an undermining of the potency of key symbols, the appeal of significant icons, and the interpretive power of unifying narratives, all of which were indispensable sources of meaning and comprehension in more stable, less complex times. This is not to suggest that other cultural constructs have not taken their place, only that—as is more widely the case in the postmodern era—these do not have, and most likely cannot take on, the broadly-based appeal and influence which could be effected in the past.

A small but significant illustration must suffice. In the final days of the twentieth century it was determined by the political elite that the Australian constitution demanded a preamble. In its draft form, co-authored by the Prime Minister and the country's leading poet, the term 'mateship' was used in order to capture some quintessential quality of how the Australian people relate to one another. For much of the twentieth century this claim would have been beyond contention. Yet no sooner did the draft document enter the public arena than the inclusion of 'mateship' was subject to derision, derogatory comment, and dismissal from many different quarters. It was an ignominious incident, yet—again paradoxically—it had a wealth of meaning inscribed within it.

See also: Aboriginal Rights; Australian Aborigines: Sociocultural Aspects; Colonialism, Anthropology of; Colonization and Colonialism, History of; Creolization: Sociocultural Aspects; Globalization and World Culture; Hybridity; Multiculturalism, Anthropology of; Multiculturalism: Sociological Aspects; Urban Growth Models; Urban Sociology

Bibliography

Bell P, Bell R J (eds.) 1998 *The Americanization of Australia*. University of New South Wales Press, Sydney, Australia

Bell S 1997 *Ungoverning the Economy: The Political Economy of Australian Economic Policy*. Oxford University Press, Melbourne, Australia

Hannerz U 1996 *Transnational Connections. Culture, People, Places*. Routledge, London

Jakubowicz A 1984 State and ethnicity: Multiculturalism as ideology. In: Jupp J (ed.) *Ethnic Politics in Australia*. Allen & Unwin, Sydney, Australia

Jupp J 1984 Power in ethnic Australia. In: Jupp J (ed.) *Ethnic Politics in Australia*. Allen & Unwin, Sydney, Australia

Jupp J 1991 Multicultural public policy. In: Price C A (ed.) *Australian National Identity*. Academy of Social Sciences in Australia, Canberra, Australia

Lawrence G 1987 *Capitalism and the Countryside: The Rural Crisis in Australia*. Pluto Press, Sydney, Australia

Malouf D 1998 *A Spirit of Play. The Making of Australian Consciousness*. ABC Books, Sydney, Australia

Price C 1988 The ethnic character of the Australian population. In: Jupp J (ed.) *The Australian People: An Encyclopedia of the Nation, its People and their Origins*. Angus & Robertson, North Ryde, NSW, Australia

Stratton J 1998 *Race Daze. Australia in Identity Crisis*. Pluto Press, Sydney, Australia

Ward R 1966 *The Australian Legend*, 2nd edn. Oxford University Press, Melbourne, Australia

A. Peace

Authenticity, Anthropology of

Within anthropology, there are two principal approaches to the question of authenticity. The first presupposes that authenticity is an attribute or prop-

erty inherent in cultural objects, and that the role of anthropology is to determine, using scientifically appropriate evidence and methods, whether particular objects are or are not authentic. The second locates authenticity not in objects but in discourses about them. This approach holds that the authenticity of phenomena is never a matter of objectively determinable fact, but is always socially constructed and, as such, subject to change.

Presently, the second approach is in the ascendance. It would, however, be short-sighted to view this trend as the culmination of a unilineal historical development in which the second approach has eclipsed the first. Rather, these opposed approaches endure because they replicate a fundamental epistemological opposition within modern societies, that between objectivism, realism or positivism, on the one hand, and constructionism or hermeneutics, on the other. Despite the recent persuasiveness of constructionism in social science, objectivist notions of authenticity remain hegemonic in many late-capitalist institutions, such as the art market, museums and courts of law.

1. *Authenticity and Reality*

Questions about authenticity should be distinguished from questions about reality and its representation. To ask whether an object is real, or an event occurred, is to ask about its existence relative to the empirical world. To declare an object or event 'unreal' or 'illusory' is to claim that it never existed or occurred. Note that such 'unreal' phenomena can exist in imagination. At least since Durkheim, anthropologists have argued that representations of objects are real-world phenomena with real-world effects, whether or not the objects of such representations can be shown (according to some sets of criteria) to exist in the world beyond those representations.

In contrast, to ask whether an object is authentic is to ask not about its existence but about its relationship to a posited social or cultural identity. The opposite of 'authentic' is not 'unreal' but 'inauthentic.' To deem an object inauthentic is to assert that it is not, despite claims to the contrary, an example of an identified class of objects, or not the creation of an identified person or group.

It is easy to confuse questions of reality and those of authenticity due to the commonsense use of the word 'real' to mean 'authentic,' as in the phrase 'this is a real Native American medicine bundle.' Such claims are not intended as predications of existence as such, but as evaluations of what is understood to be the essential nature, or the identity, of the phenomenon in question. Thus, assertions of authenticity always have embedded within them assertions of identity.

Throughout the history of their discipline, anthropologists have been asked to authenticate artifacts, customs, legends, texts, and other cultural objects, as well as the racial, ethnic, or cultural affiliation of persons. Depending on subdisciplinary and areal expertise, anthropologists might offer authoritative answers to such questions as: 'Are these stone weapons Clovis points, from Western North America and 12,000 years old, or are they modern reproductions?' 'Can this narrative be attributed to precolonial Native Americans, is it the result of European influence on Native American traditions, or is it a recent creation, and if the latter, what is the cultural identity of its author?' And most bluntly: 'Are these people Indians with unbroken ties to precolonial America, or have generations of racial mixing and cultural borrowing erased that identity?' Such assessments of the authenticity of particular objects and persons are similar to those required in fields like art history and museology, where it is often important to be able to identify objects by attributing them to named artists, historical periods, or cultural groups.

2. *Authenticity and Modern Individualism*

The commonsense notion that authenticity is an expression of the essential identity of an object stems from modern (in the sense of post-fifteenth-century) individualism. In this worldview, ultimate reality is understood to be lodged, not in a hierarchically interrelated divine cosmos, but in individuals, from the smallest particles to persons to delineated groups of persons (cultures). To quote the philosopher Cassirer on the modern ontology of nature: 'Nature … implies the individuality, the independence and particularity of objects. And from this characteristic force, which radiates from every object as a special center of activity, is derived also the inalienable worth which belongs to it in the totality of being … The part not only exists within the whole but asserts itself against it, constituting a specific element of individuality and necessity' (1965, p. 41). From such an outlook the notion of authenticity derives: each object, each individuality is itself a center of reality with a unique essence or identity that distinguishes it from all other objects.

2.1 *Authenticity and Sincerity*

In *Sincerity and Authenticity*, Trilling (1972) drew out the implications of this worldview for modern conceptions of sociality and personhood. With the emergence of modern individualism, persons were no longer defined by their position in a social hierarchy (as in medieval European realms). Instead, social positions came to be understood as 'roles' to be filled by persons whose essence, or individuality, precedes and transcends those roles. In this perspective, the inner realm

of individual character or personality is more real than the outer world of social relations.

The importance of sincerity in the early modern period, Trilling argued, derived from concerns about whether a person's social actions matched or masked inner attitudes. Indeed, it was insincerity that was at issue. As Rousseau put it, humans in the 'state of nature,' prelingual and presocial 'savages,' had no means to dissemble; their actions could be nothing other than the direct expression of their inner or essential being. But with the emergence of human language and society, 'it ... became the interest of men to appear what they really were not' (1972, p. 86). Inner authenticity diverged from outer sociality, and the 'rise' of civilization, as Rousseau saw it, entailed a process of moral corruption.

In Trilling's account, concern for sincerity, with its questioning of the relationship between society and the individual, began to cede place in the Romantic period to concern for authenticity, which privileges inner essences without relating them to an outer world of other people. The Romantics' quest for authenticity was a quest for 'essential identity,' a longing for knowledge of true inner being irrespective of any accidents of social history or circumstance. Artists, in particular, were understood to be individuals possessed of powerfully unique inner realities, uncorrupted and undiminished by social conventions. At the same time, the growing middle classes of the new mass democracies were coming to see their lives or identities as increasingly 'unreal' or 'inauthentic.'

2.2 *Authenticity and Collectivities*

In social philosophy, and later the social sciences, these concerns for individual realities were applied to human collectivities. The leap from the individual to the collective is not as great as it seems, for, as Dumont (1970) has insisted, modern ideology understands groups as 'collective individuals.' From the eighteenth century onward, folklorists and anthropologists assumed that groups of people could be characterized by unique cultural, historical, religious, linguistic or otherwise traditional features that set them off from all other groups. These models of the group were often naturalistic, using analogies to biological organisms or species: it was assumed that a group was a bounded natural unit or object, that members of the group shared its defining features, and that it was possible objectively to evaluate the cultural identities of individual persons, based on the cultural attributes they 'possessed.'

2.3 *Authenticity Endangered*

A noteworthy accompaniment of this conception of cultural authenticity is a fear of its demise. Authenticity is seen to be perpetually threatened. The authenticity of nonmodern ('savage,' 'primitive,' 'native,' 'indigenous,' 'folk,' 'peasant,' or 'traditional') cultures is imagined to depend on their pristine existence beyond the reach of advancing 'civilization.' Throughout the epoch of European colonial expansion, down to the present day, metropolitan authorities—scholars, travelers, missionaries, colonial officials, collectors—took it upon themselves to delineate unique groups in terms of their 'possession' of specific racial, linguistic and cultural attributes. The impingement of European culture on such groups, or even the influence of non-European but purportedly more 'complex' cultures, was thought to alter their unique identities and thus to destroy their cultural authenticity. Such conceptions were applied at the level of the individual as well, such that in colonial social orders, the 'half-breed' (to use a racial term) or the culturally bastardized colonial subject was often scorned as the embodiment of the worst of both the 'primitive' and 'civilized' worlds.

In sum, the European 'discovery' of 'Others' followed by ongoing 'contact' made them, almost by definition, inauthentic. Authenticity as a quality of cultures, artifacts and persons was projected by metropolitan authorities into the past, a past they might be able to document, but only just before its destruction.

3. *Constructionist Accounts of Authenticity*

Despite the pervasiveness of such assumptions about groups as bounded objects, there have been recurring social-scientific trends (historical diffusionism, for example) that privileged intergroup connections, contacts, borrowings, and syncretisms. In Anglo-American anthropology since the 1970s, there has been growing rejection of the notion that cultures, societies, and groups in general are naturally bounded and internally homogeneous. Indeed, it is now commonplace to assert the opposite, and to claim that groups are historically and situationally contingent; that the relative rigidity (or porousness) of group boundaries is a function of semiotic, not natural, factors; that collective identities salient in contemporary institutions, politics and scholarship ought not to be projected into the past; that individual persons have multiple identities that can shift according to social context, although not without regard to institutional and cultural constraints; and that the products of human creativity cannot unproblematically be assigned cultural, racial or ethnic identities. In addition, recent historical scholarship has argued that the work of colonial cultural classification was carried out in areas of the world where European notions of objective social boundaries encompassing homogeneous, authentic cultural identities made little sense in relation to local people's understanding of sociality.

3.1 Negotiating Authenticity

Constructionist approaches to authenticity question whether it's useful to specify the 'essential' identity of groups, artifacts and persons. Since human beings are historically and socially situated, they are always connected to other people. Culture exists not simply *in* them but *between* them. Any understanding of a particular person's or group's authentic identity will be a function of specific social encounters, and particular evaluations of authenticity are open to negotiation and dispute. An 'expert' (colonial administrator, art collector, anthropologist) may claim that a person is not a 'real' or 'pure' member of a posited culture, or that an artifact is not an authentic product of it, but others may reject such evaluations based on differing definitions of cultural boundaries and essences.

This does not mean it is impossible or unhelpful to describe and classify cultures. Rather, it means that the terms used to do so must be understood in relation to the discursive project of the classifiers. At any historical moment, some such terms will be relatively unproblematic. For example, given our present technology and understanding of time, we can expect the carbon 14 dating of archaeological artifacts to yield valid information. We need not question the empirical validity of the description to recognize that such data become useful, or meaningful, only in terms of more general hypotheses concerning culture history, hypotheses that can never be reduced merely to facts derived from carbon 14 dating. Moreover, terms of cultural classification that are unproblematic in one context might become contested in another, for example, present-day descendants of precolonial Native Americans might dispute carbon 14 results, claiming a greater antiquity for artifacts than anthropological methods can demonstrate.

3.2 The Politics of Authenticity

Assertions of cultural authenticity can have political consequences. Among the most serious are governmental recognition of 'tribal' or 'aboriginal' status, a status that often entitles group members to government subsidies and services, and to government (mis)management. In settler societies such as the US, Canada, and Australia, some 'indigenous' groups have governmental recognition based on treaties, some of which date to the late eighteenth century; other groups claiming to be authentically indigenous may turn to the courts to gain such recognition. In the courts, authenticity is defined in terms of temporal continuity, cultural homogeneity and distinctiveness, and geographical stability. It may or may not be possible for people who have lived for generations in a situation of 'contact' to demonstrate possession of cultural authenticity defined in such terms. Indeed, there is a cruel irony inherent in the relationship of 'native' authenticity to nation-state citizenship. On the one hand, 'indigenous' peoples are asked to assimilate, through education and economic mobility, into modern society; on the other hand, to do so is to be seen as having become inauthentic and to be scorned on that basis. In practice, both racist barriers and modern fantasies of authenticity fence people in, making it difficult for them to enjoy full citizenship.

Claims concerning the authenticity of individuals and objects can also have political (and economic) consequences. Government criteria to evaluate the affiliation of individuals to recognized tribal groups are often based on quantifications of 'blood,' that is, proportion of 'native' descent. Such criteria usually do not match those that defined social inclusion before the imposition of modern definitions of authentic personhood. Individuals' access to government benefits nonetheless depends upon satisfying the imposed criteria.

The authenticity of cultural objects is negotiated in the markets for art, museum artifacts, and touristic souvenirs. Objects deemed culturally authentic command higher prices and can claim space in more prestigious venues, such as museums and the homes of wealthy collectors, than those considered culturally 'mixed' or 'commercial.' Over time, however, objects can gain authenticity as scholars and collectors objectify and historicize what were once seen as debased cultural forms. 'Indigenous' artists and craftworkers, too, can influence standards of authenticity as they position themselves vis-à-vis the art market and museum world.

3.3 Anthropologists and Authenticity

Some anthropologists continue to approach authenticity as an objectively verifiable property of cultural objects. Others admit the validity of constructionist arguments, but worry that they undermine the claims of marginalized people to the possession of authentic culture and the accompanying political benefits. Although these anthropologists may reject objectivism in principle, they want to be able to deploy an objectivist notion of authenticity 'strategically' to support politically embattled minorities. Other anthropologists argue that recourse to concepts of authenticity inevitably reproduces the hegemonic ideology of identities that led to racial, ethnic and cultural minoritization in the first place. The fact that anthropological expertise based on any of these positions can influence the social status of objects and people is if nothing else evidence that supports a constructionist account of authenticity.

See also: Aboriginal Rights; Collective Memory, Anthropology of; Constructivism/Constructionism: Methodology; Discourse, Anthropology of

Bibliography

Adorno T W 1964 *Jargon der Eigentlichkeit: Zur deutschen Ideologie*. Suhrkamp Verlag, Frankfurt am Main, Germany (1973 *The Jargon of Authenticity*. Northwestern University Press, Evanston, IL)

Bendix R 1997 *In Search of Authenticity: The Formation of Folklore Studies*. University of Wisconsin Press, Madison, WI

Bruner E M 1994 Abraham Lincoln as authentic reproduction: A critique of postmodernism. *American Anthropologist* **96**: 397–415

Cassirer E 1965 *The Philosophy of the Enlightenment*. Beacon Press, Boston

Clifford J 1988 *The Predicament of Culture: Twentieth-Century Ethnography, Literature, and Art*. Harvard University Press, Cambridge, MA

Dumont L 1970 Religion, politics, and society in the individualistic universe. *Proceedings of the Royal Anthropological Institute* 31–45

Errington S 1998 *The Death of Authentic Primitive Art and Other Tales of Progress*. University of California Press, Berkeley, CA

Gable E, Handler R 1996 After authenticity at an American heritage site. *American Anthropologist* **98**: 568–78

Orvell M 1989 *The Real Thing: Imitation and Authenticity in American Culture, 1880–1940*. University of North Carolina Press, Chapel Hill, NC

Rousseau J-J 1973 A discourse on the origin of inequality. In: Cole G D H (ed.) *The Social Contract and Discourses*. Dent, London

Segal D A 1999 Can you tell a Jew when you see one? Or thoughts on meeting Barbra/Barbie at the museum—teaching Jewish studies. *Judaism* **48**: 234–41

Trilling L 1972 *Sincerity and Authenticity*. Harvard University Press, Cambridge, MA

R. Handler

Authoritarian Personality: History of the Concept

Around 1930 Wilhelm Reich and Erich Fromm, two Marxian psychoanalysts, developed similar psychodynamic models to account for the attraction of Germans to Nazism. Subsequent development by the exiled Frankfurt Institute linked up with empirical research on antisemitism by psychologists at Berkeley. Adorno et al. (1950) published *The Authoritarian Personality*. It presented a three-level theory: an ethnocentric, power-oriented surface ideology, undergirded by a conflicted personality structure, in turn generated by authoritarian child-rearing practices. Highly praised, the tome spawned hundreds of research reports in America and, later, abroad. But gradually the theory succumbed to methodological criticism, spotty empirical confirmation, and a changed political climate. Recent attempts to revive the theory had little impact. After promising an understanding of major historical-political developments, the idea's empirical applications failed to convince. Only the pejorative label 'authoritarian' appears to have survived.

1. *Emerging Conceptions of the Authoritarian Character*

1.1 *Reich's* Massenpsychologie des Faschismus

In the years after World War One, the core problem for leftist politicians and intellectuals in Europe was the failure of the proletarian revolution in the industrial countries. This apparent failure of Marx's theory preoccupied the young Wilhelm Reich, a Marxist activist as well as rising star of Freud's psychoanalytic circle. Combining the two revolutionary doctrines, and changing focus from capitalism to patriarchy, he diagnosed sexual repression by society as the main cause of political passivity. He called this, by 1930, '*autoritäre Unterordnung*' (authoritarian submission), personified in the *Feldwebelnatur* (sergeant-major nature), 'groveling upwards and bullying downwards.' Discussing Nazi psychology and its race theory, he proclaimed that German Fascism would be the inevitable outcome of the patriarchal-authoritarian family system. Without the label '*Autoritärer Charakter*,' Reich had assembled most elements of Authoritarianism theory by 1932.

Reich's lack of respect for orthodoxy landed him in trouble with both Communists and psychoanalysts. Soon the Nazis' rise to power, confirming his prediction, forced him to emigrate and eventually settle in America. He also elaborated his ideas of the appeals of National Socialism in his *Massenpsychologie des Faschismus* (1933). Discounting Hitler's charisma and propaganda as shallow explanations, he found the roots of the catastrophe in the character structure of lower-middle class Germans. In patriarchal society the family, as 'factory of bourgeois ideologies,' molded the child's character by implanting sexual inhibition and fear. Sexual repression and economic exploitation produced conservatism and not only submission but active support for the authoritarian order and its propaganda. The repressed sexuality turned into yearnings for mystical ideal: nation, honor, motherhood, symbols exploited eagerly by Nazi propaganda. Race theory, militarism, and anti-Bolshevism were allied themes, sadistic aggression their derivative. Sexual politics—a term apparently coined by Reich—was the key to understanding and combating Fascism.

Published on the run, the book had little effect. Its English edition (eliminating Reich's original Marxist language) was delayed until 1946. By then, Nazism was defeated, and Reich's new 'orgone' theory had made him an outcast. Earlier, Fromm had criticized Reich's overemphasis on genital sexuality in his essay

in the Frankfurt Institute's 1936 volume on *Studien über Autorität und Familie* (Horkheimer 1936)—the next development of the idea.

1.2 Fromm's Social Psychology and the Frankfurt Institute

Like Reich, Erich Fromm was a young Freudian analyst and a Marxist. In 1930, he joined the Frankfurt Institute, to provide social psychological expertise to Horkheimer's group. Fromm, too, tried to integrate Freud and Marx but still focused on the 'spirit' of capitalism, mentioning Fascism only in footnotes. Generally, Fromm's emphasis on the family as agent of authoritarian society in character formation, his references to patriarchy, sex morality, and the role of women were similar to Reich, whom he had met in Berlin and later in Switzerland. How much their ideas developed independently is unclear, although Fromm denied any influence. Reich may have been some steps ahead chronologically.

Four years later, the Frankfurt Institute, exiled to New York and renamed the International Institute for Social Research, published its *Autorität und Familie*. Horkheimer's foreword defined 'autoritär' as *autoritätsbejahend* (authority affirming). His lead essay, mentioning the 'authoritarian character' (here making its first appearance in print) as a submissive, masochistic type, dealt with freedom, reason, and authority; it invoked Hegel much more often than Marx. The Institute was moving away from historical materialism to Critical Theory. Fromm's chapter expanded his ideas into a detailed analysis of the psychology of submission to authority. He described, still in mostly orthodox terms, the anal character, which in authoritarian society metamorphoses into the 'autorite—masochistische' character—without mentioning antisemitism so important to Reich and later in Berkeley.

1.3 The Institute's Questionnaire Survey and 'Escape from Freedom'

The volume's empirical part drew a thumbnail sketch of revolutionary and authoritarian types in a summary of the Institute's survey of German workers and employees begun in 1929—years before anyone in the Institute had anticipated a Nazi takeover and before Fromm had joined the Frankfurt Institute (see Samelson in Stone et al. 1993). A complete report had to wait 50 years (Fromm 1983).

Fromm left the Institute in 1939, partly because of a quarrel with Horkheimer about scientific weaknesses in the study. He reformulated his theories in his 1941 book *Escape from Freedom*. Since the Reformation, modern man (sic) had acquired freedom from many constraints and developed several mechanisms of escape from his ensuing isolation, among them the authoritarianism of the sadomasochistic character. Fromm's 'neo-Freudian' ego psychology was responding to his new American environment as well as returning to his initial concern with capitalism. Transforming democracy's lonely citizens into more spontaneous and loving personalities, not the threat of Fascism, was now the greater challenge. Having moved away from the negations of the Critical Theorists, Fromm's more popular social psychology was shifting, like the field as a whole, toward 'interpersonal relations,' although Fromm personally remained politically engaged.

2. Empirical Research on the Fascist Character

Realizing the need for empirical data, the Critical Theorists themselves began a research project on antisemitism, a topic absent from their earlier work. In 1943, Horkheimer made contact with Nevitt Sanford, a young psychology professor at Berkeley, interested in problems of personality structure and ideology. Sanford and his graduate student Levinson were then constructing an attitude scale to measure antisemitism.

2.1 The Authoritarian Personality in Berkeley

Expanding their project, they linked up with ideas formulated in Germany a decade earlier. With Adorno representing the Critical Theory group, they developed a cooperative research plan to study what was initially called the 'Fascist character,' then 'Antidemocratic,' and finally *The Authoritarian Personality* (Adorno et al. 1950). Over five years in the making, the hefty volume described—in Horkheimer's words—a 'new anthropological species': the potential Fascist who, in contrast to the old-style bigot, combined the skills of industrial man with irrational beliefs. A cluster of ethnocentric attitudes and power orientation on the surface reflected an underlying personality structure characterized by conflict between submission and aggression, intolerance of ambiguity, and related dynamics, in their turn produced by authoritarian upbringing.

The book became an 'instant classic' (Jay 1973), praised as a model of imaginative and integrative social science research. It brought together sophisticated attitude scaling, opinion research, projective testing, and clinical interviews; it combined hard-headed empirical methods, clinical sensitivity, and original theoretical insights; and it dealt with a problem of major social significance. For two decades it produced a veritable flood of empirical follow-up studies. But a two-pronged counterattack on *The Authoritarian Personality* had also developed, the first one involving politics directly. Adapting Arendt's new

'Totalitarianism' theory of Fascism, it charged that the Berkeley group had concentrated exclusively on 'right-wing' authoritarians, while ignoring the 'authoritarians of the left': the Communists. Although differing, for doctrinal reasons, on surface attitudes such as antisemitism or ethnocentrism, fascists and communists were basically similar types. Complaining about the Berkeley group's naive left–right scheme, the critics denied any possibility of a real threat from American nativist–authoritarians.

These charges were not purely academic. At the height of the Cold War, the Berkeley group was accused of completely ignoring the authoritarianism of the left, that is, Communism, although it was the much greater threat. Such political attacks, at a time of difficulties for psychologists with real or alleged left-wing involvements, put some 'chill' on further work on this and related issues.

2.2 *The End of Ideology?*

The other line of criticism, more indigenous to psychology, tried in effect to drain the phenomenon of all political meaning. The dominant positivist-empiricist research style had already reduced the complex 'Berkeley project to the "California F Scale,"' a questionnaire handy for generating reams of data, while ignoring the rest of the theory. Whereas the outsiders had faulted the F Scale for capturing the wrong kind of political extremist, the insiders' professional consensus declared the F Scale invalid for the opposite reason. Due to a technical flaw, it identified not genuine protofascists, but only a 'response-set to agree,' that is, some very agreeable persons without strong opinions, or in one variant of this depoliticization process, only dogmatists of all stripes. Fixated on the paper-and-pencil F scale's real and alleged defects, investigators gradually abandoned this research, except for isolated individuals mostly in other countries. Repairing the scale's alleged defect proved as problematical as the original. Even Altemeyer's balanced Right-Wing Authoritarianism scale, laboriously constructed long after the F scale had fallen out of favor, tapped a much more restricted set of personality characteristics than those postulated by the original theory. Reich's original Freudo–Marxist idea of the interplay between societal organization and individual character mediated by the family had been fitted into an American empiricist, individual-psychology framework.

3. *Epilogue: Authoritarianism and the Dialectics of History*

Soon liberal, and by then staunchly anti-Communist, social science found it more appealing to attack, contrary to Horkheimer's claim, the old-style bigots from the moral high ground of antiracism. And as society changed, the predominant liberalism gave way to an apolitical scramble for practical applications capable of generating grants. Value-free empiricism had won the day—but only by reducing a genuine and important problem to a meaningless artifact, a 'response set,' drained of interest to any one except methodologists. The nuclear idea had moved from the Left to the Right until it ended in the mass media, which skillfully personalized it in the bigoted but basically harmless Archie Bunker—an utterly different, American version of the German *Feldwebel*. In the meantime, neoconservatives, in the 1950s insisting on the similarity of right- and left-wing authoritarianism, now stressed the differences between the authoritarian Right, who were possible allies, and the totalitarian Marxist–Leninists who only understood force. The psychological argument had turned into polemic and the label had become an epithet.

Recent attempts to revive the Authoritarian Personality failed to have much impact, though supporting some of the original findings (Altemeyer 1988). Whatever interesting results may be produced by tracking F scale means over time and space (Meloen, in Stone et al. 1993), they add little to our understanding of the theory. Social psychological textbooks of the 1990s usually devote only a few—respectful—paragraphs to it. Although it had been an intriguing idea, promising some understanding of major historical-political developments, a problematic initial assumption (that Hitler's main support had come from the lower middle class), the theory's erratic applications, and a changed political climate led to a shift in researchers' interests. What survived best is the pejorative label 'authoritarian,' widely used for persons or institutions considered—as in Horkheimer's original meaning—to be '*autoritätsbejahend*' (authority affirming), but stripped of the theory's complex personality dynamics and its antecedents.

See also: Anti-Semitism; Authoritarianism; Critical Theory: Frankfurt School; Ethnocentrism; National Socialism and Fascism; Parties/Movements: Extreme Right; Prejudice in Society; Xenophobia

Bibliography

Adorno T W, Frenkel-Brunswik E, Levinson D J, Sanford R N 1950 *The Authoritarian Personality*. Harper, New York

Altemeyer B 1988 *Enemies of Freedom*. Jossey-Bass, San Francisco, CA

Fromm E 1941 *Escape from Freedom*. Farrar 2 Rinehart, New York

Fromm E 1983 *Arbeiter und Angestellte am Vorabend des Dritten Reiches* (trans. Bonss W). Deutscher Taschenbuch Verlag, Munich, Germany

Horkheimer M (ed.) 1936 *Studien über Autorität und Familie*. Alcan, Paris

Jay M 1973 *The Dialectical Imagination*. Little, Brown, Boston

Reich W 1933 *Massenpsychologie des Faschismus*. Verlag für Sexualpolitik, Copenhagen, Denmark
Stone W F, Lederer G, Christie R (eds.) 1993 *Strength and Weakness. The Authoritarian Personality Today*. Springer Verlag, New York

F. Samelson

Authoritarianism

The construct of authoritarianism grew out of attempts by psychologists to understand the mass appeal of fascism in the first half of the twentieth century. As a result of a large body of empirical research, authoritarianism has become one of the major psychological explanations for the persisting finding of generalized prejudice and intolerance: the tendency for negative attitudes toward out-groups to correlate across different targets. It remains one of the most important approaches in the political psychology literature.

1. *Origins*

Research on authoritarianism was influenced by the attempts of European psychologists in the psychoanalytic tradition to understand mass support for fascism (Fromm 1941, Reich 1946). The major defining work was the publication in 1950 of *The Authoritarian Personality* by Adorno et al. (1950). This book of almost 1,000 pages combined psychoanalytic and personality theory with empirical research from in-depth interviews and questionnaires. While the project began as a study of antisemitism, the researchers quickly discovered that antisemitism appeared to be a part of a broader ethnocentrism. They located the origins of this ethnocentrism in basic personality processes as seen from a Freudian perspective.

Adorno et al. argued that punitive child-rearing practices and inconsistent parental affection produce a severe conflict within the young child. The repression of hostility toward the parents is, in adulthood, displaced onto groups perceived to be weak or inferior while authority figures (a generalization from the parents) are deferred to. Another consequence of this psychological conflict is a weak ego, which in turn leads to characteristics such as an unusual adherence to conventional social values, intolerance of ambiguity, and attraction to superstitions. From this perspective, prejudice has little to do with the actions or attributes of the target group but is a consequence of the psychological needs of those high in authoritarianism to direct repressed hostility toward subordinate groups.

Along with their explanation, Adorno et al. developed a paper and pencil measure of authoritarianism. It was labeled the F-scale because it was considered to be an indicator of susceptibility to fascist appeals. Introduced just as quantitative social science research was gaining momentum, an enormous literature soon developed that examined the correlates of authoritarianism, as measured by the F-scale or similar measure. As of 1989 there were well over 2,000 publications on authoritarianism and related constructs (Meloen 1993). Studies cataloged the relationship between the F-scale and social background characteristics, various measures of personality, and a wide range of social and political attitudes (see Kirscht and Dillehay 1967 for a typical review of this research).

2. *Criticisms*

The Authoritarian Personality was subjected to close scrutiny and criticism almost immediately after it was published (see Christie and Jahoda 1954). Methodologically, problems with the F-scale measure of authoritarianism cast doubt on much of the empirical research. Since all of the questions in this scale were worded so that an agree response indicated authoritarian characteristics, a tendency to agree with questions like this regardless of content (agreement response set) could produce high authoritarianism scores independent of any individual differences in personality. And while the questions in the F-scale were developed to tap the various syndromes hypothesized to make up authoritarianism, it is not clear that the measure reflects the underlying conceptualization. Factor analyses of the F-scale always show the measure to be multidimensional but the factors are always difficult to interpret and never cleanly correspond to the hypothesized components (see Altemeyer 1981).

Critics also argued that the authoritarianism measure, and possibly the construct itself, was ideologically biased. Although Adorno et al. were clearly responding to the history of fascism in Europe, the growing threat of communism in the 1950s led to concerns about an authoritarianism of the left (Shills 1954). However, despite observations that communists appeared to possess traits characteristic of authoritarians, there is little empirical evidence to support these claims (Stone 1980).

Perhaps most importantly, the Freudian-based theory underlying *The Authoritarian Personality* proved far less useful than the authors thought and as a result research become increasingly divorced from the theory. In addition to the general loss of status of Freudian explanations in empirical social science research, Adorno et al.'s specific explanation of authoritarianism has not fared well. Although quantitative research established relationships between the F-scale, social background characteristics, and social

attitudes, evidence on the dynamics and origins of authoritarianism consistent with the Adorno et al. theory proved elusive (see Altemeyer 1981, 1988, Duckitt 1989). The bulk of research based on *The Authoritarian Personality* thus consists of short empirical studies with only limited guidance from the original theory. Most of these studies report correlations between measures of authoritarianism and a small number of other individual level characteristics: social background, personality, values, and attitudes.

3. *More Recent Approaches*

Several attempts have been made since the publication of *The Authoritarian Personality* to develop alternative explanations for individual differences in prejudice and intolerance. Perhaps the best known of these is Rokeach's research on dogmatism reported in *The Open and Closed Mind* (1960). Rokeach moved from the Freudian perspective of Adorno et al. to a more cognitive explanation based on the structure and function of belief systems. He argued that prejudice and intolerance are a result of closed belief systems that ward off threats by limiting the scope of accepted beliefs and rejecting all challenges. Like Adorno et al., Rokeach developed a measure of dogmatism based on the same methodology of agree–disagree attitude statements. His dogmatism measure was also constructed of a diverse range of statements designed to reflect the various elements of open and closed belief systems.

While research has utilized Rokeach's dogmatism scale, most of it, like research using the F-scale, consists of narrow empirical studies that examine the correlates of the measure. There has been little reported work that has examined Rokeach's conceptualization of open and closed belief systems despite his own detailed and novel explorations reported in the book. In addition, the measurement properties of the dogmatism measure have proved to be no improvement on the F-scale. It too has items that are worded in only the agree direction and the interitem correlations and dimensionality of the measure may be even worse than the F-scale (Altemeyer 1981).

The most significant new contribution to the authoritarianism literature has been a recent body of research reported by Altemeyer (1981, 1988, 1996). From years of studies using undergraduate students, Altemeyer has both developed a new and more reliable measure of the construct and advanced a new conceptualization based on social learning theory. Rejecting the Adorno et al.'s Freudian explanation, Altemeyer (1988) prefers a simpler conceptualization that sees authoritarianism as a social attitude (or cluster of attitudes) that is learned through interactions with parents, peers, schools, the media, and through experiences with people who hold conventional and unconventional beliefs and lifestyles. His measure of right-wing authoritarianism (RWA) is more reliable and unidimensional than previous scales and has items balanced for agreement response set.

There have been several other attempts to reconceptualize the authoritarianism phenomenon. Kelman and Barkley (1963) argued that high authoritarianism reflects a limited breadth of perspective. Lipset (1960), Selznick and Steinberg (1969), and others have suggested that authoritarianism is a function of social learning associated with lower social class and less educated environments. More recently, Duckitt (1989) has developed a model in which authoritarianism reflects the desire to maintain strong in-group identifications.

4. *Limitations and Problems*

In addition to the measurement and theoretical problems discussed above, there are other significant shortcomings of the authoritarianism approach.

How do factors in the social environment exacerbate or moderate predispositions toward prejudice and intolerance? Some researchers have hypothesized that social threats or anxiety will arouse prejudice and punitiveness and encourage support for extreme right-wing parties and even mass violence (Staub 1989). There is also some evidence from archival and individual level studies that levels of authoritarianism, or the strength of the connection between authoritarianism, prejudice, and intolerance, increase during periods of threat (Sales 1973, Doty et al. 1991, Feldman and Stenner 1997). Although there appears to be strong impressionistic evidence that prejudice, intolerance, and support for right-wing groups wax and wane over time, there is nothing obvious in the present conceptualizations of authoritarianism that would help to explain this phenomenon.

More generally, authoritarianism is generally treated as a purely dispositional construct; researchers almost always examine simple additive relationships between measures of authoritarianism and indicators of prejudice and intolerance. This type of approach provides no understanding of how the reality of social and political conflicts—or any other environmental factors—interact with authoritarian predispositions. This leaves authoritarianism a purely 'personality' explanation, with only weak connections to other explanations of prejudice and intolerance.

It is also necessary to make progress on the relationship between authoritarianism and ideology, particularly since the authoritarianism construct is beginning to be used in studies of countries with a history of communist governments. Studies in the USA and Canada consistently find negative relationships between measures of authoritarianism and attitudes toward communism. In Russia, however, researchers have shown that 'right-wing authoritarianism' has been associated with support for communism (McFarland et al. 1992). This relationship can be understood if support for communism in countries

like this represents a commitment to the established, authoritarian order. Even so, a positive correlation between 'right-wing' values and communism strongly suggests that the construct of authoritarianism must be defined and measured in a way that allows it to be used in cross-national research.

5. Conclusions

Despite all of the problems that have been raised with the authoritarianism construct since the 1950s, it continues to be used in research on prejudice and intolerance. Its longevity is, in part, a function of the repeated findings of consistency in expressions of prejudice and intolerance across a wide range of targets. Individual differences account for a significant proportion of the variance in these attitudes and authoritarianism remains one of the major explanations of those individual differences.

See also: Authoritarian Personality: History of the Concept; Authority, Social Theories of; Communism; Communism, History of; National Socialism and Fascism; Prejudice in Society; Totalitarianism; Totalitarianism: Impact on Social Thought

Bibliography

Adorno T W, Frankel-Brunswick E, Levinson D J, Sanford R N 1950 *The Authoritarian Personality*. Harper, New York

Altemeyer B 1981 *Right-Wing Authoritarianism*. University of Manitoba Press, Winnipeg, MB

Altemeyer B 1988 *Enemies of Freedom*. Jossey-Bass, San Francisco

Altemeyer B 1996 *The Authoritarian Specter*. Harvard University Press, Cambridge, MA

Christie R, Jahoda M (eds.) 1954 *Studies in the Scope and Method of 'The Authoritarian Personality.'* Free Press, Glencoe, IL

Doty R M, Peterson B E, Winter D G 1991 Threat and authoritarianism in the United States. *Journal of Personality and Social Psychology* **61**: 629–40

Duckitt J 1989 Authoritarianism and group identification. *Political Psychology* **10**: 63–84

Feldman S, Stenner K 1997 Perceived threat and authoritarianism. *Political Psychology* **18**: 741–70

Fromm E 1941 *Escape from Freedom*. Rinehart, New York

Kelman H C, Barclay J 1963 The F scale as a measure of breadth of perspective. *Journal of Abnormal and Social Psychology* **6**: 608–15

Kirscht J P, Dillehay R C 1967 *Dimensions of Authoritarianism*. University of Kentucky Press, Lexington, KY

Lipset S M 1960 *Political Man*. Doubleday, New York

McFarland S G, Ageyev V S, Abalakina-Pap M A 1992 Authoritarianism in the former Soviet Union. *Journal of Personality and Social Psychology* **63**: 1004–10

Meloen J D 1993 The F scale as a predictor of fascism. In: Stone W F, Lederer G, Christie R (eds.) *Strength and Weakness: The Authoritarian Personality Today*. Springer-Verlag, New York, pp. 47–69

Reich W 1946 *The Mass Psychology of Fascism*. Orgone Press, New York

Rokeach M 1960 *The Open and Closed Mind*. Basic Books, New York

Sales S M 1973 Threat as a factor in authoritarianism. *Journal of Personality and Social Psychology* **28**: 44–57

Selznick G J, Steinberg S 1969 *The Tenacity of Prejudice*. Harper & Row, New York

Shills E A 1954 Authoritarianism: Right and left. In: Christie R, Jahoda M (eds.) *Studies in the Scope and Method of 'The Authoritarian Personality.'* Free Press, Glencoe, IL, pp. 24–49

Staub E 1989 *The Roots of Evil*. Cambridge University Press, Cambridge, UK

Stone W F 1980 The myth of left-wing authoritarianism. *Political Psychology* **2**: 3–19

S. Feldman

Authority: Delegation

Authority is a relationship between a superior or overseer and a subordinate, whereby the subordinate relies upon the superior for specific direction. The concept and practice of authority has a position of prominence in the history of human society as well as the academic fields of public policy, public administration, political science, sociology, economics, and psychology. Considerations of authority are particularly important in the policy sciences, because politics is fundamentally about who gets what, when, and how, and, perhaps most importantly, *who decides*. Since people generally would like to maximize their own decision-making authority, to promote their own interests and causes, while simultaneously minimizing the authority to which they themselves are subject, determinations of authority of necessity play a crucial role in moderating such conflicting impulses so as to organize cooperative human activity and civil society. However, authority relationships and delegations of authority are rarely self-executing, unambiguous, or nonproblematic. Often, authority is not merely granted but must be earned, and authority relationships must be negotiated and renegotiated.

1. Definition of Authority

Authority is nearly as difficult as it is important to define clearly. At a minimum, authority can be thought of as either a possession, as in claiming that someone 'speaks with authority' on a matter, or as a relationship, as in the fact that a superior exercises authority over a subordinate. The very use of the term 'exercise' in tandem with 'authority' signals the sense in which authority can be likened to a muscle, which can increase in mass and tone with use, as opposed to a

checking account balance, which inevitably decreases with use.

Flathman (1980) cuts the matter slightly differently with his distinction between 'an' authority and 'in' authority. *An authority* is a person who is generally thought of as possessing specialized knowledge regarding a topic. *In authority* is drawn from occupying an office or holding a formal title. A criminologist might be described as 'an authority on crime,' whereas one would contact a police officer if one sought 'help from the authorities.' Many contemporary normative debates over authority structures and practices highlight this difference, with the general trend in most industrialized nations being towards shifting from a reliance on 'in authority' to a reliance on 'an authority,' with the judgment regarding whether or not someone *is* an authority in relationship to another person being left increasingly to the potential subject of that authority.

1.1 Distinctions Between Authority and Power or Influence

Authority is primarily and increasingly considered to be a relationship between a superior or overseer and a subordinate, whereby the subordinate relies upon the superior for specific direction, whether it be in the form of expert advice or operational commands. A respected dictionary definition of authority is 'power to influence or command thought, opinion, or behavior' (*Webster's Ninth Collegiate Dictionary* 1983). This definition suggests that authority is related to, but distinct from, the concepts of *power* and *influence*. A widely accepted definition of *power* is that person A has power over person B when A can get B to do what A wants B to do, which is something that B would not have done on B's own (Dahl 1957). The existence of power in a relationship depends upon B's behavior actually being altered from the course that it otherwise would have taken, specifically as a result of A's intervention. *Influence* is closely related to power in that it concerns itself with an external force (e.g., a person) acting upon an object (e.g., another person) to change the object's status or behavior in some way. Influence can be distinguished from power in the sense that the outcome it produces tends to be less specifically and directly intended (e.g., the influence that a mentor has on a student) than does the outcome of the exercise of power. However, both power and influence are contingent upon outcomes.

Authority is less teleological, or outcome-focused, than power. Someone can be an authority or in a position of authority over another person before, and even without, altering the other person's behavior in any meaningful way. They may or may not influence the other person, yet remain in a position or relationship of authority over that person. The existence of authority depends more on what is brought to the situation, in terms of reputations, titles, offices, and positions, than on what actions (if any) are taken as a result of the exercise of authority.

1.2 Anticipatory Reactions

Since outcomes do not determine authority, we ought to focus our attention on the process of exercising authority in order to identify when authority does and does not exist. However, such an approach is potentially limited by the problem of anticipatory reactions. A subordinate might anticipate the preferences of a superior so perfectly that the subordinate acts in exactly the way that the authority wishes the subordinate to act without the superior needing to command the action explicitly. Anticipatory reactions to authorities can occur because a subordinate has fully internalized the preferences and values of the superior, or because the subordinate knows what the authority prefers and fears punishment if the preferred behavior is not forthcoming. For example, motorists who approach traffic lights that are changing to red choose either to speed through the light or to stop. Motorists know, at least subconsciously, that the authorities would like them to stop. Motorists might stop because they fear the possible physical consequences of not stopping, in which case authority has not affected their decision. However, motorists also might stop because they have internalized the value of orderly and predictable driving behavior, because they fear that a police officer might ticket them for running the red light, or because a traffic cop explicitly holds up a hand, ordering them to stop. All three of these responses would be reactions to authority, although the first two could be considered cases of anticipatory reactions.

2. Legitimacy

The concept of legitimacy is central to the exercise of authority. Regardless of the source of authority, if it is not perceived as legitimate by the object of authority, then there is little chance of the authority being exercised effectively. Authority is considered to be legitimate if the object of authority accepts the appropriateness of its exercise and therefore submits to its dictates (Barnard 1968). Legitimacy renders authority *authoritative* in a given case.

2.1 Weber's Ideal Types

Although Max Weber was not the first scholar to inform thinking on legitimacy as it concerns authority, his contribution to our understanding on this topic is profound. Weber (1947) argues that the legitimacy which is attached to authority stems primarily from

one of three general sources, which he characterizes as *ideal types*. The three Weberian ideal types of legitimate authority systems are traditional, charismatic, and rational-legal in design.

Under a traditional system of authority, legitimacy adheres in the station of the person who is exercising authority, and is largely unearned. The roots of traditional authority systems lie in early pagan religions and the traditional structure of authority relationships in the family. Most early religions and mythologies identified gods with forces of nature that were mysterious to, and more powerful than, the people and societies who were regularly subjected to their apparent dictates. Thor, the Norse God of Thunder, was thought to rule the sky, and the Greek god Poseidon to rule the sea. The founding event of the Jewish religious faith was a supernatural event, Abraham's witness to the burning bush that was not consumed, and from which God announced 'I am who am,' generally interpreted to mean 'I am existence itself,' something both superior to and essential to human beings. Thus began the traditional understanding of the legitimacy of God's dominion over the Jewish people and, by extension, all of creation. God neither sought nor needed to earn authority over human beings. All God had to do was declare that authority. The authority was considered to be legitimate simply because God is God.

Traditional sources of legitimate authority have also been evidenced in the power relationships of family structures ('obey me because I am your father') as well as the feudal and caste systems employed by various premodern cultures. Perhaps the clearest example of a traditional basis for the legitimacy of authority was the principle of the Divine Right of Kings. Kings were to be obeyed because they were kings, regardless of the prudence or benevolence of their commands. Kings, in particular, were thought to be appointed by God, with subsequent succession usually determined by birth or kinship, though occasionally also influenced by human schemes. Aquinas (1953, p. 160) captures the essence of the traditional source of legitimate authority when he says: 'In natural things it behooved the higher to move the lower to their actions by the excellence of the natural power bestowed on them by God; and so in human affairs also the higher must move the lower by their will in virtue of a divinely established authority.' Under a traditional system of authority, one is born into one's natural station in life, and the position of that station, from prince to slave, determines over whom one exercises legitimate authority, and to whose authority one is subject.

Weber's second ideal type of legitimate authority is termed charismatic. Under a charismatic system of authority, the legitimacy of the authority is tied to the person and is in some sense earned. Mahatma Gandhi exercised legitimate authority over the actions of many people in India, not because of any offices he held, nor due to his natural station in life, but because of who he was and what he advocated. The roots of scholarly thinking about charismatic authority go at least as far back in time as ancient Greece, when Plato argued that philosophers deserve to be kings because they possess the greatest wisdom, and Aristotle reasoned about the authority merited by effective leaders. Weber argues that charismatic authority is perceived to be legitimate because the charismatic leader binds the wills of subjects to the leader and the leader's causes. Once a leader has persuaded followers that the cause to be pursued is worthy, the followers suspend their personal judgment and follow that leader. Because the commitment to the leader and cause is both personal and complete, few if any formal systems or rules for ensuring compliance are necessary. Charismatic leaders employ the force of their personality and the power of their ideology to meld the wills of subordinates to them and their cause. As such, Weber's charismatic authority system closely resembles the prescription of Rousseau (1986) that citizens must always submit to the legitimacy of the General Will once it has been formulated.

The third and most modern of Weber's ideal types of authority systems is termed rational-legal. Under a rational-legal system, authority adheres in the office that is occupied by the person who exercises authority, a system that Oakshott (1975) has termed 'formal-procedural' or 'F-P' authority. One obeys a police officer not because of cultural tradition or the officer's charisma, but because the officer occupies a position of authority. The police officer's exercise of authority is considered to be legitimate because a rational system of training and selection places qualified officers in such positions, and clarifies and delimits their authority. Because authority adheres in the office and not the person, transitions and successions can be seamless under a rational-legal system. When the chief of police retires, a successor is appointed, often by promoting the next-highest-ranking officer in the organization. Subordinate officers and citizens continue to submit to the authority of the new chief, just as they did to the old one, because the new chief now holds the office of chief of police. Weber reasons that the rational-legal system of legitimating authority is both highly efficient and conducive to promoting the industrial and organizational development of societies, though he also fears the capacity of such systems to dehumanize workers and to assume dangerous degrees of autonomy from political and social control.

Administrative hierarchy is a central yet oft maligned feature of a rational-legal authority system. Official offices are arranged in a pyramid-shaped hierarchy to formalize and systematize authority relationships within organizations as well as to provide for a rational system for training and promoting workers. However, along with excessive formalization, a condition that is often referred to as 'red tape,' hierarchy has been a flashpoint for criticisms of

rational-legal systems of authority. Parkinson (1957) reasons that hierarchies have a self-perpetuating logic to them, as officials in a hierarchy will seek to increase the number of their subordinates, and thus enhance their relative status in the hierarchy, far beyond the number of people required to do the work of the organization. More recently, Light (1995) observes that modern governmental organizations in the USA are becoming 'thick' in the sense that their hierarchies are as wide in the middle as they are at the bottom. Contrary to Parkinson's claim, managers in thick organizations supervise few if any direct subordinates but gain great status and authority from their formal titles, such as Principal Deputy Assistant Undersecretary for Regulatory Affairs.

2.2 Legitimacy Hybrids

Another important consideration regarding Weber's ideal-type systems of authority is the existence of hybrids. Weber never claimed that his ideal-types would describe exactly the empirical reality of authority systems, so we should not be surprised that few, if any, systems reach the purity of any of his types. Nevertheless, it is strikingly common for organizational systems to employ multiple systems for legitimating authority. This reality suggests that the legitimacy of authority is so vitally important to a government, society, or organization, that no single source is trusted entirely to ensure it.

The Westminster governing system employed by the United Kingdom as well as other countries is a case in point. The most influential body of government, the House of Commons, is organized and operates based on a rational-legal model of formal rules and hierarchy that extends from the Prime Minister (PM) through the various cabinet ministers (elected Members of Parliament who are appointed to the cabinet by the PM) and down through the executive branch of government, which is staffed exclusively by career civil servants. The PM is elected to that office by fellow party members, and remains in office as long as he or she avoids a parliamentary vote of no confidence, or the electoral defeat of the majority party. Thus, the PM tends to draw upon charismatic foundations for his or her legitimacy. When the charismatic luster of a PM's standing wanes, as it did for Margaret Thatcher in 1990, the PM's authority is questioned and he or she may be replaced (as Thatcher was by John Major) in order to avoid the fall of the government due to a successful vote of no confidence.

At the same time, the UK government includes a number of institutions that draw upon traditional authority for their legitimacy. The most obvious one is the sitting Monarch, who serves as formal head of state as well as in less formal roles as a social and moral leader of the country. The second body of Parliament, the House of Lords, has traditionally drawn its membership from the titled aristocracy and retains a limited, though occasionally pivotal, role in the approval or disapproval of initiatives passed by the House of Commons. The Judiciary exercises its authority largely based on respect for the body of precedents and practices referred to as 'Common Law,' or the traditional content and practice of law in the country. Additional examples of governments that employ interesting hybrids of traditional, charismatic, and rational-legal authority systems include contemporary Russia, North Korea, Syria, Canada, and the USA, among others.

Hybrids of legitimate authority systems are also common in individual organizations, especially due to the distinction between the formal and the informal features of an organization. Formal structures, practices, and relationships are those that are contained within the official design of the organization and its written operational rules. Informal structures, practices, and relationships are those that exist outside of these official bounds, in what has been called the *culture* of an organization (Schein 1985). After acknowledging this critical differentiation, Bennis et al. (1958) proceed to argue that authority only resides in the formal aspects of an organization. Influence that is exerted through informal organizational channels, such as personal relationships, is the exercise of power, not authority. Other prominent scholars of organizations claim that informal structures can be as well developed and predictable as formal structures in organizations, and that authority relationships exist simultaneously within both of them (Crozier 1963, Downs 1967). For example, the employee at a company who is in charge of Information Technology (IT) is likely to hold a subordinate position in the formal structure of the firm. However, because IT is critical to the work of all the other employees in the organization, including the boss, and most of those other workers are not *authorities on* IT, tasks such as computer programming and Internet trouble-shooting, as is the IT professional, in practice the IT manager at a company possesses a level of informal authority that may at times exceed the formal authority even of the firm's leader.

3. Extensions of and Alternatives to Weberian Legitimacy

3.1 Procedural Legitimacy

Weber's articulation of the ideal type rational-legal bureaucracy is merely a focal point in a long tradition of thinking about *procedural legitimacy*. According to procedural legitimacy, commands are authoritative regardless of their content, as long as proper procedures were followed in developing and issuing the commands. An early example of argumentation in the procedural legitimacy tradition is the trial of Socrates, where the great Greek philosopher was found guilty

and sentenced to death for allegedly corrupting the youth of Athens. As recounted by Plato, Socrates rebuffed the entreaties of his friends to escape the harsh sentence issued by the state, claiming that he was properly tried by his peers in a court of law based on sound procedures. Since the procedures that were followed in the trial were legitimate, Socrates argued that he had no choice but to abide by the result. The role of procedural legitimacy in rendering authority authoritative is also emphasized in more recent philosophical works such as Rawls's *A Theory of Justice* (Rawls 1971), where justice is equated with fairness, and fairness is largely determined procedurally.

3.2 The Authoritarian vs. the Authoritative

More recent scholars have developed further the concept of legitimate authority. First, they have dealt with the important distinction between the authoritarian and the authoritative. Generally, an *authoritarian* approach to authority involves the issuance of commands that are expected to be obeyed automatically; whereas, an *authoritative* style of oversight relies upon explanations for rules and justifications for orders, to help render the orders and rules themselves authoritative. Along these lines, Flathman (1980) criticizes the assumption of formal-procedural theories that F-P legitimacy guarantees obedience. Drawing from Tocqueville (1842), Flathman claims that shared social values and beliefs are required to make formal rules authoritative. Absent such shared values and norms, effective authority is all but impossible. Flathman's critique leaves open the question of whether shared norms regarding the legitimacy of authority must emerge naturally from the grassroots, or can essentially be imposed on a society by the political leadership. Nevertheless, his point is bolstered by the related claim of Simon (1957) that subordinates in a hierarchical organization will only obey orders if the substance of those orders fall within a *zone of acceptability* that is bounded by the personal values and belief system of the subordinates.

Modern and postmodern commentators on authority and superior/subordinate relationships have confronted the important question of how to shape the zone of acceptability of workers so that it includes the actions that are necessary to advance the interests of an organization. Barnard (1968) writes of the exchange relationship that surrounds employment relationships, whereby the organization's executive provides material and purposive incentives to workers in exchange for their cooperation in the necessary activities of the firm.

3.3 Horizontality and Team-based Organization

Recently, public-administration theorists have promoted the idea of horizontality and team-based organization as both an incentive to induce worker cooperation, and a more efficient and effective administrative structure for organizing work activities (Osborne and Gaebler 1992). Under the principle of *horizontality*, steep rational-legal style administrative hierarchies are flattened by eliminating many middle-management positions and by establishing working relationships horizontally across functional or geographic divisions of an agency (Ostroff 1999). Under team-based systems, such as those called for by Total Quality Management, new working groups are organized to complete specific projects, comprising workers of various formal ranks and with varying skills and expertise (Cohen and Brand 1993). A team leader is appointed, who may or may not hold the highest formal rank in the group, although all team members are expected to collaborate as equals. The authority ambiguity that is a signature characteristic of team-based organization, and perhaps its most alluring feature as a management reform, also appears to be its greatest weakness. A common complaint of participants in team-based administrative operations that fail is that no one took charge of the operation, thus leading it to founder. Also, because many team-based organizations are matrix organizations that retain the formal structure of a rational-legal bureaucracy on paper, at any stage of a project a manager might reimpose the formal structure of authority on a team, thus quashing the organization's horizontality.

4. Delegation

Horizontal organization and team-based management reforms are specific manifestation of a concept that is closely associated with considerations of authority, namely, delegation. Delegation involves the transfer of authority from a superior to a subordinate. The transfer can be permanent, but often it is temporary, and the superior retains the ability to reclaim the authority if necessary. The authority that is delegated can be general, as in when a deputy secretary of a cabinet department is authorized to make all necessary decisions in the absence of the secretary, or specific, as when a subordinate is assigned the authority to review a set of contract bids and select the winner. Delegations of authority are explicit and tend to be formally documented through general rules or specific memoranda, as when a family member or friend is granted the legal 'power of attorney' to make health care or financial decisions on behalf of an incapacitated person.

4.1 Distinction Between Delegation and Discretion

Delegated authority is similar to, but distinct from, discretion. Discretion generally refers to decision-making authority within specified bounds or parameters that officials have by virtue of their position or office. Grants of discretion tend to be more informal and more implicit than delegations of authority.

Discretion often refers to the decision as to whether or not to use delegated authority in a particular situation. For example, police officers in most societies are authorized (e.g., granted the authority) to use deadly force to defend themselves and other citizens from serious threats of violence. They are then instructed to use their discretion (i.e., judgment within certain reasonable bounds) in deciding when are appropriate or inappropriate times to use the authority that has been delegated to them. Such circumstances are referred to commonly in the public administration literature as situational imperatives in which it is either impossible or unwise for superiors (at headquarters) to explicitly issue real-time commands to their subordinates in the field.

4.2 The Principle of Subsidiarity

An important question regarding delegation is where authority best lies in a given organization and for a given decision. The principle of subsidiarit*y* holds that decision-making authority is best placed (a) where responsibility for outcomes will occur; and (b) in the closest appropriate proximity to where the actions will be taken that will produce the outcomes. Authority ought to be matched with responsibility, because one tends to use authority more judiciously if one knows that one will be held accountable for the results. Moreover, once so matched, both authority and responsibility ought to be situated as closely as possible to operations, so that decisions might be well-informed by the context of real situations. A classic example of the application of the principle of subsidiarity is the position of the Roman Catholic Church that parents should have the maximum reasonable amount of authority over, and responsibility for, the raising of their children. The nature and effectiveness with which children develop into responsible adults affects the lives and interests of parents more directly and profoundly than it does the state, which the Church advises should only become directly involved in child rearing when and if individual families have clearly failed to discharge their responsibilities.

In the field of public administration more specifically, it is generally held that strategic decisions requiring a broad perspective are best made by executives near the top of an administrative hierarchy. Such top-level managers tend to have the generalist training, experience, and political sensibilities that are necessary in order to balance competing claims in light of the overall mission and goals of the agency. Tactical decisions regarding exactly how a task is to be performed on the ground are thought to be best left to the subordinates on the scene (Wilson 1989). For example, the director of a law enforcement agency confronted with an armed standoff with a suspected criminal might instruct a sharpshooter to 'shoot the suspect if he threatens anyone's life and if you can get a clear shot.' It would then be left to the operator (the sharpshooter) to use discretion in deciding whether or not the suspect's actions were threatening, and when the sharpshooter's line of fire was sufficiently 'clear' to justify pulling the trigger.

4.3 Challenges to Effective Delegation

An actual case of the above hypothetical situation underlines the ambivalence of authority, delegation, and discretion for superiors and subordinates. The Federal Bureau of Investigation in the USA faced a similar situation in a standoff with suspected criminal Randy Weaver in Ruby Ridge, Idaho. Unfortunately, the sharpshooter was mistaken in thinking that his line of fire was clear, and he shot and killed Weaver's pregnant wife while aiming for Weaver. Blame for the tragedy was placed primarily on the executive at headquarters who had issued the order to take the shot if it was clear. In that case, responsibility for the results of the operational decision remained with the superior, even though the authority to use force and the discretion to decide when the circumstances were appropriate had been delegated to the subordinate. Such complications often lead superiors to resist delegating decision-making authority or granting wide discretion to subordinates, since they want to avoid being held responsible for decisions that they no longer control. Similarly, subordinates often resist delegations of responsibility that accompany delegations of authority, since they fear being made a scapegoat for undesirable outcomes that may not be entirely their fault.

A final challenge for the effective use of delegated authority is the fact that, as with authority in general, delegated authority is only legitimate in practice if the target of authority recognizes it as such. Going back at least as far as the medieval period, nobles used their personal seal or stamp on official letters of introduction and delegations of authority to ensure that the recipient of the document would be considered legitimate in speaking on behalf of the official or exercising the authority that was granted. Although many delegations of authority are still documented in similarly formal ways, others are not so clearly codified. Moreover, just as the citizens of modern societies tend to question openly the legitimacy of authority more readily than did their forebears, the legitimacy of delegated authority can be exceedingly tenuous today. Although authority may easily be delegated, the respect that is required in order for that delegated authority to be exercised effectively often must be hard-earned.

5. Conclusion

Authority is best thought of as a relationship between a superior and a subordinate wherein the commands or preferences of the superior become authoritative for

the subordinate as long as the authority is perceived to be legitimate and the action preferred or commanded falls within the bounds of what the subordinate considers to be acceptable. The legitimacy of authority may be grounded in the natural status of the superior, the charisma or cause of the superior, or the formal office and titles that are held temporarily by the superior. People in authority and systems of authority often draw upon multiple sources for legitimacy, and the extent to which the commands of authorities become authoritative depend upon the embrace by subordinates of certain shared values and social norms.

The delegation of authority and the granting of discretion to subordinates in deciding when such authority is best used is generally thought to be desirable so long as responsibility and authority are linked properly, subordinates are likely to possess better-informed judgment regarding the decision as to when and how to use the authority, and subordinates have reason to believe that they will not be made scapegoats for unfortunate outcomes. Authority relationships are ubiquitous. Our desire to live in a state of social cooperation to achieve shared goals demands that we continue to think long and hard about how authority and its delegation ought to organize our families, work, communities, societies, and our world.

See also: Authoritarianism; Behavior, Hierarchical Organization of; Delegation of Power: Agency Theory; Influence: Social; Legitimacy; Management: General; Organizational Culture; Organizations: Authority and Power; Power in Society; Power: Political; Weber, Max (1864–1920)

Bibliography

Aquinas St T 1953 Question 104: Of obedience. *Summa Theoligica II–II* [In: Bigongiari D (ed.) *The Political Ideas of St. Thomas Aquinas*. Hafner, New York]

Aristotle 1941 *Metaphysica* [Metaphysics. In: McKeon R (ed.) *The Basic Works of Aristotle*. Random House, New York]

Barnard C I 1968 *The Functions of the Executive*. Harvard University Press, Cambridge, MA

Bennis W G, Berkowitz N, Affinito M, Malone M 1958 Authority, power and the ability to influence. *Human Relations* **11**: 143–56

Cohen S, Brand D 1993 *Total Quality Management in Government*. Jossey-Bass, San Francisco, CA

Crozier M 1963 *Les phénomène bureaucratique* [1964 *The Bureaucratic Phenomenon*. University of Chicago Press, Chicago, IL]

Dahl R A 1957 The concept of power. *Behavioral Science* **20**: 201–15

Downs A 1967 *Inside Bureaucracy*. Little, Brown, Boston, MA

Flathman R E 1980 *The Practice of Political Authority: Authority and the Authoritative*. University of Chicago Press, Chicago, IL

Hobden M, Jr. 1988 Bargaining and command by heads of US government departments. *The Social Science Journal* **25**: 255–76

Light P C 1995 *Thickening Government: Federal Hierarchy and the Diffusion of Accountability*. Brookings Institution, Washington, DC

Oakshott M 1975 *On Human Conduct*. Clarendon Press, London

Osborne D, Gaebler T 1992 *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Public Sector*. Addison-Wesley, Reading, MA

Ostroff F 1999 *The Horizontal Organization: What the Organization of the Future Looks Like and How it Delivers Value to Customers*. Oxford University Press, New York

Parkinson C N 1957 *Parkinson's Law, and other Studies in Administration*. Houghton Mifflin, Boston, MA

Plato 1968 *Res Publica* [*The Republic*. Basic Books, New York]

Plato 1969 *Crito* [In: Tredennick H (ed.) *The Last Days of Socrates*. Penguin, New York]

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Rousseau J J 1986 *Discours sur l'origine et les fondements de l'inégalité parmi les hommes* [Second discourse. In: Gourevitch V (ed.) *The First and Second Discourses Together with the Replies to Critics and Essay on the Origin of Languages*. Harper & Row, New York]

Schein E H 1985 *Organizational Culture and Leadership*. Jossey-Bass, San Francisco, CA

Simon H A 1957 *Administrative Behavior*, 2nd edn. Macmillan, New York

Tocqueville A D 1842 *De la démocratic en Amérique* [1969 *Democracy in America*. Harper & Row, New York]

Weber M 1947 *Wirtschaft und Gesellschaft* [*The Theory of Social and Economic Organization*. Free Press, New York]

Webster's Ninth Collegiate Dictionary 1983. Merriam-Webster, Springfield, MA

Wilson J Q 1989 *Bureaucracy: What Government Agencies Do and Why They Do It*. Basic Books, New York

P. J. Wolf

Authority, Social Theories of

Authority is best defined as a relation that exists between individuals. A relation of authority exists when one individual, prompted by his or her circumstances, does as indicated by another individual what he or she would not do in the absence of such indication. This I will refer to as *the authority relation*. The *legitimacy* of an authority relation is what keeps the relationship from breaking down, and is the answer to the question: Why does the one who follows do as indicated by the one who rules?

The authority relation is perhaps as old as civilization itself. The very moment in which individuals began living in communities and tried to organize the tasks required for their survival gave birth to the need to coordinate the actions of many, to delegate among the members of the group, and to trust that this coordination and delegation process would occur. Any resolution of such demands, whether successful

or not, involved the constant creation and destruction of authority relations. In this context, a social theory of authority is a collection of principles aimed at understanding: (a) how the circumstances of living in a community affect the authority relations that exist among its members, and (b) how the evolution of the community itself is affected by this web of authority relations.

1. *Origins*

The intrinsic importance of the matter of understanding authority relations as they arise in society has led to a large list of authors who have written on the subject (an excellent survey of the early meaning of authority can be found in Arendt 1958). Plato, for example, wrote that only 'those who can apprehend the eternal and unchanging [aspects of reality]' (Plato 1945, p. 190) ought to occupy positions of authority. There are two reasons why Plato wanted this tall order to be met. First, he thought that only those who are enlightened in this way could understand that which is ultimately *Good*, and for Plato the exercise of authority is only legitimate to the extent that one brings society closer to that Good. The second reason has to do with the survival of the seekers for wisdom themselves. Plato argued that whenever those uninterested about matters of truth and wisdom are the rulers in a community, the conditions for the attainment of wisdom disappear, as community life is organized in a way that leads the seekers of wisdom to become corrupted and interested in other matters.

Aristotle makes no reference to a Good order, the understanding of which qualifies one to rule in a community. He notices simply that we all have *natural* talents and dispositions toward the different tasks. One of them is that of ruling wisely and effectively. Those in possession of such disposition ought to be the ones who rule. The act of ranking those in capacity to rule according to their abilities and scope leads to a hierarchical organization, with the orders flowing from the top.

These accounts of the authority relation are more postulates of how the relation ought to be rather than descriptions of what Plato and Aristotle witnessed happening in their times. And in both accounts the authority that an individual has over another is legitimized in a manner that is *external* to the individuals: there is either the sphere of the Good, or the Natural (or the sphere of God as in the case of the early Christian writers) to mediate between the ruler and the ruled. Indeed, in these accounts both ruler and ruled have minimal importance: it is the *solidity* of the external link between them that matters. As an explanation of how the authority relation operates this seems unsatisfactory because it does not provide a causal link through which the process of legitimization takes place. It is Machiavelli who first attempts to describe in a pragmatic manner how the external links stressed by the Greeks play a role in shaping social organization.

Machiavelli begins with a simple observation: that it is not platonic or Christian virtue but the seizure of opportunity that chiefly leads to a position of authority. An accurate reading of the times, Machiavelli would say, a carefully plotted plan, and the unapologetic use of force against the correctly identified enemy, are all that is necessary to position oneself as a ruler. And once in that position it is of capital importance, according to Machiavelli, to establish appropriate *foundations* for one's ruling. Machiavelli presents his ideas through the analysis of an example that continues to be of importance today: the establishment of the Roman republic. As Arendt artfully indicates, Machiavelli 'saw that the whole of Roman history and mentality depended upon the experience of foundation, and he believed it should be possible to repeat the Roman experience' (Arendt 1958, p. 108). What Machiavelli understood was that there is not one single (natural or metaphysical) foundational cosmogony against which all authority relations should be measured, but that these cosmogonies can be crafted to support the authority of a particular individual or group. Once in place, a foundational cosmogony that supports the ruler's actions can be tremendously useful for the ruler, for no individual in a community whose way of thinking is deeply rooted in it could be in a position to challenge the ruler's authority, not even conceptually. For anyone who sees this clearly it becomes no longer possible to understand the legitimacy of an authority relation as externally given. This is indeed what happened during the twentieth century as a deep crisis of authority bred in the midst of serious skepticism regarding the validity of the traditionally accepted cosmogonies.

Max Weber scrutinized the authority relation more as a scholar than as a political strategist. Weber distinguished between forms of authority according to the type of legitimacy underlying each form. According to Weber,

> the validity of the claims to legitimacy may be based on: (1) Rational grounds—resting on a belief in the legality of enacted rules and the right of those elevated to authority under such rules … (2) Traditional grounds—resting on an established belief in the sanctity of immemorial traditions and the legitimacy of those exercising authority under them … or (3) Charismatic grounds—resting on devotion to the exceptional sanctity, heroism or exemplary character of an individual person, and of the normative patterns or order revealed or ordained by [the person] (Weber 1978, pp. 213–14).

The importance of Weber's analysis does not rely on his particular classification of forms of authority but on his emphasis on the study of the legitimacy underlying an authority relation 'within the framework of the concrete actor and his actions' (Easton 1958, p. 74). This perspective is crucial because it is one from which one can begin to understand the

motivations underlying a particular *design* of an authority relation: one is invited by Weber to study the authority relation by focusing on the motivations, desires, and circumstances of those who rule. In a sense, Weber's work picks up the analysis of authority right at the point where Machiavelli left it.

2. *Contemporary Uses*

There is wide disagreement among contemporary scholars as to how the authority relation should be understood. Many even abstain from venturing operational definitions of authority such as the one presented at the beginning of this article. (An example of this is Robert Peabody's 1968 article on authority in the *International Encyclopedia of the Social Sciences*.) Disagreement stems from many sources. Consider the following examples: Arendt believes that the authority relation is an implausible one in a modern society, due to the absence of a unique and uncontradicted cosmogony upon which the legitimacy of a ruler may be founded. William Connolly agrees with Arendt in that a modern understanding of authority is problematic, but argues that the web of conventions regarding how social customs are interpreted and implemented is the modern equivalent of the premodern cosmogony. Writers like Nancy Rosenblum, in turn, argue that women and men understand relations of authority in very different ways. Rosenblum, following Carol Gilligan, argues that 'for women, the typical moral ideal is cooperation and interdependence rather than fulfillment of a formal obligation.' The implication is that 'authority, rules, and procedures will be less imperative for [women] than empathy or the impulse to be responsible for maintaining concrete personal relations' (Rosenblum 1987, pp. 116–17). Lucian Pye has said something of that sort regarding the differences between notions of authority that occur due to cultural differences between countries (Pye 1985). Stanley Milgram and others have studied extensively how deeply rooted the practice of obedience to authority figures is in the context of small groups (a survey of this literature can be found in Hare 1976). As the examples show, one is left not with one but with multiple perspectives for the analysis of the authority relation.

2.1 *The Multiple Perspectives View*

Rather than viewing the multiplicity of perspectives as a nuisance Stephen Lukes organized them in such a way that they, collectively, form an insightful view for the study of authority.

The multiple perspectives view cuts through an old problem in the analysis of authority: that of whether one can ask analytical questions about authority in a relativized manner, that is, independently of normative issues regarding the legitimacy of authority. For some, 'the nonrelativized notion is primary and is presupposed by the relativized notion. On this view, to analyze authority is to analyze legitimate or justified authority' (Lukes 1987, p. 60). For others, the relativized analysis is possible; one thing is to discover and scrutinize instances of obedience and a different one to assess their validity and desirability. From the standpoint of the multiple perspectives view, however, in identifying relations of authority one always operates at the *interpreted* level of the social interaction. Therefore, according to Lukes, 'every way of identifying authority is relative to one or more perspectives, … there is no objective, in the sense of perspective neutral, way of doing so' (Lukes 1987, p. 60). The multiple perspectives view provides a unified foundation for both analytical and normative inquiries.

The multiple perspectives that arise in the present times with regard to the analysis of authority can be associated with the different degrees of solidity attributed to social composites such as culture, language, social groups, and the individual. The previous section presented the view that it was with respect to a solid cosmogony that the premodern authority relation could be understood. With this perspective gone, Weber's analysis relies on the solidity of the internal world of those who occupy the position of the ruler in the relation. But the perspective of the ruler is one of many that can be held. Lukes identified many other relevant perspectives: that of the one who obeys, that of an external observer, that which describes the community as either formally or informally constituted, that describing the collective wisdom of the community, and the *objective* perspective (if it can exist at all) 'from which all other perspectives may be assessed' (Lukes 1987, p. 62). In a world devoid of a unique cosmogony, they may not all be the same.

According to Lukes, the multiple perspectives view shows the need of going beyond the classical Weberian analysis of the authority relation. A command, according to Weber is obeyed 'either through empathy or through inspiration or through persuasion by rational argument' (Weber, quoted by Lukes 1987, p. 63). But Weber never asks the question: What legitimizes the command in the eyes of the one who obeys? This failure to ask the question is unproblematic when the ruler's perspective happens to be shared by rulers and followers alike. While this is a tenable point of view on many occasions, this identification of the perspectives between rulers and followers is anything but solid, as the modern study of rebellion shows. Weber's analysis is therefore deemed incomplete due to his failure to see anything but the perspective of the 'ruling class' as important in understanding authority relations.

Richard Friedman offers an approach that includes the perspectives of both the ruler and the follower. According to Friedman, for there to be a relation of

authority one needs to understand the common ground that exists between a ruler and a follower because it is this common ground that enables the obedience. In Lukes' words: 'Legitimation claimed and the according of legitimacy coincide in a shared recognition of entitlement' (Lukes 1987, p. 65). Lukes sees two problems with this point of view. First, one can think of cases where authority seems to go unrecognized by either the ruler or the follower. Richard Flathman provides a solution to this problem. He suggests that the authority relation belongs to a wider web of practices and beliefs that provide meaning to all relations in a community, even unwittingly (see Lukes 1987, pp. 66–7). The second problem, according to Lukes, is that Friedman gives no indication as to how to distinguish between the common grounds that enable obedience from the ones that do not. In other words, it still leaves unanswered the critical question regarding obedience: What legitimizes the command in the eyes of the one who obeys?

When is, then, authority legitimate? Joseph Raz advances the *dependence thesis*, namely that 'all authoritative directives should be based, in the main, on reasons which already independently apply to the subjects of the directives and are relevant to their action in the circumstances covered by the directive' (Raz 1990, p. 125). In other words, authority is legitimate when there are reasons, from the follower's perspective, that already compel the follower to obey. Obedience is then investigated from the point of view of the motivations, desires, and circumstances of both the ruler and the follower.

2.2 The Game Theoretic View

The point of view put forward by Raz is an advance both on Weber, as it includes the perspective of the follower as well as that of the leader, and on Friedman, as it displays all the elements necessary to investigate the question of legitimacy. To see this it is convenient to adopt the terminology and analysis of *game theory* (for an introduction to this mode of social analysis see Dixit and Skeath 1999). In a basic game theoretic formulation of an interaction between individuals one first begins by identifying the choices available to each, and then by delineating how the situations (that is, the motivations, desires, and circumstances) of all those involved affect each individual's choice. The goal is to be able to understand the choices of all individuals in the interaction by focusing on what beliefs are reasonable for each individual to hold about what will unfold in the interaction given each individual's situation. The beliefs held by the individuals in an interaction are in equilibrium when they are not contradicted by the choices that they justify given each individual's situation. While controversial, the game theoretic mode of analysis is gaining recognition as a valid and powerful mode of social analysis. See Bates et al. (1998) and Binmore (1994, 1998) for examples of game theoretic applications to the fields of political science and political philosophy, respectively. For these authors, social composites such as language, culture, and social groups are as solid as the equilibrium beliefs underlying the individual choices that, collectively, support the social composite in question.

The perspective given by beliefs that are in equilibrium is one with respect to which one can evaluate the plausibility and legitimacy of an authority relation: an individual can be in a position of authority with respect to another individual to the extent that there are equilibrium beliefs that support choices that an analyst of the relation denotes as 'ruling' for the ruler and 'following' for the follower (Zambrano 1999). Hobbes understood this well when he wrote: 'The power of the mighty hath no foundation but in the opinion and belief of the people' (Hobbes, quoted by Binmore 1998, p. 457). Equilibrium beliefs: (a) are at the interpreted level of the social interaction (Lukes); (b) serve as the common ground that sustains an authority relation (Friedman); (c) are determined with respect to both the ruler and the follower's situation (Raz); and (d) are part of the web of practices and beliefs that provide meaning to all relations in a community (Flathman).

Now, the problem is: 'how are we to ascertain what the reasons that apply to authority's subjects are and in what 'success' in acting on them or guiding us to them consists?' (Lukes 1987, p. 69) In other words, can we recover equilibrium beliefs that legitimize the authority relation by looking at the situation of the individuals involved? The answer seems to be: in general one could recover *many* such equilibria, not simply one. How is one to decide on which leads to the right justification? According to Lukes, the answer seems to be: there is no right justification. Different justifications suggest themselves depending on the perspective that one uses to interpret the authority relation. *There is no single social theory of authority*; there are only multiple perspectives that are plausible and valid to the extent to which they are equilibrium beliefs that legitimize the authority relations in a community. In Michel Foucault's words: 'Each society has its regime of truth, its 'general politics' of truth: that is, the types of discourse which it accepts and makes function as true' (Foucault 1980, p. 131).

The game theoretic analysis of authority helps us understand the kinds of belief systems that could support the authority of one individual over another. But there where its strength lies, lies its weakness as well. As mentioned before, in this mode of analysis one evaluates the plausibility of a belief system by checking its consistency with the choices that it justifies given each individual's situation. There are two problems with this. First, the analysis is silent as to how beliefs configure themselves in equilibrium proportions. One is at the mercy of history, as it perhaps should be, to teach us about how different institutions of authority

have emerged in different cultures. Part of the narrative of the evolution of authority can be thought of as occurring spontaneously (Aristotle), but part of its narrative can only be understood with respect to a design functional to some (Machiavelli). Game theory provides only a language in which the story can be parsimoniously told and formally analyzed.

2.3 The Cognitive Science View

The second problem is more severe. Just as recent social science challenges the solidity of social composites such as truth, language, culture, social groups, and authority (as seen in this article), recent research on cognitive science challenges the solidity of individuals themselves. As Francisco Varela puts it:

> What we call 'I' can be analyzed as arising out of our recursive linguistic abilities and their unique capacity for self-description and narration … the selfless 'I' is a bridge between the corporeal body which is common to all beings with nervous systems and the social dynamics in which humans live. My 'I' is neither private nor public alone, but partakes of both. And so do the kinds of narratives that go with it, such as values, habits and preferences (Varela 1999, p. 62).

This poses problems for a game theoretic understanding of authority, since this mode of analysis is grounded on the solidity of the individual and his or her cognitive situation.

Notice, however, the similarities between both approaches: the self, just as the other social composites, arises at the interpreted level of a collection of behaviors centered on a single body and its social history. Varela himself points out this similarity: 'whenever we find regularities such as laws or social roles and conceive of them as externally given, we have succumbed to the fallacy of attributing substantial identity to what [like the self] is really an emergent property of a complex, distributed process mediated by social interactions' (Varela 1999, p. 62). For Varela, the self, and the authority of a self over another self, are notions that give meaning to each other through their arising in the context of a collection of bodies and social histories in a way that preserves the identity of the selves and of the relations between them. In other words, they are equilibrium beliefs of a larger game, with respect to which the actions of all can be legitimized; and those actions, in turn, reinforce the apparent solidity of the authority link, and of the individuals themselves.

See also: Aristotle (384–322 BC); Cognitive Science: Philosophical Aspects; Collective Beliefs: Sociological Explanation; Conventions and Norms: Philosophical Aspects; Game Theory: Noncooperative Games; Legitimacy: Political; Legitimacy, Sociology of; Power in Society; Power: Political; Rational Choice Explanation: Philosophical Aspects; Rational Choice in Politics; Weber, Max (1864–1920)

Bibliography

Arendt H 1958 What was authority. In: Friedrich C (ed.) *NOMOS I: Authority*. Harvard University Press, Cambridge, MA, pp. 81–112

Aristotle 1992 *The Politics*. Penguin Classics, London

Bates R, Greif A, Levi M, Rosenthal J, Weingast B 1998 *Analytic Narratives*. Princeton University Press, Princeton, NJ

Binmore K 1994 *Game Theory and The Social Contract, vol. 1: Playing Fair*. MIT Press, Cambridge, MA

Binmore K 1998 *Game Theory and The Social Contract, vol. 2: Just Playing*. MIT Press, Cambridge, MA

Connolly W 1987 Modern authority and ambiguity. In: Pennock J, Chapman J (eds.) *NOMOS XXIX: Authority Revisited*. New York University Press, New York, pp. 9–27

Dixit A, Skeath S 1999 *Games of Strategy*. Norton, New York

Easton D 1958 The perception of authority and political change. In: Friedrich C (ed.) *NOMOS I: Authority*. Harvard University Press, Cambridge, MA, pp. 170–96

Flathman R 1980 *The Practice of Political Authority*. University of Chicago Press, Chicago

Foucault M 1980 *Power/Knowledge*. Pantheon Books, New York

Friedman R 1990 On the concept of authority in political philosophy. In: Raz J (ed.) *Authority*. New York University Press, New York, pp. 56–91

Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA

Hare A 1976 *Handbook of Small Group Research*, 2nd edn. The Free Press, New York

Lukes S 1987 Perspectives on authority. In: Pennock J, Chapman J (eds.) *NOMOS XXIX: Authority Revisited*. New York University Press, New York, pp. 59–75

Machiavelli N 1999 *The Prince*. Penguin Putnam, New York

Peabody R 1968 Authority. In: Sills D (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York

Plato 1945 *The Republic*. Oxford University Press, Oxford, UK, pp. 473–7

Pye L 1985 *Asian Power and Politics: The Cultural Dimensions of Authority*. Harvard University Press, Cambridge, MA

Raz J 1990 Authority and justification. In: Raz J (ed.) *Authority*. New York University Press, New York, pp. 1–19

Rosenblum N 1987 Studying authority: Keeping pluralism in mind. In: Pennock J, Chapman J (eds.) *NOMOS XXIX: Authority Revisited*. New York University Press, New York, pp. 102–30

Varela F 1999 *Ethical Know-How: Action, Wisdom and Cognition*. Stanford University Press, Stanford, CA

Weber M 1978 *Economy and Society*. University of California Press, Berkeley, CA

Zambrano E 1999 Formal models of authority: Introduction and political economy applications. *Rationality and Society* **11**: 115–38

E. Zambrano

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Autism, Neural Basis of

Autism is recognized as a genetic disorder with a profound impact on the development of the central nervous system. Affected individuals display a variety of behavioral and neuropsychological deficits. Autism strikes before the age of three years, and it affects boys at least four times as often as girls. Its early onset, symptom profile, familial pattern, and chronicity strongly argue for a biological basis, and in fact, there is now substantial data implicating core biological mechanisms. However, there is still no biological marker that is pathognomonic of autism, that is, invariantly present in all cases, but absent in the normal population. Scientific progress has been slowed by changing diagnostic standards during the last two decades of the twentieth century, making it difficult to accumulate a coherent corpus of knowledge, and by clinical heterogeneity that appears to be inherent to the disorder. Current conceptualizations emphasize a broader autism phenotype that includes Asperger syndrome and subsyndromal variants. It is not known whether these autism spectrum conditions share common neural foundations, but given their highly overlapping symptomatology, it is generally assumed that they do.

Research with modern neuroimaging techniques such as magnetic resonance imaging (MRI), and positron emission tomography (PET) is beginning to map out the neural systems affected by autism. However, this work is still in its infancy. Functional MRI (fMRI), in particular, promises to speed the discovery process, because it is noninvasive and applicable to children of all ages. Recent progress has been made in identifying select brain systems that appear as part of the pathophysiology. These include brain areas responsible for emotional and social functions, perceptual systems specific to face and affect recognition, and cognitive systems involved in understanding the intentions of others. Working models of the pathophysiology of autism commonly include the amygdala as a key component of a distributed cortical-subcortical system, but other competing models exist. Moreover, attention is now focused on the earliest stages of child development, and many now believe that a set of initial neural deficits may be compounded by transactions with the social environment during the first years of life, thereby creating a more widespread neural pathology.

1. *Core and Associated Features*

Autism was first described by Leo Kanner, a psychiatrist at The Johns Hopkins University, in 1943. Although his patients suffered from a multitude of problems, Kanner's description emphasized the social and emotional features of the disorder. He described persons with autism as suffering from an inborn deficit in their ability to make 'affective contact' with other people. The prevailing zeitgeist in the years after Kanner's report provided fertile ground for unwarranted speculation about cold and distant parenting styles causing autism (i.e., the 'refrigerator mother' theory). Today autism is recognized as a biologically based disorder.

Modern conceptualizations continue to highlight the social–emotional impairments in autism as the core deficit. European and American diagnostic criteria for autism require a triad of deficits that include developmental problems with communication, reciprocal social interaction and the presence of repetitive, rigid and odd behaviors. Of these three, the profound deficit in social functioning is the unique feature that distinguishes autism from other neuropsychiatric disorders.

1.1 *Developmental Course and Neural Hypotheses*

The social, language, and behavioral problems that occur with autism suggest that the syndrome has affected a functionally diverse and widely distributed set of neural systems. At the same time, however, the affected systems must be discrete, because autism spares many perceptual and cognitive systems; autism is not incompatible with superior skills in many areas, including the visual arts, music and even general intelligence. Even though the full syndrome likely involves 'hits' to multiple systems, it remains possible that the initial insult is localized. Attempts to explain the neural basis of autism should take close account of the developmental course. Autism is never identified at birth. In fact, the large majority of cases go unnoticed during the first year of life. One possibility is that the emergence of the syndrome toward the end of the first year and into the second is an unmasking of a complex set of congenital liabilities. However, the rule of parsimony demands that simpler alternatives first be considered. There is important evidence on this score from animal models of autism. These show that an initial insult to one system can propagate to affect other systems during the earliest months and years of development. These two developmental possibilities are currently under active investigation.

2. *Etiology*

2.1 *Genetic Factors*

Autism is a familial disorder. Statistical modeling of twin and family data suggests that autism is one of the most strongly genetic of all the neuropsychiatric disorders. Specific susceptibility genes, however, have yet to be discovered. When autism is more broadly

defined, identical twins show a concordance rate of about 90 percent, compared to a concordance rate in fraternal twins of 10 percent or less. Autism is thought to be caused by a small number of genes acting in concert, rather than any single gene. Preliminary data from family linkage studies suggest that susceptibility genes for autism may lie on chromosome 7, with some other evidence also pointing to chromosome 15.

2.2 Medical Conditions Associated with Autism

In addition to polygenic pathways to autism, perhaps 10 percent of all cases are caused by a known medical condition, such as Tuberous Sclerosis, Fragile X syndrome, and prenatal illnesses (e.g., rubella). While children with autism are at a somewhat greater risk to endure pre- and peri-natal hardships, this mechanism explains only a tiny fraction of the cases. Moreover, autism is associated with other conditions that are not implicated as causes of the disorder, and that are not mandatory features of the disorder. For example, about three-quarters of individuals with autism are mentally retarded, and many never develop functional language. About half of all lower functioning persons with autism also develop a seizure disorder, often around adolescence, a time when neural connections are typically instantiated through a final wave of neural pruning. Some individuals, however, have normal or even superior intellectual development, and thus general intellectual impairment is not a necessary component. General intellectual impairment, therefore, can be a barrier to specifying the neural basis of the disorder. Moreover, those with high IQs are not necessarily less impaired with regard to the defining triad of symptoms. The high rates of mental retardation and seizure disorder, nevertheless, demand a neural explanation and it suggests that the pathobiological mechanism of autism is such that 'collateral' damage is common.

3. Neural Theories

Nearly every neural system in the brain has been proposed at some point as the cause of autism. Theories typically derive from beliefs about the most salient behavioral and psychological features of the disorder. For example, those who emphasize difficulty with complex information processing as the principle characteristic of autism, postulate widespread cortical abnormalities sparing early sensory processes as the neural basis of autism. Alternatively, those who focus on the emotional deficits and their role in social difficulties often highlight the limbic system in the pathogenesis of autism. Currently, the research data suggest that select aspects of the temporal and frontal lobes, and portions of the amygdala are key nodes in systems affected by autism. These data are not yet specific enough to point to any one theory in clear favor over others.

Before reviewing the research evidence, it is important to note that clinical examination and laboratory tests have mixed success in identifying anomalies. Routine clinical brain scans, including MRIs to picture brain anatomy and Single Positron Emission Computed Tomography (SPECT) to assess brain metabolism typically fail to document neurological problems in the vast majority of patients. It is likely that the detection rate can be improved once research better defines the specific systems affected, allowing implementation specific image acquisition and analysis protocols in clinical practice. On the other hand, up to one third of all persons with autism develop seizures at some time in their life, often in adolescence, and these can be documented with electroencephalography (EEG). Absent of epilepsy, the majority of persons with autism have abnormal EEGs. The abnormalities are often bilateral without a consistent regional focus. Moreover, EEG evoked potential studies consistently show deficits in cognitive information processing components, such as the P3, but intact early components. These results have been interpreted to mean that autism is a disorder of complex information processing that spares early perceptual processes.

3.1 Brain Size, Functional Connectivity, and Cortical Morphology

One of the more intriguing findings to emerge in the past few years is that overall brain size appears to be increased in autism (by about 5 to 10 percent on average). It is not yet understood whether all brain regions and systems are equally affected by the expansion. It also is not clear how whole brain enlargement would serve as a risk factor for autism; it could be merely a marker for a disturbance in the fine structure of the brain that actually causes autistic symptoms. Increased brain size might come at the expense of interconnectivity between specialized neural systems, giving rise to a more fragmented processing structure. In fact, some evidence suggests that the corpus callosum, the major fiber pathway between the hemispheres, is reduced in size, in autism. Moreover, one PET study found a reduction in coordinated brain activity. Less neural integration would be consistent with Uta Frith's (1989) influential theory that attributes autistic symptoms to a lack of 'central coherence,' a processing style that makes integration of parts into wholes problematic (see *Autistic Disorder: Psychological*).

Debate continues as to whether the growth abnormality is postnatal or prenatal. Specifying the developmental epoch with the most abnormal growth rate would provide better clues as to the underlying mechanism. An origin at particular times of fetal brain

development, could suggest, for instance, disturbances in the regulation of neuronal or glial cell proliferation, neuronal migration, or apoptosis (programmed cell death). Prenatal origins of disturbed brain development have been suggested by studies finding an increased frequency of morphological abnormalities of the cerebral cortex in autistic individuals (e.g., regional alterations in the size and number of gyri). Such abnormalities stem from disturbances in neuronal migration during fetal brain development. These gross neuroanatomical abnormalities are much more common among autistic individuals with mental retardation, but still occur in a minority of cases. Thus, these findings do not appear to be specific to the core social deficit in autism.

3.2 Posterior Fossa Abnormalities

Postmortem studies of a small number of persons with autism have revealed a range of abnormalities, including a significant decrease in the number of Purkinje cells and granule cells in the cerebellum (see Bauman and Kemper 1994). The precise nature of these abnormalities, including a lack of gliosis indicative of scarring, suggests a prenatal origin. A focus on the cerebellum would be consistent with some neuroimaging evidence. A variety of posterior fossa abnormalities have been reported in autism from MRI. These include abnormalities of the pons, fourth ventricle, and cerebellar vermis, the midline portion of the cerebellum. One influential set of findings ties autism to hypoplasia of the neocerebellar vermis, but this abnormality has not consistently been observed across studies. Moreover, it seems likely that posterior fossa abnormalities are not specific to autism, but rather evident in many persons with developmental disabilities and mental retardation. Thus, specificity of the findings for the core autistic features seems doubtful.

3.3 The Amygdala and the Temporal Lobe

Of the specific brain regions implicated in the pathobiology of autism spectrum conditions, none has attracted as much interest as the limbic system, especially the amygdala and its functional partners in the temporal and frontal cortices. The limbic system lies largely within the medial and ventral aspect of the temporal lobe, providing a girdle around the phylogenetically older, deep brain structures. The amygdala, in particular, plays a critical role in emotional arousal, assigning significance to environmental stimuli, and mediating the formation of visual-reward associations or 'emotional' learning.

Postmortem examination of the brains of persons with autism finds consistent evidence for abnormalities in size, density, and dendritic arborization of neurons in the limbic system, including the amygdala, hippocampus, septum, anterior cingulate, and mammillary bodies. There is a stunting of neuronal processes and an increased neuronal packing density, suggesting a curtailment of normal development. These affected regions are strongly interconnected, and together they comprise the majority of the limbic system. The limbic system, especially the amygdala, is part of a neural structure that supports social and emotional functioning. These postmortem findings, therefore, are often heralded as the first good entrance points for understanding the pathobiology of the autism spectrum disorders.

There is supportive evidence for an amygdala theory of autism from the experimental monkey work of Jocylene Bachevalier and co-workers (Bachevalier 2000). She has produced an animal model of autism by lesioning the amygdalae of monkeys shortly after birth. Across the first year of life, these animals developed patterns of behavior reminiscent of autism, that is, social isolation, lack of eye contact, expressionless faces and motor stereotypies. Similar lesions in adulthood fail to produce autistic-like sequelae. These findings are consistent with the idea that autistic symptoms are in part a function of faulty, early emotional learning mediated by limbic system pathology. Moreover, as monkeys with early lesions to the amygdala and surrounding entorhinal cortex mature into adulthood, additional abnormalities are found in the neurochemistry of the frontal cortex and in the frontal—subcortical regulation of dopaminergic activity. Thus, early discrete damage can produce widespread abnormalities across development.

Persons with autism have deficits in their ability to recognize and discriminate faces, and to understand facial expressions (Klin et al. 1999). Functional neuroimaging and lesion data show that the fusiform gyrus, a region on the underside of the temporal lobe, is normally a nexus area for face perception, while neighboring regions in the posterior aspects of the middle and superior temporal gyri are important for reading facial expressions and social intent through eye gaze direction. Several fMRI studies have now shown hypoactivation of the fusiform gyrus during face perception tasks (e.g., Schultz et al. 2000). Preliminary evidence also links hypoactivation of the amygdala and lateral temporal cortices to autism. One hypothesis is that the principal pathology in autism resides in limbic regions, and that disturbance in affective orientation early in life causes a cascade of neurodevelopmental events, including failure to develop perceptual competence for faces and for visual and auditory displays of emotion.

Other evidence implicating the temporal-amygdala system in autism includes case studies of patients with temporal lobe lesions and autistic-like sequelae and reduced functional activity on SPECT scans in the temporal lobes in autism. An association between tuberous sclerosis and autism has long been recog-

nized. More recently, autism has been specifically related to tubers of the temporal lobe vs. tubers elsewhere in the brain.

3.4 Frontal Lobe Involvement

Aspects of frontal lobe integrity and function have been implicated in the pathogenesis of autism. Older studies using lower resolution neuroimaging techniques reported general hypoactivation of the frontal lobes. Functional neuroimaging data collected in the last decade are converging to show that subregions of the prefrontal cortices with especially strong connectivity to limbic areas are critical for 'social cognition,' that is, thinking about other's thoughts, feelings and intentions. Deficits in such 'theory of mind' abilities are common in autism (see *Autistic Disorder: Psychological*). Theory of mind ability has been linked to functional activity in the medial aspect of the superior frontal gyrus (primarily Brodmann area 9) and to prefrontal cortex immediately above the orbits of the eyes (i.e., orbital frontal cortex). The orbital and medial prefrontal cortices have dense reciprocal connections with the amygdala providing the architecture for a system that could regulate social-emotional processes. Bilateral lesions to orbital and medial prefrontal cortex lesions cause deficits on theory of mind tasks. Moreover, nonhuman primate studies have documented abnormal social responsivity and loss of social position within the group following lesions to orbital and medial prefrontal cortex.

Preliminary functional imaging evidence in autism spectrum conditions suggests altered functional representation in prefrontal cortex regions during theory of mind tasks. Moreover, medial prefrontal dopaminergic activity as measured by flourine-18-labeled fluorodopa PET has been found to be significantly reduced in autism. Reduced glucose metabolism during memory activities has also been reported in a subdivision of the anterior cingulate gyrus, a region that lies along the medial surface of the frontal lobe.

3.5 Hyperserotonemia

The best-documented biological marker of autism is elevated blood level of the neurotransmitter serotonin, found in 25 to 30 percent of the patients. However, the association between autism and platelet hyperserotonemia is complicated by the fact that peripheral and central serotonin levels are independent, with different synthesis and metabolism. Other evidence suggestive of a possible role for serotonin includes the fact that powerful psychiatric drugs that affect the serotonergic system are frequently used in autism with some benefits. However, these medications, known as selective serotonin uptake inhibitors (SSRIs) are not effective for the primary symptoms of autism, but rather are of modest help for some of the secondary symptoms (such as aggression and stereotypies). Moreover, preliminary neuroimaging data from PET suggests that persons with autism may have altered serotonin synthesis capacity; this could disrupt synaptic connectivity in the sensory cortices. Other types of neurochemical analysis in the blood, urine and spinal fluid shows inconsistent abnormalities, and psychopharmacological efforts related to the compounds have generally failed to show much benefit in autism.

4. Future Directions

The major findings to date on the neural basis of autism concern aspects of the limbic system, and functionally related and connected regions of the orbitomedial prefrontal cortex, and visual areas of the temporal lobe. While good progress toward understanding the neural basis of autism has been made in recent years, much work still needs to be done. fMRI is revolutionizing psychiatry and systems level neuroscience, and there is little doubt that this exciting new tool will be a powerful device at the disposal of autism researchers in the years to come. It ultimately should enable researchers to define dynamic brain processes that give rise to each specific symptom and feature of autism.

A challenge for research in this area will be to adapt *in vivo* neuroimaging techniques so that they are applicable to the developing infant and toddler. Studying younger children may be a prerequisite for a comprehensive understanding of the neural basis of autism, because the disorder evolves into its full form in rather a short period of time around the second year of life. Capturing this process would clearly be valuable. Moreover, newer MRI techniques, such as diffusion tensor imaging, which spotlights previously unresolved fiber pathways intermingled in the brain's white matter, could also prove quite valuable for this area of research. It could be especially important for studies of young children, since diffusion imaging is more forgiving of patient movement than fMRI, and does not require compliance with a task.

Finally, with the completion of the first complete mapping of the human genome, the promise of identifying susceptibility genes for autism in the near future seems likely. Genetic advances naturally will lead to novel treatments, some of which will be instituted early in brain development. This will open a host of new research problems and opportunities, including characterizing the effect of genes and gene combinations on the developing brain.

See also: Amygdala (Amygdaloid Complex); Autistic Disorder: Psychological; Developmental Psychopathology: Child Psychology Aspects; Functional Brain

Imaging; Infant and Child Development, Theories of; MRI (Magnetic Resonance Imaging) in Psychiatry; Social Learning, Cognition, and Personality Development

Bibliography

Bachevalier J 2000 The amygdala, social cognition and autism. In: Aggleton J P (ed.) *The Amygdala: A Functional Analysis*. Oxford University Press, Oxford, UK, pp. 509–43

Bauman M L, Kemper T L 1994 Neuroanatomic observations of the brain in autism. In: Baumen M L, Kemper T L (eds.) *The Neurobiology of Autism*. Johns Hopkins University Press, Baltimore, MD, pp. 119–45

Cohen D J, Volkmar F R (eds.) 1997 *Handbook of Autism and Pervasive Developmental Disorders*, 2nd edn. Wiley, New York

Ernst M, Zametkin A J, Matochik J A, Pascualvaca D, Cohen R M 1997 Reduced medial prefrontal dopaminergic activity in autistic children. *Lancet* **350**: 638

Frith U 1989 *Autism: Explaining the Enigma*. Blackwell, Oxford, UK

International Molecular Genetic Study of Autism Consortium 1998 A full genome screen for autism with evidence for linkage to a region on chromosome 7q. *Human Molecular Genetics* **7**(3): 571–8

Kanner L 1943 Autistic disturbances of affective contact. *Nervous Child* **2**: 217–50

Klin A, Sparrow S S, de Bildt A, Cicchetti D V, Cohen D J, Volkmar F R 1999 A normed study of face recognition in autism and related disorders. *Journal of Autism and Developmental Disorders* **29**(6): 499–508

Schultz R T, Gauthier I, Klin A, Fulbright R, Anderson A, Volkmar F, Skudlarski P, Lacadie C, Cohen D J, Gore J C 2000 Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and Asperger syndrome. *Archives of General Psychiatry* **57**: 331–40

R. T. Schultz

Autistic Disorder: Psychological

Autism is a disorder of social and communicative development. Although biologically based, with a strong genetic component, autism is diagnosed by behavioral criteria. Qualitative impairments in social interaction and communication, with rigid and repetitive interests and activities, are necessary features for diagnosis of 'autistic disorder' according to current manuals. This set of core features, however, covers a spectrum of behaviors, and the manifestation of autism varies greatly with age and ability, from the silent, aloof, and developmentally delayed child, to the verbose, overfriendly, and intelligent adult. With recognition of this spectrum, the incidence of autism is estimated at one per 1,000, and the male to female ratio at least 3:1. Most of these individuals will also have general learning difficulties and low IQ, but a striking feature of the disorder is the existence of otherwise highly intelligent people with autism who nevertheless show severe social and communicative impairments and a lack of flexible behavior.

Autism typically is diagnosed around three years or later, although recent research has suggested that screening at 18 months may be possible. Young children with autism typically fail to communicate by showing and pointing, fail to follow others' eye gaze, and do not develop pretend play. Autism lasts throughout life, and although there is no cure, the disorder is not degenerative and developmental progress and compensation can continue into adulthood, especially where specialist educational intervention is given.

1. History

There is little doubt that autism has always existed. Frith, among others, has pointed to evidence from folktales and 'wise fools' to show the presence of people with autistic features throughout history (Frith 1989). Autism was first named, however, by Leo Kanner in 1943, who reported a group of puzzling children who showed as core impairments an 'extreme isolation and the obsessive insistence on the preservation of sameness.'

Following Kanner's first description, autism (then also referred to as childhood schizophrenia or childhood psychosis) quickly became established as an important clinical entity in child psychiatry (see *Child and Adolescent Psychiatry, Principles of*). In the absence of knowledge of biological causes, and prompted perhaps by the normal physical appearance of the children, the first interest in autism focused largely on psychogenic explanations. However, psychological and epidemiological work in the 1960s and 1970s, pioneered by Beate Hermelin and Neil O'Connor, established autism firmly in the discipline of mental handicap, separating it from mental illness in general and from schizophrenia in particular. The idea that you might make your child autistic by being an unloving 'refrigerator mother' has received no support from research.

Epidemiological work in the late 1970s, by Lorna Wing and Judith Gould in the UK, established autism as a true syndrome, marked by a triad of reliably co-occurring impairments in socialization, communication, and imaginative or flexible behavior. They also introduced the notion of an autism spectrum, which included not only children who were withdrawn and silent (like many of Kanner's cases) but also children who showed their social impairment in unusually passive, or in active-but-odd social behavior, and their communication impairment in pedantic, overformal

speech (see Wing 1996). Wing's triad formed the basis for the criteria set out in subsequent diagnostic manuals, and raised awareness that the manifestation of autism varies with age and ability.

2. *Differential Diagnosis and Asperger's Syndrome*

Current diagnostic schemes place autistic disorder within the category of pervasive developmental disorders (disorders characterized by severe impairments in more than one area of development), which also includes Rett's disorder, childhood disintegrative disorder, and Asperger's disorder. Other categories from which autism is distinguished include disorders of receptive language (with secondary emotional/social consequences) and early onset schizophrenia. Current descriptive categories, whose validity remains controversial and whose relation to autism is as yet unknown, include nonverbal learning disability, schizotypal personality disorder, and semantic–pragmatic disorder.

Research recently has focused on the distinction between autism and Asperger's syndrome—a disorder taking its name from Hans Asperger, a Viennese clinician who described children similar to Kanner's cases contemporaneously. The status of Asperger's syndrome as a separate entity is somewhat contentious. However, present consensus is that this syndrome falls at the high ability end of the autism spectrum, but deserves its own label because the course of the disorder, and consequently prognosis and management, may diverge somewhat from typical autism. People with Asperger's syndrome tend to be less aloof, more socially interested, more verbal in a pedantic style, and tend to have special interests to the point of obsession (see Frith 1991). Diagnostic manuals, which introduced 'Asperger's disorder' as a category in the early 1990s, stipulate that cases must show social impairments and restricted interests as for autism, but should not show significant delay in language or cognitive skills (distinguishing them from classic autism). This latter criterion is problematic, in view of developmental change and late diagnosis. The only epidemiological study to date, by Ehlers and Gillberg in Sweden, reflects this difficulty in using strict criteria, but estimated the prevalence of Asperger's syndrome at 3–7 in 1000 school-age children, with a male to female ratio of, perhaps, 8:1.

3. *Etiology*

In contrast to early psychogenic explanations of autism, the disorder is now considered among other organically caused developmental disorders. Around three quarters of individuals with autism also have general developmental delay (i.e., IQ below 70), epilepsy is present in approximately a third of all cases, and a number of different medical disorders are associated with autism (e.g., Fragile-X syndrome, tuberous sclerosis).

Because autism (at least in its nuclear form) is a rare disorder, and because people with autism are unlikely to have children, the familial clustering of autism was at first overlooked. However, siblings, and especially identical twins, of people with autism do have a significantly raised incidence of the disorder. There are suggestions, too, that some relatives may show mild features reminiscent of autism, and these may have prompted the erroneous notion of 'refrigerator' parents causing autism.

Many mothers report that the pregnancy of their child with autism was affected by pre- and perinatal problems. The causal status of these adverse factors is uncertain, however; they may be effects, rather than causes, of some pre-existent abnormality in the child. Various infectious agents (including vaccines) have been suggested to play a role in the etiology of autism, although it appears at present that such agents could account for only a small number of cases.

There has been a great deal of research aiming to discover the brain basis of autism, with biochemical, anatomical, and neurological investigations. Almost every brain region has been proposed, at one time or another, as the site of the core anomaly in autism. Current contenders include the frontal lobes, limbic system, and cerebellum. To date, however, findings are equivocal and the specificity to autism of many of the features must still be in doubt, since appropriate control groups equated for degree of developmental delay, are often missing.

4. *Psychological Theories*

While limited progress has been made in discovering what is different in the brains of those with autism, much more has been discovered about the mind in this condition. The rise of psychological accounts can be traced back to Hermelin and O'Connor, who first contrasted autism with other forms of mental and sensory handicap. Subsequent theoretical approaches to autism have followed O'Connor and Hermelin in looking for what is specific to autism, in contrast to general intellectual impairment. At present, the major psychological accounts can be divided into those that propose a specific social deficit, and those that posit a core non-social anomaly at the root of autism (see Happé 1995).

4.1 *The 'Theory of Mind' Account*

In the 1980s interest swung from investigations of language, perception and memory, to explorations of the social impairments in autism. The task for psycho-

logical theories became to explain the concurrence of imagination, communication, and socialization impairments. Perhaps the most influential of these attempts has been the 'theory of mind' deficit account; the hypothesis that people with autism are unable to represent the mental states (e.g., beliefs, desires) of themselves and others, and to understand and predict behavior in terms of these states. Baron-Cohen, Leslie, and Frith in 1985 proposed such a deficit on the basis of failure on a simple test of theory of mind, or 'mentalizing'; children with autism, unlike normal 4-year-olds or children with Down's syndrome, were unable to predict where a character would look for an object moved in their absence. Instead of taking into account the character's mistaken belief about the object's location, the children with autism answered on the basis of the real state of affairs (see Frith 1989). This failure to attribute mental states independent of reality and of the child's own belief, has now been replicated in a number of studies. Importantly, children and adults with autism succeed on closely matched tasks not requiring theory of mind (see Happé 1995).

The notion of a deficit in mentalizing seems to account well for the triad of impairments. Children with autism lack insightful social reciprocity, but often like company and physical contact. A person with autism may transmit a message accurately, verbatim, but fail to recognize who does and does not already possess the information. Pretend play, in which real and imagined states are contrasted, is strikingly absent in autism, but physical play is seen. A theory of mind deficit thus accounts for the core features of autism, while allowing intact skills in other areas. Whether the mentalizing deficit is itself the result of impairments in earlier social processes (e.g., emotional interaction, imitation, sharing attention), is as yet unclear (see Baron-Cohen et al. 1993).

To date, a delay in theory of mind development appears to be a universal feature of autism. A minority, especially those with Asperger's syndrome, do develop some understanding of other minds in adolescence, perhaps using their general intelligence to puzzle out what others' think and feel. There is considerable interest in trying to teach theory of mind to children with autism, although to date generalization of these taught skills has been limited. However, the use of concrete props and supports, such as cartoon strips and thought bubbles, appears to be effective in helping some children develop greater social understanding (see Howlin 1998).

4.2 Explaining Nonsocial Features

The interest in social impairments in autism during the 1980s left unexplained the nonsocial aspects of the condition; clinical features such narrow interests, insistence on sameness, and special skills. More recently, psychological accounts have attempted to encompass these aspects of autism. One nonsocial feature that forms a diagnostic criterion for autism is restricted and repetitive behavior. Researchers have been struck by the parallel between the perseverative and repetitive behavior in autism and that seen in patients who have suffered injury to the frontal lobes of the brain—and this has given rise to a theory of autism as a disorder of frontal functions. Neuropsychological studies suggest that the frontal lobes are responsible for higher cognitive processes of control, planning and flexible adaptive behavior, so called 'executive functions.' There is evidence that individuals with autism, even those of normal IQ, have problems in at least some of these functions (Russell 1997). Executive function impairments are not, however, specific to people with autism. A variety of other groups with developmental disorders also show these difficulties (e.g., Attention Deficit disorder, Tourette's syndrome). How these disorders—so different in clinical presentation—differ in their executive impairment is as yet unclear.

Problems in executive functions cannot explain all aspects of the nonsocial impairments in autism and, more importantly, cannot explain why people with autism are so good at certain things. From the beginning, autism has been linked with special skills in rote memory, jigsaw-type puzzles, and the 'savant' abilities in calculation, drawing, and music. Savant skills, superior to the individual's other functioning or even to normal performance, occur in approximately 1 in 10 people with autism—much higher than in other groups. No current deficit account of autism can explain why people with autism tend to be so unusually good at certain things.

Different from other theories of autism, in that it does not present a deficit account, is the notion of weak 'central coherence.' It is striking that normal and mentally handicapped groups process meaningful and patterned information far better than random and meaningless stimuli. This benefit from meaning, which Frith (1989) termed 'central coherence,' appears to be reduced in autism. For example, while ordinary children recall meaningful sentences better than random word strings, children with autism are almost as good at recalling the latter as the former. In the same way, while normal subjects typically extract the gist of a passage or story while forgetting the details, people with autism may retain the actual words used but fail to extract the meaning.

The idea that people with autism make relatively little use of context and pay preferential attention to parts rather than the whole, can go some way towards explaining the assets seen in autism, as well as some of the deficits. So, for example, people with autism may be good at certain sorts of jigsaw-type tests because they can ignore the whole picture and focus on the specific details. Savant skill in music may begin with a tendency towards perfect pitch, as a result of attention

to details (actual pitch) rather than wholes (melody, irrespective of specific pitch). Frith and Happé have suggested that this 'weak central coherence' has both benefits and disadvantages, and as such is better characterised as a cognitive style rather than deficit. Indeed, it may be this aspect of autism that is transmitted genetically—some fathers of children with autism also excel on visual or verbal tasks where detail-focus is advantageous (Happé 1995).

It remains to be seen how these three psychological accounts fit together, or which anomaly is primary. Possible causal relations between executive functions, social insight, and cognitive style are being investigated. It remains a possibility, however, that autism is a disorder of multiple primary psychological deficits, or that different subgroups within the spectrum show different core deficits.

5. *Treatments and 'Cures'*

Given the devastating impact which autism can have on the individual and family, it is not surprising that every year sees a new claim for a miracle cure, be it biochemical or behavioral. Such claims have so far failed to find objective support, although significant improvement and compensation can occur. Programs of behavioral management can be successful in controlling problem behavior, and educational approaches that give structure and concrete prompts in a stepwise program are demonstrably effective. Use of visual material has proved helpful for the many children with little or no language, and self-injurious behavior and distress can be reduced by making environmental events and social demands easier to understand and predict (Howlin 1998).

A feature of the 1990s has been the recognition of autism in very able people, and the introduction of the Asperger's syndrome diagnosis. This has led to the emergence of groups of highly articulate individuals who can speak about their autism. This is a new voice, which in many cases challenges the 'neurotypical' perspective; some individuals with autism/Asperger's syndrome would strongly resist the notion of 'cure.' Clearly there is a considerable gray area between clinical-level difficulties and personality type—it is likely that a great many people with Asperger's syndrome never come to clinical attention.

6. *Impact and Importance*

In the last quarter of the twentieth century autism has attracted more media interest and research attention than would be warranted simply on the basis of its incidence in the population. This, no doubt, has a number of causes; at a personal level, autism challenges us to look at what we consider fundamentally human—social interaction, communication, and imagination. Autism also tests a number of mainstream psychological notions. The existence of savant skills, for example, in otherwise severely intellectually impaired individuals challenges our notions of general intelligence. The fact that people with autism may, on the other hand, have very high IQ while lacking social insight and a sort of cultural common sense, suggests that 'theory of mind,' and perhaps something like social intelligence, is quite distinct from general reasoning ability and knowledge acquisition. Thus, autism has become a test case for a number of theories of normal development.

7. *Future Directions*

Autism is the focus for research using cutting-edge scientific methods. The search is on for genes acting in the disorder. Advising families on the meaning of genetic susceptibility will, however, be problematic. At present we do not know why one child develops severe autism requiring life-long care, while another becomes an intelligent but socially impaired adult, whom colleagues might merely call eccentric.

Sophisticated brain imaging technology will give us pictures of brain functioning, to compare the developmental trajectory in autism and normal development. This may highlight brain regions for further biochemical investigation and, ultimately, connection to genetic anomalies. In the meantime, educational approaches are the most likely source of hope, with earlier identification and intervention reducing the vicious cycle in which children with autism retreat from an incomprehensible social world, thus losing opportunity to learn about minds and feelings. Finally, recognition of the unique skills and perspective of people with Asperger's syndrome should bring an appreciation of the positive aspects of autism, and the value and rights of people across the autism spectrum.

See also: Autism, Neural Basis of

Bibliography

Baron-Cohen S, Tager-Flusberg H, Cohen D J (eds.) 1993 *Understanding Other Minds: Perspectives from Autism*. Oxford University Press, Oxford, UK

Frith U 1989 *Autism: Explaining the Enigma*. Blackwell, Oxford, UK

Frith U (ed.) 1991 *Autism and Asperger Syndrome*. Cambridge University Press, Cambridge, UK

Happé F 1995 *Autism: An Introduction to Psychological Theory*. Harvard University Press, Cambridge, MA

Howlin P 1998 *Children with Autism and Asperger Syndrome: A Guide for Practitioners and Carers*. Wiley, Chichester, UK

Kanner L 1973 *Childhood Psychosis: Initial Studies and New Insights*. V H Winston, Washington, DC
Russell J ed. 1997 *Autism as an Executive Disorder*. Oxford University Press, Oxford, UK
Schopler E, Mesibov G B, Kunce L J (eds.) 1998 *Asperger Syndrome or High-Functioning Autism?* Plenum, New York
Sigman M, Capps L 1997 *Children with Autism: A Developmental Perspective*. Harvard University Press, Cambridge, MA
Wing L 1996 *The Autistic Spectrum: A Guide for Parents and Professionals*. Constable, London

F. Happé

Automaticity of Action, Psychology of

Automatic thoughts and behaviors are ones that occur efficiently, without the need for conscious guidance or monitoring. Most of our thoughts and behaviors tend to be automatic or have automatic components, and for good reason. These processes are fast, allowing us to do things like drive to work without having to think about how to turn the steering wheel each time we get into a car. There are two main categories of automaticity defined by how the thought or behavior is initiated: Some automatic processes are triggered quite unconsciously, often by stimuli in the environment, whereas others require a conscious act of will to get started.

1. Unconscious Automaticity

Some automatic processes do not require any willful initiation and operate quite independently of conscious control. These processes can be instigated by stimuli of which we are not yet conscious, or by stimuli of which we were recently conscious but are no longer (Bargh 1994). Research has often used *priming* as a technique to trigger these automatic processes. A prime is a stimulus that biases further processing of the same or related material. An everyday example might be buying Tide laundry detergent after having recently seen a nature program about the ocean. Thoughts of the ocean may have primed you to choose Tide, perhaps without any conscious knowledge of the connection.

In a study by Bargh et al. (1996), some participants solved scrambled sentences containing words related to the concept of elderly (e.g., Florida, gray, wrinkles), while other participants solved sentences with neutral words. Each participant was then surreptitiously timed walking down a hallway on the way out of the experiment. The researchers wanted to test whether priming participants with the concept of elderly would automatically and unconsciously change their behavior to become more like that of the elderly. They found that participants who were primed with the elderly concept walked out of the experiment more slowly than the other participants. Careful questioning immediately afterwards revealed that the participants were not conscious of the concept of the elderly or of their reaction to it. In a related study, Dijksterhuis and Van Knippenberg (1998) found that priming the concept 'professor' made participants more successful at answering trivia questions compared to participants who were primed with 'soccer hooligan.' And as in the previous study, participants were unaware that the prime had affected their behavior.

Much of our behavior in social life is unconsciously automatic. There is evidence that people can respond automatically and unthinkingly to facial expressions, body gestures, hints about a person's sex, ethnicity, or sexual orientation, information about someone's hostility or cooperativeness, and a variety of other social stimuli (Wegner and Bargh 1998). People also have unconscious automatic responses to things they like and dislike, from foods or books to ideas and social groups. Although people may have conscious responses to all these items as well, this rich and detailed array of unconscious automatic responses provides a background of reaction to the social world. When we do not have time, inclination, or the ability to study or consciously correct these reactions, we may still find that we are behaving quite satisfactorily on 'autopilot' nonetheless.

2. Conscious Automaticity

Many of the automatic behaviors we do every day are things of which we are perfectly aware—at the outset. We know we are getting in the car and heading off to work, for instance, or we know we are beginning to take a shower. Yet because we have done the act so often—driving to work every day, showering every darn year, whether we need it or not—we no longer need to think about the act after we have consciously launched it. These behaviors are often acquired skills, actions that become automatic only after significant repetition.

When we begin to learn an action, such as driving, we think of the action at a very detailed level (Vallacher and Wegner 1987). We think 'engage clutch, move gear shift down into second, lift left foot off the clutch and right foot onto gas.' Skill acquisition starts off as labored, conscious learning and after consistent, frequent practice becomes more automatic and unconscious. Once the action is well learned, the behavior becomes automatic in the sense that it does not require constant conscious monitoring. This automaticity allows us no longer to think about the details, and instead to think about the act at a higher level ('I am driving to work. Gosh.'). It is as though practice leads to a mental repackaging of our behavior, a chunking together of formerly stray details into a fluid sequence

that can then be set off with only a brief conscious thought rather than a continuing commentary of them. Once the conscious decision is made to drive to work, the drive itself can be quite unconscious and automatic—as we chat on the cell phone along the way—and we may remember very little of the experience once we arrive at our destination.

When we have conscious thoughts prior to our behaviors, we typically experience these behaviors as willed. So, even though 'driving to work' is largely automatic throughout its course, the fact that we thought of doing it just before it started makes us interpret the entire sequence as consciously caused. However, the more frequently we notice our intentions occurring as we act, the more we experience behavior as consciously willed and nonautomatic. If we do something that requires a lot of thinking (such as a difficult math problem, or driving when we don't know how), for example, we are more likely to feel that we have consciously willed what we have done. Behaviors that happen without any conscious thoughts at all, in turn, are not likely to be experienced as willed. Although it is common to assume that automatic behavior is the opposite of consciously controlled behavior, this analysis suggests that automaticity can characterize both behaviors we experience as consciously caused and those we experience as involuntary (Wegner and Wheatley 1999).

3. *Benefits and Costs of Automaticity*

Automatic processes do not need constant conscious guidance or monitoring, and therefore use minimal attention capacity. For this reason, they are very fast and efficient. Sometimes we might wish our automatic actions or reactions were different, such as when we mindlessly say 'fine' after a waiter asks about our inedible meal. Metaphorically speaking, it is as if the waiter had come out with a little rubber hammer and struck just below our knee. Such automatic behaviors are so often mindless that they can pop out in inappropriate contexts. For the most part, however, the fact that many of our behaviors become automatic is extremely beneficial. If all our actions required conscious thought, we would spend time planning every step instead of just 'walking.' Everything would take as much time and be as difficult to do as the first time we did it.

Automaticity allows a familiar and comfortable interaction with our environments. With experience, we learn what is likely to happen in different situations. When we walk into a grocery store, we know automatically how things are supposed to go. We go in, grab a cart, pick food off the shelf, line up for a cashier who will take our money for the food, and we can go home. It is not as if we walk into the store and think 'OK, what happened the last time I was here' or 'Why are people looting food off the shelves?' We automatically know the proper assumptions of the situation based on our experience. This automatic activation of norms makes the world a much more predictable place. We are thus free to think about Bob's annoying table manners and Jane's infectious laugh as we wander down the aisles, selecting all the necessary ingredients for the dinner party the next night.

It is the very ease and fluency of automatic thought and behavior, however, that brings with it important costs. One such pitfall comes from thinking about things the same way over and over again such that a particular way of thinking becomes the default. For example, if you learn that black men are not only male and black, but may be hostile and lazy, your responses to a particular black man could be determined by automatic processes quite beyond your conscious control. You could hate him or avoid him or treat him poorly without any knowledge of his actual characteristics. Automatic responses to people and groups may be based on stereotypes—characterizations of persons based on their membership of a particular group (e.g., Asian, Jewish, basketball player, etc.). Stereotypic ideas may be so well learned that they pop into mind automatically.

If there is plenty of time to think, as well as no distraction, a stereotype that pops into mind does not always have to be acted upon and can be corrected. Gilbert's two-factor theory of attribution suggests that automatic attributions of why someone behaved a certain way tend to be dispositional in nature (e.g., thinking someone is lazy because they are watching TV). However, with enough mental resources, we can correct those attributions for situational causes (e.g., realizing that the person had a hard day and is trying to unwind; see Gilbert and Malone 1995, for a review). In the same way, stereotypes may be automatically activated but can be countered by consciously thinking about why that stereotype is false, about other characteristics of the person that do not fit the stereotype, or about explanations that take into account the person's situation.

The attempt not to think about a stereotype can, however, ironically make that stereotype come more readily to mind. This is because weak yet measurable automatic processes regularly arise to monitor the failure of conscious intentions. When a person tries not to think about a white bear, for example, thoughts of the white bear are likely to come back repeatedly despite the attempted control. The theory of ironic processes of mental control (Wegner 1994) suggests that such ironic processes are produced whenever people try to control their thoughts—and particularly when they do so under conditions of stress or mental load. These processes are required to search for the failure of mental control and reinstate the control process when this is necessary—but they also introduce an unfortunate sensitivity to the very thoughts the person desires to suppress.

This ironic effect on stereotyping has been observed in experiments by Macrae et al. (1994). These researchers asked participants to suppress stereotype thoughts in imagining the life of a person belonging to a stereotyped group (a 'skinhead'), and then later gave these participants the opportunity to write their impressions of another person of this group. As compared to the impressions of participants who did not first suppress stereotyping, these participants formed more stereotypical impressions of the second target. Another study examined the effects of this manipulation on participants' choices of how close to sit to a target just after having controlled their stereotypes of the target in an earlier impression-formation session. Participants instructed to suppress stereotyping succeeded in creating less stereotypical imaginings about the target, but they subsequently chose to sit at a greater distance from the target than did other participants who had not been instructed to suppress the stereotype.

4. *Summary*

The automaticity of social thought and behavior is both a blessing and a curse. On the blessing side, our ability to respond unconsciously and effortlessly to a range of social settings, people, and events allows us the luxury of speedy responses that are largely appropriate. And because the conscious initiation and practice of responses can shape them yet further, we can, over the course of interaction, become skilled social agents who can interpret and react to social settings with remarkable aplomb. The curse of automaticity inheres in the lack of flexibility and control that results when we learn things too well and are not conscious of doing them. We may make maladaptive or immoral unconscious responses that we then regret or simply fail to notice. And we may find under conditions of mental load or stress that the automatic processes that occur to monitor the failure of our conscious intentions ironically create that failure. When this happens, we find ourselves thinking or acting in social situations in precisely the ways we wish we would not.

5. *Future Directions*

Automaticity researchers have just begun to examine the underlying brain mechanisms associated with automatic and controlled processes. By studying these mechanisms, we may better understand *how* thoughts and behaviors become automatic, and what brain systems underlie automatic versus consciously controlled thoughts and behaviors.

Wegner's ironic-process model is one model of how unwanted automatic thoughts may be generated and influenced by controlled processes. Brain-imaging techniques offer direct testing of such models with the goal of understanding how automatic and controlled processes influence each other. For example, conscious deliberation may be most effective at determining what becomes an automatic process but less effective at influencing deeply ingrained automatic processes. Brain imaging may be a useful tool to shed light on which processes are likely to be automatic from their inception, when processes cross the threshold between control and automaticity, and how that crossover can occur.

See also: Action Planning, Psychology of; Attention and Action; Heuristics in Social Cognition; Motivation and Actions, Psychology of; Schemas, Social Psychology of; Stereotypes, Social Psychology of

Bibliography

Bargh J A 1994 The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In: Wyer Jr. R S, Srull T K (eds.) *Handbook of Social Cognition*, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, NJ, Vol. 1, pp. 1–40

Bargh J A, Chen M, Burrows L 1996 Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology* **71**: 230–44

Dijksterhuis A, Van Knippenberg A 1998 The relation between perception and behavior, or how to win a game of trivial pursuit. *Journal of Personality and Social Psychology* **74**: 865–77

Gilbert D T, Malone P S 1995 The correspondence bias. *Psychological Bulletin* **117**: 21–38

Macrae C N, Bodenhausen G V, Milne A B, Jetten J 1994 Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology* **67**: 808–17

Vallacher R R, Wegner D M 1987 What do people think they're doing? Action identification and human behavior. *Psychological Review* **94**: 3–15

Wegner D M 1994 Ironic processes of mental control. *Psychological Review* **101**: 34–52

Wegner D M, Bargh J A 1998 Control and automaticity in social life. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, Vol. 1, pp. 446–96

Wegner D M, Wheatley T 1999 Apparent mental causation: Sources of the experience of will. *American Psychologist* **54**: 480–92

T. Wheatley and D. M. Wegner

Automation: Organizational Studies

1. *The Concept of Automation*

The word automation, as a contraction of automatic production, was first used in both by John Diebold—author of the book *Automation: The Advent of Automatic Factory* (1952)—and by D. S. Harder, vice-president of manufacturing at Ford Motor Company.

According to Gallino ( ), automation as a popular term

> indicates different situations where the human physical and intellectual work is replaced by machines and by mechanical, hydraulic, pneumatic, electrical, electronic servomechanisms which may automatically perform sequences of operations

The main components of automation indicated by most authors are:

(a) control systems based upon closed-loop control mechanism of feedback (on standards) and feed forward (on goals or week signals), performed by any kind of technological device;

(b) integration of different devices, processes into a unitarian architecture at the level of a factory, firm, network, achieving continuity of processes and management control; and

(c) system adaptation and innovation, through rapid detection both of the internal state of the system and of the environment (technical, economic, commercial, etc).

In any case, those components include tight collaboration among Information and Communication Technologies, advanced managerial systems, and operative work.

Applications and literature indicate that automation is in all cases a dynamic process. In fact:

(a) different levels of automation depend upon the quantity of human tasks embedded in the technical system,

(b) different patterns of automation result from forms and degrees of collaboration between the technical and the human system, and

(c) systems are developed along discrete stages and through change management processes.

Essentials of automation are:

(a) the need for human work because without it there would be only automatic machines;

(b) the need for collaboration among men and machines because, without interrelated activities, there is not a man-machine system; and

(c) integration of technological, fiscal, organizational, social system adopted in any single case.

As a result of 50 years of discussions about a changing phenomenon, we define automation as a stage in the process towards integrated systems of processes, technology, organization, roles, and values where technology performs a large variety of existing and new tasks, while cooperation is designed among men and technical systems with the goal of achieving optimal products and services.

2. *Applications*

The essence of automation is automatic control. This concept has been developed as an academic discipline and as a set of technologies applied in a large variety of intelligent human artifacts, that is, machines that can read complex data, signals, and symbols, have feedback loop and feed-forward mechanisms, interpret external and internal variables, adjust their conduct accordingly; and that, sometimes, can find new solutions, act without human intervention, do something which resembles thinking. We live in an environment where most of the artifacts display these features to different degrees: an elevator, a remote-controlled automatic train, an auto-piloted aircraft, a personal computer, a robot, and many others. In the following discussion, we concentrate only upon automation related to organizations that initial literature designated as industrial automation. We will adopt the term 'socio-technical automation' which includes applications that go beyond factories.

Concept, research, and debate on automation evolved since 1952 according to the available technologies and areas of applications. Factory automation came first when mechanical engineering, together with production engineering, mechanized both machine tools and conveyors in the automobile industry, giving rise to that primitive but seminal type of automation called Detroit Automation, i.e., transfer lines where products, such as engines, are manufactured and moved without human effort. Electromechanical engineering developed ingenious devices for automatic control of continuous process industries—oil, chemical products, cement—and of discontinuous flow processes—steel mills, casting. They were applied also to traffic control of material flows—in railway control rooms—and of energy flows—in power plants.

Automation introduced major disruptions in assembly, packaging, bottling in various manufacturing industries, actually turning them from discontinuous processes into continuous ones. The development of electronic and computer sciences in the 1960s increased the depth and breadth of this evolution. Computer control in the 1970s also extended automatic control into individual operating machines of numerically controlled machines (CNC), advanced robots, integrated CNC, and robots in flexible manufacturing systems (FMS). CNC and FMS were a breakthrough: manufacturing operations were automated and at the same time it was possible to manufacture small batches at costs of economy of scale, since the optimization of planning processes was embedded in the system. In recent cases of automation, processes of research and development (R&D), manufacturing, logistics, and distribution were integrated into internal and external functions or external units. To achieve this, a new architecture of telecommunication and information systems and new network organization structures were designed.

Task displacement and production integration pursued two distinct models. The first was the unmanned factory aimed at cutting costs and controversies with blue-collar workers through advanced mechanization and use of robots, as in the case of Toyota and Fiat

Cassino of the 1970s. The second model was the integrated factory which deliberately kept a meaningful portion of human tasks away from machines, as in the late Toyota developments and in Fiat Melfi. Cost savings and productivity gains were reported in Japanese firms making considerable steps back from automation, as in Toyota plant by 66 percent.

In the integrated factory, an optimal distribution of skilled work and automatic control takes place where advantages are not related to the reduction of direct manpower, but to the integration of various functions and units to achieve continuity and flexibility in business processes.

Another side of automation was office automation. Because of the development of large mainframe computers, electronic data processing (EDP) automated large areas of the operations of data collection, retrieval, and processing in public and private administrations: the work and organization in accounting, planning and control, and commercial departments were deeply affected. Electronic data exchange (EDI) supported communication among different departments and firms. First video terminals, then stand-alone personal computers (PCs), and finally networked PCs in the 1990s deeply affected the typical clerical and professional operations such as calculating, typing, editing, copying, printing, retrieving, mailing, etc. Automation of planning and control procedures gave rise to decision support systems (DSS) for managers, ranging from pioneering applications of management information systems (MIS) to the recent applications of electronic resource planning (ERP) such as the well-known SAP.

Computers in the 1980s could perform parts of very skilled tasks, such as f.i., scientific calculus, drafting, planning for designers, and manufacturing planners through computer-aided design (CAD/CAM), or project management for project leaders (Project Management SW). Automation for professionals was intended to substitute and accelerate routine tasks of scientists and other professionals. Computer graphics and multimedia first, followed by the Internet, did not automate tasks but offered opportunities to inform and communicate which did not exist before (Zuboff 1988).

Automation in services is a recent appearance. The activity of a cashier at a bank, or a booth in a public office is now widely performed by automatic teller machines (ATM) with which customers interact. Behind an ATM, there are complex changes in the banking processes, such as the advent of electronic money and electronic banking. Customer relations management (CRM) is frequently run by call centers that are going to employ 2 percent of the working population. CRM systems allow the delivery of a quick and reliable answer to customers through an integrated system of distributed intelligence including human beings, automatic telephone answering systems, web resources, and agent software.

It is uncertain whether the advent of the Internet may efface the concept of automation or will revamp it. A huge amount of functions are run automatically by Internet technology, one of the most powerful ever to have appeared. In addition to the instant and worldwide communication of texts, images, videos, and sound, the Internet will include functions and structures which in the past were work and organizational systems. The search engines automate f.i., the process of research of data displacing bibliographic searches and allowing access to data mines that were inconceivable without that technology. 'Click' is the action which instantly allows an individual to activate complex processes and actions such as retrieving, downloading, mailing, chatting, signing, buying, and many others. Disruptive applications, such as e-business, may encompass some previous stages of automation in organizations. Business-to-consumer f.i. creates a direct connection between the firm and the final customer, making the purchasing process a matter of a click. Business-to-business applications connect managers, workers, suppliers, co-maker companies, and clients using extranets to link the internal EDP with external EDPs, encompassing in the system ICT applications of the 1980s and the 1990s as EDI, ERP, CSCW, and making various organizations a unique network organization, a real organization through 'virtual organizing,' as described by Venkatraman and Handerson.

3. *Social Effects of Automation*

The literature on the social effects of automation has had discontinuous progress. In the late 1950s, development of automation in various industries gave rise to a wide program of research on the social effects of automation, financed by the governments of USA and other industrialized countries. Under union pressure, this research gave rise to concerns about an imminent wide diffusion of automation. This resulted in fears of potential consequences in terms of unemployment, deskilling, working conditions, work organization, and union/management relations. Outstanding research was carried out in this field but, by the mid 1960s, this research suddenly stopped when the US National Commission for Technology, Automation, and Economic Progress came to the conclusion that employment would not be affected, due to difficult implementation and limited diffusion of automation in most industries. In the 1970s, automation continued to increase incrementally but did not receive much attention from social scientists, except where it appeared relevant for ongoing studies and criticisms of Taylor–Fordism. In the 1980s, automation grew at a faster pace, bringing new concerns about employment and working conditions. In the 1990s, discussion and research focused more on information and communication technologies and their impact on the

service sector. By the beginning of 2000, use of the Internet was attracting great interest with a large number of case studies and prophecies, but little solid research.

Discussion about the nature and scope of automation may be grouped around two basic divides. The first focuses on the issue of continuity/discontinuity of technical development. Bright (1958) offers a classification of types of mechanization based upon degree of task allocation among men and machines, including automation. For economists, automation is no more than a particular case of technical development and no different social effects should be expected:

> the term automation is used in a loose way and there is not any advantage in distinguishing it from advanced forms of technical change.

Wild sees automation encompassing the task displacement of traditional mechanization, while he sees it as a new phenomenon as far as automatic control is concerned. For the British Social Science Council (1980) the novelty of automation lays in control systems. Drucker sees automation as 'the use of machines to run machines' ( ).

Crossman (1960), one of the first major scholars on the subject, stated that automation is the second stage in the displacement of human work substitution by information processing to machines, since in automation

> to steer, control, operate a machine, calculate, solve problems, take decisions, discussion may be performed by machines as by the human brain.

Artificial intelligence made computers more able to cope with complex problems. Development of advanced robotics gave machines sensor-motor skills (Sheridan 1988).

However, a breakdown was envisaged by Diebold (1952) who first saw automation as a new concept of integrated systems allowing continuous flow in the production process. When computers where not yet in operation, Pollock defined automation as 'the integration of continuous and discontinuous processes under the coordination of electronic systems'. For Buckingam, automation is a new pattern of manufacturing.

For sociologists such as Touraine (1955), Woodward (1965), Blauner (1964), Meissner (1969), Braverman (1974), automation is a breakdown, because massive displacement of skilled and unskilled work to machines overcomes both the craft-based organization model and the mass-production model. New tasks, roles and organizations are created, while the amount of work force required per product unit sharply diminishes: men undertake mainly control tasks. This occurred particularly in the 1980s and 1990s, with supervisory control of systems fully controlled by a computer and men controlling the computer, with tasks ranging from, at a minimum, switching the computers on and off, to at a maximum, controlling how the computer controls the computer.

Automation, in this discussion, becomes a breakdown: it not only substitutes any kind of human work, but also generates new functions not affordable by the human being. Automation is the integration of complex systems by means of automatic devices. Human work on the other hand does not disappear. Knowledge workers increase and concentrate on the design, development and supervision of 'intelligent machines.' Automation in this view is the creation of distributed knowledge among men and machines, and human-machine cooperation (HMC) according to Bagnara (Hoc 2000).

The second great divide concentrates on the pessimistic and optimistic views of automation. The pessimistic view holds that a new type of Taylorism is under way, the technological Taylorism. The 'intelligence of the system' as a whole would shift upwards towards functions which will become more specialized. This will result in a new wave of de-skilling amongst operatives and employees who will become 'transitional ancillary workers' and 'data entry workers.' As a consequence, there will be a polarization of the labor force, with a growing divergence between the unskilled masses and a super-skilled elite (Braverman 1974, Rifkin 1995).

The opposite view—the optimistic view—sees technology as a way of freeing workers from monotonous or dangerous tasks and from physical fatigue. Machines will do the work and men will supervise them (Blauner 1964). Although new technologies require less manpower, they should boost the cycle of expansion, giving rise to new firms and new professions (Freeman 1997).

Research has supported both the above-mentioned positions. It was clear since the beginning that different levels of allocation of tasks to machines affect differently the amount and quality of operative and managerial work (Bright 1958), with potential unemployment. Authors such as Jaffe and Froomkin assumed that the continuity of technical innovation would not generate dramatic changes in the content and amount of work. Rifkin (1995), Forester, and De Masi argued that technological development results in large unemployment. Economists note that job losses are due not only to technological development but in larger proportion to international trade, business process reengineering, and reorganization. Studies have reported that technical innovation creates new jobs and new enterprises with a potential gain in jobs. Sylos Labini (1997) maintains that a steady growth in national GNP of more than 3.5 percent will reduce unemployment even during a period of intense technological innovation.

Early research on skills focused first on the change in the nature of tasks in automation. Braverman

(1974) suggested that a fatal de-skilling should take place, the first Kern and Shuman (1970) found data supporting the dequalification thesis. Empirical research found that manipulation, traditional craft skills, dexterity, and operational ability (Bright 1958) were fading. On the contrary, tasks of supervision (Friedman 1971), 'control skills' (Crossman 1960), responsibility, and relationship with the entire production system (Touraine 1955), mental or visual jobs (Blauner 1964), sophisticated skills to respond to stochastic events (Davis and Taylor 1976) were all increasing.

Undesirable effects and risks of automation in dangerous processes have been studied as normal accidents (Perrow 1984), human and organization errors (Reason 1990), and misinterpretation and sense-making of technological events (Weick 1990).

Most early research into automated plants found that job design criteria moved away from the Taylor–Fordist tradition and that new criteria were emerging: interdependence of role and fewer levels of supervision, group behavior, multiskilling and shared responsibility, decline of strict job specifications and close supervision, communication, cooperation and influence, and group autonomy to cope with uncertainties. Recent developments of advanced computers enhance face-to-face and remote group behavior, human–computer cooperation (Hoc 2000), and, towards the beginning of the twenty-first century, the human–agent interaction (Lewis 2000).

The impact of automation on the composition of the workforce has been studied. Mallet ( ) found the potential development of technicians as a new working class. Kern and Shuman (1984) corrected the dequalification thesis but reported about potential polarization of the workforce. This thesis returns with Rifkin (1995) who charged modern technology with the responsibility of bringing work to an end, with the emergence of a small fraction of privileged individuals. Burris ( ) indicates that computerized work generates overall up-skilling, some de-skilling and skill bifurcation; Barley and Butera, Donati, Cesaria found that knowledge workers, managers, professionals, and technicians were rapidly increasing in industrial countries from the present 30 percent to a likely 50 percent.

Early studies on the Quality of Working Life had shown that with automation physical fatigue was decreasing, but there were increasing risks for workers: relaxation, nervous strain (WHO), problems of mental health, stress due to isolation and responsibility for catastrophic events, alienation, and powerlessness. The huge work edited by Martin Helander (1988) gives an exhaustive review of the different instances and differentiated impacts.

Workforce education influences wages and grades more than jobs and technology, coming back to the early findings of the Aix-en-Provence school of Maurice, Sellier, Silvester (*effect societal*). Plants adopting automation have more skilled workers, but this was found pre- and post-adoption. Wage premiums associated with technological change are actually a result of educational level. The wages of computer users are 15–20 percent more than for nonusers, but these were better compensated before the introduction of NTs.

4. *Conclusions: Effects or Design?*

Technology in itself has a direct influence on the physical universe of work (light, noise, heat, vibrations, radiation) and the operational universe of work (psychosocial conditions). Design of equipment has a major effect on:

(a) the amount of work per product unit (work displaced by technology);

(b) the operational 'tasks' (man/machine task allocation); and

(c) the coordination and control functions left to the managerial system.

Technology has only a moderate impact, if any, on components of work systems such as roles, organizational structures, communications, work allocation systems and policies, work management systems (such as skills, pay, working hours, regulations), and in particular the cultural and value systems which govern cooperation and conflict (Butera 1990).

Automation implies more than sheer displacement of human tasks because it includes the creation of new tasks of design, maintenance, and coordination and generates new forms of organization of production and restructuring of the labor market. Automation is at the same time a process of social and technical change, a measurable degree of distribution of knowledge and operational properties among men and technical equipment, and a set of advanced forms of socio-technical systems in manufacturing and services. Automation is a socio-technical system as a combination of business processes, technology, organization, and work, according to Trist and Murray (1993). Different patterns of supervisory control (Sheridan and Hennessey 1984), human–computer interaction (Norman 1986), and human–computer collaboration (Hoc 2000) may make a difference at the same level of automation. Participation of stakeholders in the design of the system and of new business, along with the process of introduction of the new technology, may change the final pattern of socio-technical systems (Davis and Taylor 1976).

If technology upsets work, only design resets work. Social and technical methodologies take the opportunity of 'room for maneuver' for planning, design, and experimentation of systems (such as territories, firms, factories, offices, and professional units), which may jointly meet economic, social and technical goals (Butera 1990). In synthesis, social dimensions in technological development are not effects but instead a matter of design. Designing integrated architectures of socio-technical systems, patterns of organization and

work, and interaction among men and computers is a social event and should be a participative one.

See also: Computers and Society; Demand for Labor; Employment and Labor, Regulation of; Expert Systems in Medicine; Human–Computer Interaction; Mode of Production; Psychological Climate in the Work Setting; Stress at Work; Technology and Organization; Work and Labor: History of the Concept; Work, Sociology of

Bibliography

Blauner R 1964 *Alienation and Freedom: The Factory Worker and His Industry*. The University of Chicago Press, Chicago

Braverman H 1974 *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. Monthly Review Press, New York

Bright J R 1958 *Automation and Management*. Division of Research, Graduate School of Business Administration, Harvard, Boston

Butera F, Thurman J E (eds.) 1984 Automation and work design. North Holland, Amsterdam

Butera F 1990 Options for the future of work. In: Butera F, De Martino V, Koheler E (eds.) *Options for the Future*. Coogan, London

Crossman E R F W 1960 *Automation and Skill*. HMSO, London

Davis L E, Taylor J C 1976 Technology, organization and job structure. In: Dubin R (ed.) *Handbook of Work, Organization and Society*. Rand McNally, Chicago

Diebold J 1952 *Automation: The Advent of Automated Factor*. Van Nostrand, New York

Freeman C 1998 Unemployment and long waves of economic development. In: *Sviluppo Tecnologico e disoccupazione*. Convegno Accademia dei Lincei

Helander M (ed.) 1988 *Handbook of Human–Computer Interaction*. North Holland, Amsterdam

Hoc J M J 2000 *Human–machine Interaction to Human–machine Cooperation*. Taylor and Francis, London

Kern H, Schumann M 1984 *Das Ende der Arbeitsteilung. Rationalisierung in der industriellen Produktion*. C H Beck Verlag, Munich, Germany

Meissner M 1969 *Technology and the Worker. Technical Demands and Social Processes in Industry*. Chandler, San Francisco

Norman D A, Draper S W (eds.) 1986 User centered system design: New perspectives on human computer interaction. Erlbaum, Hillsdale, NJ

Perrow C 1984 *Normal Accidents: Living with High-risk Technologies*. Basic Books, New York

Reason J 1990 *Human Error*. Cambridge University Press, Cambridge, UK

Rifkin J 1995 *The End of Work: The Decline of the Global Labor Force and the Dawn of the Post-market Era*. Putnam, New York

Sheridan T B, Hennessy R T (eds.) 1984 *Research and Modeling of Supervisory Control Behavior*. National Academic Press, Washington, DC

Sylos Labini P 1998 Nuove Tecnologie e occupazione. In: *Sviluppo Tecnologico e disoccupazione*. Convegno Accademia dei Lincei

Touraine A (ed.) 1955 *L'évolution du travail ouvrier aux usines Renault*. CNRS, Paris

Trist E, Murray H 1993 *The Social Engagement of Social Science*. University of Pennsylvania Press, Vol. 2

Weick K E 1990 Technology as equivoque: Sensemaking in new technologies. In: Goodman P S, Sproull L S (eds.) *Technology and Organizations*. Jossey Bass, San Francisco

Woodward J 1965 *Industrial Organization: Theory and Practice*. Oxford University Press, London

Zuboff S 1988 *In the Age of the Smart Machine: The Future of Work and Power*. Basic Books, New York

F. Butera

Autonomic Classical and Operant Conditioning

Autonomic classical conditioning is a type of conditioning in which the conditioned response (CR) is expressed through one or more organs that are targets of the autonomic nervous system. Among these are the cardiovascular system, digestive organs, and endocrine glands. The CR that develops during conditioning involving the autonomic nervous system has been characterized as either a discrete autonomic CR (Pavlov 1910, Dworkin 1993) or a nonspecific autonomic CR (Weinberger et al. 1984). A discrete autonomic CR is one that has been elaborated from an unconditioned reflexive response to a highly specific unconditioned stimulus (US). The autonomic CR is considered to be nonspecific when it is one of a cluster of concurrent responses elicited by a CS associated with a noxious US (Weinberger et al. 1984). This nonspecific autonomic CR is thought to acquire functional utility within the context of a cluster of CRs that develops during conditioning. The term conditioned fear is used to refer to the change in the CNS that accounts for the cluster of autonomic-mediated responses that, as a result of conditioning, is elicited by the CS.

1. Functional Utility of Discrete CRs

A fundamental observation made by Pavlov (1910) was that an autonomic CR, as it grows in strength, becomes increasingly similar to the UR. The development of a conditioned autonomic reflex was seen as a way to augment the homeostasis-promoting effect of the UR. For example, placing an irritating acidic substance in the mouth will trigger an (unconditioned) reflexive salivary response. The saliva lubricates the mouth and diminishes the impact of the acidic substance upon the mucosa, thereby promoting homeo-

stasis. The CR to a CS that precedes the US is also salivation and the latency of this CR decreases, and its magnitude increases, as conditioning progresses. Because this secretion (the CR) precedes the US it improves the effectiveness of the unconditioned reflex.

Pavlov saw autonomic conditioning as a way to study central nervous system (CNS) processes, and most of his research efforts were dedicated to studying these processes rather than the adaptive functions of discrete autonomic CRs. One of his students, Bykov, was among the first investigators to study the functional utility of discrete autonomic CRs. The focus of his research was understanding the role of classical conditioning in the control of visceral functioning, specifically regulatory mechanisms involved in conditioned diuresis (Bykov 1957). The UR in these studies was an increase in urine secretion in response to an intrarectal infusion of water. His observations regarding the temporal relationship between the CR and US that developed during conditioning were similar to those made by Pavlov in that he found that the latency of the urine flow decreased after repeated exposure to the US. Eventually, urine flow was initiated prior to the presentation of the US: a discrete autonomic CR was developing and the response was anticipatory in that it preceded the onset of the US. The CS in this experiment was the laboratory setting. Bykov (1957) also conducted studies which demonstrated that conditioned autonomic responses could be developed in the cardiovascular system. In these studies the dependent measure was the increased cardiac output that occurs when an animal is required to run on a treadmill. A discrete tone was used as the CS for this experiment. Similar to the results of the conditioned diuresis studies, Bykov found that the latency of the cardiac response (increased cardiac output) became shorter with repeated testing on the treadmill. Eventually, increases in cardiac output occurred well in advance of running. This discrete autonomic CR was smaller than the UR but it was adaptive in that it anticipated the increases in metabolic demands that are associated with running.

The observations made by Pavlov and by Bykov regarding the temporal relation between the CR and the US suggest that the conditioning of a discrete autonomic CR is a form of predictive homeostasis. Predictive homeostasis is distinguished from reactive homeostasis by the time at which the autonomic regulatory response mechanism is engaged. Regulatory response mechanisms that mediate reactive homeostasis are engaged after the occurrence of a regulatory challenge. For example, the pancreas secretes insulin in response to a rise in blood glucose level. Similarly, cardiac output and blood flow to the skeletal muscles increase in response to the elevated metabolic requirements of exercise. The regulatory response usually remains engaged until the controlled variable is returned to a preset reference level. In contrast to a reactive homeostatic mechanism, a predictive homeostatic mechanism is engaged prior to the onset of the regulatory challenge, and thus results from conditioning. It is a preemptive response that mitigates (or totally nullifies) the impact of the forthcoming stimulus that poses a challenge to homeostasis. Thus, based on the observations of Pavlov and Bykov, one possible adaptive function of a discrete autonomic CR is that it promotes homeostasis by mitigating, in advance, the impact of the regulatory challenge (the US).

It is axiomatic that if conditioned autonomic responses are involved in predictive homeostasis, the CR must be very similar to the UR. The results of studies assessing CRs to drug injections appear to be paradoxical with respect to this view. In general, the CR to a drug is opposite to the pharmacological effect of the drug (Siegel 1983). For example, epinephrine causes a decrease in gastric secretion, whereas the autonomic CR is an increase in gastric secretion (Guha et al. 1974). The CRs to almost two dozen drugs have been studied and in each case the direction of the CR is opposite to the pharmacological effect of the drug (Siegel 1983). These findings provided the foundation for a model of drug tolerance that is based on the idea that learned responses to environmental cues (the CS) associated with drug ingestion are conditioned compensatory responses (Siegel 1983).

If the discrete autonomic CR is to be viewed as a component of a predictive homeostatic process, it is essential to delineate the regulatory characteristics of the unconditioned reflex; otherwise it is not possible to determine if the CR is a replica of the UR. A question that naturally arises from studies involving conditioned autonomic responses to drugs is: Among the various physiological changes that occur as a result of drug ingestion, what is the effective US and the effective UR? For example, eating candy eventually will lead to a conditioned hypoglycemic response that precedes the arrival of the candy in the stomach. There are a number of physiological reactions to the ingestion of a piece of candy including salivary secretions, secretions in the stomach, an initial rise in blood glucose level, the secretion of insulin by the pancreas, and a subsequent lowering of blood glucose level. Which of these reactions is the US and which one is the UR? The tradition has been to refer to the drug as the US and the effect of the chemical (drug) as the UR. Thus, in the candy example the increase in glucose would be identified as the UR. Viewed from this perspective, the hypoglycemic CR is opposite in polarity to the UR, and hence appears to be a compensatory response. Viewed in terms of regulatory mechanisms and homeostasis, the effective UR is the insulin response (which is reflected by the hypoglycemic response) and the effective US is the increase in glucose. The rise in glucose is considered to be a 'challenge' to the control system involved in the regulation of blood glucose level and this system responds to this challenge by secreting insulin. In

general, most autonomic URs are homeostatic in that they diminish the impact of the US. Thus, the UR serves to limit the disruptive effect of the US upon the physiological state of the organism. Viewed in this way, the effective US is always the physiological change that engages a homeostatic response—the effective UR—that serves to return the controlled variable (blood glucose in this example) to its reference level. The CR is also an increase in insulin, and like the UR, is reflected by the hypoglycemic response. The CR is a response to the sweet taste of candy, the CS; it precedes the absorption of glucose (Deutsch 1974) and thus promotes homeostasis by anticipating a rise in blood glucose.

In our view, the reason why the autonomic CR appears to be a compensatory response to a drug is rooted in the manner in which the results of these experiments have been conceptualized (see Dworkin 1993 for a thorough analysis). More specifically, the pharmacological effect of the ingested drug has been considered to be the UR. However, if autonomic conditioning is viewed in terms of homeostatic mechanisms, the UR is always a corrective homeostasis-promoting response that serves to return a controlled variable to its reference level. The (effective) US is always a change in the physiological state of the organism that poses a regulatory challenge. If unconditioned reflexes are viewed as participants in reactive homeostasis and conditioned reflexes participants in predictive homeostasis, the CR and (effective) UR will be similar.

2. *The Adaptive Utility of Conditioned Fear and Nonspecific Autonomic Conditioned Responses*

In general, autonomic CRs develop more rapidly than most CRs mediated by skeletal muscles (referred to as somatic CRs) and in many cases this rapid development has adaptive utility. This observation, along with others, has led to the theoretical position that conditioning involving an aversive US may occur in two stages. The initial phase consists of the development of a changed state of the CNS, referred to as conditioned fear. Autonomic CRs and behavioral changes, such as CS-elicited behavioral freezing, are a reflection of this changed state of the CNS. Conditioned fear is followed by the acquisition of discrete, skeletal motor responses that mitigate or totally nullify the impact of the US (Miller 1948, Mowrer 1947). Weinberger (1982) used the term nonspecific CR to refer to a response learned in the initial phase and the term specific CR to refer to the responses mediated by skeletal muscles during the later phase. By nonspecific he meant that it was not linked to any particular US and that its functional utility is only gained within the context of the cluster of CRs elicited. Thus, an aversive US, regardless of it specific characteristics (e.g., airpuff to the cornea vs. loud noise), elicits a similar pattern of URs. The constellation of nonspecific autonomic CRs that develop during conditioning are usually similar to their corresponding URs. Conditioned fear is the term used to refer to the CNS state that generates the cluster of nonspecific CRs.

3. *Conditioned Fear and the Development of Somatic CRs*

According to one view (Weinberger et al. 1984), conditioned fear responses are thought to facilitate the development of somatic CRs that attenuate or totally nullify the impact of the US. The nonspecific autonomic CRs associated with conditioned fear are thought to reflect a general change in behavioral state. It not only leads to changes in activity in the autonomic nervous system, it alters the organism's orientation towards the environment in a way that enhances the reliability and efficiency of feature extraction from sensory input (Sokolov 1963), thereby facilitating the detection of stimuli that are most relevant to the development of adaptive somatic CRs. For example, conditioned decreases in heart rate (bradycardia) to a tonal CS are observed well in advance of the development of eye blink CRs to the same CS. The bradycardia CR is thought to be one of many nonspecific autonomic CRs that emerge as a result of the CS-US pairings. The altered CNS state (i.e., conditioned fear) that underlies the constellation of nonspecific autonomic CRs is thought to facilitate the development of the eye blink response. Once this adaptive somatic CR develops, the previously conditioned bradycardia disappears (Powell et al. 1990). Within this context, conditioned fear may be considered a CR that can be involved in predictive homeostasis because it motivates behavioral changes that lead to avoidance, in advance, of the presentation of an aversive stimulus that would disrupt homeostasis. The CNS changes associated with conditioned fear serve to improve the organism's ability to detect sensory stimuli that may be linked to danger and threat or may guide coping responses such as instrumentally conditioned responses that allow the organism to avoid an aversive stimulus.

Alternatively, the autonomic CRs in aversive conditioning studies may be interpreted solely in terms of autonomic responses involved in predictive homeostasis (Schneiderman 1972). The UR to an electric shock US in the restrained rabbit, for instance, involves an increase in arterial blood pressure, whereas the CR consists of bradycardia. Thus, the CR appears to mitigate the blood pressure increase elicited by the US, which becomes attenuated after several conditioning trials. In contrast, the UR to an aversive US in the unrestrained rat consists of an increase in arterial pressure as does the CR. Thus, the blood pressure CR

appears to be facilitating the metabolic needs of the organism (i.e., predictive homeostasis) by what Obrist and Webb (1967) referred to as the cardiac–somatic linkage. In conclusion, it appears that much of autonomic classical conditioning can be interpreted within the context of predictive homeostasis although its relationship to such concepts as conditioned fear remain to be explored.

4. *Operant Conditioning of Autonomic Responses*

In addition to classical conditioning, operant (instrumental) control has been demonstrated over a wide variety of autonomic responses in humans, including heart rate (e.g., Engel and Hansen 1966), blood pressure (e.g., Brener and Kleinman 1970), and vasomotor activity (Snyder and Noble 1968). Similar observations have been made in nonhuman animals (e.g., Miller and Banuazizi 1965).

Although there is ample evidence for instrumental conditioning of autonomic responses, the level of specificity of operant control over these responses remains an issue. When human subjects are asked to increase their respiratory activity, for example, there are changes in heart rate, but equivalent changes in heart rate occur when subjects are asked to alter their blood pressure. Equivalent heart rate changes are produced by instructions to alter heart rate, blood pressure or respiration. Moreover, the heart rate changes observed in this type of experiment have been found to be inversely proportional to the degree of somatomotor restraint imposed upon the subjects. In view of the cardiac–somatic linkage (Obrist and Webb 1967), it seems reasonable to hypothesize that the CNS programs for somatomotor and cardiorespiratory responses are coupled in the CNS, and these functionally related efferent systems are influenced in parallel by the same processes. This type of organization would lead to integrated cardiorespiratory/somatomotor responses whose characteristics are dictated by the metabolic requirements of the task at hand (see Brener (1974a) for a more thorough discussion of the issue). The results of studies by DiCara and Miller (1969) support this view. Using a shock-avoidance procedure, these investigators conditioned increases in heart rate in one group of curarized rats and decreases in heart rate in a second group. Subsequently, when tested in the noncurarized state, the rats previously reinforced for heart rate increases showed substantially higher levels of behavioral activity and respiration rate than animals reinforced for decreasing their heart rates.

Afferent feedback signals from skeletal muscles are extremely important to the regulation of motor activity. One view is that the activation of the mnestic trace of the central representation of these feedback signals is necessary for the development of instrumentally conditioned somatomotor responses (James 1890). Instrumental conditioning of cardiovascular responses is thought to involve a similar mechanism (Brener 1974b). In view of the somatic-cardiac linkage (Obrist and Webb 1967), a question that naturally arises is whether operant control of cardiovascular responses can develop in the absence of afferent feedback from skeletal muscles. Initial studies by Miller and co-workers (DiCara and Miller 1968; Miller and Banuazizi 1965) revealed that operant conditioning of autonomic responses occurs in curarized animals, thereby providing evidence that afferent feedback from skeletal muscles is not necessary for the development of instrumentally conditioned autonomic responses. Subsequent studies from the same laboratory (Miller and Dworkin 1974), however, failed to replicate these findings.

Summary

When viewed in terms of homeostasis, both discrete and nonspecific autonomic CRs are seen as homeostasis-promoting responses that prepare an organism for an impending unconditioned aversive stimulus. From this perspective, the unconditioned autonomic reflex is considered to be a reactive homeostatic mechanism, and discrete URs as compensatory regulatory responses. The conditioned autonomic reflex is considered to be a predictive homeostatic mechanism because the discrete CR that develops mitigates (or totally nullifies), in advance, the impact of the regulatory challenge (the US), thereby improving the regulatory properties of the unconditioned reflex. Autonomic conditioning involving a nonspecific CR is also considered to be a predictive homeostatic process. The constellation of nonspecific CRs that develop during autonomic conditioning is thought to be produced by an altered CNS state, referred to as conditioned fear, which results from the conditioning process. The change in behavioral state associated with conditioned fear serves two purposes: it facilitates the development of somatic CRs that mitigate the impact of the US, or the development of a instrumentally learned response that allows the organism to actively avoid the US.

Operant control over a variety of autonomic responses has been demonstrated, though these responses are usually linked to somatic responses. Thus, it appears that CNS programs for somatomotor and cardiorespiratory responses are coupled in the CNS, and that these functionally related efferent systems are influenced in parallel by the same processes.

See also: Cardiovascular Conditioning: Neural Substrates; Classical Conditioning, Neural Basis of; Fear Conditioning

Bibliography

Brener J 1974a Factors influencing voluntary cardiovascular control. In: DiCara L V (ed.) *Limbic and Autonomic Nervous Systems Research*. Plenum, New York

Brener J 1974b Learned control of cardiovascular processes: Feedback mechanisms and therapeutic applications. In: Calhoun K S, Adams H E, Mitchell K M (eds.) *Innovative Treatment Methods in Psychopathology*. Wiley, New York, pp. 245–72

Brener J, Kleinman R A 1970 Learned control of decreases in systolic blood pressure. *Nature* **226**: 1063–4

Bykov K M 1957 *The Cerebral Cortex and the Internal Organs* [*Kora Golovngo Mozga I Vnutrennie Organy*] (Trans. and ed. Gantt W H). Chemical Publishing Co., New York

Deutsch R 1974 Conditioned hypoglycemia—A mechanism for saccharin-induced sensitivity to insulin in rat. *Journal of Comparative and Physiological Psychology* **86**: 350–8

DiCara L V, Miller N E 1968 Instructional learning of vasomotor responses by rats: Learning to respond differentially in the two ears. *Science* **159**: 1485

DiCara L V, Miller N E 1969 Transfer of instrumentally learned heart rate changes from curarized to non-curarized state: Implications for a mediational hypothesis. *Journal of Comparative Physiology and Psychology* **68**: 159–62

Dworkin B R 1993 *Learning and Physiological Regulation*. University of Chicago Press, Chicago

Engel B T, Hansen S P 1966 Operant conditioning of heart rate slowing. *Psychophysiology* **3**: 176–87

Guha D, Dutta S N, Pradhan S N 1974 Conditioning of gastric secretion by epinephrine in rats. *Proceedings of the Society for Experimental Biology and Medicine* **147**: 817–19

James W 1890 *The Principles of Psychology*. Holt, New York

Miller N E 1948 Studies of fear as an acquirable drive: I. Fear as motivation and fear-reduction as reinforcement in the learning of new responses. *Journal of Experimental Psychology* **38**: 89–101

Miller N E, Banuazizi A 1965 Instrumental learning by curarized rats of a specific visceral response, intestinal or cardiac. *Journal of Comparative Physiology and Psychology* **65**: 1–7

Miller N E, Dworkin B R 1974 Visceral learning: Recent difficulties with curarized rats and significant problems for human research. In: Obrist P A, Black A H, Brener J, DiCara L V (eds.) *Cardiovascular Psychophysiology*. Aldine-Altherton, Chicago

Mowrer O H 1947 On the dual nature of learning: A reinterpretation of 'conditioning' and 'problem-solving.' *Harvard Educational Review* **17**: 102–50

Obrist P A, Webb R A 1967 Heart rate during conditioning in dogs: Relationship of somatic-motor activity. *Psychophysiology* **4**: 7–34

Pavlov I P 1910 *The Work of the Digestive Glands*, (*Pawlow*) *and Estimation of Pepsin Digestion by Modern Instruments of Precision*, 2nd English edn. (trans. Thompson W H). Griffin, London

Powell D A, Buchanan S L, Gibbs C M 1990 Role of the prefrontal-thalamic axis in classical conditioning. *Progressions in Brain Research* **85**: 433–65

Schneiderman N 1972 The relationship between learned and unlearned cardiovascular responses. In: Black A H, Prokasy W F (eds.) *Classical Conditioning. II. Current Research and Theory*. Appleton-Century-Crofts, New York

Siegel S 1983 Classical conditioning, drug tolerance, and drug dependence. In: Israel Y, Glaser F B, Kalant H, Popham R E, Schmidt W, Smart R G (eds.) *Research Advances in Alcohol and Drug Problems*. Plenum, New York, Vol. 7, pp. 207–43

Snyder C, Nobel M 1968 Operant conditioning of vasoconstriction. *Journal of Experimental Psychology* **77**: 263–8

Sokolov E N 1963 Perception and the conditioned reflex. Pergamon, New York, pp. 5–19, 49–53, 295–303

Weinberger N M 1982 Sensory plasticity and learning: The magnocellular medial geniculate nucleus of the auditory system. In: Woody C D (ed.) *Conditioning: Representation of Involved Neural Functions*. Plenum, New York, pp. 697–718

Weinberger N M, Diamond D M, McKenna T M 1984 Initial events in conditioning: Plasticity in the pupillomotor and auditory systems. In: Lynch G, McGaugh J L, Weinberger N M (eds.) *Neurobiology of Learning and Memory*, Guilford, New York, pp. 197–227

R. W. Winters and N. Schneiderman

Autonomous Agents

Autonomous agents are software programs which respond to states and events in their environment independent from direct instruction by the user or owner of the agent, but acting on behalf and in the interest of the owner. The term agent is not defined precisely, and agent software can range from simple programs composed of a small number of rules to large and complex systems. Agent technology was developed in artificial intelligence (AI) research, and can include complex AI techniques. Important application areas are those where the human user can benefit from continuous data analysis, monitoring of data streams and large databases, and where routine reactions to events are required. Many applications are related to user interface technology and the Internet. A weak position sees agents as tools which relieve humans from routine tasks suited to computational solutions, but there is also a strong position which asserts that agents can be constructed to mimic or even surpass cognitive functions of the human.

1. *What are Autonomous Agents?*

Agents are defined as software programs which:

(a) sense events and analyze data streams in a defined environment;

(b) react to certain events and states with defined responses;

(c) exhibit goal-oriented behavior representing the intentions of the owner or user of the agent;

(d) act proactively on their own initiative;

(e) function in a temporally continuous manner.

The following functions may also be implemented in agents.

(a) Learning, most often adaptation, which is the simplest form of learning, based on statistical data analysis and observation of process parameters (potentially more complex forms of learning can be included).

(b) Communication with other agents or humans, including the exchange of data, parameters, and goal functions.

The autonomy of agents is limited by the interests of the owner. The degree of autonomy ranges from task-specific agents with a restricted domain of application to viruses which proliferate widely and do not ever report back to their owners.

1.1 Distinction Between Robots and Agents

Autonomous robots and agents are closely related, but agents are software programs only, whereas robots comprise hardware designed for the specific functions of the robot. Robot behavior can be optimized by tuning sensors, effectors, or the control program; agents are composed of software functions for data analysis and communication only. Robot behavior must also be optimized with regard to energy and physical performance parameters, which are usually not important for agents.

1.2 Specific Properties of Agents with Respect to Software in General

Agents differ from software programs in general by their continuous operation, their reactive behavior, and their (partial) autonomy. It should be noted that feedback control systems also have some of these properties, but are generally restricted to the monitoring of continuous variables and the control of set points or simple forcing functions (e.g., the thermostat).

2. Constructing Agents

Agents are task-specific, and the design of a specific agent is optimized for its application to maximize performance and reliability, to minimize error probability, cost, and other relevant properties. For a well-understood task in a predictable context agent software can be specified as a discrete control system or a state machine. Autonomous agents are expected to be 'intelligent.' Intelligence refers to the extent to which the agent responds in an optimal manner to the events which he or she monitors. The quality criteria are correctness, speed, and reliability of responses. The goal of an agent is to optimize the benefit for the owner of the agent. It is thus important to note that it is the behavior only which is judged to be intelligent, not whether the underlying computational processes are implemented as a fixed control algorithm, or include planning, learning, and reasoning (McFarland and Bösser 1993).

2.1 Design Principles for Agents

The main principles for designing agents are as follows.

(a) Development of a reactive system based on profound analysis and understanding of the tasks and context in which the agent is operating. An agent of this type may be implemented as a limited number of rules, or a state machine. The intelligent behavior is not a function of the real-time computations of the agent, but of the optimal selection of a limited number of rules or behaviors which overall generate successful behavior.

(b) AI principles, including planning and learning, can be used to make an agent react successfully in many different states which need not be anticipated in the design phase.

(c) Use of inspirations and parallels to biological system.

2.2 Implementation of Agents

Agents may be implemented in widely varying forms, ranging from simple event–response formalisms or state machines (where the specification translates directly into implementation) to complex AI architectures with the aspiration to represent human behavior such as SOAR (Tambe et al. 1995). Brooks (1991) has argued that robots do not necessarily have to use cognitive functions, reasoning, or planning to behave efficiently and to appear intelligent, but that the selection of optimal features for sensors and effectors and simple rules governing the response and an appropriate architecture permits the implementation of robust and effective robots. The bottom-up composition of robots in a hierarchical fashion from separate components with their own independent behavior is called subsumption architecture, using for example legs comprising mechanical structures, sensors, and motors plus basic wiring which allows them to move independently, or a vision system which includes preprocessing and active control of the sensors as building blocks. The advantage of the subsumption architecture is that there is no need for a centralized representation of all sensor data, and complete planning of all behavior in detail. The central control system may be much smaller, and the behavior of the entire robot is more robust. As components are combined, the resulting behavior becomes more complex than that of the components, which is often called emergent behavior.

Agents, by analogy, would use existing data capture and analysis functions and initiate actions by calling

other programs, and thus can be restricted to the decision-making algorithms which are the essence of the agent's behavior.

2.3 Adaptation and Learning

The combination of a form of learning with autonomy is highly attractive because the ability of the agent to adapt to changing environmental conditions further decreases the need for supervision and instruction by the owner. In most instances simple forms of learning (adaptation of parameters) are used. More complex forms of learning (concept formation, elaboration of new response patterns based on planning) are hard to implement, require massive amounts of information, and are much less reliable and secure than pre-programmed responses. The ability to learn can have advantages, but also carries high cost during design and run-time.

2.4 Multiagent Systems and Cooperating Agents

Autonomous agents have particular strengths due to their task-specific and encapsulated nature, but this also limits them to the particular tasks and contexts for which they are designed. As the functionality of an individual agent increases, the optimization effort and maintenance effort is likely to increase unproportionally. There are more good reasons for developing distributed systems: robustness and reliability increase, access to local data and users demands fewer common resources, and the need for communication decreases. The composition of large systems as a collection of agents is attractive because of the potentially open architecture, where new agents can be added as required without the need to modify the existing system. The amount of communication and coordination in the system remains restricted due to the autonomous operation of agents. The technical field of multiagent systems is growing rapidly and receives particular interest for safety-critical systems engineering.

When there is not one owner for all agents, but many agents from different owners become active in the same context, these agents may communicate, negotiate, and share their resources (i.e., cooperate). Preconditions for multiagent systems and cooperative agent systems are defined communication languages and rules which govern a fair and beneficial exchange of information between agents.

2.5 Anthropomorphic and Mentalistic Concepts

The output from agents to human users is not restricted to text and figures, but can include natural language, speech, and animations using virtual reality techniques. The agent mimics the communication behavior of humans, making it easier for some users to understand messages (Babski and Thalmann 2000). This can help users, but may also be irritating, especially when the load on human communication resources conflicts with other needs (display space, sound). A different approach is to use mentalistic concepts in the design of agent features, such as motivation, emotions, consciousness, or moods, which are not only presented to the user but also are meaningful concepts to describe the internal state of the agent (Sloman 1997). At this stage it may be best to consider these concepts as metaphors for some aspects of communication, but it would be hard to equate these to human communication or even human mental processes.

3. Applications for Autonomous Agents

Interesting application domains for agents are those where the relative strengths of agent technology can be used to augment or partly replace direct interaction of the human user with an application. Where continuous monitoring of data streams, repeated quick responses to events, statistical analysis and pattern detection in large amounts of data, and fast access to databases are required, agents can be more efficient and reliable. The WWW and other networks are interesting domains for the use of agents because they provide access to large amounts of information, create the need to search and select data, and make it attractive to enable automatic responses. The advantages are to relieve the user from routine tasks, and also to improve the quality of his or her performance.

3.1 Intelligent User Interfaces

User interface agents (Lieberman 1997, Maes 1994) observe user behavior and have access to all information in the system. Agents can interpret user requests, suggest actions, and automate routine functions based on observations of previous user behavior and knowledge of defaults and standard options. A drawback is that users may no longer be aware of available functions, and have no opportunity to learn to use them.

3.2 Information Filtering and Access

Access to large amounts of information, for example catalogues, the CD and book market or scientific publications, but also TV programs, can be supported by agents which use instructions from the user in the form of keywords, reference to relevant similar items, and user profiles elaborated from previous user requests to actively search and display information selectively. Other agents (such as the web spiders employed by search machines) analyze and index proactively large amounts of information and make this index available for queries, resulting in much faster responses than is possible if a full search had to be carried out in response to individual queries.

3.3 E-commerce Agents

In an open e-commerce environment agents may implement some functions of a market, carrying out negotiations about prices and conditions, and for the optimization of logistics. Beyond technical and legal issues remaining to be solved, these applications also require appropriate formalization of the functions of a market. The design of agent-based systems also stimulates the development of a better understanding of market mechanisms (Sandholm 2000).

3.4 Help, Performance Support, Training, and Tutoring

Help functions in user interfaces can include agents which analyze previous user input and provide specific help for the user in his or her application context. More specific performance support uses a database of procedures specific to, for example, a particular company, a complex technical device, or an application domain. Performance support consists of guiding users to select and carry out prescribed procedures, relieving the user from the need to memorize and practice these procedures previously, and also to standardize procedures and to reuse knowledge in an organization (Chin 1991).

An important trend is to make tutoring and training accessible to users specific to their need (sometimes called 'just-in-time-learning'). Archiving relevant knowledge (knowledge management) is often not sufficient; the learner also needs guidance to direct him or her to learn the most appropriate information, and to evaluate his or her progress appropriately. Agents provide guidance in on-line learning situations.

Operators for safety-critical tasks such as power station control or piloting are trained extensively in simulators. Where the nature of the task includes cooperation with other individuals, these are implemented as agents, representing a wide array of human behavior (Tambe et al. 1995).

4. Evaluation of Agent Performance

Agent performance is judged by the owner and user according to the advantages obtained from the agent. Most agents incur only minimal costs for storage and computing power, but may have considerable costs for communication with other agents, their environment, and the user. The main advantage should be reduced communication effort for the user in instructing his or her systems to perform his or her tasks.

The bandwidth of user–agent communication should be much lower than the bandwidth required for direct user–application interaction. The quality of performance must also be considered. Quality criteria for the information obtained and for the decisions made by the user include the number of hits and errors compared to other means of accessing the same information, the erroneous decisions made, and opportunities missed. Precise and detailed analyses of the benefit of agent applications are still scarce, but should become an essential part of the design process for agent-based systems.

Studies where users were asked to assess their subjective impression of agents indicate that users are pleased to receive active support in situations where they feel helpless, but that they are easily irritated by verbose and redundant information when they have relatively mature knowledge of a domain. Agents should include strategies and thresholds for presenting information in adequate form. Games and entertainment, where agents are opponent players, gain in popularity with improved display and reality of presentation.

4.1 User Models and Manipulation

Agents which are instructed by users and observe user behavior subsequently attempt to predict user behavior. The ability to predict user reactions is a precondition for influencing, and potentially a means to manipulate, user behavior. Robots, since their conception in the twentieth century, are postulated to have a built-in rule not to harm humans. Agents, however, are mediators of human-to-human communication: an agent represents the interests of an owner, and in this role may pursue goals which are in conflict with the interest of other users. Because ethical behavior is hard to enforce universally by rules, users are likely to be cautious against giving too much information to agent-based systems which are outside their control.

This is a particularly relevant aspect for e-commerce where the integration of data-processing systems opens the possibility to test and model the behavior and state of a partner or even a competitor's system. The ability to predict the behavior of other players in the market offers considerable advantages: the own strategy can be adapted: for example, when it is possible to observe that the production facilities of a supplier are not fully used, it is likely that lower prices can be negotiated. The stability and safety of agent-based systems must also take the time dynamics of the processes into account.

5. Grand Visions and the Ontological Incongruence Between Agent Implementations and Biological Entities

Some authors (among these Moravec 1988) advocate a strong point of view of agent technology: they are seduced by the expected further growth of computa-

tional power to believe that agent technology will help to automate most of human work within the lifetime of some of us, and even surpass human cognitive abilities. These visions should be viewed with caution: the ontological incongruence of the theories of natural and biological processes on the one hand, and the engineering principles for constructing autonomous and intelligent systems may not permit this. When the core concepts in the two domains are examined, an essential difference between theories of cognition and the design specifications for artificial intelligent systems becomes apparent: formal and mathematical theories of natural processes are understood to be abstractions of a qualitatively different reality; they are generally not assumed to be able to represent the entire essence of the processes modeled, not even when extended dramatically in size. Specifications of computational artifacts must be assumed to be self-contained and complete—they could not work otherwise and would be useless for applications. Unfortunately the terms used to describe mental processes, such as emotions or consciousness, are often only names to describe entities which we recognize when we see them, but do not yet understand sufficiently in a reductionist sense. These theories may mislead some to believe that implementing these theories as computing machines will produce the full behavior of the entities in question. If not hubris, the ambition to build human-like intelligence still represents a goal which transcends today's realistic possibilities. The challenge is either to develop much more precise knowledge of the cognitive mechanisms which we fail to understand with precision today, or to develop methods to construct complex systems from imprecise and incomplete specifications.

6. *Further Developments*

The concept of agents is still evolving, but offers promise that truly intelligent tools to help humans to carry out tasks beyond their natural cognitive abilities can be implemented. This is likely to be the support for increasingly complex routine tasks. As Marvin Minsky (1986) stated, 'We have built cranes to lift 10-ton weights, and computers will enable humans to carry out tasks which are beyond their ability so far.' Given the specific advantages of agent technology, agents will automate routine cognitive skills where reliable monitoring of large amounts of data and statistical analyses of information are carried out as a basis for decisions and actions. The main advantage for human users is that with human–machine communication at reduced bandwidth, complex tasks can be carried out successfully. It will probably be difficult to deploy systems for general use which are so complex that humans are unable to understand and if necessary correct their behavior.

The visions of the future abilities of agents are associated with worries that human work will be taken over by machines. Rather than a threat, this should be seen as an opportunity to free the cognitive capacity of the human, which can be directed towards more creative work and more sophisticated decision making. This development should increase the efficiency of the human user, and create value by making the use of working time, energy, and scarce resources more efficient.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence in Cognitive Science; Artificial Social Agents; Human—Computer Interaction; Human—Computer Interface; Intelligence: History of the Concept; Network Models of Tasks

Bibliography

Babski C, Thalmann D 2000 3D on the WEB and virtual humans. *Software Focus* **1**: 6–14

Brooks R A 1991 Intelligence without representation. *AI* **47**: 139–59

Chin D N 1991 Intelligent interfaces as agents. In: Sullivan J W, Tyler S W (eds.) *Intelligent User Interfaces*. ACM Press, New York

Lieberman H 1997 Autonomous interface agents. In: Pemberton S (ed.) *Human Factors in Computing Systems*. ACM Press, New York

Maes P 1994 Agents that reduce work and information overload. *CACM* **7**: 31–40

McFarland D, Bösser T 1993 *Intelligent Behavior in Animals and Robots*. MIT Press, Cambridge, MA

Moravec H 1988 *Mind Children. The Future of Robot and Human Intelligence*. Harvard University Press, Cambridge, MA

Minsky M 1986 *The Society of Mind*. Simon and Schuster, New York

Sandholm T 2000 Agents in electronic commerce: component technologies for automated negotiation and coalition formation. *Automomous Agents and Multi-Agent Systems* **3**: 73–96

Sloman A 1997 What sort of control system is able to have a personality? In: Trappl R, Petta P (eds.) *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*. Springer (Lecture Notes in AI), Berlin

Tambe M, Johnson W L, Jones R M, Koss F, Laird J E, Rosenbloom P S, Schwamb K 1995 Intelligent agents for interactive simulation environments. *AI Magazine* **1**: 15–39

T. Bösser

Autonomy at Work

Autonomy may be defined as the condition or quality of being self-governing or free from excessive external control. According to Immanuel Kant, the eighteenth-century German philosopher often credited with laying the groundwork for modern philosophy, autonomy is important to human beings because it is the foundation of human dignity and the source of all

morality (Hill 1991). In most Western cultures, an adult person's dignity (poise and self-respect) is usually based on possessing some degree of control in making the decisions that deeply affect their life. Likewise, the source of their morality lies in the volitional self that rises above socially conditioned desires and partisan attachments in order to adopt impartial ethical principles and standards of conduct.

The theme of self-governance and autonomy has been central in theories of human growth and development in society for several decades. More recently, it has moved to the center of theory and research on work and organizational behavior. In this article, two general models concerned with self-governance and human well-being will be reviewed briefly and attention will focus on four streams of research on autonomy at work.

In one prominent approach to theorizing autonomy, the internal–external locus of control model of human personality, it is proposed that individuals perceive that their actions can affect the events in their lives, but vary in the degree to which they believe forces outside their control typically influence the events (Langer 1977). Some people (externals) have a relatively consistent tendency to perceive that the outcomes in their lives are governed by external forces, such as powerful others and social conditions, fate, luck, or chance. Others (internals) have a relatively consistent tendency to believe they control their own destiny; they perceive themselves as masters of their own fate and as in control of the actions and events that affect their lives. This model has been researched extensively, leading to the conclusion that people who do not see themselves as self-governing (externals) are handicapped in numerous ways.

Among the models of human growth and development that are centered on autonomy, the most theoretically sophisticated approach has been developed around the concepts of self-regulation and intrinsic motivation (Ryan and Deci 2000). Self-determination theory proposes that 'higher behavioral effectiveness, greater volitional persistence, enhanced subjective well-being, and better assimilation of the individual within his or her social group' result when individuals act from motivations that emanate from the inner self (intrinsic motivation) rather than from sources of external regulation (Ryan and Deci 2000 pp. 72–3). Although not without controversy (due to the entrenched, commonsensical belief that tangible rewards govern behavior), research in a wide array of social conditions has demonstrated that individuals develop a perceived external locus of causation (i.e., a sense of diminished autonomy) when tangible rewards are administered contingently, based on evaluated performance at a task. For self-determination theorists, it is the experience of an external locus of causation (or the belief that one's actions are controlled by external forces) that undermines the most powerful source of natural motivation and that (when chronic) also can lead to stultification, weak self-esteem, anxiety and depression, and alienation. Thus, health and well-being as well as effective performance in social settings are closely related to the experience of autonomy. Interestingly, Ryan and Deci (2000, p. 74) point out that 'autonomy refers not to being independent, detached, or selfish, but rather to the feeling of volition that can accompany any act, whether dependent or independent, collectivist or individualist.'

During the 1960s, amidst massive social upheaval in many of the advanced industrial nations of the world, organizational and management theorists began to turn their attention to the sources of discontent in the workplace, manifesting in 'accelerating absenteeism, job turnover, shoddy workmanship, sabotage, and the greatest outbreak of wildcat strikes since the 1930s' (Greenberg 1975, p. 192). Several now classic texts called attention to the neo-Marxist thesis of alienation and the notion that lack of control over the product and process of work would lead to intense and widespread feelings of powerlessness, meaninglessness, social isolation, and self-estrangement (e.g., Blauner 1964).

In one of the first systematic attempts to address the phenomenon of work alienation and freedom, Hackman and Oldham (1976) developed the Job Characteristics Model, a framework that focused attention on autonomy and four other key factors involved in designing enriched work. Work designed to be complex and challenging (characterized by high levels of autonomy, skill variety, identity, significance, and feedback) was theorized to promote high intrinsic motivation, job satisfaction, and overall work performance. Two decades of research in this tradition have shown that job scope or complexity, an additive combination of autonomy and the four other job characteristics: (a) is correlated significantly with more objective ratings of job characteristics; (b) may be reduced to a primary factor consisting of autonomy and skill variety; and (c) has substantial effects on affective and behavioral reactions to work, mostly indirectly through critical psychological states such as experienced responsibility for the outcomes of the work. There is some evidence suggesting that the positive effects of complex and challenging work have thresholds, but the main lesson we can draw up from this line of research is that the experience of autonomy at work has positive consequences ranging from higher job performance to enhanced general well-being.

A second stream of research in which autonomy at work is central was launched by Karasek's (1979) study of national survey data from Sweden and the USA. In this research, he proposed the Job Demands–Job Control Model that included a thesis about the causes of physical and mental strain and health problems. Job demands include requirements for working fast and hard, having too much to do and too little time to do it, conflicts among demands, and

personal conflicts with one's role. Job control is specified in terms of decision latitude and relates to the theme mentioned earlier about possessing some degree of control over decisions that deeply affect one's life. Karasek (1979) proposed two dimensions to job control (authority to make decisions on how to work and having say over what happens; and latitude to use a variety of skills on the job), which parallel Hackman and Oldham's (1976) two central concepts of autonomy and skill variety. The main proposition in Karasek's model is that when jobs are simultaneously high in demands and low in decision latitude, physical and mental health problems are more likely to occur—the 'high strain job' (Karasek 1979, p. 288). Demanding jobs that permit high decision latitude are challenging and create opportunities for arousal, learning, and healthful response, thus mitigating the stressful effects of demands and generating the potential for job satisfaction and life fulfillment.

Subsequent research on Karasek's model has been mixed. Most studies have reported that demands have negative effects regardless of the degree of job control, but in other studies, the predicted interaction effect has been found. Researchers have not yet determined what accounts for mixed support of the Job Demands–Job Control Model, but current studies have begun to develop the model further by examining the role of individual differences in moderating the importance of decision latitude. For example, the role of self-efficacy and the role of proactive personality have been explored. It was found that high strain was mitigated when decision latitude was exercised by more self-confident and proactive individuals. As with the Job Characteristics Model, research on Karasek's (1979) framework has underscored the importance of autonomy at work while also revealing some of the intricacies involved in explaining the specific processes through which autonomy exerts its positive effects.

The theme of autonomy at work also runs through five decades of research on groups and teams in organizations. Of primary concern here are the streams referred to by Moldaschl and Weber (1998) as the sociotechnical systems theory approach (STS) and the lean management approach. Both approaches focus on teams that are, to varying degrees, autonomous or self-organizing. Self-organizing teams have been theorized as having processes of self-management and collaborative teamwork. In self-organizing teams, members experience relatively high autonomy by collectively controlling, for example, the pace of work, task allocation, performance evaluation, recruitment and selection of new members, training, and even reward distributions. As noted by several researchers, because groups can undertake much larger pieces of work than individuals can, the potential to experience positive outcomes through heightened control is greatly enhanced.

Beginning with Trist and Bamforth's (1951) classic study of autonomy among underground coal miners, through recent reviews and meta-analyses of research in the STS tradition, some positive results have been reported. Despite 'substantial variance' in research findings regarding the consequences of autonomous work groups (Guzzo and Dickson 1996, p. 326), positive impacts on motivation and productivity (especially performance quality and cost savings), attitudes and perceived quality of work life, and (to a lesser extent) attendance and retention are routinely reported. Failures of some STS projects and the fact that STS interventions in organizations have not been used as widely as initially anticipated have been attributed to poorly designed (imbalanced) programs and internal political conflicts.

The lean management approach, used initially by Japanese manufacturers such as Toyota, differs from the STS approach primarily in the degree of autonomy experienced by team members. The lean management approach tends to be more supervisor-centered as team leaders are appointed rather than elected, decision making is consultative rather than consensual, and the domain of decision making authority for the team is more limited. As pointed out by Ilgen et al. (1993), when team-based structures are created in organizations, there is a tendency to retain an overarching hierarchical structure and to require teams to report to leader-managers in a very formal and traditional way. In situations where teams are teams mostly in name, autonomy is very limited, leading to negative results. When interventions empower teams to become more self-directed and autonomous, results tend to be positive across a number of indicators, including financial performance.

Although autonomous and self-organizing teams can produce positive results, the approach assumes that freeing members from hierarchical control automatically creates organizational democracy and emancipated practice. Some writers contend that while workers in the typical self-managed team situation might feel a measure of influence, upon closer inspection, it is only an illusion of self-control. This theme is developed further by Ezzamel and Willmott (1998) and others who show that teams, through peer pressure and behavioral regulation, can coerce their members in ways that are perhaps even more tyrannical than are the forms of control exercised by managers.

The fourth and final stream of research on autonomy at work that we will touch on is best labeled organizational democracy, and is more familiar to researchers in Europe than in North America (see the academic journal *Economic and Industrial Democracy* 1980–2001 volumes, for a review of this approach). This tradition is vitally important because it helps sharpen distinctions among the varieties of macrofactors operating in the world that shape self-regulating organizational systems and their effectiveness. In line with this approach is the conclusion that self-directed teams tend to be more effective in

nations other than the USA. Why should this be so? Studies of autonomy at work from the standpoint of organizational democracy provide some answers to this question and open the field up in at least two important ways. First, they suggest how individualistic and collectivist nation-state cultures shape attitudes and other reactions toward team-based organizing, in effect constraining and prefiguring the effectiveness of any intervention aimed at enhancing autonomy at work. This point corresponds with Guzzo and Dickson's (1996) statement that the effects of team-based organizing are highly situationally dependent and extends it into the realm of nation-state contingencies. Second, they suggest a need to develop programs of inquiry that go beyond the psychology of autonomy, programs that describe and explain the emergence of objectively different structures of control that are anchored in political–economic, legislative, and cultural factors (see Forbes and Jermier 1995). This is important because so long as studies of autonomy at work rely exclusively on micro thinking, the politics of 'empowerment' in any given culture at any given time will probably be limited to regimes of illusory control.

The four streams of research on autonomy at work reviewed in this essay all support the idea that being truly self-governing and free from excessive external control produce positive outcomes for individuals, groups, and work organizations. The more general thesis, that autonomy promotes human growth and development as well as positive health and well-being, is also supported. Autonomy sometimes is considered to be synonymous with individualism and egoistic willfulness but there is increasing evidence supporting the idea that collectivist autonomy is equally beneficial. The concept of autonomy at work has received considerable attention from researchers in several traditions of study. While more research in each of the traditions is needed, a more pressing need is to begin integrating insights from all the traditions to develop a general theory of autonomy at work

See also: Alienation: Psychosociological Tradition; Alienation, Sociology of; Autonomy, Philosophy of; Group Processes in Organizations; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Industrial Sociology; Job Analysis and Work Roles, Psychology of; Job Design, Psychology of; Self-efficacy; Self-monitoring, Psychology of; Self-regulation in Adulthood; Work: Anthropological Aspects; Work, Sociology of

Bibliography

Blauner R 1964 *Alienation and Freedom*. University of Chicago Press, Chicago

Ezzamel M, Willmott H 1998 Accounting for teamwork: A critical study of group-based systems of organizational control. *Administrative Science Quarterly* **43**: 358–96

Forbes L C, Jermier J M 1995 Industrial democracy in Europe: A review essay. *Organization Studies* **16**: 1080–6

Greenberg E S 1975 The consequences of worker participation: A clarification of the theoretical literature. *Social Science Quarterly* **56**: 191–209

Guzzo R A, Dickson M W 1996 Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology* **47**: 307–38

Hackman J R, Oldham G R 1976 Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance* **16**: 250–79

Hill T E Jr 1991 *Autonomy and Self-respect*. Cambridge University Press, Cambridge, UK

Ilgen D R, Hollenbeck J R, Sego D J, Major D A 1993 Team research in the 1990s. In: Chemers M M, Ayman R (eds.) *Leadership Theory and Research*. Academic Press, San Diego, CA, pp. 245–70

Karasek R A Jr 1979 Job demands, job decision latitude, and mental strain: Implications for job redesign. *Administrative Science Quarterly* **24**: 285–308

Langer E J 1977 The psychology of chance. *Journal for the Theory of Social Behavior* **7**: 185–208

Moldaschl M, Weber W G 1998 The 'three waves' of industrial group work: Historical reflections on current research on group work. *Human Relations* **51**: 347–88

Ryan R M, Deci E L 2000 Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* **55**: 68–78

Trist E, Bamforth K 1951 Some social and psychological consequences of the longwall method of coal-getting. *Human Relations* **4**: 3–38

J. M. Jermier and C. E. Michaels

Autonomy, Philosophy of

The notion of personal autonomy (or free agency) is ubiquitous within the philosophical literature. It is construed variously as a fundamental right which justifies further rights, such as those of abortion or free speech; as a basic political good of liberal political theory; as a character ideal on a par with courage or integrity; and as a necessary condition of full moral agency. Although there are many different conceptions of autonomy, there is a common theme. Being autonomous is usually taken to mean being self-governing in one way or another; an autonomous state is one in which a person's will is genuinely *their own*.

Individual autonomy (including moral autonomy) is to be distinguished from political autonomy. Political autonomy is a property of nations or states, and is usually characterized as requiring the *absence* of certain autonomy-undermining conditions. For instance, a nation is self-governing or self-determining

when it is not subject to interference or coercion by others. Individual autonomy requires that agents satisfy some positive conditions in addition to the absence of interference. Consider, for example, two typical failures of personal autonomy: first, agents who habitually and uncritically adopts others' social preferences and values as their own; and, second, agents who have so internalized the outlooks and morés of the family in which they grew up that these are treated as authoritative without critical reflection (Feinberg 1986, p. 32). Although in neither case is there any overt interference or coercion, agents in both cases lack autonomy because they fail to exhibit a characteristic necessary for autonomy—for example, a capacity for 'subjecting these governing ideals to rational criticism and then modifying them when necessary' (Feinberg 1986, p. 32). The development of such positive capacities can be undermined by oppression, and therefore the articulation of a conception of autonomy is an important tool for feminists and other social critics.

Conceptions of autonomy in the contemporary literature divide roughly into two classes: Kantian conceptions and nonKantian—or procedural—conceptions. Broadly speaking, each reflects a different approach to the following question: How can agents be genuinely self-governing in the face of the physical and social forces in which they are embedded, and which appear to determine their states and capacities? Traditional Kantian conceptions endorse a version of *incompatibilism*, namely the thesis that genuine free agency cannot be reconciled with socialization or physical determinism. They characterize autonomy by introducing either idealized conditions of free agency or metaphysical conceptions of the autonomous self. Non-Kantian conceptions are *compatibilist*, and attempt to explicate autonomy within the network of physical and social forces in which agents operate. The key task of these conceptions is to distinguish the conditions in which socialization undermines autonomy from the conditions in which it does not.

This article traces the main theories of autonomy in the current philosophical literature. The first section outlines the traditional Kantian account, which conceives of moral autonomy as an ability of rational agents to formulate universal moral principles. These moral principles are said to originate in the will of the agent and to be, in this sense, the agent's 'own.' The second section turns to three procedural accounts of autonomy: endorsement, historical, and self-knowledge accounts. Each proposes a condition on critical reflection which distinguishes autonomous from nonautonomous preference formation. The third section describes neo-Kantian responses to procedural analyses. These propose that autonomy requires a kind of normative competence, namely an ability to identify norms, for example moral norms, or norms that are 'correct' for an agent to endorse. The final section outlines two further approaches—psychological and self-realization—both of which focus on the relationship between conceptions of autonomy and the nature of the autonomous agent.

1. *The Kantian Conception of Autonomy*

A major strand in theories of autonomy is derived from the philosophy of Immanuel Kant. For Kant, autonomy has a special meaning: 'Autonomy of the will is the property the will has of being a law unto itself (independently of every property belonging to the objects of volition)' (1785/1948, p. 108). Rational beings make the moral law for themselves, and can regard themselves as authors of the law. Thus autonomy is manifested when rational agents 'will' the moral law. An important feature of the moral law is that it is a categorical, not a hypothetical, imperative. When agents formulate a categorical imperative, they formulate a maxim of the form 'I ought to do this,' not of the form 'I ought to do this if I desire that.' For example, agents formulate the moral law when they will 'I ought not to lie,' but formulate a hypothetical imperative when they will 'I ought not to lie if I desire not to be punished for lying.' The act of formulating the moral law is an act of a pure autonomous will: it is untainted by the influence of the desires and interests that an agent may have relative to a particular situation. The pure autonomous will amounts to a metaphysical conception of the self, to be distinguished from empirical conceptions of the self. The metaphysical self is unconstrained by desires, interests and social forces. Thus Kant's conception of freedom is incompatibilist.

A contemporary parallel of Kant's conception of autonomy is John Rawls' influential notion of free and rational agents formulating principles of justice in the 'original position' (Rawls 1971). Rawls argues that rational agents formulate principles of justice from behind a 'veil of ignorance,' namely from a position in which they are making decisions about how a society will function before they know who in that society they will turn out to be. In the original position, they have abstracted themselves from a particular location in the society to which the principles of justice apply. This means that such agents, like Kantian agents, are uninfluenced by the particular desires and preferences that arise from being embedded in an actual situation. Because of this, their formulation of the principles of justice is genuinely *self*-originating.

The Kantian conception of autonomy has two aspects. First, autonomy is a function of a conception of morality that is encapsulated in the categorical imperative, or, in Rawls' account, in the principles of justice. The categorical imperative provides a standard according to which we can judge whether a particular example of moral reasoning is autonomous. If the moral reasoning issues in the categorical imperative, then it is autonomous. Otherwise, it is not. Second,

autonomy is a function of a metaphysical conception of the self, because autonomy manifests itself when the Kantian 'true' self is expressed through the formulation of the moral law. Both aspects of the Kantian view are contentious. Critics challenge the notion of a universally valid moral law as well as the concept of a pure autonomous will, which is said to be implausibly metaphysical. In addition, Rawls' account of the agent operating in the original position has been criticized as being unfaithful to empirical reality. Feminists and communitarians have argued that this agent is 'atomistic'; that is, existing in a vacuum divorced from the social relations in which actual agents are embedded (e.g., Sandel 1982). No such agent exists, because agents are always the product of family influences, socialization, and so on. Moreover, an agent stripped of desires, interests, and preferences cannot engage in the imaginative moral reasoning required to formulate principles of justice. A further difficulty is that the Kantian conception is said to lead to an impoverished account of the *value* of autonomy. It is associated with the claim that autonomous agents are, and ought to be, substantively independent or self-sufficient, which in turn is associated with the character ideal of the 'self-made man.' Critics—especially feminists—have challenged this character ideal, and questioned whether substantive independence is really a value that a theory of autonomy, or normative theories in general, should promote. They have argued that a conception of autonomy should not exclude the possibility of agents having significant relationships of interdependency, such as those obtaining with family members (see Friedman 1997).

2. *Procedural Theories of Autonomy*

Procedural theories eschew substantive independence in favor of procedural independence. It is enough that agents' reasoning processes satisfy certain formal and procedural conditions for the outcomes of those processes (decisions, preferences, and desires) to count as autonomous. Procedural accounts are 'content neutral.' There are no specific beliefs, goals, or values that the autonomous agent must adopt. Nor are there any standards or conditions external to the agent that operate as touchstones of autonomous reasoning or place constraints on the contents of agents' preferences and desires. Procedural theories, therefore, are congenial to the concerns of critics of the Kantian conception.

2.1 *Autonomy as Endorsement*

Perhaps the most prominent procedural approaches are those of Gerald Dworkin and Harry Frankfurt (Dworkin 1970, 1988, Frankfurt 1971). Both argue that what distinguishes autonomous from nonautonomous critical reflection is endorsement—the agents' sense of engagement or identification with their preferences or desires. In Frankfurt's terms, an autonomous will is *the will the agent wants*. Consider the difference between a willing and an unwilling drug addict. Frankfurt argues that although both are addicted, only the unwilling addict is nonautonomous. This person is a 'passive bystander' with respect to the desire to use drugs, whereas the willing addict engages and identifies with that desire, and hence endorses it. In Frankfurt's account, the idea of endorsement is spelled out using a hierarchical distinction between different levels of desire. First-order desires—such as a desire to inject heroin—are autonomous just in case they are endorsed by second-order desires—such as a desire to desire to inject heroin (Frankfurt 1971).

Important objections have been raised to endorsement accounts. The first focuses on their hierarchical nature. Even if first-order desires can be characterized as endorsed with respect to second-order desires, this is not sufficient to ensure autonomy unless the second-order desires themselves are endorsed. The problem results in a regress unless we introduce a qualitative distinction among levels, for example, between higher-order *values* and first-order *desires* (Watson 1975). However, the regress objection can be generalized to many hierarchical accounts, because whatever the nature of the higher-order level, it can always be questioned whether it truly represents free agency. The second-order level itself will always be the outcome of the social forces influencing an agent. A second objection is that the endorsement condition is too stringent, because it seems that certain unendorsed desires should count as autonomous. Suppose that, out of habit, an agent chooses to go on vacation to the same destination every year. The fact that the agent fails to attend critically to the desire and does not endorse it each time should not rule out its being autonomous. Third, it has been suggested that the endorsement condition is too weak. The endorsement account relies on the idea that although the unwilling addicts inject themselves intentionally with heroin, they prefer not to. Since they prefer not to inject themselves, they do not endorse the desire to inject themselves. It has been argued that (all things considered) intending to do something and preferring to do something are indistinguishable, and thus if agents intend to so something, they also prefer to do it, and endorse it. Therefore, insofar as the unwilling addicts intend to inject themselves with the drug, they also endorse it. The objection claims that endorsement merely distinguishes intentional from nonintentional agency, and not autonomous from nonautonomous agency. The apparent lack of autonomy of the unwilling addicts is therefore not due to a lack of endorsement (Buss 1994).

2.2 *Autonomy as Historical Critical Reflection*

Most hierarchical approaches to autonomy are structural. They propose conditions that an agent's existing

motivational state must satisfy. Alternative approaches introduce conditions that allow reflection on the processes of development of the motivational state. For example, Gerald Dworkin analyzes autonomy as 'a second-order capacity of persons to reflect critically on their first-order preferences, desires, wishes and so forth and the capacity to accept or attempt to change these in the light of higher-order preferences and values' (1988, p. 20). A capacity allows reflection on the processes by means of which an agent acquires desires and preferences. It also distinguishes between an agent's autonomy on the one hand, and a particular exercise of autonomy on the other. A related approach is the *historical criterion of autonomy*. This proposes that agents are autonomous with respect to a certain preference or desire only if they do not resist its development when attending to the process of its development, or would not have resisted had they attended to the process (Christman 1990, 1991). Notice that this criterion can account for preferences formed out of habit because, although in most cases of habit agents do not attend each time to the process of development of their preferences, it is true that they would not have resisted the process had they attended to it.

The historical criterion is difficult to maintain in the face of apparent counterexamples. Consider agents who have internalized an oppressive ideology—for example, a young student who is so influenced by the norms of the fashion industry that an excessive amount of time and energy is devoted to the student's appearance. Such agents are usually characterized as nonautonomous with respect to the norms of the ideology, yet the historical condition cannot adequately explain why. Even if the student had reflected on the fashion norms that have been internalized, as well as the processes of development of the student's preferences and their relationship to the norms, the norms or the development of the preferences would probably not have been resisted. The very fact that the ideology is effectively internalized makes it unlikely that an agent would resist the development of preferences based on the ideology (Benson 1991). The historical condition therefore does not account for cases that are thought to be paradigms of nonautonomous preference-formation.

2.3 Autonomy as Self-knowledge

Another set of procedural conceptions proposes that authenticity is the key factor in explicating autonomy. Having one's own will is having the will one *really* wants. One important approach characterizes authenticity in terms of self-knowledge. For example, Diana Tietjens Meyers argues that self-discovery is a necessary capacity for achieving autonomy. Suppose an agent—Martin—has 'always been groomed to follow in his father's footsteps [and] his parents have taught him to feel deeply guilty whenever he disappoints their expectations' (Meyers 1989, p. 46). He completes medical school dutifully and becomes a successful surgeon. Although there are opportunities along the way for Martin to 'probe' his feelings and reconsider his career path, it does not occur to him to do so. Meyers suggests that Martin's 'insensitivity to his own responses' and 'blindered approach to life' indicate that whereas 'He may be doing what he wants ... there is no reason to suppose that he is doing what he really wants' (Meyers 1989, p. 47). The distinction between doing what one wants and doing what one really wants is important for a range of cases in which authenticity is questionable. For instance, conventional gender roles can shape values and desires through being omnipresent in culture, or instilled since birth by family socialization. Meyers points out that the enforcement of gender roles can interfere with a person's ability to distinguish between apparent and authentic desires.

This conception of autonomy must give a plausible account of the difference between 'real' and 'apparent' desires, as well as identify how self-knowledge fails in cases of failures of autonomy. One possibility is to argue that agents lack self-knowledge by having inaccurate beliefs about themselves, specifically about the desires and motivations in question. Perhaps Martin's 'real' wishes are being covered up by his failure to scrutinize. The desires to be a doctor operate on the surface, but his deep, 'true' desires are to follow some other career path. Since his beliefs that he wants to be a doctor are based on apparent desires, they are false beliefs about himself, and he lacks self-knowledge. A difficulty for this characterization however is that it may seem to employ a metaphysical notion of the self as the bearer of real desires. Even if it does not, it must rely on the agent experiencing internal conflict among desires, and failing to probe or scrutinize in response to such experiences. The apparent desires are the unscrutinized ones, and the real desires are the ones that would have emerged had the agent scrutinized the apparent desires. Often, however, there is no experienced dissonance among desires, and no failed opportunity for real desires to emerge. How should we characterize the lack of self-knowledge in these cases?

A second way in which agents could lack self-knowledge is through failing to acquire their beliefs in the right way. Suppose there is no opportunity for Martin's desires to change because he does not experience appropriate internal conflict. Nevertheless, *if* Martin's desires or motivations *were* to change, his beliefs about himself ought to change as well. We can say that: if Martin's beliefs about himself would not have changed if his desires had changed, then his beliefs have not been acquired in the right way. Although they happen to be accurate, they do not

constitute self-knowledge. This approach revises the concept of authenticity, because it does not employ a clear distinction between real and apparent desires. Rather, it relies on articulating the ways in which beliefs must be acquired to constitute self-knowledge.

3. *Neo-Kantian Theories*

Procedural accounts usually assume that the conditions they identify, which are internal to agents' own thought processes, are both necessary and sufficient for autonomy. Neo-Kantian conceptions argue that agents' preferences are subject to certain *external* constraints. Hence, although procedural criteria may be necessary for autonomy, they are not sufficient.

The central idea of recent neo-Kantian approaches is that of normative competence: the failure of autonomy corresponds to a failure of a capacity to identify the difference between right and wrong. Consider JoJo, the son of an evil and sadistic tyrant. JoJo is raised to respect his father's values and emulate his desires, so that he thoroughly internalizes his father's evil and sadistic world view. Let us suppose that, on procedural theories, JoJo counts as autonomous because he identifies in the appropriate ways with his first-order desires, has the desires he really wants, etc. It has been proposed that he is neither free nor morally responsible because his upbringing has blocked his capacity to distinguish right from wrong (Wolf 1987). Similarly, the young student who has internalized the ideology of the fashion industry can be characterized as having internalized certain *false* norms, such as that most people's natural physical appearance is deficient. The internalization of the norm blocks the agent's capacity to criticize effectively this 'false construal of personal value,' and hence the ability to distinguish correct from incorrect norms—the student's normative competence—is flawed (Benson 1991). It is the failure of normative competence that explains the failure of autonomy.

One controversial issue raised by normative competence accounts is that of the status of the moral and other norms that autonomous agents identify. For example, it has been argued that moral norms are objective because they derive from the requirements of objective reason (Wolf 1990). If this is right, the capacity required for autonomy is a capacity to 'track' objective features of the world. Normative competence theories therefore place strong constraints on the contents of agents' preferences. Preferences for incorrect norms or morally wrong courses of action—that is, preferences that do not match the relevant objective features of the world—are not autonomous. The idea that morality is objective—moral realism—is controversial, despite being a well-known and widely supported position. Moreover, realism about non-moral norms is often thought to be less plausible than moral realism. An additional difficulty for normative competence accounts is the charge that they conflate autonomy with moral responsibility. In the case of JoJo, for example, it may be that the comprehensiveness of his socialization absolves him from full moral responsibility for his acts, but does it follow that the acts are not the product of his own agency? The wish to maintain a conceptual distinction between responsibility and autonomy has led to a revision of the normative competence approach (Benson 1994).

4. *Psychological and Self-realization Approaches*

Neither procedural nor neo-Kantian conceptions of autonomy give much emphasis to the nature of the autonomous agent, and how psychological and other features of the agent may be necessary to produce and maintain autonomy. Several recent theories developed by feminists and others address this aspect of the problem, focusing particularly on the putative incompatibility between autonomy and oppression. On psychological approaches, certain kinds of psychological impairment, such as impairments of self-confidence and self-esteem, are construed as undermining autonomy. Consider a person who is believed (erroneously) by family and the medical establishment to be mentally unstable because of being passionate, excitable, and 'prone to emotional outbursts in public' (Benson 1994). Since these external judgments of character and behavior are trusted by the person, this causes disorientation, and loss of self-confidence and self-esteem. The person's will is still intact, because preferences and desires can still be put into effect, but due to the lack of a sense of self-worth, the person also lacks certain important attitudes to decision-making that seem to be necessary for decisions to be autonomous. The person is not 'behind' those decisions and does not treat them as authoritative. Such failures of a sense of assurance about the capacity to form preferences and make decisions (for example, failures of self-confidence, self-trust, or a sense of self-worth) can be caused by a variety of psychological states, including many arising from oppression.

Self-realization conceptions also focus on the inauthenticity of desires in the context of oppression. Meyers's account, for example, argues that an autonomous self is one that has been expressed or 'realized' through the exercise of the capacities of self-definition and self-direction as well as that of self-discovery. The account emphasizes the way in which these capacities can be undermined or enhanced by different kinds of socialization, especially oppressive socialization. Meyers's account is meant to be procedural, because it does not place constraints on the outcomes of the

exercise of the capacities necessary for autonomy. As long as agents' capacities for self-discovery, self-direction, and self-definition are fully developed, they are fully autonomous. This suggests that preferences endorsing one's own oppression may in principle be fully autonomous on Meyers's account. It has been questioned whether such preferences are compatible with self-realization, and hence whether self-realization accounts can be purely procedural (Mackenzie and Stoljar 2000b, pp. 18–19). Self-realization has also been linked to a substantive notion of human flourishing. It has been proposed that an influence on an agent undermines autonomy when it interferes with the agents' normal functioning (Buss 1994, p. 106). 'Normal functioning' is spelled out using a normative notion—that of well-being, or human flourishing—and thus the agents' reasoning processes are subject to a normative constraint.

Psychological and self-realization accounts have several advantages. First, they do not rely on metaphysical conceptions of the self, but rather situate agents within the network of social relations that affect them. As a result, they take an important step towards understanding the effects of oppression, and how oppression is incompatible with autonomy. Second, these accounts do not require that autonomous agents have preferences and desires with specific contents, and nor do they appeal to objective conceptions of value. They do, however, place weaker external constraints on the contents of desires and preferences. For example, if a lack of self-trust undermines autonomy, desires with contents that are incompatible with self-trust are nonautonomous. Thus these accounts are not purely procedural or content neutral.

There are nevertheless questions that can be raised for both psychological and self-realization approaches. For example, how exactly is the psychological impairment (that is, the lack of self-esteem, self-trust, or self-confidence) tied to the failure of autonomy? One suggestion is that the absence of self-worth erodes agents' sense of themselves as competent 'to answer for one's conduct in light of normative demands that, from one's point of view, others might appropriately apply to one's actions' (Benson 1994, p. 660). Agents therefore are not autonomous, because they do not see themselves as authoritative with respect to the normative demands the community places on them. However, in certain typical cases of nonautonomous decision-making, this link between the lack of a sense of self-worth and the lack of normative competence does not seem to be present. In the case of the young student who has internalized fashion ideology, one may suppose that a low sense of self-worth accompanies commitment to these norms. Yet it is not clear that the student's lack of self-esteem interferes with the sense of self as being normatively competent to answer the fashion world's expectations. Indeed, the student seems to wish to increase the sense of self-worth by living up to these norms. In this sense, the student is very much 'behind' the commitment to these norms. A second issue is whether the psychological impairment itself is constitutive of the lack of autonomy, or rather if it causes the breakdown of some other constitutive process. Psychological conceptions claim that the reasoning capacities of agents are fully intact despite the psychological impairment, and hence the failure of autonomy is due to the absence of a necessary psychological or emotional state, such as that of sense of self-worth. A proponent of a procedural conception may respond that an agent's lack of a sense of self-worth makes it impossible for critical reflection to operate appropriately. Hence it leads to a procedural failing that is responsible for the agent's lack of autonomy. A third issue concerns the notion of human flourishing that is central to self-realization accounts. The value of flourishing operates as a constraint on autonomous reasoning in the same way as do the moral and other norms of Kantian conceptions, and the same questions can be raised. For example, is the self-realization conception committed to some objective and universal notion of flourishing that autonomous reasoning must reflect? If so, it may be appealing to an implausible—or at least contentious—theory of flourishing.

See also: Freedom/Liberty: Impact on the Social Sciences; Freedom: Political; Identity in Childhood and Adolescence; Justice and its Many Faces: Cultural Concerns; Kant, Immanuel (1724–1804); Personal Identity: Philosophical Aspects; Self-knowledge: Philosophical Aspects

Bibliography

Benson P 1987 Freedom and value. *Journal of Philosophy* **84**: 465–86

Benson P 1991 Autonomy and oppressive socialization. *Social Theory and Practice* **17**: 385–408

Benson P 1994 Free agency and self-worth. *Journal of Philosophy* **91**: 650–68

Buss S 1994 Autonomy reconsidered. In: French P A, Uehling T A, Wettstein H K (eds.) *Midwest Studies in Philosophy,* XIX. University of Minnesota Press, Minneapolis, MN

Christman J (ed.) 1989 *The Inner Citadel: Essays on Individual Autonomy*. Oxford University Press, New York

Christman J 1990 Autonomy and personal history. *Canadian Journal of Philosophy* **20**: 1–24

Christman J 1991 Liberalism and individual positive freedom. *Ethics* **101**: 343–59

Dworkin G 1970 Acting freely. *Nous* **3**: 367–83

Dworkin G 1988 *The Theory and Practice of Autonomy*. Cambridge University Press, New York

Feinberg J 1986 Autonomy. In: (eds.) *Harm To Self*. Oxford University Press, Oxford, UK; reprinted in *The Inner Citadel*

Frankfurt H 1971 Freedom of the will and the concept of a person. *Journal of Philosophy* **68**: 5–20 (reprinted in *The Inner Citadel*)

Frankfurt H 1988 *The Importance of What We Care About*. Cambridge University Press, Cambridge, UK
Friedman M 1997 Autonomy and social relationships: Rethinking the feminist critique. In: Meyers D T (ed.) *Feminists Rethink the Self*. Westview Press, Boulder, CO
Hill T E Jr The Kantian conception of autonomy. In: *The Inner Citadel*
Kant I 1785/1948 *Groundwork of the Metaphysic of Morals* (trans. and analyzed Paton H J). Harper & Row, New York
Mackenzie C, Stoljar N (eds.) 2000a *Relational Autonomy. Feminist Essays on Autonomy, Agency and the Social Self*. Oxford University Press, New York
Mackenzie C, Stoljar N 2000b Introduction: Autonomy refigured. *Relational Autonomy*
Meyers D T 1989 *Self, Society and Personal Choice*. Columbia University Press, New York
Rawls J 1971 *A Theory of Justice*. Oxford University Press, Oxford, UK
Rawls J 1980 Kantian constructivism in moral theory: The Dewey lectures. *Journal of Philosophy* **77**: 515–72
Reath A 1998 Autonomy, ethical. In: *Routledge Encyclopedia of Philosophy*. Routledge, London
Sandel M 1982 *Liberalism and the Limits of Justice*. Cambridge University Press, Cambridge, UK
Watson G 1975 Free agency. *Journal of Philosophy* **72**: 205–20
Wolf S 1987 Sanity and the metaphysics of responsibility. In: Schoeman F (ed.) *Responsibility, Character and the Emotions*. Cambridge University Press, New York
Wolf S 1990 *Freedom Within Reason*. Oxford University Press, New York

N. Stoljar

Avant-garde Art and Artists

Since the concept of the avant-garde appeared at the beginning of the nineteenth century, it has been applied to a wide variety of aesthetic and social practices. Analysis of different uses of the term shows that it has been applied to three types of changes: in the aesthetic content of art, in the social content of art, and in the norms surrounding the production and distribution of artworks (Crane 1987, pp. 14–15). For example, the term avant-garde is applied to the aesthetic content of art objects: (a) when those works represent a redefinition of conventions for creating art, often in such a way that they are perceived as violating taboos and as shocking or offensive; (b) when they involve the use of new tools and techniques, or concern the nature or use of techniques *per se*; and (c) when they redefine what can be considered an artwork. The term is applied to the social content of artworks when they express social or political values that are critical of or different from the dominant culture, when they attack art institutions, and when they attempt to redefine the boundaries between high and popular culture. Finally, the term is applied to creators of artworks when they attempt to alter the social context for the production of art (for example, appropriate role models, critics, and publics for artists), the organizational context in which art is displayed and distributed, and the social role of the artist in terms of his or her participation in other social institutions such as education, religion, and politics. This article examines the evolution of artistic avant-gardes since the nineteenth century, theories of the avant-garde, and sociological research on the social organization and reception of avant-gardes.

1. *Origins and Evolution of the Term*

Originally referring to a section of an army that marched ahead of the troops, the term avant-garde was first used in France in the 1830s to refer to artists as leaders and creators of a new social order (Nochlin 1968, p. 5). Until the middle of the nineteenth century, the term generally meant that an artist was both politically and aesthetically progressive. The artist was expected to lead society in new directions and to be in conflict with tradition and the Establishment. Subsequently, the term began to refer to artists whose works expressed alienation from bourgeois society in the form of ironic or destructive commentaries on social and artistic values rather than commitment to specific programs for social or aesthetic change. These different elements associated with avant-gardes—political rebellion, aesthetic rebellion, and alienation—continued to be important in the twentieth century.

At the beginning of the twentieth century, the idea of the avant-garde as alienation was given a new dimension by Marcel Duchamp who created artworks or 'ready-mades' by making minor changes in existing artworks or by arbitrarily declaring that commonplace objects, such as a mass-produced urinal, were artworks. This conception of the avant-garde became the basis for the Dadaist movement consisting of groups of artists, located in several European cities and in New York, whose members attempted to destroy artistic conventions and to shock the public. In this period, the Dadaists shared with the Italian Futurists (whose aesthetic innovations were very different) 'a belief that they were ahead of the society in which they lived and that they were breaking social bonds of restraint' (Taylor 1968, p. 97). This attitude led the Futurists to engage in flag-burnings and to stage assaults on theaters and opera houses, although they were not actually interested in changing the social order but in aesthetic change and in 'continuous revolution as a way of art' (p. 100).

During the twentieth century, the term was applied to groups of artists who developed new aesthetic programs, many of them concerned with problems of visual perception and the interpretation of visual experience. This type of art dealt with aesthetic and

technical issues and was detached from the concerns of everyday life. Although the term avant-garde appeared before the development of modernism in the arts, there was a pronounced affinity between avant-gardes and modernism which was reflected in the modernist's commitment to the development of a particular style. At the end of the 1930s, Clement Greenberg, an American art critic, provided a theoretical justification for the type of avant-garde that has proven to be most characteristic of the second half of the twentieth century, one which engages in art for art's sake and which is concerned with aesthetic values rather than with social conflict or political struggle (Orton and Pollock 1981, p. 322). Another American art critic, Harold Rosenberg (1968), argues that the suppression of avant-garde art by the Communists and later by the Nazis marked the beginning of the separation of political and artistic avant-gardes. As an example of this rupture during the postwar period, he cites the 1968 Venice Biennale where militant students denounced the artworks as 'art for dealers and the rich.'

In the late twentieth century, the success of Pop art, Conceptual art, and Neo-Expressionism signaled both the triumph of the ironic, alienated conception of the avant-garde and a decline in the priority attached to aesthetic innovation, reflecting the increasing influence of postmodernism in all forms of culture. Unlike the modernist, the postmodernist is not interested in style conceived as a consistent, integrated set of aesthetic elements or in using a style to express opposition to the dominant culture. Instead, the postmodernist engages in *pastiche*, bringing together disparate elements from many previous texts, regardless of whether they produce a coherent unity. According to Goldman (1992, p. 214), the meaning of postmodernist texts cannot generally be discerned from analyzing the texts themselves but only by asking where the elements in the text come from. Postmodernist works tend to expose or reinterpret rather than criticize or attack the dominant culture.

That the concept of the avant-garde is still relevant today appears to be confirmed by the intensity of controversies surrounding government funding for art works that challenge conventional attitudes and values. The allocation of funds by the United States government agency, the National Endowment for the Arts, to support exhibitions of artists such as Robert Mapplethorpe and André Serrano, whose works were perceived as having transgressed moral and religious boundaries, created enormous opposition and even pressure to dissolve the agency itself. However, it is significant that these artists were not members of art movements and that their activities did not engender such movements. Dunn (1991, p. 124) argues that, in the highly fragmented and ambiguous context of postmodern culture, artistic avant-gardes have been replaced by cultural avant-gardes to which the term avant-garde is seldom applied. These movements engage in collective protest and challenge tradition and hegemonic structures on a wide range of social and cultural issues such as class, gender, sexuality, race, peace, and ecology.

2. *Theories of the Avant-garde*

Theorists who have written about the avant-garde have generally based their analyses on certain types of avant-gardes, particularly those that have rebelled against the dominant culture and excluded others from consideration. A major strand of theory concerning the avant-garde derives from the critical theorists, known as the Frankfurt School. Adorno despised popular culture produced by culture industries on the grounds that it provided shallow entertainment that distracted and manipulated the unsophisticated public. He interpreted the role of high culture and its avant-gardes as that of challenging dominant institutions and conventional ways of thought. For Benjamin (1968), the significance of Dadaism as an avant-garde was that the Dadaists deliberately attempted to produce art works that lacked an aura of uniqueness and authenticity, thus undermining the status of the arts as sacred.

Two leading theorists of the avant-garde are Bürger (1984) and Poggioli (1968). A contemporary critical theorist, Bürger examines the activities of avant-gardes and the social content of their work, arguing that what is most characteristic of avant-gardes is their challenge to art institutions. Instead of creating works of art that were autonomous, complete, and separate from social life, the avant-garde attempted to create works that challenged the public to make sense of them and to relate them to their daily lives. On this basis, he restricts the title of avant-garde to a small number of movements, including Dadaism, early Surrealism, and some aspects of Futurism.

Poggioli (1968) argues that the dominant characteristic of avant-garde art is alienation in all its forms—psychological, social, economic, historical, aesthetic, and stylistic. Bourgeois, capitalist, and technological society provides avant-garde artists with freedom to create while the tensions underlying this type of society provide them with 'a reason for existing.' In Poggioli's view, the avant-garde artist is motivated by nostalgia for an earlier period when the artist was treated as a creator rather than as a parasite and a producer.

Huston (1992) is critical of these theories on the grounds that they are unable to explain and often do not even consider factors affecting the production and social reception of avant-garde arts. He argues that most styles of avant-garde art ultimately succeed in winning the support of the art establishment which initially rejected them but that theories of the avant-garde are not helpful in explaining these developments. In his view, these theories tend to legitimate avant-

garde art as *true* art and 'a certain vision of the artist as a great artist' (p. 78). In other words, he sees the concept of the avant-garde as an ideological category that accepts uncritically artists' conceptions of their situation and tends to be used in such a way as to legitimate certain types of art rather than others (p. 79).

These theories of the avant-garde can also be faulted on the grounds that they treat the distinction between high culture and popular culture as unproblematic and view high culture and its avant-gardes as being the most influential and prestigious forms of culture. In fact, the enormous proliferation of different forms of popular culture (film, television, and popular music) has made it difficult to ignore the aesthetic influence of these cultures on traditional arts and in everyday life, and has, at the same time, decreased the influence of the traditional arts such as painting, theater, dance, poetry, and experimental music. The latter tend to be appreciated by distinct and relatively small subgroups within the middle and upper classes, generally located in a few major cities.

Bauman (1997, p. 95) argues that avant-gardes have disappeared in this postmodern context because there is no longer any clear-cut distinction between styles that can be considered more or less advanced or more or less progressive than others. Instead, while there is a great deal of change in contemporary arts, these changes are 'random, dispersed, and devoid of clear-cut direction.' The importance of an artwork is determined less by aesthetic elements than by publicity and the notoriety that results from it.

3. *Sociological Studies of Avant-gardes*

While theorists of the avant-garde such as Bürger and Poggioli emphasize the aesthetic, political, and ideological programs of avant-gardes, sociological research has focused on (a) the relationships among artists and (b) the reception of artworks by arts institutions and by the public. The dominant view underlying recent studies in the sociology of art is that the significance of all types of art is socially constructed. The question becomes one of identifying the social processes through which this takes place. A major concept is that of the art world, in which art is produced through the efforts of artists in collaboration with members of many other occupations (Becker 1982).

White and White (1965) document changes in social and institutional structures that supported the activities of artists in nineteenth-century Paris. These changes facilitated the emergence of avant-garde styles. Specifically, the transition from academic art to Impressionism, a style that transformed the aesthetic conventions of the period, necessitated the development of a new occupation for buying and selling paintings, that of the art dealer, and the appearance of a new type of collector, drawn from the expanding middle class.

Becker (1982, pp. 304–5) does not use the term avant-garde, preferring to use the concept of 'artistic revolution.' The latter is set in motion by ideological and organizational activities among art world participants. Ideological changes are revealed by the appearance of 'manifestos, critical essays, aesthetic and philosophical reformulations, and revisionist histories,' while organizational changes lead to shifts in control over 'sources of support, audiences, and distribution facilities.' Like White and White, Becker argues that the most important factor affecting the success of revolutions is the extent to which their proponents are able to take over the art world's organizational structure and facilities. He says (p. 310): 'Ideas and visions are important, but their success and permanence rest on organization, not on their intrinsic worth.'

Crane (1987, pp. 137–43) traces changes in the roles of artistic avant-gardes during the twentieth century and particularly in the decades after World War II. She identifies three artistic roles that existed in the prewar period, that of the aesthetic innovator who analyzed visual reality in terms of its constituents such as color and form, the iconoclast, who attacked bourgeois conventions and art institutions, and the social rebel whose target was political and economic institutions. In the postwar period, these three roles were gradually transformed. The role of the aesthetic innovator emerged as the dominant role but in response to the expansion of public, private, and corporate expenditures and the creation of new cultural institutions, these artists increasingly identified with the middle class in terms of lifestyle and the content of their works. Iconoclasts thrived during the 1960s but after that decade concentrated upon exploring the limits of art itself or eliminating the boundaries between the arts and the media. Artists who moved in the latter direction came to see themselves as entertainers rather than members of an alienated avant-garde. There was little support for the role of social rebel. Political art and a humanistically oriented representational art remained on the periphery of the art world. The role of social rebel was replaced by the democratic artist who communicated social and aesthetic ideas to audiences in poor urban neighborhoods and who had virtually no contact with the art world. (For an examination of art avant-gardes in Israel see Greenfeld 1988.)

At the end of the twentieth century, in the absence of strong avant-garde movements within the art world, alternatives to the aesthetic and social values of the art establishment often came from outside the system, specifically from the works of untrained and often uneducated artists, including folk and ethnic artists, the homeless, prison inmates, elderly people in nursing homes, and hospice patients (Zolberg and Cherbo (1997, p. 1). The so-called 'outsider' artists who

worked alone and who had formerly been completely ignored by the art establishment were now supported by a network of dealers, curators, and collectors.

Drawing on poststructuralist theories, sociologists view the meaning of an artwork as being constructed in the process of reception as much as in the process of creation. The response of the audience constitutes part of the meaning of the artwork. Consequently, the meaning of an avant-garde movement changes in different social contexts as different social groups with particular institutional interests engage in 'a struggle over the text.' Using this perspective, Halley (1985) examined the reception of the Dada movement and showed that, in its own time, its importance was not recognized. Audiences were very small and the movement was overshadowed by Surrealism. Subsequently, different social groups such as art historians, curators, critics, and artists themselves developed competing and conflicting interpretations of Dada and often misunderstood or ignored the original intentions of Dadaist works. From the beginning, there was more interest in Dada in the United States than in Europe, which helps to explain its influence on American avant-garde groups in music, dance, and art in the postwar period. These groups incorporated Dadaist ideas into their own artworks, contributing to wider recognition of the movement.

Bourdieu's (1984) theory argues that support for art movements is associated with social class but that a specific aesthetic statement will appeal to different social strata, depending on its significance in the definition of taste in different time periods. The social positions of supporters of an emerging avant-garde movement differ from the social positions of those that admire the movement after it has received recognition. For Bourdieu, the important aspect of the avant-garde is not its aesthetic characteristics but its structural position in the art world which is subject to change over time.

Attempts to show the relationship between social class and support for artworks that have been considered avant-garde in the past, such as abstract art, have yielded ambiguous results (Halle 1992). However, appreciation of artworks that are still considered avant-garde is confined to relatively small groups even within the middle and upper classes, as is suggested by the intensity of the conflicts that arise when such works are displayed in public places (Heinich 1997).

4. *Avant-gardes and Popular Culture*

In the past two decades, the electronic media have appropriated most of the visual strategies and tactics of artistic avant-gardes from the prewar period. Caldwell (1995, p. 8) claims that 'every framework of the avant-garde…had become highly visible in some form in the corporate world of the new television.' Kaplan (1987, p. 55), in an attempt to classify music videos in the mid-1980s, found that all the categories she identified used avant-garde strategies. The originators of punk music used many of the tactics identified with early-twentieth-century avant-gardes, such as the blurring of boundaries between art and everyday life, intentional provocation of the audience, and disorganization of accepted styles and procedures of performance (Henry 1984).

However, assimilation by the mass media of oppositional stylistic devices associated with avant-gardes raises the question of the extent to which oppositional themes are being disseminated. High-culture institutions frame oppositional messages in such a way as to highlight their effect: an entire evening at the theater devoted to a particular playwright, an entire gallery or museum wing devoted to the work of a particular artist for several weeks. By contrast, the sheer volume of messages of all kinds being transmitted by the mass media tends to obliterate the effect of oppositional messages when they do appear. Kaplan (1987, p. 65) points out that oppositional messages are likely to be 'overridden by the plethora of surrounding texts.'

The commitment of popular culture creators to the tenets of the avant-garde may be questionable. Fashion designers who create luxury clothing for wealthy clienteles often incorporate avant-garde devices in their styles as part of a marketing strategy to attract attention, either to the clothes or to products licensed under their names (Crane 2000). Benetton, an Italian clothing manufacturer, has commissioned advertisements in which photographs, often reprinted from the press, graphically portray social and political crises and calamities in order to increase public awareness of social, political, and environmental issues (Guerrin 1998). However, critics have not accepted this advertising strategy at face value but have claimed that the company is exploiting tragic situations as a means of publicizing its products.

Curiously, as some of the properties of high culture are being assimilated or co-opted by certain forms of popular culture, the old hierarchy of high and low is reappearing in the latter. For example, in popular music, certain genres are deemed superior to others and some artists' works are considered 'classics' in comparison with all the rest (Regev 1994).

5. *The Contemporary Relevance of Avant-gardes*

An analysis of the evolution of avant-gardes in art history, as well as a review of theories of the avant-garde and of relevant sociological research, suggests that the concept of the avant-garde applies to the behavior and activities of artists during a period of about 150 years. At the present time the term is being less frequently used because it is more difficult to identify avant-gardes in an era when diverse and fragmented postmodernist styles predominate. Ironi-

cally, the term now tends to be applied to works created for popular consumption that incorporate techniques but not ideological content from avant-garde movements that existed earlier in the century. Meanwhile, the avant-garde artist's mission as iconoclast and social rebel has been assumed by a broad range of 'micropolitical' movements outside the arts.

See also: Art and Culture, Economics of; Art: Anthropological Aspects; Art, Sociology of; Censorship and Transgressive Art; Cultural Policy: Outsider Art; Popular Culture

Bibliography

Agee W 1968 New York Dada 1910–1930. In: Hess T B, Ashbery J (eds.) *Avant-Garde Art*. Collier Books, Collier–Macmillan, London

Bauman Z 1997 *Postmodernity and its Discontents*. New York University Press, New York

Becker H S 1982 *Art Worlds*. University of California Press, Berkeley, CA

Benjamin W 1968 *Illuminations*. Schocken Books, New York

Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA

Bürger P 1984 *Theory of the Avant-Garde* (trans. Shaw M). University of Minnesota Press, Minneapolis, MN

Caldwell J C 1995 *Televisuality: Style, Crisis and Authority in American Television*. Rutgers University Press, New Brunswick, NJ

Crane D 1987 *The Transformation of the Avant-Garde: The New York Art World, 1940–1985*. University of Chicago Press, Chicago

Crane D 2000 *Fashion and its Social Agendas: Class, Gender, and Identity in Clothing*. University of Chicago Press, Chicago

Dunn R 1991 Postmodernism: populism, mass culture, and avant-garde. *Theory, Culture and Society* **8**: 111–35

Goldman R 1992 *Reading Ads Socially*. Routledge, New York

Greenfeld L 1988 Professional ideologies and patterns of 'gatekeeping': evaluation and judgment within two art worlds. *Social Forces* **66**: 903–25

Guerrin M 1998 Oliviero Toscani, l'âme damnée de Benetton. *Le Monde* **July 5–6**: 8

Halle D 1992 The audience for abstract art: class, culture and power. In: Lamont M, Fournier M (eds.) *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago

Halley J A 1985 The sociology of reception: the alienation and recovery of Dada. In: Herek L, Rupel D (eds.) *Alienation and Participation in Culture*. Research Institute of the Faculty of Sociology, Political Science and Journalism. University of Edvard Kardelj, Ljubljana

Heinich N 1997 Outsider art and insider artists: gauging public reactions to contemporary public art. In: Zolberg V L, Cherbo J M (eds.) *Outsider Art: Contesting Boundaries in Contemporary Culture*. Cambridge University Press, Cambridge, UK

Henry T 1984 Punk and avant-garde art. *Journal of Popular Culture* **17**: 30–6

Huston L 1992 The theory of the avant-garde: an historical critique. *Canadian Review of Sociology & Anthropology* **29**: 72–86

Kaplan E A 1987 *Rocking Around the Clock*. Methuen, New York

Nochlin L 1968 The invention of the avant-garde: France. In: Hess T B, Ashbery J (eds.) *Avant-Garde Art*. Collier Books, Collier-Macmillan Ltd., London

Orton F, Pollock G 1981 Avant-gardes and partisans reviewed. *Art History* **4**: 305–27

Poggioli R 1968 *The Theory of the Avant-Garde* (trans. Gerald Fitzgerald). Harvard University Press, Cambridge, MA

Regev M 1994 Producing artistic value: the case of rock music. *The Sociological Quarterly* **35**: 85–102

Rosenberg H 1968 Collective, ideological, combative. In: Hess T B, Ashbery J (eds.) *Avant-Garde Art*. Collier Books, Collier-Macmillan, London

Taylor J 1968 Futurism: the avant-garde as a way of life. In: Hess T B, Ashbery J (eds.) *Avant-Garde Art*. Collier Books, Collier-Macmillan, London

White H, White C 1965 *Canvases and Careers*. Wiley, New York

Zolberg V L, Cherbo J M (eds.) 1997 *Outsider Art: Contesting Boundaries in Contemporary Culture*. Cambridge University Press, Cambridge, UK

D. Crane

Aviation Safety, Psychology of

Aviation psychology arose as a unique discipline within the field of psychology as the result of technological developments during World War II. Prior to this point in history, most aviation accidents occurred as a result of some structural failure of the aircraft or the failure of the engine to continue to produce power (Koonce 1999). Because of the military demands of World War II, however, the sophistication and reliability of aircraft were improved considerably. Nonetheless, these technological improvements, like many more recent advancements in aircraft capabilities, increasingly challenged the abilities of pilots. Consequently, aircrew error began to play a progressively larger role in aviation accidents, as aircraft became more sophisticated and reliable (Shappell and Wiegmann 1996). Consequently, the need to address the psychological or 'human' side of aviation safety sparked the emergence of aviation psychology.

At the beginning of the twenty-first century, aviation psychology continues to be a primary discipline responsible for addressing and preventing human error in aviation. Specifically, aviation psychologists attempt to understand the fundamental nature and underlying causes of aircrew error and unsafe acts that significantly impact flight safety. However, just as many professionals in other areas of psychology have different opinions or theories about human behavior, aviation psychologists also have varied perspectives

on pilots' performance in the cockpit (Hollnagel 1998). There are primarily four perspectives: (a) cognitive, (b) ergonomics, (c) psychosocial, and (d) organizational. In turn, these different perspectives have historically led to different approaches for addressing pilot error in the cockpit. The purpose of this article therefore, is to provide an overview of these error perspectives and their associated approaches to improving aviation safety. A brief discussion of more recent applications of aviation psychology to safety issues outside the cockpit will then be provided.

1. Cognitive Perspective

The principal feature of the cognitive perspective is the assumption that the pilot's mind can be conceptualized as an information-processing system, much like a modern computer. As such, the cognitive perspective assumes that once information from the environment makes contact with one of the senses (e.g., eyes, ears, nose, etc.), it progresses through a series of stages or mental operations to produce a response or action (Wickens and Flach 1988). These intervening mental operators include such things as information recognition, problem diagnosis, goal setting, and strategy selection (O'Hare et al. 1994). Other factors such as attention, memory capacity, and prior knowledge or experience with similar environmental conditions also affect pilots' interpretation of information and their reactions to it. Consequently, pilot errors are believed to occur when one or more of these mental operations fail to process information appropriately (Wiegmann and Shappell 1997).

According to the cognitive perspective, reducing aircrew errors and improving aviation safety necessarily requires an enhancement of pilots' information-processing abilities. However, unlike computers that can be improved by upgrading the hardware, the information-processing hardware of the human (i.e., the brain) is generally fixed inside the head! Therefore, in order to improve performance, cognitive psychologists have attempted to improve the manner in which pilots process information. One way this is accomplished is through improved training methods. Better training methods are often developed by examining the techniques used by expert pilots to solve problems and allocate their attention to the different sources of information in the cockpit. This information is then used to train novice pilots to improve their techniques and flight performance. Another way of improving information processing is through the standardization of procedures and the use of checklists. These methods facilitate information processing and performance by reducing mental workload and task demands on pilots' memories during normal operations and emergencies, thereby reducing the potential for errors and accidents.

2. Ergonomics and Systems Perspective

According to the ergonomics and systems perspective, pilot performance (both good and bad) is the result of a complex interaction among several factors (Edwards 1988). After all, flying an airplane is a very complicated and dynamic task. Indeed, military and commercial pilots interact with high-tech airplanes and cockpit equipment and must safely operate their aircraft in all types of weather conditions. According to the systems perspective, therefore, a pilot should not be viewed as a sole source of an error or cause of an accident. Rather, pilot error is believed to occur when there is a mismatch or breakdown in the interface between the aircrew and the technology. Consequently, aircrew errors are often referred to as 'design-induced' because they are viewed as resulting from a failure to design the interface in a way that optimizes the pilot–airplane interaction.

The approach to reducing errors and improving safety taken by most ergonomists and systems theorists is to improve the design of the interface between the pilot and the airplane. Such interface issues include the design of equipment used to manually control the airplane, such as the yoke and rudders. However, they more often include the design of better flight instruments that display the status of the aircraft, such as the altimeter and airspeed indicator. As technology has increased and airplanes have become more computerized, the tasks performed by the pilot and airplane have also been redesigned through the use of automation. A ready example is the development of the autopilot and other flight management systems (FMS) that can navigate and fly the airplane without pilot input. By sharing the responsibility with automation, the opportunity for pilots to commit errors is presumably reduced. However, even when errors do occur, they can often be 'caught' by the airplane's computer. Therefore, the system as a whole has become more 'error tolerant' and the negative consequences of pilot errors (i.e., accidents) are reduced.

3. Psychosocial Perspective

According to the psychosocial perspective, flight operations are best viewed as a social endeavor that requires aircrew to interact with one another, as well as with a variety of other flight support personnel, such as air traffic controllers, dispatchers, ground crew, maintenance personnel, and flight attendants. Aircrew performance, therefore, is directly influenced by the nature or quality of these interactions (Helmreich and Foushee 1993). These social interactions, however, are often influenced by both the personalities and attitudes of the individuals within each group. The major theme of the psychosocial perspective, therefore, is that errors and accidents occur when personality difference and conflicting

attitudes disrupt group dynamics and interpersonal communications.

The psychosocial approach to reducing errors in the cockpit has focused on improving the social interactions among aircrew. One method used to achieve this goal is through systematic crew-pairing procedures. These procedures attempt to match aircrew based on their level of experience, flight skills, and personalities. Another method that has been developed is crew resource management (CRM) training. This training attempts to challenge and change pilots' traditional attitudes about differences in authority between the captain and the other aircrew (e.g. the co-pilot or first officer) that have been shown to hinder communication and cause accidents (Wiegmann and Shappell 1999). Other aspects of CRM training involve educating and training aircrew to use techniques for more effectively communicating problems, dividing task responsibilities during high workload situations, and resolving conflicts in the cockpit. Such improvements in aircrew coordination and communication ultimately result in fewer errors and improved aviation safety.

4. *Organizational Perspective*

According to the organizational perspective, pilot performance must be viewed in terms of the organizational context in which it takes place (Heinrich et al. 1980). Indeed, all professional pilots in both the military and commercial aviation industries operate within an agency or company that regulates their time and performance in the cockpit. Aviation organizations are therefore responsible for ensuring that only those pilots with the 'right stuff' are hired to fly their aircraft. In addition, these organizations are also responsible for instituting appropriate procedures that ensure safe operations of the aircraft (Shappell and Wiegmann 2000). From the organizational perspective, therefore, aircrew errors and subsequent accidents are believed to occur when managers and supervisors fail to set up basic conditions within the organization that promote flight safety (Reason 1990).

Given it is the organization's responsibility to ensure that only skilled and safe pilots get into the cockpit, a primary method used by organizational psychologists to prevent aircrew errors is the use of pilot selection tests. For those organizations that train their own pilots, as do many militaries around the world, these selection tests attempt to 'weed out' those applicants who exhibit less than adequate mental aptitudes or psychomotor skills necessary for learning how to fly. Other commercial organizations that hire trained pilots often use background and flight experience as employment criteria, while others also use medical screenings and interviews to select their pilots. In addition to selection techniques, however, another organizational approach to reducing errors in the cockpit is through the establishment of policies or rules that regulate what pilots can and cannot do in the cockpit. Such rules may restrict the type of weather in which pilots may operate their aircraft, or may limit the number of hours pilots can spend in the cockpit, in order to avoid the possible detrimental effects of fatigue on performance. By placing only safe and proficient pilots in the cockpit and limiting aircraft operations to only safe flying conditions, organizations are able to reduce the likelihood that pilots will make mistakes and cause accidents.

5. *Future Directions: Aviation Safety Outside the Cockpit*

Historically, the majority of safety efforts by aviation psychologists have focused on the performance of aircrew. However, there is a growing concern within the aviation community over safety issues that arise outside the cockpit, and an increasing number of aviation psychologists are being called upon to address some of these issues. Two such areas of growing concern are air traffic control and aircraft maintenance.

During the early years of aviation, aircrew avoided becoming lost by using simple cockpit instruments and visual landmarks on the ground. However, both military and commercial demands gradually required pilots to fly in poor visibility conditions and at night. The job of air traffic control was subsequently established to help maintain safe separation between aircraft and to ensure that pilots would not fly their planes into the ground or other obstacles (Hopkin 1995). Still, as the number of aircraft and demands on air traffic control services has increased over the decades, so has the number of accidents, incidents, and runway incursions (loss of safe separation among aircraft and other ground vehicles). As with most aviation accidents today, most of these occurrences have not been due to faulty control equipment, but rather to human error, including mistakes made by air traffic controllers.

Another important factor affecting aviation safety is aircraft maintenance. Indeed, despite all the technological advances and improvements in the reliability of aircraft equipment and systems, modern aircraft still need maintenance. This maintenance often requires that the aviation maintenance technician repeatedly disassemble, inspect, and replace millions of removable parts over the long working life of the aircraft (Reason 1997). During the early years of aviation, aircraft equipment and engines were 'simple,' compared to their modern-day counterparts. As such, these maintenance inspections were relatively easy and resulted in frequent detections and replacement of failed components that often caused accidents. Today,

however, aircraft components and systems are very complex and hardly ever fail. Still, the intricate nature of inspecting and maintaining modern aircraft often lead to errors by the mechanics doing the work. As a result, the contribution of maintenance and inspection errors to aviation accidents and fatalities is on the rise.

Given these growing concerns for safety issues outside the cockpit, aviation psychologists have begun to examine the 'human' side of both air traffic control and aircraft maintenance. However, compared to the long history of efforts by aviation psychologists to address aircrew error in the cockpit, efforts to address human error in the air traffic control and maintenance arenas have only recently begun. Nevertheless, just like aircrew performance in the cockpit, the methods used by aviation psychologists to address errors made by controllers and maintenance personnel depend heavily upon the perspective they take concerning the underlying nature and causes of these errors. These perspectives are generally very similar to those taken by aviation psychologists when addressing aircrew error in the cockpit. Consequently, these perspectives will undoubtedly produce a variety of approaches for addressing human error in air traffic control and aircraft maintenance and will each make an important contribution to future improvements in aviation safety.

6. *Conclusion*

Aircrew error has played a progressively more important causal role in aviation accidents as aircraft have become more reliable. The field of aviation psychology, therefore, studies aircrew error and other pilot actions that can significantly jeopardize the safety of flight. More recently, aviation psychologists have also begun to address errors in air traffic control and aircraft maintenance, which are quickly becoming additional safety concerns within the aviation industry. Aviation psychologists, however, often have different viewpoints or perspectives when it comes to explaining the causes of human errors, both inside and outside the cockpit. These viewpoints are not necessarily incompatible. Rather, they focus on the different cognitive, engineering, social, and organizational factors that frequently contribute to a breakdown in human performance. As such, each perspective has uniquely contributed to the development of intervention techniques and methods for reducing errors and increasing aviation safety. More information regarding these approaches outside the aviation domain can be found in this volume, including discussions of the ergonomics and systems approach (e.g., *Engineering Psychology*), group dynamics and team work (e.g., *Group Processes in Organizations*; *Teamwork and Team Training*), and personnel selection (e.g., *Personnel Selection, Psychology of*).

Finally, there are other fields outside of aviation psychology that have also contributed specifically to improvements in aircrew performance but are not discussed here. One of the most significant contributors is the field of aviation medicine that has revealed several aeromedical factors that impact aircrew performance during flight (Reinhart 1996). For example, the field of aviation medicine has led to a better understanding of the physiological and behavioral factors that lead to stress and fatigue in the cockpit. This knowledge has led to more 'biocompatible' flight schedules for aircrew, particularly when traveling overseas or across several time zones. More information about the general effects of stress on human performance can be found in *Stress in Organizations, Psychology of*.

In conclusion, the field of psychology has a long history of involvement in the process of improving aviation safety. These efforts, combined with the efforts for those in other behavioral, social, and biological sciences, have contributed significantly to the reduction of human error and aviation accidents, making aviation one of the safest modes of transportation.

See also: Ergonomics, Cognitive Psychology of

Bibliography

Edwards E 1988 Introductory overview. In: Wiener E, Nagel D (eds.) *Human Factors in Aviation*. Academic Press, San Diego, CA, pp. 3–25

Heinrich H W, Petersen D, Roos N 1980 *Industrial Accident Prevention: A Safety Management Approach*, 5th edn. McGraw-Hill, New York

Helmreich R L, Foushee H C 1993 Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In: Wiener E L, Kanki B G, Helmreich R L (eds.) *Cockpit Resource Management*. Academic Press, San Diego, CA, pp. 3–45

Hollnagel E 1998 *Cognitive Reliability and Error Analysis Method* (*CREAM*). Alden Group, Oxford, UK

Hopkin D 1995 *Human Factors in Air Traffic Control*. Taylor and Francis, Bristol, PA

Koonce J M 1999 A historical overview of human factors in aviation. In: Garland D J, Wise J A, Hopkin V D (eds.) *Handbook of Aviation Human Factors*. Erlbaum, Mahwah, NJ, pp. 3–13

O'Hare D, Wiggins M, Batt R, Morrison D 1994 Cognitive failure analysis for aircraft accident investigation. *Ergonomics* **37**: 1855–69

Reason J 1990 *Human Error*. Cambridge University Press, New York

Reason J 1997 *Managing the Risks of Organizational Accidents*. Ashgate, Brookfield, VT

Reinhart R O 1996 *Basic Flight Physiology*, 2nd edn. McGraw-Hill, New York

Shappell S, Wiegmann D 1996 U.S. naval aviation mishaps 1977–92: Differences between single- and dual-piloted aircraft. *Aviation, Space, and Environmental Medicine* **67**: 65–9

Shappell S, Wiegmann D 2000 *The Human Factors Analysis and Classification System* (*HFACS*) (Report Number DOT/FAA/AM-00/7). Federal Aviation Administration, Washington, DC

Wickens C, Flach J 1988 Information processing. In: Wiener E, Nagel D (eds.) *Human Factors in Aviation*. Academic Press, San Diego, CA, pp. 111–15

Wiegmann D A, Shappell S A 1997 Human factors analysis of post-accident data: Applying theoretical taxonomies of human error. *The International Journal of Aviation Psychology* **7**: 67–81

Wiegmann D A, Shappell S A 1999 Human error and crew resource management failures in Naval aviation mishaps: A review of U.S. Naval Safety Center data, 1990–96. *Aviation, Space, and Environmental Medicine* **70**: 1147–51

D. A. Wiegmann

Avoidance Learning and Escape Learning

Avoidance learning in animals is studied using an instrumental (operant) training paradigm that was created and first reported by L. H. Warner (1932a) in a study of the 'association span of the white rat.' In this procedure, a warning signal (WS) predicts the subsequent occurrence of an aversive event, typically a mild electric shock delivered to the feet of the animal. A response defined by the experimenter, e.g., pressing a lever in a 'Skinner box' or running and/or jumping over a hurdle to the opposite side of a two-compartment chamber (a 'shuttle box'), terminates the WS and prevents the occurrence of the shock. This is an avoidance response. Failure to avoid in the presence of the WS results in the predicted shock, which, in most procedures, can be terminated by the 'same' response, which is then classified as an escape response. Warner's procedure had its roots in one used by Yarbrough (1921), in which a mild shock was used as a cue for turning right or left for food reward in a maze; he found that the rats would respond similarly to a light signal that preceded the shock cue. Warner used what is now known as a 'trace-conditioning' procedure in which the warning signal was presented for 1 s, and the shock occurred 1, 10, 20, or 30 s after the onset of the WS in separate groups of animals (the trace intervals were 0, 9, 19, and 29 s, respectively). Difficulty of learning increased as a function of the WS-shock interval, which is, of course, confounded with the duration of the trace interval. Essentially no learning occurred in the 30-s group.

1. *Procedural Variations*

There are two forms of avoidance training: active and passive. In the active form, pioneered by Warner's early work noted above, the avoidance contingency requires the occurrence of a specific response, whereas in the passive form the avoidance contingency requires the nonoccurrence, i.e., the suppression, of some specific response. This is often called punishment. The punished response may occur because it is 'spontaneous,' i.e., innate, or because of prior reward or avoidance training. One form of the passive procedure that depends on innate behavior is the so-called one-trial passive avoidance procedure. A rat or mouse, both of which are nocturnal species, will readily leave a brightly lit, elevated platform to enter a dark compartment. If this photophobic response is then punished with a brief shock in the dark compartment, latency to re-enter the dark compartment is increased on subsequent tests, which is taken as a measure of the strength of the memory of the previous experience. This procedure has been used extensively to study the neuropsychological and neuropharmacological bases of memory, because the learning is exceptionally fast (one trial often suffices) and the learning event that establishes the memory is relatively fixed in time. The procedure has been especially useful in the study of retrograde amnesic events, such as electro-convulsive shock, stress, and hypothermia (see Duncan 1949, for an early example of this approach). If the punished response is learned and based on reward, the passive avoidance contingency usually results in suppression of the response, whereas if it is based on prior avoidance or escape training, the response is often facilitated or enhanced, at least temporarily, which can be viewed as a somewhat paradoxical effect of punishment. Note that in the active form, what the animal does to avoid the aversive event is well-defined and measured, whereas in the passive form, the avoidance response is not defined and rarely measured; only the punished response is defined and measured.

There are two kinds of avoidance training procedures: discrete-trial and free-operant. The experiment by Warner is an active form of the discrete-trial kind, but the passive form may also be of this kind. In such a case, a response in the presence of a discriminative stimulus or cue is punished. For example, an animal such as a rat may be trained to press a lever for food reward, and each response may be rewarded. However, in the presence of a specific cue, the responses may also be punished, either at every response or only some responses (partial punishment). In the active form of the free-operant kind of procedure, which originated with Sidman's (1953) initial publication of this method, there is typically no WS to predict the impending shock, only the passage of time since the last response or shock. Operationally, two timers control events: a response-shock (R-S) timer set, for example, at 30 s and a shock-shock (S-S) timer set, for example, at 5 s. Training begins with the S-S timer in control, and brief, e.g., 0.5 s, shocks occur every 5 s. A response interrupts the S-S timer and starts the R-S timer. If the R-S interval elapses, a shock is presented and the S-S timer takes over until another

response is made. Typically, rats pressing a lever under these conditions make a burst of responses immediately after a shock, and only with extensive training do they distribute some of their avoidance responses into the R-S interval. When a running response, as in the shuttle box, is used, long inter-response times well into the R-S interval are frequent, and the avoidance response is quite efficient. A variant of this procedure does not use the S-S timer, but terminates all shocks contingent on the response, so that the animal determines their duration. Another variant introduces a cue part way through the R-S interval, in which case responses tend to occur shortly before or after the onset of the WS.

2. *Theories of Avoidance Learning*

Early theorists had trouble accounting for avoidance learning because it appeared not to be a simple product of Pavlovian conditioning. One interpretation of Pavlovian conditioning, especially as it was applied to escape/avoidance learning, is that of stimulus substitution. The idea here is that the warning signal is a Pavlovian conditioned stimulus (CS) that is paired with shock on escape trials and, by association with the shock, becomes a substitute for that eliciting stimulus. An avoidance response is simply an escape response that has been transferred to the WS. This implies that the avoidance and escape responses should be the same, but in a companion piece to his original paper, Warner (1932b) showed that although the two responses accomplish 'the same end-result' they are 'clearly unlike' each other. He went on to demonstrate this fact in three different test situations which utilized three different forms of motor behavior. In each case the avoidance response appeared to be somewhat different from the escape response. Subsequently, Bolles and Tuttle (1967) showed that the avoidance and escape responses need not even be topographically similar, although there are some apparently innate constraints that make some combinations of escape and avoidance responses easier to learn than others.

Because avoidance learning appeared not to be a simple Pavlovian phenomenon, alternative interpretations were sought. These were based on the early work of behaviorists such as Watson (1913) and Thorndike (1913), who postulated that responses are learned as a result of their consequences. Thorndike's law of effect (Thorndike 1932) posits that responses are learned (reinforced) because of their effects or consequences, now termed 'response-contingent outcomes.' Simply put, responses increase if they are rewarded by desirable outcomes and decrease if punished by aversive consequences. But what is the reward for escape/avoidance learning? Perhaps it is the 'negative reinforcement' accruing from termination of the painful electric shock that clearly supports the escape response. What about the avoidance response, which seems to be acquired because of the nonoccurrence of a future aversive event. Hilgard and Marquis (1940), two early learning theorists, stated:

> Learning in this [avoidance] situation appears to be based in a real sense on the avoidance of the shock. It differs clearly from other types of instrumental training in which the conditioned response is followed by a definite stimulus change—food or the cessation of shock. In instrumental avoidance training the new response is strengthened in the absence of such a stimulus. Absence of stimulation can obviously have an influence on behavior only if there exists some sort of preparation for or expectation of the stimulation.

Several theorists answered the question, in various ways, of what mediates the avoidance response. With theoretically important variations as to detailed mechanisms, all approached the problem by postulating conditioned or learned fear (Mowrer 1950, Miller 1951, Solomon and Wynne 1954, Rescorla and Solomon 1967) or, more operationally, conditioned aversion (Anger 1963), as the mediating process. For Miller, fear was learned, along with the escape response, as a result of negative reinforcement accruing from the instrumental termination of pain from the electric shock. For the others, fear was a Pavlovian conditioned response that came to be elicited by the WS as a result of the temporal contiguity of the WS and shock on escape trials. Thus, two learning processes were assumed to account for the phenomena of escape/avoidance learning: (a) acquisition of fear during escape trials, either by Pavlovian conditioning (S-S contiguity) or operant learning reinforced by shock termination, and (b) acquisition of the instrumental escape response reinforced by shock termination and acquisition of the instrumental avoidance response reinforced by termination of the now aversive warning signal, and reduction of conditioned fear. Kamin (1956) showed that both termination of the WS and avoidance of the shock are sufficient for avoidance learning, although termination of the WS is not necessary. Additional experiments that support this two-process model have employed a transfer procedure in which animals undergo Pavlovian conditioning in one situation, and the CSs of that procedure are superimposed on an operant baseline of responding in another situation. A CS^+ from excitatory aversive conditioning facilitates a Sidman avoidance response during transfer, and, conversely, a conditioned inhibitory CS^- from various forms of discriminative conditioning inhibits the operant baseline of responding (Rescorla and LoLordo 1965).

Countering the two-process model are more recent 'cognitive' approaches, e.g., Bolles (1972) and Seligman and Johnston (1973). For example, Bolles hypothesized that two expectations are learned during avoidance training: (a) some stimuli in the situation are followed by shock and become signals of danger, and (b) a certain response will lead to safety and

becomes a signal of that safety. In this model, reinforcement is irrelevant, as all avoidance responses are innate species-specific defense responses (SSDR) (Bolles 1971), and avoidance responding consists simply of these innate (prewired) responses: withdrawal from cues of danger and approach to cues of safety. Despite the prevalent cognitive 'zeitgeist,' not all learning theorists find such mentalistic models adequate.

3. *Individual Differences*

Bolles' SSDR hypothesis, however, raises the question of what are the biological bases of avoidance learning. The general finding is that at least some animals of most species can learn the avoidance contingency, but the proportion that can learn varies among species and among the various forms and kinds of training procedures. Dogs are particularly good at learning in an active, discrete-trial or free-operant procedure in a shuttle box, and they also typically show strong resistance to extinction (Solomon and Wynne 1954). In contrast, rats are especially difficult to train using an active, lever-pressing discrete-trial procedure, although with the aid of relatively long WS-shock intervals and strong exteroceptive safety signals following each avoidance response, most rats learn rapidly (Berger and Brush 1975). An additional general finding is that with most procedures, individual differences are rather large, with some individuals learning in just a few trials, whereas others never learn to avoid but behave unremarkably in other respects. There are also large differences in the proportion of animals that learn among different strains of rats, among animals of the same strain from different commercial suppliers, and even between successive shipments of the same strain from the same supplier.

4. *Genetic Contributions*

Given this rather great variability, several researchers have bred rats selectively for extreme differences in avoidance learning. Bignami (1965) developed the Roman High- and Low-Avoidance strains in Italy, Brush et al. (1979) developed the Syracuse High- and Low-Avoidance strains in the USA, and Bammer (1978) first reported on the Australian High- and Low-Avoidance strains. In all three experiments, selection criteria were based on extreme differences in avoidance learning using discrete-trial, active, two-way shuttle-box procedures that were remarkably similar but differed in important details. Bidirectional selection was successful in all three experiments, and the strains, particularly the Roman and Syracuse animals have been used extensively in behavioral, anatomical, physiological, and neurochemical studies to identify characters that might be genetically correlated with and possibly mediate genetic expression in the selected phenotypes. Overall, these studies indicate clearly that avoidance learning has a major genetic component and that although the strains do not differ in general learning ability, they do differ in complex ways in the affective domain, i.e., in emotional reactivity or lability, in innate fearfulness or timidity, and in ease of conditioning fear. Not surprisingly, differences in brain anatomy and function, especially in those structures known to be involved in aversive conditioning and learning, and differences in the stress-related endocrine systems of the brain and body have also been found. Current research with these strains is now using population-level and state-of-the-art molecular genetic analyses to identify patterns of inheritance of phenotypic correlations and genetic polymorphisms between animals of these strains.

See also: Fear Conditioning; Fear: Psychological and Neural Aspects; Operant Conditioning and Clinical Psychology

Bibliography

Anger D 1963 The role of temporal discrimination in the reinforcement of Sidmanavoidance behavior. *Journal of the Experimental Analysis of Behavior* **6**: 477–506

Bammer G 1978 Studies on two new strains of rats selectively bred for high or low conditioned avoidance responding. Paper presented at the Annual Meeting of the Australian Society for the Study of Animal Behavior, Brisbane, Australia

Berger D F, Brush F R 1975 Rapid acquisition of discrete-trial lever-press avoidance: Effects of signal–shock interval. *Journal of the Experimental Analysis of Behavior* **24**: 227–39

Bignami G 1965 Selection for high rates and low rates of avoidance conditioning in the rat. *Animal Behavior* **13**: 221–7

Bolles R C 1971 Species-specific defense reactions. In: Brush F R (ed.) *Aversive Conditioning and Learning*. Academic Press, New York

Bolles R C 1972 Reinforcement, expectancy and learning. *Psychological Review* **79**: 394–409

Bolles R C, Tuttle A V 1967 A failure to reinforce instrumental behavior by terminating a stimulus that had been paired with shock. *Psychonomic Science* **9**: 155–6

Brush F R, Froehlich J C, Sakellaris P C 1979 Genetic selection for avoidance behavior in the rat. *Behavior Genetics* **9**: 309–16

Duncan C P 1949 The retroactive effect of electroshock on learning. *Journal of Comparative and Physiological Psychology* **42**: 32–44

Hilgard E R, Marquis D G 1940 *Conditioning and Learning*. Appleton-Century-Crofts, New York, pp. 58–9

Kamin L J 1956 The effects of termination of the CS and avoidance of the US on avoidance learning. *Journal of Comparative and Physiological Psychology* **49**: 420–4

Miller N E 1951 Learnable drives and rewards. In: Stevens S S (ed.) *Handbook of Experimental Psychology*. Wiley, New York

Mowrer O H 1947 On the dual nature of learning—A reinterpretation of 'conditioning' and 'problem solving.' *Harvard Educational Review* **17**: 102–48

Rescorla R A, LoLordo V M 1965 Inhibition of avoidance behavior. *Journal of Comparative and Physiological Psychology* **59**: 406–10

Rescorla R A, Solomon R L 1967 Two process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review* **74**: 151–82
Seligman M E P, Johnston J C 1973 A cognitive theory of avoidance learning. In: McGuigan F J, Lumsden D B (eds.) *Contemporary Approaches to Conditioning and Learning*. Wiley, New York, pp. 69–110
Sidman M 1953 Avoidance conditioning with brief shock and no exteroceptive warning signal. *Science* **118**: 157–8
Solomon R L, Wynne L C 1954 Traumatic avoidance learning: The principles of anxiety conservation and partial irreversibility. *Psychological Review* **61**: 353–85
Thorndike E L 1913 *Educational Psychology: The Psychology of Learning*. Teachers College, New York
Thorndike E L 1932 Reward and punishment in animal learning. *Comparative Psychology Monographs* **8**(39): 1–65
Warner L H 1932a The association span of the white rat. *Journal of Genetic Psychology* **41**: 57–90
Warner L H 1932b An experimental search for the 'conditioned response.' *Journal of Genetic Psychology* **41**: 91–115
Watson J B 1913 Psychology as the behaviorist views it. *Psychological Review* **20**: 158–77
Yarbrough J U 1921 The influence of the time interval upon the rate of learning in the white rat. *Psychological Monographs* **30**(135): 1–52

F. R. Brush

Axiomatic Theories

Axiomatic development of theories is common practice in pure mathematics and is also now widely used in many sciences. The main ingredients of the methods for axiomatizing theories are the following: statement of the primitive concepts of the theory, statement of the prior mathematics basis assumed, statement of the axioms, characterization of models of the theory and a definition of two such models having the same structure. Theories formulated in this way can easily satisfy the standard set-theoretical approach to axiomatization. The further step of formalizing the language of the theory, especially in the case of elementary theories, can lead to specific positive and negative results about the axiomatizability of theories in restricted languages.

1. *Historical Background*

Of all the remarkable intellectual achievements of ancient Greek civilization, none has had greater subsequent impact than the development of the axiomatic method of analysis. No serious traces are to be found in the earlier civilizations of Babylon, China, Egypt, or India. The exact history of the beginnings is not known, but elements that can now be identified emerged in the fifth century BC. A good reference is Knorr (1975). What can be said, and is important for the subsequent discussion here, is that, already in the next century, the fourth century BC, the detailed and elaborate theory of proportion of Eudoxus emerged in quite a clear and definite axiomatic form, most of which is preserved in Book V of Euclid's *Elements* (Euclid 1925). What is important about Eudoxus's work, and even the commentaries of the work of this time, for example, by Aristotle, in the *Posterior Analytics* (1994 74a–17), is that the theorems were proved, not for single geometric objects, but for magnitudes in general, when applicable. The recognition of the correct abstraction in the general concept of magnitude, its technical and thorough implementation by Eudoxus and the philosophical commentary by Aristotle, represent a genuinely new intellectual development. The language of Eudoxus's famous Definition V in Book V of Euclid's *Elements* matches in its abstractness and difficulty the standards of modern axiomatic theories in mathematics and the sciences.

Definition 5 *Magnitudes are said to be in the same ratio, the first to the second and the third to the fourth, when, if any equimultiples whatever be taken of the first and third, and any equimultiples whatever of the second and fourth, the former equimultiples alike exceed, are alike equal to, or alike fall short of, the latter equimultiples respectively taken in corresponding order.*

The codification of the Greek axiomatic approach in Euclid's *Elements* was a great success and remained almost unchallenged until difficulties in the details were found in the eighteenth and nineteenth centuries, as discussed later.

Various scientific examples of axiomatic theories existed already in ancient Greek times and, of course, from the Greek standpoint, they were regarded as essentially homogeneous with the axiomatic theory of geometry, no sharp distinction being made between geometry and mechanics, for instance. A good example is Archimedes' set of partial qualitative axioms for measuring weights on balances. This is undoubtedly the first partial qualitative axiomatization of conjoint measurement, a form of measurement that has received both much axiomatic attention and manifold applications in modern theories of measurement in the social sciences. (For detailed treatment of conjoint measurement, see Krantz et al. 1971. Also see *Measurement Theory: Conjoint*.)

Other early examples of axiomatic theories aimed at empirical matters can be found in the large medieval literature on qualitative axioms for the measurement of weight (Moody and Clagett 1952). Even more impressive examples are to be found in the medieval literature on physics. Some of the most subtle and interesting work is that of Nicole Oresme (1968) in the fourteenth century. (See especially his treatise, *Tractatus de Configurationibus Qualitatum et Motuum*.) What is surprising is Oresme's unexpectedly subtle approach in a geometrical framework to the phenomena of intensive qualities and motions. An

example for Oresme would be that of a surface being more or less white.

Finally, in the geometric axiomatic tradition of analyzing phenomena in natural science, the two great late examples of deeply original work were the seventeenth-century treatises of Huygens' *The Pendulum Clock* (Huygens 1673/1986). And, of course, as the second example, Newton's *Principia* (Newton 1687/1946). Both Huygens and Newton formulated their axioms in the qualitative geometric style of Euclid and other Greek geometers two thousand years earlier.

1.1 Axiomatic Geometry in the Nineteenth Century

Without question, development in axiomatic methods in the nineteenth century was the perfection and formalization of the informal Greek methods that had dominated for so many centuries. The initial driving force behind this effort was certainly the discovery and development of non-Euclidean geometries at the beginning of the century by Bolyai, Lobachevski, and Gauss. An important development later in the century was the discovery of Pasch's (1882) axiom as a necessary addition to the Euclidean formulation. Pasch found a gap in Euclid which required a new axiom, namely, the assertion that if a line intersects one side of a triangle, it must also intersect a second side. This was just the beginning for Pasch. He created the modern conception of formal axiomatic methods, which has been the central aspect of the model for axiomatic work up until present times. The axiomatic example in geometry that had the most widespread influence was Hilbert's *Grundlagen der Geometrie*, first published in 1897 and still being circulated in later editions (Hilbert 1956).

2. Ingredients of Standard Axiomatic Practice

Building, especially, on the work in the later part of the nineteenth century in axiomatizing geometry, early in the twentieth century there was widespread axiomatization in mathematics and, to a lesser extent, in the empirical sciences. The main ingredients of the methods for axiomatizing theories were the following: statement of the primitive concepts of the theory, statement of the prior mathematics basis assumed, statement of the axioms, characterization of models of the theory and a definition for isomorphism of two such models, proof of a representation theorem when possible, and, finally, some analysis of invariance of the models of the theory. Before turning to an explicit discussion of these ingredients, it is worth noting that the emphasis should really be on models of the theory, which is what axiomatizing a theory makes clear. For it is models of the theory, i.e., structures which satisfy the theory, as explained in Sect. 2.3, that exhibit the nature of the theory, whether it be in geometry, economics, psychology or some other science. (For an elementary, but detailed, account of the concepts discussed in this section, see Suppes (1957/1999, Chaps. 8 and 12).)

2.1 Primitive Concepts of a Theory

The first point to recognize in axiomatizing a theory is that some concepts are assumed as primitive. Their properties are to be stated in the axioms, and, therefore, it is important to know how many such concepts there are and what is their general formal character, as relations, functions, etc. In Hilbert's axioms for elementary plane geometry, the five primitive concepts are those of point, line, plane, betweenness, and congruence. In contrast, in psychological theories of measurement, an ordering of the stimuli or other phenomena is almost always assumed—in this case, a weak ordering; that is, a binary relation that is transitive and connected, rather than the geometric ordering of betweenness. In addition to a primitive concept of ordering, there are other relations, for example, a primitive concept of the comparison of stimulus differences or an operation of combination for extensive measurement, as in the case of subjective probability. These measurement examples also apply to a large literature on utility and subjective probability in economics. In psychological theories of learning, even of the simplest nature, we would need such concepts as that of stimulus, response and a conditioning relation between stimulus and response. Theories without much more elaboration would require the important concept of similarity or resemblance, in order to develop a theory of the fundamental psychological phenomena of generalization, discrimination and transfer.

There are also important theories that use much simpler primitive concepts. A good example would be the theory of zero-sum, two-person games in normal form. The primitive concepts are just that of two nonempty sets X and Y that are arbitrary spaces of strategies for the two persons and a numerical function M, defined on the product space $X \times Y$. The intuitive interpretation of M is that it is the payoff or the utility function for the player whose space is X. The negative of that is the payoff for the player whose space is Y. Later a representation theorem is given for finite games of this sort.

2.2 Axioms as Defining Theories

As is widely recognized, axioms are what intuitively characterize a theory, whether it be of geometry, game

theory or a psychological theory of measurement. From a formal standpoint, something more can be said. The essence of the matter is that to axiomatize a theory is to define a certain set-theoretical predicate. This is just as valid in the empirical sciences as in pure mathematics. The axioms, of course, are the most important part of such a definition.

Because the theory of weak orderings is used widely in both economics and psychology, as part of the theory of choice, it will be useful to give a formal definition and, therefore, the axioms for a weak ordering.

Definition 1 *Let A be a nonempty set and R a binary relation on A, i.e., let R be a subset of the product space $A \times A$. A structure (A, R) is a weak ordering if and only if the following two axioms are satisfied, for every a, b and c in A:*

Axiom 1 *R is transitive on A, i.e., if aRy and bRc then aRc.*

Axiom 2 *R is connected on A, i.e., aRb or bRa.*

There are various general methodological and folklore recommendations about the way in which axioms should be written—clarity, lack of ambiguity and simplicity are familiar. More substantive results about the form of axioms are discussed in Sect. 4 on first-order formalization. There are many definite mathematical results about the form of axioms, some of which are also discussed later.

2.3 Independence of Axioms

An early and important substantive recommendation is that the axioms be independent; i.e., none can be derived from the others. The question then arises from a methodological standpoint: How is independence to be established? The answer is in terms of *models* of the axioms. To make the matter still more explicit, we consider *possible realizations* of the theory. These are set-theoretical structures that have the right form. For example, in the case of weak orders, a possible realization is an ordered pair consisting of a nonempty set and a binary relation on that set. Such an arbitrary pair is called a possible realization because as a possible realization it is not a *model* of the theory unless the axioms of the theory are also satisfied. An obvious example of a model of the theory of weak ordering is the pair consisting of the set of positive integers and the relation of weak inequality $\geqslant$.

The independence of a given axiom of a theory is established by finding a possible realization of the theory in which all the axioms, except the particular one in question, are satisfied. The deductive argument to show that this yields a proof of independence is intuitively obvious and will not be made in a more formal manner here. But the essence is that if the concept were not independent, but definable, then it would necessarily hold in the model, just like the remaining axioms, and so a contradiction of its both holding and not holding in the given model would be obtained. So, by *reductio ad absurdum* it must be independent. It will be useful to consider an example or two.

To show that the axiom of connectedness for weak orders is independent of the axiom of transitivity, it is necessary to take a possible realization that is also a transitive relation. So, let $A = \{1, 2, 3\}$ and $R = \{1, 2\}$. In other words, R is just the strict numerical $<$ relation on the subset $\{1, 2\}$ of A. Then it is obvious that this relation is transitive but not connected for the number 3, which is in the set A, and does not stand in the relation R to any other element in A. To show that transitivity is independent of connectedness, that is, Axiom 1 for weak orders is independent of Axiom 2, it is sufficient to take the same set A as before, but now, the relation R is the set of the ordered pairs $\{(1, 2), (2, 3), (3, 1), (1, 1), (2, 2), (3, 3)\}$. Then it is clear that any two elements in the set A are connected by the relation R, but the relation R is not transitive, for, in order to be transitive, it must also have the pairs (1, 3), (2, 1) and (3, 2).

The examples of independence given are trivial, but it is to be emphasized that it can often be a difficult task to establish whether or not an axiom is independent of the remaining axioms; that is, whether or not it can be derived from the remaining axioms. A very familiar mathematical example, with a long history in the twentieth century, is the proof that the axiom of choice is independent of the other standard axioms of set theory.

2.4 Padoa's Principle for Proving Independence of Primitive Concepts

Less familiar than proving the independence of axioms is the method of using models of a theory to prove independence of the primitive concepts of the theory. To prove that a particular primitive concept of a theory, for example, the notion of congruence in Euclidean geometry, is independent of the other primitive concepts, it is sufficient to find two models of the theory, such that the domain of both models is the same, the two models are the same for all the other primitive concepts of the theory, but the two models differ only in their realization of the concept in question. Thus, to prove congruence independence, what is required are two different models of the axiom in question, which are the same for all the other concepts, such as point, line, etc., but which have two distinct notions of congruence.

In the case of weak orders, Padoa's principle can be used in an obvious way to show that the concept of the binary relation is independent of the given set A. It suffices, for example, to use the two different orderings $\leqslant$, $\geqslant$ on the set of positive integers. On the other hand, since R is connected, the set A is definable in terms of the relation R for the special case of weak

orderings by taking the union of the domain and range of the relation. But for different orderings, for example, partial orderings, which are reflexive, antisymmetric and transitive on the set A, it is easy to show that the set A is an independent concept. Just let A in one model be a proper subset in the other, and elements in the relation come only from the first set.

3. Isomorphism of Models of a Theory

The separation of the purely set-theoretical characterization of the possible realizations of a theory from the axioms proper, which characterize models of a theory, is significant in defining certain important concepts. For example, the notion of two models of a theory being isomorphic is often said to be axiom-free, since the definition of isomorphism of the models of a theory really depends on just the set-theoretical characterization of the possible realizations. A satisfactory general definition of isomorphism for the structures that are possible realizations of any sort of theory is difficult, if not impossible, to formulate. The usual practice is to formulate a special definition for the possible realizations of some particular theory. This is what will be done here, as already illustrated in the case of binary relations. A possible realization is a set-theoretical structure that is a nonempty set and a binary relation whose domain and range are included in that set. In this case the definition of isomorphism is as follows.

Definition 2 *A binary relation structure (A, R) is isomorphic to a binary relation structure (A', R') if, and only if, there is a function f such that*

(a) $D(f) = A$ and $R(f) = A'$

(b) f is a one–one function

(c) For a, b in A, aRb if and only if $f(a)R'f(b)$.

The definition of isomorphism for possible realizations of a theory is used to formulate a representation theorem, which has the following meaning. A certain class of structures or models of the axiomatized theory is distinguished for some intuitive or systematic reason and is shown to exemplify within isomorphism every other model of the theory. This means that, given any model of the theory, there exists in this distinguished class an isomorphic model. A good example of this can be found in the axiomatic theory of extensive measurement. Given any empirical structure that represents the data about subjective judgments of probability and satisfies the axioms of the theory, there is a numerical structure satisfying the axioms that is isomorphic to the empirical structure. Note, of course, that there is not any one single such isomorphism. Different individuals can have different empirical structures realizing their subjective probabilities, but there will be for each of them a particular numerical model of the axioms that is isomorphic to each given empirical structure.

Here is a simple and obvious, but useful, example of a representation result for two-person, zero-sum games that are finite. A game (X, Y, M), as introduced earlier, is finite just in case the sets X and Y are finite. So, with the definition of isomorphism obvious from the result now to be stated, any finite game $G = (X, Y, M)$ with $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_n)$ is isomorphic to the game $G' = (I_m, I_n, M')$, and $M'(i, j) = M(x_i, y_j)$, where the notation I_k denotes the set of positive integers $1, \ldots, k$. Therefore, any finite two-person zero-sum game may be represented by a game where X and Y are initial segments of the positive integers, even if vivid substantive descriptions of the individual strategies x_i and y_j have been given.

3.1 Invariance Theorems for Axiomatized Theories

In addition to having representation theorems for the models of a given axiomatized theory, it is also significant and useful to have explicit invariance theorems. The intuitive idea of invariance is most naturally explained either in terms of geometric theories or measurement theories.

Given a class of models, for example, analytical models in geometry and numerical models in measurement, with respect to which any other model of the theory can be represented, the invariance theorem is an expression of how unique is the representation in terms of such an analytic model. For example, in the case of the representation theorem for axiomatic formulations of Euclidean geometry, the invariance theorem states that any analytic model is unique only up to the group of Euclidean motions. This means that for any analytic representation of Euclidean geometry, any reflections, rotations or translations of the analytic model will produce a new analytic model, also isomorphic to the set-theoretical model satisfying the qualitative synthetic axioms of Euclidean geometry. In fact, the most general theorem widens the scope to what is often called the group of generalized Euclidean transformations, namely, those that also permit a change in scale, so that no unit of measurement is fixed.

The situation is very similar in the case of theories of measurement. Given an empirical structure and a representing numerical model isomorphic to that structure, then, in the case of intensive quantities such as cardinal utility, for example, the numerical model is unique only up to the group of affine transformations; that is, transformations that multiply the unit of measurement by a positive real number and add a constant as well. This means that in fundamental measurement of a cardinal utility there is nothing but convention in the choice of a unit of measurement or a zero value.

Classical examples of important invariant theorems in physics are that structures of classical physics

are invariant up to Galilean transformations, and in special relativity, invariant up to Lorenz transformations—in both these cases, further generalizations arise by also permitting changes in the units of measurement.

4. Theories with Standard Formalization

The set-theoretical framework for axiomatizing theories just discussed is used implicitly throughout the mathematical social sciences. Most of the axiomatic work that takes place in the social sciences can be put in a straightforward way within the framework just described. On the other hand, there is a tighter framework for discussing the axiomatization of theories that leads to clarification of problems that arise in the conceptual or qualitative thinking about theories in the social sciences. The purpose of this section is to describe in an informal way this more narrowly defined framework and to give a sense of the kind of results that can be obtained. Most of the results are of a negative sort, important because they show what cannot be done by apparently very natural qualitative formulations of theories of preference, of subjective probability or related kinds of qualitative measurement, or scaling problems, especially in economics and psychology, but also in anthropology, political science, and sociology.

A language with standard formalization is a language that is given a precise formulation within first-order logic. Such a logical framework can be characterized easily in an informal way. This is the logic that assumes:

(a) one kind of variable;

(b) logical constants, mainly the sentential connectives such as *and* ;

(c) a notation for the universal and existential quantifiers; and

(d) the identity symbol $=$.

A language formulated within such a logical framework is often called an elementary language. Ordinarily, three kinds of nonlogical constants occur in axiomatizing a theory in such a language—the relation symbols, also called predicates, the operation symbols and the individual constants.

The grammatical expressions of the language are divided into terms and formulas, and recursive definitions of each are given. The simplest terms are variables or individual constants. New terms are built up by combining simpler terms with operation symbols in a recursive manner. Atomic formulas consist of a single predicate and the appropriate number of terms. Compound formulas are built up from atomic formulas by means of sentential connectives and quantifiers.

Possible realizations of elementary theories, i.e., theories formulated in an elementary language, assume an especially simple form. First, there is a nonempty domain for the structure. Second, corresponding to any relation symbol of the theory, there is a corresponding relation, including sets representing predicates as one-place relations. Corresponding to any operation symbols in the theory are operations defined on the domain of the structure, and, finally, individual objects of the domain correspond to the individual nonlogical constants of the theory. It is worth noting in passing that the definition of isomorphism for such elementary structures is straightforward and simple, which is not always the case for theories formulated in more complicated set-theoretical languages.

4.1 Some Positive Results About Axiomatizability

The first positive result uses two concepts. First, the set of all the finite models of a theory is called a *finitary* class of elementary structures. Second, a theory is recursively axiomatizable when there is an algorithm for deciding whether or not any formula of the language is an axiom of the theory in question.

Theorem 1 *Any finitary class of models of an elementary theory is axiomatizable, but not necessarily recursively axiomatizable.*

The importance of this result is showing that the expressive power of elementary languages is adequate for finitary classes but not necessarily for the stating of a set of recursive axioms.

A more special positive result about finitary classes of models can also be given:

Theorem 2 *Let K be the finitary class of measurement structures with respect to an elementary language L and with respect to a numerical model $\mathscr{R}$ of L such that K includes all finite models of L homomorphically embeddable in $\mathscr{R}$. If the domain, relations, functions, and constants of $\mathscr{R}$ are definable in elementary form in terms of $(Re, \leqslant, +, \cdot, 0, 1)$ then the set of sentences of L that are satisfied in every model of K is recursively axiomatizable.*

A *universal sentence* is one that has only universal quantifiers at the beginning with the scope of a quantifier the remainder of the sentence. In practice, such sentences are written as quantifier-free statements to simplify the notation. The axioms of weak ordering given earlier are of this kind. To be explicit, the conjunction of the two given axioms is a universal sentence, and this single universal sentence is, in this form, the single axiom for a weak ordering.

Theorem 3 *(Vaught's Criterion 1954). Let L be an elementary language without function symbols. A finitary class K of measurement structures (with respect to L) is axiomatizable by a universal sentence iff K is closed under submodels and there is an integer n such that if any finite model M of L has the property that every submodel of M with no more than n elements is in K, then M is in K.*

The intuitive idea of Vaught's criterion for finitary classes of models of a theory is easy to explain. Consider again weak orderings. Because the axioms

involve just three distinct variables, it is sufficient to check triples of objects in the domain of an empirical structure to determine if it is a model of the theory. Generally speaking, the number of distinct variables determines the size of the submodels that must be checked to see if universal axioms are satisfied. To have a universal axiom for a theory, or what is equivalent, a finite set of universal axioms, it is necessary that the number of distinct variables be some definite number, say n. By examining submodels involving no more than n objects, it is then sufficient to determine satisfaction of the axiom or axioms, and this is the intuitive idea of Vaught's criterion.

4.2 Some Negative Results About Axiomatizability

To begin with, it is useful to state as a theorem what is a corollary of Vaught's criterion. It is simply the negation of it for determining when a finitary class of models of a theory is not axiomatizable in an elementary language by a universal sentence.

Theorem 4 *Let L be an elementary language without function symbols and let K be a finitary class of measurement structures (with respect to L) closed under submodels. Then K is not axiomatizable by a universal sentence of L iff for every integer n there is a finite model M of L that has the property that every submodel with no more than n elements is in K, but M is not in K.*

It is natural to interpret the negative result stated in the general form here as showing that the complexity of relationships in finitary classes satisfying the hypotheses of the theorem is unbounded, at least unbounded when the theory must be expressed by quantifier-free formulas of elementary logic.

The first application is something rather surprising. A semi-order is a structure $(A, \succ)$ satisfying the following three axioms.

Axiom 1 It is not the case that $(a \succ a)$.

Axiom 2 If $(a \succ b \ \& \ a' \succ b')$ then $(a \succ b' \vee a' \succ b)$.

Axiom 3 If $(a \succ b \ \& \ b \succ c)$ then $(a \succ d \vee d \succ c)$.

When the set A is finite, the following numerical representation holds:

$$\text{for every } a \text{ and } b \text{ in } A,\ a \succ b$$
$$\text{if and only if } f(a) > f(b) + 1.$$

What is now surprising is a result about the indistinguishability relation for semiorders, that is, the relation $\sim$ that is the negation of the characterization of the semiorder, namely, for a and b in A

$$|f(a) - f(b)| \leqslant 1 \text{ iff } a \sim b.$$

The following theorem is due to Roberts (1969).

Theorem 5 *Let L be the elementary language whose only nonlogical symbol is the binary relational symbol $\sim$. Then the finitary class J of measurement structures for the indistinguishability relation $\sim$ is not axiomatizable in L by a universal sentence.*

The next case of a negative result applying the negation of Vaught's criterion (Theorem 3) is the proof that the qualitative theory of utility differences or the qualitative theory, more generally, of various psychometric sensations is not axiomatizable by a universal sentence, contrary, of course, to the simple theory of order. This result is due to Scott and Suppes (1958).

Consider the elementary language whose only symbol is the quaternary relation symbol 'D' with the intended numerical interpretation

$$abDcd \text{ iff } f(a) - f(b) \geqslant f(c) - f(d).$$

We then define the finitary class of measurement structures for algebraic difference as consisting of all models (A, D) such that

(a) A is a nonempty finite set;

(b) D is a quaternary relation on A; and

(c) (A, D) is isomorphic to (A', Δ), where A' is a finite set of numbers and Δ is the quaternary numerical relation such that for real numbers x, y, u and v, $xy\Delta uv$ iff $x - y \geqslant u - v$.

Theorem 6 *Let L be the elementary language whose only nonlogical symbol is the quaternary relation symbol D. The finitary class $\mathscr{D}$ of measurement structures for (algebraic) difference is not axiomatizable in L by a universal sentence.*

The intuitive idea of the proof can be seen from construction of a ten-element structure, all of whose substructures have a numerical representation, but which does not itself have such a representation. The idea of this construction can then be generalized for arbitrary n in order to apply Theorem 4.

Using the same ideas, Titiev (1972) extended the results of Theorem 6 to additive conjoint measurement and to multidimensional scaling with the Euclidean metric. Titiev (1980) also gives a negative proof for the city block metric for $n \leqslant 3$ dimensions. Using more sophisticated logical results, it is possible to extend most of the results just stated to not being finitely axiomatizable. This means that existential quantifiers can be introduced, but the number of axioms must be finite in character. The main results here are due to Per Lindstrom, which together with the other results mentioned in this section are presented in detail in Luce et al. (1990).

Theorem 7 *The finitary class of measurement structures for algebraic difference is not finitely axiomatizable in the elementary language whose only nonlogical symbol is the quaternary relation symbol D.*

In a way, a still stronger negative result about axiomatizability can be proved for Archimedean axioms. Here is one standard formulation of such an axiom. If $a \succcurlyeq b$ then for some n, $nb \succcurlyeq a$, where nb is the combination of n copies of b. Notice that it is necessary

to introduce a quantifier for integers not just for the empirical objects of the domain. Because the exact formulation of the negative result is rather complicated, the following informal theorem is given.

Theorem 8 *For any standard elementary language used to formulate a theory of measurement there is no set of elementary formulas of the language equivalent to an Archimedean axiom for the theory.*

The best way to think about the Archimedean axiom in this context is that it is a second-order axiom and, therefore, cannot be formulated by an equivalent set of formulas, in first-order logic.

Still another way of looking at this result is that to characterize the real numbers we need some sort of second-order axiom such as Dedekind completeness, the Cauchy sequences or the least-upper bound axiom, but none of these axioms, including the Archimedean axiom, can be formulated in a first-order language whose variables take real numbers as values. Narens (1974) provides a general account of Archimedean axioms in various forms.

For deeper and more general results on axiomatizability there is much recent work that can be cited, especially concerning the definability of linear orders for classes of finite models and the problem of the complexity of the class (Stolboushkin 1992, Gurevich and Shelah 1996, Hella et al. 1997).

See also: Mathematical Models in Philosophy of Science; Measurement, Representational Theory of; Measurement Theory: Conjoint; Measurement Theory: History and Philosophy; Ordered Relational Structures

Bibliography

Aristotle 1994 *Posterior Analytics*. Clarendon Press, New York

Euclid 1925 *Euclid's Elements Book V*. [Transl. T L Heath] Cambridge University Press, Cambridge, UK [1956 Dover Publications, New York]

Gurevich Y, Shelah S 1996 On finite rigid structures. *Journal of Symbolic Logic* **61**: 549–62

Hella L, Kolaitis P G, Lousto K 1997 How to define a linear order on finite models. *Annals of Pure and Applied Logic* **87**: 241–67

Hilbert D 1956 *Grundlangen der Geometrie* 8th edn. Teubner, Stuttgart, Germany

Huygens C 1673/1986 *The Pendulum Clock, or Geometrical Demonstration of Pendula as Applied to Clocks*. trans. Rockwell R J. Iowa State University Press, Ames, IA

Knorr W R 1975 *The Evolution of the Euclidean Elements*. Reidel, Dordrecht, The Netherlands

Krantz D H, Luce R D, Suppes P, Tversksy A 1971 *Foundations of Measurement*. Academic Press, New York, Vol. I

Luce R D, Krantz D H, Suppes P, Tversky A 1990 *Foundations of Measurement Vol. III, Representation, Axiomatization and Invariance*. Academic Press, New York, CA

Moody E A, Clagett M 1952 *The Medieval Science of Weights (Scientia de Ponderibus)*. University of Wisconsin Press, Madison, WI

Narens L 1974 Measurement without Archimedean axioms. *Philosophy of Science* **41**: 374–93

Newton I 1687/1946 *Principia*. trans. Cajori F. University of California Press, Berkeley, CA

Oresme N 1968 *Tractatus de Configurationibus Qualitatum et Motuum*. trans. Clagett M. University of Wisconsin Press, Madison, WI

Pasch M 1882 *Vorlesungen über Neuere Geometrie*. Verlang von Julius Springer, Leipzig, Germany

Roberts F S 1969 Indifference graphs. In: Harary F (ed.) *Proof Techniques in Graph Theory*. Academic Press, New York

Suppes P 1957/1999 *Introduction to Logic*. Van Nostrand, New York

Scott D, Suppes P 1958 Foundational aspects of theories of measurement. *Journal of Symbolic Logic* **23**: 113–28

Stolboushkin A P 1992 Axiomatizable classes of finite models and definability of linear order. In: *Proceedings of the 7th Annual IEEE Symposium on Logic in Computer Science*. IEEE Computer Society Press, pp. 64–70

Titiev R J 1972 Measurement structures in classes that are not universally axiomatizable. *Journal of Mathematical Psychology* **9**: 200–5

Titiev R J 1980 Computer-assisted results about universal axiomatizability and the three-dimensional city-clock metric. *Journal of Mathematical Psychology* **22**: 209–17

P. Suppes

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

American Studies: History

1. *Why Study American History?*

As a nation of immigrants with ongoing processes of invention and reinvention of their national identity, America continues to capture the attention of people throughout the world. Since the founding of the new Republic the American people have questioned what meaning America and its history has for them. Professional historians in particular have repeatedly attempted to answer this important question by explaining the stories of American experiences. 'History' for Americans is a key element in the construction of their national consciousness and national identity.

Aiming to discover and define their own self-identity as individual citizens, as well as those of the members of the other diverse racial, cultural, and ethnic groups that make up America, those who study American history constitute a distinct and integral part of an American Studies that is, by definition, interdisciplinary, an Area Studies. American Studies aims to comprehend what makes the American people uniquely American in the unfolding annals of mankind and attempts to answer J. Hector St. John Crevecoeur's famous question, 'What then is the American, this new man?'

American Studies remains fascinating not only for the American people themselves but also for people living abroad as well. Perhaps as often as their American counterparts, American Studies specialists abroad have also often raised the same old but important question 'What is America?' They have sought to understand the totality of and thus appraise the quality of American civilization in the context of the history of mankind, and also, in so doing, to define themselves as individual citizens in their own country.

At the same time, in light of one of the unwritten objectives of American Studies, to contribute to world peace through mutual understanding, the field remains enormously important, particularly when one considers the position that the United States now holds in the present world. In this sense, American Studies is one of the fields that is as rewarding and challenging for specialists as laymen throughout the world.

2. *Why American Studies in Japan?*

Scholarship in American history throughout the world has considerably come to resemble its American counterpart, even if it lags behind scholarship in the United States. Even so, it seems that some approaches and conceptualizations that America specialists abroad invent and utilize in their American Studies must be original and, therefore, different from their American counterparts. If this observation is correct, it follows that Japanese scholarship may be distinct from that in the United States.

Many America specialists in Japan posit that the United States is a nation of ideas and ideals, such as freedom, equality, and democracy, and see the country as a nation that not only discusses such universalistic ideas and ideals but also aspires to achieve those high goals, as Nagayo Homma, the dean of America specialists in Japan, states (Homma 1976, 1999). In fact, the United States has been a unique country, having been engaged in a significant experiment in the history of mankind. In studying American history, Japanese scholars of American Studies endeavor to grasp the essential components that make up American civilization and make efforts to elucidate the uniqueness that is America. They pursue American Studies not simply because the subject matter they have chosen is stimulating but because they believe that United States–Japan relations are vitally important for their country and that by deepening mutual understanding they will contribute to maintaining amicable relations between these two peoples across the Pacific Ocean.

At the same time, American Studies specialists in Japan explore the American past for lessons to present to their audience. As Japan's distinguished America specialist Tadashi Aruga describes in his article in the *Journal of American History*, Japanese intellectuals invoked the example of the United States to encourage the development of a liberal constitutional ideal in Japan (Aruga 1992, Thelen 1992a). America specialists in Japan also use the American past to help them define alternatives for their society and contribute to improving their own society in these ways. As experts on the United States, they are expected to participate actively in discussions on contemporary American affairs by introducing to the public their historical perspectives on America. America specialists in Japan play a key social role in making knowledge and information gleaned from their scholarly work available to public discussions on American affairs and US–Japan relations. In addition, these scholars hope to gain a more accurate sense of the future direction in which humanity in general is moving, as well as to be able to better see where they now stand.

This short article intends to describe the present state of American history as one of the components of American Studies. With particular emphasis on American Studies in Japan, it surveys a 30-year US historiography of American history, and, then, reviews, from a Japanese perspective, some of the important Japanese works that have been written in American Studies since the 1970s. It examines what Japanese scholars have studied in American history and how they have done it, delineating changes in the milieu of their profession, as well as current trends that are seen in Japan and the United States. Perhaps

it is hardly necessary to explain that there is no good reason to justify the dividing of the 30-year period exactly by decades except for the reasons of artificiality and convenience, as history proceeds like a seamless web. In this essay, our survey starts with a review on the New Left historiography in the United States in the 1960s. In the present writer's view, an attempt to discuss New Left historiography is necessary before any appraisal of American history in the United States in the 1970s is made, because both of them are closely interrelated.

3. *American History in the 1960s and 1970s as a Background*

Revisionism is inherent in the historian's craft. American history is by no means an exception. Open to challenge and response, the stories of the American past have been written and rewritten ceaselessly, while at the same time, the field of American history has grown increasingly specialized in recent years. The intensification of specialization and the attendant fragmentation in American history have largely reflected a part of the vast expansion of higher education in post-World War II America. To survey the recent trend of historiography in American history, let us begin with the New Left revisionists among the historians of the 1960s.

Buoyed by the rising tide of social protest movements at home, New Left historians challenged all the assumptions and the entire interpretive framework that the so-called 'consensus' historians of the 1950s had constructed in understanding the American past (Hofstadter 1955, Boorstin 1953, Hartz 1955). Anxious to use the American past to make social changes, New Left scholars blamed the 'consensus' historians for celebrating the status quo of a middle class liberal society. Revisionists criticized the 'consensus' historians severely for placing too much emphasis on historical continuity in American history and overly stressing unity, the stability of institutions, and agreement on central value systems among the American people.

Those young New Left scholars, who gathered around their scholarly journal, *Studies on the Left*, regarded social and economic conflicts as major forces in American history and stressed the importance of economic factors and class analysis in interpreting the American past, as the Progressive historians of the 1930s and 1940s had similarly done previously. The revisionists of the 1960s placed new emphasis on minority and lower socioeconomic groups and saw a pressing need to reconstruct American history as 'seen from the bottom up,' urging that 'the history of the powerless, the inarticulate, the poor' be written (Lemisch 1967).

America was in great turmoil then. Intellectuals seriously questioned the legitimacy of authority, and many of them had lost faith in it. New Left historians, for instance, called upon the American people to undergo a soul-searching re-examination of the record of imperial America's expansionist foreign policy (Williams 1959). On the domestic front as well, they critically analyzed the quality of American liberalism by utilizing the key concept of 'corporate liberalism' as an analytical device. Moreover, they cast serious doubt on the American tradition of reform, while, at the same time, leveling scathing criticism at America's racism, sexism, economic exploitation, and power structure, the most typical example of which was the military–industrial complex (Williams 1961). (For a further discussion of historical writing of the New Left revisionists, see Williams 1989, Buhle 1990, van der Linden 1996; for a critical review on New Left Revisionism, see Tucker 1971.) In the midst of America's turbulence at home and abroad, New Left Revisionism presented not only a penetrating critique of the American present and the past but also an alternative theoretical framework and compelling explanatory power to make better sense of American experiences. Today, however, with the rise of trans-nationalism and globalization, the explanatory power that New Left Revisionism once enjoyed seems to have waned considerably and somehow needs to be reinvigorated. In addition to the rise of trans-nationalism and globalization, the reasons why the influence of New Left revisionism waned seem to be the following: (1) New Left revisionists lost concrete targets with which to criticize America and its imperialistic expansionism, particularly after the Vietnam War which they saw as a glaring example of American imperialism came to an end in 1975; (2) As the declensionist's idea of America losing comparative advantage in several sectors of strategic importance in the world economy came into vogue in the 1970s and 1980s, overly harsh criticism leveled at American business became unpopular, if not entirely unacceptable, among the American people; (3) Coupled with New Left revisionists' tendency to study unilateral national history, their orientation toward analyzing elite leaders' *weltanschaunng* came to be under fire as elitist, while their political dreams to create a 'community of love' in America cooled and came to be totally dismissed as romantic and visionary when America became more conservative in the 1970s; and (4) New Left revisionists lost considerable power and influence in the historical world in America as they fought among themselves and split into factions in the 1970s. For example, the *Studies on the Left* group became the nucleus of a theoretically minded cadre for a New Left, while the *Radical America* group was more or less inclined toward activism.

Historians of the 1970s in America succeeded largely to these legacies that New Left revisionists had bequeathed them and made their own careers by building on some of the works that the historians of

the 1960s had embarked on. Receiving insights from the works of European theorists (these include, for example, Claude Levi-Strauss, Roland Barthes, Michel Foucault, and Jacques Derrida, to name only a few) and scholars, particularly British historians and the French *Annales* school (British scholars are, for example, Edward P. Thompson whose influential work is *The Making of the English Working Class* (New York: Vintage Books, 1963) and Eric J. Hobsbawm whose main work, among others, is *Primitive Rebels* (Manchester, England: University Press, 1959). French historians of the *Annales* school are, for example, Fernand Braudel whose major work, among others, is *Capitalism and Material Life, 1400–1800*, Trans. Miriam Kochan (New York: George Weidenfeld and Nicolson Ltd., 1973) and Emmanuel Le Roy Ladurie whose main work is *Le Territoire de l'historien* (Paris: Edition Gallimard, 1973, 1978). For discussions on the French *Annales* school related to historical writing about the United States, see Richard Mowery Andrews, 'Some Implications of the *Annales* School and Its Methods for a Revision of Historical Writing About the United States,' *Review* 1.3/4 (1978): 165–80), many American historians of the 1970s saw politics as fleetingly transient, turning their back on traditional 'event-oriented' political history and their attention toward slow but important changes in social structure and their relationship to political, economic, and social events in America. (Scholars who placed new emphasis on the importance of studying social structure and institutional bureaucratization in modern America are identified as the historians of the 'organizational school.' They include Samuel P. Hays, Robert H. Wiebe, Louis Galambos, Alfred D. Chandler, Jr., and Robert D. Cuff, to name only a few.)

Historians of the 1970s redefined history to include entire areas of social life not before thought to be part of the subject. An increasing number of scholars, much interested in such issues as race, class, ethnicity, and gender, paid more attention to the common people, as well as to members of racial, ethnic, and cultural minority groups in America who had previously been either neglected or slighted as subjects of serious historical inquiry. Convinced that the 'new' social history approach was highly promising as a mode of systematic inquiry for future scholarship in American history, scholars and social historians placed new emphasis on the role of culture in the formation of class and class-consciousness, analyzing the dominant culture in society as something directly related to social structure. They richly described the private as well as the public spheres of people's lives by marshalling sophisticated analytical tools borrowed from the social and behavioral sciences, such as sociology, demography, econometrics, psychology, and cultural anthropology (Geertz 1973). As a result, historians of the 1970s made significant contributions toward enriching the understanding of American history (although their claim to the rediscovery of the extraordinary diversity of social groups and cultures that compose America was by no means the monopoly of the 'new' social historians (the works of social scientists such as Talcott Parsons and Edward Shils and those of the 'consensus' historians revealed the pluralistic nature of American society, taking cognizance of the vast multiplication of diverse social groups in America)).

The profound influence of the social and behavioral sciences, coupled with the use of new quantitative methods and systematic analysis by employing computer technology (Swierenga 1974, Erickson 1975, Fogel 1975) resulted in the multiplication of subdivisions in the field of American history. American history came alive again and social history in particular became vibrant. This 'new' social history spawned many other subspecialties in the field of American history. It reinvigorated the historical profession by multiplying the variety of new subfields that represented social groups within American society (the histories of Afro-Americans, Native Americans, labor, immigrants, and women, to name only the most outstanding). Few could deny the fact that this spectacular rise of the 'new' social history represented a general trend in the profession of American history in the 1970s and even beyond (for discussions on promises and problems in a 'new' social history approach, see Henretta et al. 1979, Tilly 1984).

No doubt the historians of the 1970s made remarkable achievements in enhancing our understanding of the American past. Yet, however remarkable their positive achievements were, the scholars of the 'new' social history brought with them certain attendant liabilities and serious consequences. Most scholars, with a few exceptions, continued to embrace the nation-state as a unit of analysis. They still had the luxury of thinking exclusively in national terms and indulged in an excessive focus on unilateral national history. But was it not true that a national history required the context of other national histories? In other words, as has been pointed out well, American scholarship was marred by its myopic parochialism with a unilateral national perspective divorced from the rest of the world.

Scholarship in American history suffered from something of an irony as well. It paid the heavy cost of the breakup of history into many disconnected histories. The more subjects were probed and the more monographic studies of women, immigrants, lesbians and gays, workers, children, Native Americans, Asian and Hispanic Americans, and so forth were produced, the more fragmented the images of America became and the more obscure the whole picture of American civilization became. The rise of the 'new' social history proliferated knowledge and

diverse categories, but it only fragmented the field of American history into a multitude of subdivisions that were not necessarily related to each other (Megill 1991).

Take studies in ethnic and labor history, for instance. A sharp dichotomy was observed between the studies that emphasized the importance of cultural analysis, on the one hand, and those that stressed the importance of a structural understanding, on the other (Higham 1982). The same can be said of Afro-American women, too. Given fine and fixed categories, such as race and gender, writing a history of African-American women in a coherent and organized way became increasingly difficult, with, for example, Afro-American women classified by race in African-American history, while in women's history being analyzed by gender (Higginbotham 1992). Despite the fact that such richly described lives of the American people came to be more fully appreciated, little was gained in deepening an understanding of the more complex relations of race, ethnicity, gender, and class within the totality of American society, as long as those studies remained disconnected from each other. Therefore, it was still difficult to find a satisfactory answer to the fundamental question, 'What is America?' While continuing their work, historians came to realize that there existed an urgent need for a new paradigm that would make it possible to bring these variegated pieces together with the goal of encompassing and synthesizing them. (For further discussions on historical writing in the United States in the 1970s, see Kammen 1980, Amerikashi-Kenkyukai 1985.)

4. *Japanese Scholarship in American Studies: The 1970s*

Strange as it may sound, a quick look at the table of contents of professional Japanese journals of American Studies gives one the impression that Japanese scholarship in American history, with a few exceptions, such as Afro-American history and Native American studies, follows its counterpart in the United States after about ten years. Take the 'new' social history, for instance. It was symbolically the year 1979 when *Shiso* (*Thought*), one of Japan's most prestigious journals, published a special issue featuring the promises and problems of social history. Equally curiously, it was the same year when *Tokyo Daigaku Amerika Kenkyu Shiryo Sentah Nenpoh* (The *Bulletin of the Center for American Studies of the University of Tokyo*) featured the studies of American social history for the first time. The former was published in response to the growing interest in social history among Japanese historians in general, while the latter, which introduced to its Japanese readers the current state of the field in the United States, was issued primarily with a view to capturing American Studies specialists' attention to the rise of social history.

Yet, it was not until the 1980s, especially after 1982, the year the late Professor Herbert Gutman, the then leading American scholar of social history, lectured in the Kyoto American Studies Summer Seminar, when Japanese scholarship in social history began to bear full fruit, as Natsuki Aruga, one of Japan's pioneering social history scholars, later recalled (Aruga 1979, 1998). The 1989 publication of *Amerika Shakaishi no Sekai* (*The World of American Social History*) edited by Sozo Honda was a case in point. The book was a brilliant anthology in which a group of 15 scholars including the editor worked together to cover a variety of important problems in American social history ranging from the liberation of Afro-Americans to feminism in America (Honda 1989).

The same can be said of the influence of American New Left historiography and of the so-called Organizational School of American history. From 1970 to the early 1980s, less than ten years after their American counterparts had started, Japanese scholars discussed and hotly debated the extent to which American New Left Revisionism and Organizational Historiography contributed to deepening an understanding of modern America, focusing around the key phrase 'corporate liberalism.' The concept of 'corporate liberalism,' which brings America's political economy at home and its foreign policy together, was believed to be quite helpful in not only exploring the meaning of reform anew in the context of the emergence of corporate America but also in synthesizing modern American history as well (Amerikashi Kenkykai 1980). (Historian Akira Takahashi is one of the first and leading scholars who introduced American New Left historiography to his Japanese peers. See Takahashi 1982; see also Takahashi's brilliant book, especially part III (Takahashi 1999).)

Another feature characterizing Japanese scholarship in the 1970s was that America specialists carried out a considerable number of historical analyses, having been stimulated by dramatic events that had happened in the United States in the previous years. In other words, every time a sensational affair or a festivity to celebrate a historic event was reported in the United States, Japanese scholars made it an occasion to trace its origin in the American past and sought to give new meanings by placing the subject in historical perspectives.

For instance, they re-examined US–China and US–Japan relations immediately after America's dramatic shift in China policy and Sino-American reconciliation in 1971. After going through the tumult of President Nixon's Watergate scandal and his subsequent resignation in 1974, Japanese historians and social scientists probed the quality of American democracy and its institutions by placing the problem in a historical context. Likewise, the bicentennial

celebration of the American Revolution in 1976 was made an opportunity to study that great historical event that lasted from 1763 to 1787. Some of the impressive pieces on the American Revolution produced by historians such as Makoto Saito, Tadashi Aruga, Takeshi Igarashi, and Akira Imazu deserve special mention (Saito 1976, 1992, Aruga 1976, 1988, Igarashi 1976, 1984, Imazu 1976; see also Imazu's earlier work on the American Revolution (Imazu 1960)).

In the diplomatic history field, Japanese scholars began to take advantage of the new access to previously classified State and Defense Department documents that were made available in the latter part of the 1970s. In addition, the transition to the floating exchange rate system made it easier for Japanese scholars and students alike to visit the United States and conduct research there. As a result, the latter half of the decade began to reap the rich harvest of studies of Occupied Japan and postwar US–Japan relations, complementing those joint bi-national studies in the prewar relations between Japan and the United States (Iokibe 1975, Igarashi 1979a, 1979b, 1979c; for collaborative joint works of Japanese and American scholars on the prewar bilateral relations, see Hosoya and Saito 1978, Hosoya et al. 1971–1972; for a critical review of these US–Japanese collaborative joint works, see Matsuda 1986). New findings obtained by these studies helped deepen an understanding of the US occupation of Japan, revealing an incessant and intricate interaction between the occupiers and the occupied.

5. *American History in the United States in the 1980s*

Historians in America remained intensely interested in the 'new' social history approach during the 1980s. More scholars joined the ranks of social historians. The sharp increase in the number of female historians and of monographs on women's history especially attests to that fact. In 1960, women had only 10 percent of the Ph.Ds conferred, but 30 years later they earned 37 percent of the Ph.Ds (Levine 1993). More diversification in themes, perspectives, and methodologies, the phenomenon of which had started during the previous decade, characterized scholarship in the American history of the 1980s. This resulted in even further fragmentation of the larger picture of American civilization.

Needless to say, the changing academic landscape represented historians' response to changes in demographic composition and the continuing fragmentation of American society itself. As a result, the sense of wholeness with regard to the American past seemed almost completely lost, as if the images of America had exploded into a multitude of isolated pieces. Nothing was more deeply felt than the need for synthesis in American history so as to meet the needs of the American people and the society, as well as to maintain the legitimacy of the profession.

Major scholarly journals, such as the *Journal of American History* and the *American Historical Review*, called upon historians to reflect on the 'unhealthy' conditions that afflicted the field of American history, publishing special issues addressing the central problem of synthesis in American history (e.g. Bender 1986, 1987, Megill 1991). These journals sought to remove the barriers that divided American history into diverse specialized subdivisions, while suggesting ways to help cross the walls of disciplinary specializations and to explore and reinvent a larger synthetic history (for a superb and informative book that systematically treated the 1980s state of American scholarship in American history, see Foner 1990).

Coupled with the general trend mentioned above, a set of new developments that took place in the 1980s considerably changed the texture of scholarship in American history and an approach with culture as the central focus became dominant from this decade. For one thing, the so-called conservative Reagan revolution of the 1980s forced many scholars to confront themselves with the need to explain how conservatives had succeeded in mobilizing the American people so that they could reverse the egalitarian victories won by the oppositional movements of the 1960s. Scholars then came to recognize the importance of the role of culture as a powerful political force. As a result, culture became the defining and dominant theme in the field of American history.

At the same time, as the term 'culture wars' implied, a bitter controversy over multiculturalism swept through the United States during the 1980s and 1990s, transforming the landscape of American academia. (Because it is impossible to list all the literature on the controversy over multiculturism here, it may be sufficient to name only a few. They are the following: Asante 1987, Schlesinger 1992, Takaki 1993, Bell 1992.) The controversy, in part, represented historians' awareness of and their response to change in demographic compositions that had been taking place in American society since the passage of the Immigration and Naturalization Act of 1965. It should be noted, however, that behind all this was a larger change in the cultural climate, in which cultural populism and even cultural relativism took hold.

Scholars were influenced profoundly by the work of cultural anthropologists, especially Clifford Geertz, while gaining significant insights from the works of European philosophers and theorists, such as Antonio Gramsci (i.e., his concept of 'hegemony') and Michel Foucault (i.e., his concept of 'discourse'), in particular (for discussions on the influence of Antonio Gramsci, see, for example, Lears 1985, Fink et al. 1988; with regard to Michel Foucault, see

Foucault 1969; see also Said 1978). They placed new emphasis on the role of culture in their studies, while continuing to pay sufficient attention to other factors, such as politics and economics. The end result was an approach that dominated the study of American history during the 1980s and 1990s with culture as the key theme and that, with further influence from the United Kingdom and France, would eventually come to be the new and influential field of Cultural Studies. But it does not necessarily mean that an approach with culture as the central focus had no problem. While intensively engaged in 'discourse' analyses, historians regrettably showed little interest in either reconstructing a larger synthetic history or presenting the future direction in which humanity in general is moving.

6. *Japanese American Studies Scholarship in the 1980s*

The 1980s witnessed American Studies, especially studies in American social history, in full bloom in Japan. Subtle sophistication in empirical work based on primary source materials characterized the Japanese state of the field now that primary documents became more easily accessible than previously thanks to the appreciated value of the Japanese yen.

There existed two forces at work in opposite directions. While one force was diversifying American history in terms of its topic, approach, methodology, and perspective, the other was an effort to search for an organizing theme and a synthetic American history.

The former was, of course, the 'new' social history in which American labor, immigrants, women, Afro-Americans, Native Americans, other minority groups, children, education, and so forth were each scrutinized from a new perspective. In a special issue of *Amerikashi-Kenkyu* (1981) that featured 'American Democracy Revisited,' for instance, a scholar of Afro-American history picked up the problem of the origins of racial discrimination in America and raised an important question whether discrimination against minorities including blacks and other ethnic or other racial groups should be understood as something that was primarily based in the American capitalist structure in the context of world capitalism, or whether it was something that represented an American form of Europe's colonial domination of the peoples of the non-European world (Ohtsuka 1981, Ikemoto 1987).

Another historian stressed the importance of a 'new' social history approach and placed new emphasis on elucidating relations between ethnicity and class in order to describe the world in which immigrants lived in America (Nomura 1982, 1995). As the influence of America's 'new' social history was strongly felt and became quite noticeable in the 1980s, Japanese scholarship in American history became sub-divided and the image of American civilization consequently became increasingly fragmented.

Evidence of a synthesis appeared particularly in the 1986 issue of *Amerikashi-Kenkyu*, which featured problems focusing on a structuralist understanding of American history and also in the publication of Tadashi Aruga's *Amerikashi Gairon* (*An Introduction to American History* 1987). Aruga picked up themes in American history that he believed to be important in grasping the totality of American civilization, and elaborated on the current state of American scholarship around each theme with the aim of describing a holistic image of America (*Amerikashi-Kenkyukai* 1986, Aruga 1987). These publications, although dwarfed in number by specializing studies, represented the powerful voice and persistent effort of American Studies specialists of Japan to reconstruct and synthesize American history.

In addition, Japanese scholars made the bicentennial year 1987 an occasion to re-examine and discover new meanings in the writing of the American Constitution. For instance, the November 1987 issue of *Shiso* featuring the two-hundredth anniversary of the Constitution of the United States carried impressive articles that dealt with the problem of American Constitutionalism from new and Japanese perspectives ('The Bicentennial of the Constitution of the United States.' *Shiso* 761 (1987)).

Last, but by no means the least, thanks to the new accessibility of documents and collaborative joint work between Japanese and American specialists, the 1980s saw studies in Japanese Occupation and American foreign policy covering the 1940s and 1950s making great progress both quantitatively and qualitatively, shedding new light on US–Japan relations and US global strategy in the Cold War (good examples of such works include Yui, 1985, 1989, Iokibe 1985, Igarashi 1986, Ishii 1989; see also Tokyo Daigaku Amerika Kenkyu Shiryo Sentah 1988).

7. *American History in the United States in the 1990s*

Added to studies based on cultural, post-colonial, and feminist critique, two phenomena helped define the general trends in scholarship in American history in the late 1980s and 1990s. One of those phenomena was the increasing internationalization of American history. As indicated previously, the effort to internationalize US history in part corresponded to demographic changes in American society. But the idea was primarily conceived with the hope that an internationalized American history focusing on interactions between different cultures would not only help broaden the visions of US scholars but also help

American historians rediscover new meanings in the American past by learning from America specialists abroad (Telen 1992b).

The other phenomenon was the innovation of new approaches to the American past. Scholars placed new emphasis on comparative history for one thing, and looked anew at American history from macro-history perspectives, for another. Historians sought to enrich understanding by broadening their vision and comparing ideas, institutions, and social developments in the historical settings of more than two different countries and cultures. A good example of such work might be a group of studies around the topic of slavery, although comparative studies in slavery were not entirely new by any means (for example, see Patterson 1982, Kolchin 1986, 1987, Davis et al. 2000; for discussions on comparative history, see Frederickson 1995, 1998).

At the same time, scholars began to look at the American past by placing the subject in the context of the evolving modern world-capitalist system and historic change over time. Looking at America from transnational perspectives, some immigration scholars, for example, analyzed their field from the perspective of the international movement of labor, as well as in the context of the evolving modern capitalist world system (for references on world-systematic analysis, see Wallerstein 1974, 1980, 1988, Shannon 1989, McCormick 1990, 1995; for discussions on the methodology of macro-historical sociology, see Skocpol 1984). Historians of the slave trade did likewise. By comparing that infamous trade in different countries and regions, they examined the trans-Atlantic slave trade by placing it in the context of the development of the modern world-capitalist system (e.g., see Rawley 1981, Lovejoy 1986, Miers and Roberts 1988). While doing so, these scholars put new emphasis on the forces, constraining as well as promoting, that the external world imposed on the United States and studied the complex and unforeseen consequences that American expansionism had brought to the United States. These scholars hoped that novel approaches and new perspectives would help them rethink American history and raise new questions that had not been asked previously.

With the end of the Cold War and the dismantling of the Soviet Union, American historians responded to the new issue of the rise of ethnic nationalism and other challenges that globalization and transnationalism had aroused for American society and the profession. In fact, American historians spent a considerable amount of time in the 1990s hotly debating critical issues such as 'American exceptionalism' and the writing of a national history vs. the writing of an international history (Iriye 1988, 1989, McMahon 1990, Tyrrell 1991, McGeer 1991, Appleby 1992, Kammen 1993, Nelles 1997, Koschmann 1997, Nolan 1997, 1999). During the same decade, American scholars also tackled the vexing problem of 'objectivity' in historical writing, a problem that primarily resulted from professionalization and specialization in American history. They also sought to ameliorate some of the past parochialism and national bias of their writings (Novick 1988; see chap. 4 'Objectivity in Crisis'; Kloppenberg 1989, Hexter et al. 1991, Megill 1991). As was mentioned earlier, the Organization of American Historians launched the project to internationalize American history as part of the efforts to rectify its members' tendency toward excessive focus on unilateral national history.

8. *Japanese Scholarship in American History in the 1990s*

Japanese scholarship in American history in the 1990s paralleled its American counterpart in several important ways. First, Japanese studies in American history became increasingly subdivided, and so gravely diversified in terms of topics, themes, analytical viewpoints, and methodologies that it became next to impossible to visualize the whole image of the United States. To illustrate the situation, one might note the new procedure that Amerika Gakkai (the Japanese Association for American Studies, hereafter abbreviated as JAAS) adopted at its Annual Meeting from 1998, whereby a number of small sectional meetings were simultaneously divided according to topics and themes in order to cope with problems arising from the diversification of scholarly interests among Japan's America specialists. The acute sense of crisis manifested itself in a symposium held at the JAAS's 1999 Annual Meeting in which these scholars discussed whether or not an integrative image of America was even possible. The provisional conclusion that was reached at that meeting was not very optimistic. Against this backdrop, Nagayo Homma's trilogy, written with the specific aim of presenting a comprehensive picture of America, deserves special mention (Homma 1991a, 1991b, 1991c).

Second, both women's history and critical cultural approaches that constituted a part of the 'new' social history became fashionable and dominant in the profession. These helped make scholarship in American history truly interdisciplinary. To illustrate these trends, it may be sufficient to point out that the number of female scholars specializing in women's history has increased by leaps and bounds during the decade. To meet the needs and interest of these scholars, for instance, Amerikashi-Kenkyukai (The Society of American Historical Studies) published its 1997 issue of *Amerikashi-Kenkyu* with articles on 'American History and Gender,' while in 1995 JAAS prepared a special number of its professional journal featuring articles on 'Family, Children, and Education.' In the following year, the 1996 issue of *Amerikashi-Kenkyu* featured articles on 'multi-culturalism and minority groups,' while the prestigious

journal *Shiso* devoted its entire space to discussing 'Cultural Studies' in the same year. In this connection, among several works using critical cultural studies approaches, books of high quality that are worthy of note included *Esunikku Jokyo no Genzai* (*The Current State of Ethnicity*) edited by Tadashi Aruga and the anthology, *Tabunkashugi no Amerika* (*Multiculturalism in America*), coedited by Daizaburo Yui and Yasuo Endo (Aruga 1994, Yui and Endo 1999).

Third, that historians broadened their horizons and treated their subject from transnational and comparative perspectives characterized Japanese scholarship in American history in the decade. It meant that scholars broadened not only the scope and breadth of the subject of historical inquiry but also the visions in terms of time and space in order to rectify their proclivity toward excessive focus on unilateral national history.

In 1992, for instance, a group of historians made the year an occasion to look anew at the American past, placing each one of their subjects in the context of the entire Western Hemisphere, including Canada, the Caribbean Basin, and Latin America. Their collaborative joint work produced a five-volume history of North and South America to commemorate the quincentenary of the 'discovery' of the New World and elucidated several different paths of development in each country and region treated (Rekishigaku Kenkyukai 1993; for a work that analyzed the nineteenth-century American expansionism from world-systems perspectives, see Matsuda 1993).

Historians also sought to revise previous interpretations tainted with Euro-centric views of history by posing critical questions about linear developmentalist perspectives and the range of assumptions which had been uncritically accepted previously. The 1997 publication of Iwanami Koza's world history volume entitled *Kan Taiseiyo Kakumei* (*The Atlantic Rim Revolution*) was a case in point (Iwanami Koza Sekai Rekishi 17 1997, Matsui 1991). (Needless to say, this was, by no means, the first book of this kind written by Japanese scholars. In fact, some time before a group of economic historians had already produced a couple of brilliant works by looking at the subject from a transnational and global perspective; see Kawano and Iinuma 1967, 1970; for American scholarship that has a trans-Atlantic perspective, see Kraus 1949, Palmer 1959, McCuster and Menard, 1985, Greene 1986, Greene et al. 2000.) Influenced by the work of European historians, such as Fernand Braudel and Immanuel Wallerstein, scholars who participated in the project sought to elucidate the structure of the eighteenth to nineteenth century trans-Atlantic world and the unfolding of the major revolutionary movements that had happened in a hierarchically but interdependently structured trans-Atlantic world from the latter part of the eighteenth century to the early nineteenth century. In their analysis of the American Revolution, for example, they treated the subject from a perspective of international relations, placing new emphasis on the interconnectedness in the development of the countries that comprised the Atlantic Rim Community. It was an ambitious project by any standard of historical writing, seeking to reinterpret modern history from a fresh and transnational perspective.

Several important studies in comparative history were carried out as well (Yui et al. 1994, Toyoshita 1992, Rekishigaku Kenkyukai 1989). Space precludes the present writer from dwelling on the content of each work here. Suffice it to mention the names of a couple of excellent works that treated the problem of occupation reform from a transnational perspective. *Senryo Kaikaku no Kokusai Hikaku* (*International Comparative Studies in Reform in Occupied Areas*), a joint work of 14 scholars, compared different shades and aspects of postwar reform in occupied areas of Europe and Japan as well as other Asian countries, such as the Philippines and Korea. Naruhiko Toyoshita's *Nihon Senryo Kanritaisei no Seiritsu* (*The Establishment of the Control System of Occupied Japan*) is another work worthy of note. The author explored similarities and differences in occupation policy by comparing the case of the US occupation of Japan with that of the postwar occupation of Italy, Germany, Hungary, and Bulgaria. This work set a good example of a comparative study conducted by an individual scholar for those scholars who would come after him.

The commitment of Japanese scholars to internationalize their work was the fourth and last feature that, characterized Japanese scholarship in American history in the 1990s. The English language publication by the Japanese Association for American Studies of its professional journal of American Studies illustrated the point. It is true that JAAS had started doing this from 1980 as part of the effort to promote international exchange with its foreign scholars, yet what distinguished the record of the 1990s from that of the previous decade was the diversification of featured topics, not limiting them to US–Japan relations and their related themes alone. A wide range of topics were presented, for example, from 'Nature and Environmental Issues' to 'Another "American Century"?' Japanese scholars displayed their intention to live up to the global standard of scholarship as well as to contribute to deepening an understanding of American civilization by sharing scholarly findings with their foreign peers (for a historigraphical essay that succinctly reviews Japanese scholarship in the history of US–East Asian relations, see Aruga 1996).

This feature, together with others mentioned above, seems to suggest simultaneity and interaction in the scholarship in American history that has taken

place in and between Japan and the United States, but at the same time it shows the independence of one from the other. In the age of globalization, American Studies specialists in Japan seem at last to be catching up with their American counterparts in terms of the times as well as in the quality of the scholarship they produce. From now on, non-American scholars like the Japanese are increasingly expected to exchange ideas and thoughts freely with US scholars on an equal footing, contributing to a deeper understanding of the United States. In order to do this, it goes without saying that what is required is an ongoing effort to produce original, creative work by questioning, constantly and rigorously, both uncritical assumptions and conventional views, opening up new horizons in the field of American Studies.

Bibliography

Amerikashi Kenkykai (ed.) 1980 Nyu Refuto Shigaku to Sono Shuhen (New Left Historiography and Its Neighboring Fields). *Amerikashi-Kenkyu* 3

Amerikashi-Kenkyukai (ed.) 1985 *Amerikashi-Kenkyu* (*Studies of American History*) 8. Tokyo, Japan

Amerikashi-Kenkyukai (ed.) 1986 *Amerikashi-Kenkyu 9*. Tokyo, Japan

Appleby J 1992 Recovering America's historic diversity: Beyond exceptionalism. *The Journal of American History* **79**(2): 419–31

Aruga N 1979 Studies of American social history: An examination from the viewpoint of 'new social history.' *The Bulletin of the Center for American Studies of the University of Tokyo* **2**: 18–30

Aruga N 1998 Social history. In Abe H, Igarashi T (eds.) *Amerika Kenkyu Annai (A Guide to American Studies)*. Tokyo Daigaku Shuppankai, Tokyo, pp. 67–85

Aruga T 1976 Amerika Kakumei to Amerika Gaikoseisaku no Kigen (The American Revolution and the origins of American foreign policy). *Amerika Kenkyu* (*The American Review*) A Special Issue: 19–28

Aruga T 1987 *Amerikashi Gairon (An Introduction to American History)*. Tokyo Daigaku Shuppankai, Tokyo

Aruga T 1988 *Amerika Kakumei (The American Revolution)*. Tokyo Daigaku Shuppankai, Tokyo

Aruga T 1992 Japanese scholarship and the meaning of American history. *The Journal of American History* **79**(2): 504–14

Aruga T (ed.) 1994 *Esunikku Jokyo no Genzai (The Current State of Ethnicity)*. Nihon Kokusaimondai Kenkyusho, Tokyo

Aruga T 1996 Japanese scholarship in the history of U.S.–East Asian relations. In: Cohen W I (ed.) *Pacific Passage: The Study of American–East Asian Relations on the Eve of the Twenty-First Century*. Columbia University Press, New York, pp. 36–87

Asante M K 1987 The *Afrocentric Idea*. Temple University Press, Philadelphia

Bell D 1992 The cultural wars: American intellectual life 1965–1992. *The* Wilson *Quarterly* (Summer): 74–107

Bender T 1986 Wholes and parts: The need for synthesis in American history. *The Journal of American History* **73**(1): 120–36

Bender T 1987 Synthesis in American history: A round table. *The Journal of American History* **74**(1): 107–30

Boorstin D J 1953 *The Genius of American Politics*. The University of Chicago Press, Chicago

Buhle P (ed.) 1990 *History and the New Left: Madison, Wisconsin 1950–1970*. Temple University Press, Philadelphia

Davis D B et al. 2000 AHR forum: Crossing slavery's boundaries. *The American Historical Review* **105**(2): 451–84

Erickson C 1975 Quantitative history. *The American Historical Review* **80**(2): 351–65

Fink L et al. 1988 A round table: Labor, historical pessimism, and hegemony. *The Journal of American History* **75**(1): 115–61

Fogel R W 1975 The limits of qantitative methods in history. *The American Historical Review* **80**(2): 329–50

Foner E (ed.) 1990 *The New American History*. Temple University Press, Philadelphia

Foucault M 1969 *L'Archeologie du savoir*. Gallimard, Paris

Frederickson G M 1995 From exceptionalism to variability: Recent developments in cross-national comparative history. *The Journal of American History* **82**(2): 587–604

Frederickson G M 1998 America's diversity in comparative perspective. *The Journal of American History* **85**(3): 859–75

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Greene J P 1986 *Peripheries and Center: Constitutional Development in the Extended Polities of the British Empire and the United States 1607–1788*. W. W. Norton and Company, New York

Greene J P et al. 2000 AHR forum: Revolutions in the Americas. *The American Historical Review* **105**(1): 92–152

Hartz L 1955 *The Liberal Tradition in America: An Interpretation of American Political Thought since the Revolution*. Harcourt, Brace & World, Inc., New York

Henretta J A et al. 1979 AHR forum: Social history as lived and written. *The American Historical Review* **84**(5): 1293–333

Hexter J H et al. 1991 AHR forum: Peter Novick's That Noble Dream: The objectivity question and the future of the historical profession. *The American Historical Review* **96**(3): 675–708

Higham J 1982 Current trends in the study of ethnicity in the United States. *The Journal of American Ethnic History* **2**(1): 5–15

Higginbotham E B 1992 African-American women's history and the metalanguage of race. *Signs* **17**(2): 251–74

Hofstadter R 1955 *The American Political Tradition and the Men Who Made It*. Vintage Books, New York

Homma N 1976 *Rinen no Kyowakoku (The Republic of Ideas)*. Chuokoronsha, Tokyo

Homma N 1991a *Amerika Shizo no Tankyu (In Search for an American Historical Image)*. Tokyo Daigaku Shuppankai, Tokyo

Homma N 1991b *Utsuriyuku Amerika (Changing America)*. Chikuma Shobo, Tokyo

Homma 1991c *Amerika Bunka no Hiirohtachi (Heroes in American Culture)*. Shinchosha, Tokyo

Homma N 1999 Amerika no Imi (The meanings of America). *Doshisha Amerika Kenkyu (Doshisha American Studies)* **35**: 5–17

Honda S (ed.) 1989 *Amerika Shakaishi no Sekai* (*The World of American Social History*). Sanseido, Tokyo

Hosoya C, Saito M, Imai S, Rohyama M (eds.) 1971–1972 *Nichibei Kankeishi 1932–1941*, 4 Vols. Tokyo Daigaku Shuppankai, Tokyo

Hosoya C, Saito M (eds.) 1978 Washington Taisei to *Nichibei Kankeishi (The Washington Treaty System and Japanease–American Relations)*. Tokyo Daigaku Shuppankai, Tokyo

Igarashi T 1976 Pensirubenia Kyouwaha no Seiji Shido (1) (2) (3) (Political leadership of Republicans in Pennsylvania). *Kokka Gakkai Zasshi* **89**(3–8)

Igarashi T 1979a American-Japanese peace-making and the Cold War 1947–1951. *Amerika Kenkyu* **13**: 166–187

Igarashi T 1979b Tainichi Senryo Seisaku no Tenkan to Reisen. In Nakamura T (ed.) *Senryoki Nippon no Keizai to Seiji (The Economy and Politics of Occupied Japan)*. Tokyo Daigaku Shuppankai, Tokyo, pp. 25–57

Igarashi T 1979c George Kennan to Tainichi Senryo Seisaku no Tenkan. In: Nakamura T (ed.) *Senryoki Nippon no Keizai to Seiji (The Economy and Politics of Occupied Japan)*. Tokyo Daigaku Shuppankai, Tokyo

Igarashi T 1984 *Amerika no Kenkoku (The Founding of America)*. Tokyo Daigaku Shuppankai, Tokyo

Igarashi T 1986 *Tainichi Kowa to Reisen (Amreican–Japanese Peace Making and the Cold War)*. Tokyo Daigaku Shuppankai, Tokyo

Ikemoto K 1987 *Kindai Doreishakai no Shiteki Tenkai (The Historical Evolution of Society Based on Modern Slavery)*. Mineruva Shobo, Kyoto

Imazu A 1960 *Amerika Kamumeishi Josetsu (An Introduction to the American Revolution)*. Hohritsu Bunkasha, Kyoto

Imazu A 1976 *Amerika Dokuritsu no Hikari to Kage (The Light and Shade of American Independence)*. Shimizushoin, Tokyo

Iokibe M 1975 Beikoku niokeru Tainichi Seisaku no Keisei Katei (The policy-making processes of the United States toward Japan). *Kokusaiho Gaiko Zasshi* **74**(3–4)

Iokibe M 1985 *Beikoku no Nihon Senryo Seisaku (The U.S. Occupation Policy of Japan)*, 2 Vols. Chuokoronsha, Tokyo

Iriye A 1988 Exceptionalism revisited. *Reviews in American History* **16**(2): 291–97

Iriye A 1989 The internationalization of history. *The American Historical Review* **94**: 1–10

Ishii O 1989 *Reisen to Nichibei Kankei (The Cold War and Japanese–American Relations)*. Japan Times, Tokyo

Iwanami Koza Sekai Rekishi 17 1997 *Kan Taiseiyo Kakumei (The Atlantic Rim Revolution)*. Iwanami Shoten, Tokyo

Kammen M (ed.) 1980 *The Past Before Us: Contemporary Historical Writing in the United States*. Cornell University Press, Ithaca

Kammen M 1993 The problem of American exceptionalism: A reconsideration. *American Quarterly* **45**(1): 1–43

Kawano K, Iinuma J (eds.) 1967 *Sekai Shihonshugi no Keisei (The Formation of World Capitalism)*. Iwanami Shoten, Tokyo

Kawano K, Iinuma J (eds.) 1970 *Sekai Shihonshugi no Rekishi Koso (The Historical Structure of World Capitalism)*. Iwanami Shoten, Tokyo

Kloppenberg J T 1989 Objectivity and historicism: A century of American historical writing. *The American Historical Review* **94**(4): 1011–30

Kolchin P 1986 Some recent works on slavery outside the United States: An American perspective. *Comparative Studies in Society and History* **28**: 767–77

Kolchin P 1987 *Unfree Labor: American Slavery and Russian Serfdom*. Harvard University Press, Cambridge

Koschmann J V 1997 The nationalism of cultural uniqueness. *The American Historical Review* **102**(3): 758–68

Kraus M 1949 *The Atlantic Civilization: 18th Century Origins*. Cornell University Press, Ithaca

Lears T J J 1985 The concept of cultural hegemony: Problem and possibilities. *The American Historical Review* **90**(3): 567–93

Lemisch J 1967 The American Revolution seen from the bottom up. In: Bernstein B J (ed.) *Towards A New Past: Dissenting Essays in American History*. Vintage Books, New York

Levine L W 1993 Clio, canon, and culture. *The Journal of American History* **80**(3): 862

van der Linden A A M 1996 *A Revolt Against Liberalism: American Radical Historians, 1959–1976*. Rodopi B.V., Amsterdam

Lovejoy P E (ed.) 1986 *Africans in Bondage: Studies in Slavery and the Slave Trade*. University of Wisconsin Press, Madison

Matsuda T 1986 The coming of the Pacific War—Japanese perspectives. *Reviews in American History* **14**(4): 629–52

Matsuda T 1993 Beikoku Bochoshugi no Tenkai (The unfolding of American expansionism). In Takahashi A, Kamo Y (eds.) *Kindaika no Wakaremichi* (*The Crossroads of Modernization*). *Nanboku Amerika no Gohyakunen*. Aoki Shoten Tokyo: Vol 2, pp. 270–300

Matsui T 1991 *Sekai Shijo no Keisei (The Formation of the World Market)*. Iwanami Shoten, Tokyo

McCormick T J 1990 World systems. *The Journal of American History* **77**(1): 125–32

McCormick T J 1995 *America's Half-Century: United States Foreign Policy in the Cold War and After*, 2nd edn. Johns Hopkins University Press, Baltimore

McCuster J J, Menard R R 1985 *The Economy of British America 1607–1789*. University of North Carolina Press, Chapel Hill

McGeer M 1991 The price of the 'new transnational history. *The American Historical Review* **96**(4): 1056–67

McMahon R J 1990 The study of American foreign relations: National history or international history. *Diplomatic History* **14**(4): 554–64

Megill A 1991 Fragmentation and the future of historiography. *The American Historical Review* **96**(3): 693–98

Miers S, Roberts R (eds.) 1988 *The End of Slavery in Africa*. University of Wisconsin Press, Madison

Nelles H V 1997 American exceptionalism: A double-edged sword. *The American Historical Review* **102**(3): 749–57

Nolan M 1997 Against Exceptionalisms. *The American-Historical Review* **102**(3): 769–74

Nolan M 1999 The nation and beyond: A special issue. *The Journal of American History* **86**(3)

Nomura T 1982 Rodoshi to Iminshi no Ketsugo ni tsuite (On integrating labor history and immigration history). *Amerika-shi-Kenkyu* **5**: 8–13

Nomura T 1985 *Yudaya Imin no Nyu Yohku (New York Where Jewish Immigrants Lived)*. Yamakawa Shuppan, Tokyo

Novick P 1988 *That Noble Dream: The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK

Ohtsuka H 1981 Amerika Minshushugi to Jinshu Sabetsu (American democracy and racial discrimination). *Amerika-shi-Kenkyu* **4**: 14–21

Palmer R R 1959 *The Age of the Democratic Revolution: A Political History of Europe and America 1700–1800*. Princeton University Press, Princeton

Patterson O 1982 *Slavery and Social Death: A Comparative Study*. Harvard University Press, Cambridge

Rawley J A 1981 *The Atlantic Slave Trade: A History*. Norton, New York

Rekishigaku Kenkyukai (ed.) 1989 Senryo to Kyutaisei—Sono Kokusai Hikaku (Occupation and old Regime—its international comparisons). *Reshikigaku Kenkyu* (*Journal of Historical Studies*) **600**

Rekishigaku Kenkyukai (ed.) 1993 *Nanboku Amerika no Gohyakunen (A 500-Year History of North and South America)*, 5 Vols. Aoki Shoten, Tokyo

Said E W 1978 *Orientalism*. Georges Borchardt Inc., New York

Saito M (ed.) 1976 Amerika Kakumei (The American Revolution). *Sogo Kenkyu Amerika 3, Minshusei to Kenryoku*. Kenkyusha, Tokyo, pp. 3–29

Saito M 1992 *Amerika Kakumeishi Kenkyu (A Study of the American Revolution)*. Tokyo Daigaku Shuppankai, Tokyo

Schlesinger Jr A M 1992 *The Disuniting of America: Reflection on a Multicultural Society*. W. W. Norton & Co New York

Shannon T R 1989 *An Introduction to the World-System Perspective*. Westview Press, Boulder, CO

Skocpol T (ed.) 1984 *Vision and Method in Historical Sociology*. Cambridge University Press, Cambridge, UK

Swierenga R P 1974 Computers and American history: The impact of the 'new' generation. *The Journal of American History* **60**(4): 1045–70

Takahashi A 1982 Amerika Kaikaku Seiji Kenkyu no Seika to Kadai (The achievements and assignments of studies in American reform politics.) *Tokyo Daigaku Amerika Kenkyu Shiryo Sentah Nenpoh (The Bulletin of the Center for American Studies of the University of Tokyo)* **5**: 8–16

Takahashi A 1999 *Amerika Teikokushugi Seiritsushi no Kenkyu (A Study of the Establishment of American Imperialism)*. Nagoya Daigaku Shuppankai, Nagoya

Takaki R 1993 *Different Mirror: A History of Multicultural America*. Little, Brown and Company, Boston

Thelen D 1992a Of audiences, borderlands, and comparisons: Toward the internationalization of American history. *Journal of American History* **79**(2): 434

Thelen D 1992b Of audiences, borderlands, and comparisons: Toward the internationalization of American history. *The Journal of American History* **79**(2): 432–53

Tilly C 1984 The old new social history and the new old social history. *Review* **7**(3): 363–406

Tokyo Daigaku Amerika Kenkyu Shiryo Sentah (ed.) 1988 Sengo Nichibei Kankei no Kenkyu. *Tokyo Daigaku Amerika Kenkyu Shiryo Sentah Nenpoh (The Bulletin of the Center for American Studies of the University of Tokyo)* **11**: 39–82

Toyoshita N 1992 *Nihon Senryo Kanritaisei no Seiritsu (The Establishment of the Control System of Occupied Japan)*. Iwanami Shoten, Tokyo

Tucker R W 1971 *The Radical Left and American Foreign Policy*. Johns Hopkins University Press, Baltimore

Tyrrell I 1991 American exceptionalism in an age of international history. *The American Historical Review* **96**(4): 1031–55

Wallerstein I 1974 *The Modern World System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

Wallerstein I 1980 *The Modern World-System II: Mercantilism and the Consolidation of the European World-Economy*. Academic Press, New York

Wallerstein I 1988 *The Modern World-System III: The Second Era of Great Expansion of the Capitalist World-Economy 1730–1840*. Academic Press, San Diego, CA

Williams W A 1959 *The Tragedy of American Diplomacy*. The World Publishing Company, Cleveland

Williams W A 1961 *The Contours of American History*. The World Publishing Company, Cleveland

Williams W A 1989 A round table: What has changed and not changed in American historical practice? *The Journal of American History* **76**(2): 399–478

Yui D 1985 *Sengo Sekai Chitsujo no Keisei (The Formation of the Postwar World Order)*. Tokyo Daigaku Shuppankai, Tokyo

Yui D 1989 *Mikan no Senryo Kaikaku (Incomplete Occupation Reform)*. Tokyo Daigaku Shuppankai, Tokyo

Yui D, Endo Y (eds.) 1999 *Tabunkashugi no Amerika (Multiculturalism in America)*. Tokyo Daigaku Shuppankai, Tokyo

Yui D, Nakamura M, Toyoshita N (eds.) 1994 *Senryo Kaikaku no Kokusai Hikaku (International Comparative Studies in Reform in Occupied Areas)*. Sanseido, Tokyo.

Takeshi Matsuda

Copyright © 2004 Elsevier Ltd.
All rights reserved.

 ISBN: 0-08-043076-07

Amnesia

1. Introduction

The term *amnesia*, as a description of a clinical disorder, refers to a loss of memory for personal experiences, public events, or information, despite otherwise normal cognitive function. The cause of amnesia can be either primarily *organic*, resulting from neurological conditions such as stroke, tumor, infection, anoxia, and degenerative diseases that affect brain structures implicated in memory; or it can be primarily *functional or psychogenic*, resulting from some traumatic psychological experience (see *Amnesia: Transient and Psychogenic*). This article will focus on organic amnesia. The following questions are addressed: What are the characteristics of amnesia? What structures are involved in forming memories (and whose damage causes amnesia) and what function does each serve in the process? Does amnesia affect recent and remote memories equally and, by implication, are memory structures involved only in memory formation and shortly thereafter, or are they also implicated in retention and retrieval over long intervals? Are all types of memory impaired in amnesia or is amnesia selective, affecting only some types of memory and not others? What implication does research on amnesia have on research and theory on normal memory?

2. Characteristics of Organic Amnesia

The typical symptoms of organic amnesia are the opposite of those of functional amnesia: old memories and the sense of self or identity are preserved but the ability to acquire new memories is severely impaired. Though capturing an essential truth about organic amnesia, this statement needs to be qualified in important ways in light of new research.

The scientific investigation of organic amnesia effectively began with Korsakoff's (1889) description of its symptoms at the turn of the century, during what Rozin (1976) called the 'Golden Age of Memory Research.' Likewise, it can be said that the modern era of neuropsychological research on memory and amnesia was ushered in by Scoville and Milner's (1957) publication of the effects of bilateral medial temporal lobectomy to control intractable epilepsy in a single patient, H.M. (see Fig. 1.

Some aspects of the disorder Korsakoff described are peculiar to a kind of amnesia that now bears his name (amnesia related to vitamin (thiamine) deficiency typically associated with alcoholism), while others are common to all forms of amnesia, including the one described by Scoville and Milner. The symptoms are best described by contrasting impaired abilities with preserved ones. They are as follows:

(a) Memory is impaired for various types of material, which is why the amnesia is often referred to as global, though as we shall see not all memories are affected equally. Perception, intelligence, and other cognitive functions are relatively preserved. Thus, amnesic people perform normally on standard tests of intelligence, but are impaired by comparison on standard tests of memory. They can play chess, solve crossword and jigsaw puzzles, comprehend complex instructions, and reason logically.

(b) The memory that is impaired is only *long-term* or *secondary memory*; that memory which lasts long after the information has been received and registered, and not *short-term or primary memory* which is used to hold information briefly in mind. Unaffected also is *working memory* (Baddeley 1986) which is used to operate on the information held in mind. As an example, amnesic people have a normal digit span which refers to the number of digits one can repeat immediately in sequence (typically seven, plus or minus two) and even a normal backward digit span in which the digits must be recalled in reverse order. They are impaired, however, at retaining and recalling even a sub-span list of words after a short interval that is filled with distracting activity, even if given ample opportunity to rehearse the material before. In a more ordinary life example, amnesic people can remember sentences well enough to respond to them, but cannot follow a conversation to the end if reference is made to utterances that occurred at the beginning. The same dissociation between long- and short-term memory holds for words, stories, complex visual patterns, faces, melodies, and some esthetic stimuli (Kolb and Whishaw 1996).

(c) Long-term memory loss in amnesia is most noticeable for events and information acquired after the onset of the disorder and into the future as well as in the period immediately preceding it, but not for information acquired long before that. That is, they have an *anterograde amnesia* that extends into the future but a *retrograde amnesia* limited to the time preceding the onset of the disorder. Thus, amnesic people have difficulty remembering what they learned or experienced since the onset of the disorder, even their current address and neighborhood if they had moved to a new place, but they can remember their old home and neighborhood, and some old experiences and events (Milner 1966, Squire and Alvarez 1995). Retention of such old memories, however, may be more selective and not as well-preserved as had once been believed (see below).

(d) Anterograde long-term memory loss in amnesia applies only to information that can be recollected consciously or *explicitly*. Acquisition, retention, and recovery of memory without awareness or *implicitly*

is normal. For example, having studied a set of words or pictures, amnesic patients perform poorly when their memory is tested explicitly with recognition (Which of the following items do you recognize as having been presented to you?) or recall (Tell me the items you had studied). Their memory for studied items is normal if they are tested implicitly by seeing how performance is altered by the study experience

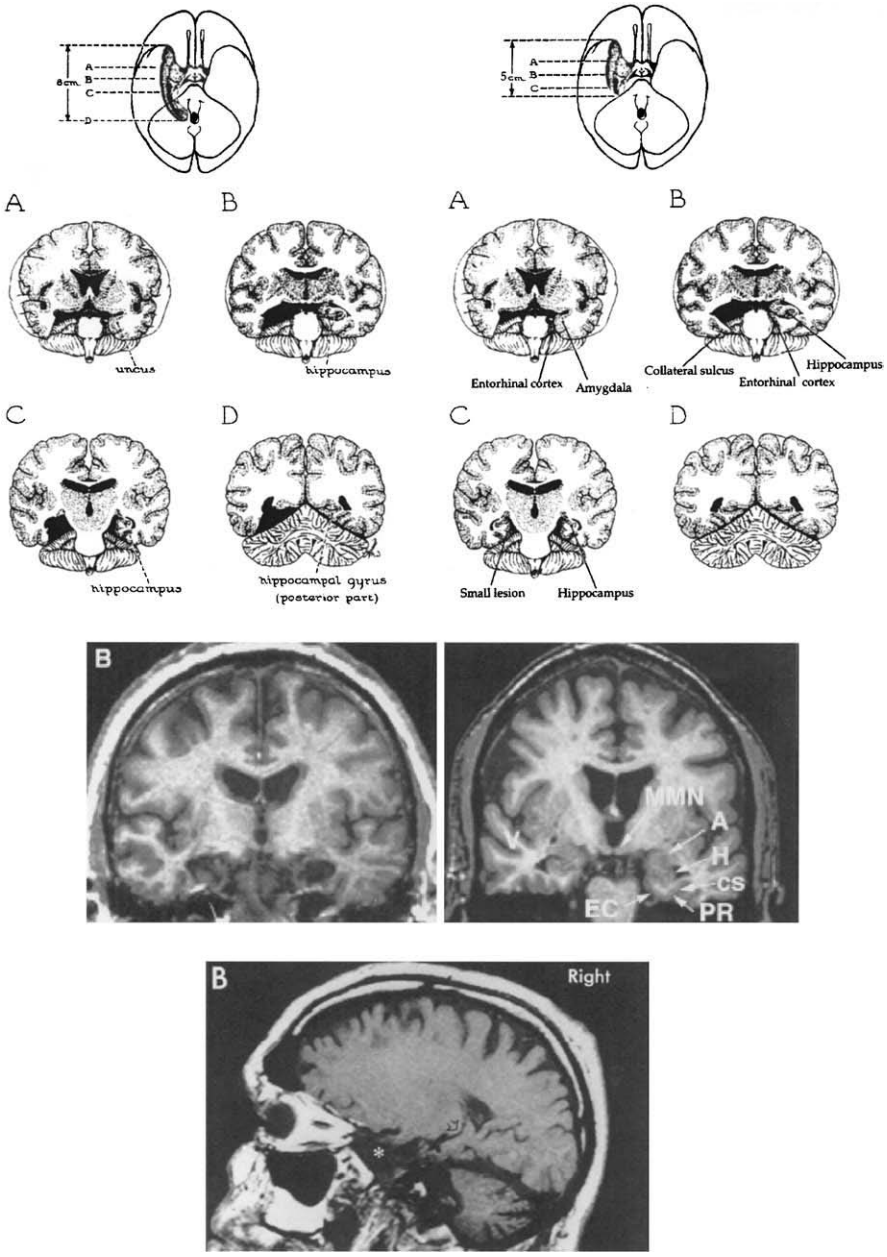

Figure 1
A recreation of H.M.'s lesion from the surgeon's report (left drawings A, B, C, D) and from a recent MRI scan (see bottom panels) of the lesion (right drawings, A, B, C, D). The surgeon overestimated the posterior extent of the lesion. The right side of each drawing is intact for comparison purposes. The two MRI scans labeled B are of H.M.'s brain, and the remaining one is of a control's brain. MMN = Mammilary bodies; A = Amygdala; H = Hippocampus; CS = collateral sulcus; PR = Perirhinal cortex; EC = Entorhinal cortex

without making any explicit reference to the study episode. Thus, though they cannot recall or recognize the items they studied, amnesic people will perceive them more quickly and accurately than items they did not study, or complete them better when they are degraded, such as by filling in the missing letters in a word or completing the lines in a picture. Performance on these implicit tests of memory indicates that information about the studied items is held in memory though they are not aware of that (Tulving and Schacter 1990, Schacter and Buckner 1998).

These characteristics will serve as the foundation for later discussion of empirical and theoretical investigations of amnesia, normal memory, and brain function. Indeed, research on amnesia, fascinating in its own right, has had a powerful impact on memory research and theory, especially since the landmark discovery of Scoville and Milner. This research can be divided into two interacting streams: a *functional neuroanatomical* one, concerned with identifying the neuroanatomical substrates of memory and determining their precise function; and a *(neuro)psychological* one, concerned with the implication that amnesia has for understanding normal memory at a functional level. Each is dealt with in turn.

3. Neuroanatomy of Amnesia

Amnesia is caused by bilateral damage to structures in the limbic system and to adjacent cortex in the medial temporal lobes (see Figs. 2 and 3). The hippocampal formation, in the medial temporal lobes, is the most prominent of the memory structures. It consists of the hippocampus proper with its various subfields and regions, plus the dentate gyrus and the subilculum. Communication between the hippocampus and neocortex occurs thorough a series of relays. The hippocampus is connected directly to the entorhinal cortex which in turn is connected to the parahippocampal gyrus and

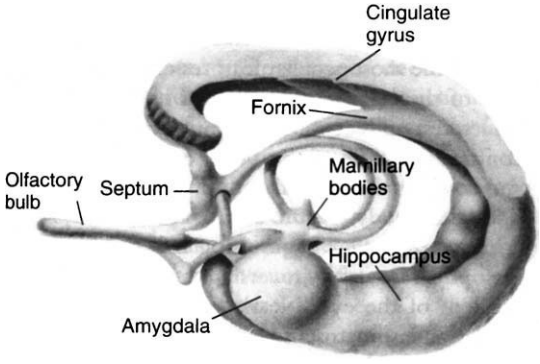

Figure 2
The limbic system and memory circuit (from Hamilton, 1976)

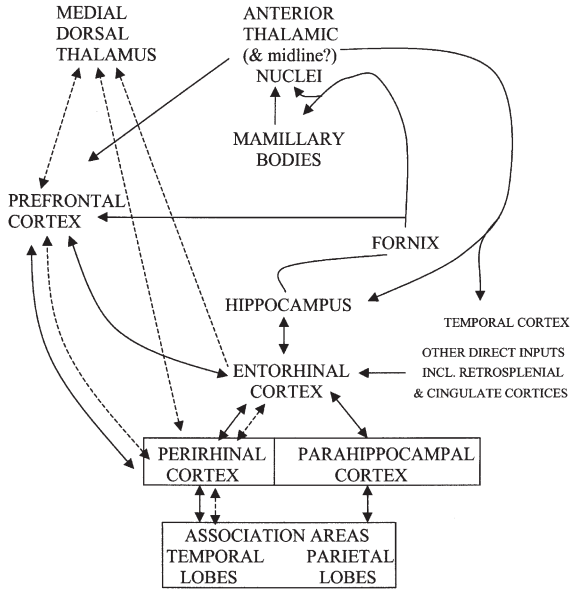

Figure 3
The hippocampal-diencephalic systems (modified from Aggleton and Brown 1999). There are two inter-related systems: The *hippocampal-fornix-anterior thalamic system* indicated by solid lines and the *perirhinal-medial dorsal thalamic system* indicated by dashed lines

perirhinal cortex which project bidirectionally primarily to the temporal and parietal lobes of neocortex, respectively. There are also projections from the hippocampus and perirhinal cortex via the fornix and anterior cingulate to the mammillary bodies and the dorsomedial nucleus of the thalamus, which are in the diencephalon. The loop of medial temporal and diencephalic structures constitute the limbic system. The hippocampus is thus ideally situated to collate information both about the cognitive (neocortex) and emotional (limbic) state of the organism and bind that information into a memory trace that codes for all aspects of a consciously experienced event (Moscovitch 1995).

3.1 Functions of the Neuroanatomical Substrates of Memory

There is considerable debate about the role that each of the components of the memory system have and how they interact with one another. Korsakoff's amnesia is associated with damage to the diencephalon whereas amnesia caused by anoxia, encephalitis, and some degenerative disorders such as Alzheimer's Disease is associated most often with medial temporal lobe damage. Although close inspection reveals differences in memory loss among the various conditions, some have not been substantiated reliably

and others could not be attributed with certainty to differences associated with diencephalic and medial temporal lobe damage, as compared to damage to other structures that often accompanies the various disorders. For example, it had been proposed that the medial temporal lobes were necessary for consolidation and retention of new information whereas the diencephalon was needed for encoding. Investigators claimed that people with diencephalic amnesia had difficulty acquiring information, but once they did so they retained it normally. People with medial temporal damage, however, showed abnormally rapid forgetting. Other investigators, however, found no difference in forgetting rates between amnesic groups, or even between them and controls (see review in Aggleton and Brown 1999).

Others noted that people with Korsakoff's, but not medial temporal, amnesia showed a number of memory deficits in addition to loss of memory for the content of an experienced event. These included confabulation, poor memory for the source and temporal order of events, susceptibility to interference, and poor meta-memory (knowledge about memory) (Moscovitch and Wincour 1992). These deficits, all of which involve strategic aspects of memory, were shown to be related more to frontal dysfunction that often accompanies Korsakoff's amnesia, rather than to diencepahlic damage *per se*.

Influenced by research on animal models of memory, investigators focused on differences among the structures in the medial temporal lobe itself, and its projections to the diencephalon. O'Keefe and Nadel (1978) proposed that the hippocampus is needed for representing allocentric spatial information, or cognitive maps but not for representing egocentric representations, routes and landmarks. While acknowledging the role of the hippocampus in allocentric spatial memory, other investigators disputed that its only function is spatial. Instead, they claimed that the hippocampus is needed for memory for all types of relational information, whether it be among spatial elements, among objects, or among words (Cohen and Eichenbaum 1993). In support of the latter idea, investigators working with animal models showed that memory for single objects is affected little, if at all, by hippocampal lesions, but is disrupted severely by lesions to the perirhinal cortex, whereas memory for the association between objects and locations is affected by parahippocampal lesions (Murray 1996, Aggleton and Brown 1999). This division of labor is consistent with the connection these regions have to neocortex. The perirhinal cortex is connected primarily to the temporal lobes which are concerned with processing objects, whereas the parahippocampus is connected to the parietal lobes which specialize in processing spatial information.

Aggleton and Brown (1999) proposed that the functional differentiation among medial temporal lobe structures extends to the diencephalon, thus forming two integrated medial–temporal–diencephalic memory systems. One system, consisting of the hippocampus and its connections to the mamillary bodies and anterior thalamic nuclei, mediates memory for recall which relies on relational information; the other, consisting of the perirhinal cortex and its connections to the dorsomedial nucleus of the thalamus, is necessary for item recognition which depends on familiarity judgments.

The lesion evidence in humans is roughly consistent with these proposals based on animal models where dissociations of function are observed along the lines predicted by the models (but see commentary in Aggleton and Brown 1999). Because amnesic people with circumscribed lesions are rare, investigators have turned to neuroimaging studies in normal people to test the hypothesis that different aspects of memory are mediated by different regions of the medial temporal lobe. In general, the evidence has been supportive of the hypotheses that have been advanced here, with greater activation in the right hippocampus on tests of spatial (Maguire et al. 1998) and relational memory (Henke et al. 1999) and in the parahippocampal gyrus (possibly entorhinal cortex) on tests of object-location memory (Milner et al. 1997) and navigation (Agguire and D'Esposito 1999).

Whatever the final verdict is regarding the role of the various regions of the medial temporal lobe and diencephalon, the evidence indicates that the type, extent, and severity of anterograde amnesia is a function of the size, side, and location of the lesion. This rule applies as well to the deficits indicative of retrograde amnesia.

4. *Retrograde Amnesia and Memory Consolidation: Where and When Are Memories Stored?*

Whereas studies of anterograde amnesia tell us about memory acquisition, studies of retrograde amnesia provide clues about the time course involved in consolidating long-term memories and the physiological processes and neural substrates which contribute to consolidation and storage. Until recently, it was widely believed that retrograde amnesia associated with medial temporal and diencephalic damage was short-lasting and temporally graded, such that memory loss was more severe for information acquired near the time of amnesia onset than for that which was acquired long before (see Sect. 2.2 above). Accordingly, the medial temporal lobes, particularly the hippocampus, and possibly the diencephalon, were considered to be temporary memory structures, needed only for memory retention and retrieval until memories were consolidated in neocortex and other structures. They were then permanently stored there and could be retrieved directly from those regions.

Nadel and Moscovitch (1997) and Nadel et al. (2000) noted a number of problems with the accepted view. Though the duration of retrograde amnesia sometimes is short, more often retrograde amnesia for details of autobiographical events after large medial temporal (or diencepahlic) lesions can extend for decades, or even a lifetime (Warrington and Sanders 1971), far longer than it would be biologically plausible for the consolidation process to last. Retrograde amnesia for public events and personalities, however, is less extensive and often is temporally graded; this is truer still of semantic memory which includes knowledge of new vocabulary and facts about the world and ourselves (our address, the names of our friends, our job), what some have called *personal semantics* to distinguish them from autobiographical episodes (see Fig. 4). The distinction between temporally extensive and temporally limited retrograde amnesia also applies to spatial memory. Schematic cognitive maps of old neighborhoods that are adequate for navigation are retained but they lack topographical details and local environmental features, such as the appearance and location of particular homes, that would allow the person to have detailed cognitive maps of their locale (see Nadel et al. 2000).

Based on this evidence, Nadel and Moscovitch concluded, contrary to the traditional consolidation model, that the function of the medial temporal system is not temporally limited but that it is needed to represent even old memories in rich, multifaceted detail, be it autobiographical or spatial, for as long as the memory exists. Neocortical structures, on the other hand, are sufficient to form domain-specific and semantic representations based on regularities extracted from repeated experiences with words, objects, people, environments, and even of autobiographical episodes that one recollects repeatedly, creating a gist of each episode. The medial temporal lobe system may aid in the initial formation of these neocortical representations, but once formed they can exist on their own. Recent evidence from studies of children whose hippocampus was damaged at birth or shortly thereafter supports this view. Vargha-Khadem et al. (1997) found that they acquired sufficient general knowledge (semantic memories) to complete high school even though their memory for autobiographical episodes was impaired. Corroborating evidence is also provided by neuroimaging studies of recent and remote autobiographical and semantic memory. These studies found that the hippocampus is activated equally during retrieval of recent and remote autobiographical memories, but not of memory for public events or personal semantics (Ryan et al. in press, Maguire 2001) (see Fig. 5).

To account for this evidence, Nadel and Moscovitch (1997) and Nadel et al. (2000) proposed a Multiple Trace Theory (MTT) according to which a

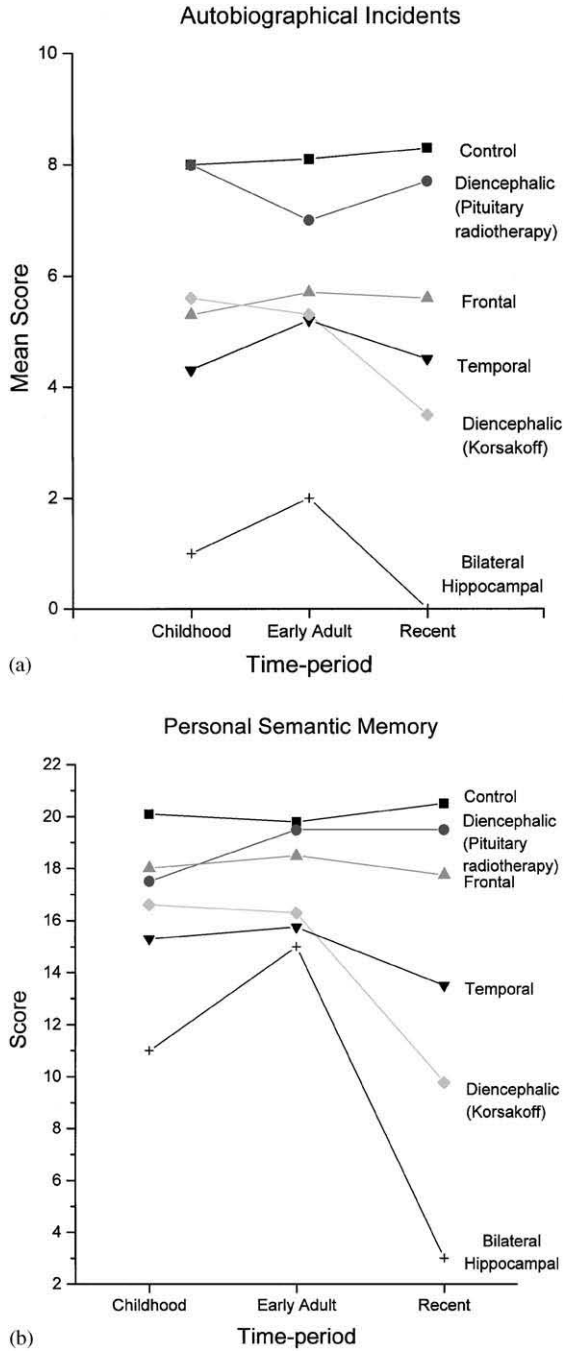

Figure 4
Example of (a) temporally-graded retrograde amnesia and (b) temporally-extended retrograde amnesia for autobiographical incidents and personal semantics in patients with bilateral medical temporal-lobe, hippocampal, and other lesions (modified from Kopelman et al. 1999 and Cipilotti et al. 2001)

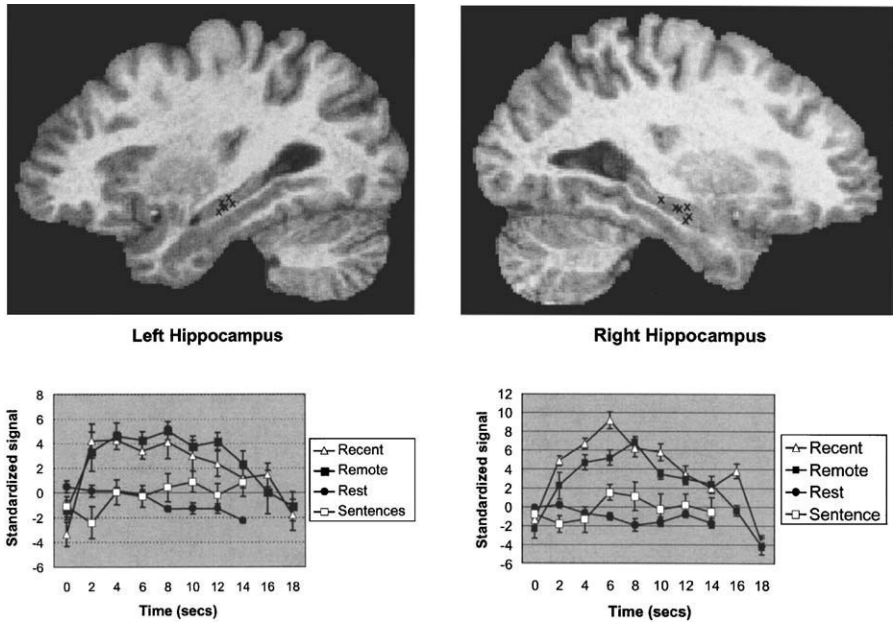

Figure 5
Hemodynamic response of the hippocampus during recall of recent and remote memories, and two baseline conditions (rest and sentence completion) (Ryan et al. in press)

memory trace of an episode consists of a bound ensemble of neocortical and hippocampal/medial temporal lobe (and possibly diencephalic) neurons which represent a memory of the consciously experienced event. Formation and consolidation of these traces, or cohesion (Moscovitch 1995), is relatively rapid, lasting on the order of seconds or at most days. Each time an old memory is retrieved, a new hippocampally mediated trace is created so that old memories are represented by more or stronger traces than are new ones, and therefore are less susceptible to disruption from brain damage than more recent ones. With respect to autobiographical episodes, the extent and severity of retrograde amnesia and perhaps the slope of the gradient are related to the amount and location of damage to the extended hippocampal complex. Remote memories for the gist of events, and for personal and public semantics, are not similarly dependent on the continuing function of the hippocampal complex (see McClelland et al. 1995 for a computational account for the usefulness of having complementary hippocampal and neocortical learning and memory systems).

Proponents of the standard consolidation model, however, argue that severe and temporally extensive retrograde amnesia is observed only when the lesion extends beyond the hippocampus to include neocortical structures where remote memories, both autobiographical and semantic, are represented (Squire and Zola 1998, but see Cipilotti et al. 2001). It remains to be determined what specific contributions the different regions of the medial temporal lobes and diencephalon make, and how they act in concert with the neocortex and other brain areas, to form and retain both detailed, contextually rich representations and context-independent knowledge (McDonald et al. 1999, Rosenbaum et al. in press).

5. *Amnesia and Neuropsychological Theories of Normal Memory: Which Types of Memory Are Affected?*

Research on amnesia and other memory disorders has influenced theories of normal memory at least since the end of the nineteenth century (Rozin 1976), but at no time has this been more apparent than in the last quarter of the twentieth century. Because amnesia is selective, affecting some memories and not others, research on amnesia has been used to promote the view that memory is not unitary, but rather consists of dissociable components, each governed by its own principles and mediated by different structures. For example, evidence showing that retrograde amnesia affects detailed autobiographical memory more than semantic memory supports the idea that episodic and semantic memory are distinct both

functionally and neurologically (Tulving 1983, Kinsbourne and Wood 1975, but see Squire and Zola 1998).

One of the characteristics of amnesia is that it affects long-term or secondary memory more than short-term or primary memory. This observation was one of the crucial pieces of evidence used in the 1960s and early 1970s to argue for a functional separation of memory into at least these two major components. The idea has since become almost universally accepted and opened the field to investigation of the functional differences between the two major components, to development of the concept of working memory (Baddeley 1986), and to identification of the mechanisms supporting working and primary memory in frontal and posterior neocortex (Smith and Jonides 1999, Petrides 2000).

Beginning in the late 1960s, research on animal models and humans indicated that the formation and retention of some types of long-term memory are spared in amnesia though there is continuing debate on how best to characterize them. Generally it is accepted that only conscious recollection is affected by amnesia. Memory retrieval without awareness seems to be intact. For example, it was noted that people with bilateral medial temporal lobe lesions, such as the patient H.M., could learn and retain motor skills for months or years, though he had no memory of the learning episode minutes after it was over. The same was true of perceptual skills involved in reading mirror-reversed words or in identifying degraded or fragmented pictures and words. That more than just a skill was involved became apparent when patients could complete or identify items to which they had been exposed more accurately and more quickly than new items, suggesting that they stored information peculiar to that item even though at a conscious level they could not recall or recognize that they had studied it (Warrington and Weiskrantz 1970) (Fig. 6). At the same time, similar phenomena were reported in normal people for material they could not consciously recollect, suggesting that the dissociation between memory with and without awareness was not peculiar to amnesia but was indicative of something fundamental about the organization of memory (see reviews in Cermak 1982, Cohen and Eichenbaum 1993, Kolb and Whishaw 1996, Tulving and Schacter 1990, Schacter and Buckner 1998, Squire and Knowlton 1999). These observations not only had a major impact on our understanding of normal memory, but also were instrumental in revitalizing research on unconscious processes in cognition and emotion, an area of investigation that had been abjured by mainstream experimental psychology for almost a century.

A number of terms have been used to refer to memory with and without awareness, including declarative and nondeclarative (or procedural memory), direct and indirect memory, memory and habit,

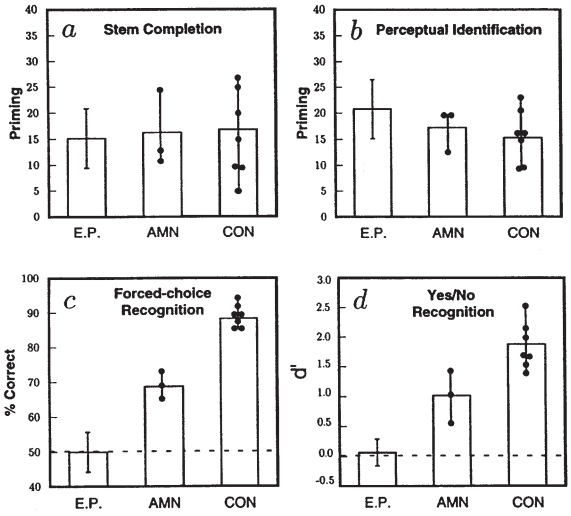

Figure 6
Example of amnesic and control performance on two implicit (stem completion and perceptual identification) and two explicit (forced choice and yes–no recognition) tests of memory (from Squire and Knowlton 1999)

controlled and automatic. We prefer the terms explicit and implicit memory because they are descriptively accurate, refer to the types of tests used to assess memory, and are close to theoretically neutral as to the processes and mechanisms involved.

Since the 1980s, research on normal people and on people with amnesia has identified a number of characteristics of implicit memory. The structures implicated in amnesia, the medial temporal lobes and diencephalon, are not needed for normal implicit memory. Instead, performance is mediated by a variety of structures, depending on the type of implicit memory that is being tested. Like explicit memory, implicit memory is not unitary and consists of a variety of different subtypes. Although a detailed description of all them is not possible, we mention three types that have been identified and about which we know a great deal: *perceptual implicit memory or priming, conceptual implicit memory, and procedural memory* (Knowlton and Squire 1999, Moscovitch 1992, Moscovitch et al. 1993, Schacter and Buckner 1998, Tulving and Schacter 1990).

On tests of perceptual implicit memory, the test stimulus resembles the studied target perceptually (e.g., a perceptually degraded version of it, or even an identical repetition of it) whereas on tests of conceptual implicit memory, the test stimulus resembles the target semantically (e.g., having studied the word 'horse,' the participant may be asked to make a semantic decision to the word at test, or asked to produce it in response to the word 'animal'). Performance is measured by speed or accuracy of

the response to the test stimulus, without explicit reference to the studied items. The implicit nature of the test is corroborated if the person is not aware that the response or test item refers to a studied stimulus. Increases in accuracy or decreases in response latency to old studied items as compared to new ones are indicative of implicit memory for the studied items.

Research on perceptual implicit memory suggests that performance is mediated by the same perceptual modules or representation systems in posterior neocortex that are involved in perceiving the stimuli. These modules are modified by the act of processing some given material, leaving behind a record of that process. As a result, processing is faster and more accurate when the system is re-exposed to that material as compared to new material (Wiggs and Martin 1998). The modules are domain specific, in that separate ones exist for objects, faces, words, and possibly places. They are also presemantic in that they do not represent the meaning of the item but only its structure. Thus, performance on perceptual implicit tests is sensitive to changes in perceptual or structural aspects of the stimulus but not to changes in semantic aspects.

The opposite holds for performance on conceptual implicit tests which is mediated by conceptual systems in the lateral temporal lobe and inferior frontal cortex (for reviews on implicit memory see Moscovitch et al. 1993, Gabrieli 1998, Schacter and Buckner 1998). As with perceptual modules, improvement in performance results from modifications in the conceptual systems themselves.

Tests of procedural implicit memory involve learning rules, motor sequences, conditioned responses, and subjective probabilities of stimulus–response associations without explicit memory for any of them. The structures that have been identified as crucial for procedural learning are the cerebellum for classical conditioning, and the basal ganglia for the others, with some indication of prefrontal involvement if learning rules or motor sequences involve strategic, sequential, or inhibitory components (Moscovitch et al. 1993, Squire and Knowlton 1999). Changes in these structures during execution of procedures underlies the changes in performance observed on implicit tests of procedural memory.

In studying implicit memory, great care must be taken to insure that performance on tests of memory that are ostensibly implicit are not contaminated by explicit components, such as might occur when asking participants to identify degraded stimuli they had studied earlier. It is for this reason that studies of amnesic patients is so useful. Because their explicit memory is so poor, equivalent performance between amnesic and normal people on an implicit test is taken as evidence that the test in question was not contaminated by explicit memory.

6. *Amnesia and Beyond: A Neuropsychological Component Process Model of Memory*

Studies of memory in amnesia have indicated that memory is not unitary but rather consists of a variety of different forms, each mediated by different component processes which in turn are subserved by different neural mechanisms. Because neither short-term or working memory, nor implicit memory is impaired in amnesia, it can be concluded that these types of memory are not dependent on the medial temporal lobes and diencephalic structures which are damaged in amnesia. The latter structures, however, are necessary for conscious recollection of long-term, episodic memories. It has been proposed that any information that is consciously experienced is picked up obligatorily and automatically by the hippocampus and related structures in the medial temporal lobes and diencephalon. These structures bind into a memory trace those neural elements in neocortex (and elsewhere) that mediate the conscious experience of an event. The episodic memory trace thus consists of an ensemble of hippocampal and neocortical neurons. The hippocampal component of the trace acts as an index or file entry pointing to the neural elements in neocortex that represent both the content of the event and the conscious experience of it. 'Consciousness' is, therefore, part of the memory trace. Retrieval occurs when an external or internally generated cue triggers the hippocampal index which in turn activates the entire neocortical ensemble associated with it. In this way we recover not only the content of an event but the consciousness that accompanied our experience of it. In short, when we recover episodic memories, we recover conscious experiences (Moscovitch 1995).

According to this model both encoding and retrieval of consciously apprehended information via the hippocampus and related structures is obligatory and automatic, yet we know from experience and from experimental investigation that we have a measure of control over what we encode and what we retrieve from memory. Moreover, if encoding is automatic and obligatory, the information cannot be organized, yet memory appears to have some temporal and thematic organization. How can we reconcile this model of memory with other facts we know about how memory works? One solution is that other structures, particularly those in the frontal lobes, control the information delivered to the medial temporal and diencephalic system at encoding, initiate and guide retrieval, and monitor, and help interpret and organize the information that is retrieved. By operating on the medial temporal and diencephalic system, the frontal lobes act as *working-with-memory* structures that control the more reflexive medial temporal and diencephalic system and confer a measure of intelligence and direction to it (Fig. 7). Such a complementary system is needed if

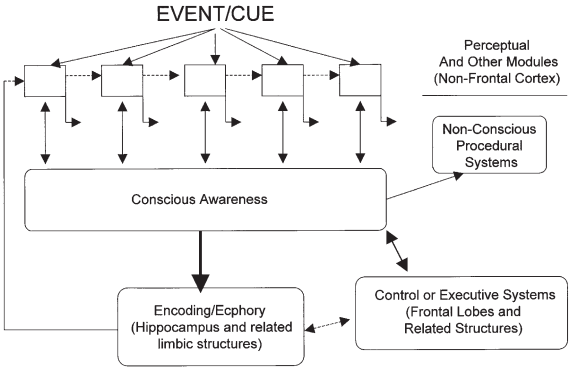

Figure 7
A schematic model of hippocampal complex-neocortical-frontal interaction during encoding and retrieval (from Moscovitch 1989, See text)

memory is to serve functions other than mere retention and retrieval of past experiences (Moscovitch 1992). As invaluable as studies of amnesia have been to our understanding of memory, those studies need to be supplemented by investigations of memory in people with other types of disorders, particularly those implicating the frontal lobes, to have a full appreciation of how memory works.

See also: Amnesia: Transient and Psychogenic; Declarative Memory, Neural Basis of; Dementia: Overview; Dementia, Semantic; Implicit Learning and Memory: Psychological and Neural Aspects; Learning and Memory, Neural Basis of; Memory, Consolidation of; Technology-supported Learning Environments

Bibliography

Aggleton J P, Brown M W 1999 Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral Brain Science* **22**: 425–89

Agguire G K, D'Esposito M 1999 Topographical disorientation: A synthesis and taxonomy. *Brain* **122**: 1613–28

Baddeley A 1986 *Working Memory*. Oxford University Press, Oxford, UK

Cermak L S (ed.) 1982 *Human Memory and Amnesia*, Erlbaum, Hillsdale, NJ

Cipilotti L, Shallice T, Chan D, Fox N, Scahill R, Harrison G, Stevens J, Rudge P 2001 Long-term retrograde amnesia … the crucial role of the hippocampus. *Neuropsychologia* **39**: 151–72

Cohen N J, Eichenbaum H 1993 *Memory, Amnesia and the Hippocampal System*. MIT Press, Cambridge, MA

Corkin S, Amaral D G, Gonzalez R G, Johnson K A, Hyman B T H.M.'s medial temporal lobe lesion: Findings from magnetic resonance imaging. *The Journal of Neuroscience* **17**: 3964–79

Eichenbaum H 1999 The hippocampus and mechanisms of declarative memory. *Behavioral Brain Research* **103**: 123–33

Gabrieli J D E 1998 Cognitive neuroscience of human memory. *Annual Review of Psychology* **49**: 87–115

Hamilton C W 1976 *Basic Limbic Anatomy of the Rat*. Plenum, New York and London

Henke K, Weber B, Kneifel S, Wieser H G, Buck A 1999 The hippocampus associates information in memory. *Proceedings of the National Academy of Science USA* **96**: 5884–9

Kapur N 1999 Syndromes of retrograde amnesia: A conceptual and empirical analysis. *Psychological Bulletin* **125**: 800–25

Kinsbourne M, Wood F 1975 Short-term memory processes and the amnesic syndrome. In: Deutsch D, Deutsch A J (eds.) *Short-term Memory*, Academic Press, New York

Kolb B, Whishaw I Q 1996 *Fundamentals of Human Neuropsychology*, 4th edn. Freeman, New York

Kopelman M D, Stanhope N, Kingsley D 1999 Retrograde amnesia in patients with diencephalic, temporal lobe or frontal lesions. *Neuropsychologia* **37**: 939–58

Korsakoff S S 1889 Etudes medico psychologique sur une forme du maladi de la mémoire. *Review Philosoph* **28**: 501–30 (trans. and republished by Victor M, Yakovlev P I 1955 *Neurology* **5**: 394–406)

Maguire E A (in press) Neuroimaging studies of autobiographical event memory. *Philosophical Transactions of the Royal Society of London Series B–Biological Sciences*

Maguire E A, Burgess N, Donnett J G, Frackowiack R S J, Frith C D, O'Keefe J 1998 Knowing where and getting there: A human navigation network. *Science* **280**: 921–4

McClelland J L, McNaughton B L, O'Reilly R C 1995 Why are there complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–57

McDonald R M, Ergis A-M, Winocur G 1999 Functional dissociation of brain regions in learning and memory: Evidence for multiple systems. In: Foster J K, Jelicic M (eds.) *Memory: Systems, Process, or Function*. Oxford University Press, Oxford

Milner B 1966 Amnesia following operation on the temporal lobe. In: Whitty C W M, Zangwill O L (eds.) *Amnesia*, Butterworth, London

Milner B, Johnsrude I, Crane J 1997 Right temporal-lobe contribution to object-location memory. *Philosophical Transactions of the Royal Society of London Series B* **352**: 1469–74

Moscovitch M 1992 Memory and working with memory: A component process model based on modules and central systems. *Journal of Cognitive Neuroscience* **4**: 257–67

Moscovitch M 1995 Recovered consciousness: A hypothesis concerning modularity and episodic memory. *Journal of Clinical Experimental Neuropsychology* **17**: 276–91

Moscovitch M, Vriezen E, Goshen-Gottstein Y 1993 Implicit tests of memory in patients with focal lesions or degenerative brain disorders. In: Boller F, Spinnler H (eds.) *The Handbook of Neuropsychology*, Vol. 8. Elsevier, Amsterdam, The Netherlands pp. 133–73

Moscovitch M, Winocur G 1992 The neuropsychology of memory and aging. In: Craik F I M, Salthouse T A (eds.) *The Handbook of Aging and Cognition*. Erlbaum, Hillsdale, NJ, pp. 315–72

Murray E A 1996 What have ablation studies told us about the neural substrates of stimulus memory? *Seminars in Neuroscience* **8**: 13–22

Nadel L, Moscovitch M 1997 Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinions in Neurobiology* **7**: 217–27

Nadel L, Samsonovich A, Ryan L, Moscovitch M 2000 Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results. *Hippocampus* **10**: 352–68

O'Keefe J, Nadel L 1978 *The Hippocampus as a Cognitive Map*. Oxford University Press, Oxford, UK

Petricles M 2000 The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research* **133**: 44–54

Rosenbaum R S, Winocur G, Moscovitch M (in press) New views on old memories: Reevaluating the role of the hippocampal complex. *Experimental Brain Research*

Rozin 1976 The psychobiological approach to human memory. In: Rozenzweig R Bennett E L (eds.) *Neural Mechanisms of Learning and Memory*. MIT Press, Cambridge, MA

Ryan L, Nadel L, Keil K, Putnam K, Schnyer D, Trouard D, Moscovitch M (in press) The hippocampal complex and retrieval of recent and very remote autobiographical memories: Evidence from functional magnetic resonance imaging in neurologically intact people. *Hippocampus*

Schacter D L, Buckner R L 1998 Priming and the brain. *Neuron* **20**: 185–95

Scoville W B, Milner B 1957 Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry* **20**: 11–21

Smith E E, Jonides J 1999 Storage and executive processes of the frontal lobes. *Science* **283**: 1657–61

Squire L R, Alvarez P 1995 Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinions in Neurobiology* **5**: 169–77

Squire L R, Knowlton B J 1999 The medial temporal lobe, the hippocampus and the memory systems of the brain. In: Gazzuniga M S (ed.) *The Cognitive Neurosciences*, 2nd edn., MIT Press, Cambridge, MA, pp. 765–80

Squire L R, Zola S M 1998 Episodic memory, semantic memory, and amnesia. *Hippocampus* **8**: 205–11

Tulving E 1983 *Elements of Episodic Memory*. Clarendon Press, Oxford, UK

Tulving E, Schacter D L 1990 Priming and human memory systems. *Science* **247**: 301–6

Vargha-Khadem F, Gadian D G, Watkins K E, Conneley A, Van Paesschen W, Mishkin M 1997 Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**: 376–80

Warrington E K, Sanders H I 1971 The fate of old memories. *Quarterly Journal of Experimental Psychology* **23**: 432–42

Warrington E K, Weiskrantz L 1970 Amnesic syndrome: Consolidation or retrieval? *Nature* **228**: 628–30

Wiggs C L, Martin A 1998 Properties and mechanisms of perceptual priming. *Current Opinions in Neurobiology* **8**: 227–33

M. Moscovitch

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

B

Baby Booms and Baby Busts in the Twentieth Century

Baby booms and busts are not a phenomenon peculiar to the twentieth century (Easterlin 1961). There appear to have been marked cycles in fertility rates and population growth in earlier eras (see *Population Cycles and Demographic Behavior*), and in virtually every immediate postwar period there has been a compensatory 'postwar baby boom' that resulted when military personnel were reunited with their families. But the baby boom experienced by the industrialized nations after World War II is unique for at least two reasons. It occurred after a prolonged secular decline in fertility (see *Fertility Transition: Economic Explanations*), when rates were already at or below replacement level and were expected to remain low indefinitely—and yet it was too prolonged and extreme to be merely a postwar 'catching up' phenomenon; and it was much more widespread than any previous boom.

It is generally assumed that there have been two significant baby busts in the twentieth century (again, apart from wartime blips), one in the 1930s and the other in the last quarter of the century. As with the baby boom these constitute a departure from past experience in that fertility rates on both occasions fell below replacement level in a majority of countries. However, it remains to be seen whether the most recent experience constitutes a true baby bust (implying eventual 'recovery' to higher levels) or simply normal post-transition behavior (see *Demographic Transition, Second*).

Figure 1 illustrates the general pattern of fertility in the industrialized nations (except Japan: see *Fertility Transition*: *East Asia*) since around 1850 using as examples the experience in England and Wales and the USA. Depicted in Fig. 1 is the Total Fertility Rate (TFR): a synthetic measure representing the number of children a woman would have in a lifetime spent at current age-specific fertility rates. Although this synthetic measure gives an accurate picture of aggregate fertility behavior at any given point in time, it is not necessarily a good measure of changes in completed family size for cohorts of women (see *Fertility Change: Quantum and Tempo*).

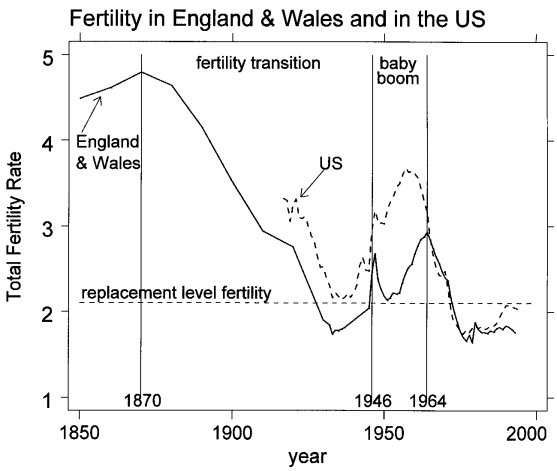

Figure 1
Using the USA and England and Wales to illustrate the general pattern of fertility in the industrialized nations in the twentieth century. Source: Coleman (1996) and Glass (1967) for England and Wales; various official sources for the USA

1. Historical Patterns

The post-World War II baby boom tends to be defined as having occurred between 1946 and 1964, since the immediate postwar booms generally occurred in the 1946–48 period and the subsequent prolonged baby booms in virtually all nations had reached their peaks by 1964. But there is a wide diversity in the patterns exhibited by the baby booms in different countries, as illustrated in Fig. 2. (For coverage of the Eastern European nations, see *Fertility Trends in the Formerly Socialist Countries of Europe*.)

Although rates have been fairly stable since the boom and bust cycle, interest has been aroused by signs of a possible resurgence in several countries since about 1985. This is most apparent in Sweden, Denmark, Luxembourg, and Iceland, with increases of 25 to 35 percent since around 1990—but also appears in Finland, Norway, North America, New Zealand, and West Germany, with increases of around 15 percent in the same period. There has been some tendency to dismiss these changes as simply the result of 'catch-up' fertility among women who postponed childbirth at younger ages, but changes in younger women's fertility rates have contributed to these mini-booms, as well (based on age-specific fertility rates in Coleman 1999).

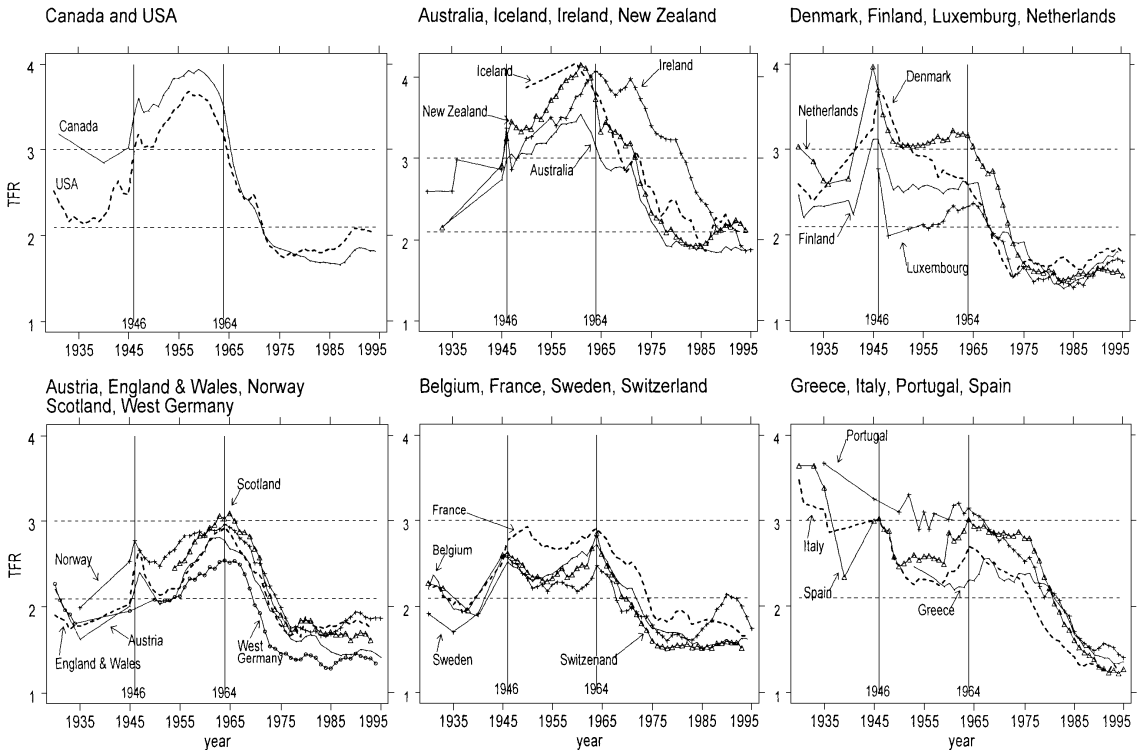

Figure 2
The Baby Boom was most pronounced in North America, Australia, New Zealand, Iceland, and Ireland, where the TFR increased by 80 percent or more and exceeded 3.5 for an extended period. It was nearly as strong, in percentage terms, in England and Wales, Scotland and Austria, where the TFR rose by about 70 percent from its prewar low. The boom was least apparent in Denmark, where the TFR increased by only 20 percent—but in the rest of northern and western Europe the increase was 40 percent or more. Although strong increases were felt in most of the English-speaking countries in the early 1950s immediately following the postwar spike, the TFR in Europe did not begin to increase again until the later 1950s–1960 in the southern European nations. Similarly, the baby boom first peaked in North America in 1957, and then in Australia and New Zealand in 1961, followed by virtually all of Europe in 1964. Even the pattern of fertility decline after the peak differed markedly among countries. It occurred rapidly in North America and northern and western Europe (apart from Ireland), with only a minor hesitation in the late 1960s, so that rates had bottomed out in those regions by about 1975. But the hesitation in the late 1960s was much more pronounced in Australia and New Zealand, where the TFR did not bottom out until about 1980. And in Ireland and Southern Europe, despite an initial peak in 1964, a pronounced fertility decline did not occur until the late 1970s, continuing through the late 1980s. Source: Coleman (1996) supplemented by Coleman (1999), Keyfitz and Flieger (1968), Glass (1967), Lee (1979) and official sources for the USA

2. *Theories: 'Price of Time'*

Two competing, but to some extent complementary theories, developed largely with respect to US experience, are generally referred to in explaining the baby boom and bust in the second half of the twentieth century. One—the 'price of time' model—juxtaposes the desire for children, which is assumed to be positively related to family income, with the price of time spent in caring for children, and emphasizes the importance of women's labor force participation and their wages relative to men's in determining that price (Becker 1981, Butz and Ward 1979; see *Fertility Transition: Economic Explanations*). It is hypothesized that during the postwar 1940s and 1950s men's wages rose more rapidly than women's, as women were displaced by men returning from the military, so that the price of children fell relative to the families' ability to support them, encouraging a baby boom. This situation was assumed to reverse itself in the late 1960s and into the 1970s as labor market opportunities for women increased and pushed up their wage, leading to

an increase in the relative price of children and hence the baby bust.

There have been critiques of the econometrics and data employed in the analysis most closely associated with this theory (Macunovich 1995), demonstrating that the true pattern of women's wages in the USA does not support the theory during the baby bust. However, several studies—especially those by Ermisch using European and Japanese data (see his summary in Ermisch 1996)—suggest that the theory is supported by the pattern of women's wages when viewed relative to that of men's.

3. Theories: 'Relative Cohort Size'

The competing theory, usually referred to as the Easterlin or 'relative cohort size' hypothesis, also assumes a positive relationship between the desire for children and family income but juxtaposes this with a couple's material aspirations, which are assumed to be influenced strongly by the standard of living experienced by young adults when they were growing up. A couple feels able to afford children only if family income surpasses some threshold determined by material aspirations. It is hypothesized that young adults in the 1950s, who were born and raised in the Depression and war years, set a lower threshold on average than young adults in the 1960s and 1970s, who were raised in the affluent postwar years. Compounding this effect of changing tastes, young adults in the 1950s were members of a very small birth cohort (product of the 1930s baby bust) relative to the size of the rest of the labor force, so that their wages were driven up relative to those of older workers in their parents' generation (Macunovich 1999, Welch 1979). The result was higher wages relative to their own (already low) material aspirations, making children appear very affordable. The large baby boom cohorts had the opposite experience when they entered the labor force, leaving them with reduced wages relative to inflated aspirations, and the baby bust resulted.

4. Combined Theories

Although the Easterlin model appeared to fit North American postwar data extremely well, cross-country tests of the hypothesis turned up fairly mixed results. It tended to fall out of favor when fertility failed to increase in the early 1980s among members of the baby bust cohorts (see Macunovich 1999 for a comprehensive review). Pampel (1993) argues that the Easterlin effect is real but is contingent on institutional differences not controlled for in cross-country analyses to date.

A recent set of analyses suggest that a model combining these two theories can be used to explain the path of fertility in the USA both before and after 1980 (Macunovich 1996, 1998b, forthcoming). In addition, this work demonstrates that while fertility has not followed the path set by relative cohort size since 1980, it has conformed with movements in a measure designed to capture Easterlin's relative income concept. It remains to be seen whether this close conformity between fertility and relative income can explain in other countries, as it does in the USA both the upturn in fertility that began after 1985 and its attenuation in the 1990s.

See also: Children, Value of; Family as Institution; Family Size Preferences; Fertility Change: Quantum and Tempo; Fertility: Institutional and Political Approaches; Health in Developing Countries: Cultural Concerns; Mortality, Biodemography of; Mortality: The Great Historical Decline; Multistate Transition Models in Demography; Population Cycles and Demographic Behavior; Population Cycles, Formal Theory of; Population Dynamics: Theory of Stable Populations; Public Health

Bibliography

Becker G 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Butz W P, Ward M P 1979 The emergence of countercyclical US fertility. *American Economic Review* **69**(3): 318–28

Coleman D 1996 New patterns and trends in European fertility: International and sub-national comparisons. In: Coleman D (ed.) *Europe's Population in the 1990s*. Oxford University Press, Oxford, UK pp. 1–61

Coleman D 1999 Zip Files of Total and Age-Specific Fertility Rates made available as part of the *Oxford Population Project* at http://marx.apsoc.ox.ac.uk/oxpop/

Easterlin R A 1961 The American baby boom in historical perspective. *American Economic Review*, December, **LI(5)**: 869–911, and reprinted (with an additional appendix) as *Occasional Paper* **79** (1962), NBER

Easterlin R A 1987 *Birth and Fortune*, 2nd edn. University of Chicago Press, Chicago

Ermisch J F 1996 The economic environment for family formation. In: Coleman D (ed.) *Europe's Population in the 1990s*. Oxford University Press, Oxford, UK pp. 144–62

Glass D V 1967 *Population: Policies and Movements in Europe*. Frank Cass and Co., London

Keyfitz N, Flieger W 1968 *World Population: An Analysis of Vital Data*. The University of Chicago Press, Chicago

Lee W R 1979 *European Demography and Economic Growth*. St. Martin's Press, New York

Macunovich D J 1995 The Butz-Ward fertility model in the light of more recent data. *Journal of Human Resources* **30**(2, Spring): 229–55

Macunovich D J 1996 Relative income and price of time: Exploring their effects on US fertility and female labor force participation. *Fertility in the United States: New Patterns, New Theories, Population and Development Review*, supplement to Volume **22**: 223–57

Macunovich D J 1998a Fertility and the Easterlin hypothesis: An assessment of the literature. *Journal of Population Economics* **11**: 53–111

Macunovich D J 1998b Race and relative income/Price of time effects on US fertility. *Journal of Socio-Economics* **27**(3): 365–400

Macunovich D J 1999 The fortunes of one's birth: Relative cohort size and the youth labor market in the US. *Journal of Population Economics* **12**(2): 215–72

Macunovich D J 2000 Relative cohort size: source of a unifying theory of global fertility transition? *Population and Development Review* **26**(2): 235–61

Pampel F C 1993 Relative cohort size and fertility: The socio-political context of the Easterlin effect. *American Sociological Review* **58**(4): 496–514

Welch F 1979 Effects of cohort size on earnings: The baby boom babies' financial bust. *Journal of Political Economy* **87**: S65–S97

D. J. Macunovich

Balance of Power, History of

The concept 'balance of power' can be confusing. It has two different, though interrelated, meanings. It can refer first to a situation in which the Great Powers of a particular figuration of states—whether regional or global—possess roughly equal power resources. A power equilibrium is unlikely; the differences must just be small enough to make the risk of going to war too high. To achieve hegemony or empire a highly uneven power balance is required.

A balance of power can be bilateral as well as multilateral. It was considered crucial in the hegemonial rivalry between the Soviet Union and the United States. The rivals saw arms competition as the primary indicator of the balance between them. However, by balance they in fact meant superiority. The unintended result was stalemate. The nuclear danger, at the time called a 'balance of terror,' chained them together and made 'superiority' meaningless.

Balance of power in the first sense thus refers to the state of the power balance between interdependent, rivaling states or their predecessors. The first theory about the power balances in competing kingdoms was probably formulated by Kautilya, around 450 BC in India. When states were first formed in Europe, their rivalry was guided by the maxims of 'reason of state.' But gradually their perception of common interests became stronger, as reflected in the development of diplomacy and international law and the creation of international organizations.

Balance of power in its second meaning was part of the process in which cooperation between Great Powers increased without removing their rivalry. It came to mean a policy to prevent a bid for hegemony of one of the Great Powers, by the threat of the other Great Powers allying against it. If deterrence of a hegemonist failed the alliance could become a war coalition to restore the balance. But the purpose of balance of power policy itself was to maintain the status quo. To succeed, the power differentials between the Great Powers should remain small enough to foreclose hegemonic aspirations. The first meaning of balance of power was therefore a condition for the second.

1. Background

How and why did balance of power policy emerge? The European peace conference in 1648 to end the many conflicts comprising the Thirty- and Eighty-Years wars was not yet a joint attempt to develop a more stable balance of power in Europe. But the conference brought the representatives of the hitherto separate northern and western power balances together. Though the participating states in Europe recognized one another as such by treaty, repudiating the prior claims to supreme authority of empire and church, dynastic rivalry remained. As Dame Wedgwood remarks about the long delay of actual negotiations in Munster: 'The ruling powers … asked for peace always in a general sense: when it came to practical action they were always prepared to fight for a little longer in order to gain their own particular end' (Wedgwood 1963, p. 474). The peace conference became possible after expansionist ends could no longer be reached. Popular sentiment against war, very strong in the devastated German lands, did not make much difference however. War in the dynastic period was characterized by Kant as a 'pleasure party' for princes, who themselves did not suffer from them as the people did. For durable peace, dynastic states should be replaced by republics, by their nature responsive to the people, and willing to form a commonwealth together. Indeed, when the power distribution between dynastic states was still fluid and difficult to assess as in the sixteenth and seventeenth centuries, wars were frequent.

The peace of Westphalia simultaneously ended a plurality of conflicts. It made France more powerful and reduced Habsburg power. It also contributed to the rise of new Great Powers like Russia and Prussia and to the fall of Spain, Sweden, and later the Netherlands. One European power balance came into being. The monopolization processes that had led to the forming of dynastic states out of feudal fragmentation continued in the rise of Great Powers.

In the larger dynastic states, the increase of the king's power and authority had to be at the expense of other centers of power, such as the warrior nobility and autonomous towns. The domestic consolidation of dynastic regimes made stability in interstate relations more desirable. Balance of power policy came in useful. But it was precarious. Power resources changed all the time, so attempts at acquiring hegemony and war between Great Powers could not be excluded.

2. *The Classical Balance of Power in Europe*

Balance of power policy was practiced before it became an explicit principle. The power of Louis XIV's France grew to such an extent after the Peace of Westphalia that the other European powers feared it might acquire hegemony. A French attempt to use the Spanish succession to increase its power aroused Britain into forming in 1701 in The Hague an alliance of the other Great Powers against France. This coalition won the war and the Treaty of Utrecht of 1713 considerably reduced the power of France. That treaty also readjusted power relations all over Europe, and France lost some colonies to Britain. But it followed the prescription of balance of power policy that one's adversaries should not be destroyed, as that would upset the power balance between the allies.

In the eighteenth century five Great Powers emerged that would remain until 1914. They included one global and colonial sea power (Britain), two continental powers (Prussia and Austria), one power with both continental and maritime colonial ambitions (France), and an imperial power, with its territory stretching beyond Europe (Russia). Though the Ottoman Empire was in practice part of the European power balance, it was not a member of the quintet. Though the power resources of the five Great Powers were quite different, they were not too widely divergent most of the time to make balance of power policy impossible.

What exactly was classical balance of power policy? As the pamphlet, *Europe's Catechism* (1741) defines it:

> The Balance of Power is such an equal Distribution of Power among the Princes of Europe, as makes it impracticable for the one to disturb the Repose of the other … It may be destroyed by Force or Fraud, by the Pusillanimity of some, and the Corruption of all … When any Potentate hath arrived to an exorbitant share of power, the Rest (ought) to league together and reduce him to his due Share of it. Otherwise there is but one Potentate, and the others are only a kind of Vassals.

The *Catechism* shows that balance of power policy was well understood at the time. It clearly describes its functions: to prevent hegemony includes maintaining the 'Repose' of the participating powers, implying that they were no longer interested in expansionist adventures and had become status quo powers.

Balance of power policy requires that the Great Powers are able and willing to shift their alliances to deter hegemonic aspirations. That goal can be supported by a balancer. Britain, as a global sea power, could fulfill this role by always aligning itself against the leading power on the continent.

In the long run, however, balance of power policy can only be successful if the power balance between the Great Powers does not change too drastically. It still allows for cooperative expansion, as the three divisions of Poland between Russia, Austria, and Prussia in the late eighteenth century make clear. Conquest of Poland by one of them would have disturbed the balance of power.

3. *The Congress of Vienna and the Concert of Europe*

Balance of power policy did not always deter all Great Powers. Revolutionary France no longer respected the dynastic order in Europe. First, the Jacobin regime and then Napoleon set out on conquest, supported by the new power of popular conscript and volunteer armies, motivated by nationalism. Then balance of power policy had to change into a war coalition to undo the conquests of the hegemonist and restore the balance. As France occupied most of the continent, this took time. The decisive turn in the war came when Napoleon overextended his power and failed to subdue Russia.

The main members of the war coalition (Russia, Britain, Prussia, and Austria) expressed their war aims as 'the restoration of a just balance of power and a redivision of their respective forces suitable to assure this equilibrium' (Gulick 1955, p. 128). The Congress of Vienna followed the rules of balance of power policy: France was readmitted as one of the Great Powers and as member of the Conference. As part of the postwar power balance the newly created Kingdom of the Netherlands—being in between Britain, France, and Prussia—was strengthened by the inclusion of the Southern Netherlands in the Kingdom.

With 'A World Restored' (Kissinger 1957) in Vienna, a relatively stable period began again. No great war occurred until 1914. Still, in the nineteenth century maintaining the balance went together with more intense competition, whether for military and industrial power resources or for colonial possessions. There were a few wars between a limited number of Great Powers, such as the Crimean war. Yet, rulers recognized a common interest in preserving a European balance of power.

4. *The Concert of Europe and its Global Successors*

Balance of power policy was extended at the Conference of Vienna by the balancer Britain. In 1815 its Foreign Secretary Castlereagh drafted Article VI of the Quadruple Alliance Treaty (to which France later acceded), which provided for periodic meetings between the Great Powers: 'for the consideration of the measures which shall be considered the most salutory for the repose and prosperity of nations and the peace of Europe' (Kissinger 1957, p. 186).

This 'conference system' broke down in 1822 when Castlereagh's successor Canning refused to support France's proposal for joint armed intervention in

Spain. But the idea of a Concert of Europe to prevent and solve conflicts did not disappear. In 1855, Austria called a conference to prevent a war between Russia and Turkey. But it failed and the Crimean war broke out. That was ended by the Conference of Paris (1856), at which the British Foreign Secretary Lord Clarendon tried to strengthen the role of conference diplomacy by proposing the rule that states would seek 'good offices' before going to war. Two other important conferences were held in Berlin (1878 and 1884–5), both at the initiative of Bismarck. In the first, he acted with success as 'honest broker' between Russia, Britain, Austria-Hungary, and the Ottoman empire to settle conflicts over the Balkan and in the Near East, resulting from the decline of the Ottoman empire. The most important result of the second was that the Great Powers agreed to give control over the Congo to King Leopold of Belgium. As with the Netherlands in 1815, balance of power policy required that the Congo would not be controlled by one of the Great Powers, no matter what that would mean for the Congo itself.

The idea of a Concert of Europe went further than balance of power policy. It could imply collective Great Power domination. But it could not prevent a second complete breakdown of the balance of power. The rivalry for colonial possessions, symbolized by the scramble for Africa, became closed toward the end of the century. The world had been divided. Latecomers such as Germany and Italy could not be satisfied. At the same time, the power balance in Europe became more uneven. After three wars, (with successively Denmark, Austria, and France) Prussia made itself the nucleus of a powerful new German empire. Its first Chancellor Bismarck still thought in terms the balance of power. But after he left in 1890, Emperor William II began to harbor hegemonic ambitions. The other Great Powers did not form an anti-hegemonic alliance against Germany's aspirations and growing capability. On the contrary, Austria-Hungary sided with Germany, whereas Britain, France, and Russia formed a Triple Alliance. Two inflexible alliances stood against each other. Czar Nicholas' attempt to revive the conference idea in the form of general peace conferences, held in The Hague in 1899 and 1907, was not successful. The outbreak of World War I in July 1914 demonstrated again that balance of power policy was precarious.

Still, the winning alliance of World War I depended on the balancing effect of the entry of the United States. But restoring a balance of power was not a stated purpose of the postwar conference of Versailles. President Wilson believed that balance of power policy was responsible for the continual scourge of war in Europe. He looked for a lasting basis for peaceful conduct of states in collective security and national self-determination. Collective security was based on all members of the League of Nations uniting against an aggressor. But it proved impotent even against Italy, itself not a Great Power. Wilson's critical attitude towards balance of power policy was representative of the idealist thinking about international politics that was influential after the horrors of World War I. After the even worse horrors of World War II its counterpart realism, stressing the importance of power and military preparedness, became dominant again.

In the Security Council of the UN, the idea of a Concert of Great Powers became institutionalized. American President Roosevelt, who had again balanced the hegemony of Germany and its allies, saw the role of Great Powers in the Security Council as 'international sheriffs.' But a permanent Conference embodied in an international organization could not solve the problem of Great Power competition either. Hegemonic rivalry started again after a few years between the United States and the Soviet Union, thereby paralyzing the Security Council. Their rivalry did not lead to another great war primarily because of the consequences of the shared danger of a nuclear war. After the peaceful end of the hegemonic rivalry and of the Soviet Union, consensus for a time reigned in the Security Council, making a mandate for the use of force against Iraq possible. But China abstained. Rivalry again sets the tone in the Security Council.

Development beyond balance of power policy has thus not come very far. It has not been replaced by international organization. The dynamic of rivalry between states and especially Great Powers has not fundamentally changed, even though their domestic development now requires more cooperation.

5. *Balance of Power in other Parts of the World*

Power balances have been part of the development of state systems everywhere, so have wars as part of hegemonic struggles. Thinkers about interstate relations such as Kautilya in India or the Legalists in the China of the Warring States were, like Machiavelli, advisors to princes, primarily interested in increasing the power of their own state rather than in a balance of power. That was also the case with dynastic states in Europe. But in Europe balance of power policy replaced reason of state as the dominant policy orientation. Why? The explanation may be found:

(a) in the relatively even Great Power balance that developed at the end of the seventeenth century;

(b) in the importance of 'Repose' for dynastic regimes;

(c) in colonial expansion (and Russia's frontier in Asia) as a substitute for territorial expansion within Europe; and

(d) in the increasing repugnance for war in established states, though in the nineteenth century it could be replaced by nationalism.

The European balance of power became global in the twentieth century. The bipolar balance between the two remaining Great Powers after 1945 was

perceived by both as covering the world as a whole. The smallest power advantage obtained by the one had to be compensated by the other.

In the post Cold War world regional balances became more important. In South East Asia, for example, the members of ASEAN had transformed their economic cooperation in a security alliance directed to what they saw as the hegemonic ambitions of Vietnam, demonstrated by Vietnam's intervention and occupation of Cambodia. Recently Vietnam joined ASEAN because of an unspoken common interest in balancing China's possible bid for hegemony in the region. American foreign policy has been constantly concerned with regional balances of power, in Europe, but also in East Asia and the Middle East.

6. *Conclusion: A Permanent Feature of International Politics?*

Since the Congress of Vienna all attempts to create an international order were based on an extension of balance of power policy. The exception was the League of Nations. But its impotence can be explained by the neglect of balance of power considerations and overestimation of the effectiveness of the all against one requirement of punishing an aggressor, especially if that were a Great Power.

After 1991, the United States was the only global Great Power left. It is not likely that it will remain so indefinitely. Potential challengers, such as China or the European Union, are already there, even though still much too weak. Will balance of power policy become relevant again in the future? As long as a monopoly of violence at the international level does not come into being, balance of power policy may well remain relevant.

See also: First World War, The; International Law and Treaties; International Organization; International Relations, History of; Military History; Peacemaking in History; Second World War, The; Warfare in History

Bibliography

Carr E H 1946 *The Twenty Years' Crisis 1919–1939*. Macmillan, London

Elias N 1982 *The Civilising Process, Vol. II: State Formation and Civilization*. Basil Blackwell, Oxford

Gulick E V 1955 *Europe's Classical Balance of Power: A Case History of One of the Great Concepts of European Statecraft*. Cornell University Press, Ithaca, NY

Kant I 1957 *Perpetual Peace*. Bobbs Merril, New York

Kautilya 1992 *The Arthasastra*. Rangarajan L N (ed.) Penguin Books, New Delhi, India

Kissinger H A 1957 *A World Restored: Metternich, Castlereagh and the Problems of Peace*. Houghton Mifflin, Boston, MA

Meinecke F 1929 *Die Idee der Staatsraison in der neueren Geschichte*. Oldenbourg, Munich, Germany

Nicholson H 1947 *The Congress of Vienna*. Phoenix Publishing, Bern, Switzerland

van Benthem van den Bergh G 1992 *The Nuclear Revolution and the End of the Cold War: Forced Restraint*. Macmillan, Houndmills, Basingstoke, UK

Vincent R J, Wright M (eds.) Special issue on the balance of power. *Review of International Studies* **15**(2)

Watson A 1992 *The Evolution of International Society: A Comparative Historical Analysis*. Routledge, London

Wedgwood C V 1963 *The Thirty Years War*. Routledge, London

G. van Benthem van den Bergh

Balance of Power: Political

As Kenneth Waltz has noted, 'If there is any distinctively political theory of international politics, balance-of-power theory is it. And yet one cannot find a statement of the theory that is generally accepted' (Waltz 1979, p. 117; for surveys of the meaning of balance of power see Claude 1962, Haas 1953, Wight 1968, 1973). But cutting through the welter of possible meanings and making a few simple and undemanding assumptions leads to a conception that explains a number of outcomes which, while familiar, cannot otherwise be readily explained: no state has come to dominate the international system; few wars are total; losers rarely are divided up at the end of the war and indeed are reintegrated into the international system; small states, who do not have the resources to protect themselves, usually survive. There is then a form of stability in international politics. Although the fates of individual units rise and fall, states and much of the pattern of their interaction remain. The system is never transformed from an anarchical into a hierarchical one. (Note that this says nothing about whether wars are more likely when power is evenly distributed among the units or whether one state, although not dominant, is clearly stonger than the others (Kugler and Lemke 1996). Although those engaged in this debate often frame it in terms of balance of power, in fact the question is quite a separate one.)

The outcomes will follow if four assumptions hold. First, there must be several independent units. Second, the units must want to survive. They can seek to expand and indeed many usually will, but at minimum they must want to maintain their independence. Third, any unit must be willing to ally with any other on the basis of calculations of interest, which means that ideology and hatreds must not be strong enough to prevent actors from working together when it is necessary for them to do so. Fourth, war must be a viable tool of statecraft. Under these conditions, the

system will be preserved even as states press every advantage, pay no attention to the common good, adopt ruthless tactics, and expect others to behave the same way. Put differently, states do not strive for balance; the restraints are not internal in the sense of each state believing that it should be restrained. Rather, restraint and stability arise as ambition checks ambition and self-interest counteracts self-interest.

The basic argument about how this happens is well known, if contested. For any state to survive, none of the others can be permitted to amass so much power that it can dominate. Although states do not invariably join the weaker side, if they are to safeguard their own independence and security they must balance against any actor that becomes excessively menacing. In a way analogous to the operation of Adam Smith's invisible hand, the maintenance of the system is an unintended consequence of states seeking to advance themselves, not the product of their desire to protect the international community or a preference for balance. Balance is then maintained by negative feedback: movement toward dominance calls up forces that put dominance out of reach.

The theory obviously passes one important test in that no state has been able to dominate the international system. But this is not definitive. Few have tried: Napoleon, Hitler, perhaps the Kaiser and Louis XIV. Although others may not have made the effort because they anticipated that they would be blocked, the small number of challenges must undermine our confidence that the system could have been maintained had there been more of them. Furthermore, although the overall balance of power system has never failed, local ones have. Not only have some countries come to dominate their regions (this can perhaps be accommodated within the theory), but isolated systems have fallen under the sway of one actor. While we consider it natural for China to be unified, in fact for centuries it consisted of independent states. Rome's neighbors did not unite to check its power, and the British conquest of India was also made possible by the failure of a local balance. But these cases were still geographically limited and did not produce a world empire and bring international politics to an end.

The other restraints and puzzles mentioned earlier—the fact that few wars become total and that losers and small states are not divided up—also follow from the dictates of self-interest within the constraints imposed by the anarchical system. Since any state can ally with any other, states do not have permanent friends and enemies. Because today's adversary may be tomorrow's ally, crippling it would be foolish. Furthermore, while the state would gain territory and wealth from dividing up the loser, others might gain even more, thus putting the state at a disadvantage in subsequent conflicts.

The knowledge that allies and enemies are not permanent and the expectation that losers will be treated relatively generously reinforce one another. Because the members of the winning coalition know that it is not likely to remain together after the war, each has to fear accretions to the power of its allies. Because winners know that they are not likely to be able to dismember the loser, why should they prolong the war? Each state's knowledge that its allies have reason to contemplate a separate peace provides it with further incentives to move quickly. The result, then, is a relatively moderate outcome not despite but because of the fear and greed of the individual states. This is one reason why international wars are much more likely to end in negotiated settlements than civil wars (Licklider 1993).

There is something wrong with this picture, however. Wars against hegemons can become total, losers sometimes are divided up, and the postwar relations among states are often very different from those prevailing previously (Jervis 1985). The reason is that a long and bitter war against the hegemon undermines the assumptions necessary for the operation of the balance. States are likely to come to believe that wars are so destructive that they cannot be a normal instrument of statecraft and to see the hegemon as inherently evil and aggressive, which means that it is not a fit alliance partner and the winning coalition must stay together. As a result, wartime allies are not regarded as being as much of a potential threat as balance of power reasoning would lead us to expect. Postwar politics may then be unusually moderate and a concert system may evolve in which the states positively value the system, develop longer-run conceptions of their self-interests, and forego competitive gains in the expectation that others will reciprocate. Ironically, then, a war against a would-be hegemon that epitomizes the operation of the balance of power is likely to produce a system in which the actors consciously moderate their behavior and restrain themselves.

1. *An Alternative View*

The model of the balance of power presented here is clearly a version of systems theory in that it sees a radical separation between intentions and outcomes. An alternative view of the balance of power that sees more congruence is summarized by Edward Gulick when he says that 'balance-of-power theory demanded restraint, abnegation, and the denial of immediate self-interest' (Gulick 1955, p. 33). Morton Kaplan's conception of the balance of power similarly posits internalized moderation as two of his six rules call for self-restraint: 'stop fighting rather than eliminate an essential national actor,' and 'permit defeated or constrained essential national actors to re-enter the system as acceptable role partners' (Kaplan 1957, p. 23). For Kaplan, these rules not only describe how states behave, they consciously guide statesmen's actions. In contrast to the version of balancing

discussed earlier, Kaplan points out that in his computer model, 'if actors do not take system stability requirements into account, a "balance of power" system will be stable only if some extra systemic factor …prevents a roll-up of the system' (Kaplan 1979, p. 136). In other words, stability and restraint are not likely unless the actors seek stability.

Here the system is preserved because states want to preserve it and there is little conflict between a state's short-run and long-run interest. The two are harmonized because the norms have been internalized through socialization as the actors watch and interact with their peers. Indeed, Paul Schroeder's important study of the transformation of European international politics caused by the Napoleonic wars stresses that stable peace and the concert were produced not only by the defeat of the aggressor, but also by the painful learning that led the victors to understand that others' interests had to be respected, that smaller states could play a valuable role, and that the eighteenth century practice of compensation and indemnities led to endless cycles of warfare (Schroeder 1994; also see Schroeder 1992 and Jervis 1992). But this view cannot readily explain how the system can be maintained in the face of actors who have interests in exploiting others' moderation.

Nevertheless, it is certainly possible that states feel internal restraints and that, if they do not, the system will be torn apart by high levels of warfare. If the proponents of the version of balance of power set forth here draw on the analogy of Smith's invisible hand, critics can respond that unalloyed capitalism, like an engine out of control, will produce so much unconstrained energy that it will soon destroy itself. Just as economic liberalism must be embedded in broader societal norms if capitalism is to be compatible with a well-functioning society (Polanyi 1944), perhaps the pursuit of narrow self-interest can yield stability and a modicum of productive peace only if it is bounded by normative conceptions that limit predatory behavior.

2. *Anticipation of the Operation of Balance of Power*

States may be restrained by the expectation that if they are not, they will be faced with intense opposition. These cases fall in between the two models discussed above. Indeed, if the view of the balance as automatic is correct it would be surprising if decision-makers heedlessly sought to expand; awareness of the likely feedback would lead them to be restrained even though this impulse does not flow from internalized norms and the desire to preserve the international order. Much has been written about self-defeating expansion, but we should not neglect the fact that leaders may be inhibited by the anticipation of these processes. These cases are literally countless—that is, they cannot be counted because they do not leave traces in the historical record. But perceptive leaders realize that the balance of power makes it dangerous for their countries to be too powerful. Edmund Burke made the point eloquently at the end of the eighteenth century:

> Among precautions against ambition, it may not be amiss to take one precaution against our own. I must fairly say, I dread our own power and our own ambition; I dread our being too much dreaded. It is ridiculous to say we are not men, and that, as men, we shall never wish to aggrandize ourselves in some way or other. Can we say that even at this hour we are not invidiously aggrandized? We are already in possession of almost all the commerce of the world. Our empire in India is an awful thing. If we should come to be in a condition not only to have all this ascendant in commerce, but to be absolutely able, without the least control, to hold the commerce of all other nations totally dependent upon our good pleasure, we may say that we shall not abuse this astonishing and hitherto unheard-of power. But every other nation will think we shall abuse it. It is impossible but that, sooner or later, this stage of things must produce a combination against us which may end in our ruin (quoted in Morgenthau 1978, pp. 169–70).

Finally, if we think of balance of power in the broadest sense of power checking power, these dynamics are built into the basic forms of domestic politics. The American Constitution was built on the concept of checks and balances because the founders believed that potentially dangerous power could best be tamed by countervailing power, to use the term that Galbraith later applied to many aspects of American political and economic life (Galbraith 1952). Other aspects of domestic politics illustrate negative feedback as the unintended consequences of the pursuit of narrower self-interest in a way even more analogous to the automatic balance. Most obviously, it is hard for any political party to gain a monopoly of power because the competition can mount matching or competing claims. If the political pendulum swings in one direction, those losing influence usually will increase their unity and activity.

See also: Alliances: Political; Diplomacy; First World War, The; International Relations, History of; International Relations: Theories; National Security Studies and War Potential of Nations; Nations and Nation-states in History; Peacemaking in History; Realism/Neorealism; War: Causes and Patterns; War, Sociology of; Warfare in History

Bibliography

Claude I 1962 *Power and International Relations*. Random House, New York

Galbraith J K 1952 *American Capitalism: The Concept of Countervailing Power*. Houghton Mifflin, Boston

Gulick E V 1955 *Europe's Classical Balance of Power*. Cornell University Press, Ithaca, NY

Haas E B 1953 The balance of power: prescription, concept or propaganda. *World Politics* **5**: 442–77

Jervis R 1985 From balance to concert. *World Politics* **38**: 58–79
Jervis R 1992 A political science perspective on the balance of power and the concert. *American Historical Review* **97**: 716–24
Kaplan M A 1957 *System and Process in International Politics*. Wiley, New York
Kaplan M A 1979 *Towards Professionalism in International Theory*. Free Press, New York
Kugler J, Lemke D (eds.) 1996 *Parity and War*. University of Michigan Press, Ann Arbor, MI
Licklider R (ed.) 1993 *Stopping the Killing: How Civil Wars End*. New York University Press, New York
Morgenthau H J 1978 *Politics Among Nations*, 5th edn. Rev. Knopf, New York
Polanyi K 1944 *The Great Transformation*. Farrar & Rinehart, New York
Schroeder P W 1992 Did the Vienna settlement rest on a balance of power. *American Historical Review* **97**: 683–706
Schroeder P W 1994 *The Transformation of European Politics, 1787–1848*. Oxford University Press, New York
Waltz K N 1979 *Theory of International Politics*. Addison-Wesley, Boston, MA
Waltz K N 1991 America as a model for the world? A foreign policy perspective. *PS: Political Science and Politics* **24**: 669
Waltz K N 1993 The emerging structure of international politics. *International Security* **18**: 44–79
Wight M 1968 The balance of power. In: Butterfield H (ed.) *Diplomatic Investigations*. Harvard University Press, Cambridge, MA
Wight M 1973 The balance of power and international order. In: James A (ed.) *The Bases of International Order*. Oxford University Press, London

R. Jervis

Bankruptcy

Bankruptcy procedures are intended to provide an efficient and fair mechanism for the reorganization or liquidation of the assets of insolvent debtors. Debtors include both individuals and business entities. A second important objective of some bankruptcy procedures is the financial rehabilitation of overindebted individuals. Rehabilitation sometimes includes discharge of indebtedness.

1. Fair and Efficient Administration of Insolvent Estates

1.1 The Need for Bankruptcy

Reliance on the usual legal procedures to collect debts from non-paying debtors can lead to inefficient and unfair results when applied to an insolvent debtor. When there are not enough assets to pay all creditors, non-bankruptcy law commonly favors the creditor who seizes and sells assets before other creditors do. The resulting race to seize the debtor's assets can lead to inefficient dispositions of debtor's assets. For example, if a business is making a profit even while insolvent, it may be more efficient to allow it to continue to operate and pay debts from future profits, yet competition between creditors can lead to sale of assets that makes continued operation of the business impossible (Jackson 1985). Competition between creditors also leads to unnecessarily duplicative collection activities. Many people also believe it is unfair to give priority in the distribution of the limited assets of an insolvent estate to the creditors who are the first to initiate formal collection actions.

Bankruptcy procedures address each of these difficulties. Once a bankruptcy proceeding is initiated, unsecured creditors are automatically enjoined from using non-bankruptcy collection procedures. In some countries, secured creditors are similarly enjoined. These injunctions eliminate duplicative collection efforts and permit an orderly disposition of the debtor's assets. In bankruptcy, the debtor's assets constitute a bankruptcy estate, to be managed in the interests of the estate's beneficiaries, the creditors. When the bankruptcy estate makes distributions, creditors with similar contractual priorities are usually paid pro rata according to the amount they are owed. Contractual promises to subordinate or to privilege particular creditors in the distribution of debtor's assets (including security agreements) are normally respected, and a few creditors (e.g., tax creditors) receive priority payments by statutory mandate. However, priorities are not usually given to creditors who have initiated collection activities before the bankruptcy filing.

1.2 Liquidation vs. Reorganization in Business Bankruptcy

The assets of an insolvent business estate may be sold, either as a unit or in separate parts, to the highest bidder(s), which is called a liquidation. Alternatively, the assets may be retained by an entity and operated as a continuing business, which is called a reorganization. In a liquidation, creditors are paid the proceeds of the sale(s). In a reorganization, creditors are given securities (debt instruments and/or shares) in the new reorganized entity, which represent rights in the future income of the continuing business. Bankruptcy creditors often have conflicting interests in the decision whether to liquidate or reorganize an insolvent business estate.

Creditors with contractual priority over other creditors (including, most importantly, secured creditors) usually prefer rapid liquidation if the anticipated proceeds will pay them in full. Reorganization both delays repayment and introduces an element of risk, because the continuing business might lose money, thus depriving these senior creditors of full payment. This preference for liquidation exists even when reorganization seems the better option from the

perspective of creditors generally, because a reorganized business is more likely than not to make profits, yet in an immediate sale a buyer is not likely to pay for those prospective gains.

Contractually subordinated creditors, on the other hand, may favor reorganization, even when it seems more likely than not that a reorganized business will suffer losses. Immediate liquidation of the assets may not yield any payments to subordinated creditors, and there is always a possibility of payments from the future profits of a continuing business. Three other groups may also favor reorganization, even when immediate liquidation seems a better prospect for maximizing payments to creditors. Trade creditors may prefer to continue the business under the former management because of the prospect of making future sales. Management of the business is often identified as a separate interest group favoring reorganization, both to preserve their jobs and to provide an opportunity to rehabilitate professional reputations harmed by business insolvency. Finally, the owners of the insolvent business will normally favor reorganization. They will receive nothing if the business is liquidated while insolvent, but in a reorganization future profits of the continuing business may be sufficient to return the entity to solvency.

1.3 Governance Issues in Business Bankruptcy

Conflicts in interest between creditors can generate controversy about control over decision-making within a bankruptcy. In most countries, senior creditor interests control the appointment of the effective decision makers. Often, secured creditors are not even enjoined from foreclosing on their collateral by the initiation of bankruptcy, and they may be able to force liquidation simply by doing so. In these circumstances, reorganization within bankruptcy may seem unlikely. In response, management and owners of insolvent businesses may take desperate steps in order to avoid a bankruptcy filing. For example, a business might sell non-essential assets quickly at lower prices than could be obtained if more time were taken, in order to pay off a senior creditor who is threatening to seize assets (Baird 1991). Junior creditor interests may also purchase the claims of senior creditor interests, even after a bankruptcy filing, in order to forestall an immediate liquidation.

In the United States, bankruptcy procedures often allow management of the insolvent business to exert substantial control over decision-making within bankruptcy. It is widely believed that this system leads to decisions to reorganize an insolvent business when immediate liquidation would be in the overall interests of creditors as a whole. Furthermore, it causes senior-creditor interests to agree to compromise their contractual priority rights, in favor of junior interests, in order to facilitate a quicker liquidation or a reorganization in which senior interests receive debt instruments with a high likelihood of ultimate payment. There has been considerable academic criticism of the governance of business bankruptcy in the United States, describing it as inefficient for allowing insolvent businesses to continue without sufficient prospects of returning to profitability, and as unfair in failing to protect fully contractual priority rights (Bradley and Rosenzweig 1992). Its defenders point out that there are many non-creditors with an interest in reorganization—most importantly, the employees of the insolvent business (Warren 1993).

1.4 Law and Practice in Business Bankruptcy

Although frequently subject to the same formal rules, there are often important practical differences between the administration of bankruptcy estates of large and small businesses. In the case of large businesses, there are likely to be substantial assets and many creditors. Creditors with different priorities are organized into groups and represented by lawyers. Cases tend to be litigious and lengthy, but most interests are represented vigorously. Because of the influence of management, there may be a slight favoritism towards reorganization, but it is not great (LoPucki and Whitford 1993). In small-business bankruptcies, however, apart from possibly one large creditor which is likely to be secured, creditors are more likely to believe that the combination of the amount of debt at stake and the prospect of a substantial ultimate payment is insufficient to justify investment in legal fees. As a result, governance in small-business cases often is left to management and perhaps the one large creditor, and they may agree to a reorganization or to delay liquidation, to the substantial detriment of creditor interests (LoPucki 1983). There have been calls for greater judicial or administrative oversight of the governance of small cases to insure more appropriate representation of smaller creditor interests.

2. Rehabilitation of Individual Debtors

The automatic injunction upon bankruptcy filing against non-bankruptcy collection actions can help the financial rehabilitation of individual debtors. By protecting the debtor's paychecks and bank accounts from judicial attachment during the course of the bankruptcy proceedings, the automatic injunction gives the debtor time to work out financial problems. Most countries allow the debtor to extend the duration of the automatic injunction for several years by proposing and then carrying out a plan for the repayment of some or all of the debt through periodic payments.

A few countries, most notably the United States, allow an individual debtor to obtain a discharge of

many unsecured debts in a bankruptcy proceeding. Before getting the discharge, the debtor must turn over for the benefit of unsecured creditors all property that is not encumbered by a security interest, or considered exempt from this requirement because it is essential to the debtor's everyday life. In practice, individual debtors usually enter bankruptcy with little or no property that is neither exempt nor unencumbered, so unsecured creditors usually receive nothing in these bankruptcy proceedings.

The availability of discharge has been justified historically as a means of providing individuals with a 'fresh start,' so that they might again become productive economic participants in society. An overindebted debtor without realistic hope of repaying all debts in the foreseeable future has little incentive to seek new economic opportunities, because the primary beneficiary of increased income will be the creditors (Whitford 1997). Relief of individual hardship is another rationale sometimes offered for the discharge.

The majority of countries that allow for the discharge of individual debt make the availability of this relief conditional upon completion of a plan for payment of debts over a three- to ten-year period (Huls 1994). Conditional discharge availability in this manner is justified as discouraging the incurring of debt in anticipation of discharge. It is also believed that payment plans can provide debtors with the opportunity to learn the disciplined living and household budget skills that will contribute to the overall financial rehabilitation objective, but there is little information available about how well payment plans achieve that objective. It is known that in the United States, where debtors have certain incentives to enter a payment plan voluntarily rather than seek an immediate discharge, more than two-thirds of payment plans are not completed successfully. Countries vary in whether they will grant a discharge to a debtor who has attempted, but not completed, a payment plan. Where discharge is denied, the result is a significant restriction of the availability of a fresh start to individual debtors.

The use of credit cards and unsecured indebtedness by individuals has increased dramatically throughout the world since the 1980s. Probably as a result of this, many countries are now reconsidering their laws concerning individual bankruptcy. Many countries that have not previously permitted discharge of consumer indebtedness now allow an individual to gain a fresh start, provided the debtor makes an attempt in good faith to complete a payment plan (Niemi-Kiesilainen 1997). In the United States, where traditionally debtors have been able to obtain a discharge without first attempting a payment plan, there has been a rapid increase in the bankruptcy filing rate. As a response, there is a significant political effort to limit the availability of the discharge for debtors whose income exceeds certain standards. These debtors would be required to complete a five-year payment plan before receiving a discharge of any amount not paid under the plan. If adopted, the result is likely to be a significant decrease in the bankruptcy filing rate, and, because most payment plans will not be completed satisfactorily, an even greater decrease in the number of discharges in bankruptcy.

See also: Business Law; Consumer Economics; Ethical Codes, Professional: Business Codes

Bibliography

Baird D 1991 The initiation problem in bankruptcy. *International Review of Law and Economics* **11**: 223–32

Bradley M, Rosenzweig M 1991 The untenable case for Chapter 11. *Yale Law Journal* **101**: 1043–95

Huls N J H 1994 *Overindebtedness of Consumers in the EC Member States*. Centre de Droit de la Consommation/Diegem, Brussels, Belgium

Jackson T H 1985 *The Logic and Limits of Bankruptcy Law*. Harvard University Press, Cambridge, MA

LoPucki L 1983 Debtors in full control—systems failure under Chapter 11 of the Bankruptcy Code. *American Bank Law Journal* **57**: 99–126, 247–73

LoPucki L M, Whitford W C 1993 Patterns in the bankruptcy reorganization of large, publicly held companies. *Cornell Law Journal* **78**: 597–618

Niemi-Kiesilainen J 1997 Changing directions in consumer bankruptcy law and practice in Europe and USA. *Journal of Consumer Policy* **20**: 133–42

Warren E 1993 Bankruptcy policymaking in an imperfect world. *Michigan Law Review* **92**: 336–87

Whitford W 1997 Changing definitions of fresh start in US Bankruptcy Law. *Journal of Consumer Policy* **20**: 179–98

W. C. Whitford

Basal Ganglia

1. Basal Ganglia Anatomy and Neurochemistry

As an anatomical entity, the basal ganglia have undergone significant revision. In 1664 anatomist Thomas Willis originally termed this subcortical region of the telencephalon 'corpus striatum.' The development of neuronal tracing techniques by Nauta and colleagues in the mid-1950s (e.g., silver degeneration) allowed for elucidation of efferent connections within the broadly defined corpus striatal region, and the term 'basal ganglia' was eventually adopted to refer to a relatively restricted group of subcortical structures that originally included the caudate nucleus and putamen, the globus pallidus, the claustrum, and the amygdala. As the amygdala became more closely identified as a component of the limbic system, its inclusion as a basal ganglia structure was dropped.

Heimer and colleagues adopted the term 'ventral striatum' to delineate the most ventral aspects of the striatum from more dorsal regions (i.e., caudate-putamen), and this striatal region includes the nucleus accumbens and portions of the olfactory tubercle. Today many neuroanatomists would agree that the core structures of the mammalian basal ganglia include the caudate nucleus, putamen, ventral striatum, and globus pallidus. In addition, the substantia nigra, ventral tegmental area, and the subthalamic nucleus are considered associated basal ganglia structures via their reciprocal connections with the core structures.

The basal ganglia receive input from virtually all regions of the cerebral cortex, and these corticostriatal pathways are topographically organized. The presence of cortico-basal ganglia loops is an important feature of the anatomical and functional organization of the basal ganglia. Specific cortical regions project to the basal ganglia, and output from the basal ganglia 'loops' back into these same cortical regions via various thalamic nuclei. An example of one such pathway is the projection from the primary motor cortex $\gg$ putamen $\gg$ globus pallidus $\gg$ ventrolateral thalamus $\gg$ primary motor cortex. Evidence suggests that at least four other parallel cortico-basal ganglia loops exist.

Neurochemically, the basal ganglia are characterized by a substantial input from midbrain dopaminergic pathways originating in the substantia nigra and ventral tegmental area, and primarily innervating the dorsal and ventral striatum, respectively. Medium spiny output neurons of the neostriatum utilize gamma-amino butyric acid as a neurotransmitter, and corticostriatal afferent projections are predominantly glutamatergic. Several neuropeptides are localized in the basal ganglia including various opioids, cholecystokinin, substance P, somatostatin, and neurotensin. A final prominent component of basal ganglia neurochemistry is a large population of cholinergic interneurons.

In the 1980s, a series of important findings demonstrated the existence of two neural 'compartments' in the mammalian neostriatum that are neurochemically and anatomically differentiated, and are commonly termed the striatal 'patch' and 'matrix.' Neurochemically, the patch compartments of the striatum (also termed striosomes) are characterized by low levels of acetylcholine and high levels of various opiates and substance P. In contrast, the matrix compartment is characterized by cholinergic and somatostatin-containing neurons. Both compartments receive dopaminergic input, although dopamine pathways originating in the ventral tegmental area and substantia nigra primarily innervate the patch and matrix, respectively. Anatomically, corticostriatal and thalamostriatal projections are closely associated with the striatal matrix, whereas projections from limbic structures to the striatum (e.g., hippocampus, amygdala) appear primarily to innervate striatal patches. Elucidation of the functional significance of the neurochemical and anatomical differentiation observed between these two striatal compartments represents an important and evolving area of basal ganglia research.

2. *Basal Ganglia and Neurodegenerative Disease*

In 1817 the English physician James Parkinson first described a neurological disorder characterized by limb rigidity, tremors, and difficulty initiating movement. Parkinson's disease is a progressive neurodegenerative disorder that may affect up to 1 percent of the population and increases in incidence with advancing age. Fundamentally, Parkinson's disease is a disorder of basal ganglia function that involves the loss of neurons in the substantia nigra, leading to a concomitant decrease of the neurotransmitter dopamine within the nigrostriatal dopamine pathway. Although there is no known cure for Parkinson's disease, a major breakthrough in the treatment of this disorder came with the discovery that administration of the dopamine precursor 3,4-dihydroxyphenylalanin (L-DOPA) improved the symptoms of Parkinson's patients, presumably by increasing the availability of dopamine within the neostriatum. In recent years, dopamine-containing neural tissue from human fetus or adrenal cortex has been stereotaxically implanted directly into the neostriatum of Parkinson's patients in an attempt to provide a long-term reversal of dopaminergic dysfunction. The results of some of these studies have been encouraging, and investigation of the efficacy of neural grafting techniques in Parkinson's disease appears a promising avenue for further research.

A second neurological disorder that involves the basal ganglia is Huntington's chorea, first described in a group of patients by George Huntington in 1872. Huntington's disease has a clear genetic basis and is characterized by the production of involuntary 'choreic' movements. In contrast to Parkinson's disease, the nigrostriatal dopamine system is largely intact in Huntington's disease. Instead, degeneration of intrinsic GABAergic and cholinergic cells in the basal ganglia may disinhibit the nigrostriatal dopamine system, resulting in excessive inhibition of pallidal neurons projecting to the thalamus, and leading to involuntary movements.

3. *Basal Ganglia, Reward, and Drug Addiction*

In the mid-1950s, James Olds and colleagues initiated the neuroscientific study of brain reward processes with their fortuitous discovery of self-administration of electrical stimulation in the mammalian brain. In addition to implicating hypothalamic pathways,

subsequent research revealed that animals readily self-stimulate the medial forebrain bundle, a neural pathway includes fibers containing the dopamine projection from the ventral tegmental area to the ventral striatum. Research conducted over the ensuing four decades has implicated the basal ganglia in reward processes mediated not only by 'artificial' electrical brain simulation, but also by natural reinforcers such as food, water, and sex. In particular, the nucleus accumbens may be a critical component of the neural circuitry by which various drugs with abuse potential in humans (i.e., addictive drugs) produce rewarding affect. Several lines of evidence are consistent with this hypothesis. First, damage to the nucleus accumbens and/or the ventral tegmental area influences self-administration of several drugs of abuse in experimental animals. Second, accumbens damage attenuates conditioned place preference behavior, in which animals learn to approach environmental cues that have been previously associated with drug administration. Finally, conditioned place preference behavior and self-administration of various drugs are affected by intracerebral administration of neurotoxic agents that selectively damage dopamine neurons, and by injection of dopamine antagonist or agonist drugs directly into the nucleus accumbens. These findings have in large part led to the prominent view that dopamine acts as a reward signal in the mammalian nucleus accumbens. Although not without controversy, this hypothesis continues to have a tremendous impact on basic research investigating the neurobiological bases of drug addiction, and the basal ganglia remain a prominent focus in this field.

4. *Basal Ganglia and Learning and Memory*

Early studies on the possible role of the basal ganglia in learning and memory were guided in part by anatomical evidence demonstrating the existence of the corticostriatal pathways. One hypothesis that has gained increased support is that the participation of the basal ganglia in learning and memory is organized by the nature of the topographical cortical input that these subcortical structures receive. For example, in experimental animals lesions of either the frontal cortex or the medial region of the dorsal striatum to which it projects produce similar impairments in performance of delayed alternation tasks. In addition, lesions of regions of the caudate-putamen that receive visual or olfactory input selectively impair conditioned emotional responding based on visual or olfactory stimuli, respectively.

Within a multiple memory systems framework, Mishkin and Petri (1984) have hypothesized that the dorsal striatum mediates a noncognitive form of stimulus-response (S-R) or 'habit' memory, in which associations between stimuli and responses are acquired in an incremental fashion. This hypothesis was based in part on early findings in monkeys of impaired simultaneous visual discrimination learning following lesions of the ventral putamen, and is now supported by studies in rats, monkeys, and humans. For example, double dissociation experiments indicate that lesions of the caudate-putamen of rats selectively impair S-R visual discriminations in rats, while leaving cognitive forms of memory mediated by other brain regions (e.g., spatial memory/hippocampal memory) intact. Selective deficits in habit learning have also been demonstrated in early to mid stages of Parkinson's disease and Huntington's chorea. For example, Parkinson's patients show impaired probabilistic classification learning that is based on arbitrary S-R associations. Functional neuroimaging studies have also revealed a role for the human neostriatum in performance of motor habits and motor sequence learning, and various computational approaches to information processing in the basal ganglia have been recently developed.

Mogenson and colleagues (1980) proposed that projections from limbic brain regions including the basolateral amygdala and hippocampus to the ventral striatum (e.g., nucleus accumbens) provide an interface between motivational states and behavioral action. This general hypothesis has received increasing support, and is reflected by recent evidence suggesting that the nucleus accumbens plays a selective role in stimulus-reward learning and memory via connectivity with the amygdaloid complex. Similarly, the nucleus accumbens may participate in hippocampus-dependent memory processes (e.g., spatial memory) via projections from the hippocampal subiculum.

Behavioral pharmacology experiments indicate a role for cholinergic, dopaminergic, and glutamatergic neurotransmission in learning and memory processes mediated by the neostriatum, in particular during memory consolidation. Long-term potentiation and long-term depression, two forms of synaptic plasticity that putatively mediate learning and memory processes occurring in other brain regions (e.g., hippocampus and cerebellum), are also experimentally inducible in the neostriatum via stimulation of corticostriatal fibers. The extent to which these forms of striatal plasticity mediate the selective learning and memory functions of the neostriatum is not currently known.

5. *Basal Ganglia and Psychopathology*

In addition to the long recognized role of the basal ganglia in neurodegenerative diseases compromising human motor behavior, recent evidence indicates a role for the neostriatum in obsessive-compulsive disorder (OCD), an anxiety disorder characterized by obsessive thought processes and compulsive behavioral acts. Structural neuroimaging studies have implicated the caudate nucleus in OCD pathology, although such studies have variously reported in-

creases or decreases in caudate volume. The overall pattern of findings from functional neuroimaging studies suggests increases in activity of the caudate nucleus and associated afferent cortical structures, including primarily the orbitofrontal and anterior cingulate cortices. One important study found that either medication or behavioral treatments resulted in similar reductions of activity in the caudate nucleus and frontal corticostriatal regions of OCD patients.

An interesting 'evolutionary' link between basal ganglia function and motoric aspects of OCD has been suggested by Rapoport and colleagues, who have pointed to similarities between the role of the caudate-putamen of lower animals in stereotypical species-specific behaviors (e.g., grooming), and the high incidence of washing and cleaning behaviors exhibited in OCD. In addition to OCD, basal ganglia dysfunction has also been associated with the compulsive and involuntary motor tics and vocalizations characterizing Tourette's syndrome. Finally, involuntary movements and OCD-like symptoms also accompany Sydenham's chorea, an autoimmune disease that attacks various brain regions including the basal ganglia.

The basal ganglia have also been hypothesized to play a role in schizophrenia, a disorder characterized by disturbances in thought, affect, and sensory processes. The discovery that the phenothiazene class of antipsychotic drugs act as dopamine receptor blockers led to an examination of the role of brain dopamine pathways innervating the basal ganglia in schizophrenia. In particular, the mesolimbic/mesocortical dopamine pathways originating in the ventral tegmental area (VTA) and innervating the nucleus accumbens and cortical regions including the prefrontal cortex may be an important component of the neural circuitry by which dopamine antagonist drugs exert their antipsychotic effects.

6. *Summary*

The basal ganglia are one of the most well studied groups of brain structures in the mammalian brain. Anatomical considerations and the long documented role of the basal ganglia in neurodegenerative diseases such as Parkinson's and Huntington's chorea have in part led to the traditional association of the basal ganglia with motor behavior. However, as this brief survey of basal ganglia research indicates, neuroscientists today recognize that the role of these structures in behavior is not restricted to motor functions. Rather, progress in understanding the neurobiological bases of numerous psychological processes is being made in basal ganglia research, including drug addiction and reward, learning and memory, and psychopathology. The anatomical and neurochemical heterogeneity reflected by the presence of cortico-basal ganglia loops and striatal patch/matrix compartmentalization may ultimately provide clues to understanding these various functional roles, and the study of basal ganglia–behavior relationships will continue to be an exciting and promising field of psychological research.

See also: Cognitive Control (Executive Functions): Role of Prefrontal Cortex; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Neurotransmitters; Vision for Action: Neural Mechanisms

Bibliography

Alexander G E, DeLong M R, Strick P L 1986 Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience* **9**: 357–81

Beiser D G, Hua S E, Houk J C 1997 Network models of the basal ganglia. *Current Opinion in Neurobiology* **7**: 185–90

Cador M, Robbins T W, Everitt B J 1989 Involvement of the amygdala in stimulus-reward associations: Interaction with the ventral striatum. *Neuroscience* **30**: 77–86

Gerfen C R 1992 The neostriatal mosaic: Multiple levels of compartmental organization in the basal ganglia. *Annual Review of Neuroscience* **15**: 285–320

Graybiel A M 1990 Neurotransmitters and neuromodulators in the basal ganglia. *Trends in Neurosciences* **13**: 244–54

Graybiel A M 1995 The basal ganglia. *Trends in Neurosciences* **18**: 60–2

Graybiel A M 1996 Basal ganglia: New therapeutic approaches to Parkinson's disease. *Current Biology* **6**: 368–71

Knowlton B J, Mangels J A, Squire L R 1996 A neostriatal habit learning system in humans. *Science* **273**: 1353–4

Koob G F, Sanna P P, Bloom F E 1998 Neuroscience of addiction. *Neuron* **21**: 467–76

Mishkin M, Petri H L 1984 Memories and habits: Some implications for the analysis of learning and retention. In: Squire L R, Butters N (eds.) *Neuropsychology of Memory*. Guilford, New York, pp. 287–96

Mogensen G J, Jones D L, Yim C Y 1980 From motivation to action: Functional interface between the limbic system and the motor system. *Progress in Neurobiology* **14**: 69–97

Packard M G, Hirsh R, White N M 1989 Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *The Journal of Neuroscience* **9**: 1465–72

Parent A 1986 *Comparative Neurobiology of the Basal Ganglia*. Wiley, New York

Rapoport J L 1998 The new biology of obsessive-compulsive disorder: Implications for evolutionary psychology. *Perspectives in Biology and Medicine* **41**: 159–75

Rauch S L, Savage C R 1997 Neuroimaging and neuropsychology of the striatum: Bridging basic science and clinical practice. *Psychiatric Clinics of North America* **20**: 741–68

Schultz W 2000 Multiple reward signals in the brain. *Nature Reviews Neuroscience* **1**: 199–208

Schwartz J M, Stoessel P W, Baxter L R, Martin K M, Phelps M E 1996 Systematic changes in cerebral glucose metabolic rate after successful behavior modification treatment of obsessive-compulsive disorder. *Archives of General Psychiatry* **53**: 109–13

Weickert C S, Kleinman J E 1998 The neuroanatomy and neurochemistry of schizophrenia. *Psychiatric Clinics of North America* **21**: 57–75
White N M 1997 Mnemonic functions of the basal ganglia. *Current Opinion in Neurobiology* **7**: 164–9
Wichmann T, DeLong M R 1996 Functional and pathophysiological models of the basal ganglia. *Current Opinion in Neurobiology* **6**: 751–8

M. G. Packard

Bayesian Graphical Models and Networks

1. Introduction

A graphical model represents the probabilistic relationships among a set of variables. Nodes in the graph correspond to variables, and the absence of edges corresponds to conditional independence. Graphical models are becoming more popular in statistics and in its applications in many different fields for several reasons. For example, they provide a language to facilitate communication between a domain expert and a statistician, provide flexible and modular definitions of families of probability distributions, and are amenable to scaleable computational techniques. (e.g., Pearl 1988, Lauritzen 1996, Whittaker 1990).

The subject of this article is directed acyclic graphical (DAG) models. These models have numerous uses including data analysis (e.g., Whittaker 1990, Spiegelhalter and Thomas 1998), modeling of causal relationships (e.g., Pearl 2000, Spirtes et al. 2001), and representing and reasoning about uncertainty in expert systems (e.g., Cowell et al. 1999). Specific applications of graphical models include diagnosis and troubleshooting, medical monitoring, genetic counseling, information retrieval, natural language processing, weather forecasting, manufacturing, digital communication, and machine vision.

In Sect. 2, we define the DAG model and describe its basic properties. In Sect. 3, we describe the use of DAG models in expert systems and decision analysis. In Sect. 4, we describe methods for constructing the structure and distributions of DAG models from data. In Sect. 5 we conclude with additional pointers to the literature.

2. Definition and Basic Properties

Let us begin with some notation. We denote a variable by a capitalized token (e.g., X, X_i, Θ, *Age*), and the state or value of a corresponding variable by that same token in lower case (e.g., x, x_i, θ, *age*). We denote a set of variables by a bold-face capitalized token (e.g., $\mathbf{X}$, $\mathbf{X}_i$, $\mathbf{Pa}_i$). We use a corresponding bold-face lower-case token (e.g., $\mathbf{x}$, $\mathbf{x}_i$, $\mathbf{pa}_i$) to denote an assignment of state or value to each variable in a given set. We use $p(x|y)$ to denote the probability or probability density that $X = x$ given $Y = y$. We also use $p(x|y)$ to denote the probability distribution for X given Y. Whether $p(x|y)$ refers to a probability, probability density, or a probability distribution will be clear from context. Finally, we use $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$ to mean that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$.

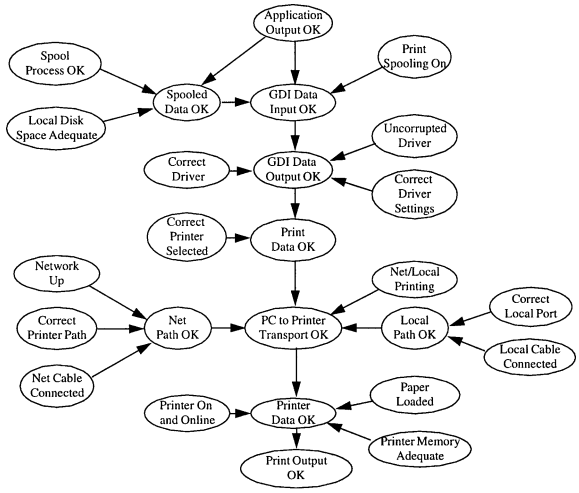

Figure 1
A DAG model for troubleshooting a printing problem

A DAG model for a set of variables or *domain* $\mathbf{X} = \{X_1, \ldots, X_n\}$ consists of (a) a network structure $\mathbf{s}$ that encodes a set of conditional independence assertions about variables in $\mathbf{X}$, and (b) a set $\mathbf{p}$ of local probability distributions associated with each variable. Together, these components define the joint probability distribution for $\mathbf{X}$. The network structure $\mathbf{s}$ is a directed acyclic graph, that is, each edge is directed and there are no paths along these directed edges that lead from a variable to itself. The nodes in $\mathbf{s}$ are in one-to-one correspondence with the variables $\mathbf{X}$. We use X_i to denote both the variable and its corresponding node, and $\mathbf{Pa}_i$ to denote the parents of node X_i in $\mathbf{s}$ as well as the variables corresponding to those parents. The *lack* of possible directed edges in $\mathbf{s}$ encode conditional independencies. In particular, given structure $\mathbf{s}$, the joint probability distribution for X is given by

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i | \mathbf{pa}_i) \tag{1}$$

The *local probability distributions* P are the distributions corresponding to the terms in the product of Eqn. (1). Consequently, the pair $(\mathbf{s}, \mathbf{p})$ encodes the joint distribution $p(\mathbf{x})$.

As an example, consider Fig. 1, which illustrates the structure of a DAG model for troubleshooting a

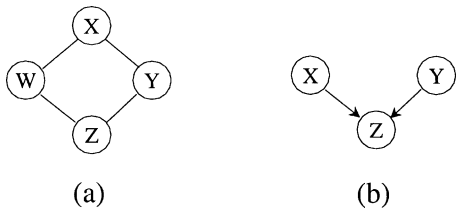

Figure 2
Undirected (a) and directed (b) graphical-model structures

printing problem on the Microsoft Windows 95 operating system. This model structure was constructed from expert knowledge using judgments of cause and effect. Note that, in general, the probability distributions in a DAG model may be subjective or objective.

In the remainder of this section, we discuss several basic issues concerning DAG models. Equation (1) implies that, for all X, the nondescendants of X are independent of X given the parents of X. Many additional independencies are entailed by this basis of independencies. In Fig. 1, for example, we have that *Print Output OK* and *Application Output OK* are conditionally independent given *Print Data OK*. Pearl (1988) describes a condition, called *d-separation*, that allows one to read off any independency from a DAG model in a computationally efficient manner.

Another commonly used graphical model is the undirected graphical (UG) model, also known as a Markov Random Field or Markov Network (e.g., Whittaker 1990, Lauritzen 1996). The independencies that can be encoded by DAG and UG models overlap, but there are independencies that can be encoded by UGs and not DAGs, and vice versa. For example, consider the UG model in Fig. 2a for the domain (X, Y, Z, W). This graph encodes the independencies $X \perp Z | (Y, W)$ and $Y \perp W | (X, Z)$. There is no DAG model for these four variables that can encode both of these independencies. In contrast, consider the DAG model in Fig. 2b for the domain (X, Y, Z). This graph encodes the marginal independence of X and Z and the possible dependence of X and Z given Y. No UG model for these variables can encode this set of relationships.

The relationships encoded by the 'v structure' or 'collider' in Fig. 2b has an intuitive explanation. For example, consider the subgraph in Fig. 1 for the variables *Correct Local Port*, *Local Cable Connected*, and *Local Path OK*, which has this structure. Before one observes the value of *Local Port OK*, the remaining two variables are marginally independent. When we observe that *Local Port OK* is false, however, the remaining two variables become dependent. If—say—*Correct Local Port* subsequently is observed to be false, then *Local Cable Connected* is more likely to be true, because the faulty local path has been explained by the first observation. In general, this pattern of relationships is sometimes referred to as the 'explaining away' phenomenon (Pearl 1988).

Another interesting point about DAG models is that two or more DAG models can be equivalent in the sense that they represent the same independencies. Given the domain (X, Y, Z),for example, the DAG models $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$, and $X \leftarrow Y \leftarrow Z$ encode the same independency: $X \perp Z | Y$. Chickering (1995) describes a simple and computationally efficient algorithm for determining whether two DAG models are equivalent. First, he defines a directed edge from X to Y to be *covered* if the parents of X and Y are the same with the exception that X is also a parent of Y. He then proves that two DAG models are equivalent if and only if there exists a sequence of reversals of covered directed edges that transforms one model to the other.

Finally, we note that the definition of DAG model does not restrict the form of the local distributions $p(x_i|\mathbf{pa}_i)$. Later, we describe several commonly used models for local distributions. Here, we mention that, when a variable and its parents are continuous, a commonly used local distribution is the linear regression with Gaussian error. When all variables in a domain are continuous and the error terms are mutually independent, the resulting DAG model is equivalent to a structural equation model with independent errors (e.g., Pearl 2000).

3. *DAG Models for Expert Systems and Decision Analysis*

The modern use of DAG models began with researchers in Management Science and AI communities (e.g., Howard and Matheson 1981, see *Artificial Intelligence: Uncertainty*) who used them to encode prior or 'expert' knowledge for decision making and expert-system inference. These communities refer to this class of model as Knowledge Maps and Bayesian Networks, respectively. In this section, we examine this use of DAG Models.

First, let us consider how a person—who we will call the decision maker—constructs a DAG model for a given domain $\mathbf{X}$ from their knowledge. One method for construction starts with a total ordering on the variables provided by the decision maker. For simplicity, let us relabel the variables $\mathbf{X} = (X_1, \ldots, X_n)$ so that X_i is the ith variable in this ordering. From the chain rule of probability, we have

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1}) \tag{2}$$

For every X_i there will be some subset $\Pi_i \subseteq$

$\{X_1, \ldots, X_{i-1}\}$ such that X_i and $\{X_1, \ldots, X_{i-1}\}\backslash\Pi_i$ are conditionally independent given Π_i, that is, for any $\mathbf{x}$

$$p(x_i|x_1, \ldots, x_{i-1}) = p(x_i|\pi_i) \tag{3}$$

Combining Eqns. (2) and (3), we obtain

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\pi_i) \tag{4}$$

Comparing Eqns. (1) and (4), we see that the variables sets $(\Pi_1, \ldots, \Pi_n)$ correspond to the DAG-model parents $(\mathbf{Pa}_1, \ldots, \mathbf{Pa}_n)$, which in turn fully specify the directed edges in the network structure $\mathbf{s}$. Consequently, to determine the structure of a DAG model, a decision maker (a) orders the variables somehow and (b) determines the variables sets that satisfy Eqn. (3) for $i = 2, \ldots, n$. Finally, the decision maker assesses the local distributions $p(x_i|\mathbf{pa}_i)$.

The approach for constructing model structure just described has an important drawback. If one chooses the variable order carelessly, then the resulting network structure may fail to reveal many conditional independencies among the variables. Fortunately, there is another often-used technique for constructing DAG models that does not require an ordering. The approach is based on two observations: (a) people can often readily assert causal relationships among variables, and (b) absent causal relationships typically correspond to conditional independencies. Thus, to construct a DAG model for a given set of variables, one simply draws arcs from cause variables to their immediate effects. In almost all cases, doing so results in a network structure that satisfies the definition Eqn. 1. The DAG model in Fig. 1 was constructed in this manner.

Once a DAG model is constructed (whether from prior knowledge, data, or both), we can use it to answer probabilistic queries. In the context of decision analysis, the answers to these queries can be used for decision making. In the context of expert systems, these answers convey 'expert advice.' For example, given the DAG model in Fig. 1, we may be interested in the probability that *Printer On* and *Online* is true, given *Print Output OK* is false. We can obtain this probability by applying the rules of probability to the joint distribution defined by the DAG model. Many researchers refer to this process as *probabilistic inference*.

Direct computation via the joint distribution is expensive. Investigators, however, have developed probabilistic-inference algorithms that use the conditional independencies in the DAG model to speed up inference, sometimes dramatically. Many of these algorithms are limited to DAG models in which all variables are discrete. For example, Howard and Matheson (1981) and Shachter (1988) developed an algorithm that reverses arcs in the network structure until the answer to the given probabilistic query can be read directly from the graph. In this algorithm, each arc reversal corresponds to an application of Bayes's theorem.

Pearl (1988) developed a message-passing scheme that updates the probability distributions for each node in a DAG model in response to observations of one or more variables. Lauritzen and Spiegelhalter (1988) created an algorithm that first transforms the DAG model into a tree where each node in the tree corresponds to a subset of variables in $\mathbf{X}$. The algorithm then exploits several mathematical properties of this tree to perform probabilistic inference. Lauritzen (1992) extended this algorithm to include DAG models that encode multivariate-Gaussian or Gaussian-mixture distributions.

When the edges of a DAG model structure form a tree, these methods for exact probabilistic inference scale linearly in the number of domain variables. In general, however, these algorithms are NP-hard (Cooper 1990). The source of the complexity lies in undirected cycles in the DAG model—cycles in the structure where we ignore the directionality of the edges. When a DAG model contains many undirected cycles, inference is intractable. For many applications, however, structures are simple enough (or can be simplified sufficiently without sacrificing much accuracy) so that inference is efficient. For those applications where generic inference methods are impractical, researchers are developing techniques that are custom tailored to particular network topologies or to particular inference queries. In addition, approximate algorithms involving Monte Carlo techniques (e.g., Spiegelhalter and Thomas 1998) and variational methods (e.g., Jordan 1998) are proving useful in many applications.

4. *Learning DAG Models from Data*

There are a variety of methods for constructing DAG models from data or a combination of data and prior knowledge. We concentrate here on the Bayesian approach, but briefly consider other approaches.

First, suppose the structure $\mathbf{s}$ of the DAG model is known and we are interested in inferring the local distributions from our prior knowledge and a set of data $\mathbf{d} = (\mathrm{x}^1, \ldots, \mathrm{x}^N)$ considered to be a sample from an infinitely exchangeable sequence. For simplicity, we assume that each local distribution is parametric so that the data form a random sample from

$$p(\mathbf{x}|\boldsymbol{\theta}_s, \mathbf{s}) = \prod_{i=1}^{n} p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{s}) \tag{5}$$

where $\boldsymbol{\theta}_s = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ are the parameters of the local distributions.

If the parameters in $\boldsymbol{\theta}_s$ are mutually independent and there are no missing data, then the parameters in $\boldsymbol{\theta}_s$ are independent *a posteriori* and we can infer

each local distribution separately. Because a local distribution can be thought of as a probabilistic classification or regression model, learning the local distributions of a DAG model under the assumptions above is a familiar task. Examples of classification/regression models that have been used in conjunction with DAG models include probabilistic decision trees, generalized linear models, neural networks, and probabilistic support-vector machines.

One class of local distributions used frequently are distributions from the exponential family. For example, when X_i and its parents are discrete, $p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{s})$ can be a set of mutually independent multinomial distributions, one for each instance of the parents of X_i. When X_i is continuous, $p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{s})$ can be a set of linear regressions with mutually independent error terms, one for each instance of the discrete parents of X_i. In these cases, the posterior uncertainty can be determined and integrated over in closed form (e.g., Heckerman and Geiger 1995).

When the parameters in $\boldsymbol{\theta}_s$ are not mutually independent, when there are missing data (and/or latent variables), when local distributions from outside the exponential family are used, or when complex expectations over posterior uncertainty are required, closed-form expressions are unavailable and Markov Chain Monte Carlo methods are often employed. The BUGS system (Spiegelhalter and Thomas 1998) is a popular software tool designed for these situations. If maximum-likelihood or maximum *a posteriori* (MAP) approximations are sufficient, then optimization-based algorithms such as Expectation–Maximization and gradient descent typically are used.

Now, let us suppose that we are also uncertain about the structure of the DAG model. Again concentrating on the Bayesian approach, we model this uncertainty with a probability distribution $p(\mathbf{s})$ over the possible structures (see e.g., *Bayesian Statistics*). Given data $\mathbf{d}$, we update these uncertainties using Bayes's rule

$$p(\mathbf{s}|\mathbf{d}) \propto p(\mathbf{s})\, p(\mathbf{d}|\mathbf{s}) = p(\mathbf{s}) \int p(\mathbf{d}|\boldsymbol{\theta}_s, \mathbf{s})\, p(\boldsymbol{\theta}_s|\mathbf{s}) d\boldsymbol{\theta}_s \quad (6)$$

where $p(\mathbf{d}|\mathbf{s})$ is the marginal likelihood of the data given the model structure. Given the posterior uncertainties, we can either 'construct' a single DAG model by—say—selecting the most likely structure, or average over some or all of the model structures.

Under the assumptions described above—(a) mutual independence of the parameters in $\boldsymbol{\theta}_s$, (b) complete data, and (c) local distributions from the exponential family—the marginal likelihood can be computed efficiently and in closed form (e.g., Heckerman and Geiger 1995). Furthermore, simple heuristic search methods, such as greedy hill climbing, can be used to find a structure whose posterior is relatively large. In applications where computational efficiency is important such as data mining, where both the number of domain variables and the number of data samples is large, these techniques are used extensively (Heckerman 1996).

There are several non-Bayesian methods for learning structure from data. One set of techniques mimics Bayesian model selection, but incorporates non-Bayesian criteria for selection such as Akaike's information criterion (AIC). Another, developed independently by Pearl (2000) and Spirtes et al. (2001) uses local non-Bayesian tests for conditional independence together with search to identify model structures compatible with the data.

5. *Pointers to the Literature*

Perhaps the first person to use DAG models was Wright (1921),who employed what are now called structural equation models for the analysis of crop failure. Good (1961) also developed the representation. In both cases, researchers attributed causal semantics to DAG models. In fact, this model class remains quite popular for encoding and reasoning about causal interactions (Glymour and Cooper 1999, Pearl 2000, Spirtes et al. 2001; see *Latent Structure and Casual Variables*).

In addition to directed models, researchers have developed and used graphical models containing undirected edges, bidirectional edges, and mixtures of directed and undirected edges (e.g., Wermuth 1976, Whittaker 1990, Lauritzen 1996, Jordan 1998). These representations are discussed in the article *Graphical Models: Overview*.

Finally, there are numerous software packages available for the knowledge-based and data-based construction of DAG models. A summary is provided at http://http.cs.berkeley.edu/murphyk/Bayes/bnsoft.html.

See also: Graphical Methods: Presentation; Graphical Models: Overview

Bibliography

Chickering D 1995 A transformational characterization of equivalent Bayesian network structures. In: Besndard P, Hanks S (eds.) *Proc. 11th Conf. Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Mateo, CA, pp. 87–98

Cooper G 1990 Computational complexity of probabilistic inference using Bayesian belief networks (Research note). *Artificial Intelligence* **42**: 393–405

Cowell R, Dawid A P, Lauritzen S, Spiegelhalter D 1999 *Probabilistic Networks and Expert Systems* (*Statistics for Engineering and Information Science*). Springer Verlag, New York

Glymour C, Cooper G (eds.) 1999 *Computation, Causation, and Discovery*. MIT Press, Cambridge, MA

Good I 1961 A causal calculus (I). *British Journal of Philosophy of Science* **11**: 305–18. Also in Good I 1983 *Good Thinking: The Foundations of Probability and its Applications*. University of Minnesota Press, Minneapolis, MN, pp. 197–217

Heckerman D 1996 Bayesian networks for data mining. *Data Mining and Knowledge Discovery* **1**: 79–119
Heckerman D, Geiger D 1995 Learning Bayesian networks: A unification for discrete and Gaussian domains. In: *Proc. 11th Conf. on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, pp. 274–84. See also Technical Report TR-95-16, Microsoft Research, Redmond, WA, February 1995
Howard R, Matheson J 1981 Influence diagrams. In: Howard R, Matheson J (eds.) *Readings on the Principles and Applications of Decision Analysis*. Strategic Decisions Group, Menlo Park, CA, Vol. II, pp. 721–62
Jordan M (ed.) 1998 *Learning in Graphical Models*. Kluwer, Dordrecht
Lauritzen S 1992 Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association* **87**: 1098–108
Lauritzen S 1996 *Graphical Models*. Clarendon Press, Oxford, UK
Lauritzen S, Spiegelhalter D 1988 Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B* **50**: 157–224
Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA
Pearl J (ed.) 2000 *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK
Shachter R 1988 Probabilistic inference and influence diagrams. *Operations Research* **36**: 589–604
Spiegelhalter D, Thomas A 1998 Graphical modeling for complex stochastic systems: The BUGS project. *IEEE Intelligent Systems and their Applications* **13**: 14–5
Spirtes P, Glymour C, Scheines R 2001 *Causation, Prediction, and Search*, 2nd edn. MIT Press, Cambridge, MA
Wermuth N 1976 Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**: 95–108
Whittaker J 1990 *Graphical Models in Applied Multivariate Statistics*. Wiley, New York
Wright S 1921 Correlation and causation. *Journal of Agricultural Research* **20**: 557–85

D. Heckerman

Bayesian Statistics

Bayesian statistics refers to an approach to statistical inference characterized by two key ideas: (a) all unknown quantities, including parameters, are treated as random variables with probability distributions used to describe the state of knowledge about the values of these unknowns, and (b) statistical inferences about the unknown quantities based on observed data are derived using Bayes' theorem (described below). The Bayesian approach shares many features with the traditional frequentist approach to inference (e.g., the use of parametric models, that is to say, models dependent on unknown parameters, for describing data) but differs in its reliance on probability distributions for unknowns (including the parameters). The traditional approach to inference relies on the repeated sampling distribution of the data for a fixed but unknown parameter value, essentially asking what would happen if many new samples were drawn; the Bayesian approach treats the parameter as a random variable and assigns it a probability distribution. Qualitatively, the Bayesian approach to inference begins with a probability distribution describing the state of knowledge about unknown quantities (usually parameters) before collecting data, and then uses observed data to update this distribution. In this article the basic elements of a Bayesian analysis are reviewed: model specification, calculation of the posterior distribution, model checking, and sensitivity analysis. Additional sections address the choice of prior distribution, and the application of Bayesian methods. Additional details about most of the topics in this article can be found in the books by O'Hagan (1994), Gelman (1995), Carlin and Louis (2000), and Gilks et al. (1998).

The earliest developments related to the application of probability to questions of inference date to the contributions of Bayes and Laplace in the second half of the eighteenth century (Stigler 1986). Prior to that point researchers focused on the traditional pre-data probability calculations, i.e., given certain assumptions about the random process, what is the probability assigned to various possible outcomes for a variable in question? Bayes and Laplace receive independent credit for 'inverting' the probability statement to make probability statements about parameter values, given observed data values. There was little activity after that time, though some individuals, notably the physicist Jeffreys (1961), continued to develop the field of inductive inference. Modern Bayesian inference developed in the period around and after World War II (e.g., see *Statistics: The Field*). The name Bayesian inference replaces 'inverse probability' only at this later time. Some key contributions: Savage (1954) is an influential book using decision theory to justify Bayesian methods; de Finetti (1974) contributed crucial work concerning the role of exchangeability (which plays a role analogous to that of independent identically distributed observations in the traditional frequentist approach to inference); Raiffa and Schlaifer (1961) developed the use of conjugate distributions in detail; Lindley (1971, 1990) and Box and Tiao (1973) contributed greatly to the popularization of the approach. Most recently, the last decade of the twentieth century saw the discovery (or rediscovery) of computational algorithms that make it possible to apply the Bayesian approach to a wide array of scientific fields.

1. Model Specification

The basic ideas needed to carry out a Bayesian analysis are most easily described with a simple and general notation. In this article y denotes all of the data

collected and θ denotes all of the associated unknown quantities. The unknown quantities are assumed to be parameters in a probability model. There are, of course, limits to what can be done using this simple notation. For example, there may be covariates x in the data set, variables that help explain y but are not modeled as random variables. Similarly the unknowns, θ, may include missing data, or latent variables in addition to model parameters. It is straightforward to extend the basic notation to accommodate such situations.

The specification of a Bayesian model for data y requires two probability distributions, a probability distribution for the data given (or conditional on) the parameters θ, and a probability distribution for the parameters. The distribution of the data conditional on the parameters is denoted $p(y \mid \theta)$, and identified as the data distribution, the data model, the sampling distribution, or the likelihood function. The latter two names serve as a reminder that this portion of the model is also used in the traditional approach to statistics. The distribution for the parameters θ, denoted $p(\theta)$, is known as the prior distribution. It can be thought of as describing the relative frequencies for different values of the model parameters before considering any of the data values. For the initial part of this article, it is assumed that a prior distribution has been specified. Details related to the choice of prior distribution are considered in Sect. 4.

In practice, the data distribution can have several components describing a number of variables of interest. Moreover, the prior distribution for θ is often specified hierarchically with the distribution of parameters θ depending on still other parameters. The simple description provided here is convenient for demonstrating the Bayesian approach. Sect. 5 briefly describes applications of Bayesian methods in more realistic models.

Example. The following very simple example will be carried through the article. Suppose that a study is carried out to determine the mean intelligence score (using a standardized scale) for a population of interest, perhaps siblings of children with a psychological disorder. A random sample of n such children is obtained. Let $y_1, \ldots, y_n$ denote the test scores for the n children. Assume that test scores in this population can be modeled as coming from a normal distribution with unknown mean μ. For convenience it is assumed that the standard deviation σ is known. (This assumption is generally avoided in practice but is helpful here.) Thus the data model is the traditional model of independent identically distributed random variables, given the unknown mean μ. Under the Bayesian paradigm the unknown mean μ is also a random variable. To complete the specification of the model, a probability distribution, the prior distribution, for the unknown parameter μ must be specified. Suppose the μ is assumed to follow a normal distribution with mean μ_0 and variance τ^2 where μ_0 and τ^2 are specified constants. This means that, *a priori*, the value of μ is expected to be near μ_0, i.e., the probability under the prior distribution is 0.95 that μ is between $\mu_0 - 2\tau$, and $\mu_0 + 2\tau$. This prior distribution is merely one possible choice; others are considered in Sect. 4.

2. *Calculation of the Posterior Distribution*

The information about θ contained in the prior distribution, and the information about θ contained in the data y, are combined in forming the posterior distribution $p(\theta \mid y)$. This distribution, the conditional distribution of the unknown quantities given the observed data, is obtained directly using the laws of probability, namely Bayes' theorem,

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} = \frac{p(y \mid \theta)p(\theta)}{\int p(y \mid \theta)p(\theta)\,\mathrm{d}\theta} \qquad (1)$$

The posterior distribution is fundamental to Bayesian inference for θ. The Bayesian approach makes probability statements concerning the unknown quantities, after fixing the things that have been observed at their known values. Note that Bayesian methods differ from traditional methods which are justified by repeated sampling considerations—only the sample at hand is considered relevant in the Bayesian approach.

The posterior distribution describes the state of knowledge about θ after observing the data. Graphical displays or numerical summaries of the posterior distribution provide the basis for inferential statements about the parameters. A point estimate is a one-number summary of the posterior distribution, e.g., the posterior mean or the posterior median. To derive a single optimal point estimate it is necessary first to specify a loss function that gives the cost of estimation errors. The optimal point estimate is the value that minimizes the expected loss (computed from the posterior distribution); see *Decision Theory: Bayesian*. The posterior distribution allows us to identify intervals that contain a given parameter with any specified probability; such intervals are analogous to the standard confidence intervals. One difference is that the posterior intervals have a natural probabilistic interpretation; they contain the unknown quantity with the specified probability. Formally, probability statements are not possible for a single instance of the classical confidence intervals, where the stated confidence level describes behavior over many repeated samples. Tests of hypotheses are possible using the Bayes factor, a ratio that measures the degree to which two competing models explain the observed data.

In relatively simple problems, it is possible to analytically determine the form of the posterior distribution and subsequently obtain statistical inferences. In more sophisticated models it can be quite difficult to determine the form of the posterior dis-

tribution. For those cases it is sometimes possible to approximate the posterior distribution, e.g., by a normal distribution. One important approximation approach involves studying the posterior distribution by generating a sample of simulations from the distribution (see *Monte Carlo Methods and Bayesian Computation: Overview*), thereby obtaining an empirical approximation to the posterior distribution. Then inferences are drawn empirically using the simulations.

Example. For normally distributed measurements with mean μ, and with a normal prior distribution for the parameter μ, the posterior distribution is also normal with

$$\text{mean} = \left(\frac{n/\sigma^2}{(n/\sigma^2)+(1/\tau^2)} \right) \bar{y} + \left(\frac{1/\tau^2}{(n/\sigma^2)+(1/\tau^2)} \right) \mu_0 \quad (2)$$

and

$$\text{variance} = \frac{1}{(n/\sigma^2)+(1/\tau^2)} \quad (3)$$

where $\bar{y}$ is the mean of the sample of observations. It is possible to interpret the posterior mean as a weighted average of the sample mean, $\bar{y}$, and the prior mean, μ_0, with the weight attached to each piece related to the precision (reciprocal of the variance) of that source of information. If the prior information is extremely precise (τ^2 small), then the posterior mean for the unknown μ will be heavily influenced by the prior mean (μ_0). If on the other hand the prior information is vague (τ^2 large), then the posterior mean will be primarily determined by the data. The precision of the posterior distribution is just the sum of the precisions of the prior distribution and the data distribution. It is also noteworthy that for a given choice of prior distribution, as the sample size becomes large, the posterior distribution does not depend on the prior distribution. In fact, for large samples, the Bayesian posterior distribution resembles the sampling theory distributional result that leads to classical statistical procedures.

The simple example demonstrates the use of a conjugate family of prior distributions, where the prior distribution (a normal distribution) and the likelihood function (based on a normal distribution) combine to yield a posterior distribution in the same family as the prior distribution (a normal distribution). The interpretation of the posterior distribution is especially convenient for conjugate families; the posterior distribution can be seen to combine the prior information and the information provided by the data. In fact this result holds more broadly than with conjugate families; a Bayesian analysis represents a compromise between the prior information available and the information derived from the data. The more data that is available, the less influence the prior distribution has on the form of the posterior distribution. One complication is that, in sophisticated models, the amount of data for different parameters will vary and consequently so does the importance of different components of the prior distribution.

3. *Model Checking and Sensitivity Analysis*

It is essential in all model-based statistical inference to examine the validity of the assumptions made in the model. For the Bayesian approach to inference the assumptions include the form of the prior distribution, the form of the data model, and any other assumptions embodied in the data model (independence, constant variance, etc.). Model checking can take the form of a sensitivity analysis where the prior distribution (or other aspect of the model) is varied and the effect on the posterior inferences assessed. Model checking can also be done by embedding the original model into a bigger family of models, e.g., a normal model might be replaced by the Student-t family of models that includes the normal distribution as a limiting case. Finally, model checking can resemble model checking in traditional statistics with residuals (now functions of the parameters) and diagnostic measures used to assess the fit of the model.

4. *The Prior Distribution*

For many users of statistical methods the largest question about the use of Bayesian methods concerns the choice of prior distribution. One key result, mentioned earlier, is that in large samples the prior distribution becomes irrelevant. For the normal distribution example, as the sample size become large, the properties of the posterior distribution do not depend on the parameters of the normal prior distribution. In the limit, the posterior distribution would behave as if there were no prior information; $p(\mu|y)$ is the normal distribution with mean $\bar{y}$ and variance σ^2/n. This limiting result is similar to the usual sampling distribution result, except of course that in the Bayesian approach μ is random and $\bar{y}$ is fixed. A consequence of this result is that although the conclusions obtained from a Bayesian analysis in the normal example may differ from those obtained using traditional methods, the two approaches will tend to agree asymptotically. This result holds more generally as well. Of course, data analysts can not always rely on large sample results, so a review of methods for selecting prior distributions is provided next.

The subjective approach to choosing a prior distribution is to take the prior distribution to be an

honest assessment of prior beliefs about the values of the model parameters. Although people are sometimes hesitant to supply such subjective prior distributions it is often the case that some prior information is available. In the example, psychological or biological theory may suggest plausible values for the mean test score in the subpopulation of interest. If it is possible to specify a prior distribution, then the Bayesian paradigm provides the way to update that prior distribution given observed data. In multiparameter problems one can argue that prior information is not as rare as one might initially think. For example, prior information may suggest that a set of parameters can be treated as a sample from a common population (as in a random-effects or hierarchical model). Tools for helping researchers develop prior distributions is one area of Bayesian research (e.g., see *Elicitation of Probabilities and Probability Distributions*).

Often the choice of a prior distribution is made primarily because of considerations of computational convenience. Conjugate families, families of prior distributions that combine with a given data distribution to produce posterior distributions in the same family as the prior distribution, are useful in this regard. As described earlier, conjugate prior distributions are convenient to use because they make calculations easy, and because they can be easily interpreted. Conjugate prior distributions are capable of supporting a variety of prior opinions (e.g., by making different choices of the parameters within the conjugate family) but of course not all. Choosing a specific prior distribution from the conjugate family remains a bit of a stumbling block for some. Empirical Bayes methods use the data to help choose the parameters of the prior distribution; this approach may be applied in more complex models (e.g., analyses that incorporate random effects).

The desire to avoid using subjective prior information and/or arbitrary distributional forms has motivated some to use vague or 'noninformative' prior distributions. Formally, a vague prior distribution is one that assigns roughly equal probability to a wide range of possible values. In the normal example, a prior distribution with large variance would be considered a vague prior distribution. A vague prior distribution will not have a strong influence on the ultimate form of the posterior distribution. In this way, the use of vague or noninformative prior distributions represents the pursuit of a form of objective Bayesian inference. One difficulty is that, in the limit, vague prior distributions may become so vague as to no longer be proper distributions (they may not integrate to one!). It is 'legal' to use such improper prior distributions as long as it is verified that the resulting posterior distribution is a proper distribution. Improper prior distributions are popular because sometimes they appear to be noninformative in the sense that they rely only on the likelihood. It is probably best to think of improper prior distributions as approximations to real prior distributions. If the improper prior distribution leads to a proper posterior distribution and sufficiently accurate conclusions, then one might use the results from this analysis without needing to work any harder to select a more appropriate prior distribution.

A final word is in order here. It is disturbing to some that individuals with different prior distributions will obtain different analyses. Of course this occurs every day in the real world when individuals with different information make different decisions. There, as in the normal example, as everyone obtains the same large pool of information, a consensus is reached independent of the prior distribution.

5. *Bayesian Methods in Applications*

The relatively simple example discussed here, normally-distributed test scores with a normal prior distribution for the population mean, is useful for demonstrating the Bayesian approach; however, it is not a particularly compelling application. Standard methods for such data are well known and widely applied. More to the point, the posterior distribution of μ in the example resembles the frequentist sampling distribution that underlies standard methods when the sample size is large or the prior distribution is vague. There can be large differences between the two approaches when an informative prior distribution is used, but most investigators appear to avoid reliance on informative prior distributions. How well do Bayesian methods do in more substantial applications? Fortunately, there is a large body of literature on such applications in diverse fields such as medicine, archeology, and political science.

Two social science applications are described in some detail next. Raftery (1995) introduced the use of Bayesian hypothesis testing via Bayes factors to sociologists in the 1980s; he developed the simple BIC (Bayesian information criterion) approximation which made Bayesian model selection as straightforward as other methods. This work was motivated by a number of concerns about traditional significance tests and the P-values used to summarize them. Specifically a number of researchers had found that P-values were not satisfactory in large sample sizes when all models tend to be rejected, nor could they accommodate situations in which a large number of models were under consideration. Bayes factors or BIC addressed these issues by relying on Bayesian principles and calculating the quantity that is most relevant the marginal probability of the data given the model.

A recent application in psychology by Rubin and Wu (1997) models responses in schizophrenic eye-tracking experiments using a sophisticated mixture model. Their model reflects psychological theory, in allowing some schizophrenic subjects to be susceptible

to disruptions that delay responses while others are not. The indicators to which subjects are susceptible are not observed, but are incorporated in the model. In the same manner, unobserved indicators identify which specific trials are affected by disruptions for those individuals that are susceptible. The model also incorporates between-subject factors like gender and within-subject factors such as the type of eye-tracking task. There appear to be at least three benefits of the Bayesian approach in this application: model specification is reasonably straightforward using conditional distributions and incorporating latent variables; it is possible to obtain inferences without relying on large sample approximations; and methods for model checking are easily developed for a model which does not satisfy the regularity conditions required for traditional tests.

The specific details of every application differ, but a few points occur often enough in serious Bayesian applications to be noteworthy. The models underlying such applications often incorporate both fixed and random effects (see *Hierarchical Models: Random and Fixed Effects*), with the random effects given a prior distribution that depends on additional parameters. In addition realistic applications often need to accommodate cases where some data-providing units are missing observations for some variables (see *Statistical Data, Missing*) or for all variables (see *Factor Analysis and Latent Structure: Overview*). In those cases the missing values are treated as unknown quantities just like the parameters of the underlying model. Finally, these applications are also characterized by the need for advanced computational methods for computing or approximating the posterior distribution (see *Monte Carlo Methods and Bayesian Computation: Overview*). Specific examples of substantive applications can be found in Gelman (1995), Carlin and Louis (2000), Gilks et al. (1998), and the published proceedings of the *Case Studies in Bayesian Statistics* workshops (e.g., Bernado et al. 1998) and of the Valencia International Meetings (published under the title *Bayesian Statistics*) (e.g., Bernado et al. 1999).

The arguments favoring the use of the Bayesian approach for data analysis follow from its use or probability distributions to describe uncertainty about unknown quantities. There is a natural probability-based interpretation for Bayesian results (e.g., interval estimates) and great flexibility in the types of inferences that can be obtained (e.g., one can easily obtain a posterior distribution on the ranks of a set of parameters). The reliance on formal probability distributions also means that it is possible to draw valid Bayesian inferences in finite samples without relying on large sample results.

Bibliography

Bernado J M, Berger J O, David A P, Smith A F M 1999 *Bayesian Statistics 6*. Oxford University Press, London

Box G E P, Tiao G C 1973 *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA

Carlin B P, Louis T A 2000 *Bayes and Empirical Bayesian Methods for Data Analysis*. Chapman and Hall, London

de Finetti B 1974 *Theory of Probability: A Critical Introductory Treatment*. Wiley, New York

Gelman A 1995 *Bayesian Data Analysis*. Chapman and Hall, London

Gilks W R, Richardson S, Spiegelhalter D J (eds.) 1998 *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London

Jeffreys H 1961 *Theory of Probability*, 3rd edn. Clarendon Press, Oxford, UK

Kass R E, Carlin B, Carriquiry A, Gelman A, Verdinelli I, West M 1998 *Case Studies in Bayesian Statistics*. Springer, New York, Vol. 4

Lindley D V 1971 *Bayesian Statistics, A Review*. Society for Industrial and Applied Mathematics, Philadelphia

Lindley D V 1990 The 1988 Wald Memorial Lectures: The present position of Bayesian statistics. *Statistical Science* **5**: 44–89

O'Hagan A 1994 *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*. Edward Arnold, London

Raftery A E 1995 Bayesian model selection in social research. *Sociological Methodology* **25**: 111–63

Raiffa H, Schlaifer R 1961 *Applied Statistical Decision Theory*. Harvard University, Boston, NJ

Rubin D B, Wu Y N 1997 Modeling schizophrenic behavior using general mixture components. *Biometrics* **53**: 243–61

Savage L J 1954 *The Foundations of Statistics*. Wiley, New York

Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, Cambridge, MA

H. S. Stern

Bayesian Theory: History of Applications

Bayesian theory concerns the generation of ideas about inductive or probabilistic reasoning that stem from studies of a certain well-defined probabilistic process that were performed over two centuries ago by the Reverend Thomas Bayes, FRS (1702–61). Bayes was a Unitarian minister who lived in Tunbridge Wells, England and whose mathematical accomplishments were sufficient to justify his election as Fellow of the Royal Society in 1742. The conclusions Bayes reached in his studies have led to what is now called Bayes' rule, a canon for the inductive reasoning process of revising our probabilistic beliefs about hypotheses in light of relevant evidence. In spite of lingering controversy about how some of the probabilistic ingredients in this rule are ever to be determined, Bayes' rule, and extensions of it, are now repeatedly applied in a variety of situations, many of which involve choices made in the face of uncertainty. Indeed, some persons believe that Bayes' rule is the canon or prescription for how we should revise our probabilistic beliefs about hypotheses in light of relevant evidence.

1. Foundations

Thanks to the works of George Boole (1815–64) and John Venn (1834–1923), we now have a very rich and flexible language for formulating and solving probabilistic problems that Bayes himself never employed. Consequently, Bayes' rule is commonly described in ways that do not appear in Bayes' original work. The probabilistic process Bayes studied involves what we now call a Bernoulli trial sequence in which n independent binary trials are performed, each of which results in either a success (S) or failure (F). Further, we suppose there is some probability of success $P(\mathrm{S}) = p$ that, under the independence assumption, remains the same from trial to trial. What we only know for sure is that p can take on any value in $0 < p < 1.0$. Under ordinary conceptions of probability, we suppose that $P(\mathrm{F}) = 1 - p$. Suppose we do not know the exact value of p and can only estimate its value by performing n Bernoulli trials. In these n trials, suppose we observe r successes. The problem Bayes addressed was how do we determine the probability that p lies between any two values, say $t > 0$ and $u < 1.0$, given that we have observed r successes in n Bernoulli trials? In modern terms we express Bayes' problem as the conditional probability

$$P(t \leqslant p \leqslant u \,|\, r \text{ successes in } n \text{ trials})$$

In this expression our hypothesis H is that

$$t \leqslant p \leqslant u$$

our evidence E consists of the observation of r successes in n trials. In modern terms, we require the conditional probability $P(H|E)$, called the posterior probability of H in light of evidence E.

Bayes' solution to the problem of finding $P(t \leqslant p \leqslant u \,|\, r \text{ successes in } n \text{ trials}) = P(H|E)$ troubled him and so he never tried to publish his findings; he died in 1762. But the executor of Bayes' will, Richard Price, discovered Bayes' unpublished paper, recognized its value, and communicated it to the Royal Society of London. Bayes' paper was subsequently published (Bayes 1763). Reprinted versions of Bayes' original paper can be found in Pearson and Kendall (1970) and in Barnard (1958). In his enthusiastic prefatory comments about Bayes' work, Richard Price observed that Bayes' studies represented the very first attempt to formalize the process of inductive reasoning (Bayes 1763, pp. 370–5).

One problem that troubled Bayes, and that has troubled so many persons since Bayes, is that, in order to determine the posterior probability that p lies between t and u, given r successes in n trials, we must also determine the probability that p lies between t and u, before we have obtained this evidence. In symbols, we must somehow determine $P(t \leqslant p \leqslant u) = P(H)$. Such probabilities are called prior probabilities. Under enumerative conceptions of probability, mentioned again later, $P(H)$ cannot be determined empirically; it can only be judged or estimated subjectively, something that many persons cannot accept. In Bayes' original problem, how are we to determine $P(H) = P(t \leqslant p \leqslant u)$ before we have made any empirical observations? In other words, in probabilistic reasoning about hypotheses, based on evidence, how do we get this inferential process started before we have obtained any specific evidence? Issues raised by this nagging question also influenced the directions taken in research on decisions made in the face of uncertainty.

In addition to probabilistic assessments, decisions also require assessments of the worth or value of consequences of decisions together with rules or algorithms for combining these two kinds of assessments. Credit for establishing the first formal rules for choice under uncertainty is usually assigned to Daniel Bernoulli (1700–82). Bernoulli and other mathematicians at the time were interested in determining rules for deciding which wagers or gambles a person ought to accept or reject (e.g., see Todhunter 1965, pp. 213–38). The rule Bernoulli advocated involves ordinary mathematical expectation for random variables:

$$E[X] = \sum_{i=1}^{t} x_i P(x_i)$$

where x_i are the possible values or random variable X and $P(x_i)$ is the probability of observing value x_i. These probabilities must sum to 1.0 across all the t possible values that X can take. So, in a simple gamble G, involving a win (W) or a loss (L), the expected dollar value of G can be stated as $E(\mathrm{G}) = \$_{\mathrm{W}} P(\mathrm{W}) + \$_{\mathrm{L}} P(\mathrm{L})$. According to this rule, it seems that we ought to accept gambles for which $E(\mathrm{G}) > 0$, reject gambles for which $E(\mathrm{G}) < 0$, and be indifferent regarding gambles for which $E(\mathrm{G}) = 0$, the latter being 'fair bets.' But Bernoulli recognized that different persons would value a certain amount of money in different ways. In particular, for you, the prospect of winning (or losing) 10 dollars depends upon how many dollars you already have. Bernoulli proposed that we consider what he termed the 'moral worth' of monetary consequences of gambles or wagers. He conjectured that the moral worth of a certain number of dollars varies as the logarithm of this number of dollars: i.e. that $U(\$) = k \operatorname{Log} \$$. Today, we would use the terms 'utility' or 'value' in preference to Bernoulli's term 'moral worth.' So, for the simple gamble just described, what we should determine is the expected utility of this gamble $E(U(\mathrm{G})) = U(\$_{\mathrm{W}})P(\mathrm{W}) + U(\$_{\mathrm{L}})P(\mathrm{L})$.

Any single stage decision (D) made in the face of uncertainty can be represented as a random variable X having, say, t possible outcomes or consequences (C_i)

that occur when hypothesis H_i is true. The expected utility of D can then be represented by the following expression, which will also illustrate how problems with Bayes' rule affected the acceptability of Bernoulli's canon for deciding upon which choices to accept or reject:

$$E[U(D)] = \sum_{i-1}^{t} U(C_i)P(H_i \mid E)$$

where H_i is the hypothesized event that produces consequence C_i, when action D is taken, and E represents the body of evidence bearing on the likeliness of the hypotheses being considered. In this expression it is vital to note that the probabilities $P(H_i|E)$ must sum to 1.0 across the t hypotheses being considered. The rule for choice implied by the above expression is simple: Choose that action or decision for which $E(U(\mathrm{D}))$ is greatest. In the situation just posed, posterior probabilities according to Bayes' rule are given as follows. Suppose we wish to determine $P(H_j|E)$, where H_j is any one of the t mutually exclusive and exhaustive hypotheses under consideration:

$$P(H_j|E) = \frac{P(H_j)P(E|H_j)}{\sum_{i-1}^{t} P(H_i)P(E|H_i)}$$

where E is some item or body of relevant evidence in the inference at hand. The denominator of the expression above is called a *normalizing constant* and it ensures that the sum of the posterior probabilities across the t hypotheses will equal 1.0. We must also have a distribution of prior probabilities $P(H_i)$ that also sums to 1.0 across the t hypotheses. The other terms in Bayes' rule, $P(E|H_i)$, are called likelihoods and, as discussed later, they grade the inferential force, weight, or strength of evidence E. These likelihoods are not required to sum to 1.0 across the t hypotheses. These facts about Bayes' rule are very important in showing why Bayes' rule for inductive reasoning along with Bernoulli's expected utility canon for choices under uncertainty were dismissed for nearly 200 years. Bernoulli's canon requires probabilities that normalize and likelihoods alone do not satisfy this requirement.

2. *Applications in Statistics*

In 1933, Kolmogorov provided the first axiomatic treatment of probability (Kolmogorov 1956). We all learn about his three basic axioms of probability:

For any event A, $P(\mathrm{A}) \geqslant 0$

For an event S that is sure to happen, $P(\mathrm{S}) = 1.0$

For any two mutually exclusive events A and B, $P(\mathrm{A \text{ or } B}) = P(\mathrm{A}) + P(\mathrm{B})$.

In this initial and in his later discussions of this formal system (Kolmogorov 1969), it is quite apparent that he considered only enumerative conceptions of probability, i.e., those in which probabilities are determined by counting.

There are two cases of enumerative probabilities, the first of which concerns aleatory probabilities or chances that arise in well-defined games of chance. In such situations we have a finite collection (S) of outcomes, each of which is assumed to be equally likely. In such instances, for any event A, defined as a subset of outcomes in S, $P(\mathrm{A}) = n(\mathrm{A})/\ n(\mathrm{S})$, where $n(\mathrm{A})$ is the number of outcomes favoring A and $n(\mathrm{S})$ is the total number of outcomes. The other case involves the relative frequencies encountered in statistical inferences. The relative frequency of event A, $f(\mathrm{A})$, is defined to be $f(\mathrm{A}) = n(\mathrm{A})/N$, where $n(\mathrm{A})$ is the number of observed cases of A and N is the number of observations taken. As recognized, $f(\mathrm{A})$ is just an estimate of actual $P(\mathrm{A})$ since N may be small relative to the total number of observations that might be taken. Observe that in either case we have restricted attention to processes that are replicable or repeatable. Under strict enumerative conceptions of probability, it is difficult to see how the prior probabilities required in Bayes' rule can ever be determined. But Bayes' rule follows as a direct consequence of the three axioms just mentioned along with Kolmogorov's own definition of a conditional probability: $P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A\&B})/P(\mathrm{B})$, where $P(\mathrm{B}) > 0$.

Generations of statisticians adhering to an enumerative and relative frequency view of probability have acknowledged the formal adequacy of Bayes' rule at the same time they have argued that this rule can never be applied on grounds that subjective judgments of probabilities have no place in fields such as science. Examples of this view are to be found in the works of Fisher (1959, pp. 8–16, 1960, pp. 6–7). One result has been that frequentistic statisticians have been forced, in statistical inference, to rely on the remaining terms in Bayes' rule, those concerning the likelihoods such as $P(E|H_i)$ mentioned above. Examples are to be found in any textbook that discusses statistical hypothesis testing from a frequentistic view. The trouble here is that these likelihoods are not required to sum to 1.0 across the t hypotheses being considered. Thus, by themselves they are not sufficient in any calculation of expected utilities according to Bernoulli's canon for choice.

The works of several individuals in quite different fields resulted in a renewed interest in Bayes' rule and in Bernoulli's expected utility canon for choice. In works by Ramsey (1990), de Finetti (1972), Savage (1972), and others, subjective, personal, or epistemic interpretations of probability were defended as being both reasonable and necessary in probabilistic reasoning. On this view of probability, resistance to the use of Bayes' rule in statistical and other contexts vanishes. In addition, the work of von Neumann and Morgen-

stern (1947) was instrumental in restoring interest in Bernoulli's expected utility canon for choice. They showed how, from a collection of what they regarded as axioms for asserting preferences in assessing the worth of consequences, Bernoulli's expected utility canon follows. One result of the works just mentioned is that there has emerged a very wide array of statistical inferential methods involving Bayes' rule that can be applied in nearly every inferential problem in statistics. Examples are to be found in the works of Winkler (1972), Box and Tao (1973), O'Hagan (1988), and Howson and Urbach (1989).

3. *Other Applications*

All statistical reasoning is probabilistic, but not all probabilistic reasoning is statistical. In many contexts people routinely make probabilistic judgments about events that are unique, singular, or one of a kind, and for which no relevant statistics exist. In short, there is a necessity for nonenumerative conceptions of probability. For example, we cannot play the world over again 1000 times to tabulate the number of occasions on which defendant committed the crime or a witness reported an event that actually occurred. Along with increased interest in applying Bayesian theory in statistics, applications of this theory began to be made in many contexts in which relevant statistics are difficult if not impossible to obtain. One of the first such applications was Ward Edwards's proposals for a PIP (probabilistic information processing) system (Edwards 1962). In such a system people make judgments of prior probabilities and likelihoods and computers calculate required posterior probabilities from these judgments. In its original forms, Edwards' PIP idea did not survive. One reason was the gradual recognition that, in other than contrived problem situations, all inferences based on evidence involve complex arguments or chains of reasoning linking the evidence to hypotheses of interest. Such inferences are said to be cascaded, catenated, or hierarchical. By such chains of reasoning we defend the three major credentials of evidence: its relevance, credibility, and inferential force.

At this point, great assistance in thinking about complex probabilistic reasoning was provided by the rich legacy of evidential scholarship and experience in the field of law. A US evidence scholar named John H. Wigmore was the first to study systematically the task of constructing defensible and persuasive arguments based on masses of evidence (Wigmore 1913, 1937). His studies provide the very first examples of what today are termed inference networks. Wigmore's inference networks are extremely useful in the study and analysis of the many recurrent and substance-blind forms and combinations of evidence that exist. The likelihood terms in Bayes' rule provide very useful and informative metrics for grading the inferential force of evidence. A very wide array of evidential and inferential subtleties lie just below the surface of even the simplest of probabilistic inferences. A recent work gives an account of these subtleties and how they can be captured in Bayesian terms (Schum 1994). The major vehicle for capturing evidential and inferential subtleties or complexities by Bayes' rule involves the concept of conditional nonindependence. At its simplest level this concept involves the idea that two or more things considered jointly can mean something quite different than they might do if considered separately or independently. Another recent work illustrated how evidential subtleties are captured in a Wigmorean and Bayesian analysis of the mass of trial and post-trial evidence in a celebrated US law case (Kadane and Schum 1996).

One of the most significant recent advances in Bayesian theory concerns probabilistic analyses of complex inference networks. In the work of Pearl (1988), Lauritzen and Spiegelhalter (1988), and others, various attempts have been made to develop computationally efficient means for propagating and aggregating large numbers of probabilities that are required in the analysis of complex inference networks. A number of very useful software systems, based on the formal developments just mentioned, are now available for such analyses and they are finding ready application in a variety of contexts including science, medicine, intelligence analysis, and business.

During the past three decades alternative formal systems of probability have been proposed including the Shafer–Dempster system of non-additive belief functions (e.g., Shafer 1976) and the Baconian system of probabilities (e.g., Cohen 1977). Probabilistic reasoning is a remarkably rich intellectual task and it is perhaps too much to expect that any one formal system of probability can capture all of this richness. One matter that divides opinion across the alternative formal systems concerns what is meant by the weight of evidence. In the theory of belief functions the weight of evidence is related to the support it provides hypotheses being considered. In the Baconian system the weight of evidence concerns how completely existing evidence covers recognized relevant questions that could be asked and how many of these questions remain unanswered by existing evidence.

Since the very beginnings of interest in the mathematical calculation of probabilities in the early 1600s, there have been skeptics who have argued that mathematical systems can never capture the true complexities of probabilistic reasoning in the variety of contexts in which it occurs. Though the Bayesian theory of probabilistic reasoning is not complete in answering all questions that arise during probabilistic reasoning, it is nevertheless capable of capturing a wide array of elements of complexity as they have been recognized recently in the emerging science of complexity (e.g., Cowan et al. 1994, Coveny and Highfield 1995). All complex processes are dynamic, hierarchi-

cal, and recursive, in which complex patterns of interaction occur among the elements of these processes. Phenomena emerging from such processes are the result of distinctly nonlinear combinations of these elements and cannot be inferred or predicted just from knowledge of the elements themselves. In linear models wholes are always equal to the sum of their parts and do not produce any surprises. As noted recently (Schum 1999, pp. 183–209), Bayes' rule incorporates these elements of complexity and, as a nonlinear model, it will produce many surprises. The results of Bayesian probability calculations made in complex inference networks do not always correspond to what intuition may suggest. In such cases we must either examine the adequacy of the argument structures to which Bayes' rule has been applied, or examine the adequacy of our intuitions. No complex mathematical system is guaranteed to produce results that always correspond to intuition.

See also: Bayesian Statistics; Decision Theory: Bayesian; Frequentist Interpretation of Probability; Game Theory and its Relation to Bayesian Theory; Neural Networks, Statistical Physics of; Probability: Interpretations; Utility and Subjective Probability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies

Bibliography

Barnard G A 1958 Studies in the history of probability are statistics. 9. Bayes, Thomas essay towards solving a problem on the doctrine of changes. *Biometrika* **45**: 293–5

Bayes T 1763 An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* **53**: 370–418

Box G E P, Tao G C 1973 *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA

Cohen L J 1977 *The Probable and the Provable*. Clarendon Press, Oxford, UK

Coveny P, Highfield R 1995 *Frontiers of Complexity: The Search for Order in a Chaotic World*. Fawcett-Columbine, New York

Cowan G A, Pines D, Meltzer D (eds.) 1994 Complexity: metaphors, models, and reality. *Proc. Santa Fe Institute Studies in the Sciences of Complexity*. Addison-Wesley, New York, Vol. XIX

de Finetti B 1972 *Probability, Induction and Statistics: The Art of Guessing*. Wiley & Sons, New York

Edwards W 1962 Dynamic decision theory and probabilistic information processing. *Human Factors* **4**: 59–73

Fisher R A 1959 *Statistical Methods and Scientific Inference*. Oliver & Boyd, Edinburgh, UK

Fisher R A 1960 *The Design of Experiments*. Hafner Publishing, New York

Howson C, Urbach P 1989 *Scientific Reasoning: The Bayesian Approach*. Open Court, LaSalle, IL

Kadane J B, Schum D A 1996 *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. Wiley & Sons, New York

Kolmogorov A N 1956 *Foundations of a Theory of Probability*. Chelsea Publishing, New York

Kolmogorov A D 1969 The theory of probability. In: Aleksandrov A, Kolmogorov A, Lavrent'ev A (eds.) *Mathematics: Its Content, Methods, and Meaning*. MIT Press, Cambridge, MA, Vol. 2, pp. 229–64

Lauritzen S L, Spiegelhalter D J 1988 Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society Series B Methodological* **50**: 157–224

O'Hagan A 1988 *Probability: Methods and Measurement*. Chapman & Hall, London

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kaufmann, San Mateo, CA

Pearson E, Kendall M (eds.) 1970 *Studies in the History of Statistics and Probability—a series of papers*. Charles Griffin, London

Ramsey F P 1990 Truth and probability. In: Mellor D H (ed.) *F. P. Ramsey: Philosophical Papers*. Cambridge University Press, Cambridge, UK, pp. 53–109

Savage L J 1972 *The Foundations of Statistics*, 2nd edn. Dover, New York

Schum D A 1994 *The Evidential Foundations of Probabilistic Reasoning*. Wiley & Sons, New York

Schum D A 1999 Probabilistic reasoning and the science of complexity. In: Shanteau J, Mellers B, Schum D A (eds.) *Decision Science and Technology: Reflections on the Contributions of Ward Edwards*. Kluwer Academic Publishers, Boston, MA, pp. 183–209

Shafer G A 1976 *Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ

Todhunter I 1965 *A History of the Mathematical Theory of Probability: From the Time of Pascal to that of Laplace*. Chelsea Publishing, New York

von Neumann J, Morgenstern O 1947 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

Wigmore J H 1913 The problem of proof. *Illinois Law Review* **8**: 77–103

Wigmore J H 1937 *The Science of Judicial Proof: As Given by Logic, Psychology, and General Experience and Illustrated in Judicial Trials*, 3rd edn. Little, Brown & Co, Boston, MA

Winkler R 1972 *Introduction to Bayesian Inference and Decision*. Holt, Rinehart, and Winston, New York

D. A. Schum

Beauvoir, Simone de (1908–86)

Novelist, philosopher, essayist, and autobiographer, the author of one of the founding texts of feminism in the twentieth century, Simone de Beauvoir can be considered to be one of the major intellectual figures of her generation. Born in Paris on January 9 1908 into a conservative middle-class milieu, she received the narrow primary and secondary education deemed appropriate for a marriageable girl of her class. However, after a sudden deterioration in her family's financial circumstances, she was permitted to study philosophy at the Sorbonne, and in 1929 she became the youngest person ever to succeed in the prestigious *agrégation de philosophie* examination, a success which enabled her to earn her own living from a series of teaching posts which she held first in Marseilles, then

in Rouen and, from 1936, in Paris. Occupying a series of cheap hotel rooms, she led a remarkably independent life, free from domestic obligations, and, despatching her teaching in the minimum time necessary, she spent her days writing in cafes and participating in a complex web of relationships, at the center of which was the couple she formed with Jean-Paul Sartre. The two never set up home together, setting their face against the idea of marriage and children, but referred to their close circle of former pupils and ex-lovers as 'the family.' The outbreak of the Second World War disrupted both this network of relationships and Beauvoir's assumption that her freedom was virtually total. As she writes in her autobiography she had to 'face the weight of reality.' Continuing her life in occupied Paris without Sartre she completed the manuscript of her first novel, *She Came to Stay* (1943), an exploration of the Hegelian dictum 'Each consciousness pursues the death of the other,' and went on to publish in quick succession *The Blood of Others* (1945), exploring some of the ethical and political dilemmas posed by the Resistance, and *All Men Are Mortal* (1946) which deals with the problem of the historical dimension of human existence. Sartre, meanwhile, had completed his philosophical essay *Being and Nothingness*, also published in 1943; in postliberation Paris the couple became famous figures as existentialism, as Sartre's philosophy came to be known, dominated the Parisian intellectual scene. In 1946 Beauvoir turned her attention to the situation of women, and in 1949 published the two volumes of radical and multidisciplinary thinking which constitute *The Second Sex*. Its frank treatment of sexuality and its critique of the maternal role expected of women made her a figure of some notoriety, and though she went on to write bestselling novels, and won the Goncourt prize for *The Mandarins* in 1954, her name was forever to be associated with her seminal work of feminist theory. From 1958 onwards she also began to publish volumes of autobiography, constructing an account of life as a twentieth-century intellectual woman followed by millions of readers worldwide. The 1950s also opened up a phase of active political engagement as she campaigned against French atrocities in Algeria during the Algerian war, supported the Maoist groups in the events of 1968, and, in the 1970s, joined with the newly formed MLF (Women's liberation movement) in campaigns on abortion, contraception, sexism, and violence against women. After Sartre's death in 1980, she published his letters to her, and her own letters both to Sartre and to her American lover Nelson Algren, were published in 1990 and 1997, respectively, after her own death in 1986. It is thus as philosopher, feminist, activist, novelist, memorialist, and correspondent that Beauvoir has made a deep mark on her era.

Nonetheless, it is principally as the author of *The Second Sex* that Beauvoir has come to be considered as one of the most significant feminist voices of the twentieth century. Beauvoir's study of women began life as an autobiographical question (what had it meant to her personally to be a woman), developed into an essay about the cultural myths which construct femininity and eventually became an ambitious multidisciplinary work, studying women's situation from a panoply of directions to demonstrate that 'one is not born, but rather becomes, a woman.' This crucial distinction between sex and gender was to open up the entire field of women's and gender studies. Beauvoir stresses in the introduction that her approach is ontological and existentialist: if there is no fixed human nature then neither can there be a fixed nature for women. This intellectual origin does indeed provide a strong impetus to the work, driven by the assumption that, like all human beings, women can and should shape their own destinies. However, the existentialist approach stresses individual responsibility rather than the way in which patriarchy functions as an institution and, taken to the extreme, it could suggest that women's oppression is a form of individual bad faith.

One of the strengths of *The Second Sex* is that although Beauvoir does not admit explicitly the extent to which social and economic analyses complement her ontology, she does in fact offer a balance between the weight of social, cultural, and economic oppression on the one hand, and the free construction of the subject on the other. Her argument that women historically have been positioned as the object or 'other' to men as subject is supported by a socially constructionist view of patriarchy.

The book begins with a series of attacks on the notion that there is a fixed destiny for women. In part 1 she investigates what biology, psychoanalysis, and Marxism appear to prescribe about female destiny to conclude that women are not defined or inevitably doomed to oppression by the laws of nature, by their psychoanalytic situation, or by economic history. The section on psychoanalysis attacks Freud for his definition of female sexuality in terms of penis envy, an attack that was responsible for a long feminist mistrust of Freud. The section on biology as destiny has proved particularly controversial since, despite insisting that female physiology offers no justification for the hierarchy of the sexes, Beauvoir paints a dispiriting picture of the daily experience of the body for women. In her account the female body is organized to such a high degree to ensure the reproduction of the species that it constitutes a burden against which women must continually struggle in their bid for freedom. Part 2 is devoted to a wide-ranging survey of women's historical role from Ancient Greece and Rome up to the 1940s. It describes women as fundamentally handicapped by maternity from the earliest times, stresses women's role as a medium of exchange, and analyses how tools of patriarchy such as Christianity and the feudal system have corralled the vast majority of women into the margins of historical narratives.

Moving on to the myths of femininity, in part 3, Beauvoir dissects a rich set of constructs ranging from the Christian images of Mary and Eve to the pervasive myth of feminine mystery. The section ends on a discussion of five male writers, showing how literature has served as a vehicle for the creation and circulation of myths about women. This analysis later became the inspiration for the work of Kate Millett and the branch of feminist literary criticism that reads male-authored texts with a deconstructive eye.

Parts 4 and 5 of the essay turn to the social and sexual roles which women are called upon to play, tracing the way in which the girl gradually becomes socialized into her role as wife and mother, under the eye of her own mother, the unwitting accomplice of patriarchy. Up to the age of puberty Beauvoir nevertheless credits the young girl with a strong sense of her own autonomy; it is at the moment when the traits of female sexuality install themselves in her body that the girl sees the social trap closing, according to Beauvoir. A section considering female sexual response and case histories of women's first experiences of sexuality follows. Drawing on the Kinsey Report, and on other studies of sexuality, Beauvoir gives serious and detailed consideration to female sexual pleasure, including a discussion of the supposed distinction between the vaginal and clitoral orgasm. She also devotes a chapter to lesbianism in which she rejects the idea that lesbianism is an anatomical or physiological destiny. On the contrary, she presents it as positive choice for women that can lead to an authentic exercise in freedom just as much as heterosexuality. Unsurprisingly, she does not produce a theory of lesbian identity, since she consistently presents identity as following from choice and acts, not as preceding action and being.

Part 5 examines the two central roles which patriarchy calls upon women to play: those of the wife and the mother. Romantic love is in itself a dangerous condition for women, Beauvoir argues, because it tends to encourage their dependency on men and discourage them from pursuing their own goals. Marriage she describes as based on an obscene principle, because it makes a duty of what should be a freely given exchange. Her attack on the role of wife is closely linked to the role of women as maintainor of the home—the thankless round of domestic tasks that have by their nature to be repeated endlessly without achieving any final aim are described graphically by Beauvoir, and the illusory joys of the gleaming interior are analyzed with penetrating insights. Motherhood, another concomitant of marriage, is presented as even more problematic. Pregnancy acts as a positive incitement to women to sink passively into a fixed destiny; after the birth, the mother–child relationship is described as a battlefield yielding few positive results for either party. She concludes that the upbringing of children care is unsafe within maternal hands and should be organized within a collective framework. Beauvoir's analyses of the dangers and social pressures of the roles of wife and mother bring sharply into focus her dual insistence on, on the one hand, the weight of social conditioning to which women are exposed, and, simultaneously, individual women's responsibility to resist such pressures. Women are both victims and complicitous.

The final two parts of the book are entitled 'Justifications' and 'Towards liberation.' In part 6, 'Justifications,' Beauvoir considers some of the false routes which women have taken in a doomed attempt to flee from their responsibility by glorying in their imprisonment. One route is that of narcissism, in which the woman cultivates her body in an impossible attempt to be simultaneously subject and object, simultaneously both source of the gaze and object of it. A second false route is that of mysticism, the attempt to achieve sovereign being through an identification with God—also, for Beauvoir, an impossibility. The third route is that of the *amoureuse*, the woman who makes a cult of her love for a man whom she has appointed as her savior. She aspires to authenticity by attempting to live vicariously through him. Beauvoir demonstrates both the philosophical impossibility of this stance and the inevitable destruction that it brings about of the relationship on which the *amoureuse* depends, as she becomes increasingly tyrannical.

Part 7, entitled 'Towards Liberation,' offers a synthesis of women's situation and a portrait of 'The Independent Woman.' Beauvoir shows how women's blocked situation even creates what we think of as the female character. The reader is urged to accept that femininity is a social product—if we can accept this, she argues, then we can change society to change the possibilities open to women. She insists on the importance of economic independence for women, and pleads for men and women to work together towards a better future, but she has no confidence that anything of substance can be achieved without a radical restructuring of society. The independent woman is not in a position to achieve much without the socialist revolution that Beauvoir assumes, in 1949, would create a society in which everyone would be free. This rather unsatisfactory ending is indicative of the exploratory, theoretical rather than campaigning spirit in which Beauvoir wrote the book; it was not until over 20 years later, in the context of French second wave feminism, that Beauvoir declared the necessity of actively pursuing the women's struggle independently of any broader political movement.

On its publication in French in 1949 the book caused moral outrage; the Catholic novelist François Mauriac, led a newspaper campaign to have it banned as pornography, and it was in fact put on the Index. Its status as a daringly sexy book written by a woman seems to have partly obscured its political thrust, and the French postwar climate was particularly inimical to the language of women's rights. In America the

book was translated in a truncated form by the zoologist Parshley, and it was this version, published in 1953 and still the only English language version available today, which a generation of American and English women including Kate Millet, Betty Friedan, Germaine Greer, and Anne Oakley read during the 1950s and 1960s. Through them the work began to exert its influence, although they did not always acknowledge their debt. Betty Friedan, for example, whose book *The Feminine Mystique* (1963), is often considered to have been the launch pad for the American women's movement, offered her readers a watered down version of Beauvoir's analysis of the housewife and has only one mention of her in the index, which refers to the charge by an American critic that Beauvoir 'didn't know what life was all about' and 'besides she was talking about French women.' Where Beauvoir offers a radical critique, insisting that the institution of the family must be destroyed, that women must become economically independent, and that a radical restructuring of society will need to take place, Friedan advises women to think about their personal growth. However, it is atleast arguable that in translating some of Beauvoir's analyses into terms acceptable to a more conservative audience, Friedan did act as a channel of communication between two cultures in which political radicalism historically have played very different roles.

The key phrase 'one is not born, but rather becomes, a woman' is, as Haraway (1991) writes, at the root of 'all the modern feminist meanings of gender' and has been developed in many different directions. Butler (1986), for example, in a reading of the phrase, stresses Beauvoir's use of the verb 'becomes,' arguing that it allows for a process of gender acquisition that is a daily act, an interplay between individual choice and acculturation of which 'styles of the flesh' are a constitutive part. This important article prepares the way for some of the influential theoretical texts which Butler has gone on to write in the 1990s: *Gender Trouble* (1990) and *Bodies That Matter* (1993) propose a performative theory of gender which build on and develop Beauvoir's original insight. Second-wave feminism in France had a complicated relationship with Beauvoir. The construction of 'French feminism' in the Anglo-Saxon world as being constituted uniquely by the theorists of difference, Cixous, Irigaray, and Kristeva, whose theories are at odds with Beauvoir's stress on individual agency and freedom, has tended to obscure the work of other French feminists such as Christine Delphy and Monique Wittig working in Beauvoir's radical tradition. Feminists reading *The Second Sex* 50 years after its first publication encounter a number of problems: the text's ethnocentricity, its tendency to regard masculinity as unproblematic, the distaste displayed for the female body, and the devaluation of the maternal are largely at odds with contemporary sensibility. Despite this, there is considerable evidence in the 1990s of a renewed interest in *The Second Sex*, as the feminism of difference finds itself in turn in difficulties. Thus, Moi (1994) writes: 'In an intellectual field dominated by identity politics, *The Second Sex* represents a real challenge to established dogmas: if we are to escape from current political and theoretical dead ends, feminism in the 1990s cannot afford to ignore Beauvoir's pioneering insights.'

The Second Sex, with its confidence in the possibility of individual agency and a freely chosen identity, combined with an awareness of how patriarchal structures are at work in the construction of female subjectivity, offers a third way through the choice between essentialism on the one hand and postmodern dissolution of the subject on the other. *The Second Sex* founded many of the feminist strands of scholarship that are being pursued—in women's history, in philosophy, in literary criticism, in social studies. It offers a model of the multidisciplinary approach that the study of women requires, and has made the discussion and theorization of gender possible. It was not intended originally as a manual of liberation or a political tool, but it became one through the impact it had on others. In her autobiography, Beauvoir describes her own astonishment at discovering, in her forties, that there was a way of looking at women with new eyes; that vision is one which has now become a part of our consciousness.

See also: Family and Gender; Feminist Epistemology; Feminist Movements; Feminist Theory; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Gender and Feminist Studies; Gender, Class, Race, and Ethnicity, Social Construction of; Gender Differences in Personality and Social Behavior; Gender History; Rationality and Feminist Thought; Sexuality and Gender; Social Class and Gender

Bibliography

de Beauvoir S 1943/1978 *She Came to Stay* [trans. Moyse Y, Senhouse R]. Penguin, Harmondsworth, UK

de Beauvoir S 1945/1978 *The Blood of Others* [trans. Moyse Y, Senhouse R]. Penguin, Harmondsworth, UK

de Beauvoir S 1947/1976 *The Ethics of Ambiguity* [trans. Fretchman B]. Citadel Press, New York

de Beauvoir S 1948/1952 *America Day by Day* [trans. Dudley P]. Duckworth, London

de Beauvoir S 1949/1986 *The Second Sex* [trans. Parshley H]. Penguin, Harmondsworth, UK

de Beauvoir S 1954/1986 *The Mandarins* [trans. Friedman L]. Fontana, London

de Beauvoir S 1958/1987 *Memoirs of a Dutiful Daughter* [trans. Kirkup J]. Penguin, Harmondsworth, UK

de Beauvoir S 1960/1986 *The Prime of Life* [trans. Green P]. Penguin, Harmondsworth, UK

de Beauvoir S 1990/1991 *Letters to Sartre* [trans. Hoare Q]. Arcade, New York

de Beauvoir S 1998 *Beloved Chicago Man. Letters to Nelson Algren*. Gollancz, London

Bergoffen D 1996 *The Philosophy of Simone de Beauvoir. Gendered Phenomenologies, Erotic Generosities.* State University of New York Press, Albany, NY
Butler J 1986 *Sex and Gender in Simone de Beauvoir's 'The Second Sex.'* Yale French Studies, pp. 35–49
Fallaize E (ed.) 1998 *Simone de Beauvoir. A Critical Reader.* Routledge, London
Haraway D J 1991 *Simians, Cyborgs and Women. The Reinvention of Nature.* Free Association Books, London, p. 131
Lundgren-Gothlin E 1996 *Sex and Existence. Simone De Beauvoir's The Second Sex.* Athlone, London
Moi T 1994 *Simone de Beauvoir. The Making of an Intellectual Woman.* Blackwell, Oxford, UK
Simons M 1995 *Feminist Interpretations of Simone de Beauvoir.* Penn State University Press, Pennsylvania, PA

E. Fallaize

Behavior Analysis, Applied

Applied behavior analysis consists of the use of basic behavioral processes, research methods, and derivative procedures, in order to prevent and alleviate problems of social importance. Applied behavior analysis is accountable on seven dimensions: Its procedures are applied to problems of relatively immediate social importance; its behavioral measures are valid and reliable; its procedures are described in sufficient technological detail for replication; its research methods are analytic; its effectiveness is socially significant; its generality is demonstrated across time, settings, and behaviors; and it is relevant to an overall conceptual system of behavior. This article presents the foundations of applied behavior analysis, its main techniques, and the principal areas of application. Innovative procedures and techniques are pointed out.

1. *What is Applied Behavior Analysis?*

Applied behavior analysis has been defined as the application of the methods, procedures, and techniques derived from the experimental analysis of behavior to a wide range of clinical and social institutions. The main assumption of behavior analysis is that behavior is a function of its consequences. Environmental variables are utilized in order to produce behavioral changes. A wide range of intervention techniques has been developed based on the principles of laboratory research, such as stimulus control, reinforcement, extinction, and punishment.

In the context of applied behavior analysis, the term 'behavior,' is interpreted broadly to include overt responding and also covert responses (cognitions, emotions). Behavior is defined as everything that the organism does or says, in the case of human beings. Behavior is a broad term, with different ways of being interpreted. In applied behavior analysis it includes muscular contractions, visceral secretions, cognitive factors, emotional responses, language, and social behavior.

Applied behavior analysis is mainly associated with 'Skinnerian' (i.e., operant) psychology. The term is contrasted with behavior therapy, behavior modification, behavioral engineering, cognitive behavior therapy, multimodal therapy, social learning, and others. The differences are in some cases just a matter of terminology, and there is not complete consensus in relation to the use of each term.

In applied behavior analysis, the emphasis is on a body of experimental work dealing with the relationships between changes in the environment and changes in the subject's responses. Great importance is given to the experimentally established procedures and principles. The applications to socially relevant issues are considered to be derived from basic laboratory research. Clinical problems are studied by means of experimental research studies.

The model of applied behavior analysis is psychological rather than medical. The emphasis on behavior (maladaptative in the case of clinical populations) implies that the therapist is not interested in underlying or disease factors that 'cause' the symptoms. In the case of clinical or maladaptative behavior, the term 'behavior therapy' centers on behavior that is observable and definable.

Applied behavior analysis can be traced back to several conceptualizations and experimental strategies. Particularly important is the Skinnerian analysis of behavior, leading to behavioral shaping, token economies, and environmental design. The emphasis upon the application of experimental methodology to individual cases is a distinctive feature of applied behavior analysis. Social learning, modeling, and classical conditioning procedures are also taken into account.

It is possible to state that the main techniques used in applied behavior analysis are operant conditioning, extinction, systematic desensitization, modeling, self-control procedures, and techniques for the modification of cognitions—a wide range of procedures that are tailored to the particular individual, the specific situations, and the behaviors to be modified.

2. *Behavior Analysis and Applied Behavior Analysis*

Behavior analysis—the natural science from which applied behavior analysis is derived—comprises three subdisciplines: the experimental analysis of behavior for basic research elucidating fundamental behavioral processes; applied behavior analysis for the application of these processes, derivative technologies, and research methods for clinical and community

problems; and the conceptual analysis of behavior, for historical, philosophical, theoretical, and methodological investigations.

The experimental analysis of behavior (EAB) involves conceptual analysis as a matter of course, that is, ontological assumptions. The EAB seeks: (a) to describe functional relations between classes of responses and classes of stimuli, and (b) to demonstrate the reliability and generality thereof, and thus their lawfulness. The dependent variable is the probability of an organism's response as a function of independent variables, which occur in real time.

On the other hand, applied behavior analysis employs basic behavioral processes, research methods, and derivative procedures, in order to prevent and alleviate problems of social importance. Applied behavior analysis is accountable on seven dimensions: Its procedures are applied to problems of relatively immediate social importance; its behavioral measures are valid and reliable; its procedures are described in sufficient technological detail for replication; its research methods are analytic; its effectiveness is socially significant; its generality is demonstrated across time, settings, and behaviors; and it is relevant to an overall conceptual system of behavior (Baer et al. 1987). Applied behavior analysis has derived from the EAB with 'their procedual emphases of reinforcement, punishment, and discriminative-stimulus contingencies as behavior-analytic environmental variables, their reliance on single-subject designs as the formats of analysis and proof, and their consistent use of the Skinner box as their arena' (p. 313). However, in the last decades, applied behavior analysis has been characterized by seven key terms: applied, behavioral, analytic, technological, conceptual, effective, and capable of appropriately generalized outcomes.

The importance of a conceptual analysis has been pointed out since the beginning of the EAB. The conceptual analysis includes metatheory and philosophy, history and historiography, methodology, and system and theory.

3. *Development of the Area*

Applied behavior analysis began in the 1950s. The first decade was characterized by polemics and paradigmatic clashes. The second decade was the time of consolidation. The third saw the development of a refined methodology and new conceptual models. The last period has been characterized by applications to new fields, even in the so-called nontraditional areas (see Mattaini and Thyer 1996).

A very relevant problem refers to the role of conditioning in applied behavior analysis and in behavior therapy. This area was sometimes called 'conditioning therapy,' and it was considered as the application of the laws of 'modern learning theory.' Discussion has arisen concerning the differences between conditioning in the animal laboratory, in the clinic, and in daily life. Conditioning as the explanatory foundation of all applied behavior analysis has been very controversial. Some authors have insisted on the need to include knowledge drawn from psychobiology and social psychology, instead of relying only upon conditioning-learning theory.

The role of cognition in applied behavior analysis has been a central topic of discussion. Due to the desire to eliminate all reference to mentalism and to inner processes, early behavioral analysts rejected all forms of cognitive influence. But in the 1980s the 'cognitive revolution' led to some reformulations and reconceptualizations, and there was discussion about the role of cognition in behavior change. The term cognitive behavior therapy was used, rather than just behavior therapy. However, if behavior includes cognitive processes, the addition of 'cognitive' is redundant. The controversy continues at the beginning of the twenty-first century, and several positions and disputes exist concerning this issue. The precise relationship between behavior and cognition remains equivocal.

4. *Techniques*

A large number of techniques have been developed in the area of applied behavior analysis. Some of them are based on operant conditioning, others on classical conditioning, social learning, or a combination of them. Some involve principles derived from social psychology, developmental psychology, physiological, and cognitive psychology. Only the main techniques are presented here, and the list is not exhaustive.

4.1 *Systematic Desensitization*

This is a term applied to a class of methods used for gradually weakening unadaptative anxiety response habits through the use of responses that compete with anxiety. As an example, a physiological state inhibitory of anxiety is produced in the patient by means of muscle relaxation; the patient is then exposed for a few seconds at a time to a stimulus arousing weak anxiety. With repetition of the exposure, the stimulus progressively loses its ability to evoke anxiety. In a progressive way, stronger stimuli are presented.

The standard procedure in systematic desensitization consists of several sets of operations: (a) introduction to the subjective anxiety scale, (b) training in deep muscle relaxation, (c) construction of anxiety hierarchies, and (d) counterposing relaxation and anxiety-evoking stimuli from the hierarchies.

Counterconditioning is the basis of the change that follows systematic desensitization, according to Wolpe and other researchers and clinicians.

4.2 Biofeedback

This term refers to a number of clinical techniques which provide feedback to patients concerning changes in a particular physiological response, so that they can learn to modify the physiological response. Two methods of providing information feedback to the patient have been developed: binary feedback (yes/no information) and analogue feedback (proportional). A number of physiological responses have been shown to be modifiable using biofeedback. The most common physiological responses to be used clinically are electromyogram activity (EMG), skin temperature, blood pressure, EEG (electro-encephalogram), vasomotor, and heart rate.

Biofeedback is used in lowering the blood pressure of essential hypertensives, for achieving relaxation, for epilepsy, migraine headaches, and many other clinical problems (see Carrobles and Godoy 1987).

4.3 Punishment

Punishment is an operant conditioning procedure in which the future probability of occurrence of a response is reduced as a result of response-dependent delivery (positive punishment) or removal (negative punishment) of a stimulus (punisher). An example of positive punishment is spraying water mist in the face of a self-injurious child each time an instance of self-injurious behavior is observed. An example of negative punishment is taking an earned token from a person each time he or she curses. In the first case the aim is to reduce the occurrence of self-injury; in the second case it is to reduce the rate of cursing.

Punishment may weaken, strengthen, or have no effect on other response classes. Its effects are controversial and subject to discussion. In the early stages of applied behavior analysis, punishment procedures were used more frequently than now. Alternatives to punishment are usually preferred.

4.4 Escape Training

In this training procedure an individual learns to perform a response in order to terminate a punishing stimulus.

4.5 Avoidance Training

This is a training procedure in which an individual learns to keep away from a designated stimulus in order to prevent punishment. The first step consists in pairing the designated stimulus with a punisher. Then the individual is negatively reinforced by performing a response that terminates the punishing stimulus. The individual subsequently learns to prevent punishment from occurring by either responding with an alternative behavior (active avoidance) or not responding at all (passive avoidance) prior to coming in contact with the designated stimulus.

4.6 Extinction

Extinction is a response-weakening procedure which can involve consistent failure to deliver a reinforcer following an operant response which had previously produced that reinforcement (operant extinction). It eventually results in nearly complete elimination of the operant response, although a transient rate increase is frequently observed soon after operant extinction is arranged.

This technique is widely used in behavior modification and is preferred to punishment, avoidance training, escape training, and similar aversive procedures.

4.7 Shaping

This technique consists of reinforcing successive approximations to the desired response. Usually a simple response is required initially, and criteria for reinforcement are gradually made more stringent so as to produce more complex or refined response. Initially a simple response (in the direction of the desired goal) would be sufficient to receive reinforcement. After this behavior is performed reliably, reinforcement is given only for a more complex or difficult response. The next step consists of giving reinforcement for the slightly more complex response. The pattern continues until the final desired behavior is achieved regularly.

4.8 Contingency Management

Contingency management involves the analysis and change of the functional contingencies in the environment that determine a person's behavior. As a result, the behavior changes from its initial rate to a different rate of occurrence. In clinical work, this involves the development of new or alternative forms of adaptative behavior. Within a certain environmental context, a contingency describes a functional relationship among stimuli antecedent to a particular behavior (the target) and the consequences that follow the occurrence of the behavior.

The functional environment for a particular person is often different from the functional environment of another person. Many stimuli or events that occur in a person's environment do not have any effect on the person's behavior. In those cases we can say that the particular part of the environment is not functional, it is just present.

Contingency management involves the analysis and modification of contingencies involving discriminative stimuli, reinforcement, punishment, and extinction. It

focuses on public and observable behaviors and controlling environments. The environment of interest is primarily that part which functions to increase, decrease, or control the occurrence of certain behaviors. Contingency management in clinical practice typically is concerned with appropriate discrimination and generalization of adaptative behavior changes.

After a comprehensive functional analysis of environmental contingencies, the therapist plans a program of intervention to modify contingencies to change the rate of the target behavior. Contingency management involves a thorough functional analysis of controlling contingencies, the rearranging of the functional analysis of controlling contingencies, the rearranging of the functional environment, and a careful monitoring of outcome.

4.9 Environmental Design

This consists of the application of empirically derived principles of behavior for the modification and design of the environment. It involves the planning of a coherent program and set of procedures to affect the total human and nonhuman environment in ways that increase the probability that certain goals will be achieved.

The goal of environmental design relates to social behavior, for instance, planning a therapeutic or educational system. It includes linking applied behavior analysis and environmental psychology. It involves training people to conceptualize the environment so that they can apply the general principles of behavior analysis to design planning. Social design and physical design should be considered. The larger context is the planning of a whole society, such as *Walden Two* (Skinner 1948), or *Walden Three* (Ardila 1990).

4.10 Token Economy

In the token economy intervention, reinforcers are delivered for specific target behaviors. Reinforcers are tokens (tickets, coins, points, stars) that can be exchanged for a variety of other rewards. Tokens help bridge the delay between the person's performance of a desired behavior and delivery of a reward.

As an example of the use of token economies in complex settings we can mention the University of Kansas Experimental Living Project, a behaviorally managed experimental community of 30 college students (see Johnson et al. 1991). Token economies have also been used in prisons, in nursing homes, on psychiatric wards, in normal classroom settings from preschool to university classes, on wards for the treatment of drug addicts and alcoholics, in various work settings to increase safety behavior, in community settings to increase recycling of waste and improve energy conservation, and in many other contexts.

4.11 Premack Principle

This principle says: For any pair of responses the more probable one will reinforce the less probable one. The technique derived from the Premack Principle consists of a special class of reinforcing operations in which access to high probability behaviors is made contingent upon clinically targeted activities. For instance, children may be allowed to eat candy after (contingent upon) doing their homework.

4.12 Modeling

This consists of the presentation of a live demonstration, a filmed or pictorial presentation, or an imagined one. Based on the presentation, the viewer learns a method of responding or the contingencies under which the target response will fall.

Observational learning, imitation, and modeling are terms frequently used with respect to this phenomenon. These techniques are useful in the treatment of phobias, and in forming social skills and many other desired behaviors.

4.13 Cognitive Procedures

Cognitive therapy or cognitive applied behavior analysis refers to a wide range of approaches, based on two explicit hypotheses: (a) cognitive factors such as thoughts, images, and memories, are intimately related to dysfunctional behavior, and (b) modification of such factors is an important mechanism for producing behavior change.

4.14 Behavioral Self-control

This set of techniques involves training patients to control their own behavior through the systematic use of behavior principles. The therapist educates patients regarding the technical aspects of various behavior change procedures. In a self-reinforcement program, patients are taught to administer consequences to themselves contingent upon the performance of the target behavior.

In the majority of cases behavioral self-control involves self-monitoring, self-evaluation, and self-delivered rewards. There are many potential benefits from teaching patients self-control rather than using external agents to administer treatment contingencies. Self-control therapy allocates a great responsibility to patients. It is clear that teaching individuals to be their own therapists will contribute to generalization of treatment effects because patients will be able to

continue to administer the program or develop new treatment following the termination of therapy.

5. Main Areas of Application

Applied behavior analysis is a broad term and covers many fields of applied psychology, both traditional and nontraditional. A brief description of some of these areas follows.

5.1 Behavior Analysis Applied to Clinical Problems

This field is also referred to as behavior therapy, and has a long history. Behavior therapy has been defined as follows:

> 'Behavior therapy is the attempt to utilize systematically that body of empirical and theoretical knowledge which has resulted from the application of the experimental method in psychology and its closely related disciplines (physiology and neurophysiology) in order to explain the genesis and maintenance of abnormal patterns of behavior; and to apply that knowledge for the treatment or prevention of those abnormalities by means of controlled experimental studies of the single case, both descriptive and remedial.' (Yates 1970, p. 18)

Lovaas (1966, 1977, 1993) has developed behavioral treatment for autistic children. He focused on strategies to teach social behaviors, eliminate self-stimulatory behaviors, and develop language skills. The behavioral treatment has produced long-lasting effects, and no alternative treatment for autistic children has been shown to be as successful as behavior modification.

5.2 Behavior Analysis in Education

A number of applications have been carried out in several important educational aspects: behavior modifications in special populations (mental retardation, visual and auditory limitations, people with physical limitations, children of superior intelligence, and so forth). Also work has been done in order to maximize the efficiency of teaching techniques, learning in the classroom, and in general academic learning from preschool to university education. The techniques of reinforcement, contingency management, environmental design, token economy, and so forth, previously described in this article, have been applied to the educational area.

5.3 Industrial/Organizational Behavior Analysis

In the early 1970s applied behavior analysis entered the industrial arena. At the heart of this approach lies the belief that most complex human behaviors are learned, and by altering either the conditions preceding the behavior, or especially the consequences following it, that behavior can be altered and new behavior patterns can be learned. A wide variety of industrial-relevant behavior has been successfully altered: teaching the skills required for the job, motivating the performer, reducing absenteeism, reducing turnover, learning to work in groups, improving communication skills in the work setting, reinforcing adequate behavior, extinguishing undesired behavior, increasing productivity, improving morale, overcoming resistance to change, and others (see Hersey et al. 1996).

5.4 Behavioral Community Psychology

As a 'marriage' of the methodology of behaviorism and the strategies and conceptual framework of community psychology, the area of behavioral community psychology (Nietzel et al. 1996) uses classical conditioning, operant conditioning, social learning, and cognitive/self-control procedures. Examples of operant behavior analysis are prompts, feedback, reinforcement and response cost. Examples of social learning and cognitive/self-control procedures are modeling, behavioral rehearsal and role-playing, self-instruction and self-regulation, especially with respect to the development of skills and self-help competences in target populations.

5.5 Forensic Applications

Most applications of behavior analysis to psychology and the law have been in connection with criminal justice. Work has been carried out concerning the behavior of the delinquent individual, including juvenile delinquents and predelinquents; in the prevention of criminal behavior; and in the areas of legal processing, the courts, the jury, witnessess, and so forth. Special importance has been given to the behavior modification of delinquents using behavior analysis procedures.

5.6 Sport Psychology

In this area the contribution of applied behavior analysis has been centered on motivation of athletes, athletic performance, leadership, optimal coaching strategies, audience effects, sport violence, and other similar issues. The main behavioral technology procedures used in sport psychology are positive reinforcement, shaping, chaining, extinction, time out, response cost, fading, prompting, performance feedback, self-monitoring, self-instructions, deep muscle relaxation, modeling, imaginal rehearsal, behavioral rehearsal, thought stopping, and response-induced aids (see Goldstein and Krasner 1987).

5.7 Larger Sociocultural Applications

A number of 'big issues' have been addressed with the procedures of behavior analysis: youth violence,

racism, sexism, poverty, health problems, destruction of the environment, the breakdown of schools, lack of productivity in the workplace, child maltreatment. According to Mattaini and Thyer 1996) these behavioral problems 'could ultimately threaten the survival of the human race and other species' (p. 1)

See also: Behavior Therapy: Psychological Perspectives; Behaviorism; Environmental Psychology: Overview

Bibliography

Ardila R 1990 *Walden Three*. Carlton Press, New York

Ardila R 1998 *Manual de Análisis Experimental del Comportamiento*. Editorial Biblioteca Nueva, Madrid, Spain

Baer D M, Wolf M M, Risley T R 1987 Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis* **20**: 313–27

Caballo V E 1998 *International Handbook of Cognitive and Behavioural Treatments for Psychological Disorders*. Elsevier Science, Oxford

Carrobles J A, Godoy J 1987 *Biofeedback*. Ediciones Martínez Roca, Barcelona, Spain

Goldstein A P, Krasner L 1987 *Modern Applied Psychology*. Pergamon Press, New York

Hersey P, Blanchard K H, Johnson D E 1996 *Management of Organizational Behavior*. Prentice Hall, Upper Saddle River, NJ

Johnson S P, Welch T M, Miller L K, Altus D D 1991 Participatory management: Maintaining staff performance in a university housing cooperative. *Journal of Applied Behavior Analysis* **24**: 119–27

Kazdin A E 1994 *Behavior Modification in Applied Settings*. 5th edn., Brooks/Cole Pub. Co., Pacific Grove, CA

Labrador F J, Cruzado J A, Muñoz M 1993 *Manual de Técnicas de Modificación y Terapia de Conducta*. Ediciones Pirámide, Madrid, Spain

Lattal K A, Perone M 1998 *Handbook of Research Methods in Human Operant Behavior*. Plenum Press, New York

Lovaas O I 1966 A program for the establishment of speech in psychotic children. In: Wing J K (ed.) *Early Childhood Autism*. Pergamon, Elmsford, NY

Lovaas O I 1977 *The Autistic Child: Language Development Through Behavior Modification*. Irvington, New York

Lovaas O I 1993 The development of a treatment-research project for developmentally disabled and autistic children. *Journal of Applied Behavior Analysis* **26**: 617–30

Masters J C, Burish T G, Hollon S D, Rimm D C 1987 *Behavior Therapy*. 3rd edn., Harcourt Brace Jovanovich, San Diego, CA

Mattaini M A, Thyer B A 1996 *Finding Solutions to Social Problems. Behavioral Strategies for Change*. American Psychological Association, Washington, DC

Plaud J J, Eifert G H 1998 *From Behavior Theory to Behavior Therapy*. Allyn and Bacon, Boston

Skinner B F 1948 *Walden Two*. Macmillan, New York

Tolan P, Keys C, Chertok F, Jason L 1990 *Researching Community Psychology*. 1st edn., American Psychological Association, Washington, DC

Yates A J 1970 *Behavior Therapy*. Wiley, New York

R. Ardila

Behavior, Hierarchical Organization of

The analysis of most complex objects of scientific study reveals a hierarchical structure. At the bottom are elementary units (e.g., the elementary substances in chemistry), which are combined to form complex units (e.g., radicals in chemistry), which are themselves combined to form still more complex units (e.g., amino acids), and so on. The bodily movements that we call behavior are no exception. Complex behavior is created by coordinating elementary units of behavior to form units of modest complexity, then coordinating these units to form more complex units, and so on, up the scale of complexity. The extraordinarily protean character of behavior—the ability of even a basic behavior like locomotion to take on a seemingly infinite variety of forms—makes the existence of units of control and their hierarchical arrangement hard to recognize. This protean character of behavior is itself a consequence of this hierarchical structure and of the control principles that operate within it.

1. The Underlying Neurobiological Elements of a Unit of Behavior

Sherrington (1947/1906) defined a unit of behavior as an assemblage of neuromuscular and neuroendocrine elements sufficient to explain a naturally occurring pattern of movement or exocrine secretion. He recognized three necessary categories of neurobiological elements in any such assemblage—initiator elements, conducting elements, and effector elements. The neural signals that lead to movements originate in initiator elements. They are translated into movements or secretions by effector elements. Conductor elements convey them from the initiator elements to the effector elements.

The currently known initiator elements are sensory receptors and pacemakers. Sensory receptors convert stimulus energy into neural signals. Pacemakers generate rhythmic discharges by means of the intrinsic structure of the voltage-dependent ion channels in the membranes of some elements (pacemaker cells) and the structure of the neural circuits in which these elements are embedded. The effector elements are muscles and the exocrine glands, which are glands like tear glands, sweat glands, and salivary glands that secrete substances outside the body—bearing in mind that the alimentary tract is outside the body from a physiological perspective. The conductor elements are the axons of neurons, which conduct electrical impulses over long distances, the synaptic junctions between cells, where signals are conducted across the very short gaps between cells, and endocrine secretions, which are chemical signals conveyed from their release site to their target sites in the blood. Initiators, conductors and effectors are essential components in

any unit of behavior. Thus, for example, the vertebrate skeletal motor unit—a motor neuron together with the striated muscle fibers that it innervates—is not a unit of behavior, because this assemblage lacks an initiator element. The question is whether there are additional neurobiological elements that play essential roles in many units of behavior beyond the simple reflexes that Sherrington largely considered. In the second half of the twentieth century, the information-processing character of the nervous system has become clearer. This has led to the postulation of additional categories of sub-behavioral elements—memory elements, which hold information over time, and computational elements, which combine information-carrying signals in accord with elementary computational operations, such as the additions and multiplications that mediate and coordinate transforms. These elements, whose cellular nature remains undetermined, mediate the computational functions of the central nervous system. As will be seen, the need for such elements is apparent even in simple movements, such as the scratching movements of a hind limb or the smooth pursuit movements of the eyes.

2. *Distinguishing Between Elementary and Complex Units*

An assembly of neurobiological elements sufficient to explain a naturally occurring pattern of movement or glandular action is an elementary unit of behavior if it cannot be broken down into subassemblies that are themselves capable of explaining a component of that pattern. Take, for example, the scratch reflex in which an animal raises a hind limb so as to place the ends of the digits at the site of an irritating stimulus and then rhythmically wipes the digits repeatedly across the irritation. Although this is a very simple behavior, it is not an elementary unit of behavior, because it can be decomposed into experimentally separable movements, which are themselves units of behavior—the movement that brings the digits to the site of irritation and the rhythmic scratching movement (Sherrington 1947/1906). These two units cannot, so far as we know, be themselves decomposed into movements that are also units, so they are elementary units. Their coordination constitutes the complex unit that mediates the scratching of an irritated spot. The neural tissue that effects this coordination belongs to the complex unit, not to the elementary units that compose it. Thus, directed scratching is one level up in the control hierarchy.

Locomotion is an example of a complex unit much higher in the hierarchy (Gallistel 1980). It involves the coordination of the stepping of each of the limbs. The circuitry that controls the stepping of a single limb is itself a complex unit, coordinating the operation of an oscillator, several reflexes, and one or more servomechanisms. The coordination of the stepping movements of different limbs is effected by means of neural tissue that carries timing signals between the pacemakers that set the stepping rhythms for each limb and by command signals descending from higher neural centers to set the periods of these pacemakers. The complex neural circuitry that effects the coordination we see in locomotion constitutes itself a (complex) unit of behavior for two reasons: (a) it imposes a definite structure on the patterns of muscular activation, a structure that appears whenever the animal locomotes (Weiss 1941) and (b) this structure is activated in the course of many different higher level behaviors.

3. *Kinds of Elementary Units*

Just as there are different elementary substances in chemistry, so, too, there are different elementary units of behavior. Figure 1 gives the functional structures of five that have been well characterized (Gallistel 1994).

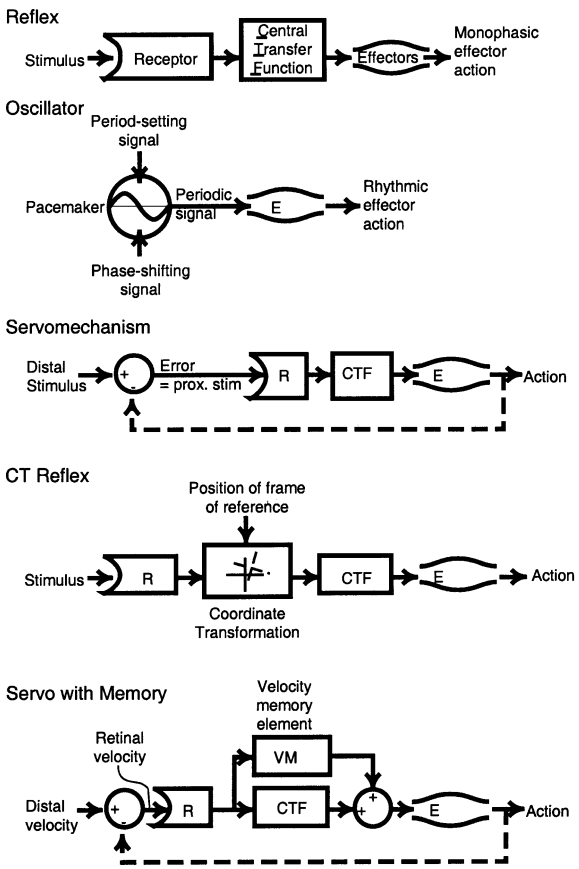

Figure 1
Functional structures of five elementary units of behavior (after Gallistel 1994)

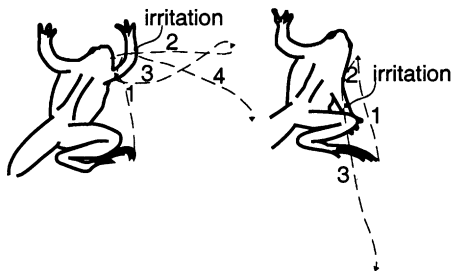

Figure 2
The scratching motion of the hind limb of the spinal frog (spinal cord separated from brain) varies depending on the position of the forelimb relative to the hind limb. Numbers indicate the sequence of motions (after Fukson et al. 1980)

In a simple reflex, the initiator element is a sensory receptor (R), sensory axons conduct the action potentials into the central nervous system (CNS), where the rise and fall of the resulting motor axon firing is determined (the central transfer function or CTF), and muscles or exocrine gland effectors (E) translate the signal into action. In an oscillator, a pacemaker is the source of the rhythmic signals driving a rhythmic action. Inputs to the pacemaker adjust its period and phase. In a simple servomechanism, like the optokinetic reaction, the proximal stimulus (in this case, the slippage of the image of the visual field across the retina) is jointly determined by the movement of the visual field (distal stimulus) and the rotation of the eyes and head (animal's response). This retinal image motion generates an error signal, which is shaped by the CTF to produce an effector action—rotation of the eyes and head in the direction of visual field movement—which reduces the proximal stimulus (retinal image motion) and hence the error signal. Thus, there is negative feedback from output to input. The parameters of the CTF presuppose the existence of the negative feedback; the unit does not operate properly in its absence (Collewijn 1969).

These three kinds of elementary units were characterized by the late 1960s, if not earlier, before the computational character of the nervous system was widely recognized. Elementary units containing computational elements and memory elements have been identified more recently.

The reflex movement that brings the digits of the hind leg to the site of irritation on the skin of the forearm in vertebrates has been shown to involve a coordinate transformation, which is why the functional structure of this kind of elementary unit includes a stage that performs this computation (Fukson et al. 1980). The transformation is necessitated by the fact that the spatial relation between the skin of the forearm and the joints of the hind limb changes as the position of the forelimb changes (Fig. 2). There is some reason to think that simple directed limb movements in the frog are controlled by four or five different populations of premotor neurons, each premotor population producing a characteristic dynamic force field when activated (Bizzi et al. 1995). Thus, to a first approximation, the endpoint of the movement may be thought of as being specified by a neural vector, with four or five dimensions (four or five different neuronal populations, with firing rates varying more or less independently between the populations). The set of all possible combinations of firing rates—hence all possible endpoints—constitutes a vector space. The signal indicating the locus of irritation on the skin of the forelimb may be thought of as coming from a two-dimensional vector space because it takes two dimensions to specify a location on the skin surface. The CNS maps the two-dimensional skin space into the four- or five-dimensional premotor space. This mapping itself varies when the position of the forelimb changes; a stimulus at the same point on the skin elicits a movement of the digits to a different point in space, because that point on the skin is now at that different point in space. The notion of a variable coordinate transformation in simple reflex orienting reactions has been most fully developed in the study of the saccadic orientation of the eyes (Gallistel 1999).

Memory elements are required in more sophisticated negative feedback mechanisms. In a simple error-actuated servomechanism, like the optokinetic reaction, there must always be a small error (in this case, a small amount of image slippage), because it is the error signal that drives the reaction. In more sophisticated negative feedback systems, like the smooth pursuit circuit, which mediates the smooth movement of the eyes in following a small foveated target, there is often no error; the eyes rotate at exactly the rate that keeps the target centered on the fovea. This requires a sample-and-hold mechanism, which is to say a mechanism that has a memory element incorporated in it (Krauzlis and Lisberger 1994). When the movement of the target is first detected, its retinal velocity is measured, the measured value is stored in a memory, and this remembered value drives the smooth rotation of the eyes. The incorporation of the memory element means that the effector action is no longer driven by the ongoing error signal, allowing this signal to become effectively zero.

4. *Selective Potentiation and Depotentiation*

A major principle of coordination, which may be observed at every level of the hierarchy, is the selective potentiation of units whose actions functionally cohere with the other actions going on at a given moment and the depotentiation of the units whose actions would be disruptive (Gallistel 1980). Potentiation

raises the potential for an action, but does not actually trigger that action. The triggering of a potentiated action is controlled by stimuli signaling the exact moment at which it becomes appropriate. There are, for example, two different stumble-preventing reflexes in the hind limb of the cat, one producing flexion of the limb, one producing extension (the exact opposite action). Both of these incompatible actions are triggered by a sudden onset tap on the dorsum of the paw (Forssberg et al. 1975), but they are not in practice both activated together. The flexor reflex is potentiated during the swing phase of each stepping cycle, while the opposing extension reflex is depotentiated. This pattern of selective potentiation and depotentiation reverses during the stance phase, when the paw is planted on the ground. The flexion reflex lifts the paw over an obstacle that would otherwise arrest its forward swing. The extension reflex thrusts the weight rapidly onto other legs, enabling the cat to lift the paw sooner when a moving object (e.g., a stone) threatens to sweep it from beneath the cat. Thus, which of the two reflexes is activated when the stimulus comes is determined by the selective potentiation and depotentiation so as to make the elicited action complement the ongoing phase of the step cycle, rather than hinder it.

At the top of the action hierarchy, the arousal of sexual motivation in a female rat through manipulation of her hormone levels potentiates the many different components of her sexual behavior (Adler 1973) and depotentiates reactions that would interfere with sexually oriented actions. For example, the hormonal state potentiates the lordosis response to a squeeze of her flanks, such as the male makes when he mounts. Her lordosis response firmly roots her to the floor and presents her genitals, facilitating penile entry. This same hormonal state depotentiates her flinch responses to painful stimuli delivered to her paws, presumably because such a reaction interferes with the male's intromission (Gordon and Soliman 1999).

5. *Origins of the Multiformedness of Behavior*

Complex units of behavior often have many different surface manifestations. For example, systems of coupled oscillators, of which the locomotory system is an example, can produce many different rhythmic sequences (in this case, different gaits, that is, sequences of stepping actions), depending on simple control parameters, such as the rate at which the stepping pacemakers are cycling (Wilson 1966). This is one reason why behavior takes on so many different forms. Reliance on coordination by selective potentiation and depotentiation within a hierarchical structure is a second major reason. It means that the actual spectrum of component actions that mediate a complex behavior like copulation vary depending on local circumstances particular to a given occasion, because different circumstances trigger different combinations of the potentiated actions.

See also: Cognitive Control (Executive Functions): Role of Prefrontal Cortex; Comparative Neuroscience; Motivation, Neural Basis of; Neural Systems and Behavior: Dynamical Systems Approaches; Neural Systems: Models of Behavioral Functions; Perception and Action; Word Meaning: Psychological Aspects

Bibliography

Adler N T 1973 The biopsychology of hormones and behavior. In: Dewsbury D, Rethlingshafer D A (eds.) *Comparative Psychology: A Modern Survey*. McGraw-Hill, New York, Chap. 9, pp. 301–44

Bizzi E, Giszter S F, Loeb E, Mussa-Ivaldi A, Saltiel P 1995 Modular organization of motor behavior in the frog's spinal cord. *Trends in Neuroscience* **18**(10): 442–6

Collewijn H 1969 Optokinetic eye movements in the rabbit: Input–output relations. *Vision Research* **9**: 117–32

Forssberg H, Grillner S, Rossignol S 1975 Phase dependent reflex reversal during walking in chronic spinal cats. *Brain Research* **85**: 103–7

Fukson O I, Berkinblit M B, Feldman A G 1980 The spinal frog takes into account the scheme of its body during the wiping reflex. *Science* **209**: 1261–3

Gallistel C R 1980 *The Organization of Action: A New Synthesis*. Erlbaum, Hillsdale, NJ

Gallistel C R 1994 Elementary and complex units of behavior. In: d'Ydewalle G, Celen P, Bertelson P (eds.) *Current Advances in Psychological Science: An International Perspective*. Erlbaum, Hillsdale, NJ, pp. 157–75

Gallistel C R 1999 Coordinate transformations in the genesis of directed action. In: Bly B O M, Rummelhart D E (eds.) *Cognitive Science*. Academic Press, New York, pp. 1–42

Gordon F T, Soliman M R I 1999 The effects of estradiol and progesterone on pain sensitivity and brain opioid receptors in ovariectomized rats. *Hormones & Behavior* **30**(3): 244–50

Krauzlis R J, Lisberger S G 1994 A model of visually-guided smooth pursuit eye movements based on behavioral observations. *Journal of Computational Neuroscience* **1**(4): 265–83

Sherrington C S 1947/1906 *The Integrative Action of the Nervous System*. Scribner, New York

Weiss P 1941 Self-differentiation of the basic patterns of coordination. *Comparative Psychology Monographs* **17**(4): 1–96

Wilson D M 1966 Insect walking. *Annual Review of Entomology* **11**: 103–22

C. R. Gallistel

Behavior Psychotherapy: Rational and Emotive

Rational emotive behavior therapy (REBT) was originated in January 1955 as a pioneering cognitive-

experiential-behavioral system of psychotherapy. It is heavily cognitive and philosophic, and specifically uncovers clients' irrational or dysfunctional beliefs and actively-directively disputes them. But it also sees people's self-defeating cognitions, emotions, and behaviors as intrinsically and holistically connected, not disparate. People disturb themselves with disordered thoughts, feelings, and actions, all of which importantly interact with each other and with the difficulties they encounter in their environment. Therefore, with emotionally and behaviorally disturbed people, REBT employs a number of thinking, feeling, and action techniques that are designed to help them change their self-defeating and socially sabotaging conduct to self-helping and socially effective ways (Ellis 1994, 1998, Ellis and Dryden 1997, Walen et al. 1992).

REBT theorizes that virtually all humans consciously and unconsciously train themselves to be to some degree neurotic; and that, with the help of an effective therapist and/or with self-help materials, they can teach themselves to lead more satisfying lives—if they choose to do so and work hard at modifying their thinking, feeling, and behaving.

1. *Philosophical Background*

Albert Ellis, the originator of REBT, was trained in Rogerian person-centered therapy in graduate school in clinical psychology (1942–7), found it too passive and abandoned it for psychoanalytic training and practice (1947–53). But psychoanalysis, too, he found ineffective because it was too much insight-oriented and too little action-oriented. His clients often saw how they originally became disturbed—supposedly because of their family history. But when he stayed with typical psychoanalytic methods, he failed to show them specifically how to think and act differently and to thus make themselves more functional.

So Ellis went back to philosophy, which had been his hobby since the age of 16, and reread the teachings of the ancient philosophers (especially Epicurus, Epictetus, Marcus Aurelius, and Gautama Buddha) and some of the moderns (especially Dewey, Russell, and Heidegger) and found that they were largely constructivists rather than excavationists. They held that people do not merely get upset by adverse life conditions, but instead often choose to disturb themselves about these adversities. Fortunately, a number of philosophers also said, people could choose to 'unupset' themselves about minor and major difficulties; and, if they made themselves anxious and depressed, they could reduce their dysfunctional feelings and behaviors by acquiring a core philosophy that was realistic, logical, and practical.

Following these philosophers, Ellis started to teach his clients that they had a choice of experiencing healthy negative emotions about the misfortunes they countered—such as feelings of sorrow, disappointment, and frustration; or they could choose to experience unhealthy negative reactions—such as panic, depression, rage, and self-pity. By using rational philosophy with troubled clients, he saw that when they faced adversities with self-helping attitudes they made themselves feel better and functioned more productively. But when they faced similar adversities with irrational (self-defeating) philosophies they made themselves miserable and acted ineffectively. When he convinced them that they almost always had the choice of helping or hindering themselves, even when their desires and goals were seriously blocked, they often were able to make that choice.

2. *The ABCs of Rational Emotive Behavior Therapy*

During the 1950s, Ellis put this constructivist theory into the now well-known ABCs of REBT. This theory states that almost all people try to remain alive and achieve basic goals (G) of being reasonably content by themselves, with other people, productively working, and enjoying recreational pursuits. When their goals are thwarted and they encounter adversities (A) they are then able to construct consequences (C)—mainly feelings and actions—that either help or hinder them satisfying these goals. They largely (though not completely) do this by choosing to follow rational, useful beliefs (B) or to follow irrational, dysfunctional beliefs. Therefore, although the adversities (A) they experience are important contributors to their emotional and behavioral consequences (C), they do not directly or solely cause these consequences. When at C, people feel and act neurotically or self-defeatingly, their irrational beliefs (B) and their experienced adversities (A) bring on their disturbed reactions. So A does not by itself lead to C. A interacts with B to produce C; or $A \times B = C$. However, people tend to be aware that C follows A, but not that B is also included in the process. They therefore think that As automatically lead to disturbed Cs—that their internal reactions are controlled by external events.

Ellis noted in his first paper on REBT, at the Annual Convention of the American Psychological Association in Chicago in August 1956, that when people feel and act disturbedly (C), they have 12 common irrational or dysfunctional beliefs (B) about the undesirable things that happen to them (A). When they change these to rational or functional beliefs (in therapy or on their own) they become significantly less disturbed. Both these hypotheses have been supported by many empirically-based studies, first by followers of REBT (Lyons and Woods 1991), and then by other cognitive behavior therapists who largely follow and have tested the ABC theory of REBT (Barlow and Craske 1994, Hollon and Beck 1994, Meichenbaum

1997). Hundreds of published studies have given much support to this theory.

After using REBT for a few years in the 1950s, Ellis came up with clinical evidence for Karen Horney's (1950) hypothesis about the 'tyranny of the shoulds.' He realized that the many irrational beliefs with which people often disturb themselves can practically always be put under three major headings, all of which include absolutistic shoulds, oughts, and musts. With these three core dysfunctional ideas, people take their strong preferences for success, approval, power, freedom and pleasure, and elevate them to dogmatic, absolutistic demands or commands.

The imperatives that frequently accompany dysfunctional feelings and behaviors seem to be: (a) 'I *absolutely must* perform well at important tasks and be approved by significant others—or else I am an inadequate person!' (b) 'Other people *absolutely must* treat me kindly, considerately, and fairly—or else they are bad individuals!' (3) 'Conditions under which I live *absolutely must* provide me with what I really want—or else my life is *horrible*, I *can't stand* it, and the world's a *rotten place*!'

These three common irrationalities lead to innumerable derivative irrational beliefs and frequently are accompanied by disturbed emotional and behavioral consequences. In fact, REBT hypothesizes that people would find it difficult to make themselves neurotic without taking one or more of their major preferences and transforming them into absolutistic demands. Individuals with severe personality disorders and psychosis also disturb themselves by turning their healthy preferences into unhealthy musturbating, but they often have other biochemical and neurological characteristics that help make them disturbed.

REBT also theorizes that the tendency to elevate healthy preferences to insistent demands, and thereby to think, feel, and act unrealistically and illogically, is innate in humans. People naturally and easily take some of their strong goals and desires and often view them as necessities. This self-defeating propensity is then exacerbated by familial and cultural upbringing, and is solidified by constant practice by those who victimize themselves with it. Therefore, especially with seriously disturbed people, psychotherapy and self-help procedures can, but often only with difficulty, change their dysfunctioning.

Many therapy techniques—such as meditation, relaxation, a close and trusting relationship with a therapist, and distraction with various absorbing activities—can be used to interrupt clients' musturbatory tendencies and help them feel better. But in order for them to get and stay better, REBT holds, they usually have to consciously realize that they are destructively escalating their healthy desires into self-sabotaging demands and then proceed to D: to actively and forcefully dispute the irrational beliefs that are involved in their disturbances. By vigorously and persistently disputing these beliefs—cognitively, emotively, and behaviorally—they can change their self-destructive shoulds and musts into flexible, realistic, and logical preferences. They thereby can make themselves significantly less disturbed.

3. *Rational Emotive Behavior Therapy Techniques*

To help people specifically achieve and maintain a thoroughgoing antimusturbatory basic outlook, REBT teaches them to use a number of cognitive, emotive, and behavioral methods. It helps them gain many insights into their disturbances, but emphasizes three present-oriented ones:

Insight No. 1: People are innate constructivists and by nature, teaching, and, especially, self-training they contribute to their own psychological dysfunctioning. They create as well as acquire their emotional disabilities—as the ABC theory of REBT notes.

Insight No. 2: People usually, with the 'help' and connivance of their family members, first make themselves neurotic when they are young and relatively foolish. But then they actively, though often unconsciously, work hard after their childhood and adolescence is over to habituate themselves to dysfunctional thinking, feeling, and acting. That is mainly why they remain disturbed. They continue to construct dysfunctional beliefs.

Insight No. 3: Because of their natural and acquired propensities to strongly choose major goals and values and to insist, as well as to prefer, that they must achieve them, and because they hold these self-defeating beliefs and feelings for many years, people firmly retain and often resist changing them. Therefore, there usually is no way for them to change but work and practice for a period of time. Heavy work and practice for short periods of time will help; so brief rational emotive behavior therapy can be useful (Ellis 1999). But for long-range gain, and for clients to get better rather than to feel better, they require considerable effort to make cognitive, emotive, and behavioral changes

REBT clients are usually shown how to use these three insights in the first few session of psychotherapy. Thus if they are quite depressed (at point C) about, say, being rejected (at point A) for a very desirable job, they are shown that this rejection by itself did not lead to their depression (C). Instead they mainly upset themselves with their musturbatory beliefs (B) about the adversity (A). The therapist explores the hypothesis that they probably took their desire to get accepted and elevated it into a demand—e.g.,'I *must not* be rejected! This rejection makes me an *inadequate person* who will *continually* lose out on fine jobs!' (Ellis 1998, 1999).

Second, clients are shown—using REBT Insight No. 2—that their remembering past adversities (A),

such as past rejections and failures, do not really make them depressed today (C). Again, it is largely their beliefs (B) about these adversities that now make them prone to depression.

Third, clients are shown that if they work hard and persistently at changing their dysfunctional beliefs (B), their dire needs for success and approval, and return to mere preferences, they can now minimize their depressed feelings—and, better yet, keep warding them off and rarely fall back to them in the future. REBT enables clients to make themselves less disturbed and less disturbable.

4. Multimodal Aspects of Rational Emotive Behavior Therapy

To help clients change their basic self-defeating philosophies, feelings and behaviors, REBT practitioners actively-directively teach and encourage them to use a good many cognitive, experiential, and behavioral techniques, which interact with and reinforce each other. Cognitive methods are particularly emphasized, and often include: (a) *active disputing of clients' irrational beliefs* by both the therapists and the client; (b) *rational coping self-statements* or effective philosophies of living; (c) *modeling* after people who coped well with adversities similar to, or even worse than, those of the clients; (d) *cost–benefit analyses* to reveal how some pleasurable substances and behaviors (e.g., smoking and compulsive gambling) are self-sabotaging and that some onerous tasks (e.g., getting up early to go to work) are unpleasant in the short term but beneficial in the long term; (e) *REBT cognitive homework forms* to practice the uncovering and disputing of dysfunctional beliefs; (f) *psychoeducational materials*, such as books and audiovisual cassettes, to promote self-helping behaviors; (g) *positive visualizations* to practice self-efficacious feelings and actions; (h) *reframing of adversities* so that clients can realize that they are not catastrophic and see that they sometimes have advantages; (i) *practice in resisting overgeneralized, black and white, either/or thinking*; (j) *practical and efficient problem-solving techniques*.

REBT uses many emotive-experiential methods and materials to help clients vigorously, forcefully, and affectively dispute their irrational demands and replace them with healthy preferences (Bernard and Wolfe 2000). Some of its main emotive-expressive techniques include: (a) *forceful and persistent disputing of clients' irrational beliefs*, done in vivo or on a tape recorder; (b) *experiencing a close, trusting, and collaborative relationship* with a therapist and/or therapy group; (c) *steady work at achieving unconditional other-acceptance* (*UOA*), the full acceptance of other people with their failings and misbehaviors; (d) *using visualizations or live experiences to get in touch with intense unhealthy negative* feelings—and to train oneself to feel, instead, healthy negative feelings; (e) *roleplaying difficult emotional situations* and practicing how to handle them; (f) *using REBT's shame-attacking exercises* by doing 'embarrassing' acts in public and working on not denigrating oneself when encountering disapproval; (g) e*ngaging in experiential and encounter exercises* that produce feelings of discomfort and learning how to deal with these feelings.

REBT uses many activity-oriented behavioral methods with clients, such as: (a) *exposure or in vivo desensitization* of dysfunctional phobias and compulsions; (b) *taking deliberate risks of failing at important projects* and refusing to upset oneself about failing; (c) *staying in uncomfortable situations and with disturbed feelings* until one has mastered them; (d) *reinforcing oneself to encourage self-helping behaviors* and penalizing oneself to discourage self-defeating behaviors; (e) *stimulus control* to discourage harmful addictions and compulsions; (f) *relapse prevention* to stop oneself from sliding back to harmful feelings and behaviors; (g) *skill training* to overcome inadequacies in assertion, communication, public speaking, sports, and other desired actions that one is inhibited about.

These are some of the cognitive, emotive, and behavioral techniques that are employed frequently in rational emotive behavior therapy. Many other possible methods are individually tailored and used with individual clients.

The main therapeutic procedure of REBT is to discover how clients think, feel, and act to block their own main desires and goals, and to figure out and experiment with ways of helping them get more of what they desire and less of what they abhor. As they make themselves less disturbed and dysfunctional, they are helped to actualize themselves more—that is, to provide themselves, idiosyncratically, with greater satisfactions. At the same time, clients are helped to stubbornly refuse to define their preferences as dire necessities and thereby tend to reinstitute their disturbances.

When Ellis originated it in 1955, rational emotive therapy was unique. It was followed by somewhat similar forms of cognitive behavior therapy (CBT) in the 1960s and 1970s, particularly the cognitive therapy of Beck (1976), rational behavior therapy of Maultsby (1975), cognitive behavior modification of Meichenbaum (1977), and multimodal therapy of Lazarus (1978). REBT and CBT were soon supported by numerous published studies that showed their effectiveness, with many different types of clients (Hollon and Beck 1994, Lyons and Woods 1991, McGovern and Silverman 1984, Silverman et al. 1992). Consequently, they now are widely employed throughout the Western world. Their use of multimodal methods of therapy has also encouraged the recent movement toward integration in psychotherapy and will most probably continue to do so in the future. It looks likely that rational emotive behavior therapy (REBT) and cognitive behavior therapy (CBT) will continue to thrive in the twenty-first century.

See also: Behavior Therapy: Psychological Perspectives; Cognitive Therapy; Experiential Psychotherapy; Psychological Therapies: Emotional Processing; Psychological Treatment, Effectiveness of; Psychotherapy Integration; Psychotherapy Process Research

Bibliography

Barlow D H, Craske N G 1994 *Mastery of Your Anxiety and Panic*. Graywind Publications, Albany, NY

Beck A T 1976 *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York

Bernard M, Wolfe J L 2000 *The REBT Resource Book for Practitioners*. Albert Ellis Institute, New York

Ellis A 1994 *Reason and Emotion in Psychotherapy*, revised and updated. Carol Publishing Group, Secaucus, NJ

Ellis A 1998 *How to Control Your Anxiety Before it Controls You*. Carol Publishing Group, Secaucus, NJ

Ellis A 1999 *How to Make Yourself Happy and Remarkably Less Disturbable*. Impact Publishers, Atascadero, CA

Ellis A, Dryden W 1997 *The Practice of Rational Emotive Behavior Therapy*. Springer, New York

Hollon S D, Beck A T 1994 Cognitive and behavioral therapies. In: Bergin A E, Garfield S L (eds.) *Handbook of Psychotherapy and Behavior Change*. Wiley, New York, pp. 437–90

Horney K 1950 *Neurosis and Human Growth*. Norton, New York

Lazarus A A 1978 *The Practice of Multimodal Therapy*. Johns Hopkins University Press, Baltimore, MD

Lyons L C, Woods P J 1991 The efficacy of rational-emotive therapy: A quantitative review of the outcome research. *Clinical Psychology Review* **11**: 357–69

Maultsby M C 1975 *Help Yourself to Happiness*. Esplanade Books, Boston

McGovern T E, Silverman M S 1984 A review of outcome studies of rational emotive therapy from 1977 to 1982. *Journal of Rational-Emotive Therapy* **2**: 7–18

Meichenbaum D 1977 *Cognitive-Behavior Modification*. Plenum, New York

Silverman M S, McCarthy M, McGovern T E 1992 A review of outcome studies of rational-emotive therapy from 1982–1989. *Journal of Rational-Emotive and Cognitive Behavior Therapy* **10**: 111–86

Walen S, DiGiuseppe R, Dryden W 1992 *A Practitioner's Guide to Rational-Emotive Therapy*. Oxford University Press, New York

A. Ellis

Behavior Therapy: Psychiatric Aspects

Behavior Therapy refers to a range of treatments and techniques which are used to change an individual's maladaptive responses to specific situations. A useful definition is provided by Meyer and Chesser (1970): 'Behaviour therapy aims to modify current symptoms and focuses attention on their behavioural manifestations in terms of observable responses. The techniques used are based on a variety of learning principles. Although behaviour therapists adopt a developmental approach to the genesis of symptoms, they do not think it is always necessary to unravel their origin and subsequent development.'

In psychiatric practice, behavior therapy has been used and demonstrated to be useful in the treatment of anxiety disorders, obsessive-compulsive disorders, habit disorders, and in the modification of challenging or antisocial behaviors. Cognitive Therapy, which examines an individual's thinking patterns and aims to help the person to alter any maladaptive thoughts, can often be used effectively in conjunction with Behavior Therapy. Combination treatment has been shown to be effective in depression, panic, generalized anxiety, posttraumatic stress, and a variety of other psychiatric conditions.

Behavior Therapy works on the premise that, in psychiatric practice, maladaptive behaviors cause and exacerbate psychological distress. Such behaviors can be replaced or unlearned. Although the behaviors may be problematic and result in distress to an individual, initially they are rewarded by a consequence of the behavior. This reward or reinforcement of the behavior strengthens the link between the specific situation or stimulus and the resultant undesirable behavior. This is through Classical Conditioning. Behavior Therapy techniques alter the behavior and the reinforcing feedback of its consequences.

1. History of Behavior Therapy

Although a range of ancient and historical documents can be found which describe treatments of both children and adults with techniques which would now be described as Behavior Therapy, the treatment of anxiety disorders and neuroses was with analytical psychotherapy at the early part of the twentieth century. The prevailing view in psychiatric practice was that such disorders were due to unconscious conflicts which could take years of psychoanalysis to resolve. One of the earliest challenges to this belief was the now famous experiment of Watson and Rayner on Little Albert (Watson and Rayner 1920). Little Albert was a young child who had shown no fear of animals. In their experiment, Watson and Rayner placed the child in a room with a white rat. Any approach of the child towards the rat resulted in a loud aversive noise being made by the researchers. Thus, through this experiment in classical conditioning, the researchers developed a fear of white rats which subsequently generalized to other furry objects and animals in a matter of a few hours. The experimenters then hoped to reverse this experimentally-induced phobia by offering rewards to the boy while in the presence of a rat. Not surprisingly, however, Albert was discharged from hospital before this further research was completed.

A few years later, Jones (1924) tried reinforcement techniques in children with pre-existing phobias. She used food with hungry children and brought the feared object into increasing proximity with the child as it ate. Any fear-response caused her to move the object further away until the child started eating again. This was repeated until the child was happy in close proximity to the feared object.

Despite these clear examples of behavioral principals, behavior therapy was not actually defined as a discipline until the 1950s when Joseph Wolpe was popularizing treatment using systematic desensitization for phobic disorder in the USA. Systematic desensitization is based on the observation that an individual cannot be relaxed and also experience anxiety. The fear response to a phobic stimulus could thus be reduced and eventually extinguished by teaching the patient deep muscular relaxation. A strict hierarchy of fear-provoking stimuli was obtained from the patient and initially, after the patient had induced relaxation, they were asked to imagine the least fear-inducing of these situations. Once this was achieved the patient was asked to move a step at a time with increasingly threatening images. If the patient became anxious the therapist intervened and reintroduced relaxation before asking the patient to imagine a stimulus lower on the hierarchy of fear (Wolpe 1958). Since that time there have been a large number of studies of systematic desensitization. For example, Gelder et al. (1967) compared systematic desensitization with individual psychotherapy and with group psychotherapy in patients with phobic disorders. Systematic desensitization was consistently better than either form of psychotherapy. Although systematic desensitization was shown to be effective in the treatment of specific phobic disorders, it involved a long and laborious process of extremely gradual exposure for both the therapist and the patient.

A more rapid method was found with real life exposure methods which were originally called flooding (Stampfl and Levis 1967). Real-life exposure involves asking the patient to face up to the feared stimulus and without using relaxation or distraction. Even though this causes anxiety, if the patient remains in contact with the stimulus for a length of sufficient time (approximately 1 h), then the anxiety does reduce. Because of the difficulty in encouraging a frightened individual to face up to their worst fear, this is also performed in a hierarchical manner, starting with a situation which causes moderate anxiety before progressing. The hierarchy does not have to be as structured as with systematic desensitization and it was soon found to be effective after only a few hours of therapist time and had the added benefit of teaching the patient how to cope with anxiety which could be used in other anxiety-provoking situations. Real-life exposure methods were pioneered with specific phobias (Watson and Marks 1971) and caused a sensation when first reported. A variety of specific phobics were treated with rapid exposure in real life with great success. Fears by some therapists that anxious patients would be made worse and other symptoms would appear proved unfounded. The great success of these techniques was also criticized because they had not yet been used for the commonest phobia, agoraphobia.

Studies of exposure treatment for agoraphobia were then carried out, for example, Mathews et al. 1976, and the treatment was found to be effective. From 1970 to 1980 exposure treatments for agoraphobia were refined and crucial treatment factors were isolated. For example, the importance of the duration of exposure in treatment of agoraphobia was discovered (Stern and Marks 1973). Real-life exposure was found to be more effective than even very rapid exposure performed in the patient's imagination (Emmelkamp and Wessels 1975).

Over the next few years refinement of psychological treatment factors in agoraphobia continued. It was thought that if treatment could be done in groups it would be more cost effective. Hand et al. (1974) compared 'structured' with 'unstructured' groups, and found progress was greater in the structured group. Only the structured group continued to improve after treatment and to 3-month follow-up. Group exposure required less professional time and so was an advance in treatment. However, in practice it is often difficult to get a group of similar phobic patients together at the same time and the progress of the group as a whole is linked to the speed of progress of it's slowest member.

More recent work in the treatment of anxiety disorders has been in the development of self-exposure where the patient is given detailed instructions about how to perform exposure treatment but the therapist is not present during the exposure therapy. Another development which led to more effective and rapid therapy was the development of home-based treatment where the therapist would carry out the treatment in the patient's home which was subsequently continued by the patient in self-exposure. Marks et al. (1983) showed that brief therapist-aided exposure was a useful adjunct to self-exposure homework instructions, but even self-exposure homework instruction alone is a potent treatment for agoraphobia. Experience to date supports the use of this approach, although involvement of family members in the treatment wherever possible is thought to be beneficial.

The treatment of obsessive-compulsive disorder was developed at the same time as that for agoraphobia. Marks et al. (1975) reported the results over 2 years of 20 patients with chronic obsessive rituals treated by real life exposure. This treatment was as effective for obsessive rituals as it was for agoraphobia. The only difference between the real life exposure used in obsessive-compulsive disorder and phobic disorder is that with the former the patient has to be asked not to perform the anxiety-reducing rituals or compulsions which are a feature of this disorder. These anxiety-

reducing behaviours would otherwise interfere with the natural habituation to anxiety which occurs with prolonged exposure. Although in patients with rituals it was often combined with modeling and response-prevention.

What is at first surprising is that a similar treatment for obsessive-compulsive disorder had been described in the early 1900s. A French psychiatrist, Janet treated a patient with obsessional rituals (1903) and the principles of his treatment are clearly described in the following quotation:

> The person who assists in the performance of these actions, has a very complicated part to play. He must aid in the performance of the action without actually doing it himself, although the latter would be very much easier; and he must do his utmost to conceal his own contribution to the action, for it is essential that the patient should feel he does the action himself and does it unaided. The guide has chosen the action, has overcome the patient's hesitations, and has taken the responsibility … by continual repetition to perform the action; by words of encouragement at every sign of success however insignificant, for encouragement will make the patient realise these little successes, and will stimulate him with the hope aroused by glimpses of greater successes in the future.

The principles outlined by Pierre Janet were refined in a series of studies over the years (Levy and Meyer 1971, Rachman et al. 1971). Until the present time when a truly effective therapy can now be offered to most patients with compulsive rituals.

As well as a treatment for phobic and obsessive-compulsive disorders, behavior therapy has been applied to range of other complaints as varied as schizophrenia and marital/sexual dysfunction. The use of behavior therapy in schizophrenia and learning disability started when Ayllon and Azrin (1968) applied operant reinforcement principles to produce changes in patients with chronic schizophrenia. Nowadays behavior therapy is used to counteract the secondary handicaps of schizophrenia and to assist in patients discharged from hospital to the community.

In 1970 the pioneering work of Masters and Johnson (Masters and Johnson 1966, 1970) led to a range of behavioral techniques for sexual dysfunction.

There are a multitude of other conditions where behavior therapy has been demonstrated to be effective but many of these are outside the realms of this short article. There is a trend now for behavior therapy to be used for an increasing range of disorders, and much progress has been made since the first tentative exposure techniques.

2. *Modern Behavior Therapy*

For the purposes of this article this section will be divided into three main sections. The first will deal with the current use and application of exposure methods of treatment which are used mostly in the anxiety disorders. The second will examine the methods used to teach alternative skills where these have not been learned. Skill training is used in social skills training, treatment of sexual dysfunction, marital therapy, and problem-solving. Third, there will be a description of the use of reinforcement techniques. These techniques are used in learning disability, with chronic patients, and sometimes in child psychiatry.

2.1 *Exposure Treatments*

Exposure treatment has been shown to be effective in 66 percent of agoraphobics (Mathews et al. 1981), between 75 and 85 percent of obsessive-compulsives (Foa and Goldstein 1978, Rachman et al. 1979), and highly effective (Marks 1981) in a mixture of specific and social phobics.

There is often fear of using exposure treatment outside specialist centres. This appears to arise from erroneous views about its applicability, its success rates, the time commitment required by the therapist and also fear of the unknown. In fact, graduated exposure is a remarkably quick and cost-efficient treatment which can be easily applied in many general practice and hospital settings (Marks 1981, 1986, Stern and Drummond 1991). Although some basic training is required, this can easily be obtained by reading about clinical techniques (Hawton et al. 1989, Stern and Drummond 1991) and by obtaining supervision from a trained behavioral psychotherapist.

The most effective exposure has been shown to be:

(a) Prolonged rather than of short duration (Stern and Marks 1973).

(b) In real life rather than in fantasy (Emmelkamp and Wessels 1975).

(c) Regularly practiced with self-exposure homework tasks (McDonald et al. 1978).

One of the concerns about exposure treatment has been that it requires considerable professional input to accompany an anxious patient into fear-provoking situations. Fortunately, it has been demonstrated that instruction in self-exposure techniques can be all that is required for many patients with phobic anxiety and obsessive-compulsive disorder (Ghosh et al. 1988, Marks et al. 1988). The efficacy of self-exposure has led to the development of a number of self-help manuals. However, few patients can complete a treatment program successfully without some professional guidance. The patient needs to be seen initially for education about anxiety and its treatment and for help in devising treatment targets. Subsequent meetings are required to monitor progress, give encouragement and advise on any difficulties which may arise (Ghosh et al. 1988).

Exposure is the cornerstone of treatment of obsessive-compulsive disorder when it is combined with response-prevention. Response-prevention means encouraging the patient not to perform rituals and compulsions. Rituals are overt behaviors or internal

thought patterns which are used to counteract the obsessional fears. Rituals usually can be prevented or reduced substantially by demonstrating to the patient how they interfere with exposure. In exposure treatment, the aim is to produce prolonged periods of contact with the feared situation until the anxiety reduces (habituation). Compulsions or rituals reduce the anxiety and this serves to reinforce the ritual. However, the reduction in anxiety produced by a ritual tends to be small and the effect temporary. In effect rituals prevent therapeutic exposure and instead increase the tendency to ritualize further.

Exposure methods have also been used to treat a variety of other anxiety based disorders to good effect including posttraumatic stress disorder (Keane et al. 1989, Foa et al. 1991) and hypochondriasis (Warwick et al. 1996).

2.2 Teaching New Skills

In a variety of situations some people do not have the appropriate skills to meet the challenges of the environment in which they find themselves. This is, of course, true to some extent for all of us. However with some individuals, specific training can resolve a crisis and remove psychological distress.

Skill training involves the therapist in detailed analysis with the patient to establish what skills are at fault and what effect this defect has. This is followed by sensitive discussion about how the skills could be improved. Frequently the therapist will then demonstrate the skill to be achieved. With a complex skill or where the patient has a very low level of skill in the area, the skill may have to be broken down into a series of stages with the stages being presented one at a time. The patient is then asked to rehearse the new skill and feedback is given by the therapist. It is important that this feedback is encouraging and highlights the positive aspects of the performance as well as making suggestions about any factors which could be improved. In some cases, video or audio tape can also be used to give additional feedback to the patient about progress. The patient is then asked to practice the skill. This continues until the patient is able to incorporate the new skill in everyday life.

These techniques frequently are used in social skills training which is often applied to groups. In couple therapy, they can be used in behavioral exchange therapy to encourage the couple to do positive things for each other using a 'give to get' principle. They may also be used to encourage positive communication between couples. Similar skills are also used in sexual therapy which is based partly on the exposure principle in anxious individuals and also aims to improve basic sexual knowledge and skill. In the case of sexual therapy, the therapist does not demonstrate the skill but describes it using diagrams and the couple are asked to practice at home. In problem-solving techniques, which can be used as an adjunct to other therapies for many patients, the individual is taught how to describe a problem and to weigh up the pros and cons of a decision. A discussion of what is an appropriate response which will produce the best result then follows. This is followed by rehearsal of the required skills.

Due to the diverse nature of the skill-training techniques, there is limited research on outcome. However, studies on sexual skills training have shown it to be an effective treatment for a whole range of sexual problems (Masters and Johnson 1970, Bancroft 1983) and social skills training has variable results depending on the patient mix but many of these are favorable (e.g., Stravynski et al. 1982).

2.3 Applying Operant Techniques to Chronic Problems

In patients with chronic behavior problems, treatment techniques based on operant conditioning are often used. Operant methods can often be more difficult to apply due to confusion about what may be reinforcing or rewarding to a patient. Premack's (1959) principle addresses this finding by observing that high-frequency preferred activity can be used to reinforce lower-frequency, nonpreferred activity. In other words, if a teenage girl spends most of her time surfing the net, this high-frequency preferred activity could be used to reinforce the lower frequency nonpreferred activity of tidying her bedroom.

The role of this type of treatment aimed at reducing undesired behaviors and increasing socially acceptable behavior has increased in the past few years with the increasing closure of many of the older psychiatric institutions and a move towards community care. Discharge to the community or even admission in a psychiatric unit of a district general hospital often means that bizarre or socially unacceptable behavior is not tolerated as well as in the older institutions which were often segregated from the rest of society.

By far the most commonly applied form of reinforcement to be used in positive reinforcement. Negative reinforcement or even punishment are hardly ever used and usually only in dangerous or life-threatening situations due to the obvious ethical dilemmas presented by such methods. Ethical problems may also arise with positive reinforcement, particularly if essential items such as food are used as reinforcers and the patient has not earned a meal that day.

2.3.1 Reinforcers which increase desirable behaviors

Positive reinforcers

(a) Social approval, for example, therapist's approval in phobic who has complied with exposure task.

(b) Higher frequency preferred activities.

(c) Feedback reinforcement, for example, social skills group.

(d) Food reinforcers.

(e) Tokens which are awarded for certain behaviors and can be used by the patient in exchange for a range of rewards.

Negative reinforcement

This means the removal of an aversive event after a specific response is obtained. There are clear ethical problems with such methods and they are rarely used in clinical practice. One exception is when they are used with a consenting adult and applied covertly. For example, a man with illegal sexual preference could be trained to have aversive images following arousing deviant images. He is also taught that he can gain relief from these aversive experiences by switching to neutral thoughts.

2.3.2 Reinforcers which reduce undesirable behaviors

Punishment

(a) Time out, that is, removal of the individual from reinforcing environment for up to 3 minutes.

(b) Overcorrection, for example a patient smashes his cup to the floor, he is asked not only to clear up the mess but then to wash the entire ward floor.

(c) Positive punishment, for example, child puts its finger towards an electric socket and receives a sharp slap on the hand.

Response cost

(a) Penalty involving some time and effort in response to certain behaviors.

(b) A positive reinforcer is removed if certain nondesired activities are indulged in.

3. The Future of Behavior Therapy

Although pure behavior therapy in the form of exposure remains the psychological treatment of choice for uncomplicated phobias and obsessive-compulsive disorder, in many clinical situations it is often combined with cognitive techniques. Many recent studies in a range of conditions as varied as schizophrenia (Tarrier et al. 1993, Kingdon et al. 1994) and hypochondriasis (Warwick et al. 1996), describe such cognitive behavioral treatment. In the future further work needs to identify the specific therapeutic factors of these treatment packages.

The treatment of anxiety disorders has developed rapidly since the 1980s. Anxiety can be treated using behavioral or drug treatments. Unfortunately, there has been a tendency for workers in this field to polarize and confine themselves to their own approach. Both forms of treatment do have difficulties as well as advantages. Similarly criticism can be made of much of the research in this area which has tended to show advantages of drug treatment by those interested psychopharmacology and advantages in behavioral-cognitive therapy by those interested in psychological approaches. This is partly because behavioral-cognitive therapists tend to be more interested in behavioral and cognitive changes, whereas biological therapists are more interested in affective change. Much of the existing research in this area suffers from underpowered studies and poor methodology (Fineberg and Drummond 1995). Despite this, behavior therapy frequently is used in patients who are taking drugs (e.g., Drummond 1993). Future research needs to be done into the effect of combining these two approaches.

Overall, behavior therapy is a proven therapy of known efficacy for a variety of psychiatric and psychological symptoms. It can be usefully combined with cognitive techniques to improve the patient's compliance or to improve the outcome. Drugs may also be used in conjunction with behavior therapy but research in this area has shown conflicting results. Future research needs to examine each aspect of a treatment package and discover which combinations are the most beneficial to patients.

See also: Behavior Therapy: Psychological Perspectives; Behavior Therapy with Children; Behaviorism; Health Behaviors; Obsessive–Compulsive Disorder; Operant Conditioning and Clinical Psychology; Pavlov, Ivan Petrovich (1849–1936); Psychological Treatment, Effectiveness of; Skinner, Burrhus Frederick (1904–90); Watson, John Broadus (1878–1958)

Bibliography

Ayllon T, Azrin N 1968 *The Token Economy, A Motivational System for Therapy and Rehabilitation*. Appleton Century Crofts, New York

Bancroft J H J 1983 *Human Sexuality and its Problems*. Churchill Livingstone, New York

Drummond L M 1993 The treatment of severe, chronic, resistant obsessive-compulsive disorder: An evaluation of an inpatient-programme using behavioural psychotherapy in combination with other treatments. *British Journal of Psychiatry* **163**: 223–29

Emmelkamp P M G, Wessels H 1975 Flooding in imagination v flooding *in vivo* – comparison with agoraphobics. *Behavior Research Therapy* **13**: 7–15

Fineberg N, Drummond L M 1995 Anxiety: Drug treatment or behavioural-cognitive psychotherapy. *CNS Drugs* **3**: 448–66

Foa E B, Goldstein A 1978 Continuous exposure and complete response prevention in the treatment of obsessive-compulsive neurosis. *Behavior Therapy* **9**: 821–29

Foa E B, Rothbaum B O, Riggs D S, Murdoch T B 1991 Treatment of posttraumatic stress disorder in rape victims: a comparison between cognitive-behavioral procedures and counselling. *Journal of Consulting and Clinical Psychology* **59**: 715–23

Gelder M G, Marks I M, Wolff H H 1967 Desensitization and psychotherapy in the treatment of phobic states: a controlled enquiry. *British Journal of Psychiatry* **113**: 53–73
Ghosh A, Marks I M, Carr A C 1988 Therapist contact and outcome of self-exposure treatment for phobias: a controlled study. *British Journal of Psychiatry* **152**: 234–8
Hand I, Lamontagne Y, Marks I M 1974 Group exposure (flooding) *in vivo* for agoraphobia. *British Journal of Psychiatry* **124**: 588–602
Hawton K, Salkovskis P M, Kirk J, Clark D M 1989 Cognitive Behaviour Therapy for psychiatric problems. In: *A Practical Guide*. Oxford University Press, Oxford, UK
Janet P 1903 *Les Obsessions et la Psychasthenie*. Bailliere, Paris
Jones M C 1924 Elimination of children's fears. *Journal of Experimental Psychology* **7**: 328
Keane T M, Fairbank J A, Caddell J M, Zimering R T 1989 Implosive (flooding) therapy reduces symptoms of P.T.S.D. in Vietnam combat veterans. *Behavior Therapy* **20**: 245–60
Kingdon D, Turkington D, John C 1994 Cognitive Behaviour Therapy of Schizophrenia: The amenability of delusions and hallucinations to reasoning. *British Journal of Psychiatry* **164**: 581–7
Levy R, Meyer V 1971 New techniques in behaviour therapy. Ritual prevention in obsessional patients. *Proceedings of the Royal Society of Medicine* **64**: 115
Marks I M 1981 Cure and care of neurosis. In: *Theory and Practice of Behavioural Psychotherapy*. Wiley, New York
Marks I M 1986 Behavioural psychotherapy. In: *Maudsley Pocket Book of Clinical Management*. Wright, Bristol, UK
Marks I M, Gray S, Cohen D, Hill R, Mawson D, Ramm E, Stern R 1983 Imipramine and brief therapist-aided exposure in agoraphobics having self-exposure homework. *Archives of General Psychiatry* **40**: 153–62
Marks I M, Hodgson R, Rachman S 1975 Treatment of chronic obsessive–compulsive neurosis by *in vivo* exposure. Two-year follow-up and issues in treatment. *British Journal of Psychiatry* **127**: 349–64
Marks I M, Lelliot P, Basoglu M, Noshirvani H, Monteiro W, Cohen D, Kasvikis Y 1988 Clomipramine, self-exposure and therapist-aided exposure for obsessive-compulsive rituals. *British Journal of Psychiatry* **52**: 522–34
Masters W H, Johnson V E 1966 *Human Sexual Response*. Churchill, London
Masters W H, Johnson V E 1970 *Human Sexual Inadequacy*. Churchill, London
Mathews A M, Gelder M G, Johnston D W 1981 *Agoraphobia: Nature and Treatment*. Guilford Press, New York
Mathews A M, Johnston D W, Lancashire M, Munby M, Shaw P M, Gelder M G 1976 Imaginal flooding and exposure to real phobic situations – treatment outcome with agorophobic patients. *British Journal of Psychiatry* **129**: 362–71
McDonald R, Sartory G, Grey S J, Cobb J, Stern R, Marks I M 1978 Effects of self-exposure instructions on agoraphobic outpatients. *Behaviour Research and Therapy* **17**: 83–5
Meyer V, Chesser E S 1970 *Behaviour Therapy in Clinical Psychiatry*. Penguin, London
Premack D 1959 Toward empirical behavior laws: 1. Positive reinforcement. *Psychological Review* **66**: 219–33
Rachman S J, Cobb J, Grey S, McDonald R, Mawson D, Sartory G, Stern R 1979 The behavioural treatment of obsessive-compulsive disorders with and without clomipramine. *Behavior Research and Therapy* **17**: 462–78
Rachman S, Hodgson R, Marks I M 1971 Treatment of chronic obsessive-compulsive neurosis. *Behavior Research Therapy* **9**: 237–47
Stampfl T J, Levis D G 1967 Essentials of implosive therapy: a learning theory based psychodynamic behavior therapy. *Journal of Abnormal Psychology* **72**: 496–503
Stravynski A, Marks I M, Yule W 1982 Social skills problems in neurotic outpatients. *Archives of General Psychiatry* **39**: 1378–85
Stern R S, Drummond L M 1991 *The Practice of Behavioural and Cognitive Psychotherapy*. Cambridge University Press, Cambridge, UK
Stern R S, Marks I M 1973 Brief and prolonged flooding: A comparison in agoraphobic patients. *Archives of General Psychiatry* **28**: 270–76
Tarrier N, Beckett R, Harwood S, Baker A, Yusupoff L, Ugarteburu I 1993 A trial of two cognitive-behavioural methods of treating drug-resistant residual psychotic symptoms in schizophrenic patients: I. Outcome. *British Journal of Psychiatry* **162**: 524–532
Warwick H M C, Clark D M, Cobb A M, Salkovskis P M 1996 A controlled trial of cognitive-behavioural treatment of hypochondriasis. *British Journal of Psychiatry* **69**: 189–95
Watson J P, Marks I M 1971 Relevant and irrelevant fear in flooding—a crossover study of phobic patients. *Behavior Therapy* **2**: 275–93
Watson J P, Rayner P 1920 Conditioned emotional reactions. *Journal of Experimental Psychology* **3**: 1
Wolpe J 1958 *Psychotherapy by Reciprocal Inhibition*. Stanford University Press, Stanford, CA

L. M. Drummond

Behavior Therapy: Psychological Perspectives

A new way of treating psychopathology, called behavior therapy, emerged in the 1950s. In its initial form this therapy was restricted to procedures based on classical and operant conditioning. Therapists who employ operant conditioning as means of treatment often prefer the term 'behavior modification.' Over the years many writers have broadened the scope of behavior therapy beyond conditioning to include any attempt to change abnormal behavior, thoughts, and feelings by applying the research methods used and the discoveries made by experimental psychologists in their study of both normal and abnormal behavior.

Over a number of years people in the clinical field —among them Joseph Wolpe and Arnold Lazarus in South Africa, H. J. Eysenck in Great Britain and B. F. Skinner, Albert Bandura, Albert Ellis, and Aaron T. Beck in the United States—began to formulate a new set of assumptions about dealing with clinical problems. Although there are areas of overlap, it is helpful to distinguish four theoretical approaches in behavior therapy—counterconditioning and exposure, operant conditioning, modeling, and cognitive behavior therapy.

1. Counterconditioning and Exposure

Counterconditioning is relearning achieved by eliciting a new response in the presence of a particular stimulus. A response (R_1) to a given stimulus (S) can be eliminated by eliciting a new response (R_2) in the presence of that stimulus. For example, in an early and now famous demonstration in 1924, Mary Cover Jones successfully treated a young boy's fear of rabbits by feeding him in the presence of a rabbit. The animal was at first kept several feet away and then gradually moved closer on successive occasions. In this way the fear (R_1) produced by the rabbit (S) was replaced by the stronger positive feelings evoked by eating (R_2).

The counterconditioning principle, deriving from earlier work by Pavlov and Guthrie, forms the foundation of an important behavior therapy technique, systematic desensitization, developed by Joseph Wolpe (1958). A person who suffers from anxiety works with the therapist to compile a list of feared situations, starting with those that arouse minimal anxiety and progressing to those that are the most frightening. Over a number of sessions and sometimes with the help of taped at-home practice, the person is also taught to relax deeply. Step-by-step, while relaxed, the person imagines the graded series of anxiety-provoking situations. The relaxation tends to inhibit any anxiety that might otherwise be elicited by the imagined scenes. The fearful person becomes able to tolerate increasingly more difficult imagined situations as he or she climbs the hierarchy over a number of therapy sessions. Reduction of fears in real life usually follows.

Wolpe hypothesized that counterconditioning underlies the efficacy of desensitization: a state or response antagonistic to anxiety is substituted for anxiety as the person is exposed gradually to stronger and stronger doses of what he or she fears. Some experiments (Davison 1968) suggest that counterconditioning accounts for the efficacy of the technique, but a number of other explanations are possible (see *Social Cognitive Theory and Clinical Psychology*). Most contemporary theorists attach importance to exposure *per se* to what the person fears; relaxation is then considered a useful strategy to encourage a frightened individual to confront what he or she fears, rather than a response that is substituted for the maladaptive anxiety. Whatever its mechanism of action, systematic desensitization and other exposure techniques have been shown to be effective in reducing a wide variety of fears, from specific phobias like fear of snakes and closed spaces to more complex fears such as social anxiety and agoraphobia (see *Psychological Treatments, Empirically Supported*).

Another type of counterconditioning, aversive conditioning, also played an important historical role in the development of behavior therapy. In aversive conditioning, a stimulus attractive to the patient is paired with an unpleasant event, such as a drug that produces nausea, in the hope of endowing it with negative properties. For example, a problem drinker who wishes to stop drinking might be asked to smell alcohol while he or she is being made nauseous by a drug. Aversive techniques have been employed to reduce smoking and drug use, and the socially inappropriate attraction that objects have for some people, such as the sexual arousal that children produce in pedophiles.

Aversion therapy has been controversial for ethical reasons. A great outcry has been raised about inflicting pain and discomfort on people, even when they ask for it. For example, in its early days, aversion therapy was used to try to change the sexual orientation of homosexuals. But in the late 1960s, gay liberation organizations began to accuse behavior therapists of impeding the acceptance of homosexuality as a legitimate lifestyle. Currently, aversion therapy is rarely used as the only treatment for a particular problem.

2. Operant Conditioning

Several behavioral procedures derive from operant conditioning, an analysis of overt behavior in terms of the conditions under which it occurs and the consequences that it elicits from the environment (see *Autonomic Classical and Operant Conditioning*; *Behavior Analysis, Applied*). Much of this work has been done with children, perhaps because a great deal of their behavior is subject to the oversight and control of others. Treatment typically consists of altering the consequences of problem behavior. For example, if it was established that the problem was motivated by attention seeking, the treatment might be to ignore it. Alternatively, the undesired behavior could be followed by time-out, a procedure wherein the person is banished for a period of time to a dreary location where positive reinforcers are not available.

Making positive reinforcers contingent on behavior is used to increase the frequency of desirable behavior. For example, a socially withdrawn child could be reinforced for playing with others. Similarly, positive reinforcement has been used to help children with autistic disorder develop language, to remediate learning disabilities, and to help children with mental retardation develop necessary living skills. Other problems treated with these methods include bedwetting, aggression, hyperactivity, disruptive classroom behavior, and tantrums (Kazdin and Weisz 1998).

2.1 The Token Economy

An early example of work within the operant tradition is the token economy, a system in which tokens (such as poker chips or stickers) are given for desired

behavior and can later be exchanged for pleasing items and activities. Ayllon and Azrin (1968) set aside an entire ward of a mental hospital for a series of experiments in which rewards were provided for activities such as making beds and combing hair, and were not given when behavior was withdrawn or bizarre. The patients, who averaged 16 years of hospitalization, were systematically rewarded for their ward work and self-care with plastic tokens that could later be exchanged for special privileges, such as listening to records, going to the movies, renting a private room, or enjoying extra visits to the canteen. The life of each patient was as much as possible controlled by this regime.

The rules of a token economy—the medium of exchange, the chores and self-care rewarded and by what number of tokens, the items and privileges that can be purchased and for how many tokens—are carefully established and usually posted so that the patients can understand what the payoff is for behaving in a particular way. These regimes have demonstrated how even markedly regressed adult hospital patients can be significantly helped to achieve more normal functioning by systematic manipulation of reinforcement contingencies. The role of cognitive factors, discussed below, was not formally acknowledged in the early operant conditioning work.

Token economy work demonstrates the positive impact of directing staff attention to rewarding self-care and recreational behaviors, and on the acquisition of social skills, in contrast to the more typical situation in which patients get attention more when they are acting maladaptively and sometimes dangerously. The beneficial effects of carefully constructed token economies have been shown to be markedly superior to routine hospital management, including antipsychotic drugs (Paul et al. 1997).

3. *Modeling*

Modeling has also been used in behavior therapy (see *Social Cognitive Theory and Clinical Psychology*). For example, people can reduce their unrealistic fears by watching both live and filmed encounters in which others gradually approach and successfully confront the things they are afraid of. Modeling is also part of the treatment for children with autistic disorder, helping them develop complex skills. Films depicting actors having pleasurable sex have been used to help sexually inhibited people overcome their discomfort with sexuality and learn sexual techniques (see *Sex Therapy, Clinical Psychology of*).

In an analogous fashion, some behavior therapists use role-playing in the consulting room. Particularly with patients who lack social skills, therapists demonstrate patterns of behaving that might prove more effective than those in which the patients usually engage and then have the patients practice them. In his behavior rehearsal procedures, Lazarus (2000) demonstrates exemplary ways of handling a situation and then encourages patients to imitate them during the therapy session. For example, a student who does not know how to ask a professor for an extension on a term paper might watch the therapist portray a potentially effective way of making the request. The clinician would then help the student practice the new skill. Similar procedures have helped patients with schizophrenia acquire social skills to allow them to deal more effectively with others, and with nonpsychotic patients to encourage greater assertiveness.

4. *Cognitive Behavior Therapy*

There is nothing either good or bad, but thinking makes it so.

(*Hamlet*, Act II, Scene 2)

The mind is its own place, and in itself
Can make a Heav'n of Hell, a Hell of Heav'n.

(*Paradise Lost*, line 247)

Behavior therapy initially eschewed any appeal to cognitive processes (Wolpe 1958), perhaps as part of efforts to distinguish it from insight-oriented therapies like psychoanalysis and its many variations, as well as humanistic and existential approaches. But it became increasingly apparent in the mid to late 1960s that an empirically-based understanding of therapeutic change would be inadequate without formal inclusion of cognitive variables (Bandura 1969, Davison 1966, London 1964) (see *Social Cognitive Theory and Clinical Psychology*).

One of the ways cognition entered into behavior therapy was via research on modeling. The question was how the observation of a model is translated into changes in overt behavior. In his original writings on modeling, Bandura asserted that an observer could somehow learn new behavior by watching others. Given the emphasis that much of experimental psychology places on learning through doing, this attention to learning without doing was important. But it did not delineate the processes that could be operating. A moment's reflection on the typical modeling experiment suggests the direction that theory and research have taken.

The observer, a child, sits in a chair and watches a film of another child making a number of movements, such as hitting a large, inflated plastic doll in a highly stereotyped manner, and hears the child in the film uttering peculiar sounds. An hour later the youngster is given the opportunity to imitate what was earlier seen and heard. The child is able to do so, as common sense and folk wisdom would predict. How can we understand what happened? Since the child did not do anything of interest in any motoric way while watching the film, except perhaps fidget in the chair, it would not

be fruitful to look at overt behavior for a clue. Obviously, the child's cognitive processes were engaged, including the ability to remember later on what had happened. Data such as these led some behavioral researchers and clinicians to include cognitive variables in their analyses of psychopathology and therapy.

4.1 Approaches to Cognitive Behavior Therapy

Cognitive behavior therapy (CBT) applies theory and research on cognitive processes to alter cognition in the interests of effecting favorable change in emotions and behavior (see *Cognitive Theory: ACT*). CBT has become a blend of the cognitive and behavioral perspectives. Cognitive behavior therapists pay attention to private events—thoughts, perceptions, judgments, self-statements, and even tacit (unconscious) assumptions—and have studied and manipulated these processes in their attempts to understand and modify overt and covert disturbed behavior. But they do not neglect the behavioral factors reviewed above that influence emotion, cognition, and overt behavior (Bandura 1969).

4.1.1 Beck's cognitive therapy. The psychiatrist Aaron Beck is one of the leading cognitive behavior therapists. He developed a cognitive therapy for depression based on the idea that depressed mood is caused by distortions in the way people perceive life experiences (Beck 1976). For example, a depressed person may focus exclusively on negative happenings and ignore positive ones. Beck's therapy aims to persuade patients to change their opinions of themselves and the way in which they interpret life events. For example, when a depressed patient expresses feelings that nothing ever goes right, the therapist uses Socratic strategies to help the person identify counter-examples. The general goal of Beck's therapy is to provide patients with experiences, both inside and outside the consulting room, that will alter their negative schemas, their general beliefs about themselves and their environment. This therapy has shown its value, particularly in alleviating depression (DeRubeis et al. 1999), but elements of Beck's approach can be found as well in effective cognitive–behavioral interventions for such problems as bulimia nervosa, panic disorder (see especially Barlow et al. 2000), social phobia, and generalized anxiety disorder (see *Clinical Psychology: Manual-based Treatment*; *Cognitive Therapy*; *Psychological Treatments, Empirically Supported*).

4.1.2 Ellis's rational–emotive behavior therapy. Albert Ellis is another leading cognitive behavior therapist (see *Behavior Psychotherapy: Rational and Emotive*). His principal thesis is that sustained emotional reactions are caused by internal sentences that people repeat to themselves and that these self-statements reflect sometimes unspoken assumptions—irrational beliefs—about what is necessary to lead a meaningful life. In Ellis's rational–emotive behavior therapy (REBT) (Ellis 1962), the aim is to eliminate self-defeating beliefs through a rational examination of them. Anxious persons, for example, may create their own problems by making unrealistic demands on themselves or others, such as, 'I must win love from everyone.' Ellis proposed that people interpret what is happening around them, that sometimes these interpretations can cause emotional turmoil, and that a therapist's attention should be focused on these beliefs rather than on historical causes or, indeed, on overt behavior.

Ellis used to list a number of irrational beliefs that people can harbor, such as the assumption that they must be thoroughly competent in everything they do. More recently, he has shifted from a cataloguing of specific beliefs to the more general concept of 'demandingness,' that is, the musts or shoulds that people impose on themselves and on others. Thus, instead of wanting something to be a certain way, feeling disappointed, and then perhaps engaging in some behavior that might bring about the desired outcome, the person demands that it be so. It is this unrealistic, unproductive demand, Ellis hypothesizes, that creates the kind of emotional distress and behavioral dysfunction that bring people to therapists, and that should be altered in order to create a more realistic, less absolutistic approach to life's demands. Research supports the value of REBT in alleviating a wide range of anxiety-related problems, including interpersonal performance anxiety, test anxiety, anger, and depression; and it may also be of use in a preventive way by teaching children that their self-worth is not utterly dependent on their endeavors.

4.1.3 Behavioral medicine. A wide range of cognitive–behavioral strategies have been applied with success in a field called 'behavioral medicine,' defined as the study and application of empirically supported techniques for the prevention and amelioration of physical problems (see *Behavioral Medicine*). For example, relaxation training has been found effective in reducing blood pressure in borderline hypertension, possibly by lessening the anger that patients experience when frustrated or provoked (see *Hypertension: Psychosocial Aspects*). Cognitive and behavioral interventions are useful also in encouraging people, even older adults, to alter their lifestyle in ways that contribute to better health and even to help cancer patients cope with their illnesses and with the pain associated with the treatment of them. There is ever-increasing appreciation of the importance of

psychological factors in encouraging people to adopt healthier lifestyles, adhere to sometimes difficult treatment regimens, and to cope with negative emotions that, if unchecked, can exacerbate the course of a physical illness as well as affect the success of a medical intervention (see *Self-efficacy and Health*).

4.2 Conceptual Issues in Cognitive Behavior Therapy

Some criticisms of CBT should be noted. The concepts on which it is based (e.g., schema) are somewhat slippery and not always well-defined. Furthermore, cognitive explanations of psychopathology do not always explain much. For example, that a depressed person has a negative schema tells us that the person has a pessimistic outlook on things. But such a pattern of thinking is actually part of the diagnosis of depression. What is distinctive in the cognitive paradigm is that the thoughts are given casual status; that is, the thoughts are regarded as causing the other features of the disorder, such as profound sadness. Left unanswered is the question of where the negative schema came from in the first place, and to what extent it creates negative emotion and maladaptive behavior rather than being only a correlate of them.

Is the cognitive point of view basically different and separate from the behavioral paradigm? Much of the preceding discussion suggests that it is. But the growing field of cognitive behavior therapy gives one pause, for its workers study the complex interplay of beliefs, expectations, perceptions, and attitudes on the one hand, and overt behavior on the other. For example, Albert Bandura (1977), a leading advocate of the cognitive viewpoint, argues that different therapies produce improvement by increasing people's sense of self-efficacy, a belief that they can achieve desired goals. But, at the same time, he argues that changing behavior through behavioral techniques is the most powerful way to enhance self-efficacy.

Therapists such as Ellis and Beck emphasize direct alteration of cognitions through argument, persuasion, Socratic dialogue, and the like to bring about improvements in emotion and behavior. Complicating matters still further, Ellis, Beck, and their followers also place considerable importance on homework assignments that require clients to behave in ways in which they have been unable to behave because they are hindered by negative thoughts (see *Cognitive Therapy*). Indeed, one study failed to find added benefit from the cognitive components of Beck's cognitive therapy (CT) as compared with the behavioral components alone (Jacobson et al. 1996). Ellis even changed the label for his approach from rational–emotive therapy to rational–emotive behavior therapy in order to highlight the importance of overt behavior. Therapists identified with cognitive behavior therapy work at both the cognitive and behavioral levels, and most of those who use cognitive concepts and try to change beliefs with verbal means also use behavioral procedures to alter behavior directly.

This issue is reflected in the terminology used to refer to people such as Beck and Ellis. Are they cognitive therapists or cognitive behavior therapists? The latter term is preferred by behavior therapists because it denotes both that the therapist regards cognitions as major determinants of emotion and behavior and that he or she maintains the focus on overt behavior that has always characterized behavior therapy. Nonetheless, Beck, even though he assigns many behavioral tasks as part of his therapy, is usually referred to as the founder of cognitive therapy (CT), and Ellis's rational–emotive therapy (RET) used to be spoken of as something separate from behavior therapy.

5. Concluding Comment

Towards the end of the twentieth century, the increasing role of managed care in the United States—entering the picture much later than in other Western countries—made mental health professionals more aware of the need to employ the best-validated and most efficient interventions available. This greater level of accountability is having a revolutionary effect on who gets treated, for how long, and at what cost. Not all patients are helped with behavioral and cognitive-behavioral procedures, and the data are not fully available as to what kinds of problems respond better to such approaches than to others (such as psychoanalytically oriented, humanistic/existential, or drug and other somatic therapies). An advantage of the approaches reviewed in this article is that they possess a high degree of accountability: procedures are clearly spelled out, links to the science of behavior change are a defining characteristic, and evaluations of outcome are routine. These features of behavioral and cognitive therapies would seem to bode well for the continuing development of more and better interventions, to the benefit both of the science and the profession as well as the communities that are served.

See also: Behavior Therapy: Psychiatric Aspects; Behavior Therapy with Children; Cognitive and Interpersonal Therapy: Psychiatric Aspects; Cognitive Therapy; Operant Conditioning and Clinical Psychology

Bibliography

Ayllon T, Azrin N H 1968 *The Token Economy: A Motivational System for Therapy and Rehabilitation*. Appleton-Century-Crofts, New York

Bandura A 1969 *Principles of Behavior Modification*. Holt, Rinehart & Winston, New York

Bandura A 1977 Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* **84**: 191–215
Barlow D H, Gorman J M, Shear M K, Woods S W 2000 Cognitive–behavioral therapy, imipramine, or their combination for panic disorder: A randomized controlled trial. *Journal of the American Medical Association* **283**: 2529–74
Beck A T 1976 *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York
Davison G C 1966 Differential relaxation and cognitive restructuring in therapy with a 'paranoid schizophrenic' or 'paranoid state.' *Proceedings of the 74th Annual Convention of the American Psychological Association*. American Psychological Association, Washington DC
Davison G C 1968 Systematic desensitization as a counterconditioning process. *Journal of Abnormal Psychology* **73**: 91–9
DeRubeis R J, Gelfand L A, Tang T Z, Simons A D 1999 Medications vs. cognitive behavioral therapy for several depressed outpatients: A mega-analysis of four randomized comparisons. *American Journal of Psychiatry* **15**: 1007–13
Ellis A 1962 *Reason and Emotion in Psychotherapy*. Lyle Stuart, New York
Jacobson N S, Dobson K S, Truax P, Addis M, Koerner K, Gollan J, Gortner E, Prince S 1996 A component analysis of cognitive behavioral treatment for depression. *Journal of Consulting and Clinical Psychology* **64**: 295–304
Kazdin A E, Weisz J R 1998 Identifying and developing empirically supported child and adolescent treatments. *Journal of Consulting and Clinical Psychology* **66**: 19–36
Lazarus A A 2000 Multimodal therapy. In: Dumont F, Corsini R J (eds.) *Six Therapists and One Client*. Springer, New York, pp. 145–66
London P 1964 *The Modes and Morals of Psychotherapy*. Holt, Rinehart & Winston, New York
Paul G L, Stuve P, Cross J V 1997 Real-world inpatient programs: Shedding some light: A critique. *Applied and Preventive Psychology: Current Scientific Perspectives* **6**: 193–204
Wolpe J 1958 *Psychotherapy by Reciprocal Inhibition*. Stanford University Press, Stanford, CA

G. C. Davison

Behavior Therapy with Children

Child behavior therapy, including behavior modification and cognitive therapy, consists of a group of diverse but related scientifically-based approaches to the assessment and treatment of children experiencing behavioral and emotional difficulties. Although contemporary behavior therapy with children is characterized by a plurality of viewpoints, techniques, and theoretical rationales, a distillation of these differing perspectives reveals several common features that define the essence of child behavior therapy. These features include '(1) principles of behavioral psychology, most notably principles of learning, (2) use of strategies or procedures that are methodologically sound and empirically validated, and (3) application of such principles and procedures to adjustment problems of children and adolescents' (Ollendick 1986). Presented here is an overview of the historical roots of child behavior therapy, an explication of key concepts underlying the approach, and examples of classic and contemporary uses of behavior therapy with children.

1. Historical Roots

The historical roots of child behavior therapy can be traced to the work of John B. Watson, a key figure in the rise of behaviorism in America during the early 1900s. Watson, who rejected the trends of the time emphasizing mentalistic causes of behavior, maintained that learning processes were the basis of all human behavior. In perhaps the first historical antecedent to modern behavior therapy, one of Watson's students, Mary Cover Jones, applied classical conditioning principles to alleviate a young boy's intense fear of rabbits, by rewarding an incompatible response to fear, and using successive approximations to introduce the rabbit (Jones 1924) (see *Classical Conditioning and Clinical Psychology*).

With the exception of a few isolated examples, such as the classic 'bell and pad' method of treating nocturnal enuresis (bedwetting) (Mowrer and Mowrer 1938), there was a lag of some 30 years before interest in behavioral approaches in the treatment of children re-emerged in psychology. This resurgence was precipitated in large part by discontent with psychoanalysis, the prevailing therapeutic model of the day. Institutional settings in which developmentally delayed children often resided provided the ideal, tightly controlled environments in which to evaluate emergent behavioral treatments. In most cases, early behavioral techniques were developed in animal laboratories by experimental psychologists with the objective of demonstrating the utility of applying specific learning principles to rather circumscribed child behavioral symptoms (Mash 1998).

During the 1960s, behavior therapy relied primarily upon operant conditioning procedures, the principles of which had been pioneered by B. F. Skinner in the 1940s (see *Operant Conditioning and Clinical Psychology*. These approaches were used successfully with developmentally delayed children in institutional settings. Lovaas and Simmons (1969), for example, were able dramatically to improve the behavior of autistic children by applying operant learning principles to manipulate the environmental antecedents and consequences of behavior. In the 1960s and 1970s, behavior therapy was extended to school settings where it was applied to classroom misbehavior. Through this work it was found that ignoring disruptive behavior, while reinforcing positive behavior through praise (or atten-

tion), although the opposite of what many teaching manuals suggested, was useful in reducing the inappropriate classroom behavior (Becker et al. 1967).

Yet, even as behavior modification techniques with children were meeting with initial success, the field was criticized for ignoring some crucial issues, such as whether improvements made in the treatment settings could be maintained across multiple situations over an extended period of time (Abidin 1975). By the middle 1970s, however, the field largely had responded to these shortcomings. Because behavioral changes in response to environmental contingencies tended to be short-lived, the field shifted in emphasis toward the teaching of adaptive behaviors, general problem-solving strategies, and coping skills that generalize more easily across settings. In the 1980s, the behavioral technique of 'time out,' which involves removing children from positively reinforcing situations for a specified period time, became widely used by parents and teachers alike in response to aggressive, destructive, and otherwise unmanageable child behaviors. The modeling of appropriate behavior was also shown to be very effective, and was applied to real world problems for children such as in the use of puppets to demonstrate coping strategies and appropriate behavior for children (Knell 1997).

2. *Principles of Learning and Behavior Change*

Behavior therapists generally hold that both normal and abnormal behavior emanates from the same principles of learning, derived from laboratory studies. These tenets have been extended and applied in a spectrum of strategies to aid in the modification problematic child behaviors. Three types of learning form the building blocks of these techniques and are central to the application of behavior therapy with children. Presented here are brief descriptions of the principles of classical conditioning, operant conditioning, and social learning.

2.1 *Classical Conditioning*

Many are familiar with the pioneering work of Russian scientist Ivan Pavlov (1849–1936) who conditioned dogs to salivate in response to the simple sound of a tone after repeatedly pairing the tone with the presentation of food. In an early example of classical conditioning with children, Watson and Rayner (1920) demonstrated that a fear response could be conditioned in a young boy by repeatedly pairing a loud noise with the presentation of a white rat, of which the child had no naturally occurring fear. After a number of pairings, the presentation of the rat alone became a conditioned stimulus, eliciting a startle response from the boy that had previously only occurred in the presence of the loud noise.

2.2 *Operant Conditioning*

In contrast to classical conditioning, which maintains that behaviors can be *elicited* by preceding conditioned stimuli, operant learning principles hold that behaviors are *emitted* from within, in response to the environmental stimuli that follow them. Operants themselves consist of actions that are performed on the environment that produce some consequence. Operant behaviors that bring about reinforcing environmental changes (i.e., if they provide some reward to the individual or eliminate an aversive stimuli) are likely to be repeated. In the absence of reinforcement, operants are weakened. Removing consequences (ignoring) can decrease or completely eliminate many annoying child behaviors such as whining.

B. F. Skinner, an experimental psychologist considered to be the primary proponent of operant learning theory, distinguished between two important learning processes: reinforcement (both positive and negative) and punishment. Positive reinforcement is the process by which a stimulus or event, occurring after a behavior, increases the future occurrence of a behavior. Negative reinforcement also results in an increase in the frequency of a behavior, but through a process of contingently removing an aversive stimulus following a behavior. Punishment refers to the introduction of an aversive stimulus, or removal of a positive one, following a response, resulting in a decrease in the future probability of that response. Skinner also observed that extinction occurs when the absence of any reinforcement results in a decrease or a reduction in response frequency.

2.3 *Social Learning*

In the 1960s, Albert Bandura defined the process of social learning, in which a learner observes and imitates the behavior of another (known as the model). In contrast to the radical behaviorism of Skinner, Bandura's social learning theory posited that cognitive expectations of reinforcement alone may motivate much of human behavior, resulting in action without the necessity of the actual experience of reinforcement (Bandura 1969) (see *Social Cognitive Theory and Clinical Psychology*). Thus, behavior is the product of a reciprocal interaction between the individual and environmental contingencies. Infants as young as 16 months are known to imitate their peers (Eckerman and Stein 1982), and by the time children reach kindergarten, they are experienced social learners. Modeling can increase negative behavior (e.g., fear of thunderstorms exhibited by parents) or positive responses (e.g., appropriate social skills).

3. *Diagnosis and Behavioral Assessment*

Historically, behavior therapists have eschewed traditional classification schemes used in mental health

diagnosis, maintaining that such systems demonstrate insufficient reliability and validity for many specific disorders to warrant widespread use. Consequently, behavior therapists have developed an alternative set of techniques, known as behavioral assessment strategies, as a means to gather information and precisely delineate behaviors to be targeted for change. These strategies are also used for the collection of data to measure the effectiveness of interventions both during and after treatment. Comprehensive behavioral assessments typically occur along multiple dimensions, which include investigations of children's observable behaviors, cognitive and emotional processes, physiological responses, and the various environmental influences impacting the child.

Three of the most common types of methods employed during behavioral assessments with children are interviews, behavior rating scales and checklists, and direct observation of variables of interest (Van Hasselt and Hersen 1993) (see *Behavioral Assessment*). Multiple informants are utilized, including the child, parents and other family members, teachers, and daycare providers. Interviews are useful in establishing rapport with the child and adults important to historical assessment and treatment, in gaining information about the presenting problems and their maintaining conditions, and in selecting target behaviors to address. Child self-report measures typically utilize a checklist or rating scale format to obtain information about the child's own behaviors, cognitions, and feelings. Parents and teachers may complete behavior checklists or rating scales to provide information about the nature and severity of a problem. Checklists or rating scales can be broad in scope, assessing global problem areas, or may be narrowly focused upon a specific problem behavior. Finally, direct observation of the child and/or family may be conducted by clinicians; frequently, a structured observational coding system is used.

4. *Contemporary Applications of Child Behavior Therapy*

Behavior therapy currently is utilized successfully to treat a variety of behavioral and emotional problems in children that arise in various settings and situations (e.g., in the home, at school, and as a result of medical conditions). Presented next are several examples of childhood problems that can be ameliorated through behavioral techniques. These include fears and phobias, attentional and hyperactivity problems, conduct problems, and various medically-related conditions. For each example, a brief explanation of the problem is provided, followed by a description of typical assessment techniques and discussion of common behavioral treatment approaches. This presentation is by no means a comprehensive catalogue of behavioral interventions; it merely illustrates prototypical applications of basic behavioral principles to the treatment of childhood disorders.

4.1 *School Phobia*

Fears and anxiety disorders are a common occurrence for many young children. However, when a fear of school is involved, the undesirable outcome may be refused attendance on the part of a child. Assessment of this problem involves the use of a child interview, self-report measures, and parent checklists to help distinguish between children whose school refusal stems from difficulty separating from parents, is secondary to conduct problems, or is simply a fearful reaction to school-related stimuli. A typical behavioral approach is to gradually shape children's behavior by rewarding him or her for attending school for progressively longer periods of time. Conversely, punishment in the form of removal of reinforcers and privileges for non-attendance may also be effective.

4.2 *Attentional and Hyperactivity Problems*

Attentional deficits and hyperactive behavior are often most evident in school settings, where children experiencing these difficulties have trouble staying on-task during classroom activities. Behavioral assessment of attention deficits and hyperactivity typically include the completion of parent and teacher rating scales, in combination with a structured interview and direct observation of children's classroom behaviors, plus tests of child attention in the presence of certain distractions. The use of operant conditioning procedures to manage children's behavior has been common in the treatment of attentional and impulse control problems. This procedure, known as contingency management, may include the use of positive reinforcement, in the form of tokens, to increase the frequency of appropriate classroom behaviors, such as working quietly, remaining seated, and paying attention to the teacher. Tokens which can be earned or lost (also known as response cost) in conjunction with social contingencies (e.g., teacher praise) and negative reprimands may be the most effective means of reducing off-task behavior (Rapport et al. 1982). Attention deficit disorders in children are often most effectively treated in combination with psychopharmacological intervention in the form of tranquilizers or more frequently, stimulants, such as Ritalin (Barkley et al. 1984).

4.3 *Oppositional and Defiant Behavior*

Oppositional and defiant behavior involves a collection of disruptive, noncompliant, and negativistic

behaviors that find a child at frequent odds with authority figures. The nature of these behaviors exceeds what would be expected from a 'typical' child at the same developmental stage. Oppositionality and defiance are assessed by means of interviews with the parents (with an emphasis on situational factors) and with the child, as well as behavior checklists completed by the child, parents, and teachers and observation in the typical 'oppositional' settings. Peer relation measures and measures of parental/family functioning may also be utilized. Parent training in effective child behavior management strategies typically is provided and involves provision of immediate, salient, and consistent consequences for appropriate and inappropriate behavior, an explicit rewards program, a proactive, nonemotional approach to misbehavior, and activities to improve the parent–child relationship. A structured didactic approach (e.g., Barkley 1997) or a 'coaching' model may be utilized (Hembree-Kigin and McNeil 1995). Also, reciprocity in negative family interactions is addressed, as families are taught to use effective problem-solving and communication skills.

4.4 Pediatric Medical Issues

Behavior therapy has also been applied to assist medically ill and injured children. In the medical arena, behavior therapists have developed a number of treatments to aid children in coping with the behavioral and emotional components of medical illness and its treatment. Behavioral interventions have targeted children with chronic medical conditions (e.g., asthma, arthritis, cancer) as well as terminal illnesses (e.g., AIDS, cystic fibrosis) and unintentional injuries. The conditions themselves, as well as the aversive treatment regimens they often entail, can negatively impact child and family functioning. Behavior therapists work with children to alleviate symptoms such as chronic pain, to adopt necessary self-care skills, and to increase compliance with prescribed treatments. The modeling of insulin injections for diabetic children, biofeedback techniques to control pain, and the use of relaxation training and distraction procedures to reduce distress during difficult diagnostic procedures, are examples of effective behavioral interventions.

5. New Directions in Child Behavior Therapy

5.1 Prevention

The traditional delivery model utilized by behavior therapists has focused upon the treatment of children with pre-existing behavioral or emotional difficulties. In recent years, however, behavior therapists have expanded the scope of their work by applying behavioral principles to address broader social problems through systemic-level preventive interventions. The concept of intervening with groups of high-risk individuals prior to the onset of a problem is particularly compelling when applied to children who can potentially be spared years of psychological distress through these efforts.

Preventive interventions have targeted a number of problem areas at various points along the developmental continuum. At the earliest stage, prenatal interventions have targeted children indirectly by teaching high-risk mothers about proper care and protection of the developing fetus, in order to prevent prematurity, low birth weight, and specific disabilities and disorders in newborns. In the first years of life, attempts have been made to spare young children from the pervasive harm resulting from child abuse and neglect by identifying those at risk and intervening with parents to prevent the onset and repeated occurrence of maltreatment. Unintentionally occurring injuries, which are the leading cause of death between one and 18 years of age, have also been the focus of successful preventive interventions. Finally, among school aged children, behavioral approaches have targeted the early risk factors signaling the likelihood of poor academic performance, inadequate social skills, as well as stranger abduction and sexual abuse.

5.2 Developmental Issues

In recent years, behavioral therapists have paid increasing attention to developmental issues in the treatment of children. Developmental factors that are of clinical relevance include the use of normative data about child development for the purpose of assessment, and the consideration of age, gender, and ethnic issues in treatment (Van Hasselt and Hersen 1993). There is also increasing recognition that interventions targeting chronic medically related issues occur within and are evaluated against an ever-changing backdrop of growth and development.

6. Summary

Since its inception, the hallmark of behavior therapy has been an adherence to treatment procedures that can be validated scientifically. Thus, behavior therapy with children is a constantly evolving form of treatment, which continually draws upon new findings from empirical research for the development and refinement of therapeutic interventions. In its early incarnations, behavior therapy with children was limited to relatively simplistic applications of basic experimental learning theory to circumscribed behavior problems. However, fueled by increased evidence of its effectiveness, behavior therapy has evolved into a more comprehensive and complex, yet mainstream therapeutic approach, with widespread ap-

plicability to a variety of children's mental health and behavioral problems. Behavior therapy with children continues to expand its purview to encompass broader social problems with a greater sensitivity to developmental processes of youth.

See also: Anxiety Disorder in Children; Attention-deficit/Hyperactivity Disorder (ADHD); Conduct Disorder

Bibliography

Abidin R R 1975 Negative effects of behavioral consultation: 'I know I ought to, but it hurts too much'. *Journal of School Psychology* **13**: 51–7

Bandura A 1969 *Principles of Behavior Modification*. Holt Rinehart, and Winston, New York

Barkley R A 1997 *Defiant Children*, 2nd edn. Guilford Press, New York

Barkley R A, Karlsson J, Strzelecki E, Murphy J 1984 Effects of age and Ritalin dosage on the mother-child interactions of hyperactive children. *Journal of Consulting and Clinical Psychology* **52**: 750–58

Becker W C, Madson C H Jr, Arnold C R, Thomas D R 1967 The contingent use of teacher attention and praise in reducing classroom behavior problems. *The Journal of Special Education* **1**: 287–307

Eckerman C O, Stein M R 1982 The toddler's emerging interactive skills. In: Rubin K H, Ross H S (eds.) *Peer Relationships and Social Skills in Childhood*. Springer-Verlag, New York, pp. 41–71

Hembree-Kigin T L, McNeil C B 1995 *Parent–Child Interaction Therapy*. Plenum, New York

Jones M C 1924 A laboratory study of fear: The case of Peter. *Journal of Genetic Psychology* **31**: 308–15

Knell S M 1997 Cognitive–behavioral play therapy. In: O'Connor K, Braverman L M (eds.) *Play Therapy Theory and Practice: A Comparative Presentation*. Wiley, New York, pp. 79–99

Lovaas O I, Simmons J Q 1969 Manipulation of self-destruction in three retarded children. *Journal of Applied Behavior Analysis* **2**: 143–57

Mash E J 1998 Treatment of child and family disturbance: A behavioral systems perspective. In: Mash E J, Barkley R A (eds.) *Treatment of Childhood Disorders*, 2nd edn. Guilford Press, New York, pp. 3–51

Mowrer O H, Mowrer W M 1938 Enuresis: A method for its study and treatment. *American Journal of Orthopsychiatry* **8**: 436–47

Ollendick T H 1986 Behavior therapy with children and adolescents. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York, pp. 565–624

Rapport M D, Murphy H A, Bailey J S 1982 Ritalin vs. response cost in the control of hyperactive children: A within-subject comparison. *Journal of Applied Behavior Analysis* **15**: 205–16

Van Hasselt V B, Hersen M 1993 Overview of behavior therapy. In: Van Hasselt V B, Hersen M (eds.) *Handbook of Behavior Therapy and Pharmacotherapy for Children: A Comparative Analysis*. Allyn and Bacon, Boston, pp. 1–12

Watson J B, Rayner R 1920 Conditional emotional reactions. *Journal of Experimental Psychology* **3**: 1–14

D. DiLillo and L. Peterson

Behavioral Assessment

Behavioral assessment is the identification and measurement of a behavioral problem and its controlling variables (environmental, biological, or personal) for the purposes of manipulating variables relevant to target responses and, ultimately, evaluating and measuring change. From this commonly accepted definition two important characteristics of behavioral assessment arise: it is strongly linked to the clinical context, and it has emerged from a specific theoretical approach of clinical psychology (see *Behavior Therapy: Psychological Perspectives*).

This article deals with the following issues: (a) a historical overview; (b) the nature of behavioral assessment; (c) the process of behavioral assessment, treatment, and evaluation; and (d) the future of behavioral assessment.

1. A Historical Overview

The history of behavioral assessment cannot be separated from behaviorism as a paradigm of scientific psychology and its clinical application. Thus, behavioral assessment began when complex human behaviors were first considered as a subject for scientific study (e.g., Staats 1963, Bandura 1969). Since the 1960s, four main phases have been identified.

1.1 The Starting Point

Although there are several origins of behavioral assessment, it is commonly accepted that Kanfer and Saslow's article *Behavioral Analysis* (1965) represents the foundation stone, with its seven-step guide to behavioral assessment (problematic behaviors and situation, motivational analysis, developmental analysis, self-control analysis, social relationships, and sociocultural and physical environment). This paper introduced behavioral assessment as a main issue—chapters devoted to behavioral assessment were included in books on behavior therapy/behavior modification, and references to it appeared in important behavioral and clinical journals.

1.2 The Period of Growth

In the late 1970s there began what Nelson (1983) called 'the honeymoon' of behavioral assessment. Hundreds of articles, several handbooks, and two journals totally devoted to behavioral assessment (*Behavioral Assessment* and *Journal of Behavioral Assessment*) were published. The main topics covered were: similarities and differences between behavioral and psychological assessment; theoretical supporting models (from radical behaviorist assumptions to the cognitive-behavioral perspective); functional analysis

as the main strategy; idiographism as the principal methodological perspective; the triple response classes of behavior as the basic classification system for operationalizing behavioral problems; the psychometric principle applied to behavioral measurement techniques; the types of method used; and the development of new behavioral assessment measurement devices for assessing clinical problems.

1.3 The Crisis

From the 1980s, we can identify two main profiles of behavioral assessment: the first supports the notion of behavioral assessment as a homogeneous corpus of knowledge, methodological tools and strategies, and a set of techniques; the second paints a diffuse picture of behavioral assessment as merely a 'content' field of psychological assessment, not as a theory-based model within assessment. Indeed, as Bellack and Hersen (1988) pointed out, there was 'neither a single accepted definition of the field, nor a consistent set of methods' (pp. 612–13).

This crisis, referred to by Cone (1993) as a 'schism,' and by Nelson as a period of 'disillusionment,' had several important repercussions, reflected in changes in two important publications: the *Journal of Behavioral Assessment* was renamed *Journal of Psychopathology and Behavioral Assessment*, and *Behavioral Assessment* disappeared in 1992, becoming subsumed as a section in the journal *Behaviour Research and Therapy*.

1.4 Overcoming the Crisis

Crises sometimes have positive net results, and this crisis was overcome. Over the 1990s, linked to the success of the behavioral paradigm, the output of behavioral assessment literature has been stable, and it is even embedded in other general publications of psychological assessment and behavior modification and therapy, and taught in doctoral and clinical practice training programs (Cone 1993, Fernández-Ballesteros 1993, Haynes 1998). Behavioral assessment has expanded its focus and continued its integration within a generic, scientifically based bio-psycho-social model of human behavior and behavior disturbances (Bellack and Hersen 1988).

2. The Nature of Behavioral Assessment

Although, throughout its evolution, the characteristics of behavioral assessment have changed somewhat, it should be emphasized that, at all times, its main goal has been behavioral change. Therefore, all conceptual and methodological features emerge from the attempt to define and measure a particular behavioral problem and its causal or controlling conditions and design the best treatment. The following five traits could be considered the essence of behavioral assessment: functional analysis, triple response modes and multicausality, idiographism, multimethodism, and the experimental method.

2.1 Functional Analysis

The main tool of behavioral assessment is functional analysis. It was defined by Skinner (1953) as '…the (stipulation of) external variables of which behavior is a function provide for what may be called a causal or functional analysis. We undertake to predict and control the behavior of the organism. This is our "dependent variable"—the effect for which we are to find the cause. Our "independent variables"—the cause of behavior—are the external conditions of which behavior is a function' (p. 35). Over the history of behavioral assessment this definition has been extended, since other behaviors (besides overt behavior as dependent variable) and other sites of causation than external conditions (as independent variable) have been integrated; nevertheless, functional analysis continues to represent the core (see *Behavior Analysis, Applied*).

2.2 Triple Response System and Multicausality

The second characteristic of behavioral assessment arose from the complexity of behavioral problems both in their definition and their causation. Behavioral problems were to be defined through the triple response mode—motor, cognitive, physiological—and diverse potential causes—multicausality—were to be accepted. In other words, behavioral problems require specification through several manifestations, and are usually caused by several conditions that cannot be reduced to those external circumstances, since relevant biological and personal (as well as environmental) conditions also contribute to their explanation.

For example, a person's depression should be defined through cognitive (feelings of loneliness, attention and concentration problems), physiological (sleep disturbance) and motor (low rate of social behaviors and physical activity) behaviors. This problem can be explained functionally by several conditions of the subject, such as a reinforcement system deficit, inadequate motivational system (personal condition), or a dysfunction in biological conditions—and usually by the interaction of all of these factors.

2.3 Idiographism

The object of study of behavioral assessment is the behavior of a given person. This means identifying and defining behavioral problems and their controlling variables idiographically. In other words, psychologists should consider a single case and its specific circumstances.

This position does not mean that there are no general principles applicable to human behavior, but it does mean that every behavioral problem may have specific components and causal conditions that should be considered in order to proceed to behavioral change. Idiographism does not reject general classification systems (for communication purposes) or the administration of standard tests (for comparing our case with a sample of the general population), but it demands the measurement of the subject's particular behavioral problem and its potential causal variables.

2.4 Multimethodism

Since overt behaviors and external conditions were the basis of Skinnerian functional analysis, at the very beginning of behavioral assessment, observational procedures were also the fundamental methods proposed. However, considering the evolution of behavioral psychology, since behavioral problems should be described through the triple response mode—motor, cognitive, physiological—other methods of assessment than observation of overt behavior and other informants (than the subject) should be considered. In sum, multimethodism is one of the most important characteristics of behavioral assessment. This technological change has also, of course, occurred in other social sciences, and today multimethodism is the most widespread approach.

Methods in behavioral assessment are diverse. Nevertheless, observation (by experts or relatives) of motor behavior continues to be a commonly used method of assessment for assessing both overt behavior and its external antecedents and consequences. Other methods, such as self-report, self-monitoring, etc., are also administered for assessing public and private events (feelings of sadness, automatic thoughts, etc.). The most important characteristics in the administration of these techniques are that they usually include environmental antecedents and consequences of these private events, which are not considered as intrapsychic traits. Finally, when physiological responses are involved in a behavioral problem, physiological equipment is used and psychophysiological variables measured. Therefore, we can conclude that in the utilization of instruments and procedures in data collection the only restriction comes not from the method but from the level of inference of the assessed variables; in other words, in behavioral assessment, a lower level of inference is prescribed. In sum, multiple response modes, multiple methods, and multiple informants are the most common measurement devices and strategies for assessing variables with low levels of inference.

2.5 Experimental Approach

Behavioral assessment provides the basis for behavioral change, behavioral change requires treatment, and treatment demands experimental manipulations and evaluation (Hayes and Nelson 1986). This is why, throughout the history of behavioral assessment, the experimental method has been a constant among its basic characteristics.

Psychological assessment has several purposes; a subject can be assessed with the purpose of description in psychological/behavioral terms, or of predicting future behavior (for personnel selection or counseling). However, and mainly in clinical settings, behavioral change through treatment is the most important goal, and one of the features that distinguish behavioral assessment from other types of psychological assessment is, precisely, the use of the experimental method.

3. The Process of Behavioral Change: Assessment, Treatment, and Evaluation

One of the most important features of behavioral assessment is its role in behavior modification or behavioral change. Thus, a person with behavioral problems should be assessed in order to specify and measure behavioral excesses, deficits, and inadequacies, to identify the relevant conditions (biological, social, or personal) that are hypothetically functionally related—causing, controlling, or maintaining—to those deficits or inadequacies, to plan the best treatment for manipulating those relevant conditions (or independent variables), and, during and/or after treatment, to evaluate behavioral change. In sum, one of the hallmarks of behavioral assessment is an interactive relationship between assessment, treatment, and evaluation.

3.1 Assessment

Behavioral assessment begins with the identification of the subject's problematic situation and the identification of its controlling, maintaining, or explaining conditions (Fernández-Ballesteros and Staats 1992).

Thus, the first task is the specification of a person's problematic behaviors in term of deficits, excesses, or inappropriateness parameters (frequency, duration, etc.), and of the circumstances in which the behavior began and has occurred since. This initial task requires the description, operationalization, and measurement of the problematic behavioral functioning, and sometimes its classification. Behavioral assessors should also consider the person's assets as a supporting element of future treatment. In sum, psychologists should make decisions about treatment targets and the subject's positive conditions.

Functional analysis also demands identification of the relevant conditions that, hypothetically, explain, control, or maintain the behavioral problem. These conditions should be assessed, at this stage of the process, in terms of their observed association and/or interaction with the subject's problematic target

behaviors, but are selected because of their well-established relationships with the problematic situation. Thus, assessment begins with a case formulation or the conceptualization of the problem (Haynes and Follette 1993).

3.2 Treatment

Although psychological interventions are usually guided, to some extent, by a preassessment of the client, the behavioral model was the first to claim that selection of treatment should be supported by previous assessment. Thus, the selection of the most appropriate treatment should be based on the relevant conditions observed as being associated with the client's target behaviors.

There are two important decisions related to assessment. As Barrios (1988) indicated, selecting the most appropriate treatment for a problem requires information on both the variables that maintain, control, or explain the problem behavior and the variables that produce the most effective treatment.

3.3 Evaluation

One of the hallmarks of behavioral assessment and therapy is evaluation of the treatment, (both during it and afterwards).

For evaluation, which should be planned at the end of the assessment phase, we need to know which measures of the target behaviors will be selected as outcome or dependent variables, and which design will provide the basis for evaluating the treatment (see *Single-case Experimental Designs in Clinical Settings* and *Clinical Treatment Outcome Research: Control and Comparison Groups*). Depending on the selected design, target behaviors will be assessed during or after treatment.

Several problematic issues arise from this final step of the behavioral assessment process. Various authors have stressed the difficulty of measuring change, the importance of change generalization (through stimuli, situations, persons, and time), the relevance of cost-benefit analysis, and the contribution of assessment to treatment effectiveness (Silva 1993).

4. The Future of Behavioral Assessment

Several challenges can be identified for the forthcoming decades. First, even if at the outset behavioral assessment was claimed to be applicable to all fields, it has remained linked to the clinical field. Although parallel frameworks have developed in counseling or personnel selection (e.g., 'assessment centers'), much more effort should be made to extend behavioral assessment.

Moreover, it is extremely important to proceed to an evaluation of behavioral assessment's reliability in terms of interassessor accuracy, and to make cost-benefit analyses with respect to treatment outcomes (Silva 1993).

Behavioral assessment needs measurement instruments and, although thousands of questionnaires and observational scales have been developed, very few have been adequately evaluated in scientific studies. Several advances in psychometrics and statistics can be applied to behavioral assessment technology, and in the near future we can expect improved measurement devices (Barrios 1988).

Classification has been a contradictory operation for the behavioral assessor and therapist. Staats' (1963) and Bandura's (1969) early behavioral classification systems were not commonly accepted. Psychiatric classification systems have been influenced by the behavioral approach in the sense of arriving at a behavioral description of symptoms and syndromes, but these systems have not been transformed into behavioral functional categories. Progress should be made in this direction (Haynes and Follette 1993).

Assessment involves a process of decision-making that is well known in terms of the operations undertaken. Nevertheless, this process is not prescriptive. We might expect that in future standards or guidelines for the assessment process will be developed and supported by scientific associations; indeed, several steps in this direction have already been taken.

Currently, clinical psychology considers not only dysfunctional behavior but also high behavioral functioning. As pointed out by Haynes (1998), the behavioral model should also be committed to promoting successful living throughout life, which, as always, will require sound assessment.

See also: Behavior Analysis, Applied; Behavior Therapy: Psychological Perspectives; Clinical Assessment: Interview Methods; Clinical Treatment Outcome Research: Control and Comparison Groups; Single-case Experimental Designs in Clinical Settings; Skinner, Burrhus Frederick (1904–90)

Bibliography

Bandura A 1969 *Principles of Behavior Modification*. Holt, Rinehart and Winston, New York

Barrios B A 1988 On the changing nature of behavioral assessment. In: Bellack A S, Hersen M (eds.) *Behavioral Assessment. A Practical Handbook*, 3rd edn. Pergamon, New York

Bellack A S, Hersen M 1988 Future direction of behavioral assessment. In: Bellack A S, Hersen M (eds.) *Behavioral Assessment. A Practical Handbook*, 3rd edn. Pergamon, New York

Cone J D 1993 The current state of behavioral assessment. *European Journal of Psychological Assessment* **9**: 175–81

Fernández-Ballesteros R 1993 Behavioral assessment: dying, vanishing, or still running. *European Journal of Psychological Assessment* **9**: 159–74
Fernández-Ballesteros R, Staats A W 1992 Paradigmatic behavioral assessment, treatment and evaluation: answering the crisis in behavioral assessment. *Advances in Behaviour Research and Therapy* **14**: 1–27
Hayes S C, Nelson R M 1986 Assessing the effects of therapeutic interventions. In: Nelson R O, Hayes S C (eds.) *Conceptual Foundations of Behavioral Assessment*. Guilford Press, New York
Haynes S N 1998 The changing nature of behavioral assessment. In: Bellack A S, Hersen M (eds.) *Behavioral Assessment. A Practical Handbook*, 4th edn. Allyn and Bacon, New York
Haynes S N, Follette W C 1993 The challenge faced by behavioral assessment. *European Journal of Psychological Assessment* **9**: 182–8
Kanfer F H, Saslow G 1965 Behavioral analysis. *Archives of General Psychiatry* **12**: 529–38
Nelson R O 1983 Behavioral assessment: past, present and future. *Behavioral Assessment* **5**: 195–206
Silva F 1993 Treatment utility: a reappraisal. *European Journal of Psychological Assessment* **9**: 222–6
Staats A W 1963 *Complex Human Behavior*. Holt, Rinehart and Winston, New York
Skinner B F 1953 *Science and Human Behavior*. Macmillan, New York

R. Fernández-Ballesteros

Behavioral Economics

It says something interesting about the field of economics that there is a subfield called behavioral economics. Surely all of economics is meant to be about the behavior of economic agents, be they firms or consumers, suppliers or demanders, bankers or farmers. So, what is behavioral economics, and how does it differ from the rest of economics?

Economics traditionally conceptualizes a world populated by calculating, unemotional maximizers that have been dubbed *Homo economicus*. In a sense, neoclassical economics has defined itself as explicitly 'antibehavioral.' Indeed, virtually all the behavior studied by cognitive and social psychologists is either ignored or ruled out in a standard economic framework. This unbehavioral economic agent has been defended on numerous grounds: some claimed that the model was 'right'; most others simply argued that the standard model was easier to formalize and practically more relevant.

Behavioral economics blossomed with the realization that neither point of view was correct. Empirical and experimental evidence mounted against the stark predictions of unbounded rationality. Further work made clear that one could formalize psychological ideas and translate them into testable predictions. The behavioral economics research program has consisted of two components: (a) identifying the ways in which behavior differs from the standard model and (b) showing how this behavior matters in economic contexts.

This article gives a flavor of the program. We begin by discussing the most important ways in which the standard economic model needs to be enriched. We then illustrate how behavioral economics has been fruitfully applied to two important fields: finance and savings. However, first, we discuss why the market forces and learning do not eliminate the importance of human actions.

1. *Is Homo Economicus the Only One Who Survives?*

Many economists have argued that a combination of market forces (competition and arbitrage) plus evolution should produce a world similar to that described in an economics textbook: do only the rational agents survive? Or, do the workings of markets at least render the actions of the quasi-rational irrelevant? These are questions that have been much studied since the 1980s, and the early impressions of many economists that markets would wipe out irrationality were optimistic.

Consider a specific example: human capital formation. Suppose that a young economist, named Sam, decides to become a behavioral economist, perhaps because he mistakenly thinks this will lead to riches, or because he thinks it is going to be the next fad, or because he finds it interesting and lacks the willpower to study 'real' economics. Whatever the reason for the choice, let us assume for the sake of argument that this decision was a mistake for Sam by any rational calculation. So, what will market forces do? First, Sam may be poorer because of this choice than if he had sensibly chosen to study corporate finance, but he will not be destitute. Sam might even realize that he could switch to corporate finance and make much more money but is simply unable to resist the temptation to continue wasting his time on behavioral economics. So, markets *per se* do not necessarily solve the problem: they provide an *incentive* to switch, but they cannot force Sam's hand.

What about arbitrage? In this case, like most things we study in economics outside the realm of financial markets, there is simply no arbitrage opportunity available. Suppose a wise arbitrageur is watching Sam's choices, what bet can he or she place? None. The same can be said if Sam saves too little for retirement, picks the wrong wife, or buys the wrong car. None of these irrational acts generates an arbitrage opportunity for anyone else.

Indeed, economists now realize that even in financial markets there are important limits to the workings of arbitrage. First, in the face of irrational traders, the

arbitrageur may privately benefit more from trading that helps push prices in the wrong direction than from trading that pushes prices in the right direction. Put another way, it may often pay 'smart money' to follow 'dumb money' rather than to lean against it (Haltiwanger and Waldman 1985, Russell and Thaler 1985). For example, an extremely smart arbitrageur near the beginning of the tulip mania would have profited more from buying tulips and further destabilizing prices than by shorting them. Second, and slightly related, arbitrage is an inherently risky activity and consequently the supply of arbitrage will be inherently limited (De Long et al. 1990). Arbitrageurs who did decide to short tulips early would probably have been wiped out by the time their bets were proven to be 'right.' Add to this the fact that in practice most arbitrageurs are managing other people's money and therefore judged periodically, one can see the short horizons that an arbitrageur will be forced to take on. This point was made forcefully by Shleifer and Vishny (1997), who essentially foresaw the scenario that ended up closing Long-term Capital Management.

So, markets *per se* cannot be relied upon to make economic agents rational. What about evolution? An old argument was that individuals who failed to maximize should have been weeded out by evolutionary forces, which presumably operated during ancient times. Overconfident hunters, for example, presumably caught less prey, ate less, and died younger. Such reasoning, however, has turned out to be faulty. Evolutionary arguments can just as readily explain over-confidence as they can explain appropriate levels of confidence. For example, consider individuals playing a war of attrition (perhaps in deciding when to back down during combat). Here overconfidence will actually help. Seeing the overconfidence, a rational opponent will actually choose to back down sooner. As can be seen from this example, depending on the initial environment (especially when these environments have a game theoretic component to them), evolution may just as readily weed out rational behavior as it does weed out quasi-rational behavior. The troubling flexibility of evolutionary models means that they can just as readily argue for bounds on rationality.

The final argument is that individuals who systematically and consistently make the same mistake will eventually learn the error of their ways. This kind of argument has also not stood up well under theoretical scrutiny. First, the optimal experimentation literature has shown that there can be a complete lack of learning even in infinite horizons. The intuition here is simple: as long as there are some opportunity costs to learning or to experimenting with a new strategy, even a completely 'rational' learner will choose not to experiment. This player will get stuck in a non-optimal equilibrium, simply because the cost of trying something else is too high. Second, work on learning in games has formally demonstrated Keynes' morbid observation on the 'long run.' The time required to converge to an equilibrium strategy can be extremely long. Add to this a changing environment and one can easily be in a situation of perpetual nonconvergence. In practice, for many of the important decisions we make, both arguments apply with full force. The number of times we get to learn from our retirement decisions is low (and possibly zero). The opportunity cost of experimenting with different ways of choosing a career can be very high.

The upshot of all these theoretical innovations has been clear. One cannot defend unbounded rationality on purely theoretical grounds. Neither arbitrage, competition, evolution, nor learning necessarily guarantees that unbounded rationality must be an effective model. In the end, as some might have expected, it must ultimately be an empirical issue. Does 'behavior' matter? Before evaluating this question in two different fields of application, we explore the ways in which real behavior differs from the stylized neoclassical model.

2. *Three Bounds of Human Nature*

The standard economic model of human behavior includes (at least) three unrealistic traits: unbounded rationality, unbounded willpower, and unbounded selfishness. These three traits are good candidates for modification.

Herbert Simon (1955) was an early critic of modeling economic agents as having unlimited information processing capabilities. He suggested the term 'bounded rationality' to describe a more realistic conception of human problem solving capabilities. As stressed by Conlisk (1996), the failure to incorporate bounded rationality into economic models is just bad economics—the equivalent to presuming the existence of a free lunch. Since we have only so much brainpower, and only so much time, we cannot be expected to solve difficult problems optimally. It is eminently 'rational' for people to adopt rules of thumb as a way to economize on cognitive faculties. Yet the standard model ignores these bounds and hence the heuristics commonly used. As shown by Kahneman and Tversky (1974), this oversight can be important since sensible heuristics can lead to systematic errors.

Departures from rationality emerge both in judgments (beliefs) and in choice. The ways in which judgment diverges from rationality is long and extensive (Kahneman et al. 1982). Some illustrative examples include overconfidence, optimism, anchoring, extrapolation, and making judgments of frequency or likelihood based on salience (the availability heuristic) or similarity (the representativeness heuristic).

Many of the departures from rational choice are captured by prospect theory (Kahneman and Tversky

1979), a purely descriptive theory of how people make choices under uncertainty (see Starmer (2000) for a review of literature on non-EU theories of choice). Prospect theory is an excellent example of a behavioral economic theory in that its key theoretical components incorporate important features of psychology. Consider three features of the prospect theory value function. (a) It is defined over changes to wealth rather than levels of wealth (as in EU) to incorporate the concept of adaptation. (b) The loss function is steeper than the gain function to incorporate the notion of 'loss aversion'; the notion that people are more sensitive to decreases in their well being than to increases. (c) Both the gain and loss function display diminishing sensitivity (the gain function is concave, the loss function convex) to reflect experimental findings. To describe fully choices prospect theory often needs to be combined with an understanding of 'mental accounting' (Thaler 1985). One needs to understand when individuals faced with separate gambles treat them as separate gains and losses and when they treat them as one, pooling them to produce one gain or loss.

A few examples can illustrate how these concepts are used in real economics contexts. Consider overconfidence. If investors are overconfident in their abilities, they will be willing to make trades even in the absence of true information. This insight helps explain a major anomaly of financial markets. In an efficient market when rationality is common knowledge, there is virtually no trading, but in actual markets there are hundreds of millions of shares traded daily and most professionally managed portfolios are turned over once a year or more. Individual investors also trade a lot: they incur transaction costs and yet the stocks they buy subsequently do worse than the stocks they sell.

An example involving loss aversion and mental accounting is Camerer et al.'s (1997) study of New York City cab drivers. These cab drivers pay a fixed fee to rent their cabs for 12 h and then keep all their revenues. They must decide how long to drive each day. A maximizing strategy is to work longer hours on good days (days with high earnings per hour such as rainy days or days with a big convention in town) and to quit early on bad days. However, suppose cabbies set a target earnings level for each day, and treat shortfalls relative to that target as a loss. Then, they will end up quitting early on good days and working longer on bad days, precisely the opposite of the rational strategy. This is exactly what Camerer et al. found in their empirical work.

Having solved for the optimum, *Homo economicus* is next assumed to choose the optimum. Real humans, even when they know what is best, sometimes fail to choose it for self-control reasons. Most of us at some point have eaten, drunk, or spent too much, and exercised, saved, or worked too little: such is human nature. People (even economists) also procrastinate. We are completing this article after the date on which it was due, and we are confident that we are not the only guilty parties. Although people have these self-control problems, the are at least somewhat aware of them: they join diet plans and buy cigarettes by the pack (because having an entire carton around is too tempting). They also pay more withholding taxes than they need to (in 1997, nearly 90 million tax returns paid an average refund of around $1,300) in order to assure themselves a refund, but then file their taxes near midnight on April 15 (at the post office that is being kept open late to accommodate their fellow procrastinators).

Finally, people are boundedly selfish. Although economic theory does not rule out altruism, as a practical matter economists stress self-interest as the primary motive. For example, the free rider problems widely discussed in economics are predicted to occur because individuals cannot be expected to contribute to the public good unless their private welfare is thus improved. In contrast, people often take selfless actions. In 1993, 73.4 percent of all households gave some money to charity, the average dollar amount being 2.1 percent of household income. Also, 47.7 percent of the population does volunteer work with 4.2 h per week being the average time volunteered. Similar selfless behavior is observed in controlled laboratory experiments. Subjects systematically often cooperate in public goods and prisoners dilemma games, and turn down unfair offers in 'ultimatum' games.

3. *Finance*

If economists had been polled around 1980 and asked to name the domain in which bounded rationality was least likely to find useful applications the likely winner would have been finance. The limits of arbitrage arguments were not well understood at that time, and one leading economist had called the efficient markets hypothesis the best established fact in economics. Times change. As we begin the twenty-first century finance is perhaps the branch of economics where behavioral economics has made the greatest contributions. How has this happened?

Two factors contributed to the surprising success of behavioral finance. First, financial economics in general, and the efficient market hypothesis in particular, generated sharp, testable predictions about observable phenomena. Second, there are many data readily available to test these sharp predictions. We briefly summarize here a few examples.

The rational efficient markets hypothesis makes two classes of predictions about stock price behavior. The first is that stock prices are 'correct' in the sense that asset prices reflect the true or rational value of the security. In many cases this tenet of the efficient market hypothesis is untestable because intrinsic values are not observable. However, in some special

cases the hypothesis can be tested by comparing two assets whose relative intrinsic values are known. One class of these is called 'Siamese Twins': two versions of the same stock that trade in different places.

A specific well-known example is the case of Royal Dutch Shell as documented in Froot and Dabora (1999). The facts are that Royal Dutch Petroleum and Shell Transport are independently incorporated in The Netherlands and the UK, respectively. The current firm emerged from a 1907 alliance between Royal Dutch and Shell Transport in which the two companies agreed to merge their interests on a 60:40 basis. Royal Dutch trades primarily in the USA and The Netherlands and Shell trades primarily in London. According to any rational model, the shares of these two components (after adjusting for foreign exchange) should trade in a ratio of 60:40. They do not; the actual price ratio has deviated from the expected one by more than 35 percent. Simple explanations such as taxes and transactions costs cannot explain the disparity (see Froot and Dabora 1999). This example illustrates that prices can diverge from intrinsic value because of limits of arbitrage. Some investors do try to exploit this mispricing, buying the cheaper stock and shorting the more expensive one, but this is not a sure thing, as many hedge funds learned in the summer of 1998 (when, at the time hedge funds were trying to get more liquidity, the pricing disparity widened).

The Royal Dutch Shell anomaly is a violation of one of the most basic principles of economics: the law of one price. Another similar example is the case of closed-end mutual funds (Lee et al. 1991). These funds are much like typical (open-end) mutual funds except that to cash out of the fund, investors must sell their shares on the open market. This means that closed-end funds have market prices that are determined by supply and demand, rather than set equal to the value of their assets by the fund managers as in an open-end fund. Since the holdings of closed-end funds are public, market efficiency would lead one to expect that the price of the fund should match the price of the underlying securities they hold (the net asset value or NAV). Instead, closed-end funds typically trade at substantial discounts relative to their NAV, and occasionally at substantial premiums. Most interesting from a behavioral perspective is that closed-end fund discounts are correlated with one another and appear to reflect individual investor sentiment. (Closed-end funds are owned primarily by individual investors rather than institutions.) Lee et al. found that discounts shrink in months when shares of small companies (also owned primarily by individuals) do well, and in months when there is lots of IPO activity, indicating a 'hot' market. Since these findings were predicted by their theory, they move the research beyond the demonstration of an embarrassing fact (price not equal to NAV) toward a constructive understanding of how markets work.

The second principle of the efficient market hypothesis is 'unpredictability.' In an efficient market it is not possible to predict future stock price movements based on publicly available information. Many early violations of this had no explicit link to behavior. Thus it was reported that small firms, firms with low price–earnings ratios earned higher returns than other stocks with the same risk. Also, stocks in general, but especially stocks of small companies, have done well in January and on Fridays (but poorly on Mondays).

An early study by De Bondt et al. (1985) was explicitly motivated by the psychological finding that individuals tend to over-react to new information. For example, experimental evidence suggests that people tended to underweight base-rate data (or prior information) in incorporating new data. De Bondt et al. hypothesized that if investors displayed this behavior, then stocks that had performed fairly well over a period of years will eventually have prices that are too high. Individuals over-reacting to the good news will drive the prices of these stocks too high. Similarly, poor performers will eventually have prices that are too low. This yields a prediction about future returns: past 'winners' ought to underperform while past 'losers' ought to outperform the market. Using data for stocks traded on the New York Stock Exchange, De Bondt et al. (1985) found that the 35 stocks that had performed the worst over the past 5 years (the losers) outperformed the market over the next 5 years, while the 35 biggest winners over the past 5 years subsequently underperformed. Follow-up studies have shown that these early results cannot be attributed to risk (by some measures the portfolio of losers is actually less risky than the portfolio of winners), and can be extended to other measures of over-reaction such as the ratio of market price to the book value of equity.

More recent studies have found other violations of unpredictability that have the opposite pattern from that found by DeBondt et al., namely under-reaction rather than over-reaction. Over short periods of time, e.g., 6 months to 1 year, stocks display momentum—the stocks that go up the fastest for the first 6 months of the year tend to keep going up. Also, after many corporate announcements such as large earnings changes, dividend initiations and omissions, share repurchases, splits, and seasoned equity offerings, there is an initial price jump on the day of the announcement followed by a slow drift in the same direction for as long as 1 year or more (Shleifer 2000).

These findings of under-reaction are a further challenge to the efficient markets hypothesis, but also to behavioral finance. Do markets sometimes over-react, and sometimes under-react? If so, can *any* pattern be explained, at least *ex post*? Is there a unifying framework that can bring together these apparently opposing facts and ideas?

Work is only now beginning to attack this extremely important question. Several explanations have been

offered recently (Shleifer (2000) summarizes them). All rely on psychological evidence (in one way or another) in unifying the facts. They all explain the anomalies by noting that under-reaction appears at short horizons whereas over-reaction appears at longer horizons, but each paper provides its own distinctive explanation. Which (if any) of these is the best one has yet to be decided. But the facts discovered so far, combined with these models, demonstrate the changing nature of finance. Rigorous empirical work building on psychological phenomena has given us new tools to unearth interesting empirical facts. Rigorous theoretical work, on the other hand, has been trying to address the challenge of incorporating these empirical facts into a psychologically plausible model.

The research discussed so far has been about asset prices and the controversy about the efficient market hypothesis. There is another stream of research that is just about investor behavior, not about prices. One example of this stream is motivated by mental accounting and loss aversion. The issue is whether investors are reluctant to realize capital losses (because they would have to 'declare' the loss to themselves). Shefrin and Statman (1985) dubbed this hypothesis the 'disposition effect.' The prediction that investors will be less willing to sell a loser than a winner is striking since the tax law encourages just the opposite behavior. Nevertheless, Odean (1998) finds evidence of just this behavior. In his sample of the customers of a discount brokerage firm, investors were more likely to sell a stock that had increased in value than one that had decreased in value. While around 15 percent of all gains were realized, only 10 percent of all losses are realized. This hesitancy to realize gains came at a cost. Odean shows that the stocks the loser stocks held under-performed the gainer stocks that were sold.

4. *Savings*

If finance was the field in which a behavioral approach was least likely, *a priori*, to succeed, saving had to be one of the most promising. The standard life-cycle model of savings abstracts from both bounded rationality and bounded willpower, yet saving for retirement is both a difficult cognitive problem and a difficult self-control problem. It is then, perhaps less surprising that a behavioral approach has been fruitful here. As in finance, progress has been helped by the combination of a refined standard theory with testable predictions and many data sources on household saving behavior. One crisp prediction of the life-cycle model is that savings rates are independent of income. Suppose twins Tom and Ray are identical in every respect except that Tom earns most of his money early in his life (e.g., he is a basketball player), whereas Ray earns most of his late in life (e.g., he is a manager). The life-cycle model predicts that Tom the basketball player ought to save his early income to increase consumption later in life, whereas Ray the manager ought to borrow from his future income to increase consumption earlier in life. This prediction is completely unsupported by the data, which show that consumption very closely tracks income over individuals' life cycles. Furthermore, the departures from predicted behavior cannot be explained merely by people's inability to borrow. Banks et al. (1998) show, for example, that consumption drops sharply as individuals retire and their incomes drop. They have simply not saved enough for retirement. Indeed, many low- to middle-income families have essentially no savings whatsoever. The primary cause of this lack of saving appears to be self-control. One piece of evidence supporting this conclusion is that virtually all saving done by Americans is accomplished in vehicles that support what are often called 'forced savings,' e.g., accumulating home equity by paying the mortgage and participation in pension plans. Coming full circle, one 'forced' savings that individuals may choose themselves may be high tax withholdings, so that when the refund comes they can buy something they might not have had the willpower to save up for.

One of the most interesting research areas has been devoted to measuring the effectiveness of tax-subsidized savings programs such as IRAs and 401(k)s. The standard analysis of these programs is fairly simple. Consider the original IRA program from the early 1980s. This program provided tax subsidies for savings up to a threshold, often $2,000 per year; there was no marginal incentive to save for any household that was already saving more than $2,000 per year. Therefore, those saving more than the threshold should not increase their total saving, they will merely switch some money from a taxable account to the IRA for the tax gain. Moreover, everyone who is saving a positive amount should participate in this program for the infra-marginal gain. The actual analysis of these programs has shown that the reality is not so clear. By some accounts at least, these programs appear to have generated considerable new savings. Some argue that almost every dollar of savings in IRAs appears to represent new savings. In other words, people are not simply shifting their savings into IRAs and leaving their total behavior unchanged. Similar results are found for 401(k) plans. The behavioral explanation for these findings is that IRAs and 401(k) plans help solve self-control problems by setting up special mental accounts that are devoted to retirement savings. Households tend to respect the designated use of these accounts, and their self-control is helped by the tax penalty that must be paid if funds are removed prematurely. (Some issues remain controversial. See the debate in the Fall 1996 issue of the *Journal of Economic Perspectives*.)

An interesting flip side to IRA and 401(k) programs is that they also generated far less than the full participation one would have expected. Many eligible

people do not participate, foregoing in effect a cash transfer from the government (and, in some cases, from their employer). O'Donoghue and Rabin (1999) present an explanation based on procrastination and hyperbolic discounting. Individuals typically show very sharp impatience for short-horizon decisions, but much more patience at long horizons. This is often referred to as hyperbolic discounting by contrast with the standard assumption of exponential discounting, in which patience is independent of horizon. In exponential models, people are equally patient at long and short horizons. O'Donoghue and Rabin argue that hyperbolic individuals will show exactly the low IRA participation that we observe. Although hyperbolic people will want eventually to participate in IRAs (because they are patient in the long run), something always comes up in the short run (where they are very impatient) that provides greater immediate reward. Consequently, they may delay joining the IRA indefinitely.

If people procrastinate about joining the savings plan, then it should be possible to increase participation rates simply by lowering the psychic costs of joining. One simple way of accomplishing this is to switch the default option for new workers. In most companies, when employees first become eligible for the 401(k) plan, they receive a form inviting them to join; to join they have to send the form back and make some choices. Several firms have made the seemingly inconsequential change of switching the default: employees are enrolled into the plan unless they explicitly opt out. This change has often produced dramatic increases in savings rates. For example, in one company studied by Madrian and Shea (2000), the employees who joined after switching the default to being in the plan were 50 percent more likely to participate than the workers in the year prior to the change. (They also found that the default asset allocation had a strong effect on workers' choices. The firm had made the default asset allocation 100 percent in a money market account, and the proportion of workers selecting this allocation soared.)

Along with these empirical facts, there has also been substantial theoretical work. To cite a few examples, Laibson (1997) and O'Donoghue and Rabin (1999) both examined the effects of hyperbolic discounting on the savings decisions. They highlighted that hyperbolic individuals will demand commitment devices (savings vehicles which are illiquid) and generally fail to obey the Euler equation.

5. *Other Directions*

We have concentrated on two fields here in order to give a sense of what behavioral economics can do, but we do not want to leave the impression that savings and financial markets are the only domains in which a behavioral approach has been or could be effective. In labor economics, experimental and empirical work has underlined the importance of fairness considerations in setting wages. For example, wages between industries differ dramatically, even for identical workers and, most interestingly, even homogeneous workers (such as janitors) earn higher wages when they work in industries where other occupations earn more. In law and economics, we have seen the importance of 'irrelevant' factors in a jury's decision to sentence or in the magnitude of awards they give. In corporate finance, one can fruitfully interpret a manager's decision to acquire or diversify as resulting from overconfidence. One could go on. There is much to be done.

See also: Bounded Rationality; Economic Sociology; Rational Choice Theory: Cultural Concerns; Savings Behavior: Demographic Influences; Stock Market Predictability; Tversky, Amos (1937–96)

Bibliography

Banks J, Blundell R, Tanner S 1998 Is there a retirement-savings puzzle? *American Economic Review* **88**: 769–88

Camerer C, et al. 1997 Labor supply of New York City cabdrivers: One day at a time. *Quarterly Journal of Economics* **112**: 407–41

Conlisk J 1996 Why bounded rationality? *Journal of Economic Literature* **34**: 669–700

De Bondt, Werner F M, Thaler R 1985 Does the stock market overreact? *Journal of Finance* **40**: 793–805

DeLong B, Shleifer A, Summers L, Waldman R 1990 Noise trader risk in financial markets. *Journal of Political Economy* **98**: 703–38

Froot K A, Dabora E M. 1999 How are stock prices affected by the location of trade? *Journal of Financial Economics* **53**: 189–216

Haltiwanger J, Waldman M 1985 Rational expectations and the limits of rationality: An analysis of heterogeneity. *American Economic Review* **75**: 326–40

Kahneman D, Slovic P, Tversky A 1982 *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK

Kahneman D, Tversky A 1974 Judgement under uncertainty: Heuristics and biases. *Science* **185**: 1124–31

Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263–91

Laibson D 1997 Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* **112**: 443–77

Lee C M C, Shleifer A, Thaler R H 1991 Investor sentiment and the closed-end fund puzzle. *Journal of Finance* **46**: 75–109

Madrian B, Shea D 2000 *The Power of Suggestion: Inertia in 401(k) Savings Behavior*. Mimeo, University of Chicago, Chicago

Odean T 1998 Are investors reluctant to realize their losses? *Journal of Finance* **53**: 1775–98

O'Donoghue E, Rabin M 1999 Procrastination in preparing for retirement. In: Aaron H (ed.) *Behavioral Dimensions of Retirement Economics*. Brookings Institution

Russell T, Thaler R 1985 The relevance of quasi rationality in competitive markets. *American Economic Review* **75**: 1071–82

Shefrin H, Statman M 1985 The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance* **40**: 777–90

Shleifer A 2000 *Inefficient Markets: An Introduction to Behavioral Finance*. Clarendon Lectures. Oxford University Press, Oxford, UK
Shleifer A, Vishny R 1997 The limits of arbitrage. *Journal of Finance* **52**: 35–55
Simon H A 1955 A behavioral model of rational choice. *Quarterly Journal of Economics* **69**: 99–118
Starmer C 2000 Developments in non-expected utility: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* in press
Thaler R 1985 Mental accounting and consumer choice. *Marketing Science* **4**: 199–214

S. Mullainathan and R. H. Thaler

Behavioral Genetics: Psychological Perspectives

Behavioral genetics, the study of genetic influences on behavior, includes the full gamut of genetic levels of analyses. Increasingly, it includes molecular genetic studies of DNA as it relates to behavior. Most behavioral genetic research focuses on the genetic and environmental origins of differences between individuals—that is, why people differ in psychopathology, personality, and cognitive abilities and disabilities. However, 99.9 percent of the 3.5 billion DNA sequences in the human genome are identical for all individuals. Although this leaves about three million DNA sequences that vary among individuals, many genes, especially those genes which are the most important evolutionarily, do not vary. It is not yet possible to study the effects of nonvarying genes in the human species but in mice differences in such genes can be created that knock out the genes' functioning, which makes it possible to study the effects of such gene knock-outs (Plomin and Crabbe 2000).

Evolutionary psychology also considers universal (nonvarying) genes but in an evolutionary timescale. It seeks to understand the adaptive value of universal aspects of human behavior such as our species' natural use of language, our similar facial expressions of emotion, and cross-cultural similarities in mating strategies (Goldsmith et al. in press). It is especially difficult to pin down the effects of nonvarying genes in an evolutionary timeframe. The main line of evidence for evolutionary psychology is to compare average differences between species. This is a very different level of analysis from most behavioral genetics research, which focuses on differences among individuals within a species rather than aspects of behavior that do not vary within a species.

Although molecular genetic and evolutionary analyses that consider universal aspects of our species represent important directions for behavioral genetic research, the vast majority of behavioral genetic research focuses on variations on these universal themes. Since the 1920s, so-called quantitative genetic methods such as twin and adoption designs have been used to estimate the net effect of genetic and environmental factors on individual differences in behavior. As explained in the following section, quantitative genetic research converges on the conclusion that genetic factors contribute importantly to behavioral differences among individuals across a broad range of behaviors. However, the same research provides the best available evidence for the importance of environmental influences, which can only be clearly demonstrated when genetic factors are taken into account.

The case for genetic influence on individual differences in behavior is so strong that molecular genetic studies are now beginning to identify specific genes responsible for the genetic effects seen in quantitative genetic research. The last section considers molecular genetic research on behavior. Although quantitative genetic and molecular genetic strategies are far more powerful using nonhuman animal models (e.g., Plomin and Crabbe 2000), this article is limited to research on human behavior.

1. *Quantitative Genetics*

The twin and adoption methods were developed in the 1920s as tools to screen for genetic influence on complex traits like behavior. These methods emerged from quantitative genetic theory that worked out the expected genetic relationships among different kinds of relatives for complex quantitative traits that are influenced by many genes and many environmental factors (Plomin et al. 2001). Quantitative genetics always focused on naturally occurring genetic variation, in contrast to early work in molecular genetics that centered on single-gene mutations, especially new mutations created using animal models.

Some caveats and cautions are discussed later. However, because the results of quantitative genetic analyses are often misinterpreted, it should be emphasized at the outset that even if a given attribute is highly heritable in the sense of accounting for a large share of differences between persons, the mechanisms by which that attribute develops could be completely dependent on environmental conditions. Take reading as an example. There are individual differences in reading ability that are associated with genetic differences. However, learning to read clearly requires exposure to reading conditions. Genetic influence in reading ability means that some persons somehow more readily pick up reading skills. Moreover, genetic influence only refers to average results across a population. A particular child's poor reading ability might be entirely due to environmental circumstances

that make it especially difficult to learn to read or other children's excellent reading skills might be entirely due to extraordinary efforts on the part of their parents or teachers.

1.1 The Twin Method

The accident of nature that results in identical (monozygotic, MZ) twins or fraternal (dizygotic, DZ) twins provides one way to address the question of genetic and environmental origins of individual differences in behavior. MZ twins are like clones, genetically identical with each other because they came from the same fertilized egg. DZ twins, on the other hand, develop from two eggs that happen to be fertilized at the same time. Like other siblings, DZ twins are only half as similar genetically as MZ twins. To the extent that behavioral variability is caused by environmental factors, DZ twins should be as similar for the behavioral trait as are MZ twins because both types of twins are reared by the same parents in the same place and at the same time. If the trait is influenced by genes, then DZ twins ought to be less similar than MZ twins.

Possible biases in the twin method have been explored. MZ twins often share the same chorion (the outermost membrane surrounding the fetus) during prenatal development, which might make them more similar than DZ twins, who never share the same chorion. The scanty evidence relevant to this issue is mixed (Sokol et al. 1995). Another possibility is that twins may not be representative of the nontwin population because of adverse intrauterine environments caused by twinning (Phillips 1993). However, the statistical distributions for most behavioral dimensions and disorders for twins and nontwins are generally similar, suggesting that twins are reasonably representative of nontwin populations (e.g., Christensen et al. 1995). A subtle but important issue is that MZ twins might have more similar experiences than DZ twins after birth. The twin method is based on the assumption that the environments of DZ twins reared in the same family are approximately as similar as the environments of MZ twins reared in the same family. This assumption has been tested in several ways and appears reasonable for most traits (Bouchard and Propping 1993). Although the possibility remains that MZ twins may be treated more alike by their parents because they are more similar in appearance and behavior, the twin method provides a rough but useful screen to unpack the 'bottom-line' effects of genes and environment (Martin et al. 1997).

1.2 The Adoption Method

The adoption method is another quasi-experimental design that has a different set of assumptions and potential problems. Because the twin and adoption methods are so different, greater confidence is warranted when results from these two methods converge on the same conclusion—as they usually do. One issue for adoption studies is that adoptees and their adoptive families might not be representative of the population as a whole, either because they have distinctive characteristics or because they span a narrower range. There is also the possibility that adopted children might be selectively placed with adoptive parents matched to the birth parents. These issues can be examined empirically. For example, in a longitudinal prospective adoption study of behavioral development that began in 1975, adoptive families were found who are reasonably representative of the population and in this study selective placement was negligible (Plomin et al. 1997).

2. Genetic Influence on Behavior

Consider schizophrenia. Until the 1960s, schizophrenia was thought to be environmental in origin, with theories putting the blame on poor parenting to account for the fact that schizophrenia clearly runs in families. The idea that schizophrenia could run in families for genetic reasons was not seriously considered. Twin and adoption studies changed this view. Twin studies showed that MZ twins are much more similar than DZ twins, which suggests genetic influence. If one member of an MZ twin pair is schizophrenic, the chances are 45 percent that the other twin is also schizophrenic. For DZ twins, the chances are 17 percent. Adoption studies showed that the risk of schizophrenia is just as great when children are adopted away from their schizophrenic parents at birth as when children are reared by their schizophrenic parents, which provides dramatic evidence for genetic transmission. There are intense efforts to identify some of the specific genes responsible for genetic influence on schizophrenia.

In the 1960s, when schizophrenia was thought to be caused environmentally, it was important to emphasize the evidence for genetic influence such as the concordance of 45 percent for MZ twins. Now that evidence for the importance of genetic influence throughout the social and behavioral sciences has largely been accepted, it is important to make sure that the pendulum stays in the middle, in between nature and nurture. We need to emphasize that MZ twins are *only* 45 percent concordant for schizophrenia, which means that in more than half of the cases these pairs of genetically identical clones are discordant for schizophrenia. This discordance cannot be explained genetically—it must be due to environmental factors.

It should be noted that the word *environment* in genetic research really means *nongenetic*, which is a much broader definition of environment than is usually encountered in the social and behavioral sciences. That is, environment denotes all nonheritable

factors, including possible biological events such as prenatal and postnatal illnesses, not just the psychosocial factors that are usually considered in the social and behavioral sciences. The point is that genetics can often explain half of the variance of behavioral traits but this means that the other half of the variance is *not* due to genetic factors.

For milder behavioral disorders, too, a more balanced view that accepts the role of nature as well as nurture is beginning to prevail, and researchers are making progress in identifying specific genes. Moderate genetic influence has also been found for most behavioral disorders that have been studied, including mood disorders and anxiety (Plomin et al. 2001). Indeed, behavioral disorders tend to show greater genetic influence than common medical disorders such as breast cancer or heart disease (Plomin et al. 1994). The reason for the greater genetic influence for behavioral disorders may be that many different biological pathways can affect behavior. Behavior can be considered as the downstream outcome of many different biological systems; genetic influences can operate on behavior via all of the upstream systems.

Genetic influences are not limited to psychopathology; they also contribute to normal variation in personality (Loehlin 1992) and in cognitive abilities and disabilities. Behavioral genetic research to date has only scratched the surface of possible applications in the social and behavioral sciences, even within the best-studied domains of psychopathology, personality, and cognitive disabilities and abilities. For psychopathology, genetic research has just begun to consider disorders other than schizophrenia and the major mood disorders. Developmental psychopathology has recently become an active area of genetic research (Rutter et al. 1999). Personality is so complex that it can keep researchers busy for decades, especially as they go beyond self-report questionnaires to use other measures such as observations and ratings by others (Riemann et al. 1997). A rich territory for future exploration is the links between psychopathology and personality (Nigg and Goldsmith 1994). New directions for genetic research on cognitive abilities and disabilities include the systematic analysis of psychological theories of cognition (Mackintosh 1998) and the use of information-processing measures (e.g., Deary and Caryl 1997, Deary 2000) and brain-imaging measures.

The vast majority of human genetic research in the social and behavioral sciences has focused on psychopathology, personality, and cognitive disabilities and abilities because these areas have long been the focus of research on individual differences. Three new areas of psychology that are beginning to be explored genetically are psychology and aging, health psychology, and evolutionary psychology (Plomin et al. 2001). Some of the oldest areas of psychology—perception and learning, for example—have not emphasized individual differences and as a result have yet to be explored systematically from a genetic perspective. Entire disciplines within the social and behavioral sciences, such as economics, education, and sociology, are still essentially untouched by genetic research.

Genetic research, and especially heritability, is often misinterpreted. As mentioned earlier, a common mistake is to assume that heritability refers to an individual rather than to individual differences in a population. The heritability of height is 90 percent, but this does not mean that you grew to 90 percent of your height for reasons of heredity and that the few remaining inches were added by the environment. Another caution is that heritability like all other descriptive statistics is limited to a particular population as a particular time. Different cultures or cohorts could yield different results. One of the most common misinterpretations is that finding evidence for heritability implies that environmental factors are not important. This is wrong for several reasons. First, heritability of complex traits is seldom greater than 50 percent, which means that most of the variance is due to environmental rather than genetic factors. Indeed, behavioral genetic research provides the strongest available evidence for the importance of the environment. Furthermore, even if the heritability of a trait were 1.0 (which it never is for complex traits), a new environmental intervention could completely change the behavior. High heritability only means that environmental factors that currently vary for the sample do not make much of a difference. An important corollary of this point is that heritability does not constrain environmental interventions such as psychotherapy. Another mistake is to think that genetic influence means that there is a single gene that determines behavior. Single-gene disorders are deterministic—if you inherit the dominant allele for Huntington's disease you will die from the disease regardless of your other genes or your environment. However, complex traits like behavior are never caused by a single gene. Genetic effects are due to many genes that operate as probabilistic propensities rather than predetermined programming. Genes are not destiny. Finally, finding genetic influence on behavior has no necessary policy implications. Such data do not compromise the value of social equality, for example. Knowledge alone does not account for societal and political decisions because values are just as important in the decision-making process.

Genetic research in the social and behavioral sciences is moving beyond heritability. Asking whether and how much genetic factors affect behavioral dimensions and disorders is an important first step in understanding the origins of individual differences, but such questions are only a beginning. The next steps involve the question 'how'—that is, the study of the mechanisms by which genes have their effects. Avenues for genetic research in the social and behavioral sciences include developmental change and continuity,

links between the normal and abnormal, multivariate genetic analysis of heterogeneity and comorbidity, and the interplay between genes and environment (Plomin et al. 2001). An especially exciting direction for research is identification of some of the specific genes responsible for the heritability of behavioral disorders and dimensions.

3. DNA

Now that the contribution of genetic factors to individual differences in behavior is widely accepted, molecular genetic techniques are being applied that can identify some of the genes responsible for this genetic variation. Associations between DNA and behavioral differences have a unique status in the behavioral sciences because the DNA differences clearly cause the behavioral difference, that is, the DNA differences cannot be caused by the behavioral difference. For other factors associated with behavior, even biological factors such as neurotransmitters, it is always possible that behavior was the cause rather than the effect. For example, a certain neurotransmitter might be at higher levels in individuals with schizophrenia but that does not mean that the neurotransmitter causes schizophrenia. Having disordered thought processes characteristic of schizophrenia might have led to the higher level of the neurotransmitter.

The heritability of complex traits like behavior is likely to be due to multiple genes of varying, but small, effect size rather than one gene or a few genes with major effect. Genes in such multiple-gene systems are inherited in the same way as any other gene but they have been given a different name—quantitative trait loci (QTLs)—in order to highlight some important distinctions. Single-gene effects are necessary and sufficient for the development of a disorder such as phenylketonuria (PKU) that, left untreated, causes mental retardation, or Huntington's disease that causes general neural degeneration later in life. In contrast, QTLs contribute interchangeably and additively, analogous to probabilistic risk factors. If there are multiple genes that affect a trait, it is likely that the trait is distributed quantitatively as a dimension rather than qualitatively as a disorder. From a QTL perspective, there are no disorders, just the extremes of quantitative traits caused by the same genetic and environmental factors responsible for variation throughout the dimension. In other words, the QTL perspective predicts that genes found to be associated with complex disorders will also be associated with normal variation and vice versa (Plomin et al. 1994).

The QTL perspective is the molecular genetic version of the quantitative genetic perspective which assumes that genetic influence on complex traits is due to many genes of varying effect size. The goal is not to find *the* gene for a particular trait, but rather *some* of the many genes that make contributions of varying effect sizes to the variance of the trait. Perhaps one gene will be found that accounts for 5 percent of the variance, five other genes might each account for 2 percent of the variance, and 10 other genes might each account for 1 percent of the variance. If the effects of these QTLs are independent, together these QTLs would account for 25 percent of the variance, or half of the heritable variance for a trait whose heritability is 50 percent. All of the genes that contribute to the heritability of the trait are unlikely to be identified because some of their effects may be too small or complicated to detect. Finding QTLs require powerful genetic designs that can detect genes of small effect. Space does not permit a description of the various methods used to detect QTLs (eg., Plomin and Crabbe 2000).

The memory loss and confusion of dementia called Alzheimer's disease was the first behavioral disorder for which a QTL was found. Although single-gene effects have been identified for rare early-onset cases, late-onset Alzheimer's disorder, which strikes as many as one in five individuals in their eighties, is not caused by a single gene. In 1993, a gene on chromosome 19 called apolipoprotein E was found that predicts risk for the disorder better than any other known risk factor (Corder et al. 1993). If you inherit one copy of a particular form (*allele*) of the gene, your risk for Alzheimer's disorder is about four times greater than if you have another allele. If you inherit two copies of the allele (one from each of your parents), your risk is much greater.

Reading disability was the next common behavioral disorder to which a QTL approach was successfully applied. A linkage between reading disability and DNA markers on the short arm of chromosome 6 (Cardon et al. 1994) has been consistently replicated (Grigorenko et al. 1997, Fisher et al. 1999, Gayán et al. 1999). This QTL appears to be broad in its effect, involving both phonological (auditory) and orthographic (visual) aspects of reading disability (Fisher et al. 1999). Several molecular genetic studies have indicated a role for dopamine genes in the etiology of hyperactivity (Thapar et al. 1999). Genes responsible for the heritability of personality (Hamer and Copeland 1998), and of cognitive abilities (Fisher et al. 1999) are also beginning to be identified.

Although attention is now focused on finding specific genes associated with complex traits, the greatest impact on social and behavioral sciences will come after genes have been identified. Few social and behavioral scientists are likely to join the hunt for genes because it is difficult and expensive, but once genes are found, it is relatively easy and inexpensive to use them. DNA can be obtained inexpensively from cheek swabs rather than blood. One cheek swab yields enough DNA to genotype thousands of genes. Microarrays the size of a postage stamp, called DNA chips, are becoming available that can genotype thousands

of genes in a few minutes at costs that will eventually be very low per individual.

Although some psychology departments already have DNA laboratories, it is likely that most research using DNA in the social and behavioral sciences will be accomplished through collaborations with molecular geneticists or through commercial arrangements. It is critical for the future of the social and behavioral sciences that we be prepared to use DNA in our research. What has happened in the area of dementia in the elderly will be played out in many areas of the social and behavioral sciences. Although the association between the apolipoprotein E gene and late-onset Alzheimer's dementia was only recently identified, most research of any sort on dementia now genotypes subjects for apolipoprotein E in order to ascertain whether the results of interest differ for individuals with and without this genetic risk factor.

Among geneticists, it is generally believed that we will be awash in genes associated with complex traits including behavior in the next few years, especially as the Human Genome Project completes sequencing all 3.5 billion bases of DNA in the human genome and identifies all genes and the several million DNA bases that differ among us. The future of genetic research lies in moving from finding genes (genomics) to finding out how genes work (functional genomics). Functional genomics is usually considered in terms of bottom-up molecular biology at the cellular level of analysis. However, a top-down behavioral level of analysis may be even more valuable in understanding how genes work at the level of the intact organism, in understanding interactions and correlations between genes and environment, and in leading to new treatments and interventions. The phrase 'behavioral genomics' has been suggested to emphasize the importance of top-down levels of analysis in understanding how genes work (Plomin and Crabbe 2000). Bottom-up and top-down levels of analysis of gene-behavior pathways will eventually meet in the brain. The grandest implication for science is that DNA will serve as an integrating force across diverse disciplines.

Social and behavioral scientists should participate constructively in the discussion of scientific, clinical and social implications of the advances brought about by DNA research. Students in the social and behavioral sciences must be taught about genetics in order to prepare them for this future. The search for genes involved in behavior has led to a number of ethical concerns: There is fear that the results will be used to justify social inequality, to select individuals for education or employment, or to enable parents to pick and choose among their fetuses. These concerns are largely based on misunderstandings about how genes affect complex traits (Rutter and Plomin 1997). Otherwise, this opportunity for the social and behavioral sciences will slip away by default to geneticists, and the genetics of behavior is much too important a topic to be left to geneticists!

See also: Developmental Behavioral Genetics and Education; Genetic Factors in Cognition/Intelligence; Genetic Studies of Behavior: Methodology; Genetic Studies of Personality; Genetics and Development; Genetics and Mate Choice; Genetics of Complex Traits Through the Life Cycle; Human Evolutionary Genetics; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Mental Illness, Genetics of; Sexual Preference: Genetic Aspects; Temperament: Familial Analysis and Genetic Aspects

Bibliography

Bouchard T J, Propping P 1993 *Twins as a Tool of Behavioral Genetics*. Wiley, Chichester, UK

Cardon L R, Smith S D, Fulker D W, Kimberling W J, Pennington B F, DeFries J C 1994 Quantitative trait locus for reading disability on chromosome 6. *Science* **266**: 276–9

Christensen K, Vaupel J W, Holmn N V, Yashlin A I 1995 Twin mortality after age six: fetal origin hypothesis versus twin method. *British Medical Journal* **310**: 432–6

Corder E H, Saunders A M, Strittmatter W J, Schmechel D E, Gaskell P C, Small G W, Roses A D, Haines J L, Pericak Vance M A 1993 Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**: 921–3

Deary I 2000 *Looking Down on Human Intelligence: From Psychometrics to the Brain*. Oxford University Press, Oxford, UK

Deary I J, Caryl P G 1997 Neuroscience and human intelligence differences. *Trends in Neurosciences* **20**: 365–71

Fisher P J, Turic D, Williams N M, McGuffin P, Asherson P, Ball D, Craig I, Eley T C, Hill L, Chorney K, Chorney M J, Benbow C P, Lubinski D, Plomin R, Owen M J 1999 DNA pooling identifies QTLs for general cognitive ability in children on chromosome 4. *Human Molecular Genetics* **8**: 915–22

Fisher S E, Marlow A J, Lamb J, Maestrini E, Williams D F, Richardson A J, Weeks D E, Stein J F, Monaco A P 1999 A quantitative-trait locus on chromosome 6p influences different aspects of developmental dyslexia. *American Journal of Human Genetics* **64**: 146–56

Gayán J, Smith S D, Cherny S S, Cardon L R, Fulker D W, Brower A W, Olson R K, Pennington B F, DeFries J C 1999 Quantitative-trait locus for specific language and reading deficits on chromosome 6p. *American Journal of Human Genetics* **64**: 157–64

Goldsmith H H, Buss K A, Lemery K S in press Toddler and childhood temperament: expanded content, stronger genetic evidence, new evidence for the importance of environment. *Developmental Psychology*

Grigorenko E L, Wood F B, Meyer M S, Hart L A, Speed W C, Shuster A, Pauls D L 1997 Susceptibility loci for distinct components of developmental dyslexia on chromosomes 6 and 15. *American Journal of Human Genetics* **60**: 27–39

Hamer D, Copeland P 1998 *Living with Our Genes*, 1st edn. Doubleday, New York

Loehlin J C 1992 *Genes and Environment in Personality Development*. Sage, Newbury Park, CA

Mackintosh N J 1998 *IQ and Human Intelligence*. Oxford University Press, Oxford, UK

Martin N, Boomsma D, Machin G 1997 A twin-pronged attack on complex trait. *Nature Genetics* **17**: 387–92
Nigg J T, Goldsmith H H 1994 Genetics of personality disorders: perspectives from personality and psychopathology research. *Psychological Bulletin* **115**: 346–80
Phillips D I W 1993 Twin studies in medical research: can they tell us whether diseases are genetically determined? *Lancet* **341**: 1008–9
Plomin R, Crabbe J C 2000 DNA. *Psychological Bulletin* **126**(6): 806–28
Plomin R, DeFries J C, McClearn G E, McGuffin P 2001 *Behavioral Genetics*, 4th edn. Worth Publishers, New York
Plomin R, Fulker D W, Corley R, DeFries J C 1997b Nature, nurture and cognitive development from 1 to 16 years: a parent–offspring adoption study. *Psychological Science* **8**: 442–7
Plomin R, Owen M J, McGuffin P 1994 The genetic basis of complex human behaviors. *Science* **264**: 1733–9
Riemann R, Angleitner A, Strelau J 1997 Genetic and environmental influences on personality: a study of twins reared together using the self- and peer report NEO–FFI scales. *Journal of Personality* **65**: 449–76
Rutter M, Silberg J, O'Connor T G, Simonoff E 1999 Genetics and child psychiatry: I. Advances in quantitative and molecular genetics. *Journal of Child Psychology and Psychiatry* **40**: 3–18
Sokol D K, Moore C A, Rose R J, Williams C J, Reed T, Christian J C 1995 Intrapair differences in personality and cognitive ability among young monozygotic twins distinguished by chorion type. *Behavior Genetics* **25**: 457–66
Thapar A, Holmes J, Poulton K, Harrington R 1999 Genetic basis of attention deficit and hyperactivity. *British Journal of Psychiatry* **174**: 105–11

R. Plomin

Behavioral Geography

1. Introduction

1.1 Definition

Behavioral geography includes the study of the processes involved in spatial decision making and the consequent traces of human decisions and movements in the environment. Two components, labeled 'spatial behavior' and 'behavior in space,' are generally recognized. 'Spatial behaviors' are the spatially manifested or overt acts of people performing a range of daily or other episodic activities (e.g., journey to work, shopping, recreation, education, and so on). These acts yield data such as distance and direction of movement, directional bias, trip frequency, episodic interval, and repetitiveness, and are represented and analyzed as occurrences in space. 'Behavior in space' focuses on supplying reasons for these overt acts and requires understanding processes such as decision making and choice, spatial cognition and cognitive mapping, spatial knowledge acquisition, risk aversion, uncertainty, habit, search and learning, emotional state, attitudes, cognitive representations, and values and beliefs.

1.2 The Precursors

Behavioral geography began a critical growth phase in the early 1960s, but isolated researchers in human and physical geography had, during the earlier part of the twentieth century, pointed to the need for linking physical reality and human images of that reality. Gulliver (1908) talked of 'perceived spatial orientation' as an important factor in understanding human activities. Trowbridge (1913) discussed the notion of 'imaginary maps,' making it clear that geography existed not only in the objective physical environment but also in our imagination; that imagination could produce vivid verbal or pictorial descriptions, as were found in numerous novels, poetry, and stories told. A 1997 translation of a book by the Finnish geographer (Granö 1997, first published in 1930) shows that he distinguished between 'proximal' and 'distant' landscapes based on the psychophysics of perception, and urged that geographers concentrate on perceived realities. The worlds of the imagination in the past and the present were emphasized further by Wright (1947) and Kirk (1951). By the early 1960s, therefore, the idea of subjective reality and its distinct geographies had been planted in the discipline.

1.3 History

Geographer Kates and psychologist Wohlwill edited a special issue of the *Journal of Social Issues* in 1966 (Kates and Wohlwill 1966). This publication represented 'an attempt by the issue editors to relate behavioral science to an important societal concern: the quality of the environment' (p. 21). This new emphasis joined the geographer's concern with the physical and built environments and the psychologist's concern for human behavior. In his paper, Kates suggested that a concern for human–environment relations was essential in that it could play a role in mediating between behavioral science and professional participation in environmental protection and conservation programs. Kates suggested that geography could assist in building a bridge between the social and behavioral sciences and the professional disciplines of architecture, design, city planning, and others that were directly concerned with the relations between the environment and human behavior.

Following up on a previously published paper on experience and imagination in geography, Lowenthal

(1967) edited a collection that explored human–environment relations from viewpoints in geography, psychology, and planning. Lowenthal's aim was to show that what he called 'the three realms' of geography could be brought together in common study. These realms included the nature of the environment, what we think and feel about the environment, and how we behave in and alter the environment. Behavioral geography generally has followed Lowenthal's suggestion of integrating these three realms, as well as pursuing more detailed examinations of each of them separately.

A third significant publication, edited by Cox and Golledge (1969), formalized the idea of a behavioral stream in geographic research that emphasized individual rather than aggregate behaviors, and suggested replacing ties to economics with ties to psychology. Together with a Penguin book on *Mental Maps* (Gould and White 1974), and a 1972 review of behavioral research in geography (Golledge et al. 1972), these publications established behavioral geography as a distinct and definable subfield of geography.

The message in each of these publications was that people gave life and meaning to different environments by virtue of their interaction with them and their behaviors in them. It was but a short step to arguing that human transformed environments were the result of deliberated decision-making activities, that the resulting structure of the environment reflected human values, beliefs, wants, and needs, and that human actions in those environments were preserved by immediate perceptions and stored in and recalled from long-term memory. As part of understanding human actions and activities, it was necessary to examine how human–environment relations were mediated by social, economic, cultural, political, psychological, and other constraints, each of which produced a more or less 'hidden' structure, which constrained the world as we knew it and experienced it. The next step simply was to acknowledge that to know an environment required learning about critical subsets of information from the mass of experiences open to each individual. Different people received, stored, and used different sensory input (or 'messages') from the different environments in which humanity lived. This differential access and use of information was translated into perceivable differences in the everyday task of living (Hanson and Hanson 1993).

1.4 International Spread

Behavioral geography emerged in the USA as a cohesive unit in the 1960s, consolidated in the 1970s, fragmented somewhat in the 1980s, and has had resurgent growth in the 1990s. Research in the UK (Downs 1970), which defined cognitive dimensions of people's images of shopping centers, stimulated more research using personal construct theory borrowed from psychiatry on human behavioral processes relating to how images are created. Recent activity has been stimulated by Kitchin's (1994) paper on cognitive maps, and by the place-based work on children's environments by Matthews (1992). In Japan, according to Wakabayashi (1996), the environmental perception component of behavioral geography was first developed, with spatial cognition and cognitive mapping emerging much later. In Spain, Aragonés and Arredondo (1985) summarized the growing behavioral interest by examining how cognitive maps developed and how imagery rather than fact seemed to influence planning policy and its implementation in urban areas. Early behavioral research in Australia emphasized cultural perception of the environment as perceived in various settings (e.g., Brookfield 1969). In Sweden, Hägerstrand's (1957) work on migration and mobility, followed by research on diffusion of innovations, initiated yet another stream of behaviorally based work including research on the process of spatial diffusion and the spatial adoption patterns of new ideas and new products. Again following Hägerstrand's lead, the 1970s saw the emergence of another behavioral stream focused on time–space analysis (Carlstein et al. 1978). Here time paths and space–time budgets were defined for daily activities and summarized in time–space frames called 'prisms.' In India, Singh and Singh Rana (1980) pioneered work on cognitive mapping and influenced research on other developmental and cognitive concepts.

During the 1970s, geographers in the USA strengthened their interdisciplinary links as a means of developing better concepts and research tools and models for behavioral geographic research. Much of this occurred via the multidisciplinary activities of the Environmental Design and Research Association (EDRA) that began publishing the journal *Environment and Behavior* in 1969. There is now an active 250 member Specialty Group of the Association of American Geographers (Environmental Perception and Behavioral Geography Specialty Group) and an average of 25 Masters theses or Ph.D. dissertations are produced in this area each year. Although early work in the USA was concentrated in the research activities of relatively few geographers, in the 1970s there was an escalated search for process based and behaviorally based explanations, particularly drawing on material from perceptual, developmental, and cognitive psychology, and from the work on emotion, affect, and attitudes from social psychology. In 2000, the subdiscipline is more proactive than reactive, reaching out to environmental and cognitive psychology (particularly spatial cognition), cognitive science, artificial intelligence, computer science, robotics, information sciences and technologies, virtual environments, and other newly developing and frontier areas of human science.

2. Major Subdivisions

2.1 Spatial Decision-making and Choice Behavior

The geographic emphasis now known as behavioral geography was prompted in part by a 1960's reaction against the unrealistic assumptions of normative economic–geographic models of human decision-making and choice behavior (Golledge 1967). Assuming perfect information, these theories claimed that humans were spatially and economically 'rational,' invariably selecting least cost and closest place alternatives in economic decision-making situations.

At the time that the theoretical and quantitative revolution was under way in geography, the major theories being investigated and used covered industrial, agricultural, and urban location, where economic models emphasizing perfect information and producing optimal and equilibrium solutions dominated the research domain. A critical assumption of all these models was that of economically rational behavior, usually interpreted in geography as least-effort or distance-minimizing behavior. The theories were geographically appealing because it was possible to hold economic variables such as price, quantity, production costs, labor costs, and so on constant and to focus only on transportation costs—the cost of overcoming distance of shipping raw materials or products—as the critical locating variable. In the geographic domain, the economically rational behavior assumption was translated into a spatially rational behavior assumption, where it was hypothesized that the nearest place of supply or demand would invariably have an advantage (by minimizing transport costs) and thus increase the probability of being chosen. The geographer's dissatisfaction with this as a fundamental assumption arose as empirical evidence showed that spatially rational behavior occurred only in a relatively few specialized cases. For example, Gould's (1963) introduction of game theory to geographers showed that as one changed one's attitude towards risk and uncertainty, the optimal solution to a geographic problem (e.g., which crops to grow in a subsistence economy given different environmental conditions) could also change, while Wolpert (1965) showed that Swedish farmers were far from being economically rational in their annual choice of crop types and land use combinations.

Spatial decision-making and choice behaviors have continued to be an important focus. This interest includes locational decisions (e.g., for new shopping centers, emergency services, group houses, child care centers) and choices that influence activities such as consumer behavior, migration, mobility, recreation, tourism, busing, school closure or expansion, and many others. It has also generated an interest in transportation varying from simple mode choice modeling and the building of computational process models of household activity patterns, to the use of global positioning systems (GPS) to track accurately daily behavior in road networks as inputs for an advanced traveler information component of an intelligent transportation system. Using survey diaries to collect data on attitudes towards and preferences for pretrip and en-route advice, behavioral geographers are contributing to a more complete understanding of today's complex travel patterns (e.g., by showing that, given delay information due to congestion or accident in the morning, people will change departure times rather than change routes of travel, whereas the reverse decision is more common in the evening).

2.2 Hazard Research

Following the lead of Gilbert White, who researched people's attitudes towards and reactions to flood hazards (White 1945), geography researchers at the University of Chicago began examining human attitudes towards and adjustments to extreme environmental events. For example, Kates (1966) examined risk and uncertainty as it influenced human responses to severe coastal storms. He showed that increased exposure to storms changed residents' subjective probabilities of storm occurrence periodicity and intensity. Burton, Kates, and White examined recidivist behavior of persons living on flood plains or other hazardous areas (Burton et al. 1978), showing how inertia brought about by investment in land, inheritance, family history, and experience influenced people's attitudes towards and expectations about future flood events. Research on perceptions of drought (Saarinen 1965) initiated a flurry of work labeled 'environmental perception' that illustrated methods for producing attitudinal data on drought and its acceptance as part of the cycle of human–environment relations. These works had an important impact on the worldwide interest in environmental protection and conservation that continues to stimulate a large part of behavioral research today.

Risk assessment, particularly in the context of human responses to hazardous events, has been a fundamental part of the growth of research in this area. It represents one of the founding pillars of interest (Kasperson and Dow 1993). It emphasizes examining subjective probabilities associated with catastrophic events, occurrences, problems of evacuation at times of risk, and the preferred behavior of persons exposed to environmental risk.

2.3 Spatial Knowledge Acquisition

In the early 1960s, geographers and psychologists at Clark University began joint work on objective and subjective spaces, eventually leading to an emphasis

on children's learning activities in a developmental context (Blaut and Stea 1969). They showed that even preschoolers could interpret aerial photos of places. It was suggested by Hart and Moore (1973) that humans passed through stages of spatial knowledge acquisition ranging from egocentric and topologically structured spaces to abstract and fully metric spaces in a manner similar to that outlined in developmental psychology studies of intellectual development. Much research followed on children's environments ranging from playgrounds to urban and rural neighborhoods. Research illustrated that female children observed and remembered more local details (e.g., landmarks), whereas male children traveled further and developed a better metric understanding of space, producing more exocentric rather than egocentric representations of experienced environments.

2.4 *Travel Activity Analysis*

In the USA, Marble (1967) led behavioral researchers to the study of human movement over transportation networks. Perry Hanson (1977) and Susan Hanson (1984) investigated disaggregate behavior by linking transportation research and activity analysis (i.e., daily movement behavior), while Australia's K. P. Burnett examined spatial choice of shopping centers (Burnett 1973), emphasizing the different criteria used by men and women to select shopping destinations. Other researchers investigated urban movement behavior, defining action and activity spaces which summarized, respectively, the total potential area of daily interaction and the actual pattern of a household's daily travel behavior. Consumers searching for products called 'spatially durable/shopping goods' or 'spatially inflexible convenience goods,' respectively, were shown to generate a variety of trips to different places at varying distances from the home base. Convenience shopping often conformed to minimum distance or closest place assumptions, while the search for places at which to purchase shopping goods resulted in 'variety-seeking behavior' which evinced spatially and economically 'irrational' behavior. The results began undermining the acceptability of the normative theories borrowed from economics and launched a search for more realistic assumptions about human spatial behavior. The result was endorsement of the idea that spatial behavior was best conceived as being 'boundedly rational.'

2.5 *Spatial Cognition and Cognitive Maps*

Research on cognitive maps was initiated in response to the question 'what do individuals know about places?' Cognitive maps were defined as a person's internally stored knowledge of the world and its events and processes. Using tasks ranging from sketch mapping to nonmetric multidimensional scaling, attempts were made to discover what people knew—or could recall—about places. External representations of this knowledge (e.g., sketches, verbal descriptions, photo identification, interpoint distance knowledge) showed that the geographic information stored and recalled by people usually was incomplete, fragmented, and distorted. In turn, this necessitated research on cognitive mapping processes, including encoding, decoding, recall, and internal manipulation of sensed information (e.g., estimating distances between real or imagined places).

Much effort was spent on designing meaningful ways to externalize one's spatial knowledge and to represent it as a spatial product that could be evaluated by others. In particular, geographers used multidimensional methods (e.g., nonmetric scaling) to represent the latent structure of geographical knowledge and developed methods of association (e.g., bidimensional regression) to provide a measure of the matching between objective geographic reality and the spatial product obtained from individuals. In part, this research aimed at answering questions as to why people went to various places to perform selected activities when alternative places seemed to be more rational alternatives (e.g., bypassing one nearby store to shop at a further one). Cognitive map research often showed that the location of the spatially 'nearby' place was not known or was wildly distorted, making a further place 'seem' near and more accessible.

Spatial cognition, spatial knowledge acquisition, environmental perception, and cognitive mapping have all been a primary focus of international attention during the 1980s and 1990s (Portugali 1996, Saarinen 1998). Emphasis has been placed on the use of cognitive maps in human wayfinding. Particular success has been achieved in comprehending the distorting effects on human knowledge of geographic alignment (e.g., is Reno, Nevada east or west of San Diego, California?), map orientation and positioning (e.g., as in determining the optimal facing direction of 'you are here' maps), and in discovering variations in spatial knowledge acquisition (e.g., differences between children's, adults', and disabled persons' environmental learning activities).

2.6 '*Place*' *and* '*Landscape*'

Throughout this period of growth there was continuous attention to the concepts of 'place' and 'landscape.' Related research covered a wide range of situations including examining the place of imagination in geographic experiences, defining both human and environmental personality traits and fostering a renewed interest in concepts of 'place' (Tuan 1977, Relph 1976). Researchers examined 'landscapes of

fear' in institutional and urban environments where criminal activity was perceived to endanger normal person–environment interactions. Behavioral geographers defined the sets of environmental features whose presence gave criminals (e.g., residential burglars or street drug sellers) a sense of security when carrying out their trade. Researchers also investigated how places developed images that encouraged or discouraged human visitation (e.g., places of beauty vs. places of fear), which in turn were incorporated into explanations of recreational park use and tourist destination choice.

The geographic study of variations in emotional reactions to different environments continues to play a key part in the practice of behavioral geography and remains significant (Amedeo 1993). In particular, emphasis on place stereotypes and the role of emotional appeal in residential site selection and neighborhood choice stands out. In association with these emphases, there has been a concurrent rethinking of the role of attitudes as a factor influencing human behavior, and an examination of fundamental belief and value systems that mediated human behaviors and which gave meaning to different environments. The latter has been an important part of the study of the ethics of environmental degradation.

2.7 *Adoption and Diffusion of Innovations*

Of perhaps equal importance has been the concern with information processing and the diffusion of both information and innovations. This has generated a strong tie with medical geography, as the spatial and temporal spread of phenomena (e.g. HIV/AIDS, influenza, and other communicable diseases) have been mapped and analyzed to highlight and explain their underlying geography (e.g., Gould 1993).

2.8 *Spatial Behavior of Special Groups*

Since the early 1980s, a growing interest in the behavioral problems of particular social groups has emerged. These groups include underrepresented minorities, the elderly, the homeless, and disabled people (Dear 1991, Golledge 1993). Outstanding in this area is the research on the NIMBY ('not in my back yard') syndrome used to explain negative reactions to local government policy decisions related to the location of what are perceived to be 'noxious' facilities such as group home and rehabilitation centers for drug users, sex offenders, and developmentally disadvantaged people.

Another new and powerful interest is in the geography of female roles and activities, in particular clarifying questions of 'spatial mismatch' between female home and work locations (Hanson and Pratt 1995). Analysis of distances traveled by ethnic and socioeconomic disadvantaged groups has effectively shown that poor female ethnic workers typically have to travel further to their jobs than do other group members.

2.9 *Cognitive and Behavioral Geography and Geographic Information Science*

Cognitive behavioral researchers have joined with researchers in geographic information science (GISc) to study visualization, cognitive cartography, scale, frames of reference, spatial language, verbal protocols for communicating with computers, and multimedia representational formats for digitized data. For example, investigations of whether verbal or mapped representations of geographic information facilitate environmental learning clearly indicate the superiority of maps or graphics wherever possible to impart spatial information, and these results have contributed to the growth of graphic user interfaces (GUI) as the primary human–computer interface.

2.10 *Residential Movement*

Mobility and migration research has been strongly influenced by behavioral research concepts relating to neighborhood preference, destination images, and perceived dimensions of environmental quality (Clark 1986). In particular, human reactions to segregation of housing, ethnic neighborhood growth, and busing of ethnic schoolchildren have been dominant themes for behavioral inquiry.

3. *Promising Future Research Directions*

As the world surges into the twenty-first century—hailed as the Age of Information Technology—behavioral geography is surging with it, no longer being pulled along by other fields of interest. One of the more promising areas of relevance to society at large includes a growing interest in geographic education. This is in part stimulated by the need to investigate spatial knowledge acquisition and new teaching practices tied to gender-free methods and computer-based distance learning technologies. As part of the twenty-first century educational scene, behavioral geographers are beginning to explore virtual environments, particularly the similarities and differences involved in undertaking spatial behavior in real and virtual worlds (e.g., comparing the effectiveness of learning an environment via an immersed virtual trip or a real world experience).

A promising area of future research is the further development of behavior-friendly digitized spatial databases together with analytical tools that explicitly assist in human understanding of data visualization processes. This exploration (via the perceptual and cognitive processes) of the significance of visualiza-

tion, audition and haptics as modes for expressing and representing spatial behavioral activities links behavioral research with cartography and geographic information systems (GIS). Behavioral and cognitive research is providing an important theoretical and analytical component of the emerging field of GISc. For example, exploring the nature of place cells for the neurophysiological evidence that will help decide if there is an understandable geography that the brain uses to store spatial information is becoming a multidisciplinary effort in which behavioral geographers are participating. A tie with spatial linguistics has also developed as researchers examine how language structures thought about space, how verbal interactions influence the way we understand space, and how natural language interfaces can be developed to communicate with computers that access geocoded data and use digital earth database metaphors.

In essence, much of tomorrow's behavioral geography will focus on an examination of the geography of everyday life, including comprehension of individual and group movements. For example, research involved with unpacking the implications of individual and household activity scheduling is promoting the development of intelligent transportation systems. This work has played a major part in establishing the International Association of Travel Behavior Researchers (IATBR). In this context, traditional behavioral research on travel and activities is being complemented by research on human navigation and wayfinding processes in different environments and their relationship to the design and building of artificial intelligences and robots.

A final area of fruitful research includes discovering the structure and influence on spatial behavior of the various realities in which groups of disabled and disenfranchised people must live out their lives. This area assesses the impact on travel of environmental obstacles (e.g., lack of ramps or curb cuts for wheelchairs) that inhibit movement and cause lengthy detours between otherwise easily accessible origins and destinations.

4. *Summary*

Behavioral geographers introduced a number of important changes to the discipline. These included (a) the introduction of new models of human behavior that went far beyond the assumptions of economically and spatially rational behavior to include beings who develop attitudes, emote, are more or less risk averse, have values and beliefs, and are part of sociocultural systems; (b) the expansion of the term 'environment' to include not only the objectively observable natural, biotic and built environments, but also the many hidden environments of human social, economic, physical, linguistic, religious, and other systems that were embedded in human existence; (c) the introduction of the notions of perceptual and cognitive environments which mediated processes of human–environment interaction and provided reference frames for knowing about places; (d) a changed emphasis from the aggregate macroscale focus to disaggregate microscale focus in which individual differences and personal characteristics and habits were seen to be important explanatory factors; (e) the development of new data based on human responses rather than simple counts or objective measures of the quantity of behavior occurring at places; (f) the introduction of different research designs (e.g., ethnomethodology, phenomenology, semiotics, and case studies) that facilitated the collection of behavioral data with reduced fear of experimenter contamination; (g) the importation and development of new methods of analysis (both quantitative and qualitative) that were suited to the subjective, individual, and disaggregate data that was being generated by interview, survey, observation, and focus group discussion; (h) the adoption of multimedia approaches that combined subjective and objective data collection and analysis procedures; (i) the emphasis on the need for valid and reliable research results; (j) the fostering of a growing suspicion that results of behavioral research cannot easily be logically connected into the axioms, theorems, laws, and deductive theories more typical of traditional scientific method as practiced in the natural and physical sciences; and (k) the suggestion of developing integrated place-based science that went beyond mere 'location' in geographic domains.

In sum, behavioral geography has a wide avenue to follow and, if followed successfully, can significantly impact the discipline of geography and enhance the quality of life of those who inhabit this world. Behavioral geographers in the twenty-first century are recognized outside their own discipline for work on cognitive mapping and spatial cognition, environmental learning, preference modeling, discrete choice and decompositional preference modeling, space–time geography, understanding place, contextual analysis, environmental risk assessment, responses to hazards, and many other areas of human–environment relations.

See also: Cognitive Maps; Diffusion: Geographical Aspects; Evolution: Diffusion of Innovations

Bibliography

Amedeo D 1993 Emotions in person-behavior episodes. In: Gärling T, Golledge R G (eds.) *Behavior and Environment*. Elsevier, Amsterdam, pp. 83–116

Aragonés J, Arredondo J 1985 Structure of urban cognitive maps. *Journal of Environmental Psychology* **5**: 197–212

Blaut J M, Stea D 1969 *Place Learning*. Place Perception Report No. 4. Graduate School of Geography, Clark University, Worcester, MA

Brookfield H 1969 On the environment as perceived. In: Board C, Chorley R, Haggett P, Stoddart D (eds.) *Progress in Geography*. Arnold, London, Vol. 1

Burnett K P 1973 The dimensions of alternatives in spatial choice processes. *Geographical Analysis* **5**: 181–204
Burton I, Kates R W, White G F 1978 *The Environment as Hazard*. Oxford University Press, New York
Carlstein T, Parkes D, Thrift N 1978 *Timing Space and Spacing Time I: Making Sense of Time*. Arnold, London, Vols. I–III
Clark W A V 1986 *Human Migration*. Sage, Beverly Hills, CA
Cox K R, Golledge R G (eds.) 1969 *Behavioral Problems in Geography: a Symposium*. Studies in Geography, No. 17. Northwestern University Press, Evanston, IL
Dear M 1991 Gaining community acceptance. In: Poster C, Longergan E (eds.) *Resolving Locational Conflicts, Not-in-My-Backyard. Part 1: Human Service Facilities and Neighborhoods*. Roy P. Drachman Institute for Land and Regional Development Studies, University of Arizona, and Information and Referral Services, Tucson, AZ, pp. 33–43
Downs R M 1970 The cognitive structure of an urban shopping center. *Environment and Behavior* **2**: 13–39
Golledge R G 1967 Conceptualizing the market decision process. *Journal of Regional Science* **7**: 239–58
Golledge R G 1993 Geography and the disabled: a survey with special reference to vision impaired and blind populations. *Transactions of the Institute of British Geographers* **18**: 63–85
Golledge R G, Brown L, Williamson F 1972 Behavioral approaches in geography: an overview. *Australian Geographer* **12**: 59–79
Gould P 1963 Man against his environment: a game theoretic framework. *Annals of the Association of American Geographers* **53**: 290–297
Gould P 1993 *The Slow Plague: a Geography of the AIDS Pandemic*. Blackwell, Oxford, UK
Gould P, White R 1974 *Mental Maps*. Penguin, Baltimore, MD
Granö J G 1997 In: Granö O, Paasi A (eds.) [Hicks M (transl.)] *Pure Geography*. Johns Hopkins University Press, Baltimore, MD (original work published as *Reine Geographie* by the Geographical Society of Finland in *Acta Geographica*, vol. 2, 1929, and as *Puhdas Maantiede* by Werner Söderström, Porvoo, 1930)
Gulliver F P 1908 Orientation of maps. *Journal of Geography* **7**: 55–8
Hägerstrand T 1957 Migration and area. In: Hannerberg D, Hägerstrand T, Odeving B (eds.) *Migration in Sweden: a Symposium. Series B, Human Geography, No. 13, Lund Studies in Geography*. Gleerup, Lund, Sweden, pp. 27–158
Hanson P 1977 The activity patterns of elderly households. *Geografiska Annaler*, Series B **59**: 109–24
Hanson S 1984 Environmental cognition and travel behavior. In: Herbert D T, Johnston R J (eds.) *Geography in the Urban Environment: Progress and Research in Applications*. Wiley, Chichester, UK, pp. 95–126
Hanson S, Hanson P 1993 The geography of everyday life. In: Gärling T, Golledge R G (eds.) *Behavior and Environment*. Elsevier, Amsterdam, pp. 249–69
Hanson S, Pratt G 1995 *Gender, Work, and Space*. Routledge, New York
Hart R A, Moore G T 1973 The development of spatial cognition: a review. In: Downs R M, Stea D (eds.) *Image and Environment: Cognitive Mapping and Spatial Behavior*. Aldine, Chicago, pp. 246–88
Kasperson R E, Dow K 1993 Hazard perception and geography. In: Gärling T, Golledge R G (eds.) *Behavior and Environment: Psychological and Geographical Approaches*. North-Holland, Amsterdam, pp. 193–222
Kates R W 1966 Stimulus and symbol: the view from the bridge. *Journal of Social Issues* **22**(4): 21–8
Kates R W, Wohlwill J F 1966 Man's response to the physical environment. *Journal of Social Issues* **22**(4): 1–140
Kirk W 1951 Historical geography and the concept of the behavioral environment. In: Kuriyan G (ed.) *Indian Geographical Journal, Silver Jubilee Edition*. Indian Geographical Society, Madras, India
Kitchin R M 1994 Cognitive maps: what are they and why study them? *Journal of Environmental Psychology* **14**: 1–19
Lowenthal D (ed.) 1967 *Environmental Perception and Behavior*. Research Paper No. 109. Department of Geography, University of Chicago, Chicago
Marble D F 1967 A theoretical exploration of individual travel behavior. In: Garrison W L, Marble D F (eds.) *Quantitative Geography, Part I (Economic and Cultural Topics)*. Studies in Geography, No. 13. Department of Geography, Northwestern University, Evanston, IL, pp. 33–53
Matthews M H 1992 *Making Sense of Place: Children's Understanding of Large-scale*. Barnes and Noble, Savage, MD
Portugali J (ed.) 1996 *The Construction of Cognitive Maps*. Kluwer, Dordrecht, The Netherlands
Relph E 1976 *Place and Placelessness*. Pion, London
Saarinen T F (ed.) 1965 *Perception of the Drought Hazard on the Great Plains*. Research Paper 106. Department of Geography, University of Chicago, Chicago
Saarinen T F 1998 *World Sketch Maps: Drawing Skills or Knowledge*. Discussion Paper 98-7. Department of Geography and Regional Development, University of Arizona, Tucson, AZ
Singh R L, Singh Rana P B 1980 Cognizing urban landscape of Varanasi: a note on cultural synthesis. *National Geographic Journal of India* **25**(3–4): 113–23
Trowbridge C C 1913 On fundamental methods of orientation and imaginary maps. *Science* **38**: 888–97
Tuan Y-F 1977 *Space and Place: the Perspective of Experience*. University of Minnesota Press, Minneapolis, MN
Wakabayashi Y 1996 Behavioral studies on environmental perception by Japanese geographers. *Geographical Review of Japan, Series B* **69**: 83–94
White G F 1945 *Human Adjustment to Floods: a Geographical Approach to the Flood Problem in the United States*. Department of Geography Research Paper No. 29. University of Chicago, Chicago
Wolpert J 1965 Behavioral aspects of the decision to migrate. *Papers and Proceedings in the Regional Science Association* **15**: 159–69
Wright J K 1947 Terrae incognitae: the place of the imagination in geography. *Annals of the Association of American Geographers* **37**: 1–15

R. G. Golledge

Behavioral Medicine

Behavioral medicine is the interdisciplinary field of research and practice that integrates behavioral, biological, and medical science in order to understand, prevent, and treat a broad range of physical health problems. Although behavioral medicine is a relatively new field, its research and clinical programs are now established in most major universities and medical

centers. This article will trace the evolution of behavioral medicine, describe its theoretical foundations, and review its clinical applications.

1. The Evolution of the Concept of Medicine

Traditionally, Western conceptions of the term 'medicine' refer to a substance that is administered into the body that carries out disease-fighting activities autonomously, acting without patient volition or even awareness of its activities. In a broader sense, medicine refers to a field or group of caregivers who heal the sick and injured, again largely beyond the active involvement of the patient. Underlying these conventional views are basic assumptions that medicine and disease processes operate on strictly biological levels, independent of thoughts, emotions, or behaviors. Medicine and health have not always been viewed in this way, but over the past two centuries the emergence of biological models of health and illness have contributed to highly mechanistic and biologically-oriented perspectives on health and well-being. This perspective is exemplified by passive patients who believe that they are made ill by pathogens or imbalances in their bodies and that they can be made well by treatments that are applied by health professionals. In this view, the sick individual bears little responsibility for the etiology of their malady or for getting well.

It is now clear that these notions of medicine and patient roles in health and illness processes are limited. This does not mean that the massive research and technological revolutions in medicine have been for naught; physicians are better prepared to maintain and restore human health than ever before. However, the emergence of behavioral medicine reflects the scientific recognition that health and disease are not just biological phenomena—that how people act, think, and feel are critical in determining who gets sick, how sick they become, and how well they can be treated. To some extent this reflects a return to doctrine that characterized the early history of medicine, before the biomedical revolution. Ancient Greek physicians and philosophers believed that the mind and body were mutually interdependent, and this interactionist approach was evident in medical practice as late as the Renaissance. Greek physicians prescribed such treatments as relaxation-inducing baths and massages, and medieval remedies included listening to music and thinking comforting thoughts, with objectives resembling modern approaches to relaxation and stress management.

1.1 The Emergence of the Biomedical Model

Despite the predominance of these early holistic approaches, Western philosophical traditions later embraced a radical and categorical separation of mind and body. By the seventeenth century, scholars such as French philosopher Rene Descartes argued that mental events and the physical body were independent of one another, overlapping only occasionally. By this account, bodily disease was properly understood as a biological event, and physicians of the day concentrated on this realm when attempting to understand or treat disease. This reductionist approach found strong empirical support in the scientific revolutions of the nineteenth and early twentieth centuries, marked by discoveries of disease-causing pathogens and the development of vaccines and antibiotic medications. A biologically focused biomedical model of health and disease emerged, and has dominated contemporary thinking about disease origin and healing.

It is probably best to view this enduring biological bias in medicine as an overcorrection, a response to centuries of unfounded dogma guiding medical theory and practice. Prior to the nineteenth century, medical authority was saturated with superstition and guesswork. Theories of the origins of disease had centered on curses, sins, or imbalances of bodily humors, and physicians had touted specious interventions such as leeches, purgatives, bloodletting, pouring boiling oil over wounds, and mixing precious stones with wine. As recently as the mid-nineteenth century, physicians did not know to wash their hands and instruments before performing surgery. But the dramatic advances in biological medicine that ensued were remarkable. Agostino Bassi and Robert Koch demonstrated that infectious diseases were caused by pathogens that could spread; Edward Jenner and Louis Pasteur exploited biological derivatives of these pathogens to produce the first vaccines; and Alexander Fleming discovered the antibiotic properties of penicillin. With these advances, physicians became increasingly able to base their authority on empirical evidence and reason. The medical model that emerged from these and other discoveries was received as a welcomed alternative to more primitive models, and as advances in biological sciences revolutionized medicine, it is not surprising that medicine embraced a strongly biologically oriented model.

1.2 The Emergence of Behavioral Models of Health and Illness

Regardless of the reasons for its pre-eminence, the biomedical model is too narrow and simply cannot explain a panoply of diseases that have replaced infectious illness as the major public health problems over the course of the twentieth century. A number of nineteenth century physicians, including Henry Holland, Benjamin Rush, Claude Bernard, and Sir William Osler, recognized these limitations and reintroduced the idea that psychological or behavioral

influences were also important for health and disease. However, these urgings were largely ignored until well into the twentieth century. First, it became apparent that pathogen models of disease were ill suited to address modern health challenges. At the dawn of the twentieth century, microorganisms were the principal sources of disease and mortality—tuberculosis, cholera, diphtheria, influenza, and other major infectious diseases were caused by viral or bacterial invaders and were best treated with medicines that attacked these pathogens or prepared the body to attack them. As these diseases were conquered, heart disease, stroke, cancer, HIV disease, diabetes, and arthritis emerged as the major public health challenges. At the same time, research from the fields of psychosomatic medicine and behavioral and experimental psychology began to provide compelling evidence of mind–body, or psychophysiological, interactions. For example, Pavlov's (1927) classical conditioning paradigm demonstrated that physiological responses could be influenced by non-biological cues. This important work was continued and broadened by Miller's (1978) seminal work on conditioning and control of autonomic functions such as heart rate, brain waves, or body temperature, as well as by Cohen and Ader (1988) who demonstrated that immune system functions could also be conditioned to neutral stimuli. By advocating a rigorous scientific approach to the simultaneous measurement of physiological and psychological indices, Wolff (1953) made important advances in the area of stress and adaptation in disease. Alexander (1950), Graham (1972), and Mason (1975) also achieved important insights into the relationship between emotional, psychological, and physiological responses. Finally, major epidemiological studies were completed that indicated significant associations between certain behavioral or psychosocial factors and health status (e.g., Berkman and Syme 1979).

Towards the latter part of the twentieth century, this evidence made it clear that biological states and disease processes are multidetermined—caused by a confluence of genetic, biological, environmental, behavioral, and psychosocial factors. This spawned more inclusive, sophisticated, integrative theories of health and illness, culminating in the advent of biobehavioral models of health and disease, which represent an expansion of biomedical models to include psychosocial and behavioral components.

2. *Behavioral and Psychosocial Influences on Health*

The significance of behavior in the field of behavioral medicine is apparent in two related principles about the relationship between behavior and disease, each of which is distinct from more traditional biomedical models of disease. First, behavior and other psychosocial factors contribute to disease etiology. Second, behavior and other psychosocial factors can be used for disease treatment and prevention. The first of these principles has been a primary focus of behavioral medicine research; the second is the basis for behavioral medicine clinical applications.

Prevention and treatment approaches in behavioral medicine are generally based on substantive empirical literatures. These investigations have defined three general pathways through which behavioral or psychosocial factors may influence the etiology, progression, treatment, or prevention of disease: (a) direct effects of cognitive, emotional, or behavioral processes on bodily systems (e.g., nervous, endocrine, or immune systems); (b) health-enhancing or health-impairing behaviors, such as diet, exercise, or smoking; and (c) responses to perceived or actual illness, such as screening behaviors or adherence with treatment recommendations. Each of these pathways will be examined in turn.

2.1 *Biobehavioral Influences in Disease Processes*

Research on the biological correlates of emotional or psychological processes has shown that the nervous, immune, and endocrine systems interact in a reciprocal fashion and regulate a variety of states and conditions. The recent development of the fields of psychoneuroendocrinology and psychoneuroimmunology are evidence of the growing recognition of these bidirectional relationships. Moreover, it has become increasingly clear that these pathways can directly or indirectly affect the etiology or progression of disease and affect disease resistance or vulnerability (e.g., see *Emotions and Health*).

The biological effects of stress in particular have been the focus of considerable research. Although the definition of 'stress' and its usefulness as a construct have been the subject of much debate (e.g., see *Stress: Measurement by Self-report and Interview*), it generally refers to a characteristic pattern of physiological, affective, cognitive, and behavioral changes that mobilize resources and support adaptation to environmental threat. Stress involves appraisal of threat, harm/loss, or challenge (Lazarus 1966), and is associated with a cascade of biological changes including activation of two major neuroendocrine systems: the sympathetic adrenome dullary (SAM) and the hypothalamic–pituitary–adrenocortical (HPA) axes. This activation is manifested by alterations in cardiovascular and respiratory function, digestion, metabolism, and skeletal muscle tone, which are coordinated to mobilize or liberate stored energy in preparation for behavioral response to threat.

Most theorists believe that these systemic changes are adaptive and facilitate immediate survival. This value is most apparent in acute stress situations, particularly those where strength and/or speed are critical coping elements. In situations in which activation of these systems is prolonged or unusually

intense, the value of arousal and activation is less clear and may instead exert a toll on health. The pathophysiology of cardiovascular disease serves as an example.

Cardiovascular disease is the leading cause of death in industrialized countries during this century. In the US, it has been the leading cause of death every year except one since 1900. A primary type of cardiovascular disease is coronary artery disease (CAD), which results from atherosclerosis, the accumulation of hardened deposits or plaque on the inner lining of coronary arteries. When enough plaque forms, blood to the heart is restricted or cut off, resulting in clinical manifestations of CAD including heart attack and sudden cardiac death. Although the pathogenesis of atherosclerosis is not completely understood, a number of lines of evidence suggest that sustained or frequent stress-related neuroendocrine activation is involved. Research suggests that individuals whose personalities are marked by impatience, hostility, and antagonism appear to be at risk for CAD (e.g., see *Coronary-prone Behavior, Type A*), as are individuals who have recently undergone stressful life events. In experimental studies with nonhuman primates, Manuck et al. (1995) demonstrated that disruption of the social environment, heightened cardiovascular reactivity, and behavioral factors such as dominance promoted atherosclerosis. Further, these researchers found that neuroendocrine pathways likely mediated this effect. Sustained or frequent increases in endocrine activity may foster atherosclerosis by damaging arteries and coronary vessels as well as by promoting the formation of blood clots. The physiological consequences of stress and some negative emotional processes appear to contribute significantly to the development of cardiovascular disease (e.g., see *Coronary Heart Disease (CHD): Psychosocial Aspects*).

Immune system activity also appears to be altered by behavioral and emotional processes (e.g., see *Psychoneuroimmunology*). Human and animal studies have demonstrated that the immune system is regulated by the central nervous system, by way of a broad array of nervous and endocrine system pathways (Ader et al. 1991). Several psychosocial factors have been related to changes in immune function, although the complexity and plurality of the immune system make it difficult to generalize or interpret these changes. These factors include loneliness, poor social support, negative mood, disruption of marital relationships, bereavement, and various other forms of pre-existing or new acute and chronic stress (reviewed in Cohen and Herbert 1996). These findings suggest that social and interpersonal support and contact are important in the stability or maintenance of immune functioning (e.g., see *Social Support and Health*).

Although the implications of stress-related immune changes for disease vulnerability have generally not been established, there are some empirical and theoretical reasons to suspect that these changes influence the incidence or progression of infectious disease (Cohen et al. 1991) (see *Infectious Diseases: Psychosocial Aspects*), HIV (Leserman et al. 1997) (see *Sexually Transmitted Diseases: Psychosocial Aspects*), and cancer (Andersen et al. 1994) (see *Cancer: Psychosocial Aspects*). Stress-related changes in immune cells known as natural killer (NK) cells may be particularly relevant for some of these conditions since NK cells have the ability to spontaneously destroy virally infected cells and cancer cells. Psychosocial factors, such as social support, have been linked to better NK activity as well as to cancer survival. Stress-related changes in immune function represent a plausible biobehavioral pathway through which psychosocial factors may influence the development or spread of cancer or other illnesses. Other immune-mediated illnesses such as asthma, arthritis, irritable bowel syndrome, and psoriasis may be influenced through similar biobehavioral pathways.

2.2 Health-protective and Health-impairing Behaviors

In addition to the direct biological effects of stress or other psychosocial factors, a number of health-protective or health-impairing behaviors can influence physical health and disease processes (see *Health Behaviors*; *Health Behavior: Psychosocial Theories*). Diet and obesity have been implicated in several major health problems, including cardiovascular disease, stroke, cancer, and adult onset diabetes. Atherosclerosis appears to be influenced by dietary behavior; the effects of obesity, substantial weight gain, or loss, and caloric restriction on atherosclerosis have been studied. The consumption of saturated fat and trans fatty acids contributes to low-density lipoprotein (LDL) cholesterol and triglyceride levels that promote atherosclerosis. Conversely, some dietary components may confer protection against the incidence or progression of cardiovascular disease. Studies suggest that consumption of antioxidant supplements, fruits and vegetables, and substances containing omega-3 fatty acids or other types of unsaturated fats may be beneficial in this regard. Dietary behavior appears to be important in both the pathophysiology and progression of cardiovascular disease.

There is a growing body of research that suggests that diet also affects the development of cancer. Epidemiological evidence suggests that diets that are rich in plant foods (i.e., fruits, vegetables, legumes, and grains) are associated with lower incidence of several types of cancers (reviewed in Potter and Steinmetz 1996). Laboratory and animal research suggests that constituents of plant foods, including vitamins and phytochemicals, have the capacity to inhibit the initiation, promotion, and/or progression of cancer, which may account for some of the epidemiological associations. On the other hand, a

number of dietary factors may promote the development of certain cancers, notably the consumption of total fat, saturated fat, and total calories.

Exercise appears to be important as a source of physical fitness, as a stress management strategy, and as a source of direct effects on health or disease (see *Physical Activity and Health*). A sedentary lifestyle is a risk factor for a number of diseases, including cardiovascular disease and cancer. Exercise may reduce the risk of cardiovascular disease by reducing stress or physiological reactivity to stress.

Tobacco use is perhaps the best-studied and most notorious health-impairing behavior (see *Smoking and Health*). It is the number one cause of preventable premature death and illness in industrialized nations; approximately half a billion people worldwide who are alive today will likely die from diseases related to tobacco use, which include cardiovascular disease, stroke, hypertension, pulmonary disease, kidney disease, and several different types of cancers. Psychological, behavioral, social, and biological factors are known to contribute to the initiation, maintenance, and relapse of cigarette smoking behavior. Modeling, social reinforcement, perceived cigarette availability, cost, and mental health all contribute to the initiation of smoking behavior, while maintenance of this behavior is largely due to the addictive properties of nicotine. The processes involved in nicotine addiction are similar to that of other drugs such as heroin and cocaine, and involve complex pharmacological, behavioral, and psychological components.

Sun exposure is another behavior that has major health consequences (see *Sun Exposure and Skin Cancer Prevention*). Basal-cell and squamous-cell skin cancers are associated with cumulative sun exposure, while melanomas are more closely associated with infrequent, periodic sunburn. Although the majority of skin cancers could have been prevented by regular sunscreen use or other management of sun exposure, better educational or prevention efforts are needed to bring about the lasting behavioral changes necessary to reduce skin cancer morbidity.

These and other behaviors represent potential contributors to illness and health, and they may interact to form an important level of analysis in the prediction of health and disease outcomes. Stress likely plays a role in many of these behaviors. Stress can also alter cognitive functioning, mood, problem-solving ability, social relationships, task performance, attention, and quality of life, which can in turn affect a number of health-related decisions or behaviors. Stress may thus act through a number of direct and indirect pathways to influence maintenance of health and susceptibility to disease.

2.3 Responses to Actual or Perceived Illness

A third general pathway through which behavior affects health is reflected in how people behave when they are ill, suspect they are ill, or learn that they are at risk for a serious illness. Obviously, early disease detection depends upon appropriate and timely human behavior. For cancer, the time at which the disease is detected can be the most crucial factor in determining treatment outcomes and disease prognosis (see *Cancer Screening*). Screening and self-detection practices are particularly relevant for breast cancer. Considerable research has been aimed at identifying and decreasing barriers to appropriate surveillance. Socioeconomic factors, quality of physician support, risk perceptions, health beliefs, and emotional responses to disease risk appear to influence screening behaviors.

Delays in seeking medical attention after symptoms are detected are also critical aspects of patient care and recovery (see *Symptom Awareness and Interpretation*). Similarly, adherence to medical recommendations and treatment regimens is another important way in which this category of behavior affects health (see *Patient Adherence to Health Care Regimens*). Nonadherence can disrupt treatment, affecting outcomes and interfering with future assessment and diagnosis. Unintentional nonadherence may occur when the patient fails to understand or remember the prescribed treatment. Cases of intentional nonadherence may reflect choices to pursue alternate forms of treatment, naïve theories of illness, or poor satisfaction (Morris and Schulz 1993). The quality of provider-patient interactions, supervision of adherence behaviors, and reminders to comply are among the factors that appear to influence adherence.

3. Behavioral Medicine Applications

The discovery of behavioral and biobehavioral pathways to disease has given rise to behavioral interventions for the treatment and prevention of medical illness and has increased the role of psychologists and behavioral specialists in medical settings. Behavioral medicine interventions generally are aimed at reducing stress or altering health-related behaviors.

3.1 Relaxation and Stress Reduction

One major emphasis of behavioral medicine interventions is the alleviation of stress and the promotion of relaxation. Techniques to achieve physical and mental relaxation have existed for thousands of years. Because of the broad impact of stress on disease pathogenesis, resistance to disease, and health behaviors, the reduction of stress may work in a variety of ways to prevent or delay disease onset or progression. Relaxation and stress reduction interventions consist of an array of techniques. Progressive muscle relaxation is one common method, and involves training the individual to recognize and reduce muscular tension, promoting cognitive and somatic relaxation. Biofeed-

back is the most technical of the relaxation methods, and consists of procedures that increase a person's ability to control physiological activities by providing information, or feedback, about these activities. Other procedures used to reduce stress or induce relaxation include autogenic training, diaphragmatic breathing training, physical exercise, guided imagery, meditation, and cognitive behavioral strategies such as coping skills and problem-solving training (see *Stress Management Programs*). Such interventions can result in the antithesis of the stress response; whereas stress elicits physiological changes that prepare the body for action, stress reduction procedures can elicit nervous and endocrine system changes that induce a state of relaxation, calmness, and decreased arousal and activity.

Stress reduction interventions have been useful in preventing or minimizing a variety of conditions (e.g., cardiovascular disease, arthritis, postsurgical outcomes, bowel disorders, ulcers, headache, and various pain conditions) (see also *Pain, Health Psychology of*). Stress reduction interventions have also been shown to reduce smoking relapse, and to reduce the number of physician visits among a healthy adult sample. The utility of stress reduction interventions is further established by studies that have documented reduced health care costs associated with these interventions (Friedman et al. 1996).

Psychosocial group interventions that provide education and social support may reduce stress or buffer the physical effects of stress (Helgeson and Cohen 1996). For example, individuals at risk for stress-related immune system changes have exhibited more favorable immune outcomes following various psychosocial interventions (e.g., Fawzy et al. 1990). The possibility that such interventions may favorably influence the course of cancer has been suggested by two provocative studies (Fawzy et al. 1993, Spiegel et al. 1989). Although the efficacy of behavioral medicine interventions for delaying cancer progression remains unknown and must await further investigation, the importance of behavioral medicine interventions as adjunct treatments for people with cancer, AIDS, or other serious life-threatening illnesses seems clear (see *Social Support and Recovery from Disease and Medical Procedures*). Providing information about coping, helping to build social support, and other behavioral or psychological strategies for stress and disease management are important in helping patients and family members adjust to the various challenges associated with these diseases.

3.2 Changing Unhealthy Behaviors

Given the deleterious health effects of behaviors such as smoking, physical inactivity, and poor diet, behavior change and modification of unhealthy behaviors is another important focus of behavioral medicine interventions. Behavioral medicine programs use a variety of techniques to promote behavior change. Self-monitoring is a straightforward technique that entails keeping a written record of specified health behaviors as a means of changing the frequency of the behavior. Stimulus control is a technique used to reduce the likelihood of an undesirable behavior by preventing or controlling antecedents of that behavior. Contingency management is a procedure that aims to change the frequency of a health behavior by carefully managing the consequences of that behavior, including rewards or punishments. These and other behaviorally-oriented procedures have been effective in increasing health promoting behaviors, such as exercise or adherence with medical treatment, and decreasing health impairing behaviors, such as smoking or alcohol consumption.

As noted, several different behaviors have been associated with the development or progression of major diseases such as cardiovascular disease and cancer. Consequently, behavioral medicine intervention programs often simultaneously target several different behavioral risk factors. For example, comprehensive behavioral change programs that promote changes in diet, exercise, smoking, social support, and stress management have achieved success in reducing modifiable risk factors for cardiovascular disease, slowing or reversing pathogenic signs of cardiovascular disease, and reducing cardiovascular disease mortality (e.g., Ornish et al. 1998).

Reduction of smoking is itself a frequent target of behavioral medicine intervention programs, because of the often-severe consequences of tobacco use. The same psychological and pharmacological properties that are responsible for the maintenance of smoking behavior are thought to be important considerations when attempting to achieve smoking cessation (see *Smoking Prevention and Smoking Cessation*). Unstated in such strategies and often overlooked are factors that bring people to the point that they want to quit smoking. Smoking cessation procedures are relatively ineffective when applied in situations where smokers do not intend to quit, and considerable effort should be directed towards motivation and contemplation of real cessation attempts. Achieving cessation is extremely difficult for many individuals; research suggests that up to 75 percent of smokers relapse within one year (Shumaker and Grunberg 1986). Because tobacco use is so difficult to stop once it has started, prevention efforts are often considered a more effective strategy to reduce smoking. However, a recent meta-analysis suggests that such programs have only a modest impact (Rooney and Murray 1996). Clearly, the reduction or prevention of tobacco use continues to represent an important challenge for behavioral medicine.

Eating behavior is another consequential health behavior that is addressed by behavioral medicine programs that target specific food consumption or

overall weight gain or loss. Such programs have been successful in reducing consumption of dietary carcinogens and increasing consumption of cancer-protective foods. Dietary change recommendations for decreasing cancer risk are more likely to be adhered to when people believe that what they eat does actually influence the development of cancer, when they understand what recommendations to follow, and when these changes are supported by others.

4. Summary and Conclusions

Behavioral medicine is an interdisciplinary scientific and medical specialty that is based on holistic notions of mind–body interactions and that seeks to prevent, treat, and overcome disease and disability by minimizing biobehavioral conditions that threaten good health. It assumes that the brain, together with the various sensory systems of the body, integrates and regulates physical and psychological events and that many of these events are so intertwined that separation of their elemental properties is impossible. Emotions, for example, consist of both thoughts and visceral events; fear is associated with feelings of fear as well as with changes in heart rate and other autonomic functions. Historically, theorists have argued about whether the cognitive or somatic components of emotions come first. For behavioral medicine, such distinctions are not particularly important or productive; consistent with its holistic approach, behavioral medicine views emotional states as simultaneous changes associated with learned or innate action patterns, and seeks to understand the health-related physiological and behavioral aspects of such experiences. Behavioral medicine is a true hybrid of many life and behavioral sciences, bringing together several disciplines and blending their approaches into something entirely new. The potential of this new science to help solve major public health problems is potentially great, and the early success of interventions derived from behavioral medicine is consistent with the recognition that many of the most pressing modern health problems are largely behavioral in nature, born of genetic and lifestyle origins rather than of microorganisms that invade the human body.

See also: Dental Health: Psychosocial Aspects; Diabetes: Psychosocial Aspects; Disability: Psychological and Social Aspects; Gastrointestinal Diseases: Psychosocial Aspects; Gynecological Health: Psychosocial Aspects; Health Education and Health Promotion; Health Psychology; Health: Self-regulation; Hypertension: Psychosocial Aspects; Injuries and Accidents: Psychosocial Aspects; Psychobiology of Stress and Early Child Development; Respiratory Disorders: Psychosocial Aspects; Rheumatoid Arthritis: Psychosocial Aspects; Self-monitoring, Psychology of; Self-regulation in Adulthood; Stress and Health Research; Stress, Neural Basis of; Stress: Psychological Perspectives; Weight Regulation: Psychological Aspects

Bibliography

Ader R, Felten D L, Cohen N 1991 *Psychoneuroimmunology*, 2nd edn. Academic Press, San Diego

Alexander F 1950 *Psychosomatic Medicine*, 1st edn. Norton, New York

Andersen B L, Kiecolt-Glaser J K, Glaser R 1994 A biobehavioral model of cancer stress and disease course. *American Psychologist* **49**: 389–404

Berkman L F, Syme S L 1979 Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda County residents. *American Journal of Epidemiology* **109**: 186–204

Cohen N, Ader R 1988 Immunomodulation by classical conditioning. *Advances in Biochemical Psychopharmacology* **44**: 199–202

Cohen S, Herbert T B 1996 Health psychology: Psychological factors and physical disease from the perspective of human psychoneuroimmunology. *Annual Review of Psychology* **47**: 113–42

Cohen S, Tyrell D A, Smith A P 1991 Psychological stress and susceptibility to the common cold. *New England Journal of Medicine* **325**: 606–12

Fawzy F I, Fawzy N W, Hyun C S, Elashoff R, Guthrie D, Fahey J L, Morton D L 1993 Malignant melanoma: effects of an early structured psychiatric intervention, coping, and affective state on recurrence and survival 6 years later. *Archives of General Psychiatry* **50**: 681–89

Fawzy F I, Kemeny M E, Fawzy N W, Elashoff R, Morton D L, Cousins N, Fahey J L 1990 A structured psychiatric intervention for cancer patients. II Changes over time in immunological measures. *Archives of General Psychiatry* **47**: 729–35

Friedman R, Sobel D, Myers P, Caudill M, Benson H 1996 Behavioral medicine, clinical health psychology, and cost offset. *Health Psychology* **14**: 509–18

Graham D 1972 Psychosomatic medicine. In: Greenfield N S, Sternbach R A (eds.) *Handbook of Psychophysiology*. Holt, Rinehart, & Winston, New York

Helgeson V S, Cohen S 1996 Social support and adjustment to cancer: Reconciling descriptive, correlational, and intervention research. *Health Psychology* **15**: 135–48

Lazarus R S 1966 *Psychological Stress and the Coping Process*. McGraw-Hill, New York

Leserman J, Petitto J M, Perkins D O, Folds J D, Golden R N, Evans D L 1997 Severe stress, depressive symptoms, and changes in lymphocyte subsets in human immunodeficiency virus-infected men: A 2-year follow-up study. *Archives of General Psychiatry* **54**: 279–85

Manuck S B, Marsland A L, Kaplan J R, Williams J K 1995 The pathogenicity of behavior and its neuroendocrine mediation: An example from coronary artery disease. *Psychosomatic Medicine* **57**: 275–83

Mason J W 1975 A historical view of the stress field. *Journal of Human Stress* **1**: 22–36

Miller N 1978 Biofeedback and visceral learning. *Annual Review of Psychology* **29**: 373–404

Morris L S, Schulz R M 1993 Medication compliance: the patient's perspective. *Clinical Therapeutics* **15**: 593–606

Ornish D, Scherwitz L W, Billings J H, Brown S E, Gould K L, Merritt T A, Sparler S, Armstrong W T, Ports T A, Kirkeeide R L, Hogeboom C, Brand R J 1998 Intensive lifestyle changes for reversal of coronary. *Journal of the American Medical association* **280**: 2001–7

Pavlov I P 1927 *Conditioned Reflexes* (trans. Anrep G V). Oxford University Press, London

Potter J, Steinmetz K 1996 Vegetables, fruit and phytoestrogens as preventive agents. In: Stewart B W, McGregor D, Kleihues P (eds.) *Principles of Chemoprevention*. International Agency for Research on Cancer, Lyon, France

Rooney B L, Murray D M 1996 A meta-analysis of smoking prevention programs after adjustment for errors in the unit of analysis. *Health Education Quarterly* **23**: 48–64

Shumaker S, Grunberg N 1986 Introduction to proceedings of the National Working Conference on Smoking Relapse. *Health Psychologist* **5**: 1–2

Spiegel D, Bloom J R, Kraemer H C, Gottheil E 1989 Effect of psychosocial treatment on survival of patients with metastatic breast cancer. *Lancet*. **2**(8668): 888–91

Wolff H G 1953 *Stress and Disease*. Thomas, Springfield, IL

B. N. Henderson and A. Baum

Behavioral Neuroscience

Behavioral and cognitive neuroscience are two sides of the same coin. The field as a whole is concerned with the neuronal and biological bases of behavior and experience, that is to say, psychology. Indeed, the distinction between 'behavioral' and 'cognitive' is rather arbitrary. Operationally, behavioral neuroscience might be defined by the kinds of papers published in the journal *Behavioral Neuroscience* and like journals. The emphasis is on biological studies of such basic phenomena as sensation and perception, motivation, learning, and memory, phenomena that occur similarly in humans and other animals. Most studies are done on infrahuman animals, including non-human primates. By the same token, most of the papers in the journal *Cognitive Neuroscience* and like journals are on humans, with imaging studies predominating. But many behavioral neuroscientists study basic psychological processes in humans, and many who consider themselves cognitive neuroscientists study such processes in animals.

The field of neuroscience is relatively recent as a 'unified' discipline; the first meeting of the Society for Neuroscience was held in 1971. The goal was to merge the disparate fields concerned with the study of the nervous system, e.g., neuroanatomy, neurophysiology, neurochemistry, neuropharmacology, neuroendocrinology, physiological psychology, clinical neurology, etc., into one interdisciplinary field.

There was a very long tradition in psychology of the study of animal behavior and its biological bases, initially termed comparative psychology. For many years the flagship journal devoted to comparative behavior and its biological bases was the *Journal of Comparative and Physiological Psychology*, published by the American Psychological Association. But increasingly the fields of comparative studies of behavior and physiological analysis of behavioral phenomena went separate ways. Consequently, in 1983 the journal was divided in two separate journals. Following is a portion of the editorial justifying this action by the then editor:

> The separation of the *Journal of Comparative and Physiological Psychology* into two companion journals, *Behavioral Neuroscience* and the *Journal of Comparative Psychology*, acknowledges the current state of the sciences concerned with biology and behavior. Behavioral neuroscience is the broader contemporary development of physiological psychology. In all animals above the level of the sponge, the nervous system is the organ of behavior. All biological and behavioral variables that exert any influence on behavior must do so by acting on and through the nervous system.
>
> Traditionally, physiological psychology emphasized certain approaches and techniques, particularly lesions, electrical stimulation, and electrophysiological recording. Study of the biological bases of behavior is now much broader and includes genetic factors, hormonal influences, neurotransmitter and chemical factors, neuroanatomical substrates, effects of drugs, developmental processes, and environmental factors, in addition to more traditional approaches. All these variables act ultimately through the nervous system. All these areas of investigation are entirely appropriate for *Behavioral Neuroscience* – all studies in which biological variables are manipulated or measured and relate to behavior, either directly or by implication.
>
> The contemporary meaning of the term 'behavioral neuroscience' is almost as broad as 'behavior' itself. Behavioral neuroscience is the field concerned with the biological substrates of behavior. Judging by its current rate of development, behavioral neuroscience could well become a dominant field of science in the future (Thompson 1983, p. 3).

By the time this occurred, the term 'behavioral neuroscience' was increasingly used to describe that aspect of the interdisciplinary field of neuroscience concerned with the biological basis of behavior. By the same token, 'cognitive neuroscience' was adapted more recently to emphasize that aspect of neuroscience concerned with the more complex phenomena of the 'mind.'

1. *Historical Developments*

The development of the two aspects of the study of 'brain, mind and behavior' is best viewed historically. As noted above 'behavioral neuroscience,' the subject of this article, is the modern equivalent of 'physiological psychology.' Since Wilhem Wundt, the founder of the modern field of psychology in the 1870s and himself a physiologist, titled his epoch making text *Grundzüge der Physiologische Psychologie* (1874), we

begin with his era. Wundt insisted on the experimental approach to all questions of psychology: the challenge was to find experimental methods to understand psychological processes. The most appropriate methods, e.g., measures of brain activity, did not of course exist in his day. Indeed, the only methods available were simple measurements of behavior like reaction time, and verbal descriptions of experience by introspection, both of which were used in Wundt's laboratory.

Wundt's career illustrates well the roots of the developing field of physiological psychology. After a year of study at Heidelberg in basic medical sciences, he moved to Berlin and worked for two years with Johannes Müller, a founder of modern, experimental physiology. Müller, incidentally, was a vitalist and believed that it would never be possible to measure the speed of the nerve impulse; his own student Helmoltz proved him wrong. Wundt then took his doctorate in medicine and was appointed Dozent in physiology at Heidelberg. Helmholtz then joined the same institute and they were colleagues, but not collaborators, for a period of 13 years. Wundt later moved to Leipzig and established the first formal psychological laboratory, in 1879.

Wundt's text and research at Leipzig represented a brilliant experimental physiologist grappling with problems of the mind and mental events, applying the experimental methods of science as best he could. In fact, the bulk of the research done in the Leipzig laboratory was on sensation and perception, and on reaction time.

Another laboratory of physiological psychology was founded at about the same time, that of William James. James went to Harvard and Harvard Medical School and studied with the great naturalist, Louis Agassiz. James was then, after a year of study abroad, appointed an instructor in physiology at Harvard College in 1872. He actually established an informal physiological psychology laboratory there in 1877 and taught a graduate course on the relations between physiology and psychology.

In sum, the two founders of modern psychology, Wundt in Germany and James in America, were both physiologists by training. More than a third of William James' extraordinary and influential text, *Principles of Psychology* (1890), was devoted to the nervous system. However, this does not necessarily mean that Wundt and James attempted to reduce psychology to physiology; rather they proposed that the subject matter of psychology should be studied scientifically, as in physiology. This view could equally apply to behavioral and cognitive neuroscience as it exists today.

There were powerful intellectual pressures moving bright young physiologically inclined psychologists at the turn of the century. The theory of evolution was a major force. The field of physiology, in particular neurophysiology, in the work of Sir Charles Sherrington and then Lord Adrian, together with clinical neurology and neuroanatomy, were vigorous and exciting fields at the beginning of the twentieth century.

The development of instrumentation was still another very important factor. The human EEG was rediscovered in 1929 (Berger) and the method applied to animal research in the 1930s. At the turn of the century the major experimental techniques for the study of brain function were ablation and electrical stimulation. Neuroanatomy was in its descriptive phase; the monumental work of Ramon y Cajal was published over a period of several decades beginning near the end of the nineteenth century. Neurochemistry was in its purely descriptive phase.

Merely because techniques and basic knowledge were limited does not mean that the field was quiescent. On the contrary, there was ferment over a number of basic issues ranging from the mind-brain problem to localization of function and the nature of neuronal interactions. Add to this the ferment developing in psychology as John Watson took on the older establishment and behaviorism began to dominate. This is the background from which modern physiological psychology developed.

2. *Lashley and the Engram*

Karl S. Lashley is the most important figure in the development of physiological psychology in America (see *Lashley, Karl Spencer (1890–1958)*). He obtained his Ph.D. in zoology at Johns Hopkins University. At Hopkins he studied with John Watson and was heavily influenced by Watson's developing notions of behaviorism. While there he also worked with Shepherd Franz at a government hospital in Washington; they published a paper together in 1917 on the effects of cortical lesions on learning and retention in the rat. Lashley then held teaching and research positions at the University of Minnesota (1917–26), the University of Chicago (1929–35) and then at Harvard from 1935 until his death in 1958. During the Harvard years he spent much of his time at the Yerkes Primate Laboratory in Orange Park, Florida.

Lashley devoted many years to an analysis of brain mechanisms of learning, using the lesion-behavior method, which he developed and elaborated from the work with Franz. During this period, Lashley's theoretical view of learning was heavily influenced by two congruent ideas: localization of function in neurology and behaviorism in psychology.

Localization of function was the major intellectual issue concerning brain organization at the turn of the century. An extreme form of localization was popular at the beginning of the nineteenth century with Gall and phrenology. The neurologist Flourens moved away from that position by arguing for no specific localization within the cerebrum. However, Broca's discovery of a speech center in 1861 began to move the

pendulum back. The critical and classic study on localization was that of Fritsch and Hitzig (1870). They stimulated the cerebral cortex electrically and defined the motor area: not only was it localized to a particular region of the cortex but there were specific organization and localization within it.

In Watson's behaviorism, the learning of a particular response was held to be the formation of a particular set of connections, a series set. Consequently, Lashley argued, it should be possible to localize the place in the cerebral cortex where that learned change in brain organization was stored: the engram. (It was believed at the time that learning occurred in the cerebral cortex). Thus, behaviorism and localization of function were beautifully consistent: they supported the notion of an elaborate and complex switchboard where specific and localized changes occurred when specific habits were learned.

Lashley set about systematically to find these learning locations—the engrams—in a series of studies culminating in his 1929 monograph *Brain Mechanisms of Intelligence*. In this study he used mazes differing in difficulty, and made lesions of varying sizes in all different regions of the cerebral cortex of the rat. The results of this study profoundly altered Lashley's view of the brain organization and had an extraordinary impact on the young field of physiological psychology. The locus of the lesion is unimportant; the size is critically important, particularly for the more difficult mazes. These findings led to Lashley's two theoretical notions of equipotentiality and mass action, i.e., all areas of the cerebral cortex are equally important (at least in maze learning), what is critical is the amount removed.

Lashley's interpretations stirred vigorous debate in the field. Walter Hunter, an important figure in physiological-experimental psychology at Brown University who developed the delayed response task, argued that in fact the rat was using a variety of sensory cues: as more of the sensory regions of cortex were destroyed, fewer and fewer cues became available. Lashley and his associates countered by showing that removing the eyes has much less effect on maze learning than removing the visual area of the cortex. Others argued that Lashley removed more than the visual cortex. Out of this came the long series of lesion-behavior studies analyzing behavioral 'functions' of the cerebral cortex. Initially, studies focused on sensory areas of the cerebral cortex. Beginning in the 1940s several laboratories, including Lashley's and those of Harry Harlow at the University of Wisconsin and Karl Pribram at Yale, took up the search for the more complex functions of association cortex using monkeys. Lashley's interests were not limited to brain mechanisms of learning. His influence is felt strongly through the many eminent contemporary behavioral neuroscientists who worked or had contact with him; Austin Riesen, Donald O. Hebb, Roger W. Sperry and Karl Pribram are examples. Hebb's book *The Organization of Behavior* (1949) is a landmark, as is Sperry's work on consciousness.

We now treat the development of the major areas of research that characterize modern behavioral neuroscience, namely neural and biological substrates of learning and memory, motivated behaviors, and sensory processes.

3. Learning and Memory

3.1 Pavlov

Lashley's pessimistic conclusions in his 1929 monograph put a real damper on the field concerned with brain substrates of memory. But there were other major traditions developing. Perhaps the most important of these was the influence of Pavlov. His writings were not readily available to Western scientists, particularly Americans, until the publication of the English translation of his monumental work *Conditioned Reflexes* in 1927. It is probably fair to say this is the most important single book ever published in the field of behavioral neuroscience. Pavlov developed a vast and coherent body of empirical results characterizing the phenomena of conditioned responses, what he termed 'psychic reflexes.' He argued that the mind could be fully understood by analysis of the higher order learned reflexes and their brain substrates.

W. Horsley Gantt, an American physician, worked with Pavlov for several years and then established a Pavlovian laboratory at Johns Hopkins. He trained several young psychologists, e.g., Roger Loucks and Wolf Brogden, who became very influential in the field. Perhaps the most important modern behavioral analyses of Pavlovian conditioning are the work of Robert Rescorla and Allan Wagner.

Although Pavlov worked with salivary secretion, most studies of classical conditioning in the West tended to utilize skeletal muscle responses, à la Bechterev. Particularly productive have been Pavlovian conditioning of discrete skeletal reflexes (e.g., the eyeblink response) characterized behaviorally by Isadore Gormezano and Allan Wagner and analyzed neuronally by Richard Thompson and his many students; and fear conditioning, first characterized behaviorally by Neal Miller and analyzed neuronally by several investigators (see Sect. 3.6).

3.2 Consolidation

Carl Duncan's discovery of the effects of electroconvulsive shock on retention of simple habits in the rat, in 1949, began the modern field of memory consolidation. Hebb and Gerard were quick to point out the implication of two memory processes, one

transient and fragile and the other more permanent and impervious. James McGaugh and his associates have done the classic work on the psychobiology of memory consolidation. McGaugh and his colleagues (see McGaugh 2000) demonstrated memory facilitation with drugs and showed that these effects were direct and not due to possible reinforcement effects of the drugs (and similarly for ECS impairment).

Chemical approaches to learning and memory are recent. The possibility that protein molecules and RNA might serve to code memory was suggested some years ago by such pioneers as Gerard and Halstead. The RNA hypothesis was taken up by Hyden and associates in Sweden and by several groups in America. An unfortunate byproduct of this approach was the 'transfer of memory' by RNA. These experiments, done by investigators who shall remain nameless, in the end could not be replicated.

At the same time, several very productive lines of investigation of neurochemical and neuroanatomical substrates of learning were developing. In 1953, Krech and Rosenzweig began a collaborative study on relationships between brain chemistry and behavior. Their initial studies concerned brain levels of AChE in relation to the hypothesis behavior and included analysis of strain differences (see, e.g., Krech et al. 1954). More recently, they and their collaborators, Bennett and Diamond, discovered the striking differences in the brains of rats raised in 'rich' vs. 'poor' environments. William Greenough, at the University of Illinois, replicated and extended this work to demonstrate dramatic morphological changes in the structures of synapses and neurons as a result of experience.

3.3 Model Systems

The use of model biological systems has been an important tradition in the study of neural mechanisms of learning. This approach has been particularly successful in the analysis of habituation, itself a very simple form or model of learning (see *Habituation*). Sherrington did important work on flexion reflex 'fatigue' in the spinal animal at the turn of the century. Sharpless and Jasper (1956) established habituation as an important process in EEG activity. Modern Russian influences have been important in this field: the key studies of Evgeny Sokolov, first on habituation of the orienting response in humans and more recently on mechanisms of habituation of responses in the simplified nervous system of the snail (see Sokolov 1963).

The defining properties of habituation were clearly established by Thompson and Spencer in 1966, and the analysis of mechanisms began. Several laboratories using different preparations (Aplysia withdrawal reflex: Kandel and associates, see Kandel 1976; vertebrate spinal reflexes: Thompson and Spencer; crayfish tail flip escape: Krasne, Kennedy, see Krasne and Bryan 1973) have all arrived at the same underlying synaptic mechanism: a decrease in the probability of transmitter release from presynaptic terminals of the habituating pathway. Habituation is thus a very satisfying field: agreement ranges from defining behavioral properties to synaptic mechanisms. In a sense, the problem has been solved. Habituation also provides a most successful example of the use of the model biological systems approach to analysis of neural mechanisms of behavioral plasticity (see Groves and Thompson 1970).

3.4 Hippocampus

Study of the role of the hippocampus in learning and memory has become an entire field. It began with Brenda Milner's extraordinary studies of patient H. M., who following bilateral temporal lobectomy lost the ability to remember his own life experiences, dating to a year or so before his operation and continuing to the present (see Milner 1966). However he is able to learn motor and even complex sensory-motor skills normally. A large animal literature grew in unsuccessful attempts to develop animal models of H. M.'s syndrome. In 1978 Mortimer Mishkin, at NIMH, developed a striking analogue of H. M's syndrome in monkeys by removing the hippocampus and amygdala bilaterally. The animals still had normal visual pattern discriminations but were markedly impaired in recognition memory (delayed non matching to sample task). From recent work, e.g., by Mishkin, Larry Squire, Stuart Zola and others, it is clear that the most extreme form of the syndrome in both monkeys and humans involves the hippocampus and related cortical areas but not the amygdala. Study of hippocampal function, particularly in humans in the context of declarative or 'what' memory is a major topic in cognitive neuroscience today.

Another facet of hippocampal study in the context of behavioral neuroscience is long-term potentiation (LTP), discovered by Bliss and Lomo in 1970. Brief tetanic stimulation of monosynaptic inputs to the hippocampus causes a profound increase in synaptic excitability that can persist for hours or days. Many view it as a leading candidate for a mechanism of memory storage, although direct evidence is still lacking.

Yet another major impetus to study of the hippocampus is the remarkable discovery of 'place cells' by John O'Keefe (1976). When recording from single neurons in the hippocampus of the behaving rat, a given neuron may respond only when the animal is in a particular place in the environment (i.e., in a box or maze) reliably and repeatedly. There is great interest now in the possibility that LTP may be the mechanism forming place cells. A number of laboratories are

making use of genetically altered mice to test this possibility.

3.5 Cerebellum

Yet another brain structure that has become a minor industry in brain substrates of learning and memory is the cerebellum. Masao Ito and his associates in Tokyo discovered the phenomenon of long-term depression (LTD) in cerebellar cortex in 1982. Repeated conjunctive stimulation of the two major projection systems to the cerebellum, mossy-parallel fibers and climbing fibers, yields long-lasting decreases in excitability of parallel fiber-Purkinje neuron synapses. Ito developed considerable evidence that this process underlies plasticity of the vestibulo-ocular reflex. Thompson and associates and others showed that the memory traces for classical conditioning of discrete responses (e.g., eyeblink conditioning) are stored in the cerebellum (Thompson 1986). LTD appears to be a critical process of such memory formation in the cerebellar cortex. The output of the cerebellum to other brain structures is from the cerebellar nuclei and Purkinje neurons inhibit nuclear neurons; hence cortical LTD would decrease inhibition on nuclear neurons.

3.6 Amygdala

Still another structure to assume a major role in the behavioral neuroscience of learning and memory is the amygdala. It was known for some time that electrical stimulation of the amygdala could elicit autonomic and emotional responses. An elegant literature developed by, e.g., Michael Davis, Joseph Le Doux, Michael Fanselow and others demonstrated clearly that the amygdala was the key structure in conditioned fear (see Davis 1992). Indeed, this literature is at least consistent with the view that the essential memory traces for classical fear conditioning are established in the amygdala.

The amygdala is also critical for instrumental learning of fear. James McGaugh and his associates demonstrated that for both passive and active avoidance learning (animals must either not respond, or respond quickly, to avoid shock) amygdala lesions made immediately after training abolished the learned fear (see McGaugh 1989). Surprisingly, if these same lesions were made a week after training, learned fear was not abolished, consistent with a process of consolidation. The apparent differences in the role of the amygdala in aspects of learned fear is an interesting issue.

4. Motivation and Learning

Physiological and neural mechanisms of motivation and emotion have been a particular province of physiological psychology since the 1930s. In recent years, the fields of 'motivation' and 'emotion' have tended to go separate ways. However, motivation and emotion have a common historical origin. In the seventeenth and eighteenth centuries, instinct doctrine served as the explanation for why organisms were driven to behave (at least infrahuman organisms without souls). Darwin's emphasis on the role of adaptive behavior in evolutionary survival resulted in the extension of instinct doctrine to human behavior. Watson rebelled violently against the notion of instinct and rejected it out of hand, together with biological mechanisms of motivation.

In 1928, Bard showed that the hypothalamus was responsible for 'sham rage.' In the 1930s, Ranson and his associates at Northwestern, particularly Magoun and Ingram, published a classic series of papers on the hypothalamus and its role in emotional behavior. Somewhat later, Hess and his collaborators in Switzerland were studying effects of stimulating the hypothalamus in freely moving cats.

4.1 Motivated Behaviors

It is against this backdrop that the modern field of psychobiology of motivation developed (the term 'motivated behaviors' is now preferred to 'motivation'). Karl Lashley was again the prime mover. His paper on *The experimental analysis of instinctive behavior* in 1938 was the key. In it, he argued that motivated behavior varies, and is not simply a chain of instinctive or reflex acts, that it is not dependent upon any one stimulus, and that it involves central state. His conclusions, that 'physiologically, all drives are no more than expressions of the activity of specific mechanisms' and that hormones 'activate some central mechanism which maintains excitability and activity' have a very modern ring.

A most important series of studies was originated in a paper by Klüver and Bucy on *Psychic blindness and other symptoms following bilateral temporal lobectomy in rhesus monkeys* in 1937. This came to be known as the Klüver–Bucy syndrome. The animals exhibited marked changes in motivation and aggressive behavior. The Klüver–Bucy syndrome, together with the suggestion of Papez (1937) that the structures of the limbic system formed a basic circuit for motivation and emotion, led to the extensive field concerned with the behavioral functions of the 'limbic system' (see Isaacson 1982).

Lashley's general notion of a central mechanism that maintains activity was developed by Beach in an important series of papers in the 1940s into a central excitatory mechanism and ultimately a central theory of drive. This view was given a solid physiological basis by Donald Lindsley from the work he and Magoun and associates were doing on the ascending reticular activating system. Lindsley sketched his activation theory of emotion in his important chapter in the *Stevens Handbook* (1951). D. O. Hebb and Elliot

Stellar pulled all these threads together into a general and influential central theory of motivation.

In his empirical work, Beach focused on brain mechanisms of sexual behavior. As the study of sexual behavior developed, hormonal factors came to the fore and the modern field of hormones and behavior developed. Beach played a critical role in the development of this field. Even within the field of hormones and behavior, several fields have developed. Sexual behavior has become a field unto itself. Another important field today is the general area of stress. The endocrinologist Hans Selye was an important intellectual influence. Kurt Richter is the pioneering figure in this field in physiological psychology. Richter took his Ph.D. from Johns Hopkins in 1921 and established his laboratory there. He worked in several areas, including the role of the adrenal gland in stress. The modern field of stress focuses on hormonal-behavioral interactions, e.g., the work of Seymore Levine.

Neal Miller represents a uniquely important tradition in behavioral neuroscience. From the beginning of his career, Miller was interested in physiological mechanisms of both motivation and learning. He took his Ph.D. from Yale in 1935 and stayed on at Yale for many years, with a year out in 1936 at the Vienna Psychoanalytic Institute. Miller was a pioneer in early studies of punishing and rewarding brain stimulation and their roles in learning. He was the first to demonstrate conditioned fear. In more recent years, his work focused on mechanisms of instrumental conditioning of autonomic responses. The impact of his work is much wider than physiological psychology, influencing psychiatry and clinical medicine as well.

4.1.1 Electrical self-stimulation of the brain. James Olds, whose untimely death in 1976 cut short an extraordinary career, made perhaps the most important discovery yet made in the field of behavioral neuroscience: rewarding electrical self-stimulation of the brain. He got his Ph.D. at Harvard and worked with Richard Solomon. Solomon, incidentally, although primarily a behavioral student of learning, had great impact on physiological psychology through his theoretical-experimental analysis of hypothetical central factors in learning. As a graduate student Olds read and was much influenced by Hebb's *The Organization of Behavior* and obtained a postdoctoral with Hebb at McGill in 1953. He began work with Peter Milner and they discovered electrical self-stimulation of the brain. The recent history of this field is well known to everyone. The brain reward system now appears closely similar if not coextensive with the 'addiction' system in the brain: the medial forebrain bundle dopamine projection system to the nucleus accumbens, prefrontal cortex, and other structures.

5. *Sensory Processes*

A major topic area in physiological psychology has been the study of sensory processes: sensation and perception. Indeed, this is perhaps the original field of psychology, dating back at least to Newton. Although specialization has resulted in separation between the fields of psychophysics and sensory physiology, in the sense that few individual scientists do research in both fields, they remain closely interlocked. From the beginning, explorations of sensory and perceptual phenomena have always involved hypothetical physiological mechanisms, e.g., the Young–Helmholtz 3 receptor theory of color vision and the Hering opponent process theory (see Hurvich and Jameson 1974).

This has been a field of extraordinary progress in the twentieth century. Techniques have been critically important. Early in the century there were really no tools, other than rather crude anatomical methods, for analyzing the organization of sensory systems in the brain. The pioneering studies of Adrian (1940) in England and Marshall, Woolsey and Bard (1941) at Johns Hopkins were the first to record electrical evoked potentials from the somatic sensory cortex in response to tactile stimulation. Woolsey and his associates developed the detailed methodology for evoked potential mapping of the cerebral cortex. In an extraordinary series of studies, Woolsey and his colleagues determined the localization and organization of the somatic sensory areas, the visual areas and the auditory areas of the cerebral cortex in a comparative series of mammals. They initially defined two areas (I and II) for each sensory field.

5.1 *Organization of the Sensory Systems*

In the 1940s and 1950s the evoked potential method was used to analyze the organization of sensory systems at all levels from the first order neurons to the cerebral cortex. The principle that emerged was strikingly clear and simple; in every sensory system the nervous system maintained a receptotopic map or projection at all levels from receptors to cerebral cortex: skin surface, retina, basilar membrane. The same organization held for the second sensory areas. The receptor maps in the brain were not one-to-one, rather they reflected the functional organization of each system: fingers, lips, and tongue areas were much enlarged in primate somatic cortex, half the primary visual cortex represented the forea, and so on.

The evoked potential method was very well suited to analysis of the overall organization of sensory systems in the brain. However, it could reveal nothing about what the individual neurons were doing. This had to wait development of the microelectrode. Indeed, the microelectrode has been the key to analysis of the fine-

grained organization and feature detector properties of sensory neurons. Metal microelectrodes were developed in the early 1950s: Davies at Hopkins developed the platinum-iridium glass coated microelectrode, Hubel and Wiesel at Harvard developed the tungsten microelectrode, and the search for feature detectors was on. The pioneering studies were those of Mountcastle and associates at Hopkins on the organization of the somatic-sensory system, those of Hubel and Wiesel at Harvard on visual system and Rose, Hind, Woolsey and associates at Wisconsin on the auditory system. Thanks to the microelectrode and to modern pathway tracing techniques, we now know that each sensory modality is represented multiply in the cerebral cortex.

Meanwhile, pioneering work was being done on receptors. Hartline and Ratliff analyzed receptor responses in a simple visual system, Limulus (horseshoe crab), and discovered lateral inhibition. Von Bekesy discovered the standing wave patterns in the cochlea (working in the psychology department at Harvard). Dark adaptation was explained in biochemical terms by Wald at Harvard. The role of eye movements in visual perception was elucidated by Riggs and associates at Brown.

In the recent past, progress in analysis of sensory processes has been quite remarkable. The ability to measure, with the microelectrode, the activity of single receptors or sensory neurons precisely, with tight stimulus control, has been matched by the great precision of psychophysical methods and results in humans and animals.

Analysis of the physiological properties of neurons in sensory systems, and their psychophysical concomitants, has become highly productive, sophisticated, and elegant. It is without question the most advanced field in behavioral neuroscience. Many would claim it as a separate field, or more precisely, a separate set of fields, i.e., the hearing sciences and the vision sciences. This, by the way, seems characteristic of behavioral neuroscience. The great questions are first raised within the field. As techniques develop and answers begin to appear, separate fields may be created.

6. Conclusion

Several earlier aspects of behavioral neuroscience have developed into separate fields in their own right. We noted the vision sciences and the hearing sciences (see Sect. 5.1). Study of physiological responses in humans has become the field of psychophysiology, with its own society and journal. Another example is the study of human brain damage, particularly in terms of higher functions. This field is now termed neuropsychology and is a part of modern cognitive neuroscience. Yet another example is the field of behavioral genetics, which grew from early studies of selective breeding of behavioral properties. Thanks to the extraordinary developments in genetics and molecular biology, behavioral genetics has become a component of neurogenetics.

The field of behavioral neuroscience (physiological psychology) has a long history of asking fundamental questions about biology, behavior, and experience and developing approaches to deal with these questions. The consequence is often that entire new fields of scientific endeavor develop and became distinct subspecialties in their own right. Perhaps in a more general sense the same can be said of psychology.

See also: Amygdala (Amygdaloid Complex); Cerebellum: Cognitive Functions; Cognitive Neuroscience; Comparative Neuroscience; Emotion, Neural Basis of; Hebb, Donald Olding (1904–85); Hippocampus and Related Structures; Lashley, Karl Spencer (1890–1958); Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Mind—Body Dualism; Motivation, Neural Basis of; Pavlov, Ivan Petrovich (1849–1936); Psychophysiology; Visual Perception, Neural Basis of; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Adrian E D 1940 Double representation of the feet in the sensory cortex of the cat. *Journal of Physiology* **98**: 16

Bard P 1928 A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system. *American Journal of Physiology* **84**: 490–515

Berger H 1929 Über das Elektroenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten* **87**: 527–70

Bliss T V P, Lomo T 1970 Plasticity in a monosynaptic cortical pathway. *Journal of Physiology* **207**: 61

Broca P 1861 Perti de parole, remollissement chronique et destruction partielle du lobe antérieur gauche du cerveau. *Bulletin de la Société anthropologique* **2**: 235–8

Davis M 1992 The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience* **15**: 353–75

Duncan C P 1949 The retroactive effect of electroshock on learning. *Journal of Comparative and Physiological Psychology* **42**: 32–44

Franz S I, Lashley K S 1917 The retention of habits in the rat after destruction of the frontal portion of the cerebrum. *Psychobiology* **1**: 3–18

Fritsch G, Hitzig E 1870 Über die elektrische Erregbarkeit des Grosshirns. *Archiv für Anatomie, Physiologie und Wissenschaftliche Medizin* **37**: 300–32

Groves P M, Thompson R F 1970 Habituation: A dual-process theory. *Psychological Review* **77**: 419–50

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Hurvich L M, Jameson D 1974 Opponent processes as a model of neural organization. *American Psychologist* **29**: 88–102

Isaacson R L 1982 *The Limbic System*. 2nd edn. Plenum Press, New York

Ito M, Sakurai M, Tongroach P 1982 Climbing fiber induced depression of both mossy fiber responsiveness and glutamate sensitivity of cerebellar Purkinje cells. *Journal of Physiology* **324**: 113–34

James W 1890 *Principles of Psychology*. Vols. 1, 2. Holt, New York

Kandel E R 1976 *The Cellular Basis of Behavior: An Introduction to Behavioral Neurobiology*. W. H. Freeman, San Francisco
Klüver H, Bucy P C 1937 'Psychic blindness' and other symptoms following bilateral temporal lobectomy in rhesus monkeys. *American Journal of Physiology* **119**: 352–3
Krasne F B, Bryan J S 1973 Habituation: Regulation through presynaptic inhibition. *Science* **182**: 590–2
Krech D, Rosenzweig M R, Bennett E L, Krueckel B 1954 Enzyme concentration in brain and adjustive behavior patterns. *Science* **120**: 994–6
Lashley K S 1929 *Brain Mechanisms and Intelligence*. University of Chicago Press, Chicago
Lashley K 1938 The experimental analysis of instinctive behavior. *Psychological Review* **45**: 445–71
Lindsley D B 1951 Emotion. In: Stevens S S (ed.) *Handbook of Experimental Psychology*. Wiley, New York
Marshall W H, Woolsey C N, Bard P 1941 Observations on cortical sensory mechanisms of cat and monkey. *Journal of Neurophysiology* **4**: 1–24
McGaugh J L 1989 Involvement of hormonal and neuromodulatory systems in the regulation of memory storage. *Annual Review of Neuroscience* **12**: 255–87
McGaugh J L 2000 Neuroscience: Memory: a century of consolidation. *Science* **287**: 248–51
Milner B 1966 Amnesia following operation on the temporal lobes. In: Whitty C W M, Zangwill O L (eds.) *Amnesia*. Butterworths, London, pp. 112–5
Mishkin M 1978 Memory in monkeys severely impaired by combined but not by separate removal of the amygdala and hippocampus. *Nature* **273**: 297–8
O'Keefe J 1976 Place units in the hippocampus of the freely moving rat. *Experimental Neurology* **51**: 78–109
Papez J W 1937 A proposed mechanism of emotion. *Archives of Neurology and Psychiatry* **38**: 725–43
Pavlov I P 1927 *Conditioned Reflexes*. Trans. Anrep G V. Oxford University Press, London
Sharpless S, Jasper H 1956 Habituation of the arousal reaction. *Brain* **79**: 655–80
Sokolov E N 1963 Higher nervous functions: The orienting reflex. *Annual Review of Physiology* **25**: 545–80
Thompson R F 1983 Editorial. *Behavioral Neuroscience* **97**: 3
Thompson R F 1986 The neurobiology of learning and memory. *Science* **233**: 941–7
Thompson R F, Spencer W A 1966 Habituation: A model phenomenon for the study of neuronal substrates of behavior. *Psychological Review* **73**: 16–43
Watson J B 1913 Psychology as the behaviorist views it. *Psychological Review* **20**: 158–77
Wundt W 1874 Grundzüge der physiologisch psychologie (Foundations of physiological psychology). Trans. Titchener E B. Macmillan, New York

R. F. Thompson

Behavioralism: Political

Behavioralism is a current or approach within the discipline of political science. In very general terms, it can be described as theory-led empiricism which aims at the establishment of laws, using as its epistemological model the methodology of the natural sciences. Its goal is the description, explanation, and prediction of political processes based on quantifiable data. Despite initial bitter resistance, behavioralism infiltrated most of the major American political science departments during the 1950s. From the early 1960s, it significantly molded the character of American political science for the next 15–20 years. Since then, there have been hardly any self-declared behavioralists, but the spirit of empiricism and methodical professionalism has shaped the discipline of political science within and outside of the United States to this day.

1. Historical Background

Viewed as an academic movement intended to reorient political science, behavioralism mainly represents a phenomenon of the 1950s and 1960s. As an intellectual school of thought, it can be traced back to the New Science of Politics movement of the 1920s and the Chicago School of Charles E. Merriam and Harold D. Lasswell. Both advocated a political science discipline modeled after empirical sociology, behavioristic psychology and modern economics. The New Science of Politics movement began as a reaction to the then-dominant institutionalist approach, known as 'institutional realism' or 'realist institutionalism.' The realist approach centered on the analysis of the actual (as opposed to simply the legal) relationships between branches of government, political parties, and interest groups. In general, the analysis of political institutions was oriented less toward theory-building than toward fact-finding, and a strict distinction between factual statements and value judgments was rare.

Improved collaboration between political scientists and administrators and politicians during the time of the New Deal and especially during World War II revealed the shortcomings of the realist approach. The gulf seemed too wide between what was required for effective political consulting and what political scientists were able to provide. In particular, the limited predictive capabilities were considered a major shortcoming (Dahl 1961). For example, realists fell short in explaining the success of European fascism through institutional analysis only. To many political scientists, the insights taken from neighboring disciplines, especially from psychology, sociology, and the economics, seemed theoretically and methodologically more sophisticated than those taken from political science.

Primarily younger political scientists became sharp critics of the realist approach but none formulated it with more urgency than David Easton. In programmatic writings dating back to the early 1950s, he criticized the theoretical deficit of the research at that time, the lack of methodological reflection and of

conceptual clarity, the absence of modern methods of data collection and evaluation, as well as the one-sided emphasis on institutional data. Owing to its lack of theoretical orientation and of a standard terminology, political science with its one-sided concentration on facts had failed in keeping pace with the development of sociology and psychology. Indeed, it had not established a canon of empirically sound knowledge that would allow the explanation and prediction of political events (Easton 1953).

Important sociological and socio-psychological studies such as *The People's Choice* (Lazarsfeld et al. 1944), the four volumes of *The American Soldier* (Stouffer et al. 1949) and *The Authoritarian Personality* (Adorno et al. 1950) proved to be important stimuli for those who criticized the institutional approach. Using newly developed data collection methods such as random sampling, attitude measurement based on psychometric scales, panel analysis, and sophisticated statistical methods, those books paved the way for successful behavioral research in the field of political science. Concurrently, a small group of researchers associated with the University of Chicago published several works along the lines of the emerging behavioralism approach. For example, in *Administrative Behavior* Herbert A. Simon (1947) advocated an administrative science based on Max Weber's postulate of value-free judgment and the spirit of logical empiricism. In 1949, V. O. Key Jr. published *Southern Politics*, soon to assume the status of a classic. In 1950, Gabriel Almond's empirical study of US attitudes towards foreign policy, *The American People and Foreign Policy*, followed. 1951 saw the publication of David Truman's analysis of politics as a process of group conflicts, entitled *The Governmental Process* (Truman 1951). In these studies, which were closely based upon methodological individualism, an alternative to the still dominant realistic institutionalism of traditional political science was developed. To many within the rising generation of academics these studies served as a paradigm of further research.

2. *The 'Behavioralist Revolution'*

In the early 1950s, the adherents of behavioralism were a small, scattered band of lone fighters, not held together by any organization, but loosely connected through their understanding of science. They met at seminars and gathered in conjunction with meetings of the American Political Science Association (APSA) to discuss new methods of empirical political science. The restructuring of the entire discipline of political science along the lines of behavioralism and not simply the addition of one more subdiscipline was the goal of these 'young Turks.' Not surprisingly, the demand for such a fundamental reorientation of political science met with stiff opposition from many established political scientists.

The supporters of the behavioralist approach increased steadily in number. They enjoyed early notable successes with the elections of Harald D. Lasswell, perhaps the most important intellectual pioneer of behavioralism, as President of APSA in 1955–6, and of V. O. Key Jr. as his successor in 1957–8. Under Lasswell, the APSA for the first time established a working group of its own concerned with issues of political behavior. By 1960, the number of behavioralist-oriented working groups at APSA conventions had grown substantially. At the same time, a series of important research studies tailored to behavioralism was published. In addition to the electoral studies conducted at Columbia University and at the University of Michigan (*Voting* and *The Voter Decides*, respectively), several programmatic and empirical studies were aimed at revamping the subfields of Comparative Politics and International Relations. After 1956, the profile of the *American Political Science Review* began to change fundamentally, a sign of the gradual transformation of the discipline. The quantitative element was slowly pushing itself to the fore. Practically all of the major American universities appointed behavioralists to their political science departments, starting a sort of chain reaction that accelerated the spread of the behavioralist conviction. After 1960, studies written in the spirit of behavioralist research began to shape the profile of the leading political science publications and the institutional policies within APSA. With only a very few exceptions, all Presidents of the APSA after 1964 were elected from the ranks of the behavioralist movement. Within a decade and a half, behavioralism had succeeded in remodeling the field of political science in the United States.

The rise of behavioralism was also evident in the dramatic increase in research output after 1960. Not a single area of political science was left untouched by the behavioralist wave, although the extent to which behavioralist thought penetrated the various subdisciplines differed. At the center was the study of decision-making and behavioral patterns. In behavioralist research, the individual served as the basis for data collection, and the individual actor, the group, the institution, or even the political system in its entirety formed the basis upon which statements were made. The core concepts of behavioralist research were *political decisions* (Simon 1947), *roles* (Eulau 1963), *interest groups* (Truman 1951), the *political system* (Easton 1965), or *power* itself (Lasswell and Kaplan 1950).

Thus, theoretical concepts and not any longer the analysis of day-to-day-politics as promoted by many institutional realists became the core of intensive discussion. That its main representatives were renowned scientists gave credence to behavioralism. Their research helped to promote the behavioralist approach probably more effectively than all programmatic studies combined that had preceded them.

Among the most renowned studies are *The American Voter* (Campbell et al. 1960), *The Civic Culture* (Almond and Verba 1963) and *The Nerves of Government* (Deutsch 1963). The financial assistance of major scientific foundations and corporations such as Carnegie, Ford, and Rockefeller proved crucial to the success of behavioralist studies that are often time consuming and expensive in their implementation.

3. *The Behavioralist Research Program*

The most important aims and assumptions of behavioralism are summarized and discussed below. Largely owing to the split between a more theoretical branch and a more empirical one, few behavioralists would have agreed with every one of these statements, but the catalogue of behavioralist principles intends to capture the theoretical and epistemological nucleus of the various behavioral science branches within American political science.

(a) In general terms, behavioralism can be defined as *theory-oriented empiricism*. Its aim is not the pure description of political processes, but their explanation and prediction through empirical laws that postulate political and social regularities.

(b) *A systematic and cumulative method*. Empirical investigations should build upon the findings of earlier studies, thereby increasing the range of theoretical statements.

(c) *A focus on the social and psychological determinants of political behavior*. The latter determinants include all types of personality traits such as attitudes or behavioral intentions. The use of mental concepts and the acceptance of introspectively obtained information differentiates behavioralism from psychological Behaviorism. Thus, treating both currents as equivalent seems unwarranted.

(d) *The inductive method*, i.e., the establishment of scientific laws by generalizing from observed social regularities.

(e) *Interdisciplinarity*, i.e., the use of approaches, theoretical concepts and results from other social sciences; this is one of the central elements of the behavioralist research program.

(f) *Orientation toward basic research*. A demand for applied research that is either too forceful or untimely leads according to the 'behavioral creed' to a theoretical abstinence that obstructs cognitive progress.

(g) *The demand for verification*. According to behavioralism, all statements in political science must be in principle verifiable.

(h) *Strict operationalism*. All theoretical terms must be translatable into an observational language. In order to meet the demand for empirical verification, all statements must therefore be attributable to observable behavior.

(i) *A focus on the behavior of individuals as the basic empirical unit of analysis*. Behavioralist research centers more on the informal aspects of the political process and the behavior of individual actors and political groups than on institutions and formal rules. In contrast to psychological behaviorism, the behavioralistic notion of behavior includes not only directly observable behavioral acts but also the attitudinal antecedents and the products of behavior as well as speech acts.

(j) *The use of empirical methods of data collection and evaluation*. In order to meet the demand for verification, suitable collection and evaluation methods should be used, such as attitude scales or complex statistical techniques. Quantitative concepts should be employed wherever possible. According to the research program (although not always to behavioralistic research practice), when a numerical description is not possible, one should turn to qualitative methods. Quantification should not be conducted to the detriment of the substance or of the relevance of the subject matter.

(k) *Adherence to scientific quality criteria*. For reasons of precision and of verification, the instruments of measurement employed must meet the demands of reliability, validity, and objectivity (understood as independence of results from the individual researcher). Furthermore, the representativeness of data is precondition for generalizations. Finally, to avoid contradictions, the concepts employed should be clearly and unequivocally defined and used consistently.

(l) *The principle of value neutrality*. In empirical research, value judgments cannot be justified by scientific method. Owing to their normative character, they interfere with the cognitive content of empirical statements. Thus, such statements should not contain any value judgments.

Over these propositions, the so-called behavioralism controversy erupted during the 1960s. Adherents of normative–philosophical and neo-Marxist standpoints attacked the basic tenets of behavioralism Most self-proclaimed behavioralists, however, opted to stay out of the dispute in favor of concrete, empirical research. Thus, the debate was somewhat lopsided. As for behavioralism itself, it survived this dispute about methods relatively unscathed (Waldo 1975, Falter 1982).

4. *The Imprint of Behavioralism*

In practice, the goals of the behavioralism research agenda have been implemented to differing degrees. For example, the principle of interdisciplinarity has been adopted widely and successfully. As far as the theoretical foundation or the value neutrality was concerned, however, some of the research fell short of its goals. Despite earlier efforts by David Easton (1953, 1965) and Karl W. Deutsch (1963), and exceptions notwithstanding, the development of termino-

logical frameworks with theoretical relevance have remained largely without consequence for empirical research. With regard to verificationism, behavioralism opposes modern analytical theory of science that—in view of the logical shortcomings of the verification principle—only insists on the indirect falsification or confirmation of statements. The same is true for the behavioralistic postulate of strict operationalism; modern analytical theory of science allows as well the indirect empirical interpretation of theoretical concepts.

Even if hardly any modern political scientists see themselves as behavioralists in the narrow sense, American political science, and in its wake major sections of the discipline worldwide, were shaped and permeated by behavioralism. Therefore, one can indeed speak, to use Thomas Kuhn's expression, of a scientific revolution (Kuhn 1962). Contemporary electoral and attitudinal research, policy analysis, and numerous representatives of the rational choice approach owe their convictions to behavioralism.

See also: Behaviorism; Behaviorism, History of; Key, Valdimer Orlando (1908–63); Participation: Political; Political Science: Overview; Political Sociology; Polling; Voting, Sociology of

Bibliography

Adorno T W et al. 1950 *The Authoritarian Personality*. Harper, New York

Almond G A 1950 *The American People and Foreign Policy*. Harcourt Brace, New York

Almond G A, Verba S 1963 *The Civic Culture. Political Attitudes and Democracy in Five Nations*. Princeton University Press, Princeton, NJ

Campbell A F et al. 1960 *The American Voter*. Wiley, New York

Dahl R A 1961 The behavioral approach in political science—epitaph for a monument to a successful protest. *American Political Science Review* **55**: 763–72

Deutsch K W 1963 *The Nerves of Government: Models of Political Communication and Control*. Free Press of Glencoe, New York

Easton D 1953 *The Political System—An Inquiry into the State of Political Science*, 1st edn. Knopf, New York

Easton D 1965 *A Systems Analysis of Political Life*. Wiley, New York

Eulau H 1963 *The Behavioral Persuasion in Politics*. Random House, New York

Falter J W 1982 *Der 'Positivismusstreit' in der Amerikanischen Politikwissenschaft. Entstehung, Ablauf und Resultate der Sogenannten Behavioralismus-Kontroverse in den Vereinigten Staaten 1945–1975*. Westdeutscher Verlag, Opladen

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Lasswell H D, Kaplan A 1950 *Power and Society—A Framework for Political Inquiry*. Yale University Press, New Haven, CT

Lazarsfeld P F et al 1944 *The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign*. Columbia University Press, New York

Simon H A 1947 *Administrative Behavior: a Study of Decision-making Processes in Administrative Organization*, 2nd edn. Macmillan, New York

Stouffer S A 1949 *The American Soldier Vol. 1–4*. Princeton University Press, Princeton

Truman D B 1951 *The Governmental Process*, 1st edn. Knopf, New York

Waldo D W 1975 Political science—tradition, discipline, profession, science, enterprise. In: Greenstein F I, Polsby N W (eds.) *Handbook of Political Science. Vol. 1. Political Science: Scope and Theory*. Addison-Wesley, Reading, MA

J. W. Falter

Behaviorism

Historically, behaviorism was the 'school' of psychology, associated with the name of John Broadus Watson, that rejected mental states and treated all psychological phenomena in terms of stimuli, responses, and stimulus–response associations. Watson launched the system in 1913, with his behavioristic manifesto, *Psychology as the Behaviorist Views It*. He spelled out implications and applications in his 1924 book, *Behaviorism*. This article describes behaviorism as Watson conceived it and presents the criticisms and reactions that led to revisions and a collection contemporary behaviorisms that have replaced Watson's version.

1. Classical Behaviorism

Watsonian behaviorism came into psychology as a protest against structural psychology ('structuralism'), the dominant theory of the time. Structuralism asserted that the purpose of psychology is to understand the mind and consciousness and that the road to such understanding is the introspective method. Watson (1924, p. 3) rejected both of these ideas: 'Consciousness is neither a definable nor a usable concept … it is merely another word for "soul."' As for introspection, 'we find as many introspections as there are individual psychologists. There is no way of experimentally attacking and solving psychological problems and standardizing methods' (Watson 1924, pp. 5–6).

One way to introduce the substantive ideas in classical behaviorism is to organize them around what, for Watson, were the hallmarks of the scientific method: empiricism, determinism, and analysis.

1.1 Empiricism

Instead of consciousness, Watson asked psychology to limit its inquiry to observables: 'Why don't we make

what we can observe the real field of psychology? Now what can we observe? Well, we can observe behavior (responses)—what the organism says or does [and the situations (stimuli) in which behavior occurs'] (Watson 1924, pp. 6–7). This insistence on observation had consequences that often were unpopular.

1.1.1 Rejection of mental states. The psychology of Watson's time often analyzed the contents of the mind into Plato's three human talents: knowing, feeling and doing, which modern psychology calls cognition, affect, and reaction tendencies (Kimble 1996). Watson objected to such mentalistic concepts and translated them into conditioned reflexes: 'laryngeal (language) reflexes' for cognition, 'visceral (emotional) reflexes' for affect, and 'manual (motor) reflexes' for reaction tendencies.

1.1.2 Rejection of physiological states. Watson put physiological mechanisms in the same category of unobservables as mental states. He had little hope that behavior can be reduced to physiology and blamed the structuralists for the promotion of the fiction that it can be: '[For a structuralist] the nervous system … has always been a mystery box—whatever he couldn't explain in "mental" terms he pushed over into the brain' (Watson 1924, p. 43). 'Until [the physiologist has reduced the various phenomena of psychology] to electrical and chemical processes … he cannot help us very much' (p. 169).

1.1.3 Radical environmentalism. Watson's methodological empiricism led naturally to the conclusion that most human traits are learned. He expressed this sentiment in his famous statement:

> Give me a dozen infants, well-formed, and my own specified world to bring them up in and I'll guarantee to take any one at random and to train him to become any type of specialist I might select—doctor, lawyer, merchant-chief and yes, even beggar-man and thief, regardless of his talents, penchants, tendencies, abilities, vocations, and the race of his ancestors' (Watson 1924, p. 82).

1.2 Determinism

Watson came to his position from an interest in animal research, a fact that had two major consequences for his treatment of psychology: First, it accounts for his preference for laboratory experimentation over introspection, which becomes undisciplined anthropomorphism in the study of animal behavior. Second, it emphasized the fact that human beings are animals: 'The behaviorist, in his attempts to get a unitary scheme of animal response, recognizes no dividing line between man and brute' (1924, p. 16). As is true of other animals, human behavior is determined: 'the child or adult has to do what he does' (p. 144).

1.3 Analysis

Watson was explicit on the point that analysis is inherent in the behavioristic position: 'All through our study of behavior [we will have] to dissect the individual. My excuse is that it [is] necessary to look at the wheels before we [can] understand what the whole machine is good for' (1924, p. 216).

2. Criticism and Revision

Even in Watson's time, many psychologists objected to behaviorism, because it seemed to dehumanize humanity, to make the individual a robot, and to ignore the personal realities of consciousness, will, and sense of self. Later, in the middle of the twentieth century, some members of the scientific community (particularly those who fought against behaviorism in the 'cognitive revolution') criticized the position in other terms: as merely scientistic, a mimicry of science. Because it concentrates on laboratory investigation with white rats, it produces knowledge that is superficial, oversimplified, and irrelevant to human behavior.

Such criticism led psychologists who leaned toward behaviorism to the development of several more liberal behaviorisms, which revised Watson's theory in ways that can be presented by revisiting the topics of empiricism, determinism, and analysis.

2.1 Empiricism

Beginning in the 1920s psychology came under the influence of the philosophy of science called logical positivism, which insisted that the basic data of science reside in public observation, and that to be acceptable to science concepts must have what came to be called operational definitions. In the words of Kenneth Spence some years later:

> Accepting the requirement that all terms employed by the scientist, no matter how abstract, must ultimately be referable back to some primitive or basic set of terms that have direct experiential reference, [the logical positivists] suggested that such a basic ['observable thing-language'] for science is provided by the terms designating directly observable physi-

> cal objects such as tables, chairs, and cats, their properties, e.g., green, hard and loud, and the relations [among] them such as between, before, and below (Spence 1956, p. 13).

Operationism allowed acceptance of most of what Watson had rejected along with the subjectivity of structural psychology.

2.1.1 Mental states. As early as 1922, Edward C. Tolman offered a 'new formula for behaviorism' which encouraged psychology to treat mental states as intervening variables that are tied to observation by operational definitions. Later on (1925), for example, he defined one such state for a nonhuman organism—a rat's 'purpose' to obtain food at the end of the maze that the animal is learning—in terms of observable quality of the animal's behavior that one might call 'persistence until' the goal is reached. Once it honors its commitment to base its science on stimuli and responses, psychology can use those data to define concepts that are as biological, cognitive, or humanistic as it chooses.

2.1.2 Physiological states. Since Watson's time, advances in neuroscience (PET scans, MRI imaging, computerized electroencephalography) have provided ways of determining whether behavioral concepts have representations in the brain. When they do, most psychologists take these representations as evidence of the physiological reality of their concepts. A minority, with Watson, believe that this is all they are, that the final determiner of the value of a concept for psychology is its relevance to behavior. All that physiology can contribute is the comfort provided by any confirmation of a theoretical idea.

2.1.3 Biological–environmental interaction. Also since Watson, advances in the study of animal behavior (ethology) and behavioral genetics have tempered the behaviorists' radical environmentalism. Now, the standard position is that behavior results from an interaction between environmental and biological influences. The relative importance of environment and biology vary. For instinctive acts, biology makes the stronger contribution; in social behavior environment is more important.

2.2 Determinism

This interpretation of human behavior as the consequence of interactions between external forces leaves some psychologists, along with ordinary people, uncomfortable because it violates the notion that behavior is caused by resident agents with tangible reality. For those who take this point of view, to explain behavior is to identify those causal agents. Many experimental psychologists find causality in the new-found biological mechanisms mentioned earlier. For many cognitive psychologists, causality is in some mental agent who is busily constructing perceptions, searching memory, parsing text, and using algorithms and heuristics to solve problems. Other psychologists, including many behaviorists, question both conceptions because, in science, explanation is not in terms of indwelling physiological—much less mental—causes, but in terms of laws relating the dependent variables of that science (behavior for psychology) to independent variables.

2.3 Analysis

Although Watson accepted the scientific principle of analysis, contrary to common opinion, he was also recognized that the responses resulting from this analysis are responses of a total individual:

> Some psychologists seem to have the notion that the behaviorist is interested only in the recording of minute muscular responses. Nothing could be further from the truth. The behaviorist is … interested in the behavior of the whole man. … When [a man] reacts he reacts with each and every part of his body (Watson 1924, pp. 14, 75).

What was missing in Watson's theorizing were the overarching organizing mechanisms that, later on, Lashley (1951) called a 'syntax of behavior' and Miller et al. (1960) referred to as 'plans.' That omission is a flaw that remains in most modern versions of behaviorism.

3. Current Behaviorism

For Watson, behaviorism was 'a natural science that takes the whole field of human adjustments as it own' (Watson 1924, p. 11). He gave behavioristic interpretations of child rearing, education, business management, social programs, psychopathology, and psychotherapy. Throughout his long career, B. F. Skinner made similar claims for widespread relevance of behaviorism. He applied behaviorism to education, verbal behavior, psychotherapy, and even an imagined utopian society, Walden Two (Skinner 1948).

However, most psychologists with behavioristic sympathies abandoned comprehensive theory and turned to narrower applications, often in the field of learning. In that context, it is possible to identify a

family of behaviorisms that accept Watson's insistence that scientific psychology is stimulus–response psychology but differ on several other such points.

Radical behaviorism (Skinner 1938) places a strong emphasis on the environmental control of action. It rejects both mental states and physiology as determiners of behavior, and asserts that theories of learning are unnecessary.

Operational behaviorism (Tolman 1922, Spence 1956) accepts mental states when they are defined operationally. It recommends deductive theory and underemphasizes physiology.

Physiological behaviorism (Hull 1943, and most modern behaviorists) combines operational behaviorism with a preference for concepts identified with physiological processes and taken as effective causes of behavior.

Functional behaviorism (Thorndike 1911, Skinner 1981, Kimble 1996) treats behavior as a product of evolution. Both the behavior that has been preserved in evolution and the behavior that is acquired by learning have adaptive value.

Theoretical behaviorism (Staats, 1999, Staddon, 1999) treats internal states as theoretical constructions which may or may not be causal agents.

Only a few psychologists (Kimble, 1996, Staats, 1999) carry on the Watsonian tradition of behaviorism as a general theory of behavior.

See also: Animal Cognition; Autonomic Classical and Operant Conditioning; Bernard, Jessie (1903–96); Classical Conditioning and Clinical Psychology; Classical Conditioning, Neural Basis of; Experimentation in Psychology, History of; Watson, John Broadus (1878–1958)

Bibliography

Hull C L 1943 *Principles of Behavior*. Appleton, New York

Kimble G A 1996 *Psychology: The Hope of a Science*. MIT Press, Cambridge, MA

Lashley K S 1951 The problem of serial order in behavior. In: Jeffress L A (ed.) *Cerebral Mechanisms in Behavior: The Hixon Symposium*. Wiley, New York, pp. 112–46

Miller G A, Galanter E, Pribram K H 1960 *Plans and the Structure of Behavior*. Holt, New York

Skinner B F 1938 *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century, New York

Skinner B F 1948 *Walden Two*. Macmillan, New York

Skinner B F 1981 Selection by consequences. *Science* **213**: 501–4

Spence K W 1956 *Behavior Theory and Conditioning*. Yale University Press, New Haven, CT

Staats A W 1999 Unifying psychology requires new infrastructure, theory, method, and a research agenda. *Review of General Psychology* **3**: 3–13

Staddon J E R 1999 Theoretical behaviorism. In: O'Donohue W, Kitchener R (eds.) *Handbook of Behaviorism*. Academic Press, New York, pp. 217–41

Thorndike E L 1911 *Animal Intelligence*. Macmillan, New York

Tolman E C 1922 A new formula for behaviorism. *Psychological Review* **29**: 44–53

Tolman E C 1925 Purpose and cognition: The determiners of animal learning. *Psychological Review* **32**: 285–97

Watson J B 1913 Psychology as the behaviorist views it. *Psychological Review* **20**: 158–77

Watson J B 1924 *Behaviorism*. Norton, New York

G. A. Kimble

Behaviorism, History of

Behaviorism was gradually developed by some psychologists during the first decade of the twentieth century. There were several streams that motivated this. One was the effort to introduce methods of 'objective' natural science into psychology. The second was the effort to make explicit the argument against introspectionism. In fact, for many, behaviorism was akin to an intellectual war against introspective methods. The need for an objective, more scientific psychology was expressed in several countries: V. Bechterev in Russia, H. Piéron in France and J. B. Watson in USA called for a 'new psychology,' closely related to physiology. Their success depended on the historical, religious and scientific context in each country. But soon supporters of mentalism tried to crowd the psychological laboratories again, producing such a muddle that B. F. Skinner engaged in a crusade to restore the purity of behaviorism, conceived of as the philosophy of the science of behavior. Unfortunately, he encountered a prominent champion of mentalism, Noam Chomsky. The psycholinguist took advantage of the troubled times of the Vietnam War and the students' protest against the establishment to provide an alternative look at human beings, that was apparently much richer and deeper.

1. The Contexts of Emergence

1.1 The Nature of Voluntary Acts

At the end of the nineteenth century, psychology was deeply influenced by philosophical issues such as will or intelligence. Most psychologists tried to encompass these complex issues by using introspection, asking subjects to report their feelings or mental operations during some experimental task. In the context of an increasing appeal to scientificity, the subjectivity inherent in such a method soon turned into an obstacle hard to overcome. At the same time, evolutionary ideas, and particularly those of Darwin, exerted a

considerable pressure on the conception of natural beings. As Darwin proposed, the distinction between animal and human intelligence was one of degree, not of kind. Thus, the study of human intelligence or of human learning faculties may begin with animals. Several researchers, such as G. Romanes (one of Darwin's students) and C. Lloyd Morgan (one of Huxley's students), devoted their work to understanding the instinctive capacities of animals, and particularly their learning capacities. Emphasis was put on imitation and on trial and error learning, as discovered by A. Bain, professor of logic in Aberdeen. The question Bain raised was a decisive one and concerned the basis of the voluntary act: when someone performs an action for the first time, does he or she do it in response to an exterior solicitation, or because he or she wants it? For the time, Bain's answer was highly original: it is by accident. If that accident is followed by a benefit, the individual will repeat that response whenever the solicitation occurs again. If the 'sanction' is painful, he or she will not. According to Bain, it is through this process that a spontaneous act becomes a voluntary one (Bain 1855). Of course, Bain's conceptions evoke the Darwinian selection of types as well as the utilitarianism of his friend Stuart Mill. In just the same way that biological types emerge, an act is evaluated from the perspective of its usefulness, in terms of the pleasure or pain it engenders.

1.2 The Beginnings of Animal Psychology in the United States

The works of these first British psychologists did not receive a real response in Great Britain. However, they considerably influenced American psychology and the birth of behaviorism. One man in the United States was especially receptive to the beginnings of animal psychology: E. L. Thorndike. Thorndike was a student of William James at Harvard and was inevitably influenced by the pragmatism and the functionalism of James. After reading the works of Morgan, Thorndike decided to study the mechanisms of learning for some kinds of animal behavior that were not included in their instinctive endowment. In James' cellar, he built some strange boxes, hard to open (the device to exit included, for instance, a simple rotating catch), in which he put a hungry animal and then observed the animal's behavior. These puzzle boxes marked the beginning of American behaviorist psychology. Thorndike studied trial and error learning in a systematic and quantitative manner (he quoted the number of trials necessary to open the door and the response latency for each trial, and thus provided the first learning curves). On the basis of several experiments, he was convinced that the animal learned to solve the problem by connecting the visual perception of the opening mechanism with what he called the successful motor impulse. The implication of this 'stimulus–response,' or S–R, view of learning is that animals learn what to do without in any sense knowing the consequence of the action. The sole effect of reward (something to eat) was to 'stamp in' the S–R connection. Thorndike called this phenomenon the 'Law of Effect' (Thorndike 1898).

1.3 Psychology in Russia and France

But the works of Thorndike did not initially receive a significant response in the United States. He wrote about his researches to a physiologist who was also working on animal learning, far from the United States: I. Pavlov. At the end of the nineteenth century, Pavlov studied the gastric secretion and conditional reflex in dogs. The study of conditioning began during a period of dramatic political developments in Russia, and universities were at the center of political strife. The medical school of St. Petersburg diffused materialistic ideas. In this school, a psychiatrist colleague of Pavlov, V. Bechterev, proposed to stop asking patients to report their feelings and subjective experience in the study of mental diseases. He was convinced that an objective study of the changes occurring during mental diseases would be more instructive. He presented a synthetic report of his convictions in a famous paper of 1906 called 'La psychologie objective,' in which he expressed a radical rejection of introspection. He asserted that introspection had no contribution whatsoever to make to the wider range of problems which psychology should be concerned with (such as mental testing, psychiatric disorders, and animal behavior). Following Sechenov and Pavlov, Bechterev considered the reflex as a key concept of this new objective psychology.

This trend towards an objective point of view in psychological studies was also important in France. In 1908, a psycho-physiologist, H. Piéron, asserted that, to be scientific, psychology had to be centered on behavior and had to ignore consciousness. Piéron was very influential, and founded many institutions and psychological reviews. His objective psychology remained, however, too physiological to change the French way of considering human conduct significantly. Psychoanalysis and the clinical psychology of P. Janet were much more influential and French psychology never became truly 'behavioristic.' Powerful obstacles towards a materialistic conception of man included questions of philosophical context (such as spiritualism), biological choices (Piéron, for example, was a neo-Lamarckian), and religious background (Catholicism remained silently important).

Meanwhile, comparative aspects entered into the battle for objective psychology in the United States. American psychology expanded rapidly around the turn of the century, becoming a prominent aspect of the general growth of higher education and scientific research. Thorndike had stopped working with ani-

mals soon after completing his thesis, and animal psychologists were not receiving financial support, in contrast to Pavlov's school in Russia. Animal psychologists studied first the sensory abilities of animals. Most of them went to Germany to learn the strict methodological rules of experimentation. It was first in France, then in Germany, that a strange but decisive debate concerning invertebrates' behavior took place.

2. *The Psychology of Microorganisms and 'Behavior Man'*

2.1 *A New Object to Study: Protozoa*

Although most psychologists believed that their subject was, or should be, a science, they continued to debate questions dealing with the mind–body problem. As a consequence, psychology here and there still appeared more as a branch of philosophy. The question of where mind and inner life appear on the evolutionary scale became of interest when A. Binet, a French psychologist, published in 1889 a book entitled *La Vie Psychique des Micro-organismes*. He suggested that single-celled animals could perceive and discriminate objects and perform purposive actions. Binet agreed with the rather romantic assertion of E. Haeckel, the German evolutionist, that 'all nature was ensoulled.'

Jacques Loeb, a young German student, imported such preoccupations to the United States. The year Binet's book was published, Loeb proposed a general theory of tropism, first concerning plants, then lower animals. He carried out experiments in the new University of Chicago, and a lively group of young experimentalists did the same at Harvard, recording reactions to systematic changes in various forms of stimulation (chemical, electrical, and by temperature change, for example). Loeb's aim was to explain behavior along purely mechanistic lines, according to a rather reductionist viewpoint. His conceptions were strongly attacked by H. S. Jennings, who asserted that the life of protozoa was far more complicated than Loeb maintained. Jennings joined the biology department of Johns Hopkins University two years after J. M. Baldwin had been appointed to revive psychology at this university. Some of Jennings' ideas about evolution were based on the 'principle of organic selection' that Baldwin and Lloyd Morgan had proposed in 1890.

2.2 *The 'Crisis' in American Psychology*

It is in such a paradoxical context of interest in the evolution of the mind, on the one hand, and in the study of mindless creatures, on the other, that J. B. Watson came to animal psychology and then founded behaviorism. At this time, American psychology was characterized by a lack of strong direction. It had to find a way of reconciling three conflicting aims: (a) to explain the nature of subjective experience and study the mind; (b) to be scientific and objective and use the kind of empirical methods employed in other sciences; (c) to be interesting for questions in everyday life or to make significant practical contributions (specially in education or mental health). When Watson defended the inclusion of animal studies within psychology, he presented the analysis of behavior as promising a combination of usefulness and scientific respectability that the previous alternative, the psychology of consciousness, had failed to provide.

Watson was an atypical American scientist. He came from a small farm in South Carolina and had a wild adolescence. Unexpectedly, he developed a strong interest in philosophy. At 22, after having written to John Dewey, he went to the University of Chicago where a psychological laboratory, included in the philosophy department, was directed by James Angell. There he began working on the process of myelinization of nerve fibers in the brain of rats. He became a close friend of R. Yerkes, who had been one of Thorndike's assistants and then championed comparative psychology as the field capable of providing criteria for the objective scientific verifiability necessary to imbue the whole discipline with sustained faith. Yerkes tried to resolve the disciplinary 'crisis' which affected psychology. He studied discrimination in 'dancing mice' in Harvard and embarked on a collaborative project with Watson on methods for studying vision in animals. Soon, Watson became much more radical and confident than Yerkes, especially concerning the rejection of mentalism. For several reasons, Watson had rapidly attained institutional and professional security: at the age of 31, he replaced J. M. Baldwin both as the head of the journal, *Psychological Review*, and as the chairman of the department of psychology at Johns Hopkins University. To be placed in this professional position was rather strange: Watson was not really well informed about human psychology, and he pursued a line of research which most of his colleagues believed to belong to biology, far from central issues of psychology. 'It is not up to the behavior man to say anything about consciousness,' Watson wrote to Yerkes in 1909. Nevertheless, the 'behavior man' was going to use that key position in American psychology to promote his standpoint.

3. *Watson's Manifesto of Behaviorism*

3.1 *'Psychology as the behaviorist views it'*

In 1913, Watson had the fortunate occasion to make explicit his numerous convictions about psychology. J. McKeen Cattell offered him a forum for their dissemination by inviting him to give a series of eight public lectures at Columbia University, the American center of applied psychology. First of all, Watson

announced that psychology should become an objective experimental science of the prediction and control of behavior. A behaviorist is a psychologist who asserts that there is no dividing line between man and brute, and who strives to arrive at a unitary scheme of animal response, including the behavior of man, with all its refinement and complexity. According to him, the introspective method was to be discarded and to be replaced by the observation of stimulus-response connections, which influence behavior without resort to consciousness. If psychologists followed the plan he suggested, the physician, the jurist and the businessman could utilize psychological data for practical purposes. Watson expressed optimism about the benefits psychology would obtain from taking the behaviorist standpoint. He knew that his conceptions had to confront very complex aspects of human psychology, such as imagination, judgment or reasoning, but he showed himself confident that behaviorism would succeed. A year later, his book *Behavior: An Introduction to Comparative Psychology* presented a central Watsonian conviction: 'there are no centrally initiated processes.' Implicit behavior consists of word movements and thought is a sub-vocal speech according to Watson.

If his colleagues at Johns Hopkins University, A. Meyer, H. Jennings, K. Dunlap, or K. Lashley, shared most of the opinions Watson expressed in 1913, they did not follow Watson on these new elements of behaviorism. First of all, they were Darwinians. But Watson, in contrast, and like Loeb, saw no interest in placing psychology within any evolutionary context. They rejected the search for any single principle, such as tropism or habit, to explain every aspect of behavior. And they did not share his faith in the peripheral origins of all psychological events. Moreover, Watson considered that science had to predict and control, when other psychologists claimed that the goal of science was to understand phenomena. Not all the 'behavior men' were Watsonian behaviorists.

Watson devoted a large part of his subsequent career to human behavior, and especially to the behavior therapy of some disorders that Freudians then called 'neurosis.' His most famous experiment in the field of emotional conditioning was conducted with an eleven-month-old infant, named Albert, whom Watson taught by conditioning to fear white rats (see Sect. 6.1). The case became widely cited as an early piece of evidence in support of conditioning theories of phobias. It was an occasion for Watson to criticize Freudian concepts and to begin a series of studies about children presented in a book he wrote with his assistant and wife, R. R. Watson, *The Psychological Care of an Infant Child*, published in 1928. In the long run, this book probably had a more profound effect on more people than anything else Watson did. The book was dedicated to the first mother of a happy child, the latter being described as a child educated by the behaviorist method. It was probably this book that encouraged the trend towards inflexible patterns of childcare that appeared in North America and Britain during the 1930s, until the publication of B. Spock's book, *Baby and Child Care*, in 1946.

3.2 *The Response to Watsonian Behaviorism*

E. Boring declared that, in the 1920s, it seemed as if all America had gone behaviorist. Even the 'behavior men' like Yerkes declared that, in spite of many criticisms of Watson's radicalism, they were still behaviorists. The manifesto changed the attitudes of human psychologists, in particular about objectivity, the rejection of introspection, and the need to experiment. The applicability of Thorndike's and Watson's theories to human educational psychology or to the therapy of human mental disorders supported the tendency to define psychology as the science of behavior. Behaviorism entered the real world and the term was widely employed to signify a practical attitude towards human problems.

4. *Neo-Behaviorism*

4.1 *Purposive Behavior*

During the second quarter of the twentieth century, a handful of second-generation behaviorists developed a new version of the science of behavior. The evolution of their conceptions was the result of some discord within behaviorism: reading Thorndike's and Watson's books, students were confronted with opposing views on how habits are formed. Thorndike was convinced that habits are entirely based on S–R connections and that such connections are strengthened by the occurrence of a satisfying award. Watson denied the Law of Effect and the role of reward because of its mysterious retroactive effect. This disagreement between the two founders of behaviorism stimulated further experimental work with animals, particularly in the University of California at Berkeley where Zing Yang Kuo and E. G. Tolman arrived in 1918. They studied the influence of reward or punishment on the learning behavior of rats in mazes. But Tolman's real aim was not to evaluate the relative merits of Thorndike's and Watson's conceptions, but to propose another approach to the problem of learning. He asserts that behavior has to be studied from a molar point of view rather than a molecular one (as a series of physiological events).

Tolman was very influenced by Gestalt theoreticians, such as W. Köhler, who went to Berkeley in 1925 and whose detour experiments bore a strong resemblance to Tolman's study of insight in a maze. The experiments run at Berkeley led Tolman to his most significant book, *Purposive Behavior in Animals*

and Men (1932), in which he proposed to reintroduce some inner elements, such as the organization and anticipation of behavior, into the conception of action towards an end. Tolman suggested that his rats constructed a cognitive map of the maze. According to him, behavior was not governed by simple one-to-one S-R bonds, but by complicated patterns of adjustment. These patterns of adjustment constituted purposes, and the organism had to be conceptualized as a functional process. In so doing, Tolman reintroduced into psychology mental concepts that Watson had wanted to eliminate from science. Tolman asserted that purpose was a descriptive feature, not a mentalistic entity supposed to exist parallel to, and to run along side with, behavior, but was out there in the behavior. But when Tolman introduced into behaviorism the concept of intervening variables lying between the stimulus and the response, such variables undoubtedly corresponded to what ordinary people called 'mind.'

4.2 *Psychology as a Hypothetico-Deductive Science*

A decade after Tolman's innovative arguments, methods of statistical inference became standard procedure in the analysis of data from psychological experiments. Then, scientific papers expressed an extreme formalism, such as can be found in C. L. Hull's works. By the mid-1930s, a group of learning theorists at Yale University became a powerful force in American psychology. They benefited from the Behavior Research Fund, which was launched in 1926 and sustained by contributions from private citizens. It was based at the University of Chicago where J. Angell had directed the psychology department. When Angell became president of Yale University in 1921, he also benefited from the Rockefeller Foundation. These flows of money into Yale supported the creation of the Institute of Human Relations of which C. L. Hull became an appointed member in 1929. There he developed a learning theory based on a conditioning model and on research on maze-running rats. Like Pavlov, whose work was translated into English in 1927, Hull conceived of thought as habits and of conditioning experiments as the way to study all aspects of thought.

Hull argued that psychologists had to construct a systematic theory of learning by integrating the two sources of certainty: observed facts and deduction. Psychology had to postulate provisional axioms from facts, to deduce consequences from these axioms and to test these consequences by experiments. Psychology as a science must be a hypothetico-deductive activity. Of course, the hypothetical constructs Hull introduced in his theory looked like Tolman's intervening variables. For Hull, they consisted in inferred states of the organism or nervous system, like 'strength of habit' or 'drive,' whose reduction produces a reinforcing effect on behavior. His behaviorism was then different from the strictly descriptive one of Watson. Hull's aim, to transform the science of man into a mathematical science, (expressed in *The Principles of Behavior: An Introduction to Behavior Theory* in 1943), led him to a very formalistic theory, although one which called for a great number of experiments at the same time. Hull's theory was certainly the most famous behaviorist approach. Neo-behaviorism slowly began to falter in the 1940s, because of the numerous problems it failed to resolve, and because of the increasing influence of operationism and of Gestalt theory in the United States. These factors contributed to the development of studies of problem-solving and to the restoration of interest for what happens inside the 'black box.'

5. *The Skinnerian Crusade*

5.1 *Return to Pure Objectivism*

In 1927, B. Russell published a brisk and readable book called *Philosophy*, devoted to the basic conceptual issues arising in contemporary physics and psychology, and particularly in behaviorism. Russell was generally favorable to Watson's ideas. Being an eminent philosopher, he thus gave behaviorism considerable prestige and deeply influenced a young dilettante student, who promised to have a great future in behaviorism: B. F. Skinner. In 1928, at the age of 24, Skinner entered the Department of Psychology at Harvard University, headed by E. G. Boring who remained an opponent to Watsonianism. In 1936 Skinner was offered a teaching position at the University of Minnesota. In this university, by World War I, applied interests had grown so steadily that separate departments of educational psychology and child welfare had been established. Learning theories provided models for education and childcare. On the basis of numerous experiments with rats and pigeons, Skinner diverged from mainstream learning theories. He rejected the notion of trial-and-error learning and argued instead that all behavior involves one basic principle of reinforcement: an organism makes movements some of which are followed by a reward and will be repeated and others are not and will not be repeated. Thus, Skinner argued, it is possible to devise schedules of reinforcement, to construct circumstances that lead to one pattern of behavior rather than to another. His approach became known as operant conditioning, which he defined as 'the making of a piece of behavior more probable.' Skinner never referred to mind, drives, habits, or any intervening variable: his behaviorism was radical. It was underwritten by four major principles: (a) science needs only systematic observations and does not need theory; (b) the physiological factors which initiate animal behaviors do not concern psychology, i.e., psychology is the science of observable consequences; (c) behaviorism must not refer to mental states or ethical concepts; (d)

behavioral evolution during the life of an organism obeys the same laws as phylogenetic evolution because behavior is a form of biological adaptation (Skinner 1938, 1969).

5.2 Victory and Decline of Skinnerianism

Skinner returned to Harvard in 1948, where he remained for the rest of his career. The same year he wrote a utopian novel, *Walden Two*, inspired by the community described by H. D. Thoreau in the late nineteenth century. The Skinnerian community lives according to the principles of operant conditioning, without any negative reinforcement, positively reinforced each day, and purely programmed. In this utopia as well as in his scientific books, Skinner appeared as a combatant. He fought against the slow resurgence of mental entities or constructs in the description of human action. According to him, such concepts had the perverse effect of obstructing the way to the analysis of what goes on, in assigning a causal status to inferred inner states. He accused 'methodological behaviorists' like Watson but above all he accused Tolman of using experimental and objective principles to describe behavior, while at the same time resorting to inner states to explain behavior. They did so because they admitted that mental life was out of the reach of psychology. Skinnerian radical behaviorism contended that mental life, 'the world inside the skin,' is not very different from observable behavior, that mental life is made up of internal events, and is a behavior.

The gate methodological behaviorism opened was used by cognitivism, that is to say, by a branch of psychology which tried to explain what happens in the black box by resorting to inner, inferred states or operations. Of course, the rise of this psychology triggered frequent and firm reactions by Skinner, who was a provocative and formidable debater. He fought, for instance, against N. Chomsky who defended Cartesian convictions concerning language and attacked the Skinnerian reduction of language to an inner behavior, obeying the same contingencies of reinforcement as any other behavior (Chomsky 1957). Chomsky played a decisive role in the rise of cognitivism and in dismissing behaviorism among the ranks of university psychologists. For many years, in the departments of psychology of several universities in the United States and in Europe, the decay of behaviorism was frequently asserted. It seemed that Chomsky had the last word. But behaviorism was far from extinct: it remained alive in 'the real world.'

6. The Triumph of Applied Behaviorism

To this day, behaviorism remains a decisive source of social methods of control of behavior, even if the scientific indices of the activity of this psychological movement became, above all in Europe, less obvious. Since the beginnings of behaviorism, mental health and education are its main fields of application. The different conceptions of learning and of behavior conditioning led to various techniques of behavior modification and control. As well as Pavlovian conditioning, Watsonian or Skinnerian behaviorisms engendered several kinds of behavioral technologies. All of them were dedicated to the modification of human conduct, especially of abnormal or undesirable forms of such conduct. They were also designed to construct socially desirable behavior.

6.1 Behavior Therapies

Pavlov was probably the first psycho-physiologist to understand the social interest in conditioning. He conceived the normal cortical functioning as based on a precarious equilibrium between the mechanisms of inhibition and of excitation. Pathological behavior results from a disruption of this equilibrium, which is produced by the lack of some substance. Consequently, for example, a deficiency in physiological inhibition processes may be treated by an injection of caffeine or bromide. The administration of such substances obviously constitutes a way of behavior control, and Pavlov was the first to study the activity of such psychotropics on learning. Pavlov also proposed a model of 'experimental neurosis' in animals, induced by a conflict between two excitation processes and leading to animal agitation, aggression, and decline. During the twentieth century, the Soviet School would develop Pavlovian concepts and behavior therapies.

Watson's manifesto was soon followed by preoccupations with the role of conditioning in the construction of habits in childhood. In 1920, Watson and Rayner carried out phobia experiments on an eleven-month-old baby named Albert: they 'taught' Albert to avoid a white rat by associating this animal (which a baby spontaneously does not fear) with a metallic noise (that babies find very frightening). After several associations, the baby was afraid of the rat, and the experimenters generalized that fear to any fur-coated object. This experiment became largely cited as an early piece of evidence in support of conditioning theories of phobias. It was followed by numerous other such experiments; all used as arguments against Freudian theory, which was becoming famous in the West. The opposition between the behaviorist model of neurosis and the psychoanalytic one was much deeper in Europe than in the United States. Behavior therapists considerably developed their techniques (flooding, reciprocal inhibition, aversion, and token economy, for example,) which were used to modify antisocial as well as pathological behavior, or to fight obesity or drug addiction for example. Because they looked like the imposition of constraints on the self, they were frequently criticized and denounced as a

danger for freedom. In spite of such criticism, however, behavior therapies developed fast, even in France or Italy where Freudian conceptions remained very influential.

6.2 Educational Concerns

Since the first formulation of the S–R Law and of the Law of Effect by Thorndike, the field of education has been a key concern for 'behavior men'. Just after the publication of his thesis in 1898, Thorndike concentrated on the American educational system. At this time, in the United States, school learning was being turned into a professional activity; teachers were to be viewed as people possessing special skills and scientifically proven methods. All psychologists, and not only in the United States, were confronted with that social demand from teachers, but behaviorists were then able to furnish some simple formulas to resolve teaching problems.

Whatever the first behaviorists' contribution to education may have been, Skinner's work remains the more decisive and the best-known one. His concern for educational problems has been persistent, and consistent with his wider conceptions of culture and society. In 1954, he developed the principles of programmed teaching, made of linear sequences in the construction of correct answers. According to the principles of operant conditioning, the correct answer has to be reinforced. Teaching machines have considerably developed with computers, which offered unforeseen possibilities.

See also: Behavior Therapy with Children; Behaviorism; Evolutionism, Including Social Darwinism; Habit: History of the Concept; Intelligence: History of the Concept; Introspection: History of the Concept; Learning Theories and Educational Paradigms; Mind–Body Dualism; Skinner, Burrhus Frederick (1904–90); Utilitarian Social Thought, History of; Watson, John Broadus (1878–1958);

Bibliography

Bain A 1855 *The Senses and the Intellect*. Parker, London

Bechterev V 1906 La psychologie objective. *Revue Scientifique* **6**: 353–7

Binet A 1889 *La Vie Psychique des Micro-organismes*. Alcan, Paris

Boakes R 1984 *From Darwin to Behaviourism. Psychology and the Minds of Animals*. Cambridge University Press, Cambridge, UK

Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands

Hull C L 1943 *Principles of Behavior. An Introduction to Behavior Theory*. Appleton Century Crofts, New York

O'Donnell J 1985 *The Origins of Behaviorism. American Psychology, 1870–1920*. New York University Press, New York

Skinner B F 1938 *The Behavior of Organisms. An Experimental Analysis*. Appleton Century Crofts, New York

Skinner B F 1948 *Walden Two*. Macmillan, New York

Skinner B F 1969 *Contingencies of Reinforcement. A Theoretical Analysis*. Appleton Century Crofts, New York

Smith R 1997 *The Fontana History of Human Sciences*. Harper Collins, London

Thorndike E L 1898 Animal intelligence: an experimental study of the associative processes in animals. Monograph Supplement No 8. *Psychological Review* **2**

Tolman E C 1932 *Purposive Behavior in Animals and Men*. Century, New York

Watson J B 1913 Psychology as the behaviorist views it. *Psychological Review* **20**: 158–177

Watson J B 1914 *Behavior: an Introduction to Comparative Psychology*. Henry Holt, New York

Watson J B, Watson R R 1928 *The Psychological Care of an Infant Child*. Norton, New York

F. Parot

Belief, Anthropology of

'Belief,' along with 'believe, beliefs' and 'belief systems,' has served as a kind of 'odd job' word for anthropologists: a word commonly used for the analysis of a society's culture, religion, or ideas about the world, but seldom defined or explicitly theorized. Belief is often included in omnibus definitions of culture, such as Tylor's classic 'Culture, or civilization, … is that complex whole, which includes knowledge, belief, art, law, morals, custom, and any other capabilities and habits acquired by man as a member of society' (1871, from Kroeber and Kluckhohn 1952 p. 43) or Sapir's 'culture, that is, … the socially inherited assemblage of practices and beliefs that determines the texture of our lives … (1921, Kroeber and Kluckhohn 1952 p. 47). The term often refers explicitly to religion, as in Tylor's 'minimum definition of religion' as 'the belief in Spiritual Beings' (1873), but is used equally for claims people make about the empirical world or their social institutions.

Accounts of central tenets of culture or world-view in a society are often reported as beliefs. Lawrence's (1964) discussion of cargo cults in New Guinea is typical: 'The Ngaing believed that their world was brought into being by their deities' (p. 16); 'all these peoples believed in a creator God' (p. 21); 'the natives believed they could obtain cargo … largely by ritual' (p. 235). As Needham (1972 p. 3) points out, Lawrence did not find it necessary to define belief or describe how he knows 'that men in New Guinea believe what they say'; he simply assumed that this 'psychological category in the English language denotes a common human capacity.' This is typical of anthropological writing. Despite its ubiquity, 'belief' seldom appears in the Index of ethnographies, and the *International*

Encyclopedia of the Social Sciences (Sills 1968) has no entry for Belief, simply referring the reader to '*See* Attitudes; Ideology; Myth and Symbol; Religion; Values.'

Although there is a long history of analyzing other societies through a systematic description of their 'beliefs and behaviors,' recent ethnographic accounts have reflected ironically on the meaning of belief. 'Take an ethnographer,' writes Jean Favret-Saada (1980 p. 4):

> She has spent more than thirty months in the Bocage in Mayenne, studying witchcraft … 'Tell us about the witches,' she is asked again and again when she gets back to the city … confirm that *out there* there are some people who can bend the laws of causality and morality, who can kill by magic and not be punished; but remember to end by saying that they do not really have that power: they only believe it because they are credulous, backward peasants … No wonder that country people in the West are not in any hurry to step forward and be taken for idiots …

Favret-Saada concludes: 'To say that one is studying beliefs about witchcraft is automatically to deny them any truth: it is just a belief, it is not true.'

'Belief' is thus a ubiquitous term in much of anthropology, a common sense category used for the analysis of the tenets of culture or religion, but a term that is increasingly contested. A review of the meaning and history of the term belief is needed to make sense of this change in its anthropological usage.

1. *Defining 'Belief'*

The Oxford English Dictionary (1989) defines 'belief' as:

(a) The mental action, condition, or habit, of trusting to or confiding in a person or thing; trust, dependence, reliance, confidence, faith.

(b) Mental acceptance of a proposition, statement, or fact, as true, on the ground of authority or evidence …

(c) The thing believed; the proposition or set of propositions held true …

Hahn (1973 p. 208) attempts to clarify anthropological understandings of belief, defining beliefs as 'general propositions about the world (consciously) held to be true.' Tooker (1992 p. 808) points out that such anthropological conceptions of belief typically have two elements: 'propositionality' ('a mental state or conviction in which a doctrine or proposition concerning one's world-view is affirmed as true as opposed to false') and an assumption that such a 'prepositional relationship to tradition is an 'interiorized' one because of its reference to mental states.'

The *OED* indicates that the definition of belief as a proposition held to be true, a definition apparently taken over by anthropologists, rests on an older definition of belief as the condition of trusting in a person or thing, as 'faith.' At the same time, while a 'belief' is defined as a 'mental acceptance of a proposition … as true,' Favret-Saada and others suggest that labeling a proposition a 'belief' implies doubt or even falsehood, that 'studying beliefs about witchcraft is automatically to deny them any truth.' A review of the history of the semantics of 'belief' suggests that this complexity is grounded in changes in the meaning and use of the terms 'believe' and 'belief' in English (and their counterparts in other European languages).

2. *A History of 'Belief' in Euroamerican Language Use*

Anthropologists often talk with members of other societies about some aspect of 'their' world which does not exist in 'ours' and which anthropologists are comfortable asserting is not part of empirical reality. An examination of anthropological texts suggests that such phenomena—the witches of the Bocage or the Azande, the 'cargo' of the Papua New Guinea cults—are discussed in the language of belief, and that the beliefs of others are juxtaposed to 'our' knowledge of the empirical world (see Good 1994 Chap.1). Western civilization has long given great importance to 'beliefs,' with wars, church schisms, persecutions, and martyrdom revolving around Credos and demands for the renunciation of false beliefs. How is it that this same term, 'belief,' came to be so central to anthropological analysis, and what is implied by the juxtaposition of belief and knowledge?

The richest discussion of the history of the concept belief is to be found in two books by Wilfred Cantwell Smith, the historian of religion, which explore the relation between 'belief' and 'faith' historically and across religious traditions. Through historical and linguistic analysis, Smith comes to the startling conclusion that 'the idea that believing is religiously important turns out to be a modern idea,' and that the meaning of the English words 'to believe' and 'belief' have changed dramatically. Since around 1700, leading to misunderstanding of both the Christian tradition and the religious faith of others.

Smith demonstrates that the Old English words which evolved into the modern 'believe' (*geleofan*, *geliefan*) meant 'to belove,' 'to hold dear,' 'to cherish,' 'to regard as lief,' and reflect the Latin root *libet*, 'it pleases,' or *libido*, 'pleasure' (Smith 1977 pp. 41–6; Smith 1979 pp. 105–27). In Chaucer's Canterbury Tales, the words 'accepted by bileve' mean simply 'accept my loyalty; receive me as one who submits himself to you.' Thus Smith argues that 'belief in God' originally meant 'a loyal pledging of oneself to God, a decision and commitment to live one's life in His service' (Smith 1977 p. 42). Its counterpart in the medieval language of the Church and in the Credo was

'I renounce the Devil,' belief and renunciation being parallel and contrasting performatives or actions, rather than states of mind. Today, however, Smith argues, the ritual acclamation 'I believe in God' means for many people 'Given the uncertainty as to whether there be a God or not, as a fact of modern life, I announce that my opinion is 'yes, I judge God to be existent' (1977 p. 44).

Smith argues that this change in the language of belief can be traced in the grammar and semantics of English literature and philosophy, as well as popular usage. Three changes serve as indicators of the changing semantics of the verb 'to believe.' First, Smith finds that grammatically, the *object* of the verb 'to believe' shifted from a person (whom one trusted or had faith in), to a person and his word (his virtue accruing to the trustworthiness of his word), to a proposition. This latter shift began to occur by the end of the seventeenth century, with Locke and was firmly represented by the mid-nineteenth century in John Stuart Mills' philosophy (Smith 1977 p. 48). The twentieth century has seen a further shift, as beliefs have come to mean 'presuppositions,' and it was in this sense that 'belief systems' became nearly synonymous with 'culture' in some strands of anthropology (see Black 1973).

A second shift has occurred in the *subject* of the verb 'to believe,' from an almost exclusive use of the first person—'I believe'—to the predominant use of the third person, 'he believes' or 'they believe.' In anthropology, the impersonal 'it is believed that' parallels the discussion of culture as belief system or system of thought. This change in subject subtly shifts the nature of the speech act involved—from existential or performative ('I believe') to descriptive ('the Ngaing believe that').

Third, Smith observes that an important and often unrecognized change has occurred in 'the relation of belief to truth and knowledge,' as these are historically conceived. Bacon wrote in 1625 of 'the belief of truth,' which he defined as the 'enjoyment of it.' Belief maintains its sense here of holding dear, of appropriating to oneself that which is recognized as true. By the nineteenth century, however, 'to believe' had come to connote doubt, and today it suggests outright error or falsehood. Knowledge requires both certitude and correctness; belief implies uncertainty, error, or both. Smith's favorite illustration of the juxtaposition of belief and knowledge is an entry in the Random House dictionary which defines 'belief' as 'an opinion or conviction,' and at once illustrates this with 'the belief that the earth is flat'! It is virtually unacceptable in popular English usage to say that members of some society 'believe' the earth is round; if this is part of their world-view, then it is knowledge, not belief.

Smith goes on to argue that failure to recognize this shift in meaning has led to mistranslation of texts in the Christian tradition and ultimately to 'the heresy of believing,' the deeply mistaken view that belief in this modern sense is the essence of the religious life rather than faith. *Credo*, in the Latin, is literally 'I set my heart' (from Latin *cordis* or heart). *Credo in unum Deum* was correctly translated in the sixteenth century as 'I believe in one God,' when it meant 'I formally pledge my allegiance to God,' Whom we of course all acknowledge to be present in the world. Today, it is a mistranslation, suggesting that the Credo consists of propositions the veracity of which we assert. This is historically inaccurate and profoundly misrepresents the traditional ritual acclamation. Equally importantly, for Smith, the misplaced focus on belief as the primary dimension of religious life has led to mistranslations and misunderstandings of other religious traditions. It is this claim that has special relevance for anthropology.

3. *'Beliefs' and 'Believing' in Diverse Anthropological Traditions*

The term 'belief' functions quite differently in diverse anthropological traditions. A brief review of several traditions will indicate why 'belief' has emerged as a contested term in recent anthropological writing.

3.1 *The Intellectualist Tradition and Rationality Debates*

'Beliefs' and 'rationality' are central terms in the Intellectualist tradition most closely associated with E. B. Taylor and J. G. Frazer, as well as a series of 'Neo-Tylorians' and a set of rationality debates in the 1970s and 1980s. Tylor and Frazer set out to trace the evolution of beliefs from our 'rude forefathers' to the present, investigating forms of rationality, errors in reasoning, and the history of the evolution of culture from magic to religion and the emergence of science. An example from Frazer's Gifford Lectures, published as *The Belief in Immortality* (Frazer 1913), will illustrate the logic of this program, as well as the evolutionist language that led to its rejection.

> … the question of the validity or truth of religious creeds cannot, perhaps be wholly dissociated from the question of their origin. If, for example, we discover that doctrines which we had accepted with implicit faith from tradition have their close analogies in the barbarous superstitions of ignorant savages, we can hardly help suspecting that our own cherished doctrines may have originated in the similar superstitions of our rude forefathers; … The doubt thus cast on our old creed is perhaps illogical, since even if we should discover that the creed did originate in more superstition, … this discovery would not really disprove the beliefs themselves, for it is perfectly possible that a belief may be true, though the reasons alleged in favor of it are false and absurd …

Beliefs in this tradition were investigated as rational propositions and forms of explanation for natural and

religious phenomena. The intellectualist paradigm was grounded in philosophical empiricism, and judgements concerning validity of propositions and the evolution of explanatory frames from those which were false, even absurd, to those which correctly represent the natural world were essential to the analysis.

British structural-functionalism explicitly rejected the search for origins, the evolutionary claims, and the focus on individual beliefs and their rationality as a basis for social analysis. Nonetheless, classic texts, such as Evans-Pritchards' *Witchcraft, Oracles and Magic among the Azande* (1937) maintained the basic structure of inquiry, investigating witchcraft as a form of rational explanation for misfortune and distinguishing those explanations which attribute 'supra-sensible qualities' to phenomena (described as beliefs) from those which accord to contemporary science (described as knowledge) (cf. Good 1994 pp. 11–3).

In the 1970s and 1980s, the intellectualist tradition was revived in British anthropology under the rubric of Neo-Tylorianism and the rationality debates (e.g., Wilson 1970, Hollis and Lukes 1982). Reacting to the perceived vacuousness of functionalist explanations, a group of anthropologists returned to investigations of the rational structure of non-Western religious systems and their similarities to science as systems of explanation for the natural world. At the heart of these debates was a central question: 'when I come across a set of beliefs which appear *prima facie* irrational, what should be my attitude toward them?' (Lukes 1970 p. 194). For example, 'what does one do with the exotic utterances of primitives, who may say such things as "my brother is a cockatoo,"' asks Turner (1979 p. 401). 'Some things, such as a flat contradiction, could not be "believed,"' in any useful sense of "believe." … Can the fact that these utterances cannot be translated into anything that we would call "rational" warrant a claim that, e.g., these primitives possess a non-logical mentality?' 'Beliefs' in this paradigm were understood as rational propositions, held to be true by members of a society, and analysis focused on the identification, translation, and rational structure of belief systems, examining characteristics of systems 'open' to falsification and thus to change, in contrast to 'closed systems,' Which remain impervious to contradictory data that would lead to change and evolution in the direction of more rational explanation.

3.2 Structural Functionalism

Reacting against intellectualists' search for origins and focus on individual states of minds, Radcliffe-Brown and generations of British anthropologists placed investigations of the structure of social institutions at the center of anthropological theorizing. In this context, individual beliefs are of little interest, and a focus on the relationship between rational propositions and the natural world is replaced by an interest in the structural relation between the symbolic and the social order. Leach made this explicit: 'To ask questions about the content of beliefs which are not contained in the content of ritual is nonsense … Ritual action, and belief are alike to be understood as forms of symbolic statement about the social order' (1954 p. 14).

3.3 The Boasian Tradition

Inheritors of the German philosophical romantics' theorizing about culture and its *Geist*, American anthropologists of the Boasian tradition were more interested in the 'spiritual possessions of the group rather than the individual' (Sapir 1924), in the role of language and culture in constituting phenomenal experience and the 'behavioral environment' (Hallowell 1955) of members of a society. Though the term belief appears in this tradition, it has little theoretical salience. Sapir notes this explicitly: ' … the cultural conception we are now trying to grasp aims to embrace in a single term those general attitudes, views of life, and specific manifestations of civilization that give a particular people its distinctive place in the world. Emphasis is put not so much on what is done and believed by a people as on how what is done and believed functions in the whole life of that people, on what significance it has for them' ([1924] 1970 p. 83).

3.4 Cognitive Anthropology

The term 'belief' appeared as a core analytic category in American anthropology beginning in the 1950s among anthropologists intent on creating a new, more scientific paradigm for studying culture, who gathered under the banner of 'The New Ethnography,' ethnosemantics, or cognitive anthropology. For cognitivists, 'cultures … are not material phenomena; they are cognitive organization of material phenomena. … The object of study is not these material phenomena but the way they are organized in the minds of men (Tyler 1969 p. 3). In this tradition, culture is conceived as generative models, and beliefs are those cognitive structures which take propositional form: 'Beliefs are propositions about the relations among things to which those who believe have made some kind of commitment … for pragmatic or emotional reasons' (Goodenough 1990 p. 597). Although 'beliefs' assume a core analytic function in this tradition, many cognitivists share more with the Boasian tradition than with the intellectualist tradition, in that they are interested in investigating how semantic categories and linguistic structures shape perception of the phenomenal world rather than in normative claims about rationality. (They differ sharply in methodology and forms of ethnographic writing.) 'Beliefs' and 'knowledge' are often used interchangeably in this

tradition, rather than explicitly juxtaposed (see Black 1973). Nonetheless, the term 'belief' often subtly connotes counter-factuality in this tradition as in others.

3.5 Medical Anthropology, Public Health, Health Psychology

Studies of health beliefs was a key element of early forms of medical anthropology, particularly those devoted to applied work in the public health field, though this has not been true in recent years (see Good 1994). However, analyses of beliefs about illness etiologies, the risks of particular behaviors, or the benefits of particular treatments are still ubiquitous in public health work and health psychology. Popular or folk health beliefs are explicitly juxtaposed to medical knowledge, and educational campaigns are devoted to correcting mistaken ideas with the hope this will produce more rational behavior.

4. Critical Reflections

The language of belief has been widely used for ethnographic descriptions of culture, dating back to at least nineteenth century. In some anthropological traditions it has been a key analytic concept; in others, it has been used infrequently and in a common sense way. In some of the behavioral sciences, 'beliefs' continues to be used with little self-consciousness. However, in each of these domains, 'beliefs' are explicitly or implicitly juxtaposed to 'knowledge,' and the use of 'believe' or 'beliefs' connotes assertions to be counter to what is 'known' about empirical reality. This is particularly true in fields such as medicine, where the knowledge claims of the natural sciences run headlong into the historicist convictions of anthropology (Good 1994).

In recent years, the claims implicit in this formulation have been sharply challenged. The representation of others' culture as 'beliefs' ('their' beliefs and 'our' knowledge) authorizes the position and knowledge claims of the anthropological observer. Such authority has been challenged by a variety of forms of critical theory—post-structuralism, feminist writing, post-colonial theorizing, and subaltern studies. The place of the ethnographer as objective scientific observer, both in research and in ethnographic texts, is increasingly contested, and along with it, the use of the language of 'belief.' At the same time, developments in the history, philosophy, and sociology of sciences challenges any views of science as a progressive, normative 'mirror of nature,' the arbiter between belief and knowledge. As a consequence, the labeling of others' assertions about the world as 'beliefs' is increasingly challenged as pejorative, as failing to recognize the 'local' character of knowledge, and as hegemonic. The very use of the term is thus increasingly contested.

See also: Attitudes and Behavior; Attitude Formation: Function and Structure; Boas., Franz (1858–1942); Cognitive Anthropology; Cognitive Archaeology; Evans-Pritchard, Sir Edward E (1902–73); Religion, Sociology of; Ritual; Values, Anthropology of

Bibliography

Black M B 1973 Belief systems. In: Honigmann J J (ed.) *Handbook of Social and Cultural Anthropology*. Rand McNally College Publishing Co., Chicago, pp. 509–577

Favret-Saada J 1980 *Deadly Words. Witchcraft in the Bocage*. Cambridge University Press, Cambridge, UK

Frazer J G 1913 *The Belief in Immortality*, Vol. 1. Macmillan and Co., London

Good B J 1994 *Medicine, Rationality, and Experience: An Anthropological Perspective*. Cambridge University Press, Cambridge, UK

Goodenough W H 1990 Evolution of the human capacity for beliefs. *American Anthropologist* **92**: 597–612

Hahn R A 1973 Understanding beliefs: an essay on the methodology of the statement and analysis of belief systems. *Current Anthropology* **14**: 207–29

Hollis M, Lukes S 1982 *Rationality and Relativism*. Basil Blackwell, Oxford, UK

Kroeber A L, Kluckhohn C 1952 *Culture: A Critical Review of Concepts and Definitions*. Papers of the Peabody Museum of American Archaeology and Ethnology, Harvard University, Vol. 47, No. 1. Cambridge, MA

Lawrence P 1964 *Road Belong Cargo*. Manchester University Press, Manchester, UK

Lukes S 1970 Some problems about rationality. In: Wilson B R (ed.) *Rationality*, pp. 194–213

Needham R 1972 *Belief, Language, and Experience*. University of Chicago Press, Chicago

Sills D 1968 *International Encyclopedia of the Social Sciences*. Macmillan Co. and The Free Press, New York

Simpson J A, Weiner E S C 1989 *The Oxford English Dictionary*. Clarendon Press, Oxford, UK

Smith W C 1977 *Belief and History*. University of Virginia Press, Charlottesville, VA

Smith W C 1979 *Faith and Belief*. Princeton University Press, Princeton, CT

Tooker D E 1992 Identity systems of Highland Burma: belief, Akha Zan, and a critique of interiorized notions of ethno-religious identity. *Man* **27**: 299–819

Turner S P 1979 Translating ritual beliefs. *Philosophy of the Social Sciences* **9**: 401–23

Tyler S 1969 *Cognitive Anthropology*. Holt, Rinehart and Winston, New York

Tylor E B 1873 *Primitive Culture*, 2 Vols. John Murray, London

Wilson B R 1970 *Rationality*. Harper and Row, New York

B. J. Good

Benedict, Ruth (1887–1948)

A US cultural anthropologist, Ruth Fulton Benedict is best known for two books, the 1934 *Patterns of Culture* and the 1946 *The Chrysanthemum and the*

Sword. Each contributes importantly to the central premises of 'culture and personality' studies in US anthropology. Both books demonstrate the close connection between descriptive intensity and ethnographic reliability. Both, too, treat the theories and methods of a science in terms of their significance for ensuring human rights in any society. *Patterns of Culture* gave a wide audience a new concept of culture: an integrated and distinctive whole that molds the temperaments and talents of individuals. *The Chrysanthemum and the Sword* penetrated into the mechanisms through which 'culture' and 'personality' reinforce and replicate a dominant ethos. Both books assert the importance of individual creativity, a bow to the humanitarian impulse Benedict never lost.

1. *The Life*

Ruth Fulton was born on June 5, 1887, in New York City, the daughter of a Vassar graduate and a doctor of medicine. Two years after her birth, her father died and the vision of his lost opportunity for serving humankind haunted her for the rest of her life (Modell 1983). After a childhood spent moving from city to city, in 1905 Ruth and her younger sister Margery entered Vassar College. After graduation, Ruth returned to her mother's house in Buffalo, where she worked for the Charity Organization Society of New York. The experience increased her sensitivity about her partial deafness and exposed her to the struggles of immigrant families in an industrial city. Dissatisfaction persisted, unresolved by teaching in a private school and living with Margery and her husband Robert Freeman in Pasadena, California. A marriage proposal from Stanley Rossiter Benedict seemed to solve all problems; like other women of her generation, Ruth hoped the role of wife would fulfill her ambitions.

A month after their marriage, World War I broke out. The crisis accentuated Benedict's despair at the limitations set for women, nonconformists, and outsiders in her milieu. She tackled the problem through essays on three 'rebellious' women, Mary Wollstonecraft, Margaret Fuller, and Olive Schreiner; never published, the manuscript reveals the writer's impatience with the conventions that restricted her own options. In 1919, Ruth Benedict enrolled in courses at the New School for Social Research, where encounters with sociologist Elsie Clews Parsons and anthropologist Alexander Goldenweiser transformed her thinking. A meeting with Franz Boas at Columbia University completed the process, and in 1921 Benedict enrolled in his anthropology program. A remarkable two years later, she received her Ph.D. for a dissertation on 'The Concept of the Guardian Spirit in North America.'

During the 1920s, Ruth Benedict viewed her commitment to the discipline in the light of her quest for identity as a woman in US society. Personal concerns provided an impetus for anthropological inquiries, while at the same time detouring her from a conventional academic career path. During this period, too, the strains between her and Stanley increased, and in 1930 they agreed on a formal separation. By then Benedict had formed strong bonds with her student Margaret Mead and with Edward Sapir, her colleague. Conversations about the poetry all three wrote prompted Benedict to prove that anthropology could incorporate the voice she had developed as a poet. Her esthetic principles contributed as much to her anthropology as the advice about culture and person offered by Boas.

Ruth Benedict took on teaching duties at Columbia, trained students for the field, and in 1924 assumed editorship of the *Journal of American Folk-Lore*, a position she held until 1940. Her marginality to success in academia became apparent when, after Boas retired in 1937, she did not receive the headship of Columbia's Anthropology Department. The appointment of the anthropologist, Ralph Linton, surprised her colleagues, seeming to many to indicate the continued discrimination against women in US society. In reaction to Linton's appointment, to her abiding restlessness, and to her equally abiding belief in the scientist's duty as citizen, Ruth Benedict moved to Washington to serve her government during the war. She died on September 17, 1948, in New York City, where she had returned. She was 61 years old and in the midst of formulating a major theory of self and society.

2. *The Anthropolosy*

The 1934 publication of *Patterns of Culture* placed Ruth Benedict in the forefront of American anthropology. The book pulled together ideas she had developed in earlier pieces: the significance of coherence and of integrating customs, the social construction of marginality and deviance, the uses of a comparative method for establishing a relativistic stance. In the book's five discursive and three ethnographic chapters, the anthropologist set the foundation for her theoretical and methodological contribution to contemporary social sciences.

The book is best remembered for the notion that 'culture is personality writ large.' The theoretical claim receives support in the ethnographic chapters, which treat three distinct cultures as unitary, integrated, and patterned entities with characteristic personalities. Benedict's concept of culture did not exclude individuals, but explained the links between person and group. The links, she wrote, could be harmonious, leading to the 'successful' individual, or they could be rough, driving an individual into nonconformity, marginality, or madness. Three cultures form the triptych for an application of her theory, the Pueblos

of the American South West, the Dobu Islanders, and the northwest coast Kwakiutl. Each of these cultures receives a 'tag,' the famous Apollonian, paranoiac, and megalomaniac designations that for many readers obscured Benedict's concern with the processes by which a cultural configuration comes into being, persists over generations, bends to change, or collapses under crisis.

Methodologically *Patterns of Culture* introduces interpretive techniques that remain significant to anthropological analysis today. Ruth Benedict's reluctance to spend time in the field went hand in hand with an imaginative reading of available evidence from disparate details of behavior reported to her by informants, stored in ethnographies, or enshrined in items of material culture. Constant reference to US culture further extends her method, drawing comparison away from a simple juxtaposition of traits and towards the assessment of diverse 'arrangements' that became central to her work of the next two decades. In *Patterns of Culture*, Benedict shows brilliantly how the social scientist can maintain relativism in description while taking a stand on the implications of an 'arc' of traits (1934, p. 24).

Over the next decade, she published articles on classic anthropological topics in the 1933 *Encyclopedia of the Social Sciences* and simultaneously probed into the causes and consequences of the Great Depression for a general audience. In periodicals like the *American Scholar* and the *Partisan Review*, she brought an anthropological perspective to the subjects of, for example, unemployment, the 'problem of youth,' and the value of freedom of speech. Drawing on her discipline's comparative method, she mounted a stern critique against US culture, where human resources were wasted by narrow-minded institutions and persons whose capacities fell out of the conventional mold shunted aside. Professional and popular pieces preached an enlightened social engineering that drew lessons from cultures in which no individual was denied social participation.

Benedict's inquiry into the patterns of American culture led her to take on issues of race and class—related, in her framework, to democracy and social justice. Consequently, she examined the spread of totalitarianism under the lens of a comparative analysis of institutions and ideologies that deny the humanity of selected groups. From her perspective, understanding the rise of Nazism and Fascism meant discovering the conditions under which individual citizens detect no disjunction between their commonplace assumptions and the dictates of a political ideologue.

On the eve of World War II, Ruth Benedict reconsidered the implications of cultural relativism and its nonjudgmental stance. Like her mentor Franz Boas, she had argued that every culture should be treated in terms of its own values and purposes. By the end of the 1930s, she sought a scheme for evaluating cultures that respected difference while disclosing the danger some arrangements clearly posed. Danger came in terms of the suppression of individual creativity, fulfillment, and satisfaction; danger came, more dramatically, in acts of murder and genocide. Reacting to world events, Benedict fashioned the material for a theory of self and society in three separate endeavors: an article on childrearing, a book on race and racism, and a series of lectures on the theme of 'synergy.'

In 1938 Ruth Benedict published a seminal piece on childrearing, entitled 'Continuities and Discontinuities in Cultural Conditioning.' The article appeared in *Psychiatry*, a journal that fused psychiatry and social science. The piece reveals Benedict's ambivalence towards Freudian theory and her conviction that social conditions provide the basis for individual well-being. Her argument then sounds commonplace now: cultures should foster consistency between childhood and adulthood, so that traits learned early are appropriate to behaviors required later. At every turn, 'Continuities' criticizes US culture where, the anthropologist argued, sharp discontinuities between childhood and adulthood produce the 'neurotic personality' Karen Horney described. Under ideal conditions in any culture, Benedict wrote, children are conditioned into responsible social participation, their capacities fully utilized.

Simultaneously Benedict chafed under an invitation she had accepted to lecture on 'the individual and society' at Bryn Mawr College in February 1941. Right up her alley, the forum challenged her to delineate systematically the relationships between individual capacity, favored temperaments, social institutions, and cultural values. By then, too, the anthropologist explicitly recognized the urgency of calculating 'the consequences for human life of different social inventions' (1970, p. 322).

Only notes remain from these crucial and complicated lectures to show how far Benedict had moved from *Patterns of Culture*. The move took three forms: greater attention to the psychodynamics of personality development; an acute analysis of the institutions through which individual behaviors are organized; an effort to develop a plan for social change that acknowledged the significance of customary beliefs and behaviors. She chose the word 'synergy' to represent the effective working together of parts, and she compared low synergy with high synergy societies. The lecturer praised 'high synergy' societies, in which cooperation is at a maximum and all acts are mutually reinforcing. In such societies, individuals achieve a strong sense of responsibility and well-being. The unanswered question appears as well: what factors produce high synergy, which low synergy in a society? Her wartime writings pushed her closer to an answer.

A third endeavor refined Benedict's theory of self and society. In the 1940 *Race: Science and Politics*, Benedict analyzed discrimination and racism in her own society. She mounted her strongest attack against

a presumed scientific fact of innate endowment, the clothing for prejudices that condemned 'Negroes' (in the language of the times) to marginal participation in US society. Using comparative examples as well as historical references, she accused her country of fostering divisiveness, crushing the will of individuals, and undermining its own stamina.

The book acknowledged a debt to Boas in insisting that anthropology turn its methods and theories to the solution of contemporary problems. War made the obligation crucial for Benedict, and in the winter of 1942 she assumed a position at the Office of War Information. Obituaries she wrote when Boas died, in December 1942, convey the viewpoint that brought her into the government: 'He never understood how it was possible to keep one's scientific knowledge from influencing one's attitudes and actions in the world of affairs' (quoted in Modell 1983, p. 274). In the OWI, Benedict pursued 'national character' studies; the last assignment put on her desk was 'Japan.'

'The Japanese were the most alien enemy the United States had ever fought in an all-out struggle.' So reads the first sentence of Benedict's 1946 publication, *The Chrysanthemum and the Sword*. The wartime study validates the place Benedict already had in her discipline and indicates the directions she would have pursued had she lived longer than two years after the book's publication. The study of Japan continues the themes she raised throughout the 1930s: how to ensure the well-being of members of diverse cultures without violating the integrity of each culture, how to turn anthropology's comparative method to the service of international organizations devoted to peace, how to exercise responsibility as a social scientist without transgressing the rules of evidence and neutrality. The failings of the book, like its strengths, can be attributed to the complexity of Benedict's goals.

Despite thickly textured descriptions, the book is less an ethnography of Japanese culture than an expansion of anthropological method and theory. Benedict turned the impossibility of doing fieldwork into a virtue, stressing the importance of interpretation over immersion. The anthropologist, she wrote, 'can draw up his hypotheses and get his data in any area of life with profit. He can learn to see the demands any nation makes, whether they are phrased in political, economic, or moral terms, as expressions of habits and ways of thinking which are learned in their social experience' (1946, p. 13). Her data are the 'commonplaces' any Japanese would recognize, the behaviors and beliefs an ordinary person deems natural.

The result is a persuasive display of the dominant 'patterns' of Japanese culture without the homogenizing stamp of her earlier efforts. Benedict shuns the isomorphism implied by 'culture and personality,' while recognizing the importance of consistency between 'society' and 'self' for maintaining high synergy. Japan in World War II was notably synergistic: every act and every attitude seemed to the observer-interpreter to flow harmoniously, the Emperor's commands coordinating with the motives and feelings of Japanese citizens. This was Benedict's test case, the high synergy society she had praised in 1941. The book explores the processes through which personal impulses, tightly constructed cultural demands, and the needs of a nation are channeled smoothly together, without frustration or conflict.

Critics of *The Chrysanthemum and the Sword* gloss over Benedict's delineation of process, and mistakenly consider the designation of Japan as a shame culture the book's major thesis. In the study of Japan, however, *emotion* is an analytic concept that allows the anthropologist to link the drives of individuals to the maintenance of social order and cultural consistency. 'Shame' pushed Benedict's ongoing exploration of self and society into its most radical formulation yet, as she demonstrated the delicate and daily ways in which persons internalize the pressures of their surroundings without experiencing lethargy, anger, suffocation, or despair. Chapters display the abundance of reinforcing activities and attitudes that make shame both a powerful sanction and a barely perceptible burden for the Japanese individual.

Ruth Benedict bequeathed to social science a theory of the social construction and the social uses of an emotion. She argues that mechanisms for instituting and fostering an emotional pattern crucially support the contemporaneous conditions in which a society finds itself; change in the conditions can render the emotion useless—a conclusion she reached about Japan, as the nation encountered devastating defeat in the war. She offers, too, the analogous proposition that an emotional pattern may waste the resources and erode the capacity for social action of members of a culture.

A comparison between Japan and the USA runs through the volume, permitting a sophisticated analysis of shame and guilt that anticipates later recognition by social scientists of the twinning of these emotions. At the same time, Benedict's concern with the collapse in Japanese society after the war biases the comparison, so that shame comes to seem an immature and confining emotion. On a theoretical level, however, her discussion of emotion is value free. She does not posit a hierarchy, in terms either of human or social development, but rather assesses the function of an emotional configuration as it releases or represses expression of human creativity—the humanism in her science.

An emphasis on the social construction of emotion for Benedict eliminated the inevitability of a Freudian model; if personal emotions are socially constructed, then they are alterable. Recognizing how much she had left undone, Benedict undertook several projects after the war, including childrearing studies for UNESCO, participation in the navy's 'cultures at a distance' project, and further comparative inquiry into emotional configurations.

3. The Accomplishment

Ruth Benedict's contribution to social science comes from her ability to combine critical theory with cultural relativism. She accomplished this by constructing a scheme for judging cultures that recognized the full scope of their unique purposes and traits. The scheme included inserting the 'human being' into an analysis of social structures in a way that foreshadows late twentieth- century notions of 'actor' and 'agency.' Moreover, Benedict never loses sight of the constituents of actor and agency: as *The Chrysanthemum and the Sword* demonstrates, the individual is composed at once of idiosyncratic drives and the shaping forces of a culture. Her attempts to explore systematically the conditions under which the ensuing tension is productive rather than destructive banished the simplistic 'culture and personality' equation. A founder of that subfield of US anthropology, Benedict pushed the theoretical formulations linking self and society toward the critical social theory of today's humanistic and social scientific disciplines.

Inasmuch as her comparative analyses have at core the human being—creative, cooperative, resistant, rebellious—they impose a rigorous critique on all cultures. The body of her work asks, 'does a particular social arrangement suppress or exploit diverse individual capacities?' The question pushes Ruth Benedict into the vanguard of critical social theory in two respects: she does not abstract the individual from the material conditions that frame social responsibility, and her emphasis on the diversity of purposes and temperaments in any setting anticipates contemporary rejection of the concept of a bounded society with a definable cultural core.

Ruth Benedict embraced cultural relativism while insisting that the anthropologist interpret cultures in terms of their human benefits. The findings of any science had to be devoted to improving the human condition: preventing war, ensuring peace, and, above all, granting anyone in any milieu access to resources and to rights. Stressing the importance of directed change, Benedict did not preach an imperialistic social engineering. She argued instead that the social scientist had unique skills with which to delineate the cultural values and social institutions upon which change must be based in order to last.

In Ruth Benedict's writings, humanism extends from theory to methods of inquiry and styles of presentation. In her view the social scientist had an obligation not only to investigate circumstances in which some were deprived of the means of survival but also to offer findings in an accessible form. Here again she makes a major contribution: writings that appeal to a general audience, to policy makers and to soldiers, to social scientists and to artists, presage a contemporary scholarship that refuses to remain in an ivory tower.

See also: Anthropology; Anthropology, History of; Cross-cultural Psychology; Cultural Relativism, Anthropology of; Culture as Explanation: Cultural Concerns; National Character; Racial Relations; Racism, History of; Relativism: Philosophical Aspects

Bibliography

Benedict R F 1934 *Patterns of Culture*. Houghton Mifflin, Boston, MA

Benedict R F 1938 Continuities and discontinuities in cultural conditioning. *Psychiatry* **I**: 161–7

Benedict R F 1940 *Race: Science and Politics*. Modern Age Books, New York

Benedict R F 1946 *The Chrysanthemum and the Sword*. Houghton Mifflin, Boston, MA

Benedict R F 1970 Synergy: Some notes of Ruth Benedict. Selected by Maslow A and Honigman J J *American Anthropologist* **72**: 320–33

Modell J S 1983 *Ruth Benedict: Patterns of a Life*. University of Pennsylvania Press, Philadelphia, PA

J. S. Modell

Bentham, Jeremy (1748–1832)

1. Life and Writings

Bentham's father was a prosperous attorney. Hoping that his precocious child would have an illustrious legal career, he had him educated at Westminster, the Queen's College, Oxford, and Lincoln's Inn (called to the Bar in 1769). For Bentham each of these three institutions came to epitomise aspects of the prevailing orthodoxy: conservatism, complacency, hierarchy, superstition, insincerity, and the elevation of authority over reason. As a child, he conceived a passionate antipathy to the inadequacies and injustices of English law and procedure. Later, 'English common law and equity seemed to Bentham archaic, uncodified, incomprehensible, arbitrary, irrational, cruelly vindictive, tortuously dilatory, and so ruinously expensive that nine out of ten men were literally outlawed' (Mack 1968, p. 55). He very soon abandoned legal practice in disgust. Unmarried, and comfortably off, he devoted himself to criticism and reform of legal, political, and social institutions on the basis of utility or 'the greatest happiness principle.'

Bentham's most important diversion from writing was his attempt to lobby the government for a contract to design and manage a model prison, the panopticon. His proposals to William Pitt in 1791 received some

encouragement, but then encountered repeated delay and opposition. The project failed, in the process nearly bankrupting him, alienating him from the establishment, harming his reputation with posterity, and pushing him towards democracy. He was eventually granted a sum in compensation for his efforts and expenses in 1813, but the project was effectively dead by 1803. Thereafter, he stayed mainly at home, or in various retreats, writing, conversing with his friends, acolytes and other visitors, and conducting an extensive international correspondence.

Bentham's life was relatively uneventful, but enormously productive. Nearly 70,000 pages of manuscripts survive, mostly in University College London, the British Library, and Geneva. By January 2000, 20 of a projected 68–70 volumes of the scholarly edition of *The Collected Works of Jeremy Bentham* had been published. Preliminary work had been completed on several more. Many of Bentham's extant writings are not yet accessible in a readable form, including important ones on sex, religion, fallacies, logic, and language, and many specific legal subjects.

The main outlines of Bentham's life and ideas are known and there have been some important specialized studies. However, the unworked manuscripts contain second thoughts, refinements, and new departures which may confirm the view that Bentham was less doctrinaire, more subtle, original, and genuinely radical than has traditionally been supposed. Revisionism in Bentham studies is at an early stage.

Bentham was a secular child of the Enlightenment. The principle of utility came mainly from Helvetius and he owed much to Bacon, David Hume, (see *Hume, David (1711–76)*). John Locke (see *Locke, John (1632–1704)*), Priestley and European thinkers, including Voltaire, Beccaria, and Charles Montesquieu (see *Montesquieu, Charles, the Second Baron of (1689–1755)*). But he differed from his predecessors in several respects. He insisted that 'the art-and-science' of legislation required as preliminaries exhaustive taxonomy based on precise concepts and exact analysis of social facts. He had an acute awareness of the importance and limitations of language as the main instrument for expressing abstract ideas, and applied his version of utility and relentlessly detailed analysis to a remarkably wide range of subjects. He adopted clear, often extreme, positions on some issues. For example, he maintained that the only criterion of good and bad, of right and wrong, is the principle of utility; all other criteria are either meaningless or perversions of utility or utility disguised or pure subjectivism; that it is easier to justify torture than punishment; that talk of natural rights is both nonsensical and dangerous; that no artificial rule of evidence can promote rectitude of decision; that the common law is not really law, but a fiction based on judicial usurpation of power; and that the interests of the legal profession and of those in power are in general opposed to those of the community.

Despite this seeming clarity, many of Bentham's ideas invite conflicting interpretations. 'Benthamic ambiguity' is an important thread in reading Bentham. For example, he is often treated as a forerunner of both *laissez faire* and the modern welfare state, yet he does not fit either; he was a genuine political radical, but he placed a high value on security, and on the protection of existing property entitlements; he favored both strong government and strong democratic control; he virulently rejected most features of the common law, yet defended some salient institutions such as orality and cross-examination; the great codifier opposed all binding rules of evidence and procedure. There is no more ambiguous symbol than 'the auto-icon,' Bentham's skeleton, fully clothed and topped with a wax head, that sits enigmatically in public in a cupboard in University College London, stimulating endless speculation about its possible meanings. Some seeming ambiguities can be resolved textually; some suggest honest and perceptive puzzlement; others hint that behind the benign exterior lay some deep inner conflicts.

2. *Bentham as Jurist*

Bentham explicitly aspired to be to jurisprudence what Luther was to Christianity and Newton was to science. His contributions to philosophy, political theory, and social reform were offshoots of his concern with law. Utility was a tool both for criticizing existing laws and institutions and guiding legislation. His ideas on linguistic analysis, on evidence and inference, and on democracy grew out of specifically juristic concerns.

Jurisprudence is the most extensive, original, and neglected part of Bentham's legacy. Yet he is better known as a philosopher, political theorist, and social reformer, for three reasons: few Bentham scholars have been jurists; many of the legal writings have only recently started to become accessible; and Bentham's disciple, John Austin, was for a century treated as 'the Father of English Jurisprudence.' Austin's command theory of law is inferior to Bentham's, but Austin was preferred because he was more sympathetic to the common law, less critical of the legal profession, and less dangerously radical than his master. His general ideas were also easier to grasp and to criticize.

Bentham's science of legislation is based on five main pillars: the greatest happiness principle, the theory of fictions, his positivist command theory of law, the pannomion, and, as a late but crucial addition, his constitutional and democratic theory.

2.1 *Bentham as Utilitarian: The Greatest Happiness Principle*

Bentham's best-known expositions of utility are in two early works, *A Fragment on Government* (1977, here-

after *Fragment*), first published in 1776, and *An Introduction to the Principles of Morals and Legislation* (1996, hereafter *IPML*), first published in 1789. The former is primarily an attack on the complacency of William Blackstone's idealization of the common law and his espousal of natural law and social contract theory at the start of his *Commentaries*. *IPML* was conceived as a prolegomenon to a penal code dealing especially with the classification of offences and the principles of punishment. The opening words of *IPML* are: 'Nature has placed mankind under two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do.'

As a utilitarian Bentham is generally regarded as more representative than original. The rhetorical opening of *IPML* is often treated as epitomizing classical utilitarianism, a vehicle for considering standard criticisms on, for example, promising, punishment of the innocent, torture, and egoism. However, not all such objections apply to Bentham. For example, pleasure refers to all human values, not just to sensual or hedonistic pleasures or to wealth; the felicific calculus was not intended as an exact measure, but rather as providing a checklist of factors relevant to weighing the costs and benefits of any action; every individual equally counts for one, but the ultimate test is aggregate happiness (i.e., the general welfare). 'Extent' ('whose happiness is to be taken into account?') applies not just to humans, but to all sentient beings, for in respect of animals 'the question is not, Can they *reason*?, nor can they *talk*?, but can they *suffer*?' (*IPML* Chap. XVII, pp. 282–6)

The first six chapters of *IPML* are not fully representative of Bentham's utilitarianism, as it developed over time. They are pedagogically convenient, because they are easy to read and to criticize. Here Benthamic ambiguity seems to begin at the very root of his thought: Does pleasure refer to satisfaction or to preference? Was he a direct or an indirect utilitarian? How different was Bentham's utilitarianism from Hume's? Does rooting utility in human nature make Bentham a closet natural lawyer? Do Bentham's utilitarian actors maximize their own pleasure or is every human choice, in individual morality and collective decision making, to be judged by its tendency to maximize the general welfare? Is the primary concern of sovereigns, their subordinates, and individuals rightly confined to one's own community, subject to certain exceptions?

Such questions are much debated. It is dangerous to impose on Bentham ideas that emerged in later debates about utilitarianism, for example, modern distinctions between act-utilitarianism and rule-utilitarianism, between preference and satisfaction, or between aggregate and average utility. Bentham's ideas developed over time and he was not always consistent. He might also have rejected sharp versions of these distinctions as false dichotomies.

Bentham's utility is a principle of both individual psychology and of social and political morality, of morals and legislation. For the jurist the latter is both more important and less vulnerable to criticism. Humans generally seek to maximize their own interests, and the primary task of legislation is to harmonize individual interests with the general welfare by a system of rewards and punishments. Similarly, the interests of the governors need to be harmonized with the interests of the governed, especially through systematic design and implementation of securities against misrule. Bentham rejected any idea of principles independent of utility, but gave an important place to four principles subordinate to utility: security, subsistence, abundance, and equality. Of these security, or settled grounds of expectation, was the principal one because it implies extension in point of time and is a precondition for achieving the other ends of government. Bentham gave a high priority to both property and order, and hence for liberty in the sense of absence of coercion. Security is also the basis for the important motto of the good citizen, 'To obey punctually; to censure freely' (*Fragment*, Preface). But security can be overridden by utility in given circumstances. Security and subsistence generally have a higher priority than equality, but in the calculus each individual counts for one, the idea of diminishing marginal utility serves a distributive function, and Bentham believed that as a psychological fact 'the nearer the actual proportion approaches to equality, the greater will be the total mass of happiness.' (*Civil Code*; Harrison (1983, pp. 243–50)).

The originality and interest of Bentham's utilitarianism lie mostly in his *applications* of the basic ideas. Utility gives coherence and bite to his detailed treatment of evidence, classification of crimes, the rationalization of punishment, and numerous other specific topics. His utilitarianism should be judged as much by its potential for illumination as an analytical and critical tool applied to concrete issues as by its philosophical defensibility.

2.2 Fictions

Bentham's theory of logic and language anticipated key later developments in English analytical philosophy, including the idea that sentences rather than words are the primary unit of meaning. To clarify basic abstract legal conceptions, such as 'obligation,' 'right,' and 'property,' he substituted for the traditional definition *per genus et differentiam* the technique of 'paraphrasis' and 'phraseoplerosis.' Abstract nouns with no direct counterpart in the world of fact are names of fictitious entities. They can be elucidated by placing them in a whole sentence, such as 'X has a right against Y,' and then asking under what conditions this sentence is true. It is true if Y has a duty. 'Y has a duty' is true if the sovereign of an existing legal system has commanded that there shall be a sanction

(pain) for nonperformance in given circumstances, and Y's situation is subsumed under this command. Thus the terms 'right' and 'duty' are linked through 'sovereign' and 'sanction' to pains and pleasures, things that can be directly experienced.

Bentham's theory of fictions contains his basic ideas on ontology, epistemology, and logic. We construct our knowledge of the world through the lens of language, itself a human construct, which is both necessary because it is our window on the world, and dangerous because it is a distorting lens. Language is the best tool we have for knowledge and understanding. This constructivist epistemology based on utility has affinities with later developments in pragmatism and the sociology of knowledge.

Bentham often used 'fiction' pejoratively. He attacked natural law, social contract theories, and technical legal fictions as dangerous mystifications. One example is his famous attack on natural rights as pestilential nonsense: to say that 'X has a right' logically implies some existing general rule or command. 'X has a legal right' presupposes a law commanded by a sovereign and backed by a sanction. Legal rights are creatures of positive laws. But since natural law does not exist nor do natural rights. The idea of natural rights is nonsensical, like the idea of a son without a father. To talk of imprescriptible (i.e., unalterable) rights is also meaningless, 'nonsense upon stilts,' for the sovereign's power cannot be legally limited. Since rights conflict, talk of 'absolute' rights is contradictory. Talk of nonlegal rights is also mischievous, raising expectations that cannot be fulfilled, inviting anarchy. Bentham's objections relate to the mode of discourse rather than to particular values. There may be moral claims and legal rights based on utility. He would not necessarily have opposed all modern legal bills of rights. His objection was to talk of nonlegal rights, not to the enactment into positive law of enforceable legal rights. A modern Benthamite can argue that talk of 'a right to food' or 'a right to work' refers to important aspirations, but begs questions about the allocation of correlative duties, about enforcement if not legally binding, and about the feasibility of implementing such 'rights.' They represent wishful thinking. As Bentham put it: 'Want is not supply, hunger is not bread.' Bentham started a line of analysis of 'rights,' 'powers,' and 'obligations' that was developed by Austin, Hohfeld (see *Hart, Herbert Lionel Adolphus (1907–92)*), Hart, and others. He still poses a challenge to loose rights talk.

2.3 Theory of Law

Bentham's legal theory operates on at least three levels: first, his general theory of law and its philosophical underpinnings; second, subtheories on, for example, codification and nomography, constitutional law, punishment and reward, and adjective law; third, the pannomion, a comprehensive body of codes of general application, subject to particular, essentially minor, modifications according to time, place and social context.

Bentham's jurisprudence starts with a strong positivist distinction between law as it is and law as it ought to be. 'To the province of the *Expositor* it belongs to explain to us what, as he supposes, the Law is: to that of the *Censor*, to observe what he thinks it ought to be' (*Fragment*, 397). Exposition can be particular or general; censorial jurisprudence is general. Bentham's main interest was in universal censorial jurisprudence.

The fullest account of Bentham's general theory of law *Of Laws in General* (1970 hereafter *OLG*), was first published in a definitive edition, edited by H. L. A. Hart, after Hart's *The Concept of Law* had become the leading text of English legal positivism, partly through its critique of Austin's imperative theory. Ironically, Hart conceded in a late essay that Bentham's version contained an important core of truth, 'the notion of a content-independent, peremptory reason for action' as part of the general notion of authority. (Hart 1982, Chap. 10, p. 261). *OLG* involved a deeper exploration of difficulties that he had dimly recognized earlier, especially the distinction between the civil and the penal branches of jurisprudence and individuation (what constitutes a single law?). The outcome was a dense, but incisive work, which includes original ideas on the logic of the will, a precursor of modern deontic logic.

Bentham elaborated his conception of law in terms later made familiar by John Austin: commands backed by threats made by a sovereign who is subject to a habit of obedience. Yet Bentham's ideas were subtly different. Bentham allowed for limited and divisible sovereignty, whereas Austin did not. The habit of obedience might be partial. Commands, such as judicial orders, need not be general; some laws can be backed by moral or religious sanctions, some laws by rewards. The idea of command has a more complex and limited role in Bentham's theory.

Burns, Hart and others have explored aspects of Bentham's jurisprudence in detail, including universal censorial jurisprudence; legal positivism and sovereignty; and his analysis of fundamental legal concepts, such as rights and obligations. Less abstract and less well known, but more extensive, are his theoretical writings on codification, adjective law (evidence and procedure), judicial organization, civil law, criminal law, indirect legislation, international law, and many other topics, whose importance is often underestimated. For example, Bentham's writings on evidence represent the most original and wide-ranging contribution to evidence theory in the Anglo-American tradition.

2.4 Codification and the Pannomion

From the early 1780s Bentham's main aspiration was to produce a complete body of laws—a pannomion—

to rationalize and systematize English law and, indeed, any legal system. At first he thought that the main weaknesses of English law were lack of organization, incompleteness, and inaccessibility, linked to the prejudice that legislation was an inferior form of law to common law—defects that might be remedied by a comprehensive Digest. By 1782 he was convinced of the necessity for a systematic structure of codes based on utility and satisfying the tests of coherence, completeness, and comprehensibility. The pannomion encompassed integrated codes on civil, penal, procedural, and constitutional law, backed by detailed legislation on specific topics.

He started with penal law, but was diverted by theoretical concerns into 'the metaphysical maze' of *OLG*. By the late 1780s he had completed important work on the general theory of law and legislation, but no actual code. Though codification remained his long-term aim, he only returned to the work of actually preparing codes much later. Starting in 1811 he offered his services as a codifier, first for several jurisdictions in the United States, then for Russia, Poland, and Geneva, and latterly for Spain, Portugal, Tripoli, and various Latin American countries. He now prepared some important general works on codification and started drafting several codes. His ideas excited considerable interest, but did not lead to the adoption of a single code. In 1830 he published the first volume of his magisterial *Constitutional Code* (1983). This, together with a procedure code, was only completed after his death by a former amanuensis, Richard Doane. Shortly before his death, Bentham was still 'codifying like any dragon.' He believed strongly that a code should be the product of a single mind, but his failure to complete a single draft reinforced the view that the pannomion was impracticable.

2.5 Constitutional Law and Democratic Theory

Bentham was not an instinctive democrat. When young, he was mainly concerned with how legislators should govern. He was prepared to deal with aristocrats, enlightened despots, revolutionaries, and democrats in promoting his ideas. His constitutional theory was a late development. The timing and reasons for his 'conversion to democracy' have been much debated. He was sympathetic to democratic ideas at the start of the French Revolution, but he only became involved in the movement for parliamentary reform after 1809, stimulated by his experience of general resistance and indifference to reform, especially the panopticon. Bentham firmly located sovereignty in the people. He recognized the problem of the tyranny of the majority, but he never satisfactorily solved it, perhaps because of his virulent antipathy to talk of nonlegal rights.

His elaborate constitutional design is based on three axioms: maximization of the happiness of all as the proper end of government; the maximization of the self-interest of the governors as the actual end of government in every political community; and the harmonization of the interests of the rulers and the ruled. His scheme involves vesting real and sustained power in the sovereign people to locate, monitor, and dislocate officials through a system of securities against misrule, especially publicity.

Several of the constitutional writings have only recently become available. The main elements in his code were 'expense minimized, aptitude maximized,' universal manhood suffrage (subject to a literacy test), secret ballots, frequent elections, a unicameral legislature, limited judicial power, accountability and transparency, and a power of recall of office holders, including the military, vested in the sovereign people. He supported ideas backed by other radical reformers of the time. In principle he favored votes for women, but considered it inexpedient to press for this. Bentham conceived of a constitutional code as a logical and engineered product of a single mind, based on articulated principles, and designed to provide popular control over misrule, patronage, and other forms of corruption.

3. Bentham and Social Science

Bentham did not draw sharp distinctions between jurisprudence, psychology, political economy, and other fields that have since become semiautonomous social sciences. Utilitarianism requires a concern for social facts. Legislation is the social science. To make calculations the legislator needs reliable information about the existing situation, the likely consequences of a given course of action, and causal relations. He emphasized 'the statistic function' of government; his stepbrother, Charles Abbott, played a leading role in the first population census (Population Bill, 1800); and Bentham himself tirelessly argued for better social data, especially regarding the Poor Laws. This Benthamite enthusiasm was satirized in Charles Dickens's *Hard Times* in Mr. Gradgrind, who believed that 'Facts alone are wanted in life.' Bentham believed no such thing. Utility was his guiding principle and he had a sophisticated awareness of the complexities of concepts such as 'fact' and 'causation.' To his credit, he was among those who first recognized and campaigned for the systematic collection of reliable social data as a necessary, but not a sufficient, condition for policy making and legislation.

On 'the art and science of political economy,' Bentham was a follower of Adam Smith, but criticized his defense of usury laws, and on some issues differed from him in either being more averse or more amenable to interventionism. Bentham saw the main role of government as promoting abundance indirectly through ensuring security, the basic presumption being against government involvement in economic affairs 'without some special reason.' But his proposals included provision for many areas later associated with the welfare state, such as public health, education,

and the relief of indigence. He argued for the emancipation of colonies both on economic and constitutional grounds and, rather than being a doctrinaire supporter of *laissez faire*, consistently applied cost–benefit analysis to particular issues. Bentham had some impact on the economic writings of James Mill, John Stuart Mill (see *Mill, John Stuart (1806–73)*, David Ricardo, and later W. S. Jevons, Alfred Marshall (see *Edgeworth, Francis Ysidro (1845–1926)*), and F. Y. Edgeworth (see *Edgeworth, Francis Ysidro (1845–1926)*). There is a useful, but incomplete, edition of Bentham's economic writings (1952–4).

One cannot do justice here to Bentham's psychology, which is regarded by some as his Achilles heel, nor to his extensive and detailed writings on education, poor law, social administration, and other aspects of social policy. There are some interesting affinities with Marx's (see *Marx, Karl (1818–89)*) ideas, for example, the clash between the interests of the few and the many (though with less emphasis on social class), interest-begotten prejudice, and demystification. But Marx, like Foucault, criticized Bentham without having studied him carefully. Marx treated him as the epitome of 'commonplace bourgeois intelligence.' Foucault interpreted the Panopticon as a symbol of repressive authoritarian control, ignoring the fact that Bentham's concern with transparency applied to those with power as well as those subject to it. Both images are, at best, half-truths. Bentham was also a pioneer in the study of bureaucracy. His standing as a social theorist is ripe for reappraisal.

4. *Influence*

Bentham first gained recognition through the efforts of a Genevan, Etienne Dumont (1759–1829) who between 1802 and 1828 produced five elegant French 'recensions' from manuscripts and printed texts. These were more lucid, succinct, and diplomatic than the originals. Some were translated into several languages, including English. The first, *Traité de legislation civile et penale* (1802) made Bentham's reputation in Europe, whence it spread to Latin America and eventually back to Britain.

In the late nineteenth century Bentham was hailed as the most influential reformer in English history. Subsequently there has been a revisionist debate about the relationship between his ideas and particular reforms. Bentham had both committed disciples and more pragmatic followers who became influential, including Romilly, Brougham, Chadwick, James Mill, and John Stuart Mill; his ideas contributed to the climate of reform in the nineteenth century both in Britain and internationally; many changes, for example in the Anglo-American law of evidence, moved in directions he had advocated, but in a more piecemeal fashion than he would have approved. He hated jargon, but coined neologisms. Some—codify, maximize, international—have become part of the language. Others, fortunately, have not. Many of his proposals, for instance on codification, have not yet been implemented. How far responsibility for specific changes should be attributed to him or his followers remains controversial.

Similar questions arise about his contemporary 'significance.' It is anachronistic to apply early nineteenth-century ideas directly to twenty-first-century issues. Yet, Bentham's views are still relevant to topics as varied as human rights, sovereignty, constitutional reform, penal policy, torture, corruption, patronage, transparency and accountability, freedom of information, political rhetoric, the legal profession, access to justice, inheritance, the free market, globalization, and 'the farther uses of the dead to the living.' On these and many other topics he is still worth consulting, not least because he often takes provocatively radical positions. As Herbert Hart (see *Hart, Herbert Lionel Adolphus (1907–92)*) remarked, 'where Bentham fails to persuade, he still forces us to think' (Hart 1982, p. 39).

See also: Constitutionalism; Economics, History of; Enlightenment; Justice: Political; Law and Democracy; Law: History of its Relation to the Social Sciences; Natural Law; Positivism, History of; Prisons and Imprisonment; Ricardo, David (1772–1823); Smith, Adam (1723–90); Utilitarian Social Thought, History of

Bibliography

Bentham J 1838–43 *The Works of Jeremy Bentham*, 11 vols. Bowring J (ed.) W. Tait, Edinburgh

Bentham J 1952–4 *Jeremy Bentham's Economic Writings*, 3 vols. Stark W. (ed.). Allen & Unwin, London

Bentham J 1970 *Of Laws in General*. Hart H L A (ed.). Oxford University Press, Oxford, UK

Bentham J 1977 *A Comment on the Commentaries and A Fragment on Government*. Burns J H, Hart H L A (eds.). Oxford University Press, Oxford, UK

Bentham J 1983 *Constitutional Code*, Vol. I. Burns J H, Rosen F (eds.). Oxford University Press, Oxford, UK

Bentham J 1989 *First Principles Preparatory to the Constitutional Code* [Schofield P (ed.)]. Oxford University Press, Oxford, UK

Bentham J 1995 *Colonies, Commerce and Constitutional Law: Rid Yourselves of Ultramaria and Other Writings on Spain and Spanish America*. Schofield P. (ed.). Oxford University Press, Oxford, UK

Bentham J 1996 *An Introduction to the Principles of Morals and Legislation*. Burns J H, Hart H L A (eds.). Oxford University Press, Oxford, UK

Bentham J 1997 *De l'ontologie et autres textes sur les fictions* 1997, P. Schofield P, Clero J-P, Laval C (eds.) Editions du Seuil, Paris

Bentham J 1968–81 *Correspondence*, vols. 1–5. Athlone Press, London

Bentham J 1984–2000 *Correspondence*, vols. 6–11. Clarendon Press, Oxford, UK

Dinwiddy J 1989 *Bentham*. Oxford University Press, Oxford, UK

Halevy E 1952 *The Growth of Philosophic Radicalism*, 2nd edn., Morris M (trans.). Faber and Faber, London
Harrison R 1983 *Bentham*. Routledge and Kegan Paul, London
Hart H L A 1982 *Essays on Bentham*. Oxford University Press, Oxford, UK
Kelly P J 1990 *Utilitarianism and Distributive Justice*. Oxford University Press, Oxford, UK
Mack M 1968 Bentham, Jeremy, in *International Encyclopedia of Social Science*, Vol. 2, pp. 55–8. Macmillan, New York
Parekh B (ed.) 1993 *Jeremy Bentham: Critical Assessments*, 4 vols. Routledge, London
Postema G J 1989 *Bentham and the Common Law Tradition*. Oxford University Press, Oxford, UK
Rosen F 1983 *Jeremy Bentham and Representative Democracy*. Clarendon Press, Oxford, UK
Twining W 1985 *Theories of Evidence: Bentham and Wigmore*. Stanford University Press, Stanford, CA

W. L. Twining

Bereavement

The term 'bereavement' is derived from the Latin word *rumpere* (to break, to carry, or tear away), and refers to the objective situation of a person who has suffered the loss of someone significant. 'Grief' is derived from the Latin *gravare* (to weigh down), and refers to the emotional experience of a number of psychological, behavioral, social, and physical reactions to one's loss. The word 'mourning' is derived from the Latin word *memor* (mindful). It refers to actions expressive of grief which are shaped by social and cultural mourning practices and expectations. Pointing to the timeless message of the original meanings of these terms, Jeter (1983) commented that 'as the ancients, people today surviving the death of a family member do feel robbed, weighted down, and are mindful of the past, knowing that life will never be the same' (p. 219). But how do individuals cope with such an experience? To address this question, the models and approaches that seem most influential to the current understanding of bereavement and grief are examined. The discussion begins with a review of the traditional views on bereavement, and then turns to the more recent developments in research and theorizing that have changed basic ways of looking at grief.

1. Traditional Views

Several different theoretical formulations have made important contributions to the current state of knowledge about loss and grief (for a more detailed review see Rando 1993). The first major contribution that is generally referred to as a classic in the field of bereavement was Freud's paper, 'Mourning and melancholia.' According to Freud (1957), the psychological function of grief is to withdraw emotional energy (cathexis) and become detached from the loved one (decathexis). The underlying idea of this formulation is that people have a limited amount of energy at their disposal. Consequently, only by freeing up bound energy will the person be able to reinvest in new relationships and activities. Freud believed that the mourner has to work through the grief (grief work hypothesis) by carefully reviewing thoughts and memories of the deceased (hypercathexis). He maintained that although the process of working through causes intense distress, it is necessary in order to achieve detachment from the loved one.

The second theoretical formulation that has been highly influential was advanced by John Bowlby. In his attachment model of grief, Bowlby (1980) integrates ideas from psychoanalysis, ethology, and from the literature on human development. Fundamental to his view is the similarity between the mourning behavior of adults and primates, and children's reaction to early separation from the mother. He considers grief to be a form of separation distress that triggers attachment behavior such as angry protest, crying, and searching for the lost person. The aim of these behaviors is maintenance of the attachment or reunion, rather than withdrawal. However, in the case of a permanent loss the biological function of assuring proximity with attachment figures becomes dysfunctional. Consequently, the bereaved person struggles between the opposing impulses of activated attachment behavior and the need to survive without the loved one. Bowlby believed that in order to deal with these opposing forces, the mourner must go through four stages of grief: initial numbness, disbelief, or shock; yearning or searching for the lost person, accompanied by anger and protest; despair and disorganization as the bereaved gives up the search, accompanied by feelings of depression and lethargy; and reorganization or recovery as the loss is accepted, and an active life is resumed. Emphasizing the survival value of attachment behavior, Bowlby was the first to give a plausible explanation for responses such as searching or anger in grief.

A number of other theorists have proposed that bereaved individuals go through certain stages in coming to terms with the loss. One stage theory that has received a great deal of attention is Kubler-Ross' model, which addresses people's reaction to their own impending death. Kubler-Ross claims that individuals go through stages of denial, anger, bargaining, depression, and ultimately acceptance. It was her model that has popularized stage theories of bereavement. For the past several years, stage models like Kubler-Ross' have been taught in medical, nursing, and social work schools. These models also have appeared in articles in newspapers and magazines written for bereaved persons and their families.

As a result, stage models have strongly influenced the common understanding of grief in Western society. There is evidence that health-care professionals tend to use the stages as a yardstick to assess the appropriateness of a person's grieving. A negative consequence of this, however, is that people who do not follow the expected stages may be labeled as responding deviantly or pathologically. For example, a person who does not reach a state of resolution after a certain time may be accused of 'wallowing in grief.' Also, legitimate feelings such as being angry because one's spouse died of receiving a wrong medication may be discounted as 'just a stage.' Such a rigid application of stage models has the potential of causing harm to bereaved persons. Therefore, many researchers have cautioned against taking any 'staging' too literally. Because of the widespread use and acceptance of stage models, Wortman and Silver (1989) systematically examined all empirical studies that appeared to provide relevant data on the topic of coping with loss. What they found was that the available evidence did not support and in some cases even contradicted the stage approach. In contrast to the notion of an orderly path of universal stages, the reviewed evidence showed that the reaction to loss varies considerably from person to person, and that few people pass through stages in the expected fashion. As a result of this critique of the stage approach, bereavement experts to date at most endorse the idea of grief as a series of flexible phases instead of a set of discrete stages. However, the main weakness of both stage and phase models seems to be that they cannot account for the immense variability in grief response, and that they do not take into consideration outside influences that may shape the course of the grieving process.

2. *Current Understanding of Grief*

Although most grief researchers now agree that the notion of 'stages' is too simplistic, many still hold assumptions about the grieving process that have been influenced by the contributions of Freud and Bowlby. One such assumption is that following a loss, individuals will go through a period of intense distress. Positive emotions, such as happiness, are implicitly assumed to be absent during this period. Second, it is often assumed that failure to experience such distress is an indication that the grieving process is abnormal, and that subsequent mental or physical health problems will occur. Third, it is assumed that successful adjustment to the loss requires that individuals confront and work through their feelings. Fourth, continued attachment to the person who died has generally been viewed as pathological, and the necessity of breaking down the attachment to the loved one is often considered to be a key component of the mourning process. Finally, it is believed that over time, individuals will reach a state of acceptance regarding what has happened and return to prior levels of functioning.

Although these assumptions about the grieving process are widely held among clinicians and the general public, a careful review of the evidence reveals a surprising lack of support for any of them (see Wortman and Silver in press for a review). In one study addressing the notion that most individuals go through a period of intense distress, elderly widows' and widowers' ratings on symptom inventories were used to classify them into *DSM-IV* (*Diagnostic and Statistical Manual of Mental Disorders*) categories of depression. Two months after the loss, 20 percent were classified as showing major depression, 20 percent were classified as exhibiting minor depression, 11 percent were classified as evidencing subsyndromal depression, and 49 percent were classified as evidencing no depression. These findings provide compelling evidence that following the loss of a spouse, a substantial minority of respondents show few signs of depression (see *Depression*).

In direct contradiction to the idea that it is necessary to experience a period of depression, the expression of negative emotion in the first few months following the loss has repeatedly been shown to portend subsequent difficulties. Expressions of positive emotions, in contrast, are quite common among the bereaved and have been consistently associated with less severe and long-lasting symptoms. These findings have emerged whether emotions are assessed through self-report, coded from narratives, or coded from facial expressions. Similarly, many studies have tested the grief work hypothesis, defining the construct in various ways. Most of these studies have produced evidence that flatly contradicts it, with those working through the loss showing worse outcomes over time (see Bonanno and Kaltman 1999, for a more detailed discussion). For example, Bonnano and associates reported from their study on conjugal loss that those bereaved individuals who evidenced emotional avoidance six months after the loss showed low levels of interviewer-rated grief throughout the study, and this response pattern was linked to few symptoms of any kind over a two-year period.

In contrast to the notion of the necessity of breaking down the attachment, numerous studies show that a continuing attachment to the deceased is normal. For example, in their study of how children cope with the loss of a parent, Silverman and Nickman (1996) reported that four months after the death, it was common for children to maintain an active connection to the deceased. The clear majority of children (74 percent) located their parent in heaven, and almost 60 percent of the children reported talking with the deceased. Types of connections identified in other studies include incorporation of virtues of the deceased into one's own character, using the deceased as a role model, and turning to the deceased for guidance regarding a particular problem. Evidence suggests

that in the majority of cases, such continuing bonds are comforting, although some respondents may find them frightening or disturbing. For example, 57 percent of the children in the study by Silverman and Nickman indicated that they were 'scared' by the idea that their parents could watch them from heaven.

It is widely believed that over time, as a result of working through their loss, individuals will achieve a state of acceptance regarding what has happened, move on with their lives, and resume normal functioning. Yet, available evidence suggests that this is not the case. For example, in a study on conjugal bereavement, Parkes and Weiss (1983) found that 61 percent of respondents whose spouses died suddenly, and 29 percent of those who had forewarning were still asking why the event had happened, two to four years postloss. Similarly, in a study on the long-term impact of losing a spouse or child in a motor vehicle accident, Lehman et al. (1987) found little evidence of acceptance even after four to seven years. Despite the amount of time that had elapsed, a majority of the respondents were unable to find any meaning in the loss, had thoughts that the death was unfair, and had painful memories about their spouse or child during the previous month. Furthermore, this study revealed that the bereaved respondents reported higher levels of depression, lower quality of life, and, in case of the loss of a child, more marital problems as well as a higher divorce rate than respondents in the non-bereaved comparison group.

Because the assumptions about the grieving process and recovery discussed above have not been supported by empirical data, we have come to call them myths of coping with loss (e.g., Wortman and Silver 1989). Instead, the available findings demonstrate the enormous variability in response to loss. They also highlight the importance of thinking about the grieving process in frameworks other than those based on breaking down the attachment or working through the loss.

3. *New Directions in Grief Research*

3.1 *The Stress and Coping Paradigm*

At present, much of the research on bereavement is loosely guided by a theoretical approach that might be called the stress and coping approach (e.g., Dohrenwend and Dohrenwend 1981). This paradigm is based on the assumption that once a stressful life event such as loss of a loved one is encountered, the appraisal of the stressor, as well as mental and physical health consequences, will depend on the individual's vulnerability or resistance factors (see *Stress and Coping Theories*). Those factors most reliably associated with adaptation to a major loss include suddenness of the loss, prior mental health problems, concomitant stressors such as ill health, and lack of social support (see Cook and Oltjenbruns 1998 for a more comprehensive review). However, there have been many inconsistencies across studies regarding the most important risk factors for poor outcome. It has been suggested that this inconclusiveness may be due to the interaction of different contextual factors. For example, a sudden death has often been shown to result in more physical and mental health problems for a bereaved individual than an anticipated death. However, the perception of suddenness appears to be related to the age of the deceased, in the sense that even the sudden death of an older person tends to be not as 'unexpected' as the death of a younger person because it constitutes more of an on-time life event. Thus, to understand the role of risk or protective factors, it is important to look at the overall constellation of contextual variables, and at the ways in which they interact.

The stress and coping paradigm not only points to the importance of contextual factors but also adds a perspective that goes beyond the issue of dealing with grief. Keeping in mind that a newly bereaved individual faces a life without the loved one, it becomes obvious that grief itself is only a part of what may be involved in coping with loss. Drawing on the stress and coping literature, Stroebe and Schut (in press) advanced a dual-track model involving two modes of coping that attend to the specific features of the bereavement process: Loss-oriented coping involves an effort to confront feelings of grief and loss itself, while restoration-oriented coping is an attempt to appease pain in some way, or distance oneself from one's grief in order to focus on the demands of daily life and to be able to keep going. This may enable a person to deal with grief in smaller dosages and, at the same time, create some space and save energy to attend to aspects of restructuring life. The idea is that usually both modes are needed and used at some point in the grieving process, and that a certain oscillation (and balance) between them is most likely to constitute an adaptive coping style.

In support of the two coping modes, Stroebe and associates have been able to show that bereaved mothers tended to be more loss-oriented than their husbands, following the death of their child. Another study revealed that men benefited more from interventions 'teaching' them to be more loss-oriented whereas women gained more from improving their restoration-oriented coping skills. Thus, Stroebe and associates provide evidence suggesting that the two coping modes are useful concepts to explain differential bereavement patterns, such as gender differences, and that bereaved individuals benefit from using both loss- and restoration-oriented strategies. What remains to be examined, however, is the process of oscillation and its adaptiveness.

3.2 *Attachment History*

Another innovative line of bereavement research is focusing on the attachment history of the bereaved.

Evidence suggests that some people show an intense and prolonged grief reaction to the loss of their spouse, and that this response is influenced by their attachment history. In an important series of papers, Prigerson and her associates have identified a type of mourning called traumatic grief (Prigerson et al. 1997), and have shown that its symptoms (e.g., preoccupation with thoughts of the deceased) are distinct from the symptoms of bereavement related depression (e.g., apathy). In one study of elderly bereaved respondents, these investigators found that symptoms of traumatic grief, assessed six months post loss, predicted critical health outcomes such as incidences of cancer, high blood pressure, and cardiac events one year later. Based on their research, the authors propose that a background of physical or sexual abuse, neglect, hostile conflict, or early parental loss or separation, typically leads to attachment disturbances such as excessive dependency or compulsive caregiving. In such cases, the loss of a marital relationship that was stabilizing may result in traumatic grief, even if the loss itself did not occur under traumatic circumstances. At present, preliminary research findings support this model.

3.3 Trauma Theory

In recent years, there has been considerable interest in applying concepts from the trauma literature to help understand the diversity of reactions to loss (see Janoff-Bulman 1992). Many theorists have maintained that losses are particularly likely to result in intense and prolonged grief if they shatter the survivor's most basic assumptions about the world. Particularly if the loss is sudden and traumatic, it may lead most mourners to question assumptions that they previously took for granted. These include assumptions that the world is predictable and controllable, that the world is meaningful and operates according to principles of fairness and justice, that one is safe and secure, that the world is benevolent, and that, generally speaking, other people can be trusted. Consistent with these ideas, there is some evidence to suggest that sudden, traumatic deaths pose particular difficulties for the survivor, especially if the death was intentional (i.e., murder). It appears that such deaths are more likely than those brought about by natural causes to result in post-traumatic stress symptoms such as intrusive thoughts and nightmares (see *Post-traumatic Stress Disorder*).

4. Treatment for Grief

Is it worthwhile to seek treatment following a major loss? How do people know whether treatment is warranted in their case? What types of treatment have been found to be effective? It has been suggested that if the acute symptoms of grief subside after a few months, treatment is probably not warranted. However, if a bereaved person continues to experience intense feelings of yearning and depression, anxiety, or post-traumatic stress symptoms beyond the first six months, there is some consensus that it is wise to consider treatment. Jacobs (1993) has reported that during the first year following a loss, approximately 30 percent of the 800,000 who lose a spouse will experience a diagnosable psychiatric problem. However, less than one in five of these individuals seek help. Those who would particularly benefit from professional help include parents who have lost children, people who have experienced the untimely loss of a spouse, those who are experiencing concomitant stressors such as poor health, and those who have little social support available. Help should ideally be sought from a therapist who is experienced in dealing with grief. For those who have experienced a sudden, traumatic loss, it is important that the therapist be knowledgeable about trauma as well. Individuals may also benefit from support groups. Such groups can help the bereaved to understand that their feelings and reactions are normal, and can impart information about how to deal with specific problems. When one's grief is intense or prolonged, support groups are best considered as an ancillary treatment rather than a substitute for psychotherapy.

5. Conclusion and Outlook

Taken together, the ways in which researchers think about bereavement seem to have changed in a number of crucial regards. As pointed out above, a shift can be observed from the idea of a universal pattern of stages towards a recognition of immense variability and the importance of contextual influences; from the breaking bond orientation towards the concept of transforming the connection to the loved one; and from a model that is marked by a unitary focus on grief toward a comprehensive approach that includes the challenges of dealing with daily life after the loss. However, traditional ways of thinking about grief are still surprisingly prevalent among laypeople, practitioners, and some researchers, despite the lack of supportive evidence. It seems that the more recent evidence contradicting the traditional views of grief will have to be replicated in order to further move along the described shift in understanding bereavement, to the point that outdated models and concepts lose their 'myth' character.

In addition, research is needed to examine further and refine the new ways of understanding bereavement that were outlined above. For instance, while the literature is clear in suggesting that it is common for individuals to maintain a connection to the deceased,

it is unclear what kind of connection tends to be more or less adaptive. Second, a further investigation of the impact of a person's attachment history on adaptation to loss could provide helpful insights into the phenomenon of traumatic grief and may give cues for effective interventions. Third, it would be a contribution of future research to investigate further the use and interplay of the restoration- and loss-oriented modes and their adaptiveness in the context of different kinds of loss and outside conditions. In this context, the controversial question of the role of grief work needs to be further assessed. One important question, for example, is to what extent an oscillation or a certain balance between loss- and restoration-orientation is necessary for positive adjustment. Another issue of great interest is how this dynamic coping process may work in families or other social situations. Some families, for example, may show a certain level of synchrony with respect to their balance of coping. Others may have more of a teamwork style, which means that different family members take over different coping tasks, depending on their current resources and state of mind (Baltes and Staudinger 1996).

As these thoughts demonstrate, there is still much to be learned and understood about the ways in which people cope with a major loss, and those who suffer from a major loss need and deserve an open-minded attitude toward their plight. After all, the death of a loved one constitutes one of the most challenging experiences that we all encounter at some point in our lives.

See also: Adult Psychological Development: Attachment; Adulthood: Developmental Tasks and Critical Life Events; Coping across the Lifespan; Coping Assessment; Death and Dying, Psychology of; Death and Dying, Sociology of; Death, Anthropology of; Stress and Health Research

Bibliography

Baltes P B, Staudinger U M 1996 *Interactive Minds: Life-Span Perspectives on the Social Foundation of Cognition*. Cambridge University Press, New York

Bonanno G A, Kaltman S 1999 Toward an integrative perspective on bereavement. *Psychological Bulletin* **125**: 760–76

Bowlby J 1980 *Loss: Sadness and Depression*. Basic Books, New York, Vol. 3

Cook A S, Oltjenbruns K A (eds.) 1998 *Dying and Grieving: Life Span and Family Perspectives*, 2nd edn. Harcourt Brace, Fort Worth

Dohrenwend B S, Dohrenwend B P 1981 Life stress and illness: Formulation of the issues. In: Dohrenwend B S, Dohrenwend B P (eds.) *Stressful Life Events and their Contexts*. Prodist, New York, pp. 1–27

Freud S 1957 Mourning and melancholia. In: Strachey J (ed. and trans.) *The Standard Edition of the Complete Work of Sigmund Freud*. Hogarth, London, Vol. 14

Jacobs S 1993 *Pathologic Grief: Maladaptation to Loss*. American Psychiatric Press, Washington, DC

Janoff-Bulman R 1992 *Shattered Assumptions: Towards a New Psychology of Trauma*. Free Press, New York

Jeter K 1983 Analytic essay: Family, stress and bereavement. In: McCubbin H I, Sussman M B, Patterson J M (eds.) *Social Stress and the Family*. Haworth Press, New York

Lehman D R, Wortman C B, Williams A F 1987 Long-term effects of losing a spouse or child in a motor vehicle crash. *Journal of Personality and Social Psychology* **52**: 218–31

Parkes C M, Weiss R S 1983 *Recovery from Bereavement*. Basic Books, New York

Prigerson H G, Bierhals A J, Kasl S V, Reynolds C F, Frank E, Jacobs S 1997 Traumatic grief as a risk factor for mental and physical morbidity. *American Journal of Psychiatry* **154**: 616–23

Rando T A 1993 *Treatment of Complicated Mourning*. Research Press, Champaign, IL

Silverman P R, Nickman S L 1996 Children's construction of their dead parents. In: Klass D, Silverman P R, Nickman S L (eds.) *Continuing Bonds: New Understandings of Grief*. Taylor and Francis, Washington, DC

Stroebe M, Schut H in press Models of coping with bereavement: A review. In: Stroebe M, Hansson R O, Stroebe W, Schut H (eds.) *Handbook of Bereavement Research: Consequences, Coping, and Care*. American Psychological Association Press, Washington, DC

Wortman C B, Silver R C 1989 The myths of coping with loss. *Journal of Consulting and Clinical Psychology* **57**: 349–57

Wortman C B, Silver R C in press The myths of coping with loss revisited. In: Stroebe M, Hansson R O, Stroebe W, Schut H (eds.) *Handbook of Bereavement: Consequences, Coping, and Care*. American Psychological Association Press, Washington, DC

K. Boerner and C. B. Wortman

Bernard, Jessie (1903–96)

Jessie Bernard's role as an amiable radical perfused both her personal history and sociological *oeuvre*. Her traditional sociological training and her marriage to a major figure in the field originally brought Bernard into the sociological establishment. Subsequently, through a series of personal 'revolutions,' she emerged as a feminist critic of mainstream sociological paradigms, social institutions, and public policy, particularly those concerning women's roles. Nonetheless, by the 1970s, Bernard's work was recognized internationally and honored by the entire field, as well.

In 15 solo books, seven co-authored or edited volumes, 25 book chapters, and 60-plus journal articles, her humane and unpretentious style created a medium in which her increasingly radical message was accepted without raising traditional hackles. Bernard's private history may be read as an amicable, but undaunted, rebellion against the expectations that family and society pressed upon a female who was born shortly after the turn of the century and died four years shy of the millennium.

Her belief in social positivism profoundly shaken by the Nazi Holocaust, Bernard resonated to societal

upheavals—particularly the Women's Movement—that were to mark the second half of the twentieth century and dramatically alter her *Weltanschauung*. She was, by turns, a professional academic woman, a single mother, a revered figure in the field of sociology, a feminist author whose work reached an international audience far beyond academic sociology, and a beloved mentor extraordinaire.

1. *Biography*

Born in Minneapolis, Minnesota on June 8, 1903, to parents of Romanian Jewish heritage, Jessie Bernard was the third of Bessie Kanter and David Solomon Ravitch's four children. In January 1920, Bernard entered the University of Minnesota, where she studied with Pitirim Sorokin, Alvin Hanson, N. S. B. Gras, Karl Lashley, and Luther Lee Bernard. She received both her B.A. (1923) and M.A. (1924) from the University of Minnesota, before marrying Bernard, 21 years her senior, against her family's objections to age and religious differences. The Bernards moved to Washington University, St. Louis, where Jessie completed her Ph.D. in 1935. During a marital separation, she worked for the US Bureau of Labor as a social science analyst. Bernard taught at Lindenwood College for Women, St. Charles, Missouri (1940–7) and served as professor of sociology at Pennsylvania State University (1947–64).

Influenced by L. L. Bernard's social positivism as 'the best way to achieve desired societal ends' (1978), Bernard grew to intellectual maturity trained 'strictly in the measurement tradition' (personal communication 1979). In the mid-1940s, the shock of the Nazi Holocaust led her to reject that social positivist paradigm. She moved squarely into the functionalist school, where she was to remain until her feminist epiphany.

L. L. Bernard's death in 1951 left Jessie Bernard responsible for three young children. Bernard's personal life as widow and female family head provided the experiential background for empathic insight into a major emerging sociological problem. By the fall of 1953, Bernard left for Europe to study postwar trends in sociological research and to heal her emotional wounds.

After her year abroad, Bernard returned to Pennsylvania State to work on two new volumes: *Remarriage: A Study of Marriage* (1957a/1971) and *Social Problems at Mid-century: Role, Status, and Stress in a Context of Abundance* (1957b). She spent 1959–60 as a visiting professor at Princeton University, which had just allowed women to enter the front door of the university library. By 1964, Bernard decided to leave academia, where she sensed it was necessary to 'censor (her) ideas to fit the pattern of ideas surrounding (her)' (personal communication 1979). Bernard moved to Washington, DC, where she maintained an active scholarly and professional life as a participant in local and international conferences and policy debates through the late 1980s. She died on October 6, 1996 in Washington, DC.

2. *Major Contributions*

Bernard's scholarship is eclectic, covering a wide range of ideas and using a complex palette of methodologies. Bernard's earliest works, such as *American Family Behavior* (1942a/1973), through *The Future of Marriage* (1972), utilize macrodata, including macrostatistics, to compare sociological patterns and to gauge the strength and trends of sociological phenomena. Later, she turned to microdata in the form of the letters she and her children exchanged over more than 25 years. In *Self-Portrait of a Family* (1978), a courageously self-revealing effort, she chronicled the intimate infrastructure of family life. In her last two volumes, *The Female World* (1981) and *The Female World from a Global Perspective* (1987), Bernard combined macro- and microanalysis to provide a rich texture to her work.

In 1942 Bernard published her first solo book *American Family Behavior* (1942a/1973). In the early years at Penn State, she wrote *American Community Behavior* (1949/1962). Both volumes foreshadowed later work in which Bernard measured the American family's performance, success in marriage, and the 'institutionalization' of marriage and family norms. Despite her defection from social positivism to functionalism, Bernard's reliance on research findings as the best, or only available, evidence remained a hallmark of her later analyses.

Bernard's break from academia came almost simultaneously with the publication of her most significant work to that time, *Academic Women* (1964/1974). Although *Academic Women* received both the Pennsylvania State University Bell award and the Kappa Gamma honorary award, the larger academic establishment, according to Bernard, responded initially with 'a great big yawn' (1964/1974).

Academic Women explored the condition of women in academe, dispassionately examining the evidence for and against discrimination. That book presaged Bernard's later attempts to deal with significant social issues—abortion, battered women, child abuse, the culture of poverty, female-headed families, and sexism—long before they became incendiary social questions. Widely influential, both despite and because of the feminist criticism leveled against it, *Academic Women* scrutinized the factors that accounted for women's subordinate role in academia. Bernard concluded that sex was more salient than role as a determinant of status. She attributed women's unequal condition to their propensity to teach in colleges rather than universities, to teach rather than undertake research, to act as transmitters of established knowledge rather than 'men of knowledge,' and to follow patiently rather than innovate boldly.

Bernard recognized that scientific productivity was a function of a researcher's position in the communication system. She depicted the 'stag effect' as a subtle process excluding women from the informal communication system along which emerging scientific knowledge is disseminated. Although Bernard described the palpable procedures and processes of sexism (a term not yet coined), she stopped short of recognizing them as the informal underpinnings of discrimination. Bernard did not yet perceive, as she later did, the crippling relationship between the informal practices and attitudes of sexism and the structural manifestations of discrimination.

Academic Women touched the beachhead of concern about women's condition in society just ahead of the swelling wave of feminism. Radical feminists railed against Bernard's conclusion that no formal discrimination existed. Later, after her own feminist 'revolution,' Bernard applauded her critics' analyses. Nonetheless, *Academic Women* was the beginning of a new surge of intellectual strength, which even the male-dominated academic establishment could no longer ignore.

Over the next 14 years, Bernard broke through the functionalist paradigm that had held mainstream sociology in its thrall. She entered the unexplored terrain of the feminist perspective. Beginning in her sixth decade, Bernard began to mine a new, richer multidisciplinary vein—feminism—that would later fuel her strongest and most prodigious intellectual contributions.

The most basic theme that snakes through Bernard's work is biculturality, first discussed in terms of Jewish immigrants in America (1942b) and later expanded in *Marriage and Family among Negroes* (1966). Bernard developed the notion of two black cultures—one externally adapted, the other acculturated—within the larger white context. Over her remaining career, Bernard would apply this theme to the female world, as well.

The Sex Game (1968) elaborated biculturality by depicting the sexes as two large, relatively unstructured collectivities living in sexual apartheid. In *Women and the Public Interest: An Essay on Policy and Protest* (1971), Bernard introduced the concept of 'women's sphere' and the pervasive influence of women's 'stroking' function. Bernard marshaled evidence to show that women's specialization in stroking throughout the world contributes to their subordinate position. Through this behavior, the stroker shows solidarity, raises the status of others, gives help, rewards, agrees, concurs, complies, understands, passively accepts (1971). Such stroking behavior, Bernard argued, is incompatible with high-level occupational roles, in which instrumental, aggressive, and, often, competitive behavior is required.

The theme of biculturality is further explored in *The Future of Marriage* (1972), probably her most widely read work. There, Bernard conceptualized each marital union as two different, noncoinciding marriages or worlds (his and hers). With Durkheimian precision, Bernard used census data to demonstrate that marriage benefited men more than women.

Other major themes texture the fabric of Bernard's work: formal vs. informal discrimination; power; women as a subordinate, dependent group; 'stroking' as a fundamental unit of social action; the conflict between women's family and work roles; sexism and power relationships; the deteriorating effect of housework and total responsibility for childcare on women's mental health; homosociality, that is, the preference for social interaction with people like oneself; sex differences, including typical vs. characteristic differences between males and females; biological sex roles vs. cultural gender roles; the function of stereotypes as mechanisms for papering over the lack of fit between gender roles and individual differences; tipping points and turning points; social policy as an instrument for alleviating the disadvantages women face; and the historical youthfulness of the nuclear family.

In the late 1960s, Bernard moved agonizingly, but irrevocably, from a functionalist to a feminist perspective. Surveying sociological phenomena through a feminist prism cast a different, more vivid spectrum of colors than Bernard had seen before. The cool 'objectivity' of social science could now be perceived as a rationalization for remaining emotionally uninvolved in the inequities and moral dilemmas of social life. Until the feminist perspective emerged, male and female researchers alike had to refract society's ills through a single distorting male lens. Missing was the female lens, essential for the social stereoscope through which the two worlds could be seen together in their true dimensionality.

The Female World, Bernard's (1981) *magnum opus*, is arguably the most complex, in-depth, multidisciplinary analysis of women's separate world that has ever been undertaken. *The Female World* includes a complex assessment of women's station. Not only do women live in different marriages than their husbands, they live in different families, educational systems, social strata, occupational structures, and political and cultural realms. Bernard does not deny that women also live in the world of men, but argues that they are not of that world, in much the same way that Americans living in Paris are not bona fide Parisians.

In *The Female World*, Bernard probed the anatomy of that world's unique class structure. She concluded that the very subconcepts underlying the notion of social class (long accepted in mainstream sociology) failed to reflect the complex class structure of the female world. Bernard reported many new feminist historical discoveries, boldly endorsing the emerging options that would reduce the loss of women to meaningless lives. Until then, the female world had been mostly ignored. When it was not completely *terra incognita*, the female world has been analyzed as a by-

product of the male world, using male concepts, male methodologies, and male values. Bernard's goal in *The Female World* was to redress this intellectual inequity. Here, Bernard conceptualized the female world *sui generis*, as an autonomous entity, rather than one created from the rib of male sociological analysis.

The volume, spanning prehistoric times to the late twentieth century, is monumental in scope. It reports on previously ignored female worlds: from the third and fourth century convents to the beguinages, their lay analogue in the thirteenth and fourteenth centuries, to the female vagabonds of the Middle Ages, to the Lowell girls in the nineteenth century, to the twentieth century's institutionalized and homeless women. Bernard moves seamlessly from subhuman primates to humans, from sociology to anthropology, history, psychology, and economics with a breathtaking grasp of concepts, methodologies, central themes, ideological underpinnings, and weaknesses.

To describe the class structure of the female world, Bernard relied in part upon the traditional sociological tripartite model of education, income, and occupation. Where those components were not applicable (e.g., for women who were not in the paid labor force), however, she supplemented them with 'subjective behavioral criteria of status,' such as social climbing, influence, modeling (i.e., who want to be like whom), and even husband's income. The classes Bernard derived were 'society, media-created celebrities, intellectuals, white-collar and pink-collar workers, unpaid housewives, blue-collar and service workers, welfare recipients, and demoralized, broken, homeless, outcast women.'

In *The Female World*, Bernard built upon the bicultural perspective which informed much of her earlier work, using several conceptual polarities to identify the structured differences between these worlds. Following Tönnies, she described the *Gemeinschaft* of the female world, based on kinship and locale. It stands in sharp contrast to the *Gesellschaft* ethos of the male world, structured by money and politics. The *Gemeinschaft*, according to Bernard, unfits women for living in the *Gesellschaft*. The *Gesellschaft*, however, is permeated by a male ethos that Bernard concludes is maladaptive for men as well as women.

The second polarity is that between the integry and the economy, originally described by Kenneth E. Boulding. The integry/economy dichotomy parallels *Gemeinschaft*/*Gesellschaft*. The integry forms that segment of the social system that attends to community, identity, love, trust, altruism, intimacy, and loyalty. The integry belongs to women, encompassing a set of compassionate activities and norms that holds the social structure together.

By contrast, in the economy, monetary exchanges create the context for relationships. The economy, built and governed primarily by men, is based on rationality, self-interest, competition, aggression, and impersonality. Bernard catalogued the costs to women of living in their own world colored by an ethos of agape, that is, of social activism or humanitarianism. She argued that only women who derive their economic support from others—fathers or husbands—can afford to devote themselves to the humanitarian needs of society. Although the total society may benefit from their contributions, the personal price women pay, ranging from guilt to dependency, is heavy.

The Female World is a serious, documentary celebration of women's unique strengths on which Bernard rests the chances for the future salvation of both sexes in a postindustrial society. She offers no apologia for women or for the sex differences whose irreducibility she insisted upon both before and after her feminist conversion. Earlier, Bernard politely predicted that sex role transcendence and shared roles might be the wave of the future (Bernard 1972, 1974, 1975, Lipman-Blumen and Bernard 1979). In this volume, however, Bernard analyzed the entire female world with its complicated structure and ethos, past and present, in still well mannered, but now clearly forceful, tones.

In her final work, *The Female World from a Global Perspective*, Bernard posed two integrating metaphors: the Feminist Enlightenment and equitable integration. Bernard perceived the Feminist Enlightenment, beginning in the 1960s, as an analogue to the eighteenth century French Enlightenment: an intellectual and political movement that redefined history, rejected the authoritarian status quo, and created a new basis for understanding the world. The Feminist Enlightenment, according to Bernard, ground a fresh and sharper scalpel than 'objective' science had provided two centuries before. Feminism offered a laser-hot tool to cut through the stereotypes, sexism, ghettoization, and blindness to the 'racism, classism, and status impediments' that women around the globe still experienced in mid-twentieth century. Bernard apprehended the importance of Women's Studies, a new multidisciplinary field within higher education, as a critical mainspring of the Feminist Enlightenment.

The reconceptualization of equitable integration is the second organizing metaphor of *Global* (as Bernard referred to this volume). Bernard unflinchingly analyzed the political and geographic separations among the segments of the female world no less rigorously than the partitioning of the male and female worlds. She also fathomed the serious difficulties in bridging the chasm between Western feminism and feminism in other parts of the globe. Nonetheless, Bernard worked toward a conceptualization of equitable integration that would allow for solidarity *cum* diversity. Equitable integration would be bolstered by strategies yet to be developed and implemented through realistic policies for reducing, if not totally eradicating, racism, classism, and status impediments.

Communication and technologies for promoting all forms of communication among women loomed large to Bernard. She welcomed enthusiastically the new

possibilities for communication among women, through newsletters, films, TV, and videos. Long before the Internet had become the *modus communicandi*, she understood the importance of electronic communication in breaking down the ideological and cultural barriers among women. Bernard also recognized both the significance and the difficulties of face-to-face meetings among women, particularly the three International Conferences on the Status of Women, sponsored by the United Nations during the Decade of Women, 1975–85. Even when results were disappointing, even when these conferences were nearly capsized by the tsunamis of international politics, Bernard remained clear-eyed but optimistic. She based her optimism on the scholarship and ferment that were sure to flow from the developing Feminist Enlightenment.

3. *Influence, Impact, and Current Significance*

Recognition of Jessie Bernard as a major contributor to the discipline, as well as to the education of the broader public, has come from every quarter. One review of Bernard's work concluded that she was 'one of the most important sociologists of the twentieth century' (Howe and Cantor 1994). Bannister's (1991) less positive evaluation has been rejected uniformly by numerous scholars as too error-ridden to be taken seriously. Ironically, Bernard's ability to speak lucidly and incisively, without academic pretensions, to a larger public probably delayed the enormous outpouring of professional acclaim until her later years.

Bernard's influence is evident in contemporary work on female/male relationships. Her considerable contributions to the field were analyzed in two posthumous panels at the annual meetings of the Eastern Sociological Society and the American Sociological Association. From 1966 to 2000, *The Social Sciences Citations Index* listed 2,994 citations to Bernard's work. Her concepts and insights so widely permeate scholarly thinking that they often appear unacknowledged in the work of contemporary researchers. Many years after Bernard concluded that men benefit more than women from marriage, this topic continues to be debated by contemporary scholars, another sign of the amiable radical's impact. Her work is required reading on current syllabi in sociology, psychology, and women's studies courses, many of which appear on the Internet.

Bernard's influence is not limited to the scholarly world. For example, the central idea of the best seller *Women Are From Venus*, *Men Are From Mars*, by John Gray is vintage Bernard. Various searches of the Internet reveal approximately 450 websites with re ferences to Bernard.

Although she served as president of the Society for the Study of Social Problems (SSSP) (1963–4), after leaving academia Bernard declined nominations to various professional offices, including the presidency of the American Sociological Association. Nonetheless, she received the merit award from the Eastern Sociological Society and the Burgess award from the National Council of Family Relations (1973), the Kurt Lewin award from Pennsylvania State University (1976), the outstanding achievement award from the American Association of University Women (1976), and honorary doctorates from 10 major universities. In 1989, she received the American Sociological Association's (ASA) Distinguished Career Award.

Several awards have been established in Bernard's name to honor those who similarly have contributed intellectually, professionally, and humanely to the world of scholarship and feminism: the Jessie Bernard awards, created by the American Sociological Association (1976), the District of Columbia Sociologists for Women in Society (1978), and the Wise Woman Award by the Institute for Women Policy Studies. These reflect the esteem in which her intellectual brilliance and eclecticism, personal courage, amiable radicalism, unpretentious humanity, and feminist sisterhood are held.

See also: Community Sociology; Family and Gender; Family as Institution; Family Theory: Feminist–Economist Critique; Feminist Theory; Feminist Theory: Liberal; Gender and Feminist Studies; Gender and Feminist Studies in Sociology; Gender History; Marriage

Bibliography

Bannister R C 1991 *Jessie Bernard: The Making of a Feminist*. Rutgers University Press, New Brunswick, NJ

Bernard J 1942a/1973 *American Family Behavior*. Russell & Russell, New York

Bernard J 1942b An analysis of Jewish culture. In: Graeber I, Britt S H (eds.) *Jews in a Gentile World*. Macmillan, New York pp. 243–62

Bernard J 1949/1962 *American Community Behavior: An Analysis of Problems Confronting American Communities Today*, rev. edn., Holt, Rinehart and Winston, New York

Bernard J 1957a/1971 *Remarriage: A Study of Marriage*. Russell & Russell, New York

Bernard J 1957b *Social Problems at Midcentury: Role, Status, and Stress in a Context of Abundance*. Dryden Press, New York

Bernard J 1964/1974 *Academic Women*. Meridian, New York

Bernard J 1966 *Marriage and Family among Negroes*. Prentice-Hall, Englewood Cliffs, NJ

Bernard J 1968 *The Sex Game: Communication Between the Sexes*. Prentice-Hall, Englewood Cliffs, NJ

Bernard J 1971 *Women and the Public Interest: An Essay on Policy and Protest*. Aldine-Atherton, Chicago

Bernard J 1972 *The Future of Marriage*. World Publishing Company, New York

Bernard J 1973 *The Sociology of Community*. Scott-Foresman, Glenview, IL

Bernard J 1974 *The Future of Motherhood*. Dial, New York

Bernard J 1975 *Women, Wives, Mothers: Values and Options*. Aldine, Chicago

Bernard J 1978 *Self-portrait of a Family*. Beacon Press, Boston

Bernard J 1981 *The Female World*. Free Press, New York
Bernard J 1987 *The Female World from a Global Perspective*. Indiana University Press, Bloomington, IN
Howe H, Cantor M G 1994 Jessie Bernard: the unfolding of the female world. *Sociological Inquiry* **64**(1): 10–22
Lipman-Blumen J, Bernard J 1979 *Sex Roles and Social Policy*. Sage, Beverley Hills, CA

J. Lipman-Blumen

Bernoulli, Jacob I (1654–1705)

Jacob I. Bernoulli, born December 27, 1654 in Basel, died 1705 August 16, in Basel, Swiss mathematician and first scholar of the famous Bernoulli family with important contributions to the infinitesimal calculus, mechanics, theories of variation and probability.

Jacob Bernoulli was born into a family of bankers and merchants from Amsterdam who later moved to Basel. In 1671 he finished his studies in philosophy with a master (Magister) degree, 1676, in theology with a liscentiat. During 1676–80 and 1681–82, he visited France, the Netherlands, and England, and started his studies in applied mathematics, inspired by the philosophical and mathematical works of R. Descartes and N. Malebranche and the physical theories of R. Boyle and R. Hooke. In 1682 he refused an offer to become a preacher in Strasbourg in order to continue his mathematical studies. His brother Johann I. Bernoulli (1667–1748) was instructed by him in mathematics. Johann became also a famous mathematician who later on competed with him in the solution of several mathematical problems. Initially, Jacob gave lectures in experimental physics at the university of Basel. From 1687 he was professor of mathematics at his home university. His '*Opera omnia*' (1744 posthum published) contains scientific articles published by himself and 32 edited treatises.

Since 1677, Jacob Bernoulli wrote a scientific diary, called 'Meditationes.' His first articles were a (false) theory of comets and a treatise on the gravity of the ether. In physics he started with inquiries about the atmospheric density and the perpetuum mobile of D. Papin. During 1684–86, he also worked about logic. In mathematics he started with Descartes' geometry and published algebraic and geometric articles in the '*Acta Eruditorum*.' In competition with J. Wallis he constituted the principle of complete induction in 1686. In 1687, Jacob Bernoulli proved that the area of a triangle can be divided into four parts by two orthogonal straight lines. The work of J. Wallis, I. Barrow, and G. W. Leibniz on the infinitesimal calculus had a great impact on Jacob Bernoulli. As young professor of mathematics, he asked Leibniz to explain to him the foundations of the infinitesimal methods. Behreen 1684 and 1704, he wrote several treatises concerning the theory of mathematical series. Exponential series were considered as the inverse of logarithmic series. In a paper on the divergence of harmonic series he independently repeated the so-called Bernoulli inequality $(1+x)^n \geqslant nx$ $(x \geqslant 0)$, which had already been introduced by I. Barrow in 1670.

In 1691 he published papers about problems in infinitesimal mathematics (e.g., the slope of a tangent line, quadratures, rectifications). Jacob Bernoulli applied infinitesimal methods systematically to geometric problems. For example in 1692, he solved the so-called Florentine problem which means the quadrature of a certain part of the semi-sphere. He analyzed several curves such as the parabolic and logarithmic spiral and introduced the Bernoulli differential equation

$$y' + f(x)y + g(x)y^n = 0.$$

For these problems, he found a formula for the radius of curvature at a point of a curve and used polar coordinates. In 1690 Jacob Bernoulli solved a problem of Leibniz (1686) to determinate the curve along which a body falls with uniform velocity (isochrone). He introduced the word 'integral' when suggesting the name '*calculus integralis*' instead of Leibniz's original '*calculus summatorius*' for the inverse of '*calculus differentialis*.'

In 1696, Johann I. Bernoulli stated his famous problem of the Brachystochrone: along which curve does a mass point move on a vertical plane from a point to a deeper one in the shortest time only by its gravity? Johann I. Bernoulli himself only delivered a restricted solution. His brother and rival Jacob solved the problem in a general way. His main idea was the following: one has to find a curve with an extremum property (i.e., maximum or minimum) from a set of curves. Therefore Jacob Bernoulli assumed that every part and especially every element of the curve also has the extremum property. Bernoulli's postulate inspired L. Euler in a new mathematical discipline which is now called the theory of variations. In his famous inquiry '*Methodus* inveniendi lineas curves maximi minimive proprietate gaudentes' (1744), he reduced problems of variation to problems of maxima and minima. The calculus of variation considers a new kind of problem in infinitesimal calculus: one has not only to find the solution of a function (curve) with an extremum property, but of a functional with real functions as arguments (set of curves). In 'Analysis magni problematis isoperimetrici' (1701), Jacob analyzed isoperimetric curves of a given length enclosing a maximal area. In the analytical mechanics of the eighteenth century (e.g., J. L. Lagrange's *Mécanique analytique* 1788), the calculus of variation became an important procedure for solving mechanical equations of motion.

Today, problems of variation are analyzed in the more general framework of functional analysis. Although it was Jacob Bernoulli who recognized the general concept of variation in the problem of the Brachystochrone, Johann acknowledged his brother's merit only after his death.

In physics, Jacob I. Bernoulli found important results in the mechanics of solid and elastic bodies. In 1691 he started a series of theorems on an elastic beam which is bent by an external force. The problem had already been analyzed by G. Galilei, E. Mariotte, G. W. Leibniz, and others. So, it was a challenge for Jacob to criticize the false or restricted assumption of his famous forerunners and to apply and test his infinitesimal methods. Leibniz praised Bernoulli's approach because of its generality. Some months before his death in 1705, Jacob I. Bernoulli finished his last work about the curvature of a bent beam. He foresaw that his solution was the beginning of a general theory on the elastic continuum. In 1704, he also anticipated D'Alembert's famous principle of virtual velocities by an own version.

Since 1685, Jacob I. Bernoulli had been interested in the theory of probabilities. This early interest is documented in his correspondence with Leibniz. Jacob I. Bernoulli was influenced by the studies of C. Huygens on games of hazard (1657) which he enlarged to a theory with the name '*ars conjectandi*' (art of conjecturing). The *ars conjectandi* which was published post-1713 by Jacob's nephew Niklaus I. Bernoulli had a great impact on the development of the theory of probability (e.g., A. de Moivres' *Doctrine of Chances* 1718). Jacob I. Bernoulli generalized Huygen's examples of numbers and derived general formulas of combinatorics with skillful notations. Referring to Leibniz's combinatorics, he analyzed the so-called Bernoulli polynoms

$$B_m(z) = \sum_{k=0}^{m} \binom{m}{k} B_k \cdot z^{m-k}$$

with $m \geqslant 0$ (integer) and coefficients B_k which can recursively be calculated as so-called Bernoulli numbers

$$B_0 = 1, \quad \sum_{k=0}^{m-1} \binom{m}{k} B_k = 0 \quad \text{where } m \geqslant 2.$$

They are used for solving difference equations of the form

$$u(z+1) - u(z) = m \cdot z^{m-1} \quad \text{where } m \geqslant 1.$$

In some hints of the *ars conjectandi*, one can even find the transition to statistics. The famous law of large numbers can also be found in the *ars conjectandi*. Let us consider a Bernoulli sequence of tests, i.e., a sequence of independent tests by random which always have the constant probability p for the occurrence of an event A. An example is the repeated throw of a coin with probability $p = 0.5$ for the occurrence of the event face side. According to Bernoulli's theorem, the probability that the relative frequency r_n of the occurrence of A after n tests is arbitrarily near to p converges to 1 with increasing p, i.e.

$$\lim_{n \to \infty} P(|r_n - p| < \varepsilon) = 1 \quad \text{for all } \varepsilon > 0.$$

E. Borel (1881–1956) introduced a stronger version of the law of large numbers. The probability that r_n converges to p is 1, i.e.

$$P(\lim_{n \to 0} r_n = p) = 1.$$

Borel's version is stronger than Bernoulli's one, because Borel's version follows from Bernoulli's one, but not vice versa. In the modern theory of probability which is no longer restricted to Bernoulli sequences of tests, Bernoulli's and Borel's laws are special cases of more general theorems about random variables according to certain conditions. Therefore, we prefer to speak of laws of large numbers. In this mathematical tradition, R. von Mises (1883–1953) tried to reduce the concept of statistical probability to sequences of relative frequencies. But from a modern point of view, Bernoulli's and Borel's laws are only probabilistic propositions referring to a measure P of probability. Thus, they presuppose a concept of probability. Besides probability, the *ars conjectandi* also contains reflections about certainty, necessity, random, moral, and caculable expectation.

The work of Jacob I. Bernoulli is distinguished by great algorithmic skills and conceptual depth. In cooperation and competition with famous mathematical contemporaries, he developed the infinitesimal calculus with great mastership. His lifetime is a mathematical period of consolidation and exploitation of the great discoveries of the seventeenth century and of their application to the investigation of scientific problems, for instance, in mechanics. With his brother Johann he realized the transition from the mathematical pioneer Leibniz to the age of L. Euler who became the dominant mathematical figure of the eighteenth century. But Jacob's mathematical contributions are not only restricted to the mathematical fundaments of the eighteenth century. The theory of variation also has great importance in the twentieth century physics of quantum theory and theory of relativity and in the twentieth century mathematics of functional analysis. With his *ars conjectandi*, Jacob I. Bernoulli even founded a mathematical theory with interdisciplinary application in the natural and social sciences. As son of a banking and merchant family, he was aware of the monetary, economic, and social developments of his age, as general interest increased in problems of mathematical expectations, decision, random, probability, interest rate, and statistics,

characterizing human society. Contrary to the indeterminism of human society, the natural science of the seventeenth and eighteenth centuries assumed a completely mathematical determinism of nature which was successfully computable in analytical mechanics. The infinitesimal calculus delivered the mathematical fundament of the new mathematical disciplines. Philosophically, the concept of nature and society in the eighteenth century was discussed later on by Immanuel Kant. In what way is acting and deciding on the basis of free will in a completely determined nature possible? The strict distinction between a completely determined nature and a human world of random, expectation and probability has been overcome in the twentieth century when quantum mechanics with its probabilistic quantum effects became the new fundamental of physics.

Bibliography

Bernoulli J I 1713 *Ars conjectandi* Basel [repr. 1968, Brussels]

Bernouli J I 1969 *Die Werke*. 1st edn. Naturforschende Gesellschaft

Bernouli J I 1959 *Opera I-II*, Geneva [repr. 1967, Brussels]

Dietz P 1959 Die Ursprünge der Variationsrechnung bei Jakob Bernoulli. *Verhandlung der Naturforschenden Gesellschaft Basel* **70**: 81–146

Fueter O (ed.) 1939 Jacob Bernoulli *Großе Schweizer Forscher*. Zurich, pp. 86–8

Hacking I 1971 Jacques Bernoulli's Art of Conjecturing. *British Journal for the Philosophy of Science* **22**: 209–29

Hofmann J E 1970 Jacob Bernoulli. In: *Dictionary of Scientific Biography II*. New York. pp. 46–51

Mainzer K. 1980 Jacob Bernoulli. In: Mittelstraβe J (ed.) *Enzykiopädie Philosophie und Wissenschaftstheorie I*. Mannheim, pp. 291–2

Spiess O 1955 Jacob Bernoulli. In: *Neue Deutsche Biographie II*. Berlin, pp. 130–1

Sylla E D 1998 The emergence of mathematical probability from the perspective of the Leibniz–Jacob Bernoulli correspondence. *Perspectives on Science* **6**: 41–76

Thiele R 1990 Jacob Bernoulli. In: Gonwald S, Ilgands H-J, Schlote K-H (eds.) *Lexikon bedeutender Mathematiker*. Leipzig, pp. 48–9

K. Mainzer

Big Man, Anthropology of

Referring to achieved leadership, the term 'big man' has come to stand for a type of polity (distinguished, e.g., from types identified with inherited rank). It is associated closely with, though not limited to, the ethnography of Melanesia. While prominent in ethnography and in developmental theories from the 1950s through the 1980s, this ethnographic construct has become less important in recent years. Contemporary work reflects a methodological turn away from typology building (and the functional and developmental theories it served). Emphasis has shifted instead towards historically and culturally situated understandings of power and agency, attentive to gender and emergent class relations, and to national and transnational processes.

1. Early Work and Developing Analyses

Big man is the Anglicization of a descriptive phrase *bikpela man*—meaning 'prominent man'—common in some variants of *Tok Pisin* (an important Melanesian *lingua franca*). The term was adopted widely in post-World War II Melanesian ethnography (e.g., Oliver 1955, Read 1959, Strathern 1971) to refer to male leaders whose political influence is achieved by means of public oratory, informal persuasion, and the skillful conduct of both private and public wealth exchange. The anthropology of big men comprises both regional ethnography (the sociocultural interpretation of case materials derived from extended field research) and—as this ethnography is puzzling from a comparative perspective—theoretical debate.

1.1 Ethnography

While a variety of political systems have been observed in Melanesia from the early colonial period to the present day, the big man is a notable feature of political life in highland regions of Papua New Guinea and Irian Jaya. These regions are distinctive for high population densities and settled, loosely kin-centered communities (e.g., 'clans' associated with named 'places'). Their horticultural economies characteristically are based on intensive tuber (sweet potato, taro) cultivation and pig raising.

Big man leadership has intrigued anthropologists as a vantage for understanding how economic intensification might be possible in the absence of institutionalized political structures. As it has been observed ethnographically, highland Melanesian sociopolitical life is decentralized, informal, and participatory in spirit. That is, access to garden land is universal (a flexible concomitant of kin relationships), and systematic differences in access to subsistence means, spouses (and other household labor resources), and valuables are muted or absent. While direct participation in clan events is a more or less exclusively male prerogative and affects male social standing, it is not obligatory. Even for men, clan interests do not take precedence necessarily over personal relationships with kin and other exchange partners, in which women are also involved. Finally, in many communities, leadership is a personal achievement associated with organizing events held in the names of clans and tribal alliances. Leaders neither inherit their status by virtue of seniority, lineage membership, or ritual sanction, nor are they formally elected or instated. Communities do not typically institutionalize decision-making councils or offices that must be filled.

Despite the lack of institutionalized political structures, during precolonial and early colonial times communities did affiliate themselves regionally for warfare and ritual performances (e.g., initiations and fertility cults). People also came together frequently to sponsor competitive exchanges of indigenous valuables (e.g., pigs and pearlshells), events in which thousands participated as recipients and donors. Periodic exchange festivals of different scales and degrees of social and political complexity have driven intensive garden production for generations in the highlands. During the colonial period and nowadays, exchanges also involve money and introduced valuables (e.g., cows, trucks), and may partially motivate participation in wage-work and commercial endeavors. Held to mark important events like deaths, and to constitute or reorder regional alliances between groups, these festivals require skillful organizing and long-term planning, as well as an intensive production base.

Early in Highlands research Read (1959) noted a tension between culturally sanctioned assertions of personal autonomy and of collective purpose. Western constructs resolve an apparently analogous relation by subsuming naturalized 'individual' interests within institutionalized social structures (like the state). In contrast, Melanesian cultures implicate both differentiating and collectivizing interests as social possibilities in a distinctive concept of personhood. Autonomy is enacted in the elaboration of personal networks of kin and exchange partners (affines, friends). More or less idiosyncratic, each person's social network embodies interests that converge partially—but may regularly conflict—with the collective projects of clans.

Big men make their names by successfully orienting their clansmen to collective ends. While the means they use varies in different parts of highland Melanesia, their power is personal and ephemeral relative to leadership power in societies with inherited rank or with formal councils. This absence of structurally reproduced power gives Melanesian communities a reputation for 'egalitarianism' (among men, if not between men and women). Men with leadership ambitions work to develop personal access to resources within and outside their communities. They achieve fame and influence by using local and regional social networks as bases for organizing collective wealth prestations: events that make and remake clans and tribal alliances. By means both of public oratory and private persuasion, big men work to add a collective significance to their own and other's actions, which would otherwise be construed only as diversely personal.

1.2 Political Typologies and Developmental Puzzles

In post-World War II anthropology, ethnographic accounts from Melanesia and Polynesia were influential in the modeling of sociopolitical ideal types (respectively exemplifying 'tribes' and 'chieftainships'). Such typologies were central to theories of cultural evolution (development), of special interest to archeologists but also prominent over the past century both within and outside of sociocultural anthropology. Developmental theories assumed that less economically productive, less socially differentiated and politically centralized social types give way, over time, to more productive, more centralized ones. They posited a series of functional interdependencies among variables like population density, technology, the organization of production, 'surplus' production, and sociopolitical stratification.

Sahlins's (1963) typological comparison of the Polynesian chief and Melanesian big man—perhaps the single most influential argument concerning Pacific polities—established the figure of the leader as key to arguments concerning their historical fortunes. With Polynesian chiefs as a standard, Sahlins emphasized the limited coercive power of the big man in mobilizing wealth for public prestations, and identified these limits in the refusal of his clansmen (understood as his political supporters in regional inter-clan prestations) to put up with his increasingly unreciprocal behavior.

Viewed from this perspective, evidence of regular, large-scale prestations of pigs, pearl shells, and other wealth in highland Melanesia was puzzling. Ethnographic arguments by Mervyn Meggitt, Andrew Strathern, and others subsequently demonstrated that—while relations between men (dominant in wealth exchange) and women (mainstays of food and pig production) were conflictual—relations between leaders and their fellow clansmen were not predominantly extractive, but were mitigated by big men's ability to tap labor and other resources in groups other their own by means of exchange networks.

Meanwhile, a convergent line of research—bent more on understanding economic change than sociopolitical systems—developed a comparison between Melanesian exchanges and capitalist markets. Dubbing Melanesians 'primitive capitalists,' this work emphasized apparent similarities between local orientations and Euro-American cultural values like individual achievement, competition, material wealth, and investment. For example, Finney (1973) argued that the similarities culturally and psychologically 'preadapted' the big man, in particular, for capitalist development. Indeed, during the colonial period, highland Papua New Guineans were notably entrepreneurial (rather than simply spending cash incomes on consumer goods).

1.3 Big Man on the Margins

These lines of research came to an ambivalent resolution in the late 1980s (e.g., Godelier and Strathern 1991). Most prominently, Maurice Godelier centered

attention on the structure of marriage exchanges in an argument meant to suggest that the big man type—with its emphasis on the manipulation of wealth—is not as typical of highland Melanesia as earlier ethnography implied it was. In communities where marriage conventions de-emphasize bridewealth (transfers of wealth for persons or their capacities) in favor of 'sister exchange' (transfers of persons for persons), Godelier proposed that one observes 'Great Men' not big men: varieties of male prominence (notably initiation cult leadership) founded on ritual expertise.

Godelier's position followed the classic pattern of using ethnographic cases to construct political economic types as elements in a hypothetical developmental sequence. It made diverse cases comparable (capable of being organized as a progression) by assuming their common commitment to the maintenance of male collective (clan) interests. Also like earlier work, it placed the big man type structurally on the frontier of market capitalism. However, these approaches were already under siege, ethnographically and theoretically, as they were articulated in the mid/late 1980s.

2. Recent Trends

The progressive decline in research on big manship since the mid-1980s echoes a now long-standing disciplinary trend away from typological, developmental/functional comparison and toward nuanced cultural interpretation in the context of historical and ethnographic analyses of intra- and intercultural engagements. On the whole, Melanesian anthropology has not only reflected but also spearheaded these transformations in the 'culture' concept and its uses. In this reworked arena, the rich lines of research that the puzzles of big manship stimulated have been absorbed into other projects.

These projects include, for example, increasingly serious attention to gender relations and meanings and to political economic transformations. Melanesian gender studies direct attention to divergent perspectives and relations within and between communities. They render the male-centered (not to say leader-centered) typification of cultures analytically unusable and big manship a decidedly qualified value (e.g., Lederman 1986, 1990, Godelier and Strathern 1991).

Complementary work seeks insight into the contradictory engagements of men and women, differently positioned in emerging class relations, as Melanesians situate themselves in a roiling global economy (e.g., Gewertz and Errington 1999).

Attention to the adoption of unfamiliar class and national identities has in turn sharpened understanding of the reinvention of culturally familiar relations, values, and meanings. Thus, studies of the articulation of market and gift exchange relations, and their associated political cultures, have clarified their differences even as their mutual entanglements have elaborated over the past generation. For example, one highlands Papua New Guinea people personify the state as a big man to insist on a relationship of equivalence, not hierarchy, between the 'local' and 'national' (Clark 1992). Ethnographic observations like this show what is at stake for Melanesians as the intramale egalitarianism associated with big manship confronts—ambivalently, unevenly, and contingently—structural inequalities associated with state and global involvement, as well as new versions of familiar alternatives and conflicts (notably those between men and women: Sexton 1993).

These studies no longer identify themselves as contributions to the anthropology of the big man or functional/developmental typology building. Instead, they aim to contribute to a more thoroughly historical perspective on the politics of shifting contexts for meaningful action. The rise and ultimate dispersion of big man studies was characteristic of Melanesian anthropology over the past century. A similar interplay of local interpretive ethnography and comparative cultural analysis will continue around political values well into the twenty-first century.

See also: Economic Anthropology; Exchange in Anthropology; Exchange: Social; Kula Ring, Anthropology of; Melanesia: Sociocultural Aspects; Trade and Exchange, Archaeology of

Bibliography

Clark J 1992 Imagining the state: Idioms of membership and control in the Southern Highlands Province of Papua New Guinea. Paper presented at the conference 'Nation, Identity, and Gender,' Australian National University, Canberra, ACT

Finney B 1973 *Big-Men and Business: Entrepreneurship and Economic Growth in the New Guinea Highlands*. Australian National University Press, Canberra, ACT

Gewertz D, Errington F 1999 *Emerging Class in Papua New Guinea: The Telling of Difference*. Cambridge University Press, Cambridge, UK

Godelier M, Strathern M (eds.) 1991 *Big Men and Great Men: Personifications of Power in Melanesia*. Cambridge University Press, Cambridge, UK

Lederman R 1986 *What Gifts Engender: Social Relations and Politics in Mendi, Highland Papua New Guinea*. Cambridge University Press, Cambridge, UK

Lederman R 1990 Big men large and small? Towards a comparative perspective. *Ethnology* **29**: 3–15

Oliver D 1955 *A Solomon Island Society*. Cambridge University Press, Cambridge, UK

Read K E 1959 Leadership and consensus in a New Guinea society. *American Anthropologist* **61**: 425–36

Sahlins M 1963 Poor man, rich man, big man, chief: Political types in Melanesia and Polynesia. *Comparative Studies in Society and History* **5**: 285–303

Sexton L 1993 Pigs, pearlshells, and 'women's work': Collective response to change in Highland Papua New Guinea. In: Lockwood V S et al. (eds.) *Contemporary Pacific Societies* Prentice-Hall, Englewood Cliffs, NJ, pp. 117–34

Strathern A 1971 *The Rope of Moka: Big Men and Ceremonial Exchange in Mount Hagen, New Guinea*. Cambridge University Press, Cambridge, UK

R. Lederman

Bilingual Education: International Perspectives

The majority of the world's population speaks more than one language. Given the human cognitive capacity of managing multiple linguistic systems, curricula employing two or more languages of instruction and drawing on the linguistic and cultural resources of bilingual or multilingual individuals should find wide acceptance. In reality, the situation is more complicated. After a definition of bilingual education, an overview of the objectives and major types of programs is offered, and parameters determining success are identified. Throughout, L1 refers to the child's first language(s) and L2 to any languages acquired after age 3 years.

1. *Definition*

For educational programs to qualify as 'bilingual,' two conditions should be met: (a) more than one language serves as medium of instruction and (b) bilingualism and biliteracy are explicit goals. In practice, however, discussions of bilingual education often relax these conditions and include policies considering bilingualism a transitional state.

2. *Settings and Objectives*

Since the 1960s, bilingual curricula have been developed all over the world (see the comprehensive overview in Baker and Jones (1998) and García (1997)). Sometimes, as in Asia and Africa, more than two languages are involved. Besides globalization and the need for intercultural communicative competence, the following factors can be singled out as most conducive to this development: (a) co-existence of two or more official languages, as in Belgium, Canada, Luxembourg, and Switzerland; (b) co-existence of different local languages with a non-indigenous (colonial) language functioning as a neutral official language, such as English in Ghana or French in the Ivory Coast; (c) in-migration and high drop-out rates of minority children; and (d) revival of interest in ethnic cultures and languages.

Against this background, at least three educational objectives can be identified:

(a) assimilation of the child of in-migrant or indigenous minority groups into mainstream society;

(b) maintenance and development of the first language of minority children; and

(c) enrichment and empowerment of minority and majority children.

3. *Types of Bilingual Education*

Baker (1996) (see also Baker and Jones 1998) distinguishes 'strong' and 'weak' educational policies, with only the former satisfying both conditions mentioned in the definition.

3.1 *Strong Forms of Bilingual Education*

An intensively researched program is the *immersion* of groups of majority children in a second language, such as English speakers in French programs in Canada, which may begin in kindergarten or at later grade levels, and be partial in the beginning (with only some subjects taught in L2) or total, with a shift of some contents to L1 in later years.

In *maintenance* or *heritage* programs, minority children are taught in their L1 at least 50 percent of the time. Successful programs have led to the strengthening of Navajo in the USA, of Catalan, Gaelic, Finnish, and Welsh in Europe, of Maori in New Zealand, and of aboriginal languages in Australia.

In *two-way* or *dual-language* programs, which teach through both minority and majority languages (or rely on more than one majority language), balanced numbers of native speakers share a classroom. Languages alternate, either by subject, day, or some other consistent principle.

Second-language medium instruction is also an important feature of *International* and *European schools*, which, like the Swiss 'finishing schools,' are typically attended by children of socioeconomic and intellectual elites.

3.2 *Weak Forms of Bilingual Education*

The 'sink or swim' policy of *submersion* is the most widespread way of dealing with minority children, both in-migrant and indigenous. The child is placed in mainstream classrooms, sometimes assisted by additional instruction in the majority language (in withdrawal or pull-out classes). 'Structured immersion' L2 classes, not to be confused with the immersion program for majority children (e.g., in Canada), contain only minority children. *Transitional* bilingual education starts by teaching in the minority child's L1 and as soon as possible moves over to instruction via L2. It is possible to distinguish 'early exit' (after 2–3 years) and 'late-exit' options.

In many countries, in-migrant children may well experience a mix of different methods. Parental initiatives, religious organizations, and consulates often provide additional L1 classes focusing on language development, ethnic history, and culture.

4. Findings and Problems

Bilingual education programs are subject to many interacting variables. Because of differences in expectations and biases, different evaluations of findings come as no surprise. Indeed, it is difficult to decide which dependent variables (such as degree of competence in various language skills, achievements in other subjects, self-esteem, success in mainstream culture, etc.) are crucial and how they should be measured and compared across studies (compare the recent meta-analysis in Greene (1998)). Nevertheless, a number of conclusions can be drawn (Baker 1996, Baker and Jones 1998, Brisk 1998, Crawford 2001, García 1997, Hakuta 1986).

Prestige of the languages involved and majority/minority status of the child appear to be most significant, as shown in French immersion for English-speaking Canadian children. Positive results can also be achieved where the school locally grants equal status to different languages and where the numbers of majority and minority children are balanced, as in American Spanish–English dual-education programs. In addition to highly qualified and committed bilingual teachers, parental and community support has been vital in all successful cases. Of course, in terms of funding, what matters most is a society's attitude. Xenophobia and concerns about social disintegration often taint discussions and sway political decisions against providing the funds needed for quality curricula.

Scientific studies show that successful *immersion* education of majority children is possible without damage to the child's L1 and without deficits in other academic areas taught in L2. Within the spectrum of immersion options (early vs. late, total vs. partial), best results are achieved in early total immersion. *Maintenance* programs also show promising results. However, while they support in-migrant children's L1 and ethnic identity, they may also segregate these children from their majority peers. An advantage of *two-way/dual-language* schools is that they bring together students from both cultural backgrounds and create a forum for interaction.

Submersion is generally viewed critically (e.g., Baker and Jones 1998, Greene 1998). It leaves the average child with limited L2 proficiency and therefore does not provide the resources needed for coping with other academic subjects. Both *submersion* and early-exit *transition* programs may lead to subtractive effects and eventual loss of L1 without compensatory integration into mainstream society. Moreover, shift to L2 at the expense of L1 may result in alienation from the family.

According to Cummins (1991), for additive rather than subtractive effects to become possible, critical thresholds of L1 proficiency and cognitive development have to be attained; the more advanced children's L1, the better is the basis for positive transfer of skills (including literacy) and concepts acquired through L1.

5. From Paranoia to Empowerment Pedagogy

Bilingual education must be seen against the backdrop of conflicting political ideologies and different attitudes towards bilingualism in general. As pointed out by Brisk (1998, p. 1), 'The paradox of bilingual education is that when it is employed in private schools for the children of elites throughout the world it is accepted as educationally valid […]. However, when public schools implemented education for language minority students over the past 50 years, bilingual education became highly controversial.' Opponents of US bilingual education, in particular, view bilingualism not only as 'a costly and confusing bureaucratic nightmare' (Hayakawa 1992, p. 44) but also as likely to lead to political disloyalty and destabilization.

In his overview of the history of bilingual education in the USA, Baker (1996) distinguishes four overlapping periods characterized by (a) permissiveness, (b) restrictiveness, (c) opportunity, and (d) dismissal (see also Baker and Jones 1998, García 1997, Crawford 2001). Before World War I, *tolerance* of linguistic diversity prevailed. Many ethnic communities taught children in languages other than English. But the Nationality Act of 1906 required immigrants to the USA to learn English in order to become naturalized citizens. World War I and increased numbers of immigrants led to calls for assimilation and Americanization (an attitude strongly reinforced by World War II) and to *restriction* of public-school instruction in languages other than English. This attitude relaxed in the wake of the Civil Rights and Equal Education movements and the growing interest in ethnic values and traditions. In addition, the academic failure of limited English proficient children raised the issue of equal *opportunity* and compensatory measures. Courts mandated bilingual education for minority children; amendments to Education Acts made federal funds available. The *dismissive* period starting in the 1980s took another turn in favor of submersion and transition programs, supported by pressure groups such as US English, English First or English for the Children (see respective websites).

From an international perspective, bilingualism and bilingual education have a more promising future, with countries envisioning themselves as global players aware of their need for interculturally competent mediators. Israel, for example, has moved from submersion to a successful multilingual curriculum. Within the European Community, new employment opportunities have led to an increasing demand for internationally accredited degrees and to recognition of the desirability of L2 skills. More and more parents of majority children are seeking bilingual programs at the nursery and primary school levels. While the current focus is on internationally marketable languages (and the languages of immediate neighbors, as along the border of France and Germany), it is to be expected that this interest will reinforce bilingualism in

general and contribute to the development of effective pedagogical concepts of empowerment for majority and minority children alike.

6. Conclusion

Supporters of bilingual education have to contend not just with those who fear that multilingualism and cultural diversity lead to societal disintegration, but also with those who are concerned about negative consequences for children's cognitive, linguistic, and emotional development. Research has shown that under favorable conditions children can thrive in bilingual programs, at no expense to academic excellence in other domains and without risk for their mother tongue. These results are reinforced by research on bilingualism in general and by the successful simultaneous acquisition of two first languages in early childhood (Grosjean 1982, De Houwer 1996). Where problems arise, they should not be blamed on bilingualism *per se* but rather on the conditions under which it develops.

In the past, schools have often not just served as agents of submersion but prohibited and punished the use of minority languages on their grounds (Grosjean 1982, pp. 27ff). Given our changing world, schools are now called upon to meet the challenge of bilingualism not just as an educational objective but as a school ethic, with highly qualified educators as believable role models. This in turn requires institutions of higher education and their researchers to take more active roles in communicating their findings to practitioners, parents, politicians, and the public.

See also: Bilingualism and Multilingualism; Bilingualism: Cognitive Aspects; First Language Acquisition: Cross-linguistic; Language Acquisition; Language Policy: Linguistic Perspectives; Language Policy: Public Policy Perspectives; Second Language Acquisition

Bibliography

Baker C 1996 *Foundations of Bilingual Education and Bilingualism*. Multilingual Matters, Clevedon, UK

Baker C, Jones S P 1998 *Encyclopedia of Bilingualism and Bilingual Education*. Multilingual Matters, Clevedon, UK

Brisk M E 1998 *Bilingual Education: From Compensatory to Quality Schooling*. Lawrence Erlbaum, Mahwah, NJ

Crawford J 2001 Language politics in the U.S.A.: The paradox of bilingual education. In: Ovando C, McLaren P (eds.) *The Politics of Multiculturalism and Bilingual Education: Students and Teachers Caught in the Cross-Fire*. McGraw-Hill, New York

Cummins J 1991 Interdependence of first- and second-language proficiency in bilingual children. In: Bialystok E (ed.) *Language Processing in Bilingual Children*. Cambridge University Press, New York, pp. 70–89

De Houwer A 1996 Bilingual language acquisition. In: Fletcher P, MacWhinney B (eds.) *Handbook on Child Language*. Blackwell, London, pp. 219–50

García O 1997 Bilingual education. In: Coulmas F (ed.) *The Handbook of Sociolinguistics*. Blackwell, Cambridge, MA, pp. 405–20

Greene J P 1998 *A Meta-analysis of the Effectiveness of Bilingual Education*. http://ourworld.-compuserve.com/homepages/JWCRAWFORD/-greene.htm

Grosjean F 1982 *Life with Two Languages*. Harvard University Press, Cambridge, MA

Hakuta K 1986 *The Mirror of Language. The Debate on Bilingualism*. Basic Books, New York

Hayakawa S I 1992 Bilingualism in America: English should be the only language. In: Goshgarian G (ed.) *Exploring Language*. Harper Collins, New York, pp. 42–7

R. Tracy

Bilingualism and Multilingualism

1. Introduction

The topic of competence in more than one language is large and complex and—because it is both an individual and a societal phenomenon—it has an extensive literature within sociolinguistics and the sociology of language. The brief overview provided here can do nothing more than alert the reader to the most relevant strands of the subject. Many important matters—growing up bilingual, informal second-language acquisition, formal second-language learning at school—have to be omitted altogether. Matters of definition, degree and assessment are dealt with first: what is bilingualism and how can its strength be assessed? Second, are there different types of bilingualism—if so, how are the varieties best understood? Third, how are bilingual or multilingual fluencies related to personal characteristics such as age and intelligence? Fourth, how are abilities across languages interrelated? Here, questions of cross-variety borrowing and interference, and 'code-switching'—the use of more than one variety in the same context—are considered. Fifth, there is a brief look at how multiple fluencies arise and are sustained at the social level, and at how these necessitate accommodations—the use of translation, or the emergence of lingua francas, for example.

2. Bilingualism and Multilingualism: Definition, Degree and Assessment

Giuseppe Mezzofanti, chief curator in the Vatican Library in the early nineteenth century, was reportedly fluent in 60 languages, and had some translating ability in about twice that number. Georges Schmidt,

head of the terminology division at the United Nations in the 1960s, knew about 20 languages. The contemporary writer George Steiner claims full and equal competence in English, French, and German. Your cousin in Birmingham can say only *c'est la vie* or *guten tag* or *por favor*. How are these fluencies related? Mezzofanti would seem to be more multilingual than Steiner, for instance—but perhaps the latter has a native-like depth in his three varieties that exceeds Mezzofanti's more superficial, if broader, abilities. Would we then situate Schmidt somewhere between these two and, if so, on what grounds? And what of your Birmingham cousin—does she count as bilingual at all?

In his seminal *Language* (1933), Leonard Bloomfield described bilingualism as the addition of a perfectly learned foreign language to one's maternal and undiminished variety (he did acknowledge that 'perfection' was difficult to assess). Uriel Weinreich, in *Languages in Contact* (1953), noted more simply that bilingualism was the alternate use of two languages. In the same year, Einar Haugen suggested that bilingualism involved the ability to produce complete and meaningful utterances in the second medium. These definitions are representative of a much larger collection, and show that bilingualism has labeled fluencies at the Schmidt or Steiner end of the continuum—but also at the *guten tag* level. In general, earlier definitions tended to restrict bilingualism to equal mastery of two languages (or more, in the multilingual individual), while later ones have allowed for greater variations in competence. Since, however, any relaxation of the concept's status—to cover your Birmingham cousin, for instance—can obviously prove as unsatisfactory a guide to some sort of general definition as earlier arguments from perfection, most contemporary scholars acknowledge that any meaningful discussion can only proceed within specific contexts and for specific purposes.

Further complicating the question of where bilingualism may be said to start is the fact that any definitional lines drawn must cross several specific threads of ability—not just one overall language dimension. Consider, first, that there are four basic language skills: listening, speaking, reading, and writing. Consider further the possible subdivisions: speaking skill, for example, includes what may be quite divergent levels of expression in vocabulary, grammar, and pronunciation. Overall, there are at least 20 linguistic dimensions or features which could or should be assessed when determining bilingual proficiency (Mackey 1962, Weinreich 1953).

We have moved here from definition to assessment—and it follows that, if the former is problematic, the latter becomes difficult. Nonetheless, a number of standard measures of bilingualism have emerged; these include rating scales, and tests of fluency, flexibility, and dominance. In the first category are interviews, language-usage measurements and self-assessment procedures. In some ways, relying upon self-ratings has a lot to recommend it, but the strengths here rest upon the capacity of an individual to be able to self-report accurately, a roughly equivalent sense across respondents of what competence means, and (of course) a willingness to reveal proficiency levels. The possibilities of error here are not entirely absent from apparently more objective tests. One type involves asking people to respond to instructions in more than one language, measuring their response times and, on this basis, attempting to ascertain which variety is dominant. As well, subjects can be presented with picture-naming or word-completion tasks, can be requested to read aloud, or can be asked to pronounce a word common to more than one language (*pipe*, for instance). More straightforward assessments involve measuring the extent of vocabulary, or seeing how many synonyms can be produced, and so on. A great deal of research shows that, although the results of such measurements often intercorrelate, the tests are far from perfect.

3. *Varieties of Bilingualism*

Even if it were possible to accurately measure bilingual and multilingual abilities, labeling problems would remain; after all, it is hardly to be expected that individuals would neatly fall into a small number of categories. To complicate matters further, different writers have used different terms in trying to capture degrees and types of linguistic fluency. For example, balanced bilingual, ambilingual, and equilingual have all been used to designate speakers whose bilingual capacities are great. Baetens Beardsmore (1986) described the ambilingual as one who—in all settings—can function equally well in either variety, and who shows no trace of one language while speaking the other. Since, however, such an individual is rare (if not actually nonexistent), then the balanced or equilingual designations become more likely, insofar as they refer to more roughly equivalent abilities—which need not approach perfection. Of course, most bilingual or multilingual speakers do not even warrant the labels balanced or equilingual—far from it, since most bilingual capacities are quite uneven and usually very susceptible to context and topic.

A useful distinction has been made between receptive (or passive) and productive (or active) bilingual competence. Some can understand a language—spoken or written—without being able to produce it themselves (that is, in anything other than the most halting and rudimentary manner); others can do both. Receptive competence alone has also been referred to as semibilingualism, which does have a certain logic. (Unfortunately, however, the term is sometimes confused with another, semilingualism, which means something else entirely: a lack of fluency in either of two languages. That any speaker of normal intel-

ligence should speak 'no language tolerably'—as Bloomfield (1927) famously observed of an Indian informant—is now widely discredited.)

Another important distinction is that between additive and subtractive bilingualism. In some circumstances, for example, learning a new language represents an expansion of the linguistic repertoire; in others, it may lead to a replacement of the first. At the simplest level, additive bilingualism is the outcome when both languages continue to be useful and valued—consider the bilingualism of academics, of social élites, or aristocracies. Subtractive bilingualism, on the other hand, can be expected in social settings in which one variety is more valuable or dominant than the other. There is little need to maintain two languages indefinitely, if one comes more and more to serve all functions in all domains.

A third dichotomy involves primary and secondary bilingualism. The first reflects the acquisition of a dual competence which has come about naturally—through early learning in the bosom of the family, for example, or because of real and obvious social–contextual demands. The second refers to abilities gained through more self-conscious efforts, notably systematic and formal instruction at school. These are not watertight compartments: one could pick up a good conversational command of a second language at work, say, and then add more formal reading and writing skills through education. This is, incidentally, redolent of the process by which a mother tongue is developed, and it is noteworthy that the more enlightened educational language curricula have tried to capture this in their second-language programs. Nonetheless—and especially at the societal level—there are some obvious differences between primary bilinguals (native speakers of Gaelic in the Scottish highlands and islands, for example) and secondary ones (those who have made a formal commitment to learn Gaelic at school in Glasgow or Edinburgh.

4. *Age and Intelligence*

The expansion of linguistic repertoires—either actual or potential—is often seen to be related to age and intelligence. Smart, young people are popularly considered to possess an advantage here, while older and less gifted individuals find things rather more difficult (There is also the view, particularly prevalent in Anglophone societies, that whole groups can be 'poor at languages,' but this is another matter entirely, albeit an interesting one.) There is some neurological justification for the idea that the young brain is more plastic and flexible than the older one, and hence acquiring another language (or anything else, for that matter) is easier. On the other hand, an overemphasis upon early acquisition and brain malleability—and, more pointed still, the idea of a 'critical period' for languages—can be criticized. Older learners, for example, obviously have a wealth of cognitive experience lacking in younger ones and, provided the motivation is sufficient, often prove to be faster and more fluent learners. If one could combine the maturity and necessity of the older with the imitativeness and spontaneity of the young, one would surely have a recipe for proficient bilingualism.

Many have understood brain capacity to be finite. Two implications follow from this: first, the more capacity you have, the more knowledge (of languages, for example) you can develop; second, there is a danger that taking in something new means losing something already there. At the simplest level, this is reminiscent of the old craniometric axioms that crudely linked brain size to intelligence. At a more subtle level, it is now evident that accepting a finite-container model need have no implications for most people—the brain is large enough to eliminate any worries about exceeding its limits. This is a relatively modern notion, however, and one need not go back very far to find prominent scholars expressing reservations about bilingualism. Firth noted, in 1930, that while bilinguals have two strings to their bow, one is slacker than the other; unilinguals, he felt, had the advantage (see Firth 1970). Jespersen, in his *Language* (1922), observed that bilingualism involved considerable cost—a sort of semilingualism could well result, and a brain overly devoted to languages must necessarily become less available for other tasks.

These opinions were once, as already implied, part of the prevailing intellectual discourse. In general, early studies—particularly those that were part of the American intelligence-testing movement during the great waves of European immigration—tended to associate bilingualism with lowered intelligence. These studies were typically crude and methodologically deficient, and were undertaken at a time when fear and prejudice were rampant. There is space here for only one example; in 1926, Goodenough observed that 'the use of a foreign language in the home is one of the chief factors in producing mental retardation.' It was, of course, just as noticeable in the 1920s as it is today that most people in the world are bilingual or multilingual—not perfectly 'balanced,' perhaps, but not retarded, either.

Later and more careful research tended to show essentially no relationship between intelligence and bilingualism, while investigations beginning in the 1960s demonstrated a positive one. An important study here was that of Peal and Lambert (1962), in which bilingual children outperformed their monolingual counterparts on intelligence tests. The authors concluded that bilingualism was associated with greater mental flexibility and superior conceptual skills. Importantly, though, they acknowledge that their work could not illuminate the causal direction of the apparent bilingualism–intelligence link. As well, cogent criticisms have been made of such studies, on methodological grounds and—more commonly—on

the basis of the nonrepresentativeness of the subjects (usually children, and usually more or less 'balanced' bilinguals). A further problem is that both bilingualism (as already noted) and intelligence are hard to define: consequently, trying to establish links between them is even harder. Strong conclusions about such links are obviously not warranted but it is not unfair to suggest that, if marked cognitive advantages do not flow from bilingualism, at least no cognitive price needs to be paid for it. It cannot be denied, however, that expansion of the language repertoire represents another dimension of individual capacity—at the minimum then, linguistic growth means experiential growth, in the same way that any new intellectual acquisition does.

5. *Borrowing, Interference, and Code-switching*

When a person possesses more than one language, there are obvious possibilities for mixing within a single conversation. The term 'interference' suggests unwelcome or unthinking intrusions from one variety into another. However, while this undoubtedly occurs in the speech of bilinguals, switches often take place for more positive reasons: for emphasis, because the *mot juste* is found in only one of languages, or because of the setting in which conversation occurs (where relevant matters include topic differentiation, variable linguistic competences of interlocutors, desired levels of intimacy or formality, and so on). Some, therefore, have felt that the more neutral 'transference' is an apter term. In any event, code-switching is extremely common and is generally a nonrandom occurrence, triggered by factors which—while not always immediately obvious—usually reveal themselves to careful analysis.

'Sometimes I'll start a sentence in English *y termino en español*' (Poplack 1980) is both an example and the title of an article which attempts to categorize code switches.

The situation of languages in contact often leads to more permanent transferences than those which the previous paragraph has implied. At the lexical level, for instance, one might find a Brussels francophone using the Dutch *vogelpik*, rather than the French *fléchettes* for a game of darts. In this context the Dutch term is a loanword, an item borrowed and used in unchanged form (although, of course, it may be given some elements of a French pronunciation). Another variety of lexical transfer involves loan translation—where the English *skyscraper* becomes *wolkenkrabber*, *wolkenkratzer*, *gratte-ciel* and *rascacielos* (in Dutch, German, French, and Spanish, respectively). Sometimes a word is more fully embraced by another language: in morphological transfer, the Dutch *kluts* ('dollop') and *heilbot* ('halibut') become *une clouche* and *un elbot* in Belgian French. It is also common to find syntactic transfer: for instance, a Dutch speaker might say, in French, '*Tu prends ton plus haut chiffre*'—making his adjectives precede the noun, as they would in Dutch (*Je neemt je hoogste cijfer*), but not as they do in French. Extremely common is phonological transfer—think of fluent adult speakers with 'horrible' accents in their nonmaternal variety.

6. *Multilingualism and its Social Consequences*

Simple observations of the number of languages in the world, and their degree of spread and contact, suggest that bilingual and multilingual competences are commonly required. The forces which give rise to these phenomena are, in general, very easily understood. They include migration, imperialist or colonialist expansion, sociopolitical union among different language groups, the long-term existence of linguistic border regions. In these scenarios, language competences expand (or, occasionally, contract) as a consequence of more basic changes. It is also possible for multiple competences to arise on a more purely voluntary basis—cultural and educational motivations, for example, can expand repertoires, even if they are unaccompanied by either the desire or the possibility to use the new acquisition in ordinary conversational ways.

Multilingualism as a social phenomenon implies both heightened and lessened opportunities for interpersonal exchange. Individuals who know each other's languages can obviously converse, but it is equally clear that a world of many languages is a world in which communicative problems abound. Historically speaking, two major methods of bridging language gaps have exited: translation, and the use of lingua francas. The latter can be conveniently divided into three varieties. First, there are existing languages which have achieved regional (or global) power or status—the so-called 'languages of wider communication.' Greek and Latin were the classical lingua francas—mediums which all educated persons knew (in addition, of course, to their own local speech). Other important lingua francas have included Arabic, Italian, French, and English. Indeed, the status of the last has so far outstripped that of all others that it now poses—in many people's eyes—a distinct threat to many other languages.

Second, limited forms of languages have arisen as lingua francas in settings (trading settlements, for instance) in which ease of acquisition and simple communication are the imperatives. A pidgin variety, for example—often reflecting a mixture of some European colonial language and an indigenous form—is one in which vocabulary and grammar are restricted. Many pidgins have a relatively short life, but some achieve considerable longevity and become expanded and enriched themselves; they have then evolved into creoles. Nobody's mother tongue has become somebody's mother tongue.

Third, the idea of an 'artificial' or 'constructed' language has appealed to many as the ideal lingua franca. Esperanto is the most well-known constructed language but, in fact, hundreds have been developed over the centuries. They are usually quite easy to learn—their vocabularies and grammars are typically free of irregularities—and they have logical appeal as more or less neutral universal mediums. None, however, has managed more than a vestigial existence; apart from certain technical criticisms and some unfortunate associations, it would seem that most people have simply not taken them seriously.

The other great bridge over multilingual chasms is translation and, although much can be said about it, it is only its existence which really needs noting here. Translators and their work are indeed a bridge between linguistic solitudes, but perhaps it is not surprising that their obvious usefulness has sometimes occasioned misgivings. The translator necessarily has a foot in more than one camp and, as Steiner (1992) noted, may take 'hoarded dreams [and] patents of life across the frontier.' The old Italian proverb was blunter: *traduttore*, *traditore*.

See also: Bilingual Education: International Perspectives; First Language Acquisition: Cross-linguistic; Foreign Language Teaching and Learning; Language Acquisition; Language Contact; Language Development, Neural Basis of; Second Language Acquisition

Bibliography

Baetens Beardsmore H 1986 *Bilingualism: Basic Principles*. Multilingual Matters, Clevedon, UK

Baker C, Jones S 1998 *Encyclopedia of Bilingualism and Bilingual Education*. Multilingual Matters, Clevedon, UK

Bloomfield L 1927 Literate and illiterate speech. *American Speech* **2**: 432–9

Bloomfield L 1933 *Language*. Holt, New York

Edwards J 1995 *Multilingualism*. Penguin, London

Firth J 1970 *The Tongues of Men and Speech*. Oxford University Press, London

Goodenough F 1926 Racial differences in the intelligence of school children. *Journal of Experimental Psychology* **9**: 388–97

Grosjean F 1982 *Life with Two Languages*. Harvard University Press, Cambridge, MA

Hakuta K 1986 *Mirror of Language*. Basic, New York

Hamers J, Blanc M 1983 *Bilingualité et Bilinguisme*. Mardaga, Brussels, Belgium

Haugen E 1953 *The Norwegian Language in America*. University of Pennsylvania Press, Philadelphia, PA

Hoffmann C 1991 *An Introduction to Bilingualism*. Longman, London

Jespersen O 1922 *Language*. Allen & Unwin, London

Kelly L 1969 *Description and Measurement of Bilingualism*. University of Toronto Press, Toronto, ON

Mackey W 1962 The description of bilingualism. *Canadian Journal of Linguistics* **7**: 51–85

Paulston C 1988 *International Handbook of Bilingualism and Bilingual Education*. Greenwood, Westport, CT

Peal E, Lambert W 1962 The relation of bilingualism to intelligence. *Psychological Monographs* **76**: 1–23

Poplack S 1980 Sometimes I'll start a sentence in English *y termino en español*: Toward a typology of code-switching. *Linguistics* **18**: 581–616

Romaine S 1995 *Bilingualism*. Blackwell, Oxford, UK

Steiner G 1992 *After Babel: Aspects of Language and Translation*. Oxford University Press, Oxford, UK

Weinreich U 1953 *Languages in Contact*. Mouton, The Hague

J. Edwards

Bilingualism: Cognitive Aspects

'Bilingualism' is the regular use of two languages, and bilinguals are those people who use two languages. More specifically, bilingualism refers to the individual competence of comprehension and production of two (natural) languages (language variants like dialects included). Balanced full competence in both languages (balanced bilingualism) is less frequent than dominance of one language (imbalanced bilingualism). Balanced imperfect competence in both languages is labeled semibilingualism. The two languages may be acquired together from infancy on (early bilingualism) or may be learned sequentially (late bilingualism or first language, L1, acquisition and second language, L2, learning). The first language is not always the dominant language when viewed over the life span. The first language of childhood tends to be lost when the child has to emerge into a second language environment without a chance to develop and use its first language further. In general, bilingualism also includes cases with one active and one inactive language. Basic cognitive aspects of bilingualism include (a) issues of neural representations of bilingual language processing, (b) simultaneous or successive learning or acquisition processes of the two languages, (c) representations of linguistic forms and meanings in long-term and working memory, (d) language loss and forgetting, (e) metalinguistic awareness associated with L1 and L2 processing, and (f) code-switching and language mixing.

Two major themes have dominated since the early 1980s in these six aspects: Are a bilingual's language systems separate modules, and if not how much interaction is between them? And in what way are language forms (codes) and their meaning (knowledge structures) represented in the cognitive system? Grosjean (1992) distinguishes two fundamental views on bilingualism: a fractional (monolingual) and a holistic (bilingual) view. According to the fractional view a bilingual should have two separate language

competencies, each single competence similar to that of a corresponding monolingual. The holistic view states that the two language systems of a bilingual are processed in an interdependent way. The blending of the two languages creates a speaker with a specific competence of using the languages together or separately.

1. *Multiple Language Processing and Neurophysiological Evidence of Language Representations*

An individual's obvious capacity to process more than one language without constant interferences requires explanations by mechanisms in the cognitive subsystems responsible for language processing in general and for the separation of the languages during processing, more specifically. These cognitive subsystems have specific neuroanatomical bases. The processing of more than one language on the neurophysiological level seems to be indicated by a greater involvement of the right hemisphere when the weaker language is processed as opposed to the stronger first language (Schönpflug 2001). The involvement of deeper (older) structures of the brain for the first language acquired as opposed to the second language learned later is suggested by Paradis (1994).

Fabbro (1999) discusses differential learning mechanisms involved in the acquisition and learning of L1 and L2. Simultaneous recovery of L1 and L2 after brain injuries are cited as evidence for the involvement of the overlapping structures in the brain processing a bilingual's two languages. Differential recovery of first L1 and then L2 or vice versa are equally frequent, but less frequent than simultaneous recovery. These recovery patterns hint at partially differential organizational structures responsible for a bilingual's language processing in the brain.

2. *Acquisition vs. Learning of L1 and L2*

Learning a language without being explicitly taught is referred to as language acquisition. Language acquisition is a case of implicit learning, that is (a) the acquisition of information is casual, with little direct focalization of attention, nor voluntary concentration; (b) the information is memorized by means of implicit strategies, which do not make it available for conscious introspection; (c) information is automatically used independently of conscious control, letting the subject establish the objectives to be reached, but not consciously carry out the program (execution is automatic)—the subjects' attention can thus be focalized on the result, ignoring the process; (d) this type of learning occurs slowly and improves with practice (for procedural knowledge only). Studies have suggested that in their first year of life children have implicit learning, and in the third year there is still preponderance of implicit over explicit learning.

Several researches report a decline of the ability to learn a L2 up to a native-like performance after the age of eight to 10 (Schönpflug 2001). This suggests a biological critical period that is limited at the beginning or in early puberty. Maturational growth of the structures involved in L2 learning is observed up to puberty, followed by a decline of skills for language learning and great variation in learning effects. These findings leave still open the precise nature of such a critical period. Furthermore, in line with neurophysiological evidence, studies involving phonem discrimination in L2 as compared to L1 support the hypothesis that L1 shapes the perceptual system at early stages of development in such a way that it will determine the perception of nonnative phonemic contrasts, even if there is extensive and early exposure to L2 (e.g., Sebastian-Galles and Soto-Faraco 1999). Cutler et al. (1992) found that adult French monolinguals employ a syllable-based speech segmentation, and monolingual English speakers use a stress-based segmentation procedure. French–English bilinguals with French dominance showed syllable segmentation only with French material, whereas the English-dominant group showed no syllable segmentation in either language. However, the English-dominant group showed stress-based segmentation with English language material, whereas the French-dominant group did not. Thus the two approaches to language segmentation are mutually exclusive, as a consequence of which speech segmentation by bilinguals is, in this respect, functionally monolingual.

The parallel learning of two languages from early childhood on should have effects on the cognitive development of children which should still be noticeable in adulthood. A threshold model of bilingual language acquisition predicts that the acquisition of the two languages has to pass a certain threshold level in order to support cognitive development in other areas (e.g., Cummins 1987). A semilingual language competencies below the threshold for both languages are detrimental for the development of other cognitive domains. However, in a recent well-controlled study no relationship was found between degree of bilingualism and nonverbal intelligence, contrary to the level of bilingualism hypothesis. The results suggest that the effects of bilingualism on cognitive development are not solely dependent on the level of second language proficiency (Jarvis et al. 1995).

3. *Bilingual Language Forgetting and Language Loss*

Memory research has contributed little toward understanding maintenance or loss of complex knowledge systems like the two languages of bilinguals. This is so

because such systems are acquired and maintained over long time periods that cannot be accommodated by traditional research methods. Continued maintenance of knowledge depends on periodic access. Bahrick (1984) tested retention of an L2 learned in school over a 50-year period. The degree of language loss is predictable from the intensity and the success of the original L2 training because there is usually little rehearsal over the life span. In the first three to six years retention exponentially declines; after that there is little loss up to a period of 30 years. But maintenance of retention cannot only be reappearance of the same material but should also be seen as a (re)constructive process. This interpretation implies that metaknowledge of L2 in form of a language schema would provide the frame for reconstruction.

Clinical case studies (e.g., Fabbro 1999, Chap. 26) show that when necessary structures of the brain are still intact and language loss is observed, hypnotic treatment may disinhibit activation processes to understand and produce either a lost L1 or L2. Thus, 'lost' language systems seem to be covered or inhibited but not totally erased. Eliciting factors were familiarity, language of the environment, and affective conditions.

4. *Cognitive Processing of L1 and L2*

A central question in theories of cognitive aspects of bilingualism concerns how a bilingual's two languages are stored in long-term memory. Does each language possess its own memory store or do languages share a single representational system? At the lexical level, words in each language appear to be stored independently, but at the conceptual level, words in each language appear to access a common semantic representation. Dufour and Kroll's (1995) hierarchical model of bilingual representation in memory adds a learning component to the single conceptual store: In adults, early stages in the acquisition of second language vocabulary appear to be characterized by reliance on L1. That is, during the initial phase of L2 learning, individuals seem to rely on associations between words in their two languages in order to get access to the meaning. With greater proficiency in the second language, they become able to access concepts directly without L1 involvement. L1 and L2 connections are asymmetrical.

Additional evidence has come from cross-language conditions; they probably promote the activation of both lexical stores, and less fluent bilinguals may not be able to properly inhibit the influence of the dominant language (see also Schönpflug 2000) .

Paivio (1986) forwarded the most elaborate published model of bilingual memory functioning. Paivio's bilingual dual coding model is in one sense a specific version of the independence approach to bilingual cognition, but it also includes a common representational system that provides a basis for interpreting some findings that appear to support the dependence hypothesis. His basic bilingual memory model claims that in long-term memory representations are coded verbally and nonverbally associated with imagery. The representations associated with the two language codes are conceived to be independent on the verbal coding level. The nonverbal representations or images are shared, however. Abstract representations are predominantly verbally coded whereas concrete representations are coded verbally and in an imagery code. The two verbal systems are functionally independent of the imagery system. The assumptions imply that bilinguals can perceive, remember, and think about nonverbal objects and events without the intervention of either verbal system, and conversely, that they can behave or think verbally without constant input from the nonverbal system. Translation equivalents do not necessarily have identical referential or imaginal meanings.

Anderson (1978) has convincingly argued that the issue of representational format cannot be answered independently of the processes used in accessing those representations. Depending on the particular assumptions chosen by a theorist for accessing memory, results often taken to support imaginal coding might equally well be interpreted within a propositional theory, and vice versa. But no need exists to limit coding dimensions to one or two forms.

The notion is akin to that of the transfer-appropriate processing approach in that performance on some test will benefit to the extent that procedures required during the test recapitulate those employed during encoding. Free recall experiments usually show evidence for a single store view of bilingual language representations. It is a 'conceptually driven' retrieval task. In free recall, individuals are presented no overt cues or hints to guide performance and thus must rely on stored concepts to facilitate remembering. On the other hand, the word fragment completion test, priming in perceptual identification, and lexical decision can be considered as tapping data-driven processing. Finding both language-independent and language-dependent patterns in an experiment under identical study conditions, but with different memory tasks, implies that the question of bilingual representation cannot be answered without a model of both storage and retrieval processes (Anderson 1978). Thus the issue of language dependent or independent representations is indeterminate until the tasks are specified for bilinguals' functioning.

An interesting suggestion was made by French and Ohnesorge (1995). In their model, each group of activated language units of either L1 or L2 will be differentially activated according to the amount of activation coming from the languages with which they are associated. If the L1 context is fully active and the L2 context is only weakly active then the L1 units will have higher overall activation than the weaker units of L2. The former will win the competition and determine

the activation of the entire representation. The advantage of this activation-based network model of bilingual memory is that it fits into an established framework of automatic spreading activation between verbal units, in which interlanguage lexical priming could be more easily explained. Just as activation in a monolingual L1 lexicon initially spreads automatically from a word to all of its related items in a context-independent manner, the activation-based description of bilingual lexical access could be used to explain the type of cross-lingual priming reported by some authors (see also de Groot and Comijs 1995).

Working memory (short-term memory) studies involving bilinguals are quite rare. There is, however, evidence that the phonological component of working memory, the phonological loop, plays a crucial role in supporting the long-term learning of the sound patterns of new words of both native and foreign languages (Baddeley et al. 1998). There is also evidence of strong lexical influence on working memory processing of new phonological structures via the phonological loop. Familiarity with L1 appears to have beneficial consequences for short-term storage of material conforming to the phonotactic and lexical properties of that language. The impact observed is on a lexical and a sublexical level (e.g., digit span). These effects are not found when L2 working memory processing was tested. Thus, the phonological loop is more effective in maintaining representations of words and nonwords from highly familiar languages than from languages that are less well known (e.g., Thorn and Gathercole 1999). In addition, Cheung et al. (2000) report support for Baddeley's model of working memory which specifies a phonological loop including an articulatory (rehearsal) control process and a nonarticulatory phonological store for acoustic memory traces left by verbal input. Differential effects of L1 and L2 on recall of pseudowords from L1 and L2 reveal that the language effect is phonological in nature, rather than having to do with the bilinguals' imbalanced proficiencies in the respective languages.

5. *Metacognition and Bilingualism*

Metalinguistic awareness, one type of metacognition, may be defined as an individual's ability to focus attention on language as an object in and of itself, to reflect upon language, and to evaluate it. Types of metalinguistic abilities are: phonological awareness, word awareness, syntactic awarenesss, and pragmatic awareness, or in other words analyses of knowledge and the control of cognitive operations involving language processing (Bialystock 1988). As language develops, the child structures and organizes an implicit body of language and gradually moves toward representations of knowledge that include explicit features for the structure of that knowledge. Bilingual children show accelerated metalinguistic (phonological) awareness in the preschool years when compared to monolingual children.

The development of linguistic awareness may help with learning a second language. As Thomas (1992) demonstrated, the formal training of a second language has a positive impact on the learning of a third language. This might be due to the fact that metalinguistic awareness determines the use of effective language learning strategies, such as relying on 'natural' informal situations for second language acquisition as opposed to formal instructions.

6. *Code-switching and Language Mixing*

It is critical to know in which speech mode a bilingual is in before making any claims about the individual's language processing or language competence. Interferences between languages during speech production may be a perfectly conscious borrowing or code switching in the bilingual speech mode. Rare are the bilingual corpora that clearly indicate the speech mode. In the monolingual speech mode, bilinguals adopt the language of the monolingual interlocutor. They deactivate more or less successfully the other language.

In the bilingual speech mode, both languages are activated. Bilinguals usually choose a base language to use with other bilingual interlocutors, but can within the same conversation switch to the other language. The switches take place at particular locations like at morpheme, word, phrase, or sentence(s) boundaries. Code-switching is an intentional process whereas language-mixing is the consequence of indiscriminative learning and the bilingual's lack of intra-language associations in one or both languages.

This survey of basic issues in bilingualism research reveals that bilingualism is an important competence in times of globalization and as such is worth being carefully analysed, but apart from this applied aspect it holds promises for the advancement of research and theorizing in some areas of cognitive science.

See also: Bilingual Education: International Perspectives; Bilingualism and Multilingualism; First Language Acquisition: Cross-linguistic; Language Acquisition; Language Development, Neural Basis of; Second Language Acquisition

Bibliography

Anderson J R 1978 *Language, Memory and Thought*. Erlbaum, Hillsdale, NJ

Baddeley A, Gathercole S E, Papago C 1998 The phonological loop as a language learning device. *Psychological Reviews* **105**: 158–73

Bahrick H 1984 Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General* **113**: 1–26
Bialystock E 1988 Levels of bilingualism and levels of linguistic awareness. *Developmental Psychology* **24**: 560–7
Cheung H, Lemper S, Leung E 2000 A phonological account for the cross-language variation in working memory processing. *Psychological Record* **50**: 375–86
Cummins J 1987 Bilingualism, language proficiency, and metalinguistic development. In: Homel P, Palij M, Aaronson D (eds.) *Childhood Bilingualism: Aspects of Linguistic, Cognitive, and Social Development*. Erlbaum, Hillsdale, NJ, pp. 57–73
Cutler A, Mehler J, Norris D, Segui J 1992 The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology* **24**: 381–410
de Groot A, Comijs H 1995 Translation recognition and translation production: Comparing a new and an old tool in the study of bilingualism. *Language Learning* **45**: 467–509
Dufour R, Kroll J 1995 Matching words to concepts in two languages: A test of the concept mediation model of bilingual representation. *Memory & Cognition* **23**: 166–80
Ellis N (ed.) 1994 *Implicit and Explicit Language Learning*. Academic Press, London
Fabbro F 1999 *The Neurolinguistics of Bilingualism. An Introduction*. Psychology Press, Hove, UK
French R M, Ohnesorge C 1995 Using non-cognate interlexical homographs to study bilingual memory organisation. In: Moore J D, Lehman J F (eds.) *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Erlbaum, Mahwah, NJ, pp. 31–36
Grosjean F 1992 Another view of bilingualism. In: Harris R J (ed.) *Cognitive Processing in Bilinguals*. North-Holland, Amsterdam, pp. 51–62
Harris R J (ed.) 1992 *Cognitive Processing in Bilinguals*. North-Holland, Amsterdam
Homel P, Palij M, Aaronson D 1987 (eds.) *Childhood Bilingualism: Aspects of Linguistic, Cognitive, and Social Development*. Erlbaum, Hillsdale, NJ
Jarvis J S, Newport E L 1989 Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* **21**: 60–99
Jarvis L H, Donks J H, Merrimen W E 1995 The effects of bilingualism on cognitive ability: A test of the level of bilingualism hypothesis. *Applied Psycholinguistics* **16**: 293–308
Paivio A 1986 *Mental Representations. A Dual Coding Approach*. Oxford University Press, Oxford, UK
Paradis M 1994 Neurolinguistic aspects of implicit and explicit memory implications for bilingualism and second language acquisition. In: Ellis N (ed.) *Implicit and Explicit Language Learning*. Academic Press, London, pp. 393–419
Schönpflug U 2000 Word fragment completions in the second (German) and third (English) language: A contribution to the organisation of the trilingual speaker's lexicon. In: Cenoz J, Jessner U (eds.) *English in Europe. The Acquisition of a Third Language*. Multilingual Matters Ltd, Clevedon, UK, pp. 121–42
Schönpflug U 2001 Zweisprachigkeit. Biologische und neurophysiologische Aspekte (Bilingualism. Biological and neurophysiological aspects). In: Götze L, Helbig G, Henrici G, Krumm H-J (eds.) *Handbuch Deutsch als Fremdsprache* (*Handbook of German as a Second Language*). DeGruyter, Berlin, Vol. 1
Sebastian-Galles N, Soto-Faraco S 1999 On-line processing of native and non-native phonemic contrasts in early bilinguals. *Cognition* **72**: 111–23
Thomas J 1992 Metalinguistic awareness in second- and third-language learning. In: Harris R J (ed.) *Cognitive Processing in Bilinguals*. North-Holland, Amsterdam, pp. 531–45
Thorn A S, Gathercole S E 1999 Language-specific knowledge and short-term memory in bilingual and non-bilingual children. *Quarterly Journal of Experimental Psychology* **52A**: 303–24

U. Schönpflug

Bill of Rights

1. *Overview*

Bills of rights in the sense of a listing of popular claims and admonitions to rulers and legislators are largely the legacy of late medieval and early modern English constitutional history. Their elevation into robust, court-enforced restrictions often trumping powers clearly assigned to political players is largely the product of twentieth-century America. Many subterranean steps both in the USA and abroad preceded its full-blown emergence in the 1930s. Current worldwide attraction to the new tradition of judicially supervised rights is a product of experience as to their universal appeal, their value in nation building and civic culture, and their efficient underwriting of a participatory political system. Still in contestation is the relationship of the older rights of political expression and protected procedure, and the claims of twentieth century ideologies of rights to economic achievement and/or economic equality.

2. *The English Antecedents*

Although largely a list of maddeningly narrow perquisites, circumscribed in time, space, and, social class, some of the provisions of Magna Carta have a universalistic King James Bible ring to them. Magna Carta has unmistakably been the great precursor of bills of rights both in popular image and legal reality. The other great English instruments—the Petition of Right (1628) and the document that came to be known as the Bill of Rights (1689)—are mainly allegations of kingly violations of Magna Carta, although they add some new claims as well. Besides being political rallying points these documents gave judges leverage in creatively interpreting and limiting statutes and royal actions.

Even before the conflicts that preceded the revolution American colonists were proud and jealous of their rights as Englishmen, as spelled out in the first

charter of Virginia (1609) and claimed in Maryland's Act for the Liberty of the People (1639). The prototype of American claims, however, was the Massachusetts Body of Liberties (1641), with a listing of specific rights culled from English documents and colonial experience. It contains a guarantee of free speech at town meetings and freedom of women from chastisement by husbands among its innovations. Other colonies enacted similar but varied listings. The colonists also effectively used such justifications as cause for dissatisfactions beginning with the Declaration of Rights and Grievances (1765) and culminating in the Declaration of Independence (1776). Among these was the Address to the Inhabitants of Quebec (1774), which vainly sought Canadian support by listing established freedoms in the colonies threatened by English authorities.

3. *The Road to the Bill of Rights*

In the postrevolutionary era, fully nine of the 13 states had bills of rights, usually in constitutional preambles. The others had some major liberties listed as did the Northwest Ordinance (1787). Only the Articles of Confederation avoided such enumerations.

The Philadelphia Convention's rejection of a Bill of Rights was therefore somewhat out of step with prevailing practice. The chief pro-Federalist defenses—that a Bill of Rights was unnecessary in a government of limited powers, and that enumeration of some rights might be seen as derogation of others—seemed lame since the proposed Constitution already included many traditional liberties, some even limiting state power.

As it became clear that the Constitution would be ratified, the anti-Federalists, prodded by Jefferson from his ambassador's post in Paris, accepted ratification in states conventions, but concentrated on proposed amendments. For their part the Federalists in the late-ratifying states worked mainly to limit amendments in the form of nonbinding recommendations to Congress. Six conventions called for amendments (seven if Rhode Island's belated ratification is counted) and in Pennsylvania a minority report did the same (Rutland 1991, Schwartz 1971).

The Bill of Rights has been aptly characterized as the last great compromise of the constitution, and it has served as the legacy of the states-rights anti-Federalists completing the constitutional structure designed by the nationalizing Federalists. The job of drafting fell largely to Madison, uniquely fit for the task. Then the unofficial floor leader of the Federalists, confidante of Jefferson and later an anti-Federalist president, major author of *The Federalist*, and key player at Philadelphia, he presented a set of proposals which deftly combined ratifying convention resolution, states' bills of rights, and creativity of his own. The House made major changes, deciding to affix the amendments as an appendix to the Constitution rather than intersperse them as in Madison's proposals. It also pruned and reworded them judiciously, adding some proposals generally rejected by the Senate. That later body reduced 17 proposals to 12, fusing some and rejecting others. It dropped Madison's pet proposal, which would have prohibited states as well as nation from infringing on rights of conscience and expression.

Ratification was swift but the first two proposed amendments were rejected. (In one of history's oddities the second was confirmed two hundred years later and is now the Twenty-second Amendment.) Despite much blather about the 'firstness of the First Amendment' it was as sent to the states as the third. Essentially Madison succeeded in satisfying most of the anti-Federalists, while only 'true-blue' Federalists dared sneer at the Bill of Rights, mainly in their private journals or correspondence.

4. *The American Application*

Although a showpiece in early US adulation of 'liberty', the Bill of Rights was barely evident as a legal barrier prior to the Civil War. The Marshall Court decisively slammed the door shut on any effort to use the amendments as checks on state government in Barron vs. Baltimore (1833). At the federal level it was congressional not judicial decision making that 'first crystalized the meaning of the rights of free speech and press' by returning fines, assessed under the Alien and Sedition laws. The due process clause of the Fifth Amendment is mentioned in cloudy passage in the Dred Scott case (1857), but it hinges on issues of citizenship and legal standing rather than liberty.

Not until the twentieth century era, in which the Supreme Court began to apply national standards to state regulation of such matters, criminal law, and the rights of freedom of expression, did the Supreme Court seriously implement the Bill of Rights against the national government. This was partly a product of the relatively limited reach of the national government during the nineteenth century, but it also reflected a shift in judicial sensibility in the twentieth. Nineteenth-century adjudication centered its allocation of power between nation and state. The concern for liberty dealt primarily with property and limits on regulation. The twentieth century has witnessed a constitutional sea change with the Bill of Rights providing the impetus. This has been a two-tier development. Under the rubric of 'liberty,' which states under the Fourteenth Amendment must provide, the Supreme Court had required states to meet the standards of most of the Bill of Rights (on some procedural matters, e.g., size of juries, the standard is less rigorous than that applied to the federal government). At the same time, the federal government has also been held to more rigorous

supervision under a concept (best expressed in footnote 5 of *The Carolene Products* decision of 1937) that courts have a special duty to protect those groups and individuals who are disadvantaged politically, especially where the Constitution confers specific rights.

5. *Experience Abroad*

Some liberal theorists have suggested that development from declaratory to enforceable rights has an imperative that unfolds under modern conditions of growing governmental ubiquity and the need to accommodate cultural diversity. So far only Canada has traversed the full path. Partly to placate growing French-Canadian demands for cultural recognition, Canada first promulgated a set of declaratory rights (1960) and later an explicitly justiciable Charter of Rights in its 'expatriated' constitution of 1992.

European developments also paralleled this but, as with much European Community law, the protection of rights is in fact more problematic and complex than it appears. Continental civil law countries in the nineteenth century generally followed the prototype of the French Declaration of Human Rights in their respective constitutions, but also followed the French in treating these rights as exhortatory rather than enforceable. Like the British, the continent rejected 'government by the judiciary.' True democratic rights were best protected—and least endangered—by popular sovereignty and the people's agent, the legislature.

The easy victory of Nazism and fascism in overnight subversion of rights alarmed and puzzled constitution makers after World War II. Several European countries installed constitutional courts—essentially halfway houses having special responsibility to maintain governmental boundaries. But in many instances only governmental agencies could bring cases. This and other restrictions have kept such courts significant but not encroaching on legislative power in important ways.

Another by-product of post-Hitlerian concerns was the promulgation of the European Convention on Human Rights in 1950 and later the European Court of Human Rights to enforce its provisions. The European Court of Justice, noting that all members of the community were signatories of the European Convention, assimilated it and its jurisprudence into European law. The action was presumably part of the court's conscious attempt at community building (giving individuals as well as businesses a stake in the system). But it also thereby undercut the German courts' resistance to the supremacy of European Community law on the asserted grounds that Germany had stronger protection of rights.

In effect the European Community has a Bill of Rights subscribed to by its members and defined mainly by a separate non-Community entity. In Great Britain, Parliament has already incorporated such rights into English law as of October 2000. This simplifies the task of the English judge, but also raises paradoxes that will be resolved only by future litigation or legislation.

While judicial protection would seem to armor-plate rights, libertarians even in the USA have suggested popular support for liberty may be diminished by reliance on judicial enforcement. Critics of the USA are as impressed with failures at crucial points over two centuries—Dred Scott, Red Scares—as by positive achievements such as desegregation. And certainly normative lists of rights have had great impact in Britain, France, and other European countries.

6. *What do Bills of Rights Protect?*

Many bills of rights are cynically promulgated or fall into obsolescence, often in countries where military juntas take over. The most cynical of all were probably the Soviet and other East European efforts, which were often painstakingly crafted by experts, incorporating the most forward-looking thinking, without any intent of implementation, but to stake a claim of progressive leadership to impress external adherents. Even within the Eastern bloc, however, dissidents carefully memorized provisions, acting as if they expected their rights to be respected.

The contents of bills of rights vary, but they generally accentuate rights of expression (including religion) and procedural protection as to criminal proceedings or governmental taking of property. Equal treatment under the law and gender equality are also usual, as are guarantees of rights to participate politically. Some bills of rights guarantee cultural autonomy or preservation of families (Duchacek 1973).

These so-called 'negative freedoms' have for over a century been seen as inadequate by liberal intellectuals and parties who call for guaranteed 'positive' freedoms—economic opportunity, access to medical services, and employment. A right to universal free education is perhaps the least controversial. Not found in the US Constitution, it is standard in most countries and many American state Constitutions. More controversial are such 'rights' as full employment or guaranteed family income. Communist and Socialist societies have contended economic rights are precursors and prerequisites of procedural and participatory rights, which are in a sense luxuries resulting from economic satisfaction (MacPherson 1966). Liberal theorists argue that history has it right: participatory rights and restraints on governmental naked power unleash marked creativity which permits extension of economic rights.

The fall of 'the evil empire' in Eastern Europe, paralleled by governmental deficits in welfare states in

the West has kept those stressing economic equality rights on the defensive. Even communist China has announced that 'iron rice bowl' guarantees are gone and workers must not only work but scurry to gain employment. At the same time protection against arbitrary confinement and rights of expression has been demonstrated to have wider appeal than some have thought, although perhaps still valued most by the middle classes.

See also: Common Law; Constitutional Courts; Constitutionalism; Human Rights, Anthropology of; Human Rights, History of; Human Rights: Political Aspects; Justice and Law; Justice: Philosophical Aspects; Justice: Political; Rights; Rights: Legal Aspects

Bibliography

Amar A R 1998 *The Bill of Rights*. Yale University Press, New Haven, CT

Brant I 1967 *The Bill of Rights*. New Home Library, New York

Duchacek I D 1973 *Rights & Liberties in the World Today*. ABC-CLIO, Santa Barbara, CA

Dumbauld E 1957 *The Bill of Rights and What it Means Today*. University of Oklahoma Press, Norman, OK

Hoffman R, Albert P J (eds.) 1997 *The Bill of Rights*. University Press of Virginia, Charlottesville, VA

MacPherson C B 1966 *The Real World of Democracy*. Clarendon Press, Oxford, UK

Rutland R A 1991 *The Birth of the Bill of Rights*. North Eastern University Press, Boston

Schneider H W 1938 The philosophical difference between the Constitution and the Bill of Rights. In: Reed C (ed.) *The Constitution Reconsidered*. Columbia University Press, New York

Schwartz B (ed.) 1971 *A Documentary History*. Chelsea House, New York

S. Krislov

Binding Problem, Neural Basis of

For a large part of the twentieth century, the issue of how physical brain states represent mental objects was dominated by one idea, that of single units as elementary symbols of our mind. According to this idea, an individual's mental state is fully described by the ensemble of single units active in a given moment. This symbol system would be most peculiar in having only one composition rule, that of simultaneous activity, whereas all other known symbol systems have flexible means of building hierarchical structures. The mind, meeting place and source of all other symbol systems, surely must possess a mechanism for combining its semantic atoms into molecules and aggregates in a way richer than anything yet contrived on paper or in electronics. Identifying this mechanism is the binding problem.

1. The Binding Problem

One of the solid results of research on brain function is the localization of mental themes in the brain in a hierarchical manner down to the assignment of definite semantic meaning to individual neurons. It is, therefore, a broadly accepted view that neurons can be treated as elementary symbols with fairly fixed meaning. The assignment of meaning to neurons has been a very successful enterprise for decades, and it is probably a permissible extrapolation to apply this picture to every neuron in our cerebrum, if not in our whole nervous system. It has, furthermore, long been an uncontested view that the state of a neuron can be characterized by an activity variable, which in a not very precisely definable way corresponds to its current rate of firing (one speaks of 'rate coding') . Semantic meaning is routinely assigned to neurons by temporally associating their rate of firing with sensory stimuli or motor responses.

A neuron is thus taken as an elementary symbol that is either active or dormant in a given mental state. It is elementary in the sense of having no internal structure in terms of subsymbols (although a neuron's semantic referent invariably is a composite entity). As a rule there seem to be groups of equisemantic neurons. To allow for this possibility, the term 'single unit' will be used instead of 'single neuron.' In spite of all the internal anatomical and physiological structure of single units, the mental symbols associated with them are taken to be elementary, i.e., as having no internal degrees of freedom.

At issue here are the laws by which higher symbols are composed from single units. With rate coding there is only one composition rule: mere additive, unstructured combination of the active units' elementary meanings into one amorphous lump. This is a very poor composition rule, and a somewhat inadequate basis for cognitive and mental operations. A few examples may illustrate this inadequacy. Assume a first mental object (e.g., an imagined hand) to be represented by the set A of single units, and a second mental object (e.g., an apple) by set B (for simplicity let A and B have no units in common). Now assume there is reason to hold both mental objects simultaneously (as, for instance, when the hand is grasping the apple). The above composition rule states that the superset $C = A \cup B$ will be active. This, however, creates a problem in that C doesn't contain any information as to grouping of its elements in terms of its constituents A and B, and there now may be several ways to decompose C into part symbols. This ambiguity has

been termed the 'superposition catastrophe.' In a more concrete example, let a person perceive a blue letter A and a yellow letter B on a sheet of paper, and assume there are units to represent 'A,' 'B,' 'blue' and 'yellow' in the person's brain. The simultaneous activity of all four units represents the situation incompletely, as a yellow A and a blue B would evoke the same situation. This type of 'conjunction errors' are actually committed by human subjects when not given enough time to inspect a situation (for review see Wolfe and Cave 1999). The experimental examples show that the binding of elements into groups can be a real problem for our brain if it is not given sufficient time.

If single units are the mind's 'bricks,' what is the 'mortar' to erect its cognitive edifices—to group features into objects (such as in figure–ground separation), to attach attributes to referents (such as grammatical roles to the words of a sentence), to represent temporal or spatial arrangements of elements such as in a spoken sentence or in a visual scene, or to point out correspondences between structures found to be analogous to each other?

2. *Binding by More Single Units*

Single units can code for (fixed) combinations of other units, and in fact this is true for almost all units. In the above example, the colored letter ambiguity would be dispelled immediately if there were single units encoding blue As and yellow Bs. So why shouldn't the cognitive architecture of the whole brain be based just on combination-coding units? A correct short answer to this question probably is that indeed it is to a very large degree, but that for the purposes of a small but vital subset of operations it cannot. The reason is that whatever a brain's complement of combination-coding units, it is bound to run continuously into situations which call for new combinations and which would only be represented with dangerous ambiguities. Each individual ambiguity could be stopped by more combination-coding units, but those units happen not to be present (yet). Thus, vital flexibility in handling new situations will need a binding mechanism that transcends a code based entirely on single units.

Several serious obstacles stand in the way of achieving anything like completeness in terms of single-unit coding. The number of composite symbols that are required over a lifetime is too large to be covered ahead of time in any combinatorially complete way. To take just one example, there are an infinite variety of feature combinations through which physical objects can manifest themselves in our visual system, and which must be handled as composite mental entities. It is impossible to represent them one by one in terms of single units. Each one of them is new and unforeseeable, and a single unit to represent it would have to be manufactured on the spot. All that can be hoped for is that certain subpatterns are common to many situations, get represented in single units, and collectively reduce by large factors the combinatorial ambiguity the brain has to cope with, by representing a given mental object in terms of relatively few units.

Representing composite entities (like 'my grandmother') by single units is problematic in itself. It is as if a mathematician wanted to replace all composite expressions by single-letter symbols. We can operate with composite objects (especially when they appear for the first time) only with reference to their parts (and their relations) with which we have had previous experience. Consequently, those parts (and relations) have to be explicitly represented themselves! If, for instance, all colored letters were, for the sake of avoiding ambiguity, represented by colored-letter coding cells (blue-A etc.), anything that I have learned in the past about green As would not be available for blue As and I would have to learn it anew. If, on the other hand, all part-representing units were visible along with the units representing the whole (blue, yellow, A, B along with blue-A and yellow-B), the potential confusion would be even bigger than without the combination-coding units. Each client (that is, group of units) would have to learn to pay exclusive attention to the appropriate symbol level, by disconnecting from units on all levels that would cause confusion. For this and other reasons, installing a new unit is a complicated business that is warranted only for representing frequently recurring symbols. Finally, a brain based entirely on single units could not represent explicitly the hierarchical structure of complex symbols and would consequently not be able to discover structural relations between mental entities.

In summary, a nervous system based entirely on a flat symbol system composed of single units would be totally inflexible and uncreative and would not be able to deal with novel situations.

3. *Temporal Binding*

According to this idea sets of units express mutual binding by synchronizing their signals. In the above example, the units 'blue' and 'A' would express their binding by firing in synchrony, or the two sets of units *A* and *B* would avoid the superposition catastrophe by firing with strong positive signal correlations within each set and zero or negative correlation between them. This idea as such does not go beyond the generally accepted fundamental principle, stated above, that the composition rule for unit symbols is simultaneous activity; it extends it, however, down to the millisecond time scale, and it requires the processing of whole signal sequences if complex binding structures are to be expressed. Temporal signal structure is induced by the presence and structure of connections between units, with strong excitatory links generating positive correlations. Temporal signal

structure is interpreted by structured circuits in that the correlation structure of a set of signals may or may not fit the internal connectivity of a circuit upon which it impinges (two units with mutual excitation are easily excited by correlated signals, for instance, but would respond less readily if they inhibited each other).

Random connectivity structures would neither produce clear-cut signal structures nor could they selectively respond to structured signals. Appropriate connectivity patterns can be formed by network self-organization on the basis of (possibly rapid and reversible) synaptic plasticity: an initially random circuit would create correlation patterns, which would act back on the circuit by strengthening or weakening connections in a feed-forward fashion (strong correlations leading to strong connections, for instance), rapidly converging to an attractor state in which signal correlation and circuit structure are optimally reinforcing each other.

The binding problem and its solution by signal synchrony was first discussed as a fundamental issue of brain function by Von der Malsburg (1981) (although the idea of synchrony coding has been briefly mentioned in several earlier references), and the same reference has proposed network self-organization as its basis. Temporal binding has been applied successfully to a range of problems of brain function, as reviewed in Von der Malsburg 1999. Among these are logical reasoning, figure–ground segregation in the visual, auditory and olfactory modalities, and invariant object recognition. Experimental evidence for temporal signal structure relevant for binding and its occurrence under appropriate circumstances is reviewed in Gray 1999 and in Singer 1999. Main objections to the temporal binding hypothesis, reviewed in Shadlen and Movshon 1999, concern still insufficient experimental evidence and doubts about the ability of cortical tissue to process temporal signal structure on an interestingly fast timescale.

A strong limitation on temporal binding is the low bandwidth of neural signals. According to optimistic estimates, the temporal resolution with which correlations are evaluated in the cerebral cortex is one millisecond. This leaves little space for many time slices to be kept separate in typical processing times of 100 milliseconds. Are there other, more efficient means of binding? One possibility that will have to be explored in the future is based on multicellular units: the neurons in a unit may all code for the same elementary symbol but they may differ in their connectivity to other units. By proper control of a unit's internal activity distribution it may be made to dynamically change its connection pattern and thus express selective binding to other units. This will require highly specific connectivity patterns. Before these are installed, temporal binding and rapid reversible synaptic plasticity will have to act as the 'fire brigade' suppressing binding ambiguities as they emerge.

Once the neural community will start to build up momentum and methodology to fully utilize the strengths of a neural-based symbol system with full binding capability, the remaining deep riddles of our cognitive apparatus may be ready for assault.

See also: Neural Synchrony as a Binding Mechanism; Object Recognition: Theories; Perceptual Organization; Visual Perception, Neural Basis of

Bibliography

Gray C M 1999 The temporal correlation hypothesis of visual feature integration: Still alive and well. *Neuron* **24**(1): 31–47

Shadlen M N, Movshon J A 1999 Synchrony unbound: A critical evaluation of the temporal binding hypothesis. *Neuron* **24**(1): 67–77

Singer W 1999 Neuronal synchrony: A versatile code for the definition of relations? *Neuron* **24**(1): 49–65

Von der Malsburg C 1981 *The Correlation Theory of Brain Function*. MPI Biophysical Chemistry, Internal Report 81–2 (Reprinted in: *Models of Neural Networks II*. In: Domany E, van Hemmen J L, Schulten K (eds.) 1994 Springer, Berlin, Chap. 2, pp. 95–119)

Von der Malsburg C 1999 The what and why of binding: The modeler's perspective. *Neuron* **24**(1): 95–104

Wolfe J M, Cave K R 1999 The psychophysical evidence for a binding problem in human vision. *Neuron* **24**(1): 11–17

C. von der Malsburg

Binet, Alfred (1857–1911)

1. A Brief Biography

Alfred Binet was born in the French town of Nice on July 11, 1857. According to his daughter Madeleine (Avanzini 1974), Alfred was a bright child who succeeded so well in school that his mother decided to send him away to the capital when he was barely twelve years old, in order that he might study at one of the best schools in France. Upon graduation from high school, he completed a law degree in 1877, but then decided to pursue studies in medicine and biology. Under the supervision of Balbiani, who was later to become his father-in-law, he started writing a dissertation on 'The sub-intestinal nervous system of insects.' At the same time he wrote plays for the theater. He finally managed to combine his different areas of interest in one discipline, psychology. To simplify, one can distinguish three phases in his career as a psychologist, although his publications prove that the stages were overlapping rather than strictly separate (Delay 1958): psychopathology, experimental psychology, and child psychology.

French psychology in the second half of the nineteenth century was focused mainly on psychopath-

ology, and Binet's early interests were no exception. During this first stage of his career, the main influences on Binet were Théodule-Armand Ribot and Jean-Martin Charcot, and Binet's first major writings explored the domains of sensations, hysteria, hypnotism, and personality disorders. Binet worked closely with Ribot between 1880 and 1907, and published over 40 articles in Ribots's *Revue Philosophique*. During this period, he also began to study hypnosis under the famous neurologist Charcot at the Salpetrière hospital in Paris.

In 1891 Binet received tenure at the Physiological Psychology Laboratory at the Sorbonne University in Paris, where his research focus switched from clinical to experimental psychology. The main contemporary influences in psychophysiology and experimental psychology were German, namely Wilhelm Wundt, whose laboratory in Leipzig served as a model for the one Binet created in Paris. Binet completed his thesis in biology in 1895, the same year in which he became the head of the Physiological Psychology Laboratory and published an *Introduction to Experimental Psychology*. It is also the year in which he founded a journal that is still published today, *L'Année Psychologique*. In 1899 Binet met Théodore Simon, who was to become his closest and most constant collaborator, and that same year he joined the *Free Society for Psychological Child Studies*, of which he became the President in 1902. Like many researchers of his time, Binet decided to follow in the footsteps of neurologist Paul Broca. Binet explored intelligence through phrenology, believing that there is a direct relation between the size of the cranium and a person's level of intelligence. Binet's empirical investigations measured the crania of school children considered by their teachers to be bright and compared these measures with the measures obtained from low-achieving children. In 1901 Binet conceded that 'the idea of measuring intelligence by measuring the head … seemed ridiculous' (Binet 1901, p. 403), but he then discovered that there were noticeable differences in the cranial size and shape of children from the contrasted groups. Very bright students had larger skulls than very low-achieving students. Binet himself, in the same 1901 volume, reached the conclusion that the initial measures showing a cranium size difference between low- and high-achieving students were probably due in part to his own suggestibility. With a rare sense of honesty, he admitted that, having looked for differences, he found them where he expected them. His faith in phrenology shaken, he abandoned it. He also discarded hypnosis, and moved to explore the human psyche through new tests and puzzles that he initially developed for, and then tested on, his two daughters (Binet 1890). As was the case with Piaget several years later, most of his ideas on cognitive development sprang from the observation of his own children.

After a first period devoted to psychopathology, and a second centered on experimental psychology, one can distinguish a third period in Binet's life, dedicated primarily to child psychology. The start of this period can be dated to 1905, the year in which Binet and Simon published several articles on the diagnosis of abnormality (Binet and Simon 1905a, 1905b, 1905c) and the year in which Binet opened a laboratory of experimental pedagogy. In this laboratory, Binet worked closely with teachers. It was during this period that Binet made his major contributions to the study of academic underachievement, but also to psychometrics, a term coined by Sir Francis Galton in 1879. Another major contribution of Binet to child psychology was the view that children are not, as generally thought previously, miniature adults, but obey different rules of cognitive functioning (Zazzo 1958). The rest of this essay will focus on Alfred Binet's contributions to the understanding of abnormal children and the development of the cognitive-assessment tools to which this interest led. This historical account will end with a brief discussion of the current state of the field.

On October 28, 1911, Alfred Binet passed away unexpectedly, at the age of 54, and was laid to rest in the Montparnasse cemetery in Paris.

2. *The Study of Abnormal Children*

Starting with the beginning of the nineteenth century, Binet showed great interest in the study of 'abnormal children' (Binet and Simon 1907). His *Free Society for the Psychological Study of Children* pushed the Government to start evaluating children, with the aim of early identification of abnormal children and the provision to them of special education. In 1904, Binet was officially appointed by the French Minister of Public Education to devise a means of identifying retarded school children. It was based on this interest in abnormal children that the Metric Intelligence Scale was developed in 1905 (Avanzini 1974), with the help of Binet's doctoral student Théodore Simon. Through the use of this scale, children with learning impairments, who were nevertheless thought to be able to profit from education, were selected and put into special-education classes. The first 'perfectioning class' was opened in 1907, and two more such classes followed that year. Although Binet himself was fully convinced that abnormality could to some extent be overcome, Binet always stressed the importance of empirically assessing the correctness of his ideas. And although Binet worked primarily on applied problems, he never forgot the importance of strict experimental methodology. He thus clearly stated the need experimentally to assess the usefulness of the special-education classes before any legislative measures for their institutionalization would be taken.

Before further describing the Metric Intelligence Scale and its most widespread adaptation, known as the Stanford–Binet Scale of Intelligence, let us first go

back to Binet's theory of abnormal children and the special-education classes that led Binet to believe in the need for psychometric assessment.

In their 1907 publication *Abnormal Children*, Binet and Simon proposed a new view of mental retardation. The abnormal child is not a normal child that has either stopped or slowed down in development, but rather, a child with a different developmental pattern. This pattern is unbalanced, with some aspects identical to those of normal children, and others, different. It is this lack of balance between the different developmental aspects that constitute the abnormality (Binet and Simon 1907, 1922). Although there might be an important delay in the acquisition of certain aptitudes, the development of other aspects (cognitive and somatic) might be normal.

Binet and Simon argued that the implications and impact of an acquisition are linked to the age at which it is acquired. For example, learning to read at six does not have the same implications as learning to read at twelve. The abnormal child who learns to read at twelve has a wider vocabulary and more extra-scholastic interests than does the normal six-year-old child. Intellectual impairment can be characterized by three major components: a global developmental delay, the imbalance of this delay depending on the aspects measured, and the resulting lack of co-ordination in the functioning of the mind. Because abnormality is a distinctive developmental pattern, not a mere retardation, special education is possible and needed.

3. *The Metric Intelligence Scale*

The Intelligence Scale devised by Binet and Simon was published in 1905. It contained 30 tests, some of which were expressly created for the scale, and some of which were adaptations of already existing cognitive tests, namely, those developed by the French physicists Blin and Damaye. All tests had been piloted on samples of 'normal' and 'retarded' samples of children of ages two through twelve. The different tests measured everything from such basic skills as co-ordination of movement and imitation to such complex processes as comprehension, judgment, and abstract reasoning. The tests were presented in order of increasing difficulty.

The easiest tests distinguished severely and profoundly mentally retarded children (then labeled as 'idiots') from the rest; tests of intermediate difficulty distinguished between the severely retarded and moderately retarded (then labeled as 'imbeciles'); and higher-level tests between mildly retarded (then labeled as 'morons') and normal children of the same age. The tests also distinguished, of course, between younger and older children. As a member of the ministerial committee on education, Binet's main purpose was to identify the mildly retarded children in order to provide them with the special education, which was thought to increase their cognitive functioning.

In the first publication of the Assessment Scale, the authors emphasized that the Scale should be used only as an indication of the child's cognitive level at the time of administration, and that this performance level is subject to change through an appropriate education. A revised version of the Scale was published in 1908, offering the possibility of establishing a child's *mental age*, a notion first introduced by Chaillé in 1887. If 65–75 percent of the children from a given age group succeeded on a test, the test was classified as being appropriate for and measuring mental performance at that age level. A given child can thus succeed on tests that are at a mental age level that is lower, equal to, or higher than that targeted at her or his biological age. Note that the mental age simply corresponds to the performance norm of a given age group. It does not, as was thought by the American adapters of the Scale, provide an indication of the stage in developmental progression reached by the child. Having a given mental age does not imply functioning as a child of that age. It implies only that an individual's performance on a given test corresponds to what the majority of children at a given age will achieve.

Binet explicitly warned against the use of performance on an IQ test as a fixed measure of intelligence. Indeed, by the nature of the test construction, 25–35 percent of children of a given age will not succeed on a test intended for their age (Lippman 1922).

Binet made it clear that the scale was just one assessment tool, and that the observation of the 'global person' and the test taker's reaction to the situation, as well as his or her behavior, was as important as the test results *per se*. More than the end results, it is the paths by which they are reached that interested Binet.

3.1 *Misinterpretations and Misuses of the Metric Intelligence Scale*

To paraphrase the title of an article by Sarason (1976), the fate of Alfred Binet and school psychology was unfortunate. As described above, Binet's aim was to develop an instrument that would contribute to the identification of those children who needed special help in order to make the most of these children's potential. His aim was to prevent further rejection of these children by supplying them with the special education that would eventually enable them to return to the regular school system. But this second aspect of Binet's program was all too often forgotten and his scale used to weed out rather than to select children for special programs.

It is important to remember that Binet did not view intelligence as a given quantity fixed at birth, but rather, as incremental. It is precisely because of the elastic nature of intelligence and the possibility of

developing it through education that children with specific needs needed to be identified and given the special education that would help them develop their intellectual abilities. But the researchers who imported his instrument to the United States and Great Britain had different views. Lewis Terman, who in 1916 introduced a US version of the Metric Intelligence Scale known as the Stanford–Binet, stressed the importance of hereditary factors in explaining performance on intelligence tests. Terman considered this genetic influence too strong to be altered through education, and advocated placing children with low performance scores in special classes, not in order that they get the special attention they needed to progress, but because there was no hope of their being able to integrate into and profit from normal schooling. Terman 1916) believed that the State would have to assume guardianship for those with intelligence levels in the mentally retarded range.

4. *Strengths and Weaknesses of Binet's Contributions*

Binet's strengths and weaknesses are very much the two sides of the same coin: his strength is in the applied nature and the large social-cultural impact of his work, his weakness, in the lack of a complete theory. Binet's research was particularly important for its pragmatic approach and its attempt directly to connect to social and educational concerns. His research on abnormal children and on cognitive assessment started not from a theoretical construct, but rather in response to a concrete problem.

Binet is perhaps to be remembered for two major theoretical contributions. First, he broke with the Galtonian tradition of studying intelligence through psychophysical processing. He suggested instead that intelligence be studied through higher order processing in complex everyday tasks. This viewpoint, evident as it may seem today, was novel, and almost provocative, to many of his contemporaries. Before Binet, most psychologists, following Wundt's example, thought that intelligence was best explored by studying lower order, very simple, processes. Binet not only decided to explore higher order cognitive functions, but to do so with the same experimental rigor that had previously been applied to simpler processes. Second, he introduced a new methodology for studying intelligence.

It is difficult to find what could be called a true theory of intelligence in Binet's work (Reuchlin 1977). Nevertheless, Binet did think in theoretical terms. In his publications, intelligence is sometimes assimilated to judgment, common sense, practical sense, initiative, or faculty to adapt. At other times, intelligence is decomposed into four 'functions': comprehension (understanding), invention, direction, and self-censorship. But neither Binet's heart nor his head was in theorizing about intelligence. He even went so far as to state that he was only interested in understanding the specific facts that he had gathered and not in any general theory of intelligence.

5. *Psychological Assessment Today*

The importance of psychometrically sound assessment scales to measure level of cognitive development was soon generalized beyond the population of abnormal children and constitutes one of Binet's greatest legacies to the field of cognitive evaluation. The high positive correlation between children's scores on the Binet–Simon Metric Intelligence Scale and their school performance, largely due to the very scholastic nature of the test items, was seen as a proof that the scale was indeed measuring intelligence, and contributed to its widespread success. The Stanford–Binet scale is still widely used for testing of intelligence, although its prominence has been somewhat eclipsed by the success of the Wechsler scales.

See also: Cattell, Raymond Bernard (1905–98); Cognitive Development in Childhood and Adolescence; Galton, Sir Francis (1822–1911); Hebb, Donald Olding (1904–85); Immigrants: Economic Performance; Intelligence: Historical and Conceptual Perspectives; Intelligence: History of the Concept; Piaget, Jean (1896–1980); Psychology: Historical and Cultural Perspectives

Bibliography

Avanzini G 1974 *A Binet: Écrits psychologiques et pédagogiques.* Privat, Toulouse, France, pp. 6–176

Binet A 1890 Recherches sur les mouvements de quelques jeunes enfants. *Revue Philosophique* **29**: 297–309

Binet A 1901 Recherches de céphalométrie sur 26 enfants d'élite et enfants arrières des écoles primaries de Seine-et-Marne. *L'Année Psychologique* **7**: 403–11

Binet A, Simon T 1905a Sur la nécessité d'établir un diagnostic scientifique des états inférieurs de l'intelligence. *L'Année Psychologique* **11**: 163–90

Binet A, Simon T 1905b Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique* **11**: 191–244

Binet A, Simon T 1905c Application des methods nouvelles au diagnostic du niveau intellectuel des anormaux d'hospice et d'école primaire. *L'Année Psychologique* **11**: 245–336

Binet A, Simon T 1907 *Les enfants anormaux*. A. Colin, Paris

Binet A, Simon T 1922 *La mesure du développement de l'intelligence chez les jeunes enfants*. Société pour l'Étude Psychologique de l'Enfant, Paris

Delay J 1958 La vie et l'oeuvre d'Alfred Binet. *Psychologie Française* **3**: 85–95

Lippman W 1922 The mental age of Americans. The New Republic, October 25, 1922. In: Jacoby R, Glauberman N (eds.) 1995 *The Bell Curve Debate: History, Documents, Opinions.* 1st edn. Times Books, New York, pp. 561–65

Reuchlin M 1977 *Psychologie*. Presses Universitaires de France, Paris
Sarason S B 1976 The unfortunate fate of Alfred Binet and school psychology. *Teachers College Record* **77**: 579–92
Terman L M 1916 *The Measurement of Intelligence*. Houghton Mifflin, Boston
Zazzo R 1958 Alfred Binet et la psychologie de l'enfant. *Psychologie Française* **3**: 113–21

R. J. Sternberg and L. Jarvin

Binocular Space Perception Models

Human binocular vision gives rise to the perception of a unitary and stable three-dimensional space of compelling phenomenological reality, but in many situations perception is not veridical. Binocular space perception models aim at predicting perceived spatial relations—such as perceived position, shape, or size—under conditions excluding monocular cues and experiential factors, so that only binocular cues are effective. Binocular cues are the vergence position of the eyes and binocular disparity, which is the difference of the retinal projections in the two eyes caused by their lateral separation.

In the traditional experimental setup observers were presented with isolated point-like light sources of low intensity in complete darkness, with the position of the head kept fixed. The stimuli were confined to the horizontal plane at eye level, taking for granted that disparities in the vertical direction do not contribute to perceiving depth. Sect. 2 illustrates the diversity of psychophysical models that have evolved within this tradition.

With the use of random dot stereograms, which were developed in 1960, it has been shown that disparities on their own are sufficient to produce three-dimensional vision. This impression of depth emerging from binocular disparity, which is known as *stereopsis*, has been the focus of much recent research on binocular vision. Howard and Rogers (1995) provide a comprehensive state-of-the-art report, including a treatment of the optical and physiological basis of stereopsis, and an excellent survey of the vast experimental literature (see also Regan 1991). Sect. 3 highlights some theoretical conceptions relevant to modeling stereoscopic surface perception, and Sect. 4 indicates how this approach may be integrated into a psychophysical framework.

1. Distal and Proximal Stimuli

There are various coordinate systems that not only allow the specification of the position of distal stimuli in three-dimensional physical space but also lead to a characterization of proximal stimuli, which are the retinal images (see Howard and Rogers 1995). In a longitudinal-azimuth/latitudinal-elevation system the position of each distal stimulus is determined by angles of azimuth α_h, β_h and elevation α_v, β_v, describing horizontal and vertical directions relative to the rotation centers of the right and the left eye, respectively (see Fig. 1). Idealizing the relevant optics by identifying the center of rotation and the optical node with the center of curvature of a spherical eye, these coordinates directly specify the proximal stimuli.

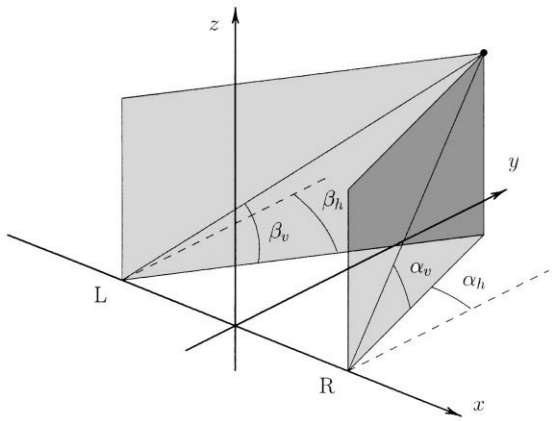

Figure 1
The coordinates α_h, β_h denote the monocular azimuth of a point with respect to the rotation centers of right and the left eye and relative to the dashed lines parallel to the y-axis. The coordinates α_v, β_v denote its elevation relative to the place $z = 0$

In the horizontal plane $z = 0$ at eye level the locus of stimuli that are projected onto corresponding retinal points (points that are congruent when the two retinas are superimposed) is a circle through the fixation point and the rotation centers of the two eyes (see Fig. 2). This so-called Vieth–Müller circle is the trajectory of constant binocular parallax γ, which is defined as the difference of the monocular azimuth angles. Accordingly, the locus of symmetric retinal points, which deviate to the same extent but in opposite directions from the foveæ, is known as the hyperbola of Hillebrand (see Fig. 2). This hyperbola can be characterized as the trajectory of constant binocular azimuth φ, which is the average of the monocular azimuth angles. Thus, we have

$$\gamma = \alpha_h - \beta_h \quad \text{and} \quad \varphi = \frac{\alpha_h + \beta_h}{2} \tag{1}$$

With these definitions the relative horizontal disparity between two stimuli may be identified with the difference of their respective binocular parallaxes.

In stereopsis the stimuli, like stereoscopic surfaces, are characterized by a two-dimensional disparity vector field that is to be interpreted as the trans-

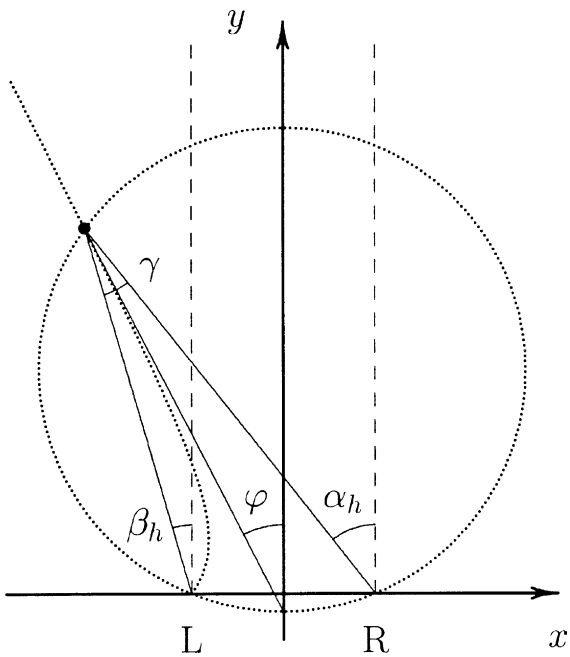

Figure 2
In the horizontal plane $z = 0$ at eye level the binocular parallax γ is the angle subtended by the visual axes when the eyes coverage on a point, while the binocular azimuth φ describes its eccentricity. The dotted curves are the Vieth–Müller circle and the hyperbola of Hillebrand incident with the point

formation of the respective retinal images from one eye into the other. The horizontal and vertical component of a disparity vector are given by the differences of $\alpha_h - \beta_h$ and $\alpha_v - \beta_v$ to the respective values of the fixation point.

2. *Psychophysical Models*

To allow for a compact presentation of their basic rationale, the discussion of the models is confined to perceived egocentric distance, and to binocular visual direction, both in the horizontal plane at eye level.

A highly sophisticated approach is due to Rudolf K. Luneburg (1903–49), a physicist, who is given credit for achieving the final mathematical identification of optics with electromagnetics. After becoming acquainted with the striking empirical phenomena in binocular space perception known at that time Luneburg (1947) developed a theory in which depth perception is related to a non-Euclidean geometrical structure of binocular visual space through certain psychophysical assumptions. In particular, he assumed that the points lying on a Vieth–Müller circle are perceived as being equidistant from the observer, and that the hyperbolas of Hillebrand are perceived as radial lines of constant direction. The concrete form of the psychophysical function of egocentric distance $\rho_{K,\sigma}$ is given by the metric radial distance

$$\rho_{K,\sigma}(\gamma) = \tan_K^{-1}(2e^{-\sigma\gamma})$$

where

$$\tan_K^{-1}(r) = \begin{cases} \dfrac{2}{\sqrt{K}} \tan^{-1}\left(\dfrac{\sqrt{K}}{2} r\right) & \text{for } K > 0 \\ r & \text{for } K = 0 \\ \dfrac{2}{\sqrt{-K}} \tanh^{-1}\left(\dfrac{\sqrt{-K}}{2} r\right) & \text{for } K < 0 \text{ and } |r| < 2/\sqrt{-K} \end{cases}.$$

Egocentric distance depends on the geometry through the Gaussian curvature K taken to be constant, and on an observer-specific parameter σ indicating depth sensitivity. Veridical perception is assumed for binocular visual direction, i.e., the respective psychophysical function ϕ takes on the form $\phi(\varphi) = \varphi$.

Numerous experimental studies have revealed that, although the Luneburg theory captures many of the qualitative properties of the data, it is less successful at the quantitative level. Moreover, there is ample evidence that the loci of perceived equidistance exhibit systematic deviations from the Vieth–Müller circles, which, in terms of binocular parallax differences, do not remain constant but tend to decrease in size with increasing distance. This finding led to the conclusion that relative egocentric distance depends not only on disparities, but also on binocular parallax.

Foley (1980) obtained a psychophysical function expressing this dependence by utilizing a heuristic principle sometimes called 'vision is inverse optics.' A formal characterization of how the percept emerges from the proximal image is derived from an optical equation relating the proximal image to the distal stimulus simply by replacing the physical magnitudes by corresponding perceptual quantities. For the physical radial distance r from the origin the approximate (physical) relationship $r \approx (i/\gamma) \cos \varphi$ holds, where γ denotes the binocular parallax, φ the binocular azimuth, and i the interocular distance. By substituting the perceived egocentric distance ρ for r, the 'effective binocular parallax' γ_e for γ, and by identifying the binocular visual direction ϕ with its physical counterpart φ as in the Luneburg theory, Foley obtained the equation

$$\rho = \frac{i}{\gamma_e} \cos\varphi$$

He argued that the egocentric distance ρ to a reference point in the stimulus configuration is misperceived,

while differences of binocular parallaxes are veridically perceived. The linear functional dependence $\gamma_e = p + q\gamma$, with parameter values restricted to move in the ranges $0 < p < 2°$ and $0 < q < 1$, is interpreted to reflect a tendency to overestimate short distances and to underestimate large distances. Consequently, the parallax differences of the locus of perceived equidistance to the corresponding Vieth–Müller circle are a linear function of its parameter γ. Although Foley's theory closely predicts experimental data, its derivation deserves some criticism from the viewpoint of representational measurement theory. No justification is offered for replacing the relevant physical quantities by corresponding perceptual quantities and for treating them as extensive magnitudes (see *Memory Models: Quantitative*). This critique applies to a number of approaches (e.g., Wagner 1985) that convert physical equations into psychological theories by employing the same heuristic. Moreover, meaningfulness problems arise whenever physical and perceptual quantities are compared with each other, which is done by stating that distances are over- or underestimated.

Heller (1997) provides a measurement-theoretic approach to account for the empirically observed locus of perceived equidistance. Binary relations $\preccurlyeq_\rho$ and $\preccurlyeq_\phi$ are introduced to describe the perceived ordering with respect to egocentric distance, ρ, from near to far, and direction, ϕ, from left to right. Within the Luneburg theory these orderings are directly induced by the physical orderings with respect to the binocular parallax $\gamma = \alpha_h - \beta_h$, and the binocular azimuth φ (or, equivalently the sum $\alpha_h + \beta_h$). This motivates the assumption that there exist strictly increasing functions f and g, such that for all stimuli (α_h, β_h), (α_h', β_h')

$$(\alpha_h, \beta_h) \preccurlyeq_\rho (\alpha_h', \beta_h') \text{ iff } f(\alpha_h) - g(\beta_h) \geqslant f(\alpha_h') - g(\beta_h')$$
$$(\alpha_h, \beta_h) \preccurlyeq_\phi (\alpha_h', \beta_h') \text{ iff } f(\alpha_h) + g(\beta_h) \geqslant f(\alpha_h') + g(\beta_h'). \quad (2)$$

The functions f, g can be thought of as implementing the optics of a more realistic eye model, but may additionally specify the contribution of nonoptical causes. A set of sufficient conditions to prove the existence of such a representation is provided within a conjoint measurement approach (see *Measurement Theory: Conjoint*). There is empirical evidence speaking to the validity of the basic axioms, and the empirically observed locus of perceived equidistance is closely predicted with an appropriate choice of functions. With

$$\Gamma = f(\alpha_h) - g(\beta_h) \quad \text{and} \quad \Phi = \frac{f(\alpha_h) + g(\beta_h)}{2} \quad (3)$$

we obtain a psychologically significant recoordinatization of physical space that generalizes Eqn. (1), where the coordinates Γ and Φ characterize the loci of perceived egocentric equidistance and constant binocular direction, respectively. Substituting Γ, Φ for γ, φ at each of their occurrences in the Luneburg theory leads to a generalization of that model.

3. *Computational Approach*

There is currently no comprehensive model of binocular space perception based on stereopsis. On the one hand, it is still necessary to experimentally identify those attributes of the two-dimensional disparity vector field which are relevant to the perception of depth, and to determine their impact. On the other hand, there are a number of criteria indicating depth, with perceived slant and inclination of stereoscopic surfaces having received most attention. Much of the modeling is from 'computational vision', which, however, is concerned primarily with the information that in principle can be extracted from a given disparity vector field irrespective of its relevance for human binocular vision. In search for invariants that allow for characterizing slant, inclination, eccentricity, or distance of a stereoscopic surface, among other quantities, the spatial derivatives of the disparity vector field are considered. The decomposition of the first-order derivative into three components (rotation, uniform expansion or contraction, and an area-preserving linear combination of expansion and contraction in orthogonal directions), the invariances of which can easily be interpreted in optical terms, is a celebrated result by Koenderink and van Doorn (1976).

The following example demonstrates that the computational approach may actually be successful in identifying cues that the visual system could possibly exploit. Mayhew and Longuet-Higgins (1982) showed that the vertical disparities of a few nonmeridional points provide information about the distance and direction of the fixation point, which was thought to be available only from oculomotor cues. Rogers and Bradshaw (1995) argue that, at least in the case of large-field stimuli, this visual information is actually used in binocular space perception. They found that vertical disparities together with vergence contribute to a veridical perception of flat and fronto-parallel surfaces.

4. *Future Research*

It is left as an open problem to provide a comprehensive model of binocular perception of complex stimuli in three-dimensional space. Progress may be expected from integrating the computational results into a psychophysical theory. Showing that the identified optical invariances are relevant to binocular vision amounts to nothing else but showing that they actually constitute psychophysical invariances, which

are transformations of the stimuli leaving perception invariant. Within a representational framework as in Eqn. (2), psychophysical invariances induce functional equations that restrict the possible form of the psychophysical function (see Aczél et al. 1999).

See also: Fechnerian Psychophysics; Functional Equations in Behavioral and Social Sciences; Neural Plasticity in Visual Cortex; Psychophysical Theory and Laws, History of; Psychophysics; Vision, High-level Theory of; Vision, Low-level Theory of; Vision, Psychology of; Visual Perception, Neural Basis of; Visual Space, Geometry of; Visual System in the Brain

Bibliography

Aczél J, Boros Z, Heller J, Ng C T 1999 Functional equations in binocular space perception. *Journal of Mathematical Psychology* **43**: 71–101

Foley J M 1980 Binocular distance perception. *Psychological Review* **87**: 411–34

Heller J 1997 On the psychophysics of binocular space perception. *Journal of Mathematical Psychology* **41**: 29–43

Howard I P, Rogers B J 1995 *Binocular Vision and Stereopsis*. Oxford University Press, New York

Koenderink J J, van Doorn A J 1976 Geometry of binocular vision and a model for stereopsis. *Biological Cybernetics* **21**: 29–35

Luneburg R K 1947 *Mathematical Analysis of Binocular Vision*. Princeton University Press, Princeton, NJ

Mayhew J E W, Longuet-Higgins H C 1982 A computational model of binocular depth perception. *Nature* **297**: 376–8

Regan D 1991 *Binocular Vision*. Macmillan, London

Rogers B J, Bradshaw M F 1995 Disparity scaling and the perception of frontoparallel surfaces. *Perception* **24**: 155–79

Wagner M 1985 The metric of visual space. *Perception and Psychophysics* **38**: 483–95

J. Heller

Bioarchaeology

Bioarchaeology is the study of human remains from archaeological settings, viewing these remains in relation to biocultural adaptation, life history, lifeway reconstruction, and population and demographic history (after Buikstra 1977). Human remains can include any body tissue, ranging from artificially or naturally preserved bodies—such as mummies from Egypt or bog bodies from Denmark—to fragments of bones and teeth. Because skeletons are preserved far more often than other types of remains, they comprise most of the subject matter of bioarchaeology (Larsen 1997).

1. Origins and Paradigm Shifts: from Racial Classification to Biological Process

Bioarchaeology originated in the field of osteology, the anatomical study of bones and teeth. Much of the history of study of ancient skeletons focused on classification and racial typology. Biological differences between populations from different time periods—such as differences in skull form—were interpreted to represent replacement of one 'race' of people by another. Various lines of evidence reveal the adaptability of bone tissue, especially with regard to influence of physical activity on shape and structure. For example, skulls representing earlier populations who lived exclusively on wild plants and animals tend to be larger and more robust than skulls representing later farmers (e.g., Carlson and Van Gerven 1977). In consideration of the extrinsic factors that influence skull form (e.g., shift from chewing hard- to soft-textured foods), the temporal variation for many settings is best interpreted in light of underlying biological processes that influence musculoskeletal development and not racial history.

The reinterpretation of temporal changes in skull form is emblematic of the paradigmatic shift that took place in bioarchaeology in the last 30 years of the twentieth century (Larsen 1997). Namely, ancient bones and teeth are no longer viewed as objects for classification only. Rather, bioarchaeologists seek to understand ancient skeletons as though the populations they represent are alive today, subject to the myriad of environmental, cultural, and social influences that affect human biology.

Skeletal and dental tissues are remarkably sensitive to a variety of environmental conditions affecting their form and composition. Owing to this sensitivity during life, human remains recovered from archaeological settings are informative about quality of life (health) and life-style (activity and work) in particular, and behavior in general. Skeletons also contain a fund of information about biological relatedness and demographic history.

2. Quality of Life and Health

Quality of life refers to a variety of factors that influence physical well-being—health is the most important. Health is assessed from human remains via chemical, stress, pathological, and morphological indicators.

2.1 Reconstructing Diet and Inferring Nutrition: Stable Isotopes and Bone Chemistry

The food humans eat and the nutrition that it provides is fundamental to health. Until the 1980s, most

knowledge about ancient diets was based on plant and animal remains from archaeological contexts (see *Paleodemography: Demographic Aspects*; *Zooarchaeology*). In the late 1970s, Vogel and Van der Merwe (1977) identified the timing and intensity of the shift from diets based on wild plants to domesticated plants (maize) by analyzing stable isotope ratios of carbon ($^{12}C/^{13}C$) extracted from archaeological human bone samples. This path-breaking study provided a new basis for reconstructing past diet and set the stage for analysis of stable isotopes of other elements (e.g., nitrogen, strontium, oxygen) and drawing inferences about variation in diet, food use, weaning practices, and climate.

2.1.1 Stress and deprivation. The cells responsible for tooth and bone development are easily disrupted if physiological stress caused by malnutrition or disease occurs while these tissues are forming. In teeth, macroscopically visible areas of enamel deficiency (hypoplasia) or microscopically visible defects (accentuated Retzius lines) indicate growth disruption. In bones, growth arrest (Harris) lines are seen in x-ray images of the ends of leg and arm bones. In addition to having elevated frequencies of indicators of growth arrest, populations that experience chronic illness or malnutrition tend to be short in stature and have less bone mass in comparison with healthy, well-nourished populations.

Many human populations experience iron deficiency, caused by iron-poor diet, parasitic infection, or poor living circumstances involving poor sanitation. Anemia often develops, resulting in increased production of a person's red blood cells. The skull bones of anemic persons are often thickened and display porosity (porotic hyperostosis). The incidence of growth arrest indicators, anemia, and other related morbid conditions increased in the last 10,000 years as human populations shifted to agriculture and became sedentary.

2.1.2 Illness and infection. Throughout the entire five million or so years of human evolution, human beings have been exposed to many different microbes, many of which cause disease. Skeletons record the effects of a number of chronic diseases, such as the treponematoses (the disease group that includes the nonvenereal and venereal forms of syphilis), leprosy, and tuberculosis. Identification of deoxyribonucleic acid (DNA) in archaeological skeletons has verified that the lesions found in ancient remains result from the disease-causing microbes (e.g., *Mycobacterium tuberculosis*; cf. Salo et al. 1994).

Leg (tibia) and, to a lesser extent, other bones in archaeological skeletal samples sometimes display periosteal reactions, areas of pathological bone apposition. These lesions are nonspecific because they can be caused by different infectious agents or by trauma to the lower leg. For many settings, these lesions are likely due to population crowding and poor hygiene, conditions conducive to the maintenance and spread of infectious disease. Sedentary, crowded populations in prehistory have generally higher frequencies of nonspecific and specific infectious disease than mobile, dispersed populations.

2.2 Violence and Injury

All human populations experience physical confrontation of one type or another at some point in time. In the Santa Barbara Channel Island region of the southern California Pacific coast, numerous skulls display indentations (depressed fractures) caused by being struck on the head with a club (Lambert and Walker 1991). In late prehistory in this setting, the frequency of cranial fractures increased, and lethal projectile wounds became more common than before. This temporal shift coincides with deterioration in climate and reduced availability of food resources, suggesting that as environmental stress increased, people became more competitive and more violent. The skeletal evidence derived from the victims of violence may be more informative about the causes and consequences of confrontation than other archaeological evidence (e.g., fortifications, iconography, and weaponry).

3. Lifestyle and Activity

Physical activity is a defining characteristic of different human adaptive regimes. Hunter-gatherers are often characterized as highly mobile, living demanding lifestyles that require a great deal of work in order to acquire food. Agriculturalists, on the other hand, are often thought of as being more 'advanced,' leading lives of leisure. Study of ancient skeletons allows us to address these assumptions and to test hypotheses about labor, workload, and physical activity in past societies.

3.1 Activity and Degenerative Pathology

The articular surfaces of the joints of the skeleton are adapted to withstand mechanical stress, such as those that derive from lifting or walking, or a combination of activities. Over the course of a person's lifetime, the cartilage and bone in the articular joints begins to erode in response to mechanical loading (e.g., knees). In addition, spicules of bone form along the joint margins, which may impede movement. These bone changes are part of a disorder called osteoarthritis. In archaeological settings, there is a tendency for skeletons of prehistoric hunter-gatherers to have more osteoarthritis than those of prehistoric farmers, and males generally have more osteoarthritis than females (see Bridges 1992, Larsen 1995). However, the pattern

and frequency of osteoarthritis is highly variable from one region to another, suggesting that the disorder is influenced by local circumstances involving a complex interplay between culture and lifestyle.

3.2 Activity and Bone Structure

Bone tissue remodels itself in response to mechanical stimulation (Wolff 1892/1986). In areas of the skeleton (or a particular bone) that are subjected to high levels of physical demand, more bone tissue is added in order to add strength and to resist forces that would cause fracture. Breadths and circumferences (e.g., for leg and arm bones) can be used to characterize the size of bones, and to infer levels and types of physical activity. Engineers have developed a more precise way of inferring levels or types of physical activity. Civil and mechanical engineers measure the strength of beams that are used in the construction of buildings. The distribution of material in cross-sections of these beams is analyzed in relation to how they are able to resist bending, twisting (torsion), and other loading modes. Similarly, the human skeleton provides the body with a superstructure for the support of body weight and the action of muscles. From cross-sectional measurements of bones (derived from CAT scans or by direct cutting to reveal the cross-sections), bioarchaeologists apply these engineering principles and methods to determine the strength of bones (Ruff 1992). Biomechanical analysis of skeletons from around the world shows a general reduction in bone strength over the course of human evolution, especially in the last 10,000 years, suggesting that physical activity has declined.

3.3 Activity Above the Neck

The size and morphology of the face and jaws is influenced by mechanical loading and activities involving mastication of food, and in the use of teeth in nondietary roles (e.g., use of front teeth for preparing animal hides for clothing). Human faces and jaws have become smaller in recent human evolution, reflecting the increased tendency for eating softer, more processed foods. Consumption of softer foods and more carbohydrates has also contributed to increase in dental caries (cavities) in recent humans. Patterns of tooth wear are informative about how people use their teeth. Scanning electron microscopy of the chewing surfaces of teeth reveals microwear patterns relating to diet and tooth use (Teaford 1991).

4. Estimating Biological Relatedness: Bone Shapes, Tooth Traits, and Ancient DNA

The shapes and structures of bones and teeth contain key information about population history, since these shapes and structures are at least partly genetically determined. There are some 30 different dental (discrete morphological) traits that are especially useful for identifying inter- and intrapopulation biological relatedness (Scott and Turner 1997).

Because bone shape is influenced by environmental factors (see above), and many dental traits are obliterated by wear on chewing surfaces of teeth, identification of biological relatedness is sometimes problematic. A potentially more informative approach to reconstructing biological history came about in the mid-1980s with the discovery that DNA is often preserved in ancient bones. The development of polymerase chain reaction (PCR) made it possible to amplify short segments of DNA into millions of copies that could be used to identify key parts of the genome (O'Rourke et al. 1996). The analysis of DNA has allowed the testing of hypotheses about biological history, such as long-standing issues relating to population origins and continental settlement (e.g., the origin of native Americans and migration to the Americas), in ways not possible by study of morphology and traits in bones and teeth alone.

5. Population Profiles and Demographic History

Physical anthropologists have developed various methods for identifying sex and estimating age at death from skeletons (Ubelaker 1989). Identification of sex (a biological attribute) allows insight into gender (a social/cultural attribute) and its relationship to health, activity, and social and cultural behavior (Grauer and Stuart-Macadam 1998). Mortality and fertility history of a past population can be reconstructed from age composition of collections of skeletons from archaeological cemeteries. In a stable population with minimal in- or out-migration, age composition is especially informative about birth rates of past populations. A skeletal series having many infants and young children and relatively fewer old adults suggests relatively high birth rate. Conversely, a series containing few children and many older adults suggests low birth rate (e.g., Buikstra et al. 1986). In some regions of the world, the increase in population size that accompanied the shift from foraging to farming appears to be related to decreased birth spacing and increased fertility.

6. Challenges

Research results derived from the study of ancient skeletons—once mostly relegated to the appendices of obscure, unpublished archaeological reports—are now viewed as integral to archaeology. Reflecting this increased role is the growing visibility of bioarchaeology in the popular press and in the scientific literature. Past human biology provides an essential context for understanding present-day diseases, health, and life-style. As the discussion of cultural 'repatriation' continues to build in the USA, Canada,

Australia, Israel, and elsewhere, the archaeological community is increasingly called upon to explain to the larger public why ancient remains are important. Public education about bioarchaeology is important and should be pursued if an informed understanding of the past using the human biological component of the archaeological record is to progress.

See also: Plagues and Diseases in History; Race: History of the Concept

Bibliography

Bridges P S 1992 Prehistoric arthritis in the Americas. *Annual Review of Anthropology* **21**: 67–91

Buikstra J E 1977 Biocultural dimensions of archeological study: A regional perspective. In: Blakely R L (ed.) *Biocultural Adaptation in Prehistoric America*. University of Georgia Press, Athens, GA, pp. 67–84

Buikstra J E, Konigsberg, L W, Bullington J 1986 Fertility and the development of agriculture in the prehistoric Midwest. *American Antiquity* **51**: 528–46

Carlson D S, Van Gerven D P 1977 Masticatory function and post-Pleistocene evolution in Nubia. *American Journal of Physical Anthropology* **46**: 495–506

Grauer A L, Stuart-Macadam P 1998 *Sex and Gender in Paleopathological Perspective*. Cambridge University Press, Cambridge, UK

Lambert P M, Walker P L 1991 Physical anthropological evidence for the evolution of social complexity in coastal southern California. *Antiquity* **65**: 963–73

Larsen C S 1995 Biological changes in human populations with agriculture. *Annual Review of Anthropology* **24**: 185–213

Larsen C S 1997 *Bioarchaeology: Interpreting Behavior from the Human Skeleton*. Cambridge University Press, Cambridge, UK

O'Rourke D H, Carlyle S W, Parr R L 1996 Ancient DNA: A review of methods, progress, and perspectives. *American Journal of Human Biology* **8**: 557–71

Ruff C B 1992 Biomechanical analysis of archaeological human skeletal samples. In: Saunders S R, Katzenberg M A (eds.) *Skeletal Biology of Past Peoples: Research Methods*. Wiley-Liss, New York, pp. 37–58

Salo W L, Aufderheide, A C, Buikstra J, Holcomb T A 1994 Identification of *Mycobacterium tuberculosis* DNA in a pre-Columbian Peruvian mummy. *Proceedings of the National Academy of Sciences* **91**: 2091–4

Scott G R, Turner C G II 1997 *The Anthropology of Modern Human Teeth: Dental Morphology and Its Variation in Recent Human Populations*. Cambridge University Press, Cambridge, UK

Teaford M F 1991 Dental microwear: What can it tell us about diet and dental function? In: Kelley M A, Larsen C S (eds.) *Advances in Dental Anthropology*. Wiley-Liss, New York, pp. 341–56

Ubelaker D H 1989 *Human Skeletal Remains: Excavation, Analysis, Interpretation*, 2nd edn. Taraxacum, Washington, DC

Vogel J C, Van der Merwe N J 1977 Isotopic evidence for early maize cultivation in New York State. *American Antiquity* **42**: 238–42

Wolff J 1892/1986 *The Law of Bone Remodelling* [trans. Maquet P, Furlong R]. Springer-Verlag, Berlin

C. S. Larsen

Bioethics: Examples from the Life Sciences

Bioethics is a relatively new field of learning, drawing on many established academic disciplines, such as philosophy, jurisprudence, psychology, sociology, and others. Though medical ethics is a central part of bioethics, the latter has a broader scope: in addition to the main theme of classical medical ethics—doctor–patient relationship and doctor–doctor relationship—bioethics is concerned with genuine problems in connection with philosophical ethics as well as practical questions arising from the medical and non-medical life-sciences and affecting public policy and the direction and control of science.

1. The History of Bioethics

The main stimuli for the increase of research in bioethics around the 1960s arose from the technical progress in medicine, societal changes in the western world, and a revitalized interest in the discipline of philosophical ethics. Around that time, progress in medical technology—e.g., kidney dialysis, artificial respirators, organ transplantation, medically safe abortions, and the contraceptive pill—radically changed the practice of medicine. In addition, the first steps towards a powerful gene-technology were made. Roughly at the same time, profound changes occurred in western societies. The civil rights movements in the USA and elsewhere, as well as the transformation of traditional institutions such as the family and the church, caused moral questions concerning an individual's rights and duties in modern society.

Finally, the philosophical discipline of ethics received a fresh impetus. For a long time there had been consensus amongst most philosophers that judgments about moral problems could not be rational in principle. In the second half of the twentieth century, much work has been done, however, which successfully aimed at providing scientific methods for moral reasoning (Arrington 1997, Stingl 1997). In combination, these developments formed the basis of the extraordinary rise of bioethics in the second half of the twentieth century. In the meanwhile, bioethics has become institutionalized in academic departments, professional journals, and international associations.

2. The Scope of Bioethics

Though medical ethics forms the core of bioethics, the latter has a broader scope and is understood to encompass questions arising in the context of the life-sciences, i.e., medicine and the biological disciplines, as well as in some fields of the environmental and social sciences.

The scope of bioethics can be divided roughly into three fields of research. First, there is the project of a rational justification of moral norms. It forms the methodological core of bioethics, since it supplies the strategies of moral reasoning that are applied to the respective problems at issue.

Second, there is what might be called clinical ethics; this group encompasses moral problems that individuals, e.g., nurses, physicians, or researchers, are confronted with in their daily practice. Examples of this kind involve decisions about the end of life, e.g., decisions regarding abortion or medically-assisted suicide.

Finally, there is a group of problems that should be dealt with on the regulatory or policy level. Examples of problems of this kind involve the necessity to develop operationalizable juridical guidelines for awarding patents on biological materials, policies for the allocation of medical resources or the protection of biodiversity.

3. The Method of Bioethics

3.1 Morality and Philosophical Ethics

The task of professionalized ethics is not to be seen in gaining insights from some special source of knowledge—insights that would entitle the ethicist to demand a certain course of action or insist on certain conduct being prohibited. Rather the proper task of ethics is the critical assessment of existing moral convictions and norm systems established from them. This assessment should be done by using a method that is in principle accessible to everyone. Moral convictions are subject to such an assessment when they generate conflicts, e.g., when they are incompatible with other moral convictions. In the case of moral conflicts there is a need for an ethical criterion, that allows us to establish which conviction should be followed. A simple example for such a criterion is the 'Golden Rule': 'We should do unto others as we would wish them to do unto us.' This rule does not provide us with a material action-guide but provides a formal criterion that any action guide has to fulfil in order to be morally acceptable.

Traditionally it has been the domain of philosophical ethics to design strategies for the solution of moral conflicts (Gethmann 1989). In this sense bioethics is only a new area because the conflicts to be solved were unknown until progress in medical technology and societal changes directed public attention to the ethical problems of the life sciences.

It should be the role of a philosopher wishing to contribute to the solution of bioethical issues to provide a clarification of the moral concepts involved and to reconstruct the arguments used in the discussion. On this basis he can develop strategies of moral reasoning that help determine which moral convictions one ought to accept.

It should be made clear, however, that these strategies are developed on the basis of an underlying ethical theory and that a variety of different ethical theories has been developed during the history of ethics (for an introduction, see Frankena 1973): consequentialism claims that, if anything determines the moral status of an action, then it is the consequences of that action, whereas deontological theories hold that there are at least some actions that are morally advisable whatever their consequences. In revitalizing ancient approaches, virtue ethics aims at developing action guides that tell agents what they ought to do in concrete situations.

Principle-based approaches hold that some general moral norms are central to moral reasoning; contrary to this, case-oriented approaches advocate starting with the examination of concrete cases and proceeding from that to more general norms. An example of the latter is the ethics of care: it aims at an expressly practice-oriented approach to bioethics by engaging emotionally in the moral problems at issue. The above list of ethical theories is neither complete nor does it present a proper classification of such theories, but it names those theories most relevant in the recent debate on bioethics (see e.g., Beauchamp and Childress 1994).

Such theories often share certain presuppositions, e.g., the demand for 'universalizability' and 'consistency,' but differ in others, so that—depending on the theory applied—different strategies of moral reasoning can be developed in the discussion of a certain moral problem at issue.

It is often used as an argument against the use of moral reasoning for practical questions that there are no indisputable ultimate arguments for moral norms achievable on a rationally understandable basis. Though true, this should lead to modesty, not resignation, concerning the power of moral argumentation, since we should not demand from the field of ethics what is not available in other areas too: for example, in applying the laws, a judge is always left with a degree of discretion within which he has to examine carefully the special circumstances of the case at issue.

The ability to analyze the moral concepts involved and to apply one or the other ethical theory to bioethical problems will help the physician, policy-maker, or whoever else is confronted with these questions to think about them in a sophisticated way and to develop moral convictions acceptable to all parties concerned.

3.2 Bioethics and the Natural and Social Sciences

The methodology supplied by philosophical ethics is only one, albeit essential, component in the attempt to solve bioethical problems. The philosopher is trained in examining the arguments proposed, whether they are valid relative to some standards of rationality or whether they fall back upon metaphysical or emotion-based beliefs. Substantial arguments about bioethical problems should be based, however, on knowledge from those sciences in the domain in which those problems originated. A discussion of the ethical problems of genetic counseling, for example, without the participation of physicians and others involved in this counseling would be futile. Similarly, it would be pointless to establish the impact that the results of a new technique, such as the genetic manipulation of food, will have on society without the participation of those social sciences that have the necessary empirical methods at their disposal. Finally, it would be vain to develop recommendations for societal regulations, e.g., economically effective and juridically sound allocation procedures for medical resources, without the backing of health economics and jurisprudence. Bioethics is, therefore, an auxiliary discipline supporting the discussions of other disciplines.

3.3 The International and Intercultural Aspects of Bioethics

The experience of a conflict between existing moral convictions in view of the consequences of new techniques is not confined to any cultural or geographical frontier: it is a universal experience. The actual object of the conflict, however, may be highly dependent on, for example, the culture in which the conflict arises.

This context-dependency of bioethical problems is often taken as the prime argument against there being a single method of bioethics, since solutions to moral problems are—so it is claimed—valid only relative to the particular situation for which they were developed. But although it is often the case that material moral convictions do differ between cultures, this does not in itself exclude the possibility of there being a formal criterion acceptable in both cultures—in the same way that the differing views of plaintiff and defendant do not in themselves exclude the possibility of an acceptable judgment.

4. Major Issues in Contemporary Bioethics

The great number and variety of issues discussed in contemporary bioethics compels us to make a rigorous selection. The following section throws a spotlight on some topics that were intensely discussed in the recent debate and that furthermore aptly illustrate an interdisciplinary approach to bioethics (see Chadwick 1998, Kuhse and Singer 1999, Reich 1995).

4.1 Medical Ethics

Since the time of Hippocrates and the ancient Greek medico-philosophers, the emphasis of medical ethics has been on the doctor-patient relationship. Since the 1960s, this has been debated mainly in terms of the improvement of patient autonomy in view of an allegedly prevailing paternalism on the physician's side (Faden and Beauchamp 1986). Today the respect for autonomy is widely accepted as an action guide of overriding importance. Topics still controversially debated are moral problems arising at the beginning of life, the end of life, and the allocation of resources in and to the health care system.

4.1.1 Beginning of life. The moral acceptability of abortion as well as the fate of handicapped newborns is one of the most extensively and intensely discussed issues in bioethics. Basic to this discussion is what quality an entity must possess in order to be a moral subject endowed with certain rights, e.g., a right to life, and how we know when an entity actually possesses this quality.

For some authors, this quality is simply that of belonging to the species *homo sapiens*—a position called speciesism and criticized because affiliation to a certain biological class is as morally irrelevant as the affiliation to a certain race, nation, or social class.

According to another position, this quality is that of being a person—where being a person is understood to involve some kind of self-consciousness. A third position claims that being an agent, i.e., causing one's own actions, makes an entity a moral subject. Characteristic for the latter two positions is that according to them there is the possibility that some members of the species *homo sapiens* are not persons or agents (e.g., a human foetus) while some persons or agents are not members of species *homo sapiens* (e.g., gorillas or dolphins). Once what counts as a moral subject has been defined, it still has to be clarified what rights this moral subject is endowed with. It is, for example, not plausible to claim that a foetus is endowed with an absolute right to life, when even adults are usually not awarded such a far-reaching right: e.g., in the event of war and in self-defense (Tooley 1972).

Over and above this, some criterion has to be found which is applicable in practice and which allows us to determine whether an entity belongs to the class of moral subjects or not. Here an interdisciplinary approach is necessary that combines the above moral considerations concerning the quality characterizing a moral subject with state-of-the-art knowledge from the natural sciences about when an entity actually possesses this quality.

4.1.2 End of life. It has long been (and still is) a rule of most professional codes in medicine to forbid active euthanasia whereas passive euthanasia is some-

times allowed. This distinction between a morally reprehensible action (active euthanasia) and a morally acceptable action (passive euthanasia) is based on an alleged moral difference between 'killing' (act) and 'letting die' (omission). There have been numerous attempts to point out the morally relevant difference between these—without much success so far (Steinbock and Norcross 1994). Though active euthanasia seems acceptable from the point of view of ethics, the satisfactory implementation of codes of conduct concerning active euthanasia in medical practice remains difficult. The reason for this might be that both performing (or even only advocating) active euthanasia can not only jeopardize the patient's trust in his doctor, but moreover generate a threat to legal integrity.

Another controversially debated topic in bioethics is the brain-death criterion: modern medical techniques allow us to sustain the blood circulation of patients, who could not breath spontaneously due to the loss of certain basal brain-functions. According to the traditional definition of death—the standstill of the cardiovascular system—these patients are not dead (though they depend on the continuous support of a respirator). But most people's intuitions are at odds with a patient's body not showing any behavior expected from a living human being. Since the traditional death criterion was no longer useful, it was necessary to adopt a new one (Singer 1994). The Ad Hoc Committee of Harvard Medical School to Examine the Definition of Brain Death in 1968 developed the brain-death criterion that has since developed into the standard death criterion. It is important, however, to notice that the brain-death criterion is not the only possible definition of death—indeed the traditional criterion of complete cessation of blood circulation is another one. In a certain sense, death is not naturally given but defined and much of the controversy about bioethical issues at the end of life arises from neglecting this basic insight.

4.1.3 Allocation of resources. Whenever limited resources are allocated (rationing), refraining from taking desirable actions of one kind becomes the price of performing actions of another kind. Health care is no exception to this universal truth. Fixing a budget in the health care sector means that an increase in spending for, say, preventive medicine has to be accompanied necessarily by decreasing spending in another sector, say, transplantation centers or, on another level, higher spending in a state's education or military sector necessitates lower spending in other sectors, e.g., the health system. The lack of even a basic public health care system makes allocation of resources to the health care system an important topic in many developing countries. But increasing costs of health care have brought the topic of allocation of resources in the health care system onto the general agenda even in those countries that are equipped with a public health care system. The moral questions arising from this are among other things: is there a right to health care? If yes, what has to be provided by the state: a decent minimum or optimal provision? Who shall allocate the scarce resources and by what criteria? Is there a moral obligation to redistribute health care resources on a global level?

4.2 Ethics of Genome Research and Manipulation

Human genome research—best known via the International Human Genome Project, but actually consisting of a multitude of different research programs—has experienced an extraordinary boom during the last decades. Though promising unprecedented options for the prevention, diagnosis, and therapy of diseases, this research bears considerable risks. In this context it is useful to differentiate medico-technical problems of human genome research (e.g., the safety of gene therapy protocols) from moral problems (e.g., the moral acceptability of intervening into the human genome). Only the latter will be considered in the following.

4.2.1 Genetic diagnostics. Genetic tests performed before and after birth as well as the genetic screening of populations are debated controversially because of the potential misuse of the information acquired. Even if the test is used only for medical purposes, the results might lead to stigmatization and discrimination of individuals or ethnic groups. Though not based on state-of-the-art knowledge from the biomedical sciences, this geneticization of medicine should be countered by very careful implementation of genetic diagnostics in the medical practice, including a stronger than usual emphasis on the protection of the privacy and autonomy of the tested (Chadwick et al. 1999).

Another possible domain of use of genetic diagnostics is the testing of individuals applying for insurance coverage (Sorell 1998): it is feared that individuals with an increased genetic risk might have to pay higher premiums or, worse, might be unable to obtain insurance coverage at all. From an ethical point of view, the main question to be discussed is whether there is a right to insurance coverage. A right to the provision of elementary health insurance can be legitimated by an argument concerning the necessary conditions required to allow the full exercise of an individual's civil rights. On the other hand, it seems implausible to postulate a right to life insurance—as long as a life insurance is understood as covering needs beyond the basic needs.

Furthermore, genetic tests might be used in pre-employment (employment selection) tests. Though beneficial in some respects, e.g., saving allergic people

from exposure to allergens, this application may end in severe discrimination of prospective employees: not only that employment could be denied in the light of scientifically unsound tests such as genetic markers for intelligence or criminal behavior, moreover, acquiring a DNA sample of a person makes it possible to perform a whole range of genetic tests that fall outside the employer's legitimate interest in hiring healthy persons.

4.2.2 Gene-therapy. It is hoped that the incorporation of human or nonhuman genes in body cells (somatic gene-therapy) or the human germline (germline gene-therapy) will be a major contribution to the therapy of diseases with a genetic component—hereditary as well as non-hereditary. Somatic gene therapy is seen as a relatively unproblematic method from the ethical (not medico-technical) point of view, since the transferred genes are not inherited by the descendent of the treated individual.

In stark contrast to this, germline gene-therapy is controversially debated. Since the transferred genes to germline cells are inherited by all offspring, the question arises whether such an intervention in the genetic constitution of an unborn individual can be legitimated and, on a more general level, whether germline therapy leads to a new form of eugenics or genetic enhancement. Since most authors hold that if germline gene-therapy were to be applied, it could not be exclusively confined in the long run to medical application, the debate therefore centers on the question whether the intervention into the lives of future persons by genetic enhancement is different in any morally relevant way from classic interventions such as education (Agar 1998)

4.2.3 Cloning. The first cloning—i.e., asexual reproduction—of a mammal in 1996 raised moral question concerning the application of this technique to non-human animals and humans (Harris 1997, Wilmut 1998). Animal cloning is debated mainly in view of the legitimation of manipulating the genome of wild-type animals and the effect of a change in the genetic pool of the biosphere that might have considerable effects on biodiversity.

The debate on human cloning has focused on the ethical problems of producing copies of existing human beings (reproductive cloning). Concerning reproductive cloning, the emphasis of the moral debate lies on whether there are morally acceptable aims for the actual application of this technique, whether the genetic identity of a clone and its donor amounts to a morally problematic personal identity, and whether human cloning leads to an infringement of the dignity of the clone or the donor.

The cloning only of parts of humans (cells or organs) is technically still in its infancy but has caused an intense debate concerning the moral status of the embryonic cells that are necessarily consumed in the embryo experimentation needed for a further development of this technique (see Lauritzen 2001).

4.2.4 Genetically modified organisms. An area of application for the results of genome research is the manipulation of organisms. Well established is the introduction of genes in bacteria, plants or animals. The transgenic organisms produced employing this method are used either to perform further research on the transferred gene or with the purpose of producing certain rare substances in a bio-factory, e.g., the generation of insulin in transgenic bacteria. Additionally, it is hoped that transgenic organisms will be helpful in improving food-supply in developing countries.

The moral problems of genetically modified organisms arise from possibly deliberate interventions in the 'natural' biodiversity as well as from alleged infringements of animal rights. The destruction of biodiversity is a controversially discussed problem, which is tightly connected with the development of modern societies. Usually it is presupposed that the description and quantification of biodiversity is scientifically and unambiguously clarified. In contrast to this expectation, neither biology nor other disciplines dealing with corresponding parameterizations, such as economics, have presented uniform concepts which can serve as a basis for measurability and comparability of biodiversity.

4.2.5 Bio-patenting. Genetic research and its commercial use have stimulated a debate on the moral acceptability of patenting human and non-human genes. Though there exists a number of national and international conventions on the patentability of bio-materials there remains considerable uncertainty on exactly what can be patented and what rights are connected to awarded patents.

The moral debate on patents centers around three major questions: are 'patents on life' immoral and is, therefore, the patenting of plants and animals, but in particular of human genes, morally unacceptable? Does the patenting of bio-materials—in particular human genes—lead to a massive hindrance to biomedical research? Does bio-patenting lead to an unfair distribution of resources between the rich, industrialized countries and the developing countries? None of these questions can be said to have been solved at the moment and strong political and economic interests make this issue important well beyond the academic area

See also: Animal Rights in Research and Research Application; Bioethics: Philosophical Aspects; Ethical Dilemmas: Research and Treatment Priorities; Ethical Issues in the 'New' Genetics; Ethics Committees in Science: European Perspectives; Ethics for Biomedical Research Involving Humans: International Codes; Euthanasia; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Testing and Counseling: Educational and Professional Aspects; Intellectual Property Rights: Ethical Aspects; Pecuniary Issues in Medical Services: Ethical Aspects; Reproductive Medicine: Ethical Aspects; Research: Ethical Aspects of Long-term Responsibilities; Research Ethics, Cross-cultural Dimensions of; Research Subjects, Informed and Implied Consent of; Risk Screening, Testing, and Diagnosis: Ethical Aspects

Bibliography

Agar N 1998 Liberal eugenics. *Public Affairs Quarterly* **12**, 2(April): 137–55

Arrington R L 1997 Ethics I (1945 to the present). In: Canfield J V (ed.) *Routledge History of Philosophy Volume X*. Routledge, London

Beauchamp T L, Childress J F 1994 *Principles of Biomedical Ethics*. Oxford University Press, New York

Chadwick R (ed.) 1998 *Encyclopedia of Applied Ethics*, 4 Vol. Academic Press, San Diego, CA

Chadwick R, Shickle D, ten Have H, Wiesing U (eds.) 1999 *The Ethics of Genetic Screening*. Kluwer Academic Publishers, Dordrecht, Germany

Faden R R, Beauchamp T L 1986 *A History and Theory of Informed Consent*. Oxford University Press, New York

Frankena W K 1973 *Ethics*. Prentice-Hall, Englewood Cliffs, NJ

Gethmann C F 1989 Proto-ethics. Towards a formal pragmatics of justificatory discourse. In: Butts R E, Burrichter J R (eds.) *Constructivism and Science. Essays in Recent German Philosophy*. Kluwer Academic Publishers, Dordrecht, Germany

Harris J 1997 Goodbye Dolly? The ethics of human cloning. *Journal of Medical Ethics* **23**: 353–60

Kuhse H, Singer P (eds.) 1999 *Bioethics. An Anthology*. Blackwell Publishers, Oxford, UK

Lauritzen P (ed.) 2001 *Cloning and the Future of Human Embryo Research*. Oxford University Press, New York

Reich W T (ed.) 1995 *Encyclopedia of Bioethics*, 5 Vols. Macmillan, New York

Singer P 1994 *Rethinking Life & Death*. The Text Publishing Company, Melbourne, Victoria

Steinbock B, Norcross A (eds.) 1994 *Killing and Letting Die*. Fordham University Press, New York

Stingl M 1997 Ethics I (1900–45). In: Canfield J V (ed.) *Routledge History of Philosophy Volume X*. Routledge, London

Sorell T (ed.) 1998 *Health Care, Ethics and Insurance*. Routledge, London

Tooley M 1972 Abortion and infanticide. *Philosophy and Public Affairs* **2**(1): 37–65

Wilmut I 1998 Cloning for medicine. *Scientific American*, December(12): 58–64

F. Thiele

Bioethics: Philosophical Aspects

Etymologically, the term 'bioethics' derives from the Greek words *bios* (life) and *ethike* (ethics), and literally means the ethics of life. In practice, however, it denotes a field largely concerned with the ethical analysis of normative problems in the biomedical sciences, medical practices, and health care management. Reflecting this focus, this article addresses the nature and history of bioethics, the dominant ethical theories employed in bioethical reasoning, and their influence in the main areas of bioethical research.

1. Bioethics—Scope and History

Bioethics, both as an academic discipline and as a professional activity, has evolved from the domain of medical ethics. It is a multidisciplinary field that extends far beyond the spheres of healthcare and medical ethics. It encompasses a wide range of ethical problems in the life sciences. These include issues related to genetics, theories of human development, behavioral psychology, and resource allocation in healthcare management.

Bioethical expertise is sought by courts, parliaments, research ethics committees, and is used in clinical consultations to guide the behavior of medical professionals. Despite its practical appeal, however, disagreement exists about the nature and scope of bioethics as a professional/theoretical discipline. Bioethicists come from a diverse range of professional backgrounds, including healthcare professions, philosophy, jurisprudence, sociology, and theology. Debate about the proper relations between the disciplines contributing to bioethical analysis is continuing (Carson and Burns 1997).

1.1 Expanding Issues in Medicine

The reasons for the progression of bioethics as a multidisciplinary activity can perhaps best be understood by considering the societal context in which it began. The rapid developments with regard to both the availability of biomedical technology and dramatically advancing biomedical research required a rethinking of traditional approaches to medical ethics. The 1960s saw not only the first successful organ transplantations, the contraceptive pill, and the arrival of prenatal diagnosis, but also a shift to the deployment of highly technological medicine. Particular people began to demand greater liberties, such as equal rights for those who were non-Caucasian, female, and gay. Medical paternalism, and the attitude that the 'doctor knows best' were no longer considered sacrosanct. The Western world saw the progression

towards civil liberties, and changes in the practices of the biomedical sciences began to reflect the modern emphasis on autonomy. Informed consent, the protection of human subjects, and patient self-determination began to assume paramount importance in both theoretical and legal contexts.

1.2 Bioethics Teaching

Bioethics is currently taught in many diverse forms throughout both the developed and developing world. Bioethicists teach in medical faculties, medical humanities, and philosophy programs, and in one of several hundred research centers in postgraduate degree programs. The socioeconomic and cultural contexts of teaching influence the topics that constitute a given syllabus. For example, in developing countries, issues such as exploitation in organ transplantation or human experimentation feature more prominently than in developed nations. Teachers try to instill heightened sensitivity in their students to ethical issues and values in medicine, as well as provide them with skills in dealing with ethical problems arising in healthcare practice. Courses are often taken as part of obligatory continuing medical education requirements.

2. Theoretical Approaches to Bioethics

Philosophical ethics, the theoretical heart of bioethical analyses, consists of a variety of competing ethical theories. Utilitarians, deontologists, casuists, communitarians, contractarians, virtue ethicists, and ethicists of other persuasions appeal to differing modes of ethical reasoning. This has consequences for bioethical analyses since the type of ethical theory or religious framework to which a given bioethicist subscribes will be reflected in the practical advice advocated.

2.1 Principle-based Bioethics

Principle-based bioethics typically refers to an approach developed by Beauchamp and Childress (1994). They propose a system of bioethics comprised of four major principles: autonomy, beneficence, nonmaleficence, and justice. These *prima facie* principles primarily concern respect for the choices people make, the obligation to help, but not harm, other parties, and the requirement to act in a fair and equitable manner with regard to the distribution of medical burdens and benefits. Critics charge that principle-based bioethics is unsuitable for practical decision making because the lack of hierarchical order of principles renders the ranking in any given situation somewhat arbitrary. Because it is able to be taught in a comprehensive and accessible manner, however, this approach is favored in many bioethics teaching programs.

2.2 Utilitarian Bioethics

Utilitarians have developed a consequentialist type of theory, judging the rightness or wrongness of a given action exclusively by its consequences. Of all the ethical theories, utilitarian modes of reasoning are most easily suited to problem solving in bioethics. This is because their guiding principle is singular and unambiguous, providing a clear procedure for decision making. The basic utilitarian premise is that both individual action and public policy should maximize utility (which is normally defined in terms of happiness or preference-satisfaction) for the greatest number of people. Its patterns of analysis are congruent with traditional forms of reasoning in public policy. Utilitarians have contributed sophisticated works to central problems in bioethics, such as abortion, voluntary euthanasia, embryo experimentation, and resource allocation.

Critics of utilitarianism doubt that it is possible to quantify interest or preference satisfaction, and question the feasibility of the utilitarian calculus as a decision directive. Also controversial is the utilitarian rejection of two distinctions which are central to other approaches in medical ethics: the intention/foresight and acts/omissions distinctions. These distinctions are important in terms of motives and responsibilities for action, both of which utilitarians ignore, focusing exclusively on consequences as criteria of right action.

2.3 Deontological Bioethics

Secular deontological approaches to bioethics tend to be based on Kantian and neo-Kantian ethical theories, and feature most prominently in the areas of resource allocation and general social health policy. Daniels (1985), for example, uses a neo-Kantian form of contractarianism to support his influential argument for a universal right to healthcare. The Kantian moral agent is quite different from the utilitarian agent who acts in order to satisfy interests or desires.

Kant was concerned with the motivation of action, and argued that duty alone should motivate morally adequate action. An 'action done from duty has its moral worth not in the purpose to be attained by it, but in the maxim in accordance with which it is decided upon' (Kant 1976). These maxims are constructed as absolute imperatives of the sort, 'don't kill,' 'don't lie,' etc. Furthermore, Kant developed

Categorical Imperatives, the most influential of which demands that the moral agent never treat other people as mere means to ends (however noble these may be), but rather as ends in themselves.

There are some reasonably clear differences between Kantian and utilitarian bioethical decision making. Utilitarians and Kantians arrive at very different answers to the question of whether it is ethically justifiable to kill one 'innocent' human being in order to save 10 similar others who are in need of organ transplants. All other things being equal, utilitarians would conclude that it is right to kill one person to save the lives of 10 others, while a Kantian would argue that it is unacceptable to kill an innocent person. Kantians are not concerned about the consequences of action, but rather with the question of whether one can consistently wish to be treated by other rational agents in the same manner as one desires to act in a comparable situation. Kantian bioethicists tend to defend absolutist positions, such as the rejection of voluntary euthanasia, irrespective of the suffering this may cause.

2.4 Feminist Bioethics

Feminist approaches to bioethics are unified by a common concern for women's oppression. Feminist bioethics raises the question of unequal distribution of social power, and the subsequent biases that manifest in the life sciences, medical practices, and philosophical ethics. There are numerous feminist approaches to these fields. Feminist projects provide a critique of traditional approaches to bioethics, focusing on the effects that the historical exclusion of women and their experiences have had on theory production. Some feminists aim merely to provide a *corrective* to dominant theories and practices, while others embark on the more radical project of creating positive feminist theories and practices which differ *fundamentally* from those which are the objects of critique. Whether one stops at critique, or progresses to creating novel theories founded on new premises, depends largely on how deep the level of masculine bias is held to be. Bioethics provides fertile ground for feminist analysis, since scientific claims have long been used as tools to justify women's oppression as part of the natural order of things. Feminist interactions with the sciences aim to expose the value-laden nature of scientific practices; how our social context effects the sort of questions asked, the methodologies employed, and the sort of answers that are considered coherent. Feminists have been central to debates on abortion and reproductive technologies. The scope of feminist argument is not, however, limited to those domains explicitly related to women; it spans the entire field of bioethics. Some theorists have demanded that feminist theory also consider the ways in which race, ethnicity, and class bear upon theory construction. For these reasons, many feminists reject the goal of one universal moral theory in favor of a more pluralistic approach to knowledges and their application.

Commonly, feminists take issue with the *sort* of human agent that moral theory assumes. Specifically, it is argued that certain features of this subject (i.e., its participation in an alleged objective, universal human Reason) reflect a mode of reasoning which depends on the exclusion/devaluation of women's experiences. Contemporary models of moral reasoning drawn from experiences in the private sphere (such as mothering, caring for the aged, and preparing and distributing food) are substantially different from those based on hitherto ignored experiences common to the public arena, which has historically been the domain of men. Indeed, care- (Gilligan 1982, Noddings 1984) and maternal- (Ruddick 1995, Held 1993) based are perhaps the main challenges to traditional moral theory produced by Anglo-American feminists. From a utilitarian perspective, their implications have been most thoroughly explored in the domain of nursing (Kuhse 1997).

2.5 Communitarian Bioethics

In the context of bioethics, communitarianism constitutes a critical discourse that challenges dominant approaches which are based on Enlightenment philosophies. The idea of an isolated knowing subject, who defines ethical truth in accordance with the dictates of human reason, is rejected. Communitarians privilege the interests and values *of the community* in negotiating morally optimal courses of action. They argue that any moral theory is determined by socio-historical context, including the traditions, religions, and culture of a community. Within traditional moral theories, these influences are often seen as extraneous, contaminating the goal of objective moral truth. For communitarians, however, attempts to produce an impartial, universal moral theory are not only futile, but represent a mistaken concept of what moral thought *should* involve. They emphasize that certain decisions require a view of the good, of what is worthwhile and valuable for a community.

The goal of defining goods and values that are genuinely shared by any given community is problematic. In increasingly multicultural societies and global communities, identifying shared conceptions of the moral good might only be possible if communities are defined in a narrow sense, resulting in many communities with perhaps competing views of the good life.

2.6 Religious Bioethics

Religious ethics is not a unified ethical theory, like Kantianism. Religious ethical analyses will always

depend on the particular religious scripture that guides the ethicist. Monotheistic religions are traditionally less open-minded with regard to pluralistic views that might be held in secular societies. Still, even within Christianity and Judaism, there is a range of diverse views held by bioethics scholars. Catholic bioethicists tend to be in strict agreement with the teachings of their organization. This is largely due to the fact that Catholicism is the only mainstream religion that has a body of papal ex cathedra teaching. Even here, however, the Biblical commandment 'thou shalt not kill' has undergone some theological reinterpretation. Despite this, there is unanimity among Catholic scholars that active euthanasia and abortion are ethically unacceptable. In many countries, Islamic medical associations have produced their own code of ethical conduct.

3. *Issues in Bioethics*

3.1 *Research Ethics*

Research ethics came about largely because of revelations regarding the gruesome medical experiments conducted by Nazi doctors in German concentration camps during the Third Reich. The most important international normative framework regulating the standards of research clinical trials is the Declaration of Helsinki (WMA [World Medical Association] 1964). In 1993, it was supplemented by international research ethics guidelines produced by the CIOMS (Council of Medical Organisations of Medical Sciences) in collaboration with the WHO (World Health Organisation). This set of documents contains a series of important protections for people participating as research subjects. These are especially important for people in developing countries, where exploitative trials continue to take place. Bioethical analyses of research ask questions such as the following:

(a) Ought researchers be held responsible for the use of their research results by other parties?

(b) What is the appropriate point in time to stop research clinical trials?

(c) Are placebo controls defensible in trials with terminally ill patients?

(d) Are terminally ill patients justified in breaking protocols in placebo controlled trials, given that many only joined the trial in order to access the experimental agent (Schüklenk 1998)?

(e) Ought women of childbearing age be enrolled in research clinical trials?

(f) Ought prisoners be asked to participate in nontherapeutic research?

A key issue in research ethics is informed consent, which is considered a precondition for any ethical research, given that it is voluntary and autonomous. Justifications for this premise derive from the idea that individual autonomy is of intrinsic and/or instrumental value. Neo-Kantians argue that autonomy is of great instrinsic value. Young (1986), for example, understands the value of autonomy as a character ideal, rather than as a means to some further good. Utilitarians also support the idea of respect for informed consent because they, too, value personal autonomy. In this instance, personal autonomy is of instrumental value, necessary to satisfy one's own preferences, desires, and interests. Research ethics protections and standards pertaining to research subjects in developing countries are also issues of paramount importance. AIDS research in particular has led to large numbers of international collaborative research projects. International research ethics guidelines require that every patient (including those in any control group) should be assured of access to the best proven diagnosis and therapeutic method. This requirement was designed to prevent the exploitation of research subjects in developing countries who participate in projects undertaken by Western researchers. Some bioethicists have argued that the enforcement of Western norms in developing countries constitutes cultural imperialism. Utilitarians are most likely to inquire after the consequences of accepting a lower standard of care for research subjects in developing countries. If such a strategy would lead to the development of cheaper, perhaps affordable, drugs for people living in developing nations, utilitarians would accept such a policy change.

Deontologists argue that the primary obligation of physicians is to their research subjects and not to future generations of patients. This makes them more ideal advocates of current research subjects' interests, as opposed to utilitarians who would have to take into account the interests of all concerned, including those of future generations. Principalists would have to balance the duties of nonmaleficence and beneficence against each other in this case. They suggest that we have an absolute duty not to harm anyone, but that we have no absolute duty to do good. This view could be used to justify providing local standards of care in international collaborative research efforts, because providing the highest attainable standards of care would require providing more than what is already available. This is arguably doing good over and above the call of duty.

3.2 *Beginning and End of Life*

The moral status of embryos, fetuses, and infants continues to dominate beginning of life debates. Some ethicists suggest that fetuses deserve moral standing and consideration by virtue of their belonging to the human species. Others reject this proposal as speciesist, arguing that the mere fact of being human does not give fetuses special status (Singer 1990).

Rather, their moral standing ought to depend on their dispositional capacities.

The debates about the moral status of embryos and fetuses primarily affect issues surrounding new reproductive technologies and abortion.

Catholic bioethicists hold that nothing can justify the killing of innocent human beings, irrespective of whether they are fetuses, infants, adults, or people suffering from terminal illnesses. Abortion is considered unethical, even when the mother's life is at stake (Sacred Congregation for the Doctrine of Faith 1980). These bioethicists argue that fetuses are human beings from conception, and ought to be accorded the same moral entitlements to life as other humans. In turn, embryo experimentation is also deemed unacceptable. This view is often supported by deontological ethicists who propose an absolute sanctity of all life. The vast majority of feminists support (and have been crucial in creating and maintaining) the legal right to abortion on the basis of the mother's interest in/right to bodily autonomy and integrity. This interest/right is often seen as overriding any moral status that the fetus might have. Feminists have also argued that maternal–fetal conflict models inaccurately characterize what is at stake in pregnant women's decisions which affect their fetus, and alternate models which do not centralize conflict have been proposed.

Bioethical debates concerning the end of life tend to be preoccupied with the permissibility of physician-assisted suicide and voluntary euthanasia. Many people suffering from terminal illnesses find their quality of life during their last months of life so unbearable that they ask physicians to help them die, either by means of supplying them with lethal doses of drugs, or by killing them actively. The Hippocratic Oath requires doctors to 'neither give a deadly drug to anyone if asked for it, nor make a suggestion to this effect.' Accordingly, medical ethics has traditionally seen doctors' assistance in patients' attempts to die as incompatible with their roles as lifesavers. The sanctity-of-life doctrine in medicine, strongly influenced by religious views, rejects the view that some lives might be so bad that they are not worth living. Utilitarians argue that we ought to abandon the sanctity of life doctrine and substitute it with a quality of life ethics instead (Kuhse 1987).

3.3 The 'New Genetics'

Molecular genetics, and its associated research and engineering techniques, are often called the 'new genetics.' This field facilitates the manipulation of living organisms to a hitherto unimaginable degree. We are able to clone animals, research treatments of diseases with both somatic and germ-line gene therapy, select animal and human embryos and fetuses on the basis of their genotype, create microorganisms of desired genotypes, and have them, as well as plants, express both human and animal genes which code for desired protein products. Questions such as the following are now widely debated:

(a) Is reproductive/therapeutic *human* cloning morally acceptable?

(b) Does selecting against people with disease genes through prenatal screening constitute discrimination?

(c) Can we define disease genes in morally neutral ways?

(d) Does the desire for positive human traits such as physical beauty or intelligence justify the selective abortion of fetuses with 'suboptimal' genotypes?

(e) What ecological effects are likely as we cross species barriers by creating hybrid organisms?

(f) How can we control the creation of genetically engineered pathogens tailored for particular human populations in the service of military or terrorists' agendas?

Human society remains deeply troubled by genocidal wars, and discrimination against people on the basis of race, ethnicity, sex, and sexual orientation is rife. The identification of genetic markers of marginalization opens new possibilites for eugenic programs, racial 'cleansing,' and genetically targeted biological warfare. It is thus clear that the 'new genetics' may be used for malign purposes as well as for the good of humanity. An example of the ways in which the application of moral theories to genetics differs can be found in discussions of prenatal testing and subsequent selective abortion for nonmedical reasons. What constitutes a *medical* reason is itself open to debate, and the question of whether abortion, based on the presence of, say, a gene predisposing to late-onset cancer, is justified, is far from settled. Screening and abortion for nonmedical reasons are far more controversial. A dominant utilitarian approach to this issue would look to individual preferences, and would be likely to endorse a liberal approach, where uptake of such technology would be a matter of individual choice. For as long as fetuses are not considered entities either with interests able to outweigh those of parents, or with the capacity to suffer significantly, there would be no reason against the provision of selective abortion. The only utilitarian argument against such technology would be one which held that the indirect consequences for society in general would be so detrimental that they outweighed the direct preferences of potential users.

A deontologist's approach might oppose the provision of prenatal screening in order to facilitate selective abortion for nonmedical reasons. Motives for action are crucial for a deontologist. Deontologists emphasize respect for human dignity, which entails that all people be treated equally, and that they not be used merely as means to another's ends, but rather, always as ends in themselves. It is possible that abortion *per se* would be opposed on the basis of respect for human potentiality. Differentiating be-

tween people on the basis of sex or hair/eye color would be incompatible with respect for human dignity, and associated selective abortion would certainly be a case of using a potential person for another's, perhaps trivial, ends (Davis 1997).

A liberal feminist bioethicist's response might endorse eugenic technology in accordance with respect for a woman's right to autonomy, and control over the pregnancies with which she chooses to continue (Warren 1985). Feminists would also be concerned with the social conditions in which technology is developed, and the demand for its provision generated. Attentive to the ways in which science might work in the interests of the dominant sex, race, and class, any technology which facilitates discrimination against these markers of marginalization would be viewed with suspicion. Thus, prenatal testing for sex might not be supported in societies where there is entrenched oppression/devaluation of women, and a concomitant preference for male children.

3.4 Healthcare Professional–Patient Relationship

Some writers suggest that the relationship between healthcare professionals (HCPs) and patients is best understood as a contractual agreement, similar to that between a customer and a professional selling service. A contract usually includes protections to the partners such that if the contract is broken by one party (i.e., by providing substandard treatments, or by not paying the due fee), legal recourse can be sought by the wronged other. The advantage of this model is that it breaks with authoritarian models of the physician. Instead it requires respect for the patient's autonomy, usually first-person voluntary informed consent to treatment, and agreement between doctors and patients about a proposed course of action. It has been proposed that medical practice ought to rely on a covenant model instead of the code-based model that it tends to rely on in actuality (May 1975). Others have seen doctors' service provision and associated medical paternalism as an extension of their professional duty of beneficence (Downie 1994). This traditional attitude has been challenged by claims of patients' rights to medical care and doctors' obligations to satisfy these rights.

US bioethicists have attempted to address ethical concerns of the HCP–patient relationship in the context of the principle-based approach. This strategy turned out to be quite problematic because of the conflicts between simultaneously having to respect autonomy, act beneficently, and act justly.

Utilitarians defend upholding patient confidentiality by pointing to the negative consequences for the physician–patient relationship once patients cannot trust that their medical details will be kept confidential. They suggest that maintaining patient confidentiality is necessary for achieving the health and well-being of patients. Deontologists support this conclusion on the basis of a distinctly different rationale. They argue that respect for patient autonomy requires that patients keep control over what type of intimate, private information is available to parties other than those to whom they disclose it.

3.5 Resource Allocation

Determining the optimal means of allocating scarce resources for healthcare is an intractable problem requiring constant negotiation as new technologies and treatments become available. While the UN Declaration of Human Rights states that every person 'has the right to a standard of living adequate for the health and well-being of himself (sic);' defining the criteria of such adequacy is open to debate. As is the case with any limited resource, there are always situations in which the number of people desirous of a particular form of healthcare outweighs what is available. On a macrolevel, decisions must be made regarding the extent to which healthcare should be state-funded, and the way in which such provisions should be distributed between medical research programs, available treatments, and technologies. On a microlevel, medical staff must confront issues such as how to allocate hospital beds, order of, and eligibility for, treatments. Prominent criteria for resource allocation decisions are need, merit, desert, order of demand, social utility, and expected benefit from treatment. Some theorists suggest that it is only when consensus about degree of need cannot be reached that other factors come into play (Harris 1998). Where a maximizing strategy is endorsed, however, it is largely the consequences of treatment that matter, and medical need is not necessarily the primary criterion. Resolving some way of considering these issues is particularly pertinent in the face of an aging population, where healthcare becomes increasingly expensive, and quality of life considerations become pressing. For example, how should funding be distributed between geriatric and newborn-infant care?

Callahan (1990) argues that this question is necessarily linked to the goals of the community. A communitarian approach might articulate the goals of medicine in terms of particular stages of the life cycle, while also thinking about what the goals and virtues of the elderly should be. This entails the possibility that elderly persons might be morally required to forego certain life-prolonging treatments in the interests of providing healthcare for younger generations. A utilitarian method uses quality adjusted life years (QALYs) in order to help determine who has most claim on limited treatments. QALYs provide a means of calculating which individuals, with treatment, can expect to live the longest, where the number of years

are adjusted for quality. Thus a person predicted to live 10 *actual* years after treatment with a very low quality of life would have less than 10 QALY years to live, and would lose out to a patient predicted to live, say, eight actual years with a good quality of life. Similarly, if we imagine an elderly person and a youth, both in need of an organ transplant, and both with similar qualities of life expected postoperation, it is clear that the QALY method would favor the young person. A utilitarian may, of course, choose to include other factors beside QALYs in their calculation of who should get the organ. If the elderly person had, say, a large number of refugees dependant on them for survival, then this may count as a compelling reason to favor their receipt of the organ. Nonetheless, whether or not the initial favoring of youth amounts to ageism is debatable. Indeed, whether or not the QALY method discriminates against disabled people is also questionable, since mental/physical disabilities are generally assumed to reduce quality of life. Any method of assessing quality of life remains controversial. Some argue that quality of life is a subjective concept which can only be meaningfully assessed by the individuals in question. Other theorists who maintain that quality of life can be judged objectively disagree over the relevant criteria in such a decision. Thus it has been suggested that some measure of both subjective and objective evaluations are needed in order to measure health states (Wikler and Marchland 1998).

See also: Consequentialism Including Utilitarianism; Ethical Codes, Professional: Business Codes; Ethics and Values; Rights

Bibliography

Beauchamp T L, Childress J F 1994 *Principles of Biomedical Ethics*. Oxford University Press, New York

Callahan D 1990 *What Kind of Life?* Simon and Schuster: New York

Carson R A, Burns C R (eds.) 1997 *Philosophy of Medicine and Bioethics: a Twenty-Year Retrospective and Critical Appraisal*. Kluwer, Dordrecht, The Netherlands

Daniels N 1985 *Just Health Care*. Cambridge University Press, Cambridge, UK

Davis D S 1997 Genetic dilemmas and the child's right to an open future. *Hastings Center Report* **27**(2): 7–15

Downie R S 1994 The doctor–patient relationship. In: Gillon R (ed.) *Healthcare Ethics*. Wiley, Chichester, UK, pp. 343–7

Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA

Harris J 1998 Micro-allocation: deciding between patients. In: Kuhse H, Singer P (eds.) *A Companion to Bioethics*. Blackwell Publishers, Oxford, UK

Held V 1993 *Feminist Morality: Transforming Culture, Society and Politics*. University of Chicago Press, Chicago

Kant I 1976 *Critique of Practical Reason*. Garland, New York, pp. 67–8

Kuhse H 1987 *The Sanctity-of-Life Doctrine in Medicine*. Oxford University Press, Oxford, UK

Kuhse H 1997 *Caring: Nurses, Women and Ethics*. Blackwell, Oxford, UK

May W F 1975 Code, covenant, contract, or philanthropy. *Hastings Center Report* **5**(6): 29–3

Noddings N 1984 *Caring: a Feminine Approach to Ethics & Moral Education*. University of California Press, Berkeley, CA

Ruddick S 1995 *Maternal Thinking: Towards a Politics of Peace*. Beacon Press, Boston, MA

Schüklenk U 1998 *Access to Experimental Drugs in Terminal Illness: Ethical Issues*. Pharmaceutical Products Press, New York

Sacred Congregation for the Doctrine of Faith 1980 *Declaration on Euthanasia*. Vatican City, Rome

Singer P 1990 *Animal Liberation*. Review of Books, New York

Steinbock B 1998 Mother–fetus conflict. In: Kuhse H, Singer P (eds.) *A Companion to Bioethics*. Blackwell, Oxford, UK, pp. 135–46

Warren M A 1985 The ethics of sex preselection. In: Humber J, Almeder R (eds.) *Biomedical Ethics Reviews*. Humana Press, Clifton, NJ

Wikler D, Marchland S 1998 Macro-allocation: dividing up the health care budget. In: Kuhse H, Singer P (eds.) *A Companion to Bioethics*. Blackwell, Oxford, UK

World Medical Association 1964 Declaration of Helsinki. In: Jansen A R, Veatch R M, Walters L (eds.) *1998 Source Book in Bioethics*. Georgetown University Press, Washington, DC, pp. 13–15

Young R 1986 *Personal Autonomy: Beyond Negative and Positive Liberty*. St. Martin's Press, New York

U. Schüklenk and J. Kerin

Biogeography

Biogeography is not easy to be defined precisely and this is for several reasons. Some scientists think that biogeography is a biological discipline, firmly rooted in nothing but the biological sciences. Biogeography, or the two common subdivisions phytogeography and zoogeography, is the field for studying the distribution of plants and animals, respectively. Due especially to the higher diversity and the greater mobility of animals, zoogeography has developed less than plant geography. Moreover, plants are more directly affected by their physical habitat and exert at the same time a greater influence on climatic and edaphic properties than animals. They better indicate visibly the environmental condition they live in.

On the other hand, many geographers claim biogeography to be a geographical discipline and regard it as a most important link between physical and human geography, because man's effects on plants and animals are included. They stick firmly to the ecosystem concept and regard biogeography as a vital part of ecological geography or geoecology. And they underline the often neglected meaning of scales in time and space as well as the dynamics of the ecosystems.

The struggle between biologists, who prefer a more

historical view and underline the importance of evolution, and geographers, who accentuate ecological relations and anthropogenic changes, is not very productive. As the term indicates, biogeography is at the same time a biological and a geographical science. The roots of biogeography lie in the classical 'Natural History.' Probably Alexander von Humboldt was the first biogeographer, a generalist who was able to do research in botany and zoology as well as in geology, geography, climatology, and soil science. He recognized the very complex nature of plant and animal distribution patterns and was aware of the necessity of a broader, interdisciplinary view on the basis of the integration of data and information from numerous disciplines of natural and even social and economic sciences. And for this tremendous diversity of biogeography, a holistic approach is indispensable. Only biologists and geographers together can bridge the gap between the topics involved.

For these reasons a short and comprehensive definition of biogeography is rather impossible. The main contents, tasks, and scientific approaches of biogeography can be summarized as follows.

Biogeography is the science of living organisms in time and space. Biogeographers combine geographical and biological methods in order to discover, to present, and to explain the development of the spatial distribution patterns of single organisms and of organisms associated in phyto-, zoo- and biocenosis. In this approach, evolutional dynamics and historical developments have to be taken into consideration as well as the ecological background with abiotic factors such as temperature, light, water, nutrients, and mechanical influences, which the environment offers in a given quality and quantity, and the demands of plants and animals respectively. Biotic factors, as far as they concern competition and especially human influences, are of outstanding importance as well. Basic knowledge about the dynamics of the area of species/biocenosis is a prerequisite for a better understanding of ecosystem dynamics as well as for treating properly applied problems, e.g., concerning bioindication in a polluted environment, protection of natural reserves, or the maintenance of biodiversity.

In order to understand the distribution patterns of organisms and biocenosis, which is the main objective of biogeography, all relevant historical and ecological factors on all temporal and spatial scales must be integrated. Because of the complex nature of biogeography and the countless interrelations between different elements of different scientific disciplines, it is almost impossible to find a logical order for treating the most important aspects separately. The following subdivision into several sections is highly artificial and is only included for practical reasons. And for reasons of interest and knowledge, attention is primarily given to the study of plants rather than animals.

Evolutional and historical developments and the ecological background of distribution patterns will be treated in the following sections. Aspects of regional biogeography and some general rules and models follow. In a concluding section, topics of applied biogeography are addressed.

1. *Historical Biogeography*

Historical biogeography is concerned with the evolutionary processes which contributed to the actual distribution patterns of plant and animal life. The chorological dynamic is focused over very large temporal and spatial scales. The effects of past events on the present distribution and all the environmental changes that have taken place during the time of origin of a species until today are studied. Evolution for all biota means change, too, and this change—due to environmental changes—can be accomplished by genetical adaptation and/or by spatial adaptation. If changes of the environment are slow and continuous, species adapt and speciation can take place; if they happen too fast, extinction may be the consequence. For some biogeographers this approach is not only important but is the whole content of biogeography. Most important is the perception that biogeography has a vital component that has been often neglected: dynamics in time and space. Distribution patterns can only be explained by understanding the underlying processes.

Basic processes are all dispersal mechanisms. They can be decisive for the migration potential of organisms; the spatial range a species may colonize. Active dispersal (autochory) and different modes of passive dispersal (allochory) are responsible for the distribution of diaspores. Among the latter, anemochory (dispersal by wind) and different types of zoochory (dispersal by animals, especially birds and numerous insects) are most successful. But there are more dispersal modes; more recently, man has an increasing importance (anthropochory).

Not only the dispersal mode but also the actual and the former constellations of the continents and oceans highly influence a more or less successful expansion of the area of a given species. Terrestrial plants may cross marine barriers, where animals did not succeed. Some seeds tolerate salt water, others do not. For climatic and orographic reasons, mountain ridges are sometimes invincible barriers. Species with a successful dispersal mode, which have been able to cross all such barriers by land bridges, stepping stones, long-distance jumps, or special transportation can achieve a cosmopolitan range. Others stay restricted to their area of origin and remain endemics. Also, the area of formerly widespread organisms can be split up by newly opened oceans or other barriers. The result is disjunction and probably vicariance.

Not only dispersal mechanisms but also continental drift, plate tectonics, landmass fragmentation, and marked climatic changes contributed greatly to form-

ing the actual distribution patterns of plants and animals. Geological changes are mainly linked with their large-scale distribution, climatic changes also with finer-scale patterns. Especially important during the Paleozoic has been the initiation of the breakup of the old Pangea. Mesozoic times are marked by the great extinctions of formerly important groups (dinosaurian) and the rise of flowering plants. Partly due to sophisticated research techniques like radiocarbon data and palynological analysis, knowledge about the influence of the Ice Ages on the distribution of plants is rich and detailed. Temperature fluctuations and eustatic changes of the sea level have opened new dispersal pathways for some species or forced others to look for refuges.

Best documented is the dynamics of plant and animal distribution in the last few thousands of years. Nevertheless, there is now one element which is new and complicates the interpretation of a given distribution pattern. This element, which must also be regarded as an ecological factor of growing importance, is man and his manyfold activities and influences. Domestication and Neolithic agriculture have been the impact of early human culture. Now not only all energy and nutrient cycles can be manipulated but also any still-existing geographical barrier can easily be bridged. Man can change the genetic setup of organisms as well as create new site conditions. Pollution in a very large sense is a completely new aspect in biogeography. 'Greenhouse-effects,' 'forest die-back,' 'eutrophisation,' and more are the new keywords. The problems arising with this development show that there are growing challenges for modern biogeography.

Research on historical biogeography has always been accompanied by theoretical and modeling efforts. The best known are linked to the different succession theories and to the idea of a final stage of development ('climax') and to plant-strategy types ('r'- and 'K'-strategists). More recently, vegetation dynamics are discussed on the basis of models, which take the role of competition, biodiversity, and plant functional types more into account. The stability of communities of plants and animals, and finally of complete ecosystems is intensely discussed. Furthermore, the problem of temporal scales is gaining increasing recognition.

2. *Ecological Biogeography*

Ecological biogeography is concerned with ecological processes, which contribute to the distribution of plant and animal life in the biosphere. It is important to distinguish the historical elements within the actual patterns from the ecological ones. After the understanding of the geological history and evolution of life on earth, it is now the present abiotic and biotic environment, the major ecological factors on short temporal and often fine spatial scales, which is focused on. Especially, geographers preferred this ecological concept to the historical concept.

The first and basic ecological approach (autecology) is concerned with relations between organisms and the abiotic complexes like climate, soil, relief forms, or abiotic factors such as light, temperature, water, and chemical and mechanical factors of their habitat. The most important factors affecting the large-scale distribution of plants and animals are climatic factors. Climate is generally considered as the basis for explaining biogeographical patterns since long-standing close relations between climatic parameters and the distribution of large vegetation units such as tropical savannas or boreal forests are known. Timberline studies especially underline these relations. On finer scales, soil chemistry and moisture levels, or mechanical disturbances like fire or avalanches may be more decisive. Autecologists always ask two questions. The first one concerns the quality and quantity of a chosen factor that affects the environment. Geographers often concentrate on this question whereas biologists prefer working on the second one. This concerns the ecological amplitude of plants or animals with respect to the chosen factor; for instance, temperature requirements for seed germination or availability of a minimum amount of water.

Today, the most important ecological factors affecting plant and animal life are certainly the manifold influences of man. As already pointed out in the chapter about historical biogeography, human impact on purpose or by chance often directly changed the close natural relations between organisms and their environment by changing site qualities as well as the genetic constellation of organisms. Man is regarded not only as an integral part of the global biocenosis but also as the ecologically dominant organism in it. Results of recent research prove that the role of man is still of increasing relevance for our future.

A second, more complex and more important approach of ecological biogeography is concerned with the relations between organisms (synecology). Already man's influences may be regarded as synecological processes. In general, synecology concerns all mutualistic relations (e.g., symbiosis, parasitism) and often most complex interactions between individuals or species of plants and animals including the important plant–animal interactions. Inter- and intraspecific competition as well as the responses of plants, animals, and their communities to every kind of stress have to be understood in order to explain distribution patterns. Competition can eliminate those species that are only poorly adapted to their environment and may determine which individuals will survive or fail. Some species may tolerate a broad range of stress factors and can survive; others are unable to colonize in the same situation.

The ecosystem concept, to organize the very complex topic of ecological interactions and functional

relations, which has been developed by geographers as well as by biologists, is a fundamental integrating concept. It shows the interactions between the living and non-living world and helps to understand the way in which the natural environment and its abiotic and biotic components operate. Here, the fluxes of energy, the cycling of nutrients, and the exchanges of information can be demonstrated. It is possible to give an overview on food chains and trophic levels and to show the participating producers, consuments, and destruents in the systems context. So far, the ecosystem concept quite often has been used for modeling ecological processes.

The term 'ecosystem' is also used for spatial entities but is not restricted to a certain scale. This concept can be adapted as well to a complete climatical zone as to the flowerhead of a thistle, where different insects struggle for the same resources. It is often utilized for lower levels of spatial and organizational hierarchies such as biotopes and ecotopes or populations and communities. The most important fact is that ecosystems are open systems. New members may move into such a system and may leave it. Their boundaries are not fixed; there are transitional zones (ecotones) between adjacent ecosystems, no sharp dividing lines.

Today, modern technical equipment and new statistical and mathematical methods for collecting and analyzing data from the field as well as from the laboratories lead to a growing knowledge about ecological relations (e.g., competition, stress reaction, plant–animal interaction) and prove that the ecosystem is still the most valuable conceptional framework in biogeography.

3. *Aspects of Regional Biogeography: Classification and Ordination*

The classification of natural regions has long been an important objective for physical geographers. Decisive criteria for these classifications at a broad scale are climate, soils, and vegetation and their zonal distribution. Biogeography early became the discipline that was concerned with the large world vegetation types. The biosphere was divided into three-dimensional bioms or the large vegetation zones closely related to climate such as 'tropical rain forests' or 'boreal coniferous forests.' Biologists sometimes followed another hierarchical concept that they related to spatial scales, too: the concept of individuals, species, populations, communities, and ecosystems. Here, concepts are proposed, which are commonly accepted and widely used, bearing in mind that the chosen scale is largely responsible for interpretation and explanation of spatial patterns of organisms.

Since the nineteenth century, botanists and geographers contributed to a regional classification that is based on the distribution of species, a so-called floristic classification. The distribution area of a species and of others, which occupy the same or nearly the same area, is regarded as a biogeographical unit. Such a unit may be called, e.g., 'alpine,' 'Atlantic,' 'sub-Mediterranean.' In a hierarchical order these units (area types) form provinces, regions, and finally plant kingdoms, the Holarctic, palaeotropic, Neotropic, Australian, capensian, and Antarctic plant kingdoms. The plant kingdoms, especially, reflect the geological development more than the recent ecological background.

Plant sociology produced a classification on the basis of the composition of plant species, too, but these are ecologically clearly related to one another. The basic unit is the plant association. Species living together in these associations need more or less the same light, temperature, water, and nutrient supply and arrange themselves according to their competitive ability. Associations are characterized by a significant species group combination and character, and differential species, respectively. According to the floristic composition and the environmental conditions, associations are grouped to higher elements in the plant sociological hierarchy, e.g., to alliances and to classes. Examples of the latter are the rock, grassland, or swamp communities of Central Europe.

Physiognomy is the keyword for a vegetation classification that is based on plant life forms. Here, morphological structural characteristics of plants are more important than their taxonomy. Life forms—annuals, perennials, xerophytes, mesophytes—or the components of the traditional concept with phanerophytes, chamaephytes, hemikryptophytes, geophytes, and therophytes, have a clear ecological background which is proved by convergent traits, e.g., succulence, hairy leaves, thin bark. This approach is particularly useful because it facilitates the identification of functional relations between plants and animals. The vegetation units with a characteristic composition of life forms are called formations, such as deserts, forests, and grasslands, dominated by certain life forms (and not species) such as therophytes, grasses, or trees. The advantage of such a classification is obviously the possibility to compare vegetation units under the same macroclimatological conditions from continent to continent, without looking at species composition, and to understand better the functional meaning of morphological plant and vegetation structures, for example, leaf characteristics or stratification.

There is at least one complete method other than the traditional classifications of vegetation as a means for better relating and understanding plant species and their environment: the ordination. Plant sociology and more recently sigma sociology (analyses of vegetation complexes) have mainly developed in Central Europe, ordination—following the continuum concept—like other numerical methods developed in North America. Classification adopts discontinuous models and presumes that discrete units exist. Ordination is a

model for continuous vegetation distribution and needs gradient analysis. The results of an ordination are not discrete units but distribution trends. Today, the character and sharpness of boundaries between vegetation units and the stochastic aspects of plant and animal distribution are intensely discussed.

4. General Rules and Models in Biogeography

Biogeography is not only concerned with historical or ecological questions in particular studies but also with general rules and models, which provide a general framework of understanding patterns and processes, that may later be used for prognostic purposes. One such long-standing generalization is Buffon's law, which says that different regions of the earth that have the same ecological conditions are nevertheless inhabited by different species of plants and animals. Certain biogeographical rules are well known, such as Bergmann's rule, which says that animal species from the same genus are bigger in high latitudes than in lower ones. Allen's rule points out that for the same examples, the extremities in colder regions are smaller than in warmer. And according to Averill's rule, the wings of birds in higher latitudes are narrower than those of birds in lower latitudes. More recently, general rules like the latitudinal diversity gradient or models like the Sloss model ('single large or several small' areas for natural conservation) are discussed.

A special field of research has been dedicated to island biogeography. Life on islands played a vital part in Darwin's ideas about natural selection and adaptive radiation of birds. More recently, a model or theory of island biogeography has been proposed by McArthur and Wilson (1963). This theory says that on each island a state of dynamic equilibrium between species immigration and extinction exists and that the present species number depends largely on island size and distance of the island to the next continent or nearby island. The theory of island biogeography has been very stimulating for research and further modeling, although the results of many studies did not support the predicted relationships. The individual ecological abilities and needs of plant and animal species have been largely neglected as well as the fact that, for example vegetation development is not static and that it is important to choose the right moment for immigration in order to compete successfully with other plant species. Just counting species and calculating species turnover seem to be simple. To reach the main objective of biogeography, the understanding of pattern and processes requires a more comprehensive approach. For such an approach, the existing knowledge and the availability of sound data, however, does not justify mathematical sophistication so much today. It is more important to strengthen the empirical basis.

5. Applied Biogeography

During recent decades there has been a growing awareness of human impact on ecosystems and especially on vegetation. Man not only altered but also often destroyed the plant cover; his activities also had many negative consequences for site conditions. Looking for solutions to restore the plant cover and re-stabilize the whole ecosystem is one example for research in applied biogeography. In this case, plants or the vegetation cover as a whole are regarded as an important ecosystem element. If this element has been destroyed or altered, consequences arise for many other elements of the system, e.g., the soil properties or microclimatological conditions. The most important fields of applied research in this respect are almost all types of environmental degradation (e.g., soil erosion, desertification). Detailed afforestation planning—not necessarily with exotic tree species—can be a result of this research.

Another field for applied research in biogeography is linked with the idea that all biota are humankind's most valuable resource on earth. There is an urgent need to develop plans for a sustainable use of these renewable or non-renewable resources in agriculture, including range management and forestry. Furthermore, resources are both biomass and genetic resources. The topics dealt with thus relate to either land use or protection of plants and animals; and maintenance of biological diversity. Biogeographers contributed a lot, especially in natural conservation planning of reserves of the biosphere as well as of smaller areas of protection, to develop not only special management concepts but also general rules. In this context, the study of rare plants is as important as that of introduced and invasive species.

A last important subject of applied biogeography concerns bioindication. Plants, animals, and microorganisms can be regarded as simple instruments, which show by their mere presence or absence a certain environmental quality or the stress any kind of contamination has on a certain environment. Organisms as indicators may have advantages over chemical analysis, because they indicate more average conditions. Bioindication is already very useful to characterize more or less natural conditions of ecosystems. It has, nevertheless, growing importance in cases of environmental pollution in industrial areas, towns, and in aquatic ecosystems. And bioindication queried earlier—once the chosen organisms have been calibrated with the polluting substances—is certainly a valuable instrument in those countries where more sophisticated data collecting is too expensive.

See also: Ecology and Health; Ecology, Cultural; Ethnobiology; Human Development, Bioecological Theory of

Bibliography

Blondell J 1979 *Biogéographie et Ecologie*. Masson, Paris
Cox C B, Moore P D 1993 *Biogeography: An Ecological and Evolutionary Approach*, 5th edn. Blackwell, Oxford, UK
Ellenberg H 1988 *Vegetation Ecology of Central Europe*, 5th edn. Cambridge University Press, Cambridge, UK
Hallam A 1994 *An Outline of Pjanerozoic Biogeography*. Oxford University Press, Oxford, UK
Hengefeld R 1990 *Dynamic Biogeography*. Cambridge University Press, Cambridge, UK
Müller P 1980 *Biogeographie*. Ulmer, Stuttgart, Germany
Pielou E C 1979 *Biogeography*. Wiley & Sons, New York
Sedlag U, Weinert E 1987 *Biogeographie, Artbildung, Evolution*. Gustav Fischer, Stuttgart, Germany
Simmons I G 1979 *Biogeography, Natural & Cultural*. E. Arnold, London
Tivy J 1979 *Biogeography. A Study of Plants in the Ecosphere*. Oliver and Boyd, Edinburgh, UK

K. Muller-Hohenstein

Biographical Methodology: Psychological Perspectives

Lytton Strachey, the eminent artist of biography, once wryly observed: 'it is perhaps as difficult to write a good life as to live one.' Expressing the same opinion more forcefully, Virginia Woolf remarked of biographical art: 'Writing lives is the devil' (quoted in Edel 1984, p. 17). Scientific counterparts express similar ambivalence toward the biographical enterprise (Mc Adams and West 1997). Indeed, many would concur that 'near anarchy' captures the current state of narrative analysis, including that of biography.

Ironically, the greatest ambivalence may characterize personologists, a subgroup of personality psychologists, who draw systematically from psychological theory and research to write lives. Most personologists continue to endorse the core belief that the 'real business of psychology' concerns the long unit—that is, 'the life cycle of a single individual' (Murray et al. 1938, p. 39)—in principle.

The call has sounded yet again to resume and refine the personological tradition of Henry Murray in practice (McAdams and West 1997). Although personology has progressed considerably since 1990 (Craik 1997, Runyan 1997), scientific writing of individual lives has not kept pace (e.g., McAdams and West 1997, Nasby and Read 1997). Shoda and Mischel (1996) astutely note: 'generally not until the end of papers [do] researchers express the conviction that what we really need now is to return to the basic mission of the field, and to return to the commitment to understand intra-individual functioning and dynamics' (1996, pp. 414–15).

The personological study of life narratives has simply not grown to the potential that methodological and theoretical advances permit. We now hear more and more about the biographical study of individual lives, but only a handful of personologists have actually heeded the call to produce scientific biographies of particular individuals (e.g., Nasby and Read 1997).

A nagging irony arises here, because understanding individual lives constitutes one of the four tasks of personality psychology (Runyan 1997). Nevertheless, the field has not witnessed the optimal use of the scientific biography (e.g., McAdams and West 1997). Despite exhortations to write lives, ambivalence continues to stifle implementation of scientific biography (McAdams and West 1997, Rosenwald 1988). Why, then, does the slim and unacceptable ratio of heat to light persist?

A partial answer simply reflects the fact that academic training to pursue biographical science remains limited (Nasby and Read 1997), which almost further begs the question. Similarly, personality psychology receives a less-than-generous share of available funding, personology even less. Not surprisingly, therefore, the field of personality continues to emphasize individual and group differences, where the limited funds disproportionately flow. Moreover, detailed study of an individual life poses daunting challenges, technical as well as conceptual.

Considering the specific challenges to writing a life from the perspective of personology may also clarify the impediments that other disciplines confront when conducting biography from a scientific vantage point. The potential generalizability of the current analysis owes much to the strategic location of personality (Mayer 1998) vis-à-vis adjoining disciplines (e.g., anthropology, history, psychological subfields other than personology, sociology, and even neurology). Insights from personology may also elucidate the difficulties of biographical endeavors that follow the artistic, rather than scientific, path. A detailed characterization and analysis of the problem follows.

More generally, the problem consists of multiple elements, running the gamut from data, theory, and methodology, to the unit of analysis and the reward structure of academic institutions. However accurate, colorful allusions to demonic impediment and 'near anarchy' serve to identify, but not overcome, the problem. The remainder of this article, therefore, examines the nature and scope of each challenge more extensively. Ultimately, articulating how best to write a life first requires careful examination of the problematic details—where, after all, the devil resides.

1. Nature and Scope of the Problem

1.1 Data

When studying an individual life, data often pose the central difficulty. The challenges of data can prevent a

biographical investigation from exiting the starting gate. Most agree that the single-case study requires a wealth and variety of data, but investigators must typically confront inaccessibility of subjects. To complete the task, psychobiographers often can only consult archival material; the investigator cannot obtain data through interviews or assessments, essentially reducing the project to psychobiography. Given inadequate data, production of psychobiography almost inevitably falls prey to projection and other varieties of countertransference. The consequences of inadequate data partially explain the multiplicity of embarrassing work that plagues the genre.

Biographies that permit creative investigation *in vivo* typically include clinical cases that suffer from other limitations, most notably a pathological focus. Furthermore, studying a life ideally means conducting an investigation over time, which poses practical difficulties that intimidate all too many. Often, investigators come no closer to the ideal than studying college sophomores over a semester.

Recent developments, however, illustrate that the task of gathering adequate data, although difficult, need no longer derail a biographical project. For example, personologists have outlined guidelines according to which a biographer can extract case data from narrative sources and reveal the underlying order therein. Personologists have also profitably applied coding schemes to analyze the content of narrative material. For example, personologists have devised coding systems that yield quantitative measures of important motives, including intimacy, as well as achievement, power, and affiliation-intimacy, and the broader concerns of identity, intimacy, and generativity. One may also reliably evaluate affect or affective tone through ratings of narrative material.

Applying each of the aforementioned techniques, Nasby and Read (1997) reported an integrative case study of the solo circumnavigator, Dodge Morgan. In addition, the investigators applied concomitant time series analysis (CTSA) to the quantitative measures of motives, broader concerns, and affect as well as performance measures of daily progress. CTSA permitted detection, modeling, and removal of statistical artifacts (long-term trends, cycles, and serial dependencies) from each variable over time. Once decomposed, accurate calculation of cross-correlation functions that assessed synchronous and lagged relations between variables occurred, which permitted valid statistical tests of hypotheses about the circumnavigator's functioning throughout the life-defining event of the voyage.

Similarly, Simonton (1998) investigated 'Mad' King George of England, first performing content analyses of the historical record to obtain quantitative measures of stress and health, and then decomposing each series before finally calculating the cross-correlations (synchronous and lagged) between the multiple indices of stress and health. Of considerable importance, the 'historiometric' approach illustrates that a biographer can often derive quantitative indices from the qualitative or narrative sources of information that dominate historical records, and then apply sophisticated statistical techniques, including but not restricted to CTSA, to test explicit hypotheses about historical figures.

1.2 Theory

Many psychologists have now accepted the premise that clarification and elaboration of theory best justifies conducting the case study (and, through extension, the biography). A principal responsibility of psychologists entails generating and testing increasingly powerful theoretical insights. Although theory development and testing may require sustained activities that extend beyond the confines of the case study or the biography, either proves well suited to clarify theory.

The detail and depth of the biography and the case study do much to promote understanding of theoretical foundations. Indeed, the application of theory to the individual means that an individual life becomes text that one must interpret. Inevitably, any comprehensive fund of biographical data will reveal discrepancies from theoretical expectations. One must not avoid or discard inconsistencies, which, disturbingly, occurs often. Effectively, the latter practice distorts evidence and wastes information. Instead, discrepancies and inconsistencies provide opportunities—not only to locate where the theoretical foundation contains cracks, but also to integrate through a higher-order synthesis or dialectic.

More important, but rarely appreciated among biographers, the dividends of applying multiple theories to the same case deserve far more attention. For example, Nasby and Read (1997) illustrated the process of applying two diverse theoretical perspectives—the Five Factor and Life Story Models—to multiple sources of data, qualitative as well as quantitative. In addition, Wiggins (1997), who provided commentary, reinterpreted some of the case data from a third perspective—that of the Interpersonal Circumplex Model. The multiple theories served the role of multiple methods (McAdams and West 1997), each providing a unique perspective from which to construe data from the case. Considering the three accounts jointly, a complex and coherent portrait of the circumnavigator emerged, one far richer than would have resulted from the application of a single perspective.

In addition, applying multiple theories provided explicit bases of deriving hypotheses, some complementary, others contradictory. The approach forced the researchers to consider the full range of available

evidence supporting or refuting each hypothesis, 'rather than foreclosing on a conceptually pleasing but incomplete or possibly erroneous understanding' (Mc Adams and West 1997, p. 758).

1.3 Methodology

That no definitive methodology exists to design and conduct biographical research has led critics to conclude that such endeavors lack scientific rigor. Runyan (1997) has repeatedly and effectively, however, countered each criticism. From an alternative perspective, the richness of methodological choices simply creates distinctive technical challenges (e.g., Nasby and Read 1997). Any biographer can misuse any method; but the expanded degrees of freedom that ensue from the complexity of the endeavor and the flexibility of methods gives single-case studies, including biographies, 'probing and testing power' that can validly disclose unanticipated effects that more structured approaches might very well miss.

Moreover, one can recast the methodological concerns to reveal strengths of the biographical approach. Through integration, the methodological concerns point to utilizing multiple sources of data and multiple strategies of measurement, multi-operationalizing multiple constructs, analyzing the data across varying levels of abstraction, and including both qualitative and quantitative methods (e.g., McAdams and West 1997, Nasby and Read 1997).

2. The Long Unit of Analysis: The Whole Person Over a Lifetime

The underachievement that characterizes the personological agenda of studying the whole person ensues also from improperly appreciating the implications of adopting the *long unit* (Murray et al. 1938). Confusing the two meanings of the rallying cry, one methodological, one conceptual, may have hindered identification of the source of difficulty. The methodological meaning concerns the comprehensive assessment of personality. Few doubt that comprehensive assessment should remain the aspiration of the personologist, even though '[c]apturing the 'whole personality' remains beyond our grasp' (Craik 1997, p. 1095).

The trouble can be accorded to the conceptual meaning. Allport (1937) observed a long while ago that (p. 343):

> [Holistic conceptions of personality] do little more than assert personality to be an 'Indivisible Whole,' 'a total integrated pattern of behavior,' … Personality … is like a symphony. Granted; but does not the comprehension of symphonic unity come only though an understanding of the articulate weaving of motifs, movements, bridge-passages, modulation, contrasts, and codas? Nothing but empty and vague adjectives can be used to characterize the work as a whole. If a totality is not articulated, it is likely to be an incomprehensible blur; it can then be extolled, but not understood.

No less germane today, the critique that Allport (1937) marshaled suggests a solution. Craik (1997), for example, refers to the current revival of the personality system framework (Mayer 1998, Nasby and Read 1997, Shoda and Mischel 1996). Unlike previous efforts, the current revival appreciates that one cannot study the undifferentiated whole. Studying a complex system (e.g., a person) first requires systematically engaging the task of differentiation, a task that specifies both the components of the system and the organization of the components.

Mayer (1998) has considerably advanced the initial and necessary process of differentiation, locating personality amongst adjoining systems (biological, situational, sociological). More specifically, Mayer presents a structural model that not only specifies the components of a personality system, but also accommodates all four tasks of personality psychology (Runyan 1997): general theory, individual and group differences, specific processes and behavior, and individual persons and lives. The structural model achieves inclusive and integrative power through a variety of means, which bridge data, theory, and methodology.

The structural model offers a preliminary specification of how systemic components relate to one another statically. Refining the model, however, will require a treatment of systemic organization from a dynamic perspective (e.g., external control, distributed control, hierarchical control). Articulating how the components function jointly over time awaits further efforts. My colleagues and I have initiated development of a metaperspective that takes us beyond the metaphor of a systems framework to a formal (mathematical) implementation. Interestingly, formalizing a dynamic systems model returns personology full circle to the long unit (Murray et al. 1938): the preferred means of examining and testing a dynamic model tracks the functioning of an individual system (person) over time.

3. Reward and Value Structure of Academic Institutions

The technical challenges of data, theory, and methodology as well as the conceptual obstacles of espousing and pursuing the long unit of 'the whole person' share a common denominator. Each identifies an internal source of impediment to writing a life. Internal factors explain the difficulties of optimally pursuing artistic biography. Beyond internal impediments, however, the institutional context of pursuing biography scientifically presents additional hurdles.

Personality psychologists continue to pay lip service to the personological agenda in principle, and avoid accepting the challenge in practice. Rosenwald (1988)

remarked, 'when tendered an invitation to examine any life in detail, the psychologist usually declines … the case study has come to be regarded as a scholar's trap' (p. 239). Along parallel lines, Carlson (1971) chided much earlier that: 'Personality psychology would seem to be paying an exorbitant price in potential knowledge for the security afforded by preserving norms of convenience and methodological orthodoxy' (p. 207). The situation has not appreciably changed.

Summarizing the current state of personality research, Endler and Speer (1998) concluded that (p. 667):

> Success in an academic climate seems to involve the adoption of methods aimed at publishing a maximum number of articles in a minimum amount of time … [T]he choice of research method in our field is more the product of the academic marketplace than of scientific considerations … The question remains as to whether these constraints have served to compromise the overall quality and scientific contributions of our discipline.

Institutional standards and values of academic reward set stringent criteria that emphasize quantity and speed. Here the internal factors and external factors converge (or collide). Accordingly, most psychologists skirt the 'scholar's trap' of the case study and the scientific biography, opting instead for cost- and time-efficient investigations of group and individual differences or specific processes and classes of behavior.

Realistically, the prospects of reforming university philosophy range between slim and none. A few, therefore, have sought to establish climates more hospitable to the personological agenda. McAdams, for example, has established the Foley Center for the Study of Lives, 'a mini-oasis of personology in the Midwest dust bowl.' (from a personal communication, March 7, 2000). The Foley Center, which exclusively pursues a personological agenda through the financial support of the Foley Family Foundation, nevertheless operates under the auspices of a traditional university. Therefore, the center does not replace the reward structure of mainstream academia. My colleagues and I have adopted a different tack, searching beyond the confines of traditional academia to establish, also through private funding, the California Personological Research Institute (CaPRI). The mission of CaPRI exclusively concerns the personological study of lives, and the reward structure matches the core purpose of the institution accordingly.

4. Conclusion

Through the current analysis, CaPRI has begun to develop a biographical/personological study, one that draws inclusive and integrative strength from the application of dynamic systems theory. The metaperspective may bring personology closer to the goal of more efficiently and satisfactorily addressing the technical challenges to scientific biography that pertain to data, theory, and methodology as well as the conceptual obstacles that surround the adoption of the long unit (Murray et al. 1938). Overcoming institutional obstacles may await the establishment of biographical or personological centers that follow ad hoc missions and reward investigators accordingly.

The current analysis may help to overcome some ambivalence among personologists, thereby generating more heat relative to light, and should extend beyond the boundaries of personology and adjoining disciplines to include biographical artistry. We may never remove the devil from either the art or the science of biography, but we may stand a better chance of overcoming the technical and conceptual impediments to writing a life well. Although the odds of living a good life will remain unchanged, the odds of writing one could improve substantially.

See also: Archival Methods; Biography and Society; Biography: Historical; Case Study: Logic; Case Study: Methods and Analysis; Case-oriented Research; Life Course: Sociological Aspects; Narratives and Accounts, in the Social and Behavioral Sciences; Single-subject Designs: Methodology

Bibliography

Allport G W 1937 *Personality*. Holt, New York

Carlson R 1971 Where is person in personality research? *Psychological Bulletin* **75**: 203–19

Craik K H 1997 Circumnavigating the personality as a whole: The challenges of integrative methodological pluralism. In: McAdams D P, West S G (eds.) *The Inner and Outer Voyages of a Solo Circumnavigator: An Integrative Case Study by William Nasby and Nancy W. Read. Journal of Personality* **65**(Special issue): 1087–111

Edel L 1984 *Writing Lives: Principia Biographica*. Norton, New York

Endler N S, Speer R L 1998 Personality psychology: Research trends for 1993–1995. *Journal of Personality* **66**: 621–69

Mayer J D 1998 A systems framework for the field of personality. *Psychological Inquiry* **9**: 118–44

McAdams D P, West S G 1997 Introduction: Personality psychology and the case study. In: McAdams D P, West S G (eds.) *The Inner and Outer Voyages of a Solo Circumnavigator: An Integrative Case Study by William Nasby and Nancy W. Read. Journal of Personality* **65**(Special issue): 757–83

Murray H A et al. 1938 *Explorations in Personality: A Clinical and Experimental Study of Fifty Men of College Age*. Oxford University Press, New York

Nasby W, Read N W 1997 The life voyage of a solocircumnavigator: Integrating theoretical and methodological perspectives. In: McAdams D P, West S G (eds.) *The Inner and Outer Voyages of a Solo Circumnavigator: An Integrative Case Study by William Nasby and Nancy W. Read. Journal of Personality* **65**(Special issue): 785–1068

Rosenwald G C 1988 A theory of multiple-case research. In: McAdams D, Ochberg R (eds.) *Psychobiography and Life Narratives*. *Journal of Personality* **56**(Special issue): 239–64

Runyan W M 1997 Studying lives: Psychobiography and the conceptual structure of personality psychology. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 41–69

Shoda Y, Mischel W 1996 Toward a unified, intra-individual dynamic conception of personality. *Journal of Research in Personality* **30**: 414–28

Simonton D K 1998 Mad King George: The impact of personal and political stress on mental and physical health. *Journal of Personality* **66**: 443–66

Wiggins J S 1997 Circumnavigating Dodge Morgan's interpersonal style. In: McAdams D P, West S G (eds.) *The Inner and Outer Voyages of a Solo Circumnavigator: An Integrative Case Study by William Nasby and Nancy W. Read*. *Journal of Personality* **65**(Special issue): 1069–86

W. Nasby

Biography and Society

In 1959 the American sociologist C. Wright Mills wrote in *The Sociological Imagination*: 'Social science deals with problems of biography, of history, and of their intersections within social structures,' adding that 'these three—biography, history, society—are the coordinate points of the proper study of man' (Mills 1959). Indeed in disciplines such as history or ethnography the association between 'biography' and 'society' is taken for granted. Neither historians nor ethnographers need to be reminded that the societies they study have an history which extends up to the present times; nor that each of their members is a 'singular universal' (Sartre 1960), differently embedded in social structures and historical time, but all participating in the constant remaking of their own society.

Sociology, however, as an *institution*, with shared beliefs handed down to younger cohorts through textbooks, and with norms of scientificity, seems to forget this again and again. Within the short history of empirical (sociological) research scholars have organized the more or less systematic production of autobiographies and more recently, of recorded life stories (the 'biographical method'). But the legitimacy of such date has been recurrently denied. Not so long ago, for example, Bourdieu (1986) argued that life history had been 'smuggled' from common sense into sociology. He was echoing a long line of earlier similar statements—the 'common sense' view of sociology as institution—denigrating life stories as subjective and therefore(?), unreliable, data. The dramatic changes of the biographical method's status in the history of empirical sociology cannot be understood without taking into account this background of diffuse hostility. We shall return to this point in the discussion below, after a recall of the main events of the biographical method's history.

1. *A Short History of the Biographical Method in Sociology*

In the history of sociology the use of autobiographies written on request coincides with the beginnings of empirical research at the University of Chicago. The authors of the first major work, *The Polish Peasant in Europe and America* (Thomas and Znaniecki 1918–1920), went as far as stating that the life history provides the best kind of sociological material. Indeed its greatest virtue is to provide access to and knowledge about how people in a given situation; experience that situation (perceive it, evaluate it, feel about it) within the totality of their own life experience; and how they react, interact and act in its context.

The life history allows the expression of lived experience, of the meanings it takes (in retrospect) for the person; of the values and attitudes that shape such subjective meanings; of the affects that go with it; of motives for action. Such an 'insider's view' is rich in hints, clues and insights to the outsider (the researcher) who is trying to understand situations and processes, social milieus and their inner dynamics. While Thomas and Znaniecki used mostly sets of letters exchanged between emigrants and their families back home, they also commissioned an extensive autobiography of a young Polish emigrant to Chicago; finding it so interesting that they published it as one of the four volumes of their work, commenting it by way of numerous footnotes.

While this autobiography itself reads as well today as it did in the 1920s, Thomas and Znaniecki's footnotes, however, sound dated. The authors' sociological theorizing of the processes of cultural disorganization and reorganization which necessarily accompany flows of emigration/immigration were remarkably innovative and proved very influential; but in retrospect they seem to stem much more from the authors' general knowledge of such processes than from the personal documents they used.

In the 1920s and 1930s several Chicago scholars trained by Robert E. Park, who had taken over from Thomas, produced a stream of very valuable studies on various marginal groups living in Chicago. Some of the best ones used autobiographies as one of their main tools (see especially Shaw 1927; for an historical account see Bennett 1981). However a consistent methodological paradigm did not emerge until Lindesmith's (1947) study of heroin addicts. By then it was too late. A new method, the survey, was raising considerable hopes, monopolizing attention, scientific legitimacy and research funds. So-called (by contrast)

'qualitative' methods, including case histories were marginalized. If part of the spirit of the Chicago School survived in symbolic interactionism it was with a focus on direct observation of interaction, which also meant losing sight of the historical, societal and biographical roots of social action.

While after World War II the life history was disappearing from the scene of American sociology, a major work was in the making in American anthropology: Oscar Lewis's *The Children of Sanchez* (1961). Before Lewis, American anthropology had already generated a long list of published life histories, some of them great classics (such as Simmons's *Sun Chief* 1942). To collect the information, however, researchers had either been taking notes while interviewing, or—in Simmons's case—had asked a literate Indian, Don Talayesva, to write his own autobiography. The technological innovation of the tape recorder now allowed researchers to record verbatim the speech of persons who could not write, thus opening immense possibilities.

Lewis was the first to fully use the new technique. Indeed *The Children of Sanchez* became an instant best-seller worldwide. The body of the book is made of the life stories of four siblings who are growing up in a deprived area of Mexico City. Their young lives are already full of dramas and emotion, of adventures and experiences, which they narrate as if they were all born storytellers. Middle-class readers all over the world felt that for the first time they were listening directly to the fresh voices of persons they would normally never meet; persons very distant in geographical and social terms, but who appeared very close psychologically. Reading this book in Paris in 1962, de Beauvoir (1965) (see *Beauvoir, Simone de (1908–86)*) wondered aloud whether there was still some room for fiction writers, since now the characters of Balzac, Dickens, and Zola could speak by themselves. (It must be added, however, that the vividness of *The Children of Sanchez's* narratives owes much, perhaps not in content but certainly in form, to the highly skilled rewriting of Lewis.)

It is in Western Europe that the life history reappeared, partly through the influence of Lewis's volume, in the 1970s. The Italian scholar Franco Ferrarotti was one of the first postwar sociologists to realize that what ordinary people have to say on the basis of their own life experience is worth listening to. His *Storia e storie di vita* (History and life histories) is an all-out effort, deeply inspired by Sartre's humanism, to identify what he calls the *specificity* of the biographical method (Ferrarotti 1981; see also his paper in Bertaux 1981).

In France Catani (1982) inspired by the anthropologist Louis Dumont, was the first to try using one single life history to capture the whole 'cultural model' of a given culture. This was nothing less than the French cultural model, which Catani, as an Italian, was looking at from the outside. His *Tante Suzanne* contains the verbatim transcription of the six interviews through which an aging woman, Suzanne Mazé, tells him her life story. Born in a rural environment, she had worked most of her life as a worker in small shops, had married an artisan and retired as a petty-bourgeois: a rather typical trajectory for her generation. Catani, a precursor of current 'narrativism,' analyses both content and form of her interviews: the *content* yields her core values, which can plausibly be extended to the social milieus of which she had been part. The recurrent *forms* her narrative takes up give access to another kind of implicit meanings. The latter may also be embodied in objects, not words: for instance in her garden Suzanne has purposefully planted the 'same' trees as those which adorned the castle's park where her mother worked as servant when herself was a child.

Another important work is by Abdel Malek Sayad (1991) on the emigration of Kabyles from Algerian villages to French cities in the 1950s and 1960s. Sayad interviewed dozens of such emigrants and published some of their life stories, which—after translation and no doubt some rewriting, as in Lewis's case–are so self-expressive they hardly need commentary. His analysis addresses other levels of reality including directly collective ones, such as the complex chain of myths, dreams, disillusions, and lies through which emigration becomes a self-sustaining process; the secret complicity of French and Algerian governments in organizing such migration flows to the benefit of both States; the collective dilemma of living between two societies while belonging to neither; and the silent drama of ageing emigrants when realizing that their children, growing up in France, are becoming alien to them. While resting on a limited number of cases, Sayad's sociologically dense descriptions of the processes accompanying emigration sound very convincing, all the more so because he was himself part and parcel of the process he was studying.

Bertaux's work exemplifies a third style of research using life stories, the one focusing on a social world: in his case French artisanal bakery (Bertaux and Bertaux-Wiame 1981). His initial focus on the structural relationships of production between artisans and bakers, on their consequences on everyday life and his subjects' respective life chances, came from Marx and Bourdieu; but his decision to collect life stories of aging or retired bakery workers, of bakers and bakers' wives, was inspired by *The Children of Sanchez*. This unlikely combination resulted in the discovery of the hidden mechanisms through which the supposedly obsolete artisanal form of production in France had successfully resisted the recurrent attempts by industrial bakeries to take over the bread market (Bertaux and Bertaux-Wiame 1981). Coupled with data such as historical documents, juridical regulations, and (scarce) statistics, the life stories gave a historical depth and biographical thickness to the sociological inquiry. Also the recurrence, from one life story to the

next, of descriptions of the same situations and actions led the authors to identify, via the phenomenon of 'saturation' emerging after a limited number of case studies, a methodological principle allowing them to put forward plausible generalizations. Saturation has thus become, for case studies, the equivalent of the representative sample for survey research.

Bertaux also took the initiative to organize an *ad hoc* group on life history within the program of the 1978 World Congress of Sociology in Uppsala; a group to which an unexpectedly high number of sociologists turned up. This eventually led to the creation of the Biography and Society Research committee within the International Sociological Association and contributed to the return of life history as a legitimate sociological method. The 20 years that followed have seen a flowering of research projects using this method, particularly in Germany: the German Sociological Association has a *Biographieforschung* Research Committee. Some of its regular debates are devoted to the study of life course, focusing on the issue–framed by Martin Kohli—of whether 'biography' is presently becoming more institutionalized, or less (See *Drugs: International Aspects*; *Age, Sociology of*; *Disability: Sociological Aspects*; *Age Policy*).

For a reflexive overview of recent trends in Germany see Apitzsch and Inowlocki 2000. The case history of whole families, an extension of the biographical method, has also been used by some European scholars in researching specific social mobility processes—for example, on what happened to Russian families expropriated by the October Revolution, or on the very differing fates of descendants of poor peasants and craftsmen in three neighbouring villages in Tuscany: the aim of such research projects being to show that the study of social mobility, long the exclusive province of survey research, can benefit from research based on case histories (Bertaux and Thompson 1997).

One recent trend in Europe is the comparative study across Western Europe of poverty and precariousness by means of case studies and case histories (Chamberlayne et al. 2000, chap. 18). Together with the very dynamic field of European comparative research on Welfare States (see *Welfare*), such studies provide 'views from below.' They usefully complement the statistical surveys of poverty research, which prove inadequate in capturing the complexity of local situations and, especially, the survival tactics and strategies of individuals and households confronted with the precariousness of their conditions of life.

2. Discussion

Contemporary sociology as an institution manifests a reluctance to grant full status to the use of life stories. Why is it so? A short semantic analysis of the expression 'Biography and Society' allows to get at the root of the issue. When the 'revival' of the biographical method took place in the early 1980s, sociology as institution was organized around the hegemony of survey research and, more generally, of a self-conception which borrowed its core elements from the natural sciences and the hypothetico-deductive method (the considerable efficiency of survey research is not to be denied, but the structural similarity between sociology and natural sciences is highly questionable, if only because the controlled experiment, the core method of natural sciences, is out of question for sociology). At the 1978 World Congress of Sociology, an *ad hoc* group on 'the individual and society' would have attracted few sociologists and aroused no expectations. But 'life history' brought to the fore elements which had been forgotten or repressed by the 'natural science' conception of sociology. While 'the individual' is a generic term, 'life history' denotes the *singularity* of destinies: 'biography' in that sense is the biography of one given singular person who exists somewhere in time and space, a person with a gender and embedded in a context.

Biography also immediately refers to time, both biographical time and historical time, and all levels of intermediate temporalities. Biography also speaks, in Sartre's words (1960), of 'what people do of what has been done of them,' hence of action (which radically distinguishes the world of human affairs from the world of inanimate objects). Finally biography includes the sphere of meaning, which Weber has shown to be constitutive of action. Singularity, historicity, action, and meaning (not to mention emotions and feelings) are part of social-historical reality, as Dilthey recognized (and theorized about) long ago; but they do not fit with the view of sociology as a natural science in the making, the view which has long been the core legitimacy resource of the institution of sociology.

See also: Life Course in History; Life Course: Sociological Aspects; Macrosociology–Microsociology; Mobility: Social; Traditions in Sociology

Bibliography

Apitzsch U, Inowlocki L 2000 Biographical analysis: a 'German' school? In: Chamberlayne P, Bornat J, Wengraf T (eds.) *The Turn to Biographical Methods in Social Science. Comparative Studies and Examples*. Routledge, London, pp. 53–70

Beauvoir S de et al 1965 *Que peut la littérature*? 10/18, Paris

Bennett J 1981 *Oral History and Delinquency. The Rhetoric of Criminology*. University of Chicago Press, Chicago

Bertaux D (ed.) 1981 *Biography and Society: The Life History Approach in the Social Sciences*. Sage, London

Bertaux D, Bertaux-Wiame I 1981 Artisanal bakery in France: How it lives and why it survives. In: Bechhofer F, Elliott B (eds.) *The Petite Bourgeoisie. Comparative Studies of the Uneasy Stratum*. Macmillan, London

Bertaux D, Thompson P 1997 *Pathways to Social Class. A Qualitative Approach to Social Mobility*. Clarendon Press, Oxford, UK
Bourdieu P 1986 L'illusion biographique. *Actes de la Recherche en Sciences Sociales*. **62/63**: 69–72
Catani M, Mazé S 1982 *Tante Suzanne*. Librairie des Méridiens-Klinsieck, Paris
Chamberlayne P, Bornat J, Wengraf T 2000 *The Turn to Biographical Methods in Social Science. Comparative Studies and Examples*. Routledge, London
Ferrarotti F 1981 *Stori e storie di vita*. Laterra, Basi, Roma
Lewis O 1961 *The Children of Sanchez: Autobiography of a Mexican Family*. Random House, New York
Lindesmith A 1947 *Opiate Addiction*. Indiana University Press, Bloomington, IN
Mills C W 1961 *The Sociological Imagination*. Oxford University Press, New York
Rosenthal G (ed.) 1998 *The Holocaust in Three Generations: Families of Victims and Perpetrators of the Nazi Regime*. Cassell, London
Sartre J P 1960 *Questions de méthode*. Gallimard, Paris
Sayad A 1991 *L'immigration ou les paradoxes de l'altérité*. De Boeck, Brussels, Belgium
Shaw C 1930 *The Jack Roller: a Delinquent Boy's Own Story*. University of Chicago Press, Chicago
Simons L (ed.) 1942 *Sun Chief: Autobiography of a Hopi Indian*. Yale University Press,
Thomas W I, Znaniecki F (1918–1920) reprinted 1958 *The Polish Peasant in Europe and America*. Dover, New York

D. Bertaux

Biography: Historical

1. Definition and Parameters

Historical biography (from the Greek *grafein*: writing; *bios*: life; *historia*: inquiry, knowledge) is an individual, historical life in writing, written by someone from a later era. The historical biography as scholarly biography manifests a somewhat contradictory predicament inasmuch as biography places the individual at the center, whereas history focuses on common structures and events. Historical biography is a genre which is characterized by variety and diversity, both in historical outlook and methodology. This article will present the Western, historical biography. In terms of genre, historical biography verges on autobiography, literary biography, the traditional tale, and the biographical novel. In specialist professional terms, the historical biography verges on psychology, sociology, anthropology, history, and literature. In terms of subject, the historical biography is open to everyone and anyone, statesmen and farmers, generals and artists, philosophers and scientists, heroes and villains, women and men. As regards time-scale, historical biography has existed since the first century AD and up to the threshold of the second millennium. All this makes the genre both extremely complex and deeply fascinating. Historical biography diverges from popular biography's interest in the private lives of the famous, but shares the popular genre's interest for getting behind the myth. It diverges from fiction, but shares genre with the modern novel, which has long since broken with linear narrative in recognition that the passage of real life is fragmented.

2. Two Thousand Years of the Historical Biography: Changing Focus

Since antiquity Western historical biography has been regarded as a branch of historiography and has developed from being an ethical-humanistic genre to being a genre consisting of various methodologies, forms, and styles within twentieth-century specialized scholarship. As a genre, biography in the West is considered to have been established by the Greek Plutarch, who published the comparative lives of Greek and Roman statesmen *Bioi Parallelloi* (AD 125). This work, together with Tacitus' *De vita Agricolae* (AD 98) and Diogenes Laërtius' biographies of Greek philosophers (3rd century AD), characterized what has been called *the classical biography*, built upon the fundamental principles of ethics: the central figures were either commended for having fulfilled their duty or censured for falling into the trap of ambition or arrogance. The classical biography-type from antiquity was maintained throughout the Middle Ages. A sidelong consideration must here be given to China's great historian Sima Qian, who, during the period of classical antiquity, developed a biographical form which belongs to a much later Western epoch: *Shiji* (185–45 BC). This work contains biographies, not just of eminent statesmen and soldiers, but also of individuals such as fortune-tellers, courtiers, and murderers. A completely modern approach, in today's terms, was demonstrated by Sima Qian's contemporary, the poet and literary historian Liu Xiang, whose work *Lienü zhuan* (79–78 BC) shows that female personalities in China were already at this stage considered worthy of biographical study.

In accordance with the requirements of the church and spiritual needs, the Middle Ages saw the development of the martyr biography (e.g. John Foxe's *Book of Martyrs*, 1563) or lives of holy men and women, known as hagiographies. The purpose was didactic and the central character was presented as a model of Christian propriety and public virtue: God's creation was portrayed as an individual who, in the course of a lifetime, developed as a moral example to others, or whose destiny was first fulfilled in death. This biographical model was also used for biographies of princes and emperors, with Einhard's *Vita Caroli*

Magni (829) frequently cited as the most important. But until twentieth-century medievalists began producing historical biographies, the Middle Ages essentially lay in biographical darkness. History, seen as the contemporary environment, was of no interest in the Middle Ages, unless contributing to an insight into the central figure's moral attributes.

The approach to historical biography changed during the Italian Renaissance, a change which continued further afield from the seventeenth up to the nineteenth century. This type of biography has been called *romantic-linear development history* and developed from Boccaccio's sense for the specific in *Vita di Dante Alighieri* (1354–5) to a glorification of brilliant individualities in connection with the much later liberal individualism in society. Biographical literature was so extensive in the nineteenth century that E. M. Ottinger's (1854) bibliography of biographical literature was only rudimentary, even though it was wide-ranging. From the turn of that century, James Boswell's *Life of Samuel Johnson* (1791) is still singled out as the pioneering biography of the era, as it was based on empirical material in the form of letters, private papers, conversations, interviews, and personal observation of Samuel Johnson's comportment. This was the first biography to construct a nuanced, candid personality. In the historical biography of the time, the history of human achievement was fundamentally the story of imposing male heroes (e.g., Thomas Carlyle 1841). It was still the exceptional person and the developmental process which characterized the 'life-and-letters' biography at the end of the nineteenth century (e.g., Dilthey 1870, Grimm 1873). Even though ideas from the community at large were incorporated into the biographies via reference to letters, there was no trace of historical reflection on the individual and society.

The modern life story, as twentieth-century historical biography has been called, came about as the result of the crisis for humanism, Christianity, and rationalism which followed in the wake of Darwinism and psychoanalysis as it developed during the twentieth century. This category of biography displays the following characteristics: (a) a multitude of biographical methodologies evolved; (b) the number of published historical biographies increased; (c) historical–biographical literature as a genre became broader, with the emergence of the 'life-and-times' biography in which the individual is presented in the light of contemporary society; (d) male heads of state were, with a few exceptions, the focus of interest throughout the century; (e) biographical critique and the methodology debate intensified at regular intervals, but were not carried out systematically; (f) a new aspect was incorporated into the genre—demythologization, unmasking; (g) historical biography was influenced by both academics and artists. Serious biographical literature thus covered the entire spectrum from not very accessible treatises to stylistically inviting reading.

The sociological and historical professions regarded the biography critically throughout the century. As early as 1898, G. V. Plekhanov published a short tract in which he called to account the perception of the imposing personality as being the only driving force of history and at the same time proposed the dynamics between personality and society as historical transformation factor (Plekhanov 1898). He maintained that an imposing personality only had historical significance if the person in question was possessed of attributes which were necessary at that specific point in history and which gave historical impact to this particular talent. This observation has had profound repercussions for historians' development of the historical biography. Sociologists have used historical biography as methodology in a qualitative analysis of the significance of subjective experience for social reality. Pioneers in this respect were F. Znaniecki and W. Thomas (1918–20) who presented the individual as both creative and created in social evolution. This was followed up theoretically by the Italian sociologist Franco Ferrarotti (1983) who saw the individual as an active pivot in respect of the structures and history of society. Life histories have also been found interesting by late twentieth-century anthropologists wishing to promote subjective experience as a factor in the creation of personal identity (e.g., Crapanzano 1980). Anthropologists have considered biography to be a participant observator in a crosscultural and cross-epochal dialogue with the historical subject. It was the mentalities-historians who influenced the historical biography, especially the French Annales School and its preoccupation with personification and its hermeneutic, mentalities-historical studies. In her innovative study of sixteenth-century *Martin Guerre* (1983), Natalie Zemon Davis used new ideas, concepts and methodologies from anthropology, ethnography, and literary criticism to examine the dilemma of truth and doubt in historical research. Perception of a central figure became the result of a communicative process between two cultures and two people, not an objective description, as Giovanni Levi (1989) has also expressed it. Jean-Paul Sartre's concept of interaction between the life lived and the written text inspired the development of the dialectic biography. In Denmark, the historical figure within this tradition is interpreted as both the bearer of a cultural convention and a cultural agitator (Possing 1992). Sociologists, ethnologists, and historians have all used reminiscence, the oral-history tradition, and private papers as empirical source material through which to understand social action in a wider perspective as a confrontation with the positivist, Marxist, and structuralist thinking which has starved the biography of life. The objective has been to get behind the mythologization of major figures: 'A modern biographer may or may not choose to reveal the intimate, the amorous details of a life, but he must, if he is good at what he does, probe beneath its public, polished self' (Pachter 1979).

The modern life story in literary history was introduced with Lytton Strachey's *Eminent Victorians* (1918), which in the twentieth century came to play the same revolutionary role for historical biographers of all professions as Boswell had for biography in the eighteenth and nineteenth centuries: the introduction of the artistic, interpretative biography, given form via selection, concentration, and interpretation of the sources. The idea was to get behind the myth of the subject of the biography and it became possible to write biographies of both men and women. Leon Edel (1957) published the most influential post-war study of the biography, which he honed throughout his life's work under the motto: 'A writer of lives is allowed the imagination of form but not of fact' (Edel 1984). Owing to, among other factors, the achievement of Edel, literary biography became a well-established genre in its own right. Within this genre, the post-modernists Ira Bruce Nadel (1984) and William H. Epstein (1987) have sought for a poetics of biography by pointing out that definitive life portraits are not to be found.

3. *The Intellectual Position of the Historical Biography on the Eve of the New Millennium: Quest and Ambiguity*

Historical-sociological biographers and literary biographers began to draw a little closer together after the 1930s. This drawing together has continued throughout the rest of the century: literary-framed biography has become more scholarly-framed and historical-, sociological-, anthropological-, and natural science-framed biography has become more literary-framed in recognition that the more objective a biography is, the more lifeless and hollow it becomes (Kendall 1965, Madélenat 1984). On the other hand, the psycho-historical biography has not really established a firm footing within the humanities and the social sciences, primarily because it is generally based on extremely insubstantial source matter.

In the light of a quarter-century's interest in social- and mentalities-history and historical narrative, the historical biography, with its focus on the individual above community, class, and social group, has undergone a renaissance. The role of historical biography as a point of orientation in human life could be one of several reasons for its popularity as a genre at a time when belief in the great utopias has disappeared: it has become the prism for a multitude of preoccupations. The 1990s have seen a desire for images of women and men as creative, reflective, doubting, and determining individuals; a desire to get away from the tendency to reduce specific singularity to social regularity. The French historian Giovanni Levi (1989) has expressed this need for open reflection by pointing out that, more than ever, the ambiguous biography is a central preoccupation for historians. The biography is a sanctuary for dramatic, old-fashioned narrative amidst a deconstructing and fragmented era. It facilitates reflection on the human ideal and the flawed, so that biography does not become eulogy.

The renaissance of biography is a manifestation of a genuine renewal of the genre, with its source in inspiration from advanced literature such as the works of Proust and Joyce, and in the critical question about the extent and manner in which history is created by people and how a life can be decoded. On the one hand, historical biography makes a justifiable demand to set the agenda via a rehumanization of the humanities, which means that passion, irrationality, and human idiosyncrasy are drawn into an understanding of an individual's life work. On the other hand, the genre's respectability is contested time and again, and it has not secured a serious, scholarly profile in those humanities occupied with historical change. Some maintain that biography chose them while they were engaged on other research. Ambivalence in relation to placing the biography within historiography would seem to be the result of opposition to a consideration of single characters. Individual lives, especially the private, are considered by some as irrelevant or inappropriate to the understanding of history, scholarship, or art, because the genius of these individuals has forced them to keep the world at a certain distance. Their significance, maintain the anti-biographers, is to be found in their work(s), not in their person.

That the genre gets left out in the cold, but has still imposed itself, is further illustrated by the reaction of postmodern and poststructuralist critics such as Derrida, Foucault, and Barthes, who have called the historical biography 'impossible to use as reference,' 'spurious,' 'a feature of the exercise of power,' 'profit-mongering in intimacy' or, conversely, a thanatography—an account of a person's death. Pierre Bourdieu has used the expression 'the biographical illusion' (1986), by which is meant that a life story has no direction and cannot be construed in a chronological order. A human individual, meanwhile, is not conceivable without the surrounding society and a biography does not need to present a life as a continuum. A biography can both de- and reconstruct a life. Historical biography will, in any case, easily create new myths in the process of demythologization. This situation is a quandary which historical biography will have to live with when it creates what could be called a kind of authentic fiction.

4. *The Interdisciplinary, Historical Biography: Need for Systematic, Methodical Reflection*

At the year 2000, the historical biography is inherently interdisciplinary. It has developed from a conscious mythologization of prominent, historical figures to a

critical humanization of both well-known and forgotten historical lives. Despite the abundance of publications, certain common features can be detected in historical biography of the 1990s:

The subjects of the great majority of biographies are men living in the twentieth century, most of them being politicians or scientists. Besides statesmen and princes, figures from earlier periods include religious leaders and pilgrims, intellectuals, philosophers and scientists, inventors, adventurers, lawyers, painters, composers, plus a queen or two, and a couple of early feminists.

The familiar gender-imbalance in historiography is pronounced in the historical-biographical genre: less than 4 percent of biographies reviewed in specialized journals have a woman as the central subject and less than 8 percent are written by female biographers.

The scholarly history community's parameter for the evaluation of a historical biography is still the 'life-and-times' model, some biographies having a great deal of 'life' in relation to 'times,' others having little. Most are based on archival material and some even on detective-like research. Some historical biographies skew the individual perspective at the expense of historical analysis, in others the person disappears. Generally, the relationship between the nomothetic and the idiographic is not erased in the historical biography, no more than it is in the rest of historiography.

The target readership for historical biography is often a broad public beyond the specialized forum.

Nevertheless, historical biography has not been the subject of systematic, theoretical-methodological discussion in professional journals in Europe and the USA during the 1990s. Only specialized history journals in the former East Germany have provided the framework for an independent biography-historical discussion. The content of the American journal *Biography*: *An Inter-disciplinary Quarterly*, which has been published since 1977, consists primarily of reviews. In the small Danish-language sphere, a castigation of the historical biography genre proceeded throughout the 1990s (*Historisk Tidsskrift*).

At the beginning of the new millennium a series of primary historical-biographical forms can be identified:

the classical historical biography, in which the perspective concentrates on the central figure's work and in which personalism is irrelevant (e.g., Young-Bruehl 1982).

the existential biography, where the subject is seen in his/her own creative context and the life and work as a coherent, existential whole; where the linear perspective is restored and the personal-public are inseparable (e.g., Söderquist 1998).

the historical biography as historical prism, in which the light of history is refracted and the perspective raises the central figure as representative of a time, a historical situation, a type or social phenomenon. This is often seen as a bridge-builder between research and exposition (e.g., Tuchman 1966, 1986).

the historical biography as a cultural reflector, where a popular culture is analyzed via an exceptional case, by means of which a singular personality can prove to reveal a representative aspect in the historical culture (e.g., Ginzburg 1976).

the historical life-and-times biography, where the portrait painted is colored by society's palette (e.g., Sklar 1973, Fox Keller 1983). In some French and a few Scandinavian biographies, this type has been expanded to:

the historical life-work-times biography, also known as 'the total' or 'the dialectical biography,' in which the dynamic interplay and power structure between the individual, the work and the surroundings becomes a driving force (e.g., Possing 1992, Nilsson 1994, Le Goff 1996).

relational biography, which focuses on the relationship between two or more central figures (e.g., Rose 1983, Haavet 1998).

the historiographical biography, which discusses all the others' ways of writing biographies of the same central figure (e.g., Rosenbaum 1998).

The same methodological criteria apply for the historical biography as for all other historical analysis. But greater caution and ethical awareness are required when writing the biography of a person than when analyzing a theory or a social movement. In other words, there is a difference between the critique of a collective process and that of an individual. The twenty-first century will not just need historical biographies of prominent men and (more) women. There will also be a need for biographies of the grand fiasco, the oddity, the ordinary normal citizen, and the profoundly unoriginal opportunist, because our knowledge of the mentalities of the past will thereby be deepened. The significance of biographical myth in the formation of national and supranational identities will be central to future historical research. Myths have not infrequently proved to be more creative in the view of history than have objective facts, and thus deserve historiographical treatment. The historical biography can redevelop '*l'écriture historique*,' historical writing. The fundamental question for historical biography in the twenty-first century will be if rehumanization of the humanities and social sciences is compatible with clear historical analysis.

See also: Biography and Society; History and Literature; History and the Social Sciences; Life Course in History; Life Course: Sociological Aspects.

Bibliography

Boccaccio 1354–5 *Vita di Dante Alighieri* [*Das Leben Dantes*, trans. Taube O. Leipzig, Germany]

Boswell J 1791 *The Life of Samuel Johnson*. London
Bourdieu P 1986 L'illusion biografique. *Actes de la Recherche en sciences sociales* **62/63**
Carlyle T 1841 *On Hero, Hero-Worship and the Heroic in History. Six Lectures*. London
Crapanzano V 1980 *Tuhami. A Portrait of a Maroccan*. Chicago
Davis N Z 1983 *The Return of Martin Guerre*. Cambridge, MA
Dilthey W 1870 *Leben Schleiermachers*. Berlin
Edel L 1957 *Literary Biography*. London
Edel L 1984 *Writing Lives, Principia Biographica*. New York
Einhard 829 *Vita Caroli Magni*
Engelberg E, Schleier H 1990 Zu Geschichte und Theorie der Historischen Biographie. *Zeitschrift für Geschichtswissenschaft* **38**: 195–217
Epstein W H 1987 *Recognizing Biography*. Philadelphia
Ferrarotti F 1983 Biography and the social sciences. *Social Research* : 56–80
Foxe J 1563 *Book of Martyrs*
Fox Keller E 1983 *A Feeling for The Organism. The Life and Work of Barbara McClintock*. San Francisco
Ginzburg C 1976 *Il formaggio e i vermi: Il cosmo di un mugnaio del '1500*. Torino [1980 *The Cheese and the Worms: The Cosmos of a Sixteenth-Century Miller*]
Grimm H 1873 *Leben Michelangelos*, 1–2
Haavet I E 1998 *Nina Grieg. Kunstner og Kunstnerhustru*. [Nina Grieg: Artist and Artist's Wife]. Bergen, Norway
Kendall P M 1965 *The Art of Biography*. London
Laërtius, Diogenes (3rd century AC) *Vitae Philosophorum*
Le Goff J 1996 *Saint Louis*. Paris
Levi G 1989 Les usages de la biographie. *Annales ESC* **6**: 1325–36
Liu Xiang (BC 79–8) *Lienü zhuan*. [1953 *Biographies des Femmes Illustres*, trans. Kaltenmark M. Peking]
Madélenat D 1984 *La Biographie*. Paris
Nadel I B 1984 *Biography: Fiction, Fact and Form*. London
Nilsson J O 1994 *Alva Myrdal—en virvel i den moderna strömmen*. [Alva Myrdal—a Whirlpool in the Modern Stream]. Stockholm
Ottinger E M 1854 *Bibliographie biographique, 1–2*. Bryssel
Pachter M 1979 *Telling Lives: The Biographer's Art*. Washington DC
Plekhanow G V 1898 K Voprosu o Roli Lichnosti v Istorii. [The Role of the Individual in History]. *Nauchnoe Obozrenie* **3**–4
Plutarch (AD 125) *Bioi Parallelloi*
Possing B 1992 *Viljens Styrke. Natalie Zahle. En Biografi, 1–2*. [Strength of Will: Natalie Zahle. A Biography, 1–2]. Copenhagen, Denmark
Possing B 1992 & 1997 Den historiske biografi og historievidenskaben; Biografien—en frisk eller en skæv bølge. *Historisk Tidsskrift*
Rose P 1983 *Parallel Lives. Five Victorian Marriages*. New York
Rosenbaum R 1998 *Explaining Hitler. The Search for the Origin of His Evil*. New York
Shortland M, Yeo R (eds.) 1996 *Telling Lives in Science. Essays on Scientific Biography*. Cambridge University Press, Cambridge, UK
Sima Qian (BC 145–85) *Shiji* [*Historical Memoirs*]
Sklar K K 1973 *Catharine Beecher. A Study in American Domesticity*. Yale University Press, New Haven, CT
Söderquist T 1998 *Hvilken Kamp For At Undslippe*. [What a Struggle to Escape]. Copenhagen, Denmark
Strachey L 1918 *Eminent Victorians*. London
Tacitus (AD 98) *De vita Agricolae*
Thomsen Niels 1992 & 1997 Historien om frk. Zahle—er det historie?; Biografiens nye bølge—en skæv sø? *Historisk Tidsskrift*
Tuchman B 1966 *The Proud Tower. Portrait of the World Before the War 1890–1914*. London
Young-Bruehl E 1982 *Hannah Arendt. For the Love of the World*. New Haven, CT
Zimmer D 1991 Der Mensch in der Geschichte und die Biographie. Entropie eines klassischen Streitpunktes? *Zeitschrift für Geschichtswissenschaft* **39**: 353–61
Znaniecki F, Thomas W I 1918–20 *The Polish Peasant in Europe and America: Monograph of an Immigrant Group, Vol. 1–5*. Chicago

B. Possing

Biology's Influence on Sociology: Human Sociobiology

In their early development, sociology and evolutionary biology were complementary. Soon, however, biology was subject to certain controversial issues, and most sociologists became alienated from it. Then, starting a few decades ago, biological theory made compelling strides toward the study of social, including human, behavior. As a result, while sociologists' estrangement is still widespread, a small and growing number have been accepting the challenge. This article begins with a brief statement on the early relationship. It then highlights the following major stages in the growth of evolutionary biology in view of their special relevance to sociological theory: (a) the synthesis of Darwinian theory and genetic science and (b) sociobiology, or the study of evolution and behavior. Underscored are the theory of altruism, the fitness principle, and the theories of parental investment and sexual selection.

1. *Evolutionary Biology and Sociology: The Early Stage*

Both sciences are creatures of the nineteenth century. For several decades they were also mutually influential. For instance, The Division of Labor in Society, sociologist E. Durkheim's classic, argues in a Darwinian key that labor becomes more divided in large part 'because the struggle for existence [Darwin's concept] is more acute.' In the process, Durkheim produces a theory of solidarity that foreshadows the current theory of altruism. Likewise, in 1852 H. Spencer argued, again in Darwinian fashion, that the evolution of human society was a result of competition for scarce resources in which 'the more adapted [prevailed] over the less adapted.' In short, social evolution was the result of the 'survival of the fittest.'

Darwin (1859) liked this expression in preference to his own concept of 'natural selection,' first used in 1838.

1.1 Darwin's Theory of Evolution by Natural Selection

Darwin (1859, p. 75) was at pains to point out that his great discovery (the theory of evolution by natural selection) owed much to another early sociologist, the English demographer T. Malthus. In fact, if we focus on this link of scientific progress, we can grasp immediately the basics of Darwin's theory. In 1798, Malthus had argued that the growth of human population tends to be steeper than the growth of the resources necessary for its sustenance. The inevitable result: competition for scarce resources.

In 1837 Darwin had been on a fact-finding voyage around the world for nearly five years, and had returned to England fully convinced of the fact of evolution or, as he called it, 'descent with modification.' But he had searched in vain for the ultimate, or uniform, mechanism of evolution. A reading of Malthus in the winter of 1837–8 gave him the key, leading him to reason that Malthus' argument applied to all species. Accordingly, he defined his theory as 'the doctrine of Malthus applied with manifold force to the whole animal and vegetable kingdoms; for in this case there can be no artificial increase of food, and no prudential restraint from marriage' (1859, p. 75). In short, life is a 'struggle for existence' in which, given the scarcity of resources and different innate 'variations' (later replaced by 'genes'), some individuals are more likely than others to survive and reproduce. Moreover, such variations are heritable. Natural selection, therefore, refers to a process of differential reproduction in which, through heredity, the genes that are better adapted to the competition, and thus better respond to environmental challenges, are conveyed through the generations, while the less adapted are at the same time discarded. In time, this process, along with genetic novelties later termed 'mutations,' produces descent with modification. Consider, for example, a population of ancestral females. Those who mated with individuals who excelled in the competition for resources were at a reproductive advantage in relation to the mates of less successful males. As a result, today's females are the heirs of the more successful ancestors. To a substantial degree, their reproductive strategy may also be expected to reflect the ancient one.

1.2 Sociologists' Subsequent Estrangement

The revolution in physical science took nearly two centuries to be widely accepted. Darwin's theory fared almost as poorly. It clashed with the theological view of the creation. Moreover, from the beginning there was some equivocation in the concept of natural selection. Equating it with 'survival of the fittest' was an error, both scholarly and human. In fact, as C. Badcock (1994) notes, the use of 'fit' and 'fitness' (though technical terms) invites ideological interpretations. Further, Darwin's revolution coincided with the classical stage of another revolution, and this was accompanied by an ideology that celebrated the rich and mighty: the capitalists. Survival of the fittest came to be the catchword of the new order, and 'social Darwinism' was born.

Few social scientists bothered to read Darwin carefully, if at all. Else they would have known that, for Darwin, the fittest were likely to be the members of the highly reproductive 'lower order.' They fancied Darwin through the errors of such statements as the following by Rockefeller: 'The growth of a large business is merely the survival of the fittest .... the working out of a law of nature and a law of God.' From the pulpit, others attributed the abject conditions of the working masses to the will of God.

Even great scholars were unable to escape the virus of the ideological pestilence, and as a minimum were guilty of careless pronouncements. In his Statics, Spencer, who heavily influenced the great American sociologist, W. G. Sumner among others, wrote such irritants as the following: 'The whole effort of nature is to get rid of [feeble elements], to clear the world of them.' At about the same time, starting around 1883, a movement arose—and soon compromised the IQ test and the rising science of genetics—according to which, as one enthusiast put it, 'human matings could be placed on the same high plane as that of horse breeding.' Capitalist ideology, Darwinian vernacular, and eugenics all wrapped up in one. Sociologists fell under it.

In the meantime, Europe-born sociology was migrating to the United States where the cities were teeming with humble immigrants. Even to many a great professor, these were the unfit, biologically destined to failure. Indeed, as the Great Depression crashed in, not a few of them were reduced to quasi-imbecility by hunger, prejudice, and humiliation. But the children and the grandchildren of these 'failures' soon started filling the halls of the universities. Many of them studied sociology; even today an exceptional percentage of sociology students were raised in relatively poor ethnic groups. The old epithets had stung, and were never forgotten. One of their first acts was to expunge the very word biology from the curriculum on human behavior.

Sociology today faces a marvelous challenge. The many branches of evolutionary biology, e.g., neuroendocrinology, human genetics, and primatology, leave little doubt that a fuller understanding of human behavior requires an opening to biology. The aversion to biology persists, but a small and growing number of sociologists are accepting the challenge, avidly conscious of richer frontiers in behavioral science.

2. *The Modern Synthesis*

Natural selection is one of the two atlantes of evolutionary biology. Genetics is the other. Genetics provides information on the units of heredity, the materia prima of evolution (DNA). It was discovered around 1865 by the botanist Gregor J. Mendel, but it lay dormant until 1900. Indeed, until about 1930, the two pillars of what has become a prodigious scientific revolution developed with near disregard of each other—a grave lesson on the obstacles facing the construction of scientific bridges. By 1930, however, the inevitable happened. A group of great minds formally wedded the evolutionary twins into the 'Synthetic Theory' or 'Modern Synthesis' (Huxley 1942). The particulars are numerous and complex. But T. Dobzhansky and his associates (1977, p. 18) so captured the crucial premise of the momentous synthesis: 'All biological ogranization, down to the level of molecules, has evolved as a result of natural selection acting upon genetic variation.'

Whether natural selection 'acts on' genetic variation is a knotty question that need not be treated here (but see Lopreato 1989). Technically, the action on genes comes from environmental pressures; natural selection 'records and orders' the ensuing effects through the generations. But the proposition is most remarkable, representing a splendid case of scientific inevitability. To grasp the point, we need only glance at some three centuries and behold the cornucopia of laws that constituted physical science. They too have been epitomized by a synthesizing premise, as follows: 'The universe is a system of matter in motion obeying immanent, natural laws' (Kuhn 1957). Returning to the general proposition of the modern synthesis, it is possible to show that it is a sort of corollary of this Newtonian proposition. Suppose we translate the former as follows: the universe of biological organizations is a system of genetic matter in motion obeying the immanent, natural laws of natural selection and genetic variation (Lopreato 1984). We have not disturbed the logical structure. We have merely specified (an avoidable deed) the type of matter and the relevant laws. The modern synthesis was inherent in the scientific revolution.

3. *Sociobiology: The New Synthesis*

The goal above was to exemplify the inherent tendency of sciences to unite, and thus to suggest two other inevitabilities: (a) the expansion of the modern synthesis toward the evolutionary study of behavior (sociobiology) and (b) the 'consilience' (Wilson 1998) of sociobiology and social science, namely the rise of human sociobiology, which to almost imperceptible degrees branches off into 'evolutionary anthropology,' 'evolutionary psychology,' 'evolutionary sociology,' and so forth.

The forging of the modern synthesis gave exceptional impetus to the study of social behavior and organization. By the early 1970s, four zoologists in particular (W. D. Hamilton, R. L. Trivers, G. Williams, and E. O. Wilson) had published compelling discoveries. Pre-eminent was the finding that in evolutionary history organisms, human ancestors included, appear to have behaved in fair harmony with their genetic structure (genotype). Among social scientists, this is a troublesome thought, in part because we have little experience with remote concepts. The problem is, in turn, exacerbated by a tendency to grasp the idea in teleological terms and to reject the unrestrained stress on selfishness and genetic determinism presumably implied in it. In fact, nothing in sociobiology suggests unidirectional biological determinism, or that human beings are obsessed with their 'selfish genes' (Dawkins 1989). In classical scientific mode, sociobiology stresses the heuristic value of a remote (distal, ultimate, relatively constant) explanatory device. Certainly it does not neglect dimensions of sociability. Living in society requires sentiments that act subtly in antagonistically cooperative efforts, such as devotion, a sense of fairness, and the need for vengeance. D. Thiessen (1996, p. 7) has captured the poetry of the seeming paradox as follows: 'We cannot view love, loyalty, and altruism without also seeing hate, abandonment, and selfishness.... We may wish for better, but when we accept one of our sides, we accept the other. This is our bittersweet destiny, to be noble, debased.'

In 1975, the entomologist E. O. Wilson published a hugely influential work that creatively codified the evolutionary knowledge on social behavior and formally launched sociobiology as the 'new synthesis.' For social scientists, the basic theme was the quest for a comprehensive theory of human nature: What are the fundamental forces driving human beings, and how are they expressed in group living? An intuitive quest for a theory of human nature had been central to the work of all founders of sociology; the time had now come to rely on more rigorous techniques.

3.1 *Altruism*

A major step forward was taken by W. D. Hamilton (1964) in a seminal paper on 'the genetical theory of social behaviour.' Among eusocial insects, e.g., ants, only a very small number of individuals are reproductive—a bit of a puzzle in view of Darwin's accent on direct reproduction. Such insects have a peculiar system of reproduction and sex determination (haplodiploidy); the vast majority (females) are diploid (have two parents), while males are haploid (have only a mother). In examining the kinship coefficients (r), Hamilton discovered that the diploids are more than normally related to one another. Accordingly, they have evolved to cater very diligently to the needs of the

queen and the young and, except for rare cases, to forego reproduction. Given the very close kinship, this 'altruism' redounds to their own genetic benefit. In short, eusocial species have evolved according to what is termed kin selection; the genetic fitness of organisms is inclusive of the share accruing to them from reproductive kin.

To varying degrees, kin selection applies to all social species. Hamilton eventually summarized by concluding that the probability of altruism increases (a) as the *r* between benefactor and beneficiary increases, and (b) as the benefit accruing to the beneficiary exceeds the cost incurred by the benefactor. In short, the more kindred two individuals are, and the less it costs to do each other good, the more likely they are to do it. Numberless studies show that this rule covers the typical case. Does it also apply to human beings? The answer is strongly suggested by last wills and testaments, which show that rarely do we bequeath our (fitness-enhancing) resources to others than our kin (Clignet 1992). Facts of great subtlety are even more convincing. For instance, most of us would expect a child's grandparents to be equally caring of the child. On average, they are equally related to the grandchild. But—and here is an evolutionary drama—in our ancient brain the paternity of a child is not so certain as the child's maternity, and 'paternal uncertainty' modifies the probability of altruism. Thus, a recent study shows that the most caring is the mother's mother; second is her husband, not the paternal grandmother. Dad's father comes last (Euler and Weitzel 1996). The difference in altruism between grandmother and grandfather reflects, as we shall presently see, another major theory of sociobiology, namely 'differential parental investment.'

Kin selection and altruism theories are very promising tools for sociologists. Consider briefly ethnic conflict, a phenomenon that time and again has grievously plagued the human peace in all corners of the earth. To explain it, social scientists tend to search for facts peculiar to given times and places. This is not a useless strategy. But if a phenomenon is universal and recurrent, it also requires a universally applicable explanation. We are the descendants of individuals whose social horizon was delimited by a clan of 25–50 souls. Beyond the clan was the 'out-group,' often an enemy. This condition lasted until very recent times. In the meantime our clannish brain has not changed. It is still the product of kin selection and kin altruism. Hence, even in the midst of the megasociety, we are subject to a force that whispers from within: 'they' are like 'us'; the 'others' are the enemy—if not just now, then in the time of the 'Fathers.' J. Piaget's famous studies showed that children grow up with us-vs.-them notions. Ultimately, human beings are 'ethnic' and genocidal because we have an ancient, clannish brain. Its deadliness is touched off by a variety of contextual pretexts. A more mature sociology must consider both types of causes.

Altruistic favoritism is not limited to the kin group, though it reaches its greatest intensity there. It may be observed wherever it serves the selfish needs of its agents. In 1971, Robert Trivers expanded the theory of altruism to account for acts of reciprocity even across species. For instance, a number of species of fish perform a service analogous to that of the dental technician. Their 'customers' have been observed queuing up and, as the occasion arises, defending their health providers from predators. So, the ones profit from having their mouth and gills cleared of parasites. The others receive food and protection from predators. Complex societies are possible largely because of reciprocal altruism.

3.2 The Fitness Principle

Scientific theories are held together by at least one general principle or law. The general principle of sociobiology, often referred to as the fitness or maximization principle, is implicit in the above theories and others that will be sampled below. It may be expressed as a probability statement as follows: *organisms tend to behave so as to maximize their inclusive fitness*. This is shorthand for saying that organisms are, to varying degrees, observed behaving as if they were intent on conveying the maximum possible portion of their genotype to future generations, partly or fully through kin selection. In time, however, purely probabilistic statements typically undergo contingent refinements, as in the form: if X, then Y, provided that Z.... Newton's laws of motion, for example, take this shape. Something of this nature is needed for a more finished fitness principle. A first approximation for human sociobiology is available, and its contingencies are (a) the pursuit of 'creature comforts,' (b) self-deception, and (c) 'autonomization'—as when, e.g., the quest for riches becomes an end in itself and ceases to be a means to reproductive success (Lopreato 1989). Sociology is perfectly suited to improve further.

3.3 Differential Parental Investment and Sexual Selection

The fitness principle is a broad umbrella covering a growing number of theoretical statements. This final section introduces several crucial ones specific to sex differences. It will be helpful to first recall a few widely noted facts. In general, female behavior is more cautious than its male counterpart. Thus, at all ages except the 80s women are subject to far fewer traumas resulting from accidents or agonistic activity. They are less likely to stress 'skin-deep' qualities in their mates. Young women prefer slightly older men who offer the prospects needed to raise successful children. In the market place they suffer various disadvantages. Yet there is evidence that female medical students prefer life partners who will be more successful than they.

Females are more apt to want a commitment before agreeing to sex, have fewer sexual partners, and are less disposed to cheat on them (Laumann et al. 1994). Above all, they devote more time to child care, even when they are employed in the market place.

Sociologists tend to attribute such differences to socialization, and many 'feminists' single out a system of patriarchy as the culprit designed to favor male interests. There is some validity to this view; once such a system is in place, it inevitably abets socialization processes consistent with the feminist criticism. But this strategy takes facts as given. The job of science is to explain them. Socialization is only part of the story; by itself it cannot explain why such 'sex roles' are universal and persistent. It can also result in stupefying absurdities. By its logic, we should be able, for example, to raise boys as girls and girls as boys. In fact, studies show that even 'gender stereotypes' endure unabated, despite recent deep changes in sex roles.

Nobody knows for sure how patriarchy developed. Chances are good that it is millions of years old. Hence, the best bet is that it is wired in a brain that has not changed for at least 30,000 years. What, then, is the complement of socialization? There is a major physio-anatomic difference between males and females, across the creation. Anisogamy, the name given to it, refers to the inequality between female gametes (eggs) and male gametes (sperm). In our species, eggs are about 85,000 times larger than sperm and are responsible for the nutrition of the embryo in its early stage. Produced once in a lifetime, *in utero*, only a tiny number can be promoted into offspring; pregnancy and child rearing are time-consuming. Females, in short, make a greater parental investment from the very beginning, and then continue to contribute disproportionately through lactation and child-care at least during the tender years. It follows that a mistake—mating with one who will contribute little or nothing to the well-being of the offspring—is very costly. Accordingly, evolutionists hold that women have been selected to specialize in the caring, cautious, 'choosy' reproductive strategy that has been encountered above.

The male scenario is quite different. The miniscule sperm contribute only genes. They are also produced in huge quantities continuously from the onset of sexual maturity. The reproductive potential of males is enormous, and some men are known to have sired literally thousands of children (Betzig 1986). It is no mere effect of socialization that, for example, middle-aged men frequently 'play around,' nor that they only too often fail to pay legally required child support. There is a hedonistic streak in the male endocrine system. More than females, on average males have evolved to prefer quantity and to philander. That may be detrimental to abandoned spouses and children. But evolution has not been kind to males. Given anisogamy, there is a prohibitive scarcity of candidates to philander with. The male reproductive system has encouraged polygyny, or marriage with multiple wives, so widespread until very recently. But the other side of polygyny is male celibacy. Even today a study from Sweden (Lindquist Forsberg and Tullberg 1995) shows that there is an overlarge percentace of unskilled men without mates. If we now recall the female preference for relatively rich males (e.g., Buss 1989)—resulting in what is termed 'hypergyny'—it is hard to avoid the conclusion that patriarchy, so hastily taken as a given by some sociologists, most probably evolved with the aid of female complicity. There is no intention here to pass judgment. Understanding certain facts helps to know certain others. At any rate, as Doreen Kimura (1992) has put it, boys and girls enter the world with differently wired brains.

Polygyny, hypergyny, and patriarchy are ultimately the results of the differential parental investment and sexual selection inherent in anisogamy. But these are the landmarks, as it were, of a large landscape of human nature and its rich repertoire of human facts. On sex differences the following are useful guides to such awe-inspiring abundance.

The Law of Anisogamy: the two sexes are endowed with different reproductive strategies, and their behaviors reflect that difference in direct proportion to their relevance to it (Lopreato and Crippen 1999). In other words, the closer we get to the fundamentals of life (e.g., sex, sex roles, courting, family), the more readily observable are the effects of anisogamy.

Corollary 1. Males, the sex investing less in parenting, compete more for mates than females, the sex that invests more (Trivers 1972).

Corollary 2. Females have been selected to specialize in choosy behavior, while males have been selected to specialize in agonistic behavior (Darwin 1871).

Together with a variety of culture and socialization factors, the above propositions explain large classes of behavioral differences between the sexes. But sociology will derive enormous profit from an evolutionary perspective on all its subject matter.

See also: Darwin, Charles Robert (1809–82); Darwinism: Social; Evolutionism, Including Social Darwinism; Exchange: Social; Methodological Individualism in Sociology; Population Ecology; Social Evolution, Sociology of; Sociobiology: Overview; Spencer, Herbert (1820–1903); Theory: Sociological

Bibliography

Badcock C R 1994 *PsychoDarwinism: The New Synthesis of Darwin and Freud*. Flamingo, London

Betzig L L 1986 *Despotism and Differential Reproduction: A Darwinian View of History*. Aldine, New York

Buss D M 1989 Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures (with commentaries). *Behavioral and Brain Sciences* **12**: 1–14

Clignet R 1992 *Death, Deeds, and Descendants: Inheritance in Modern America*. Aldine de Gruyter, New York

Darwin C 1959 *The Origin of Species*. Mentor Books, New York
Darwin C 1871 *The Origin of Species and The Descent of Man and Selection in Relation to Sex*. Random House, New York (undated collection)
Dawkins R 1989 *The Selfish Gene*. Oxford University Press, Oxford, UK
Dobzhansky T, Ayala F J, Ledyard Stebbins G, Valentine J W 1977 *Evolution*. Freeman, San Francisco
Euler H A, Weitzel B 1996 Discriminative grandparental solicitude as reproductive strategy. *Human Nature* **7**: 39–59
Hamilton W D 1964 The genetical theory of social behaviour: I & II. *Journal of Theoretical Biology* **7**: 1–52
Huxley J 1942 *Evolution: The Modern Synthesis*. G. Allen & Unwin, London
Kimura D 1992 Sex differences in the brain. *Scientific American* **267**: 119–25
Kuhn T S 1957 *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought*. Harvard University Press, Cambridge, MA
Laumann E, Michael R, Michaels S, Gagnon J 1994 *The Social Organization of Sexuality: Sexual Practices in the United States*. University of Chicago Press, Chicago
Lindquist Forsberg A J, Tullberg B S 1995 The relationship between cumulative number of cohabiting partners and number of children for men and women in modern Sweden. *Ethology and Sociobiology* **16**: 221–32
Lopreato J 1984 *Human Nature & Biocultural Evolution*. Allen & Unwin, Boston
Lopreato J 1989 The maximization principle: A cause in search of conditions. In: Bell R W, Bell N J (eds.) *Sociobiology and the Social Sciences*. Texas Tech University Press, Lubbock, TX
Lopreato J, Crippen T 1999 *Crisis in Sociology: The Need for Darwin*. Transaction, New Brunswick, NJ
Thiessen D D 1996 *Bittersweet Destiny: The Stormy Evolution of Human Behavior*. Transaction, New Brunswick, NJ
Trivers R L 1971 The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–47
Trivers R L 1972 Parental investment and sexual selection. In: Campbell B H (ed.) *Sexual Selection and the Descent of Man, 1871–1971*. Aldine, Chicago
Wilson E O 1975 *Sociobiology: The New Synthesis*. Belknap Press of Harvard University Press, Cambridge, MA
Wilson E O 1998 *Consilience: The Unity of Knowledge*. Knopf, New York

J. Lopreato

Biomedical Sciences and Technology: History and Sociology

Despite the recent emergence of a few major, unifying themes, the history and sociology of the biomedical sciences and technology (HSBST) can hardly be said to constitute a homogeneous, well-defined field of inquiry. This diversity probably accounts for the vitality of the field. Two reasons explain this state of affairs, namely, the scattered, somewhat twisted (sub)disciplinary origins of the field, and the ambiguities surrounding the nature and identity of its subject matter, biomedicine.

1. *Origins and Definitions of the Field*

As with other interdisciplinary endeavors, HSBST lies at the intersection of several specialized (sub)fields. In this case, the intersection is somewhat tangled. The history and the sociology of medicine can only partly be defined as the field's parent disciplines insofar as they have both experienced serious trouble in dealing with science as a subject matter. In this respect, criticism of the two disciplines' shortcomings has often served as a negative cast for delineating HSBST's programmatic content. As Warner (1985) has noted, during the 1970s and 1980s a new brand of historian of medicine, reacting against narrow internalist histories of medical theories, recentered the domain within the social history of medicine. Unfortunately, the resulting accounts have, more often than not, left out medicine and science, concentrating instead on the role of science as an ideology, a source of authority, or a mere legitimization device. Two consequences followed. The first, more obvious one, was to prevent any investigation or understanding of the actual content of medical practices. The second was to take for granted the dominant, laboratory-based definition of what constitutes medical science, thus disregarding the multiplicity and variety of clinical, laboratory, and public health practices that, at different times and in different locales, have been defined as scientific. A similar situation has characterized the development of medical sociology, insofar as the latter grew out of the distinction between disease and illness, i.e., between the biological and psycho-social aspects of pathological processes, a dichotomy that effectively left a major portion of medical activities outside of the purview of sociological analysis (Lock and Gordon 1988). One of medical sociology's key notions, namely medicalization, a term used to refer to the medical redefinition of previously nonmedical problems and behaviors, rested on the possibility of establishing *a priori*, unproblematic distinctions between social and medical problems, thus missing the opportunity to investigate the actual work that goes into the production of medical (and social) problems *qua* medical (and social) problems.

Before the 1980s, apart from a remarkable, pioneering study by Ludwik Fleck (1935) of the development of the Wassermann test for syphilis, the content of biomedical practices was, from the point of view of the social sciences, largely *terra incognita*. The situation changed with the development of science and technology studies as a semiautonomous field and the ensuing constitution of a rich body of empirical analysis of scientific beliefs and practices. The articles

collected in Wright and Treacher (1982) exemplify early attempts to apply insights from science and technology studies to medicine, by adopting the, by now, largely superseded approach of correlating a specific body of medical knowledge with the social and political interests of professional or ruling groups (see *Strong Program, in Sociology of Scientific Knowledge*). Several of the founding contributions to science and technology studies, especially those providing ethnographic analyses of laboratories (see *Laboratory Studies*: *Historical Perspectives*), took biomedical laboratories as their field site. However, for most of these early studies, biomedical activities acted as examples of (generic) scientific practices, rather than starting points for the investigation of biomedicine.

Moreover, medical practices encompass more than laboratory research: the clinic and its derivatives, including clinical research, are a major part of the picture. The relation between the laboratory and the clinic constitutes, arguably, the essential tension of the biomedical enterprise. As a result, work in HSBST, while unfolding in close relation to developments in science and technology studies, far from severing links with more traditional contributions to the history and sociology of medicine, has continued to draw from their insights, albeit by reinterpreting them. Connections across dividing lines have been made easier by the fact that social scientists in different subfields sometime share theoretical approaches, as was the case, for instance, with symbolic interactionism (compare, e.g., Strauss et al. (1985) with Star (1989)).

At this point, an additional difficulty arises. HSBST could be defined, so to speak, from without, by compiling a list of contributions that, because of their subject matter, can be construed as referring to biomedicine, even in the absence of (strong) co-citation patterns. Or, it could be more cogently characterized by the presence of a core of mutually citing contributions that bear witness to the emergence of a common set of questions. By adopting this latter avenue, it becomes possible to argue that the field's coherence is provided by its focus on the biomedical enterprise as defined by the mutually constitutive relations between the laboratory and the clinic, or, more precisely, between investigations of (normal) biology and investigations of pathological structures and mechanisms. This theme should not to be understood as a purely epistemological one, for, in addition to epistemological issues, it simultaneously raises institutional, economic, and political considerations. To visualize this point fully, a few additional remarks on the nature of biomedicine are needed.

2. *What is Biomedicine?*

Recent dictionary entries define biomedicine as a branch of medicine that is combined with research in biology or, in other words, as the application of the natural sciences, especially the biological and physiological sciences, to clinical medicine. While, indeed, in present-day discourse the term 'biomedicine' (and its adjectival form 'biomedical') generally refers to modern, scientific Western medicine, casual use of the term disguises two important facts, namely that the term itself is of recent origin (its widespread use postdates World War Two) and, more important, that the relation between biology and the clinic is, even in the twenty-first century, far from being as seamless and as unproblematic as the term would lead one to suppose. The term first surfaced, rather idiosyncratically, in the 12th edition (1923) of Dorland's medical dictionary, but its career took off in the 1950s, when users gave it two related, yet distinct meanings. On the one hand, there was the epistemological idea of a crossing over of medicine with biology, with all its institutional consequences. On the other hand, the term referred to a substantive area, namely the study of environmental stresses on the human body—a normal body in a pathological environment—as in the case of space travel (NASA, with its manned space program, was established in 1958) and radioactive fallout from nuclear experiments. Both meanings are, of course, linked, insofar as they are mirror images of the attempt to relate biology to pathology. To speak of biomedicine is to take for granted the claim that the relationship between biology and pathology, rather than a matter of quality, is simply one of quantity, i.e., of bodily mechanisms gone awry by reaching beyond the upper or lower limits of the normal spectrum. As forcefully argued by Canguilhem (1966), such a claim is highly questionable. This is not to say that such a thing as 'biomedicine' is an epistemological monster and thus a practical impossibility. Rather, it is to point to the fact that the relationship between biology and pathology is not a simple one, and that, instead of taking their relationship for granted, one should analyze their practical, ongoing articulation.

But is the biomedical project really that novel? Did not pathology already attempt to model itself on biology since Thomas Sydenham's seventeenth century proposal that diseases be treated as autonomous entities, like plants and animals, and be subjected to similar schemes of description and classification? Is it not the case that contemporary 'biomedicine' goes back to the bacteriological and laboratory revolution of the second half of the nineteenth century, or to Rudolf Virchow's contemporary program of a cellular pathology applying cell biology to pathological processes (Vogel and Rosenberg 1979, Cunningham and Williams 1992)? While the idea that the pathological can be reduced to the normal is indeed a common theme in nineteenth century medicine, at the practical level, the normal and the pathological remained the products of distinct experimental and institutional practices. From this point of view, modern biomedicine constitutes a novel institutional and scientific activity that is incommensurable with the biology and

medicine of, say, a Louis Pasteur or a Robert Koch. Biology and medicine are now tightly intertwined research enterprises. Practitioners of the activity known as biomedicine can no longer say beforehand whether a particular research project, clinical investigation or even clinical intervention will result in biological or medical facts. Somewhat like the distance between pure and applied research, that between biology and medicine has collapsed without, of course, erasing the distinction between the two activities.

3. Socio-historical Accounts of Biomedicine

The exponential rise in public funds made available to university and para-public research since 1945 has put a lot of biology on the market and has generated numerous public and private schemes to translate that knowledge into clinically useful practices. Social scientists have described a number of these large-scale movements of money and personnel, for instance by analyzing the influence of large science-based funding institutions and cancer policy on biomedical research in terms of 'crusades' (Panem 1984), 'bandwagons' (Fujimura 1996), and 'wars' (Proctor 1995). In the same vein, the wave of reform that ran through the Western biomedical world in the post-War period vastly expanded the notion of public health to include funding for basic medical research. In the UK, for example, the creation of the National Health Service in 1948 gave the Ministry of Health the power to fund and conduct research. In the USA, the National Institutes of Health, founded in 1948, quickly became the leading source of continuous federal funding for medical research. The rise in funds was accompanied by a concomitant rise in clinical personnel in university departments of medicine. As clinical departments became more academic, medical schools became increasingly research-oriented to the extent that by 1980, they awarded 42 percent of all PhDs given in the biological sciences. Useful overviews of this post-War institutional expansion can be found in Bynum and Porter (1993) and Krige and Pestre (1997).

Thus, once a handmaiden to pathology, biology has come to count as the ultimate description and account of disease origins and mechanisms. Prediagnostic biological tests for individuals and screens for groups as well as therapeutic monitoring procedures now pervade the practice of medicine. This trend has been analyzed in Reiser's (1978) seminal overview of clinical instruments and machines, that emphasized a number of themes that have since become standard. It is generally held, for example, that new scientific technologies have tended to remove the patient from the clinical picture with clinicians' attention being increasingly led from the bedside to the laboratory. As a consequence, modern medicine can be characterized by its high degree of specialization, often based on new scientific technologies. In addition, it is maintained that scientific technologies have downgraded or eliminated skills embodied in what is termed clinical judgment with the consequence being that diagnosing physicians became more and more subservient to technologies whose development they cannot control (Berg 1997). Although biology, as characterized by the investigation of quantifiable, biological variables, has yet to substitute itself, in the reductionist sense, for pathology, it has to a certain extent encircled medicine. While the qualitative, synthetic judgment of the pathologist or clinician recognizes a specific disease in an individual, biological variables, by definition, vary continuously and thus apply to populations. They consequently open the door to screening and automation in medicine, two practices that bring pathology and biology together within a single space of representation.

In analyzing the institutional and conceptual innovations that have created the dynamic interchange between biology and pathology, historians and sociologists have isolated a series of trends. A major trend is the alignment of biomedical research along molecular and cell biology lines (Moulin 1991, de Chadarevian and Kamminga 1998). The trend predates the meteoric development of molecular biology and molecular genetics during the second part of the twentieth century. The latter, however, is generally regarded as the agent of an epochal transition to a new kind of medicine, 'predictive medicine,' whose full effects have yet to be felt (Wailoo 1997) but that has the potential of extending medical intervention from the sick to the potentially sick person, and from individual to collective patients (families and populations).

One of the consequences but also one of the causes of the use of biological variables in the diagnosis of diseases, and thus of the renewed alignment of the biological and the medical, has been that the technologies for the investigation of the normal and the pathological have become increasingly the same. Historical studies have shown, for example, that an experimental system instituted for the study of cancer can easily become a system for the study of protein synthesis (Rheinberger 1997). As medical problems thus find their solution through biological innovation, the transfer of concepts and techniques from fundamental biology to medical practice has been vastly accelerated. Conversely, medical problems have become the starting point for many fundamental biological investigations (Sinding 1991). Finally, clinical trials, the hallmark of so-called evidence-based medicine, are no longer restricted to the testing of therapies. As a spate of recent histories has shown, since the 1960s, an expanded notion of clinical trial now includes the study of biological markers and etiological research. No longer solely a determination of the efficiency of treatment, clinical trials are sim-

ultaneously an exploration of human biology (Löwy 1996).

While in the early twentieth century several areas of medical diagnostics became routinely based on technologies derived from chemistry and physics (Howell 1995), in the second half of the century, and despite the contributions of nuclear medicine and biophysics (Blume 1992), biomedical technologies have become relatively independent of physics and chemistry. Indeed, the post-War alliance of biology and medicine has made possible a new kind of medical technology that is truly biomedical: generated within biology and medicine, these biotechnologies use fragments of living systems such as cells, enzymes, antibodies, and gene segments. Whether or not genetic engineering techniques, and the results obtained through their application to, say, the study of the human genome, have the potential of transforming biomedicine by collapsing ontological distinctions between nature and society (see the debate between Rheinberger and Rabinow in Lock et al. (2000)) is still an open question. What is undisputed is that, in the absence of outside, physico-chemical, standards, the effective use of biotechnologies requires intensive standardization and new forms of collaboration between biomedical practitioners (Fujimura 1996). In a similar vein, Gaudillière (in de Chadarevian and Kamminga 1998) has suggested that engineering and management practices have played an important role in the spread of a number of experimental models of cancer etiology, and Marks (1997) has shown how clinical research has been transformed by clinical trials and how, during World War Two, standardization and cooperation came to be seen as necessary although not sufficient conditions for the conduct of clinical research. In short, it has become increasingly clear that regulatory interventions—ranging from tacit agreements to the stabilization of practices, the establishment of norms, the setting of *de facto* standards and guidelines, and the promulgation of formal, explicit regulations—have become part and parcel of the biomedical enterprise.

Nonconventional actors such as patient support groups and activist organizations increasingly play a role in these regulatory activities, and thus in the production of biomedical knowledge, insofar as the former are constitutive of the latter (Epstein 1996). Moreover, support group activities are consonant with the development of new group identities based on the sharing of an actual or potential condition, as in the case of hereditary or genetic diseases. These new 'bio-social' identities question traditional distinctions between scientists and lay people or between experts and politicians, thus challenging the regulatory and policy mechanisms to which governmental bodies had grown accustomed (Rabeharisoa and Callon 1999, Rabinow 1999).

In keeping with leading approaches in science and technology studies, much work in HSBST has focused on the *material culture* of biomedicine, that is, on the local, contingent production of biomedical practices and representations, examining in often painstaking detail the various elements and tasks that enter in their constitution. Much of this work, in turn, has gone into the analysis of the (interactive) production of medical talk, images, inscriptions, standards, classifications, and the mobilization of tools, skills, and bodies (e.g., Atkinson 1995, Cambrosio and Keating 1995, Fujimura 1996, Berg 1997, Berg and Mol 1998). Yet it would appear that the emerging coherence of HSBST comes from its focus on biomedicine as a distinctive social, institutional, material, and epistemological configuration.

See also: Bioethics: Examples from the Life Sciences; Biotechnology; Ethics for Biomedical Research Involving Humans: International Codes; Medical Sociology; Medicalization: Cultural Concerns; Medicine, History of

Bibliography

Atkinson P 1995 *Medical Talk and Medical Work*. Sage, London

Berg M 1997 *Rationalizing Medical Work. Decision-support Techniques and Medical Practices*. MIT Press, Cambridge, MA

Berg M, Mol A (eds.) 1998 *Differences in Medicine. Unraveling Practices, Techniques and Bodies*. Duke University Press, Durham, NC

Blume S S 1992 *Insight and Industry. On the Dynamics of Technological Change in Medicine*. MIT Press, Cambridge, MA

Bynum W F, Porter R (eds.) 1993 *Companion Encyclopedia of the History of Medicine*. Routledge, London

Cambrosio A, Keating P 1995 *Exquisite Specificity: the Monoclonal Antibody Revolution*. Oxford University Press, New York

Canguilhem G 1966 *Le Normal et le Pathologique*. P.U.F., Paris [1989 *The Normal and the Pathological*. Zone Books, New York]

Cunningham A, Williams P (eds.) 1992 *The Laboratory Revolution in Medicine*. Cambridge University Press, Cambridge, UK

de Chadarevian S, Kamminga H (eds.) 1998 *Molecularizing Biology and Medicine: New Practices and Alliances, 1910s–1970s*. Harwood, Amsterdam

Epstein S 1996 *Impure Science. AIDS, Activism, and the Politics of Knowledge*. University of California Press, Berkeley, CA

Fleck L 1935 *Entstehung und Entwicklung einer Wissenschaftlischen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv*. Benno Schwabe, Basel [1979 *Genesis and Development of a Scientific Fact*. University of Chicago Press, Chicago]

Fujimura J H 1996 *Crafting Science: a Sociohistory of the Quest for the Genetics of Cancer*. Harvard University Press, Cambridge, MA

Howell J D 1995 *Technology in the Hospital: Transforming Patient Care in the Early Twentieth Century*. Johns Hopkins University Press, Baltimore, MD

Krige J, Pestre D (eds.) 1997 *Science in the Twentieth Century*. Harwood Academic, Amsterdam

Lock M, Gordon D R (eds.) 1988 *Biomedicine Examined*. Kluwer, Dordrecht, The Netherlands
Lock M, Young A, Cambrosio A (eds.) 2000 *Living and Working with the New Medical Technologies. Intersections of Inquiry*. Cambridge University Press, Cambridge, UK
Löwy I 1996 *Between Bench and Bedside. Science, Healing, and Interleukin-2 in a Cancer Ward*. Harvard University Press, Cambridge, MA
Marks H M 1997 *The Progress of Experiments: Science and Therapeutic Reform in the United States, 1900–1990*. Cambridge University Press, Cambridge, UK
Moulin A M 1991 *Le Dernier Language de la Médicine. Histoire de l'Immunologie de Pasteur au Sida*. P.U.F., Paris
Panem S 1984 *The Interferon Crusade*. Brookings Institution, Washington, DC
Proctor R N 1995 *Cancer Wars. How Politics Shapes What We Know and Don't Know About Cancer*. Basic Books, New York
Rabeharisoa V, Callon M 1999 *Le Pouvoir des Malades. l'Association Française Contre les Myopathies et la Recherche*. Presses de l'École des Mines, Paris
Rabinow P 1999 *French DNA: Trouble in Purgatory*. University of Chicago Press, Chicago
Reiser S J 1978 *Medicine and the Reign of Technology*. Cambridge University Press, Cambridge, UK
Rheinberger H J 1997 *Toward a History of Epistemic Things. Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford, CA
Sinding C 1991 *Le Clinicien et le Chercheur: des Grandes Maladies de Carence à la Médecine Moléculaire* (*1880–1980*). P.U.F., Paris
Star S L 1989 *Regions of the Mind: Brain Research and the Quest for Scientific Certainty*. Stanford University Press, Stanford, CA
Strauss A L, Fagerhaugh S, Suczek, B, Wiener C 1985 *The Social Organization of Medical Work*. University of Chicago Press, Chicago
Vogel M J, Rosenberg C E (eds.) 1979 *The Therapeutic Revolution. Essays in the Social History of American Medicine*. University of Pennsylvania Press, Philadelphia, PA
Wailoo K 1997 *Drawing Blood. Technology and Disease Identity in 20th Century America*. Johns Hopkins University Press, Baltimore, MD
Warner J H 1985 Science in medicine. *Osiris* (*Second Series*) **1**: 37–58
Wright P, Treacher A (eds.) 1982 *The Problem of Medical Knowledge. Examining the Social Construction of Medicine*. Edinburgh University Press, Edinburgh, UK

A. Cambrosio and P. Keating

Biopsychology and Health

In countries and periods of history dominated by war and poverty, the attitude towards health predominantly concentrated on survival, by caring for sufficient food supplies. But in more affluent societies and historical periods, attention to mere alimentation changed into interest in nutrition and its significance for health and disease. In the twentieth century, nutrition turned into a concern of medicine when it became obvious that a deficiency of certain nutrients or the wrong diet could cause disease. Therefore the analysis of food components (for example, artificial preservatives or flavorings) became an important aspect of general programs on risk assessment concerned with a variety of environmental risk factors, for example, exposure to air pollution, radiation, chemicals, or drugs of abuse.

Similarly, at the beginning of the twentieth century, the discovery of bacterial infections by Semmelweis, Koch, and Virchow led to the improvement of public hygiene. Respective laws for protection from bacterial infection as well as for food safety were enacted. They were later followed by regulations on permissible doses of chemicals or radiation at workplaces and the safety of drugs.

Another line in health psychology came from stress research developed from experiments on animals and observations in psychosomatic diseases demonstrating the mechanisms by which mental and physical strain affect somatic processes. At present, however, there is more concern about public health and safety rules than about the doses and biological mechanisms by which environmentally noxious stimuli may affect the organism. Therefore, this article is devoted to illustrating biological mechanisms by which health-related behaviors affect somatic systems involved in the maintenance and restoration of health. Furthermore, the role of personality and the genetic roots of these mechanisms will be discussed.

1. *Classes of Health-related Behaviors and Classes of Somatic Mediators*

There are three major classes of behavior related to health: (a) avoiding exposure to or protecting the organism from noxious influences and environmental risks; (b) the application of drugs, nutrients, or psychological pleasure in order to increase health; and (c) challenging the organism by training techniques. The first two behaviors try to adapt the environment to the organism, while the third behavior adapts the organism to the environment. All the behaviors more or less act via the following somatic mediators: (a) the autonomic nervous system; (b) the endocrine system and its neurotransmitter-related feedback mechanisms; (c) the central nervous system; (d) the musculoskeletal and peripheral nervous system; (e) the immune system; (f) metabolism; and (g) systems of cell biology relevant for genetic information, growth, and repair functions.

Most papers on behavior and health do not communicate details on the somatic systems involved, and those on psychophysiology usually relate behavior to the somatic mediators, but cannot prove that they are directly related to later beneficial or detrimental effects

on health outcome, so that no causal conclusions are possible. Restoring health through behavior can be more validly related to health outcome than preventive behavior, because comparison of treatment is possible in small samples, whereas beneficial effects of preventive behavior are hard to prove and require larger samples and longer observation times for identifying the signal to noise ratio. Furthermore, in questionnaire studies, reported health behavior is highly confounded by personality factors (e.g., neuroticism) that are related to the behavior as well as to reporting symptoms of disease. Therefore the emphasis will be on results obtained from experimental and clinical sources.

2. *Avoiding Noxious Influences*

In order to demonstrate the beneficial influence of avoiding noxious stimuli, their detrimental effects have to be outlined. It must be emphasized, however, that most of the stimuli avoided (like stress, excess calories, certain food components, or even alcohol) can also have beneficial effects, if exposure is moderate.

2.1 *Avoiding Stress*

Exposure to social, psychological, and work-related stress relates to all the mediator systems listed and associated diseases (for a review, see Brown et al. 1991, Krantz et al. 1985).

The major mediator is the *sympathetic nervous system* (*SNS*), with its influence on cardiovascular reactions leading to hypertension and myocardial ischemia or infarction. Constant sympathetic tone leads to constriction of vessels, inadequate perfusion, anaerobic metabolism, and diminished contraction of the heart—and therefore to reduced oxygen supply. Furthermore, the stress-induced increase in adrenaline increases cell metabolism, which requires more oxygen, leading to a discrepancy between supply and demand. In addition, stress-induced increases in cortisol, in combination with increased cholesterol, may increase the danger of infarction by increasing the tendency of blood clotting. Mediating processes for hypertension consist of continuous contraction of vessels and false shifts in feedback mechanisms from baroreceptor signals to the brain (see *Stress and Health Research*).

With respect to the digestive system, the stress-induced sympathetic tone will first suppress parasympathetic secretion of hydrochloric acid with a subsequent rebound effect. Furthermore, adrenaline is known to increase secretion of gastrin, an acid-stimulating peptide in the mucosa. Finally, protective factors like prostaglandines are reduced, and the blood supply of the mucosa in the stomach is decreased due to sympathetic vasoconstriction. This may lead to necrosis of mucosa areas. Moreover, stress-induced gastric ulcers may develop from hemorrhagic erosions facilitated by stress-induced cortisol increase. Furthermore, sympathetically-induced adrenaline increases plasma glucose levels and inhibits insulin, resulting in hyperglycemia. This contributes to the development of diabetes. Also, disturbances in bowel movements (constipation) may result from sympathetic arousal.

All *endocrine systems* respond to physical and psychological stress, in particular if the stressor involves uncertainty (see Brush and Levine 1989, Nemeroff and Loosen 1987). All hormones are organized according to feedback loops: high blood levels lead to reduced release of peptidergic releasing factors in the brain, with subsequent reduced production of respective glandotropic hormones from the pituitary. Therefore, imbalance in the system usually elicits a cascade of disturbances which may also affect other endocrine systems. So, sustained stimulation of the hypothalamo-pituitary-adrenocortical (HPA) axis leads to considerable increases in cortisol. Due to its close interaction with the hypothalamo-pituitary-gonadal axis, hyperstimulation of the HPA axis will result in a suppressed functioning of reproductive hormones and subsequent disturbances of the menstrual cycle and fertility. Due to its antagonistic function to insulin in glucose metabolism, cortisol increase may also lead to insulin resistance and diabetes. Prolonged cortisol increase may lead to a degeneration of hippocampal cells and therefore even to a decline in memory. Furthermore, the corticotropin releasing factor (CRF) of the brain also inhibits gastric and colonal transit time, leading to constipation.

Relations between stress and the *immune system* have been widely described. They are closely related to the SNS as well as to the HPA axis activities (e.g., see Lewis et al. 1994, Glaser and Kiecolt-Glaser 1994). This is due to the fact that immune cells (macrophages, lymphocytes, and natural killer (NK) cells) bear receptors for neurotransmitters, hormones, and neuropeptides, and because the lymphoid tissues for cell storage (e.g., spleen, lymph nodes, and bone marrow) are directly innervated by sympathetic nerves. Although activation of immune cells is mainly directed against antigens discovered as 'nonself' (e.g., viruses, bacteria, neoplastic, or damaged cells), it also occurs upon nonantigen-related stressors. In a very simplified manner, it could be said that the SNS has a primarily activating effect (increasing the number and activity of these cells), as observed in acute stress, whereas the hormones of the HPA axis have more immune-suppressing than activating effects. Immunosuppression is only observed after chronic stress (e.g., hopelessness, helplessness). The reduction in macrophage and NK cell activity diminishes the defense mechanisms, increasing the probability of infection and growth of cancer cells because environmental and internal antigens can no longer be eliminated. More-

over, the capacity of the immune system to discriminate between 'self' and 'nonself' or between harmless and dangerous antigens is reduced by stress factors. In combination with reduced cortisol release as a consequence of chronic stress, this may lead to immunological overreactions, as observed in allergies and autoimmune diseases. In these diseases, the production of soluble mediators of immune cells, like cytokines and growth factors, contribute to the aggravation of symptoms.(see *Psychoneuroimmunology*).

In summary, the relationship between the immune system and stress follows an inverted U-function, which means that avoiding stress as well as exposure to an extremely high stress load (intensity, duration) are both accompanied by low immune functions. The most important issue in immune functioning is therefore to maintain a balance between activation and suppression of immune functions. A disturbed immunological homeostasis is the most serious factor for developing immunological diseases.

The *musculoskeletal system* is affected by physical stressors, like muscular constraint induced by monotonous repetitive physical work. This increases muscular tension by sustained activation of small low-threshold motor units, which leads to degenerative processes. If, in addition, carbon dioxide (CO_2) is increased by hyperventilation, neuronal excitability and muscular tension will develop, a process which has also been shown to result from mental stressors.

Protection from stimuli like *radiation*, *sun exposure*, *or air pollution* aims at reduction of the risk of cancer. These stimuli exert direct damage of DNA in the cells, for instance, in melanomas developing after heavy sun exposure and after involuntary exposure to radiation. Among polluting substances, lead is one of those most thoroughly investigated, since it produces subclinical central nervous disturbances such as attention deficits.

2.2 Avoiding Unhealthy Food

Beliefs about what kind of food is unhealthy refer to an item's calorie, salt, fat, sugar, and chemical food additive content, as well as to its likelihood of being bacterially contaminated.

Restricting calorie intake in general to avoid obesity aims at reducing heart failure, hypertension, and stroke, as well as secondary insulin deficiency and diabetes. It is less known that cancerogens are also reduced by a restricted diet. This results from an increase in the production of enzymes involved in eliminating free radicals from the cell, which play a major role in DNA damage, mutagenic changes, and in the development of cancer (Joannides 1998). Low-calorie supply also supports apoptosis, the death of cells, which is necessary to keep cell numbers constant, to remove DNA-damaged cells, and to protect the organism from toxicants and allergens. Therefore, oncogen formation has been shown to be lower in food-restricted animals.

Salt is accused of contributing to hypertension because of its sensitizing effect on catecholamines and its effect on water retention, both of which are relevant to hypertension. This is supported by intercultural correlations between the consumption of salt and the incidence of hypertension, as well as by effects of salt reduction in the diet of hypertensives.

Fat is avoided not only for its high caloric value, which contributes to obesity, but in particular for its cholesterol content, which leads to precipitation of arteriosclerotic plaques in the arteries, increasing the risk of coronary heart disease, particularly in combination with low density lipoprotein (LDL). Avoiding fat containing saturated fatty acids (e.g., those found in pork) as opposed to unsaturated fatty acids (those found in olive oil) reduces cholesterol formation, which is only increased by saturated fatty acids.

Avoidance of *sugar* aims at reducing calorie intake and preventing bacterial growth, which leads to dental caries. Bleached white sugar in particular is avoided because of its suspected cancerogenic property. According to animal research, this is claimed to develop from the formation of free radicals induced by the hypochlorites used for bleaching. But questionnaire studies relating health problems to consumption of sweet food did not reveal associations, although this food is consumed more in conditions of stress.

Smoked food is also avoided for its possible carcinogenesis. Its consumption leads to formation of nitrosamines and benzpyrines thought to damage DNA function, which again may lead to abnormal cell proliferation. *Food preservatives* like biphenyls (e.g., those sprayed on fruit) may also cause DNA damage. However, international food safety regulations that define the admissible daily intake (ADI) of food additives and the permitted procedures for smoking, radiation, or genetic manipulation of food are so stringent nowadays that food from countries subjected to these laws may be considered as generally safe.

The fear of *food contaminated* by microorganisms is widespread and leads to avoidance of unwashed or uncooked fruit and vegetables or to abstinence from food which is beyond its recommended date of consumption. However, no studies have been conducted to demonstrate noxious consequences to the health of people who neglected these dates, because there are wide safety ranges based on differing national regulations. However, since the discovery of the BSE disease in cows (suspected to lead to the central nervous condition Creutzfeldt-Jakob Disease, or CJD), and after reports of salmonella infections in chickens, avoiding respective sources of beef and eggs was felt to be justified even by government regulations.

Persons with respective experiences will avoid *food causing allergies*, such as fish, shrimps, eggs and other proteins, or fruit (particularly strawberries, tomatoes,

or nuts), but also food additives (such as benzoates or antioxidants, colorants, and flavorings). These allergies are mediated by intensive production of immunoglobulin E antibodies formed against the food allergens, which lead to severe antigen antibody reactions upon reexposure of antigens (for details, see Ziegler and Filer 1996).

2.3 Avoiding Drugs of Abuse

Alcohol is avoided because of several risk factors (see Verschuren 1993). The most prominent is drunken driving, but people are also afraid of developing liver cirrhosis, a degenerative induration of liver cells upon heavy long-term consumption. It results from inhibition of prolinoxidase, which leads to formation of fibrotic tissue. Furthermore, alcohol leads to higher concentrations of free fatty acids and to inhibition of protein synthesis, resulting in accumulation of lipoproteins and fat in liver cells (see *Drug Use and Abuse: Psychosocial Aspects*).

Thirty-two epidemiological studies proved a linear or j-shaped relationship between alcohol intake and blood pressure, even when age, weight, smoking, coffee consumption, and physical activity were partialled out. The regression becomes linear with consumption of 30–40 g of ethanol per day. Although acute alcohol administration and low doses may lower blood pressure by dilation of capillaries and reduced peripheral resistance, chronic intake may reduce the diameter of blood vessels by incorporation of cholesterol. Similar mechanisms are involved in the association between chronic consumption and stroke, whereas the protective effects of lower doses against coronary heart disease may be due to its fibrinolytic (and therefore antithrombotic) properties and its capacity to increase the antisclerotic high density lipoprotein (HDL). A further hypertensiogenic mechanism by higher doses is the activation of the renin-angiotensin system, which, in combination with a reduced venous return of the blood and neurogenic mechanisms, may exert pressor effects on the cardiovascular system. The relationship between alcohol and colonorectal cancer, however, is weak. Cell damage of the central nervous system after long-term exposure is well-known in psychiatry as the Korsakow disease (with symptoms of tremor and amnesia).

There is also a direct effect of ethanol on endocrine functions (decrease of testosterone, but increase of weak androgens and reduction of gonadotropin secretion from the hypothalamus and the pituitary) resulting in reproductive deficits.

Smoking in most cases is not a matter of decision to abstain from a primarily desired activity, because its reinforcing properties have to be learned, but quitting is related to avoiding risks. These are threefold: (a) nicotine induces adrenaline release, leading to vasoconstriction and the danger of coronary and peripheral arterial obstruction; (b) the formation of carbon monoxide-induced methemoglobin reduces oxygen supply, which increases ischemic pain; (c) the numerous carcinogenic substances contained in tobacco smoke, primarily benzyprens, are the major source of lung cancer, the prevalence of which among smokers is 5.2 times higher than in nonsmokers, even when correcting for confounding factors (age, social class, and alcohol consumption) (see *Smoking and Health*).

3. Application of Health-promoting Stimuli

Behavior of this type mainly refers to oral application of '*healthy food*,' vitamins, minerals, or drugs. Fruit and vegetables are considered healthy because of their low calorie and cholesterol content and richness in vitamins and fiber. Fiber promotes the motility of the intestines and protects against cancer. Proteins are preferred because they need more endogenous energy to be metabolized and are therefore less fattening than carbohydrates and fat.

With respect to *vitamins*, both under- and overdosing is related to disease. Therefore, health behavior may also result in symptoms caused by overdosing. *Vitamin A* (retinol) is a major product of rhodopsin necessary for vision. Deficiency not only results in problems adapting to the dark, but also sclerosis of the cornea, skin, and mucous membranes. Hypervitaminosis, among other symptoms, has teratogenic effects (malformations in the offspring). *Vitamins B1* (thiamin), *B2* (riboflavin), *B5* (nicotinamide), pantothenic acid, and *B6* (pyridoxine) are usually combined in vitamin pills and have the common function of serving as cofactors in different enzymes of the intermediate metabolism. Therefore, a multitude of symptoms may results from vitamin deficiencies: neurological (B1, B5, B6) and cardiovascular disturbances (B1), arteriosclerotic (B2) and dermatological/epithelial changes (B5, B6). *Vitamin B12* (cyanocobolamin) in combination with *folic acid* is involved in hematopoesis and nucleotide formation necessary for DNA formation. Therefore deficiency leads to pernicious anemia.

Vitamin C (ascorbic acid) is the most popular vitamin because of its well-known protective effects against scurvy. This is due to its ability to support collagen biosynthesis. The same mechanism is suspected to operate against invasion of microorganisms during infections. This vitamin, therefore, is consumed in large doses in order to avoid infectious diseases.

Vitamin E (tocopherol), by virtue of its antioxydative potential, protects the cells from free radicals and lipid peroxides which cause instability of cell membranes. Therefore, Vitamin E is said to be protective against carcinogenicity, and also against Alzheimer's disease.

Among *minerals and trace metals*, health-minded individuals are best familiar with the necessity of ingesting enough *iron* ions, available in meat and vegetables or medication, because of its central role in formation of hemoglobin, which transports oxygen. A deficiency may result not only from diminished intake, but also from increased need, such as during infections, pregnancy, and cancer; from dysfunction in storage in liver or bone marrow; and from increased loss, such as bleeding. *Fluoride* is known to be beneficial for protection against caries. Many other elements, such as calcium and phosphorus, are basic requirements for maintaining cell function and hence are constituents of diet supplementation. Also, the relevance of trace metals, for example iodine, copper, manganese, cobalt, zinc, and selenium, which fulfil important biochemical functions as cofactors in biotransformation enzymes (selen, zinc) or as components of hormones (iodine in thyroxin), are targets of dietary enrichment in persons concerned about health.

It must be summarized, however, that except for states of deficiency in vitamins, minerals, or trace metals as perhaps in malnutrition, disturbances of internal absorption, pregnancy, or infectious diseases, no health-promoting or disease-prohibiting effects of vitamins have been proven by epidemiological studies.

4. *Challenging the Organism by Training*

Health-minded persons attempt to adapt their organisms to environmental requirements for health reasons. They achieve this through training techniques, such as *somatic exercises*, relaxation, or behavior therapy. Somatic exercise, like sports, act on all the mediator systems listed above. Any isometric or isotonic muscular activity increases heart rate. But long-term training also increases the number and diameter of myocardial muscle fibers and the growth of collaterals of the coronaries. This results in the lowering of blood pressure and increase in stroke volume. The acute effect of running is said to produce short-term euphoria by release of endogenous β-endorphins.

There is a clear negative correlation between amount of exercise a person takes per week and their relative risk of mortality. It is supposed that this is due to production of lower values of LDL cholesterol. However, it must be kept in mind that performing exercise is not only a cause, but also a consequence of better health. Effects of physical exercise on the immune system follow the same inverted U-shape relationship as that of mental stressors. Increases in number and activity of NK cells (Shephard and Shek 1994), mediated by adrenaline and β-endorphin release, result from short-term activity but will be suppressed again upon longer physical strain, mainly by release of cortisol. Long-lasting or extensive sport activities may lead to higher rates of infections (see *Physical Activity and Health*).

Relaxation training and *biofeedback techniques* are usually applied to alleviate rather than to prohibit symptoms. These techniques aim at influencing the autonomic nervous system (blood pressure, thermoregulatory effects, or visceral symptoms) and musculoskeletal pains. Sleep disturbances may be alleviated because feedback loops from the SNS and baroreceptors to the brain may decrease neuronal firing. The effects may be acutely observed, but their enduring effects after cessation of training are never very long-lasting. Also, relaxation-induced effects on the immune system are well documented (e.g., in a review by van Rood et al. 1993), and include an increase in the number and activity of NK cells and macrophages and in immunoglobulin A. The clinical relevance for reduction of cancer or infections, however, has not been shown in the same populations because large prospective epidemiological intervention studies would be required.

Compliance with the physician's recommendations and willingness to take part in *behavior therapy* are, of course, similar active processes suitable to influence somatic functions.

5. *The Genetic Basis for Behavior and Mediator Systems*

As outlined in the introduction, many behaviors are determined by personality traits, and this again is partly of a genetic nature. Thus, risk-taking behaviors, such as seeking dangerous activities, exposing oneself to drugs of abuse (particularly alcohol), careless driving, unsafe sex, and risky dangers in connection with crime, have been clearly shown to have a genetic basis of about up to 56 percent. These behaviors may already be predicted from traits measured at early adolescence. The traits are directly related to lower expectations about the person's beliefs about their susceptibility to health risks, and are negatively associated with the personality dimensions agreeableness and conscientiousness. Traits such as neuroticism, anxiety, conscientiousness, and the coping style of sensitization are associated with cautious, health-minded behavior and avoidance of noxious stimuli. Also, these traits have been shown to have a genetic basis of about 30–50 percent. Undertaking training activities is probably more learned, but is also based on hereditary traits, such as conscientiousness.

On the other hand, somatic mediators are also very much influenced by genetic factors, for instance body weight or dispositions for heart and circulatory diseases. Genes responsible for the development of certain types of cancer have been traced to their location on certain chromosomes; the development of

diabetes, disturbances in lipoprotein formation, and the disposition for allergies and autoimmune diseases, as well as for Alzheimer's disease, have all been proven to have a genetic basis, as demonstrated by twin studies and by molecular genetic methods. Of course, genetic factors are never related to one gene or one chromosome, but result from multiple interactions between different genes.

The major conclusions drawn from above for future concern about health research are threefold:

(a) Health research should focus on the process characterized by the Allostatic Load Model developed by McEwen (1998), which refers to the ability 'to achieve stability through change,' i.e., activation of physiological systems in order first to cope with the stressor and then to shut off the allostatic response as soon as the stress is terminated. This means that not the size of response to a noxious stimulus, but the adaptability of the organism has to be the major concern.

(b) Epidemiological and psychophysiological/biochemical research should become more closely related.

(c) Scientific facts about the pathophysiology of risk factors should be provided by administrative agencies, instead of restrictive rules, in order to prevent health behavior ridden by irrational beliefs.

See also: Health Behaviors; Psychoneuroimmunology; Stress and Cardiac Response; Stress and Health Research

Bibliography

Brown M R, Koob G F, Rivier C 1991 (eds.) *Stress: Neurobiology and Neuroendocrinology*. Dekker, New York

Brush F R, Levine S 1989 (eds.) *Psychoendocrinology*. Academic Press, San Diego, CA

Glaser R, Kiecolt-Glaser J 1994 (eds.) *Human Stress and Immunity*. Academic Press, San Diego, CA

Joannides C 1998 (ed.) *Nutrition and Chemical Toxicity*. Wiley, Chichester, UK

Krantz D S, Grunberg N E, Baum A 1985 Health psychology. *Annual Review of Psychology* **36**: 349–83

Lewis C E, O'Sullivan C, Barraclough J 1994 (eds.) *The Psychoimmunology of Cancer*. Oxford University Press, New York

McEwen B S 1998 Protective and damaging effects of stress mediators: Allostasis and allostatic load. *New England Journal of Medicine* **840**: 33–44

Nemeroff C B, Loosen P T 1987 (eds.) *Handbook of Clinical Psychoneuroendocrinology*. Guilford Press, New York/London

Shephard R J, Shek P N 1994 Potential impact of physical activity and sports on the immune system—a brief review. *British Journal of Sports Medicine* **18**: 340–69

Van Rood Y R, Bogaards M, Goulmy E, van Houwelingen H C 1993 The effects of stress and relaxation on the *in vitro* immune response in man: a meta-analytic study. *Journal of Behavioral Medicine* **16**: 163–82

Verschuren P M 1993 (ed.) *Health Issues Related to Alcohol Consumption*. ILSI Press, Washington, DC

Ziegler E E, Filer L J 1996 (eds.) *Nutrition*. ILSI Press, Washington, DC

P. Netter

Biosphere Reserves

1. Definition

Biosphere reserves are natural protected areas included in a global network organized by the United Nations Educational, Scientific, and Cultural Organization (UNESCO). Participating countries propose land and water sites within their boundaries as potential biosphere reserves, and accepted sites are designated at the international level by UNESCO's Man and the Biosphere (MAB) Program. To qualify for acceptance as a biosphere reserve, a protected area must have global or regional significance for biological conservation, one or more inviolate core zones, and one or more surrounding buffer zones or transition zones where human communities utilize natural resources in ecologically sustainable ways.

Approval of a proposed site by the MAB Program morally commits the host country to manage the protected area according to designated standards and to participate in MAB's international biosphere reserve network. On-the-ground management of biosphere reserves continues to be the obligation of the individual countries, according to their national laws. Participation in the MAB biosphere reserve system is voluntary; countries are not required to designate biosphere reserves. UNESCO's role in the biosphere reserve network focuses on facilitating information exchange, providing guidance and technical assistance, encouraging international cooperation, and promoting financial and technical support from governments and international organizations. A small secretariat staffed by professional conservationists coordinates the biosphere reserve program from UNESCO's headquarters in Paris. As of mid-1999, 357 biosphere reserves had been established in 90 countries.

2. The Concept

The central tenet of biosphere reserves is the conservation of natural resources alongside their utilization for human benefit. The overall goal of biosphere reserves is the protection of biological diversity, but they differ from strictly protected areas such as national parks and wilderness areas by accepting

human settlement as a feature of the landscape. Biosphere reserves do not represent an attempt to wall out the external world, but to incorporate human populations into sustainable land-use systems. In most biosphere reserves, land ownership remains unchanged upon acceptance of the reserve into the global network: private landowners retain their holdings, but agree to manage them in ways compatible with the reserve's maintenance. In other biosphere reserves, all land areas are owned and controlled by the nation's government.

Although no single, land-use-zoning system exists for biosphere reserves, most contain three elements: a core zone, a buffer zone, and a transition zone. Valuable ecological resources are preserved in core zones, which are surrounded by buffer zones and transition zones characterized by increasing intensity of human use. These buffer zones and transition zones may include seminatural ecosystems and agricultural systems. Core zones are securely protected land or marine areas designed to conserve biological diversity, provide locations for monitoring minimally-disturbed ecosystems, and serve as sites for nondestructive research and education. A core zone usually represents a major ecosystem of world significance and is large enough to allow *in situ* conservation of the site's biological diversity. A biosphere reserve's core zone serves as a global benchmark of ecological health, setting the standard for comparison with the effects of human uses occurring elsewhere in the reserve. A clearly defined buffer zone, which surrounds or adjoins the core zones, may be used for activities compatible with sound ecological practices, including environmental education, tourism, and applied research. The buffer zone serves to integrate the biosphere reserve into the geographical region it represents.

The buffer zone may be used for education, training, and manipulative research on ecosystem management, and may include traditional land-use practices such as grazing, fishing, and even timber extraction. Where possible, a second form of buffer zone, the transition zone, surrounds the first for experimentation in alternative land-uses, education, training, and recreation, while simultaneously benefiting local human populations. The flexible transition zone may contain human settlements, agriculture, and other uses in which communities, management agencies, scientists, and other stakeholders work together to sustainably develop the area's resources (UNESCO 1996). In some biosphere reserves, human communities exist in both the buffer zone and transition zone. In other reserves, humans occupy only the transition zone.

Although the zoning system of biosphere reserves originally was envisioned as a series of concentric rings with the core zone in the center, zoning systems have been implemented in various ways, according to local conditions. In some countries, for example, a zone designed for sustainable resource extraction lies between core zones. Nonetheless, most biosphere reserves follow the usual model of having one or more well-protected core zones surrounded by a buffer zone and transition zone.

Because biosphere reserves are intended to emerge from a participatory process with local communities and natural resource harvesters, they have social, as well as spatial, components. Interest groups—or stakeholders—affected by the reserve may be invited to participate in planning the biosphere reserve's design and management. This participation is aimed at gaining community support for the reserve and commitment to its long-term success.

Sites selected by MAB for inclusion in the biosphere reserve network are intended to carry out three basic functions: conservation, development, and logistic support for research, education, and monitoring. More specifically, these three functions are designed to:

(a) Contribute to the conservation of landscapes, ecosystems, species, and genetic variation;

(b) Foster ecologically, socially, and culturally sustainable development; and

(c) Provide logistical support for environmental education, training, demonstration projects, monitoring, and research relating to local, regional, national, and global issues of conservation and sustainable development (Gregg 1999).

UNESCO's MAB program maintains a computerized data base, the MAB Information System, which provides details on the geographical, ecological, and administrative characteristics of the global network of biosphere reserves. The database includes relevant publications and information on research projects within the reserves.

3. *Process of Selecting Biosphere Reserves*

Proposals for new biosphere reserves are sent to the Paris office of UNESCO from the participating country's MAB National Committee or MAB focal point, or (if such entities have not been established) from the country's UNESCO National Commission. Formally established in 1971, the MAB is an intergovernmental United Nations program of research and training aimed at developing an interdisciplinary scientific basis for the conservation and rational use of the natural resources of the global biosphere. The MAB program is designed to provide continuous training of professional conservation personnel and to acquire, transfer, and apply the information needed to manage the world's natural resources. The World Network of Biosphere Reserves forms the backbone of the MAB Program.

Countries proposing biosphere reserve status for natural areas within their boundaries are asked by UNESCO's MAB Secretariat to submit maps and to complete a questionnaire about the site. The ques-

tionnaire allows MAB specialists to evaluate the suitability of the site as a biosphere reserve and supplies information for the network's computerized database. The database allows for the interchange of standardized information about accepted biosphere reserves. Data included in the questionnaire include latitude and longitude of the proposed reserve, minimum and maximum altitudes, total size, status of existing protection, land tenure, physical characteristics, vegetation, fauna, scientific research, and human modifications. MAB's International Coordinating Council, comprising representatives from 34 member states, gives final approval of protected areas for biosphere reserve status. The General Conference of UNESCO elects these representatives on a rotating basis. A few countries (e.g., Mexico and Guatemala) have enacted specific legislation declaring new protected areas as national biosphere reserves, which are later submitted to UNESCO for inclusion in the global network. However, most biosphere reserves encompass areas previously protected under other categories, such as national parks or wildlife reserves. Biosphere reserves may also be recognized through other international designations, such as World Heritage Sites and Ramsar wetland sites.

4. *Characteristics*

Michel Batisse, Senior Science Advisor to UNESCO and one of the pioneers of the biosphere reserve concept, defines the primary characteristics of biosphere reserves as follows (Batisse 1982):

(a) Biosphere reserves are protected areas of land and coast environments; together they should constitute a worldwide network linked by international understanding on purposes, standards, and exchange, of scientific information.

The network of biosphere reserves should include significant examples of biomes throughout the world.

(b) Each biosphere reserve should include one or more of the following: representative examples of natural biomes; unique communities or areas with unusual features of exceptional interest; examples of harmonious landscape resulting from traditional patterns of land-use; and/or examples of modified or degraded ecosystems that are capable of being restored to more-or-less natural conditions.

(c) Each biosphere reserve should be large enough to be an effective conservation unit and to accommodate different uses without conflict.

(d) Biosphere reserves should provide opportunities for ecological research, education, and training; they will have particular value as benchmarks or standards for measurement of long-term changes in the global biosphere as a whole.

(e) A biosphere reserve must have adequate long-term legal protection.

(f) In some cases biosphere reserves will coincide with, or incorporate, existing or proposed protected areas, such as national parks, sanctuaries, or nature reserves.

5. *History of Biosphere Reserves*

The term biosphere reserve emerged during a 1969 UNESCO-sponsored gathering of 80 scientists from 30 countries, aimed at coordinating a worldwide network of protected areas to ensure the conservation of genetic resources. The first official definition of biosphere reserves came in 1970 in a plan proposed to the UNESCO General Conference, which launched the Man and the Biosphere Program. A task force organized by UNESCO in Paris in 1974 formulated criteria and guidelines for the selection and establishment of biosphere reserves. In 1976, the MAB Bureau of the MAB Coordinating Council designated the first cohort of 57 biosphere reserves. By the following year, the network had grown to include 118 biosphere reserves in 27 countries.

During a 1981 international conference sponsored by UNESCO to review the MAB program, delegates focused on the need to emphasize the development and integration of the functions of biosphere reserves, rather than the designation of a maximum number of sites. The conference also set the stage for the first International Biosphere Reserve Congress, held in Minsk, Belarus, in 1983. International specialists at the conference produced the first detailed compilation of biosphere reserve case studies and made recommendations for a program framework. UNESCO used these recommendations to publish its first Action Plan for Biosphere Reserves, adopted by the organization in 1984. The plan recommended a minimum set of activities for implementation in each biosphere reserve, including baseline inventories of flora and fauna, preparation of a history of research, establishment of research facilities and research programs, and preparation of a management plan that addresses the three defined biosphere reserve functions of conservation, sustainable development, and logistic support for research, education, and monitoring (UNESCO 1984).

Through time, the concept of biosphere reserves has been elaborated and clarified in international meetings. In 1982, a MAB Scientific Conference celebrating the tenth anniversary of the MAB program added the phrase 'representative ecological area' as a subtitle to the term biosphere reserve. The biosphere reserve network is designed to include ecosystems that are representative of the major biomes of the world, as defined by a global classification of biomes based on 193 biogeographical provinces belonging to 14 types of biomes within 80 biogeographical realms. The biogeographer Udvardy created this classification at

the request of UNESCO and the International Union for the Conservation of Nature (IUCN) in 1975. During its first 10 years of existence, conservationists promoting the biosphere reserve attempted to include at least one biosphere reserve in each biogeographical province. However, the requirement that participating countries must nominate sites within their borders before they can be included in the network has hindered the goal of full coverage.

In 1995, in Seville, Spain, UNESCO sponsored an international conference of experts to recommend a new strategy for developing effective biosphere reserves. The Seville Strategy (1996) establishes goals, objectives, and actions for implementing the biosphere reserve concept at international, national, and biosphere reserve levels. The Strategy includes a checklist of indicators for use in evaluating progress in implementation. The document also presents a statutory framework of the World Network of Biosphere Reserves, setting forth definitions, functions, selection criteria, and designation and review procedures for biosphere reserves. One new provision in the documents requires governments to submit status reviews of each biosphere reserve every 10 years for comments and recommendations by MAB's International Coordinating Council. The Seville Strategy also highlights the role of biosphere reserves in responding to Agenda 21, an agreement that resulted from the United Nations Conference on Environment and Development (the Earth Summit), in Rio de Janeiro, Brazil, in 1992, as well as the contribution of biosphere reserves to implementation of the Convention on Biological Diversity.

6. *Trends and Criticisms*

From its first list of 57 sites in 1976, the biosphere reserve network grew to include 214 reserves in 58 countries by 1982, and by 2000 to 368 biosphere reserves in 91 countries. New biosphere reserves are added to the network each year. More than 110 countries have joined the program and formed national MAB committees. The development of the biosphere reserve concept helped create a new movement in environmental conservation that focuses on combining the protection of biological diversity with sustainable economic development for human benefit. The link between biological diversity conservation and the development needs of human communities is now recognized as a central component in the successful management of protected areas. Because they are designed to be relevant to human needs, biosphere reserves have helped make the conservation of biological diversity more scientific, more systematic, and more socially and economically acceptable to human populations.

Within biosphere reserves themselves, recent innovations include new methodologies for involving stakeholders in decision-making processes and resolving conflicts. New kinds of biosphere reserves are also evolving, such as cluster and transfrontier reserves, which transcend national boundaries (e.g., the La Amistad Biosphere Reserve in Costa Rica and Panama).

During the late 1990s, information exchange emerged as a priority goal for the network of biosphere reserves. New international networks, facilitated by the Internet and expanded access to computers, are improving communication and cooperation between biosphere reserve managers in different countries. Funding from the Global Environmental Facility (GEF) has linked biosphere reserves in five Eastern European countries via the Internet. The technology firms Intel and NEC have joined with UNESCO and Conservation International to provide communications training and equipment for 25 biosphere reserves in Africa, Asia, and Latin America. The EuroMAB Network sponsors workshops that link biosphere reserves in Europe with reserves in North America. Regional networks have been established for Spanish and Portuguese-speaking biosphere reserve managers in Latin America and for French and English-speakers in Africa. These regional efforts are connected to one another through the UNESCO-MABnet Internet site (http://www.unesco.org/mab).

Some conservationists have criticized the sustainable development aspect of biosphere reserves as 'wishful thinking' (e.g., Terborgh 1999). In his view, project managers who try to invigorate the local economy (e.g., in buffer zones) of protected areas, end up attracting new settlers to a reserve's perimeter, thereby increasing external pressure on the protected area's natural resources. Terborgh warns that such projects 'shoot at the wrong target'—local communities rather than the protected area and its resources (Terborgh 1999).

Other researchers have noted that while biosphere reserves 'look like a politically expedient and socially acceptable way of meeting livelihood and conservation needs,' in reality, the management of biosphere reserves is more socially complicated than that of conventional parks (Brandon 1998). To be successful, this perspective holds, biosphere reserves require agreements with local people to stop using the resources of the core zones, to alter use patterns in the buffer zones, and to agree to abide by a new management authority. At the same time, the biosphere reserve concept facilitates conservation efforts in some protected areas precisely because it emphasizes agreements with local populations in the reserve's design and management.

The 1996 Seville Strategy declared that biosphere reserves are now poised to take on a new role. The rapid growth of the human population and increasing consumption of natural resources are forcing conservationists to search for examples of how to merge conservation and sustainable development. According

to the Seville Strategy, 'Biosphere reserves offer such examples,' becoming theaters for reconciling people and nature by showing the way to a more sustainable future. The Seville Conference concluded that the three basic functions of biosphere reserves—conservation, sustainable development, and logistic support for research, education, and monitoring—will be as valid as ever in future years. The primary challenges to biosphere reserves today are adding new sites to improve world coverage and ensuring that biosphere reserves fulfill their designated functions.

See also: Conservation: Wetlands; Deforestation–Forestation; Environmental Challenges in Organizations; Environmental Planning; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; Environmentalism: Preservation and Conservation; Resource Geography

Bibliography

Batisse M 1982 The biosphere reserve: a tool for environmental conservation and management. *Environmental Conservation*. **9**(2): 101–11

Batisse M 1997 Biosphere reserves: a challenge for biodiversity conservation and regional development. *Environment* **39**(5): 7–33

Brandon K 1998 Perils to parks: the social context of threats. In: Brandon K, Redford K H, Sanderson S E (eds.) *Parks in Peril: People, Politics, and Protected Areas*. Island Press, Washington, DC, Chap.14

Gregg W P 1999 Environmental policy, sustainable societies, and biosphere reserves. In: Peine J D (ed.) *Ecosystem Management for Sustainability: Principles and Practices Illustrated by a Regional Biosphere Reserve Cooperative*. Lewis Publishers, Boca Raton, FL, Chap. 2

Gregg W P, McGean B A 1985 Biosphere reserves: their history and their promise. *Orion Nature Quarterly* **4**(3): 40–51

Halffter G 1985 Biosphere reserves: conservation of nature for man. *Parks* **10**(3): 15–18

Terborgh J 1999 *Requiem for Nature*. Island Press, Washington, DC

UNESCO 1984 Action plan for biosphere reserves. *Nature and Resources* **20**(4): 1–12

UNESCO 1996 *Biosphere Reserves: The Seville Strategy and the Statutory Framework of the World Network*. UNESCO, Paris

Walker R T, Solecki W D 1999 Managing land use and land-cover change: the New Jersey Pinelands Biosphere Reserve. *Annals of the Association of American Geographers* **89**(2): 220–37

J. D. Nations

Biotechnology

In the closing decades of the twentieth century, biotechnology emerged as a site of rapid change in science and technology and as an arena of social and institutional transformation. The development of new techniques for studying, manipulating, and redesigning living things produced important applications in medicine and agriculture, and generated massive investment. In the world of research, biotechnology was often at the forefront of change in scientific institutions and practices. More broadly, biotechnology—widely perceived as having 'revolutionary' implications—inspired both intense enthusiasm and determined opposition. As an area of science and technology with the explicit goal of intervening in the machinery of life, biotechnology often disrupts traditional ways of distinguishing 'nature' from 'culture,' calling into question settled social arrangements and presenting societies with unfamiliar risks, unprecedented ethical dilemmas, and novel opportunities. As a result, biotechnology poses difficult challenges of governance. This article examines the rise of biotechnology and considers its technological and epistemic structure, its institutional dimensions, and its problematic position in contemporary politics.

1. The Term Biotechnology

Defining biotechnology poses challenges, for the word is less a tightly-defined, technical term than a loose umbrella category, or even a slogan, that conveys—sometimes simultaneously—visions of unbounded progress and unregulated tampering with nature. Many authors have tried to capture biotechnology within their own well-crafted definitions, but these attempts cannot neatly contain this expanding network of activities and its increasingly dense connections to diverse social worlds. Although the word has a long history (Bud 1993), in most contemporary contexts biotechnology refers to a novel and growing collection of techniques, grounded in molecular and cell biology, for analyzing and manipulating the molecular building blocks of life. The term also designates products, such as pharmaceuticals or genetically-modified foods, created using these techniques. At times, it refers not to products or techniques but to an economic sector or area of research. Biotechnology acquired these intertwined meanings toward the end of the 1970s, coming into widespread use in the early 1980s, as molecular biology was increasingly understood not only as a 'science' for learning about nature but also as a 'technology' for altering it.

2. The Biotechnology 'Revolution'

One of the ironies of the rise of biotechnology is that a sense of revolutionary potential energizes both the enthusiasm and the opposition it engenders. Supporters and critics alike often fit biotechnology into a narrative of radical discontinuity (e.g., Conway 1997, Kevles and Hood 1992, Rifkin 1983). Biotechnology advocates claim that it will completely transform medicine, spawn entirely new industries, and supply

adequate food to the growing world population. Critics—who warn of unanticipated consequences and the hazards of altering human identity—also cast the new biology as a revolutionary force, arguing that genetic engineering is ushering in a risky new stage in evolution, as nature itself comes under technological control.

Alongside the notion of revolution, the discourse on biotechnology also features an opposing frame that stresses continuity with the biological and historical past. Thus, supporters of biotechnology sometimes downplay its novelty, defining it inclusively (e.g., as the harnessing of biological agents to provide goods and services) and portraying it as merely the latest twist on age-old methods of plant breeding, animal husbandry, and brewing. Such moves help to make the exotic familiar, to suggest that ancient precedents justify apparently novel practices, and to undermine the notion of an untouched 'natural' order that must be protected from human intervention. Critics of biotechnology, for their part, often contend that predictions of unprecedented progress overstate the benefits, suggesting that biotechnology will perpetuate, rather than eliminate, long-standing inequalities in agricultural and health care systems.

As these observations suggest, narratives of continuity and discontinuity provide opposing frames that participants in debates about biotechnology can selectively deploy to mobilize support for their views. But for social scientists, the relevant question is whether the concept of a biotechnology revolution has analytic utility. Although some might be tempted to dismiss this notion as overheated rhetoric, many people in contemporary societies apprehend biotechnology in precisely these terms. Like information technology and computers, biotechnology seems perpetually to stand at the threshold of the next qualitative transformation (see *Science and Technology, Social Study of: Computers and Information Technology*). Biologists who pause to marvel at the pace of progress, like ordinary citizens who reflect on the stream of announcements from the research front, experience biotechnology as an area of ongoing change where old boundaries are continually broken and the unprecedented rapidly becomes the mundane.

2.1 A New Technological and Epistemic Space

Whether framed as revolutionary change or incremental evolution, biotechnology clearly emerged from a complex constellation of phenomena, among them the development of increasingly powerful tools for representing and manipulating the molecular building blocks of life. The development of such techniques as recombinant DNA technology, monoclonal antibodies, in vitro fertilization, the polymerase chain reaction (PCR), and high-throughput DNA sequencing have dramatically extended human control over living organisms. Although the arrival of each of these techniques was widely heralded as 'revolutionary,' the increasing power of biotechnology cannot be attributed to any single tool, but stems from the ability to combine and recombine such tools, creating new assemblages capable of building and taking apart molecules, rearranging the genetic code of organisms, and reading and writing the 'language' of DNA. No less significantly, such techniques have consolidated and accelerated epistemic change in the life sciences, increasing the centrality of metaphors from information theory and cybernetics (e.g., code, control, and text) for understanding living things (Kay 1999, Keller 1995). More precisely, these tools provide practical means that, quite literally, lend substance to these metaphors, giving them material form and making them into epistemic things embedded in experimental systems (Rheinberger 1997) and sociotechnical networks (see *Experiment, in Science and Technology Studies*).

The field of molecular biology expanded rapidly after World War II, and by 1970, scientists knew how genes specify protein structures and had found enzymes that cut, join, and extend DNA molecules. This set the stage for recombinant DNA technology (rDNA)—the emergence of which is often taken as the starting point for modern biotechnology. Recombinant DNA techniques enabled scientists for the first time to insert or delete genes from the genetic code of living things, for example, by removing a piece of DNA from one organism and splicing it into the DNA of a distantly related species. The development of these methods generated great excitement and provoked controversy (discussed below) about the wisdom and safety of genetic engineering. Beyond opening up a wide range of new research strategies, this technology allowed scientists to design genetically modified organisms to produce commercial products. In the late 1970s, researchers began using genetically engineered bacteria as microscopic manufacturing plants for making valuable proteins, such as human insulin and human growth hormone.

In the last two decades of the twentieth century, the power of the biotechnology tool kit continued to increase. The development of monoclonal antibodies, which permitted scientists to construct extremely accurate assays for detecting specific proteins, yielded applications in research and medical diagnostics (Cambrosio and Keating 1995). New methods for transferring genes into developing embryos allowed scientists to alter the genetic code of animals, thus extending genetic engineering from microorganisms to higher forms of life. Parallel techniques enabled scientists to produce transgenic plants, modifying the genomes of agricultural crops to enhance disease resistance, shelf life, and other properties. Molecular biology and biotechnology also played an increasingly central role in the development of pharmaceuticals, including

vaccines, antibiotics, and drugs used to combat heart disease, cancer, and AIDS.

Tools and practices for representing and analyzing DNA molecules were also rapidly developed (Fujimura 1996). Methods for mapping the location of genes grew increasingly powerful; the polymerase chain reaction (PCR) became important in research, diagnostics, and forensics (Rabinow 1996); and DNA sequencing—a technique for representing the nucleotide sequences of DNA molecules as written inscriptions (strings of the letters A, C, G, and T)—rapidly developed. In the 1990s, the human genome project, along with other concerted mapping and sequencing efforts, produced an exponentially increasing volume of biomolecular data on a variety of organisms of medical and agricultural importance (Cook-Deegan 1994, Fortun 1998, Hilgartner 1998). These data—along with increasingly automated means of producing, standardizing, ordering, and analyzing them—constituted the core of genomics, a field that further united information sciences and biology. In effect, by opening up new computational approaches to biology, genome research created a new technological and epistemic 'space,' allowing researchers to perform 'experiments' *in silico* or in a domain linking computational work to the 'wet lab.'

2.2 New Institutional Spaces

Large infusions of capital, promotional state policies, and novel institutional arrangements played a significant role in spurring these developments (Thackray 1998). By the end of the 1970s, government, industry, and academic elites in both Europe and the United States saw investment in 'high technology'—especially computers and biotechnology—as the long-term prescription for economic growth. On both sides of the Atlantic, policymakers sought to spur innovation and speed the translation of basic science into marketable products. The result: a significant shift in the political economy of research that eased the commodification of knowledge (see *Intellectual Property, Concepts of*). Governments encouraged universities and corporations to form new kinds of academic–industry alliances, and the scope of intellectual property protection grew considerably broader, expanding to encompass, for example, genetically engineered organisms (Boyle 1996, Jasanoff 1995a).

In the United States, significant numbers of academic biologists became entrepreneurs for the first time. The most dramatic institutional innovation was the startup company founded by venture capitalists and university professors—many of whom kept their academic posts (Kenney 1986). Increasingly, biotechnology research took place in spaces that were not neatly lodged in either an academic or a corporate milieu, but in complex hybrids of university and industry, public and private, basic and applied (see *Science and Industry*; *Universities and Science and Technology: United States*). Some observers warned that these hybrids would threaten the independence of academic biologists, compromising their ability to offer a credible, critical voice in public decision-making. Ultimately, however, these institutional innovations proved irresistible, given competitive pressures and the scientific and economic opportunities.

Owing to international variation in research systems (see *National Innovation Systems*), the commercialization of biotechnology followed distinct paths in different nations. In Europe and Japan, the large multinational corporations (MNCs) that control production of pharmaceuticals, chemicals, and agricultural inputs built biotechnology directly into their in-house research operations. But in the United States, with its unique pattern of venture capital financing, a separate 'biotechnology industry' composed of free-standing firms took shape (Kenney 1998). The first major wave of biotechnology firms appeared between 1979 in 1981, and during the following two decades, entrepreneurs founded successive waves of firms, whipping up investor enthusiasm around a series of 'revolutions,' including monoclonal antibodies, gene therapy, and genomics.

Even in the United States, however, MNCs—which have poured huge sums into genetic engineering, rational drug design, and genomics—became the dominant players in efforts to exploit the new biology commercially. Indeed, given the structural connections between the biotechnology firms and the MNCs, the image of a separate biotechnology industry is somewhat misleading: Biotechnology startups often derive the bulk of their revenue from contracts with giant companies; MNCs often acquire successful firms; and many biotechnology firms are best understood as research shops that hope to sell intellectual property to large corporations.

3. The Politics of Biotechnology

New institutional spaces have been important not only to the commercialization of biotechnology but also to its politics. Since the early 1970s, biotechnology has posed difficult problems of governance, confronting societies with a stream of potentially controversial issues. More deeply, because biotechnology often seems to undermine basic categories of social order, such as the 'natural' or the 'human,' it tends to pose problems that do not fit neatly into the cosmologies and routines that guide public action. Democratic societies thus face the challenge of building new discursive regimes, regulatory mechanisms, and forums for public discussion that can produce legitimate settlements in this context. In this respect, the institutionalization of risk management and bioethics

have been particularly important. Although the precise techniques for engineering consent have varied across nations (Gottweis 1998, Jasanoff 1995b), states have consistently responded to opposition in ways that allowed the development of biotechnologies to proceed. Despite a cascade of visible controversies, states have so far managed to convert volatile mixtures of concerns into more stable forms. Unstable blends of technical uncertainties, science fiction scenarios, critiques of corporate power, and misgivings about hubris and slippery-slopes, have been discursively distilled into categories amenable to bureaucratic management and expert decision-making (Jasanoff 1995b).

3.1 Risk Management

The process of 'thinning' multidimensional controversies into manageable technical problems susceptible to rqegulatory solutions is well illustrated by what has come to be known as the recombinant DNA debate (see *Risk, Sociology and Politics of*). During the 1960s and early 1970s, scientists and other observers often construed the arrival of genetic engineering—viewed, once again, through the frame of radical discontinuity—as a portentous development that raised many troubling choices for society. As researchers became convinced that molecular techniques would soon make it possible to modify the genomes of living organisms, a number of prominent scientists warned that developments in biology could be dangerously misapplied, and they called for broad public discussion of such possibilities as genetic engineering of humans. However, soon after researchers began splicing genes into bacteria and viruses, the focus of discussion shifted from broad concerns about the long-term implications of genetic engineering to much narrower and immediate questions about the laboratory hazards of recombinant DNA (rDNA) research (Wright 1994). In particular, scientists familiar with the early gene splicing experiments worried about accidentally creating dangerous microorganisms, resulting in catastrophic epidemics or disrupting ecosystems. The problem of providing credible scientific arguments that rDNA research could be safely conducted posed significant challenges given limited evidence and technical debate about the magnitude of the hazards, and in 1974 ten prominent biologists published a call for a partial moratorium on certain rDNA experiments.

The debate about the 'moratorium' and the physical hazards of rDNA research threw the future of genetic engineering into doubt, but it also allowed broader ethical and political questions to slip into the background (Gottweis 1998, Wright 1994). Moreover, scientific experts soon developed a discursive framework that provided a technical strategy for controlling the laboratory risks. This strategy, which took shape at a famous international meeting in Asilomar, California, relied on a system of classification (experiments would be grouped according to their expected level of hazard) and a gradient of controls (e.g., physical barriers, such as negative air pressure, and biological controls, such as 'disarmed' bacterial strains unable to survive outside the laboratory). On both sides of the Atlantic, the strategy of requiring increasingly stringent containment for increasingly risky experiments formed the basis for guidelines or regulatory controls, designed by expert committees, that allowed research to proceed.

Containing the microbes within the laboratory became a means of containing fears and allowing genetic engineers to pursue the unbounded possibilities they believed to lie before them. As the 1970s progressed, these controls were progressively relaxed as a growing number of scientists concluded that the danger of devastating epidemics was remote. Ultimately, the resolution of the controversy inspired a second-order debate about whether scientists, policymakers, and publics had responded appropriately or excessively to the risks of novel organisms. In contemporary debates over biotechnology, the rDNA debate has acquired an almost mythic status: for some, it serves as a cautionary tale about the dangers of irrational fear of technology; for others, it provides a commendable example of scientists taking responsibility for the social implications of their work.

In the 1980s and 1990s, new regulatory challenges emerged as agricultural biotechnologists sought to move genetically modified organisms (GMOs) from the laboratory to the field—first on an experimental basis and later in full-scale production (see *Agricultural Sciences and Technology*). In this context, risk management shifted from containment to deliberate release into the environment, a move that engaged new sources of expertise, such as ecologists, in risk assessment. Fears grew that agricultural biotechnology might accidentally produce new superweeds, alter ecosystems, trap the poor farmers of the developing world in dependency on imported seeds, and expose unsuspecting consumers to the risks of allergic reactions. Although the political culture of GMO regulation differed substantially internationally, in the United States, Germany, France, and the United Kingdom the dominant policy narratives defined the regulatory challenge in terms that allowed development of the technology to proceed, framing agricultural biotechnology as potentially risky but beneficial and indispensable given international competition, defining hazards as amenable to expert solutions, and developing experimental systems for measuring and monitoring risks (Cambrosio et al. 1992, Gottweis 1998, Jasanoff 1995). Whether, and under what conditions, these patterns of accommodation will prove durable, however, remains an open question—especially in some European Union countries, where public trust in GMO regulation to protect consumers and the environment has been lacking.

The global marketing of seeds and the rise of social movements opposed to GM foods in Asia, Latin America, and the United States, also suggest that the transnational dimensions of the politics of biotechnology are likely to remain important.

3.2 Bioethics

The policy narratives and practices of risk management offered one means for translating the unruly politics of biotechnology into governable forms; bioethics offered another—especially regarding human biotechnology and molecular medicine. The potential of the new biology to destabilize, redefine, or deliberately alter the identities of individuals, groups, or 'human nature' often generated apprehension or resistance (see *Human Sciences: History and Sociology*; *Biomedical Sciences and Technology: History and Sociology*). Thus, some observers worried that predictive genetic testing would threaten privacy and deprive people of uncertainty about fate; others feared that genetics would reinforce racial prejudices or produce new stigmatized groups; still others objected to the desacralization and commodification of human life. Such worries, which did not neatly fit into the technical and economistic discourse of risk management, were typically addressed through the interpretive frameworks and institutions of bioethics.

During the 1970s and 1980s, bioethics emerged as an institutionalized area of inquiry and technique for decision-making. The bioethics phenomenon, like the growth of risk management, was by no means confined to biotechnology. In medicine, health care, and clinical research, bioethics—performed in such diverse sites as academic literatures, national commissions, hospital ethics committees, and, most recently, biotechnology corporations—has grown into a prominent feature of public life (DeVries and Subedi 1998, Kleinman et al. 1999, Rothman 1991). The rise of biotechnology contributed to the rise of bioethics, fueling its growth with a cascade of controversies about genetic testing, gene therapy, and reproductive technology—matters that increasingly came to be seen as part of the professional jurisdiction of a new cadre of bioethics experts (see *Science and Technology Studies: Experts and Expertise*). Indeed, by the late 1980s, when the Human Genome Project was proposed, the notion that bioethical analysis could play a major role in addressing the societal dimensions of the new biology was sufficiently widespread that national governments flocked to build 'ethics' into their programs of genomics research (Cook-Deegan 1994).

The peculiar relationship between the domains of bioethics and biotechnology is nicely illustrated by the politics of governmental programs on the ethical, legal, and social implications (ELSI) of genomics. Here, a broad consensus supporting ELSI programs often coexisted with deep disagreement about what these programs should aim to accomplish. The US ELSI program, for example, was celebrated as an important innovation in science policy and presented as proof that social concerns would not be neglected, yet the program was also beset by criticism from many directions. Thus, the ELSI program was accused of being both hypercritical and unduly promotional of genetic technology; of taking an overly academic and long-term view and of narrowly focusing on the immediate; of being too small to address the pace of change and of being a waste of valuable research funds. But such debates about the shape of these programs missed the more pervasive, if subtle, ways that bioethics functions as a quasi-regulatory institution in contemporary states. Bioethics provided an idiom and institutional space for defining controversies about human genetics and genomics as ethical issues, better addressed through moral reflection than mobilization. More deeply, bioethics shored up the legitimacy of political institutions, suggesting that states could find rational and secular means for drawing moral boundaries to replace the apparently natural ones that the biotechnology revolution erased. In this way, bioethics—like other forms of expert advice (Jasanoff 1990, Hilgartner 2000)—has become an important tool not merely for making decisions but also for reordering societies challenged by rapid technological and social change.

See also: Bioethics: Examples from the Life Sciences; Bioethics: Philosophical Aspects; Biomedical Sciences and Technology: History and Sociology; Cultural Evolution: Overview; Health Care Technology; Medicine, History of; Research and Development in Organizations Technological Innovation

Bibliography

Boyle J 1996 *Shamans, Software, and Spleens: Law and the Construction of the Information Society*. Harvard University Press, Cambridge, MA

Bud R 1993 *The Uses of Life: A History of Biotechnology*. Cambridge University Press, Cambridge, UK

Cambrosio A, Keating P 1995 *Exquisite Specificity: The Monoclonal Antibody Revolution*. Oxford University Press, New York

Cambrosio A, Limoges C, Hoffman E 1992 Expertise as a network: A case study of the controversies over the environmental release of genetically engineered organisms. In: Stehr N, Ericson R V (eds.) *The Culture and Power of Knowledge: Inquiries into Contemporary Societies*. Walter de Gruyter, New York

Conway G 1997 *The Doubly Green Revolution: Food for All in the 21st Century*. Cornell University Press, Ithaca, NY

Cook-Deegan R 1994 *The Gene Wars: Science, Politics, and the Human Genome*. W. W. Norton, New York

DeVries R, Subedi J (eds.) 1998 *Bioethics and Society: Constructing the Ethical Enterprise*. Prentice-Hall, Upper Saddle River, NJ

Fortun M 1998 The human genome project and the acceleration of biotechnology. In: Thackray A (ed.) *Private Science: Biotechnology and the Rise of the Molecular Sciences*. University of Pennsylvania Press, Philadelphia, PA

Fujimura J H 1996 *Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer*. Harvard University Press, Cambridge, MA

Gottweis H 1998 *Governing Molecules: The Discursive Politics of Genetic Engineering in Europe and the United States*. MIT Press, Cambridge, MA

Hilgartner S 1998 Data access practices in genome research. In: Thackray A (ed.) *Private Science: Biotechnology and the Rise of the Molecular Sciences*. University of Pennsylvania Press, Philadelphia, PA

Hilgartner S 2000 *Science on Stage: Expert Advice as Public Drama*. Stanford University Press, Stanford, CA

Jasanoff S 1990 *The Fifth Branch: Science Advisers as Policymakers*. Harvard University Press, Cambridge, MA

Jasanoff S 1995a *Science at the Bar: Law, Science, and Technology in America*. Harvard University Press, Cambridge, MA

Jasanoff S 1995b Product, process, or programme: three cultures and the regulation of biotechnology. In: Bauer M (ed.) *Resistance to New Technology: Nuclear Power, Information Technology, and Biotechnology*. Cambridge University Press, Cambridge, UK

Kay L E 2000 *Who Wrote the Book of Life? A History of the Genetic Code*. Stanford University Press, Stanford, CA

Keller E F 1995 *Refiguring Life: Metaphors of Twentieth-Century Biology*. Columbia University Press, New York

Kenney M 1986 *Biotechnology: The University-Industrial Complex*. Yale University Press, New Haven, CT

Kenney M 1998 Biotechnology and the creation of a new economic space. In: Thackray A (ed.) *Private Science: Biotechnology and the Rise of the Molecular Sciences*. University of Pennsylvania Press, Philadelphia, PA

Kevles D J, Hood L (eds.) 1992 *The Code of Codes: Scientific and Social Issues in the Human Genome Project*. Harvard University Press, Cambridge, MA

Kleinman A, Fox R C, Brandt A M 1999 Introduction: Biothics and Beyond. *Daedalus: Journal of the American Academy of Arts and Sciences* **128**(4): vii–x

Rabinow P 1996 *Making PCR: A Story of Biotechnology*. University of Chicago Press, Chicago, IL

Rheinberger H-J 1997 *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford, CA

Rifkin J 1983 *Algeny: A New Word—A New World*. Viking Press, New York

Rothman D J 1991 *Strangers at the Bedside: A History of How Law and Bioethics Transformed Medical Decision Making*. Basic Books, New York

Thackray A (ed.) 1998 *Private Science: Biotechnology and the Rise of the Molecular Sciences*. University of Pennsylvania Press, Philadelphia, PA

Wright S 1994 *Molecular Politics: Developing American and British Regulatory Policy for Genetic Engineering, 1972–1982*. University of Chicago Press, Chicago, IL

S. Hilgartner

Bipolar Disorder (Including Hypomania and Mania)

1. Introduction and Historical Background

Bipolar disorder, also known as manic-depressive illness, is a typically episodic, often severe, and sometimes disabling psychiatric illness characterized by marked changes in mood, thinking, and behavior (Goodwin and Jamison 1990, Sadock and Sadock 2000). Bipolar disorder is a major public health problem, although sometimes under-recognized. Characteristically, its symptomatic changes present in extremes of mania and depression, or with simultaneous excited and depressive features. Although mania and severe depression (melancholia) have been described as distinct conditions since antiquity, the unique aspect of bipolar disorder is their occurrence in the same person over time.

The concepts of mania and melancholia were well established by Hippocratic physicians (c. 400 BC). Aretaeus of Cappadocia (AD 150) is credited with having first proposed the association of mania and melancholia in the same person (Adams 1856). Elaborations of this concept did not appear in the European medical literature until the early 1800s, including descriptions of double or circular insanity by Jules Baillarger and Jean-Pierre Falret in 1854. Emil Kraepelin (1921) later proposed a more systematic, comprehensive, and influential description of manic-depressive insanity based on his observation of many patients with recurrent melancholic episodes and fewer cases with manias as well as depressions.

The contemporary distinction of bipolar (manic and depressive) and unipolar (depressive) forms of recurrent major mood disorders evolved in the writings of Karl Kleist and Karl Leonhard in the 1950s. In the 1970s David Dunner and Ronald Fieve proposed the further subdivision of bipolar disorder into types I and II, based on the presence of mania vs. hypomania, as well as recurrent depression. In the 1980s, Hagop Akiskal proposed an even broader bipolar spectrum disorder, ranging from minor to severe forms (Goodwin and Jamison 1990).

2. Mania and Bipolar Depression

Episodes of mania present with an increased rate of thinking and speaking, increased physical and sexual energy, decreased need for sleep, and changes in appetite. Thinking is typically rapid, pressured, overproductive, and semilogical, often with a grandiose flavor. As a result, attention and concentration are impaired. Manic persons are often distractible, with continuous shifting of topics that are only loosely connected or hard to follow. Attempts to interrupt pressured behavior often provoke irritable or angry responses. Overactivity includes incessant but disorg-

anized and aimless motion, often without clear or well-sustained aims. Dysfunctional behavior derives from grossly excessive self-confidence, and includes lack of inhibition, intrusiveness, impulsivity, recklessness, irritability, and sometimes aggressiveness. Risk taking is common, and may include unrealistic plans or implausible business schemes, impulsive overspending or gambling, disinhibited sexual activity, or reckless driving, with or without abuse of alcohol or illicit drugs.

Depending on the intensity of the mood elevation, two different types of manic episodes occur. Fully expressed mania usually involves disruptive or dangerous behavior requiring emergency treatment, white hypomania is a milder form of sustained, unnatural mood elevation that may not disrupt occupational and interpersonal relationships, and may require intervention or treatment, but rarely on an emergency basis. In hypomania, as well as in mania, judgment is often impaired.

Bipolar disorder also includes major depressive episodes in forms ranging from profound inertia to agitation. Characteristically, bipolar depression presents with generalized psychomotor retardation or slowing of thought and activity, with loss of interest in usual or pleasurable activities and impaired self-care. Along with a low level of energy, lassitude, loss of sexual libido, and abnormalities of sleep and appetite are usually found. Sluggish thinking, impaired concentration, and inability to make decisions are common, and feelings of guilt, worthlessness, and self-blame may be present. Self-assessments usually are severely distorted, unrealistically negative, and pessimistic. Sometimes, somatic complaints such as migraine-like headaches, aching muscles or joints, gastrointestinal, respiratory, cardiac, and other vegetative symptoms, may be more prominent than psychic symptoms. Despair and hopelessness are common, and there is a greatly increased risk of suicide during depressive or dysphoric-agitated phases of the illness, with much less risk during hypomania or mania.

3. *Other Psychopathological Features*

Bipolar disorder can present with mixed manic-depressive states that include both depressive and manic features at the same time or in rapid alternation. Indeed, many episodes of mania and most episodes of mixed states are characterized by striking lability of mood, with rapid shifts from one extreme to the other within minutes or hours.

In addition, psychotic symptoms can be found in perhaps three-quarters of bipolar disorder patients at some time. Bipolar disorder is, arguably, the most frequent form of idiopathic psychotic illness, exceeding even schizophrenia in frequency. Most commonly, the delusions of bipolar disorder are paranoid, with either grandiose or persecutory themes (often pertaining to religious or other fame- or power-based exaggerations), particularly in manic or mixed episodes. Nihilistic or guilty delusions and other pessimistic exaggerations are more common in depressive phases of the illness. Self-referential misinterpretations of random events, as well as auditory hallucinations, may occur in both types of episode. Psychotic features are often consistent with the current dominant polarity of mood, and so may be mood-congruent (olothymic). Alternatively, mood-incongruent delusions may involve impressions that thoughts are being externally controlled, inserted by, or broadcast to other persons. A small minority of patients are described as schizoaffective because they show an excess of psychotic over affective features. They often become chronically ill with severe disability, and such disorders are considered a type of schizophrenia in current nosological systems.

Persons diagnosed with bipolar disorder often show features of other psychiatric syndromes, such as major anxiety disorders, including panic disorder, obsessive-compulsive disorder, or lesser forms of generalized anxiety or particular phobias, and may meet diagnostic criteria for more than one disorder. In addition, abuse of alcohol or drugs occurs often in patients with bipolar disorder, and can severely complicate clinical management and worsen the overall prognosis. Whether substance abuse is a contributing cause of bipolar disorder, or the disorder favors addictive behavior remains unresolved.

Psychodynamic characterizations of patients with bipolar disorder recognize prominent narcissistic features, lack of empathy, instability of interpersonal relationships, use of denial and activity defenses in mania, and guilt-driven self-aggressiveness in depression. Mania has been considered an attempt to avoid or to compensate for depressive phases. Depression has been interpreted as a manifestation of self-aggressiveness in response to the loss of a loved object (psychoanalytic perspective), as a consequence of learned helplessness after repeated unavoidable experiences (developmental-behavioral model), and as an expression of negative self-evaluation activated by stress and learning (cognitive approach). Both polarities share common traits that may include emotional dependence, deficient awareness of internal emotional responses in oneself or others, and a lack of appreciation of differences between actual and idealized situations.

4. *Diagnosis*

The presence of at least one episode of mania is sufficient for a diagnosis of type I bipolar disorder, although most cases involve a series of manic and depressive recurrences over several years. Persons with recurrences of clinical depression and hypomanic episodes are considered to have type II bipolar

disorder. Both types are recognized in the World Health Organization's *International Classification of Diseases* (*ICD-10*, World Health Organization [WHO] 2000) and the American Psychiatric Association's *Diagnostic and Statistical Manual* (*DSM-IV-TR*, American Psychiatric Association [APA] 2000). The latter also recognizes rapid-cycling bipolar disorder, with at least four episodes of mania or depression within 12 months of clinical assessment (Dunner and Fieve 1974, APA 2000).

These standard diagnostic systems also recognize secondary forms of bipolar disorder that may arise in association with exposure to a variety of drugs or toxins, or emerge with various neurological or general medical disorders. Cyclothymia is also recognized as a less severe form of bipolar disorder whose diagnostic criteria are not as well established and overlap with those of type II bipolar disorder as well as certain personality disorders.

5. *Course of the Illness*

The onset of bipolar disorder is often unrecognized as such, and several years may precede diagnosis and appropriate treatment. Many cases start in adolescence, and sometimes before puberty, when their distinction from attention or conduct disorders of childhood is often difficult (Papolos and Papolos 1999). More than half of bipolar cases present initially with an episode of mania or hypomania, particularly in men, whereas initial depressions are more common in women. Women also tend to have more type II bipolar illness and prominent depressive episodes. In some cases, presenting with recurrent depression, mania or hypomania emerge later, sometimes during treatment with an antidepressant medicine, but late emergence of bipolarity is unlikely after three or more depressive episodes.

The natural duration of untreated episodes of mania or bipolar depression is about six to nine months, with approximately yearly recurrences, on average. With modern psychopharmacological treatments, acute episodes of mania last about one to two months, and depression, two to four months. Without treatment, episodes of mania are slightly more common than depressive episodes, but the total time in depressions exceeds that in manias over years of bipolar illness, regardless of treatment (Angst 1978, Tondo et al. 2001).

Single episodes of mania with or without depressions, as well as single episodes of depression with one or more hypomanic or manic episodes occur in small proportions of cases. Instead, episodic recurrences are the rule in adolescent and young adult bipolar disorder, and their course over years tends toward worsening. Initial intervals of more or less normal mood (euthymia) may average two to three years in young adults, but these intervals usually shorten with successive recurrences, leveling off at eight to nine months after six to eight recurrences (Angst 1978). About 15 percent of bipolar disorder patients eventually follow a more or less chronic course.

Pediatric-onset bipolar disorder often does not follow a clearly episodic course, but may instead involve highly unstable mood and behavior over periods of months or years. A pattern of discretely recurring episodes of mania and depression becomes established during adolescence and is usual in young adults (Faedda et al. 1995, Shulman et al. 1995, Papolos and Papolos 1999). Late-onset bipolar disorder also tends to recur with relatively high frequency from the start, and many geriatric cases become more or less continuous or chronic, and may respond unfavorably to treatment (Shulman et al. 1995).

The succession of manic and depressive episodes often follows fairly consistent or repetitive patterns in the same person over time, with manias preceding depressions or vice versa, or with a continuously circular course with no free intervals; fewer cases follow an erratic or unpredictable course (Angst 1978, Kukopulos et al. 1980). Rapid cycling occurs in about 15 percent of cases, but only a fraction of such persons sustain high rates of recurrence over many years. Rapid cycling is somewhat more common among women and persons with type II bipolar disorder. An uncertain proportion of cases of bipolar disorder follows a rather consistent hemicircannual or seasonal course, commonly but not always, with depressions in fall or winter and manias in spring or summer (Rosenthal and Wehr 1987). This recurrence pattern is more likely with greater distance from the Earth's equator, among women and in younger persons.

A majority of bipolar-disorder patients experience some degree of recovery between episodes of acute illness, and euthymic or stable periods often allow them to re-establish previous activities and relationships. Nevertheless, even with modern treatment, bipolar disorder can lead to sustained occupational and interpersonal dysfunction, particularly following youthful onset, psychotic symptoms, and early hospitalization. At least 5 percent of bipolar-disorder patients shift from a classic early course to a schizoaffective disorder marked by chronic psychotic symptoms and severe dysfunction. Poorer outcomes tend to associate with either very early or very late onset age, frequent presentation of mixed episodes and psychotic features, severe and sustained depression, and perhaps with a tendency toward sustained relatively rapid cycling and a relatively large number of untreated episodes (Goodwin and Jamison 1990). An important contribution to long-term outcome in bipolar disorder is the secondary psychological and social consequences of the illness, and its often profound impact on self-esteem, independent functioning, and interpersonal relations.

6. Increased Mortality

Premature mortality is a strikingly common outcome in bipolar disorder, in part owing to its strong association with suicide. More than one-third of bipolar disorder patients make at least one serious suicide attempt, and 15–20 percent of deaths in bipolar disorder patients are due to suicide (Goodwin and Jamison 1990. Suicide attempts have a particularly ominous implication in bipolar disorder (and in severe major depression), in that the ratio of attempts to fatalities is probably less than 5:1, whereas in the general population the ratio is approximately 20:1. The depressive phase of a bipolar disorder carries the highest suicide risk, and accounts for at least three-quarters of all attempts and fatalities. Mixed states with both agitated turmoil and dysphoria are also highly risky, but account for only a minority of suicides associated with bipolar disorder. Suicidal ideation arises from a desire to give up in the face of difficulties that seem otherwise inescapable, or intense anguish to which no end can be imagined other than in death. Risk of completed suicide in bipolar disorder is greater for men than women, and attempts are more frequent among women. Other risk factors include previous attempts, older age, abuse of alcohol or drugs, and probably a family history of psychiatric illness that includes suicide.

In addition to suicide, mortality rates are elevated in bipolar disorder owing to the effects of risk-taking behaviors and associated substance abuse leading to accidental deaths, medical complications of prolonged abuse of alcohol and other substances, and the impact of stress-sensitive cardiovascular and pulmonary disorders, in which mortality is at least three times higher than in otherwise comparable patients lacking bipolar disorder.

7. Epidemiology

Bipolar disorders have a lifetime prevalence ranging from 1.6 percent (for mania) to as much as 2–5 percent of the general population if type II and cyclothymic cases are included. These disorders represent a quarter to a third of all major mood disorders. The annual incidence of type I disorders is at least 0.05 percent. Women have a moderately excessive risk for bipolar disorder (sex ratio about 1.4:1) that is much less than in nonbipolar major depressive disorder. The median age at onset is about 20 years (Goodwin and Jamison 1990, APA 2000). There have also been secular trends for all major mood disorders to be recognized more commonly, and at younger onset ages over the past century. However, the contributions of improved case finding, broadening conceptualizations of the disorders, and the probable impact of substance abuse and other environmental factors to these trends are not clear. Cyclothymic and hyperthymic temperaments, as well as childhood attention deficit hyperactivity disorder may be predisposing factors for bipolar disorder.

8. Biology

The risk of depression, psychosis, or suicide in first-degree relatives of index cases of bipolar disorder is 20–25 percent. The same risk is found in fraternal twins, but it is about 70 percent in identical twins, leading to a monozygotic–dizygotic risk ratio of more than 3:1. Rare adoption studies carried out to evaluate the role of genetic versus environmental factors have found more affective illness and suicides among biological relatives of index cases of major affective disorders than in the adopting family. Despite this substantial epidemiological evidence of a familial contribution to the risk of affective illness, a specific mode of inheritance, let alone a discrete genetic marker or molecular defect specific to bipolar disorder has yet to be identified.

Other biological findings also are far from definitive (Nemeroff 1999, Baldessarini 2000b). Structural and functional brain-imaging techniques are receiving much attention, but the findings remain limited and inconclusive. Older findings of changes in biochemical measurements of stress-sensitive neurotransmitters and neurohormones have proved nonspecific by diagnosis, or incidental to altered general metabolic activity. There is also a tendency toward early appearance of the rapid-eye-movement phase of sleep in mania as well as in severe depression, with other alterations of ultradian or circadian biorhythms, but these and the biochemical markers are mainly state-dependent descriptors, and not necessarily clues to biological causes of the disorder (Nemeroff 1999). Relative cerebral deficiency of serotonin has been associated with aggression and suicide, but this is not specific to bipolar disorder. Cognitive and other neuropsychological deficits can also be identified in bipolar disorder, and some may be sustained between acute episodes of illness (Mayberg et al. 1997).

9. Clinical Management

Mania and severe episodes of bipolar depression require careful clinical management. Manic patients need protection if they show impulsive, hostile, or excessively uninhibited behavior, or if they refuse treatment. Patients need protection from exploitation by others who may take advantage of their hypersexuality and inappropriate generosity in manic or hypomanic states. In extreme cases, patients may need to be committed to specialized hospital units where they can be treated in structured, calm, secure settings. Depressed patients may require hospitalization because of a high risk of suicide or inability to care for themselves. In both cases, early transfer to partial

hospital, day treatment, or intensive outpatient treatment is often feasible.

Psychotherapeutic efforts are much less extensively evaluated in bipolar than in depressive disorders, in part reflecting traditional views that bipolar patients are often less able to accept their condition as a disorder requiring treatment. Nevertheless, psychotherapy methods that have been developed and scientifically evaluated for depressive disorders may also have utility in bipolar disorder when acute symptoms are adequately controlled medically. These include interpersonal psychotherapy, based on improving coping strategies for social and interpersonal relationships. Cognitive-behavioral approaches attempt to modify irrational or exaggerated concepts by encouraging development of more flexible schemas, and rehearsal of new cognitive and behavioral responses. Psychoeducational and group-based or individual supportive treatments are advisable, and usually feasible with medication (Huxley et al. 2000).

Patients and their families also benefit from education about the illness, its treatment, and expected course. Counseling is also useful to plan for the possible impact of the illness on career and financial and family planning. Genetic counseling regarding risks to offspring can be offered, and advice to women to avoid pregnancy during mood-stabilizing treatment (especially with potentially teratogenic anticonvulsants or lithium), as well as anticipation of a high risk of recurrences in the postpartum period, are also necessary.

10. Treatment

Treatment of single episodes of bipolar disorder primarily involves antidepressants for bipolar depression and antimanic antipsychotics or other central depressants (including some anticonvulsants and potent sedatives) for mania (Baldessarini and Tarazi 2001). Common overuse of antidepressants in bipolar depression can have a destabilizing effect in predisposed patients, manifest as an increased rate of mood cycling or sudden switching from depression to mixed, dysphoric states or potentially dangerous manic-psychotic excitement (Goodwin and Jamison 1990). Older (mainly tricyclic or monoamine oxidase inhibitor) antidepressants have a particularly high risk of inducing adverse mood shifts, but such reactions can occur with any mood-elevating drug. Modern antidepressants are more commonly employed now because of their far greater safety on acute overdose with suicidal intent. There is little evidence that antipsychotics or other central depressants used to treat mania can induce depression in bipolar disorder. Electroconvulsive therapy may be a temporary alternative in severe cases, and can be beneficial in both manic and depressive episodes.

Since the main feature of a bipolar disorder is recurrence of acute episodes, the principal aim of treatment is prevention of recurrences of both depression and mania. Lithium salts (carbonate or citrate) are the best established option for this purpose (Cade 1949, Baldessarini and Tarazi 2001). Lithium treatment has a slow onset of action (about ten days) and alone is usually insufficient for treating acute mania. Lithium has complex effects on interneuronal signal transduction and on other molecular regulatory events in nerve cells (Manji et al. 2000). This agent is effective in preventing, delaying, or lessening manic and depressive recurrences in about two-thirds of bipolar disorder patients over many years. Nevertheless, a minority of patients continue to experience some illness each year regardless of the treatment provided. Owing to its limited safety margin, serum concentrations of lithium are monitored regularly, and are safe and effective in the range of 0.6–1.0 mEq/L. Adverse effects include tremor, polyuria, diarrhea, nausea, weight gain, hypothyroidism, and occasional severe impairment of renal function with uremia. Increased tremor, fatigue, dizziness, hyperreflexia, confusion, lethargy, or convulsions indicate intoxication. A serious problem of long-term lithium treatment is that abrupt or rapid discontinuation is often followed by temporarily greatly increased risk for recurrences of mania, depression, or suicidality, even after years of effective treatment (Baldessarini et al. 1999b). Lithium treatment is a very rare psychiatric intervention with compelling evidence of reduction of suicidal behavior.

Potential alternatives and adjuncts to lithium include a growing number of anticonvulsants. A particularly potent and rapid antimanic effect has been found with valproic acid iand its salts. It is also employed empirically for maintenance treatment, although the long-term protective effects of this agent against bipolar depression and mania are not proved (Bowden et al. 2000). Virtually all antipsychotic drugs have short-term antimanic effects, and therefore are employed empirically to treat breakthrough episodes of mania or to manage patients with both affective and psychotic symptoms, but their ability to prevent recurrences of bipolar depression or mania for prolonged periods is not established. Testing of additional anticonvulsant, antipsychotic, antidepressant, and other experimental agents for the treatment of bipolar disorder is ongoing. However, none has yet proved superior to lithium for long-term management of depression and suicidal behavior as well as mania. Sometimes, combinations of two or more treatment options are employed when a monotherapy proves inadequate (Baldessarini and Tarazi 2001). Finally, cost-effective group, family, and individual psychosocial and rehabilitative interventions are increasingly employed, and have some research support (Huxley et al. 2000).

See also: Bleuler, Eugen (1857–1939); Childhood Depression; Depression; Depression, Clinical Psycho-

logy of; Depression, Hopelessness, Optimism, and Health; Freud, Sigmund (1856–1939); Kraepelin, Emil (1856–1926); Mania; Schizophrenia and Bipolar Disorder: Genetic Aspects

Bibliography

Adams F 1856 *The Extant Works of Aretaeus, the Cappadocian*. Sydenham Society, London

Angst J 1978 The course of affective disorders II. Typology of bipolar manic-depressive illness. *Archiv. für Psychiatrie Nervenkranken*. **226**: 65–73

American Psychiatric Association 2000 *Diagnostic and Statistical Manual*, 4th edn., text revision (*DSM-IV-TR*). American Psychiatric Association, Washington, DC

Baldessarini R J 2000b American biological psychiatry and psychopharmacology 1944–1994, 1st edn. In: Menninger R W, Nemiah J C (eds.) *American Psychiatry After World War II (1944–1994)*. American Psychiatric Press, Washington, DC, Chap. 16, pp. 371–412

Baldessarini R J, Tarazi F I 2001 Drugs and the treatment of psychiatric disorders: Antipsychotic and antimanic agents. In: Hardman J G, Limbird L E, Molinoff P B, Ruddon R W, Gilman A G (eds.) *The Pharmacological Basis of Therapeutics*, 10th edn. McGraw-Hill, New York, Chap. 20

Baldessarini R J, Tondo L, Viguera A C 1999b Effects of discontinuing lithium maintenance treatment. *Bipolar Disorders* **1**: 17–24

Bowden C L, Calabrese J R, McElroy S L, Gyulai L, Wassef A, Petty F, Pope H G Jr, Chou J C, Keck P E Jr, Rhodes L J, Swann A C, Hirschfeld R M A, Wozniak P J 2000 Randomized, placebo-controlled 12 month trial of divalproex and lithium in bipolar treatment of outpatients with bipolar I disorders. *Archives of General Psychiatry* **57**: 481–9

Cade J F J 1949 Lithium salts in the treatment of psychotic excitement. *Medical Journal of Australia* **2**: 349–52

Dunner D L, Fieve R R 1974 Clinical factors in lithium carbonate prophylaxis failure. *Archives of General Psychiatry* **30**: 229–33

Faedda G L, Baldessarini R J, Suppes T, Tondo L, Becker I, Lipschitz D S 1995 Pediatric-onset bipolar disorder: A neglected clinical and public health problem. *Harvard Review of Psychiatry* **3**: 171–95

Goodwin F K, Jamison K R 1990 *Manic-Depressive Illness*. Oxford University Press, New York

Huxley N A, Parikh S V, Baldessarini R J 2000 Effectiveness of psychosocial treatments in bipolar disorder: State of the evidence. *Harvard Review of Psychiatry* **8**: 126–40

Kraepelin E 1921 *Manic-depressive Insanity and Paranoia* [ed. Robertson G M, trans Barclay R M]. E and S Livingstone, Edinburgh, UK

Kukopulos A, Reginaldi D, Laddomada P, Floris G, Serra G, Tondo L 1980 Course of the manic-depressive cycle and changes caused by treatments. *Pharmakopsychiatrie Neuropsycho-pharmakologie* **13**: 156–67

Manji H K, Bowden C L Belmaker R H (eds.) 2000 *Bipolar Medications: Mechanisms of Action*. American Psychiatric Press, Washington, DC

Mayberg H S, Mahurin R K, Brannan S K 1997 Neuropsychiatric aspects of mood and affective disorders. In: Yudofsky S C, Hales R E (eds.) *American Psychiatric Press Textbook of Neuropsychiatry*, 3rd edn. American Psychiatric Press, Washington, DC, Chap. 21, pp. 883–902

Nemeroff C B (ed.) 1999 Mood disorders. In: Charney D S, Nestler E J, Bunney B S (eds.) *Neurobiology of Mental Illness*. Oxford University Press, New York, pt. 4, 291–435

Papolos D, Papolos J 1999 *The Bipolar Child*. Broadway Books, New York

Rosenthal N E, Wehr T A 1987 Seasonal affective disorders. *Psychiatric Annals* **15**: 670–74

Sadock B J, Sadock V A (eds.) 2000 *Kaplan and Sadock's Comprehensive Textbook of Psychiatry*, 7th edn. Lippincott Williams and Wilkins, Philadelphia

Shulman K, Tohen M, Kutcher S P (eds.) 1995 *Bipolar Disorder Through the Life-cycle*. John Wiley, New York

Tondo L, Baldessarini R J, Floris G 2001 Lont-term effectiveness of lithium maintenance treatment in types I and II bipolar disorders. *British Journal of Psychiatry* **178**(suppl. 40): 1–7

World Health Organization 2000 *Manual of the International Statistical Classification of Diseases, Injurie, and Causes of Death (ICD-10)*, 10th edn. World Health Organization, Geneva, Switzerland

R. J. Baldessarini and L. Tondo

Birdsong and Vocal Learning during Development

Songbirds are one of the few groups of organisms other than humans that learn sounds for vocal communication during development, and the neural substrate that controls vocal learning and behavior is highly localized, making these circuits amenable to experimental manipulation and analysis. The song system is a model one for studying the neural and hormonal bases of vocal learning, and for examining plasticity of structure and function in neural systems generally. Similarities between vocal learning in birds and humans suggest that many aspects of the learning process have evolved to meet demands imposed by vocal communication (Marler 1976, Doupe and Kuhl 1999). Therefore, an understanding of the mechanisms underlying vocal learning in songbirds may have immediate application to problems in human vocal learning. In this review I concentrate on what is known about neural and hormonal mechanisms of vocal learning in zebra finches (*Taeniopygia guttata*), a well-studied species in which juvenile males gradually learn to produce a specific vocal pattern during a sensitive period of development, and by adulthood are normally incapable of altering that pattern or learning new ones.

1. A Brain–Behavior System

Zebra finches learn the sounds used for vocal communication during a sensitive period of development. Specific auditory experience is necessary for vocal

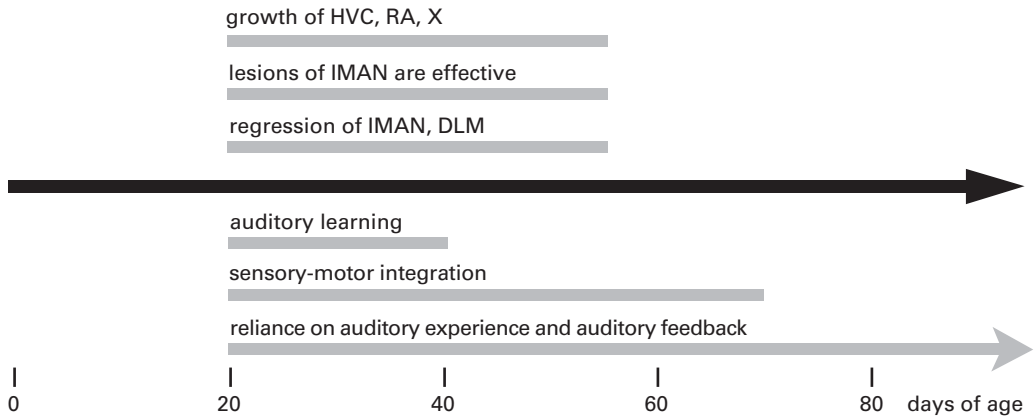

Figure 1
A time line of zebra finch vocal development showing approximate timing of some aspects of song learning. Birds fledge from the nest around 20 days of age and start singing soon thereafter. They are dependent on their parents to feed them up until 35–40 days of age. The exact timing of auditory learning from an external model (tutor song) and reliance on auditory feedback of self-produced sounds are not completely known. Juvenile birds use auditory feedback to refine subsong until it matches the template of the tutor song. Once song learning is complete, adult birds also require auditory feedback to maintain production of stable, stereotyped song. Large-scale growth and regression of brain regions involved with vocal learning and behavior occurs during the sensitive period (see text)

learning to proceed normally: juveniles must hear a 'tutor' song model (normally their father) from approximately 20 to 50 days of age in order to reproduce an accurate copy of that song later in development (e.g., Böhner 1990). Zebra finches begin to produce their first song-related vocalizations around 30–35 days, and gradually refine their utterances until they achieve a close match to the external tutor sounds they have heard earlier (Fig. 1). During this period of auditory–motor integration, birds must hear their own vocalizations: auditory feedback is necessary for juvenile males to learn to adjust the motor patterns giving rise to vocal output so that the latter gradually comes to match the tutor song (Price 1979). By 80–90 days, zebra finches are sexually mature and produce a stereotyped song pattern that is maintained without changes throughout adult life. Once a stereotyped song is produced, adults also rely on auditory feedback to maintain that stable song pattern (Nordeen and Nordeen 1992). Surprisingly, although it has been known for many years that there is a sensitive period for hearing external auditory sounds during early stages of vocal learning, it is not known whether there is also a sensitive period for hearing self-produced sounds (i.e., for experiencing auditory feedback).

Song behavior is highly sexually dimorphic: normal adult females never sing, even if treated with testosterone. However, females treated with sex hormones soon after hatching produce male-typical songs as adults (Pohl-Apel 1985, Simpson and Vicario 1991). The behavioral dimorphism reflects a profound neural sex difference: the sizes of cortical song-control regions are several times larger in males than in females (Nottebohm and Arnold 1976). The neural substrate for vocal learning and behavior is highly localized in males, comprising several interconnected brain regions (Fig. 2) that are highly amenable to investigation of structure–function relationships. For example, HVC and RA are part of a direct efferent system linking the cortex with descending motor circuitry that activates vocal musculature, and lesions of this pathway disrupt stereotyped song behavior in adult males (Nottebohm et al. 1976, Simpson and Vicario 1990, Wild 1997). The activity of HVC neurons in awake, singing birds correlates precisely with production of specific song syllables, indicating that the HVC → RA pathway is part of the on-line circuitry for producing learned vocal patterns in adult birds (McCasland and Konishi 1981). Furthermore, the activity of RA neurons in singing birds correlates with subsyllabic elements of vocal production, whereas HVC neurons apparently encode larger chunks of information, indicating a hierarchical organization in the neural control of vocal behavior (Yu and Margoliash 1996, Vu et al. 1994).

There is also a multirelay route from HVC to RA: a distinct population of neurons in HVC projects to Area X in the basal ganglia, which relays through the thalamic nucleus DLM to the cortical region lMAN, which projects onto the same RA neurons that receive inputs from HVC. RA-projecting neurons in lMAN send an axon collateral to Area X, which creates a forebrain loop connecting X, DLM, and lMAN (Vates and Nottebohm 1995). Lesions of the X → DLM → lMAN pathway have no effect on song production in adult birds, but profoundly disrupt song behavior in

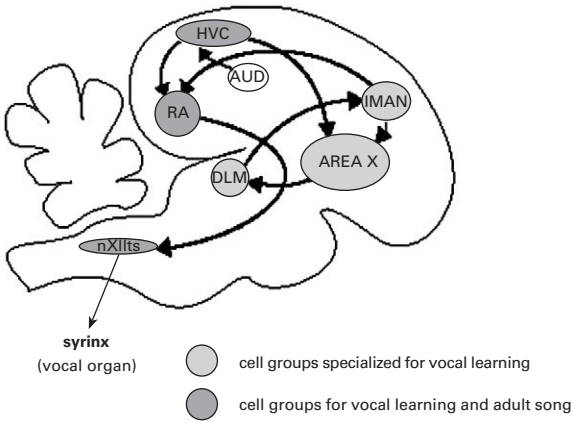

Figure 2
A side view of adult male zebra finch brain showing a highly simplified schematic of the major circuits controlling vocal learning and behavior. HVC contains two separate populations of projection neurons, one sends axons to RA and the other to Area X. The HVC → RA pathway regulates production of already-learned songs in adult birds, and is assumed to be involved with later stages of song learning in juveniles. The HVC → X → DLM → lMAN → RA pathway is necessary for normal song production during early stages of vocal learning, but is clearly not on the main motor pathway for vocal behavior in older juveniles and adults (see text). RA projects to the motor neurons (nXIIts) that activate the vocal muscles.
Abbreviations: lMAN (lateral magnocellular nucleus of the anterior neostriatum), X (Area X of the avian basal ganglia), HVC (high vocal center), AUD, auditory cortex, RA (robust nucleus of the archistriatum), DLM (medial dorsolateral nucleus of the thalamus), nXIIts (tracheosyringeal portion of the hypoglossal nucleus)

juveniles during early stages of vocal learning (Bottjer et al. 1984, Sohrabji et al. 1990, Scharff and Notte bohm 1991). There is a dramatic decrease in the effectiveness of lMAN lesions in juvenile males around 55–60 days of age, which seems to correlate with the development of song as a motor pattern in the sense that the temporal sequence of notes becomes fairly regular at this point. These results suggest that lMAN circuitry is necessary for learning during early stages of vocal development, but may play no role in stereotyped vocal production. The decrease in the effectiveness of lesions within this pathway suggests a change in its function, whatever that may be. One possibility is that the function subserved by this circuit is either no longer important for vocal learning and production, or is taken over by other song-control circuits, such as the HVC → RA pathway. However, an alternative interpretation of the developmental decline in the ability of lMAN lesions to disrupt behavior is that the function of this pathway is conserved, but somehow masked or suppressed at later stages of vocal learning (see below). Either way, this pathway may be similar to mammalian basal ganglia pathways in terms of its involvement in motor aspects of learning, planning, and coordination of movement and motivation (cf. Alexander 1994, Graybiel et al. 1994, Bottjer and Johnson 1997).

2. *Steroid Hormones and Vocal Development*

Much evidence suggests that sex hormones are involved in regulating one or more aspects of behavior during the sensitive period for song learning (Bottjer and Johnson 1997). Blocking the action of androgens in juvenile zebra finches prevents the development of stereotyped vocal behavior (but has no effect in adults). That is, syllable order fails to become stereotyped in juvenile zebra finches that are castrated and treated with anti-androgens, such that the overall structure of song is unstable and abnormal. Interestingly, subsequent exposure to physiological levels of testosterone (T) in adulthood induces stereotyped song production over a period of weeks, suggesting that the final stages of auditory-motor learning can be carried out in delayed fashion upon exposure to high levels of T. The enhanced behavioral and neural plasticity associated with the sensitive period for vocal learning appears to depend, at least in part, on the relatively low levels of circulating hormones that normally occur during early stages of juvenile development. In accord with the idea that reduced levels of plasma testosterone are actually necessary for normal song development, chronic exposure to normal adult levels of T during juvenile development severely impairs vocal learning in zebra finches. Juvenile males receiving systemic T exposure at the onset of vocal learning (20–40 days) develop fewer song syllables than controls, suggesting that the ability to memorize tutor syllables is impaired by exposure to adult levels of T during this interval. These data indicate that the proper timing and amount of exposure to hormones are important factors in the normal development of vocal behavior. Because many song-control neurons contain hormone receptors (particularly in HVC and lMAN), it is likely that T may exert direct effects on song-control circuitry to alter functional properties of neurons during vocal learning.

One possible speculation suggested by this pattern of results is that blocking the action of androgens extends a sensitive period for vocal learning, whereas normal exposure to rising titers of androgens closes one or more windows of opportunity for learning. This pattern of results has strong parallels to studies of human language acquisition, which have shown that the capacity for language learning decreases sharply following puberty (e.g., Johnson and Newport 1989).

It is also known that the sensitive period for song learning can be extended by rearing zebra finches in isolation. Although birds that remain isolated produce highly abnormal vocalizations, birds that are provided with tutors in adulthood can learn some song syllables long after the normal sensitive period has ended (Eales 1987, Morrison and Nottebohm 1993). Interestingly, juvenile birds that have been socially and acoustically isolated have abnormally low levels of T, although hormonal titers recover by the time birds reach late stages of song learning (Livingston et al. 2000). Thus the extended capacity for learning produced by isolation may be due to a prolonged period of low hormonal titers, although lack of experience may also be a contributing factor. It is possible that any conditions that prevent learning tend to keep the sensitive period open. This idea suggests the hypothesis that the loss of plasticity associated with the closing of the sensitive period is due to learning; if vocal learning is prevented then the sensitive period remains open.

3. *Developmental Changes in Song-control Nuclei: Growth and Regression*

The neural substrate for song control undergoes large-scale morphological changes during the period of vocal learning in juvenile male zebra finches (Nordeen and Nordeen 1997, Bottjer and Arnold 1997). Substantial (e.g., threefold) changes in the overall size of song-control nuclei can include changes in neuron number and/or size and complexity of axon terminals and dendritic arbors. For example, HVC and RA grow substantially in males during song learning, more than doubling in volume, whereas lMAN shows substantial regression (Fig. 1). Regression of lMAN correlates with the loss in effectiveness of lMAN lesions, whereas the growth of HVC and RA are consistent with the idea that the HVC → RA pathway assumes greater control of vocal behavior as song learning progresses. The profound growth and regression of song-control brain regions suggest highly dynamic roles for vocal circuitry in one or more aspects of song learning, and must be accompanied by substantial remodeling of synaptic connectivity. An interesting feature of these changes is that their occurrence during development is delayed relative to circuits that are not related to song control. Thus, the major neural scaffold controlling song behavior is created as song is being learned.

The gross changes in volume of brain regions observed during song learning suggested the possibility that experiential factors might play a role in inducing such morphological changes in the substrate for song, such that these large-scale changes actually reflect vocal learning. However, the volume and neuronal number of telencephalic song-control nuclei is the same in hearing and deaf birds, showing that auditory experience has no influence on regulation of neuron number, and that normal patterns of growth and regression occur even in birds that do not engage in normal song learning (Burek et al. 1991). This result indicates that gross morphological changes in song-control circuits are in fact *not* a product of the learning process, but rather may be prerequisite to learned changes in behavior. In accord with this idea, Brenowitz et al. (1995) demonstrated that tutoring two groups of marsh wrens with either many syllables or few syllables led to the development of large versus small learned song repertoires, respectively, although the size (and neuronal number) of HVC and RA in the two groups did not vary. However, within the group tutored with many syllables, only birds with a relatively large HVC learned large repertoires. Thus, a large HVC may be permissive for a larger vocal repertoire, but is clearly not a consequence of such. It seems likely that the basic neural scaffold underlying vocal learning is dependent on innate factors such as endogenous levels of hormones and cytokines (Johnson et al. 1997, Rasika et al. 1999). However, the idea that large-scale changes in song-control nuclei are a necessary prerequisite for vocal learning has not been tested directly (i.e., the occurrence of these changes during the sensitive period for vocal learning is correlational; cf. Bottjer and Arnold 1997).

4. *Neural Signatures of Learning and Development*

If large-scale patterns of growth and regression in the neural substrate for vocal learning are not a reflection of learning, then what changes in the brain signify learning and the close of the sensitive period? Several studies have described normative developmental changes that occur during the sensitive period, and may underlie specific aspects of the learning process. For example, the axonal projections of the DLM → lMAN → RA pathway show exuberant growth during early stages of song learning followed by pruning (Nixdorf-Bergweiler et al. 1995, Iyengar et al. 1999). In addition, NMDA (*N*-methyl-D-aspartate) receptors decrease in density within lMAN during song learning, and the duration of NMDA-R-mediated synaptic currents in individual lMAN neurons becomes significantly shorter (Aamodt et al. 1995; Livingston and Mooney 1997). These changes indicate that NMDA-R carry a greater proportion of the synaptic current in juveniles than adults. NMDA-R figure prominently in activity-dependent adjustment of synaptic connections: because they require both a ligand (glutamate) and depolarization of the post-synaptic cell, they serve as molecular detectors of coincident pre- and post-synaptic activation and therefore provide a mechanism for Hebbian learning (e.g., Cline 1991). Thus, it is

possible that NMDA-R in lMAN neurons could detect correlated activity of auditory and motor aspects of song to reinforce synapses that correspond to the acoustic structure of tutor song. In general, highly precise patterns of synaptic connectivity as well as the behaviors they subserve are established during sensitive periods guided by activity-dependent mechanisms based on sensory experience (Knudsen 1999), so it therefore seems likely that experience-dependent synaptic rearrangements in song-control circuits, such as refinement of the DLM → lMAN → RA pathway, may underlie the emergence of specific vocal patterns.

If developmental changes in synaptic rearrangements and NMDA-R currents reflect learning, then one prediction is that they should be experience-dependent. This is at least partially true: isolation of juvenile birds to prolong the sensitive period (see above) does delay the shortening of NMDA-R currents, as well as loss of the modulatory subunit that gates the corresponding slow kinetics normally seen in juveniles (White et al. 1999, Singh et al. 2000). In addition, the number of dendritic spines on lMAN neurons decreases during song learning in normal birds, but remains high in age-matched birds that have been reared in isolation (Wallhaüser-Franke et al. 1995), and the exuberant growth of the lMAN → RA connection is prolonged in birds that have been deafened or raised in white noise (Iyengar et al. 1999). However, isolation does not prevent these normative changes permanently. For example, although NMDA-R currents are longer in isolate-reared birds compared to normals at early stages of song learning, they recover to normal levels at later stages (Livingston et al. 2000). Thus, although lMAN neurons of isolate-reared birds seem to remain in a more juvenile state initially, they eventually achieve normal levels despite continued isolation. Furthermore, because isolated birds can learn some new syllables as adults, this pattern indicates that the enhanced NMDA-R currents seen in normal juveniles during the sensitive period are not necessary for at least some types of learning to occur. However, it is possible that the greater contribution of NMDA-R to synaptic currents in young birds could subserve some early aspect of learning that is necessary for later song development. For example, longer NMDA-R currents might enable young birds to learn an association between specific articulatory gestures and the vocal sounds they produce.

One way to assess this latter idea would be to prevent the longer NMDA-R currents that normally occur at the onset of the sensitive period. Exposing juvenile males to adult levels of testosterone starting at 20 days severely disrupts vocal learning and causes premature shortening of NMDA-R currents as well as an accelerated loss of the slow modulatory subunit of the NMDA-R (White et al. 1999, Singh et al. 2000). This pattern is consistent with the idea that enhanced NMDA-R synaptic currents are necessary for some, perhaps unknown, aspects of vocal learning. The delay seen in several neural parameters caused by experiential and hormonal manipulations may provide neural correlates that contribute to the extended sensitive period for song learning in isolated zebra finches. Furthermore, the relative decrease in the contribution of NMDA-R currents could be responsible for closing one or more aspects of the sensitive period.

One particularly interesting neural signature of vocal learning is the development of auditory tuning in song control neurons (Doupe and Solis 1997, Margoliash 1997). For example, lMAN neurons in adult birds that project to RA are selectively responsive to playback of each bird's own song, and this selectivity emerges at about 60 days of age, around the time when birds are beginning to produce a relatively stable song pattern and lMAN lesions are losing the ability to disrupt vocal behavior (Solis and Doupe 1999, Rosen and Mooney 2000). This correlation suggests that the development of auditory tuning coincides with a decreased role for lMAN neurons in vocal development, and is consistent with an idea raised above—namely that learning itself acts to curtail the sensitive period. Perhaps auditory selectivity of lMAN neurons reflects the fact that the bird has achieved a match to the tutor song (or developed their own individual template), such that lMAN no longer contributes actively to regulating the song pattern. Indeed, Solis and Doupe (2000) found that birds that were prevented from developing a good match to the tutor song (via damage to the vocal motor nerve) showed lower auditory selectivity in Area X than did birds that were able to mimic the tutor song. Thus, it seems as if the development of selectivity to bird's own song is dependent on learning, and perhaps a reflection of it. As indicated above, refinements of axonal connectivity such as that seen in the DLM → lMAN → RA pathway may represent an important neural signature of learning: remodeling of synaptic connections in this pathway may be necessary for the increased auditory selectivity of lMAN neurons.

5. *Recent Findings*

As indicated above, adult male zebra finches normally are incapable of altering their previously learned, stereotyped vocal pattern. This behavioral pattern correlates well with neural data showing that lesions of the X → DLM → lMAN pathway are effective only during a restricted period of vocal learning, and not in adult birds. These results encouraged hypotheses of vocal learning that stressed the idea that 'age-limited learners' such as zebra finches develop a fixed central motor program, and that the basal ganglia-thalamo-cortical circuit from X to DLM to lMAN is no longer able to regulate behavioral and neural plasticity in

adult birds. However, traditional hypotheses have been challenged by recent data. (a) Exposing adult birds to delayed auditory feedback causes gradual deterioration of stable song, but original song patterns may be recovered when normal auditory feedback is restored (Leonardo and Konishi 1999). (b) Although lesions of Area X and lMAN do not impair adult song production, the act of producing 'undirected' song (i.e., not directed to a female) induces strong expression of the immediate early gene *zenk* in X and lMAN of adult males (Jarvis et al. 1998), and multiunit activity recorded in lMAN and X is consistently higher during production of undirected song (Hessler and Doupe 1999). (c) Although lesions of lMAN in normal adults do not cause behavioral disruption, lMAN lesions in adult birds that are not producing a normal stereotyped song pattern do produce changes in song behavior, and prevent delayed learning from a tutor in isolate-reared birds (Morrison and Nottebohm 1993). (d) Williams and Mehta (1999) demonstrated that vocal plasticity induced by injury to the vocal motor nerve (N XIIts) in adult birds is blocked by lesions of lMAN, and similarly Brainard and Doupe (2000) showed that lMAN lesions in adult deafened birds prevent gradual deterioration (i.e., the song remains stereotyped despite the absence of auditory feedback, suggesting that lMAN may monitor auditory feedback and respond if an error signal is detected; if lMAN is gone, the error signal produced by altered auditory feedback would not be detected in deafened birds).

In the aggregate, this pattern of findings suggests that the X → DLM → lMAN pathway is not irrevocably 'shut down' in adults after vocal learning is complete, despite the fact that lesions of this pathway are ineffective in normal adults. Furthermore, one or more functions of this circuit may be 're-invoked' by certain types of changes in learned song behavior. In addition, these results encourage the idea that adult zebra finches might be capable of new learning following the sensitive period, if appropriate conditions can be found (e.g., if the bird 'unlearns' his initial song pattern; Zevin et al. 2000). Loss of a stable vocal pattern may be a prerequisite to (re)inducing a greater degree of neural and behavioral plasticity in adulthood, perhaps by 'unmasking' an apparently latent function of lMAN circuitry that may be conserved in adult birds.

6. Conclusions and Questions

Vocal learning, like other developmentally acquired skills, is shaped by the interplay of innate factors and experiential influences during a sensitive period. The substantial growth and regression of song-control nuclei that occur during the period of vocal learning are not dependent on experience, and are likely regulated by innate factors that specify the amount of brain space allocated to vocal learning, thereby influencing how much can be learned (Nottebohm et al. 1981). Experience-independent specification of the basic neural scaffolding for vocal learning may also provide specific innate constraints on vocal learning, such as predispositions to hear or produce certain vocal sounds (Marler and Pickert 1984, Marler and Sherman 1985; cf. Seidenberg 1997). In addition, the fact that large-scale patterns of growth and regression occur as learned vocal behavior is being acquired may set temporal constraints on learning. Perhaps this is why humans, like songbirds, are much more adept at vocal learning as juveniles than as adults. If large-scale neural changes are necessary concomitants of vocal learning, then we must ask how experiential factors influence the fine scale of song-control circuitry during such changes. Is it possible that certain types of experience-dependent mechanisms, such as refinement of specific synaptic patterns, can operate only in the context of the overall growth and regression that occurs during song learning? Answers to such questions will begin to provide a better mechanistic understanding of sensitive periods in general.

See also: Animal Cognition; Cerebellum: Associative Learning; Classical Conditioning, Neural Basis of; Communication and Social Psychology; Comparative Neuroscience; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Neural Plasticity; Sensitive Periods in Development, Neural Basis of; Speech Production, Neural Basis of

Bibliography

Aamodt S M, Nordeen E J, Nordeen K W 1995 Early isolation from conspecific song does not affect the normal developmental decline of N-methyl-D-aspartate receptor binding in an avian song nucleus. *Journal of Neurobiology* **23**: 76–84

Alexander G E 1994 Basal ganglia-thalamocortical circuits: their role in control of movements. *Journal of Clinical Neurophysiology* **11**: 420–431

Böhner J 1990 Early acquisition of song in the zebra finch. *Animal Behavior* **39**: 369–374

Bottjer S W, Arnold A P 1997 Developmental plasticity in neural circuits for a learned behavior. *Annual Review of Neuroscience* **20**: 459–481

Bottjer S W, Johnson F 1997 Circuits, hormones, and learning: Vocal behavior in songbirds. *Journal of Neurobiology* **33**: 602–18

Bottjer S W, Miesner E A, Arnold A P 1984 Forebrain lesions disrupt development but not maintenance of song in passerine birds. *Science* **224**: 901–3

Brainard M S, Doupe A J 2000 Interruption of a basal ganglia-forebrain circuit prevents plasticity of learned vocalizations. *Nature* **404**: 762–6

Brenowitz E A, Lent K, Kroodsma D E 1995 Brain space for learned song in birds develops independently of song learning. *Journal of Neuroscience* **15**: 6281–6

Burek M J, Nordeen K W, Nordeen E J 1991 Neuron loss and addition in developing zebra finch song nuclei are independent

of auditory experience during song learning. *Journal of Neurobiology* **22**: 215–23

Cline H 1991 Activity-dependent plasticity in the visual systems of frogs and fish. *Trends in Neurosciences* **14**: 104–11

Doupe A J, Kuhl P K 1999 Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience* **22**: 567–631

Doupe A J, Solis M M 1997 Song- and order-selective neurons develop in the songbird anterior forebrain during vocal learning. *Journal of Neurobiology* **33**: 694–709

Eales L A 1987 Song learning in female-raised zebra finches: Another look at the sensitive phase. *Animal Behavior* **35**: 1356–65

Graybiel A M, Aosaki T, Flaherty A W, Kimura M 1994 The basal ganglia and adaptive motor control. *Science* **265**: 1826–31

Hessler N A, Doupe A J 1999 Social context modulates singing-related neural activity in the songbird forebrain. *Nature Neuroscience* **2**: 209–11

Iyengar S, Viswanathan S S, Bottjer S W 1999 Development of topography within song control circuitry of zebra finches during the sensitive period for song learning. *Journal of Neuroscience* **19**: 6037–57

Jarvis E D, Scharff C, Grossman M R, Ramos J A, Nottebohm F 1998 For whom the bird sings: Context-dependent gene expression. *Neuron* **21**: 775–88

Johnson F, Hohmann S E, DiStefano P S, Bottjer S W 1997 Neurotrophins suppress apoptosis induced by deafferentation of an avian motor-cortical region. *Journal of Neuroscience* **17**: 2101–11

Johnson J S, Newport E L 1989 Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* **21**: 60–99

Knudsen E 1999 Early experience and critical periods. In: Zigmond M J, Bloom F E, Landis S C, Roberts J L, Squire L R (eds.) *Fundamental Neuroscience*. Academic Press, San Diego, CA, pp. 637–54

Leonardo A, Konishi M 1999 Decrystallization of adult birdsong by perturbation of auditory feedback. *Nature* **399**: 466–70

Livingston F S, Mooney R 1997 Development of intrinsic and synaptic properties in a forebrain nucleus essential to avian song learning. *Journal of Neuroscience* **17**: 8997–9009

Livingston F S, White S A, Mooney R 2000 Slow NMDA-EPSCs at synapses critical for song development are not required for song learning in zebra finches. *Nature Neuroscience* **3**: 482–8

Margoliash D 1997 Functional organization of forebrain pathways for song production and perception. *Journal of Neurobiology* **33**: 671–93

Marler P 1976 Sensory templates in species-specific behavior. In: Fentress J C (ed.) *Simpler Networks and Behavior*. Sinauer, Sunderland, MA, pp. 314–29

Marler P, Pickert R 1984 Species-universal microstructure in the learned song of the swamp sparrow (Melospiza georgiana). *Animal Behavior* **32**: 673–89

Marler P, Sherman V 1985 Innate differences in singing behaviour of sparrows reared in isolation from adult conspecific song. *Animal Behavior* **33**: 57–71

McCasland J S, Konishi M 1981 Interaction between auditory and motor activities in an avian song control nucleus. *Proceedings of the National Academy of Sciences of the United States of America* **78**: 7815–19

Morrison R G, Nottebohm F 1993 Role of a telencephalic nucleus in the delayed song learning of socially isolated zebra finches. *Journal of Neurobiology* **24**: 1045–64

Nixdorf-Bergweiler B E, Wallhausser-Franke E, DeVoogd T J 1995 Regressive development in neuronal structure during song learning in birds. *Journal of Neurobiology* **27**: 204–15

Nordeen K W, Nordeen E J 1992 Auditory feedback is necessary for the maintenance of stereotyped song in adult zebra finches. *Behavioral and Neural Biology* **57**: 58–66

Nordeen K W, Nordeen E J 1997 Anatomical and synaptic substrates for avian song learning. *Journal of Neurobiology* **33**: 532–48

Nottebohm F, Arnold A P 1976 Sexual dimorphism in vocal control areas of the songbird brain. *Science* **194**: 211–13

Nottebohm F, Kasparian S, Pandazis C 1981 Brain space for a learned task. *Brain Research* **213**: 99–109

Nottebohm F, Stokes T M, Leonard C M 1976 Central control of song in the canary, Serinus canarius. *Journal of Comparative Neurology* **165**: 457–86

Pohl-Apel G 1985 The correlation between the degree of brain masculinization and song quality in estradiol treated female zebra finches. *Brain Research* **336**: 381–3

Price P H 1979 Developmental determinants of structure in zebra finch song. *Journal of Comparative and Physiological Psychology* **93**: 260–77

Rasika S, Alvarez-Buylla A, Nottebohm F 1999 BDNF mediates the effects of testosterone on the survival of new neurons in an adult brain. *Neuron* **22**: 53–62

Rosen M J, Mooney R 2000 Intrinsic and extrinsic contributions to auditory selectivity in a song nucleus critical for vocal plasticity. *Journal of Neuroscience* **20**: 5437–48

Scharff C, Nottebohm F 1991 A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implications for vocal learning. *Journal of Neuroscience* **11**: 2896–913

Seidenberg M S 1997 Language acquisition and use: learning and applying probabilistic constraints. *Science* **275**: 1599–603

Simpson H B, Vicario D S 1990 Brain pathways for learned and unlearned vocalizations differ in zebra finches. *Journal of Neuroscience* **10**: 1541–56

Simpson H B, Vicario D S 1991 Early estrogen treatment alone causes female zebra finches to produce learned, male-like vocalizations. *Journal of Neurobiology* **22**: 755–76

Singh T D, Basham M E, Nordeen E J, Nordeen K W 2000 Early sensory and hormonal experience modulate age-related changes in NR2B mRNA within a forebrain region controlling avian vocal learning. *Journal of Neurobiology* **44**: 82–94

Sohrabji F, Nordeen E J, Nordeen K W 1990 Selective impairment of song learning following lesions of a forebrain nucleus in the juvenile zebra finch. *Behavioral and Neural Biology* **53**: 51–63

Solis M M, Doupe A J 1999 Contributions of tutor and bird's own song experience to neural selectivity in the songbird anterior forebrain. *Journal of Neuroscience* **19**: 4559–84

Solis M M, Doupe A J 2000 Compromised neural selectivity for song in birds with impaired sensorimotor learning. *Neuron* **25**: 109–21

Vates G E, Nottebohm F 1995 Feedback circuitry within a song-learning pathway. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 5139–43

Vu E T, Mazurek M E, Kuo Y C 1994 Identification of a forebrain motor programming network for the learned song of zebra finches. *Journal of Neuroscience* **14**: 6924–34

Wallhäusser-Franke E, Nixdorf-Bergweiler B E, DeVoogd T J

1995 Song isolation is associated with maintaining high spine frequencies on zebra finch 1MAN neurons. *Neurobiology of Learning and Memory* **64**: 25–35
White S A, Livingston F S, Mooney R 1999 Androgens modulate NMDA receptor-mediated EPSCs in the zebra finch song system. *Journal of Neurophysiology* **82**: 2221–34
Wild J M 1997 Neural pathways for the control of birdsong production. *Journal of Neurobiology* **33**: 653–70
Williams H, Mehta N 1999 Changes in adult zebra finch song require a forebrain nucleus that is not necessary for song production. *Journal of Neurobiology* **39**: 14–28
Yu A C, Margoliash D 1996 Temporal hierarchical control of singing in birds. *Science* **273**: 1871–5
Zevin J, Seidenberg M S, Bottjer S W 2000 Song plasticity in adult zebra finches exposed to white noise. *Society for Neuroscience Abstracts* **26**: 723

S. W. Bottjer

Birth, Anthropology of

An anthropology of birth is based on the common human experiences of insemination, conception, pregnancy, and birth of an infant. Characteristic of the human species and central to the origins of culture are the immaturity of the newborn human and its long-term dependence on caretakers for survival. These experiences exist in every society, but details vary in different cultural environments.

1. *Cultural Anthropological Perspectives*

Anthropological perspectives on birth are concerned with the context, variety, and complexity of reproductive life, modes of caring for a newborn infant, and cultural factors including rules for behavior that are learned, shared, and transmitted within families and cultural groups. There is cross-cultural variation among societies during these universal human processes. Because birth and recruitment to society are important to all human groups, customs and rules surround fertility behavior in all societies.

The influence of culture appears in definitions, for example, of which individuals may become parents (maternal age or married status), or the beginning of life (the point of conception, pregnancy to a viable birthweight, or delivery of a live infant). Throughout history and across cultures, certain variations have occurred in the definition of when life begins—whether it is at birth, or from motion or other indication of consciousness of an infant, or at a formal tenth day ceremony. Naming an infant provides a universal identification of the beginning of social life. Membership in society is defined and protected by both rules and traditions of social behavior. For example, international differences in mate selection include monogamy, or single partners, in most of the Christian Western world, and polygamy, or multiple wives in religious groups such as Mormons in the United States, or in some Islamic countries.

Differences occur also in sexual behavior, timing of pregnancy, and reproduction of society over time. It is the basis of identity, including how many people will be members of each society. The concept of 'legitimacy' refers to the acceptability of infants born to married mothers in most societies. 'Illegitimacy' refers to birth to an unmarried mother. Despite the language of legality, this is not generally an issue of law, but of social or religious acceptability.

Additional international and intercultural differences are discussed in the following sections, including fertility and fertility regulation, prenatal care, a comparative view of birth rates in different countries (Table 1), infant and under 5 child mortality (Table 2), and finally, the impact of the modern international infectious disease of HIV (human immunodeficiency virus, Table 3).

2. *Fertility*

Fertility refers to the capability of reproduction in the human species from sexual relations of male and female to conception and pregnancy in the female. Sex and social continuity are of concern in all societies. This concern is a part of the ancient and universal desire to bring order to membership in society as well as the behavior of individual members. 'Infertility' of either partner indicates an incapacity to reproduce.

Pregnancy occurs as a result of the connection of male sperm to female ovum initiating cell division leading to conception and beginning development of an embryo, until around the third month of gestation when it begins to develop body structures recognizable as human, and enters the fetal stage. Cells differentiate and the fetus then grows and develops the human form of an infant.

Gestation is the period of intrauterine development. Human variation occurs in the length of gestation for live birth. The biologically expectable period of gestation is around nine months of pregnancy. However, there can be differences in the timing, location, and length of the birth process. In some societies infants are delivered by a midwife or nurse, or a birth specialist either at home, or in the institutional setting of a hospital or nursing home. In modern urban societies, delivery of an infant is usually conducted in a hospital by a nurse or a physician specializing in obstetrics. Even here, however, there are significant cultural differences. For example, in urban areas, the surgical procedure of caesarean section is used for an impeded delivery, or for selective timing of birth.

In developed countries medical care may include prenatal care with a medical specialist in obstetrics or a nurse practitioner in a professional medical en-

vironment. In many areas of the world, reproductive health and training for motherhood are carried out by a nurse midwife or an experienced female family member, although there is increasing professionalization of birth attendants internationally. Attention to the process of gestation is perceived as important to the quality of life of the potential human being in all societies.

3. Fertility Regulation

Fertility regulation refers to intentional methods of avoidance of pregnancy used by either male or female. There have been methods of fertility regulation at least as far back as the beginnings of recorded history, as families have sought to control the number and timing of births. Male methods of family planning can range from a barrier method such as the condom, or withdrawal, or the surgical procedure of vasectomy. Female methods include the oral contraceptive pill or the barrier method of the female condom. Extended breastfeeding of the previous infant has long been known to suppress ovulation and enable spacing between births.

Social concerns regarding birth and reproduction vary internationally, with age, ethnic identity, or religion. Cultural factors, such as assuring virginity of marriageable females in some societies, religious opposition to fertility regulation in others, or in a few societies, female genital surgery for removal of the clitoris to reduce female enjoyment of sexual relations, have an impact on quality of life as well as reproductive capacity.

Ending unwanted pregnancy is one of the most contentious issues of reproduction. It may be terminated by induced clinical abortion in those countries where it is legal and medically provided. The psychosocial effects of unwantedness of a newborn infant are difficult to measure, but outcomes appear consistently negative in terms of later parental disinterest, child abuse, or abandonment. However, while clinical abortion is legal in such countries as Sweden, Japan, or the United States, it is illegal in most Muslim and Catholic areas. In these countries, family planning is more likely to include abstinence during the expectable time of ovulation (sometimes difficult to determine), and birth rates may be comparatively higher.

4. Prenatal Care

Most developed societies have some form of preparation for birth. The importance of prenatal care is in testing for health of the fetus, and in teaching the pregnant woman how to maintain the best possible health for herself and her expected infant. This includes protective health care, measuring regular weight gain, providing training for the birth process, and social support.

An important component is also prevention of behaviors detrimental to infant development. For example, prenatal smoking, alcohol and drug use, or starvation to limit weight gain, can all have damaging effects on the developing fetus. These can range from decreased blood flow to the fetus from smoking to fetal alcohol syndrome (FAS), which can result in abnormal physical or mental development (Espy et al. 2000). As risks from these behaviors have become better known over the last several decades, preventing risks of learning impairment in the resulting child is an important component of effective prenatal care before birth (Newman 1996).

5. Premature Birth

Not all pregnancies continue to a viable birth. Loss can occur at any time. Abortion refers to expulsion of a malformed or incompletely developed embryo within the first three months of gestation. Later loss can result from fetal death. For example, if loss occurs in the first weeks of pregnancy, it is considered a 'miscarriage' of a non-viable fetus. Early delivery of a viable infant at a birth weight of 2500 g (about 5.5 16) can happen as early as seven months gestation, and is termed 'low birth weight.' 'Very low birth weight' refers to the infant born around 1500 g.

In developed and industrialized countries, extensive efforts are undertaken to ensure survival of the lowest birth weight infants. A viable low birth weight infant in developed countries is kept at a neonatal intensive care nursery in an incubator for clear observation, to maintain body heat, and to provide oxygen, if necessary to breathing, as well as to enable constant surveillance by special care staff. The purpose of enclosure in all countries is to keep the premature infant safe from infection. In developing countries without intensive care technology, such babies are seen as likely to suffer mental and physical limitations, and may not be given special care so that family and community resources can be focused on those infants perceived as more viable.

6. Birth Rate

Birth rates differ from country to country based on religion, cultural expectation, issues such as female participation in the workforce, or religious and political views of fertility limitation through birth control. It can also be influenced by political programs such as China's 'one child family' policy previously under-

Table 1
Total population and birth rate

Nation	Total population (millions)	Birth rate (per woman)
Australia	18.6	1.8
Brazil	165.9	2.3
China	1238.6	1.9
France	58.8	1.8
India	979.7	3.2
Indonesia	203.7	2.7
Mexico	95.8	2.8
Nigeria	120.8	5.3
Russian Federation	146.9	1.2
South Africa	41.4	2.8
United Kingdom	59.1	1.7
United States	270.3	2.0

Source: The World Bank, *1998 World Development Indicators*, Washington DC, 2000.

taken by the nation with the highest population in the world.

Table 1 indicates total population for 12 nations and the birth rate, or births per woman, for each country.

7. *Infant Mortality*

Infant mortality refers to death in the first year of life. While it has lessened worldwide in the twentieth century, the rates vary in different countries, ranging from 5 per 1000 live births in developed countries such as Australia and France, to as many as 76 per 1000 live births in Nigeria. Rates of Under 5 Child Mortality, while higher, are comparable nationally. Table 2 compares Infant and Under 5 Child Mortality in 12 countries.

Table 2
Infant and under 5 child mortality

Nation	Infant mortality (per 1000 live births)	Under 5 child mortality (per 1000)
Australia	5	6
Brazil	33	40
China	31	36
France	5	5
India	70	83
Indonesia	43	52
Mexico	30	35
Nigeria	76	119
Russian Federation	17	20
South Africa	51	83
United Kingdom	6	7
United States	7	7

Source: The World Bank, *1998 World Development Indicators*, Washington DC, 2000.

Table 3
International prevalence of HIV infection

Nation	Percent of adults	Estimated number infected (in 000's)
Australia	0.14	11
Brazil	0.63	580
China	0.06	400
France	0.37	110
India	0.82	4100
Indonesia	0.85	52
Mexico	0.35	180
Nigeria	4.12	2300
Russian Federation	0.05	40
South Africa	12.91	2900
United Kingdom	0.09	25
United States	0.76	820

While infant mortality can occur from prenatal causes, or in the process of delivery, it can also occur through accident or by endemic infectious disease later within the first year of life.

In many areas of the world, fertility regulation and birth control methods have been used to increase the time between births and enable optimum developmental opportunity for each infant born. Exceptions occur in those countries where fertility regulation is forbidden by religious law. Birth spacing is accomplished in some countries by later marriage and in some by abstinence during the fertile period of ovulation.

8. *Sexually Transmitted Disease*

The most pervasive sexually transmitted diseases in the world have been syphilis, gonorrhea, and chlamydia. However, a powerful new influence on the culture of pregnancy and birth is the spread of the new infectious disease of the modern world, human immunodeficiency virus, or HIV infection, identified in 1980. While HIV can be transmitted by sharing needles among drug users, it can be transmitted sexually to both males and females. It is also transmitted from mother to infant prenatally, during delivery, or postnatally through breast feeding (WHO and UNAIDS 1998). This has had a devastating effect internationally on the number of infants born with HIV infection or being infected within the first year of life. While the prevalence of HIV differs geographically, and in many countries is less than 1 percent, the rate in some African countries is higher, ranging from 4 percent to 12 percent, as indicated in Table 3.

The fear of becoming HIV positive is beginning in some countries to change behavior toward careful

choice of sexual partners, and using barrier methods of protection. The failure to change in other areas has led to devastating epidemics of HIV.

9. *Conclusion*

Anthropological perspectives on birth enable cross-cultural comparison of healthy birth practices and effective family planning methods based on cultural acceptability. Effective fertility regulation and prenatal care enable continuing reproductive health. Cultural variation in reproductive rights, values, and education demonstrate differences in fertility, contraceptive use, and birth practices, as well as expected family size internationally.

See also: Fertility Control: Overview; Fertility Control: Prevalence and Consequences of Breastfeeding; Fertility of Single and Cohabiting Women; Fertility: Proximate Determinants; HIV and Fertility; Infant and Child Mortality: Central and Eastern Europe; Infant and Child Mortality in Industrialized Countries; Infant and Child Mortality in the Less Developed World

Bibliography

Andrews Espy K, Francis D J, Riese M L 2000 Prenatal cocaine exposure and prematurity: Neurodevelopmental growth *Journal of Developmental and Behavioral Pediatrics* **21**(4)

Newman L F 1996 *Preventing Risks of Learning Impairment*. Education Commission of the United States, Denver, CO

Rosen J E, Shanti R C 1998 *Africa's Population Challenge: Accelerating Progress in Reproductive Health, Country Study Series No. 44*. Population Action International, Washington, DC

UNICEF 2000 *The Progress of Nations: The Nations of the World Ranked According to their Achievements in Fulfilment of Child Rights and Progress for Women*. UNICEF, New York

World Bank 2000 *World Development Indicators, 2000*. The World Bank, Washington, DC

World Health Organization 2000 Reproductive health. *Bulletin of the World Health Organization* **78**(5): 563–714

WHO and UNAIDS 1998 *HIV and Infant Feeding*. 1. Guidelines for Decision-Makers. 2. A Guide for Health Care Managers and Supervisors. 3. A Review of HIV Transmission through Breastfeeding

L. F. Newman

Bleuler, Eugen (1857–1939)

Eugen Bleuler is a central figure in the history of European psychiatry. His importance is founded on the fact that he proposed *and* practiced a critically mediating link between two basic psychiatric positions which stood at two controversially opposite poles at the beginning of the twentieth century (and still do in part up to the present day); that is, the psychoanalytical depth psychological school on the one hand, which was situated predominantly outside the universities, and academic psychiatry on the other, which was principally oriented towards the descriptive somatic approach. In this respect Eugen Bleuler—to put it in layman's terms—stood, as far as his basic psychiatric conviction was concerned, as it were 'between' the two similarly aged exponents of the aforementioned competing approaches, namely Sigmund Freud (see *Freud, Sigmund (1856–1939)*) (1856–1939) and Emil Kraepelin (see *Kraepelin, Emil (1856–1926)*) (1856–1926).

Born in Zollikon, near Zurich, on April 30, 1857, Bleuler came from an old-established Swiss farming family on both his father's and his mother's side. After completing his medical studies between 1875 and 1881, he worked first as an intern at the psychiatric clinic Waldau, near Bern. Study visits took him to Paris (with Charcot and Magnan) and to Munich (with Von Gudden). In 1885 he went to work with Forel at the Zurich clinic named the 'Burghölzli' which had been founded in 1870, only to be appointed director of the Psychiatric Hospital Rheinau in the canton of Zurich the following year, a position which he held for 12 years, that is, until 1898. The professional position which followed also covered an unusually long period of time, and brought with it more far-reaching repercussions and to a large extent had a formative influence on the way Bleuler was to conceive the whole field of psychiatry. He took over the chair in psychiatry at the University of Zurich in 1898, succeeding August Forel (1848–1931), a position which was attached to the directorship of the 'Burghölzli.' Bleuler held this position for almost three decades until 1927.

At the end of his academic life—he was at that time also the Rector of the University of Zurich—he was said, in an account cited by the Swiss psychiatrist J. Klaesi, to have remarked, not without a hint of maliciousness, 'that the state forces one to be an old man at 72' (Klaesi 1970). Eugen Bleuler died in Zollikon in his 83rd year on July 15, 1939.

As far as content is concerned, Bleuler's work is characterized by its tireless attempt, influenced by a strong sense of tolerance, to approach the enigma of psychotic illnesses on as many levels as possible, while at the same time taking into consideration general or 'normal' psychological findings. On the one hand he wanted to avoid making any absolute one-sided dogmatic assertions, such as in the sense of Karl Jaspers' still derided, though at that time (and not only then!) highly popular 'brain mythologies' which, with no further reflection, identified the psychic with material mental processes along ingenuously materialistic lines. On the other hand Bleuler was concerned with regarding the psychotic person *not only* as ill, as

deviant, or as disturbed, but rather directed his attention to the fundamental comparability to healthy mental life, a standpoint which interested him above all because of the possibility for its therapeutic exploitation. This did not, however, mean that Bleuler in any way placed little value on the then-prevailing 'brain pathological'—in today's terminology, neurobiological—school of thought and the research approaches arising from it. On the contrary; like the vast majority of contemporary psychiatrists, he too was convinced that the brain function in the aetiology and pathogenesis of endogenously psychotic illnesses—predominantly, therefore, schizophrenic and (bipolar) affective disorders—played a decisive role. Unlike the many 'simplifiers,' however, Bleuler always insisted that it was of great importance for the handling of patients, for the diagnosis, and above all for the therapy to *understand* the clinical picture and each individual, concrete symptom, independent of the suspected aetiology.

Bleuler's handling of the basic concepts of association theory was not unified, and was hence liable to give rise to misunderstandings and controversial statements. On the one hand he repeatedly took up a clear position against the 'association psychology' in the narrow sense advocated earlier in the nineteenth century, in which (healthy as well as ill) human mental life was merely the sum total of individual intellectual associations prompted by sensory perception. Against this he set an affective psychology which aimed at comprehending the personality in its individual totality, but which was nonetheless coupled substantially with elements from association psychology in the field of the schizophrenias. Indeed, not only did Bleuler see no contradiction in this, he even saw it as objectively necessary. It was here that the (then as now) topical controversy surrounding the quantifying scientific *explanation* and the interpretative humanistic *understanding* of mental and, here in particular, psychotic manifestations arose. Bleuler also attempted in this context to overcome a rigid distinction between the two approaches, quite unlike Karl Jaspers, who strongly criticized him over this on several occasions (for instance in Jaspers 1923).

Unlike most of his university colleagues in his field, Bleuler saw in very general terms *no* fundamental contradiction between (explanatory) brain research in the broadest sense and the interpretative access to the psychotic of depth psychology which was at that time almost exclusively advocated by the still young Sigmund Freud. On the contrary, Freud's theory seemed to Bleuler to be virtually predestined for critical implementation into clinical psychiatry. For Freud himself, who fought a lifelong struggle to achieve 'academic' recognition for his theory, such a competent and influential researcher like Bleuler naturally took on a particularly great importance. Bleuler, however, emphasized the aspect of the *critical* adoption of psychoanalysis already mentioned so much over the years that it led to a rift. Much has been written about the escalating debate between Freud and his particularly creative and original disciples such as Carl Gustav Jung (see *Jung, Carl Gustav (1875–1961)*) or Alfred Adler. It is less well-known that Eugen Bleuler must also be mentioned in this connection. The psychiatry historians Alexander and Selesnick (1969) even went so far as to call him the 'first victim' of this controversy, as Bleuler resigned from the International Psychoanalytic Society in 1910 because of his discontent at its, in his view, authoritarian style of leadership. The fundamental importance of the depth psychology approach for psychiatry was for him in no way diminished because of this. Scherbaum (1992) points out the paradoxical situation that Bleuler was 'too psychoanalytical' for academic colleagues in his field, whereas the psychoanalysts for their part viewed him as being too critical, even 'disloyal,' and he poses the question of whether Bleuler's understanding of psychoanalysis might not also have been shaped by the same 'scientistic self-misconception,' which Habermas (1979) had observed in Freud himself.

One article was of decisive importance for the development of *clinical psychiatry*, an article which Bleuler had written for the *Handbuch der Psychiatrie* (*Handbook of Psychiatry*), edited by Gustav Aschaffenburg, entitled *Dementia praecox oder Gruppe der Schizophrenien* (1911; *Dementia praecox, or the Group of Schizophrenias*). The agenda was already set out in its title. Unlike Kraepelin, who was searching for 'natural illness units' and through whose influential textbook the concept of 'dementia praecox' had become both popular and influential in the German-speaking world (Hoff 1994), Bleuler did not speak of schizophrenia as a *single* unit, but of the *group* of schizophrenias. However, he was concerned with the question of which characteristics the otherwise very heterogeneous schizophrenic patients had in common. Bleuler was here describing first and foremost *psychopathological* phenomena, and was avoiding any all too narrow (and speculative) analogies to the brain function.

Two psychopathological distinctions which Bleuler introduced are important with regard to schizophrenic syndromes and are frequently confused: the basic and accessory symptoms on the one hand, and the primary and secondary symptoms on the other. *Basic symptoms* were, according to Bleuler, those things which were peculiar to every schizophrenic illness process, independent of its other clinical manifestation. Among these he counted schizophrenic autism (not to be confused with the syndrome of the same name in child and adolescent psychiatry), ambivalence (a volitional disorder), the disorder of formal thought (in particular the aforementioned disorder of associative thought processes), and affectivity. *Accessory*, that is, not necessarily existing and also frequently observed in non-schizophrenic and above all exogenous psychoses,

meant symptoms such as hallucinations, delusion, or catatonic states. While this differentiation is still of great heuristic value for psychopathological research today, the differentiation between the *primary symptoms*—which result directly from the organic illness process, as Bleuler also suspected—and the *secondary symptoms*—which constitute a subjective psychological reaction to the illness by the person affected—has lost considerable importance.

In contrast to the (as the term implies) prognostically very pessimistic concept of 'dementia praecox' such as in Kraepelin's work, Bleuler constantly stressed that there were also benign schizophrenic courses—and by no means only as an exception—which after one or more episodes turned into an extensive or complete recovery.

The *Lehrbuch der Psychiatrie* (*Textbook of Psychiatry*), which first appeared in 1916, achieved a virtually legendary reputation. In this work, Bleuler (1916) put forward his deeply personal view of this field in a systematically, vividly, and clinically oriented manner. He had worked five years on this book. It went through many editions, for which he himself took on the task of updating as necessary up to and including the sixth edition. After his death (1939) his son Manfred Bleuler took over the task.

Eugen Bleuler's writings on general psychology and philosophy have, on the other hand, all but been forgotten today. In his later work—such as in the work *Die Psychoide als Prinzip der organischen Entwicklung* (1925; *Psychoids as a Principle of Organic Development*)—he attempted to develop a comprehensive natural philosophical concept which was, however, not adopted on any great scale at the time. His basic idea of a memory which was independent of conscious processes and did not only refer to the individual person but—an evolutionary biological perspective—to the whole species ('Mneme') was, in its idiosyncratic embracing of biogenetic vitalistic ideas of the late nineteenth century, of virtually central importance for Bleuler's whole scientific understanding. Although Bleuler considered inappropriate the classification of his (later) work as 'vitalistic,' from the point of view of the history of psychiatry today the parallels are unmistakable. This is true in particular if one looks at Bleuler's objective of a comprehensive 'science of life' which did not separate the categories of physical, biological, mental, and even social phenomena from each other, but rather sought to understand them as scientifically equal forms to express a single integrative (life) principle. A weakening or even lifting of the aforementioned categorical dividing lines (though not *this* particular way) is, what is more, a highly topical subject in psychiatric research nowadays which, indeed, appears less and less able to maintain the classical differentiation of 'organic' and 'endogenous' as well as 'psychogenous' illnesses.

Despite the approaches in his later work which are irritating and speculative to today's reader, and the very clear fixing, especially in the context of questions of forensic psychiatry which are also in his work, in the dubious modes of discussion of his time such as in the 'degeneration theory,' Eugen Bleuler was without question one of the most original and consistent psychiatric clinicians and researchers of his time. His thought is also of interest to today's research into psychosis, and has been receiving increased attention in recent literature (Möller and Hell 1999, 2000, Stotz-Ingenlath 2000). What is particularly gratifying about this is the fact that research into the history of psychiatry has turned away extensively from the bad habit, which one often came across in the early years, of uncritically hagiographic and therefore unscientific excessiveness from eminent advocates in the field, and is now concerning itself discernibly with a comprehensive understanding of the history of our field supported by the original literature and other sources, and embedded in contexts from social history and the history of ideas.

See also: Dementia: Overview; Dementia: Psychiatric Aspects; Freud, Sigmund (1856–1939); Kraepelin, Emil (1856–1926); Mental and Behavioral Disorders, Diagnosis and Classification of; Mental Health and Normality; Nosology in Psychiatry; Philosophy, Psychiatry, and Psychology; Psychiatry, History of; Psychoanalysis, History of; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Schizophrenia; Schizophrenia and Bipolar Disorder: Genetic Aspects

Bibliography

Alexander F G, Selesnick S T 1969 *Geschichte der Psychiatrie*. Diana, Konstanz, Germany

Bleuler E 1911 Dementia praecox oder die Gruppe der Schizophrenien. In: Aschaffenburg G (ed.) *Handbuch der Psychiatrie*. Deuticke, Leipzig Spezieller Teil, Sect. 4, Germany

Bleuler E 1916 *Lehrbuch der Psychiatrie*. Springer, Berlin

Bleuler E 1925 *Die Psychoide als Prinzip der organischen Entwicklung*. Springer, Berlin

Habermas J 1979 *Erkenntnis und Interesse*, 5th edn. Suhrkamp, Frankfurt/Main, Germany

Hoff P 1994 Emil Kraepelin und die Psychiatrie als klinische Wissenschaft. Ein Beitrag zum Selbstverständnis psychiatrischer Forschung. In: Hippins H, Janzanik W, Müller C (eds.) *Monographien aus dem Gesamtgebiete der Psychiatrie*. Springer, Berlin, Vol. 73

Jaspers K 1923 *Allgemeine Psychopathologie*, 3rd edn. Springer, Berlin

Klaesi J 1970 [1956] Eugen Bleuler (1857–1939). In: Kolle K (ed.) *Große Nervenärzte*, 2nd edn. Thieme, Stuttgart, Germany, Vol. 1, pp. 7–16

Möller A, Hell D 1999 Das allgemeinpsychologische Konzept im Spätwerk Eugen Bleulers. *Fortschritte der Neurologie und Psychiatrie* **67**: 147–54

Möller A, Hell D 2000 Prinzipien einer naturwissenschaftlich begründeten Ethik im Werk Eugen Bleulers. *Nervenarzt* **71**: 751–7

Scherbaum N 1992 Psychiatrie und Psychoanalyse—Eugen Bleulers Dementia praecox oder Gruppe der Schizophrenien (1911). *Fortschritte der Neurologie und Psychiatrie* **60**: 289–95
Stotz-Ingenlath G 2000 Epistemological aspects of Eugen Bleuler's conception of schizophrenia in 1911. *Medicine, Health Care and Philosophy* **3**: 153–9

P. Hoff

Bloch, Marc Léopold Benjamin (1886–1944)

The French historian Marc Bloch is one of the authors most frequently cited by those working in contemporary historical disciplines around the world. This immense influence is founded on both scholarly and biographical grounds. As one of the founders of the field of the history of society and mentality, as a historical theoretician, and a critical intellectual who was executed by the Gestapo in 1944, Bloch has become an inspiration and model for generations of researchers working in a diverse range of scientific fields and from a variety of different political perspectives.

1. *Biographical Outline*

Marc Bloch was born in Lyon on 6 July, 1886 into a family of practicing Jews of Alsatian heritage. Bloch himself was an agnostic, and in his final will and testament he explicitly requested a secular funeral. Because his father was professor for Roman history at the Sorbonne, Bloch grew up in Paris, where he went to the elite secondary school Louis-le-Grand. From 1904 to 1908 he studied history and geography at the Ecole Normale Supérieure in Paris; following this period he spent two semesters in Berlin and Leipzig, where he attended lectures by scholars such as Gustav Schmoller, Karl Bücher, and Karl Lamprecht. From 1909 to 1912 he was the recipient of a Fondation Thiers scholarship, before becoming a teacher at secondary schools in Montpellier and Amiens.

Bloch fought on the front as an enlisted soldier in World War I, and in the course of being repeatedly wounded in action and decorated for his service he was promoted to the rank of captain. In 1919 he married and was appointed lecturer for medieval history at the newly founded French University of Strasbourg. In 1921 he was appointed associate professor, and in 1927 he received his full professorship. In 1929, together with his Strasbourg colleague Lucien Febvre (1878–1956), Bloch founded the journal *Annales d'histoire économique et sociale*, which rapidly grew to become a widely influential forum for innovative historiography. In 1936, after several unsuccessful attempts to apply for a position in Paris, Bloch was granted a Sorbonne professorship in economic history.

Immediately following the outbreak of World War II, he volunteered for military service to defend his homeland against Nazi Germany. As an officer sent to reinforce the 1st Army, he experienced both the anticipation of the *Sitzkrieg* and then the unexpectedness of the *Blitzkrieg* in the French coastal canal region. Like thousands of other French soldiers, Bloch was rescued from Dunkirk in June of 1940 in the emergency evacuation across the channel. From England, he returned immediately to France, where he was able to avoid imprisonment. Because Nazi law prohibited him from reassuming his teaching position at the Sorbonne, the Vichy government sent him to rejoin the faculty of the University of Strasbourg, which had been relocated to Clermont-Ferrand. At the same time, Bloch sought, and eventually attained, an appointment at the New School of Social Research in New York. Because of the complicated provisions for obtaining a visa, however, he would have been compelled to leave behind part of his family in France—a condition Bloch, the father of six children, was entirely unwilling to accept.

In December of 1940 Bloch was barred from government service by the 'Jewish Statute,' only to be exempted several weeks later along with a number of other prominent Jewish scholars, and allowed to resume his teaching duties. For the sake of his wife's health, Bloch applied a year later for a transfer to Montpellier, even though this meant that he would no longer be permitted to hold public lectures. It was at this point in time that Bloch's active involvement in the Resistance movement began. Following the occupation of the Vichy zone by the German army in November 1942, Bloch fled Montpellier with his family for his countryside home near Guéret in the Creuse district of France. From this time on he worked and lived for the resistance, operating under pseudonyms like 'Chevreuse,' Arpajon,' and 'Narbonne' he belonged to the masterminds of the *Franc-Tireur* movement. In 1943 he moved to Lyon, where he belonged to the regional leadership of the *Résistance*. On 8 March, 1944, Bloch was arrested by the Gestapo and brutally tortured before being thrown in prison. On 16 June, several days after the Allied troops landed at Normandy, Bloch and 29 other prisoners were taken to a field on the outskirts of Lyon and shot.

2. *Bloch's Work*

The thematic emphasis of Bloch's scholarly production lies in the social and economic history of the medieval period. Although Bloch was initially concerned primarily with the topic of serfdom, in the early 1920s he began to devote his attention to what

appeared to be a highly specific question in the field of political theology: namely, the issue of the widespread belief in feudal France and England in the healing power of the king; in particular, when he laid his 'holy hands' upon the sick who had gathered before the cathedral immediately following the coronation ceremony. This anthropological approach, however, so clearly inspired by Durkheim's sociology of religion, was one Bloch would only use intermittently in later years. Instead, he turned his focus to a comparative historiography of European agricultural structures and societal structures that spanned the period from late antiquity to early modern times.

As he wrote in his 1934 application for a teaching position at the Collége de France, Bloch no longer considered himself to be first and foremost a 'medievalist,' but rather a general historian of 'European societies.' His two-volume work *Feudal Society* (Bloch [1939–40] 1994) can be understood as an exemplary attempt at putting this perspective of an *histoire total* into practice. In what amounted to an intellectual testament, Bloch himself summed up his book as follows:

> 'I have given an example of something … that I chose to call "the dissection of a social structure." In the course of the development of Western society there was a phase, which we call feudal society, which had a certain social tone. It was this tone that I wanted to investigate. … Furthermore, I have attempted to bring into play in a European context the multifarious experiences which the comparative method allows us to grasp. If my work is truly original in one regard, then in my opinion its originality lies in both of these efforts: structural analysis and the incorporation of comparative experiences' (*Cahiers Marc Bloch*, No. 2, 1995, p. 16).

Bloch's originality lies indeed in this 'structural' and comparative approach, but also in the manner in which he conducts his scholarly interventions. Precisely because he took seriously the approach of Durkheim's disciples such as François Simiand, who in their work had critically engaged the traditional praxis of political history, Bloch had been confronted with the necessity of making a fundamental change in the theory and practice of historiography. In countless essays and reviews he called not only for a new, interdisciplinary and internationally oriented historiography that would address itself to new themes by means of new methods, but also new ways of educating historians and new forms of communication within the historian's guild. An important instrument, a sort of editorial 'lever,' in his efforts to this end was the journal *Annales*, which he published together with Lucien Febvre. It was in the course of this cooperative venture, and also as a result of the ensuing friendly competition between him and Febvre that Bloch developed his remarkable ability to quickly and succinctly formulate new scholarly perspectives. The fight between the two coeditors was thus all the more painful for Bloch, when, in the aftermath of the French defeat in 1940, they were forced to decide if and how the journal should continue to be published. Whereas Bloch argued in favor of discontinuing the publication of the *Annales* until the end of the war, Febvre was prepared to adapt—at least on the surface level—to the new conditions under occupation in order to continue to have a public voice. In the end, Bloch accepted this strategy and continued to publish in the *Annales* under the pseudonym 'Fougères' until the very end. At the same time, he continued to work on two booklength manuscripts, which were edited by Febvre after the war: an analysis of the French defeat (*Strange Defeat*) and a reflective work on the intellectual preconditions of historiography as science, craft, and art (*The Historian's Craft*, Bloch 1953).

3. *Reputation and Influence*

The beginning of Bloch's considerable reputation in French as well as international scholarly circles can be traced back to the 1930s. It was really after World War II, however, particularly in the closing decades of the twentieth century, that Bloch's ascendancy to the ranks of one of the world's most quoted historians took place. His books have been translated into all of the world's major languages: the number of printed copies of *The Historian's Craft* extends well into the six-figure range, one-third of which have appeared in Spanish translation. Bloch's heroic death has elevated him to the status of an icon, adored by intellectuals and academics representing a broad spectrum of disciplines and political convictions alike. Accordingly, one encounters a range of conflicting images of 'Marc Bloch': the Marxist or the conservative, the Jew or the nationalist, and recently even the 'postmodernist' Bloch, who supposedly drew on Nietzsche as a source of inspiration. Whereas such portrayals threaten to bury the author's historical work under ever-increasing layers of reception, recent years have seen an increased effort towards precise scholarly contextualization of this work in order to better reconstruct Bloch's contribution to the historical discipline of the twentieth century.

See also: Economic History; Historiography and Historical Thought: Current Trends; Middle Ages, The; Social History

Bibliography

Bloch M [1924] 1983 *Les rois thaumaturges. Étude sur le caractère -An Impossible Biography*. Culture & Patrimoine en Limousin, Limoges, France

Bloch M [1924] 1983 *Les rois thaumaturges. Étude sur le caractère surnaturel attribué à la puissance royale, particulièrement en France et en Angleterre*. Gallimard, Paris

Bloch M [1939–40] 1994 *La société féodale*. Albin Michel, Paris
Bloch M [1949] 1953 *The Historian's Craft*. Knopf, New York
Bloch M 1963 *Mélanges historiques*. Sevpen, Paris, 2 Vols.
Cahiers Marc Bloch 1994–(annual)
Fink C 1989 *Marc Bloch. A Life in History*. Cambridge University Press, Cambridge, UK
Friedman S W 1996 *Marc Bloch, Sociology and Geography. Encountering Changing Disciplines*. Cambridge University Press, Cambridge, UK

P. Schöttler

Blocking, Neural Basis of

Blocking is a conditioning paradigm, first described by Leon Kamin (1969), in which previous conditioning to a stimulus reduces the degree to which a second stimulus can be conditioned during compound conditioning.

A typical blocking experiment consists of two training phases and a testing phase. During the first training phase, animals in the experimental (blocking) group learn that a conditioned stimulus (CS; e.g., a tone) predicts the occurrence of the unconditioned stimulus (e.g., an electric shock), and as a result develop a conditioned response (e.g., blinking) to the tone. A different group of animals (the control group) experience no training during this phase. During the second phase, both groups are trained with a compound stimulus, which consists of the previously trained stimulus (e.g., the tone) and a new stimulus (e.g., a light) presented simultaneously (tone + light). Later, during the testing phase, both light and tone are individually tested in their capacity to elicit the conditioned response (CR). The typical results are that, whereas the animals in the control group learned about both stimuli (i.e., both the light and the tone elicit a conditioned response), those animals in the experimental group did not learn about the light. It appears that previous experience with the tone 'blocked' or attenuated subsequent learning about the light.

1. *Theoretical Interpretations*

Before blocking was described, a common assumption in learning theory was that temporal contiguity was sufficient for an association to be formed between stimuli. This assumption was challenged by blocking, since it implied that other factors, such as previous training experience with stimuli, influence the way an animal learns about those stimuli.

Kamin (1969) proposed a cognitive explanation for how such previous experience influences learning. According to him, animals form expectations about the world, compare current input to those expectations, and learn only when something previously unpredicted happens. Thus, he claimed that stimuli support learning only to the extent that the outcomes they signal are surprising.

In the case of the previously mentioned example of blocking, by the end of the first training phase, the unconditioned stimulus (US) is already fully predicted by the tone (CS1). Consequently, when the animal is presented with the light (CS2) during the compound phase (tone + light) the shock is not surprising anymore (the animal already knows, just by hearing the tone, that the shock is coming next) and therefore an association between the light and the tone does not occur.

This idea was formalized in a mathematical model proposed by Rescorla and Wagner (1972). Although it is beyond the scope of this article to present details of this model, the key idea is that the amount of learning that occurs on each trial is proportional to the difference between the animal's expectation of the US (based on all the CSs present on that trial), and the presence (or absence) of the US. The more surprising the US, the bigger this difference, and therefore the larger the amount of learning (or associative strength) that a stimulus can accrue. As training progresses, the CS becomes increasingly more predictive of the US. This makes the US less surprising at each trial, resulting in progressively smaller gains in associative strength to the CS.

2. *Neural Substrates*

Although blocking is observed in virtually all classical conditioning paradigms, it would be simplistic to assume that there is only one set of neural structures underlying blocking in all such paradigms. Rather, it is quite likely that different paradigms require the involvement of different neural systems mediating blocking.

One approach to studying the neural basis of blocking consists of identifying a biologically plausible mechanism that performs the kind of computations called for by a learning model such as Rescorla and Wagner's.

The formulation made by the Rescorla and Wagner model—that the associative strength acquired by a CS is progressively smaller as training progresses—implies the existence of a negative feedback mechanism regulating the amount of US information provided to the brain throughout training. Neural mechanisms that resemble such computations have been observed during some classical conditioning paradigms, notably fear conditioning and eye-blink conditioning.

During fear conditioning, a CS is paired with a painful electric shock. Once associated, the CS elicits

fear-related behaviors, one of which is a state of analgesia mediated by opioids (Fanselow 1984). Given that fear conditioning depends on the painfulness of the US, the analgesia mediated by the CS could progressively diminish the reinforcing efficacy of the US. During a blocking task, this would provide negative feedback on the acquisition of fear conditioning, automatically performing calculations that resemble those in the Rescorla and Wagner model. Initial evidence for this negative feedback mechanism was provided by the finding that administration of the opioid antagonist naloxone during phase two of blocking attenuates blocking (Fanselow and Bolles 1979).

Further exploration of a negative feedback mechanism for blocking occurred in the context of eyeblinking conditioning. During this task, a CS is paired with an airpuff US, which elicits an unconditioned response (UR) consisting of reflexive closure of the eyelid. Through conditioning, the CS comes to elicit a conditioned response (CR), which mimics the US and precedes it in onset time (for review see Gormezano et al. 1983). The basic circuitry for the formation of CS–US association is well characterized in this paradigm (Thompson 1986). The cerebellum receives information about the CS from the pontine nuclei, whereas the inferior olive (IO) provides it with US information (Mauk et al. 1986). Interestingly, neurons in the IO show evoked activity to the US early during eye blinking training but not later, once the animal starts performing CRs (e.g., Sears and Steinmetz 1991).

This reduction in US input as conditioning progresses resembles a negative feedback mechanism involved in the regulation of the US information from the inferior olive to the cerebellum. Such a mechanism is likely to be provided by the inhibitory projection from the interpositus nucleus in the cerebellum back to the inferior olive. The activation of this negative feedback mechanism seems to be regulated by the establishment of the conditioned response, so that once the animal has learned about the CS–US relationship, the information about the US starts to be suppressed. Kim et al. (1998) demonstrated that the CR-induced inhibition of the inferior olive activity is mediated by the inhibitory neurotransmitter Gamma-amino butyric acid (GABA), because when picrotoxin, a GABA antagonist, was directly infused in the inferior olive in well-trained rabbits, the Purkinje cells in the cerebellum responded to the unconditioned stimulus, even though the animals continued to perform CRs.

Such a negative feedback mechanism could account for blocking during eye-blinking conditioning. In order to test such a possibility, Kim et al. (1998) implanted rabbits with unilateral guide cannulae directed to the IO and trained the animals in a blocking procedure using a tone as CS1 and a light as CS2. During the compound phase (tone + light), some animals underwent infusions of picrotoxin into the IO, while the control animals received infusions of artificial cerebrospinal fluid. The results showed that whereas the control animals exhibited normal blocking, those animals receiving picrotoxin learned to respond to both the tone and the light, thus showing no evidence of blocking.

According to Kim et al. (1998), the infusion of picrotoxin in the IO during compound conditioning prevented the tone-induced neural activity in the Purkinje cells of the cerebellum from inhibiting the IO activity, allowing it to keep providing US information during the presentations of the compound, thus making it possible for an association to occur between the light and the US.

3. *Hippocampus and Blocking*

Since blocking has been considered an indication that animals actively select and process information from their environment according to its behavioral significance (Rickert et al. 1981), it is possible to infer that those brain regions that play a relevant role in the processing and selection of stimuli might also be involved in blocking.

One structure that has been proposed to play a critical role in stimulus processing is the hippocampus. Pyramidal cells in this brain region increase their firing rate in response to CS–US pairings in anticipation of the CR (Hoehler and Thompson 1980). This has been interpreted as evidence that the hippocampus might function to mark behaviorally significant stimuli. It has also been suggested that one function of the hippocampus in associative learning is the reduction of attention to stimuli that are irrelevant or for which the behavioral consequences are already known (Han et al. 1995).

There is indeed evidence for a hippocampal involvement in blocking. For example, blocking has been shown to be disrupted in eye-blink conditioning in rabbits when the dorsal region of the hippocampus was lesioned (Solomon 1977); in a conditioned suppression paradigm in rats with a lesion in the ventral hippocampus (Rickert et al. 1981), as well as in taste aversion learning in rats with dorsal hippocampal lesions (Gallo and Candido 1995).

4. *Summary and Conclusion*

Distinct neural regions might participate in blocking according to the particular paradigm in which it is being evaluated. However, the finding of a common mechanism that can support blocking would greatly improve our understanding of the kind of computations the brain performs during learning, which in turn give rise to phenomena such as blocking. Evidence for such a mechanism could also provide support for models of learning, if the neural mechanism resembles

the computations claimed by the model to occur during learning.

As of 2000, there is strong evidence for the existence of negative feedback mechanisms underlying blocking in two paradigms: fear conditioning and eye-blink conditioning. The biological plausibility of such mechanisms should motivate further research in this direction. Additionally, more efforts should be directed to elucidate the involvement of other brain regions known to be important for stimulus selection, such as the hippocampus. The experiments to date strongly suggest an important role for this structure in blocking, but the variability in the exact location and extension of the lesions across studies imposes some difficulty in the interpretation of the results. Future studies with more restricted lesions will be required in order to clarify the hippocampal involvement in blocking.

See also: Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Eyelid Classical Conditioning; Fear Conditioning; Hippocampus and Related Structures; Mathematical Learning Theory

Bibliography

Fanselow M S 1984 What is conditioned fear? *Trends in Neuroscience* **7**(12): 460–62

Fanselow M S, Bolles R C 1979 Triggering of the endorphin analgesic reaction by a cue previously associated with shock: Reversal by naloxone. *Bulletin of the Psychonomic Society*. **14**(2): 88–90

Gallo M, Candido A 1995 Dorsal hippocampal lesions impair blocking but not latent inhibition of taste aversion learning in rats. *Behavioral Neuroscience* **109**: 413–25

Gormezano I, Kehoe E J, Marshall B S 1983 Twenty years of classical conditioning research with the rabbit. *Progress in Psychobiology and Physiological Psychology* **10**: 197–275

Han J S, Gallager M, Holland P C 1995 Hippocampal lesions disrupt decrements but not increments in stimulus processing. *Journal of Neuroscience* **15**: 7323–9

Hoehler F K, Thompson R F 1980 Effect of the interstimulus (CS–UCS) interval on hippocampal unit activity during classical conditioning of the nictitating membrane response of the rabbit (*Oryctolagus cuniculus*). *Journal of Comparative Physiology and Psychology* **94**: 201–15

Kamin L J 1969 Predictability, surprise, attention and conditioning. In: Campbell B A, Church R M (eds.) *Punishment and Aversive Behavior*. Appleton-Century-Crofts, New York

Kim J J, Krupa D J, Thompson R F 1998 Inhibitory cerebello-olivary projections and the blocking effect in classical conditioning. *Science* **279**: 570–3

Mauk M D, Steinmetz J E, Thompson R F 1986 Classical conditioning using stimulation of the inferior olive as the unconditioned stimulus. *Proceedings of the National Academy of Sciences USA* **83**: 5349–53

Rescorla R A, Wagner A R 1972 A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black A H, Prokasy W F (eds.) *Classical Conditioning II: Current Research and Theory*. Appleton-Century-Crofts, New York

Rickert E J, Lorden J F, Dawson R, Smyly E 1981 Limbic lesions and the blocking effect. *Physiology and Behavior* **26**: 601–6

Sears L L, Steinmetz J E 1991 Dorsal accessory inferior olive activity diminishes during acquisition of the rabbit classically conditioned eyelid response. *Brain Research* **545**: 114–22

Solomon P R 1977 Role of hippocampus in blocking and conditioned inhibition of rabbit's nictitating membrane response. *Journal of Comparative Physiology and Psychology* **91**: 407–17

Thompson R F 1986 The neurobiology of learning and memory. *Science* **233**: 941–7

I. Orduña and M. Gluck

Bloomfield, Léonard (1887–1949)

Leonard Bloomfield, one of the key figures of American linguistics, was born in Chicago on April 1, 1887. He studied German philology and linguistics at the University of Wisconsin and Chicago, where he got his Ph.D. in 1909. In the same year, he got married to Alice Sayers. After serving as teaching assistant of German at the Universities of Cincinnati and Illinois, he went for graduate studies in 1913 and 1914 to the Universities of Leipzig and Göttingen. From 1913–21 he was Assistant Professor of Comparative Philology and German at the University of Illinois, and from 1921–27 Professor of German and Linguistics at Ohio State University. In 1925 Bloomfield was one of three linguists who founded the Linguistic Society of America, which soon grew into the most influential organization of the field. From 1927–40, Bloomfield was Professor of Germanic Philology at the University of Chicago.

During this period, the publication of his widely read book *Language* (1933), his growing influence on the field of general linguistics, and his international reputation made him one of the leading representatives of American linguistics. Although he was not a charismatic person, easily dominating an audience, but rather of reserved and unimposing personality, he still had an enormous impact on American linguistics. He had few students, but many linguists regarded themselves as members of a school bearing his name. His unique effect was completely due to sober argument and the strength and clarity of his writings. In 1940, he became Sterling Professor of Linguistics at Yale University. Applying his theoretical insights to practical concerns, Bloomfield participated actively in the Intensive Language Program of the American Council of Learned Societies, preparing textbooks and

courses of strategically important languages during World War II. In May 1946 he suffered a heavy stroke from which he never fully recovered. He died on April 18, 1949 in New Haven, Connecticut.

1. *Background of American Linguistics*

Although grounded in the great tradition of European historical and comparative linguistics, which had revealed the systematic relationship among the Indo-European languages, American linguistics developed its own, independent orientation during the first decades of the twentieth century. Two major factors were characteristic. First, the interest in the analysis of a wide variety of languages outside the Indo-European family, to which European linguistics was largely restricted. The primary motive for this extended perspective was the concern for the language and culture of American Indians, the systematic study of which was initiated by the anthropologist Franz Boas (1858–1942). Among the active followers of Boas were Edward Sapir (1884–1939) and Bloomfield, scholars who shaped the subsequent development. Growing interest in languages without written records raised methodological issues unfamiliar to European his torical linguistics. While Hermann Paul (1846–1921) in his influential book *Prinzipien der Sprachgeschichte* programmatically denied the possibility of a scientific approach to language other than the historical one, the American-Indian languages called for exactly such an approach, for an analysis that is independent of written historical sources. The systematic foundation of synchronic description as a scientific approach leads to the second characteristic of American linguistics, viz. the development of a specific descriptive methodology, whose most rigorous form became known as 'distributionalism.'

This interest in methodological clarification of descriptive linguistics corresponds to a large extent to similar developments in European linguistics at the turn of the twentieth century, initiated especially by Ferdinand de Saussure (1857–1913) with his *Cours de linguistique générale* (1916). American linguistics made, however, its specific contribution to the overall development called 'structural linguistics.' Many of the inspiring ideas stimulating this development were due to Edward Sapir. The classical formulation of the framework, however, that by its systematic character eventually shaped the development, was due to Leonard Bloomfield. In spite of their mutual respect and scientific appreciation, there is an interesting tension between the personality as well as the intellectual approach of these two important men. While Sapir was burning with brilliant ideas, Bloomfield shaped the field by careful and systematic argument. The tension is inherent, moreover, in Bloomfield's own contribution, where some friction between theoretical foundation and factual insights will be observed.

Trained in Indo-European philology with a special emphasis on Germanic languages, Bloomfield aimed from the very beginning at a systematic foundation of his field of interest, including the need of coming to grips with problems posed by languages outside the standard canon of traditional linguistics. In his first monograph, *An Introduction to the Study of Language* (1914), he still assumed that this task should be accomplished within the overall framework of European linguistics, with Wilhelm Wundt's notion of *Völkerpsychologie* as theoretical background for the study of language.

2. *Behaviorist Orientation of Linguistics*

A radical change in Bloomfield's attempt to turn linguistics into a science by general methodological standards is marked by the axiomatic approach he adopts in *A set of postulates for the science of language* (1926), where the psychological orientation of European linguistics is replaced by a strictly operationalist conception, with which he became familiar through the work of his colleague Albert Paul Weiss at the University of Chicago. This perspective, systematically elaborated in Bloomfield's most influential *chef d'œvre, Language*, has a characteristic ambivalence, or even paradoxy.

On the one hand, Bloomfield criticizes Hermann Paul and other European linguists not only for their exclusively historical orientation, but even more for their imprecise psychological theory on which their linguistic concepts are based, a perspective that Bloomfield himself had adopted in his *Introduction* in 1914. He now considered the scientific analysis of language as a strictly objective matter, without any commitment to psychological concepts. Under this perspective, he took it to be completely irrelevant for descriptive linguistics, which sort of psychology one adheres to. Linguistics should be established as an autonomous discipline, that need not rely on any concepts other than its own.

On the other hand, Bloomfield was strictly opposed against one particular type of psychology, insisting on a rigorous exclusion of all mentalistic concepts from scientific description and explanation. He considered any resort to internal states and mental accounts of linguistic (or other types of) behavior as unwarranted at the present state of knowledge, and presumably unnecessary and hence unpermissible for scientific accounts in principle. In fact, Bloomfield advocated what he called a mechanistic explanation of behavior, opting more specifically for Weiss' version of behaviorism. Both aspects of this ambivalent position—the behaviorist account of language as well as the principle of autonomous, formal analysis—are

worked out carefully and applied to all kinds of concrete problems in the 28 chapters of *Language*, supporting his unusual influence during the decades to come.

Bloomfield's proposal to extend the standard behavioristic stimulus-response schema to linguistic behavior is original, instructive—and misguided. Assuming that the conditioned response R caused by a stimulus S in the standard schema S → R can be mediated by a linguistic act, Bloomfield arrives at the following proposal: An external stimulus S causes a linguistic response r of the speaker, which acts as a linguistic stimulus s on the hearer causing the practical response R. This yields the complex schema S → r … s → R. Bloomfield illustrates this by a situation, where S is the condition of being hungry and seeing an apple on a tree, while R is the act of plucking and eating it. This connection can now be mediated by the act r of making some appropriate noise, transcribed as, e.g., 'I'm hungry,' and the perception s of that very noise. Thus, the dotted line between r and s represents the interindividual relation connecting the production and perception of a given phonetic form, while the arrow between S and r as well as s and R represents the intraindividual association of this form with the practical stimulus imposed on the speaker and reaction produced by the hearer, respectively. What is misleading in this proposal is in particular the fact that the features of S and R, which Bloomfield considers and defines as the meaning of the form produced by r and perceived as s, cannot have any systematic relation to the respective s and r, whose formal features must, of course, be identical in the relevant respects. Thus, the meaning of 'I'm hungry' would not be different, if there is no apple present, or if the hearer, being on a diet, doesn't want to eat. This observation is by no means restricted to Bloomfield's original example, but applies to the relation between language and the conditions of use in a rather principled way. In other words, the behavioristic attempt to account for linguistic meaning in terms of conditioned use of linguistic forms was doomed to fail, even though it was taken for granted by a whole generation of American linguists and psychologists, including Skinner (1957) in his programmatic *Verbal Behavior*.

One particular consequence of Bloomfield's approach was his strong conviction that a systematic theory of meaning could only be developed on the basis of a complete, explicit account of all possible aspects of speaker and hearer's environment and the physiological mechanisms determining the pertinent reactions. For this reason, he considered the scientific treatment of semantics out of reach, at least for the time being, and perhaps in principle, although he had no doubts about the central role of meaning for any satisfactory approach to language. For three decades, skepticism towards the possibility of a serious theory of meaning became a characteristic feature of American as opposed to European linguistics. There are residues of this skepticism even in Chomsky's *Syntactic Structures* (1957), initiating a radically different approach that paved the way of an explicitly mentalistic theory of language.

3. *Structural Analysis of Language*

Restricting the central task of linguistics to the descriptive analysis of (r … s), i.e., to the properties of linguistic form, Bloomfield developed a systematic and comprehensive account of the properties linguistic expressions might possibly exhibit. The central idea of the framework carefully laid out in *Language* is the assumption, that all relevant aspects of verbal utterances can be captured on the basis of strictly formal criteria, identifying their parts in terms of articulatory and perceptual distinctions and their subsequent classification according to their possible occurrences. More specifically, basic linguistic forms, called morphemes, are characterized as sequences of phonemes, where phonemes are classes of minimal segments identified by their articulatory properties (roughly corresponding, incidentally, to letters in alphabetic writing). Morphemes can then be grouped into form classes on the basis of their distribution, i.e., observable combination, giving rise to the properties of so called constructions, which are the invariant skeleton of increasingly complex utterances.

Methodologically, this framework was based on systematic segmentation and classification in terms of strictly acoustic and distributional criteria, applied to actual utterances in the sense of well-defined types of physical events. In this way, segments of the acoustic stimulus are classified as phonemes, sequences of phonemes are identified as morphemes, morphemes are grouped into several classes, particular sequences of which are classified as constructions. The details of this elegant framework—the relevant criteria for segmentation and classification, the properties of primitive and complex elements, and the specificity of the pertinent levels of structural analysis—have been discussed intensively and worked out carefully by a whole generation of linguists in the decades following the publication of Bloomfield's influential book.

A rigorous and comprehensive version of the overall approach was arrived at in Harris's (1951) *Methods in Structural Linguistics*, where the canon of distributionalism is formulated in the most systematic and radical form. Bloomfield's own exposition developed the major ideas with remarkable sensitivity and respect for traditional grammatical analyses, preserving their insights where possible and correcting them where necessary, identifying their appropriate place under the new, systematic perspective. This holds for traditional notions of descriptive grammar like parts of speech, morphological categories like gender, number, tense, etc., but particularly for the important discoveries establishing the historical relationship among

different Indo-European languages. In fact, more than one third of Bloomfield's seminal book deals with topics of comparative and historical linguistics, with different types of language change, and problems of dialect geography, i.e., issues that go beyond purely descriptive analysis. Bloomfield's intimate familiarity with the history of the field, combined with his capacity to fundamentally reconstruct its aims and methodology, explains to a large extent his unique impact.

In order to accommodate various grammatical phenomena treated in traditional linguistics, Bloomfield applies the behaviorist notion of habits controlling the speaker's behavior. A habit corresponds to a pattern of linguistic forms. Thus, speakers of English have a habit according to which finite verb forms agree with their subject, as in *the boy walks* vs. *the boys walk*. The paradigms of inflectional forms like *walk*, *walks*, *walked*, *walking* involved in these habits can be accounted for by systematic combinations of 'free' and 'bound' morphemes, such as *walk* and *–s*, *–ed*, or *–ing*. The patterns set up on this basis can then be extended to 'irregular' forms like *sing*, *sings*, *sang*, *sung*, *singing*, where the free form *sing* is combined with an alternation of the vowel instead of the affix *–ed*, yielding the form *sang*, and to even further irregularities like *go*, *goes*, *went*, *gone*, *going*, where went substitutes the combination *go-ed*. In a similar way, the systematic character of habits is supposed to control patterns of variation in the phonetic realization of linguistic forms. Thus, Bloomfield argues that the masculine and feminine form of French adjectives like *plat*vs. *platte*, phonetically [pla] vs. [plat] 'flat,' *gris* vs. *grise*, phonetically [gri] vs. [gri:z] 'gray,' *frais* vs. *fraîche*, phonetically [frɛ] vs. [frɛ:š] 'fresh,' and many others, are best be described by taking the feminine as the underlying form plus 'the simple statement that the masculine form is derived from the feminine by means of a minus feature, namely, loss of the final consonant' (Bloomfield 1933). A second operation must furthermore shorten the long vowel in the resulting forms [gri:] or [frɛ:], etc. in order to derive the actual masculine [gri], [frɛ], etc. In other words, an additional vowel shortening is necessary after the previous deletion of the final consonant.

4. *Beyond the Limits of Distributionalism*

These apparently minor details of descriptive technology are worth mentioning because of their far-reaching consequences. The assumption of an underlying form plus operations deriving its surface realization clearly stretches the notion of habit determining linguistic behavior, as the actual response must now be taken to depend on previous operations applying to underlying forms that control the articulatory patterns at best indirectly, if at all. Assuming underlying forms plus operations or rules of realization is obviously a step towards the recognition of internal, which means mental, representations, and operations, a notion that Bloomfield wanted to rigorously exclude from scientific methodology. But he goes even further in the prohibited direction. In his ingenious *Menomini morphophonemics* (1939), an analysis of an Algonquian dialect, he set up underlying forms and various operations of realization with by far more intricate conditions on their relative dependence than the simple case of French masculine adjectives. This dependence, which Bloomfield was very much concerned with, requires an ordering of the operations up to five steps. It is obvious that internal operations of this complexity cannot be accommodated to the notion of stimulus–response–habit in any serious sense. What is fascinating here, is first the depth of analytic insight of Bloomfield's analysis, and second his decision in case of a conflict between behaviorist methodology and descriptive adequacy. Although he was, of course, not aware of the conflict, clearly he discarded the orthodoxy in favour of the insight.

In the same spirit, Bloomfield (1933) argued implicitly for an ultimately mentalistic theory of sound change, summarizing the core of his opinion by the slogan 'Phonemes change.' This means in fact that the underlying pattern, rather than the acoustic realization, is relevant for the processes of linguistic change. In a far-sighted conclusion, based on his *Menomini morphophonemics*, he even envisaged the possibility that sound change may consist of the introduction of rules of alternation: 'our basic forms do bear some resemblance to those which would be set up for a description of Proto-Algonquian, some of our statements of alternation … approximate the historical development from Proto-Algonquian to present-day Menomini (1939). Chomsky and Halle (1968) have later developed this view systematically in generative phonology.

5. *Roots of a Mentalistic Conception of Language*

Thus, Bloomfield insisted on a strictly antimentalistic program of linguistic analysis, according to his opinion that 'Non-linguists … constantly forget that a speaker is making noise, and credit him, instead, with the possession of impalpable "ideas".' It remains for the linguist to show, in detail, that the speaker has no 'ideas' and that the noise is sufficient (1936). But while he defended this dogma by practically excluding semantics from the actual linguistic research agenda, he was, surprisingly, not able to exclude mentalistic concepts from the domain of his very central concern—the analysis of linguistic form: underlying forms, derivational rules, precedence order of operations, and even phonemes as parts of underlying forms are concepts that go beyond the mechanistic elements extracted from noise. This conflict between rigid methodological principles on the one side and

explanatory insights on the other mirrors one aspect of the above mentioned tension between Bloomfield and Sapir, in the following sense: although both scholars agreed on the general principles of linguistics as a scientific enterprise, including insights in relevant detail such as the description of Indian languages, or the nature of sound change, Sapir always insisted on the psychological nature of linguistic facts. This is clear from his *Sound patterns in language* (1925), and even more programmatically from *The psychological reality of phonemes* (1933). As a matter of fact, the orientation of Sapir's inspiring essays amounted to a mentalistic research strategy.

Thus, although Bloomfield in 1940 was Sapir's successor as Sterling Professor of Linguistics at Yale University, it was Sapir's notion of language as a psychological reality that dominated the subsequent period of post-Bloomfieldian linguistics, as shown, e.g., by the title and orientation of *The Sound Pattern of Russian* by Halle (1959) and *The Sound Pattern of English* by Chomsky and Halle (1968).

But the tension persists. While Sapir and Jakobson provided ideas and orientation of this development, it is still Bloomfield's radical clarification of the conceptual framework and descriptive technology that is the indispensable foundation of modern linguistics. And perhaps more importantly, his *Menomini morphophonemics* is the paradigm case not only of a complete descriptive analysis given in terms of an explicit, complex rule system, but also of an impressive victory of clear insight over methodological orthodoxy. Conceding the amount of mentalistic machinery implied in this analysis of sound structure, strong theories of syntax and even semantics have in the meanwhile been shown to be possible and productive—in spite of Bloomfield's antimentalistic bias.

Bibliography

Bloch B 1949 Leonard Bloomfield. *Language* **25**: 87–98

Bloomfield L 1914 *An Introduction to the Study of Language*. Holt, New York

Bloomfield L 1926 A set of postulates for the science of language. *Language* **2**: 153–64

Bloomfield L 1933 *Language*. Holt, New York

Bloomfield L 1936 Language or ideas? *Language* **12**: 89–95

Bloomfield L 1939 Menomini morphophonemics. *Traveaux du cercle linguistique de Prague* **8**: 105–15

Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands

Chomsky N, Halle M 1968 *The Sound Pattern of English*. Harper and Row, New York

Halle M 1959 *The Sound Pattern of Russian*. Mouton, The Hague, The Netherlands

Harris Z S 1951 *Methods in Structural Linguistics*. University of Chicago Press, Chicago

Hockett C F 1970 *A Leonard Bloomfield Anthology*. Indiana University Press, Bloomington, IN

Paul H 1880 *Prinzipien der Sprachgeschichte*. Niemeyer, Halle, Germany

Sapir E 1925 Sound patterns in language. *Language* **1**: 37–51

Sapir E 1933 La Réalité psychologique des phonèmes. *Journal de Psychologie Normale et Pathologique* **30**: 247–65

de Saussure F 1916 *Cours de linguistique générale*. Payot, Paris

Skinner B F 1957 *Verbal Behavior*. Appleton-Century-Crofts, New York

Weiss A P 1924 *A Theoretical Basis of Human Behavior*. R. G. Adams, Columbus, OH

M. Bierwisch

Boas, Franz (1858–1942)

1. Biography

Franz Boas was born on 8 July, 1858, in Minden, in the German province of Westphalia. He was the son of freethinkers of Jewish origin who still cherished the ideals of the democratic revolution that had taken place in Germany in 1848. Coming from such an educated bourgeois background (*Bildungsbürgertum*), he had access to books on the natural sciences and geography, and this environment had a profound impact on his future aspirations. From 1877 onwards, he studied natural sciences and geography, first in Heidelberg and then in Bonn and Kiel. After completing his Ph.D. in physics (on the color of water) he went to Berlin, where he began to move in the academic and social circles characteristic of much of the German scientific community at that time. A major role model for him was Rudolf Virchow, the professor of medicine and anthropology.

The most decisive period for Boas was that from 1881 to 1887. In Berlin he frequented the meetings of a learned society, the *Berliner Gesellschaft für Anthropologie, Ethnologie, und Urgeschichte*, and sought to raise funds for research in North America while preparing intellectually for such a trip. In 1883–4, after a short period of service in the German army, he began to conduct field research among the Inuit of Baffin Island. This prompted him to shift his focus of interest from the landscape to its inhabitants. He found that the geographical determinism of his time did not hold up against his observations of the central Inuit, in that some of the behavior he observed was exhibited in spite and not because of the environment. After a stay in New York, he returned to Berlin to become assistant at the Museum of Ethnology in 1885. There he met a group of Bella Coola Indians who had been brought to Berlin as a living exhibit (*Völkerschau*). This encounter prompted his decision to study the American Indian peoples of the northwest coast. In 1886 he embarked on the first of his many research trips to this region, where he gained rich insights into the social organization and self-expression of the Kwakiutl. On his way back he stopped in New York, where he met the editor of *Science* magazine and was immediately appointed assistant editor. Thus, he now had the means to marry

Marie Krackowitzer, a woman from New York's liberal German-speaking community, to whom he had been introduced by his uncle. He decided to remain in America, despite the fact that in 1882 he had written that 'his dearest aim' was to become a German professor. Even then, he had known that if the USA offered him 'a better chance … I would go without hesitation.' A strong motivation for staying was that American science was, as he saw it, 'so far behind in comparison to Europe that a young man as I am, is able to accomplish much more' (Liss 1996). There was less teaching of anthropology or ethnology than in German universities, but the strong role of the natural sciences was to impact on the development of anthropology as an academic discipline. Boas' understanding of anthropology was that of an empirical science, part of a wider trend in the social sciences, oriented towards the works of Mach, Dewey, and James (Lowie 1956, p. 1015).

Boas' experiences in the field resulted in a rather rare intellectual approach, that of a commuter who kept travelling between different nations and their intellectual environments. He thus saw different milieus as an outsider and observer, sometimes integrated and engaged, sometimes detached. Over the course of his life he showed a strong sympathy for other such commuters. For example, he greatly encouraged George Hunt, brought up among the Kwakiutl by a Tlingit mother and a British father, to do his own fieldwork among the Kwakiutl, and also provided the funding for this work. Later, too, he was active in support of those moving between different environments, be they immigrants to the USA or people living on the fringes of their home community. Boas himself experienced the advantages of being an outside observer, maintaining that such a status sharpens one's perception of those features of a culture which are self-evident to those living within it. He was very conscious of diversity, variation, and individual differences.

Boas' basic research motivation was one for which no assumption of cultural heritage was needed: curiosity combined with ambition. For Boas, it was a particular thrill to find complex answers to simple-looking phenomena that others considered to be obvious. He stated that he was 'motivated by the effective appeal of a phenomenon that impresses us as a unit, although its elements may be irreducible to a common chance. In other words the problem that attracted me primarily was the intelligent understanding of a complex phenomenon (cf. Herskovits 1953, p. 10).

Boas assumed an entrepreneurial role in the organization of US anthropology when, in 1896, he became assistant curator of the American Museum of National History in New York (curator, 1901–5), and lecturer in physical anthropology at Columbia University (professor of anthropology, 1899–1936). Among other activities, he modernized the learned journal *The American Anthropologist* in 1898, and helped to found the American Anthropological Association in 1902. He was not only interested in founding institutions (indeed, others tried to do this as well), but in insuring their professionalization. It was this specific approach (which also brought him into conflict with those who had hoped for a broader amateur membership of the AAA), which resulted in the lasting success of his institutional activities. It should not be forgotten that the period between the Civil War and the 1920s was a great era for entrepreneurs in and around the sciences. Many institutions of higher learning, new disciplines, scientific journals, and learned societies were founded, and funding raised, but most of these efforts failed in the long run. Today, it is difficult to appreciate that Boas' efforts were unusually successful, as we only notice those institutions which have survived.

2. *Contributions to Anthropology*

Boas' major and lasting contribution lies in the methodological domain. He was always on the look out for differences and, as such, was less interested in the contemporary transformations of Kwakiutl Indian culture than in what he perceived to be its remembered past. His quest for difference also made him sensitive to variations within the ensemble of aspects which the informants or outsiders described as a culture. Any boundaries and standards were suspect to him. He raised certain issues 90 years before they were advocated by the postmodernists of the 1980s (Jacknis 1996). He was particularly skeptical about classification, and naive subjective classification aroused his anger (see for e.g., 'The principles of ethnographic classification'). A systematic discussion of the problem can be found in his linguistic paper on 'Alternating sounds'. Boas contended that people tend to classify according to familiar categories, but that anthropologists should be concerned with what was later called emic distinctions—the distinctions made by the people themselves.

Boas' empiricism shielded his work from the speculative thinking which dominated the discipline in the German-speaking countries. This methodological rigor contributed to the strong and lasting reputation acquired by American anthropology, particularly in comparison to other social and natural sciences. It was manifest in the way Boas and many of his pupils discussed and validated the results of their fieldwork. More than any other author of his time, he was conscious of two influences on anthropological sources: that of the method chosen and that of the observers' perspective, including artifacts created by the methods. He gave preference to texts emanating from the culture itself, and asked his interlocutors to produce and record texts of all kinds, not only ceremonial ones. Moreover, he carefully recorded the effects of his methods, e.g., how the phonograph was

used, and how different informants influenced the style and content of dictated or written texts. The aim of this never-ending endeavor was to achieve objectivity. For Boas, a command of the native languages was essential in order to 'understand directly, what the people … speak about, what they think, and what they do' (Boas 1906, pp. 183–8). His concept of objectivity introduced a clear frame of reference, distinguishing his approach from that of his German and American contemporaries—a frame of reference rooted within the culture itself.

The results of Boas' fieldwork may suggest to many readers that he spent prolonged periods within the cultures under investigation, participating intensively in their everyday activities. The reality was different. His most important results were yielded by repeated short stays in the field—up to three weeks in one place. As the letters to his family show, he profited from the contrasts he experienced, seeing the surprising and the colorful in what was self-evident to the local population. The value of his reports is rooted in his insistence that each phenomenon should be viewed from different perspectives, and evidence gathered from different informants. He also integrated the latest technological methods into his approach—photography as early as 1883, and sound recording in 1887.

One of Boas' research topics was largely disregarded by the subsequent generation, namely, the place of the individual within society. Boas' interest in this subject may be seen as a legacy of the liberal German milieu in which he grew up. It was in front of the German Society of New York that he stated in 1880: 'The invention is not difficult. Difficult is the retention and further development … . It is important to observe the fight of individuals against tribal customs. The same kind of struggle the genius has to undergo … in his battle against dominant ideas' (cf. Herskovitz 1953, p. 75). He continued to address this question later on, in his interpretation of esoteric cults, insisting that cults should be studied in connection with the broader base of belief in society as a whole. It would be an error, he argued, to see esoteric beliefs 'as the only true form of the inner life. They might as well be considered as a reaction … to the general cultural environment.'

The anthropological subject most intimately linked to Boas' name, and the one which prompted most references to his work in the second half of the twentieth century, was the *potlatch*. 'Potlatch' is a Chinook word for festivities, which Boas used to describe a group of redistributive ceremonies performed by the Kwakiutl. Redistribution was not the dominant feature in all cases, however. As he emphasized in his very first discussion of the subject in 1899, the creation of credit and obligation was also a central element. Moreover, the redistribution of wealth created prestige as well as obligation.

To judge from the attention he gave to the potlatch, Boas was clearly fascinated. His curiosity about otherness was captured by a phenomenon which, with historical distance, can be seen to be a travesty of one particular feature of (non-native) American civil life—too conspicuous consumption. (Both private universities and museums were dependent on such redistributions of wealth, and on the fact that benefactors were, in turn, endowed with prestige.) The Canadian administration did not recognize society's need to have ways of creating prestige. Nor did it understand that the indebtedness created by such ceremonies was, in fact, a creation of wealth and capital not entirely different from that of the modern banking system, in which the capital mobilized as credit by far outweighs the liquid capital deposited. Boas brought these arguments into the foreground in order to defend the Kwakiutl and neighboring groups against intrusion by the Canadian administration.

Later critics observed that the potlatch, as portrayed in Boas' works, was just one momentary impression in a rapidly changing situation (Mauzé 1986, pp. 21–63), and that Boas neglected other aspects of the economy, mainly land tenure and the livelihood of classes other than the nobility. However, these criticisms cannot diminish the importance of his discovery of a system of attributing merit which was, in itself, economically organized.

Boas was also skeptical about the evolutionism of his times, insofar as it implied a ranking of more or less progressive cultures. How was one to classify a culture which had a seemingly simple technology, but an extremely intricate social structure? For Boas, it was not acceptable to label these as 'early forms' of humanity. Assumptions about the coincidence of linguistic communities, biological inheritance, and culture were implicit in evolutionism and explicit in historical reconstructions of the 'genealogies of cultures.' Yet using material gathered from American Indians, Boas was able to show that linguistic communities, groups of descent, and culture do not necessarily coincide, and that they have in fact varied independently of each other in many cases. Thus the evolution of a given language or language group may run across different groups of descent which, over the course of time, may become members of different linguistic families.

In the first half of the twentieth century, Boas' contribution to science was most strongly identified with his work on physical anthropology. Here, he introduced the perspective of regarding differences in the human races as a phenomenon of domestication, rather than in analogy to differences in animal species. He maintained that nutrition, the instrumental and ornamental use of the body, crossing, and selection are the social facts shaping the variant physical appearance of humankind. Of his many empirical investigations, a study published in *The American Anthropologist* in 1912 is particularly notable. In this study, based upon a huge sample, he compared the head shapes of immigrants and their children. The shape of the head had, until then, been considered the best

indicator of hereditary lines. Yet Boas was able to show that these measurements changed considerably in the children's generation. This challenged the dogma that the physical type was determined by heredity alone.

Boas did as much as he could to collect texts, especially songs, tales, and myths, and to encourage others to do the same. He maintained that such materials have the merit of bringing out 'those points which are of interest to the people themselves' (Boas 1916, p. 393). He objected to other authors classifying genres according to their subjective standpoint or the standpoint prevalent in their own culture, and referred instead to the categories used by the people themselves (emic categories)—a procedure which, even today, is not always followed. He rejected explanations of symbolism which were not based on explicit data from the culture itself as pure speculation, whether they be such as those of the German *Kulturhistorische Schule*, or Freudian. In contrast, he placed particular weight on the recognition of myths and tales as artistic creations, arguing that these should not only be a source of anthropological research, but valued as a product of human creativity.

Boas' attitude to politics was interested but rather detached. The liberal German milieu in which he grew up advocated the separation of science from religion and politics. The sociologist Max Weber, who later matured in the same milieu, formulated the idea of value freedom in science. Having experienced the repression of such scientific freedom by the institutions of power and religion in the nineteenth century, the liberal scientific community saw this conception of science as an important basis for the growth of scientific endeavor and discourse—a neutral medium for the way society saw itself and for societal development. Detachment did not mean indifference, however. Boas had already described the goal of science to be 'the ice-cold flame of the passion for seeking the truth.' Moreover, observations of prejudice called for scientists to speak out. On several occasions, Boas spoke out against racial prejudice, defending the American 'Negro,' underlining the evidence of high cultural achievement in the African past, criticizing the US and Canadian policies which discriminated against Indian customs (namely the potlatch), censuring war and, last but not least, condemning anti-Semitism. His texts against anti-Semitism were printed in a widely read German newspaper (*Frankfurter Zeitung*) in 1926 and later distributed as leaflets by the German anti-Nazi resistance.

3. *Influences*

Boas' legacy cannot be identified with a particular theory. In contrast to other authors, after his investigation of the Inuit he produced few texts which attempted to provide a general ethnography. The organization of findings in his texts often obscures information rather than elucidating it. Central theoretical arguments are typically found neither in the introduction nor in the summary, but somewhere in the middle of the text, sometimes as footnotes. He did, however, introduce some important methodological approaches. It was Boas who cross-bred physical anthropology with ambitious statistical methods. He and his pupils were the first to take linguistic approaches, looking for the categories by which language was organized in the mind. This later lead to the study of phonemics with his pupil Edward Sapir, and later still, to the idea of the categorial grammar which takes Boas' perceptive of trying to find by induction categories immanent to the language itself, rather than transplanting the categories of Indo-German grammars.

In the decade after his death, more was written about Boas than about any American anthropologist before him. This may have created the impression of a Boas School. Indeed, his pupils included such anthropologists as Ruth Benedict, Melville Herskovits, Robert Lowey, Margaret Mead, Ashley Montague, and Edward Sapir. Later, however, harsh criticisms such as those of Lesley White exposed some limitations in Boas' work. His didactical methods apparently did not meet the standards of the North American universities. Moreover, considering that those who studied under him represented 'the greatest variety of interests, methods and theoretical orientation,' one cannot but agree with Herskovits' early statement (1953, p. 23) that 'the term "Boas' school" … is a misnomer.'

However, one cannot overlook the fact that functionalism and its later variations such as conflict theory owe their success to the methodological rigor introduced in Boas' aversion to speculative anthropology, and to the institutional foundations he laid. Evidently, Boas and Malinowski—the leading Anglo-Saxon anthropologist of the time—tended to ignore each other, although both were trained in the German approach to the natural sciences. In the few remarks he made on Malinowski and functionalism, Boas named the close observation of daily life as a merit, though underlining that fieldwork was already an established practice in US anthropology before the beginning of the twentieth century. He criticized the explanation of human behavior on the basis of the actual social environment alone, the approach taken in earlier British functionalism. Boas felt reference to historical data to be indispensable for an understanding of ongoing processes (cf. Boas 1938).

It was thanks to Boas that exotic presentations of unexplainable behavior and the reduction of unfamiliar cultural forms to an irreducible genetical heritage were gradually excluded as forms of scientific explanation. He brought the assumption of rationality to the foreground, making this the starting point of

anthropological explanation. This does not imply that all human behavior is rational, or that all social institutions are functional (in fact, Boas abhorred such generalizations). However, it does mean that explanations of individual behavior and culturally encoded behavior should relate to the goals, cultural strategies, and environmental conditions of the informants. Of course, behavior may still fail to achieve its goals or may even be erratic in view of all possible goals.

The 1980s and 1990s saw a return to the skepticism which distinguished Boas' approaches from those of most earlier theorists. His legacy is an antidote to reification (*Verdinglichung*), reminding us that any culture is multifaceted and continuously changing, and that individuals may row against its currents.

See also: Anthropology, History of; Diffusion: Anthropological Aspects; Diffusion: Geographical Aspects; Diffusion, Sociology of; Ethnography; Evolution: Diffusion of Innovations; Exchange in Anthropology; Exchange: Social; Potlatch in Anthropology

Bibliography

Boas F 1885 *Baffin-Land: geographische Ergebnisse einer in den Jahren 1883 und 1884 ausgeführten Forschungsreise*. Petermanns Mitteilungen 17, Supplement 80. Perthes, Gotha, Germany

Boas F 1888 *The Central Eskimo*. 6th Annual Report, US Bureau of American Ethnology, Washington, DC, pp. 399–669

Boas F 1906 Some philological aspects of anthropological research. In: Stocking G W (ed.) 1974 *A Franz Boas Reader: The Shaping of American Anthropology 1883—1911*. University of Chicago Press, Chicago, pp. 267–81

Boas F 1909 *The Kwakiutl of Vancouver Island*. Yesup North Pacific Expedition, Brill, Leiden, Germany, Vol. 5, pt. 2.

Boas F 1911 *The Mind of Primitive Man*. Macmillan, New York

Boas F 1912 Changes in bodily form of descendants of immigrants. *American Anthropologist* **14**: 530–62

Boas F 1916 *Tsimshian Mythology*. 31st Annual Report, 1909—1910, US Bureau of American Ethnology, Washington, DC

Boas F (ed.) 1938 *General Anthropology*. Heath, Boston and New York

Boas F 1940 *Race, Language and Culture*. Macmillan, New York

Herskovits M J 1953 *Franz Boas: The Science of Man in the Making*. Scribner, New York

Lowie R 1956 Reminiscences of anthropological currents in America half a century ago. *American Anthropologist*. 995–1015

Mauzé M 1986 Boas, Les kwagul et le potlatch. *L'homme* **100**: 21–63 [including comments by Claude Meillassoux and other authors]

Ryding J 1975 Alternatives in 19th century German ethnology: A case-study in the sociology of science. *Sociologus* **25**, **1**: 1–28

Stocking G W (ed.) 1974 *A Franz Boas Reader: The Shaping of American Anthropology 1883—1911*. University of Chicago Press, Chicago

Stocking G W jr. (ed.) 1996 *Volksgeist as Method and Ethic. Essays on Boasian Ethnography and the German Anthropological Tradition*. University of Wisconsin Press, Madison, WI

G. Elwert

Body: Anthropological Aspects

The theoretical status of 'the body' in cultural anthropology has evolved in roughly four stages: (a) in earlier decades of the twentieth century it was an implicit, taken-for-granted background feature of social life; (b) beginning in the 1970s the body became an explicit topic of ethnographic concern; (c) from the 1980s the body began to become a problem to be accounted for with respect to its cultural and historical mutability; and (d) by the century's end the body offered a theoretical locus for rethinking various aspects of culture and self. This conception of the changing status of the body in anthropology guides the exposition that follows.

1. Early Reflections on the Body

Although the body was traditionally implicit in anthropological writing, a rereading of older sources is nevertheless likely to offer a surprising richness of insights on bodiliness. Paul Radin, for example, explicitly discusses indigenous philosophy of the body in a chapter on 'The Ego and Human Personality' (1927). In what we would today call an attempt to formulate an ethnopsychology of the person, he emphasized the Maori distinction between material (substance) and immaterial (form) aspects of the body as the resting place of various components of the person. In its anatomical aspect the Maori body was associated with psychical rather than physiological functions. After going on to discuss the person among the Oglala Dakota and the Batak, he concludes that these ethnotheories display an 'inability to express the psychical in terms of the body,' but must project it onto the external world (1927, pp. 273–4). Robert Lowie, in a chapter on 'Individual Variability' in religious experience, dealt with issues relevant to embodiment under the heading of 'sensory types' (1924, pp. 224–31). His data for this discussion were instances of visionary or revelatory imagery. Noting the absence of studies in this area by anthropologists, he was concerned with the manner in which imagination is concretely engaged with the various sensory modalities. Drawing primarily on Crow material supplemented by evidence from other North American tribes and from the African Bushmen, he suggested that the cross-cultural data conformed with then-current psychological findings that visual ima-

gery predominates in frequency over imagery in other sensory modalities, though certain individuals might show a propensity toward auditory or even motor, tactile, and kinesthetic imagery. Maurice Leenhardt (1947) reports a conversation between himself and an elderly New Caledonian philosopher about the impact of European civilization on the indigenous cosmocentric world. His interlocutor said that 'What you've brought us is the body.' For Leenhardt this pronouncement upends a stereotype that presumes the body lies on the side of nature (the primitive) and spirit on the side of culture (the civilized). Leenhardt vividly suggests that the very possibility of individuation has as its condition of possibility a particular mode of inhabiting the world as a bodily being.

Some of the earliest sustained treatments of the body in its cultural and social dimensions were Robert Hertz's (1909/1960) study of the symbolic pre-eminence of the right hand and Marcel Mauss' (1934/1950) justly famous article on techniques of the body. Mauss' work is almost universally cited as a precursor of the contemporary interest in the body. It anticipates the notions of practice and habitus later elaborated by Pierre Bourdieu (1977), though the notion of habitus had already passingly been used with a similar sense by Max Weber in his discussion on religions of ethical salvation. Important historically is the relation between this article on the body and Mauss' fragmentary but influential discussion of the person, appearing only four years later (1938/1950). In the latter article, he suggested that all humans have a sense of spiritual and corporal individuality, and saw the person as associated with the distinction between the world of thought and the material world as promulgated by Descartes and Spinoza.

2. *Anthropology of the Body*

In later decades, attention to the body largely took the form of studies of gesture, nonverbal communication, kinesics, and proxemics (Benthall and Polhemus 1975). In these studies, interest in the body in its own right was subordinated to interest in communication as a cultural process, with the body serving as the means or medium of communication. In other words, rather than beginning with a concern for bodiliness *per se*, these analyses took language as their model, using a linguistic analogy to study various types of languages of the body. Perhaps for this reason, they are rarely cited in the literature on the body and embodiment that has appeared in abundance since the late 1980s.

Two books by Mary Douglas define the threshold of a true anthropology of the body. In 'Purity and Danger' (1966), she made her famous argument that the dietary rules in the biblical book of Leviticus summarized the categories of Israelite culture, and made the general claim that social system controls induce consonance between social and physiological levels of experience. In 'Natural Symbols' (1973), Douglas made her equally famous argument that the bodily state of trance in different societies corresponds to the state of social organization characterized in terms of the 'grid' of cultural classifications and the 'group' control over the ego as an individual actor. Building on Mauss, she rejected both Edward Hall's work on nonverbal communication and Claude Levi-Strauss' work on mythical thought, on the grounds that they lacked hypotheses to account for variation across cultures. In presenting such a hypothesis, she committed herself to understanding the body as a 'medium of expression,' arguing that a drive to achieve consonance across levels of experience requires the use of the body to be 'coordinated with other media' so as to produce distinct bodily styles, and that 'The physical body can have universal meaning only as a system which responds to the social system, expressing it as a system' (1973, p. 112). Overall, Douglas' legacy to the anthropology of the body is encapsulated in her conception of 'two bodies,' the physical and the social, 'the self and society' (1973, p. 112).

Blacking's (1977) introduction to his edited volume on anthropology of the body offers a programmatic outline that is both more concerned with the body *per se* and its contribution to social processes than with the manner in which it reflects or expresses those processes, and which is more explicitly concerned with the relation between the biological and the cultural. Its 'chief concern is with the cultural processes and products that are externalizations and extensions of the body in varying contexts of social interaction,' and it rejects the distinction between biological and cultural anthropology on the grounds that culture has shaped the physical body, while features of culture such as language are biologically based (1977, p. 2). Blacking sketches four premises for the anthropology of the body: that society is a biological phenomenon, that all humans possess a common repertoire of somatic states, that nonverbal communication is fundamental, and that the mind cannot be separated from the body (1977, p. 18). Blacking emphasized the liberatory potential in anthropology of the body insofar as it could contribute to de-alienation and 'ownership of our senses.'

Once the body emerged from theoretical anonymity to become a recognized topic in its own right, its omnipresence in social life led to a multiplication of ways to organize its study. Following Douglas' inspiration to recognize 'two bodies,' Nancy Scheper-Hughes and Margaret Lock (1987) suggest we instead consider 'three bodies,' including the individual body, the social body, and the body politic. John O'Neill (1985) suggested thinking in terms of 'five bodies,' including the world's body (i.e., the anthropomorphized cosmos), the social body, the body politic, the consumer body, and the medical body.

3. The Body as a Theoretical Problem

It was not long between the emergence of the body as a topic and the transformation of the body into a problem. This aspect of the body's career in anthropology is deeply bound up with the career of anthropology in interdisciplinary studies. Scholars from virtually every branch of the human sciences have been influenced by reformulations of basic understandings of the body through the work of Michel Foucault on the discursive formations that have constituted the hospital (1973), the prison (1977), and sexuality (1978); the work of Pierre Bourdieu (1977) on practice and habitus; and the work of Maurice Merleau-Ponty (1962) on perception and embodiment. Equally, scholarly understandings of the body have been transformed by feminist works including Luce Irigaray's (1985) extended critique of psychoanalysis and Donna Haraway's (1991) studies of the encounter between gendered bodies and technology. Although it is a simplification to formularize the scope and nature of the theoretical sea change effected through such works, two aspects stand out. First, though work such as Blacking's called into question the distinction between mind and body, and between biological and cultural anthropology, subsequent work has gone even farther in calling into question the degree to which biological nature can be considered a stable substrate of human existence. Second, though Douglas encouraged us to think in terms of two bodies corresponding to self and society, subsequent work has gone beyond her tendency to treat the body-self as a passive lump of clay or tabula rasa upon which society imposes its codes, toward understanding it as a source of agency and intentionality, taking up and inhabiting the world through processes of intersubjective engagement. This radical shift has made the body a central problem, and one of some urgency, across a range of disciplines. Thus, Emily Martin, whose own work on gender, science, and technology (1987, 1994) has contributed significantly to the body's move to center stage in social theory, has posed the question of the 'end of the body' as we have known it.

In addition, there has been a chorus of statements about the historical malleability of the body. The radical 'situatedness' of the body extends even to the domain of biology itself, as is evident in recent feminist theory that eliminates 'passivity' as an intrinsic characteristic of the female body, and reworks the distinction between sex and gender (Haraway 1991, pp. 197–8), as well as the decoupling of female sexual pleasure with the act of conception. With biology no longer a monolithic objectivity, the body is transformed from object to agent (Haraway 1991, pp. 197–8).

The contemporary cultural transformation of the body can be conceived not only in terms of revising biological essentialism and collapsing conceptual dualities, but also in discerning an ambiguity in the boundaries of corporeality itself. Some have suggested that in contemporary civilization the human body can no longer be considered a bounded entity, due to the destabilizing impact of social processes of commodification, fragmentation, and the semiotic barrage of images of body parts. Others have explicitly problematized bodily boundaries between animal and human, between animal/human and machine, and between human and divine. Exploring these cultural boundaries can be incredibly rewarding and remarkably problematic given the circumstances of corporeal flux and bodily transformation sketched above.

4. Embodiment and Culture

This radical rethinking has also created an opportunity for a rethinking of culture and self from the standpoint of the body and embodiment. If the body was often a background feature in traditional ethnographies, so it has often remained implicit in anthropological theories of culture, which historically have been cast in terms such as symbols, meanings, knowledge, practices, customs, or traits. The problematizing of the body and its movement to center stage in social theory has also led to the emergence of studies that do not claim to be about the body *per se*, but instead suggest that culture and self can be understood from the standpoint of embodiment as an existential condition in which the body is the subjective source or intersubjective ground of experience.

This development can be seen taking place concretely over time in the work of a variety of anthropologists. For example, in the early 1970s, Strathern and Strathern produced a conventional monograph on body decoration in Mount Hagen, New Guinea. More recently, each author has developed their thought in light of a reformulated understanding of embodiment (A. Strathern 1996) and its gendered nature (M. Strathern 1992). Terence Turner (1980) published an influential essay on body decoration among the Brazilian Kayapo, and since then has moved from this conception of the 'social skin' to a more thorough development of the place of 'bodiliness' among the Kayapo and more generally in anthropological thinking (1995). James Fernandez, following his powerful treatment of metaphor and body symbolism in the Bwiti religion among the Fang in Gabon, later turned his attention to the place of the body in Bwiti explicitly with respect to developing theorizations of embodiment (Fernandez 1990). In work on ritual healing, ritual language, and the cultural constitution of self among Catholic Charismatics in contemporary North America, Thomas Csordas elaborated earlier arguments cast in terms of rhetoric with analyses made from the standpoint of embodiment (Csordas 1997). In her work on South Africa, Jean Comaroff (1985) refocused the traditional anthropological interest in body symbolism in the context of contemporary

approaches to the analysis of political economy and ritual healing, tracing these themes across the precolonial, colonial, and postcolonial periods. In two brief but compelling passages, Michelle Rosaldo (1984) did much to open the way for an understanding that collective symbols acquire 'power, tension, relevance, and sense' through embodiment (p. 141), and that emotions are 'embodied thoughts' (p. 143).

The theoretical crux of these new syntheses is a critique of tenacious conceptual dualities such as those between mind and body, subject and object, sex and gender, body and embodiment. These developments must be seen in the theoretical context of the prominence since the 1970s of notions such as the interpretive turn, the linguistic turn, the move to cultures defined as systems of symbols, and concepts such as textuality, discourse, representation, and semiotics broadly conceived. Understandings of the body itself in terms of the 'metaphor of the text' is evidenced in common scholarly phrases like 'the body as text,' 'the inscription of culture on the body,' or 'reading the body.' Recent work on embodiment offers the opportunity to balance what is thus, to some, an overemphasis on representation. When this tack is chosen, the key theoretical term that comes to take its place alongside representation is 'being-in-the-world.' The latter notion, drawn from phenomenology, does not supplant representation, but offers it a dialogical partner: in brief, semiotics offers the notion of textuality in order to understand representation, and phenomenology offers the notion of embodiment in order to understand being-in-the-world. 'The body' *per se*, can then be construed both as a source of representations and as a ground of being-in-the-world (Csordas 1994). Meanwhile, 'embodiment' becomes an avenue of approach to culture and self, just as textuality is an approach to the study of culture and self. Thus, to work in a 'paradigm of embodiment' is not to study anything new or different, but to address familiar topics—healing, emotion, gender, or power—from a different standpoint. There is not a special kind of data or a special method for eliciting such data, but a methodological attitude that demands attention to bodiliness even in purely verbal data such as written text or oral interview.

In summary, the body's career in anthropology has been on an upswing. From an early anonymity as a taken-for-granted background feature of social life, it has emerged in recent decades first as an explicit topic of anthropological research, then as a problem as its cultural and historical instability as a natural object became increasingly evident. More recently, embodiment has presented itself as an opportunity for reformulating earlier interpretations and rethinking fundamental concepts of culture and self. Through the 1990s the problem posed by the body of the relation between representation and being-in-the-world has been an increasingly prominent site at which anthropologists have become engaged in the wider interdisciplinary discourse of the human sciences. Relevant studies can be found under the rubrics not only of anthropology of the body, but also of medical anthropology and anthropologies of the senses, of space, of dance, of violence, and of science and technology (for a review see Lock 1993). On a wide variety of fronts, the body continues to advance.

See also: Body, History of; Body Image and Gender; Cultural Relativism, Anthropology of; Ecology, Cultural; Feminist Epistemology; Feminist Theory; Gender and Feminist Studies; Mind–Body Dualism; Psychological Anthropology; Reflexivity in Anthropology

Bibliography

Benthall J, Polhemus T (eds.) 1975 *The Body as a Medium of Expression*. E.P. Dutton, New York

Blacking J (ed.) 1977 *The Anthropology of the Body*. Academic Press, London

Bourdieu P 1977 *Outline of a Theory of Practice* [Nice R trans.]. Cambridge University Press, Cambridge, UK

Comaroff J 1985 *Body of Power, Spirit of Resistance: The Culture and History of an African People*. University of Chicago Press, Chicago

Csordas T J (ed.) 1994 *Embodiment and Experience: The Existential Ground of Culture and Self*. Cambridge University Press, Cambridge, UK

Csordas T J 1997 *Language, Charisma, and Creativity: The Ritual Life of a Religious Movement*. University of California Press, Berkeley, CA

Douglas M 1966 *Purity and Danger*. Routledge and Kegan Paul, London

Douglas M 1973 *Natural Symbols*. Vintage, New York

Fernandez J 1990 The body in Bwiti: Variations on a theme by Richard Werbner. *Journal of Religion in Africa* **20**: 92–111

Haraway D 1991 *Simians, Cyborgs, and Women: the Reinvention of Nature*. Routledge, New York

Hertz R [1909]1960 The preeminence of the right hand. *Death and the Right Hand* [Needham R, Needham C trans.]. Aberdeen University Press, Aberdeen, UK

Irigaray L 1985 *This Sex Which Is Not One* [trans. Porter C, Burke C]. Cornell University Press, Ithaca, IL

Leenhardt M 1947/1979 *Do Kamo: Person and Myth in a Melanesian World* [trans. Miller B Gulati]. University of Chicago Press, Chicago

Lock M 1993 Cultivating the body: Anthropology and epistemologies of bodily practice and knowledge. *Annual Review of Anthropology* **22**: 133–55

Lowie R H 1924 *Primitive Religion*. Grosset and Dunlap, New York

Martin E 1987 *The Woman in the Body: A Cultural Analysis of Reproduction*. Beacon Press, Boston

Martin E 1994 *Flexible Bodies: The Role of Immunity in American Culture from the Days of Polio to the Age of AIDS*. Beacon Press, Boston

Mauss M [1934]1950 *Les Techniques du Corps. Sociologie et Anthropologie*. Presses Universitaires de France, Paris

Merleau-Ponty M 1962 *Phenomenology of Perception* [trans. Edie J]. Northwestern University Press, Evanston, IL

O'Neill J 1985 *Five Bodies: The Shape of Modern Society*. Cornell University Press, Ithaca, IL

Radin P 1927 *Primitive Man as Philosopher*. Dover Publications, New York
Rosaldo M 1984 In: Shweder R, Levine R (eds.) *Toward an Anthropology of Self and Feeling. Culture Theory*. Cambridge University Press, Cambridge, UK, pp. 137–57
Scheper-Hughes N, Lock M 1987 The mindful body: A prolegomenon to future work in medical anthropology. *Medical Anthropology Quarterly* **1**: 6–41
Strathern A J 1996 *Body Thoughts*. University of Michigan Press, Ann Arbor, MI
Strathern M 1992 *Reproducing the Future: Essays on Anthropology, Kinship, and the New Reproductive Technologies*. Routledge, New York
Turner T 1995 Social body and embodied subject: Bodiliness, subjectivity, and sociality among the Kayapo. *Cultural Anthropology* **10**: 143–70

T. J. Csordas

Body, Evolution of

Major changes in body form have taken place during human evolution, related to the initial adoption and increasing commitment to bipedal locomotion, and continuing adaptations to climatic, dietary, and other environmental factors. Although all modern humans share a similar body plan, there is still substantial variability in body size and body shape within and between populations that is in part a reflection of environmental adaptation. This article reviews the evidence for both the evolution and modern distribution of body form in humans, and its behavioral and ecological significance.

1. *The Bipedal Transformation*

It is widely agreed that the transition from quadrupedal to bipedal locomotion was one of the most important innovations in human evolution; indeed, bipedality defines our lineage better than any other single characteristic (Clark 1964, Aiello and Dean 1990, Fleagle 1999). Efficient bipedal locomotion requires changes in a number of anatomical characteristics, including lengthening of the lower limb, restructuring of the pelvis, and other alterations. Some of these characteristics are apparent as early as 3–4 million years ago (McHenry 1991, Leakey et al. 1995). However, it is also becoming increasingly clear that the transition to fully terrestrial bipedality occurred in a number of stages, and that some anatomical adaptations to arboreality (tree climbing), especially in the forelimb, persisted for millions of years (McHenry 1991, White et al. 1994, Clarke and Tobias 1995, Leakey et al. 1995, Leakey et al. 1998, McHenry and Berger 1998, Asfaw et al. 1999). It is not until about 1.5 million years ago, with *Homo erectus* (or *ergaster*), that fully modern human body proportions, and by implication, fully modern bipedalism, are evident in the fossil record (Walker and Leakey 1993).

Thus, the adoption of terrestrial bipedality was gradual, and probably involved increasing reliance on this form of locomotion while still retaining the ability to use trees, e.g., for escape from predators. There are many theories on why and how human bipedalism evolved, including ecological and dietary shifts, and changes in social organization (for a recent review, see Fleagle 1999). While distinctively human proportions of the body, such as forelimb to hindlimb length, were established by 1.5 million years ago, within this basic body plan variations in other aspects of body size and shape are apparent throughout the more recent fossil record and among living humans.

2. *Environmental Adaptation*

2.1 *Ecogeographical 'Rules' and their Application to Modern Humans*

One aspect of the environment that has probably had a profound effect on human body size and shape is climate. Systematic relationships between climate and body form among homeothermic animals have been recognized for more than a century, and have been codified as Bergmann's and Allen's ecogeographical 'Rules' (Mayr 1963). Bergmann's Rule states that within a species or closely related group of species, populations living in colder climates will be larger in body mass than those in warmer climates, while Allen's Rule states that colder climates will be associated with shorter extremities and warmer climates with longer extremities. Both of these observations are actually consequences of a more general relationship between body mass and body surface area, in which both a larger body mass and relatively shorter extremities decrease the ratio of surface area to body mass, thus conserving heat, and vice versa. The same principle can also be applied to the overall shape of the human trunk: it can be shown that an absolutely wide trunk (and thus wide body) will lead to a decreased surface area/body mass ratio, regardless of body height, while an absolutely narrow trunk will lead to an increase in the ratio, again regardless of variation in height (Ruff 1991). This explains why both tall and short populations inhabit tropical Africa; both maintain an absolutely narrow body and thus a high surface area/body mass ratio.

Ecogeographical clines that conform to physiological (climatic) expectations have been demonstrated for many aspects of living human body form, including body mass, body breadth, relative limb length, and direct estimates of surface area to body mass (Roberts

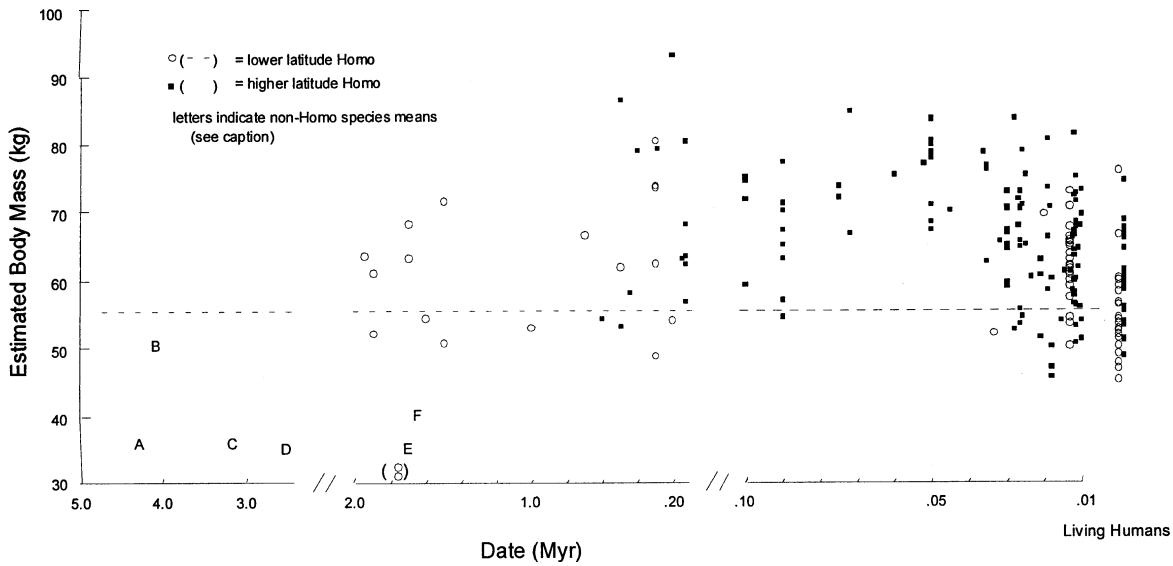

Figure 1
Changes in estimated body mass (weight) during human evolution. Living humans are sex-sample-specific population means (Ruff 1994), other *Homo* data points are individual fossils (see text for explanation of two data points in parentheses), letters are non-*Homo* sample means: A: *Ardipithecus ramidus*, B: *Australopithecus anamensis*, C: *Australopithecus afarensis*, D: *Australopithecus africanus*, E: *Australopithecus* (*Paranthropus*) *robustus*, F: *Australopithecus* (*Paranthropus*) *boisei*. Estimates for *Homo* from Ruff et al. 1997 (plus two additional data points), for C-F from McHenry 1992, for B from Leakey et al. 1995, and for A from White et al. 1995 (my estimate based on similarity in humeral head size with middle of range for *A. afarensis*). Note changes in temporal scale (millions of years). (Modified from Ruff et al. 1997.)

1978, Ruff 1994). Other environmental factors, such as diet, obviously also can affect body form, but do not adequately account for these general trends (Ruff 1994). It is more difficult to evaluate such morphological variability in our fossil ancestors, but recent paleoanthropological discoveries combined with developments in methodology have provided some new insights.

2.2 Methods for Reconstructing Body Form from Skeletal Remains

Body mass (weight) for fossil specimens has been estimated using various parts of the skeleton, including craniodental dimensions, but it is clear that postcranial dimensions provide the least biased and potentially most accurate means for reconstructing body size, especially in an evolutionary lineage where changes in relative tooth and cranial size have been dramatic (Pilbeam and Gould 1974). Two basic types of methods have been employed: a mechanical approach, which uses features that are related to the mechanical load-bearing function of the skeleton, such as lower limb articular size, and a morphometric approach, in which features that actually reflect the size and shape of the postcranial skeleton are used. For the latter, it has been shown that stature, estimated from limb bone length, and body breadth, measured across the pelvis, together provide a good estimate of body mass (Ruff 1994). Aspects of body shape in fossil humans can be evaluated through intra and inter-limb bone length proportions (Trinkaus 1981), proportions of limb bone lengths to vertebral column (trunk) length (Holliday 1997), proportions of body breadth to body height, and reconstructed body mass to body height (Ruff 1994).

2.3 Temporal Changes in Body Size

Changes in body mass over 4.5 million years of human evolution are shown in Fig. 1. Because of large temporal differences in available data density, three different temporal scales are used. Also, because it is known that body mass varies ecogeographically in modern populations (see above), the data for the genus *Homo* are plotted separately for higher (30°) and lower latitudes (all non-*Homo* fossils are from lower latitudes). Body masses were estimated from both

lower limb articular size and using the stature/pelvic breadth 'morphometric' approach; references are given in the figure caption.

Several trends are apparent in Fig. 1. First, body size was relatively small in the earliest period of human evolution, and increased with the appearance of *Homo* at least two million years ago. This may have been associated with significant changes in ecology in *Homo*, i.e., increased diurnal travel through open country, possibly indicating an increased reliance on hunting (or at least increased foraging distance) (Wheeler 1992, McHenry 1994). The two low outlying *Homo* data points at about 1.77 million years ago are specimens belonging to a species (*Homo habilis sensu stricto*) that shows some distinctly non-*Homo*-like traits and may be better referred to another taxon (Wood 1999).

There was a further increase in average body size at about half a million years ago, coinciding with the first (postcranial) evidence for *Homo* from higher latitudes. As noted above, colder temperatures favor larger body size, so this may very well be an ecogeographical phenomenon. Finally, body size decreased beginning about 50,000 years ago, reaching modern values by 10,000 years ago in higher latitude samples, although lower latitude samples from that time period are still larger than living populations.

Overall, body mass in *Homo* prior to 10,000 years ago was about 10% larger than in living humans inhabiting the same areas today (Ruff et al. 1997). Larger body size may be related in part to the hunting of large game (Thieme 1997); a decline in body size over the last few tens of thousands of years may reflect increasing technological sophistication and decreasing reliance on brute strength for performing subsistence tasks (Klein 1989). There is evidence that body size continued to decline, at least in some regions, through the Mesolithic and Neolithic (Frayer 1984), perhaps because of decreased nutritional quality and/or increased disease transmission associated with the agricultural revolution (Cohen and Armelagos 1984). Body size has increased over the past century in many higher latitude populations, probably due to improvements in the environment (Eveleth and Tanner 1990), while the opposite trend has been observed in some lower latitude (Third World) populations (Tobias 1985).

2.4 Temporal Changes in Body Shape

As noted earlier, vestiges of arboreality, particularly a relatively longer and/or more robust forelimb, remained in the skeleton long after the first evidence for bipedality. In the few specimens where we can evaluate it, the trunk was relatively wide in these small-bodied early fossil taxa (Ruff 1991). About 1.5–1.7 million years ago, stature suddenly increased in *Homo* (along with body mass) and the trunk became relatively narrower as well. This coincided with a major ecological change in East Africa (the location of the first known *Homo* fossils) at this time, in which the climate became much more arid and the landscape more open (Potts 1998), favoring body linearity (Ruff 1991, Wheeler 1993). Alternatively or in conjunction, a dietary shift in *Homo* towards more meat eating may have required a less bulky abdomen and thus permitted a relatively narrower body (Aiello and Wheeler 1995).

Where they can be evaluated, the body proportions of later human ancestors appear to mirror the same ecogeographic clines found among living populations. So, for example, Western European Neandertals, who lived in a very cold climate for many millennia, were quite stocky and short-limbed, while their replacements, Upper Paleolithic-associated 'anatomically modern humans,' who very likely migrated in from farther south, were much more linearly built (Trinkaus 1981, Ruff 1994, Holliday 1997). Wide bodies and/or short limbs are not limited to Neandertals but rather are a general phenomenon of higher latitudes (colder climates), as evidenced by other recently described specimens from Great Britain, Spain, and northern China dating from 200,000–500,000 years ago (Arsuaga et al. 1999, Rosenberg et al. 1999, Trinkaus et al. 1999). In fact, it can be argued that pre-modern humans from higher and lower latitudes exhibit 'hyper-arctic' and 'hyper-tropical' body proportions, respectively, probably because of less efficient cultural buffering against the environment (Ruff 1994, Trinkaus et al. 1999).

3. Body Size as a Baseline for Other Comparisons

In addition to its ecological and behavioral significance, body size is important in human evolutionary studies because it is often used as a 'denominator' against which to evaluate other physical traits. One prime example is brain size, which is commonly expressed as a ratio against body mass, or encephalization quotient (EQ) (e.g., Pilbeam and Gould 1974). EQ increased in early *Homo* (1.5–2.0 million years ago) from previous estimated values, but then remained constant for at least a million years (to 0.5 million years ago), after which it increased again, exponentially, to modern values (McHenry 1994, Ruff et al. 1997). Skeletal robusticity has also been evaluated relative to body size, and shown to follow an inverse trend relative to brain size (Ruff et al. 1993).

Finally, the body size of earlier humans can be used as a baseline for interpreting the health of modern populations. Comparing present-day measures of stature and body mass with those of ancestral populations may give better estimates of potential body size limits and provide guidance in setting nutritional and other health standards (WHO 1995).

See also: Adaptation, Fitness, and Evolution; Body, History of; Brain, Evolution of; Darwin, Charles Robert (1809–82); Evolution, History of; Evolution: Optimization; Human Behavioral Ecology; Human Cognition, Evolution of

Bibliography

Aiello L, Dean C 1990 *Human Evolutionary Anatomy*. Academic Press, London

Aiello L C, Wheeler P 1995 The expensive-tissue hypothesis. *Current Anthropology* **36**: 199–211

Arsuaga J-L, Lorenzo C, Carretero J-M, Gracia A, Martinez I, Garcia N, Bermudez de Castro J-M, Carbonell E 1999 A complete human pelvis from the Middle Pleistocene of Spain. *Nature* **399**: 255–8

Asfaw B, White T, Lovejoy O, Latimer B, Simpson S, Suwa G 1999 *Australopithecus garhi*: A new species of early hominid from Ethiopia. *Nature* **284**: 629–35

Clark W E L 1964 *The Fossil Evidence for Human Evolution*. University of Chicago Press, Chicago

Clarke R J, Tobias P V 1995 Sterkfontein member 2 foot bones of the oldest South African hominid. *Science* **269**: 521–4

Cohen M N, Armelagos G J 1984 *Paleopathology at the Origins of Agriculture*. Academic Press, New York

Eveleth P B, Tanner J M 1990 *Worldwide Variation in Human Growth*. Cambridge University Press, Cambridge, UK

Fleagle J G 1999 *Primate Adaptation and Evolution*. Academic Press, New York

Frayer D W 1984 Biological and cultural change in the European Late Pleistocene and Early Holocene. In: Smith F H, Spencer F (eds.) *The Origins of Modern Humans: A World Survey of the Fossil Evidence*, Wiley-Liss, New York

Holliday T W 1997 Body proportions in Late Pleistocene Europe and modern human origins. *Journal of Human Evolution* **32**: 423–47

Klein 1989 *The Human Career*. University of Chicago Press, Chicago

Leakey H G, Feibel C S, McDougall I, Walker A 1995 New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. *Nature* **376**: 565–71

Leakey M G, Feibel C S, McDougall I, Ward C, Walker A 1998 New specimens and confirmation of an early age for *Australopithecus anamensis*. *Nature* **393**: 62–6

Mayr E 1963 *Animal Species and Evolution*. Harvard University Press, Cambridge, MA

McHenry H M 1991 First steps? Analyses of the postcranium of early hominids. In: Coppens Y, Senut B (eds.) *Origine(s) de la bipédie chez les hominidés*. Centre National de la Recherche Scientifique, Paris

McHenry H M 1992 Body size and proportions in early hominids. *American Journal of Physical Anthropology* **87**: 407–31

McHenry H M 1994 Behavioral ecological implications of early hominid body size. *Journal of Human Evolution* **27**: 77–87

McHenry H M, Berger L R 1998 Body proportions in *Australopithecus afarensis* and *africanus* and the origin of the genus *Homo*. *Journal of Human Evolution* **35**: 1–22

Pilbeam D, Gould S J 1974 Size and scaling in human evolution. *Science* **186**: 892–901

Potts R 1998 Environmental hypotheses of hominid evolution. *Yearbook of Physical Anthropology* **41**: 93–136

Roberts D F 1978 *Climate and Human Variability*, 2nd edn. Cummings, Menlo Park, CA

Rosenberg K R, Lu Z, Ruff C B 1999 Body size, body proportions and encephalization in the Jinniushan specimen. *American Journal of Physical Anthropology* Suppl. **28**: 235

Ruff C B 1991 Climate, body size and body shape in hominid evolution. *Journal of Human Evolution* **21**: 81–105

Ruff C B 1994 Morphological adaptation to climate in modern and fossil hominids. *Yearbook of Physical Anthropology* **37**: 65–107

Ruff C B, Trinkaus E, Holliday T W 1997 Body mass and encephalization in Pleistocene *Homo*. *Nature* **387**: 173–6

Ruff C B, Trinkaus E, Walker A, Larsen C S 1993 Postcranial robusticity in *Homo*, I: Temporal trends and mechanical interpretation. *American Journal of Physical Anthropology* **91**: 21–53

Thieme H 1997 Lower Palaeolithic hunting spears from Germany. *Nature* **385**: 807–10

Tobias P V 1985 The negative secular trend. *Journal of Human Evolution* **14**: 347–56

Trinkaus E 1981 Neanderthal limb proportions and cold adaptation. In: Stringer C B (ed.) *Aspects of Human Evolution*, Taylor and Francis, London

Trinkaus E, Stringer C B, Ruff C B, Hennessy R J, Roberts M B, Parfitt S A 1999 Diaphyseal cross-sectional geometry of the Boxgrove 1 Middle Pleistocene human tibia. *Journal of Human Evolution* **37**: 1–25

Walker A, Leakey R (eds.) 1993 *The Nariokotome Homo Erectus Skeleton*. Harvard University Press, Cambridge, MA

Wheeler P E 1992 The thermoregulatory advantages of large body size for hominids foraging in savannah environments. *Journal of Human Evolution* **23**: 351–62

Wheeler P E 1993 The influence of stature and body form on hominid energy and water budgets; a comparison of *Australopithecus* and early *Homo* physiques. *Journal of Human Evolution* **24**: 13–28

White T D, Suwa G, Asfaw B 1994 *Australopithecus ramidus*, a new species of early hominid from Aramis, Ethiopia. *Nature* **371**: 306–12

WHO 1995 *Physical Status: The Use and Interpretation of Anthropometry*. World Health Organization, Geneva

Wood B A 1999 The human genus. *Science* **284**: 65–71

C. Ruff

Body, History of

1. The Problem of Corporeality and the Two Approaches to the History of the Body

In a text fragment entitled 'Das Interesse am Körper', (the interest for the body), Max Horkheimer and Theodor W. Adorno argued that the 'love–hate relationship to the body' and the refusal of corporeality were the hidden references of the modern society after the Enlightenment. This oblivion of the body was related to the long-lived dichotomy between body and soul, i.e., with the Cartesian opposition of *res extensa* and *res cogitans*, which met with such success for many centuries (Danto 1999). The consequence was the naturalization of the body and its exclusion from social processes. This 'modern

body,' an idea which has been emerging since the eighteenth century and was supposed to be a 'stable experience' (Armstrong 1983), did not follow historical evolution: it could even be defined as the antithesis of cultural change. With the progress of secularization, mankind learned to consider society, its institutions and modes of representation as historical realities which man himself had called into existence. But the human body was not man made, and as such it could be of no interest as an object of historical interpretation. Even in the nineteenth and twentieth centuries, the body still remained a bastion of nature that opposed a limit to the general historicizing of thought. Since the middle of the nineteenth century, the history of the body had become the stronghold of evolutionary biology, while a number of other disciplines (anatomy, physiology, biomedicine, physical anthropology, occupational and nutritional science, hygiene, molecular biology, genetics, etc.) individually began to research into this seemingly immutable physical entity, using exact methods and laboratory experiments. Optimization of physical fitness, rationalization, normalization, healing, or maintaining in good health were the main objectives of their efforts. The fact that the body was thought to be independent of culture and that, at the same time, the modern age was literally obsessed by the impact of carnal corporeality on both individuals and society, explains the existence of a 'love–hate relationship' towards the body diagnosed by the Critical Theory of Adorno and Horkheimer.

Given this background, the history of the body proves to be a sensitive subject. However, it does not deal exclusively with the presence of the body in history. At this point, it is necessary to define two different conceptions:

(a) It must be stressed that the body was never absent from historiography. 'Either "the body" is not really a subject of its own, or it contains nearly all other subjects' (C.W. Bynum 1996). Historical works are full of phenomena which would be impossible to describe without frequent references to the human body. Demographers work with vital statistics to produce indicators on birth rate, morbidity, or mortality; some of these studies reduce the drama of life to quantitative relationships, other authors (such as A.E. Imhof (1992)) combined empirical findings on the extension of the average life expectance and the tendency towards a decline in the number of children with historico-cultural questions, from which was derived that the experience of the body and the meaning of life were omnipresent. Neither the history of political philosophy, nor the history of societies and mentalities as it was suggested by the School of the Annales in France from the 1920s onwards would make any sense without the presence of the body. As early as 1938, Lucien Febvre suggested writing a history of sensitivities and of the changing systems of emotions in order to understand human agency.

(b) It is clear that a continuous, conscious, explicit reflection on the role of the body only began in the last three decades of the twentieth century in social sciences and humanities. Once a static, biological, and naturalistic idea of the body started being criticized, body-related questions developed as a trend among historians and constituted a new field of research. As such, it is sometimes seen as an interdisciplinary meeting-point of several different approaches (discourse analysis, semiotics, gender studies, cultural studies, science, and technology studies, etc.); others consider it to be an independent sub-discipline of history.

These two conceptions of body history give rise to frequent misunderstandings. While one party insists on the innovative potential that body history has proved to offer since the 1970s, the other party dismisses it as a fashion trend that conceals the evidence that history, as a science focusing on the cultural interaction of human beings, has always been body history. From a historic–scientific perspective, however, there exist a number of arguments in favor of highlighting the shift of paradigms that took place in the 1970s, and asking how this new, theoretically-based examination of the body has influenced older fields of research like demography and social history.

2. *Continuities and Shift of Paradigms: Four Periods*

The historicizing of the body can be split up into *four phases* (which are not clearly distinct, but overlap frequently with one another).

In the first phase, which lasted until the early 1970s, a number of innovative studies looked at the human body, and the imagery that had crystallized around it, as objects of theoretically elaborated analyses. Later, they became 'classics' in the domain, although their initial intention had nothing to do with the new speciality named 'history of the body,' which began to emerge in the 1970s. The historical studies concentrated principally on the Middle Ages and the early modern era, whereas anthropology, philosophy, and psychoanalysis dared to delve into contemporary phenomena. Some of the most significant works of this period are Sigmund Freud's *Civilization and its Discontents* (1929/30), whose central theme is the extension of the body with the help of technical artifacts and the promotion of the human being to a god of prostheses, Mauss (1935), in which the body is analyzed as the first and most natural instrument at man's disposal; Paul Valéry's *problem of the three bodies* (1943), which looks into the various, parallel ways of speaking of the body; Maurice Merleau-Ponty (1945) and Samuel Todes (1963), which deals with the body as the condition of awareness, in compliance with a philosophy of finiteness; and Kantorowicz (1957), whose intention is to contribute to the political theology of the Middle Ages and looks at the mystical fiction of an immortal *political body* existing in parallel to the mortal *natural body* of the king.

A second phase began in the 1960s—accompanying the removal of the body's taboos in the

alternative culture of the time—giving rise to many publications and a new theoretical thinking in the 1970s. The principal references of this phase are the works of Michel Foucault (1977) (in particular *The Birth of the Clinic* and *Discipline and Punishment: The Birth of the Prison*), as well as Norbert Elias' (1969) *magnum opus*, *The Process of Civilization*, which was published in the late 1930s, but met with significant success only three decades later. Another important contribution was Mary Douglas' *The Two Bodies* (1973), in which the author examines the metaphorical relationships and cognitive interactions between the body as a social construction and the individual physical body. The new studies of the human body were based on the thesis of a constant interaction between knowledge and power. Foucault's concepts of a bio-power and of a political economy of the body were set against a liberal-Marxist scenario of disciplinary force which represses the body. Contrary to the assumption of a disciplinary society (in which the subject is oppressed by power), in Foucault's view the body is produced by knowledge-based power, which shapes everyday practices and generates regimes of visibility and discourses. In the concept of a 'normalizing society,' the body is the medium which binds together the macro-organization of power and the micro-practices through which social structures are stabilized. This perspective is directly connected to Norbert Elias' concept of the individual internalization of social constraints (force to force oneself?). The production of a modern human being can be defined in terms of an ever-growing self-control of the human being, mainly through the decoupling of emotion and bodily expression. The 1970s also saw the publication of interesting studies in various domains of social sciences notably by Helmuth Plessner (philosophical anthropology of the senses), Heinrich Schipperges (outlines for a philosophy of the body), Dietmar Kamper and Christoph Wulf (return of the body and fading of the senses), Rudolf zur Lipe (On the Own Body. About the Economy of Life) and Utz Jeggle (preliminary reflection on a popular knowledge of corporeality).

The third phase was strongly influenced by the feminist criticism of science, which admitted gender as a category for analysis and introduced new standards into body history. Thanks to Yvonne Verdier, Carolyn Bynum, Gianna Pomata, Ludmilla Jordanowa, Marie-Christine Pouchelle, and many others, new interpretations of the body appeared at the end of the 1970s, emphasizing its relation to gender history and clearing the ground for a far-reaching methodologically highly interesting discussion. In Germany, these endeavors are connected above all to the name of Barbara Duden, who published an innovative study, *The Woman Beneath the Skin: A Doctor's Patients in Eighteenth Century Germany* (Duden 1987). In this book about the interior, invisible corporealness of women in Eisenach (Germany) and their doctor Johann Storch, the historian and sociologist Barbara Duden examines the sociogenesis of the modern body in relation to a history of the lived body's receptivity to experiences. She analyses the loss of control over the body as a gate towards physical experience, and speaks in favor of a hermeneutic interpretation of the body's reality in the past. Although we cannot use our body, which we have possessed since the eighteenth century, as a bridge to interpret the experiences of people before the Enlightenment, the corporeality of every human existence gives us access to the context in which women and men experienced feelings in the past.

The fourth phase is characterized by a turn towards discursive, semiotic, and performative approaches and marks the beginning of a new era in the history of the body, which also lead to a post-structuralist reassessment of key authors of the 1960s, especially of Foucault. The 1990s saw the upcoming of new authors, in particular, Laqueur (1990) and Butler (1993) fulfilled the *linguistic turn* in body history and provoked the supporters of the traditional, experience-based study of corporeality. Laqueur examines the constitution of the body through discourse and action; from this point of view, sex is situational; it is explicable only within the context of battles over gender and power. In contrast, Butler—who, after her book was published, was wrongly considered to be an advocate of disembodiment—criticized the subject-centered constructionist approach by relating the materiality of the body to the performativity of gender: What I would propose in place of these conceptions of construction is a return to the notion of matter […] as a process of materialization that stabilizes over time to produce the effect of boundary, fixity, and surface we call matter?

In the meantime, several positions have been clarified. In the last few years, there has been a growing tendency to criticize the dissolving of the body into language (Bynum). However, this does not herald a return to an unreflected essentialism of corporeality. In contrast, analytical efforts converge in the concrete, mortal human body, without forgetting that the corporal existence of human beings will never be independent of the empire of signs, which is permanently created and transformed by cultural change.

3. *Fields of Research and Analysis*

For a long time, researchers in the field of body history were not sure whether they were looking into a completely new domain, or whether they had simply discovered a new and interesting perspective which offered the possibility of an innovative approach towards well-known social phenomena. The question becomes more complex: from the very beginning, body history was the fruit of interdisciplinary efforts. This can have two consequences: the loss of clarity due to friction between the disciplines, and the positive effects of synergy (cf. among others Feher et al. 1984). Authors who had already attempted a systematic approach to body history in the mid 80s presented an overview of the broad portion of

research based on social sciences. From a phenomenological point of view, Turner (1984) describes four tasks with which any social organization confronts bodies (and vice versa): (a) the reproduction of populations in time, (b) the regulation of bodies in space, (c) the constraints of the interior body through control and discipline, (d) the representation of the exterior body in the symbolic space of the society. O'Neill (1985) provides and alternative, more institution-based conceptualization, distinguishing (a) the world's body, (b) social bodies, (c) the political body, (d) consumer bodies, and (e) medical bodies.

In the narrower field of historiography, although there have been frequent variations in the main fields of interest, one can observe an obvious concentration of research on the Middle Ages and the Early Modern Era. Starvation and fasting, posture, body language, clothing habits, disciplining of the human body through the art of dancing, training of soldiers, seduction, sexuality, pregnancy, and childbearing: these and other themes were examined in the light of cultural anthropology and iconography and resulted in a new view of everyday life and bodily experience in past centuries. The 'education of the senses' (Peter Gay), the history of 'sensory awareness' (Rudolf zur Lippe), and an 'anthropology of the senses' (Lucien Febvre) turned out to be especially interesting fields of research. The conflict about the hierarchy of the senses, which began in antiquity, the theses of a mythopoetical constitution of the individual (Hartmut Böhme) and of a privilege of the eye at the end of the twentieth century (Paul Virilio) remain of symptomatic value. The history of sensory experience is closely linked with the history of wisdom and the remembering of the body, as a materialized matrix for inscriptions of cultural experience, the body is a part of the memory of a society. Step by step, performativity relates the signification process with the materiality of the body. It is the reiteration of practices and the regulation of discourses which generate the feeling of continuity and reliability of the real world.

Another area that offers many possibilities for research is the history of the perception of the body and of body-related imagery and esthetics (Elkin 1999). In the course of secularization, the image of the body as 'clockwork,' the perfect functioning of which was the proof of God's creation, was replaced by the idea of a machine working according to the laws of nature (with a male connotation: the man is considered to be the standard human being). Along with this evolution came the growing conviction that the body is a closed entity, as the medieval image of a 'grotesque', open body 'Michail Bakhtin' was gradually replaced by the modern, closed skin-self (Anzieu 1992). This evidence depended on new schemes of visibility: Barbara Stafford (1993) describes this historical shift toward visualization with the help of somatic metaphors and she points out the important role played by visual arts when it came to solving the representational problem of the embodiment or personification of knowledge. Crucial for this development was the eighteenth century, which saw the breakthrough of a *possessive individualism* in the perception of the body. This resulted in a degradation of the notion of a self extended into a unique and inviolable corporeal volume, to one in which the self only loosely possessed a body (Karl Figlio). The problem for the historical investigation into this body, which we possess and which is analyzed by science, is not how something that is obvious today had remained hidden for so long, but how the body had become so evident in the first place (Armstrong 1983). Thanks to the modern perceptions of the body, new ways of seeing collective bodies were also brought to the fore. Jacob Rogozinski showed that the incarnation of the individual in an all embracing global body is an old phenomenon, but in the nineteenth and twentieth centuries, the political hypostatization of this collective body became still stronger. While society depended more and more on technology, strengthening the thesis that social cohesion results from the linking-up of the elements of society by a powerful communicative, nervous system, nationalism encouraged the phantasm of a body of the people and of a blood community. The imagery of the body of the nation (Svenja Goltermann) e.g., proved to be mediated by the practice of male gymnastics, which aims at the appropriation of national order through individuals and social groups.

Medicine, dietetics inspired from humoral theory, and the modern conception of hygiene which emerged in the middle of the nineteenth century, also offer fertile ground for research in body history. Traditional, epidemiological approaches of research on disease are completed by analyses of illness, suffering, pain, and death of human beings that were increasingly considered as patients in the course of the nineteenth century, and thus were submitted to medical experts' power of definition. A similar process of professionalization and scientization took place in the areas of occupational and nutritional science. The interaction of physiology and the industrial society show, for instance, how epistemic regimes, technological artifacts, and social interests are intertwined in the production of knowledge, and how the thus-produced knowledge facilitates the appearance of individuals who consider their bodies as instruments of self-perfection (Sarasin 2001). In this context, the modification of individuals' self-images under the influence of scientific knowledge is studied in parallel with the unwillingness of the body faced with the exigencies of a working society, which is subject to the paradigms of efficiency and rationalization. In the leisure society of the post-war period, in which the body functions as a site of consumption, the questions arising from an invariably incomplete integration of the individual body into the social structure remain to be answered by further historical research.

4. *Controversies and Prospects*

Generally, in social sciences and humanities, experience-based concepts of the body implying a lived

corporeality are being challenged by poststructuralist–deconstructivist models that consider the body primarily as a discoursive effect or a product of the knowledge–power complex accumulated since the Enlightenment. Nonetheless, today, the consensus is that it must be essential to overcome this opposition of discourse and experience (Canning 1994). These debates frequently overlook the fact that the history of the body is in the midst of the fundamental discussion on the blurring of the boundaries between nature and culture. The ontological status of the traditional categories natural and social has become a critical question and this makes it obvious to consider the products of technology—and therefore the world we live in—as hybrid artifacts. From this point of view, the human body, too, would be a hybrid, i.e., a crossbreed in which social, technological, cultural and natural components cannot be held apart with certainty.

The materiality of the body plays a crucial role in this debate. This is obvious in disciplines such as biomedicine, cognitive sciences (CS), and artificial intelligence (AI), which lie in a paradoxical area of tension between hyperembodiment and disembodiment. Whilst scientists in the domain of AI start out from the assumption that intelligence can be modeled only on a material, body-bound base, which implies that machines that have emerging properties and therefore the ability of learning must by definition be techno-bodies (Lakoff and Johnson 1999), other authors (e.g., Flusser 1994) follow the principle from the subject to the project in order to develop visions according to which the conception of bodies has become a creative activity of mankind, escaping the blind contingencies of nature. It is essential to realize that the concept of information technology, genetic engineering, and xenotransplantation have made natural boundaries permeable and aims at putting mankind in a position to gain control over the evolutionary process of life which started billions of years ago.

It is precisely because of these perspectives, which are at once technocratic and playful, that the findings of body history have gained a major importance today. Although they are for various reasons not comparable, the concepts of the body suggested by biomedicine, information technology, and genetic engineering are in compliance with the more modern approaches of social sciences and humanities according to which the body is not an invariable, unchangeable entity, independent of time and history. When we are willing to reassess the body as a product of society, we have to think about how it is culturally constructed, or constituted. If historiography demonstrates that bodies were always modeled by culture, this means that the scientific and technological interventions in the physiology and (for the time to come) in the genome of the human being are nothing new in principle, they are simply an intensification of a traditional way in using bodies. In this sense, Donna Haraway pleads in favor of the curious and innovative application of the new possibilities offered by technology, and sees the creation of natural–cultural hybrids as a promising path towards the creative self-realization of mankind. This 'Cyborgs-optimism' is contradicted by another: distinguishing herself from technocratic feasibility fantasies, Barbara Duden supposes that although human beings are partially formed by culture, the somatic fundaments of life are an inescapable condition of their existence. Based on the paradigm of a 'historical somatology,' the examination of the time-bound forms of experiencing corporeality concentrates on the relation of human beings to their body, which changes in the course of history, as a *sine qua non* of experience. Such an approach considers the current colonization of the body by science and technology as an unprecedent process and an attack on the human condition itself. No doubt that Duden remains skeptic against the mythopoesis of a technology which models the desires of individuals in a way that they anticipate the possibilities of science and technology.

These divergent positions are the consequence of two conflicting evaluations of what must be the most fundamental challenges of mankind. In this context, body history becomes—whether it intends to or not—a moral authority and a source of criteria for the legitimacy of technological interventions on the human body. This clearly implies a heavy responsibility, but at the same time, it is a proof of the persistant relevance of the history of the body.

See also: Body: Anthropological Aspects; Body, Evolution of; Body Image and Gender; Cultural Relativism, Anthropology of; Ecology, Cultural; Emotions, History of; Feminist Theory; Gender and Feminist Studies; Gender History; Medicine, History of; Mind—Body Dualism; Psychohistory; Psychological Anthropology; Reflexivity in Anthropology

Bibliography

Anzieu D 1992 *Das Haut-Ich*. Suhrkamp, Frankfurt/M

Armstrong D 1983 *Political Anatomy of the Body: Medical Knowledge in Britain in the Twentieth Century*. Cambridge University Press, Cambridge

Bielefelder Graduiertenkolleg Sozialgeschichte (ed.) 1999 *Körper Macht Geschichte. Geschichte Macht Körper*. Verlag für Regionalgeschichte, Bielefeld

Butler J 1993 *Bodies that Matter. On the Discursive Limits of Sex?* Routledge, New York/London

Bynum C W 1996 Warum das ganze Theater mit dem Körper? Die Sicht einer Mediävistin. In: *Historische Anthropologie*. (6. Jg. No. 1), S 1–33

Canning K 1994 Feminist history after the linguistic turn: historicizing discourse and experience. *Signs* **19**: 368–404

Danto A C 1999 *The Body/Body Problem. Selected Essays*. University of California Press, Berkeley

Douglas M 1973 *Natural Symbols. Exploration in Cosmology*. Barrie & Jenkins, London

Duden B 1987 *Geschichte unter der Haut. Ein Eisenacher Arzt und seine Patientinnen um 1730*. Klett-Cotta, Stuttgart

Elias N 1969 *Über den Prozess der Zivilisation: soziogenetische und psychogenetische Untersuchungen*. Bern, Francke

Elkins J 1999 *Pictures of the Body Pain and Metamorphosis*. Stanford University Press, Stanford

Febvre L 1953 *La sensibilité et l'histoire*. In: Do., Combats pour l'histoire, Colin, Paris, pp. 221–38
Feher Michel (ed.) 1984 *Fragments for a History of the Human Body*, Vol. 3, Urzone, New York
Flusser V 1994 *Vom Subjekt zum Projekt: Menschwerdung*. Bollmann, Bensheim/Dusseldorf
Foucault M 1977 *Discipline and Punishment: The Birth of the Prison*. Penguin Books, London (1979-edition)
Foucault M 1978 *The History of Sexuality*. Penguin Books, New York (1979-edition)
Imhof A E (ed.) 1992 *Leben wir zu lange? Die Zunahme unserer Lebensspanne seit 300 Jahren - und die Folgen Köln [etc.]*. Böhlau
Jones C, Porter R 1994 *Reassessing Foucault. Power, Medicine and the Body*. Routledge, London/New York
Kamper D, Wulf C (eds.) 1982 *Die Wiederkehr des Körpers*. Suhrkamp, Frankfort/M
Kantorowicz E H 1957 *The King's Two Bodies. A Study in Mediaeval Political Theology*. Princeton University Press, Princeton
Lakoff G, Johnson M, Rothstein E 1999 *Philosophy in the Flesh: the Embodied Mind and its Challenge to Western Thought*. Basic Books, New York
Laqueur T 1990 *Making Sex, Body and Gender from the Greeks to Freud*. Harvard University Press, Cambridge, MA
Lorenz M 2000 Leibhaftige Vergangenheit. In: *Einführung in die Körpergeschichte*. Tübingen (edition diskord).
Mauss M 1989 Die Techniken des Körpers. In: *Do., Soziologie und Anthropologie*, Vol. 2. Fischer-Taschenbuch-Verlag, Frankfurt/M, pp. 197–220 (first edition 1935)
Merleau-Ponty M 1966 *Phänomenologie der Wahrnehmung*. de Gruyter, Berlin
Öhlschläger C 1997 *Körper-Gedächtnis-Schrift. Der Körper als Medium kultureller Erinnerung*. Schmidt, Berlin
O'Neill J 1985 *Five Bodies*. Cornell University Press, Ithaca, NY
Perrot P 1984 *Le corps féminin, XVIIIme-XIXme siècle. Le travail des apparences*. Édition du Seuil, Paris
Revel J, Peter J-P 1974 Le corps. L'homme malade et son histoire. In: LeGoff J, Nora P (eds.) *Faire de l'histoire. Nouveaux objects*. Gallimard, Paris, pp. 169–91
Rogozinski J 1996 Wie die Worte eines berauschten Menschen. Geschichtsleib und politischer Körper. In: Nagl-Docekal H (ed.) *Der Sinn des Historischen. Geschichtsphilosophische Debatten*. Fischer-Taschenbuch-Verlag, Frankfurt/M, pp. 333–72
Sarasin Ph 2001 *Reizbare Maschinen eine Geschichte des Körpers 1765–1914*. Suhrkamp, Frankfurt am Main
Schreiner K, Schnitzler N (eds.) 1992 *Gepeinigt, begehrt, vergessen Symbolik und Sozialbezug des Körper, im späten Mittelalter ind in der frühen Neuzeit*. Fink, Munich
Stafford B M 1993 *Body Criticism. Imaging the Unseen in Enlightenment Art and Medicine*. MIT Press, Cambridge, MA
Todes S 1990 *Body and World*, Cambridge/Mass./London 2001 (with introductions by Hubert L. Dreyfus and Piotr Hoffman) (= Rev. Ed. of: The human body as material subject of the world, New York (Garland) (Thesis 1963)
Turner B 1984 *The Body and Society*. Blackwell, Oxford
Virilio P 1994 *Die Eroberung des Körpers: vom Übermenschen zum überreizten Menschen Frankfurt am Main 1996*. Fischer-Taschenbuch-Verlag, München/Wien

J. Tanner

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Body Image and Gender

Body image refers to perceptions, thoughts, and feelings about physical appearance. Although most people are relatively satisfied with their bodies, evaluative measures of body image indicate that many individuals, both men and women, from young to old, are dissatisfied and wish they could change some aspect of their bodies. On average, women are more dissatisfied with their bodies than men. This article reviews the measurement of body image, the factors that influence its development, and gender differences in which aspects of physical appearance are the strongest determinants of body image.

1. Assessment of Body Image

The assessment of body image is based either on perceptual estimation or on affective reactions to self-evaluation. Unfortunately, the two measures are poorly related to each other. Perceptual estimations refer to the accuracy of judgments about the relative size of various body components whereas subjective evaluations refer to how people feel about those judgments.

1.1 Perceptual Estimates Are Not Reliable

Perceptual techniques are used to assess the accuracy of judgments about physical size. For example, researchers distort physical images (e.g., photographs, silhouettes, or mirror images) and then have subjects select their actual image from among the distortions. The difference between perceived and actual body size is used as an indicator of body image. Unfortunately, perceptual distortions of body image are common and occur with equal frequency across many diverse populations. Moreover, the perceptual distortions are unrelated to body image satisfaction and do not indicate any pathological condition. Relatively little research has found these measures to have adequate reliability or validity, and therefore they cannot be recommended at this time (Polivy et al. 1990).

1.2 Subjective Evaluations Reflect Satisfaction With One's Body

Body image is also assessed through subjective evaluation, such as asking people to rate satisfaction or dissatisfaction with their bodies or parts of their bodies. Common techniques include the following: self-ratings of physical attractiveness; self-ratings of specific body part satisfaction (e.g., hips, thighs, nose, and chest); self-ratings of weight, size, or shape satisfaction; and self-reports of affective reactions (such as anxiety or dysphoria) to thoughts about the body. People's feelings about their bodies are often unrelated to objective reality—many young girls who are objectively underweight feel that they are fat and are actively trying to lose weight. Most people are also typically happy with some of their physical features while being unhappy with others. Although people who dislike many of their body parts tend to have

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

more negative body image, how the various parts contribute to the whole has yet to be precisely determined.

1.3 Excessive Concerns About Body Image May Reflect Psychopathology

Some individuals are so preoccupied with trivial or imagined defects in their appearance that it interferes with normal psychological functioning. *Body dysmorphic disorder* is a pathological disturbance in body image in which individuals feel extreme distress about minor flaws in some part of the body, such as the size or shape of the ears, eyebrows, mouth, hands, feet, fingers, or buttocks. These thoughts can be so intrusive that individuals avoid work and public places, going out only at night when they cannot be seen. Those who have body dysmorphic disorder often undergo cosmetic surgery, but unfortunately for some, the surgery fails to alleviate body image dissatisfaction. Indeed, in some cases it increases their concerns as a doctors' willingness to provide surgery validates their views of abnormality, which may give rise to intensified or new preoccupations.

2. Development and Components of Body Image

Many different physical characteristics affect body image, including skin color, nose and ear size, hair loss, facial acne, pregnancy, wrinkles, varicose veins, straightness of teeth, and so on. Indeed, nearly every body part could influence overall body image, especially for those who perceive that body part to be unusual in some desirable or undesirable way. The features most closely associated with body image change over the course of lifespan development and differ as a function of gender.

2.1 Body Image Develops Early

Facially attractive infants receive more positive attention (e.g., increased smiling, eye contact, greater expectations of intelligence) than unattractive infants. For instance, mothers of attractive babies are more affectionate than are mothers of unattractive babies (Langlois et al. 1995). This differential treatment continues throughout childhood. School teachers, nurses, and parents rate attractive children as having better personalities, greater academic ability, and being more likely to be successful than unattractive children. These ratings are likely to have a strong impact on the self-esteem and body image of unattractive or overweight children. The children of parents whom themselves are preoccupied with body-weight issues or dieting or who have symptoms of disordered eating are at greater risk of developing body image dissatisfaction.

2.2 Adolescence is a Critical Period

Adolescence is a time of increased self-reflection and self-attention, and teenagers are especially concerned with how they are viewed by their peers (see *Adolescent Development, Theories of*). The physical changes that accompany adolescence, such as secondary sexual characteristics, oily complexion and acne, and tremendous individual variation in rate of growth lead most adolescents to be particularly conscious of physical appearance. Many girls become obsessed with body image issues, and by age 16 nearly all female adolescents report having at some point dieted in an attempt to lose weight. At the same time, boys are often concerned with being too short or not sufficiently muscular, especially when compared with popular male peers.

Negative comments from peers, particularly in the form of teasing, are important predictors of body image dissatisfaction (Grilo et al. 1994). Unattractive and obese adolescents are liked less, excluded from social events, and viewed by their peers as possessing more negative traits (e.g., lazy, sloppy) than their slimmer and more attractive peers. This social exclusion may promote a self-fulfilling prophecy, in that ostracized adolescents have fewer opportunities to acquire social skills (because of their limited social opportunities) and, in turn, their diminished social skills reinforce people's avoidance of them. In general, body image is at its most negative during adolescence.

2.3 Adulthood to Aging: Satisfaction to Dissatisfaction

As individuals mature into adulthood and focus on family and career issues, most individuals experience a reduction in body image concerns. However, individuals who gain large amounts of weight during adulthood experience increased body dissatisfaction, which may motivate efforts to improve a healthful lifestyle or lead to unhealthful dieting practices and weight cycling.

Changes in appearance and physical stamina that accompany old age may have a negative influence on body image. Older men may feel a decline in their body satisfaction because of their declining physical abilities, whereas women may be more concerned about excess weight as well as their wrinkling skin and hair loss. Both men and women may also be concerned about being too thin, since being frail may indicate poor or failing health.

3. Gender

Across the lifespan, women tend to have lower body image satisfaction than men. Women are more likely than men to evaluate specific body features negatively,

to attempt weight loss, to report anxiety about the evaluation of their physical appearance, and to have cosmetic surgery.

3.1 Women Are Especially Concerned With Body Weight

Body image dissatisfaction among women is usually related to self-perceptions of overweight. More than three-quarters of women would like to lose weight and almost none would like to gain weight. Believing oneself to be overweight, whether one is or not, is closely related to body image dissatisfaction. Beginning in early adolescence, women compare their body shape and weight with their beliefs about cultural ideals. A discrepancy from the cultural ideal often motivates people to undertake dieting in order to achieve a more attractive body size. Dieting is rarely successful, with fewer than 1 percent of individuals able to maintain weight loss over five years. Repeated dietary failures may exacerbate body image dissatisfaction and feelings of low self-esteem (Heatherton and Polivy 1992). Perfectionistic and low self-esteem women are particularly affected by body dissatisfaction, such that these personality traits in combination have been linked to increased bulimic symptoms (Vohs et al. 1999).

Black women are much less likely to consider themselves obese and are much more satisfied with their weight than are white women, despite the fact that black women are twice as likely to be obese. Black women also rate large black body shapes much more positively than do white women rating large white body shapes.

3.2 Men are Concerned with Size and Strength

Men are more likely than women to view their bodies as instruments of action (Franzoi 1995). Men who are physically large are viewed as more powerful than men who are physically slight, and many men try to increase their physical size. Very thin men are likely to experience body image dissatisfaction. Many men have a desire to be taller and tend to exaggerate their height. This is not surprising given the well-documented link between men's height and positive social outcomes. Many short men report dissatisfaction with their stature and some evidence suggests that they are more likely to experience decreased self-esteem and more negative body image relative to their taller male counterparts.

3.3 Physical Attractiveness Cues Vary by Gender

The specific attributes that are found attractive differ by gender. Facial features that indicate youthfulness (e.g., large eyes, small nose, big lips) and body features that are petite and thin (e.g., long legs, flat stomach) tend to be desirable for women. For men, facial features that imply maturity (e.g., square jaw, visible cheekbones) and body features that indicate mass and largeness (e.g., height, mesomorph build) tend to be desirable. There is only a modest association between objectively rated physical attractiveness and body image. However, subjective ratings of physical attractiveness (i.e., personal beliefs) are closely linked to body image satisfaction.

4. Culture

Throughout history, cultural influences have played significant roles in determining body image: the Greeks revered the male body, the Romans valued thinness, and people from the Middle Ages showed a preference for larger, rounder female body types. Thus, the determinants of body image change over time as a function of cultural and societal influences (Fallon 1990).

4.1 There is Variation Across Cultures

A limited number of physical attributes have defined attractiveness over time (e.g., fleshiness rather than flabbiness, cleanliness, and symmetry in one's body); the rest are defined within cultures. For instance, the stigma of obesity varies greatly across cultures: Fijians, Kenyans, Samoans, Mexicans and Israelis stigmatize obesity less than do Americans, Canadians, and the British. Individuals often conform their bodies to match local norms and standards. Extreme examples of this include the Burmese tradition of women affixing brass rings around their necks in order to stretch their necks to lengths of up to 40 cm, the East Indies tradition of filing teeth down to the gums, the South American (Abipone) tradition of inflicting deep wounds on the face, breasts, and arms, and the proliferation of Americans piercing and tattooing various body parts.

4.2 There is Variation within Cultures Over Time

Within a single culture, mandates for what is beautiful and desirable also undergo substantial variation across time, particularly for women. For instance, during the 1820s some women drank vinegar to lose weight and stayed up all night to look pale and fragile; in the mid-nineteenth century, a big, voluptuous figure was in vogue and women often worried about appearing too thin; in the early twentieth century a more slender but very sturdy physique was desirable. Although the preference for body size has fluctuated during the twentieth century, Western societies have been obsessed with thinness since at least 1970. Thus, body image is determined to a great extent by current societal norms and expectations.

See also: Adolescent Health and Health Behaviors; Body: Anthropological Aspects; Body, History of;

Cerebellum: Cognitive Functions; Culture as a Determinant of Mental Health; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Obesity and Eating Disorders: Psychiatric

Bibliography

Fallon A 1990 Culture in the mirror: sociocultural determinants of body image. In: Cash T F, Pruzinsky T (eds.) *Body Images: Development, Deviance, and Change*. Guilford, New York, pp. 80–109

Franzoi S L 1995 The body-as-object versus body-as-process: gender differences and gender considerations. *Sex Roles* **33**: 417–37

Grilo C M, Wilfley D E, Brownell K D, Rodin J 1994 Teasing, body image, and self-esteem in a clinical sample of obese women. *Addictive Behaviors* **19**: 443–50

Heatherton T F, Polivy J 1992 Chronic dieting and eating disorders: a spiral model. In: Crowther J H, Hobfall S E, Stephens M A P, Tennenbaum D L (eds.) *The Etiology of Bulimia Nervosa: the Individual and Familial Context*. Hemisphere, Washington, DC, pp. 133–55

Langlois J H, Ritter J M, Casey R J, Sawin D B 1995 Infant attractiveness predicts maternal behaviors and attitudes. *Developmental Psychology* **31**: 464–72

Polivy J, Herman C P, Pliner P 1990 Perception and evaluation of body image: the meaning of body size and shape. In: Olson J M, Zanna M P (eds.) *Self-inference Processes*. L Erlbaum Associates, Hillsdale, NJ, pp. 87–114

Vohs K D, Bardone A M, Joiner T E, Abramson L Y, Heatherton T F 1999 Perfectionism, perceived weight status, and self-esteem interact to predict bulimic symptoms: a model of bulimic symptom development. *Journal of Abnormal Psychology* **108**: 695–700

T. F. Heatherton

Borderline Personality Disorder

The term borderline personality is used in two main ways:

(a) Borderline personality disorder, one of the personality disorders described in the classification system of DSM-IV (American Psychiatric Association 1994), that is, a circumscribed constellation of pathological personality traits that constitutes approximately 15 percent of all diagnosed personality disorders, and, at least in the USA, probably around 15 percent of all patients in psychiatric hospitals.

(b) Borderline personality organization, that is, the psychostructural features that together characterize all severe personality disorders, and that differentiate them from less severe personality disorders (that is, from patients presenting neurotic personality organization). Borderline personality organization includes the borderline personality disorder as one dominant constellation of pathological personality traits among others.

This terminological issue has practical implications: Borderline personality disorder, with its circumscribed symptomatology, has been subject to research regarding specific treatment approaches. The present trend is to treat these patients with a supportive psychotherapeutic approach, based either on psychoanalytic principles or on cognitive behavioral ones, plus a psychopharmacological approach to certain target symptoms. Dialectic behavior therapy (DBT) is one form of cognitive behavioral therapy that has proven effective in reducing suicidal and parasuicidal behavior in borderline patients (Linehan et al. 1991), and psychodynamic psychotherapy has proven effective in treating borderline patients in a day hospital setting (Bateman and Fonagy 1999). Low dose neuroleptics are effective in reducing severe anxiety and cognitive disorganization characteristic of borderline patients, and antidepressive medication, particularly the use of SSRI medication, has been effective in palliating their symptomatic depression (Soloff 1998).

1. The Characteristics of Borderline Personality Disorder

The DSM-IV description of borderline personality disorder defines it as a pervasive pattern of instability of interpersonal relationships, self-image, and affects, with marked impulsivity beginning by early adulthood and present in a variety of contexts. It is indicated by five (or more) of the following:

(a) Frantic efforts to avoid real or imagined abandonment.

(b) A pattern of unstable and intense interpersonal relationships alternating between extremes of idealization and devaluation.

(c) Identity disturbance: persistent and markedly disturbed, distorted, or unstable self-image and/or sense of self.

(d) Impulsivity in at least two areas that are potentially self-damaging, (e.g., spending, sex, substance use, reckless driving, binge eating).

(e) Recurrent suicidal threats, gestures, or behavior, or self-mutilating behavior.

(f) Affective instability due to a marked reactivity of mood (e.g., intense episodic dysphoria, irritability, or anxiety, usually lasting a few hours and only rarely more than a few days).

(g) Chronic feelings of emptiness.

(h) Inappropriate, intense anger or lack of control of anger (e.g., frequent displays of temper, constant anger, recurrent physical fights).

(i) Transient, stress-related paranoid ideation, or severe dissociative symptoms.

This personality disorder has substantial comorbidity or overlap with other severe personality disorders, particularly the histrionic, avoidant, dependent, paranoid, and narcissistic personality disorders.

Although there are no precise statistics, it is reasonable to assume that between 0.2 percent and 1.8 percent of the general population suffer from this personality disorder. Factor analytic studies have concluded that the dominant three factors underlying this symptomatic constellation are identity diffusion, affect dysregulation, and impaired impulse control (Clarkin et al. 1993). In the study of the changes in symptomatology that occur with treatment, there is a tendency for impulse control to improve first, affect dysregulation later, and identity disturbance, the most resistant aspect of this pathology, only after extended periods of treatment, if ever.

2. *Classification Problems: Categorical vs. Dimensional Models of Personality Disorders*

A currently dominant dimensional model, the five-factor model, has synthesized numerous factor analyses into the proposal that neuroticism, extroversion, openness, agreeableness, and conscientiousness constitute basic factors that may describe all 'officially' accepted personality disorders in DSM-IV (Costa and Widiger 1994, Widiger et al. 1994). The problem is whether these are really fundamental determinants of the organization of the normal personality or even of the personality disorders. An 'equalization' of these character traits seems strange when applied to the subtleties of the clinical features of specific personality constellations. To develop factorial profiles for each personality disorder on the basis of those five factors has an eerie quality of unreality for the experienced clinician.

Those researchers who are inclined to maintain a categorical approach to the classification of personality disorders, usually clinical psychiatrists motivated to find specific disease entities, tend to proceed differently. They study the clinically prevalent combinations of pathological personality traits, carry out empirical research regarding the validity and reliability of clinical diagnoses, attempt to achieve a clear differentiation between personality disorders, and, of course, keep in mind the clinical relevance of their approaches (Akhtar 1992, Stone 1993). This approach, pursued in DSM-III and DSM-IV, has helped to clarify—or at least to permit the clinical psychiatrist to become better acquainted with—some frequently seen personality disorders. The approach has been plagued, however, by the high degree of comorbidity of the severe types of personality disorders, and by the unfortunate politicalization of decision making, by committee, of what personality disorders to include and exclude in the official DSM system, and under what labels (Jonas and Pope 1992, Kernberg 1992, Oldham 1994).

Recent studies of alteration in neurotransmitter systems in severe personality disorders, particularly in the borderline personality disorder, although still tentative and open to varying interpretations, point to the possibility that neurotransmitters are related to specific distortions in affect activation (Stone 1993). Abnormalities in the adrenergic and cholinergic systems, e.g., may be related to general affective instability; deficits in the dopaminergic system may be related to a disposition toward transient psychotic symptoms in borderline patients; impulsive, aggressive, self-destructive behavior may be facilitated by a lowered function of the serotonergic system.

These aspects of inborn dispositions to the activation of aggression mediated by the activation of aggressive affect states are complementary to the now well-established findings that structured aggressive behavior in infants may derive from early, severe, chronic physical pain, and that habitual aggressive teasing interactions with mother are followed by similar behaviors of infants, as we know from the work of Galenson (1986) and Fraiberg (1983). The impressive findings of the prevalence of physical and sexual abuse in the history of borderline patients confirmed by investigators both here and abroad (Marziali 1992, Perry and Herman 1993; van der Kolk et al. 1994) provide additional evidence of the influence of trauma on the development of severe manifestations of aggression.

3. *A Psychostructural Nosology*

Borderline personality organization is characterized by lack of integration of the concept of self and significant others, that is, identity diffusion, a predominance of primitive defensive operations centering around splitting, and maintenance of reality testing. The defensive operations of splitting and its derivatives (projective identification, denial, primitive idealization, omnipotence, omnipotent control, devaluation) have as a basic function to maintain separately the idealized and persecutory internalized object relations derived from the early developmental phases predating object constancy: that is, when aggressively determined internalizations strongly dominate the internal world of object relations, in order to prevent the overwhelming control or destruction of ideal object relations by aggressively infiltrated ones. This primitive constellation of defensive operations centering around splitting thus attempts to protect the capacity to depend on good objects and escape from terrifying aggression.

Reality testing, as mentioned before, is present in borderline personality organization. It refers to the capacity to differentiate self from nonself, intrapsychic from external stimuli and to maintain empathy with ordinary social criteria of reality, all of which are typically lost in the psychoses, and manifested particularly in hallucinations and delusions (Kernberg 1984). All patients with psychotic personality organ-

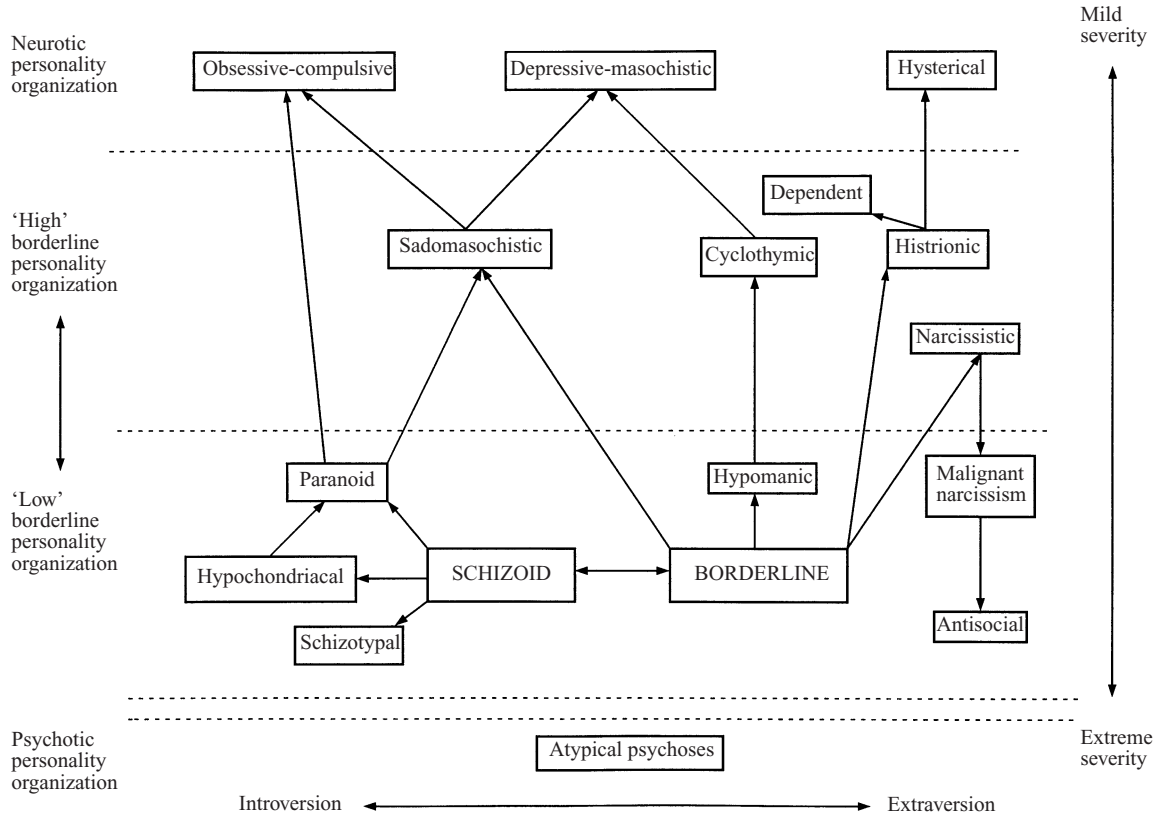

Figure 1
Personality disorders: their mutual relationships

ization really represent atypical forms of psychosis. Therefore, strictly speaking, psychotic personality organization represents an exclusion criterion for the personality disorders in a clinical sense.

Borderline personality organization includes all the severe personality disorders in clinical practice. Typical personality disorders included here are the borderline personality disorder in the DSM IV sense, the schizoid and schizotypal personality disorders, the paranoid personality disorder, the hypomanic personality disorder, hypochondriasis (a syndrome which has many characteristics of a personality disorder proper), the narcissistic personality disorder (including the syndrome of malignant narcissism), and the antisocial personality disorder.

All these patients present identity diffusion, the manifestations of primitive defensive operations, and many evince varying degrees of superego deterioration (antisocial behavior). A particular group of patients typically suffer from significant disorganization of the superego, namely, the narcissistic personality disorder, the syndrome of malignant narcissism, and the antisocial personality disorder.

All the personality disorders within the borderline spectrum present, because of the identity diffusion, severe distortions in their interpersonal relations—particularly problems in intimate relations with others, lack of consistent goals in terms of commitment to work or profession, uncertainty and lack of direction in their lives in many areas, and varying degrees of pathology in their sexual life. They often present an incapacity to integrate tenderness and sexual feelings, and they may show a chaotic sexual life with multiple polymorphous perverse infantile tendencies. The most severe cases, however, may present with a generalized inhibition of all sexual responses. All these patients also evince nonspecific manifestations of ego weakness, that is, lack of anxiety tolerance, of impulse control, and of sublimatory functioning in terms of an incapacity for consistency, persistence, and creativity in work.

An additional group of personality disorders also presents the characteristics of borderline personality organization, but these patients are able to maintain more satisfactory social adaptation, and are usually more effective in obtaining some degree of intimacy in

object relations and in integrating sexual and tender impulses. Thus, in spite of presenting identity diffusion, they also evince sufficient nonconflictual development of some ego functions, superego integration, and a benign cycle of intimate involvements, capacity for dependency gratification, and a better adaptation to work that make for significant quantitative significant differences. They constitute what might be called a 'higher level' of borderline personality organization or an intermediate level of personality disorder. This group includes the cyclothymic personality, the sadomasochistic personality, the infantile or histrionic personality, and the dependent personalities, as well as some better functioning narcissistic personality disorders.

The next level of personality disorder, namely, neurotic personality organization, is characterized by normal ego identity and the related capacity for object relations in depth, ego strength reflected in anxiety tolerance, impulse control, sublimatory functioning, effectiveness and creativity in work, and a capacity for sexual love and emotional intimacy disrupted only by unconscious guilt feelings reflected in specific pathological patterns of interaction in relation to sexual intimacy. This group includes the hysterical personality, the depressive-masochistic personality, the obsessive personality, and many so-called 'avoidant personality disorders,' in other words, the 'phobic character' of psychoanalytic literature (which remains a problematic entity).

Figure 1 summarizes the relationship among all the personality disorders mentioned, and represents their overall classification into neurotic and borderline personality organization.

What follows is a summary of the psychoanalytically based psychotherapy for borderline personality organization as developed and manualized by a team of psychoanalysts, psychoanalytic psychotherapists, and researchers at the Department of Psychiatry of the Cornell University Medical College (Clarkin et al. 1999).

4. *Therapeutic Strategy*

From a therapeutic perspective, the main objective of the psychodynamic psychotherapy to be described is to focus upon the syndrome of identity diffusion, its expression in the form of the activation of primitive object relations in the transference, and the exploration of these primitive tranferences as they reflect early internalized object relations of an idealized and persecutory kind. The goal of this strategy is to identify such primitive transference paradigms and then to facilitate their gradual integration, so that splitting and other primitive defensive operations are replaced by more mature defensive operations, and identity diffusion is eventually resolved (Kernberg 1984).

The essential strategy takes place in three consecutive steps: First, the dominant primitive object relation is identified in the transference, and is described in an appropriate metaphorical statement that includes a hypothesized relation between two people linked by a dominant peak affective state. Second, within this dominant relationship, the patients' representation of self relating to the representation of a significant other ('object representation') is described, and patients are shown how that self-representation, linked to its corresponding object representation by a specific affect, is activated with frequent role reversals in the transference. These role reversals show themselves in the patients' alternatively enacting their representation of self or of the corresponding object, while projecting the other member of the internalized object relationship into the therapist. In this second phase patients learn not only to understand the different ways in which the same transference disposition may show in completely contradictory behaviors, but also to gradually tolerate their identification with both self and object representations in this interaction. Thirdly, the idealized internalized object relations are interpretively integrated with their corresponding, opposite, split-off persecutory ones, so that the patients, who already have learned to accept their identification with contradictory internalized representations of self and object at different points of their treatment experience, now learn to integrate them, to accept that they harbor both loving and hateful feelings toward the same object, that their self concept is both 'good' and 'bad,' and that their objects are neither as exclusively good or bad as they originally perceived them.

This gradual integration of the internal world of object relations leads towards the tolerance of ambivalence, a toning down and maturing of all affective experiences and emotional relations with significant others, a decrease in impulsive behaviors, and a growing capacity for self-reflection and empathy with significant others as the patients' self-concept consolidates in an integrated view of themself, and they experience relationships with significant others in a new, integrated way (see *Transference in Psychoanalysis*).

5. *Therapeutic Techniques*

The essential techniques taken from psychoanalysis that, in their respective modification, characterize the technique of this psychodynamic psychotherapy, are: interpretation, transference analysis, and technical neutrality.

The technique of interpretation includes the clarification of the patient's subjective experience, the tactful confrontation of those aspects of the patient's nonverbal behavior that are dissociated or split off from his or her subjective experience, the interpret-

ation in the 'here and now' of hypothesized unconscious meanings of the patient's total behavior and their implicit conflictual nature, and the interpretation of a hypothesized origin in the patient's past of that unconscious meaning in the here and now.

Transference analysis refers to the clarification, confrontation, and interpretation of unconscious, pathogenic internalized object relations from the past that are typically activated very early in the relationship with the therapist. In simplest terms, the transference reflects the distortion of the initial therapist–patient relationship by the emergence of an unconscious, fantasized relationship from the past that the patient unwittingly or unwillingly enacts in the present treatment situation.

Technical neutrality refers to the therapist's not taking sides regarding the patient's unconscious conflicts, and helping the patient to understand these conflicts by maintaining a neutral position. Therapists, in their total emotional reaction to the patient, that is, their countertransference reaction, may experience powerful feelings and the temptation to react in specific ways in response to the patient's transference challenges. Utilizing their countertransference response to better understand the transference without reacting to it, therapists interpret the meanings of the transference from a position of concerned objectivity, which is the most important application of the therapist's position of technical neutrality.

The therapist's emotional response to patients at times reflects empathy with the patients' central subjective experience (concordant identification in the countertransference), and reflects at other times the therapist's identification with what the patients cannot tolerate in themselves, and are projecting onto the therapist (complementary identification in the countertransference). Both reactions, when the therapist is able to identify and observe them, serve as valuable sources of information.

Countertransference analysis is in fact an essential aspect of this psychotherapy. The countertransference, defined as the total emotional reaction of the therapist to the patient at any particular point in time, needs to be explored fully by the therapist's self-reflective function, controlled in the therapist's firmly staying in role, and utilized as material to be integrated into the therapist's interpretive interventions. Thus, the therapist's 'metabolism' of the countertransference as part of the total material of each hour, rather than its communication to the patient, characterizes this psychotherapeutic approach.

The tendency to severe acting out of the transference characteristic of borderline patients has been mentioned already; in addition to its management by the modification of technical neutrality and limit setting in the hours mentioned before, the treatment begins with the setting up of a treatment contract, which includes not only the treatment setting and frame, but also specific, highly individualized conditions for the treatment that derive from life-threatening and potentially treatment-threatening aspects of the patient's psychopathology. Particularly, the establishment of realistic controls and limit setting that protect the patient from suicidal behavior and other destructive or self destructive patterns of behavior are typical objectives of contract setting.

In the course of the treatment, it will become unavoidable to face very primitive traumatic experiences from the past reactivated as traumatic transference episodes in which, unconsciously, the patient may express traumautophilic tendencies in an effort to repeat past traumas in order to overcome them. Primitive fears and fantasies regarding murderous and sexual attacks, primitive hatred, efforts to deny all psychological reality in order to escape from psychic pain are the order of the day in the psychodynamic psychotherapy of these patients.

This internalized object relation, that has transformed the primitive affect of rage into a characterologically anchored, chronic disposition of hatred is activated in the transference with alternating role distribution: the patients' identification, for periods of time, with their victim self while projecting the sadistic persecutor onto the therapist, will be followed, rapidly, in equally extended periods of time, by the projection of their victimized self onto the therapist while the patients identify themselves, unconsciously, with the sadistic perpetrator. Only a systematic interpretation of the patient's unconscious identification with both victim and perpetrator may resolve this pathological constellation and lead to a gradual integration of dissociated or split-off self representation into the patient's normal self. The effects of the traumatic past reside in the patient's internalized object relations; the key to its therapeutic resolution is coming to terms with this double identification.

See also: Behavior Therapy: Psychiatric Aspects; Behavior Therapy: Psychological Perspectives; Differential Diagnosis in Psychiatry; Mental Health and Normality; Personality Disorders; Personality Theory and Psychopathology

Bibliography

Akhtar S 1992 *Broken Structures*. Jason Aronson, Northvale, NJ

American Psychiatric Association (APA) 1968 *Diagnostic and Statistical Manual of Mental Disorders: DSM-II*, 2nd ed. American Psychiatric Association, Washington, DC

American Psychiatric Association (APA) 1980 *Diagnostic and Statistical Manual of Mental Disorders: DSM-III*, 3rd ed. American Psychiatric Association, Washington, DC

American Psychiatric Association (APA) 1994 *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*. American Psychiatric Association, Washington, DC

Bateman A, Fonagy P 1999 Effectiveness of partial hospitalization in the treatment of borderline personality disorder: A

randomized controlled trial. *American Journal of Psychiatry* **156**: 1563–69
Clarkin J F, Hull J W, Hurt S W 1993 Factor structure of borderline personality disorder criteria. *Journal of Personality Disorders* **7**(2): 137–43
Clarkin J F, Yeomans F E, Kernberg O F 1999 *Psychotherapy for Borderline Personality*. Wiley, New York
Costa P T, Widiger T A 1994 Introduction. In: Costa P T, Widiger T (eds.) *Personality Disorders and the Five-factor Model of Personality*. APA, Washington, DC, pp. 1–10
Fraiberg A 1983 Pathological defenses in infancy. *Psychoanalytic Quarterly* **60**: 612–35
Galenson E 1986 Some thoughts about infant psychopathology and aggressive development. *International Review of Psychoanalysis* **13**: 349–54
Jonas J M, Pope H G 1992 Axis I comorbidity of borderline personality disorder: Clinical implications. In: Clarkin J F et al. (eds.) *Borderline Personality Disorder*. Guilford Press, New York, pp. 149–60
Kernberg O F 1984 *Severe Personality Disorders: Psychotherapeutic Strategies*. Yale University Press, New Haven, CT
Kernberg O F 1992 *Aggression in Personality Disorder and Perversions*. Yale University Press, New Haven, CT
Linehan M M, Armstrong H E, Suarez A, Allmon D, Heard H 1991 Cognitive-behavioral treatment of chronically parasuicidal borderline patients. *Archives of General Psychiatry* **48**: 1060–4
Marziali E 1992 The etiology of borderline personality disorder: Developmental factors. In: Clarkin J F et al (ed.) *Borderline Personality Disorder*. Guilford Press, New York, pp. 27–44
Oldham J M 1994 Personality disorders. *Journal of the American Medical Association* **272**: 1770–6
Perry J C, Herman J L 1993 Trauma and defense in the etiology of borderline personality disorder. In: Paris J (ed.) *Borderline Personality Disorder*. American Psychiatric Press, Washington, DC, pp. 123–40
Soloff P H 1998 Algorithms for pharmacological treatment of personality dimensions: Symptom-specific treatments for cognitive-perceptual, affective, and impulsive-behavioral dysregulation. *Bulletin of the Menninger Clinic* **62**: 195–214
Stone M 1993 *Abnormalities of Personality*, 1st ed. Norton, New York
van der Kolk B A et al. 1994 Trauma and the development of borderline personality disorder. In: Share I (ed.) *Borderline Personality Disorder: The Psychiatric Clinics of North America*. W. B. Saunders, Philadelphia, PA, pp. 715–30
Widiger T A et al. 1994 A description of the DSM-III-R and DSM-IV personality disorders with the five-factor model of personality. In: Costa P T, Widiger T (eds.) *Personality Disorders and the Five-factor Model of Personality*. American Psychological Association, Washington, DC, pp. 41–56

O. F. Kernberg

Borders, Anthropology of

1. Introduction

The anthropology of borders is characterized by three perspectives which, though they sometimes overlap, can be distinguished by their relative emphasis on the cultural, territorial, and social dimensions of borders, respectively. These three dimensions may distinguish different types of border, or they may be aspects of a single border. Indeed, anthropologists are rarely interested in only one type of border or the other, and in many cases the borders they analyze have all three dimensions.

Anthropologists have a long-standing interest in borders. In the past, this was reflected in a concern with bounding their field of study. Cultural variation was considered a function of geographic and social isolation and early ethnographers advocated analyzing this diversity by treating each social unit as a discrete organic whole. But as the mobility of people dramatically increased (with urbanization, industrialization, etc.), culture contact rather than cultural isolation became the focus of the day. Diffusionists and acculturationists had always been interested in culture contact, but their work was overshadowed and ultimately displaced by the rapid rise and supremacy of functionalism. The belated 'rediscovery' of the value of their insights for contemporary theoretical questions about global cultural interconnectedness indicates a major shift within the discipline: from a concern with what borders encompass to an interest in the borders themselves.

Some anthropologists have been primarily interested in cultural borders which separate and connect different worlds of meaning and identity, others in borders which mark out geopolitical space, and yet others in borders which order social relations and indicate membership of 'community.' All three have been integral elements in the emergence of an anthropology of borders.

2. Cultural Borderlands

Borders have become a metaphor for the cultural flux and indeterminacy of much contemporary life. In anthropology the border metaphor was introduced by critics of the classic anthropological view of culture as shared, consensual, and discrete. Often with personal experience of cultural contradictions—as members of sexual, ethnic, or other minorities—these critics sought ways to study the differences within and the spaces between cultures, ways that could incorporate the changes, inconsistencies, and incommensurabilities of everyday life. These inter-cultural spaces are often referred to as 'borderlands,' a usage that evokes the geopolitical and the metaphorical, the literal and the conceptual. In this view, borders and borderlands exist not just at the edges of the nation-state, but anywhere cultures meet (Rosaldo 1989).

Borderlands are zones of cultural overlap characterized by a mixing of cultural styles. They are liminal spaces, simultaneously dangerous and sites of creative cultural production open to cultural play and experimentation as well as domination and control. The fusion of registers in borderlands may be ultimately empowering for those who inhabit these zones, but it

need not always be so, as testified by the lives of many Mexican migrants in the American Southwest.

Not everyone agrees that it is helpful to extend the use of 'border' and 'borderland' in this way. When cultural encounters share some of the specific socio-political processes characteristic of borders between states, such as unequal access to official forms of power, the borderland metaphor may be appropriate. But where there is no strong analytical connection to state border processes, understanding of the border can become reductive and delocalized (Heyman 1994).

Cultural differences and the juxtaposition of different worlds of meaning come to be emphasized at the expense of stressing inequalities of power. Heyman and others feel uneasy about this differential emphasis, reflecting wider tensions within the discipline over the relative prominence to be given to culture and power. The force of the image, they feel, takes over from the analysis.

In a sense, of course, borders are always metaphors, since they are arbitrary constructions based on cultural convention. Metaphors, moreover, are part of the 'discursive materiality of power relations' and in this respect no less concrete in their consequences than state borders (Brah 1996, p. 198). Metaphorical and state borders are not, therefore, as far apart as some studies imply. In fact, the anthropology of borders has benefited from the productive interaction of approaches that emphasize one or the other. The following sections outline other anthropological analyses of borders from which the border-as-metaphor draws its force and resonance.

3. *Territorial and Political Borders*

State borders entail a mapping out in geographic space and recognition in international law. They mark the limits of sovereignty and of state control over citizens and subjects, limits that may be upheld by force or by the threat of force. They are often highly visible physically, which has led some scholars to refer to them as 'real' borders in contrast to borders without territorial counterparts. This can be misleading, and the materiality of state borders should not blind us to their cultural and symbolic dimensions (just as we should not assume that cultural and symbolic borders automatically have no materiality).

Apart from the Mexican–US border, state borders have not until recently been subject to systematic comparative scrutiny by anthropologists. Where they did appear in ethnographic accounts, they often figured only as a backdrop to some other line of inquiry. Prior to the 1970s even studies of the Mexican–US border rarely included the border as a variable in the analysis. As a result, the anthropology of state borders was slow to develop.

A fledgling 'school' of anthropological border studies briefly emerged at the University of Manchester, UK, in the 1960s, but subsequently fizzled out. During this period, several books appeared specifically with borders in their titles; each influenced by Max Gluckman who was Departmental Chair at the time (e.g., Cohen 1965, Harris 1972). Taken together, they began something that took another 20 years to bear fruit. But though these books did not problematize the state borders near which their authors did research, and though they did not cross-refer to one another, they nevertheless identified some common themes in the political anthropology of borders which laid a solid foundation for contemporary anthropological analyses of nations and states. They pointed out, for instance, how proximity to a state border can intensify local conflicts, which can easily escalate into conflicts over nationality, a threat which only cross-cutting linkages (of kinship, common residence, and shared social and economic interests) help to keep in check. Understanding these borders thus requires local ethnographic knowledge, and not just knowledge of state level institutions and international relations.

The value of localized studies for understanding how cultural landscapes are superimposed across social and political divides was developed by Cole and Wolf (1974) whose field site in the Italian Tyrol was specifically chosen because its successive historical partitioning allowed them to explore the transformation of local political loyalties in relation to nation-building. What particularly interested Cole and Wolf about the South Tyrol was the durability of a cultural frontier long after the political borders of state and empire had shifted. National boundaries had clearly survived the demise of state borders, and remained important in everyday life. Despite their similarities, the two villages studied had followed a different political and cultural course since World WarI. Villagers minimized these ideological and cultural differences in public encounters, but in private quickly resorted to ethnic stereotypes to explain the actions of the others. Cole and Wolf thus reiterate Barth's (1969) observation that ethnic boundaries may be maintained despite relations across them. Their major contribution, however, was in demonstrating the need to combine the study of local and extra local influences to understand and explain this process.

Here, then, is an example of where ethnic boundaries result from, and evolve with, the rise and demise of state borders. Each can only be understood by reference to the other. In this respect Cole and Wolf combined symbolic border studies (see section 4) with a political economy perspective that situates ethnographic knowledge of local boundaries within wider historical and political processes, a novel combination that marked an important transition in the anthropology of borders and heralded the beginning of a new form of inquiry. Subsequent anthropological research on national and international borders was to draw explicitly or implicitly upon this groundbreaking work

(see Donnan and Wilson 1999 for examples). Anthropologists began to use their field research at state borders as a means of widening perspectives in political anthropology to encompass the formal and informal ties between local communities and the larger polities of which they are a part. Some studied border areas as a way of examining how proximity to an international border could influence local culture or could create the conditions that shape new rural and urban communities. Others focused on the voluntary and involuntary movement of people across borders as traders, migrants, and refugees. Yet others concentrated on the symbols and meanings which encode border life. The Mexican–US border, where the issues of underdevelopment, transnationalism, and the globalization of power and capital, among other aspects of culture, increasingly concern the growing number of historically informed and wide-ranging ethnographic accounts, offers excellent examples of all of these trends (see Alvarez 1995, Stoddard 1975 for overviews).

Regardless of theoretical orientation or locale, however, most of these border studies have focused on how social relations, defined in part by the state, transcend the territorial limits of the state and, in so doing, transform the structure of the state at home and in its relations with its neighbors.

This new anthropological interest in how local border developments can have an impact on national centers of power and hegemony was partly stimulated by historical analyses of localities and the construction of national identities (e.g., Sahlins 1989). Cole and Wolf's insistence on the need to view the anthropology of borders as historical anthropology is recalled. As the South Tyrol case so clearly shows, borders are spatial and temporal records of relationships between local communities and between states. Ethnographic explorations of the intersection of symbolic and state borders have salience beyond anthropology because of what they may reveal about the history of cultural practices and about the role of border cultures and communities in policy-making and diplomacy. This kind of research highlights the growing importance of a border perspective in political anthropology, a perspective in which the dialectical relations between border areas and their nations and states take precedence over local culture viewed with the state as a backdrop.

The next section outlines the work of some of the scholars whose innovative theorizing of the symbolic dimensions of social boundary making has provided a cornerstone to the anthropology of state borders worldwide.

4. *Social and Symbolic Borders*

One of the most influential theorists in the anthropology of borders has been Fredrik Barth (1969), who argues that people may traverse group boundaries and maintain regular relations across them without affecting the durability and stability of the boundaries themselves. Barth's focus is on ethnic groups, which he argues are social constructions, the result of individuals strategically manipulating their cultural identity according to context, rather than the pre-determined outcome of some objective list of culture traits. Membership of ethnic groups should thus be understood as based on self-ascription and ascription by others: individuals claim membership in a particular group and others confirm or question that claim. Attention is thereby drawn to the boundary between groups: to how and why people distinguish themselves from others, and to the rules of behavior that sustain such boundaries in the face of relations that transcend them. Barth's boundaries are above all 'social' boundaries: they may or may not have a territorial dimension, and cultural differences become significant only in so far as they are useful for organizing social relations.

Barth's work both revolutionized the study of ethnic groups and stimulated further research on borders. Some of this research demonstrates that borders mark affective and identificatory as well as structural and organizational disjunctures: those on the inside not only inhabit a separate social system from those on the outside but they identify themselves differently. Any social border is thus a consequence of the various possible relationships between these two dimensions on each side of the border as well as across it (Wallman 1978). This highlights the relational nature of social boundaries, drawing attention to what they mean to people and to how they are marked. These are issues developed by Anthony Cohen (1985). Echoing Barth, Cohen suggests that borders are constructed by people in their interactions with others, from whom they wish symbolically to distinguish themselves. However, we cannot predict what the distinguishing features of these symbolic borders will be, or exactly where the lines will be drawn. Moreover, they may mean different things to different individuals, both to those within a border as well as those outside it. In fact, borders recognized by some may be invisible to others. The anthropologist's task is to uncover these borders and their meanings in order to grasp what it involves to belong or be excluded.

Cohen has been criticized for devoting more attention to one side of the border than the other. For him, the significance of symbolic borders is that they allow people to retain a sense of distinctiveness in a world where the diversity of local communities is increasingly threatened by cultural and structural incorporation into the nation-state. This inevitably leads him to emphasize what goes on within a border, since this is how he can identify the symbolic beliefs and practices that people deploy to mark and celebrate their sense of difference. When he does consider what lies beyond the border, it is often only to show how

external events can be manipulated to symbolic advantage at the local level.

Similar criticisms have been leveled at Barth. As we have seen, Barth also emphasizes that boundary making involves both self-ascription and ascription by others. But he too tends to focus on only one side of the border, emphasizing internal identification rather than external categorization and the shaping influence of wider structures. This minimizes the relative power relations upon which the external ascriptive process depends, and underplays the fact that some people can impose their categorizations on other people. Relationships of domination and subordination can only be restored to theoretical centrality in border studies by recognizing that those on one side of a border may be better able than those on the other side to determine where and how the line is drawn (cf. Jenkins 1997). Indeed, we saw earlier that some anthropologists who focus on state borders have attempted to build this dimension into their analyses.

5. *Conclusion*

Three reasonably distinct but mutually enriching streams in the anthropology of borders have been identified. All three show that anthropologists have brought to the study of borders sensitivity to what borders mean to those whose lives they enframe. It is in this emphasis on how borders are constructed, negotiated, and viewed from 'below' that the value and distinctiveness of an anthropology of borders arguably resides

See also: Ethnicity, Sociology of; Ethnocentrism; Frontiers in History; Groups, Sociology of; Nationalism: General; Racism, Sociology of; State, History of; Symbolic Boundaries: Overview; Symbolism in Anthropology; Xenophobia

Bibliography

Alvarez R R 1995 The Mexican–US border: The making of an anthropology of borderlands. *Annual Review of Anthropology* **24**: 447–70

Barth F 1969 Introduction. In: Barth F (ed.) *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*. George Allen and Unwin, London

Brah A 1996 *Cartographies of Diaspora: Contesting Identities*. Routledge, London

Cohen A 1965 *Arab Border-Villages in Israel: A Study of Continuity and Change in Social Organization*. Manchester University Press, Manchester, UK

Cohen A P 1985 *The Symbolic Construction of Community*. Tavistock, London

Cole J W, Wolf E 1974 *The Hidden Frontier: Ecology and Ethnicity in an Alpine Valley*. Academic Press, New York

Donnan H, Wilson T M 1999 *Borders: Frontiers of Identity, Nation and State*. Berg, Oxford

Harris R 1972 *Prejudice and Tolerance in Ulster: A Study of Neighbours and 'Strangers' in a Border Community*. Manchester University Press, Manchester, UK

Heyman J 1994 The Mexico–United States border in anthropology: a critique and reformulation. *Journal of Political Ecology* **1**: 43–65

Jenkins R 1997 *Rethinking Ethnicity: Arguments and Explorations*. Sage, London

Rosaldo R 1989 *Culture and Truth: The Remaking of Social Analysis*. Beacon Press, Boston

Sahlins P 1989 *Boundaries: The Making of France and Spain in the Pyrenees*. University of California Press, Berkeley, CA

Stoddard E R 1975 The status of borderlands studies: sociology and anthropology. *The Social Science Journal* **12**: 29–54

Wallman S 1978 The boundaries of 'race': Processes of ethnicity in England. *Man* **13**: 200–17

H. Donnan

Boserup, Ester (1910–99)

Ester Boserup (1910–99) summed up her research as focusing '… on the interplay of economic and non-economic factors in the process of social change, both today and in the past, viewing human societies as dynamic relationships between natural, economic, cultural, and political structures, instead of trying to explain them within the framework of one or a few disciplines' (1999). Boserup's book, *Woman's Role in Economic Development* (1970) generally is acknowledged as initiating the formal field of study: Women in Development (WID) frequently referred to as Gender and Development (GAD) since the 1980s. While the WID/GAD field is multidisciplinary and built on the research and practice of numerous women and men, Boserup's contribution is substantial. By demonstrating that women are often marginalized in the course of development rather than gaining from the process, her work helped shift the focus of development research from growth of income toward people's wellbeing.

Boserup was born in Copenhagen, Denmark, May 18, 1910. She studied economics at the University of Copenhagen during the Great Depression (1929–35) which influenced her search for theoretical explanations of economic processes that went beyond the equilibrium approach. Her major field of study stressed theoretical economics but also allowed her to take courses in sociology and agricultural policy. After graduating, she worked for 12 years for the Danish government writing on regulations and agricultural policy in Europe. From 1947 to 1957, Boserup worked in Geneva for the United Nations' Economic Commission for Europe and began collaborative work with the FAO (UN Food and Agriculture Organization). During this period, her research focus expanded to developing countries.

In 1957, Boserup and her husband (Mogens Boserup) left the UN and went to India for the *Asian*

Drama research project with Gunnar Myrdal. Ester Boserup drafted chapters on agriculture, Gandhism, and development planning. Her travels in South and Southeast Asia exposed her to different types of farming and land tenure systems. She began to formulate her own concepts of the process of development that countered much of the received wisdom of the time; for example, she observed that labor was still a constraint in traditional agricultural systems so that the marginal product of labor could not be zero (or close to it) as commonly assumed and that rural-urban migration was often caused more by pull factors, such as attractive money incomes, than by push factors. In 1960, Boserup and her husband resigned from the *Asian Drama* project because their ideas about the development process had grown too far from Myrdal's.

For the rest of her life, Boserup worked as a consultant for organizations such as the FAO Trade Division in Rome and the Center for Industrial Development (later called UNIDO). This was the period when she published her most influential works: *The Conditions of Agricultural Growth* (1965/1993) and *Woman's Role in Economic Development* (1970). She played a central role in the UN World Population Conference in Bucharest in 1974 and the first UN World Conference on Women in Mexico City in 1975. In the 1990s, she stopped participating in conferences because of her health, but she continued to be active in research and writing, stressing dynamic, multidisciplinary approaches to development issues, until her death September 24, 1999 in Ascona, Switzerland.

1. *Boserup's Theoretical Contributions*

Boserup's work reflects her background in classical and neoclassical economics. In 1936, she began building her analysis with a comparison of Marx's theory of underconsumption and Keynes's theory of propensity to consume (1999). Boserup established her reputation as a researcher with *The Conditions of Agricultural Growth* in the mid-1960s. Her main thesis in this work is that technological change and property rights are endogenous. The book was published at a time when growth theory was dominated by neoclassical models that treated technology as exogenous. Boserup hypothesized that increasing population density led to transformation of the environment and introduction of new technology, in particular in a change of agricultural systems from shifting (slash and burn) techniques to use of the plow and irrigation. This view countered the Malthusian and neo-Malthusian positions on population growth that took the carrying capacity of the earth and technology as given and predicted that continued population growth would lead to soaring prices for grain, wars, disease, and at best a return to subsistence survival for most people.

While Boserup's analysis was a bold break with much of the received wisdom of the time, it still contained elements of the linearity present in the works of writers like Marx and Rostow and implicitly posited a view of development as progress from family production to modernity characterized by labor specialization and monetization of the economy. Statements that Boserup made at times in her writing demonstrate that her views were more complex; she noted, for example, that reversals can occur such as the fall of the Roman Empire (1999). The complexity, however, showed up more consistently in her multidisciplinary, dynamic approach rather than in addressing issues of non-linearity and variety in development processes. Although there are differences in current views about what motivates technological change, the conceptualization of technology as endogenous has been a significant contribution to development theory. It is the basis of much of the new growth theory of the 1980s and 1990s, but new growth theorists are more likely to explain the technological adaptation in terms of human capital and positive externalities rather than population pressure.

Her book on agricultural growth gave Boserup the credentials in the academic community necessary to stir interest in her next major work, *Woman's Role in Economic Development* (1970). The timing was also good since activism among women was growing and interest in changing public policy was increasing among men and women. Boserup brought gender into development analysis as an integral factor that has remained central to much subsequent development work. She pointed out that both colonialism and modern development programs had frequently marginalized women from rights that they previously had.

She brought in her ideas from the agricultural growth book to identify shifting agriculture with low population density, tools such as the hoe, common property, and women taking an active role in subsistence farming. This system, common in Africa, she identified as a female farming system. She contrasted this with the system often found in more densely-populated Asia of plowed fields, irrigated crops, and private property, which she termed a male farming system. She noted that the gender roles in economic activity were distinct in all the societies she studied and although considered 'natural' by members of each group, they varied across cultures—illustrating that the gender roles were constructed socially rather than biological. Despite her awareness of these variations in occupation and time use by gender, Boserup focused on constructing a general view and downplayed the variation among agricultural systems. She stated that the agricultural system would tend to progress from the female farming, shifting system to the male farming, plowed system as population density increased. Continuing growth of population would lead to more intensive farming of irrigated land by both women and men (a mixed system) (1970). Much of the

research that had focused on women and gender at that time had been done by anthropologists who frequently use in-depth case studies as a methodological approach. As a trained economist, Boserup used the approach of collecting large amounts of data on the roles of women and men from many countries (mainly in sub-Saharan Africa and South/Southeast Asia) and making inferences—thus, the 268 tables in her book.

2. *Continuing Influence of Boserup's Pioneering Effort*

Woman's Role in Economic Development has been cited often and has generated debate and controversy that has deepened our understanding of human development. The publication of the book in 1970 marks the beginning of the multidisciplinary WID field that brought together researchers, practitioners, and policy-makers/advocates initially to work to integrate women into the development planning process (see Tinker 1990). Boserup's work clearly demonstrated that she recognized that women had long been actively participating in the actual development process (Jaquette 2001).

The WID/GAD field has evolved gradually to focus extensively on relative gender analysis of how women and men fare in the costs and opportunities associated with economic transformation policies and the process of globalization and how differences in age, class, religion, and location influence the results. Significant critiques of *Woman's Role* were published by Huntington (1975) and Beneria and Sen (1981). Huntington challenged Boserup's classification of female and male farming systems and suggested that women in Africa had been in subordinate relations with men, more like peasants and aristocrats, long before the colonial period. Beneria and Sen acknowledged Boserup's contributions but critique the lack of clear theoretical framework, the dominance of the capitalist model of development in her work, and the lack of feminist analysis of women's subordination and their reproductive roles (1981). Boserup's publications also motivated the *Persistent Inequalities* volume (Tinker 1990) that remains a leading contribution to the WID/GAD field. Debate continues both over the rhetoric of what to call this evolving field and appropriate theories and policies that recognize diversity and give voice to the people affected throughout the world.

Boserup's pioneering publications have contributed significantly to our understanding of how economic policies and processes affect people's well-being and agency—all people, not just the men of the society. While best known for her work on gender and agricultural systems, Boserup also wrote about the 'bazaar and service occupations' (1970) now called the informal sector that has gained much attention with economic restructuring policies since the 1980s. In addition, she wrote on rural–urban migration, polygamy, property rights, and urban employment. Her work is global and multidisciplinary in perspective, addressing the links between culture, environment, technology, population, employment, and family. She framed the argument for women's rights in terms of development that was more acceptable to the mainstream at the time (1999). She brought efficiency as well as equity into the argument for women's rights and a just society, but she did not depict women as victims (Jaquette 1990). Boserup used her publications to make women's work visible.

See also: Agriculture, Economics of; Economic Development and Women; Economic Globalization and Gender; Gender and Environment; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender and Technology; Gender, Economics of; Gender History; Household Production; Land Rights and Gender; Poverty and Gender in Developing Nations; Rural Industrialization in Developing Nations and Gender; Rural Sociology

Bibliography

Beneria L, Sen G 1981 Accumulation, reproduction, and women's role in economic development: Boserup revisited. *Signs* **7**(2): 279–98

Boserup E 1965/1993 *The Conditions of Agricultural Growth: The Economics of Agriculture under Population Pressure*. G. Allen and Unwin, London

Boserup E 1970 *Woman's Role in Economic Development*. St. Martin's Press, New York

Boserup E 1996 Development theory: An analytical framework and selected applications. *Population and Development Review* **22**(3): 505–15

Boserup E 1999 *My Professional Life and Publications 1929–1998*. Museum Tusculanum Press, University of Copenhagen, Denmark

Huntington S 1975 Issues in woman's role in economic development: Critique and Alternatives. *Journal of Marriage and the Family* **37**(4): 1001–11

Jaquette J 1990 Gender and justice in economic development. In: Tinker I (ed.) *Persistent Inequalities: Women and World Development*. Oxford University Press, New York

Jaquette J 2001 From academia to the WID Office: Crossing the line. In: Tinker I, Fraser A (eds.) *Women Affecting International Development: The Personal and Political*. Feminist Press, New York

Tinker I (ed.) 1990 *Persistent Inequalities: Women and World Development*. Oxford University Press, New York

G. Summerfield

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Boundaries and New Organization Forms

Organizational boundaries, the demarcation between an organizational entity and its external environment, underwent considerable remaking in the last decades of the twentieth century. Long thought to be relatively stable, boundaries were once only redrawn either to increase efficiency or gain more leverage over external factors. But organizational boundaries have become much more fluid and permeable as organizations experiment with a variety of new organizational forms to access critical knowledge, skills, and resources.

New systems of organizing do not arrive on the scene ready made and announce their availability. The historian David Hounshell (1984) illustrated how the mass production model emerged piecemeal in the USA in the latter half of the nineteenth century, beginning with the use of interchangeable parts in rifles made at the armories. Subsequently, the manufacture of sewing machines and then bicycles, and later meatpacking and beer brewing all played a critical role in the eventual development of the assembly line by Henry Ford. Similarly, new forms of organizing have recently emerged incrementally, in fits and starts, and are now visible in a variety of guises. Several notable lines of research have stressed that the boundaries of organizations are shaped largely by considerations of efficiency and power. For much of the second half of the twentieth century, arguments drawn from economics and sociology stressed that organizational boundaries sharply demarcated an organization from its external environment. Indeed, many organizations operated like medieval castles, walled off and protected from external influences. When boundaries were redrawn, considerations of scale and scope loomed large as decisions to shift boundaries were based on weighing the costs of internalizing a transaction vs. conducting it in the market (Coase 1937, Williamson 1985, Chandler 1990). In another line of work, emphasizing the ability of an organization to maintain autonomy and control over its environment, boundaries were recast to reduce uncertainty, to reduce dependency on external parties (Pfeffer and Salancik 1978), and to reshape the production chain in order to enhance strategic position within an industry (Porter 1980). These lines of research have been challenged in recent years by a wide proliferation of alternative governance arrangements and new types of organizational forms.

Organizations now seem to worry more about their adaptability to a changing landscape than their adaptive fit to a specific environment (see Powell and Smith-Doerr 1994, Stark 1999). These new forms enhance absorptive capacity (Cohen and Levinthal 1990), the capability to access skills and knowledge located outside the organization, and the ability to reconfigure organizational practices and structures to respond to a rapidly changing environment in which internal activities are more and more interdependent with external parties. Considerable effort is underway to explain both the causes and consequences of this transformation.

The unraveling of the old system of bureaucratic employment in the large firm is now widely recognized, but developing a set of metrics to assess the spread of new forms of work and organization is a challenge. The developments underway in various capitalist economies are multilevel in nature, involving a transformation in the ordering of work at the point of production, a profound change in the boundaries of many organizations, and a remaking of relations with competitors. These developments are discontinuous because there is no clear stopping point in the process and no road back to the previous system (Helper et al. 2000). Performance is replacing seniority as the condition of employment. Learning and speed are replacing quantity as the criteria for evaluating organizations. These shifts bring new actors and identities and new business models to the fore, and push aside incumbents.

1. *New Forms of Workplace Organization*

Jobs emerged in the late nineteenth and early twentieth centuries as a way to package work in settings where the same task was done repeatedly. But work today is changing into short-term projects often performed by teams. Consequently, the future organization of work is likely to be much less frequently honeycombed into a pattern of highly specified jobs. Work is now more commonly organized around a team or work group charged with responsibility for a project. Sabel (1994) terms this process of joint exploration 'learning by monitoring.' The activities of work teams are coordinated by a process of iterated goal setting. General projects, such as the design of a new car, are initially determined by study of best practices and prospects for competing alternatives. Then broad plans are in turn decomposed into tasks for work groups. The goals are subsequently modified as work teams gain experience in executing the required tasks. Through these revisions, changes in the parts lead to modifications in the conception of the whole, and vice versa. The same procedure of monitoring decentralized learning, moreover, allows each party to observe the performance of the other collaborators closely enough to determine whether continued reliance on them, and dedication of resources to the joint projects, are warranted (Sabel 1994).

This form of production integrates conception and execution, with design and production running on parallel tracks. With concurrent design and development, participants constantly evaluate one another's work. If project groups decide who supplies their inputs, they need not choose the traditional

internal unit but instead may turn to outside suppliers if they provide better value. This reconceptualization of work is designed to reduce and expose fixed costs, to make the expenses of all units dependent on their contribution, and to fuse the knowledge housed in different parts of the organization.

These new arrangements are deeply corrosive of the old system of sequential steps, linear design, and vertical integration that provided worker and manager alike with security. A key consequence of the remaking of the division of labor is that important tasks no longer need be performed inside the organization. This change remakes not only the organization of work, but also the boundaries of organizations.

2. *Flattening of Hierarchies. Spread of Networks*

Just as the changed conception of work as organized around project teams transforms firms internally, the growing involvement of organizations in an intricate latticework of collaborations with 'outsiders' blurs the boundaries of the firm, making it difficult to know where the firm ends and where the market or another firm begins. The former step redraws internal lines of authority, while the latter spreads the core activities of the firm across a much wider array of participants, with an attendant loss of centralized control. Astute observers of these developments, such as Rosenbloom and Spencer (1996), suggest that industrial competition today resembles less a horse race and more a rugby match in which players frequently change uniforms.

Various forms of interorganizational collaboration have grown rapidly in recent years (Gomes-Casseres 1996, Doz and Hamel 1998, Mowery and Nelson 1999). So intensive are these partnering efforts that the interorganizational network is increasingly the relevant unit of production. This form of collaboration does not dampen rivalry, but instead shifts the playing field to sharp competition among rival networks with fluid membership. The growth of alliances and partnerships entails novel forms of complex collaboration with suppliers, end-users, distributors, and even former competitors.

3. *Research and Development*

Recent empirical research shows significant change in the boundaries of organizations, most notably in the area of research and development (R&D). A recent US National Research Council analysis of trends in industrial R&D reports that the innovation process has undergone a significant transformation in the past decade, a change that is both 'substantial' in magnitude and consequential to economic performance (Merrill and Cooper 1999). There are four components of this reorienting of R&D: (a) a shift in the industries and sectors that dominate R&D toward new emerging technologies and nonmanufacturing industries; (b) a change in the time horizons of R&D, with industry focusing more on shorter-term development and relying more on universities for basic research; (c) a change in the organizational structure of R&D, with greater decentralization of research activities and increased reliance on both outsourcing and collaboration among firms, universities, and government laboratories; and (d) changes in the location of R&D, with successful research increasingly dependent on geographic proximity to clusters of innovative organizations. A companion National Research Council survey of 11 industries, purposefully diverse in character and technology but all resurgent in the 1990s, notes that common to each industry are: (a) increased reliance on such external sources of R&D as universities, consortia, and government laboratories; and (b) greater collaboration with domestic and foreign competitors, as well as customers, in the development of new products and processes.

The motives for the upsurge in collaborations are varied. In one form, they are an effort to reshape the contours of production by relying more on subcontractors, substituting outside procurement for in-house production. The subcontractors work under short timeframes, provide considerable variety of designs, spend more on R&D, and deliver higher quality, while the 'lead' firm affords reciprocal access through data-sharing and security through longer-term relationships (Dyer 1996). There is no natural stopping point, however, in this chain of decisions to devolve centralized control. Thus, fixing the boundaries of an organization becomes a nearly impossible task, as relationships with suppliers, subcontractors, and even competitors evolve in unexpected ways. As these network ties proliferate and deepen, it becomes more sensible to exercise voice rather than exit. A mutual orientation between parties may be established, based on knowledge that the parties assume each has about the other and upon which they draw in communication and problem solving. Fixed contracts are thus ineffectual, as expectations, rather than being frozen, change as circumstances dictate. At the core, then, of this form of relational contracting are the entangling strings of reputation, friendship, and interdependence.

Much sophisticated technical knowledge is tacit in character—an indissoluble mix of design, process, and expertise. Such information is not easily transferred by license or purchase. Moreover, passive recipients of new knowledge are less likely to appreciate fully its value or be able to respond rapidly. In research on the commercial field of biotechnology, an industry rife with all manner of interorganizational collaborations, we have argued that learning is closely linked to the conditions under which knowledge is gained (Powell 1996, Powell et al. 1996). Thus, regardless of whether collaboration is driven by calculative motives, such as

filling in missing pieces of the value chain, or by strategic considerations to gain access to new knowledge, network ties become admission tickets to high-velocity races. Connectivity to an interorganizational network and competence at managing collaborations have become key drivers of the new logic of organizing.

The growth of new organizational forms is driven by divergent factors and pursued in a different manner by a wide array of organizations. Larger organizations are making their boundaries more permeable in order to procure key components or critical R&D. Subcontracting and outsourcing are steps taken to reduce fixed overheads. Organizations cooperate with ostensible competitors in order to take on projects too risky or challenging for one entity to pursue alone. Clusters of small organizations collaborate, cohering into a production network to create what no single small entity could on its own. In sum, organizations are coming to resemble a network of treaties because these multistranded relationships encourage learning from a broad array of collaborators and promote experimentation with new methods, while at the same time reducing the cost of expensive commitments. These developments do not mean that competition is rendered moot, instead the success of organizations is linked to the nature and depth of their ties to organizations in diverse fields.

See also: Authority: Delegation; Bureaucracy and Bureaucratization; Management: General; Network Analysis; Organization: Overview; Organizational Decision Making; Organizations: Authority and Power; Organizations, Sociology of

Bibliography

Chandler A D 1990 *Scale and Scope: The Dynamics of Industrial Capitalism*. Harvard University Press, Cambridge, MA

Coase R 1937 The nature of the firm. *Economica* **4**: 386–405

Cohen W, Levinthal D 1990 Absorptive capacity: a new perspective on learning and innovation. *Administrative Science Quarterly* **35**: 128–52

Doz Y L, Hamel G 1998 *Alliance Advantage: The Art of Creating Value Through Partnering*. Harvard Business School Press, Boston

Dyer J 1996 Specialized supplier networks as a source of competitive advantage: Evidence from the auto industry. *Strategic Management Journal* **17**(4): 271–92

Gomes-Casseres B 1996 *The Alliance Revolution: The New Shape of Business Rivalry*. Harvard University Press, Cambridge, MA

Helper S, MacDuffie J P, Sabel C 2000 Pragmatic collaborations: advancing knowledge while controlling opportunism. *Industrial and Corporate Change* **9**(3): 443–88

Hounshell D A 1984 *From the American System to Mass Production 1800–1932: The Development of Manufacturing Technology in the US*. Johns Hopkins University Press, Baltimore, MD

Merrill S A, Cooper R S 1999 Trends in industrial research and development: evidence from national data sources. In: *Securing America's Industrial Strength*. National Research Council Board on Science, Technology and Economic Policy. National Academy Press, Washington, DC, pp. 99–116

Mowery D C, Nelson R R (eds.) 1999 *Sources of Industrial Leadership: Studies of Seven Industries*. Cambridge University Press, New York

Pfeffer J, Salancik G R 1978 *The External Control of Organizations: A Resource Dependence Perspective*. Harper and Row Publishers, New York

Porter M E 1980 *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. Free Press, New York

Powell W W 1996 Inter-organizational collaboration in the biotechnology industry. *Journal of Institutional and Theoretical Economics* **152**: 197–215

Powell W W, Koput K, Smith-Doerr L 1996 Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly* **41**: 116–45

Powell W W, Smith-Doerr L 1994 Networks and economic life. In: Smelser N J, Swedberg R (eds.) *Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ, pp. 368–402

Rosenbloom R S, Spencer W J 1996 The transformation of industrial research. *Issues in Science and Technology* **12**(3): 68–74

Sabel C F 1994 Learning by monitoring: the institutions of economic development. In: Smelser N J, Swedberg R (eds.) *Handbook of Economic Sociology*. Russell Sage Foundation, New York, pp. 137–65

Stark D 1999 Heterarchy: distributing authority and organizing diversity. In: Clippinger J H III (ed.) *The Biology of Business*. Jossey-Bass, San Francisco, pp. 153–80

Williamson O E 1985 *The Economic Institutions of Capitalism*. Collier Macmillan, London

W. W. Powell

Bounded and Costly Rationality

Some kind of model of rational decision making is at the base of most current economic analysis, especially in microeconomics and the economics of organization. Such models typically assume that economic decision makers have sufficient cognitive capacities to solve the problems they face, and have preferences and beliefs that are consistent in a rather strong sense. In particular, it is typically assumed that decision makers (a) do not make logical errors, and can solve any relevant mathematical problems; (b) can process all available information in the time required (including computation, storage, and retrieval); and (c) have an adequate understanding of the decision problems they face, which includes having precise beliefs about the relevant uncertainties and precise preferences among the various consequences of their actions. However, this model has been criticized as inadequate from both normative and descriptive viewpoints. The various strands of this critical movement form the topic known as 'bounded rationality.' This article sketches the

historical roots and some current developments of this movement, distinguishing between attempts to extend the standard models and the need for more radical departures.

Unease with mainstream models of *homo economicus* had already been voiced by J. M. Clark at the beginning of the twentieth century, but the work of Jacob Marschak and Herbert Simon provided the stimuli for a more intense level of activity. The term 'bounded rationality' was coined by Simon: 'Theories that incorporate constraints on the information-processing capabilities of the actor may be called *theories of bounded rationality*' (Simon 1972, p. 162). (For a review of empirical evidence of bounded rationality, see Conlisk 1996.)

The current mainstream theory of rational individual decision making in the face of uncertainty was elaborated by Savage (1954). This will here be called the 'Savage Paradigm,' and will be the main starting point of this article. Extensions of this theory to describe rational strategic behavior in multiperson situations are the subject of the theory of games (see below). The notion of optimizing is central to all these models of rationality.

The general concept of bounded rationality covers the two rather different approaches of Marschak and Simon. Marschak emphasized that the Savage Paradigm could be extended to take account of costs and constraints associated with information acquisition and processing in organizations, without abandoning the notion of optimizing behavior. This approach will here be called 'costly rationality,' and is elaborated in Sect. 4.

Simon was more concerned with behavior that could not so readily be interpreted, if at all, as optimizing. However, in some of his publications he apparently considered bounded rationality to be a broader concept, subsuming costly rationality as a particular case. The narrower concept will here be called 'truly bounded rationality' (Sect. 5).

1. *Uncertainty*

Discussions of rational decision making—unbounded and otherwise—have been closely tied to uncertainty. The very beginnings of formal probability theory were in part stimulated by questions of how to act rationally in playing games of chance with cards and dice. During the first half of the twentieth century a number of alternative views were developed concerning the nature of uncertainty, the possibly of different kinds of uncertainty, and whether, or in what circumstances, it could be measured (Arrow 1951, Savage 1954, Chap. 4). One could be uncertain about natural events, the consequences of action, the laws of nature, or the truth of mathematical propositions. There was general (but not universal) agreement that, if uncertainty could be measured (quantified), then that quantification should obey the mathematical laws of probability. However, the 'frequentist school' reserved the legitimacy of probabilistic reasoning for 'experiments' (planned or naturally occurring) that were repeated indefinitely under identical conditions. The 'personalist school,' which included a diverse set of methodologies, argued that the concept of probability was applicable to events outside the frequentist realm. Some personalists went so far as to deny that the frequentist view could be applied meaningfully to any events at all, i.e., all probability judgments were 'personal.' (See the accounts of Arrow 1951 and Savage 1954 of the work of Ramsey, Keynes, Jeffries, Carnap, and De Finetti. In some sense, the personalist view might also be ascribed to earlier authors, such as Bayes and Laplace.) The personalist view was given a solid foundation by Savage (1954).

Central to the development of thinking about uncertainty was the simple idea that uncertainty about the consequences of an action could (or should) be traced to uncertainty about the 'state of the world' in which the action would be taken. An essential feature of the concept of the 'state of the world' is that it is beyond the control of the decision maker in question. A further clarification was provided by the theory of games (put forward by J. von Neumann and O. Morgenstern in 1944, and further elaborated by J. Nash, J. Harsanyi, R. Selten, and others). In multiperson decision-making situations in which the participants have conflicting goals, this theory distinguishes between two aspects of the state of the world from the point of view of any single decision maker, namely (a) the 'state of Nature,' which is beyond the control of any of the persons involved; and (b) the actions of the other persons, the latter being called 'strategic uncertainty.' (For material on the theory of games, especially noncooperative games, see *Game Theory*; *Game Theory: Noncooperative Games*; *Game Theory and its Relation to Bayesian Theory*)

The remainder of this article concentrates on a critique of the theory of rational behavior as it is applied to single-person decision making, or to multiperson situations in which the persons do not have conflicting goals. (For discussions of bounded rationality in a game-theory context, see Rubinstein 1998 and Radner 1997.)

2. *The Savage Paradigm*

A sketch of the Savage Paradigm is needed here in order to understand the notions of costly and bounded rationality. (For a systematic treatment, see *Utility and Subjective Probability: Contemporary Theories*; *Utility and Subjective Probability: Empirical Studies*.) The essential building blocks of the model are (a) a set of alternative states of the world, or simply states, which are beyond the decision-maker's control; (b) a set of alternative actions available to the decision maker, or as Savage calls them, 'acts'; and (c) a set of alternative consequences. An act determines which

consequence will be realized in each state (of the world). Hence a parsimonious way to think about an act is that it is a function from states to consequences.

The decision maker (DM) is assumed to have preferences among acts. These preferences reflect both the DM's beliefs about the relative likelihood of the different states, and the DM's tastes with regard to consequences. A few axioms about the independence of beliefs from consequences enable one to impute to the DM two scales: (a) a probability measure on the set of states, reflecting the DM's beliefs; and (b) a utility scale on the set of consequences, reflecting the DM's tastes. Using these two scales, one can calculate an expected utility for each act, in the usual way, since an act associates a consequence with each state. Thus, for each state, one calculates the product of its probability times the utility of the associated consequence, and then adds all of the products to obtain the expected utility of the act. One proves that expected utility represents the DM's preferences among acts in the following sense: *the DM prefers one act to another if and only if the first act has a higher expected utility*. (This theorem is sometimes called the expected utility hypothesis.) The rational DM is assumed (or advised) to choose an act that is most preferred among the available ones, i.e., has the highest expected utility; this is the assumption of optimization.

The simplicity of this formulation hides a wealth of possible interpretations and potential complexities. First, the axioms of the theory enable the DM to infer preferences among complicated acts from those among simpler ones. Nevertheless, the required computations may be quite onerous, even with the aid of a computer.

Second, if the decision problem has any dynamic aspects, then states can be quite complex. In fact, a full description of any particular state will typically require a full description of the entire history of those features of the DM's environment that are relevant to the decision problem at hand.

Third, the description of the set of available acts reveals—if only implicitly—the opportunities for the DM to acquire information about the state of the world and react to it. The laws of conditional probability then determine how the DM should learn from observation and experience. In fact, this is what gives the Savage Paradigm its real 'bite.' An act that describes how the DM acquires information and reacts to it dynamically is sometimes called a strategy (plan, policy). In a model of a sequential decision problem, the space of available strategies can, of course, be enormous and complex. This observation will be a dominant motif in what follows.

(The formula for learning from observation is sometimes called 'Bayes's theorem,' after the eighteenth-century author Reverend Thomas Bayes. Hence, the method of inference prescribed by the Savage Paradigm is called 'Bayesian learning.')

3. *The Simon Critique*

As Herbert Simon emphasized in his work, the cognitive activities required by the Savage Paradigm (and its related precursors) are far beyond the capabilities of human decision makers, or even modern human/computer systems, except with regard to the simplest decision problems. This led Simon and his colleagues (especially at Carnegie-Mellon University) to investigate models of human decision making that are more realistic from the point of view of cognitive demands, and yet do not entirely abandon the notion of rationality. This research also had an impact on the emerging field of artificial intelligence (see Simon 1972, 1981, and references therein). Savage himself was aware of the problem of bounded rationality, but he nevertheless felt that his model was a useful one for thinking about rational decision making (Savage 1954, pp. 16, 17).

4. *Costly Rationality and the Extended Savage Paradigm*

As just sketched in the previous section, the Savage Paradigm does not appear to take account explicitly of the costs of decision making. However, nothing prevents the DM from incorporating into the description of the consequences of an act the costs—in terms of resources used—of implementing the corresponding actions. The costly activities involved in decision making include:

observation and experimentation;
information processing, i.e., computation;
memory; and
communication.

The last category may be important when the decision-making process is undertaken by a team of individuals.

If the resources used by these decision-making activities are limited, then those limits may impose binding constraints on the activities themselves—constraints that must be taken into account in the DM's optimization problem. If the constraints are on the rate of resource use per unit time, then more extensive decision-making activities may cause delays in the implementation of the eventual decisions. To the extent that a delay lowers the effectiveness of a decision (e.g., by making it more obsolete), one may think of delay as an 'indirect cost.' Extending the Savage Paradigm to incorporate the costs of decision making may in some cases be natural, and in other cases problematic. The first class of cases is here called 'costly rationality.'

The notion that observation is costly was implicit in the Neyman–Pearson theory of hypothesis testing, and was made explicit by Abraham Wald in his pioneering studies of sequential statistical procedures (see Wald 1950 for an influential codification of his general approach). The cost of observation also figures in

more classical (nonsequential) statistical problems such as the design of sample surveys and agricultural experiments. Given some model of the costs of observation, the DM chooses the kind and amount of observation, optimally balancing the expected benefits of additional observations against their costs. Such decision problems fit naturally into the Savage Paradigm, although taking account of these costs typically complicates the analysis. For example, in the case of clinical trials and similar problems, the calculation of optimal policies quickly becomes computationally intractable for many problems of realistic size.

Even after the information has been collected, it still must be further processed to produce the required decisions. This information-processing task may be quite demanding. Examples include (a) computing a weekly payroll; (b) scheduling many jobs on many machines; (c) managing multiproduct inventories at many locations; and (d) project selection and capital budgeting in a large firm. Such tasks are typically too complex to be handled by a single person, even with the aid of modern computers. In such circumstances the required processing of the information is decentralized among many persons in the organization. The theoretical study of decentralized decision making in an organization whose members have identical goals was introduced by J. Marschak in the theory of teams (Marschak and Radner 1972).

Computer science has provided a number of useful models of information processing by both computers and humans, and the decentralization of information processing in human organizations finds its counterpart in the theories of parallel and distributed processing in computer systems. T. A. Marschak and C. B. McGuire, in 1971, were probably the first to suggest the use of a particular model of a computer (the finite automaton) to represent the limited information-processing capabilities of humans in economic organizations. S. Reiter and K. R. Mount were early contributors to this line of research, and went further in analyzing economic organizations as networks of computers. (For more recent developments, see Radner 1997, Van Zandt 1999, and references therein.) One conclusion from this literature is the iron law of delay for networks of processors of bounded individual capacity. This 'law' can be paraphrased in the following way: as the size of the information-processing task increases, the minimum delay must also increase unboundedly, even for efficient networks, and even if the number of available processors is unlimited (Radner 1997, and references therein).

Memory storage and communication among humans and computers are also resource-using activities, and cause further delays in decision making. Both the storage and transmission of information and the results of information processing seem to be relatively 'cheap' compared with observation and processing, at least if we consider computer-supported activities. The proliferation of large data banks, and the flood of junk mail, telephone calls, and e-mail, lend support to this impression. It appears that today it is much cheaper, in some sense, to send, receive, and store memos and papers than it is to process them. (For game-theoretic models of players with limited memory, see Rubinstein 1998. For models of costly communication in organizations, and some implications for organizational structure, see Marschak and Reichelstein 1998.)

5. *Truly Bounded Rationality*

Many real decision problems present difficulties that prevent the DM from usefully treating them as optimization problems. Among these difficulties are:

inconsistency;
ambiguity;
vagueness;
unawareness; and
failure of logical omniscience.

As will be seen, these difficulties are somewhat related and overlapping. In particular, it is difficult to distinguish in practice between 'ambiguity' and 'vagueness.'

Regarding inconsistency, Savage (1954, p. 57) wrote:

> According to the personalistic view, the role of the mathematical theory of probability is to enable the person using it to detect inconsistencies in his own real or envisaged behavior. It is also understood that, having detected an inconsistency, he will remove it. An inconsistency is typically removable in many different ways, and the theory gives no guidance for choosing.

Some 'inconsistencies' have been observed so frequently, and have been so 'appealing,' that they have been used to criticize the Savage axioms, and to form a basis for a somewhat different set of axioms (e.g., the so-called 'Allais paradox' and 'Ellsberg paradox' (see *Utility and Subjective Probability: Contemporary Theories*; *Measurement Theory: Conjoint*). In other cases, it has been argued that inconsistent preferences arise because the DM is forced to articulate preferences about which they are not 'sure.' (This explanation is related to 'vagueness'; see below.) In particular, this unsureness may be related to uncertainty about what are the states of the world, a circumstance that can lead to a preference for 'flexibility' (see below). Finally, it has been observed in experiments that inconsistencies are more frequent the closer the alternatives are in terms of preference. This observation led J. Marschak and others to the elaboration of models of 'stochastic choice' (see *Decision and Choice: Random Utility Models of Choice and Response Time*).

Allusion has already been made to the DM's possible vagueness about his/her preferences. However, he/she could also be vague about any aspect of his/her

model of the decision problem, and is likely to be so if the problem is at all complex. Vagueness can be about the *interpretation* of a feature of the model, or about its *scope*, or both. Unfortunately, there has been little if any formal theorizing about problems of vagueness.

At a moment of time, the DM may be *unaware* of some aspect of the problem: for example, he/she may be unaware of some future contingencies that could arise, or of some actions that are available to him/her. This phenomenon is particularly interesting if the DM is aware of the possibility that he/she may be unaware of something. For example, if the DM anticipates that in the future he/she will become aware of new acts of which he/she is currently unaware, then he/she may prefer present actions that allow for 'flexibility' of choice in the future. (This idea was formalized by T. C. Koopmans, D. Kreps, and others: Kreps 1992. For other theoretical treatments of unforeseen contingencies, see Dekel et al. 1998.)

As a consequence of the preceding considerations, decision theorists recognize that it is impossible for a DM to construct a complete model of his/her 'grand decision problem,' i.e., for their whole life! A common research strategy is to suppose that the DM can break up the grand decision problem into subproblems that can be solved independently without (much) loss of utility. Savage called this the device of constructing 'small worlds,' but showed that the conditions for this to be done without loss are unrealistically stringent (Savage 1954, pp. 82–91).

Finally, I come to what is perhaps the most difficult aspect of truly bounded rationality. Up to this point it has been assumed—if only implicitly—that the DM has no difficulty performing mathematical calculations or other logical operations. In particular, having formulated a decision model, he/she will be able to infer what it implies for their optimal strategy. As has already been pointed out, this assumption is absurd, even for small-world models, except for 'Mickey Mouse' problems that are constructed for textbooks and academic articles. The crux of the matter is that, in any even semirealistic decision problem, *the DM does not know all of the relevant logical implications of what he or she knows*. This phenomenon is sometimes called 'the failure of logical omniscience.' Examples of the failure of logical omniscience are:

(a) A DM who knows the axioms of arithmetic is uncertain about whether they imply that 'the 123rd digit in the decimal expansion of pi is 3,' unless he/she have a long time to do the calculation and/or has a powerful computer with the appropriate software.

(b) Twenty years ago, a DM who knew the axioms of arithmetic was still uncertain about whether they imply Fermat's last theorem.

The following examples are closer to practical life, and possibly more intimidating:

(a) Given all that a DM knows about the old and new drugs for treating a particular disease, what is the optimal policy for conducting clinical trials on the new ones?

(b) Given all that is known, theoretically and empirically, about business organizations in general, and about telecommunications and AT&T in particular, should AT&T reorganize itself internally, and if so, how?

Savage (1954, p. 7fn) commented on this class of problems:

> The assumption that a person's behavior is logical is, of course, far from vacuous. In particular, such a person cannot be uncertain about decidable mathematical propositions. This suggests, at least to me, that the tempting program sketched by Polya of establishing a theory of the probability of mathematical conjectures cannot be fully successful in that it cannot lead to a truly formal theory …

(For further discussion and references, see Savage 1972.)

In spite of some interesting efforts (see, for example, Lipman 1995), it does not appear that there has been significant progress on what it means to be rational in the face of this kind of uncertainty.

6. *Satisficing, Heuristics, and Non-Bayesian Learning*

In view of the difficulties posed by the various manifestations of 'truly bounded rationality,' a number of authors have proposed and studied behavior that departs more or less radically from the Savage Paradigm. These will be discussed under three headings: satisficing, heuristics, and non-Bayesian learning.

The term 'satisficing' refers to behavior in which the DM searches for an act that yields a 'satisfactory,' as distinct from an optimal, level of expected utility. The target, or 'satisfactory,' level of expected utility is usually called the DM's 'aspiration level.' In the simplest model, the aspiration level is exogenous, i.e., a given parameter of the model. More ambitious models describe some process whereby the aspiration level is determined within the model, and may change with experience (Simon 1972, Radner 1975). Such aspiration levels are called 'endogenous.' In some problems even optimal behavior bears a resemblance to satisficing. One category is the 'secretary problem' (Radner 2000).

The term 'heuristics' refers generally to behavior that follows certain rules that appear to produce 'good' or 'satisfactory' results most of the time in some class of problems (Simon 1972, see *Heuristics for Decision and Choice*). For example, the calculation of an optimal schedule for assigning jobs to machines is typically intractable if the numbers of jobs and machines are even moderately large. Nevertheless, human schedulers routinely construct 'satisfactory' schedules with such numbers, using various rules of

thumb that have been developed with experience. Heuristics are central to many artificial intelligence applications. Satisficing plays an important role in many heuristic methods, and also in the processes of their modification.

The discussion of heuristics leads naturally to the consideration of non-Bayesian learning (NBL). Bayesian learning (i.e., the application of the calculus of conditional probability) is of course part of the Savage Paradigm in any decision problem in which the DM conditions his/her action on information about the state of the world. Many standard statistical methods use NBL. For example, the use of the sample mean to estimate a population mean is typically inconsistent with the Savage Paradigm (although in some cases the latter can be shown to be a limit of Bayesian estimates, as some parameter of the problem goes to infinity). Most psychological theories of learning postulate some form of NBL. A central question in the theory of NBL is: under what conditions, if any, does a particular NBL procedure converge asymptotically to a procedure that is Savage-Paradigm optimal as the DM's experience increases? (Rustichini 1999).

Again, one must ask: is there any satisfactory meaning to the term 'rationality' when used in the phrase 'bounded rationality'? The convergence of NBL to optimal actions could provide one (weak) meaning. Nevertheless, the problems raised by the various phenomena grouped under 'truly bounded rationality' may eventually lead students of decision making to answer this last question in the negative.

Bibliographic Notes

Many important works on bounded rationality have been omitted from the bibliography because of space limitations. The references cited in the body of the entry have historical interest, provide an overview of a topic discussed, and/or provide a key to other literature. The following provide additional references and information on the application of notions of bounded rationality in economics and management: Arrow 1974, Radner 2000, Shapira 1997, Van Zandt 1999.

See also: Bounded Rationality; Decision and Choice: Paradoxes of Choice; Decision Research: Behavioral; Game Theory; Heuristics for Decision and Choice; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Rational Choice Explanation: Philosophical Aspects

Bibliography

Arrow K J 1951 Alternative approaches to the theory of choice in risk-taking situations. *Econometrica* **19**: 404–37

Arrow K J 1974 *The Limits of Organization*. Norton, New York

Conlisk J 1996 Why bounded rationality. *Journal of Economic Literature* **34**: 669–700

Dekel E, Lipman B, Rustichini A 1998 Recent developments in modeling unforeseen contingencies. *European Economic Review* **42**: 523–42

Kreps D 1992 Static choice in the presence of unforeseen contingencies. In: Dasgupta P, Gale D, Hart O, Maskin E (eds.) *Economic Analysis of Markets and Games*. MIT Press, Cambridge, MA, pp. 258–81

Lipman B L 1995 Decision theory without logical omniscience: toward an axiomatic framework for bounded rationality. Unpublished thesis, Queens University, Canada

Marschak J, Radner R 1972 *Economic Theory of Teams*. Cowles Foundation, New Haven, CT

Marschak T A, Reichelstein S 1998 Network mechanisms, informational efficiency, and hierarchies. *Journal of Economic Theory* **79**: 106–41

Radner R 1975 Satisficing. *Journal of Mathematical Economics* **2**: 253–62

Radner R 1997 Bounded rationality, in determinacy, and the managerial theory of the firm. In: Shapira Z (ed.) *Organizational Decision Making*. Cambridge University Press, Cambridge, UK

Radner R 2000 Costly and bounded rationality in individual and team decision-making. *Industrial and Corporate Change* **9**: 623–58

Rubinstein A 1998 *Modeling Bounded Rationality*. MIT Press, Cambridge, MA

Rustichini A 1999 Optimal properties of stimulus-response learning models. *Games and Economic Behavior* **29**: 244–73

Savage L J 1954 *The Foundations of Statistics*. Wiley, New York. Dover, New York

Savage L J 1972 *The Foundations of Statistics*, 2nd edn. Dover, New York

Shapira Z (ed.) 1997 *Organizational Decision Making*. Cambridge University Press, Cambridge, UK

Simon H A 1972 Theories of bounded rationality. In: McGuire C B, Radner R (eds.) *Decision and Organization*. North-Holland, Amsterdam, pp. 161–76

Simon H A 1981 *The Sciences of the Artificial*. 2nd edn. MIT Press, Cambridge, MA

Van Zandt T 1998 Organizations with an endogenous number of information processing agents. In: Majumdar M K (ed.) *Organizations with Incomplete Information*. Cambridge University Press, Cambridge, UK, pp. 239–305

Van Zandt T 1999 Real-time decentralized information processing as a model of organizations with boundedly rational agents. *The Review of Economic Studies* **66**: 633–58

Wald A 1950 *Statistical Decision Functions*. Wiley, New York

R. Radner

Bounded Rationality

1. Introduction

The central ideas of bounded rationality (BR) are straightforward. First, humans are cognitively constrained in various ways, e.g., we can consciously attend to only one choice problem at a time. Second, these

mental properties have behavioral consequences: they significantly affect decision-making. Third, the harder the problem the more likely it is that a decision-maker's information-processing constraints will matter (Simon 1981).

The first claim is uncontroversial. Cognitive psychologists have demonstrated, for example, that our short-term or 'working' memory holds only a few (between about five and nine) chunks of information (Miller 1956). The second claim follows close on the heels of the first. Since our working memory is small, our conscious attention must be selective. Hence, e.g., an agency head can work on only one problem at a time, no matter how many his subordinates present to him.

The third claim is important for demarcating domains in which BR theories have empirical 'bite', i.e., their predictions differ from those of other theories—especially those of BR's main competitor, rational choice (RC) theories. Consider Simon's formulation: 'the capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behavior in the real world—or even for a reasonable approximation to such objective rationality' (1957, p. 198). Thus, bounded rationality describes a relation between a person's mental abilities and the complexity of the problem he or she faces. It is not a claim about the brilliance or stupidity of human beings, independent of their task environments. It is easy to miss this central point and to reify the idea of bounded rationality into an assertion about our absolute capacities (e.g.,'people are dumb'). The fundamental notion is of cognitive limits, and as is true of any constraint, if cognitive constraints do not bind in a particular situation, then they do not matter: they will not affect the outcome. And whether they bind depends partly on the demands placed on decision-makers by the problem at hand. Simon has called the joint effects of 'the structure of task environments and the computational capabilities of the actor … a scissors [with] two blades' (1990, p. 7). Theories of bounded rationality have cutting power—especially relative to RC theories—only when both blades operate.

Since this is the program's foundation, it is worthwhile exploring its implications via an example. Consider a group of normal adults, randomly paired up to play either chess or tic-tac-toe. A pair can either play their assigned game or they can stipulate a particular outcome, take their money for participating in the experiment and leave. The intuitive prediction is that more people assigned to tic-tac-toe would stipulate an outcome (a draw): after all, that game is trivial.

However, so—in a sense—is chess: Zermelo proved long ago that there exists an optimal way to play chess, and if played optimally then, as in tic-tac-toe, the same outcome always occurs: either white wins or black does or they draw. Indeed, for classical game theory, which ignores cognitive constraints, chess and tic-tac-toe belong to the same class: they are zero-sum, finite games of perfect information. Hence they are game theoretically equivalent. In the real world, of course, these games are not equivalent. Normal adults do not play tic-tac-toe; it is pointless. But they do play chess; indeed, for some it is a career. The point is simple but vital: our mental abilities are a binding constraint in chess but not in tic-tac-toe. Accordingly, BR and RC theories make observationally equivalent predictions about the latter but not the former. Hence BR theories have cutting power in chess—knowing the players' cognitive capacities gives us predictive leverage—but not in tic-tac-toe.

The example provides a more subtle point. Chess simplifies towards a game's end, but the players' cognitive capacities remain the same. Since the task's demands are falling while mental resources stay fixed, at some point *those resources may no longer be a binding constraint*. If that happens, the players will play optimally, and so a standard RC theory will accurately predict their behavior from that point on: a chess master burdened by a big disadvantage in the endgame will resign, for he knows it is hopeless. Once the players can identify the optimal strategies, the game becomes completely predictable, hence continuing is as pointless as playing tic-tac-toe.

This example reveals the subtlety of the contest between BR and RC theories. Theories from these two research programs predict different behaviors in chess only when the game is sufficiently complex so as to make players' mental capacities binding constraints. For expert players, this condition holds early in the game but not at the end. Thus, the two types of theories make observationally equivalent predictions about the endgame behavior of experts.

The lesson is that the theoretical significance of bounded rationality turns on the *difference* between cognitive resources and task demands, not on the former's absolute level.

2. *The Main Branches in Political Science of BR*

Bounded rationality should not be confused with a theory (e.g., of satisficing), much less with a specific formal model (e.g., Simon 1957). It is best considered a research program: a sequence of theories with overlapping sets of assumptions, aimed at solving similar problems. Political scientists have often conflated BR with the specific theory of satisficing, but that conflation produces a serious underestimation of the program's substantive content. In principle, the program's empirical domain is vast—it is as imperialistic as the rational choice program—and so its set of possible theories is also very large.

Many research programs contain multiple branches. So it is for BR: political science alone exhibits two main orientations toward bounded rationality. The first orientation sees the glass as half full, emphasizing how people manage to do 'reasonably well' *despite* cognitive limitations, *even in complex tasks*. Here belong Simon's line of work and also Lindblom's theory of incrementalism (Lindblom 1959, Braybrooke and Lindblom 1963, Bendor 1995). In the second branch, the glass is half empty: it emphasizes how people make mistakes *even in simple tasks*. Here is the Tversky–Kahneman (T–K) research tradition on heuristics and biases. (For a superb overview see Dawes 1998.)

2.1 The Glass Is Half Full

As noted above, one of Simon's key premises is that a decision maker's 'inner environment' (Simon 1981) of information processing will become manifest only when the task is sufficiently difficult. This implies that in order to discover the important types of mental limits, one should study people facing hard problems, such as chess.

But analyzing chess revealed more than mental limits: it generated new questions for the Simonian line, especially about performance. People's performance in chess varies tremendously, and this variation demanded an explanation. Thus, a new research focus emerged: how do some decision-makers *do so well*, given that they face similar cognitive constraints? Thus, scholars in this branch came to see the glass as half full: on some hard problems some humans perform very well, relative to an empirically sensible aspiration level. Given that objectively optimal play in chess-like problems is clearly impossible, what is interesting is not whether people behave fully rationally—we know they don't—but how do relatively competent agents work around their cognitive limits. (See, e.g., Simon 1990, and the associated references.)

2.2 The Glass Is Half Empty

In sharp contrast to the Simonian program's focus on 'good' performance in difficult domains, the prototypical experiment of the T-K tradition shows that even highly trained subjects answering simple questions can perform suboptimally. Kahneman and Tversky's goal—clearly stated in their seminal paper (Tversky and Kahneman 1974, p. 1124)—was to map subtle mental processes that can cause cognitive illusions, similar to the study of perceptual illusions. This project's results are more striking if they are shown to occur even in simple tasks, just as perceptual illusions are more striking if demonstrated with simple stimuli. The results are still more striking if one shows that even experts are vulnerable to illusions.

Subsequently, however, the field has tended to emphasize more the errors themselves, rather than the underlying cognitive mechanisms: 'The … approach triggered an explosion of research on inferential error, and the list of illusions, foibles, flops, and bloopers to which ordinary people were apparently prone became rather long … soon, if there was a mistake to be made, someone was making it in the presence of a social psychologist' (Gilbert 1998). (After reading many of these studies, the glass probably does look half empty—or worse.) This focus on mistakes triggered a scholarly backlash which has tried to show that humans are better decision makers than the T–K program apparently claims (e.g., Gigerenzer 1991; for a reply see Kahneman and Tversky 1996). Unfortunately, this debate has clouded Tversky and Kahneman's original intention, which was not to show that *Homo sapiens* are dolts but rather to uncover fundamental cognitive mechanisms that leave their imprint nearly everywhere. Given the primary objective of demonstrating that certain mental processes are ubiquitous, T–K's approach made sense. But it had the unfortunate side effect of making the program largely inattentive to variations in performance. With a few exceptions (e.g., the study of how people calibrate subjective probabilities), specialists in the T–K branch have shown little interest in performance variation; in particular, genuinely expert performance, at the level of chess masters, has been neglected. Instead, performance evaluations in this branch mostly use dichotomous theoretical standards: do people make choices in accordance with the axioms of expected utility theory? Do they revise beliefs in a Bayesian manner? The answers—generally no, they don't—are less informative than would be answers based on a quantitative scale measuring degrees of sophistication, even if the empirical best was far short of the theoretical ideal.

However, these differences in how the branches' study performance reveal that in one respect the T–K research agenda is more ambitious: it is harder demonstrating that cognitive limits show through on simple problems than on hard ones. Thus, the T–K program has worked on *pushing out the boundaries of BR* by showing that even quite subtle problem representations ('framing') can induce suboptimal performance. (Establishing which problem-representations are transparent—i.e., cognitively obvious—and which are opaque has been important throughout the T–K program (Kahneman and Tversky 1996).) One needn't go all the way to chess to uncover our mental limitations: humans are sufficiently sensitive (hence vulnerable) to framing so that judgmental or decisional imperfections appear even when a task exhibits no combinatorial explosion. Thus, a main finding of the T–K program is that BR has predictive bite in a larger domain than once thought.

3. Essential Properties of Humans as Information-processors

What are our essential properties as information-processors? Synthesizing Hogarth (1987) and Simon (1990) gives us the following list.

(a) Selective perception of information: our objective environments always contain far more information than we can perceive or attend to. Thus, perceptions must be guided by anticipations and other forms of top-down processing (e.g., schemas).

(b) High order information-processing, especially conscious thinking and attention, is largely serial. This has significant implications for the real-time behavior of busy officials.

(c) Compared with modern computers, humans process many kinds of information slowly, partly due to physiological limits (neurons are much slower than electronic circuits).

(d) Compared with computers, people are poor at calculation.

(e) Memory is actively reconstructive, not photographic. It therefore lacks photographic fidelity.

(f) Although there is no known limit to long-term memory, working memory is very small (Miller 1956). Thus, because everything that enters the former goes through the latter, short-term memory is a key bottleneck in information-processing.

The other crucial type of cognitive property are ways in which humans work around the above constraints. Simon calls these mechanisms for procedural rationality (1981, 1990). Three are especially worth noting.

(a) Recognition processes. After long experience, experts have stored a great many patterns about their tasks in long-term memory, both about situations ('this is situation type x') and actions ('in situation x, do y'). For familiar problems a fundamental trick is to *substitute recognition for search*.

(b) Heuristic search. When recognition fails, experienced decision-makers fall back on search. For reasonably complex problems search is heuristic: optimality is not guaranteed. Instead, heuristics make complicated problems manageable: they cut them down to (human) size and, in the hands of experts, often yield 'good' solutions. There are two types of heuristics. If the task is highly and recognizably structured, experts will use powerful task-specific heuristics. When these are unavailable they use general (but weak) heuristics such as satisficing (stop searching for alternatives when you find one that satisfices—exceeds your aspiration level—rather than requiring an optimal solution). Note the place of satisficing here. Far from being the heart of BR, it is merely one of several general-but-weak heuristics.

(c) Heuristic search occurs in *problem spaces*: mental representations of the task at hand. Experts learn to use economical and sometimes powerful problem representations.

4. Applications: Theories of Politics

4.1 Budgeting

Building on Lindblom, Wildavsky and his colleagues constructed an incremental model of budgeting (Davis et al. 1966) which received substantial empirical support. Crecine (1969) argued that the general heuristic of decomposition—when faced with a hard problem, decompose it into smaller, easier subproblems (Simon 1981)—is central to budgeting, and developed a model of city budgeting based on that idea. Padgett (1980) created a model of federal budgeting that continued to use the BR notion of heuristic search but which implied that allocations will sometimes be nonincremental.

4.2 Organizational Reliability

Landau (1969) has shown that certain types of structural redundancy can make organizations more reliable than any of their subunits. (Consider several R&D teams working in parallel.) Relatedly, Ladha (1992) and others have shown how using Condorcet's jury theorem can reduce both type 1 errors (e.g., the Food and Drug Administration's approving bad drugs) and type 2 errors (rejecting good drugs). Significantly, these results hold even if individual decision-makers face quite severe cognitive constraints.

4.3 The Evolution of Cooperation

Axelrod (1981) and Bendor and Swistak (1997) have used evolutionary game theory to analyze how cooperation might emerge and stabilize when central authority cannot solve Hobbes' problem. (Evolutionary game theory assumes that people learn to use strategies that work and to discard those that fail; agents need not optimize.)

4.4 Subgovernments

Bendor and Moe (1985) constructed a model of subgovernments with myopic actors (legislators, an agency and interest groups). The simulation outcomes were typically 'pluralist equilibria,' characterized by a balance of power between the interest groups.

4.5 Electoral Competition

In Kollman et al. (1992), adaptively rational political parties compete for the support of voters in two-party elections. In unidimensional policy spaces the parties converge to the median voter (Pages, personal com-

munication, September 1999); this result appears to hold for a wide variety of search heuristics. Hence Downs' famous result can be derived with weak cognitive assumptions. In a multidimensional policy space, the parties behave 'sensibly' but can get hung up on local optima. This suboptimization depends crucially on nonobvious aspects of voters' preferences.

4.6 Prospect Theory

The key premise of prospect theory, Tversky and Kahneman's most important theoretical contribution, is that choices are evaluated relative to a reference point, e.g., the status quo. The second assumption is that people are risk-averse about gains (relative to the reference point) but risk-seeking about losses. The third premise is loss-aversion: losing x hurts more than gaining x helps. This simple and elegant structure has yielded suprisingly powerful predictions in the study of bargaining (Neale and Bazerman 1991), international politics (Levy 1997), and voting (Quattrone and Tversky 1988).

For an excellent and accessible review of the BR program in economics see Conlisk (1996). Rubinstein (1998) is very instructive but much more technical.

See also: Bounded and Costly Rationality; Decision Research: Behavioral; Decision Theory: Classical; Psychology and Economics; Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Rational Choice in Politics; Rational Choice Theory: Cultural Concerns; Rational Choice Theory in Sociology; Rational Theory of Cognition in Psychology; Rationalism; Rationality in Society

Bibliography

Axelrod R 1981 The emergence of cooperation among egoists. *American Political Science Review* **75**: 306–318

Bendor J 1995 A model of muddling through. *American Political Science Review* **89**: 819–40

Bendor J, Moe T 1985 An adaptive model of bureaucratic politics. *American Political Science Review* **79**: 755–74

Bendor J, Swistak P 1997 The evolutionary stability of cooperation. *American Political Science Review* **91**: 290–307

Braybrooke D, Lindblom C 1963 *A Strategy of Decision*. Free Press, New York

Conlisk J 1996 Why bounded rationality? *Journal of Economic Literature* **34**: 669–700

Crecine J 1969 *Governmental Problem-Solving*. Rand McNally, Chicago, IL

Davis O, Dempster M, Wildavsky A 1966 A theory of the budgetary process. *American Political Science Review* **60**: 529–47

Dawes R 1998 Behavioral decision making and judgment. In: Gilbert D et al. (eds.) *The Handbook of Social Psychology*. McGraw-Hill, Boston, MA, Vol. I pp. 497–548

Gigerenzer G 1991 How to make cognitive illusions disappear: beyond heuristics and biases. *European Review of Social Psychology* **2**: 83–115

Gilbert D et al. (eds) 1998 Ordinary personology. In: *The Handbook of Social Psychology*. McGraw-Hill, Boston, MA, Vol. II, pp. 89–150

Hogarth R 1987 *Judgement and Choice*, 2nd edn. Wiley, New York

Kahneman D, Tversky A 1996 On the reality of cognitive illusions: a reply to Gigerenzer's critique. *Psychological Review* **103**: 582–91

Kollman K, Miller J, Page S 1992 Adaptive parties in spatial elections. *American Political Science Review* **86**: 929–38

Ladha K 1992 The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science* **36**: 617–34

Landau M 1969 Redundancy, rationality, and the problem of duplication and overlap. *Public Administration Review* **29**: 346–58

Levy J 1997 Prospect theory, rational choice, and international relations. *International Studies Quarterly* **41**: 87–112

Lindblom C 1959 The science of 'muddling through'. *Public Administration Review* **19**: 79–88

Miller G 1956 The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* **63**: 81–97

Neale M, Bazerman M 1991 *Cognition and Rationality in Negotiations*. Free Press, New York

Padgett J 1980 Bounded rationality in budgetary research. *American Political Science Review* **74**: 354–372

Quattrone G, Tversky A 1988 Contrasting rational and psychological analyses of political choice. *American Political Science Review* **82**: 719–36

Rubinstein A 1998 *Modeling Bounded Rationality*. MIT Press, Cambridge, MA

Simon H 1957 *Models of Man*. Wiley, New York

Simon H 1981 *The Sciences of the Artificial*, 2nd edn. MIT Press, Cambridge, MA

Simon H 1990 Invariants of human behavior. In: Rosenzweig M, Porter L (eds.) *Annual Review of Psychology*. Annual Reviews, Palo Alto, CA, Vol. 41, pp. 1–19

Tversky A, Kahneman D 1974 Judgment under uncertainty: heuristics and biases. *Science* **185**: 1124–31

J. Bendor

Bourgeoisie/Middle Classes, History of

1. Introduction

The history of the bourgeoisie and the middle classes concerns the development and the transformation of social formations that are made up of those occupying intermediary social positions. These social formations and their members take on mediating or leading functions in society, economics, politics, and culture. The historiography of the bourgeoisie and middle classes takes as its point of departure the development of the modern bourgeoisie out of the older urban

bourgeoisie (burghers) of Europe. It concentrates on Western and Central Europe and on the USA, the history of which has, from the eighteenth through the twentieth centuries, been shaped to a large degree by the bourgeoisie and the middle classes (Kocka 1995a, 1995b, Zunz et al. 2001). It asks how the functions and meaning of the 'bourgeoisie' and the 'middle classes' have changed in these core areas, how, since the late nineteenth century, the conception has diffused to the periphery of Europe and worldwide, and how, in the process, it has transformed and run up against certain limits. Finally, it analyzes the attempts to eliminate or transform the bourgeoisie in socialist countries and the reconstruction of modern middle classes in post-socialist societies (Hoepken and Sundhaussen 1998).

Historical research on the middle classes investigates symbolic and social structures of long duration, historical continuities and ruptures, and regional and national similarities and differences. Of special interest are the particular characteristics of different formations and actors in various historical contexts. Research topics include the social, economic, professional, political, cultural, and legal characteristics and structures of the middle strata, whether in hierarchically differentiated societies, under conditions of social and cultural inequality, or in processes of class formation. Historians are also concerned with symbolically mediated practices and processes, in which individuals and groups come together in historically identifiable middle classes such as the traditional 'middle estates,' the exclusive 'modern bourgeoisie,' the economic and political 'bourgeoisie,' and the 'broad and inclusive middle classes.' Common to both historical and systematic concepts of the bourgeoisie and the middle classes, however diverse they are otherwise, is the preoccupation with the central region of the society in question, that is with that which takes shape between the extremities. The fundamental assumption is that society is formed and integrated not 'from above' (aristocracy, upper class, oligarchy), or 'from below' (working class, the propertyless, the uneducated), or 'from outside' (foreign domination).

2. *Definitions*

The concepts 'bourgeoisie,' 'middle class,' 'middle classes,' and many similar terms applying to particular eras and cultures, have more or less in common, depending on the historical period, the society in question, and the historical situation. A typical characteristic of the historical scholarship on these topics is the tension between exclusive and inclusive concepts of the bourgeoisie and the middle classes.

Definitions vary with regard to the composition of the bourgeoisie, the criteria for inclusion and exclusion, the social and symbolic processes of integration and delimitation, and with regard to the functions and aspirations of the middle classes in society and history. In the fields of social and economic history, scholars may employ classifications that are descriptive, historically and culturally specific, and statistical; or they may, for heuristic purposes, adopt abstract definitions, which are derived from particular theories of the bourgeoisie or from general theories of society. In either case, the focus is on the correspondence between social position and interests. In contrast, cultural historians and culturally oriented social historians understand the bourgeoisie and the middle classes primarily in terms of images of self and other, strategies, performances, discourses, rituals, institutions, sociocultural practices, habitus, and mentalities. In the history of ideas and of philosophy, as well as in theories of the state, law, and constitutions, scholars usually rely on normative and systematized conceptions of the bourgeoisie and middle classes. Social historians also draw on *Begriffsgeschichte*, the history of ideas, and the history of mentalities in order to show how, under what conditions, and why the concepts 'middle class,' 'bourgeoisie,' and 'middle classes' have been employed for the self-definition of individuals and groups, or for the purposes of establishing boundaries between self and other (Kocka 1995a, p. 12, Koselleck et al. 1991). Of special interest, in this regard, is how such classificatory activity affects consciousness and social action. A central element in all definitions and research agendas is the question of exclusion and inclusion.

2.1 *Exclusive Formations: Middle Class and Bourgeoisie*

Exclusive formations of the middle classes include wealthy property owners, entrepreneurs, those with high salaries, the highly educated and highly qualified, as well as administrative, political, and cultural elites. These groups are characterized by similar economic and political situations. They are also linked through social networks and have similar cultural orientations and life-styles. Historical research has been directed toward eliciting the meaning of the concepts and categories just cited in concrete historical situations, and discovering the historically and regionally specific terms that allow them to be grouped together in a single formation. In England and the USA, this social phenomenon has, since the eighteenth and nineteenth centuries, been referred to as the 'upper middle class' or 'middle class' (singular). In francophone Europe, the relevant terms are *notables*, *bourgeoisie*, or *classe moyenne* (singular). The German-speaking peoples of Europe have employed the terms *Bürgerstand*, *Mitt*elstand, or Bürgertum. The corresponding labels in Italy are *borghesia* or *ceti medi*, and in Spain *burguesía* (Kocka 1995a, Zunz et al. 2001). In the nineteenth century, the bourgeoisie in this narrow sense made up approximately 5 percent of the population, although this percentage was somewhat higher in cities.

In contemporary scholarship, these formations are understood in terms of Max Weber's sociology of classes and status groups, and are referred to using the historical-systematic concept of the 'modern bourgeoisie' or 'middle class' (Kocka 1995a, 1995c, Zunz et al. 2001). Understood in this way, the bourgeoisie is a heterogeneous formation, which includes economically independent and dependent classes (both entrepreneurs and civil servants, both professionals and employees), the members of which have similar values and styles of life. The concept of the 'modern bourgeoisie' differs from the Marxian concept of the 'economic bourgeoisie.' The latter is defined in terms of the ownership of capital, access to the means of production, kinds of revenue, and strategies of capital deployment and commercialization. The economic bourgeoisie, which is the counterpart of the working class, is the central category of political-economic development theory. Historical scholarship that has been influenced by Karl Marx and Marxian theory distinguishes among the precapitalist bourgeoisie (speculative capital, merchant capital, usury) those land owners and agriculturalists who have a dynamic relationship to property and capital, the petite bourgeoisie of artisans, the bureaucratic bourgeoisie, the modern commercial and industrial bourgeoisie, and the monopolist bourgeoisie (Radandt et al. 1981). Since the late nineteenth century, the Marxian conception of the bourgeoisie has, in many variations, become established in scholarship and in public discussion. From the late nineteenth century and up until the 1980s, Marxian theory was also very influential in non-Marxian sociology and social theory as well.

More recently, the sociological and economic conceptions of the middle classes and the bourgeoisie have been challenged by cultural historians, who insist upon the central importance of phenomena which, in classical Marxian theory, are part of the superstructure. From a cultural historical perspective, membership in the middle classes depends primarily on particular ideas, values, attitudes, and sociocultural practices. Emphasis falls on the symbolic mediation of that which is bourgeois or middle class through discourses, images, institutions, and places of socialization and sociability such as the family, the school, the public sphere, the theater, literature, voluntary associations, coffee houses, festivals, architecture, and monuments. Research of this type leads to further important insights into the bourgeoisie. If, however, bourgeois culture is not linked back to the social groups that foster it, this kind of approach may become merely a socially diffuse cultural history of modernity.

2.2 Inclusive Middle Classes

The history of the exclusive middle class has been shaped by the need to establish distinct boundaries with reference to traditional elites, the working class, and the 'inclusive' or 'broad' middle classes. The 'inclusive middle classes' may seem to be less homogeneous than the exclusive bourgeoisie in terms of profession, status, education, income, style of life, attitudes, and values. Nevertheless, from the viewpoint of both historical actors and historians, they have often been thought to belong together. Their unity is often characterized with qualifying expressions such as 'finely graduated' or 'leveled.'

The factors that unite the inclusive middle classes vary from one land, historical period, and social situation to another. These common factors may include professional and economic interests, similar legal claims, traditions and mentalities, political viewpoints and voting behavior, similar hopes and fears with regard to upward or downward mobility, or attitudes with regard to style of life, which may in turn be oriented toward those of the exclusive bourgeoisie. The social spectrum of the inclusive middle classes extends from mid-level entrepreneurs to independent professionals, mid-level and low-level civil servants and employees, mid-level farmers and tenants, and even to workers with higher salaries and qualifications. When social, professional, and economic criteria are used to assess the size of this group in the Western and Central European countries of the nineteenth century, it may be estimated to have made up 15–30 percent of the total population. When other criteria are employed, this percentage may rise or sink.

In the USA and in France, the inclusive middle class is referred to as 'middle classes' or *classes moyennes* (plural), respectively. The corresponding term in Germany is *Mittelstand*, and in Switzerland *Mittelstand* or *Bürgertum* in the broadest sense of the term. In the German-speaking countries, the term *Volk* was also used in this sense, especially in the socially and politically most inclusive societies of the early twentieth century (Kocka 1995a, Tanner 1995, Savage 2001, Charle 2001). In the second half of the twentieth century, the concept of the inclusive middle classes has been extended even further. As it gains in influence and in integrative potential, the older concept of the exclusive middle class seems to be falling increasingly into disuse.

It is common for both historical actors and historians to distinguish between historically older and newer groups of the broad middle classes. Master artisans and small businessmen are referred to as the *alter Mittelstand* (old middle estate) or the 'old middle class' (Haupt and Crossick 1998). From the early nineteenth century and up to the 1970s, they have often been viewed as a historical relict and as potential victims of modernization. Employees i.e., members of the strata of white-collar workers, which has been expanding since the late nineteenth century and which continues to become increasingly differentiated, are referred to as the 'new middle class.' They are often thought to have developed a distinctive class con-

sciousness at a relatively late date, due especially to their intermediate position between the bourgeoisie and the working class. In different societies and in different historical periods, employees may be understood either as agents or as victims of modernization (Mills 1956, Kocka 1977).

3. *History of the Bourgeoisie and the Middle Classes*

Recent scholarship on the eighteenth and nineteenth centuries has concentrated on the conditions and forms of the genesis of the modern exclusive bourgeoisie. The historiographical literature on the twentieth century focuses especially on the transformation, the crisis, and the declining significance of the exclusive bourgeoisie. With reference to the development of the modern bourgeoisie in Central and Northwestern Europe, Jürgen Kocka distinguishes between three phases, including an ascendant phase from the eighteenth to the early nineteenth century, a culminating phase from the middle of the nineteenth century to World War I, and a period of differentiation, dissolution, and even renaissance, which lasts into the late twentieth century (Kocka 1995a, 1995c). Attempts to assign the history of the broadly inclusive middle classes to historical periods usually posit a period of decline in the nineteenth century and a new period of ascendance in the twentieth century, which culminates in the inclusive middle-class society. In the following summary of the history of the bourgeoisie and the middle classes, reference is made to periods of 'genesis,' 'expansion,' and 'contraction.'

3.1 *The Genesis of the Modern Bourgeoisie and Middle Classes* (*Eighteenth to Nineteenth Centuries*)

The modern bourgeoisie took shape in the eighteenth and nineteenth centuries, though the precise time varied by region and by nation (Kocka 1995b, Schrader 1996, Gall 1993, Charle 2001). It included the following social groups: (a) the more dynamic part of the older urban bourgeoisie of the cities and communes; (b) those civil servants of the modern territorial state who were not part of the feudal hierarchy; (c) merchants, bankers, industrial entrepreneurs, and those members of the independent artisanry who were able to remove themselves from the order of the corporative estates; (d) professionals, literati, and artists; (e) enlightened or economically innovative portions of the landholding nobility, as well as non-noble estate owners and agriculturalists (in Southern and Western Europe the landowners often lived in cities, which is why they were counted among the bourgeoisie); (f) participants in emancipatory movements that were cultural or confessional, liberal or democratic, or regional or national in character.

Communication, symbolic representation, and social integration among the ranks of the bourgeoisie were promoted through institutional changes, which were implemented partly by the bourgeoisie itself, and partly by the monarchs and the state in the context of reforms and revolutions (freedom of trade, religious freedom, freedom of association, freedom of speech and of the press, political participation and accountability). The bourgeoisie developed its self-understanding and its language in the fields of the economy, culture, politics, and law. Bourgeois praxis developed in the representative organs of the estates, in free associations, in scientific academies, in literary circles and salons, in the print media, and in the reformed institutions of higher education, administration, law, legislation, and government. In these contexts, 'bourgeois' notions of collective and individual autonomy, representation, and the free play of public opinion were formed. In both the workplace and in public debates, new understandings of property and interest, ability and performance, self-cultivation and virtue, individual and society, contract and constitutions, and tradition and progress took shape. New experiences and discourses allowed the emerging middle class to criticize existing conditions, conceive of new social and symbolic orders, and lay claim to leading or mediating positions as well as to central functions and competences in society.

The bourgeoisie improved its chances for gaining favorable positions in a system of material, social, political, and cultural inequality by generalizing its own understandings of law and removing traditional privileges. It distanced itself from the traditional 'upper estates' of the court, the landed nobility, and the urban patriciate by adopting the role of a modern, dynamic, and achievement-oriented cultural and functional elite. It constituted itself in opposition to the nobility and the clergy as a 'third estate,' which represented the general public and was, in principle, open to everyone. The 'bourgeoisie' and the 'middle class' are concepts that imply a program and that serve to mobilize and justify certain forms of social action in the transition from estate society to class society. In principle, they ignore traditional differences of status, profession, and confession.

Typically, however, new forms of differentiation and new tensions arise fairly rapidly within the bourgeoisie, for example between the economic bourgeoisie (*Wirtschaftsbürgertum*) and the educated bourgeoisie (*Bildungsbürgertum*), between professional groups and status groups, between rural and urban groups, along local, regional, and national lines, among different confessions, or between political traditionalists and progressives or liberals and democrats. The integration of the bourgeoisie is a complicated and lengthy process, which usually succeeds only partially. The unity of the bourgeoisie is more a

postulate than a reality (Banti 1996, Siegrist 1996). It is most evident in critical historical situations or in conflicts with others, to whom the bourgeoisie feels itself to be superior, whether morally, economically, politically, or in terms of performance (nobility, manual laborers, the propertyless, the uneducated, the disenfranchised, and strangers or foreigners).

3.2 Contraction and Expansion: Bourgeosie and Middle Classes in the Nineteenth and Twentieth Centuries

In the process of the formation of the modern bourgeoisie, some groups of artisans and merchants, who previously belonged to the inclusive urban bourgeoisie, were economically marginalized and socially downgraded. In contrast, the educated bourgeoisie, merchants, and entrepreneurs belonging to ethnic and confessional minorities, and rural industrialists acquired new rights and gained in status. In England, the concept of the 'middle class' served to mobilize business people, industrialists, and mine owners from the provinces in the context of the electoral reform movement of 1832. It helped to integrate in a single group all of those who had not had the right to vote for the Lower House and who had felt themselves to be neglected by the upper middle class (London wholesalers and financiers, aristocratic landowners). Since the 1830s, in England as on the European continent, the interconnection among political, social, and legal rights possessed by the bourgeoisie had begun to emerge (Koselleck et al. 1991). This bundle of rights was viewed by bourgeois philosophers and social theorists as the very basis of political virtue in a well-ordered society. In the course of the nineteenth century, in Europe and in the USA, perceptions of the 'bourgeois' or 'middle-class' male person became closely linked with the concept of the politically enfranchised 'citizen' (*citoyen*, *Staatsbürger*). In those parts of Central and Eastern Europe in which the political rights of citizens remained limited, property owners and the educated often shared political power at the local level. The towns and cities, in which various property-based or tax-based voting laws remained in effect until World War I, were, in many parts of Europe, the domain of the bourgeoisie. In Central Europe, the concept of the bourgeoisie lost much of its political content after the failure of the 'bourgeois revolutions' of 1848/9. Under these conditions, emphasis shifted to the cultural and aesthetic connotations of the concept. Throughout Europe, the bourgeoisie experienced significant gains in processes of 'inner' nation-building and cultural nationalization in the old nation-states, and through nationalist movements and the formation of new nation-states in Northern, Southern, and Eastern Europe.

Since the late nineteenth century, the bourgeoisie has continually renewed itself by incorporating members of new and expanded professional, functional, or status groups such as managers, engineers, planning and transportation experts, journalists, and high-level employees. These groups proliferated under conditions of intensified industrialization, commercialization, and urbanization or with the concentration of capital, the expansion of the role of science in society, increasing bureaucratization, and the rise of new communication media. The exclusive bourgeoisie allied itself increasingly, though not unreservedly, with social groups whose middle-class status was dubious, but who offered themselves as supporters of the established order. Tenants, agriculturalists with medium-sized holdings, and independent artisans and merchants were offered greater protection and drawn into the bourgeois alliance. Employees gained social and legal privileges over workers and were, thus, prevented from cooperating more closely with the socialist and trade-union organizations of the working class (Kocka 1977).

At no time has the bourgeoisie been unchallenged. In the nineteenth century, the critique of the bourgeoisie was formulated from aristocratic, corporativist, Catholic, anti-Semitic, and socialist points of view. Since the late nineteenth century, the traditional, exclusive conception of the bourgeoisie has come increasingly under attack from members of social movements that arose within the bourgeoisie and the middle classes themselves. The leading concepts of the bourgeois progressives, social reformers, life-style reformers, and democrats are science, achievement, rationality, political equality, and social justice. Progressive and democratic critiques of the bourgeoisie contrast, however, with those of the authoritarian and conservative-elitist movements for renewal of the early twentieth century. The latter often attracted members of the bourgeoisie and petite bourgeoisie, even though they cultivated a specifically 'antibourgeois' discourse, in which the established and increasingly pragmatic bourgeoisie was accused of decadence and the inability to lead. In the twentieth century, antibourgeois sentiment formed a broad spectrum that ranged between conservative-corporativist and fascist programs, on the one hand, and communist ideology, on the other.

The concept of the exclusive bourgeoisie and antibourgeois discourses were supplemented in the twentieth century by conceptions of a broad and leveled-off middle class, which is typical of advanced industrial society. In the history of Europe and the USA, this perspective drew upon older ideas, such as those of social democratic, social-liberal, Christian-social, egalitarian, and communitarian traditions. After World War II, the USA made the concept of the inclusive middle class into a guiding principle for the reconstruction of the defeated states of Europe and Japan (Zunz et al. 2001). This concept unites socially, culturally, politically, and economically heterogeneous groups in a common vision of political and social citizenship in a democratic system and a market

economy. During the era of the conflict between the East and the West, this vision was also motivated by a broadly based anticommunism i.e., by a rejection of a model of society, economy, and government that explicitly excludes the middle and claims to constitute society from below, under the leadership of the party of the working class.

In the noncommunist countries of Europe, in the first decades after 1945, those who wanted to conserve or renew the exclusive bourgeoisie of the past mounted an attack against the concept of the leveled and encompassing middle-class society. The exclusive bourgeoisie had lost prestige because of the Great Depression and due to the role that it had played in the fascist states. Now, however, its representatives mourned the waning and the demise of the bourgeoisie. Up until 1970 they warned of the decline of the bourgeoisie and the resulting threat posed to the nation and to Europe by 'mass culture,' 'consumer society,' 'Americanization,' 'materialism,' the 'welfare state,' the 'excesses' of democracy, the 'weakening of bourgeois values,' the 'loss of individuality,' and the 'leveled middle-class society' (Siegrist 1994, Savage 2001). The debate over the bourgeoisie experienced a new surge around 1968, which had, however, other causes, as the 'establishment,' on one hand, and students, intellectuals, socialists, and communists, on the others, battled one another in the language of anti-Marxism and Marxism. Since the 1970s, the discussion about the bourgeoisie has been superseded by debates over 'elites' and over traditional cultural, ethnic, and gender-based differences and discriminations within and outside the middle classes.

4. *Middle Classes, Functional Equivalents, and the Limits of the Concept*

The development of the middle classes in Eastern and Southeastern Europe, Latin America, Japan, and in the colonies and postcolonial societies of Africa and Asia differs considerably, both temporally and qualitatively, from that of core regions in the history of the middles classes. The concepts have their limits.

In many parts of the world in the nineteenth century, the bourgeoisie was hardly present. The social groups and the corresponding discourses, mentalities, and institutions were largely lacking. When concepts are imported and extended to phenomena that are more or less similar to the bourgeoisie, their meaning changes significantly (Kocka 1995a, 1995c). Marxian or economic and social historical studies, which, on the basis of theoretical premises and European concepts, identify bourgeois and middle-class groups in these areas, rarely do justice to their materials.

Recent approaches in the cross-national and cross-cultural comparative history of the middle classes focus on cultural and historical differences, and also on the specific meaning of similarities. Comparisons may show, for example, that institutional, economic, and technical innovations, which in Western and Central European history are associated with the bourgeoisie, are planned and carried out by other actors and groups in those areas without a bourgeoisie. In such cases, it is not very helpful to apply the concept of the bourgeoisie to such groups. Modernizing landowners in estate societies, civil servants in authoritarian states without a bourgeoisie, and rural and urban oligarchies and military leaders in preindustrial societies may be understood, at best, as 'functional equivalents' of the bourgeoisie. Latin American estate owners and the mercantile elite of large cities and ports may, in the course of the nineteenth century, have come to approximate the economic bourgeoisie. In societies that are sharply differentiated by status group and ethnicity, however, they form a political-social 'oligarchy,' rather than a 'modern bourgeoisie.' With the growth in the twentieth century of new agrarian, mercantile, bureaucratic, and industrial middle strata, the middle classes in such societies have expanded, and the traditional oligarchies have been forced to include them, at least to a degree, into the political elite (Florescano 1985). Even then, however, the situation differs significantly from that in Europe.

In Eastern Europe in the nineteenth century, progressive groups of nobles, members of ethnic and confessional minorities, and Western-educated intellectuals and civil servants assimilated to the pattern of the bourgeoisie of Western and Central Europe. The formation of the bourgeoisie did not, however, proceed beyond the initial stages. In Southeastern Europe, modern middle classes began to emerge in the late nineteenth century, when, following the demise of the Ottoman empire, the segmented and not extremely hierarchical society changed into a national, centralized, and bureaucratic one. The national functional and cultural elites made the state into a resource and became a specific type of bourgeoisie (Höpken and Sundhaussen 1998). In the communist societies of Eastern Europe, both genuinely and ostensibly bourgeois and middle-class groups were marginalized in the context of the communist program of the classless society under the leadership of the working class. Those who remained were, along with the new functionaries, redefined as members of the functional middle strata. The leading middle stratum was termed 'nomenclatura,' whereas the others were known as the 'service class' and the 'socialist intelligentsia.' Given this historical background, the problems involved in forming postsocialist middle classes are quite significant.

In twentieth-century Japan, where, previously, a historical bourgeoisie had hardly existed, there was a dramatic development of the inclusive middle classes. In the early twentieth century, the word *sarariiman* (salary-man), which at first meant scribe, office employee, and civil servant, became a catch-all term for

members of the qualified, salaried middle classes (Gordon 2001). Since the 1960s, the concept of *churyu* has become popular. This means 'mainstream' and includes all those who consider themselves to be modern, average, and normal with regard to attitudes toward work, family, consumption, and life-style. They make up the statistical and social 'middle mass' (Kelly 2001), with which, in public opinion polls, 90 percent of the population aligns itself. The broad conception of the middle classes in Japan has its roots both in national ideas of community and in the inclusive view of the middle class that has been imported from the USA in the postwar period.

5. *Historiography and Emphasis in Current Research*

The exclusive bourgeoisie and (to a lesser degree) the broad middle classes have, from the very beginning, employed historical writing as a means of self-presentation, legitimation, community building, and self-reflection. History serves the bourgeoisie as a means of locating itself in space and time, and orienting itself in a modern, dynamic society, which it seeks constantly to transform and to stabilize. The historiography of the middle classes is characterized, first, by teleological notions of progress and modernization; second, by a narrative of crisis and decadence and of danger and dissolution, which draws upon older religious and aristocratic motifs but is subsequently transformed and takes on a life of its own; and, third, through the motifs of conservation and stabilization. Central elements of this history are elevated by the bourgeoisie and middle classes to the status of norms and, thereby, dehistoricized. They then take on the form of myths or theories of law, society, and development, which are insensitive to variations in era and context. The bourgeoisie displays a tendency toward idealizing, sacralizing, and naturalizing its own past. For this reason, its historiography has often been the target of historians of various ideological or social provenance, who are critical of or opposed to the bourgeoisie.

Until the 1970s, the history of the bourgeoisie and the middle classes was, by and large, written from the perspective of either proponents or opponents. Since the 1980s, historical research on the bourgeoisie has, in some important senses, become more impartial, more independent, more scientific, and more innovative. Traditional versions of history at local and national levels have been relativized and revised with reference to interdisciplinary, international, and comparative approaches in historical research (Kocka 1995b, 1995c, Siegrist 1996). As a result of a stronger focus on the dimensions of 'gender' and 'religion' (Hall and Davidoff 1987) as well as 'ethnicity,' new perspectives have been achieved with regard to the periodization and evaluation of historical materials. In Europe and, to a somewhat lesser degree, the USA, the bourgeoisie and the middle classes rank among the best researched topics in historical scholarship.

See also: Business History; Capitalism; Elites: Sociological Aspects; Social Stratification; Industrialization, Typologies and History of; Marx, Karl (1818–89); Weber, Max (1864–1920); Western European Studies: Gender and History; Western European Studies: Society

Bibliography

Banti A M 1996 *Storia della borghesia italiana. L'età liberale*. Donizelli, Rome

Charle C 2001 The middle class in France: Social and political functions of semantic pluralism 1870–2000. In: Zunz O et al. (eds.) *Social Contracts Under Stress*. Russell Sage, New York

Davidoff L, Hall C 1987 *Family Fortunes: Men and Women of the English Middle Classes 1780–1850*. Routledge, London

Florescano E (ed.) 1985 *Origines y desarollo de la burguesía en America Latina*. D. F., Mexico

Gall L (ed.) 1993 *Stadt und Bürgertum im Übergang von der traditionalen zur modernen Gesellschaft*. Oldenbourg, Munich, Germany

Gordon A 2001 The short happy life of the Japanese middle class. In: Zunz O et al. (eds.) *Social Contracts Under Stress*. Russell Sage, New York

Haupt H-G, Crossick G 1998 *Die Kleinbürger. Eine europäische Geschichte des 19. Jahrhunderts*. C. H. Beck, Munich, Germany

Hoepken W, Sundhaussen H (eds.) 1998 *Eliten in Südosteuropa. Rolle, Kontinuitäten, Brüche in Geschichte und Gegenwart*. Südosteuropa-Gesellschaft, Munich, Germany

Kelly W W 2001 At the limits of new middle class Japan: Beyond mainstream consciousness. In: Zunz O et al. (eds.) *Social Contracts Under Stress*. Russell Sage, New York

Kocka J 1995a Das europäische Muster und der deutsche Fall. In: Kocka J (ed.) *Bürgertum im* 19. *Jahrhundert*. Vandenhoeck and Ruprecht, Göttingen, Germany, Vol. 1

Kocka J (ed.) 1995b *Bürgertum im* 19. *Jahrhundert*. Vandenhoeck and Ruprecht, Göttingen, Germany

Kocka J 1995c The middle classes in Europe. *The Journal of Modern History*. **67**: 783–806

Kocka J 1977 *Angestellte zwischen Faschismus und Demokratie. Zur politischen Sozialgeschichte der Angestellten. USA 1890–1940 im internationalen Vergleich*. Vandenhoeck and Ruprecht, Göttingen, Germany

Koselleck R, Spree U, Steinmetz W 1991 Drei bürgerliche Welten? Zur vergleichenden Semantik der bürgerlichen Gesellschaft in Deutschland, England und Frankreich. In: Puhle H-J (ed.) *Bürger in der Gesellschaft der Neuzeit*. Vandenhoeck and Ruprecht, Göttingen, Germany

Mills C W 1951 *White Collar. The American Middle Classes*. Oxford University Press, New York

Radandt H et al. (eds.) 1981 *Handbuch der Wirtschaftsgeschichte*. Deutscher Verlag der Wissenschaften, Berlin, pp. 692–700, 856–61, 968–76

Savage M 2001 Professional dominance and middle class culture in Britain. In: Zunz O et al. (eds.) *Social Contracts Under Stress*. Russell Sage, New York

Schrader F 1996 *Die Formierung der bürgerlichen Gesellschaft 1550–1850*. Fischer, Frankfurt, Germany

Siegrist H 1994 Ende der Bürgerlichkeit? Die Kategorien 'Bürgertum' und 'Bürgerlichkeit' in der westdeutschen Gesellschaft und Geschichtswissenschaft der Nachkriegsperiode. *Geschichte und Gesellschaft*. **20**: 549–83

Siegrist H 1996 *Advokat, Bürger und Staat. Sozialgeschichte der Rechtsanwälte in Deutschland, Italien und der Schweiz (18.–20.Jahrhundert)*. Klostermann, Frankfurt, Germany

Tanner A 1995 *Arbeitsame Patrioten—wohlanständige Damen: Bürgertum und Bürgerlichkeit in der Schweiz 1830–1914*. Orell-Füssli, Zurich, Switzerland

Zunz O et al. (eds.) 2001 *Social Contracts Under Stress*. Russell Sage, New York

H. Siegrist

Bowlby, John (1907–90)

Edward John Mostyn Bowlby was born in London on 26 February, 1907, to Mary Bridget Mostyn and Anthony Alfred Bowlby. They had six children, who were usually referred to in groups of two: 'the girls' (Winnie and Marion), 'the boys' (Tony and John), and 'the babies' (Jim and Evelyn). Only 13 months apart, 'the boys' had a close relationship, including intense rivalry. But in addition to keeping his end up with a slightly older brother, John was caring towards a younger brother with chronic problems due to a thyroid deficiency, discovered too late to be cured. John was born at a time when it was customary for an upper-class mother to hand over her children to a nanny and nursemaids. The nanny in charge practiced a firm, regular, and disciplined routine. A particular nursemaid took care of John on a daily basis. He loved her, and years later she said that she was always fondest of John. However, she left the family when John was about four. The effect of this on John is unknown, as he never discussed it (van Dijken 1998). The Bowlby children saw their mother for a short time each day, and their father on Sundays. Their father, who was knighted in 1911 as a result of being a surgeon to royalty, was absent even more during World War I. However, there were long family holidays in the country to the New Forest in July and to Scotland in late summer. This remained a family tradition with John, his wife Ursula, their four children, and later their grandchildren retreating to the Isle of Skye each summer. John did much of his writing there, as well as walking and enjoying the 'season,' including the Skye Ball with its Scottish reels. Skye is where John died on September 2, 1990, and where he is buried, on a remote hillside overlooking the cliffs and sea.

When John was 11, he and Tony were sent off to boarding school. From there, John went to the Royal Naval College, Dartmouth. But at 17, he decided that the Navy was not for him, writing to his mother that he wanted a job which would 'improve the community as a whole' (van Dijken 1998, p. 46). At Trinity College, Cambridge, John followed his father by studying medicine, with a departure into psychology in his final year, graduating in 1928. Rather than going straight into clinical school, John spent a year teaching in two boarding schools, including one for disturbed children. Their early disrupted childhoods impressed Bowlby, and he decided to combine his medical training with psychoanalytic training. After becoming medically qualified at University College Hospital in 1933, Bowlby went to the Maudsley to train in Adult Psychiatry, and then to the London Child Guidance Clinic in 1936. He became an Army psychiatrist in 1940, and after the war he went to the Tavistock Clinic, where he remained (Holmes 1993).

At the Institute of Psycho-Analysis, where he had been accepted when only 22, Bowlby was assigned to Joan Riviere and then to Melanie Klein, who supervised him after his analytical qualification in 1937. Differences grew concerning the importance of real-life experiences over unconscious fantasies. For example, when treating a hyperactive and anxious three-year-old boy, Bowlby noted that the boy's mother had been admitted to a mental hospital and was having a another breakdown. Klein, however, did not allow Bowlby to treat the mother, devaluing what was actually happening in the child's life. Yet Bowlby was keenly interested in what happened, thanks to his love of nature, his scientific training, and his clinical observations.

1. *The Theme of Separation*

Bowlby was particularly interested in what happened around separation. This interest may have sprung from his own childhood experiences, as well as later when working with maladjusted children. 'I was struck by the high incidence of severely disrupted relationships with mother-figures in the early histories of children and adolescents who had been referred to the clinic on account of repeated and apparently incorrigible stealing.' He continued, 'First-hand observations of young children admitted to residential nursery or hospital and not visited by parents revealed the intense and prolonged distress they suffer. Moreover, visits home showed that the child's relationship with the mother is seriously disturbed for weeks or longer after reunion. These observations lent considerable support to the hypothesis that the experience of separation and loss of mother-figure in the early years could have adverse effects on development, at least in some cases. That raised a question: if the disruption of a child's relationship with mother-figure in the early years creates much distress and anxiety, what is so special about the relationship that has been disrupted?'

(1991b, pp. 302–3). The prevailing answer was that bond formation stemmed from the association of mother with the provision of food, thereby satisfying a primary need. But in Bowlby's view, this 'cupboard love theory' (p. 303) was insufficient.

2. *An Ethological Perspective*

During the 1950s Bowlby ran a weekly workshop in a dingy room in the old Tavi, in Beaumont Street. The group was entirely heterogeneous: a Freudian and a Kleinian analyst, a Hullian and a Skinnerian learning theorist, a Piagetian, an ethologist, sometimes an antipsychiatrist, psychiatric social workers, having in common only an interest in parent–child relations. John Bowlby's intellectual dynamism and judicious enthusiasm held the group together. During that period Bowlby formulated his ideas on 'the child's tie to his mother' and 'separation anxiety' with inputs from ethology and systems theory' (Hinde 1991, p. 216).

With an ethological perspective, the occurrence of species-characteristic behavior suggests that such behavior may have been selected for during the course of evolution. Bowlby applied this thinking to attachment behavior defined as any form of behavior which attains or maintains proximity to a caregiver in times of need or stress. He argued that individuals who exhibited attachment behavior would have been more apt to survive and leave offspring who in turn would reproduce, compared with those who did not. Obviously selection for attachment behavior could not have happened without a similar pressure on its complement, caregiving behavior. 'During the course of time, the biologically given strategy of attachment in the young has evolved in parallel with the complementary parental strategy of responsive caregiving, the one presumes the other' (Bowlby 1991c, p. 293).

Bowlby further suggested that attachment behavior reflects a distinct motivational system in its own right, not dependent on prior association with any other motivational system such as hunger, and usually subserving a particular biological function. Bowlby postulated a function of protection from harm, by keeping a child in touch with one or more caregivers. With this approach, activation of a fear behavior system would lead to activation of the attachment behavior system. Attachment behavior would lead to proximity to the caregiver, which would in turn deactivate the fear system, enabling activation of an exploratory or social system (e.g., Bowlby 1969/1982).

Bowlby realized that an evolutionary argument could provide insight into behavior which otherwise appeared abnormal. 'It is against this picture of man's environment of evolutionary adaptedness that the environmentally stable behavioural equipment of man is considered. …not a single feature of a species' morphology, physiology, or behaviour can be understood or even discussed intelligently except in relation to that species' environment of evolutionary adaptedness' (Bowlby 1969/1982, p. 64). Bowlby used this reasoning to understand as 'natural' the occurrence of 'irrational fears' as well as the occurrence of attachment behavior throughout the life span. In considering the fears of childhood such as fear of strangeness, separation, or darkness, Bowlby disagreed with influential analysts such as Klein or Freud who had labeled them as irrational. Instead he argued that our tendency to fear what had been dangers in the environments in which we evolved is 'to be regarded as a natural disposition of man…that stays with him in some degree from infancy to old age…Thus it is not the presence of this tendency in childhood or later life that is pathological, pathology is indicated either when the tendency is apparently absent or when fear is aroused with unusual readiness and intensity' (Bowlby 1973, p. 84).

Similarly, attachment behavior in times of stress should not be viewed as a sign of weakness, even beyond childhood. Long before research began on the role of attachment in adulthood, Bowlby wrote 'such tendencies…are present not only during childhood but throughout the whole span of life. Approached in this way, fear of being separated unwillingly from an attachment figure at any phase of the life-cycle ceases to be a puzzle, and instead, becomes classifiable as an instinctive response to one of the naturally occurring clues to an increased risk of danger' (Bowlby 1973, p. 86).

In one of his final contributions, Bowlby summed up thus, 'Once we postulate the presence within the organism of an attachment behavioural system regarded as the product of evolution and having protection as its biological function, many of the puzzles that have perplexed students of human relationships are found to be soluble. …an urge to keep proximity or accessibility to someone seen as stronger or wiser, and who if responsive is deeply loved, comes to be recognised as an integral part of human nature and as having a vital role to play in life. Not only does its effective operation bring with it a strong feeling of security and contentment, but its temporary or long-term frustration causes acute or chronic anxiety and discontent. When seen in this light, the urge to keep proximity is to be respected, valued, and nurtured as making for potential strength, instead of being looked down upon, as so often hitherto, as a sign of inherent weakness' (Bowlby 1991c, p. 293).

3. *Development of an Attachment Bond*

Thus, an attachment behavior system, organized as a distinct motivational system, gives rise to attachment behavior when activated by stress, particularly threats

or perceived threats of separation. An infant is born with 'a marked bias to respond in special ways to several kinds of stimuli that commonly emanate from a human being' (Bowlby 1969/1982, p. 265). As attachment behavior develops, it forms the basis for an inferred attachment bond. Bowlby described particular phases in its development: pre-attachment (from birth to about 2 months), involving signaling without discriminating one person from another; attachment-in-the-making (2–6 months) where signals become directed to particular persons; clear-cut attachment (0.5–4 years) with locomotion and goal-corrected behavior; and finally a goal-corrected partnership (4 years onwards) with perspective taking, communication skills, and sharing mutual plans. Although additional attachments may develop throughout life, these early attachments endure.

What is not common to all individuals is the quality of attachment, which tends to be consistent throughout life, and is reflected in the three basic patterns: avoidant, secure, and ambivalent. Although Bowlby has been mistakenly quoted as stating that such patterns are indelibly 'fixed' over the first year, he in fact stressed that any continuity must depend on a continuous flow of interactions. 'It is evident that the particular pattern taken by any one child's attachment behaviour turns partly on the initial biases that infant and mother each bring to their partnership and partly on the way that each affects the other during the course of it' (Bowlby 1969/1982, p. 340). Bowlby continues, 'By the time the first birthday is reached both mother and infant have commonly made so many adjustments in response to one another that the resulting pattern of interaction has already become highly characteristic' (Bowlby 1969/1982, p. 348). In addition to the immediate characteristics of the infant, the behavior of the caregiver is determined by 'the particular sequence of environments, from infancy onwards, within which development takes place' (Bowlby 1969/1982, p. 378). Thus, while a consistent cycle of interactions tends to maintain a given quality of attachment, if the cycle were to change, then so would the pattern of attachment.

As to which pattern of attachment is desirable, Bowlby was concerned with what might be called 'psychological desiderata.' However, other perspectives may prevail, such as 'biological desiderata' for increasing inclusive fitness, or 'cultural desiderata.' Although influencing each other, the three desiderata, biological, psychological, and cultural, may differ, especially in modern industrialized societies (Hinde and Stevenson-Hinde 1991). Nevertheless, it was 'psychological well-being' which Bowlby as a practicing psychiatrist regarded as basic. Making an analogy with 'physical well-being,' Bowlby argued that 'psychological well-being' had an absolute meaning, with security of attachment an essential ingredient. Security in turn depends on antecedent interactions with a 'sensitively responsive' mother, as found in Mary Ainsworth's pioneering Baltimore study (Ainsworth et al. 1978) and subsequently in many others (DeWolff and van IJzendoorn 1997).

4. *Assessing the Quality of Attachment*

Bowlby's collaboration with Mary Ainsworth (1913–99) remains a model for creativity in science. Their friendship spanned 40 years, from 1950 when Ainsworth joined Bowlby at the Tavistock. After her longitudinal study of mother–infant behavior in Uganda, Ainsworth joined the faculty at the Johns Hopkins University in 1955. Appreciating that assessments of attachment must involve how a child uses mother as a 'secure base' when the attachment behavior system is activated, Ainsworth developed a Strange Situation procedure for infants (12–18 months old: Ainsworth et al. 1978). This was modified for older children, with her guidance (2.5–4.5 years: the Cassidy and Marvin system; 6 years: the Main and Cassidy system; reviewed by Solomon and George, in Cassidy and Shaver 1999). Ainsworth's procedure paved the way for empirical studies to explore the framework provided by Bowlby.

The Strange Situation procedure involves a series of episodes in an unfamiliar laboratory setting, including two separations and two reunions from an attachment figure. The coding emphasis is on the reunion episodes, which show how the child organizes behavior to mother following the stress of separation. A Secure pattern involves the child greeting mother with full gaze and positive affect. Interactions are calm, while also intimate and indicative of a special relationship. The two main insecure patterns may be contrasted with this, and indeed with each other. Whereas the Avoidantly attached child shows minimal responses and maintains neutrality, the Ambivalently attached child emphasizes dependence on mother, with angry/whiny resistance and/or immature behavior. A Disorganized pattern reflects confusion and apprehension. Beyond infancy, one may see a transition from Disorganization to a Controlling pattern, reflecting an effort to reduce uncertainty by taking charge.

The move beyond behavioral observations to the level of representation was largely inspired by one of Mary Ainsworth's former Ph.D. students, Mary Main. Once children reach Bowlby's 'goal-corrected partnership' stage, their verbal behavior may be used to index representations of attachment (reviewed by Solomon and George, in Cassidy and Shaver 1999). Mary Main and colleagues developed the Adult Attachment Interview (AAI), based around 18 questions designed to access one's current state of mind with respect to attachment. Correspondence between parents' AAI classifications and their infants' Strange Situation classifications has now been reported in over 18 studies (reviewed by Hesse, in Cassidy and Shaver 1999).

5. *Emotional Communication*

In Bowlby's own words, 'Thus what is happening during these early years is that the pattern of communication that a child adopts towards his mother comes to match the pattern of communication that she has been adopting towards him. Furthermore, from the findings stemming from the Adult Attachment Interview, it seems clear that the pattern of communication a mother adopts towards her infant and young child is modeled on the pattern characteristic of her own intrapsychic communications. Open and coherent intrapsychic communication in a mother is associated with open and coherent two-way communication with her child, and vice versa' (1991c, pp. 295–6). To Bowlby, 'the concern of a psychoanalyst is not only with the extent to which a patient feels free to express his emotions openly, but with the prior questions of whether he knows what his feelings are and what has aroused them' (1991c, p. 294).

One reason for the success of attachment theory is that it lies at the heart of our emotional life. 'Many of the most intense emotions arise during the formation, the maintenance, the disruption and the renewal of attachment relationships. The formation of a bond is described as falling in love, maintaining a bond as loving someone, and losing a partner as grieving over someone. Similarly, threat of loss arouses anxiety and actual loss gives rise to sorrow; while each of these situations is likely to arouse anger. The unchallenged maintenance of a bond is experienced as a source of security and the renewal of a bond as a source of joy' (1991b, p. 306).

6. *Implications*

6.1 *Normative Development and Parenting*

The implications of attachment theory for the development and expression of emotions within close relationships, including how patterns may be transmitted across generations, is clear from the above. Furthermore, the sense of security which must underlie any appreciation of emotions within one's self and their expression to others depends upon having attachment figures who are sensitively responsive to emotional needs. A number of studies have revealed what types of parental interactions promote security, thereby paving the way for methods of parental guidance and childcare policies in general.

6.2 *Psychopathology*

With children who have already developed disorders, particular patterns of insecure attachment shed light on aspects of parenting that might be modified to promote security. For example, the Disorganized pattern, often seen in clinical samples, appears to be associated with fearfulness in either parent or child or both, thereby suggesting a window for intervention (Lyons-Ruth, in Cassidy and Shaver 1999). Additionally, attachment theory has direct implications for disorders related to separation or loss, such as abnormal grief, neurotic depression, or agoraphobia (see Holmes 1993).

6.3 *Social Policy Issues*

When a child goes to hospital, we now take it for granted that parents may visit or even live in. However, before Bowlby's influence, hospital practice involved leaving caregiving to the experts and keeping relatives away. Obstetric practice followed similar lines, and Bowlby (1953) did not hesitate to challenge this: 'reflect for a moment on the astonishing practice which has been followed in obstetric wards of separating mothers and babies immediately after birth and ask ... whether this is the way to promote a close mother–child relationship. It is hoped that this madness of western society will never be copied by so-called less developed societies.' His concern over separation from attachment figures remains relevant to our present provision of care for adults, such as the elderly and mentally ill.

Regarding daycare, Bowlby was keen to put right the misinterpretation of what he wrote. In stressing the importance of early caregiving, he did not mean that a parent must be with the child all the time. From the above it will be clear that attachment theory is about the quality of a relationship rather than amount of time spent together. Similarly with adoption, attachment theory does not say that early adverse experience cannot be overcome. Indeed the implication is that adoptive parents, if sensitively responsive to the child's needs, may do a great deal to set interactions in a cycle that will promote security.

Finally, attachment theory has been applied to the level of society. As a professor of social planning has urged, 'We need also to institute a style of governing our relationships with each other which takes as its first principles reciprocity of commitment, predictability, and respect for the unique structure of meaning and attachment which makes life worthwhile for each of us.' Such an argument implies the 'need to create a world in which we dare become attached' (Marris 1991, p. 89).

7. *Conclusion*

Thus, Bowlby produced a framework for understanding a core aspect of behavior, as set out in his trilogy, *Attachment and Loss* (1969/1982, 1973, 1980). The fruits of his theory are gathered together in the *Handbook of Attachment* (Cassidy and Shaver 1999), dedicated to John Bowlby and Mary Ainsworth. With 36 chapters and an inspired Epilogue by Mary Main, the Handbook is organized into six sections: Overview, Biological perspectives, Attachment in infancy and

childhood, Attachment in adolescence and adulthood, Clinical applications, and Emerging topics and perspectives. The last major publication of Bowlby himself was a biography of Charles Darwin (Bowlby 1991). With a convincing application of attachment thinking, Bowlby explained Darwin's lifelong psychosomatic symptoms in terms of unmourned loss of mother at age eight. Additionally, a keen reader may find parallels between the two men. Like Darwin, Bowlby had a distant father who was a doctor, both were younger sons, both were open to evidence, and both lacked rancor towards their detractors. Finally, it might be said of their theories that they have the quality of immediacy and 'obviousness.' In retrospect it seems obvious that species have evolved by natural selection, that people are attached to one another and suffer when they separate 'but it took child-like simplicity of vision combined with mature determination and attention to detail to root out the obvious and to create for it a secure theoretical base' (Holmes 1993, pp. 34–5; see also van Dijken 1998).

See also: Adult Psychological Development: Attachment; Attachment Theory: Psychological; Emotion and Expression; Emotions, Children's Understanding of; Infancy and Childhood: Emotional Development; Klein, Melanie (1882–1960); Love and Intimacy, Psychology of; Psychiatry, History of; Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis, History of

Bibliography

Ainsworth M D S, Blehar M C, Waters E, Wall S 1978 *Patterns of Attachment*. Erlbaum, Hillsdale, NJ

Bowlby J 1953 *Child Care and the Growth of Love*. Penguin, Melbourne, Australia

Bowlby J 1969/1982 *Attachment and Loss, Vol. I: Attachment*. Hogarth Press, London

Bowlby J 1973 *Attachment and Loss, Vol. II: Separation, Anxiety and Anger*. Hogarth Press, London

Bowlby J 1980 *Attachment and Loss, Vol. III: Loss, Sadness and Depression*. Hogarth Press, London

Bowlby J 1991a *Charles Darwin: A New Life*. Norton, New York

Bowlby J 1991b Ethological light on psychoanalytical problems. In: Bateson P (ed.) *Development and Integration of Behaviour*. Cambridge University Press, Cambridge, UK, pp. 301–13

Bowlby J 1991c Postscript. In: Parkes C M, Stevenson-Hinde J, Marris P (eds.) *Attachment Across the Life Cycle*. Tavistock-Routledge, London, pp. 293–7

Cassidy J, Shaver P R (eds.) 1999 *Handbook of Attachment: Theory, Research, and Clinical Applications*. Guilford Press, New York

DeWolff M S, van IJzendoorn M H 1997 Sensitivity and attachment: A meta-analysis on parental antecedents of infant attachment. *Child Development* **68**: 571–91

Hinde R A 1991 Obituary: John Bowlby. *Journal of Child Psychology & Psychiatry* **32**: 215–17

Hinde R A, Stevenson-Hinde J 1991 Perspectives on attachment. In: Parkes C M, Stevenson-Hinde J, Marris P (eds.) *Attachment Across the Life Cycle*. Routledge, London, pp. 52–65

Holmes J 1993 *John Bowlby and Attachment Theory*. Tavistock-Routledge, London

Marris P 1991 The social construction of uncertainty. In: Parkes C M, Stevenson-Hinde J, Marris P (eds.) *Attachment Across the Life Cycle*. Tavistock-Routledge, London, pp. 77–90

van Dijken S 1998 *John Bowlby: His Early Life*. Free Association Books, London

J. Stevenson-Hinde and R. A. Hinde

Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences

1. Introduction

Many of the effects of aging on behavior depend on changes that occur in the brain. But because the effects of normal aging on behavior, cognition, and personality are not completely determined by neurobiological factors, it is important to consider not only the age-related limitations imposed by inevitable biologically based losses, but also the extent to which there is plasticity of brain function during the later years. To understand fully the consequences of brain aging for behavior, it is important to keep in mind that human aging occurs along multiple dimensions simultaneously, and that there is dynamic interplay between brain aging, physical health and various behavioral and socioenvironmental influences. This entry draws on a developmental framework for summarizing and organizing current research findings and conceptual issues in the study of brain aging and its consequences for human behavior. From this perspective (see Baltes 1997), it is recognized or presumed that there are simultaneous gains and losses in function along various dimensions of human aging, and that there are potentials as well as limits in function throughout the adult life span.

2. Normal and Pathological Brain Aging

The substantial increases in the relative and absolute numbers of older individuals throughout the world pose serious challenges for medicine as well as for the behavioral and social sciences (e.g., Fries 1997). In trying to understand the relationships between brain aging and behavior, it is important to distinguish between the effects of brain aging on behavior that occur for all individuals, referred to as normal brain aging, and the effects of disease on behavior that are more prevalent in older populations. Even the study of normal brain aging reveals an enormous range of differences both among species and within populations in the course of biological senescence. All neurophysiological systems gradually decline in efficiency throughout the adult years. Although senescence

affects all individuals and all biological systems, different rates of senescence can be observed for different systems within individuals. For particular individuals, for example, long-lived macromolecules can become damaged through racemization of amino acids or glycation of proteins. Alternatively, individual differences might be observed in terms of the amount of shrinkage of large neurons and the buildup of amyloid deposits developed in extracellular spaces of neurons in the cortex and hippocampus.

Unsurprisingly, there are qualitative as well as quantitative differences between the effects of normal aging and dementing illness. The hippocampus is, or appears to be, only mildly affected by normal aging, yet there is substantial atrophy of hippocampal structures and in the entorhinal in patients in the early stages of Alzheimer's disease. At different levels of organization within individuals, it is useful to consider differences in the rates of degenerative changes for different structures and functions. However, it is also important to consider that a threshold model may be more useful than a continuous decrement model for describing some of the functional relationships between decline of cognitive functions and age-related changes in brain function; that is, significant cognitive deficits may appear somewhat abruptly only after a critical amount of structural deterioration has occurred.

3. *What is Brain Aging the Aging of?*

Until recently, it was generally accepted that progressive loss of numbers of neurons was an inevitable consequence of normal aging, and that age-related declines in cognitive function could be attributed to neuronal death. Early studies reported that age-related neural loss was 25–50 percent in most neocortical areas and in some areas of the hippocampus, but new studies have revealed that neuronal viability and not neuronal death is significantly involved in normal brain aging. Shrinkage of dendritic branching, and other changes in neuronal viability and in the neural architecture with aging are central to the understanding of age-related changes in cognitive function (e.g., Morrison and Hof 1997, Raz 2000).

As already mentioned, brain aging is both global and specific or selective. Various aspects of biological aging, together or separately, conspire to produce age-related losses in behavioral and cognitive function. Current theories of biological aging emphasize genetically programmed deficits, the cumulative negative consequences of mutation, and the cumulative negative effects of wear and tear on function (for reviews, see Finch 1996). Among the brain areas most affected by aging are the frontal-striatal circuits of the caudate. Other areas that are known to be substantially affected by normal aging are the prefrontal cortex and the dense network of projections connecting the prefrontal cortex and subcortical monoaminergic nuclei. The extent of age-related deterioration is also substantial for the posterior association areas and the neostriatum. The primary sensory areas of the brain are largely unaffected by aging. In light of findings of preserved brain function in the sensory areas, reliable declines in sensory efficiency with aging are attributed to receptor function and peripheral neural function.

The most prominent behavioral deficits associated with aging have to do with a general slowing in the speed of performance. Age-related slowing in information processing speed accounts for a substantial part of the variance associated with age differences in performance across a wide range of cognitive tasks (e.g., for comprehensive reviews, see Cerella 1991, Salthouse 1996). The performance of older adults can be precisely described in terms of a ratio of about 1.4–1.7 of the performance of young adults. To quote Cerella:

> The effects in some 288 experimental conditions are primarily determined by a single aspect of the information processing requirement, namely, task duration. The evidence is near-to-overwhelming that age is experienced, at least to a first approximation, as some sort of generalized slowing … success over such a diversity of data suggests that aging effects stem from some elementary aspect of the biology of the nervous system. (Cerella 1991, pp. 220–1)

Age-related slowing in cognitive performance across tasks may be associated with a number of brain-aging phenomena such as changes in neurotransmitter substances at the synapse, neural demyelination, disruptions in neural circuitry related vascular lesions, and increases in the extent or amount of recruitment and activation of brain volume. With regard to possible age differences in the amount of recruitment of brain areas, the results of several recent fMRI studies suggest that more cognitive resources are required for older adults than for younger adults to carry out cognitive tasks. In these studies, younger persons showed a relatively well-defined pattern of activation involving the inferior prefrontal-orbitofrontal cortex during the encoding of episodic information, and involving the right prefrontal cortex during the retrieval of episodic information. In contrast, older adults showed more widespread activation patterns at encoding and retrieval.

Much of the work in the cognitive neuroscience of aging has focused on age-related memory loss. Some aspects of memory function show large deficits associated with normal aging, whereas other aspects of remembering appear to be relatively unimpaired in healthy older adults (e.g., semantic priming, implicit memory). Relatively substantial age-related declines in cognitive function can be observed in tasks that entail the active and simultaneous processing and storing of information. This kind of processing, referred to as working memory, is a key component of cognitive tasks that require juggling lots of infor-

mation at once. Research bearing on the neural substrates underlying age-related deficits in working memory suggests that working memory requires cooperation among scattered areas of the brain. Which specific areas are activated depends on whether the task requires spatial or semantic memory. Working memory has been associated with the prefrontal cortex. Structural imaging studies indicate shrinkage of the prefrontal cortex with aging. However, functional imaging studies indicate increased activation in the prefrontal areas for older adults than for younger adults. These findings can be interpreted as suggesting that older adults use more cognitive resources to carry out demanding cognitive tasks.

New developments in behavioral genetics bear on the understanding of cognitive changes associated with aging. For example, in a recent study of 5,888 older adults (65 years old and above), the presence of apolipoprotein E epsilon 4 allele (APOE-ε4) was associated with cognitive decline over a seven-year period (Haan et al. 1999). APOE-ε4 is a plasma protein involved in cholesterol transportation. The gene coding for APOE is located on chromosome 19. Relatively little cognitive decline was found for healthy individuals, but systolic blood pressure, atherosclerosis, diabetes mellitus, and subclinical cardiovascular disease were predictive of cognitive decline.

In describing the consequences of brain aging for behavior, cognition, and personality, it is important to consider the consequences of age-related changes in immune function and in hormonal function (e.g., Uchino et al. 1996). Recent research indicates changes in estrogen levels in women is an important factor in both normal brain-cognitive aging and in neurodegenerative disease. Clinical studies show that estrogen has a protective effect in regard to the onset of Alzheimer's disease and perhaps in regard to the course of normal age-related neurodegeneration. It seems that estrogen can protect neurons against amyloid induced toxicity and other toxic effects. The relationship between loss of estrogen and risk of degenerative disease is of obvious relevance to postmenopausal women. Examination of the links among N-methyl-D-aspartate (NMDA) receptors, estrogen levels, hippocampal circuitry, and memory function is a significant area for future research.

The role of the cerebellum in cognition and behavior also deserves mention in this review. The cerebellum serves to coordinate voluntary movement, and may also play a role in the progression of dementia, schizophrenia, and other psychiatric disorders (see Rapoport et al. 2000).

Recent descriptions of the evolution of the human brain might also bear on the understanding of brain aging and its consequences for behavior. For example, Rapoport (1999) suggested that the evolution of primates is characterized by the emergence of neural networks in the neocortex that are activated by mental activity *per se*, absent of sensory and motor stimulation. Thus, even though the heritability of longevity is small, less than 35 percent, it is possible that repeated selection of genotypes has enabled development of brain organizations responsive to higher-order cognitive functions.

4. *Brain Plasticity and Brain Reserve Capacity*

Considering the inevitable aspects of brain aging and the increased prevalence of dementing diseases in later life, how and to what extent is it possible to improve, sustain, or restore effective behavioral and cognitive function in later life? The term, 'brain plasticity' refers to an individual's potential for change, especially the potential for growth or maintenance or restoration of function in response to loss or disease. Recent conceptions of brain plasticity are based on a number of findings showing that some aspects of neural circuitry and synaptic connectivity are capable of growth and repair throughout life. Brain plasticity and effective cognitive function may be interdependent in that continued active involvement in cognitive activities might serve to facilitate brain plasticity, or at least lead to the use of alternative strategies for effective cognitive and behavioral functioning. That is, brain plasticity is probably an outcome of deliberate cognitive efforts to restore, protect, or promote function plasticity more so than a normally occurring or spontaneous phenomenon in later life.

There is abundant evidence demonstrating that associative learning, especially deliberate or explicit learning, produces changes in cortical sensory and motor function. Among the many important questions still to be answered is how intraneuron and interneuron function is altered in response to stimulation. Along these lines, recent work examining long-term potentiation shows how changes in the sensitivity of neural transmission are affected by repeated stimulation and how such changes in sensitivity are maintained across time. Repeated stimulation produces alterations of presynaptic and postsynaptic neurons. Studies of changes in neurochemical function as a result of stimulation promise to contribute to the understanding of the potentials and limits of brain function in later life.

Early approaches to neuropsychological rehabilitation by Luria and others emphasized recovery of function by reorganization of surviving neural circuits. Various forms of rehabilitation achieve moderate degrees of success in restoring or improving behavioral function after acute brain injuries, through methods designed to produce reorganization or adaptation of brain circuitry (Robertson and Murre 1999). Traditionally, the focus of rehabilitation has been on teaching brain-injured individuals to regain function by performing actions in ways that use different neural apparatus. However, new methods and approaches in rehabilitation medicine are emerging that are based on

neural network theory and take into account the possibility of restoration of damaged networks (e.g., Plaut 1996).

Some researchers have suggested that there are links between brain plasticity and vulnerability to and protection from cognitive/functional impairment due to strokes and neurodegenerative diseases such as Alzheimer's disease (e.g., Satz 1993). The extent to which there is brain plasticity or brain reserve capacity could determine the threshold levels at which particular neurological diseases produce observable symptoms, and recent research has focused on the development of measures that index protection from the deleterious effects of normal brain aging and neurological disorders such as Alzheimer's disease. For example, the neural correlates of processing speed, working memory, learning rates, and sensory function are likely to be differential predictors of cognitive and personality function for different aged adults. Brain plasticity or brain reserve capacity may be called upon when an individual faces high levels of cognitive demand, or when the individual faces challenges associated with normal brain aging or with rehabilitation in response to neurological damage or disease.

5. *Conclusions*

Age-related changes in behavior, cognition, and personality during adult years and in later life depend on the interplay of brain aging and behavioral factors (i.e., learning, regularly occurring adaptive organism–environment transactions). Age-related losses occur for many aspects of brain aging, but there is also growth and stability for some aspects of brain function throughout life. Concepts of brain plasticity and brain reserve capacity give emphasis to the potential of individuals to continue to improve, maintain, or optimize behavioral function in response to insidious losses associated with normal aging, and to relatively abrupt losses associated with disease.

See also: Aging and Health in Old Age; Aging Mind: Facets and Levels of Analysis; Aging, Theories of; Cognitive Aging; Lifespan Theories of Cognitive Development; Memory and Aging, Cognitive Psychology of; Memory and Aging, Neural Basis of; Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

Bibliography

Baltes P B 1997 On the incomplete architecture of human ontogeny. *American Psychologist* **52**: 366–80

Baltes P B, Lindenberger U 1997 Emergence of a powerful connection between sensory and cognitive functions across the adult lifespan: A new window to the study of cognitive aging? *Psychology and Aging* **12**: 12–21

Cerella J 1991 Age effects may be global, not local: Comment. *Journal of Experimental Psychology: General* **120**: 215–23

Finch C E 1996 Biological bases for plasticity during aging of individual life histories. In: Magnusson D (ed.) *The Life-span Development of Individuals: Behavioral, Neurobiological, and Psychosocial Perspectives*. Cambridge University Press, New York, pp. 488–511

Fries J F 1997 Can preventive gerontology be on the way? *American Journal of Public Health* **87**: 1591–3

Gabrieli J D E, Brewer J B, Desmond J E, Glover G H 1997 Separate neural bases of two fundamental memory processes in the human medial temporal lobe. *Science* **276**: 264–6

Haan M N, Shemanski L, Jagust W J, Manolio T A, Kuller L 1999 The role of APOE epsilon4 in modulating effects of other risk factors for cognitive decline in elderly persons. *Journal of the American Medical Association* **282**: 40–6

Morrison J H, Hof P R 1997 Life and death of neurons in the aging brain. *Science* **278**: 412–19

Plaut D C 1996 Relearning after damage in connectionist networks: Towards a theory of rehabilitation. *Brain and Language* **52**: 25–82

Rapoport M, Van Reekum R, Mayberg H 2000 The role of the cerebellum in cognition and behavior. *Journal of Neuropsychiatry and Clinical Neurosciences* **12**: 193–8

Rapoport S I 1999 How did the human brain evolve? A proposal based on new evidence from in vivo brain imaging during attention and ideation. *Brain Research Bulletin* **50**: 149–65

Raz N 2000 Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In: Craik F I M, Salthouse T A (eds.) *The Handbook of Aging and Cognition*, 2nd edn. Erlbaum, Mahwah, NJ, pp. 1–90

Robertson I H, Murre J M J 1999 Rehabilitation of brain damage: Brain plasticity and principles of guided recovery. *Psychological Bulletin* **125**: 544–75

Rowe J W, Kahn R L 1997 Successful aging. *The Gerontologist* **37**: 433–40

Salthouse T A 1996 The processing speed theory of adult age differences in cognition. *Psychological Review* **103**: 403–28

Satz P 1993 Brain reserve capacity on symptom onset after brain injury: A formulation and review of evidence for threshold theory. *Neuropsychology* **7**: 273–95

Uchino B N, Cacioppo J T, Kiecolt-Glaser J K 1996 The relationship between social support and physiological processes: A review with emphasis on underlying mechanisms and implications for health. *Psychological Bulletin* **119**: 488–531

West R L 1996 An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin* **120**: 272–92

W. J. Hoyer

Brain Asymmetry

Brain asymmetry or cerebral asymmetry refers to anatomical, physiological, or behavioral differences between the two cerebral hemispheres. The hemisphere that is larger, more active, or superior in performance is dominant. The scientific study of cerebral dominance is recent and dates back to Paul Broca's discovery in 1865, based on observing acquired language deficit (aphasia) following left hemisphere (LH) stroke, that the left cerebral hemisphere of right-handed people is dominant for language. Until the mid-1940s, the general neurological consensus was that the LH of right-handers is dominant for all higher

functions and that the RH is dominant in left-handers. European neurologists and neuropsychologists, such as Hecaen, Piercy, McFie, and Zangwill, then noted that the LH is dominant for language and planned movements (praxis), whereas the RH is dominant for visuospatial functions. This led to replacing the dogma of exclusive LH specialization by the theory of complementary hemispheric specialization (Denes and Pizzamiglio 1999, Heilman and Valenstein 1993).

Beginning in the 1960s, research on Vogel and Bogen's split brain patients in Sperry's lab at Caltech confirmed dramatically the model of complementary hemispheric specialization by comparing the positive competencies of the two hemispheres instead of inferring them from deficit. More importantly, split brain research showed that each disconnected cerebral hemisphere constituted a complete cognitive system with its own sensations, perceptions, memory, even language, personality, and consciousness (Zaidel et al. in press). The two hemispheres can process information simultaneously and independently. This is the thesis of hemispheric independence (Zaidel et al. 1990).

Split brain research, in turn, gave an impetus to research on asymmetric functions in the normal brain, notably by Kimura and Bryden at McGill. Earlier studies by Shephard Ivory Franz in 1933 on complementary normal hemispheric specialization for reading and nonsense shape recognition were ignored because of his untimely death and due to his own antilocalizationist stance on mind/brain relations. Curiously, both Franz's work and subsequent work in Hebb's lab at McGill in the early 1950s were motivated by issues of transfer of training (across the retina or across the hands) and were confounded by issues of attentional scanning from left to right.

Today, cerebral asymmetry remains a cornerstone of human neuropsychology and serves as a model system for a fundamental question in cognitive neuroscience: how do separate subsystems of the mind/brain maintain their independence, on the one hand, and interact, on the other?

In this article, the focus is on asymmetries in the brains of normal right-handers. Asymmetries observed after unilateral brain-damage or in the split brain are discussed elsewhere in this book. Also omitted are discussions of individual differences in cerebral asymmetry associated with left-handedness and with sex. Finally, the coverage excludes abnormal asymmetries associated with congenital disorders, such as dyslexia, and with psychopathology, such as schizophrenia.

1. *Anatomical Asymmetries*

In a seminal paper published in 1969, Geschwind and Levitsky described a dramatic anatomical asymmetry: the left planum temporale, underlying the sylvian fissure near Wernicke's area, was larger than the right in 70 percent (really 80 percent) of right-handers. Correspondingly, the sylvian fissure is larger on the left in 70 percent of right-handers and this is said to be associated with LH specialization for language. The association is imprecise: lesion studies suggest that about 98 percent of right-handers have language specialized in the LH. A similar anatomical asymmetry is already present in the fetus by 31 weeks of gestation. These asymmetries occur whether the area is defined by anatomical landmarks or cytoarchitectonically. They appear to be reduced in left-handers. Lesion, PET and fMRI studies suggest that planum temporale asymmetry is related to phonological decoding and auditory comprehension.

Geschwind and Galaburda also found that pars opercularis (Broca's area) is larger on the left. More grossly, the frontal region is wider and larger in the RH whereas the occipital region is wider and larger in the LH.

Scheibel described an asymmetry in dendritic organization in the left and right operculum: Higher order dendritic branching was greater on the left, whereas lower order dendritic branches were longer on the right. This was attributed to quicker early RH development but greater LH development after the first year of life. This dynamic epigenetic pattern of changing histological asymmetries contrasts with the stable gross asymmetries described above.

2. *Neurochemical Asymmetries*

The neurotransmitter dopamine is involved with control of motor behavior and higher integrative functions, and is lateralized to the LH. By contrast, noradrenergic innervation and the neurotransmitters norepinephrine, as well as serotonin, are associated with up regulation and down regulation of autonomic and psychological arousal, respectively, and are lateralized to the RH. This is said to confer LH dominance to the 'activation' regulatory system, but RH dominance to the 'arousal' regulatory system. Activation refers to 'readiness for action,' and leads to motor functions. Arousal refers to 'orientation to input stimuli' and leads to attentional control. These, in turn, are conjectured to determine hemispheric specialization for perception, action, emotion, and cognition.

There are also asymmetries in neuroendocrine activity. Thus, there is asymmetry in cerebral control of cortisol secretion during emotional situations (Wittling in Davidson and Hugdahl 1995). There is also asymmetry in neuroimmunologic activity. LH activation seems to be immunoenhancing, whereas RH activation is immunosuppressive.

Fluctuating levels of steroid hormones, such as changes of estrogen during the menstrual cycle, may

affect LH function by modulating regional dopamine and other catecholamines. The corpus callosum is further believed to mediate dopaminergic asymmetry. Thus, increased callosal function may be associated with reduced dopaminergic, and consequently functional, asymmetry. It is, therefore, postulated that stable steroid hormones, notably testosterone, affect the organization of hemispheric asymmetry prenatally, whereas fluctuating hormones, notably estrogen, affect the activation of neurotransmitter tracts and selectively change callosal function.

Asymmetries in brain receptors, such as nicotinic acetylcholinergic receptors, may underlie asymmetric hemispheric activation, predominantly on the right, during uptake of recreational drugs, including alcohol, caffeine, and especially nicotine. Asymmetric hemispheric effects have also been shown for pharmacological agents, such as lithium bicarbonate, methylphenidate, and chlorpromazine. This is important for biological psychiatry. For example, changes in behavioral hemispheric asymmetry following floxetine treatment successfully predicts therapeutic response in depression.

3. *Physiological Asymmetries*

3.1 *PET and fMRI*

These techniques measure hemodynamic response and have a good spatial resolution (~ 4 mm) but poor temporal resolution (several seconds). Cabeza and Nyberg (2000) provide a valuable review of imaging during cognition. In general, these studies display bilateral activations that contrast with the asymmetric effect of lateralized lesions.

3.1.1 Attention. Sustained attention (vigilance) tasks yield prefrontal and parietal activation, preferentially in the RH. *Spatial compatibility* (Stroop) tasks selectively engage the anterior cingulate and left prefrontal cortex. *Orienting of visuospatial attention* selectively engages right posterior parietal cortex. *Divided attention* tasks selectively activate left prefrontal cortex. Thus, attention involves a bilaterally distributed network whose components are lateralized to one or the other hemisphere.

3.1.2 Motor. Unimanual choice reaction time selectively activates left prefrontal-intraparietal areas compared to simple reaction time, consistent with LH specialization for action. Spatial compatibility tasks activate dorsal premotor and superior parietal areas, predominantly on the left. Imitation of hand movements and of object manipulation activate bilateral inferior frontal areas, more on the left, and bilateral parietal areas, more on the right (mirror neurons).

3.1.3 Perception. Perception of *objects* and *faces* activates the same ventral pathway but there is greater right lateralization for faces than for objects. Perception of *smell* activates olfactory orbitofrontal cortex, especially on the right. Right occipitotemporal (fusiform) cortex is selectively involved in processing specifitj c visual form information (same view of same object), whereas the left is more involved in processing categorical information (different exemplars). *Music* perception engages right superior temporal cortex for melodies, but left Broca's area/insula for rhythms. Recognition of pictures of self selectively activates right prefrontal cortex.

3.1.4 Imagery. Generation of images from words activates left inferior posterior temporal lobe, although more complex visual imagery selectively engages right temporal cortex. Imagery of maps engages right superior occipital cortex, whereas mental navigation engages the left middle occipital gyrus.

3.1.5 Language. Spoken word comprehension activates bilateral medial and superior temporal gyrus but activation during written word recognition is left lateralized. Surprisingly, left prefrontal cortex is activated equally during written word recognition with or without spoken responses. Written words also selectively activate left posterior temporal cortex. There is a progression of hemispheric specialization: feature conjunction in letters occurs in medial extrastriate cortex contralateral to the stimulus visual field (VF). Letter shapes activate posterior fusiform gyrus opposite the processing hemisphere, whereas letter names activate anterior and posterior fusiform gyrus bilaterally. Letter strings activate left posterior fusiform area, whereas legal words (visual word forms) activate left anterior fusiform area with stimuli in either VF. Lexical access via the visual word form activates the posterior temporal lobe bilaterally, whereas phonological recoding activates a left-lateralized network of posterior (fusiform gyrus, presumably involved in sub-lexical graphemic processing) and anterior (inferior frontal gyrus, presumably involved in phonological recoding) structures. This supports the 'hemispheric dual route model': both hemispheres have lexical routes ('sight reading'), but only the LH has a nonlexical ('phonetic') route.

Deaf native users of American Sign Language show normal LH activation during language processing. In addition, they show recruitment of RH fronto-

temporal-parietal regions presumably for processing location and motion information present in ASL. This RH recruitment may be subject to a critical period.

Listening for the moral of a fable selectively activated frontal and temporal regions of the RH.

3.1.6 Memory. The *phonological loop* component of working memory distinguishes between a phonological store, associated with left parietal activation, and a rehearsal buffer, associated with activation in Broca's area. *Working memory* engages prefrontal and parietal regions. These are lateralized by type of material: Broca's area and left parietal areas are activated for verbal/numerical tasks; activations for nonverbal material tend to be bilateral.

Tulving's hemispheric encoding/retrieval asymmetry (HERA) model posits that (a) the left prefrontal cortex is selectively involved in semantic memory retrieval, (b) the left prefrontal cortex is selectively involved with episodic memory encoding (personally experienced past events), and (c) the right prefrontal cortex is selectively involved in episodic memory retrieval. However, recent data show material-specific laterality: LH activation during both encoding and retrieval of words, and RH activation during both encoding and retrieval for patterns, faces, scenes, or nonverbal sounds.

Different semantic domains engage different cortical regions. For example, retrieval of animal information engages left occipital regions, perhaps reflecting processing of physical features, whereas retrieval of tool information engages left prefrontal regions, perhaps reflecting processing of linguistic or motor information. Similarly, generating action words engages left temporo-occipital regions, close to motion perception regions. In other words, knowledge about object attributes appears to be stored close to the regions involved in perceiving those attributes.

There may be a dissociation in right prefrontal cortex between dorsal cortex, associated with free recall, and ventrolateral cortex, associated with cued recall. Finally, autobiographical retrieval is associated with activation in a right fronto-temporal network.

3.1.7 Emotions. Activation during perception of different *emotions* converges in the left prefrontal cortex. Dolan (in Gazzaniga 2000) reports lateralized right amygdala activation during masked (unconscious) presentation of conditioned fearful faces, but left lateralized amygdala activation during unmasked (conscious) presentation.

3.2 The Electroencephalogram (EEG)

The EEG has a good temporal resolution but poor spatial resolution and, therefore, it provides a relatively crude measure of hemispheric asymmetry. It is now beginning to be used together with imaging techniques like fMRI, that have a good spatial resolution but poor temporal resolution. EEG measures tend to focus on ongoing changes in the frequency domain whereas event-related potential (ERP) measures focus on stimulus-yoked changes in the voltage wave pattern over a second or so. Different frequency ranges of the EEG have different functional significance. Early work by Galin and Ornstein in the 1970s focused on the α band (8–12 cycles per second or Hz), whose power is inversely related to cognitive engagement. They found a larger right-to-left power ratio during verbal (LH) tasks and a smaller ratio during spatial (RH) tasks. They also found differential stable asymmetries in different professionals: lawyers with larger R/L ratios (more LH activity), and artists with larger L/R ratios (more RH activity).

More recently, focus has shifted to asymmetries in higher frequency bands of the EEG. For example, the gamma band (25–35 Hz) showed a dissociation between word and nonword reading or listening in the LH but not in the RH. Also promising, are measures of intrahemispheric and interhemispheric connectivity and coherence from large (128) electrode arrays that permit better spatial localization and provide indices of network dynamics during cognition.

Davidson (in Davidson and Hugdahl 1995) used EEG measures of α suppression to characterize anterior hemispheric emotional asymmetries and individual differences in emotional reactivity, mood, and temperament. Approach-related positive affect was associated with left anterior activation, whereas withdrawal-related negative affect was associated with right anterior activation.

3.3 Et-related Potentials (ERP)

Separate components of the ERP are associated with semantic and syntactic components of postlexical integration in language comprehension (Brown et al. in Gazzaniga 2000). The left perisylvian area is critical for syntactic processing, as well as for aspects of higher-order semantic processing. Within 200 msec after stimulation, processes related to lexical meaning and integration emerge in the ERP waveform:

(a) A transient negativity over left anterior electrode sites (left anterior negativity, LAN) emerges 200–500 msec after word onset. The LAN is associated with initial parsing and processing of syntactic word category information.

(b) A transient bilateral negativity (N400) develops between 200 and 600 msec after word onset. The N400 is associated with contextual effects of semantic processing (e.g., semantic unexpectedness of a noun in a sentence or discourse).

(c) A transient bilateral positivity develops between 500 and 700 msec, called the syntactic positive shift (SPS) or the P600, and is associated with syntactic

correction operations, such as arise from syntactic violations.

(d) A slow positive shift over the front of the brain develops across the sentence, and is associated with its overall meaning. The slow shift has separate left and right lateralized components.

Anterior LH specialization for grammar but not for semantics normally develops at age 36–42 months and is absent in later language acquisition.

3.4 Psychophysiological Asymmetries

Consistent with the thesis of hemispheric independence, brain asymmetry applies to cognition, emotion, and regulation of autonomic physiological processes.

Neurotransmitter, neuroendocrine, immunomodulatory, cardiovascular, and electrodermal activity, skin temperature, and vasomotor activity are all asymmetric.

Neural control of cardiovascular function is lateralized. Peripheral autonomic chronotopic control of cardiac activity (heart rate) is controlled by right-lateralized sympathetic and parasympathetic pathways. By contrast, AV conduction and cardiac contractivity are left-lateralized. Emotional stimuli in the LVF yield larger heart rate and blood pulse volume changes than stimuli in the RVF. These peripheral asymmetries are not associated with corresponding cognitive asymmetries. Finally, there may be an asymmetry in pain sensitivity, greater in the RH.

In sum, the LH appears to specialize for motor functions and in the regulation of the body's defense against invading agents, whereas the RH appears specialized for control of vital functions supporting survival and coping with stress.

4. Behavioral Asymmetries

4.1 Models of Behavioral Asymmetry

Behavioral asymmetries, or laterality effects, in the normal brain are measured by restricting input stimuli or behavioral responses or both to one sensory or motor side. By far the most common experimental paradigm is hemifield tachistoscopy with unimanual responses. Stimuli are flashed for less than 180 msec to one visual hemifield (projecting to the opposite hemisphere) to prevent involuntary saccades towards the stimulus, and response choices are signaled by different fingers of one hand, again controlled by the opposite hemisphere. For the sake of simplicity, the following methodological discussion refers to a hemifield experiment, but similar comments apply to experiments in other modalities.

Split brain research motivates two limited case models of behavioral laterality effects in the normal brain. The first is the exclusive specialization model, where only one hemisphere, say the LH, can process the stimuli, so that stimuli in the left visual field (LVF) must be relayed from the RH to the LH through the corpus callosum prior to stimulus processing. This model is called the callosal relay (CR) model. The second is the relative specialization, or hemispheric independence (of representation and strategy) model, where each hemisphere can process the stimuli, though usually not equally well, and the input VF determines the processing hemisphere. This is called the direct access (DA) model. Consider the example of a lateralized lexical decision task with orthographically legal nonwords and unimanual word/nonword responses. A callosal relay model predicts a main effect of input VF (RVF advantage or RFVA) as well as a main effect of an independent stimulus variable (say, wordness) but no significant interaction between the two (assuming that words and nonwords are equally complex and equally susceptible to delay or degradation in callosal relay). By contrast, a main effect of VF (RVFA) and a significant interaction of VF and an independent variable, say wordness, implies that the DA model holds.

Some behavioral laterality tasks, like phonetic perception in dichotic listening to nonsense consonant-vowel (CV) syllables, are CR, exclusively specialized in the LH. Surprisingly many, like lateralized lexical decision, are DA. For reasons still unclear, it appears that the VF of the target, rather than the response hand, determines the hemisphere in control of processing. Even when DA obtains, there may be implicit interhemispheric priming effects and resource borrowing, leading to the concept of 'degrees of hemispheric independence.' Many tasks probably engage both DA and CR components.

4.2 Methodology

4.2.1 Signal detection. For reasons still unclear, behavioral laterality effects are more likely to occur with accuracy than with latency as the dependent variable. The signal detection measure of bias, β, often shows a more negative bias in the LVF than in the RVF. A speed-accuracy tradeoff analysis of some DA tasks suggests that the two hemispheres have similar speed-sensitivity functions, suggesting similar decision strategies, but different speed-bias functions, implying hemispheric differences at the response choice stage. This highlights the need to apply signal detection measures to behavioral laterality experiments. That is rarely done.

4.2.2 Previous trial effects. Common laterality tests show previous trial effects involving the correctness,

the VF, and the target identity of the previous trial. They exhibit a selective implicit RH error monitoring, and same-VF or same-target priming effects across trials, indicating the importance of context effects and the ubiquitousness of hemispheric momentum effects in time.

4.2.3 Bilaterality effect. Bilateral displays, with targets in one VF and distractors in the other, increase hemispheric independence and the behavioral asymmetry.

4.2.4 Reliability and validity. This is a sorely neglected topic. A standard measure of the laterality index is $(L-R)/(L+R)$, where L = accuracy in the LVF, etc. Voyer (1998) carried out a meta-analysis of 88 experiments and found an average test-retest or split-half reliability of the laterality index of 0.68. Dichotic listening had greater reliabilities than VF studies, and auditory verbal tests had greater reliabilities than nonverbal tests. Verbal dichotic tests had an average reliability of 0.70. Dichotic listening to stop consonant-vowel (CV) syllable pairs showed the highest reliability, 0.80. Accuracy yielded higher reliabilities than latency. Number of items affected split-half but not test–retest reliability. Split-half reliabilities were higher than test-retest reliabilities. Extensive testing with dichotic CVs yielded reliabilities around 0.83 and testing with lateralized lexical decision including lateralized distractors yielded reliabilities around 0.81.

The reliability of the difference between the two VF scores is maximized when they are negatively correlated with each other, but when each is based on a large number of trials that are positively intercorrelated. Indeed, the correlation between the two ear scores in dichotic listening varies from 0.1 to -0.72 and the use of bilateral stimuli in VF studies reduces the correlation between the two VFs.

Voyer's estimate of criterion validity from cross-correlations of laterality measures between tests yielded a low validity of 0.26. However, the procedure assumed a true global measure of hemispheric specialization. This is unlikely. Different functions even within the same domain, such as language, are lateralized to different degrees. Thus, phonetic perception appears to be exclusively lateralized whereas visual and especially auditory word recognition are more bilaterally represented. Instead, one can use the split brain as establishing construct validity. By administering the same behavioral laterality tests to split and normal brains, it is possible to establish the underlying degree of lateralization of the relevant functions (Zaidel et al. 1990). This procedure reveals that lateralized lexical decision measures relative hemispheric specialization, but dichotic listening to nonsense CV syllables is exclusively specialized in the LH. Consequently, individual differences in the right ear advantage reveal individual differences in LH competence or in callosal relay (from RH to LH), or both.

4.2.5 Laterality and ability. There is another sorely neglected aspect of the validity of behavioral laterality tests: the degree to which lateralized performance predicts the underlying or associated ability in free vision. It is possible that a RVFA reflects better LH performance overall and thus better scores on any test predominantly controlled by the LH. This concept is captured by Levy's individual differences in stable hemispheric arousal asymmetry. Such characteristic asymmetry is said to account for half of the variance in asymmetry scores across tasks, and it contributes more to asymmetry in tasks with bilateral stimuli. However, no relationship was found between characteristic arousal asymmetry and performance on various tests of cognitive ability.

The question remains whether performance in a lateralized version of a task predicts free vision performance on the same task? It does, but surprisingly weakly. For example, there was an unexpected RVFA in a test for recognizing facial emotions and a moderate correlation ($r = 0.433$) between the RVFA and performance on a free vision version of the test. In a lateralized lexical decision task, there was the expected RVFA but free vision measures of vocabulary correlated positively with accuracy in either VF (LVF: $r = 0.524$, RVF: $r = 0.59$), and reading comprehension correlated only with LVF accuracy ($r = 0.385$). Voyer also observed a positive correlation between the RVFA in a lateralized mental rotation test and free vision performance on a test of spatial ability. Since the correlations are relatively small, it is misleading to infer hemispheric specialization of a cognitive test from the VFA in a lateralized version of the same or a similar test.

4.3 Domains of Specialization

Consistent with the theory of hemispheric independence, hemispheric differences can occur at any level of information processing.

4.3.1 Attention. Hemispheric engagement by lateralized stimuli creates an activational bias to the opposite half of space and enhances processing of material in that location. Task demands, such as processing facial emotions, may selectively activate the specialized hemisphere. Clinical and imaging studies suggest that the RH is specialized for alerting, arousal, and vigilance. Indeed, testing of covert orienting of spatial attention using the Posner paradigm discloses a greater cost for detecting LVF targets with invalid cues in the RVF than for detecting RVF targets with invalid cues in the LVF. There is some suggestion that the RVF shows selective sensitivity to

object-based cueing, whereas the LVF is selectively sensitive to location-based cueing.

4.3.2 Perception. Older views held that hemispheric specialization is divided by modality, material, or stage of processing, with the LH specialized for auditory, verbal, or output processing; the RH for visual, nonverbal, or input processing. A more recent view is in terms of information-processing styles: the LH is analytic and the RH is synthetic, and a related current view holds that the LH is local and the RH is global, for example, in processing hierarchic patterns. A more precise statement of the latter view is that the LH is specialized for processing relatively high spatial and temporal frequencies whereas the RH is specialized for relatively low frequencies, given the context (Ivry and Robertson 1998). The evidence for hemispheric specialization in the normal brain, in terms of relative frequencies, is still controversial, and the evidence from normals for hemispheric specialization in global/local processing is weak. The global/local hemispheric dissociation has better support from neuroimaging studies in normals and from behavioral studies in hemisphere-damaged patients.

The RH is said to be specialized for processing facial emotions and emotional prosody. In space localization, the LH is said to be specialized for processing categorical relations, such as on-off, inside/outside, above/below, or near/far; the RH is said to be specialized for coordinate, metric relations. The claim appeals to selective RH involvement in magnocellular vision, still a controversial view. The claim received mixed empirical support (Hellige 1993).

4.3.3 Imagery. The componential theory of visual imagery distinguishes image generation, maintenance, scanning, and transformation. There is controversial evidence for LH specialization for image generation and for RH specialization in image rotation. Kosslyn believes that categorical image generation (by description of parts) is specialized in the LH, whereas image generation by stored metric memories is specialized in the RH.

4.3.4 Emotions. There is a LVFA (RH specialization) for recognizing and, perhaps, expressing facial emotions, as well as emotional intonation and emotional gestures. The RH may have a selective role in expressing negative emotions and the LH in expressing positive ones. Heller (in Banich and Heller 1998) distinguishes emotional valence (pleasant, unpleasant), controlled by anterior cortical regions, (pleasant more on the left, unpleasant more on the right), from emotional arousal (high, low) controlled by right posterior regions. Heller also distinguishes two groups of individuals with high trait anxiety: those with anxious apprehension (worry), and increased LH EEG activation (α suppression), and those with anxious arousal (panic), and increased RH activation. Pettigrew believes that bipolar disorder is the result of a slow interhemispheric switching mechanism that becomes stuck on the left (mania) or on the right (depression).

4.3.5 Language. Findings from neurological patients guide the design and interpretation of behavioral laterality experiments in normal subjects: the anterior LH is specialized for phonology and syntax, more posterior (temporal) language areas in the LH control semantics, and the RH is specialized for pragmatics.

Dichotic listening studies suggest that phonetic perception (e.g., distinguishing stop consonant-vowel syllables by place of articulation and voicing) is specialized in the LH. Ivry finds that the perception of voice onset time is specialized in the RH, whereas the perception of place of articulation is specialized in the LH, because voicing is carried by low speech frequencies and place is carried by high frequencies.

Word repetition priming is case- and font-sensitive in the LVF but not in the RVF and the word superiority effect (detecting a letter is faster in a word than in a nonsense string) is greater in the RVF, suggesting that word recognition in the LH is quicker and deeper. However, format distortion, for example, a LtErNaTiNg case or vertical presentations, selectively impairs word recognition in the RVF, more for words than for nonwords, suggesting a selective contribution of pattern recognition to word recognition in the LH. Word length can affect word recognition in both hemispheres, more in the RH than in the LH. Both hemispheres show comparable word frequency effects, but the LH shows a selective word regularity effect (better recognition of words with regular than with exceptional spelling-to-sound correspondence). The preceding observations are consistent with the view that both hemispheres can recognize words via a lexical route ('sight vocabulary') but that the LH has selective access to a nonlexical route (spelling-to-sound, 'phonics'). However, when taxed, the RH does show regularity effects, suggesting that it has latent competence for spelling-to-sound conversion.

When we fixate during reading, we look ahead, thus selectively attending to the RVF in English (read left-to-right), but to the LVF in Arabic (read right-to-left). Indeed, hemifield tachistoscopic studies show that while both types of languages are specialized in the LH, early visual stages of word recognition engage the RH more in Farsi than in English.

Semantic priming experiments suggest that the RH activates automatically more distant semantic relationships and multiple meanings of ambiguous words.

The LH may be more specialized for function words (closed class items) than for content (open class items). Sentences are more effective primes for the LH than for the RH, showing that the LH is selectively specialized for syntax. Still, both hemispheres show sensitivity to grammatical agreement, suggesting that the RH does have some grammatical competence. Finally, RH activation of multiple distant meanings may help it make predictive inferences in discourse, especially when breaks in coherence occur.

Based on lateralized priming experiments, Chiarello concluded that lexical orthographic relationships are more available to the RH, phonological relationships are more available to the LH, and semantic relationships are available to both. Automatic lexical semantics is more available to the RH, whereas controlled, strategic, lexical semantics is more available to the LH.

4.3.6 Olfaction. Unlike vision, hearing, and somesthesis, the olfactory system is mainly uncrossed: left nostril → LH, right nostril → RH. Studies of normal subjects suggest that the olfactory thresholds of the two hemispheres are the same. Odor discrimination and intensity judgments show an advantage in the right nostril (RH), but odor naming shows a left nostril superiority. However, results are inconsistent and behavioral imaging and lesion studies reach conflicting conclusions. Nonetheless, there is agreement that interhemispheric interaction is critical for olfaction: birhinal presentation (both nostrils) increases perceived odor intensity and facilitates odor naming, although it does not affect detection thresholds.

5. Convergence

There are systematic discrepancies between profiles of hemispheric specialization observed in the normal brain (Bradshaw and Nettleton 1983, Bryden 1982), in the split brain (e.g., Zaidel et al. in press), and in hemisphere-damaged patients (Denes and Pizzamiglio 1999, Heilman and Valenstein 1993). Laterality effects are typically smallest in the former and largest in the latter. Thus, unilateral lesions can lead to devastating cognitive deficits not mirrored in the disconnected, let alone the normal, hemispheres (Benson and Zaidel 1985). For example, large perisylvian LH lesions can lead to global aphasia and yet the disconnected RH is neither word deaf nor word blind. Similarly, right posterior lesions can lead to neglect of the left half of space and denial of illness (anosognosia) and yet the disconnected LH does not exhibit neglect or denial. In general, it seems that unilateral lesions incorporate loss of function in diseased tissue, diaschisis, and pathological inhibition of residual competence in the healthy hemisphere. Those effects are minimized in the split brain and absent in the normal brain. Consequently, the disconnected hemispheres exhibit greater functional competence than inferred from lesion studies. The normal hemispheres, in turn, evince even greater competence than the disconnected hemispheres, presumably because they can borrow resources from each other even when they process information independently.

Hemispheric representation for language is illustrative. Phonological and syntactic deficit follow exclusively LH lesions, semantic deficits follow predominantly LH lesions and pragmatic deficits follow predominantly RH lesions. But the disconnected RH has substantial auditory language comprehension, including some grammar, a modicum of lexical (but not sublexical) reading, and on occasion even some speech (Zaidel et al. in press). Consequently, appeal is often made to RH language in trying to account for perplexing symptoms in aphasic syndromes, including covert reading in pure alexia, semantic errors in deep dyslexia, or good miming in the presence of misnaming written words in optic aphasia. The normal brain suggests that RH language competence is even richer, including grapheme-phoneme correspondence. The challenge for future research is to release latent RH language capacities for clinical recovery in aphasia, and to modulate degree of hemispheric specialization in the normal brain for cognitive optimization.

See also: Asymmetry of Body and Brain: Embryological and Twin Studies; Brain, Evolution of; Cerebral Cortex: Organization and Function; Electroencephalography: Basic Principles and Applications; Learning and Memory, Neural Basis of; Split Brain

Bibliography

Banich M T, Heller W (eds.) 1998 Evolving perspectives on lateralization of function. *Current Directions in Psychological Science* **7**: 1–37

Benson D F, Zaidel E (eds.) 1985 *The Dual Brain: Hemispheric Specialization in Humans*. Guilford Press, New York

Bradqhaw J L, Nettleton N (eds.) 1983 *Human Cerebral Asymmetry*. Prentice Hall, Englewood Cliffs, NJ

Bryden M P 1982 *Laterality: Functional Asymmetry in the Intact Brain*. Academic Press, New York

Cabeza R, Nyberg L 2000 Imaging Cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience* **12**: 1–47

Davidson J, Hugdahl K (eds.) 1995 *Brain Asymmetry*. MIT Press, Cambridge, MA

Denes G, Pizzamiglio L (eds.) 1999 *Handbook of Clinical and Experimental Neuropsychology*. Psychology Press, UK

Gazzaniga M S (ed.) 2000 *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA

Heilman K M, Valenstein E (eds.) 1993 *Clinical Neuropsychology*, 3rd edn. Oxford University Press, New York

Hellige J B 1993 *Hemispheric Asymmetry: What's Right What's Left?* Harvard University Press, Cambridge, MA

Ivry R B, Robertson L C 1998 *The Two Sides of Perception*. MIT Press, Cambridge, MA

Zaidel E, Clarke J M, Suyenobu B 1990 Hemispheric independence: A paradigm case for cognitive neuroscience. In: Scheibel A B, Wechsler A F (eds.) *Neurobiology of Higher Cognitive Function*. Guilford Press, New York, pp. 297–355

Zaidel E, Iacoboni M, Zaidel D W, Berman S M, Bogen J E in press Callosal syndromes. In: Heilman K M, Valenstein E (eds.) *Clinical Neuropsychology*, 4th edn. Oxford University Press, New York

E. Zaidel

Brain Damage: Neuropsychological Rehabilitation

Rehabilitation in its broadest sense means 'the restoration of patients to the highest level of physical, psychological and social adaptation attainable' (WHO 1986). This includes measures to reduce handicap caused by social disadvantage, improve social integration, designing disability-favorable environments and many others. For the purposes of this article, however, attention will be concentrated on the subset of rehabilitation activities whose aim is to augment cognitive functions, and use the following definition of specifical neuropsychological rehabilitation: 'the systematic use of instruction and structured experience to manipulate the functioning of cognitive systems such as to improve the quality and/or quantity of cognitive processing in a particular domain of mental activity' (Robertson 1999, p. 703).

Neuropsychological rehabilitation is, therefore, a specialized component of more general rehabilitation, whose aim is the maximization of the functional independence and adjustment of the brain damaged individual. Rehabilitation in general may include a wide range of goals, the following of which are a tiny subset of examples: achieve standing balance; diminish family tension by reducing incidence of temper outbursts; achieve independent use of the lavatory; allow verbal communication via keyboard-operated voice synthesizer; reduce depression by increasing level and range of day-to-day activities; achieve supported participation in a workplace.

While neuropsychological rehabilitation may have a part to play in achieving some of the above goals, its domain is that of cognition—namely attention, memory, perception, gnosis, praxis, reasoning, and executive control. Setting goals for cognitive rehabilitation is therefore a superficially more straightforward task than in the wider realm of rehabilitation. Of course, some more general rehabilitation interventions (e.g., reducing depression or increasing level of participation in everyday activities) may have effects on cognition; these, however are not neuropsychological rehabilitation *per se*, because the result is a by-product of a more general goal, and the intervention was not specifically and directly planned to alter cognition.

1. History

While the attempt to foster recovery in brain function has a long history, it only properly took scientific root in conjunction with the great breakthroughs in the understanding of brain function made by the great European neurologists in the late nineteenth and early twentieth centuries. Paul Broca, for instance, described attempting to teach an aphasic patient to read again, stating: ' … I am convinced that, without returning to aphemics (aphasics) the part of their brain, one could, with enough perseverance, treating them with the infatigable consistency of the mother that teaches her infant to speak, one could, I say, obtain considerable results' (cited in Ryalls and Lecours 1996, p. 240).

The deadly technological advances of the twentieth century took a hand, however, in forcing a more active interest in neuropsychological rehabilitation. During World War I, Kurt Goldstein in Frankfurt and Walter Poppelreuter in Cologne were pressed into service trying to rehabilitate the bullet- and shrapnel-blasted brains of the German infantry. As early as autumn 1914, Poppelreuter founded an institution, a military hospital devoted to rehabilitation. Goldstein developed a theoretical model for neuropsychological rehabilitation, arguing that *restitution* of cognitive function may be possible through repetitive stimulation of the impaired brain regions, but only if some of the substrate of the cognitive functions in question were spared. Where the damage was more severe and widespread, *compensation* should be the aim of rehabilitation. As early as 1917, electrically powered memory devices were in use, providing repeated stimulation of stimuli that patients were required to learn by heart (for review, see Poser et al. 1996).

It was during the carnage of World War II, however, that another great neuropsychologist A. R. Luria turned his attention to the rehabilitation of brain damaged Russian servicemen and women. He developed four main principles of rehabilitation: (a) respect for the unique and variable nature of the brain's functional systems across individuals; (b) making use of the intact brain system to replace the damaged one; (c) allowing for previously internalized acts to be externalized, through speech or other external aids, until the behavior can again be performed automatically; and (d) using feedback constantly to provide patients with information about their strengths and weaknesses (Luria 1963). Luria emphasized the interaction between assessment of the nature of the cognitive impairment and the development of individually tailored methods to foster recovery.

2. Effectiveness of Neuropsychological Rehabilitation

The pioneering work of the first half of the twentieth century was carried out without controlled clinical

trials of the effectiveness of the therapeutic procedures being carried out. This was not surprising, because the notion of controlled clinical trials of behavioral or psychotherapeutic methods was at that time relatively unknown, and it was really only in the late 1960s and early 1970s that the effectiveness of various types of neuropsychological rehabilitation began to be formally assessed.

Even after the instigation of controlled trials in the 1970s, however, the evidence about the effectiveness of neuropsychological rehabilitation for a range of neuropsychological disorders is still not unequivocal. This is probably because (a) the number of trials is still very small and (b) the statistical power and methodological adequacy of most trials is not optimal. Furthermore, with only a few exceptions, rehabilitation methods have not been tied closely to underlying theoretical models of cognitive function. Nevertheless, there are some grounds for optimism in the clinical trial literature. Robey (1994), for instance, carried out a meta-analysis of 21 studies of aphasia rehabilitation and concluded that there was an effect of therapy, particularly when it was started relatively early following brain damage. More recently, one study has found changes in cerebral metabolism that may have been associated with training-induced improvements in language comprehension (Musso et al. 1999).

Another pioneering series of studies took place in New York in the 1970s, where a purely behavioral approach to rehabilitation of unilateral left neglect—a loss of attention and/or responding to the left side of space—was attempted (Weinberg et al. 1979). This strategy involved training patients to engage in habitual scanning towards the left side of space. Promising results in the New York studies were supported in subsequent studies of a more extensive and adapted form of training carried out in Rome. Pizzamiglio and colleagues subsequently showed that in three patients with unilateral neglect following primarily subcortical lesions, recovery was mainly associated with cerebral activation in right hemisphere cortical regions known to subserve the relevant visuoperceptual functioning in normal subjects.

3. Compensatory vs. Restitutive Approaches to Goal-setting in Cognitive Rehabilitation

In the 1990s, advances in our understanding about the plasticity of the adult central nervous system (Nudo et al. 1996) required that a new attempt be made to formulate a theory of recovery of function which allows not only for compensation as a mechanism of recovery but also for partial restitution of the impaired neuropsychological processes themselves. This was particularly true given the evidence that cell genesis was found to be possible in the adult human (Eriksson et al. 1998) and that in mice the extent of such new cell proliferation in the dentate gyrus was influenced by hippocampal-dependent associative learning (Gould et al. 1999).

It is, however, still acknowledged that behavior changes can occur following lesions that are determined by compensatory changes in the way the behavior is subserved by the brain. For instance, it was shown that while unilateral neglect following right hemisphere damage apparently recovers on measures such as line bisection, patients still showed very marked distortion in their representation of contralateral space as measured by a closely similar task requiring no motor response but only perceptual judgment (Harvey and Milner 1999). It is probable that the basic problem originally causing the neglect remains, but that patients have learned to compensate for this by adjusting their motor responses to the external world.

Compensatory approaches that target specific functional behaviors, without attempting to foster restitution in the underlying damaged brain circuits, have been argued for (Wilson and Watson 1996), and in the case of memory rehabilitation, there is as yet no evidence for direct and lasting improvement of memory through restitution-oriented therapies. Hence compensatory approaches to memory problems appear to be—for the time being at least—the treatment of choice.

The theoretical underpinnings for restitution-oriented rehabilitation were discussed in a review paper (Robertson and Murre 1999) arguing that such approaches are feasible in the case of lesions that spare a proportion of connections in a lesioned circuit. In such cases, targeted input that respects the facilitatory and inhibitory architecture of the cognitive systems may potentially 'rescue' circuits that would otherwise disintegrate. Such 'rescue' had indeed been demonstrated in monkeys (Nudo et al. 1996).

The question of how to choose between a restitutive and a compensatory approach to rehabilitation is crucial and not a question of purely academic interest. Faced, for instance, with someone who cannot move her arm, does one devote limited rehabilitation resources to trying to get her to move that arm, or to developing compensatory strategies to allow as normal a life as possible while avoiding the use of that arm? Faced with someone showing severe expressive dysphasia, do we try to train him to produce spoken words or do we teach him alternative means of communication? Confronted with 'dysexecutive' problems, do we focus our efforts on structuring the environment to support more organized behavior, or do we struggle to retrain at least some internally mediated executive control skills?

Quantitative theoretical models have been proposed that suggest ways of choosing between restitutive and compensatory approaches (Robertson and Murre 1999), but until we can routinely image the living human brain in such a way to assess the residual

connectivity in a lesioned circuit, rehabilitationists must rely on other methods. The presence of residual behavioral function apparent under certain circumstances may represent a reasonable proxy basis for assessing which individuals may potentially be amenable to restitution via guided recovery of primary lesioned circuits. Such an approach has already been adopted in the successful Constraint Induced Movement Therapy for hemiplegia (Taub and Wolf 1997, Miltner et al. 1999). Specifically, patients with a minimal degree of hemiparetic limb function could have that function partially restituted through a combination of restraint of the nonimpaired limb and graded movements of the partially hemiparetic limb.

4. *Theory–Practice Links in Neuropsychological Rehabilitation*

The last two decades of the twentieth century saw increasingly strong links between cognitive neuroscience theory on the one hand and the development of rehabilitation methods on the other. One good example of this comes from studies of the interaction of the brain's perceptual and motor systems. The theoretical basis of this arose from findings that certain types of visual information may have privileged access to the control of motor responses, yet not be available to awareness. The notion that the so-called dorsal stream (Ungerleider and Mishkin 1982) may provide visual information for the motor system that is not available to awareness has received strong evidence from a series of experiments by Goodale and Milner (Milner and Goodale 1995). They have shown how cortically damaged patients who are incapable of consciously discriminating perceptual features (e.g., orientation) may nevertheless be able to make appropriate motoric responses that are sensitive to these features.

In the light of these results, we predicted that the manifestation of unilateral neglect may be alterable by changing the purpose of otherwise very similar responses to spatially extended objects. Spatial-neglect-based deviation to the right of center was, for instance, significantly less when subjects reached towards metal rods as if to pick them up in the center compared to when they pointed to their centers. This, and other experiments, suggested that prehensive movements towards objects allow 'leakage' of information about their spatial extent via an unaffected stream of information available for motor-manipulative responses. This finding was compatible with data showing that while a conscious, 'ventral' representation of a stimulus can contaminate or over-ride the short-lived motor representation, no such reciprocal influence seems to occur (Rossetti 1998). In other words, this latter finding is compatible with data showing that activation of the dorsal stream appears to have effects on the ventral stream. In a rehabilitation outcrop of this study, people suffering from unilateral left neglect were actually encouraged repeatedly to pick up rods. In so doing, not only would one be repeatedly facilitating the putative 'crosstalk' between the dorsal and ventral systems, but one would also induce perceptual conflict in neglect patients: even though reaching to grip reduced neglect, their grip was still biased to the right of center. Hence when they picked up the rod at the point which visually seemed to them to be the center, they would experience contradictory feedback from both the proprioceptive and visual modalities to show that this was not in fact the case. The authors predicted short-term improvements in neglect as a result of a brief number of such exposures. Indeed, significant short-term improvements in neglect were found on neglect tasks after subjects experienced proprioceptive and other feedback discrepant from the judgments they made based on visual information alone (Robertson et al. 1997). The ventral/dorsal concept in cognitive neuroscience is not however the only one that has generated rehabilitation methods. Other examples include theories of frontal lobe/executive function leading to rehabilitation methods (Levine et al. 2000).

In the field of learning and memory, another theory–practice link was shown in work by Baddeley and Wilson (1994) and Wilson et al. (1994). They demonstrated that people with severe memory disorders learn better when trial and error methods are avoided during learning. Errorless learning is a teaching technique whereby people are prevented as far as possible from making mistakes while they are learning a new skill or new information. Instead of teaching by demonstration which may involve trial and error, the correct information or procedure is presented in ways that minimize the possibility of erroneous responses. Errors are likely to be reinforced by people with poor episodic memory because of their reliance on implicit memory which is inefficient at error elimination. Errorless learning has proved effective over a range of tasks, including learning names, new information, identification of line drawings, and programming an electronic aid. It is also effective for a range of diagnostic groups including head injury, stroke, encephalitis, and Korsakoff's syndrome, and a range of times post-insult from post-traumatic amnesia to 12 years post-infection.

In short, rehabilitation made significant progress during the twentieth century, but the progress in the next century will likely depend on much closer links being forged with the basic cognitive neuroscience of the brain as well as with combining behavioral and pharmacological methods for accelerating brain repair. The possibility of gene therapy also beckons, but it is entirely likely, as in the case of neural transplants, that the requisite improvement in behavior only occurs when such methods are combined with the best behavioral stimulation.

See also: Cognitive Neuropsychology, Methodology of; Memory Problems and Rehabilitation; Recovery of Function: Dependency on Age

Bibliography

Baddeley A D, Wilson B A 1994 When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia* **32**: 53–68

Eriksson P S, Perfilieva E, Bjork-Eriksson T 1998 Neurogenesis in the adult human hippocampus. *Nature Medicine* **4**(11): 1313–17

Gould E, Beylin A, Tanapat P, Reeves A, Shores T J 1999 Learning enhances adult neurogenesis in the hippocampal formation. *Nature Neuroscience* **2**(3): 260–5

Harvey M, Milner A D 1999 Residual perceptual distortion in 'recovered' hemispatial neglect. *Neuropsychologia* **37**: 745–50

Levine B, Robertson I H, Clare L, Hong J, Wilson B A, Duncan J, Stuss D T 2000 Rehabilitation of executive functioning: An experimental-clinical validation of goal management training. *Journal of the International Neuropsychological Society* **6**: 299–312

Luria A R 1963 *Restoration of Function After Brain Injury*. Pergamon, Oxford, UK

Mayer E, Brown V J, Dunnett S B, Robbins T W 1992 Striatal graft-associated recovery of a lesion-induced performance deficit in the rat requires learning to use the transplant. *European Journal of Neuroscience* **4**: 119–26

Milner A D, Goodale M A 1995 *The Visual Brain in Action*. Oxford University Press, Oxford, UK

Miltner W H, Bauder H, Sommer M, Dettmers C, Taub E 1999 Effects of constraint-induced movement therapy on patients with chronic motor deficits after stroke: A replication. *Stroke* **30**(3): 586–92

Musso M, Weiller C, Kiebel S, Mueller S P, Buelau P, Rijntjes M 1999 Training-induced brain plasticity in aphasia. *Brain* **122**: 1781–90

Nudo R J, Wise B M, SiFuentes F, Milliken G W 1996 Neural substrates for the effects of rehabilitative training on motor recovery after ischemic infarct. *Science* **272**: 1754–91

Pizzamiglio L, Perani D, Cappa S F 1998 Recovery of neglect after right hemispheric damage: H215O positron emission tomographic activation study. *Archives of Neurology* **55**: 561–8

Poser U, Kohler, J A, Schoenle P W 1996 Historical review of neuropsychological rehabilitation in Germany. *Neuropsychological Rehabilitation* **6**: 257–78

Robertson I H 1999 Setting goals for rehabilitation. *Current Opinion in Neurology* **12**: 703–8

Robertson I H, Murre J M J 1999 Rehabilitation of brain damage: Brain plasticity and principles of guided recovery. *Psychological Bulletin* **125**: 544–75

Robertson I H, Nico D, Hood B M 1997 Believing what you feel: Using proprioceptive feedback to reduce unilateral neglect. *Neuropsychology* **11**: 53–8

Robey R R 1994 The efficacy of treatment for aphasic persons: A meta-analysis. *Brain and Language* **47**: 582–608

Rosetti Y, Rode G, Pisella L 1998 Prism adaptation to a rightward optical deviation rehabilitates left hemispatial neglect. *Nature* **395**: 166–69

Ryalls J, Lecours A R 1996 Broca's first two cases: From bumps on the head to cortical convolutions. In: Code C, Wallesch W-W, Joanette Y, Lecours A R (eds.) *Classic Cases in Neuropsychology*. Psychology Press, Hove, UK, pp. 235–42

Taub E, Wolf S L 1997 Constraint induced movement techniques to facilitate upper extremity use in stroke patients. *Topics in Stroke Rehabilitation* **3**: 38–61

Ungerleider L G, Mishkin M 1982 Two cortical visual systems. In: Ingle D J, Goodale M A, Mansfield R J W (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA, pp. 549–86

Weinberg J, Diller L, Gordon W, Gerstman L, Lieberman A, Lakin P, Hodges G, Ezrachi O 1979 Training sensory awareness and spatial organization in people with right brain damage. *Archives of Physical and Medical Rehabilitation* **60**: 491–6

Wilson B A, Baddeley A D, Evans J J, Shiel A 1994 Errorless learning in the rehabilitation of memory impaired people. *Neuropsychological Rehabilitation* **4**: 307–26

Wilson B A, Watson P C 1996 A practical framework for understanding compensatory behaviour in people with organic memory impairment. *Memory* **4**: 465–86

Woldorff M G, Gallen C C, Hampson S A, Hillyard S R, Pantev C, Sobel D, Bloom F E 1993 Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proceedings of the National Academy of Sciences USA* **90**: 8722–6

World Health Organization 1986 Optimum care of disabled people. Report of a WHO meeting. WHO, Turku, Finland

Zangwill O L 1947 Psychological aspects of rehabilitation in cases of brain injury. *British Journal of Psychology* **37**: 60–9

Zangwill O L 1975 Excision of Broca's area without persistent aphasia. In: Zulzch K J, Creutzfeldt O, Galbraith G C (eds.) *Cerebral Localization*. Springer-Verlag, Berlin

I. H. Robertson

Brain Development, Ontogenetic Neurobiology of

Human infants have the longest period of dependence upon caregivers of any mammal, and changes in behavior during their development are obvious to all of us. Over time, the largely reflexive movements of the newborn infant are replaced by the highly purposive activity of the adult. The long period of human postnatal development allows the developing nervous system to reflect variability in the environment experienced by the infant as well as variability in genetic endowment. Much of the structural development of the nervous system has taken place before birth. Studies of primates have led to important advances in our understanding of how brains develop, including information on the prenatal development of the cerebral cortex (Rakic 2000).

Within common postnatal development there is variability among infants and children. One infant is fearful when presented with new people or objects; another does everything possible to seek out new levels of stimulation. Questions of cognitive and emotional development and individual differences have had a long history within psychology and education. There has also been considerable progress in exploring cognitive, social, and emotional behavior and in

specifying the knowledge and processing mechanisms involved in behavior at varying ages (Ruff and Rothbart 1996).

However, despite some impressive starts (e.g., Johnson and Morton 1991), links between psychological and biological development remain largely unexplored. One reason for the current high level of interest in human brain development is to establish connections between the psychological development of children and the study of the genetics and neurobiology of how brains become organized before and after birth. During the 1990s, the adult human brain has become increasingly available for analysis with new methods of neuroimaging. The effort to apply functional imaging methods to infants and children has barely begun and important obstacles to overcome in making these adaptations remain (Thomas and Casey 1999). In this essay we briefly summarize progress in developmental biology and psychology, concentrating on efforts to adapt current methods to achieve linkages between their separate findings.

1. Neurobiology

1.1 Cortical Structure

The human brain develops before birth largely under the instructions contained in the genetic message of 48 chromosomes containing tens of thousands of genes. For all humans, the gross organization of the brain is very similar. For example, the cerebral cortex, containing the primary sensory and motor systems, is a six-layered structure. Although differing in detail between brain regions, the cerebral cortex shows remarkable similarity in cellular structure across the entire brain, suggesting the strong influence of genetics on the brain's gross organization (Rakic 2000).

From the study of non-human animals, neurobiologists have shown that nerve cells follow a pathway from their place of origin to their eventual location within the cortex. This process of cell migration continues during the early years of life. From birth through childhood, there is evidence from infant autopsies on how the structure of the human cortex develops in each sub-region (Conel 1939). Although brain areas develop at different rates, in all areas cell migration proceeds from the deep layers to the more superficial layers of the cortex (Rakic 2000).

Despite the uniformity in the overall structure of the brain, there is evidence of the importance of experience even in the size and shape of sub-regions. Identical twins, for example, who share the same genetic endowment, have similar brain structure as revealed by imaging studies, but there are often striking differences in the shape of structures, particularly within the right cerebral hemisphere (Posner and Raichle 1994).

1.2 Connections

As brain cells establish layers of the cortex, they also form connections (synapses) with other neurons, forming the sensory and motor systems of the brain. The density of synapses rises rapidly during the first few years of life to reach a peak at about age two to three years. Synapse density then stays roughly constant until declining following puberty. Brain regions appear to differ in the detailed time course of these events (Huttenlocher and Dabholkar 1997).

Connections between neurons are thought to be central to the networks underlying thought and emotion, and findings of rapid changes in synaptic density over the first few years of life can thus be of great importance in shaping how we think about human development. To date, however, few methods have been available to relate synaptic density directly to child behavior, nor are methods available for connecting the details of brain changes with the obvious behavioral changes. For these reasons, few empirical data have been available to constrain theoretical swings in the relative emphasis placed on experience and genes as the basis for changes observed both in the brain and in behavior. Some theorists have thought experience was of primary importance in acquired behavior and others have stressed genetic constraints on what can be learned. New methods for imaging the human brain should eventually provide data to consider the roles of both genetics and experience in shaping the brain and behavior.

The formation and loss of connections between nerve cells continue throughout life, and are widely thought to form the major basis for changes in behavior during life. The early plasticity of the human brain is revealed in cases of brain injury. Even the removal of an entire hemisphere in early infancy may result in very little, if any, deficit, while a much smaller lesion in an adult may cause paralysis of one side of the body and, depending upon the hemisphere, specific deficits of cognition. As we shall see, in some systems there appear to be sensitive periods during which input is extremely important in shaping brain connections. However, in general, the ability to form new connections and learn new facts and procedures continues throughout life.

2. Neuroimaging

2.1 Adult Studies

Neuroimaging studies have used changes in blood flow or chemistry to examine parts of the brain active while normal adults perform tasks. These studies have provided a general picture of where cognitive and emotional processing is carried out by the brain (Posner and Raichle 1994). They give hope of helping us understand how genetics and experience work to-

gether to produce the behavior patterns of the adult. Neuroimaging studies have examined in the brains of normal adults such processes as perception, working memory, language, spatial attention, attentional selection, arousal, and happy and sad emotional states. The nearly universal picture that emerges is that tasks within each domain activate a specific network of neural areas that often include both posterior and anterior areas of the cortex as well as subcortical structures. As the complexity of tasks increases, more areas are recruited. Most tasks involve both bottom-up activity driven by sensory input and top down activity driven by the motivations, expectations, and attention of the person. Because measures based on changes in blood flow lag neural activity by hundreds of milliseconds, however, it is not possible to measure their time course directly by vascular changes (Posner and Raichle 1994).

Time dependent imaging methods can be obtained by recording electrical or magnetic fields from outside the skull, providing a measure of the time course of mental processing in the millisecond range. These studies have shown that the top down processing is usually delayed for a time following an unexpected target, although if a cue indicates the aspect of the stimulus the subject is to select, top down processing may be rapid.

2.2 Pediatric Studies

Measures of changes in blood flow have just begun to be used in child studies. Methods that require exposure to some amount of radiation cannot be used with children unless there is a medical condition that justifies it. However, structural and functional magnetic resonance imaging (fMRI) techniques have opened the field of imaging to developmental research.

Nevertheless, many special issues arise in the application of the fMRI technique to the study of normal children. Data from fMRI are very sensitive to motion artifacts. When dealing with children, this problem is especially acute, since for most, lying still for long periods of time is extremely difficult. Some fMRI studies have involved infants who are sedated for medical reasons. This prevents motion artifacts, but prevents the use of any functional task. In addition, sedation makes interpretation of results more difficult, especially when they are compared with those of non-sedated adults. Finally, children may experience anxiety and distress upon being in an unknown, hospital-like environment and especially seeing and entering the scanner magnet (Thomas and Casey 1999).

Electrical signals have long been recorded from infants and young children. They provide evidence of the time course of information processing and some evidence on anatomy. There is also hope that optical methods might be particularly useful with infants because of the low density of the skull at this age.

2.3 Marker Tasks

Another method that has been useful for studying development during early childhood is the use of marker tasks. Marker tasks are behavioral tasks that have been shown in adults to involve particular brain systems and even specific portions of these systems. By studying infants' ability to perform different marker tasks with increasing age, it is possible to make inferences about the developmental course of these systems and networks. There must be some caution here because the location of cognitive processes may itself change with development.

A clear example is the use of reaching during the first year of life to study children's ability to inhibit proponent actions elicited by the situation (Diamond 1990). Usually, infants reach where they are looking. A toy is placed under a transparent box. The opening of the box is on the side so that the subject can reach it only if the tendency to reach directly through the transparent top of the box is inhibited. At six months, infants appear to be constrained to reach along the line of sight, but by 12 months they are able to reach through the opening even while looking through the top. A comparison of infants with the performance of monkeys and adult humans with specific brain lesions on similar tasks suggests this task is sensitive to the development of the dorsolateral prefrontal cortex. Maturation of this brain area seems to be critical for the development of this form of inhibition.

3. Perception and Attention

During the first year of life, the visual system undergoes remarkable development supporting increases in visual acuity, the ability to control foveation and attention voluntarily, and recognition of faces and other visual objects. Infant visual acuity has been traced month by month during the first year of life by recording looking preferences (Ruff and Rothbart 1996).

Visual experience is critically important for the normal maturation of the visual system. For example, if one eye of a cat is closed, the representation of that eye in the cortex and its ability to drive visual cells is greatly reduced. Human infants who undergo unilateral removal of cataracts show clear effects of the time of this deprivation period on the development of the visual field (Maurer et al. 1999). Thus, the genetic plan for visual system development includes a critical role for exposure to the visual world.

There are also many indications from animal and human studies that sensory systems maintain the

ability to change with experience well into adulthood. In one set of studies, monkeys were trained to make discriminations with one of their fingers. The representation of that finger in their cortex was altered by the experience, with more cortex devoted to the trained digit than to the others (Merzenich and Jenkins 1995). Patients who have lost a limb due to amputation often develop feelings of the missing limb at other body locations that are represented in parts of the somatosensory cortex adjacent to those of the missing limb (Ramachandran and Blakelee 1998). These changes appear to occur as the result of the loss of input from the missing limb, allowing innervation from adjacent parts of the cortex.

Infants as young as four months can learn to anticipate the location of an event by moving their eyes to a location at which the target will occur. Thus, caregivers can teach important aspects of where to attend and use orienting to counteract distress well before the infant begins to speak. Infants also show preferences for novel objects in the first few months of life. Later in childhood more complex forms of attentional control (effortful control) begin to emerge. Infants first show the ability to reach away from the line of sight (described above), and later the toddler begins to develop the ability to choose among conflicting stimuli and courses of action (Ruff and Rothbart 1996).

4. *Learning and Cognition*

There has been an explosion of information about cognitive processes in infancy and childhood. The infant brain seems to have primitive capabilities for face and object recognition, number processing, language, and imitation from early infancy (Spelke 1998). The nature of these innate achievements is under active debate. Some views stress the infant as having a primitive ability to represent aspects of visual objects, including their solidity (Spelke 1998). Others feel infants have only low level specific tendencies based on aspects of the stimulus, so that more detailed representation emerges only with experience (Johnson 2000). In the case of numbers, while a primitive notion of the quantity of very small numbers seems to be present in animals and infants, a lengthy process including connections to language is needed to allow precise computation (Dehaene 1997).

4.1 *Language*

Experience begins to shape the direction of language even before the child is able to babble. Newborns enter the world with the ability to discriminate phonemes of all the world's languages. However, exposure to their own language leads to development of structures that emphasize the phonemes of their native language and reduces the ability to discriminate phonemes not present in that language (Kuhl et al. 1992). Nonetheless, the ability for relearning phonemes remains present even as an adult. There is no single sensitive period for language, but in some domains there is reason to believe that the degree of plasticity in learning new skills can be reduced or lost as the organism develops. For example, immigrants to the US from China show grammatical skills that reflect the age of their first exposure to English, but a different language skill, word knowledge, is independent of the age of first exposure (Weber-Fox and Neville 1996). These findings suggest that only careful empirical studies can help us understand the skills for which exposure at a particular time in brain development might be of special importance.

In the 1960s, the so-called Whorfian hypothesis held that the nature of thought was constrained by a person's language. Eleanor Rosch (1973) tested this idea in the domain of color names by working with a New Guinea tribe that used only two names for colors (dark and light). She nevertheless found they were better able to learn nonsense names for prototypical hues determined by locating the hue at the center of the set of colors spanned by English category name (for another view see Roberson et al. 2000). The Rosch result was interpreted as meaning that the color names used in English were not arbitrary; they surrounded particular hues that were highly memorable. Subsequently, anthropologists at Berkeley showed that there was a very nice relationship between names used for colors in the world's languages. Although languages differed in the number of color names, the particular spectral frequencies given names tended to emerge in the same order and were clearly related to the brain's systems for color analysis. Far from language specifying what could be best perceived, names seemed to depend upon the structure of the human visual system. The words we use influence the communication of our percepts and even the ability to remember them, but our perception seems to have laws related to the structure of the brain.

4.2 *Categories and Skills*

The development of the human brain is closely related to the acquisition of the many skills that children learn. Of particular importance in all cultures is the learning of categories. Work by Rosch and others extended the prototype analysis described in Sect. 4.1 to many other categories besides color (Smith and Medin 1981). Priming experiments showed that when people thought about a category like animals, they could easily retrieve information about prototypical animals like dogs and cats, but the category name actually slowed them down in thinking about salmon or fly, since fish and insects do not come close to the usual prototypical animals. Most people are familiar with basic categories, which, like dog,

couch or chess, describe a class of things that have physical similarity and common use. They do less well with abstract categories like game, which are harder to think about in terms of a central tendency. Expertise allows people to make finer distinctions than are present in the basic category, so they might think of dogs in terms of separate breeds, each with its own prototypical structure.

Priming experiments give us a good way of seeing what items are automatically activated when we think of one member of a category. Of course, we can also use our attention to overcome or restrict the boundary of activated items. In one experiment, people were taught to think body part when they saw the category name animal. When they did so, targets like arm and finger were responded to quite rapidly. However, if the target was introduced very shortly after the prime, the person would still be fast to COW, indicating that the animal category was automatically activated by the prime, but its access was reduced in efficiency as attention was directed to the learned associations (Neely 1977).

A basic tenet of the cognitive study of expertise has been the idea that experts differ from novices based on the information they have stored in memory about their domain of expertise. Early work in this area involved the memory of chess masters compared to more ordinary players. The chess master could retain the position of many more pieces, provided they were in the orderly position expected from an actual master's level game. If the pieces were randomly placed, the master was little better than the ordinary player (Chi et al. 1988).

Of course, few people become chess masters, but the study of expertise illuminates the learning of many skills acquired in school. Skills such as reading, processing number, or science are based upon relatively automatic access to a rich domain of information in semantic memory. These skills are often closely related to language.

We now have a basis for understanding something about the neural systems that lie behind this acquisition of skill. This work may begin to provide hints about how complex information of the type acquired by school subjects might be organized within the brain. It has been possible to show that the amount of cortical tissue given over to an area of processing can reflect experience with that kind of processing. This was mentioned in our discussion of cortical changes in the representation of a finger used by monkeys in learning tactile discriminations (see Sect. 3). The same finding has also been shown to apply to pianists and violinists who appear to have sensory and motor representations for the fine movements of their instruments that are far beyond those found in nonmusicians. In some laboratory tests, a few hours of training with a particular sequence of movement was sufficient to increase the amount of the cortex activated by the sequence (Karni et al. 1998).

One reason these results with monkeys and humans are so important is that there is now evidence of mapping related words and pictures within semantic areas of the left hemisphere. When subjects were shown pictures of tools, animals, etc., in an fMRI experiment, different brain regions were active within both frontal and posterior cortex for the different categories (Martin et al. 1995). Actual storage of this information may involve brain areas that process attributes such as motion and color. Observations also suggest that brain lesions of the left temporal lobe can disrupt the ability to recognize specific categories of items while leaving other categories completely intact, and reports from animal studies show that experiences stored at a common time may activate common cells (Erickson et al. 2000).

The priming, lesion, and neuroimaging data all suggest that concepts with common semantic and perceptual features may be stored together. It would be expected, based on the monkey results, that the size of these areas would change with recent experience. This experience might influence how easy it is to learn new material or how likely information in that domain will come to mind. We know that maps of important features are very common in the visual, auditory, and motor systems, so it should not be too surprising that related items are stored in adjacent areas within semantic networks.

We are a long way from understanding the details of how the storage of concepts takes place. We know that at least some aspects of dealing with such concepts as object and number categories are present at birth. However, even a primitive understanding of the structural brain changes that occur during the acquisition of concepts should provide teachers and others who work with children new perspectives on methods to use for teaching skills.

4.3 Reading

Reading is a high-level skill that is common in modern society, but not characteristic of all human communities. Neuroimaging studies have shown that the processing of English words involves posterior and anterior areas of the left cerebral cortex related to English orthography, phonology, and meaning (Posner and Raichle 1994). These imaging studies fit well with models in which letters are automatically integrated into a single visual unit (visual word form). The visual word form is probably a series of neural areas in the occipital and temporal cortex that becomes organized around the presentation of visual words. It represents an important adaptation of the visual cortex to performance of a high level skill not learned until childhood.

In acquiring reading, visual words are transformed into a speech-based code. Imaging studies suggest that problems with this speech-based code form the basis

of many forms of dyslexia (difficulty in learning to read). Speech-based codes play an important role in adult reading as well, although there is some dispute as to whether some words look up their meaning more directly from the visual word form system. Whatever the fact of the adult state, it seems likely that the skill of reading English depends both upon learning the alphabetic principles that relate visual letters to sound and development of the visual word form system (Posner and McCandliss 1999). Nevertheless, complete language systems, e.g., American Sign Language, can be constructed without the use of a speech-based code. Neuroimaging studies of sign suggest that many similar left hemisphere areas are recruited in interpreting sign language, but there is much more right hemisphere activity than in reading English (Neville et al. 1998).

5. *Temperament and Emotion*

Human brains differ from one another in gross anatomy. As noted in Sect 1.1, even identical twins whose brains are generally similar may show quite striking differences in the size and shape of particular areas. Moreover, we know that infants come into the world with prepared reactions to their environment and that even siblings can be very different in their reaction to events. On infant, for example, is easily frustrated, has only a brief attention span and cries with even moderate levels of stimulating play. Another may tolerate even very rough play, frequently seek out exciting events, and maintain attention over extended periods. Differences in this reactivity to the environment, together with the mechanisms that regulate them, constitute temperament. As experience accumulates with socialization, temperament becomes the basis for the adult personality (Rothbart et al. 2000).

Much of early temperament relates to positive and negative emotions such as fear or smiling and laughter. The brain's fear system has been widely studied in animals and by neuroimaging in adults. It involves a very specific anatomy including subcortical brain areas such as the amygdala as well as orbital frontal cortex. Differences in temperament and personality depend upon genetic endowment as expressed in brain structure and function and as influenced by experience. These dimensions are not fixed but develop over the life course.

For example, the full fear reaction is not present at birth, but requires some months to develop. Fearful infants have difficulty in exhibiting approach both to strange objects and people. Thus, fear serves a regulatory function. Fear as measured by responses to novel toys in infancy predicts fearfulness at age seven years. In addition, shyness or fear in social situations also appears to be an enduring temperamental variable. Later in childhood, maturation of the frontal lobe produces more reliance on effortful control systems, allowing increased scope for methods of socialization. The strength and effectiveness of this later developing effortful control system are also an important temperamental difference. Fear and effortful control are both major regulatory systems in later childhood and play an important role in the development of high level systems of morality and conscience (see *Temperament and Human Development*).

6. *Summary*

Human brain development seeks to relate basic perspectives from developmental neurobiology and psychology. Studies involving methods of brain imaging now provide opportunity for examining how the development of brain areas influences the expression of cognitive and emotional behavior, and how the brain itself is modified by experience. The study of brain development should help illuminate both general principles of brain plasticity and the basis for human variability.

See also: Cognitive Development in Childhood and Adolescence; Cognitive Development in Infancy: Neural Mechanisms; Infancy and Childhood: Emotional Development; Language Development, Neural Basis of; Neural Development: Mechanisms and Models; Neural Plasticity; Prefrontal Cortex Development and Development of Cognitive Function; Sensitive Periods in Development, Neural Basis of; Temperament and Human Development; Visual Development: Infant

Bibliography

Chi M T H, Glaser R, Farr M J 1988 *The Nature of Expertise*. Erlbaum, Hillsdale, NJ

Conel J L 1939–63 *The Postnatal Development of the Human Cerebral Cortex*. Harvard University Press, Cambridge, MA, Vols. I–VII

Dehaene S 1997 *The Number Sense*. Oxford University Press, New York

Diamond A (ed.) 1990 *The Development and Neural Basis of Higher Cognitive Functions*. New York Academy of Sciences, New York

Erickson C A, Jagadeesh B, Desimone R 2000 Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nature Neuroscience* **3**(11): 1143–8

Gazzaniga M S 1992 *Nature's Mind*. Basic Books, New York

Huttenlocher P R, Dabholkar A S 1997 Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology* **387**: 167–78

Johnson M H, Morton J 1991 *Biological and Cognitive Development: The Case of Face Recognition*. Blackwell, Oxford, UK

Johnson S 2000 The development of visual surface perception: Insights into the ontogeny of knowledge. In: Rovee-Collier, Lipsitt L, Haynes H (eds.) *Progress in Infancy Research*. Erlbaum, Mahwah, NJ, Vol. I, pp. 113–54

Karni A, Gundela M, Rey-Hipolito C, Jezzard P, Adams M M, Turner R, Ungerleider L G 1998 The acquisition of motor performance: Fast and slow experience-driven changes in primary motor cortex. *Proceedings of the US National Academy of Sciences* **95**(3): 922–9

Kuhl P, Williams K A, Lacerda F, Steven K N, Lindblom B 1992 Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* **255**: 606–8

Martin A, Haxby J V, Lalonde F M, Wiggs C L, Ungerleider L G 1995 Discrete cortical regions associated with knowledge of color and knowledge of action. *Science* **270**: 102–5

Maurer D, Lewis T L, Brent H P, Levin A V 1999 Rapid improvement in the acuity of infants after visual input. *Science* **286**: 108–10

Merzenich M M, Jenkins W M 1995 Cortical representation of learned behaviors. In: Anderson P, Ohvalby O, Paulsen O, Hokfelt B (eds.) *Cortical Representation of Learned Behaviors*. Elsevier, Amsterdam, pp. 47–451

Neely J H 1977 Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attentional mechanisms. *Journal of Experimental Psychology: General* **106**: 226–54

Neville H J, Bavelier D, Corina D, Rauschecker J, Karni A, Lalwai A, Braun A, Clark V, Jezzard P, Turner R 1998 Cerebral organization for language in deaf and hearing subjects: Biological constraints and effects of experience. *Proceedings of US National Academy of Sciences* **95**(3): 922–9

Posner M I, McCandliss B D 1999 Brain circuitry during reading. In: Klein R M, McMullen P (eds.) *Converging Methods for Understanding Reading and Dyslexia*, MIT Press, Cambridge, MA, pp. 305–37

Posner M I, Raichle M E 1994 *Images of Mind*. Scientific American Books, New York

Rakic P 2000 Setting the stage for cognition: Genesis of the primary cerebral cortex. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*, MIT Press, Cambridge, MA, pp. 7–21

Ramachandran V S, Blakeslee S 1998 *Phantoms in the Brain*. Fourth Estate, London

Roberson D, Davies I, Davidoff J 2000 Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General* **129**(3): 369–98

Rosch E 1973 On the internal structure of perceptual and semantic categories. In: Moore T E (ed.) *Cognitive Development and the Acquisition of Language*. Academic Press, New York, pp. 111–44

Rothbart M K, Ahadi S A, Evans D E 2000 Temperament and personality: Origins and outcomes. *Journal of Personality and Social Psychology* **78**: 122–35

Ruff H A, Rothbart M K 1996 *Attention in Early Development*. Oxford University Press, New York

Smith E E, Medin D 1981 *Categories and Concepts*. Harvard University Press, Cambridge, MA

Spelke E S 1998 Nativism, empiricism and the origins of knowledge. *Infant Behavior and Development* **21**: 181–200

Thomas K M, Casey B J 1999 Functional magnetic resonance imaging in pediatrics. In: Badetinni P, Moonen C (eds.) *Medical Radiology: Functional Magnetic Imaging*. Springer-Verlag, New York, pp. 513–23

Weber-Fox C M, Neville H J 1996 Maturational constraints on functional specialization for language processing ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience* **8**: 231–56

M. I. Posner and M. K. Rothbart

Brain, Evolution of

When did the human brain evolve, and how did it happen? Obviously, to answer this question will require a time machine, and thousands of generations of observations to ascertain both the variability and direction of selection pressures in the past. We can, however, flesh out an initial understanding of how we got to be the animal *par excellence* that utilizes its brain for intelligent rationalizations based largely on the use of arbitrary symbol systems. The evidence consists of two components: (a) the 'direct' evidence from the fossil record; and (b) the 'indirect' evidence of the comparative neuroscientific record of extant living animals, particularly those most closely related to us, such as the chimpanzee. There is a third possibility: since the human genome project has sequenced almost all of the genetic code, the future study of evolutionary neurogenetics might provide more data about the actual genetic history of our Genus through time, as well as that of the great apes mentioned above. As this latter possibility is simply a mote in our eye at present, this article must concentrate on the evidence provided by the first two components.

1. Lines of Evidence

1.1 Direct Evidence

The term *paleoneurology* is used to describe evidence appraising the size and morphology of the casts made from the inside of actual fossil cranial remains. Occasionally, the casts are 'natural,' i.e., where fine sediments have filled the inside of the cranial cavity, and became compacted through time. These casts often retain whatever morphological details were imprinted on the internal table of bone of the cranium. The famous australopithecine Taung child's skull, described by Dart (1925), is one of the best-known examples. Curiously, these 'natural' endocasts are only found in the S. African australopithecines, of which several exist, and date from about 2.5 MY to about 1.5 MY. Most often, the paleoneurologist makes a cast of the inside of the fossil skull using rubber latex, or silicone rubber, and extracts this from the cranium. The partial cast is reconstructed by adding plasticine (modeling clay) to the missing regions. The whole is then measured by immersion into water, and the amount of water displaced is regarded as the volume

of the once-living brain. Other measurements and observations are made on the original cast.

During life the brain is surrounded by three dural sheaths (dura mater, arachnoid tissue and its cerebrospinal fluid, and pia mater) that interface between the actual brain tissue (cerebral cortex, mostly) and the internal table of bone of the skull. The gyri and sulci (convolutions) of the once-pulsating cerebral cortex are rarely imprinted on the interior of the skull, and the degree of replication often varies in different regions, i.e., sometimes the frontal lobe imprints more details than the parietal lobe. The degree of replication also varies in different animals. Two extremely important considerations emerge from this: (a) the resulting imprints are never complete and thus 'data poor,' and never include subcortical structures; and (b) controversial interpretations of what the underlying brain once looked like are guaranteed. Nevertheless, these endocranial brain casts do provide extremely important information regarding the size, shape, rough estimates of the lobar dimensions of the brain, and cortical asymmetries that have relationships to hemispheric specializations, including handedness. In addition, if the imprints of the underlying gyri and sulci are available, these can provide important information regarding the organization of the cerebral cortex, and whether the patterns of these are the same or different as in known extant primate brains. The infamous 'lunate sulcus' is a good example, as it is a demarcation boundary between purely sensory primary visual striate cortex (PVC) and multi-modal association cortex. When the lunate sulcus appears in an anterior position, it is most similar to the condition known in modern apes. When it is found in a posterior position, it is more in a human-like condition. Ascertaining its correct position is essential in deciding whether or not such a fossil hominid had a brain organized along human or ape lines. Finally, meningeal arteries and veins that nourished the dura mater also imprint on the internal table of bone and these sometimes show patterns that are useful for deciding taxonomic issues; these have no known relationship to behavioral functions of the brain.

1.2 Indirect Evidence

This line of evidence is 'data rich,' providing comparative neurological information on living species, such as brain size (both absolute and relative, i.e., related to body size), the actual makeup of the brain from the gross to microscopic levels, including neural nuclei, fiber systems and interconnections, and distribution of neurotransmitters and neuroreceptors. Additionally, the brain can be studied ontogenetically, and neuroscientists can actually study the relationships between how the brain varies neurologically and how these variations relate to the behavioral variation. This richness is simply lost to the paleoneurologist as it is not available as direct evidence. However, it is necessary to realize that the extant living species, e.g., chimpanzee, macaque, are end points of their own evolutionary lines of development and are not our ancestors, however closely related. It is the blending and complementation of these two approaches which provide the best set of evidence for when and how our brains evolved.

2. Characteristics of the Human Brain

2.1 Brain Size, Absolute and Relative

The human animal is obviously obsessed with size, and those who study the brain comparatively, perhaps more so. With an average brain weight of 1,330 grams (Tobias 1971), the human species has the largest absolute brain size within the primate order, but is actually dwarfed by elephants and some of the Cetacea, where brain weight can exceed 7,500 grams. Of course, the body weights are very much higher in elephants and whales. But even for its body weight, *Homo sapiens* does not have the largest relative brain weight, being outdone by several monkeys, some rodents, and even some fish. Normal modern human brain size varies between roughly 900 and 2,000 grams, although some very small number of exceptions does occur, with sizes ranging from about 750 to 900 and 2,000 to 2,200 grams. There exist both human population variation as well as differences between the sexes. In general, Arctic peoples have larger brains than those living in the tropics, and the smallest brains appear to be found among Ituri forest pygmies, also displaying small stature. Males in all populations for which good autopsy data have been gathered show brain sizes on the average of 100–150 grams greater than females, an amount roughly the same as the range of modern human racial variation. It should be pointed out that these differences, and their possible relationship to cognitive skills, are highly controversial and simple correlations are deceptive (Holloway 1996). Table 1 provides a listing of the major fossil hominid taxa and their respective brain sizes. Notice that the range of values from the earliest australopithecine to modern *Homo* is roughly 1,000 ml, or about the same amount as the normal range of variation within our species.

2.2 Encephalization (Encephalization Coefficient, E.Q.)

Nevertheless, the human animal does come out on top of the evolutionary heap when its absolute brain and body weights are considered together. When the log (base 10) of brain weight is plotted against the log 10 of body weight for a group of relevant taxa, the result is a linear line, where (log 10) brain weight = a + b (log 10) body weight. For a large array of primate data (Stephan et al. 1981), the slope of the line (b in the

Table 1
Some fossil hominid brain volumes

Group	Number	Location	Average Brain Volume	Range	Dating (MY)
A. afarensis	3	E. Africa	435	400–500	3–4
A. africanus	8	S. Africa	440	420–500+	2–3
A. aethiopicus	1	E. Africa	410	410	2.5
A. robustus	6	E. & S. Africa	512	500–530	1.6–2.0
H. rudolphensis	2	E. Africa	775	752–800	1.8
H. habilis	6	E. Africa	612	510–687	1.7–2.0
H. ergaster	2	E. Africa	826	804–848	1.6
H. erectus	2	E. Africa	980	900–1067	1.0–1.6
H. erectus	8	Indonesia	925	780–1059	1.0
H. erectus	8	China	1029	850–1225	0.6
Archaic *H. sapiens*	6	Indonesia	1148	1013–1250	0.13
Archaic *H. sapiens*	6	Africa	1190	880–1367	0.125
Archaic *H. sapiens*	7	Europe	1315	1200–1450	0.5–0.25
H. sapiens (Neand.)	25	Europe, M. East	1415	1125–1740	0.09–0.03
H. sapiens sapiens	11	World	1506	1250–1600	0.025–0.01

Source: Holloway 1997.

Table 2
Some examples of encephalization quotients

Species	Brain wt. (g)	Body wt. (g)	EQ Homo[a]	EQ Jerison[b]	EQ Primates[c]	EQ Stephan[d]
Lemur	23.3	1,400	21	1.56 (22.6)	0.94 (32.7)	5.66 (19.6)
Baboon	201	25,000	28	1.97 (28.5)	0.90 (31.3)	7.94 (27.5)
Gorilla	465	165,000	23	1.56 (22.5)	0.61 (21.2)	6.67 (23.2)
Orang	370	55,000	31	2.15 (31.1)	0.91 (31.7)	8.90 (30.9)
Chimp	420	46,000	39	2.63 (28.1)	1.81 (41.1)	11.3 (39.3)
Human	1,330	65,000	100	6.91 (100)	2.87 (100)	28.8 (100)

Source: Holloway 1997.
Formulae:
a EQ Homo = Brain wt/1.0 Body $wt^{0.64906}$, b EQ Jerison = Brain wt/0.12 Body $wt^{0.66}$, c EQ Primates = Brain wt/0.0991 Body $wt^{0.76237}$, d EQ Stephan = Brain wt/0.0429 Body $wt^{0.63}$
Note: Each formula is based on a different set of data. The EQ Homo equation simply uses the average brain and body weight for *Homo sapiens*, and assumes an intercept where both brain and body weights are zero. The value of whichever animal is calculated is then given as a direct % of modern *Homo sapiens*. EQ Jerison is based on data for almost 200 mammals, while the EQ Primates is based on Martin's (1983) data set for primates only. The EQ Stephan equation is based on insectivores only. The numbers in the parentheses are the % of the *Homo sapiens* value.

equation above) is about 0.76, and the correlation coefficient is 0.98, indicating that the relationship is almost perfect. This relationship will naturally vary depending on the databases and the transformations used. This is known as an allometric equation, and these are used frequently in biology to assess the underlying relationships between the size of parts of the body and the whole. The slope sometimes has an interpretation suggesting functional relationships between the brain and other variables. For example, in the above example, the slope is 0.76, extremely close to 0.75 or 3/4, which often describes a metabolic relationship (Martin 1983). The slope of 0.666, or 2/3, has been championed by some (e.g., Jerison 1973) as indicating an important geometric relationship between volume and surface area. It is important to realize that these slopes vary depending on the taxa examined. In general, as the taxa become more similar, the slope decreases. Species within a genus generally have a slope around 0.3; within a species the slope is smaller yet, being about 0.2, and the correlation coefficient is also reduced.

Just as the human animal is curious, it is also vainglorious, always trying to find a measure that places it at the top. Thus we can fabricate a device, the *Encephalization Coefficient*, or E.Q., which shows that relative to any database, the human animal is the most encephalized animal living. The point for *Homo sapien*s shows a clear positive residual above the expected regression line, and in fact the human value is about three times that expected for a primate with its body weight. Table 2 provides a number of different equations based on differing databases, which happily give *Homo sapiens* the highest value. (Actually, young

immature dolphins will provide a higher number, but when compared to an immature human, the value is higher in the latter.) Two additional points should be made: (a) E.Q.s are relative to the databases used, and thus there is an inherent 'relativity' to relative brain sizes; and (b) E.Q.s do not evolve, only brain weight/body weight relationships do, and E.Q.s are simply a heuristic device enabling comparisons between taxa; they have no reality outside of the database chosen, or species within a taxa, and are not designed to discuss within-species variation. For example, female humans are 'more' encephalised than males, given their smaller body sizes and smaller brains but the relationship is simply a statistical artifact with no known behavioral manifestation given their equal overall intelligence.

We will discuss somewhat later how the processes of hypertrophy and hyperplasia have been positively selected for in the course of the last 2 MY of hominid evolution. (Hypertrophy refers to increases in size of the neural components, e.g., neurons, dendritic branching, nuclei, fiber tracts; hyperplasia refers to increased production of cells through mitotic division.) It is most probably the case that these processes are controlled by regulatory genes, and one of the major differences between ourselves and our closest nonhuman primate relative, the chimpanzee (brain size = ca. 385 grams), relates to the schedules by which hyperplasia and hypertrophy are turned on and off during ontogenetic development (Holloway 1995).

2.3 Brain Organization and Reorganization

It is well known that the brains of most animals are extremely similar to each other in terms of their overall organization, by which are meant neural nuclei and fiber systems. The human animal does not appear to show any different structures when compared to Old World monkeys such as the macaque, or the Great Apes, including chimpanzee, gorilla, and orangutan. Even the neural fiber tracts that are involved in human language appear in these primates (Deacon 1997). One might ask, then, given the obvious species-specific repertoires that exist in all animals, how can these behaviors differ without differences in the underlying nervous systems? This is one of the major challenges of studying brain evolution, and in particular understanding what neural organizations account for the specificity of, say, human behavior, the ability to use language composed of arbitrary symbols. In other words, all mammals have a cerebral cortex, a thalamus, cerebellum, hypothalamus, etc., and basically these structures possess almost identical divisions of nuclei and do the same neural tasks. Clearly, brain size alone will never explain species-specific behavior, and the relationships between neural nuclei and fiber tracts will only go so far in explaining behavioral differences.

Allometric equations showing the relationship between one bodily component and the whole are instructive here. If we were to plot the logs (base 10) of primary visual cortex (PVC) against brain volume, we would find that the human PVC is 121 percent less than predicted, and similarly, the lateral geniculate nucleus of the thalamus is about 144 percent less than expected for a primate of our brain size. In contrast, if one were to plot the amount of cerebral cortex against brain weight the result is a straight line, and the human point lies almost exactly on the line. In short, the human cerebral cortex is as large as would be expected for a primate of its brain size. But do portions of the cerebral cortex vary in size between different primates? In humans, the residuals mentioned above suggest that compared to chimpanzees, the amount of PVC is significantly smaller in humans, or alternatively put, the posterior association cortex of the parietal and temporal lobes is relatively larger in humans. Since there are no essential differences between chimpanzees and humans in their visual abilities and competencies, these differences reflect selection for expanded functioning of the association cortex in humans. This is precisely what is meant by 'reorganization.'

When used in a comparative or evolutionary context reorganization means changes in the sizes and proportions thereof of neural nuclei and their fiber tracts. Given that chimpanzees and hominids last had a common ancestor some 5–7 MY, and that chimpanzees appear to have large PVC cortices, we infer that one aspect of human brain evolution has been some reorganization of the cerebral cortex, namely an increase in posterior association cortex (or, equally, a reduction in PVC) involved in polymodal cognitive tasks, where visual, auditory, and motor information are brought together in a synthetic whole. The trick, of course, is to demonstrate objectively when, where, and why these changes took place. This example of PVC has been purposefully chosen because one of the sulcal landmarks of the cortex that defines the anterior border of PVC is the 'lunate' sulcus, named for its crescentic shape, and there is some hope of identifying its position on some of the early hominid brain endocasts.

Neuroanatomists have been trying for many decades to demonstrate the major differences between us and other primates, and aside from gross brain size, very little has been shown. The frontal lobe, and particularly its prefrontal portion, has been a favorite target and indeed, Brodmann (1909) claimed it was proportionally larger in humans, a view most recently championed by Deacon (1997). Unfortunately, other works have shown that the human brain has just as much frontal lobe as would be expected for a primate of its brain weight (Semendeferi et al. 1997, Uylings and van Eden 1990), although the picture regarding prefrontal cortex has yet to be determined objectively using cytoarchitectonic criteria, which is how prefrontal cortex is differentiated from the pure motor cortex behind it. Hominid brain endocasts do not, alas, provide any sulcal landmarks with enough

reliability to determine the boundaries of prefrontal cortex, which is so important to impulse control, and higher cognitive functions such as planning and abstraction. Thus these regions cannot be measured in a phylogenetic sequence. The Neandertals, living from about 200,000 to about 28,000 years ago have frequently been described as having smaller frontal lobes; this is not based on objective measurements, but rather a perception that the large brow ridges on these humans were constraining frontal lobe development. Studying the Neandertal brain endocasts and comparing them to modern humans, I have failed to see any significant difference between these two groups.

Similarly, regions such as 'Broca's and Wernicke's areas,' anterior and posterior association cortical regions involved in motor (Broca's) and receptive (Wernicke's) aspects of speech, are not determinable on most fossil brain endocasts, although we can determine, for example, that Broca's region is more human-like on the brain cast of early *Homo*, some 1.8 MY. This is the famous KNM-ER 1470 endocast of *Homo rudolphensis* from Kenya, which had a brain volume of 752 ml, but which may not be a direct ancestor to our own line of *Homo*.

While the concept of reorganization has a heuristic value in directing our attention to changing quantitative relationships between different neural nuclei and fiber tracts, we cannot yet describe behavioral differences between closely related animals such as chimpanzee, gorilla, and orangutans, or different species of the genus Macaca, or indeed different breeds of dogs with their different temperaments, aptitudes, and sociality. We simply do not know what magic level of neural description is necessary to describe species-specific behavior. More recent research on prairie and mountain voles suggests that the difference in the females' ability to retrieve pups back to the nest depends on the distribution and number of neuroreceptors for the hormone oxytocin found in several nuclei of the brain, particularly the thalamus. Otherwise, their brains appear identical (Insel and Shapiro 1992). In addition, it is necessary to remember that the brain possesses aspects of plasticity that we did not appreciate except within the past decade, and that as the brain's organization unfolds ontogenetically, interactions with environmental stimuli are always occurring, and the brain builds its organization partly through its plasticity. It is difficult enough to study and understand such patterns in laboratory animals, let alone in our fossil ancestors! While the above suggests a somewhat pessimistic tone, we should remember that advances in noninvasive technology such as MRI, fMRI, and PET scanning have enormously increased our understanding of how the brain works, and how neural systems integrate and dissect data from the environment, always providing us with newer paradigms for further exploration about our brains and behavior, and in time, those of our closest relatives, the apes, in particular the chimpanzee.

2.4 Human Brain Asymmetry

The cerebral cortices of the human brain are usually asymmetrical, and tend to grow in a torque manner, reflecting minor differences in maturation rates. The hemispheres are seldom, if ever, equipotential in terms of functioning. Our left hemisphere is often characterized as 'analytic' and involved with language tasks, while our right hemisphere appears most competent in visuospatial integration, and is often thought of as the 'intuitive' or 'gestalt' hemisphere. These characterizations, while crude, hold up fairly accurately for right-handers and many ambidextrals. From radiographic studies, it was possible for LeMay (1976) to ascertain different petalia patterns for right- and left-handed humans with a high degree of precision. These petalias are small extensions of cerebral cortex that extend farther in one part of a hemisphere than on the other side. For example, we speak of a left-occipital right-frontal torque pattern of petalias as occurring with high frequency in right-handed individuals. This means that the left occipital lobe bulges somewhat more posteriorly on the left hemisphere while the right hemisphere is somewhat broader in width in the frontal lobe. In true left-handers, who are represented in human populations by about 8–10%, the pattern is reversed, meaning they exhibit a right-occipital left-frontal pattern. Petalia patterns for a large collection of apes indicated that while chimpanzees, gorillas, and orangutans sometimes demonstrated asymmetries, they did not show the particular torque pattern described above for humans. The gorilla, incidentally, was the most asymmetrical of the apes (Holloway and de LaCoste-Lareymondie 1982). On the other hand, brain asymmetries, particularly in the *planum temporale* (temporal cortex) of the chimpanzee, show a strong left-hemispheric size difference compared to the right (Gannon et al. 1998). This is simply puzzling as we do not have any evidence that chimpanzees use this structure in communication as do humans, and the fact that we share this difference with chimpanzees suggests that brain organizational features relating to complex cognitive functioning has been around for at least 5–7 MY. As our non-invasive scanning techniques become more sophisticated, we can expect to learn how these asymmetries function is animals other than ourselves.

Hominid brain endocasts, when complete for both sides (unfortunately, this is very rare) allow the paleoneurologist to assess the cerebral asymmetries, and indeed, even australopithecines appear to show beginnings of the right-handed torque pattern found in humans, and, as one progresses through time, the petalia patterns become more accentuated in the modern human direction. If we add to these observations those of Toth's (1985) studies on the early stone tools of about 2 MY, which strongly suggest right-handedness, this underlines the fact that our early ancestors, brains, despite their small size (sometimes

within the extant apes' range), were reorganized, and that they probably had some modes of cognition very similar to our own.

3. *Synthesis: Putting Together Size, Organization, and Asymmetry During Human Evolution*

As mentioned earlier, human brain evolution has clearly been a process of integrating neurogenetic processes that led to increased size of the brain (hyperplasia and hypertrophy) and these neurogenetic changes also played roles in the reorganization (quantitative shifts) of neural nuclei, fiber tracts, and cortical cytoarchitectonics. In addition, it is probable that other changes occurred at the neurochemical level, involving neurotransmitters and receptor sites, but these are not well known from the comparative record, let alone the fossil one. This integration was sometimes gradual, sometimes 'punctuated,' at least based on the fossil hominid record currently available. The only reliable evidence from paleoneurology suggests that Brodmann's area 17 (PVC) was reduced early in hominid evolution, signs of the reduction being clear in *Australopithecus afarensis* some 3 to 3.5 MY. While this would have meant a relative increase in posterior parietal cortex (area 39) and peri- and parastriate cortex (areas 18 and 19 respectively), the faithfulness of sulcal impressions does not allow for unambiguous definition of these areas. Similarly, it is not possible at this time to measure and delineate remaining areas of the temporal cortex and superior parietal lobule unambiguously. What is suggested, however, is that visuospatial abilities were most probably cognitively enhanced early in hominid evolution. It is not until we come to *Homo rudolphensis* at ca. 1.8 MY that a case can be made for some frontal lobe reorganization in the third inferior frontal convolution, Broca's area. Thus it would appear there was a gradient of cerebral reorganizational changes starting posteriorly, and progressing anteriorly. Table 3 outlines these changes.

Table 4 outlines the major size changes in the human brain during its evolutionary odyssey. Paleoneurological data simply are not detailed enough to integrate the two tables of size and reorganizational changes into one holistic sequence of events. Basically, the paleontological record supports an early reorganizational change resulting in an increase in posterior cortex associated with visuospatial processing, perhaps accompanied by a relative small allometric increase in brain size from *A. afarensis* to *A. africanus*. This would correlate well with geological and paleontological evidence that shows that early hominids were expanding their ecological niches and becoming more diverse in their subsistence patterns in mixed habitats. We know this based on the fact that stone tool types are becoming standardized in form, tool inventories grow larger, and right-handedness is highly probable. With the advent of *Homo*, we find strong evidence for a major increase in brain size, both allometric (related to body size) and non-allometric, and a reorganized frontal lobe, broader, and showing a more modern human-like Broca's area. This suggests that indeed there had been some strong and dramatic selection pressures for a somewhat different style of sociality, one most probably based on a primitive proto-language that had some arbitrary symboling elements as the standardization of stone tools (e.g., Acheulean handaxes) increases, suggesting social cohesion and control mediated through symbolically-based communication. Needless to say, this is but one speculative account of the evidence. But from about 1.8 to roughly 0.5 MY, we think there were minor allometric brain size increases to the earliest *Homo erectus* hominids of Indonesia and China, where brain sizes ranged from 750 to 1250 ml in volume. We have very little evidence for body sizes, but we believe, on the basis of the KNM-WT 17,000 Nariokotome youth from Kenya at ca. 1.6 MY, that these did not differ significantly from our own.

This is also a time during which cerebral asymmetries are becoming more strongly pronounced. With the advent of Archaic *Homo sapiens*, about 0.15–0.2 MY, we find brain sizes well within modern human values, and no evidence for further allometric increases, except possibly for the Neanderthal humans in which it can be argued that their larger brain and body sizes (lean body mass: bone and muscle) were adaptations to colder conditions. If further changes took place in cerebral and/or subcortical organization, they are simply not apparent from a paleoneurological perspective. Yet, the Upper Paleolithic is the time when cave art makes its appearance, and one cannot help but wonder whether the explicit use of art involving symbolization might not also have been the time for the emergence of language. In fact, however, there is nothing in the direct fossil evidence, and in particular paleoneurology, which provides any evidence for such views. Claims for a single mutation are ridiculously speculative. Finally, it would appear that there has actually been a small reduction in brain size, probably allometric in nature, from about 0.015 MY to the present.

The totality of evidence shows that the brain has always been evolving during our evolutionary journey, with myriad changes taking place at different tempos during different times. As suggested recently (Holloway 1997, p. 200):

> In sum, the major underlying selectional pressures for the evolution of the human brain were mostly social. It was an extraordinary evolutionary 'decision' to go with an animal that would take longer to mature, reach sexual maturity later, and be dependent for its food and safety upon its caretakers (parents?) for a longer period of time. The benefits for the animal were many, including a longer learning period, a more advanced, larger, and longer-growing brain, and an increasing dependence on social cohesion and tool making and tool

Table 3
Summary of reorganizational changes in the evolution of the human brain

Brain Changes	Taxon
(1) Reduction of primary visual striate cortex, area 17, and a relative increase in posterior parietal cortex	*Australopithecus africanus* *Australopithecus afarensis*
(2) Reorganization of frontal lobe (3rd inferior frontal convolution, Broca's area)	*Homo rudolphensis*
(3) Cerebral asymmetries, left-occipital right frontal petalias	Australopithecines and early *Homo*
(4) Refinements in cortical organization to a modern *Homo sapiens* pattern	*Homo erectus* to present

Source: Holloway 1997.
Note: (4) is inferred, as brain endocasts cannot provide that level of detail necessary to demonstrate the refinements in cortical organization from surface features alone.

Table 4
Brain size changes in hominid evolution

Brain changes	Taxon	Time (MY)	Evidence
1. Small increase, allometric[a]	*A. afarensis* to *A. africanus*	3.5–2.5	Brain endocast increase from 400 to 450 ml.
2. Major increase, rapid, both allometric and non-allometric	*A. africanus* to *H. habilis*	2.5–1.8	KNM–1470, 752 ml (300 ml increase)
3. Modest allometric increase in brain size to 800–1000 ml	*H. habilis* to *H. erectus*	1.8–0.5	*H. erectus* brain endocasts and postcranial bones
4. Gradual and modest size increase to archaic non-allometric	*H. erectus* to *H. sapiens neanderthalensis*	0.5–0.075	Archaic *Homo H. sapiens*, Neandertal endocasts 1200–1700 + ml.
5. Small reduction in brain size among modern allometric	*H. sapiens H. sapiens sapiens*	0.015–present	Modern endocranial volumes

Source: Holloway 1997.
a Related to increase in body size only.

using to cope with the environments that they encountered. Needless to say, language abilities using arbitrary symbol systems were an important ingredient in this evolution.
The fossil record shows us that there was a feedback between the complexity of stone tools (which must be seen as a part of social behavior) and increasing brain size and the expansion of ecological niches. The 'initial kick,' however, the process that got the ball rolling, was a neuroendocrinological change affecting regulatory genes and target tissue-hormonal interactions that caused delayed maturation of the brain and a longer growing period, during which learning became one of our most important adaptations.

These ideas have been detailed elsewhere (Holloway 1996), where more details may be found.

4. *And to the Future?*

There appear to be two common presumptions about our future brain evolution. One is that our biological evolution has stopped. The second is that our brains will continue to grow in size, with bulging frontal lobes (sort of a cross between E.T. and X-Files …), to handle our growing dependence on technology. What we have witnessed from the past fossil record is that our brains and bodies work largely in allometric fashion, and given the high metabolic cost of operating bigger brains (about 20–25% of our metabolic resources go to supporting our brains, which constitute only 2% of our total body weight), the second scenario seems highly unlikely. The first scenario is simply untrue, but it would require vast amounts of information from each generation of many living populations, something feasible perhaps, but not currently being collected. Furthermore, it is quite controversial whether brain size has any close relationship to intelligence, however intelligence is actually measured. Recent research based on MRI determinations of brain volume and selected batteries of cognitive tests have shown correlations between test scores and brain volume ranging from 0.4 to 0.6. (Andreasen et al. 1993). This is a figure significantly larger that previously reported (e.g., Van Valen 1974), and will need more replication studies. But if protein resources were to nosedive throughout the world for a significant period of time, selection would probably favor smaller body sizes in our species, and that would result in smaller brains, given an allometric relationship of

roughly 0.3 between stature and brain size. While genetic engineering may well provide some respite between the ever-increasing mass of humanity, ecological and nutritive degradement, this too is likely to be nothing more than short-term fending off of the unstoppable future. These degradations are part and parcel of the human brain's capacity to ignore warnings that should curtail greed and stupidity. The paleontological record for most mammals suggests that at the taxonomic level of the genus (such as *Pan*, *Homo*, *Canis*, *Notocherus*, etc.), one finds a recognizable record of that genus spanning approximately 5 to 10 million years. Our genus has a duration of about 2 MY. We, as a genus, despite our largish highly encephalized brains, have another 3 MY to go if we wish to be as successful in the paleontological longevity game.

See also: Body, Evolution of; Evolution of Cognition: An Adaptationist Perspective; Human Cognition, Evolution of; Intelligence, Evolution of

Bibliography

Andreasen N C, Flaum M, Swayze H V, O'Leary D S, Alliger R, Cohen G, Ehrhardt N, Yuh W T C 1993 Intelligence and brain structure in normal individuals. *American Journal of Psychiatry* **150**: 130–4

Brodmann K 1909 *Vergleichende Lokalizationzlehre der Grosshirnrinde*. J.A. Barth, Leipzig, Germany

Dart R 1925 Australopithecus africanus: The man-ape of South Africa. *Nature* **115**: 195–9

Deacon T 1997 *The Symbolic Species: The co-evolution of language and the brain*. Norton, New York

Gannon P J, Holloway R L, Broadfield D C, Braun A R 1998 Asymmetry of chimpanzee planum temporale: Humanlike pattern of Wernicke's brain language area homolog. *Science* **279**: 220–2

Holloway R L 1984 The taung endocast and the lunate sulcus: A rejection of the hypothesis of its anterior position. *American Journal of Physical Anthropology*. **64**: 285–7

Holloway R L 1995 Toward a synthetic theory of human brain evolution. In: Changeux J P, Chavaillon J (eds.) *Origins of the Human Brain*, Clarendon Press, Oxford, UK, pp. 42–54

Holloway R L 1996 Evolution of the human brain. In: Lock A, Peters C (eds.) *Handbook of Human Symbolic Evolution*. Oxford University Press, New York, Chap. 4, pp. 74–116

Holloway R L 1997 Brain evolution. In: Dulbecco R (ed.) *Encyclopedia of Human Biology*, Academic Press, New York, Vol. 2, pp. 189–200

Holloway R L 2000 Brain. In: Delson E, Tattersall I, Van Couvering J, Brooks A S (eds.) *Encyclopedia of Human Evolution and Prehistory*, 2nd edn, Garland Publishing, New York, pp. 141–9

Holloway R L, de LaCoste-Lareymondie M C 1982 Brain endocast asymmetry in pongids and hominids: some preliminary findings on the paleontology of cerebral dominance. *American Journal of Physical Anthropology* **58**: 101–10

Insel T, Shapiro L E 1992 Oxytocin receptors and maternal behavior. *Annals of the New York Academy of Sciences* **652**: 448–51

Jerison H J 1973 *Evolution of Brain and Intelligence*. Academic Press, New York

LeMay M 1976 Morphological cerebral asymmetries of modern man, fossil man, and nonhuman primates. *Annals of the New York Academy of Sciences* **280**: 349–66

Martin R D 1983 *Human Evolution in an Ecological Context*. James Arthur lecture (1982). American Museum of Natural History, New York

Semendeferi K, Damasio H, Frank R, Van Hoesen G W 1997 The evolution of the frontal lobes: A volumetric analysis based on three-dimensional reconstructions of magnetic resonance scans of human and ape brains. *Journal of Human Evolution* **32**: 375–88

Stephan H, Frahm H, Baron G 1981 New and revised data on volumes of brain structures in insectivores and primates. *Folia Primatologia* **35**: 1–29

Tobias P V 1971 *The Brain in Hominid Evolution*. Columbia University Press, New York

Toth N 1985 Archaeological evidence for preferential right-handedness in lower and middle Pleistocene, and its behavioral implications. *Journal of Human Evolution* **14**: 607–14

Uylings H B M, van Eden C G 1990 Qualitative and quantitative comparison of the prefrontal cortex in rats and primates, including humans. *Progress in Brain Research* **85**: 31–62

Van Valen L 1974 Brain size and intelligence in man. *American Journal of Physical Anthropology* **40**: 417–24

R. Holloway

Brain Implants and Transplants

The concept of brain repair with cells or tissue transplants is immensely attractive. After all, tissue and organ transplantation has progressed significantly over the past 30 years, and allowed extended life to patients with deadly, and previously incurable, diseases of the heart, lungs, liver, kidneys, and other organs. Transplantation to the brain, however, presents an altogether different, complex series of problems. Even with our rapidly evolving knowledge about the brain, its intricate neural networks, sensitivity to injury, and delicate chemical balance, make this most complex organ so much more difficult to treat.

These problems notwithstanding, our understanding of the brain is advancing. It is no longer considered a stable, unchanging organ in the adult and the long-held dogma that repair and regeneration of the brain cannot take place is wrong. Trophic factors, neural transplants, and stem cell grafts are just a few of the many ways by which the central nervous system can be modified.

1. How do Transplants Work?

There are several ways in which implants or transplants can work in the central nervous system: (a) provide replacements for lost cells, with integration

into the brain circuitry and reconstruction of the damaged brain; (b) function as a biological mini-pump—a way to deliver desired substances locally, avoiding the blood–brain barrier (BBB) and side effects; or (c) supply trophic signals and metabolic support for existing neurons, thus providing stimuli for neuronal survival and regeneration. While the first mechanism of action requires a neural transplant, the other two do not.

Thus, direct transplants of neurons into the brain are only one possible therapeutic strategy. Other types of implants are possible, which may avoid some of the problems associated with fetal cell transplantation for neurodegenerative diseases, or offer new therapeutic strategies for other conditions.

2. *Parkinson's Disease*

Successful grafting of brain tissue has been reported since the late nineteenth century. The modern era of neural transplantation, however, can be traced to the work of several pioneers who started their work before 1970. Parkinson's disease (PD) was the first clinical target for which transplantation to the central nervous system was tried. This is a common chronic degenerative disorder, affecting about one million people in the USA alone. A relatively small group of cells in the area of the brain called the *substantia nigra*—'the dark substance'—undergoes a gradual degeneration, until the vast majority of the neurons in that area is lost. These cell bodies produce the neurotransmitter dopamine, which is secreted at the terminals of these neurons located in a brain region called the striatum. Once the levels of dopamine in the striatum decreases by 80 percent, Parkinsonian symptoms occur. The patients suffer from a chronic deterioration, with characteristic tremor and slow movements progressing to rigidity, inability to initiate movements, and eventually complete disability and premature death. Medical treatment is available, but is far from perfect. L-DOPA in various forms increases the availability of dopamine and can dramatically ameliorate the symptoms for five to seven years. However, this drug does not forestall the disease process or protect the neurons of the *substantia nigra*. PD became the first target for trials of neural transplantation because in this disease major symptomatology can be traced to the loss of a single neurotransmitter, and this neurochemical system has a well-defined target. Also, excellent animal models of PD exist.

2.1 *Rodent Models*

Critical to the development of the field is the demonstration that transplantation into the brain is not only technically feasible, but also has functional effects. For the neural grafting field to develop, good animal models were required. Ungerstedt (1968) developed a unilateral 6-hydroxydopamine (6-OHDA) model of PD in which rats received lesions of the nigrostriatal system on one side. This resulted in the animals spontaneously rotating to the side of the lesion. The number of these rotations would increase in the presence of amphetamine that increases dopamine secretion from presynaptic terminals. Further, rotations occur in the opposite direction if the animals receive the dopaminergic receptor agonist apomorphine.

One of the most influential early papers demonstrating the potential of neural grafting was co-authored by Stenevi and Bjorklund (1979) who grafted into the rat brain dopaminergic cells obtained from various peripheral nervous system and central nervous system sources. Adrenal tissue, cervical ganglia, ventromedial mesencephalon, dorsolateral pons, median pontine raphe, from fetal, newborn, and adult rats, were transplanted into different CNS locations. These experiments served to elucidate the specific parameters critical for predicting graft viability. In 1979, using the rotational model, Bjorklund and Stenevi (1979) from Lund University in Sweden, as well as Freed (1980), Hoffer (1983), and their co-workers from the National Institute of Health in the USA, demonstrated that the experimental disease could be ameliorated by transplantation of fetal cells taken from a specific area of the brain—the ventral mesencephalon (see also Perlow 1979). The rotational imbalance induced by the unilateral 6-OHDA lesion was completely reversed by the graft, and recurred when the graft was destroyed.

Following these initial landmark demonstrations, a series of logical experiments in rodent models was pursued to elucidate the parameters required for optimal transplantation. These included the appropriate quantity of transplanted tissue, number of transplants, the age of the donor fetus, and the preferred target area. Issues of graft survival were critical—both keeping the donor tissue alive until the implantation, and survival within the host, where the graft has to obtain blood supply and avoid rejection by the host immune system. For ventral mesencephalic neurons, the optimal donor age in the rat is constrained to a specific time window of two to three days. The grafts appeared to survive especially well when transplanted directly into the striatum, their usual target area, where they integrated with the host brain and secreted dopamine. In the earliest experiments, a cavity was created in the recipient's cortex, and a solid graft of embryonic rat ventral mesencephalon was placed on the dorsal surface of the striatum. This method provided limited recovery of function because the solid graft was able to only integrate with the dorsal part of the target. To circumvent this problem, Dunnett et al. (1981) developed a cell suspension grafting technique, in which trypsin was used to dissociate a solid graft into a single cell or aggregate

suspension. This allowed stereotactic placement of the cells into several locations deep within the striatum. This resulted in a more robust behavioral recovery. The time between tissue dissection and graft is a crucial factor. Graft viability *in vitro* prior to implantation depends on several factors, including the handling of the suspension following trypsin treatment. Studies were undertaken to establish the minimal number of surviving grafted neurons required in order to produce a functional effect in the rat model. Brundin et al. (1985) established that 120 grafted dopamine neurons, or about 2 percent of the normal number of dopamine neurons in the rat nigrostriatal system, are needed to mediate a functional response. The grafted cells consistently reinnervated the host brain, but the fiber outgrowths radiated only 1–2 mm from the site of the transplant. Multiple grafts were therefore needed to achieve good coverage of the striatum. The grafts were innervated by host neurons. The tonic secretion of dopamine was established, and the electrophysiological and metabolic properties of the grafted cells were equal to those of normal dopaminergic cells in the *nigra*.

Rejection by the host immune system was not a severe obstacle. Allografts (grafts between individuals of the same species) and even xenografts (cross-species transplants) survive much better in the brain compared to other organs. In many respects, the brain is a relatively 'immunologically privileged site.' This is because neural cells lack MHC antigens. Further, the brain has no lymphatic drainage system and is protected in part by the blood–brain barrier. The blood–brain barrier tends to keep cells of the immune system out of the brain, and some of the astrocytes (an important supporting cell type in the brain) actually secrete substances that suppress the immune response. Nonetheless, trauma can activate the immune system and abolish the relative tolerance.

Although the studies in the rodent model established the feasibility and basic parameters of neural grafting, it is very difficult to translate results from rodent experiments directly into human patients. Our brains are vastly larger, there are differences in anatomy and embryology, and it is difficult to model complex behaviors in rodents. Some stepping stones were needed to cross this wide gap.

2.2 Primate Models

In California, several cases of rapid onset Parkinsonism were seen among intravenous narcotic addicts. Their doses of an illicit synthetic drug contained n-methyl 4 phenyl, 1,2,3,6 tetrahydropyridine (MPTP), which was later determined to be a potent toxin that destroys the dopaminergic neurons of the *substantia nigra*. The patients developed a syndrome that resembled advanced Parkinson's disease. The discovery of MPTP provided a means by which an animal model of the disease could be established. Interestingly, there are species differences in the sensitivity of an organism to the toxic effects of MPTP. Monkeys are perhaps the most sensitive. When MPTP is given to monkeys, they display many of the cardinal clinical features of Parkinson's disease and pathologically display lesions of the nigrostriatal system. Transplantation in the monkey model confirmed many of the findings in the rat model. The importance of appropriate cell source and target area was confirmed. Furthermore, recovery of symptoms such as rigidity, bradykinesia, and tremor was demonstrated. Functional recovery was long-term (see Sladek et al. 1988, Bankiewicz et al. 1990). In most studies, integration of the graft with the host brain was observed. Sladek and others concluded that less than optimal graft placement and creation of a cavity for the graft predisposed to glial scar formation and can inhibit fiber outgrowth. Nonhuman primate models employing focal and systemic administration of MPTP had a high predictive value as to what will work in humans. Monkeys have a relatively large brain, which closely resembles the human brain anatomically. Moreover, more complex behaviors could be demonstrated in monkeys compared with rodents (see Bakay et al. 1987).

2.3 Clinical Studies

Even though scientific data clearly favored fetal tissue as the best donor source for transplantation, adrenal medullary autografts were the first donor tissue tried clinically in patients with PD. In animals, adrenal medullary tissue had been shown to poorly survive in rats and monkeys and only elicit modest functional effects. However, use of this cell type did not induce the same social and ethical debate that fetal tissues did. Patients initially received adrenal medullary autografts, (tissue from the same organism). In these studies, solid grafts of adrenal medulla were transplanted into a cavity in the brain of the recipient. After the 1987 report of success in the study from Mexico, further clinical trials were disappointing, because the grafts survived very poorly if at all. A more gradual approach was needed, making use of the advances made in animal studies.

The next clinical trials were based on the success of the fetal cell suspension transplantation paradigm. Several groups, from Sweden, the UK, the USA, and France conducted carefully monitored studies of human fetal mid-brain cell transplantation into patients with Parkinson's disease. The reports from the different teams were mixed. Whereas some investigators reported marked improvement, others could not demonstrate a significant positive effect. For example, the Yale team reported no change in the condition of the patients after transplantation. They transplanted cryopreserved cells from a single fetus to each patient. An autopsy case from their series failed to demonstrate

surviving cells. The Lund University program in Sweden started clinical transplantation in 1984. They used cells from three to seven fetuses (six to nine weeks gestational age) per patient. From three to seven passes of the stereotactic needle for each treated side were required. Results showed marked variability between patients. The researchers used a positron emission tomography (PET) scan to demonstrate high metabolism of dopamine precursors in the area of the graft. Importantly, all cases demonstrated PET scan evidence of graft survival, and some of the patients were followed up for more than 10 years with continuing benefit from the transplantation. Similar outcomes were seen in patients from the University of South Florida in Tampa study (see Freeman et al. 1998). Clinical improvement was documented in all of the first six patients. These patients were transplanted with cells from six to eight fetuses per patient; the cells had been kept in cool hibernation medium for up to two days. While the patients still needed their medication (L-DOPA), all of them were more active, with fewer signs of Parkinson's disease. PET scans suggested high dopamine metabolism in the grafts. This activity persisted and even increased with time. After one of these patients died from an unrelated problem 18 months after the transplant procedure, his brain was studied. The graft indeed survived, was integrated into the brain, and was producing dopamine. There were no signs of rejection, even though immunosuppressant medications were stopped six months after the transplantation, a year before the patient died.

These results of preliminary clinical trials established that neural transplantation could be a valid therapeutic option for a human neurodegenerative disease. Also, the predictive value of animal models, particularly the monkey MPTP model, was confirmed. The methodology of clinical neural transplantation was refined. Nevertheless, even when good graft survival was demonstrated, none of the patients was completely cured. The variability between the studies and between patients in the same study was significant. Several controlled randomized studies are going on to determine the efficacy of transplantation to treat Parkinson's disease. More research is also required in order to further improve functional outcomes.

3. *Huntington's Disease*

Based on the experience gained with PD, another devastating brain disease was approached—Huntington's disease (HD)—a severe genetic disorder of the brain. It is inherited in an autosomal dominant fashion, meaning that the children have a 50: 50 chance of inheriting the disease from an affected parent. Most patients are normal until the fourth decade of life, when a relentless cognitive, psychiatric, and motor deterioration starts. The disease primarily affects the striatum—part of the basal ganglia. The largest neuron population in the striatum—GABA-ergic inhibitory neurons—degenerate first. The most prominent feature is a severe chorea, manifesting as involuntary jerking of the body and extremities, which after several years is replaced by rigidity and total disability. Death comes after an average of 15–17 years. There is no treatment.

The first step to human transplantation for HD was to develop an animal model. Although no perfect model has yet been found, there are several methods to produce relatively selective damage to the cell population that is typically affected in HD. Importantly, there are differences between PD and HD. In PD, only the tonic secretion of dopamine needs to be reestablished; the transplanted cells function as little more than biologic minipumps. In HD, however, the transplants have to reconstitute the host circuitry in order to achieve their full effect.

A logical series of investigations, in rodent and then primate models of HD, has proven that grafts of fetal striatum (the affected part of the basal ganglia) survive, reinnervate the host, and restore function. The first reports of successful fetal striatal grafts in a rat model of HD were by Deckel et al. (1983) and by Isacson et al. (1986). In these studies, rats received injections of fetal striatal cells bilaterally into previously lesioned striata. The grafts survived well and partially corrected the behavioral impairments seen in this model, such as nocturnal hyperactivity, cognitive impairments on a T-maze test, and abnormal responses to amphetamine and apomorphine. Further studies delineated the parameters for successful transplantation, including the optimal timing of the transplantation, the appropriate age of the donor fetus, and the precise source of the transplanted fetal tissue.

Studies in nonhuman primate models revealed many similarities between fetal allografts and normal neonatal striatal development. Primate models also allowed the study of the behavioral outcomes and the use of computerized axial tomography (CAT) scans, positron emission tomography (PET) and magnetic resonance imaging (MRI) to obtain information for comparison with the human disease (Helm et al. 1992, Peschanski et al. 1995). Several studies of fetal cell transplantation for HD are now underway. Initial reports have been encouraging (Philpott et al. 1997).

4. *Challenges on the Road to Large-scale Clinical Application*

There are reasons for cautious optimism regarding the potential to successfully treat PD and HD with fetal cell transplants. However, in the present form, this treatment in not ready for large-scale human application. Currently, transplants are only performed in several centers around the world, by highly specialized teams as part of clinical studies. The experience gained

so far provides 'proof of concept,' but the reproducibility and the therapeutic value will need to be better established before the methodology can spread beyond a few academic centers.

There are several challenges that are facing the field of neural transplantation in the twenty-first century. The methods of transplantation are still developing. Research is continuing to establish the optimal anatomic location of the implant for each condition, the correct amount of cells to transplant, and the appropriate timing of transplant therapy for optimal effect in the human patient. Easily reproducible methods are needed to ensure correct identification of the grafted cells.

Another challenge is assuring maximal graft survival and integration into the host brain. The transplanted tissue is subject to stresses such as lack of nutrients, oxidative stress, and lack of trophic supportive signals. Although neural grafts tend to be rejected less than do grafts of other tissues, this 'immunological privilege' is only partial. Thus, allografts and particularly xenografts can be rejected.

The greatest obstacle to large-scale clinical implementation of neural transplantation is probably the source of the cells: six to eight spontaneously aborted fetuses are needed for each successful treatment with fetal brain cells. In many countries, any use of aborted fetuses and abortion itself are highly controversial. Therefore, in the future alternative strategies must evolve. The ideal source of cells for transplantation should be available, reliable, survive well after grafting, and demonstrate functional effects.

5. *Sources of Cells*

5.1 *Embryonal Stem (ES) Cells*

There are a number of potential sources of cells for transplantation. Rapid advances in our ability to grow and manipulate stem cells in culture makes this technology a highly promising source of cells. Stem cells can provide a limitless source of neural cells for grafting. However, with human ES cells some ethical and religious issues may arise. Research on human ES cells is gradually gaining acceptance and legal support. A law was passed in the UK which allows the use of human ES cells for research, and Pope John Paul II, when addressing the 18th International Society for Transplantation meeting in Rome, expressed the approval of the Roman Catholic Church for ES research.

5.2 *Immortalized Cells*

The use of adult-derived, immortalized cell lines may circumvent the use of embryonal cells. These cell lines can be derived from rodent neuroepithelial stem or progenitor cell population, or even from human neuroepithelial tumor cell lines. When implanted into the rodent brain, these cells differentiate into both mature neurons and glia. Moreover, the implants take on the characteristics of the surrounding brain tissue—resembling cortex, cerebellum, or hippocampus when transplanted into each of these areas, respectively (Borlongan et al. 1998).

5.3 *Xenografts*

The use of xenografts may obviate the need to use any human tissue, and may be readily and cheaply available. A number of preclinical studies have demonstrated successful xenografts from rat to mouse, rat to monkey, human to rat, and monkey to human. Because the brain tends to reject grafts less than do other organs, transplantation of porcine fetal neurons into the human brain may be possible, with adequate immunosuppression. Still, lifelong immunosuppression carries significant risks. Another challenge is the risk of transmission of retroviruses from other species to the human population. Pigs have relatively few retroviruses in their genome compared with primates, yet the porcine retrovirus PERV is common enough to cause concern. Despite this controversy, a clinical study with fetal porcine neural transplants was initiated in 12 patients with PD. Initial results showed a very limited clinical response, with no evidence of infection with PERV. An autopsy case from this trial revealed very few surviving cells. A similar study in 12 patients with HD showed a favorable safety profile, but no clinical improvement after one year (Fink et al. 2000). Major challenges remain to the routine use of xenotransplantation for neurodegenerative diseases.

5.4 *Polymer-encapsulated Grafts*

Xenografts can be placed into polymer capsules which immunoisolate the cells and prevent them from being rejected. The membrane is made of a polymer (PAN-PVC or others) and engineered to have a specific morphology. It allows nutrients and active substances to pass freely, but blocks cells and large molecules such as immune globulins. Encapsulating the transplanted cells allows expansion to the sources of cells for transplantation-xenografts, which may be performed without immunosuppression. This type of graft may be utilized when continuous local delivery of active substances such as dopamine is desired. The implant can be retrieved if side effects occur (Emerich 1997). Encapsulated grafts producing dopamine resulted in behavioral recovery in the rat model of PD. Since the gene for HD is known and can be tested for, there is an opportunity to try a preventive strategy. The onset of the disease may be delayed by delivery of trophic factors, which support the cell population at risk. Experiments with encapsulated grafts in a monkey model of HD demonstrated neuroprotection by the graft. The neuroprotection strategy may be

applied to other degenerative diseases. The principle of this intervention is delivery of a trophic factor known to support the brain structure or cell population most at risk in that disease. Nerve growth factor (NGF) was used in models of Alzheimer's disease (AD), and ciliary neurotropic factor (CNTF) was utilized in preclinical and clinical studies in the treatment of amyotrophic lateral sclerosis (ALS).

5.5 Other Sources

Research of sources of cells and tissue for transplantation into the brain is very active. New directions are sought and old ideas are revisited in the search for treatments for the most devastating diseases of the brain. We can only mention some of the promising developments:

(a) Stem cells derived from adult rat bone marrow appeared to differentiate into macroglia and to improve outcomes in a rat model of stroke.

(b) Autologous dopaminergic cells may still be used to treat Parkinsonism if their function and survival can be improved. The carotid body is a source of dopaminergic cells, and some graft survival was reported after transplantation of these cells into rhesus monkeys.

(c) Neural progenitor cells are further along their development into mature neurons and glia than stem cells. Intensive research is going on to improve our ability to manipulate these cells and use them in transplantation.

(d) More than one type of cell can be transplanted together. Co-grafting of neural cells with various types of supporting cells may increase graft survival. Co-transplantation of neurons with Schwann cells from peripheral nerve in rodent models of spinal cord injury may enhance graft survival. Sertoli cells from the testis are known to secrete substances ranging from growth factors to immunomodulatory factors and may provide nutrients to the neighboring neurons. Co-grafting with Sertoli cells enhanced survival of dopaminergic neurons. Similar beneficial effects were seen when fetal kidney cells were used (Granholm et al. 1998, Hoffer and Olson 1997).

Interestingly, even grafts of supporting cells alone can be beneficial. Local delivery of a variety of trophic factors may account for this effect. Thus, whether we transplant neurons or not will depend on what we want to achieve with the graft, or on our treatment goal.

6. Other Implant Strategies

6.1 Gene Therapy and Brain Implants

Neuroprotection and other treatment goals may be achieved by cell implants combined with gene therapy. Direct implants of viral and other vectors are possible, in order to stimulate the host cells to improve or restore lost function. In a recent study, implants of a modified lentivirus (a human retrovirus modified so that it cannot replicate *in vivo*), engineered to carry a gene for human glial-derived neurotrophic factor (GDNF) were injected into the striatum of rhesus monkeys. Robust transduction of striatal cells with the gene was seen. The majority of the cells incorporating the gene were neurons. GDNF production and increased activity of dopaminergic neurons were seen both in aged monkeys with naturally low function of the dopaminergic system and in monkeys which received MPTP (Kordower et al. 2000).

Gene transfer may be combined with cell transplantation or used to enhance the success of other types of transplants. Adrenal medullary tissue is a source of dopaminergic cells, but the grafts displayed very poor survival and no significant therapeutic benefit was derived. Transfection of these cells with genes for trophic factors may enhance their survival.

All types of gene therapy are still experimental. Concerns include long-term safety and stability of the constructs. Expression of transgenes tends to diminish over time. In order to be useful for chronic human illness, extended expression and therapeutic effect have to be demonstrated.

7. Other Diseases and Conditions

So far, we have concentrated on two degenerative diseases of the brain. These were the first conditions in which experience was gained with cell transplantation in humans. In Parkinson's disease a clear benefit and success of neural transplantation was demonstrated. However, as the sources of cells and tissue for transplantation expand, and as new methods develop, transplantation research is expanding to include various conditions in the central nervous system (CNS). Among these are degenerative diseases such as AD and ALS, but also brain injury and stroke, and spinal cord injury. These are common conditions, and the leading causes of disability in the developed world.

7.1 Transplantation for Stroke

Nishino et al. (1993), and Grabowski et al. (1993) demonstrated successful transplantation of fetal cells in a rat model of stroke, with improvement in function both with a striatal and a cortical graft. More recent studies (Borlongan et al. 1998) show similar functional improvement with transplantation of a human immortalized neural cell line in a rat model of striatal infarction. The rationale for these studies was based on the success of transplantation for striatal damage in animal models of PD and HD. Local secretion of trophic factors and neurotransmitters probably played a role in the rapid functional gains. Further extensive

experimentation in rodent and nonhuman primate models is needed in order to understand the neurophysiology of transplantation for stroke and learn the parameters for successful transplantation.

7.2 Transplantation for Chronic Pain

The most common goal of neural transplantation is restoration of function in conditions where irreversible loss of neurons has occurred. However, if we consider that transplanted cells can deliver multiple beneficial substances locally, other medical problems become amenable to treatment with cell transplantation. Chronic pain syndromes are very common. Back pain alone permanently disables eight million Americans. Adrenal medullary transplants produce a variety of neurotrophic factors in addition to dopamine and opioid peptides. In the human spinal cord, endogenous opioids (endorphins) block pain sensation and dopamine potentiates opioid action. Experiments in animals and in a small number of human subjects demonstrated moderate to good improvement in pain after adrenal medullary allografts into the spinal canal (Sagen 1998). Encapsulated chromaffin cell grafts were also tried and may provide a similar effect without inducing an immune response (Aebischer et al. 1994).

8. The Future of Transplantation in the Brain

Cell transplantation and related implant strategies into the brain are constantly developing, expanding both the potential sources of cells for transplantation and the range of conditions that may be treated by these methods. This growth parallels our understanding of neuropathophysiology and human genetics, and our ability to manipulate mammalian cells in culture. Cell transplantation and other therapeutic biological implants have the potential to treat some of the most common and devastating diseases. However, even for Parkinson's disease, the technology has not yet advanced beyond experimental use.

Treatment of all types of brain injury with transplantation will require a better understanding of the conditions required for regeneration in the brain as well as further advances in our understanding of the pathophysiology of these diseases. It is likely that future transplantation efforts will make use of a combination of strategies under development at the beginning of the twenty-first century. For example, stem cells may be transplanted with peripheral cells (co-transplantation), and with a vector of gene therapy to improve the local conditions for survival and connectivity.

The long-term outlook for transplantation in the CNS primarily depends on our ability to improve the convenience, reliability, and availability of cells for transplantation, and on the efficacy of transplantation to correct functional deficits. Ultimately, the development of this field parallels advances in our knowledge of the basic mechanisms of cell development and differentiation.

See also: Brain Damage: Neuropsychological Rehabilitation; Neural Plasticity.

Bibliography

Aebischer P, Buchser E, Joseph J M, de Tribolet N, Lysaght M, Rudnick S 1994 Transplantation in humans of encapsulated xenogenic cells without immunosuppression: a preliminary report. *Transplantation* **58**: 1275–77

Bankiewicz K S, Plunkett R J, Jacobowitz D M, Porrino L, di Porzio U, London W T, Kopin I J, Oldfield E H 1990 The effect of fetal mesencephalon implants on primate MPTP-induced Parkinsonism. *Journal of Neurosurgery* **72**: 231–44

Bakay R A E, Barrow D L, Fiandaca M S, Iuvone P M, Schiff A, Collins D C 1987 Biochemical and behavioral correction of MPTP Parkinson-like syndrome by fetal cell transplantation. *Annals of the New York Academy of Sciences* **495**: 623–40

Borlongan C V, Tajima Y, Trojanowski J O, Lee V M, Sanberg P R 1998 Transplantation of cryopreserved human embryonal carcinoma-derived neurons (NT2N cells) promotes functional recovery in ischemic rats. *Experimental Neurology* **149**(2): 310–21

Bjorklund A, Stenevi U 1979 Reconstruction of the nigrostriatal dopamine pathway by intracerebral nigral transplants. *Brain Research* **177**: 555–60

Bjorklund A, Stenevi U, Dunnett S B, Iversen S D 1981 Functional reactivation of the deafferented neostriatum by nigral transplants. *Nature* **289**: 497–99

Bjorklund A, Lindvall O 2000 Cell replacement therapies for central nervous system disorders. *Nature Neuroscience* **3(6)**: 537–44

Brundin P, Barbin G, Isacson O, Mallat M, Chamak B, Prochiantz A, Gage F H, Bjorklund C 1985 Survival of intracerebrally grafted rat dopamine neurons previously cultured in vitro. *Neuroscience Letters* **61**(1–2): 79–84

Deckel A W, Robinson R G, Coyle J T, Sanberg P R 1983 Reversal of long-term locomotor abnormalities in the kainic acid model of Huntington's disease by day 18 fetal striatal implants. *European Journal of Pharmacology* **93**: 287–8

Dunnet S B, Bjorklund A, Stenevi J, Iverson S D 1981 Behavioral recovery following transplantation of substantia nigra in rats subjected to 6-OHDA lesions of the nigrostriatal pathway: II. Bilateral lesions. *Brain Research* **14** 229(1): 209–17

Dunnett S B, Boulton A A, Baker G B (eds.) 2000 *Neural Transplantation Methods*. Humana Press, Totowa, NJ

Emerich D F, Dean R L III, Sanberg P R (eds.) 2000 *Central Nervous System Diseases—Innovative Animal Models from Lab to Clinic*. Humana Press, Totowa, NJ

Emerich D F, Winn S R, Hantraye P M, Peschanski M, Chen E Y, Chu Y, McDermott P, Baetge E E, Kordower J H 1997 Protective effect of encapsulated cells producing neurotrophic factor CNTF in a monkey model of Huntington's disease. *Nature* **386**: 395–9

Fink J S, Schumacher J M, Ellias S L, Palmer E P, Saint-Hilaire M, Shannon K, Penn R, Starr P, VanHorne C, Kott H S, Dempsey P K, Fischman A J, Raineri R, Manhart C, Dinsmore J, Isacson O 2000 Porcine xenografts in Parkinson's disease and Huntington's disease patients: Preliminary results. *Cell Transplant* **9**: 273–8

Freed W J, Perlow M J, Karaim F, Seiger A, Olson L, Hoffer B J, Wyatt R J 1980 Restoration of dopaminergic function by grafting of fetal rat substantia nigra to the caudate nucleus: long term behavioral, biochemical and histochemical studies. *Annals of Neurology* **8**(5): 510–19

Freeman T B, Olanow C W, Hauser R A, Nauert G M, Smith D A, Borlongan C V, Sanberg P R, Holt D A, Kordower J H, Vingerhoetz F J 1995 Bilateral fetal nigral transplantation into the post commissural putamen in Parkinson's disease. *Annals of Neurology* **38**(3): 379–88

Freeman T B, Widner H 1998 *Cell Transplantation for Neurological Disorders*. Humana Press, Totowa, NJ

Grabowski M, Brundin P, Johansson B B 1993 Functional integration of cortical grafts placed in brain infarcts of rats. *Annals of Neurology* **34**(3): 362–8

Granholm A C, Henry S, Hebert M A, Eken S, Gerhardt G A, van Horne C 1998 Kidney co-grafts enhance fiber outgrowth from ventral mesencephalic grafts to the 6-OHDA lesioned striatum and improve behavioral recovery. *Cell Transplant* **7**: 197–212

Helm G A, Palmer P E, Simmons N E, DiPierro C, Bennett J P Jr 1992 Descriptive morphology of developing fetal neostriatal allografts in the rhesus monkey: A correlated light and electron microscopic Golgi study. *Neuroscience* **50**: 163–79

Hoffer B, Freed W, Olson L, Wyatt R J 1983 Transplantation of dopamine-containing tissues to the central nervous system. *Clinical Neurosurgery* **31**: 404–16

Hoffer B, Olson L 1997 Treatment strategies for neurodegenerative diseases based on trophic factors and cell transplantation techniques. *Journal of Neural Transmission* **49**: 1–10

Isacson O, Dunnett S B, Bjorklund A 1986 Graft-induced behavioral recovery in an animal model of Huntington's disease. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 2728–32

Kordower J H, Freeman T B, Snow B J, Vingerhoets F J G, Mufson E J, Sanberg P R, Hauser R A, Smith D A, Nauert G M, Perl D P, Olanow C W 1995 Neuropathological evidence of graft survival and striatal reinnervation after the transplantation of fetal mesencephalic tissue in a patient with Parkinson's disease. *New England Journal of Medicine* **332**: 1118–24

Kordower J H, Emborg M E, Bloch J, Ma S Y, Chu Y, Leventhal L, McBride J, Chen E Y, Palfi S, Roitberg B Z, Brown W D, Holden J E, Pyzalski R, Taylor M D, Carvey P, Ling Z, Trono D, Hantraye P, Deglon N, Aebischer P 2000 Neurodegeneration prevented by lentiviral vector delivery of GDNF in primate models of Parkinson's disease. *Science* **290**(5492): 767–73

Langston J W, Ballard P, Tetrud J W, Irwin I 1983 Chronic Parkinsonism in humans due to a product of meperidine-analog synthesis. *Science* **219**: 979–80

Nishino H, Aihara N, Czurko A, Hashitani T, Isobe Y, Ichikawa O, Watari H 1993 Raconstruction of GABAergic transmission and behaviour by striatal cell grafts in rats with ischemic infarcts in the middle cerebral artery. *Journal of Neural Transplantation and Plasticity* **4**(2): 147–55

Perlow M J, Freed W J, Hoffer B J, Seiger A, Olson L, Wyatt R J 1979 Brain grafts reduce motor abnormalities produced by destruction of nigrostriatal dopamine system. *Science* **204**(4393): 643–7

Peschanski M, Cesaro P, Hantraye P 1995 Rationale for intrastriatal grafting of striatal neuroblasts in patients with Huntington's disease. *Neuroscience* **68**: 273–85

Philpott L M, Kopyov O V, Lee A J, Jacques S, Duma C M, Caine S, Yang M, Eagle K S 1997 Neuropsychological functioning following fetal striatal transplantation in Huntington's chorea: Three case presentations. *Cell Transplant* **6**: 203–12

Sagen J 1998 Transplantation strategies for the treatment of pain. In: Freeman T B, Widner H (eds.) *Cell Transplantation for Neurological Disorders*. Humana Press, Totowa, NJ

Sladek J R, Redmond D E Jr, Collier T J, Blount J P, Elsworth J D, Taylor J R, Roth R H 1988 Fetal dopamine neural grafts: Extended reversal of methylphenyltetrahydropyridine induced Parkinsonism in monkeys. *Prog. Brain Res.* **78**: 497–506

Ungerstedt U 1968 6-Hydroxydopamine induced degeneration of central monoamine neurons. *European Journal of Pharmacology* **5**: 107–10

B. Z. Roitberg and J. H. Kordower

Brain: Response to Enrichment

Can experience produce measurable changes in the brain? The hypothesis that changes occur in brain morphology as a result of experience is an old one. In 1815 Spurzheim asked whether organ size could be increased by exercise. He reported that the brain as well as muscles could increase with exercise 'because the blood is carried in greater abundance to the parts which are excited and nutrition is performed by the blood.' In 1874 Charles Darwin mentioned that the brains of domestic rabbits were considerably reduced in bulk in comparison with those from the wild because, as he concluded, these animals did not exert their intellect, instincts, and senses as much as did animals in the wild. However, it was not until the 1960s that the first controlled studies in animals demonstrated that enriching the environmental condition in which they were confined could alter both the chemistry and anatomy of the cerebral cortex and, in turn, improve the animals' memory and learning ability.

In these early experiments only the brains of young animals were studied. Although many were impressed to learn that the cerebral cortex could increase its thickness in response to enriched living conditions, they raised the question about whether enrichment might similarly affect older animals. Once middle-aged rats' brains showed positive responses to enrichment, the next step was to experiment with very

old animals. Once again, increases in cortical thickness were found. It then became important to discover what was responsible for these changes.

One step at a time, the level of morphological changes—from neuronal soma size, to number and length of dendrites, to types and numbers of dendritic spines, to synaptic thickening, to capillary diameter, and to glial types and numbers—was examined. Age, gender, duration of exposure, etc. were critical variables that had to be tested in new experiments.

Most of the basic data reported on the enrichment paradigm and its impact on brain and behavior have accumulated through studies on the rat. Effects of enriched and impoverished environments on the nerve cells and their neurotransmitters in the cerebral cortex have now been generalized to several mammalian and avian species (Rosenzweig and Bennett 1996). Some corroborating studies mentioned herein involved cats and monkeys, as well as isolated studies in human subjects. For example, Jacobs et al. (1993) using an isolated portion of the human cerebral cortex responsible for word understanding, Wernicke's area, compared the effects of enrichment in tissue from deceased individuals who had had a college education and from those who had had only a high school education. They demonstrated that the nerve cells in the college-educated showed more dendrites than those in the latter. (Tissue was obtained from the Veterans' Hospital in west Los Angeles.) Experiments on human tissue frequently support the data obtained from studies in the rat, and, in turn, benefit from these animal studies. We can now safely say that the basic concept of brain changes in response to enrichment holds true for a wide variety of animals and for humans.

1. *The Effects of Enrichment on the Cerebral Cortex*

What do we mean by *enrichment* for the rats that have served as the animal of choice for most of these studies? Thirty-six Long-Evans rats were sorted into three experimental conditions using 12 animals in each group: (a) enriched, (b) standard, or (c) impoverished environments. All animals had free access to food and water and similar lighting conditions. Eventually, it was determined that animals maintained in their respective environments from the age of 30 days to 60 days developed the most extensive cerebral cortical changes. For the enriched environment, the 12 animals lived together in a large cage ($70 \times 70 \times 46$ cm) and were provided five or six objects to explore and climb upon (e.g., wheels, ladders, small mazes). The objects were changed two to three times a week to provide newness and challenge; the frequent replacement of objects is an essential component of the enriched condition. The combination of 'friends' and 'toys' was established early on by Krech as vital to qualify the experiential environment as 'enriched' (Krech et al. 1960). For the standard environment, the animals were housed three to a small cage ($20 \times 20 \times 32$ cm) with no exploratory objects. For the impoverished environment, one animal remained alone in a small cage with no exploratory objects.

The numbers of animals placed in these separate conditions were based on the manner in which the routine housing was established in the rat colony. Three rats in a cage has been considered standard for all experimental work over the decades. Since prior to these experiments no one had designed studies to examine brain changes in response to different environmental conditions, the decisions about what represented 'impoverishment' and what represented 'enrichment' was more arbitrarily than scientifically reasoned.

After 30 days in their respective environments, all animals were anesthetized before the brains were removed for comparison among the three groups. Twenty micra frozen sections were cut and stained, and the thickness of the frontal, parietal, and occipital cortices were measured. Results indicated clearly that the cortex from the enriched group had increased in thickness compared with that living in standard conditions, whereas the brains from the impoverished group decreased compared to the standard. Because the nerve cells were farther apart in the enriched vs. the impoverished brains, it was thought that the major component of the brain changes due to enrichment had to do with alterations in the dendritic branching. With more detailed studies, the cortical thickness increases were found to be due to several factors, including increased nerve cell size, number and length of dendrites, dendritic spines, and length of postsynaptic thickening as measured on electron microscopic pictures of synapses (Diamond et al. 1964, Diamond 1988).

In the initial experiments designed to explore the impact of an enriched environment on the brain of postweaned rats, only enriched and impoverished groups were used. Rats were maintained in their respective environments from 25 to 105 days of age because there were no available data on how long it would take to create chemical or structural changes in the cortex. Chemical and anatomical measurements taken from these animals showed significant differences between the two groups in cortical thickness, cortical weight, acetylcholinesterase, cholinesterase, protein, and hexokinase levels (Bennett et al. 1964, Diamond et al. 1964). In these initial experiments, however, it was not clear if the changes were due to enrichment or impoverishment because there were no standard conditions established as controls.

Nonetheless, the differences in cortical thickness with this 80-day exposure to the two environmental conditions were not as great as during the 30-day exposure. Consequently, in subsequent experiments, the period of exposure to the experimental conditions was reduced from 80 days to 30 days, then 15 days,

seven days, and finally to four days. At each of these intervals, animals from the enriched environment showed increases in cerebral cortical thickness in some areas but not in others. For example, in the male animals exposed for 80 days to enriched conditions, the somatosensory cortex did not show significant changes, whereas male animals exposed for 30 days did develop significant differences in the somatosensory cortex. The occipital cortex showed significant changes for both the 80- and the 30-day experiments, but, again, the differences were greater at 30 days than at 80 days. It is possible that the longer exposure served to increase cortical thickness in the early days of enrichment but that over time the environmental condition became monotonous and this effect decreased. In later experiments the experimental conditions were modified to try to establish what the major factors were that created the observed cortical changes. For example, was the effect associated with the number of rats exposed or to the presence of stimulus objects?

The new conditions included one rat living alone in the large enrichment cage with the objects that were changed several times each week. The cortices of these rats did not show a significant effect of enrichment. Twelve rats living together in the large cage without the stimulus objects did not show as great an effect as 12 rats living with the stimulus objects. In other words, the combination of social conditions and frequent exposure to new stimulus objects was necessary for the animals to gain the full effect of enrichment.

Establishing what constitutes 'enrichment' for human beings is more problematic. Not only are controlled experiments not feasible, but no two human brains are identical. Individuals differ in their genetic backgrounds and environmental inputs. Furthermore, what is considered enrichment for one individual may be quite different for another. Yet, as mentioned earlier, the enrichment effect was evident in Wernicke's area from measurements of the amount of dendritic branching in brain tissue from college-educated individuals versus that from high school-educated people. The basic finding of dendritic growth in response to environmental stimulation appears in all brains studied to date. It would appear that newness and challenge are important for the human cortex as well as for that of animals.

2. *Independent Variables: Age and Gender*

Among the many variables researchers must consider as they seek to understand and accurately interpret the effects of enrichment on the brain, age and gender are important considerations. Enrichment has been shown to enhance many aspects of cortical structure at any age—from prenatal to extremely old rats (904 days of age). The amount of change varies with the age of the animal. For example, when a 30-day-old rat is put in an enriched environment for four days, the effects are not as pronounced as they are in the 60-day-old rat maintained in enriched conditions for four days. Is four days too short a time for the very young animal to adjust and benefit from enrichment? A young animal maintained for 30 days in an impoverished environment shows reduced morphological development of its cortex when compared to that of an adult animal maintained in impoverished conditions for 30 days. In further age-related experiments, another component was added to the enrichment conditions of old rats. Despite significant increases in the length of the dendrites in the brains of 600-day-old rats that had been placed in an enriched environment for 30 days (600 to 630 days), several of the old rats in this population died.

To determine whether the enrichment conditions could be modified to extend the animals' lifespan, the investigators added a new component: hand-holding the rats each day for several minutes while the cages were cleaned. In an attempt to increase the lifespan of the rats, rats were placed three to a cage after weaning at 25 days of age, and maintained in these standard conditions until they reached 766 days, at which time half went into enriched conditions until they reached 904 days of age and half stayed in the standard conditions. The only variable added was the daily hand-holding of the rats as they aged. Is it possible that handling the rats had extended their lifespan? Indeed, many investigators have been amazed that these rats survived to 904 days of age. The 904-day-old rats in enriched conditions developed a cortex significantly thicker than the cortex of rats living in the standard conditions (Diamond 1988). These experiments offered support to the thesis that the cerebral cortex is capable of responding positively to an enriched environment at *any* age.

Experiments comparing the effects of enrichment on male and female brains are few. Most enrichment studies have been carried out on male brains to avoid the compounding factors associated with the estrous cycle. In one study focused on gender, the female neocortex was found to respond differently from the male neocortex exposed to the same type of enrichment conditions (Diamond 1988). The male showed significant changes in cortical thickness in the occipital cortex, but no significant changes in the somatosensory cortex. (Although the right cerebral cortex in the brain of the male rat is thicker than the left, especially in the visual or occipital region, an enriched environment appears to alter both the right and left cortex similarly.) In the female, the thickness of the occipital cortex increased significantly in response to enrichment, although not as much as in the male, but the thickness of the somatosensory cortex increased significantly more in the female than in the male. In a follow-up experiment, however, in which obstacles were piled up in front of the female's food cup to provide a greater challenge to her already enriched

environment, the thickness of the occipital cortex increased as much as did that of the male without the additional challenge.

In rats whose testes were removed, either at birth or at 30 days of age before the rats were placed in an enriched environment for 30 days, the increases observed in cortical thickness were similar to those of their littermates with intact testes (Diamond 1988). These findings suggested that testosterone is not implicated in the increases in cortical thickness observed in the brains of rats living in enriched environments. Since sex differences were evident in the responses of the animals to enrichment, interest was now focused on the brains of pregnant rats, in which the concentrations of sex steroid hormonal concentrations are greatly altered. The brains of female rats living in the enriched environment from 60 to 90 days and then becoming pregnant and returning to enrichment until 116 days of age were compared between nonpregnant and pregnant animals living in an impoverished environment for the same time periods. When animals from the two groups were autopsied at 116 days, no significant differences in cortical thickness were found. Evidently, pregnancy has an effect on the cerebral cortex regardless of whether the environment is impoverished or enriched.

These initial experiments, all of which were replicated, clearly indicate gender differences in the brain's response to enrichment. Having dealt with the independent variables, we turn to the impact of dependent variables in the enrichment paradigm. For these studies, one must look at: duration of exposure, brain anatomy and chemistry, presence of lesions or fetal neocortical grafts, negative air ions, stress, physical activity, and nutrition, as well as behavioral effects. These are discussed in turn below.

2.1 Duration

The duration of exposure to the enriched environment is clearly a significant dependent variable that must be factored into research in this area. As short a period as 40 minutes of enrichment has been found to produce significant changes in RNA and in the wet weight of cerebral cortical tissue sampled. One day of enrichment was insufficient to produce measurable changes in cortical thickness, whereas four consecutive days of exposure (from 60 to 64 days of age) to an enriched environment did produce significant increases in cortical thickness, but only in the visual association cortex (area 18) (Diamond 1988).

When young adult rats were exposed to 30 days of enrichment, however, the entire dorsal cortex, including frontal, parietal, and occipital cortices, increased in thickness. Extending the duration of the stay in enriched conditions to 80 days did not produce any greater increase in cortical thickness than that seen at 30 days (in fact, it was often even less); however, the longer the rat remained in the enriched conditions, the longer the cortex retained its increased dimensions following return to the standard environment (Bennett et al. 1974). When we looked at age-related differences in the context of duration of stay in the enriched environment, we found that old rats (766 days of age) placed in enriched conditions for 138 days showed an increase in cortical thickness that was quite similar to that observed in young adult rats (60 days of age) that had lived in enriched conditions for 30 days.

2.2 Anatomical and Chemical Components

Early experiments, and those to follow in subsequent years, again demonstrated significant differences in brain chemistry and anatomy associated with enriched living conditions. Anatomical increases include all of the structural constituents measured in the cerebral cortex to date, such as cortical thickness (Diamond et al. 1964), nerve cell soma size, nerve cell nuclear size (Diamond 1988), dendritic dimensions (Holloway 1966, Greenough et al. 1973), dendritic spines, synaptic size and number (Mollgaard et al. 1971, Black et al. 1990), number of glia, capillary diameter (Diamond 1988), dendritic number after lesions (McKenzie et al. 1990), and successful tissue grafts (Mattsson et al. 1997). Chemical increases include: total protein, RNA-to-DNA ratio, cholinesterase-to-acetylcholine ratio, Nerve Growth Factor mRNA, cyclic AMP, choline acetyltransferase, cortical polyamines, NMDA (N Methyl D Aspartate) receptors, and hexokinase, etc.

2.3 Lesions

Another variable has to do with the impact of enriched conditions on purposefully incurred brain lesions. In a 1990 study, 60-day-old rodents were exposed for 30 days to either an enriched or standard environment two days after having received a lesion in the left frontal cortex that created a motor dysfunction in the right forepaw. Animals living in the enriched condition showed significant increases in cortical dendritic branching in both hemispheres, the lesioned and the non-lesioned sides, along with a significant return of motor function in the right forepaw compared to those animals living in standard conditions (McKenzie et al. 1990).

2.4 Fetal Neocortical Graft

Similarly, providing an enriched environment to rats that had undergone fetal neocortical grafts one week

after lesioning was found to improve behavior and to reduce the atrophy in the thalamus, a major structure beneath the cortex that supplies neural input to the cortex (Mattsson et al. 1997). The fact that the fetal neocortical graft when placed in the lesioned cerebral cortex could prevent atrophy in the underlying thalamus as a consequence of enrichment is of great interest to researchers considering the future possibility of using such grafts for brain-damaged individuals.

2.5 Air Ions

The possibility that physical environmental stimuli other than those classically regarded as 'sensory' could have an effect on the brain was tested experimentally by exposing rats living in enriched or standard environments to high concentrations of negative air ions. The experiments were undertaken to determine whether the effect of negative ions on serotonin, the putative second messenger cyclic-AMP, and on cyclic GMP in the cerebral cortex, differ depending on whether the animals lived in enriched or standard conditions. Studies demonstrated that rats placed in the enriched environment in the presence of enhanced negative air ions (ion density of 1×105) showed a significant decrease in serotonin, an effect not found in the brains of animals living in standard conditions (Diamond et al. 1980). Measurements of cyclic AMP decreased as well in the brains of the animals living in the enriched conditions, but cyclic GMP did not. These results indicate the importance of considering air quality and atmospheric conditions in determining the brain's response to enrichment.

2.6 Stress

The presence or absence of stress represents yet another variable to be taken into consideration in such studies, certainly so in any extrapolation of these findings to humans. Stress is a major factor in contemporary, fast-moving urban life. Crowding, for example, is deemed stressful under conditions where competition for space or food is likely. Experiments were set up to assess the effect of crowding on the brains of rats maintained in an enriched environment. To create a condition in which crowding would be experienced as stressful, 36 rats were placed in an enrichment cage usually housing only 12 rats, and kept there for 30 days. The results indicated that, compared with rats living in standard conditions, the thickness of the medial occipital cortex increased significantly whether the enrichment cage housed 12 or 36 animals (Diamond et al. 1987). One hypothesis to come from this study was that the animals' interaction with the toys might be diverting their attention or entertaining them sufficiently to mitigate the stress of the crowded condition.

Chronic stress has been reported by Meaney et al. (1988) to produce excess glucocorticoids, which are toxic to neurons—especially those of the hippocampus. Aged rats are particularly vulnerable to chronic stress. The investigations of Meaney showed that enriching the living conditions of old rats, or handling them in their infancy, helps to prevent stress-related hippocampal damage.

It is possible that stress can be produced by increasing the frequency with which the various objects in the enrichment cage are changed. In all previous studies, objects had been replaced daily or at least several times each week. Then the question was asked whether increasing the frequency of changing the objects would further increase the growth of the cortical thickness, or, alternatively, would it be experienced as a stress factor, given that the animals were inhibited from interacting with them in the more leisurely manner to which they were accustomed? For these experiments, rats 60 to 90 days of age found their objects changed every hour for three hours on four nights of each week for four consecutive weeks. Under this regime, the cerebral cortical thickness did not grow significantly compared to cortices from rats whose objects were changed several times each week for four weeks (Diamond unpublished).

Corticosteroids, released under stress, have been shown to reduce cortical thickness and future experiments would be necessary to compare differences in corticosteroid levels in animals exposed to these differing conditions.

2.7 Behavior

Psychologists have known for a long time that early experience influences the adult performance of an animal. In experiments in the 1950s (Bingham and Griffths 1952, Forgays and Forgays 1952) investigators were interested in determining how much experience in complex environments was necessary to produce a highly intelligent adult animal and when, specifically, during early life these experiences had to occur. These studies showed that all of the animals maintained in enriched conditions were better problem-solvers than those with no enrichment; however, in some other occasions, using other tests, enriched rats did not perform significantly better than controls.

One of the most robust effects of environmental enrichment on the behavior of rats appears in the areas of learning and memory. Investigators (York et al. 1989, Kempermann et al. 1997) studying the effects of enrichment in the rat brain have reported that new nerve cells develop in the adult dentate gyrus, an area dealing with recent memory processing. In the York

experiments the rats were 60 to 90 days of age (truly adult animals) during the enrichment experience, whereas in the Kempermann experiments the mice were 21 to 40 days of age. These findings are significant because neurogenesis had not previously been found in the cerebral cortex of the mammalian adult. Earlier studies had found that enriched environments stimulate the growth of dendrites in the dentate gyrus, and only in female rats (Juraska et al. 1985).

2.8 Physical Activity

One component of enrichment is the physical exercise involved in the animals' having to move about the cage, interacting with and climbing upon the novel objects. These activities appear to influence the motor cortex as well as the hippocampus. Olsson et al. (1994) showed that rats living in enriched environments at 50 days of age showed higher expression of the gene-encoding glucocorticoid receptors and induction of genes for Nerve Growth Factors in the hippocampus.

2.9 Nutrition

Nutrition is clearly an important variable to consider in all studies dealing with brain and behavior. Environmental enrichment and impoverishment have pronounced effects on nutritionally deficient animals. One study compared the effects of environmental enrichment on the offspring of mother rats living on protein-rich or protein-deficient diets during pregnancy (Carughi et al. 1990). The protein-rich diet proved beneficial for the healthy development of the cerebral cortical dendrites in young rats and even more so when combined with an enriched environment.

The cerebral cortical dendrites in rat pups from mothers with a protein-deficient diet were significantly less well developed than those of their counterparts, but, of greater importance, the cortex from the protein-deficient animals did not significantly increase with enrichment. However, when protein-deficient pups were fed a protein-rich diet and maintained in an enriched environment during their early postnatal life, cortical development improved almost to the level seen in rat pups from mothers on a high-protein diet during pregnancy followed by postnatal enrichment. These data are very encouraging, because they suggest the possibility of making up for lost brain growth during pregnancy by enriching both the diet and the environmental conditions during the postnatal period.

Another dietary factor significant to optimal brain function is glucose. The brain depends almost exclusively on glucose for its energy. Synapses use a great deal of energy and glucose supplies this energy. Although we know that different parts of the brain use glucose at different rates, to learn which of 30 discrete brain regions were most active in adult rats placed in enriched living conditions from 57 to 87 days of age, we studied their radioactive glucose uptake during this 30-day period and compared it with that of rats raised in standard conditions (Diamond 1988). Again, the cerebral cortex showed the greatest differences between enriched and nonenriched groups, but, surprisingly, of the two groups, glucose uptake was lower in rats maintained in enriched conditions. We concluded from this finding that glucose uptake is more efficient in the brain of animals living in enriched environments. Out of the 30 areas of the brain measured, including the cortex, only one area showed significantly greater glucose uptake in the enriched animals: the corpus callosum, specifically, the large mass of axons connecting the nerve cells between the two cerebral hemispheres. Could the axons forming the corpus callosum from the nerve cells in the cerebral cortex be more active than the nerve cell bodies from which they arise? Yet the right and left cerebral cortices show comparable cortical thickness increases with enrichment due to the effects on dendritic branching, but now the data show that the rates of glucose utilization in both the frontal and parietal cortices were 13 percent lower in the enriched rats than in the standard control rats, a paradox to be untangled in the future.

3. Methodological Issues Associated with Enrichment Research in Humans

Of the vast number of animal studies that yield results of interest to human research, studies on the impact of an enriched environment on brain development and behavior can be of enormous interest to humans. Despite similarities in some key respects between the brain of the rat and other mammals, replicating or extrapolating from anatomical and chemical studies conducted in animals is fraught with difficulty, for obvious reasons. Not only is it not presently possible to control all of the experimental variables at work in humans, but the diversity and complexity of human experience militates against designing experiences comparable to those used with lower animals. Nevertheless, these studies and those few human studies that have been done suggest that there are measurable benefits to enriching an individual's environment in whatever terms that individual perceives his immediate environment as enriched. At the very least, this work indicates that there are many opportunities for enhancing brain activity and behavior at all ages, and that they can have pronounced effects throughout the lifespan.

See also: Learning and Memory, Neural Basis of; Neural Plasticity; Synapse Formation

Bibliography

Bennett E L, Diamond M C, Krech D, Rosenzweig M R 1964 Chemical and anatomical plasticity of the brain. *Science* **164**: 610–19

Bennett E L, Rosenzweig M R, Diamond M C 1974 Effects of successive environments on brain measures. *Physiology and Behavior* **12**: 621–31
Bingham W E, Griffiths W J 1952 The effect of different environments during infancy on adult behavior in the rat. *Journal of Comparative Physiology and Psychology* **45**: 307–12
Black, J E, Isaacs K R, Anderson B J, Alcantara A A, Greenough W T 1990 Learning causes synaptogenesis, whereas motor activity causes angiogenesis, in cerebellar cortex of adult rats. *Proceedings of the National Academy of Science* **87**: 5568–72
Carughi A, Carpenter K J, Diamond M C 1990 The developing cerebral cortex: Nutritional and environmental influences. *Malnutrition and the Infant Brain*. Wiley-Liss, pp. 127–39
Darwin C 1874 *The Descent of Man*, 2nd edn. Rand McNally, Chicago
Diamond M C, Krech D, Rosenzweig M R 1964 The effects of an enriched environment on the rat cerebral cortex. *Journal of Comparative Neurology* **123**: 111–19
Diamond M C, Connor J R, Orenberg E K, Bissell M, Yost M, Krueger A 1980 Environmental influences on serotonin and cyclic nucleotides in rat cerebral cortex. *Science* **210**: 652–4
Diamond M C, Greer E R, York A, Lewis D, Barton T, Lin J 1987 Rat cortical morphology following crowded-enriched living conditions. *Experimental Neurology* **96**: 241–7
Diamond M C 1988 *Enriching Heredity*. The Free Press, New York
Forgays G, Forgays J 1952 The nature of the effect of free environmental experience in the rat. *Journal of Comparative Physiology and Psychology* **45**: 322–8
Greenough W T, Volkman R, Juraska J M 1973 Effects of rearing complexity on dendritic branching in fronto-lateral and temporal cortex of the rat. *Experimental Neurology* **41**: 371–8
Holloway R L 1966 Dendritic branching: some preliminary results of training and complexity in rat visual cortex. *Brain Research* **2**: 393–6
Hubel D H, Wiesel T N 1965 Binocular interaction in striate cortex of kittens reared with artificial squint. *Journal of Neurophysiology* **28**: 1041–59
Jacobs B, Schall M, Scheibel A B 1993 A quantitative dendritic analysis of Wernicke's area in human. II. Gender, hemispheric, and environmental changes. *Journal of Comparative Neurology* **327**: 97–111
Juraska J M, Fitsch J M, Henderson C, Rivers N 1985 Sex differences in the dendritic branching of dentate granule cells following differential experience. *Brain Research* **333**: 73–80
Kempermann G, Kuhn H G, Gage F H 1997 More hippocampal neurons in adult mice living in an enriched environment. *Nature* **386**: 493–5
Krech D, Rosenzweig M R, Bennett E L 1960 Effects of environmental complexity and training on brain chemistry. *Journal of Comparative Physiology and Psychology* **53**: 509–19
Mattsson B, Sorensen J C, Zimmer J, Johansson B B 1997 Neural Grafting to experimental neocortical infarcts improves behavioral outcome and reduces thalamic atrophy in rats housed in enriched but not standard environments. *Stroke* **6**: 1225–31
McKenzie A, Diamond M C, Greer E R, Woo L, Telles T 1990 The effects of enriched environment on neural recovery following lesioning of the forelimb area of rat cortex. Paper presented at the American Physical Therapy Annual Conference, Anaheim, CA
Meaney M J, Aitkin D H, Bhatnagar S, Van Berkel C, Sapolsky R M 1988 Postnatal handling attenuates neuroendocrine, anatomical and cognitive impairments related to the aged hippocampus. *Science* **283**: 766–8
Mollgaard K, Diamond M C, Bennett E L, Rosenzweig M R, Lindner B 1971 Quantitative synaptic changes with differential experience in rat brain. *International Journal of Neuroscience* **2**: 113–28
Olsson T, Mohammed A H, Donaldson L F, Henriksson B G, Seckl J R 1994 Glucocorticoid receptor and NGFI-A gene expression are induced in the hippocampus after environmental enrichment in adult rats. *Molecular Brain Research* **23**: 349–53
Rampon C, Jiang C H, Dong H, Tang Y-P, Lockhart D J, Schultz P G, Tsien J Z, Hu Y 2000 Effects of environmental enrichment on gene expression in the brain. *Proceedings of the National Academy of Science of the USA* **97**(23): 12880–4
Rosenzweig M R, Bennett E L 1996 Psychobiology of plasticity: Effects of training and experience on brain and behavior. *Behavioral Brain Research* **78**: 57–65
Spurzheim J C 1815 *The Physiognomical System of Drs Gall and Spurzheim*, 2nd edn. Baldwin Cradock and Joy, London, pp. 554–5
York A D, Breedlove S M, Diamond M C 1989 Housing adult male rats in enriched conditions increases neurogenesis in the dentate gyrus. *Society for Neuroscience Abstracts* No. 383.11, p. 962

M. C. Diamond

Brain Stimulation Reward

1. *The Discovery and its Context*

Brain stimulation reward was discovered in 1953 by James Olds and Peter Milner (1954; see also Olds 1973, Milner 1989), who had come to McGill University to work with D. O. Hebb, inspired by his groundbreaking theoretical work. In contrast to the dominant behaviorist orthodoxies of the day, which isolated psychology from the emerging disciplines of neuroscience and cognitive psychology, Hebb's views linked brain, mind, and behavior.

New findings and ideas about the neural bases of motivation and reinforcement provided a rich context for the discovery of BSR (Valenstein 1973). For example, the demonstration that lesions of different hypothalamic nuclei could lead either to massive overeating or starvation contributed to a seminal physiologically based theory of motivation (Stellar 1954). In that era, English-speaking investigators became familiar with Hess' use of electrical stimulation to map the hypothalamic control of autonomic outflow and behavior. Among Hess' observations were stimulation-elicited responses suggesting states of rage or fear. Shortly before Olds and Milner's discovery, Delgado, Roberts and Miller (Delgado et al. 1954) showed that cats would learn to escape electrical stimulation of certain deep forebrain structures. An-

other key component of the context was research linking brainstem circuitry to the control of arousal and vigilance. The pattern of connections between nerve cells in the brainstem region of interest was seen as net-like ('reticulated'); thus the region was dubbed the 'reticular formation.'

At the time that Olds joined Hebb's group, Milner and Seth Sharpless, another graduate student, had developed a theory linking reticular activation to positive reinforcement, the process by which desirable outcomes strengthen the behaviors that produce them. However, their experimental tests of this idea proved disappointing: if anything, the rats tended to avoid locations in a maze where stimulation of the reticular formation had been delivered.

A strikingly different reaction was noted in a rat prepared by Olds, in which the electrode somehow missed its reticular-formation target and ended up instead in the basal forebrain. This subject returned repeatedly to the location where stimulation had been delivered. The behavior of this rat was readily shown to be the product of learning. As the experimenters altered the location in the maze where the stimulation was delivered, the animal's behavior changed accordingly, and it directed its searching towards the site where it had most recently hit pay dirt. Within hours of the initial discovery, Olds succeeded in training the rat to press a lever that triggered delivery of the stimulation. This animal worked indefatigably to stimulate its own brain; such behavior has since been termed 'intracranial self-stimulation' (ICSS).

In the initial report published by Olds and Milner (Olds and Milner 1954), they presciently discuss their findings 'as possibly laying a methodological foundation for a physiological study of the mechanisms of reward.' Indeed, their paper was the well-spring for thousands of subsequent experiments.

Why did the report by Olds and Milner have such a large and enduring impact? Perhaps paramount is the promise their results offered that psychological phenomena such as learning, reinforcement, and motivation could be investigated fruitfully by physiological means. The authors also pointed out that the phenomenon of BSR could be used to distinguish between competing behavioral theories. For example, a prominent theory at the time held that reinforcement arose from the reduction of a drive, such as the abatement of hunger. This view is not easily reconciled with the findings that rewarding stimulation of many sites arouses the animal rather than calming it and that stimulation of some BSR sites can also induce sated animals to engage in consummatory behaviors such as eating and drinking.

2. *Lines of Inquiry*

Many of the principal questions about BSR can be subdivided under the headings, 'where,' 'why,' and 'how.' The 'where' questions include the mapping of BSR sites in the brain, the identity of the key neural populations, and the interconnections between these cells. The 'why' questions concern the functional significance of the phenomenon: how is BSR related to the rewarding effects of natural stimuli and of habit-forming drugs? The 'how' questions are mechanistic in nature: what set of neural and psychological processes translate the transient stream of nerve impulses elicited by the stimulation into goal-directed behavior and an enduring record of the rewarding effect?

2.1 *Where?*

The sites where BSR can be produced occupy a substantial proportion of the brain volume. Positive sites are distributed from the olfactory bulbs, at the front of the brain, to the nucleus of the solitary tract, at the back, and are found both in cortical and subcortical regions of the forebrain.

The identification of the cells underlying BSR continues to be a major challenge. Progress has been made in characterizing such neurons (Shizgal and Murray 1989; Yeomans 1990; Shizgal 1997), but their identity has not yet been established firmly at any BSR site. The best-characterized sites lie along two major pathways, the medial forebrain bundle (MFB) and fibers coursing near the mid-line of the brainstem. In both cases, the neurons that carry the reward-related signal away from the electrode tip include cells with relatively fast-conducting, longitudinally oriented, myelinated axons. This characterization narrows down the field of plausible candidates, but it leaves open many possibilities.

Lesions of several structures have been shown to weaken the rewarding effect. For example, following damage to a basal forebrain region that includes the sub-lenticular extended amygdala, the strength of MFB stimulation must be increased markedly in order to support self-stimulation. However, there is substantial across-subject variation in the effect of these and other lesions on ICSS. Explanations of this variability have been offered but have not yet been substantiated (Shizgal 1997).

Pharmacological methods have implicated several neurotransmitter systems in BSR. There is a large body of evidence tying dopamine-containing neurons to the rewarding effect (Wise and Rompré 1989). For example, manipulations that enhance transmission in these cells potentiate BSR and manipulations that decrease dopaminergic neurotransmission weaken the rewarding effect. The dopamine-containing neurons that project from the ventral tegmental area of the midbrain to the nucleus accumbens, a structure at the base of the anterior forebrain, have been singled out as particularly important. The modulation of BSR by manipulation of acetylcholine- and serotonin-containing neurons in the brainstem may be due to facilitatory and inhibitory interactions of these cells with dopaminergic neurons. Despite the considerable evidence

linking dopaminergic neurons to BSR, the nature of their contribution has not yet been established firmly. The directly activated neurons responsible for self-stimulation of the MFB and periaqueductal gray include non-dopaminergic cells (Shizgal and Murray 1989). Do dopaminergic neurons relay the output of the directly activated neurons to later stages of the circuitry or do the dopaminergic neurons serve to gate transmission at key synapses? Difficulty in reconciling certain key findings is a further indication that the last word about the dopaminergic contribution to BSR has not yet been written. For example, injection into the nucleus accumbens of dopamine receptor blockers decreases the effectiveness of rewarding MFB stimulation (Nakajima and O'Regan 1991), yet failure to detect release of dopamine in the nucleus accumbens following prolonged self-stimulation of the MFB has been reported, despite the use of very sensitive measurement methods (Garris et al. 1999).

2.2 *Why?*

A rat may spend hours on end doing little else but pressing a lever to deliver brief trains of electrical pulses to MFB neurons. Showing few signs of fatigue or habituation, the rat continues to approach and vigorously press the lever. If the lever is in a large enclosure and the rat is displaced to a distant corner, it will gallop back immediately. What is it about the brief burst of nerve impulses triggered by the electrode that causes the rat to focus so resolutely on obtaining more stimulation?

Since the early days of research on BSR, it has been suspected that the stimulation mimics effects of natural goal objects, such as food and water. Indeed, MFB stimulation has been shown to compete with, summate with, and substitute for rewards such as sugar solutions or food pellets. Differing views have been offered as to the nature of the effects that the stimulation mimics. For example, it has been argued that the stimulation mimics the overall utility of natural goal objects rather than specific sensory attributes such as texture, taste, and temperature (Shizgal 1999). This view is based on evidence that BSR is represented by a unidimensional neural code and that natural rewards of different types (e.g., solutions of sugar or salt) can summate with the rewarding effect of stimulating a single brain site. However, it has also been shown that food or water deprivation can alter preference between stimulation of different brain sites (Gallistel and Beagley 1971) and that the potentiating effect of chronic food restriction is seen only at some MFB sites (Blundell and Herberg 1969). Moreover, the fat-signaling hormone, leptin, has opposite effects on self-stimulation of different brain sites, depending on whether or not the rewarding effect is sensitive to chronic food restriction (Fulton et al. 2000). The unidimensional coding hypothesis could be reconciled with such findings if different subsets of neurons represented the utility of different categories of natural goal objects and the outputs of these multiple populations of cells converged onto a final common path.

Much work has been done to assess the effects on BSR of drugs such as cocaine, amphetamine, heroin, and nicotine (Wise 1996). These drugs have been shown to increase the effectiveness of the rewarding stimulation; reward effectiveness has also been shown to decrease during drug withdrawal. Such findings suggest that the neural circuitry responsible for BSR may play an important role in the habit-forming effects of drugs. Results of a study employing functional magnetic resonance imaging are consistent with such a view. Regions containing cells suspected of contributing to the rewarding effect of MFB stimulation in rats, such as the sub-lenticular extended amygdala and nucleus accumbens, showed activation during the expectation and/or the experience of cocaine-induced euphoria in humans (Breiter et al. 1997).

2.3 *How?*

The signal injected by the stimulating electrode can be regarded as having two destinations. The first is in the real-time control of ongoing behavior. This is illustrated particularly clearly in the case of electrodes that produce 'pure' rewarding effects without recruiting significant aversive effects as well. If the experimenter delivers a prolonged stimulation train through such an electrode and then turns it off, the rat will immediately turn it back on. Such observations have led theorists to link BSR with homing mechanisms that potentiate ongoing behaviors yielding high payoffs while terminating behaviors with negative outcomes. In such a view, the electrode activates the neural circuitry that estimates ongoing payoffs so as to signal a high return (Shizgal 1999).

The results of numerous experiments demonstrate that information about the strength of the rewarding effect is stored. This can be seen informally in the eagerness of experimental subjects to gain entry to the test cage and formally in the dependence of approach tendencies on the strength of the stimulation delivered on preceding trials. Thus, it has been argued that the second destination of the signal injected by the stimulating electrode is a memory record of some kind (Gallistel, Stellar et al. 1974).

3. *Methodological Issues*

In order to identify the neurons responsible for the rewarding effect, the investigator must have a good *a priori* sense of the quarry. Thus, methods for inferring characteristics of the reward signal and the neurons that produce it are of paramount importance to identifying the cells responsible for BSR (Yeomans 1990). Another aspect of the problem of reward

measurement concerns the distinction between performance and competence. When a rat ceases to respond for BSR following a lesion or administration of a drug, has the value of the BSR decreased or is the subject simply having trouble performing the required response? Practical approaches for distinguishing changes in performance capacity from changes in reward have been developed (Miliaressis et al. 1986), and these methods appear far superior to those used in the early days of research on BSR.

Research on the determinants of instrumental performance has evolved in parallel with work on BSR. Since the pioneering work of Olds and Milner, the 'matching law' was introduced (Davison and McCarthy 1988), offering a reasoned basis for inferring the strength of reward from performance. According to the matching law, the investment of behavior in competing activities is proportional to their relative payoffs. If so, relative reward strengths can be inferred from indices of performance such as time allocation or response rates. Although ideas based on the matching law have been applied to the scaling of BSR (Gallistel and Leon 1991), the validity of the law remains controversial. One promising line of attack extends the matching law by application of ideas from microeconomics, in particular the theories of consumer choice and labor supply (Kagel et al. 1995).

In parallel with the development of scaling methods, there has been dramatic progress in the development of physical methods for observing and manipulating neural activity. Methods for measuring the activity of single neurons and the release of neurotransmitters in behaving subjects have been applied successfully to the study of BSR. A wide array of lesion methods is now available, including methods for targeting specific neurotransmitter systems and for damaging cells on the basis of their projection fields. The promise of applying these methods to the study of BSR is far from having been exhausted. Dramatic progress has also been made in neuropharmacology, providing investigators with precise descriptions of receptor types and distributions, and increasingly selective tools for manipulating neurotransmission. Application of techniques from molecular biology has been used to visualize neurons activated by rewarding stimulation, and new methods for genetic manipulation promise very specific means of testing the roles in BSR played by particular neurotransmitters and receptors.

4. Future Directions

One likely direction for future work on BSR would combine advances in the scaling of reward with new techniques for measuring and manipulating neural activity. Such work offers the promise of achieving a fundamental goal that has as yet proved elusive: identifying the directly activated neurons responsible for the rewarding effect. Finding these cells would provide a solid base from which to trace inputs and outputs, to test theories of the natural function of the circuitry underlying BSR, and to determine how dopaminergic neurons contribute to BSR.

Electrical brain stimulation provides fine control over the timing and strength of reward, thus rendering BSR a promising tool for investigating basic questions in the decision and behavioral sciences. For example, there is a lively debate in decision science concerning the role of temporal patterning in the evaluation and choice of hedonic stimuli. This issue can be addressed in laboratory animals via the BSR paradigm.

The development of functional brain imaging methods has seeded many fruitful collaborations between cognitive scientists and neurobiologists studying human subjects. Applying these methods in behaving laboratory animals should make it possible to map neural activation evoked by rewarding brain stimulation and to compare it to the distribution of activation induced by natural rewards.

Solving the puzzle of BSR has likely proved harder than was anticipated in the heady years following the initial discovery. However, progress continues to be made, and there are grounds for cautious optimism concerning future prospects. Working out the mechanisms and circuitry underlying this intriguing phenomenon promises to pay rich dividends for our understanding of the neural foundations of both adaptive and maladaptive behaviors.

See also: Arousal, Neural Basis of; Electrical Stimulation of the Brain; Hebb, Donald Olding (1904–85); Reinforcement: Neurochemical Substrates; Reinforcement, Principle of

Bibliography

Blundell J E, Herberg L J 1969 Relative effects of nutritional deficit and deprivation period on rate of electrical self-stimulation of lateral hypothalamus. *Nature* **219**: 627–8

Breiter H C, Gollub R L, Weisskoff R M, Kennedy D N, Makris N, Berke J D, Goodman J M, Kantor H L, Gastfriend D R, Riorden J P, Mathew R T, Rosen B R, Hyman S E 1997 Acute effects of cocaine on human brain activity and emotion. *Neuron* **19**(3): 591–611

Davison M, McCarthy D 1988 *The Matching Law*. Lawrence Erlbaum Associates, Hillsdale, NJ

Delgado J M R, Roberts W W, Miller N E 1954 Learning motivated by electrical stimulation of the brain. *American Journal of Physiology* **179**: 587–93

Fulton S, Woodside B, Shizgal P 2000 Modulation of brain reward circuitry by leptin. *Science* **287**(5450): 125–8

Gallistel C R, Beagley G 1971 Specificity of brain stimulation reward in the rat. *Journal of Comparative and Physiological Psychology* **76**: 199–205

Gallistel C R, Leon M 1991 Measuring the subjective magnitude of brain stimulation reward by titration with rate of reward. *Behavioral Neuroscience* **105**: 913–25

Gallistel C R, Stellar J R, Bubis E 1974 Parametric analysis of brain stimulation reward in the rat: I. The transient process and the memory-containing process. *Journal of Comparative and Physiological Psychology* **87**: 848–59

Garris P A, Kilpatrick M, Bunin M A, Michael D, Walker Q D, Wightman R M 1999 Dissociation of dopamine release in the nucleus accumbens from intracranial self-stimulation. *Nature* **398**(6722): 67–9
Kagel J K, Battalio R C, Green L 1995 *Economic Choice Theory: An Experimental Model of Animal Behavior*. Cambridge University Press, Cambridge, UK
Miliaressis E, Rompré P-P, Laviolette P, Philippe L, Coulombe D 1986 The curve-shift paradigm in self-stimulation. *Physiology and Behavior* **37**: 85–91
Milner P M 1989 The discovery of self-stimulation and other stories. *Neuroscience and Biobehavioral Reviews* **13**(2–3): 61–7
Nakajima S, O'Regan N B 1991 The effects of dopaminergic agonists and antagonists on the frequency-response function for hypothalamic self-stimulation in the rat. *Pharmacology Biochemistry and Behavior* **39**: 465–8
Olds J 1973 Commentary. In: Valenstein E S (ed.) *Brain Stimulation and Motivation; Research and Commentary*. Scott Foresman, Glenview, IL, pp. 81–99
Olds J, Milner P M 1954 Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology* **47**: 419–27
Shizgal P 1997 Neural basis of utility estimation. *Current Opinion in Neurobiology* **7**(2): 198–208
Shizgal P 1999 On the neural computation of utility: implications from studies of brain stimulation reward. In: Kahneman D, Diener E, Schwarz N (eds.) *Well-being: The Foundations of Hedonic Psychology*. Russell Sage Foundation, New York, pp. 502–26
Shizgal P, Murray B 1989 Neuronal basis of intracranial self-stimulation. In: Liebman J M, Cooper S J (eds.) *The Neuropharmacological Basis of Reward*. Oxford University Press, Oxford, UK, pp. 106–63
Stellar E 1954 The physiology of motivation. *Psychological Review* **61**: 5–22
Valenstein E S 1973 *Brain Stimulation and Motivation; Research and Commentary*. Scott Foresman, Glenview, IL
Wise R A 1996 Addictive drugs and brain stimulation reward. *Annual Review of Neuroscience* **19**: 319–40
Wise R A, Rompré P P 1989 Brain dopamine and reward. *Annual Review of Psychology* **40**: 191–225
Yeomans J S 1990 *Principles of Brain Stimulation*. Oxford University Press, New York

P. Shizgal

British Cultural Studies

One of the markers of contemporary studies in the humanities and social sciences is the close, critical attention directed onto the mass media and popular culture. British cultural studies, emerging out of the expanded English university system in the 1970s, were among the first to develop and broadly disseminate the methodologies and approaches which made such attention possible. Focusing initially on the relations between language and ideology through the close analysis of media texts, before diversifying into a broad range of interdisciplinary activities—cultural theory, cultural history, ethnography, media studies, and the analysis of the social construction of gender, race and identity—British cultural studies has played a fundamental role in establishing the conceptual and theoretical agenda for studies of popular cultural forms and practices.

1. *The Analysis of Popular Culture*

Academic disciplines in the humanities and social sciences took some time to accept the need to understand properly the power and influence of the mass media in Western societies. The spread of a transnational commercial popular culture, apparently disconnected from specific local cultures and disseminated through advertising and marketing, was seen by many in the academy as a threat to traditional or civilized values which should simply be repudiated. Within the British university system until the end of the 1960s, the task of analyzing and evaluating popular culture had been taken on by the discipline of English. However, the methods supplied by training in literary studies were suited for making esthetic and moral distinctions that proved inappropriate as a means of understanding the appeal of popular cultural forms such as film, television, or popular music. It was gradually acknowledged that popular cultural forms served a very different social function for their audiences than that served by high art for its audiences. Cultural studies was a product of the consequent recognition of the necessity of finding new ways of understanding the specific social functions of popular culture. Although it was partly a product of literary studies, the methods it developed had to be very different.

The starting points for this project are usually seen (Turner 1996) to be Richard Hoggart's account of the clash between an 'organic' English working class culture and mass mediated popular culture, *The Uses of Literacy* (1958), and Raymond Williams' history of cultural change in an industrializing England, *Culture and Society 1780–1950* (1958). The debates generated by these books set the conditions within which Hoggart established the Centre for Contemporary Cultural Studies (CCCS) at the University of Birmingham in 1964. In 1968, he was succeeded as Director of the CCCS by Stuart Hall, who played a major role in developing methods of analysis for popular cultural forms over the next two decades. In particular, Hall's work during the 1970s provided a very important conduit through which theoretical developments from France, by Barthes around post-Saussurean semiotics and by Althusser around the workings of ideology (Hall 1977), reached an English readership.

Initially, the analyses of popular culture which emerged from this tradition were highly political critiques of the ways ideology was embedded in language. Close analysis of media 'texts' (specific examples of media practice exposed to detailed study)

was employed as means of explaining how social understandings were carried and reinforced by the mass media, and how these social understandings were structurally inclined to be those supported by the state or the interests of market capitalism (Hall et al. 1978). However, by the latter half of the 1980s, British cultural studies had distanced itself from the notion that the media was simply a means of reinforcing the ideological interests of capital or the state. Instead, it had acknowledged the alternative possibilities suggested by such writers as Dick Hebdige (1979) during the late 1970s and began to investigate how popular cultural forms might also serve the interests of those who consume them. This produced a more positive reading of popular culture, and a more nuanced and contingent set of expectations about the specific interests served by popular cultural forms at particular historical conjunctures (Fiske 1987). Among the results of this focus over the 1980s was a significant expansion of the field, as a broad range of studies of television, popular music, the print media, fashion, media audiences, and youth subcultures were published, and as cultural studies established itself at undergraduate and graduate level in universities in the UK, the USA, Canada, Australia, and elsewhere. As the various forms of cultural studies proliferated (borrowing methods from history and anthropology as well as literary studies and sociology), the relative distinctiveness of the British tradition began to dissipate.

By the early 1990s, this reorientation of the field and modes of analysis had exposed cultural studies to the accusation of 'cultural populism' (McGuigan 1992): of complicity with the media and cultural industries that reduced its capacity to provide a critical analysis of the interests these industries served. This provoked a further reassessment of the study of popular culture which served to reassert the complicated set of understandings required: to understand how popular cultural forms worked as communication; to understand the competing political potentials implicit in the consumption of popular culture; and to understand how the political economy of the media and cultural industries operates as a determining frame around the relations between producers and consumers.

Crucial to the development of the analysis of popular culture within this tradition has been the focus on the media product as a 'text,' and the complementary insistence on the importance of understanding the processes through which these texts are consumed by audiences.

2. *Texts and Audiences*

The idea of treating media products as 'texts' comes from literary studies but the early means of doing so came from semiotics (see, for example, *Semiotics*). Semiotics was especially useful for two reasons. First, while it was interpretative as well as descriptive, it did not produce an aesthetic or moral evaluation of the text and was thus differentiated from the only other discipline where textual analysis was highly developed—literary studies. Second, it was the first method which offered the means for analyzing all forms of communication—written, verbal, and visual, either singly or in conjunction with other forms. As a result, semiotics enabled the detailed and comparatively objective analysis of popular cultural products such as advertising, television news, and newspaper layouts.

For most of the 1970s and much of the 1980s, British cultural studies busied itself with developing modes of textual analysis from a combination of semiotics, certain aspects of film theory, and Althusserian theories of ideology. The interest in the text was initially in critically unraveling its carriage of ideology. This slowly gave way to an attention that was more intrinsically interested in the text or genre of texts under examination—the appeal of soap opera, for example, or the construction of the news. Such work initially assumed that the readings generated by the cultural studies critic were immanent in the text, repeating an elitist assumption also made within literary studies that the critic had a privileged insight into what the 'true meanings' of popular culture might be.

The flaws behind this assumption were addressed by a number of interventions, the most significant being Hall's argument that the process of encoding or producing the text was separate and different from the process of decoding or 'reading' the text (1980). Indeed, he argued that there were at least three kinds of reading positions consumers might take when decoding the text. Consumers might simply accept what was described as the 'preferred' meaning; that is, the meaning apparently preferred by the producers or senders of the message or what is often referred to in the literature as the 'dominant' meaning. They might take an 'oppositional' position, which flatly rejects or subverts what they nevertheless would recognize as the preferred reading in order to construct their own, alternative reading of the message. Or they might produce a 'negotiated' reading, developed out of a relatively complicated consideration of the preferred reading and its alternatives. Hall's schema emphasized the contingency of the act of consumption, the varying ways in which audiences might understand any text. In practice, it was accepted, most would develop a negotiated reading rather than either of the other two options. This also emphasized how individualized the processes might be through which culture constructs its meanings. Consequently, cultural studies began to focus on these broader cultural processes rather than the processes of the texts themselves; this in turn was a shift away from the opinion of the critic and towards a serious consideration of the evidence from the audience.

Research into audiences' understanding of texts had occurred within the CCCS during the late 1970s, and the Centre's interest in subcultures had also produced a number of studies which used the methods of ethnography as a means of investigating how subcultural groups made sense of their social experiences (Hobson 1980, Willis 1977). However, it was not until Ien Ang's innovative analysis of fans of the US soap opera, *Dallas* (1985), that audience studies began the process of becoming a fundamental component of the practice of cultural studies. Ang's study used the letters from a group of *Dallas* fans both as texts to be analyzed and as ethnographic evidence which helped substantiate the nature of the soap opera's appeal to, for instance, intelligent and educated women who, at that time, would not have been reckoned to belong to the target audience for that kind of television. Something of an explosion of audience studies occurred, usually claiming to be in some way 'ethnographic' (that is, in this context, involving some direct observation of the group under analysis). The appropriation of what was a much reduced version of their research methodology provoked stern critiques from social science disciplines where the established protocols of ethnography were taken more seriously and employed more rigorously.

While the trend for what one might call 'thin' ethnography eventually waned, the usefulness of ethnography as a means of examining not only media audiences but other cultural formations has been accepted in cultural studies. David Morley (1986, 1992) moved audience studies away from largely providing a complementary strategy for those trying to understand the workings of media texts, and situated it as a means of, for instance, understanding the function of television within the home, and its implication in the web of social relations involving age and gender that made up families' everyday lives (see *Audiences*).

3. *Representation and Identity*

While the above account might imply that cultural studies' interest in the text has been displaced by the focus on audiences, there is at least one area where that is not true. Textual analysis in cultural studies was initially almost an end in itself; the development of a process which could unpack the construction of meaning within, particularly, media texts was rightly seen as a significant methodological advance. As a result, for quite some time, the performance of textual analysis was a dominant genre of research and writing within the literature of the field. As the authority of the critic was challenged by audience studies, however, the point and practice of textual analysis had to change.

Cultural studies had always been interested in how the world was 're-presented' to us in other forms: as images, as sounds, as writing, and as narrative. The influence of poststructuralist theories of language and culture situated representation as our only access to the real: we understood the world through languages, and through the ways in which it is represented to us. The problem of culture for many, therefore, is overwhelmingly the problem of representation.

One of the shifts in approach emerging from the mid-1980s interest in the audience was a focus on how cultural identity was represented. Feminist cultural studies was the primary instigator here, with its emphasis on how women's lives were determined and constrained by the dominant representational patterns within which the feminine is constructed (Franklin et al. 1991). The interest in identity as a contingent product of representation, and of the need for minority groups to therefore regard their representation as a political issue, becomes a major concern for cultural studies theory and a major genre of textual analysis.

It is also probably the location where the two complementary streams of British cultural studies—on the production and textual side, and on the consumption or audience side—have been most effectively and productively combined (Hall 1997). The focus on the representation of particular minority groups, and the interrogation of the effect of that representation upon particular examples of those groups, is precisely the territory that feminist cultural studies has developed. It has also been taken up by other groups—gays and lesbians, ethnic and racial minorities, diasporean communities—as a means of understanding the construction of their identities, and of contesting the political effects of that construction through a critique of the representational patterns involved. Finally, in the 1990s, the focus on representation has assisted in opening up the issue of the construction of national identity for British cultural studies.

4. *The Politics of British Cultural Studies*

One of the ways in which British cultural studies have conceptualized their project is to see it as inherently political. That is, British cultural studies set out to understand how culture works in order to intervene in that process and produce change. Williams, Hall, and other key figures were identified with New Left Marxism, and had played the roles of activitists within British politics. The process of developing the field, the work of 'theoretical clarification,' as it was called, has routinely addressed the problem of how to make academic knowledge matter within contemporary party politics. Indeed, when Hall (1988) addressed the problem of the Left's incapacity to combat the working class alignment with (then) Prime Minister Margaret Thatcher's policies, he presented it as an issue both for party policy and cultural theory.

The centrality of politics to the British cultural studies project is among the factors which may still

differentiate it from those more recent cultural studies traditions which have developed, for instance, in the USA. Indeed, among the concerns expressed by Hall and others about the success British cultural studies achieved in exporting itself to the USA at the end of the 1980s was the likely displacement of politics from the theoretical field. The widespread pattern of institutionalization of cultural studies, in the UK and elsewhere, has also been proposed as a threat to this fundamental principle. Similarly, if a little paradoxically given its direct engagement with political institutions, the incorporation of cultural studies academics into the process of state policy formation, particularly in the UK and Australia, and the consequent interest in the related field of cultural policy studies (Bennett 1998), has been criticized as a form of co-option.

The politics of British cultural studies itself is not without its blind spots, of course. It was reluctant to admit feminist concerns and paid little attention to issues of race, ethnicity or nationality until the 1990s. Its resistance to empirical research methods has exposed it to criticism from within the social sciences. However, its determination to make a highly theoretical enterprise prove itself in the material world of practical politics continues to mark it out as a distinctive intellectual and social project.

See also: Adolescents: Leisure-time Activities; Cultural Expression and Action; Cultural Geography; Cultural History; Cultural Policy: Outsider Art; Cultural Psychology; Cultural Studies: Cultural Concerns; Cultural Studies of Science; Culture, Sociology of; Mass Media and Cultural Identity; Media, Uses of; Popular Culture; Television: Genres; Western European Studies: Culture

Bibliography

Ang I 1985 *Watching Dallas: Soap Opera and the Melodramatic Imagination*. Methuen, London

Bennett T 1998 *Culture: A Reformer's Science*. Allen and Unwin, Sydney, Australia

Fiske J 1987 *Television Culture*. Methuen, London

Franklin S, Lury C, Stacey J 1991 *Off-Centre: Feminism and Cultural Studies*. Harper Collins, London

Hall S 1977 Culture, the media and the 'ideological effect'. In: Curran J, Gurevitch M, Woollacott J (eds.) *Mass Communication and Society*. Edward Arnold, London, pp. 315–48

Hall S 1980 Encoding/decoding. In: Hall S, Hobson D, Lowe A, Willis P (eds.) *Culture, Media, Language*. Hutchinson, London, pp. 128–38

Hall S 1988 *The Hard Road to Renewal: Thatcherism and the Crisis of the Left*. Verso, London

Hall S (ed.) 1997 *Representation: Cultural representations and Signifying Practices*. Sage, London

Hall S, Critcher C, Jefferson T, Clarke J, Robinson B 1978 *Policing the Crisis: Mugging, the State, and Law and Order*. Macmillan, London

Hebdige D 1979 *Subculture: The Meaning of Style*. Methuen, London

Hobson D 1980 *Crossroads: The Drama of a Soap Opera*. Methuen, London

Hoggart R 1958 *The Uses of Literacy*. Penguin, London

McGuigan J 1992 *Cultural Populism*. Routledge, London

Morley D 1986 *Family Television: Cultural Power and Domestic Leisure*. Comedia, London

Morley D 1992 *Television, Audiences and Cultural Studies*. Routledge, London

Turner G 1996 *British Cultural Studies: An Introduction*. Routledge, London

Williams R 1958 *Culture and Society 1780–1950*. Penguin, London

Willis P E 1977 *Learning to Labour: How Working Class Kids Get Working Class Jobs*. Saxon House, Farnborough, UK

G. Turner

Broadbent, Donald Eric (1926–93)

During the second half of the twentieth century Donald Broadbent's research and theories, communicated in 247 publications, including four books, shaped the direction of applied human experimental psychology. In Britain he exerted a broader influence, fostering the entire subject of psychology by publicly representing it as a significant science. He achieved this through advice on appointments to many influential chairs of departments and public committees and through representations to funding bodies that secured recognition, and thus funding for academic and applied psychology and, later, particularly for cognitive science. His influence continues through the careers of his many research students and postdoctoral collaborators. His death, distressingly soon after his retirement, deprived his family and friends of emotional and intellectual comfort and support, and curtailed sharp insights on how the intellectual history of the subject of psychology may determine its future.

Donald described a boyhood troubled by sharply fluctuating family fortunes, the early loss of a father whose talents greatly exceeded his educational and social opportunities, and the determination of his mother that Donald should transcend similar problems by gaining a scholarship to a prestigious public school, Winchester. There he found the physical sciences dispiriting because of the volume of existing knowledge to be mastered before any original work was possible. Nevertheless, a brief wartime course in engineering arguably provided the skills and mindset that developed his particular slant on psychology. He was called up to the RAF, where pilot training in the USA forced recognition that poor engineering design causes catastrophic human errors. Also that, in a

richer society than postwar Britain, these problems were beginning to be addressed by a profession to which he too might aspire. Experience of RAF personnel selection intensified his conviction of the value of occupational psychology, and he resumed his interrupted undergraduate career at Pembroke College, Cambridge. Here, against the strong advice of his tutors, he read psychology in Sir Frederick Bartlett's Department.

This was then, arguably, the best place in the world for him to be. Bartlett, and all his staff, had distinguished themselves in applied human factors research and personnel selection during the war, and so gained respect from other scientists for what was then a minute and misunderstood discipline. The Cambridge department was uniquely in the mainstream of new ideas. Donald has described it as an intellectually 'explosive' environment, similar to that in the Cavendish laboratories and the 'Bun Shop' pub just across Downing Street, where DNA was being discussed. Among the volatile ingredients were the immediate local legacy of Kenneth Craik's seminal applications of control theory to human performance; William Hick's brilliant use of Shannon and Weaver's new information theory to model decision making and his interpretation of its use in the recent wartime work of Egan, Garner, and Pierce; the speculative computational neuropsychology of Hebb's (1949) 'Organisation of Behaviour,' and demonstrations by Grey Walter and Ross Ashby that automata theory could simulate and explain the behaviour of simple organisms. This paradigm shift was still imperceptible anywhere else in Europe. After gaining a first-class degree, Donald had to explain to Bartlett that he could not afford to follow his advice to support himself independently to work on any problem that interested him. So he was appointed as a member of scientific staff at the Medical Research Council Applied Psychology Research Unit (APRU) in Cambridge, which Bartlett had founded and Norman Mackworth then directed.

The APRU was funded to solve applied problems for government agencies, principally the Royal Navy, who determined the research agenda. Mackworth asked Donald to work on noise, both as an obstacle to effective speech communication and a factor in attentional distraction, arousal, and stress. Noise became a main theme of Donald's research, from his first paper, 'Some effects of noise on visual performance' in 1951, to a concluding review, 'Human Performance and Noise,' published with Dylan Jones just before Broadbent's retirement 40 years later. It brought an exceptionally congenial and fruitful early collaboration with Peter Ladefoged on speech perception, and a lifelong interest in auditory psychophysics, and in the subjective and objective measurement of the effects of noise and other stressors on human comfort and performance. Most importantly, it provided applied problems that cued Donald's seminal work on auditory multitasking and dichotic listening, which crystallized in his pivotal achievement, his book *Perception and Communication* (Broadbent 1958).

Modern readers appreciate immediately that the remarkable impact of this book came partly from a wonderful lucidity of style and inspired choice of illustrative metaphors that clarifies difficult arguments. It is now impossible to convey its fresh delight for his contemporaries. For all outside a small circle in Cambridge this was the first insight into how an entirely new style of conceptual models, derived loosely from information-flow diagrams used by computer scientists, could make sense of exciting new discoveries in momentary and sustained selective attention, short-term memory, arousal, and the effects of stress. The subtext was that the dominant conceptual framework of S–R connectionism could not possibly account for these new data. Donald's next book, *Behaviour* (Broadbent 1961) illustrates his recognition of the power of the zeitgeist that *Perception and Communication* had to exorcise. It showed how the behaviorist framework derived from animal learning experiments had become inconsistent internally, and that much of the data it had engendered could now be reinterpreted usefully in terms of simple cybernetic models such as those currently being suggested by scientists such as Tony Deutsch and Stuart Sutherland.

In 1958 Donald was appointed director of the APU and attacked energetically the hard tasks he inherited, and even harder ones that he set himself. The funding base had to be widened beyond insecure reliance on Royal Navy contracts. Civilian clients, such as the Post Office, the Department of Transport, British Rail, and the Royal Mint were, arduously, brought to recognize practical benefits from human factors research. Part of a self-imposed task of improving public understanding and support in the UK for a tiny and misapprehended subject was public appearances on radio and television. Effective representation of psychology through and within the British Association for the Advancement of Science, work with committees and boards of the MRC and the other, newer, Research Councils enlarged academic respect, and secured better funding for psychology in Britain. Similar efforts in the British Psychological Society and the Experimental Psychology Society brought him presidency of both bodies—and membership of even more committees. There was also the day-to-day administration of a rapidly expanding, and increasingly demanding, unit.

Donald always insisted that it would have been impossible to continue his personal research without Margaret Gregory, who began to be his lifelong collaborator in 1961, and his wife in 1972. Apart from a diversity of papers on auditory and visual selective attention, decision processes and stress, his main work during the 1960s became reinterpretation of data from vigilance tasks in terms of the new metric of Tanner

and Swets' Signal Detection Theory. Donald and Margaret found that small shifts in operator bias, which could be related neatly to maintenance of arousal and motivation, brought about counter-intuitively large changes in patterns of errors. This suggested fresh ways of analyzing the effects on sustained attention of stressors such as fatigue and noise, and of personality variables such as introversion and extraversion. It also led to the speculations on distinctions between selective perceptual sensitization for particular signals and preparatory bias for particular responses that became central to his next book, the monumental *Decision and Stress* (Broadbent 1971), and guided much of his later research. However, administrative loads and personal problems became overwhelming. A sabbatical at All Souls College, Oxford allowed personal reassessment and completion of 'Decision and Stress' which appeared in 1971. Perhaps election to the Royal Society also brought new confidence in public esteem for his individual contributions, independently of the background successes of the unit he had built up. He gave three years' notice of resignation, and in 1974 moved to the Department of Experimental Psychology in Oxford, which was then emerging into a period of pre-eminence in Britain under Larry Weiskrantz. Here, until his retirement in 1991, he at last had his chance to follow Bartlett's advice to work full time on whatever problems he wished.

Decision and Stress achieved immediate widespread respect for its scholarship and incomprehension of its message. Its central theme is a model for the way information is processed, remembered, and retrieved, expressed metaphorically as a distinction between 'pigeonholing' and 'categorization.' At first sight this seems as helpful as all Donald's other heuristic devices, but generations of students have failed to discover precisely what it means. It seems partly to contrast control of attention by selective tuning of perceptual, or attentional, sensitivity, and by response bias. It is stressed that these functional processes must be guided 'from the top down,' by the complex mental representations that people make of their environments, their momentary and continuing goals, and of the information that they need and the actions that they must perform to achieve these, rather than from 'the bottom up' by passive assimilation of sensory input. There is still no better reference source for the entire body of experimental work in applied experimental psychology between 1958 and 1970, but the metaphor did not provide the necessary guiding thread through this vast literature, or extract from it messages as concise and fruitful as those that *Perception and Communication* had drawn from no more than 20 key papers. Hindsight shows that the real problem was the familiar one in any rapidly developing science: the style of models that had been so successful in 1958 no longer worked for the new data base that they had generated. Issues of the inevitably increasing clumsiness of explanatory frameworks as further data accumulate, and of the best level of description at which to work, cut deep in applied cognitive psychology.

For the rest of his working life, Donald continued his commentary on intellectual navigation between finicky debates about the increasingly sterile details of models based on single experimental paradigms, and theoretical frameworks that are so general that it is unclear what they explain or how they can be tested. This was one of the themes of his William James lectures at Harvard, published as *In Defence of Empirical Psychology* (Broadbent 1973). A background stimulus was a fashion, during the 1970s, for rejection of empiricism in favour of 'cognitivist' descriptions. Donald's response to the new 'cognitive psychology' named, but not defined, by Neisser in 1966, had been cautiously tepid. He feared that it might encourage less gifted practitioners to a psychology based on pseudophilosophical speculation about phenomenology rather than on empirical evidence. This motivated an expression of concern at these trends in British psychology in a paper 'In defence of empirical psychology' (Broadbent 1970) published in the *Bulletin of the British Psychological Society*. Donald detested academic affectation, particularly grandiloquent posturing to win the generous hearts of students who had not yet been given the critical apparatus necessary to distinguish efflorescing cults of personality from the necessary tool kit of methodology and hard-won information in a science that many of them had resisted opposition to be able to study. He saw a threat to civilized values when, as he said, 'If one doesn't argue from evidence it must come to bludgeons.' Perhaps much of his distaste came from what he called a puritanical streak, which took him directly athwart the spirit of the 1970s: 'I decided then, and still believe, that self-realisation and the development of personal experience are neither dignified nor respectable goals in life, that most of the world lives within extremely tight economic margins and that positions of privilege, (such as the conduct of psychological research) demand obligations in return.'

During the 1970s and 1980s he commented shrewdly on early connectionist models by McClelland and Rumelhardt, on parallels between human experts and expert computer systems, and became concerned with levels, hierarchies and the locus of control, and the effects of emotional states and personality traits on selective attention and decision making. He resonated to the strong neuroscience bias that Larry Weiskrantz had brought to the Oxford University Department. He proposed a 'Maltese Cross' model to describe interactions between modality-based memory subsystems and their mutual control to achieve selective attention. Empirical research continued on a remarkable breadth of topics. Work on how the emotional valence of words affects the speed with which they can be recognized, on more general problems of word perception, on visual perceptual selectivity under

stress and distraction with Margaret Broadbent and Sue Gathercole, on the processing of rapid strings of successive visual stimuli, and on grouping effects in short-term memory with other distinguished collaborators.

Donald never lost the determination to apply psychology to economically and socially useful ends, and to enrich theory by studying complex real-life tasks rather than artificially simple experimental paradigms. With Diane Berry he investigated how managers develop and use mental representations of the systems that they try to control, providing empirical evidence that their decisions are affected strongly by assumptions, biases, and even computational procedures, which they do not articulate, and of which they may even remain unaware. Parallel work on errors and lapses took place in hospital wards and on automobile production lines, in collaboration with his student Peter Cooper, and with Dennis Gath of the Oxford University Department of Psychiatry. It is a pity that this work has gained far less attention than it deserves, because it teaches precisely those lessons of empirical methodology and acceptance of social responsibility that he most wished to pass on.

Donald felt that his so-called 'puritanical streak' discomforted his younger colleagues, but this was not quite the issue. He could not endure gladly the company of people whom he took to be insincere, or perverse, in their determination to avoid recognition of logical weaknesses in their positions, but he was always as fair and generous to them as he could force himself to be. Discomfort did not come from his humane responses but from glimpses of the considerable effort that they sometimes cost him. His intellectual acuity was often intimidating, as was occasional looming impatience with remarks or behaviour that he interpreted as histrionic or unkind. His genuine shyness was disconcerting in one so obviously, and exceedingly, clever and eminent, and sometimes compounded problems of communication, but his underlying kindliness was unmistakable, totally reliable, and amply assuaged any occasional unease. He was frugal with praise, perhaps because he was suspicious of merely verbal gestures, but he was endlessly generous and sensitively proactive in offering practical help. It is excellent that he retired sure of the intellectual esteem of his colleagues and students, but sad that he remained unaware that he also earned far more affection than he ever realized.

See also: Attention: Models; Behaviorism; Behaviorism, History of; Cognitive Neuropsychology, Methodology of; Cognitive Psychology: History; Cognitive Psychology: Overview; Experimentation in Psychology, History of; Hebb, Donald Olding (1904–85); Psychology: Historical and Cultural Perspectives; Psychology: Overview; Stress and Coping Theories; Stress: Psychological Perspectives

Bibliography

Broadbent D E 1958 *Perception and Communication*. Pergamon Press, Oxford, UK

Broadbent D E 1961 *Behaviour*. Basic Books, New York

Broadbent D E 1970 In defence of empirical psychology. *Bulletin of the British Psychological Society* **23**(79): 87–96

Broadbent D E 1971 *Decision and Stress*. Academic Press, London

Broadbent D E 1973 *In Defence of Empirical Psychology*. Oxford University Press, Oxford, UK

Broadbent D E 1984 The Maltese Cross: A new simplistic model for memory. *Behavioural and Brain Sciences* **7**(1): 55–94

P. Rabbitt

Broadcasting: General

Broadcasting refers to the social application of two related technologies of communication; radio and, later, television. Radio broadcasting, which began in the USA and Europe in the early 1920s, was a by-product of wireless telephony which was in turn an extension of wired telephony and telegraphy. All earlier electronic technologies of communication, whether wired or wire less, were designed expressly to transmit messages from point A to a distant receiver at point B. They were conceived as closed systems whose information was only for specific senders and receivers. Wireless telephony was developed as an extension of wired telephony with the obvious advantage that it could transmit messages to distant points (to ships at sea, for instance, or to armies in the field) that could not be linked by wire. However, this new technology had an apparent drawback. Anyone within range of the transmitted wireless signal could also 'tap into' the message if they had appropriate equipment. This became a major problem for all contending armies in World War I, all of whom used wireless telephony to maintain links between headquarters and their fronts and all of whom were obliged to design sophisticated codes to encrypt their messages which otherwise would give information to the enemy. It was only after the Great War that the *broadcast* effect of wireless telephony, which scattered its message across a wide radius, began to be exploited as a quite new kind of communication for general social purposes.

Radio broadcasting became a popular hobby, in the USA and Europe, in the early twentieth century. The technology was, in principle, a two-way system of transmission, permitting live interaction between two sender-receivers. In countless suburban bedrooms and garden sheds male members of the household, young and old, became enthusiastic radio 'hams.' Some put out programs of music at particular times to which

others, within range of reception, could tune in and listen. At the same time the electronics industry, then in its infancy, began to perceive the possibilities of a market for the sale of radio receiving equipment, but this presupposed that there was something to which purchasers of sets might listen. Broadcasting as we know it today began with the separation of transmission and reception. People would buy radio sets that only received, but could not transmit a broadcast signal, if there was something 'in the air' to which they could tune their receiving set and listen.

Simple though this sounds it contains issues of enormous complexity. The delivery of broadcast services depends on technological, legal, political, economic, and cultural factors all of which are enmeshed in each other. Take the most basic. How is a broadcast program service to be financed? The pay mechanism for other informational or entertainment services is obvious; you pay for the newspaper or at the box office for entry to the cinema or concert hall. But radio and television developed as services delivered to households, rather like gas or water. People paid for their radio sets. But how could they be made to pay for what they listened to? Today the 'box office' mechanism is in place and receivers of cable and satellite services operate on a pay-per-channel or pay-per-view principle. But for many years terrestrial broadcasting was a 'free to air' service, and the cost of the program service was met either by the sale of on-air advertising or by some kind of license fee, an annual charge levied and collected by central government on households with radio or television sets. In the USA broadcasting developed via the first route: as a commercial system funded by advertising and sponsorship revenue. In Europe, and classically in Great Britain, broadcasting developed as a public service, funded by a license fee which was levied by the Post Office, a department of state.

These two alternative models for the delivery of broadcast services have defined the first two 'ages' of broadcasting; the first age being the radio era, the dominant broadcast medium until the beginning of the 1950s in the USA and the end of the 1950s in Europe, when it was surpassed by 'the age of television.' Separately and together broadcast radio and television expanded worldwide in the second half of the twentieth century to become truly global systems of communication. The broadcast services that developed in the USA and Britain in the 1930s, and which today are still major players in an increasingly complex media environment, form the focus of this article.

Although radio broadcasting started up in different ways in different countries the nature of the program service on offer turns out to be everywhere the same: a mix of information and entertainment delivered through the day (and night) day by day and every day. The nature of this service imposes fundamental constraints upon the agencies that deliver it. Broadcasting is labor intensive. The delivery of news is a complex, costly affair if the broadcaster is in the business of actually gathering and producing the news that it transmits, and the same is true of broadcast entertainment, if it draws on professional, 'star' talent, and especially if it produces fictional drama. Thus we find, in early histories, that although radio started up in many places as a small local service, sooner or later there was a process of convergence whereby national systems of broadcasting developed—the commercial networks in the USA, and the British Broadcasting Corporation (BBC) and similar public service institutions in Europe and elsewhere. This convergence was driven largely by the necessary economies of scale that underpin the delivery of labor-intensive program services to very large audiences through the day and continuously.

When broadcasting began, electricity was a very new energy source that was not widely available throughout societies. In the 1920s most houses were wired only for electric light. In the 1930s they began to be wired for mains electricity so that electric appliances could run from them. But before the World War II, the 'wired home' as we know it today was restricted largely to urban centers and more affluent households. Electric goods such as cookers, refrigerators, record players, and washing machines only became widely available throughout societies in the 1950s. The development of a culture in which radio and television sets were part of everyday existence for everyone was part of the more general social process whereby domestic life was transformed in the twentieth century, as the home increasingly became the primary focus of leisure, relaxation, and entertainment across all social sectors.

But why should people want to listen to radio or watch television? First, it should be something that is easily available for them to do. Second, it should be something that they will want to go on doing. The availability of broadcast services is, in the first place, a technological matter. On the one hand there is the delivery of the signal, and on the other, the design of the receiving equipment. In terms of transmission there are two key issues: distance and quality. The greater the reach of the signal the more economic the delivery of the service. But at the same time the quality of the signal must be such that listening or watching is of sustained high standard, free of technical 'noise,' fade or interference. In terms of the receiving equipment it must be simple and easy to use and not liable to go wrong or to need frequent renovation.

It took time to resolve these basic problems. Early transmitters sent out a low-power signal with a reach of no more than 20 miles. Soon more powerful transmitters could transmit over much greater distances, but coverage of a continent such as the USA required a vast network of transmitters and telephone 'long lines' to relay the signal over thousands of miles. At the same time interference from neighboring

stations was initially a serious problem in both Europe and the USA where it was known as 'the chaos of the ether.' The Federal Radio Commission was established in 1927 as a central regulatory body responsible for the allocation of station frequencies across the whole of the United States. In Europe continent-wide agreement on the number of frequencies allowed to each nation state was reached by the leading broadcasters from 16 countries at Geneva in 1926 and formally ratified by their governments at the Prague Conference in 1929. The BBC which, until then had been broadcasting on 20 different frequencies, agreed to halve this number and undertook a fundamental redesign of its services, aiming to achieve nation-wide coverage with five high-powered twin transmitters strategically sited throughout the United Kingdom.

While these matters were being resolved the radio set as we know it today began to emerge. The earliest radio sets were sold as kits of electronic equipment which required some knowledge to assemble and some skill to tune in and pick up a signal. As a domestic good, rather than a scientific hobby, the radio set needed to be not only simple to use but acceptable to look at. It must appear as part of the domestic space in which it was situated. By the 1930s radio sets with a frequency tuner and volume control and housed in a molded plastic case were in mass production, and became the biggest selling electric appliance designed for domestic use. By 1939, 75 percent of British households had a radio, far ahead of the electric iron and carpet cleaner which were the next most purchased electric appliances.

What then did people want to listen to, having purchased their radio set? The whole of the broadcasting industry worldwide exists in order to deliver something that will make people turn on their receiving sets and keep them on. What is involved in this? We must first understand the basic characteristic of both radio and television as *time-based* media. When you turn on the set you expect there to be something to listen to or watch whenever you do so, not just occasionally or only at certain times. There must, in fact, always be something 'on' for listeners and viewers. So that from the start broadcasters were impelled to provide a service that ran continuously through the day and into the evening, and eventually all through the night as well. The most fundamental problem facing broadcasters is how to fill *air time*. How do you supply a program service that will run day in day out and through the day and night without ceasing?

Let us link this to the expectations of listeners and viewers. What did people want to listen to? What did they get? Radio in its very early days was a hit-and-miss amateurish affair. It could supply talking voices and music drawn from whatever local resources the harassed station operator could muster, and that usually was not very much. There was a demand from the start, everywhere, for broadcasting to link people to the centers of existing public life from which the vast majority were, of necessity, excluded. The time base of broadcast media is real time, *now*. The real-time property of radio and television meant that, from the start, listeners and viewers could in principle be linked to major events even as they happened: to a baseball or soccer game, a presidential inauguration, a royal wedding, a major 'breaking' news event live and as it happened. More generally, broadcasting could connect listeners to hitherto inaccessible cultural resources. Before radio only a tiny minority had ever heard a full symphony orchestra playing the classical repertoire. By the mid-1930s the Boston Symphony Orchestra, the Manchester Hallé, or the Berlin Philharmonic could be presented on air and heard throughout the land. Contemporary urban, urbane music—jazz, swing, dance music live from night clubs and smart metropolitan hotels—resounded in places far removed from big city life.

The demand for access to these things was irresistible and everywhere impelled broadcasting to centralize and concentrate its resources in order to provide high-cost entertainment, news, and access to great events for rapidly growing listening publics. At the same time broadcasting was compelled to invent new forms of entertainment that could meet the insatiable demands of continuous daily program services. Before broadcasting professional entertainers (singers, comedians) made their living by repeating the same limited repertoire as they toured the country, appearing in local variety theatres or concert halls. Star performers were signed up for radio only to discover that after one or two appearances they were burned out. Their repertoire, which sufficed for a lifetime of performing to audiences of at most hundreds, did not bear repetition anywhere after it had been heard once, simultaneously, by millions. Thus, the basic production problem that broadcasting faced was to discover serial formats that permitted the indefinite repetition of the same basic material yet which would be different each and every time. Key inventions in this respect were the serialized 'never ending' fictional narrative (or soap opera), and the 'situation comedy' which allowed for endless variations on the same basic comic theme. In time program formatting would extend to news, game shows, quizzes, and talk shows. Today all broadcasting, with the exception of singular live-to-air events (and even these are routinized in American news programs), is formatted on a principle that allows any program, in principle, to run indefinitely.

Program formatting is one key solution to the fundamental problems which time-based daily media pose for broadcasters. The other, and intimately linked to this, is scheduling. Scheduling is to do with the sequencing of programs through the day, and is crucial to the familiarity of daily broadcast services. This familiarity is given by the ways in programs are so arranged that they appear, as it seems to us, on the

right channel on the right day at the right time of day (or night) and go on for the right length of time. The routinization of schedules into settled, knowable sequences was accomplished by the 1930s and has the effect of adjusting the times of broadcasting to the times of daily life and its routines. Thus, the time of broadcasting interacts with and supports the temporal structures of day-to-day life in modern societies. The time of any program is always attuned to the time of day. Early morning radio and television shows have names like *Good Morning America*, *Today*, or *The Big Breakfast Show* and the fundamental task they perform is that of orienting whole populations to the ordinary business of the day ahead. What has happened since yesterday? What is happening now or later today? What's the weather and the traffic like 'out there'? What's the time? (It's time to get going). In dealing with all this such shows are not just *at* a certain time of day. They are *for* that time. The schedules are not just one damned thing after another, but have a temporal dynamic; not just the time of day, but time through the day: time to go to work and time to return (drive time and its attendant programs) and time to relax.

There is never 'empty' time in broadcasting. It is always filled with something going on. In the world as shown by broadcasting things are always happening, and thus what radio and television have accomplished unobtrusively is the routinization of history, day by day and every day. The formatting of programs and schedules combine to overcome the problem of the ever same (about which there is nothing to say). Broadcasting is characterized by difference in sameness. Radio and television combine and interact with each other (the times for listening and watching are not the same) to respecify each and every day as *this* day in which this is happening and these things are going on. Each day is like any other day and yet it is particular.

Particular for whom? Broadcasting began by separating production and transmission from reception. In so doing it created quite new kinds of audience, thereby redefining the character of public life and the boundaries between public and private existence. An audience usually is understood as a collectively gathered for some common purpose of witnessing, for instance, a concert, a play, or a movie. To be an audience member is to be present at a public event of some kind. Participation in public life, before broadcasting, had of necessity demanded some investment of time, effort, and often money. It meant journeying to the location or building dedicated to the event. It meant, moreover that the audience was always a particular kind of public, constituted by the nature of the occasion: a church congregation, a football crowd, a cinema audience. Now, via radio and television, the world 'out there'—the great world of public events and persons—entered into the spaces of everyday existence for untold millions. Whereas previously public life had been available only to particular and self-selecting publics, it now became generally available to a new kind of general public equivalent, on certain occasions, to the whole society and today even the whole world. Events themselves, hitherto discrete and catering to specific 'taste publics,' now became part of a shared common life enjoyed by millions who beforehand—of necessity or lack of interest, time, or money—had not participated in them. The common culture of broadcasting is defined by its nonexclusive, 'free to air,' general availability. Radio and television straddle two worlds: the great world of public life and affairs and the private worlds of individual social members. These two worlds are quite distinct and separate. Before broadcasting the horizon of concern for most ordinary social members extended no further than the immediate environment in which they dwelt. Public life and affairs were, for the majority, remote from day-to-day existence and impinged upon it only occasionally and unexpectedly. But broadcasting links these two worlds, routinely and continuously. How does it do this?

In an obvious way what links the public world of broadcasting and the private worlds of viewers and listeners is speech; the on-air talk that goes out on radio and television every day. This talk is, by now, so familiar and seemingly natural that it is hard to see that the question of how to address their audience was once a crucial issue for broadcasters. Broadcasting, as we have noted, separates production and reception. Broadcast talk is produced in the studio or on location but, either way, is always heard elsewhere. This means that although broadcasters control the talk which they produce, they cannot control how their absent audiences will react to that talk. This is more generally true of all live-to-air events and occasions broadcast either from the studio or on location. Live and present audiences in any event are called upon to adjust their behaviors to the nature of the occasion which is controlled and managed by the public speaker-performers. But no such constraints fall upon listeners or viewers. Thus, the key to understanding the communicative character of broadcasting is to see that the relationship between broadcasters and audiences is essentially noncoercive, because the former cannot control the behavior of the latter.

What are the consequences of this? It reverses the prevailing norms of public life in which audiences are required to adjust their behaviours to the nature of the occasion. This means, in effect, that they submit to being controlled by the performers who manage the performance. Broadcasters learnt that on the contrary they must adjust their behaviours to chime in with the situations and circumstances of listeners and viewers. Overwhelmingly watching and listening are not collective 'mass' activities. It is true that, in the early years of radio when only few people had sets, certain public events and ceremonies were relayed by loudspeakers in public spaces in which thousands congregated to

listen. Fascist radio was used in this way, in the 1930s, in Germany and Italy. Today the crowds at a major sporting event or ceremony may have simultaneous access to coverage of that at which they are present via giant television screens. But by and large people in the past listened to radio and today watch television in, as we say, 'the privacy of their own homes.' Viewing and listening are not work-related activities but are linked to what we think of as 'free time,' to leisure, and domesticity. Moreover, they are essentially individual activities which may be shared with family or friends.

In such circumstances, how would one expect to be spoken to? Most forms of public speech have some propagandist or persuasive aim. But in the contexts of domestic life and leisure people do not want to be preached to or lectured at when they turn on the television or radio. They want to relax and they expect to be spoken to as equals, and as individuals in their own right, not as anonymous members of a collectivity. Thus, the communicative style of broadcasting on radio and television everywhere came to orient itself to the norms of ordinary conversation, or talk. Those norms, in turn, have unobtrusively transformed the character of public life and entertainment in the twentieth century.

Consider the transformation of singing by the microphone. The microphone is a technology that amplifies and relays sound. It is fundamental to the recording and broadcasting industries. Before its invention the norms of public singing in Europe and America were defined by the requirements of opera, in which the voice had to project over a full orchestra and to the farthest reaches of large opera houses. In this style of singing, necessarily, the power and vocal range of the voice were especially prized. The operatic voice was designed specifically for performance to live, mass audiences of hundreds if not thousands.

When, in the early years of broadcasting, such singers came to the microphone to perform they almost short circuited the equipment when they gave voice. It was discovered immediately that to produce their sound on air or on disc the singers needed to be positioned several feet away from the microphone. However, a new style of singing began to develop which exploited the potential of the microphone. Here the singer stood close to the mike and, instead of bellowing, seemingly spoke or whispered into it. This 'talking' voice repersonalised singing. The crooner did not need a trained operatic voice in order to perform. He or she sang in their own, natural voice. So that what you heard was the voice *of* Bing Crosby or Frank Sinatra and they sang as if to you alone. Crooning brought intimacy into the public domain: instead of loudness, softness; instead of singing to many, singing to one; instead of the impersonal, the personal; instead of extraordinary, ordinary voices. For the singing voice, broadcasting and the record industry made sincerity the authenticating measure of intimacy as it was transposed from the private into the public domain. Just as the movie (and later, television) close-up offers to countless millions an intimate access to the human face and an exchange of looks hitherto available only between lovers or a parent and their little child, so close-miked singing produces similar effects for us as we listen.

The transformation of singing by the microphone is indicative of a much more general process whereby the personal norms of interpersonal life—of intimacy, sincerity, and authenticity—have been transposed into the public domain. Franklin Roosevelt was the first major political leader to exploit these communicative characteristics. His 'fireside chats' with American listeners in the 1930s were the reverse of the fiery public rhetoric of European dictators. They exploited the intimacy of broadcasting and established a more egalitarian relationship with the electorate. In the course of the twentieth broadcasting has become folded inextricably into the political process everywhere. It provides politicians with unparalleled access to those they wish to influence and, reciprocally, it makes them visible to their constituencies in ways that were hitherto impossible.

The growing involvement of politics with broadcasting is a decisive characteristic of the age of television in the second half of the twentieth century. Television makes politicians visible, and in close-up, and thus the management of visibility becomes increasingly important for them. Viewers judge by appearance and performance, as much as the substance of policy or the soundness of argument. They evaluate the person as well as the policies. The famous Nixon–Kennedy debate on television in 1960 was perhaps the defining moment when politicians, not just in America, came to realise that they must attend to their self-presentation and image on television if they wished to win electoral support. Margaret Thatcher, before she came to power in 1979, had a make-over, largely for television, which involved restyling her hair, her wardrobe, and her voice. Image management has become increasingly important in the last quarter of the twentieth century as individual politicians, parties, and governments all seek to put the best spin on themselves and their activities as they appear on television.

1. Coda

In *Waiting for the Rain*, written in 1975, the Zimbabwean novelist, Charles Mungoshi, describes the astonishment of Sekuru, an old man who lives up country, when he hears his grandson's transistor radio. He is amazed to hear his own language, Shona, magically emanating from this small black box and wonders (to the vast amusement of his city dwelling grandson) if the voices he hears are spirits who somehow live inside it. He listens to what the voice is saying:

> 'And to end the news, here are the main points again. Israeli troops have bombed an oil dump in the Suez Canal area. President Karanga has been shot dead. Rebel soldiers are still searching for President Bomba who escaped after a successful coup at dawn … '
> 'What's all this?'
> 'It's the news, Sekuru.'
> 'Where is all this happening?'
> 'All over the world, Sekuru. All over the world.'

The technologies of broadcasting, through their live immediacy, make the world available *as* the world. But it is not just that what is happening in the world now becomes generally available with an unprecedented immediacy. The relationship between events and their telling changes. History *is* this relationship. Unless a happening is brought to articulation—made into, reported, and recorded as a tellable story—it has no historic significance. Narrative is the discovery *of* the historical. It is this that has been routinized, on a daily basis, by modern news media. But radio and television have a specific temporal dynamic that earlier news media did not possess.

In written news stories (in news photographs or film) there is a necessary, inevitable gap between events as they happen and their postproduction as record or narrative. The temporality of written history moves on an axis from present to past. The time of the event and the time of its telling are always moving away from each other, and hence history is always the history of a receding past. But with radio and television the time of the event and the time of its telling coincide. Both exist in the same phenomenal real-time now. Broadcasting does not stand apart from the events that it covers. It is caught up in them, interacting with what is happening and, on ocasion, affecting how they unfold in real present time. This was widely noted in the dramatic sequence of events, in one country after another, as communism collapsed in Russia and Eastern Europe at the end of the 1980s. In these events television did not merely show what was happening. It contributed to what was happening. In this way broadcasting contributes to the making of history rather than acting merely as its recording angel. The temporal axis of both event and narrative in live broadcasting is one that moves from present to future. History is no longer 'then.' It is 'now.' It is no longer 'there,' but 'here.' The now-and-then, the here-and-there come together in the live immediacy of broadcast news and events which are structured in expectancy of what is to come.

Thus, the real-time basis of radio and television is a specific kind of temporality in which individual lives and the historical life of societies intersect with and interact upon each other routinely in the common public, available now of daily ratio and television output. How they do so is everywhere a matter of continuing concern. The impact of broadcasting on all our lives has been from the start and will continue to be something that is argued over and evaluated positively or negatively from many differing perspectives all over the world. But this in itself is indicative of what broadcasting has delivered unobtrusively in the course of the twentieth century: a world in which the past and the future come together and encounter each other in the live, immediate, unfolding now as a common, available world of concern.

See also: Audience Measurement; Entertainment; Mass Media, Political Economy of; Media Effects; Media, Uses of; News: General; Political Communication; Radio as Medium; Television: History; Television: Industry

Bibliography

Dayan D, Katz E 1992 *Media Events*. Harvard University Press, Cambridge MA

Forty A 1986 *Objects of Desire, Design and Society 1750–1980*. Thames and Hudson, London

Heidegger M 1962 *Being and Time*. SCM Press, London

Mungoshi C 1975 *Waiting for the Rain*. Heinexann Educational, London

Peters J D 1999 *Speaking Into the Air. A History of the Idea of Communication*. University of Chicago Press, Chicago

Ricoeur P 1984–88 *Time and Narrative*, 3 vols. University of Chicago Press, Chicago

Scannell P (ed.) 1991 *Broadcast Talk*. Sage, London

Scannell P 1996 *Radio, Television and Modern Life*. Blackwell, Oxford, UK

Schudson M 1998 *The Good Citizen. A History of American Civic Life*. Martin Kessler, New York

Thompson J B 1995 *The Media and Modernity*. Polity Press, Oxford, UK

Williams R 1974 *Television: Technology and Cultural Form*. Fontana, London

P. Scannell

Broadcasting: Regulation

By *regulation* we refer to political decisions and legal and economic instruments by which broadcasting activities are governed. These instruments include laws, rules, statutes, concessions, licenses, and traditions that make up the legal and economic framework; that is, the *regulatory regimes* within which broadcasting activities are carried out. Broadcasting regulations are outcomes of political decisions on broadcasting policy. Naturally, the regulation of broadcasting activities differs in time and space. First and foremost, broadcasting regulations vary between countries, as they reflect *national* political, cultural, social, and economic conditions. Still, in 2000, when the globalization of the media industry has become

conspicuous, is it fair to state that regulation of broadcasting to a considerable degree remains within the hands of individual states. This is true even if we consider the attempts of international organizations to establish international regulations for transnational television and trade in audiovisual products.

1. Distinctive Criteria

One main distinction between regulatory systems can be drawn between *democratic* and *authoritarian* states, which typically will regulate broadcasting differently with regard to freedom of expression but not necessarily with regard to ownership and market behavior. That is, the systems will vary according to the editorial freedom granted to media institutions: in democratic systems, the media are guaranteed a high degree of autonomy, editorial freedom, and freedom from censorship, persecution, and sanctions. Authoritarian states have generally repressive politics towards the media: either the media become instruments for the incumbent regimes, or they are heavily censored. The distinction between authoritarian and democratic not coincidental with the degree of state intervention but is characterized by the *distance* between authorities and media institutions in terms of editorial influence. To illustrate: there are many examples of state-owned broadcasting institutions that have or have had a national monopoly and which have had a high degree of editorial freedom (e.g., the British Broadcasting Corporation (BBC); the Swedish Broadcasting Corporation, Sveriges Radio), the Japanese Broadcasting of postwar Japan, NHK). There are also examples of countries where privately owned and market-driven broadcasting institutions have been heavily censored and worked as instruments of authoritarian governments, such as was the case with Brazilian Globo during the military government in the 1960s.

Another main distinction can be drawn between states with a *free-market economy* and states with a more *interventionist* tradition. The first type tends towards regulating broadcasting as an industry with the objective of securing competition and avoiding market failure. The latter type of states tends to regard broadcasting and other communications as a *public good*. Accordingly, *broadcasting policy* or *broadcasting legislation* in these countries is a more common term than regulation. In most European countries the state has owned and run broadcasting institutions operating with a national monopoly or duopoly. In these countries, there were few or no distinctions between the state as the regulator of broadcasting markets and the state as actor in the broadcasting sector. Instead, media policy and broadcasting policy were treated as a central topic within the larger area of *cultural policy*; that is, together with political issues related to maintaining and strengthening national identities, language, history, and culture. As such, in Europe broadcasting was and to a large extent still is an area considered as a sector for production of cultural goods more than cultural commodities.

In countries with a less interventionist tradition, such as the USA, the willingness to place broadcasting under cultural policy has been less prevalent. Broadcasting and other media activities are regarded as an area for free enterprise; that is, as a *market* for exchange of audiovisual products. Ideally, markets should regulate themselves, but since problems such as scarcity of frequencies, dominance of one or a few actors, and over-establishment led to market failure, broadcasting markets have been regulated with respect to *entrance* and *dominance*.

2. Regulatory Models: Public Service, Market Regulations, and Mixed Systems

2.1 Broadcasting Regulated as a Public Service

In Western Europe, broadcasting was early regulated as a *public service*; that is, as a service provided by the state or a publicly regulated institution. Public services are goods provided to the entire population on fairly equal conditions and that should be accessible for all at a relatively low cost. Most European countries centralized their broadcasting services in one or a few institutions either when broadcasting was introduced in the national context or shortly after. The discourse on public service broadcasting sometimes seems to imply that there exists *one* model, yet it is worth noting that a number of different systems and institutions are subsumed under the concept. The public service model thus varied from country to country, reflecting the cultural and political characteristics of different countries (Skogerbø 1996). In some, but not all, European countries they held a national monopoly and provided universal coverage and a program menu consisting of enlightenment, education, and entertainment that should appeal to different social and cultural groups. In some countries, license fees were the sole source of revenues for the public broadcasters; in others they drew their incomes from advertising or collected revenues from both these sources.

Accordingly, public broadcasting institutions were designed to cater to a diverse number of national interests (management of scarce resources, security, education, nation-building, news) as they were set up in many European countries in the 1920s and 1930s. In that period, the use of centralized institutions in order to solve problems that were regarded as in the national interest was quite common, especially in Europe and in particular in communications. Many countries had centralized telegraph and telephone services, and if not, the idea that communication services should be provided on a universal basis was widespread.

Within the public service model, there was room for quite large variations. Not all countries had a centralized broadcasting institution. The Dutch public

service system, for instance, incorporates a number of different broadcasters that have in common that they are affiliated with the traditional social *pillars* of Dutch society; that is, the religious and social cleavages that until the 1960s divided Dutch society into a number of strictly separate groups. Each of these groups or pillars established 'broadcasting societies' that were allowed the right to operate their own radio station and, at a later stage, produced programs for television. These broadcasters had the right to collect license fees, and from 1967 they also had revenues from advertising, allocated according to a complicated system. In 1967, a broadcaster providing news and sports was established in order to provide independent programming (*Nederlandse Omroep Stichting*, NOS), too. The allocation of, first, radio frequencies and later 'windows' on the national television channels to 'broadcasting societies' took care of the needs to cater to different audience segments, whereas NOS provided the audience with common references to the national culture, politics, and society.

2.2 Broadcasting Regulation as Market Regulation

Whereas the regulation of broadcasting as a public service can be said to be a European invention that has been exported to a number of other countries in Africa and Asia, the model of regulating of broadcasting as a competitive industry has its deepest roots in the USA. In the USA, a distinct *broadcasting policy* was never formulated. According to some scholars and judges the First Amendment of the American Constitution (which states that the federal government shall 'make no law abridging freedom of speech, or of the press' prohibits even a government's statement of broadcast goals and protects as far as possible the rights of private persons and even private corporations in the marketplace to speak as they choose. The European invention of vesting the responsibility for production of diverse programs in one or a few public institutions was adverse to the American conception that diversity and pluralism is a product of freedom of expression interpreted as freedom to establish media institutions.

Regulations were implemented in order to prevent, or rather to avoid, 'chaos on the ether' that resulted from the shortage of frequencies and lack of regulation in the 1920s. In the USA, the early period of chaos led to implementation of a regulatory regime aimed at correcting market failure in the form of chaos and scarcity, and to avoid monopolization and dominance. The networks, however, soon dominated national broadcasting whereas in the local and regional areas hundreds of radio and television stations operated.

The Communications Act that was implemented in 1934 was the backbone of broadcasting regulations in the USA. The act set up the regime that was in power until the late 1990s, and more or less defined the framework within which the American media and broadcasting industry was to develop. Broadcasters were regulated according to *dominance* in the market by the restrictions on ownership in terms of how many radio and television stations that could be owned by one company or individual; by *entry* regulations that foreclosed the networks from dominating program production and that divided broadcasting and communications into separate markets. As in all other countries, the legislation was amended and adapted to changing circumstances but these principles remained in force until the Communications Act of 1996 replaced them.

The Communications Act of 1996 has been regarded as a major liberalization of the communications market in the USA, and has influenced market developments within the country as well as outside. Basically, the new legislation removed most of the entry barriers between different markets, most notably between the telephone market and broadcasting market, and lifted or relaxed ownership restrictions on radio and television stations. In effect, these changes were aimed at opening and establishing *one* communications market, taking into account that the changes in information technologies had undermined the rationale for implementing different regulations for, for example, broadcasting and telecommunications.

The model of regulating broadcasting by means of the market has been adopted in other countries too, although the specific regulations vary. In Europe, Luxembourg was the exception to the dominance of the public service system. In Latin America, many countries adopted commercial systems, among them Brazil and Mexico.

2.3 Mixed Systems

If the pure public service model and the pure market regulation model are considered two opposite poles, it also becomes evident that there are a number of countries that have adopted hybrid variants or mixed models; that is, some system combining public service models with markets models. Canada and Australia are states that early set up mixed systems, Britain introduced its duopoly consisting of the BBC and the independent broadcasters in 1956, and most other European countries followed suit during the 1960s and 1970s. The Scandinavian countries and Belgium were the last countries to introduce nationwide commercial broadcasters in the early 1990s, the so-called *hybrid channels*. Hybrid channels are private broadcasters granted the privilege of tapping the national advertising market in return for complying with public service obligations concerning programming and coverage. The group of mixed systems is large, diverse, and increasing, reflecting technological as well as political changes in the broadcasting sector.

3. Research into Broadcasting Regulations

3.1 Research on Broadcasting Regulations before 1980

As a field of systematic research, the politics and regulation of broadcasting have developed gradually from the 1970s onwards. Although the broadcasting industry and the broadcasting institutions received much attention from the moment they were established, research into broadcasting regulation and the variations of broadcasting regimes were not studied systematically until the 1980s onwards. There are several explanations for this. The first and most obvious one is that communication studies in many countries were not established as a separate academic field. To the degree that there existed a research tradition on media politics in general and broadcasting regulation in particular, studies tended to be either descriptive country studies accounting for one specific broadcasting system, or they were efforts to typologize regulatory regimes across the world by a set of simple criteria, such as degree of state intervention as opposed to the degree of privatization, or the existence of commercial broadcasters as opposed to noncommercial systems.

Examples of such typologies that gained much attention were the book *Four Theories of the Press* by Siebert et al. (1956) who, writing from a liberal perspective at the peak of the Cold War, divided the world into two main categories of regulatory regimes, liberal with a contemporary variant termed 'social responsible' and authoritarian with the correspondent communist model. In spite of the fact that the model did not encompass variants such as the public broadcasters, the book, probably because of its simplicity, remained a standard reference in central textbooks (e.g., McQuail 1994) for several decades. Williams presented another typology in his book *Communications* (1960). Williams distinguished between authoritarian, commercial, and democratic systems. This was also an effort to set up a universal typology over media regimes, with the exception that Williams argued that fully democratic regimes were nonexistent and needed to be established. Both typologies were prescriptive rather than descriptive and did not in any detail describe the conditions under which different regulatory systems developed.

The relative stability of national regimes between 1950 and 1970 probably contributed to the lack of studies investigating, comparing, and contrasting national broadcasting systems in order to arrive at more precise models. In spite of the fact that television in this period replaced radio as the dominant electronic mass medium in most of the Western world, the introduction of the new medium in general did not change the existing regulatory regimes. In countries that had a national monopoly on broadcasting, television services were opened within the existing institution (e.g., Britain, Scandinavia), in countries that had a commercial system television services were provided by the existing actors (USA), and in authoritarian states television was controlled and used as an instrument for government propaganda in the same way as radio and newspapers.

3.2 Research on Broadcasting Regulations 1980–2000

As a research area, media politics and regulation had its breakthrough in the late 1980s. The increased interest among media researchers for studies of the political and regulatory aspects of the media replaced the preoccupation with power and influence of media institutions, media corporations, and media content that dominated research in political communication in the preceding periods. The 1980s initiated a period of major changes in broadcasting structures and markets all over the world leading to a liberalization of broadcasting markets, entrance of transnational television channels, and a diversification of national broadcasting structures. As a direct consequence, media politics and regulations were moved high on the agenda not only for national policy makers but were also discussed on regional and international forums such as the Council of Europe, the European Union, the World Trade Organization, and UNESCO. These changes, which were conceived of as dramatic by the research community as well as the legislators, spurred a renewed interest into describing and evaluating the processes.

One of the outcomes of renewed interest into media regulation was an increase in the number of *comparative studies* of broadcasting regulation. Whereas research into broadcasting typically had meant country studies, the late 1980s and 1990s saw an upsurge of works that compared and contrasted national systems (e.g., Sepstrup 1989, Siune and Truetzschler 1992, Syvertsen 1992, Humphreys 1996, Hoffmann-Riem 1998, McQuail and Siune 1998). These studies put different systems in relief to each other and reconceptualized the characteristics of national systems into general models, such as the coining of the 'public service broadcasting' model as an analytical concept in the early 1990s.

There exists a number of different attempts to grasp the characteristics of the public service system. The British Broadcasting Research Unit (BRU) assembled a list of eight characteristics, whereas the British political communication researcher Jay Blumler operated with six characteristics. Both were attempts at creating a conceptual tool that could be applied to the different national variants of public service broadcasters. The problem with these lists, however, was the fact that they were too complex and included criteria that applied to some but not all variants of what were conceived of as public service systems.

Another problem was that these criteria did not encompass the subsumption of the newcomers to the public service system—the hybrid channels—under the concept.

An alternative concept of public service broadcasting as a specific regulatory model was set forth by the Norwegian researcher Trine Syvertsen, who, on the basis of a comparison between the British Broadcasting Corporation and the Norwegian Broadcasting Corporation, described public service broadcasting as a reciprocal system between the broadcasting institution and the state. This model is described in terms of a *contract* that grants public broadcasters *privileges* from the state and in return for which the broadcasters take on a set of *obligations* (Syvertsen 1992). The privileges consist, mostly, of being given an exclusive position on the market, such as a monopoly on national terrestrial broadcasting, exclusive access to the license fee, or as in the case of the hybrid channels, exclusive access to the advertising market on national terrestrial broadcasting. The obligations consist, varying between national contexts, of providing universal access to the programs for the entire population, providing a diverse program menu, catering for the interests of majority and minority groups, and strengthening national culture and identities, to mention some of the commonest ideas.

Syvertsen's model allows for the national variations in the public service model. The fact that regulating broadcasting as a public service did not mean that identical models were set up in several countries, rather that each country tried to design a system of publicly governed broadcasting within the frames of external constraints and national characteristics (Skogerbø 1996) is also a feature that has been emphasized during the 1990s (Humphreys 1996, Hoffmann-Riem 1998, Smith and Patterson 1998). The model has also proven productive as it incorporates changes in the political developments of the public service systems.

At the beginning of the twenty-first century there is also an increasing amount of research on globalization of the media industry and the following regionalization and internationalization of regulatory regimes for broadcasting. Globalization can be interpreted as the outcome of a number of different processes, among them the digitization and convergence of different technological domains, such as broadcasting, print and publishing, information technology, and telecommunications. Liberalization of national broadcasting and telecommunications markets that started in the early 1980s contributed to transforming the communications industry into a global industrial sector. These developments have led to an increased focus both on the processes of change, on the causes, and on the effects that they have had on national and regional regulatory regimes.

In Europe, the efforts by the European Union to establish a common regulatory regime for transnational television have been followed critically and discussed extensively. A similar process took place in the Council of Europe culminating in the *European Convention on Transfrontier Television*. These international regulations subsumed a hitherto unregulated emerging industry of transnational satellite and cable television under a common regional regime. Similar provisions are baked into other regional trade agreements (e.g., North American Free Trade Agreement (NAFTA). A theme specific to the European debate on broadcasting regulations, however, was the objectives laid out in the Green Paper on *Television Without Frontiers* that was published by the European Union in 1984 as a precursor to the Directive with the same title that was adopted in 1991. The Green Paper stated that the attempt to harmonize broadcasting regulation across Europe was one of several ways to create a basis for a common European identity and a common European culture. These attempts were debated hotly in the European political community, and have also been an issue for research and debate among European media researchers. The gist of the academic debate was a profound skepticism towards the idea that international regulations could create the basis for a common European identity (e.g., Morley and Robins 1995, Schlesinger 1997). The early 1990s also displayed a conspicuous clash of interests between the USA and the European Union regarding regulation of trade in audiovisual products, of which television programs are a major part, that have attracted interest from scholars as well as political actors. The closing of the Uruguay Round in the negotiations on the General Agreement on Tariffs and Trade (GATT) threatened to break down because of major conflicts of interest between these two parties.

In the late 1990s less attention was given to these issues mostly owing to the fact that European agencies turned their attention to attempts at setting up a regulatory regime that adapted to the emerging *convergence* of infrastructure, services, and markets between broadcasting, telecommunications, publishing, and information technology. The changes foreseen were that the separate regulatory regimes for broadcasting and telecommunications could no longer be maintained, and should be replaced by one common regime for all types of communications.

4. *Methodological Issues*

The study of media politics and regulations is still a young field within communication research. During the 1990s the number of empirical studies comparing different national broadcasting systems has increased markedly. This development also signals that there is a methodological reorientation in the field, emphasizing descriptive and explanatory analyses of a few cases at the expense of normative evaluation of universal models. The comparison of national systems yields

insights into the conditions that have shaped broadcasting institutions and provides insights necessary to understand the current political conflicts on how to regulate broadcasting under new conditions, both nationally and internationally. In media studies there is a long tradition for adopting an implicit, not explicit, normative (critical or liberal) approach when phenomena are analyzed, which sometimes has created problems of analytical clarity. With the increased focus on comparisons, these problems may be overcome, as normative judgments will have to be made more explicit.

Bibliography

Hoffmann-Riem W 1998 *Regulating Media: The Licensing and Supervision of Broadcasting in Six Countries*. Guilford Press, New York

Humphreys P 1996 *Mass Media and Media Policy in Western Europe*. Manchester University Press, Manchester, UK

McQuail D 1994 *Mass Communication Theory: An Introduction*. Sage, London

McQuail D, Siune K (eds.) 1998 *Media Policy: Convergence, Concentration and Commerce*. Sage, London

Morley D, Robins K 1995 *Spaces of Identities: Global Media, Electronic Landscapes and Cultural Boundaries*. Routledge, London

Schlesinger P 1997 From cultural defence to political culture: media, politics and collective identity in the European Union. *Media, Culture and Society* **19**: 369–91

Sepstrup P 1989 *Transnationalization of Television in Western Europe*. Libbey, London

Siebert F S, Peterson T, Schramm W 1956 *Four Theories of the Press*. University of Illinois Press, Urbana, IL

Siune K, Truetzschler W (eds.) 1992 *Dynamics of Media Politics: Broadcast and Electronic Media in Western Europe*

Skogerbø E 1996 External constraints and national resources: Reflections on the Europeanisation of communications policy. *The Nordicom Review* **1**(96): 69–80

Smith A, Patterson R (eds.) 1998 *Television: An International History*. Oxford University Press, Oxford, UK

Syvertsen T 1992 Public television in transition: A comparative and historical analysis of the BBC and the NRK. Thesis. Levende bilder 5/92, Norges allmennvitenskapelige forskningsråd, Oslo

Williams R 1969 *Communications*. Chatto and Windus, London

E. Skogerbø

Buddhism

Of the world's three great missionary religions (Buddhism, Christianity, and Islam), Buddhism is the oldest. It was one of a number of universalist soteriologies which emerged in the Gangetic plain of north India in the sixth and fifth centuries BCE along with cities and kingdoms. It spread throughout Asia between 250 BCE and 1200 CE. Thereafter it began to go into decline in some places, eventually dying out in India, the land of its birth; in other places it went through cycles of decline and revival, culminating in newly invigorated and often innovative revivals in the modern era. In the nineteenth century—partly because of perceived similarities to and contrasts with Christianity—it proved to be powerfully attractive to many Westerners, an attraction that continued and grew in the twentieth century. But there are also underlying differences between Christianity and Buddhism which, when overlooked, lead to serious misunderstandings.

From a sociological point of view Buddhism can be seen as an individualist and humanist soteriology (salvation religion) which developed a variety of communal forms and worldly orientations in different settings. It has legitimated many elaborate hierarchies; in modern times, many Asian intellectuals interpret it as an early form of egalitarian socialism. For many, meditation is the heart of Buddhism (Carrithers 1983); in other places and times elaborate ritual has been, and remains, more important. Some see vegetarianism as central to Buddhism; many Buddhists eat meat. Everywhere Buddhism is centrally concerned with death and salvation; but in some places, notably (but by no means only) in East Asia, Buddhist temples have become deeply involved in providing rituals for worldly protection and winning personal advantages. Most Buddhists take very seriously the notion of rebirth in accordance with the morality of one's actions, but in East Asia the influence of Confucianism means that moral notions are usually divorced from the idea of rebirth.

1. The Buddha, his Teaching (Dharma), and his Community (Sangha)

Buddhist tradition is unanimous that the Buddha was born in Lumbini in southern Nepal and died, aged 80, in Kushinagara, India. The Theravada tradition maintains that he died in 543 BCE, and this date has become widely accepted by Buddhists: in 1956 the 2500th anniversary of the Buddha's death was celebrated in Nepal with delegations from around the world. Western historians of Buddhism have long rejected this date as too early and thus many textbooks give dates between 486 and 460 BCE. Recently it has been argued that the correct date is probably around 400 BCE.

'The Buddha' means 'the awakened one'; the historical Buddha's family name was Gautama and his personal name was Siddhartha. Buddhists believe that in principle anyone can become enlightened and teach others, and that there have been many other Buddhas in different world ages and in different worlds. According to the tradition, the Buddha's father knew from a prophecy that his son might abandon his kingdom to become a great spiritual teacher and

therefore isolated him within the palace, attempting to surround him with nothing but pleasure and ease. One day, however, he managed to escape and saw an old man, a sick man, and a corpse. These 'three signs' alerted him to the existence of suffering and death, and he resolved to seek a way beyond them. He left the palace at night, cut off his own hair, abandoned his clothes, sent away his servant and horse, and set out to seek enlightenment. He spent six years experimenting with different teachers, before coming to see that mortification of the body was as bad as self-indulgence. He realized that a Middle Way was necessary—eating once a day, enough to maintain the body without overindulgence—and this led him to attain Enlightenment in Bodh Gaya, India, at the age of 35. Being enlightened—an aim that he shared with other renouncers of the time—means escape from the otherwise endless cycle of rebirths.

He gave his first sermon, to his first five disciples, in Sarnath, near Banaras in India, on the Four Noble Truths: the existence of suffering, the arising (cause) of suffering, the end of suffering (nirvana), and the path to achieve it. Nirvana is the 'blowing out' of the passion, hatred, and delusion with which we are on fire (and from which we suffer). The metaphor of the fire refers to Brahmanical thinking: the Buddha took over much of the the Brahmans' vocabulary while radically changing its meaning by ethicizing and universalizing it. The path to nirvana consisted of three stages: morality, meditation, and wisdom. In order to practise the second two stages, and to attempt to attain enlightenment in this life, it was (until modern times) usually thought necessary to renounce the world, that is, to abandon, like the Buddha himself, the life of a householder, and live with other celibates, monks or nuns. The role of lay people was to practise morality and to support monks and nuns; by doing so they could improve their karmic state and attain better rebirths in future, eventually going on to attain enlightenment themselves. Later many Theravada Buddhists came to believe that the world had declined to such an extent that even monks and nuns could not attain enlightenment, and they too should seek better rebirths.

At his death the Buddha entered full nirvana and was not reborn. His remains were cremated and the ashes were divided between eight different lay groups and enshrined in *stupa*s. Later the remains of other prominent Buddhist clerics were similarly enshrined and these places, or smaller *caitya*s modeled on them, became the characteristic places of Buddhist worship.

To facilitate a life of self-control and meditation, the Buddha founded his Sangha or Monastic Community. All Buddhists 'take refuge in' the Three Jewels, which are the Buddha, the Dharma (his Teaching), and the Sangha. Members of the Sangha are governed by 227 Vinaya rules (in the Theravada tradition; other traditions have 250 or more). The rules cover the correct behaviour of monks (the nuns have their own Vinaya, with even more rules). This way of life appealed mainly to those of high status: although there were low-caste converts within the Sangha, high castes predominated. It was, as Max Weber (1958 [1916–17], p. 371) remarked, a 'genteel'(*vornehm*) soteriology.

Every fortnight monks are supposed to come together to confess any infractions of the rules and then to recite the rules together. However, in Theravada Buddhism there was originally little enforceable hierarchy. Monks do not promise to obey anyone. Deference is owed on the basis of seniority to monks who were ordained earlier than themselves, and to one's teachers, but that is all. Ecclesiastical hierarchies have emerged in many Buddhist countries, but they are not prescribed in the Vinaya.

Despite the fact that in practice monks frequently maintain close ties with their kin and often carry on secular occupations, the institution of celibacy was to have weighty consequences for the societies which adopted the Buddha's teaching. In Tibet it is estimated that before the 1950s 10–12 percent of the male population were monks, and a further 10–15 percent were non-celibate Buddhist practitioners. In Tibet and the Theravada countries the history, as written in the premodern period, was the history of the Sangha. The issue of monastic property, and its transmission, is important in all Buddhist countries, with a variety of solutions being adopted.

By the standards of his day, the Buddha's teaching was individualistic, humanist, egalitarian, and rationalist. It was individualistic in that individual persons had their own karmic trajectory, which they could work to better by good actions or to harm by bad. There was no question of any higher power or other person granting salvation. It was humanist in that it debunked the Brahmans' claims to higher purity and spiritual status on the basis of their birth; it denied the efficacy of ritual (especially animal sacrifice), ecstatic trance, or worshipping gods in order to obtain salvation. Only one's own efforts will do. It was egalitarian in that all human beings have an equal chance to start out on this path (and humans are superior to gods in this respect: gods, though powerful, have to be reborn as humans to have a chance of achieving nirvana). It was rationalist in that the Buddha favored openness (in contrast to the secretiveness of the Brahmans, whose learning was available only to other Brahmans); and he stressed the need to try out doctrines for oneself and a rational ethic of helpfulness to others.

Thus modernist reformers in the nineteenth and early twentieth centuries could relatively easily represent Buddhism as a rational and scientific belief system, as a philosophy and not a religion, as unconcerned with the existence of God or gods, and as implying social reform and development. However, most Buddhists have believed in gods, have practised ritual in order to overcome misfortune and in order to benefit the dead, have regarded their faith as a religion,

and have accepted socio-religious hierarchies. All the evidence is that these practices are as old as Buddhism itself, and the modernist interpretation of Buddhism should be viewed as just that: as modernist and as an interpretation.

2. *The Spread of Buddhism*

At the beginning Buddhism was just one of many different schools of renunciatory philosophy. With its universalism and its teaching of careful attention, it seems to have appealed to townspeople and merchants, and to have spread with them throughout the Indian subcontinent. It emerged to prominence through the actions of the Emperor Ashoka (reigned c. 268 BCE to 239 BCE) who ruled most of present-day India. Though he protected all religions within his realm, his inscriptions advocate the Buddhist virtues of non-violence, hard work, and consideration for others. He may well have seen Buddhism as a means of binding his enormous Empire together and he sent out Buddhist missionaries in all directions, including to West Asia (Middle East). Buddhist tradition remembers Ashoka as the ideal Buddhist ruler and he became the model for many subsequent rulers attempting to reform the Sangha.

As Buddhism spread, different schools and ordination traditions emerged naturally simply through geographical separation. The Buddha's teachings were preserved orally. This explains their repetitious character: they were only preserved because monks learned to recite them, and preservation of the canon quickly became a prime raison d'être of monastics. Writing probably did not yet exist in the Buddha's day. The earliest compilations of teachings were not systematically written down until the first century BCE, by which time they had long been organized into three 'baskets' (*pitaka*): the Sutta Pitaka, preserving the Buddha's teachings; the Vinaya Pitaka, preserving his rulings on the way of life of the Sangha; and the Abhidhamma Pitaka, which claims to present the teaching literally and systematically, and dates from after Buddhism was already established. (On internal evidence some of the Sutta Pitaka and much of the Vinaya Pitaka must also post-date the Buddha himself.) From time to time monks would come together to chant and check on each other's recitations. The early Buddhist councils were about fixing the Buddhist canon and debating minor differences that had arisen over monastic discipline. As far as the Buddhist laity was concerned, the central teachings were encapsulated in narratives about the Buddha's life, and particularly those concerning his 550 past lives contained in scriptures known as Jatakas and Avadanas; many of the latter were composed much later than the canon.

Among the many early Buddhist traditions, one was the Theravada, meaning 'doctrine of the elders,' so-called because they considered themselves the most conservative. This is now found in South-East Asia and Sri Lanka. Other schools had more adherents in India, especially in the north-west from which China and Tibet received their missions, but Buddhism in what is now Pakistan, Kashmir, and Central Asia died out after the arrival of Islam. Consequently Theravada Buddhism is the only form of pre-Mahayana Buddhism to survive into the modern world.

Although there are some disagreements, the vast mass of Theravada scriptures, preserved in the Pali language, are shared and agreed upon in the Theravada countries. Pali was close to the local vernaculars of the Buddha's environment and the Buddha rejected use of the Brahmans' holy language, Sanskrit, because of his policy of openness. But as Buddhism spread, and as the spoken languages of north India evolved, Pali quickly became a purely scriptural and liturgical language, just like Sanskrit, and lay people became dependent on monks to expound the teaching to them.

2.1 *Mahayana Buddhism*

Around the turn of the common era an entirely new form of Buddhism arose in north-west India called the Mahayana. This is conventionally translated as 'The Great Vehicle' but equally means 'the great path or way.' It introduced a whole raft of new scriptures, now composed in Sanskrit, which were supposed to date from the time of the Buddha himself; some claimed to have been hidden by him until people would be ready for them. The Mahayana reinterpreted the term *bodhisattva*, which for non-Mahayanists essentially means 'a future Buddha.' Mahayana Buddhism posited the existence of numerous powerful, celestial *bodhisattva*s. The most important of these was Avalokiteshvara, the embodiment of compassion, who would appear to people in their hour of need and save them. In China he was thought of as female and became the goddess Kwan Yin (Kannon in Japan). The Mahayana, unlike earlier forms of Buddhism, was most definitely based on a cult of the book: by this time scriptures were holy objects to be enshrined and worshipped as well as read.

The Mahayana introduced a revolution into the moral code of Buddhism, because it encouraged all Buddhists, whether monks or lay people, to aspire themselves to become Buddhas: this is called 'taking the *bodhisattva* vow.' Selfless service to others is part of this path; thus non-Mahayana Buddhists are said to be following the 'Hinayana' or Lesser Way. Great stress is laid on the Buddha's doctrine of 'skill in means' i.e., adapting the teaching to the level of the hearers. This, together with the philosophical teaching

of the essential 'emptiness' of all worldly phenomena, acted as a charter for developing new forms of ritual to attract lay Buddhists and non-Buddhists onto the path. These doctrines also served to legitimate the relaxation of some monastic rules e.g., the ban on monks' handling gold and silver.

Mahayana Buddhism did not reject the pantheon, the scriptures, or the monastic discipline of earlier Buddhism; it simply added further scriptures and objects of worship. Two of its most influential scriptures were the Lotus Sutra (Saddharma-Pundarika) and the Perfection of Wisdom (Prajna-Paramita). It was primarily in this form that Buddhism then spread to China, Tibet, Mongolia, Japan, Korea, and northern Vietnam.

2.2 *Vajrayana* (*Tantric*) *Buddhism*

There was a still later current which is known as Vajrayana or Tantric Buddhism. This is based on scriptures known as Tantras, which appeared in India between 400 CE and 900 CE primarily within Hinduism and Buddhism. (In fact, there were several successive waves of Buddhist Tantras.) These scriptures were esoteric, placed great emphasis on ritual, and used magical and antinomian symbols and practices, which, they claimed, enabled the spiritual seeker to achieve rapid results not possible by other methods under the guidance of a guru. The earlier forms of Tantric Buddhism spread to China and Japan (and are still carried on in Japan by the Tendai and Shingon sects). Later forms of Tantrism, once found from India to China, survive today only as part of Tibetan and Nepalese (Newar) Buddhist traditions. The very latest Tantric Buddhist scripture, the Kalachakra, can be dated to the early eleventh century and, very unusually, contains numerous references to Islam.

The development of the Tantric (Vajrayana) Buddhism encouraged the emergence of Buddhist priests, that is to say, full-time ritual specialists devoted to the religious needs of lay people, who are frequently married and pass on their position to their sons. Such specialists are found in Tibet, Nepal, and Japan. The ideology of Tantric Buddhism attaches paramount importance to the guru. He is frequently included alongside the Three Jewels of Buddhism or identified with the Buddha himself. In order to receive the full Tantric teachings it is necessary to take special initiations (Skt. *diksa*); celibacy is not a prerequisite; in fact, given the sexual nature of much of the symbolism and some of the practices, it may be an impediment. Thus Tantric Buddhism introduced a criterion of spiritual hierarchy that was sometimes in conflict with the monk/lay distinction, and at the very least served to undermine it. The necessity of a guru introduced a clear rationale for hierarchy among clerics. However, Tantric Buddhism, just like Mahayana Buddhism before it, did not reject what had gone before, but rather added to it; Tantric Buddhism therefore became the esoteric heart of a highly complex and layered set of traditions. Different schools and teachers disagreed on what exact place to give it. Monastic practice continued to be important in most forms of Buddhism. Tantrism has often been interpreted as a 'tribal,' peripheral, or 'shamanic' form of religion: however, whatever its origins, within Buddhism Tantrism was a highly sophisticated, learned, and scholastic tradition that attempted to tap some of the power of popular practices, both for salvation and for worldly benefits, while maintaining an elite distance from them.

2.3 *Later Developments*

In East Asia whole sects of Mahayana Buddhism grew up devoted to a particular scripture, a development which did not occur in India or Tibet. Thus, in Japan there are sects whose primary allegiance is to the Lotus Sutra, practising a devotional Buddhism, while others, e.g., the Shingon, have as their core practice the Tantric ritual of fire sacrifice, and yet others (the various Zen sects) focus on meditation. However, the Tendai attempts to teach and combine all the scriptures, while holding the Lotus Sutra to the final and highest truth. In Japan, until the Meiji restoration, most Buddhist practitioners were celibate, except those of the Jodo Shinshu, founded by Shinran (1173–1263); it was one of the most radical sects focused on the worship of the Amida Buddha of the Western paradise (Pure Land): simply uttering obeisance to Amida was all that was required for salvation; all other rules (including celibacy) were beside the point. A strong opponent of Shinran was Nichiren (1222–1282), who advocated devotion to the Lotus Sutra and mounted vituperative attacks on Buddhists who did not agree with him.

In Tibet yet another innovation, dating from the 13th century, was the institution of reincarnate lamas (teachers), of whom the most famous is the Dalai Lama. The first Dalai Lama lived in the sixteenth century; the present Dalai Lama is the fourteenth. When a Dalai Lama dies, after the interbirth period his successor is sought according to the indications of the state oracle. The boy chosen is believed to be both the reincarnation of the previous Dalai Lama and of the *bodhisattva* Avalokiteshvara, the special protector of Tibet. There are many other reincarnate lamas throughout Tibet and in neighboring areas where Tibetan Buddhism is practised.

2.4 *The Contemporary Situation*

Nowadays Theravada Buddhism is practised in Burma, Thailand, Laos, Cambodia, and Sri Lanka. Tibet was and is a stronghold of both Mahayana and

Vajrayana Buddhism, with both married and (more prestigious) celibate practitioners. Chinese Buddhism was predominantly non-Tantric Mahayana of various sorts, as are Korean and Vietnamese Buddhism. In East Asia generally different sects grew up on the basis of different Mahayana scriptures; thus in Japan there are both Tantric (Shingon, Tendai) and non-Tantric Mahayana schools, and even a tiny school (the Ritsu) based entirely on the practice of the monastic code. Thus when new religious movements arose in this century in Japan, naturally some were strongly Buddhist: some, like Soka Gakkai and Reiyukai, have spread through missionary activity to many different countries. The Tibetan diaspora has carried Tibetan Buddhism around the globe as well, and Theravada Buddhists have also proselytized widely. There have also been attempts to build a global ecumenical Buddhist movement.

There is a striking contrast between the Theravada countries plus Tibet, where Buddhism became the dominant cultural tradition and came eventually to define a nationality, and East Asia, where Buddhism, though at periods dominant and favored by the elite, never achieved unquestioned acceptance. In China, Japan, and Korea the influence of Confucianism and Daoism meant that Buddhists often had to defend celibacy against the argument that it undermined society; by contrast the spiritual value of avoiding marriage and reproduction is rarely questioned in South Asia and Buddhist Southeast Asia.

3. *Buddhism and Political Authority*

Although a minority of the first generations of monks chose to wander for nine months of the year, most later monks and nuns settled in monasteries, working at preserving the scriptures and serving the laity. Buddhist monasteries rapidly acquired economic assets, and some were wealthy, owning villages, buying and selling land, and involved in trade. Property brought problems, requiring managers, both lay and monastic (in Theravada Buddhism monks are not supposed to handle money). On the other hand, monks and nuns are obliged to act as 'a field of merit' and accept whatever the laity offer them. The suspicion that many were only interested in a comfortable and secure life, free from labor in the fields, was common.

A king was meant to be the foremost layman of the kingdom and ensure the purity and safety of the Sangha. All ordinations took place under his authority. Following the example of Ashoka, kings were supposed to expel monks who broke the rules or fomented splits within the Sangha. It has been argued by Samuel (1993) that it was the relative strength of the monarchy in Southeast Asia which enabled Theravada to remain dominant and prevented other sects, based on the Mahayana scriptures for example, from becoming established. By contrast, in Tibet the effective political control of Lhasa never spread far beyond the heartland, thus enabling a rich variety of Buddhist philosophies and schools to flourish, including many that gave a much greater emphasis than the dominant Gelukpa sect to shamanic and non-celibate practice.

Buddhist kings were faced with the problem that simply in order to carry out their duty as a king—protecting the people, maintaining order—they had to commit sins (primarily killing) and were therefore disbarred from attaining salvation. At the same time, in both Theravada and Mahayana countries, the king is often regarded as a *bodhisattva* (Buddha-to-be). Ashoka is remembered as having killed many enemies in war in Kalinga, and then, after becoming full of remorse, vowed not to do it again. But other ideal kings either do one or the other, and in Sri Lanka the heroes of the Sinhalese chronicles (composed by monks) are those kings, Dutthagamani (101–77 BCE) and Parakkamabahu I (1153–86), who fought on behalf of Buddhism. The Sangha offered a way out: earning merit to counterbalance their sins though supporting the Sangha. At the same time, one of the most popular and (for Buddhists) most poignant Jataka (Rebirth) stories tells how King Vessantara (the Buddha in a previous life) renounced and gave away, not only his kingdom and wealth, but his wife and children as well.

In short, the Sangha enabled the king to legitimate himself in the eyes of the people, and refused to ordain fugitives from royal justice. In return, the king supported the Sangha, granting it land and other privileges, and when necessary carried out purifications, suppressed noncanonical innovations, and supported revivals. Only in Tibet did the Sangha eventually take over and *become* the state.

4. *Religious Specialists and the Laity*

Lay Buddhists take the Five Precepts: undertakings not to take life, not to steal, not to commit sexual wrongdoing, not to lie, and not to take intoxicants which lead to carelessness and the breaking of the other four precepts. On special days lay people take the Eight Precepts: this includes the Five (while interpreting the third to mean a complete cessation of sexual activity), and adds three others: not to eat after midday, not to adorn oneself or go to entertainments, and not use luxurious beds (i.e., to sleep on the floor). In this way lay people, mostly women, take up the same discipline as a monastic novice, but on a temporary basis.

The Buddhist laity do not have to foreswear all other religious relationships. Providing they approach the gods for mere worldly matters, this is usually of no concern to the monks, and has nothing to do with their faith in Buddhism as a path to salvation. It is hard for Buddhist laity to continue to practise animal sacrifice but even this is done by some Buddhists in Tibet,

Nepal, and Thailand in what they define as non-Buddhist contexts.

The facts that Buddhism is centred so firmly on salvation and the afterlife, and that lay Buddhists almost inevitably have other, worldly interests, mean that 'Buddhism in real life is *accretive*' (Gombrich 1971, p. 49, original emphasis). Buddhism always coexists—often in a structured hierarchy that places the Buddha at the top of the pantheon—with other religious systems which satisfy the worldly needs of ordinary Buddhists. Thus when Theravada Buddhists celebrate the harvest, seek a cure for illness, or get married, they do so in a non-Buddhist idiom with the help of non-Buddhist specialists (Tambiah 1970). Hindu gods are included in the Buddhist pantheon as powerful spirits, subordinate to the Buddha, and travelled with Buddhism to other parts of Asia where they were sometimes identified with local deities, all of whom became part of the Buddhist cosmology.

This soteriological focus explains why, traditionally, there is no such thing as a Buddhist wedding ceremony, although pious lay people may optionally invite a monk or monks to chant and be fed after the ceremony. Even in Mahayana countries the close relationship between Buddhism and death continues. In Japan death ritual is almost exclusively a Buddhist concern, whereas birth and marriage are celebrated in a Shinto idiom. However, many Buddhist temples in East Asia also offer help for worldly needs; and in Nepal, among the Newars, Buddhist priests have even established a complete set of Buddhist life-cycle rituals, including marriage.

This interrelationship between Buddhism and other systems has frequently been mislabelled 'syncretism' by Western observers using an exclusivist definition of religion. But this is an ethnocentric judgement which ignores the fact that Buddhist lay people who worship non-Buddhist spirits or gods do nothing wrong, providing they do not do so in the hope of obtaining salvation. Ironically, however, modernist Buddhists have themselves adopted the Western-derived exclusivist definition of religion in their attempts to reform what they see as backwardness and superstition.

The radical individualism of the Buddha's message was socially and psychologically hard to accept. Most Buddhists, including monks, have wished to make some kind of offering to benefit their dead parents, for example. The ritualized offering of merit to others, which may include dead relatives, is common in all forms of Buddhism. In Theravada Buddhism the canonical explanation of the protective chanting ritual performed for the dead is that the dead can only benefit if they happen to be in the form of ghosts and happen to be nearby so that they can gain merit from listening. That such an abstruse justification might not be shared by all lay people can easily be imagined. In a similar way, many lay people have a straightforwardly magical attitude to the use of Buddhist amulets. Mahayana Buddhist specialists do not insist so unequivocally on maintaining a rationalist, commemorative explanation of Buddhist rituals; but they agree that the magical attitude corresponds to the lowest level of understanding.

5. *Reforms, Revivals, and National Identity*

Buddhism has always been subject to cycles of decline and reform. All forms of institutionalized Buddhism experience a tension between well-endowed institutions, which specialize in scholarship and ritual, on the one hand, and wandering or forest-dwelling monks or practitioners, on the other. Frequently, the laity (including the elite) perceive the freelance practitioners as closer to the Buddhist ideal and beat a path to their door. Over time, and thanks to the donations they cannot refuse, the freelancers themselves become well endowed, and inspire a further cycle of reform aimed at recapturing true asceticism and/or meditative practice.

5.1 *The Rise of Buddhist Modernism*

In the nineteenth and twentieth centuries, faced everywhere with competition from Christian missionaries and taking advantage of the new technology of the printing press, reform movements took a new and sometimes overtly political turn. Buddhist institutions experienced very different trajectories depending on the country. In Japan Buddhism had been closely tied to the state in pre-Meiji times and Buddhist priests had been responsible for government record keeping; after 1868 Buddhism was forcibly separated from Shintoism (with mixed religious sites being forced to choose one or other identity) and much land was expropriated (Ketelaar 1990). In China and Tibet, Buddhism, like all autonomous civil institutions, suffered enormously under Communist rule, being permitted a limited revival in the 1980s and 1990s.

The Theravada countries have experienced a new form of rationalist, scripturalist Buddhism which has been dubbed Buddhist modernism or Protestant Buddhism (Gombrich and Obeyesekere 1988). In Sri Lanka this was was a protest both against the Protestant missionaries and against traditional Theravada Buddhism: it criticized monks for being unlearned ritualists, for being inactive on issues of social and political concern, for not practising meditation, and for encouraging superstition. More positively, it stood for universal literacy, Buddhist Sunday schools, and a middle-class morality of strict monogamy and hard work. It sought to base Buddhist teachings strictly on the Pali canon, and to purify practice of anything that could be dubbed Hindu or Mahayana influence. It stressed the scientific nature of Buddhist teachings, often claiming that major scientific advances were already to be found in the Pali

canon. In Sri Lanka it also aimed to make Buddhists proud to be Buddhist, and to defend their cultural and religious heritage against the inroads of the Protestant missionaries who confidently predicted that in fifty years' time the Buddha would have gone the same way as Wodin and Thor. The rise of Protestant Buddhism went hand in hand, therefore, with the emergence of Sinhala nationalism. The Sinhalese race-nation came to be defined in terms of its Buddhism, and Buddhist historical chronicles were taken to demonstrate its long association with, and right to dominate, the island. The key figure in this Buddhist revival in Sri Lanka was Don David Hewawitarana (1864–1933), who took the title Anagarika Dharmapala, founded the Maha Bodhi Society, and campaigned ceaselessly to awaken the Sinhalese to their religious and nationalist mission.

5.2 Western Views of Buddhism and the Protestant/Catholic Parallel

The story of Buddhist revival in Sri Lanka intersects crucially with the story of the Western discovery of Buddhism. Many of the Victorians had a deep sympathy for Buddhism, because they thought it was, like Christianity, a simple and rational faith founded by a pure-hearted reformer. Edwin Arnold's poem, *The Light of Asia* (1879), had an enormous impact. Counter-attacking Christian apologists dubbed Buddhism 'nihilistic' and 'anti-life.'

There was a pervasive tendency, which has lasted up to the present in some quarters, to view Mahayana and Theravada Buddhisms as the equivalent of Roman Catholicism and Protestantism within Christianity. (Hostile descriptions of Buddhism borrowed a well-worn vocabulary, honed in many Protestant descriptions of Catholicism, in which the words 'priestcraft,' 'superstition,' and 'mumbo-jumbo' loomed large.) It is true that Mahayana Buddhism has a baroque pantheon, numerous elaborate rituals, and more elaborate priestly hierarchies; but the contrast was exaggerated by adopting the modernist interpretation of Theravada Buddhism as a simple, rationalist philosophy whose only ritual was the commemoration of its founder. The parallel breaks down as a serious tool of analysis for several reasons:

(1) historically Theravada Buddhism represents the earlier form, against which Mahayana rebelled;

(2) Theravada Buddhism is traditionally built on monasticism, and therefore on spiritual hierarchy, and it is modernist Buddhism which has introduced a Protestant-style belief in the equality of all believers;

(3) the scriptures of the Theravada were preserved in Pali, and it is only in the modern period that they began to be translated into the vernaculars;

(4) the worship of relics is very important in Theravada Buddhism, and sponsoring their worship has been an important source of legitimacy for rulers, both traditionally and in the modern period;

(5) traditionally Theravada Buddhists believed that nirvana was impossible to attain at present, because of the degeneracy of the age, which made the quest for salvation considerably less urgent than in Protestantism.

These differences between Protestantism and actually existing Theravada Buddhism were obvious to those Westerners, such as Madame Blavatsky and Colonel Olcott, founders of the Theosophical Society, who travelled to India in search of the wisdom of the East. Olcott came to Sri Lanka in 1880 and set about reforming Buddhism, so that it could become what they claimed it had always been. It was Olcott who first set up Buddhist schools, drew up a Buddhist catechism and encouraged the annual celebration of the Buddha's enlightenment with Buddhist carols. He was instrumental in the setting up of Buddhist Sunday schools and the Young Men's and Young Women's Buddhist Associations. He also designed the Buddhist flag which is now flown around the world. It was Olcott and Blavatsky who first encouraged Dharmapala. Many of the early Buddhist revivalists were Theosophists as well. Dharmapala, along with D. T. Suzuki, who first popularized Zen in the West, and Vivekananda, the leader of neo-Hindu revivalism, attended the 1893 World Parliament of Religions in Chicago, a key event in the history of Buddhist modernism.

5.3 Government Control in Thailand

In Thailand, which avoided European colonialism, modernist reforms were, by contrast, largely a top-down process. The reforming King Rama IV (Mongkut) (reigned 1852–68) had spent 27 years a monk before ascending the throne. He emphasized strict adherence to the Pali canon, excluding later, popular works that encouraged the laity to view the Buddha as a miracle worker. It was Mongkut's monastery that was the center of what was to become the Dhammayutika Nikaya, a new monastic order that stressed adherence to the Vinaya rules. Increasing centralization led in Thailand to the Sangha Act of 1902, by which the government took effective control not just of the royal monasteries in the capital but, for the first time, of all monasteries and monks in Thailand. A national exam syllabus was established with textbooks written by Rama IV's son and the king's half-brother, Prince Wachirayan, who himself became the Supreme Patriarch of the Thai Sangha in 1910. In Thailand, but not in Sri Lanka, it is the custom for all young men to spend at least 2–3 months as a monk, around the age of 20. (In Burma temporary ordination tends to be taken much younger, and can last anything from a day to several years.) Giving up the status of monk and returning to lay life is not seen

as in any way shameful; many men spend 10 or more years receiving an education as a monk, and then return to lay life. One consequence of tightening up monastic education was that monasteries, which had previously had a monopoly on learning and literacy, had to stick to a more strictly religious curriculum, and many forms of knowledge—medicine, astrology, history writing—that had been passed on in them, now found a home outside.

5.4 Buddhism and National Identity

As in Sri Lanka, the reform of Buddhism in Thailand was intimately tied to the construction of a sense of national identity. Muslim and Christian minorities have not been persecuted, but Buddhism receives official backing. However, the detailed government control of religion has meant that, when dissident movements arise, it is faced with the difficult choice between suppressing them in line with its traditional responsibility for the purity of Buddhism, and permitting them in the name of religious freedom.

In Burma, Thailand, and Sri Lanka the relationship between the ruling elite, the people, and Buddhism was so close that it is possible to speak of proto-nationalism there even before the modern period; that is to say, the people were defined in terms of their adherence to Buddhism and any ruler who failed to protect Buddhism could not be accepted as legitimate. Certainly, when the modern period dawned and nation-states were created, it was hard to be accepted as a Burmese or a Thai unless one professed Theravada Buddhism, and many of the ethnic troubles of Sri Lanka have flowed from the fact that most Sinhalese Buddhists feel the same way.

There is a similarly tight connection between Tibetan nationalism and Buddhism, but in other Mahayana countries the position is very different. In Japan, Buddhism was persecuted as a foreign religion in the 1870s, and in Korea restrictions were continuous from 1392 until the twentieth century: the ruling Choson dynasty, which based itself on Confucian ideas, banned Buddhist monasteries from the major towns.

5.5 Lay Meditation and the Movement for Female Ordination

The rise of the lay meditation movement marked a major change in Buddhist practice in the twentieth century. Traditionally few monks meditated. Most performed ritual services for the laity and some studied the scriptures. Highest status went to learned monks in urban centres or to rare forest meditators. Lay life was thought to be incompatible with meditation. Part of the modernization of the religion has seen meditation being taught to the laity, often by laymen and women. For many middle-class Buddhists, this means that they bypass monasteries altogether and base their religious practice in meditation centers. Burma led the way in the development of meditation for the laity, and Vipassana (insight) meditation has spread throughout the world; for example, the efforts of an ethnic Indian born in Burma, S. N. Goenka, have been very influential well beyond the community of those born Buddhist. Zen Buddhism has perhaps been even more influential in the West, though whether the popular understanding of it accords with the way it is practised in East Asia is open to question. In India Bhimrao Ambedkar, leader of India's Untouchables and one of the framers of India's Constitution, began a major movement of social liberation by urging his followers to embrace Buddhism and converting himself at the end of his life in 1956 (Zelliot 1992).

Just as the Buddhist laity in general are not as submissive to monastic authority as they once were, so there has been a general movement for the emancipation of women in Buddhism. Historically the order of nuns has always been subordinate to monks, the highest ranking nun being considered, formally, as inferior to the most junior monk. Furthermore, within Theravada countries the nun's ordination tradition died out in the eleventh century (Sri Lanka) and thirteenth century (Burma): since five nuns are needed to ordain a new one, the formal status of nun was no longer available to women. All that they could do was to take the Ten Precepts, like a monastic novice, on a permanent basis, and live a celibate life, a precarious and unprestigious possibility attractive mainly to widows (though often also done by elderly men). As part of the revival of Buddhism in the Theravada countries, many women started to adopt a more monastic way of life and to be treated as fully fledged nuns. In the 1980s and 1990s Western feminist Buddhists have argued that Theravada women should take full ordination from the Chinese tradition where it has been preserved, and some have done so. This has been extremely controversial; many of those (mainly men) who have opposed them doing so argue that they should thereby become Mahayanists or refrain from calling themselves nuns. While the theological objections are very different, the issue arouses similar passions to the ordination of women priests in the West.

See also: Buddhism and Gender; East Asia, Religions of; Religion: Definition and Explanation; Religion: Evolution and Development; Religion, Psychology of; Religion, Sociology of; Southeast Asian Studies: Religion

Bibliography

Almond P C 1988 *The British Discovery of Buddhism*. Cambridge University Press, Cambridge, UK

Bartholomeusz T J 1994 *Women under the Bo Tree: Buddhist Nuns in Sri Lanka*. Cambridge University Press, Cambridge, UK

Bechert H, Gombrich R (eds.) 1984 *The World of Buddhism*. Thames and Hudson, London
Buswell R E Jr. 1992 *Zen Monastic Experience: Buddhist Practice in Contemporary Korea*. Princeton University Press, Princeton, NJ
Carrithers M 1983 *The Forest Monks of Sri Lanka*. Oxford University Press, Delhi, India
Collins S 1982 *Selfless Persons: Imagery and Thought in Theravada Buddhism*. Cambridge University Press, Cambridge, UK
Gellner D N 1992 *Monk, Householder, and Tantric Priest: Newar Buddhism and its Hierarchy of Ritual*. Cambridge University Press, Cambridge, UK
Gombrich R F 1971 *Precept and Practice: Traditional Buddhism in the Rural Highlands of Ceylon*. Clarendon Press, Oxford, UK [1991 reissue as *Buddhist Precept and Practice*. Motilal Banarsidass, Delhi, India]
Gombrich R F 1988 *Theravada Buddhism: A Social History from Ancient Benares to Modern Colombo*. Routledge, London
Gombrich R F, G Obeyesekere 1988 *Buddhism Transformed: Religious Change in Sri Lanka*. Princeton University Press, Princeton, NJ
Goodwin J R 1994 *Alms and Vagabonds: Buddhist Temples and Popular Patronage in Medieval Japan*. University of Hawaii Press, Honolulu, HI
Gunawardana R A L H 1979 *Robe and Plough: Monasticism and Economic Interest in Early Medieval Sri Lanka*. The University of Arizona Press, Tucson, AZ
Hardacre H 1984 *Lay Buddhism in Contemporary Japan: Reyukai Kyodan*. Princeton University Press, Princeton, NJ
Ishii Y 1986 *Sangha, State, and Society: Thai Buddhism in History* tr. P Hawkes. University of Hawaii Press, Honolulu, Hawaii
Ketelaar J E 1990 *Of Heretics and Martyrs in Meiji Japan: Buddhism and its Persecution*. Princeton University Press, Princeton, NJ
Lamotte E 1988 *History of Indian Buddhism: From the Origins to the Saka Era* tr. S. Webb-Boin. Oriental Institute. Université Catholique de Louvain, Louvain-la-Neuve, Belgium
Lopez D S Jr. 1999 *Prisoners of Shangri-La: Tibetan Buddhism and the West*. The University of Chicago Press, Chicago
Malalgoda K 1976 *Buddhism in Sinhalese Society, 1750–1900*. University of California Press, Berkeley, CA
Mumford S R 1989 *Himalayan Dialogue: Tibetan Lamas and Gurung Shamans in Nepal*. University of Wisconsin Press, Madison, WI
Ortner S B 1989 *High Religion: A Cultural and Political History of Sherpa Buddhism*. Princeton University Press, Princeton, NJ
Samuel G 1993 *Civilized Shamans: Buddhism in Tibetan Societies*. Smithsonian Institution Press, Washington, DC
Schopen G 1997 *Bones, Stones, and Buddhist Monks: Collected Papers on the Archaeology, Epigraphy, and Texts of Monastic Buddhism in India*. University of Hawaii Press, Honolulu, HI
Silber I F 1995 *Virtuosity, Charisma and Social Order: A Comparative Sociological Study of Monasticism in Theravada Buddhism and Medieval Catholicism*. Cambridge University Press, Cambridge, UK
Spiro M E 1982 [1970] *Buddhism and Society: A Great Tradition and its Burmese Vicissitudes*. 2nd expanded edn. University of California Press, Berkeley, CA
Tambiah S J 1970 *Buddhism and the Spirit Cults in North-East Thailand*. Cambridge University Press, Cambridge, UK
Tambiah S J 1976 *World Conqueror and World Renouncer: A Study of Buddhism and Polity in Thailand against a Historical Background*. Cambridge University Press, Cambridge, UK
Tambiah S J 1984 *The Buddhist Saints of the Forest and the Cult of Amulets*. Cambridge University Press, Cambridge, UK
Tambiah S J 1992 *Buddhism Betrayed? Religion, Politics and Violence in Sri Lanka*. University of Chicago Press, Chicago
Weber M 1958 [1916–17] *The Religion of India: The Sociology of Hinduism and Buddhism*, tr. and eds. H H Gerth and D Martindale. The Free Press, Glencoe, IL
Welch H 1967 *The Practice of Chinese Buddhism: 1900–1950*. Harvard University Press, Cambridge, MA
Zelliot E 1992 From Untouchable to Dalit: Essays on the Ambedkar Movement. Manohar, Delhi, India

D. N. Gellner

Buddhism and Gender

Studies on Buddhism and gender can be divided into two categories. First, scholars are beginning to look into historical Buddhism, both to ascertain the history of women in Buddhism and to find out what classical Buddhist texts say about gender. Second, some contemporary Buddhists are re-evaluating their tradition and suggesting changes that would make their religion more equitable both in terms of gender and in terms of sexual orientation.

As is the case for all major religions, scholars who want to explore topics related to Buddhism and gender inherited androcentric scholarship about a male-dominated religion. There is no question that Buddhism was historically male-dominated, and remains so in most of its contemporary forms. And there is no question that until recently, scholars who studied Buddhism studied Buddhist men and ignored Buddhist women, for the most part. Buddhist studies and Buddhist constructive thought are both relatively conservative enterprises; therefore gender studies in Buddhism are less developed than is the case for other religions, especially Christianity.

Because the conceptual framework for gender studies evolved in the context of Western thought, care must be taken in applying that conceptual framework to Buddhism. It is easy to misinterpret or misrepresent Buddhist ideas and practices about gender by assuming that they mean the same things in a Buddhist context as they would in a Western context. For example, the popular assumption that it is 'bad karma' to be a woman can easily be taken to mean that women are 'bad' or evil, an assumption common in some forms of Western religions. But in a Buddhist context, the meaning of this assumption is that women are less fortunate than men precisely because of limitations placed on them by male-dominated cultures—an assessment with which Western feminists would agree!

1. Traditional Buddhist Understandings of Gender: Historical Perspectives

Despite the varied, often contradictory statements about gender found in classical Buddhist texts and throughout Buddhist history, several generalizations are tenable. There have been two major themes and one minor theme that have dominated Buddhist teachings and practices surrounding gender in all periods of Buddhist history, in all schools of Buddhism, and in all Buddhist cultures.

The first theme is that gender is irrelevant, 'something about which weak worldlings are confused,' to paraphrase a common scriptural theme. This strand of Buddhist thought is found in the textual traditions of all forms of Buddhism, though it is strongest in Indian Mahayana texts, which began to be written about five hundred years after the historical Buddha, and in Indian and Tibetan Vajrayana texts, which date from several centuries later. Though there are many variants of this theme, all of them proclaim that a person's gender is irrelevant to their understanding of Buddhist teachings and ability to attain enlightenment.

The second theme contradicts the first, declaring that gender matters a great deal and that it is far more fortunate to be reborn as a man than as a woman. The texts that present this point of view very clearly recognize that women's lives were far more circumscribed and difficult than men's lives in the cultural settings with which they were familiar. Cultural limitations on women's ability to study Buddhist teachings and practice Buddhist meditation were only the most onerous of their burdens. In addition, women were required always to be formally subordinate to male authority, even if they were nuns, and most women faced heavy domestic and childbearing responsibilities. The men who evaluated this situation came to the conclusion that women were, indeed, quite unfortunate. They explained this misfortune by means of common teachings about rebirth, according to which one's present life, filled with positive or negative circumstances, resulted from actions (*karma*) done in past lives. Thus, people who commit negative acts are reborn as women rather than men, with all the difficulties outlined above. Some have even claimed that there is no gender inequity in classical Buddhism because it teaches that 'deserving' women are reborn as men in future lives. Indeed, even in contemporary times, Buddhist women in some cultures engage in religious practices that would help them be reborn as men in the future.

The third, more minor theme in Buddhist literature is a strand of misogyny, or fear and hatred of women, that runs through Buddhist texts. This theme is especially strong in the writings of male monastics and in those forms of Buddhism that most valued celibacy. Thus it is no surprise that this theme is strong in the earliest Buddhist literature which describes the life of the Buddha and his immediate disciples. Significantly, some forms of Tibetan Vajrayana Buddhism specifically forbid their followers from disparaging women. This command clearly indicates the prevalence of antiwomen rhetoric in Buddhist contexts, but it is also noteworthy that some Buddhists authorities condemned that practice by prohibiting their followers from indulging in such prejudice.

How can these contradictory positions on gender be reconciled, if at all? Most knowledgeable Buddhists would conclude that the first view is the most normative for Buddhists. It is clearly in accord with the major teachings of Buddhism, whereas the second and third views actually contradict the most basic teachings of Buddhism. However, the second view is far more prevalent and popular among ordinary followers of Buddhism. Far more problematic is that all Buddhists institutions, especially the highly valued monasteries and universities, were organized in accord with the second, rather than the first view. Sex segregation was almost complete and men received far more of the education and livelihood provided by monasteries and universities than did women. So complete was the favoring of men's institutions over women's that the once flourishing nuns' order died out in all parts of the Buddhist world except China and Korea. In the other parts of the Buddhist world, only a much less prestigious form of lay ordination for women survived.

2. Key Buddhist Doctrines and Gender

Given this historical legacy, contemporary Buddhist analysts concerned with gender have studied the classical doctrines of Buddhist tradition, asking what implications they might have for contemporary gender issues, especially gender inequity and controversies surrounding sexual orientation. These contemporary analysts are especially concerned with finding firm doctrinal foundations for suggesting changes in traditional Buddhist practices that favor men over women and assume heterosexuality.

All forms of Buddhism trace human misery to grasping or clinging, and state that the most basic and pain-producing form of grasping is clinging to belief in a fixed and permanent underlying core of selfhood—what is called 'soul' in many religions. This is the famous doctrine of *anatman*, frequently translated as egolessness or 'no-self,' which is often the most confusing Buddhist idea to insiders and outsiders alike. Nevertheless, traditional Buddhist doctrine is firm and clear. There is no permanent, independent, real self that can be found by any means at all; the practice of positing such a self and acting as if there were such a self is the ultimate cause of all human suffering.

Thus, according to Buddhism, there is not even a permanent enduring sexless soul attached to any individual's existence, let alone any permanent, real, or lasting identity based on gender. Thus fixed norms

regarding gender roles or stereotypes based on gender are completely contradictory with the most basic Buddhist doctrines, which means that tenacious, emotionally intense clinging to those gender norms and stereotypes is even more inappropriate. Furthermore, the suffering experienced both by those who do not conform to gender norms, as well as those who do, is to be expected if clinging to rigid and fixed norms about gender persists. All that classical Buddhist doctrines would seem to warrant is a recognition that someone is male or female biologically, but that no inherent or necessary implications about appropriate behavior and psychology can be attached to that biology.

This doctrine of *anatman* is common to all forms of Buddhism. Mahayana Buddhism, which began to develop about 500 years after the Buddha, includes doctrines that reinforce the implications of early Buddhist doctrines and texts that specifically state that gender is irrelevant to one's ability to attain enlightenment—Buddhism's highest goal. Briefly, the central Mahayana doctrine of emptiness (*shunyata*) states that not only does a fixed permanent self not exist, but that nothing whatsoever has any permanent, fixed, or independent and real existence. Therefore, one should cling to nothing whatsoever; texts about emptiness state explicitly that gender does not inherently exist and is, therefore, irrelevant to the religious life. Some schools of Mahayana Buddhism also propose the doctrine of Buddha-nature, which claims that everything is potentially enlightened Buddha. All that needs to occur is for beings to realize this indwelling Buddha-nature. Schools of Buddhism that hold this doctrine commonly assert that all sentient beings, including all women and men, are equally endowed with Buddha-nature. Thus, these two Mahayana doctrines, when combined, assert that gender traits do not exist inherently and the men and women are equally endowed with the only trait that matters—indwelling Buddhahood.

3. *Contemporary Issues and Suggestions*

Contemporary analysts of Buddhism and gender have concluded that when Buddhist institutional practice regarding gender and Buddhist doctrines pertinent to gender are placed side by side, a massive contradiction is obvious. Thus it is no surprise that objections to practices of gender inequity in traditional Buddhism have come from many quarters, both Asian and Western, in the late twentieth century. Important factors in this development have been the rapidity with which Buddhism is becoming a religion of choice for Western intellectuals and the rise of Western feminism, which has resulted in a vigorous Buddhist feminist movement.

For many, the most obvious issue is the status of Buddhist nuns throughout the Buddhist world. For traditional Buddhism, monasticism is the preferred lifestyle and the prestige of monastics can hardly be exaggerated. Thus, a reliable marker of gender equity, at least in more traditional forms of Buddhism, is the presence of a flourishing community of well-educated nuns. In contemporary Taiwan and Korea, this is the case, but the situation is not so sanguine in other parts of the Buddhist world. Only the novice ordination has survived in Tibetan Buddhism and in Therevada Buddhism, even that has been lost. All that remains is an informal lay ordination for women, who then wear distinctive robes, but not of the ochre color that is the time-sanctioned mark of a Buddhist monastic. In addition, these novice or lay nuns are rarely well educated in classical Buddhist disciplines and often become mainly servants of the monastery.

How to reinstitute the nuns' ordination has been discussed by many; the mechanics of doing so are not difficult and ordinations have been offered by those Buddhist communities that still retain the ability to ordain nuns. However, in many cases, changes have not yet occurred while male monastic authorities investigate sectarian differences and the legitimacy of the ordination lineages being offered. In the meanwhile, especially among Tibetan Buddhists in exile, the status of novice nuns is improving and they are being taught disciplines not usually taught to nuns in the past. In Theravada countries, there are vigorous pro-ordination movements and some Theravadin women go abroad to receive monastic ordination. However, simply reviving nuns' ordination will not address all the issues brought up by contemporary analyses of Buddhism and gender. The monastic rulebook subordinates all nuns to all monks, with no regard for seniority or accomplishments, thus reduplicating the gender hierarchy of the societies in which Buddhism developed.

In Western countries, where Buddhism is rapidly spreading, monastic issues are less central to the daily concerns of most Buddhists. They have not become, and do not expect to become, monks and nuns, and whether monasticism will become part of Western Buddhisms remains an open question. Following a pattern more common in Japanese Buddhisms than in other Asian Buddhisms, even the most serious Western Buddhist practitioners remain laypersons, usually attempting to combine the demanding meditation and study programs of serious Buddhist practice with careers and family life. What is distinctive about Western Buddhism is that women are just as enthusiastic in such pursuits as are men. In many meditation programs, the number of women participants is as high as, or even higher than, the number of men participants, and, at least in some Buddhist denominations, the teacher is as likely to be a woman as a man. This is a radical departure from traditional Asian practices. If this pattern of lay Buddhists being seriously involved in Buddhist study and practice persists and becomes the norm for Western Buddhism, it will represent a significant paradigm shift in the 2500

year history of Buddhism, not only because of the role women are playing, but also because of the role laypeople are playing.

For both Western and Asian Buddhism, perhaps the key issue determining whether or not gender equity becomes the Buddhist norm is the presence or absence of women teachers at the highest ranks. Buddhism is more dependent on the teacher–student relationship than are most religions and in this nontheistic religion, no authority is higher than the heads of teaching lineages. In the past, such authorities have, with very few exceptions, been men, which has probably contributed more to the persistent gender inequity of Buddhism than any other single factor. As was discussed above, in both Asian and Western contexts, large numbers of Buddhist women are becoming better educated than in the past. In many forms of Western Buddhism, women are common among the senior teachers at the penultimate ranks, but very few women have been authorized to teach at the highest level, especially among those who practice Tibetan forms of Buddhism. American Zen Buddhists now authorize women as independent teachers somewhat more frequently and there are a significant number of women teachers in the Vipassana movement as well.

The presence of women teachers is important for several reasons beyond the requirement for equal opportunity. First is the importance of role models and the subtle messages communicated to newcomers by the presence or absence of members of various groups in leadership positions. More important is the fact that Buddhist teachers have an enormous impact on the development of the tradition. Like all religions, Buddhism is open-ended and unfinished, and the evolution of Buddhism largely depends upon the teachers of any given generation. No matter how realized and accomplished a woman may be as a meditator or philosopher, if she does not teach, she will have no impact on the development of Buddhist thought and practice. The lack of women teachers in the past explains why the insights of women's culture have not been integrated into Buddhist thought previously. They will not become part of Buddhist tradition until large numbers of women are authorized as teachers and speak in their own voices.

In fact, it could be argued that the androcentric character of inherited Buddhist thought and institutions is largely due to such an absence of women functioning as teachers. All periods of Buddhist history do record some highly accomplished women and probably there were many more, memories of whom were lost due to androcentric record-keeping practices. But we rarely encounter stories of them as teachers either in the monastic or the yogic traditions and few texts have been attributed to them.

The intersection of Buddhism, especially Western Buddhism, with the current attention to issues of gender promises to rectify this situation. Especially beginning with the decade of the 1990s, many thoughtful and innovative books and articles have been written on the topic of Buddhism and gender. Many of the authors of these texts openly identify as practicing Buddhists seeking to make their tradition more equitable. In this emerging body of literature, women who have begun to voice their concerns as Buddhist teachers focus on many issues relevant to the lifestyles of lay women practitioners, such as shared domestic responsibilities and arrangements for children at meditation programs. They have explored the importance of community (*sangha*) in Buddhist life and discussed alternative forms of community. They also have searched the historical records looking for accomplished women of the past whose stories had been lost to Buddhist collective memory. A matter of critical importance for many has been translating traditional Buddhist liturgies into gender-neutral and gender-inclusive language. Finally, some have contemplated the visual imagery common in Buddhist settings and sought to rectify the overwhelming preponderance of male images in the typical Buddhist meditation hall or sacred site.

More recently, the topic of sexual orientation has also begun to be discussed publicly in Buddhist circles. Because Buddhism has always required celibacy for its monastics and has not devised strict and detailed law codes governing the everyday behavior of its laity, the topic has not received significant formal attention in the past. The codes of behavior for monks and nuns would forbid homosexual activity simply through the precise and detailed descriptions of forbidden sexual activities. It is difficult to obtain much information historically about homosexual practices of lay Buddhists. But in contemporary times, especially among Western Buddhists, concern that Buddhism recognize the legitimacy of homosexuality has become an important issue. Because Western Buddhism is usually socially liberal, Western Buddhists are generally accepting of their gay and lesbian peers, and a number of important spokespersons for Western Buddhism are gay or lesbian. However, on the basis of relatively obscure traditional norms about forbidden sexual practices, some Asian teachers have objected to the practice of homosexuality, though not to homosexual sexual orientation. To date, less has been written on the topic of gender as an issue of sexual orientation than on the topic of gender as an issue of equity between men and women. With the growing realization that gender is an omnipresent facet of all peoples' experience, not something only women possess, it will no longer be possible to limit discussions of Buddhism and gender to issues of women's place in Buddhism. Such a development will signal a major conceptual breakthrough, for nothing is more indicative of androcentrism than the tendency to collapse the categories of 'gender' and 'women.'

See also: Buddhism; Catholicism and Gender; Family and Gender; Feminist Theology; Goddess Worship

(Old and New): Cultural Concerns; Islam and Gender; Judaism; Judaism and Gender; Protestantism and Gender; Religion and Gender; Religion: Family and Kinship; Religion: Mobilization and Power; Religion: Morality and Social Control; Women's Religiosity

Bibliography

Allione T 1984 *Women of Wisdom*. Routledge and Kegan Paul, London

Cabezon J I (ed.) 1992 *Buddhism, Sexuality, and Gender*. State University of New York Press, Albany, NY

Dresser M (ed.) 1996 *Buddhist Women on the Edge: Contemporary Perspectives from the Western Frontier*. North Atlantic Books, Berkeley, CA

Friedman L, Moon S (eds.) 1997 *Being Bodies: Buddhist Women on the Paradox of Embodiment*. 1st edn. Shambhala, Boston

Gross R M 1993 *Buddhism After Patriarchy: A Feminist History, Analysis, and Reconstruction of Buddhism*. State University of New York Press, Albany NY

Gross R M 1998 *Soaring and Settling: Buddhist Perspectives on Contemporary Social and Religious Issues*. Continuum, New York, NY

Horner I B 1930 *Women Under Primitive Buddhism*. Routledge, London

Klein A C 1995 *Meeting the Great Bliss Queen: Buddhists, Feminists, and the Art of the Self*. Beacon Press, Boston

Layland W 1998 *Queer Dharma: Voices of Gay Liberation*. Gay Sunshine Press, San Francisco

Paul D Y 1979 *Women in Buddhism: Images of the Feminine in Mahayana Tradition*. Asian Humanities Press, Berkeley, CA

Rhys-Davids C A F, Norman K R (trans.) 1989 *Poems of the Early Buddhist Nuns: (Therigatha)*. Pali Text Society, Oxford, UK

Shaw M 1994 *Passionate Enlightenment: Women in Tantric Buddhism*. Princeton University Press, Princeton, NJ

Tsomo K L 1988 *Sakyadhita: Daughters of the Buddha*. Snow Lion Press, Ithica, NY

Tsomo K L (ed.) 1999 *Buddhist Women Across Cultures: Realizations*. State University of New York Press, Albany, NY

Willis J D (ed.) 1987 *Feminine Ground: Essays on Women and Tibet*. 1st edn. Snow Lion, Ithica, NY

R. M. Gross

Budgeting and Anticipatory Management

All organizations must:

(a) coordinate the way resources will be used by the various (often specialized) decision makers; and

(b) learn from the observed deviations between intended and actual.

They need these two activities in order to create value for their stockholders by increasing the probability of satisfying their key stakeholders such as customers, employees, suppliers, and any outside agency granting the organization the right to continue to operate.

For the past 75 years or more, budgets have been the prime instrument in providing such a process of *ex ante* coordination and *ex post* analysis. They did so by defining, often from 'above' (the leadership sends a 'command'), the actions that are likely to lead to goal achievement, combined with *ex post* evaluation of variances or deviations from plan ('control' to bring conformance).

A budget is generally defined as a formal document that quantifies an organization's plan for achieving its goal (Jiambalvo 2001). A budget, in other words, is a description of the financial implications of a sequence of coordinated actions and specialized targets that will allow an organization to achieve its objectives in a changing environment. Most budgets, until recently, assumed implicitly, once the budget had been defined—i.e., the actions and targets selected—that uncertainty about the environment could be essentially 'ignored' for a period of implementation (and the managers were expected to manage to the budget and not necessarily to pay attention to the competitive environment). Frequently, such a period used to be about 12 months and has recently frequently been reduced to six and, in certain cases, three or four months, given the turbulence of the competitive environment. However, many behavioral habits about managing to the budget have been kept alive, and many still try to 'achieve the budget,' thus still assuming the estimate made some 6 to 18 months before is still valid. A budget is in itself a fundamental document and budgeting an essential process in the operation of any organization.

How the budget has been used in practice has led to many abuses and negative effects. This short article will define the principles of budgeting and 'budgets' before showing the limitations of current practice and will explore how the 'concept' (i.e., a tool to support ongoing adaptive anticipation or anticipatory management) ought to be used today to meet current economic and competitive challenges.

1. A Philosophy of Anticipation

'As for the future, your task is not to forsee it, but to enable it' (Antoine de Saint Exupery in *Le Petit Prince*) and, to do so means to be able to mobilize all resources in the organization continuously to adapt to and exploit new opportunities that appear. This means the path to 'success' cannot be completely defined before hand.

Despite such an observation, 'command and control' are still the most commonly used form of management, essentially in contradiction to the reactivity and agility required for success in a competitive world. It seems odd that command-and-control-type budgets still occupy such an important

role in the ways of coordinating the behavior of decision makers. The whole organization, and especially its management team collectively, must continuously update the best way to create value for stakeholders in continuously changing competitive environments.

Budgets and the budgeting activity (process of establishing and validating action plans) are processes and procedures that aim at contributing to the creation of value. A budget is a statement—a declaration—of actions intended to be carried out by several organizational decision makers in a coordinated way. Action plans are needed in both 'for profit' and 'not for profit' organizations. A budget reflects the answers the organization and its leadership give at one point in time (and it may possibly be a valid answer only for a short period of time) to one key question: how are we going to go from where we are to where we want to be? A budget and budgeting process are one of the key ways through which energies are mobilized and channeled for the successful implementation of the business strategic intent.

We will first identify the purposes and the process of creation of the budget, then the key requirements for the budget to be an effective tool that fulfills its promises, before elaborating on the criticisms that have been formulated about budgets and their current use.

2. *Purposes of Budgets and of Budgeting Processes*

The budgeting process and budgets contribute to the process of anticipation in that they:

(a) Provide a forum in which the various organizational members (who do not have identical skills, experience, and expertise) can identify and share their views of opportunities and threats to the value creation process of the business or organization.

(b) Provide a locus for the construction of a shared business model (identify the key success factors and how they interact in creating the success of the firm).

(c) Allow management to discuss, select, and communicate values (to guide behavior), intent (defining what value creation or success means), and plans (coordinated sequence of interlocked actions and identification of resources created and consumed).

(d) Create a forum for coordinated decision making between actors that may otherwise act in a disconnected fashion that would result from their referring to different business models. For example, the sales manager might consider developing sales at a rate that might not be compatible with the action plan imagined by the manufacturing manager (who foresees difficulties in bringing the required new technology on line) or with the action plan of the financial manager (who would rather increase the dividends than finance the increase in working capital that the increased sales might create).

(e) Provide for the development of transfunctional action plans describing how to achieve continuous alignment of the firm or organization 'behavior' with competitive pressures.

(f) Identify what resources are needed at what time. Resources ought not be limited to tangible or financially measurable resources: know-how, intellectual capital, customer, supplier or employee loyalty, morale, image and style, etc. are resources that must be used appropriately to create and maintain long term success.

(g) Allow for the identification and allocation of the business' scarce resources required to carry out the action plan effectively and efficiently.

(h) Reinforce a structure of delegation, devolution and, 'subsidiarity' in co-locating 'rights to make decisions' (on the use of resources) and knowledge and in supporting initiative and autonomy while maintaining coordination. The term subsidiarity is not in the dictionaries we have consulted. It is, however, a commonly used term in the context of the partition of rights between States and the leadership—executive and legislative—of the European Union (EU). We use it here in the same logic to refer to the fundamental idea behind the word as it is used in the EU dialogues: decision making should be as close as possible to the 'customer' or to the market, i.e., any unnecessary attempt at going up in the delegation chain to reach a decision is slowing down the organization's reaction and runs the risk of creating a solution which is maladapted to the need.

(i) Help build commitment by individuals or groups of individuals to achieving time-based results that move the organization in the desired direction. Note that the principle is to obtain commitment to a result and not necessarily to the way through which this result is to be achieved. This aspect is the source of great misuse of budgets in that many organizations confuse 'commitment to the use of resources as agreed upon' for 'commitment to results.' Often the source of confusion is due to a combination of lack of trust on the part of the supervisor and lack of self-confidence on the part of the subordinate.

(j) Contribute to establishing benchmarks for performance measurement and evaluation (for both business units and individuals). Here again it is important to draw attention to the fact that the benchmarks need not be limited to financial measures of results or resource consumption.

(k) Provide a basis for learning from deviations. This aspect has often been misunderstood to mean that conformance to the plan is expected. Since, by definition, no plan can be perfect, conformance for its own sake can be a source of non-competitiveness.

(l) Create a context for the deliberate management of performance improvement.

(m) Allow for the forecasting of cash flow needs on the basis of the action plans that have been retained (cash is an essential scarce resource in any organization and must be monitored at all times).

(n) Offer a process for an ongoing estimation of year-end financial results and more specifically EPS & ROE for commercial organizations.

This whole array of purposes cannot be broken down and each aspect dealt with separately. They are all to be pursued simultaneously.

3. *The Budgeting Cycle*

Generally, the budgeting process follows a pre-established sequence but many iterations are possible with today's simulation instruments:

(a) Identify and communicate estimated future demand and markets conditions and business environment evolution to establish for every participant the hypotheses that all should assume in researching their contribution to the action plans. Such a task should, most of the time, originate with upper management that holds a bird's eye view of the whole business.

(b) Jointly identify threats, opportunities, weaknesses, and strengths of the organization in the perceived (or declared) future environment.

(c) Identify and evaluate the various possible action plans (sales, production, sourcing, human resources, innovation and investments, R&D, etc.) and verify their compatibility with both intent or goals and resources.

(d) Repeat the process until compatibility is obtained and acceptance by everyone in the organization has been secured (this step may require the superior to act as a binding arbitrator to avoid subordinates wasting time in useless battles). An arbitrator listens to all parties, reflects on the best collective solution, and selects a preferred course of action that is both explained and justified to all parties who must adhere to it.

(e) Once the budget or action plan is established, results are matched against expected (intermediate or final) results and the analysis of differences (or deviations) leads to learning and thus a reevaluation of the appropriateness of the action plan and its possible updating.

4. *Components of an Action Plan*

An action plan reflects the answers of the concerned persons to at least the following questions:

(a) Where are we and why do we need to do 'something' (understand the cause of the action plan by identifying a possible gap between where we are and where we are headed, and where we think we ought to be, given the intent or goals of the firm or organization)?

(b) What is the 'intention' behind doing something?

(c) How does whatever will be done relate to the achievement of the firm's mission?

(d) Who are the various actors who need to be involved (basis for both the richness of analysis and delegation in the implementation process)? An action plan is always teamwork, even though each person on the team is personally responsible for one 'component' of the plan, often overlapping with someone else's. For example, an action plan aimed at 'generating orders that the firm can deliver to the customer's satisfaction and that help the firm grow' will require, for example, the cooperation of such specialized managers as: sales manager, sales people, marketing manager, production scheduling, assembly manager, legal department, sourcing manager, management accountant, etc. If the plan were not the result of an agreement between all these required skill holders, it would have little chance of success.

(e) What causal model (also known as business model) do the 'team members' agree on to help them identify the 'action variables' they will activate through their decisions and behavior?

(f) What possible actions can be envisioned? (it is important to review several possible alternatives critically).

(g) Which one will we choose, and why? What will we actually do in this action plan?

(h) How will we do it? What process? What resources will be consumed and created and when? What delegation structure will be developed, etc.?

(i) Are the required and generated resources coherent with the availability and needs of the organization? If the answer is negative, there must be a return for adjustment to any or all questions (a) through (h).

(j) Why will we do it this way? What were the alternatives' pros and cons? How realistic is the alternative selected and how realistic were the other ones? What are the risks attached to each alternative and especially the one selected?

(k) Coherence with strategic intent, strategy, and other action plans.

(l) Expected impact of the selected actions (several may apply simultaneously):

(i) development of a competitive advantage;

(ii) ambitious yet realistic: placing the organization under tension;

(iii) challenging the ways of thinking; and

(iv) all of the above or other.

(m) When will these actions be carried out (calendar of actions)?

(n) How often will these action need to be carried out (if it is a recurring action plan)? This is an important point as a multistage action plan allows for better learning and continuous adjustment.

(o) To what do each of the actors actually commit (intermediate or final results, non-financial indicators, relationships, dates, intensity, etc.)?

(p) How will the achievement of the commitments

be measured? There are at least three domains for which measures should be agreed upon beforehand by all parties involved in the action plan. These measures will serve as a basis for learning from deviations and for continuously updating the action plan so that it remains aligned with the intention selected, if it is still the right one):

(i) Process, i.e., the way the action plan should and does unfold during implementation (the equivalent aerospace term would be 'trajectory'):

evolution of the conditions that created the need for the action;

resource productivity and effectiveness;

resources consumed and created;

organization and relationships between actors;

mastery over risks identified in step (j);

efficiency of use of resources; etc.

(ii) Progress (against the time scale selected by the actors in the plan to state by which date the target will be achieved):

timing: date of start, selection of milestones, date of expected completion;

positioning of progress on a PERT diagram;

validation of the critical path;

definition of the method of calculation of the percentage achievement; etc.

(iii) Impact:

effect of actions on objectives and key performance indicators (KPIs) for both short and long term; and

evaluation of the continued relevance of the intent.

(q) What priority should be given to this action plan and how much of the firm will be involved (if applicable)?

Traditional budgeting seems to have focused essentially on steps (h) and (i) and therefore has led to a very partial view of what is meant by an action plan. Essential to the effectiveness of the budgeting process is the measurement system evoked in step (p). The measures should be balanced between areas of expertise, perspectives, and time horizons and allow simultaneously both effective implementation and continuous validation of the relevance of the whole action plan given the evolution of the competitive environment. Budgeting and processes of anticipation link, therefore, directly to the area of management now commonly described as the Balanced Score Card or *Tableau de Bord*.

5. *Budget Process Requirements*

5.1 *A Sense of Direction*

There must be goals that preexist the beginning of the budgeting process. Each goal must have a time frame and be a milestone towards achieving the organization's mission or intent. In specifying its goals, the organization states who the stakeholders are that are taken into consideration.

5.2 *A Continuously Updated Shared Causal Model*

Decision makers and personnel must be able to understand some key 'cause and effect' relationships that affect the organization's output or performance. The budget rests on the identification of a causal model that helps understand what consequences might obtain from what actions in the perceived competitive environment. Much time is needed in a business to identify and maintain the model(s) or causal representations that will help understand how we could go from point X to point Y in the map of opportunities.

5.3 *A Monitoring System*

The monitoring system must be able to tell managers how well they are 'getting there' from 'here.' Some of the measures must be anticipatory or leading in the sense that they allow the managers to anticipate inappropriate or unacceptable results and therefore head off the undesired trend through appropriate corrective actions, while other measures are more historical and serve to keep track of what has been achieved.

5.4 *A Set of Criteria for Trade-off between Goal Achievement and Resource Consumption and Creation*

Whether the trade-off will be carried out in economic terms only or using other criteria such as market share, employee satisfaction, or citizenship of the firm, the firm needs to have some kind of preference ordering system that will allow for the identification of a preferred way each time a choice is to be made. It is, however, very difficult in any multi-person setting to come to an agreement on criteria for preference ordering. This difficulty should not prevent the managers from encouraging participation and expression on the part of everyone in the organization since that is the only way to open dialogue and communicate values and orientations.

5.5 *Above All, a Budget Requires Trust Between Participants*

A budget is grounded in delegation. Each person holds responsibility for a 'piece of the action' and must be able to count on the others doing what they promised to do. That promise includes a very crucial element: let others, who depend on the achievement of a milestone or of a result, know as early as feasible (through shared leading indicators) the actual changes in the probability to fulfill commitments. Budgets therefore require trust between participants (both vertically and laterally) and self-confidence combined with a reasonable willingness to take risks and not adhere to a prewritten script.

6. Budgets, 'Opportunity Gap,' and 'Performance Gap'

In projecting itself into the future, any organization faces two simultaneous challenges:

(a) Improve the 'productivity' (effectiveness and efficiency) of resources consumed in carrying out currently retained action plans.

(b) Identify new opportunities for continued growth that will require modification of current action plans and possibly the development of entirely new ones. Opportunities have a limited lifespan. If no new opportunities are identified and addressed, the organization will cease to grow and will decline into oblivion. Single-purpose organizations such as the March of Dimes or the Tennessee Valley Authority in the US have had to reorganize their mission completely to continue existing. A long-lasting organization such as the Kirin Beer Brewery in Japan is always reinventing itself because product attractiveness has a time limit and customers tend to be fickle unless the producer changes with the evolution of customer tastes.

We define the first challenge as that of managing the 'performance gap' while we will call the second one that of managing the 'opportunity gap.' As business environments grow turbulent, the continued relevance of any given action plan is less and less likely. Simultaneously, the need for continuous identification of new opportunities grows greater.

Figure 1 shows that, at any point in the continuum of turbulence, the required anticipation effort of an organization is always broken into two parts: managing the performance gap (increasing effectiveness and efficiency) and managing the opportunity gap (preparing the organization for the future). At each point on such a continuum, managers need to be able to hold two very different mental maps simultaneously, in differing proportions, as shown in the two 'wedges' at the top of Fig. 1. The coexistence of such differing relations to time and to uncertainty is not an easy task and the budgeting process offers a forum where these two mental maps can coexist.

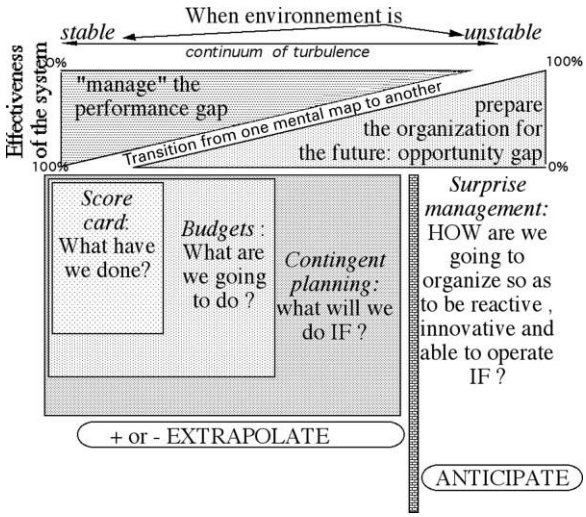

Figure 1
The tools of anticipation change with the level of turbulence

The effectiveness of the system requires that at each level of turbulence (competitive environment) the manager and the supporting processes take into account the appropriate mix of attention paid to the performance gap and to the opportunity gap. At each level of turbulence, different tools (including the ones relevant at the previous level of turbulence) meet the requirements of management. If the world is very stable, a simple scorecard about past results will suffice as the next action plan is a mere extrapolation or adaptation of the previous one. As the world becomes less stable, 'budgets' answering the question 'what will we do?' suffice to help the firm project itself into the future in a coherent way. The critical issue is to limit the action to have a balance between the resources available (increased by the resources created by the action plan) and the resources needed by the action plan. That vision of 'budgets' is still quite prevalent, as mentioned previously, despite the fact that turbulence conditions make it appropriate no longer.

When the world becomes really very turbulent, a business requires two simultaneous tools: contingent planning (answering the question 'what will we do if?') and 'surprise management' which seeks an answer to the question 'how can we organize ourselves now to be reactive, innovative, and able to operate under future conditions we cannot even imagine right now'? Contingent planning is largely about understanding the relationships between possibilities in the world in which the organization operates and lining up possible 'baskets of competence' for each identified discrete possibility, while surprise management is essentially about understanding relationships between internal capabilities and how they can be mobilized.

If we go from a less turbulent to a most turbulent environment, the first three phases described in Fig. 1 refer to various degrees of extrapolation (the past is full of information relevant for the future), while the last phase, once again, completely changes the mode of thinking. It also forces management to hold simultaneously two mental maps because the future cannot be forecasted.

Turbulence affects each level of responsibility differently in the organization. To a large extent, it is the responsibility of a superior to 'filter' some of the uncertainties and turbulence she or he faces for his or her subordinate. The higher the person is in an organization, the higher the degree of uncertainty she or he has to deal with. Therefore, the higher the person is, the more her or his anticipation efforts should be addressing the 'opportunity gap.' As Fig. 2 shows, even at the top of the organization, although the effort

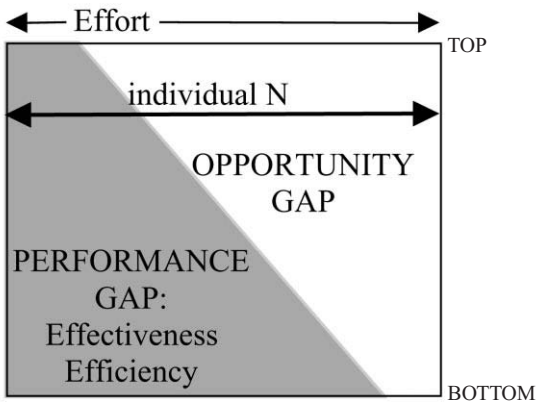

Figure 2
The effort is applied differently depending on the level of the individual in the hierarchy

of the individual is dedicated essentially to the management of the opportunity gap, there is still a part of the effort that is addressing the performance gap (effectiveness and efficiency).

As individual N moves from the bottom to the top of the organization, her or his preoccupation is more and more with the efficiency and effectiveness of the organization, but there is always a non-negligible preoccupation with some form of attention paid to the opportunity gap, recognizing that new opportunities can be identified at all levels of the organization.

7. Current Budgeting Practice is not Always Aligned with the Principles

As mentioned previously, it is a challenge to build a budgeting process that considers the issue of competitiveness not only from the point of view of the performance gap (efficiency and effectiveness, cost control, etc.) but also from that of the opportunity gap. The causes for such a difficulty can be traced to the following elements:

Objectives and processes to get to the target are often defined only *in* and *on* financial terms.

Partition of the process to achieve the target tends to be only hierarchical and functionally defined instead of being developed along the lines of transfunctional business processes.

The processes of alignment, dialogue, and team building through the budget construction are frequently ignored and the exercise is too often reduced to a scarce resource allocation leading to competitive behavior between organizational entities that ought to cooperate to serve the stakeholders best.

Value creation is mostly defined in terms of financial profit.

The time horizon of an action plan should be linked to the business cycle but is based, too often, on fiscal year constraints (which are defined by reporting needs and have little link with the need to have visibility of the future). Budgeting should be carried out almost as an ongoing process, provided information technology makes it economically feasible. Budgeting is a state of mind. The budget construction should be reengineered in most firms to take as little time as is possible (for example, by concentrating on the relationships in the causal model and on leading descriptors of the status of these relationships). If the budget were an ongoing process, frequent debates would take place that would continuously replace the firm on the most appropriate trajectory or path to its objectives and targets, thus developing confidence and commitment on the part of all in the organization.

Many times the budget has been highjacked as a substitute for an anticipatory form of financial accounting used in estimating year-end accounting results of the firm so that management can comfort stockholders about forecasted profit.

Focus is more on cost reduction than cost management (the 'right' cost is not always the least cost: costs are driven by product and process design, and by people eliminating non-value-adding activities).

Most action plans assume, usually erroneously, that costs are driven by the volume of sales and that costs are best controlled at the point where they can be measured (Lebas 1996). This leads to conformance budget (focused on resource allocation) thus limiting initiative and customer satisfaction.

Many budget processes merge and confuse anticipating and defining benchmarks for reward systems (especially if the latter are based on conformance with resource consumption) thus creating gaming behavior and budget slack which immobilize resources uselessly. Budget slack occurs when managers overestimate resource consumption or required effort and underestimate value created with the customers. Managers indulge in such behavior because most of the time they do not trust the leadership either to allocate resources or support action plans that are compatible with the manager's perception. Such behaviors immobilize resources and prevent the exploitation of opportunities. A performance management tool called 'flexible budgeting' consists in identifying clearly, before the fact, the areas of uncertainty each manager is responsible for (and which ones are not under her or his control) and in recalculating, after the facts are known, the elements of the action plan (consumption of resources as well as outcome or output) that would have been determined had the 'uncontrollable' elements been forecasted with pinpoint accuracy. Such an approach allows for total freedom in really anticipating because the manager knows she or he does not have to take precautions (through the building of budget slack) to cover risks she or he cannot control.

Often, the budgeting process does not implicate everyone and thus a frequent reaction is to consider that 'spending' resources allocated in the context of an

action plan becomes the objective rather than continuously validating the alignment with the intent of the plan of the actions that led to the original allocation.

The main weaknesses of many budget processes currently in place in organizations is that they are not built on a well thought-out system of:

(a) shared representations;

(b) subsidiarity and delegation which would rest on trust in people's ability to make the right decision at the right time, as close to the point of need as possible;

(c) representation of the firm as a network of business processes; and

(d) customer orientation.

8. *In Conclusion*

Anticipatory tools today are not constrained by the fiscal time-horizon; they are rolling and cover a period often closely related to product lifecycles. Anticipatory practices recognize the preeminence of the human aspect in the success of business processes. Choosing the right people and motivating them in well- thought-out processes make for satisfied customers who, in the end, drive the financials of the firm. Such a sequence requires flexibility and continuous adaptation.

Table 1 summarizes the main characteristics of the traditional and the 'new budgeting' approaches in a synthetic format. Budgets and budgeting processes, as they have been used in supporting the development of businesses since the end of World War I and mainly since the post World War II economic boom, seem to have reached their limit. New ways are emerging in coordinating energies in the firm to exploit opportunities. We have described their common denominator by outlining what is required to fulfill that purpose. What we have called new budgeting is far from being settled and probably never will be. Businesses are experimenting in many parts of the world; no solution seems, to date, to be vastly superior to any other, however. Firms like Svenska Handelbanken, Volvo Cars, or Skandia in Sweden, Schlumberger, Air Liquide, or Bull Computers in France, Borealis in Denmark, Boots the Chemist in the United Kingdom, or Sprint in the US have abandoned or are abandoning the 'old budget' approaches and inventing the 'new budgeting' (Hope and Fraser 1999). The real revolution is that we are entering an era of continuous adaptation and learning. Businesses need to adapt the tools too, and the 'budget' is not spared here. Because they have to be completely reinvented, budget processes should be even more fascinating to students, management researchers, and managers alike than they were before.

Table 1
Comparison of traditional and new budgeting (adapted from CAM-I Europe 1994, Lebas 1996)

	Traditional budgeting	'New' budgeting
Purpose	Financial performance forecast	Create and protect value
Structure	Multiple cost centers	Business processes
Resources measured	Costs only	Cost, quality, time, motivation, intellectual capital, loyalty, etc.
Resources view	Scarce, need to be partitioned	Can be obtained if creates value
Classification	General ledger items	Activity resource needs
Approach	Top-down command and control	Emerging and adaptive business process options
Objective	Control and minimize resource consumption	Create customer value through coordinated transverse activities
Mechanism	Historic costs extrapolation	Activity relationships and causality
Cost drivers	Single: output volume	Volume, product design, customer value offering, etc.
Data quality	Internal: implied accounting precision	Internal and external: ok to have approximation and non financial
Cost focus	Cost center spending	Hidden costs consumption
Logic	Simple and mechanical	Complex and dynamic
Reporting and time frame	Periodic and *ex post* analytical	Continuous, flexible, and explanatory
Responsibility	Budget holders	Process team
Deviation/variance	Return to conformance	Understand, learn, and adapt
Behavior induced	Competition	Cooperation
Rewards	Linked to conformance to budget	Linked to outcome long term coherence with strategic intent
Status	Entrenched and well known	Rare and still experimental
Vision	Incremental and bureaucratic	Radical and creative

See also: Budgeting: Political Science Aspects; Organization: Overview; Organizational Decision Making; Organizations and the Law; Organizations, Sociology of

Bibliography

CAM-I Europe 1994 *A Journey to Advanced Management Systems: A Report on the Research and Findings of the CAM-I Advanced Budgeting Study Group*. R-94-AMS-0.1.1 Poole, UK

Hope J, Fraser R 1999 Beyond budgeting. *Management Accounting (UK)* January

Jiambalvo J 2001 *Managerial Accounting*. John Wiley and Sons, New York

Lebas M 1996 Budget control. In: Warner M (ed.) *International Encyclopedia of Business and Management*. International Thomson Business Press, London; Routledge, New York

Lebas M 1999 Which ABC? Accounting based on causality rather than activity-based costing. *European Management Journal* **17**(5): 501–11

M. Lebas

Budgeting: Political Science Aspects

Budgeting is a process by which governments allocate public resources to bureaucracies and client groups. This process usually takes place on an annual basis, involves procedural rules and institutional arrangements, often invokes intense political conflict, and, together with taxation, forms the basis for fiscal and macroeconomic policy. Because it fundamentally determines political winners and losers through its allocative function, budgetary processes and outcomes serve as distinct guide to a political state's philosophy of government. The act of budgeting, therefore, lies at the heart of politics and government.

1. *The Study of Budgeting*

Budgeting may be studied from several perspectives. From the vantage point of professional schools of public administration, the practical elements of budgeting include the actual drafting and formatting of budget documents, the application of cost-benefit and other tools of program analysis, the projection of revenues and expenditures, the division of spending into its operational and capital components, and the management and implementation of programs funded by the budget. Economists principally focus on the macroeconomic consequences of budgeting, such as the effects of deficit spending on interest rates, inflation, the crowding out of private investment, foreign exchange rates, and national debt management. At a more microeconomic level, economists may also consider the effectiveness of specific programs funded in the budget, such as determining the influence of public jobs programs on the labor market. These questions may also be of interest to and directly addressed by political scientists, but the centrality of political science research looks at budgeting as a guide to a broader analysis of politics, relations between institutions, the behavior of political actors within those institutions and specific public policies.

2. *Budgeting as Incrementalism*

Modern theoretical understandings of budgeting begin with the idea of 'incrementalism.' Formulated in the late 1960s, incrementalism attempts to explain how budgetary decisions are made, how budgetary actors carry out their roles, and what the consequences are for politics by structuring budgetary decisions in one form as compared to alternative ones (Wildavsky 1964, Fenno 1966). Incrementalism is grounded in certain assumptions about human nature, that people are inherently limited by cognitive, time and resource constraints from being truly rational and synoptic in their ability to make budgetary decisions. Budgetary actors 'muddle through' and 'satisfice' rather than 'optimize.' They are constrained from making decisions in the way that microeconomically rational consumers might, where all information sources and alternatives are weighed before a decision is reached. To compensate, these actors rely on 'aids to calculation'. In particular, instead of reviewing the entire budget to determine what is funded, the focus of analytical and political decision making is on the annual change, or increment, in spending. The budget base, or last year's spending, is taken as a given, and the politics of budgeting revolve around determining the size of incremental 'fair shares' allocated to budgetary claimants.

Incrementalism takes into account the role of institutions and actors within those institutions. Budgeting is regarded as a decentralized, 'bottom-up process,' where budget proposals emerge from within executive branch agencies. These agencies in turn act as 'advocates' for larger increments before the Appropriations Committees in the House of Representatives and the Senate. Committee members evaluate these funding claims and act according to a set of behavioral norms. The House Appropriations Committee, for example, operates as the 'guardian of the purse,' while the Senate committee acts as a 'court of appeals.' The overarching norm is that budgets should be balanced.

Incrementalism views budgeting above all else as a political activity, and it gives limited consideration to notions of administrative efficiency or economic rationality in the determination of budgetary out-

comes. Both agencies and politicians instead rely upon incremental aids to calculation. For bureaucrats, incrementalism implies stability and long-term growth in resources. For politicians, incrementalism reduces societal tension and political conflict, as interested groups quickly learn that once their claims are included as part of the budget base, they may count on regular, presumably growing increments.

3. *Challenges to Incrementalism*

Critics of incrementalism soon argued that the model had become obsolete. First, alternative methods of budgeting came into vogue that directly challenged incrementalism's assumptions of limited human rationality and constrained organizational resources. Zero-Based Budgeting (ZBB) and then program budgeting in the form of Planning-Program-Budgeting Systems (PPBS) were designed to improve the choices of decision makers (Schick 1966). The supporters of ZBB declared that the entire budget, rather than just the increment, could be analyzed each year by breaking up the budget into discrete 'decision units' and then reprioritized and reassembled. The defenders of PPBS argued that multiyear planning rather than annual increments enabled decision makers to set broad programmatic goals.

Second, incrementalism, its critics claimed, neglected to account for the rise in entitlement spending, generally focused on the nondefense share of the budget, and ignored the role of the president in setting the budgetary agenda and in shaping outcomes. Where incrementalism depended upon budgetary growth for its conflict-reducing fair shares and growing budget bases, the rise of deficit spending, recessions and ideological demands for reductions in the size of the public sector, caused governments to engage in 'decrementalism' rather than incrementalism. Instead of providing for orderly budgetary growth, by the mid- and late 1970s, governments confronted the highly difficult political task of cutback budgeting.

Throughout the 1980s and 1990s, political scientists noted that governments throughout the world coped with fiscal stress by adapting similar budgetary techniques (Schick 1986, 1988). Budgeting became more centralized, 'top-down' and 'front-loaded' to reduce the influence of interest groups and budgetary claimants. Macrobudgetary guidelines set at the beginning of the budgetary process constrained micro-decisions during the remainder of the process. Governments employed hard targets, ceilings and caps to limit spending; calculated long-term inflationary and program costs through baseline budgeting; created new legislative committees and support agencies to impose greater oversight on the budgetary process; enacted budget resolutions and reconciliation to control entitlements; and strengthened the powers of finance ministers in constraining agency spending demands. Budget balancing reasserted itself as the focus of American politics, and these efforts at deficit reduction were characterized by the enactment of the Balanced Budget and Emergency Deficit Control Act of 1985, better known as Gramm–Rudman–Hollings, the Budget Enforcement Act of 1990, and the Balanced Budget Agreement of 1997 (Savage 1988, White and Wildavsky 1989). In Europe, the 1992 Maastricht Treaty imposed deficit and debt targets on countries seeking to join the European Monetary Union (von Hagen 1992, Wildavsky and Zapico-Goni 1993).

These changes in law not only altered budgetary decisions and institutional arrangements, they signaled a retreat from governments pursuing an activist fiscal policy through budgetary deficits. Spending might indeed violate caps and targets, but these expenditures were often incorporated into the budget through accounting gimmicks, rather than some publicly justified notion of employing Keynesian budgets to stimulate aggregate demand. In this sense, the drive towards deficit reduction and balanced budgets reflected the ascendancy of moderate and conservative governments and the pressures of globalization on public policies.

4. *The Rational Choice Approach To Budgeting*

Since the early 1980s, many political scientists have come to rely upon economically inspired rational choice models, rather than incrementalism's social psychological approach to study budgeting. The basic assumption of rational choice is that political actors rationally seek to maximize their self-interest. Holding power is central to political self-interest, and in democratically elected governments this means being re-elected. Consequently, political institutions and public policies, including budgetary processes and outcomes, are structured by politicians for electoral purposes.

Symptomatic of re-election budgeting is pork-barrel spending, where members direct appropriations to their states and districts to win political credit with their constituents. Pork-barrel projects are also allocated as 'side payments' to create coalitions of legislators in sufficient numbers to pass legislation. Depending upon the model, the size of these coalitions of legislators vary from 'minimum winning coalitions' that contain a majority plus one legislator, to 'universal coalitions,' where all legislators who want a project are given one. The budget is therefore characterized by programs where projects are provided for each legislative district. As a result of pork-barrel spending, the government spends money on economically dubious projects that create large, systemic budget deficits with what are presumed to be negative consequences for the macroeconomy (Shepsle and Weingast 1981).

Like incrementalism, the basic rational choice model has not gone unchallenged. Although politicians do engage in pork-barrel spending, the cost of

these 'distributive' projects is simply insufficient to account for the huge budget deficits incurred by the industrialized democracies in the 1980s and 1990s. Entitlement spending, principally in the form of pensions and health care expenditures, rather than pork-barrel spending is clearly the primary source of budgetary growth. The rise of large-scale deficit spending in the USA, moreover, may also be traced to stagflation, recession, tight monetary policy, tax cuts and the buildup in defense spending during the 1980s. Furthermore, politicians in many countries have risked their own re-election chances by pursuing politically costly fiscal policies, including cutbacks in distributive spending, in order to reduce these deficits.

The minimum winning and universal coalition elements of the rational choice models, which have usually been expressed in formal, mathematical terms, have also been challenged on empirical grounds. Individual, nonentitlement government programs appear to benefit only a limited number of legislative districts. Even when bundled together for purposes of creating voting coalitions, most program 'portfolios' fail to provide universal benefits for all districts. While the rational choice assumes that pork-barrel budgeting takes place for re-election purposes, individual constituents overwhelmingly receive their government benefits from entitlement and general benefit programs, not low visibility pork-barrel funding (Stein and Bickers 1995).

5. *Institutions, Divided Government and Budgeting*

Where rational choice, with its focus on institutions, and incrementalism, which examines roles and norms, experience some convergence is in the analysis of divided government's effect on budgeting. Political institutions and roles within those institutions matter. Empirical evidence indicates that when competing political parties control the executive and legislative branches of government, these governments experience greater difficulties in controlling budget deficits and expenditures than unified ones (Alesina and Rosenthal 1995). Leftist parties spend more than rightist parties, and both parties differ in their spending priorities. Changes in the American budgetary process since 1974 enhanced the power of the parties to set overall budget and fiscal priorities, though the detailed spending decisions remain the prerogative of the committee system. When control of the government is divided, these priorities conflict and fiscal stalemate likely results. Similarly, the more unified and conservative the government, the greater the priority is budget balance (Kiewiet and McCubbins 1991).

Incrementalists add that divided government also influences the advocate and guardian roles of budget makers. Whereas the early incremental models regarded specific roles as enduring, the behavior of budgetary actors during the rise of divided government indicated they were in fact mutable (Cox et al. 1993). When rightist parties control the executive branch, for example, the executive becomes less of an advocate of spending and more of a guardian or protector of perceived budgetary excess. The more leftist oriented the House Appropriations Committee becomes, the less its emphasis on guardianship, particularly in the presence of a conservative executive. A decreasing level of guardianship is associated with the growing size of budget deficits. The Senate Appropriations Committee, meanwhile, tends to remain in its role as a court of budgetary appeals.

6. *Budgeting and Political Science*

Incrementalism and rational choice theories have, despite their limitations, greatly contributed to political science's analysis of the institutional forces and politics that create budgets. These two theories also offer important policy implications of budgeting. Incrementalism, for example, points to the norm of balanced budgets as a reoccurring policy goal of American politics. Rational choice, for instance, accounts for the continued presence of often wasteful pork-barrel spending, which threaten the success of government programs and public policies.

See also: Business and Society: Social Accounting; Financial Accounting and Auditing; Public Administration: Organizational Aspects; Public Administration, Politics of; Public Bureaucracies; Public Sector Organizations; Rational Choice in Politics

Bibliography

Alesina A, Rosenthal H 1995 *Partisan Politics, Divided Government, and the Economy*. Cambridge University Press, New York

Cox J, Hager G, Lowery D 1993 Regime change in presidential and congressional budgeting: role discontinuity or role evolution?*American Journal of Political Science* **37**: 88–118

Fenno R F Jr. 1966 *The Power of the Purse: Appropriations Politics in Congress*. Little, Brown, Boston, MA

Kieweit D R, McCubbins M D 1991 *The Logic of Delegation*. University of Chicago, Chicago

Savage J D 1988 *Balanced Budgets and American Politics*. Cornell University Press, Ithaca

Schick A 1966 Planning Programming Budget System—Symposium: The road to PPB: the stages of budget reform. *Public Administration Review* **26**: 243–58

Schick A 1986 Macro-budgetary adaptations to fiscal stress in industrialized democracies. *Public Administration Review* **46**: 124–34

Schick A 1988 Micro-budgetary adaptations to fiscal stress in industrialized democracies. *Public Administration Review* **48**: 523–33

Shepsle K A, Weingast B R 1981 Political preferences for the pork barrel: a generalization. *American Journal of Political Science* **25**: 96–111

Stein R M, Bickers K N 1995 *Perpetuating the Pork Barrel*. Cambridge University Press, New York
von Hagen J 1992 Budgetary procedures and fiscal performance in the European communities. *Economic Papers* **16**: 1–79
White J, Wildavsky A 1989 *The Deficits and the Public Interest*. University of California Press, Berkeley, CA
Wildavsky A 1964 *The Politics of the Budgetary Process*. Little, Brown, Boston, MA
Wildavsky A, Zapico-Goni E 1993 *National Budgeting for Economic and Monetary Union*. Martinus Nijhoff Publishers, Dordrecht, The Netherlands

J. D. Savage

Burckhardt, Jacob (1818–97)

Jacob Burckhardt was born in Basle, Switzerland on 25 February 1818 as the son of the pastor of Basle cathedral. At first he studied theology in Basle, although after a religious crisis he changed to history and history of art, and continued his studies in Bonn and Berlin from 1841 to 1843. He was most influenced, especially in Berlin, by Franz Kugler, Johann Gustav Droysen, August Boeckh, and Leopold Ranke. He habilitated at Basle in 1844, held lectures in history and history of art there, and edited the 'Basler Zeitung' from 1843 to 1845. After the uprisings in Switzerland of 1844–5, an early manifestation of the midcentury revolutionary unrest in Europe, his previously liberal political outlook changed into a decided conservatism. With this he turned away from his political and publishing activities in favor of the 'education of old Europe.' In 1855 he was appointed to a chair at the technical university of Zurich, and in 1858 at Basle, where he taught history until 1886, and history of art from 1874 to 1893. In 1860 he stopped publishing and concentrated on his lectures, although he completed a number of manuscripts after this point. He died in Basle on August 8, 1897.

Jacob Burckhardt's historical writing can be understood as a critical diagnosis of the present. This found its most clear expression in his lectures on the history of the revolutionary era and *Über das Studium der Geschichte* (*On the Study of History*) as well as in references to the present throughout his work. Burckhardt rejected the liberal historians' optimism in progress, and saw a danger for his ideal of culture in the contemporary processes of mechanization, industrialization, and fundamental democratization. He also felt the modern development of large states to be a threat, particularly that of the new German Reich from 1871, and he made accurate prognoses of the political catastrophes of the twentieth century. However, his historical pessimism was in contradiction at least to his early religiously influenced opinions on and admiration of art. In numerous long tours through Italy he worked on the basis of his extensive works on art history, and these journeys through Italy, France, and The Netherlands remained an important part of his life until his old age. In 1855 he published *Cicerone. Eine Anleitung zum Genuß der Kunstwerke Italiens* (*Cicerone. An Introduction to Enjoying Italy's Art*), a work which visitors to Italy interested in art still use today. In 1867 his work, *Kunst der Renaissance* (*Art of the Renaissance*), was brought out by his friend Wilhelm Lübke. In this work he deliberately avoided a history of the artists themselves and instead concentrated on a 'history of art according to functions.' His late masterpiece *Erinnerungen an Rubens* (*Memories of Rubens*) was published from his papers after his death, in which the re-evaluation of the Baroque which he had completed in his later years found its expression. His portrayal of Renaissance art had an important influence on the educated German public's reception of the Renaissance and Italy. With his concept of a history of art ordered according to functions he guided the emerging study of art on to new methodological paths.

Burckhardt's basic historiographical principles are most clearly illustrated in his introductory lecture *On the Study of History*, which he held three times between 1868 and 1873, and which his nephew Jacob Oeri edited and published in 1905 under the now famous title *Welthistorische Betrachtungen* (*Reflections on History*). Burckhardt drew up his images of the historical world from a universal historical perspective, assuming three basic human needs: (a) the political (state), (b) the 'metaphysical' (religion), and the 'critical' (culture). In this scheme of 'potencies' he relativized the importance of the nation and the nation state in conscious contrast to German historicism, and allotted an equally important role to culture as the expression of human spontaneity as to the 'stable' powers of state and religion. As culture he understood 'the quintessence of all that which has come into existence spontaneously for the promotion of material life, and as the expression of the spiritual and emotional–traditional life, all social gathering, all technologies, arts, writing and sciences.' Burckhardt thereby rediscovered an enlightenment tradition of cultural history, and consolidated it in that he saw the polarity of freedom and violence in a developed form not only in state and religion, but also in culture. However, the potencies theory also enabled him to relativize the stress on a typological basis. The interaction of the three potencies adds depth and complexity to historical reality, as Burckhardt experienced it.

In his three historical masterpieces, *Die Zeit Constantins des Großen* (*The Age of Constantine*), (1853), *Die Kultur der Renaissance in Italien* (*The Civilisation of the Renaissance in Italy*) (1860) and *Die griechische Kulturgeschichte* (*The Cultural History of Greece*) (*3 volumes, 1898–1900*) Burckhardt varied the potencies scheme in that he added a further definition

of culture to state and religion. Culture is then the entirety of the achievements of a human creative urge, which is, however, permanently under threat of being destroyed. In all three works Burckhardt assumes a state of acute breakdown of authority and an increased societal and moral antagonism. In this latent chaos the respective forms of state and religion represent the relatively stable potencies, which, however, remain fragile because the 'spirit' strives persistently as a 'subversive,' as a destructive, and also creative ancient force, to bring forth new forms.

In *The Civilisation of the Renaissance* Burckhardt focused on Italy of the the fifthteenth and early sixteenth century, in contrast to Sismondi and Michelet, and thereby lent the term 'Renaissance' the temporal and geographical boundaries it has retained up to the present day. Burckhardt drafted the image of an epoch in the sway of violence and crime, lawlessness and tyranny, in which, however, equally the rational calculation of seizing and retaining power became established and so created the specific modern form of the 'state as artform.' The Humanists established themselves alongside the tyrants as a temporally specific human type, in whom the modern individual was embodied just as it was in the rulers. For Burkhardt they were the 'most obvious examples and victims of unbridled subjectivity' with a new intensity and sophistication of emotional life as well as a new method of rationalizing their relationship to the world. In *The Cultural History of Greece* Burckhardt describes similarly to the image of the state in *The Civilisation of the Renaissance* the Greek state, the polis, as a place of 'state slavery of the individual' in which democracy degenerated into a tyranny of the majority. Burckhardt developed the principle of agon as an essential structural feature of life in the Greek polis, the competition between individuals, yet also the bitter struggle of the poleis against one another. This principle does have destructive effects, yet at the same time it brings about the preconditions for the development of cultural freedom and a singular creative wealth. In Burckhardt's view the Greek form of religion was, however, just as constitutive as the principle of agon, as it knew neither theology, theocracy nor the rule of priests, unlike the states of the old orient, and so enabled a free development of spiritual life. As leading experts on ancient history criticized sharply, Burckhardt made a decisive contribution to the demythologization of the neohumanist–idealist reception of Greece with his portrayal of Greek society.

Burckhardt's greatest success by far among his audience, apart from *Cicerone*, was with *The Civilisation of the Renaissance in Italy*. This is due to a consciously narrative and nonscientific diction, the consistency of his analysis and references to contemporary society kept continuously alive in the subtext, which consistently refer to the dialectic of the rationalization of the individual relation to the world and the development of modern subjectivity. Burckhardt hereby made an important contribution to a structural analysis of the modern age. He overcame the thought of individuality of German historicism with his typological thinking, which orders material according to the aspect of 'cultural importance,' and anticipated Max Weber's cultural historical method. The potencies theory is reminiscent of the conception of equal systems which, however, exercise an influence over each other, or of historical areas of reality in modern sociology (Talcott Parsons for example). Burckhardt's potencies theory differs though, in that economy and society are not taken into account, and in the important position of religion and the specific function of culture, among other factors. Nor did Burckhardt, who always portrayed himself as an unsystematic thinker, develop a 'historical method.' Yet he did make an essential contribution to both the theoretical and methodological reflection of historical research. He analyzed the relationship of prejudice and historical judgement and the effects of the emancipation of knowledge from political rule and religious belief more precisely than most of his contemporaries. In the accelerated social and cultural change of the 'Revolutionary Age' he discovered an elementary compensatory need for the preservation of traditions. He encountered every tendency to political–social instrumentalization of historical knowledge, out of progressive or conservative interest, with extreme skepticism. In conscious contrast to the contemporary professionalization of historical science he retained an older, personal understanding of learning and education. Historical education should not contribute to an alteration of the outer reality, but widen the mental horizon of those seeking knowledge. With this maxim, his independent cultural historical method, and his art of portrayal he not only in fact influenced the educational horizon of those interested in history to a great degree, but also had a stimulating effect on the increasingly specialized areas of historical science (ancient history, modern history, history of art).

See also: Art History; Civilization, Concept and History of; Disciplines, History of, in the Social Sciences; Historicism; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Renaissance

Bibliography

Burckhardt J *Gesamtausgabe*, Dürr E et al.

von Boch S, Hartau J, Hengeross-Dürkop K, Warnke M (eds.) 2000 *Kritische Gesamtausgabe*, Vol. 6: Beiträge zur Kunstgeschichte von Italien, Vol. 10: Aesthetik der bildenden Kunst.

Burckhardt J 1949–86 *Briefe. Vollständige und kritisch bearbeitete Ausgabe*, Burckhardt M (ed.) 10 vols.

Burckhardt J 1974 *Vorlesung über die Geschichte des Revolutionszeitalters in den Nachschriften seiner Zuhörer, Rekonstruktion des gesprochenen Wortlautes*. Ernst Ziegler, Basel/Stuttgart

Burckhardt J 1982 Über das Studium der Geschichte. *Der Text der weltgeschichtlichen Betrachtungen aufgrund der Vorarbeiten von E. Ziegler, nach den Handschriften*. Munich
Gilbert F 1992 *Geschichte, Politik oder Kultur? Rückblick auf einen klassischen Konflikt*.
Große J 1997 *Typus und Geschichte*. Eine Jacob-Burckhardt-Interpretation
Hardtwig W 1974 *Geschichtsschreibung zwischen Alteuropa und moderner Welt*.
Guggisberg H R (ed.) 1994 *Umgang mit Jacob Burckhardt*.
Hardtwig W, Burckhardt J 1990 Weber M zur Genese und Pathologie der modernen Welt. In: *Geschichstkultur und Wissenschaft*, pp. 189–223
Jaeger F 1994 *Bürgerliche Modernisierungskrise und historische Sinnbildung. Kulturgeschichte bei Droysen, Burckhardt und Max Weber*
Kaegi W, Burckhardt J 1947 *Eine Biographie*.
Löwith K, Burckhardt J 1966 *Der Mensch inmitten der Geschichte*, 2nd edn.
Schulin S 1983 *Burckhardts Potenz- und Sturmlehre*.
Siebert I, Burckhardt J 1991 *Studien zur Kunst- und Kulturgeschichtsschreibung*.

W. Hardtwig

Bureaucracy and Bureaucratization

Bureaucracy has several meanings, some positive, most negative. The sociologist Max Weber (1946) coined the term 'bureaucracy' and treated it as synonymous with rational organization: bureaucracy embodies the ideals of rational–legal authority, under which all decisions, except executive decisions, are based on rules that are internally consistent and stable over time. Political scientists tend to think of bureaucracy as governance by bureaus having the following characteristics: bureaus are large, are staffed by full-time employees who have careers within the organization, and rely on budget allocations rather than revenues from sales since their outputs cannot be priced in voluntary *quid pro quo* transactions in the market (Downs 1967, Wilson 1989). There is a third definition of bureaucracy, which is far less flattering: bureaucracy is inefficient organization, is inherently antidemocratic, cannot adapt to change, and, worse, exacerbates its own errors (Crozier 1964). Discussions of bureaucracy tend to be ideologically tinged. The political left emphasizes the rationality and neutrality of government while downplaying the power of bureaucracy itself, while the right uses bureaucracy as an epithet or shibboleth and focuses on bureaucracy's antidemocratic tendencies and inefficiencies. Outside of academic circles, the rightist view of bureaucracy as inefficient and antidemocratic administration prevails. At least in the USA, very few positive references to bureaucracy can be found in the press. For most people, bureaucracy is unbusinesslike administration, an epithet or an accusation.

1. Bureaucracy vs. Traditional Organizations

Weber's analysis of bureaucracy begins by comparing the structure of bureaucracies with traditional organizations. This comparison is made along four dimensions: differentiation, integration, constraints, and incentives.

Compared to traditional organizations, bureaucracies are highly differentiated. There is horizontal division of labor: jobs are specialized and responsibilities are strictly delimited. There is vertical division of labor as well: there are higher and lower offices, the latter subordinate to the former, all of which are ultimately accountable to the head of the organization. Perhaps most importantly, there is a clear differentiation of official duties from personal interests and obligations, what Weber calls separation of home from office. Traditional organizations, by contrast, are undifferentiated. They do not delimit responsibilities, do not separate higher from lower offices, and do not distinguish personal from official capacities-instead, the organization mirrors the stratification system of the larger community (Dibble 1965).

Compared to traditional organizations, bureaucracies have numerous integrating mechanisms, among them written rules and regulations, procedures for selection and advancement of officials, and a specialized administrative staff charged with maintaining these rules and procedures. Written rules, formal procedures, and specialized administrative staffs are largely absent from traditional organizations—actions taken by traditional organizations need not be consistent.

Compared to traditional organizations, bureaucracies constrain the conduct of officials while offering powerful incentives for compliance. There are several types of constraints: actions must be justified in terms of the larger purposes of the organization; actions must be guided by the norm of impersonality, that is, detachment and objectivity is required in all decisions; and advancement is contingent on contributing to the purposes of the organization. Traditional organizations have no distinctive purposes apart from the purposes of the people participating in them; norms of impersonality and objectivity are practically nonexistent because the organization and the community are indistinct; and there is no possibility of advancement within the organization because, again, stratification within the organization is determined by the stratification of the community.

The incentives offered by bureaucracies include the prospect of a lifetime career, salaries paid in cash rather than in kind, and (in Europe if not the USA) a modicum of prestige attached to the status of the official. Careers, salaries, and prestige based on position within the organization do not exist under traditional administration. The elements of differentiation, integration, constraints, and incentives render bureaucracies more powerful than traditional organizations yet more responsive to central authority. The

power of bureaucracies results from their capacity for coordinated action. Traditional organizations are incapable of coordinating large-scale action, except temporarily when the interests of elites coincide. The responsiveness of bureaucracies to central authority arises from top-down control of ultimate decision premises, the policies and objectives of the organization, which is impossible in traditional organizations where authority is fragmented. Paradoxically, the same elements that make bureaucracies powerful yet responsive to central authority render individual bureaucrats nearly powerless. Limited responsibilities, subordination to higher authority, and numerous rules and regulations, as well as the norm of impersonality deprive bureaucrats of latitude in decision-making. The powerlessness of individual bureaucrats, moreover, is exacerbated by their dependence on the organization for their income and social standing. Weber claims that the capacity for coordinated action, responsiveness to central authority, and dependence of individual bureaucrats on the organization promote administrative efficiency. 'Precision, speed, unambiguity, knowledge of the files, continuity, discretion, unity, strict subordination, reduction of friction and of material and personal costs—these are raised to the optimum point in the strictly bureaucratic administration...' (Weber 1946, p. 214). The context in which Weber was writing is critical. Weber was not comparing modern businesses with bureaucracies. Rather, he was comparing bureaucracies capable of coordinated large-scale action with traditional organizations that were not.

2. *Bureaucracy vs. Business Administration*

Many of the elements of bureaucracy identified by Weber are present in modern business organizations, but some are not. Like bureaucracies, large businesses are highly differentiated—jobs are specialized, a chain of command runs from higher to lower offices, and people are expected to ignore personal interest when acting in their official capacities. As in bureaucracies, many of the integrative mechanisms used are also present in large businesses, among them written regulations, procedures governing selection and advancement (today called human resource management), and specialized administrative staff. And similar constraints operate in bureaucracies and businesses—actions must be justified by the purposes (today called strategy) of the organization, norms of impersonality prevail, and advancement depends on contributing to the organization's purposes. The similarity of the administrative structures of bureaucracies and business firms occurs for several reasons. Many administrative practices arise from bounded rationality limits, people's limited capacity to make and execute complicated decisions. No one person is capable of assembling an airplane or winning a war. Instead, complicated activities are coordinated through administrative hierarchies that transform broad decision premises (maximize shareholder value, win the war) into specific programs of action (marketing plans, battle plans) and guided by general rules that people accept as conditions of employment (Simon 1976).

Research on organizational structures also shows that organizational size is the principal determinant of administrative structure for both bureaucracies and businesses (Blau and Schoenherr 1971). Both bureaucracies and businesses, moreover, are constrained by norms of rationality, which means that people expect to find a chain of command, consistent rules and purposes, and rewards based on qualifications and accomplishments whether or not they are technically necessary (Meyer and Rowan 1977).

In other respects, however, businesses and bureaucracies have evolved differently in the last 50 years. Public and private administration were similar at the time Weber was writing in the early 1900s. Much of the US public sector was modeled explicitly after the private sector at that time. It is not accidental that the reform movement in the USA, which called for governmental administration devoid of politics (or, to use another term, 'businesslike' administration), coincided with the emergence of scientific management, which called for rational selection, training, and supervision within firms. Nor is it accidental that in the 1940s, the same theory of administration was believed to apply to public- and private-sector enterprises (Gulick and Urwick 1937). Public and private organizations have diverged in the last 50 years, however. Divergences have occurred in several domains, most notably in organizational design, accounting practices, and incentives and performance measurement. With respect to organizational design, firms now choose from among diverse organizational designs. Many larger firms have alternated between functional and divisional organizational designs, that is, between designs where the principal units are specialized by function (such as purchasing, manufacturing, and sales) so as to maximize scale economies, and designs where the principal units are self-contained businesses containing a full array of functions so as to focus attention on bottom-line objectives. Some firms have experimented with matrix organization, where people report to both product and functional managers, and still other firms have more complicated designs where some functions are centralized while others are dispersed across geographic or product units. Public bureaucracies, in contrast to business firms, exhibit very little variety in organizational design. Almost all remain organized by function as they were 90 years ago. To be sure, some central government functions have been decentralized to local governments (this is called 'loadshedding' in the USA), contracted to the private sector, or privatized outright, but the core organizational design for public bureaucracies remains functional.

With respect to accounting, public-sector agencies have departed substantially from business practice. At the beginning of the twentieth century, public entities issued consolidated financial reports and maintained capital accounts just like private businesses. Consolidated accounting gave way to much more complicated fund accounting during the 1920s, when it was believed necessary to segregate revenues and expenditures intended for different purposes into separate funds. Capital accounting has all but disappeared from the public sector—few governments distinguish between current and capital expenditures—though accounting for long-term indebtedness is retained because it is required by creditors.

With respect to incentives and performance measurement, public sector bureaucracies face more complicated issues than private businesses and, as a consequence, appear to lag substantially behind business. Businesses use several approaches to measure and reward performance. To begin with, most publicly-held businesses define performance as total return to shareholders and reward performance so defined using techniques like value-based management. Under value-based management, executive compensation is based largely on gains in share prices, and the compensation of middle managers is based on a metric (for example, return on invested capital or economic value added) believed to drive shareholder value. The non-financial drivers of financial performance are captured in measures of internal process and customer satisfaction. Moreover, firms' internal operations are gauged against industry benchmarks assembled by third parties, usually consultants or trade associations. There is no analog to total shareholder returns in the public sector, at least none that can be easily measured—the electoral process is a coarser measure of performance than the stock market, and it is difficult to reward civil servants for electoral performance. Public agencies are taking some initial steps toward measuring non-financial performance by setting objectives and gauging progress toward them—performance measurement of this sort is part of the initiative, in the USA, to 'reinvent government' (Osborne and Gaebler 1992) and also by measuring customer (constituent) satisfaction. Even so, competitive performance assessment hardly exists for government—the concept of benchmarking has not yet penetrated the public sector. In the USA, at least, performance comparisons across governmental units are strongly resisted, mainly on the grounds that the functions and outputs of governmental units cannot be compared across jurisdictions.

3. *Liabilities of Bureaucracy*

There is widespread agreement that bureaucracies suffer some liabilities in comparison with business. Many of these liabilities of bureaucracy are systemic in that they arise from the structure of bureaucracy rather than from people's predilections. Sociologists and political scientists have focused on what they call dysfunctions of bureaucracy, among them displacement of goals, so-called vicious circles in which dysfunctions feed on one another, and spiraling bureaucratic growth. Economists, by contrast, have drawn attention to the efficiency disadvantages of bureaucracies as compared to firms, asking whether, in general, nonmarket transactions are inefficient compared to market transactions, and, specifically, whether the funding of bureaucracies through budgets rather than market transactions results in overproduction of bureaucratic services. These potential liabilities of bureaucracy should be reviewed *seriatim*.

3.1 *Displacement of Goals*

Bureaucracies are known for rigid adherence to rules and procedures, even when rules and procedures appear to impede the objectives of the organization. The notion of goal displacement provides both a description and an explanation for this seemingly nonrational conduct. Goal displacement, following Robert K. Merton (1958), describes the process whereby means become ends in themselves, or 'an instrumental value becomes a terminal value'. The displacement of goals is especially acute in settings—such as bureaucracies—where the following conditions exist: the technical competence of officials consists of knowledge of the rules; advancement is contingent on adherence to the rules; and peer pressure reinforces the norm of impersonality, which requires rules and procedures to be applied with equal force in all cases. What is important is that goal displacement, at least as originally conceived, argues that bureaucracies are efficient in general—under conditions anticipated by their rules and procedures—but inefficient in circumstances that cannot be anticipated. The implications of goal displacement for innovation and new product development have been realized only gradually: bureaucracy can be antithetical to innovation.

3.2 *Vicious Circles*

A more thorough critique of bureaucracy argues that dysfunctions are normal rather than exceptional and, moreover, that dysfunctions accumulate over time such that organizational stasis is the expected outcome. The elements of the vicious circle of bureaucratic dysfunctions are impersonal rules that seek to limit the discretion of individual workers, centralization of remaining decisions, isolation of workers from their immediate supervisors as a consequence of limited decision-making authority, and the exercise of unofficial power in arenas where uncertainty remains.

Thus, as Michel Crozier observes (1964), maintenance people exercise undue influence in state-owned factories because their work is inherently unpredictable and cannot be governed by rules. The logic of vicious circles, it should be pointed out, yields several consequences. To begin, new rules will arise to eliminate whatever islands of power remain in the organization, but these rules will trigger further centralization, isolation, and power plays as new sources of uncertainty arise. Additionally, to the extent that the organization is opened to uncertainties arising externally, line managers have the opportunity to reassert power that would otherwise erode through the dynamics of vicious circles. External crisis, in other words, may be an antidote to bureaucracies' tendency toward rigidity over time.

3.3 Spiraling Growth

Bureaucratic systems also tend toward growth, other things being equal (Meyer 1985). Until recently, growth of government and of administrative staff in private firms was endemic. The causes of growth lie in several factors, but chief among them are people's motives for constructing organizations in the first place. People construct formal organizations in order to rationalize or make sense of otherwise uncertain environments; organizations, in fact, succeed at making the world more sensible; as a consequence, there is continuous construction of bureaucracy and hence bureaucratic growth as people attempt to perfect their rationalization of an inherently uncertain world. Three comments are in order.

(a) The logic of bureaucratic growth is built into neoclassical administrative theory developed by Simon (1976) and others. Irreducible uncertainty in the environment, in conjunction with the belief that administrative organization can rationalize uncertainty, will result in continuous growth in administration.

(b) The growth imperative is so strong that deliberate campaigns to 'downsize' or 'restructure' organizations must be launched in order to achieve meaningful reductions in staff.

(c) Inefficiency. Economists have asked persistently without resolution whether bureaucracies are inherently less efficient than private-sector enterprises. Several answers have been proffered, none fully satisfactory.

From the 1940s to the present time, the Austrian school of economics, von Mises (1944) and others, have argued that any departure from market principles yields both inefficient transactions and antidemocratic tendencies. This position has proved difficult to reconcile with contemporary transaction-cost theories (Williamson 1975), which argue that hierarchies may be more efficient than markets under some circumstances. In the 1970s, the efficiency question was cast somewhat differently: might bureaus, which depend on budgets for their sustenance, overproduce compared to firms subject to the discipline of the market? (Niskanen 1971) Here too the answer was equivocal, as analysis showed that rent-maximizing monopolists would have similar incentives to overproduce whether they were located in public bureaucracies or private firms. Despite the uncertain analytic underpinnings for the belief that bureaucracies are more apt to harbor inefficiencies than private-sector organizations, privatization of governmental functions is occurring rapidly and with positive results in many countries. It is unclear whether the liabilities of public bureaucracies are simply the liabilities of established organizations that have been shielded from extinction for too long, or whether bureaucracies suffer disadvantages in comparison with private organizations because of weak incentives and inconsistent objectives.

4. Research on Bureaucracy

Organizational research and research on bureaucracy were once synonymous or nearly so, as the bureaucratic model was believed descriptive of all organizations, for-profit, non-profit, and governmental. Case studies of bureaucracy written during the 1950s and 1960s encompassed government agencies and industrial firms alike as evidenced by titles like Gouldner's (1954) *Patterns of Industrial Bureaucracy*. Early quantitative research on organizations, such as the work originating at Aston University and the University of Chicago, focused mainly on relations among elements of organizational structure (size, hierarchy, administrative ratio, formalization, centralization, etc.) that flowed from the bureaucratic model implicitly if not explicitly. As attention shifted to external causes of organizational outcomes, however, the bureaucratic model has become less relevant to organizational theory. Thus, for example, the key causal variable in resource dependence models of organizations (Pfeffer and Salancik 1978) is control of strategic resources, which is more germane to businesses than to government bureaus. The key dependent variables in organizational population ecology (Hannan and Freeman 1984) are births and deaths of organizations, which are infrequent in the public sector (see Kaufman 1976). The new institutional theory of organizations has downplayed Weber's notion of bureaucracy as rational administration and has substituted for it the notion that all organizations, bureaucratic and nonbureaucratic alike but especially the former, pursue social approval or legitimation by appearing to be rational rather than actual efficiency outcomes. Almost alone among major organizational theorists, Philip Selznick argues that the problem of bureaucracy remains a central issue in organizational theory: '… the ideal of an effective, fair, and responsive bureaucracy remains elusive. Our society desperately needs organized ways of dealing with social problems; we cannot rely solely on market strategies. Yet the specter of bureaucracy

still haunts and repels, still saps public confidence and weakens support for collective action' (Selznick 1996, p. 276).

Several streams of research on bureaucracy remain, although they are at the periphery of organizational theory. One stream of research, largely in the field of public administration, focuses on the possibility of bureaucratic reform, of reducing the otherwise endemic dysfunctions of bureaucracy. This research does not promise solutions; quite the opposite, it finds bureaucratic dysfunctions intractable. For example, reform initiatives almost inevitably create new positions for people to oversee the reform, resulting in new layers of administration and 'thickening' of each layer (Light 1997). Decentralizing managerial authority in public bureaucracies is offset by a corresponding centralization of political authority, with little net improvement in efficiency or responsiveness (Maor 1999). Attempts to 'reinvent government' by reshaping bureaucracies into so-called performance-based organizations are usually frustrated by conflict over what the performance standards should be (Roberts 1997).

Another stream of research, pursued mainly by economists, focuses on the impact of incentives and performance standards on performance outcomes in bureaucracies. The tentative conclusions of this research are, again, unpromising. Only weak incentives are available in bureaucracies. Not only do bureaucracies perform multiple tasks, but they also answerable to multiple constituencies or, in the argot of agency theory, principals. Multiple tasks weaken incentives in all organizations because the outcomes of some tasks can be more readily observed than the outcomes of others. Multiple constituencies or principals, however, are unique to bureaucracies and further weaken incentives because each constituency will bargain separately with the bureaucracy (Dixit 1996). Specific performance standards built into legislation do discipline the behavior of bureaucrats somewhat, but measures of short-term results can work against long-term objectives and are frequently gamed (Heckman et al. 1997).

See also: Bureaucratization and Bureaucracy, History of; Industrial Sociology; Organizational Climate; Organizational Culture; Organizational Culture, Anthropology of; Organizational Decision Making; Organizations: Authority and Power; Organizations, Sociology of; Weber, Max (1864–1920)

Bibliography

Blau P M, Schoenherr R 1971 *The Structure of Organizations*. Basic Books, New York

Crozier M 1964 *The Bureaucratic Phenomenon*. University of Chicago Press, Chicago

Dixit A 1996 *The Making of Economic Policy*. MIT Press, Cambridge, MA

Downs A 1967 *Inside Bureaucracy*. Little, Brown, Boston

Gouldner A W 1954 *Patterns of Industrial Bureaucracy*. Free Press, Glencoe, NY

Heckman J, Heinrich C, Smith J 1997 Assessing the performance of performance standards in public bureaucracies. *American Economic Review* **87**: 389–95

Kaufman H 1976 *Do Government Organizations Ever Die?* The Brookings Institution, Washington DC

Light P C 1997 *Thickening Government: Federal Hierarchy and the Diffusion of Authority*. The Brookings Institution, Washington DC

Maor M 1999 The paradox of managerialism. *Public Administration Review* **59**: 5–18

Merton R K 1958 Bureaucratic structure and personality. In: Merton R K (ed.) *Social Theory And Social Structure*, 2nd edn., Free Press, Glencoe, NY, pp. 195–206

Meyer M W 1985 *Limits To Bureaucratic Growth*. de Gruyter, Berlin

Niskanen W 1971 *Bureaucracy And Representative Government*. Aldine, Chicago

Roberts A 1997 Performance-based organizations: Assessing the Gore plan. *Public Administration Review* **57**: 465–78

Selznick P 1996 Institutionalism 'old' and 'new.' *Administrative Science Quarterly* **41**: 270–77

Simon H A 1976 *Administrative Behavior*. 3rd edn., Free Press, New York

von Mises L 1944 *Bureaucracy*. Yale University Press, New Haven, CT

Weber M 1946 Bureaucracy. In: Gerth H, Wright Mills C (eds.) *Essays In Sociology*. Free Press, Glencoe, NY, pp. 196–244

Wilson J Q 1989 *Bureaucracy*. Basic Books, New York

M. Meyer

Bureaucracy, Sociology of

The word 'bureaucracy' has been, for a hundred years, one of the keywords of the social sciences. The problems of bureaucracy were one of the favorite battle grounds of intellectual debates in sociology and in political science. Curiously enough, bureaucracy is a very fuzzy word which does not correspond to an acceptable definition. There has never been a rigorous concept, nor a precise definition but this lack was itself a reason for its success since fuzziness provided scope of much debate.

The traditional meaning of bureaucracy seems to have been 'government by specialized and obedient bureaus.' This supposes the existence of a state apparatus composed of appointed, i.e., not elected, civil servants, organized in a hierarchy and reporting to the sovereign authority, originally to the king, then to the nation as represented by its elected representatives.

Bureaucratic power, in this definition, implies the '*Etat de droit*'—the state of law. At the same time, it rejects the citizen's direct participation. Parallel to this traditional use which may be relatively precise, another much broader meaning has gradually taken

hold and is related to the theory of bureaucratization. Bureaucratization means the rationalization of human activities into bureaucratic organizations, i.e., organizations arranged hierarchically and staffed by well trained obedient personnel. Bureaucratization supposes the concentration of bigger and bigger production units and the massive extension of the state apparatus. The theory is built upon a deterministic postulate: it is ineluctable because of the constraints of complexity and growth. Bureaucratic organizations are a necessity also because they look as if they were the only solution to our technical problems. But bureaucratization does not necessarily refer only to public organizations. The same trend can be found in all forms of activity, including industrial, commercial, social, or political organizations.

The two meanings of the word 'bureaucracy' are very often mistaken: the problem of political choice between electing and appointing office holders and the problem raised by the inevitability of new modern forms of organization which are necessarily bureaucratic. This confusion has been compounded by the emergence of a third, more popular, meaning. Bureaucracy, in common language, has become synonymous of dilatory procedures, clumsiness, routine, and procedural complications. Whatever the context, bureaucracy calls to mind the difficulty in adapting organizations or institutions to the needs and demands of the people they are meant to serve. Bureaucracy, in this context, is a permanent evil. The emotional content of the word in current language cannot be removed from the intellectual debate and this has prevented the emergence of a neutral scientific definition.

One cannot discuss the growth of bureaucratic forms of organization as a natural phenomenon since it means the growth of an evil. The confusion between these three different meanings has created one of the basic nightmares of social forecasting. If one admits that the rise of big and complex organizations will inevitably encourage the development of nondemocratic forms of management and government and the proliferation of dehumanizing experiences for employees and for customers, one will have to predict the end of our unresponsive civilization or the coming of a redemptory crisis.

Logically however, there is no real link of causality between these different facts which are, on the other hand, only partially acceptable. Research in social sciences over the last fifty years has focused on disentangling the summary logic results of such confusion and has provided empirical data with which to answer the various questions that the ideological debate has raised:

Is the concentration of productive and administrative activities in large units really inevitable?

Does it carry with it a bureaucratic model of management? Other models of managing people may be possible.

Is there a strong link between these bureaucratic models and the evils of routine and alienation about which the ordinary citizen complains?

Is there a link between the concentration of organizations and the evolution of the model of political authority?

Three intellectual currents have successively emerged from these major questions. The first of these is the Weberian current which appeared between 1900 and 1930. Max Weber was both the instigator and the person who formulated its most refined form. Drawing on a rich philosophy of history and a rationalistic view of human relations, he built a well documented and sensible analysis of the evolution of the models of organization and their basic consequences. A second more empirical wave, emerged in the 1930s and 1940s in the US and a few years later in Europe. Based on observations of the reality of contemporary organizations and of the way it was experienced by employees and managers, it has focused especially on the gap between theory and reality and has thus developed facts and interpretations about the 'dysfunctions' of the Weberian model. Since the 1960s, a new wave has emerged. This acknowledged the gap between theory and practice but has sought to understand the system which results from the pressure of managers on employees so that they will comply with the theoretical model and from the pressure of employees who succeed in changing this model when adjusting to it. This neo-rationalist analysis makes it possible to put bureaucratic forms into the perspective of a broader postindustrial model, for which rationalization is no longer the dominant logic.

Not only is it difficult to recognize the partial inefficiency of bureaucratic organization, but it becomes now indispensable to conceive and put into practice nonbureaucratic organizations that could be efficient and profitable. In such a new context, at a time when more and more activities are appearing, the traditional dilemma of bureaucracy tends to decline in the public debate. However, we can observe a new cleavage between public administration, which has become the domain of a more sophisticated analysis of public choice, and business firms, whose complexity of management escapes the simple dilemma of the traditional debate.

1. The Weberian Ideal Type

The reigning concept of bureaucracy has been shaped by the success of large scale organizations in the 1900s, be they American and German corporations, state apparatuses, or the organization of social democratic parties in Europe. Max Weber, the founder of the theory of bureaucracy, was especially influenced by the achievements of the Prussian state and its mirror image, German social democracy. For Weber, the ideal type of a bureaucratic organization, similar to

the Prussian state, comes from the combination of a legal rational model of social control and a hierarchical model of organizational authority.

This national specificity was neglected since the model was rational and could be used to understand all modern administrations. The emergence of the bureaucratic form was the consequence of disenchantment (*Entzauberung*) with the world, the basic characteristic of modern societies. Max Weber's theory is basically historical. We have moved from a patrimonial to a charismatic concept of authority, and then to a rational-legal one.

The rational-legal order requires that functions, rules, and procedures be impersonal. Agents (civil servants) must be specialized and well-trained and they must report to superiors within a hierarchical system. Order could be exercised by a collegial type of organization but Weber believed in the superiority of the pure bureaucratic type, which was monocratic. Its success, he thought, was unavoidable, in the same way that machine tools had proved a success in mass production industry. This type of organization, which started in traditional public service organization, invaded big capitalist enterprises and went on to conquer hospitals, political parties and churches as well.

Weber perceived, especially at the end of his life, the dangers of servility which may come from such a system and which may, as a consequence, undermine democracy but he never questioned its technical efficiency. At the same time, a young union official, Robert Michels, who was active in Italy and Germany, pushed this reasoning further. Starting with an analysis of the mechanism of power in labor unions and social democratic parties, he formulated another law, the iron law of oligarchy. According to Michels, militants who wanted to bring deep social transformations, had to operate through collective action which means organizations, i.e., bureaucracies. But, the existence of bureaucracies was not compatible with democratic goals and values, which were the only justification of collective action. This proved to be a tragic dilemma and led to one of the great political debates in Europe and in the world about ends and means.

But the weakness in Weber's reasoning, i.e., his implicit admiration for the infallible superiority of the bureaucratic machine, was never questioned. The great political leaders and thinkers of the revolutionary period of the 1930s believed in the machine. For them, the problem lay in seizing power, not exercising it. They thought that, once the state power was taken over by the proletariat, no basic problem would ever exist. The results were appalling but the leaders and thinkers of the revolutionary effervescence of the 1920s and 1930s were blind to them. They reacted by emphasizing the danger of a catastrophic term of human society and thus revolutionary pessimism became the dominant model of the period and developed into arguments for the coming revolution. The logic was very simple. By exaggerating the danger of the evolution of oppression, one is pushed to a desperate wager: either socialism, or the coming of barbary. Because of this contradiction, one calls for revolution which will dialectically overcome the problem.

The revolutionary current was, of course, a minority current, but its cultural success made a profound impact on society and on its politics. It can still be found as an overtone in the best-selling social novels from the 'iron heel' of Jack London to James Burnham, the era of organizers and even in William H Whyte Jr.'s *The Organization Man*.

2. *The New Paradigm of Dysfunctions*

According to the rationalist theory, it was argued that evolution would establish the dominance of the bureaucratic form of organization, as it was the only form able to provide the stability and predictability which was needed in the modern world. Using this logic, one could try to predict the development, conditions, and consequences of its evolution.

The deep trauma of the Great Depression disturbed these confident beliefs in the US and led to new and quite different research questioning this theory. Gradually, a new paradigm developed around the concept of dysfunctions, based on the discovery of the severe gap that existed between employees' behavior in reality and behavior as prescribed by the model. New authors showed that bureaucracy was not as effective as the rationalist model predicted, and sought to understand why this was happening and what were the mechanisms of ineffectiveness.

The new dysfunction paradigm can be summarized in the following way. Decisions taken from a rational perspective bring about a chain of secondary consequences that run counter to the established objectives. This very practical logic was so pervasive that the negative properties of the bureaucratic model were stabilized into 'dysfunctions' that were as characteristic of the bureaucratic model as its rational capacities of action.

The emergence of this new approach was stimulated by the discovery of the importance of the human factor in the way organizations function, as seen from Western Electric's experience at Hawthorne. It was also echoed the political discussion about ends and means which had raged in revolutionary and reformist circles in Europe. The seminal contribution in this line was Robert K Merton's article (1936) on 'The unexpected consequences of purposeful action,' in which Merton argued that the discipline necessary to obtain the standardized behavior indispensable to achieving the objectives in a bureaucratic structure would provoke a major displacement of goals among the civil servants, who would behave in a ritualistic manner.

What was to be means to attain the end became an end in itself. This brought about a strong rigidity which prevented civil servants from responding realistically to the concrete demands of their tasks. The caste behavior of civil servants thus created a basic gap with the public which accounted for the ineffectiveness of their action and prevented the attainment of agreed official goals.

Empirical research carried out during the 1940s and early 1950s by American sociologists and in the 1950s by European sociologists provided a number of cases validating this reasoning. The bureaucratic model of action created behavioral rigidities among subordinates who had to apply the bureaucratic rules. This created conflicts between superiors and subordinates and between performers and their public, and introduced a need for control and regulation. Thus the unexpected consequences and the dysfunctions of the bureaucratic model tend to reinforce the model.

Why does such a model persist since it is not effective? Alvin Gouldner (1954) has shown that it reduces tensions generated by subordination and control but, at the same time, it maintains their existence.

3. *The Neo-rationalists*

The first proponents of bureaucratic rationality borrowed without questioning the experience of Taylorian engineers who considered the members of an organization as mere cogs in the machine. Against this very narrow view, the proponents of the dysfunction paradigm saw them as affective human beings whose behavior was influenced by the conditions they were enduring.

At the end of the 1950s, a new wave of analysis developed into a real reversal of perspective, in which each human being was recognized as also being a free and active human agent inside the organization. This led to a new theory of action and the re-emergence of the problem of power.

The new theory of action resulted from consideration of a new approach: decision analysis. Herbert Simon was the inspirer of this breakthrough. He questioned the principle of the 'one best way,' according to which, once an objective is established, there is only one best way to attain it. The engineer's job is to discover and to calculate that best way. After analyzing the practice of decision-making in different settings, Simon showed conclusively that human beings—and engineers are human beings—were incapable of reaching the one best way because there was no absolute rationality. Rationality is bound by the costs of information and even more by the limitation of cognitive problems. The discovery of these limitative problems led Simon to propose a new theory of limited rationality. Since human beings are incapable to reach absolute rationality, they will stop their search when they will reach a 'satisfying' rationality. With his colleague, James C March, they applied this model to the field of organization and bureaucracy. Their book *Organizations* was a real landmark. According to this new perspective, it was possible to return to rational analysis. Human relations analysis cannot be placed in opposition to rational analysis but can be integrated with it inasmuch as it will define the conditions based on which the satisfying point can be calculated.

Meanwhile a new theory of power relations was being formulated based on game theory and using experimental research on groups. Political science became the cutting edge of research. The battleground shifted from bureaucracy to public problems.

4. *The Decline of the Bureaucratic Paradigm*

If one wishes to summarize the state of research and the major trends, it seems quite clear that the bureaucratic paradigm is no longer central to sociology and political science. But past research on bureaucracy has given rise to a number of connecting fields:

(a) First of all, the debate has moved from the questioning of the trend towards bureaucratization to analysis of the complexity of new large scale organizations that cannot work under the principles of hierarchical authority at the time of NTIL, networking and constant restructuring.

(b) A new cleavage has developed between public organizations and other organizations. Public organizations are less dominated by organizational paradigm as decentralization and subsidiarity. Debates turn around public policy problems, costs-advantage 'techniques,' accountability, and evaluation.

(c) In private organizations, the basic paradigm has become more managerial. After a long period during which the basic problems centered on strategy, from the end of the 1980s, the basic paradigm has been applied to the functioning of the internal system. The principle of hierarchical authority as the central axis of organizational order is now being questioned more effectively. With the progress of information technology, the argument for a nonhierarchical, lateral principle of organization has partially won. Re-engineering has proved to be essential for the re-development of organizations.

(d) The most promising fields are now to be found in the experience of intervention and change and the roles of innovators and change agents.

One may surmise however that bureaucracy will reemerge as an active field of research. The problems of implementing new practices of cost-advantage calculus and evaluation cannot be solved simply by imposing best practices because they are more rational. It becomes a problem of innovation in a bureaucratic setting, i.e., a problem of restructuring bureaucracy. The same is true of re-engineering in private organizations. Intervention requires understanding bureaucratic resistance. The strategy of

change brings us back—at least partially—to the old debate on dysfunctions.

See also: Authority: Delegation; Bounded Rationality; Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Hierarchies and Markets; Management: General; Oligarchy (Iron Law); Organization: Overview; Organizations: Authority and Power; Organizations, Sociology of; Weber, Max (1864–1920)

Bibliography

Bennis W 1966 *Changing Organizations*. McGraw Hill, New York

Burns T, Stalker G M 1961 *The Management of Innovation*. Tavistock, London

Crozier M 1965 *The Bureaucratic Phenomenon*. University of Chicago Press, Chicago

Crozier M, Fiedberg E 1977 *Actors and Systems*. University of Chicago Press, Chicago

Drucker P 1969 *The Age of Discontinuity*. Harper and Row, New York

Eisenstadt S M 1963 *The Political System of Empires*. Free Press, Glencoe, NY

Erikson E 1964 *Child Hood and Society*, 2nd edn. W W Norton, New York

Friedberg E 1997 *Local Orders – Dynamics of Organized Action*. JAI Press,

Gouldner A 1954 *Patterns of Industrial Bureaucracy*, Glencoe, NY

Maccoby M 1990 *Why Work*

March J C, Simon H 1958 *Organizations*. Wiley, New York

Melton R K 1952 *Reader in Bureaucracy*. Free Press, New York

Michels R 1949 *Political Parties – A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. Free Press, New York

Moss Kanter R 1989 *When Giants Learn to Dance*. Simon & Schuster, New York

Peters T, Waterman R 1982 *In Search of Excellence*. Harper and Row, New York

Selznick P 1949 *T.V.A. and the Grass Roots*. University of California Press, Berkeley, NY

De Tocqueville A 1952 *The Old Regime and the Revolution*. Basil Blackwell, Oxford, UK

Toffler A 1980 *The Third Wave*. Morrow, New York

Weber M 1965 *Wirtschaft und Gesellschaft*. Cologne-Berlin (English 4th edition)

M. Crozier

Bureaucratization and Bureaucracy, History of

Bureaucratization is a process, and the term in itself thus refers to a historical development. It is, however, used in several distinct meanings, some of which refer to moral and political values, which are endangered by a growth of bureaucracy according to a common belief. As a descriptive term, apart from its evaluative content, bureaucratization may mean: (a) a growth of numbers of administrators and administrative organs between points of reference; (b) a growth in administrative complication in the handling of matters; and (c) an increase in administrative competence and independence and in the functional division of administration. These three definitions may be inter-related, though this is not necessarily the case. The third one comes close to the analysis first made by Max Weber of the development of a bureaucratic principle (Weber 1922).

There is also a great difference in the reference of the term in different contexts. Here one may distinguish first (a) a reference to state administration only or (b) a reference to any sort of administration, whether public or private.

A third distinction may also be made. Sometimes the term bureaucratization is used in regard to states and societies irrespective of their political system. On other occasions the use is restricted to states and societies where a separation is made, at least in principle, between political and administrative duties and responsibilities. The latter limited scope of the term is, of course, much more related to the near history than to past centuries, with the nineteenth century as an intermediate period when this separation first came to be more commonly observed.

The different aspects of the concept of bureaucratization and their close relationship to political ideals regarding the scope of state interference in relation to citizen's rights and the division of responsibility between politicians and bureaucrats make it difficult to use the concept in historical and social science analysis. Yet, in regard to changes in power and control, it covers an important reality, which has to be differentiated from the use of administration in general. The following account goes out from Weber's analysis combined with the distinction between political and administrative responsibilities.

1. From Antiquity to the Nineteenth Century

There is a striking difference between the state of Egypt in the Antiquity, or the Mesopotamian empires of Babylon and Assur, or Rome compared to the Medieval states of Europe including the vast empire of Charles the Great. This difference has to do with the ambitions of the state. While delegation to subprinces was the rule in the feudal system that emerged in medieval Europe, the state was in principle centralized as to the important decisions in the early empires. It is often said that these empires were highly bureaucratized. The Russian historian of the Antiquity Michail I. Rostovcev in the first half of the twentieth century wrote several books in which he tried to show that the process of bureaucratization was going on in Egypt,

the Hellenistic Empires and the Roman Empire (Rostovcev 1936–8, 1941, 1957). In spite of his brilliance his views are not current among ancient historians of today. Some would rather stress the limits of state power in the Antiquity and others the specific structure of the ancient system of government in different parts of the world. What was common to these regions was the fusion of politics with administration. No real difference was made, and only the king or emperor exerted a political power meant to express a deliberate policy. He headed an organization where the highest officials had some political duties but their policy-making had mostly to be concealed under administrative clothing. Further down in the hierarchy, the administrative chores were the real occupation of the employees. Often, administrators were neither full-timers nor independent, and regulations tended to vary with the will of the ruler. Further, it is important that the state in the Antiquity was highly centralized and that no clear difference was made between public law and private law until the late Roman Empire and, thus, no real limit was set between state obligations and rights and a private sector. Mostly the kings and emperors commanded only 'their' servants and demanded only what belonged to them, though if these relations were of private or public nature was often not quite clear. This 'bureaucratic' state was, thus, vague in its boundaries but, as a centralized state, it needed an efficient bureaucracy. It is hardly evident that bureaucracy should have grown or extended its procedures from one period to another or from one empire to another during the period before AD 400.

China was in many ways the same sort of bureaucratic-political state as imperial Rome but the Chinese system started earlier and continued in the dynasties all up to the end of the empire in 1911. The Confucian class of state servants was hierarchically ordered under the emperor, who had the political power in his hands. No real delegation of political power was used, but rather a blend of political and administrative obligations, which included a strict rule regimen. The bureaucracy was organized in a refined system already in the Ch'in and Han dynasties (221 BC–220 AD). There were ministries and ministers, but even these, and some senior officials or 'excellencies' who were the highest in rank, could influence policy via their channels to the emperor. Their responsibilities did not include the task to form a policy even for the field for which they were responsible. Hierarchy was enormously developed both within each ministry or office and also in the country as a whole. Local administration took place via organs in at least five different levels. Yet, it will seem that all political decision-making stayed with the emperor in spite of a system of checks and balances within the bureaucracy (Loewe 1986, Bielenstein 1986).

The politico-bureaucratic system of China, Egypt and imperial Rome under the Antiquity did not change in principle, which makes it difficult to use the term bureaucratization in each of the cases. The fundamental structure of bureaucratic government did not develop even though specific modifications were introduced in the division of tasks and responsibilities. Except in China, Korea and Japan, a fundamental debureaucratization took place in the centuries after AD 400. In the nomad empires of which the one of the Mongols was by far the most extensive, virtually no bureaucracy existed. When the Mongols had taken China, the Chinese system of government became theirs but not in the parts of the empire that broke away from Mongol China. Bureaucracy was also kept small in the Arab and Islamic states, which were rapidly formed in the seventh to tenth centuries. European feudal princes had very few officials who were employed to administer and such officials seldom had a full time employment for their administrative work and they could almost never rely on a rule system, which might ensure a field for decision-making without interference and with some autonomy. Only within the Catholic Church a system resembling a bureaucracy survived.

Only with the rise of central power in the fifteenth to seventeenth centuries administrative institutions grew again in Western Europe, and in Eastern Europe the same system existed for much longer, in Russia until 1905 and partly until 1917. There are many similarities between this system of bureaucratic forms of government and the one of ancient China and Rome. A specific instruction for future bureaucrats emerged. Universities played a role in this, and China had such educational institutions from an early date. In Europe two different types of reforms increased the theoretical competence of the bureaucracy. First, in centralized monarchies there was a need for educated secretaries and fiscal employees, who could master the subtleties of foreign and internal relations and international and national law. This need was partly met through recruitment of low-estate university men, who were sometimes promoted into nobility (the 'noblesse de robe'). Second, nobles who wanted to preserve to nobles influential posts in the immediate vicinity of central power, arranged specific educational institutions for young nobles, to make them fit for bureaucratic leadership. It also became customary to send young nobles to universities in a tour over Europe to polish their education in a fitting way. Centralized monarchy in this manner again made bureaucracy important in the European way of handling public affairs. The well-functioning of the administration of the state was a pre-condition for efficiency in foreign policy and warfare, and these were the ultimate aims of most of these monarchies (Raeff 1983). In the middle of the seventeenth century military entrepreneurs were finally made servants of the state and a military bureaucracy evolved. Later 'cameralism', the 'science' of civil administration, developed mainly in the German states (Mann 1986/1993). The relatively

poor Swedish state became a power of significance partly because of the efficiency of its bureaucratic apparatus, organized in 'collegia' defined according to functional criteria, and with (in principle) salaried employees. This organization impressed Peter I of Russia, so that he reformed the Russian bureaucracy according to the same model, soon modified by his successors (Peterson 1979, Shepelyov 1999).

The new European bureaucracies represented some innovations in the process of a bureaucratization of state and society. Some of the state bureaucrats of the seventeenth and eighteenth centuries had a real full-time employment and were salaried for their services. Rule systems were given to ensure that the administrations worked in an efficient manner and on a daily basis. A formalized routine of handling incoming matters and a hierarchical organization were intended to give the rational flow of work, which was the aim of the monarchs who set them up. However, obstacles to a rational decision-making process were universal. In Russia, complaints were common that copying in triplicate and more, of documents took a long time and slowed down the process (Shepelyov 1999). The heads of administrations also had a privileged position and were normally not regarded as bound by regulations in the same way as the lower echelons of bureaucracy. Often salaries were paid to nominal civil servants that were allowed to employ substitutes to perform the real duties. Finally, autonomy was restricted in the bureaucracies. Top bureaucrats often combined a role as a counselor to the monarch with the administrative leadership. Lower bureaucratic strata of course tried to observe the wishes of their superiors. This tended to mean that political considerations and political ambitions mixed into the work of the bureaucracy and this was regarded as natural (Yeroshkin 1997). As long as no ambition was found to separate politics and administration and to differentiate, for example, between a minister and a head of an administration, this seems to have been unavoidable.

2. *Bureaucratization and Parliamentary Politics from the Early Nineteenth to the End of the Twentieth Century*

The development of bureaucratization took a new turn from the nineteenth century in Europe and the US and, because of European imperialism and the domination of these countries in world politics and economy this new turn made a global impact. The first half of the century was filled with parliamentary and bureaucratic reforms and the second half saw different stages of democratization take place. Politics and bureaucracy thus became centers of societal change. Constitutional rule was the first political demand from the 'left' (relative to the political scale of the time) all over Europe, and its fundamental issue was to get rid of 'despotism' in the form of monarchic rule through bureaucracy. This demand meant to give a decisive influence to elected bodies at least in fiscal and educational matters. It became usual to contrast bureaucratic and political responsibilities.

In this situation, during the first half of the nineteenth century, a long-standing request for reform of the civil service in several countries was finally agreed to in several countries. In Germany, France and Britain these demands led to results and extensive reforms of their civil service took place (Mann 1986/1993, Gillis 1971, Charle 1980, Thuillier 1980, Campbell 1955, Mackenzie and Grove 1957). These reforms had as their main objectives to reduce or get rid of nepotism in the appointment of personnel, to abolish corruption and thus to establish a certain independence and objectivity among bureaucrats. Those who have investigated these reforms have not contended that they were entirely successful in their aims, but it is generally said that a huge improvement took place. Since these days West European bureaucracy has established a reputation for being normally impartial and honest.

During part of the nineteenth century when the ideas of economic liberalism flourished, state intervention was normally confined to a restricted area, though practice never conformed to theory. When the security of the state or vital interests of common nature to the whole society were at stake, the state should intervene. This meant that military purposes were generally excluded from the restriction for the state to engage in economic activities and also some public education for the supply of civil servants and, in fact, common interests could be extended into new fields. In principle, however, the civil servants had to protect state interests through legal means and the military to protect such interests through means of force and violence. The bureaucracy in the strict sense thus was an instrument for a 'legal domination,' one of the forms of power acknowledged by Max Weber. During most of the nineteenth century bureaucracy was regarded with distrust by the liberal forces in Europe and America and its growth was seen as an evil. Conservatives, with another evaluation of law and legal forms, tended to appreciate bureaucracy as a counterweight to democratic rule and as a guarantor of equal treatment and long-term obligations of the state system. This did not imply, however, that conservatives generally were favorably inclined to any growth of bureaucracy but, rather, that they wanted it to be strong and limited.

What holds for the bureaucratic development in Europe and the US is not valid for the European colonies. The European powers pursued another policy in their colonies, which were ruled almost entirely through bureaucracy. Real political bodies existed only in the 'mother' countries and the bureaucracy in the colonies were responsible to the

governments in their political center and, indirectly to their political mandators. Political elements in the colonies were few, elected bodies had little or no influence of the administration and, most important, administration and military power were closely united in terms of command and responsibility. For example, in India the heads of bureaucracy (the civil administration, Indian Civil Service) in each district were intertwined with the military forces in the districts. Though they were separate organizations they were used together, if required, in command and operations, and the military personnel was used to enforce the implementation of the regulations of the civil service. This system, slightly different in the colonies of different European powers, was based on bureaucratic rule in the colony. Thus, it differed completely from what was tolerated in the colonial states at home and it meant a bureaucratization, which was never only a legal domination but always also domination by violence.

At the close of the nineteenth century the economy of most European countries and North America started to grow rapidly after a protracted slump. Investments increased and trade and industry grew. As shown by Jürgen Kocka the structure of the industrial and commercial firms changed. Very big companies arose, and these companies met new organizational problems. The structure of the Siemens and Halske Company had been, earlier, one of a patrimonial type with an allegiance of the salaried employees to the owners directly. Their loyalty was ensured through bonuses and other benefits given on a personal basis from the members of the Siemens family. In the 1880s a total reorganization took place. The Company replaced the family. General regulations were given instead of personal instructions from the leadership. Benefits were regulated and depended on the firm (Kocka 1969). Although this may have been an unusually clear case, several other examples show that this type of impersonalized bureaucratization took place in many industrial and commercial enterprises during the late nineteenth and early twentieth century (Torstendahl 1991) (see *Business History*).

Not only did bureaucratization get an inroad in each company but industries also started to organize themselves into federations both as employers and as producers and traders. Such employer's associations and industrial organizations were intended to give support to their members through a general knowledge of industries and branches. This required an administrative staff. In this way private sector enterprises both as individual firms and as organizations of firms took a decisive step into a formal bureaucratization of their activities in the late nineteenth or early twentieth century. Thereby they also forced the states to act. It was impossible for states to rely only on their traditional administration, when organizations of industry made demands on the state and its policy. States had to engage competent personnel, in the first hand technically educated persons who could meet the demands from industry and trade. These new bureaucrats should also assist states in their demands on industry, in the first hand in questions of security—inspectors of ships, elevators, constructions, bridges, railroads etc. were engaged and organized into new administrative units. Security of labor became another source of state activity (Winkler 1974, Bruguière et al. 1985, Torstendahl 1991, Chap. 4). Voluntary organizations and, not least, parties were also drawn into the organizational mainstream and got a bureaucratic structure (Michels 1911).

It is well known that World War I (see *First World War, The*) meant an enormous growth of state intervention in the economy all over Europe. In order to make resources suffice for the most urgent needs the state had to regulate vast areas of the economy entailing an extensive bureaucratization. After the war most of the new administrations were abolished. The recollection of the war years and the large-scale bureaucratization of society during these years was, however, alive with many people and it formed a source of inspiration to some both on the left and the right (Rials 1977).

The reshaping of Europe after World War II (see *Second World War, The*) brought a new stage to bureaucratization: the social welfare administrations. They were foreshadowed in Scandinavia and France before the war but became a general phenomenon in West Europe only after the war. States then took the lead in the economies of European societies and, with or without nationalization of industries, wanted to carry through extensive planning. The variations between West European states were considerable in regard to political will. Enthusiasm for planning and welfare was great in the socialist and social democratic camps in countries like France, Britain, and Sweden (Rousso 1986; Rousso 1987, Morgan 1984) (see *Socialism: Historical Aspects*). In Germany and Italy, for a long period dominated by their Catholic and conservative parties, governments also accepted some state intervention. (Baring 1982, Mammarella 1985), accentuated with political change in the late 1960s. All over Western Europe administrations grew in size and importance.

Around the end of the 1960s complaints about bureaucratization had become numerous and bureaucracies had begun to behave independently of politicians, not any longer implementing political demands but also presenting more and more their own policies. (Aberbach et al. 1981, Mayntz and Scharpf 1975). This new phase of bureaucratization meant that politicians began a drive to reduce the influence of bureaucracy on policy-making. The cutting up of big administrations became usual in Europe and efforts to strengthen the influence of politicians in relation to bureaucracies were frequent in the 1970s and 1980s. Curtailed bureaucracies however, showed their tenacity through

branching off and re-establishing on a local or regional level. Political initiatives may have been transferred more to the politicians but in pursuing their policy politicians were dependent on negotiations with different interest groups and different bureaucratic bodies within the public sector. In the new negotiating economy bureaucracies played a distinctive role though mostly in the background or as participants in the groups forming policy solution in complete packages. Bureaucratization was no longer as obvious as earlier, but it did not mean that its development was reversed (Torstendahl 1991).

Outside of Western Europe strong presidential systems have often given rise to a different process of bureaucratization from the one in Western Europe. In the US the president is regarded as the head of the administration as well as the head of politics, while in Western Europe the two spheres have in principle been kept apart. Bureaucratization in Europe has meant that the superiority of politics over administration has been challenged through bureaucratic expansion, while in the US bureaucratization has meant, rather, the increasing influence of federal (public) decisions in the private sector of economy and life, balanced by the equally strong judicial sector. Strong presidential systems have arisen in many other countries in the period after World War II but often without the strong judiciary that exists in the US. Russia is one recent example.

The parallelism of the development of bureaucratization in Europe and North America—and in the twentieth century also Australia and New Zealand—thus became less visible after around 1950. Russia, on the other hand, has not made away with large parts of its bureaucracy from Soviet times after *perestroika* starting in 1985, which might have been expected. The strong presidential rule in new Russia relies heavily on the administrative organs, and these are still intertwined with the leadership of the privatized huge industrial conglomerates. In domestic matters the president is regarded and seems to regard himself more as the leader of an administration than as a political leader and he has no party to rely on. The Duma and the Federation Council (and their predecessors in the previous constitution) have little influence on the government and the administration. Bureaucratic rule is still prevailing, as in Soviet and Czarist times (Shevtsova 1999).

The process of bureaucratization in societies where a distinction is made between politics and bureaucracy has tended to develop in similar forms in different countries. The changes have been going in the same direction at approximately the same points of time. It will then seem that we meet the same type of dynamics and it is tempting to refer this to basic economic structures of these societies. This assumption might be well founded for the West European societies, for they have been fundamentally like each other. It is more difficult to say that it holds also for the US. It is still more difficult to transfer the argument to India, where bureaucratization of both the private and the public sector has been fairly similar to the process in Western Europe in spite of a different type of economy and another basic economic structure.

3. *Bureaucratization as a General Phenomenon and Object of Theory-making*

The interest in the problem of bureaucratization seems to have been directly dependent on the relation between politics and administration. In regard to the interest in bureaucracies surprisingly few books have been devoted to the process of bureaucratization as a general phenomenon (primarily Jacoby 1969, Rizzi 1939).

A wide literature has been produced on the bureaucratization of different parts of modern Western societies. Eisenstadt, with several investigations on bureaucratic developments, gave an overview of the literature in 1958, now somewhat outdated (Eisenstadt 1958). A general approach has been tried in regard to the bureaucratization of socialism (Hodges 1981) but most discussions of bureaucratization have been particular in regard to the selection of societies and time periods discussed.

Max Weber brought forward the basic criteria for a discussion of bureaucratization as a social phenomenon, answering to the need for rationalization in the sphere of government. His standpoint is often rendered as an inevitability of bureaucratization and an impossibility to evade the 'iron cage' of bureaucracy. This hardly makes justice to Weber's sophisticated argument about men's possibility to counteract what seem to be probable predictions (Mommsen 1974, 1989, Schluchter 1989). However, several critics of contemporary culture in post-World-War-II Europe and America have taken bureaucratization as their point of departure for a demand to change a development into a civilization run by administration.

See also: Administration in Organizations; Bureaucracy and Bureaucratization; Civil Service; Democracy, History of; Development and the State; Parliaments, History of; Public Bureaucracies; State Formation; Weber, Max (1864–1920); Weberian Social Thought, History Of; Welfare State

Bibliography

Aberbach J D, Putnam R D, Rockman B A 1981 *Bureaucrats and Politicians in Western Democracies*. Harvard University Press, Cambridge, MA

Baring A 1982 *Im Anfang war Adenauer. Die Entstehung der Kanzlerdemokatie*. DTV, Munich, Germany

Bielenstein H 1986 *The Institutions of Later Han, in Cambridge History of China, Vol. 1: The Ch'in and Han Empires, 221*

BC–AD 220. Cambridge University Press, Cambridge, UK, pp. 491–519
Bruguière M, Clinguart J, Guillaume-Hofnung M, Machelon J-P, Plessis A, Rials S, Thuillier G, Tulard J 1985 *Administration et Contrôle de l'Économie 1800–1914*. Libr. Droz, Geneva, Switzerland
Campbell G A 1955 (new edn. 1965) *The Civil Service in Britain*. Duckworth, London
Charle C 1980 *Les hauts fonctionnaires en France au XIXe siècle*. Gallimard/Julliard, Paris
Eisenstadt S N 1958 Bureaucracy and bureaucratization: a trend report and a bibliography. *Current Sociology* **7**: 2
Gillis J R 1971 *The Prussian Bureaucracy in Crisis, 1840–1860. Origins of an Administrative Ethos*. Stanford University Press, Stanford
Hodges D C 1981 *The Bureaucratization of Socialism*, University of Massachusetts Press, Amherst, MA
Jacoby H 1969 *Die Bürokratisierung der Welt: ein Beitrag zur Problemgeschichte, Soziologische Texte*, Vol. 64, Neuwied [1973, *The Bureaucratization of the World*, University of California Press, Berkeley, CA
Kocka J 1969 *Unternehemensverwaltung und Angestelltenschaft am Beispiel Siemens 1847–1914*. Zum Verhältnis vom Kapitalismus und Bürokratie in der deutschen Industrialisierung, Stuttgart, Germany
Korzhikhina T P 1995 *Sovietskoye gosudarstvo i yego uchrezhdeniya 1917–1991* [*The Soviet State and its Administrations, 1917–1991*]. RGGU, Moscow
Loewe M 1986 *The Structure and Practice of Government, in Cambridge History of China, Vol. 1: The Ch'in and Han Empires, 221 BC–AD 220*. Cambridge University Press, Cambridge, UK, pp. 463–90
Mackenzie W J M, Grove J W 1957 *Central Administration in Britain*. Longmans Green, London
Mammarella G 1985 *L'Italia contemporanea (1943–1985)*. Il Mulino, Bologna, Italy
Mann M 1986/1993 *The Sources of Social Power*. Cambridge University Press, Cambridge, UK, Vols. 1–2
Mayntz R, Scharpf F W 1975 *Policy-making in the German Federal Bureaucracy*. Elsevier, Amsterdam
Michels R 1911 *Zur Soziologie des Parteiwesens in der modernen Demokratie*. W. Klinkhardt, Leipzig, Germany
Mommsen W J 1974 *The Age of Bureaucracy. Perspectives on the Political Sociology of Max Weber*. Blackwell, Oxford, UK
Mommsen W J 1989 *The Political and Social Theory of Max Weber, Collected essays*. Polity Press, Oxford, UK
Morgan K O 1984 *Labour in Power, 1945–1951*. Clarendon Press, Oxford, UK
Peterson C 1979 *Peter the Great's Administrative and Judicial Reforms: Swedish Antecedents and the Process of Reception*. Nordiska bokhandeln, Stockholm [Rättshistoriskt bibliotek, Vol. 29]
Raeff M 1983 *The Well-ordered Police State: Social and Institutional Change Through Law in the Germanies and Russia, 1600–1800*. Yale University, New Haven, CT
Rials S 1977 *Administration et organisation. De l'organisation de la bataille à la bataille de l'organisation dans l'administration française*. Beauchesne, Paris
Rizzi B 1939 *La bureaucratisation du monde*. Paris [1985, *The Bureaucratization of the World*, 1st Am. Ed. Free Press, New York]
Rostovcev M I 1936–38 *A History of the Ancient World*. Clarendon Press, Oxford, UK, Vols. 1–2
Rostovcev M I 1941 *The Social and Economic History of the Hellenistic World*. Oxford University Press, Oxford, UK, Vols. 1–3
Rostovcev M I 1957 *The Social and Economic History of the Roman Empire*. Clarendon Press, Oxford, UK, Vols. 1–2
Rousso H (ed.) 1986 *De Monnet à Massé. Enjeux politiques et objectifs économiques dans le cadre des quatre premiers Plans (1946–65)*. Editions du Centre National de la Recherche Scientifique, Paris
Rousso H (ed.) 1987 *La planification en crises (1965–1985)*, Editions du Centre National de la Recherche Scientifique. Paris
Schluchter W 1989 *Rationalism, Religion, and Domination. A Weberian Perspective*. University of California Press, Berkeley, CA
Shepelyov L E 1999 *Chinovnii mir XVIII–nachalo XX v* [*The World of the Civil Service from the 18th to the Beginning of the 20th Century*]. Iskusstvo-SPB, Saint-Petersburg, Russia
Shevtsova L 1999 *Yeltsin's Russia: Myths and Reality*. Carnegie Endowment for International Peace, Washington, DC
Thuillier G 1980 *Bureaucratie et bureaucrats en France au XIXe siècle*. Droz, Geneva, Switzerland
Torstendahl R 1991 *Bureaucratisation in Northwestern Europe, 1880–1985: Domination and Governance*. Routledge, London
Weber M 1922 *Wirtschaft und Gesellschaft*, Mohr, Tübingen [1978] *Economy and Society*. Bedminster Press, New York, Vols. 1–3
Winkler H A (ed.) 1974 *Organierter Kapitalismus. Voraussetzungen und Anfänge*. Vandenhoeck & Ruprecht, Göttingen, Germany
Yeroshkin N P 1997 *Istoria gosudarstvennykh uchrezhdenii dorevolyutsionnoi Rossii* [*A History of State Administration in Prerevolutionary Russia*]. Tretii Rim, Moscow

R. Torstendahl

Burnout, Psychology of

Burnout is a psychological syndrome that develops in response to chronic emotional and interpersonal job stressors. The three defining components of this syndrome are: (a) exhaustion, (b) cynicism and detachment from the job, and (c) a sense of inefficacy and reduced accomplishment. Burnout impairs both personal and social functioning. While some people may quit the job as a result of burnout, others will stay on, but will only do the bare minimum rather than their very best. This decline in the quality of work and in both physical and psychological health can be costly—not just for the individual worker, but for everyone affected by that person.

1. Development of the Burnout Concept

The relationship that people have with their work, and the difficulties that can arise when that relationship goes awry, have been long recognized as a significant

phenomenon of the modern age. The use of the term 'burnout' for this phenomenon began to appear with some regularity in the 1970s in the USA, especially among people working in the human services. This popular usage was presaged by Greene's 1960 novel, *A Burnt Out Case*, in which a spiritually tormented and disillusioned architect quits his job and withdraws into the African jungle. Even earlier writing, both fictional and nonfictional, described similar phenomena (for example, the protagonist in Mann's *Buddenbrooks* (1900) displays the core features of burnout, including extreme fatigue and the loss of idealism and passion for one's job). What is noteworthy is that the importance of burnout as a social problem was identified by both practitioners and social commentators long before it became a focus of systematic study by researchers.

The evocative power of the burnout term to capture the realities of people's experiences in the workplace is what has made it both important and controversial in the research field. As the 'language of the people,' burnout was more grounded in the complexities of people's relationship to work, and gave new attention to some aspects of it. However, burnout was also derided as nonscholarly 'pop psychology.' Unlike other research on the workplace, which used a top-down approach derived from a scholarly theory, burnout research initially used a bottom-up or 'grass-roots' approach derived from people's workplace experiences. At first, the popular, nonacademic origins of burnout were more of a liability than an advantage. However, given the subsequent development of theoretical models and numerous empirical studies, the issue of research scholarship has now been laid to rest.

Research on burnout has gone through distinct phases of development. In the first, pioneering phase, the work was exploratory and had the goal of articulating the phenomenon of burnout. The initial articles appeared in the mid-1970s in the USA, and their primary contribution was to describe the basic phenomenon, give it the identifying name of 'burnout,' and show that it was not an uncommon response. This early writing was based on the experience of people working in human services and health care—occupations in which the goal is to provide aid and service to people in need. Such occupations are characterized by high levels of emotional and interpersonal stressors, so it is not surprising that the concern with burnout originated there. The initial articles were written by Freudenberger, a psychiatrist working in an alternative health care agency, and by Maslach, a social psychologist who was studying emotions in the workplace. Freudenberger (Freudenberger 1975, Freudenberger and Richelson 1980) provided direct accounts of the process by which he and others experienced emotional depletion and a loss of motivation and commitment, and he labeled it with a term being used colloquially to refer to the effects of chronic drug abuse: 'burnout.' Maslach (1976, 1982) interviewed a wide range of human services workers about the emotional stress of their job, and discovered that the coping strategies had important implications for people's professional identity and job behavior.

The clinical and social psychological perspectives of the initial articles influenced the nature of the first phase of burnout research. On the clinical side, the focus was on symptoms of burnout and on issues of mental health. On the social side, the focus was on the relationship between provider and recipient, and on the situational context of service occupations. Most of this initial research was descriptive and non-empirical—it was more qualitative in nature, using such techniques as interviews, case studies, and onsite observations.

In addition, this first phase was characterized by a strong applied orientation, which reflected the particular set of social, economic, historical, and cultural factors of the 1970s. Both Farber (1983) and Cherniss (1980) provided analyses of how these factors had influenced the professionalization of the human services in the USA, and had made it more difficult for people to find fulfillment and satisfaction in these careers. Burnout workshops became a primary mode of intervention, as exemplified in the work of Pines et al. (1981).

A second phase of burnout research developed in the 1980s, in which the emphasis shifted to more systematic empirical research. This work was more quantitative in nature, using questionnaire and survey methodology, and studying larger subject populations. A particular focus of this research was the assessment of burnout, and several different measures were developed. The scale that has had the strongest psychometric qualities and has been used in over 90 percent of the research studies is the Maslach Burnout Inventory (MBI) developed by Maslach and Jackson (see Maslach et al. 1996 for the most recent versions of this measure).

The shift to greater empiricism was accompanied by theoretical and methodological contributions from the field of industrial–organizational psychology. Burnout was viewed as a form of job stress, with links to such concepts as job satisfaction, organizational commitment, and turnover. The industrial–organizational approach, when combined with the prior work based in clinical and social psychology, generated a richer diversity of perspectives on burnout and strengthened the scholarly base via the use of standardized tools and research designs.

Another aspect of this empirical phase is that burnout extended beyond its original American borders. At first, the phenomenon drew attention in English-speaking countries, such as Canada and the UK. Soon articles, books, and research measures were translated into numerous languages, and subsequently research on burnout emerged in many countries of Europe, as well as Israel. Because burnout research in these countries started after the concept and measures

had been established in the USA, that work built on an already established theoretical and methodological base. Hence, the initial conceptual debate on burnout was less broad, and alternative measures were rarely developed. However, by the 1990s, the intellectual contributions of non-Anglo-Saxon authors in terms of theory, research, and intervention were considerable (most notable is the work of Schaufeli and his colleagues; see Schaufeli et al. 1993, Schaufeli and Enzmann 1998).

In the 1990s this empirical phase continued, but with several new directions. First, the concept of burnout was extended to occupations beyond the human services and education. Second, burnout research was enhanced by more sophisticated methodology and statistical tools. The complex relationships among organizational factors and the three components of burnout led to the use of structural models in much of burnout research (Leiter is the leading contributor here; see Leiter 1993), and also spurred developmental proposals of the merits of sequential versus phase models (for the latter, see Golembiewski and Munzenrider 1988).

A third direction has been the development of new theoretical perspectives on burnout. For example, burnout has been conceptualized as a lack of reciprocity in social exchanges, as emotional contagion, as a failed quest for existential meaning, and as a mismatch between person and job in six areas of worklife (see Schaufeli et al. 1993, Maslach and Leiter 1997). Another new development has been the focus on the opposite of burnout, namely the productive and fulfilling state of engagement with work. Engagement has been defined in terms of the positive end of the three dimensions of burnout (i.e., energy, involvement, and efficacy), but future research may suggest a more complex model of this phenomenon.

2. *A Research Overview*

The current body of research evidence yields a fairly consistent picture of the burnout phenomenon (see Schaufeli and Enzmann 1998 for the most recent review of this literature). Because burnout is a prolonged response to chronic job stressors, it tends to be fairly stable over time. It is an important mediator of the causal link between various job stressors and individual stress outcomes. The exhaustion component of burnout tends to predict the rise of cynicism, while the inefficacy component tends to develop independently.

2.1 *Contributing Factors to Burnout*

The primary antecedents of burnout are work overload (both quantitative and qualitative) and personal conflict at work. A lack of resources to manage job demands also contributes to burnout. The most critical of these resources have been social support among colleagues, and the opportunity to exercise control over one's work. Although the bulk of burnout research has focused on the organizational context in which people work, it has also considered a range of personal qualities. Burnout scores tend to be higher for people who have a less 'hardy' personality or a more external locus of control, or who score as 'neurotic' on the Five-Factor Model of personality. People who exhibit Type A behavior tend to be more prone to the exhaustion dimension of burnout. (See *Personality and Health*)

There are few consistent relationships of burnout with demographic characteristics. Although higher age seems to be associated with lower burnout, it is confounded with both years of experience and with survival bias. The only consistent gender difference is a tendency for men to score slightly higher on cynicism.

2.2 *Outcomes of Burnout*

The exhaustion component of burnout is more predictive of stress-related health outcomes than the other two components. These physiological correlates mirror those found with other indices of prolonged stress. Parallel findings have been found for the link between burnout and various forms of substance abuse.

In terms of mental, as opposed to physical, health, the link with burnout is more complex. It has been assumed that burnout may result in subsequent mental disabilities, and there is some evidence to link burnout with greater anxiety, irritability, and depression. However, an alternative argument is that burnout is itself a form of mental illness, rather than a cause of it. Some of this research has focused on the distinction between burnout and depression: burnout is job-related and situation-specific, as opposed to depression, which is general and context free.

Burnout has been associated with various forms of job withdrawal—absenteeism, intention to leave the job, and actual turnover. However, for people who stay on the job, burnout leads to lower productivity and effectiveness at work. Consequently, it is associated with decreased job satisfaction and a reduced commitment to the job or the organization.

People who are experiencing burnout can have a negative impact on their colleagues, both by causing greater personal conflict and by disrupting job tasks. Thus, burnout can be 'contagious' and perpetuate itself through informal interactions on the job. There is also some evidence that burnout has a negative 'spillover' effect on people's home life.

3. *Implications for Interventions*

The personal and organizational costs of burnout have generated various intervention strategies. Some try to treat burnout after it has occurred, while others focus on how to prevent burnout. The interest in job engagement has led to a focus on how to promote this

positive state (rather than just on how to reduce the negative state of burnout). Although some interventions have been implemented, there has been almost no evaluation of their effectiveness.

The primary emphasis has been on individual strategies to deal with burnout, rather than social or organizational ones, despite the fact that research has found that situational and organizational factors play a bigger role in burnout. Future progress in burnout interventions will depend on the development of strategies that focus on the job context as well as the people who work within it.

See also: Job Stress, Coping with; Occupational Health; Stress at Work; Workplace Safety and Health

Bibliography

Cherniss C 1980 *Professional Burnout in Human Service Organizations*. Praeger, New York

Farber B (ed.) 1983 *Stress and Burnout in the Human Service Professions*. Pergamon, New York

Freudenberger H J 1975 The staff burnout syndrome in alternative institutions. *Psychotherapy: Theory, Research, and Practice* **12**: 72–83

Freudenberger H J, Richelson G 1980 *Burnout: The High Cost of High Achievement*. Doubleday, Garden City, NY

Golembiewski R T, Munzenrider R F 1988 *Phases of Burnout: Developments in Concepts and Applications*. Praeger, New York

Leiter M P 1993 *Burnout as a developmental process: consideration of models*. In: Schaufeli W B, Maslach C, Marek T (eds.) *Professional Burnout: Recent Developments in Theory and Research*. Taylor and Francis, Washington, DC, pp. 237–50

Maslach C 1976 Burned-out. *Human Behavior* **5**: 16–22

Maslach C 1982 *Burnout: The Cost of Caring*. Prentice-Hall, Englewood Cliffs, NJ

Maslach C, Jackson S E, Leiter M P 1996 *Maslach Burnout Inventory Manual*, 3rd edn. Consulting Psychologists Press, Palo Alto, CA

Maslach C, Leiter M P 1997 *The Truth About Burnout*. 1st edn; Jossey-Bass, San Francisco, CA

Pines A, Aronson E, Kafry D 1981 *Burnout: From Tedium to Personal Growth*. Free Press, New York

Schaufeli W, Enzmann D 1998 *The Burnout Companion to Study and Practice: A Critical Analysis*. Taylor and Francis, Philadelphia, PA

Schaufeli W B, Maslach C, Marek T (eds.) 1993 *Professional Burnout: Recent Developments in Theory and Research*. Taylor and Francis, Washington, DC

C. Maslach

Business and Society: Social Accounting

To succeed, any modern business manager needs to understand all the interfaces between the business and wider society no less than all the internal operations of the business itself. For no business can survive, let alone flourish, unless it provides goods or services for which customers outside the business will be prepared to pay. Business success depends on meeting market demand. Awareness of market conditions and of how the market may be influenced is a key part of business management. In this article the various interfaces between businesses and their social context are briefly examined, using a form of analysis which has come to be known by the acronym SLEPT, in which the different letters stand respectively for the Social, Legal, Environmental, Political, and Technological factors influencing business organizations.

1. Social Factors

Market demand depends on the needs, desires, and interests of people in society, and knowledge of these is essential if a business is successfully to meet that demand. Such factors may be determined by the physical climate, by the language and culture, or by the demographic structure of a society. There is likely to be little demand, for example, for heavy overcoats in tropical countries, or for umbrellas in arid lands. In some cultural settings it is still rare for women to seek paid employment. In others the increase in life expectancy may have resulted in increasing demand for leisure activities for pensioners who have retired from employment. Changes to birth rates cause fluctuations in demand for goods and services required by children. These are examples of the ways in which knowledge of the market place will help a business entrepreneur to decide what goods to produce and offer for sale to meet demand.

Part of the cultural context of a business is the prevailing ethos or climate of moral opinion about what are right standards for business activity and trading. During the last two decades, such questions of business ethics have figured increasingly in public concern and on the agendas of business organizations. These highlight the need for complete probity and accountability in business, and for more openness in the operations of boards of directors. Ethical considerations may involve how a firm exploits its markets, how it obtains raw materials or cheap labour for its operations in different parts of the world, or how its activities affect the physical environment (which will be explored more fully in Sect. 3).

The ethical standing of a business will affect its public relations, which in turn may impact upon the demand for its products, one way or the other. Modern businesses, especially larger ones, pay increasing attention to their public relations and how their images—and the brand images of their products—appear to the general public. Part of that public image is determined by how a business is presented by the news media, part by the reputation which products have among customers, and part by the business' own public advertising of its products. No successful

business can ignore its need for a continual campaign to advertise its products, for in this way it can significantly influence market demand and maintain its share of markets with strong competition. The outlets for advertising are constantly spreading, using newspapers, magazines and journals, individual mail-shots, public hoardings, television and radio, and more recently, the Internet. This leads to the whole field of electronic commerce to be considered more fully in Sect. 5. At the same time, businesses need to be conducting market research, assessing public opinion about their existing products and exploring what other demands there are in the marketplace for potential new products. Market research enables a business to maintain direct links with any community where it seeks to trade.

The social factors relating to a business company have been recognized in many circles by introducing the concept of stakeholders of the company, who are all the groups of people with a legitimate interest in the operations of the company, not just the shareholders. Stakeholders equally include the employees of a business, the customers, local residents and communities, and representatives of environmental interests. For many years there have been schemes of co-ownership and copartnership which have involved employees in the actual ownership of companies. More recently, and particularly in the United States, some corporations have introduced special constitutional features so that there are boards with representation from all stakeholder groups. A growing number of large firms are now publishing regular social or environmental reports as a means of accounting to all their stakeholders; and independent professional bodies are now issuing reviews of whole industries in this way, such as the Turnbull Report of the Association of Chartered Accountants of England and Wales. These arrangements vary in their levels of formality, but they are all attempts to involve wider social interests in the decision-making processes of the company. The growth of consumerism and of specific organizations established to protect and further consumer interests has inevitably brought increasing pressures to bear on the decisions of many businesses.

2. *Legal Factors*

The framework for business activity in any society may be constrained by legal requirements in that society. As long ago as 1570, the English Parliament enacted a law requiring every member of the 'lower classes' over the age of six years to wear a woollen cap on Sundays and holidays. Such legislation would today be regarded as a gross infringement of personal freedom, but it was introduced to further the interests of the politically powerful sheep farmers and wool traders at the time.

In later centuries in Britain there was significant dismantling of such legal constraints on business, with the ending of the mercantile system and the growth of '*laissez faire*.' Free market economists like Adam Smith, Thomas Malthus, David Ricardo, and John Stuart Mill advocated the freeing of business markets from legal restraint and control, so that the 'invisible hand' of free market forces could determine the price levels at which supply would equal demand. It was maintained that the interests of both producers and consumers were optimised by the free play of market forces, and that this required the absence of legal constraints, to enable businesses best to serve society by pursuing their own interests as vigorously as possible.

During the last 180 years, however, the political pendulum has swung very much the other way. Acts of Parliament in Britain ended the slave trade, banned the employment of women and children in coal mines, and introduced legislation for the health and safety of employees in factories and other places of work. Other statutes provided for the registration of business companies, giving them legal entity as corporate bodies rather than partnerships of all their various owners, and limiting the financial liability of shareholders to what they paid for their shares. Increasingly under governments of differing political shades, more and more legislation has come into force in virtually all countries of the world which sets the standards and requirements to be observed in business trading. Business managers need a very full awareness of the legal framework in which they are operating, both of the statutes established by political processes and of the decisions of courts in applying those statutes and developing case law.

3. *Environmental Factors*

Concern for the health and safety of employees is one example of a limitation on business operations to curb the potentially harmful effects of those operations. Just as people within the workplace may be injured and exposed to health risk, so too the activities of a business, in both production and distribution, may be detrimental to people living in the vicinity, along transport routes, and in the community generally. The discharge of waste materials from industrial processes, be they solid, liquid, or gaseous, on to waste land, into rivers, or into the atmosphere can have harmful effects reaching far into other countries, and into the ozone layer above the atmosphere. There is global concern that raw materials are being depleted at rates which will restrict or even exhaust their availability for future generations. For instance, the rain forests of Brazil are being decimated by the worldwide demand for timber and paper. Similarly, known reserves of fossil fuels in the world are finite. Public opinion is increasingly aware of the environmental effects of business activities in the widest contexts and longest timescales, and there is growing demand that all economic activity

should be environmentally sustainable and without permanent damage to the planet.

Most goods supplied to the market today are packaged in some form, and difficulties may arise in the disposal of the packaging. Cardboard, paper, bottles, and some metallic containers can now be recycled for reuse, but containers made of some plastics have no future use. Manufacturers therefore need to be more aware of the environmental impact of the way they package their products. And in this age of the throw-away society, there are immense problems arising from the disposal of goods which are no longer of use to their owners. It is not easy for society to get rid of motor cars and household applicances which have outlived their useful lives, and there are limits to how the world can dispose of its growing mountains of rubbish. All such considerations challenge business operators to develop products which are easier to recycle or get rid of, or which will last much longer. Political pressures on governments will almost certainly lead to more legal constraints on business to meet these challenges.

4. *Political Factors*

Section 2 highlighted the ways in which business activities can be constrained by laws in force in society, and in most cases those laws are made, amended, or repealed through the established political processes of the country in question. Political structures vary between countries, and any business manager needs to understand the working of those structures in any country where the business has operations, in securing raw materials, in production, or in marketing and distribution. In countries which have democratic structures, a business needs to be constantly alert to what changes may come about in public policies. The programs of potential alternative governments need to be scrutinized and possibly influenced through the use of open democratic processes, so that when new laws are enacted they do not take the business by surprise and may be more helpful for the business than would otherwise have been the case.

Equally important is an understanding of the administrative processes of government through which laws and statutes are implemented. A business must be especially alert to the taxation implications of all its activities, and be ready to rearrange and retime those activities to reduce its tax liabilities. Similarly, the day-to-day administration of other departments and agencies of government may bear on the activities of a business, and it is essential for the manager to have an easy means of communication with appropriate public servants in order to sort out any problems which may crop up for the business. The best means of cooperation in this respect will vary from country to country, but one general principle is that all representations to officials must be open and never underhand nor illegal.

5. *Technological Factors*

In Sect. 1 it was noted that technological changes provide another important influence on how a business is run, introducing new products to be manufactured, new methods of manufacturing them, new means of packaging and distribution, new methods of advertising and marketing, and new means of communication to be employed in all operations. The pace of technological change appears to be accelerating all the time, and the contemporary electronic revolution in communicating information is introducing a new way of business life for all large organizations. The mobile phone, the fax, and email enable decisions to be taken and implemented instantly.

Technological changes can bring vast benefits to producers and consumers alike, but they may also pose difficult problems, particularly in the management of human resources. As old machinery becomes obsolete, staff need to be retrained to use new technology or redeployed to new activities because the new technology is not so labor-intensive. Many new machines take the place of people, so it is not unusual for established workforces to resist the introduction of new technology, and industrial relations may become fraught. Nowhere was this shown more dramatically than in the English newspaper industry in the 1980s, when decisions to introduce new technology for print-setting led to mass picketing and some violence at Wapping in east London.

At the same time, the modern revolution in electronic communications enables many more people to work at home in more conducive surroundings and avoiding the traffic congestion associated with commuting to work. Word processors, fax machines, and emails enable the workforces of many businesses to be geographically scattered, with many individuals working on a contract or consultancy basis on tasks which had previously to be done in central offices. Add to all this new technology the far-reaching potential of the Internet, and it becomes possible for more and more business to be conducted electronically, with consumers ordering, and paying for, goods on the Internet; the effectiveness of such commerce depends partly on the strength and security of computer systems, and partly on the reliability of networks used for the prompt distribution of goods.

6. *Social Accounting*

Whereas the foregoing description of the interfaces between business and society was in terms of the links between an individual business and its social context, social accounting is more concerned with macroeconomics, involving a study of how whole industries and institutions relate to the wider economy and society, and of how the production of goods and services, income formation, and income distribution relate to each other. Social accounting techniques have

been widely applied in the fast developing economies of the third world, such as Iran, Malaysia, Sri Lanka, Pakistan, Indonesia, Ecuador, and Swaziland.

In the application of these techniques, Social Accounting Models, or SAMs, can be constructed for specific countries and used as integrated databases for understanding the effects of national economic policies. While the British Treasury has used Economic Models for over a quarter of a century, in most of the less developed countries similar SAMs have now been built, varying greatly according to the economic characteristics of local economic conditions. Such models may be diagrammatically represented as shown in the literature, and are increasingly used in the formation of public economic policy.

Social accounting concerns microeconomics as well, since a growing number of corporations are now active in social accounting, auditing and reporting. They publish social and environmental reports on a yearly basis, which provide detailed information about the relations with their stakeholders, such as investors, employees, consumers, suppliers, central and local governments, and environmental associations. Financial rating agencies evaluate firms not only in terms of their economic and financial performance, but also in terms of their social, ecological and ethical performance.

See also: Advertising and Advertisements; Advertising: General; Business Law; Environmental Risk and Hazards; Market Research; Marketing Strategies; Technology and Social Control

Bibliography

Alarcon J, van Heemst J, Keuning S, de Ruijter W, Vos R 1990 *The Social Accounting Framework for Development*. Avebury, Aldershot, UK

Jefkins F 1987 *Public Relations for your Business*. Mercury Books, London

Marshall E 1993 *Business and Society*. Routledge, London

Pearce D 1991 *Corporate Responsibility and the Environment*. British Gas plc.

E. I. Marshall

Business History

Strictly speaking, Business History (company history) as the history of private and/or public corporations with specific legal structures must be distinguished from Entrepreneurial History (business biography), the history of entrepreneurially active personalities. In the first instance, the concern is with the development of organizations in which questions of economic success loom paramount. In the second case, the object of study is a certain social type and/or individual personality whose behavior inside and outside of the enterprise can be researched, for example, in its social, cultural, or political effects. In practice, however, institutional and biographical approaches tend to overlap. This article starts out by sketching the history of the subject, above all in the USA and in Europe, before then going on to a discussion of contemporary methods and approaches.

1. History of the Academic Discipline

Modern business and/or entrepreneurial history has two sources: First, the need of corporations and businessmen for image cultivation and self-affirmation, which has brought forth an increasing number of commissioned biographies and company histories, as well as autobiographical accounts. Business History received academic impetus above all from the German Historical School of Economics. This school questioned the explanatory power of classic liberalism and attempted to replace it through alternative concepts in which long-term perspectives as well as the entrepreneur played a key role in explaining economic growth. Of primary significance was the work of the German scholars Gustav Schmoller, Werner Sombart, and Lujo Brentano.

Against this theoretical background, scholars around 1900 began to write academic business histories, thus establishing a counterweight to the celebratory usage to which these chronicles had been put in the last third of the nineteenth century initially by banks and publishing houses, soon thereafter by the large industrial firms. Both genres were reacting to the mounting criticism of the conduct and profits of the more successful businessmen. In such a way, business history served, among other things, to legitimize a rising social class that was rapidly acquiring both affluence and influence. In 1912 Richard Ehrenberg, whose *History of the Siemens Brothers* (1906) was one of the first company histories based on archival documents, collaborated with Hugo Racine on *The Worker Families of Krupp*, which defended the company politics of the Ruhr's heavy industry sector. Also during this period was the founding of the first systematically organized business archives (Krupp 1905, Siemens 1907. In Great Britain in 1904, George Unwin published his pioneering work on the origins of the modern enterprise, which traced the decline of the guilds through to the emergence of functionally differentiated management structures in the seventeenth century.

Academic approaches to business history received a further impulse from the History of Technology and the burgeoning field of 'management studies' which specialized in private enterprise. Here historical methods were combined with the self-promotional interests of new academic disciplines. Both fields of study wanted not only to secure their academic standing, but also harboured naive hopes—from a modern perspec-

tive—of being able to derive general laws of historical development.

In nineteenth-century America there appeared numerous non-academic business histories in which banks, canal and railroad companies were the leading players. The subject first established itself academically in the 1920s with the Graduate School of Business Administration of Harvard University, which sought to use Business History for training purposes. Future managers were to be trained through an analysis of historical case studies and through a comparison of past and current methods of executive leadership. With the appointment of Norman Gras to the first ever chair for Business History in 1927, Harvard laid the groundwork for the institutionalization of American business history. There subsequently appeared such periodicals as the *Journal of Economic and Business History*, 1928–32; *Bulletin of the Business Historical Society*, 1938–53; and *Business History Review*, since 1954. In his magnum opus *Business and Capitalism*, inspired by the work of Sombart, Ehrenberg, and Unwin, Gras (1939) divided the economic history of America into various stages, naming them for each period's distinctive business type: petty, mercantile, industrial, financial, and national capitalism. The *Casebook in American Business History*, which Gras co-authored with his Harvard colleague Henrietta M. Larson, also appeared in 1939. In 1948 Larson published an extensive bibliography in her *Guide to Business History*. Already the year before, with the registration of the *Business History Foundation*, Gras and Larson had gone a step further down the road to firmly establishing their discipline. In 1954 the *Business History Conference* was founded, which has remained the subject's central American organization to the present day.

The subject received an additional stimulus through the favourable reception of Joseph Schumpeter's ideas, which stressed the dynamic innovations of entrepreneurs as the driving force of economic development. Following his programmatic address to the Economic History Association, Schumpeter himself helped in establishing the Research Center in Entrepreneurial History (1948–58), funded by the Rockefeller Foundation and the Carnegie Corporation. Its first director Arthur H. Cole, along with Thomas C. Cochran, Leland H. Jenks and the German economist Fritz Redlich, grouped their work around the concept of 'entrepreneurship,' and anchored it in a broader social context. The Center also published its own journal, *Explorations in Entrepreneurial History*. In 1959, after 10 years, Cole summarized the Center's most significant findings, which went far beyond those of mere biographical studies. The main emphasis was on the development of entrepreneurial typologies, as well as on the relation of the entrepreneur to the corporation and its environment.

As of the 1940s, more and more American universities, colleges, and business schools introduced courses and chairs in Business History. In Germany, the discipline has been unable to achieve a similar rank within the academic canon, and is primarily taught by history and economics faculties under the rubric of economic and social history.

Appearing in the mid-1950s were many periodicals that reflected a growing interest in the subject, such as *Business History* (1956, UK), *Tradition* (1956, West Germany; since 1978 *Zeitschrift für Unternehmensgeschichte*), *Technology and Culture* (1958, USA), and *Histoire des Entreprises* (1958, France). Ironically, *Business History* remained a force even when its object of study was undergoing severe periods of crisis. In the wake of the student movements of the 1960s, entrepreneurs and companies were subjected to harsh criticism. Then in the 1970s the pioneer countries of the industrial age lapsed into a structural crisis that within a few years had shrunk those long dominant sectors of mining, iron, steel, and textiles. These upheavals led to the founding of numerous museums and business archives that have since become lively research sites. Added to this were newly established academic organizations such as Germany's *Gesellschaft für Unternehmensgeschichte* (Society for Business History, est. 1976).

At the same time, *Business History* began to subdivide, there emerging special organizations and journals for individual aspects of the field, sporting names such as *Accounting*: *Business and Financial History* (1991), *Financial History Review* (1994) and *Journal of Industrial History* (1998). By contrast, the successor of *Business and Economic History Review* (1962–99), *Enterprise and Society* (2000), began to pose questions whose impulse came from the broader fields of social and cultural history. And since 1998, the *European Yearbook of Business History* has likewise represented a wide thematic outlook.

Due to the regional embeddedness, further initiatives arose that not infrequently culminated in the founding of museums and archives. Prominent examples are the Centre for Business History in Scotland (Glasgow), the Hagley Museum and its affiliated Center for the History of Business, Technology, and Society in Wilmington, Delaware (USA), and the Society for Westphalian Economic History in Germany's Ruhr district. In the 1990s, various international organizations were founded, such as the European Association for Banking History (1990), the European Business History Association (1994), and the Society for European Business History (1997).

2. *Methods and Approaches of Business History*

From the moment it entered academia, *Business History* began to borrow methodologies from neighbouring disciplines, particularly economics and management studies, but also from history, sociology, political science, and ethnology. Equally variegated are the research areas and methodologies of modern

business history. Roughly speaking, one can distinguish nine new areas of study:

2.1 The Enterprise as a Commercial Endeavour

In an analysis of the commercial success of firms, the instruments of management studies are of primary importance. Costs, investments, yield, cash-flow, and profit margins are studied over prolonged periods. Company development is essentially traced with the help of in-house accounting files—themselves objects of research, as the methods of accounting have been subject to historical change. The external relations of the enterprise are only of marginal interest, or are treated indirectly. However, in the same way as trade cycles, fiscal, legal, or political intervention show up in the final balance sheet.

2.2 Finance Capital and Industrial Production

Stimulated chiefly by Rudolf Hilferding's thesis of the dominance of big banks over industry, the relationship between the financial and the industrial sectors is likewise an object of analysis. For European and American industrialization, it has been revealed that banks and money markets at first played only a subordinate role in the financing of business. However, when these firms reached certain growth thresholds—for example, merging with or investing in capital-intensive technologies—external funding was required, in which way they would enter into a close relationship to the banks.

Of Great Britain it is said that between 1880 and 1930 London financial institutions neglected industrial interests, thus hindering the modernization and concentration of the country's industrial base. This was in contrast to the big banks of Paris and Berlin. On the other hand, Collins sees the aloofness of the money market to industry as a consequence rather than a cause of slow corporate growth. Davis argues that the American money market circa 1900 was less efficiently organized than its British counterpart, and the selective access to credits explains the concentration of capital in the hands of multi-millionaires like Rockefeller and Carnegie. By comparison, Chandler emphasizes the dominance of internally-generated cash-flow as the result of successful integration. Case studies on the close relationship between big banks and industry in Germany show that one cannot speak of a general ascendancy of financial institutions.

2.3 Businesses as Innovators and Users of Technology

Technological innovations, either self-generated or initiated from outside the firm, are central to the development of many enterprises. Thus, *Business History* undertook the description of technical processes; however, these were often uncritical lists of glorious inventions and paeans to the genius of the particular entrepreneur. The new *Business History* discusses questions as to the timely application of technical innovations or their adequacy for specific conditions of production and sales. In the debate on the decline of that pioneer of the Industrial Revolution, Great Britain, comparative investigations have been particularly enlightening. While some researchers brand the British businessman's clinging to outdated technology as a failing, others see this as perfectly rational behavior in view of fundamental differences in the market as well as in the conditions of supplies and sales (Wengenroth 1986).

The mass production technology that had been suitable for the large and relatively homogenous American market and a predominantly unskilled labor force was non-transferable—at least with the same level of success—to the completely different economic terrain of Europe. It is also not always an advantage to be in the forefront of technological innovation. Wait and see—then imitate. Later investments in more mature technologies has proven itself in many ways a superior strategy, especially for countries lagging behind the pioneers of industrial development.

Further significant themes in this field are: mechanisms of technology transfer, research and development, the relationship of companies and academic research, as well as the social, economic, and mounting ecological consequences of technical progress (Hounshell 1984).

2.4 Organizational Capability and Comparative Advantage

Belonging to the central themes of *Business History* is the quest for the optimal organizational structure with respect to strategy, technology, and the market. It were along these very lines that Alfred D. Chandler of the Harvard Business School forwarded theses that were to preoccupy entire generations of scholars. In Chandler's first major work, *Strategy and Structure* (1962), he asserted that changes in the 'market-cum-technological environment' modify the 'basic long-term goals and objectives of an enterprise,' which themselves determine the shape of the company's organization. It was Chandler's (1977) *The Visible Hand* which stressed the superior performance of managers i.e., salaried entrepreneurs, in comparison to the owner-entrepreneuer. His third notable work, *Scale and Scope*, extended the study to Great Britain and Germany, concluding that differences in national competitiveness were the result of divergent utilization of economies of scale. In employing this approach, Chandler elevated the American model of managerial capitalism and its large-scale 'three-pronged investment in production, management, and marketing facilities' to a universal prescription for success.

Chandler's theses met with enthusiastic response. He had moved beyond the stage of individual case

studies in favour of a general paradigm for the analysis of industrial capitalism. At the same time, his claim to general validity provoked vehement protest. The ensuring debate was unusually productive, and issued most significantly in a criticism of Chandler's idealization of the American model of big industry. 'The usual Chandler bracketings apply. Labor, culture, state, politics, and all industrial activity outside the top 200 are set aside as secondary or irrelevant.' (Scranton 1991)

The latter point was especially raised by Michaele Piore and Charles Sabel who pointed to the viability of 'Historical Alternatives to Mass Production.' According to their 'theory of dualism,' mass production is not the only paradigm of industrial capitalism. Small-scale enterprises, always representing the majority of companies, were also able to operate successfully. The most efficient were those specializing in small market segments and using flexible production technologies as well as highly skilled labor. Fluctuations in demand and design could be easily coped with. In a way, this 'modern craft production' was a 'necessary complement to mass production' and explains the persistence of the small firm. Philip Scranton and Gary Herrigel have demonstrated for the USA and Germany respectively that multiple paths to industrial profit did coexist and that successful specialty production of small and medium-sized firms hinged on quality, differentiation, human skills, and local networks rather than on high throughput technologies and size.

Associated with this debate are questions concerning the optimal size of firms, the reasons for and success of mergers and acquisitions, the relation of transaction costs and corporate structure, as well as solutions to the principle-agent problem.

2.5 Multinationals in the Process of Internationalization and Globalization

Closely connected to the question of the appropriate organizational structures is the internationalization process within firms (Wilkins 1970, Jones 1996). Why do companies prefer foreign direct investments with direct control as opposed to portfolio investments or cooperation with independent foreign partners? Firms like Nobel and Vickers invested abroad in order to exploit those competitive advantages already acquired at home. By contrast, Shell and Lever secured access to vital raw materials. Others were desirous of skirting customs barriers, of operating more closely to the market, or securing foreign know-how, locational advantages, and subsidies.

Buckley and Casson maintain that excessive transaction costs are the main reason for substituting cross-border market cooperation through multinational hierarchical organization. Internalization is a superior method of coordination across national frontiers.

In addition, the detailed structures of these firms are examined i.e., the choice between sales and production subsidiaries or the emergence of particular types of firms like the Free Standing Company, which is characterized by a small corporate headquarter in the country of origin and the location of all productive assets abroad. Related hybrid structures such as joint ventures, cartels, and strategic alliances are also analyzed with the use of similar questions.

It is necessary to distinguish between internalization and globalization. Whereas international firms mainly operate from a national basis, global enterprises are difficult to locate precisely, distributing their activities over the entire globe both fiscally and in terms of market strategy and costs.

2.6 Firms as a Venue of Social Interaction

The history of work relations was given particular impetus by the classic school of social history. This school discovered that in the process of industrialization, many entrepreneuers rose from related occupations from which they derived important social and financial capital, as well as technological and/or organizational know-how. Their degree of social mobility has long been overestimated. Social history further asks how owners of capital, managers, employees and workers, and men and women organize their social relations within firms. In what ways are individual interests bound together organizationally, and how do they factor into the bargaining process? Thus the focal points of research are the history of unions, of worker participation of company social policy, and methods of in-house company relations.

Other objects of study are the mechanisms with which top management uses its power and control in the implementation of directives. In these, how is the principal-agent problem solved? That is, what way is the self-interest of manager-employees curbed with regard to the goals of the entire firm and the capital owners? Chandler, however, sees the rise of better educated managers bereft of family interests as one of the positive and decisive competitive advantages of the American model. He sees the main reason for the relative decline of British industry in its insistence on 'personal capitalism.' Numerous critics have warned against generalized statements concerning the different potential of managers and owner-entrepreneurs, respectively. There is little doubt, however, that outside of the USA the emergence of large company structures frequently coincided with the rise of managers who had no familial connection to the capital owners.

2.7 Corporations as Cultural Institutions

Firms are bearers of specific corporate cultures that play a key role in internal and external communication. As concerns the internal structure, it is a question of the 'software of organization,' of shared rules and values. While Chandler's work suggests that the formal structures—the 'hard facts of organi-

zation'—are decisive, proponents of cultural business history assert that internal integration and coordination is essentially dependent on the strength of the given corporate culture. This assures loyalty, team spirit, and communication skills, while also lowering transaction costs. It consolidates the results of long-term learning processes, especially firm-specific routines and tacit knowledge.

Corporate culture is frequently influenced by strong individuals, chiefly company founders, but also by external socio-cultural factors. Organizational traditions like the German bureaucracy (Kocka 1978), those of the military, or agrarian paternalism were especially effective in the industrialization process. Following mergers and takeovers, there is often a clash of cultures and incompatibility between differing value systems. Furthermore, it is precisely the strong corporate cultures that have a tendency to develop moments of inertia during which adaptation to external change is impeded. The research has shown that organically-evolved corporate cultures determine the inner quality of a company in essential ways. However, social engineering—i.e., the arbitrary, short-term modification of corporate cultures—is impossible.

Corporate culture is not only effective as an internal link of identity and cohesion, but also serves as an exogenous label and a seal of quality. It determines the firm's image in the customer's mind and in that of the broader public, as well as defining the self-projection of the enterprise and its advertised products. The image-forming power of corporate culture extends to brands that rationalize the communication process with customers, while also implying an adherence to quality standards. The strong trust of such arrangements is central to all forms of marketing, but can also be transformed into its diametrical opposite when that trust is abused.

2.8 Companies and Entrepreneurs as Political Actors

Modern companies operate within a regulated framework. The state establishes legal forms that superintend ownership laws and property rights, responsibilities, and information duties. Aside from company law, also of prime importance are labor law, wage law, environmental law, trade law, and tax law. Therefore it is in the vested interest of the firm to influence the development of regulatory frameworks and to involve itself in the political process. The range of its dealings extends from simple forms of lobbying to the participation in *coup d'états*.

Central themes are the effects of organized interests, the activity of companies in political parties, parliaments, administration, governments, the thin line between fund-raising activities and corruption, private-public partnerships, the sale of products to the state, the nationalization or privatization of enterprises, and the correlation between business goals and politics in general.

The view of the political role and ethical responsibility of the economy has gained in significance through research on the behavior of German and foreign businessmen in relation to the National Socialist regime. In reaction to the decades-long taboo placed on studies of the Nazi past, numerous works have now emerged rejecting ideologically-motivated generalizations and accusations, while at the same time confirming the deep involvement of many businessmen in National Socialism.

2.9 The Biographical Approach (Entrepreneurial History)

Entrepreneurial History is concerned with the people behind the strategic decisions. What socio-cultural variables determined their world view and thereby their business personality? Such questions were at first chiefly posed by biographical investigations that frequently sought to portray their subjects in heroic terms. Simultaneously, there was the counter-tradition of critical and even defamatory entrepreneurial biography.

Since the 1970s, social history has introduced quantifying methods into entrepreneurial biography. The great advantage of collective-biographical studies consists in the ability to distinguish between exceptional facts and general structures. Drawing upon selective and representative samples, they investigate heritage, education and upbringing, religion, marital and social status, financial circumstances, the chances of ennoblement and political influence (Augustine 1994, Berghoff and Möller 1994, Daumard 1987). One of the most heatedly discussed questions is whether European businessmen became gentrified by imitating the aristocracy and abandoning their own identity, a debate that has led to an overwhelming rejection of this thesis.

Emerging parallel to the wave of collective-biographical works have been several excellent academic compilations with a wealth of individual-biographical items (Jeremy 1984). With the current wave of historiography addressing everyday life and mentalités, individual-biographical methods are currently on the rise. Vivid and atmospheric narratives now complement more representative statements on general structures which are analyzed through statistical methods.

3. Conclusion

On the whole, despite its varying worldwide status, *Business History* is experiencing an unmistakable upsurge and continues to evolve and expand. As before, there is the ever present danger of a particularistic narrowing of the subject to a series of individual case studies often composed with an antiquarian mind-

set, and which enlarge on their theme with a wealth of material but with little regard for historical context. This unhappy state of affairs—typical of many commissioned studies—can only be countered through a determined commitment to macrohistorical and macroeconomic findings and the inclusion of theoretical approaches from related disciplines.

See also: Business and Society: Social Accounting; Business Law; Corporate Culture; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; Ethical Codes, Professional: Business Codes; Policy Process: Business Participation; Poverty, Culture of; Technological Innovation

Bibliography

Augustine D L 1994 *Patricians and Parvenus. Wealth and High Society in Wilhelmine Germany*. Berg, Oxford, UK

Berghoff H, Möller R 1994 Tired pioneers and dynamic newcomers? A comparative essay on German and British entrepreneurial history. *Economic History Review* **47**: 262–87

Brentano L 1907 *Der Unternehmer*. Simion, Berlin, Germany

Buckley P J, Casson M 1976 *The Future of the Multinational Enterprise*. Macmillan, London

Chandler A D 1962 *Strategy and Structure, Chapters in the History of the Industrial Enterprise*. MIT Press, Cambridge, MA

Chandler A D 1977 *The Visible Hand. The Managerial Revolution in American Business*. Belknap Press, Cambridge, MA

Chandler A D 1990 *Scale and Scope. The Dynamics of Industrial Capitalism*. Harvard University Press, Cambridge, MA

Cole A H 1959 *Business Enterprise in its Social Setting*. Harvard University Press, Cambridge, MA

Collins M 1995 *Banks and Industrial Finance in Britain 1800–1939*. CUP, Cambridge, UK

Daumard A 1987 *Les Bourgeois et al Bourgeoisie en France Depuis 1815*. Aubier-Montaigne, Paris

Davis L E 1966 Capital markets and industrial concentration: The US and UK, a comparative study. *Economic History Review* **19**: 255–72

Dunning J H 1993 *The Globalization of Business: The Challenge of the 1990s*. Routledge, London

Ehrenberg R 1902/1905 *Grosse Vermögen, Ihre Entstehung und ihre Bedeutung*. Fischer, Jena, Germany, 2 Vols.

Goodall F, Gourvish T, Tolliday S (eds.) 1997 *International Bibliography of Business History*. Routledge, London

Hayes P 1987 *Industry and Ideology: IG Farben in the Nazi Era*. CUP, Cambridge, UK

Herrigel G 1996 *Industrial Constructions. The Sources of German Industrial Power*. CUP, Cambridge, UK

Hounshell D A 1984 *From the American System to Mass Production, 1800–1932. The Development of Manufacturing Technology in the United States*. Johns Hopkins University Press, Baltimore

Jeremy D J (ed.) 1984 *Dictionary of Business Biography*. Butterworth, London, 5 Vols.

Jeremy D J 1998 *A Business History of Britain 1900–1990s*. OUP, Oxford, UK

Kocka J 1978 Entrepreneurs and managers in German industrialization. In: Mathias P, Postan M M (eds.) *The Cambridge Economic History of Europe*. CUP, Cambridge, UK, Vol. 7, pt. 1

Piore M, Sabel C F 1984 *The Second Industrial Divide: Possibilities for Prosperity*. Basic Books, New York

Pollard S 1965 *The Genesis of Modern Management. A Study of the Industrial Revolution in Great Britain*. Edward Arnold, London

Redlich F 1964 *Der Unternehmer, Wirtschafts—und Sozialgeschichtliche Studien*. Vandenhoeck, Göttingen, Germany

Schumpeter J A 1911 *Theorie der wirtschaftlichen Entwicklung. Eine Untersuchung über Unternehmergewinn, Kapital, Kredit, Zins und den Konjunkturzyklus*. Duncker & Humblot, Munich, Germany

Scranton P 1991 Review of scale and scope. *Technology and Culture* **32**: 1102–4

Scranton P 1998 *Endless Novelty: Speciality Production and American Industrialization, 1865–1925*. Princeton University Press, Princeton, IN

Sombart W 1913 *Der Bourgeois. Zur Geistesgeschichte des modernen Wirtschaftsmenschen*. Duncker & Humblot, Munich, Germany

Unwin G 1904 *Industrial Organization in the Sixteenth and Seventeenth Centuries*. Clarendon Press, Oxford, UK

Wengenroth U 1986 *Unternehmensstrategien und technischer Fortschritt. Die deutsche und die britische Stahlindustrie 1865–1895*. Vandenhoeck & Ruprecht, Göttingen, Germany

Wilkins M 1970 *The Emergence of Multinational Enterprise: American Business Abroad from the Colonial Era to 1914*. Harvard University Press, Cambridge, MA

Wilson J F 1995 *British Business History, 1720–1994*. Manchester University Press, Manchester, UK

H. Berghoff

Business Law

1. Contracts

1.1 Elements

A contract is a legally enforceable promise. Formation of a valid contract generally requires four basic elements. First, there must be an agreement between the parties formed by an offer and acceptance of that offer. Second, the parties' promises must be supported by something of value, known as consideration. Third, both parties must have the capacity to enter into a contract. Fourth, the contract must have a purpose that is legal.

1.2 Intent to Be Bound

As business transactions have grown more complex, courts have wrestled with the issue of when preliminary negotiations have ripened into a contract (Farnsworth 1987). Perhaps the leading case arose out of Pennzoil Company's failed attempt to merge with Getty Oil Company (*Texaco, Inc. v. Pennzoil Co.* 1987). The court held that a five-page memorandum of

agreement, signed by the majority shareholders of Getty Oil and approved (but never signed) by the board, constituted a binding contract under New York law even though a definitive agreement for the $5 billion transaction was never signed.

In Japan, businesspersons are more likely to rely on oral contracts or on short contracts with a broad catch-all provision stating that the parties will consult in good faith to resolve any questions of interpretation or matters not covered by the contract (Yanagida et al. 1994). As a result, Japanese courts are more likely than their US counterparts to conclude that negotiations have ripened into a binding contract. In addition, as the Japanese High Court explained in a case decided in 1987:

> In the event that preparation between two parties progresses toward conclusion of an agreement and the first party comes to expect that the agreement will surely be concluded, the second party becomes obligated under the principles of good faith and trust to try to conclude the agreement, in order not to injure the expectation of the first party. Therefore, if the second party, in violation of its obligation, concludes that the agreement is undesirable [absent certain circumstances], it is liable for damages incurred by the first party as the result of its illegal acts (Yanagida et al. 1994).

US law imposes more limited precontractual liability under the doctrine of promissory estoppel for the out-of-pocket reliance damage suffered when one party reasonably and foreseeably relies on a promise made during negotiations.

1.3 Consideration

In most jurisdictions each side must provide something of value to form a valid contract. The thing of value, known as consideration, can be money, property, a promise, or a service. Consideration is present if a party agrees to do something that he or she is not otherwise legally required to do or if the party agrees to refrain from doing something that he or she was otherwise legally entitled to do. Courts usually will not scrutinize the value of the consideration, thus giving rise to the adage that even a peppercorn can be adequate consideration. Indeed, under Japanese law, there is no requirement for consideration at all.

1.4 Unconscionability

A court will refuse to enforce a contract if it is so oppressive or fundamentally unfair as to shock the conscience of the court. This doctrine of unconscionability dates back to Aristotelian theory and Roman law (Gordley 1981). The doctrine of *laesio enormis* provided a remedy when the contract price for the sale of goods deviated by at least half from the just price, defined as the market price for similar goods under similar circumstances. The German and French laws that relieve a party from its obligations under a contract that is not for the just price have their roots in this doctrine of *laesio enormis*. Similarly, in England and the United States, the general principle of freedom of contract has always been limited by the doctrine of unconscionability.

1.5 Genuineness of Assent

Even if a contract meets all the requirements of validity (agreement, consideration, capacity, and legality), a court will not enforce it if there has been no true 'meeting of the minds' between the parties. This may occur when there is fraud, duress, or a mistake of fact.

1.5.1 Fraud and duress. There is no meeting of the minds if a contract is tainted with fraud or if one party was forced to enter into it under duress. Thus, a contract is voidable, that is, subject to being undone by one or more parties, if one party induced another to enter into the contract by making false statements, by failing to disclose information when under a legal duty to do so (e.g., when the parties are partners), or by engaging in blackmail or extortion.

1.5.2 Mistake of fact. A contract is voidable if it was predicated on a substantial mistake of fact. For example, in the classic case *Raffles v. Wichelhaus* (1864), two parties had signed a contract in which Wichelhaus agreed to buy 125 bales of cotton to be brought by Raffles from India on a ship named *Peerless*. Unbeknownst to both parties, there were two ships named *Peerless* both sailing out of Bombay during the same year. Raffles meant the *Peerless* sailing in December, and Wichelhaus meant the *Peerless* sailing in October. When the cotton arrived in the later ship, Wichelhaus refused to complete the purchase, and Raffles sued for breach of contract. The court concluded that the mutual mistake of fact had resulted in there being no meeting of the minds and, therefore, no contract.

1.5.3 Mistake of judgment and commercial impracticability. While a mistake of fact makes a contract voidable, a mere mistake of judgment does not. Therefore, even if a party is mistaken about some aspect of what is bargained for, the contract is still enforceable. For example, the Wisconsin Supreme Court held that a contract to sell a stone for $1 was enforceable when neither party knew at the time that the stone was in fact a diamond (*Wood v. Boynton* 1885).

Some jurisdictions will relieve a party of its contractual obligations if performance is commercially

impracticable, that is, if performance is made financially impracticable by the occurrence of an event unforeseen by the contract (Eagan 1980).

1.6 Statutes of Fraud

1.6.1 United States. Although most oral contracts are enforceable, almost every state in the United States has statutes requiring that certain types of contracts be evidenced by some form of writing. These Statutes of Fraud require that any (a) contract for the transfer of any interest in real property (such as a deed, lease, or option to buy), (b) promise to pay the debt of another, (c) agreement that by its terms cannot be performed within a year, (d) prenuptial agreement (that is, an agreement entered into before marriage that sets forth the manner in which the parties' assets will be distributed and the support to which each party will be entitled in the event of divorce), and (e) contract for the sale of goods priced at $500 or more must be evidenced by some writing signed by the party against whom enforcement is being sought.

1.6.2 England, France, and other jurisdictions. Although the original Statute of Frauds was adopted in England in 1677, England has repealed all of it except the provisions relating to transfers of interests in land and promises to pay the debt of others. Many other countries that once had a Statute of Frauds have repealed all or parts of it. France, Japan, and other civil law countries never had a writing requirement.

1.7 Remedies

In the event that a valid contract is breached, the other party is entitled to monetary damages or to a court order requiring performance. The nonbreaching party is generally entitled to expectation damages, that is, such amount of money as is necessary to put the plaintiff in the cash position it would have been if the contract had been fulfilled.

Consequential damages, that is, compensation for losses that occur as a foreseeable result of the breach, are also often available. However, if the party suing for consequential damages has suffered lost profits that were not reasonably foreseeable or a natural consequence of the breach, then there will be no award of consequential damages.

2. Agency

Agency law specifies under what circumstances the contracts entered into by one party (the agent) will bind another (the principal) on whose behalf the agent is acting. Agency law also addresses the question of when the principal will be liable for the torts or civil wrongs of the agent. The employer/employee relationship (historically described as the master/servant relationship) is perhaps the most common type of agency in business settings.

2.1 Authority to Bind the Principal

A contract entered into by an agent on behalf of the principal is binding on the principal only if the agent had authority to enter into the contract or if the contract is later ratified by the principal. Authority to enter into a contract can be either actual or apparent. Actual authority can be given expressly or can be implied from the nature of the responsibilities the principal has given the agent to fulfill. Apparent authority exists when the principal, either directly or indirectly, leads a third party to believe that the agent has power to act on the principal's behalf. If a contract is one subject to the Statute of Frauds then, under the Equal Dignities Rule, execution of that contract by an agent would be binding on a principal only if the agent/principal relationship itself were embodied in a writing.

2.2 Respondeat Superior

Under the doctrine of *respondeat superior*, the principal is liable for any torts committed by the agent while acting within the scope of employment. A minor detour, whereby the agent deviates from the assigned task, is still considered to have been within the scope of employment unless the deviation is so great as to constitute a frolic, that is, an excursion that in no way furthers the business of the principal.

A principal is more likely to be responsible for activities that otherwise might be deemed frolics when the harm done by the agent was made possible by an instrumentality provided by the principal. For example, if the principal provides the agent a truck to drive, the principal is more likely to be responsible for automobile accidents occurring while the agent is driving the truck even if at the time of the accident the agent was not directly furthering the business of the principal. Similarly, courts have found the city liable when a police officer arrests then rapes a female suspect, because much of the police officer's power over the woman to begin with derived from the badge of office given by the city.

In Japan, the employer is not liable for the misdeeds of an employee if the employer exercised due care selecting the employee and supervising the employee's work or if the damages would have arisen even if due care had been exercised.

2.3 Employment Laws

Most jurisdictions have statutes and case law setting the minimum wage, imposing limits on the employ-

ment of children, and specifying whether an employee without a written employment contract is an employee at-will who can be terminated at any time for any or no reason, or an employee who can be fired only for just cause. In addition, in many jurisdictions, the employment relationship is subject to statutes banning discrimination based on race, color, national origin, religion, sex, or age. For example, effective April 1, 1999, Japan prohibited discrimination based on sex in the hiring of women. Previously, Japanese law banned sexual discrimination once the person was hired, but did not extend that protection to the hiring decision itself.

3. *Forms of Business Entities*

A third body of law fundamental to the conduct of business prescribes the types of entities by which business can be conducted. A business may be conducted as a sole proprietorship owned by one person, a general partnership, a limited partnership, a limited liability company, or a corporation.

3.1 *Corporations*

The most prevalent entity used for large businesses is the corporation. A corporation is a fictitious person that has capacity to enter into contracts and to hold property. The equity of a corporation, evidenced by shares of stock, is owned by the shareholders. The shareholders elect the board of directors, which has the responsibility for managing the corporation. The board elects the officers, such as the president, treasurer, and secretary, who report to the board and serve at its pleasure.

3.1.1 Corporate governance. The respective rights and obligations of shareholders, directors, and officers of a corporation are dictated by the statutes of the jurisdiction in which the corporation is formed. For example, in the state of Delaware, where more than half of the Fortune 500 companies are domiciled, the board of directors has broad latitude to determine when it is in the best interests of the corporation to rebuff a hostile takeover offer (Bagley 1999).

Jurisdictions vary significantly in their treatment of corporate stakeholders, such as employees, customers, suppliers, and communities. In large German corporations, co-determination statutes require that half of the supervisory board be elected by the employees and labor unions and councils. In Japan, boards traditionally included no nonexecutive directors and put the welfare of the employees before shareholder return. In the United States, shareholder primacy tends to be the rule, at least once a decision has been made to sell control of the corporation or to break it up (Bagley and Page 1999).

3.2 *Limited Liability*

A key advantage of a corporation is the limited liability it provides to its owners. Absent facts justifying piercing the corporate veil, such as the commingling of personal and corporate assets or the shareholders' disregard for the corporate form, the shareholders are not liable for the debts of the corporation. In contrast, in a general partnership, each of the general partners is personally liable for the debts and other obligations of the partnership. Similarly, the owner of a sole proprietorship is personally liable for all debts of the business. In a limited partnership, the managing partner has the same unlimited liability as a general partner in a general partnership, but the limited partners have no liability beyond their capital contribution unless they actively participate in management. In the case of a limited liability company, the owners (called members) have limited liability even if they are active in the management of the business.

3.3 *Tax Considerations*

The choice of business entity can often be driven by the different tax treatment accorded. For example, in the United States, a corporation is generally taxed as a separate entity so that the corporation pays tax on its income; when any corporate property, including dividends, is distributed to the shareholders, the shareholders pay a tax on those distributions. In contrast, with a limited liability company or a limited partnership, there is no entity-level tax; instead, the income earned by the firm is taxed directly to the owners even if in fact none of that income is distributed to them.

4. *Commercial Laws Governing the Sale of Goods, Negotiable Instruments, and Secured Lending Transactions*

Because of the key role played by purchases and sales of goods in virtually every type of business, many jurisdictions have specific statutes designed to facilitate sales transactions. Commercial laws also govern the creation and transfer of negotiable instruments and the lending of money secured by property.

4.1 *Uniform Commercial Code*

In the United States, every state other than Louisiana has adopted a version of the Uniform Commercial Code (UCC).

4.1.1 Sale of goods. The provisions in UCC Article 2 concerning the sale of goods eliminate many of the formalities required under the common law. For example, the UCC abolishes the 'mirror image rule,' which requires that an acceptance be identical in terms to the offer. Instead, the UCC acknowledges that sales transactions are often represented by an exchange of preprinted forms with additional or conflicting terms (the 'battle of the forms'), and it provides guidance as to which terms will be deemed to be part of the contract.

4.1.2 Negotiable instruments. Article 3 of the UCC governs negotiable instruments, including drafts, checks, notes, and certificates of deposit. The most common type of draft is a check: the writer of the check is the drawer; the bank on which the check is drawn is the drawee; and the person to whom the check is payable is the payee. Article 3 sets forth the requirements for negotiability: a writing signed by the maker or drawer containing a definite and unconditional promise or order to pay a fixed amount of money on demand or at a definite time. It also specifies what endorsements are required to negotiate an instrument payable to order, such as a check.

The UCC prescribes the rights given a holder in due course, i.e., the holder of a negotiable instrument who took it for value, in good faith, and without notice that it had been dishonored or that any person has a defense against or claim to it. Unlike an ordinary holder who has only those rights that the transferor had, a holder in due course takes the instrument free of most of the defenses and claims to which the transferor was subject.

4.1.3 Secured lending transactions. Article 9 of the UCC governs transactions in which payment of a debt is guaranteed, or secured, by property owned by the debtor. It permits the secured party to perfect its security interest by filing a financing statement or, in some cases, by possession of the collateral. In general, the first to perfect has priority over other third parties who may wish to have their debts satisfied out of the same collateral.

4.2 Convention on Contracts for the International Sale of Goods

Most sales of goods involving parties in more than one country are governed by the Convention on Contracts for the International Sale of Goods (CISG), unless the parties expressly opt out of CISG. Unlike the UCC, CISG requires that the terms of the acceptance mirror those of the offer. It provides that an offer is irrevocable if the offeror orally states that it is irrevocable, or if the offeree reasonably relies on the offer as being irrevocable. In contrast, the UCC requires that an irrevocable offer not supported by consideration be in writing. Under CISG, the parties must specify the price or provisions for its specification, whereas the UCC requires that only the quantity be specified.

Some commentators maintain that CISG imposes a duty for merchants to conduct themselves in good faith (Koneru 1997). If this interpretation is correct, then, for example, a buyer could not arbitrarily deny a seller's request for additional time to deliver goods. This contrasts sharply with the UCC's perfect tender rule, which permits a buyer to reject goods not conforming perfectly to the contract.

5. Product Liability

Most jurisdictions impose liability on the sellers of goods for defective products. For example, the Council of Ministers of the European Union adopted a directive in July 1985, which made manufacturers and producers strictly liable for injuries caused by defects in their products. Previously, all the member states other than France required the injured consumer to prove both negligence and privity of contract (Bagley 1999). In 1963, the Supreme Court of California became the first in the United States to impose strict liability for product defects as a matter of common law (*Greenman v. Yuba Power Products, Inc.* 1963).

In 1997, the American Law Institute (ALI) adopted the *Restatement (Third) of Torts: Product Liability*, which abandoned the notion of strict liability in defective design cases and imposed on the plaintiff the burden of proving that there was a defect in the product that could reasonably have been avoided by the use of an alternative design. Although judges often regard the ALI's Restatements as persuasive authority, they are free to ignore them. The Supreme Court of Connecticut declined to impose a burden on the plaintiff to prove a feasible alternative design (*Potter v. Chicago Pneumatic Tool Co.* 1997), and it remains to be seen whether other courts will adopt the Restatement (Third) view of product liability.

6. Intellectual Property

Intellectual property is any product or result of a mental process that is given legal protection against unauthorized use. It includes patents, copyrights, trademarks, and trade secrets.

6.1 Patents

A patent is a government-granted right to exclude others from making, using, or selling an invention for 20 years after the date the patent application is filed. A utility patent, the most common type of patent, protects any novel, useful, and nonobvious process,

machine, manufacture, or composition of matter. In the United States, human-made living organisms can be patented as can computer software.

6.1.1 Business process patents. In 1998, the US Court of Appeals for the Federal Circuit reversed prior precedent and held that a business process could be patented (*State Street Bank & Trust Co. v. Signature Fin. Group, Inc.* 1998). Several companies engaged in electronic commerce, including Amazon.com (with a patent on one-click technology for placing orders online) and Priceline (with a patent on using computers for reverse auctions whereby the buyer names a price and seeks a willing seller), filed suits in late 1999 to protect their methods of conducting electronic business. If upheld, these patents could chill innovation in the fast-moving market for electronic commerce.

6.2 Copyrights

Unlike a patent, which can protect an idea, a copyright protects original expressions fixed in a tangible medium, such as books, records, musical scores, and films. Although registration of a copyright gives the author certain additional rights, the copyright is created as soon as the work is fixed in a tangible medium of expression. Under the Berne Convention, most industrialized countries give persons with a copyright under the laws of another jurisdiction the same rights that a national would have under local law.

6.3 Trademarks

A trademark is a word, name, symbol, or device used by a manufacturer to identify and distinguish its goods from those manufactured by others and to indicate the source of the goods. Examples include Coca-Cola, Marlboro, and Nescafe. A service mark, such as Citigroup, is similar but is used in connection with services.

To avoid consumer confusion, trademarks are granted for a particular field of use, e.g., nonalcoholic beverages or automobiles. Thus, more than one company can register the same mark. The importance of a company's domain name, or electronic address, for Internet branding purposes and the fact that domain names contain no field-of-use restriction have required courts to consider when use of a trademark within a domain name constitutes trademark infringement. Courts have ruled against 'cybersquatters,' who register a trademarked name as a domain name to force the trademark owner to buy the domain name. They also have found infringement when the company using the domain name offers the same or similar products as the trademark owner or uses the name in a way that dilutes or tarnishes a famous mark. For example, Hasbro, owner of the trademark Candyland for a children's game, was able to force the owner of Candyland.com to stop using the domain name for a pornography site. Effective January 1, 2000, the U.S. Anticybersquatting Consumer Protection Act bans registration of domain names in bad faith.

6.4 Trade Secrets

A trade secret is information, including a formula, compilation, or process, that derives independent economic value from not being generally known and is the subject of reasonable efforts to maintain its secrecy. Perhaps the most famous trade secret is the formula for Coca-Cola. In response to perceived problems facing US businesses from foreign theft of trade secrets, the United States adopted the Economic Espionage Act in 1996, which imposes criminal liability (including fines and prison sentences) on any person who knowingly steals a trade secret or knowingly receives a wrongfully obtained trade secret.

In 1997, Volkswagen AG (VW) paid General Motors Corporation (GM) $100 million to settle allegations that VW's purchasing chief stole trade secrets, including plans for future GM models and car-building techniques, when he left GM to join VW. Under the emerging doctrine of inevitable disclosure, an employer can challenge a former employee's decision to work for a competitor if the new position would result in the inevitable disclosure or use of the former employee's trade secrets.

7. Other Laws Affecting the Legal Environment of Business

Other laws affecting the legal environment in which businesses operate include antitrust and competition law; environmental law; consumer protection laws; regulations governing the offer and sale of securities; and bankruptcy law.

7.1 Antitrust and Competition Law

The wide-ranging case brought in 1998 by the US Justice Department against Microsoft Corporation promises to help define how courts in the twenty-first century will apply the Sherman Act's 110-year ban on restraints of trade to the digital economy and its network effects, whereby the value of owning a product rises as other consumers own the same product. That case and others will have to wrestle with the question of how to reconcile the exclusionary rights arising out of intellectual property law with the bans on anticompetitive behavior and abuse of dominant position contained in such laws as the Sherman Act in the United States, Articles 85(1) and 85(2) of the Treaty of Rome in the European Union, and the Act Concerning Prohibition of Private Monopoly and Maintenance of Fair Trade in Japan.

7.2 Environmental Laws

In response to high-cost hazardous waste spills and clean-ups, a number of jurisdictions have adopted laws governing the storage, transport, and disposal of hazardous waste. The European Union has adopted the policy that the polluter pays with the objectives of protecting the quality of the environment and human health and promoting the prudent and rational utilization of natural resources. The United States imposes liability not only on the person who stored, transported, or disposed of the waste, but also on the owner or operator of the property at the time the waste is discovered.

7.3 Consumer Protection

Jurisdictions continue to adopt laws and regulations to protect consumers from deceptive sales practices and advertising. The rise of the Internet and the ease with which electronic databases can be used to collect personal data have sparked new concerns about privacy.

7.4 Securities Law

The United States has the most restrictive rules regarding the offer and sale of securities, such as common and preferred stock and bonds. Generally, any public offering of securities must be made pursuant to a registration statement declared effective by the Securities and Exchange Commission (SEC). The most important part of the registration statement is the prospectus, which contains detailed information about the company issuing the securities and its management, principal shareholders, business, and financial condition. The US laws provide civil and criminal penalties for any offer or sale of a security, regardless of whether it is registered, involving a misstatement of a material fact or an omission that makes the facts stated materially misleading. SEC Regulation S facilitates the offering of securities outside of the United States by US firms by exempting many such offerings from the burdensome US registration requirements that would be applicable if the securities were sold within the United States.

7.5 Bankruptcy Law

The principle that a debtor should be permitted to discharge certain debts has ancient roots. As early as 1400 BC, Mosaic Law required unconditional forgiveness of debts every seven years. The British Parliament under King Henry VIII passed a Bankruptcy Act in AD 1542 to regulate failing English merchants.

A key objective of bankruptcy law is to ensure that creditors and equity holders of financially distressed businesses are treated fairly in a judicial setting designed to ensure the efficient handling of claims. In general, creditors must be paid in full before equity holders receive anything.

The United States provides perhaps the most pro-debtor bankruptcy relief available. An insolvent firm in the United States generally can elect to liquidate its assets or to reorganize the business, often with the debtor still in control. In contrast, in Australia, creditors can force the liquidation of a company that cannot pay its debts even if the debtor argues that the business would be worth more as an operating whole. Some countries, such as Germany, do not provide for the complete discharge of debts in bankruptcy, only arrangements for their payment over time.

See also: Area and International Studies: Political Economy; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; International Law and Treaties; International Organization; International Trade: Commercial Policy and Trade Negotiations; Law and Economics; Law and Economics: Empirical Dimensions; Markets and the Law; Monetary Policy; Regulation, Economic Theory of; Regulation: Working Conditions; Venture Capital

Bibliography

American Law Institute 1997 *Restatement (Third) of Torts: Product Liability*. American Law Institute, Philadelphia

Bagley C E 1999 *Managers and the Legal Environment: Strategies for the 21st Century*, 3rd edn. West Educational Publishing, Cincinnati, OH

Bagley C E, Page K L 1999 The devil made me do it: Replacing corporate directors' veil of secrecy with the mantle of stewardship. *San Diego Law Review* **36**: 897

Eagan W 1980 The Westinghouse uranium contracts: Commercial impracticability and related matters. *American Business Law Journal* **18**: 281–302

Farnsworth E A 1987 Precontractual liability and preliminary agreements: Fair dealing and failed negotiations. *Columbia Law Review* **87**: 217–94

Gordley J 1981 Equality in exchange. *California Law Review* **69**: 1587

Greenman v. Yuba Power Products, Inc. 1963 377 P. 2d 897 (Cal.)

Hadley v. Baxendale 1854 156 Eng. Rep. 145

Koneru P 1997 The international interpretation of the UN convention on contracts for the international sale of goods: An approach based on general principles. *Minnesota Journal of Global Trade* **6**: 105

Potter v. Chicago Pneumatic Tool Co. 1997 694 A. 2d 1319 (Conn.)

Raffles v. Wichelhaus 1864 159 Eng. Rep. 375 (Exch.)

Texaco, Inc. v. Pennzoil Co. 1987 729 S.W. 2d 768 (Tex. Ct. App.)

Wood v. Boynton 1885 25 N.W. 42 (Wis.)

Yanagida Y, Foote D H, Johnson Jr E S, Ramseyer J M, Scogin Jr H T 1994 *Law and Investment in Japan: Cases and Materials*. Harvard University Press, Cambridge, MA

C. E. Bagley

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

C

Campaigning: Political

At the microlevel, political campaigning involves an individual candidate attempting to win political office. At the macrolevel, political parties seek to gain seats in government for their partisans and ultimately to gain power or at least to share power in the relevant governmental arena. At each level decisions must be made about candidacies, organizations must be built, messages must be refined, strategies must be set, and the public must be convinced to support those waging a particular campaign. Only certain aspects of political campaigns are viewed by the public. Frequently, the actions taken away from public view are more determinative of success and failure.

1. The Choice of Candidates

Decisions about who will run for office involve decisions by potential candidates as to whether or not they will run and decisions by political parties, acting under a variety of rules, concerning who their candidates will be.

1.1 The Decision to Run

Individuals who run for political office must exhibit political ambition beyond the desire to simply serve their community. Campaigning for and serving in political office involves personal costs and public exposure that deter many from seeking office. However, even those ambitious for elective office must make decisions about what offices to seek and when to seek those offices (Jacobson and Kernell 1981). These decisions are based on a complex, often-implicit cost-benefit analysis. Potential candidates weigh the costs and benefits of both winning and losing. Their decision calculus can be summarized:

$$p_r = p_w\ (B_w - C_w) + p_l(B_l - C_l)$$

where p_r = probability of running; p_w = probability of winning; p_l = probability of losing; B_w = benefit of winning; C_w = cost of winning; B_l = benefit of losing; C_l = cost of losing.

The difficulty that potential candidates face in making these decisions relates to the uncertainty regarding virtually all of the relevant variables (Maisel 1986, Kazee 1994). Generally the rather obvious benefits far outweigh the costs of winning for those with the requisite political ambition; but costs must be weighed—loss of privacy, loss of time with family, frequently loss of income. Similarly, the evident costs of losing are almost always greater than the benefits; but some benefits do accrue to losers—making a name for oneself, building up good will with party officials. Beyond these broad generalizations, little is known about how specific candidates make individual decisions given the dazzling array of difficult-to-estimate variables (Maisel and Stone 1997).

1.2 Parties' Choices of Candidates

The ways in which political parties choose their candidates for office varies significantly from political system to political system and from party to party. The key variable in this aspect of the process relates to the extent to which party organization controls places on the ballot contrasted with voters acting independently. In the former, the process is often described as candidate recruitment, in the latter, candidate emergence.

At one extreme are those systems in which political parties slate their candidates for office, ranking candidates or placing them in safe constituencies. Most parties operating in electoral systems with proportional representation follow this procedure. It is also followed in systems with strong party systems, in which citizens vote more for parties than they do for individual candidates. Many European systems fit this description (Hix and Lord 1997).

At the other extreme are systems, of which the United States is prototypical, in which voters choose party candidates in primary elections. The shape of the electorate in primary elections varies considerably from polity to polity. In theory, primary elections are open only to members of one political party or another, but in actual practice how membership is defined varies according to local rules and can be quite restrictive or totally open. In either case, voters in the primary elections determine the party nominee, even if that nominee would not have been the choice of party officials (Maisel 1999).

Many systems are hybrids, with party officials and the voters sharing power. The US presidential nominating system stands as perhaps the most complex means of choosing party candidates. Delegates elected to national party conventions make the actual nom-

inations. Some delegates are selected because of their positions within the party or as elected officials with party affiliations; others are chosen at party meetings (or caucuses) held in local communities; still others are chosen by the voters in primary elections, based on lists of delegate candidates approved by the presidential campaigns they seek to support (Polsby and Wildavsky 2000, Wayne 2000).

2. The Campaign Organization

In polities with strong party systems, campaign organizations tend to be party-centered. In these cases, the voters cast their ballots for candidates as representatives of a party. The party defines the campaign message and communicates it to the voters. The party raises the money to fund the campaign, does the polling necessary to understand what the electorate is thinking, and structures the campaign strategy and tactics. Typically full-time, year-round party workers perform all these functions or they are contracted out to professionals.

Where strong party systems do not exist, individual candidates must build their own organizations. In these cases, candidates are most concerned with their own elections, not that of those who share this party label. Candidates choose campaign managers whose sole goal is victory for that one candidate. They build personal organizations to handle the important functions of campaigning—doing research to develop issue positions and to counter opponents' proposals; polling the public to ascertain perceptions of the candidate and reactions to strategies; writing speeches; communicating with the press; developing and placing advertisements; sending out direct mail appeals, telephoning and leafleting the district; scheduling and preparing for candidate appearances; and perhaps most important, raising money. For larger and more expensive campaigns, all of these functions are performed by paid workers or volunteers in a candidate's own organization, or bought with campaign contributions.

Recent campaign organizations have been less able to handle all of these functions, and two means of coping have emerged. For campaigns in smaller constituencies, the candidate often performs all of these functions him or herself. They are done more or less well depending on the skill of the candidate, the time available, the level of competition, and the sophistication that is expected in campaigning for office (Maisel 1999).

Even those with relatively high budget campaigns often find that they cannot hire staff to perform all of the necessary campaign activities. These candidates frequently turn to paid political consultants, experts from outside the campaign who specialize in certain aspects of the campaign process and sell their services to many campaigns during any election cycle. Firms specialize in media advertising, polling, direct mail, telephone banks, fundraising, and virtually every other aspect of campaigning; other firms are structured to take over all or many aspects of a campaign. Some work in particular geographic areas. Others are ready to sell their services anywhere throughout an entire nation or even internationally.

3. Defining Messages and Setting Strategy

A political campaign requires getting a message to voters. To do that, a candidate must define what the message is and must devise a strategy to reach the voters who will be swayed by that message. Again, how this is to be done varies with whether the campaign in question is party-centered or candidate-centered.

For party-centered campaigns, the message is the party platform, and the voters to be reached at those in the party's core constituency and others to whom the message might appeal in a particular year. Neither the message nor the strategy vary much from year to year, though marginal changes are made as the context changes—and these marginal changes might well spell the difference between victory and defeat.

For candidate-centered campaigns these decisions are among the most crucial. Candidates must decide whether to stick with the party message or to devise one of their own; they must decide what particular issues will play best to their voters; and they must figure out who precisely those voters are. Campaign strategists divide the electorate in a variety of ways—by demographic characteristics, by economic interests, by geographic locations. The goal is to use research and polling data to determine what appeals will work with which audiences and to have the candidate address those audiences appropriately. For incumbents seeking re-election, the strategic consideration often involves relying upon strengths demonstrated earlier. For challengers, the strategy must be to find an opponent's weakness and to emphasize one's own strengths.

Effective campaign strategies remain more art than science. Professional political campaigners know what has worked in the past. But many strategists learn those lessons. The successful strategists are those who can see how to change the approach as the context changes, as is often said in the military, 'not to refight the last war but to plan for the next one.'

4. Communicating with the Voters

To be effective, a political campaign must be able to send the message it has devised to the audience it has targeted. Campaign messages are communicated in two ways, through free media and paid media.

4.1 Free Media Coverage

Political campaigns seek free media coverage in whatever form they can, whenever they can. Candidates want to be reported on in newspapers and newsmagazines; they want to grant interviews on radio; they want their events covered on television. Free media coverage is advantageous not only because there is no cost but also because the viewer sees news coverage of a candidate as giving that candidate a sense of legitimacy; the candidate is seen as part of the day's news, not as part of a paid advertisement. However, free media is uncontrolled exposure. Campaign strategists can try to structure what topics will be covered—and campaigns are exploring increasingly sophisticated means to do this—but too frequently the message delivered is very different from the message the candidate seeks to convey.

4.2 Paid Media Advertising

The opposites apply to paid advertising. Paid media—whether in the form of television, radio, or print advertisements, of Internet sites, or of direct mail—is costly and is often viewed by the prospective voter as the slanted message it in fact is. However, these paid media also have the distinct advantage of allowing a candidate to emphasize exactly the message desired. These messages can be narrowcast, that is designed for and directed to specific target audiences. These ads can create images, discuss positions, or—increasingly in recent years—attack an opponent, all in precisely the way strategists feel will be most effective for a specified audience.

The critical concerns about political campaigns today relate to paid media. Many fear that effective political messages cannot be conveyed in 30-second advertisements, the medium preferred by advertising executives. Others complain that the emphasis on negative, attack advertising has poisoned the atmosphere that surrounds political life, keeping some of the best candidates from running (Ansolabehere and Iyengar 1995). Virtually everyone bemoans the exorbitant costs of campaigns, costs spurred by reliance on paid media. However, political campaigns will continue to rely on paid media so long as they are effective in communicating a candidate's message to the target audience; for that is the means to electoral success—the ultimate message of a campaign's effectiveness.

See also: Advertising: Effects; Electoral Systems; Electronic Democracy; Mass Media, Political Economy of; Media and Social Movements; Media Effects; Media, Uses of; Party Identification; Political Machines; Political Money and Party Finance; Political Parties; Polling; Primary Elections; Voting, Sociology of; Women's Suffrage

Bibliography

Ansolabehere S A, Iyengar S 1995 *Going Negative: How Attack Ads Shrink and Polarize the Electorate*. Free Press, New York

Hix S, Lord C 1997 *Political Parties in the European Union*. St. Martin's Press, New York

Jacobson G C, Kernell S 1981 *Strategy and Choice in Congressional Elections*. Yale University Press, New Haven, CT

Kazee T A (ed.) 1994 *Who Runs for Congress? Ambition, Context, and Candidate Emergence*. Congressional Quarterly, Washington, DC

Maisel L S 1986 *From Obscurity to Oblivion: Running in the Congressional Primary*, rev. edn. University of Tennessee, Knoxville, TN

Maisel L S 1999 *Parties and Elections in America: The Electoral Process*, 3rd edn. Rowman & Littlefield, Lanham, MD

Maisel L S, Stone W J 1997 Determinants of candidate emergence in U. S. House elections: an exploratory study. *Legislative Studies Quarterly* **22**: 79–96: Current work from the Candidate Emergence Project can be found at http://socsci.colorado.edu/CES/home.html

Polsby N W, Wildavsky A 2000 *Presidential Elections: Strategies and Structures of American Politics*, 10th edn. Chatham House Publishers, New York

Wayne S J 2000 *The Road to the White House 2000*. St. Martin's Press, New York

L. S. Maisel

Campbell, Donald Thomas (1916–96)

Donald Thomas Campbell, born November 20, 1916 in Grass Lake, Michigan, died May 6, 1996 in Bethlehem, Pennsylvania, thus ending a career marked by an array of superlatives. Prolific scholar and author, original thinker, ebullient teacher with a notable twinkle in his eye, and generous colleague, Campbell was widely regarded as the most important social science methodologist of the twentieth Century. Several of Campbell's articles vie with one another as among the most widely cited pieces of social science scholarship. Repeatedly he coined phrases any one of which most scholars would be proud to have conceived, for example, 'quasi-experiments,' 'unobtrusive measures,' 'internal and external validity,' and 'plausible rival hypotheses.' His concepts are incorporated as fundamental in several fields—psychology, sociology, anthropology, organization and management sciences, public policy, evaluation, education, and philosophy—and in common use by scholars unaware of their inventor. Moreover, Campbell's welcoming and self-critical personal style, modeled on his scholarship, led him to embrace and examine all 'plausible rival hypotheses' and endeared him to students and colleagues alike.

Campbell received his B.A. (1939) and Ph.D. (1947), both in Psychology, from the University of California (Berkeley), and taught at several institutions, including

Ohio State University (1947–50), the University of Chicago (1950–53), Syracuse University (as N.Y. State Board of Regents Albert Schweitzer Professor, 1979–82) and Lehigh University (as Distinguished University Professor of Sociology-Anthropology, Psychology and Education, 1983–96), but the majority of his scholarly years were spent as professor of psychology at Northwestern University (1953–79) with which Campbell's name and reputation are 'permanently and inextricably co-identified' (Campbell 1981).

Campbell's honors include a Fulbright Visiting Professorship in Social Psychology at Oxford (1968), the Distinguished Scientific Contribution Award from the American Psychological Association (1969), election to the National Academy of Sciences and the American Academy of Arts & Sciences (1973), the Kurt Lewin Memorial Award from the Society for the Psychological Study of Social Issues (1974), presidency of the American Psychological Association (1975), the Williams James Lectureship at Harvard (1977), the Myrdal Prize in Science (1977), and the Distinguished Contribution Award from the American Educational Research Association (1981). In addition, two annual awards are given in his name: The Donald Campbell Award for Significant Research in Social Psychology (1982–) and the Donald Campbell Award from the Policy Studies Organization, to an 'outstanding methodological innovator in public policy studies' (1983–). Campbell received honorary degrees from numerous universities, including Oslo, Michigan, Chicago, and Southern California.

But Campbell felt most honored by the vast array of books dedicated to him, a list that grew steadily in several disciplines before and after his death.

1. *Major Contributions*

When he died, Campbell's resumé listed more than 230 published books, monographs, and articles. Any brief discussion of his contributions must necessarily focus on a few. Best known as a methodologist, with considerable scholarly work in the areas of experimental design, measurement, and social experimentation, Campbell is perhaps remembered especially for his explorations of the concept of validity. He was also a well-regarded epistemologist with a keen interest in the sociology of science. To each domain, Campbell brought his unique mind to bear on the problem of knowledge production. In a remarkable array of works, Campbell explored from several perspectives the inevitable fallibilities inherent in both observers and methods in accurately portraying the world.

In his earliest publication, 'The Indirect Assessment of Social Attitudes' (1950), Campbell evidenced an aspect of the interests that would become the overall focus of his scholarly life: the imperfections introduced by human observers and methods, including the scientific method, in the search for veridical description of the physical world and mechanisms for compensating for these errors. (Campbell later declared profound 'ambivalence' toward the use of disguised measures; see Kidder and Campbell 1970.) Campbell explored the sources and loci of bias and developed and refined methods for illustrating and for minimizing these biases. For example in Jacobs and Campbell (1961) he demonstrated how consensus in the 'reality' of an arbitrarily invented but shared group norm could persist over generations of experimental subjects. In a series of crosscultural explorations Campbell and co-workers (e.g., Segall et al. 1966) explored sources of validity and invalidity in perceptions of 'in-groups' and 'out-groups,' and the differential susceptibility to perceptual illusions between European and non-European cultures.

Campbell and Fiske (1959) published 'Convergent and Discriminant Validation by the Multitrait, Multimethod Matrix,' a complex and frequently cited analysis of the need for multiple measures of underlying constructs (traits) and of the need for the demonstrated capacity of methods to distinguish among traits if one is to minimize irrelevant measurement artifacts. Since every measure is partially invalid, multiple and distinctive methods are called for to yield a 'heterogeneity of irrelevancies.' Campbell and Fiske explored the value of simultaneously deploying maximally different measures of an underlying trait (in pursuit of convergent validation) and of different traits measured the same way (in pursuit of discriminant validation). Only by demonstrating that an underlying trait can be measured in various distinct ways and that the chosen measurement strategies distinguish among measured traits, can one minimize the impact of misleading artifacts.

By the early 1960s Campbell had established an enviable reputation for the vigor and reach of his campaign for advancing the methods and theory associated with issues of validity. His reputation was further advanced with the publication with (Campbell and Stanley 1963) of 'Experimental and Quasi-experimental Designs for Research,' a widely circulated work (elaborated as Cook and Campbell 1979) in which they popularized the terms 'internal' and 'external' validity:

> Internal validity is the basic minimum without which any experiment is uninterpretable: did in fact the experimental treatment make a difference in this specific experimental instance? External validity asks the question of generalizability: To what populations, setting, treatment variables, and measurement variables can this effect be generalized? (Campbell and Stanley 1963).

Another widely cited treatise, *Unobtrusive Measures: Non-reactive Research in the Social Sciences* (Webb et al. 1966) evolved during affable lunchtime conversations with colleagues from different departments. Over several years these colleagues sought to identify relatively unbiased, nonreactive ways of mea-

suring behavior. A frequently cited example describes ways a Chicago museum tried to identify the popularity of its exhibits. 'Obtrusive measures' included questioning visitors leaving the museum; an 'unobtrusive measure' was the frequency with which museum staff had to replace worn tiles in front of exhibits. By that index, the exhibit of baby chicks hatching live and wet from quivering eggs was by far the most popular.

Campbell's participation in an interdisciplinary conference spawned his brilliant foray into the sociology of science entitled 'Ethnocentrism of Disciplines: A Fishscale Model of Omniscience' (1969a). On reading a paper 'in an area of high relevance' to his own work by esteemed fellow social psychologist William McGuire, Campbell realized and then acknowledged that he had not 'read or read-at even half of McGuire's citations, and was not at all aware of the existence of another sizeable proportion' (Campbell 1969a). While many scholars might react with chagrin, Campbell instead reflected on how the 'highly arbitrary' organization of universities and academic fields skewed knowledge production:

> Thus anthropology is a hodgepodge of all novelties venturing into exotic lands—a hodgepodge of skin color, physical stature, agricultural practices, weapons, religious beliefs, kinship systems, history, archeology, and paleontology … Thus psychology is a hodgepodge of sensitive subjective biography, or brain operations, or school achievement testing, of factor analysis, of Markov process mathematics, of schizophrenic families, of laboratory experiments on group structure in which persons are anonymous … Thus economics is a hodgepodge of mathematics without data, of history of economic institutions without mathematics or theory, of an ideal model of psychological man … (Campbell 1969a)

Campbell took a breathtaking leap when he conceived the 'myth of unidisciplinary competence,' demonstrating how social, administrative, structural, and political supports which accrue over time around arbitrary disciplines ultimately create incentives that reinforce these arrangements and punish others of equal epistemological worth. Campbell detailed the inevitable consequence—biased knowledge production—and offered a model of 'interdisciplinary narrowness' for mitigating this 'ethnocentrism of disciplines.'

By the late 1960s, following Vietnam War protests, campus unrest, and a national focus on urban inequities, Campbell began trying conscientiously to influence the design, implementation, and evaluation of social policies. In 'Reforms as Experiments' (1969b) Campbell introduced a range of real-world quasi-experiments (attempts to approach the standards of experimentation even where random assignment of subjects to conditions is unachievable)—each tested against standards of 'internal and external validity'—to the practical world of social reform. That article frequently is cited as the most important single work in the field now known as 'program evaluation.'

In a brilliant tongue-in-cheek section of this article Campbell advised 'trapped administrators whose political predicament will not allow the risk of failure' how to capitalize on 'threats to validity' to assure positive results (Campbell 1969b). For example, were they to accept the occasional 'grateful testimonial' as if it were a representative outcome, to reserve interventions for carefully selected subpopulations most likely to succeed, and to eliminate from their analyses all those who prematurely quit the program, 'successful' program outcomes could be virtually assured.

Campbell continued in various fora to insist that social scientists have both obligations and opportunities to test the relevance of their theories and methods in the service of public good. In this, Campbell's unswerving concern was with the biases affecting the generation of knowledge. For example, in his too-infrequently cited 'Assessing the Impact of Planned Social Change' (1975) Campbell again displayed remarkable methodological and sociological insight in generalizing the problem of data validity in the policy arena in the following way:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social processes it is intended to monitor (Campbell 1975).

Corrupted indicators remain a persistent and only vaguely explored source of real-world invalidity (e.g., Cochran et al. 1980).

Another highly influential, widely circulated piece went unpublished for nearly 20 years: in 'Methods for the Experimenting Society' Campbell attempted specifically to design a strategy for a society willing to test innovations and their intended effects. He argues for a society that would

> vigorously try out possible solutions to recurrent problems and would make hard-headed, multidimensional evaluations of outcomes, and when the evaluation of one reform showed it to have been ineffective or harmful, would move on to try other alternatives (Campbell, in Overman 1988).

Surveying the multitude of societal reforms underway at the time, many of which he endorsed (e.g., Alexander Dubcek in Czechoslovakia, Salvador Allende, in Chile), Campbell argued:

> There is no such (experimenting society, committed to reality testing, self-criticism, to avoiding self-deception) anywhere today. While all nations are engaged in trying out innovative reforms, none of them are yet organized to adequately evaluate the outcomes of these reforms (Overman 1988).

After specifying requirements of such an 'experimenting society' Campbell attempted in various ways to determine conditions under which these results might be achievable. For example, after Allende was assassinated in Chile, Campbell invited Chilean social scientist Ricardo Zuñiga to Northwestern University

specifically to reflect on the conditions that might allow vigorous, large-scale reforms to be tested.

Campbell made other major contributions in his application of evolutionary theory to the realm of ideas. Campbell, the eternal epistemologist, focused in particular on the joint operation of the 'blind-variation-and-selective-retention process fundamental to all inductive achievements, to all genuine increases in knowledge.' In his 'evolutionary epistemology,' which perhaps received its fullest flowering in his essay honoring Karl Popper (in Overman 1988), Campbell endorsed the natural selection paradigm for explaining the production of knowledge. Through the selective winnowing of ideas, Campbell argued, what survive over time are those ideas that are most meritorious, particularly as each is subjected to 'critical realism.' In his William James lectures, Campbell (1977) summarized his perspective in this way:

> I am a fallibilist and antifoundationalist. No part of our system of knowledge is immune to correction. There are no firm building blocks, either as indubitable and therefore valid axioms, or in any set of posits that are unequivocal once made. Nor are there any unequivocal experiences or explicit operations on which to found a certainty of communication in lieu of a certainty of knowing.
> I am some kind of a realist, some kind of a critical, hypothetical, corrigible, scientific realist. But I am against direct realism, naive realism, and epistemological complacency (Campbell, in Overman 1988).

2. *Legacy*

Campbell's legacy is vast and pronounced. It affects a wide range of social scientists as well as public policy scholars and practitioners, and is reflected in the work of hundreds of his students and more distant heirs. He introduced fundamentally new ways of thinking that are taken for granted today about how we perceive, discover, and assess accumulated knowledge. Many of his concepts are now fundamental throughout the social and policy sciences. Campbell is the most important intellectual figure in the history of evaluation research. The strategies he employed in 'Reforms as Experiments' for unearthing and illustrating threats to interpretations of policy implementation are an often-cited model for exploring complex data (see Tufte 1983).

In remarkably vivid, fetching but precise language, Campbell repeatedly captured the central importance of fallibilities inherent in human observers and their methods as they accumulate knowledge about the world, and thus of the importance of multiple perspectives, the triangulation of methods, and an open process whereby ideas are tested against one another to promote the winnowing effects of competition. Thus, in the James lectures, Campbell stated that

> We are cousins to the amoeba, and have received no direct revelation not shared with it. How then, indeed, could we know for certain? … Our only hope as competent knowers is that we be the heirs of a substantial set of well-winnowed propositions (Campbell, in Overman 1988).

In the heady intellectual mix he created and sustained, Campbell was known for changing his mind when persuaded by carefully tested evidence, and for challenging others to do the same. For example, over the years Campbell was persuaded by sociologist Howard S. Becker and others to appreciate the falsifiability of carefully wrought case studies, in spite of their failure to live up to his long-cherished methodological criteria (Campbell, in Overman 1988). In return, he prompted qualitatively-oriented colleagues to persist in reconsidering the validity claims inherent in their own work (see, for example, Becker 1986). For this open-mindedness, Campbell is revered by quantitative and qualitative methodologists alike (see Patton 1999).

Campbell also influenced legions of colleagues to reflect honestly on their own work. He talked and wrote openly about his own insecurities and frustrations, including how late in his career he began to write, droughts in his own productivity, his discomfort with the production pressure and relentless expectation of genius when he was a young assistant professor at the University of Chicago, his frequent bouts of depression, and the times when he believed his students and colleagues 'carried him.'

He explained his messy desk cheerfully by saying that he 'filed archeologically,' and as long as no one moved things around he could find whatever he wanted. Amazed colleagues would watch him reach into a pile and pull out the precise papers relevant to their discussions.

Although its publishers have sold thousands of copies of *Experimental and Quasi-Experimental Designs for Research* and students return to it year after year, Campbell never received any royalties; he asked only for unlimited copies to give freely to his colleagues.

Consideration of Campbell's legacy would be incomplete without a discussion of his personal style, for it, too, promoted the production of knowledge while providing his colleagues with a vital and successful paradigm. It would understate Campbell's colossal influence if there were no mention of his towering warmth, kindness, and generosity, and his deeply honest and self-critical perspective on his own work (Campbell 1981). Campbell was remarkably able to reflect on his own fallibilities and to create conditions that would model his epistemology, and improve both his work and the work of others. Examples are diverse and legendary: he would often send a requested reprint accompanied by a colleague's critical rejoinder. He was always in pursuit of contrary evidence and counter-intuitive examples, particularly if they confronted his own work. He encouraged students to write minority sections to project reports if they

disagreed with him or with an emerging majority. Campbell was resolutely encouraging of others and in fact was thought by some to 'suffer fools gladly.' He gave three grades: A, B, and incomplete. If the data collected in dissertations failed to support anticipated hypotheses, or failed to replicate published findings in the literature, Campbell urged students to write up the apparent reasons for the obtained results.

Over time, Campbell connected his early deep philosophical explorations to a program of practical work that will continue to guide generations of scholars and practitioners. Campbell lives on through his voluminous work in many fields, through his students, and through their adoption of his style.

See also: Evolutionary Epistemology; Experimental Design: Large-scale Social Experimentation; Experimental Design: Randomization and Social Experiments; Experimentation in Psychology, History of; External Validity; Foucault, Michel (1926–84); Internal Validity; Panel Surveys: Uses and Applications; Pearson, Karl (1857–1936); Quasi-Experimental Designs; Reform: Political; Unobtrusive Measures

Bibliography

Brewer M, Collins B 1981 *Scientific Inquiry and the Social Sciences*. Jossey-Bass, San Francisco

Campbell D T 1950 The indirect assessment of social attitudes. *Psychological Bulletin* **47**(1): 15–38

Campbell D T 1969a Ethnocentrism of disciplines and the fish-scale model of omniscience. In: Sherif M, Sherif C (eds.) *Interdisciplinary Relationships in the Social Sciences*. Aldine, Chicago

Campbell D T 1969b Reforms as experiments. *American Psychologist* **24**: 409–29

Campbell D T 1975 Assessing the impact of planned social change. In: Lyons G (ed.) *Social Research and Public Policies*. Public Affairs Center, Dartmouth College, Hanover, NH

Campbell D T 1977 *Descriptive Epistemology: Psychological, Sociological, and Evolutionary. William James Lectures*. Harvard University, Cambridge, MA

Campbell D T 1981 *Another Perspective on a Scholarly Career*. In: Brewer M B, Collins B E (eds.) *Scientific Inquiry and the Social Sciences*. Jossey-Bass, San Francisco

Campbell D T, Fiske D A 1959 Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin* **56**: 81–105

Campbell D T, Stanley J C 1963 Experimental and quasi-experimental designs for research on teaching. In: Gage N (ed.) *Handbook of Research on Teaching*. Rand McNally, Chicago

Cochran N, Gordon A C, Krause M 1980 Proactive records: Reflections on the village watchman. *Knowledge* **2**(1): 5–18

Cook T D, Campbell D T 1979 *Quasi-experimentation Design and Analysis Issues for Field Settings*. Rand McNally, Chicago

Jacobs R C, Campbell D T 1961 The Perpetuation of an arbitrary tradition through several generations of a laboratory microculture. *Journal of Abnormal & Social Psychology* **62**: 649–58

Kidder L, Campbell D T 1970 The indirect testing of social attitudes. In: Summers G (ed.) *Attitude Measurement*. Rand McNally, Chicago

Overman E S (ed.) 1988 *Methodology and Epistemology for Social Science: Selected Papers of Donald T. Campbell*. Chicago

Patton M Q 1997 *Utilization–Focused Evaluation*, 3rd edn. Sage, Beverly Hills, CA

Segall M H, Campbell D T, Herkovits M J 1966 *The Influence of Culture on Visual Perception*. Bobbs-Merrill, Indianapolis, IN

Tufte E R 1983 *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT

Webb E J, Campbell D T, Schwartz R D, Sechrest L 1966 *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Rand McNally, Chicago

A. C. Gordon

Cancer-prone Personality, Type C

Type C has emerged as a behavioral pattern, coping style, or personality type that predisposes people to, or is a risk factor, in the onset and progression of cancer. Type C has been described as being over-cooperative, stoical or self-sacrificing, appeasing, unassertive, patient, avoiding conflict, compliant with external authorities, unexpressive, suppressing or denying negative emotions, self-sacrificing, and predisposed to experiencing hopelessness and depression (Bleiker 1995, Eysenck 1994, Temoshok 1990).

Since the mid-twentieth century, the contribution of psychosocial factors to cancer has been an important research topic. A search of the literature shows an exponential growth: while only nine publications appeared from 1951 to 1955, between 1995 to 1999 more than 300 hundred documents dealing with this topic have been published in scientific journals (Psychlit and Medline database). This article deals with the 'state of the art' of Type C and its role as a risk factor in the onset and progression of cancer.

1. Historical Overview

The belief that personality contributes to the etiology of disease is as old as the history of human thought. In the third millennium BC, Hippocrates stated that melancholic humor (caused by a surplus of black bile) caused cancer, and Galen (second century AD) observed that neoplasm was more likely to occur in melancholic and depressed than in sanguine women.

It is in the twentieth century that experimental evidence about relationships between cancer and personality are found. During the first half of the twentieth century, investigations were made from the perspective of psychoanalytical theory, and within the field of psychosomatic medicine. The most important

characteristics in cancer patients found through empirical research mentioned: loss of important relationships, inability to express negative emotions, tension concerning parental relationships, and sexual problems.

Finally, it is from the second half of the twentieth century that the role of behavioral and psychosocial conditions in the onset and progression of cancer has been investigated systematically in research programs, groups, and laboratories, following a track parallel to research in Type A, Coronary prone behavior. A good deal of clinical observations and research data has been produced about a systematic link between some psychosocial factors and cancer. At the same time, there have been comprehensive reviews of psycho-oncology research, and strong criticism of its methodological flaws. Thus, before presenting the main traits constituting Type C, the most important difficulties involved in its research will be discussed.

2. *Difficulties in the Study of Cancer-prone Personality, Type C*

Temoshok and Heller (1984) described Type C research in a very illustrative fashion, as 'comparing apples, oranges and fruit salad,' expressing its complexity and methodological difficulties. In other words it involves different constructs, assessed through different instruments, in different samples of subjects with different types of cancer and other diseases, and using different designs.

The heterogeneity of factors contributing to the noncomparability of psycho-oncology literature can be considered in the following ways: (a) the nature and measurement of psychological phenomena, (b) cancer characteristics (type, site, and stage in the process), (c) sample characteristics, and (d) designs used.

2.1 *The Nature and Measurement of Psychological Phenomena*

Personality type, coping style, stress or emotional response, or behavioral pattern are constructs with a variety of conceptualizations in psychology. This problematic issue has been overcome by Eysenck (1994); in short, the subject's response to a given stressful situation cannot be understood without taking into consideration their stressor perception and the behavioral pattern used in coping with stress—both are stable personality conditions.

To this conceptual problem, a methodological one should be added. As constructs, all types of variables must be operationalized through psychological measures. From semistructured interviews and projective techniques to well-developed questionnaires, hundreds of measures have been used in psycho-oncology research, which have different psychometric properties and different levels of generalizability, and their scores can, therefore, be generalized to different universes, presenting difficulties for the comparability of results. Over recent decades much more attention to Type C measurement has been paid, and these methodological problems have been partially overcome.

2.2 *Cancer Characteristics*

Cancer is not a homogeneous entity; both cancer types (melanoma, sarcoma, etc.) and cancer sites (lung, breast, skin, etc.) vary with regard to etiology, course, mortality, heritability, and risk factors. Also, cancer always implies a process, and psychological factors may act at different stages of this process; this fact introduces important research problems. For example Kreitler et al. (1993) stated that emotional repression may be a response to the threat posed by the cancer diagnosis. Several studies have dealt with the timing of psychological assessment and its relationship to cancer, but the relationships between psychosocial conditions and neoplastic processes require sophisticated and complex research designs, and these have yet to be fully developed.

2.3 *Subjects/Sample Characteristics*

Several authors have pointed out that sociodemographic variables may be associated with medical and environmental risk factors. Among sociodemographic characteristics, age seems to be the most important, since cancer incidence and prevalence increases with age, and several personality characteristics are also associated with age (e.g., introversion). When cancer groups are compared with other participants, age should be controlled (e.g., Fernández-Ballesteros et al. 1998).

In order to test the psychological risk factors of a given disease or illness through psychological covariation, comparison between healthy and ill subjects is not enough. It is essential to distinguish also between the target illness and other illnesses. Cancer patients have been compared, as well as with normal patients, with patients suffering from cardiovascular and digestive problems, chronic diseases and benign pathologies, and with accident victims, among others.

2.4 *Designs*

Science requires strategies or designs for testing hypotheses; that is, when and under what conditions units (subjects) will be observed and measured. Authors agree that the best design for studying a given risk factor is a prospective one. Subjects (recruited at random in the community or as a member of a given

group) are assessed in the relevant target behaviors and psychological constructs, and are monitored until some of them develop cancer and others do not. Also, in longitudinal studies subjects already diagnosed with cancer are monitored during different stages of the illness process, from diagnosis to progression, treatment, and survival or death.

In quasiprospective designs psychological evaluation is carried out before cancer diagnosis, but this period is usually very short. This type of design is commonly used within cancer prevention programs in which all subjects are assessed in psychosocial characteristics and, after diagnosis, subjects with cancer are monitored. This type of design has two main flaws: psychological characteristics are measured when subjects have already developed cancer, and subjects attending a prevention campaign cannot be considered to be in a neutral context. Furthermore, this situation is quite heterogeneous, since while several subjects may already be suspicious about lumps or lesions, others are merely participating in a preventive—or 'normal'—situation.

In retrospective designs, subjects already diagnosed or being treated for cancer are compared with control groups in a set of psychological characteristics. Such retrospective designs are highly criticized because they do not tell us whether psychological characteristics are the cause or the result of cancer (e.g., Kreitler et al. 1993).

All of these designs can be considered as 'descriptive' or 'correlational' research strategies. However, there are also experimental designs; in these, the 'independent' variable (a psychological or personality condition) is manipulated in order to measure its causal or functional relationship with the 'dependent' variable (cancer).

In sum, several of these conceptual and methodological problems have been overcome, while others continue to be the object of criticism.

3. *Personality Characteristics in Cancer Patients*

Evidence from prospective, quasiprospective, retrospective, and experimental studies converges on a set of psychological characteristics that seem to act in the onset and progression of all types of cancer. Let us examine these characteristics.

3.1 *Emotional Expression*

This appears to be at the core of Type C. That is, suppression, repression, inhibition and/or denial of negative emotions seem to be the central characteristics of the cancer-prone personality (see *Emotions and Health*). Cancer patients (in comparison to controls) are described as anti-emotional and alexitimic, with a tendency to control their negative feelings (mainly anger, aggressiveness, hostility); they report not feeling anger or irritation, and in their relationships they report avoiding arguments with others by using reason and logic, often contrary to their feelings. For example, in a quasiprospective study, within a breast cancer prevention program, Fernández-Ballesteros et al. (1997) compared the rationality (as well as other somatic characteristics) of healthy women ($N = 96$), women with benign breast disease ($N = 90$) and women with breast cancer ($N = 122$). Healthy women were matched with women with benign breast pathology or with breast cancer. All were assessed before diagnosis. Results show that women with breast cancer differed significantly from healthy women and those with benign breast pathology. Any other review on this topic yields the same results: emotional expression is one of the psychological hallmarks of cancer (Bleiker 1995, Eysenck 1994, Spiegel and Kato 1996, Temoshok 1990).

Unfortunately, however, experimental data does not describe the nature of these psychological characteristics, nor whether cancer subjects fail to perceive or feel their emotional internal physiological reactions or whether they simply fail to express them (verbally or otherwise).

3.2 *Depression*

Consideration of medical anamneses also suggests that depression is a frequent precursor of cancer; feelings of helplessness and hopelessness are also found in several studies. For example, the Western Electric Health longitudinal study monitored 2,020 employees with the purpose of investigating coronary disease risk factors. As a measure of personality and personal and social adjustment the Minnesota Multiphasic Personality Inventory (MMPI) was administered. At the 17-year follow-up, 4 percent had died of cancer. As reported by Shekelle et al. (1981), there was a twofold increase in the odds of death from cancer in men with psychological depression. This result was consistent across sites and types of cancer. Also, this association persisted after adjustments had been made for age, cigarette smoking, use of alcohol, family history of cancer, and occupational status (see *Depression*, *Hopelessness*, *Optimism*, *and Health*).

In spite of these impressive results, it is unclear whether these states are antecedents or consequences of cancer, and research data is inconsistent in this area.

3.3 *Interpersonal Style*

In the medical literature, cancer patients have been described by their doctors as 'extremely pleasant' in their interactions with them. There is evidence from retrospective, quasiprospective, and prospective studies, as well as from experimental studies, that cancer patients—in comparison to control subjects—yield significantly higher scores in tests assessing need

for harmony, self-sacrifice, and unassertiveness. For example, in a quasiprospective study Fernández-Ballesteros et al. (1998) compared, in terms of need for harmony (as well as rationality), three groups of women with breast cancer (before diagnosis (74), during treatment (105), and during the follow-up (132)) with healthy women. Both personality variables correctly classified 86 percent of the participants (87 percent of the breast cancer patients and 82 percent of the healthy participants) and, moreover, no significant differences in rationality and need for harmony were found to be related to the stage of the cancer process.

This behavioral pattern can be understood as a coping style for stressful situations, as a personality characteristic, or as a defense mechanism against anxiety.

3.4 Stressors, Stress, Strain, and Coping Mechanisms

Cancer patients' anamneses usually show that an antecedent of cancer disease is a subject's life event. However, as already mentioned, it is important to differentiate between the external stressor, the subject's appraisal and coping style, and the stress reaction (see *Stress and Health Research*). As underlined by several authors, stress is the response of a given subject to their perception of a stressor—a stimulus or situation perceived as threatening. Also, however, subjects have different mechanisms for coping with stress. For example, Cooper and Faragher (1993) pointed out that subjects who experienced a stressful situation they perceived as threatening had a significantly higher cancer risk. In a retrospective study, Watson et al. (1984) compared breast cancer patients and healthy controls. Breast cancer patients were more likely than controls to report a tendency to control emotional reactions, particularly anger.

However, evidence about psychosocial influences on cancer incidence and progression comes not only from descriptive studies—there are also important experimental results. For example, Grossarth-Maticek carried out an experimental study within the Heidelberg prospective study. In 1971, 1,026 healthy persons from a random sample of the population were assessed in a wide set of psychosocial and medical variables. Also, 1,537 subjects were selected on the basis of their increased psychosocial and clinical risk of cancer or cardiovascular or circulatory disease. Criteria used included chronic depression and hopelessness, chronic excitement and anger, heavy smoking, hypertension, high blood cholesterol, and high blood sugar. From the 1,537 high-risk subjects, 100 were selected who were thought to have an especially high risk of cancer (and also of CV disease). Half of these subjects were selected randomly and given psychological therapy intended to decrease the respective risk. This therapy, which has been called 'creative novation behaviour therapy,' was developed by Grossarth-Maticek and Eysenck (1990) on the basis of cancer-prone characteristics (emotional expression, assertiveness, coping, etc.), and it was attempted to manipulate them through behavior therapy techniques. In 1983, mortality and causes of death for the 2,563 subjects of the study were ascertained. Of the selected subjects, 38 percent of risk subjects had died, as compared to 10 percent of the random sample.

In the cancer therapy group deaths were significantly fewer than in the control group, and there were no deaths from cancer. In marked contrast, 60 percent of the deaths in the control group were from cancer. Although Grossarth-Maticek's studies have been criticized for certain methodological shortcomings (questionnaires used were not in a standard form, and there was no report on psychometric properties), their impressive results are in accordance with the well-established effects of psychosocial treatment, which significantly increases survival in cancer patients (e.g., Spiegel and Kato 1996).

From descriptive and experimental studies it can be summarized that cancer-prone or Type C personality is a psychological configuration with two main conditions: (a) suppression of negative emotions (such as anger and anxiety), which includes not only lack of verbal expression of emotions but also corresponding interpersonal behaviors and (b) inappropriate stress coping mechanisms, leading to feelings of hopelessness, helplessness, and depression.

Finally, what mechanisms can be assumed to link psychological characteristics with cancer?

3.5 Mechanisms that Link Psychological Factors with Cancer

From a mentalistic and/or dualistic perspective, psychological characteristics observed in cancer patients have been considered as though the 'soul' (or mental entities) were affecting the body. This position has been brought up to date through empirical evidence from both psychological and biological literature that support the relationship between personality and emotions and the autonomic, endocrinological and immune systems.

O'Leary (1990), in her review of stress response and the immune system, referred to the fact that chronic stress is associated with a suppression of the immune function, and that personality and coping styles may enhance or degrade the immune response. As pointed out by Spiegel and Kato (1996), chronic and acute stress are associated with reductions in various measures of the immune function, and, moreover, psychological intervention positively influences hallmarks of the immune system.

In sum, as Spiegel and Kato (1996) conclude, 'there is a nonrandom relationship among various psychosocial factors and cancer incidence and progression that can only partially (underline added) be explained by behavioral, structural, or biological factors.'

4. Conclusions

Throughout the twentieth century, the empirical study of the relationship between personality and cancer has been an important research topic. During the 1980s and 1990s, several conceptual and methodological problems have been overcome.

Various designs, assessment devices, and samples have shown that a set of psychological characteristics appears to play a role in cancer onset and progression. These characteristics can be reduced to the suppression of negative emotions and to inappropriate stress coping mechanisms. Much more attention should be paid to the psychological nature of these characteristics.

Psychological treatments (manipulating the above psychological characteristics) appear to act on the cancer process, preventing cancer onset and increasing patient survival possibilities.

Scientific literature supports the hypothesis of a link between personality characteristics, stress reaction coping styles and biological systems. That is, there is evidence of a connection between personality, stress and cancer, as well as between personality, stress, and the autonomic, endocrinological, and immune systems.

These psychological characteristics can be considered as cancer risk factors. Nevertheless, Type C or cancer-prone personality should be understood in terms of its synergic interaction with other genetic, biological, and environmental conditions.

See also: Cancer: Psychosocial Aspects; Cancer Screening; Depression, Hopelessness, Optimism, and Health; Emotions and Health; Gender and Cancer; Personality and Health; Stress and Health Research

Bibliography

Bleiker E 1995 *Personality Factors and Breast Cancer*. ICG, Dordrecht, The Netherlands

Cooper C L, Faragher E B 1993 Psychosocial stress and breast cancer: the interrelationships between stress events, coping strategies and personality. *Psychological Medicine* **23**: 653–62

Eysenck H J 1994 Personality, stress, and cancer: Prediction and prophylaxis. *Advances in Behavior Research and Therapy* **16**: 167–215

Fernández-Ballesteros R, Ruiz M A, Garde S 1998 Emotional expression in healthy women and those with breast cancer. *British Journal of Health Psychology* **3**: 41–50

Fernández-Ballesteros R, Zamarrón M D, Ruiz M A, Sebastián J, Spielberger C D 1997 Assessing emotional expression. *Personality and Individual Differences* **22**: 719–29

Greer S, Watson M 1985 Towards a psychobiological model of cancer: Psychological considerations. *Social Science and Medicine* **20**: 773–7

Grossarth-Maticek R, Eysenck H J 1990 Prophylactic effects of psychoanalysis on cancer-prone and coronary heart disease-prone probands, as compared with control groups and behaviour therapy groups. *Journal of Behavioral Therapy and Experimental Psychiatry* **21**: 91–9

Kreitler S, Chaitchik S, Kreitler H 1993 Repressiveness: cause or results of cancer? *Psycho-oncology* **2**: 43–54

O'Leary A 1990 Stress, emotion, and immune function. *Psychological Bulletin* **108**: 363–82

Shekelle R B, Raynor W J, Ostfeld A M, Garron D C, Bieliauskas L A, Liu S C, Maliza C, Paul O 1981 Psychological depression and seventeen-year risk of death from cancer. *Psychosomatic Medicine* **43**: 117–25

Spiegel M D, Kato P M 1996 Psychosocial influences on cancer incidence and progression. *Harvard Review of Psychiatry* **4**: 10–26

Temoshok L 1990 On attempting to articulate the biopsychosocial model: Psychological-psychophysiological homeostasis. In: Friedman H (ed.) *Personality and Disease*. Wiley, New York

Temoshok L, Heller B W 1984 On comparing apples, oranges and fruit salad: A methodological overview of medical outcome studies in psychosocial oncology. In: Cooper C L (ed.) *Psychosocial Stress and Cancer*. Wiley, New York

R. Fernández-Ballesteros

Cancer: Psychosocial Aspects

1. Introduction

Psychosocial research in oncology has been conducted since the 1950s. The initial interest in cancer onset has changed, perhaps due to the many methodological problems, to research in psychosocial consequences of a cancer, and the impact of psychosocial factors in cancer progression. Although there are some studies, in several fields of research, which show negative results, most well-controlled studies have shown positive correlations between specific psychosocial factors and quality of life during cancer treatment, cancer progression, and survival.

2. Quality of Life in Cancer Patients

The concept of quality of life (QOL) was first referenced in *Index Medicus* in 1977. Since that time, it has emerged as an important concept in cancer research in the description of conditions in cancer diagnosis and treatment that make the quality of life of patients worse, and in the use of quality of life measurements for the evaluation of new methods of treatment. QOL is defined as a concept which refers to the individual's own perceptions about the degree of satisfaction with treatment, and ability to perform in their daily life. Although impairment in several dimensions of quality of life are primarily caused by biological aspects of the disease, psychosocial research has focused on psychosocial factors that additionally modulate the intensity of subjectively experienced impairments in quality of life. Extensive psychosocial research has been done with cancer patients in the

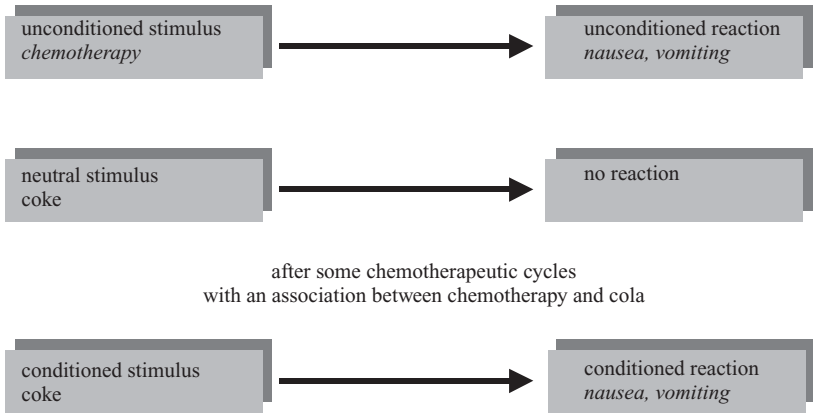

Figure 1
Learning model of classical conditioning, of anticipatory nausea and vomiting. See text for further details

initial phase of diagnosis and treatment, whereas little information is available about the problems and concerns that persist for long-term survivors. It can be assumed that, for all cancer survivors, what begins as a crisis involving diagnosis and treatment gradually becomes a chronic illness with life-long follow-up medical care, long-lasting psychological effects, and changes in social and employment relationships.

2.1 Physical Symptoms

2.1.1 Treatment-related nausea and vomiting. It is estimated that 75 percent of the patients receiving chemotherapy will suffer from nausea or vomiting in spite of an antiemetic therapy (Morrow 1992). Nearly 25 percent of patients suffer from so-called anticipatory nausea. In these cases nausea or vomiting occur prior to the infusion of chemotherapy and is triggered by particular special odors or situations similar to stimuli experienced in previous treatment. Whereas post-chemotherapy nausea and vomiting are caused by neuropharmacological agents, anticipatory nausea and vomiting are assumed to be caused by classical processes of conditioning. Neutral stimuli (e.g., the taste of a drink like cola) that were associated with the infusion of chemotherapy (cola was mixed into the antiemetic drug that was given prior to chemotherapy), will be associated with post-chemotherapy nausea. As a consequence, its taste will release nausea even if the neuropharmacological effects of chemotherapy have passed. Taste becomes a learned stimulus for anticipatory nausea or vomiting (see Fig. 1).

Several predictors were identified as high risk factors for the development of nausea and vomiting during cancer chemotherapy: (a) antiemetic drugs; (b) physiological factors, such as age, constitutional predisposition; and (c) psychological factors, such as treatment-related state anxiety or general psychological distress (Andrykowski and Gregg 1992). There are various hypothesized pathways by which a state of anxiety or distress may contribute to the experience of treatment-related nausea. First, due to the proximity of structures in the higher brain stem responsible for the release of vomiting, and noradrenergic structures that are influenced by anxiety and distress, it can be assumed that distress triggers release and intensity of nausea and vomiting through a higher noradrenergic activity. Second, it is well known from learning psychology that anxiety and distress may facilitate the process of classical conditioning. Therefore it can be assumed that distress may enhance the conditioned anticipatory nausea and vomiting.

2.1.2 Pain. Pain is one of the most prevalent physical complaints of cancer patients. Bonica published the most comprehensive overview in 1985 (see Fig. 2). Fifty to 70 percent of cancer patients suffer from pain syndromes during the late phases of their disease; 20 to 50 percent suffer from cancer- or treatment-related pain even in early phases of their disease. In 60 to 80 percent of the cases pain syndromes are attributable to the malignancy of the tumor (tumor compression, infiltration of nerves); in 10 to 25 percent they are long-term results following operation, chemo- or radiotherapy; in 3 to 10 percent of cases the pain syndromes can be assumed to be unrelated to the cancer or treatment (e.g., migraine, arthritis). In many cases, however, it is difficult to discover the main cause of pain. It can be assumed that several psychological factors, known modulators in nonmalignant pain also influence the intensity and aversiveness of cancer-related pain. Among others, factors of depression, chronic daily stress in private life or at work, maladaptive pain-related patterns of

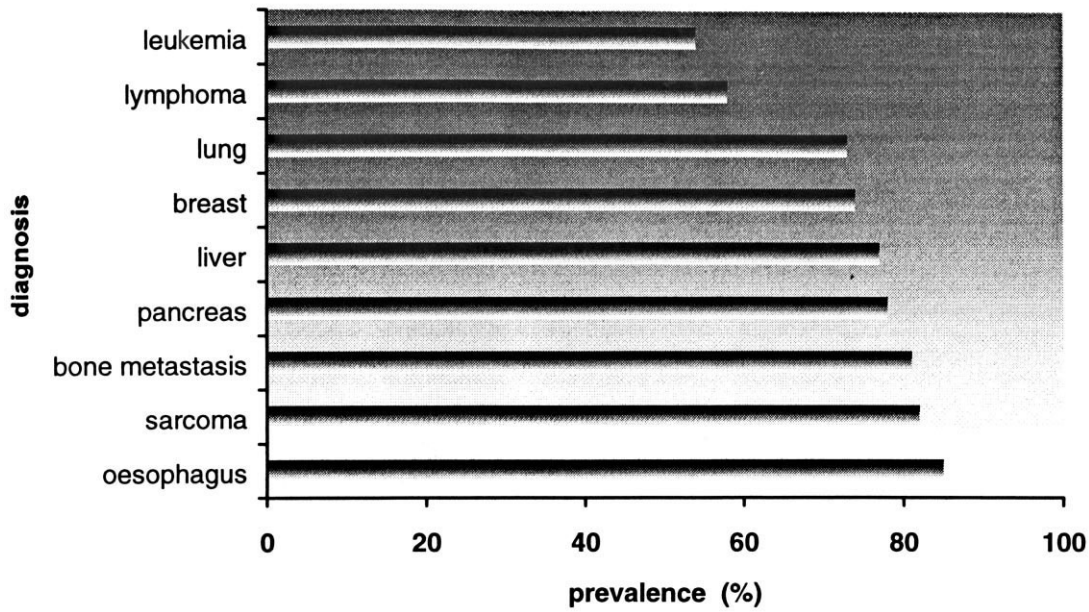

Figure 2
Prevalence of pain in different cancer diagnoses (see Bonica 1985).

thought like catastrophizing or suppressive cognitions; maladaptive pain coping strategies belong to the most important psychological factors that contribute to the chronicity of pain, e.g., chronic low pack pain (Hasenbring 1998).

2.1.3 Fatigue. Only in the last years of the twentieth century did cancer research focus on fatigue as another cancer- or treatment-related physical complaint; a national survey in the USA showed a prevalence of problems with fatigue in 76 percent of 419 chemotherapy and/or radiation-treated patients, with impairment in walking, exercising, house cleaning, or climbing stairs (Curt 2000). Cancer-related fatigue had also a significant emotional component, with 90 percent of patients reporting that fatigue contributed to a loss of control, hopelessness, isolation, lack of motivation, sadness, and frustration. A significant number of patients also reported fatigue-related social impairment, which included difficulties in shopping, expressing intimacy with a loved one, playing with children, or spending time with friends. Curt has shown that the physician recommendations tended to be nonspecific: 40 percent of physicians recommended doing nothing, 37 percent prescribed rest. Further research should focus on the biological and psychological mechanisms by which fatigue is produced and maintained, in order to develop effective interdisciplinary treatment modalities.

2.2 Body Image

Impairments of body image are mostly a direct consequence of cancer treatment. Significant impairment was seen in patients with a tumor of the head, neck, or larynx; in breast cancer patients with a resection of the mamma; and in patients with colorectal cancer and an artificial anal/sphincter praeter. Between 10 and 56 percent of the breast cancer patients develop problems in sexuality and in their partnerships. More than 50 percent of the patients with cancer of the head or neck experience a decrease in self-esteem, combined with anxiety and depression, which leads to withdrawal and social isolation (Holland and Rowland 1998).

2.3 Physical Functioning

A number of cancer patients suffer from impairment in sexual functions, in the short- and/or long-term, after cancer treatment. For instance, 75 percent of women with cancer of the cervix showed intense, continuous sexual difficulties; 33 percent gave up their sexual contacts. Patients with cancer of the prostate experience an impaired erection, which is a generally reversible consequence of injury to the autonomous nerves. Anxiety, dysfunctional patterns of thought, and feelings of shame are psychological factors eliciting avoidance behavior, which leads to persistent sexual difficulties. A study of 407 long-term survivors of bone-marrow transplantation (BMT) found sexual

problems in 29 percent of the men and in 80 percent of the women (Syrjala et al. 1998). The authors recommend behavioral as well as hormonal or mechanical therapy.

2.4 Emotional Distress

Confrontation with a diagnosis of cancer is experienced as a crisis, which leads to distress in the form of, for instance, expectations of chronic suffering, acute, or long-lasting fear of death, and feelings of anxiety, depression, or anger. Patients with dysfunctional coping strategies are at risk of developing chronic anxiety and other feelings of distress; patients with intense or long-lasting emotional distress suffer more from the physical complaints described above. Extensive clinical studies have shown that 20 to 40 percent of cancer patients suffer from intense or long-lasting emotional distress (Holland and Rowland 1998). Some studies have seen intense feelings of anxiety in 80 percent of the patients immediately subsequent to cancer treatment (e.g., Hasenbring et al. 1993). Patients suffered from fear of pain caused by unfamiliar methods of treatment, by claustrophobia during computed tomography, or by anxiety because they did not know what the next steps of their therapy would entail.

2.5 Cognitive Functions

Clinical impressions suggest that a significant number of cancer patients suffer from impairment to various cognitive functions, although only a few studies have assessed cognitive functioning in cancer patients objectively. Cull et al. (1996) found 49 percent of cancer patients, with various diagnoses, suffering from impairment of concentration and memory. Cancer itself (e.g., tumor of the brain, brain metastases in breast cancer patients) and its treatment (neurotoxicity of cytostatic therapy) are known causes of cognitive impairment. Anxiety, depression, fatigue, and low psychosocial support are further possible causes of impairment of cognitive functions.

2.6 Social Relations

Fear of death, and impairment of emotional and physical functioning of cancer patients, lead to differing degrees of disturbance in their social relationships. Social roles and responsibilities in a family have to be redefined where patients suffer from significant complaints. Manne et al. (2000) reported, in a cross-sectional study with 219 cancer patients and their spouses, a significant relationship between increasing patient functional impairment and spousal negative behaviors related to greater restrictions on the activity of the spouse and by depression.

3. Coping With Cancer-related Distress

Coping research in cancer patients has focused on cognitive and behavioral strategies that patients employ to reduce emotional distress or to change a distressing situation. A major question in cancer research has been whether specific coping responses would decrease or increase survival time. Results of different prospective cohort-studies (i.e., with two or more assessment points) indicate that several aspects of avoidance behavior, for instance, avoidance of distressing social contacts (social withdrawal), avoidance of expressing feelings of fear, anger, or regret, as well as the minimization or denial of the potential threat, are predictors of reduced survival time. For instance, Greer et al. (1979) conducted a retrospective study in women with breast cancer including a ten-year follow-up. They found that women with an active, confronting coping behavior ('fighting spirit') showed an increased survival time when compared to those who showed denial or stoic acceptance behavior. Women with signs of help-/hopelessness showed the lowest survival time. The tendency to minimization and anger were independent predictors of survival time in a further prospective study in 125 patients with metastatic melanoma, published by Butow et al. (1999).

4. Social Support

There has been extensive research into the role played by positive social relationships in the adaptation to stress. Among the variables (e.g., the number of available persons, frequency of contact with relatives, friends, etc., perceived social support, and degree of satisfaction with perceived support), perceived social support seems to be the best predictor of adaptation to cancer. Other qualitative dimensions are emotional support, instrumental, self-esteem, and appraisal support. Social support might influence patients' ability to adapt to the illness and its treatment ('stress-buffer hypothesis') or might have a direct influence on the progression of the disease ('main-effect hypothesis'). Retrospective studies in cancer patients have revealed some evidence of a positive relationship between social support and progression of the disease (Spiegel and Kato 1996). Nevertheless, there is also some evidence from recent studies for differential aspects. In a group intervention study, Helgeson et al. (2000) found that peer discussion groups in 230 women with breast cancer were helpful for women who lacked support from their partners or physicians, but harmful for

women who had high levels of support. More research is needed to clarify the relationship between special aspects of social support, social networks, and the adaptation to the disease.

5. *The Impact of Psychosocial Factors on Cancer Progression*

Whereas psychosocial predictors of the onset of cancer have not yet been reliably identified (at the end of the twentieth century), perhaps due to methodological problems, the literature supports the idea that psychosocial variables predict cancer progression. Besides the coping styles, and the quality and amount of social support, as described above, cancer-independent distress in daily life has also been investigated in relation to cancer progression. In a study of 86 breast cancer patients, Forsén (1991) interviewed the women after surgery. Patients with more stressful life events during the year before the operation revealed an increased risk of recurrence (relative risk: 3.48) and of death (relative risk: 4.37). The role of chronic daily stress at work or in private life was seen as a high risk factor for tumor progression and for survival, after adjusting for age, diagnosis and tumor stage was explored in a further study in 51 cancer patients during their first chemotherapy (Hasenbring et al. 1993). However, some studies could not find a correlation between life events and tumor progression or survival.

6. *Possible Links between Psychosocial Variables and Tumor Progression*

There are a number of possible links between psychosocial factors and the progression of cancer. Chronic daily stress, the lack of social support and maladaptive strategies of coping with stress, especially when accompanied by depression (a) may change individual health behavior, which is known to be related to cancer onset or progression or (b) may directly influence endocrinological and/or immunological factors that are assumed to be precursors of cancer progression and death.

6.1 *Health Behavior*

Patients with increased depression as a persistent reaction to cancer, who experience intense or continuous stress (related to or independent of cancer treatment), or who lack sufficient social support may reveal more maladaptive health behaviors; for example, they tend to be less motivated to practice healthy eating habits, they are more often reliant on alcohol and smoking, and their sleep will be more frequently disturbed (Spiegel and Kato 1996). A number of epidemiological studies have shown that several of these health behaviors are significantly related to cancer progression.

6.2 *Psychobiological Hypotheses*

There has been extensive research that corroborates a connection between psychological distress, neuroendocrine effects, and progression of cancer (Garssen and Goodkin 1999). Chronic stress, depression, and the lack of social support are seen to be related to hyperactivity of the hypothalamus—pituitary—adrenal axis. The possible links between stress, neuroendocrine hyperactivity, and the progression of cancer are many. It is possible that increased cortisol can increase the onset of glucosis, which leads to a selective decrease in the growth of normal cells and enables the increase in growth of tumor cells. Another effect could be that cortisol, prolactin, or another stress-sensitive hormone, could stimulate the tumor growth directly in hormone sensitive tumors such as breast cancer. A third effect extensively researched, assumes that stress-induced hyperactivity of neuroendocrine functions will lead to a suppression of immune functions (Andersen et al. 1994).

7. *Psychosocial Intervention*

Psychosocial intervention has been studied in the acute in-patient setting, primarily in order to reduce treatment-related side effects, such as nausea, vomiting, and pain, as well as in late phases of the disease in order to enhance the ability to cope with cancer-related emotional distress and to mobilize social support.

7.1 *Intervention for the Reduction of Treatment-related Side Effects*

There are a number of studies which have investigated the efficacy of different behavioral methods for reducing chemo- or radiotherapy-induced nausea and vomiting, and for enhancing quality of life during the acute in-patient setting, e.g., relaxation, systematic desensitization, hypnosis, and biofeedback (Carey and Burish 1988). Whilst all methods show positive effects in the reduction of nausea, vomiting, and depression, relaxation, especially in combination with guided imagery, proved to be the most effective treatment.

7.2 *Interventions for Enhancing the Coping Repertoire and Quality of Life*

Several single therapeutic methods were investigated which focused on emotional and social support during cancer treatment. Spiegel et al. (1989) published a first controlled, randomized intervention study of group therapy offered to women with metastatic breast cancer. For a year, sessions of seven to ten women met weekly under professional supervision. These focused on discussions about death, problems in families, on the communication with their doctors, and on methods of enhancing the quality of life during the last phases

of their disease. Follow-up studies after a year revealed lower anxiety and depression, decreased pain and fatigue, and more adaptive coping strategies, when compared to a control group with nonspecific psychological support. In a ten-year follow-up significantly different survival times were observed: 36.2 months after the end of therapy in the experimental group, and 18.9 months in the control group.

Another startling group therapy study was published by Fawzy et al. (1993). They treated patients with a malignant melanoma for six weeks with a one-and-a-half-hour group session once a week. These included educational elements regarding coping with stress, seeking for social support, and relaxation/imagination. Besides a significant reduction of anxiety, depression, and fatigue, for the first time modifications in immunological variables were demonstrated. The authors found significant increases in lymphocytes, natural-killer-cell activity, and alpha-interferon induced natural-killer cytotoxicity compared to a randomized control group. In the six-year follow-up, the experimental group showed a significantly higher survival rate than did the controls, with higher survival rates in patients who showed an increase in active coping behavior.

7.3 How Can We Find Out Who Needs Professional Psychological Support?

In order to incorporate psychosocial therapy into the routine treatment of cancer patients, the question of which patients might need professional psychosocial support, and which patients would respond to the offer, has to be answered. For instance, Wellish and Wolcott (1994) tried to define criteria from their clinical experience of bone marrow transplantation that would enable the assessment of the need for psychosocial therapy based on individual coping abilities, stress in daily life, and psychiatric history. An empirical approach would be the analysis of responders and nonresponders and their determinants within randomized controlled studies (see Helgeson et al. 2000). The question 'Who needs psychological therapy?' should possibly be modified to 'Who needs which sort of psychological therapy?' because it has been shown that more than 95 percent of patients in an acute phase of their cancer treatment will accept this kind of support (Hasenbring et al. 1999).

8. Future Implications

Further research is needed to replicate the important findings of Spiegel et al. (1989) and Fawzy et al. (1993), and to answer the question of how these interventions could be implemented in the routine treatment of cancer patients. A further field of research is the investigation of the tendency towards the chronicity of physical complaints such as pain and fatigue in long-term survivors, and to identify predictors of these. It can be assumed that psychosocial factors play a significant role in association with biological predictors, which would demonstrate the necessity for adjunct psychological therapy.

See also: Cancer-prone Personality, Type C; Childhood Cancer: Psychological Aspects; Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Chronic Pain: Models and Treatment Approaches; Gender and Cancer

Bibliography

Andersen B L, Kiecolt-Glaser J K, Glaser R 1994 A biobehavioral model of cancer stress and disease course. *American Psychologist* **49**: 389–404

Andrykowksi M A, Gregg M E 1992 The role of psychological variables in post-chemotherapy nausea: Anxiety and expectation. *Psychosomatic Medicine* **54**: 48–58

Bonica J J 1985 Treatment of cancer pain: Current status and future needs. In: Fields H L, Dubner R, Cervero F (eds.) *Proceedings of the 4th World Congress on Pain: Seattle*. Raven, New York, pp. 589–616

Butow P N, Coates A S, Dunn S M 1999 Psychosocial predictors of survival in metastatic melanoma. *Journal of Clinical Oncology* **17**: 2256–63

Carey M P, Burish T G 1988 Etiology and treatment of the psychological side effects associated with cancer chemotherapy: A critical review and discussion. *Psychological Bulletin* **104**: 307–25

Cull A, Hay C, Love S B, Mackie M, Smets E, Stewart M 1996 What do cancer patients mean when they complain of concentration and memory problems? *British Journal of Cancer* **74**: 1674–9

Curt G G A 2000 The impact of fatigue on patients with cancer: Overview of fatigue 1 and 2. *The Oncologist* **5**(suppl 2): 9–12

Fawzy F I, Fawzy N W, Hyun C S, Elashoff R, Guthrie D et al. 1993 Malignant melanoma: Effects of an early structured psychiatric intervention, coping and affective state on recurrence and survival 6 years later. *Archives of General Psychiatry* **50**: 681–9

Forsén A 1991 Psychological stress as a risk factor for breast cancer. *Psychotherapy and Psychosomatics* **55**: 176–85

Garssen B, Goodkin K 1999 On the role of immunological factors as mediators between psychosocial factors and cancer progression. *Psychiatry Research* **85**: 51–61

Greer S 1991 Psychological response to cancer and survival. *Psychological Medicine* **21**: 43–9

Greer S, Morris T, Pettingale K W 1979 Psychological response to breast cancer: Effect on outcome. *Lancet* **2**111: 785–7

Hasenbring M 1998 Predictors of efficacy in treatment of chronic low back pain. *Current Opinion in Anaesthesiology* **11**: 553–8

Hasenbring M, Gassmann W, Arp K, Kollenbaum V, Schlegelberger T 1993 Belastungen, Ressourcen und Krankheitsverarbeitung im Verlauf einer Polychemotherapie: Ergebnisse einer prospektiven Längsschnittstudie. [Distress, resources and coping behavior during a polychemotherapy: results of a prospective study]. In: Muthny F A, Haag G (eds.) *Onkologie im psychosozialen Kontext*. Roland Asanger, Heidelberg, Germany

Hasenbring M, Schulz-Kindermann F, Hennings U, Florian M, Ramm G, Zander A R 1999 The efficacy of relaxation/imagery, music therapy and psychosocial support for pain

relief and quality of life: First results from a randomized controlled clinical trial. *Bone Marrow Transplantation* **23**(Suppl. 1): 549

Helgeson V S, Cohen S, Schulz R, Yasko J 2000 Effects of psychosocial treatment in prolonging cancer: Who benefits from what? *Annals of the New York Academy of Science* **840**: 674–83

Holland J C, Rowland J H (eds.) 1998 *Handbook of Psychooncology* (2nd edn.). Oxford University Press, New York

Manne S L, Alfieri T, Taylor K L, Dougherty J 2000 Spousal negative responses to cancer patients: The role of social restriction, spouse mood, and relationship satisfaction. *Journal of Consulting and Clinical Psychology* **67**: 352–61

Morrow G R 1992 Behavioural factors influencing the development and expression of chemotherapy induced side effects. *British Journal of Cancer-Supplement 19: S54–60* **66**: 54–61

Spiegel D, Bloom J R, Kraemer H C, Gottheil E 1989 Effect of psychosocial treatment on survival of patients with metastatic breast cancer. *Lancet* **2**: 888–91

Spiegel D, Kato P M 1996 Psychosocial influences on cancer incidence and progression. *Harvard Review of Psychiatry* **4**: 10–26

Syrjala K L, Roth-Roemer S L, Abrams J R, Scanlan J M, Chapko M K, Visser S, Sanders J E 1998 Prevalence and predictors of sexual dysfunction in long-term survivors of marrow transplantation. *Journal of Clinical Oncology* **16**: 3148–57

Wellish D K, Wolcott D L 1994 Psychological issues in bone marrow transplantation. In: Forman S J, Blume K G, Thomas E D (eds.) *Bone Marrow Transplantation*. Blackwell-Scientific, Boston, pp. 556–570

M. I. Hasenbring

Cancer Screening

Cancer is the uncontrolled growth and spread of abnormal cells (see *Cancer: Psychosocial Aspects*; *Cancer-prone Personality, Type C*; *Childhood Cancer: Psychological Aspects*; *Gender and Cancer*; *Sun Exposure and Skin Cancer Prevention*). Typically cancer develops slowly and has a long preclinical phase. Early detection of cancer means better prognosis because of early and lighter treatment. This can lead to improved quality of life and savings for the society.

1. *Difference between Screening and Testing*

Screening for cancer is any kind of test performed for systematic detection or exclusion of cancer, risk factors, or susceptibility to cancer assuming that the *initiative comes from outside the individual*: from the healthcare or screening organizer. Screening sorts out apparently healthy people who probably have the screened condition from those who probably do not. A screening test need not take the form of a large-scale population intervention. It is not intended to be diagnostic. Persons with positive or suspicious findings are checked further by diagnostic work-out. Screening is different from *testing* where the *initiative comes from the individual*, or from a health professional, *in the course of a diagnostic process* (Wilson and Jungner 1968, Health Council, Netherlands 1994). People screened are less prepared for the diagnosis and less aware about the condition they are screened for than those tested. Screening is aimed at unaware, healthy, asymptomatic people, and thus has heavier public health, social, psychological, and ethical implications than testing.

Prerequisites for successful screening (Wilson and Jungner 1968 for WHO) have been modified to be in accordance with developments in genetics (Nuffield Council of Bioethics), and specific European guidelines for mammography (x-ray of the breasts) screening have been developed (Kirkpatrick et al. 1993). In very few countries launching new screening programs is regulated by law, in some guided by recommendations. The criteria of the National Screening Committee (1998), which are based on the refined WHO criteria of 1968, are expressed as follows:

(a) The *condition* screened should be an important health problem and its natural history should be understood; it should be recognizable either at a latent or early symptomatic stage.

(b) The *test* should be simple, safe and reliable, inexpensive, and acceptable to those screened. The distribution of test values should be known, and the cut-off levels agreed upon. There should be an agreed policy for diagnostic evaluation of those with positive screening findings. The chance of physical or psychological harm to those screened should be less than the chance of benefit.

(c) The *treatment* or intervention should be effective with evidence that early treatment leads to a better outcome.

(d) The *screening program* should be clinically, socially, and ethically acceptable respecting the equity of access principle; it should also be cost-effective and managed and monitored according to quality assurance principles.

The public health rationale behind cancer screening is to reduce mortality and morbidity. The idea is to find cancer in a preclinical asymptomatic phase. Screening for risk factors or genetic susceptibility to cancer is based on the assumption that knowledge of risk affects lifestyle choices and interventions.

2. *Screening Programs for Cancer*

An example of cancers amenable for screening is breast cancer (the most common female cancer) screened by mammography, and also cervical cancer (by Pap smear screening). More than 1 in 10 women in industrialized countries get breast cancer in their lifetime. The natural history of breast cancer is known—there is an asymptomatic stage; treatment is more effective after early diagnosis, and based on

randomized controlled studies the disease-specific mortality among women over 50 has been reduced by a third in the screened group (Kerlikowske 1995, deKoning 2000). The test is acceptable and cost-effective. When screening is carried out as a public health program, the equity of access principle is fulfilled, too. Among over 100 known types, other potential cancers for screening include prostate cancer (in many countries already more common than lung cancer among males) by serum prostate specific antigen (PSA); colorectal cancer by fecal occult blood test and sigmoidoscopy, and to a lesser extent lung, stomach, and ovarian cancer by different technologies.

Different countries implement different screening policies, some more enthusiastically than others. This shows value judgments in the interpretation of scientific evidence, as well as differences in health culture and the availability of resources. The cost of large-scale high-technology screening exceeds the resources of most societies. Low-technology options, like clinical breast examination for breast cancer and visual inspection for cervical cancer, in use in some developing countries, are often the next best, although less effective options.

3. *Social and Psychological Aspects of Cancer Screening*

3.1 *Determinants of Screening Provision and Uptake*

In society, screening can be studied both as a health service offered and an act of an individual. Different cultural and resource factors on both society and service level *determine the screening provided* (Table 1). These, plus individual factors, determine the *uptake of screening* (Marteau 1993). A tradition of centrally organized public healthcare encourages compliance, but also provides equity in access, whereas services based on self-initiation attract those people who usually take care of their health. Examples of individual level determinants are social background, health behavior in general, beliefs, perceived risk, and knowledge (Aro et al. 1999). Previous screening experience influences adherence to later screening rounds.

Theoretical models used to study determinants of cancer screening uptake have mostly been cognitive expectancy value models predicting self-initiated participation. Reasons why they are not so good at predicting behavior are probably due to the fact that, where social and service contexts are essential, the models are too individualistic, cognitive (e.g., concept of intention), and general for cancer screening. The stages of change theory can be applied to adherence to repeated screening rounds, especially if the context factors are included in the model.

3.2 *Implications of Screening Provision and Uptake*

Person-years saved, preferably also quality adjusted life years, are the main implications of screening uptake at the society level (Table 1). Uptake as well as provision of screening raise costs and awareness, and have an impact on values in the society. Medicalization

Table 1
Screening in a society: examples of determinants and implications of screening provision and uptake on different levels of society

	Screening provision		Screening uptake	
	Determinants	Implications	Determinants	Implications
Society level				
	values health policy healthcare resources	awareness costs medicalization	culture healthcare context	lives saved costs demand awareness
Service level				
	healthcare insurance tradition values resources	organization education access	staff invitation follow-up cost context	workload treatment
Individual level				
	NA[a]	awareness worry (nonattenders)	beliefs health behavior experience	inconvenience relief worry false reassurance

a not applicable

is intensified if the screened condition is seen as something bad and unwanted. Diagnostic work-out and treatment increase the amount of work needed. Since screening also detects cancers which otherwise would never become clinical, or which would appear later (overdiagnosis), the workload of the service provider is further increased.

People *invited* to, or otherwise made aware of screening, might get worried about the screened condition, and therefore refrain from participating. An invitation to screening may raise worry and a screening test may be inconvenient or bring other hazards. A *normal screening finding* brings relief, but a finding of any other sort can cause long-lasting anxiety. For those with a *false positive* finding, worry, especially cancer-specific worry, and its behavioral indicators, preoccupation with symptoms, and increased use of health services, might last long past the reassuring diagnosis (Rimer and Bluman 1997, Aro et al. 2000). Intrusive diagnostic procedures, like surgery, can be very stressful. *A true positive diagnosis* of cancer due to screening may both prolong life and improve its quality. However, this impact cannot be evaluated individually but only at a group level. Finding cancer, which would never have become clinical in a lifetime, is clearly a major adverse effect of screening. *A false negative* finding misleadingly reassures and delays diagnosis. *A true negative* finding is reassuring, assuming that screening has no adverse side effects. A good quality organized mammography screening program (for ages 50+ every two or three years) finds breast cancer in five out of 1,000 screened women, has 1:1 malignant:benign ratio of breast biopsies (Kirkpatrick et al. 1993), and a very low (even less than 1 percent) false positive rate per screening round. For a woman attending several screening rounds in a lifetime, the cumulative risk of a false positive finding rises considerably—close to 50 percent in 10 screening rounds in the US screening (Elmore et al. 1998).

While the screening uptake literature is partly based on theoretical models in predicting self-initiated attendance, literature on implications of the screening process and findings has a nearly nonexistent theoretical framework. The cognitive-behavioral models of illness anxiety rising from the tradition of psychosomatic studies have mostly been used to guide research questions and the development of methods.

4. Challenges in Studying Cancer Screening

The study of cancer screening is enriched by the examination of the determinants and implications of both provision and uptake, including the levels of society, service, and the individual. A prospective design—most easily built in organized screening programs with a planned schedule—enables the measurement of predictive determinants. Context-specific measures, in addition to standard ones, reveal valuable factors, such as improved cancer awareness and enhanced health behavior.

Screening, however, is not the solution for cancer control. It is costly, cost-effective only among certain age or risk groups, and not globally feasible. Prevention would be better than screening for an already existing disease, but causal factors of cancer are still poorly understood. Of the predicted 20 million new cancer cases every year by 2020, up to 14 million will occur in developing countries. At present one in three to four people in industrialized countries fall ill with cancer during their lifetime. The overall five-year survival rate for cancers is over 50 percent, but for small (<1 cm) breast cancers it exceeds 90 percent.

Screening for risk factors or susceptibility to cancer is closer to prevention than screening for disease, but has different implications. Screening for disease detects few cases, most of which would become clinical later without screening, whereas screening for risk factors finds a lot of people who will never develop the disease. Eventual screening for genetic susceptibility to cancer has to meet the criteria of good quality screening as well. Informed consent is essential and, in addition to risk notification, social and psychosocial consequences need to be understood and intervention options (lifestyle changes, choosing a partner, medication, preventive surgery, other kinds of therapy, and combinations or interactions of these) clarified. Potential screening programs need to be evaluated against accepted criteria (Hakama 1991) by healthcare technology assessment methods. Public debate and the empowerment of lay people to take part in discussions on values, ethics, screening policies, and priorities need to be encouraged. Modeling software provides one way of forecasting effects and impacts of screening where experimental studies are not acceptable or feasible. People need knowledge and skills to control the cancer problem both in the roles of decision-makers and service providers, and as individuals.

See also: Cancer-prone Personality, Type C; Cancer: Psychosocial Aspects; Childhood Cancer: Psychological Aspects; Gender and Cancer; Genetic Screening for Disease-related Characteristics; Health Behavior: Psychosocial Theories; Health Care Technology; National Health Care and Insurance Systems; Risk Screening, Testing, and Diagnosis: Ethical Aspects; Screening and Selection; Sun Exposure and Skin Cancer Prevention; Vulnerability and Perceived Susceptibility, Psychology of

Bibliography

Aro A R, Absetz P S, van Elderen T M, van der Ploeg E, van der Kamp L J T 2000 False-positive findings in mammography screening induce short-term distress—breast cancer-specific concern prevails longer. *European Journal of Cancer* **36**: 1089–97

Aro A R, de Koning H J, Absetz P, Schreck M 1999 Psychosocial predictors of first attendance for organised mammography screening. *Journal of Medical Screening* **6**: 82–8
de Koning H J 2000 Assessment of nationwide cancer-screening programmes. *Lancet* **355**: 80–1
Elmore J G, Barton M B, Moceri V M, Polk S, Arena P J, Flectcher S W 1998 Ten-year risk of false positive screening mammograms and clinical breast examinations. *New England Journal of Medicine* **338**: 1089–96
Hakama M 1991 Screening. In: Holland W W, Detels R, Knox G, Fitzsimons B, Gardner L (eds.) *Oxford Textbook of Public Health*. Oxford University Press, Oxford, UK, Vol. 3
Health Council of the Netherlands: Committee Genetic Screening 1994 *Genetic Screening*. Health Council, publication No. 1994/22E, The Hague
Kerlikowske K, Grady D, Rubin S M, Sandrock C, Ernster V L 1995 Efficacy of screening mammography. A meta-analysis. Review. *Journal of the American Medical Association* **273**: 149–54
Kirkpatrick A, Törnberg S, Thijssen M A O (eds.) 1993 *European Guidelines for Quality Assurance in Mammography Screening*. Office for the Official Publications of the European Communities. Report EUR 14821. Luxembourg
Marteau T M 1993 Health-related screening: Psychological predictors of uptake and impact. In: Maes S, Leventhal H, Johnston M (eds.) *International Review of Health Psychology*. Wiley, Chichester, UK, Vol. 2
Morrison A S 1992 *Screening in Chronic Disease*, 2nd edn. Oxford University Press, New York
National Screening Committee 1998 *First Report of the UK National Screening Committee*. Department of Health, London
Rimer B K, Bluman L G 1997 The psychosocial consequences of mammography. *Journal of the National Cancer Institute Monograph* **22**: 131–8
Wilson J M C, Jungner G 1968 *The Principles and Practice of Screening for Disease*. WHO, Geneva

A. R. Aro

Cannibalism

Cannibalism is the eating of one's own kind. Of the 75 mammalian species known to eat conspecifics, three are higher primates: chimpanzees and gorillas, usually under conditions of social or environmental stress—and humans. Among these species the rarity of cannibalism, relative to the myriad opportunities and temptations to engage in it, suggests a biologically-based aversion that may have evolutionary utility. Many cultures have formulated this aversion as a taboo, or as an attitude which passively regards the behavior as unthinkable or faintly ridiculous. In most cultures where cannibalism was institutionalized (and as such is now thought to be extinct in the world), the element of taboo appeared in the guise of restrictions on the practice of cannibalism: the categories of persons eligible to eat and be eaten, the body parts consumed, and various procedural rules marking the eating of one's own kind as a special, often ritually significant, form of behavior. In societies in which cannibalism was a component of initiation into a religious cult or secret society, it was frequently the very act of violating the taboo and overcoming the fear and loathing associated with cannibalism that provided the element of ordeal common in such induction rites. In symbolic contexts, the cultural treatment of cannibalism is akin to that of incest, and in mythology the two images are often closely associated; they also raise similar explanatory problems, in that each rests upon the interaction of biological and cultural factors.

If a phylogenetic aversion toward cannibalism does exist, it is one that can be modified—though perhaps not fully abolished—by culture. Thus, from the actor's point of view the objective definition of cannibalism as applied to humans may be complicated by locally defined differences between own kind and other kind. If members of alien groups are taken to be nonhuman, eating them may not be perceived or psychologically registered as cannibalistic. Ontological distancing of this sort is not uncommon among cultures practicing cannibalism and although objectively such maneuvers may be seen as rationalizations, it is important to grant that cultural understandings are capable of creating a situation in which eating the other is palatable, even commendable, whereas eating one's own is repulsive, horrific, or insane.

As the quintessential uncivilized Other, and a projected inversion of Self, the Cannibal has been an object of fascination in many societies and for much of recorded history. Writing in the fifth century BC, Herodotus reports that man-eating is a custom among the Anthropophagi, dwelling in the wild lands north of the Black Sea, and among other groups living beyond the light of Greek civilization. In the sixteenth century, Montaigne's ironic essay *Of Cannibals* used the man-eating attributed to tribes in the New World as a foil to criticize certain European barbarities of his time. The implied 'otherness' of cannibalism appears in later literary works, such as Daniel Defoe's *Robinson Crusoe* (1719), Herman Melville's *Typee* (1846), and Joseph Conrad's *Heart of Darkness* (1902), to epitomize the exoticism of the faraway Americas, South Seas, and Africa, respectively. Today, novelists and filmmakers sometimes employ cannibal themes to convey the grotesque, nightmarish, or psychotic nature of certain situations and characters. The same fascination excites the public's appetite for accounts of present-day incidents of cannibalism under conditions of starvation or psychopathology.

As a staple figure in the world's folklore traditions, the Cannibal's otherness is usually certified by imagining the creature as not quite human: a witch, ghost, demon, giant, god, ogre, vampire, were-animal, or other monster. By negative example, the Cannibal discloses the nature of morality, sociality, and other virtues. Freudian theorists (e.g., Klein 1975) interpret

such images as projections of unconscious ideas originating in early childhood. Thus, the infantile wish to devour the parent is converted into the fear of being devoured by the parent—a fear made acceptable to the ego by virtue of the parent-figure's fantastical disguise.

Types of literal cannibalism vary according to motive and circumstance; so great is the diversity, in fact, that it tends to overwhelm the common feature of ingestion and confound efforts to understand cannibalism as a unitary phenomenon. Auto-cannibalism conceivably applies to the act of eating cast-off parts of oneself—hair, nail clippings, mucous, excrement, placenta—but is perhaps better reserved for instances in which, under torture or other duress, individuals partake of their living flesh, raw or cooked. Greek mythology imagines this horror in the fate of Erisichthon, who unwisely violates a grove sacred to Ceres. As punishment, the goddess calls for Famine to visit the offender and deliver upon him a hunger so insatiable that, eventually, he devours his own body, limb by limb, unto death. Gustatory cannibalism refers to the unceremonious eating of human flesh, invariably that of an enemy, simply as food. Although reported for parts of Africa and Melanesia, some authorities doubt the existence of such culturally unadorned cannibalism; for, as Sahlins (1983, p. 88) remarks, 'cannibalism is "symbolic," even when it is "real."' Phrased slightly differently: noncannibals are fascinated by cannibalism because they don't practice it; cannibals are fascinated by cannibalism because they do. Accordingly, while human flesh is undoubtedly a potential source of high-quality protein (Harner 1977), cannibalism as a cultural practice cannot be satisfactorily explained purely as a response to particular material conditions. Epicurean cannibalism regards human flesh as a delicacy, an aesthetic normally requiring the victim to be perceived as something less than—at any rate, other than—human. Medicinal cannibalism is the ingesting of human tissue, usually that of an executed criminal, as a supposed medicine or tonic; this practice existed throughout Europe during the sixteenth, seventeenth, and eighteenth centuries (Gordon-Grube 1988). Innocent cannibalism is so named because the perpetrator is unaware that he or she is eating human flesh; thus, in Greek legend, Atreus punishes his brother Thyestes for seducing his wife, by tricking him into eating his own children. Survival cannibalism occurs under starvation conditions, such as shipwreck, military siege, and famine, in which persons normally averse to the idea are driven to the act by the will to live. That some persons under such conditions have been known to accept death instead of resorting to cannibalism attests to the psychological strength of the taboo and the corresponding revulsion many individuals feel toward the cannibal act. Survival cannibalism, especially when combined with murder, raises legal and moral issues having to do with the 'defense of necessity,' the first judicial analysis of which occurred in British courts when, in 1884, survivors of the wrecked yacht *Mignonette* were successfully prosecuted for murder (Simpson 1984).

In works of ethnology and ethnohistory, aggressive cannibalism, usually in the context of warfare, is the most frequently reported form. For many peoples of Africa, the Americas, and the South Pacific, cannibalism is the supreme expression of hostility: killing and eating someone is the ultimate act of annihilation; reducing the hated enemy to feces is the ultimate insult. Such motives appear, also, in the large-scale cannibalism reportedly perpetrated upon 'class enemies' in China's Guangxi Province during the Cultural Revolution of 1966–1976 (Zheng 1996). In some societies the act aggressively appropriates the victim's spiritual strength, by consuming either the entire body or particular parts thereof, such as the heart, liver, genitals, or head. Early Spanish explorers reported that the Tupinamba of coastal Brazil adopted captives into their community, where they received excellent treatment and sometimes even married, before being slaughtered and eaten. Similarly, the Aztecs are said to have treated war captives with great honor and solicitude, as a prelude to sacrificing them to the gods and giving their bodies over to be eaten by the nobility. In these cases, then, cannibalism occurs in conjunction with sacrifice, but only after the victim is ennobled and made precious as an offering through incorporation into the group. Such sacrificial cannibalism is widely noted (e.g., Hogg 1966) and is shown graphically, if symbolically, in the holiest of Christian rituals—the eucharistic 'eating' of the body and blood of Christ.

As implied, most types of volitional cannibalism involve eating an enemy or, at the least, someone from outside the group ('exocannibalism'). Eating someone from within the group ('endocannibalism') usually occurs as a form of mortuary (or funerary) cannibalism, in which all or part of the body is consumed as an act of piety or affection, as a source of spiritual strength or continuity, as a kind of recycling of the spirit for the purpose of group renewal or reproduction, or as the means for assisting the spirit of the deceased to attain a desirable state in the afterlife. A premortem variant may be found in the Chinese idea (and practice) of ultimate piety, that is, a child giving a piece of his own flesh to nourish an ailing parent. The unconscious desire to incorporate the loved object—a common psychoanalytic observation—supports the notion that mortuary practices in many societies include symbolic elements that are sublimations of cannibalistic impulses or, in a more speculative vein, vestiges of actual cannibalism in the past (Sagan 1974).

As a topic of scholarly and popular interest, cannibalism was jolted by the publication of a book doubting that institutionalized cannibalism ever existed, anywhere. In *The Man-eating Myth*, author Arens (1979) argues that the centuries of reports about cannibalism in remote times or places rest on not a shred of reliable, eyewitness evidence. Instead of

describing literal practices, the widespread attribution of 'cannibalism,' Arens suggests, is a derogatory racial or ethnic stereotype—a device for debasing or demonizing 'the other,' thus proving the need to govern or convert them. Scholars, he concludes, have been too quick and uncritical in accepting such attributions at face value.

The debate generated by Arens's book comes down to whether or not one is prepared to accept eyewitness accounts from missionaries, administrators, and adventurers, or the circumstantial evidences of ethnography or ethnohistory. Although few anthropologists agree with Arens's blanket dismissal of such sources, the controversy has usefully heightened both scholarly awareness of the ideological potential of 'cannibalism' and empirical rigor in studies of cannibalism as a culturally embedded, institutionalized practice (e.g., Brown and Tuzin 1983, Sanday 1986, Goldman 1999); it has also helped to inspire a body of literary and historical studies (e.g., Lestringant 1997) which examine 'cannibalism' as a metaphor in colonial discourse or in explicitly imaginary treatments of the cultural 'other.'

Of particular interest, insofar as it offers the only remaining possibility of direct evidence of willful cannibalism, is a resurgence of archaeological attention to prehistoric cannibalism and a refinement of techniques to test for it. Thus, modern taphonomic analyses of skeletal remains recovered at Anasazi sites (AD 900–1300) in the US Southwest (Turner and Turner 1999) and at Fontbrégoua Cave, a Neolithic site in southeastern France (Villa et al. 1986), strongly indicate that cannibalism occurred in those places in association with violent homicide, probably in the context of war or political terrorism. It is hoped that these new techniques may be used to test long-standing suspicions of cannibalism on the part of more ancient hominids—*Australopithecus africanus*, *Homo erectus*, *Homo neanderthalensis*, and archaic *Homo sapiens*.

See also: Ritual; Taboo

Bibliography

Arens W 1979 *The Man-eating Myth: Anthropology & Anthropophagy*. Oxford University Press, Oxford, UK

Brown P, Tuzin D (eds.) 1983 *The Ethnography of Cannibalism*. Society for Psychological Anthropology, Washington, DC

Goldman L R (ed.) 1999 *The Anthropology of Cannibalism*. Bergin and Garvey, Westport, CT

Gordon-Grube K 1988 Anthropophagy in post-renaissance Europe: The tradition of medicinal cannibalism. *American Anthropologist* **90**(2): 405–9

Harner M 1977 The ecological basis for Aztec sacrifice. *American Ethnologist* **4**(1): 117–35

Hogg G 1966 *Cannibalism and Human Sacrifice*, 1st American edn. Citadel Press, New York

Klein M 1975/1932 *The Psycho-analysis of Children* (trans. Strachey A). Dell, New York

Lestringant F 1997 *Cannibals: The Discovery and Representation of the Cannibal from Columbus to Jules Verne*. University of California Press, Berkeley, CA

Sagan E 1974 *Cannibalism: Human Aggression and Cultural Form*. Harper & Row, New York

Sahlins M 1983 Raw women, cooked men, and other 'great things' of the Fiji Islands. In: Brown P, Tuzin D (eds.) *The Ethnography of Cannibalism*. Society for Psychological Anthropology, Washington, DC, pp. 72–91

Sanday P R 1986 *Divine Hunger: Cannibalism as a Cultural System*. Cambridge University Press, Cambridge, UK

Simpson A W B 1984 *Cannibalism and the Common Law: The Story of the Tragic Last Voyage of the Mignonette and the Strange Legal Proceedings to Which it Gave Rise*. University of Chicago Press, Chicago

Turner C G II, Turner J A 1999 *Man Corn: Cannibalism and Violence in the Prehistoric American Southwest*. University of Utah Press, Salt Lake City, UT

Villa P, Bouville C, Courtin J, Helmer D, Mahieu E, Shipman P, Felluomini G, Branca M 1986 Cannibalism in the Neolithic. *Science* **233**: 431–6

Zheng Y 1996 *Scarlet Memorial: Tales of Cannibalism in Modern China* (trans. Sym T P). Westview Press, Boulder, CO

D. Tuzin

Capitalism

Capitalism is an economic system based on free market, private enterprise and ownership. The term and concept of capitalism as an exploitative socio-economic system was initially introduced by Marxist social sciences. The spread of the term, however, was strengthened and its concept changed by the Western critiques of Marxism and socialism. Although rarely used, sociology and economics accepted the term and historiography broadly discuss its half-a-millennium development and various stages from commercial to globalized capitalism.

1. Concepts and Theories

The term, derived from *capital* (root in Latin *caput*, 'head') and related to *capitalist*, appeared in the mid-late nineteenth century in the English, German, French, and other European languages. The word *capital* had been used from the seventeenth century onward in the economic sense, meaning 'fund of money,' and later 'principal' (as opposed to *interest*). The term *capitalist* appeared in mid–late eighteenth century French, English, and American texts. The first recorded occurrence of the form *capitalism* in English is 'The sense of capitalism sobered and dignified Paul de Florac' in Thackeray's *The Newcomes* II (1854).

The spread of the term *capitalism* in common language use paralleled its use in Marxist socio-

economic theory, the appearance of *socialist/communist* economies in various countries, and the expansion of Cold War rhetoric based on the opposition of the two socioeconomic systems. The term and concept of *capitalism* as a socioeconomic system was basically introduced by the first theorists of Marxist social sciences. The spread of the concept, however, was strengthened by Western critiques of socialism and Marxism. Although rarely used, sociology and economics accepted the term and historiography broadly discussed the transition from feudalism to *capitalism*, and the various stages of capitalist development from its early commercial phase via industrial and financial *capitalism*, to its late, twentieth century characteristics. The capitalist world system is often discussed according to the pattern of advanced core and dependent, backward periphery. Development studies and economics analyze late twentieth century globalization and the increased role of transnational companies in a more inter-related, but at the same time even more polarized capitalist world economic system.

Karl Marx, who often used the terms capital, capitalist, and capitalistic, did not use the term capitalism as a noun in such basic writings as the *Communist Manifesto* and *Das Kapital*. The term appeared only later in his correspondence on the development of Russian capitalism during the late 1870s. The concept finally gained scholarly interpretation and became widely used by Marxist theory and the socialist movement in the last third of the nineteenth century. The concept was broadly used in historiography and sociology, but much less frequently in non-Marxist economics, to describe a phase of history or a set of ideas and mentality. During the twentieth century, capitalism became a popular term, especially to contrast the private-market system with socialism.

Marxist theorists interpreted the term capitalism as a socioeconomic system, a 'mode of production,' in which 'capital is not a thing but a social production relation belonging to a definite historical formation of society' (Marx 1952, Chap. 48, English trans.; first published in 1867). Capitalism is a class society basically consisting of two antagonistic classes: bourgeoisie and proletariat. The system historically came about through 'primitive accumulation,' the separation of peasantry from the means of production, and the creation of a free proletariat deprived of property, whose only saleable commodity was its labor power. The ruling class, the bourgeoisie, owns the means of production, buys the labor power, and controls all economic transactions and production processes. In the capitalistic reproduction process, capitalists who own the capital buy various types of commodities (raw material, machinery, labor power, etc.) for producing new commodities for sale. In this reproduction process money is transformed into commodities, then again into money (M–C–M). The capitalist, however, is motivated by making profit and wants to get more money back at the end of the reproduction process (M–C–M'). The source of profit is one of the commodities—the labor power, which creates a 'surplus value.' The value of the goods produced by the workers is higher than the value of the labor power bought for a wage, which covers the expenses of the reproduction of labor power (including the reproduction of the class itself). Capitalists expropriate this 'surplus,' thus exploiting the proletariat. In capitalism, therefore, social production is in sharp contrast with individual expropriation. As a consequence, wealth is accumulated in the hands of capitalists, while the proletariat is kept in poverty. Capitalism, however, creates its own 'grave diggers'—the ever-increasing proletariat who, uniting together in the advanced West, rises up to destroy the capitalist system through a proletarian revolution. Bourgeois expropriators become expropriated and a just communist society without classes and private ownership is established.

Post-Marxian Marxism revised this theory. Eduard Bernstein, the leading theorist of the powerful German Social Democratic Party, rejected the concept of proletarian revolution in his landmark article series, *Probleme des Socialismus* (1896). He argued that capitalism has developed in a direction different from the one indicated by Marx. Poverty has not grown but diminished, and welfare reforms have been accepted peacefully. Bernstein and the Western Social Democratic movement maintained that the proletariat could realize its interests and goals by organized mass movement and parliamentary reforms. Another leading German socialist, Rudolf Hilferding, in his *Das Finanzkapital* (1909), added that modern economy created 'consciously regulated social relations.' Later he concluded in the theory of 'organized capitalism,' thus, a state-controlled welfare system.

Revolutionary socialism emerging during the first decades of twentieth century redefined contemporary capitalism as the higher or the 'last' stage of capitalism, otherwise called *imperialism*. In her *The Accumulation of Capital*, Rosa Luxemburg (1951, English trans.; first published in 1913) maintained that C would not be able to expand its markets, recruit new labor force, and mobilize new resources without a 'third person,' i.e., noncapitalist (independent farmer) sectors of the advanced countries, and, most of all, the agricultural economy of the colonies and peripheries. The Russian Vladimir I. Lenin defined imperialism as a new, expansionist stage of capitalism, which was also characterized by the domination of 'monopolies,' where the competition of small individual capitalists is replaced by the deadly competition of monopolies, not only locally but also in the international arena. According to this view, capital exports became a new instrument for exploiting backward peripheral countries. Imperialism has led to the fight for the redistribution of the colonies and spheres of interest, and this unavoidably generates wars, and leads to the

proletarian revolution of the backward world. This version of Marxism was the leading ideology of the Bolshevik revolution in Russia, a revolutionary wave in Central and Eastern Europe after World War I, and the communist takeover in a series of Central and Eastern European, Asian, and African countries after World War II.

Non-Marxist scholarship did not build on the Marxist concept and interpretation of capitalism, but the term itself became commonly used in scholarship. Many twentieth century sociologists and economists used the term capitalism to mean a society and economy based on free market and entrepreneurial interest, and accepted these characteristics as the essence of capitalism, in contrast to socialism. Critiques of Marxism, therefore, contributed just as much to the spread of the term and concept of capitalism. Because of his work on the role of Protestant ethic in the development of capitalism, Max Weber played an instrumental role in making capitalism a commonly used term in scholarship. Ludwig von Mises and his disciple Friedrich von Hayek frequently used the term capitalism when launching the most radical attacks against Marxist theory and socialist economy and defending an undisturbed free market. Joseph Schumpeter's book *Capitalism, Socialism, and Democracy* (1943) emphasized the changing evolutionary character of capitalism because of its entrepreneurial interest that 'incessantly revolutionarizes the economic structure from within' and with a 'creative destruction' destroys the old, outdated and creates the new.

The broadly accepted non-Marxist meaning of capitalism is rather different from the Marxist interpretation. According to the most widespread definition, capitalism is a complex socioeconomic system where the bulk of the means of production is in private hands guaranteed by modern property rights and laws. An 'invisible hand' (Adam Smith), the operation of market, regulates the entire economy, production, distribution, and labor. The market operates according to the play of supply and demand, which influence price fluctuations and automatically regulate investments and output. As the Say-law phrased it: each supply creates its demand. The market mechanism, therefore, is a self-regulating system. Capitalist entrepreneurs are acting individually and freely in a rational organization and control themselves with double-entry bookkeeping, motivated by gain and increase of profit. Capitalism, in this sense, is in sharp contrast with both its rigidly regulated feudal predecessor and contemporary rival socialism. Hereditary, almost unchangeable social status, basically noncommercialized production for local consumption, strict communal and guild regulations, and the lack of free private property, labor and market make economic relations in Feudalism rigid opposed to capitalism; and non-market economy based on state ownership and central planning contrast socialism to capitalism.

While capitalist society in its early stage has been, and in contemporary backward countries is characterized by extreme social polarization and income inequity, capitalism granted equal rights and, in principle, freedom of choice later in the advanced countries. From the 1870s on, but mostly from the 1950s, the industrial working class began shrinking, and the 'white-collar' layers, including the middle class, gradually expanded and became the most dominant layer of society. During the 1960–80s, a two-thirds to three-quarters white-collar and middle class majority dominated advanced societies.

2. *Early Stages of Development*

Capitalism, broadly accepted as a phase of history, is dominating the last half-a-millennium. It has also become conventional that capitalism itself had various phases and stages. Its antecedents go back to ancient history. Barter and exchange of one thing for another, as Adam Smith, the main advocate and ideologue of market capitalism argued, are a part of human nature. The pockets of private-market economy and its institutions, especially in urban settlements, flourished in medieval Europe. Moneylenders and urban merchants were its main representatives.

Its continuous development, however, began mostly from the sixteenth century. Historiography produced a vast literature on the transition from feudalism to capitalism. These writings maintain that declining productivity, a demographic crisis, and the scarcity of peasant labor generated the lowering of rents, a decrease of labor services, and an increase in peasant mobility and freedom in Western Europe. In most of these countries, consequently, serfdom was gradually loosened and disappeared by the sixteenth century. Commercialization of agriculture played a central role in the development of capitalism. The flourishing textile industry in Britain with an excess production over consumption led to the enlargement of productive capacities in an unheard scale and made England the 'workshop of the world.'

This stage between the sixteenth and eighteenth centuries is often called 'commercial capitalism' in history. The system was based on private ownership of warehouses and ships and buying and selling goods all over the world. The merchants also lent out money at interest and established connections with production. They bought and distributed raw materials for peasants who worked at home in the traditional way (putting-out system), and gradually subordinated the rural cottage industry. It led to a proto-industrialization, the development of a decentralized putting-out system coordinated by merchants, and later the establishment of (nonmechanized) factories, which introduced division of labor in the production procedure.

The emergence of capitalism, as various historians have discussed it, also gained incentive from the discovery of the 'New World.' The unlimited inflow of gold and silver from Latin America led to the accumulation of wealth. Although Spain was the main beneficiary at the beginning, most of the wealth got into the hands of Dutch and English merchants. 'It is not the importation of gold and silver—stated Adam Smith—that the discovery of America has enriched Europe … [It] made a most essential [change] … by opening a new inexhaustible market to all the commodities of Europe, it gave occasion to a new division of labor.' The new markets and emerging Atlantic trade enriched the commercial powers of northwest Europe and made them the core of the emerging modern world system from the sixteenth century on. This period saw the rise of modern world trade, characterized by the trade of mass consumption goods of food and textile, instead of the medieval trade of luxury goods from the near East and India. While the core countries of Western Europe sold processed industrial products, they bought unprocessed raw materials and agricultural produces from the peripheral countries of the world system: Latin America and Eastern Europe.

The rise of capitalism was strongly influenced by cultural-ideological and institutional factors. The Protestant reformation of the sixteenth century created a new work ethic and lifestyle, which was a prime mover of capitalist development. Hard work and thrift became virtues, which was in sharp contrast to parasitic noble values and lifestyle, the so-called *Hidalgo* (Spanish) and *Szlachta* (Polish) attitude. Meanwhile, the new moral code legitimized capitalist inequity: character was idealized, richness considered to be the triumph of will and work, poverty became a moral failing.

In the absolute state between the sixteenth and eighteenth centuries, centralizing policy creating uniform monetary systems, legal codes, unified and defended home markets without internal tariffs, also played an important role. The isolation of the domestic market by limiting imports was accompanied by a tremendous effort to build up a large colonial empire and a system of various privileges and stimuli for exports. The state itself became a promoter of the development of market, industry, trade, and transportation by the creation of road systems, canals, navies, and even state-owned industries. All these served to strengthen the state, but also historically paved the way for dramatic capital accumulation and modern capitalism.

2.1 Industrial Capitalism and Beyond

In the last quarter of the eighteenth century, Adam Smith published his powerful theory and ideology of the free market system, the *Wealth of Nations*, which became instrumental in the development of laissez-faire capitalism. Smith argued that from the selfish activity of the entrepreneurs, driven by profit motivation, the entire society profited, and that the individual actions, because of the self-regulating impact of the market, created an economic harmony and well being for the nation. Smith established the economic thought and popular thinking on capitalist market economy for two centuries. Political thinkers and economist of latecomer countries, however, attacked British *laissez-faire* ideology. The American Hamilton and Germans Johann-Gottlieb Fichte and Friedrich List argued against *laissez-faire* in favor of protective state interventionism.

The end of the eighteenth century is considered to be a turning point in the development of capitalism. As a consequence of the British Industrial Revolution, a new stage of more mature, as social scientists and historians called it, 'industrial capitalism' emerged and dominated the nineteenth century. Starting in Britain, the industrialization process transformed the core countries of the capitalist world system into industrialized economies with a high, sustained economic growth. Industrial capitalism, in the first period of its existence, institutionalized long working days, and exploited female and child labor ruthlessly in the factory system. Living conditions of the workers in crowded and polluted industrial cities, as described by Charles Dickens and Friedrich Engels, were extremely poor. Capitalism, however, reached its dynamic period of development. A banking revolution in Belgium and France, then, most of all, in Germany, introduced the Crédit Mobiler type, then the German type of investment banks with huge industrial portfolios. These banks played an instrumental role in industrialization. The relatively small, family-managed factories enlarged, and most of them became joint stock companies with professional management. From the mid-nineteenth century on, a European railroad system created a united European transportation network and an integrated marketplace. In this period virtually the entire globe became part of the capitalist world system. The core countries offered an unlimited market for food and raw materials. Britain and a handful of Western European countries absorbed nearly 70 percent of the world's food and raw material exports. New institutions strengthened international capitalism, such as a free trade system and the gold standard—a fiscal underpinning of the free trade system making the national currencies convertible, which was gradually introduced from the 1860s to the 1870s.

The expansion of capitalism around the turn of the nineteenth/twentieth centuries led to considerable changes in the system. Werner Sombart introduced the term of 'late capitalism,' and Rudolf Hilferding named this new period 'finance capitalism.' The concept of 'Imperialism' (Hobson) also appeared and became a leading concept of revolutionary Marxism. Some of the socialist theorists, however, such as Karl Kautsky,

considered imperialism as a 'policy' of capitalism, similar to long workdays and child labor in its first phase. He maintained that capitalism could exist without colonies by introducing new policies.

Twentieth-century capitalism was characterized by the challenge of major wars, worldwide economic depression, and a significant slowing down of economic growth. More importantly, capitalism had to face the challenge of socialism, first as a movement, then as a rival socioeconomic system. A great part of the twentieth century was characterized by the struggle of the two socioeconomic systems. Capitalism reacted by a highly flexible adjustment to the new challenges. During the wars and depression, the development of state-capitalism with a strong and effective state interventionism, a return to protectionism, inflationary financing, and even state investments and planning, helped the system to respond to the new challenge. Large state-owned sectors emerged in Italy, Spain, and Poland. National planning was introduced in Germany and Hungary to cope with economic decline and unemployment. Self-regulating, *laissez-faire* market capitalism was replaced everywhere by a regulated market system. Adam Smith's theory was replaced by John Maynard Keynes' concept of state interventionism to create additional demand.

3. *The Welfare State, Mixed Economy and Back to Globalized Laissez-Faire*

The socialist challenge of capitalism required an effective response. The first chapter of this rivalry was opened in the late nineteenth century. Chancellor Otto von Bismarck, in his struggle with the strong German socialist movement, turned towards the introduction of welfare institutions. Pension system, health care, and other welfare institutions were introduced for workers to take the wind out of the sail of the emerging socialist movement. For these developments, the Great Depression of the early 1930s brought about a turning point. The Swedish Social Democratic government, elected in 1932, started building a welfare state. President Franklin D. Roosevelt also initiated major social legislation and the introduction of the social security system. The comprehensive system of the welfare state, however, emerged only after World War II in Western Europe. Capitalism, in contrast with its nineteenth century features, initiated widespread social legislation. Paid vacation, short workweek, nationwide pension systems, free health care, education, maternity leave, cheap housing, and many other institutions were established. Human rights were re-evaluated and the right to work, i.e., a full-employment policy, became dominant. Inequity, at least in the middle layers of the society, markedly diminished in the advanced capitalist countries.

The period after the Second World War also saw the spread of mixed economies. Almost all of the advanced Western European countries, led by France and Britain, nationalized entire sectors of their economies and established new state-owned firms in various industries.

Capitalism underwent a spectacular transformation. In his article on *The Instability of Capitalism* (1928), the leading Austrian-American economist, Joseph Schumpeter, prophesied the transformation of the system and the fading of the original meaning of capitalism. He maintained that capitalism is inherently stable and able to recover from great crises, but generates socio-intellectual effects discordant with its spirit and institutions. As a consequence, 'Capitalism is … in so obvious process of transformation into something else … [that] it will be merely matter of taste and terminology to call it Socialism or not.'

Towards the end of the twentieth century, capitalism adjusted to the new historical situation and markedly transformed but preserved entrepreneurial interest, market flexibility, efficiency, competitiveness, and reached its highest prosperity and fastest growth rate ever. The Western core countries of the world system successfully adjusted to a new technological-communication revolution and became the leader of new technologies and the rise of postindustrial society.

During the long period of postwar prosperity, a part of the former Asian and South European peripheries of the capitalist world system reached a higher than average growth and became an integrated and equal part of the core countries. After a long period of rivalry, capitalism emerged victorious while the parallel world system of centrally planned state socialism, which conquered one-third of the world and influenced an even greater part by its policy after World War II, collapsed in 1989–91. The former Soviet Bloc countries introduced radical market reforms, thus re-establishing capitalism. During the last third of the twentieth century, *laissez-faire* capitalism became the leading model again. Neo-liberal economics triumphed and challenged mixed economy and welfare state.

Capitalism in late twentieth century reached a new stage in its history. The main trend of this new age is globalization. National boundaries and national economies rapidly lost their importance. Multi- or transnational companies penetrated previously independent economies. Foreign direct investment became instrumental all over the world, including the advanced core countries. Towards the end of the 1990s nearly one-third of American exports and two-thirds of imports were intrafirm deliveries. About one-third of French, Dutch, and British industrial output was produced and roughly 25–40 percent of their research and development expenditure was financed by affiliates of transnational companies. In Ireland, and former state-socialist Hungary, foreign affiliates produced two-thirds of industrial output and financed roughly 70 percent of research and development expenditures.

While globalization, in some cases, contributed to a successful catching-up process with the core, it also preserved and even strengthened the core–periphery inequity. The gap between advanced core and the peripheries was growing considerably: intercountry income spread was 10:1 in 1913; 26:1 in 1950, but it increased to 40:1 at the end of the twentieth century. A newly globalized but even more polarized capitalist world system, dominated by an expanding Western core, was emerging at the turn of the millennium.

See also: Capitalism: Global; Economic History; Industrialization, Typologies and History of; Marxian Economic Thought; Marxist Social Thought, History of; Polanyi, Karl (1886–1964); Social History; Weberian Social Thought, History Of

Bibliography

Aston T H, Philpin C H E 1985 *The Brenner Debate*. Cambridge University Press, Cambridge, UK

Baldwin P 1990 *The Politics of Solidarity. Class Bases of the European Welfare State 1875–1975*. Cambridge University Press, Cambridge, UK

Hacker L M 1940 *The Triumph of American Capitalism*. Simon & Schuster, New York

Heimann E 1964 *History of Economic Doctrines*. Oxford University Press, New York

Hobson J A 1902 *Imperialism. A Study*. Nisbet & Co, London

Kautsky K 1914 *Der Imperialismus*. Neue Zeit, Berlin

List F 1904 *The National System of Political Economy*. Green and Co., London

Luxemburg R 1951 *The Accumulation of Capital*. Routledge, London

Marx K 1952 *Capital*. Encyclopedia Britannica, Chicago

Schumpeter J 1943 *Capitalism, Socialism, and Democracy*. Allen & Unwin, London

Smith A 1976 *An Inquiry into the Nature and Causes of the Wealth of Nations*. University of Chicago Press, Chicago

Sombart W 1921–8 *Der moderne Kapitalismus*. Duncker and Humbolt, Munich, Germany, Vols. 1–III

Sweezy P M 1942 *The Theory of Capitalist Development. Principles of Marxian Political Economy*. Oxford University Press, New York

Wallerstein I 1974 *The Modern World System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

Weber M 1992 *The Protestant Ethic and the Spirit of Capitalism*. Routledge, London

I. T. Berend

Capitalism: Global

While the idea of international capitalism has been part of political economy and related disciplines for centuries, the concept of global capitalism is of more recent origin. Despite the fact that many scholars have used the terms interchangeably, with the widespread use of the concept of globalization in the 1990s, the force of the argument for a firm distinction between these concepts has increased. International capitalism has generally been conceptualized in state-centrist terms, focused mainly on how national capitalists based on competing national economies and working through national companies operated across borders. The distinctive concept of global capitalism takes its departure from the idea of a global economy dominated by globalizing corporations and those who own and control them, and those in influential positions who serve their interests. The article traces the development of this conception of global capitalism since the 1950s, from the transitional idea of capitalist world-economy through several attempts to establish a genuinely global conception of capitalism not grounded in national economies and societies.

1. Introduction

Researchers on global capitalism have focused on three inter-related phenomena, increasingly significant since the 1960s. These are, first, the ways in which transnational corporations (TNCs) have facilitated the globalization of capital and the production of goods and services; second, the rise of new global forms of organization of the capitalist class; and third, transformations in the global scope of TNCs that own and control the mass media, notably television channels and the transnational advertising agencies and their role in promoting global brand consumer goods and the emergence of a global culture and ideology of consumerism. Theory and research on each of these three phenomena roughly coincide with attention to the economic, political and culture–ideology spheres of global capitalism.

2. Global Capitalism as an Economic System

It is no coincidence that interest in a global capitalist system in contrast to competing national capitalisms increased perceptibly from the 1950s. The context of theoretical and empirical interest in competing national capitalisms was (and for many still is) the history of colonialism and imperialism. This is overlaid with several versions of the theory that capitalist states could more or less successfully plan their own economic futures (see, for example, Keynesianism, regulation theory, and the developmental state). The concept of international capitalism, therefore, refers to a system of interacting and competing national capitalist economies, in which national elites of various types use 'their' big business (and businesses) to further national interests around the world. As direct imperialism and colonialism came to an end and as increasing numbers of very large TNCs began to

emerge in the 1960s, attention began to shift decisively from national to global capitalism.

The dependency approach to development and underdevelopment of Gunder Frank and the related world-systems approach of Wallerstein, both highlighted the systemic nature of capitalism as a worldwide phenomenon over several centuries. While these theoretical innovations can be said to have prepared the ground for it, neither entirely succeeded in establishing a coherent concept of global capitalism for what came to be termed the 'age of development' from the 1950s onwards. This was due to ambivalence over their units of analysis and insufficient focus on the role of the major corporations in development in general. For example, the analysis of core, semiperiphery and periphery in the world-systems approach is based on national economies, as is the theoretically more ambitious concept of commodity chains. The simple assertion that these take place within a world-economy or global economy does not suffice for an analysis of global capitalism.

Theories of global capitalism, in the sense used here, take off from the proposition that capitalism entered a new, global phase in the second half of the twentieth century. By the end of the century, the largest TNCs had assets and annual sales far in excess of the gross national products (GNPs) of most of the countries in the world. In 2000, only about 70 countries out of a total of around 200 for which there were data, had GNPs of more than 10 billion US dollars. By contrast, the *Fortune Global 500* list of the biggest corporations by turnover reported that around 450 of them had annual sales greater than US$10 billion. This comparison, however, underestimates the economic scale of major corporations compared with sovereign states, as most TNC revenues are counted as part of the gross domestic product (GDP) of some countries. A more appropriate measure is to compare TNC revenues with government revenues. Gray (1999) has calculated that for 1997–99, the seven largest economic entities in the world were the budgets of the governments of the USA, Germany, Japan, China, Italy, UK and France, but nine of the top 20 were corporations, and of the richest 100 economic entities by revenues, 66 were TNCs. Thus, in this important sense, such well-known names as General Motors, Shell, Toyota, Unilever, Volkswagen, Nestle, Sony, Pepsico, Coca Cola, Kodak, and Xerox, the huge Japanese trading houses (and many other corporations most people have never heard of?) have more economic power at their disposal than the majority of the countries in the world. These figures indicate the gigantism of TNCs relative to the state budgets of most countries.

Not only have TNCs grown enormously in size since the 1950s, but their global reach has expanded dramatically. Many TNCs regularly earn more than half of their revenues outside the countries in which they are legally domiciled. This is not only the case for TNCs from countries with relatively small domestic markets (for example, Switzerland, Sweden, Canada, Australia), but also for those legally domiciled in the USA and Japan. While most of the biggest corporations are still headquartered in the First World, several dozen companies originating in what is conventionally called the Third World or that part of it known as the newly industrializing countries have been numbered among the 500 biggest companies by revenues in the world. This group has included the state-owned oil companies of Brazil, India, Mexico, Taiwan and Venezuela (owned by the state but increasingly run like private corporations), banks in Brazil and China, and the Korean manufacturing and trading conglomerates (chaebol), some of which have attained global brand-name status (for example, Hyundai and Samsung).

Writers who are skeptical that capitalism is a global system argue that because most major TNCs are legally domiciled in the USA, Japan and Europe, and because they trade and invest mainly between themselves, capitalism is still best analyzed in terms of national corporations. For such writers, the global economy is a myth and, consequently, there is no global capitalist system as such. Against this conclusion, proponents of the salience of a global capitalist system argue that an increasing number of corporations operating outside their countries of origin see themselves as developing global strategies of various types, as is obvious from the contents of their annual reports and other corporate publications. While all parts of all economies are clearly not equally globalizing, an increasing volume of empirical research indicates that the production and marketing processes of most major industries are being de-territorialized from their countries of origin and that these processes are being driven by the TNCs. The central issue for economic globalization is the extent to which TNCs domiciled in the USA, Japan, European and other countries can be more fruitfully conceptualized as expressing the national interest of their countries of origin (the globo-skeptic argument) or what can be conceptualized as the private interests of those who own and control them, aggregated as the interests of global capitalism. Even if historical patterns of TNC development have differed from country to country and region to region, it does not logically follow that TNCs and those who own and control them express any type of 'national' interest or national character.

The formal ownership of capital and the corporations has been transformed since the 1960s. The ownership of share capital has increased throughout the world by means of greater (though still a tiny minority in most communities) participation of the general population in stock markets and the indirect investments that hundreds of millions of people have through their pension funds and other forms of savings. This has led some to argue that economic globalization has created a popular capitalism, though others argue the more elitist thesis that the real drivers

of the global capitalist system are the managers of unit trusts and pension funds. However, formal ownership does not necessarily mean effective control over capital and the resources of TNCs.

The globalization of the international financial and trading system can be fruitfully analyzed in terms of the progressive weakening of the nation-state and the growing recognition that the major institutions of global capitalism, notably TNCs and globalizing financial and trading organizations, are setting the agenda for these weakened nation-states. Theory and research on this issue has, not surprisingly, led to an increased interest in the politics of global capitalism.

3. *Global Capitalism as a Political System*

The politics of global capitalism is debated intensely inside and outside the social sciences. Since the disintegration of the Soviet empire from the late 1980s, the struggle between capitalism and communism has been largely replaced by the struggle between the advocates of capitalist triumphalism and the opponents of capitalist globalization. Many theorists have discussed these issues within the triadic framework of states, TNCs and international economic institutions. From this perspective, the global capitalist system is dominated by the relations between the major states and state-systems (USA, the European Union, and Japan), the major corporations, and the World Bank, the International Monetary Fund, World Trade Organisation (supplemented in some versions by other international bodies, major regional institutions, and so on).

The idea of a transnational ruling class has been suggested by several authors, notably in Cox's thesis (1987) on the emergence of a global class structure and in the work of Gill (1990) on the Trilateral Commission, where he identifies a 'developing transnational capitalist class fraction.' Sklair (2001) proposes a more explicit concept of a transnational capitalist class, and it plays a central role in his theory of the capitalist global system. Here, the transnational capitalist class is the characteristic institutional form of political transnational practices in the global capitalist system (paralleling the role of transnational corporations in the economic sphere and consumerism in the culture–ideology sphere). In this formulation the transnational capitalist class is analytically divided into four main fractions: (a) TNC executives and their local affiliates; (b) globalizing bureaucrats and politicians; (c) globalizing professionals; (d) consumerist elites (merchants and media).

A transnational capitalist class might be transnational in at least three senses. First, its members would tend to have outward-oriented global rather than inward-oriented national perspectives on a variety of issues, for example, support for free trade and neoliberal economic and social policies. Second, they would tend to be people from many countries, more and more of whom begin to consider themselves as citizens of the world as well as of their places of birth and residence (these might differ). And third, they would tend to share similar lifestyles, particularly patterns of higher education (in cosmopolitan business schools) and consumption of luxury goods and services. Theory and research to support this argument are as yet in a very embryonic phase. Nevertheless, and despite real geographical and sectoral conflicts, the whole of the transnational capitalist class shares a fundamental interest in the continued accumulation of private profit on a global scale.

Many other Marxist, anti-Marxist and Marx-inspired scholars see capitalism as a global system, but tend to conceptualize globalization in much wider terms and, thus, minimize the importance of global capitalism as an explanatory variable. The most important of these are the geographer, Harvey (1989), whose notion of time–space compression has been very influential, and the sociologist, Giddens (1990), whose conception of global capitalism is but one element in his theory of globalization as a product of late (and reflexive) modernity. Both of these contributions are significant for their attempts to build a bridge between the debates around economic, political and cultural globalization.

4. *Global Capitalism as a Culture–Ideology System Dominated by Large Globalizing Media and Advertising Corporations*

The third distinctive aspect of capitalism as a global system is the worldwide diffusion and increasingly concentrated ownership and control of the electronic mass media, particularly television. For example, the number of TV sets per capita has grown so rapidly in developing countries (from fewer than 10 per 1,000 population in 1970 to 145 per 1,000 in 1995, according to UNESCO) that many researchers argue that a globalizing effect due to the mass media is taking place all over the world.

Ownership and control of television, including satellite and cable systems, and associated media (like newspaper, magazine and book publishing, films, video, records/tapes/compact disks, and a wide variety of other marketing media, notably the internet), are concentrated in relatively few very large TNCs. The predominance of US-based corporations is being challenged by Japan, Europe and Australia-based groups globally, and even by Third World corporations like the Brazil-based media empire of TV Globo.

The culture–ideology of consumerism prioritizes the exceptional place of consumption and consumerism in contemporary capitalism, increasing consumption expectations and aspirations without necessarily ensuring the income to buy. The extent to which economic and environmental constraints on the pri-

vate accumulation of capital challenge the global capitalist project in general and its culture–ideology of consumerism in particular, is a central issue for theory and research on the capitalist global system. Nevertheless, it should be pointed out that most scholars studying these issues of global culture do so not in terms of capitalism but in terms of the potential impact of global culture on national and local cultures and identities.

5. *Crises of Global Capitalism*

While Marxist and Marx-inspired theories of the inevitability of a fatal economic crisis of capitalism appear to have lost most of their adherents, at least two related but logically distinct crises of global capitalism have been identified. The first is the simultaneous creation of increasing poverty and increasing wealth within and between societies (the class polarization crisis), not to be confused with Marx's emiseration thesis which failed to predict enormous increases in wealth for rapidly expanding minorities all over the world. The second is the unsustainability of the system (the ecological crisis). These crises are often interpreted through the prism of consumerism inherent in a global capitalist system based on globalizing corporations (Sklair 2001). Globalizing corporations (in some cases rather more clearly than national governments) recognize the class crisis, but largely in marketing terms. In most communities around the world the absolute numbers of people who are becoming global consumers have been increasing rapidly over recent decades. However, it is also true that in some communities the absolute numbers of the destitute and near-destitute are also increasing, sometimes alongside the new rich consumers. The best available empirical evidence (see United Nations Development Programme Human Development Report, published annually since 1990) suggests that the gaps between rich and poor have widened since the 1980s in many parts of the world. The very poor cannot usually buy the goods and services that global capitalists sell. While there is a long way to go before consumer demand inside the rich First World is satisfied, the gap between the rich and the poor all over the world is not welcome news for TNCs. In addition to the profits lost when poor people who want to buy goods and services do not have the money or even the credit to do so, the increasing visibility of the new rich and the new poor in an age of constant global media exposure directly challenges capitalist claims that everyone eventually benefits from economic globalization.

The ecological crisis is also directly connected with consumerism, encapsulated in the political and ideological struggles over the concept of sustainable development. Since the 1980s most major TNCs have developed detailed policies in response to the challenges of environmental harm and declining stocks of resources essential for the maintenance of global capitalist consumerism. However, the persistence of problems of pollution, health risks, environmental degradation and waste management intrinsic to the system, suggests that it will be difficult to avoid ecological crisis. In this context, the attempt by TNCs and their supporters in international bureaucracies, the professions, government (globalizing bureaucrats and politicians), and the mass media to capture the idea of sustainable development and to reconcile it with capitalist globalization is worth further study.

6. *Resistance to Global Capitalism*

Global capitalism is often seen in terms of impersonal forces (notably market forces, free trade) wreaking havoc on the lives of ordinary and defenseless people and communities. It is not coincidental that interest in economic globalization has been accompanied by an upsurge in what has come to be known as New Social Movements (NSM) research. NSM theorists, despite their substantial differences, argue that the traditional response of the labor movement to global capitalism, based on class politics, has failed. In its place, a new analysis based on identity politics (notably of gender, sexuality, ethnicity, age, community, belief systems) has been developed, directed towards resistance to sexism, racism, environmental damage, war-mongering, capitalist exploitation and other forms of human rights abuses.

The globalization of identity politics involves the establishment of global networks of people with similar identities and interests outside the control of international, state and local authorities. There is a substantial volume of research and documentation on such developments in the women's, peace, indigenous peoples' and environmental movements, some of it in direct response to perceived TNC malpractices. This provides a series of research-rich connections for scholars influenced by postmodernist and global capitalist theories.

Serious challenges to global capitalism in the economic sphere have also come from those who 'think global and act local.' This normally involves disrupting the capacity of TNCs and global financial institutions to accumulate private profits at the expense of their workforces, their consumers and the communities that are affected by their activities. An important aspect of global capitalism is the increasing dispersal of the manufacturing process into many discrete phases carried out in many different places, popularized by Dicken (1998) as global shift. Being no longer so dependent on the production of one factory and one workforce gives capital a distinct advantage, particularly against the strike weapon that once gave tremendous negative power to the working class. Global production chains can be disrupted by strategi-

cally planned stoppages, but this is generally more of an inconvenience than a real weapon of labor against capital. The global division of labor builds flexibility into the system so that not only can capital migrate anywhere in the world to find the cheapest reliable productive sources of labor but also few groups of workers can any longer decisively 'hold capital to ransom' by withdrawing their labor. At the level of the production process, globalizing capital has a substantial advantage over labor. In this respect, the global organization of the TNCs and allied institutions like the World Bank and the WTO have generally proved to be too powerful for local labor and community organizations.

Nevertheless, global capitalists, if we are to believe their own propaganda, are continuously beset by opposition, boycott, legal challenge, and moral outrage from the consumers of their products, concerned citizens, and by disruptions from their workers. There are also many ways to be ambivalent or hostile about cultures and ideologies of consumerism, some of which the Green movement has successfully exploited.

The issue of democracy is central to the prospects for global capitalism and the struggle against it. The rule of law, freedom of association and expression, freely contested elections, as minimum conditions and however imperfectly sustained, are as necessary in the long run for mass market based global consumerist capitalism as they are for alternative social systems.

While most theory and research on capitalism continues to be state-centrist, focusing largely on how it works within specific countries, the growing influence of globalization in the social sciences appears to be encouraging more scholars to consider capitalism in the global as well as the local and/or national context.

See also: Capitalism; Economic Growth: Theory; Global History: Universal and World; Globalization and World Culture; Globalization: Geographical Aspects; Globalization: Political Aspects; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; International Business; International Marketing; International Trade: Economic Integration; International Trade: Geographic Aspects; Marx, Karl (1818–89); Multinational Corporations; World Systems Theory

Bibliography

Cox R W 1987 *Production, Power, and World Order: Social Forces in the Making of History*. Columbia University Press, New York

Dicken P 1998 *Global Shift: Transforming the World Economy*, 3rd edn. Paul Chapman, London

Gereffi G, Korzeniewicz M (eds.) 1994 *Commodity Chains and Global Capitalism*. Praeger, Westport, CT

Giddens A 1990 *The Consequences of Modernity*. Polity Press, Cambridge, UK

Gill S 1990 *American Hegemony and the Trilateral Commission*. Cambridge University Press, Cambridge, UK

Gray C 1999 *Corporate Cash*. WEP, Eugene, OR

Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Cambridge, MA

Herman E S, McChesney R 1997 *The Global Media: The New Missionaries of Corporate Capitalism*. Cassell, London

Lechner F J, Boli J (eds.) 1999 *The Globalization Reader*. Blackwell, Malden, MA

Ross R J S, Trachte K C 1990 *Global Capitalism: The New Leviathan*. State University of New York Press, Albany, NY

Sklair L 1995 *Sociology of the Global System*, 2nd edn. Johns Hopkins University Press, Baltimore, MD

Sklair L 2001 *The Transnational Capitalist Class*. Blackwell, Malden, MA

Strange S 1996 *The Retreat of the State: The Diffusion of Power in the World Economy*. Cambridge University Press, Cambridge, UK

United Nations Development Programme 1990–onwards *Human Development Report*, Oxford University Press, New York

van der Pijl K 1998 *Transnational Classes and International Relations*. Routledge, London

Wallerstein I 1979 *The Capitalist World Economy*. Cambridge University Press, Cambridge, UK

L. Sklair

Cardiovascular Conditioning: Neural Substrates

Our existence relies on our ability to alter our future behavior as a function of our past experiences, a phenomenon known as learning. This requires that a permanent record of our experiences, our memories, be stored in the brain to be recalled to guide our future actions. The last three decades of the twentieth century witnessed intense investigations of the brain areas where memories are formed and stored, and the structural changes that represent them. Many neuroscientists have adopted Pavlovian learning paradigms in animals to aid in these investigations. These relatively simple paradigms present an opportunity to identify the brain circuitry and mechanisms that contribute to learning as reflected in the acquisition of specific learned responses. Among these responses are learned cardiovascular responses, and particularly learned heart rate responses, which are the focus for this article.

1. Pavlovian Heart Rate Conditioning: A Model to Assess the Brain Circuits that Contribute to Memory

1.1 Pavlovian Conditioning

During Pavlovian conditioning the relationship among events is learned. For example, an animal learns that one event, such as the presentation of an

auditory or visual stimulus, that repeatedly precedes a second event, such as the presentation of an electric shock or a pleasant morsel of food, provides information about the occurrence of the second event.

The auditory or visual stimulus is called the conditional stimulus (CS), and the electric shock or morsel of food is called the unconditional stimulus (US). This learned relationship or association is stored in memory such that subsequent presentations of the CS will elicit expectations regarding the occurrence of the US. These expectations elicit responses, called conditioned responses (CRs), which are a specific consequence of the formation and storage of the association between CS and the US. Implicit in the search for the structural changes that form the substrates for Pavlovian associative memories is the assumption that if an association between the CS and US is made, neuronal information concerning both must converge at a common brain structure(s). Thus, research directed at identifying structures involved in memory has been guided by an identification of structures where CS and US information converge.

1.2 Heart Rate Conditioning

Researchers have used heart rate CRs as model CRs to identify the areas where memories are formed and stored. These CRs are advantageous because the locations of the motor neurons that produce them are known. Thus, structures that send projections to activate these neurons leading to the expression of the CR can be identified using anatomical methodology. The subsequent identification of areas that send information to the areas that in turn activate the motor neurons, including the pathways by which CS and US information, access these areas, exposes an entire circuit that contributes to memory. Identification of the circuit permits the identification of sites of CS and US convergence and sets the stage for analyses of the structural changes that form the substrates for memory of the CS–US association.

1.3 An Early Model: Heart Rate Conditioning in the Pigeon

David Cohen (1980) was the first to use a heart rate CR to identify the components of a circuit that contributes to memory. He measured the accelerative heart rate CR in the pigeon that developed in response to a visual CS which preceded an electric current (US) applied to the foot. He believed that a systematic analysis of the flow of neuronal information, from CS and US input to CR output, should eventually lead to the identification of sites of CS and US convergence. While the pigeon model is no longer used, Cohen's strategy proved successful and was adopted by others using other models. Importantly, he identified an area where the structural changes responsible for the memory of the association potentially may occur. That area was the avian homologue of the mammalian amygdala. Its destruction blocked the development of the CR. This finding guided future research focused on the mammalian amygdala as a component of a brain circuit that contributes to learning and memory, using a heart rate CR in the rabbit as a model response.

2. Heart Rate Conditioning in the Rabbit

2.1 Circuit Components

The heart rate decelerative, or bradycardic, CR in the rabbit has been widely used to identify the components of the circuit that contribute to memory formation and storage (see Kapp et al. 1998 and Powell 1994 for reviews). This CR develops when an acoustic CS immediately precedes either an aversive or appetitive US. The neurons comprising the final motor pathway for the expression of this CR are located in the medulla, and their axons travel via the vagus nerve to the heart. The identification of this CR pathway created the opportunity to identify the structures that may activate it. As noted in Sect. 1.3, Cohen's work pointed to the amygdala, and research in the rabbit has demonstrated an important contribution of a group of neurons in this structure to CR expression. These neurons are located in the central nucleus (ACe) and project directly and indirectly to the final motor path neurons. Electrical stimulation of the ACe produces bradycardia and lesions of the ACe markedly attenuate the development and expression of the CR (Kapp et al. 1998). Further, increases in the activity of ACe neurons developed to the CS during conditioning (Applegate et al. 1982, McEchron et al. 1995). These results suggest that these neurons excite the motor neurons leading to CR expression and are an important component of the circuit for the development and expression of the CR (Fig. 1). Recall that an assumption in the search for the neural substrates of Pavlovian associative memories is that CS and US information must converge at a common site(s). Thus, identification of the pathways conveying CS and US information to central structures such as the ACe is an important component of the analysis. The most peripheral components of the acoustic CS pathway have yet to be identified. However, research using an acoustic CS in the rat has demonstrated that destruction of neurons in the inferior colliculus, a structure that receives information from the most peripheral components of the auditory system (Fig. 1), blocks the development of several Pavlovian CRs (LeDoux 1995). The inferior colliculus in turn projects to several other auditory structures including the magnocellular component of the medical geniculate nucleus (MGm). MGm destruction blocks CR development in the rabbit (McCabe et al. 1993). The MGm projects to the lateral amygdaloid nucleus (AL), and AL neurons respond to acoustic stimuli (Quirk et al. 1995). Neurons in the AL project both directly and

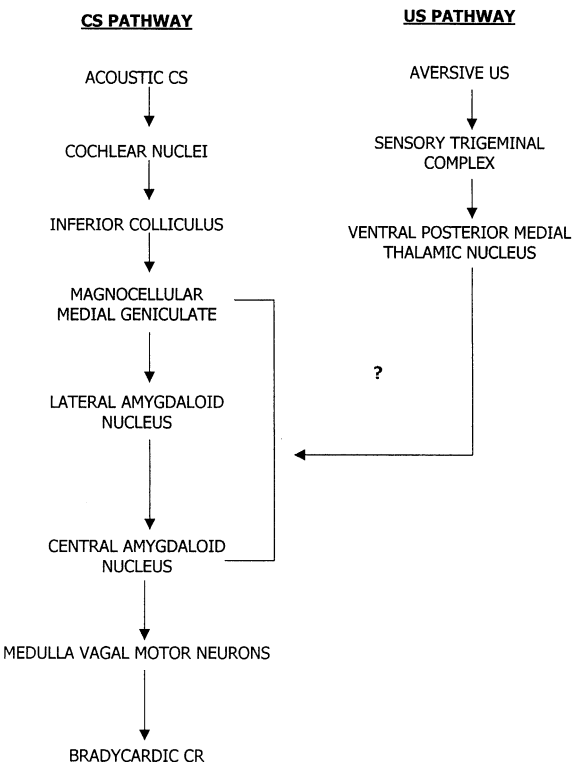

Figure 1
A simplified diagram of the putative structures and pathways comprising the brain circuit that contributes to the acquisition of the bradycardic CR in the rabbit

indirectly via intra-amygdaloid connections to the ACe (Fig. 1).

Less is known about the pathway by which US information gains access to central structures to converge with CS information. This pathway may be comprised of projections from spinal trigeminal complex neurons, which are activated by US presentations (McEchron et al. 1996b) and which project to the ventral posterior medial thalamic nucleus. The latter's destruction in the rabbit prevents the development of the bradycardic CR (McEchron et al. 1995). The pathway by which the US information is transmitted to converge with CS information in more central structures is at present unknown. However, recordings of neurons in the CS pathways to determine if they respond to CS and US presentations indicates sites where convergence occurs, as described in Sect. 2.2.

2.2 Potential Sites of Memory Formation: The Convergence of CS and US Information

The above research suggests that CS information accesses the amygdala, and that the ACe produces CR expression via an influence on vagal motor neurons (Fig. 1). Does CS and US convergence occur in the amygdala, making it a potential candidate for the site where the memory is formed? Or, does convergence occur in other areas as well? With respect to the ACe, neurons within this structure are responsive to both the CS and US (McEchron et al. 1995), and the responses that develop to the CS in this nucleus during conditioning may well reflect a structural change(s) that represents the memory for the association. However, the mere development of these responses is not conclusive evidence that it is a site where this change occurs. Such responses may reflect their relay from other brain regions. For example, neurons in the rabbit and rat MGm, and in the rat AL, respond to both the CS and US (Bordi and LeDoux 1994, Romanski et al. 1993. McEchron et al. 1996a). Furthermore, responses in the MGm develop to the CS in many species during Pavlovian conditioning (Edeline et al. 1988), McEchron et al. 1995), and such changes also develop in the rat AL (Quirk et al. 1995). Thus, the responses that develop in the ACe may represent responses relayed from the MGm to the AL and from the AL to the ACe. Importantly, increases in synaptic conductivity that indicate a structural change have been demonstrated in both the MGm and AL during Pavlovian conditioning McEchron et al. 1996a, Rogan et al. 1997). The as-yet-unidentified structural substrates for these increases may well represent the neural basis for the memory of the association formed between the CS and US, and the resultant heart rate CR that develops. The extent to which increases in synaptic conductivity occur in the ACe has yet to be demonstrated. To the extent that it does occur in the ACe, then the overall picture is one in which the structural change(s) that form the substrate for the associative memory during Pavlovian conditioning may occur at several sites within the essential circuit.

As reviewed in Sect. 2.1 considerable research devoted to an identification of the essential circuit for conditioning bradycardia has focused on the amygdala and MGm. It is important to realize, however, that other brain structures also appear to contribute importantly to the development of this CR. Two in particular, the prefrontal cortex and the cerebellar vermis, make important contributions (Powell 1994, Supple and Kapp 1993). Lesions of both areas severely retard CR development, and neuronal activity develops to the CS in both areas. The functional interactions of these structures with the MGm and amygdala in the development of the bradycardic CR will be an important focus for research in the early twenty-first century. An equally important focus will be (a) a determination of the exact contribution that each of these components makes to CR development, and (b) the exact site(s) and nature of the cellular substrates for Pavlovian associative memories. Obviously, much needs to be accomplished if we are to completely understand their neural basis.

See also: Autonomic Classical and Operant Conditioning; Classical Conditioning and Clinical Psychology; Classical Conditioning, Neural Basis of; Coronary Heart Disease (CHD): Psychosocial Aspects; Eyelid Classical Conditioning; Pavlov, Ivan Petrovich (1849–1936)

Bibliography

Applegate C D, Frysinger R C, Kapp B S, Gallagher M 1982 Multiple unit activity recorded from amygdala central nucleus during Pavlovian heart rate conditioning in rabbit. *Brain Research* **238**: 457–62

Bordi F, LeDoux J E 1994 Response properties of single units in areas of rat auditory thalamus that project to the amygdala. II. Cells receiving convergent auditory and somatosensory inputs and cells antidromically activated by amygdala stimulation. *Experimental Brain Research* **98**: 275–86

Cohen D H 1980 The functional neuroanatomy of a conditioned response. In: Thompson R F, Hicks L H, Shvyrkov V B (eds.) *Neural Mechanisms of Goal Directed Behavior and Learning*. Academic Press, New York

Edeline J M, Dutrieux G, Neuenschwander-El, Massioui N 1988 Multiunit changes in hippocampus and medial geniculate body in freely behaving rats during acquisition and retention of a conditioned response to a tone. *Behavior and Neural Biology* **50**: 61–79

Kapp B S, Silvestri A J, Guarraci F A 1998 Vertebrate models of learning and memory. In: Martinez J L, Kesner R P (eds.) *Neurobiology of Learning and Memory*. Academic Press, New York

LeDoux J E 1995 Emotion: Clues from the brain. *Annual Review of Psychology* **46**: 209–35

McCabe P M, McEchron M D, Green E J, Schneiderman N 1993 Electrolytic and ibotenic acid lesions of the medial geniculate-prevent the acquisition of classically conditioned heart rate to a single acoustic stimulus in rabbits. *Brain Research* **619**: 291–8

McCabe P M, McEchron M D, Green E J, Schneiderman N 1995 Destruction of neurons in the VPM thalamus prevents rabbit heart rate conditioning. *Physiology and Behavior* **57**: 159–63

McEchron M D, Green E J, Winters R, Nolen T G, Schneiderman N, McCabe P M 1996a Changes of synaptic efficacy in the medical geniculate nucleus as a result of auditory classical conditioning. *Journal of Neuroscience* **16**: 1273–83

McEchron M D, McCabe P M, Green E J, Hitchcock J M, Schneiderman N 1996b Immunohistochemical expression of the c-Fos protein in the spinal trigeminal nucleus following presentation of a corneal airpuff stimulus. *Brain Research* **710**: 112–20

McEchron M D, McCabe P M, Green E J, Llabre M M, Schneiderman N 1995 Simultaneous single unit recording in the medial nucleus of the medial geniculate nucleus and amygdaloid central nucleus throughout habituation, acquisition, and extinction of the rabbit's classically conditioned heart rate. *Brain Research* **682**: 157–66

Powell D A 1994 Rapid associative learning: Conditioned bradycardia and its central nervous system substrates. *Intergrative Physiological and Behavioural Science*. **29**: 109–33

Quirk G J, Repa J C, LeDoux J E 1995 Fear conditioning enhances short-latency auditory responses of lateral amygdala neurons: Parallel recordings in the freely behaving rat. *Neurons* **15**: 1029–39

Rogan M J, Staubli V V, Le Douse J E 1997 Fear conditioning induces associative long-term potentiation in the amygdala. *Nature* **390**: 604–7

Romanski L M, Clugnet M C, Bordi F, LeDoux J E 1993 Somatosensory and auditory convergence in the lateral nucleus of the amygdala. *Behavioral Neuroscience* **107**: 444–50

Supple W F Jr, Kapp B S 1993 The anterior cerebellar vermis: Essential involvement in classically conditioned bradycardia in the rabbit. *Journal of Neuroscience* **13**: 3705–11

B. Kapp

Care and Gender

Care consists of the physical, emotional, and intellectual processes that enable humans to maintain their lives and activities; such activities usually are distinguished from economic production. Care is a central aspect of human life, but social scientists have only recently begun to pay careful attention to the connection of care and gender. Since most activities of care historically have belonged properly in the same sphere as the family and been the activity of women, slaves, servants, and working-class people, care has been beneath the concern of most social theorists and scientists. Changes in the roles of family and public institutions for care, as well as feminist scholars' impetus to study the concerns of women's lives, have led to the emergence of a robust field of study that explores the physical, social, political, economic, and philosophical implications of care.

1. The Growing Importance of Care for Public Concern

The long-standing devaluation of care grows out of the devaluation of its dual central aspects. On the one hand, care is about the actual physical work required to maintain and repair objects and people. On the other hand, care also denotes mental states of engagement that make central the concerns of the person, object, idea, etc. toward which the care is directed. Thus, in the English language care is associated with both burdens and woes. Even in current discussions of care some scholars emphasize the physical concerns of care, others see it as a psychological or philosophical category, while others try to combine these two elements.

The historical exclusion of care from public life and from serious scientific consideration also reflects the duality of care's nature (Tronto 1996). Aristotle began

his *Politics* by distinguishing between the worlds of politics and household and relegating all the banal and daily concerns of the household as outside of the concerns of citizens. To such thinkers, care was beneath the dignity of citizens. Others have viewed care as too lofty for any serious association with public life. Realist thinkers from Augustine to Max Weber have been suspicious of attitudes of care because it was associated with ideals of Christian charity and agape, and have relegated such concerns to realms beyond public concern, either to spiritual communities or to individuals' consciences.

A major cause for this greater attention from social scientists is that care itself has changed during the industrialization since the 1850s (see *Family and Gender*; *Household Production*). Households used to be almost entirely self-contained. They produced foodstuffs, domestic products such as soap, energy, etc. for themselves. Birth, death, and illness were treated in the household with minimal assistance from others. Households consisted of people who were in various stages of their life. With the rise of industrialization, households have grown smaller, most household goods have become commodified, and professionals have begun to assist in processes of birth, death, and illness. Thus, activities that used to be conceived of as essentially private and personal became social, public, and political concerns. The development of public institutions since the eighteenth century to aid in caring for the ill and infirm (Foucault 1965) has accelerated with the development of welfare state bureaucracies (Bussemaker and van Kersbergen 1994). As care has become a more central aspect of social and political life, these questions have become more important for scholars: who cares for whom, how is care organized and paid for in different societies, and what are the practical and normative questions entailed in practices of care?

2. *Conceptualizing Care*

Scholars use different formulations of care depending upon the different conceptual purposes to which they employ the concept. Among these uses are: care as a sociological category; i.e., an account of how societies organize care work; care as a psychological orientation, i.e., a framework for understanding psychological differences between men and women; care as an ethical category, i.e., a basis for making ethical judgments in care settings; care as a political and philosophical perspective, i.e., a paradigm in which to understand an alternative account of human nature and values. The empirical work that ties care to gender is strongest in the psychological and sociological approaches, but within all of these accounts of care, the relationship of care to gendered activities and to the relative status of men and women in society remains central.

2.1 *The Sociological Approach*

Among the first issues that feminist scholars and activists noted at the beginning of the second wave of feminism was that, even in industrial societies, a deep division of labor persisted that left to women most of the tasks of caring. Marxist scholars frequently distinguished this realm of reproduction from production. Although conservatives argue that childcare and tending to the household were naturally the realm of women's work, feminists began to challenge both the gendered division of labor and the devaluation of women's work. Political platforms throughout the West provided analyses of these discrepancies and called for policy solutions such as available day care and wages for housework. Scholars soon began to explore such 'labors of love' (Finch and Groves 1983). A large sociological literature has now emerged that explores the boundaries and nature of personal care within and outside of traditional family contexts and institutionalized care in society.

2.2 *The Psychological Approach*

Gilligan's path breaking work, *In a Different Voice* (1982), is often taken as the starting point for the discussion of care. Gilligan's work challenged the basic assumption of Lawrence Kohlberg's influential account of cognitive moral development by arguing that not all moral development could be measured by a single path. Gilligan posited that, in addition to Kohlberg's account of the development of justice reasoning, psychologists needed to observe as well the development of an orientation to care. To Gilligan, three qualities describe this ethic and distinguish it from the justice approach: (a) it revolves around the moral concepts of responsibility and relationships rather than rights and rules; (b) it is tied to concrete circumstances rather than to abstract rules of morality; (c) it is an activity rather than a set of principles. Gilligan insisted that these three orientations need not be associated with gender, but her admonition was widely disregarded. Despite the absence of empirical evidence to support the claim that these two orientations distinguish between men and women, a broader public audience seized the idea that there was a gender difference in morality that distinguished men and women; men were more interested in impartial and principled conceptions of justice, and women were more interested in maintaining particular relationships of care. Arguments for care thus became associated with a political strategy to valorize the lives and experiences of women.

2.3 *The Ethical Approach*

Noddings (1984) transformed the assumption of women's more caring nature into a moral position.

Arguing that care was always a dyadic relationship, Noddings posited that there was a limit to the kinds of concerns that caring could address; they were limited to the kinds of relationships and life activities that occurred in intimate settings.

Ruddick's (1990) important work on 'maternal thinking' reinforced the notion that caring had its own set of moral norms and practices that might be distorted if applied in a broader context (see *Motherhood: Social and Cultural Aspects*). It is perhaps not surprising that proponents of this account of care were often exploring care relationships within social institutions and professions where women predominated, such as mothering (Ruddick 1990), nursing (Benner and Wrubel 1989), and in education (Noddings 1984).

2.4 The Political and Philosophical Perspective

Critics of the psychological and narrow ethical positions soon emerged, arguing that care was necessarily neither gendered nor limited to a narrow sphere of life (Sevenhuijsen 1998, Tronto 1993). These critics began to define care much more broadly and to think about it as an essential human activity, whose particular contours would be shaped by broader political and social decisions made in a society. Writing about the development of institutions for caring in the Scandinavian welfare states, Waerness (1984) argued that a distinctive 'rationality of care' could be discerned in agencies and practices that were involved in issues of care. This broader conception of care combines the previous frameworks: it understands care to be both work and a framework of values about work, to be both concrete physical activity and normative orientations and concerns about that activity. It combines the critical dimension of the sociological approaches that recognize the devalued nature of care with the psychological and ethical approaches that seek to reconstruct the values of the caring work and activities that are done. From this perspective, the persistence of the gendered nature of care is a sign of a broader problem in devaluing certain necessary human traits and activities, not only a devaluation of that which is seen as feminine or residing in women's sphere. Furthermore, scholars have begun to connect the devaluation of care with the devaluation of women's lives and activities. The relative powerlessness of care-givers is tied to the relatively low status of the work that they perform (Tronto 1993).

3. Care as a Challenge to How We Organize Social and Political Life

Scholars have begun to apply these various care frameworks to explore a number of arenas of social life. Concerns about care have informed how sociologists have studied the organization of institutions to care for children, the elderly, the disabled, the infirm. Scholars in communication have begun to use a care perspective to explain how care can alternatively frame listening and communicating with others. Legal scholars have used an ethic of care to explain alternative ways to understand the legal process. Economists have begun to explore how economists' theories and policy proposals have been distorted by the failure to recognize the fact that women bear most of the costs of rearing children, while all share in the public benefits of well-reared workers and citizens (Folbre 1994).

More generally, political and social theorists have noted that to take the care perspective seriously challenges some of the basic assumptions that seem to permeate Western value systems about human life. The care perspective stresses human vulnerability and dependence. It argues that all individuals are dependent upon others and that an understanding of individuals as interdependent rather than as autonomous, is actually a more realistic portrayal of their existence. Since most contemporary political philosophies rest upon an assumption of the individual as a rational, autonomous individual, the care perspective thus raises a fundamental challenge to them.

Another set of challenges concerns how we organize social and political life by dividing the public sphere and the private sphere. Feminist scholars have long argued that the division of life into public and private spheres is one of the most important demarcations in reducing women, associated with the private, to second-class status (Pateman 1988). In contemporary industrialized societies, care is not divided easily into public and private: many previously private care concerns have become part of the purview of welfare state institutions, and many public activities are carried out to accomplish care.

4. Applying Care Perspectives: The Work of Care in Society

Since scholars actually are thinking about a variety of concerns when they discuss care, they might use the term in a variety of ways. No standard definition of care has yet emerged, in part because the following problems seem to be essentially contested within the meaning of the term: (a) does care apply to care-work that is directed at another, or can it also be directed toward the self? To include the self adds dimensions of spirituality to the conception of care. It also makes it more difficult to associate care with altruism. Yet advocates of incorporating the care of the self with other forms of care posit that introspections about the meaning of care are often important sources for thinking about care, and that there is another place within the framework of talking about care to distinguish between care for the self and care for others. (b) Does care primarily concern mental and emotional

attitudes or does it primarily concern physical activities of care giving? Scholars who tend to emphasize the mental and emotional aspects of care often do so by ignoring the range of sites of care in society and emphasizing only the most familiar aspects of care, such as mothering. Scholars who focus on the activities of care, especially as carried out in social institutions often emphasize care as work to the exclusion of seeing other dimensions of care. (c) What are the best ways to describe different aspects of care? Fisher and Tronto (1990) provided a framework for dividing care into four phases: caring about, caring for, care giving, and care receiving. By thinking of care as an ideal of a complete process in which all of the phases are present, it is possible to provide a standard by which one can evaluate how effectively care is rendered in any particular setting. (d) How can one account for the power differentials in different forms of care? Other scholars have emphasized that it makes sense to classify care according to the relative power positions of those involved in the caring relationship. Thus, Waerness (1984) distinguishes between 'personal service' and 'necessary care.'

Personal service is work that one could have performed by oneself but calls upon others to do. Necessary care is the care needed by people who cannot provide such care to themselves. Necessary care may require particular expertise (e.g., providing medical or therapeutic care) or it might be care that is not specialized (e.g., helping the frail elderly with transportation needs). (e) How are needs for care determined? While on some basic level the needs for care are obvious, in any given society there can be a variety of ways to meet caring needs and to conceptualize and prioritize caring needs. For example, though elderly people may need assistance in moving about, different societies will determine that elderly parents should reside with their grown children, that public services should provide assistants to help the elderly in their homes, or that elderly people should live in specially designed communities or institutional settings designed for them. Indeed, any particular society may decide to allow people to choose among many different options in deciding how best to meet the needs of care. A society may decide to leave such decisions to individuals, to individual families, to the market, or to governmental agencies.

5. *Applying Care: Philosophical Concerns About Care in Society*

A number of philosophical issues has arisen as scholars have begun to think about care and gender.

5.1 *The Care–Justice Debate*

One central way to frame the philosophical dispute about the value of care has been to see it as a question of care versus justice. In this debate, the framework delineated by Gilligan's (1982) psychological theory is taken as definitive of a difference in orientations toward the moral world. Thus, the question becomes, are care and justice perspectives compatible? Throughout the 1980s, scholars emphasized the incompatibility of the approaches, often to the detriment of taking the care approach seriously. Kohlberg himself described care as a secondary kind of moral orientation that came into play to resolve local moral issues once the larger issues of justice had been settled. Some feminists tried to argue that the model of care was preferable to one of justice (Held 1993) while others insisted that the qualities of localness inevitably made care a flawed conception for understanding ethical issues (Jaggar 1995). Increasingly, feminist theorists resolve the care–justice debate by asserting that the two approaches were compatible and could complement each other, for example, by insisting upon the rights of care workers (Bubeck 1995, Kittay 1999) or by discussing caring in terms of rights (Tronto 1993).

5.2 *Problems With Care Theories*

Nevertheless, advocates of care theory recognize several serious problems in using care as a moral approach. One problem is the problem of parochialism. If people care most for those who are closest to them, how can they make judgments about the (perhaps more serious) needs for care of those who are more distant from them? Many care theorists acknowledge that such problems of parochialism, of being too partial to the needs of those to whom we are closest, require that care be supplemented with more abstract principles of justice. Care theorists might demand, for example, that justice requires that everyone be provided with minimal levels of care. At its most profound level, the problem of parochialism raises the question of what are the appropriate limits and boundaries of care. Does the ethical imperative to care stop at the family, local community, group, or nation, or is it global (Robinson 1999)?

A second serious problem is the problem of paternalism. How can approaches of care avoid the problem that care-givers often have a degree of power over the people for whom they care? The problems of abuse in care relationships are serious concerns. Even when abuse is not present, however, there is often conflict among the parties in a situation of care about how to best provide the care. Since the care-giver, especially in cases of necessary care, is usually in a position of providing the care receiver with something that the care receiver needs, givers have an upper hand in defining the situation and in determining the nature and extent of care. A third serious problem is in trying to determine a standard for good care. If care must be understood as relative to different circumstances, then on what basis can one ever judge what constitutes

adequate or good care? Again, since care receivers are often in positions of receiving care, they are often in a weak position to argue for their own views of what caring should be provided and how. Even if care-givers are relatively devalued in society, they may still be in positions of power vis-à-vis their charges and possibilities for abuse or mistreatment are serious. A fourth serious problem concerns how we define equality in the light of the need for care for dependent individuals in democratic societies. If some individuals are highly dependent upon care, then in what sense can they count as the equals of other democratic citizens? Indeed, Kittay (1999) has noted that the care-givers of highly dependent people also are excluded from some aspects of social and political life by their need to care for their charges. For example, single mothers of disabled children provide exhausting hours of care which renders them unable to earn money for the household. Care workers are especially vulnerable to exploitation (Bubeck 1995).

5.3 *Care and Citizenship*

At another level, scholars have begun to question whether the very definition of what constitutes a citizen has to be conceived according to changes in caring practices. A model of citizenship that makes many benefits contingent on participation in paid employment disadvantages women (and those few men) who (must) leave the paid work force to fulfill their continuing care obligations to children, elders, or disabled relatives (Knijn and Kremer 1997).

6. *Areas for Future Research*

Scholars throughout the industrialized world have begun to question how care fits into the framework of modern industrialized countries. As the model of the family as headed by a breadwinner and homemaker continues to decline, as more caring functions leave the household and become marketed commodities or services (Ungerson 1997) or provided by social institutions, care practices will continue to evolve. But the meaning and significance of care will also continue to evolve as individuals perceive that the relationships between care-givers and care receivers continue to be intensely personal and intimate.

One important part of care is the interpersonal relationship that it represents. To what extent can advanced societies continue to provide personal care? The professionalization of care in the last century (the growth of specialties such as doctoring, nursing, teaching, etc.) may in fact be matched by a process of deskilling of care in the next century as tasks are further differentiated and separated from one another.

Yet as important as the interpersonal dimensions of care remain, we must keep in mind how easily such interpersonal relationships of care can mask the vast inequalities that exist when we ask, who cares for whom? People of relative privilege in the industrialized world are able to command vast amounts of necessary care and personal service. On the other hand, those who are the poorest members of the global society not only receive less essential care, they are often hired at subsistence wages to work and provide care for others. The globalizing economy has also commodified and globalized the traffic in care workers.

Even as theorists of care have warned against the parochialisms that often creep into care analyses, they have only begun to confront the question of whether there is a moral imperative to organize care on a global level. To what extent does one nation owe care to the citizens who live within another nation? To what extent can such concerns be extended before they are stretched too far to bear any resemblance to the kinds of concerns that we usually associate with 'caring'?

How much provision of care is enough? Tragically, there will always be more needs for care than there is care available in human societies. How then are difficult decisions about the allocation of care best made? In order for care to take place, there must first be a perception of the need for care. Yet as Fraser (1989) observed, needs are determined through a highly political process of needs interpretation. How can such a process occur fairly?

What constitutes good care will be variable in any society. Different people have different notions of what constitutes the best care for them. The challenge of care will be to offer options for care so that everyone can be well cared for in a manner that they find acceptable. Finally, the feminist impulse to revalue care grew out of a desire to end gender-based disparities in the provision of care. The greatest challenge to theorists of care is to see that, not only do gender-based forms of caring change to make the burdens and blessings of care equally distributed among men and women, but to do so in a way that does not reinscribe care as unequal burdens for the relatively disadvantaged.

See also: Caregiver Burden; Family and Gender; Fatherhood; Household Production; Masculinities and Femininities; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Social Welfare Policies and Gender; Time-use and Gender

Bibliography

Benner P, Wrubel J 1989 *The Primacy of Caring: Stress and Coping in Health and Illness*. Addison-Wesley, Menlo Park, CA

Bubeck D 1995 *Care, Justice and Gender*. Oxford University Press, Oxford, UK

Bussemaker J, van Kersbergen K 1994 Gender and welfare

states: Some theoretical reflections. In: Sainsbury D (ed.) *Gendering Welfare States*. Sage, London
Finch J, Groves D 1983 *A Labour of Love: Women, Work and Caring*. Routledge and Kegan Paul, London
Fisher B, Tronto J C 1990 Towards a feminist theory of care. In: Abel E, Nelson M (eds.) *Circles of Care*. State University of New York Press, Albany, NY
Folbre N 1994 *Who Pays for the Kids? Gender and the Structures of Constraint*. Routledge, New York
Foucault M 1965 *Madness and Civilization*. Pantheon Books, New York
Fraser N 1989 *Unruly Practices: Power, Discourse, and Gender in Contemporary Social Theory*. University of Minnesota Press, Minneapolis, MN
Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA
Held V 1993 *Feminist Morality: Transforming Culture, Society and Politics*. University of Chicago Press, Chicago
Jaggar A 1995 Caring as a feminist practice of moral reason. In: Held V (ed.) *Justice and Care: Essential Readings in Feminist Ethics*. Westview Press, Boulder, CO
Kittay E 1999 *Love's Labor*. Routledge, New York
Knijn T, Kremer M 1997 Gender and the caring dimension of welfare states toward inclusive citizenship. *Social Politics* **4**: 328–61
Noddings N 1984 *Caring: A Feminine Approach to Ethics and Moral Education*. University of California Press, Berkeley, CA
Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA
Robinson F 1999 *Globalizing Care*. Westview, Boulder, CO
Ruddick S 1990 *Maternal Thinking: Towards a Politics of Peace*. Beacon Press, Boston
Sevenhuijsen S L 1998 *Citizenship and the Ethics of Care*. Routledge, London
Tronto J C 1993 *Moral Boundaries*. Routledge, New York
Tronto J C 1996 The Political concept of care. In: Hirschmann N, Di Stefano C (eds.) *Revisioning the Political: Feminist Reconstructions of Traditional Concepts in Western Political Theory*. Westview Press, Boulder, CO
Ungerson C 1997 Social politics and the commodification of care. *Social Politics* **4**: 362–81
Waerness K 1984 The rationality of caring. *Economic and Industrial Democracy* **5**: 185–211

J. Tronto

Career Development, Psychology of

1. The Traditional Career Perspective

A career can be defined as a pattern of work experiences comprising the entire life span of a person and which is generally seen with regard to a number of phases or stages reflecting the transition from one stage of life to the next (= sequences of work roles). Therefore 'career' may be conceptualized with regard to the development of skills and expertise, the ability to learn, the developmental identity, and the self-concept as 'career anchor.'

The models of career development and the study of careers are always linked to the existing environmental conditions (cf. Whyte 1986 vs. Hall 1986), to a 'relationship concept' (= psychological contract) and to prevailing points of view (e.g. intra- or interorganizational) of the different sociological disciplines and the different fields of psychology (e.g., occupational, organizational, personnel, and managerial psychology). The orthodox perspective consists of the concentration on the big organization, associated with the 'bounded career' which develops within a single organization.

However, it needs to be pointed out that the situation in the field of career development research and in developing career theories must be assessed as extremely unsatisfactory, not least because it is a recent, interdisciplinary field within the field of social sciences. Several academic disciplines and their 'subfields' are working in it—a circumstance which has led to a 'fractionation' of the field which has therefore not been able to produce integrative theories of career development with the disciplines ignoring the results of the others (psychology, sociology, management science, anthropology, etc.).

The biggest part by far of empirical career research within the field of applied psychology (as also of the interdisciplinary studies) is dominated by the perspective of the 'intraorganizational career.' Here two issues play a central role: 'careers in organizations' and the 'matching' between the needs of the employee on the one hand and those of the organization on the other (cf. Holland 1997). This 'intrafirm perspective' has supported the so-called 'bounded career assumptions' for decades (as opposed to the 'boundary-less assumptions') having stressed for a major part the 'single organizational settings' (= research into career development within a single, stable organization) until today. Unfortunately, it can be observed in career research that this orthodox approach has been kept for a major part—in spite of significant changes in the organizations.

This traditional perspective of work and career, as it is also reflected by most of the classical 'models of career development' (cf. Brown and Brooks 1990, Osipow 1983, Super 1957), obviously comprises principles such as stability in the working environment (e.g., continued existence of the organization and the work role), movement by hierarchical promotion and interorganizational mobility, and a constant availability of positions and work roles in accordance with the interests, talents, and the lifestyle preferred by the person. It regards the organizations as a 'benevolent' unity playing an active role in designing the career of the individual employee or the executive, respectively. This traditional career perspective was accompanied by two institutions defining the background for the 'locus of career responsibility.' (a) The 'psychological contract,' the mutual relationships and expectations between the employee and the organization (=

unwritten contract of relationships): what an employee thinks he/she owes to the organization and what he/she thinks he/she can expect the organization to owe him for this. A psychological contract thus produces a long-term security (for both sides) and a high level of commitment and loyalty to the organization on the part of the employee. To the employee this had the advantage of having a valued job and a position which enabled him to climb up a hierarchical career ladder in scheduled steps. To the organization this model guaranteed a constant supply of talented employees with a high level of commitment, striving to take up new and enlarging roles—possibly even involving putting personal interests last. 'Downsizing,' 'head-count reductions,' and 'large-scale layoffs' have rendered this contract invalid. Employees can no longer count on a 'link' to their company going beyond the contractually agreed salary.

The second institution consists of the (b) assumption of big organizations which held a great fascination especially for psychological career research of the 1970s and 1980s and which was regarded as the basis for lifelong employment and as a model for a career context encompassing the entire working life (Ouchi 1981, Pascale and Athos 1981). Thus the 'locus of career responsibility' was determined as well. The organization was required to create stable social working conditions, and also to create an environment which offered ideal conditions to both the employees and the organization itself and which was also taking care of the personal development and the welfare of its employees. The focal point in career research was placed—even until the beginning of the 1990s—mainly on intraorganizational issues and limits itself to executive, management careers, or professional and hierarchical careers, respectively.

2. *The No-boundary Organization*

In the meantime, this traditional perspective of career development in organizations has dramatically changed, most of all with regard to the implicit work of the psychological 'relationships contract' involving a high level of commitment in the long term. A significant number of factors relating to changes in the design and the structure of organizations influence the nature of careers within organizations. Among other things, these include the changed work relationships and work contracts, the trend towards leaner and flatter organizations, and the increasing distribution of work to teams. The 'relational contract' became a 'transactional contract' designed for shorter terms, including a performance-based payment, a lower level of mutual commitment which can relatively easily be terminated by either side (MacNeill 1985). To the extent that organizations grow flatter, also the number of people required for leadership functions in the middle and upper management level will be reduced. However, most of all this significantly limits the traditional possibility for career development by 'climbing up the ladder.' New environmental conditions or conditions of the surrounding field with their high degree of uncertainty are forcing organizations to abandon the practice of guiding their employees and executives over long-standing career paths and through prescribed development sequences. Within the 'boundary-less' organization this responsibility is shifted towards the person who is thus itself responsible for its career development—via different employment relationships in different organizations (= boundary-less career). It is the opposite of an inter-organizational career (= bounded career), with several meanings to this term: (a) this career surpasses the 'bounds' of different organizations; (b) there are no (classically) hierarchical principles in career development; (c) the marketability and network relations are predominant. Hall (1976) has coined the term 'protean career' for this new scenario. By this we understand a proactive process which is controlled by the person—and not by the organization. It consists of all the different experiences of a person in the fields of training and the work experience in different organizations, the change between different groups of training occupations or job roles, etc. to different employment relationships in different organizations.

Besides, it can also be observed that the growing trend towards teams, leads towards assigning management functions to these teams (e.g., in the medium managementb level) as their members are increasingly acquiring the ability of self-management. So-called 'high involvement work teams' coordinate, schedule, and distribute the work without depending on firmly established 'supervisor positions' (see *Teamwork and Team Training*). A stronger orientation towards teams in organizations should result in a less clear definition of individual tasks or work roles which are not inseparably linked to certain persons. The team members are expected to fulfill different tasks at different moments and, if required, to share their expertise with others. The ideal member of a team should have the highest possible number of different abilities and should be able to work without direct leadership. Therefore in such a team there is not much room for people who only want to give instructions or for those who are only able to carry out one single task or who want to specialize on certain tasks. This circumstance, too, will result in a more difficult career planning in the sense of a linear career path. Finally, the changes in work relations and in the 'psychological contract' have reduced the stability of the traditional career, as well as the attitudes, the loyalty, and the expectations of the employee or the executive. The employee can no longer count on a long-term commitment on the part of the employing organization, and neither can he or she necessarily count on the fact that his or her abilities will be assessed and valued on the job market.

Changes in career development require us to look at the changes in the organizational, professional, and industrial or economic context at the same time. The evolution of the forms of organizations has always been the impulse for the content and the way of professional careers. That is why career theories always reflect the current assumptions on organizations (Arthur et al. 1989). Organizational structures dictate the necessary core competencies and different organizational practices require a different 'mix' of competencies. Therefore changed 'boundary-less' organizations will employ employees with changed values (e.g., a limited commitment) and with a 'boundary-less' career oriented more towards the self. 'Boundary-less' organizations change existing career patterns. Here some 'critical dimensions' can be identified, such as the 'size' and the 'degree of centralization' determining the evolution of organizations and having various effects on the developing career concept and career systems.

3. *The Reorganization of Organizations*

Large-scale companies and organizations have rapidly lost to an increasing extent—at least since the end of the 1980s—their dominance, the importance of their expansion, and their fascination. Instead of vertical coordination of dominant companies, other models have been found which are more efficient and competitive at being able to cope with the fast, complex, global developments of the information and service society more effectively. Big organizations of the future will consist of multiple 'divisions' and 'joint ventures,' 'regional alliances,' private and public partnerships. In addition to this, entrepreneurs, franchisers, and smaller businessmen will supply supporting services, technologies and materials, and distribute goods. These 'boundary-less organizations' continuously change their dynamic forms and structures by expanding and reorganizing, acquiring new parts or selling old ones and even entering into partnerships with other organizations or companies for certain periods of time. During this process the old ideas of vertical coordination will increasingly vanish, just because horizontal coordination models as an alternative are better able to adapt to the specific and constantly changing interests of every company.

This revolution in the reorganization of organizations which was triggered by the global competition has given a new definition to both the work relationships and the forms of career development and newly challenges the classical career theory to paths not yet taken. In order to cope with a highly turbulent external environment which can be predicted only to a limited degree and in the short run, organizations have tried to adapt to the changed circumstances by 'reorganizing,' 'downsizing,' 'delayering,' 'flattening the pyramid,' 'teaming,' and 'outsourcing.'

However, the efforts of the organizations concerning the work relationships between employees and executives on the one hand and the companies on the other have also caused a high degree of destabilization. Additionally, a higher degree of diversification among both employees and executives can be observed (the percentage of women is constantly increasing and—most of all among executives—an increasing internationalization can be recognized). But also the attitudes and the behavior of the individual employee and executive towards the work situation has significantly changed, most of all in the fields of commitment, loyalty, and work and family.

4. *New Career Competencies*

Instead of 'bemoaning' these changes and to wish back work relationships and career developments from the past, it is necessary to point out chances for future development and to bring about clarity. This can best be done by newly arranging work roles, setting competencies required for the future, clarifying different career concepts and their conceptual meaning (e.g., whether it is firm-centered or person-centered), and by newly defining career success as well as career responsibility.

Usual work roles are often described as artifacts of the industrial age, as 'vehicles' which were used in order to split up work in 'packages.' By continuous reorganization, downsizing, teaming, etc., many of the old work roles or 'jobs' have been significantly changed—or totally deleted—by 'job enrichment' or 'job enlargement' measures. The changes in the working world also require new career competencies and strategies: a 'portfolio of skills' which can be translated from one work situation to others, with the focus on 'interorganizational employability'; lifelong learning in order to maintain the skills 'relevantly'; experience in 'team and project work' (so-called 'collaborative abilities'); building reputation over one's career; and, not least, the career should be consistent with the self-identity. In order to be successful in a turbulent working world, in the future more than mere 'job skills' will be required. New 'career competencies' will be necessary, so-called 'metaskills.' They are required in order to acquire new skills. Here are some examples: adaptability, tolerance of ambiguity, and uncertainty (Handy 1990).

Changes in the required core competencies can best be shown in the field of leadership and management, at the point of transition from the 'traditional organization' to the current 'network organization.' Careers in the traditional organization often used to comprise only a single technical expertise (e.g., production or sales)—but only in a few cases the understanding of several business functions. The more complicated the organizations became, the more competencies—in different compositions—were required (e.g., technical,

commercial, and administrative competencies). In today's network organizations, in which independent companies are linked to each other, the objective is to provide critical expertise required for a specific project or product. Network organizations gain their particular influence by relying on their internal and external partners who, on their part, contribute to the value added chain. By doing this products and projects (from RandD to sale) can be tackled which could not be handled by a single organization on its own but which can only be handled with the help of network partners.

In order to do this, executives need 'collaborative' competencies. These consist of three components: (a) the ability to analyze a problem and develop solutions with the help of partners and networks; (b) 'partner abilities,' i.e., to develop concepts, to negotiate, and to carry them out to mutual benefit; (c) relationship management, i.e., to identify the needs of clients and partners and to meet them precisely. These career competencies, depending on the organizational structures, can finally be continued to the currently developing 'cellular organization' which no longer includes any kind of management hierarchy. The 'cells' of these organizational forms consist of, among other elements, 'self-managing teams' with their own management responsibilities which would be able to survive autonomously. However, they share their knowledge and their information with other 'cells' in order to be stronger and more competent. This organizational form does not work as 'employer' but it is a helpful mechanism in order to promote the knowledge-intensive abilities of their members in their application and enlargement.

5. *New Career Concepts*

So this leads us back to career responsibility as in the cellular organization it is completely placed with its members (= the locus of responsibility). The old theories for career management regard career development in the sense of narrow, sequenced, and methodical activities. Today's working world, on the other hand, is characterized by continuous changes (for both sides), by uncertainty, turbulence, limited obligations, and a lack of borders. Therefore the traditional 'organizational career,' in which the employees were expected to put their personal interests last, was replaced by a so-called 'self-centered career.' So the career model now changes into an individual process of self-responsibility. The individual explores his or her career opportunities, sets career goals, develops strategies, and searches for a relevant definition from time to time (Greenhaus et al. 1995).

In order to do this, Schein (1990) proposes a differentiation of the term 'career' into an 'internal' and an 'external' career—with the internal career relating to the subjective sense for where, i.e., in which direction, the working life will lead. The 'external' career relates to the formal phases and roles as they are defined by the organization's practices and the concepts of society regarding what an individual can expect within the professional organization. Closely related to the term of the 'internal' career is the concept of the 'career anchors' which can be equated with the self-concept of a person. This develops during the course of gaining increasing experience of life and work. It consists of talents and abilities which a person recognizes for himself or herself, of a structure of values, and of job-relevant motives and needs. The self-concept constructed in this way has a stabilizing influence on the actions and decisions of a person and becomes part of the personality—with these career anchors not being cemented forever but able to change (among other things by a change of values in the society). We differentiate between several categories of such career anchors which, according to the occupational group, can be part of the individual self-concept in order to express the respectively existing central motives, values, and needs (cf. also the works by Hogan et al. 2000). The essential categories of career anchors are: (a) autonomy/independence; (b) security/stability; (c) technical–functional competency; (d) general management competency; (e) entrepreneurial creativity; (f) lifestyle; (g) service in the sense of a service to a higher task; (h) challenge. Unlike the construction of 'career resilience,' the concept of the 'career anchor' implies that the definitions by different persons in organizations of their career regarding developing talents, motives, and values vary significantly.

Careers are currently changing in several ways. Today we cannot, or much less than in the past, assume that people will spend all their life in one job or in one special field. That is why we have to find answers to, among other things, the following questions: how can we find a transition from technical to management roles? In order to achieve an effective performance lifelong learning is required and the acquisition of ever new abilities. Neither can we assume that people will always stay employed by one company. These changes create an environment in which employees and executives regularly have to evaluate their potentials and career plans again and again in order to keep themselves 'marketable'—in connection with the opportunity to appear professionally 'attractive' to a company. Careers will develop laterally and diagonally. However, loyalty and commitment from the part of the organizations will decrease which will shift the center of responsibility and control of the career development (as mentioned before) towards the employee. The working individual is to develop a higher degree of 'resilience' during this transitional stage of shifting responsibility and control in order to become 'career self-resilient' in the working world. This construction of 'career resilience' measures the ability of an individual to change his or her

career, to cope with uncertainty and with disappointments in organizational processes. This concept can be applied to all individuals and all careers.

6. Challenges for the Field of Applied Psychology

In these enormous processes of change in the field of career development the above-mentioned fields of applied psychology are given a number of essential tasks. However, here it is indispensable that the individual 'subfields' do not neglect the results of the others (neither those of other social sciences)—as has been done up to now. These tasks include, for one thing, the development of assessment methods and development tools in order to identify individual abilities, potentials, and talents, to determine the state of development and to provide creative approaches for 'skill-building.' This task gets all the more difficult if at the moment it is not yet known how the new requirements and roles will look like in 5–7 years. The same applies to development plans, job descriptions, and incentive systems. However, they are also included in the field of 'multi-rater feedback' tools such as the 360° feedback. A special importance is attached to 'cross-training' in order to create opportunities to acquire new abilities and skills or to enlarge existing ones. The activities in the field of 'human resources planning' must be significantly enlarged. One focal point here should be put on career planning in order to develop career paths and 'mentoring programs' (also over different organizational levels).

The development of training courses for 'boundary-less career management' or the problematic of 'career adaptability' must not be forgotten. The latter must be intensively studied by psychologists. For example, is 'career adaptability' a function of the personality? Also 'coping strategies' must be developed. Mentors must be enabled to help employees and executives to always steer their career safely in an ever-changing working world. Psychologists are able to help working people in their career development by replacing their (conservative) expectations of a continuous mobility 'upwards' by a rather cyclical and lateral career development. In 'boundary-less organizations' and 'boundary careers,' major role conflicts caused by complicated working conditions must be expected, and a substitute must be found for identifying with the organization.

With all of these psychological career activities the reliable determination of a specific need for development (especially the future need in growth areas), advisory activities, and the conduct of evaluation programs is of major importance. Here a 'dual focus,' i.e., individual and organization, is an essential requirement for successful work on the part of the psychologist. Career development programs offer their services to the employee for his or her career planning and give incentives for learning. The organization benefits from this program by developing in-house talents, by supporting the planning for succession, and by countering discontent and decreasing commitment caused by uncertainty and instability of new work relations.

7. Summary and Implications for Future Research

Global competition has triggered a revolution in the reorganization of the organization. In contrast to the traditional form of the 'bounded' organization with its centralized decision making processes, vertical models of coordination, hierarchies, and career ladders, the present organization emphasizes a boundary-less concept, a vertically coordinated approach, and boundary-less career principles. New forms of organizations have dynamic shapes and structures. They rearrange themselves to adapt to changes in the environment. It was outlined how evolution of the forms of organizations has been the impulse for content and way of professional careers. In this scenario several trends were described such as the shift from bounded to boundary-less concepts; stability to instability of employment; long-term to short-term career goals; generic to firm-specific competencies; firm-centered to person-centered career approaches; relational to transactional employer–employee relationships; and the shift of career responsibility from the employer to the employee. In order to describe the evolving career management focus, a number of different concepts and terms created by researchers over the past several years have been presented: career anchors; psychological contract; protean career; internal vs. external career; self-centered career; destabilization of relationships between people and organizations; interorganizational concepts; networking; career self-resilience; interorganizational employability; high involvement work teams; collaborative competencies.

All of this raises some important questions for future research. For instance, to what extent will boundary-less organizations alter career theories and actual career mobility patterns? To what extent is career adaptability a function of personality (disposition toward proactive behavior) or age vs. a skill that can be developed? Can people learn to be adaptable? There is need for more research on what personal and developmental growth training best prepare people to engage in new career cycles and disengage from old ones? How can organizations help employees to regularly assess their skills, interests, and values so that they can figure out what kind of work experience to seek? As information technology makes new organizational forms possible, and as social values shift priorities, what should a given job consist of and how should one hire and train people for the ambiguous and changing roles? It is necessary to consider whether a career goal is helpful or an unnecessary

restriction. Research is needed on new standards of career success, on new forms of work identity, on role overload (the work vs. family conflict), and new substitutes for organizational identification have to be found. It is also worth studying what support systems are needed for employees under the transactional contract and how the development of a 'spot market' mentality can be avoided? What is the impact of choice in the different employment contracts (relational vs. transactional)? A major implication of these ideas for future research is that the study of careers must be better connected to a turbulent, complex, rapidly changing environment, and it has to become multidisciplinary.

Bibliography

Arthur M B, Hall D T, Lawrence B S (eds.) 1989 *Handbook of Career Theory*. Cambridge University Press, New York

Brown D, Brooks L (eds.) 1990 *Career Choice and Development*. Jossey-Bass, San Francisco

Greenhaus J H, Callanan G A, Kaplan E 1995 The role of goal setting in career management. *The International Journal of Career Management* **7**(5): 3–12

Hall D T 1976 *Careers in Organizations*. Scott, Foresman, Glenview, IL

Hall D T 1996 *The Career is Dead: Long Live the Career*, 1st edn. Jossey-Bass, San Francisco

Hall D T (ed.) 1986 *Career Developments in Organizations*. Jossey-Bass, San Francisco

Handy C 1990 *The age of Unreasoning*. Harvard University Press, Boston

Hogan J, Hogan R, Weinert A B 2000 *The Values and Interests Inventory*. University of the Federal Armed Forces, Hamburg, Germany

Holland J L 1997 *Making Vocational Choices*, 3rd edn. Psychological Assessment Resources, Odessa, FL

MacNeill I R 1985 Relational contracts: what we do and do not know. *Wisconsin Law Review*. **3**: 483–525

Osipow S H 1983 *Theories of Career Development*, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ

Ouchi W G 1981 *Theory Z*. Addison-Wesley, Reading, MA

Pascale R T, Athos A G 1981 *The Art of Japanese Management*. Simon and Schuster, New York

Schein E H 1990 *Career Anchors*. University Associates, San Diego, CA

Super D E 1957 *The Psychology of Careers*, 1st edn. Harper and Row, New York

Whyte W F 1986 *The Organization Man*. Simon and Schuster, New York

A. B. Weinert

Caregiver Burden

1. Definition of Family Caregiving

The provision of assistance and support by one family member or friend to another is a pervasive aspect of everyday human interactions. Providing help to a family member with chronic illness or disability is not very different from the tasks and activities that characterize interactions among families and close friends without the presence of illness or disability. Thus, when a wife provides care to her husband with Alzheimer's disease (AD) by preparing his meals, it may be an activity she would normally do for an unimpaired husband. However, if a wife also assists her cognitively impaired husband with bathing and dressing, few would question whether or not caregiving is taking place. The difference is that providing assistance with bathing and dressing or assisting with complex medical routines clearly represents 'extraordinary' care and exceeds the bounds of what is 'normative' or 'usual.' Similarly, parents caring for a child with a chronic illness may need to assist with daily medical routines (e.g., insulin injections or chest physical therapy) that are time consuming and difficult and are in addition to normal parenting responsibilities. Caregiving involves significant expenditure of time and energy often for months or years, requiring the performance of tasks that may be physically demanding and unpleasant, and frequently disrupting other family and social roles of the caregiver (see *Gender Role Stress and Health*).

Although caregivers may perform tasks similar to those carried out by paid health professionals, they perform these services for no compensation and do so either voluntarily or because they feel there are no other alternatives. Because the physical and mental health consequences of taking on this role are sometimes severe, and because caregivers represent an invaluable resource to the well-being of our population, research on caregiving has become a high priority among scholars in many disciplines as well as among policy makers.

2. Prevalence of Caregiving

Although the definition and boundaries of what is meant by the term caregiving often vary depending on the purpose for which such definitions are used, there is strong consensus that regardless of how caregiving is defined, its prevalence is high. A broadly inclusive approach might argue that a caregiver is needed for every person with health-related mobility and self-care limitations which makes it difficult to take care of personal needs, such as dressing, bathing, and moving around the home. Current estimates indicate that 4 percent of the non-institutionalized US population under the age of 55 meet these criteria. Beyond the age of 55, the proportion of persons with mobility and/or self care limitations increases dramatically; fully half of the population falls into this category after age 85 (US Bureau of the Census 1990). If we assume that these individuals minimally require one caregiver, these estimates yield over 15 million caregivers in the US. Indeed, these estimates are somewhat lower than results reported in a recent national survey of caregivers which reported that there were 22.4 million

households that met broad criteria for the presence of a caregiver in the past 12 months (National Alliance for Caregiving 1997).

Caregiving is not just a late life phenomenon involving the care of disabled older persons. It is estimated that 10–14 percent of children and adolescents (7.5 million) in the US have some type of chronic illness or disability; of these individuals, approximately 20–25 percent (1.5 million) have serious health conditions that impair daily functioning and thus require a caregiver. Additionally, 4.1 million individuals between the ages of 21 and 64 require personal assistance in ADLs/IADLs.

3. Who Provides Care and What Type of Care is Provided?

Caregivers to elderly individuals generally are differentiated by age and relationship to the care recipient. One large group consists of adult children, usually daughters or daughters-in-law, in their 50s and 60s; the second group of caregivers comprises spouses of care-recipients and is generally older and has a higher proportion of male caregivers than adult children caregivers (see *Gender Role Stress and Health*).

The roles and functions of family caregivers vary by type and stage of illness and include both direct and indirect activities. Direct activities can include provision of personal care assistance, such as helping with bathing, grooming, dressing, or toileting; healthcare assistance such as catheter care, giving injections, or monitoring medications; and checking and monitoring tasks, such as continuous or periodic supervision, and telephone monitoring. Indirect tasks include care management such as locating services, coordinating service use, monitoring services, or advocacy, and households tasks, such as cooking, cleaning, shopping, money management, and transportation of the family member to medical appointments or day care programs (Biegel and Schulz 1999). The intensity at which some or all of these caregiving activities are performed varies widely, with some caregivers having only limited types of involvement for a few hours per week while other caregivers might provide more than 40 hours a week of care and be on call 24 hours per day (see *Social Support and Stress*).

4. Conceptual Approaches to the Study of Caregiving

The dominant conceptual model for caregiving assumes that the onset and progression of chronic illness and physical disability is stressful for both patient and caregiver and, as such, can be studied within the framework of traditional stress-coping models. Indeed, some researchers have likened caregiving to being exposed to a severe, long-term, chronic stressor (Pearlin et al. 1995). Within this framework, objective stressors include measures of patient disability, cognitive impairment, and problem behaviors, as well as the type and intensity of caregiving provided. Key outcome variables for the caregiver include psychological distress and burden, often referred to as caregiver burden, psychological and physical morbidity, and patient outcomes such as institutionalization and death.

Although the literature consistently shows a moderate relationship between level of patient disability and psychological distress of the caregiver, there is considerable variability in caregiver outcomes which is thought to be mediated and/or moderated by a variety of factors including economic and social support resources available to the caregiver (Haley et al. 1996), a host of individual difference factors, such as gender, personality attributes (optimism, self-esteem, self-mastery), coping strategies used, and the quality of the relationship between caregiver and care recipient (see Quittner et al. 1990, Schulz et al. 1990, 1995). Researchers have further extended basic stress-coping models to include examination of secondary stressors, such as role conflict engendered by caregiving demands (Pearlin et al. 1997), and have applied many additional theoretical perspectives borrowed from social and clinical psychology, sociology, and the health and biological sciences to help understand specific aspects of the caregiving situation. Finally, researchers interested in the health consequences of caregiving have focused on a variety of physiological mechanisms including the pituitary-adrenal axis, the sympathetic nervous system, and the immune system in their effort to identify biological modulators (Haidt and Rodin 1995, Kiecolt-Glaser et al. 1991, Vitaliano et al. 1997) of the stress-health relationship. Researchers have also shown that the stresses associated with caregiving are a risk factor for the caregiver's death (Schulz and Beach 1999). In sum, caregiving clearly provides a rich platform for the application of much of the theoretical and methodological expertise of researchers in many disciplines.

A wide range of caregiving effects have been described in the literature including disruption of family routines, psychological distress, and psychological and physical morbidity including mortality, financial hardship, and work-related problems (see *Stress and Health Research*). Feeling burdened or distressed by the demands of caregiving is the most frequently reported outcome associated with caregiving, although this is not a universal phenomenon, particularly among spousal caregivers (Zarit et al. 1986). Psychiatric morbidity such as depression and anxiety are also common. Physical health effects such as increased susceptibility to illness have been more difficult to demonstrate, although they are likely to occur in high demand situations among vulnerable (e.g., frail) caregivers. Possible mediators of illness effects are increased depression associated with caregiving and changes in health related behaviors such as

sleeping and eating patterns and medical compliance. Positive effects of caregiving such as increased self-esteem, the satisfaction of knowing that one's relative is being properly cared for, as well as improved mental health have also been reported (Beach et al. 2000).

The demands and negative impacts of dementia caregiving are generally higher than nondementia caregiving (Schulz 2000). Indeed, a recent report by Ory et al. (1999) documents the ways in which dementia care is different from other types of family caregiving. Not only do dementia caregivers spend significantly more hours per week providing care than nondementia caregivers, they also report greater employment complications, caregiver strain, mental and physical health problems, reduced time for leisure and other family members, and family conflict. Factors that are likely to account for this greater level of strain include having to contend with behavioral problems of the care-recipient (e.g., wandering, aggressiveness), and the unpredictable nature and course of dementing illnesses. A number of health psychologists have focused this type of caregiving as a platform for studying mind-body phenomenon linking chronic stress to physical morbidity.

5. *The Future of Caregiving*

A number of important interrelated demographic, health, and social trends will shape the caregiving agenda in the future. First, there will be a worldwide increase in the number of older individuals, and possibly increased numbers of disabled individuals with longer life expectancies due to medical interventions. A key question will be the extent to which increases in life expectancy are associated with increasing years of disability. Some early evidence suggests that future cohorts of elderly individuals will be healthier and more functional than current cohorts, thus adding years to life may not necessarily increase the need for caregiving assistance. Alternatively, one can speculate that some types of medical interventions (e.g., drug therapies that enable AD patients to spend more years at home) will extend the family caregiving career. Second, future cohorts of elderly will have smaller families and fewer children available to provide care. The supply of caregivers may be further depleted by the increased and sustained labor-force participation of adult daughters, making them less available to provide care. Third, the trend toward shifting care from formal to informal care providers is likely to continue and accelerate. Caregivers increasingly will be asked to provide complex postacute care in addition to the chronic care they have provided traditionally. As these examples illustrate, caregiving issues are linked inextricably to broader demographic trends and the health and disability of our older population. At the macrolevel, caregiving and caregiver burden will have to be addressed through government policy and private sector programs.

See also: Aging and Health in Old Age; Gender Role Stress and Health; Social Support and Stress; Stress and Coping Theories; Stress and Health Research

Bibliography

Beach S R, Schulz R, Yee J L, Jackson S 2000 Negative and positive health effects of caring for a disabled spouse: Longitudinal findings from the Caregiver Health Effects Study. *Psychology and Aging* **15**: 259–71

Biegel D E, Schulz R 1999 Caregiving and caregiver interventions in aging and mental illness. *Family Relations* **48**: 345–54

Haidt J, Rodin J 1995 *Control and Efficacy: An Integrative Review*. A report to the John D. and Catherine T. MacArthur Foundation Program on Mental Health and Human Development

Haley W E, Roth D L, Coleton M I, Ford G R, West C A C, Collins R P, Isobe T L 1996 Appraisal, coping, and social support as mediators of wellbeing in black and white family caregivers of patients with Alzheimer's disease. *Journal of Consulting and Clinical Psychology* **64**: 121–9

Kiecolt-Glaser J K, Dura J R, Speicher C E, Trask O J, Glaser R 1991 Spousal caregivers of dementia victims: Longitudinal changes in immunity and health. *Psychosomatic Medicine* **53**: 345–62

National Alliance for Caregiving and the American Association of Retired Persons 1997 *Family Caregiving in the US Findings from a National Survey. Final Report*. National Alliance for Caregiving, Bethesda, MD

Ory M G, Hoffman III R R, Yee J L, Tennstedt S, Schulz R 1999 Prevalence and impact of caregiving: A detailed comparison between dementia and non-dementia caregivers. Dementia and non-dementia caregiving. *The Gerontologist* **39**: 177–85

Pearlin L I, Aneshensel C S, Le Blanc A J 1997 The forms and mechanisms of stress proliferation: The case of AIDS caregivers. *Journal of Health and Social Behavior* **38**: 223–36

Pearlin L I, Aneshensel C S, Mullan J T, Whitlatch C J 1995 Caregiving and its social support. In: Binstock R H, George L K (eds.) *Handbook of Aging and the Social Sciences*, 4th edn. Academic Press, New York, pp. 283–302

Quittner A L, Glueckauf R L, Jackson D N 1990 Chronic parenting stress: Moderating vs. mediating effects of social support. *Journal of Personality and Social Psychology* **59**: 1266–78

Schulz R (ed.) 2000 *Handbook of Dementia Caregiving*. Springer, New York

Schulz R, Beach S 1999 Caregiving as a risk factor for mortality. The caregiver health effects study. *Journal of the American Medical Association* **282**: 2215–9

Schulz R, O'Brien A T, Bookwala J, Fleissner K 1995 Psychiatric and physical morbidity effects of Alzheimer's Disease caregiving: Prevalence, correlates, and causes. *The Gerontologist* **35**: 771–91

Schulz R, Visintainer P, Williamson G M 1990 Psychiatric and physical morbidity effect of caregiving. *Journals of Gerontology: Psychological Sciences* **45**: P181–P191

US Bureau of the Census 1990 *The Need for Personal Assistance with Everyday Activities: Recipients and Caregivers. Current Population Reports* (*Series P-70. Household Economic Studies*). Department of Commerce, Washington, DC

Vitaliano P P, Schulz R, Kiecolt-Glaser J, Grant I 1997 Research on physiological and physical concomitants of caregiving:

Where do we go from here? *Annals of Behavioral Medicine* **19**: 117–23
Zarit S H, Todd P A, Zarit J M 1986 Subjective burden of husbands and wives of caregivers: A longitudinal study. *The Gerontologist* **26**: 260–6

R. Schulz

Caregiving in Old Age

1. *Caregiving, the Area*

Caregiving in old age is the provision of assistance to an elder when his or her health deteriorates, whether it is physical or mental health, or a combination of the two. Caregiving typically refers to unpaid care from members of the informal network, that is, from family or friends. It is an aspect of the more general concept of social support. However, there is no precise scientific definition of the term caregiving, and research reveals a diversity of views. For example, spouses who are providing assistance to their loved one with instrumental activities of daily living are less likely than others, such as children or siblings, to consider themselves as caregivers (more frequently referred to as carers in the UK). Think of the instance where a daughter takes her mother grocery shopping. Either one or both may define it as a chance to socialize, or as assistance, or as simply something they do together. The literature adds to the confusion with the use of a variety of terms, including caregiving, caring, assistance, interaction, and support, sometimes used synonymously and sometimes not. Despite these difficulties, the area of caregiving in old age has received much research attention from 1970 onwards.

2. *In the Beginning*

Interest in the area arose during the 1960s and 1970s as a practical concern. During this time, gerontologists documented the extensiveness of social ties, including caregiving, during later life. This was important within the context of the times when it was commonly believed that seniors in Western industrialized societies were largely isolated from their family, living alone, and often housed in long-term care institutions. Gerontological research recorded the falsity of these images by studying the strong interactional ties within the lives of most seniors. Researchers reported the preferred normality of 'intimacy at a distance,' in which neither seniors nor their children wish to live together but have a desire for contact. Usually seniors lived geographically proximate to at least one of their children. This period also examined the prevalence of informal, unpaid caregivers as the dominant source of assistance for elders. The informal network was the first resort for care when health declined. This was important because it established the strength of the family as a major source of interpersonal support and care during old age (Chappell 1990).

3. *Reaching Maturity*

By the late 1970s and especially in the 1980s, research on caregiving was burgeoning. It examined earlier assumptions of the research. For example, early research assumed contact meant assistance or contact included positive support. Research started to investigate whether and under what circumstances social interaction was indeed supportive, when assistance was positive. Research distinguished who provides support and studied the critical role of spouses among married couples and children when a spouse is not available frequently because of death. While caregiving frequently focuses on tasks of activities of daily living, either instrumental activities of daily living (IADL) such as shopping and banking, and basic activities of daily living (ADL) such as going to the toilet, eating, personal mobility, and other activities necessary for survival, the emotional aspects received little attention. Yet it was recognized that it is this emotional element that distinguishes informal care from formal or paid caregiving.

Caregivers were labeled the 'hidden victims,' 'sandwich generation,' 'generation-in-the-middle' (Brody 1981), raising the public recognition of caregiving. It was during this period that Cantor's (1979) hierarchical compensatory model, also known as the substitution model, received much attention. She argued for an orderly hierarchical selection of caregivers, determined by the primacy of the relationship between the giver and the recipient. According to this view, the most preferred caregiver was the spouse, followed by daughters, sons, other relatives, friends, and neighbors, in that order. Litwak's (1985) competing hypothesis of task specificity also gained popularity. In this instance, it was argued that persons differentially placed within society provide different types of assistance (spouses can provide emotional and other long-term needs on a continuous basis; neighbors provide short-term, sporadic, and instrumental assistance).

There was less of a focus on support from siblings, friendships, and grandchildren; most studies examined care from the spouse and children, who are the most prevalent care providers. Significant gender differences in caregiving were revealed: the fact that women tend to do the emotional and hands on work while men tend to provide advice and financial assistance (Horowitz 1981). Other studies reported that men are more likely to rely exclusively on their spouse for emotional support whereas women are more likely to rely on friends (Hess and Waring 1980); that caregivers experience burden as a result of their involvement in this role (Zarit et al. 1980); that there is generally one main care provider who does most of the work (Stone

et al. 1987); and that working daughters do not provide fewer hours of caregiving than those who are not in paid labor (Brody et al. 1984).

Still other research indicated that some ethnic groups have different and more extensive caregiving than others, but disentangling how much of the variation is due to 'culture' is difficult because ethnic minority status and social class are correlated empirically. It was also during this time that the full extent of caregiving was exposed. In a review of scientific studies throughout industrialized worlds, Kane and Kane (1985) estimated that between 75 and 85 percent of all personal care to seniors comes from the informal network. Chappell (1985) reported that almost all community living elders receiving any type of assistance do so from the informal network. The lack of recognition of the care provided by caregivers and the lack of support for caregivers within formal health care systems was also documented (George 1988).

The sheer volume of research on caregiving established it as a major area and much was learned. Informal caregiving emerged as the indisputable dominant system of care in post-modern society—despite all that has been said about our individualism and lack of concern for one another. Caregiving, it became obvious, is significantly a woman's issue. Women predominate as the caregivers and the care-receivers. Caregivers are burdened and women who work do not shirk from caregiving.

4. *Growth in the 1990s*

By the late 1980s and early 1990s caregiving had become a popular area of gerontological research, producing studies on a multitude of facets of this topic. Studies on burden and stress continued, revealing either no differences in levels between male and female caregivers or, when differences were reported, that female caregivers experience more stress. Research on male caregivers showed their more instrumental approach to the role than women have. Positive aspects of caregiving such as feeling useful and bringing comfort to a loved one were being studied. Caregivers, furthermore, are reluctant users of formal services. However, the last decade of the twentieth century did not produce simply more of the same. It also saw another substantive shift in this area of inquiry. Caregiving became politicized.

Before this time, it was recognized within academic and practice circles as important, but it was not recognized at the political level. This changed for two major reasons. One was heightened awareness from a feminist perspective that aging and caregiving, as a woman's issue, had been more or less invisible as part of the private domain and not part of a public debate. Public policy had been operating by assuming that women's caregiver roles would simply continue (Hooyman 1990). A second major factor was the prolonged economic recession throughout industrialized countries. This was related to a perceived crisis in health care system funding and paved the way for a new political rhetoric, recognizing, for the first time in the twentieth century, caregivers and community care. Indeed, family care emerged in the 1990s as a cornerstone of the new vision for health reform throughout the industrialized world.

This new-found political awareness of caregivers brought an urgency to research in this area, highlighting the need for an adequate understanding of caregiving if policies and programs were going to make assumptions about their capacities and their needs. It could not be assumed that the family is necessarily the most appropriate place for caregiving. The area of elder abuse provided an example. Programs that insist seniors stay with families could prolong abusive situations.

The current political interest in caregiving represents a rediscovery of caregiving. Although government had not previously acknowledged the role played by caregivers, they had never replaced private arrangements. In the past, governments intervened only when families and individuals were not coping and came to their attention. At the present time, governments' embrace of caregivers can be viewed skeptically as a means to cost-shift from the public purse to caregivers, largely women. A concern with increased burden for caregivers is prominent. New questions are being addressed. What are the assumptions about family caregivers in current policies? What is the impact of different service interventions on caregivers? How can the formal system support caregivers and how can caregivers be integrated with formal health care delivery? The answers to these questions promise exciting new knowledge to assist seniors and their families.

In addition, increasingly sophisticated analyses are being conducted and testing of earlier explanations continues. For example, Penning (1990) demonstrates a lack of empirical support for the concepts of the sandwich generation and generation-in-the-middle. Her research supports serial caregiving as a more apt descriptor of this phenomenon, since individuals are usually involved in raising their children, then caring for their parents, then caring for their husband rather than being engaged in all of these roles at one time. In addition to the characteristics of caregivers and the burdens of caring, researchers are beginning to study the meaning of caring. Wenger et al. (1996) suggest conceptualizating caregiving in terms of purposes and relationships rather than tasks, taking the everyday experiences of caregivers and care recipients into account. For example, one can examine a preparatory stage of anticipating care; being involved in having to provide preventive care, such as ensuring the person eats well and exercises; and being involved in the supervision of care.

New theoretical insights into the socioemotional context of relationships has direct relevance for care-

giving. Lang et al. (1998) for example, finds increasingly discriminating choices in our social interactions as we age. Baltes (1996) studies etiologies (causes) of behavioral dependency such as learned helplessness, learned dependency and selective optimization with compensation and the prominent role the social environment plays in this regard. Her research directs attention to overdependency among care recipients, an under-researched area to date.

5. *Conclusions*

Interest in caregiving in old age arose and has continued from a pragmatic applied interest. It began with assumptions about supportive relationships and about the role of the family, and over time has focused on the complexity of both the definition of caregiving as well as its contextual fields. Only recently, has there been conceptual development in this area with current attempts to examine the meaning of the term 'caregiving' and how that varies from group to group. At the present time, complex conceptual issues are starting to be addressed—When is interaction caregiving? Whose definition counts? How are caregivers taken into account within the health care system? How can they be taken into account? Whose definitions of caregiving are appropriate for service delivery? How can the autonomy of caregivers be maintained? The conceptual issues relate to methodological issues. If the mother does not consider the daughter's efforts to be caregiving but the daughter does, whose definition does the researcher accept, if either? If the wife is cooking the meals anyway and always has, is this part of caregiving when her husband's health declines? When computing the economic value of caregiving, do we include those times and tasks governments would not provide?

Future directions for caregiving research in the short term seem more or less clear. Renewed interest in the area has been fueled by the recognition of caregivers at the political level in the new vision of health reform.

See also: Aging, Theories of; Care and Gender; Caregiver Burden; Ecology of Aging; Health Care Delivery Services; Health Care Markets: Theory and Practice; Health Care, Rationing of; Health Care Systems: Comparative; Lifespan Development: Evolutionary Perspectives; Lifespan Theories of Cognitive Development; Population Aging: Economic and Social Consequences; Psychoneuroimmunology

Bibliography

Baltes M M 1996 *Many Faces of Dependency in Old Age*. Cambridge University Press, Cambridge, UK

Brody E M 1981 Women in the middle and family help to older people. *Gerontologist* **21**: 471–80

Brody E M, Johnsen P T, Fulcomer M C 1984 What should adult children do for elderly parents? Opinions and preferences of three generations of women. *Journal of Gerontology* **39**: 736–46

Cantor M H 1979 Neighbors and friends: an overlooked resource in the informal support system. *Research on Aging* **1**: 434–63

Chappell N L 1985 Social support and the receipt of home care services. *Gerontologist* **25**: 47–54

Chappell N L 1990 Aging and social care. In: Binstock R H, George L K (eds.) *Handbook of Aging and the Social Sciences*, 3rd edn. Academic Press, San Diego, CA, pp. 438–54

George L K 1988 Why won't caregivers use community services? Unexpected findings from a respite care demonstration/evaluation. In: George L K, Fillenbaum G G, Burchett B M (eds.) *Respite Care: A Strategy for Easing Caregiver Burden: Final Report*. Duke University, Center for the Study of Aging Family Support Program. Durham, North Carolina

Hess B B, Waring J M 1980 Changing patterns of aging and family bonds in later life. In: Skolnick A, Skolnick J H (eds.) *Family in Transition*. Little, Brown, and Co., Boston, pp. 521–37

Hooyman N 1990 Women as caregivers of the elderly: social implications for social welfare policy and practice. In: Biegel D E, Blum A (eds.) *Aging and Caregiving: Theory, Research and Policy*. Sage, Newbury Park, CA, pp. 221–41

Horowitz A 1981 Sons and daughters as caregivers to older parents: Differences in role performance and consequences. Paper presented at the annual meeting of the Gerontological Society of America, Toronto

Kane R A, Kane R L 1985 The feasibility of universal long-term care benefits. *New England Journal of Medicine* **312**: 1357–64

Lang F R, Staudinger U M, Carstensen L L 1998 Perspectives on socioemotional selectivity in late life: How personality and social context do (and do not) make a difference. *Journal of Gerontology* **53B**: 21–30

Litwak E 1985 *Helping the Elderly: The Complementary Roles of Informal Networks and Formal Systems*. Guilford Press, New York

Penning M J 1990 Receipt of assistance by elderly people: hierarchical selection and task specificity. *The Gerontologist* **30**: 220–7

Stone R, Cafferata G L, Sangl J 1987 Caregivers of the frail elderly: A national profile. *The Gerontologist* **27**: 616–26

Wenger C, Grant G, Nolan M 1996 Older people as carers as well as recipients of care. In: Minichiello V, Chappell N, Kendig H, Walker A (eds.) *Sociology of Aging*. International Sociological Association, Research Committee on Aging, Australia, pp. 189–206

Zarit S H, Reever K E, Bach-Peterson J 1980 Relatives of impaired elderly: Correlates of feelings of burden. *Gerontologist* **20**: 649–55

N. L. Chappell

Cargo Cults

1. *Preliminary Definition*

Cargo cults or movements are socio-magico-religious activities which have been occurring mainly in Melanesia (that part of Oceania comprising the island

archipelagoes from Irian Jaya in the west, to the east and southeast through Papua New Guinea, the Bismarcks, Solomon, Vanuatu, Fiji, and New Caledonia) since the 1850s (Steinbauer 1971, p. 181) until the present day. They include genuine sociopolitical and economic aspirations, but are in essence millenarian in nature. That is, the activities, usually initiated by someone generally like a prophet, are directed towards obtaining or greeting an expected state of bliss or contentment, the latter being envisaged as a free access to cargo (the goods and foodstuffs offloaded from the ships and, more recently, the aircraft of industrialized countries).

2. *Alternative Nomenclature*

Despite the name 'cargo' the cults may be considered a subset in a local idiom of all those movements of socio-religious reform, revival, and renewal which, endemic to the Christian ambience, have been variously called millenarian, messianic, or enthusiastic movements. Although similar activities have occurred where Christian influences have been marginal, including one cargo cult (Berndt 1952), the vast bulk of recorded instances have taken place where Christianity has been more or less established.

Still, avoiding Christian associations and the local cargo idiom, bringing cargo movements into line with similar kinds of activities found elsewhere in Oceania and among the indigenous peoples of the Americas, sub-Saharan Africa, Australasia, and parts of Asia (areas which Europeans have colonized or settled), a number of alternative terms to cargo, according to apparent main emphases, have been coined. Thus Accommodation, Acculturation, Adaptation, and Adjustment, also as verbs, reveal the activities as attempts to reconcile tradition with the ways of a more powerful and intrusive culture: Crisis and Disaster stress the effects of prior traumatic natural or cultural events; Nativistic, Militant, and Denunciatory emphasize the movements as protests against foreign rule together with the attempt to resolve problems by reaching back into tradition; Dynamistic, Vitalization, and Revitalization accent a positive cultural renewal in the face of what is perceived as decay or decadence; Prophetic, Charismatic, Messianic stress the importance given to the leader of a movement; and Christian influences are more directly evoked by Holy Spirit and Salvation movements.

While some cargo cults have occurred elsewhere in Oceania, they are usefully regarded as a Melanesian phenomenon. More than 400 instances have been recorded since the 1850s, and one must assume that others, in secluded areas, have gone unrecorded. The decade after the end of World War II, during which allied forces had deployed vast quantities of war materiel and general supplies throughout Melanesia, saw acceleration in the number of instances.

3. *Background*

From early in the nineteenth century the many Melanesian coastal communities have been subject to the increasing pressures of Euro-American imperialism. Nonliterate and differentiated by many languages and dialects but using Pidgin (*Tok pisin*) as a *lingua franca*, these hunting, fishing, and horticultural communities with stone age technologies, came into contact with Euro-American industrialized civilizations. Colonial forms of order were imposed. New infectious diseases ravaged populations. Money, taxes, plantations, and forms of indentured and contract labor were introduced. From the earliest years, Christian missionaries of different denominations communicated their varying messages and doctrines. Fortune hunters, labor recruiters, traders, and prospectors tended to be 'rough' as well as 'ready.' In turn, Germans, British, Australians, Japanese, and Americans have warred and brought their own goods and modes of government. Through all this Melanesians tried to live within their own community organizations based upon simple subsistence economies (defined as without money, a common and factorial medium of exchange) plus, in many areas, plantation or other service labor for cash—which alone could buy cargo. Some Melanesians, vulnerable to unstable world markets, engaged in cash cropping.

4. *Generalized Course of a Movement*

One of the more engaging qualities of Melanesians is their phlegmatic acceptance of the wonders of modern technology. Still, what seems to have impressed them most about white people has been not their particular human vices and virtues or their military prowess, but their apparent free access to cargo: all those useful artifacts, sacks of flour and rice, canned foods, and frozen carcasses. A persistent theme of day to day life was, and in places still is, the 'secret' of that access. Often, one who alludes to the subject is regarded as a bore and ignored. Still, the problem itself remains: speculation and supposition are joined to traditional myths and, indeed, Bible stories—particularly Noah's curse on the sons of Ham (Genesis 9:25)—to explain why whites seemed to have free access to cargo while Melanesians did not. Such stories and speculations make up what Burridge (1963, p. 147) has called a myth-dream: a compost of hopes, desires, and possibilities which a prophet or visionary may bring into a focus and coherence through action instructions.

Given the context of the myth-dream, the prophet usually reports a peculiar experience: a dream or vision, an encounter with a traditional spirit entity, an ancestor perhaps, or a Christian representation such as the archangel Gabriel, Virgin Mary, Holy Spirit, or other entity. Such an experience may become part of the myth-dream even if, as often occurs, it is shrugged

off. Sometimes, however, prophet, mystical encounter, and action instructions come together into cogent and persuasive synthesis. What it is that transforms a possibly overwrought imagination into an active cult or movement, which includes those who are otherwise sober, practical, and businesslike, is, so far, elusive.

Sometimes visionary, prophet, and leader are contained in the same person. At other times, depending on capacities for organizing others, they separate. The explicit objective of a movement, access to cargo, as well as the means thereto have led to the activities being called 'bizarre.' These have included: destruction of crops and/or traditional sacra; marching about with wooden 'rifles'; taking scribbled pieces of paper ('cheques') to a store; signaling for cargo with 'radios' of palm thatch; forms of 'baptism' (taken from Christian rites); orgiastic dances and sexual promiscuity (re-enacting traditional creation myths in which chaos and disorder give way to the moralities and rules of sexual access and marriage). Often there is a rider: those who do not participate will be doomed to destruction.

Failure of the cargo to arrive, or administrative action, usually bring activities to an end. Still, they may smolder and develop, as the Jon Frum movement on Tanna in Vanuatu (Guiart 1962) has done over the years into stable, if syncretic, sects. However, discussions continue, the myth-dream is kept alive. The question of the 'secret' remains: how do white people get cargo?

5. *Interpretations*

The kinds of interpretation of cargo cults are implicit in the nomenclature: reactions with differing emphases to the cultural stresses of Melanesian history that have been summarized in the previous sections. Socioeconomic analyses going ultimately to kinds of deprivation, especially problems attending the advent of cash, have dominated. Such interpretations tend to be historically particularist. They do not necessarily suffice for very similar kinds of movement in, for example, California where, instead of cargo, the idiom may be access to spacecraft or redemption by aliens from space. Although one is forced to consider the conditions of life of participants, these conditions are not exhausted by the purely socioeconomic or even political.

For example, in *Tok pisin* the word *wok* (= work) refers to magical rites and spells, cult activities and also to work in gardens, forests, or at sea, which last translates into food, giving feasts, making exchanges, gaining and maintaining relative status, and so political influence. Further, the symbolisms in the activities bespeak an ending of the old or present ways of life and a fresh start. This may be a movement from an economy based on exchanges of foodstuffs and valuables such as shell necklaces or chaplets of dogs' teeth, to one based upon cash which, in turn, would provide access to cargo and, perhaps, an equivalence in criteria of status in relation to whites.

Cargo movements thus envisage the creation of a 'New Man,' with access to cargo and capable of competing with whites on their own terms. If a new heaven seems cloudy, it is mainly embodied in a typical Melanesian way in a new earth where access to cargo will reward the faithful, with more for the most astute.

The incidence of cargo cults, in particular the high concentration in former New Guinea where *Tok pisin* is/was most used, hints at other factors at work. The few occurrences in Polynesia and Micronesia, where *Tok pisin* is absent, and in the Papua New Guinea Highlands where, despite a more ordered colonial penetration, much the same socioeconomic conditions in relation to whites existed, presents problems. One answer is that in the context of the myth-dream the *Tok pisin* word *kago* (= cargo) accretes to itself a transcendent sense of redemption from perceived inequities which the English 'cargo' does not.

Furthermore, it is noticeable that cargo cults tend to occur most frequently in those areas, such as in former New Guinea excluding the Highlands, where (due in part to population depletion from introduced diseases) community organizations involved shifting political and social leaderships. In consequence, infirm political authority and ambiguities of obligation and identity—just that 'anomy and incertaintie' noted in the seventeenth century of English enthusiastic movements (Burridge 1960, p. 13). Elsewhere in Oceania as well as in the Highlands of Papua New Guinea not only were (and are) communities much larger and more stable, but the loci of obligation and political

6. *Conclusion*

Over the space of 70 years for the Highlands of Papua New Guinea, and an added near century for most of the peoples in the coastal and intermediate areas, Melanesians have moved from a stone age technology rich in symbolism (where every quality of human character and nuance of change in climate, vegetation, or animal behavior was charged with meaning) into an ambience where bureaucracy, science, and reason are paramount. The former qualities are discounted in favor of quantity, and symbolism is derided as superstition. In such a world, except where access to money and the goods it can buy are difficult, for most Melanesians cargo cults are becoming an anachronism.

As a note of caution, long ago cult leaders and influential Melanesians were taken to Australia where they were conducted around factories to see how cargo was made. Yet this only took the central question one step deeper: whence the ability to make cargo? Why were the necessary resources not available in New authority have persisted in reasonable certainty.

Guinea? Perhaps, following American example and in contemporary idiom, future cult activities may focus on the Book of Revelation from the Bible or be directed toward access to space vehicles. Still, they will pose much the same basic questions for social scientists, as have cargo movements. The question of just why a cult does or does not occur is still a mystery. Melanesian scholars will no doubt reveal many new nuances, but whether they will do any better in relation to their movements than Europeans have done for theirs, remains to be seen.

See also: Belief, Anthropology of; Cognitive Anthropology; Conflict: Anthropological Aspects; Economic Anthropology; Exchange in Anthropology; Horticultural Societies; Melanesia: Sociocultural Aspects; Millennialism; Political Anthropology; Political Economy in Anthropology; Symbolism in Anthropology

Bibliography

Belshaw C S 1954 *Changing Melanesia*. Oxford University Press, Melbourne, Australia

Berndt R M 1952 A cargo movement in the central Highlands of New Guinea. *Oceania* **XXIII**: 40–65, 137–58, 202–34

Burridge K O L 1960 *Mambu*. Methuen, London. [1970 Harper Torchbooks, New York; 1995 Princeton University Press, Princeton, NJ]

Burridge K 1969 *New Heaven, New Earth*. Basil Blackwell, Oxford, UK

Guiart J 1962 *Les Religions de l'Oceanie*. Presses Universitaires de France, Paris

Knox R A 1950 *Enthusiasm*. Clarendon Press, Oxford, UK

Lawrence P 1964/1971 *Road Belong Cargo*. Manchester University Press, London

Lindstrom L 1990 *Knowledge and Power in a South Pacific Society*. Smithsonian Institution Press, Washington, DC

Steinbauer F 1971 *Melanesian Cargo Cults* [trans. Wohlwill M]. George Prior Publishers, London

Trompf G W (ed.) 1990 *Cargo Cults and Millenarian Movements*. Mouton de Gruyter, Berlin

Wallace A F C 1956 Revitalization movements. *American Anthropologist* **LVIII**: 264–81

Worsley P 1957 *The Trumpet Shall Sound*. MacGibbon & Kee, London (Rev. Edn. 1968)

K. Burridge

Caribbean: Sociocultural Aspects

1. A String of Isles

The Caribbean archipelago, sometimes called the Antilles or the West Indies, comprises nearly a thousand identifiable islands and islets that spread over 2,500 miles around the Caribbean Sea. The archipelago spans southeasterly from southern Florida to the Gulf of Paria, then turns westward along the coast of Venezuela.

The islands are divided into four groups. The most northern group is the Bahamas, a string of several hundred islets, of which 29 are inhabited, that stands apart from the Antilles. The next two groups, the Antilles proper, are the demographic, economic, and cultural heart of the region. The Greater Antilles consist of the islands of Cuba, Hispaniola (comprised of Haiti and the Dominican Republic), Jamaica, Puerto Rico, and their respective dependencies. They encompass the bulk of the Antillean lands and house three-quarters of the region's inhabitants. The Lesser Antilles, which include the Leeward and Windward Islands, span like a crescent at the eastern end of the archipelago, and count about 40 inhabited isles. A fourth and looser grouping includes the islands off Venezuela, from Tobago and Trinidad to Aruba. Beyond the archipelago, most social scientists now include in the Caribbean the Colombian islands off the coast of Nicaragua and the mainland territories of the Guianas (Guyana, Suriname, Belize, and Cayenne, also known as French Guiana).

2. An Obvious Heterogeneity

Scholarly treatments of the Caribbean as a single object of study, and the related conceptualization of the region as a distinguishable sociocultural area, are both recent and controversial. Poets and novelists, such as Cuban writers Alejo Carpentier and Nicolas Guillen, were the first to herald Caribbean sociocultural unity early in the twentieth century. This call was renewed by recent literary figures such as Antonio Benitez-Rojo and Edouard Glissant. Most social scientists, however, took the old colonial boundaries for granted, dividing the region into linguistic spheres that duplicated the European dominions and across which they saw few similarities. By the early 1960s, international symposia on the region (e.g., Rubin 1960) and the comparative sketches that appeared in the *Nieuwe West Indische Gids*, *Caribbean Studies*, and *Social Economic Studies* implicitly acknowledged the overall unity of the Caribbean. Soon after, explicit treatments of the region as a single object of study—and parallel efforts to conceptualize its structural similarities—emerged in European geography (e.g., Lowenthal 1972) and North American anthropology (e.g., Mintz 1966, 1984).

The earlier reluctance of social scientists to view the Caribbean as a whole is understandable, and not just because their research evolved within European colonial studies and often reproduced the insularism of the local elites. With 40 million people spread over some 90 territories, most of which are tiny and surrounded by sea, the Caribbean displays a wide range of similarities and differences. No single sociological or cultural feature stands out as the defining

essence of the region. Thus, complexity in regard to size and population and social heterogeneity within and across political boundaries are two defining themes of Caribbean Studies, even among the social scientists who question the singleness of the region (Trouillot 1992).

Social scientists today see the division of labor, past and present, as the root cause of Caribbean diversity both within and across territories (Cross and Heuman 1988). In the last five centuries, the Caribbean has experienced a wide range of labor regimes: slavery, indentured labor, sharecropping, peasant agriculture, simple commodity production in family-based enterprises, agricultural and industrial wage labor. Except for the barely known practices of its earliest inhabitants, the Arawaks and Island Caribs who probably specialized in swidden agriculture and small-scale fishing respectively, these labor regimes all reflect the incorporation of the Caribbean in the capitalist world economy. Today, while most Caribbean men and women still derive the bulk of their income from agriculture, some of the islands have specialized in offshore banking. Others have become favored hubs in assembly industries or electronic communications. Others rely heavily on tourism, especially from the United States, which now exercises unmatched economic and political power over most independent states within the region.

The political scene is as varied as the economic landscape (Stone 1985). The Caribbean bears the mark of a variety of political systems: colonies and neocolonies with different degrees of limited autonomy; three monarchical experiments in nineteenth-century Haiti; Westminster-style parliamentarism in the former British colonies; civil and military dictatorships notably in Cuba, the Dominican Republic, Haiti, and Suriname; and various unusual and self-labeled versions of socialist rule. Elections are now the norm in most countries, but the formal adherence to democratic procedures barely hides a wide variety of practices from blatant fraud and unabashed clientelism to single-party rule and electoral divisions along ethnic lines.

This mosaic of economic and political formulas parallels an impressive variety of languages and religions. Caribbean languages include Spanish, French, English, Dutch, and Creole languages of all kinds, echoes of the many European nations that influenced the area (Taylor 1977, Christie 1996) Likewise, almost all the variations of Christianity can be found in the area. Jewish influence is not to be dismissed; nor is that of Islam. Hinduism and Buddhism strive in Trinidad and Guyana. Native Caribbean religions complete the denominational mosaic: Haitian Vodoun, Trinidadian Shango, Cuban Santeria, Jamaican Pocomania and Rastafarianism, belief systems and practices where the Old World meets the New. The influence of the Old World reaches its peak with the African contribution. Caribbean populations descend directly from regions of Africa that correspond to at least 18 contemporary states.

3. *Family Resemblance: The Shapes of History*

Beyond this profusion of traits and descent lines, the unity of the Caribbean is one of family resemblance. Different characteristics account for the likeness between any two territories within and across linguistic boundaries. Yet the web of parallelisms, relationships, and genealogies that spans the entire archipelago and spills over into the Guianas makes the area, as a whole, quite distinct from any other world region.

Caribbean family resemblance is governed by the shared experience of power. Caribbean societies evolved in the shadow of power yielded almost always vertically from within and almost always resting on the ultimate domination of Europe and, later, of Europe and the United States. Six overlapping features typify this exercise of power: (a) the decimation of the native population, which created an effective *terra nullius*, a land literally up for grabs; (b) a subaltern integration in the international order characterized by the duration and extent of external rule (thus the depth of intrusion in local life); (c) the extreme regimentation of populations that paralleled this intrusion; (d) plantation slavery as the epitome of the last two features; (e) the continuity of institutionalized—often state-enforced—exclusions and hierarchies; and (f) the no less continuous regimentation of cultural practices. Scholarship on the region now tends to address these themes singularly or in combination.

Modern Caribbean history starts in the early sixteenth century with the swift decimation of the original inhabitants of the Antilles from disease, warfare, mistreatment, and forced labor. From then on, Europeans moved through the islands as if they were empty lands to be fashioned exclusively for goals that originated elsewhere. Indeed, the Caribbean stands out in the world as an exceptional product of European colonialism. Caribbean territories have experienced Western European influence longer and more profoundly than any other area outside of Europe itself. Nowhere else have European states held onto dependencies for so long and shaped them so deeply without having to take into account the strength or relevance of native institutions. Almost everything that we now associate with the Caribbean—from sugar cane, coffee, mangoes, bananas, donkeys, and coconuts, to the people themselves, whether African or Asian in origin—was brought there as part of the European conquest.

External rule dictated the coercion of labor and the regimentation of local populations, both of which reached their peak during the 370 some years that African slavery lasted in the Caribbean. The slave trade itself lasted from about 1518 to 1873. During

that time, the Caribbean imported at least four million African slaves, perhaps one third of all the Africans who came to the Americas. In comparison, the United States imported about half a million.

Caribbean slavery was plantation slavery, sponsored by European capitalists and geared toward the production of tropical crops for export. Sugar cane and coffee dominated the system, but tobacco, cotton, and indigo were also important. The impact of slavery on the cultural and social life of Caribbean populations may be the most evident feature behind Caribbean family resemblance. The role of particular territories varied during the centuries of slavery. Yet the entire region was deeply molded by that experience.

In 1791, the slaves of Saint-Domingue/Haiti, then France's most lucrative colony and perhaps the most profitable dependency of any European power, started an uprising that augured a new phase in Caribbean history. After 13 years of war, they defeated the formidable army sent by Napoleon to restore slavery and proclaimed the independence of Haiti in 1804. The Haitian Revolution signaled the end of Atlantic slavery. Throughout the century, European powers successively abolished either the slave trade or slavery itself in a process completed in the Antilles with the abolition of slavery in Puerto Rico (1873) and Cuba (1880).

Abolition did not bring an end to the plantation system or the coercion of the labor force. Hundreds of thousands of indentured laborers, mainly from China and South Asia, were brought to replace the blacks on sugar plantations. Caribbean people of South Asian descent now constitute important ethnic subgroups from Guadeloupe to Suriname and, most notably, in Trinidad and Guyana. In the latter two places, the Afro-Caribbean descendants of the former slaves found new niches in the agro-social system, becoming independent peasants or, later, part of a growing middle class tied to the state. The peasantry was even stronger in the mountainous islands of the Windwards, and in Haiti and Jamaica. At the other end of the spectrum the plantation system outlasted slavery, even without massive input of Asian labor, in places like Barbados, Cuba, Antigua, or Puerto Rico—though in Puerto Rico it coexisted with a peasantry whose roots anteceded freedom itself.

Whether paid laborers or independent peasants, Caribbean rural dwellers tend to be marginalized politically, socially, and culturally by local elites that replaced European colonizers. Indeed, the whole region is deeply marked by social exclusion based on markers such as skin color, ethnic affiliation, class origins, occupational status, religion, and language. That exclusion, often institutionalized, intensifies a regimentation of cultural practices which dates back to slavery. Explicit or tacit codes limiting the reach and value Creole speech, the denigration of Afro-Caribbean religions throughout the region, the persecution of Rastafarians in Dominica or Jamaica, of Maroons in Suriname and of homosexuals in Cuba, are all evidence that cultural coercion continues in the Caribbean.

4. *The Caribbean in the Social Sciences*

Caribbean social science started long before the institutionalization of current disciplines, with the field observations and commentaries of colonists such as Bartolomé de las Casas, Jean-Baptiste Labat, Bryan Edwards, John Gabriel Stedman, and Moreau de Saint-Méry. Since then, local elites, colonial and postcolonial, have generated a huge literature that prefigures and continues to fuel the themes favored today by professional academics (Lewis 1983). Haiti and Cuba held the lead in the social sciences until the second half of the twentieth century. Since the 1960s, however, the anglophone Caribbean has generated a spectacular amount of social scholarship. The impact of the various local schools and the widespread acknowledgement outside of the region that its colonial history profoundly shaped its present give Caribbean social science a decisive historical bent.

History generates the greatest amount of scholarship about and from the region. The record begins with the colonists, includes the Haitian pioneers of postcolonial history, Thomas Madiou and Beaubrun Ardouin and mushrooms in the twentieth century with the works of Elsa Goveia, C. L. R. James, Eric Williams, Manuel Moreno Fraginals, Jean Fouchard, Walter Rodney, Barry Higman, and a number of historians from the University of the West Indies. Williams set the connection between capitalism and slavery as a central theme of Caribbean and Afro-American studies. James made the Haitian Revolution a legitimate object of study outside of Haiti. Fouchard pioneered twentieth-century studies of maroon slaves.

Today, the anglophone world—England and especially the United States—produces the greatest number of titles on Caribbean history, although contributions from Dutch researchers remain significant. The state of Caribbean historical research is summarized in UNESCO's six-volume *General History of the Caribbean*. Slave studies remain a highlight of current research, with increasingly detailed accounts of plantation life and coordinates. In recent years, the Haitian Revolution, slave resistance, the rise of peasantries, women's history, postslavery life in general, and US interventions in the region have attracted increased attention.

History's lead is not confined to the number of titles it generates. The disciplines next in line—anthropology in particular, but also geography, sociology, economics, and even political science—tend to frame their objects against a strong historical background. Because Caribbean social science cannot ignore the

role of North Atlantic power in shaping the region, it constantly arks back to the consequences of colonialism, of external domination, and of the region's integration in the capitalist world economy. For instance, economics focuses as much on the plantation system, past and present, as on the global mechanisms that sustain the Caribbean's incorporation in the Atlantic world, often echoing dependency theory as exposed in continental Latin America (Beckford 1999).

Even economists not aligned with theories of dependency take power and history into account. Thus, while Sir Arthur Lewis's focus on 'the dual economy' and the 'traditional sector' won him a Nobel Memorial Prize, fellow Caribbeanists are acutely aware that the work of the St. Lucian economist was deeply involved in the history that produced the dualism he described (Lewis 1978). Within Caribbean Studies, the social sciences in general move in tandem with historiography, always taking into account the state of knowledge about where these societies come from even when the explicit purpose is to explain where they are now or where they might be going in the future.

Not surprisingly, a leading theme in Caribbean social studies has been the continuous assessment of the specific heritage of various Old World regions on particular territories when not on the entire area. Here again, social scientists have to navigate between obvious signs of blending and no less obvious signs of heterogeneity to account for their observations. Thus, local observers have produced a number of theoretical schemes—such as Cuban Fernando Ortiz's notion of 'transculturation,' later recycled by North Atlantic anthropology—to bypass the difficulties inherent in the enterprise. On the ground, however, Ortiz himself joined a number of researchers who focused on Africa's influence on social, political, and economic institutions (from female-dominated retail trade to matrifocal families), cultural beliefs and practices (from religion to language) to the arts (most notably music). The attention to gender roles and family structure is central to the work of a number of scholars, not all of whom accept the centrality of the African legacy (e.g., Smith 1988). On the other hand, the emphasis on African continuities is central to the work of Haitian Jean Price Mars, a founder of the Négritude movement, who renewed with the call for the re-evaluation of Africa first made by early nineteenth-century Haitian writers.

The more specific search for sub-Saharan 'survivals' became institutionalized in North Atlantic universities with the work of US anthropologist Melville Herskovits who saw the Caribbean as a subset of the Afro-Americas, much like his French counterpart, Roger Bastide. Price-Mars and Herskovits also share an emphasis on the 'cultural ambivalence' of Caribbean elites, torn in their eyes between their African and European heritage—an ongoing subtheme of Caribbean critical inquiries, most recently sustained by the cultural critique of Rex Nettleford and linguist Mervyn Alleyne in Jamaica. Since the late twentieth century, however, a number of writers, notably anthropologists Mintz and Price (1992) have critically reassessed the Herskovits agenda. Many scholars now emphasize the principles and processes behind Caribbean sociocultural development rather than trace the Old World origins of singular traits, thus overlapping studies in creolization.

Creolization—the process by which newly arrived populations facing severe physical and institutional constraints generated a distinct mode of life and developed cultural beliefs and practices, including linguistic practices, that have become distinctively native to the region—has a long pedigree in Caribbean studies. Here again, the puzzle remains how to reconcile heterogeneity and family resemblance. A century after the beginning of the slave trade, local and foreign observers began to describe the features and development of Creole languages and wondered about the similarities across islands. Creole linguistics have now become a subfield with its own insular debates. Yet other aspects of the creolization process and the nature of Creole society itself remain central and cross-disciplinary themes in Caribbean social and cultural studies (Arnold 1998).

The emphasis on social and cultural blending, favored by various students of creolization, has long run counter to the emphasis on heterogeneity among observers who see Caribbean societies as unwieldy patchworks of groups solidly marked by skin color and artificially joined by power. Here again, the trend starts in colonial times. It runs through the nineteenth, and is revived in the early twentieth by writers such as Lorimer Denis and François Duvalier in Haiti. Yet it took Jamaican anthropologist Smith's application of the plural society model (1965) to give that line of research much of its firmness and exposure in Caribbean studies. Smith's work continues to fuel virulent controversies but the fundamental question it raises remains unavoidable. Given what is known of Caribbean heterogeneity, how does a social scientist relate current institutions and other elements of the social system to the cultural traditions of the peoples involved?

Yet the question leads to an impasse only if one looks for a unified content, a single essence that epitomizes Caribbeanness. The now common vision of the Caribbean as a sociocultural area where heterogeneity and family resemblance are inherently indissociable products of a history of uneven power provides the frame within which most scholars approach the region today (Mintz and Price 1985, Watts 1987)

5. *The Exploding Islands*

Just as the vision of the Caribbean as a single object of study becomes accepted in social science circles (e.g.,

Mintz and Price 1985, Cross and Heuman 1988) it faces new challenges in a world increasingly dominated by its powerful US neighbor (Serbin 1998). It must now contend with the growth in rural–urban migration, the crossing of borders within the region and the size and visibility of Caribbean diasporas in the North Atlantic.

Caribbean people have been moving within and across political borders since slavery days, and migration has been a major theme in Caribbean Studies since at least the 1960s. Today however, migratory flows within single islands, within the region, and away from it have reached unprecedented proportions. As peasants and rural proletarians rush to the cities, the expansion of urban slums now redefine the sociocultural landscape in all territories. The growing presence of intra-regional migrants—such as Haitian cane cutters in the Dominican Republic, Dominicans in Guadeloupe and Martinique—also questions boundaries once thought impermeable. Most important, outside contact has become part of daily local life. The number and visibility of Caribbean migrants in North Atlantic countries, and the fact that breakthroughs in transport and communications enhance their economic and cultural impact on their country of origins make diasporas an inherent part of Caribbean Studies (Brana-Shute 1983). Urban glut, and the frequency of outside contact have also increased the crime rate, notably infractions associated with drug use or trade. The concept of a single sociocultural area will have to accommodate this geographical and social remapping of the region and redraw its own heuristic boundaries accordingly.

See also: Colonialism, Anthropology of; Colonialism: Political Aspects; Colonization and Colonialism, History of; Creolization: Sociocultural Aspects; Historiography and Historical Thought: Indigenous Cultures in the Americas; Slavery as Social Institution; Slavery: Comparative Aspects; Slaves/Slavery, History of

Bibliography

Arnold A J (ed.) 1998 Who/what is 'creole'? *Plantation Society in the Americas* **5**(1)

Beckford G 1972/1999 *Persistent Poverty: Underdevelopment in Plantation Economies of the Third World*. The University of the West Indies, Kingston, Jamaica

Brana-Shute R 1983 *A Bibliography of Caribbean Migration and Caribbean Immigrant Communities*. University of Florida, Gainesville, FL

Christie P 1996 *Caribbean Language Issues, Old & New: Papers in Honour of Professor Mervyn Alleyne*. University of the West Indies, Kingston, Jamaica

Cross M, Heuman G (eds.) 1988 *Labour in the Caribbean*. Macmillan Caribbean, London

Lewis G K 1983 *Main Currents in Caribbean Thought. The Historical Evolution of Caribbean Society in its Ideological Aspects, 1492–1900*. The Johns Hopkins University Press, Baltimore, MD

Lewis W A 1978 *The Evolution of the International Economic Order*. Princeton University Press, Princeton, NJ

Lowenthal D 1972 *West Indian Societies*. Oxford University Press, Oxford, UK

Mintz S W 1966 The Caribbean as a socio-cultural area. *Cahiers d'Histoire Mondiale* **IX**: 916–41

Mintz S W 1974/1984 *Caribbean Transformations*. The Johns Hopkins University Press, Baltimore, MD

Mintz S W, Price R 1976/1992 *The Birth of an African-American Culture: An Anthropological Perspective*. Beacon Press, Boston

Mintz S W, Price S 1985 *Caribbean Contours*. The Johns Hopkins University Press, Baltimore, MD

Rubin V D (ed.) 1960 *Caribbean Studies: A Symposium*. University of Washington, Seattle, WA

Serbin A 1998 *Sunset over The Islands. The Caribbean in an Age of Global and Regional Challenges*. St. Martin's Press, New York

Smith M G 1965 *The Plural Society in the British West Indies*. University of California Press, Berkeley, CA

Smith R T 1988 *Kinship and Class in the West Indies*. Cambridge University Press, Cambridge, UK

Stone C 1985 The Caribbean as a political region. In: Mintz S W, Price S (eds.) *Caribbean Contours*. The Johns Hopkins University Press, Baltimore, MD

Taylor D M 1977 *Languages of the West Indies*. The Johns Hopkins University Press, Baltimore, MD

Trouillot M-R 1992 The Caribbean region: An open frontier in anthropological theory. *Annual Review of Anthropology* **21**: 19–42

UNESCO 1997–2001 *General History of the Caribbean*. UNESCO and Macmillan Caribbean, London, 6 Vols.

Watts D 1987 *The West Indies: Patterns of Development, Culture and Environmental Change since 1492*. Cambridge University Press, Cambridge, UK

M. R. Trouillot

Cartographic Visualization

1. Setting

Assume one is interested in an overview of the population distribution of a particular area. An option would be to give a written description such as: 'in the northeast only a few small towns are found, while in the southeast....' Before one gets a clear picture, several pages could have been filled, especially if a description of the relations of the towns with other geographic phenomena such as the rivers or railroad network are also incorporated. A better option is to

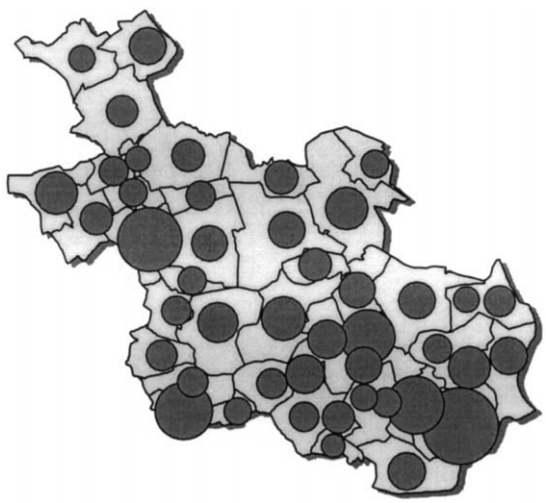

Figure 1
Population distribution

tell this story using a map. Maps are the most efficient and effective means to transfer geospatial information. The map user can locate geographic objects, while the shape and color of signs and symbols representing the objects inform about their characteristics, such as the town's location and population size. Maps reveal geospatial relations and patterns, and offer the user insight into, and an overview of, the distribution of particular phenomena, such as an area's population distribution.

The map easily allows one to answer questions of the nature 'Where…? What…?, and When…?' These questions deal with the basic components of geospatial data: location, characteristics, and time, or their combination. The map discussed above and found in Fig. 1 will quickly result in answers to questions such as 'Where are the large towns located?' or 'What town has the highest number of inhabitants?' Questions regarding 'When…?' depend on the map contents. Often the map is just a snap shot in time, for instance the map in Fig. 1 depicts the status as of 1998. In answering questions, another quality of maps will be revealed—the ability to offer an abstraction of reality. It simplifies by selection, but at the same time it puts, when well designed, the remaining information in a clear perspective. The map of the area only needs the boundaries of municipalities, and a symbol for the number of people living in each town. In this particular case there is no need for roads, mountains, or other physical features. Maps like these are often static and come as they are, especially when they are printed on paper. On-screen maps sometimes allow for some interaction, such as getting data from the database behind the map.

For more complex questions one needs more than just a static map. For instance, to study the impact of industrial development on the growth of towns one would need the possibility to generate maps depending on the nature of the question. This could be to switch layers with information on or off. This would also include the possibility to view the data in alternative visualizations. Questions are no longer the simple Where?, What?, and When?, especially if for instance specific demographic models are included, which can predict alternative future population trends based on census- and socio-economic data. These trends could be visualized in an animation. The mapping environment requires interaction and dynamics. This is what today's visualization is all about, as will become clear in the remainder of this chapter. However, according to the dictionary, visualization means 'to make visible.' From this perspective map making has always been visualization. Here the term visualization is used in the context that allows the map user to interact with the (digital on-screen) map. For a detailed discussion see (Hearnshaw and Unwin 1994, MacEachren and Taylor 1994).

2. *Visualization process*

Any map, static or dynamic, on screen or on paper, complex or simple, is created during what is called the cartographic visualization process. This process is considered to be the translation or conversion of geospatial data from a database into graphics. Predominantly these are map-like products as is schematically explained in Fig. 2. This process should be seen in the context of geospatial data handling. Geospatial data handling stands for the acquisition, storage, manipulation, and visualization of geospatial data in the context of particular applications.

During the visualization process, cartographic methods and techniques are applied. These can be considered as a kind of grammar that allows for the optimal design, production and use of maps, depending on the application. The process is guided by the phrase 'How do I say what to whom, and is it effective?' The phrase holds four key words: 'How' which refers to cartographic methods and techniques (in the case of Fig. 1 a proportional point symbol map has been

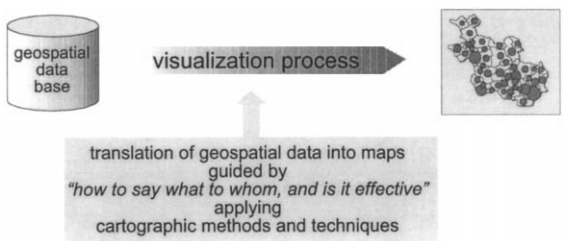

Figure 2
The cartographic visualization process

chosen); 'What' which refers to the geospatial data (Fig. 1 deals with quantitative population data); 'Whom' which refers to the map audience and the purpose of the map—(the map in Fig. 1 is rather basic and could function in a newspaper or school atlas); 'Effective' reflects the usefulness of the map (do the map readers understand the message the map intends to bring, an overview of the area's population distribution).

The producer of maps could be a professional cartographer, but could also be an expert who is mapping, for instance, vegetation stands using remote sensing images or health statistics in the slums of a city. With today's availability of web mapping tools the mapmaker could be anyone—with or without any notion of cartographic design.

The visualization process can vary greatly depending on where the visualization takes place and the purpose for which it is needed. Visualizations can be, and are, created during any phase of the geospatial data handling process. They can be simple or complex, while the production time can be short or long. Some examples are the creation of a full, traditional topographic map sheet, a newspaper map, a sketch map showing a route, a map from an electronic atlas, an animation showing the growth of a city, a three-dimensional view of a building or a mountain, or even a real-time display of traffic conditions via the World Wide Web. Other examples include 'quick and dirty' views of part of the database, the map used during the updating process, or during a geospatial analysis. The environment in which the visualization process is executed can vary considerably. It can be done on paper, on a stand-alone personal computer, or a computer linked to the World Wide Web.

Many tools are available to visualize the data. These tools consist of functions, rules, and habits. Algorithms to classify the data or smooth a coastline are samples of functions. Rules tell us, for instance, to use proportional symbols to display quantities or position an artificial light source in the northwest to create shaded relief maps. Habits, or traditions as some would call them, tell us to color the sea in blue, lowlands in green, and mountains in brown (Robinson et al. 1995, Kraak and Ormeling 1996).

3. *Visualization and cartography*

In the past, even dealing with incomplete and uncertain data, the visualization process nearly always resulted in an authoritative map. The maps created by a cartographer were good enough for the user. This shows that cartography, for a long time, has been very much driven by supply rather than demand. Somehow, this is still the case. However, nowadays, it is also accepted that just making a map is not the only purpose of cartography. Especially since 1980, many people have become involved in making maps. The widespread use of Geographical Information Systems (GIS) has significantly increased the number of maps being created (Longley et al. 1999). Even the spreadsheets used by most office workers today have mapping capabilities, although most people are probably not aware of this. The opportunities offered by the World Wide Web will again lead to an incredible increase in maps produced. Some websites, such as MapQuest produce over a million maps a day! Many of these maps are not produced as final products, but rather as intermediate products to support the user in his or her work dealing with geospatial data. The map, as such, has started to play a completely new role: it is not just a communication tool but also a tool to aid the user's (visual) thinking process.

This process is being accelerated by the opportunities offered by hardware and software developments. These have changed the scientific and societal needs for geo-referenced data and, as such, for maps. New media such as CD-ROMs and the WWW not only allow for dynamic presentation but also for user interaction. Users do expect immediate and real-time access to the data and data geospatial has become abundant. This abundance of data, welcomed in some sectors, is a major problem in other sectors. One lacks the tools for user-friendly queries and retrieval when studying the massive amount of data produced by sensors, and now available via the WWW.

These developments have given the word visualization enhanced meaning, since progress in other disciplines has linked the word to more specific ways in which modern computer technology can facilitate the process of 'making visible' in real time. Specific software toolboxes have been developed, whose functionality is based on two key words: interaction and dynamics. A separate discipline called scientific visualization, has developed around it (McCormick et al. 1987), which is having a major impact on cartography as well. If applied in cartography it offers the user the possibility of instantaneously changing the appearance of the map. Interacting with the map will stimulate the user's thinking and will add a new function to the map. As well as communication, it will prompt thinking and decision-making.

Developments in scientific visualization have stimulated a model for map-based scientific visualization (DiBiase 1990). As such, it is also known as Geographical visualization (MacEachren 1995). It covers both the communication and thinking functions of the map. Communication is described as 'public visual communication' since it concerns maps aimed at a wide audience. Thinking is defined as 'private visual thinking' because it is often an individual playing with the geospatial data to determine its significance (see Fig. 3). On a lower level, different visualization stages can be recognized: each requires a different strategy from the perspective of map use, based on audience, data relations, and the need for interaction. These

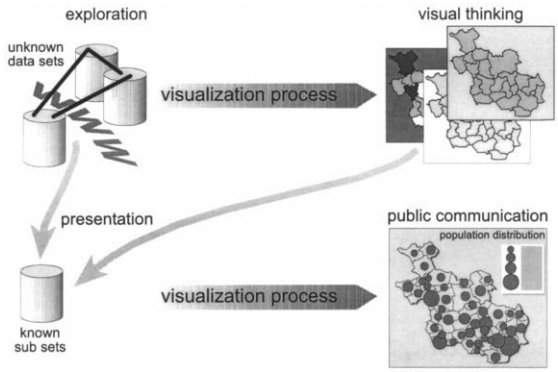

Figure 3
The visualization process: visual thinking and visual communication

stages are exploration, analysis, synthesis, and presentation.

From Fig. 3 it is obvious that presentation fits into the traditional realm of cartography, where the cartographer works on known geosptial data and creates communicative maps. These maps are often created for multiple uses. However, exploration often involves a discipline expert creating maps while dealing with unknown data. These maps are generally for a single purpose and are related to the expert's attempt to solve a problem. While dealing with the data, the expert should be able to rely on cartographic expertise provided by the software or some other means.

This process describes the 'democratization of cartography' (Morrison 1997). He explains it as 'using electronic technology, no longer does the map user depend on what the cartographer decides to put on a map. Today the user is the cartographer.' And 'users are now able to produce analyses and visualizations at will to any accuracy standard that satisfies them.'

4. *Visualization and exploration*

What is exploratory cartography? The environment has just been described: a person is trying to solve a particular geo-problem and is exploring various geospatial databases. Exploration also means working with unknown data. However, what is unknown for one is not necessarily unknown to others. For instance, browsing in Microsoft's Encarta World Atlas is an exploration for most of us because of its wealth of information. With products like these, such exploration takes place within boundaries set by the producers. Cartographic knowledge is incorporated in program wizards resulting in pre-designed maps. Some users feel this to be a constraint, but those same users will no longer feel constrained as soon as they follow the web links attached to this electronic atlas. This example shows that the environment, the data, and the type of users influence one's view of what exploration entails.

Returning to the sentence driving the visualization process 'How do I say what to whom and is it effective' some similarities as well, as differences between modes of visualization: presentation and exploration. 'How' still represents the cartographic methods and techniques. However, new technology is emerging and this offers challenges and opportunities, such as animation, the application of the third dimension and virtual reality, multimedia, etc. 'I' is no longer just the cartographer, but an expert geoscientist. In the very near future, it can probably be just anyone having access to the WWW. 'What' no longer represents a relatively well-defined and known data set; at least, certainly not from the user perspective. 'To whom' seems to be simpler than before; it is not a relatively well-defined user group, but the same person represented by 'I,' the expert geoscientist in the role of cartographer. 'Effective' raises some interesting questions. When a map is used, the information to be transferred is known and, for all problems involved, can somehow be measured. But how can the visual thinking process be measured? If it is considered positively, is it because of the efficient graphics or because of the geoscientist's clever thinking? These questions become more complex if we realize that we do not even know the initial aim of the visualization in these circumstances.

The most prominent change is the shift from supply-driven cartography to a demand-driven approach. Although having many people making maps without any cartographic knowledge might seem wrong, those non-cartographers will also introduce fresh views, as for instance described by Keller and Keller (1993). They distinguish three steps in the visualization process: the first identifies the visualization goal; the second removes mental roadblocks; and the third designs the display in detail. In the second step, Keller and Keller suggest taking some distance from the discipline in order to reduce the effects of traditional constraints. Why not choose an alternative mapping method? For instance, show a video of the landscape next to a topographic map. New, fresh, creative graphics could be the result, they might also offer different insights and would probably have more impact than traditional mapping methods. During the third step, which is especially applicable in an exploratory environment, one has to decide between mapping data and visualizing phenomena.

An exploratory visualization environment offers the tools to act in the ways previously suggested. Such an environment should allow the user to look at geospatial and other geo-referenced data in any combination, at any scale, with the aim of seeing or finding geospatial patterns (which may be hidden). Geospatial patterns can be defined as variations in location, attributes or time, or a combination of any of the three geospatial components within an area of interest. One

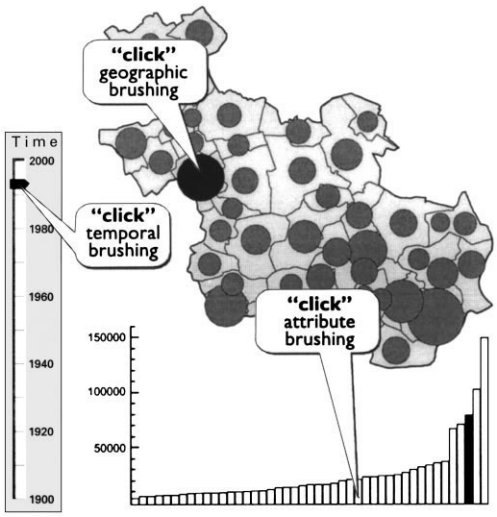

Figure 4
Brushing

of the first concepts of visual geospatial data exploration was introduced by Monmonier (1989) when he described the term brushing (see Fig. 4). This is when the selection of an object in a map automatically highlights the corresponding elements in the other graphics. Depending on the view in which one selects the objects, there is geographical brushing (clicking in the map), attribute brushing (clicking in the diagram or table), and temporal brushing (clicking on the time line). As such, the user gets an overview of the relation among geographic objects based on location, characteristics, and time.

5. *Visualization Functions*

What are the basic requirements of an exploratory visualization environment? The necessary functions described below all need high interactivity options.

5.1 *Basic Display*

Map displays need tools to allow the user to pan, zoom, scale, transform, and rotate the image contents. These geometric tools should be available and independent of the dimensionality of the displayed geospatial data. Fig. 5(a) illustrates panning and zooming facilities. Another option could be rotating the map. This last function is of particular need when displaying 3D-maps, since it also allows objects that otherwise might be hidden by other objects to be seen.

5.2 *Navigation and Orientation*

This involves the keys to the map. At any time, the user should be able to know where the view is located and what the symbols mean. To illustrate this function Fig. 5(b) shows a map with its marginal information as well as the coordinates at the cursor location in the map.

5.3 *Query Data*

During any phase of the visualization process, the user should have access to the geospatial database to query the data. The questions should not necessarily be limited to simple What? Where? or When? As Fig. 5(c) shows clicking a geographic object in the map reveals the information available in the database, as well as the hyperlinks that are attached to the object.

5.4 *Multi-Scale*

Combining different data sets is a common operation in an exploratory environment. The chance that these sets will have the same data density or the same level of abstraction is unlikely. Generalization operators to solve these multi-scale problems remain necessary, if not just to make sure zoom-in and zoom-out operators result in sensible maps. In Fig. 5(d) zooming-in results in a more detailed map.

5.5 *Re-expression*

To stimulate visual thinking, an unorthodox approach to visualization was recommended. This requires options to manipulate data behind the map or offer different mapping methods for displaying the data. An example of data manipulation is the application of several classification systems, while the use of different advanced map types to view the same data represents display differences. Fig. 5(e) illustrates how the population data shown in the figures main map can be displayed alternatively. The upper inset shows a prism map (a 3D representation of the data in which the height of the area corresponds to the attribute value; the big towns have high columns) and a cartogram (a map in which the area of a geographic unit is equal to the attribute value and not, as usually, to its geographic area; the highly populated towns take a large maps space).

5.6 *Multiple Dynamically-linked Views*

These tools represent a combination of multimedia and the brushing technique already mentioned. The

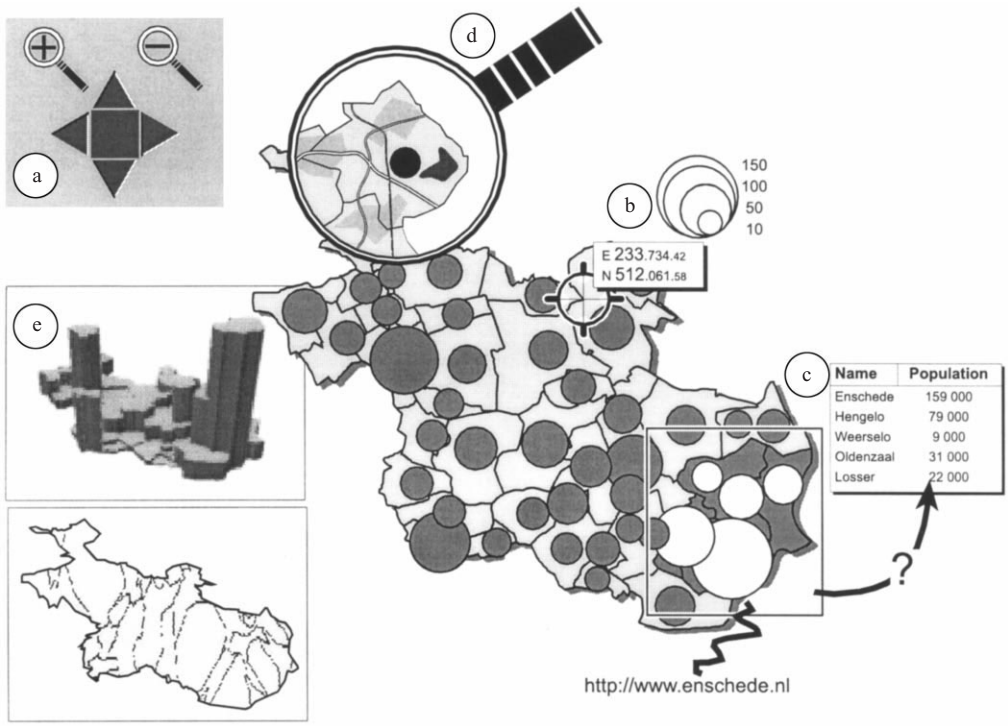

Figure 5
Exploratory cartographic functions: (a) Basic display; (b) Navigation and orientation; (c) Query; (d) Multi-scale; (e) Re-expression

user will be able to view and interact with the data in different windows, all representing related aspects of the data. These views do not necessarily contain maps; video, sound, text, etc. can all be included. Clicking an object in a particular view will show its geospatial relations to other objects or representations in all the other views (see Fig. 6).

5.7 *Animation*

Maps often represent complex processes that can be well expressed by animation. Animation can be used for temporal as well as non-temporal changes in geospatial data. Aspects to be solved are related to the interface (user interaction—navigation) and the legend. Since the exploratory user will also be the creator of the animation, he or she should be able to influence the flow of the animation.

6. *Visualization Environments*

Most cartographic components of geographical information systems can only handle one or two of the functions described in the previous section. Scientific visualization software might have the most functions available, but it lacks the typical geo-referencing options needed to deal with geospatial data. However, operational experimental exploratory visualization environments do exist. Examples are web-based Decartes (Andrienko and Andrienko 1999) and Cartographic Data Visualizer (CDV) (Dykes 1997, Dykes 1998). As an example CDV will be discussed since it is available free as a result of academic research (http://www.geog.le.ac.uk/jad7/cdv/). Both packages primarily aim at the visualization of socio-economic census data. Fig. 6 shows a typical view on a CDV-session. The program requires a base map (administrative boundaries) and statistical data. From the menu on the left the user can select active variables and visualizations. Options are choropleth maps (or polygon maps), proportional circle maps, cartograms, and non-cartographic visualization such as dotplots and scatter diagrams. When creating these different views on the data they will all be linked together. Clicking an area in a map or a symbol in a diagram will highlight the corresponding area/symbol in the other views. The user has several options available to choose the layout of each of the views. One can also execute basic calculations to create new variables or classify the data according to different methods. CDV is by the nature

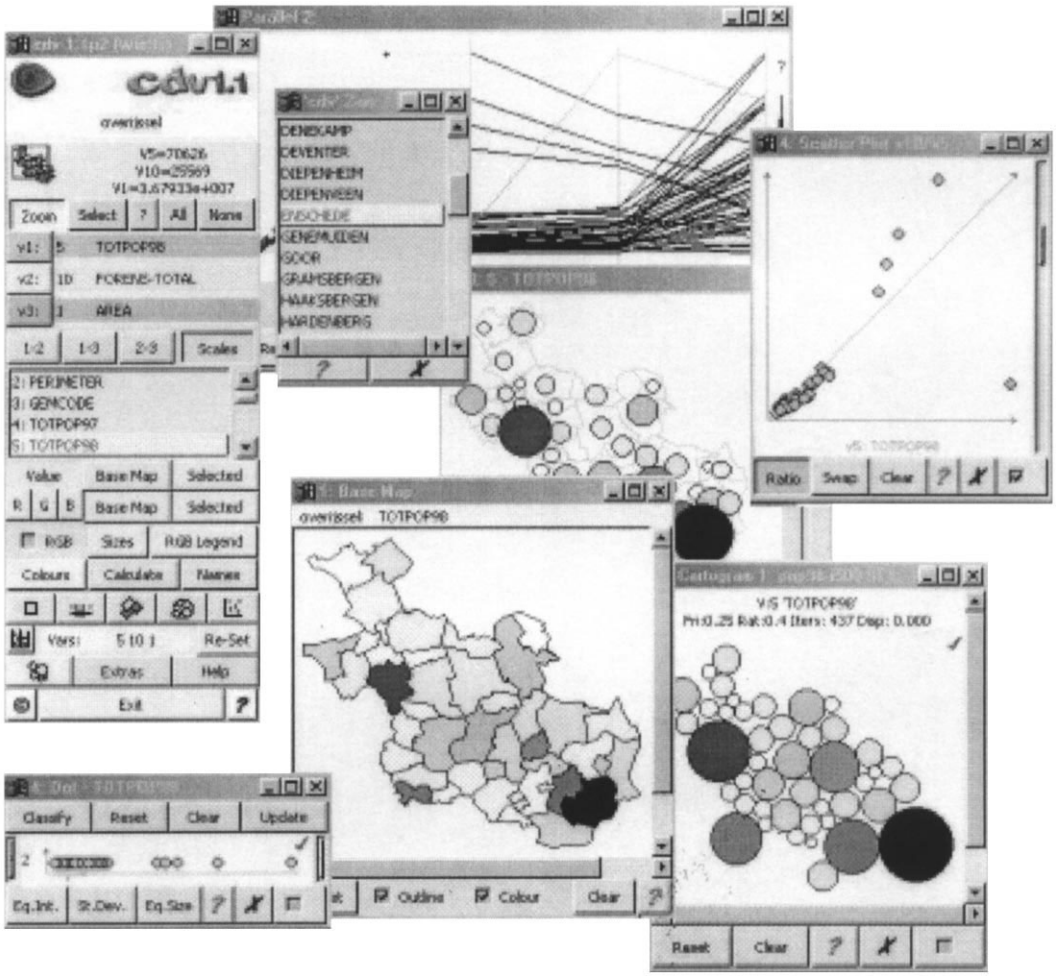

Figure 6
A typical view of a CDV session

of its functionality, capable of showing geospatial patterns that would not be apparent when single maps or diagrams would be viewed.

7. *Conclusions*

Packages with the functionality similar to CDV offer an advanced cartographic visualization environment. However, remembering 'How do I say what to whom, and is it effective' leaves the question open if it indeed works. From a technical point of view it certainly does, but can the user handle all those linked views. It is argued that if users can change their perspective on the data through selection and transformation, as well as alternative visualizations, meaningful relations among data variables are more likely to be revealed. For some products, this is easy to determine its usability, for others it is quite a job. To understand how maps work is not easy (Wood 1992), especially as we change into a demand-driven mapping environment. In the supply-driven environment, it was known what the map should tell and who the customers were. Today cartographers deal partly with providing facilitating tools to visualize geospatial data. How the tools are used remains, for the moment at least, unknown. These trends will only increase as more and more maps are produced via the World Wide Web. One has to find out why someone wants to make a map, in order to be able to judge the future tools that need to be developed.

See also: Cartography; Cartography, History of; Cognitive Maps; Dynamic Mapping in Geography; Ethnic Conflict, Geography of; Geographic Information Systems; Planetary Cartography; Tactile Maps in Geography; Thematic Maps in Geography

Bibliography

Andrienko G L, Andrienko N V 1999 Interactive maps for visual data exploration. *International Journal for Geographic Information Sciences* **13**: 355–74

DiBiase D 1990 Visualization in earth sciences. *Earth & Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences* **59**: 13–18

Dykes, J 1997 Exploring spatial data representation with dynamic graphics. *Computers and Geosciences* **23**: 345–70

Dykes J A 1998 Cartographic visualization: exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv. *The Statistician* **17**: 485–97

Hearnshaw H M, Unwin D J (ed) 1994 *Visualization in Geographical Information System*. Wiley, Chichester, UK

Keller P R, Keller M M 1993 *Visual Cues, Practical Data Visualization*. IEEE Press, Piscataway, NJ

Kraak M-J, Ormeling F J 1996 *Cartography, the Visualization of Spatial Data*. Addison Wesley Longman, London

Longley P, Goodchild M, Maguire D M, Rhind D (eds.) 1999 *Geographical Information Systems: Principles, Techniques, Applications and Management*. Wiley, New York

MacEachren A M 1995 *How Maps Work: Representation, Visualization and Design*. Guilford Press, New York

MacEachren A M, Taylor D R F (eds.) 1994 *Visualization in Modern Cartography*. Pergamon Press, London

McCormick B H, DeFanti T A, Brown M D 1987 Visualization in Scientific Computing. *IEEE Computer Graphics* **7**: 69

Monmonier, M 1989 Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis* **21**: 81–4

Morrison J L 1997 Topographic mapping for the twenty first century. In: Rhind D (ed.) *Framework of the World*. Geoinformation International, Cambridge, UK, pp. 14–27

Robinson A H, Morrison J L, Muehrcke P C, Kimerling A J, Guptill S C 1995 *Elements of Cartography*. Wiley, New York

Wood D, Fels J 1992 *The Power of Maps*. Routledge, London

M. J. Kraak

Cartography

The International Cartographic Association defines cartography as 'the discipline dealing with the conception, production, dissemination and study of maps.' It goes on to say, 'A map is a symbolized image of geographic reality, representing selected features or characteristics, resulting from the creative efforts of cartographers, and is designed for use when spatial relationships are of special relevance' (International Cartographic Association 1995). Cartography, along with disciplines such as geodesy, surveying, aerial photogrammetry, and satellite remote sensing, is a component of the mapping sciences. It is closely allied with geographic information systems and geographic information science.

Cartography is most often found in geography departments in colleges and universities in the USA and Canada. It is commonly an independent department in Europe and other parts of the world. Commercial enterprises, nonprofit entities, and government agencies that produce maps generally have large cartography departments that are central to their operations.

The term 'research' within the field of cartography has at least two distinct meanings. One is the systematic gathering of information from a variety of sources for compilation into a coherent map. The other is the quest for discovery of knowledge about maps and the processes associated with them. The former definition is closely associated with commercial and nonprofit map production and the latter with academic cartography. The subject-matter of academic cartographic research ranges from historical studies of the cultural, technical, and political context of maps to the changing processes of production to the ways in which maps are used in wayfinding and in the production of knowledge. The relationship between people and maps, including symbol perception, the development of map-reading abilities, and cognitive processes in map use, have been of special interest in cartography in the late twentieth century, as have the processes by which maps are produced by computer.

Although maps are most closely associated with geography, practitioners in many disciplines and professions use the map as a device for recording and preserving information, as a tool in research, and as pedagogical illustration. As computers are used more for the process of mapping, it is no longer as strictly the domain of professional cartographers as it once was. Virtually anyone with a computer and appropriate software can produce a map. Whether this democratizing of mapping is a positive or negative development is not altogether clear. The products of nonspecialists range from misleading to highly insightful and creative. Maps are now far more widespread and are available in a more timely way to fit immediate needs than in the past, and they are becoming less constrained by convention.

Cartography and its products have a profound effect on human thinking and behavior. The map is a metaphor for the real world and is often the model that shapes it. Existing hills and valleys, rivers, and political boundaries are recorded on maps, but decisions about surface excavation, river redirection, or political redefinitions are marked on maps and become reality. The terms 'map' and 'mapping' have become common metaphors in everyday language; 'on the map' signifies importance, and 'mapping a strategy' implies attentiveness and purpose in the planning of actions.

Cartography has responded to changing social, intellectual, and technological conditions throughout history, and the permeation of computers into all aspects of life in the late twentieth century has resulted in profound changes. Cartography has transformed from a data-poor to a data-rich field, as information

can now be gathered using aerial photography, global positioning systems, and satellite imaging. Inexpensive computer storage and the Internet have made it possible to share and receive massive amounts of data. Professional cartographers are now heavily involved in information management and in creating mechanisms for shared use of data. Scientific visualization in cartography, which implies the use of maps for discovery of knowledge, has gained considerable attention in recent years as a result of the ubiquity of computer usage in spatial representation.

1. Major Cartographic Concepts

Mapping Earth's surface involves the concepts of scale, projection, spatial relationships, generalization, symbolization and data modeling, and categories of maps. Some level of understanding of each of these concepts is inherent in every instance of the employment of a map.

1.1 Scale

The map is a scaled model of Earth's surface (or of the surface of some other planetary body), and the reduced size is one of its most useful features. Earth itself is too large for ready comprehension.

There are three types of scale expression commonly used on conventional maps: verbal, representative fraction (RF), and graphic (Fig. 1) (Robinson et al. 1995). Although each can refer either to distance (linear scale) or area (areal scale), linear scale is by far the more commonly used. Areal scale is generally indicated only if area is consistently scaled over the entire map. Since distance cannot be consistently scaled, no such constraint is associated with the appearance of a linear scale on a map.

Explicit scale expression is not the only indication of scale on maps (Eastman 1981, Brewer 1990). Map content and design give scale clues as well. Double-line streets, for example, immediately convey that the map shows a small area; a shaded relief background and sparse linework indicate that the map covers a large area. Scale clues are becoming increasingly important, as the user of an animated map cannot always use an explicit expression of scale while observing the map.

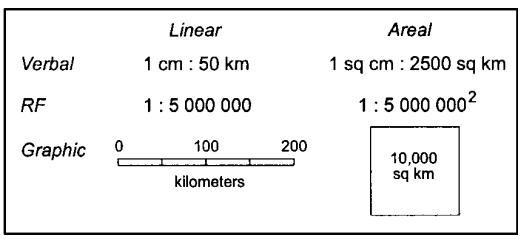

Figure 1
The three types of scale expression, illustrated in both linear and areal form

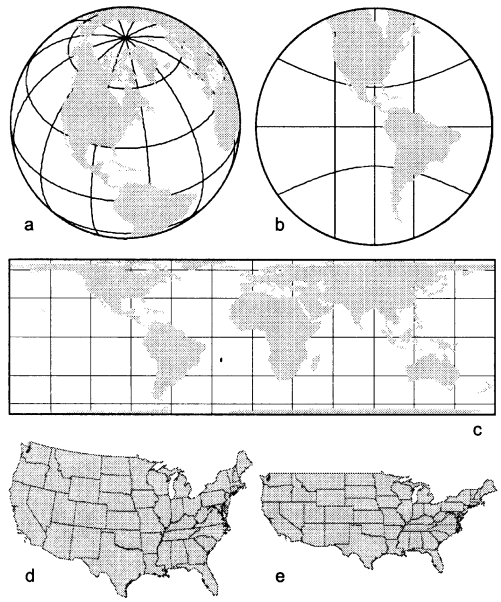

Figure 2
Five map projections showing a variety of shapes and sizes of land masses when different projection systems are applied: (a) orthographic, the projection that looks like a view of a globe, (b) gnomonic, a very distorted projection but the only one on which all great circles are straight, (c) cylindrical equal area, which preserves correct sizes of all areas but has extreme changes in linear scale, (d) Alber's, which is equal area and is commonly used for the USA, and (e) plane chart, which represents all degrees of latitude and longitude at the same length regardless of their length on Earth's surface

1.2 Projection

The term projection refers to the system by which points on a spherical or spheroidal surface are assigned to points on a flat surface. Every point on the sphere or spheroid has a corresponding point on the flat map. Perhaps the most fundamental observation about map projections is that linear scale cannot be the same everywhere; in other words, map projections are geometrically distorted representations, in contrast to globes, which maintain a shape very close to that of the planet. There are many map projections in use and the five represented in Fig. 2 illustrate a variety of shapes and relative sizes of landmasses that can result when different systems of projection are applied.

Although it is necessary to tolerate distortion when projecting Earth onto a flat surface, the distortion in map projections is systematic and in some cases highly useful. Every rectangular projection of the world, for example, has one or two straight lines on which there is no distortion and that distortion increases at right angles away from the line(s) of no distortion. Although large-area maps have major visible distortion, maps of small areas, say a city or small region have so little distortion that it has traditionally been of interest only for the most exacting measurements.

The development of geographic information systems has resulted in renewed interest in map projections because many maps have been transformed into digital files and data have to be 'unprojected' (transformed from flat map cartesian coordinates to longitude and latitude) to be compatible with other data. Even the limited amounts by which map projections affect location within small areas will affect the compatibility of multiple data sets.

Despite the impossibility of maintaining consistent linear scale on a map projection, there are two other properties that may be maintained. When, at every point on the map, maximum linear scale distortion is compensated by opposite distortion in the opposite direction, area scale is maintained. If the linear scale change is the same in every direction at each point, angles at all points will be maintained. The terms for these two categories of projection are equal-area (or equivalent), and conformal (or orthomorphic), respectively. Many projections have neither compensating nor equal changes of linear scale in opposite directions; these projections are neither equal-area nor conformal.

Some projections have special properties that make them useful for specific purposes. The Mercator projection (see *Thematic Maps in Geography*), when centered on the Equator, shows all loxodromes (or rhumb lines, i.e., lines of constant geographic direction) as straight lines. This property has been important to navigators. It was a profound development because most loxodromes on Earth (and on globes) are spirals, and the Mercator projection simplified route planning. The projection suffers from extreme deformation of area, and, like many other projections, has been used inappropriately. The Mercator maps hanging on classroom walls have led many students to think of Africa and South America as a relatively small land bodies relative to the greatly exaggerated high-latitude landmasses of Greenland and Russia.

There are over 100 named projections (see Snyder 1993), and cartographers try to choose one that is appropriate for the mapping purpose at hand. The discrete naming of projections, while providing convenient reference, obscures the fact that there are an infinite number of ways to represent Earth on flat paper. Computer construction of projections may result in revised nomenclature and increasingly tailored projections (Laskowski 1997).

1.3 Spatial Relationships

Maps show spatial relationships in a readily comprehended form. Metric relationships include distance, direction (angle), and area; topological relationships include such properties as connected to, inside, and outside. The concept of spatial relationship also includes spatial (or geometric) form; an entity may be considered to have a footprint of a point, line, or area, and if quantities are associated, the entity may be thought of as a pole, ribbon, or volume. Spatial form also includes shape (e.g., Long Island vs. Martha's Vineyard) and distribution ('more of something here than there'). The relevant spatial relationships vary with the type of map and with how they are used. Street map users want to know the street on which a feature is located and the turns and distances to reach it. Users of a world population map want to know where population is concentrated and where there are relatively empty areas.

Like the general concept of scale, spatial relationship is often taken for granted by the practiced map user. The reason for using a map is not the spatial relationship itself, but what it means. A 'hot-spot' on a map showing cancer rates suggests the need for further study and possible action; the spatial relation on the map is simply the trigger. Whether a street, population, or cancer map, it is the spatial relationships that make it a useful device. Without them, there would be no need for a map.

1.4 Generalization

As scaled representations of Earth or portions thereof, every map is generalized. Lines have fewer details in them than present in the feature on Earth's surface, and only selected features are shown. When quantities are depicted, they are usually shown by category rather than by exact value. Generalization is highly useful in focusing the map on important information instead of cluttering it with overabundant detail. It also means that some important features may not be there, not because they do not exist but because they have been generalized out of the map.

Despite the obvious necessity for generalization in making representations that are literally thousands and millions of times smaller than the original territory, it is not always understood. The small street maps of major cities that appear on road maps may be assumed by users to include every street, and a surface feature that falls between the contours of an elevation map may surprise and confuse the hiker. The failure of a map due to generalization is one of the reasons that there are very many different maps of the same area at different scales and with different selections of features. The larger the scale (i.e., the larger the representation of a given area) the less generalization is necessary and the more useful it is likely to be for

finding details. The smaller the scale (the smaller the representation of a given area), the more likely it is that broad patterns will be visible because of the lack of obfuscating detail.

The relationship between scale and generalization is not an exact one. Some maps have much more detail than others of the same physical scale. The term generally applied to the degree of detail relative to size of features on Earth's surface is resolution (Tobler 1988). A map of the USA showing data by state (50 units total) has considerably coarser resolution than a map showing data by county (3,000+ units total) whether the latter map is larger, smaller, or the same size as the former.

1.5 Symbolization and Data Modeling

The very act of representing a feature with a symbol is a form of generalization. The symbol is not the original, and every map is a set of symbols. Mapmakers have been highly creative in employing symbols.

Geometrically, mapmakers can employ only points, lines, and areas as symbols on flat maps, though they may be used in such a way as to simulate the illusion of a third dimension. The use of three-dimensional media, such as molded plastic, allows the explicit use of volume as a symbol. The development of virtual reality allows the illusion not only of a third dimension but of the intersection of surfaces. Regardless of medium, the geometry of symbols generally reflects the way in which humans think of the object relative to the scale of the map. For example, a city is a large area on a street map but a point on a world map. A river on a local, large-scale engineering map is an area, but the same river appears as a line on a smaller-scale regional map. An important exception is when numerosity of symbols represents quantity. A dot map of population, for example, uses the dot to represent a quantity of persons, not an individual feature. In effect, then, dots on the map are being used to represent conceptual volume.

There are a limited number of characteristics of symbols, called 'visual variables,' which can be used to represent the characteristics of the feature represented (Bertin 1983). Darkness–lightness of the symbol and size are the most commonly used to represent quantity. Shape, color hue, and orientation most often represent differences in kind. Patterning in area symbols may also be used to distinguish characteristics of features.

The modeling of geographic space in digital files introduces a different form of representation from that on paper or similar physical medium (Peuquet 1991). Geometric concepts of point, line, area, and volume are still valid, but there is no scaled physical representation or visual symbols associated with features. Rather, numbers represent location, and codes or descriptive terms represent what exists at a location. These digital files can be thought of as data models as opposed to a physical model. They can almost always be converted into a map, and they are usually much more flexible. A digital elevation model, for example, i.e., a file of elevation values at points, can be converted into any of the maps in Fig. 3.

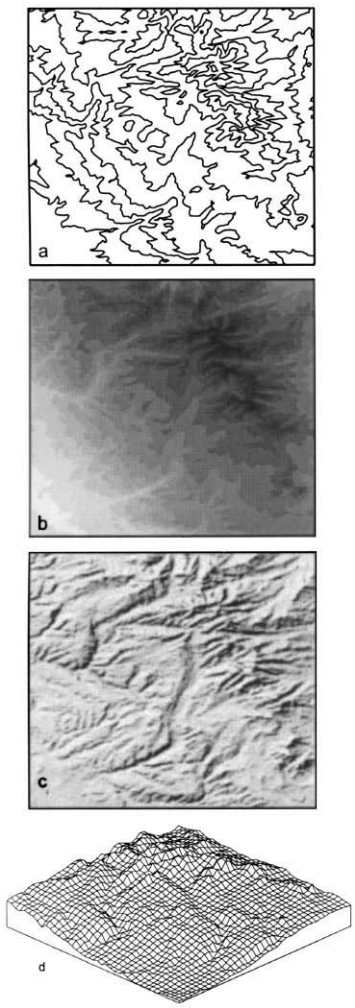

Figure 3
Four ways of representing a set of elevation data; lines are used in the contour map (a) and three-dimensional diagram (d), whereas area symbols are used in the layer shaded map (b) and in the shaded relief map (c)

1.6 Categories of Maps

Maps can be divided into two very broad categories: general reference and thematic. A general reference

Figure 4
An example of a general reference map, extracted and adapted from *The World Factbook*, CIA, 1999 ⟨http://www.cia.gov/cia/publications/factbook/⟩ (original is in color)

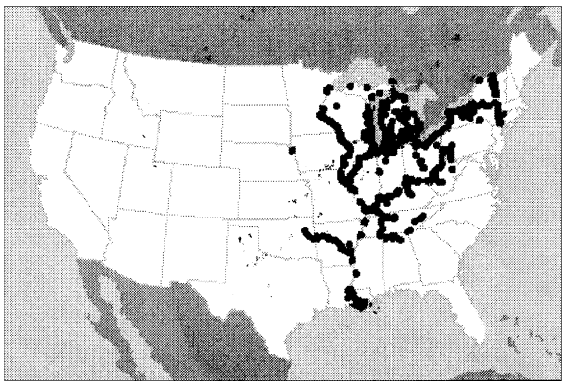

Figure 5
Zebra Mussels in the USA, an example of a thematic map, extracted and adapted from *National Atlas of the United States*, US Geological Survey, ⟨http://www.nationalatlas.gov/⟩ (original is in color)

map (Fig. 4) shows a variety of individual features and is generally used to find specific places, such as cities, rivers, or regions. The general reference map is selective of features in any category; only 'important' cities, rivers, or regions are identified, for example. Although lettering on the map often stands out visually, there is limited visual hierarchy among the features of the map, as all the features must be distinguishable from one another. A thematic map (Fig. 5) shows a distribution of one or more phenomena and is used to visualize spatial form. The thematic map is generally based on all the known instances of the phenomenon even though none may be shown individually. The thematic map generally has clear visual hierarchy, with the symbols representing the distribution standing out the most.

A simple twofold classification is limited, and sometimes a third is added: navigation maps. Others consider the general reference and thematic categories to be opposite ends of a scale, with individual maps fitting somewhere between. Navigation maps, such as a street guides, are near the general reference end because many different specific items are represented. They have certain thematic qualities, however, because of the focus on the features relevant to finding one's way. However, inventory maps, such as a map showing the residence of every patient with a certain disease, are close to the thematic end of the scale. They share the general reference characteristic that individual items are shown.

There is good argument that the general reference–thematic classification is more fitting of use than of maps themselves. One can look at a thematic map to see how large one political unit is relative to another, which is a function generally associated with the term 'general reference.' Likewise, one can look at a general reference map and observe overall pattern of drainage or evidence of population, a use generally associated with the term 'thematic' (Robinson and Petchenik 1976).

2. *The History of Maps and Cartography*

The history of cartography begins earlier than recorded time and in all likelihood had its origins in gestures and ephemeral marks in the soil (Godlewska 1997). The development of civilizations and of cartographic practice were probably highly intertwined, and at least some cartographic artifacts remain from ancient Mesopotamia and Egypt, including a clay tablet from roughly 3800 BC with mountains and a river. Extant land ownership maps date to 2000 BC. The development of grid systems, associated with less subjective and more scientific approaches to mapping, was reflected in written materials in China in the third century AD, and the earlier development of paper in China provided a flexible and highly portable physical medium for drawing maps (Thrower 1996).

The concept of the world as a sphere was basic to the development of logical representation of large areas of the world. Measurements make sense only when the curvature of the planet is taken into account. The idea seems to have had its origins with the Pythagoreans and was supported in work by Plato and Aristotle. Eratosthenes (276–196 BC), who served as the librarian at Alexandria, used the concept of the spherical Earth and the relationship between noon sun angles at different locations to measure the size of the planet. Because the exact length of a stade, the unit of measure in use at the time, is not known, it is difficult to know exactly how accurate his estimate was, but it seems to have been remarkably close and a better estimate than several that followed as scholars attempted to correct the figure. A successor in his

position in the second century AD, Claudius Ptolemy, compiled the seminal tome entitled *Geographia* that contained a guide to making maps, including instructions for several map projections and the latitude-longitude location of about 8,000 places.

Mapmaking in China and South Asia is characterized by remarkable developments in roughly the same time period as the accomplishments of the ancient Greeks. There is reference to a map on silk in the third century BC, and maps were used by the military and administrators in the Han Dynasty. In later centuries, the first printed map appeared in China well before the development of printing in the West (approximately 1153 vs. 1472).

European cartography during the Middle Ages was characterized by maps that are highly schematic by modern standards but are important cultural documents of their time, having been associated with religious institutions and beliefs. One schema was a circular map with east at the top showing Asia as the upper section of the circle, separated from Europe in the lower left and Africa in the lower right by the Don River and the Nile. Europe and Africa were separated by the Mediterranean Sea. These maps are known as T-O maps, the rivers and Mediterranean forming a rough T shape and the oceans surrounding the landmasses forming the O. The second schema was a circle with north at the top and, within the circle, temperature zones that included the torrid one along the equator, bordered by the temperate zones, which in turn were bordered by frigid zones at the poles. Route maps of the time period carried far more detail than these schematic world maps, but the most remarkable detail and the most scientific approach was found in the portolan charts. These maps were used in navigation, primarily in the Mediterranean Sea, and showed intricate coastlines with hundreds of identified coastal locations. At several locations within the water are compass roses with direction lines extending from them.

Cartography in the Islamic world during the European Middle Ages, while varied, included continuation of the Greek scientific tradition. The scientific instrument called the astrolabe, a device for measuring angles to stars, was important in finding the way to Mecca. In the early 1400s, Ptolemy's *Geographia*, which had been preserved in Byzantium, reached Italy via refugees from the invading Turks. Translated into Latin, the work had an extraordinary influence on subsequent development of cartography. Systematic projection of the spherical earth to the flat map followed Ptolemaic instructions for well over a century (Thrower 1996).

Accurate mapping depends on the ability to measure both latitude and longitude accurately. Latitude, which could be measured by angles to the sun and other heavenly bodies at specific times of the day was far easier to measure than longitude, which depends on differences in time from one location to another. In a competition sponsored by the British government, John Harrison in the later half of the eighteenth century produced a chronometer that could retain accurate time at sea (Sobel 1995). It was a turning point in the accuracy of maps covering large regions of the world.

The rise of detailed scientific mapping of Earth's surface in the form of topographic mapping owes much to the Cassini family, whose influence on cartography began with the appointment of astronomer Giovanni Domenico Cassini to the Académie Royale in Paris, which along with the Royal Society of London was concerned with mapping and other scientific problems. Under his direction, maps recorded only locations that were determined by astronomical methods, and detailed measurement of a degree of latitude was undertaken at various locations to test the theory of the earth as a prolate spheroid, which affects the placement of locations on flat maps. The detailed mapping of France was continued under three more generations of the Cassinis (Jacque, César-François, and Jean) and was completed in 1793.

In the eighteenth and early nineteenth centuries, one of the most important developments in cartography was thematic mapping (Robinson 1982). With the practice of accurate base maps established, it was feasible to focus attention on the spatial arrangement of various phenomena of interest such as elevation, population, and winds. Symbol systems such as isolines, dots, and the choropleth method came into being and are commonplace in modern mapping.

Changes in cartography in the twentieth century included the use of aerial photographs in topographic mapping and changes in production technologies. The latter half of the century saw the development of satellite sensing for gathering data for maps, the deployment of global positioning systems (constellations of satellites sending signals from which latitude and longitude are determined), and the infusion of computers into mapping including the development of geographic information systems and Internet mapping. The variety of thematic mapping methods has recently undergone rapid change as attention has focused on use of the computer to overlay multiple phenomena, produce animated maps, and develop sophisticated visualizations of spatial phenomena (Thrower 1996).

The richness of cartographic history throughout the world and over many centuries is reflected in *The History of Cartography*, a monumental multi-volume series underway as this Encyclopedia is being written (Harley and Woodward 1987). The project itself is an important intellectual event in cartographic history.

See also: Cartographic Visualization; Cartography, History of; Cognitive Maps; Dynamic Mapping in Geography; Ethnic Conflict, Geography of; Geographic Information Systems; Planetary Cartography;

Tactile Maps in Geography; Thematic Maps in Geography

Bibliography

Bertin J 1983 *The Semiology of Graphics* [English translation by Berg W J]. University of Wisconsin Press, Madison, WI
Brewer C A 1990 The effect of color on the perception of map scale. In: Everett W (ed.) *Student Honors Competition Winning Papers*, Association of American Geographers, Cartography Specialty Group, Toronto, Canada
Eastman J R 1981 The perception of scale change in small-scale map series. *The American Cartographer* **8**(1): 5–21
Godlewska A 1997 The idea of the map. In: Hanson S (ed.) *Ten Geographic Ideas that Changed the World*. Rutgers University Press, New Brunswick, NJ
Harley J B, Woodward D (eds.) 1987 *The History of Cartography*. University of Chicago Press, Chicago, IL
International Cartographic Association 1995 The Definition of Cartography, a statement adopted by the 10th General Assembly, Barcelona, Spain, September 3, http://www.icaci.org
Laskowski P 1997 *The Distortion Spectrum*, monograph 50. *Cartographica* **34**(3): 67–95
Peuquet D 1991 Methods for structuring digital cartographic data in a personal computer environment. In: Taylor D R F (ed.) *Geographic Information Systems: The Microcomputer and Modern Cartography*. Pergamon Press, Oxford, UK
Robinson A H 1982 *Early Thematic Mapping in the History of Cartography*. University of Chicago Press, Chicago, IL
Robinson A H, Petchenik B B 1976 *The Nature of Maps: Essays toward Understanding Maps and Mapping*. University of Chicago Press, Chicago
Robinson A H, Morrison J L, Muehrcke P C, Kimerling A J, Guptill S C 1995 *Elements of Cartography*, 6th edn. Wiley, New York
Snyder J P 1993 *Flattening the Earth: Two Thousand Years of Map Projections*. University of Chicago Press, Chicago, IL
Sobel D 1995 *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*. Walker, New York
Thrower N J W 1996 *Maps and Civilization: Cartography in Culture and Society*. University of Chicago Press, Chicago, IL
Tobler W 1988 Resolution, resampling, and all that. In: Mounsey H, Tomlinson R (eds.) *Building Databases for Global Science*. Taylor & Francis, London, pp. 129–37

J. M. Olson

Cartography, History of

Recent expansion in both the usages and definitions of the noun 'map' and the verb 'to map' have had consequences for the history of cartography as a field. Its focus is no longer biased almost exclusively towards flat, formally-constructed, professionally-produced, artifactual maps of the earth's terrestrial and marine surfaces, made in the European tradition. The increasingly adopted operational definition of 'maps' as 'graphic representations that facilitate a spatial understanding of things, concepts, conditions, processes, or events in the human world' (Harley and Woodward 1987) has increased the range of artifacts to include celestial maps and cosmographical maps of supposed worlds. It also embraces performance cartographies; spatial representation incorporated in gesture, procession, ritual, theatre, and dance. The linked constraints of paper and flatness have been abandoned to include many and mixed media and three-dimensional expressions of these, such as buildings, whole settlements, and massive earth sculptures.

1. The Field

The history of cartography as conceived in the 1940s in the first modern review of the field has been transformed (Bagrow 1951). Approximately 90 percent of that text focused on maps made in the Classical and European worlds. Although written primarily for map collectors, and not for scientists and scholars, the chronological sequencing of chapters and narrow cultural range gave the unintended impression of an evolutionary development. The field is no longer Eurocentric but encompasses maps and mapmaking in all cultures. An important consequence of this is that the historical development of mapmaking is no longer seen as unilinear. Establishing ultimate origins in prehistory may never be possible but they were certainly multiple, probably numerous, and almost certainly geographically widespread. Another consequence of the shift to a multicultural perspective is that mapmaking is no longer seen to be exclusively a specialist or even semispecialist activity. Though not as ubiquitous as speech, it has long been a widespread vernacular skill. Transformation of the field has also reduced the formerly excessive emphasis given to printed maps based either on surveys or systematic compilation and plotted on mathematically-generated graticules. Categories of what were once considered 'less accurate' derived maps are now accorded appropriate attention. These include early manuscript maps, many thematic maps, and most media maps including the cartographoons of political cartoonists.

The longer timescale, adoption of multicultural perspective, and recognition given to additional categories of maps have together increased actual and potential links between the history of cartography and other fields. Those with the histories of art, science, and technology are fairly obvious. Even so, these have not been developed as much as might have been expected. Others are beginning to be forged. Cultural anthropology is likely to be significant in relation to vernacular cartography (see, for example, Nabokov in Lewis 1998). Likewise, religious studies should contribute to the understanding of cosmographical maps and archaeoastronomy to resolving claims that some rock art incorporates celestial maps. Archaeologists of later periods are beginning to make use of

maps made in or derived from the cultures being investigated (see, for example, Waselkov in Lewis 1998). The historical role of maps in relation to other modes for communicating information spatially is at last beginning to be investigated (see, for example, Fletcher 1995 and Pearce in Lewis 1998). Links with the cognitive sciences (including cognitive archaeology) will be particularly important in furthering the understanding of map use and in helping to reconstruct the early prehistory of cartography. The ultimate and most elusive question is likely to be which emerged first as the language of space among *Homo sapiens*; those currently universal features of speech referred to as spatial deixis, schematic graphics (including artifactual maps), or nongraphics (including performance cartographies)? (Twyman 1982). In trying to answer this question, links with comparative and evolutionary linguistics will be vital. Whether actual or potential, all these links will be mutually beneficial.

The substance, scope, methods, and intentions of the transformed field are being revealed by an ongoing publishing venture; the multivolume *The History of Cartography* (Harley and Woodward 1987, 1992, 1994, Woodward and Lewis 1998; with further volumes scheduled for the European Renaissance, European Enlightenment, Nineteenth Century, and Twentieth Century). Whereas, for example, Bagrow (1951) covered 'Maps of primitive peoples' in four pages, *The History of Cartography* devotes a whole volume of more than 650 pages to maps in 'traditional societies' (Woodward and Lewis 1998). Likewise, Asian, including Islamic, cartography to which Bagrow devoted 21 pages is the subject of two volumes, together containing more than 1500 pages (Harley and Woodward 1992, 1994). Whereas the index to Bagrow is approximately 95 percent personal names, and virtually devoid of topics, the volume indexes of *The History of Cartography* reveal a surprising diversity of topical content including, for example, 'accuracy,' 'alphabets,' 'ancestors,' 'animal language,' 'animism,' 'anthropology,' and 'architecture,' as well as the much less surprising 'altitude,' 'arcs,' 'astrolabes,' 'axes,' 'atlases,' and 'azimuths.'

The defining criteria of the history of cartography as announced by the editors of *The History of Cartography* in the mid-1980s (Harley and Woodward 1987) are now agreed to by a majority of those active in the field:

(a) 'acceptance of a catholic definition of "map"';

(b) 'commitment to a discussion of the manifold processes that have contributed to the form and content of individual maps';

(c) 'recognition that the primary function of cartography is ultimately related to the historically unique mental ability of map-using peoples to store, articulate, and communicate concepts and facts that have a spatial dimension';

(d) 'belief that, since cartography is nothing if not a perspective on the world, a general history of cartography ought to lay the foundations, at the very least, for a world view of its own growth.'

To these, many working in the field would now add:

(e) 'a commitment to explore the roles of maps within societies, involving reconstructing their meaning.'

These defining criteria differentiate the 'history of cartography' from 'historical cartography' the practice of compiling maps in the present from historical data. Latterly, however, 'map history' has emerged as an alternative term for 'history of cartography.' This is regrettable. The neologism 'cartographe' was introduced in 1839 for the study of early maps. The word was soon applied to the map field in general and was to appear in many European languages in the second half of the nineteenth century. It is now associated worldwide with institutions, organizations, and activities involved with the making, use, and preservation of plans, charts, and globes as well as maps in the strict sense of representations of the earth's land surface at medium to small scales.

2. *Historiography*

Notwithstanding the long history and near universality of maps, mapmaking, and map use, the history of cartography is a youthful field. Although there was a long but sporadic tradition of antiquaries, collectors, and some mapmakers chronicling the subject, the field only began to emerge after 1850 with the institutionalization of geography as an academic field especially in Germany. Even then it lacked identity, being treated as part of the wider history of geographical discovery and exploration. Its beginning as a modern subject began in 1935 with the founding of *Imago Mundi*, still the only international journal devoted exclusively to the field, as distinct from map collecting. Since then, the 'emergence of cartography as an independent and practical discipline' has provided 'new theoretical frameworks as well as a reinforced *raison d'être* for the study of cartographic history' (Harley and Woodward 1987).

Many of the material resources on which the emergence of the new history of cartography were based were already in place, including archives, map libraries, and private collections, but they were to become more accessible (Skelton 1972). Epistemological and methodological developments were consequences of the wider ferment of ideas and institutional changes that followed World War II. Increasing prosperity and better and more specialist librarians and bibliographers were in part responsible for the appearance of major printed catalogues of collections and regional cartobibliographies, albeit heavily biased towards Europe and North America. New printing technologies made possible the reprinting of earlier obscure but significant publications. Among these

Acta Cartographica, Vols. 1–27 (Theatrum Orbis Terrarum, Amsterdam, 1967–81) was particularly important. Most important among these publishing initiatives was the quality reproduction, often in color, of thousands of maps hitherto only accessible to a few.

The field became somewhat less dominated by geographers and historians and began to attract some of the new generation of cartographers, some of whom had trained in the military, and were interested in modern maps and in the contexts and consequences of their use. In 1972 The International Cartographic Association created a Standing Commission on the History of Cartography from an existing working group that was already preparing a handbook of cartographic terms in use before 1900 (Wallis and Robinson 1987). An international conference on the History of Cartography now meets biennially, usually preceded by a meeting of the International Society for the Curators of Early Maps. Map collectors are an important interest group, providing the field with financial and material support. The International Map Collectors' Society meets annually.

By 1994 there were more than 500 self-designated historians of cartography in the world; unequally distributed geographically with approximately 40 percent resident in continental Europe, 25 percent in North America, and 10 percent in the United Kingdom. Research interests were equally skewed with, for example, 66 respondents registering an interest in cartobibliography and 13 in projections, but only one in mining maps, two in cartouches, three in map accuracy, four in Hebrew maps, and five in tithe maps. Likewise, 19 registered an interest in Gerardus Mercator and 13 in Claudius Ptolemy but eminent cartographers from other cultures and periods were far less prominent. For example, al-Idrīsī, Timothy Pont, and Thomas Hutchins had only one registrant each. Nevertheless, the earlier collector-led, dilettante, aesthetically motivated, Eurocentric, and somewhat insular preoccupation with Renaissance and Enlightenment printed maps and mapmakers has undergone a relative decline and interest in such topics as mathematics in cartography, techniques of map production, map publishing, and surveying is increasing. Listed under the umbrella of 'Theory and Methodology' are such interests as 'cartography and cultural theory,' 'construction of knowledge space,' 'politics of projections and maps,' and 'semiotics of prehistorical maps.' (Lowenthal 1998).

In anticipating future directions of the field, the most serious limiting factors are the paucity of specialist undergraduate courses and limited opportunities for formal graduate training. For the most part, those active in the field continue to enter by chance and often in mid-career. They contribute to a rich mix of epistemologies, interests, and skills, but it could be argued that this is delaying the emergence of an overarching paradigm. Essentially, researchers do not operate within an agreed philosophy, concur on the theoretical focus of their work, or use accepted methodological procedures to solve problems identified within a theoretical framework. Conversely, most set out to investigate topics *de novo* and only rarely on the basis of what is already known. They add to knowledge but usually without extending theory. Given the paucity of graduate courses in the history of cartography, much depends on the few that do exist if the imbalance between empiricism and theory is to change. Meanwhile, it is important that the field accepts and integrates concepts, issues, and directions being developed at or beyond what many insiders now see as its periphery, often by persons who do not identify themselves with it. Important among these are the idea of maps as power, the distinction between intended and actual roles, an increased commitment to iconology, and re-examining the idea of value-free cartography.

3. *The History of Mapmaking and Maps*

The following outline reflects something of the enormous empirical additions made during the last quarter of the twentieth century but it is neither structured nor underpinned by a body of theory. The overriding concern has been to present a multicultural review from the earliest times to the beginning of the electronic age of geographical information systems and, in so doing, to reduce the Eurocentrism that has distorted most syntheses hitherto. Although primarily organizational, the structure does imply a working hypothesis for a developmental theory.

Cartography was originally a near-universal skill. It probably had prehistoric origins in every part of the then settled world, when map content was one or a mix of terrestrial, celestial, and cosmographical, and its roles were geographical, calendrical, and making sense of the unknown universe and linking it to the world of experience and report. The ability to read random patterns as maps was probably used in divinational practices. In so-called traditional societies, these types of cartographic activities survived until recent times. Indeed, they are still practiced in some. In Western societies some of the skills characteristic of this level of ability are still used, as when experiential sketch maps are made in communicating to others or when the scaleless topologically-structured maps provided by public transport systems are referred to.

Cartography underwent a qualitative shift in most of the world's early urban societies, where mapmaking slowly became an activity of elites but rarely a specialism. Although increasingly sophisticated and often revealing superb craftsmanship in producing artifacts of artistic quality, the basic roles of maps were still threefold, though with an increasing emphasis on terrestrial, made and used in the contexts of communicating, instructing, planning, building, administering, and record keeping. Very slowly, these

societies would seem to have developed a cartographic consciousness, although the concept of 'map' as a discrete category probably remained weak.

From the later urban societies of the Middle East and the eastern Mediterranian, two further and more or less independent cartographic traditions emerged: very quickly, towards the end of the first millennium AD, Islamic terrestrial and celestial mapping; and slowly, via more complicated stages, but ultimately with greater global impact, the Classical–Christian–European tradition. Both traditions used advanced instrumentation and mathematics to achieve greater precision than in other urban societies, arguably achieved higher artistic and aesthetic standards than elsewhere, and used maps very effectively in the interests of religious, scientific, and ruling elites.

Ultimately, the European tradition achieved most, surpassing that of Islam and, without knowing it until much later, the earlier peak of cartographic achievement reached in China. The printing of maps disseminated spatially-organized information more effectively than ever before and, together with education, increasing literacy, and the mass media raised cartographic consciousness to an unprecedented level. Maps and atlases became trade items. Formal ground survey and systematically conducted censuses and inventories of many kinds replaced judicious compilation of randomly assembled intelligence as the basic sources of information. Nation-states began to map their own territories and, later on, their colonial possessions. Cosmographical mapping virtually ceased but new types of maps emerged, in particular thematic maps and large-scale national topographic map series.

After c. 1975, a fourth cartographic tradition began to emerge rapidly from the European tradition: geographical information systems. An important part of the electronic age it seems to be supplementing rather than replacing the tradition from which it emerged. If so, this will be in keeping with the historically earlier transitions between cartographic traditions.

3.1 Vernacular Terrestrial Maps

Many, perhaps most, adult members of historically traditional societies appear to have possessed the linked abilities to make certain kinds of maps and to understand those made by others in their own and similar societies. These abilities emerged in prehistory (Harley and Woodward 1987, Woodward and Lewis 1998). Artifactual, gestural, and performance maps were observed by Europeans and Euro-Americans in their early contacts with preliterate tribal societies (articles by Bassett, Davenport, Lewis, and Turnbull in Selin 1997). Less frequently, but almost as widespread, tribal peoples practicing divination were observed to interpret cartographically patterns induced on a variety of organic materials. Never constructed according to even approximate linear scale, content was determined by the usually immediate purpose a map was made to serve. For example, physically conspicuous and culturally important ground features might be omitted, whereas minor features and unique events central to a message often dominated, usually to the exclusion of all but an essential minimum of base data. Artifactual maps were made on or with a wide range of media. Most content was terrestrial, some littoral, but very little marine. The maps were almost always short-lived and often ephemeral. In the absence of writing, many artifactual examples were made to communicate to persons not present messages about events, conditions, and proposed activities elsewhere. They were functionally undifferentiated examples of much wider systems of pictographic communication. Sometimes they were made to instruct persons present or in the course of interactive planning. Very rarely did they serve as the equivalents of the reference maps of Western societies, though some were made and from time to time remade to preserve lore. Rather, their equivalents in the latter are the experiential sketch maps of small areas used by individuals to convey local information to others. It seems likely that they manifested the cognitive and behavioral substrata from which formal cartographies later emerged but direct evidence from prehistory is, and seems likely to remain, rare and contestable. Gestural and performance maps could leave no physical evidence and most artifactual examples woulds have had short lives.

In the course of geographical discovery and exploration, Europeans routinely solicited vernacular terrestrial maps from the indigenous peoples of Africa, the Americas, the Arctic, and Australasia, often incorporating them into their own maps. They did not, however, recognise important differences between their own and the indigenes' epistemologies of terrestrial space. What appeared to be crude equivalents of their own maps were fundamentally different. Appraised retrospectively according to Western concepts of accuracy, the consequences of this led to major errors in European cartography of newly discovered areas (Lewis 1986, 1993).

3.2 Celestial and Cosmographical Maps of Shamans and Tribal Leaders

Within historical nonliterate societies, not all maps were terrestrial in content or of vernacular origin. Religious and tribal leaders often had responsibility for making, replicating, and sometimes performing celestial and cosmographical maps, as well as for preserving the lore and conducting the ceremonies with which these were associated. This was particularly so in societies preoccupied with the heavens in the contexts of calendrics and astrology and with a belief in a structured cosmos extending far beyond the world of direct experience. Celestial patterns were, for

example, stitched on shamans' coats, painted on animal hides, incorporated in sand paintings, and mirrored in the organization of rooms and in the positioning of buildings within settlements and vis-à-vis each other. The rapidly developing field of archaeoastronomy is producing increasingly convincing evidence for the incorporation of celestial maps in prehistoric rock art.

The frequency, variety, and physical diversity of cosmographical maps were probably each greater than for celestial maps. In these, planar symmetry about one or more axes was a common geometrical characteristic, often incorporating an axis mundi as a third dimension. Media were very varied, including ceramics, skin drumheads, wall murals, sculpted figures, and scarification of human bodies.

There is little doubt that celestial and cosmographical maps had been made in the antecedents of these societies well before the earliest contact with Europeans. In other societies, maps had already become much more sophisticated; essentially those that had acquired scripts, evolved more advanced forms of religion, and developed large urban communities, especially Babylonia, Egypt, northern China, Persia, Islam, Mesoamerica, and the central Andes.

3.3 Cartography in Early Urban Societies

Urbanization was part cause and part consequence of increasingly complex social systems, religious institutions, civil and military administrations, and extensive trading systems. It also involved increasing specialization and the concentration of wealth. In these conditions new types and styles of maps emerged. Within each of the major urbanized regions a new cartography would appear to have emerged more or less independently. So far as is known, maps were not conceived, compiled, or made by specialist cartographers but as part only of the work of artists, scribes, engravers, sculpters, scholars, astronomers, architects, engineers, etc. The centers of map production were at first the larger urban centers, separated by hinterlands in which the vernacular, shamans', and tribal leaders' maps were almost certainly still the norm. Even in the urban centers map consciousness was probably still weak. Symptomatic of this was the absence of words for 'map' as evidenced in those languages of which there are reasonably reliable lexicons: ancient Greek and Latin, Persian, Arabic, Sanskrit, Hindi, and Chinese before 300 BC. Even so, the maps made in these societies were much more sophisticated and, by later Western standards at least, much more maplike in appearance than the vernacular shamans' and tribal leaders' maps being made elsewhere, as, perhaps, by the nonelites and reactionary groups in the towns.

Between c. 2,300 BC and 500 BC, Babylonian scribes, in addition to writing on clay tablets, used them as surfaces on which to make plans of property, land, houses, and temples, as well as small-scale maps of larger areas and the whole world as they supposed it to be (Millard in Harley and Woodward 1987).

In the fourth millennium BC, rudimentary topographic diagrams began to appear on Egyptian decorated pottery. Detailed plans were being made in ink on plaster-covered boards by the middle of the second millennium BC and cosmographical and celestial maps were being engraved in stone by the end of the first millennium BC (Shore in Harley and Woodward 1987).

In China, maps of many kinds, great complexity, and often exquisite craftsmanship were being made during and after the Warring States period (403 BC – 221 BC). Many have survived; perhaps more than for any of the other early urban societies. Among the earliest is a plan for the construction of a royal tomb, engraved on a bronze plate and inlaid with gold and silver. Others include topographic, administrative, and thematic maps on wooden boards and a map on silk showing a pattern of roads, rivers, and mountains, together with the locations of settlements, scenic sites, and places of historical interest. One of the earliest known maps to have a grid (though probably not based on it) is engraved on a stone dating from AD 1136 and may have had antecedents. There is some direct evidence for mapping according to linear scale and some textual support for this (Wanru et al. 1990 and Yee in Harley and Woodward 1994).

Surprisingly, and unexplained, there are almost no extant terrestrial maps from the Indian subcontinent for the two millennia before the advent of the Portuguese. The exceptions are a few incised potsherds from the end of the first millennium BC that have rough plans of monasteries and a few ancient sculptures depicting sacred rivers. Nevertheless, the first urban culture, the Harappan dating from the middle of the third millennium BC, had many temples and settlements that were so standard in form that they must have been based on precise plans. Furthermore, artifacts from the period have been identified as surveying instruments. Cosmographical mapping may have had its roots in temple design but the maps emerged in their complex and artistically rich forms much later in the Hindu and Jain religious traditions. Likewise, terrestrial mapping, though ultimately sophisticated, emerged late; from the middle of the seventeenth century AD onwards. In the subcontinent, therefore, early urbanization and events associated with it was not the catalyst for cartography—but two much later events were; the emergence of the great religions and contacts with Europeans (Schwartzberg in Harley and Woodward 1992).

Within months of first entering Mexico in 1519, Spaniards were presented with paintings on cloth that were itinerary maps complete with topographic details necessary for route finding. What they either did not see or recognize were three other categories of Meso-

american cartography, terrestrial maps that incorporated accounts of history (cartographic histories); cosmographical maps showing either a horizontal cosmos divided into five quadrants or a vertical cosmos divided into layers along an *axis mundi*; and celestial maps of stars and constellations. Of these, cosmographical models were probably the oldest tradition, dating back to the Olmecs who founded urban centers as early as 1200 BC. By the first millennium AD carved stone tablets showing the cosmic layout were widespread. Celestial maps were less common but cartouches in a Mayan wall mural are believed to represent the arrangements of planets and constellations on the night of an event in AD 792. In central Mexico extant cartographic histories are numerous and elsewhere known by report. One dates from c. 1542. Made on paper, cloth, animal hides, or parchment, they were complex, iconographically sophisticated, and showed what the ruling elites wanted communities to know about their pasts. Artists, scribes, and sculpters also formed an elite but at the conquest product specialization was probably restricted to the larger urban centers. In smaller communities artists were expected to paint everything from pots to maps (Munday in Woodward and Lewis 1998).

In the years after 1532, when the Spaniards first reached the central Andean empire of the Inkas, the invaders were not given and did not report seeing maps. It is possible that flat maps did exist but that native conceptions of space and their symbolizing of it were so different from the Spaniards' experience that they were not recognized. The archaeological record certainly reveals spatial and landscape representations from the previous 2,000 years. The Nazca lines (large ground drawings) may have been made to attract and spatially direct Andean gods. Ceramics of the first millennium AD represent landscapes and buildings three dimensionally. In the years immediately before the conquest, landscapes were carved three-dimensionally on stones and even on extensive rock outcrops. Knotted string devices (khipus) may, in addition to other functions, have been used as maps. Soon after the conquest maps began to be included in native chronicles. There are long traditions of ceremonies that incorporate mapping procedures. Typically, these involve the arrangement of amulets in relation to a ground drawing of a map or plan (Gartner in Woodward and Lewis 1998).

Within 200 years of Mohammed's death in AD 632, the Islamic-Arab world extended from the Atlantic coast of North Africa in the west to the Chinese sphere in the east and was soon to extend south to the Sahara. Together with religious practices, the fostering of scholarship and science, extensive geographical exploration, and urbanization, this rapid territorial expansion stimulated cartography, in part by borrowing languishing concepts and techniques from the earlier Hellenic, Byzantine, and Persian cultures and by interacting with contemporary cultures as distant and different as medieval Europe and Hindu India. Islamic cartographic achievements were numerous and diverse: artifacts of the highest craftsmanship and artistic quality, such as astrolabes in brass incorporating pierced planisphere star maps, celestial globes, and exquisitely painted birds-eye views of cities; mathematically sophisticated circular world maps centered on Mecca, constructed in such a way that the city's distance and direction could be read directly from any locality in the Islamic world; and delicately colored cosmographical maps incorporating three divine worlds, the earth surrounded by seven heavenly spheres and a legendary encircling mountain, together with a seven-layered hell.

With the possible exception of that in China, Islamic cartography was more diverse and sophisticated than that in any of the other early urban societies. Its traditions persisted for more than a millennium. With the exception of some world maps, marine charts, and celestial globes, most maps were made for or copied from manuscript texts, to which they were generally subservient. Many of the texts were translations into Arabic from Greek, Syriac, Hebrew, Persian, Sanskrit, and Turkish. These included Ptolemy's *Almagest* and *Geography*. Islamic scientists and craftsmen were creative in adopting and adapting the techniques described in these translations. Many cartographic artifacts were used in determining religious practices but few, if any, were religious icons. Cartography was encouraged and supported by imperial patrons and many maps were linked with political power. As in all other early urban societies there were no equivalents of the modern professional cartographer. Terrestrial maps, at least, were produced by the elite for the elite for the purposes of edification, illustration, and propagation of imperial glory. With the exception of marine charts and aids to religious practice very few were practical. (Harley and Woodward 1992 and Savage-Smith, Karamustafa, and King in Selin 1997).

3.4 Cartography is European and Other Later Urban Societies

The extent to which Dark Age and medieval cartography in the western Mediterranean and Europe was influenced by that in the classical world is inadequately understood. The importance and quality of maps certainly declined in the late Roman Empire but the role of the Byzantine Empire in transmitting cartographic traditions has still to be assessed. One extant monastery map from the early ninth century AD is reminiscent of the best large-scale Roman plans. Medieval *mappae mundi* were certainly in a tradition extending from the third century AD. Some of the simple world outlines known as T-O maps made from the seventh century AD onwards were influenced by Greco-Roman philosophical tradition. Portolan

charts have been claimed by a minority of authorities to have Roman, Greek, Phoenician, Egyptian, even Neolithic origins (Harvey, Woodward, and Campbell in Harley and Woodward 1987). Possible, probable, or certain though these influences from earlier cultures may have been, they were insignificant compared with that of Ptolemy (Claudius Ptolemaeus).

A second century AD Greek scholar who worked within the framework of the early Roman Empire, Ptolemy's writings were unknown in Europe until 1406, when his *Geography* was translated from Greek into Latin. Hand-copied versions, many with maps, soon began to circulate and printed editions proliferated after 1475. Written as a manual for mapmakers, the original may not have contained maps but its accounts of three projections and tables of latitudes and longitudes were sufficient for the construction of world and regional maps. The *Geography* has been called the touchstone of the Renaissance in European cartography. Before 1482 Ptolemaic maps were exclusively reconstructions but the Ulm edition of that year contained woodcut maps that incorporated recent discoveries by Europeans (Dilke 1985).

Although Ptolemaic maps continued to be published for several centuries, the influence of the ideas in the *Geography* were in the long run the more important. It insisted on carefully measured data, stressed the need for validation, promoted the idea of plotting data with reference to a graticule, and the equally important notion that the graticule could be transformed by the use of different projections, thereby preserving one property at the expense of others. Because they were in such demand, Ptolemaic maps were a factor in stimulating the map-printing technology; at first from woodblocks and, after the mid-sixteenth century, from engraved copper plates. The latter were more readily correctable and revisable than woodblocks (Woodward and Harvey in Harley and Woodward 1987).

For almost 1,000 years before the Ptolemaic impact, cartography in Europe had served the Christian church. Indeed, many, perhaps most, maps had been made by monks and clerics. The primary purpose of *mappaemundi* had been to communicate the significant events in Christian history rather than to record their precise locations, with symbolism and allegory playing major roles. The T in T-O maps separated the three known continents but also symbolized the Christian Cross, in some cases with the body of Christ superimposed on it. Many local maps and plans were of ecclesiastical buildings and church property (Woodward and Harvey in Harley and Woodward 1987). Of the main categories of maps, only the Portolan charts had a primary utilitarian function (Campbell in Harley and Woodward 1987).

The Ptolemaic impact preceded a post-Renaissance change in the role of maps from mainly religious to almost exclusively secular. It occurred during the most momentous and diverse series of developments in the history of cartography before the technical innovations of the second half of the twentieth century. From the early sixteenth century until the mid-nineteenth century Europe dominated the field. Thereafter, the USA and Canada assumed increasing importance. Voyages of discovery and overseas settlement and exploration vastly increased geographical knowledge. During and after the eighteenth-century Enlightenment, science focused attention on hitherto ignored or little known phenomena such as ocean currents and magnetic compass declination, both of which began to be mapped. New and improved instruments increased the accuracy of observations. For example, in the mid-eighteenth century the sealed spring-driven chronometer made it possible to establish longitude accurately anywhere and under virtually any condition (Sobel 1995). Navigation, travel, trade, warfare and international rivalry, settlement of newly discovered and reclaimed lands, Christian missionary activities, searching for and exploiting new and better natural resources, civil administration of increasingly complex societies with emerging senses of social responsibility, planning and civil engineering, science, tourism, popular journalism, the new academic geography that first emerged in Germany after the mid-nineteenth century, and the pedantic geography that paralleled it in the new state education systems were among the factors creating demands for and access to more, better, and different kinds of maps.

At first, those who were known to acquire maps improved their status. For example, the original commissioners of mid-sixteenth century custom-made Italian world atlases; slightly later, those who bought the standardized world atlases of the late sixteenth century Dutch cartographers Abraham Ortelius and Gerardus Mercator; and, soon after that, the purchasers of single-country atlases such as Christopher Saxton's of England and Wales. Those who knew how to use maps, rather than just placing them in their libraries or having their names engraved on them as patrons, increased their political influence, material assets, and intellectual standing. For example, in attempting to consolidate their powers, monarchs in early modern Europe used maps to better define and expand their territories (Buisseret 1992). Known as cadastral maps, cartographic records of property ownership were important tools in the extension of land-based power (Kain and Baigent 1992). From the early eighteenth century onwards, statesmen began establishing permanent departments to map national and colonial territories, systematically, accurately, and usually at much larger scales than those, if any, already available; for example, in France (Konvitz 1987) and British India (Edney 1997). Thereafter, up-to-date topographic map series were to become hallmarks of nation-states. Governors of quasistates such as the Hudson's Bay Company attempted something similar, though less systematically (Ruggles 1991). By the eighteenth century cartographic consciousness had

extended to the rural gentry and new urban middle class subscribers to monthly magazines, most of which included maps on a fairly regular basis. To conform with magazine formats and keep engraving costs down, these were small. Many illustrated current events, and were often made by eminent cartographic engravers such as Emanuel Bowen and Thomas Jefferys (Jolly 1990).

After about 1825, lithography became the popular map-printing technique in Europe. Cheaper than copper-plate engraved maps, the new maps became larger and, from the 1860s onwards, the technique was adapted to print in color. A new generation of maps and atlases began to be published, often cheaply, in large quantities, for use in schools and in the home. In the USA, however, between 1880 and 1940 wax engraving was the dominant technique in commercial cartography. One advantage was that printer's type was used for lettering on maps (Woodward 1977). Wax engraving was important in mass producing categories of widely available, cheap, and sometimes freely distributed maps that extended map consciousness among almost all sectors of the American public; railroad and domestic tourist maps after 1880 and highway maps between the two world wars. They accelerated a trend that had been stimulated by popular journalism during the Civil War. Before 1861 American newspapers rarely contained maps (Bosse 1993). Thereafter, they were to become common (Monmonier 1989).

Between 1500 and 1850 the centers of cartographic innovation and map production were in the financial cities of the densely populated, economically active regions of southern and northwestern Europe. The center of gravity of these activities were Italy (Venice, Rome, and Genoa) in the early and mid-sixteenth century, the Low Countries (Antwerp and Amsterdam) in the late sixteenth and seventeenth centuries, and France (Paris) and England (London) in the eighteenth century. Theory and techniques improved, specialisms emerged, and cartography became increasingly professionalized. By the late nineteenth century the center of gravity had crossed the North Atlantic and dispersed to cities such as New York, Washington DC, and Chicago.

One important category of maps did not conform to these general trends; maps of one or a closely related group of physical, or social, or economic phenomena. Now referred to as thematic maps, early examples were made in the late seventeenth century, rather more during the eighteenth, followed by a burgeoning in northwest Europe during the first half of the nineteenth. A consequence of growth of interest in natural history, the environment, society, and economic change, it was facilitated by scholarly periodicals and only possible once base maps existed. Much of it stemmed from the curiosity and research interests of individuals in topics as diverse as cholera, language, cereals, and volcanoes, often in attempts to demonstrate spatial correlations between two or a few phenomena (Robinson 1982). The emergence of thematic cartography heralded the great late nineteenth and twentieth century national surveys of geology, soil, landuse, etc. and the best of the national atlases published after the middle of the twentieth century.

Aerial photogrammetry, the science of obtaining accurate measurements from air photographs, was used increasingly after World War I. By the 1950s most large-scale mapping and map revision made use of it and in the 1960s and 1970s it was used to map the moon. In 1950, an even greater technical advance was heralded by the publication of the first computer-generated map. Computers are now used in every stage of mapmaking, from surveying to the compilation, design, and production of the end product. Increasingly, maps are custom made. In this respect they are similar to the vernacular maps made in traditional societies. They are not, however, grounded in individual experience or shared traditions but use software programs (geographical information systems) and digitized databases. How historians of cartography in the future will incorporate these developments into the universal history of maps and mapmaking is as yet unclear.

See also: Geography; Space and Social Theory in Geography; Spatial Analysis in Geography; Spatial Data

Bibliography

Bagrow L 1951 *Die Geschichte der Kartographie*. Safari-Verlag, Berlin. First English edition, used in preparing this article, rev. and enl. Skelton R A [trans. Paisey D L] 1964 *History of Cartography*. Harvard University Press, Cambridge, MA

Bosse D 1993 *Civil War Newspaper Maps*. Johns Hopkins University Press, Baltimore, MD

Buisseret D (ed.) 1992 *Monarchs, Ministers, and Maps: The Emergence of Cartography as a Tool of Government in Early Modern Europe*. University of Chicago Press, Chicago

Dilke O A W 1985 *Greek and Roman Maps*. Cornell University Press, Ithaca, NY

Edney M H 1997 *Mapping an Empire: The Geographic Construction of British India 1765–1843*. University of Chicago Press, Chicago

Fletcher D 1995 *The Emergence of Estate Maps: Christ Church, Oxford 1600–1840*. Oxford University Press, Oxford, UK

Harley J B, Woodward D (eds.) *The History of Cartography, Vol. 1, Cartography in Prehistoric, Ancient, and Medieval Europe and the Mediterranean, Vol. 2. book 1, 1992, Cartography in the Traditional Islamic and South Asian Societies; and Vol. 2, book 2, 1994. Cartography in the Traditional East and Southeast Asian Societies*. University of Chicago Press, Chicago

Jolly D C 1990 *Maps in British Periodicals. Part 1. Major Monthlies Before 1800*. D C Jolly, Brookline, MA

Kain R J P, Baigent E 1992 *The Cadastral Map in the Service of the State: A History of Property Mapping*. University of Chicago Press, Chicago

Konvitz J W 1987 *Cartography in France 1660–1848: Science, Engineering, and Statecraft*. University of Chicago Press, Chicago
Lewis G M 1986 Indicators of unacknowledged assimilations from Amerindian maps on Euro-American maps of North America. *Imago Mundi* **38**: 9–34
Lewis G M 1993 Metrics, geometries, signs, and language: sources of cartographic miscommunication between native and Euro-American cultures in North America. *Cartographica* **30**(1): 98–106
Lewis G M (ed.) 1998 *Cartographic Encounters: Perspectives on Native American Mapmaking and Map Use*. University of Chicago Press, Chicago
Lowenthal M A (ed.) 1998 *Who's Who in the History of Cartography: The International Guide to the Subject (D9)*. Map Collector Publications Ltd. for Imago Mundi Ltd., Tring, UK
Monmonier M 1989 *Maps With The News: The Development of American Journalistic Cartography*. University of Chicago Press, Chicago
Robinson A H 1982 *Early Thematic Mapping in the History of Cartography*. University of Chicago Press, Chicago
Ruggles R I 1991 *A Country So Interesting: The Hudson's Bay Company and Two Centuries of Mapping, 1670–1870*. McGill-Queen's University Press, Montreal.
Selin H (ed.) 1997 *Encyclopaedia of the History of Science, Technology, and Medicine in Non-Western Cultures*. Kluwer Academic, Dordrecht, The Netherlands
Skelton R A 1972 *Maps: A Historical Survey of Their Study and Collecting*. University of Chicago Press, Chicago
Sobel D 1995 *Longitude*. Walker and Co., New York
Twyman M 1982 The graphic presentation of language. *Information Design Journal* **3**(1): 1–22
Wallis H M, Robinson A H (eds.) 1987 *Cartographic Innovations: An International Handbook of Mapping Terms to 1900*. Map Collector Publications, Tring, UK in association with the International Cartographic Association
Wanru C, Xihuang Z, Shengzhang H, Zhongxun N, Jincheng R, Deyuan J (eds.) 1990 *An Atlas of Ancient Maps in China—From the Warring States Period to the Yuan Dynasty (476 BC–AD 1368)*. Cultural Relics Publishing House, Beijing
Woodward D 1977 *The All-American Map: Wax Engraving and its Influence on Cartography*. University of Chicago Press, Chicago
Woodward D, Lewis G M (eds.) 1998 *The History of Cartography Vol. 2, book 3, Cartography in the Traditional African, American, Arctic, Australian, and Pacific Societies*. University of Chicago Press, Chicago

G. M. Lewis

Case Study: Logic

1. The Case Study Defined

The case study is a research strategy with special implications for theoretical analysis and data collection. A single case of a particular phenomenon is examined intensively for the light that it can shed on a specific problem or question. Often several methods of data collection will be used to assemble a wide range of information about the case. One of the main benefits of a wide-range of data is that it permits a *triangulation* of research methods, thus, providing substantial verification of the particular phenomenon in question.

Researchers vary in why they choose to study a single case, but most often they do so in order to examine a proposed theory or to provide grounded and detailed information for a new theory (Glaser and Strauss 1967). Although many quantitative researchers consider the case study as a limited form of analysis, especially by comparison with methods that collect information on a large sample of instances, such a conclusion misconstrues fundamentally the main purpose of the case study. The analyst who selects and studies a single case, or even a handful of cases, is not primarily interested in statistical inference to a larger population of units, but rather in theoretical analysis and inference. Nevertheless, the findings from the study of a single case, like those from the study of a large sample of instances, always are to be regarded as tentative, awaiting further confirmation from additional studies of parallel theoretical cases.

2. Why Study the Single Case?

The intensive analysis of the case study is premised on the special advantages that it furnishes. First, the study of single cases enables the researcher to probe a particular question, or phenomenon, in great detail. Such in-depth examinations ultimately permit the researcher to acquire a degree of knowledge about the case that is typically impossible through the examination of a large number of cases. Moreover, such in-depth work also enables the researcher to pursue the examination of alternative theoretical ideas, thus ultimately arriving not merely at a thorough understanding of the empirical facts, but ideally a careful and correct appreciation of the most germane and effective theoretical argument to fit these facts (Campbell 1975).

Second, by studying a single case the researcher is able to take full account of the social, or historical, context of the phenomenon in question. Students of case studies regard context as essential to understanding the nature of the phenomenon. Take, for example, the study of children who are unruly in the classroom. The investigator may believe that such unruliness is the product of how the child relates both to his peers and teacher in the context of classroom activity. In order to appreciate and to fully understand the nature of the child's reactions, the investigator is compelled to study the child in the classroom situation (Stake 1995).

Third, the study of the single case permits the researcher to probe comprehensively into the empirical data at hand. In the study of the workings of a single community, for example, the case study researcher can

explore a variety of dimensions of the community and can thereby create a multidimensional, or holistic, sense of the community rather than, let us say, a unidimensional one based upon its size or territorial breadth. In so doing the researcher can also emerge with a fuller understanding of the case in question by fashioning an integrated portrait of the case into which the various pieces, or dimensions, fit. Fourth, the case study provides boundaries to the nature of the phenomenon under investigation. The case is chosen because it represents a self-contained unit that will permit the researcher to investigate the phenomenon in isolation from other forces. Thus, educational researchers will often investigate a single classroom or school because such a case represents some unusual qualities in which they are interested (Stake 1995).

Finally, the single case is sometimes chosen because it represents a special illustration of the phenomenon under investigation. Sometimes it is portrayed as the exception to the rule, or deviant case, thereby permitting the observer to understand some more general phenomenon in greater depth.

3. *How to Choose Cases*

Great care must be exercised in the choice of the single case. Sometimes social scientists believe that any case may be used to explore a particular phenomenon. Indeed, there appears to have been a great deal of faulty case study research because there are few guidelines by which to select the case. In fact, the choice must always involve certain critical decisions in advance of the data collection and analysis.

First and foremost, the researcher must decide the grounds for selecting a particular unit for study. Since such selection is never based upon issues of statistical inference, the theoretical reasons for selecting the case must always be made clear—or, at least, as clear as possible—in advance of the research. For example, the sociologist R. Stephen Warner believed that broad societal changes might have a deep influence on the life of churches in America. Thus, he chose to focus on a single church, and to examine how the internal organization of the church changed over several decades during which there was marked change in American society. Second, the researcher should choose those cases that will furnish the clearest tests of the theory, or argument, in question. If a researcher is interested in the causes of revolution, then a society should be chosen where there has been a substantial and marked revolution, to examine the possible causes.

3.1 *What Qualifies as a Case?*

Unlike certain forms of research method, such as the sample survey, case study research can employ very different units of analysis on which to focus its theoretical attention. The earliest examples of case studies were often in-depth studies of particular *social roles*, designed to explore the nature of such roles. Other famous case studies have been done of entire *cities*, including the range of different people and organizations in such cities. More recently, social researchers have studied specific *organizations* using the case study format; specific *locales* designed to understand the nature of social groups in those locales; and even specific *individuals* whose traits exemplify special qualities for investigation (Stake 1995). In addition, the case is sometimes used very effectively to explore in rich detail the nature of *social processes* at work. One of the best illustrations of using the case study in this manner was done by the political scientist, Matthew Crenson (1971). He furnished a number of important insights into the nature and exercise of power, including the dynamics of agenda-setting, simply from the study of how political decisions were made in a single city.

4. *The Case Study as a Strategy of Research*

By comparison with other research methods, the case study has certain limitations. It does not permit the easy and refined manipulation of variables in the same way as an experiment does. Nor does it permit a researcher to investigate various configurations of variables that might be associated with particular outcomes, or dependent variables, like the sample survey. Most often, the case study is done using qualitative methods, thus resulting in some confusion between case studies and qualitative research (Merriam 1998). Yet there are important examples in which quantitative data also have been collected to provide key information about the single case. For example, some researchers have used time series data on a single country to examine it in depth.

4.1 *The Case Study as an Inductive Tool*

The case study has proven most useful for the generation of new theory. Because a single case can be examined in great depth, and because it can be studied with a variety of research tools, researchers can mine it for a great deal of information. At the same time, the careful examination of such information can furnish insights not easily acquired by other research methods. Therefore the case study is perhaps at its strongest when it is used as an inductive research instrument, allowing the researcher to construct an explanation for the phenomenon under investigation (Glaser and Strauss 1967).

4.2 *The Case Study as a Deductive Tool*

Yet, the study of the single case can also furnish an important tool for those observers who wish to test, or apply, deductive theories. One of the most famous

illustrations of this occurs in the classic study of the International Typographers Union (Lipset et al. 1956). The authors had concluded, both on empirical and theoretical grounds, that trade unions were characterized by very rigid and tightly closed political administrations, with one clique maintaining its rule in office over a long period of time. They then chose deliberately to examine the ITU in detail, because it appeared to be the exception to the general situation. What was it about the ITU, they asked, that produced empirically disparate results from all other trade unions? This was a powerful, well-refined question. In effect, the ITU provided a comparison to all other cases: if its outcome was different, so too must have been the set of elements that brought about such an outcome. The answer, furnished through a variety of evidence, suggested that, because of their level of education, skill, and commitment to the ITU, union members of this union, as compared to others, simply were more likely to take an active role in union affairs.

The one consideration that is overlooked in using the case study as a deductive tool is that researchers may fail to specify the nature of the null hypothesis, or the likely outcome if the theory is incorrect. While most case studies have failed to consider this important feature of theory construction and testing, Yin (1994) argues that, if done carefully, such a strategy can be incorporated easily into the analysis of single, and even multiple, cases.

5. *The Case Study: Seminal Illustrations*

Over the years there have been a number of very significant case studies. Two, in particular, illustrate the wide variety of uses to which case study analysis can be put in the study of social phenomena.

One of the first, and surely most famous, case studies was done of the city of Muncie, Indiana (Lynd and Lynd 1929). Conducted during the course of the 1920s, the investigators and their research team wanted to uncover the full nature of a typical American city over the course of its lifetime. Part of their work involved tracing the history of the city. A large part, however, involved the collection of a wide range of information on the city's current residents, organizations, and its internal workings.

There were a variety of important discoveries, including discoveries about the nature of work life and religious life in the community, as well as about the dominance of a leading industrial family over the entire range of civic and political life of the community. It might be noted that sample surveys of city residents, or demographic data about them, would never have uncovered the full and complex portrait the Lynds were able to construct through their pioneering case study research.

A second famous case study was done of men who lived in the South End of Boston (Whyte 1943). This study, also conducted in the 1930s, was ostensibly an examination of a group of impoverished men, and how they adapted to life during the Great Depression. In fact, as Whyte so brilliantly showed, it really was a case study of a small group of men, and the way in which the nature of stratification emerged within the group, and how it influenced the behavior of individuals. It was done with great insight and care by Whyte, making it still one of the best and most influential case studies of all time.

6. *Issues of Theoretical Specification: The Use of Singular Cases to Illuminate General Principles*

Sometimes the study of the single case is done specifically to reveal the importance of a special configuration, or confluence, of elements that together help to explain the unique outcome of a case, and its difference from the general pattern (Campbell 1975). The logic underlying the analyst's concern is that the set of variables together constitutes the sufficient, perhaps even singular, explanation for a particular outcome, or effect. The research on the International Typographers Union by Lipset et al. (1956), for instance, pointed to the special configuration of conditions in the ITU that helped to promote internal democracy. It suggested that the special confluence of factors promoted democracy, leading to the obvious conclusion that only under certain special social and historical circumstances could such variables again come together in such a specific manner.

7. *Issues of Theoretical Generalization and Refinement: The Use of Parallel and Different Cases*

Because the analysis of a single case can only take the analyst so far, the best rule to follow, if possible, is to examine at least three or four pertinent cases in depth. Multiple cases permit the researcher to refine and develop theoretical arguments, and they do so precisely because case studies permit the researcher to examine cases deeply and, ultimately, to make careful comparisons among those cases.

One example will illustrate the advantages of using several cases and examining them in depth. Orum (1995) was interested in examining various theories about the origins of urban growth. Popular theories had suggested that such growth came about because of special alliances among the leading figures in different major institutional spheres, particularly local government, business, and the media. Orum suggested that such theories were limited only to contemporary circumstances and that they did not help much to explain the sources of growth in the past. Using several different cases, and studying them in great depth permitted him to come up with several important discoveries bearing upon issues of urban growth and expansion.

The first case examined was that of Milwaukee, Wisconsin. Tracing its origins back to the early nineteenth century, Orum discovered that the city seemed to go through certain stages in its growth, and that one could not simply rely on explanations of alliances among leaders of different institutional spheres. Thus, based upon intensive study, he argued that, in the earliest years, the growth of the city relied heavily on local entrepreneurs, people who promoted the city and who, among other things, sold land for it. But over time this motive force for growth changed, and the main forces turned to new economic institutions, ones that brought in growth by attracting new residents. A third stage happened when the impetus for growth shifted from the local economic enterprises, which had grown somewhat stagnant, to local government. Government sought to expand the community by adding new territory. Among other things, it also became clear in this stage that a single conception of urban growth was inappropriate; growth could take the form of new population, but also new territory. Such an insight would have been impossible had Orum relied only on the study of this single case in depth. The fourth and final phase happened as the city began to decline. In large part, the decline took place when many of its original industries left for greener pastures in other cities. Under these circumstances, a heavy burden was placed upon local government to devise ways to promote growth, either by expanding its borders or by attracting new industry through various incentives.

In effect, then, Orum's explanation for Milwaukee's growth suggested that such expansion happened by stages, and that at each stage there was a different configuration of local business and political forces responsible for such growth, and that the principal form of growth itself could vary between stages. Next, he asked whether this pattern held only for Milwaukee, or whether it could also describe the trajectory of growth and decline in other parallel industrial cities. Thus, he made comparisons in the pattern and stages of growth in Milwaukee with those in Cleveland, Ohio, a city which, on the face of it, seemed similar in many respects. In fact, the stages and forces of growth at each stage were virtually identical in Cleveland and Milwaukee. He thus argued that among American industrial cities, the nature of growth, and the forces responsible for it, showed important parallels, underscoring his argument on behalf of a stage theory of urban growth.

Finally, Orum pushed the argument one step further, wondering whether the pattern found was only characteristic of industrial growth, and cities, in America. Could the pattern of stages also apply to cities that grew up in the postindustrial era and were themselves postindustrial cities? Here he made a comparison between the patterns in Milwaukee and those in Austin, Texas, a postindustrial city he also had come to know intimately. This comparison showed that the first three stages were very similar in terms of the rapidity of growth, and in the role of entrepreneurs, local business, and local government in promoting such growth in Milwaukee and Austin alike. In the end, then, he refined his overall argument to suggest that the stage view of urban growth in America fit not only industrial-era cities, but also cities in the postindustrial era. This suggested that there appeared to be an underlying structural pattern to the growth of American cities, regardless of era. This last comparison also enabled him to draw out differences between Milwaukee and Austin, allowing for further understanding of the nature of urban growth, and difference, in the United States.

In brief, then, as the example of the various cities is intended to illustrate, the careful and systematic study of several cases can permit the researcher both to substantiate important theoretical generalizations and to refine and extend them. Again, this can be done because of the deep acquaintance the researcher gains from the study of cases, an acquaintance unlike that to be obtained with methods like surveys or experiments.

8. Conclusion

The use of a single case to develop, or to test, theoretical insights remains a very important strategy in the social sciences at the start of the twenty-first century. Indeed, it would appear to be growing in significance. But it must be executed with great care, and with a special sensitivity to theoretical issues. The major limitation in the study of the single case, or even a handful of cases, lies not in the limited number of empirical units, but in the ability of the researcher to be sensitive to issues of theory development and the gathering of relevant evidence to develop and refine that theory.

See also: Case Study: Methods and Analysis; Case-oriented Research; Human–Environment Relationship: Comparative Case Studies; Psychotherapy: Case Study; Single-subject Designs: Methodology; Time Series: General

Bibliography

Bradshaw Y, Wallace M 1991 Informing generality and explaining uniqueness: The place of case studies in comparative research. *International Journal of Comparative Sociology* **32**(1–2): 154–71

Campbell D T 1975 'Degrees of Freedom' and the case study. *Comparative Political Studies* **8**(2): 178–93

Crenson M 1971 *The Unpolitics of Air Pollution*. Johns Hopkins Press, Baltimore, MD

Feagin J R, Orum A M, Sjoberg G (eds.) 1991 *A Case for the Case Study*. University of North Carolina Press, Chapel Hill, NC

Glaser B G, Strauss A L 1967 *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, Chicago
Lipset S M, Trow M, Coleman J S 1956 *Union Democracy: the Internal Politics of the International Typographical Union*. Free Press, Glencoe, IL
Lynd R, Lynd H M 1929 *Middletown: A Study in American Culture*. Harcourt Brace, New York
Merriam S B 1998 *Qualitative Research and Case Study Applications in Education*. Jossey-Bass, San Francisco
Orum A M 1995 *City-Building in America*. Westview Press, Boulder, CO
Ragin C C, Becker H S (eds.) 1992 *What Is A Case? Exploring the Foundations of Social Inquiry*. Cambridge University Press, Cambridge, UK
Ragin C, Zaret D 1983 Theory and method in comparative research: Two strategies. *Social Forces* **61**(3): 731–54
Stake R E 1995 *The Art of Case Study Research*. Sage, Thousand Oaks, CA
Warner R S 1991 Oenology: The making of *New Wine*. In: Feagin J R, Orum A M, Sjoberg G (eds.) *A Case for the Case Study*. University of North Carolina Press, Chapel Hill, NC, pp. 174–99
Whyte W F 1943 *Street Corner Society: The Social Structure of an Italian Slum*. University of Chicago Press, Chicago
Yin R K 1994 *Case Study Research: Design and Methods*, 2nd edn. Sage, Thousand Oaks, CA

A. M. Orum

Case Study: Methods and Analysis

Case study methods have been around as long as recorded history, and they presently account for a large proportion of the books and articles in anthropology, biology, economics, history, political science, psychology, sociology, and even the medical sciences. The logic of case study methods, much like that of any historian's or detective's efforts to make inferences from patterns within cases and comparisons between them, is more intuitive than the logic of statistical inference. Until relatively recently, however, the lack of formalization of the logic of case study methods inhibited them achieving their full potential for contributing to the progressive and cumulative development of theories. It is only in the last three decades that scholars have formalized case study methods and linked them to underlying arguments in the philosophy of science. Ironically, statistical methods, though less intuitive, were standardized earlier, so that attempts to formalize case study methods often misappropriated terms and concepts from statistics (McKeown 1999). More recently, case study methods have evolved from being defined 'negatively,' via contrasts to statistical methods, to being defined 'positively,' by their distinctive logic, techniques, and comparative advantages. This continuing evolution remains a contested process, but there is growing consensus on the proper procedures for carrying out case studies and the strengths and limits of such studies. It is becoming increasingly clear through this process that the comparative advantages of case study and statistical methods are largely complementary and that the two methods can thus achieve far more scientific progress together than either could alone.

1. *The Definition of 'Case' and 'Case Study'*

Early efforts to define 'case studies' relied on distinctions between the study of a small vs. a large number of instances of a phenomenon. Case studies thus became characterized as 'small n' studies, in contrast to 'large *N*' statistical studies. Related to this, one early definition stated that a 'case' is a 'phenomenon for which we report and interpret only a single measure on any pertinent variable' (Eckstein 1975). Case study researchers have increasingly rejected this definition, however, because it wrongly implies, in the language of statistics, that there is an inherent 'degrees of freedom' problem in case studies. In other words, this definition of case studies suggests that with a greater number of variables than observations on the dependent variable, case studies provided no basis for causal inference. In the view of case study researchers, however, each case includes a potentially large number of observations on intervening variables and qualitative measures of different aspects of the dependent variable, so there is not just a 'single measure' of the variables or an inherent degrees of freedom problem. This point is increasingly recognized by researchers from the statistical tradition (King et al. 1994).

In addition, the 'small n/large N' distinction implies that large N methods are always preferable whenever sufficient data is available. As argued below, however, case studies can serve useful theory building purposes, such as the inductive generation of new hypotheses, even when instances of a phenomenon are sufficiently numerous to allow the application of statistical methods.

For present purposes, then, *a case is defined as an instance of a class of events* (George 1979a, 1979b). The term 'class of events' refers here to a phenomenon of scientific interest, such as revolutions, types of governmental regime, kinds of economic system, or personality types. A *case study is thus a well-defined aspect of a historical happening that the investigator selects for analysis*, rather than a historical happening itself. The Cuban Missile Crisis, for example, is a historical instance of many different classes of events: cases of deterrence, coercive diplomacy, crisis management, and so on. In deciding which class of events to study and which theories to use, the researcher decides what data from the Cuban Missile Crisis is relevant to their case study of it. Of course, even if one accepts the present definition of a case, there is still room in the context of particular studies to debate such questions as: 'what is this event a case of' and 'given this phenomenon, is this event a case of it?' (Ragin and Becker 1992).

There is potential for confusion among the terms 'comparative methods,' 'case study methods,' and 'qualitative methods.' In one view the comparative method, or the use of comparisons among a small number of cases, is distinct from the case study method, which in this view involves the internal examination of single cases (see *Comparative Studies: Method and Design*). For the present purposes, however, case study methods are defined to include both within-case analysis of single cases and comparisons between or among a small number of cases. This is not an effort to claim wider meaning for the term 'case studies,' but an outgrowth of the growing consensus that the strongest means of drawing inferences from case studies is the use of a *combination* of within-case analysis and cross-case comparisons within a single study or research program, although single case studies can also play a role in theory development. As for the term 'qualitative methods,' this is sometimes used to encompass both case studies carried out with a positivist view of the philosophy of science and those implemented with a postmodern or interpretive view. This present article hews to the traditional terminology in focusing on 'case studies' as that subset of qualitative methods that has adopted a largely positivist framework.

2. *The Historical Development of Case Study Methods*

Case study methods have developed through several phases over the last three decades. Prior to the 1970s, 'case studies' consisted primarily of historical studies of particular events, countries, or phenomena, with little effort to cumulate results or progressively develop theories (Verba 1967). Throughout the 1970s, however, scholars who were dissatisfied with the state of case study methods, and encouraged by the example of the formalization of statistical methods, began to formalize case study methods.

First, Adam Przeworski and Henry Teune (1970) clarified the logic of 'most similar' and 'least similar' case comparisons. In the former comparison, which draws on the logic of John Stuart Mill's method of difference and mimics the experimental method, the researcher compares two cases that are similar in all but one independent variable and that differ in the outcome variable. Such a comparison may be consistent with the inference that the difference in the single independent variable that varies between the cases accounts for the difference in the dependent variable (although for a variety of reasons discussed below, this inference may be spurious). In a comparison of least similar cases, which draws on Mill's method of agreement, the researcher compares two cases that differ in all but one independent variable but that have the same value on the dependent variable. If, for example, we find that teenagers are 'difficult' in both tribal societies and industrialized societies, we might be tempted to infer that it is the nature of teenagers rather than the nature of society that accounts for the difficulty of teenagers.

Table 1
Types of case studies

Lijphart	Eckstein
atheoretical	configurative-ideographic
interpretative	disciplined-configurative
hypothesis generating	heuristic
deviant	?
theory-confirming/ infirming	crucial, most-likely, least-likely

Arend Lijphart (1971) and Harry Eckstein (1975) contributed further to the formalization of case study methods by clarifying the differences among various types of case study research designs and theory-building goals. These authors identified similar types, although their terminology differs and Lijphart adds an important type, the 'deviant case,' for which Eckstein does not make explicit provision. Their types of case studies correspond as shown in Table 1.

The atheoretical or configurative-ideographic case study takes the form of a detailed narrative or 'story' presented in the form of a chronicle that purports to illuminate how an event came about. Such a narrative is highly specific and makes no explicit use of theory or theory-related variables. Most case studies, however, do have an explanatory purpose. These studies generally fall into the category of 'disciplined-configurative' or 'interpretive' case studies, in which general propositions are used, often implicitly, to explain specific historical cases. Another variant of such case studies is the use of cases as examples that illustrate a theory.

Heuristic case studies seek to generate new hypotheses inductively from the study of particular cases. Notably, statistical methods lack this capacity for inductively generating hypotheses, and they typically rely instead on hypotheses derived deductively or borrowed from case study research. An especially important type of case study for developing new hypotheses is the 'deviant' case study. This is the study of a case whose outcome is not predicted or explained adequately by existing theories. Unless the outcome of a deviant case turns out to be a consequence of measurement error, the case is likely to be useful for identifying variables that have been left out of existing theories. Finally, researchers can use case studies to test whether the outcomes and processes that theories predict in particular cases are in fact evident.

Eckstein's and Lijphart's contributions demonstrated that there was not just a single type of case study, but many kinds of case study research designs and many different theory-building purposes that they

could serve. Their treatments differed, however, in that Lijphart relied greatly on statistical concepts and language. He was thus skeptical of the value of single case studies for building social science theories, and, consistent with the widespread preference at the time for 'large *N*' over 'small *n*' methods, he urged researchers to consider several means of either decreasing the number of variables in their models or increasing the number of cases to be studied in order to make use of statistical rather than case study methods. This advice, however, raised the risk 'conceptual stretching' (Sartori 1970), or of lumping together dissimilar cases under the same definitions. Possibly for this reason, Lijphart later placed greater emphasis instead on the controlled comparison of most similar to cases as a basis for causal inference (Lijphart 1975, Collier 1993).

Eckstein, in contrast, focused on the use of case studies for theory testing and argued that even single case studies could provide tests that might strongly support or impugn theories. In so doing, Eckstein developed the idea of a 'crucial case,' or a case that '*must closely fit* a theory if one is to have confidence in the theory's validity, or, conversely, *must not fit* equally well any rule contrary to that proposed' (Eckstein 1975, his emphasis). Eckstein argued that true crucial cases are rare, so he pointed to the alternative of 'most likely' and 'least likely' cases. A most likely case is one that is almost certain to fit a theory if the theory is true for any cases at all. The failure of a theory to explain a most likely case greatly undermines our confidence in the theory. A least likely case, conversely, is a tough test for a theory because it is a case in which the theory makes only a weak prediction. A theory's ability to explain a least likely case is strong evidence in favor of the theory. In this way, Eckstein argued, even single case studies could greatly increase or decrease our confidence in a theory or require that we alter its scope conditions.

Alexander George (1979a, 1979b) further developed case study methods by refining 'within-case' analysis and cross-case comparisons in ways that help each method compensate for the limits of the other. George argued, as Mill himself had, that the 'method of difference' and the corresponding practice of comparison of most similar cases could lead to spurious inferences. One reason for this is that no two nonexperimental cases achieve the ideal of being similar in *all* respects but one independent variable and the outcome. Thus, there is always the danger that left-out variables or residual differences in the values of the independent variables account for the difference in the outcomes of similar cases of (see *Human–Environment Relationship: Comparative Case Studies*). In addition, as Mill recognized, phenomena might be characterized by what general systems theorists have termed 'equifinality,' or the condition in which the same outcome can arise through different causal pathways or combinations of variables. Thus, there might be no single necessary or sufficient variable for a phenomenon: it might be that either ABC or DEF causes Y, and that none of the variables A–F is itself sufficient to cause Y (see *Human–Environment Relationship: Comparative Case Studies*). In such circumstances, pair-wise comparisons of cases might wrongly reject variables that contribute to the outcome of interest in conjunction with some contexts but not with others, and might also accept as causal variables that are in fact spurious.

To compensate for these limits of controlled comparison, George developed the 'within case' methods of 'congruence testing' and 'process tracing' as means of checking on whether inferences arrived at through case comparisons were spurious (see *Pattern Matching: Methodology*). In congruence testing, the researcher checks whether the prediction a theory makes in a case, in view of the values of the case's independent variables, is congruent with the actual outcome in the case. In process tracing, the researcher examines whether the causal process a theory hypothesizes in a case is in fact evident in the sequence and values of the intervening variables in that case. Thus, process tracing might be used to test whether the residual differences between two similar cases were causal or spurious in producing a difference in these cases' outcomes. Process tracing can perform a heuristic function as well, generating new variables or hypotheses on the basis of sequences of events observed inductively in cases.

George (1979a, 1979b) also systematized case study procedures by developing what he called the method of 'structured focused comparison.' In this method, the researcher systematically: (a) specifies the research problem and the class of events to be studied; (b) defines the independent, dependent, and intervening variables of the relevant theories; (c) selects the cases to be studied and compared; (d) decides how best to characterize variance in the independent and dependent variables; and (e) formulates a detailed set of standard questions to be applied to each case. In addition, consistent with his emphasis on equifinality, George argued that case studies could be especially useful in developing what he called 'typological theories,' or contingent generalizations on '*the variety of different causal patterns* that can occur for the phenomena in question … [and] *the conditions under which each distinctive type of causal patterns occurs*' (George 1979a, his emphasis). He thus advocated a kind of 'building block' approach to the development of theories in which each case, while rendered in terms of theoretical variables, might prove to be a distinctive causal pathway to the outcome of interest.

In the 1980s and 1990s, thousands of books and articles made use of these improvements in case study methods in a wide variety of social science research programs. Meanwhile, scholars continued to elaborate case study methods and articulate the ways in which they differed from statistical methods. David Collier,

reviewing the development of case study and comparative methods, argued that these methods have advantages in defining and measuring qualitative variables in conceptually valid ways and forestalling the problem of conceptual stretching (Collier 1993). Charles Ragin argued that qualitative methods were also better than statistical methods at accounting for equifinality and complex interaction effects. Although statistical methods can model several kinds of interaction effects, Ragin noted, they can do so only at the cost of requiring a larger sample size, and models of nonlinear interactions rapidly become complex and difficult to interpret. Ragin also introduced the method of Qualitative Comparative Analysis, which uses Boolean algebra to reduce a series of comparisons of cases to the minimum number of logical statements or hypotheses that entail the results of all the cases compared (Ragin 1987). This method, he argues, makes comparisons among cases in ways that treat them inherently as configurations of variables, and that thus allow for the possibility of equifinality and complex interactions (see *Configurational Analysis*). Both Collier and Ragin also noted the limitations of case study methods, including the potential for indeterminacy when attempting to sort out rival explanations in a small number of cases, the difficulty of attaining a detailed understanding of more than a few cases, and the inability to make broad generalizations on the basis of small numbers of cases.

3. *New Developments in Case Study Methods*

The thousands of applications of case study methods in the last two decades have provided fertile ground for further methodological refinements. Three key recent developments include the strengthening of linkages between case study methods and the philosophy of science, the elaboration of the concept of typological theories, and the emergence of elements of consensus on the comparative advantages and limitations of case study methods.

3.1 *Case Studies and the Philosophy of Science*

With regard to the philosophy of science, the 'scientific realist' school of thought has emphasized that causal mechanisms, or independent stable factors that under certain conditions link causes to effects, are important to causal explanation (Little 1998). This has resonated with case study researchers' use of process tracing to uncover evidence of causal mechanisms at work. It has also provided a philosophical counterpoint to attempts by researchers from the statistical tradition to place 'causal effects,' or the expected difference in outcomes brought about by the change in a single independent variable, at the center of causal explanation (King et al. 1994). Case study researchers have argued that both causal mechanisms, which are more easily addressed by case studies, and causal effects, which are best assessed through statistical means, are essential to the development of causal theories and causal explanations (George and Bennett 2001).

Another relevant development in the philosophy of science has been the resurgence of interest in Bayesian logic, or the logic of using new data to update prior confidence levels assigned to hypotheses. Bayesian logic differs from that of most statistics, which eschew reliance on prior probabilities. Eckstein's crucial, most likely, and least likely case study designs implicitly use a Bayesian logic, assigning prior probabilities to the likelihood of particular outcomes (McKeown 1999). One new development here is the refinement of Eckstein's approach, taking into consideration the likelihood of an outcome not just in view of one theory, but in the presence of alternative hypotheses. If a case is 'most likely' for a theory, and if the alternative hypotheses make the same prediction, then the theory will be strongly impugned if the prediction does not prove true. The failure of the theory cannot be blamed on the influence of the variables highlighted by the alternative hypotheses. Conversely, if a theory makes only a weak prediction in a 'least likely' case, the alternative hypotheses make a different prediction, but if the first theory's prediction proves true, this is the strongest possible evidence in favor of the theory (Van Evera 1997). This helps address the central problem of a Bayesian approach—that of assigning and justifying prior probabilities—even if it does not fully resolve it.

The continuing development of the logic of hypothesis testing has also been relevant to case study methods (see *Hypothesis Testing: Methodology and Limitations*). On this topic, Imre Lakatos argued that a theory can be considered progressive only if it predicts and later corroborates 'new facts,' or novel empirical content not anticipated by other theories (Lakatos 1976). This criterion helps provide a standard for judging whether process tracing, the designation of new subtypes, and the proposal of new theories from heuristic case studies are being done in a progressive or regressive way. It also provides a philosophical basis for arguing that a hypothesis can be derived from one set of observations within a case and then to some extent tested against the 'new facts' or previously unexamined or unexpected data that it predicts within that same case, although independent corroboration in other cases is usually advisable as well (Collier 1993).

3.2 *Typological Theories and 'Fuzzy Logic'*

A second recent development in case study methods has been the elaboration of the concept of typological theory. Typological theories occupy a middle ground between covering laws, or highly general abstract propositions, and causal mechanisms. Typological

theories identify recurring conjunctions of mechanisms and provide hypotheses on the pathways through which they produce effects. Thus, like QCA, typological theories treat cases as configurations. Unlike QCA, they do not attempt to reduce the number of theoretical statements about the variables, but retain a diverse and admittedly complex set of contingent generalizations, with potentially one generalization per type. Consequently, typological theories are well suited to modeling equifinality.

To construct typological theories, researchers first specify the variables and use them to define the typological space, or the set of all mathematically possible combinations of the variables (this is sometimes termed a truth table in the philosophy of science). At first this may seem to produce an unmanageably large number of combinations: a model with five dichotomous variables, for example, would have 32 possible types. However, once the researcher begins to categorize extant cases in a preliminary way into particular types, it often becomes possible to narrow the range of cases of interest for study. Many types may remain empty, with no extant cases. Some types may be overdetermined for the outcome of interest, and hence not worthy of study unless they have an unexpected outcome. From among the cases and types that remain, the researcher can use the preliminary categorization of cases within the typological space to help identify most likely, least likely, most similar, least similar, and crucial cases for study. Cases in the typological space with unexpected outcomes, or deviant cases, can help identify new causal pathways that can be added to the existing theory in a kind of 'building block' approach (George and Bennett 2001). A related development concerns the concept of 'fuzzy logic' (Ragin 2000). Fuzzy logic treats cases as configurations but rather than using dichotomous or trichotomous variables and categorizations of cases, it allows the use of scaling to give a score on the extent to which a case fits into a certain type. In other respects, the use of fuzzy logic proceeds in ways much like those of typological theories.

3.3 The Emerging Consensus on the Strengths and Limits of Case Study Methods

A third development is that while several debates on case study methods continue, others have moved toward synthesis or even closure, and the overall picture is of an emerging consensus on the advantages limitations of case study methods. As noted above, researchers from a variety of methodological traditions have recognized that because case studies can include many observations, they do not suffer from an inherent degrees of freedom problem. At the same time, it is also widely agreed that particular case studies may suffer from indeterminacy, or an inability to exclude all but one explanation on the basis of available process tracing evidence (Njolstad 1990). When this occurs, it may still be possible to narrow the number of plausible explanations, and it is also important to indicate as clearly as possible the extent to which the remaining hypotheses appear to be complementary, competing, and incommensurate in explaining the case.

Second, most case study researchers have readily acknowledged the limits of Mill's methods. Ragin's alternative of qualitative comparative analysis makes less restrictive assumptions, but its results are highly sensitive to changes in the measurement or coding of a single case (Goldthorpe 1997). There has thus been a movement toward typological theories and fuzzy logic, which make still less restrictive assumptions than QCA and are not so sensitive to the results of a single case. In addition, there is growing consensus that the use of within-case methods of analysis helps provide a check on the potential spuriousness of cross-case comparisons (Collier 1993, Mahoney 1999, George and Bennett 2001). Case study researchers consequently seldom if ever rely on case comparisons alone.

Third, there is growing recognition that the case selection criteria necessary for statistical studies are in some respects inappropriate for case studies. Random selection in a case study research design, for example, can result in worse biases than intentional selection (King et al. 1994). There is also increasing understanding that, consistent with the reliance of some case study designs on a Bayesian logic, case studies are sometimes intentionally selected not to be representative of some wide population but to provide the strongest possible inferences on particular theories (McKeown 1999). There is still disagreement between those who warn against any selection on the dependent variable (King et al. 1994) and those who argue that selection on the dependent variable is appropriate for some research objectives (Collier and Mahoney 1996, Ragin 2000, George and Bennett 2001). Related to this is a continuing disagreement over whether single case studies can make only limited contributions to theory building (King et al. 1994), or whether single case studies have indeed reshaped entire research programs (Rogowski 1995). There is wider agreement, however, that selection bias is potentially more severe in case studies than statistical studies because biased selection of case studies can overstate as well as understate the relationship between the independent and dependent variables (Collier and Mahoney 1996).

On the whole discussions of these issues have moved toward an emerging consensus on the comparative advantages and limitations of case study methods. These methods' advantages include the conceptualization, operationalization, and measurement of qualitative variables (conceptual validity), the avoidance of conceptual stretching, the heuristic identification of new variables and hypotheses (often through study of deviant cases), the assessment of

whether statistical generalizations offer plausible or spurious explanations of individual cases, the incorporation of equifinality and complex interactions effects, and the inferences made possible by combining within-case and cross-case analyses (Collier 1993, Munck 1998, George and Bennett 2001). It is possible that new statistical methods may be able to improve upon the statistical treatment of equifinality and interaction effects, and at least narrow the gap in the treatment of this issue, but the other comparative advantages of case study methods appear to be inherent in their differences from statistical methods.

The limits of case study methods include their inappropriateness for judging the relative frequency or representativeness of cases, their weakness at performing partial correlations and establishing causal effects or causal weight, the necessarily narrow and contingent nature of their generalizations, and the danger that selection bias can be more catastrophic than in statistical studies (Collier 1993, King et al. 1994, Munck 1998, George and Bennett 2001). Fortunately, these are precisely the strengths of statistical studies.

4. Future Directions in Case Study Methods

Just as the formalization of case study methods in the 1970s inspired a generation of more sophisticated research, the recent further refinements in these methods are likely to lead to still greater sophistication in their use in the social, biological, and even physical sciences. The increasingly evident complementarity of case study and statistical methods is likely to lead toward more collaborative work by scholars using various methods. The recent interest among rational choice theorists in using case studies to test their theories, for example, is an important step in this direction (Bates et al. 1998). Because case studies, statistical methods, and formal modeling are all increasingly sophisticated, however, it is becoming less likely that a single researcher can be adept at more than one set of methods while also attaining a cutting-edge theoretical and empirical knowledge of their field. Collaboration might therefore take the form of several researchers working together using different methods, or of researchers more self-consciously building on the findings generated by those using different methods. In either form, effective collaboration requires that even as they become expert in one methodological approach, scholars must also become conversant with alternative approaches, aware of their strengths and limits, and capable of an informed reading of their substantive results.

See also: Biographical Methodology: Psychological Perspectives; Case Study: Logic; Case-oriented Research; Configurational Analysis; Human–Environment Relationship: Comparative Case Studies; Psychotherapy: Case Study; Single-case Experimental Designs in Clinical Settings; Single-subject Designs: Methodology

Bibliography

Bates R H, Greif A, Rosenthal J, Weingast B, Levi M 1998 *Analytic Narratives*. Princeton University Press, Princeton, NJ

Collier D 1993 The comparative method. In: Finifter A W (ed.) *Political Science: the State of the Discipline II*. American Political Science Association, Washington, DC

Collier D, Mahoney J 1996 Insights and pitfalls: selection bias in qualitative research. *World Politics* **1**: 56–91

Eckstein H 1975 Case studies and theory in political science. In: Greenstein F I, Polsby N W (eds.) *Handbook of Political Science*. Addison-Wesley, Reading, MA, Vol. 7

George A L 1979a Case studies and theory development. In: Lauren P G (ed.) *Diplomacy: New Approaches in Theory, History, and Policy*. Free Press, New York

George A L 1979b The Causal nexus between cognitive beliefs and decision-making behavior: the 'Operational Code.'. In: Falkowski L S (ed.) *Psychological Models in International Politics*. Westview, Boulder, CO

George A L, Bennett A O 2001 *Case Studies and Theory Development*. MIT Press, Cambridge, MA

Goldstone J 1997 Methodological issues in comparative macrosociology. *Comparative Social Research* **16**: 107–20

Goldthorpe J 1997 Current issues in comparative macrosociology. *Comparative Social Research* **16**: 1–26

King G, Keohane R, Verba S 1994 *Designing Social Inquiry*. Princeton University Press, Princeton, NJ

Lakatos I 1976 Falsification and the growth of scientific research programs. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK

Lijphart A 1971 Comparative politics and the comparative method. *American Political Science Review* **65**: 682–93

Lijphart A 1975 The comparable cases strategy in comparative research. *Comparative Political Studies* **8**: 158–77

Little D 1998 *Microfoundations, Method, and Causation*. Transaction, New Brunswick, NJ

Mahoney J 1999 Nominal, ordinal, and narrative appraisal in macro-causal analysis. *American Journal of Sociology* **104**: 1154–96

McKeown T 1999 Case studies and the statistical world view. *International Organization*. **1**: 161–90

Munck G L 1998 Canons of research design in qualitative analysis. *Studies in Comparative International Development* **33**: 18–45

Njolstad O 1990 Learning from history? Case studies and the limits to theory-building. In: Gleditsch N P, Njolstad O (eds.) *Arms Races: Technological and Political Dynamics*. Sage, Newbury Park, CA

Przeworski A, Teune H 1970 *The Logic of Comparative Social Inquiry*. Wiley-Interscience, New York

Ragin C C 1987 *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. University of California Press, Berkeley, CA

Ragin C C 2000 *Fuzzy-Set Social Science*. University of Chicago Press, Chicago, IL

Ragin C C, Becker H S 1992 Introduction. In: Ragin C C, Becker H S (eds.) *What is a Case? Exploring the Foundations of Social Inquiry*. Cambridge University Press, Cambridge, UK

Rogowski R 1995 The role of theory and anomaly in social-scientific inference. *American Political Science Review* **89**: 467–70
Sartori G 1970 Concept misformation in comparative politics. *American Political Science Review* **64**: 1033–53
Van Evera S 1997 *Guide to Methods for Students of Political Science*. Cornell University Press Ithaca, NY
Verba S 1967 Some dilemmas in comparative research. *World Politics* **20**: 111–27

A. Bennett

Case-oriented Research

1. Introduction

Case-oriented research focuses on interconnections among parts and aspects within single cases. In this approach, the researcher attempts to make sense of each case as a singular, interpretable entity. In-depth knowledge of the cases included in a study is considered a prerequisite for the examination of patterns that might be observed across cases. Case-oriented researchers often study one case at a time, but they may also study multiple instances of a given phenomenon (e.g., comparable instances of ethnic conflict). The distinctiveness of case-oriented research is apparent when this approach is contrasted with the variable-oriented approach, where researchers focus more exclusively on cross-case patterns, without first gaining an understanding of each case.

2. Goals of Case-oriented Research

Today social scientists tend to identify case-oriented research with specific techniques of data collection linked to the observation and analysis of singular cases (e.g., direct observation of individuals at the micro level and archival research on nation-states at the macro level). While generally useful, the identity of case-oriented research with specific techniques of data collection is unfortunate, for it obscures basic differences between case-oriented research and conventional variable-oriented research. More fundamental than differences in methods of data collection is the contrast between goals (Ragin 1987). Case-oriented strategies are distinctive in that they are centrally concerned with making sense of a relatively small number of cases, selected because they are substantively or theoretically significant in some way (Eckstein 1975). Conventional variable-oriented strategies, by contrast, are centrally concerned with the problem of assessing the relationship between aspects of cases across a large number of generic 'observations,' usually with the goal of inferring general patterns that hold for a population.

For example, a researcher might use a case-oriented approach in order to study a small number of firms in an in-depth manner. Suppose these firms were all thought to be unusually successful in retaining their best employees while at the same time investing in them and thus enhancing their potential value to competing firms. To find out how they do it, a researcher would have to conduct an in-depth study of the firms in question. By contrast, a variable-oriented researcher might study the predictors of variation in rates of 'employee retention' across a large sample of firms. Is it more a matter of firm or industry characteristics? Do these two sets of factors interact? Useful answers to these questions would be based on careful analysis of relationships between variables, using data drawn from a survey of a large number of firms—the more (and the more varied), the better.

As these two examples show, what matters most is the researcher's starting point: does the researcher seek to understand specific cases or to document general patterns characterizing a population? This contrast follows a longstanding division in all of science, not just social science. Georg Henrik von Wright argues in *Explanation and Understanding* (1971) that there are two main traditions in the history of ideas regarding the conditions an explanation must satisfy in order to be considered scientifically respectable. One tradition, which he calls 'finalistic,' is anchored in the problem of making facts understandable. The other is called 'causal-mechanistic' and is anchored in the problem of prediction. The contrast between case-oriented and variable-oriented research closely parallels this fundamental division. In the two examples just described, the first researcher uses the case-oriented approach in order to make certain facts understandable, for example, the spectacular success of a handful of firms in retaining their most valuable employees; the second researcher uses the variable-oriented approach in order to derive an equation predicting levels of retention, based on a large sample of firms, and to draw inferences from this equation to an entire population.

Once the distinction between case-oriented and variable-oriented research is established and their contrasting goals acknowledged, it is clear that the importance of techniques of data collection as bearers of the 'case-oriented vs. variable-oriented' distinction begins to fade. For example, it is clear that a researcher using case-oriented methods to study a handful of firms might benefit from conducting surveys of their employees and performing a conventional variable-oriented analysis of these data. The results of the survey would contribute to this researcher's depth of knowledge about the firms in question, just as interviewing their top executives or studying their archives would contribute useful information. Likewise, it is clear that the researcher using variable-oriented methods to predict rates of retention could benefit from interviews of top executives or personnel

officers to help interpret results found in the analysis of the predictors of retention. Still, the first researcher would remain focused on the problem of understanding the handful of firms in question, while the second would remain focused on the problem of explaining variation in retention rates across a large number of firms and making inferences from this sample to a population. The important point here is that data collection techniques *per se* are relatively neutral; what matters most is the researcher's goal and the contrasting research strategies that follow from different goals.

3. *Logic of Case-oriented Research*

The logic of case-oriented research is fundamentally configurational. Different parts of each case are understood in relation to one another and in terms of the total picture or package that they form together. The organizing idea in such research is that the parts of a case may constitute a coherent whole; that they have an integrity and coherence considered together. For example, researchers who study families often find that patterns of interpersonal accommodation are so enmeshed that a 'dysfunction' cannot be remedied without addressing many different aspects of a family all at once. Likewise, researchers who study cultures often observe that cultural traits seem to come in packages that defy easy disassembly. The 'parts' that case-oriented researchers examine can be quite varied. In macro-level research, for example, the 'parts' of a case might include institutions, path dependencies, social structures, historical patterns and trends, routine practices, cultural beliefs, singular events, event sequences, connections to other cases, the case's larger environment, and so on. What matters most is that the investigator makes sense of multiple aspects of the case in an encompassing manner, using theory as a guide.

Donald Campbell (1975) offers a rough formalization of this approach in his examination of the logic of case-oriented research. Campbell argues that at first glance the in-depth study of a single case appears to be lacking in scientific merit because there is only one case to explain and many possible explanations to choose from. He notes, however, that case-study researchers routinely 'reject' theories because they do not explain the facts of their case. Further, despite having only one case, they must often struggle to find theories that work. Why is it so difficult? The key to this puzzle is the simple fact that every theory has many implications, relevant both to features of the case in question and to causal processes and sequences operating within the case. Thus, researchers evaluate many theoretical implications relevant to their cases to see if their cases conform to expectations. Not all features of a case may be compatible with the initial theory, and the researcher must either find an alternate theory that works better, revise an existing theory, or propose an entirely new one (Walton 1992).

Campbell (1975) suggests that each separate theoretical implication can be seen as a different 'observation' for 'testing' a theory. Thus, a single case becomes many observations—some contradicting and some supporting competing theories. Collectively, competing theories and their different implications define all theoretically relevant aspects of the case in question. The researcher's task is to see which theory does the best job of explaining aspects of the case relevant not only to its own implications, but also to the implications of competing theories. Thus, after defining and selecting relevant aspects of the case, the researcher assesses the explanatory power of each theory. The theory (or combination of compatible theories) that best covers both its own implications and those of competing theories prevails and provides the basis for the investigator's representation of the case.

In the end, the researcher crafts an explanation that satisfies as many theoretical implications as possible in a coherent manner. The success of a case study hinges on (a) the number of relevant aspects of the case the researcher can encompass with his or her explanation, (b) the success of the researcher in showing that his or her portrait of the case actually makes sense of all theoretically relevant aspects, and (c) the agreement of other scholars that all relevant aspects of the case in question have in fact been addressed by the researcher in a convincing manner. Of course, the most successful case studies accomplish much more than theoretical and substantive coherence. They also may advance theory or establish important lessons for policy makers (White 1992). Still, theoretical and substantive coherence should be considered preconditions for these more ambitious goals (Yin 1994).

Campbell's (1975) argument that case-oriented researchers test multiple theoretical implications for each case underscores the configurational nature of this type of research. Because theoretical implications direct the researcher's attention to 'observations' (i.e., different aspects of a single case) that cannot be independent from one another, the researcher must make sense of them all at once, as a package. Furthermore, because different theories typically have implications about different aspects of the case, the researcher's attention is directed to a broad range of aspects, all of which may be interconnected in some way. Thus, case-oriented researchers must examine overlapping configurations of aspects as they weigh the explanatory power of competing theories.

4. *Case-oriented Research on Multiple Instances*

Case-oriented researchers sometimes conduct a series of case studies, selecting cases strategically so that the knowledge the cases generate accumulates. For example, a specific type of evidence that is not available

in one case may prompt the investigator to select for his or her next study a case that offers this evidence. In this manner, case-oriented researchers may develop theoretical arguments in much the same way that psychoanalysts build theory from the analysis of a series of patients, admitting to their practice patients that promise to enhance their thinking and their theories in some way.

This case-by-case, grounded approach can be contrasted with a research strategy that takes *multiple instances* of a phenomenon as its starting point. For example, a researcher might be interested in the formation of green parties in Western Europe and focus on a handful of parallel cases. The key contrast between the study of multiple instances and studying a series of cases is the *timing* of the selection of cases. In serial case studies, findings from one case determine the selection of the next case; in the study of multiple instances, by contrast, the researcher identifies multiple instances of a phenomenon at the outset of an investigation. Of course, the researcher's definition of the population of cases may shift in the course of the research as the investigator learns more about the phenomenon (see Ragin 1997). The key point is that the study of multiple instances begins with a preliminary specification of relevant cases.

Designs appropriate for the case-oriented study of multiple instances are varied, with the number of cases ranging from two to as many as the researcher can study in a detailed manner. Still, there are two general approaches to the study of multiple instances: 'bottom-up' and 'top-down.' Neither strategy is practiced in pure form, and there are also many mixed and hybrid strategies in between these two hypothetical extremes, just as the there are many ways to blend and mix the study of multiple instances, chosen all at once, and the grounded selection of cases one at a time.

The bottom-up approach to multiple instances attempts to develop a more or less succinct explanation of each case, in relative isolation from other cases. Of course, it is impossible for researchers to wear blinders or to forget what they have already learned, and knowledge of one case will inevitably impinge on the examination of subsequent cases. Still, the goal of the bottom-up approach is to give each case, in effect, a separate voice and thereby allow for maximum diversity. After completing each separate case study, the researcher turns to the task of comparing cases, and each case is used as a lens for viewing other cases. The goal of this cross-fertilization of case studies is to help the researcher identify all possibly relevant factors and to revise understandings of each case using insights gained from other cases. On the basis of these comparisons, the researcher builds an analytic frame for the outcome in question, which embraces all causally relevant conditions that have some degree of cross-case relevance. In the end, the researcher builds a more complete account of each case and also develops a strong basis for identifying both cross-case patterns and different types of cases (e.g., different paths to some outcome).

The top-down approach is more theory centered and more dependent upon prior knowledge and research. Based on existing theoretical and substantive knowledge, the researcher develops an analytic frame for the outcome under investigation, much like a survey researcher develops a questionnaire. This 'instrument' is then applied to each case in a relatively uniform and structured manner. That is, each case is 'interrogated' in much the same way. A central goal of the top-down approach, in addition to the goals of reliability and replicability, is to ensure that the cases selected for in-depth study have equal voice and the results are not skewed toward the most prominent cases, the cases with the most or best data, or the cases that the researcher simply happens to know best. While the analytic frame for the top-down approach is established at the beginning of the research, it is, of course, open to revision as researchers learn more about their cases.

The top-down approach also may be used when researchers seek to study a specific range of cases or specific categories of cases. In this approach, a preexisting analytic frame is used to guide the selection of instances for in-depth study. The goal is to select cases that maximize variation on causally relevant conditions while economizing on the number of cases. For example, a researcher might construct a two-by-two table cross-tabulating levels of two main causal conditions (e.g., high versus low levels of development and high versus low levels of democracy) and then seek to 'fill' each cell of this table with at least one case. The investigator would then conduct in-depth studies of the selected cases, knowing that important causal conditions are well represented in the set of instances chosen for examination.

5. *Incongruities Between Case-oriented and Variable-oriented Research*

Case-oriented research is distinct from variable-oriented research in many ways. The distinctiveness of the case-oriented approach is often overlooked, and researchers using different strategies talk past each other when discussing their methods. While some scholars emphasize philosophical differences between case-oriented and variable-oriented strategies, there are striking and profound differences between the two approaches at the 'practical level.' The practical level refers to the relatively mundane procedures researchers use when working with evidence to produce social scientific representations of what they have learned—that is, summaries of findings and patterns.

The importance of practical differences as a basis for miscommunication between case-oriented and variable-oriented researchers can be seen clearly in the contrasts between two very simple procedures: the

search for commonalities across multiple instances of a phenomenon in case-oriented research and the examination of the correlation between two variables across a sample of observations in variable-oriented research. The search for commonalities is usually part of an effort to derive limited generalizations from the in-depth study of a modest number of instances. For example, a researcher might attempt to identify common antecedent conditions for an outcome these cases share. This researcher might argue further that the common antecedents are necessary conditions for the outcome—that is, if theoretical and substantive knowledge support this interpretation. For example, a researcher conducting an in-depth study of a select group of firms—those that are able to both invest in and retain their best employees—might identify causally relevant features shared by these firms. Shared features, in turn, might point to a possible formula for success and offer lessons for other firms. The search for commonalities across a select group of cases is the most basic case-oriented strategy applied to multiple instances. By contrast, the examination of a correlation between two variables across a sample of observations in variable-oriented research is usually conducted with the goal of establishing that two conditions or aspects covary. For example, high rates of retention might be correlated with high levels of employee compensation across a sample of firms. This correlation, in turn, might be interpreted causally: the researcher would argue that one reason firms differ in their employee retention rates is that they differ in how well they compensate employees.

While both procedures seem deceptively simple and straightforward, these two ways of producing results from evidence involve sharply contrasting practical orientations toward cases, outcomes, and causal conditions:

5.1 Cases

When a researcher using variable-oriented methods computes a correlation between two variables, the relevant cases become more or less invisible and variables take center stage. Furthermore, the set of cases included in the computation must be fixed before the researcher can compute the correlation. Once this set is fixed, usually at the outset of an investigation, it is rarely altered. What matters most is that the cases (which are understood as 'observations') belong to the same general 'population' and that they be drawn from this population with an eye toward randomness or representativeness or some combination of these criteria.

In a case-oriented study of commonalities, by contrast, cases have clear identities and are usually chosen specifically because of their substantive significance or theoretical relevance. Furthermore, the set of relevant cases may shift during the investigation because the researcher may decide that one or more cases do not 'fit' with the others. For example, a researcher studying 'firms that invest in and retain their top employees' might decide that several firms originally thought to belong to this group really don't belong, and that maybe one or more that were thought to be outside this group actually belong in it. This flexibility is maintained throughout the investigation because the core concepts (e.g., 'investing in employees') may be revised as the researcher learns more about relevant instances.

5.2 Outcomes

In correlational studies researchers usually identify a 'dependent variable'—an outcome that varies across cases. Typically, such outcomes are aspects of cases that vary by level, for example, level of satisfaction, level of bureaucratization, and so on. Sometimes the outcome variable is categorical, indicating whether or not some event has occurred (e.g., filing a complaint), and sometimes it is a frequency or a rate (e.g., rate of employee retention). The important consideration, in this procedure, is that the outcome vary across 'observations.' The goal of the research typically is to explain, if possible, why cases have different values on the dependent variable. Such research is centrally concerned with the question of 'why.' For example, a researcher might seek to explain why some employees are more satisfied than average and others less so, why some firms have higher retention rates than average and others less than average, and so on.

In a case-oriented study of commonalities, by contrast, the outcome is often something that does not vary substantially across the chosen cases. In a study of firms, for example, cases might be chosen precisely because they all display the same outcome—a specific pattern of successful retention. Recall that the goal of a case-oriented study of commonalities is to identify common causal conditions linked to a specific outcome across a relatively small number of purposefully selected cases. Thus, the focus is on cases with a specific outcome, not cases that vary widely in how much they display this outcome. While the outcomes in a study of this type will not be exactly identical across cases, the researcher must demonstrate that the outcomes in the cases selected are in fact enough alike to be treated as instances of the same thing. Finally, unlike correlational studies, which are centrally concerned with the question of 'why' (as in: why some more than others?), case-oriented studies are centrally concerned with the question of 'how' (as in: how does it happen?). How do firms retain their most valuable employees?

5.3 Causal Conditions

In a correlational study, causation may be inferred from a pattern of covariation. If a variable thought to represent a cause or to be an indicator of a key causal

condition is strongly correlated with the outcome variable, then the researcher may make a causal inference. Usually, the researcher will assess the relative strength of several causal variables at the same time. The typical goal is either to find out which one explains the most variation in the outcome variable or simply to assess the relative importance of the different independent variables. In effect, variables compete with each other to explain variation. In most investigations, each causal variable is considered sufficient, by itself, for the outcome or an increment in the outcome. That is, each one is considered an 'independent' variable capable of affecting the outcome variable, regardless of the values of other causal variables.

In a case-oriented study of commonalities, by contrast, causation is typically understood conjuncturally. The goal of this type of analysis is to identify the main causal conditions shared by relevant cases. Causal conditions do not compete with each other, as they do in correlational research; they combine. How they combine or 'fit together' is something that researchers try to discern using their in-depth knowledge of cases. Because all the cases have more or less the same outcome, the usual reasoning is that the causally relevant conditions shared by cases provide important clues regarding which factors must be present to produce the outcome in question. When constructing this argument, the researcher is especially sensitive to the possibility that a given causal requirement (i.e., a necessary condition) may be met in a variety of different ways.

These and other practical differences in how researchers using case-oriented versus variable-oriented methods work with evidence to produce results provide many opportunities for disjunctures in findings. For example, from the perspective of variable-oriented work, the study of commonalities across a small number of instances is fraught with analytic sins and errors: (a) The number of cases is too small and too nonrandom to warrant any kind of inference (King et al. 1994). (b) The procedure 'selects on the dependent variable' (i.e., focuses on cases all with more or less the same value on the outcome variable). This practice may deflate otherwise robust correlations (Collier 1995, Collier and Mahoney 1996, King et al. 1994). (c) Researchers may drop cases that 'don't fit' at various stages of the analysis, which seriously undermines any effort to generalize beyond the cases that remain (King et al. 1994). (d) The most important causal factors (i.e., causally relevant commonalities) do not vary and thus are impossible to assess, and so on (Lieberson 1991, 1994; Goldthorpe 1991, 1997).

Likewise, from the perspective of case-oriented work, the examination of the correlation between a causal and outcome variable across many cases is fundamentally flawed: (a) Typically, there are so many cases that there is no way for the researcher to know if they are all really comparable and thus belong together in the same analysis. (b) Fixing a population or sample boundary also fixes the assumption of homogeneity, which usually is not warranted. Investigators should be able to redefine the set of relevant cases as they learn more about them (Ragin 1997). (c) It is difficult to determine how something 'comes about' by comparing cases with different levels of the outcome. The partial instances (i.e., those with lower scores on the outcome variable) are likely to provide many false leads (Lijphart 1971, 1975, Ragin 1997). (d) It is pointless to try to isolate the 'independent' effect of any causal condition when several factors usually must combine for a particular outcome to occur (George 1979), and so on. In short, at a practical level the two approaches seem almost antithetical. It is no wonder that findings diverge and researchers often talk past each other (Rueschemeyer 1991; Ragin 1997).

6. *Bridging Case-oriented and Variable-oriented Research*

While the gulf between case-oriented and variable-oriented research seems great, it is possible to span it. Consider the following scenario: a researcher using case-oriented methods studies several instances of an outcome (e.g., a small number of firms that are very successful in investing in and retaining top employees) documents causally relevant commonalities, and then constructs a general, composite argument about how these firms do it. This argument leads to four specific recommendations (which might be labeled X_1 to X_4) based on the observed commonalities. A second researcher reads the report of this study and decides to evaluate it with a large sample using variable-oriented methods. This researcher collects information on a random sample of firms and finds that, as independent variables, X_1 to X_4 do *not* distinguish more successful from less successful firms, using various measures of employee retention. In short, the second researcher shows that there is no statistically significant difference in the retention rates for firms with and without these four aspects, considering these aspects one at a time or in an additive, multivariate equation.

What went wrong? Usually, the researcher using variable-oriented methods will claim that the first researcher's 'sample' was 'too small' and 'unrepresentative.' Thus, the identification of X_1 to X_4 took advantage of specific aspects of the selected cases. The first researcher might counterattack by arguing that causally relevant commonalities identified through in-depth study are very difficult to represent as 'variables,' and that the second researcher's crude attempt to operationalize them fell far short. Indeed, the first researcher might argue that it would take in-depth knowledge of each firm included in the variable-oriented study to capture these conditions appropriately and contextually.

These criticisms and countercriticisms are quite common. However, the incongruity between the two

hypothetical studies can be resolved without resorting to derogation. The first researcher in this example—the one using case-oriented methods—selected on instances of the outcome and identified four causally relevant conditions shared by the firms in question. In essence, this researcher worked backwards from the outcome to causes and thus identified potential *necessary* conditions for the outcome (Ragin 2000). Are these conditions truly necessary? In part, this is an empirical question. To gain confidence, the researcher should examine as many instances of the outcome as possible, to see if they agree in displaying these four causally relevant conditions (or their causal equivalents). But it is also a question about existing knowledge. Is the argument that these four conditions are necessary consistent with theoretical and substantive knowledge? Do they make sense as necessary conditions? If the researcher's finding is consistent with existing substantive and theoretical knowledge, then the argument that these four conditions are necessary is strengthened.

How should the second researcher—the one using variable-oriented methods—respond to the argument that these four factors are necessary conditions? At a more abstract level, the specification of necessary conditions is relevant primarily to the identification of cases that are candidates for an outcome. Cases cannot be candidates for an outcome if they do not meet the necessary conditions. But many cases may meet the necessary conditions for an outcome and still not exhibit the outcome, because they lack additional conditions which, when combined with the necessary conditions, establish sufficiency for the outcome. In fact, the cases displaying the outcome may be only a small minority of those that meet the necessary conditions. Thus, while there are clear gains from specifying necessary conditions, as in the hypothetical case-oriented study, the identification of causally relevant commonalities shared by instances of an outcome does not establish the conditions that are sufficient for an outcome. Thus, the variable-oriented researcher's finding that these four conditions do not distinguish low-retention firms from high-retention firms across a large sample of firms does *not* directly challenge the argument that these conditions are necessary.

The variable-oriented analysis of these four conditions across a large sample of firms is much more directly relevant to their *sufficiency*. To show that these conditions are jointly sufficient for the outcome, it would be important to demonstrate that when these four conditions are combined the outcome follows. In other words, the variable-oriented researcher could evaluate sufficiency by examining the correspondence between the combination of the four causes, on the one hand, and the outcome, on the other, across a large sample of firms. Still, this analysis would be an evaluation of sufficiency, not necessity, and the results of this analysis would not bear in a direct manner on the claim of necessity made by the case-oriented researcher.

This sketch identifies only one of several ways to span case-oriented and variable-oriented inquiry. The general and most important point is that it *is* possible to join these two approaches if researchers are careful to distinguish between necessity and sufficiency and to separate the analysis of these two aspects. More generally, this discussion underscores the distinctiveness of case-oriented research, especially the case-oriented study of multiple instances of an outcome.

See also: Case Study: Logic; Case Study: Methods and Analysis; Classification: Conceptions in the Social Sciences; Configurational Analysis; Explanation: Conceptions in the Social Sciences; Person-centered Research; Psychotherapy: Case Study; Single-case Experimental Designs in Clinical Settings; Single-subject Designs: Methodology; Time Series: General; Triangulation: Methodology

Bibliography

Campbell D T 1975 Degrees of freedom and the case study. *Comparative Political Studies* **8**: 178–93

Collier D 1995 Translating quantitative methods for qualitative researchers: The case of selection bias. *American Political Science Review* **89**: 461–6

Collier D, Mahoney J 1996 Insights and pitfalls: Selection bias in qualitative research. *World Politics* **49**: 56–91

Eckstein H 1975 Case study and theory in political science. In: Greenstein F I, Polsby N W (eds.) *Handbook of Political Science, Volume 7, Strategies of Inquiry*. Addison-Wesley, Reading, MA

George A 1979 Case studies and theory development: The method of structured, focussed comparison. In: Lauren P G (ed.) *Diplomacy: New Approaches in History, Theory and Policy*. Free Press, New York

Goldthorpe J 1991 The uses of history in sociology: Reflections on some recent tendencies. *British Journal of Sociology* **42**: 211–30

Goldthorpe J 1997 Current issues in comparative macrosociology. *Comparative Social Research* **16**: 1–26

King G, Keohane R O, Verba S 1994 *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press, Princeton, NJ

Lieberson S 1991 Small N's and big conclusions: An examination of the reasoning in comparative studies based on a small number of cases. *Social Forces* **70**: 307–20

Lieberson S 1994 More on the uneasy case for using Mill-type methods in small-N comparative studies. *Social Forces* **72**: 1225–37

Lijphart A 1971 Comparative politics and comparative method. *American Political Science Review* **65**: 682–93

Lijphart A 1975 The comparable cases strategy in comparative research. *Comparative Political Studies* **8**: 158–75

Ragin C C 1987 *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. University of California Press, Berkeley, CA

Ragin C C 1997 Turning the tables: How case-oriented research challenges variable-oriented research. *Comparative Social Research* **16**: 27–42
Ragin C C 2000 *Fuzzy-Set Social Science*. University of Chicago Press, Chicago
Rueschemeyer D 1991 Different methods—contradictory results? Research on development and democracy. In: Ragin C C (ed.) *Issues and Alternatives in Comparative Social Research*. E. J. Brill, Leiden, The Netherlands
Walton J 1992 Making the theoretical case. In: Ragin C C, Becker H S (eds.) *What Is a Case? Exploring the Foundations of Social Inquiry*. Cambridge University Press, New York
White H 1992 Cases are for identity, for explanation, or for control. In: Ragin C C, Becker H S (eds.) *What Is a Case? Exploring the Foundations of Social Inquiry*. Cambridge University Press, New York
Wright G H von 1971 *Explanation and Understanding*. Cornell University Press, Ithaca, NY
Yin R K 1994 *Case Study Research: Design and Methods*, 2nd edn. Sage, Thousand Oaks, CA

C. C. Ragin

Cassandra/Cornucopian Debate

The Cassandra/Cornucopian debate is one of many names given to an argument between two extreme positions on the prospects for human society and the environment in the face of population and economic growth. So-called Cassandras (after the Greek princess given the power of prophecy but denied the power of persuasion) contend that unchecked growth in numbers of people and material consumption rates will inevitably lead to environmental and social catastrophe. This group is alternately called doomsayers, pessimists, catastrophists, or neo-Malthusians and stereotypically is associated with ecologists and environmentalists. Cornucopians (after the Greek legend of the 'horn of plenty' that overflows with whatever its owner wishes for) believe that human ingenuity and free markets will allow the human species to adapt to any conceivable pressures caused by growth of the human enterprise. They are referred to alternately as optimists, panglossians, or exemptionists, and stereotypically are associated with free-market economists. The debate originated at least as far back as Malthus in the eighteenth century, flared up with special intensity in the late 1960s and 1970s over the issue of exhaustible natural resources, and continues in more subdued form today as reflected in the issue of global environmental change.

1. Historical Roots

The roots of the Cassandran position extend as far back as the Reverend Thomas Malthus, who in 1798 published the first and most pessimistic version of his tract *Essay on the Principle of Population* (Malthus 1798). In it, he argued that population tends to grow at a geometric rate (i.e., a constant percent increase, but ever-larger absolute increments) while necessities (particularly food supply) tend to grow at an arithmetic rate (constant absolute increments). Therefore, sustained population growth would inevitably outstrip its means of support, unless subjected to certain 'checks.' In the first edition of the essay, these 'positive checks' were war, pestilence, disease, and famine. The second edition introduced the possibility of 'preventive checks,' which included moral restraint, celibacy, late marriage, and abstinence. If the preventive checks did not slow population growth before it overran the limits of the land to support it, then the positive checks would come into play.

Malthus principally was interested in whether aid to the poor could be expected to lift them out of poverty (he concluded that it could not). Concern over the consequences of population and economic growth for human welfare and the environment has waxed and waned several times since the publication of Malthus's essay. Particularly relevant to contemporary scholarship is the growth of neo-Malthusianism since World War II. Luten (1980) traces this recent history through concern over population growth and its impact on resource use and economic growth in the 1940s and 1950s (see, e.g., Osborn 1953, Coale and Hoover 1958), to pollution and pesticides in the 1960s (notably Carson 1962), and back to population and resources in the late 1960s and 1970s (Paddock and Paddock 1967, Ehrlich 1968, Meadows et al. 1972).

The Cornucopian position has roots that extend back to the mercantilist school of economic thought in the sixteenth century, which emphasized the positive effect of population growth on economic production. It was the mercantilist economist Jean Bodin who claimed 'There is no wealth but men.' More recently, Boserup (1965, 1981) argued that increased population size and its associated demand for food itself induced new agricultural production techniques and increased density facilitated the flow and application of new knowledge. In this way a growing population aided rather than hindered development.

These points of view and others have been reflected and expanded on in Cornucopian responses to the post-war rise of neo-Malthusian concern. For example, Barnett and Morse (1963) argued that technological progress can more than offset resource scarcity, while Simon (1981, 1996) and Kahn et al. (1976) have been the most visible Cornucopian standard-bearers in recent decades.

2. Cassandras

The most recent crest of neo-Malthusian concern occurred in the 1960s and 1970s, driven by a number of writers and researchers concerned with what they

saw as the inevitability of rising levels of pollution and increasingly scarce food supply and nonrenewable natural resources. Most of these arguments pointed to material consumption and especially population growth as the key drivers of these problems, and advocated immediate measures to reduce consumption and control population. For example, Hardin (1968) argued that population growth inevitably leads to the degradation of common property resources, the root of many environmental problems, and that the only solution to the 'population problem' was 'mutual coercion, mutually agreed upon,' that is, societies must agree to enforce limits upon their own reproductive behavior.

The most well-known advocate of the Cassandra position was Ehrlich, who in 1968 published *The Population Bomb*, a short popular book declaring that the struggle to feed humanity was over and had been lost; hundreds of millions were destined to starve in the coming decades. It foresaw a future so dire that it advocated consideration of a system of 'triage' in foreign aid policy first recommended by Paddock and Paddock (1967). The system would cut off aid to countries such as India deemed beyond help due to rapid population growth and dim prospects for improvements in agricultural production and would focus instead on countries such as Pakistan that might possibly avoid massive starvation with the right kind of assistance.

In 1972 the sense of impending doom was heightened further by the publication of *The Limits to Growth* (Meadows et al. 1972), a book by a group of researchers at the Massachusetts Institute of Technology that argued that the human population was approaching the limit of the planet's ability to sustain it, and without fundamental societal changes we would soon exceed that limit, triggering a dramatic increase in death rates, catastrophic environmental degradation, and social chaos. Their conclusion received an enormous amount of attention, in part because it was based on one of the first attempts to create a computer model of the world as a single system. They subjected their 'World3' model to an extensive sensitivity analysis, but in all cases it produced an 'overshoot and collapse' pattern: a surge in population followed sooner or later by a crash. The robustness of the result seemed to lend great credibility to their conclusions.

3. *A Storm of Criticism*

Both *The Population Bomb* and *The Limits to Growth* sold millions of copies, and this wave of academic and popular attention to a presumed approach to the planet's carrying capacity generated a storm of criticism. *The Limits to Growth* itself received special attention by academic critics (Maddox 1972, Cole et al. 1973, Nordhaus 1973, Sanderson 1994), perhaps because its basis in computer modeling gave it an aura of academic respectability that more popular treatments lacked. Criticism covered a wide range of issues, but there were several common themes centered on the perceived lack of economic principles.

3.1 *Technical Progress*

Neoclassical economics views scarcity not as absolute but as relative and as a function of technology. Technical progress lowers the cost of making a resource more accessible. For example, improvements in irrigation systems make it cost effective to expand the area of irrigated land, essentially increasing the total land available for cultivation. Thus, expressing limits in fixed physical units, as was done in *World3* and in other carrying-capacity arguments, left out the crucial role of technology in expanding the pool of resources available to society.

3.2 *Adjustment Mechanisms*

Many critics pointed out the lack of accounting for well-known adjustment mechanisms in response to scarcity, especially price systems. In an ideal market economy, as a resource becomes relatively more scarce, its price rises. This rising price induces a host of adjustments: recycling; conservation; the development of better methods of discovery, extraction, and production; and the substitution of other goods, all of which have the effect of essentially increasing supply. An often-quoted example is that although global oil consumption rates increased between the 1970s and 1990s, oil reserves grew rather then declined over this time period—a result of price-induced improvements in exploration, extraction, and efficiency.

The *World3* model did not include a price system at all. It also aggregated the world into a single region, eliminating trade as a possibility, and aggregated over many goods so that substitution was not possible. Thus, many economists put little stock in its results, or in the conclusions of arguments that did not account for the ability of a market economy to respond to scarcity.

3.3 *Methodological Weaknesses*

Several critics focused on particular aspects of the methodology of *The Limits to Growth* study. Relationships between variables in the model may have been plausible, but were based on little more than the authors' intuition, ignoring theory and evidence in economics, demography, and other fields. Also, several studies showed that the 'overshoot and collapse' behavior of their model was by no means inevitable. If changes in variables were made slowly rather than

suddenly, the model produced steady growth (Cole et al. 1973). If the prices were included or the possibility of substitution was allowed for, or alternative assumptions about population growth or technical change were made, model outcomes changed drastically, in some cases producing continuous growth (Nordhaus 1973).

4. *Cornucopians and their Critics*

In many ways the Cornucopian credo is neoclassical economic reasoning taken to an extreme. Simon was the best-known Cornucopian. In *The Ultimate Resource* (1981; a 1996 follow-up did not alter the basic message) he argued in essence that on balance people create more than they destroy, so that the more people, the better the world. While he agreed that larger populations can create short-run scarcity and environmental degradation, he believed that these problems presented challenges to which society would always adapt, improving itself in the process. In the long run, society is better off—prices are lower, the environment is cleaner, and consumption is greater—than if the problems had never arisen.

Critics of the Cornucopian position seized on several perceived weaknesses. In general, it is often argued that the theory is not testable: any scarcity or environmental degradation that is encountered can be explained away as either a market imperfection or as a short-run problem that will in fact lead to a long-run benefit. By presuming to explain everything, it explains nothing. More specific criticisms have focused on several main points.

4.1 *Discontinuities*

Many critics argue that Cornucopians have a blind faith that the future will repeat the past. Because resource prices have generally fallen, and by some measures and in some places the environment has improved, it is presumed that these trends will always continue. Cornucopians argue that there is no reason to expect the future to be different from the past in terms of society's ability to adapt to and benefit from scarcity-induced challenges. Cassandrans counter that there is every reason to expect the future to be different, since gradual increases in environmental degradation driven by the increasing scale of human impact are likely to eventually push ecosystems past thresholds beyond which decline will be rapid and perhaps uncontrollable.

4.2 *Insufficient Accounting for Natural Science*

The Cornucopian position is often accused of focusing on improvement in material consumption and human welfare at the expense of associated degradation of the environment. Many ecologists and other natural scientists considered the treatment of environmental impacts by Simon and others to be perfunctory and unconvincing at best, and deliberately misleading at worst. For example, while optimists often focused on improvements in the levels of pollution of, for example, DDT or PCBs (improvements that, Cassandras often point out, were driven in large part by the concern of environmentalists), they usually said little about larger issues such as climate change or biodiversity loss. Critics accuse them of being quick to dismiss scientific consensus when convenient to their argument, and of suggesting that such consensus was a conspiracy to obtain more research money.

4.3 *Unfounded Assumptions*

Simon's argument was often attacked for its reliance on the assumption of a rather mechanical link between human population numbers and the assumed stock of knowledge and other forms of social capital. Sanderson (1980) points out that Simon assumes social capital grows automatically with the size of the population, and that it has an unusually large effect on production. To some extent, population growth's positive long-run impact on economic production is built into Simon's model by assumption.

5. *Winners and Losers*

Although the debate is ongoing, it is possible to assess the accuracy of some of the claims made so far by the two sides. For example, Malthus has turned out to be wrong so far, because he failed to anticipate the opening up of new agricultural lands, the scope for improvements in agricultural technology, storage, and transport, and the decline in Western fertility. Similarly, the predictions of mass starvation made in the 1960s and 1970s never materialized; the Green Revolution in agriculture raised yields much more than anticipated, and developing country population growth slowed as a transition to lower fertility was begun. Nor did protracted shortages in natural resources materialize. This outcome was highlighted by a much-publicized bet made between Simon and Ehrlich in 1980. The two wagered over whether the prices of five metals (chosen by Ehrlich) would rise over the decade (indicating scarcity); Simon won on all counts as the prices of all five fell. Furthermore, as of yet there are no signs of *Limits to Growth*-type scenarios coming to pass. Supporters of the project emphasize that independent of the value of their scenarios, there was value in the attempt to model the world as a system, however crudely, as a first step toward an important new direction of research. However, Sanderson (1994) points out that *World3*

was an 'evolutionary dead-end'—even the authors' own follow-up study in 1992 presented no substantial advance over the original model.

On the Cornucopian side, Simon's argument that population growth is unambiguously positive for long-term economic growth was certainly influential in countering a focus on the negative consequences of growth, but it was not adopted wholesale by economists. Mainstream economics has come to no strong conclusion on whether population growth influences economic growth. More recently, studies in Asia suggest that age structure effects of fertility reduction in some settings can stimulate growth if combined with other economic and political measures.

Some other Cornucopian articles of faith have not fared well. Nuclear energy did not solve the energy problem, as several technological optimists predicted, and while the oil price increases of the 1970s have since abated, the fossil-fuel-driven problem of global climate change has not gone away.

6. *Resolution?*

Attempts have been made to resolve the Cassandra/Cornucopian debate in a number of ways. Some conclude that the truth lies somewhere in between the two extremes: the outlook is neither as rosy nor as dismal as these two camps believe. Others conclude that each may have particular aspects of the story right. For example, Cornucopians may be right that we can feed an expanding population without massive famine, but Cassandras may be correct in emphasizing that this will come at great environmental cost if current trends continue.

In the absence of resolution, lessons drawn from the debate could facilitate progress on contemporary issues related to global environmental change. For example, the climate change issue has given rise to a new group of models, called integrated assessment models, which share some features with the early world-systems models: they are often global in scope, they seek to incorporate feed-backs between society and the environment, and they cover long time-scales. History offers lessons on the dangers of aggregation, the importance of grounding relationships in theory and data, and the importance of transparent assumptions and treatments of uncertainty. In addition, the debate over the Earth's capacity to sustain society has not disappeared; current manifestations include the issue of human-induced loss of 'ecosystem services' such as soil aeration, pollination, air and water purification, and element cycling. Key economic concepts left out of earlier arguments are being confronted: ecologists argue that there are no practical substitutes for ecosystem services, their supply is fixed, and their constraint on production seems to be near enough to matter. In this case, the history of the Cassandra/Cornucopian debate may be contributing to a more productive discussion.

See also: Environmental and Resource Management; Environmental Policy; Environmental Vulnerability; Environmentalism: Philosophical Aspects; Environmentalism, Politics of; Environmentalism: Preservation and Conservation; Human–Environment Relationships; Malthus, Thomas Robert (1766–1834); Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Sustainable Development

Bibliography

Barnett H J, Morse C 1963 *Scarcity and Growth*. Johns Hopkins Press for Resources for the Future, Baltimore, MD

Boserup E 1965 *The Conditions of Agricultural Growth*. Aldine, Chicago

Boserup E 1981 *Population and Technological Change: A Study of Long-term Trends*. University of Chicago Press, Chicago

Carson R 1962 *Silent Spring*. Houghton Mifflin, Boston

Coale A J, Hoover E M 1958 *Population Growth and Economic Development in Low-income Countries*. Princeton University Press, Princeton, NJ

Cole H S D, Freeman C, Jahoda M, Pavitt K L R (eds.) 1973 *Models of Doom: A Critique of The Limits to Growth*. Universe Books, New York

Ehrlich P R 1968 *The Population Bomb*. Ballentine Books, New York

Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8

Kahn H, Brown W, Martel L 1976 *The Next 200 Years: A Scenario for America and the World*. Morrow, New York

Luten D B 1980 Ecological optimism in the social sciences: The question of the limits to growth. *American Behavioral Scientist* **24**(1): 125–51

Maddox J R 1972 *The Doomsday Syndrome*. McGraw-Hill, New York

Malthus T R 1798/1967 *Essay on the Principle of Population*, 7th edn. Dent, London

Meadows D H, Meadows D L, Randers J 1992 *Beyond the Limits*. Chelsea Green Publishing, Post Mills, VT

Meadows D H, Meadows D L, Randers J, Behrens W W III 1972 *The Limits to Growth*. Universe Books, New York

Nordhaus W D 1973 World dynamics: Measurement without data. *The Economic Journal* **83**(332): 1156–83

Osborn F 1953 *The Limits of the Earth*. Little, Brown, Boston

Paddock W, Paddock P 1967 *Famine—1975!* Little, Brown, Boston

Sanderson W C 1980 Economic-demographic simulation models: A review of their usefulness for policy analysis. Research Report. International Institute for Applied Systems Analysis, Austria, RR-80-14

Sanderson W C 1994 Simulation models of demographic, economic, and environmental interactions. In: Lutz W (ed.) *Population–Development–Environment: Understanding their Interactions in Mauritius*. Springer-Verlag, Berlin, pp. 33–71

Simon J L 1981 *The Ultimate Resource*. Princeton University Press, Princeton, NJ
Simon J L 1996 *The Ultimate Resource 2*. Princeton University Press, Princeton, NJ

B. O'Neill

Caste

The career of caste as an anthropological concept provides a fascinating terrain on which to examine how anthropological objects come to be invented and stabilized, and the relation they bear to patterns of sociality, needs of government, and production of knowledge. Many textbooks in sociology and social anthropology represent caste as a fundamental institution of Indian society. In this view castes are ranked in a hierarchical order and are governed by rules of endogamy, commensality, and purity–pollution that put strict limits on exchange of women and food between castes. The hierarchical order, however, permits the flow of women from lower to higher castes through marriage and the flow of food from higher to lower castes. The caste system is seen to correspond to a rough division of labor, though it is recognized that certain occupations such as agriculture cut across different castes. This textbook picture of the caste system, though influential, has come under serious scrutiny in recent years as being not only ahistorical but also ignoring the power/knowledge axis in the production of social science concepts.

1. *Caste as the Ideology of Indian society*

Louis Dumont's (1980) *Homo Hierarchicus* has long been regarded as an outstanding contribution to the understanding of caste. Dumont argued that principles of hierarchy and holism were central for explaining the caste system. The principle of hierarchy in India, he proposed, was based upon the religious opposition between pure and impure—pollution incurred in the biological processes of life and death was removed in India not through processes of reciprocity (I bury your dead—you bury mine) but through principles of hierarchy. The task of removal of pollution was assigned to the lower castes who became permanently imbued with it. Thus the separation between castes as well as their hierarchical ordering could be derived from the opposition between pure and impure. The scheme had the simplicity and elegance to make the bewildering diversity of Indian civilization immediately knowable, especially to Western readers.

Dumont's characterization of Indian society has been challenged on the ground that what he saw as a timeless ideology was itself a result of certain practices of classification and enumeration instituted in the context of colonial administration that gave a dominant place to Brahmanical texts as representatives of Indian society. An important intervention made by McKim Marriott (1990) needs mention here. Marriott provided a significant alternative to Dumont's formulation, arguing that different transactional strategies defined the position of different castes—it was not a simple case of hierarchy versus equality but rather of a universe governed by a complex set of rules and strategies regarding matching, mixing, and marking through which different regional and local configurations of castes were generated. What was at stake for both Dumont and Marriott, despite their differences, was the representation of India as the 'other' of the modern West. They were much less interested in either the concrete historical processes through which institutions were formed or the contemporary changes in the caste system. It is instructive to compare this with the way that caste was rendered in the work of the Indian anthropologist, M. N. Srinivas. Srinivas's stake in the local and his deep concern with the way caste was shaping Indian democratic politics distinguish him from these authors.

2. *The Stake in the Local*

The decade of the 1950s saw the emergence of 'village studies' as an important genre of anthropological writing in India. For Srinivas the village was an indispensable site for fieldwork because of the importance he attached to what he called the 'field-view' as distinct from the overly textual view of the caste system. The two important concepts that Srinivas (1971) formulated with far reaching consequences for the study of caste were '*sanskritization*' and 'dominant caste.'

Srinivas contended that a preoccupation with textual sources had led many to assume that the caste system was a rigid hierarchical system. Attention to historical and ethnographic processes on the ground, on the other hand, showed that while individual mobility was severely restricted in the caste system, group mobility had always been possible for castes in the middle rungs through the process of *sanskritization*. This process referred to the ability of a caste group, which had achieved economic or political mobility, to stake a claim to a higher status by recasting its customs and rituals to correspond more closely to Brahmanical rituals and practices. The process of change in ritual practices and social customs was not open to everyone. However, when a caste was successful in acquiring a measure of economic or political power through physical mobility, opening up of markets, or changed political alignments, it could translate this into claims to a higher caste status. With the decline in traditional kingship brought on by the colonial rule, such functions were shifted to colonial authorities. It is one of the exquisite ironies of the colonial rule that when caste began to be recorded systematically in the census, it became a major source

for claiming of higher status by various caste groups who adduced their *sanskritic* practices or their myths of origin as evidence of their higher status.

Unlike the concept of *sanskritization*, which assumed an interaction between local and regional levels, the concept of the dominant caste emerged out of an understanding of local and especially agricultural society. Srinivas (1971) argued that the ritual hierarchy of caste notwithstanding, the dominant role in village life was played by landowning peasant-proprietor castes who were rarely Brahmins. The relation between these rich peasant castes and the other castes in the village was characterized as a form of patron–client relations. Thus Srinivas argued that the horizontal solidarity of caste expressed by endogamy and commensality was counterbalanced by the vertical solidarity observable at the village level through patron–client relations. Many institutions such as fictive kinship within the village, ritual processes for creation of boundaries, and the division of labor and economic exchange, created the village as an economic, political, and ritual unit that established solidarities cutting across caste. As the general understanding of locality in the field of anthropological theory deepened, interest shifted to the mapping of specific histories through which localities were created. The search for overarching structural principles seemed less fruitful than understanding how the momentous changes that were taking place in Indian society impacted upon caste, at the local, regional, and national levels.

3. *Colonial Constructions*

Once the search for enduring structural principles gave way to historically grounded work, the colonial archive became an important source for understanding the nature of colonial rule and how it transformed the institution of caste. The processes of enumeration and classification played a particularly important role in this process. Although the predecessor states of the British in India did have apparatuses for counting, classifying, and controlling populations, these were tied to specific needs of the state such as revenue collection, or rising of temporary armies. The British colonial state instituted a new way of collecting information in the form of maps, settlement reports, revenue records, statistical information, censuses, folklore, narratives, to name a few.

This new form of governance, or 'rule-by-records and rule-by-reports' in the felicitous phrasing of Richard S. Smith (1986) had a decisive influence on the shaping of caste identities in the twentieth century. In the processes of classifying and enumerating the population, the British did not start with caste as if it were a natural category found on the ground. In the early phases of colonial rule the emphasis was on cadastral control. Statistics on land ownership, tenancy, crop production, and instruments of agricultural production were geared towards standardizing methods of revenue collection and it was by no means obvious that caste rather than village would be the obvious unit for the organizing of data. According to Smith it was only around 1850 that the census in the case of Punjab was transformed from an instrument of tax to an instrument of knowledge. Earlier census reports were more pragmatic and localistic in orientation. There was considerable tension between the concerns of centrist census officials collating data in an encyclopedic manner and the local officials who were concerned with recording the nature of social groups and categories for more practical purposes such as collection for revenue or settlement of disputes (see Appadurai 1996). In order to understand the concerns of the centrist census officials who recorded information on caste for which there could have been no practical use, it is necessary to turn to the scientistic notions of race in this period.

In shifting to castes as the most natural groups around which information was to be organized on Indian society, British officials relied on their notions of race and physical types. Herbert Risley ([1908], 1969) was the most vocal proponent of using anthropometrical measures in conducting the ethnographic survey of India because, according to him, anthropometry yielded particularly good results in India by reason of the caste system, which allowed marriage only within a limited circle. Relying on such symbolism as that of dark versus light skin color and shape of nose and jaw, the category of caste was thus collapsed with that of race. Ghurye (1932) was the first Indian sociologist to criticize this view of caste and to challenge its political implications.

It is not anyone's case that the process of recording caste created this institution *ex nihilo*. What it did was to freeze the ongoing processes by which caste came to be solidified in the official imagination, and to generate the conception of community as enumerative, which had a strong influence on processes of political representation (Cohn 1984). It is important to underscore the fact that the census, gazetteers, reports, and other such forms of knowledge came to represent the power of official discourse to name and fix the status of caste groups in local imaginaries. Many caste groups began to see the census as the source for claiming higher status. Thus census commissioners were besieged with petitions challenging a particular status ascribed to a caste.

It is also interesting to see that the theories of the racial origin of castes had a serious impact on social and political mobilization, as in the anti-Brahmin movements in the South of India and in the mobilization of the untouchable castes under the leadership of Ambedkar in Western India—both of which used the idea of original settlers versus Aryan invaders. The discursive formations around caste quotas in educational institutions, government jobs, and reserved constituencies, as well as the politics of mobilization

around these issues, show the interaction and circulation of categories between popular imagination and official discourse (Dirks 1992).

4. The Contemporary Context

An important paradox about caste in public life in India is that while the Constitution of India recognizes only individuals as bearers of rights and duties, and bans any discrimination based on caste or religion, the processes of democratic politics with the imperatives of grass root mobilization have created new arenas for the evolution of caste-based politics (Béteille 1996, Kothari 1970). Interestingly the legal prohibitions on caste and the political mobilization around caste reflect the contrary pulls of formal legal justice system and the imperatives of representational politics in contemporary India. This can be illustrated with the example of 'untouchability' and its contradictory careers in law, politics, and everyday life (Galanter 1972). It is not that a similar exercise cannot be done on other castes, but the example of untouchability has a particular resonance because of its unique association with the caste system.

While it could conceivably be argued that some form of caste-based discrimination is to be found in all the legal sanskritic texts, recent examination of the genealogy of the terms through which untouchability makes an appearance in discourse shows the importance of political and social processes in the negotiation of group identities in democratic societies (Charsley 1996). The term 'untouchability' is ascribed to Sir Herbert Risley ([1908], 1969) and was part of his effort to classify and rank castes in the subcontinent as a whole. While the category *shudra* occurs in the Sanskrit texts and is sometimes taken to be equivalent to 'untouchables,' its referents are varied—ranging from kings, powerful landowning castes, to castes with extreme disabilities. Prior to Risley, compilers of district gazetteers and state census reports had experimented with other terms, such as depressed classes, depressed castes, *panchamas*, *pariahs*, etc.

The use of the term 'untouchable,' in public life, owed much to the reformist politics of early twentieth century, especially to Gandhi's politics of reform and the agenda for the abolition of untouchability in the nationalist movement (Gandhi 1954). In 1931 Gandhi adopted the term *Harijan* (people of god) and increasingly substituted it for other terms in his writings. While the prototypical Harijan for Gandhi was a member of the *bhangi* caste who cleaned lavatories and was thus rendered 'unclean,' his major political opponent, Ambedkar, who himself came from the Mahar caste of Bombay Presidency, was much more interested in forging political alliances between the major agricultural dependent castes, whose low status came from their dependency rather than their polluting occupations. Yet he preferred to retain the label 'untouchable' in his politics (Charsley 1996). In the embroilment of nationalist politics with caste mobilization, Gandhi's strategy was to retain untouchables within the fold of Hinduism while Ambedkar's attempt was to forge a new identity for them, but within the Indian civilizational context—hence the choice of neo-Buddhism as the new religion for the erstwhile untouchables (Ambedkar 1946).

The term 'untouchable,' like other such generic terms masks considerable local heterogeneity. Its importance lay in the binary division between untouchable and nonuntouchable castes created through the bureaucratic imperative of creating a single category for purposes of reservations. Under the Government of India Act of 1935, the term 'scheduled castes' replaced the earlier category of depressed classes. Though untouchability was to be the criterion for inclusion of castes into the list of scheduled castes, the names of the castes that were finally included in the state lists formed a kind of unity only through a 'common relationship their members have with government' (Dushkin 1972, p. 166). The fact that the category of untouchability was basically a twentieth century creation does not detract from either the experiences of oppression of castes grouped under this category, or the importance of the grounding of struggles for greater equality under identity politics. In fact, the creativity of the new social movements can be seen in the emergence of the term *dalit* (literally 'downtrodden') among the neo-Buddhists and scheduled castes in Maharashtra, which has become now the most commonly accepted term of self-reference by such groups. The most interesting feature of this phase of the movement of *dalits* is the emphasis not only on political action but also on representation of the experiences of untouchable castes through literature, life histories, and collection of oral literature as a critique of caste society (Murugkar 1996, Omvedt 1995).

An interesting question arises as to whether politics is the only domain within which the legacy of oppression and humiliation is articulated in contemporary India. How are caste identities performed in everyday life? Khare (1984) had suggested that instead of playing the impure foil to the Brahmin, the untouchables conduct themselves even at the community level as civilizational critics, showing a complex relationship with caste ideology. Other important ethnographies of caste show how caste identity is performed in relation to everyday activities of work and worship; how elaborate strategies for management of caste identities are evolved, and how these performances come to be experienced as forms of embodiment.

Debates on caste have shown the intricate relations between forms of esthetic, political, and social representations of caste. Simultaneously much of the scholarly agenda of research in India in recent years

has been influenced by the public controversies on such issues as caste-based reservations or reintroduction of caste-based enumeration for the next decennial census. New patterns of sociality are evident in ways in which caste identity is performed in the context of family, community, economy, and polity which reconceptualize caste not as a sign of the *exotic* but of the *contemporary* repositioning of the subjects and objects of anthropology in new ways.

See also: Area and International Studies: Development in South Asia; Class: Social; Labor, Division of; Social Stratification; South Asian Studies: Culture; Subaltern History; Subaltern Studies: Cultural Concerns

Bibliography

Ambedkar B R 1946 *What Congress and Gandhi Have Done to the Untouchables*. Thacker, Bombay, India

Appadurai A 1996 Number in the colonial imagination. In: Appadurai A *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN, pp.114–39

Béteille A 1996 Caste in contemporary India. In: Fuller C J (ed.) *Caste Today*. Oxford University Press, Delhi, India

Charsley S 1996 'Untouchable': What is in the name? *Journal of Royal Anthropological Institute* **2**: 1–23

Cohn B S 1984 The census, social structure and objectification in South Asia. *Folk* **26**: 25–49

Deliege R 1992 Replication and consensus: Untouchability, caste and ideology in India. *Man* **27**: 155–73

Dirks N B 1992 Castes of mind. *Representations* **37**: 56–78

Dumont L 1980 *Homo Hierarchicus: The Caste System and its Implications*. University of Chicago Press, Chicago

Dushkin L 1972 Scheduled caste politics. In: Mahar J M (ed.) *The Untouchables in Contemporary India*. University of Arizona Press, Tucson, AZ

Galanter M 1972 The abolition of disabilities: Untouchability and the law. In: Mahar J M (ed.) *The Untouchables in Contemporary India*. University of Arizona Press, Tucson, AZ

Gandhi M K 1954 *The Removal of Untouchability*. Navajivan Press, Ahmadabad, India

Ghurye G S 1932 *Caste and Race in India*. Kegan Paul, London

Khare R S 1984 *The Untouchable as Himself: Ideology, Identity and Pragmatism Among the Lucknow Chamars*. Cambridge University Press, Cambridge, UK

Kothari R 1970 *Caste in Indian Politics*. Orient Longmans, Delhi, India

Marriott M (ed.) 1990 *India Through Hindu Categories*. Sage, Delhi, India

Murugkar L 1991 *Dalit Panther Movement in Maharashtra: A Sociological Appraisal*. Popular Prakashan, Bombay, India

Omvedt G 1995 *Dalit Visions: The Anti-Caste Movement and the Construction of an Indian Identity*. Orient Longman, New Delhi, India

Risley H R [1908] 1969 *The People of India*. Oriental Books, New Delhi, India

Smith R S 1986 Rule-by-records and rule-by-reports: Complementary aspects of the British imperial rule of law. *Contributions to Indian Sociology* **19**(1): 153–76

Srinivas M N 1971 *Social Change in Modern India*. University of California Press, Berkeley, CA

V. Das

Categorization and Similarity Models

Imagine meeting a friend's pet, and having to decide whether it is best classified as a cat or a dog. According to prototype models of categorization, the pet is compared with mental prototypes for cat and dog, and if it is more similar to the cat prototype than to the dog prototype, then it is classified as a cat. According to exemplar models of categorization, the pet is compared with all remembered instances of cats and dogs, and if it is more similar to known cats than to known dogs, then it is classified as a cat. Notice that both types of models rely on the notion of similarity. For these models to be usefully tested and compared with each other, or with other models of categorization, similarity must be formally defined and empirically assayed, along with formal and empirical treatments of categorization. This article briefly considers empirical results and formal models of similarity and categorization, concluding that the complex but systematic behavior should eventually yield to accurate models.

1. Similarity

1.1 Empirical Assays of Similarity

The similarity of items can be assessed in many different ways. The most direct method is simply to show a person two items and ask him or her to rate their similarity or dissimilarity. Alternatively, three or more items can be presented, and the viewer is asked to put them into subgroups of apparently similar items. A different method is for the viewer to see two items and then judge whether they are the same or different. A fourth method shows the viewer a single item that must be identified; presumably the degree of confusion between items, as measured either by response time or by error rate, reflects the similarity of the items.

Although in many cases these different assays of similarity are concordant, there are a variety of situations in which they do not agree (Medin et al. 1990). For example, consider the two pairs of countries, West Germany–East Germany and Ceylon–-Nepal. When asked to select the pair of countries that are most *similar* to each other, people tended to choose West Germany–East Germany. When asked to select the pair of countries that are most dissimilar from each

other, people again chose West Germany–East Germany (Tversky 1977). It has also been established that even for a single method of measurement, the relative similarities of items can change depending on context. For example, in the context of hair, gray is more similar to white than it is to black, but in the context of clouds, gray is more similar to black than it is to white (Medin and Shoben 1988). Nevertheless, similarity is not an utterly unconstrained, useless theoretical construct. Rather, there are strong regularities in how similarity is affected by context and by measurement method (Goldstone 1994b; Medin et al. 1993).

1.2 Models of Similarity

Among the many models of similarity two are most prominent. Multidimensional scaling (MDS) represents items by values in a multidimensional psychological space (Shepard 1987). For example, on the dimensions of size and ferocity, the item *lion* would have values *large* and *high*. The similarity of two items, A and B, is inversely related to the distance between the items in the psychological space, denoted 'distance(A,B).' The similarity of items A and B is formally specified as

$$S(A, B) = \exp(-\text{distance}(A, B)) = \exp\left(-(\sum_r \alpha_r |A_r - B_r|^p)^{1/p} \right) \quad (1)$$

where the sum is taken over all the relevant dimensions of the space, α_r is the attention allocated to dimensions r, A_r is the value of item A on dimension r, and p is a power typically set at a value of 1 or 2 depending on whether the dimensions can be selectively attended to or not. For a review of MDS models, extensions, and some of their applications, see Ashby (1992) and Nosofsky (1992b).

A different formalization was proposed by Tversky (1977) in his 'contrast model.' In this model, items are assumed to be represented by sets of present or absent features. If A denotes the set of all relevant features in item A, then the similarity of items A and B is specified by a linear combination (i.e., a contrast) of their shared features (denoted $\mathbf{A} \cap \mathbf{B}$), the features in A that are not in B(denoted $\mathbf{A}-\mathbf{B}$), and the features in B that are not in A(denoted $\mathbf{B}-\mathbf{A}$),

$$S(A, B) = \theta f(\mathbf{A} \cap \mathbf{B}) - \alpha f(\mathbf{A}-\mathbf{B}) - \beta f(\mathbf{B}-\mathbf{A}) \quad (2)$$

where f is a function that specifies the saliences of the various features (and sets of features), and where θ, α and β are weighting factors for the influence of the shared and distinctive features.

These models have been quite successful in addressing a wide variety of phenomena in similarity data. However, all these models take for granted that there is a prespecified range of dimensions or features that considered. For example, in judging the similarity of a lion and a dog, there are an infinite number of shared features such as 'smaller than a battleship,' but these features are presumably not included in the computation of similarity. In other words, the specification of what features or dimensions are relevant to the similarity computation is critical for theories involving similarity (Goodman 1972, Murphy and Medin 1985). These formalisms can accommodate differences in relevances of the features or dimensions (by way of the salience function, f, in the contrast model, and by way of the attentional factors, α_r in the MDS model), but the processes that determine these relevances are not specified in these models.

2. Categorization

Categorization can be measured in a variety of ways. For a given item, a person can be asked to make a discrete choice among several candidate categories. Alternatively, the person can give a rating of how typical the item is of a certain category. Response times can also be measured.

In many situations, similarity is a good predictor of categorization. For example, people (in the USA) rate a robin as being highly similar to many other birds, but people do not rate a penguin as being very similar to many other birds. This similarity difference predicts that people should categorize a robin more efficiently than a penguin, which in fact occurs; people are faster to verify that 'a robin is a bird' than 'a penguin is a bird.' Similarity can also explain why people are slow to falsify the statement, 'a bat is a bird,' because a bat is similar to a bird, yet is not a bird (Smith et al. 1974).

There are also situations in which similarity apparently does not predict categorization. Consider, for example, the category of 'things to remove from a burning house.' Children and heirloom jewelry are prominent members of this category, yet they are not very similar (Barsalou 1983). As another example, consider a hypothetical three-inch diameter disk-shaped object. People will rate it as more similar to a quarter (i.e., a 25 cent coin) than to a pizza, but they categorize it as more likely to be a pizza than a quarter (Rips 1989). Some of these apparent dissociations between similarity and categorization can be reconciled if selective emphasis on particular features is taken into account. For example, children and heirloom jewelry are very similar on the dimensions of irreplaceability and portability, which are critical dimensions emphasized by a burning house. When a hypothetical three-inch diameter object also includes other features characteristic of quarters, so that there is less selective emphasis on size, then category judgments cohere with similarity judgments (Smith and Sloman 1994).

A prominent similarity-based model of categorization is the 'generalized context model' (GCM, Nosofsky 1986). In the GCM, an item to be categorized is compared with all exemplars in memory, and each exemplar 'votes' for its category with a strength proportional to the similarity of the exemplar to the item. An important quality of the GCM is that the attention paid to each stimulus dimension can be adjusted to suit the categorization at hand. An extension of the GCM includes a mechanism by which the dimensional attention strengths are learned (Kruschke 1992). Whether or not empirical evidence yet demands inclusion of prototypes, in addition to exemplars, is a matter of current debate; Nosofsky (1992a) provides a summary of tests of prototype models versus exemplar models.

Perhaps more problematic for similarity-based theories are cases in which categorization appears to be based, at least in part, on 'rules' which specify strict, necessary, and sufficient conditions. One characteristic of 'rules' is that they can be extrapolated far beyond the training exemplars. Erickson and Kruschke (1998) trained people to categorize simple geometric forms, most of which could be classified by a simple rule regarding height, and a few of which were exceptions to the rule. When tested with novel forms beyond the range of the training instances, people responded according to the rule, despite the fact that the most similar training exemplar violated the rule. Another clear case in which an abstracted rule violates exemplar similarity is presented by Shanks and Darby (1998). However, even for rules, similarity to exemplars can have an influence. Erickson and Kruschke (1998) found that rule-obeying exemplars were classified more accurately when they were similar to a high-frequency training instance than when they were similar to a lower-frequency training instance. Several other demonstrations of exemplar influence on rule use have been reported (e.g., Allen and Brooks 1991, Palmeri and Nosofsky 1995). Categorization models that combine rules with exemplars, or graded similarity, or other modifications have recently been developed (Ashby et al. 1998, Erickson and Kruschke 1998, Vandierendonck 1995).

3. Future Progress

Despite the apparent complexity of phenomena in similarity and categorization, there are systematicities that should yield to a characterization of categorization in terms of similarity. Future models of similarity and categorization will have to include (a) mechanisms than learn feature relevance that is context-sensitive; (b) mechanisms that abstract new features; and (c) mechanisms that compare representations that are more complex than feature sets or points in psychological space. An excellent discussion of these issues is provided by Goldstone (1994b). Recent progress in these directions includes models of category learning in which the relevance of dimensions or features is learned (Kruschke 1999, Kruschke and Johansen 1999), numerous connectionist or neural-network models that create novel internal representations (for an overview see Ellis and Humphreys 1999), and models of similarity that apply to structured representations (Goldstone 1994a).

See also: Categorization and Similarity Models: Neuroscience Applications; Concept Learning and Representation: Models; Knowledge Spaces

Bibliography

Allen S W, Brooks L R 1991 Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General* **120**: 3–19

Ashby F G (ed.) 1992 *Multidimensional Models of Perception and Cognition*. Erlbaum, Hillsdale, NJ

Ashby F G, Allfonso-Reese L A, Turken A U, Waldron E M 1998 A neuropsychological theory of multiple systems in category learning. *Psychological Review* **105**: 442–81

Barsalou L W 1983 Ad hoc categories. *Memory & Cognition* **11**: 211–27

Ellis R, Humphreys G W 1999 *Connectionist Psychology*. Psychology Press, Hove, East Sussex, UK

Erickson M A, Kruschke J K 1998 Rules and exemplars in category learning. *Journal of Experimental Psychology: General* **127**: 107–40

Goldstone R L 1994a Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory and Cognition* **20**: 3–28

Goldstone R L 1994b The role of similarity in categorization: providing a groundwork. *Cognition* **52**: 125–157

Goodman N 1972 Seven strictures on similarity. In: Goodman N (ed.) *Problems and Projects*. Bobbs-Merrill, Indianapolis, IN, pp. 437–47

Kruschke J K 1992 ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* **99**: 22–44

Kruschke J K 1999 Toward a unified model of attention in associative learning. (Revision to appear in *The Journal of Mathematical Psychology*. Available online http://www.indiana.edu/~kruschke/tumaal.html)

Kruschke J K, Johansen M K 1999 A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*. **25**: 1083–119

Medin D L, Goldstone R L, Gentner D 1990 Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science* **1**: 64–9

Medin D L, Goldstone R L, Gentner D 1993 Respects for similarity. *Psychological Review* **100**: 254–78

Medin D L, Shoben E J 1988 Context and structure in conceptual combination. *Cognitive Psychology* **20**: 158–90

Murphy G L, Medin D L 1985 The role of theories in conceptual coherence. *Psychological Review* **92**: 289–316

Nosofsky R M 1986 Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General* **115**: 39–57

Nosofsky R M 1992a Exemplars, prototypes, and similarity rules. In: Healy A F, Kosslyn S M, Shiffrin R M (eds.) *Essays in Honor of William K. Estes. Vol. 2: From Learning Processes to Cognitive Processes*. Erlbaum, Hillsdale, NJ, pp. 149–67

Nosofsky R M 1992b Similarity scaling and cognitive process models. *Annual Review of Psychology* **43**: 25–53

Palmeri T J, Nosofsky R M 1995 Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition* **21**: 548–68
Rips L J 1989 Similarity, typicality, and categorization. In: Vosniadou S, Ortony A (eds.) *Similarity and Analogical Reasoning*. Cambridge University Press, Cambridge, UK, pp. 21–59
Shanks D R, Darby R J 1998 Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes B* **24**: 405–15
Shepard R N 1987 Toward a universal law of generalization for psychological science. *Science* **237**: 1317–23
Smith E E, Shoben E J, Rips L J 1974 Structure and process in semantic memory: A featural model for semantic decision. *Psychological Review* **81**: 214–41
Smith E E, Sloman S A 1994 Similarity- versus rule-based categorization. *Memory and Cognition* **22**: 377–86
Tversky A 1977 Features of similarity. *Psychological Review* **84**: 327–52
Vandierendonck A 1995 A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin and Review* **2**: 442–59

J. K. Kruschke

Categorization and Similarity Models: Neuroscience Applications

Categorization is the act of assigning objects or events to classes (i.e., categories). It is performed countless times every day, and is among the most important and basic of all decisions. Many different categorization models have been proposed. In several cases, models that make very different assumptions about how people learn new categories have been equally successful at accounting for a given set of categorization data. Recent advances in neuroscience have provided a new means with which to test among such competing models. The correct model of categorization will be consistent with the known neuroscience, as well as with observable categorization behavior. In addition, a knowledge of the underlying neuroscience makes it possible to develop categorization models for many groups of people that most current theories ignore (e.g., children, the elderly, people with various neurological disorders). This article reviews what is known about the neuroscience of categorization, and considers the implications of this knowledge for categorization models.

1. Structures

Categorization is not a single mental ability, but instead depends on several different abilities that use different brain structures and processes. As evidence of this, a wide variety of normal and pathological conditions have been shown to interfere with normal category learning. These include normal aging, major depression, Parkinson's disease, Alzheimer's disease, Huntington's disease, strokes, and schizophrenia (for a review, see, e.g., Ashby et al. 1998). This and other recent evidence sheds light on which neural structures do or do not participate in category learning.

1.1 Sensory Cortex

Sensory cortex refers to all cortical areas associated with sensory function. In the case of vision, this includes virtually all of the occipital cortex and much of the temporal and parietal cortex. An object must be perceived before it can be categorized, so an intact sensory cortex is necessary for normal categorization. It is not so clear, however, whether the characteristics of a category and the rules for distinguishing it from similar but different categories are learned and stored in the sensory cortex. Interest in this hypothesis has been sparked in recent years by reports of a variety of category-specific agnosias that result from damage to certain high-level visual cortical areas. Category-specific agnosia is defined as the ability to perceive or categorize most visual stimuli relatively normally, but a reduced ability to recognize exemplars from some specific category, such as inanimate objects (e.g., tools), or fruits. The most widely known of such deficits occur with human faces (i.e., prosopagnosia). Although such category-specific agnosias are consistent with the hypothesis that category structure is represented in visual cortex, they are also generally consistent with the hypothesis that visually similar objects are represented in nearby areas of the visual cortex. For example, it is well known that neighboring cells in the visual cortex tend to fire to similar stimuli. Thus, damage to some contiguous region of the visual cortex is likely to lead to perception deficits within a class of similar stimuli. In fact, specific tests have failed to rule out this similarity hypothesis.

Evidence against the hypothesis that category learning occurs in the visual cortex has been obtained in single cell recording experiments with monkeys. Several studies have found that the firing properties of cells in high-level visual areas (e.g., inferotemporal cortex) do not change when there is a switch in the category assignment of the visual stimulus to which the cell is most responsive (e.g., Rolls et al. 1977). If the categories were represented in the visual cortex, then the firing properties of visual cortical cells should change when the category memberships are switched. For example, similar studies have found changes in the responses of cells in other brain areas (e.g., amygdala) in such experiments.

1.2 Frontal Cortex

The sensory cortex projects directly to the frontal cortex, which consists roughly of the forward-most

third of the cerebral cortex. With respect to category learning, two especially important structures in this region are the prefrontal cortex and the anterior cingulate, which are thought to be critical for working (i.e., short-term) memory and executive (i.e., volitional) attention—operations that are important in some types of category learning. So, it is not surprising that there is abundant evidence that these structures are critically important for learning at least some types of category structures. The most well known such evidence comes from applications of the Wisconsin Card Sorting Test (WCST). This is a widely used neuropsychological test of frontal dysfunction in which participants learn a series of categories each of which differ on only a single critical attribute (i.e., symbol color, shape, or number). Patients with frontal cortical lesions are well known to have deficits on the WCST. Activation in frontal areas, including the prefrontal cortex and the anterior cingulate, has also been found in the few extant neuroimaging studies of category learning.

The frontal cortex is the locus of human reasoning. As such, the evidence that the frontal cortex participates in category learning supports classical models of category learning (which date back to Aristotle) that assume people learn categories by reasoning about them (e.g., through an explicit process of generating and testing hypotheses about category structure).

1.3 Basal Ganglia

All of the cortical structures that have so far been discussed (sensory cortex, prefrontal cortex, anterior cingulate) project directly to the basal ganglia—a collection of subcortical structures that includes the caudate nucleus, the putamen, and the globus pallidus (other structures are also included). Patients with diseases of the basal ganglia (e.g., Parkinson's disease, Huntington's disease) are impaired in category learning (Knowlton et al. 1996a, 1996b), which suggests that these subcortical structures may contribute to the learning of new categories. Perhaps the best evidence for a basal ganglia contribution to category learning, however, comes from a long series of lesion studies in rats and monkeys. In primates, virtually all of the visual cortex projects directly to the tail of the caudate nucleus, and the cells in this area then project, via the globus pallidus and thalamus, to the prefrontal cortex and more posterior motor areas (i.e., the premotor cortex). These projections place the caudate in an ideal position to link percepts and actions, and many researchers have hypothesized that this is its primary role. Many studies have shown that lesions of the tail of the caudate nucleus impair the ability of animals to learn visual discriminations that require one response to one stimulus and a different response to some other stimulus (e.g., Packard and McGaugh 1992). Because the thalamus and visual cortex are intact in these animals, it is unlikely that their difficulty is in perceiving the stimuli. Rather, it appears that their difficulty is in learning to associate an appropriate response with each stimulus alternative.

The basal ganglia are also thought to be critical structures for procedural learning and memory, which is a phylogentically ancient system in which simple associations between stimuli and responses are learned. For this reason, the evidence that the basal ganglia are important in category learning supports models that assume category learning depends on procedural learning and memory.

1.4 Medial Temporal Lobe

Learning about a new category requires the use of some form of memory. Oftentimes, when a person thinks of memory, he or she means the memory of facts and events (called semantic and episodic memory, respectively). It is now known that medial temporal lobe structures, including the hippocampus, the parahippocampal regions, and the entorhinal and pararhinal cortices, are critical for the consolidation of such memories. It might be expected, therefore, that such memory systems contribute significantly to category learning, especially because a number of current theories assume that people categorize a new stimulus by comparing it to memory representations of previously seen exemplars. For this reason, a number of studies have examined how amnesiacs with medial temporal lobe lesions perform in a variety of category-learning tasks. With only a few exceptions, these studies have found normal category learning, even in patients with dense amnesia that resulted from extensive lesions (e.g., Squire and Knowlton 1995). When presented with a stimulus, amnesiacs with medial temporal lobe damage are significantly impaired in judging whether they have seen that stimulus before, but their ability to assign it to the correct category is essentially normal.

These results support the hypothesis that medial temporal lobe structures are not critical for most forms of category learning, and although the issue is not yet resolved, they present a challenge to models that assume people access instance-based memories (i.e., detailed representations of previously seen category exemplars) during category learning.

2. Multiple Category Learning Systems

Another lesson learned from studying the neuroscience of categorization is that human category learning may be mediated by several different systems. This is suggested by the widely distributed neural structures that participate in category learning and by the overwhelming evidence that there are multiple memory systems. Because any category-learning system requires memory, the existence of multiple memory systems raises the possibility that each of several

memory systems might have their own dedicated category-learning module. The notion that there may be multiple category-learning systems has a long history, but recently has been the subject of intense scrutiny. Among those researchers postulating multiple systems, the consensus is that one system learns explicitly and at least one learns implicitly. The explicit system is accessible to consciousness and engages in an explicit reasoning process that may involve hypothesis testing or theory construction and testing. This system is almost certainly mediated by frontal cortical structures. The implicit system is not accessible to conscious awareness. Currently, there is debate as to whether this system uses a procedural- or instance-based memory system. Much evidence supports this multiple systems hypothesis. For example, a number of studies have found qualitative differences in the way people learn categories that are best separated by an explicit rule as opposed to tasks in which no salient explicit rule will succeed.

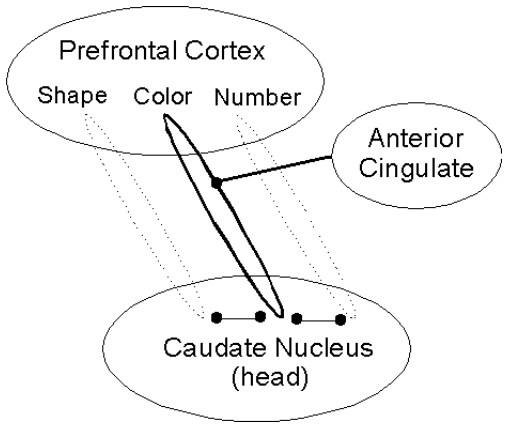

Figure 2
A schematic illustrating how the explicit system in the COVIS model of category learning operates during the Wisconsin Card Sorting Test

3. *COVIS (COmpetition between Verbal and Implicit Systems)*

Currently, there are only a few neuropsychological theories of category learning (Ashby et al. 1998, Gluck et al. 1996, Pickering 1997). A schematic illustrating the most important neural structures and pathways of one of these is shown in Fig. 1 (Ashby et al. 1998). This model, called COVIS, postulates separate, competing explicit and implicit category-learning systems that are simultaneously active at all times. Depending on the relationship between the categories to be learned, however, one system may dominate the other. There are three hierarchical levels—cortex, thalamus, and the basal ganglia. The two systems are mediated by parallel loops following the path: cortex–caudate–globus pallidus–thalamus. The posterior loop, from visual cortical areas to the tail of the caudate nucleus, mediates the implicit system. In humans, all visual areas project directly to the tail of the caudate (except Area V1), with about 10,000 visual cortical cells converging on each caudate cell. Cells in the tail of the caudate then project to the prefrontal or premotor cortex (via the globus pallidus and the thalamus). COVIS assumes that, through a procedural learning process, each caudate unit learns to associate a category label, or perhaps an abstract motor response, with a large group of visual cortical cells (Ashby and Waldron 1999).

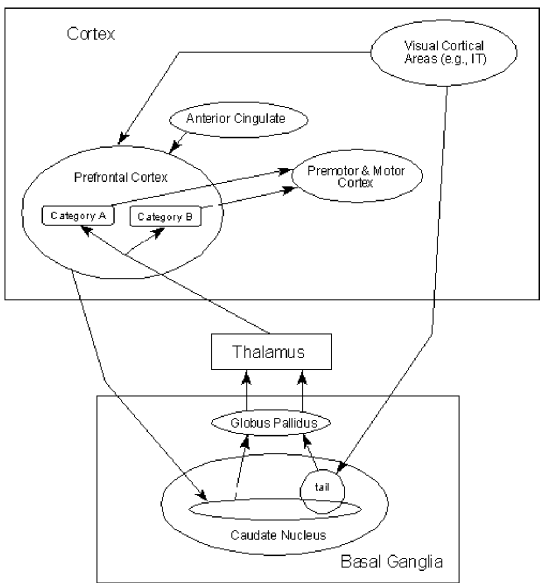

Figure 1
A schematic illustrating the COVIS model of category learning. See text for more details (IT = inferotemporal cortex)

The explicit system is mediated by an anterior loop, from the anterior cingulate and prefrontal cortex to the head of the caudate nucleus, and then back to the prefrontal cortex. Figure 2 illustrates how this system might operate for a task like the WCST. In Fig. 2, the active rule is to sort the cards by the color of the symbols pictured. This rule is maintained in working memory via the bold reverberating loop shown in the figure. If feedback indicates that this rule is incorrect, then the system must implement a different rule, perhaps one that says to sort the cards by the shape of the symbols pictured. There is evidence that implementing a new rule requires two separate operations (e.g., Owens et al. 1993). First, a new rule (e.g., sort by shape) must be selected from among the alternative

salient explicit rules, and second, attention must be switched from the previously active rule to the new rule (e.g., from a rule to sort by color to a rule to sort by shape). COVIS assumes that the selection process is mediated by the anterior cingluate and prefrontal cortex, and that the switching process is mediated within the basal ganglia (Ashby et al. 1999).

See also: Categorization and Similarity Models; Concept Learning and Representation: Models; Connectionist Models of Concept Learning; Memory Models: Quantitative; Neural Networks: Biological Models and Applications; Neural Representations of Objects

Bibliography

Ashby F G, Alfonso-Reese L A, Turken A U, Waldron E M 1998 A neuropsychological theory of multiple systems in category learning. *Psychological Review* **105**: 442–81

Ashby F G, Isen A M, Turken A U 1999 A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review* **106**: 529–50

Ashby F G, Waldron E M 1999 On the nature of implicit categorization. *Psychonomic Bulletin & Review* **6**: 363–78

Gluck M A, Oliver L M, Myers C E 1996 Late-training amnesic deficits in probabilistic category learning: a neurocomputational analysis. *Learning Memory* **3**: 326–40

Knowlton B J, Mangels J A, Squire L R 1996a A neostriatal habit learning system in humans. *Science* **273**: 1399–402

Knowlton B J, Squire L R, Paulsen J S, Swerdlow N R, Swenson M, Butters N 1996b Dissociations within nondeclarative memory in Huntington's disease. *Neuropsychology* **10**: 538–48

Owens A M, Roberts A C, Hodges J R, Summers B A, Polkey C E, Robbins T W 1993 Contrasting mechanisms of impaired attentional set–shifting in patients with frontal lobe damage or Parkinson's disease. *Brain* **116**: 1159–75

Packard M G, McGaugh J L 1992 Double dissociation of fornix and caudate–nucleus lesions on acquisition of two water maze tasks: further evidence for multiple memory systems. *Behavioral Neuroscience* **106**: 439–46

Pickering A D 1997 New approaches to the study of amnesic patients: what can a neurofunctional philosophy and neural network methods offer? In: Mayes A R, Downes J J (eds.) *Theories of Organic Amnesia*. Psychology Press, Hove, UK, pp. 255–300

Rolls E T, Judge S J, Sanghera M K 1977 Activity of neurones in the inferotemporal cortex of the alert monkey. *Brain Research* **130**: 229–38

Squire L R, Knowlton B J 1995 Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences, USA* **92**: 12470–4

F. G. Ashby

Catholicism

Catholicism signifies the central form that Christianity has taken in history from, at latest, the second century, but it is a form that has continued steadily to evolve. In the second millennium Catholicism has come to designate most particularly the character Christianity has assumed within the Roman Communion under the authority of the papacy.

'Catholic' is not a term to be found in the New Testament but it is already being used to characterize and designate the church in its early postapostolic period. Its primary meaning is 'universal,' signifying both Christianity's claim to be a faith unbounded by language, ethnic, or class boundaries and the insistence that local church communities must be in communion with one another: the church in Corinth or Ephesus is 'Catholic' because part of an association of communities sharing the same faith and rituals present up and down the Roman Empire and, potentially, everywhere. As such the term quickly became the church's most regularly used name, incorporated into the creeds. All Christians using the creeds, whether Roman Catholic, Orthodox, or Protestant, claim to believe in, and belong to, the Catholic Church.

The character assumed by Christianity in the second century and generally termed 'Early Catholicism' is one in which each local church is ruled by a bishop, assisted by priests (presbyters) and deacons, claiming 'succession' from the apostles. Entry is via the sacrament of baptism while regular worship is centered upon the weekly celebration of the Eucharist, renewing Jesus's last supper with his disciples. The Hebrew scriptures were retained as inspired, generally in their Greek Septuagintal form, to which was added a collection of the earliest Christian writings, the New Testament, to which apostolic authority was attributed. The shape of Catholicism had, however, been established well before the full canon of the New Testament was agreed upon, certain books remaining for long in dispute. While there was at first no formal machinery for linking local churches together in the one church, local councils of bishops from neighboring towns soon began to be called, while the churches in major cities, particularly those claiming an apostolic foundation, assumed a leading position. Among these the Church of Rome, where both Peter and Paul were martyred, came to be accorded an unquestionable preeminence from an early date.

While all the books of the New Testament are centered unequivocally upon the figure of Jesus Christ, they already suggest, particularly in the account of the first Christian martyr, Stephen, but also in concentration upon certain of the apostles, the importance for early Christianity of the personal 'witness' to Christ (the strict meaning of martyr). A cult of the martyrs, focused upon their tombs, began early and developed into a wider cult of the saints that remained a permanent characteristic of Catholicism. Furthermore virginity/celibacy was stressed from early times as an alternative road to exceptional holiness, a way later institutionalized in monasticism.

This 'Early Catholicism,' constituted by hierarchical ministry, sacraments, the canon of scripture, and the

cult of the saints, was common to east and west, the Greek and Latin churches, as also to Syrian, Coptic, and Armenian churches. Despite diversities of language, liturgy, and theological tradition, Catholic Christianity was held together by the fact and theology of 'communion,' a sacramental bond of fellowship from which individuals or whole churches could be excluded on account of heretical teaching or moral deviation.

After Christianity became a privileged religion with the conversion of the emperor Constantine in the early fourth century, it grew enormously in numbers. What had hitherto been a community of a voluntary sort whose members knew that they might well have to face persecution on account of their beliefs became instead quite quickly a church of the masses, many of whose members had received very limited Christian instruction. The myths and customs of the rural population were carried on, often covered over with a thin Christian veneer. Henceforth Catholicism would be characterized by the co-existence of a clerical and monastic religious upper class with a popular religion which it would preach to generation after generation but never fully convert.

Post-Constantinian Catholicism grew too in the development of an institutional framework designed to resolve conflicts and maintain unity, although the result was often to precipitate still worse division. The most striking development of this was what came to be recognized as the 'ecumenical council' beginning with that of Nicaea (325). At the same time the authority of the major churches—Rome, Alexandria, Antioch, and then Constantinople and Jerusalem—was further accentuated. By the fifth century this had been stabilized in the theory of the 'Five Patriarchates' but, among them, Rome and Constantinople held a quite privileged position. That of Rome, with its unquestionable apostolic origins, had been there from the start, but was reinforced greatly from the late fourth century. Constantinople, the new imperial capital, claimed to be for that reason also the church's second see in dignity, though Rome denied that its own primacy was due to its civic, rather than its apostolic, status. In practice, the two co-existed for centuries, at times somewhat tensely, but also co-operatively. Rome enjoyed an increasingly effective primacy of jurisdiction throughout the Latin west but of little more than honor in the east. At the same time the advance of monasticism profoundly affected the ethos of Catholicism in both east and west.

It was only after the eleventh century that the 'Catholicism' of the western church could be distinguished clearly from the 'Orthodoxy' of the eastern. The Gregorian reform (named after Gregory VII, pope from 1073–1085) profoundly affected the western church but through the accentuation of tendencies already present. Three are crucial: the first was a large extension in the claims but, still more, the effective jurisdictional power of the papacy. Its ability to impose a defined model of ecclesiastical life was turning 'Catholicism' into 'Roman Catholicism': a hitherto largely decentralized church was being made to conform ever more closely to rules laid down in Rome, rules which included insistence on higher moral and educational standards in the clergy and greater separation between clergy and laity. The second was a vast extension of canon law and its implementation through a hierarchy of church courts with appeal always possible to the highest courts in Rome. Canon law became the effective tool for implementing the new Gregorian ideals. The third was the law of celibacy for all the clergy. Hitherto most ordinary priests were, or had been, married, just as they had been and continued to be in the east. Henceforth in the west the married priest would be an offender, though in practice he would long continue to exist. The law of celibacy was a necessary tool for ensuring the segregation of clergy from laity, making it possible for the former to be controlled effectively by higher church authority. The 'church' became in practice more and more a clerical reality in which the laity were expected ideally to remain passive and obedient. The identification of church and clergy was accentuated further by language. The early church had allowed great linguistic flexibility but in the medieval west Latin came to have a wholly privileged position as the language of the church and its public services as well as of theology. As a range of vernaculars developed vigorously their own literatures, the ecclesiastical use of Latin became a further characteristic of Catholicism. It expressed a certain clerical universalism at odds with the national vernaculars used by the laity.

Two further developments, linked particularly with the thirteenth century, were also lastingly influential for the shaping of Catholicism and its differentiation from eastern Orthodoxy. One was the founding, beginning with the Franciscan and Dominican friars, of religious orders which were not purely monastic but, instead, active, mobile, and involved in the world around them, especially the urban world. Further societies of this sort, most notably the Jesuits in the sixteenth century, would multiply, providing a powerful, well-organized, and educated religious army at the service of the church, available not only internally but also for missionary service abroad, as the Franciscan mission to China in the fourteenth century already demonstrated. The second was scholasticism, formal systems of theology, heavily influenced by the rediscovery of Aristotle's works, which grew out of the teaching system of the new universities. Here formal logic was learned and applied in a systematization of theology very different from that of the patristic period or the continued traditions of the Greek east.

Thus, by the later Middle Ages the type of Christianity which we have come to call Catholicism was almost fully evolved—a mix of subapostolic, post-Constantinian, and medieval developments. The Reformation brought an attack on many of these developments, almost anything in fact which could

not be defended with firm scriptural authority. Against this assault the papacy yielded almost nothing. It admitted moral and educational, but not doctrinal or institutional, failings. The Council of Trent (1545–63) formulated the lines of the Counter-Reformation position, emphatically reaffirming almost everything that Luther, Zwingli, and Calvin had impugned. The one major new element that it added to the character of Catholicism as a social reality was the seminary. Hitherto there had been no institutions formally established for the training of diocesan priests, who had learned what they needed to know either through the general education of a university or through a measure of apprenticeship, a son often enough learning the job from his father. The Tridentine seminary would improve greatly the effectiveness and education of the clergy but also further distance them socially from the laity, particularly on account of the 'minor seminary,' effectively a secondary school in which candidates for the priesthood would be segregated from their early teens. This could lead easily on to a further gap between the clergy and the university world, and therefore also between the clergy and modern ways of thought. While with time medieval scholasticism was ousted from the university, it would in a much degraded form continue to dominate the characteristic mindset of the clergy until the twentieth century.

Tridentine Catholicism was, nevertheless, far more creative than the term 'Counter-Reformation' might suggest. While classical Protestantism with its commitment to '*sola scriptura*' was averse to religious involvement in secular culture, the Catholicism of the sixteenth and seventeenth centuries produced its last great cultural era, the Baroque, as rich artistically as the earlier Catholic art cultures, the Romanesque and the Gothic. In art and music, as in theology and spirituality, Catholicism is by nature open to the secular and developmental, even luxuriantly. This was always true, though only in the nineteenth century did John Henry Newman justify the development of doctrine as a human and Catholic necessity, to be contrasted with Protestant insistence upon not going beyond the explicit content of scripture.

Sixteenth- and seventeenth-century Catholicism was also particularly fruitful in the literature of prayer, from the mystical writings of the Spanish Carmelites, Teresa of Avila and John of the Cross, to the Jesuit *Spiritual Exercises*, Francis de Sales' *Introduction to the Devout Life*, and countless other minor classics, many of them popular also among Protestants. Catholicism then has traditionally combined institutional rigidity with spiritual and artistic creativity. In the nineteenth century with the triumph of Ultramontanism (a name used to describe a form of Catholicism in Europe north of the Alps anxious to enhance papal authority and uniformity with Rome in contrast with the more old-fashioned Gallicanism which accepted papal supremacy but limited to a minimum papal intervention in non-Italian ecclesiastical affairs) institutional rigidity grew while creativity declined. Faced with the often anticlerical, and even antireligious, ethos of the French Revolution and later liberal movements, the Catholic Church drew in on itself as a fortress against the modern world, defining the authority of the pope at the first Vatican Council of 1870 under Pius IX (1846–78) while a little later Pius X (1903–14) condemned every form of 'modernism.' In the first decades of the twentieth century Catholicism could then appear as a consistently reactionary force, intellectually, culturally and socially. Yet it was far from being a community in decline. Modern technology made Roman centralization work in a way never previously possible. The railway brought countless pilgrims to Rome and a cult of the reigning pope among ordinary Catholics quickly developed. Pius IX, antimodern as he was, became in consequence the first 'modern' pope, almost idolized by millions. At the same time a series of apparitions of the Blessed Virgin Mary mostly to children, of which the most famous were at Lourdes in 1858 and at Fatima in 1917, much reinforced a principal focus for Catholic popular devotion, both backed by and supporting papal definitions of Mary's Immaculate Conception (1854) and Assumption into Heaven (1950). The missionary orders, mostly nineteenth-century foundations, were extending Catholicism throughout the southern hemisphere; there was a move to encourage forms of lay apostolate across Europe and North America, while for the first time there had developed a vast army of nuns at work almost everywhere. This last may have been the most significant shift. In the Middle Ages nuns were few in comparison with monks and included no active orders. Only in the seventeenth century, led by heroic women like Mary Ward and Jane Frances de Chantal, did unenclosed orders of women, committed to apostolic work as well as prayer, begin to exist despite huge clerical opposition. By the close of the nineteenth century there were more nuns in the church than priests. Silent in many ways they were forced to remain, Roman Catholicism being traditionally an intensely male-dominated and clerical form of Christianity, but a changing balance between the sexes was evident even here.

Catholicism continued to present publicly an almost unchanged face until the end of Pius XII's pontificate in 1958; highly clerical, intellectually repressed, well-organized, expansive, more tightly controlled from Rome than ever, insistent upon the use of Latin, retaining apparently untouched its medieval legacy, institutional, and liturgical. Yet processes of modernization had already gone far, intellectually especially in western Europe, institutionally elsewhere with the appointment of Asian and African bishops, and the creation of numerous cardinals to represent every continent. For the first time since the fourteenth century, the majority of cardinals were not Italian. It

needed only the gentle touch of John XXIII and the calling of the Second Vatican Council (1962–5) for a truly revolutionary process to set in, only part of which was formally approved by the Council. The intellectual debate unleashed by 4 years of conciliar discussions set a movement in train which went well beyond the letter, though often not far beyond the spirit, of the Council's decrees. Within a few years Latin was replaced by the local vernacular in the liturgy and theological education; scholastic theology almost disappeared from the seminaries; hitherto frigid relationships with the Orthodox and Protestant churches gave way to a near-frenzied spate of official ecumenism.

In strictly theological terms the changes of Vatican II may seem less than revolutionary, but in the wider terms of the sociology of Catholicism, its sense of identity and shared ethos, they undoubtedly were. Symbolic ramparts, such as the denial of the cup in communion to the laity, which had been fiercely maintained against Protestantism for centuries, were dismantled quickly. It was not only that the barriers which made Catholics feel they were different—from Latin to Friday abstinence from meat—had almost disappeared, and with them the shared certainty that the pope was always right. It was also that Catholics now found themselves divided in a way that had never previously been the case in modern times. If, for Pope Paul VI (1963–1978) and the official church, Vatican II had gone just far enough this was challenged from both right and left. For the former it had gone much too far, betraying authentic Catholic values in the pursuit of ecumenism, while for the latter it had failed through evading the most awkward issues, notably that of Roman centralization. The conflict hardened particularly over two matters that refused to go away: contraception and the law of celibacy for priests. Pope Paul had prevented the Council from discussing either. Contraception was passed to a papal commission that actually recommended a change in teaching. Paul rejected the recommendation and in 1968, in the encyclical *Humanae Vitae*, reaffirmed the use of contraception to be always wrong. He was quite unprepared for the degree of open opposition this aroused and for the consequent decline in confidence in papal teaching, particularly among the more thoughtful laity, who disregarded widely the papal ban. Again, despite numerous appeals, at least to allow married men to be ordained, both he and John Paul II (1978–) repeatedly renewed insistence on the law of celibacy in face of the near-catastrophic decline in the number of priests in the western world in the 30 years following the council, due both to the withdrawal of priests from the ministry, usually to marry, and a linked fall in recruitment. A further factor was that minor seminaries were now seen as obsolete and closed almost everywhere in Europe and North America in the early postconciliar years. The credibility of the law had been internally undermined by several things: Vatican II's praise for the married clergy of Catholic Uniate churches in the east, its permission for married men to be ordained to the diaconate, and, finally, the acceptance of married priests, especially in Britain, if they were converts who had been ministers in another church previously.

Opposition to contraception set Vatican policy against almost all international attempts to curb the rapid rise in global population. This, together with papal rejection of the ordination of women and refusal to open the door to a married priesthood, did much to restore the nineteenth-century image of official Catholicism as incurably reactionary and antimodern. At no time since the Middle Ages has the papacy appeared to matter so much within public policy as in the age of John Paul II but its image has reverted to that of Pius XII. In papal history John XXIII could seem an aberration.

Shortage of priests resulted in the closing or merging into one of many parishes, especially in France, and the transfer of duties from priests to deacons or to the laity in general. Women, who in the past had never been allowed even to serve at mass, were now enrolled in thousands as 'eucharistic ministers.' Moreover, the religious orders, both male and female, whose numbers had grown enormously between 1850 and 1960, went into rapid decline in Europe and North America. In practice a rapid declericalization of the Church's ministry was taking place. While this could be seen as advantageous, its value was partially undermined by the haphazard way in which it happened, lack of a positive pastoral strategy and continued insistence by Rome that the traditional clerical pattern remained normative.

Behind the conflicts over practice and the loss of confidence in authority there were now huge differences in theological perception. If Vatican II seemed almost unbelievably radical to Catholics content with the church of Pius XII, the policies of John Paul II seemed almost unbearably reactionary to Catholics who had found the Council a harbinger of religious liberation. Many 'progressive' Catholics, while deeply loyal to the unity of a worldwide communion and to the idea of a central 'see of unity,' not only rejected much current papal teaching, particularly that of John Paul, whose personal theology often appeared fundamentally preconciliar, but had come to share so full a sense of fellowship with Christians of other churches that intercommunion had come to be seen as obviously right and was practiced frequently with the encouragement of many priests, despite condemnation by authority. 'Conservatives' accused 'progressives' of practicing an '*à la carte*' Catholicism. 'Progressives' charged 'Conservatives' with 'integralism,' confusing the substance of Catholicism with clericalism and Ultramontanism. While 'Progressives' tended to define Catholicism in terms inclusive of the principal insights of the Reformation as constituting a necessary critique of the one-sidedness of the late medieval church,

'Conservatives' continued to define it as inherently contradictory to Protestantism. Vatican II had, in fact, led to a profound questioning of the model of Catholicism as it developed across a 1000 years from the Gregorian reform through the Counter-Reformation to the decrees of Vatican I and Ultramontanism. How far that model will in consequence alter in the long run remains unclear.

It may well be claimed that the areas where this contestation is most in evidence are also areas of church decline. As against that decline, most obvious in western Europe, is to be set a huge growth of Catholicism in Africa, Latin America, and several parts of Asia. Brazil is now the country with by far the largest number of Catholics. Here too, however, there are great differences. Latin America has suffered an acute shortage of clergy for far longer than Europe and most of the issues dividing the church in Europe are evident here too, as also in the Philippines. The advance of Pentecostal Protestantism is also altering the hitherto almost exclusively Catholic character of Latin American Christianity. Elsewhere in Asia and in most of Africa the number of priests and religious has escalated. Religious orders whose recruitment plummeted after Vatican II in their traditional areas have grown remarkably in Africa, India, and Indonesia. For the Jesuits in particular India became the most flourishing of provinces, but the growth of houses of contemplative nuns in Latin America also is noteworthy. Nevertheless throughout the southern hemisphere there are profound tensions between Roman control and local processes of inculturation and political orientation. The future of Catholicism as the principal constituent within Christianity may depend upon how far a lasting *modus vivendi* is established between pope and Curia on the one hand and the young churches of the south on the other. John Paul's understandable preoccupation with Poland and eastern Europe is unlikely to be continued by his successors. Brazil, Argentina, India, the Philippines, Nigeria, and the Congo may matter more for the shaping of Catholicism in the twenty-first century than traditional pillars such as France, Spain, Ireland, and Poland, though the mediating position of the vast church of the United States may still prove to matter most of all. While the decline in church practice or missionary enthusiasm in France or Ireland is undeniable, there is as yet little evidence to suggest that world wide the strength of Catholicism is in decline and, if the reforms of Vatican II have led to a shaking of the foundations unprecedented for many centuries, they have also contributed to a massive renewal evident above all within the third world.

See also: Catholicism and Gender; Christianity Origins: Primitive and 'Western' History; Historiography and Historical Thought: Christian Tradition; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Latin American Studies: Religion; Middle Ages, The; Orthodoxy; Reformation and Confessionalization; Religion: Evolution and Development; Religion, History of; Religion, Sociology of; Renaissance; Revolutions of 1989–90 in Eastern Central Europe; Western European Studies: Religion;

Bibliography

Alberigo G (ed.) 1995 *History of Vatican II*. Orbis, Maryknoll, NY
Aubert R 1978 *The Church in a Secularised Society*. Paulist Press, New York
Cleary E L, Stewart-Gambino H (eds.) 1992 *Conflict and Competition: The Latin-American Church in a Changing Environment*. Lynne Rienner Publishers, Boulder, CO
Duffy E 1997 *Saints and Sinners: A History of the Popes*. Yale University Press, New Haven, CT
Ellis J T 1969 *American Catholicism*. University of Chicago Press, Chicago
Flannery A 1992 *Vatican Council II: The Conciliar and Post-conciliar Documents*. Dominican Publications, Dublin, Eire
Gannon T M (ed.) 1988 *World Catholicism in Transition*. Macmillan, New York
Gifford P 1998 *African Christianity: Its Public Role*. Indiana University Press, Bloomington, IN
Greeley A M 1977 *The American Catholic: A Social Portrait*. Basic Books, New York
Hastings A 1989 *African Catholicism*. SCM Press, London
Hastings A (ed.) 1991 *Modern Catholicism*. SPCK, London
Hastings A (ed.) 1999 *A World History of Christianity*. Eerdmans, Grand Rapids, MI
Hebblethwaite P 1984 *John XXIII: Pope of the Council*. Chapman, London
Hebblethwaite P 1993 *Paul VI: The First Modern Pope*. Paulist Press, New York
Hennesey J J 1981 *American Catholics: A History of the Roman Catholic Community in the United States*. Oxford University Press, New York
Hornsby-Smith M P 1987 *Roman Catholics in England: Studies in Social Structure Since the Second World War*. Cambridge University Press, Cambridge, UK
Keough D 1990 *Church and Politics in Latin America*. St Martin's Press, New York
McBrien R P 1994 *Catholicism*. Harper, San Francisco
McBrien R P (ed.) 1995 *The Harper Collins Encyclopedia of Catholicism*. HarperCollins, New York
O'Farrell P 1977 *The Catholic Church and Community in Australia: an History*. Nelson, West Melbourne, Australia

A. Hastings

Catholicism and Gender

Christianity in its beginnings appeared to offer a gender-inclusive promise of redemption. Both women

and men entered the baptismal waters to be freed from the corrupting power of the old Adam and to rise to newness of life in Christ. This vision of an inclusive community of redemption was rooted in the ministry of Jesus and the early Jesus community. Here it was suggested that women and other despised people are not just included in the existing social and cultural system, but that system itself is to be radically dismantled and turned upside down.

This vision was expressed in the baptismal formula cited in St. Paul's *Letter to the Galatians* (3:28):

> For as many of you are baptized into Christ have put on Christ. There is neither Jew nor Greek, there is neither slave nor free. There is neither male nor female, for you are all one in Christ Jesus.

Some early Christians understood this as meaning that baptism into Christ dissolved the social subordination of women to men and allowed women to preach and travel as itinerant evangelists. But this view clashed with deeply held assumptions that the female was inferior, unclean, under subjugation by nature, unable to image God.

Paul himself sought to modify this radical vision by suggesting that although this transformation of relations will happen very soon, for 'the time is later now than when we first believed,' it has not yet happened. Here and now all are to remain in their present social conditions, whether married or unmarried, slave or free.

Paul's successors sought to systematize his efforts by spiritualizing equality and splitting it from social relations, insisting that the hierarchical authority of the paterfamilias is still intact. Wives, youths, and slaves must not only continue to obey their husbands, fathers, and masters, but do so willingly.

But this conflict between egalitarian and patriarchic views of the church continued in the second century. The deutero-Pauline letter of I Timothy tries to counteract women who are claiming a public ministry by a patriarchic reading of Genesis 1–3. Women were created second and sinned first; they are under subjugation through their secondary place in God's original order of creation. Moreover, this subjugation is redoubled through their primacy in sin. They may not preach, but are to keep silence. They are saved only by accepting this subjugation in marriage, child bearing, and modest, submissive behavior.

This struggle between conflicting views of Christianity continued into the third century. By the fourth century, these more radical versions of redemptive equality had been mostly defeated. The women who flocked into celibate Christian monastic life in the fourth and fifth centuries accepted their segregation from public ministry and preaching in the church. Yet the idea continued to linger in female monasticism that the female ascetic, by virtue of renouncing marriage, was no longer under male domination.

Augustine tried to fuse these contradictions between creational patriarchy and redemptive equality. He realized that in order for woman to be redeemed in Christ she must be made in the image of God, otherwise she would have no redeemable soul. But he was also bound by Paul's dictum in I Corinthians that only the male is the image and likeness of God, and woman has only a secondary reflection of humanness under the male.

Augustine's solution was to split the image of God as a gender-neutral spiritual soul from woman's nature as female. The sex-neutral soul in woman is in the image of God and possesses a redeemable soul, but, as female, she is not in the image of God but images the body, the subjugated lower self. As female, woman was made subject to the male from the beginning, even in the original creation. This subjugation has to do with her role as wife, and sexual partner for the production of children. Furthermore, due to her primary role in leading the male into disobedience to God, woman's original subordination has now been worsened into coercive servitude. This subjugation is not changed by baptism or even by celibacy. Rather, a Christian woman should accept her subordination willingly as both her original place in creation and as punishment for her primacy in sin. This subordination will disappear only in heaven when the created order of sex, procreation, mortality, and sin is swallowed up in redemptive immortality.

This Augustinian view would be transmitted to the Middle Ages as orthodoxy to be repeated in substantially the same form by Thomas Aquinas, with the addition of the Aristotelian anthropology that defined women as biologically inferior, non-normative human beings. For Thomas Aquinas, this inferior nature of woman meant that Christ had to be a male in order to represent the fullness of human nature, only possessed by males. Only males, in turn, could represent Christ in the priesthood. Moreover, women could not exercise public authority in their own right, since they were by nature under subjugation.

Yet alternative traditions continue to lurk around the edges of the Medieval Catholic tradition, ideas of an original spiritual equality that was accessible here and now through redemptive conversion. There was also the Pentecost tradition, in which the gift of the Spirit empowers women as well as men to speak as God's prophets. In the Wisdom tradition, God coming forth to create the world is personified as female. The creating, redeeming God is imaged as both male and female.

Medieval female mystics, such as Hildegard of Bingen and Julian of Norwich, laid hold of these alternative traditions. Although accepting female subordination in marriage, they assumed that they had been freed from this subordination through celibacy and withdrawal into a female community independent of male rule, where women religiously governed their own lives. As virginal women they existed no longer as

femina, but as *homo*, as the spiritual self made in the image and likeness of God. Through the gift of the Spirit, they had been freed from silence and called to speak as God's prophetic voice to the church and society at large, calling them to account for their sins and revealing the mysteries of God's coming transformation of the world.

Hildegard of Bingen, Julian of Norwich and other female mystics also claimed the wisdom tradition to define a feminine personification of God through which God created, sustains, and reconciles creation to God-self. The wisdom tradition enabled these female mystics to overcome the assumption that God is essentially male and so woman as woman cannot image God. As divine wisdom woman images God and God is imaged as woman. This female wisdom tradition culminates in the thought of Julian of Norwich, where femaleness is brought into the Trinity itself. As Julian puts it, 'As truly as God is our Father, so truly is God our Mother.'

The Reformation era of the sixteenth to seventeenth centuries was in many ways a setback for women. Laws that allowed women membership in guilds and control of property were narrowing. The Protestant Reformation abolished celibacy while championing the patriarchal family, where women's sole vocation was that of dutiful wife and mother. Women lost many vocations that had been available through celibate female communities.

In Catholic areas, of course, monastic life for women continued, but the narrowing of women's legal and economic opportunities also happened there. The Catholic Counter-Reformation stressed the strict cloistering of women and resisted the efforts of many creative women of the period, such as Mary Ward in England, who sought to found new non-cloistered activist orders for women. Women were either to be wives obedient to their husbands or cloistered nuns obedient to the male authorities of the church.

A few Catholic humanist thinkers, such as Erasmus, championed women's expanded education. The maverick Catholic humanist Agrippa von Nettesheim even suggested that women were originally not only spiritually equal to men, but superior in their representation of the divine Wisdom of God. Male domination was recent, and manifested not God's will or female inferiority, but simply male tyranny. But this view was deeply contrary to the trends of the times and would not be heard again until expounded by late eighteenth century feminists—for example, Mary Wollstonecraft, wh ochampioned women's rights in her treatise, *The Vindication of the Rights of Woman* (1792). Such views would not appear in Catholicism until after the Second Vatican Council in the 1960s.

There was virtually no organized Catholic participation in the first wave of American feminism that ran from the first Women's Rights Convention in Seneca Falls, New York in 1848 until the winning of women's suffrage in 1920. The reasons for this lie in several factors; first, the Catholic bishops were vehemently opposed to feminism, associating it with a secular modern apostasy to Christian and 'family' values, represented by child labor laws, the Equal Rights Amendment, birth control, divorce, liberalism, and socialism. Many bishops explicitly rejected women's suffrage itself, claiming that it would take women out of their proper sphere in the family.

The view of gender that came to prevail in nineteenth and early twentieth century Catholicism was no longer one that stressed women's inferiority, but rather their complementarity to men. Women were different and even better than men, but their specific nature lay in their passive auxiliary being. Women were 'Marylike,' rather than 'Christlike.' They are to represent the church in its submission to Christ, and to masculine leadership, not taking any such leadership themselves. In these views, Catholicism echoed and put its particular stamp on the romantic views of gender complementarity common to the period.

The conflict between official Catholicism and more egalitarian views of women has moved into the Catholic community in the post-Vatican II era. Vatican II itself largely ignored women. Women, even women religious, were absent from the delegates to the Council. Protest about women's absence by Cardinal Leon-Joseph Suenens in 1962 brought a slight change. A few women religious and laywomen were allowed as 'auditors' without vote in the later sessions of the Council. Three even served on commissions that drafted the final Council statements, specifically those on 'the Church in the Modern World' and on the 'Apostolate of the Laity.'

The final documents of the Council have a few concessions to women's new role in work and legal equality in modern society. It is said that women are claiming a new equity for themselves before the law and in fact, and the participation of women in cultural life should be acknowledged and supported by the church. These statements are echoed in Pope John XXIII's encyclical *Pacem in Terris* (1963). Here it is said that 'women are becoming ever more conscious of their human dignity, they will not tolerate being used as mere material instruments, but demand rights befitting the human person in both domestic and in public life' (Sect. 41).

These concessions to women's role in public society did not include changes in the life of the church itself, in terms of women's ordination, or in teachings on sexuality and reproduction. Both of these were being challenged in the 1960s. Mainline Protestant churches in Western Europe and North America had begun to ordain women in increasing numbers in the 1960s. In 1976 the American Episcopal Church endorsed women's ordination, a change that brought women into the priesthood in a tradition more closely allied to Catholicism.

The Roman Magisterium responded with an encyclical in that same year insisting that women could

not be ordained because as women they lacked the capacity to 'image Christ.' Maleness and representation of Christ as a male were thus identified in a way that many Catholics had come to be seen as questionable. Far from ending the discussion, support groups for women's ordination among Catholics have grown since 1975 in Europe, and the United States and Canada.

Another area of contention among Catholics has been the traditional teaching that sex (allowable within marriage alone) is primarily for procreation, and cannot be separated from its procreational potential. Thus birth control (other than the rhythm method) is forbidden. As moral theologians and ordinary Catholics began increasingly to question these teachings, Pope Paul VI convened a commission to study the question in 1964. The commission met from 1964–7 and included among its members three married couples who testified to the anxiety created by lack of reliable family planning.

The commission's report voted to change the teaching to one that accepted all medically approved methods of birth control within the context of committed, child-raising families. But this official report was rejected by Pope Paul VI, who was advised by a few dissenters that such a change would threaten the church's claim to authoritative teaching. The reaffirmation of the anti-birth control teaching in the papal enclyclical *Humanae Vitae* (1968) created a storm of protest. Most lay Catholics have simply chosen to ignore the Church's teaching on this subject. Yet the Vatican, under John Paul II, has continued to insist on this teaching and even to claim that it is unchangeable, deepening the conflict between Magisterium and laity.

Recognizing the deeply conflicted state of relations of the official church to women, the American bishops voted in the early 1980s to engage in an official dialog with the US Women's Ordination Conference. At the conclusion of this dialog the bishops decided to undertake a pastoral letter on women in the church. They conducted a series of 'listening sessions' with Catholic women around the United States to hear their issues. They then began to draft an epistle that declared that men and women should be partners in home, family, and church. The epistle even condemned 'sexism' as a sin.

But the Vatican intervened in the process of drafting this letter, demanding that the condemnation of birth control and women's ordination must be strengthened, and that the model of male–female relations must be one of complementarility, not partnership. The effect of these interventions was a decision by the American bishops to table the epistle, rather than issue a document that was likely to worsen rather than improve relations with women.

This conflict between official teachings of the Vatican and a growing consciousness of their rights to equality among Catholic women (and male supporters) is by no means limited to North America. Feminist theological networks grew in Western Europe in the 1980s and 1990s, and in the Catholic church in Asia, Africa, and Latin America. With an increasing crisis in the celibate male priesthood, more and more of the ministry in Catholic parishes and chaplaincies is being done by theologically trained laywomen.

Thus the conflict between the Catholic hierarchy and women, exacerbated by the reactionary policies of the pontificate of John Paul II, showed no signs of abating at the end of the twentieth century. The issues go far beyond the inclusion of a few token women in clerical office. They point to a fundamental rethinking of ancient traditions that are rooted in the male hierarchy and a view of femaleness and sexuality as expressions of inferiority and sin to be distanced from the 'purity' of sacramental office.

A thoroughgoing rethinking of these patterns any time soon in official teaching seems unlikely. But since women who claim a more egalitarian view of themselves and their right to decide about their own bodies show no signs of departing in large numbers from the Catholic church, it is likely that this conflict will continue in the foreseeable future as a deep schism of viewpoints concerning gender identity and relations among Catholics.

See also: Buddhism and Gender; Catholicism; Family and Gender; Feminist Theology; Goddess Worship (Old and New): Cultural Concerns; Islam and Gender; Judaism and Gender; Protestantism and Gender; Religion and Gender; Religion: Family and Kinship; Religion: Mobilization and Power; Religion: Morality and Social Control; Women's Religiosity

Bibliography

Boerrsen K E 1981 *Subordination and Equivalence: The Nature and Role of Women in Augustine and Thomas Aquinas*. University Press of America, Washington, DC

Kaiser R B 1986 *The Politics of Sex and Religion*. Leaven Press, Kansas City, MO

Raming I 1976 *The Exclusion of Women from the Priesthood: Divine Law or Sex Discrimination?* Trans. N Adams, Scarecrow Press, Metuchen, NJ

Ruether R R 1995 'Catholic women'. In: *In Our Own Voices: Four centuries of American women's religious writing* 1st edn., Harper, San Francisco, pp. 17–60

Ruether R R 1998 *Women and Redemption: A Theological History*. Fortress Press, Minneapolis, MN

Swidler L, Swidler A 1977 *Women Priests: A Catholic Commentary on the Vatican Declaration*. Paulist Press, New York

Weaver M J 1985 *New Catholic Women: A Contemporary Challenge to Traditional Religious Authority*, 1st edn. Harper & Row, San Francisco

R. R. Ruether

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Cattell, Raymond Bernard (1905–98)

In describing the psychologist Raymond Bernard Cattell's comprehensive research program, Child, in a single sentence, gave an amazingly complete and accurate description: 'His major concern was to map out an integrated theory of human intellectual, temperamental and motivational characteristics within the context of hereditary and environmental influences using multivariate methods of analysis' (Child 1998, p. 356). In trying to convey the sweep of Cattell's theoretical and empirical work Goldberg (1968) referred to him as 'psychology's master strategist' (p. 617). Indeed, the sheer audacity of venturing into so many major research areas during one career is unlikely to be seen again for some time.

Raymond B. Cattell was born on March 20, 1905 in Staffordshire, England. He was raised in Devon, near the town of Torquay. Cattell began his academic career in chemistry at University College, London and received his Bachelor of Science degree with first class honors in 1924. Drawn to the study of psychology, Cattell began work on a doctorate at King's College, also in 1924. After completing his Ph.D. in 1929, Cattell did some teaching at University College, Exeter before becoming Director of the School Psychological Services and Clinic at Leicester in 1932.

In 1937, E. L. Thorndike invited Cattell to come for a year to Columbia University as a research associate. Subsequently, he also spent brief periods at Clark, Harvard, and Duke universities. During his time at Harvard, Cattell began seriously to develop the methodology for intensively studying the single case (P-technique factor analysis) in which he remained interested throughout his career. Cattell acknowledges that with his '… office … having been next door to Gordon Allport's, the latter may also have played an initiating part' but goes on to lament that although Allport, too, was '… in pursuit of uniqueness of personality … ' he had '… some unreadiness to come to grips with statistical models' (Cattell 1984, p. 141). Exhibiting a certain disdain for those who refused to use 'proper' methods, no matter how celebrated their status, was one of Cattell's stable characteristics.

Cattell worked for a time during the latter years of World War II in the Adjutant General's Office. In 1945 he became a research professor at the University of Illinois in Urbana-Champaign and remained there until 1973 when, in part due to age restrictions then in force, he retired. He moved on to the University of Hawaii where he maintained a special appointment for a few years. When that appointment ended, Cattell stayed in Honolulu and continued to work informally with several younger colleagues and to publish an occasional paper. One of his last publications (McArdle and Cattell 1994) brought to a close his 50-year affair (Cattell 1944) with the beguiling concepts of 'parallel proportional profiles' and 'confactor rotation,' more of which will be said later.

By far the longest and most productive portion of Cattell's scientific career occurred while he was Director of the Laboratory of Personality Assessment and Group Behavior (later the Laboratory for Personality Analysis) at the University of Illinois. It was from this setting that his influence on the development of the science of psychology hit its maximum. The sheer number of big ideas and level of productivity in following them up during those years are remarkable. His success in claiming his 'fair' share of CPU time on the Illiac, the university's automatic computer, was legendary. Although by today's standards, its capacity was meager, at the time the Illiac allowed Cattell to do large factor analyses in weeks and months that would otherwise have taken years.

1. *Contributions to the Methods and the Substance of Psychology*

Raymond B. Cattell is an exemplar of a class of psychological researchers who were heavily invested in the application of quantitative models to the study of behavior. Also included were the likes of Cyril Burt, Hans J. Eysenck, J. P. Guilford, Charles Spearman, Godfrey Thompson, and L. L. Thurstone. Let it be said openly that this is not a list of like-minded individuals who necessarily had more in common than not. Rather, it is a list of important contributors to the development of psychology as a science who relied substantially on the models and methods of factor analysis in their empirical research pursuits. In both strategy and tactics, they differed from each other in the use of factor analytic methods.

Because of Cattell's heavy reliance on factor analysis, it is not possible to separate the substance from the methods. In fact, in an earlier review of Cattell's research program in *Theories of Personality* (Hall and Lindzey 1978), the chapter heading is 'Cattell's Factor Theory.' Therefore, a discussion of his work can be usefully begun with a methodological introduction.

At the heart of Cattell's research program lay the basic equation of the common factor model. Cattell referred to it as the factor specification equation and wrote it as:

$$a_{ji} = b_{j1} \cdot f_{1i} + b_{j2} \cdot f_{2i} + \cdots + b_{jk} \cdot f_{ki} + u_{ji}$$

where a_{ji} is a score for person i on behavior (or act) a_j, and f_{qi} ($q = 1, 2, \ldots, k$) represent the ith person's endowments or scores on the k common factors, the b_{jq} ($q = 1, 2, \ldots, k$) are weights (factor loadings) that specify the amount to which any individual's score on the qth factor contributes to (in Cattell's theoretical framework, *determines*) his score on variable a_j, and u_{ji} is a contribution to a_{ji} that is unique to both the

behavior and the individual. The u_{ji} terms include both errors of measurement and portions of the observed score that are reliably measured, but fall outside the purview of the common factors defining the interrelations of a given set of variables.

The b_{jq} values, which Cattell referred to as situational indices (later, behavioral indices), brought the situation and the individual's traits into the same equation so that both individual and situational information could be used to predict behavior. Suppose, for example, a given individual had a very high endowment on the factor Cattell labeled Exvia, which was similar in nature to the individual differences dimension called Extraversion in many personality research and theory contexts. If, on the one hand, the b_j for Exvia with respect to performance of a given a_j was small, then that high level of Exvia had little or no bearing on that person's a_j score. On the other hand, however, if b_j for Exvia on performance of a given a_j was large, then that high level of Exvia had major influence in determining that person's a_j. By contrast, for the person with a near zero level of Exvia, little would be contributed to their performance on a_j by the Exvia factor even if the b_j for Exvia was large, because the product $b_{j,exvia} \cdot f_{exvia,i}$ would be close to zero.

This factor specification equation is thus a 'working model' of the individual differences orientation in that the differing amounts of the latent attributes possessed by different people (variation in the f_q) are 'converted' into differences in their scores on the observed variables (variation in the a_j). By knowing the amounts of a person's trait endowments and the magnitude of the contribution the traits made to manifest variables, one could predict an individual's scores on the manifest variables, that is, behavior.

Within that framework, the scientific agenda is rather clear: define the latent factors as clearly as possible, construct tests to measure them accurately, and develop empirically based estimates of the b values for the manifest variables, the a_j, of interest. Then one has made the prediction objective of science reachable and can also elaborate an explanatory system as well by detailing more and more regarding the nature of the factors (e.g., see how the factors behave across different media of observation and how stable they are across time, examine their genetic and environmental variance contributions, plot their age changes, etc.). Attacking these several objectives was a major part of Cattell's research program.

Thus, despite knowing the adaptations in the methods of the physical sciences had to be made in studying behavior, Cattell's methodological approach showed a deep reliance on fundamental principles of physical science. He believed that factors were causal influences on observable variables and that, in analytical work, when the factors had been properly rotated to their 'causal influence' identities, the factor loadings (the b_j values in the specification equation) would be proportional for a given set of behaviors (the a_j values in the specification equation) across different samples. This represented an explicit commitment to a fundamental objective of scientific research: the identification of relationships that remain *invariant* under different transformations. In his context, Cattell formalized this notion in the terms 'parallel proportional profiles' (the results) and 'confactor rotation' (the method) mentioned earlier. Along the way, he worked on the development of methods for assessing the similarity of factor loading patterns so that one could render some judgment about factorial invariance even without reaching the ideal parallel proportional profiles.

Cattell was also firmly committed to the principles of rigorous, quantitative measurement and scaling and once presented, with characteristic lack of modesty, a proposed more basic set of concepts (Cattell 1964) to replace the prevailing notions of reliability and validity. In so doing, he again challenged the status quo regarding the role of item homogeneity in evaluating the quality of measuring instruments.

Cattell placed a strict reliance on mathematical representations of phenomena of interest although, in his case, this amounted generally to using some variant of the common factor model. His insistence on the introduction of rigor into the definition of psychological concepts left its mark in a number of ways. He argued that ambiguity and confusion about concepts could be reduced by discarding as many of those in vogue as possible, discovering and defining new ones, and giving them new names to mark their unique status. In the process, he delivered such labels as 'cortertia' (cortical alertness) and 'threctia' (threat reactivity) to identify major individual differences in dimensions of personality.

A monograph on real-base, true zero scaling (Cattell 1972) illustrates the concern he held for fundamental issues of scaling, the commitment and energy with which he would attack a problem, provide a solution, and advance it for discussion and debate. Debate, however, was not always the reaction evoked by his proposals.

In addition to this strong analysis and modeling orientation, Cattell's research program was buttressed by several other key attributes. Where others might carve up the study of behavior to identify more clearly a piece for themselves in the broader scheme, Cattell did so as a way of more systematically approaching the whole. Much of his work was organized around the tripartite division of abilities, temperament, and motivation. But he enriched this classical triumvirate by distinguishing between surface and source traits, estimating heritability coefficients, and incorporating concepts of psychological states, learning theory, and situational influences. He introduced the multiple abstract variance analysis (MAVA) design to incorporate different family configurations into the estimation of heritability coefficients and in the design of correlational studies he saw not one, but six important

ways to interrelate variables, persons, and occasions of measurement. He systematized the latter in a heuristic often referred to as the 'data box' or covariation chart (Cattell 1952) which is still frequently encountered in the literature.

As an example of his capacity for orderly abstraction and the willingness and ease with which he would think 'big,' Cattell was not satisfied to name and report the results of his many factor analyses. Instead, he proposed the development of a Universal Index (U. I.) of factors, the first 15 of which (U. I. 1 through U. I. 15) were to be reserved for what he considered to be the well-established human ability factors. He then proceeded to assign the next 17 numbers (U. I. 16 through U. I. 32) for factors based on objective tests on which he had published. He invited further contributions to this indexing system but, as one might expect from even a passing acquaintance with human nature, one person's proposal for a Universal Index is hardly likely to be adopted readily by others.

In addition to a frequently observed trait of successful research scientists—a desire to be first and to receive credit for it—Cattell was gifted with a high level of intelligence and capacity for abstract thought. He made up many of the items for the Culture Fair Intelligence tests (Cattell 1954) himself; took delight in the rapidity with which he could solve items from 'would be' intelligence test item writers and, in return, relished 'stumping' them with a recently written item or two of his own.

The levels of abstraction with which Cattell felt quite comfortable were in many cases a far remove from data. For example, second-order factor analyses (factor analyzing the intercorrelations of factors) were routine in his laboratory. Third-order factor analyses, while not routine, were completed on several occasions to which published papers attest. Along the way, at least one fourth-order factor analysis was conducted.

Cattell was not inclined to be tolerant of others' lack of knowledge of his system. Those who worked in his laboratory were apt to be rebuffed when they would occasionally suggest ways by which he might strengthen the bonds of communication with his readers. 'Let them go and read my 1957 book' (Cattell 1957), was the reply heard on more than one occasion to the suggestion that a little more detail might be helpful to the reader. Nor, as was already mentioned, was he kindly diposed toward those he felt were not using the multivariate tools of the trade wisely.

Cattell was an avid reader (and writer) of poetry and also enjoyed biographies from which he enriched his own writing. His enormous legacy of books, chapters, and articles, numbering well over 500 publications, exhibit many examples of his caring about the craft of writing as well as the ideas being written about. He would search for just the right word or turn of phrase—and did so amazingly quickly! When he was well into his eighties the author once asked him in what areas of his intellectual functioning he could detect losses. He responded that he didn't act on visual input as rapidly as he used to when, for instance, driving his automobile. (This was true!) But, he very quickly added, he could still retrieve adjectives, etc., quite handily when he was writing and he was very pleased about it. Some of his 'classic' remarks, many of which are buried in footnotes, will likely go unheeded by the future generations, and that will be a stylistic as well as a substantive loss. For example, his characterization of the hyperplane as 'the footprint of a causal influence' or his accusation that by developing and making widely available the by then nearly ubiquitous orthogonal analytic rotation program, Varimax, Henry Kaiser had 'put opium on the market,' are illustrative.

Cattell's deep reliance on factor analysis was symptomatic of his belief that the workings of variables should be examined with as little interference as possible from the observer. The methods of factor analysis could then be used to tease out the key patterns of relationships. As any craftsman, Cattell felt strongly about having access to proper tools and worked to evolve factor analytic methods during the length of his career. He did this both by formalizing design with the presentation of the covariation chart or 'data box' and by restlessly pushing technological developments. The ever present need for faster, more efficient methods for conducting factor analysis if his research goals were to be realized, directly resulted in software applications such as Maxplane, Procrustes, and Rotoplot to facilitate rotation of factors to a final solution. The Rotoplot program, for instance, displayed the pairwise factor (actually, reference vector) plots on an oscilloscope and filmed them by 35 mm camera mounted on the oscilloscope. As soon as the film strip could be processed, Cattell would project the plots on his office wall and write down the tangents of the angles of rotation for another run, urgently trying to achieve 'maximum simple structure' so that factors could be interpreted and the article written and submitted. At the time that Rotoplot was invented, others who were doing graphical rotation of factors were having to have each pairwise plot rendered by triangle and T-square.

2. *Bridging Psychology's Past and its Future*

Cattell and his research program represent an important bridge between psychology's past and its future. While he was alive, Cattell was a personal, as well as scientific link to the psychology's past. On occasion, he would tell of conversations (sometimes, encounters) with the likes of Burt, Fisher, Pearson, and Spearman. Although he was not a student of Thurstone, he seized on the tool of multiple factor analysis which Thurstone was so instrumental in bringing to psychologists and future refined it to his

own ends. At the same time, his zeal in promoting, and passing along to future generations of researchers not only his substantive findings but also the methods that he thought were appropriate and necessary for psychological research, forged important links to the future. He was a prime mover in establishing the Society of Multivariate Experimental Psychology (SMEP) and the founding of the society's journal, *Multivariate Behavioral Research* (MBR). He helped to launch MBR in 1966 with a provocative lead article entitled 'Multivariate Behavioral Research and the Integrative Challenge,' in which he summarized his own version of the two general traditions of psychological research—the experimental and the individual differences traditions—and pointed out the directions for future research that would elevate the study of behavior to new levels of competence. In that same year the first edition of the *Handbook of Multivariate Experimental Psychology*, for which Cattell was both the editor and the major contributor, was published. A second edition followed 22 years later (Nesselroade and Cattell 1988).

McArdle (1984) credited Cattell with having a profound effect on the development of what is now referred to generally as 'structural modeling' the fitting of quantitative models to covariance matrices (and occasionally the associated arithmetic means, as well), which has grown to enormous proportion in current day psychological research. This connection reinforces the amazingly productive service Cattell extracted from the factor analytic model. Cattell's empirical and theoretical work, as no one else's, has helped to clarify the distinction between latent and manifest variables and the key properties of both.

The contributions that Cattell made both substantively and methodologically were many. Here, we have only touched on some of the major ones. He developed and presented one of the most systematic and encompassing theories of personality that psychology has ever seen. Along the way, he introduced a large number of methodological innovations, some of which have made their way into today's behavioral science argot with an ease that today gives no hint of Cattell's pioneering contribution. The 'scree test,' 'Procrustes' rotation, and 'P-technique' come readily to mind.

The theory of fluid and crystallized intelligence, on which Cattell collaborated closely with Horn (Horn and Cattell 1966) remains very much in currency today as far as so-called psychometrically oriented models of cognition are concerned. The apparently different age gradients of these two kinds of abilities has helped to stimulate thinking not only about the complexity of the organism and the nature of its development across the lifespan (e.g., Baltes 1987), but also about the need to consider what level of aggregation of variables is optimal for a given purpose.

Despite Cattell's strong investment in the study of human abilities (Cattell 1971), he did not neglect the temperament and motivation components of his research program. For example, the fruits of his efforts to measure temperament by questionnaire, especially, are evident today in the continuing usage of the High School Personality Questionnaire (HSPQ) and the 16 Personality Factor questionnaire (16PF). Some of the work of which Cattell was most proud and for which he had the highest hopes sprang from his contributions to the study of motivation. His concept of the dynamic calculus, which was an attempt to write a systematic prediction equation for behavior that included trait, state, and dynamic aspects was refined into what he came to call structured learning theory. In the author's view, this was a key contribution in Cattell's own evaluation of his scientific legacy.

Many of the individuals who received their scientific training in Cattell's laboratory are still active and productive today. They, in turn, are passing on to their students the concepts of the data box, fluid and crystallized intelligence, covariation designs, factorial invariance, states versus traits, etc.

Despite the comprehensive perspective, the numerous insights and innovations, and the sheer bulk of his scholarly output, which spanned parts of eight different decades, Raymond B. Cattell's place in the history of psychology will not be identified without controversy for the foreseeable future. A lifetime achievement award for his contributions to psychology to be given by the American Psychological Foundation was announced in early 1997. Although Cattell had never taken kindly to the science 'establishment,' he had reached a point in his journey where such a high level of recognition for his scientific work meant something to him. But it was not to be. Some of his more philosophical, provocative writings having to do with the roles in our society of genetics, evolution, and public policy had generated bitter and sufficient resentment that, after some 'negotiating,' the award was withdrawn prior to the 1997 annual meeting of the American Psychological Association in Chicago, where it was to have been given. Cattell died the following February 2, at the age of 92.

See also: Adult Cognitive Development: Post-Piagetian Perspectives; Behavioral Assessment; Darwin, Charles Robert (1809–82); Factor Analysis and Latent Structure: Overview; Galton, Sir Francis (1822–1911); Lifespan Theories of Cognitive Development; Multivariate Analysis: Overview; Parsons, Talcott (1902–79); Personality Structure; Personality Theories; Quetelet, Adolphe (1796–1874); Thorndike, Edward Lee (1874–1949)

Bibliography

Baltes P B 1987 Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* **23**(5): 611–26

Cattell R B 1944 'Parallel proportional profiles' and other principles for determining the choice of factors by rotation. *Psychometrika* **9**: 267–83
Cattell R B 1952 *Factor Analysis*. Harper, New York
Cattell R B 1954 *Culture Fair Intelligence Tests, Scales 1, 2, and 3, Forms A and B, rev. edn*. IPAT, Champaign, IL
Cattell R B 1957 *Personality and Motivation Structure and Measurement*. World Book Co., New York
Cattell R B 1964 Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology* **55**: 1–22
Cattell R B 1971 *Abilities: Their Structure, Growth, and Action*. Houghton Mifflin, Boston
Cattell R B 1972 Real base, true zero factor analysis. *Multivariate Behavioral Research Monographs* **72**(1): 1–162
Cattell R B 1984 The voyage of a laboratory, 1928–1984. *Multivariate Behavioral Research* **19**: 121–74
Child D 1998 Raymond Bernard Cattell (1905–1998). *British Journal of Mathematical and Statistical Psychology* **51**: 353–7
Goldberg L R 1968 Objective Personality and Motivation Tests—theoretical introduction and practical compendium—Cattel R B and Warburton F W. *Contemporary Psychology* **13**: 617–9
Hall C S, Lindzey G 1978 *Theories of Personality*, 3rd edn. Wiley, New York
Horn J L, Cattell R B 1966 Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology* **57**: 253–70
McArdle J J 1984 On the madness in his method: R. B. Cattell's contributions to structural equation modeling. *Multivariate Behavioral Research* **19**: 246–67
McArdle J J, Cattell R B 1994 Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research* **29**(1): 63–113
Nesselroade J R, Cattell R B (eds.) 1988 *Handbook of Multivariate Experimental Psychology*, 2nd edn. Plenum, New York

J. R. Nesselroade

Causal Counterfactuals in Social Science Research

The term 'counterfactual conditional' is used in logical analyses to refer to any expression of the general form: 'If A were the case then B would be the case.' In this usage, A is usually false or untrue in the world so that A is 'contrary to fact' or counterfactual. Examples abound. 'If kangaroos had no tails, they would topple over' (Lewis 1973b). 'If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone' (Rubin 1978). Perhaps the most obnoxious counterfactuals in any language are those of the form: 'If I were you, I would … .'

Lewis (1973a) observed the connection between counterfactual conditionals and references to causation. He finds these logical constructions in the language used by Hume in his famous discussion of causation. Hume defined causation twice over. He wrote 'we may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second*. Or, in other words, *where, if the first object had not been, the second never had existed*' (Lewis 1973a, italics are Lewis's).

Lewis draws attention to the comparison between the factual first definition where one object *is* followed by another and the counterfactual second definition where, counterfactually, it is supposed that if the first object 'had not been' the second object would not have been either.

It is the connection between counterfactuals and causation that makes them relevant to social science research. From the point of view of some authors, it is difficult, if not impossible, to make any sense of causal statements without using counterfactual language (Lewis 1973a, Holland 1986, Rubin 1978, Robins 1985, 1986). Other authors are concerned that using such language gives an emphasis to unobservable entities that is inappropriate in the analysis of empirical data (Dawid 1997, Shafer 1996). The discussion here accepts counterfactuals in discussions of causation and will explain their role in the estimation of causal effects based on the work of Neyman (1923, 1935), Rubin (1974, 1978), and others.

We begin with a simple observation. Suppose that we find that a student's test performance changes from a score of X to a score of Y after some educational intervention. We might then be tempted to attribute the pretest-posttest change, $Y-X$ to the intervening educational experience, that is, to use the gain score as a measure of the improvement due to the intervention. However, this is social science and not the tightly controlled 'before-after' measurements made in a physics laboratory. There are many other possible explanations of the gain, $Y-X$. Some of the more obvious are: simple maturation, other educational experiences occurring during the relevant time period, and differences in either the tests or the testing conditions at pre- and posttests. Cook and Campbell (1979) provide a classic list of 'threats to internal validity' that address many of the types of alternative explanations for apparent causal effects of interventions. For this reason, it is important to think about the real meaning of the attribution of cause. In this regard, Lewis's discussion of Hume serves us well. From it we see that what is important is what the value of Y *would have been* had the student *not had* the educational experiences that the intervention entailed. Call this score value, Y^*. Thus enter counterfactuals. Y^* is not directly observed for the student, that is, they *did have* the educational intervention of interest, so asking for what their posttest score *would have been* had they *not had it* is asking for information collected under conditions that are contrary to fact. Hence, it is not the difference $Y-X$ that is of interest, but the difference $Y-Y^*$, and the gain score has causal significance relative to the effect of the educational experience only if X can serve as a substitute for the

counterfactual Y^*. In physical-science laboratory experiments such a substitution is often easy to make, but it is rarely believable in many social science applications of any consequence.

A formal model or language for discussing the problem of estimating causal effects (of the form $Y - Y^*$ rather than $Y - X$) was developed by Neyman (for randomized experiments) and Rubin (for a wide variety of causal studies) and will be called the Neyman/Rubin model for causal effects here.

1. *Prospective Causal Studies*

The Neyman/Rubin model is most easily understood in the context of a *prospective causal study* that has the general structure specified by this sequence of events.

(a) Subjects or experimental units of study are identified.

(b) Baseline or pretest information about these units is recorded.

(c) The units are either assigned to (in a controlled study) or select themselves to (in studies without the control of assignment) exposure to one of the treatment conditions or interventions of the study.

(d) These units are then subsequently exposed to their assigned or self-selected treatment condition (and each unit is affected by this exposure in a manner that is unrelated to the exposure conditions of the other units).

(e) At an appropriate later time an outcome, endpoint, or posttest measure is recorded for each unit in the study.

The type of study that this five-part schema is intended to cover includes most randomized comparative experiments as well as many types of pretest/posttest quasiexperiments or observational studies. It should be emphasized that, properly interpreted, the Neyman/Rubin model has application to other types of causal studies (see Holland 1988, Holland and Rubin 1988, Robins 1997), but, for our purposes here, prospective causal studies are already sufficiently complicated and inclusive.

The condition mentioned parenthetically in (d), that the exposure conditions of the other units do not affect the outcomes associated with a given unit, is very important, and is clearly an assumption that would not be true in general. For example, in the study of infectious diseases your vaccination will affect my likelihood of contracting polio. Rubin explicitly identifies this assumption calling it the Stable Unit-Treatment Value Assumption, or SUTVA. SUTVA will be assumed throughout this discussion.

2. *The Neyman/Rubin Model*

The version of the Neyman/Rubin model used here is adapted from that of Holland (1986), and differs from the original versions by Neyman (1935) and Rubin (1978) mainly in its emphasis on population quantities. One of the main benefits of this model is that it identifies certain 'counterfactual conditional expectations' as the location of key assumptions about the inferential structure of any causal study and its resulting data.

The prospective causal study begins with the 'units,' 'subjects,' or 'cases' of the study, and the ith unit is denoted by the subscript 'i.' It may help the reader to imagine that this is a discussion about a very large sample of units. Denote the population of units under study by P. For the most part P will lie quietly in the background without being noticed. The baseline information that is collected or recorded for unit i will be denoted as a *vector* of numerical information, z_i.

There is a 'causal' variable denoting a set of possible 'treatments' or 'exposure' conditions to which each unit in the study could be exposed. For simplicity we assume that these are only two treatment conditions denoted by $x = 1$ (treatment) and $x = 0$ (control). A more complicated version of this would let x be a number representing the 'strength' of the treatment level, but we will just use the dichotomous case here.

An important aspect of causal variables designating such treatments levels or intervention conditions is the assumption that the level of exposure for any unit *could have been different* from what it actually was. This condition excludes 'attributes' of units (such as race, gender, age, or pretest score) as causal variables in the sense that such attributes *cannot* have 'unit-level causal effects' in the sense that we will define below. This idea is discussed more extensively in Holland (1986, 1988), and is mentioned again in another context at the end of the present discussion. Each unit is exposed to one treatment level and the value of x to which i is exposed is denoted by x_i.

Finally we come to the outcomes or dependent variables in the study, and here is where a special notation is needed. We let $Y_i(x)$ denote the (numerical) response that would be recorded for unit i if unit i were exposed to treatment level, x. For each i, $Y_i(x)$ is a function of x. It should be emphasized that $Y_i(x)$ is not directly observed unless $x_i = x$. This is an important point because it is crucial to realize that the $\{Y_i(x)\}$ do not denote *observed data* like z_i and x_i do, but rather the $\{Y_i(x)\}$ are 'potential outcomes' that lie behind the observed values of the outcome variable. For this reason, we denote the potential observations by capital letters to distinguish them from quantities that are directly observable, which we denote by lower case letters.

The connection between the potential outcomes and the actually observed outcomes is then given by the equation

$$y_i = Y_i(x_i) \tag{1}$$

where y_i is the observed outcome or value of the dependent variable for unit i. The idea behind Equation (1) is that to get from the potential outcomes

$\{Y_i(x)\}$ to an observed outcome we must select the value of x in $Y_i(x)$ to be the value to which i is actually exposed, that is, $x = x_i$, and then we obtain y_i from the potential observations, $\{Y_i(x)\}$, via Equation (1).

The *observed data* for each unit i is the vector (z_i, x_i, y_i). The potential outcomes, $\{Y_i(x)\}$, are never observed for *all* values of x for a fixed unit, i, but only for the specific x-value to which i is actually exposed, x_i. It is sometimes said that the $\{Y_i(x)\}$ are 'counterfactual' because they are not actual observations. Here they are called *potential observations* because they could have been observed had x_i been different than it was.

It is helpful to use a notation such as, $E(y)$, to mean the average value of y_i across the (large number of) units in P. Furthermore, an expression such as

$$E(y \mid x = a) \tag{2}$$

will mean the average value of y_i across all of the (large number of) units in P for which $x_i = a$. The use of the expectation notation is to make certain quantities clearer in their meaning, and may be justified from either a frequentist or Bayesian point of view. It should be noted that within the expectation notation, the subscript i, denoting the unit, is suppressed because, within the scope of the $E(\ )$ operator, i is averaged over.

3. *Using the Neyman/Rubin Model*

An important fact about the Neyman/Rubin model is the Fundamental Problem of Causal Inference (Holland 1986), which is: It is impossible in principle to *observe* $Y_i(x)$ for more than one value of x for any one unit, i. Any procedure that claims to have avoided the Fundamental Problem of Causal Inference can always be shown to be based on untestable assumptions. Sometimes such assumptions are plausible, and sometimes they are not.

A basic definition that we are now in a position to make is that of a 'unit-level causal effect.' Because we are restricting attention to the simple case of two treatment levels, $x = 0/1$ we may restrict attention to these differences

$$Y_i(1) - Y_i(0) = \text{the casual effect of } x = 1 \text{ relative to } x = 0 \text{ for unit } i \tag{3}$$

In the Neyman/Rubin notation, the unit-level causal effects are the *basic* quantities of interest in causal inference. However, the Fundamental Problem of Causal Inference is now immediately seen to *be fundamental* because it implies that unit-level causal effects are, themselves, *never directly observable*. Thus, we are always reduced to making assumptions that allow us to make some sort of conclusion or inference about these causal effects. This is the place where the Neyman/Rubin model might appear to be impractical for applied research, that is, because its most basic parameters, the unit-level causal effects, are not directly observable. Furthermore, this is exactly the place where the potential exposability of all of the levels of x to any unit is seen to be crucial to the foundations of the theory. The definition of causal effect requires this assumption so that the difference, $Y_i(1) - Y_i(0)$, is meaningful. It is the Fundamental Problem of Causal Inference and this definition of causal effect that makes causal inference both more interesting and more difficult than the simple computation of correlational and associational measures. The unit-level causal effects mimic the comparison between a 'factual' and a 'counterfactual' identified in the quote mentioned earlier from Hume by Lewis.

It is now time to introduce the notion of an Average Causal Effect (ACE). In general, an ACE is any average of unit-level causal effects. The most general ACE has the form

$$\text{ACE} = E[Y(1) - Y(0) \mid A] \tag{4}$$

where A denotes some collection of units defined in terms of either z_i or x_i or both. In Equation (4) we again suppress the subscript i because it is being averaged over. As an example of an A in Equation (4) we might use A = 'all the units in the study,' in which case the ACE is the average causal effect over all of P. But other cases might be of interest, for example, A = 'all units where i is male and for whom $x_i = 1$.' In this case the ACE is for the males in treatment group 1. Here we restrict our attention to the ACE that is called the 'effect of the treatment on the treated' in which A denotes all of the units for which $x_i = 1$, that is

$$\begin{aligned}\text{ACE} &= E[Y(1) - Y(0) \mid x = 1] \\ &= E[Y(1) \mid x = 1] - E[Y(0) \mid x = 1]\end{aligned} \tag{5}$$

Up to now, we have simply defined the basic structure of the data collection design as well as the causal connection between the potential outcomes and the causal variable x, that is, Equation (3). We have not yet identified the connection between any quantities that could be estimated with data and the causal parameters given by either the unit-level causal effects in Equation (3) or the average causal effects in Equation (5). This leads us to the '*prima facie* Average Causal Effects,' (FACEs). The FACEs are what can be estimated from the data. To parallel the ACE in Equation (5) we examine the FACE which is simply the difference between the mean of the outcome variable observed in each treatment group, that is

$$\text{FACE} = E[y \mid x = 1] - E[y \mid x = 0] \tag{6}$$

If we substitute the definition of y in terms of the potential observations, $\{Y_i(x)\}$, that is given in Equation (1) into Equation (6) we obtain

$$\text{FACE} = E[Y(1) \mid x = 1] - E[Y(0) \mid x = 0] \quad (7)$$

Finally if we combine Equation (5) and Equation (7) we obtain the following basic formula that relates the ACE to the FACE

$$\text{FACE} = \text{ACE} + \text{BIAS} \quad (8)$$

where

$$\text{BIAS} = E[Y(0) \mid x = 1] - E[Y(0) \mid x = 0] \quad (9)$$

The BIAS term contains two parts, one *factual*, that is, $E[Y(0) \mid x = 0] = E[y \mid x = 0]$, and the other *counterfactual*, that is, $E[Y(0) \mid x = 1]$. The factual part is just the mean of y_i for those units with $x_i = 0$. The counterfactual part is the mean of $Y_i(0)$ for those units for whom $x_i = 1$. $E[Y(0) \mid x = 1]$ is a quantity for which there can never be any data because the conditioning event makes the quantity being averaged over a counterfactual. Thus, it is a counterfactual conditional expectation. The value of such counterfactual parameters is that they pinpoint exactly where assumptions must be made that allows causal inference to take place using empirical data. When BIAS = 0, we have FACE = ACE and the empirical FACE equals the causal ACE.

An important condition that insures that BIAS = 0 is the construction of x_i by random assignment which forces x_i and $Y_i(0)$ to be statistically independent of each other as functions of i over P (Holland 1986). When this independence holds, $E[Y(0) \mid x = 1] = E[Y(0) \mid x = 0]$ and BIAS = 0.

4. Empty Counterfactuals

There is an unsatisfactory and rather misleading use of counterfactuals that sometimes arises in social science research. It occurs when the counterfactual condition, that is, the 'if A were the case' part could never occur in any real sense. Such empty counterfactuals arise when a nonmanipulable factor in a causal study is described as having an 'effect' on some outcome. Examples easily come about in casual causal talk. The effect of gender on salary suggests considering 'what her salary would have been had she been a man.' The effect of test performance on future employment suggests 'the job he would have had had he scored higher on the test.' The effect of English language proficiency on a math test in English suggests 'the mathematics score a non-English speaker would have received had he or she been an English speaker.' These empty counterfactuals arise when the value of a variable for a factor in a study could not have been other than the value that it was. The interesting and useful counterfactuals arise in those cases when the variable could have had a different value that it did for the individuals in a study, at least in principle. Judgment as to when a counterfactual is empty or not is not always easy and may require careful thought in many cases. Consider the examples 'if I were you, I would, …' Lewis's kangaroo and Rubin's aspirin in the opening paragraph. These represent very different kinds of counterfactuals from this perspective. The first is as empty as they get, the emptiness of the second depends on how the kangaroo might not have a tail (our imagination vs. an axe), while Rubin's aspirin could be taken or not.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Counterfactual Reasoning: Public Policy Aspects; Counterfactual Reasoning, Qualitative: Philosophical Aspects; Counterfactual Reasoning, Quantitative: Philosophical Aspects; Internal Validity; Quasi-Experimental Designs

Bibliography

Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin, Boston

Dawid A P 1997 Causal Inference without Counterfactuals. Research Report No. 188, Department of Statistical Science, University College, London

Holland P W 1986 Statistics and causal inference. *Journal of the American Statistical Association* **81**: 945–70

Holland P W 1988 Causal inference, path analysis and recursive structural equations models. In: Clogg C (ed.) *Sociological Methodology*. American Sociological Association, Washington, DC, pp. 449–84

Holland P W, Rubin D B 1988 Causal inference in retrospective studies. *Evaluation Review* **12**: 203–31

Lewis D K 1973a Causation. *Journal of Philosophy* **70**: 556–67

Lewis D K 1973b *Counterfactuals*. Harvard University Press, Cambridge, MA

Neyman J 1923 Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczki* **10**: 1–51 (in Polish: English trans. Dabrowska D, Speed T 1991 *Statistical Science* **5**: 463–80)

Neyman J 1935 Statistical problems in agricultural experimentation. *Supplement of the Journal of the Royal Statistical Society* **2**: 107–80

Robins J M 1985 A new theory of causality in observational survival studies—Application of the healthy worker effect. *Biometrics* **41**: 311

Robins J M 1986 A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**: 1393–1512

Robins J M 1997 Causal inference from complex longitudinal data. In: Berkane M (ed.) *Latent Variable Modeling with Applications to Causality*. Springer-Verlag, New York, pp. 69–117

Rubin D B 1974 Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**: 688–701

Rubin D B 1978 Bayesian inference for casual effects: The role of randomization. *Annals of Statistics* **6**: 34–58

Shafer G 1996 *The Art of Causal Conjecture*. MIT Press, Cambridge, MA

P. W. Holland

Causal Inference and Statistical Fallacies

1. Generalities

The pairing of causality and fallacies may seem idiosyncratic. In fact it nicely captures the point that many statistical fallacies, i.e., plausible-seeming arguments that give the wrong conclusion, hinge on the overinterpretation or misinterpretation of statistical associations as implying more than they properly do. The article begins by discussing three main views of causality, briefly indicating the scope for fallacious arguments and then at the end returns to discuss some fallacies in slightly more detail. See *Graphical Models: Overview*.

The very long history in the philosophical literature of discussions of causality is largely irrelevant for these purposes. It typically regards a cause as necessary and sufficient for an effect: all smokers get lung cancer, all lung cancer patients smoke. Here the concern is with situations with multiple causes, even if one is predominant, and where explicit or implicit statistical or probabilistic considerations are needed.

2. Notions of Causality

2.1 Causality as Stable Association

Suppose that a study, or preferably several different but related studies, shows that two features, C and R, of the individuals (people, firms, communities, households, etc.) under investigation are associated. That is, if we take, to be explicit, positive monotone association, individuals with high values of C tend to have high values of R and vice versa. For example C and R might be test scores at a given age in arithmetic and language, or level of crime and unemployment rate in a community.

Under what circumstances might one reasonably conclude that C is a cause of a response R, or at least make some step in the direction of that conclusion? And what would such a statement of causality mean?

2.1.1 Symmetric and directed relations. Association is a symmetric relation between two, or possibly more, features. Causality is not symmetric. That is, if C is associated with R then R is associated with C, but if C is a cause of R then R is not a cause of C. Thus the first task, given any two features C and R, is to distinguish the cases where:

(a) C and R are to be regarded as in some sense on an equal footing and treated in a conceptually symmetric way in any interpretation.

(b) One of the variables, say C, is to be regarded as explanatory to the other variable, R, regarded as a response. That is, if there is a relation, it is regarded asymmetrically.

Often significance tests for the existence of association and of dependency are identical. The distinction being studied here is a substantive one of interpretation. Failure to observe this distinction leads to the fallacy of the overinterpreted association.

2.1.2 Graphical representation. A useful graphical representation shows two variables X_1 and X_2, regarded on an equal footing, if associated, as connected by an undirected edge, whereas two variables such that C is explanatory to R, if connected, are done so by a directed edge. See Fig. 1a and Fig. 1b.

There are two possible bases for the distinction between explanatory and response variables. One is that features referring to an earlier time point are explanatory to features referring to a later time point. The second is a subject-matter working hypothesis based for example on theory or on empirical data from other kinds of investigation. Thus the weight of a child at one year is a response to maternal smoking behavior during pregnancy. In such situations the relevant time is not the time when the observation is made but the time to which the features refer, although of course observations recorded retrospectively are especially subject to recall biases.

As an example of the second type of explanatory-response relation, suppose that data are collected on diabetic patients assessing their knowledge of the disease and of their success in managing their disease, as measured by glucose control. These data may well refer to the same time point and it is not inconceivable that, for example, patients with poor glucose control are thereby encouraged to learn more about their disease. Nevertheless, as a working hypothesis, one might interpret the data assuming that knowledge, C, is explanatory to glucose control, R, considered as a response. This is represented in simple graphical form in Fig. 1b by the directed edge from C to R.

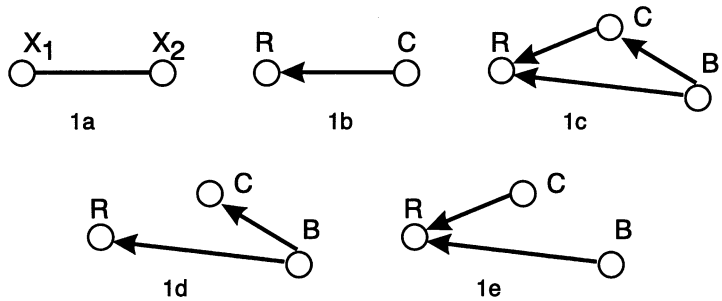

Figure 1
(a) Undirected edge between two variables X_1, X_2 on an equal footing. (b) Directed edge between explanatory variable C and response variable R. (c) General dependence of response R on B, C. (d) Special situation with $R \perp\!\!\!\perp C \mid B$. (e) Special situation with $B \perp\!\!\!\perp C$ corresponding in particular to randomization of C

In summary the first step towards causality is to require good reasons for regarding C as explanatory to R as a response and that any notion of causal connection between C and R, and there may be none, is that C is a cause of R, not the other way round.

We may talk about the 'fallacy of the incorrect direction' when the explanatory-response relation is identified in the wrong direction.

2.1.3 Common explanatory variables. Next, consider the possibility of one or more common explanatory variables. For this, suppose that B is potentially explanatory to C and hence also to R. There are a number of possibilities of which the most general is shown in Fig. 1c with directed edges from B to C, from C to R, and also directly from B to R. On the other hand, if the relation were that represented schematically in Fig. 1d, the only dependence between C and R is that induced by their both depending on B. Then C and R are said to be conditionally independent given B, sometimes conveniently written $R \perp\!\!\!\perp C \mid B$. There is no direct path from C to R that does not pass via B. Such relations are typically assessed empirically by some form of regression analysis. In such a situation, one would not regard C as a cause of R, even though in an analysis without the background variable B there is a statistical dependence between the two.

This discussion leads to one definition used in the literature of C being a cause of R, namely that there is a dependence between C and R and that the sign of that dependence is unaltered whatever variables B_1, B_2 etc., themselves explanatory to C, are considered simultaneously with C as possible sources of dependence. This definition has a long history but is best articulated by I. J. Good and P. Suppes. A corresponding notion for time series is due to N. Wiener and C. W. J. Granger. This definition underlies much empirical statistical analysis in so far as it aims towards causal explanation.

The definition entertains all possible alternative explanatory variables. In implementation via an observational study one can at best check that the measured background variables B do not account for the dependence between C and R. The possibility that the dependence could be explained by variables explanatory to C that have not been measured, i.e., by so-called unobserved confounders, is less likely the larger the apparent effect, and can be discounted only by general plausibility arguments about the field in question. Sensitivity analysis may be helpful in this: that is, it may be worth calculating what the properties of an unobserved confounder would have to be to explain away the dependence in question. For further details see Rosenbaum (1995) and the entry *Observational Studies: Overview*.

Mistaken conclusions reached via neglect of confounders, observed or unobserved, may be called 'fallacies of the neglected confounder.'

2.1.4 Role of randomization. The situation is substantially clarified if the potential explanatory variable C is a randomized treatment allocated by the investigator. Then in the scheme sketched in Fig. 1e there can be no edge between the B's and C since such dependence would be contrary to randomization, i.e., to each individual under study being equally likely to receive each treatment possibility. In this situation an apparent dependence between C and R cannot be explained by a background variable as in Fig. 1d. It is in this sense that it is sometimes stated, especially in the statistical literature, that causality can be inferred from randomized experiments and not from observational studies. It is argued here that while other things being equal, randomized experiments are greatly to be preferred to

Figure 2
(a) Intermediate variable I accounting for overall effect of C after ignoring I; $R \perp\!\!\!\perp C \mid I$. (b) Correlated variables C, C^* on an equal footing and both explanatory to response R

observational studies, difficulties of interpretation, sometimes serious, remain.

2.1.5 Intermediate variables. In Sect. 2.1.4 the variables B have been supposed explanatory to C and hence to R. For judging a possible causal effect of C it would be wrong to consider in the same way variables intermediate between C and R, i.e., variables I that are responses to C and explanatory to R. They are valuable in clarifying the nature of any indirect path between C and R, but the use of I in a regression analysis of R and C would not be correct in assessing whether such a path exists. If R is independent of C given an intermediate variable I, but dependent on I, then C may still have caused I and I may be a cause of R.

For instance suppose that C represents assignment to a new medical regimen as compared with a control regimen and that the former, but not the latter, eventually induces lower blood pressure, I, which in turn induces a reduced cardiac event rate, R; see Fig. 2a. Does the new regime cause a reduced cardiac event rate? If R is conditionally independent of C given I, it would be reasonable to say that the regimen does cause a reduction in R and that this reduction appears to be explained via improved blood-pressure control.

2.1.6 Explanatory variables on an equal footing. In some ways an even more delicate situation arises if we consider the role of variables C^* on an equal footing with the variable C whose causal status is under consideration; see Fig. 2b. If the role of C is essentially the same whether or not C^* is conditioned, i.e., whether or not C^* is included in the regression equation, there is no problem, at least at a qualitative level. On the other hand, consider the relatively common situation where there is clear dependence on (C, C^*) as a pair but that either variable on its own is sufficient to explain the dependence. There are then broadly three routes to interpretation:

(a) To regard (C, C^*) collectively as the possibly causal variables.

(b) To present at least two possibilities for interpretation, one based on C and one on C^*.

(c) To obtain further information clarifying the relation between C and C^*, establishing for instance that C^* is explanatory to C and that the appropriate interpretation is to fix C^* when analysing variations of C.

For example, suppose that C and C^* are respectively measures of educational performance in arithmetic and language of a child, both measured at the same age, and that the response is some adult feature. Then the third possibility is inapplicable; the first possibility is to regard the variables as a two-dimensional measure of educational performance and to abandon, at least temporarily, any notion of separating the role of arithmetic and language.

In summary, this first broad notion of causality is that of a statistical dependency that cannot be explained away via an eligible alternative explanation.

2.2 Causality as the Effect of Intervention

2.2.1 Counterfactuals. While the notion of causality discussed in Sect. 2.1.6 is certainly important and is strongly connected with the approach adopted in many empirical statistical studies, it does not, however, directly capture a stronger interpretation of the word causal. This is connected with the idea of hypothetical intervention or modification. Suppose for simplicity of exposition that C takes just two possible forms, to be called presence and absence. Thus presence might be the implementation of some program of intervention, and absence a suitable control state. For monotone relations, one may say that the presence of C causes an increase in the response R if an individual with C present tends to have a higher R than that same individual would have had if C had been absent, other things being equal.

Slightly more explicitly, let B denote all variables possibly explanatory to C and suppose that there are no variables C^* to be considered on an equal footing to C. Consider for each individual two possible values of R, R_{pres}, R_{abs} that would arise as C takes on its two possible values, present and absent, and B is fixed. Then presence of C causes, say, an increase in R if R_{pres} is in some sense systematically greater than R_{abs}.

We now discuss this notion, which has its origins at least in part in J. Neyman's and R. A. Fisher's work on design of experiments and in the studies of H. A. Simon and has been systematically studied and fruitfully applied by D. B. Rubin.

For a given individual, only one of R_{pres} and R_{abs} can be observed, corresponding to the value of C actually holding for that individual. The other value of R is a so-called counterfactual whose introduction seems, however, essential to capture the notion hinted at above of a deeper meaning of causality.

2.2.2 Differences in counterfactuals. The simplest and least demanding relation between the two values of R is that over some populations of individuals under study the average of R_{pres} exceeds that of R_{abs}. This is a notion of an average effect and is testable empirically in favorable circumstances. A much stronger requirement is that the required inequality holds for every individual in the population of concern. Stronger still is the requirement that the difference between the two values of R is the same for all individuals, i.e., that for all individuals

$$R_{pres} - R_{abs} = \Delta$$

This is called, in the language of the theory of the design of experiments, the assumption of unit-treatment additivity.

Now these last two assumptions are clearly not directly testable and can be objected to on that account. The assumptions are indirectly testable, to a limited extent at least. If the individuals are divided into groups, for example on the basis of one or more of the background variables B, the assumptions imply for each individual observed that the difference between the two levels of R has the same sign in the first case and that it is the same, except for sampling errors, in the second. In Sect. 2.2.3 the consequences of the possibly causal variable C having a very different effect on different individuals are discussed.

2.2.3 Ignorable treatment allocation. In Sect. 2.2.2 it has been tacitly assumed throughout that the two possible values of R for each individual depend on the value of C for that individual and would be unaffected by reallocation of C to other individuals. That is, the effects of C act independently as between different individuals. These considerations have a strong bearing on the appropriate definition of a unit of study. For example, in a comparison of different methods of class teaching of children, the unit would primarily be a whole class of students, i.e., the whole group who are taught and work together.

2.2.4 Intrinsic variables. There is an important restriction implicit in this discussion. It has to be meaningful in the context in question to suppose that C for an individual might have been different from how it in fact is. This is relevant only to variables that appear solely as explanatory variables. For example they may be variables measured at base-line, i.e., at entry into a study. Any intermediate variable by its nature of being at some point a response is potentially manipulable. Purely explanatory variables can be divided into intrinsic variables, essentially defining characteristics of the individual, and potential explanatory variables, which might therefore play the role of C in the present discussion. Intrinsic variables should not be regarded as even potentially causal in the present sense. For example the gender of an individual would in most contexts be regarded as an intrinsic characteristic. The question 'what would R have been for this woman had she been a man, other things being held fixed?' is in many, although not quite all, contexts meaningless.

2.2.5 Variables to be held fixed. Finally, care is essential in defining what is to be held fixed under hypothetical changes of C. Certainly responses I to C are not fixed. Variables, B, explanatory to C are held fixed. There is an essential ambiguity for variables C^* on an equal footing with C. This is strongly connected with the issue of which explanatory variables to include in empirical regression analyses.

2.3 Causality as Explanation of a Process

There is a third notion of causality that is in some ways more in line with normal scientific usage. This is that there is some understanding, albeit provisional, of the process that leads from C to R. This understanding typically comes from theory, or often from knowledge at a hierarchical level lower than the data under immediate analysis. Sometimes, it may be possible to represent such a process by a graph without directed cycles and to visualize the causal effect by the tracing of paths from C to R via variables I intermediate between C and R. Thus the effect of diet in an epidemiological study may be related to the physiological processes underlying the disease under study, the effect of pharmaceutical products related to the pharmaco-dynamics of their action, the effect of interventions at a community level related to ideas of individual psychology, and so on.

This last notion of causality as concerned with generating processes is to be contrasted with the second view of causality as concerned with the effects of intervention and with the first view of causality as stable statistical dependence. These views are complementary, not conflicting. Goldthorpe (1998) has argued for this third view of causality as the appropriate one for sociology, with explanation via

Rational Choice Theory as an important route for interpretation.

To be satisfactory there needs to be evidence, typically arising from studies of different kinds, that such generating processes are not merely hypothesized.

3. Special Issues

3.1 Interaction Involving a Potentially Causal Variable

We now turn to the issue of interactions with a potentially causal variable. The graphical representations used in Sect. 2.3 to show the structure of various kinds of dependency and independency holding between a set of variables have the limitation, at least in the form used here, that they do not represent interaction, in particular that an effect of C may be systematically different for different groups of individuals.

This affects the role in analysis and interpretation especially of variables B that are themselves possibly explanatory to the variable C whose causal status is under consideration. So far we have been largely concerned with whether such variables could explain the effect on response of C. In more detailed discussion we should consider the possibility that the effect of C is substantially different at different levels of B. For example if B is an intrinsic feature such as gender, we consider whether the effect of C is different for men and for women. In particular, if the effects of C are in opposite directions for different levels of B we say there is a qualitative interaction, a possibility of special importance for interpretation.

Note especially, that even when C represents a randomized treatment which is automatically decoupled from preceding variables B, the possibility of serious interactions with B cannot in general be ignored; see Sect. 5.1.

Viewed slightly differently, absence of interaction is important not only in simplifying interpretation but also in enhancing generalizability and specificity. That is, an effect that has been shown to have no serious interaction with a range of potential variables is more likely to be reproduced in some new situation and more likely to have a stable subject-matter interpretation.

3.2 Unwanted Unobserved Intermediate Variable

Consider further the role of variables I referring to time points after the implementation of C. A subject-matter distinction can be drawn between on the one hand intermediate variables that are responses to C and that are explanatory to R and are part of some natural process, and on the other hand interventions into the system that may depend on C and which may be explanatory to R, but which in some sense are unwanted or inappropriate for interpretation. In the context of clinical trials, an example is the failure of patients to comply with a treatment assigned to them. Ignoring such noncompliance can lead to inappropriate intention-to-treat analysis.

Another example is in evaluations of study programs, whenever students in only one of the programs receive intensive encouragement during the evaluation period.

3.3 Aggregation

So far, little has been said about the choice of observational units for study. At a fundamental research level it may be wise to choose individuals showing the effects of interest in their simplest and most striking form. More generally, however, the choice has to be considered at two levels. There is the level at which ultimate interpretation and action is required and the level at which careful observation of largely decoupled individuals is available. For example, a criminologist comparing different sentencing or policing policies is interested in individual offenders but may be able to observe only different communities or policing areas. A nutritional epidemiologist comparing different diets is interested in individual people but may have to rely, in part at least, on consumption and mortality data from whole countries. The assumption that a dependence established on an aggregate scale, for example at a country level, has a similar interpretation at a small-scale level, for example for individual persons, involves the assumption that there are no confounders B at a person level that would account for the apparent dependency. This will typically be very hard or even impossible to check at all carefully from country-level data.

Errors rising as a result of over-aggregated units of analysis are called 'ecological fallacies' or in econometrics 'aggregation biases.' See *Ecological Inference*.

4. Bradford Hill's Conditions

The above discussion implicitly emphasizes that, while causal understanding is the aim of perhaps nearly all research work, a cautious approach is essential, especially, but by no means only, in observational studies. The most widely quoted conditions tending to make a causal interpretation more likely are those of Bradford Hill (1965), put forward in connection with the interpretation of epidemiological studies. Bradford Hill emphasized their tentative character.

For a critical discussion of these conditions, see Rothman and Greenland (1999) and for a slightly revised version of them Cox and Wermuth (1996, Sect. 8.7).

Koch gave conditions for inferring causality when the potential cause can be applied, withdrawn, and reapplied in a relatively controlled way and the pattern of response observed.

5. *Some Fallacies in More Detail*

The previous discussion has mentioned various points at which fallacious arguments are not only possible, but relatively common. This in no way covers the wealth of fallacious arguments possible in a statistical context, perhaps the most pervasive being the comparison of rates using inappropriate or ill-defined denominators. There is not space to discuss all these possibilities. This article is therefore concluded with three specific examples related to the main discussion of causality.

5.1 *Inappropriate Reference Set for Probabilities*

The first fallacy to be discussed, unlike the others, does not center around misuse of notions connected with causality, but rather with mathematically correct but inappropriate calculations of probability. Cornfield (1969) discussed a criminal case in California, People versus Collins, in which a couple had been found guilty partly on the basis of the following argument. Eye-witnesses described the crime as committed by a man with a moustache, by an Afro-American man with a beard, by an inter-racial couple in a car, by a couple in a partly yellow car, by a girl with blond hair, and by a girl with a ponytail.

Probabilities were assigned to these six features and, assuming independence, multiplied to give a probability $P = 0.8 \times 10^{-7}$, the smallness of this being argued as evidence of guilt. The California Court of Appeal rejected this argument as inappropriate, primarily because they regarded the issue as being whether there could be one or more further matching couples in the Los Angeles area. For this they argued in effect that if N is the population of the greater Los Angeles area then the number of couples with the assigned properties has a Poisson distribution of mean NP and, since one such couple is known to exist, the number has a zero-truncated Poisson distribution with parameter NP. From this the Court of Appeal calculated a probability of about 0.4 of there being one or more further couples with matching features, too high to justify a safe conviction of the original couple.

A more formal argument would use Bayes's theorem to calculate the posterior probability of guilt, assuming that the only evidence is that stated above and that the numerical assignments are reasonably accurate.

5.2 *Interactive Effect Involving an Unobserved Explanatory Variable*

It has been reported (Zivin and Choi 1991) that early controlled clinical trials with stroke patients were discouraging because they appeared to give differing answers to the question: do thrombolytic agents effectively improve the status of stroke patients? These are substances to dissolve blood clots, i.e., thrombi. Nowadays it is known that a stroke can be caused by a thrombus or by a burst vein. Thus, a thrombolytic agent may improve the patient's status or it may worsen it considerably, depending on the reason for the stroke. With patients of both types in any study 'treatment-unit' additivity will not hold, nor can there be 'strongly ignorable treatment allocation,' even if the study is a controlled clinical trial with random allocation of patients to one group treated with a thrombolytic agent and the other with a placebo.

Instead, observations like those in Table 1 are to be anticipated. The main response is success of treatment

Table 1
Counts, percentages and odds-ratios; the two explanatory variables are independent, $C \sqcup B$, due to randomized allocation of treatments, C, to patients; depending on the patient's unobserved status, B, the thrombolitic agent has a very different chance of treatment success

	$B = 1$, Burst vein		$B = 2$, Thrombus	
R, Success of treatment	C, Thrombolytic agent yes	no	C, Thrombolytic agent yes	no
yes	6 (2 percent)	60 (20 percent)	1425 (95 percent)	300 (20 percent)
no	294	240	75	1200
sum	300	300	1500	1500
odds-ratio	0.08		16	

Table 2
Counts, percentages and the same association between R and C as measured in terms of relative chances given each level combination of A and B; strong three-factor interaction between A, B, C; response depends on each of the explanatory variables A, B, C

	$A = 1$				$A = 1$			
	$B = 1$		$B = 2$		$B = 1$		$B = 2$	
	$C = 1$	$C = 2$	$C = 1$	$C = 2$	$C = 1$	$C = 2$	$C = 1$	$C = 2$
$R = 1$	604 (30 percent)	40 (20 percent)	300 (75 percent)	2000 (50 percent)	301 (75 percent)	2015 (50 percent)	600 (30 percent)	40 (20 percent)
$R = 2$	1396	160	100	2000	99	1985	1400	160
Sum	2000	200	400	4000	400	4000	2000	200
	Relative chances for $R = 1$ comparing $C = 1$ to $C = 2$ given A, B							
	1.5		1.5		1.5		1.5	

Table 3
Counts, percentages and relative chances obtained from Table 2 by summing over the levels of A show seemingly replicated associations for R and C given B; the actual associations between R and C, as shown in the previous table, appear reversed and $(R, C) \perp\!\!\!\perp B$

	$B = 1$		$B = 2$	
	$C = 1$	$C = 2$	$C = 1$	$C = 2$
$R = 1$	905 (38 percent)	2055 (49 percent)	900 (38 percent)	2040 (49 percent)
$R = 2$	1495	2145	1500	2160
Sum	2400	4200	2400	4200
	Relative chances for $R = 1$ comparing $C = 1$ to $C = 2$ given B			
	0.78		0.78	

and the main explanatory variable is the type of treatment. In this example there is a ratio of 1:5 for the number of stroke patients with a burst vein to those with a thrombus. With successful randomization the patient's status will be independent of treatment, but the strong interactive effect of status and treatment on outcome cannot be avoided. Whenever it is not feasible to observe the patient's status, then the reported results of any clinical trial will strongly depend on the actual percentage of patients in the study having a thrombus as cause of the stroke. In any case, nowadays it would be unethical to include stroke patients known to have a burst vein in a clinical trial designed to study a thrombolytic agent.

5.3 Dependence Reversal

The artificial data in Table 2 illustrate that a strong three-way interaction among explanatory variables A, B, C can lead to replicated dependence reversals whenever the response R depends on each of the explanatory variables.

Here, treatment $C = 1$ is consistently better under four different conditions, since the chances for successful treatment $R = 1$ are higher for $C = 1$ if compared with $C = 2$; relative chances are even identical in the four conditions. But, this treatment appears to be consistently worse under two replications $B = 1$ and $B = 2$, i.e., when A is unobserved. In addition, in an analysis of only the three variables R, B, C, it appears as if randomization had been used successfully since $C \perp\!\!\!\perp B$.

These effects are an extreme example of what is often called the Yule-Simpson paradox. Although in a sense the word paradox is inappropriate, the possibility of this kind of dependence reversal reinforces the need for either carefully studying relations among explanatory variables or for using effective randomization procedures.

If a statistical association is to be judged as evidence for a causal hypothesis, then one should be certain that the observed associations do not mislead us about the actual associations. This is impossible without assumptions about unobserved variables even in trials

with randomization (see also Stone 1993). Therefore it appears that substantial progress in establishing causes can be expected only via understanding and description of the processes which generate observed effects. See *Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox)*.

6. Suggested Further Reading

The statistical aspects of causality are best approached via the discussion paper of Holland (1986), where in particular, key references to the earlier literature will be found; see also Cox and Wermuth (1996, Sect. 8.7). For general issues about observational studies, see Cochran (1965) and Rosenbaum (1995). For a philosophical perspective, see Simon (1972) and Cartwright (1989). For an interventionist view, see Rubin (1974) and for a more formal analysis still from a social science viewpoint Sobel (1995). For a development based on directed acyclic graphs, see Pearl (2000) and for the general connections with graph theory Lauritzen (2000). For an approach based on a complete specification of all independencies between a set of variables, followed by a computer-generated listing of all directed acyclic graphs consistent with those independencies, see Spirtes et al. (1993). The use of counterfactuals is criticized by Dawid (2000). The reader should be aware that many rather different interpretations of causality are involved in these discussions.

An elementary account of fallacies is given, with many interesting examples, by Huff (1954). Good (1978) gives a detailed classification of types of fallacy, again with excellent examples. See also Agresti (1983).

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Causation: Physical, Mental, and Social; Explanation: Conceptions in the Social Sciences; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Agresti A 1983 Fallacies, statistical. In: Kotz S, Johnson N L (eds.) *Encyclopedia of Statistical Sciences*. Wiley, New York, Vol. 3, pp. 24–8

Bradford Hill A 1965 The environment and disease: association or causation. *Proceedings of the Royal Society of Medicine* **58**: 295–300

Cartwright N 1989 *Nature's Capacities and their Measurement*. Clarendon Press, Oxford, UK

Cochran W G 1965 The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society* A **128**: 234–265

Cornfield J 1969 The Bayesian outlook and its application (with discussion). *Biometrics* **25**: 617–57

Cox D R, Wermuth N 1996 *Multivariate Dependencies—Models, Analysis and Interpretation*. Chapman and Hall, London

Dawid A P 2000 Causality without counterfactuals (with discussion). *Journal of the American Statistical Association* **95**: 407–48

Goldthorpe J G 1998 *Causation, Statistics and Sociology*. Economic and Social Research Institute, Dublin, Republic of Ireland

Good I J 1978 Fallacies, statistical. In: Kruskal W H, Tanur J M (eds.) *Encyclopedia of Statistics*. Free Press, New York, Vol. 1, pp. 337–49

Holland P 1986 Statistics and causal inference (with discussion). *Journal of the American Statistical Society* **81**: 945–70

Huff D 1954 *How to Lie with Statistics*. Norton, New York

Lauritzen S L 2000 Causal inference from graphical models. In: Klüppelberg C, Barndorff-Nielsen O E, Cox D R (eds.) *Complex Stochastic Systems*. Chapman and Hall/CRC, London

Pearl J 2000 *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK

Rosenbaum P R 1995 *Observational Studies*. Springer, New York

Rothman K J, Greenland S (eds.) 1998 *Modern Epidemiology*, 2nd edn. Raven-Lippincott, Philadelphia, PA

Rubin D B 1974 Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Studies* **66**: 688–701

Simon H A 1972 Causation. In: Kruskal W H, Tanur J M (eds.) *Encyclopedia of Statistics*. Free Press, New York, Vol. 1, pp. 35–41

Sobel M E 1995 Causal inference in the social and behavioral sciences. In: Arminger G, Clogg C C, Sobel M E (eds.) *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Plenum, New York

Stone R 1993 The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society B* **55**: 455–66

Spirtes P, Glymour C, Scheines R 1993 *Causation, Prediction and Search*. Springer, New York

Zivin J A, Choi D W 1991 Neue Ansätze zur Schlaganfall-Therapie. *Spektrum der Wissenschaft* **Sept**: 58–66

D. R. Cox and N. Wermuth

Causation (Theories and Models): Conceptions in the Social Sciences

Many, perhaps most problems and hypotheses in social and behavioral research concern *causal relations*. What caused the fall of communism? What caused Peter's depression? What causes aggression in general? Atkinson and Birch (1978, p. 30) consider as one of the fundamental questions in the study of motivation: 'What causes the strength of tendencies to change?' However, the idea or concept of causation is also involved in many cases in which people do not use the word 'cause.' Many concepts include the idea of causation as part of their meaning, for example, a

makes b happen, a produces b, a leads to b, a has an influence on b, or b depends on a. The central importance of causation is demonstrated by the fact that social scientists continuously emphasize the difference between causal relations and mere correlations. Many research methods have been developed for drawing valid causal inferences from observations.

Causal statements are part of *explanations*. b can be explained by demonstrating that a occurred and that there is a causal law saying that whenever an event of type A takes place an event of type B necessarily follows. Although not all explanations are causal, causal explanations play a central role in science and in everyday life. Rational action is based on causal assumptions. This holds for everyday action as well as for scientific technology. If we press a switch, talk quietly to a baby, take medicine, or give up smoking, we have causal assumptions implying that what we are doing leads to a certain goal or prevents an event we don't want to happen.

1. *The Concept of Cause*

The concept of causation expresses a *relation* between *events*, meaning that one event is the *cause* and another the *effect*. 'a causes b' means that a makes b happen. We also speak of causation when one event prevents another from happening. a and b may be qualitative events, for example, Paul losing his job or Maria passing her exam. Causation also applies to quantitative variables, for example, the raise in price caused the drop in demand. In such a case the cause and the effect are changes of quantities. It is also possible to say that X taking a certain value is the cause for Y taking a certain value.

Causal statements can be *singular* statements or *general* ones. 'This event a caused this event b' is singular. 'A always causes B' is general. An example of a general hypothesis is 'Frustration always causes aggression' (which is not true). A general causal hypothesis which has been confirmed by empirical research is called a *causal law*. Causal statements may refer to observable as well as to unobservable events. Some causal statements connect theoretical with empirical concepts, thus functioning as *bridge principles* in testing or applying theories. For example, if an experimenter presumes that a certain treatment arouses the power motive of subjects, she accepts a causal assumption. Or if reaction time is used as an indicator of certain cognitive processes, it is presupposed that these processes have a causal influence on reaction time.

Since David Hume's famous treatise on causality, any discussion of this subject is much influenced by what he taught. Hume maintained that causal relations cannot be observed; they are inferred by the observer. We see that one billiard ball touches the other immediately before the second one begins to move. We perceive the two events and the temporal relation between them, but we do not perceive something like a causal power or a causal link between them (Hume 1739, p. 636). Therefore, Hume defined a cause to be an event a, followed by a contiguous event b, where A is always followed by B (1748, p. 76). This is Hume's '*regularity theory*,' which implies that there is no causal nexus in addition to regular succession.

The view that causation is reducible to regular succession has been criticized for centuries (cf. Armstrong 1983). It has been argued, for example, that, according to the regularity theory, day causes night. Hume's view was nevertheless taken over by modern empiricists. Bertrand Russell claimed that the concept of cause was a 'relic of a bygone age' and could be replaced by the concept of functional relationship. Empiricism in turn had a considerable influence on the social sciences. Finally, however, logical empiricism (or 'logical positivism'), was severely criticized (Popper 1959, Quine 1951). Its central claims had to be given up, especially the principle that a statement is meaningful if and only if it can be verified by observation. As a consequence of this criticism, the empiricist view of causation was put into question, too, and many philosophers came to the conclusion that causation is not reducible to regular succession (Bunge 1959). The idea of regular succession does not include the idea that a *makes* b *happen*. 'Causation' has perhaps to be taken as a fundamental concept not capable of being defined by other concepts like succession. There is, however, still no general agreement on this problem.

In contemporary science and philosophy, *spatial contiguity* (e.g., one billiard ball touches the other) is not considered as necessary for causation. In fact, spatial contiguity as a necessary condition was already outdated in Hume's days, since Newton had developed his theory of gravitation which involves distant action. However, *temporal contiguity* is widely accepted. In the social sciences, causal statements usually presuppose that a cause precedes its effect. In principle, the idea of simultaneous causation also seems to apply. Backward causation, however, as discussed in some fields of physics, does not yet play a role in social or behavioral theories.

The idea of causation is closely connected to *counterfactual conditional statements*. Hume has given a second definition of causation which connects causes with conditionals. According to this definition, the statement that a caused b means that '*if the first object had not been, the second never had existed*' (Hume 1748, p. 76). Today this is called *counterfactual dependence*: if a had not happened, b would not have happened either; if a had happened, b would have happened, too. Lewis (1986, essays 17 and 21) defines *causation* in terms of counterfactual dependence. It is, however, controversial whether the concept of causation or the concept of counterfactual dependence is more fundamental. Some philosophers maintain that

conditionals should be explained in terms of causation (Sanford 1989, Chaps. 11–14). In any case, the concepts of causation and conditionals are related to each other.

How closely are causes connected to their effects? Bunge defines causation in a way that assumes a very close connection: if A happens, then (and only then) B is always produced by it (Bunge 1959, p. 46). In this case, A is not only a sufficient condition of B, but also a necessary condition. Hume, as well as Newton, and later on, Russell, were convinced that an effect always has one and only one (type of) cause. Most contemporary social scientists instead hold that different causes can have the very same effect (a view already maintained by Machiavelli). For example, John may have been fired (b) because he stole the money (a). But he could have been fired anyway, so that stealing the money was *not necessary* for b. Being fired may have many different causes.

Interestingly, causes are often *not sufficient* either. John was only fired because his stealing was observed (c) and reported to the works management (d). a alone was not sufficient for b but a together with c and d were. This set of sufficient conditions was in turn not necessary for b since there could have been quite another set of sufficient conditions. Mackie (1974, p. 62) calls an event a of this kind an *inus condition*: a is an *insufficient* but *nonredundant* part of an *unnecessary* but *sufficient* condition of b.

In the social sciences, causal statements are often interpreted as *ceteris paribus* statements: other things being equal, A produces B. '*Ceteris paribus*' does not refer to every other object and state all over the world (they never all remain unchanged, not even for a millisecond), but to those (partly unknown) states that are causally relevant for B. Consider, for example, this psychological law, known as 'the Zeigarnik effect': unfinished tasks are more easily recalled than completed ones. Obviously, this can only be true if understood as a *ceteris paribus* statement. The fact that a task has not been finished has a causal influence on recall. But there are, of course, further factors that determine whether a task will be recalled or not. Therefore, the Zeigarnik effect can only be demonstrated under certain experimental conditions: subjects work on a number of puzzles. Some of them are interrupted, some are not, and everything else is kept unchanged (as far as possible).

Since it is not always clear which additional factors have to remain unchanged, causal hypotheses are *incomplete* in a certain sense (Gadenne 1984). They can be made more complete by specifying further conditions under which the connection between A and B is supposed to hold. However, it is usually not possible to formulate a very precise hypothesis saying that A produces B provided that certain well-defined boundary conditions C hold. These 'further conditions' seem incapable of being described exactly. For example, it is usually presupposed that, during the causal process, the system (e.g., a person, a group, an institution) exposed to A is in a 'normal state,' that is, it is not destroyed, damaged, or disturbed too much. It is assumed, for instance, that there won't be a natural catastrophe or a war, that the persons involved won't become ill or crazy, and so on. If such extraneous events do happen and prevent B from happening, this is not considered as evidence against the hypothesis. The causal statement refers to the 'normal', 'undisturbed' case.

If a is considered as a cause (maybe an *inus* condition) of b, this does not necessarily mean that a is the most immediate cause. A war may cause that some people die of hunger. This may be true, though there is a more immediate cause, namely that these people had no food just before they died, and even more immediate, physiological causes could be pointed out.

The previous remarks describe the general use of the concept of causation in the social and behavioral sciences. In addition, there are special concepts that differ from the general view. It has been claimed that in everyday language the concept of cause implies *manipulation* (Collingwood 1940, von Wright 1971). 'a is the cause of b' means that we can produce (or prevent) b by making a happen (preventing a). Now it is seems to be true that many causes are actions and that people learn the concept of causation in close connection with their own behavior and its effects. However, it is implausible that this concept, once acquired, implies manipulation as part of its meaning. It is rather vice versa. The idea of manipulation by action seems to involve the idea of causation. Furthermore, people are able to use and to understand propositions about events which cannot be manipulated at all, like the explosion of a star.

In empirical social research, scientists regularly test statistical hypotheses, say, about means or correlation coefficients. In which way are these statistical hypotheses related to those causal hypotheses scientists are interested in? Usually, it seems to be presupposed that a general causal hypothesis can be tested by testing a statistical hypothesis about a parameter like a mean (cf. Erdfelder and Bredenkamp 1994). For example, if for every individual an increase in X causes an increase in Y, then for every group of individuals (and, of course, for a large population), the mean of X must be greater than the mean of Y.

But why deal with group means or correlations when the subject of research are causal relations? The reason for this is that observation and measurement are inaccurate to some degree, so that many observations have to be made in order to determine whether there is an increase or decrease in Y. In addition, to control effects of repeated measurement, it is often necessary in social research to use several research groups of individuals differing only in X (best achieved by random assignment of individuals, see below). Now, even if it is true that for every individual an increase in X causes an increase in Y, it does not follow

from this that every individual of a group higher in X has a higher *measured score* in Y than every individual of a group lower in X. Therefore, one needs statistical analysis. Usually, if the statistical null hypothesis is rejected and the alternative hypothesis is accepted, this is considered as a confirmation of the causal hypothesis being tested. However, there is no total agreement about this procedure and there are some problems, concerning, for example, the logical relation between causal and statistical statements, or the connection between deciding on statistical hypotheses and the confirmation or falsification of the causal hypotheses.

It has been asked whether causal statements themselves can be interpreted as *probabilistic* statements. Suppes (1970) has presented a probabilistic theory of causality. Let the probability that B happens be higher if A occurs before: $p(B/A) > p(B)$. In this case, A seems to be something like a cause of B, it is a *prima facie* cause. A *prima facie* cause is not always a genuine cause. Suppose that there is an event A′ that precedes A, and let $p(B/A \text{ and } A') = p(B/A')$. Thus, A is actually quite irrelevant to B. It only appeared to be a cause, it is a *spurious* cause. If a *prima facie* cause is not spurious, it is a *genuine* cause. For a criticism of this theory see Salmon (1984, p. 193). For an advancement of the statistical approach to causation, especially with respect to the social and behavioral sciences, see Steyer (1985, Cheng 1997).

2. *Causal Hypotheses and Empirical Research*

Most social scientists hold (with Hume) that we cannot perceive something like a causal link between A and B. We only can perceive the events A and B themselves and their temporal relationship. The causal relation must be concluded from what we can observe. Under what conditions is it justified to conclude that certain results of observations confirm a causal statement? This is one of the major questions in the methodology of the social and behavioral sciences.

If A has always been followed by B, this can be explained by the causal hypothesis H: 'A always produces B.' It could be, however, that in every case in which A occurred, there was another event A′, unknown to the researcher, which actually caused B. Therefore, it is required that extraneous factors like A′ have to be ruled out by means of an appropriate empirical design. This is optimally done by comparing cases that differ only with respect to A (J. S. Mill's Method of Difference). For instance, individuals are randomly assigned to an experimental and a control condition. The two conditions are as similar as possible (time, rooms), except that the individuals of the experimental group are exposed to A, whereas the individuals of the control group are not. The research question is whether A causes B, or, as social scientists often say, whether there is a causal relation between the *independent variable* X (in this simple case, X has the values A and non-A) and the *dependent variable* Y (B, non-B). If B is obtained under the experimental condition and fails to be obtained under the control condition, it is justified to consider the causal hypothesis H as confirmed (H being interpreted as a *ceteris paribus* hypothesis). This is justified because other possible causes of B were ruled out by different kinds of *control*. Some of them were *eliminated* (e.g., noise), others were *kept constant* (e.g., time, rooms, experimenter), so that they cannot be the cause of a difference between the experimental and the control condition.

In *experimental* research, X is manipulated by the researcher, whereas in *passive observational methods* (*ex post facto* design) X varies naturally. If manipulation is possible and justifiable, it has an advantage over passive variation, since it is much easier in this case to prepare two or more very similar research conditions, differing only in X. The most important method of control in social experimentation (as opposed to physical experimentation) is *randomization*. By randomly assigning individuals to experimental and control conditions, the properties of these individuals are held constant between the conditions. To be more precise, the probability distributions of these variables are held constant. In order to do this it is not necessary to know these variables, that is, one does not need to identify and measure each variable that could have an influence on Y.

In passive observational methods, one can also try to discover whether a correlation between X and Y is based on a causal relation with (the values of) X being the cause of (the values of) Y. Suppose, for instance, that people who pray regularly are healthier. It may be that praying (X) has a positive causal influence on health (Y). It could also be, however, that X and Y are both causally influenced by a third variable Z, say, a variable connected with lifestyle, or something else. In this case, the correlation between X and Y should disappear if Z is held constant. Statistically, the influence of the variable Z can be determined by calculating the *partial correlation* between X and Y adjusted for the linear regression of each on Z. If this partial correlation is close to zero, one may conclude that X is not a cause of Y. However, if one has not found a variable Z that accounts for the correlation between X and Y, this is no guarantee that such a variable does not exist.

Based on this kind of reasoning, methods have been developed for testing causal assumptions using data from passive observation (Blalock 1964). They are known under such names as *causal models*, *path analysis*, and *structural equation models*. They are especially important in sociology, economics, and political science, where it is more difficult to perform experiments than in psychology.

In addition to experimental and passive observational methods there are *quasi-experimental methods*.

These are experimental designs that allow for some control of unknown extraneous variables, though they do not use random assignment (Campbell and Stanley 1963, Cook and Campbell 1979, see also Cook and Shadish 1994).

In passive observational methods only those variables can be controlled which have been deliberately selected and measured. This is, of course, better than no control at all. Nevertheless, it cannot be ruled out that a covariation between X and Y is really due to a further variable which has not been under control. On the other hand, it should be noted that in randomized experiments, too, one can never be completely sure that every extraneous variable is under control. This follows from the trivial fact that, taken exactly, no empirical variable can be held perfectly constant, since two individuals (or groups) always exist at different space–time points. More important is that by manipulating X the experimenter may manipulate something else that, unbeknownst, has a causal influence on Y. This kind of extraneous variable cannot be controlled by randomization, but only by theoretical reflection: could the effect expected (or already found) in this experiment be caused by something else that varies with X? Is there any alternative hypothesis which also explains the difference in Y? Are there any nuisance variables connected with X? Reflections of this kind are of the utmost importance in planning experiments as well as in interpreting their results. They are not rendered superfluous by using standard designs.

In the social sciences the methodology of empirical research as presented above is usually ascribed to J. S. Mill. Mill (1843) conceived four *Inductive Methods* for discovering and proving causal relations, the most important of them being the *Method of Difference* (see above). However, he only put certain procedures already known in the Middle Ages into the form of precise rules. Robert Grosseteste (c. 1168–1253) suggested that a good way to determine whether a particular herb has a purgative effect would be to examine numerous cases in which this herb is administered under conditions where no other purgative agents are present.

Mill overrated the importance of his inductive rules. He believed that every causal law yet known in science was discovered by one of these rules. Mill's contemporary William Whewell (1794–1866), one of the intellectual fathers of Karl Popper (1959), held the opposite view: hypotheses are discovered by creative insight, a process not reducible to specific inductive rules. Today most philosophers and historians of science maintain that there are many ways of inventing hypotheses and theories and that, say, Newton and Einstein did not invent their theories with the help of Mill's rules.

However, Mill was at least partially right, since he also held that his rules play a decisive role in *proving* causal statements. The hypothetico-deductive methodology as conceived by Popper has quite similar implications for planning and conducting empirical research as Mill's inductivism has. In order to test hypothesis H a test prediction P is derived deductively from H and some auxiliary assumptions. If non-P is observed, H is falsified—unless the auxiliary assumptions are put into question. If P is observed, this does not automatically count as a confirmation of H. The result P confirms H only if H has been *tested seriously*. A serious test requires that P cannot be explained by other hypotheses yet known in this field of research. A test is especially critical if, in addition to being serious, P contradicts other hypotheses that have been brought forward (which is known as an *experimentum crucis*). In this case, it is rather difficult for H to pass the test, and if it does, it deserves an increase in confirmation. Shortly put, this hypothetico-deductive methodology requires control of alternative explanations of B, and this in turn requires methods of control of other possible causes than A, including nuisance variables.

Since Campbell and Stanley (1963) it has become common practice to discuss the problem of causal interpretation in terms of *internal* and *external validity*. Internal validity is concerned with the singular causal statement 'a causes b.' The conclusion that a caused b is *internally valid* if the research design made it possible to control those extraneous variables that could have also produced b. However, the result that a caused b is compatible with there being further, nonredundant conditions which, in connection with a, produced b, e. g., properties of persons or the situation. Thus, the question arises whether this causal relationship also holds under different conditions. *External validity* is defined as the approximate validity with which conclusions are drawn about the generalizability of a causal relationship to and across populations of persons, settings, and times' (Cook and Campbell 1979, p. 39). External validity includes the question whether a causal relation found in laboratory research also holds under more natural conditions. Cook and Campbell (1979) expanded this view by adding *construct validity* and *statistical conclusion validity*. For a further development of these methodological criteria see Campbell (1986), Cook and Shadish (1994).

It is obvious that the problem of external validity is closely related to the *problem of induction*. Actually, the theory of internal and external validity is based on principles from Mill's *inductivistic* view as well as on elements from Popper's (1959) *hypothetico-deductive* methodology. Some supporters of a pure deductivistic view have instead proposed to dispense with any question of inductive generalizability and to formulate the problem in this way: which general hypotheses concerning the causal relation between A and B were tested in this research and which is the result? (cf. Gadenne 1984, Kruglanski and Kroy 1975). The result could be, for example, that the general hypothesis 'A and C together cause B' is confirmed, whereas the hypothesis 'A alone is sufficient for B' is falsified. The

result could also be that a hypothesis which was confirmed in earlier laboratory research is now falsified in a natural setting.

Like all hypotheses and theories, causal hypotheses can be empirically confirmed though not be proven as certain. They cannot be definitely falsified either. If the result of some repeated experiments is A and non-B, the hypothesis that A *ceteris paribus* produces B is invalidated to some degree and may be given up. But even in this case it is not completely sure that the hypothesis is false. It could have happened that in all these experiments there was an uncontrolled factor F sufficient to prevent B. Furthermore, a test prediction does not usually follow from H alone; it is derived from H together with further assumptions, including singular statements describing the experimental procedure, bridge principles connecting theoretical and empirical concepts, auxiliary assumptions from other theories or even other disciplines (for example, laws concerning measuring instruments), and the assumption that extraneous variables were eliminated or held constant. Therefore, unexpected empirical results never tell us directly which assumption is false (Duhem 1954). All scientists can do is to guess which assumptions are the mistaken ones, replace them by others, and make new predictions and tests. Testing hypotheses and theories is *holistic* in this sense. Hypotheses and theories, at least nontrivial ones, are not separately testable, which renders the interpretation of results more difficult. Nevertheless, most scientists and, probably, most philosophers of science believe that although there is no verification and no conclusive falsification there is progress in empirical research. After some systematic research, it turns out that certain causal hypotheses are much better confirmed than others, and the same holds, in the long run, for whole theories. It should be noted that contemporary Popperians (often misunderstood as supporting definite falsifications) agree with this view. They do not believe that theories can be falsified easily, and do not recommend that a theory has to be given up immediately if it is contradicted by an empirical result (cf. Albert 1999).

People have a strong tendency to 'see' and to infer causal relations in their environment. Cook and Campbell (1979) speculate that this is a product of the biological evolution: causal knowledge is helpful to adapt more effectively to external circumstances, especially knowledge about manipulable causes. Though causal hypotheses in everyday life (as well as in science) are usually simplifications, they often correspond sufficiently to real processes to have survival value.

There are limits to causal thinking. First, causal statements in the social sciences, even highly confirmed ones, are *incomplete* in a certain sense (see above): whether A really produces B depends on conditions which are not yet known completely and can probably never be described in a very precise way. The causal processes in the world are so complex that even the best causal laws are presumably simplifications which do not hold without exception.

Causal thinking is compatible with the fact that there are *interactions* and *feedback processes*. Wherever people interact, that is, in the whole psychological, social, and economic area, it should be taken into account that a variable Y which is causally dependent on a variable X may have an effect on X, too. Sometimes, this effect from Y on X is so small that it can be neglected. In this case, the simplification of taking only (the values of) X as the cause and (the values of) Y as the effect is justified. But in many other cases it may lead to serious mistakes if interactions are overlooked. Even in mechanics, which is usually considered as a shining example of causal thinking, many laws are actually interaction laws, for example, Newton's third law of motion and his law of gravitation (cf. Bunge 1959).

See also: Causal Counterfactuals in Social Science Research; Causal Inference and Statistical Fallacies; Causation: Physical, Mental, and Social; Control Variable in Research; Explanation: Conceptions in the Social Sciences; External Validity; Generalization: Conceptions in the Social Sciences; Hypothesis Testing in Statistics; Hypothesis Testing: Methodology and Limitations; Internal Validity; Laboratory Experiment: Methodology; Theory: Conceptions in the Social Sciences

Bibliography

Albert H 1999 *Between Social Science, Religion, and Politics. Essays in Critical Rationalism*. Rodopi, Amsterdam

Armstrong D M 1983 *What is a Law of Nature?* Cambridge University Press, Cambridge, UK

Atkinson J W, Birch D 1978 *An Introduction to Motivation*. Van Nostrand, New York

Blalock H M Jr (ed.) 1964 *Causal Inferences in Nonexperimental Research*. The University of North Carolina Press, Chapel Hill, NC

Bunge M 1959 *Causality: The Place of the Causal Principle in Modern Science*. Harvard University Press, Cambridge, MA

Campbell D T 1986 Relabeling internal and external validity for applied social scientists. In: Tronchim W M K (eds.) 1986 *Advances in Quasi-Experimental Design Analysis: New Directions for Program Evaluation*. Jossey-Bass, San Francisco, CA, Vol. 31, pp. 67–77

Campbell D T, Stanley J C 1963 Experimental and quasi-experimental designs for research on teaching. In: Gage N L (eds.) *Handbook of Research on Teaching*. Rand McNally, Chicago

Collingwood R G 1940 *An Essay on Metaphysics*. Oxford University Press, Oxford, UK

Cook T D, Shadish W R 1994 Social experiments: Some developments over the past fifteen years. *Annual Reviews of Psychology* **45**: 545–80

Cook T D, Campbell D T 1979 *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Rand McNally, Chicago

Duhem P 1954 *The Aim and Structure of Physical Theory*. Princeton University Press, Princeton, NJ (Originally published 1908)
Erdfelder E, Bredenkamp J 1994 Hypothesenprüfung. In: Herrmann T, Tack W (eds.) *Enzyklopädie der Psychologie. Methodologische Grundlagen der Psychologie*. Hogrefe, Göttingen, pp. 604–48
Gadenne V 1984 *Theorie und Erfahrung in der psychologischen Forschung*. Tübingen, Mohr
Hume D [1739] 1978 *A Treatise of Human Nature* [Book I Selby-Bigge L A (ed.), 2nd rev. edn. Nidditch P H (ed.)]. Clarendon Press, Oxford, UK (Originally published 1739)
Hume D [1748] 1975 *An Enquiry Concerning Human Understanding*. [Selby-Bigge L A (ed.), 3rd rev. edn. Nidditch P H (ed.)]. Clarendon Press, Oxford, UK (Originally published 1748)
Kruglanski A W, Kroy M 1975 Outcome validity in experimental research: A re-conceptualization. *Journal of Representative Research in Social Psychology* **7**: 168–78
Mackie J L 1974 *The Cement of the Universe*. Clarendon Press, Oxford, UK
Mill J S [1843] 1959 *A System of Logic, Ratiocinative and Inductive*, 8th edn. Longmans, London (Originally published 1843)
Popper K R 1959 *The Logic of Scientific Discovery*. Hutchinson, London
Quine W V O 1951 Two dogmas of empiricism. *The Philosophical Review* **60**: 20–43
Steyer R 1985 Causal regressive dependencies: An introduction. In: Nesselroade J R, von Eye A (eds.) *Individual Development and Social Change: Explanatory Analysis*. Academic Press, Orlando, FL, pp. 95–124
Suppes P 1970 *A Probabilistic Theory of Causality*. North-Holland, Amsterdam
von Wright G H 1971 *Explanation and Understanding*. Cornell University Press, Ithaca, NY

V. Gadenne

Causation: Physical, Mental, and Social

We invoke the notion of causation to describe many different aspects of our experience and of the world. For example, in describing the physical domain, we might say *the earthquake caused the building to collapse*; in describing the psychological or mental domain, we might say *paranoia was the cause of Oswald shooting Kennedy*; and in describing the social domain, we might say *excessive military spending caused instability in the Soviet Union*. In addition, we often use causal notions to link these various domains together. For example, we might say *her arm rose because she wanted to signal*, linking the mental domain (wanting to signal) and physical domain (her arm rising); or *social policy causes the degradation of the environment*, linking the social domain (social policy) and the physical domain (the environment). In short, causation is a very *general* notion: it applies in and across the physical, mental, and social domains. According to many philosophers, however, there is an interesting and difficult class of problems that emerges when we focus on how causal statements in one domain—for example, the physical domain—interact with causal statements in others—for example, the social or mental domains. Problems in this class—which we will call *problems of causation in multiple domains*—is what this article is about (for causation in general, see *Causes and Laws: Philosophical Aspects*).

1. Cartesian Dualism and Gassendi's Objection

The seventeenth century philosopher and mathematician Rene Descartes famously held that, while the nature and behavior of nonhuman animals could be explained completely by the physical science of his day, various features of human beings—including in particular the fact that human thought and action are not under direct environmental control—seemed to place them beyond the reach of physical science. Descartes therefore endorsed *dualism*, the idea that every human being is a complex of a physical body located in space and subject to physical laws and an immaterial mind *not* located in space and *not* subject to physical laws. Descartes' contemporary Gassendi famously objected that dualism made a mystery of how bodily actions and movements could be caused by immaterial psychological events: 'How can there be effort directed against anything, or motion set up in it, unless there is mutual contact between what moves and is moved? And how can there be contact without a body?' (Descartes and Cottingham et al. 1984).

Gassendi's objection is an example of a problem of causation in multiple domains. It turns on the fact that there is causal contact between the mental and physical domains, something which is difficult to understand if Cartesian dualism is true. However, problems of this sort are by no means unique to the relationship between the mental and the physical, nor to Descartes' specific account of that relationship, and nor do they arise only from older ideas in science and philosophy. On the contrary, they arise from a general picture of the world which is very common in contemporary intellectual culture: *the causal hierarchy picture*.

2. The Causal Hierarchy Picture

What sort of picture of the world does contemporary science present us with? According to one common view, science presents us with a twofold picture. One part of the picture is that the world is organized into a series of levels (cf. Oppenheim and Putnam 1958). The most basic level is the physical level, which is described

by the most basic science, physics. Arranged in a hierarchy on top of this level are other levels such as the chemical level, the biological level, the psychological level, and the social level, and each of these levels are described by nonbasic sciences (sometimes called *special* sciences), such as chemistry, biology, psychology, and sociology. To put things the other way around, the picture views social entities (institutions, groups, or corporations) as being composed of psychological entities (persons or individuals), which are themselves composed of biological entities (cells), which are themselves composed of chemical entities (molecules), which are themselves composed of fundamental physical entities (atoms and subatomic particles).

Another part of the picture adds causation to the hierarchy, in two distinct ways. First, it is part of the picture that, since causation is a very general notion, causal statements are true at every level. Thus, causal claims about the behavior of corporations or persons are, from the point of view of the picture, perfectly genuine and potentially true. Second, it is also part of the picture that every time a causal statement is true at some nonbasic level, there is a corresponding causal statement that is true at the basic level.

The intuitive idea behind this second point is that causal statements at higher levels require *mechanisms* at lower levels, and, ultimately, at the physical level. Thus, consider the statement, *she raised her arm because she wanted to signal*. According to the picture, this statement (if true) is made true by, and requires for its truth, a further statement drawn from a lower level, presumably a statement about neural activity and how that activity causes movement of the limbs. Or again, consider the statement, *urbanization in Australia in the twentieth century resulted in a decline of religious practice*. This statement (if true) is made true by, and requires for its truth, a further statement drawn from a lower level, presumably a statement about the thoughts and goals of particular people that made up the relevant populations. In both cases, causal statements at upper levels require mechanisms at lower levels.

Many aspects of this causal hierarchy picture have been discussed and examined in recent philosophy of (natural and social) science (see *Reduction, Varieties of*; *Social Properties (Facts and Entities): Philosophical Aspects*). But for our purposes the important point is that it raises a problem of causation in multiple domains, and in fact a whole class of such problems. Here is a simple way of seeing the issue. Consider some causal statement *S* which is putatively true at some nonbasic level *N*. Given the causal hierarchy picture, there will be another causal statement *S** which is true at the basic level *B*. However, if *S** is true, and we know it is true given our picture of science, why should we suppose *in addition* that *S* is true? Intuitively, after all, all the causal work—all the pushing and shoving—is done at the basic level. But then it is mysterious why *S*—or any nonbasic causal claim—should be true.

While the issues are sometimes assimilated, the problem generated by the causal hierarchy picture is different from the problem that Gassendi raised for Descartes, in two respects. First, Gassendi's objection flows from the fact that, for Descartes, the physical and the mental domains are so different in nature: one is in space, the other not. But the problem raised by the causal hierarchy picture does not have its source in the fact that the mental level is so different in nature from the physical level. Its source is rather the simple fact that upper-level causal claims are *distinct from* (i.e., not identical with) lower-level claims, together with the idea that every time a causal claim is true at some nonbasic level, a corresponding causal claim will be true at the basic level. Second, the problem generated by the causal hierarchy picture is not limited to the mental and physical domains, as was Gassendi's objection. As we have seen, it arises for any nonbasic causal claim, whether in biology, psychology, or social theory. So long as one supposes that there are causal claims at these levels, a problem of causation of multiple domains can be raised.

We can now see the significance of the problem of causation in multiple domains for contemporary thought. On the one hand, we have a picture of the world, the causal hierarchy picture, which is presented to us by science and therefore has a reasonable claim to be our best and most sophisticated picture of the world. On the other hand, the problem of causation in multiple domains apparently tells us that picture cannot be right, because the picture itself makes it hard to see how any nonbasic causal statements can be true.

3. *The Exclusion Argument*

We have so far been concerned with presenting the problem generated by the causal hierarchy picture in an informal way. However, it is desirable to be a bit more explicit about the reasoning which leads to the problem. The crucial fact about this reasoning is that it exploits a principle about causation often called *the exclusion principle* (Kim 1993, Yablo 1992). As we shall shortly see, the precise formulation of this principle is a matter of controversy, but a reasonable initial formulation is (E1):

(E1) If e_1 causes e_2, then there is no event e_3 such that: (a) e_3 is not identical to e_1; and (b) e_3 causes e_2

Intuitively, the idea here is that the discovery of a cause for some phenomenon usually means that other causes have been ruled out or excluded. For example, if a doctor tells you that the pain in your foot is caused by uric acid crystals in the joint of your big toe, you do

not regard it as an open question whether the pain has some *other* cause—calcium crystals, for example.

With the exclusion principle in place, we can now be more explicit about the problem generated by the causal hierarchy picture. It is easiest to state the argument if we focus on a particular kind of causal statement, one in which a certain nonbasic event n causes a physical event p. (Later we will consider whether this assumption is misleading.) Thus, suppose that (1) is true:

(1) n causes p.

Given the causal hierarchy picture, it would seem that there must be a physical event p* which *also* causes p; hence (2) is true:

(2) p* causes p.

Now, given that n and p* are from different domains, i.e., are drawn from different levels, we may further assume that:

(3) n is distinct from (i.e. not identical to) p*.

However, the principle of exclusion—in the form of (E1)—says that, in general, no two distinct events can cause a single event. Hence:

(4) abc

(4) If n is distinct from p*, then it is not the case that both n and p* causes p.

But (1)–(4) are jointly contradictory: while any three of them might be true, all four cannot be true together. It follows (barring some subtle ambiguity) that at least one of (1)–(4) must be false. In other words, what the reasoning tells us is that the causal hierarchy picture together with the exclusion principle presents a contradiction. To solve the problem one must give up or modify either the exclusion principle or the causal hierarchy picture or both.

4. *Possible Responses*

The most common response to the exclusion argument is to modify the principle of exclusion. Before turning to that response, however, it is important to briefly consider some alternative possibilities.

4.1 *Rejecting the Hierarchy Part of the Picture*

One obvious response is to suggest that the causal hierarchy picture is a distorted picture of the world because it takes the metaphor of levels too literally. This objection might go as follows. While it might be true that there are apparently lots of different sciences, this is only a methodological or pragmatic fact, which represents our limited epistemological access to the world. In fact, there is only a single science—physics—which genuinely explains the world. The other sciences are simply convenient descriptions which, apart from their pragmatic usefulness, will or could be dispensed with in the future.

The main problem with this position is a feature of the causal hierarchy picture that so far has not been discussed: *multiple realizability*. Earlier we saw that according to the causal hierarchy picture, entities from upper levels are composed of entities from lower levels. But it seems an obvious empirical fact that the *same* entities from an upper level might be composed of many *different* entities at lower levels. Thus, the same corporation over time might be composed of different individuals. The same mental processes in different people might be made up of different neural processes—indeed, given science fiction possibilities, they might not even be made of neural processes at all. The lesson usually drawn from multiple realizability (as these sort of facts are called) is that entities from upper levels are genuinely distinct from lower-level entities and, as a corollary, that the sciences dealing with upper-level entities are genuinely distinct from lower-level sciences (Fodor 1974); see Kim (1993) for an opposite view. However, if this *is* the right lesson, one should take seriously the hierarchy part of the causal hierarchy picture.

It is important to note that drawing this lesson from multiple realizability does not represent a return to Cartesian dualism. One can see this if one introduces another feature of the causal hierarchy picture: *supervenience*. Philosophers often say that, in the causal hierarchy picture, the upper levels supervene on the basic level, where this means that, while the upper levels in the hierarchy are distinct from the basic level, they are nevertheless wholly determined by, or are entailed by, the basic level (cf. Kim 1993). A simple way to think of supervenience is by switching metaphors from levels to patterns. To say that the psychological or the social supervene on the physical is to say that these are patterns in the physical—hence are wholly physical—it is simply that they are not patterns that physicists are interested in. To put the point a little more formally, supervenience tells us that if you imagine a possible world which completely matches the actual world in all physical respects, then you have *ipso facto* imagined a possible world that *also* matches the actual world in all psychological and social respects. But this idea is something that Descartes would have denied. According to Descartes, the psychological (and presumably the social also) is only contingently related to the physical, and thus a possible world that completely matches the actual world in physical respects is not necessarily a world that completely matches it in these other respects.

4.2 Rejecting the Causal Part of the Picture

If one cannot reject the hierarchy part of the causal hierarchy picture, can one reject the causal part? One way to develop this objection starts with the observation that it is not at all obvious that particular sciences make explicit use of causal notions. Bertrand Russell (1917) famously claimed that physics had no use for causation. More recently, many have argued convincingly that causation has no place, or at least no central place, in the explanations offered by particular special sciences such as psychology and linguistics (Chomsky 1959, Cummins 1983) and anthropology (Sperber 1996). So perhaps the proper account of the picture of the world that science presents is that it is hierarchical but not causal.

However, while it might very well be true that particular sciences do not provide explicitly causal explanations, this does not mean that various causal claims will not be *true* at the various levels. We can see this if we examine briefly the dispute been Chomsky and Skinner, surely one of the central moments of twentieth-century science (Chomsky 1959). Skinner had proposed, among other things, that the basic aim of scientific psychology and linguistics should be to provide causal explanations of human behavior—functional analysis, Skinner called it. Part of Chomsky's response to Skinner was that it is premature to suppose that scientific psychology could provide such explanations. He suggested instead that a more reasonable aim for scientific psychology would be the decomposition of certain psychological capacities and abilities, and in particular the ability on the part of a speaker to speak a language. I think it is fair to say that, in this regard, Chomsky's suggestion has been enormously influential and that most cognitive psychologists follow him in not aiming at causal explanations of behavior. However, even if this is true, it also remains true both that there *are* truths about how behavior is caused, and that these truths are in all probability multiply realized. But then we have the problem of causation of multiple domains back again.

One can imagine a more radical development of this point, where the claim at issue is not that causation plays no role in the explanations offered by special sciences, but rather that no nonbasic causal claims are true. In fact this idea has been historically popular in philosophy of mind and psychology. In part because of the influence of Wittgenstein, it was at one time very common to hear that causal talk in the psychological domain either should not be taken seriously or should be interpreted very differently from similar talk deployed in the physical domain (Ryle 1949; for criticism see Davidson 1963). But suggestions of this sort consistently underestimate the sense in which causation is an extremely abstract and general notion. W.V. Quine famously argued that the difference between the statements *cows exist* and *numbers exist* is not a difference in existence, but rather a difference between cows and numbers. Similarly, one might say that the difference between causal claims in the physical domain and e.g. the psychological domain is not a difference in causation, it is simply a difference between physics and psychology.

4.3 Rejecting the Metaphysical Presuppositions of the Picture

Finally, a number of philosophers have attempted to avoid the exclusion argument by suggesting that it only arises from mistaken metaphysical views lying behind the causal hierarchy picture. For example, in stating the exclusion argument above, we assumed that an upper-level event might cause a lower-level event—(1) summarizes just that possibility. According to some philosophers, however, this sort of 'downward causation' is deeply mysterious, and causation takes place only within levels, not across levels (cf. Kim 1993). However, if one rejects downward causation, one might reject the exclusion argument right from the start.

However, the problem with this suggestion is twofold. First, it is not clear that downward causation is so mysterious. As we have noted a number of times, causation is an extremely general notion. There seems no reason in principle why we should not suppose that upper-level events might not cause lower-level ones. Second, it seems possible to formulate the exclusion argument even if one does not start by assuming that downward causation is possible. For consider: given the causal hierarchy picture, every nonbasic event will supervene on some basic event. But then one could causally explain the presence of any nonbasic event by causally explaining the basic event on which it supervenes. This brings the problem back again: given the causal hierarchy picture, there seems no rationale for supposing that any nonbasic causal statement is true.

5. Problems with Exclusion

If we cannot answer the exclusion argument by rejecting the causal hierarchy picture or some metaphysical presupposition of the argument, it would seem that the only option left is to give up or modify the exclusion principle. Indeed, objections to this principle are easy enough to find. Consider the spy shot by the firing squad: the first soldier on the firing squad caused his death, but so did the second. Hence the spy's death is overdetermined. But a little reflection shows that the exclusion principle as stated so far rules out cases of overdetermination *a priori*.

This objection to the exclusion principle is suggestive but limited. It is true that there are possible cases of overdetermination. But it still seems unlikely that

every case of nonbasic causation should be analogous to the case of the firing squad. To accommodate the possibility of overdetermination we can reformulate the exclusion principle to read:

(E2) If e_1 causes e_2, then (probably, in general) there is no event e_3 such that: (a) e_3 not identical to e_1; and (b) e_3 causes e_2

(E2) allows, where (E1) does not, the possibility of firing squad cases, but it also says that we can expect such cases to be the exception rather than the rule. On the other hand, it seems clear that we could formulate a probabilistic version of the exclusion argument based on (E2) which would be almost as bad as the nonprobabilistic version based on (E1).

A second objection to the exclusion principle cuts deeper, however. Consider the possibility of composition (or multiple realizability) all the way down: whenever we arrive at a potential basic level, it turns out that this level is *itself* composed at a lower level. Combining this idea with the exclusion principle yields some absurd results. For if there is composition all the way down, it is easy to see that the exclusion argument will tell us that *any* candidate causal statement is false. But this suggests that there really is something wrong with the exclusion principle. Surely a principle of causation should not have the consequence that no candidate causal claim is true!

In response to this objection, a proponent of the exclusion principle might take one of two options. First, one might deny the possibility that there is composition all the way down. However, this is a fairly desperate maneuver. It is surely an empirical question, not a question which can be decided *a priori*, whether the levels of the world simply go on forever. Second, and much more plausibly, one might revise the exclusion principle to accommodate the objection. The obvious way to do this is to replace 'identical to' in (E2) with 'supervenient on', invoking the relation we discussed in the context of multiple realizability. That would result in the following principle:

(E3) If e_1 causes e_2, then (probably, in general) there is no event e_3 such that: (a) e_3 is not supervenient on e_1; and (b) e_3 causes e_2

This version of the exclusion principle does not have the result that, if the world is composed all the way down, none of our paradigm causal claims are true. For it is consistent with (E3) that causal claims can be true at levels which supervene on the basic level. Thus, even if we assume that there is composition at every level, it is still consistent with (E3) that some causal claims at nonbasic levels are true.

On the other hand, if it is (E3), and not (E1) or (E2), which provides the proper articulation of the exclusion principle, then we have a response to the exclusion argument. The exclusion argument mistakes (E1) for (E3). Combining the causal hierarchy picture and (E1) yields a contradiction, but there is no problem with combining the causal hierarchy picture and (E3), for (E3) does not exclude the possibility of nonbasic causal claims.

6. *Further Questions*

Problems of causation of multiple domains threaten to undermine our picture of the world. But, as we have seen, these problems arise only from a mistaken conception of the exclusion principle, i.e. only if we formulate that principle as (E1) or (E2) and not (E3). However, it would be a mistake to conclude that this is the end of the matter. For (E3) generates some further puzzling questions of its own. I will close by briefly mentioning two of these.

The first question arises when we consider the difference between (5) and (6):

(5) For some events e_1, e_2, and e_3 , if e_1 causes e_2, and if e_3 supervenes on e_1 then e_3 causes e_2

(6) For all events e_1, e_2, and e_3 , if e_1 causes e_2, and if e_3 supervenes on e_1 then e_3 causes e_2

(E3) allows the possibility that (5) is true. But it had better not allow the possibility that (6) is. For (6) is subject to counterexamples such as the following (Jackson and Pettit 1990). Suppose living in a particular neighborhood on the north side of town (e_1) makes you happy (e_2); living in a *particular* neighborhood on the north side of town entails living on the north side of town—as we might put it, living on the north side of town (e_3) supervenes on living on a particular neighborhood on the north side (e_1). Nevertheless, it might not be true that living on the north side (e_3) makes you happy (e_2)—after all, with the exception of your own particular neighborhood, you may dislike the north side, thinking of yourself as a 'south side person.' But now we are faced with a problem: if (6) is not true and (5) is, how are we to draw the line?

In order to answer this problem, Jackson and Pettit appeal to a principle of causation they call *invariance of effect under variance of realization*. They go on to develop this principle, embedding it in a framework for thinking about the relation between causation, causal explanation, and the causal hierarchy picture (especially as that applies to the social and psychological realms) which they call *program explanation* (cf. Jackson and Pettit 1990, 1992, Pettit 1993). Unfortunately, reviewing these ideas is beyond the scope of the present discussion. But the important point is that the idea that (E3) articulates the exclusion

principle incurs an obligation: to explain what distinguishes the event triples $\langle e_1, e_2, e_3 \rangle$ which satisfy (5) from the event triples which don't.

The second question returns us to Gassendi's objection to Descartes. In response to Gassendi's objection, many Cartesian dualists are tempted to endorse epiphenomenalism. Epiphenomenalists accept that there is a correlation between mental events and behavioral events, but insist that the correlation in question is not causal. Rather the correlation is the product of two other relations: a nomological (or lawful) relation between brain events and psychological events, and a causal relation between brain events and behavioral events. Since these latter relations can obtain without a causal relation obtaining between mental events and behavioral events, epiphenomenalism is a possible position, no matter how counterintuitive it seems to suppose that mental events don't cause behavioral events.

The problem that epiphenomenalism presents for (E3) and our discussion of it is the following. If we restrict attention to the relation between the psychological level and the neurological level, we too have articulated a correlation between mental and behavioral events, and for us too the correlation is a product of two further relations: first, there is a supervenience relation between the mental event and the brain event; second there is a causal relation between the brain event and the behavioral event. However, why should we suppose that this correlation is causal? As the case of epiphenomenalism makes clear, the mere fact that a correlation is a product of a causal relation and some other relation does not make the correlation causal. Perhaps then, the answer we have given to the exclusion argument is no better than the epiphenomenalist answer to Gassendi?

Part of the answer to this question is that there are correlations and there are correlations. True, the fact that a correlation is the product of a causal relation and some other relation does not make it causal. But it had better not make it noncausal either—after all, *every* causal relation is a product of a causal relation and identity. So the structural similarity between the account we have offered and epiphenomenalism does not entail that account is no better than epiphenomenalism.

But there is also a somewhat deeper issue here. The world presents us with a myriad of correlations. Our concept of causation is in part a tool to divide these correlations into the causal and the noncausal. Obviously, there are some correlations which are clearly causal and some which are not. But it needs to be admitted that there are hard cases, cases in which it is difficult to say whether we have a causal correlation here or not. The causal hierarchy picture that we have been discussing presents us with such hard cases. Adopting a concept of causation that employs (E1), these cases will turn out not to be causal; adopting a concept of causation that employs (E3), they will. As we have seen, the latter concept *does* seem a reasonable concept of causation. But defending that choice in detail requires more investigation of the concept of causation that we can enter into here.

See also: Causal Inference and Statistical Fallacies; Causation (Theories and Models): Conceptions in the Social Sciences; Causes and Laws: Philosophical Aspects; Cognitive Science: Overview; Individualism versus Collectivism: Philosophical Aspects; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective

Bibliography

Chomsky N 1959 Verbal behavior Skinner B F. *Language* **35**(1): 26–58

Cummins R 1983 *The Nature of Psychological Explanation*. MIT Press, Cambridge, MA

Davidson D 1963/1981 Actions, reasons and causes. In: Davidson D (ed.) *Essays on Actions and Events*. Oxford University Press, Oxford, UK

Descartes R Cottingham J et al. (trans.) 1984 *The Philosophical Writings of Descartes* Cambridge University Press Cambridge, UK, Vol. 2

Fodor J A 1974/1981 Special sciences: or, the disunity of science as a working hypothesis. In: Fodor J A (ed.) *Representations* 1st edn. MIT Press, Cambridge, MA

Jackson F, Pettit P 1990 Causation and the philosophy of mind. *Philosophy and Phenomenological Research* **50**: 195–214

Jackson F, Pettit P 1992 In defense of explanatory ecumenism. *Economics and Philosophy* **8**: 1–21

Kim J 1993 *Mind and Supervenience*. Cambridge University Press, Cambridge, UK

Oppenheim P Putnam H 1958 The unity of science as a working hypothesis *Minnesota Studies in Philosophy of Science* **2**: 3–36

Pettit P 1993 *The Common Mind*. Oxford, University Press, New York

Russell B 1917/1963 On the notion of cause. In: Russell B (ed.) *Mysticism and Logic*. Penguin, New York

Ryle G 1949 *The Concept of Mind*. Hutchinson's University Library london

Sperber D 1996 *Explaining Culture*. Blackwell, Oxford, UK

Yablo S 1992 Mental causation. *Philosophical Review* **101**: 245–80

D. Stoljar

Causes and Laws: Philosophical Aspects

The questions 'what makes it the case that one event causes another?' and 'what makes it the case that something is a law of nature?' are highly controversial ones for which, amongst contemporary philosophers,

there is no answer that can claim orthodoxy. This article sketches some of the most influential theories of causation and lawhood.

1. Preliminaries

'Cause' and 'law' are perhaps two of the most important and fundamental concepts that human beings deploy in their attempts to understand and intervene in their environment, both in everyday life and in their scientific endeavors. When we want to explain why an event occurred, we seek out its causes. When we act, we do so because we believe the action will have certain effects. When we want to know why our environment and our fellow human beings behave in regular, predictable ways, we look for the laws that govern that behavior.

It is part of the scientist's job to discover what causes what, and to investigate what the laws of nature are. The philosopher's job, however, is a more abstract one: the philosopher wants to know what makes it the case that something—anything, be it the decay of a subatomic particle or the assassination of Archduke Ferdinand—causes something else, or that something is a law of nature—whether it is a physical, social, psychological, or any other kind of law. We know that when one strikes a match and it lights, the first event—the striking—caused the second. But this causal fact depends for its truth on more than the mere fact that the two events occurred one after the other: there must be some extra feature of the world which binds these two events together as cause to effect. Similarly, it seems that laws of nature must be more than mere regularities: it is doubtless true that all lumps of gold are smaller than a mile in diameter, but it is not a law that this is so. So genuine laws must have some extra feature, apart from regularity, which marks them off from mere 'accidental' regularities. The philosopher's job, or at least part of it, is to say what, if anything, this 'extra feature' might be.

1.1 The Basic Locutions

There are three kinds of locution which need to be distinguished from one another. First, we have statements of law, which can be said to have the form 'it is a law that all *F*s are *G*s'. *F* and *G* are properties or kinds of event. For instance, if it is a law that all heated metals expand, the relevant properties are being a heated metal (a property that all and only metals that are being heated instantiate) and expanding (a property that all and only expanding things instantiate). Particular heated (and therefore expanding) bits of metal are in turn said to instantiate or are instances of the law that all heated metals expand.

Laws, then, are general facts about how objects or events of a particular kind will, as a matter of law, behave in a given kind of situation. What about causal facts? Two different kinds of causal claim, one general and one particular, need to be distinguished. On the one hand, there are general (sometimes called 'generic,' 'population-level,' or 'type-level') causal claims, which have the form '*F* causes *G*' or '*F* is a cause of *G*' ('smoking causes cancer'; 'poverty is a cause of crime'). Again, *F* and *G* are properties or kinds of event or object. On the other hand there are particular causal claims of the form '*c* caused *e*' (or '*c* is a cause of *e*'): 'the striking of the match caused the fire,' 'the assassination of Archduke Ferdinand was a cause of the First World War,' and so on. *c* and *e* here are particular events (or facts): a particular assassination and war, for example, rather than assassinations and wars in general.

1.2 Determinism vs. Indeterminism

A question that is relevant to any discussion of the nature of causes and laws is that of whether the universe is fundamentally deterministic or indeterministic. One way of putting the thesis of determinism is this: the complete state of the universe at any given time together with the laws of nature determines precisely what the complete state of the universe will be at all future times. Indeterminism is simply the denial of determinism.

The view that the universe is fundamentally indeterministic has been widely (though by no means universally) held within science since the advent of quantum mechanics; however, philosophers have only relatively recently begun to take the possibility of indeterminism seriously. One consequence of this reluctance to take indeterminism on board is that most current theories of causes and laws (with the exception of theories of general causation) originally were formulated under the assumption of determinism, and were then modified to take into account the possibility that at least some laws and causal processes are fundamentally indeterministic. The theories of laws and particular causation discussed below are presented in their simpler, deterministic form. While the indeterministic versions of the theories present additional complications and problems—for example, because they employ the notion of probability—the fundamental philosophical issues addressed by the indeterministic and deterministic versions are mostly the same (see *Probability and Chance: Philosophical Aspects*).

2. Laws of Nature

What is the difference between laws of nature and merely 'accidental' regularities? (Suppose that it is a law that all metals expand when heated and that it is merely accidentally true that all (past, present, and

future) Queens of England are less than two meters tall.) A crucial difference is that laws 'support counterfactuals,' whereas accidental regularities do not. It is true of any unheated piece of metal that if it were to be heated it would expand. But it need not be true of any non-Queen of England that if she were to be Queen she would be less than 2 meters tall. If someone is currently, say, 2.05 meters tall, becoming Queen of England would not result in her losing more than 5 cm height.

Another difference is that violations of law are said to be physically impossible, whereas violations of accidental regularities are physically possible. It is physically impossible for a heated metal to fail to expand when heated; but it is perfectly possible for a Queen of England to be taller than 2 meters.

A third difference is that we tend to suppose that laws (unlike accidental regularities) 'govern' what goes on in the universe. We tend to suppose that laws of nature are rather like pieces of divine legislation: decrees that the universe must obey certain rules rather than mere general descriptions of what in fact happens.

Can an analysis of lawhood be given that does justice to these intuitions about the nature of laws? A popular view (see Armstrong 1983) is that laws are relations of necessitation between universals. To say that it is a law that all *F*s are *G*s is to say that there is a necessary relation, *N*, that holds between the universals (properties) *F* and *G*. (Following Armstrong, we can write this '$N(F, G)$.') When an object instantiates *F*, the instantiation of *F* guarantees, via *N*, that *G* will also be instantiated. Thus, $N(F, G)$—its being a law that all *F*s are *G*s—guarantees that all *F*s will in fact be *G*s.

The basic point of this 'realist' theory of laws is that it provides an 'ontological ground' for the difference between laws and accidents: accidental regularities just happen, whereas lawful regularities are grounded in and explained by a necessary relation holding between the co-occurring properties. And, on this view, sense can be made of the idea that laws govern what happens: if $N(F, G)$ holds, then *G* must be instantiated whenever *F* is instantiated.

The main rival to this conception of laws is the 'Humean' or 'regularity' conception, according to which laws are really no more than regularities. The regularity view is motivated primarily by an 'empiricist' epistemology according to which we ought not to believe in entities that we cannot perceive. Since our evidence that there are any laws of nature comes solely from the observation of regularities—when we see a law (as opposed to an accidental regularity) being instantiated we do not see any extra ontological feature—we ought not to believe that there *is* any extra ontological feature which laws have but accidental regularities lack.

A 'naïve' regularity theory, however—according to which statements of law are merely true universal generalizations—is untenable, for it simply collapses the distinction between laws and accidental regularities. If statements of law are merely true universal generalizations, then accidental regularities are themselves laws: it turns out to be a law of nature that no Queen of England is taller than 2 meters after all.

A more sophisticated version of the regularity theory is provided by the 'Ramsey–Lewis view' (after F. P. Ramsey and David Lewis (see Lewis 1973b, pp. 73–5: J. S. Mill also held a similar view.) The basic idea of the Ramsey–Lewis view is as follows. When we seek, as scientists do, to find out how the universe behaves, we are not satisfied with isolated facts about what happened in the laboratory on Tuesday afternoon, what happened when a particular patient took a particular drug, and so on. Rather, what we seek are wide-ranging generalizations that abstract as far as possible from the particularities of the laboratory or the patient. Generalizations about the heights of Queens of England or the diameters of lumps of gold apply to an extremely restricted number and range of phenomena. On the other hand, generalizations about subatomic particles, chemical reactions, the relationship between force, mass and acceleration, and so on, apply to an enormous number of diverse phenomena, and are hence much more useful and explanatorily powerful. According to the Ramsey–Lewis view, the difference between laws of nature and merely 'accidental' regularities amounts to roughly this difference between generalizations that are wide-ranging and powerful and those that are not. It is not that there is any ontological distinction between laws and accidents—laws do not have an extra feature (*N*, say) that accidental regularities lack; rather, it just so happens that some generalizations (the laws) are interesting and useful ways of codifying what happens in the universe, and others (the accidents) are not.

The Ramsey–Lewis view (and the Humean view of lawhood in general) diverges from the everyday, intuitive conception of laws of nature in that it does not accord laws the governing role that we ordinarily suppose them to have. At bottom, nothing makes metals expand when heated any more than anything makes any Queen of England be shorter than 2 meters. Some philosophers claim that this aspect of Humeanism takes the sting out of the alleged conflict between determinism and free will (see Swartz 1985, chaps. 10 and 11, Berofsky 1987) (see also *Free Will and Action*).

3. *General Causation*

3.1 *Probabilistic Theories*

Frequently it is claimed on cigarette packets that smoking causes heart disease; but what makes (or would make) such a claim true? The most popular

kind of theory of general causation takes such claims to be made true by probabilities: what makes it true that smoking causes heart disease is that smoking raises the probability of heart disease.

What does it mean to say that smoking (C) raises the probability of heart disease (E)? One can think of the matter in roughly the following way. Fix on a particular population: Australian citizens, say, or the residents of Greater London. Now divide that population into the smokers (those members of the population who instantiate C) and the nonsmokers (those who don't instantiate C). Find out what the relative frequency of heart disease is in each group, that is, the proportion of smokers who get heart disease (call this proportion x) and the proportion of nonsmokers who get heart disease (y). If $x > y$, then smoking raises the probability of heart disease in the population under investigation: $\Pr(E/C) > \Pr(E/\sim C)$ (the probability of E given C is greater than the probability of E given the lack of C).

However, while the fact that C raises the probability of E gives us *prima facie* evidence for thinking that C causes E, it does not entail that C causes E. Falling barometer readings (C) raise the probability of rain (E), that is, $\Pr(E/C) > \Pr(E/\sim C)$, even though C does not cause E. Rather, C and E are both effects of a common cause: low atmospheric pressure (F). The correlation between C and E is spurious. We, therefore, need to refine the basic intuition that causes raise the probabilities of their effects, in order to rule out cases of spurious correlation.

One way of doing this is to 'hold fixed' other relevant factors (in this case, F) when assessing C's probabilistic impact on E: rather than looking just at the correlation between C and E, we consider the correlation between C and E in the presence of F and, separately, in the absence of F. What we find in the barometer case is that $\mathrm{P}(E/C \,\&\, F) = (\mathrm{P}(E/\sim C \,\&\, F)$ and $\mathrm{P}(E/C \,\&\, \sim F) = \mathrm{P}(E/\sim C \,\&\, \sim F)$: the probabilistic correlation between C and E disappears when we hold fixed the relevant factor, and this reflects the fact that C does not cause E: falling barometer readings do not cause rain.

According to one kind of analysis (see, for instance, Suppes 1970, Skyrms 1980, Eells 1991), general causal facts are to be analyzed in terms of this kind of conditional probability. Of course, there is usually more than one factor or property (apart from the factor C whose causal influence we are trying to establish) that is relevant to an effect E, so we need to keep all such factors fixed if the probabilistic correlations are to reflect causal correlations. Call a subset of the population whose members are all the same with respect to which relevant factors (apart from C) they possess a 'background context'. For example, if there are two factors apart from C that are relevant to E (call them X & Y), there will be four background contexts: $X \,\&\, Y$, $X \,\&\, \sim Y$, $\sim X \,\&\, Y$, and $\sim X \,\&\, \sim Y$. Each member of the population under investigation will be a member of one and only one background context. Now we can assess the causal status of C with respect to E in each background context. According to Eells's analysis, C causes (or is a positive causal factor for) E in a particular population if and only if C raises the probability of E in every background context of that population, that is, if and only if $\Pr(E/C \text{ and } B_i) > \Pr(E/\sim C \text{ and } B_i)$ for each background context B_i. Similarly C is a negative causal factor for E if C lowers the probability of E in every background context, and E is causally neutral for E if C makes no difference to the probability of E in each background context. According to Skyrms's slightly weaker (1980) analysis, C causes E if and only if C raises the probability of E in at least one background context and C does not lower the probability of E in any background context.

Dupré (1984, 1990) has argued that analyses like Eells's and Skyrms's are too strong: they make it too hard for general causal claims to be true. Suppose that some tiny minority of the US population has some peculiar physiological condition P that makes smoking (C) a prophylactic against heart disease (E)—even though for everyone else in the population, smoking increases the risk of heart disease. It follows from Eells's analysis—and also from Skyrms's—that smoking does not cause heart disease in the US population, since $\Pr(E/C \,\&\, P) < \Pr(E/\sim C \,\&\, P)$. This is a result that Dupré regards as highly counterintuitive. Dupré's rival analysis takes as its starting point a method that is actually used to test causal hypotheses in the social and medical sciences: that of the controlled experiment. If one wants to know whether C causes E in a given population, one way of trying to find out is to take a random sample of the population, induce C in a random subset of that sample, and compare the results. The point of having a random sample is to try to insure that other factors that are relevant to E occur with the same relative frequency as they do in the population as a whole, and thus that the probabilistic correlation between C and E is not spurious. Dupré calls a sample that achieves this match of frequencies a fair sample. His claim is that C causes E if and only if C raises the probability of E in a fair sample of the population. This analysis yields the desired result that smoking causes heart disease in the above example, since in any fair sample those with the physiological condition P will be vastly outnumbered by those who lack P; so C will still raise the probability of E in a fair sample.

An alleged benefit of Dupré's analysis is that it makes the metaphysics of general causation—what makes general causal claims true—match up in an obvious way with actual scientific methodology, and therefore explains why the methodology of the controlled experiment can, and often does, provide a way of uncovering causal facts.

A problem with probabilistic theories of general causation, however, is that they appear to fail given

the assumption that at least some features of the world are determined by prior circumstances (see Carroll 1992). For example, suppose that C causes E, but that C itself is determined by the comination of factors XYZ. Then it is not true that $\Pr(E/C \ \& \ XYZ) > \Pr(E/\sim C \ \& \ XYZ)$, since $\Pr(E/\sim C \ \& \ XYZ)$ is undefined; hence, according to Eells's and Skyrms's analyses—and contrary to hypothesis—it is not true that C causes E.

Dupré's analysis also falls prey to this objection. Suppose that 10 percent of the population have XYZ. Then in a fair sample of Cs, 10 percent must have XYZ; similarly for a fair sample of $\sim C$s. But there can be no such fair sample of $\sim C$s, since by hypothesis there will be nobody at all who has XYZ but lacks C. So according to Dupré's analysis it is not true that C causes E.

3.2 A Deflationary View

Given the apparent failure of standard probabilistic theories of general causation, perhaps a radical change of approach is needed. The theories discussed so far seek to analyze general causation solely in terms of conditional probabilities: no mention is made of particular causation. But this may seem rather odd, for it is natural to think that general and particular causal facts are related in some way: it is natural to think that the fact that smoking causes heart disease has something to do with the fact that lots of individual smokers are caused by their smoking to get heart disease. This line of thought leads naturally to the view that general causal claims are not claims about a distinctive kind of causation—general causation—at all, but are rather merely generalizations about particular causal claims.

The generalizations in question cannot be universal generalizations: when we say that smoking causes heart disease we do not mean to imply that all smokers get heart disease as a result of smoking. But perhaps they are rather weaker generalizations, akin to 'dogs bark' or 'American men like baseball' (see Carroll 1991). According to this proposal, general causal claims do not have precisely defined truth conditions: the question of precisely what proportion of smokers need to get heart disease because they smoke in order to make it true that smoking causes heart disease no more admits of an answer than does the question of precisely what proportion of dogs need to bark in order to make it true that dogs bark.

A central issue, then, is whether there is a special, general, kind of causation that serves to make general causal claims true—a kind of causation that is distinct and relatively autonomous from particular causation—or whether general causal claims are made true solely by particular causal claims.

4. Particular Causation

4.1 Hume's View

The agenda for much of the contemporary discussion about the nature of particular causation was set by Hume (1777, Sects. IV–VII). Hume was an empiricist, believing that there could be no 'ideas' in the mind that do not somehow come from our senses. How, then, do we come to have an idea of causation: what features of the world could furnish us with that idea? Famously, Hume maintained that we do not perceive any intrinsic connection or relation between two events we judge to be causally related: we see the match being struck, and we see it light, but we do not see any causation between those two events. Causation, then, cannot be a mysterious intrinsic relation between events—for if it were, since we cannot perceive any such relation, we could have no idea of causation. (However, see Strawson 1989 for the claim that this standard interpretation of Hume is mistaken.) Hume's positive claim was that two events are related causally in virtue of contiguity (causes and their immediate effects are right next to each other in space and time), temporal priority (causes precede their effects), and constant conjunction (events similar to the cause are always followed by events similar to the effect). So, according to Hume, what makes it true that the striking caused the lighting is that the two events are contiguous (or, if there is some spatio-temporal gap between them, the events are mediated by further events so that there is a chain of contiguous events starting with the striking and ending with the lighting); the striking occurs before the lighting; and similar strikings are always followed by similar lightings. The constant conjunction requirement insures that for Hume (and for most modern day Humeans) causation is an *extrinsic* relation: a causal relation obtains between two events not purely in virtue of how those events themselves are, but in virtue of features of other events: the striking only gets to count as a cause of the lighting because other strikings are also followed by lightings.

The other notable feature of Hume's view is that it is a reductionist view: causal claims, like 'the striking of the match caused the lighting,' are made true, at bottom, not by any primitive causal feature of the world, but by noncausal features (since the fact that two events are contiguous, the fact that one occurs before the other, and the fact that they are similar to other events are not themselves causal facts).

Nowadays Hume's own analysis is regarded as untenable, for a variety of reasons. For example, the contiguity requirement rules action at a distance (one event's causing another, spatially or temporally distant event without there being a chain of events 'hooking up' the first to the last) impossible, while many philosophers think that action at a distance is at least conceptually possible. The constant conjunction re-

quirement is obviously flawed: there can be constant conjunction without causation—for example, joint effects of a common cause are constantly conjoined but neither causes the other—and, if determinism is false (which most philosophers believe to be evident from quantum mechanics), it seems there can be causation without constant conjunction. Bombarding an atom might cause it to decay 2 seconds later even though exactly similar bombardments will not always be followed two seconds later by decay.

Nevertheless, the general issues that Hume raised are still among the main foci of dispute in contemporary philosophy of causation. In particular, the question of whether causal facts reduce to noncausal facts is the driving force behind many of the theories of causation currently on offer.

4.2 *Counterfactual Analyses*

Arguably the most influential analysis of causation in recent years has been Lewis's counterfactual analysis (Lewis 1973a). According to Lewis's analysis, an event e causally depends on event c if e counterfactually depends on c—which is to say, if, had c not occurred, e would not have occurred either (see *Counterfactual Reasoning, Quantitative: Philosophical Aspects*; *Counterfactual Reasoning, Qualitative: Philosophical Aspects*). A 'chain of causal dependence' is a series of events$\langle a, b, c, \ldots, n\rangle$ such that b causally depends on a, c causally depends on b, and so on. Finally c causes e if there is a chain of causal dependence (perhaps involving only c and e themselves, but perhaps involving many hundreds of intermediate events) from c to e.

The need for chains of dependence arises when there is no counterfactual dependence between cause and effect because of a backup mechanism or 'preempted alternative.' Suppose that the bus stop is next to the taxi rank. Susan takes the bus to work (c); but if the bus had been late she would have taken a taxi. She attends a 9.30 meeting (e). e does not counterfactually depend on c, since if Susan had missed the bus she would have taken a taxi and attended the meeting just the same. But there is some intermediate event d—Susan's getting off the bus at the stop outside her office, say—which counterfactually depends on c (if she had not got on the bus in the first place, she would not have got off it either) and upon which e counterfactually depends (if she had not got off at that stop—given that she was already on the bus, and it was already nearly 9.30—she would have had to walk back to her office from the next stop, and she would have missed the meeting).

Lewis's theory has been found to be subject to a number of counter-examples; and much recent work on causation has been devoted to devising alternative analyses which, while remaining true to Lewis's basic claim that causation is to be analysed in terms of counterfactuals, add further complexities in an attempt to circumvent the problem cases (see, e.g., Menzies 1989). Lewis's theory and its successors fall squarely within the Humean tradition, since the counterfactual analysis is a reductionist project: it seeks to show how causal facts are made true by noncausal features of the world.

Not everyone, however, is a Humean. Many philosophers deny Hume's central motivating claim that causation cannot be perceived (see Anscombe 1971, Armstrong 1997, Chap. 14, Menzies 1998); and many philosophers believe for other reasons that the reductionist project is doomed. Some argue that it is subject to fatal counter-examples (see, e.g., Carroll 1994). Some argue that the features of the world to which reductionist analyses seek to reduce causal facts are themselves really causal features; hence reductionist analyses are circular and therefore unsuccessful (see Carroll 1994). Some hold that our concept or causation is so heterogeneous that no unified analysis is possible; others that if a broadly Humean analysis of causation were true it would render the pervasive regularities which the universe exhibits utterly inexplicable: it would make the apparent orderliness of nature a sort of cosmic fluke (see Strawson 1989, Chap. 5).

5. *Causes and Laws*

The above discussion of particular causation makes no explicit mention of any relationship between causation and laws; yet intuitively there is some relationship between, for instance, the law that all metals expand when heated and the fact that a particular piece of metal's expansion is caused by its being heated. In fact, however, laws of nature typically do have a role to play in theories of causation. Hume's constant conjunction requirement, for example, can be seen as a requirement that causes and effects are lawfully correlated (or 'fall under' a covering law). Laws of nature enter into Lewis's counterfactual analysis of causation indirectly, via his analysis of counterfactuals (see Lewis 1973b): for Lewis, the truth of 'if A had been the case, B would have been the case' requires that B is true at the 'closest possible world(s)' in which A is true; and closeness of possible worlds depends in part on the extent to which they share the same laws of nature.

In any case, the relationship between causes and laws is certainly more complex than the simple heated metal example may suggest. For one thing, lawful correlation is not sufficient for causation, for example, arsenic poisoning may be correlated lawfully with death, but it does not follow that anyone who is poisoned with arsenic and then dies is caused to die by the poisoning (since they may get hit by a bus just after

they take the arsenic). Moreover, it has been argued that lawful correlation is not necessary for causation either; Anscombe (1971), for example, holds that *c* may cause *e* even though *c* and *e* do not fall under any regularity. However, a unified account of causes and laws has recently been proposed by Heathcote and Armstrong (1991): according to their view, the causal relation is (in fact, but not as a matter of necessity) identical with *N*, the relation in virtue of which laws of nature obtain.

See also: Causal Inference and Statistical Fallacies; Causation (Theories and Models): Conceptions in the Social Sciences; Causation: Physical, Mental, and Social; Counterfactual Reasoning, Qualitative: Philosophical Aspects; Counterfactual Reasoning, Quantitative: Philosophical Aspects; Determinism: Social and Economic; Empiricism, History of; Logical Positivism and Logical Empiricism; Natural Law; Probability and Chance: Philosophical Aspects

Bibliography

Anscombe G E M 1971 *Causality and Determination*. Cambridge University Press, London

Armstrong D M 1983 *What Is A Law of Nature?* Cambridge University Press, Cambridge, UK

Armstrong D M 1997 *A World of States of Affairs*. Cambridge University Press, Cambridge, UK

Berofsky B 1987 *Freedom from Necessity*. Routledge & Kegan Paul, London

Carroll J 1991 Property-level causation? *Philosophical Studies* **63**: 245–70

Carroll J 1992 The unanimity theory and probabilistic sufficiency. *Philosophy of Science* **59**: 471–79

Carroll J W 1994 *Laws of Nature*. Cambridge University Press, Cambridge, UK

Dupré J 1984 Probabilistic causality emancipated. *Midwest Studies in Philosophy* **9**: 169–75

Dupré J 1990 Probabilistic causality: A rejoinder to Ellery Eells. *Philosophy of Science* **57**: 690–8

Eells E 1991 *Probabilistic Causality*. Cambridge University Press, Cambridge, UK

Heathcote A, Armstrong D M 1991 Causes and laws. *Nous* **25**: 63–73

Hume D 1777/1975 *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd edn. Clarendon Press, Oxford, UK

Lewis D K 1973a Causation. *Journal of Philosophy* **70**: 556–67

Lewis D K 1973b *Counterfactuals*. Blackwell, Oxford, UK

Mellor D H 1995 *The Facts of Causation*. Routledge, London

Menzies P 1989 Probabilistic causation and causal processes: A critique of Lewis. *Philosophy of Science* **56**: 642–63

Menzies P 1998 How justified are Humean doubts about intrinsic causal links? *Communication and Cognition* **31**: 339–64

Skyrms B 1980 *Causal Necessity*. Yale University Press, New Haven, CT

Strawson G 1989 *The Secret Connexion*. Clarendon Press, Oxford, UK

Suppes P 1970 *A Probabilistic Theory of Causality*. North-Holland, Amsterdam

Swartz N 1985 *The Concept of Physical Law*. Cambridge University Press, New York

H. Beebee

Celebrity

1. Celebrity and Critiques of Mass Culture

The term 'celebrity' in its modern meanings began to be used in the nineteenth century, but the study of the phenomenon began in earnest with the rise of mass-produced culture, and in particular with the elaboration of an industrialized Hollywood film 'star system' in the early decades of the twentieth century. (For an excellent history of fame and fame discourse from ancient times to the near-present, see Braudy 1986.) It first emerged as a sustained focus of inquiry through mid-twentieth century criticism of mass culture, from both the left and the right. Although these works were rarely based on empirical research, and were filled with unsupported assertions about those creating and receiving celebrity images, they called attention to the mass production and management of celebrities, and to the question of the social impact of industrialized celebrity culture.

1.1 The Celebrity and Capitalist Ideology

Early perspectives on celebrity were largely theoretical, emerging as part of Marxist-influenced Frankfurt School cultural criticism of 'mass culture', 'mass society', and the 'culture industry'. Celebrities were seen as mass-produced, standardized commodities posing as unique human individuals, and celebrity discourse as a major ideological support beam for consumer capitalism. In the 1940s, Theodor Adorno and Max Horkheimer (1977), for instance, saw celebrities as the products of 'the culture industry', the cultural apparatus of mass society; Hollywood stars serve as distractions from the dissatisfactions created by industrial capitalism, and to manipulate 'the masses' into capitalism's false promises of both choice (standardized, mass-produced celebrities appear to be different individuals) and universal success (celebrities appear to demonstrate the rewards available to all).

Leo Lowenthal (1968), also writing in the 1940s, researched changes in 'mass idols' in popular magazines, charting the move from 'idols of production' (business and politics) to 'idols of consumption' (entertainment and sports); he too suggested that these popular culture heroes perpetuated the myth of an open social system, such that the existing social system is celebrated along with the star. C. Wright Mills

(1956, p. 71) wrote in the 1950s of the professional celebrity as a summary of American capitalist society's promotion of competition and winning; as 'the crowning result of the star system in a society that makes a fetish of competition', the celebrity shows that rewards go to those who win, regardless of the content of the competition. The definition of celebrities as mass-produced distractions, and their ideological role in promoting consumption, competition, individualism, and the myth of open opportunity, has continued in much contemporary cultural criticism.

1.2 Celebrity Versus Heroism

For more conservative cultural critics from the 1950s onward, the role of the celebrity system as an ideological support for capitalism was less important than its reflection of a large-scale disconnection between notoriety and merit. Such a view crystallized in the 1960s with the publication of Daniel Boorstin's *The Image* (1961), which distinguished celebrity from heroism. In an argument that presaged more recent postmodernist theory on 'simulation and simulacra' and the implosion of artifice and reality (see Baudrillard 1988), Boorstin argued that, with the growth of mass media, public relations, and electronic communication, it was possible to produce fame without any necessary relationship to outstanding action or achievement. Thus, the hero, whose fame is the result of distinctive action or exceptional, meritorious character, has been superseded by the celebrity, whose notoriety is manufactured by mass media without regard for character or achievement; the signs of greatness are mistaken for its presence. In Boorstin's definition, the celebrity is a 'human pseudo-event'. The phenomenon of celebrity is a symptom of a media-driven culture in which artifice has displaced reality, and in which merit and attention have become uncoupled. Although as Leo Braudy (1986) has shown, the oppositions between pure, 'real' fame and inauthentic 'artificial' celebrity do not fully hold up—historically, 'fame and merit have never been firmly and exclusively coupled—the conservative critical approach to celebrities as false, vulgarized heroes has pointed towards historically new features. Modern media, through the increasingly sophisticated creation, management, and reproduction of images, have an unprecedented capacity to place a person on the cultural radar screen, quickly and with no necessary reliance on the person's publicly celebrated actions or character.

2. Celebrities as Symbolic Entities

The interest shown by cultural critics in what the workings of contemporary celebrity tell us about the culture that makes it so central has been taken up in many humanities-based approaches to stars and stardom, which tend to consider the symbolic activity that takes place in and through celebrity discourse. Who gets attention, the logic goes, tell us much about the core values, or ideological contradictions, of the society giving the attention. Two themes have been particularly pervasive: the tension between egalitarian and aristocratic cultural strands; and the pursuit of the authentic self.

2.1 Celebrities as a Democratic Aristocracy

One striking feature of contemporary Western celebrity discourse is the way celebrities are treated to a cultural status that is simultaneously 'above' the rest of the populace and 'of' that populace. Celebrities are culturally constructed as a sort of elected aristocracy, both elevated and brought down by the watching crowds; the celebrity has become one symbolic means through which the population of the unfamous declares its own power to shape the public sphere (Marshall 1997). Moreover, while celebrity culture certifies some people as more deserving of attention and rewards because of their difference from the rest of the population, it also continually demonstrates that such people are ordinary, just like everyone else (Braudy 1986). Thus, popular celebrity discourse embodies an ambivalence about hierarchy in Western democracies: celebrities are celebrated for being better than, but no better than, those who watch them.

2.2 The Search for the 'Real' Self

A second outstanding feature of contemporary celebrity discourse is the thematic emphasis on getting 'behind' celebrity images to the 'true' or 'real' self. Celebrity discourse, as Richard Dyer (1991, p. 135) has demonstrated, involves a 'rhetoric of authenticity': the question of what a celebrity is 'really like', what kind of self actually resides behind the celebrity image, is a constant, whether in the form of tabloid exposés, behind-the-scenes reporting, celebrity profiles, or fan activities such as autograph-seeking. In part, this is because celebrities have the unique characteristic of appearing to audiences only in media texts, while also living in the world as actual human beings—they are images, but are 'carried in the person of people who do go on living away from their appearances in the media' (Dyer 1991, p. 135). In part, the theme of realness is the result of the increasing visibility over time of celebrity production mechanisms, raising the question of whether the celebrity image has been manufactured to attract an audience, or whether it reflects a true, deserving self (Gamson 1994). Celebrity discourse, with its heavy rhetorical emphasis on authenticity, thus manifests a larger cultural anxiety about the relationship between media images and lived realities.

3. The Political Economy of Celebrity

While many of the attempts to grapple with the unique symbolic or ideological features of contemporary celebrity have been either entirely speculative or based exclusively on textual analysis, much of the empirical research on the topic has focused on celebrity as an economic and social system. Influenced by the strategies of political economists and organizational sociologists, this research has investigated not so much the cultural meaning of celebrity as the internal organization and economic logic of the celebrity system. In contrast to approaches which assume that film stars are popularly selected for attention, for instance, such analysts tend to see celebrity as the result of 'the exigencies of controlling the production and marketing of films' (King 1986, p. 155). Although celebrities increasingly emerge in other social domains (politics, academia, etc.), most attention has been given to the major celebrity production center, the entertainment industry.

3.1 The Hollywood Star System

The pursuit of celebrity, especially in the entertainment business, became highly routinized, rationalized, and industrialized over the course of the twentieth century, with the development of industries, such as public relations, specifically devoted to the generation and management of public visibility. Celebrities are, in this context, marketing tools. In the notoriously risky entertainment business, which requires high capital investment for most of its products, a star is an insurance policy against audience disinterest, used primarily to minimize the risk of financial loss. Thus, star images are typically managed in accordance with the needs of the financiers of the vehicle with which a celebrity is associated—with film stars, for instance, a movie studio. The key nexus is not so much the celebrity and his or her audience, but the celebrity's backers, who pursue publicity, and journalists, editors, and producers, who provide it (Gamson 1994).

The structure of the star system has changed significantly over its brief lifetime. The early studio star system involved tight control of the production, exhibition, and distribution of films and their associated film-star images by several major studios; stars were under studio contracts, and studio publicity operations were responsible for producing and disseminating celebrity stories and images. When the studio oligopoly was broken up by a US Supreme Court decision in the 1950s, many more parties with a financial interest in celebrities' careers became involved in the management of celebrity images—personal publicists, managers, agents, in addition to the celebrity himself or herself, joined studio publicists in battling for control of the process. With the growth of television since the 1950s, and the explosion of celebrity-driven media outlets since the 1970s, moreover, it has become both easier to build celebrity and more difficult to retain it—hence Andy Warhol's famous declaration that eventually everyone will be world famous for 15 minutes. While Hollywood movie studios still generate a large proportion of American and international celebrities, celebrity has become less centralized, and the logic of celebrity has taken hold within a wider range of social spheres, including the worlds of literature, art, business, music, sports, and scholarship.

4. Fandom and the Reception of Celebrity

Analyses of the social and economic organization of celebrity tend to bracket questions of its cultural meanings, and textual analyses of celebrity tend to operate with untested assumptions or assertions about the meaning and impact of celebrity for audiences. Methodologically, both have tended to exclude empirical research into the meaning of celebrities and celebrity in the everyday lives of the fans or audiences encountering them. As the study of culture in general began, in the 1980s, to take more of a methodological turn towards audience research, audience-related aspects of celebrity have also come more sharply into focus. This territory remains, however, underinvestigated, in part because the study of audiences is both methodologically cumbersome and costly.

Considerable thought, if sparse research, has been devoted to the question of the fans' or audiences' relationship to celebrities. One early theory, for instance, proposed that mass media such as television facilitate a 'para-social relationship' between performers and audience members, in which the spectator comes to relate to the celebrity as if they were in a face-to-face relationship, with the 'illusion of intimacy' (Horton and Wohl 1956). Since then, largely through psychoanalytic film theory, discussions have focused on the processes by which audience members identify with celebrities, especially film stars (Tudor 1974; Stacey 1991). Various typologies of identification have been set forth, emphasizing quite a range of activities, uses, and types of attachment developed by audiences in their encounters with celebrities (Marshall 1997). This has been especially important in challenging the assumption that 'the audience' for celebrity is a homogenous mass, and that celebrities mean the same thing for all of its members. Para-social identification and celebrity hero-worship, for instance, both appear to be common stances; but ironic, playful, and irreverent interpretations of celebrity images also appear to be prevalent, especially as the pursuit of celebrity itself has become a more common focus of public discussion (Gamson 1994). Particular attention has been paid to the ways groups marginalized on the basis of race, class, gender, sexuality, ethnicity, or age make use of celebrity images for their own purposes:

building group solidarity or expressing rebelliousness and alienation, for example, through the celebration of particular types of stars (strong women, for example, or stars from their own ethnic group) (Stacey 1991).

5. *The Future of Celebrity*

While there is considerable consensus on certain distinctive features of contemporary celebrity, a tremendous amount of room remains for empirical investigations of the contours of celebrity as a socially organized system of meaning, status, identification, and pleasure; theory on the phenomenon remains much sharper and more sophisticated than research. Three areas in particular are likely to prove fruitful for future research. First, audience research can continue not only to document the range and types of audience positions vis-à-vis celebrities, but also to examine various possible explanations for the variance—audience characteristics, for instance, or qualities of different celebrity domains and types. Second, the process by which the logic of celebrity can and does spread to spheres other than entertainment, and how it may operate differently in those realms, is not well understood. Finally, nearly all of the literature on celebrity has emerged from, and focused on, the USA and the UK. Cross-cultural comparative research is a very promising, and almost entirely untapped, source of insight into the cultural, economic, political, and social logic of celebrity.

See also: Advertising: General; Charisma: Social Aspects of; Elites: Sociological Aspects; Media Effects; Media, Uses of; Reputation; Sport, Sociology of; Status and Role: Structural Aspects

Bibliography

Adorno T, Horkheimer M 1977 The culture industry: enlightenment as mass deception. In: Curran J, Gurevitch M, Woolacott J (eds.) *Mass Communication and Society*. Sage Publications, Beverly Hills, LA, pp. 349–83

Alberoni F 1972 The powerless 'elite': theory and sociological research on the phenomenon of the stars. In: McQuail D (ed.) *Sociology of Mass Communications*. Penguin Books, Harmondsworth, UK, pp. 75–98

Baudrillard J 1988 *Selected Writings*. Stanford University Press, Stanford, CA

Boorstin D J 1961 Jean Baudrillard: *The Image: A Guide to Pseudo-events in America*. Harper & Row, New York

Braudy L 1986 *The Frenzy of Renown: Fame & Its History*. Oxford University Press, New York

Dyer R 1991 A star is born and the construction of authenticity. In: Gledhill C (ed.) *Stardom: Industry of Desire*. Routledge, London, pp. 132–40

Gamson J 1994 *Claims to Fame: Celebrity in Contemporary America*. University of California Press, Berkeley, CA

Horton D, Wohl R R 1956 Mass communication and parasocial interaction: observations on intimacy at a distance. *Psychiatry* **19**: 215–29

King B 1986 Stardom as an occupation. In: Kerr P (ed.) *The Hollywood Film Industry*. Routledge and Kegan Paul, London

Lowenthal L 1968 The triumph of mass idols. In: *Literature, Popular Culture, and Society*. Pacific Books, Palo Alto, CA, pp. 109–36

Marshall P D 1997 *Celebrity and Power*. University of Minnesota Press, Minneapolis, MN

Mills C W 1956 *The Power Elite*. Oxford University Press, London

Stacey J 1991 Feminine fascinations: forms of identification in star-audience relations. In: Gledhill C (ed.) *Stardom: Industry of Desire*. Routledge, London, pp. 141–66

Tudor A 1974 *Image and Influence: Studies in the Sociology of Film*. St. Martin's Press, New York

J. Gamson

Censorship and Secrecy: Legal Perspectives

1. *Secrecy*

Secrecy involves norms about the control of information, whether limiting access to it, destroying it, or prohibiting or shaping its creation. Secrecy is a general and fundamental social process known to all societies. It can characterize interaction at any level—from information that an individual withholds, to secret rites of passage of preindustrial societies, to the secrets of contemporary fraternal or business organizations, to state-held information on national security. Secrecy norms are embedded in role relationships and involve obligations and rights to withhold information, whether reciprocal or singular. In preventing or restricting communication, the legally supported form of censorship discussed here involves secrecy. Yet most secrecy (e.g., concealing information about a surprise party or aspects of one's past) does not involve formal law and the law involves secrecy in many other ways.

In a democratic society secrecy and openness reflect conflicting values and social needs and exist in an ever-changing dynamic tension. Efforts to control information occur in a rich variety of contexts. Norms about the concealment of information and restrictions on communication ideally should be considered alongside of their opposites—norms mandating the revelation of information and protecting the freedom to know and communicate. Such norms may involve formal legal rules such as Britain's Official Secrets Act or the United States' Freedom of Information Act, nonlegally binding formal policies such as a bank's refusal to reveal client information in the absence of a warrant or the consumer information voluntarily provided on some product labels, or it may involve informal expectations (close friends are expected *not to reveal* shared secrets to outsiders but are expected *to reveal* certain personal details to each other, such as

true feelings about shared interests). The correlates and consequences of such variation offer rich material for analysis of the sociology of secrecy. This article reviews some selected social forms, processes, and consequences of secrecy, and the law as applied to censorship.

There is no widely agreed upon general theoretical or conceptual framework for considering secrecy issues. Given their social importance, there is a surprising lack of empirical or explanatory research seeking to understand the contours of secrecy and openness and why, and with what consequences, some forms have the support of law. Nor has there been much research contrasting different forms of legal secrecy.

Philosophers have considered ethics, (Bok 1989) students of politics the implications for democracy, (Shils 1956, Laquer 1985, Donner 1980, Moynihan 1998) and other social scientists have studied the patterning, processes, and correlates of information control rules across institutions and societies (Simmel 1964, Goffman 1969, Tefft 1980, Wilsnak 1980, Scheppele 1988). The largest body of work is by legal scholars; it emphasizes jurisprudence in often related areas such as the First Amendment, obscenity and pornography, national security and executive privilege, freedom of information, trade secrets, privacy and confidentiality, informant identities, fraud and implied warranties, but generally fails to explain broader social processes.

2. *Censorship*

2.1 *Definitions and Differences*

Censorship of communication in the modern sense is associated with large, complex, urban societies with a degree of centralized control and technical means of effectively reaching a mass audience. It involves a determination of what can, and cannot, (or in the case of nongovernmental efforts should and should not) be expressed to a broader audience in light of given political, religious, cultural, and artistic standards. Censorship may involve withholding or editing existing information, as well as preventing information from being created. In the interest of keeping material from a broader audience, content deemed to be offensive or harmful to public welfare is suppressed or regulated.

At the most general level any rule, whether codified or customary, proscribing self-expression (e.g., nudity, hairstyles, body adornment, language use) or the surveillance and suppression of personal communication (phone, mail) can be seen as a form of censorship. But our focus is primarily on state-supported efforts to control mass communication justified by claims of protecting the public interest, a form with profound implications for a democratic society.

Censorship assumes that certain ideas and forms of expression are threatening to individual, organizational, and societal well-being, as defined by those in power, or by those involved in a moral crusade, and hence must be prohibited. It presupposes absolute standards, which must not be violated.

Much censorship assumes that all individuals, not just children, are vulnerable and need protection from offending material—whether pornography or radical criticism of existing political and religious authority. Individuals cannot be trusted to decide what they wish to see and read or to freely form their own opinions.

Some censorship is largely symbolic, offering a way to enhance social solidarity by avoiding insults to shared values (e.g., a prohibition on flag burning). It may be a form of moral education as with prohibitions on racist and sexist speech. Or masquerading under high principles of protecting public welfare and morals, it may simply involve a desire to protect the interests of the politically, economically, and religiously powerful by restricting alternative views, and criticism or delegitimation of those in power.

Among the most common historical rationales are political (sedition, treason, national security), religious (blasphemy, heresy), moral (obscenity, impiety), and social (incivility, irreverence, disorder). These of course may be interconnected. What they share is a claim that the public interest will be negatively affected by the communication.

Censorship may be located publicly relative to other legal forms of secrecy. Censorship is justified by the *protection of public welfare*. Rationales for other legally supported forms include: the *protection of private property* for trade secrets; *economic efficiency and fairness* justifications in common law disputes over secret information; the *encouragement of honest communication and/or protection from retaliation* underlying forms such as lawyer–client and doctor–patient confidentiality, the secret ballot, and a judge's *en camera* ruling that the identity of an informant need not be revealed; the protection of *intimate relations* in the case of spousal privilege; the protection against *improperly elicited confessions* underlying the Fifth Amendment; the *strategic advantage* justification of sealed warrants and indictments; and the *respect for the dignity and privacy of the person* justification for limits on the collection and use of personal information, whether involving census, tax, library, or arrest (as against conviction) records. There has been little empirical research on whether, how well, with what consequences, and under what conditions these justifications are met.

Censorship is involuntary, unlike a nondisclosure agreement that parties to a court settlement voluntarily agree to. Censorship is unitary and nondiscretionary—those subject to it don't have the option of communicating. In contrast, the dyad of a

confidential professional relationship is discretionary for one party, such as the client; with the client's permission, a doctor or lawyer may reveal confidential information. Censorship seeks to withhold information from a mass audience, rather than a given individual, as with controversial laws preventing revelation of the identity of birth parents to adoptees. Where information exists but censors prevent its release, it is intended to *remain secret*. In contrast are legal secrets involving a natural *cycle of revelation* such as sealed indictments and search and arrest warrants, which become known when executed, or an industry confidentiality agreement, which may expire after a few years. Censorship as a form of secrecy stands alone. It is not reciprocally and functionally linked with its opposite—the legal mandate to reveal. For example some civil grand juries compel testimony, but then promise to keep it confidential.

Censorship is distinct from government regulation of fraudulent or deceptive commercial communication, which, unlike opinion and artistic expression, offers a clearer basis for empirically determining truth, as with the Federal Trade Commission's truth in advertising requirements. Censorship is separate from restrictions on communication based on copyright infringements, where the issue is not secrecy, but wrongful use. It is also distinct from editorial gatekeeping based on other criteria such as quality, cost, demand, and relevance, and in the case of regulating public demonstrations and entertainment, public safety and order. These can of course mask a desire to censor that would not otherwise be legally supported.

Government-legitimated censorship is distinct from censorious outcomes that may result from the actions of private groups. With the separation of church and state, only censorship by government has the support of law. Nongovernmental organizations such as a religious group or social movement may prohibit, or attempt to dissuade, members and others from producing, disseminating, or reading, listening to, or viewing material deemed objectionable. They may request editorial changes, advocate boycotts, and lobby school boards, libraries, bookstores, and theaters to exclude such material.

When we look at social processes of information control such as withholding information and selective presentation, a form of censorship may sometimes be seen in propaganda, public relations, and advertising, as public and private sector actors pursue their interest in creating favorable public impressions. Consider for example cigarette companies withholding information on the health risks of smoking or tire manufacturers not revealing the knowledge that their tires are unsafe.

2.2 Historical Development

Interest in the topic is strongly related to developments in communications technology and current events. In the modern period, continuing a trend that began with the printing press, new technologies, involving newspapers, mass-produced books and magazines, radio, telegraph, telephone, television, film, audiocassettes, video, fax, and the Internet, with their unprecedented ability to reach large numbers of people relatively easily, inexpensively, and efficiently, have created demands from conflicting groups for greater openness and freedom of communication and greater control over it. The conflict and debate continues—note conflicts over cable TV and efforts to regulate content and access to the Internet.

Following revelations in the USA about Watergate and government spying and disruption of the civil rights and antiwar movements during the 1970s, the groundbreaking Freedom of Information Act was passed and the Supreme Court strongly reaffirmed the principle of no prior restraint on the press in the Pentagon Papers case (*New York Times Co. v. United States 1971*).

As the examples of Socrates who chose to die rather than to have his ideas censored (or Plato who argued for censorship of the arts), the Romans who censored plays and banished offending poets, Pope Gelasius in the fifth century who issued the first papal list of prohibited books, and the Inquisition beginning in 1231 indicate, technology is hardly needed to spur censorship.

However, demands for censorship of religious and political ideas gained significant momentum in the fifteenth century with the appearance of that most subversive of technologies (after the invention of writing)—the printing press and the subsequent spread of literacy. This broke the historic monopoly, however limited, of religious and government institutions on communication with the masses. Authorities tried and are still trying (often in vain) to control the new techniques of mass communication. Later with the separation of church and state and the increased power of the nation state, the reach of religious censorship declined (e.g., prosecution for blasphemy) while political censorship gained in importance, as did ideas of free expression, which both countered and provoked censorship.

In the West, cultural values from the Enlightenment elaborated on by Immanuel Kant and later J. S. Mill and others stressed the importance of freedom of expression and openness as central to finding the truth, and for the stability and effectiveness of democratic government. Individuals were optimistically assumed to be responsible and rational beings who would reach the best conclusions, whether involving normative or empirical truth, with full information and discussion. Scientific ideals involving the ability to question and the freedom to communicate fit here as well. For both government and science, visibility or transparency is believed to be a central factor in accountability. Later arguments emphasized that the psychological well-being and dignity of the person

were best served by the freedom to express one's self and form one's own opinions. The argument based on personality has been stronger in Europe than in the USA.

In the second half of the twentieth century, with the allies' victory over fascist governments in World War II, the fall of colonialism, and the ending of the cold war, the cultural force of democratic ideals involving freedom of inquiry and expression has grown stronger. The principle of freedom of expression is contained in the First Amendment to the United States Constitution, various United Nations documents, European Constitutions, and documents such as the European Convention on Human Rights and Fundamental Freedoms. In the USA, the Supreme Court's extension of the protection of the First Amendment to the states meant that numerous nineteenth- and early twentieth-century state and local laws sanctioning censorship were in principle unconstitutional, although in practice there was often strong local support for such laws. This can be seen in struggles over education (e.g., the Scopes 'monkey' trial involving the teaching of Darwinism in Tennessee in 1925), the routine denial of First Amendment rights to labor protestors up to the 1930s and civil rights protestors through the 1960s, and various local struggles over efforts to ban books from libraries.

For Western style pluralist democracies, formal government censorship is the exception rather than the rule, at least relative to absolutist authoritarian regimes, which believe they have the only truth (whether political, religious, or moral) and do not permit opposing views. In 2000, an annual survey of press freedom found that 80 percent of the world's population lives in nations with less than a free press; about one-third of the countries were considered to have free press and broadcast systems; and one-third had systems with strong government control (Freedom House 2000). An extreme example is from the government of Iran where Salman Rushdie's book *Satanic Verses* was not only banned for being blasphemous, but a reward was offered for Rushdie's death.

Direct organizational means of government censorship must be considered separately from the availability of resources to create and distribute information and from informal means of censorship, whether by government or private interests, including self-censorship.

While freedom of expression is a central component of the modern democratic state, among democracies, there is considerable variation in censorship by content, media of communication, place, and time period. Constitutional and legislative guarantees of the individual's right to freedom of expression are not absolute. In considering the social consequences of exercizing a right, courts and legislatures balance it against other rights and community needs and standards, such as the presence of the clear and present danger that Justice Holmes wrote of in *Schenck v. United States* (1919). In the 1968 case of *United States v. O'Brien* the Supreme Court held that local laws could regulate time, place, and manner of expression if done in a content neutral fashion, narrowly tailored to serve substantial government interests and if alternative channels of expression were left open.

There is often disagreement about the social consequences of expression and how material should be defined. Does exposure to sexually or violently explicit words and images result in incitation and mimicry as some research claims (Itzin 1993, National Academy of Sciences 2000), or is it a safety valve and thus an alternative to action as others claim (Segal and McIntosh 1993, Hein 1993)? Does the prevalence of violent and sexual content reflect or create public demand? Can a reasonable consensus be reached on the distinction between pornography and erotic art? Can heterogeneous, rapidly changing societies with multiply porous borders meaningfully talk of community standards?

There has been little research on variation in censorship. Political and religious expression has generally received greater legal support than other forms such as sexual expression. Printed matter has greater protection than other media. Film, live audience presentations, the Internet, and cable TV have greater protection than conventional television and radio where there is a scarcity of spectrum. Artistic expression likely to raise the concerns of censors is generally ignored until it seeks to reach a mass audience via the media or museums. Material appropriate for adults may not be suitable for children. Freedom of expression and access to information generally have greater protection in the USA than in Europe (e.g., greater tolerance of hate and other offensive speech, and stronger freedom of information laws and protections against libel suits).

2.3 Methods of Censorship

Three major means of direct censorship are *preventive* in nature. Their goal is to stop materials deemed unacceptable from appearing, or if that is not possible, from being seen or heard by prohibiting their circulation:

Formal prepublication review. This requires would-be communicators to submit their materials for certification before publicly offering them. Soon after the invention of the printing press, the church required review and approval before anything could be printed. However impractical and difficult to enforce in the contemporary period, to varying degrees such 'prior restraint' is found in authoritarian societies, whether based on secular political (as in Cuba) or religious doctrines (as in Iran and Afghanistan) at the turn of the century. It may be seen in democracies during emergency periods such as a war. There may be formal

review boards, or censors may be assigned directly to work at newspapers and broadcasting stations.

Government or interest-group monopolization of publication. Here the censors in effect are the producers and are the only ones allowed to offer mass communication. For much of its history the church was intertwined with government and was in effect the only publisher. In the former Soviet Union the press and media were government controlled and private means were prohibited.

Licensing and registration. The means of production and transmission of information may be limited to trusted groups who agree to self-censorship in light of prior restrictions. In England in the sixteenth century, printing was restricted to one official company and all books had to be cleared by religious authorities prior to publication. Four centuries later, China required that all Internet content providers be registered with the government and abide by vague content restrictions. Permission may be required to own a printing press, and in some countries, even ownership of a typewriter has been regulated.

Government subsidized programs for the arts and journalism may come with political and cultural strings attached. In the Soviet Union, sponsorship of artists' and writers' associations stressed 'socialist realism,' a doctrine that held that art should serve the purposes of the state. Those rejecting this doctrine were neither subsidized nor offered access to the public, and they risked prosecution, as with Nobel Prize winning author Alexander Solzhenitsyn.

In the USA in 1990, under prodding from Congress, The National Endowment for the Arts required that grant recipients sign a nonobscenity oath and that artistic merit be determined by taking into account general standards of decency. A federal court held that the decency clause was too broad and that public funding of art was entitled to First Amendment protection (*Karen Finley et al. v. National Endowment for the Arts and John Frohnmayer 1992*).

A related aspect involves an informal *de-licensing* in which individuals deemed to be untrustworthy relative to the official standards are prohibited from communicating. For example, during the 1950s, Hollywood film writers suspected of communist sympathies were prohibited from working in the industry via a black list.

A more subtle form of exclusion involves denying access, as when government officials provide information only to favored journalists believed to put an acceptable slant on their reports. Even where the means of communication are freely available in a legal sense, inequality in resources often means that many potential voices go unheard. Journalist A. J. Leibling has observed, 'freedom of the press is assured to those who own one.' A related issue involves the trend toward consolidation of newspapers, magazines, television, and motion pictures under fewer and fewer owners. Such monopolies are unlikely to express as wide a spectrum of viewpoints as would be found with more decentralized ownership.

A related area of licensing can be seen in local requirements that those wishing to hold a public demonstration obtain a permit. Given constitutional protections, such permits are usually granted in the USA, although there may be restrictions justified by the need to maintain public order.

In Western democracies, broadcast media (e.g., radio and television), unlike print media, are subject to licensing. The scarcity of bandwidth requires government regulation and, depending on the criteria used, can be an invitation to censorship. The US Federal Communications Commission, for example, has ambiguous rules regarding the control of broadcast content. The use of certain four letter words deemed to be indecent is prohibited. Although rarely exercised, there is the possibility of license revocation or nonrenewal for violations.

After comedian George Carlin used the word 'fuck' in a late night broadcast in 1973, the offending radio station received a warning letter from the FCC. The station then sued, claiming that FCC regulations on indecent speech violated the First Amendment. The Supreme Court (*FCC v. Pacifica Foundation 1978*) upheld the FCC action and added an additional controversial, very broad, censorship rationale known as the 'pervasiveness doctrine.' Under this doctrine, regulation is required because, 'the broadcast media have established a uniquely pervasive presence in the lives of all Americans' and offensive and indecent material delivered over the airwaves confronts the citizen, 'not only in public, but also in the privacy of the home.' Given the ease with which indecent communications may enter the home, children must be protected from unwillingly or willingly encountering them.

The elastic quality of a standard such as 'pervasiveness' could be used to justify censorship of any form, even of books and newspapers that also pervade society. Indeed, when Congress passed the Communications Decency Act in 1996 to regulate Internet content, a medium not characterized by spectrum scarcity, it was argued that the Internet pervades the home just as the radio and television do and hence must be regulated. However the Supreme Court found this Act unconstitutional in *ACLU v. Reno* in 1997. As the Internet evolves, and to the extent that it becomes a platform for delivering voice and video communications that parallel traditional broadcasts, conflicts over the appropriateness of its regulation will likely intensify.

In contrast are means applied *after the fact*, which seek literally to block or destroy communications or to punish and deter. A classic example, likely seared on the memory of anyone who has seen it on film, is the Nazis' burning of books in 1933. Materials originating from suspect sources that do not cooperate with censors and/or are from outside a country may be

categorically blocked. This may happen through technical means as when the former Soviet Union electronically jammed communications of Radio Free Europe or through the seizure of material at borders. China has created an electronic wall around its Internet to block access to material from nonapproved sources (e.g., among sites blocked are the New York Times and CNN). Until the United States Supreme Court (*US v. One Book Entitled Ulysses 1934*) found that James Joyce's *Ulysses* was a work of art, even though it contained 'dirty words,' US customs authorities routinely seized literature deemed inappropriate. The US Postal Service prohibits the importation and domestic transmission of certain forms of obscene communication. With the 1934 decision we see the seeds of later Supreme Court rulings such as *Miller v. California 1973*, which held that work could be prohibited only if, taken as a whole, it had no redeeming value as art or science, was patently offensive, and was not in keeping with local community standards.

The conflict between the principle of no prior restraint and any modern society's legitimate need to control some communication has resulted in a variety of after-the-fact sanctions (e.g., criminal offenses involving espionage, revealing national secrets, obscenity, pornography, and incitement, injunctions to cease publication and administrative sanctions; and civil remedies such as invasion of privacy, libel, and defamation).

In the USA in 2000 (in contrast to Britain, which has an Official Secrets Act), the disclosure of properly classified information (with certain exceptions such as information on the design of nuclear weapons and the names of intelligence agents) is not a crime. However disclosing such information can result in losses of security clearance, dismissal, and fines.

Communication has a special quality. Unlike most other legally regulated subjects, most communications cannot be legally prohibited before they are offered, but once offered, legal penalties may apply. Thus, there is a paradox and some uncertainty and risk for communicators, particularly those pushing boundaries. The goal here, beyond punishment for the infraction in question, is to warn and deter others in the hope of encouraging self-censorship.

On a statistical basis, the major form of censorship in western societies is self-censorship. Publishers, editors, and producers of mass communications are aware of boundaries not to be crossed, even though there are many gray areas. Communicators generally stay within the borders, whether to avoid prosecution or lawsuits, to avoid offending various social groups, to keep the channels of government information open, to please stockholders and advertisers, or out of their sense of patriotism and morality.

Press and broadcast organizations (e.g., The National Association of Broadcasters) and the major newspapers and television networks have codes of ethics and voluntary standards. Between the time a journalist writes a story and its appearance there are several levels of review. The large broadcasting companies have internal units that review everything from advertisements to program content before they appear. When Elvis Presley appeared on national television for the first time, only his upper body was shown, given censors' concern with what was then considered to be his prurient hip shaking.

Another prevalent nongovernmental means, often undertaken to avoid the threat of more stringent government controls, involves voluntary rating systems. Here the goal is not to ban the material but to give consumers fair warning so they can make up their own mind. In the 1920s, the Motion Picture Association of America created a seal of approval for films meeting its standards. In 1968, it created its rating system (expanded in 2000) for films based on nudity and violent content. Some comic books, TV programs, music videos, video games, music, and web sites are also rated. This may be welcomed as consumer information or seen as censorship that can chill expression and create an undesirable climate, opening the door to greater control.

Technical means of information control such as the v-chip, which permits programming a television set so it will not receive material deemed objectionable, and various software filters, which screen web sites for sexual and violent content, facilitate private control. Such means are seen as more efficient than a heavy-handed government censor and more consistent with an open, highly heterogeneous society. Those who want such material have access while those who might be offended can avoid exposure.

There are degrees of censorship. More common than outright prohibition, particularly in the case of erotic material (which has received increased legal protection in recent decades in the face of local legal prohibitions that appeared in the nineteenth century), are time, place, manner, and person restrictions—whether required by government or undertaken voluntarily. Potentially offensive material is segmented and walled off from those for whom it is deemed inappropriate. For example, pornographic material may be restricted to red-light districts, to adults, and to late night programming when minors are presumed not to be watching; children may be prohibited from certain concerts, and stores may refuse to sell them violent videos.

2.4 Limitations of Censorship

While government censorship makes a symbolic statement, it is often rather impractical beyond the short run, given the ubiquitous nature and continual improvements in mass communication technologies and the leaky nature of most social systems. Of course computer-based technologies may make it easier to

track whom an individual communicates with and what material they access. But on balance, technology appears more likely to be on the side of freedom of expression than the side of the censors. The ease of modern communications, in particular remote forms (such as radio, television, fax, and the Internet) whose transmission can transcend national borders and means of reproduction (such as photocopiers, scanners, audio- and videotaping, and printing through a computer) that are inexpensive and relatively easy to use and conceal limit the ability of censors. The Internet, if available, has the potential to make everyone a publisher. Its 'many to many' communication through labyrinthine networks (chat rooms, bulletin boards, and e-mail) is far more difficult to censor than the traditional 'one to many' communication of the newspaper or television station and it is less expensive.

Given the expanding scale of published material and the diffusion of communications technology that began with the printing press, government and religious bodies are forever trying to catch up. This is the case even in modern highly authoritarian settings. For example in China during the Tiananmen Square protest, fax technology kept China and the world informed of the events, and in Iran the fall of the Shah was aided by smuggled audiotapes urging his overthrow. Strong encryption, which protects messages, also makes the censor's task more difficult.

Beyond technical factors, censorship is often accompanied by demand for the censored material. Censoring material may call attention to it and make it more attractive (the forbidden fruit/banned in Boston effect). Black market demand for such material makes it likely that some individuals will take the risk of creating and distributing it, whether out of conviction or for profit. Potential communicators often find ways to avoid or deceive censors whether using satire, parable, code language, changing the name of prohibited publications, or through simply defying the law, as with the many underground '*samizdat*' presses that challenged communist rule in Eastern Europe.

It is also the case that sometimes, 'the truth will out.' In a democracy illegitimate political censorship in the name of national security or executive privilege is vulnerable to discovery—note Watergate and the Iran-Contra affair. In spite of its dependence on government sources, the mass media may play an important counterbalancing role here, watching those who seek to watch. Beyond investigative reporters often using the Freedom of Information Act, such 'dirty data' may be revealed by the legal procedure of 'discovery' in court cases, by experiments and tests, by leaks, whistleblowers, and participants with a Dostoevskian compulsion to confess and by uncontrollable contingencies such as accidents (e.g., the crash of an airplane carrying Watergate 'hush' money). The more complex and important a cover-up or illegal conspiracy, the more vulnerable it is to revelation. Even most legitimately classified government secrets have a shelf life and must be revealed after 75 years.

In the long run it is also difficult for censors to deny pragmatic outcomes and those that are empirically obvious. This raises the intriguing sociology of knowledge question of the relationship between culture with its significant, but not unlimited, elasticity and a level of reality or truth that, when in conflict with culture, may erode it over time. No matter what the power of the church to prosecute Galileo for heresy and ban his work, or the amplitude of its megaphone to assert the earth was flat, it could not suppress the truth for long.

See also: Censorship and Transgressive Art; Censorship in Research and Scholarship; Communication and Democracy; Confidentiality and Statistical Disclosure Limitation; Mass Communication: Normative Frameworks; Privacy of Individuals in Social Research: Confidentiality; Secrecy, Anthropology of

Bibliography

Aumente J 1999 The role and effects of journalism and *Samizdat* leading up to 1989. In: Aumente J, Gross P, Hiebert R, Johnson O, Mills D (eds.) *Eastern European Journalism Before, During and After Communism*. Hampton Press, Cresskill, NJ

Bok S 1989 *Secrets: On the Ethics of Concealment and Revelation*. Vintage Books, New York

Coetzee J 1996 *Giving Offense: Essays on Censorship*. University of Chicago Press, Chicago

Donner F 1980 *The Age of Surveillance: the Aims and Methods of America's Political Intelligence System*. Knopf, New York

Freedom House 2000 *Press Freedom Survey 2000*. Freedom House, Washington DC

Green J (ed.) 1990 *The Encyclopedia of Censorship*. Facts on File, New York

Goffman E 1969 *Strategic Interaction*. University of Pennsylvania, Philadelphia, PA

Hein M 1993 *Sex, Sin, and Blasphemy*. Free Press, New York

Herman E, Chomsky N 1988 *Manufacturing Consent: The Political Economy of the Mass Media*. Pantheon, New York

Itzin C 1993 *Pornography: Women, Violence and Civil Liberties*. Oxford University Press, New York

Laqueur W 1985 *A World of Secrets*. Basic Books, New York

Marx G T 1984 Notes on the collection and assessment of hidden and dirty data. In: Schneider J, Kitsuse J (eds.) *Studies in the Sociology of the Social Problem*. Ablex, New York

Moynihan D 1998 *Secrecy: The American Experience*. Yale University Press, New Haven, CT

National Academy of Sciences 2000 Promoting Health: Intervention Strategies from Social and Behavioral Science Research. National Academy Press, Washington, DC

Pool I D S 1983 *Technologies of Freedom*. Belknap Press of Harvard University Press, Cambridge, MA

Popper K 1962 *The Open Society and its Enemies*. Routledge & Kegan Paul, London

Segal L, McIntosh M (eds.) 1993 *Sex Exposed: Sexuality and the Pornography Debate*. Rutgers University Press, New Brunswick, NJ

Scheppele K 1988 *Legal Secrets: Equality and Efficiency in the Common Law*. University of Chicago Press, Chicago

Shils E 1956 *The Torment of Secrecy*. Free Press, Glencoe, IL
Simmel G 1964 *The Sociology of Georg Simmel* (ed. Wolff K). Free Press, Glencoe, IL
Strum P 1999 *When the Nazis Came to Skokie: Freedom for Speech We Hate (Landmark Law Cases and American Society)*. University of Kansas Press, St. Lawrence, KS
Sunstein C 1993 *Democracy and the Problem of Free Speech*. Free Press, New York
Tefft S 1980 *Secrecy: A Cross Cultural Perspective*. Human Sciences Press, New York
Wilsnak R 1980 Information control: A conceptual framework for sociological analysis. *Urban Life* (Jan)

G. T. Marx

Censorship and Transgressive Art

'Transgressive art' violates aesthetic or broader cultural norms. It provokes, challenges, and shocks audiences, causing anger, insult, or anxiety. Restrictions such as *censorship* may result, where parties inside or outside the production process impose modifications on work, force its suppression, or induce creators to self-impose constraints on what they make in order to avert potential problems. Artistic actions and the social reactions they call out are thus organically linked. Transgression in the world of art can also be a vital source of innovation.

1. The Term Transgressive Art: Three Meanings

Transgressive art has become a buzzword in the art world of the mid- to late 1990s. In the broadest sense, it connotes *any* artwork that offends audiences. That includes creations that could be deemed blasphemous for violating religious doctrines, be they traditional religions or dominant civic creeds; work that defies social hierarchies and would be regarded as subversive by those in power; and art that deals with 'unsavory' or 'impolite' topics. *Transgressive art* thus becomes a catch-all term, embracing a wide variety of work within its scope. In this regard, art is transgressive if it is prickly. In another age, the same material might have been condemned or dismissed as obscene, heretical, or merely 'bad art.'

But '*transgressive art*' also has two more specific, and inter-related, meanings. First, *transgressive art* corresponds to philosopher Immanuel Kant's '*sublime*,' as opposed to the 'beautiful.' Whereas the beautiful is soothing, representing order and regularity, the *sublime* is surprising, it embraces the obscure, the novel, and the terrible. The *sublime* challenges the senses and produces restlessness, as opposed to satisfaction. The *sublime* excites fear; it is unpredictable and unanticipated. Although this distinction came to prominence in the mid-eighteenth century, it has obvious relevance to the contemporary *avant garde*, for whom unsettling the status quo is a primary principle.

Second, artists have expanded the meaning of aesthetic since the mid-1980s to beyond the object, and beyond the art world. This type of art supersedes the aesthetic formalism of Clement Greenberg by actively engaging the world, critically dissecting it, and proposing an alternative future (Kester 1998). 'Art' frequently encompasses activist interventions in the art world or the larger community, uncovering disguised social processes such as power, racism, and sexism (e.g., the feminist collective Guerrilla Girls, or Hans Haacke, whose 1971 exhibition unmasking the complex operations of a Manhattan slumlord was canceled by the Guggenheim Museum because of its 'extra-esthetic' nature). It can therefore encompass process rather than object. This means expanding the concept of 'public art' beyond heroic depictions of 'great men' and their deeds, and beyond large-scale, transcendent 'masterpieces' by confirmed artistic geniuses, to face-to-face collaborations with real communities (Jacob 1995), or enlarging the definition of art to include the voices of formerly marginalized groups, presented in untraditional venues, and employing unconventional media. This can lead to the violation of norms and taboos broader than those of the art world itself. The first page of the catalogue accompanying the Whitney Museum of American Art's exhibition 'Abject Art: Repulsion and Desire in American Art' (1993) boldly represents this position: '[O]ur goal is to talk dirty in the institution and degrade its atmosphere of purity and prudery by foregrounding issues of gender and sexuality in the art exhibited' (Ben-Levi et al. 1993, p. 7). The materials in the object list included dirt, hair, excrement, dead animals, and rotting food.

1.1 The Contemporary Coalescence

Generally speaking, since the last quarter of the nineteenth century—with the rise of Impressionism, and the breakdown of the Academy system throughout Europe—artists have veered between two major aesthetic strategies. One approach has been 'art for art's sake,' where artists produce work primarily for a small group of like-minded individuals, using what could be compared to an exclusive language of expression. The other style is producing art aimed for the wider public, incorporating social and political themes. The first strategy is aesthetically daring but hermetic; the second has often been aesthetically conservative but liberally accessible. Dada and Abstract Expressionism are examples of 'art for art's sake'; the social realism that dominated art production

in the United States during the Depression of the 1930s, or the work of muralists such as Diego Rivera who were active throughout the same era in Mexico, represent socially and politically oriented art.

But contemporary artists are distinctive: they frequently have mobilized *both* approaches, combining inventive techniques with urgent social issues. What has been the impetus for this dramatic innovation? Artists have been pulled from the cloister of their studios by larger social events since the mid-1980s. The AIDS crisis hit American artists in disproportionate numbers. They responded to inadequate or biased media coverage, and the relative dearth of official political responses to the epidemic, by treating this subject with an activist stance (the now-ubiquitous 'Silence = Death' logo was originally a neon installation in the window of New York City's New Museum of Contemporary Art). Furthermore, the empowerment movements that gained strength during the 1960s—particularly the civil rights movement, feminism, and gay liberation—affected the art world by expanding opportunities for artists and art administrators from these groups, as well as making artists more self-reflexive about the structure of the art world and communities beyond. Finally, conservative religious and political leaders in the United States who reacted to the success of the disenfranchised in expressing themselves led a series of attacks to cut off public funding of the arts, targeting unconventional expression. Contemporary artists have used both traditional and *transgressive* strategies to confront these pressing issues and resist these assaults.

1.2 The Fundamental Nature of Transgression

'Transgression' signifies that rules have been violated. The most useful conceptual tool to understand the significance of this is the work of anthropologist Mary Douglas on moral pollution. In order to structure interaction, every society consolidates experience into what Douglas calls '*natural categories*,' binary oppositions that consign phenomena to one classification or another. Significant *natural categories* include *sacred* and *profane*, public and private, masculine and feminine, internal and external. Societies differ in where they assign particular behaviors, but the process is basic across human groups.

Any infraction of these *natural categories* resounds throughout society, threatening the sanctity of other distinctions as well. Controversial artists have run afoul of convention, and have been prime movers in what one observer has termed a 'category crisis' (Garber 1992), a pervasive blurring of previously precise *symbolic boundaries*. All have committed these figurative affronts in some manner or another.

Such violations are not entirely new. For example, Jean-Baptiste Carpeaux produced a sculptural group for the facade of the new Paris Opéra, *La Danse* (1865–9), featuring a cluster of frenzied female dancers surrounding a *génie de la danse*. The central figure is *sexually* ambiguous, bearing the slim body of a male youth capped by a female face. S/he troubled many viewers and became the lightning rod for a heated *public controversy*, as did *Princess X*, a sculpture by Constantin Brancusi (1916). In one respect this work resembles the classic pose of a Madonna. But it also unmistakably resembles a set of male genitalia, thus challenging the categories of male and female, *sacred* and *profane*. Furthermore, a production of *The God of Vengeance* (penned by playwright Sholem Asch, 1910) ran afoul of public opinion and the law in New York City in 1922. This is the tale of Yankel, who operates a brothel in the basement while he maintains a 'wholesome' atmosphere for his wife and chaste daughter, Rivkele, upstairs. *Sacred* and *profane*, public and private are segregated—until Rivkele falls in love with one of the prostitutes. Such cause *célèbres* were novel in their time; they are much more commonplace today.

The photographer Andres Serrano's notoriety was established by *Piss Christ* (1987), where he placed a plastic crucifix into a container of his own urine. Whereas many people were repulsed by the idea of mixing *sacred* and *profane* when this work came to wide public attention in 1989, one canny reviewer shifted the figure/ground perspective: instead of seeing the crucifix as being despoiled, perhaps more importantly the urine was being sanctified. After using a variety of vital fluids in his work (including milk, blood, and semen), subsequent projects have included *The Morgue* (photos that provided a peak into the generally forbidden realm of corpses, in 1992), *A History of Sex* (1997), which featured images of intergenerational *sexual* relations (an 87-year-old grandmother with her young lover), and a nude woman stroking the erect penis of a horse (underscoring the cultural regulation of 'proper' *sexual* expression and the *symbolic boundaries* of taboo).

At about the same time that Serrano was vilified, so was the photographer Robert Mapplethorpe. A significant portion of Mapplethorpe's work features *sexually* explicit or homoerotic images. A few nude or seminude portraits of children were also judged in *sexual* terms by some social critics. Mapplethorpe's extensive series of self portraits represents a range of *transgressions*: in 1978 he became the Devil incarnate with his leathers and a bullwhip stuck into his rectum; in 1980 he was the ethereal waif with lipstick pout, mascaraed eyes, and naked, vulnerable torso. From demon to *sexually* ambiguous figure, Mapplethorpe tweaked public sensibilities in the 1980s and 1990s.

Finally, the widespread outcry against Martin Scorcese's film *The Last Temptation of Christ* (1988) and Salman Rushdie's novel *The Satanic Verses* (1988) have similar origins in violations of *symbolic boundaries*. In both instances, major religious figures were demystified by focusing on their humanity, alongside

their divinity. But to the fundamentalist mind, this was an unbearable assault on received wisdom and sanctified territory (Dubin 1992).

1.3 Bad Girls and Gender Deviation

As demonstrated, protection of *natural categories*, and intolerance of their combination or violation, are at the base of a significant portion of modern-day controversies about art. Feminist performance artists Karen Finley and Annie Sprinkle *transgress* social expectations of femininity, the self, and the 'proper' display of the body in public: they use 'earthy' language, vault from persona to persona, and reveal intimate parts of their bodies and their functions (Sprinkle invites audience members to examine her genitals with a flashlight; Finley smears the outside of her body with raw eggs, squirts breast milk, and rubs yams over her buttocks, upsetting the presumed order of nature). In one of her best-known acts, *We Keep Our Victims Ready*, Finley slathers her nearly naked body with chocolate she scoops out in handfuls from a heart-shaped box. Next she sprinkles herself with tiny red cinnamon candies, followed by a layer of alfalfa sprouts, and topped off with Xmas tinsel. Her body has become her canvas. Id trumps everything else in this type of expression (Dubin 1992).

Finley and Sprinkle continue the tradition of feminist 'bad girls' who have pioneered performance and 'body art' since the 1960s, such as Carolee Schneeman: in *Meat Joy* (1964) she enlisted others to join her in an uninhibited romp with raw fish, chickens, and other items; in *Interior Scroll* (1975), she literally gave birth to her art as she removed a rolled text from her vagina. In more recent times, French artist Orlan has embarked on a series of seven plastic surgeries which will transform her face into an amalgam of idealized female features: Mona Lisa, the goddess Diana, and Botticelli's Venus among them. Photographer Marina Vainshtein has altered her body in a different respect: her torso is covered with tattoos evoking the Holocaust, including skeletons, swastikas, barbed wire, crematoria, and the notorious inscription on the gate to Auschwitz, *Arbeit macht frei* ('Work leads to freedom'). Her 'project' *transgresses* traditional Jewish prohibitions against altering the body, at the same time that it symbolically re-enacts and then repossesses or 'controls' the violation committed by the Nazis when they dehumanized their victims by tattooing numbers on their skin.

These defiant women chafe at the way the female body has been objectified in the past, and challenge the constraints imposed upon them by men. The 1994 exhibition *Bad Girls* (The New Museum of Contemporary Art, New York) began with a linguistically loaded title: 'bad' becomes 'good,' a badge of honor from this perspective; and 'girl,' freighted with indignity when uttered by a condescending male, becomes an affectionate code word for sisterhood. Works in the exhibition highlighted women who self-consciously work beyond the *symbolic boundaries* of prim femininity, skewering the cultural assumptions of how they should act.

Of course, male artists have used their bodies as an artistic medium as well: Vito Acconci's *Seedbed* (1971) featured the artist masturbating beneath a ramp in a New York City gallery, while visitors strolled above. Chris Burden's *Shooting Piece* (1971) involved having a friend shoot him in the arm; in *Trans-Fixed* (1974), he was crucified over the back of a Volkswagen. The late Bob Flanagan famously nailed his penis to a board. Ron Athey currently performs ritualized piercings and mutilations. These performances transgress convention, to be sure. But in vital ways they harmonize with male norms of *sexuality* and bravado, unlike their female artistic counterparts who more directly challenge *gendered* norms of behavior.

2. Transgressive Art and Social Response

During every epoch and in every conceivable place where modern cultures have developed, we can expect to observe this peculiar dialectic: a particular type of social control evolves to address specifically whatever the powers-that-be find threatening. The 'problem' inevitably calls out the 'solution.' This is what philosopher Michel Foucault described as 'perpetual spirals of power and pleasure' (Foucault 1978), which symbiotically bind opponents to one another in pitched battle. In other words, enforcing the rules and violating them both provide gratification. Prelate and infidel, prime minister and traitor, each characteristically complements the other in the struggle for ascendence.

2.1 Censorship: Different Forms

For all its flaws—it is an extremely vague term, all too promiscuously applied to depict any obstacles your opponents erect to limit, direct, and/or interfere with your absolute freedom of expression—*censorship* is the most recognizable and purposeful term we have to cover such situations. (Unlike Eskimos, with their purported multiplicity of descriptors of snow, speakers of English encounter a serious deficit of phrases characterizing constraints on expression.) '*Censorship*' is a contested term; its meanings are reckoned along a variety of axes.

Censorship's visibility is one pivotal characteristic, be it overt or covert, *de jure* or *de facto*, regulative or constitutive (Jansen 1991). Much *censorship* is occult, brought to light only when taken-for-granted procedures fail to stem the tide of forbidden thoughts, or

when they are unable to anticipate when and where new types of danger may erupt, and in what form. As the understanding of hegemony alerts us, the most efficient form of social control is through ideas, not by force. Whenever a society effectively socializes everyone into supporting the same core belief system, controls will seem natural, matter-of-fact, unremarked. This promotes the sort of stability where prescription and proscription dictate and regulate nearly all thought and behavior. An efflorescence of *censorship* therefore commonly signals some breakdown, an unraveling of established rules.

Censorship's essential nature and primary targets are additional attributes. Emile Durkheim's justly famous dictum asserts that deviance is to be expected even in a society of saints (Durkheim 1938). In a religious community, then, like the seventeenth century Italian monastery swept by a witch craze in Aldous Huxley's *The Devils of Loudun* (1952), blasphemy and heresy will be the shape given to doctrinal aberrations, and punishments will be meted out accordingly. In other societies where a puritanical moral system is to be defended, *censorship* will be directed at alleged instances of perversity. Whatever fashions the fundamental character of the community also shapes how the threats to it are perceived—those things which must be banished from view and from earshot.

The particular site or source of *censorship* is also critical. In its classic, perhaps purest form, *censorship* is a prerogative of government, be it civil or ecclesiastical. *Censorship* is something officials are entitled to impose, hence state *censorship*. Market *censorship* exists as well: certain groups are excluded from having their say because they lack the resources or access which would allow them to disseminate their ideas easily. Systematic exclusion from the marketplace of ideas typically parallels the exclusion from other opportunities and rewards a society may have to offer some of its citizens, but not to others, characteristically because of their race, religion, *gender*, class, or *sexual* orientation. Finally, we must consider self-*censorship*, where individuals hold themselves in check, consciously or unconsciously, either out of fear of bringing official action down upon their heads, or from the fatigue and frustration which results from continually trying to get past cultural gatekeepers and getting rebuffed.

These are, of course, ideal types. In any specific case, one or more of these forms of regulation might be operative, either simultaneously or sequentially. *Censorship* may occur at any point along the process of production or distribution of cultural goods, and in cases where hegemony maintains a stranglehold upon independent thinking, even at the point of reception and consumption (Dubin 1997).

The main examples of *transgressive art* cited earlier have all endured negative reactions and restrictions of one type or another. Carpeaux's sculptural *gender*-bending provoked extensive debate in periodicals of the day, suffered defacement with a bottle of ink, and even an order of banishment issued by Napoleon III. The salon president had Brancusi's *Princess X* removed from the Salon des Indépendants in 1920, preempting the possibility of adverse public response or even police action. All the cast members and the producer of *The God of Vengeance* were arrested and charged with obscenity, immorality, and indecency (they were found guilty, but their sentences were suspended; the play was forcibly closed). Theaters screening *The Last Temptation of Christ* endured bomb threats, picketing, and vandalism, and it was even banned in certain locales. Iran's Ayatollah Khomeini declared Salman Rushdie's *The Satanic Verses* blasphemous in 1989 and issued a *fatwa* or death decree against the Indian-born writer; 10 years later, Rushdie (who is now a British citizen) was tentatively beginning to make public appearances again.

Furthermore, a long and acrimonious national debate erupted in the United States over the work of Andres Serrano, Robert Mapplethorpe, and Karen Finley; it reached from the halls of Congress to backwater radio talk shows. In each case, a small amount of federal money had been appropriated by the National Endowment for the Arts (NEA) to underwrite exhibitions of their work. Senator Alphonse D'Amato ripped up the catalogue from an NEA-sponsored fellowship show that included the photo *Piss Christ*, on the Senate floor in 1989. Moreover, an enraged teenager damaged *Piss Christ* with a hammer while it was on display in an Australian museum in 1997; administrators subsequently closed the exhibition of Serrano's work 'for security reasons.' His *History of Sex* provoked bomb threats in the Netherlands.

The Contemporary Art Center in Cincinnati, and its director Dennis Barrie, were indicted for pandering pornography when it hosted a Robert Mapplethorpe retrospective in 1990 (they were acquitted). Karen Finley was denied an NEA fellowship in 1990 after she had gained notoriety as 'the chocolate-smeared woman' and became the poster girl for what conservative opponents of the NEA felt was an illegitimate use of public funds (she and three other performance artists—the 'NEA Four'—later successfully sued to have their fellowships restored). Significant threats and impediments to the unfettered circulation of these works materialized in every instance. Yet also in every instance, generally at great personal cost, the creators and their works prevailed.

3. *Summing Up*

One of the difficulties with the term *censorship* is that it commonly connotes finality. More accurately, *censorious* actions are part of an ongoing process of

action and reaction that demonstrates a variable equilibrium. Cultural products typically display a resilience that enables them to persist, despite meeting substantial obstacles to their successful completion and circulation during particular times or in specific venues (Dubin 1999). Even in extreme situations where the creator is permanently silenced or the product destroyed, history often records the deed. Culture typically outlasts those who would crush it.

Transgressive art is commonly credited with destabilizing taboos (Ben-Levi et al. 1993) and providing a source of innovation in the art world (Graves 1998). Both views undoubtedly have merit. But it is equally true that at the beginning of the new millennium, *transgressive art* has become a norm in itself. In that respect it has become somewhat banal, incorporating shock for the sake of shock. What was on the extreme margins of comprehension and acceptability not so long ago has quickly become absorbed into the mainstream. Artists who seek to be transgressors at present are increasingly hard pressed to uncover virgin territories to explore.

This provides both an opportunity and a challenge to researchers. In the past, social scientists studied a wide variety of transgressions employing a Durkheimian, normative approach. Mary Douglas' work subsequently moved investigators toward a more structural form of analysis, continuing the predilection to analyze such phenomena under the rubric of 'deviance.' But in a *postmodern* world, these earlier frames of reference are not as adequate as they once were.

As a broad range of social breaches has become routine in contemporary society, the role, the import, or even the continued existence of an artistic *avant garde* is in question. Marginal art might no longer be 'where the action is.' Researchers may turn instead to other realms to examine creativity, the defiance of *symbolic boundaries*, and the eventual social responses. The frontiers of innovation are frequently found in the domains of science and technology. New forms of communication, insight, and philosophical exploration are as likely to spring from the Internet as from artistic styles such as Impressionism, a prime example of how art that once posed a dynamic, transgressive challenge to perception has become static, conventionalized.

Future researchers will discover fresh veins of material in the social worlds people create with and around computers. Computers alter discourse (who you contact, where, for what purposes, and with what frequency); identity and perception ('presentation of self' becomes more malleable, 'reality' more manifold); the distribution of ideas and images (the availability of sexually-oriented material); and the process of creation itself (an array of raw material can be 'sampled' and reconstructed in myriad forms). This fluidity of categories denotes and defines our *postmodern* world, and researchers will need to develop new theoretical tools to understand these phenomena. The extent to which artists utilize the potential of new technologies will have a significant impact upon how viable their art remains as a generative force in the twenty-first century, and the degree to which social scientists will continue to study them.

See also: Art: Anthropological Aspects; Art, Sociology of; Cultural Studies: Cultural Concerns; Film: Genres and Genre Theory; Postmodernism in Sociology; Postmodernism: Philosophical Aspects; Television: Genres

Bibliography

Apel D 1997 The tattooed Jew. *New Art Examiner* **10**: 13–17

Ben-Levi J, Hauser C, Jones L C, Taylor S 1993 *Abject Art: Repulsion and Desire in American Art*. Whitney Museum of American Art, New York

Douglas M 1973 [1970] *Natural Symbols: Explorations in Cosmology*. Vintage, New York

Douglas M 1984 [1966] *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*. Routledge & K. Paul, London

Dubin S 1992 *Arresting Images: Impolitic Art and Uncivil Actions*. Routledge, New York

Dubin S 1999 *Displays of Power: Memory and Amnesia in the American Museum*. New York University Press, New York

Dubin S 1997 Pressed to the limit: printers and the problematics of censorship. *Journal of Visual Anthropology* **9**: 229–41

Durkheim E 1964 [1938] *The Rules of Sociological Method*. Free Press, New York

Foucault M 1978 *The History of Sexuality: Vol. 1: An Introduction*. Pantheon, New York

Garber M 1992 *Vested Interests: Cross-Dressing and Cultural Anxiety*. Routledge, New York

Graves L 1998 Transgressive traditions and art definitions. *Journal of Aesthetics and Art Criticism* **1**: 39–48

Jacob M J 1995 *Culture in Action: A Public Art Program of Sculpture Chicago*. Bay Press, Seattle, WA

Jansen S C 1991 *Censorship: The Knot That Binds Power and Knowledge*. Oxford University Press, New York

Kester G H 1998 *Art, Activism, and Oppositionality: Essays from Afterimage*. Duke University Press, Durham, NC

Rose B 1993 Is it art? Orlan and the transgressive act. *Art in America* **2**: 82–7

Tucker M, Tanner M, Goode Bryant L, Dunye C 1994 *Bad Girls*. MIT Press, Cambridge, MA

S. C. Dubin

Censorship in Research and Scholarship

Censorship is the suppression or alteration of speech or writing prior to publication in the interests of an alleged higher social good. The narrowest definition would limit the use of the term to authoritative action,

usually, but not always applied by governments. However, historically, conflicts with prevailing religious doctrine have been a major reason for censorship imposed by clerical authorities. While that remains a major motive in some areas of the modern world, censorship has come to be associated primarily with the actions of governments. The rise of totalitarian governments in the twentieth century produced perhaps the most egregious and systematic programs of censorship in history, both as a consequence of the range of behavior over which such governments claimed sovereignty and of the ability of the modern state to impose and enforce its rules on its citizens. But democratic governments too engage in acts of censorship when their central values appear to be threatened or, more accurately perhaps, when the values of the dominant political or economic forces within them are threatened.

Censorship, as defined above, is frequently associated with patronage, especially in the academic world. Sponsors of research—governments, corporations, foundations, and individuals—are often in a position to close areas of research by not funding it, or to stop or alter the publication of the results of research seen to be inimical to their interests. Patronage, or more precisely, the threat of losing it, may also lead to acts of self-censorship by scholars. As research in the natural and social sciences has become more expensive, these forms of censorship have become increasingly common.

It is important to distinguish censorship, which is always accompanied by the threat of legitimately imposed sanctions, from disagreements that may be accompanied by efforts to persuade others not to read or listen to the offending writing or speech. The two are frequently confused, though the former prevents free speech, while the latter is, itself, an exercise of free speech.

1. *Censorship and Research*

Scholarship and research, by their very nature, have always been vulnerable to censorship. The questions asked, the methods used, the answers found and their implications, may challenge a prevailing orthodoxy, and in the extreme and most dramatic cases, may threaten to undermine the fundamentals of ruling regimes, secular and clerical alike. In democratic societies social science research is unlikely to be seen as threatening to the interests of the state. It is much more likely, however, to be seen as a threat to some group interest in the society and to elicit governmental action through the political supporters of that group. Indeed, since social science research, particularly large scale projects, has come to rely on governmental funding, and since the subjects of social science research often touch on and challenge deeply-held social and religious values, resisting efforts to stifle it may, on occasion, require a degree of legislative self-restraint that democratically elected legislatures cannot be expected to muster. Two especially sensitive areas of research, sex and race, illustrate the point. For an excellent discussion of restraints on research in these areas in the United States, see Hunt *The New Know-Nothings*.

2. *Censorship and Research On Sexual Behavior and Attitudes*

In the United States, religion, politics, and the freedom to do research on the subject of sexual practices and attitudes clash more vividly than perhaps anywhere else. The result can be seen in a history of attempts, frequently successful, by the Congress to forbid studies of sexual behavior financed by public funds. In one instance, the research of Alfred Kinsey, the Congress reached into the private sector in an effort to cut off funding for the work. When Kinsey's two studies, *Sexual Behavior in the Human Male* (1948) and *Sexual Behavior in the Human Female* (1953) were published and captured wide public attention, religious, and political conservatives became alarmed at what they saw as publicizing depraved sexual practices, thereby contributing to the moral degradation of the society. Kinsey's research, however, was not publicly financed. Rather it was funded by grants channeled through the National Research Council by the Rockefeller Foundation. Congressional conservatives used the instrumentality of the Committee to Investigate Tax Exempt Foundations. The committee suggested that foundations that supported research like Kinsey's might well lose their tax exempt status. Whether there was any Constitutional basis for such a threat hardly mattered in the climate of the times. Shortly after the Committee Report was published, the Rockefeller Foundation ended its funding of the research. While the research continued, it did so with far less money, largely earned from the royalties of the books, and, as Hunt reports, '…much of its data went unpublished for many years; the findings of a survey completed in 1970, for instance, were not published until 1989.'

The Kinsey case was extraordinary in the attempt of the Congress to reach the private funding of research. In other respects, however, it was a precursor to future attempts to censor unpopular social science and art. Publicly funded research on sex has been constantly in trouble, and in the 1970s and 1980s, several large-scale surveys of sexual behavior that had been approved by the National Institutes of Health, were stopped when the Congress cut off funds for them by legislation or pressured political leaders in the Executive Branch to reverse NIH's decisions.

The most publicized case of the kind involved an exhibition at the Corcoran Gallery of Art in Washington DC of the work of the photographer Robert Mapplethorp. Included in the exhibition was a

series of explicitly homoerotic pictures. Since the Corcoran Gallery is privately owned and largely privately financed, there was no direct way for outraged religious groups and their Congressional supporters to reach the Gallery directly. However, the exhibit was funded in part by money from the National Endowment for the Arts which, as a government agency was directly reachable both through legislation and appropriations. Both avenues were used. Appropriations were drastically reduced, and legislation imposing 'decency standards' on Endowment grants was passed. The latter was subsequently declared unconstitutional by the courts, but the Endowment, itself, has yet to recover from the episode and from the continuing implacable hostility of the religious and political right wings of American life.

3. *Censorship and Research on Race*

Conservative religious groups can claim credit for restraining research into matters of sexuality. But American liberal and civil rights groups can make a similar claim with respect to research that touches on race, a topic that is guaranteed to elicit emotional controversy. Two sets of events, both taking place at the intersection of race and genetics, illustrate the point. The two are similar in some respects, but also different in ways that show important shadings in evaluations of prior restraint of research.

The alleged correlation between race and intelligence has generated more emotional controversy than perhaps any other topic in the social sciences. Educational psychologist Arthur Jensen set off the debate in 1969 with an article in the Harvard Educational Review, 'How Much Can We Boost IQ and Scholastic Achievement?' Jensen, relying heavily on the literature of studies of identical twins, concluded that there is, in fact, a large component of heritability in intelligence, measured by IQ tests, and that African Americans on average were at a genetic disadvantage. Jensen's avowed purpose, as the title of his article suggested, was to force educators to confront that finding and to devise educational strategies that took it into account.

Jensen's conclusion from the data he reviewed, quite apart from the policy prescriptions that followed from it, was highly debatable in genetic terms, and it was hotly debated. There was no question of censoring his work, though some might have wished to do so if it had been possible. Jensen's work was wholly funded from private sources. He was, however, prevented from speaking on some college campuses by disruptions or the threat of them. If the term 'political correctness' is understood to describe a form of debate in which the object is to anathematize one's adversary in order to eliminate him from the debate on grounds of moral inferiority, then the Jensen episode can be seen as a warning to social scientists who would venture into the area of race, intelligence and genetics that they did so at the risk not only of having their work attacked, which was fair enough, but of being personally vilified as pseudoscholars whose purported scholarship was merely a mask for a racist social agenda.

Twenty-five years later, the Jensen episode was replayed, with even greater intensity following the publication of *The Bell Curve* (Herrnstein and Murray 1994). The book was met by a storm of criticism, much of it scholarly and serious—the vast majority of psychologists and geneticists found their work to be wholly without merit—but some of it consisting of *ad hominum* vilification of the authors.

Neither the Jensen nor the Herrnstein–Murray cases meet the strict definition of censorship given at the outset. Rather, they are examples of what might be called 'virtual censorship' in which the threat of personal vilification may serve to prevent some scholars from addressing topics for which they can be certain that even their hypotheses, much their results, will call down on them painful personal confrontations. There is no way to know whether the Jensen and Herrnstein–Murray examples had that effect on others. However, the case of the aborted conference on Genetics and Crime demonstrates that the governmental equivalent of personal vilification, the threat of political retribution, can have that effect on even the best government agencies.

4. *Genetics Race and Violence*

On May 1, 1991, the National Institutes of Health awarded a grant for a conference on 'Genetic Factors in Crime: Findings, Uses and Implications.' The award was based on the recommendation of a peer review group, which had concluded the organizers had done a 'superb job of assessing the underlying scientific, legal, ethical, and public policy issues and organizing them in a thoughtful fashion.' Ten weeks later funds for the conference were frozen when NIH Director, Dr. Bernadine P. Healy, concluded that the review group's recommendation had not been strong enough for the meeting to go forward in the face of public criticism.

The public criticism that concerned Dr. Healy came from a coalition of psychiatrists and civil rights groups that argued that government sponsorship of the meeting was yet another step toward legalization of the racist idea that violence is genetically based. The Center for the Study of Psychiatry asserted there are no known genetic factors in crime, but merely raising the issue arouses racial prejudices and distracts us from the true causes of the high crime rate within the inner city, including poverty, unemployment, racism, dangerous and inadequate schools, drug abuse, family dysfunction, and a variety of other social and economic factors. And as one opponent put the more extreme argument, 'It is clear racism. It is an effort to

use public money for a genocidal effort against African Americans.'

The cause was taken up by the Congressional Black Caucus and the problem facing the management of NIH became not the merits of the project or the protection of the peer review process which the NIH pioneered and developed, but the protection of the agency from the fallout of a political storm. The decision was not surprising. From the point of view of agency managers, political embarrassment leads to political problems, putting at risk agency interests, programs and careers that far outweigh the value of a single project, no matter how meritorious.

5. *Industry and Censorship*

Because social and behavioral research is not heavily funded by industry, it is unlikely that it will be heavily affected by the problems that have accompanied industry's support of scientific research, especially research in the biomedical sciences. The restraints imposed by industrial sponsors are typically motivated by concern for the protection of intellectual property, and most commonly they take the form of required delays in publication for a specified period of time to allow for review by the sponsor for intellectual property implications. There have been occasions, however, in which corporations have gone further and prohibited publication because the results of the research would be publicly embarrassing to the company, a situation more analogous to that sometimes faced by social scientists.

A potentially risk does face social scientists, even though they are not heavy recipients of corporate support. It is an indirect but real risk, especially for scholars in universities. Universities are often willing to protect their scholars from publication threatened by the government. However, if they are simultaneously accepting limitations on publication as a part of industrial contracts, their ability to hold a principled position against the government is seriously compromised. Indeed, during the 1980s and early 1990s, when it was the policy of the American government to use the Export Control Act to restrict the international dissemination of unclassified but allegedly sensitive research findings, it was common for government officials to argue that universities were willing to do for money what they would not do to protect the national interest. It is an argument that resonates powerfully with legislatures and with the public, and once it is accepted it applies to any research that can be invested with a claim of national interest. Since that is an open-ended category, the willingness of universities to accept prior restraints on publication as the price they must pay for industrial money could greatly compromise their ability, or even willingness, to defend social scientists whose work is challenged.

It is in the nature of governments and other important social interests to protect themselves from challenge or embarrassment. It is in the nature of social science to provide both from time to time. Tension between the two is, therefore, unavoidable. The main protection for social scientists in protecting their right to publish freely lies in their willingness to join together to assert it when threatened and the willingness of their employing institutions, individually and collectively, to join with them.

See also: Censorship and Secrecy: Legal Perspectives; Censorship and Transgressive Art; Civil Liberties and Human Rights; Civil Rights; Ethics and Values; Freedom/Liberty: Impact on the Social Sciences; Freedom of the Press; Fundamental Rights and Constitutional Guarantees

Bibliography

Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. The Free Press, New York

Hunt M *The New Know-Nothings: The Political Foes of the Scientific Study of Human Nature*. Transaction Publishers, New Brunswick, NJ

Jensen A 1969 *Harvard Educational Review* **39**: 1–121

R. M. Rosenzweig

Censuses: Comparative International Aspects

Over 200 countries have taken censuses within the last ten years. Because there is a desire to compare data from one country to another, the United Nations has issued guidelines for the taking of censuses. Each country may have its own way of dealing with special populations, but if the methods are documented, comparisons can be made. Countries differ in whether or not they use mail primarily or rely on census-takers to visit each unit and record the information. Many countries use sampling as part of the census operation, as part of the processing to get an early look at results, as part of an evaluation, or in an effort to reduce the burden on households by not asking every household to complete every question. Because of more recent concerns about privacy and confidentiality, some countries have challenged the use of a census and are looking at new methods of obtaining population data.

1. *Distinguishing Features of a Census*

National censuses, drawing on the sponsorship and authority of national governments typically provide information for national needs such as population counts used to reapportion legislatures. National

censuses also serve local needs, such as providing population data for state and local agencies for planning for health, education, and many other services.

Because local populations may grow dramatically between censuses, many cities and states or provinces take local censuses to benefit from more up-to-date population figures. The emphasis in this article is on national censuses and their comparability. Before a census takes place, there must be agreement about what geographic areas are to be included. If new territory has been annexed since the last census, it must be clearly defined. Geographical subdivision as well as national boundaries must be clearly delineated and changes from the last census must be clearly stated.

While censuses typically strive for universality, in every country some groups are excluded. Such groups must be decided upon before the census. For example, the population residing in an embassy of a foreign country may be exempt from the census of the country in which the embassy is located. People who are living in the country as temporary workers may be exempt, though it is often a function of the amount of time they spend in the country. Military personnel and their families living outside of the country may or may not be counted. (In the USA, they have been counted in some censuses and not in others.) This is different in concept from early censuses when whole groups of citizens or residents were left out. (See *Censuses: History and Methods* for examples of censuses that counted only adult men.)

Though not essential, it has become a custom in most countries to tie the census of population to a census of housing, expanding analytical capability. However, it is also important to make provision for those persons who do not live in regular housing, such as the homeless or nomads. Provision must be made for squatters' colonies, refugee camps, defense areas, and the institutional population.

Modern censuses rely on enumeration of the individual, although often through households or institutions. This implies that data about each individual is collected and recorded, enabling cross-classification by individual characteristics. In some early censuses, or in some enumeration of hard-to-enumerate groups, aggregates from the group have been recorded, thereby prohibiting cross-classifications. In the USA, the first census of 1790 was of this type, collecting data on the number of males, number of females, but no individual data (Anderson 1988).

Another feature of a census is that it takes place at the same time, or with reference to a specific census day, at all places in the country. To get an accurate snapshot of the country, censuses must take into account all births, deaths, and immigrants as of a certain day, usually denoted as Census Day. (Weather sometimes dictates the physical enumeration of some remote areas of the country at a separate time.) Periodicity of censuses ensures that trends can be followed. Also, comparisons among countries are more meaningful if they refer to the same time period. The United Nations has been instrumental in trying to set a schedule for censuses. The final step in producing a census is the compilation and publication of the data as soon as possible. Without dissemination, a census cannot be used effectively.

2. *Coverage of Population*

One of the biggest problems in comparing censuses across countries is the difference between a *de jure* and a *de facto* enumeration. A *de jure* census counts people at their usual place of residence, whether or not they were there on Census Day. The USA, for example, defines an individual's 'usual' residence as the place where one spends the majority of one's time. A *de facto* census counts people where they were present at the time of the census. In practice, most countries combine one method with some elements of the other. *De jure* residence may not coincide with legal residence. The United Nations (1992) recommends a combination of the two methods. For example, the 1980 census in Brazil was described as both *de jure* and *de facto* (Goyer and Domschke 1983).

Using the *de jure* concept means deciding, in advance of enumeration, where to count people who have more than one residence, persons who work at one location and return to their usual residence for the weekend, students at boarding school or college, members of the military who live on base but maintain a residence off base, and other such cases. Comparisons between countries that use *de jure* and *de facto* counting rules are difficult, as are those between countries using different *de jure* counting rules. Documentation of how and where people are counted is the best hope for making correct comparisons.

Another factor to be considered when preparing for a census is how to count certain special populations. A group that is difficult to count in many developed countries is the homeless. Though efforts are made to count them where they receive services such as food or shelter, there is always the chance that they are not counted or counted more than once. In some countries, the counting of nomads presents difficulties. Many approaches have been tried, such as a tribal approach in which a chief gives the information, or a water point approach in which all nomads are enumerated at water points. The treatment of civilian aliens who work in a country as seasonal workers, who cross a frontier daily to work in a country, or who are residing in a country for some work-related reason is now an important consideration in many European countries as well as the USA. The differential effect of how countries treat the special populations has an important effect on the estimated size of the world's population (UN Statistical Office 1992).

Influencing the coverage of the census is the time of the year when it is conducted. Most countries opt for a time of the year when travel is not impeded by bad weather and when people are at their usual place of residence. This may affect the principle of census taking that calls for simultaneous data collection. In the USA, Alaska is enumerated at a different time from the continental states. If some areas are impossible to reach, coverage will be biased downward. Another time-related factor is whether the enumeration takes place early in the month or mid-month. Canada has found that mid-month is helpful in the enumeration of those people who rely on a monthly check, move out of short-term housing until the next check appears, live on the streets at the end of one month and the beginning of the next, but are usually in housing at mid-month. Tests in the USA confirmed that a mid-month start would be useful there as well, and changing from the traditional April 1 Census Day to mid-March would also provide additional time in which to conduct the census (Cohen et. al 1999).

In conducting the census, several reference periods are used. For example, questions about income usually refer to income in the previous year. Questions about labor force usually refer to the week previous to Census Day, whereas questions about who was living in a household, usually refers to the night before Census Day. Having multiple reference periods can often be confusing to respondents and may affect the quality of data. The length of the enumeration period also affects coverage. The shorter the length, the more likely the census will avoid over- or undercounting. In large countries, the time span may be longer, but the extended period often leads to overcounting. Suharto (1995) points out that no country that used a *de jure* method in censuses between 1985–1994 was able to complete the census in a day, while nine countries that used a *de facto* method completed the census in one day. Though a one-day census avoids the complexity of people moving, coverage suffers. From a survey of countries that took censuses between 1985 and 1994, Suharto (1995) reported that 31 countries completed enumeration in a week. However, countries such as Colombia and the USA took well over two months. As the time period of the enumeration lengthens, coverage errors increase.

Privacy and confidentiality concerns also affect coverage (Suharto 1995). Within the last 20 years, in several countries, concerns about confidentiality have led to decreased census participation. Because census data are used for so many purposes, and public use files are now widely disseminated, the risk that individual census records may be inadvertently disclosed has raised public concern in many countries. A strict provision of census confidentiality is necessary to reassure people concerned about issues of privacy. This feeds into the need for an effective public relations program in which the legal protection of individual census information is stressed.

3. *Basic Content*

To be able to compare countries by census characteristics, countries need to collect data on a basic list of characteristics with common definitions. The United Nations has been very instrumental in developing a list of basic content items with their definitions. Though there is not total agreement on every item or definition, the ability to make comparisons on basic characteristics has been enhanced.

An item of great interest to many countries is the percentage of people who live in urban or rural areas. These data are regarded as an aspect of the developmental stage of a country. To be able to make cross-country comparisons, the definitions must be the same. However, there is no universal definition of urban. In some countries, it depends on population size, sometimes on the presence of a city, on whether an area has urban characteristics, or on whether agriculture is the predominant economic activity. All other areas not urban are defined as rural.

Another item of interest is that of households and families. In some countries, the terms household and family are used interchangeably. In others, however, household means one or more persons living in a dwelling, whether or not related by blood or marriage, while family is two or more persons related by blood or marriage. Thus, there can be more than one family in a household. Of interest in many international comparisons are numbers of households and families, types of households, number of one-person households, and family sizes (Goyer and Domschke 1983).

Other common characteristics for which cross-country comparisons are frequently made are: sex, age, race, ethnicity, marital status, educational attainment, literacy, religion, languages spoken, citizenship, employment status, occupation, income, and fertility. Some countries will not ask certain questions. For example, the USA will not ask a question about religion. Some questions cause uneasiness and may lead to coverage problems. Questions on citizenship in a country that has many undocumented workers may not be desirable (see *Censuses: History and Methods*).

4. *Organizational Structure*

In the more developed countries, there is a permanent census office. In other countries, a census office may be temporary for the period of the census. A permanent office offers many advantages, because what was learned in one census can be factored into the plans for the next. There are more likely to be staff in place who have census experience and who are aware of census concepts, definitions, and planning. In many countries, censuses are viewed as a statistical activity, carried out by an agency that handles statistical activities. In some countries, the census is carried out

by a non-statistical agency, perhaps even the military. A country is more likely to participate in UN statistical activities and adopt UN recommendations if it has a statistical organization running the census.

5. *Type of Enumeration*

There are two main types of census enumeration. In the canvasser approach a census enumerator interviews the population in an enumeration area and records the responses on a census questionnaire. In the self-enumeration approach, the questionnaires are distributed (in advance) and a member of the household records the information about the household and its members on the questionnaire.

There are advantages and disadvantages to each method. In a country where the literacy rate is low, the canvasser approach is necessary. In the USA, the self-enumeration method is used, supplemented by the canvassing of nonrespondent households and special populations. Tests have shown that accuracy is improved in a self-enumeration approach because the household respondents are more likely to report what they know about themselves rather than being influenced by the responses from other households. In tests for the 1960 census, the US Census Bureau found that interviewers were influenced by what they thought the responses should be instead of what the facts showed. Suharto (1995) reported that in the 1985–1994 period, about three-quarters of the countries that took censuses used the canvassing approach.

6. *Use of Sampling and Administrative Records*

Sampling and administrative records are each efforts to supplement and improve upon enumeration. Sampling has been used for many years, and in a variety of contexts. One common way is to ask most of the population a minimum set of items and then ask a much larger set of questions of a sample of the population. In countries where this is done, the questionnaires are usually known as the long form and the short form. The long form has been under discussion in the USA as a possible deterrent to response. Though that view is not universally held, efforts are underway to replace the long form by more continuous survey activity, which would also have the benefit of providing more up-to-date data to local areas (Edmonston and Schultze 1995). Another use of sampling may occur during processing. It may be important to get an early indication of some census results, so a sample of questionnaires may be selected, coded for occupation, industry, and other items, and then only this sample is processed. This use of sampling may still be of special value in countries with limited computer power. Sampling is also used in the quality control of the data processing. Similarly, in evaluation studies samples are often selected for coverage or content studies. Sampling has not yet been used for the follow-up of people who did not return the census questionnaire, though it was proposed in the USA for 2000. (See Cohen et al. 1999 for a description of the sampling approach the US Bureau of the Census originally recommended for 2000.)

Many people have suggested using administrative records as the basis for a census, and, indeed, administrative records in the form of a population register, augmented by housing registers is a replacement for a census in some European countries. However, there are other ways that administrative records can be used to improve the quality of data collected, add to the data, or evaluate the data. The US Census Bureau has long used a limited set of administrative records to evaluate its census data (Edmonston and Schultze 1995). One of the experiments carried out, as part of the 2000 census in the US was a test of whether the merging of several administrative record files would produce high-quality short-form information. The results of this test will be available in 2002.

7. *Evaluation*

In comparing censuses across countries, it is important to know about their quality. Many countries carry out Post-Enumeration Surveys to estimate the coverage of the census, or checks with administrative records or additional interviews to estimate the accuracy of the data for specific items. These evaluations can be helpful in changing questionnaire items for future censuses. For example, when people are asked how old they are the responses tend to 'heap' at the terminal digits of 0 and 5. Demographers refer to this phenomenon as age heaping. When the question is about date of birth, less age heaping results. The results of evaluation studies also facilitate comparisons of fertility, disability, and language spoken in the home. Estimates of coverage by major demographic groups within countries can help in the interpretation of differences in rates of different kinds such as mortality rates, prevalence rates, home ownership rates, crime victimization rates, and many others.

8. *Census Processing and Dissemination*

The entire census data-processing system needs to have been decided upon and tested before enumeration. Overexpenditure of census funds on data collection has sometimes affected the availability of funds to process the data. Occasionally a sample of the

questionnaires is selected for coding and processing to facilitate early dissemination. The form and extent of the tabulations is also important and affects international comparability. It is important to have a plan for the release of census results to the government, to other data users, and to the general public. National results provide data for government policy and action. They also permit international comparisons and analysis of trends. Small area data are the planning tools used by local governments and businesses. Academic researchers analyze census data at each level and often find different results from government analysts, as well as puzzling results requiring deeper investigation. To facilitate broad analyses by diverse analysts, many countries have developed the tradition of releasing public-use samples of census records. This use of sampling (from the full census file) also helps address concerns regarding confidentiality once identifiers are removed. Currently, data are disseminated in many different forms. Most countries have traditionally published printed volumes of census data, but many have also prepared computer files of use to researchers, and a number of countries have begun to release tabulations and sample data on CD-ROM and in the form of files accessible via the Internet.

9. *Alternatives to a Census*

Though over 200 countries have taken censuses since 1990, several countries have recently abandoned census taking because of lack of public cooperation or vocal opposition, relying instead on other methods such as population registers, administrative records, or a combination of administrative records and household surveys to supply data needs. Also, the escalating cost of censuses has led to a consideration in many countries of other ways of satisfying data needs. The accuracy of such alternative forms on national data collection needs careful assessment, and is of special concern for international comparisons.

See also: Censuses: Demographic Issues; Policy Knowledge: Census; Population Cycles, Formal Theory of; Population Forecasts; Statistical Systems: Censuses of Population

Bibliography

Anderson M 1988 *The American Census*. Yale University Press, New Haven, CT

Cohen M L, White A A, Rust K F (eds.) 1999 *Measuring a Changing Nation: Modern Methods for the 2000 Census*. National Academy Press, Washington, DC

Edmonston B, Schultze C (eds.) 1995 *Modernizing the US Census*, National Academy Press, Washington, DC

Goyer D S, Domschke E 1983 *The Handbook of National Population Censuses, Latin America and the Caribbean, North America, and Oceania*. Greenwood Press, Westport, CT

Suharto S 1995 *Emerging Issues Related to World Population and Housing Census Program*, United Nations Statistics Division, Technical Notes

UN Statistical Office 1992 *Handbook of Population and Housing Censuses. Part 1. Planning, Organization and Administration of Population and Housing Censuses*. Department of Economic and Social Development. Statistical Office, New York

B. A. Bailar

Censuses: Demographic Issues

Censuses are designed to provide population counts for countries or areas within a country together with a range of demographic, social, and economic data on the population. Basic demographic data provide information on the size, distribution, structure (particularly age and sex), and change of populations; more broadly, demographic data systems also collect information on racial/ethnic characteristics, socioeconomic characteristics, and health (see *Demography: Twentieth-century History*). As a complete count of the population, a census differs from other demographic data systems, but serves to complement them. This article describes the essential features of a census in contrast with other demographic data systems and discusses the quality of demographic data collected in censuses, with a particular focus on age data.

1. *Characteristics of a Census*

A census of population is a complete counting of the number of persons in a country (or region) at a given point in time. Because of the amount of detail available in a census, both demographic and geographic, a well-conducted census forms the core of a modern governmental statistical or demographic data system. In particular, it provides population totals, geographic distributions, and characteristic data that can be used as benchmarks or reference points for other demographic and socioeconomic data.

The essential features of a census include reference to a well-defined territory and a specific point in time, or reference date. The census is designed to enumerate every person within the designated territory at the specific reference date and to collect information on their characteristics at that time. As ideal types, censuses may be either *de facto* or *de jure*, counting all the people present in the territory on the census date or all the people who 'usually' or legally live there

(Shryock and Siegel 1980, p. 92). Since these ideal types are difficult to implement, most censuses, in practice, are a mix of the two. At a national level, the most problematic groups to define and count tend to be: military, naval, and diplomatic personnel (both foreign personnel in the country and native personnel abroad); indigenous populations and nomadic groups; civilian nationals temporarily abroad; and civilian aliens temporarily in the country (including refugees and displaced persons). The United States census, for example, counts most people at their usual place of residence, a *de jure* concept, includes some persons temporarily in the country, such as foreign students, but excludes others such as foreign tourists (see *Censuses: Comparative International Aspects*).

For countries that either receive large numbers of temporary and permanent migrants or send large numbers, there may be a substantial difference between the *de facto* and *de jure* concepts. The alternative population concepts can be even more problematic for geographic subdivisions within a country than for the entire country. In addition to the groups previously listed, persons with multiple residences (e.g., vacation or seasonal homes, migrant workers, or persons with multiple job sites) may be difficult to assign properly to a specific subnational area. Defining the nature of a 'temporary' stay away from home can create difficulties in assigning residence for a variety of populations; for example, boarding students or residential college students may never return to their parental home but may be counted at either site (and, in fact, may often be erroneously counted at both sites). The specific set of residence rules used in a census can have a significant impact on the measured population size for areas with large numbers of these problematic groups, areas such as university towns or resort areas.

Although the principal purpose of censuses is to count the population, their main utility is generally the data collected on the characteristics of the population. The following sections of this article discuss the types of data collected in censuses and alternative data collection methods. Then the article addresses the quality of census data, the role of demographic methods in assessing data quality, and, more specifically, data on age and sex.

2. Census Content

Censuses collect a range of information on individuals and households (or families). In a census, the most basic geographic and demographic data are collected for *every individual* in the population, generally: location (or usual residence), age, sex, and relationship to head of household or family. Because of its broad population coverage, a census may be the only source for detailed demographic and socioeconomic information about subnational areas, especially small ones, or for smaller subgroups of the population, such as ethnic or immigrant populations. Such detailed information is generally unobtainable in other general-purpose demographic data collection systems.

The major expense, both in time and money, in conducting a census is the set of activities involved in compiling a list of addresses or dwelling units and actually contacting each household. Thus, the marginal cost of collecting additional demographic, social, and economic data is generally not great (e.g., Edmonston and Schultze 1995, p. 46). Further, with modern sampling techniques, sufficient precision can often be obtained by embedding a sample within the census so that it is unnecessary to burden the entire population with a lengthy set of questions. The choice of whether to collect information from the entire population or a sample is determined by several factors, including legal requirements for the data, the desired degree of precision, and the cost of data collection.

Data collected in a census provide detailed measures of the demographic make-up of the population. Typical measures include: marital status, birthplace, citizenship, children ever born (or some measure of fertility), and measures of nationality, ethnicity, or race. Most modern censuses collect a range of other information on socioeconomic status, including: education (educational attainment, school enrollment, literacy), economic activity (occupation, industry, type of activity, income, commuting activity), and other social or political characteristics (language, religious affiliation or ethnicity, living arrangements).

3. Alternative, Complementary Demographic Data Collection Systems

A census is the backbone of a national statistical system, but it is only one part of a complete system. Surveys and registrations systems fill the data gaps left by a census; each system can provide data to check the completeness and accuracy of the others. The distinction between a census and a survey is often a matter of degree. One key difference is that censuses are usually designed to determine the size of the population of an area; surveys rarely do so. Sample surveys provide the same types of demographic, social, and economic data as collected in censuses. However, because of the generally smaller scale of a survey and because it is usually administered by trained and experienced interviewers, the survey can delve more deeply into various subjects than a census. The main tradeoffs involve the greater detail (categorical and geographic) and the greater precision provided by a census.

Registration systems, the third part of a demographic data system, are generally designed to count vital events: births, deaths, marriages, and entries and exits at international boundaries. They also include

population registers of the entire population or segments of the population; examples of the latter include persons of draft age, workers covered by social insurance or government health systems, voters. Population registers, when active records are tallied, can share many features of a census: population coverage, geographic coverage, and content (Bogue et al. 1993, Sect. 3, p. 2). Registers differ in that the data are generally not collected simultaneously or with reference to a specific point in time. When a register is continuously updated (e.g., to take into account departures through death or emigration, additions through birth, immigration, or aging, and changes in status or residence), this discrepancy in timing is removed. While universal population registers are rare, partial registers are very common. In fact, comparisons of census counts with partial registers can provide critical checks on the coverage and accuracy of a census.

Vital registration systems count events, not people, yet they are essential for providing information on the demographic dynamics of a population. Birth and death rates are generally computed with vital registration data as the numerators and population data from a census as the denominators. Cumulative counts from birth, death, and immigration registration systems can be used to reconstruct the demographic history of a country and check the accuracy of census data or the consistency of censuses across time. Since the data in vital rates and census evaluations come from different collections systems (different registration systems, censuses, and surveys), such comparisons must be constructed carefully to maintain consistency with regard to definition of population groups, time references, and territorial coverage.

4. *Accuracy of Censuses*

Censuses are widely used to allocate political power, as a tool for allocating governmental resources, and a plethora of private uses. Although improvements in methods and technology have led to concomitant improvements in the quality of census data, the range of uses has sparked concerns over both the completeness of census counts and the accuracy of the information collected (Anderson and Fienberg 1999). Techniques for evaluation of completeness and accuracy can be broadly categorized as statistical in nature or demographic. The completeness of a census, or its coverage, is distinct from the accuracy of the data items collected in the census and involves both undercounts and overcounts. Undercounts or census omissions arise from failures to include persons or households in the census counts. Overcounts can arise from duplications or multiple counts of the same person, from fabrications, and from erroneous enumerations or counts of persons who are not properly part of the census universe. Problems with proper reporting of residence can present special problems both for census accuracy and for census evaluations. Persons with no usual place of residence (such as homeless populations or migratory workers), persons with multiple residences, and persons away from their families or homes of record (such as college students, members of the armed forces) are particularly prone to being omitted, double counted, or counted in the wrong place, which results in a double error—an undercount at the proper location and an overcount at the wrong location. The numerical difference between the gross undercount and gross overcount is termed the *net* undercount of a census.

Statistical techniques for measuring census completeness and accuracy generally involve matching individual census records with a sample of individual records from another data system. Other systems used for evaluation include administrative records from registration systems, previous censuses, periodic sample surveys, and specially designed coverage evaluation (or postenumeration) surveys. Statistical evaluations offer the potential for measuring coverage according to a wide range of characteristics and for measuring both gross undercounts and gross overcounts.

Demographic techniques of census evaluation involve comparison of aggregate statistics or consistency checks, either internal or with external data sources. Because demographic evaluations do not check individual records, they measure *net* undercounts (or overcounts) rather than the gross components of census error. Most demographic evaluations rely on a version of the demographic balancing equation (Shryock and Siegel 1980, pp. 105–11). This simple equation relates the population at a point in time to the population at a previous time and the demographic components of change during the intervening interval: Population at the final point (P_1) equals

Population at the initial time (P_0) plus
Births during the interval (B) minus
Deaths during the interval (D) plus
Immigrants during the interval (I) minus
Emigrants during the interval (E),

Or

$$P_1 = P_0 + B - D + I - E$$

This equation can be applied to the total population of an entire country, an area within the country, age–sex groups (where it can be called cohort analysis), racial/ethnic populations, and other subgroups of the population. Since no demographic data are completely accurate, correction of errors in the measures of the components or assessment of the potential error in the resulting estimate is an essential part of demographic analysis of census errors.

The United States Census Bureau has applied demographic evaluation techniques to measuring coverage of recent censuses with a program called Demo-

Table 1
Demographic estimates of percent net census undercount, by race, for the United States: 1940–90

Race	1990	1980	1970	1960	1950	1940
Total	1.8	1.2	2.7	3.1	4.1	5.4
Black	5.7	4.5	6.5	6.6	7.5	8.4
Not Black	1.3	0.8	2.2	2.7	3.8	5.0
Black–Not Black difference	4.4	3.7	4.3	3.9	3.6	3.4

Source: Robinson et al. 1993, Table 2.

graphic Analysis (DA). The DA estimates draw on a variety of demographic data to construct estimates of the population by age, sex, and race—historical statistics on births beginning in 1925 (corrected for under-registration as measured in a series of registration tests), historical statistics on deaths and legal immigration since 1935, administrative data from Medicare (the national health insurance system for the elderly, also corrected for underregistration), and demographic estimates of emigration and unauthorized immigration. Comparing the DA estimates with census counts provides a consistent set of coverage estimates by age, sex, and race for the censuses of 1940 through 1990 (Robinson et al. 1993); the resulting estimates are shown in Table 1. The DA estimates show a steady pattern of improvement in census coverage from 1940 (5.4 percent net undercount) through 1980 (1.2 percent) but a worsening of coverage in 1990 (1.8 percent). The same pattern of change is apparent for both the minority Black population and the balance of the US population (largely the majority White population). However, the difference in undercoverage between the Black and non-Black population shows no such trend; in fact the 4.4 percentage point difference in coverage in the 1990 Census is the highest shown.

The demographic estimates of coverage for the series of US censuses have proved to be roughly consistent with measures derived from coverage evaluation surveys, for example, a 1990 survey estimated the net undercount at 1.6 percent, only 0.2 percentage points different from the demographic estimate (Hogan and Robinson 1993). The survey estimates, like the DA estimates, showed higher undercount rates for the Black population; further, the survey showed high undercount rates for other minorities populations (Hispanics, Asians, and American Indians) that could not be separately measured with demographic techniques.

The experience of the 1990 US Census illustrates both the strengths and potential weaknesses of demographic estimation techniques. Because of the reliance on historical data series, the demographic estimates permit a degree of precision in comparisons of change in coverage across censuses that is hard to achieve with coverage evaluation surveys, for example, sampling variability, other measurement errors, and changes in survey design may limit comparisons across censuses, particularly those separated by several decades. For the demographic estimates across censuses, however, the degree of variability and potential biases tend to be consistent because the estimates are developed with consistent data, methods, and assumptions. The survey estimates are, however, more flexible in that they can provide many measures that are either not available from demographic methods or are subject to potentially large estimation errors; for example, coverage evaluation surveys can measure gross overcounts and undercounts, coverage error for racial and ethnic groups for which historical data may not be available, coverage of subnational geographic areas for which the required demographic measures of internal migration are not sufficiently precise, and other population groups for which the requisite demographic data may be completely missing or highly inadequate (such as owners and renters, family types, native and foreign-born populations).

The accuracy of demographic methods for evaluating census coverage ultimately depends on the availability of data and the degree of error and uncertainty in the vital statistics and other data from noncensus sources. A variety of techniques for assessing the variability in such demographic estimates have been proposed (Robinson et al. 1993, Anderson et al. 2000), but none is widely accepted or applied. Further, some methods for explicitly combining demographic techniques with survey results have been proposed to improve the accuracy of coverage measurements (e.g., Bell 1993).

A full evaluation of coverage with demographic techniques may not be achievable in many countries because of a lack of the requisite long historical series of demographic data plus supplementary data sources, such as those used for the DA estimates in the United States. However, demographic techniques can provide a variety of consistency checks that yield measures of the relative coverage of the census to an external standard or the relative coverage of different subpopulations within the same census rather than absolute measures of census net undercount. For example, analyzing the change in population between two censuses either in absolute terms with the balancing equation or as a rate of change provides an indication of the relative coverage of the two censuses. Likewise,

sex ratio analysis (described below) provides a measure of relative coverage of men and women in different age groups. Checks of census totals or subtotals against independent aggregates such as birth statistics, partial registers (e.g., draft registrations, national health system enrollments), or universal registers requires correcting the aggregate data for underenrollment or definitional differences from the census. If full and accurate corrections can be made, such comparisons do provide measures of net undercoverage of the census.

5. *Age, Sex, and Race/Ethnicity Data in Censuses*

Perhaps the most crucial demographic data collected in a census are age and sex; consequently, they are almost universally collected in censuses and surveys. Much of the science of demography is devoted to analyzing, accounting for, or controlling for the effects of age structure on other social and economic phenomena, so that age is a factor, either explicit or implicit, in almost every demographic analysis. Differences in age composition between populations can account for many other observed differences, either in whole or in part, including differences in fertility, mortality, sex ratios, dependency, household structure, and even employment rates.

The primacy of age data requires special attention to quality, measurement of errors, and correction. While errors of coverage and reporting occur across the entire age spectrum, studies of age reporting from many countries (Ewbank 1981, Spiegelman 1968, p. 59) have documented several specific types of errors in censuses from many countries:

(a) underenumeration of specific age groups, especially young children and young adults;

(b) age misstatement around 'threshold' ages (age of majority, age of retirement);

(c) age overstatement at older ages;

(d) preference for ages ending in certain digits;

(e) failure to report age.

Whereas underenumeration is a general problem with census data, age-specific underenumeration affects the overall quality of these essential data used in other analyses. Problems in some age groups can be traced to specific types of enumeration issues. Undercounts of very young children seem to be widespread and appear to have a variety of causes, including oversight on the part of adult respondents, devaluation of female children in some societies, and overstatement of age. The other age group that seems to suffer generally from undercounts is young adults (roughly ages 15–30). These are the ages in most countries when individuals enter adulthood, marry, form new households, and leave their parental homes. The related geographic mobility and other changes of residence can easily lead to enumeration difficulties and age-selective omissions from the census count.

Age misstatement takes a variety of forms. Advantages accruing to individuals for attaining certain threshold ages can lead to reporting of these ages on the part of persons younger than the critical age because they perceive reporting older ages to be advantageous. While these ages may vary from country to country, the age of majority (often 18 or 21) and age of retirement (often 65) often show striking patterns of overstatement. At the extreme upper end of the age distribution, especially at ages 100 and over, there is a marked tendency for individuals to report exaggerated ages; this tendency is more prevalent in countries with poorer birth registration systems and in societies and groups with lower levels of literacy.

A particularly striking type of age reporting error is the tendency of individuals to prefer reporting ages ending in certain digits; this tendency is generally exhibited as a preference for ages ending in '0' or '5.' Other types of preferences can also be found including a tendency to report even-numbered ages and a tendency to prefer reporting years of birth ending in '0' or '5' which may result in different patterns of age reporting, depending on the date of the census. Because digit preference can be severe and can have serious ramifications for the use of census data, a variety of techniques have been developed to measure the phenomenon and to correct the data. The principal correction techniques involve distributing the population across age groups according to historical data on births or, more generally, 'smoothing' the age distribution while maintaining totals for certain age groups with graduation techniques developed mainly by actuaries (Shryock and Siegel 1980, pp. 201–29).

In contrast to age, classification of individuals according to sex presents no general problems in censuses. Analytic techniques are generally straightforward, relying principally on the sex ratio, conventionally defined as the number of males per 100 females. Nonetheless, analysis of sex ratios, especially by age group, can be a very powerful evaluative and analytic tool. Many factors can affect the sex ratio of a population, but the principal ones are age structure and migration patterns. The sex ratio of births falls in a fairly narrow range from 101 to 107 and is biologically determined. Differential mortality of the sexes causes this ratio to change as a birth cohort ages. In developed countries, male mortality generally exceeds female mortality at every age so that the sex ratio decreases smoothly as a birth cohort ages.

In the absence of migration, the sex ratio reaches 100 by young adulthood. Above age 50, the sex ratio decreases more sharply with age, reflecting again the higher mortality of males. As a result of these mortality patterns, older populations tend to have lower sex ratios than younger populations.

Departures from this general pattern can be indicative of underlying demographic, social, or econ-

omic phenomena or anomalies. In some countries, very high sex ratios among young children (i.e., high proportions of boys) are thought to indicate sex-selective abortions, infanticide, neglect, or concealment of children. Higher sex ratios among young adults may indicate high levels of maternal mortality. In the United States, lower than expected sex ratios among adults aged 25–54 are indications of poorer census coverage among males than females. Low sex ratios can also result from excess male mortality in wartime; this effect will persist for the lifetime of the affected cohorts. Migration patterns can alter sex ratios if migration is sex-selective, as it often is. Areas with employment concentrations in extractive industries (agriculture, mining), heavy manufacturing, or military often have high sex ratios as men, more than women, migrate into these areas in search of jobs. On the other hand, areas with a high concentration of jobs in 'white collar' occupations and 'office work' often have low sex ratios as more women than men are employed in these jobs.

The age and sex structure of a population encapsulates its demographic history. Cohorts are born, age, and die with predictable and discernable patterns; migrants into and out of an area alter the composition. A census—a snapshot of the population at a given time—provides invaluable information on this history. Analysis of the age–sex data, either alone or in combination with data from other demographic data systems such as surveys, registration systems, and previous censuses, can tell a great deal about the quality of the census data and the nature of the population.

Data on race or other similar identifiers such as ethnicity, nationality, or religion are collected in many, but not all censuses. While race data can be extraordinarily useful in demographic analyses and are often critical for applications of census data, the quality of these data can be quite variable. Groupings based on race (or similar identifiers) are often perceived in the popular mind as equivalent to demographically distinct populations; that is, individuals enter or leave a race group only through the conventional demographic processes of birth, death, and migration. Indeed, DA as applied to race groups in the United States treats the Black and not Black populations in just this manner; and for the 1940–90 results (described above), this treatment appears quite appropriate. However, in many modern societies, particularly multiethnic ones, racial identification and ethnicity are increasingly becoming matters of personal choice and self-identification rather than of societal ascription (Waters 1990). Consequently, individual membership in or identification with a racial/ethnic group may change over time. This phenomenon introduces another 'component of change' into demographic measures of the size of a population group. For example, because of the large numbers of individuals in the United States with some American-Indian ancestry and changes the salience of such ancestry to individual self-identification, more than half of the growth of the American-Indian population can be attributed to inconsistent measurement of this racial group across censuses (Passel 1996).

High incidences of intermarriage, changes in data collection methods, and changing societal norms all contribute to difficulties in consistently measuring racial/ethnic populations over time and across different data systems. The 2000 Census of the United States encompasses virtually all of the potential pitfalls in measuring race. For the first time, respondents to the US census are allowed (even encouraged) to respond with more than one racial identification. While this procedure offers a detailed picture of the population in 2000, it does not reflect the way such data were collected in previous censuses; individuals choosing more than one race are not classified in the same category as in previous censuses. As a result, valid comparisons with past (and future) censuses will require modification of the race groupings with so-called 'bridging' techniques; unfortunately for demographic analyses, the data necessary to produce consistent group definitions across time may, at best, be approximate or may simply not be available at all.

6. Conclusion

Censuses provide the principal measure of population size for countries and geographic subdivisions within countries while also providing detailed information on the demographic, social, and economic characteristics of the population. An accurate census is a crucial feature of a national statistical system, providing baseline data and benchmarks for other collection systems. Demographic techniques, particularly for assessing age and sex structure and for analyzing changes over time, provide essential means for utilizing the data and for assessing their quality. For further information on censuses and census data, see Anderson (2000) and *Censuses: History and Methods*; *Censuses: Comparative International Aspects*.

Bibliography

Anderson M J (ed.) 2000 *Encyclopedia of the U.S. Census*. Congressional Quarterly Press, Washington, DC

Anderson M J, Daponte B O, Fienberg S E, Kadane J B, Spencer B D, Steffey D L 2000 Sampling-based adjustment of the 2000 Census: A balanced perspective. *Jurimetrics* **40**: 341–56

Anderson M J, Fienberg S E 1999 *Who Counts? The Politics of Census-Taking in Contemporary America*. Russell Sage Foundation, New York

Bell W R 1993 Using information from demographic analysis in post-enumeration survey estimation. *Journal of the American Statistical Association* **88**(423, September): 1106–18

Bogue D J, Arriaga E E, Anderton D L (eds.) 1993 *Readings in Population Research Methodology: Volume 1. Basic Tools*. United Nations Population Fund, Social Development Center, Chicago

Edmonston B, Schultze C L 1995 *Modernizing the U.S. Census*. National Academy Press, Washington, DC

Ewbank D C 1981 *Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis*. Report No. 4, Committee on Population and Demography, National Academy Press, Washington, DC

Hogan H, Robinson J G 1993 What the Census Bureau's coverage evaluation programs tell us about the differential undercount. In: *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*. US Department of Commerce, Washington, DC

Passel J S 1996 The growing American Indian population, 1960–1990: Beyond demography. In: Sandefur G D, Rindfuss R D, Cohen B (eds.) *Changing Numbers, Changing Needs: American Indian Demography and Public Health*. National Academy Press, Washington, DC

Robinson J G, Ahmed B, Das Gupta P, Woodrow K A 1993 Estimation of population coverage in the 1990 United States census based on demographic analysis. *Journal of the American Statistical Association* **88**(423): 1061–71

Shryock H S, Siegel J S 1980 *Methods and Materials of Demography*, 4th printing, rev. US Government Printing Office, Washington, DC

Spiegelman M 1968 *Introduction to Demography*, rev. edn. Harvard University Press, Cambridge, MA

Waters M C 1990 *Ethnic Options*. University of California Press, Berkeley, CA

J. S. Passel

Censuses: History and Methods

1. Definition/Overview

A census is a count of the population of a country as of a fixed date. National governments conduct censuses to determine how many people live in different areas of the country, whether the population is growing, stable, or declining in the country as a whole or in particular parts of the country. They also determine what the characteristics of the population are in terms of age, sex, ethnic background, marital status, or income. Generally governments collect the information by sending a questionnaire in the mail or an interviewer goes to every household or residential address in the country. The questionnaire asks the head of the household, or a responsible adult in the household (the respondent), to list all the people who live at the address as of a particular date, and answer a series of questions about each of them. The respondent or the interviewer is then responsible for sending the answers back to the government agency, which in turn adds up the results, or tabulates or aggregates the answers from the country overall and for political subdivisions such as states or provinces, cities, counties, or other civil divisions. The agency usually reports the results to the public a few months or years after the census; the results are considered 'news' and are reported in the media. Since censuses aim to count the entire population of a country, they are very expensive and elaborate administrative operations, and thus are conducted relatively infrequently, generally at five to ten year intervals.

Between censuses, governments estimate the size and characteristics of the population, either by extrapolating the trends in the census into the future, estimating the population from other data systems such as tax or drivers license records, or by conducting periodic sample surveys to collect information about the population. Representative probability samples collect information from a small portion of the population, and thus can be conducted frequently, even monthly. In the USA, the Current Population Survey of around 50,000 households is conducted monthly. National governments also conduct other types of censuses, particularly of economic activity, such as an agriculture, manufacturing, or business census. Such censuses collect information on the number and characteristics of farms, businesses, or manufacturing firms. In the nineteenth century, such censuses were conducted at the same time as the population census. Today, the economic censuses are generally conducted on a different schedule from the population census (see *Statistical Systems: Censuses of Population*).

2. History

Censuses have been taken since ancient times by emperors and kings trying to assess the strength of their realms. These early censuses were conducted sporadically, and generally served to measure the tax or military capacity of a particular area. They tended to count adult men, men liable for military service, or people liable to tithe (pay taxes). For census taking of an entire population to become feasible, a uniform unit of analysis was required. Hence census taking in the West had to await the emergence of the concept of commensurate households, generally seen as a development of the Medieval European West (Herlihy 1985). The household or family served as the unit or analysis or the locus for counting the members within it.

Generally speaking the modern periodic census of all persons is an invention of the early modern period in the European West and, particularly from a New World perspective, was associated with efforts by the home country to determine the success of the overseas colonies. Thus the British Crown and the British Board of Trade ordered repeated counts of the colonial American population in the seventeenth and

eighteenth centuries, starting in the 1620s in Virginia (Wells 1975, Cassedy 1969). In Canada, French efforts to count the population began in 1665–1666 when Jean Talon came to the New World on behalf of Louis XIV and took a census of what became Quebec. Censuses were continued at irregular intervals after Canada became a British colony in 1763 (Worton 1997). By the early nineteenth century, census taking began to be a regular feature of government in Western Europe and North America (Alterman 1969, Nissel 1987, Glass 1973, Desrosieres 1993, Patriarca 1996). The International Statistical Congress (established 1853) and the International Statistical Institute (established 1885) proposed and promoted the uniform regular censuses for all national states. In the twentieth century, census taking spread throughout the world. The United Nations Statistical Office compiles reports on population worldwide (Ventresca 1996).

3. *Functions and Techniques*

Censuses serve a variety of purposes in different countries. At a minimum a census provides a measure of the size of the population of a country, which can be compared with the population in the past, the population of other countries, and to make estimates of the likely population in the future. Governments use census information in almost all aspects of public policy, from determining how many children an educational system must serve, to determining where to put new roads. Censuses are also used to provide the denominators of other measures, e.g., measures of per capita income for a state or local area, or for such measures as crime rates or birth or death rates. Private businesses use census data for marketing analyses to determine where to locate new businesses, or to decide where to advertise particular products.

Other government agencies and private researchers use the census to provide the 'sampling frame' for other survey research. This is the address list that researchers use to determine where to conduct a sample or poll. In the USA, the Bureau of the Census or Census Bureau conducts the census during the tenth year of each decade. Each decade, the population count provides the data for reapportioning seats among the states in the House of Representatives and Electoral College, and for redrawing district boundaries for seats in the House, in state legislatures, and in local legislative districts. In Canada, a full census is taken during the first year of every decade and an abridged census is taken during the sixth year of the decade. The population data are also used to apportion seats among the provinces in the House of Commons and to draw electoral districts (see *Policy Knowledge: Census*).

Most nations create a permanent national statistical agency to take the census, such as the US Bureau of the Census or Statistics Canada. The agency in charge generally undertakes a public review process to determine the questions to be asked. The questions vary from nation to nation depending on the particular political and social history and conditions of the country. Most censuses include basic demographic information such as the age, sex, educational background, occupation, and marital status of the individual. Race, ethnic or national origin, and religious affiliation are important questions in many nations. Further questions can include the person's place of birth, relationship to the household head; the individual's or the family's income; the type of house the household occupies; and whether the person is a citizen, has moved in the past five years, or speaks a particular language. Questions that are quite routine in one nation may be seen as quite controversial in another, depending on the history of the country. Americans do not ask questions on religious affiliation on the census since it is considered a violation of the First Amendment right to freedom of religion. Other nations, such as India, do collect such information. Questions on the number of children born to a woman were quite controversial in China in recent years because of the emphasis on the one-child policy of population limitation. In the USA, asking a question on income was considered controversial in 1940 when it was first asked. It is no longer as problematic. Questions change in response to public debate about the state of society. Americans wanted to know which households had radios in 1930, and introduced questions on housing quality in 1940. Canadians have recently begun to ask census questions on the unpaid work done in the home.

Taking a census can be divided into several phases. The statistical agency first divides the country into geographical subdivisions to be counted, and makes maps and address lists, and prepares instructions for the local census takers. In most countries, this phase requires hiring large numbers of temporary workers to do the counting, or calling upon other government employees at a local level, such as schoolteachers, to conduct the count. The central statistical agency prepares and prints the questionnaires, and distributes them to households either through the mail, or by delivery by enumerators. During the second phase of the count a responsible adult or household head in every household, family, or equivalent institution, is asked to fill out the form or respond to the enumerator and supply the required information about each member of the household. Questions for all people usually include a brief set, on a 'short' form, for example, name, age, sex, racial or ethnic status, marital status, and relationship to the household head. Generally a smaller sample of households receive a more complicated or 'long' form which can have many detailed questions on the individual's work status, income, housing, educational background, citizenship, and recent moves. The person who fills out the form or the enumerator is responsible for returning it

to the statistical agency. The respondent then mails the form to the agency; the enumerator collects the form in person or the information by phone. During the third phase, the statistical agency enters the data onto a computer and adds up, or tabulates the responses for the nation, states, or provinces, and cities, towns, and other local jurisdictions. The agency also cross-tabulates the answers, reporting not merely on the number of people in a local area, but on the number of people, for example, in five-year age cohorts, for each sex, for local areas. The agency only publishes the tabulated results of the count, and keeps the individual responses confidential. In the USA, individual census responses are stored at the National Archives. After 72 years, the original forms are opened to the public for use. These original responses are frequently used by people researching the history of their families, or by those constructing genealogies.

Until the 1980s, statistical agencies published census results in large volumes of numeric tables—sometimes numbering in the hundreds of volumes. Since then, census results have become available electronically, on disc, magnetic tape, CD-ROM, or the Internet. The choice of census technique for a particular country is the result of its social and political traditions and technological capacities. The US census is highly automated, and is primarily conducted by mail. Canada sends enumerators to deliver the census form to each household; the household head fills it out and sends it in. Other nations use even more labor-intensive techniques of collecting and tabulating the data. Turkey, for example, currently requires people to stay home to await the census taker and counts the entire population on a specific Sunday census day (see *Censuses: Comparative International Aspects*).

4. *National Traditions*

4.1 *The United States*

The US census was mandated in the 1787 Constitution. At the time of the American Revolution the framers faced many problems of uniting the 13 separate colonies into a national government, including how to allocate political representation among the states, and how to levy taxes.

The initial government structure under the Articles of Confederation gave each state one vote in Congress, and required the states to collect taxes for the national government. This proved unsatisfactory since the states were of widely different sizes, tax capacity, and cultures. The 1787 Constitution derived the sovereignty of the state from the 'people of the US' and created a bicameral legislature with representation in the Senate based on the states; and representation in the House of Representatives based upon the population of each state. The Electoral College was composed of electors determined by summing the House and Senate members for each state. The decennial census was designed to provide the population figures for apportioning the seats in the House according to population. 'Direct taxes' levied on the states were also to be apportioned on the basis of population (Anderson 1988).

At the time of the Constitution, a racially based slave labor system also existed, and almost 20 percent of the American population were enslaved African-Americans. The framers debated whether slaves were 'persons' or 'property' and thus whether states should receive representation for their slave populations. The southern states where slavery dominated did not consider the slave population for purposes of apportioning their state legislatures, but they did consider slaves property for tax purposes. The framers could not find an easy solution to this dilemma, and developed what came to be called the Three-Fifths Compromise, which 'discounted' the size of slave population as equivalent to 60 percent of the free population when determining the apportionment of the House. The Three-Fifths Compromise thus required a separate count of slave and the free, mainly white, population. The Constitution also specified that 'Indians not taxed', that is, those Indians who were not considered part of the civil society were not to be counted in the census. The Three-Fifths Compromise was abolished with the abolition of slavery after the Civil War; the tradition of accounting for the various racial groups in the population continued after abolition.

The first census was taken in 1790. Assistant US marshals were instructed to travel the country and ask six questions at each household. These included: the name of the family head; and for each household, the number of free white males 16 and over; the number of free white males under 16; the number of white females; the number of other free people (the free colored); and the number of slaves. They totalled the figures for their local jurisdiction and sent them to the US marshal for the state that in turn totalled the figures for the state and sent them to the President. The first American census counted 3.9 million people.

In later years, the census became more elaborate, with more questions asked, and more data published (Cohen 1982). In 1850 Congress mandated a census schedule (form) with a line of questions for each person, including their name. A temporary Census Office, as it was then called, was set up in Washington to add up the tables of responses and publish the results in large volumes. By 1880 when the American population topped 50,000,000, the census was still being compiled by hand, using a primitive tally system. In 1890, the Census Office introduced machine tabulation of the responses, and each person's answers were converted to codes punched in Hollerith cards, a precursor to the IBM punch card. The cards were then

run through counting machines. This innovation was the beginning of modern data processing, and led to further innovations in tabulating large amounts of data. By the 1940s, the Census Bureau commissioned the construction of the first non-defense computer, UNIVAC, to tabulate the 1950 census. By the late 1950s, the Census Bureau dispensed with the punch cards and developed an electronic scanning system, called FOSDIC, or Film Optical Scanning Device for Input to Computers, to read the answers on the census form and input the data to computer (Eckler 1972).

In 1940, the US began to collect some census information from a sample of the population, and slowly shifted the detailed questions on the census to the long form asked of only about 15–25 percent of households. In 1970 the American census became primarily a mail enumeration, as the Census Bureau developed automated address files for the country. As of the year 2000, over 90 percent of the roughly 100,000,000 residential addresses in the US receive the census form in the mail. If the Census Bureau does not receive a response, it sends an enumerator to determine if the address is correct and get the information from the household at the address.

4.2 Canada

Regular decennial censuses began in 1851 in Canada. The British North America Act of 1867 required that the census provide population counts for the apportionment of representation in the House of Commons among the four provinces of Ontario, Quebec, Nova Scotia, and New Brunswick, and for the periodic readjustment of the boundaries of electoral districts. The first census of the Dominion was taken in 1871, and counted 3.7 million people.

The first Canadian census after Confederation was also a very elaborate affair, collecting information on agriculture, livestock, animal products, industrial establishments, forest products, shipping and fisheries, mining, and public institutions as well as on population. Canada was primarily an agricultural nation at the time, and the Department of Agriculture conducted the census. In 1905 the census bureau was made a permanent government agency in the Agriculture Department. In 1918 an independent Dominion Bureau of Statistics was created with oversight for taking the census and collecting other statistical information, headed by a Chief Statistician. In 1971, the bureau was renamed Statistics Canada. Canada has a centralized statistical system, with all major statistical work conducted in the same agency (Worton 1997).

Canada began to conduct sample surveys along with the full census in 1941, introduced a quinquennial census (a census on the sixth year of the decade) in 1956, and self-enumeration in 1971. In the 1990s, the Canadian census introduced adjustment of census results to correct for errors on the basis of post-enumeration check survey. The population of Canada in 1997 was 30.3 million.

4.3 Great Britain

The first Parliamentary proposals for regular censuses of population in Britain date to the mid-eighteenth century, but they did not lead to the institutionalization of census taking until the end of the century. In late eighteenth century Britain, the publication of Thomas Malthus' Essay on the Principle of Population and concerns that the population was declining led Parliament to pass a bill 'for taking an Account of the Population of Great Britain and of the Increase or Diminution thereof.' The first count was taken in March 1801 by the Overseers of the Poor under the direction of former clerk of the House of Commons, John Rickman, and counted 9 million people. Census taking continued every decade afterwards except for the cancellation of the 1941 Census during the Second World War.

In 1837 the General Register Office was organized. The decennial census was made a responsibility of the GRO and used the administrative structure of the registration districts to organize the field portion of the count.

The 1841 census introduced individual level enumeration. The mid-nineteenth century census was particularly influenced by the ongoing medical and public health work of the GRO leaders such as William Farr. It therefore developed detailed data on occupation, age, and locale for the denominator data required for the publication of mortality and other statistics (Glass 1973, Nissel 1987).

Machine tabulation of the census was introduced in 1911. The 1920 Census Act established a permanent legislative mandate for the census. Sampling was introduced in 1951 by way of the analysis and publication of results from a 10 percent sample of the complete 1951 forms. In 1961, the census used a short form for the entire population and an additional long form for a 10 percent sample. In 1966, a one in ten sample count was taken. The results of the 1961 census were computer tabulated. By 1966 the Census Office had its own computer facilities. The population of the UK in 2000 is 59 million.

4.4 France

France established a population registration system in the late eighteenth century, and began a tradition of quinquennial census taking under Napoleon in 1801. The first census counted a population of 33 million. For much of the nineteenth century, the administration of the count was local with a small national office

(*Statistique Generale de la France* (SGF)) publishing the results tabulated locally. Individual forms were used in Paris starting in 1817; by 1876 the entire country used three standardized forms: one for the individual, one for the family or household, and a third for the building enumerated. Tabulation was done locally until 1896 when Lucien March introduced machine tabulation. The French pioneered in the analysis of occupational statistics derived from the census as a means of understanding the evolution of the economy from an agricultural and artisan basis to one characterized by professional qualifications and employees with credentials. In 1946, after World War II, the French National Statistical Agency was titled *Institut National de la Statistique et des Etudes Economiques* (INSEE) (Desrosieres 1991). The population of France at the 1999 census was 60 million.

5. *Issues*

Censuses provide important information about the population of a country, and can become embroiled in political or social controversy simply by reporting information relevant to ongoing issues in the society.

Complaints about the census generally involve concerns about the accuracy of the count, the propriety of particular questions, and the uses to which the data are put (see *Censuses: Demographic Issues*).

Censuses require public understanding, support, and cooperation to be successful. Concerns about government interference with private life can prevent people from cooperating with what is an essentially voluntary counting process. People may be suspicious of giving information to a government agency, or may object that particular census questions are invasions of privacy.

When such public trust is lacking, people may fail to participate. In recent times, individuals doubled up in illegal housing units, undocumented immigrants who do not reside in the country legally, or individuals who do not wish to reveal their economic or social situation to a government agency, can be reluctant to respond to a census. In the most serious challenges, people claim the results will not be held in confidence and that census should not be conducted at all. During World War II, census records from countries occupied by the Nazis, for example in the Netherlands, were used to identify Jews for detention, removal, and extermination. The effect of such use was to undermine the legitimacy of the census after World War II. In the Netherlands, which took its last regular census in 1971 and collects population information through other mechanisms, the legacy of the Nazi era was one of the major justifications for ending census taking (Seltzer 1998).

At the other side of the spectrum, political challenges can be made to the census if the census does not count the population well enough. All censuses contain errors of various kinds. Some people and addresses are missed. People may misunderstand a question, or fail to answer all the questions. Census officials have developed elaborate procedures to catch and correct errors as the data are collected, but some errors tend to remain. Since census results are often used to allocate seats in legislative bodies and government funds, such errors undermine the credibility of the census as an allocation mechanism.

In recent years, developments in statistical analysis have made it possible to measure the accuracy of censuses (Choldin 1994, Anderson and Fienberg 1999). Census results may be compared with population information from other sources, such as the records of births, deaths, and marriages in vital statistics. Census officials can also determine the level of accuracy of the count by conducting a second, sample count, (a post enumeration survey or PES) shortly after the complete census, and then matching the records of the sample and the census. Census officials estimate who is missed, and who is counted twice or in the wrong geographic location, and use these estimates to evaluate the overall accuracy of the count. Canada and Australia adjust the census results for omissions and other errors. Great Britain is planning to adjust the 2001 census (Chambers and Crudas 2000, Paice and Steel 2000).

In the USA, people who live in cities, the poor, and minorities tend to be undercounted relative to the rest of the country. Since 1970 officials representing such undercounted jurisdictions claimed that these jurisdictions have suffered loss of political representation and government funding since the apportionment and funding formulas are based on incorrect data. Mayors and leaders of civil rights organizations have pressed for adjustment of the census results, and have filed lawsuits to compel adjustment of the enumeration. The courts have not ordered adjustment of the US census results, and the question of adjustment has emerged as a political controversy in Congress. Since the late 1980s, Republicans have generally opposed adjusting for the undercount, while Democrats have supported it. The nation will debate these issues anew after the 2000 Census since the Census Bureau plans to publish results based upon both adjusted and unadjusted population counts and to provide their professional judgment indicating which data set they consider the more accurate.

Congress, states, and local governments could use adjusted counts for determining legislative districts, funding allocations, and for program evaluation, though such actions are likely to draw further court challenges.

See also: Population Cycles, Formal Theory of; Population Forecasts; Statistical Systems: Censuses of Population

Bibliography

Alterman H 1969 *Counting People: The Census in History*. Harcourt, Brace & World, New York
Anderson M 1988 *The American Census: A Social History*. Yale University Press, New Haven, CT
Anderson M, Fienberg S E 1999 *Who Counts? The Politics of Census Taking in Contemporary America*. Russell Sage Foundation, New York
Cassedy J 1969 *Demography in Early America: The Beginnings of the Statistical Mind*. Harvard University Press, Cambridge, MA
Chambers R, Cruddas M 2000 A one number census for the United Kingdom. *Chance Journal* **13**(3): 38–41
Choldin H 1994 *Looking for the Last Percent: The Controversy over Census Undercounts*. Rutgers University Press, New Brunswick, NJ
Cohen P C 1982 *A Calculating People: The Spread of Numeracy in Early America*. University of Chicago Press, Chicago
Desrosieres A 1991 Official statistics and medicine in nineteenth-century France: The SGF as a case study. *Social History of Medicine* **4**: 515–37
Desrosieres A 1993 *La Politique des Grands Nombres: Histoire de la Raison Statistique*. Edition La Découverte, Paris [1998 The Politics of Large Numbers: A History of Statistical Reasoning. Harvard University Press, Cambridge, MA]
Eckler A R 1972 *The Bureau of the Census*. Praeger, New York
Glass D V 1973 *Numbering the People*. Saxon House, Farnborough, UK
Herlihy D 1985 *Medieval Households*. Harvard University Press, Cambridge, MA
Paice J, Steel D 2000 Census adjustment in Australia. *Chance* **13**(3): 41–2
Patriarca S 1996 *Numbers and Nationhood: Writing Statistics in Nineteenth-century Italy*. Cambridge University Press, New York
Nissel M 1987 *People Count: A History of the General Register Office*. Her Majesty's Stationery Office, London
Seltzer W 1998 Population statistics, the Holocaust, and the Nuremberg trials. *Population and Development Review* **24**: 511–52
Ventresca M 1996 *When States Count: Institutional and Political Dynamics in Modern Census Establishment, 1800–1993*. Ph.D. dissertation, Stanford University, Stanford, CA
Wells R 1975 *The Population of the British Colonies in America before 1776: A Survey of Census Data*. Princeton University Press, Princeton, NJ
Worton D A 1997 *The Dominion Bureau of Statistics: A History of Canada's Central Statistics Office and its Antecedents: 1841–1972*. McGill-Queens University Press, Kingston, ON

M. Anderson

Center–Periphery Relationships

Concepts of center and periphery became increasingly prominent in the social sciences during the second half of the twentieth century. They were of diverse origins and entailed varied emphases, but these differences have not always been obvious in continued use. There has actually been a cluster of related conceptual pairs: center and periphery, but also core and periphery, metropolis and satellite, and metropolis and province. But they have not been mere synonyms. Moreover, in obscure and perhaps treacherous ways, ideas of center and periphery may occasionally have overlapped in social thought and discourse with other classical contrasts, such as modernity–tradition, or urban–folk.

The sociologist Edward Shils had a major part in introducing the center–periphery pair of concepts into the vocabulary of academic social science (cf. Greenfeld and Martin 1988). His paper 'Center and Periphery' originally appeared in 1961, but there was a close affinity between it and several of his other publications in the same period. The overall conception of society and culture in Shils's perspective was consensualist; but not, he asserted in a later comment, as merely a facile expression of the mood of the era. In his view, the importance of consensus to social life had been underestimated (Shils 1975, p. xi). Thus he concentrated on notions such as 'tradition, 'ritual,' 'deference,' and 'charisma.'

Notably, in his 'Center and Periphery' paper, these two terms were used in a metaphorical sense—centrality had 'nothing to do with geometry and little with geography.' The center was identified with what was ultimate, irreducible, and sacred in the realm of symbols, values, and beliefs; it was also identified in the realm of action with those roles and institutions which embodied such cultural understandings and were most actively engaged in propagating them. Yet in 'Metropolis and Province in Intellectual Life', published in the same year, the facts of uneven spatial organization which are now usually linked to concepts of center and periphery become clear, and are set in what we would now describe as a transnational context. Drawing on his study of the situation of intellectuals in India, Shils argued that in their minds, people have varyingly extensive maps of the world significant to them, and that a major feature of such maps would be their portrayal of one's qualitative proximity to or distance from the metropolis (Shils 1972, p. 356).

In the former of these two papers, then, the center is a locus of excellence, of grace; the periphery defers to it. In the latter, the metropolis is a center of vitality, a seat of creativity. The province, on the other hand, Shils notes, is frequently taken to be in itself 'rude, unimaginative, awkward, unpolished, rough, petty, and narrow' (Shils 1972, p. 357). Cultural salvation lies in an involvement with the metropolis. The choice seems to be one between impoverished autonomy and enriching dependence.

If Shils was a theorist of consensus and cultural authority, the metropolis/satellite and core/periphery pairs, as inserted into the debates of the 1960s and 1970s, pointed in an entirely different direction, resonating with another political mood. 'Metropolis' and 'satellite' were the terms used by the economist

Andre Gunder Frank (1967), who had spent much of the 1960s in Latin America, studying what he described as the development of underdevelopment, and engaging with the growth of 'dependency theory.' Then in the 1970s, introducing 'world-system theory,' the sociologist Immanuel Wallerstein (e.g., 1974) contrasted 'core' with 'periphery.' Both Frank and Wallerstein were primarily concerned with political economy, and with the expanding control and exploitation of the material resources of the 'periphery' or 'satellite' on the part of the 'core' or 'metropolis.' In both instances, the emphasis was on conflict and domination within a global order.

The points of view represented by Shils, on the one hand, and by Frank and Wallerstein on the other, seem not often to be brought into direct confrontation, although over the years, they may have come to interact and blend with one another. By now, when center–periphery concepts are more passingly referred to, it is not always entirely clear which of the varieties is involved, if not some combination of them. Yet there is a field of tension here between studies focusing on distributions of material assets and the exercise of power on the one hand and studies of culture on the other; between views suggesting consensus and views emphasizing conflict; and between views treating the social organization of meaning and meaningful forms somewhat in isolation and other views which insist on setting it in the context of political and economic structures.

Probably all would accept, however, that centers and peripheries imply each other. We are dealing with relational phenomena: there is no center without periphery, no periphery unless there is a center. The particular way the center is a center is also reflected in the way the periphery is a periphery.

1. *Levels and Varieties*

Our conception of center–periphery relationships here is that they are relationships of inequality existing in geographical space. Such relationships, however, have been defined at very different levels of specificity. During the latter half of the twentieth century, especially in the Cold War era, the world was understood to be divided into three main segments—one Western and capitalist, one Socialist, one 'developing' and postcolonial. Here the First and the Second Worlds had their internal center–periphery structures. The Third World, on the other hand, may have had these, but also tended to be seen rather more as a periphery in its entirety, to one First or Second World center or other. At present the extreme macroview may be that of a center North and a peripheral South. Much of the cultural and political debate over center–periphery relationships remains mostly at such levels of identification. Notions such as Orientalism, Occidentalism, and Eurocentrism refer to understandings and representations of self and other in the context of these large-scale center–periphery structures.

Alternatively, identifications may indeed be made at the national level: the United States or France are seen rather more as centers than Sweden, Romania, or Burkina Faso. In other instances, centers are yet more specifically placed. Nations and regions may have their own centers in particular localities. In the study of historical civilizations, there is frequently an emphasis on centers as sites combining political power with complex ritual life and elaborated knowledge systems in the hands of specialists. A well-known paper by the anthropologists Bernard Cohn and McKim Marriott (first published in 1958), for example, delineates the complementary roles of centers and networks in the integration of Indian civilization (Cohn 1987, 78ff). In related mid-twentieth century work, as Redfield and Singer (1954), also anthropologists, discussed the cultural role of cities, they described the dynamic relationship between the 'great tradition' of early urban centers and the 'little traditions' of the surrounding peasant societies. The emphasis on the symbolic authority of the center here is clearly reminiscent of Shils's view.

Yet Redfield and Singer also noted that later urban centers, rather than refining particular local traditions, are often at the crossroads of the world, bringing together diverse traditions and serving as communication nodes for wider areas. In this the authors adumbrated recent interest in the role of particular cities in contemporary transnational center–periphery relationships. 'World cities,' such as New York, London, or Paris, may be seen as generalized, multipurpose centers, combining various kinds of power and drawing the attention of the periphery for many different reasons (Hannerz 1993, Knox and Taylor 1995). The contemporary global structure of center–periphery relationships, however, can be understood as more internally varied. Cities may be centers in particular ways to particular people, dispersed in a transnational periphery—Rome to the Catholic world; San Francisco to gay people not only from elsewhere in the United States, but also from other continents; Memphis, Tennessee, to friends of country music. As we see, attachments of very diverse kinds are involved. This means that the same people in the periphery may well be involved with several, noncompeting centers for different purposes. Yet centers of the same type can also compete in the periphery.

How do centers become centers, even drawing attention across national boundaries? Some are centers for mainly historical or mythical reasons, because something occurred there once in the past, or occurred there first, or because the memory of a charismatic figure is preserved there. They belong on the map of the significant past, or of eternal truths. Often they are intensified as centers at particular points in a ritual calendar. Many 'centers of pilgrimage' have such

characteristics. Other centers are very much of the present. They draw attention and stimulate imagination from afar because they are 'where the action is' now, with regard to one or more lines of human preoccupation. The contemporary proliferation of center–periphery relationships of transnational reach is undoubtedly related to a greater ease of transportation and communication, making centers easy to get to, and easy to stay in touch with. As the participants, practitioners, adherents, activists, and employees of more subcultures, lifestyles, ideologies, occupations, or corporations use new means to extend their circles out of local or national habitats, new faraway centers are discovered, or even made. As more diasporas are generated by more migrants, old centers may also be reaffirmed, from further away.

2. *Center–Periphery Diffusion*

In part, it may be in the passage of culture from center to periphery that the two parties to the relationship manifest themselves. In the clearest case, the center would be always the donor, the periphery always the recipient—the center is active, the periphery is passive. Yet even if the imbalance is not quite so great, some measure of net cultural export would appear to be one conception of center–periphery influence. What is prominently involved here, consequently, is diffusion.

In a wider arena of debate about the effects of global interconnectedness, since mid-twentieth century, the preoccupation with cultural diffusion from center to periphery has often been expressed in one or the other of two ways. Post-World War II modernization theory offered a forecast of where the world was heading—where the West actually already was, more or less, and where the rest would follow, with some assistance from the center. When not drily 'value-free,' modernization theory tended to take a positive view of the changes in question. The other major way of describing cultural diffusion on a global scale was less favorably inclined. The term 'cultural imperialism,' as Tomlinson (1991, 3ff), has pointed out in a review, is a combination of two complicated concepts into one, with a considerable ideological charge. It has tended to refer to a growing dominance of western, and especially American, media and consumer goods in other parts of the world. Coca-Cola, McDonald's, and Barbie dolls have become the predictable pieces of evidence in a genre of cultural critique. Both the early form of modernization theory and the critique of cultural imperialism thus tend to offer global homogenization scenarios; implying, forecasting, or warning of the end of cultural diversity.

Such scenarios, however, have been increasingly often contested. There has been a growing attention to the multicentricity of culture, to crisscrossing cultural flows, and not least to instances of cultural counterflow, from periphery to center. While some net asymmetry of flow may yet be undeniable, many commentators have also become less inclined to accept the assumption of passive reception at the periphery. They are much more likely to see an active periphery engaged in managing diffusion from the center: the periphery accepts this, rejects that, modifies one thing, and synthesizes something else with items from its own local cultural inventory. Metaphors of creolization and hybridity summarize this combination of cultural diffusion with creativity at the periphery.

3. *Mixed Feelings*

It is in the nature of their asymmetrical relationship that center and periphery take different views of it, and may even be differently aware of it. A periphery has to attend to its center. The center for its part may at times be preoccupied with its internal affairs, and may not have a clear grasp of the more widely dispersed consequences of its actions. The periphery tends to have a more developed idea of the center than vice versa.

Emotions may be attached to such linkages. Center–periphery relationships often do not leave people cold, passionless, neutral. In the consensualist view, centers generate warm, deferential attitudes at the peripheries. In a conflict perspective, in contrast, center and periphery are primarily defined by political and economic structures—and we may then see their cultural concomitants at the periphery in the responses to control, coercion, and constraint. Along such lines, the many forms of cultural resistance have been a major theme in the portrayal of center–periphery relationships. The periphery may defend itself symbolically against intrusions through a celebration of the self and its local and traditional roots, but there may also be a denigration of the other, an uncomplimentary representation of the center.

Yet the responses of the periphery to the center are not always clear-cut. For one thing, claims to center–periphery consensus now increasingly meet with skepticism. With Gramsci as a source of theoretical inspiration, observers may interpret deference in the more complicated terms of hegemony. It is likewise possible, however, that people at the periphery may be genuinely of two minds about a center. If ambivalence is a prevalent quality of social life, as Smelser (1998) has argued, perhaps this is even as common a response to a center as any more one-sidedly favorable or unfavorable stance. It may be a place one loves to hate and hates to love. One may feel that good things come from there, and at the same time resent its influence, or what may seem to be its narcissism. Especially those multipurpose centers understood to be 'where the action is,' moreover, may inspire sentiments which are positive, but instrumental or playful rather than deferential. They are among the liminal spaces of contemporary life.

All views of the center, however, and all messy feelings toward it, are not necessarily to be found in

any single inhabitant of the periphery. They may rather be complicatedly distributed among its population, and such variations can be the foci of intense debates and conflicts structuring local life at the periphery.

4. The End of Center–Periphery Relationships?

With some frequency, one now hears voices criticizing center–periphery concepts, or proposing the decline of center–periphery structures. It is said that the world economy is increasingly decentered (see, e.g., Lash and Urry 1994, p. 4). Human life, it is also argued, is increasingly deterritorialized. People move quickly between places, being rooted in neither of them, and have relationships which belong in no place in particular. 'Critical masses' for cultural elaboration may be built up in cyberspace, and do not require local face-to-face contacts. As would-be centers and would-be peripheries are closely in touch, by way of electronic media or jumbo jets, the 'cultural lag' which divided them before is no longer there.

Even if they are often closer to futuristics than to the really existing world, there may be something to such arguments, At times, however, both the concepts criticized and the alternatives to them would need to be elaborated more precisely. Indeed we must be sensitive to the ambiguity of our terms, and to the actual range of variations out there. Asymmetries may indeed only be relative, culture flows are not entirely one-way, the world is crisscrossed by more symmetrical relationships as well, and there are many centers, and many kinds of centers, rather than only one. Centers also rise and decline. Center–periphery conceptualizations now deserve to be scrutinized and developed, rather than merely rejected. It is notable, too, that the end of centers and peripheries appears often proclaimed by commentators whose own vantage point is at the center—in this case not a privileged position.

See also: Dependency Theory; Globalization and World Culture; Globalization, Anthropology of; Globalization: Political Aspects; Hegemony: Anthropological Aspects; Hegemony: Cultural; Pilgrimage; World Systems Theory

Bibliography

Cohn B S 1987 *An Anthropologist among the Historians and Other Essays*. Oxford University Press, Delhi

Frank A G 1967 *Capitalism and Underdevelopment in Latin America*. Monthly Review Press, New York

Greenfeld L, Martin M (eds.) 1988 *Center: Ideas and Institutions*. University of Chicago Press, Chicago

Hannerz U 1993 The cultural role of world cities. In: Cohen A P, Fukui K (eds.) *Humanising the City?* Edinburgh University Press, Edinburgh, UK

Knox P L, Taylor P J (eds.) 1995 *World Cities in a World-system*. Cambridge University Press, Cambridge, UK

Lash S, Urry J 1994 *Economies of Signs and Space*. Sage, London

Redfield R, Singer M 1954 The cultural role of cities. *Economic Development and Cultural Change* **3**: 53–73

Shils E 1972 *The Intellectuals and the Powers*. University of Chicago Press, Chicago

Shils E 1975 *Center and Periphery*. University of Chicago Press, Chicago

Smelser N J 1998 The rational and the ambivalent in the social sciences. *American Sociological Review* **63**: 1–15

Tomlinson J 1991 *Cultural Imperialism*. Pinter, London

Wallerstein I 1974 *The Modern World-System*. Academic Press, New York

U. Hannerz

Centers for Advanced Study: International/Interdisciplinary

In the current period, scores of institutions of scholarly learning around the world include the words 'advanced study,' or some approximate translation thereof, in their formal titles. Others that might appropriately use such a title do not, often for simple historical reasons. Moreover, some institutions using the term have little in common, in structure or function, with the leading examples of such organizations. Therefore it is useful to review the main goals that have come to be associated with 'advanced study,' before considering the variety of institutional designs used to address them.

1. The Essence of 'Advanced Study'

The phrase 'advanced study' is hardly a technical term. Its superficial meaning is obvious, and for this reason its original coinage is not easy to trace for the English-speaking world. But its popularity and more restrictive meaning among scholars is quite surely a phenomenon of the twentieth century, springing directly from the creation and rapid success of the Institute for Advanced Study at Princeton University in the United States, officially incorporated in 1930. The original design of this Institute was in turn rather clearly the innovation of Abraham Flexner. In the 1920s, Flexner had become a noted student and critic of the American system of higher education through a series of publications including the *Flexner Report*, a sharp critique of the nation's medical schools. Flexner was later approached by the Bamberger family, who wished to endow a new medical college in New Jersey, to serve as an agent in this enterprize. Flexner soon came to recommend the endowment not of a medical college but of a new kind of university. His vision was of an institution without undergraduate students, in

which advanced graduate students and top-flight professors would form a partnership of scholars who would 'be left to pursue their own ends in their own ways … in tranquillity …' As this vision crystallized, Flexner decided it should be called 'an Institute of Higher Learning or Advanced Studies.'

The phrase was launched into a realm of high visibility. It helped that one of the first recruits was Albert Einstein, soon surrounded by world-class mathematicians, physicists and cosmologists who could engage in fruitful dialogues with him.

Flexner's conception was deliberately designed as a counterweight to the conventional university. Universities, at least in America, are the long-acknowledged home for basic research and the pursuit of increasingly complex topics in the advancement of knowledge. Flexner felt that a center for advanced study, dedicated to roughly the same ends, should be different. Some said that Flexner's 'innovation' was a retreat to the simplicity of Plato's Academy. But he was attempting to address issues surrounding the interplay between the advancement of knowledge and the standard forms of modern university life.

While 'advanced study' might seem to involve the examination of extremely complex or intricate topics, a more essential definition is *research at the most remote frontiers of knowledge*, as they stand at any point in time. The core of well-established knowledge is for the most part constantly expanding. Institutions of education, including higher education, are mainly dedicated to the transmission of such bodies of 'received wisdom.' This transmission presumes at least a two-tier system of actors, being the masters who know and the young who are learning.

The frontier itself is a fuzzy region, which begins at the far edge of consolidated knowledge where doubt first rises as to the shape and meaning of what lies beyond. As one proceeds outward, fragments of knowledge become increasingly sparse, and general uncertainty increases rapidly. As with most poorly-explored zones, there is no dearth of conjecture as to what lies there, but controversy is high. There is also confidence that further exploration can in due time produce enlightenment, and hence true additions to consolidated knowledge. To explore effectively it is usually helpful to have mastered established knowledge in the terrain closest to that edge of the frontier. But once past that prerequisite—and the more advanced of graduate students have achieved that state—everyone can be an explorer at the same level. Hence Flexner's collegium of scholars was realistic, along with his welcome proviso that such a group be as liberated as possible from the routine demands of daily life, including most of the bureaucratic and teaching activities required of the typical university professor.

Conventional universities have long been organized into specific disciplines such as physics, biology, or economics. These disciplines have their own regional frontiers, and are expert at sending out fingers of inquiry which in disciplinary terms illuminate previously unknown terrain. But deeper and richer probes that consolidate knowledge of larger areas, such as linking up several neighboring peninsulas to one another, are often best carried out by teams from multiple disciplines. This has, of course, become an increasing necessity, as the bodies of already-established knowledge within given disciplines have become so large that mastery of even the local 'consolidated knowledge' of a single narrow discipline consumes much of the education process. Therefore most centers of advanced study are more or less deliberately interdisciplinary, and even take steps to ensure that resident scholars do not segregate themselves by discipline, but are exposed to enriched stimulation from broader cross-disciplinary discussions. Scholars who have spent long periods of study and teaching in conventional universities before encountering this novel environment testify that they have sorely missed such broad-spectrum intellectual interaction ever since they first devoted themselves to a single discipline by entering graduate schools. Moreover, scientific knowledge about the natural world, as well as the symbolic world of mathematics, is notably cosmopolitan. Therefore diversity in national origins is also sought by most centers for advanced studies, given the possibility of relevant genius in any part of the world, and the impress of cultural perspectives on knowledge.

2. *Illustrative Institutional Designs*

The Institute for Advanced Study at Princeton is not only the original model for advanced study centers, but is probably larger in endowment and size, as well as structural complexity, than any others. By now, it has a permanent faculty of nearly two dozen, spread over four distinct schools. These faculties review applications from further scholars to be designated as Members, who reside at the Institute for periods ranging from a single term to several years in occasional cases. Such Members normally propose research projects that fit the specialized interests of scholars with permanent faculty appointments, although some Members are encouraged to work in areas which look profitable to the current faculty, but which are under-represented in the relevant school. Other scholarly visitors, in addition to those selected as Members, are also hosted for shorter periods where mutually fruitful. In all, some 200 scholars spend significant periods of time at the Institute in any given year.

The Princeton Institute began with a special focus on mathematics and classical studies, but has broadened considerably over time. The School of Mathematics continues to be the largest collegium of scholars. The School of Natural Sciences is substantial in size as well, although it has a heavy emphasis on

astrophysics and particle physics. The School of Historical Studies continues to reflect its original dedication to ancient history and classical studies, but now enjoys significant representation in medieval and modern history. The School of Social Science was not added until the early 1970s, and remains the smallest of these units. In addition to the major social-science disciplines which have core representation, the School hosts Members and visitors in literature, philosophy, and even art history.

In 1954 the Center for Advanced Study in Behavioral Sciences, with the help of the Ford Foundation, opened its doors on land of Stanford University in Palo Alto, California, although it was administratively independent of that institution. It was inspired by the generic example of the Princeton Institute, but was also motivated as a response to the scant attention being given to the social and behavioral sciences in the Princeton Institute's program at the time. Although its broad goals are parallel to those at Princeton, its mode of operation differs in a number of particulars. Most notably, it has no permanent faculty appointments. Instead, a small secretariat invites nearly four dozen scholars as residential Fellows each year. Scholars are given office space and various forms of computer and research support, and are encouraged to pursue whatever reflections, research, and writing they wish. Various methods are employed to encourage informal intellectual communication between the disciplines present.

The disciplinary scope of these fellowships for the Center at Stanford is extremely broad. To be sure, the five primary social sciences—anthropology, economics, political science, psychology, and sociology—make up a majority of each cohort, especially as augmented by, typically, an ample representation of historians. But leaders in research in other diverse social science specialties, such as education, linguistics, statistics, geography, law, and philosophy appear with fair regularity from cohort to cohort. In addition, there has been on one hand routine representation from the boundaries in the form of literary studies or, occasionally, well-known authors of fiction, and on the other hand representatives of the natural sciences, including ornithologists, biomedical researchers, Nobel Laureates in biology, and even a chemist working on the social communication of insects. While the normal fellowship gives free rein to individual researchers, the Center at Stanford dedicates some fraction of slots within each cohort to 'special projects' prearranged for work on scientific or policy problems that require sustained multidisciplinary attention. On average there are about two such projects per year, participated in by a fifth to a tenth of the Fellows in residence.

The general model for the Center at Stanford has been adopted by newer advanced study institutions of interest to social and behavioral scientists. In 1970 the Netherlands Institute for Advanced Study in the Humanities and Social Sciences was founded through the efforts of an eminent Dutch linguist, E. M. Uhlenbeck, following his experience as a Fellow of the Stanford Center. Other institutions have been established in Europe with attention to the same model, such as the Institute for Advanced Study (*Wissenschaftskolleg*) in Berlin (1980), and the Swedish Collegium for Advanced Study in the Social Sciences (1985) at Uppsala. The National Humanities Center in North Carolina was also promoted largely by former Fellows of the Center at Stanford, and again modeled closely after it.

There are local adaptations in form across these institutions. For example, the Dutch Institute requires that half of its Fellows be from the Netherlands. The Berlin Institute hosts a few permanent Fellows as well as annual residential ones. The prominence of group projects or annual 'themes' varies from institution to institution. None of these centers, however, is designed to train graduate students or offer advanced degrees; instead, they are dedicated to promoting research of great distinction. They also enjoy a high degree of autonomy from specific universities as well as state or federal governments, although they may receive funding from, and engage in relationships with, such agencies.

In addition to the growing number of such large-scale centers, numerous smaller intrauniversity units with similar goals have emerged. While advanced study centers will always be vastly outnumbered by universities and other training institutions, the existence and popularity of these centers reflect an increasing awareness that the Flexner ideal for a collegium of research scholars is, under modern conditions, an indispensable part of the infrastructure desirable for the healthy advancement of knowledge.

See also: International Research: Programs and Databases; Specialization and Recombination of Specialties in the Social Sciences; Think Tanks

Bibliography

Flexner A 1960 *An Autobiography*. Simon and Schuster, New York

P. E. Converse

Central Africa: Sociocultural Aspects

Most authors define Central Africa as the vast area comprising speakers of the different western branches of Bantu. The great diversity of western Bantu populations and traditions have gradually extended, beginning 5,000 years ago, from southern Cameroon,

into Equatorial Guinea, Gabon, People's Republic of Congo, the Democratic Republic of Congo, and Angola. Side to side to Ubangi and Nilotic speakers, they have moreover spread across the Central African Republic, southern Sudan, Western Uganda, Rwanda, and Burundi, to Eastern Congo, Tanzania, and Northwestern Zambia. Whereas the eastern Bantu have specialized in grain growing, the western Bantu adopted cassava and yams for staple foods. The region has been the scene of complex interactions between the local and outer worlds.

1. *AD 1000 to 1880*

Probably due to its sparse population, Central Africa did not develop the type of domestic markets and specialized local industries as found in West Africa. Until the late fifteenth century, it stayed out of the lines of long-distance communication such as those that had developed from coastal East Africa or along the caravan routes of West Africa. The history of precolonial Central Africa from AD 1000 can be divided broadly into two main regional developments.

With regard to the northern half of Central Africa, comprising the great equatorial forest with its northern woodland fringe and the Congo basin, linguistic and ethnographic data—common 'words and things' (Vansina 1990)—bear witness to the interaction between the farming and pastoral peoples in the forest and savanna areas north and south of the equator. The most important hunter-gatherer groups of Northern Central Africa were the matrilineal Pygmies, interacting with their agricultural neighbours and long-distance trading groups (like the salt-making Vili of Loango, and the neighboring copper-trading Tio). The Zaire river and its tributaries have been an important corridor for the transmission of cultural influences and systems of government, as well as for the trading economies of the Atlantic zone. The farming in the thinly scattered patrilineal societies of the northern and eastern corner of the rain forest remained for long untouched by the food plants originating from South America and already imported into other parts of Central Africa.

In the central part of the equatorial forest, the Mongo patrilineal societies were strongly influenced by the neighboring agriculturalists to the northwest along the Ubangi and higher up in the Sudan. Mongo Big Men, often combining agriculture and trade, tended to privilege their patrilocal household and patriclan rather than the village community as an economic stronghold and political unit, also by way of matrimonial transactions and cultic and initiatory specializations. By the sixteenth century, long-distance exchange economies of textiles and fire-arms for ivory and local captives may have connected the equatorial forest even with Egypt.

The savanna regions in the southern and western half of Central Africa saw the emergence of important political leadership roles, including kingship. Depending on the rainfall, the wooded savanna was moderately suitable for agriculture and/or pastoralism. Priests were the guardians of all reproductive resources. The extended family was the focus of physical and spiritual well-being of both living and deceased members of the patrilineage. Lineage chiefs gained control over the rain-making cults and the territorial custody of hunting traditions and the collection of tribute. From the fourteenth century, the chiefs gained regional control over iron and smithery, and particularly from the seventeenth century, over fire-arms and market-oriented commerce. This as well as their indirect involvement in the Atlantic slave trade made their power paramount over large areas in central Angola and western Congo.

2. *Challenges from the Atlantic and Colonizing Europe*

Unlike West and East Africa, Central Africa had not entered the maritime trade prior to the arrival of the Portuguese caravels seeking trading bases, from the 1480s onwards. Reports by Portuguese missionaries and merchants as well as occasional correspondence by western-educated Kongo, show how the foreign influence brought radical changes to Kongo kingdom and culture. The kingdom then comprised prosperous farming communities linked to a regional trade of fish and craft goods. From this period, the *nzimbu* shells in the hands of Europeans took on the functions of coinage for the previously unfamiliar purposes of wage-payment and marketing.

Portuguese traders carried Mediterranean manufactures to Kongo in exchange for raffia cloth, ivory, dye wood, and copper. The redistribution of these manufactures (such as North African textiles, iron knives, glass mirrors, glass schnapps bottles, glazed bowls, Venetian beads, and glazed china) was controlled carefully by the royal court. Moreover, Portuguese teachers, artisans, lawyers, and priests enhanced the authority of the king and his closest supporters. A number of Kongo were taken to Europe for further education. From the sixteenth century the history of the southern savanna became tragically bound up with the growth of Atlantic slave-trade. Manpower, rather than landed property as in early modern Europe, was the key to value in the Central African communities. Slave-trading of servile subjects and captives gained via long-distance marketing at the Malebo Pool (where Kinshasa would later develop) became the grim solution to local needs for foreign goods. Like any slave-trading society, the Kongo kingdoms and other neighboring equatorial kingdoms became oppresssive and fractious, and deemed to collapse.

The Atlantic trade gradually came under the control of the semi-colonial Luso–African community of the Loanda hinterland, Africa's first White colony. The long-distance savanna trade-routes probably became avenues for the dissemination of European goods and influence, as well as of the South American crops: maize, cassava, tobacco, tomatoes. Soon, frontier states such as Matamba and Kasanje as well as a new class of trading entrepreneurs took control over trade relations with the European powers. Following the industrial revolution in Europe, new goods were being shipped to Southern Cameroon, Gabon, and the Lower Congo. Towards the end of the eighteenth century, with the involvement of French, English and Dutch merchants, the Atlantic slave trade reached its peak.

Swahili-Arab penetration in Northeastern Zambia and Southeastern Congo in the early nineteenth century started as trade of guns and powder for ivory and slaves. It would soon overthrow the local chiefs and gain independent political power for itself.

Following Livingstone's and Stanley's 'explorations,' the 1885 Berlin Conference legitimized Europe's colonial scramble for Central Africa. The conquest of Equatorial Africa would last until the 1920s before military power, forced labor, harsh rubber exploitation, diseases, and hunger would have decimated half of the population and broken local resistance. Colonial rule gradually deprived local societies of the institutional ability to confront the exploitative forces and exogenous 'civilizing programs' imposed upon them. In the views of the colonizing masters, the bringing about of a philanthropic medical provision and wide-spread school education in the colonies was an essential feature in the international justification of political and economic colonialism. The colony's extractive economy directed at the 'so far unutilized resources' and relying on mandatory labor in lieu of tax payment, as well as its transport and communication technology were to open up the remote areas to the larger world scene. They were to 'uplift' and integrate the local populations in the new era of 'universal civilization and progress.'

Bantu traditional rule was deeply alien to the colony's display of order. Colonial state power was imposed through regulations, enticements, and direct interference, as both a text and texture, absorbing and domesticating people and events by the very acts of *writing* and administrative *ordering*. The textual economy disassembled and inserted local realities into panoptical recording and regulation. The management of people's civil identity and geographic confinement was achieved through the identity cards and mandatory 'passbooks.' Records of marriage, descent, settlement, and land use were meant to bind the populations geographically, and were authoritative for the succession to chiefly titles and for settling matrimonial and family disputes. On the other hand, such regulatory documents were intended to modernize society by freeing the so-called *évolué* from customary collectivism. Population censuses, together with geographic and linguistic-ethnographic mapping and recording of data, allowed for a compartmentalization of languages and ethnic groups.

Unlike the nineteenth century bureaucratic nation–state in the West, the Bantu political traditions of Central Africa do not draw their inspiration from orders of visual representation and architectonic spatial models. They are moulded by organic, hydraulic, and/or animal-totemic metaphors informing political networks and strategies as an order of events, forces, sources, and relations. Membership of and alliances between particular social groups are not primarily tied to a geographic partitioning but to blood ties and to the mythical or primal space-time order in line with the constant cosmogenetic re-enactment of the reproductive and hierarchical weave between the founding ancestors, their foundational and migratory exploits, and their multiple descendents.

The functions of traditional political title holders were, and are, being thought of as prior to, and the source of, all life forms as well as the guarantee of their order. A chief represents and surpasses his subordinates by his twofold function. First, like his totemic animals (leopard, eagle, crocodile), he is a conqueror. Through enthronement, the ruler embodies the founding ancestors and represents the primal space-time order initiated by the immigration of his ancestral people and their conquest of the land. Through his body, in particular his clairvoyance and the nightly forces which he shares with his totemic animals, the ruler impersonates the founding ancestor. He imposes the qualities of a perennial hierarchical social organization, territorial unity, and moral order on his society. Second, the chief acts as the androgynous life-giver or mediator of the (re)generative processes between the land, society, man, and the 'the primal womb' in the earth. His rule is thus one of cosmic and physical regeneration, guaranteeing human, agricultural and social reproduction, and instituting commensality and sharing in his territory. In this diurnal ordering role, the chief protects his people from the nocturnal anti-rule of envy, theft, sexual abuse and sorcery. The elders, through councils, ceremonial exchange, and authoritative speech, extend the regenerative capacity into the daily weave of events and relations uniting kin groups under common rule.

In the Christian mission stations or in the plantation, mining, and industrial enterprises, ever larger numbers of individuals and families joined the workers camps and 'indigenous townships,' set up outside customary rule, on the fringes of White settlements. They thus created new ideological and physical spaces of identity and collective imagination. Indeed, from the last decades of the nineteenth century, Christian congregations set out to work side by side with the

concessionary companies and the colonial administration. Their aim was to 'bring civilization and universal salvation to the Dark Continent,' as well as to 'save the individual's soul' through the removal of 'pagan African customs.' While the first decades of the missionary endeavor were directed at the adaptation of Christianity to local society and culture, from the 1930s on, missions became increasingly involved in the social engineering of the townships, echoing the dominant mood of the plantation or mining trusts and the colonial administration. From then on, church authorities designed paternalist programs to educate converts towards assimilating modern skills and dispositions. By the 1950s, the term modernization had become a new banner for the assimilationist option.

In the Belgian Congo, the colonial government had secured the agreement of the missions to manage the schools in exchange for land concessions. Nearly all mission stations developed boarding schools with enough fields, cattle, poultry, and workshops to be self-sufficient. The colonial school presented Western progress as the source of a more dignified identity on the world scene. Vinck (1995) convincingly shows the extent to which, from the 1920s in the Belgian Congo, the school books conveyed the primary symbols and values of the West. By encouraging the pupils to develop a new identity and self-image, these manuals influenced the entire political élite-to-be in post-independence Congo. The concept of State, a key notion in the school books, was the comprehensive and abstract, depersonalized expression of the new power and authority which came to dominate the various spheres of life. According to the manuals, all Whites shared in the power of the State and thus were to be considered as authorities. Colonial school books depicted traditional society and religion as dominated by sorcery and the machinations of the devil.

3. *Endogenization*

From the 1950s, among the second generation of those having assimilated the European civilizational capital, the frustration grew with the glaring contradiction between the promises which the colonial master had put before them and their second-rank position in the colonial institutions. A few Christian priests or pastors and élite got imbued with the Third World emancipation movement enhanced by the rapid decolonization in Asia, the 1955 Afro–Asian Conference in Bandung, the independence struggle in Algeria, the *Négritude* call for the rehabilitation of African cultures, and the Pan-Africanism radiating from Ghana and Guinea.

The militant élite used a triple banner in the battle for decolonization. First, in the terms of the Christian, humanist, and/or socialist discourses of the colonizers, they claimed the right to dignity, social emancipation, and equality. Second, their political aspirations were largely drawing on European ideological discourse associating modern economic development with gradual political emancipation. They thereby became complicit with the modernizing and authoritarian endeavor of the White welfare—and nation–state, however alien to people at grassroots level. Third, in the terms of western Bantu traditions, they favored a palaver model of negotiation, ceaselessly co-opting 'brothers' and 'doing things with words.'

In the 1970s, authenticity movements in various states prompted a radical shift away from assimilation of the White civilizational models, now urging people to take possession in a militant way of their own history and a dignified self-representation on the world scene. Throughout the 1970s and 1980s, artists and customers in the discos celebrated a liberation from paternalist and moralizing colonization. Today, the people in both the rural and urban milieux have become aware how much their lifestyles, modes of production, and environments are both in rupture with their parental models, and simultaneously excluded from the current economic and informational globalization of western consumerist lifestyles. The masses in the poverty-stricken suburbs have been haunted, through television and downtown scenes, by the imageries of ease and extravagance that the transcontinental mass media as well as the few very rich nationals exhibit in fine clothing, expensive cars, and luxury goods. Many are increasingly bitter about their exclusion. However, they do not fail to develop their own proud visions of society and the way things are. In the early 1990s, through the waves of demonstrations and Dead-Town protest manifestations or even uprisings, suburban people counteracted the govermental views on modernization and the neo-colonial politics. Irony and parody allowed the populace to deconstruct imperialist twentieth century modernity.

Suburbanites straddle worlds through hybridity or the imaginary transgression of codes, in particular in the utopian fields of humor, daydream, and glottophagia. These are gradually furthering a cultural critique of the postcolonial situation: rather than fighting one another as ill-fortuned or discontented consumers of modern cash goods (increasingly brought in second-hand from the North), people in suburban neighborhoods and rural communities are re-exploring their genuine sense of communality, and their collective memory stored in body techniques and sensuous culture. Hybridity in many cultural expressions blurs the tradition/modernity, Bantu/western, precapitalist/capitalist oppositions that have been created by the Europeanizing tropes. Songs on the radio today in vernacular languages and along old rhythms, recalling both the collective frenzy and euphoria in the bars of the 1970s and 1980s and village festivities or rituals, are now unsettling the Reformist voices of the (post) colony which had connected city life with French or English speech and étiquette, with

school education, and the petty bourgeois life style, and with the well-equipped biomedical services. The pidgnization or creolization of French or English colloquial language, and of narrative styles in songs and newspapers, vitiate modernist master-tales about the literate African city dweller and the retrograde and illiterate villager. Thousands of independent prophetic healing churches of the holy spirit, through exorcizing materialist greed and westernization of public mores, or even State politics identified with the work of *sataani* (the local term for the Christian-imported notion of Satan, referring to those engaged in illegal practices and sorcery), develop a forceful critique of the catastrophic collusion between economic modernization and people's sociocultural dispossession. At the same time, through collective trances and 'Christian' forms of telepathy and clairvoyance in the name of the holy spirit, the prophetic churches explore ways towards domesticating modernizing forces. During the ceremonial offering of money to the prophet and assistant assembly leaders, adepts may circle more than a full hour around the congregation, dancing and singing as if to re-enchant one of modernity's most striking secularizations, anonymous cash trade. The offering aims to transform the subaltern's poverty and needs into what is defined as divine grace or healing. The prophetical churches thereby speak back to global capitalism and celebrate their participation in a new economy. They subject the monetary economy to the utopian ideals of equality and solidarity proper to brotherhood and sisterhood.

See also: African Studies: Culture; African Studies: Politics; Colonization and Colonialism, History of; Fourth World; Historiography and Historical Thought: Sub-Saharan Africa; Near Middle East/North African Studies: Economics; Near Middle East/North African Studies: Society and History

Bibliography

Andersson E 1958 *Messianic popular movements in the Lower Congo*. Almqvist and Wiksells, Uppsala, Sweden

Birmingham D 1981 *Central Africa to 1870: Zambezia, Zaïre and the South Atlantic*. Cambridge University Press, Cambridge, UK

Birmingham D, Martin P M (eds.) 1983 *History of Central Africa*, 2 Vols. Longman, London

Boahen A (ed.) 1990 *General Hiistory of Africa: VII Africa under Colonial Domination 1880–1935*. Abridged edn. Unesco, Paris and Currey, London

Cairns H 1965 *Prelude to Imperialism: British Reactions to Central African Society 1840–1890*. Routledge, London

De Boeck F 1996 Postcolonialism, power and identity: Local and global perspectives in Zaire. In: Werbner R, Ranger T (eds.) *Postcolonial Identities in Africa*. Zed, London, pp. 75–106

Devisch R 1995 Frenzy, violence, and ethical renewal in Kinshasa. *Public Culture* **7**: 593–629

Devisch R 1996 Pillaging Jesus: Healing churches and the villagisation of Kinshasa. *Africa* **66**: 555–86

Harms R 1981 *River of Wealth and Sorrow: The Central Zaire Basin in the Era of Slave and Ivory Trade, 1500–1891*. Yale University Press, New Haven, CT

Hunt N R 1999 *A Colonial Lexicon of Birth Ritual, Medicalization and Mobility in the Congo*. Duke University Press, Durham, NC

Mazrui A A, Wondji C (eds.) 1993 *General History of Africa: VIII Africa since 1935*. Heinemann, London

Mbembe A 1992 Provisional notes on the postcolony. *Africa* **62**: 3–37

Mudimbe V Y 1988 *The Invention of Africa: Gnosis, Philosophy, and the Order of Knowledge*. Indiana University Press, Bloomington, IN

Vansina J 1966 *Kingdoms of the Savanna: A History of the Central African States until European Occupation*. University of Wisconsin Press, Madison, WI

Vansina J 1990 *Paths in the Rainforests*. Currey, London

Vaughan M 1991 *Curing their Ills: Colonial Power and African Illness*. Polity Press, Cambridge, UK

Vinck H 1995 The influence of colonial ideology on schoolbooks in the Belgian Congo. *Paedagogica Historica* **31**: 355–405

R. Devisch

Central America: Sociocultural Aspects

Most scholars know that a great deal of violence took place in Central America in the late twentieth century, but relatively few know the reasons for it, what its consequences were, or how it affected research on the region. The immediate causes of the violence were revolutionary attempts from the 1960s through the 1980s which led to brutal military reprisals, mainly against civilians—in Guatemala, Nicaragua, and El Salvador. The violence had an enormous economic impact on all five countries of the region (which also includes Honduras and Costa Rica), leaving already poor people much poorer. Civilians suffered the greatest casualties and many of the survivors left the region permanently, while those who remained live in heavily militarized states where 'low-intensity' conflict continues. Yet to everyone's surprise social mobilization and protest by a wide variety of disadvantaged groups also continues. The vast social changes engendered by these phenomena have encouraged a new kind of social science research, such that most now deal with regional patterns and transnational flows, history, large institutions such as the state and military, and the nature of ongoing as well as past social movements. After a brief summary of what happened in the five countries of the region, this article will discuss recent comparative and historical research projects and will then treat the way in which ethnography changed and developed to deal with the questions raised by this period of history.

1. The Central American Revolutions

Revolutionary movements began in Guatemala, El Salvador, and Nicaragua following the 1959 Cuban revolution, which inspired them. The Central American insurgencies resembled those that took place in much of Latin America, although unlike the others, in the 1970s they became significantly more violent and threatening to their national regimes, which were in turn much weaker. Small guerrilla fronts led by relatively well-educated men (most of them members of the middle or elite classes) attempted to organize peasants or rural workers in most countries of Latin American in order to seize state power—whether or not locally inspired grassroots movements were already in place. The early guerrillas, whose basic strategy was *focismo* (Wickham-Crowley 1992) worked largely in regions where they assumed popular support could be generated to overturn the old social order for a new one, but did little to describe their actual sociopolitical projects to the masses. The *focistas* operating in the 1960s suffered ignominious defeats, but became much more successful when they renewed their efforts under changed leadership in the 1970s. Several different guerrilla groups developed in each country, differing on strategy (insurrectionary versus prolonged popular war), revolutionary subject (workers, peasants, or middle groups), and international line (Russian, Chinese, or Third World). The actual number of guerrillas in Central America was never very large (from 2,000 to 10,000), but they were supported by most of the rural population in the areas where they worked. The military forces in Central America were of variable size and sophistication—greatest in Guatemala, least in Nicaragua—but always outnumbered the guerrillas by at least ten to one. By 1980 they had developed into formidable counter-insurgency powers (everywhere but in Costa Rica), being provided with arms, advisors, and tactical support by the USA (everywhere but in Nicaragua).

Nicaragua's Sandinistas had the greatest popular support and took state power in 1979, thus becoming the only successful revolution in Latin America besides Cuba. The Sandinistas were defeated in elections ten years later; but most observers agree that the human and material cost of the *contra* war (Vilas 1995) instigated by the USA explains the defeat. The Guatemalan and Salvadoran revolutionaries came close to taking power shortly after the Nicaraguan victory, and were in a strong enough position to negotiate peace accords with the military governments of their countries in the 1990s. During the civil wars, the violence visited upon the rural civilian population by the military was extremely high and indiscriminate everywhere, but especially so in Guatemala, where Maya Indians were most heavily affected. In fact, both scholars and the UN Commission on Human Rights depicted Guatemala's military actions as genocidal. Violence continued in much of the region, but now is mostly carried out by ex-soldiers and other uprooted people, unemployed by the previous cycle of violence. Nonetheless, very significant social movements, especially of indigenous people, are now taking place throughout the region.

Carlos Vilas (1995) provides an excellent general treatment of all of the phenomena described here in fewer than 200 pages. Vilas, an Argentinian sociologist, treats the cases in greater depth than other social scientists because he lived in the region for more than ten years, was in intellectual contact with revolutionary and state sectors in all five countries, and participated in the Nicaraguan attempt to establish a revolutionary social and political regime. Vilas' earlier (1987) study of the Sandinista revolution in Nicaragua is the best general study of revolution in Central America.

2. Comparative Research

What took place in Central America calls for comparative research, especially on the question of why three of its countries had major revolutionary movements, and two did not. Many such studies have been done, but few go beyond the obvious economic differences among the five countries. Most point to the pattern of export-led economic growth since the 1960s, which impoverished peasants everywhere but in Costa Rica. (The case of Honduras is usually explained by the fact that fewer peasants were displaced there.) The most useful and original comparisons have been done by Robert Williams, an economist who uses sociological, historical, and ethnographic methods in his research. Williams's first book (1986) observes that cotton and cattle production for export expanded hugely in all five countries (with considerable economic aid from the USA for its own economic reasons) and led to significant dispossession of peasants everywhere, *including in Costa Rica*. But the five Central American states handled peasant protest quite differently, with both Costa Rica and Honduras carrying out land reform and expanding services while the three other states responded with repression and militarization—which led to war. In his second book, Williams (1994) examines the social and economic factors that led to two different kinds of states in Central America—the three revolutionary countries being controlled by rigid oligarchies, the other two being led by more open political groups. He finds an explanation in the social and political relations created by the coffee-export economy (among the owners, workers, merchants), the first major postcolonial export in the region, which played a critical role in state formation. (Paige's 1997 study, which is less complete, reaches similar conclusions.)

Timothy Wickham-Crowley (1992) compares where, and to what effect guerrilla movements arose,

and who participated as both leaders and cadre, as well as the military and regime responses provoked by guerrilla warfare. By treating all of the significant guerrilla-led movements in Latin America since (and including) the Cuban revolution, he provides a useful comparative context for the Central American guerrilla movements of both the 1960s and 1980s. Virtually no in-depth or ethnographic studies of one or more Central American guerrilla groups have been done. Such work may be forthcoming now that the wars are definitively over and the surviving guerrillas have returned to engage in peaceful politics in their countries. The Guatemalan and Salvadoran cases both cry out for research while the revolutionary protagonists are still alive, especially since Wickham-Crowley's controversial study is not actor based and relies on rather weak secondary materials.

An important six-volume comparative history has been produced (Rivas 1993), most of its contributors Central Americans. The coverage of the civil wars and their aftermath is quite limited and little new theoretical ground is broken in these volumes. What is innovative about this work is the consistent comparative emphasis, unusual for historians, which expands information on all of the countries and identifies the major gaps in knowledge. The economic and political events of the last quarter of the twentieth century clearly led to the comparative focus.

Two other comparative historical *and* ethnographic works are under way in which the Central and North American scholars are more equal in number. The more traditional project is being done only in Guatemala under the guidance of anthropologist Richard Adams, who has worked in Central America for nearly fifty years. The emphasis of the ethnographic project is on how ethnic relations were affected by the violence that took place in different communities. (In this regard it is little different from the early book edited by Robert Carmack in 1988.) The second project, taking place under the leadership of Jeffrey Gould, Charles Hale, and Dario Euraque, is more innovative. The historians and anthropologists conducting the research often worked in the same sites (some in all Central American countries except Costa Rica) and met frequently to discuss methods and ideas. The project focuses on the nature and meaning of the process of *mestizaje* (the creation of a single people through cultural and biological mixing, encouraged if not forced by nation-state building) during the past two centuries and into the present. Throughout Latin America *mestizaje* is assumed to have been a 'natural' and noncoercive process that took place mainly in the colonial period and thus had little to do with state building; North and Latin American historians have mostly repeated this myth as fact, despite strong evidence to the contrary. This Central American project takes the myth on with detailed historical and ethnographic case studies, which illuminate a great deal about the general pattern of ethnic, gender, and race relations in the region. An early glimpse into the conclusions of this project can be found in Gould's (1998) book on *mestizaje* in Nicaragua. Because of its many novel components—participation of both anthropologists and historians, North and Central American scholars, and group work in interpretation—this project promises to be a landmark study.

3. *Ethnographic Research*

The social scientists most challenged by the revolutionary situation in Central America have been anthropologists. Anthropologists have mainly worked in particular communities of Mayan Indians in Guatemala. Such work continues, but is impaired by the limited insight into state and subject formation, social movements, and transnational flows that such bounded research can produce. Those who have stayed within community boundaries have dealt mainly with the impact of fear and community divisions on local culture (e.g., Zur 1998); those who initially dealt with the larger issues (e.g., Smith 1990) lost most of the strengths of an ethnographic approach. But in recent years ethnographers have stretched themselves to do quite innovative, multisited ethnographic research.

Jennifer Schirmer's (1998) research broke new ground by providing an ethnography of military political perception and strategizing in Guatemala—where counterinsurgency tactics were best developed, as well as more violent and politically significant than elsewhere. Schirmer interviewed hundreds of Guatemalan army officers multiple times between 1986 and 1996 about everything from the nature of the guerrilla threat to their views about the proper political role of the military in the state apparatus. Criticized when her research began on the assumption that she could only parrot military views, Schirmer's published work on Guatemala's military project is now widely recognized as providing invaluable information on the formation of a military ideology, one that was nativistic (even though strongly supported by the USA) and extremely successful. Especially revelatory is how deliberately the military increased their control over Guatemalan politics, culture, and civil society. The first stage of pacification used strictly military means, including murders of suspects and massacres of civilians; the second stage involved economic and cultural restructuring of the communities most affected by the violence (the indigenous ones) together with reorganization of national party politics and elections; the final stage was reached with the 1994 peace accords, in which the military helped to reconfigure both state and civil society. As one military officer enthused, 'We [planned] the State in all of its ramifications!' (Schirmer 1998, p. 235).

The most important of the postwar social movements that have taken place in Guatemala has been

that of the indigenous Maya, who have always been the largest and most exploited group in Guatemala. They had not been politically active on their own specific causes or identity issues, however, until the late 1980s—although they certainly did participate in Guatemala's revolutionary movement. As it is presently constituted, the Maya movement is a cultural movement led by indigenous intellectuals, who take language, dress, and religious tradition to be the main (and probably safest) issues. But the political ramifications are much broader and a less well-known Maya movement is building on a more popular base, making economic and political issues central (Bastos and Camus 1993). Before the movement, the Maya majority rarely held important political offices, even where they outnumbered Ladinos. Today they play a major electoral role, are in charge of several major public institutions, and have to be taken into the political equation on all national issues. Among other things, the movement challenges North American ethnographers to take up issues of importance to the Maya (Warren 1998). An edited book on racism (Arenas et al. 1999), a major issue to Maya activists but a topic rarely addressed by anthropologists before the 1990s, fits this still small genre. Sieder (1998) edits a discussion of the Maya and other kinds of social movements—women's, human rights, and so forth.

Only Diane Nelson (1999) has attempted to characterize the Maya movement in relation to the state. This inventive work covers many different public signs of change in the cultural politics of ethnic relations in Guatemala (movies, signs, advertisements, jokes about Rigoberta Menchú) as well as real cultural politics over the past 15 years. Use of such a variety of materials allows Nelson to describe a complex social movement which has been extremely significant for the cultural reconstitution of Guatemala. In truth, Nelson treats state politics in only one chapter and on one issue, thus not producing a true ethnography of the state, but rather an entertaining postmodern pastiche of the multiple issues that surround ethnography in the Central American context—from new social movements, to new public cultures and discourses, to new kinds of reflexive positioning for activist analysts.

Closer to a true ethnography of the state is the as yet unpublished work of Finn Stepputat, a Danish cultural geographer. Over a long period he worked with several distinct indigenous groups, some of them returned refugees, as well as both Indian and Ladino state and military actors—all in one large municipality—on the implications of the violence, new social actors, and political movements of various kinds on indigenous political subjectivity. To summarize his thesis very briefly, he suggests that intensified interaction between indigenous groups and various representatives of the state (e.g., the military, Ladino representatives, and actors in foreign nongovernmental organizations) as well as social and religious leaders of various kinds, has evoked changes in indigenous subjectivity to the point that many now operate *in* the state rather than *apart from* the state. His ethnography promises to show how specific state-like operations and powers interact with those in civil society to produce modern citizens/states—a kind of ethnography of the state that will be of general interest to social science.

A very different kind of indigenous movement took place in Nicaragua during its revolution, one spearheaded by the Miskitu of the Atlantic coast, a large, remote, and politically neglected area. This much smaller group of Indians resisted the revolutionary Sandinistas, who wanted to assimilate them into the state. The dialog between the Miskitu and the Sandinistas over several years produced an agreement of autonomy for those living in the Miskitu region, but one whose economic and political features were ambiguous. Charles R Hale's (1994) analysis of Miskitu history and identity, together with the more contemporary negotiations between Miskitu of various positions, Sandinistas of various positions, and other ethnic groups in the region, is a rare ethnography of a multifaceted political movement, one that treats culture, social relations, and political consciousness as dynamic historical phenomena. Roger Lancaster (1992) writes a more typical ethnography of urban Mestizos in postrevolutionary Nicaragua, notable for its treatment of race, gender, and sexuality in the context of a ruined economy.

The only ethnography on El Salvador for any time period is that by Binford (1996) on El Mozote, the community that was eradicated by an elite Salvadoran brigade directly under the tutelage of US advisors. Binford tries to bring the dead to life in his ethnography, by piecing together the nature of their lived experiences before their community was destroyed. He also situates the destruction of this community in the complex environment of Salvadoran guerillas, military, and US support. There are literally no ethnographies written of Honduras and very little social science done on the country, even though Honduras represents an especially interesting Central American case in many respects. It has been divided between a modern, lowland banana enclave owned by the USA and a highland region backward in both economic and political development—in part because of the banana enclave; it was heavily militarized in the 1980s, yet its military governments engaged several times in significant land reform; and the peasants joined no revolutionary movements, even though they struggled for land (often successfully) for decades. The case asks for innovative ethnography on all of these features. Yet only one case of ethnographic research is now being written up (by Sarah England), which looks at another feature of life characteristic of all Central America, transnational migration. Though at least 10 percent of the Central American population are now transnational migrants, the phenomenon has been little

treated by social scientists working in any Central American country. England's research on the Honduran Garifuna describes the multiple cultural identities taken on by a black, indigenous group, who move between Honduras (where they hold an indigenous identity) and North America (where they are considered black).

The final ethnography to be treated here is Marc Edelman's (1999) study of 'peasants against globalization' in Costa Rica. Costa Rica is the model Central American country, with a small, very homogeneous population, an early democratic government whose social services rival the best in Latin America, and no significant military presence in recent history. While little other innovative social science has been done there recently, Edelman's study of Costa Rican social movements during the 1980s is a gem, possible only for someone who has worked in a country for more than a decade. Combining a sophisticated political economic analysis of global neo-liberalism with a strong political ethnography of a multitude of peasant organizations which dissolve and regroup repeatedly, Edelman's book is bound to be controversial because of its opinionated stance on a multitude of issues, including the literature on social movements and globalization. It remains, however, an example of the innovative ethnography being produced in Central America because of the global nature of its problems, the complexity of its social movements, and the peculiarities of its state-level institutions—which operate on a scale small enough to invite national-level ethnography.

See also: Dependency Theory; Revolutions, Sociology of; Revolutions, Theories of; South America: Sociocultural Aspects

Bibliography

Arenas C, Hale C R, Palma G (eds.) 1999 *Racismo en Guatemala? Abriendo debate sobre un tema tabu*. Facultad Latinoamericano de ciencias sociales, Guatemala City, Guatemala

Bastos S, Camus M 1993 *Quebrando el silencio: Organizaciones del pueblo maya y sus demandas (1986–1992)*. Facultad Latinoamericano de ciencias sociales, Guatemala City, Guatemala

Binford L 1996 *The El Mozote Massacre: Anthropology and Human Rights*. University of Arizona Press, Tucson, AZ

Carmack R 1988 *Harvest of Violence: The Mayan Indians and the Guatemalan Crisis*. University of Oklahoma Press, Norma, OK

Edelman M 1999 *Peasants Against Globalization: Rural Social Movements in Costa Rica*. Stanford University Press, Stanford, CA

Gould J L 1998 *To Die in this Way: Nicaraguan Indians and the Myth of Mestizaje, 1880–1965*. Duke University Press, Durham, NC

Hale C R 1994 *Resistance and Contradiction: Miskitu Indians and the Nicaraguan State, 1894–1987*. Stanford University Press, Stanford, CA

Lancaster R N 1992 *Life is Hard: Machismo, Danger, and the Intimacy of Power in Nicaragua*. University of California Press, Berkeley, CA

Nelson D 1999 *A Finger in the Wound: Body Politics in Quincentennial Guatemala*. University of California Press, Berkeley, CA

Paige J M 1997 *Coffee and Power: Revolution and the Rise of Democracy in Central America*. Harvard University Press, Cambridge, MA

Rivas E Torres (ed.) 1993 *Historia General de Centroamerica, Tomos I–VI*. Ediciones Siruela, Madrid, Spain

Schirmer J 1998 *The Guatemalan Military Project: A Violence Called Democracy*. University of Pennsylvania Press, Philadelphia, PA

Sieder R (ed.) 1998 *Guatemala After the Peace Accords*. Institute of Latin American Studies, London

Smith C A 1990 (ed.) *Guatemalans and the State: 1540–1988*. University of Texas Press, Austin, TX

Vilas C 1987 *Perfiles de la revolucion Sandinista: Liberacion nacional y transformaciones sociales en Centroamerica*. Editorial Legasa, Madrid, Spain

Vilas C 1995 *Between Earthquakes and Volcanoes: Market, State, and the Revolutions in Central America*. Monthly Review Press, New York

Warren K B 1998 *Indigenous Movements and their Critics: Pan-Maya Activism in Guatemala*. Princeton University Press, Princeton, NJ

Wickham-Crowley T P 1992 *Guerrillas and Revolution in Latin America: A Comparative Study of Insurgents and Regimes Since 1956*. Princeton University Press, Princeton, NJ

Williams R G 1986 *Export Agriculture and the Crisis in Central America*. University of North Carolina Press, Chapel Hill, NC

Williams R G 1994 *States and Social Evolution: Coffee and the Rise of National Governments in Central America*. University of North Carolina Press, Chapel Hill, NC

Zur J N 1998 *Violent Memories: Mayan War Widows in Guatemala*. Westview Press, Boulder, CO

C. A. Smith

Ceramics in Archaeology

Ceramics, or pottery, are among the oldest and most significant technological innovations in the history of humankind, the first truly synthetic material. Being highly plastic and thus virtually infinite in the range of shapes and forms possible, offering ready surfaces for decoration, and being ubiquitous in many archaeological contexts, ceramic containers and artifacts have provided archaeologists with one of their main categories of empirical data. Studies of variation in ceramic production, style, and use have assisted archaeologists both in the construction of chronologies and in the interpretation of ancient societies.

Ceramics denotes objects made from clay and fired to achieve hardness, which range from low-fired terracottas and earthenwares (firing range 900–1200 °C), through stonewares (1200–1350 °C), to porcelains

(1300–1450 °C). Most archaeological analyses of ceramics are focused on earthenwares, which dominated pottery industries throughout most of prehistory. Aside from vessels and containers of all kinds, ceramic artifacts also include such important items as bricks and tiles, figurines and models, and tableware.

1. History of Ceramics

The earliest human experimentation with clay—including the discovery that moist clay is plastic, can be shaped, and when heated or fired will retain its shaped form—is not well documented archaeologically. Fired clay figurines are known from the Dolní Vëstonice site in Czechoslovakia, dated to 30,000 BC. In the Old World, the first ceramic vessels appear in the early Jomon culture of Japan at about 10,000 BC, and in the Near East (Anatolia) at around 8500–8000 BC. Ceramics do not appear until considerably later in the New World, in several localities between 3000–2500 BC. In all cases, the origins of ceramic vessels appear to be closely associated with sedentary modes of life, and with storage either of agricultural products or of quantities of gathered plant foods (as in the Jomon case). There is a high degree of correlation between sedentism and pottery making, as evident in a world ethnographic sample where only two out of 46 pottery-making societies are nonsedentary (Arnold 1985).

The oldest known pottery in the world is the 'cord-marked' Jomon pottery industry of Japan, dating to 10,000 BC and associated with sedentary or semi-sedentary hunter-gatherers. Jomon pottery is highly distinctive, both with its cord or string-impressed surface decoration, and its elaborate modeled rims; vessel forms are typically high jars or beakers. Elsewhere in the Far East, pottery appears in parts of China by around 7000 BC, if not earlier. The Yangshao-Culture, centered in the Yellow River valley of northern China and dating between 4800–4200 BC, is noted for its beautifully formed earthenware jars and dishes, decorated with red-and-black painted geometric designs. In later time periods, the Chinese perfected many technical aspects of ceramic production, such as the horizontal and vertical updraft kilns, the potter's wheel, and high-fired stonewares and glazes. Porcelain ceramics were innovated in China sometime around the beginning of the first millennium AD, and were widely traded throughout the Old World, and later to Europe and the Americas.

In the Near East, architectural uses of clay (e.g., for walls, floors, and roofs) are widespread by around 7500 BC, in association with the domestication of plants and animals, and the origins of sedentary village life. Pottery containers appear by 8500–8000 BC in Anatolia, and slightly later in other parts of the Near East. A variety of other clay implements and artifacts are also characteristic of early Near Eastern sites, such as clay spindle whorls and loom weights, clay sickles (with obsidian or flint blades), clay toys and models, and clay stamps and cylinder seals. Ceramic technology was not independently developed in Europe, but rather was borrowed or diffused from the Near East, in association with the spread of agriculture and a sedentary mode of existence. In the classical world, of course, Greek black- and red-figured pottery (600–400 BC) represents a high level of technical and artistic achievement, as does slightly later Roman Arretine ware (100 BC–AD 400).

Ceramic technology in the New World arose considerably later in time, evidently as an independent development; moreover, the potter's wheel and the use of glazes never developed in the New World. Between ca. 2500–2000 BC, pottery appears in several localities, including the coast of Columbia, the Pacific coast of Mexico, and in the southeastern United States. Among the outstanding pottery traditions of the New World, one must count the coastal Peruvian Nazca and Moche traditions (200 BC to AD 700) with their elaborated modeled anthropomorphic and zoomorphic vessels; the Late Classic Maya pottery which included exquisite polychrome funerary vessels; and the Pueblo pottery traditions of the southwestern United States.

2. Archaeological Approaches to Ceramics

Ceramics have played a central role in archaeological method and theory, for several reasons: (a) pottery has a long history, and is virtually ubiquitous in most sedentary societies; (b) pottery is nonperishable, and is often recovered in very large quantities from archaeological excavations; (c) pottery functioned both as utilitarian cooking, storage, and serving vessels for all strata of a society, as well as special purpose functions for elite ceremonial or funerary use; and, (d) as a highly plastic material, pottery displays seemingly infinite variation in composition, manufacturing method, shape, and decoration. These variations, moreover, were culturally conditioned, resulting in particular ceramic styles or traditions, which can be traced through time and space.

Early in the development of modern archaeology, scholars such as Flinders Petrie in Egypt, and somewhat later James Ford, Irving Rouse, and James Griffin in North America, recognized the value of ceramic studies in developing relative chronologies based on changes in ceramic style over time. Prior to the development of such absolute chronometric dating methods as radiocarbon or dendrochronology (see *Chronology, Stratigraphy, and Dating Methods in Archaeology*), the construction of ceramic chronologies were essential to developing local and regional time frames for prehistoric cultures.

With the rise of absolute dating methods in the second half of the twentieth century, archaeological studies of ceramics have shifted somewhat, with less emphasis on classification, seriation, and chronology construction, in favor of new approaches. These include detailed physical and chemical studies of ceramic composition and of production techniques and sequences, which inform archaeologists about the technology of ceramic production and use, and about the distribution and movement of ceramic vessels during their life spans. Such information in turn is useful to archaeologists who are attempting to understand prehistoric economic and sociopolitical organization.

2.1 Classification of Ceramics

As perhaps the most plastic of all material culture media known in antiquity, ceramics exhibit enormous variation—in fabric, methods of production, shape, decoration, and function—offering tremendous advantages as well as challenges to the archaeologist. In order to bring some order to this variability, archaeologists must classify pottery into sets of like objects (see *Classification and Typology (Archaeological Systematics)*). The aims and methods of ceramic classification, however, depend greatly upon the particular archaeological approach. For example, a classification which is aimed at the discovery of historical types, those which show meaningful variation over time, will most likely emphasize different aspects of variation from a classification designed to exhibit key differences in manufacturing process. The literature on archaeological classification of ceramics is vast, but useful overviews may be found in Shepard (1965), Rice (1987), and Sinopoli (1991).

Ceramic classifications may be based on a number of different kinds of variables, the most common being technological dimensions, vessel shape, and decoration. Classifications based on technological dimensions are especially useful when the aim is to understand production processes, although such dimensions may also be useful in the definition of historical types. Major technological dimensions include raw materials (especially clay and nonplastic inclusions such as mineral temper), methods of vessel forming (such as coiling or slab building, and the use of the potter's wheel), secondary treatments (such as the application of slips or glazes), and methods of firing (open air firing, use of kilns). Vessel shape may be classified formally according to various systems, such as that developed by Shepard (1965), which distinguishes between restricted and unrestricted vessels, and further between composite and inflected shapes. Vessel shape, naturally, is closely linked with vessel function. The third major category of variation used in ceramic classification is that of decoration and surface treatment. Possible surface treatments include slipped, glazed, burnished, polished, paddle-impressed, and smoothed. Decorations may be applied by painting, stamping, incising, carving, or other methods, such as three-dimension appliqué or relief. Designs used in decorations typically consist of individual design elements, which are systematically organized into motifs and larger decorative panels, and generally follow culturally-determined rules. The classification and analysis of such design systems may be based on individual motif catalogs and set of design rules (e.g., Mead et al. 1965), or on analysis of underlying principles of symmetry (e.g., Washburn 1977).

Methods of ceramic classification also vary. Pottery producers have their own indigenous or folk classification systems, which have been the subject of considerable ethnoarchaeological study (see Sect. 2.4 below). Formal archaeological procedures for classifying pottery include the well-developed 'type-variety' system, first applied in North America and later extended to Mesoamerica. This system uses a 'binomial' nomenclature, in which a geographic name is combined with some specific technological descriptor (e.g., 'Barton incised'). Classification schemes range from simple paradigmatic classifications, to complex, hierarchical taxonomies. An alternative approach, much favored in the 1960s and 1970s, is quantitative or phenetic classification, in which 'types' are generated by a computer, following certain mathematical algorithms operating upon a number of qualitative and/or quantitative parameters. The prime example of such a numerical taxonomy of ceramics is that of Clarke (1970) for the Bell Beaker pottery complex of Great Britain.

2.2 Ceramics and Chronology

The emphasis accorded ceramic studies by archaeologists reflects the importance of using ceramic change as a means for constructing cultural chronologies. In the late nineteenth century, pioneering Egyptologist Flinders Petrie recognized that pottery vessels which had been placed in tomb groups as funerary objects showed subtle but continuous stylistic changes over time. Petrie arranged representative vessels in an inferred chronological sequence, resulting in the first seriation of pottery.

The methods of seriation were greatly refined by American archaeologists, such as Irving Rouse and James A. Ford (Ford 1962), working with ceramic assemblages from various New World localities, including the Mississippi Valley region, and the Viru Valley of Peru. Their work generated much debate concerning the methods of ceramic classification and the reality of ceramic 'types.' The fundamental principle of seriation was to define a set of historical types

which displayed gradual temporal changes, with a particular type arising at some point in time, gradually increasing in popularity (hence reflected by increased frequency in archaeological assemblages), and later decreasing until the type disappeared from the archaeological record. When the frequencies of such historical types were plotted as percentage diagrams, they displayed a characteristic frequency distribution resembling the plan of a battleship, and hence were known as 'battleship curves.'

The great advantage of seriation was that it permitted the construction of cultural chronologies independent of any 'absolute' method of direct dating. Surface collections of potsherds from sites of unknown age could be tabulated according to the frequency of key ceramic types, and then chronologically ordered by arranging the frequency distributions to form 'battleship curves.' Moreover, local ceramic chronologies could be linked together using trade pottery which occurred in more than one region, or by tracing the diffusion of particular stylistic traits. With the invention of radiocarbon dating and other methods of 'absolute' dating in the later half of the twentieth century, the importance of seriation has declined, although it is still important as a cross-check on radiocarbon-based chronologies.

Pottery may also be directly dated, either by radiocarbon dating of organic inclusions within the ceramic fabric (i.e., dung or chaff included as temper), or by thermoluminescence (TL) dating. TL dating is based on the fact that when clay and other geological inclusions in pottery are fired at temperatures of 500 °C or higher, electrons which had been 'trapped' in the crystal lattice structure are freed, emitting light or thermoluminescence. Following the firing process, new trapped electrons gradually accumulate in the crystalline imperfections of the pottery, as a natural consequence of radioactive decay. If ancient pottery is then reheated in the laboratory up to 500 °C, and the emitted light is measured and plotted as a 'glow curve' of intensity vs. temperature, the age of the specimen can be calculated, since the intensity of light emitted will be proportional to age. There are, of course, many possible complications deriving from the geological composition of the fabric, and from post-depositional conditions affecting the pottery. As a result, TL dating is less widely used than the radiocarbon method.

2.3 Compositional Studies

Within many ancient societies, pottery was produced by specialists who then traded, exchanged, or sold their wares to other nonpottery producing sectors of society, and/or to other villages or geographic localities. Moreover, pottery was frequently traded or exchanged over considerable distances. Tracing the production, distribution, and specialized use of pottery in ancient societies requires that the archaeologist be able to characterize the unique composition of a particular ceramic product, typically a mix of clay and other nonplastic inclusions (including purposefully added 'temper.') A key phase in archaeological pottery analysis is thus ceramic characterization (Bishop et al. 1982). Characterization may also include efforts at sourcing, in which the materials that make up a particular ceramic ware are traced to their geographic points of origin, such as local clay quarries or sources of sand used as temper.

A wide range of mineralogical and geo-chemical techniques have been applied to ceramic compositional analysis, whether for characterization or sourcing. Petrographic analysis of the nonplastic inclusions within a ceramic fabric, in which the specific mineral grains are identified by examining a thin section of the pottery under a polarizing microscope, is a widely used technique. X-ray diffraction, which identifies minerals by their crystalline structures, is another frequently applied method for the characterization of pottery on the basis on the temper or nonplastic inclusions. Other techniques which have been applied more recently include optical emission spectroscopy, X-ray fluorescence spectroscopy, atomic absorption spectroscopy, neutron activation analysis, proton-induced X-ray emission, Mössbauer spectroscopy, electronic microprobe analysis, and inductively-coupled plasma analysis (see Rice 1987 for a review of these and other methods).

2.4 Ethnoarchaeology and Ceramics

Given the importance of ceramics in archaeology, it is not surprising that archaeologists have turned to traditional pottery-making societies to learn more about potential variability in ceramic production, distribution, use, and discard. The study of contemporary peoples, using ethnographic methods of participant-observation, in order to gain knowledge of material culture variability which is potentially applicable to the interpretation of archaeological assemblages, is called ethnoarchaeology (see *Ethnoarchaeology*). Ceramic ethnoarchaeology (Kramer 1985) is perhaps one of them most important subfields within this topic.

Because archaeologists had long been concerned with the classification of ceramics, a number of ethnoarchaeological studies have focused on the ways in which traditional potters classified or categorized their own products. Indigenous potters typically pay little attention to the technological attributes often accorded emphasis by archaeologists (such as details of temper, paste, surface treatment, or decoration), but rather emphasize general function in their folk classifications. Thus, among the Kalinga of the Philippines, pottery vessels are distinguished by whether they are intended for cooking rice or for cooking

vegetables and meat, and by their sizes (Longacre 1981). The Fulani of Cameroon (David and Hennig 1972) lexically discriminate among five size classes of jars, with secondary classification based on their intended contents.

Other ethnoarchaeological studies of pottery have focused on aspects of ceramic production, including the social role or status of potters in their societies, on ceramic distribution, on the use of vessels, their life spans, and on their breakage and discard rates (see Kramer 1985 for a general review). Such studies have aided archaeologists in their interpretations of ancient ceramics by revealing the complex linkages between behavior and material culture, demonstrating that strictly utilitarian explanations for archaeological phenomena are not always preferable, and by showing that multiple lines of evidence may help to discriminate between alternative explanations.

See also: Art: Anthropological Aspects; Art History; Chronology, Stratigraphy, and Dating Methods in Archaeology; Classification and Typology (Archaeological Systematics); Ethnoarchaeology; Intensification and Specialization, Archaeology of; Trade and Exchange, Archaeology of

Bibliography

Arnold D E 1985 *Ceramic Theory and Culture Process*. Cambridge University Press, Cambridge, UK

Bishop R L, Rands R L, Holley G R 1982 Ceramic compositional analysis in archaeological perspective. In: Schiffer M (ed.) *Advances in Archaeological Method and Theory*, Vol. 5. Academic Press, New York

Clarke D L 1970 *Beaker Pottery of Great Britain and Ireland*. Cambridge University Press, Cambridge, UK

David N, Hennig H 1972 *The Ethnography of Pottery: A Fulani Case Seen in Archaeological Perspective*. Addison-Wesley, Reading, MA

Deetz J 1965 *The Dynamics of Stylistic Change in Arikara Ceramics*. University of Illinois Press, Chicago

Ford J A 1962 *A Quantitative Method of Deriving Cultural Chronology*. Technical Manual No. 1. Pan American Union, Washington, DC

Kramer C 1985 Ceramic ethnoarchaeology. *Annual Review of Anthropology* **14**: 77–120

Longacre W A 1981 Kalinga pottery, an ethnoarchaeological study. In: Hodder I, Issac G, Hammond N (eds.) *Pattern of the Past*. Cambridge University Press, Cambridge, UK

Mead S M, Birks L, Birks H, Shaw E 1975 *The Lapita Pottery Style of Fiji and its Associations*. Polynesian Society Memoir No. 38. Wellington (New Zealand)

Rice P M 1987 *Pottery Analysis: A Sourcebook*. University of Chicago Press, Chicago

Rye O S 1981 *Pottery Technology: Principles and Reconstruction*. Taraxacum, Inc., Washington, DC

Shepard A O 1965 *Ceramics for the Archaeologist*. Carnegie Institution of Washington Publication 609, Washington, DC

Sinopoli C M 1991 *Approaches to Archaeological Ceramics*. Plenum Press, New York

van der Leeuw S E, Pritchard A C (eds.) *The Many Dimensions of Pottery: Ceramics in Archaeology and Anthropology*. University of Amsterdam, Amsterdam

Washburn D K 1977 *A Symmetry Analysis of Upper Gila Area Ceramic Design*. Papers of the Peabody Museum of Anthropology and Archaeology, No. 68. Peabody Museum, Cambridge, UK

P. V. Kirch

Cerebellum: Associative Learning

Associative learning is behavioral change that accompanies the presentation of two or more stimuli at the same point in time or space. For many years, behavioral and neural scientists have studied associative learning in invertebrate and vertebrate species using standard classical and instrumental conditioning procedures in hopes of delineating neural circuits, brain structures, and brain systems that are involved in encoding learning and memory. In this article, the critical involvement of the cerebellum in associative learning is examined.

1. The Cerebellum and Classical Eyeblink Conditioning

Arguably the best-understood associative learning paradigm, from both behavioral and neurobiological perspectives, is classical conditioning of the eyeblink response. Briefly, a neutral stimulus such as a tone or light (the conditioned stimulus or CS) is presented just before an aversive stimulus such as a peri-orbital shock or corneal air puff (the unconditioned stimulus or US). Initially, the CS produces no overt movement while the US causes a reflexive eyeblink (the unconditioned response or UR). After 50–100 pairings of the CS and US, the CS begins to elicit a learned eyeblink (the conditioned response or CR). While most eyeblink conditioning experiments have involved rabbits as subjects, it appears that all mammals, including humans, learn this simple associative task at similar rates using similar brain circuitry.

For a variety of reasons that include (a) the relative simplicity of the response being monitored, (b) the great deal of control that the experimenter has over stimulus delivery, and (c) the precise timing of the learned response, this behavioral task has proven useful for delineating the neural circuitry involved in simple associative learning. Many experiments conducted since the early 1980s have demonstrated conclusively that the cerebellum contains a population of neurons that change their patterns of firing to encode the acquisition and performance of the classi-

cally conditioned eyeblink response—that is, the cerebellum's circuitry constitutes the essential learning and memory architecture for this basic associative learning procedure (see Woodruff-Pak and Steinmetz 2000, for review).

1.1 Lesion Experiments

The initial demonstrations of the involvement of the cerebellum in classical eyeblink conditioning were lesion experiments (e.g., McCormick and Thompson 1984). Lesions placed in the interpositus nucleus of the cerebellum prevented acquisition of eyeblink CRs and abolished previously learned CRs. Lavond et al. (1985) demonstrated the same lesion effect with infusions of kainic acid, which spared fibers of passage that course through or near the interpositus nucleus. Reversible lesions placed by cooling brain tissue (Clark and Lavond 1993) or injecting muscimol (Krupa et al. 1993) were also effective in abolishing CRs. Interestingly, when additional paired training was delivered without cooling or muscimol inactivation, the animals showed no savings in the rate of CR acquisition: they behaved as if they had received no previous paired training. These studies provide strong evidence that critical neuronal plasticity that underlies classical eyeblink conditioning occurs in the cerebellum.

Lesions of the cerebellar cortex have not produced as consistent results as interpositus nucleus lesions (e.g., Lavond and Steinmetz 1989). Cerebellar cortical lesions have reportedly caused retarded rates of CR acquisition, reduced CR amplitudes, or the appearance of mistimed CRs. These data indicate that the cerebellar cortex is involved in the conditioning process, but its precise role in conditioning or its interactions with the interpositus nucleus during conditioning are not well understood.

1.2 Recording Experiments

Electrophysiological recordings taken from cerebellar cortex and the interpositus nucleus have provided additional evidence for the involvement of the cerebellum in classical eyeblink conditioning (e.g., Berthier and Moore 1986, 1990). Recordings in regions of the cerebellum known to receive converging CS and US input have revealed neurons that discharge with patterns that seem to be encoding the conditioning process. Specifically, Purkinje cells were identified that discharged when the CS or US was presented. Other Purkinje cells either increased or decreased their rate of discharge in a pattern that seemed to be time-locked to execution of the behavioral CR. Purkinje cells that decreased their firing rates are particularly interesting because Purkinje cells are known to inhibit neurons in the deep cerebellar nuclei. Thus, a decrease in firing rate of a Purkinje cell could result in an increase in excitability of interpositus nucleus neurons, a result that is compatible with formation of a behavioral CR.

Similar to cerebellar cortex, neurons that developed discharge patterns highly correlated with CR performance were observed in the interpositus nucleus. Neurons that discharged to presentations of the CS and US were seen and, after learning, neurons that discharged in patterns that were time-locked with the behavioral CR were abundant. Interestingly, the onset of interpositus unit activity preceded the behavioral response by 30–60 milliseconds. These important observations provide strong evidence that *cellular activity in the interpositus nucleus is the neural substrate of the behavioral CR that is observed*. It is thought that CR-related activity generated in the interpositus nucleus activated neurons in the red nucleus which, in turn, activate neurons in the brainstem motor nuclei that are responsible for generating eyeblinks.

1.3 CS and US Pathways

Using stimulation, recording and lesion methods (e.g., Steinmetz et al. 1989), the putative pathways for projecting CS and US information from the periphery to the cerebellum have been delineated. It appears that a tone CS is projected from the ear to the cochlear nuclei which, in turn, relays tone information to the basilar pontine nuclei. The pontine nuclei then project information about the CS to the cerebellum along mossy fibers. On the US side, air puffs are known to activate corneal receptors that send projections to the trigeminal nucleus. The trigeminal nucleus, in turn, projects information about the US to the rostromedial portion of the dorsal accessory inferior olive. Climbing fibers that originate from the inferior olive then relay information about the occurrence of the US to the cerebellum.

1.4 The Cerebellum as an Associator

There is ample evidence from anatomy, electrophysiology, lesion, and microstimulation experiments that information concerning the occurrences of the CS and US converges on populations of neurons in the cerebellum. The leading models of the involvement of the cerebellum in classical eyeblink conditioning postulate that CS–US inputs converge in two locations—the interpositus nucleus and cerebellar cortex (e.g., Steinmetz 2000). Changes in the excitability of cortical and nuclear neurons, produced by convergent CS and US inputs, are thought to form the cellular bases for the learning and performance of the classically conditioned eyeblink response. In essence, the

cerebellar circuitry serves as an 'associator' for discrete stimuli that are presented in the environment. This idea is certainly not new. Computational neurobiologists such as Marr (1969) have long considered the architecture of the cerebellum to be ideal for associating environmental information with teaching or reinforcing inputs. Further, models such as those of Marr have hypothesized that mossy fibers and climbing fibers serve as the environmental and teaching inputs, respectively. This architecture maps very nicely onto the known neural circuitry involved in eyeblink conditioning where the CS appears to be carried along mossy fibers and the US appears to be carried along climbing fibers.

The previous models do not, however, predict a role for the deep cerebellar nuclei in the associative learning process; rather they postulate that the deep nuclei are passive recipients of outflow from cerebellar cortex. The data collected using eyeblink conditioning suggest differently. The nuclei appear to receive convergent CS–US input, the nuclei show neuronal responses that are related to conditioning, and the reversible lesion experiments of Clark and Lavond (1993) and Krupa et al. (1993) suggest that critical cellular plasticity related to conditioning occurs in the nuclei. At this time, the most parsimonious explanation of the available data suggest that critical plasticity that underlies classical eyeblink conditioning occurs both in cerebellar cortex and in the deep cerebellar nuclei. It has been suggested that these two areas may encode different features of the conditioning, with excitability changes in nuclear cells important for generating activity that drives brainstem motor neurons responsible for eyeblink CRs, and excitability in cortical cells important for providing gain on the response and for regulating the timing of the response (e.g., Gould and Steinmetz 1996, Steinmetz 2000).

2. *The Cerebellum and Other Associative Learning Paradigms*

There are surprisingly few other demonstrations of the involvement of the cerebellum in associative learning. This is likely not due to a general lack of involvement of the structure in this type of learning (although see below for some limitations on cerebellar involvement in associative learning) but rather that few experiments have been conducted to explore the involvement of the cerebellum in associative learning tasks. Two exceptions are briefly described here; adaptation of the vestibulo–ocular reflex (VOR) and instrumental signaled bar-press conditioning.

2.1 *Adaptation of the VOR*

The VOR is a brain system used to stabilize a visual image on the retina during movement. In this reflex, rotations of the head are detected by semicircular canals located in the inner ear, and the eyes are moved in their sockets in the direction opposite to the movement of the head. This reflex stabilizes the line of sight. The VOR is highly plastic as the gain of the reflex can be changed easily to accommodate changes in the strength or efficiency of the extraocular muscles to deal with changing levels of vestibular activation. In many respects, this can be considered to be an associative learning procedure as it is known that gain-setting of the VOR is dependent on two events: vestibular input from the semicircular canals and visual information (the relative slippage of the visual image on the retina during head movements) to determine if a change in VOR gain is needed. Over the years, a variety of studies have implicated the cerebellum and associated brainstem structures in adaptation of the VOR (Ito 1984, Lisberger 1988). In a similar way to classical eyeblink conditioning, it appears that neuronal plasticity that forms the basis of VOR adaptation occurs in discrete regions of the cerebellar cortex and in brainstem nuclei (the vestibular nuclei) that receive convergent input from the vestibular and visual systems.

Critical involvement of the cerebellum has also been demonstrated for an instrumental conditioning task that has some similarities to classical eyeblink conditioning (Steinmetz et al. 1993). In this task, rats are first shaped to press a bar to terminate a mild, pulsating foot-shock. After learning the bar-press response, the rats are placed in a signaled-training situation where they learn to avoid the foot-shocks by pressing the response bar during tone presentations. Rats reach 50–60 percent avoidance rates with about 10–12 days of training in this procedure. This associative task is somewhat similar to classical eyeblink conditioning in that a neutral stimulus (a tone) is used to signal an impending noxious stimulus (a mild foot-shock). Bilateral lesions of the dentate and interpositus nuclei in rats prevented learning of this avoidance response. Escape responding was initially high in lesioned rats, but this responding decreased over sessions. Interestingly, deep nuclear lesions seem to be effective in preventing avoidance learning only when the interval between tone onset and foot-shock onset is five seconds or less.

3. *When is the Cerebellum Critical for Simple Associative Learning?*

Another way to frame this question is to ask: when is the cerebellum *not* involved in associative learning? A number of experiments have addressed this issue.

First, the cerebellum appears to be necessary for associative learning when the interval of time between the stimuli being associated is relatively short. Classical eyeblink conditioning can only be obtained when the CS–US interval is 3–4 seconds or less. Longer

CS–US intervals do not produce eyeblink CRs, although a variety of other conditioned responses can be elicited. Adaptation of the VOR requires near-simultaneous occurrences of head rotation and retinal slip. As detailed above, signaled bar-press conditioning seems to be critically dependent on cerebellar function only when relatively short CS–US intervals are used. Avoidance conditioning is relatively easy to obtain with longer tone–foot-shock intervals, but cerebellar lesions appear to have no effect on these learned responses—suggesting that other brain areas are critical for active avoidance learning when intervals between the stimuli are relatively long. These observations are highly consistent with what is known about the role of the cerebellum in movement and posture—the cerebellum is intricately involved in making fine adjustments to ongoing movements during relatively brief periods of time (often referred to as movement error-correction).

Second, and related to the first point, the cerebellum seems to be critical for associative learning that involves relatively simple, discrete skeletal muscle responses. Conditioned eyeblinks, VOR adaptation, and conditioned bar-press responses require the rapid recruitment of relatively few muscles, especially when the time period allowed for responding is relatively brief. The idea that cerebellar involvement in learning may be limited by response requirements has been tested (Steinmetz et al. 1991). Rabbits were trained in two tasks: classical eyeblink conditioning, and a discriminative avoidance procedure. In the discriminative avoidance procedure, rabbits were presented with a tone and required to locomote in an activity wheel to avoid a foot-shock that was presented after the tone onset. Bilateral lesions of the interpositus nucleus of the cerebellum prevented acquisition of the conditioned eyeblink response and abolished already learned eyeblink CRs, but had no effect on the acquisition or performance of the discriminative avoidance response. While the discriminative wheel-turn avoidance task differs from eyeblink conditioning on several dimensions (e.g., it involves discrimination learning and uses a longer interstimulus interval), one of the major differences between paradigms lies in the response requirements (discrete eyeblink vs. a relatively complex, bipedal, locomotive response). Lavond and colleagues (1985) have also shown that the conditioned changes in heart-rate that are normally observed during classical eyeblink conditioning are not affected by cerebellar lesions. These data suggest that the encoding of autonomic responses during associative learning involves areas outside of the cerebellum, a finding that is compatible with a number of other studies. Together, these data suggest that the cerebellum may be involved in associative learning when relatively discrete skeletal muscle responses are conditioned.

Third, there is evidence that the cerebellum is not involved in encoding associative reward learning (i.e., learning that involves reinforcement). The rats described above that showed severe deficits in learning to avoid the signaled foot-shock after cerebellar lesions (Steinmetz et al. 1993) were trained in an appetitive version of the task. In the appetitive version of the task, a tone was presented for a 3–5 second period, and a bar press resulted in the delivery of food pellet reward. The cerebellar-lesioned rats easily learned the appetitive task even though they showed a complete inability to learn the aversively motivated task. In a more direct comparison of appetitive and aversive classical conditioning, Gibbs (1992) trained rabbits on two classical conditioning tasks before delivering lesions to the interpositus nucleus of the cerebellum. One task was standard classical eyeblink conditioning (an aversive task), while the second task was classical jaw-movement conditioning (an appetitive task). Jaw-movement conditioning involves pairing a tone CS with the delivery of water or juice (the US) into the mouth. The water or juice causes a movement of the jaw as the rabbit consumes the liquid. Several pairings of the CS and US produce an anticipatory jaw movement to the tone (the CR). Gibbs showed that in this within-subject experiment, lesions of the cerebellar deep nuclei abolished eyeblink conditioning but had no effect on jaw-movement conditioning. These data suggest that the cerebellum is involved in encoding aversively motivated associative learning but not appetitively motivated associative learning. This suggestion is not surprising given the large body of research that has detailed the involvement of forebrain structures and circuits in reward learning.

4. *The Cerebellum and Associative Learning*

For several years, theorists who have speculated about the function of the cerebellum have noted that the basic anatomy and architecture of the cerebellum seems to be designed for associative learning. The cerebellum receives inputs through two separate and unique systems of fibers: mossy fibers and climbing fibers, and a growing body of evidence suggests that plasticity in cerebellar neurons may be due to associative interactions that occur between these inputs. Further, classical eyeblink conditioning seems to be an ideal associative learning paradigm for studying the involvement of the cerebellum in associative learning. This procedure involves the conditioning of discrete responses, involves the presentation of an aversive or noxious US, and involves a relatively brief period of time between the presentation of the stimuli being associated. In essence, the classical eyeblink conditioning procedure could be considered the prototypical learning paradigm for engaging the cerebellum during associative learning. Further studies into cellular and systems-level cerebellar processes engaged during classical eyeblink conditioning should provide valuable data concerning the general role of the cerebellum in

associating two or more external stimuli or internal events in time.

See also: Cerebellum: Cognitive Functions; Classical Conditioning, Neural Basis of; Electroencephalography: Basic Principles and Applications; Electroencephalography: Clinical Applications; Eyelid Classical Conditioning; Long-term Depression (Cerebellum); Topographic Maps in the Brain; Vestibulo-ocular Reflex, Adaptation of the

Bibliography

Berthier N E, Moore J W 1986 Cerebellar Purkinje cell activity related to the classically conditioned nictitating membrane response. *Experimental Brain Research* **63**: 341–50

Berthier N E, Moore J W 1990 Activity of deep cerebellar nuclear cells during classical conditioning of nictitating membrane extension in rabbits. *Experimental Brain Research* **83**: 44–54

Clark R E, Lavond D G 1993 Reversible lesions of the red nucleus during acquisition and retention of a classically conditioned behavior in rabbits. *Behavioral Neuroscience* **107**: 264–70

Gibbs C M 1992 Divergent effects of deep cerebellar lesions on two different conditioned somatomotor responses in rabbits. *Brain Research* **585**: 395–9

Gould T J, Steinmetz J E 1996 Changes in rabbit cerebellar cortical and interpositus nucleus activity during acquisition, extinction and backward classical conditioning. *Neurobiology of Learning and Memory* **65**: 17–34

Ito M 1984 *The Cerebellum and Neural Control*. Raven Press, New York

Krupa D J, Thompson J K, Thompson R F 1993 Localization of a memory trace in the mammalian brain. *Science* **260**: 989–91

Lavond D G, Hembree T L, Thompson R F 1985 Effect of kainic acid lesions of the cerebellar interpositus nucleus on eyelid conditioning in the rabbit. *Brain Research* **326**: 179–82

Lavond D G, Lincoln J S, McCormick D A, Thompson R F 1984 Effect of bilateral lesions of the dentate and interpositus cerebellar nuclei on conditioning of heart-rate and nictitating membrane/eyelid responses in the rabbit. *Brain Research* **305**: 323–30

Lavond D G, Steinmetz J E 1989 Acquisition of classical conditioning without cerebellar cortex. *Behavioral Brain Research* **33**: 113–64

Lisberger S G 1988 The neural basis of learning of simple motor skills. *Science* **242**: 728–35

Marr D 1969 A theory of cerebellar cortex. *Journal of Physiology* **202**: 437–70

McCormick D A, Thompson R F 1984 Cerebellum: essential involvement in the classically conditioned eyelid response. *Science* **223**: 296–9

Steinmetz J E 2000 Brain substrates of classical eyeblink conditioning: A highly localized but also distributed system. *Behavioral Brain Research* **110**: 13–24

Steinmetz J E, Lavond D G, Thompson R F 1989 Classical conditioning in rabbits using pontine nucleus stimulation as a conditioned stimulus and inferior olive stimulation as an unconditioned stimulus. *Synapse* **3**: 225–33

Steinmetz J E, Logue S F, Miller D P 1993 Using signaled bar-pressing tasks to study the neural substrates of appetitive and aversive learning in rats: Behavioral manipulations and cerebellar lesions. *Behavioral Neuroscience* **107**: 941–54

Steinmetz J E, Sears L L, Gabriel M, Kubota Y, Poremba A 1991 Cerebellar interpositus nucleus lesions disrupt classical nictitating membrane conditioning but not discriminative avoidance learning in rabbits. *Behavioral Brain Research* **45**: 71–80

Woodruff-Pak D S, Steinmetz J E (eds.) 2000 *Eyeblink Classical Conditioning: Animal Models*. Kluwer, Boston

J. E. Steinmetz

Cerebellum: Cognitive Functions

In the nineteenth century, researchers reached a consensus on the basis of animal ablation experiments that damage to the cerebellum leads to motor disorders but does not affect sensory or cognitive functions. In the early twentieth century Stewart and Holmes (1904) demonstrated that cerebellar lesions resulting from tumors or gunshots elicit comparable motor deficits in humans such as reduction of muscle tone, impairment of movement coordination, and deficits in the regulation of gait and posture. Voluntary movement control is severely affected, the main symptom being a disturbance of movement coordination (ataxia) which may affect the control of limb muscles and ocular muscles as well as speech control. In the view of clinical neurology, the cerebellum was thus thought to be exclusively engaged in the control of motor activity.

In the 1950s, however, a broader concept of cerebellar function was suggested, with a possible cerebellar involvement in the control of autonomic and limbic activity. Clinical observations made an important contribution to this broader concept (for a summary see Daum and Ackermann 1995). Congenital cerebellar malformations, for example, were found to be associated not only with ataxia, balance problems, and other motor deficits, but also with mental retardation. In addition, disorders such as schizophrenia and autism were related to neuropathological abnormalities of the cerebellum. The hypothesis of a cerebellar involvement in the control of emotions was further supported by findings of a modulation of fear and aggression by cerebellar stimulation for the control of epileptic seizures.

More recent concepts of cerebellar function were influenced by reports of two-way connections between the cerebellum and the cerebral cortex, findings of cognitive deficits in patients with cerebellar dysfunction, and neuroimaging reports of cerebellar activation during a range of cognitive tasks (Schmahmann 1997).

1. Anatomy of the Cerebellum

Like the cerebrum, the cerebellum consists of two hemispheres. Three functional regions can be distinguished: the centrally located vermis (Latin 'worm') and the lateral and intermediate zones in each hemisphere. Mossy fibers are the major afferents to the cerebellum, receiving their input from brain stem nuclei and spinal chord neurons. The climbing fibers, which are the second excitatory input to the cerebellum, originate from a single site in the medulla and the inferior olivary nucleus. Efferent projections are mediated by the Purkinje cells and the deep nuclei.

There are several reciprocal pathways between the cerebellum and the cerebral cortex. The cerebellum receives input via pontine nuclei from the parietal cortex, the prefrontal cortex, and the superior temporal sulcus. The cerebellum projects back to the same regions via the thalamus (Schmahmann 1997). These afferent and efferent projections imply a possible role of the cerebellum in the modification of information which is projected from the cortex to the cerebellum and sent back to the cortex.

2. Motor Learning and Motor Imagery

It is well known that the cerebellum plays a critical role in motor control, with the lateral regions of the cerebellum mediating movement planning and programing, while the medial regions are involved in the execution of movement (Dichgans and Diener 1984).

Imaging studies using positron emission tomography have also demonstrated a cerebellar role during motor learning of sequential finger movements as well as in trajectorial learning. In addition, the cerebellum was found to contribute to the monitoring and optimizing of movements by using sensory (proprioceptive) feedback (Jueptner and Weiler 1998).

Motor skill learning describes the qualitative improvement of performance through practice which ensures that movements can be performed fast, accurately, and with little attentional control. Electrophysiological and lesion studies in nonhuman primates with cerebellar hemispherectomies have demonstrated the critical contribution of the cerebellum at a stage of motor learning when performance becomes fast and accurate. Similar results were observed in patients with cerebellar dysfunction, who had problems in learning the skillful execution of serial movements (Doyon 1997).

The most frequently used motor learning paradigm is classical conditioning of the eyeblink response. An acoustic stimulus, the conditioned stimulus (CS), is paired with a corneal airpuff, the unconditioned stimulus (US), which evokes an eyeblink. After repeated pairing of the CS and the US, the eyeblink occurs to the CS but before onset of the airpuff and thereby forms a conditioned response (CR). The essential neuronal circuitry involves the convergence of CS and US information in the cerebellum and an efferent projection from the cerebellum to motor nuclei in the brain stem, which control eyeblink responses (see Thompson 1991). Patients with cerebellar dysfunction are severely impaired at acquiring eyeblink CRs, although the reflex blink to the US is unaffected. Conditioning of simultaneously recorded nonmotor autonomic and electrocortical responses are also intact (Daum et al. 1993a). The critical involvement of the cerebellum in the conditioning of motor responses has now been documented by a large number of clinical investigations, functional neuroimaging studies, and studies of physiological manipulations of cerebellar functions in normal subjects (for a summary see Schugens et al. 2000).

In motor imagery, a motor program that is stored elsewhere in the CNS is activated without any overt movement. There is some evidence that the cerebellum becomes active during motor imagery, in tasks such as silent counting and imagination of tennis training movements (Decety et al. 1990).

3. Timing

The notion that the cerebellum computes timing requirements for motor performance is supported by investigations in animals, as well as by clinical data (Keele and Ivry 1991). Keele and Ivry have argued that the lateral regions of the cerebellum are critically involved in the internal timing of motor and nonmotor behavior which requires temporal computation. They attributed the deficits in classical conditioning of cerebellar patients to problems in timing the initiation of the CR. This hypothesis is supported by the finding of an inappropriately timed CR reported by Daum et al. (1993a) and by Topka et al. (1993). Key symptoms of cerebellar symptoms, such as dysmetria (problems with precise movements) or dysdiadochokinesia (problems with fluent alternating movements), can also be interpreted within the context of deficient cerebellar timing functions. Further support for this idea stems from impairments in rhythmic tapping in cerebellar patients (Ivry et al. 1988) as well as from deficits in speech production, which reflects a decline in temporal coordination of neuromuscular interaction needed for articulation (Ackermann and Ziegler 1992). Impairments of speech perception are also consistent with the idea of a timing deficit in the nonmotor domain (Ackermann et al. 1999).

4. Cognitive Functions

The results of developmental dysfunction, such as cerebellar agenesis (absence of the cerebellum) or cerebellar hypoplasia (prenatal developmental deficits which result in loss or incomplete cerebellar devel-

opment), varies from congenital apraxia to normal motor abilities. Similarly, cognitive development can range from profound mental retardation to normal status. Intellectual deficits after delayed motor development may be due to a close coupling of motor and intellectual functions in early life.

Performance of patients with cerebellar dysfunctions on standard intelligence tests is generally in the normal range. Short- and long-term declarative memory as well as priming effects are also largely unaffected in patients with cerebellar lesions (for a summary see Daum and Ackermann 1997).

With respect to skill learning, it has been argued that the neocortex may be primarily concerned with the generation/processing of specific operations, while the cerebellum serves to modulate and optimize the functions in question (Ito 1993). In support of this idea, cerebellar damage was associated with deficits in the automatization of visuomotor sequences and visuomotor skill learning (Doyon et al. 1998). As far as nonmotor skill learning is concerned, performance of patients with cerebellar dysfunction on standard perceptual and cognitive skill acquisition was largely unimpaired (Daum et al. 1993a, Helmuth et al. 1997). The acquisition and performance of language skills may be more difficult for such patients (Fiez et al. 1992); and cerebellar activations during such tasks may be related to verbal response search (Desmond et al. 1998).

The anatomical as well as functional relationship between the cerebellum and the prefrontal cortex (Kim et al. 1994) led to the investigation of cognitive functions that are thought to be associated with prefrontal or executive function. Performance on anticipatory planning tasks was found to be impaired in patients with cerebellar atrophy in some studies (Grafman et al. 1992, Hallett and Grafman 1997). Verbal fluency abilities are also associated with executive processing. In such tasks, subjects are asked to name as many items as possible of a certain semantic category or starting with a certain letter within a specific time limit. Patients with cerebellar damage may occasionally show problems with word generation tasks of this kind (Fiez et al. 1992). Such problems may, however, be influenced by slowing of speech ('dysarthria') which is a frequent symptom of cerebellar damage. This motor speech slowing may interfere with the execution of verbal fluency tasks, and lead to poorer performance in some cases.

While cerebellar activation is observed in functional neuroimaging during performance of the Wisconsin Card Sorting Test, a standard test of concept formation, perseverative tendencies do not usually occur in cerebellar patients (Daum and Ackermann 1997, Hallett and Grafman 1997). By contrast, cerebellar lesion patients had problems in attentional shifting between modalities (Akshoomhoff et al. 1992). This pattern might be explained by impaired cerebellar–prefrontal interaction where 'prefrontal' activation, which would be associated with changing attentional behavior, is deficient due to cerebellar dysfunction.

Two-way cerebellar-parietal projections led to the investigation of visuospatial abilities that are mediated by the parietal cortex. The studies carried out so far yielded no clear evidence of a general impairment of visuospatial functions in patients with cerebellar dysfunction. Findings of difficulties of such patients with the mental manipulation of three-dimensional objects in space offers some evidence of a visuospatial processing deficit consistent with possible dysfunction of cerebellar–parietal circuits (Wallesch and Horn 1984).

5. *Conclusion*

In summary, the cerebellum makes an important contribution to the control of voluntary movement and movement coordination as well as to the control of balance, gait, and posture. Motor learning abilities are also largely dependent upon the functional integrity of the cerebellum. There is also strong evidence for a cerebellar role as an 'internal clock' which comes into play during the control of movement as well as during perceptual processing.

The exact nature of the cerebellar involvement in cognitive processes is so far less well understood. Possible contributions to prefrontal or executive functions and visuospatial processing remain to be specified by studies using patients with selective cerebellar lesions, adequately clinical and nonclinical matched control groups, and the use of a wide range of tests assessing different aspects of the abilities in question. Functional neuroimaging techniques also provide a good tool to study the cerebellar contribution to different cognitive abilities. A problem of imaging techniques is, however, that it is difficult to determine which brain area is critically involved in which aspects of cognitive processing, since essential and correlated activity cannot be easily distinguished. A combination of imaging techniques and transcranial magnetic stimulation, which elicits a transient lesion, may be a promising approach in this regard.

See also: Habituation; Motor Control Models: Learning and Performance; Motor Cortex; Prefrontal Cortex

Bibliography

Ackermann H, Graber S, Hertrich I, Daum I 1999 Cerebellar contributions to the perception of temporal cues within the speech and nonspeech. *Brain Lang*. **67**(3): 228–41

Ackermann H, Ziegler W 1992 Die cerebelläre Dysarthrie: Eine Literaturübersicht. *Fortschr. Neurol. Psychiat*. **60**: 28–40

Akshoomhoff N, Courchesne E, Press G, Iragui V 1992 Contribution of the cerebellum to neuropsychological functioning: Evidence from a case of cerebellar degenerative disorder. *Neuropsychologia* **30**: 315–28

Daum I, Ackermann H 1995 Cerebellar contributions to cognition. *Behavioural Brain Research* **67**: 201–10
Daum I, Ackermann H 1997 Neuropsychological abnormalities in cerebellar syndromes—fact or fiction? In: Schmahmann J D (ed.) *The Cerebellum and Cognition*. Academic Press, San Diego, CA
Daum I, Ackermann H, Schugens M M, Reimold C, Dichgans J, Birbaumer N 1993a The cerebellum and cognitive functions in humans. *Behavioral Neuroscience* **107**: 411–9
Daum I, Schugens M M, Ackermann H, Lutzenberger W, Dichgans J, Birbaumer N 1993b Classical conditioning after cerebellar lesions in humans. *Behavioral Neuroscience* **107**: 748–56
Decety J, Sjöholm H, Ryding E, Sternberg G, Ingvar D H 1990 The cerebellum participates in mental activity: Tomographic measurements of regional cerebral blood flow. *Brain Research* **535**: 313–7
Desmond J E, Gabrieli J D E, Glover G H 1998 Dissociation of frontal and cerebellar activity in a cognitive task: Evidence for a distinction between selection and search. *Neuroimage* **7**: 368–76
Dichgans J, Diener H C 1984 Clinical evidence for functional compartmentalisation of the cerebellum. In: Bloedel J, Dichgans J, Precht W (eds.) *Cerebellar Functions*. Springer, Berlin, pp. 126–47
Doyon J 1997 Skill learning. Schmahmann J D (ed.) *The Cerebellum and Cognition. International Review of Neurobiology* **41:** 273–96
Doyon J, Laforce R Jr, Bouchard G, Gaudreau D, Roy J, Poirier M, Bedard P J, Bedard F, Bouchard J P 1998 Role of the striatum, cerebellum, and frontal lobes in the automatization of a repeated visuomotor sequence of movements. *Neuropsychologia* **36**: 625–41
Fiez J A, Petersen S E, Cheney M K, Raichle M E 1992 Impaired nonmotor learning and error detection associated with cerebellar damage. *Brain* **115**: 155–73
Grafman J, Litvan I, Massaquoi S, Stewart M, Sirigu A, Hallet M 1992 Cognitive planning deficits in patients with cerebellar atrophy. *Neurology* **42**: 1493–96
Hallet M, Grafman J 1997 The cerebellum and cognition: Executive function and motor skill learning. *International Review of Neurobiology* **41**: 297–323
Helmuth L L, Ivry R B, Shimizu N 1997 Preserved performance by cerebellar patients on tests of word generation, discrimination learning, and attention. *Learning and Memory* **3**(6): 456–74
Ito M 1993 Movement and thought: Identical control mechanisms by the cerebellum. *Trends in Neurosciences* **16**: 448–50
Ivry R B, Keele S W, Diener H C 1988 Dissociation of the lateral and medial cerebellum in movement timing and movement execution. *Experimental Brain Research* **73**: 167–80
Jueptner M, Weiler C 1998 A review of differences between basal ganglia and cerebellar control of movements as revealed by functional imaging studies. *Brain* **121**(8): 1437–49
Keele S W, Ivry R 1991 Does the cerebellum provide a common computation for diverse tasks. A timing hypothesis. *Annals of New York Academy of Sciences* **608**: 179–211
Kim S G, Ugurbil K, Strick P L 1994 Activation of a cerebellar output nucleus during cognitive processing. *Science* **265**: 949–51
Schugens M M, Topka H R, Daum I 2000 Eyeblink conditioning in neurological patients with motor impairments. In: Woodruff-Pak D S, Steinmetz J E (eds.) *Eyeblink Classical Conditioning: Vol. I Applications in Humans*. Kluwer Academic Publishers, Boston
Schmahmann J D (ed.) 1997 *The Cerebellum and Cognition*. International Review of Neurobiology Vol. 41
Stewart T G, Holmes G 1904 Symptomatology of cerebellar tumors: A study of forty cases. *Brain* **27**: 522–91
Thompson R F 1991 Are memory traces localized or distributed? *Neuropsychologia* **29**: 571–82
Topka H, Valls-Sole J, Massaquoi S G, Hallet M 1993 Deficit in classical conditioning in patients with cerebellar degeneration. *Brain* **116**: 961–9
Wallesch C W, Horn A 1984 Long-term effects of cerebellar pathology on cognitive functions. *Brain and Cognition* **14**: 19–25

B. Suchan and I. Daum

Cerebral Cortex: Organization and Function

If it is possible to distinguish a hierarchy of complexity in behavior, ranging from the simplest reflex to the so-called cognitive functions, and if behavior is related to the structural organization of the brain, there is no doubt that the cerebral cortex corresponds to the highest levels. The main reasons for this generally accepted notion are the following: (a) the cerebral cortex is not only the largest piece of cerebral gray matter in humans and in other mammals, but it is also the one with the most impressive system of internal connections, suggesting an essentially global operation; (b) there are good reasons to believe that most of the connections between cortical neurons are of the 'plastic' kind, i.e., they are modified through learning and thus incorporate knowledge about the world; and (c) lesions of the cerebral cortex often impair behavior in a realm that clearly belongs to the psychological level, such as language, orientation, and perception. This pre-eminence of the cerebral cortex in the control of complex behavior can be related to some of its anatomical and physiological peculiarities.

1. *The Structural Type of the Cortex*

The term cortex defines a class of brain structures characterized by an essentially two-dimensional layout. A 'vertical' organization along the 'thickness' of the cortex is repeated almost identically throughout the 'plane' of the cortex. The thickness of the cerebral cortex in humans varies between about 2 and 4 mm (in the mouse 1 mm) while in the plane the cortex of one hemisphere covers an area of about 1000 cm^2 (in the mouse about 1 cm^2). Besides the cerebral cortex of mammals, the cerebellar cortex, the optic tectum, and many other structures of vertebrate and invertebrate brains are built according to a similar two-dimensional

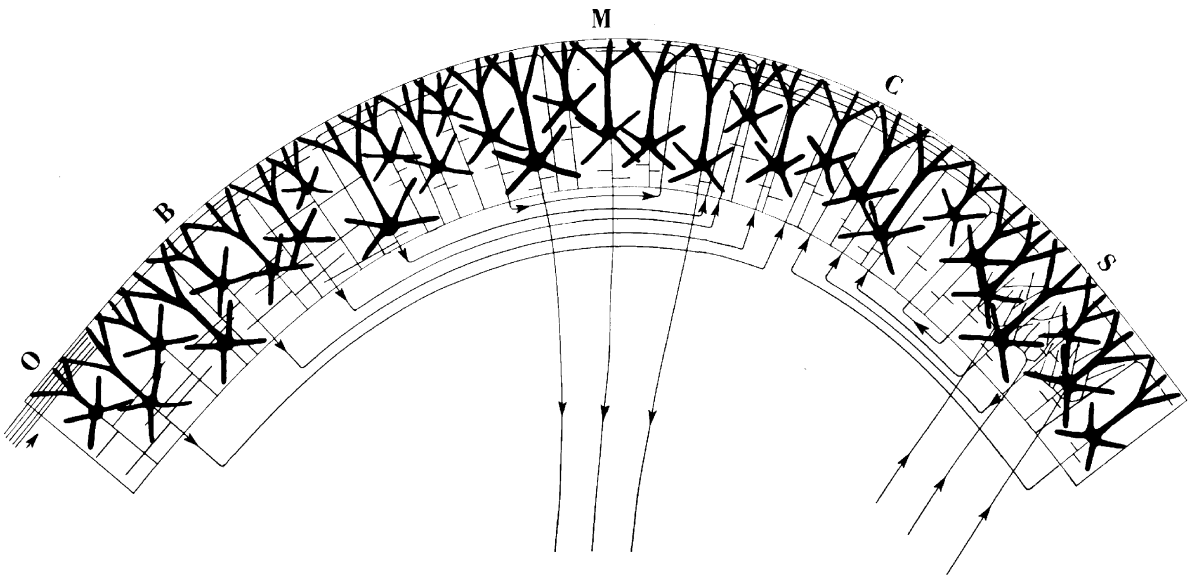

Figure 1
The basic connectivity of the cortex. Its most numerous neuronal contingent is pyramidal cells, connected into an excitatory network both locally by a rich system of axon collaterals (indicated in B) and over longer distances via the white matter beneath the cortex, terminating mainly in the upper cortical layers (indicated in C). M: primary motor cortex with axons going to the spinal cord, S: primary sensory area, receiving afferent fibers from a sensory system (via the thalamus). O: olfactory region at the edge of the cortex. From Braitenberg (1977) *On the Texture of Brains*. Springer, Berlin

scheme. What they all have in common is a diffuse projection of inputs over the whole plane, which is transformed by a unitary operation, varying from cortex to cortex, into the output. In many places the plane of the 'cortex' represents point to point some external input space such as the visual field or the surface of the body. Also, at every point of the cortical plane, output fibers leave the cortex to reach a variety of destinations, including other parts of the cortex itself. In the cerebral cortex of mammals both input and output fiber systems are on the same side of the cortical plane, forming the so-called 'hemispheric white matter' (with some minor exceptions in a marginal region, where the olfactory input enters the cortex in the uppermost layer, see olfactory input O in Fig. 1). In addition to the cortico-cortical fibers, sensory input fibers and motor output fibers shown in Fig. 1, the white matter contains a diffuse system of cortical fibers reaching the basal ganglia, as well as a system of two-way connections between cortex and thalamus.

2. *Layers*

As in other cortices, the input from distant places reaches the cerebral cortex predominantly at a special level of the cortical thickness. Similarly, the output to other parts of the brain takes its origin from another level. This distinction of input and output levels, together with the levels where most of the internal traffic of the cortex takes place, is at the origin of the well-known laminar structure of the cerebral cortex. The most common distinction is one of six layers, numbered (usually by Roman numerals) from the top down (from the free surface to the white matter). The most characteristic features of the various layers with respect to input and output are the following: the upper layers, layers I to III, are devoted to communication between distant parts of the cortex within or between hemispheres. Layer IV is the level at which sensory input fibers terminate (relayed from the thalamus), or fibers mostly from other parts of the cortex directly relaying such inputs. Layer V sends fibers to the basal ganglia and to distant parts of the brain or to the spinal cord. Layer VI communicates with the uppermost layers, as well as with the thalamus.

All of these statements are only statistically valid: a certain amount of thalamic fibers can also reach layers I, III, and VI, and layers IV to VI also participate to a certain degree in cortico-cortical communication. In spite of this, the distinction of the layers also gains support in the appearance of histological sections through the cortex. When the neural cell bodies are stained, the layers differ both in the number and the size of the neurons they contain. They also differ in the

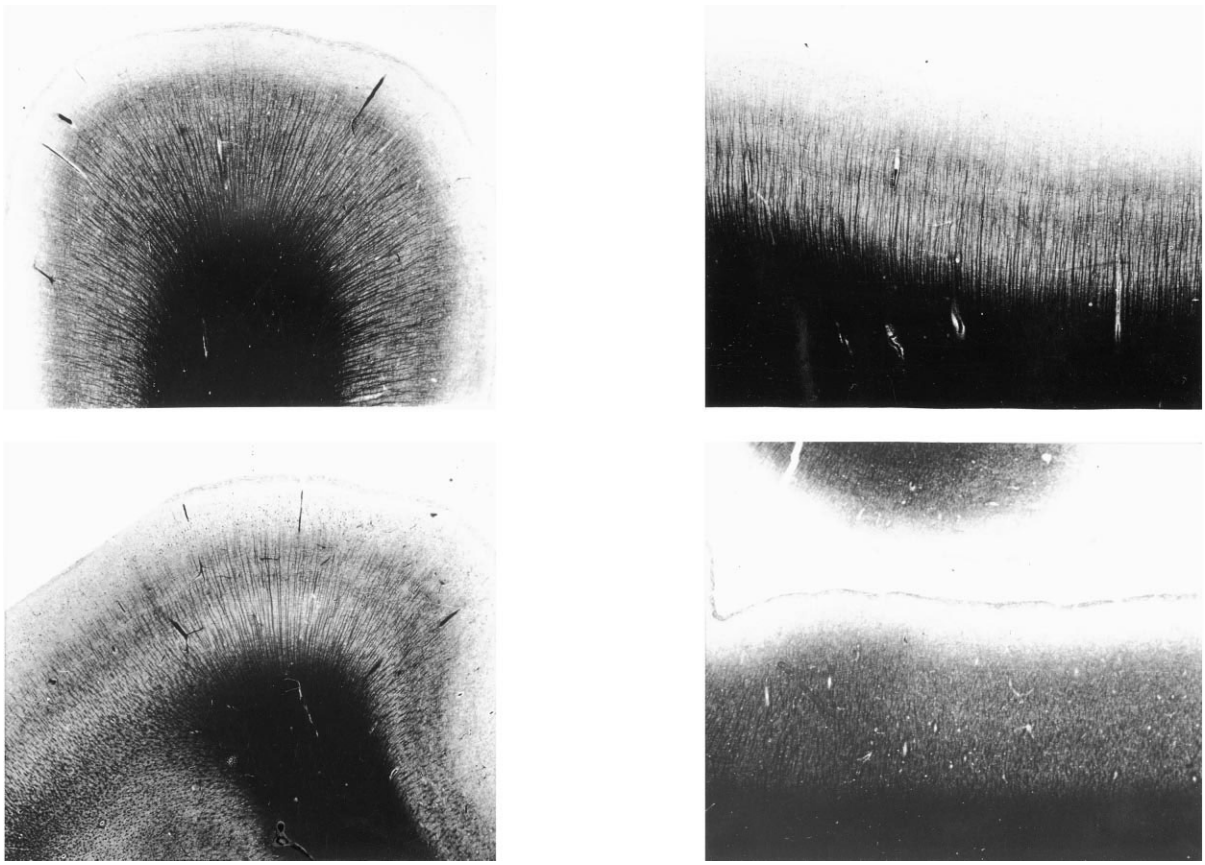

Figure 2
Examples of myeloarchitectonic differences between areas in the human cortex. Both stripes of Baillarger can be seen in the upper right (middle frontal gyrus) and lower left (Broca region), only the outer stripe can be recognized in the upper left (upper frontal gyrus) and both of them disappear in the strong myelination of the primary motor cortex (lower right). From: Braitenberg and Schüz (1989) Cortex: hohe Ordnung oder größtmögliches Durcheinander? *Spektrum der Wissenschaft* **5**: 74–86

density of fibers in myelin preparations (Fig. 2). These differences are certainly related to the different roles the layers play in cortical information handling.

The complexity of interactions both within and between layers can be appreciated by staining individual neurons in their entirety, for example with the time-honored Golgi method or with the modern techniques of intracellular injection of dyes. These methods show cortical neurons as three-dimensional devices which collect signals in the region of their dendritic trees and distribute them in the region (or regions) of their axonal terminations. The diameters of axonal and dendritic trees greatly exceed (by a factor of 10 to 100 or more) the distance between the corresponding cell bodies in the tissue. The result is an extremely dense felt of cell processes in which each neuron is interwoven with about 100,000 other neurons. Taking this into account, it follows that the borders between the layers cannot be as sharp as they sometimes appear in textbook diagrams, since they are always crossed by dendrites and axons of many neurons in adjoining, and even more distant layers (Fig. 1).

3. *Types of Neurons in the Cortex*

The pattern of axonal and dendritic ramifications varies a great deal between individual neurons, depending on their localization in different layers and in different parts of the cortex. This led to the classification of a great number of neuronal types, where one overriding distinction is now accepted by most authors, that of the spiny neurons (often subsumed under the term pyramidal cells or Type I neurons), and of the spineless neurons (often subsumed under the

term stellate or Type II neurons). This distinction is supported by differences in the fine structure of membrane specializations as they appear in the electron microscope both on the dendritic and axonal tree, and by electrophysiological findings. Briefly, spiny neurons receive most of their synaptic input on 'spines,' i.e., small processes emanating from the dendritic tree, and make 'excitatory' synapses onto other neurons via their axonal tree. Spineless neurons receive their synapses directly on their dendrites and 'inhibit' other neurons via their axonal tree. This distinction coincides with another anatomical feature: most spiny neurons have an axon which descends to the white matter and makes both, 'short-range connections' via local axon collaterals and 'long-range connections' somewhere else in the brain or—in most cases—somewhere else in the cortex (Fig. 1). In contrast, the axon of a spineless stellate cell does not enter the white matter and only contacts other neurons in its vicinity.

Spiny neurons are the great majority (about 85 percent) of all neurons in the cerebral cortex. This category largely coincides with that of pyramidal cells of older classifications, characterized by a bipartite dendritic tree, with several 'basal dendrites' distributed around the cell body, and one 'apical dendrite,' ascending vertically through the cortex and ramifying in upper layers (Fig. 1). There are, however, also spiny neurons without an apical dendrite, sometimes termed 'spiny stellate cells,' especially as recipients of primary sensory input in layer IV of primary sensory areas. Disregarding a subpopulation of these which does not project to the white matter, the basic connectivity of cortical neurons can be described as follows (Braitenberg and Schüz 1998).

4. *The Basic Connectivity of the Cerebral Cortex*

The 'skeleton cortex' (Fig. 1) consists of a large number of pyramidal cells (of the order of 10^{10} in humans), distributed throughout all layers of the cortex, and producing excitatory synapses which for the most part again contact other pyramidal cells. There is considerable divergence and convergence built into this system, since every pyramidal cell communicates with many thousands of other pyramidal cells both in its input and in its output. Thus, 'diffuse excitatory feedback' is the pre-eminent feature of the cerebral cortex. Only a small percentage of pyramidal cells project to other parts of the brain and only a few percent of the synapses in the cortex come from neurons from other parts of the brain. By far the greatest part of synaptic traffic in the cortex is internal and only involves cortical neurons.

The inhibitory stellate cells are diffusely distributed in the network of pyramidal cells. In some cases their inhibitory synapses have a strong grip on a relatively small number of cortical neurons in their vicinity (pyramidal and stellate). In other cases their inhibitory synapses are distributed more profusely, making the distinction of several classes of stellate cells possible according to the pattern of their ramification (basket cells, chandelier cells, double bouquet cells, etc.; see Peters and Jones 1984). They, too, receive their input mainly from cortical pyramidal cells. In the regions where primary sensory input (e.g., visual) enters the cortex, they may also be directly contacted by the incoming fibers.

Although the inhibitory interneurons in some cases are certainly involved in the computation which takes place in the cortical network, their role may be considered as ancillary with respect to that of the pyramidal cells. They may perhaps put a brake on neuronal activity when it threatens to explode in a runaway reaction, as may be expected in a network of elements, such as the pyramidal cells, all exciting each other. Indeed, this braking action occasionally fails, as evidenced by epileptic fits, one of the commonest forms of functional derailment in the cortex.

The comparison with other parts of the brain shows that a network of mainly excitatory connections is indeed a striking peculiarity of the cerebral cortex. In the other major parts of the brain the majority of neurons are either connected into an inhibitory network (basal ganglia) or they do not form a network among themselves but relay some input to some output in a feedforward manner (cerebellar cortex, thalamus).

5. *The Basic Function*

The question of what could possibly be the advantage of an immense network of interconnected excitatory neurons whose increase in size has accompanied mammalian evolution up to the crowning event of human culture, has received various tentative answers. The most convincing interpretation of cortical structure is based on the observation that the synapses between pyramidal cells, the majority of all synapses in the cortex, are of the special kind residing on dendritic spines, and on the supposition that spine-synapses are 'plastic,' i.e., modifiable by 'learning.' There is no definitive proof of this supposition, but enough indirect evidence to make it plausible (see review by Horner 1993), and, moreover, alternative explanations for the role of spines are less convincing. Be this as it may, there is little doubt that, of all parts of the brain, the cerebral cortex is the one most concerned with the acquisition of knowledge. Suffice it to say that lesions of the cerebral cortex impair complex acquired capacities, such as language.

In terms of engineering, the network of cortical pyramidal cells could be likened to a giant 'associative memory,' a device which connects together more strongly (through modifiable synapses) the neurons

that are often active at the same time. Thus, events of the outside world which tend to present themselves together will be represented in the brain by neurons tied together by strong synapses. The idea that the cortex incorporates knowledge by repeating, in its synaptic connectivity, the structure of reality was spelled out in two well-known theories. In 1949, D. O. Hebb proposed 'cell assemblies' as the units of cognitive operations in the brain. These are ensembles of cortical neurons connected to each other by excitatory synapses which have been strengthened in the course of a learning process. Such cell assemblies may be thought of as representing 'objects' of the world. Due to activity reverberating among its neurons, such an assembly stays active for some time once it has been activated, a property which perhaps is at the basis of 'short-term memory.' Also, a cell assembly may become active in its entirety even if only a subset of its neurons is activated initially, in a way reminiscent of the phenomenon of 'pattern completion' well known to perceptual psychologists.

'Synfire chains' (Abeles 1991) are the other theoretical proposal based on associative memory. Quite compatible with the anatomy of the cortical network and with the physiological properties of single neurons, it is possible to imagine sets of neurons, each set when activated in synchrony, activating another such set, and this in turn another one, etc., forming long chains of activity propagating through the cortex with great selectivity and temporal precision. This scheme explains how events displayed in time (e.g., the words of a language, musical themes, complex movements, etc.) are incorporated in memory. The temporal precision postulated by this theory has been impressively verified by correlation studies on spike activity of different cortical neurons (Abeles and Prut 1996).

Between Hebbian cell assemblies representing 'objects' of the world, and Abelesian synfire chains representing 'events,' there may be transitional forms of cortical activity, all based on the idea of modifiable synapses embodying the notion of synchronous events, or of events occurring in succession, acquired through a process often called 'Hebbian learning.'

6. *Cortical Areas*

From the 'macroscopic layout' of the cortical network it is also possible to gain insights into its organization. It has been known for some time that restricted lesions of the cortex produce different symptoms according to their localization. This led to the definition of 'cortical areas,' distinct regions of the cortex a few centimeters across (in humans) whose reality has since been confirmed by many detailed electrophysiological studies. There are 'sensory areas' (visual, acoustic, somatosensory, olfactory), 'motor areas,' and 'association areas.' The latter in part have been recognized as secondary and tertiary stations in the elaboration of primary sensory input, as in the case of special areas for the extraction of motion (area V5) or color (area V4) in the visual scene (Zeki 1993), or in the case of a special area (*Wernicke's area*) fed by acoustic input in the context of language, distinct from other areas concerned with acoustic perception. Evidently, besides the various sensory inputs reaching the cortex in separate regions, and the long efferent (e.g., cortico-spinal) axons emanating from other regions, it is the context in which the cortex operates locally that defines the areas.

Apart from the different effects of lesions, cortical areas were also delimited on the basis of subtle, and sometimes also quite evident, differences in the appearance of their layers and in the degree of myelination ('cortical architectonics'). For example, a small-celled layer IV is particularly well developed in primary sensory areas while it cannot be discerned at all in the primary motor area (Area 4). The primary motor area sticks out because of a population of very large cells in layer V, the cortico-spinal neurons. Large pyramidal cells in layer III are found in some association areas, including the speech centers. Two heavy bands of horizontal myelinated fibers (the 'stripes of Baillarger') are evident in some areas and only one such band in others (Fig. 2). Overall myelination tends to decrease towards higher association areas. Such structural differences were at the basis of 'cortical maps,' the best known of which, the map by K. Brodmann from 1909, distinguishes about 50 areas. Although this large number was met with skepticism at first, it is remarkable that many of the structural distinctions were later shown to correspond to different properties of neurons when cortical mapping was undertaken anew by microelectrode analysis. More recent maps tend to assume even higher numbers of areas.

Evidently, the neuronal network of the cortex, although built according to common principles throughout its extent, is subject to variations that adapt it locally to different kinds of input and to different kinds of computation, which the input is subjected to. The variations which appear in the Nissl (cell body) picture and in the myelin preparations (cyto- and myeloarchitectonics) are just two aspects of the underlying variations of the basic network, and indeed the two were shown by Hellwig (1993) to be related to each other by a set of simple rules.

7. *Columns*

Beyond the confirmation of the different nature of individual cortical areas, microelectrode neurophysiology in many cases revealed an even finer mapping within areas, the so-called 'columns.' In the primary visual area V-1 (Brodmann's area 17) certain properties to which neurons are tuned, such as responding to stimuli from the right or left eye, or to vertical,

horizontal, or oblique stripes, periodically recur over the surface of the cortex at distances of about 0.5 or 1 mm (Hubel 1988). The term 'column' was chosen for these small compartments within the areas since the locally specific properties of neurons in one column tend to be similar for neurons in all, or in several layers. Columns (or slabs) of neurons with similar properties, but different in a systematic way from column to column have also been described in secondary visual areas, in the somatosensory cortex, and in the primary acoustic cortex. In many cases, these functionally defined columns can be attributed to the way the input is organized: inputs from different sources tend to enter the cortex in alternating bundles (e.g., right eye vs. left eye in the visual cortex, fibers from the same or the opposite hemisphere in the auditory and prefrontal cortex), thus imposing periodical inputs onto a largely homogeneous looking intracortical network. However, periodicity can also be found in the pattern of axonal arborizations of pyramidal cells within cortical areas where they preferably project to columns with similar functional properties (Malach et al. 1997, Yoshioka et al. 1996). Occasionally, columns may also be visible in the arrangement of cell bodies and dendritic trees (whisker representation in rodents), or as the 'cytochrome oxidase blobs' of the visual cortex. However, neither areas nor columns can be regarded as separate entities; the fiber felt is continuous across the borders between them.

8. *Connections Between Areas*

The reality of cortical areas is also evident in the pattern of connections between them. This can be studied (in animals) by injecting certain dyes locally into the cortex, which are then transported in axons either in the direction away from the cell body ('anterograde transport') or towards it ('retrograde transport') depending on the particular dye chosen. The overall picture is the following (Young et al. 1995): individual areas are connected to several other areas, but not to all. Areas can differ considerably with respect to the number of areas they are connected to. Some primary sensory areas are connected to only a few other areas; higher association areas can be connected to more than one third of all cortical areas. The majority of the connections between areas are reciprocal. In the monkey cortex, if an area A projects to another area B, there is also a projection from B to A in 82 percent of the cases. Areas that are located close to each other in the cortex, often related to the same sensory modality, are more likely to be connected, but there are also connections between areas quite far apart. Most areas are connected to their symmetrical partners in the other hemisphere. The exceptions are in parts of the visual and somatosensory areas, which do not contribute fibers to the *corpus callosum*, the main interhemispheric fiber bundle.

The strong internal connectedness of the cortex in each hemisphere is reflected also in the large volume of the hemispheric white substance (the cerebellar hemispheres by comparison have a very scanty white substance, containing only afferent and efferent fibers). In the human brain and in the brains of other large mammals, within the white matter underlying the cortex it is possible to discern about six large bundles connecting distant parts of the cortex, such as the occipito-frontal fascicle or the arcuate fascicle between the motor and the sensory speech areas. The rest of the white matter is composed of shorter fibers between neighboring areas, and to a lesser extent, of commissural fibers, connecting both hemispheres via the *corpus callosum*, and of the fibers connecting the cortex, both ways, with subcortical centers. The fact that the volume occupied by the long cortico-cortical bundles is a small fraction of the total volume of the white substance suggests that there is a hierarchical organization, with a great amount of preprocessing between neighboring areas, often within one modality, preceding the global integration.

See also: Brain Asymmetry; Brain, Evolution of; Brain: Response to Enrichment; Cingulate Cortex; Electrical Stimulation of the Brain; Functional Brain Imaging; Long-term Potentiation and Depression (Cortex); Motor Cortex; Neural Plasticity; Neural Plasticity in Auditory Cortex; Neural Plasticity in Visual Cortex; Prefrontal Cortex; Pre-motor Cortex; Split Brain; Topographic Maps in the Brain; Visual System in the Brain

Bibliography

Abeles M 1991 *Corticonics: Neural Circuits of the Cerebral Cortex*. Cambridge University Press, Cambridge, UK

Abeles M, Prut Y 1996 Spatio-temporal firing patterns in the frontal cortex of behaving monkeys. *Journal of Physiology* **90**: 249–50

Braitenberg V, Schüz A 1998 *Cortex: Statistics and Geometry of Neuronal Connectivity*, 2nd edn. *Anatomy of the Cortex: Statistics and Geometry* 1991 Springer, Berlin, Heidelberg, New York

Hebb D O 1949/1961 *Organization of Behavior. A Neuropsychological Theory*, 2nd edn. Wiley & Sons, New York

Hellwig B 1993 How the myelin picture of the human cerebral cortex can be computed from cytoarchitectural data. A bridge between von Economo and Vogt. *Journal of Brain Research* **34**(3): 387–402

Horner C H 1993 Plasticity of the dendritic spine. *Progress in Neurobiology* **41**: 281–321

Hubel D H 1988 *Eye, Brain and Vision*. The Scientific American Library, New York

Malach R, Schirman T D, Harel M, Tootell R B H, Malonek D 1997 Organization of intrinsic connections in owl monkey area MT. *Cerebral Cortex* **7**: 386–93

Peters A, Jones E G (eds.) 1984 *Cerebral Cortex, Cellular Components of the Cerebral Cortex*. Plenum Press, New York, London, Vol. 1

Yoshioka T, Blasdel G G, Levitt J B, Lund J S 1996 Relation between patterns of intrinsic lateral connectivity, ocular dominance, and cytochrome oxidase-reactive regions in macaque monkey striate cortex. *Cerebral Cortex* **6**: 297–310
Young M P, Scannell J W, Burns G 1995 *The Analysis of Cortical Connectivity*. Springer, Berlin, Heidelberg, New York
Zeki S 1993 *A Vision of the Brain*. Blackwell Scientific Publications, London, Edinburgh

A. Schüz and V. Braitenberg

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Change: Methods of Studying

The themes of stability and change permeate from the Greek philosophers to the present. Following Heraklit, change seems to be a ubiquitous phenomenon: 'all is flowing (*Panta rei*)' or 'you cannot get into the same river twice.' The opposite position was held by Parmenides, who said, 'something is real only if we can say it *is*, if it was or will be, it is not real. Change is an illusion.' Stability and change are basic concepts within psychology, too. The aim of this article is to give an overview regarding methods of studying change. Therefore, in the first step, important concepts of change and stability are introduced. For each of the different concepts of stability, change is defined as absence of stability.

1. *Concepts of Change and a System of Categories to Differentiate Them*

1.1 *Basic Concepts of Change*

Traditionally, there are at least two important basic concepts of stability and change: absolute and normative stability (Baltes et al. 1977). Whereas normative stability is only defined at the group level, absolute stability can be defined at the individual *and* the group level. A definition of absolute stability for an individual is that the value of a variable for this individual does not change between occasions of measurement. Absolute stability at the group level is defined as the mean level of the group for a variable does not change between observations. Figure 1(c) shows absolute stability at the individual level for individual 3 but not for individual 1. At the same time there is stability at the group level, because the mean of the whole sample does not change. Figure 1(a) describes change in absolute stability for each individual and at the group level.

Normative stability is given if the rank order with respect to a variable does not change between two occasions of measurement. Imagine, for instance, that the variable control belief is measured on two occasions. Each individual starts from a different level of control belief and then shows an increase by the same amount (cf., Fig. 1(a)). The rank correlation is 1 and, therefore, normative stability is given. Figure 1(c) illustrates change with respect to normative or rank-order stability: the rank order is reversed from time 1 to time 2. In this sample, the stability coefficient measured as rank correlation between occasions is $r = -1$.

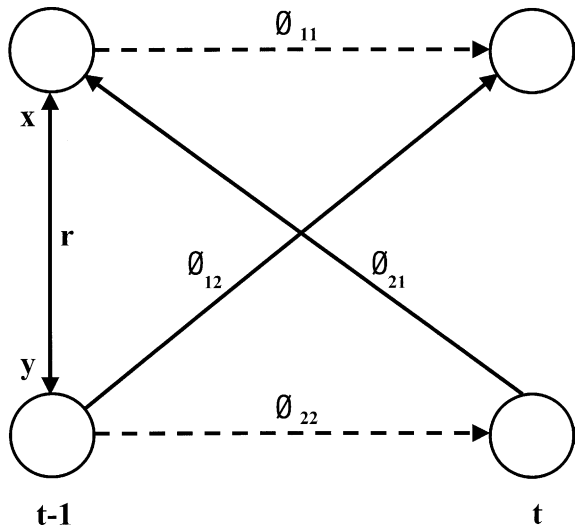

Figure 1
(a) Concepts for univariate stability/change: change of absolute value (individual and group level) but normative stability; (b) change of variability but normative stability; (c) change of rank order but absolute stability (group level)

Another concept of stability/change is stability of variability. Stability of variability requires that the variance of a variable does not change between two occasions of measurement. Figure 1(b) shows an increase in the variability across situations. The example in Fig. 1(b) demonstrates that change in the variability across occasions can occur despite absolute and normative stability. To conclude, these examples demonstrate that there exist different concepts of stability/change and that stability can be given for one concept, but not necessarily for the others.

1.2 *A Classification System for Stability/Change Concepts*

To gain an even fuller picture of change concepts, a classification system of change concepts will be outlined, which takes into account the following:

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

(a) individual change or change within groups;
(b) number of variables: univariate or multivariate change;
(c) number of and time distance between occasions: two or multiple measurements;
(d) scaling of the variables: categorial, ordinal, interval scale.

For reasons of economy, other important dimensions are not (explicitly) included within this classification:

(e) change can be located at the surface or the construct level;
(f) (measured variable or latent construct; phenotype or genotype);
(g) trend, growth, and rhythm;
(h) change of the (dimensional) structure;
(i) change in measurement error;
(j) synchronicity, asynchronicity;
(k) change and causality.

2. *Examples of Important Methods for Studying Change*

When outlining a system of categories for change concepts, it becomes evident that by combining the different dimensions of the system a large number of change concepts could be derived. Within this short description, however, only a subset can be explained in more detail.

2.1 *Analyzing Change by Using Differences Between Two Occasions*

This way of studying change is frequently used in research. The way of measuring change shown in Fig. 1(a) is based on comparing measures of one variable for a sample of individuals for two occasions. The question is whether there is a change in the mean of the variable. To analyze these questions, one computes differences between time-2 and time-1 measures. Critiques of this approach have pointed to various problems:

(a) differences are dependent on the level of the first measurement;
(b) differences are not reliable (not as much as the original variable);
(c) regression to the mean leads to systematic biases.

(a) The law of initial values states that high initial values correlate with small differences between second and first measurement and vice versa, which means that there is a negative correlation between initial status and change.

(b) From the assumptions of classical test theory, e.g., Lord (1963) derived an equation for the reliability of differences. He argued that if there are two fallible measures, the resulting measure must be worse. This means that if there is a high correlation between the first and second measurements and the reliabilities of the single measurements are medium or high, then the reliability of the difference is low (numerical example: if $r_{11} = r_{22} = 0.84$ and $r_{12} = 0.83$, then it follows that $r_{dd} = 0.06$).

(c) Regression to the mean says that extreme values for the first occasion will probably be followed by more average values in the second measurement. The mathematical formulation for regression to the mean is $s(y_{pred}) < s(x)$ (y_{pred} are the predicted values). It can be shown easily that regression to the mean is a simple mathematical tautology depending on the assumption of equal variances for pre- and post-test.

For each of these points of critique, Rogosa (1995) gives examples that they do not hold in general and that they can be subsumed under the myths of longitudinal research. Therefore, difference scores can be used if there are only two occasion measurements.

2.2 *Cross-lagged-panel Analysis*

After studying change using difference scores, we turn to the study of causal relations using change measurements. Consider, for instance, the theoretical question of whether there exists a causal relationship between control belief and academic performance and vice versa. The empirical basis for computing estimates of cross-lagged-panel relationships is that a sample of individuals is measured (at least) twice with respect to (at least) two variables.

The cross-lagged-panel model contains two important cross-lagged paths (a) between control belief (time 1) and performance (time 2) and (b) between performance (time 1) and control belief (time 2). These models can be estimated and tested for significance, yielding possible causal (in the sense of time-lagged relationships; see Schmitz 1990) relationships: no causality, unidirectional causality, or bidirectional causality (feedback).

2.3 *Analysis of Time Series*

The data basis for analysis of time series is that at least one individual is studied for at least one variable, using many occasions of measurement. Whereas the methods in Sect. 2.1 require only a small database (two occasions measurements), for time series analysis one often needs more than 20 or even 50 data points. One might think that it would be difficult to collect such data, but in psychophysiology, behavioral observations, and diary approaches, it is feasible to collect this kind of data. One of the great advantages of the time-series approach is that one can test hypotheses also for

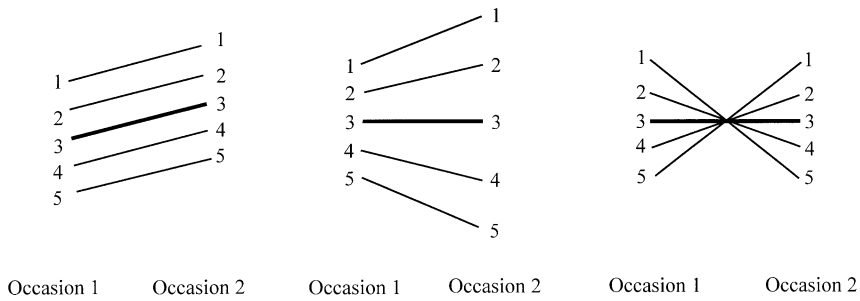

Figure 2
Cross-lagged relationships for two variables

a single individual. A typical research question may be to study the effect of an intervention (e.g., does a training have an effect on control belief?) and the identification of trends and rhythms. A variable shows a linear trend if the increase from one occasion to the next is constant. An example of a variable which shows a simple rhythm is workload measured daily: usually, the workload is high during workdays whereas at weekends people will not work. In the bi- or multivariate case one can test synchronous or asynchronous relationships (see Schmitz and Skinner 1993). Synchronous relationships occur if variables show similar patterns over time (e.g., days); they show asynchronous relationships, e.g., if one variable lags behind another variable (shows a similar pattern just one day later). One can analyze, for instance, *intra-individual* causal relationships between variables. If there are more individuals, one can perform intra-individual time-series analyzes for each individual and then combine the individual parameters at the aggregate level, as is done in hierarchical linear modeling (HLM) (Bryk and Raudenbush 1987) or in the approach proposed by Schmitz et al. (1996). Both methods can be regarded as solutions to the idiographic–nomothetic controversy. (See Fig. 2.)

3. *Other Methods and Outlook*

A special case of the general time-series models are chaos models, which are nonlinear.

One characteristic of chaos is the severe dependence on the initial conditions, which is often referred to as the butterfly effect. There are simple chaotic systems which can be described by only one so-called order parameter. These systems converge to homeostasis if the order parameter stays within certain limits. However, if the parameter changes slightly the system shows chaotic behavior (Alligood et al. 1997).

Sometimes the assumption of interval data is not fulfilled, and in these cases synchronicity and asynchronicity can be studied using Markov models (see Gottman and Roy 1990). If information is given on how long it takes until an event occurs, methods of event-history analyzes can be applied (Willett and Singer 1995).

In sum, methods of studying change based on poor information—as in two-occasion measurements—can only lead to poor results (one cannot derive conclusions that are valid for individuals), whereas methods which use the information contained in rich (e.g., time-series) data can lead to rich conclusions (e.g., which are also valid for individuals). If someone is really interested in stability and change (although it requires a great deal of effort to conduct a longitudinal study), one should not always apply only simple methods (such as differences) and small data basis (such as two measurement points) for economical reasons as there are alternative methods which may provide more differentiated results. To conclude, it is time to avoid shortcomings of studying change or, as Rogosa (1995 p. 8) put it, 'Myth 1: two observations a longitudinal study make.'

See also: Chaos Theory; Event-history Analysis in Continuous Time; Event Sequence Analysis: Methods; Longitudinal Data; Longitudinal Data: Event–History Analysis in Discrete Time; Longitudinal Research: Panel Retention; Markov Models and Social Analysis; Markov Processes for Knowledge Spaces; Time Series: General

Bibliography

Alligood K T, Sauer T D, Yorke J A 1997 *Chaos. An Introduction to Dynamical Systems*. Springer, New York

Baltes P B, Reese H W, Nesselroade J R 1977 *Life-span Developmental Psychology: Introduction to Research Methods*. Brooks/Cole, Monterey, CA

Bryk A S, Raudenbush S W 1987 Application of hierarchical linear models to assessing change. *Psychological Bulletin* **101**: 147–58

Gottman J M, Roy A K 1990 *Sequential Analysis. A Guide for Behavioral Researchers*. Cambridge University Press, Cambridge, UK

Lord F M 1963 Elementary models for measuring change. In: Harris C W (ed.) *Problems in Measuring Change*. University of Wisconsin Press, Madison, WI, pp. 21–38

Rogosa D R 1995 Myths and methods: 'myths about longitudinal research' plus supplemental questions. In: Gottman J M (ed.) *The Analysis of Change*. Erlbaum, Mahwah, NJ, pp. 3–66

Schmitz B 1990 Univariate and multivariate time-series models: the analysis of intraindividual variability and intraindividual relationships. In: von Eye A (ed.) *Statistical Methods in Longitudinal Research, Volume II: Time Series and Categorical Longitudinal Data*. Academic Press, Boston, pp. 351–86

Schmitz B, Skinner E 1993 Perceived control, effort, and academic performance: interindividual, intraindividual, and multivariate time-series analyses. *Journal of Personality and Social Psychology* **64**: 1010–28

Schmitz B, Stanat P, Sang F, Tasche K G 1996 Reactive effects of a survey on the television viewing behavior of a telemetric television audience panel: a combined time-series and control group analysis. *Evaluation Review* **20**: 204–29

Willett J B, Singer J D 1995 Investigating onset, cessation, relapse, and recovery: using discrete-time survival analysis to examine the occurrence and timing of critical events. In: Gottman J M (ed.) *The Analysis of Change*. Erlbaum, Mahwah, NJ, pp. 203–59

B. Schmitz

Chaos Theory

Chaos theory is the study of deterministic difference (differential) equations that display sensitive dependence upon initial conditions (SDIC) in such a way as to generate time paths that look random.

This article (a) explains what 'chaos' is in mathematics, (b) explains why economists became interested in this concept, (c) sketches economic forces that can 'smooth' out dynamical irregularities that can lead to chaos, (d) very briefly discusses theoretical models that generate chaotic dynamical systems as equilibria, and (e) discusses empirical testing for chaos. We shall spend most of the time on (e) because even though the vast bulk of studies has not found convincing evidence for chaotic dynamics that are short term predictable (out-of-sample) using nonlinear methods with high ability to detect chaos, the quest for chaos has generated useful statistical methods.

1. Chaos

The mathematical apparatus for rigorous treatment of chaos can be quite intimidating to the nonmathematician. Therefore we shall attempt to explain the concepts verbally as much as possible. This expositional strategy will allow the reader to decide whether it is worth their while to investigate this area further. Reference to wide-ranging surveys are given.

We follow Brock (1986) for a very brief treatment of some mathematics of chaos below. A deterministic discrete time dynamical system on n-dimensional space is a system,

$$\boldsymbol{x}(t+1) = F(\boldsymbol{x}(t)), \quad t = 1, 2, \ldots,; \ \boldsymbol{x}(0) = \boldsymbol{x}_0 \quad (1)$$

of n difference equations in n variables. Here F is a map from n-dimensional real space to itself, $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t+1)$ are n-dimensional vectors at date t and $t+1$ respectively, $\boldsymbol{x}_0$ denotes the initial condition vector, and t is time.

A useful pedagogical example is the scalar 'tent map' T from the closed interval [0, 1] to itself defined by

$$\boldsymbol{x}(t+1) = T(\boldsymbol{x}(t))$$

where

$$T(\boldsymbol{x}) = 2\boldsymbol{x}, \ \boldsymbol{x} \text{ in } [0, 0.5]$$

and

$$T(\boldsymbol{x}) = 2-2\boldsymbol{x}, \ \boldsymbol{x} \text{ in } [0.5, 1] \quad (2)$$

The hallmark of chaos is SDIC, i.e. for two nearby initial conditions, $\boldsymbol{x}(0)$, $\boldsymbol{y}(0)$, the map F magnifies the distance between them:

$$|F(\boldsymbol{x}(0)) - F(\boldsymbol{y}(0))| > |\boldsymbol{x}(0) - \boldsymbol{y}(0)| \quad \text{for } |\boldsymbol{x}(0) - \boldsymbol{y}(0)| \text{ small enough} \quad (3)$$

for most pairs of nearby initial conditions. Here $|\cdot|$ denotes a distance measure. The tent map (2) illustrates this idea nicely because $|\boldsymbol{x}(0) - \boldsymbol{y}(0)|$ is magnified by a factor of two for small enough $|\boldsymbol{x}(0) - \boldsymbol{y}(0)|$ except for the rare case where $\boldsymbol{x}(0)$, $\boldsymbol{y}(0)$ straddle the maximizer, $x = 0.5$.

A popular way to precisely capture the idea of SDIC and to measure it is the largest Lyapunov exponent, L, which is defined by the following limit as h tends to infinity,

$$L(\boldsymbol{x}(0), \boldsymbol{v}) = \text{Limit}(1/h) \ln\{|DF^h(\boldsymbol{x}(0)) \cdot \boldsymbol{v}|\} \quad (4)$$

Here DG denotes the n by n derivative matrix of map G, $F^h(\boldsymbol{x}(0))$ denotes the application of map F to the n-vector $\boldsymbol{x}(0)$ h times, $\boldsymbol{v}$ denotes a nonzero n-dimensional direction vector, '$\cdot$e denotes dot product, $|\cdot|$ denotes the Euclidean norm on n-dimensional space, and ln denotes natural logarithm.

Sufficient conditions are available for $L(\boldsymbol{x}(0), \boldsymbol{v})$ to exist and to be independent of $\boldsymbol{x}(0)$, $\boldsymbol{v}$ for most initial conditions $\boldsymbol{x}(0)$ and most vectors $\boldsymbol{v}$. A popular definition of chaos is:

Definition (Chaos) The map F is said to be chaotic if $L > 0$.

This is not the only definition of chaotic map in the literature, but it is a popular one and we shall use it here.

The value of L for the tent map T is $L = \ln(2) > 0$ so T is chaotic by this definition of chaos. As we saw before, T displays SDIC. The quantity L measures how fast (on average per iteration) a tiny measurement error (captured by $\boldsymbol{y}$) in the initial condition $\boldsymbol{x}(0)$ is magnified by the map. If the iteration h in (2) is thought of as a forecast then $|DF^h(\boldsymbol{x}(0)) \cdot \boldsymbol{v}|$ represents the error in an h-horizon forecast caused by measurement error v at date zero. If $L > 0$, this error is growing exponentially as h increases by a factor $\exp(L)$. This kind of behavior is associated with deterministic maps generating random-looking time series output. Clearly the tent map example generates random-looking time series output. Another example of a type of chaotic map is a psuedo-random number generator for computers.

If the solution $\{\boldsymbol{x}(t), t = 1, 2, \ldots\}$ to (1) is bounded for each initial condition, x_0, under regularity conditions the limiting behavior of chaotic dynamical systems is contained in a set A that is invariant under application of F which is called a 'strange attractor.' It is called 'strange' because it is not a rest point or a p-cycle. Here a p-cycle is a collection of p vectors, $\{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(p)\}$, such that

$$\begin{aligned} &\boldsymbol{x}(2) = F(\boldsymbol{x}(1)),\ \boldsymbol{x}(3) = F(x(2)), \ldots, \\ &\boldsymbol{x}(p) = F(\boldsymbol{x}(p-1)),\ \boldsymbol{x}(1) = F(\boldsymbol{x}(p)) \end{aligned} \tag{5}$$

A rest point is a p-cycle where $p = 1$.

It is useful to briefly explain the notion of bifurcations and 'routes to chaos.' Consider a dynamical system with 'fast' and 'slow' variables, $\boldsymbol{x}(t)$, and $\boldsymbol{a}(t)$ where the fast variables $\boldsymbol{x}(t)$ have a rate of change that is much faster than the slow variables. Write

$$\begin{aligned} &\boldsymbol{x}(t+1) - \boldsymbol{x}(t) = f(\boldsymbol{x}(t), \boldsymbol{a}(t)), \\ &\boldsymbol{a}(t+1) - \boldsymbol{a}(t) = \mathrm{eg}(\boldsymbol{x}(t), \boldsymbol{a}(t)),\ 0 < e \ll 1, \end{aligned} \tag{6}$$

where '$\ll 1$' means 'much less than one' to capture the idea that the second difference equation in (6) moves much more slowly than the first. Hence, it is useful sometimes to assume the fast variables have already converged to an attractor conditional on the value of the slow variables. A bifurcation value is a value of the slow variables such that passing through it leads to an abrupt change of the attractor of the fast variables. For example suppose the attractor is a rest point which abruptly changes to a p-cycle, $p > 1$; or a p-cycle which abruptly changes a more complex attractor. There is a classification theory for bifurcations (Kuznetsov (1995)).

A well-studied route to chaos in one dimensional discrete dynamical systems is the Feigenbaum cascade (and closely related Sharkovsky ordering). Here a rest point (i.e., a one-cycle) bifurcates into a two-cycle, followed by bifurcation into a four-cycle, followed by bifurcation into an eight-cycle, … , followed by bifurcation into a $2m$-cycle, … to fully developed chaos as a slow 'tuning' parameter increases. This particular route to chaos is used a lot in economics (Benhabib 1992). Some economists argue that fully developed chaos is not as useful to economic science as the analysis and classification of bifurcations themselves, because the evidence from data is stronger for the existence of bifurcations.

2. *Chaotic Economic Dynamics?*

Even though predictability of a chaotic dynamics is futile in the long term, nonlinear prediction methods can do a good job on short term prediction when chaos is present. Consider a general stylized dynamic economic model represented by the following stochastic dynamical system

$$\boldsymbol{x}(t+1) = f(R(\boldsymbol{x}(t)),\quad \boldsymbol{x}(t)) + s \cdot n(t) \tag{7}$$

where $\boldsymbol{x}(t)$ is a vector of economic quantities that represent the state vector of the economy at date t, $R(\boldsymbol{x}(t))$ represents the response of economic agents to the state of the economy at date t, and the function f produces the new state vector $\boldsymbol{x}(t+1)$ at date $t+1$. Here $n(t)$ denotes a stochastic process, called 'forcing noise,' (often chosen to be an independent and identically distributed sequence of mean zero, finite variance random variables) which represents outside shocks to the economic system and s denotes standard deviation of each random variable $n(t)$.

Many economic models (e.g., many of the models treated in Benhabib 1992) can be put into the mathematical form of Eqn.(7). If we define

$$F(\boldsymbol{x}) = f(R(\boldsymbol{x}), x) \tag{8}$$

then Eqn.(1) is a difference equation of the form (1) when $s = 0$. Since economic theory generates nonlinear dynamics it is theoretically easy to produce economic models of the form (8) that generate chaotic dynamics when external forcing noise is set equal to zero (e.g., Benhabib 1992, Boldrin's chapter in Anderson et al. 1988, Dechert 1996).

An important issue is whether, in such models, the parameter values needed to obtain chaos (especially a chaos where prediction in the near term can be improved by exploiting it) are consistent with empirical measurements in economics. For example, when real interest rates are low (as they typically are), intertemporal smoothing operations such as intertemporal arbitrage (e.g., the trading of assets across different points of time in order to profit after adjustment for interest costs and for risk bearing) tend to squash cycles and chaos in economic systems with a rich enough variety of market instruments.

Brock's chapter in Anderson et al. (1988) goes through a list of empirical plausibility checks for the magnitude of 'frictions' needed to obtain short term forecastable deterministic cycles and chaos at high to

usuably (usuably from a policy-relevant predictive point of view) low frequencies in macroeconomic and financial data. Brock's general conclusion is not encouraging for the presence of persistent deterministic cycles and chaos for economic dynamics of macro variables and financial asset returns in countries with well-developed asset markets and financial markets but maybe better for countries with poorly-developed markets. It is argued that well-developed market economies just have too many instruments through which self-interested intertemporal smoothing behavior can operate to be consistent with persistent deterministic chaos or cycles at high to medium frequencies.

After all, the usual arguments behind the efficient markets hypothesis suggests that there is money lying on the table if there are potentially predictable patterns such as deterministic chaos or deterministic cycles in asset returns. That is, arbitrage opportunities are available unless interest rates are high, market instruments for arbitrage are few, risk adjustments are high, or constraints on borrowing and lending are high. Of course, none of these arguments are germane to the presence of chaos and cycles at very low frequencies. For example, Day's chapter in Pesaran and Potter (1992) is more encouraging for the presence of persistent 'complex' economic dynamics such as chaos, especially at long-term historical frequencies. However, while theory is suggestive, there is no substitute for rigorous statistical testing.

More interesting to empirical economists is whether there is evidence in economic and financial time series data for the presence of chaos. While the evidence is weak for financial data and for macroeconomic data for the presence of chaos which can be short term predicted, there does seem to be useful evidence in favor of nonlinear structure which can be predicted in the short term, conditional on appropriate information sets. The evidence for extra unconditional out-of-sample predictability (i.e., prediction of data that was not in the sample used to fit the model) using nonlinear methods appears to be weak for asset returns data. The evidence for extra conditional out-of-sample predictability of asset returns is better (LeBaron 1994). See the review of Brock et al. (1991) with especial attention to the references to Diebold, Nason, and LeBaron. However evidence for bifurcations, complex dynamics, abrupt changes, and other nonlinear phenomena seems quite strong (Dechert and Hommes 2000, see especially Chavas for animal dynamics; Carpenter et al. 1999, for ecological dynamics; LeBaron 1994, for nonlinear patterns in financial data; Dechert 1996, for an overall review of evidence).

3. *New Statistical Methods*

The quest for evidence of deterministic chaos, deterministic cycles, and other complex deterministic persistent patterns in economic and financial data, has inspired the development of new statistical methods. These methods have turned out to be useful in areas having nothing to do with deterministic chaos. We briefly explain two of them here. The first method is a specification test called the 'BDS test' by many writers (Bollerslev et al.'s chapter in Engle and McFadden 1994, De Grauwe et al. 1993). The method emerged out of the work that culminated in Brock et al. (1996). Here is a very brief explanation.

Suppose one formulates a model that relates a set of variables to be predicted to another set of variables called predictor variables, and fits this model to data and saves the residuals. If one has done a proper job of theorizing, model formulation, and model fitting, then the residuals should be unforecastible using histories based upon observables. The BDS test is used to formulate and carry out tests for unpredictability of residuals of fitted models. (See Brock et al. 1991 for extensive discussion and Monte Carlo work.) The BDS test has been used to test the adequancy of fitted models to data (Dechert 1996, De Grauwe et al. 1993, Pesaran and Potter 1992). The BDS test plays a similar role for nonlinear and general models as the Box and Jenkins Q-test does for auto-regressive integrated moving average (ARIMA) models (Box and Jenkins 1976). Tests like the Q-test and the BDS test are useful for testing the adequacy of fitted models and evaluating whether the evidence warrants a more costly exploration of alternatives to the null hypothesis model. See Dechert (1996) for a collection of studies where the BDS test is used in this manner.

The bootstrap-based specification test is a more sophisticated specification testing method. This method is especially relevant for detecting subtle nonlinearities in settings like finance where the economic logic dictates that any patterns are likely to be hard to detect (e.g. Maddala and Li's chapter in Maddala and Rao 1996). The idea here is to use a version of bootstrap to compute the null-model distribution of statistics gleaned from various trading strategies, and to use the data values of these statistics to suggest refinements of the null model in financially relevant directions. For example, in financial applications, the null model is sometimes taken to be a random walk. This procedure emphasizes development of statistical quantities to evaluate the null model which are motivated by the behavior that one is modeling. This method was made possible by recent technical increases in computing speed, and reduction in cost, as well as advances in computationally based inference methods such as bootstrap.

4. *Conclusion and Future Directions*

The study of chaos and general complex dynamics as well as general complexity theory in economics, finance, and social studies has already been fruitful in generating new methods and new evidence. This kind of work is likely to grow in prominence as computa-

tional costs fall and advances continue in computational-based methods of theory and statistical inference.

See also; Computational Psycholinguistics; Linear and Nonlinear Programming; Neural Systems and Behavior: Dynamical Systems Approaches; Self-organizing Dynamical Systems; Stochastic Dynamic Models (Choice, Response, and Time)

Bibliography

Anderson P, Arrow K, Pines D (eds.) 1988 *The Economy as an Evolving Complex System*. Addison-Wesley, Redwood City, CA

Benhabib J (ed.) 1992 *Cycles and Chaos in Economic Equilibrium*. Princeton University Press, Princeton

Box G, Jenkins G 1976 *Time Series Analysis: Forecasting and Control* (revised edn). Holden-Day, San Francisco, CA

Brock W 1986 Distinguishing random and deterministic systems (abridged version). In: Grandmont J (ed.) *Nonlinear Economic Dynamics*. Academic Press, New York

Brock W, Dechert W, Scheinkman J, LeBaron B 1996 A test for independence based upon the correlation dimension. *Econometric Reviews* **15**(3): 197–235

Brock W, Hsieh D, LeBaron B 1991 *Nonlinear Dynamics, Chaos, and Instability Statistical Theory and Economic Evidence*. MIT Press, Cambridge, MA

Carpenter S, Ludwig D, Brock W 1999 Management of eutrophication for lakes subject to potentially irreversible change. *Ecological Applications* **9**(3): 751–71

Dechert W (ed.) 1996 *Chaos Theory in Economics: Methods, Models and Evidence*. Edward Elgar, Cheltenham, UK

Dechert W, Hommes C (eds.) 2000 Complex nonlinear dynamics and computational methods. *Journal of Economic Dynamics and Control* **24**: 651–62

De Grauwe P, Dewachter H, Embrechts M 1993 *Exchange Rate Theory, Chaotic Models of Foreign Exchange Markets*. Blackwell, Oxford, UK

Engle R, McFadden D 1994 *Handbook of Econometrics*. North-Holland, Amsterdam, Vol. 4

Kuznetsov Y 1995 *Elements of Applied Bifurcation Theory*. Springer, New York

LeBaron B 1994 Chaos and nonlinear forecastibility in economics and finance. *Philosophical Transactions of the Royal Society* **348**: 397–404

Maddala G, Rao C (eds.) 1996 *Handbook of Statistics 14: Statistical Methods in Finance*. North-Holland, Amsterdam

Pesaran H, Potter S 1992 Nonlinear dynamics and econometrics. *Journal of Applied Econometrics* **7**: S1–S195

W. A. Brock

Characterization Theorems in Random Utility Theory

In random utility theory the evaluation of a stimulus by a subject is modeled by a random variable from which a sample is taken at each presentation of the stimulus. This idea, which goes back to Gustav Fechner and was further developed in psychology by Thurstone (1927), is used to explain the often observed inconsistency in choice experiments, where a subject on repeated presentations of one particular subset of alternatives does not always select the same alternative. The term 'utility' refers to a setting where the subject is requested to choose on the basis of preference (a choice context typical for economics), but it has to be stressed that the applicability of random utility models is much wider. In psychophysical experiments the criterion of choice might be, for instance, the intensity of lights, pitch of tones, perceived weight of objects, and so on. While in most instances of random utility models the distributions of the random variables are restricted to some parametric family (following Thurstone, who considered Gaussian distributed variables), the problem addressed in this entry is to characterize random utility theory in its most general form in terms of the testable restrictions it imposes on the choice data that it is supposed to model.

1. Random Utility Representation

The data of a typical choice experiment consist of the relative frequencies, or, ideally, choice probabilities $p(i, K)$ with which an alternative $i \in K$ is chosen when the subset K of alternatives is offered. The master set of n alternatives is conveniently identified with $\boldsymbol{n} = \{1, \ldots, n\}$. In principle any system of subsets may be offered, but two special cases are of both theoretical and practical interest. In a *complete system* of choice probabilities choices are obtained for every subset of $\boldsymbol{n}$ with at least two elements; if just all two-element subsets are offered, the data can be collected in a *binary choice probability* (BCP) matrix.

According to random utility theory the internal value or 'strength' of each alternative $i \in \boldsymbol{n}$ in the choice context of the experiment is modeled by a random variable $\boldsymbol{U}_i$, with all such random variables defined on the same probability space with measure Pr. The collection $\{\boldsymbol{U}_i\}_{i \in \boldsymbol{n}}$ is a *random utility* (RU) representation of the choice probabilities $p(i, K)$ whenever

$$p(i, K) = \Pr(\boldsymbol{U}_i = \max_{j \in K}\{\boldsymbol{U}_j\}), \quad i \in K \subseteq \boldsymbol{n}. \quad \text{(RU)}$$

The fact $p(i, \{i, j\}) + p(j, \{i, j\}) = 1$ for two-element subsets entails that the collection $\{\boldsymbol{U}_i\}_{i \in \boldsymbol{n}}$ is *nonconcident*, meaning that $\Pr(\boldsymbol{U}_i = \boldsymbol{U}_j) = 0$, and thus $\Pr(\boldsymbol{U}_i \geqslant \boldsymbol{U}_j = \Pr(\boldsymbol{U}_i > \boldsymbol{U}_j)$, for $i \neq j$. This is no serious restriction; for instance, all continuous random variables have this property.

2. Representation by Rankings

The problem of interest now is to characterize RU in terms of a set of necessary and sufficient conditions on the observed choice probabilities, allow-

ing an arbitrary joint distribution for the collection $\{\boldsymbol{U}_i\}_{i\in\boldsymbol{n}}$. Obviously, in RU the choice probabilities are completely determined by the $n!$ values $\Pr\{\boldsymbol{U}_{i_1} > \boldsymbol{U}_{i_2} > \cdots > \boldsymbol{U}_{i_n}\}$ of this distribution for all rankings of the n alternatives. Any such ranking can be identified with the linear order π on $\boldsymbol{n}$ for which $i\,\pi\,j$ whenever i is ranked higher than j. Denoting the collection of linear orders on $\boldsymbol{n}$ by Π_n, choice probabilities $p(i, K)$ are said to be *induced by rankings* (IR) if there exists a probability distribution $\mathbb{P}$ on Π_n such that

$$p(i, K) = \mathbb{P}\{\pi\in\Pi_n : i\,\pi\,j \quad \text{for all} \quad j\in K\backslash\{i\}\}, \quad i\in K\subseteq\boldsymbol{n} \tag{IR}$$

As already noted by Block and Marschak (1960), RU and IR are equivalent.

In economics, where the interest is not in repeated choices by one individual, but in an aggregation of choices over many individuals, IR results immediately from the assumption that within any single person the choices derive from some fixed ranking of the alternatives (economists tend to call such choices 'rational'), with $\mathbb{P}$ collecting the relative frequencies of the rankings in the population.

The study of the representation problem for random utility models concentrates on IR rather than RU: verifying the existence of a probability distribution on a finite set of rankings turns out to be more tractable than verifying the existence of a collection of undetermined random variables.

2.1 *Characterization for Complete Systems*

Block and Marschak (1960) established that for a complete system of choice probabilities the linear inequalities

$$\sum_{K\subseteq J\subseteq\boldsymbol{n}} (-1)^{|J\backslash K|} p(i, J) \geqslant 0, \quad i\in K\subseteq\boldsymbol{n} \tag{BM}$$

are necessary conditions for IR, thus RU. Via an intricate combinatorial argument Falmagne (1978) then constructively showed how, given BM, a probability distribution $\mathbb{P}$ on Π_n satisfying IR can be constructed, thus proving that BM actually characterizes IR and RU in terms of testable restrictions on complete systems of observed choice probabilities.

3. *The Geometry of Binary Choice Probabilities*

No such definite result is available for BCP matrices, for which RU and IR specialize to

$$p_{ij} = \Pr(\boldsymbol{U}_i > \boldsymbol{U}_j) = \sum_{\{\pi\in\Pi_n : i\,\pi\,j\}} \mathbb{P}(\pi), \quad ij\in\boldsymbol{N}$$

with p_{ij} as a shorthand for $p(i,\{i, j\})$ and with $\boldsymbol{N} = \{ij\colon i, j\in\boldsymbol{n}, i\neq j\}$ denoting the collection of ordered pairs from $\boldsymbol{n}$.

Geometrically, any BCP matrix p is a point in $[0, 1]^{\boldsymbol{N}}$, the unit hypercube in $n(n-1)$-dimensional real vector space indexed by $\boldsymbol{N}$, and the collection of BCP matrices on n elements that are induced by rankings is a particular subset of this hypercube, denoted by P_n. Identifying a linear order on $\boldsymbol{n}$ with its indicator function (i.e., writing $\pi_{ij} = 1$ instead of $i\,\pi\,j$), IR can for BCP matrices be restated as

$$p\in P_n \Leftrightarrow p = \sum_{\pi\in\Pi_n} \mathbb{P}(\pi)\pi \tag{CH}$$

for some probability distribution $\mathbb{P}$, that is, for some nonnegative real numbers $\mathbb{P}(\pi)$ summing to unity. Thus, by definition, P_n constitutes the *convex hull* of the set Π_n consisting of $n!$ of the $2^{n(n-1)}$ vertices of the hypercube. By *linear duality theory* (see, e.g., Kuhn 1956) this statement is equivalent to the general criterion

$$p\in P_n \Leftrightarrow \langle c, p\rangle \leqslant \max_{\pi\in\Pi_n}\{\langle c, \pi\rangle\} \quad \text{for all integer } c\in\mathbb{R}^{\boldsymbol{N}},\ c\geqslant 0$$

where $\langle\cdot,\cdot\rangle$ is the standard inner product in $\mathbb{R}^{\boldsymbol{N}}$. Thus, is an abstract sense, this system of linear inequalities characterizes BCP matrices p satisfying RU and IR. This solution, however, involves an infinite system of inequalities for any fixed, finite number n of alternatives, and, consequently, does not yield an effective procedure for checking whether for a given BCP matrix a random utility representation exists. (This contrasts with the solution BM for complete systems of choice probabilities.) The following standard theory makes it clear that, for any fixed n, the set P_n is in fact defined by a finite subsystem of the above infinite system. The characterization problem that remains is to find, for any n, such a finite collection of linear inequalities.

4. *The Linear Description of P_n*

By CH, P_n is a polytope, i.e., the convex hull of finitely many points, in $\mathbb{R}^{\boldsymbol{N}}$. The set of vertices of P_n is Π_n, whence it is called the *linear ordering polytope*. Any polytope is alternatively described as a bounded *polyhedron*, i.e., a subset obtained as the intersection of finitely many halfspaces, or, the solution set of finitely many linear inequalities. In fact, any $p\in P_n$ satisfies the system of $\binom{n}{2}$ linear equalities (EQ)

$$p_{ij} + p_{ji} = 1, \quad ij\in\boldsymbol{N} \tag{EQ}$$

each of which constrains P_n to be contained in the

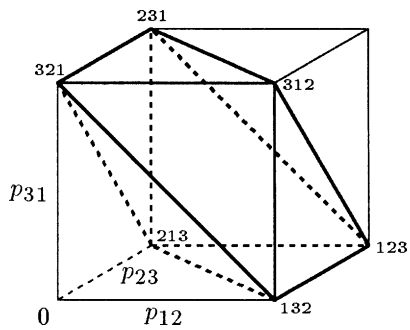

Figure 1
Unit cube with the linear ordering polytope P_3 (in bold) and vertices indicated with the corresponding rankings of {1, 2, 3}

hyperplane which is the intersection of the halfspaces defined by the two corresponding inequalities. As a consequence, P_n is not of full dimensionality in $\mathbb{R}^N$, but has (affine) dimension $n(n-1)-\binom{n}{2}=\binom{n}{2}$.

If the inequality $\langle a, p\rangle \leqslant a_0$ is valid for all $p \in P_n$ and the corresponding equality $\langle a, p\rangle = a_0$ is satisfied by some, but not all, $p \in P_n$, then the set $\{p \in P_n : \langle a, p\rangle = a_0\}$ is called a *face* of P_n. Geometrically, P_n is fully contained in the halfspace of $\mathbb{R}^N$ defined by this valid inequality and the face of P_n consists of the points at the boundary of P_n that lie in the hyperplane delimiting this halfspace. A face of maximum dimension, i.e., one less than the dimension P_n itself, is called a *facet* of P_n, and the corresponding valid inequality is called *facet defining*. A complete, nonredundant system of inequalities for P_n consists of facet-defining inequalities, one for each facet of P_n. Such a system is non-unique, since any two inequalities $\langle a, p\rangle \leqslant a_0$ and $\langle b, p\rangle \leqslant b_0$ that are the same modulo EQ (i.e., such that EQ entails $\langle a, p\rangle - \langle b, p\rangle = a_0 - b_0$) are equivalent and define the same face of P_n (if any). Thus, the representation problem is redefined as one of finding facet-defining inequalities for all facets of the linear ordering polytope P_n.

4.1 *The Case $n = 3$ and the Triangle Inequality*

The smallest nontrivial BCP-matrix $p = (p_{12}, p_{13}, p_{21}, p_{23}, p_{31}, p_{32})$ for $n = 3$ alternatives is, by (7), contained in a three-dimensional subspace of the unit hypercube of $\mathbb{R}^6$, specified, for instance, by the p_{12}, p_{23} and p_{31} coordinates. This is illustrated in Fig. 1. The polytope P_3 is the octahedron obtained as the convex hull of the six linear orders on {1, 2, 3}. These vertices and the boldface edges in Fig. 1 constitute the faces of dimension 0 and 1, respectively. The facets of P_3 are the triangular faces of the octahedron. Six of them coincide with the facets of the cube itself, correspond to the inequalities $p_{ij} \geqslant 0$, $p_{ij} \leqslant 1$, and thus represent no restriction on BCP matrices. The remaining two facets, which do exclude parts of the cube from P_3, are defined by inequalities of the form

$$p_{ij} + p_{jk} + p_{ki} \leqslant 2, \quad i, j, k \in \boldsymbol{n}$$

Since this inequality can be equivalently rendered as

$$p_{ik} \leqslant p_{ij} + p_{jk}, \quad i, j, k \in \boldsymbol{n}$$

it is known as the *triangle inequality*.

Thus, for $n = 3$, the situation is simple, with the triangle inequality as necessary and sufficient condition. Since an inequality defining a facet of some P_n remains facet-defining for all higher-dimensional linear ordering polytopes, the triangle inequality is a necessary condition for all $n \geqslant 3$. Although the geometries of the six-dimensional P_4 and the 10-dimensional P_5 are considerably more complex than that of P_3, the triangle inequality turns out to be also sufficient for $n = 4$ and $n = 5$. It can be interpreted as 'probabilistic transitivity,' since, applied to the vertices of P_n, the inequality amounts to the transitivity property of linear orders as binary relations on $\boldsymbol{n}$. In this sense, EQ in one direction ($\leqslant$) represents a(nti)symmetry and in the reverse direction ($\geqslant$) completeness. (The property of (ir)reflexivity plays no role here.)

5. *The General Case*

In this way the picture seems nice and complete and, indeed, at some point it was conjectured that the triangle inequality would be sufficient for all n. However, for $n = 6$ soon a counterexample was found of some $p \notin P_6$, but satisfying this inequality. Thus, starting with $n = 6$, there are other facet-defining inequalities. Once this was realized, a hunt for facets started, yielding more and more inequalities, in fact whole classes of inequalities, for increasing n. (Fishburn (1992) presents history and state of the art of these developments at that point in time.) Two of the most general classes found up to now are presented next, to illustrate the complexity of the problem.

5.1 *Facets Defined by Möbius Ladders*

In the field of operations research, where the polytope P_n turns up in a discrete optimization context, Grötschel et al. (1985) developed techniques for proving inequalities facet-defining and defined various schemes of facet-defining inequalities in the language of graph theory. In this setup, *directed graphs* are considered with node set $\boldsymbol{n}$. Any arc set $A \subseteq \boldsymbol{N}$ defines through its indicator function 1_A a linear functional denoted as

$$A(p) = \langle 1_A, p\rangle = \sum_{ij \in A} p_{ij}.$$

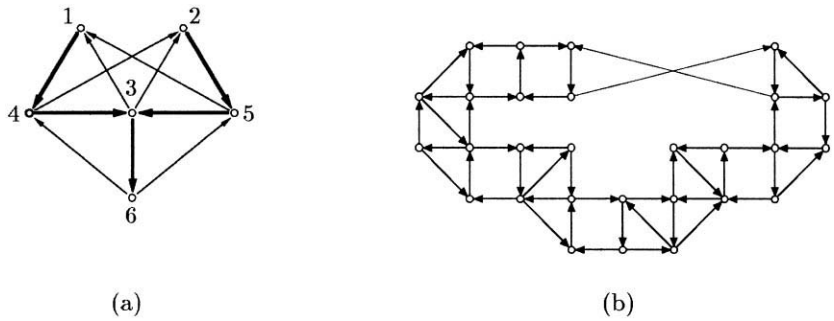

Figure 2
(a) Möbius ladder composed of one four-cycle and four three-cycles; arcs shared by two cycles in bold. (b) Möbius ladder of 21 cycles (adapted from Grötschel et al. (1985))

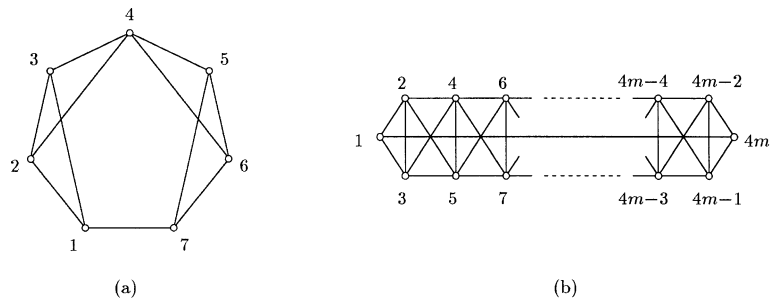

Figure 3
(a) A stability-critical graph with stability number 2. (b) An infinite family of stability-critical graphs (with stability number m)

By suitable choices for the arc set A and constant a_0 facet-defining inequalities of the form $A(p) \leqslant a_0$ may be obtained. One example follows.

Grötschel et al. (1985) call an arc set A a *Möbius ladder* basically if A is the union of an odd number k of (directed) cycles $C_1, C_2, \ldots, C_k$, such that each C_i is either a three-cycle or a four-cycle and the given sequence constitutes an (undirected) cycle on a higher level in the sense that any adjacent pair C_i, C_{i+1} (including C_k, C_1) has precisely one arc in common, while all nonadjacent pairs C_i, C_j are disjoint. (There are a few additional side-conditions.) For any such Möbius ladder A the inequality

$$A(p) \leqslant |A| - \frac{k+1}{2}$$

is facet defining for all linear ordering polytopes P_n for which it is defined (that is, with n equal to or exceeding the number of nodes appearing in A). Fig. 2(a) shows a small Möbius ladder, which, with $|A| = 11$ and $k = 5$, represents a facet-defining inequality for all P_n, $n \geqslant 6$. Fig. 2(b) illustrates the generality of the class of Möbius ladders by a more elaborate example, which also more explicitly shows the idea of the Möbius strip.

5.2 *Facets Defined by Stability-critical Graphs*

Using the techniques developed by Grötschel et al. (1985), Koppen (1995) derived another general class of facet-defining inequalities, this time in terms of *undirected graphs*. Let $G = (\boldsymbol{k}, E)$ be a connected graph with node set $\boldsymbol{k} = \{1, \ldots, k\}$, $k \geqslant 3$, and some edge set E of unordered pairs of $\boldsymbol{k}$. Let $S(\boldsymbol{k}, E)$ denote the *stability number* of G, i.e., the maximum cardinality of a *stable set* in G, where a subset of nodes of G is called stable if it does not contain an edge. G is called *stability-critical* if the removal of any one edge from G would increment the stability number (by one). Koppen (1995) proved that for any such $G = (\boldsymbol{k}, E)$ and any $2k$ distinct elements $u_1, \ldots, u_k, v_1, \ldots, v_k$ in $\boldsymbol{n}$ the inequality

$$\sum_{i \in \boldsymbol{k}} p_{u_i v_i} - \sum_{\{i,j\} \in E} (p_{u_i v_j} + p_{u_j v_i}) \leqslant S(\boldsymbol{k}, E)$$

defines a facet of the linear ordering polytope P_n, $n \geqslant 2k$, if and only if G is stability-critical.

The graph in Fig. 3(a) represents such an inequality with $k = 7$ and $S(\boldsymbol{k}, E) = 2$ (any three-set contains an edge, but any edge removed introduces a stable three-set), which is facet defining for P_n, $n \geqslant 14$. In Fig. 3(b) an infinite family of stability-critical graphs is de-

picted, each corresponding to a new facet-defining inequality. The symmetry present here is 'accidental'; being stability-critical imposes no further structural restrictions on a graph whatsoever.

5.3 Conclusion for the Binary Case

The Möbius ladders and stability-critical graphs are just two examples of the various classes of facet-defining inequalities for BCP matrices known to date. In particular, while all cases considered here have 0–1 coefficients, facet-defining inequalities with general, unbounded integer coefficients have been found (Leung and Lee 1994, McLennan 1990, Suck 1992).

The conclusion is that, starting with the triangle inequality for $n \leqslant 5$, there is, with increasing n, a combinatorial explosion of facet-defining inequalities for P_n with no apparent structural regularities—witness the special case of stability-critical graphs alone. Thus, the problem of a complete characterization for general n seems intractable at the moment. What can be achieved on a more practical level, by computer enumeration, is obtaining a complete linear description of P_n for specific small n, $n = 6, 7, 8, \ldots$.

6. Extension to General Probabilistic Measurement

Taking an abstract, measurement-theoretic point of view, Heyer and Niederée (1992) have shown that random utility theory, as discussed above, can be generalized to other instances of probabilistic measurement (see *Measurement Theory: Probabilistic*). The basic idea is as follows (considerable extensions to more complicated relational structures and data formats are to be found in Niederée and Heyer (1997)).

Let for some finite domain of n objects, identified with $\boldsymbol{n}$, a mapping p: $\boldsymbol{n}^k \to [0, 1]$ specify 'probability data' on k-tuples of objects. Let Q be a 'quantitative' k-ary relation on $\mathbb{R}$ and $\mathscr{R}$ the set of k-ary relations on $\boldsymbol{n}$ isomorphically embeddable in $(\mathbb{R}, Q)$. Then, in this general setting, it can be shown that p has a random scale representation

$$p(i_1,\ldots, i_k) = \Pr((\boldsymbol{U}_{i_1},\ldots, \boldsymbol{U}_{i_k}) \in Q)$$

for some jointly distributed collection of random variables $\{\boldsymbol{U}_i\}_{i \in \boldsymbol{n}}$ if and only if p is induced by a probability distribution on $\mathscr{R}$,

$$p(i_1,\ldots, i_k) = \mathbb{P}\{R \in \mathscr{R}: (i_1,\ldots, i_k) \in R\}$$

if and only if p is in the convex hull of $\mathscr{R}$, interpreted as vertices of the hypercube $[0, 1]^{n^k}$. Classical random utility theory in terms of RU, IR, and CH is simply the special case with $k = 2$, $Q = >$ and $\mathscr{R}$ the set of linear orders on $\boldsymbol{n}$.

The consequence is that for the general case of probability data p the characterization problem amounts to finding, for any n, a system of linear inequalities defining the facets of the polytope with vertex set $\mathscr{R}$. Some results have been obtained for other relations (Regenwetter 1996, Suck 1997). Typically, in such a case the first nontrivial facets found are the ones corresponding to the defining properties of the class of relations $\mathscr{R}$ (see the end of Sect. 4.1). The parallel with the linear order case goes further in that, quite generally, more and more facets turn up with increasing n. To date, these investigations have not resulted in characterization theorems for general n and, except for very small n, linear descriptions of the corresponding polytopes are only partly known.

See also: Decision and Choice: Random Utility Models of Choice and Response Time; Elicitation of Probabilities and Probability Distributions; Measurement Theory: Probabilistic; Preference Models with Latent Variables; Utility and Subjective Probability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies

Bibliography

Block H D, Marschak J 1960 Random orderings and stochastic theories of response. In: Olkin I, Ghurye S, Hoefding W, Madow W, Mann H (eds.) *Contributions to Probability and Statistics*. Stanford University Press, Stanford, CA

Falmagne J-C 1978 A representation theorem for finite random scale systems. *Journal of Mathematical Psychology* **18**: 52–72

Fishburn P C 1992 Induced binary probabilities and the linear ordering polytope: A status report. *Mathematical Social Sciences* **23**: 67–80

Grötschel M, Jünger M, Reinelt G 1985 Facets of the linear ordering polytope. *Mathematical Programming* **33**: 43–60

Heyer D, Niederée R 1992 Generalizing the concept of binary choice systems induced by rankings: One way of probabilizing deterministic measurement structures. *Mathematical Social Sciences* **23**: 31–44

Koppen M 1995 Random utility representation of binary choice probabilities: Critical graphs yielding critical necessary conditions. *Journal of Mathematical Psychology* **39**: 21–39

Kuhn H W 1956 Solvability and consistency for linear equations and inequalities. *American Mathematical Monthly* **63**: 217–32

Leung J, Lee J 1994 More facets from fences for linear ordering and acyclic subgraph polytopes. *Discrete Applied Mathematics* **50**: 185–200

McLennan A 1990 Binary stochastic choice. In: Chipman J S, McFadden D, Richter M K (eds.) *Preferences, Uncertainty, and Optimality*. Westview, Boulder, CO

Niederée R, Heyer D 1997 Generalized random utility models and the representational theory of measurement: A conceptual link. In: Marley A A J (ed.) *Choice, Decision, and Measurement*. L. Erlbaum, Mahwah, NJ

Regenwetter M 1996 Random utility representations of finite m-ary relations. *Journal of Mathematical Psychology* **40**: 219–34

Suck R 1992 Geometric and combinatorial properties of the polytope of binary choice probabilities. *Mathematical Social Sciences* **23**: 81–102
Suck R 1997 Probabilistic biclassification and random variable representation. *Journal of Mathematical Psychology* **41**: 57–64
Thurstone L L 1927 A law of comparative judgment. *Psychological Review* **34**: 273–86

M. Koppen

Charisma and Charismatic

'Charisma' or 'gift of grace' is a theological notion that has been widely used in the social and religious sciences to describe either the hierarchical organization of religious roles or explain the growth and development of social movements based on religious inspiration or the basis of authority and leadership in society generally. In its strictly religious context, it means a divinely conferred power, being derived from the Greek *kharisma* (*kharis* favor or grace). Charismatic power is associated with the idea of the sacred as a force in human affairs. A person in possession of charisma is thought to have a talent, for example, in terms of healing or prophecy. In anthropological research, there has been considerable interest in 'shamanism' as a form of charismatic authority that depends on a capacity to have visions and to perform healing (see M. Eliade 1964). In sociology, it is conceptually part of an analytical framework that is concerned with understanding large-scale changes in religious institutions and the foundations of authority.

1. *The Sociology of Charisma*

In the sociology of religion, the study of charisma has been closely associated with Max Weber (1864–1920) who adopted the idea from the historical and theological research of Rudolf Sohm and Karl Holl who in turn had developed the concept in their analysis of Canon Law. Weber wrote that 'the concept of charisma' (gift of grace) is taken from the vocabulary of early Christianity. For the Christian hierocracy Rudolf Sohm, in his *Kirchenrecht*, was the first to clarify the substance of the concept, even though he did not use the same terminology. Others (for instance, Holl in *Enthusiasmus und Bussgewalt*) have clarified certain important consequences of it. It is thus nothing new (Weber 1978, pp. 1, 216). Despite Weber's modesty about the adoption of the idea, it became a fundamental dimension of his analysis of power and had far-reaching consequences for the development of the sociology of religious institutions. Weber generalized the idea and grasped its radical implications for the study of political change in human societies.

In the rise of religion certain individuals are recognized as having a capacity to experience ecstatic states that are regarded as the precondition for healing, telepathy, and divination. It is primarily through 'these extraordinary powers that have been designated by such special terms as 'mana,' 'orenda,' and the Iranian 'Maga' (the term from which our word 'magic' is derived). We shall henceforth employ the term 'charisma' for such extraordinary powers' (Weber 1963, p. 2). Such charismatic power is either inherited as a natural endowment or it is acquired by extraordinary means. The religious talent in both cases is a substance that may remain dormant in a person until it is aroused by asceticism or trance.

Weber did not dwell specifically on these features of charisma in folk religion because he wanted to employ the term to understand the secular dynamic of authority and leadership in social institutions. His main intention was to compare and contrast three types of authority: charismatic, traditional, and legal-rational. As we have seen, charismatic authority rests on the ability of a leader to inspire disciples in the belief of the authenticity of a calling. In practical terms, an authentic claim is validated by a talent such as healing, but this alone cannot be the basis of authority. In the case of genuine charisma, a follower has a duty to accept the authority of a leader. In *Economy and Society* (Weber 1978, pp. 1, 241), the term charisma is applied to a certain quality of an individual personality by virtue of which he is considered extraordinary and treated as endowed with supernatural, superhuman, or at least specifically exceptional powers or qualities. Traditional authority involves the acceptance of a rule that expresses a custom, namely an established pattern of belief or practice. Finally, legal-rational authority is typical of bureaucracies in which formal rules of conduct are underpinned by procedural norms. These forms of authority are in turn forms of compliance. Tradition depends on compliance through empathy; legal-rational authority rests on rational argument; and charismatic authority and leadership require inspiration.Weber provided many diverse illustrations of charismatic leaders including Jesus Christ, Mohammed, Napoleon, Stefan George, and the Chinese Emperor. Although the cases were heterogeneous, he argued that charismatic authority is confronted by a common problem of succession with the death of the leader. Charismatic authority is thus unstable. With the demise of the charismatic leader, the disciples typically disband, but occasionally a solution for continuity will be developed. In the case of the Christian Church , Weber argued that the charismatic authority of Christ was invested in the Church itself (as the body of Christ) and thus in the bishops who control the 'keys of grace' enjoy a vicarious authority. This 'institutionalization of charisma' becomes over time increasingly formal, bureaucratic, and impersonal. As the charismatic power of Christ becomes transformed into a set of formal procedures and bureaucratic rules, Weber spoke of the 'routinization of charisma.'

2. Forms of Religious Association

Weber's account of charisma was also an important component of his sociology of religion within which he attempted to identify different religious roles and patterns of organization. For example, he distinguished between the prophet who, as a charismatic figure, has a personal call, and the priest has authority by virtue of his service in a sacred tradition. The prophets, who often emerge from the ranks of the priesthood, are unremunerated, and depend on gifts from followers. Weber also distinguished two forms of prophecy as represented on the one hand by Buddha and on the other by Zoroaster and Muhammad. The latter are involved in 'ethical prophecy' and are conceived as instruments of God. These prophets receive a commission from God to preach a revelation and demand obedience from their disciples as an ethical duty. By contrast exemplary prophets demonstrate to their followers a salvational path through the example provided by their own lives. Exemplary prophecy was, according to Weber, characteristic of Asia; ethical prophecy, of the Abrahamic religions of the Middle East. Weber's analysis of charisma with respect to ethical prophecy in the Old Testament has been subject to considerable criticism (see Zeitlin 1984), but his conceptual framework continues to influence both sociology (see Lindholm 1990) and anthropology (see Werbner and Basu 1998).

This discussion of charisma with respect to different social roles should be seen as part of a larger sociological debate about the forms of association that characterize the social organization of religious belief and practice. Weber wanted to argue that any group that is subject to charismatic authority forms a charismatic community (*Gemeinde*) and that such a community is inherently unstable. With the death of the leader, the group either dissolves or charisma undergoes a process of routinization. The disciples have no career, no formal hierarchy, no offices, and no qualifications. The Church that provides the organizational context of the priesthood is very different. Ecclesiastical organizations require a hierarchical administration of the 'charisma of office' in which there are definite stages in clerical and administrative careers. It is clearly the case that Weber's sociology of charisma should be understood as an application of Ernest Troeltsch's 'church-sect typology' in which there is an historical oscillation between the evangelical sects and the bureaucratic churches (see Troeltsch 1931).

3. Charismatic Movements

Because Weber believed that in modern societies legal rational authority would become dominant, tradition and charisma were regarded as 'prerationalistic' and thus as characteristic of premodern societies. The notion that charismatic authority was not a resilient aspect of modern society was in turn a function of Weber's pessimistic understanding of social change in terms of secular rationalism and the erosion of religious meaning. Against Weber's assessment, the late nineteenth and twentieth centuries have been profoundly influenced by charismatic movements. Charismatic renewal has been a common theme of diverse religious movements in 'primal societies' and in the industrial societies of Europe and North America. The collapse of aboriginal or tribal societies under colonial settlement saw the spread of charismatic movements against the supremacy of White-settler societies such as the Ghost Dance among the Cheyenne and Sioux tribes of the American Plains. A Paiute prophet called Wovoka had received a vision in which through ritual dance the dead would return to restore the pristine culture of native societies. This anti-White charismatic movement subsided after the murder of Sitting Bull and the subsequent massacre of his followers at Wounded Knee in 1890. Charismatic leadership has also played a significant role in those new religious movements that have been a response to the social and economic disruptions associated with the decolonization of the Third World (see P. Worsley *The Trumpet Shall Sound* 1970).

In contexts of rapid social change, charismatic leadership is an important component of so-called 'revitalization movements' that function to create an effective transition from tribal to urban society. However, charisma is also associated with religious forms that are a response to personal alienation, isolation, and meaninglessness in the developed, industrial world. One illustration is the Family who were disciples of Charles Manson and operated in southern California in the late 1960s. Manson, who was a social dropout with a criminal record, provided a message of personal liberation based on mind-expanding drugs and found a receptive set of disciples in the cultic milieux of the American counterculture. The Manson Family offered disoriented disciples the experience of an absolute community that depended on a powerful ideology and indoctrination (see Schreck 1988). Another American illustration would be Jim Jones and the People's Temple (see Weightman 1983). These manifestations of charisma in modern society are closely associated with the dislocations of youth culture, the anomie of the modern city, and the spread of personal alienation.

4. Conclusion; Charisma in the Information Age

Weber's analysis of charisma has produced a rich legacy of sociological research, but there are serious conceptual problems associated with its application in contemporary society. The concept of charismatic movement is used to describe a variety of revivalist, nativist, messianic, and healing movements. There is considerable debate in sociology and religious studies about whether charisma in modern society is degraded

and inauthentic (see Wilson 1975). The term is frequently employed to describe the fleeting popularity of political leaders whose social appeal is constructed by campaigns in the electronic media. Whereas charisma in traditional societies arises spontaneously in the collective enthusiasm for divinely inspired leaders, in modern politics it is inevitably the product of deliberate orchestration of the media. The notoriety of the Manson Family and the People's Temple was a conscious product of media attention. However, this argument indicates a persistent issue in charismatic movements, namely how to distinguish between genuine and false claims to religious authority. Biblical warnings against 'false prophets' recognized the fact that disciples could be misguided. While the term has been implausibly stretched to describe a diverse and heterogeneous range of leaders in the twentieth century from Adolf Hitler to J. F. Kennedy, it remains a basic concept in sociology and religious studies.

See also: Charisma: Social Aspects of; Folk Religion; Healing; Millennialism; New Religious Movements; Prophetism; Religion, Sociology of; Weber, Max (1864–1920)

Bibliography

Eliade M 1964 *Shamanism. Archaic Techniques of Ecstasy*. Ballingen Foundation, New York

Lindholm C 1990 *Charisma*. Blackwell, Oxford, UK

Schreck N (ed.) 1988 *The Manson File*. Amok Press, New York

Troeltsch E 1931 *The Social Teaching of the Christian Churches*. Allen and Unwin, London, 2 Vols.

Werbner P, Basu H (eds.) 1998 *Embodying Charisma Modernity Locality and the Performance of Emotion in Sufi Cults*. Routledge, London and New York

Weber M 1963 *The Sociology of Religion*. Methuen, London

Weber M 1978 *Economy and Society. An Outline of Interpretive Sociology*. University of California Press, Berkeley, CA, 2 Vols.

Weightman J M 1983 *Making Sense of the Jonestown Suicides. A Sociological History of the People's Temple*. Mellen Press, New York

Wilson B R 1975 *The Noble Savages. The Primitive Origins of Charisma and its Contemporary Survival*. University of California Press, Berkeley, CA

Zeitlin I M 1984 *Ancient Judaism Biblical Criticism from Max Weber to the Present*. Polity Press, Cambridge, UK

B. S. Turner

Charisma: Social Aspects of

The concept of charisma indicates the presence of a quality which is considered to be extraordinary and exceptional. The subject holds this quality personally, freely, and independently of his or her own will. Once this quality is socially recognized, it lends those who possess it a specific power inside the community they are part of. The explanations given for the possession of charisma—very often imputed to the arbitrary will of divinity—are crucial for the distinction between the different forms of power deriving from its possession, ranging from the authoritativeness of the prophet to the ruling of the leader, and from the magic powers of the shaman to the pure might of the fierce warrior. From the point of view of its social aspects, the concept of charisma refers to the manifestation of a specific modality of power (see *Power in Society*) inside the society, as well as to the specific form of community (see *Community Sociology*) that forms itself around the carrier of personal charisma.

1. The Origin of Charisma as a Problematic Concept

Originally, the concept of charisma appears in the New Testament literature, and almost exclusively in the epistles of the apostle Paul, where it is used with different meanings. It can in fact be: (a) a free gift from God, (b) the specific gift of a vocation, (c) a series of godly gifts following the investiture of a priest, (d) the extraordinary manifestations of the Holy Spirit that underline the evangelizing potentiality of the first Christians. Paul distinguishes between extraordinary gifts— those that do not correspond to any determined function—and exceptional spirituality, which fills the performance of ordinary functions with grace. In the spiritual life of the first communities, the charisma that resulted in the production of miraculous effects were fulfilling a function of legitimization of the acts of preaching. Moreover, they were also facilitating and increasing the efficacy of preaching. In these cases, the relation between charisma and conduct of life becomes apparent (Ducros 1937). A difference emerges, therefore, between the gift given to spread the Word, and that given to allow the subject to begin moving toward perfection. In this tradition of studies, the possession of charisma is the foundation and legitimization of the relations of the subject with the divinity that grants it, but also with the institution that guarantees the transmission of the various faith truths and of the appropriate forms of worship. Moreover, having the charisma also legitimizes its possessors in their relations with those for whom they fulfill a function of spiritual leader.

Being the leader of a community of believers for the salvation of its members is the decisive and fundamental function of the carrier of personal charisma. Having recognized the charismatic power of the leading function exercised by the 'holy people' means grasping the ideal-typical status in which every leader is a guide leading toward definitive and certain

salvation, against every possible opposition. Therefore, the carrier of personal charisma ends up coinciding, in the New Testament literature, with the man of God, who intervenes in moments of crisis to show the way of perfection: the only one that truly counts for individual or collective salvation.

The considerations regarding the use of charisma in the communities of believers can be transposed also to civil communities. A long historic tradition from Plutarch to Hegel through Machiavelli, Kant, Carlyle, and the whole German Romanticism movement prepared the basic elements of the image of the charismatic leader (Cavalli 1981) that Weber would later translate into an ideal-typical figure.

2. *The Concept of Charisma in Contemporary Sociology*

Max Weber (1864–1920) is the author who derives the best results from this dual tradition on the social aspects of charisma. While accepting Rudolf Sohm's work on the origin of the Christian Church, Max Weber underlined different aspects of it: the idea of the supernatural gift offered for the work of a specific mission; the concepts of call, of predestination, and of election. From other works on the history of Christianity, Weber derives further developments of his theory of charisma. This is the case with Karl Holl and his studies on monasticism. For Holl, as for Rudolf Sohm, the possession of charisma is the foundation of the authority inside religious institution. Whatever its concrete manifestations—from the proof guaranteed by exceptional interventions, to the exemplary manifestation of a permanently sanctified life—the objective of charisma is to lead the community of believers after having been recognized by it.

2.1 *Meanings of Charisma in Weber*

In his work, Weber defines charisma many times. First of all, it is part of the sociological categories and more precisely, it is one of the ideal types of power. In the definition of 'charismatic power' the 'charisma' is

> a certain quality of an individual personality by virtue of which he is apart from ordinary men and treated as endowed with supernatural, superhuman, or at least specifically exceptional power or qualities. These, as such, are not accessible to the ordinary person, but are regarded as of divine origin or as exemplary, and on the basis of them the individual concerned is treated as a leader. In primitive circumstances this peculiar kind of deference is paid to prophets, to people with a reputation for therapeutic or legal wisdom, to leaders in the hunt, and heroes in war (Weber 1956/1968, p. 48)

The analysis of the term 'charisma' occurs also when Weber is discussing types of religious communities. Weber defines it here—not without connections to the analyses of the Durkheimian school, and in particular, with Marcel Mauss's studies—as the collection of extraordinary powers that populate the field of magically or religiously motivated action. In this field charisma, understood as an ideal-typical form, is 'a gift that inheres in an object or person simply by virtue of natural endowment,' and it 'may be produced artificially in an object or person through some extraordinary means.' This latter definition is apparently secondary, since 'it is assumed that charismatic powers can be developed only in people or objects in which the germ already existed but would have remained dormant unless evoked by some ascetic or other regimen' (Weber 1956/1965, p. 2).

The two parts, of course, complement each other. In the analysis of charisma as a sociological category of social action, the accent is on the recognition of the exceptional quality, whatever this might be. We must also stress, however, that in the section concerning religious communities, charisma is underlined in its substantiality: It has, in other words, a specific content and expresses itself in a doctrine.

2.2 *Analysis of Charisma as to Types of Religious Communities*

The quality of charisma becomes more precise in its primary form as the capacity of the subject that carries it to awaken and govern the hidden powers that inhabit the natural world. This capacity can be consubstantial with the actors or activated as a consequence of an exceptional state of being, namely ecstasy. What for lay people cannot be other than an occasional experience, is for magicians a dimension characterizing the existence to which they have permanent access: 'The magician is the person who is permanently endowed with charisma.'

The success of charisma, understood in this sense, is directly connected to the capacity to force the invisible powers to yield to one's own will through the use of adequate stratagems. On the other hand, it disappears when supernatural powers assert themselves, powers that are indifferent to any kind of constraint but sensitive to veneration and worship acts. The magician is substituted by the priest or minister who is a functionary taking care of a 'regularly organized and permanent enterprise concerned with influencing the gods, in contrast with the individual and occasional efforts of magicians' (Weber 1956/1965, p. 28). The regularity of the functions and rites rendered priests as functionaries who often inherit their position and whose power seems to be separated from personal charisma.

> Yet another distinguishing quality of the priest, it is asserted, in his professional equipment of special knowledge, fixed doctrine and vocational qualifications, which brings him into contrast with sorcerers, prophets, and other types of religious functionaries who exert their influence by virtue of personal gifts (charisma) made manifest in miracle and revelation (Weber 1956/1965, p. 29).

The prophet is, in this analysis, 'a purely individual bearer of charisma, who by virtue of his mission proclaims a religious doctrine or divine commandment' (Weber 1956/1965, p. 46). It is, therefore, not so important if a new community emerges around prophets or if their followers refer more to their person than to their doctrine. What is decisive is the 'purely personal charisma' distinguishing prophets from priests or ministers, while the content of their preaching and the specificity of their actions, consisting in doctrines and ethical imperatives instead of magic, distinguish them from magicians.

2.3 Sociology of Power

Max Weber defines once more the concept of charisma when passing from the analysis of religious communities to that of political communities and, through these, to the sociology of power. It is during this third phase of his work that the social aspects of charisma are underlined. What is important to highlight is his further clarification of the emotional aspect of charisma, which is also, to a certain extent, meta-everyday, alien to the paths of rationalization that are needed for its social reproduction, but still necessary for the foundation of a new epoch. Weber arrives at the point of saying that every foundation moment is, in fact, charismatic. This connects the use of charisma not only to its social recognition but also to the processes of innovation and social transformation. He contrasts the permanent character of the traditional and the bureaucratic structures, to the foundational character of the charismatic action.

> Die Deckung allen über die Anforderungen des ökonomischen Alltags hinaus gehenden Bedarfs dagegen ist, je mehr wir historisch zurücksehen, desto mehr, prinzipiell gänzlich heterogen und zwar: charismatisch, fundiert gewesen. Das bedeutet: die 'natürlichen' Leiter in psychischer, physischer, ökonomischer, ethischer, religiöser, politischer Not waren weder angestellte Amtspersonen, noch Inhaber eines als Fachwissen erlernten und gegen Entgelt geübten 'Berufs' im heutigen Sinn dieses Wortes, sondern Träger spezifischer, als übernaturlich (im Sinne von: nicht jedermann zugänglich) gedachter Gaben des Körpers und Geistes (Weber 1956, p. 662).

Summarized in one sentence, this important quotation says that the answer to human needs, going beyond mere economic needs, is not given by experts or people holding public office but by bearers of qualities of the body and mind perceived as supernatural.

In Weber's methodology such distinctions have of course an ideal-typical character. It is therefore impossible to trace the 'pure forms' of each of them as well as to draw specific boundaries between one category and the other. This means, for example, that it is possible to find some elements of prophetic charisma in the behavior of a political leader, of a priest, or of a military leader. Likewise, magical elements can be found in religious action: This happens, for instance, when prophets use magic acts to demonstrate their own charisma.

At least three aspects of the Weberian analysis of charisma need to be recalled here: the nature of social recognition, the implicitly protesting character of charismatic power, the transformation of the exercise of charisma into everyday practice, or in other words, its 'routinization,' and, inside the everyday practice, the distinction between personal and official charisma.

2.4 Social Recognitions

The possession of charismatic qualities is connected to the social context in two ways. On the one hand it is in the social context that the recognition takes place and those possessing charisma need to be socially recognized as such by their followers. On the other hand, social recognition does not constitute the foundation of charisma. It is, on the contrary, a duty for those that are expected to recognize it. 'This recognition, psychologically born of enthusiasm or adversity and hope, involves complete personal devotion' (Weber 1956/1994, p. 33). The recognition refers, therefore, to a previous confident expectation. The nature of this expectation is that of a positive event such as a reconquest of political freedom or of supernatural salvation or, again, of a redemption capable of restoring to the entire people its place among the nations. A specific relation emerges between recognition of charisma, the diffused and extended expectation of the carrier of charismatic qualities, and the reunification around that person of a completely renewed community of followers, which aims at originating a process of renewal for the whole society. Since the entire Weberian discourse is based on ideal-typical configurations, it also has numerous variants: The social recognition of prophetic charisma has a completely different relation with 'proof' compared to magical charisma or to that of the political leader. Prophets proclaiming a 'religious doctrine or a divine imperative' need only occasionally to 'prove' the validity of their message through magical acts revealing their possession of prophetic charisma. Things look different for the magician and the leader. For them proof is the concrete result of their enterprise. Therefore, while magicians are constantly tied to the success of their own art, leaders are continually challenged to prove their charisma through (military or political) victory. As a consequence, while prophets can expect to be believed by virtue of their own testimony and are not under an obligation of certifying their own charisma through magical acts, magicians and charismatic leaders connect their credibility to the ongoing success of their extraordinary qualities.

Naturally, in cases like this, where it is not possible to find such ideal-types in their pure status in historic reality, there are many steps between one type and the

other. Prophets, seen as an ideal-typical representation of charismatic power, are not just those who proclaim the 'new commandment.' They are also those who, 'as a prophetic blow and a blustering flame,' evoke 'the indefinable that pervades and reinforces big communities' (Weber 1985: p. 612, my translation). This reconnects the prophet to the charismatic leader, while the position toward the divinity (of which the prophet explicitly wants to be interpreter), as well as the social function exercised, keep the two types distinguished from each other. The prophet's function is in fact limited to the announcement of the divine message in the first case, while leaders are forced to maintain their position until victory is achieved.

2.5 *Protesting Character of Charismatic Powers*

The above helps us to understand the connection that Weber draws between charismatic power and situations of social crisis: The recognition of charisma, following an expectation, emerges only when a situation of social crisis looks for a solution in exceptional individuals, of whom it is possible to be devotees. From this point of view, the possession of charisma being a direct and personal gift from God, it ends up representing a kind of power implicitly in conflict with and often alternative to traditional or bureaucratic-legal power. The crisis in which the carrier of personal charisma operates is, in fact, also and mainly a crisis of leadership. In the articulation of the ideal-typical image of the 'prophet' defined as an ideal-typical case of personal charismatic power, Weber observes that the prophet almost never comes from the clergy: '… the personal call is the decisive element distinguishing the prophet from the priest. The latter lays claim to authority by virtue of his service in a sacred tradition, while the prophet's claim is based on personal revelation and charisma' (Weber 1956/1965, p. 46). This is true also in the political sector. Charismatic leaders often operate in the passage between property power and legal power and what characterizes them is not so much the content of their message as the style they use to spread it. They act driven by a duty, by virtue of which they do not present a value, but state a truth. Their message is therefore a mission they are imposing on themselves. The charisma is therefore no longer just a quality or a gift, but a 'highly asymmetric power-relationship between an inspired guide and a cohort of followers who see in him and his message the promise and anticipated achievements of a new order, to which all adhere with greater or lesser conviction' (Boudon and Bourricaud 1989, pp. 70–1). Therefore, charismatic power is based not only on the exceptional endowments of its carriers but also on the exceptional character of the social situation inside which they establish themselves, and on the promise they proclaim. Thus the analysis of the connection between charismatic power and utopia is introduced, where utopia is understood as the realization of a new social order. Often, though not always, Weber suggests, the mission of the carrier of charisma is of revolutionary character, since it reverses every hierarchy of values and clashes against usage, law, and tradition.

In the case of the prophet, as well as in that of the charismatic leader, charismatic power is tied to moments of crisis and transition and it is a temporary power. As such it is based on an uncertain and risky economic ground (a real 'charismatic economy'), often constituted by free offerings, gifts, or war booty. Carriers of personal charisma do not create any kind of systematic and permanent production of resources through a stable and lasting economy, nor do their followers. Charismatic power is therefore characteristic of the moments of transition *in statu nascenti* and is bound to disappear when social life goes back to normal.

2.6 *Transmission of Charisma*

From Weber's point of view, getting back to everyday life is, also in this case, the result of the social action of concrete social actors. In fact, the 'adherents' of the emotional community that formed itself around the carrier of charisma, as well as the 'administrative machinery' (whether constituted by followers or trustworthy people), tend to bring it about that the relation 'ends up resting, ideally and materially on a lasting fundament of unitary character.' This becomes especially evident when the problem of succession poses itself, after the demise of the carrier of charisma. The problem can be solved in different ways. Charismatic leaders can themselves, while still alive, look for a new carrier of charisma or designate a successor; the administrative machinery 'charismatically qualified' can designate a successor, or the charisma can be identified as a quality of the blood and therefore parental inheritance can affirm itself as a mechanism of succession.

Clearly, personal charisma is not always abandoned as such but it certainly tends to be subordinated to an administrative, charismatically qualified apparatus, capable of recognizing the 'signs,' or of deciding upon the norms on the basis of which the charisma can be recognized in the most rigorous possible way. Sometimes the personal charisma disappears completely, as is the case with primogeniture, which affirmed itself in the West and in Japan, where lords or monarchs are such independently of the recognition of their subjects; from this point of view, they can be completely devoid of any personal charisma.

Among these forms of transmission of charisma, of particular importance is the possibility of transmission through ritual acts, which takes a concrete form in the

charisma of function. This is the case of priestly charisma and of royal charisma, both transmitted through purification, laying on of hands, and anointing. The charismatic capacities that have been acquired this way are independent of the personal qualities of those who possess them.

The transformation of charisma in everyday practice has many consequences. For example, the establishment of the principle of the charisma's inheritance extended to the administrative apparatus implies the hereditary transmission of the powers of lordship and administration of goods, therefore founding the type of the 'aristocratic state.' Moreover, the transformation of charisma in everyday practice implies the loss of the extraeconomic character, distinctive of voluntary militant adherence, in favor of a continual acquisition and redistribution of resources. The vassals are substituted by the 'taxable subjects,' the trustworthy supporters by the party's executives, the penitential charisma of the martyrs and ascetics is replaced by the official charisma of the bishops and priests.

Weber also establishes a relation to the economy: 'But the more developed the economic interdependencies of the monetary economy, the greater the pressure of the charismatic subject's everyday needs becomes' (Weber 1956/1994, p. 44). To conclude: 'Charisma typically appears early on in the development of a religious (prophetic) or political (conquering) authority. It will give way before long, however, to routine powers as soon as its authority has been assured and, above all, as soon as it has gained sway over the masses' (Weber 1956/1994, p. 44). This way the passage takes place between the charismatic and the ordinary administration.

3. *Charisma as an Operational Concept in Contemporary Research*

Contemporary sociology has shown interest in the Weberian concept of charisma in various ways, in the frame of a general rediscovery of this author. Important studies have been carried out in the field of political sociology where a reconstruction of the 'charismatic leader' has been accomplished. From this reconstruction a new interpretation of totalitarian dictatorships, and of the crisis of contemporary democracies, has emerged. Besides this, a second stream of research can be seen in the field of the study of the routinization processes of charisma and, in particular, of the coexistence of personal and official charisma. In fact, if we admit that among the different forms of transmission of charisma there is one that rests on ritual acts carried out by a 'charismatically qualified' authority, a substantially different situation is reached. In this new situation, the faith of the believers is not directed toward a person but toward a way of transmitting an extraordinary quality, reproduced through a ritual, which is in turn guaranteed by an institution. The functional charisma does not refer to the subject anymore but to the 'belief in a state of grace, specific to the institution' (Séguy 1988). The result is a situation opposite to the initial one: 'from being a personal, unstable, and ephemeral characteristic, and by definition nontransmissible, the charisma becomes stable and enduring' (Séguy 1988, p. 18, my translation). In this way, it becomes possible to imagine a 'charismatic grace,' characteristic of the carrier of personal charisma, as opposed to an 'institutional grace.' Weber does, of course, admit the possibility that a personal charisma may hide inside a subject—for instance a priest—who carries a functional charisma. Jean Séguy does not, however, see an exceptional opposition between the two, but more a kind of kinship or elective affinity. Functional charisma is not given in an indiscriminate fashion from the charismatically qualified institution to everyone who may ask for it. It is not just the result of a charismatic education, but is founded on the pre-existence of the charisma itself inside the subject. Briefly, the institution is only a caretaker whose task is to protect the personal vocation, and therefore the gift, that the individual has received in a completely free and personal way, and to let it grow, therefore guaranteeing its realization.

4. *Methodological Issues and Possible Developments*

The concept of charisma has gained remarkable success in ordinary language. There, it has been assumed in its weak meaning, becoming synonymous with qualities that are functionally connected to the mere communication process. The charismatic person is first and foremost a communicator and the 'charisma' consists essentially in his or her capacity to communicate and convince. The current extension and complexity of mass media communication renders such qualities crucial for media success. The decisive character of this kind of success in the contemporary communication society renders the success of such weak understanding of the concept of charisma treacherous and misleading. In fact, even if the communicative qualities that explain media success are personal and free, they do not establish real leaderships.

When passing from the revelation of pure media success to the contents of such success, it is evident that it is not possible to put on the same level the success of a television journalist and that of the media preachers who were flourishing in the USA during the 1980s and 1990s. Nor is it possible to compare the latter with the media success of a political leader. Compared to the original Weberian conceptualization, the mere art of communicating and convincing certainly does not contain the social aspects of

charisma, except in the most insignificant way. Media success does not make this carrier a 'people's leader,' nor does it make the person the carrier of a mission, which in some way transcends his or her ordinary professional task. Nothing impedes the carrier of personal charisma, spiritual as well as political, 'from the plebiscitary ruler, to the big demagog, or to the leader of a political party' (Weber, 1958, p. 50, my translation) from making constant and methodical use of the potentialities of the mass media. This does not prevent the charisma—understood here as characterized by the assumption of a leadership responsibility following an absolutely personal call and in response to a crisis situation—from expressing itself completely independently from the mediagenic qualities of the subject and, therefore, of their eventual enhancement through the mass media.

The concept of charisma, correctly interpreted in its ideal-typical meaning, is far away from being just a historically born out and conceptually concluded sociological category. Even if charisma is often connected by Weber to the infancy of societies and is supposed to come before the widespread and all-embracing process of rationalization, the 'charismatic' moment continues to cross contemporary social establishments periodically, every time they find themselves in a context of crisis. Elements of 'charismatic power' are sometimes present, not only behind official charisma but also to distinguish between the functionary of a party with a strong ideological identity and who, chosen among the holders of leadership offices, will be its actual leader. While occupying the leadership position this person will trace the boundaries of the ethic of conviction, redefining the objectives. Of course, this is not enough: In their redefinition of the objective, leaders need to lean on the main principles with which they feel themselves to be personally invested. Elements of 'personal charisma' and of 'exemplary prophecy' can therefore be found inside a party secretary as well as inside every founder of a religious movement. It was exactly among these religious movements that the term 'charisma,' among others, was reintroduced into Paul's original interpretation. The charismatic movements, which came into being in the beginning of the twentieth century within Protestant circles, have been spreading within European Catholicism since the 1970s. Based on prayer, and on biblical, theological, and spiritual growth, they consider the *charisma* as a totality of gifts given by the Holy Spirit, which is able to give life to the present Christian communities in the same way as it did to the original ones.

In many parts of his work Weber admits the possibility of a charismatic presence inside rational society, founded on instrumental rationality. This is the case with some finance managers who, thanks only to their uncommonly strong personal trustworthiness, are able to obtain conventions and agreements that others can absolutely not obtain. Such phenomena are '*in ihrer ganzen Struktur, ihrem "Geist" nach, grundverschieden von der rationalen Leitung eines regulären großkapitalistichen "Betriebs"...*' However, they are part of the 'twofold nature' of what can be called 'capitalistic spirit': '*...und ebenso das Verständnis des spezifischen Eigenart des modernen, "berufsmäßig" bürokratisierten Alltagskapitalismus ist geradezu davon abhängig, daß man diese beiden, sich überall verschlingenden, im letzen Wesen aber verschiedenen Strukturelemente begrifflich scheiden lernt*' (Weber 1956, p. 667). In a word: charismatic entrepreneurship not only is not foreign to ordinary routine bureaucratized capitalism but belongs to it, even though rationality and charisma should be conceptually distinguished.

Charisma remains therefore clearly present inside rational society, even if it is intertwined with bureaucratic-legal power and the professionally qualified bureaucracy. Rather than presenting exemplars of the pure types indicated by Weber, the concept of charisma appears to be an important conceptual tool to define and explain phenomena such as personal power and consensus in the bureaucratic-rational institutional organizations. It can work as a precious revealer of situations of latent crises as well as of conflicts—latent or open—which are growing in the society as a whole and the various institutions inside it. If it is true that we are living in an epoch 'without God or prophets' it is also true that elements of personal charisma, in its various forms, do not cease to appear, mostly, even if not exclusively, in the political and religious arenas.

See also: Authority, Social Theories of; Elites: Sociological Aspects; Ideal Type: Conceptions in the Social Sciences; Institutionalization; Judaism; Leadership in Organizations, Psychology of; Leadership, Psychology of; Legitimacy; Legitimacy, Sociology of; Organizational Climate; Organizational Culture; Organizations: Authority and Power; Organizations, Sociology of; Power in Society; Power: Political; Religion, Sociology of; Weber, Max (1864–1920)

Bibliography

Boudon R, Bourricaud F 1989 Charisma. In: Boudon R, Bourricaud F (eds.) *A Critical Dictionary of Sociology*. University of Chicago Press, Chicago, pp. 69–73

Cavalli L 1981 *Il capo carismatico*. Il Mulino, Bologna, Italy

Ducros X 1937 Charisma. In: Viller M (ed.) *Dictionnaire de Spiritualité*. G. Beauchesne, Paris, Vol. 2, pp. 503–7

Séguy J 1988 Charisme de fonction et charisme personnel: Le cas de Jean Paul II. In: Séguy J et al. (eds.) *Voyage de Jean Paul II en France*. Cerf, Paris, pp. 11–34

Shils E 1958 The concentration and dispersion of charisma: Their bearing on economic policy in underdeveloped countries. *World Politics* **11**: 1–19

Shils E 1965 Charisma, order and status. *American Sociological Review* **30**(2): 199–213

Sholem G 1974 *The Messianic Ideology in Judaism and Other Essays in Jewish Spirituality*. Schoken Books, New York
Social Compass 1982 About the theory of charisma, Special issue, **XXIX**
Weber M 1947 *The Theory of Social and Economic Organization*, 1st American edn. Oxford University Press, New York
Weber M 1956 *Wirtschaft und Gesellschaft*. J. C. B. Mohr, Tübingen, Germany
Weber M 1956/1965 *Wirtschaft und Gesellschaft*. J. C. B. Mohr, Tübingen, Germany, Vol. I, Chap. V, Sect. 1 [trans. *The Sociology of Religion*. Methuen, London]
Weber M 1956/1968 *Wirtschaft und Gesellschaft*. J. C. B. Mohr, Tübingen, Germany, Vol. I, Chap. III, Sect. 10 [trans. Eisentstadt S N (ed.) *Max Weber on Charisma and Institution Building*. University of Chicago Press, Chicago
Weber M 1956/1994 *Wirtschaft und Gesellschaft* J. C. B. Mohr, Tübingen, Germany, Vol. I, Chap. III, Sect. 11–12a [trans. Heydebrand W (ed.) *Sociological Writings*. Continuum, New York]
Weber M 1958 *Politick als Beruf*. Duncker and Humbolt, Berlin
Weber M 1985 Wissenschaft als Beruf. In: Weber M (ed.) *Gesammelte Aufsätze zur Wissenschaftslehre*. J. C. B. Mohr, Tübingen, Germany, pp. 582–613
A. Zingerle (ed.) 1993 Carisma. Dinamiche dell'origine e della quotidianizzazione-Charisma. Dynamiken des ursprungs und der veralltäglichung Special issue of *Annali di Sociologia—Soziologisches Jahrbuch* **9**

S. Abbruzzese

Chemical Sciences: History and Sociology

The chemical sciences are concerned with specific kinds of matter, and their transformations. The boundaries of chemistry, notably with physics and biology, are however social constructions varying in different times and places. Chemistry is very ancient, going back into remote prehistory with cookery, the preparation of drugs and dyes, the baking of clay into ceramics, and metal-working. Its evolution into a science, where theory guides practice, and into a profession, with formal courses and qualifications, happened in the eighteenth and nineteenth centuries (Brock 1992).

1. *Ancient Technologies*

From very remote times, people have been using techniques and processes which we would call chemical, involving careful control, as part of a craft or art passed from father to son, mother to daughter, or master to apprentice. Indeed the word 'chemistry' is supposed to come from an ancient Egyptian word '*chem*' meaning earthy: the Arabic definite article '*Al*' was added to yield our 'alchemy', and dropped to give '*chymistry*' and then by 1700, 'chemistry.' Early technologies, culminating in triumphs such as the making of porcelain and Japanese swords, include features we would regard as magical; but since the course of chemical reactions depends upon the purity of components, those involving natural products are hard to predict and to repeat. Recourse to prayers, incantations, and curious additives should not amaze us.

2. *Metallurgy, Alchemy, and Pharmacy*

Alchemy, with its objective of converting base metals into gold, which chemist-historians portrayed as absurd or dishonest, was not unreasonable. Nature was believed to be perfecting metals within her womb, and the alchemist was simply speeding up the process. If everything was composed, as Aristotle believed, of the four elements Earth, Water, Air, and Fire in different proportions, then changing these ratios would transform one substance into another, and lead might become gold. If, alternatively, Democritus and Epicurus were correct in believing that in appearance there were colors, smells, and tastes, but in reality atoms and void; then because lead, gold, and everything is made up of different arrangements of these ultimately similar atoms (or 'corpuscles'), again conversions are possible.

Alchemy began in Egypt and Babylonia, and also in China: with emphasis both upon making gold (or maybe something resembling it) using an elixir to expedite the process, and of prolonging and enhancing life by giving humans the noble and permanent qualities of gold. Pharmacy grew out of trial and error, but in the West the maverick Swiss doctor calling himself Paracelsus (1493–1541) brought alchemy into it. He introduced metallic compounds into previously herbal medicine, notably for the treatment of the new disease, syphilis, which was ravaging Europe. He publicly burnt the books of the great Greek physician, Galen, and saw chemical study as essential for medicine. His career outraged the medical establishment, but the powerful and dangerous new remedies proved irresistible to doctors and patients, and medical schools became centers for chemistry.

3. *The First Chemical Theories*

Until the mid-twentieth century, it was believed that alchemy was abandoned by the rational thinkers of the Scientific Revolution; especially Robert Boyle (1627–91) and Isaac Newton (1642–1727). Close examination of their manuscripts (Principe 1998) shows that both of them were in fact adepts, copying out and trying alchemical recipes, and believing that they were well on the way to a transmutation. But they were also adherents to the atomic view of matter, seeing hard and indestructible corpuscles or particles as fundamental. These formed very stable primary mixts, such as iron, gold, or sulfur, which in turn combined with each other. Unlike gravity which was universal, chemical affinity was elective: some substances reacted together, others did not. J. W. Goethe wrote a novel,

Elective Affinities (1809) exploring chemical and human bonding; chemists in eighteenth-century Germany and Sweden (the center of chemical activity) drew up tables of affinity in attempting to predict the outcome of reactions.

From Germany also came the first chemical paradigm. G. H. Stahl (1660–1734) proposed that everything which would burn contained '*phlogiston*' (Greek, flammable): this idea brought order into chemical understanding, whereas atomic ideas were vague and untestable. Moreover, in Germany Lorenz Croll in 1778 began the first chemical journal, *Chemische Annalen*, bringing into being a chemical community there (Hufbauer 1982). His example was followed in France and England by Antoine Lavoisier and William Nicholson.

4. *Lavoisier's Revolution*

Lavoisier (1743–94) was a wealthy man, prominent in the privatized tax system of France; his spare time he devoted to chemistry, in a splendidly equipped laboratory. Becoming a member of the Royal Academy of Sciences, the small salaried body charged with scientific research, he resolved to reform the language and theory of chemistry. As in Carl Linnaeus' botany, names should be international, clear, and free from changeable theory: while *phlogiston* should be replaced as incoherent. Stahl saw *phlogiston* emitted in burning; Lavoisier by contrast (in a classic paradigm shift) saw something absorbed from the air, leading to an increase in weight. He drew upon the work of Joseph Priestley (1733–1804), who had isolated 'vital' or 'eminently respirable' air in a British tradition of work on gases. Lavoisier christened this substance 'oxygen' (Greek, sour) because he believed that it was also responsible for acidity (generalizing from analyses of nitric and sulfuric acids). Water was a compound of oxygen with another gas, hydrogen: such elements were the basis of chemistry, rather than the hypothetical corpuscles which might concern physicists, or the Earth, Water, Air, and Fire with which Priestley's friend Thomas Jefferson (1743–1826) structured his book on Virginia. In 1794 Lavoisier was executed as a tax profiteer during Robespierre's Reign of Terror, while the left-wing views of Priestley (who continued to disagree with him over *phlogiston*) led to his exile in Pennsylvania. But their new and exciting chemistry survived and prospered (Bensaude-Vincent and Abbri 1995, Knight and Kragh 1998).

5. *Electricity and Chemistry*

In 1799 Alessandro Volta (1745–1827) showed that electricity was generated when two metals were dipped into water; there was no need for any animal tissue, as Luigi Galvani (1737–98) had supposed. His paper was an alarm bell, as Humphry Davy (1778–1829) put it, and chemists everywhere repeated and extended the experiments. But results were confusing until in 1806 Davy did the careful experiments confirming his intuition that pure water is decomposed electrically into oxygen and hydrogen only. Just as Newton had found that gravity was the force behind planetary motions, so Davy inferred that electricity and chemical affinity were manifestations of one power. In 1807 he used this insight in isolating the light and reactive metals potassium and sodium, and (putting Britain back on the chemical map) went on to demonstrate, with chlorine, that Lavoisier had been wrong about acidity.

Davy had been appointed to the newly-founded Royal Institution in London's fashionable West End, where he proved himself a lecturer of enormous attractiveness, making professing a performance art (Golinski 1992, Knight 1998). The fees which men and women paid to join, and hear him, supported a research laboratory in the basement. Davy became one of the first people in Britain to make a living out of chemical research, which had previously perforce been a hobby for an aristocrat like Boyle, a minister of religion like Priestley, or a doctor like Galvani. At the Royal Institution, Davy trained (in a kind of informal apprenticeship) his successor, Michael Faraday (1791–1867), and the pursuit of Davy's insight that chemical affinity was electrical continued there.

With Lavoisier, chemistry had acquired an exact language, closer to algebra than to the evocative terms of the alchemists; and it had testable theories, for example of acidity. It is the science of the secondary qualities, of colors, smells, and tastes; it promised to be useful (chlorine for disinfecting and bleaching, for example, and oxygen for chest complaints); and it proved popular everywhere. With its connection to electricity, it became the dynamic fundamental science, concerned not just with matter but also with force; there was as yet no unified science of physics. Mechanical explanations seemed shallow; while chemistry's connections with heat, light, and electricity went deep.

6. *A Mature Science*

J. J. Berzelius (1779–1848) in Sweden used the unsystematic Davy's insight to create a structure for chemistry, 'dualism,' based on the idea that every compound had a positive and a negative part. He also picked up John Dalton's idea that each element was composed of atoms, identical to each other and different from those of other elements: Berzelius arranged these in an electrochemical series from oxygen, the most negative, to potassium. The number of elements known steadily grew through the century with improvements in chemical analysis.

Berzelius trained a number of chemists by having them to stay in his house, where Anna the housekeeper washed up dishes and flasks. But in the 1820s Justus Liebig (1803–73) at the University of Giessen launched

the first graduate school for turning out a stream of chemists with Ph.D. degrees (Brock 1997, Morrell 1997). Liebig's success depended upon his having perfected apparatus for analyzing organic compounds; his students usually did their research on some natural product, and published it in the journal which became called *Liebigs Annalen* after its editor. They found jobs, particularly in the dye industry (Fox and Nieto-Galan 1999) and in pharmacy which were both becoming based in science rather than craft skills; many went to England, a rich country with a poor educational system. With the collapse of Napoleon's empire in 1815, the University of Berlin had emerged dedicated to research and teaching (Wissenschaft und Bildung), and the various German states began to compete in their opera houses and universities. They followed Giessen, building better laboratories and bidding for star chemists. Schools began teaching chemistry, textbooks were needed (Lundgren and Bensaude-Vincent 2000), and academic careers opened up in a field now largely separated from medicine. Universities in Britain and the USA followed the German model, usually demanding German research experience from professorial candidates.

In the 1850s chemists could agree about what things were made of, but not about formulae. Dalton had supposed that water must have the simplest possible formula, HO; Davy and others, notably Amadeo Avogadro (1776–1856), went for our H_2O formula because two volumes of hydrogen combine with one of oxygen. An atom of oxygen thus weighed either 8 or 16 times as much as one of hydrogen, and such uncertainties ran through the whole list of elements. In 1860 August Kekule (1829–96), a pioneer in working out chemical structures such as that of benzene, called for an end to this confusion through an international conference, which met in Karlsruhe. It was poorly organized, but afterwards chemists came to accept the reformulation of Avogadro's arguments by Stanislao Cannizzaro (1826–1910). With agreed atomic weights, tabular arrangement of the elements became possible; and the most successful was the Periodic Table of Dmitri Mendeleev (1834–1907).

His predictions of the properties of some hitherto undiscovered elements were startlingly accurate; and with the table (as he hoped) the student had to remember fewer brute facts. From its position an element's properties would be known. It is striking that so many bright ideas, from Dalton via Cannizzaro to Mendeleev, came from people on the periphery rather than in the great scientific centers.

7. *The Fragmentation of Chemistry*

In death, we rot: for we (and animals and plants) are then subject to chemical reactions which go differently while we are alive. Most people believed in a vital force which maintained life. It is claimed that Friedrich Woehler (1800–82), pupil of Berzelius and friend of Liebig, destroyed this vitalism when in 1828 he synthesized urea. In fact the chief interest in this reaction was that ammonium cyanate and urea turned out to have the same atomic constitution: their different properties were the result of different arrangements (Brooke 1995). So the story has more to do with understanding molecular structure; but the synthesis, and the work of Liebig and his students in analysis, showed that no gulf separated organic and inorganic worlds. Nevertheless, by 1848 when Berzelius died, it was clear that dualism did not fit organic compounds well, and as the chemical community grew it was convenient to separate organic chemistry, based upon carbon, from the inorganic branch. The expansion of universities led to new professorships and laboratories devoted to the specialism of organic chemistry, from which in the twentieth century emerged biochemistry.

Chemists had relied upon balances, test-tubes, condensers, blowpipes, and other apparatus difficult to manipulate. The chemist had to think with his (or occasionally her) fingers, and was proud of skills in glassblowing. Chemistry was essentially experimental, exciting and often dangerous, attractive. Then in 1860 came collaboration between Robert Bunsen, inventor of the controllable gas burner, and the physicist G. R. Kirchhoff, who found that elements heated to a high temperature have characteristic spectra. Analysis could be done by physical methods, and this optical spectroscopy was the first of what is now an armory of such techniques which has transformed the appearance of chemical laboratories (Morris and Travis, in Krige and Pestre 1997, pp. 715–40).

About the same time the new science of thermodynamics, based on energy and its transformations, brought together into classical physics sciences which had been separate, or had been part of the empire of chemistry, Davy and Faraday had been pioneers in what became a new specialism, physical chemistry, investigating energy changes in reactions, and the mechanisms, rates, and reversibility of processes. The leaders here were Wilhelm Ostwald (1853–1932) and J. H. Van't Hoff (1852–1911) who launched a journal, and promoted academic positions and laboratories. The new profession of chemical engineering was closely linked to the rise of physical chemistry. Whereas early in the nineteenth century chemists had been called in only as consultants or trouble shooters when something went wrong, by the end of it they were employed full-time (Bud and Roberts 1984). In industry, intellectual property belongs generally to the company and not the individual, and is secured by patents (Travis et al. 1998).

8. *The Reduction of Chemistry*

The nineteenth century was the heyday of chemistry, the golden age in which it came to maturity and seemed fundamental. The chemist and spectroscopist

William Crookes (1832–1919) followed Faraday in studying cathode rays, but J. J. Thomson in 1897 identified them as composed of subatomic corpuscles, soon named 'electrons.' The subsequent nuclear atomic model of Ernest Rutherford (1871–1937)—for whom all science was physics or stamp-collecting—and Niels Bohr (1885–1962) accounted not only for spectra, but also for the Periodic Table. Chemistry became a branch of physics (Nye 1996); the properties of gold could in principle be calculated from data about protons, neutrons, and electrons, though in practice the chemistry laboratory is essential. This meant that chemistry lost its glamor; the chemist was as ubiquitous as ever, an essential member of the teams or groups so characteristic of twentieth-century science, but playing a service role (Knight 1995).

The number of chemists has continued to grow, as has the number of new substances unknown in nature which they have synthesized. Davy wrote of the chemist being a godlike creator, and this creativity is nowadays celebrated by chemists such as Roald Hoffmann. The engineer or architect must remember the law of gravity, but to mourn that architecture has been reduced to physics would be absurd: like the poet or the painter, the chemist has to work within constraints, but that is a feature of life—indeed making creativity possible (Hoffmann and Torrance 1993).

Hoffmann, born in Poland, surviving World War II, escaping to the USA, learning chemistry there, and doing research which brought him a Nobel Prize, exemplifies another trend. Chemistry reached the West via Islam. By the eighteenth and nineteenth centuries, it was a European science, with Germany the most important center by 1900. Papers in German journals, and research experience in Germany, counted high in any pecking order; but already the USA was becoming a major power in science. There in 1916 G. N. Lewis proposed the electronic theory of chemical combination, much developed by his pupil Linus Pauling. Since 1945 the USA has been the center of things, making the English language and publication in US journals the key to prestige in research. Two world wars, and Hitler's coming to power between them, are part of the reason for this; equally important has been American prosperity, itself dependent on science. Chemistry has steadily gone West.

9. *The Status of Chemistry*

Davy and Liebig (Rossiter 1975) wrote famous books on agricultural chemistry, and in the nineteenth century chemical fertilizers and pesticides were unequivocally welcomed in a Europe of food shortages. Lavoisier improved French gunpowder, and later chemists produced high explosives making possible engineering achievements and also formidable weapons. All these things were seen as benefits. National chemical societies, their academic and professional aspects sometimes in tension, were formed and enjoyed prestige (Russell et al. 1977). Although pollution from new chemical industries (as from older ones like tanning) was palpable and led to legislation, the expectation was that the chemists would be able to cure it. It did not happen: Rachel Carson's book *The Silent Spring* (1962) alerted the world to the dangers. So in the late twentieth century, despite successes such as plastics, and the array of new drugs available for medicine, chemistry is seen as boring and its applications as threatening. Chemists feel misunderstood and underappreciated.

Twentieth-century chemistry is dominated not only by universities in the Giessen tradition, but also by big-spending international companies with research laboratories, now turning towards biotechnology (Galentos and Sturchio, and Kevles, in Krige and Pestre 1997, pp. 227–52, 301–18) and by the military. Research is carried on no longer by a Woehler or a Crookes, on their own or with an assistant, but by teams of people possessing various skills. Chemistry has been taught in an impersonal way, with less hands-on experiment in a world more conscious of health and safety.

This is a strange eventful history, which was until the mid-twentieth century mainly written by participants who looked for progress. They had the advantage of being familiar with chemicals and apparatus; but professional historians of science have come to look more closely at contexts and careers. The history which emerges deserves to be known beyond the world of chemists.

See also: Archaeometry; Behavioral Neuroscience; Biomedical Sciences and Technology: History and Sociology; Ceramics in Archaeology; Cognitive Neuroscience; History of Science; Human Sciences: History and Sociology; Physical Sciences: History and Sociology; Research and Development in Organizations; Scientific Disciplines, History of; Technological Innovation

Bibliography

Bensaude-Vincent B, Abbri F 1995 *Lavoisier in European Context*. Science History, Canton, MA

Brock W H 1992 *The Fontana History of Chemistry*. Fontana, London

Brock W H 1997 *Justus von Liebig: the Chemical Gatekeeper*. Cambridge University Press, Cambridge, UK

Brooke J H 1995 *Thinking about Matter*. Ashgate Variorum, Aldershot, UK

Bud R, Roberts G K 1984 *Science versus Practice*. Manchester University Press, Manchester, UK

Fox R, Nieto-Galan A 1999 *Natural Dyestuffs*. Science History, Canton, MA

Golinski J 1992 *Science as Public Culture: Chemistry and Enlightenment in Britain, 1760–1820*. Cambridge University Press, Cambridge, UK
Hoffmann R, Torrance V 1993 *Chemistry Imagined*. Smithsonian, Washington, DC
Hufbauer K 1982 *The Formation of the German Chemical Community 1720–1795*. California University Press, Berkeley, CA
Knight D M 1995 *Ideas in Chemistry*. Athlone, London
Knight D M 1998 *Humphry Davy: Science and Power*. Cambridge University Press, Cambridge, UK
Knight D M, Kragh H 1998 *The Making of the Chemist*. Cambridge University Press, Cambridge, UK
Krige J, Pestre D 1997 *Science in the Twentieth Century*. Harwood, Amsterdam
Lundgren A, Bensaude-Vincent B 2000 *Communicating Chemistry: Textbooks and their Audiences, 1789–1939*. Science History, Canton, MA
Morrell J 1997 *Science, Culture and Politics in Britain, 1750–1870*. Ashgate Variorum, Aldershot, UK
Nye M J 1996 *Before Big Science*. Twayne, New York
Principe L 1998 *The Aspiring Adept*. Princeton University Press, Princeton, NJ
Rossiter M 1975 *The Emergence of Agricultural Science: Justus Liebig and the Americans, 1840–1880*. Yale University Press, New Haven, CT
Russell C A, Coley N G, Roberts G K 1977 *Chemists by Profession*. Open University Press, Milton Keynes, UK
Travis A S, Schröter H G, Homburg E, Morris P J T 1998 *Determinants in the Evolution of the European Chemical Industry, 1900–1939*. Kluwer, Dordrecht, The Netherlands

D. Knight

Chess Expertise, Cognitive Psychology of

Expertise may be defined as the ability of some individuals to perform at levels vastly superior to the majority. For historical and scientific reasons, research on chess expertise has played a major role in the study of expertise in general. The first reason is that chess itself has a very long history (the modern form of Western chess goes back to the sixteenth century). This has made possible an extensive study of the game, leading to the development of several 'theories' about the proper way to play by leading players such as Steinitz, Nimzowitch, and Euwe. Next, the rules of chess offer a well-specified and constrained environment that is easily formalizable. Chess is also a game flexible enough to allow multiple experimental manipulations. In addition, the presence of a rating system (the Elo system) allows one to estimate players' skill quantitatively and precisely. Compared to most other domains of expertise, this ability to measure skill is a definite advantage. Contrast this situation with, for example, the study of experts in physics or medicine, where researchers have to use very rough classifications such as novice, intermediate, and expert. Finally, there has been rich cross-fertilization between psychological research on chess expertise and research in formal fields like computer science and mathematics.

1. Chess in the Sciences

1.1 Chess in the Formal Sciences

Unsurprisingly, chess has been a favorite subject of study in the formal sciences. On several occasions, chess has been used to explore aspects of game theory; in a celebrated paper published in 1912, Zermolo formalized the concept of game tree and introduced the method of backwards induction with reference to chess. The game has also been of interest to mathematicians, for example in the field of combinatorics. However, most of the research has been made in artificial intelligence and computer science. If one ignores chess automata, most of which turned out to be fraudulent, computer chess started in earnest in 1949 with Shannon's paper describing a computer program able to play an entire game, either by full search to a specified depth or by selective search. Since that seminal work, researchers have extensively explored various techniques for improving the efficiency of search algorithms or to make search more selective (see Newell and Simon 1972, Levy and Newborn 1991). The crowning achievement of the quest for efficient search algorithms (the so-called 'brute-force' approach) was the development of Deep Blue, the first computer to beat a world champion in an official match. Deep Blue's special-purpose hardware allowed it to consider up to 200 million positions per second. By contrast, a nice example of the selective-search approach is a program written by Pitrat (1977), which uses heuristics to cut the search tree down to the same size as humans' (about 100 positions). Recently, computer chess has seen a strong interest in database theory and in the development and testing of machine-learning algorithms.

1.2 Chess in the Social and Behavioral Sciences

Expert behavior in chess has attracted the attention of various social and behavioral sciences, including psychoanalysis, psychiatry, and sociology. Questions such as 'Does extreme practice of a skill lead to madness?' 'Why are women weaker than men at chess?', 'Can oedipal compulsions lead to creativity?', and 'Why is there a high proportion of Jews among top players?' have been asked in these fields, although the answers offered are often controversial (Dextreit and Engel 1981, Holding 1985). In addition, chess has sometimes been used not as an object of study, but as

a model. Two examples may suffice. In philosophy, Lasker developed a philosophical system (machology) based on the element of fight extant in a chess game. In linguistics, De Saussure made regular use of chess to illustrate the rule-like character of language. For instance, he drew an analogy between a chess game and the synchronic analysis of language: if somebody walks into a room where a chess game is being played, they can study and understand the position without knowing the moves leading to it.

2. Psychological Studies of Chess Expertise

However, in none of the social and behavioral sciences mentioned above has chess had such an impact as in psychology, where it has been used as a standard task environment for exploring expertise (Simon and Chase 1973). Several concepts, such as progressive deepening and selective search, and several experimental techniques, such as the use of verbal protocols and the use of recall tasks to study expertise, have their main source in chess research.

2.1 Brief History of Psychological Research

The first psychological investigation of chess expertise was carried out at the end of the nineteenth century by Binet, who was interested in masters' ability to play a game, or even several games, without seeing the board. This work, using questionnaires and focusing on the search for chess players' hypothetical, and probably nonexistent, concrete visual memory, no longer has much impact. Nor does the work carried out in 1927 by Djakow, Petrowski, and Rudik, who were the first psychologists to bring chess players into the laboratory. These Russians scientists, who gave a battery of psychometric tests to their subjects, were interested in the question of chess talent. Their tests measured various 'faculties of mind,' such as memory, attention, or combination power. It turned out that chess masters did not differ from lay people on most of these tests, the exception being tasks of visual memory where the stimuli bear a strong resemblance to chess-boards.

The next wave of research on chess psychology had a huge impact on expertise research and on cognitive psychology in general. It was the work of a single man, the Dutch psychologist and chess master Adriaan De Groot. In his doctoral thesis in 1946, De Groot (1978) introduced two key methods (recall of briefly presented positions, and analysis of verbal protocols collected during problem solving), which allowed him to uncover several key determinants of expertise. First, players at all levels are highly selective in their search. Even top-level players do not search much more than about one hundred positions during a 15-minute deliberation. Second, there exist almost no differences between the search behavior of world-class grandmasters and that of weaker players, as far as measures such as depth of search, breadth of search, or branching factors are concerned. Third, players of all levels visit the same variation several times when choosing a move, a phenomenon De Groot called 'progressive deepening.' This behavior allows cognitive systems of limited capacity, such as those of humans, both to propagate information from a given branch of the search tree to other branches, and to overcome the limits of short-term memory (De Groot and Gobet 1996). Fourth, masters are able to zoom into the key features of the problem at hand very rapidly. Fifth, chess masters normally judge a position based on a single feature, which can be either static (e.g., material balance) or dynamic (e.g., potential actions). This is in sharp contrast to computer programs, which typically use a polynomial function to combine a large number of features (Levy and Newborn 1991). Sixth, chess masters have a remarkable memory for meaningful material taken from their domain, even when this material is presented for just a few seconds.

De Groot's work, and through it the ideas of the German psychologist Otto Selz, was important in shaping the revolution in cognitive psychology (Newell and Simon 1972). As the cornerstone of most current research on expert behavior, it has spawned a large number of empirical studies, the most important of which are reviewed in the next section.

2.2 Key Empirical Results

It is common to organize chess research along the following lines: perception, memory, knowledge, look-ahead search, and general intelligence. Research on perception has confirmed De Groot's earlier results, and has shown that players can identify and memorize critical patterns in a position even with presentation times below one second. Automatization also affects lower levels of perception: As shown by Saariluoma (1995), the speed with which chess pieces are recognized is a function of the level of expertise. Finally, eye-tracking studies have shown that chess masters have faster eye movements, cover more of the important squares on the board, and tend to look at the intersection of squares more often than weaker players (De Groot and Gobet 1996).

Empirical studies on memory have shown that chess players' performance is mediated by variables such as depth of processing, presentation time, typicality, level of randomization, and age (for reviews, see Holding 1985, Saariluoma 1995, Gobet 1998). Interestingly, and contrary to widely held opinion, masters' superiority is maintained with briefly presented positions, even after their semantics have been destroyed by randomizing the location of pieces (Gobet

and Simon 2000). While the effect is small, it is also reliable, and has been found in other domains of expertise such as programming and music.

As witnessed by the formidable time chess masters spend studying books and analyzing games, knowledge plays an important role in chess expertise. Several researchers have attempted to study knowledge directly, using techniques such as sorting experiments, questionnaires, and verbal reports. There is evidence that expertise correlates both with qualitative organization of knowledge and with quantitative amount of information stored (Freyhoff et al. 1992), reflecting findings of other domains of expertise such as physics.

While De Groot's statistics about problem solving have in most cases withstood the test of time, subsequent research has identified a few skill differences. In particular, it has been found that depth of look-ahead search varies as a function of skill (e.g., Saariluoma 1995), although the effect is rather small.

Several studies have addressed the question of whether intelligence correlates with chess skill. The general conclusion is that there is a correlation with tests measuring general intelligence, but not with tests measuring visuospatial intelligence (Doll and Mayer 1987). From these studies, it is unclear whether chess practice develops aspects of intelligence measured in IQ tests (a good candidate for explanation would be ability to think under time pressure), whether attainment of a high skill requires superior intelligence, or whether both intelligence and chess skill are causally related to a third variable, such as an ability to concentrate for long periods.

Finally, chess has been a useful domain for studying the cognitive processes that support outstanding skill in problem solving across the life span. In particular, chess has been useful in identifying compensatory mechanisms used by older adults to allow high-level performance in spite of age-related declines in perceptual, memory, and cognitive abilities (Charness 1981).

3. *Theories of Problem Solving*

Given the rich set of empirical data generated by chess research, it is not surprising that chess has produced several theories of expertise. It is convenient to present these theories as mainly addressing either problem solving or memory, although they often aim at a general characterization.

The first theory devoted primarily to problem-solving behavior in chess is De Groot's (1978) elaboration of Otto Selz's framework of productive thinking. Selz proposed that thinking is a continuous activity that can be described as a linear chain of operations. De Groot showed that this framework was, with a few extensions and modifications, quite successful in accounting for chess thinking. As anticipated by Selz, players often use a hierarchy of subsidiary methods, which relates to Newell and Simon's (1972) means-end analysis. For De Groot (1978) a necessary condition for becoming a chess master is the construction through experience of two things: a highly developed and specific mode of perception, and a system of reproductory methods stored in memory.

Chess players' memory can be separated into knowledge (knowing that...) and intuitive experience (knowing how...). Selz's description of thought as a sequence of operations is also apparent in the formal models developed by the Carnegie Mellon research group centered around Herbert Simon (this research is summarized in Newell and Simon 1972). Two computer programs (written by Newell, Shaw, and Simon in 1963, and by Baylor and Simon in 1966) implemented the idea of selective search made possible by the use of heuristics. Another model, developed by Newell and Simon in 1965, used the evaluation obtained at the end of a branch to formulate six principles that dictate the generation of moves and sequences of moves. Several of these ideas have been recently combined with the chunking theory (see below) in a probabilistic model of chess thinking (Gobet 1997).

Two informal theories have also been influential in research on expertise. Holding's (1985) theory emphasizes the role of search and knowledge and suggests that human experts search in ways similar to computers. The theory discussed by Saariluoma (1995) proposes that players, while thinking about a position, access goal positions by apperception—that is, conceptual perception. They then try to close the path between the problem position and the goal position. When this is not possible, the problem space is restructured. Thus, chess thinking may be described as a sequence of apperception–restructuration cycles, which make it possible to find solutions with only limited search.

While these theories differ in their emphasis and in their specificity, going from formal computer programs to informal verbal theories, they also share a few important assumptions: They all emphasize the role of knowledge in problem solving and the high selectivity of human search.

4. *Theories of Memory*

One can identify four major theories of chess perception and memory in chess expertise (Gobet 1998). The most influential of them is Simon and Chase's (1973) chunking theory, which proposes that experts acquire a vast database of chunks (perceptual patterns that can be used as units) giving access to semantic

memory and to procedural memory. Subsets of the chunking theory were implemented in 1973 in a computer program by Gilmartin and Simon. Based on extrapolations from the simulations, it was estimated that between 10,000 and 100,000 chunks are necessary to reach a high level of expertise. However, some weaknesses of chunking theory were uncovered by later research, mainly the fact that it underestimates storage into long-term memory (LTM), and that it underestimates the role of high-level structures such as schemata (Charness 1976, Holding 1985).

Several attempts have been made to repair these theoretical weaknesses, while still accounting for the data that chunking theory successfully explained. Holding (1985) is representative of a group of researchers emphasizing the role of high-level knowledge structures, such as schemata or prototypes. In opposition to the simple and specific structures proposed by Chase and Simon, Holding emphasizes that chess masters' memories are richly organized and general. Emphasis is also given to metaknowledge, which consists of principles for efficient search and evaluation of positions.

Ericsson and Kintsch's (1995) long-term working memory theory emphasizes that experts in various fields, including chess, can encode information into LTM more rapidly than had been postulated by traditional models of human memory. In the spirit of Selz and Newell and Simon's (1972) approaches, this theory views cognitive processes as a sequence of stable states representing end products of processing. Through acquired memory skills, these end products can be stored in LTM and can be accessed from short-term memory by means of retrieval cues. Two intertwined mechanisms allow rapid storage into LTM. The first mechanism allows encoding through a hierarchical retrieval structure; in the case of chess, the retrieval structure corresponds to the 64 squares of the board. The second mechanism allows encoding through knowledge-based associations that elaborate patterns and schemata stored in LTM. Ericsson and Kintsch (1995) suggest that these two mechanisms account for chess masters' excellent memory for chess material, as well as for their ability to plan and to evaluate alternative sequences of moves.

Gobet and Simon's (2000) template theory proposes that patterns recurring often in players' practice and study lead to the creation of more complex data structures, called templates. As with classical schemata, templates have both a core (containing the same information as chunks), and slots, where variable information can be stored. The template theory is implemented as a computer program, which acquires chunks in an unsupervised way by scanning a large database of master games. The program can simulate various data from perception, such as players' eye movements during the first five seconds of presentation of a position, and from memory, such as the role of presentation time on memory recall. A version of the program also plays (poorly) by pure pattern recognition.

One of the important challenges facing researchers in the field is to provide a coherent and integrated picture of chess thinking, combining simple perceptual structures such as Chase and Simon's chunks with more complex structures such as schemata. Long-term working memory and template theory can be seen as attempts to address this challenge.

5. *Future Lines of Research*

Chess expertise has been extensively studied in the past (more research has been done in psychology about chess than about all other games put together), and it is likely that this will continue in the future. In most cases, results from chess research can be generalized to other domains of expertise. While scientific understanding of expertise has grown substantially through a large number of experiments and through a wealth of theoretical developments, there is still no single theory able both to account for most of the empirical data and to simulate human behavior by playing chess at a high level of expertise. With the rapid progress in artificial intelligence and computer science, which have already produced grandmaster-level programs by brute force, it is however realistic to expect such a theory within a decade or two. In addition to this effort in computer modeling, a main research domain is likely to be neuropsychological investigation using brain imaging techniques to study the biological substrate of chess expertise.

See also: Expert Memory, Psychology of; Expertise, Acquisition of; Medical Expertise, Cognitive Psychology of; Protocol Analysis in Psychology; Short-term Memory, Cognitive Psychology of; Sports as Expertise, Psychology of; Working Memory, Psychology of

Bibliography

Charness N 1976 Memory for chess positions: Resistance to interference. *Journal of Experimental Psychology: Human Learning and Memory* **2**: 641–53

Charness N 1981 Aging and skilled problem solving. *Journal of Experimental Psychology: General* **110**: 21–38

De Groot A D 1978 *Thought and Choice in Chess*, 2nd edn. Mouton Publishers, The Hague, The Netherlands

De Groot A D, Gobet F 1996 *Perception and Memory in Chess*. Van Gorcum, Assen, The Netherlands

Dextreit J, Engel N 1981 *Jeu d'échecs et sciences humaines* [*Chess and Human Sciences*]. Payot, Paris

Doll J, Mayer U 1987 Intelligenz und Schachleistung—eine Untersuchung an Schachexperten [Intelligence and success in chess playing—an examination of chess experts]. *Psychologische Beiträge* **29**: 270–89

Ericsson K A, Kintsch W 1995 Long-term working memory. *Psychological Review* **102**: 211–45
Freyhoff H, Gruber H, Ziegler A 1992 Expertise and hierarchical knowledge representation in chess. *Psychological Research* **54**: 32–7
Gobet F 1997 A pattern-recognition theory of search in expert problem solving. *Thinking and Reasoning* **3**: 291–313
Gobet F 1998 Expert memory: A comparison of four theories. *Cognition* **66**: 115–52
Gobet F, Simon H A 2000 Five seconds or sixty? Presentation time in expert memory. *Cognitive Science*
Holding D H 1985 *The Psychology of Chess Skill*. Erlbaum, Hillsdale, NJ
Levy D, Newborn M 1991 *How Computers Play Chess*. Computer Science Press, New York
Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ
Pitrat J 1977 A chess combination program which uses plans. *Artificial Intelligence* **8**: 275–321
Saariluoma P 1995 *Chess Players' Thinking*. Routlege, London
Simon H A, Chase W G 1973 Skill in chess. *American Scientist* **61**: 393–403

F. Gobet

Chiefdoms, Archaeology of

The use of the term 'chiefdom' to refer to pre-state complex societies is a relatively recent phenomenon, beginning with Kalervo Oberg's historical classification of South and Central American societies and Marshall Sahlins's ethnographic work on Polynesian societies in the 1950s. Elman Service included 'chiefdoms' a decade later in his neoevolutionary categories of *band*, *tribe*, *chiefdom*, and *state*, and the concept became closely associated with the cultural evolutionary paradigm. Almost immediately, archaeologists began to apply the chiefdom concept to investigations of early societies such as the Mississippian towns of the southeastern USA and the Neolithic European builders of Stonehenge, which manifested some complexity but generally lacked the long-held archaeological indicators of 'civilization' (e.g., large urban centers, massive kingly tombs, written dynastic histories and legal codes, markets, and monetary systems).

1. *Chiefdoms and Cultural Evolutionary Theory*

Chiefdoms have been defined by some archaeologists in primarily political terms, as smaller-scale complex societies with centralized decision-making hierarchies of one or two levels above the individual village (compared to three plus levels in states). Chiefly leaders obtain their political authority through both ascription and performance (compared to largely hereditary kings), they have very generalized administrative roles (judicial, economic, ritual, military), they have few bureaucratic specialists (unlike the specialized bureaucracies of states), and they generally rely on kin-based alliances to structure political relations with subordinates (rather than the nonkin-based political institutions and formal legal codes of states). Most archaeologists would expand the definition to include social ranking that has at least some hereditary component (unlike the achieved ranking of tribal societies and the almost wholly hereditary social classes of early states) and some degree of economic centralization (control over staple production, building and managing irrigation systems, tribute mobilization, control over the production and exchange of prestige goods, and/or foreign trade monopolies). Chiefs generally maintain their political power not only through varying strategies for accumulating and disbursing material resources, but also through ideological means (manipulating cosmologies, ritual, and myth, and creating power-imbuing sacred landscapes) and military coercion. Chiefdoms rarely exist or evolve in isolation, but instead are found as clusters of interacting *peer polities* that often share aspects of elite culture, have similar structures, and compete for regional supremacy. Many chiefdoms are also part of larger *world systems* that link them to more developed states and empires.

Chiefdom-level societies have been identified by cultural anthropologists, archaeologists, and historians in many parts of the world and in many time periods (Fig. 1), often as precursors to state development but sometimes as long-term, stable structures that do not transform into more complex states (see *States and Civilizations, Archaeology of*). The Polynesian chiefdoms, the pre-Roman (Late Neolithic, Bronze Age, and Iron Age) societies of Europe, and eastern North American complex societies (particularly Mississippian Period) have received the most intense archaeological study. Archaeologists have recently expanded their use of the chiefdom model to societies in the Near East, Africa, East Asia, South Asia, Mesoamerica, South America, and western North America (Fig. 1), although the 'evolutionary status' of many as chiefdoms is controversial (e.g., the Woodland Period societies of the eastern USA, the Late Neolithic societies of Europe). Early in the development of the chiefdom concept, archaeologists identified a number of common material correlates of these types of societies (e.g., two- to three-level settlement hierarchies, monumental architecture, burials with hereditary status indicators, specialist-manufactured prestige goods, regional and interhousehold differences in wealth; see Fig. 1). However, as more of these complex societies are intensively studied by archaeologists, it has become clear that material patterns, and presumably the ideological, political,

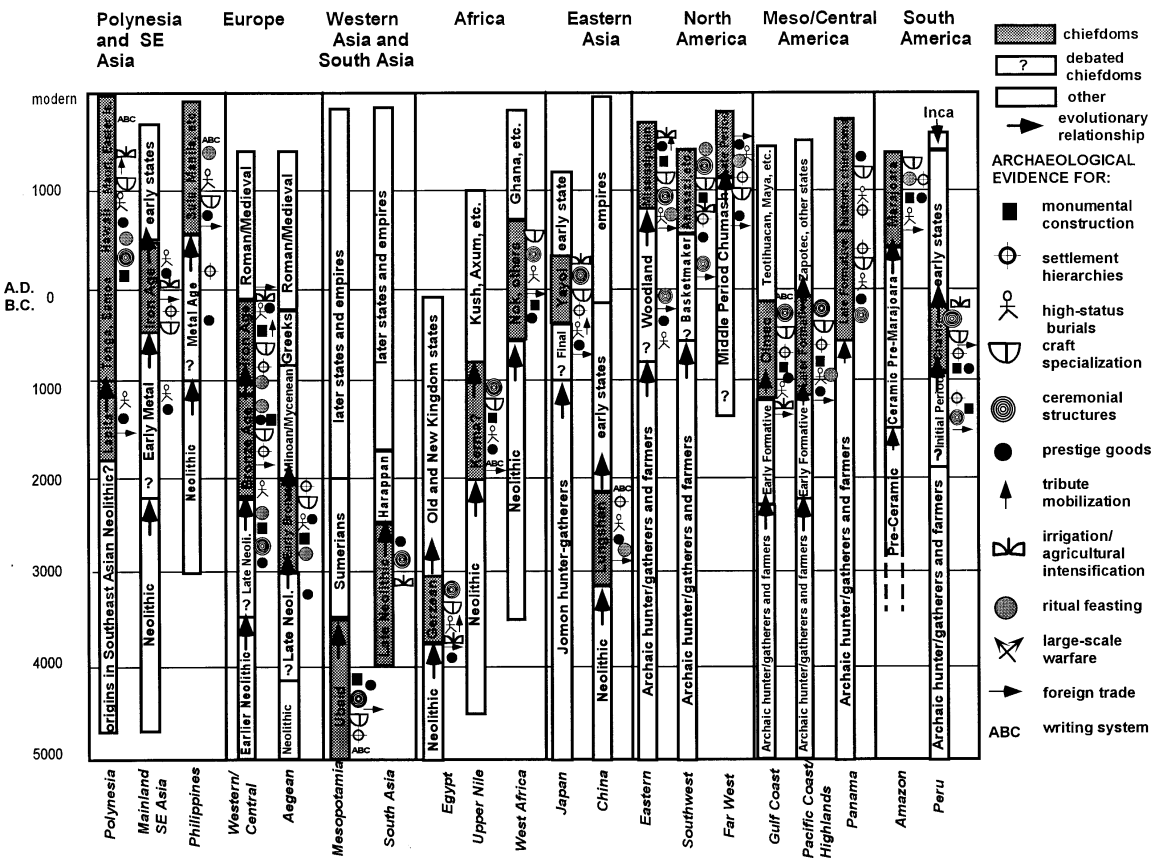

Figure 1
The chronological and geographic distribution of some chiefdoms known through archaeological and ethnographic research, with varying archaeological indicators

social, and economic structures underlying them, vary substantially (see below).

2. *Critiques and Redefinition of the Chiefdom Concept*

A number of recent critiques of the chiefdom concept, and cultural evolutionary models in general, have emphasized the organizational diversity and historical uniqueness that is ignored when anthropologists attempt to fit societies in broadly defined classifications and assume uniform trajectories of development. Many scholars have pointed out the great deal of variability in societies designated as 'chiefdoms.' One response to this recognized diversity has been to subdivide pre-state complex societies into developmentally distinct types, such as Robert Carneiro's 'minimal,' 'typical,' and 'maximal' chiefdoms (based on the complexity of political hierarchies), Timothy Earle's 'simple' and 'complex' chiefdoms (based on scales of chiefly political, economic, and ritual control), and Colin Renfrew's 'group-oriented' and 'individualizing' chiefdoms (based on differing strategies of lineage versus individual aggrandizement and social display).

Other scholars have emphasized that transformations in social, political, economic, and ideological structure that have previously defined chiefdoms do not always occur together. Some archaeologists have promoted the concept of *heterarchy*—in which both hierarchical and nonhierarchical relations operate simultaneously and separately along multiple dimensions (social, economic, political, ritual)—as a better way of modeling structure and process in these societies. Other archaeologists have favored a move away from the systemic evolutionary processes implied in the chiefdom concept to 'actor-based' approaches, Marxist theories of political control, and nonmaterialist or postprocessual analyses focused on the ideological and symbolic bases for structure. Thus, while many archaeologists favor continual refinement of the

chiefdom concept and its continued utility in archaeological research, others have advocated its abandonment.

3. *Political and Social Features*

Chiefdoms are characterized by partially inherited and partially achieved political leadership, reinforced through ideological manipulation, control of economic resources, and militarism. The political power of chiefs is often reflected materially in symbols of office (e.g., the jade axes of Maori chiefs and the stone maceheads of Gerzean Period Egyptian chiefs) and in monumental architecture which demonstrates a chief's ability to control and mobilize large labor forces (e.g., the henge monuments of Late Neolithic Europe, the monumental statuary of the Olmec and Easter Islanders). Chiefly centers, as the focal point of political, economic, and religious activities of chiefs, are generally archaeologically distinguishable from subordinate villages by their greater size, centrality, and presence of monumental architecture, wealth objects, and specialized production locales. Multiple levels of political authority (*paramount chiefs* and local chiefs), typical of larger-scale chiefdoms, are often archaeologically visible in regional settlement hierarchies of up to three levels.

A number of archaeologists have offered new conceptual frameworks for analyzing differing political power strategies in chiefdoms. One recent model, offered by Richard Blanton and Gary Feinman, contrasts *network* versus *corporate* strategies of political dominance, elaborating Colin Renfrew's earlier distinction between *individualizing* and *group-oriented* chiefdoms. These differing strategies are viewed as part of political dynamics in all complex societies, but political actors in a particular society may emphasize one mode of control more than the other. In a *network strategy*, political actors try to create personal networks of dominance through the strategic distribution of portable wealth and symbolic capital (e.g., ritual potency and religious knowledge). The emphasis on individual aggrandizement is archaeologically manifested in lavish individual burials, elaborate household wealth display, and competitive presentational events such as ceremonial feasting (e.g., Iron Age chiefdoms of Europe, early West African chiefdoms). The highly conflictive nature of leadership and long-term instability of political configurations are evidenced in rapidly changing settlement hierarchies. In contrast, a society emphasizing a corporate political strategy disperses power across different groups and sectors of society through bureaucratic institutions promoting consensus, solidarity, and collective action, often reinforced through archaeologically visible public architecture, collective tombs, and unifying emblems (e.g., Late Neolithic chiefdoms of Europe, Woodland Period chiefdoms of North America).

In terms of social structure, chiefdoms invariably have some form of ascribed social ranking. While social status differences are characteristic of all human societies, social ranks in chiefdoms are at least partially inherited intergenerationally. They are given structural rigidity through behavioral taboos and symbolic expression, and in most cases, there is distinct economic advantage (differential access to resources) conferred on an 'elite' stratum. Social stratification in chiefdoms is often manifested in archaeologically visible rank insignia, differential wealth in households, and varying complexity of residential architecture at settlements.

Ascribed rather than strictly achieved social ranking is most archaeologically evident in burial practices. Inherited status is generally expressed through variation in body positioning, body treatment, grave forms, and burial accompaniments which cross-cut age and sex categories, making status archaeologically distinguishable from mortuary variability, attributable to age differences, gender roles, and achieved status. Mortuary analyses based on this premise have been carried out most successfully at Mississippian, Chinese Lungshan, Egyptian Gerzean, and Bronze Age and Iron Age European cemeteries, where relatively large samples of well-preserved burials have allowed the identification of distinct social hierarchies. In cases where larger burial populations are lacking, archaeologists have frequently made an argument for the presence of elites based on finds of elaborate, often monumental 'chiefly' burials (e.g., jade-yielding Formative Period Mesoamerican tombs, the Nok mounded tombs of West Africa, the Yayoi stone-lined crypts of Japan). Status-related dietary and health differences between elites and nonelites have also been studied through zooarchaeological and paleobotanical analyses of food remains at settlements, and through osteological assessments of nutritional stress and disease in burial populations.

4. *Political Economy*

Just as the chiefdoms studied by archaeologists vary in political structure, they differ in the ways in which chiefs attempt to assert economic control in support of their governing institutions. Timothy Earle and Terrence D'Altroy have made a useful distinction between strategies of *staple finance* and *wealth finance* in chiefdoms. *Staple finance* involves the systematic levying of tribute payments (in the form of staple goods and/or labor) from subjugated commoners, with the revenue used to finance directly the power-enhancing activities of the chief, such as monument building, warfare, trade, and ceremonial feasting. Since control of land and its production is essential to staple finance, chiefdoms emphasizing this economic strategy generally have strongly developed institutions

for land tenure (archaeologically manifested in boundary features such as walls and corporate monumental structures) and significant chiefly investment in agricultural intensification (archaeologically manifested in large-scale irrigation works, water control, and artificial terracing). Archaeological investigations suggest that the Hawaiian chiefdoms, pre-Inca South American chiefdoms, and many pre-state Mesopotamian complex societies had political economies dominated by staple finance.

Wealth finance involves the use of prestige goods or valuables as political currencies by chiefs and other elites to cement politically strategic alliances with other elites and to reward subordinates for loyalty and service (particularly associated with a *network* political strategy). Prestige goods vary according to socially defined standards of value and can be any rare or easily controlled object (e.g., jade in Lungshan China, gold in Panamanian chiefdoms, bronze in European chiefdoms, shell and copper ornaments in eastern North American chiefdoms). Politically charged wealth can be locally manufactured or obtained through foreign trade. Emerging elites control access to this political currency through support of *attached specialists* (highly skilled artisans who produce these goods wholly for elite consumption), by monopolizing valuable raw materials and technologies, by dominating trade routes, and by defining limited social contexts for their circulation (e.g., elite-controlled ceremonial feasting events). While specific chiefdoms tend to emphasize one means of economic control over another, chiefs tend to use elements of both staple finance and wealth finance in varying combinations, resulting in unique forms of political economy.

In addition to the emergence of *attached specialists*, the rise of chiefdoms is sometimes accompanied by the growth of *independent specialists*, full-time specialists producing more mundane household goods for an unrestricted set of consumers who concentrate at chiefly centers due to economies of scale. While the archaeological recovery of large-scale craft workshops containing mass production equipment (e.g., large kilns or smelting furnaces, molds, potter's wheels) is the most direct line of evidence for specialist production, such finds are rare. More often archaeologists have assessed craft production modes through analysis of the products themselves and their regional distribution. Specialized production by a limited number of concentrated, recurrently interacting, full-time craftspersons (either *attached specialists* or *independent specialists*) generally results in a more standardized product, as measured through object form, raw material, and/or decoration. In addition, specialist-produced prestige goods are largely restricted to elite centers, while household-produced domestic goods are widely dispersed throughout a region, and the socially unrestricted products of independent specialists are concentrated at, but not restricted to, regional centers.

Particularly in chiefdoms relying heavily on wealth finance, political relationships and hierarchies of authority are reinforced through the continual circulation of prestige goods, most often in the context of bride wealth exchanges (e.g., gifts of porcelain in Southeast Asia and gold in Panama), political investitures (e.g., the presentation of feather capes in Polynesia), ritualized feasting, and other politically-charged events. Of these exchange contexts, ritual feasting is particularly evident in the archaeological record. Specialized feasting paraphernalia (e.g., bronze drinking vessels in the European Iron Age, large ceramic cooking vessels in the Mississippian and Formative Mesoamerican chiefdoms), unusual food remains (e.g., large pig concentrations in Lungshan China), and their association with ritual architecture (e.g., ball courts in Mesoamerica) have been used to identify feasts which were likely aimed at political integration and competitive status display.

5. *The Ideology of Rulership*

Ideologies are collective representations of the social and political order in particular societies, often encoded in myths, ceremonies, and various public performances, but also frequently materialized in archaeologically visible monumental architecture, iconography, and portable objects. In both state-level societies and chiefdoms, the creation of a dominant ideology and its imposition on the populace are an important basis for power. In many chiefdoms, elites developed their own language, dialect, lexicon, and/or writing systems to control the flow of esoteric knowledge and to restrict the performance of religious rites. The archaeological contexts in which written texts are found and their decipherable content (e.g., stone stela found with monumental earthen works in Formative Period Mesoamerican centers, wooden tablet inscriptions associated with Easter Island ceremonial complexes) suggest that literacy was an elite prerogative which served to institutionalize cosmological notions and to legitimate the political and social domination of elites.

In the absence of written texts, archaeologists often infer shared cosmic orders and ruling ideologies from iconography on portable objects (e.g., the were-jaguar figurines of Formative Period Mesoamerica, 'sun' motifs and other symbols on Mississippian pottery, 'patron-deity' symbols on Egyptian Gerzean stone palettes and maces), from monumental constructions (e.g., the platform temple complexes of Polynesia, the henge monuments of Late Neolithic Europe), from material symbols in burials (e.g., painted scenes in Japanese Yayoi mounded tombs), and from the spatial organization of chiefly centers (e.g., the layout of Cahokia and other Mississippian centers). An influx of exotic symbols representing foreign cosmologies or religious beliefs (e.g., Buddhist and Hindu art and architecture in Southeast Asia, widespread Olmec

styles in Formative Period Mesoamerica, Egyptian themes in the Nubian chiefly tombs of Kerma) has often been interpreted by archaeologists as a particularly effective strategy for chiefs to add to their monopolistic store of power-enhancing knowledge.

While past archaeological research has largely focused on ideologies of domination in chiefdoms, many researchers now recognize that the many individuals and social factions comprising these complex societies have distinct, and sometimes conflicting, worldviews based on personal experiences and interests. Increasingly, archaeologists are attempting to tease out diverse values and perspectives related to gender, class, ethnicity, occupation, and individuality. For example, a number of recent archaeological studies have focused on gendered views of cultural norms and social orders (e.g., Joyce Marcus's and Kent Flannery's analyses of feminine depictions in human figurines and gender-segregated work spaces in Formative Period Mesoamerica).

6. *Warfare and Militarism*

Anthropological studies have suggested that warfare in chiefdom-level societies differs from that in tribal societies by its expansionistic focus on acquisition of territory, resources, and captives, and by its concentration of military power in the hands of chiefly leaders who use it to expand their political sway. Robert Carneiro has suggested that both the rise and evolution of chiefdoms are related to warfare. Under conditions of land shortages and population stress, militarily powerful leaders systematically assault settlements along polity boundaries. Conquered enemies are incorporated into the expanding chiefdom, resulting in the elaboration of political hierarchies and the coalescence of a polity of greater complexity as well as scale. However, warfare is an almost continuous process in many chiefdoms, one of a number of forms of 'peer polity' interaction that are politically transformative. Militarism can often result in more subtle changes in power relationships among competing chiefdoms, over both the short and long term, without territorial conquest. Political loyalties may be shifted to military victors who control the ideology of warrior prestige and who increase their political currency through captured labor and valuables, resulting in larger alliance networks.

Archaeologists have focused on a variety of material evidence to document changes in the scale, intensity, and behavioral and ideological aspects of warfare in chiefdom-level societies. These include the development of fortifications or other defensive works, changes in regional settlement patterns (e.g., concentration of population in large centers, depopulation along polity boundaries, relocation of settlements to defensible positions), and systematic destruction of power-symboling architecture (e.g., dismantling of monumental architecture, toppling of statuary, burning of chiefly centers). Examples are the alternating rebuilding and burning of fortifications surrounding Iron Age European towns such as Heuneberg, the creation of 'no man's lands' between Mississippian centers, and the defacement of sacred monumental statuary on Easter Island. Osteological evidence from burials allows archaeologists to evaluate rates of violence in populations (e.g., scalping, skeletal trauma, embedded projectiles), the ritualized use of body parts in prewar and postwar ceremonialism (e.g., decapitation for trophy head taking), and debilitating physical conditions and malnutrition related to prolonged exposure to military siege. Archaeological studies of burials also allow archaeologists to identify 'warrior insignia' associated with the development of specialized warrior classes and strong ideologies of warrior prestige exemplified in the elaborate bronze horse-fittings of Bronze Age European warriors and the gold-pegged teeth of island Southeast Asian warriors. The development of new warfare technologies associated with escalating military competition can be traced archaeologically in intensified metals or stone mining, labor-saving equipment (e.g., molds) or large-scale facilities for mass production of weapons, and the adoption of horses for mounted cavalry.

7. *The Evolution of Chiefdoms*

Many archaeologists view the transition to inherited, institutionalized power and status, associated with chiefdoms but further developed in states, as the key transformation in human societies. Since the organizational dynamics of chiefdoms are viewed as similar to that of states, theories of state development are commonly applied at the chiefdom level. Early theories emphasized the managerial benefits of chieftainship as societies expanded in scale. Chiefs arose to administer irrigation works (Karl Wittfogel), coordinate localized production and exchange (Elman Service), mediate land conflicts and warfare (Robert Carneiro), ameliorate economic risk through ritual intervention (Robert Drennan), and to meet other organizational challenges.

More recently, archaeologists have criticized the functionalist nature of these theories and their simplistic emphasis on single factors, instead focusing on the varied acquisitive and power-seeking strategies of political actors as they vie for symbolic capital and control over the labor and resources of others. In this view, the origins of permanent forms of social inequality and political authority are to be found in the competitive and aggrandizing behavior of 'big men' in tribal societies. Archaeologists now examine this transformation at varying scales of analysis, ranging from the transfiguring actions and motivations of individual political actors, to multiple polities develop-

ing together through forms of *peer polity interaction*, to change-stimulating contacts with more complex states and empires (*world systems theory*). In addition, many archaeologists now recognize that not all chiefdom-level societies have unilinear evolutionary trajectories leading inevitably toward greater complexity. Many chiefdoms fail to develop state-level institutions over the long-term, instead perpetually cycling between complex and simple forms, 'devolving' into tribal societies (e.g., New Zealand's South Island Maori, possibly Amazonian societies), mysteriously collapsing (e.g., the Mississippian chiefdoms, Easter Island), or otherwise maintaining a chiefdom-level organization until eventual absorption by colonizing states and empires (e.g., many chiefdoms of the Central American Isthmus, North America, the Caribbean, island Southeast Asia, Europe and Polynesia).

See also: Big Man, Anthropology of; Conflict and War, Archaeology of; Intensification and Specialization, Archaeology of; Political Anthropology; Power in Society; States and Civilizations, Archaeology of

Bibliography

Carneiro R 1981 The chiefdom as precursor of the state. In: Jones G, Krautz R (eds.) *The Transition to Statehood in the New World*. Cambridge University Press, Cambridge, UK, pp. 37–79

Drennan R, Uribe C (eds.) 1991 *Chiefdoms in the Americas*. University Press of America, Lanham, MD

Earle T 1987 Chiefdoms in archaeological and ethnohistorical perspective. *Annual Review of Anthropology* **16**: 279–308

Earle T (ed.) 1991 *Chiefdoms: Power, Economy and Ideology*. Cambridge University Press, Cambridge, UK

Earle T 1997 *How Chiefs Come to Power*. Stanford University Press, Stanford, CA

Kirch P V 1984 *The Evolution of the Polynesian Chiefdoms*. Cambridge University Press, Cambridge, UK

Renfrew C, Shennan S (eds.) 1982 *Ranking, Resource and Exchange*. Cambridge University Press, Cambridge, UK

Yoffee N 1993 Too many chiefs. In: Yoffee N, Sherratt A (eds.) *Archaeological Theory: Who Sets the Agenda?* Cambridge University Press, Cambridge, UK, pp. 60–78

L. L. Junker

Child Abuse

Article 19 of the *Convention on the Rights of the Child*, ratified in 1989, prescribes that parties shall take all appropriate legislative, administrative, social, and educational measures to protect the child from all forms of physical or mental violence, injury or abuse, neglect or negligent treatment, maltreatment or exploitation, including sexual abuse. Until now this article has been disregarded throughout the world. In laying blame we should not only consider third-world countries (with child labor, prostitution, and soldiers) but also the rich industrial countries, where violence against children must be seen as a 'social epidemic.'

1. Forms of Child Abuse

There are four forms of child abuse:

Physical abuse. Blows or other violent actions, resulting in injuries to children, such as being beaten, hit, whipped, pushed down stairs, hurled against a wall, burned with hot water/cigarettes, jammed in doors/car windows, tortured with needles, put into cold water or pushed under water, made to eat their own faeces or drink their own urine, strangling, and poisoning.

Sexual abuse. Any action that is inflicted upon or must be tolerated by a child against their own will or any action about which the child cannot make a decision due to their physical, emotional, mental, or verbal inferiority. The offenders use their position of power and authority to satisfy their own needs at the expense of these children who thus suffer discrimination as sexual objects. Sexual abuse between children themselves is manifest when one child is much older and/or uses force.

Neglect. Considerable impact on or damage to a child's development due to a lack of care, clothing, feeding, medical care, supervision, or protection from danger.

Emotional abuse. Outright rejection, intimidation, terrorization, or isolation of a child. Actions such as verbal abuse/discrimination on a daily basis, locking a child in a dark room, tying them to a bed and many other major threats, including to their lives.

2. Prevalence of Forms of Abuse

This article draws on survey results from German-speaking countries and the United States as they are deemed to be representative of comparable statistics from modern industrial countries.

2.1 Physical Abuse

A representative survey (Pfeiffer and Wetzels 1997) of 16 to 59 year olds in Germany about their childhood experiences of physical parental violence indicates a

Table 1
Incidence of parental violence

Incidence of parental violence *Possible answers: never, rarely, occasionally, often, very often* *Multiple answers were possible, N = 3241*		
Question	Rarely	More often than rarely
1. Throwing objects	7.0%	3.7%
2. Grabbing/pushing around	17.9%	12.1%
3. Slapping	36.0%	36.5%
4. Hitting/striking with an object	7.0%	4.6%
5. Punching, kicking	3.3%	2.6%
6. Striking, beating up	4.5%	3.5%
7. Strangling	1.4%	0.7%
8. Intentional burns	0.5%	0.4%
9. Threats with a weapon	0.6%	0.4%
10. Use of weapon	0.6%	0.3%
Total physical violence (Questions 1–10)	36.1%	38.8%
Total physical violence (Questions 5–10)	5.9%	4.7%

distinction between physical *punishment* and physical *abuse*. Physical punishment is defined as intentionally inflicting pain to control a child's behavior but without intending to cause severe injuries or damage and without violating (existing!) laws. Physical abuse definitely violates laws, as injuries to the child are either intentional or tolerated as a consequence of these violent actions. The incidences of violence observed are listed in Table 1.

Of the persons interviewed, 74.9 percent confirmed having experienced physical parental violence in their childhood, and 10.6 percent definitely suffered physical parental abuse.

2.2 Sexual Abuse

Recent retrospective studies in Germany, Austria, and Switzerland (Deegener 2000) suggest that about 15 to 25 percent of women and 5 to 10 percent of men questioned had been sexually abused in their childhood or adolescence. We believe that the approximate incidence and severity of sexual abuse in both sexes is as shown in Table 2. Almost two thirds of victims were ever only abused on one occasion. One third of cases involved repeated sexual abuse. However in the latter group, about 10 percent of cases continued over periods exceeding 12 months and extending to several years. Such long-term sexual abuse is perpetrated predominately by a victim's relatives. Between 80 and 90 percent of offenders are male.

2.3 Neglect and Emotional Abuse

If defined in conservative terms, the extent of these forms of abuse in Germany far exceeds the incidence of the other two forms. If defined in broader terms, it is not an exaggeration to say that the bullying of children is widespread and commonplace.

2.4 Statistics from the USA

The National Committee to Prevent Child Abuse (1995) indicates that over three million (suspected)

Table 2
Severity of sexual abuse

Serious sexual abuse Attempted or actual vaginal, anal, or oral violation; oral satisfaction by victim or anal penetration of offender	15%
Severe sexual abuse Victim has to masturbate in front of offender; offender masturbates in front of victim; offender fondles victim's genitals; victim has to touch offender's genitals; victim has to expose genitals to offender	35%
Less severe sexual abuse Offender attempts to touch victim's genitals; offender fondles victim's breasts; sexual kisses; French kisses	35%
Sexual abuse without physical contact Exhibitionism; victim has to watch pornographic videos; offender observes victim while bathing	15%

victims were recorded in 1994 in the USA. Almost 1300 of these children died as a result of abuse and neglect. Elders (1999) reports 3,195,000 registered cases of child abuse in 1997, with one million cases being confirmed. However, only 50 to 75 percent of all registered cases were adequately investigated. Over the years the following averages have been determined: 45 to 50 percent are cases of neglect, 25 percent of physical abuse, 10 percent of sexual abuse, 3 percent of emotional abuse, and 15 percent of other causes.

2.5 Combination of Different Forms of Abuse

Multiple episodes of maltreatment during childhood are more the rule than the exception. Pfeiffer and Wetzels (1997) found that even in the youngest age group (i.e., 16 to 29 year olds), approximately one sixth of adolescents and young adults in Germany had been victims of frequent or severe physical parental violence or sexual abuse (involving physical contact and abuse outside the family). Employing a rather conservative estimate, we believe that about one fifth of this young generation has been affected by physical parental violence, sexual victimization, or frequent adult partner violence.

2.6 Assessing the Incidence of Child Abuse

The extent of child abuse invites comparison with a global epidemic. However, in contrast to the degree of professional, political, and public attention received by other diseases, child abuse is neglected. Sadler et al. (1999) for example, estimate that the incidence of abuse is ten times higher than that of all forms of cancer. The 'costs' of child abuse are tremendously high, with the 'human costs' of the sociopsychological, cognitive, and physical consequences of abuse being inestimable. Courtney (1999) quotes a 1993 estimate for the USA, indicating that approximately 11.4 billion US$ was spent on examining abused children, providing medical care to severely injured children, treating victims and their families, and caring for children in foster families. In addition, Courtney cites Westman, who calculates that expenditure due to 'incompetent parenting' (in broad terms) in the USA amounts to about 38.6 billion US$ each year.

3. Causes of Child Abuse

More specific causal models are becoming accepted in research and clinical practice (Cicchetti 1989), resulting in a marked improvement in early recognition, therapy and prevention. Four groups of influences can be identified which interact within a complex reference framework:

(a) At the individual level there is the perpetrator's life history and personality, e.g., their own experiences of childhood abuse, early separation of parents, periods in (foster) homes, emotional disturbances, or alcohol and drug abuse.

(b) At the family level there are maladjusted parent–child relationships, partner conflicts and problems, etc.

(c) At the social level there are factors involving poverty, unemployment, meager and limited housing, social ghettos, and insufficient social support, etc.

(d) At the community level there are, among other things, a high tolerance of violence and aggression and a high degree of violence in their upbringing.

There are numerous interactions between these levels that may lead to (acute or protracted) familial destabilization, resulting in neglect or physical abuse. In addition to the risk factors, which may increase the probability of child abuse, one must consider the preventative factors, which reduce the risk and consequences of abuse.

4. Consequences of Child Abuse

Not only can child abuse influence every aspect of behavior and experience, but it can also lead to psychosomatic diseases and physical injuries (Egle et al. 2000). In addition, child abuse influences the attachment of children to their parents and their relationships with their peers. However, as the short- and long-term consequences are not specific there can be various other causes.

Whereas some research suggests that the consequences of child abuse can last throughout a person's life, other research shows that abused children suffer few (if any) adverse consequences. During the 1990s, a growing number of investigators has identified many mediating and moderating factors that can either ameliorate or exacerbate the consequences of child abuse (Masten et al. 1990, Kendall-Tackett et al. 1993, Kaplan et al. 1999):

characteristics of the child's experiences (e.g., nature, frequency, severity, type, and prior history of child abuse);

resources of the child (e.g., good social and academic competencies, endearing temperament);

vulnerability of the child (e.g., psychiatric disorder, low intelligence, insecure attachment style, early onset of the abuse);

social support of the child (e.g., good relationship with the nonabusing parent, support of significant others from outside the family, psychotherapy).

Therefore, the extent of the negative consequences of child abuse depends on complex transactional processes between vulnerability or risk factors on the one hand, and resiliency or compensatory factors on the other hand. The term 'resiliency' should not be

understood to minimize the suffering of abused children or justify criticism of those children who are not resilient. However, some children from violent homes are less influenced by their abuse than others and develop effective coping skills and strategies. In this connection, the relationship between the causes and consequences of child abuse must also be considered: a low intelligence may—for example—stimulate abusive behavior by parents or caretakers, but low intelligence can also be a negative consequence of severe abusive experiences in early childhood.

Generalizing, the severity of the consequences of child abuse depends especially on the following factors: the age and developmental stage of the child, the length and severity of the abuse, and the relationship of the abuser to the victim. Therefore, long-term severe abuse of a young child perpetrated by a parent tends to produce more detrimental effects on the emotional, physical, social, and cognitive development than shorter-term abuse of an older child by a stranger. In the case of long-term physical abuse it is not uncommon to find a range of injuries inflicted at different times. The consequences of neglect remain underestimated (i.e., 'neglected neglect').

5. *Prevention*

The extent to which abusive educational measures (including physical punishment) are tolerated varies widely from country to country as a function of the sociocultural context. Except for legal prosecution in confirmed cases of child abuse, the rights of parents (or teachers, etc.) to employ a wide range of seemingly appropriate educational measures (including physical punishment) remain unchallenged.

In 1979 Sweden introduced a law banning physical punishment, addressing this grey zone between 'permitted' and 'illegal' educational measures. Countries such as Finland (1984), Denmark (1986), Norway (1987), and Austria (1989) followed suit. In Germany the following law was passed in 2000: 'Children have the right to a non-violent upbringing and education. Physical punishment, emotional abuse and other discriminatory measures are unacceptable.' This does not imply an increase in prosecutions. As Sweden has demonstrated successfully, the objective is rather to alter the value system and influence the 'legal hygiene,' while simultaneously strengthening the support for a nonviolent upbringing and education. An education conducted in accordance with the above-mentioned UN conventions and with the child-protection provisions found in many other constitutions would not tolerate physical punishment or other discriminatory educational measures.

Adequate prevention requires worldwide birth control, a reduction in poverty, an improvement in both socio-ecological and living conditions, the improvement of general education standards and lifelong learning (including parent training), and a reduction in the wide power and relationship gap existing between men and women, resulting from raising boys to be 'masters' and girls to be 'servants.'About one third of all cases of sexual abuse/violation are committed by male adolescents.

Furthermore, there should be independent, state-nominated representatives for minors, whose role is to replace parents at the community level and to fight for children's needs and rights in society. The young should also have more opportunities to co-determine what occurs by means of an institution such as a children's or young people's parliament. Special broad-scale support programs for families/parents at risk (such as home visits to teenage mothers or high-risk women during pregnancy) should be offered, and more assistance should be available to families in serious difficulty.

The predominant objectives of prevention need to be: (a) to interrupt the generational cycle of family violence, (b) to teach future generations humanitarian attitudes and civil behavior, and (c) to promote 'social parenting' and 'structural security' instead of 'structural violence' for all children.

There is widespread condemnation when child abuse or violent acts committed by (abused) adolescents become public, but this still tends to be little more than lip service. We tend to avoid reflecting upon our own violent relationships and actions, obscuring these by employing projections of monstrous child abusers. Appropriate responsibility for children and their futures is not being implemented for one simple reason: 'We are all against such things as child pornography, but only a few are willing to actually support programs that could save children's lives, because these cost money and comfort and require another form of living and life' (Sigusch 1996).

See also: Child Care and Child Development; Childhood Depression; Childhood Sexual Abuse and Risk for Adult Psychopathology; Children and the Law; Children, Rights of: Cultural Concerns; Family Health; Socialization in Infancy and Childhood; Violence and Effects on Children; Violence as a Problem of Health

Bibliography

Cicchetti D 1989 How research on maltreatment has informed the study of child development: perspectives from developmental psychopathology. In: Cicchetti D, Carlson V (eds.) *Child Maltreatment. Theory and Research on the Causes and Consequences of Child Abuse and Neglect*. Cambridge University Press, New York

Courtney M E 1999 The economics. *Child Abuse and Neglect* **23**: 975–86

Deegener G 2000 *Die Würde des Kindes. Plädoyer für eine Erziehung ohne Gewalt* [The dignity of the child. A plea for an upbringing without violence]. Beltz, Weinheim, Germany

Egle U T, Hoffmann S O, Joraschky P 2000 *Sexueller Mißbrauch, Mißhandlung, Vernachlässigung. Erkennung und Therapie psychischer und psychosomatischer Folgen früher Traumatisierungen* [Sexual abuse, maltreatment and neglect. Recognition and therapy of psychological and psychosomatic consequences of traumatizations at an early age]. Schattauer, Stuttgart, Germany

Elders M J 1999 The call to action. *Child Abuse and Neglect* **23**: 1003–9

Kaplan S J, Pelcovitz D, Labruna V 1999 Child and adolescent abuse and neglect research: A review of the past 10 years. Part I: Physical and emotional abuse and neglect. *Journal of the American Academy of Child and Adolescent Psychiatry* **38**: 1214–22

Kendall-Tackett K A, Meyer Williams L, Finkelhor D 1993 Impact of sexual abuse on children: A review and synthesis of recent empirical studies. *Psychological Bulletin* **113**: 164–80

Masten A S, Best K M, Garmezy N 1990 Resilience and development: Contributions from the study of children who overcome adversity. *Development and Psychopathology* **2**: 425–44

National Committee to Prevent Child Abuse (NCPCA) 1995 *Current Trends in Child Abuse Reporting and Fatalities. The Results of the 1994 Annual Fifty State Survey*. NCPCA, Chicago

Pfeiffer B, Wetzels P 1997 *Kinder als Täter und Opfer. Eine Analyse auf der Basis der PKS und einer repräsentativen Opferbefragung*. [Children as offenders and victims. An analysis of criminal statistics and victim interviews]. Forschungsbericht 68. Kriminologisches Forschungsinstitut, Hanover, Germany

Sadler, B L, Chadwick, D L, Hensler D J 1999 The summary chapter—The national call to action: Moving ahead. *Child Abuse and Neglect* **23**: 1011–18

Sigusch V 1996 Kultureller Wandel der Sexualität [Sexuality in cultural transition]. In: Sigusch V (ed.) *Sexuelle Störungen und ihre Behandlung* [Sexual disorders and their treatment]. Thieme, Stuttgart

G. Deegener

Child and Adolescent Psychiatry, Principles of

Child and adolescent psychiatry and psychotherapy comprises the diagnosis, treatment, prevention, and rehabilitation of neuropsychiatric and developmental disorders as well as behavior disturbances during childhood and adolescence. The need for a separated psychiatric discipline for children and adolescents results from age-dependent characteristics of mental disorders, strongly influenced by rapidly alternating stages of neurobiological and social development in this period of life. The discipline of child and adolescent psychiatry is now acknowledged as a medical speciality or subspecialty in many countries; however it will still need much effort to offer a specialized child mental health service worldwide.

This article provides an overview of the historical development of child psychiatry in different cultural regions, focusing on the development in Europe and the USA. A necessary distinction is made regarding the diagnostical classification system for mental disorders in childhood and adolescence in comparison to the classification system in general psychiatry. Principles of child-specific assessment and treatment are described. Further, some future perspectives in the development of a more biologically influenced child psychiatry are discussed as well as the needs for a modern child psychiatry in developing countries.

1. *History*

The clinical discipline of psychiatry was formed in the nineteenth century, in a situation of an increasing interest in and knowledge of psychological phenomena. In 1899, the term 'child psychiatry' was first used by the French psychiatrist M. Manheimer, who called his book *Les troubles mentaux de l' enfance*, subtitled *Précis de psychiatrie infantile*.

Four main traditions have made substantial contributions to the current body of knowledge in the field of child and adolescent psychiatry, and influenced the structure and the actual treatment concepts of child psychiatric institutions. The formerly unified disciplines of psychiatry and neurology have given rise to the tradition of neuropsychiatry, especially in some European countries. Several scientific associations still include reference to neurology. Increasing research activity in the areas of neuropsychobiology and neuropsychology confirms the need of a close linkage between these two 'brain disciplines' for a better understanding of psychiatric disorders.

A movement based on a remedial clinical tradition, promoted by Hans Asperger in Austria and Paul Moor in Switzerland, still plays a role in pediatric departments with activities in the field of child psychosomatics. In general, remedial education is an important part of the multidisciplinary children's mental healthcare service.

Developed by the pioneers of psychoanalytical work with children, like Anna Freud, Alfred Adler and Melanie Klein, the psychodynamic-psychoanalytic tradition influences etiological concepts of behavioral and personality disorders and gives implications on psychotherapeutic treatment strategies; however, behavioral therapy, often in combination with family therapy interventions, is predominant in the clinical use of psychotherapy nowadays.

The empirical, epidemiological, and statistical tradition has been established in a number of European countries, mainly the UK, Scandinavian countries, and Germany. It has been strongly influenced by the work of Michael Rutter and by research influences

from the USA. This tradition created the basis of the currently used classification systems in psychiatry.

Although first considerable activities in the field of child psychiatry started in Europe in the early twentieth century, it was not until 1954 that the first European symposium of child psychiatry took place in Magglingen (Switzerland), a consequence of World War II and the ensuing political situation. At this meeting, first attempts were made to establish a unifying scientific association, which was founded in 1960 as the Union of European Pedopsychiatrists at the first European congress in Paris. Later, at a congress in Madrid in 1979, the name of the society was changed to the European Society for Child and Adolescent Psychiatry (ESCAP) (Remschmidt and van Engeland 1999).

The crystallization of child psychiatry in the USA was influenced by the growing knowledge in the field of psychiatry in Europe. A sociocultural reform, beginning in the year 1900, was the main factor contributing to the establishment of child psychiatry in the USA. The desire to protect children from social hardship resulted in the need for mental healthcare institutions. William Healy founded the first Institute of Juvenile Research in 1909 to handle child problems occurring in association with delinquency and traumatic injuries. The Orthopsychiatric Association, founded in 1924, integrated the different medical and sociopsychological professionals working in the psychiatric field. The psychiatrist Adolf Meyer installed the first department of child psychiatry at the Johns Hopkins University in the early 1930s. Leo Kanner, an immigrant from Austria, was the first chairman and published his textbook of child psychiatry in 1935. Child psychiatry was accepted as a clinical discipline in its own right in 1950. Since then, the American Academy of Child and Adolescent Psychiatry (AACAP) has made a marked contribution to the global development of a modern child and adolescent psychiatry (Schwab and Schwab-Stone 1999).

The different professionals working in the field of child psychiatry have been organized in the International Association of Child and Adolescent Psychiatry and Allied Professions (IACAPAP), founded as an umbrella organization in the 1930s. Countries from East Asia and the Pacific region present a remarkable level of medical supply and research in the field of child and adolescent psychiatry. In recent years South America, African countries, and Australia have also contributed with increasing activities to the specialization of child mental health services.

2. *Epidemiology and Classification*

The prevalence rates of all mental disturbances in children are estimated at about 15 percent. However, mental disorders in need of treatment occur in about half of these cases. In children with primary neurological brain dysfunctions, the prevalence of psychiatric disorders increases with the degree of the brain damage or the metabolic alteration. So, children with epilepsy suffer five times more from mental problems than unaffected children.

Psychiatric classification systems are compilations of diagnostic criteria, based on clinical experience, to define and differentiate mental disorders and establish agreement and common language among healthcare professionals and researchers. The classification of mental disorders is essential for the development of treatment concepts and a necessary precondition for epidemiological, clinical, and biological research in this field.

The 10th revision of the *International Statistical Classification of Diseases and Related Health Problems (ICD-10)*, published by the World Health Organisation (WHO) in 1992, is a comprehensive classification system of medical conditions and mental disorders, used in official medical and psychiatric nosology throughout most of the world. However, some countries (e.g., France and the USA) use compatible or modified classifications. The fourth edition of *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*), published in 1994 by the American Psychiatric Association (APA), is the official psychiatric coding system used in the USA and is in part compatible with the ninth revision of the *International Classification of Diseases (ICD-9)*.

The mental development of children and adolescents, in contrast to that of adults, is strongly influenced by brain maturation processes and social environment. Considering this fact, the classification system in child and adolescence psychiatry is based on a multiaxial diagnostical view. The approach of the actual classification systems is atheoretical according to etiological concepts. Due to the limited knowledge about etiology and pathomechanism of the mental disorders, it is mainly the symptoms and course of a disorder that are used to define and classify the mental disorders. An approach to predict the clinical course of the disorder is associated with the description of a clinical feature. Additional contributing factors, like the intelligence level, the parent–child relationship, and the social environment are assembled by using different diagnostical axes. The impact of the disorder on psychosocial functions is assessed on Axis VI of the multiaxial diagnostic system (Remschmidt and Schmidt 1994).

ICD-10, regardless of the use of a multiaxial system, does not consider age of disease onset consistently. Apart from that, the predicted clinical course is basically used for many clinical decisions including treatment concepts (Rutter 1989). Taking into account these important facts, an appropriate version of the classification system for child and adolescent psychiatry has been developed, where age of onset and typical clinical course has been integrated. The disorders have been divided into early-onset disorders with a persistent course (e.g., autism), early-onset

disorders with transient manifestation (e.g., enuresis), age-related interactional disorders (e.g., separation anxiety), young age-related disorders with sometimes recurrent episodes or chronical course (e.g., eating disorders), early-onset adult-type disorders (e.g. schizophrenia).

Even though specified diagnostic criteria are provided for each mental disorder, increasing the reliability of clinicians process of diagnosis, the professionals have to pay attention to existing mixed forms or comorbid disorders (Caron and Rutter 1991), like attention deficit/hyperactivity disorder (AD/HD) with conduct disorder or eating disorder with depressive disorder (Biederman et al. 1991).

3. *Developmental Psychopathology and Pathogenesis*

Mental disorders in children and adolescents are manifested in behavioral problems, affective and cognitive disturbances, and somatic concerns. The quality and severity of these problems are influenced by the developmental state. The rapid development, simultaneously on the biological and on the social level, has a strong impact on the diverse developmental tasks of a minor (Munir and Beardslee 1999).

Biological maturation, particularly in early childhood, plays a key role in the etiology of mental problems. With increasing age social adaptation and social learning becomes more prominent. The maturation of the brain and gender-specific role taking are two main factors of becoming adult. Acceleration or retardation of the biological and social maturation results in an alteration of the developmental process with behavioral problems and social disintegration in the case of existing permanent deficits.

Developmental psychopathology describes age-related aspects of mental disorders, especially the formation and alternating pattern of symptoms over time (Cicchetti and Cohen 1995). In early childhood the ability of autonomous behavior is acquired. For that, an inner representation and judgment of the child's cognitive and behavioral activities is an important precondition. Older children compare their self-image with an imagined ideal as a kind of self-control, which modifies and limits the outcome of the social learning process. Although there is an interaction between these different components, mental disorders in childhood and adolescence can be interpreted as deficits in maturation or learning processes or in their autonomous efforts. Parallel to the differentiation of cognitive, emotional, and social functions there are increasing demands on the child. Sexual development, gender-related role taking, and academic achievement all which have to be coped with adequately.

During certain periods of development, the child and adolescent have specific responsibilities. Functions according to these specific developmental demands can be easily disrupted, because they are not well established due to the rapid developmental process. Infants suffer from regulation disorders, which are often induced by the behavior of caregivers. Preschool-age children suffer from somatic, physiological dysfunction in the area of sleep, eating, speech, and elimination control, whereas in school-age children communication problems in peer groups and in the school situation appear (Esser et al. 1990). Separation anxiety, insufficient affect control, and learning disabilities are characteristic problems with increasing age. Adolescents then have to develop self-confidence, to deal with authorities and to accept their sexual identity.

Risk and protective factors, in addition to the developmental process, are effective in the pathogenesis of mental disorders. Genetically determined factors are the somatic constitution, particularly the brain function, and the pattern of personality, the temperamental factors (Rothenberger 1990).

Pervasive developmental disorders like autism have a high genetic risk (Rutter et al. 1999). Traumatic brain injuries prenatal, perinatal, or postnatal, or chronic metabolic dysfunction increase the risk of developing a mental problem. A lack of sociocultural stimulation, socioemotional deprivation, or high pressure as a result of people's inadequate expectations could be summarized as sociogenic risk factors (Rutter 1999). Language acquisition in particular strongly depends on a sociogenic balance. Intrapsychic conflicts are a result of chronic problems, like sexual or physical abuse, mental disease of a parent, delinquency in the family, distorted familiar communication. The interactive process between parents and child can amplify or otherwise reduce behavior in both directions, to a normalization or to an increased disturbance. Life events, as acute stress factors, in general play a minor role in the pathogenesis of mental disorders of children and adolescents. In the case of strong traumatic impairment, like being a victim of a rape or having a bad accident, a posttraumatic stress disorder can appear (Costello et al. 1998).

Protective factors, like social attractiveness, verbal skill, appropriate problem-solving strategies, self-consciousness, creative intelligence and interests, a pleasant life and family situation, and stable peergroup relationships can act successfully against other biological or psychosocial stress factors to reduce the risk of developing a mental disorder or in the case of manifestation to reduce symptom severity and improve clinical outcome (Laucht et al. 1997).

4. *Assessment*

To provide a diagnosis, child and adolescent psychiatrists examine the mental status, the psychopathological profile, and behavior of the patient;

ascertain the social and biological development, and evaluate the prior medical and psychiatric history. Further useful information can be obtained by parental interviews and school reports. Different sources of information reflect different point of views, experiences, and insights (Jensen et al. 1999). Depending on the type of disorder and the age of the child, the reliability of this information is quite different as well. The summary of information from different sources, including the child, the parents, teachers, peer group members, doctor, pediatrician, and sometimes the youth welfare department, provides the most reliable and complete picture of symptoms, functional level of the child, and influencing environmental factors.

Recognizing the level of social functioning and the performance of developmental tasks related to the child's age is a key assessment for the estimation of severity and prognosis of the disorder. Structured psychiatric interviews, rating scales, and self-report forms can be used to observe, quantify, and categorize the evaluated symptoms, for example, the Diagnostic Interview Schedule for Children-Revised (DISC-R), the Child Behavior Checklist, scales for the assessment of psychotic symptoms (Positive and Negative Symptom Rating Scale—PANS) or mood disorders (Beck Depression Inventory).

Cognitive and intellectual abilities, basic factors of a healthy mind, are age-related functions. To ascertain alterations of these functions, it is necessary to determine the social level and the state of the biological maturation. Particularly in children with mental retardation, learning disabilities, or pervasive developmental disorders a comprehensive assessment of the intellectual function (examples are Kaufman-ABC and WISC-III) is necessary to create an optimized educational and treatment program (Ollendick and Hersen 1993).

In a subgroup of patients, a defined somatic dysfunction, mainly neurological diseases, is the primary cause of mental diseases. Brain diseases (tumor, epilepsy), metabolic dysfunctions (phenylketonuria, hypothyroidism) and diseases with defined chromosomal aberrations (like Down, Angelman, and Prader-Willi syndrome) can be associated with a varied picture of mental disturbances. Therefore physical examinations and technical investigations (brain computerized scan, radiology, electroencephalography, and blood and genetic screening) are standard procedures to rule out somatic disorders and to determine the developmental state of a child. The analysis of genetic mutations, polymorphisms, and associated molecular-biological dysfunctions will provide more insight into the etiology of a psychiatric disorder which is always manifested on a neurobiological level. The early assessment of a genetically caused metabolic dysfunction (like phenylketonuria) can help to prevent the emergence of a disorder or will help, in some cases of known genetic defects (like Prader-Willi syndrome) or relevant polymorphisms associated with a disorder (like in AD/HD), to develop better treatment strategies in the near future.

Assessment of the neuropsychological functions should be done routinely in children and adolescents with AD/HD (Barkley 1998). The core symptoms—inattention, hyperactivity, and impulsivity—need to be quantified by neuropsychological tests, in part with support of computerized methods (e.g., Continuous Performance Task, CPT). Learning disabilities, often associated with developmental or attention deficit disorders, need further examination by questionnaires and psychological tests. To rule out sensory dysfunction, visual and acoustic perception has to be assessed in cooperation with departments of ophthalmology, phonetics, and pediatric audiology.

If children are younger than four years, the psychopathological assessment is oriented on findings regarding the child's development and interactional behavior. This information can be received mainly from parental reports or, in the case of some early onset diseases, by the medical records. Semi-standardized play situations and the observation of peer group and parent–child interactions provide further information about the child's intellectual function, capacity for emotional reactions, attachment, and behavior.

5. *Treatment*

Psychiatric therapy in general is based on three major treatment pillars: psychotherapy, psychopharmacology, and psychosocial and family interventions. Each kind of mental disorder or disease needs a typical combination of these three therapy forms to get an optimal therapy effect. In children and adolescents the family and school environment particularly has to be taken into consideration.

For a sophisticated mental healthcare network, the cooperation of the various professional groups, including psychiatrists, psychologists, social workers, teachers, and members of administrative institutions of the government, responsible for children and adolescents, is a necessary precondition (Simeon 1990).

The treatment of mental disorders can be divided into three main groups: treatment mainly by psychotherapy, treatment mainly by a psychopharmacological intervention, and combined psychotherapeutical and pharmacological approaches. Psychosocial interventions, specially in childhood and adolescence, are in general a part of the treatment program.

A follow-up assessment over a longer time period provides information about the course of the disease. It is necessary to react early to changes in the symptomatology, to observe the social functioning of the child and adolescent, to control possible side

effects of the treatment, and not to overlook the best moment to stop the treatment.

There is a group of early-onset disorders with a persistent course like mental retardation, pervasive developmental disorders, and transient developmentally dependent disorders like learning disorders, communication disorders, selective mutism, attachment and elimination disorders (encopresis, enuresis). Treatment plans are based on behavioral therapy concepts. Contingency management, self-monitoring by protocols, training of social skills and communication, and principles of classical and operant conditioning are expanded by family therapy, educational programs and cooperation with teachers, social interventions, psychomotoric and coordinative training, and in many cases additional parental guidance (Noshpitz 1981).

At the present time, there is no disorder-specific medication on the market; however, there are some substances which are successful in the reduction of single symptoms belonging to these disorders (Houts et al. 1994), like antipsychotics in autistic children (McDougle et al. 1997). An efficient medication exists only for single early-onset disorders. In the case of AD/HD, the core symptoms can be reduced about 70–80 percent by medication, mainly by the use of psychostimulants. However, children with AD/HD need an intensive behavioral and educational treatment program as well, particularly, if they suffer from an associated oppositional defiant or conduct disorder. Often their parents need additional support from parental guidance (Barkley 1998, Kazdin 1997).

The main emphasis in the treatment of mood and schizophrenic disorders, summarized as early-onset adult–type disorders, which get manifested mainly in adolescence and early adulthood, is the psychopharmacological therapy. The exception is that younger children with mood disorders often gain a clinical improvement just from psychotherapy. Antidepressive and antipsychotic medication acts quite specifically on the disturbed brain metabolism, which is, as a primary mechanism or a biological reaction on permanent psychosocial stress, the main causal factor for the disorder. Neurotransmitters like serotonin, noradrenalin, and dopamine are regulated by these substances and neuroplastical processes, which are important for the structural function of brain, are regenerated and stabilized over a longer time. Therefore the intake of most of these medications should be continued over a longer time, if there is no contraindication by serious side effects, which might occur in some cases. Although medication is necessary and the first choice of treatment, these patients need an intensive psychosocial support and training program to prevent them from social disintegration and to cover their academic achievement.

Another large group of mental diseases can be described as behavioral, emotional, age-specific interaction disorders and personality disorders. Examples are eating, anxiety and obsessive-compulsive disorders, stress reactions, and emotionally unstable personalities, called borderline personality disorder. Various behavioral therapy methods in combination with family therapy and psychosocial interventions, continued in an outpatient therapy setting, are established as the most successful kind of treatment. In a few cases, the additional use of medication can improve or accelerate the treatment success (Piccinelli et al. 1995).

Controlled clinical trials are the basis for the development of effective treatment approaches (Vitiello and Jensen 1997). Performing clinical trials in children and adolescents, either to investigate the efficacy and safety of psychotherapeutic treatment strategies or of newer psychopharmacological substances, is difficult from the legal and ethical point of view. There is the need to get a written consent from the proband or patient before starting any investigation. The proband has to be informed about targets, efficacy, and expected side effects of the planed therapy. In the case of children, especially those suffering from mental problems, an adapted, comprehensible form of information needs to be presented, and the parental authorities have to give their written consent as well. Although double-blind placebo-controlled clinical trials are the highest standard to study the efficacy and safety of newer treatment concepts, it is quite difficult from ethical point of view to carry out such a design in children.

Further aggravating circumstances hinder a fast performance of clinical trials in children and adolescents: First, the brain of a minor is not completely maturated. Brain dysfunction and resulting mental disturbances may be transient, like separation anxiety and sleep problems in early childhood. The child psychiatrist has to be aware of this fact, especially regarding the duration of medical and psychotherapeutic treatment interventions. Further, risks of pharmacological side effects and unintentional reactions to psychotherapeutic interventions are higher in children and adolescents than in adults as a consequence of the unstable somatic and mental system. Secondly, some mental disorders in children and adolescents may be triggered by an altered familiar environment with communication problems, may be with violence or sexual abuse. In these cases the treatment of the family system and not exclusively the individual therapy is the primary aim. Unintended reactions, like separation of the parents, should be mentioned as a kind of side effect.

6. *Future Directions*

In developing countries the scarcity of trained personnel has prevented a higher specialization of child mental health services. For professionals in developing

countries the term 'child mental health' therefore covers a broad range of problems, including neurological and developmental disorders, mental retardation, educational difficulties, and psychiatric disorders. Worldwide prevalence rates, an estimation of the persons with a disorder in relation to the whole population, of about 15 percent for children with mental problems, emphasize that developing countries need support in the training and further specialization of their professionals who are responsible for child mental healthcare. Cooperative research projects, under the patronage of the WHO and of the local psychiatric associations, may be one strategy for analyzing the quantity and the quality of child mental health problems in developing countries. It is necessary to recognize characteristic patterns of mental health problems, to describe risk factors (like malnutrition, war and displacement, political oppression, poverty, child labor, urbanization, and social changes) and to develop treatment concepts in consideration of the specific sociocultural background of the community (Rahman et al. 2000).

Although the establishment of specialized mental health services in developing countries is a primary goal, improvement of the child and adolescent psychiatry in developed countries remains a permanent task. Growing insight in molecular biology increases knowledge about brain function and improves the understanding of the mind–body relation (Deutsch 1990). Genetic factors will be described, which have responsibility for the manifestation of mental diseases. Biotechnology will offer new therapeutic strategies, which have to be examined in clinical trials by child and adolescent psychiatrists.

Going one step back from the future perspectives to present problems, child and adolescent psychiatrists need to perform more controlled clinical trials assessing the efficacy and security of psychotherapeutic and psychopharmacological interventions and both in combination. It is necessary to know more about long-term courses and outcome of mental diseases in children to young adulthood and to control the efficacy of pharmacological and psychotherapeutic treatment over a longer time range. A comprehensive neuroscience research especially will increase the knowledge about disorders with an early onset and a persistent pattern of symptoms and about specific developmental-related disorders.

Some mental disorders are quite rare in children and adolescents, like manic affective disorder, so multicenter studies in many hospitals have to be designed to investigate the efficacy of therapies in a sufficient number of child and adolescent patients. Other early onset problems in children, like the different kinds of learning disabilities, need to be investigated more exactly for the development of treatment concepts with higher efficacy. Especially in developing countries with a weak school network, knowledge about learning disabilities, about efficient treatment approaches and perhaps about their prevention, might be very helpful (Schmidt and Remschmidt 1989).

At the least the vision of child psychiatry is the prevention of diseases (Greenfield and Shore 1995). Universal prevention, comparable to vaccination strategies, is hardly reachable in psychiatry. Even selective prevention by early assessment, provisional diagnoses and early therapeutic and psychosocial interventions in identified risk groups is difficult and comprises the problem of false positive assignments with the risk of stigmatizing people. Before acute clinical symptoms occur, there is mainly the registration of unspecific problems, giving some evidence of the later manifestation of a specific mental disorder. Defined criteria of premorbid symptomatology associated with prevention concepts have not yet been established. Probably the progress in biotechnology by detecting disease-specific genetic patterns, will improve the chance of a selective prevention. A more realistic concept is that of an indicated prevention in the earliest stage of a manifested disorder with minor clinical symptoms (McGuire and Earls 1991).

See also: Adolescence, Psychiatry of; Adolescence, Sociology of; Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Anxiety Disorder in Children; Behavior Therapy with Children; Childhood and Adolescence: Developmental Assets; Childhood Depression; Childhood Health; Infant and Child Development, Theories of; Mental Health Programs: Children and Adolescents; Substance Abuse in Adolescents, Prevention of

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*, 4th edn. American Psychiatric Association, Washington, DC

Barkley R A 1998 *Attention Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment*, 2nd edn. Guilford Press, New York

Biederman J, Newcorn J, Sprich S 1991 Comorbidity of attention deficit hyperactivity disorder with conduct, depressive, anxiety, and other disorders. *American Journal of Psychiatry* **148**: 564–77

Brownell K D, Fairburn C G 1995 *Eating Disorders and Obesity. A Comprehensive Handbook*. Guilford Press, New York

Caron C, Rutter M 1991 Comorbidity in child psychopathology: Concepts, issues and research strategies. *Journal of Child Psychology and Psychiatry* **32**: 1063–80

Cicchetti D, Cohen D J (eds.) 1995 *Developmental Psychopathology*. Wiley, New York, Vol. 2

Costello E J, Angold A, March J, Fairbank J 1998 Life events and post traumatic stress: The development of a new measure for children and adolescents. *Psychological Medicine* **28**: 1275–88

Deutsch S I, Weizman A, Weizman R (eds.) 1990 *Application of Basic Neuroscience to Child Psychiatry*. Plenum Medical, New York

Esser G, Schmidt M H, Woerner W 1990 Epidemiology and course of psychiatric disorders in school-age children—results of a longitudinal study. *Journal of Child Psychology and Psychiatry* **31**: 243–63
Greenfield S F, Shore M F 1995 Prevention of psychiatric disorders. *Harvard Review of Psychiatry* **3**: 115–29
Houts A C, Berman J S, Abramson H 1994 Effectiveness of psychosocial and pharmacological treatments for nocturnal enuresis. *Journal of Consulting and Clinical Psychology* **62**: 737–45
Jensen P S, Rubio-Stipec M, Canino G, Bird H R, Dulcan M K, Schwab-Stone M E, Lahey B B 1999 Parent and child contributions to diagnosis of mental disorders: Are both informations always necessary? *Journal of the American Academy of Child and Adolescent Psychiatry* **38**: 1569–79
Kaplan H I, Sadock B J 1998 *Kaplan and Sadock's Synopsis of Psychiatry: Behavioral Sciences/Clinical Psychiatry*. Williams & Wilkins, Baltimore, MD
Kazdin A E 1997 Practitioner review: Psychosocial treatments for conduct disorder in children. *Journal of Child Psychology and Psychiatry* **38**: 161–78
Laucht M, Esser G, Schmidt M H 1997 Developmental outcome of infants born with biological and psychosocial risks. *Journal of Child Psychology and Psychiatry* **38**: 843–53
Lewis M 1996 *Child and Adolescent Psychiatry*. Williams & Wilkins, Baltimore, MD
McDougle C J, Holmes J P, Bronson M R, Anderson G M, Volkmar F R, Price L H, Cohen D J 1997 Risperidone treatment of children and adolescents with pervasive developmental disorders: A prospective, open label study. *Journal of the American Academy of Child and Adolescent Psychiatry* **36**: 685–93
McGuire J, Earls F 1991 Prevention of psychiatric disorders in early childhood. *Journal of Child Psychology and Psychiatry* **32**: 129–54
Munir K M, Beardslee W R 1999 Developmental psychiatry: Is there any other kind? *Harvard Review of Psychiatry* **6**: 250–62
Nathan P E, Langenbucher J W 1999 Psychopathology: Description and classification. *Annual Review of Psychology* **50**: 79–107
Nissen G, Fritze J, Trott G E 1998 *Psychopharmaka im Kindes- und Jugendalter*. Gustav Fischer Verlag, Ulm, Germany
Noshpitz J D 1981 Psychotherapy with children: Basic principles. *Current Psychiatric Therapies* **20**: 47–59
Ollendick T H, Hersen M 1993 *Handbook of Child and Adolescent Assessment*. Allyn & Bacon, Boston
Piccinelli M, Pini S, Bellantuono C, Wilkinson G 1995 Efficacy of drug treatment in obsessive-compulsive disorder. A meta-analytic review. *British Journal of Psychiatry* **166**: 424–43
Rahman A, Mubbashar M, Harrington R, Gater R 2000 Annotation: Developing child mental health services in developing countries. *Journal of Child Psychology and Psychiatry* **41**: 539–46
Remschmidt H, Schmidt M H 1985 *Kinder- und Jugendpsychiatrie in Klinik und Praxis*. Thieme, Stuttgart, Germany
Remschmidt H, Schmidt M H 1994 *Multiaxiales Klassifikationsschema für psychische Störungen des Kindes- und Jugendalters nach ICD-10 der WHO*. Huber, Bern, Switzerland
Remschmidt H, van Engeland H (eds.) 1999 *Child and Adolescent Psychiatry in Europe*. Steinkopff, Darmstadt, Germany
Reynolds W M, Johnston H F 1994 *Handbook of Depression in Children and Adolescents*. Plenum Press, New York
Rothenberger A (ed.) 1990 *Brain and Behavior in Child Psychiatry*. Springer, Berlin
Rutter M 1989 Pathways from childhood to adult life. *Journal of Child Psychology and Psychiatry* **30**: 23–51
Rutter M 1999 Psychosocial adversity and child psychopathology. *British Journal of Psychiatry* **174**: 480–93
Rutter M, Silberg J, O'Connor T, Simonoff E 1999 Genetics and child psychiatry: II Empirical research findings. *Journal of Child Psychology and Psychiatry* **40**: 19–55
Rutter M, Smith D 1995 *Psychosocial Disorders in Young People: Time Trends and Their Causes*. Wiley, Chichester, UK
Rutter M, Taylor E, Hersov L 1994 *Child and Adolescent Psychiatry*. Blackwell, Oxford, UK
Schmidt M H, Remschmidt H (eds.) 1989 *Needs and Prospects of Child and Adolescent Psychiatry*. Hogrefe & Huber, Toronto, ON
Schwab J J, Schwab-Stone M E 1999 History of child psychiatry in the USA. From social reform and psychoanalysis to psychiatry of the family. *Zeitschrift für Kinder und Jugendpsychiatrie und Psychotherapie* **27**: 277–81
Shapiro A K, Shapiro E S, Young J G, Feinberg T E 1988 *Gilles de la Tourette Syndrome*. Raven, New York
Simeon J G, Ferguson H B (eds.) 1990 *Treatment Strategies in Child and Adolescent Psychiatry*. Plenum Press, New York
Simon B 1996 The history of psychiatry: An opportunity for self-reflection and interdisciplinary dialogue. Essay review. *Psychiatry* **59**: 336–56
Steinhausen H C 1995 *Eating Disorders in Adolescence*. Walter de Gruyter, Berlin
Vitiello B, Jensen S 1997 Medication development and testing in children and adolescents—current problems, future directions. *Archives of General Psychiatry* **54**: 871–6
Wiener J M 1997 *Textbook of Child & Adolescent Psychiatry*. American Psychiatric Press, Washington, DC
World Health Organization (WHO) 1992 *The ICD-10 Classification of Mental and Behavioral Disorders-Clinical Descriptions and Diagnostic Guidelines*. WHO, Geneva

M. H. Schmidt and A. Maras

Child Care and Child Development

Childcare is regular care provided by someone other than a child's parents. Throughout human history, grandparents, siblings, and relatives have cared for young children, but since about the 1950s, nonparental childcare has become an increasingly visible and prevalent part of the lives of young children in most industrialized societies, largely because of increasing rates of maternal employment and of single-mother families. In the US, by 1997, the majority of infants and preschool children spent some time in childcare; many of them were in full-time care (35+ hours per week) (Capizzano and Adams 2000).

This rapid and dramatic change in the ecology of children's experience has raised a host of questions about the benefits and risks associated with nonparental childcare. Strong value commitments and assumptions have often influenced the questions asked. In the first wave of research in the 1970s, for

example, researchers asked whether 'day care' was harmful to children; they rarely considered potential benefits. At the same time, others were investigating the benefits of 'early childhood intervention' without considering possible harmful effects. Both groups were examining systems of nonparental care for young children, but the ways they framed questions and the labels they used led to different conclusions.

1. Is Extensive Childcare Harmful?

Why might we expect nonparental childcare to be harmful for children? According to the 'maternal deprivation' view, having one primary caregiver (usually the mother) with whom to develop an early attachment relationship is critical to the socioemotional development of young children. This view implies that even very high quality nonmaternal care could impair the development of a secure attachment to the mother if it seriously reduced the amount of time that a child spends with its mother. The most commonly used measure of attachment is the Strange Situation, a laboratory procedure in which children are observed during and after brief separations from their mothers. Secure attachment is indicated when the child goes to the mother and is comforted by her. Insecure attachment may be manifested by avoiding or ignoring the mother or by clinging to her without being comforted.

A large number of early studies concluded that childcare in the preschool years did not have harmful effects on attachment security or later socioemotional development, but the evidence is more mixed with respect to extensive nonmaternal care in the first year of life. Although the majority of children who receive full-time childcare in infancy have secure attachments to their mothers, some studies find elevated rates of insecure attachment when comparing these children to those in exclusive maternal care or part-time childcare (Clarke-Stewart 1989). A major study of 1,153 infants across the US indicated that children who spent extensive time in childcare during the first 3 years of life had mothers who were slightly less sensitive in interactions with them, and that extensive care was associated with elevated rates of insecure attachment only for children whose mothers were insensitive (NICHD Early Child Care Research Network 1997, 1999a, 1999b). As one reviewer concluded, 'Adverse effects on infant-mother attachment appear to occur only when infant day-care co-occurs with other risky conditions, …' (Lamb 1997).

It does not appear that time in childcare, in and of itself, has long-term effects on most other aspects of socioemotional development, but children with extensive care from infancy on tend to be more aggressive and assertive and to be less compliant to adults in some settings than are those with less child care experience. Many of these differences are not apparent when children have experienced high quality care (Lamb 1997).

2. Quality of Care

A second theory predicts that the effects of childcare depend on the physical and social environment provided. Harmful effects on both social and intellectual development might be expected if children receive less attention, affection, interaction, and stimulation from nonparental adults than they would from their parents. This view implies that the effects of care depend on its quality and on the quality of the home environment. Quality can be defined by the processes that occur in the setting: (a) sensitive, responsive, positive interactions of adults with children; (b) intellectually stimulating adult actions and activities, including appropriate language, reading, and play; and (c) toys, materials, and curricula that provide age-appropriate opportunities for learning. Quality can also be defined by structural characteristics, some of which can be mandated by regulatory agencies: (a) small ratios of children to adults; (b) small group sizes; (c) caregiver education and training in child development; (d) sufficient space per child; (e) a safe and clean physical environment; and (f) continuity and wages of the staff. Structural and process measures are correlated moderately with one another.

Whether process or structural indicators of quality are used, children in high-quality care have higher levels of language and intellectual skills, and perform better on academic tasks than do children in low-quality care. Although this finding is consistent across studies, there is some disagreement about whether quality produces a large enough difference to be socially significant (Lamb 1997, Scarr 1998). In one analysis, the quality of the child care environment and characteristics of the home environment during the first 3 years of life were compared as predictors of school readiness and language skill at age 3. The size of the childcare quality effect was about half that of the family environment, suggesting that childcare makes a substantial contribution to language and academic development (NICHD 1999a, 1999b). Critics have also argued that childcare effects do not last into later childhood, but evidence for long-term effects is beginning to accumulate. In a large study following children from age 4 through grade 3, children who attended higher quality childcare centers performed better on measures of cognitive skills (e.g., math and language abilities) while they were in childcare and, in many cases, through the end of second grade (Peisner-Feinberg et al. 1999).

High quality childcare also predicts social skills with peers and social competence with adults, but the relations between quality and social behavior are

weaker than those between quality and cognitive/academic performance. Some studies suggest that children's relationships with their teachers in child care forecast social skills and relations with teachers in elementary school (Howes et al. 1998, Peisner-Feinberg et al. 1999).

One problem in evaluating childcare effects is that families with more resources, better education, and more sensitive parenting styles (to name only a few attributes) place their children in higher quality childcare settings than do less advantaged families. Most recent studies have measured both family and childcare characteristics so that the independent contributions of each could be evaluated, but there is still the possibility that other unmeasured family attributes could affect childcare choices and children's cognitive and intellectual development.

Experimental studies, in which children are assigned randomly to enriched early care experiences or to control groups, demonstrate clearly that high quality early care has lasting effects on cognitive and academic skills, at least for children from economically disadvantaged families. Experiments avoid possible confounds between family and childcare attributes. The Abercedarian project enrolled children from low-income families in educational childcare from infancy until they reached school age. Children in the treatment group performed better than children in a control group (who received social and nutritional services) on tests of intelligence and school achievement throughout childhood and adolescence (Ramey et al. 2000). Less intensive early intervention programs for children ages 3–5 also produced lasting effects on children's school progress; participating children were less likely to be retained in a grade or to need special education than were children in control groups (Lazar and Darlington 1982).

The importance of childcare does not end when children enter school. Participation in formal after-school care programs that provide cognitive stimulation and positive adult interactions is associated with high academic achievement and low levels of behavior problems, particularly among low-income children (Posner and Vandell 1999). Children without adult supervision in the out-of-school hours are at risk for behavior problems and poor adjustment, particularly if they live in low-income families or unsafe neighborhoods (Pettit et al. 1999).

3. *Type of Care*

A great deal of childcare occurs in the child's home or someone else's home, especially for infants and toddlers. Caregivers can be grandparents, other relatives, or nonrelatives. In the US, a small percentage of infants receive center-based care, but the percentage increases as children reach ages 3–5. Many western European countries have extensive center-based care for children age 3 and over, but care settings for younger children vary considerably (Kamerman and Kahn 1991). The available data suggest both advantages and disadvantages of center-based care relative to home-based care for very young children. Infants and young children in center care show better language and cognitive development by ages 2 and 3 years than those in home-based settings of comparable quality, but they also have more communicable illnesses (e.g. colds and other respiratory illnesses) (NICHD 2000, in press).

Children attending a center or a childcare home with several other children have more experience with peers than do children in other types of home-based care. Children with extensive group experience develop better skills interacting with peers, but their caregivers also rate them higher on negative behavior (e.g. aggression, disobedience) than children in other forms of care (Lamb 1997). For young children, social skills and sociability with peers are often accompanied by a certain amount of aggression; perhaps early contact with peers leads to increases in a range of social behaviors.

4. *Do Effects Differ for Children from Different Family Backgrounds?*

In the US, where most childcare is funded privately, children from affluent families receive higher quality care than children from families with low and moderate incomes do. In some cases, children from very poor families receive slightly higher quality care than children from families with modest incomes, largely because of publicly funded programs for children in poverty (Phillips et al. 1994). In countries with publicly funded childcare for children from all income levels, average quality appears to be considerably better than it is in the US, and there are not wide discrepancies in quality associated with family income (Lamb 1997).

Some researchers have proposed a compensatory hypothesis: that children from disadvantaged homes might profit from childcare of reasonable or high quality because it provides more opportunities for learning and development than they receive at home. The complementary lost resources hypothesis suggests that children from highly advantaged homes may be harmed by childcare because it provides fewer opportunities than their home environments do. The evidence for these hypotheses is mixed. Almost all of the experiments exposing children to high quality care have included only economically disadvantaged children, and these children clearly profit from such care. Some investigations have found support for the lost resources notion, but most studies support the idea that high quality care provides benefits for children from a wide range of backgrounds (Lamb 1997).

5. Public Policy Issues

5.1 Public vs. Private Funding

Childcare is expensive. In some industrialized nations, public funds pay the great majority of the costs. Sweden, France, and some other countries provide large-scale publicly funded, full-day, preschool programs. In the US, at the other end of the spectrum, limited amounts of public funds are available to low-income families, and tax credits cover a small portion of childcare costs for people earning enough to owe taxes. Overall, parents pay the vast majority of the costs of care. At the same time, childcare workers and staff receive very low wages. In 1997, teachers in childcare centers, many of whom had college degrees, earned from approximately $13,000 to $19,000 per year; salaries of teaching assistants ranged from $10,500 to $12,250 per year (Whitebook et al. 1998). As a result, there is high turnover among childcare workers.

5.2 Family Leave vs. Infant Care

Care for infants is more expensive than care for preschoolers because they require more concentrated adult attention and they fare best in small groups. Many industrialized nations provide parents with paid family leave as an alternative to supplying center-based infant care (Kamerman and Kahn 1991). In the US, some workers receive paid family leave, but federal law requires only that parents receive up to 12 weeks of unpaid leave with guaranteed job security and medical benefits. Even this policy applies only to full-time, full-year workers in large organizations. As a result, many infants enter nonmaternal care between ages 2 and 3 months. There is some evidence mothers who return to work very early and whose babies have extensive nonmaternal care are less sensitive during interactions with their infants than are mothers who have more time at home (Clark et al. 1997). It may be more difficult to learn your infant's signals and respond appropriately when you spend a great deal of time at work starting very early in the child's life.

6. Conclusion

Most children in modern industrialized societies will spend a significant portion of their early lives in childcare, and many will require such care during the early school years as well. Initial fears that child care *per se* would be harmful have not been supported, but there is some evidence that extensive care, beginning in early infancy, can make it more difficult for mothers to relate sensitively to their infants. Whether childcare has positive or negative effects on children's development depends primarily on its quality. Quality is defined by sensitive, responsive, and stimulating caregivers and age-appropriate activities and curriculum. High quality care promotes cognitive, academic, and social development, and its effects last beyond the preschool years. High quality care for children from all economic levels is provided more successfully in countries that make large public investments in early care than in countries that require parents and child care staff to pay the costs of care.

7. Future Directions

Nonmaternal care is here to stay. Policy debates about what is 'good enough' care abound, and research is needed to identify thresholds of quality that make large differences, with particular attention to individual child and family characteristics. We understand how early childcare can contribute to cognitive and academic performance better than we understand how it can influence positive social behavior with both peers and adults. New designs, including experiments with random assignment, could help to answer these important questions.

See also: Divorce and Children's Social Development; Nontraditional Families and Child Development; Parenting: Attitudes and Beliefs; Parents and Teachers as Partners in Education

Bibliography

Capizzano J, Adams G 2000 *The Hours that Children under Five Spend in Child Care: Variations across States. A Report of Assessing the New Federalism*. The Urban Institute, Washington, DC

Clark R, Hyde J S, Essex M J, Klein M H 1997 Length of maternity leave and quality of mother-infant interactions. *Child Development* **68**: 364–83

Clarke-Stewart K A 1989 Infant day care: Maligned or malignant? *America Psychologist* **44**: 266–73

Howes C, Hamilton C E, Philipsen L C 1998 Stability and continuity of child-caregiver and child-peer relationships. *Child Development* **69**: 418–26

Kamerman S B, Kahn A J 1991 *Child Care, Parental Leave, and the Under-3s: Policy Innovation in Europe*. Auburn House, New York

Lamb M E 1997 Nonparental child care: Context, quality, correlates, and consequences. In: Damon W, Sigel I, Renninger K A (eds.) *Handbook of Child Psychology: Vol. 4. Child Psychology in Practice*, 5th edn. Wiley, New York, pp. 73–134

Lazar I, Darlington R 1982 Lasting effects of early education: A report from the Consortium for Longitudinal Studies. *Monographs of the Society for Research in Child Development* **47**(2-sup-3): 1–151

NICHD Early Child Care Research Network 1997 The effects of infant child care on infant-mother attachment security: Results of the NICHD study of early child care. Child Development **68**: 860–79

NICHD Early Child Care Research Network 1999a Child care and mother-child interaction in the first 3 years of life. *Developmental Psychology* **35**: 1399–1413

NICHD Early Child Care Network 1999b Child outcomes when child-care classrooms meet recommended standards for quality. *American Journal of Public Health* **89**: 1072–7
NICHD Early Child Care Research Network 2000 The relation of child care to cognitive and language development. *Child Development* **71**: 960–80
NICHD Early Child Care Research Network in press. Child care and common communicable illnesses: results from the NICHD Study of Early Child Care. *Archives of Pediatrics and Adolescent Medicine*
Peisner-Feinberg E S, Burchinal M R, Clifford R M, Culkin M L, Howes C, Kagan S L, Yazejian N, Byler P, Rustici J, Zelazo J 1999 *The Children of the Cost, Quality, and Outcomes Study Go to School*. University of North Carolina, Chapel Hill, NC
Pettit G S, Bates J E, Dodge K A, Meece D W 1999 The impact of after-school peer contact on early adolescent externalizing problems is moderated by parental monitoring, perceived neighborhood safety, and prior adjustment. *Child Development* **70**: 768–78
Phillips D A, Voran M, Kisker E, Howes C, Whitebook M 1994 Child care for children in poverty: Opportunity or inequity? *Child Development* **65**: 472–92
Posner J K, Vandell D L 1999 After-school activities and the development of low-income urban children: A longitudinal study. *Developmental Psychology* **35**: 868–79
Ramey C T, Campbell F A, Burchinal M R, Skinner M L, Gardner D M, Ramey S L 2000 Persistent effects of early childhood education on high-risk children and their mothers. *Applied Developmental Science* **4**: 2–14
Scarr S 1998 American child care today. *American Psychologist* **53**: 95–108
Whitebrook M, Howes C, Phillips D 1998 *Worthy Work, Unlivable Wages. The National Child Care Staffing Study, 1988–1997*. Center for Child Care Workforce, Washington, DC

A. C. Huston

Childbearing, Externalities of

In any family, lower fertility would raise income per family member at least in the short run, since the number of family members to share income would be smaller and women of reproductive age might do more market work. Would this higher per capita income justify a government policy to reduce the rate of population growth by lowering fertility? Not necessarily. By choosing to have a child, people express a preference for the child over the additional consumption that would otherwise be possible for family members. However, there may be costs and benefits to the additional child that are not directly borne by the decision-making parents, but are rather passed on to other families and to society as a whole. Such consequences, if not mediated by the market, are known as pure or technical externalities to childbearing. When they are present, then individually optimal childbearing decisions need not add up to socially optimal fertility. In this case, there is a role for government intervention to influence fertility decisions. In fact, externalities are pervasive in all societies. This article will discuss their sources, their size, and their policy implications.

1. *Why Childbearing Externalities Matter*

When consequences of childbearing are mediated by the market, they are called pecuniary externalities, as when an additional child reduces the wages of future workers by increasing their number. It has been shown that in this case, socially optimal fertility will not diverge from the individually optimal level (conditional on additional assumptions and a restrictive concept of social optimality; see Nerlove et al. 1987). Technical externalities do not pass through the market, as would be the case if an additional child meant higher taxes for others or a bigger hole in the ozone layer.

Economic theory asserts that in the absence of technical externalities, collective welfare will be maximized (in the sense of Pareto optimality) by individuals pursuing their own self-interest in the context of a competitive market (subject to some further conditions). In the presence of technical externalities, this is no longer so. Disregarding technical externalities, the sum of individually optimal fertility decisions should lead to the same outcome as a collective societal decision about fertility levels. With technical externalities, the outcomes would differ. Thus Garret Hardin, in a famous article, *The Tragedy of the Commons*, which first gave this issue prominence, called for 'mutual coercion, mutually agreed upon' (Hardin 1968, p. 1247). Subsequent articles by Demeny (1972), Blandy (1974), Ng (1986), Nerlove et al. (1987), and Willis (1987) developed the theory from a more rigorous economic standpoint.

Externalities to childbearing are salient in several modern policy contexts, as follows. First, in industrialized nations, fertility is on average about 1.5 children per woman, well below replacement level, leading to rapidly aging populations. Why is it so low? When old age support is provided to elders by their own adult children, this enters a couple's cost/benefit calculus and provides an incentive for higher fertility. In industrialized nations, although old age support is provided by the younger generations, this is done generally through public sector tax and transfer pension programs, rather than by the elders' own children. Thus the creation of public pension programs has created a large positive externality to childbearing: higher fertility is socially beneficial through the pension programs, but this social benefit does not impinge on individual decisions since it is the general level of fertility that matters.

Second, in developing nations, the public sector provides education and healthcare for children, so

additional children impose tax costs on others, a negative externality. Governments typically spend little on the elderly in these countries, and also there are few elderly, so the public costs of children dominate. This gap between the private and social costs of children is a negative externality that is often taken to justify government intervention.

Third, in all countries additional children will place additional demands on the environment, now and in the future, and most environmental amenities (clean air, fresh water, the ozone layer, climate and CO_2 emissions, biodiversity, forest cover) are outside the market. These negative environmental externalities to childbearing are therefore potentially very important, particularly in the industrialized nations where consumption per capita is greater.

Focusing on just a few issues like these can be misleading. A proper understanding of childbearing externalities requires a more comprehensive approach.

2. Sources of Externalities

Sources of technical externalities may be grouped into (a) common resources or collective wealth, (b) public sector transfers from one age group to another, and (c) provision of public goods or social infrastructure. In addition, there is a wide range of pecuniary externalities, of which the most important is a potential adverse effect on future wages or per capita incomes. Each of these three will be briefly discussed, before we turn to a quantitative assessment.

The existence and magnitude of externalities depend on institutional context, in particular on the existence of property rights in resources and on the size of the public sector. With full property rights and no public sector, most childbearing externalities would vanish.

2.1 Common Resources or Collective Wealth

The most basic externality to childbearing occurs when an asset is commonly owned and all members of the population have free access to it. Suppose a group shares a common pasture for its cows. The larger the group, the fewer cows each will be able to graze without degrading the pasture and the cows. Each birth increases the size of the group, but if there are many families, then this diluted effect will not count in the parents' self-centered cost benefit analysis. Thus there is a negative externality. Fertility will be higher, the group larger, and all worse off than with a collective fertility decision (Hardin 1968). A similar argument could be made about environmental amenities (water, air, climate, ozone layer, etc.). Nationally-owned land, parks, and mineral or fossil fuel deposits can likewise be important sources of negative externalities.

The argument also works in the opposite direction, when collective costs are shared, as with national debt. The US national debt is roughly $20,000 per capita. Additional members of the population due to births or immigration take on a share of this obligation as taxpayers, thereby reducing what must be paid by the balance of the population. This is a positive externality to childbearing.

2.2 Public Sector Inter-age Transfers

When parents rear their children and thereby transfer income to them, and when adults transfer income to their elderly parents, no externalities to childbearing arise: the need for transfers enters into the fertility decision. When transfers take place through the public sector, however, they do lead to externalities. The most important public transfers in this context are for health, education, and pensions.

2.3 Public Goods and Social Infrastructure

The costs of providing a network of roads rises with the size of the population served within a given area, but less than proportionately to the population increase; the same applies for communications networks. These are called quasi-public goods. In addition, there are pure public goods, which by definition cost no more to provide to many people than to few. The leading example is a nation's military force, which can protect a larger population just as well as a smaller one; other examples are broadcasting, weather forecasting, and scientific research. Provision of a given level of public goods is cheaper per capita for a larger population, since the tax bill per head will be lower. Public and quasi-public goods give rise to positive externalities of childbearing.

3. Evaluation

The first attempt to evaluate childbearing externalities appears to be have been made by Lee (1991). The following discussion will draw on the more comprehensive work by Lee and Miller (1990), which included estimates for six developing nations and the USA for the early to mid-1980s. Begin with collective wealth or debt. Great variations in natural wealth, particularly for oil in Saudi Arabia and land in Brazil, lead to negative externalities that dominate other sources for these countries. For the USA, national debt is an important source of positive externalities, as it would be for many OECD nations. An effort was made to evaluate other items, but none appears to be very important. Environmental externalities were not

addressed in Lee and Miller, for lack of evidence. Since then, a group of ecologists and economists have estimated the value of all the services provided by natural resources worldwide to fall within the range 16 trillion to 54 trillion dollars per year (Roush 1997). The midpoint estimate implies a negative externality of −$175,000 per birth worldwide, very approximately. For most nations, this number would dominate all externalities. Despite the great uncertainty surrounding the calculation, it warns us that environmental externalities can be very large and should not be ignored.

Turn now to public sector transfers, from taxpayers to beneficiaries of various programs. When the direction of transfers is downward from older to younger ages, such as for public education, then incremental births are costly and there is a negative externality. When the direction of transfers is upward from younger to older ages, such as for public pensions and healthcare for the elderly, then incremental births reduce the old age dependency ratio, and there is a positive externality. When these three major transfer programs and other smaller ones are summed, the net direction of transfers can be found. In the USA, the net direction is strongly upward, from younger to older. These results would be even stronger for other OECD countries, since their populations are typically older and their pensions more generous than in the USA. In the developing nations evaluated other than Brazil, the net direction of transfers is downward. Like a number of other Latin American countries, but unlike most countries in Asia and Africa, Brazil has a strong pension program. It is also likely that the net direction of public transfers throughout the developing nations is strongly downward, except for those countries in Latin America that have generous unfunded public pension programs.

Now consider public goods. Expenditures on the military are the main item in this category for the USA and Saudi Arabia, but for other countries social infrastructure dominates. Public goods and social infrastructure advantages generate positive externalities to childbearing for all countries, falling in the range of two to ten times the level of gross national product (GNP) per capita.

These three broad kinds of externalities can be summed to find the net externality.

If we ignore environmental externalities, there is a substantial positive net externality to childbearing in the USA, three times as great as per capita GNP. This would probably be approximately true for other OECD countries as well. None of the developing nations evaluated has a positive externality, while Brazil and Saudi Arabia have very large negative externalities. It is striking that Kenya and Bangladesh, two countries that are generally regarded as having serious population problems, both have externalities near zero relative to GNP per capita. If global environmental externalities were taken into account at the level discussed earlier, then all net externalities would be negative.

The estimates just discussed were based on a simple model and the analysis was done by comparative steady states. In more recent work, Lee and Miller (1997) examined fiscal impact externalities in a more detailed and nuanced manner. They projected tax payments over a very long horizon according to whether or not there was an additional birth in the base year, including the impacts of all descendants of the incremental birth as well. They found a positive net externality of US$170,000, which is six times per capita GNP for the reference year (which compares to 3.6 in Lee and Miller 1997, adjusted for non-fiscal effects). Given the increase in the cost of healthcare benefits and differences in some key assumptions, the numbers are in reasonable agreement.

4. *Broader Views*

The evaluations just reported are based on the narrow concept of technical externalities, following the main development of the theoretical literature. Some analysts argue for a broader view of externalities, even though the welfare implications are then no longer clear. Some stress pecuniary externalities: an additional birth will mean less land and less capital for future workers, and therefore lower productivity and wages. Nerlove et al. (1987) analyze this case, and show that under certain assumptions there is no technical externality and collective fertility decisions would not improve welfare. Nerlove et al. view parents as deriving utility from their own consumption, the number of children they have, and the future utility or consumption for their children. But it is possible that the goals of society differ from those of the individual parents. In particular, society may give the welfare of future generations greater weight than do the individual parents, or value wilderness or other aspects of the environment more highly. Society may care more about the distribution of income, or give a different weight to the welfare of women or preexisting children than does the household decision-maker. In these cases, there is no reason to expect that the sum of individual decisions will be socially desirable, and there will be many other reasons besides technical externalities to expect a gap between the individually and socially optimal level of fertility. There may also be a different kind of externality, in which one person's fertility may influence decisions by others. For example, if one couple uses contraceptives to control fertility, that may convey information to other couples, enabling them to make a better-informed decision. Or it may alter the social norms that influence the fertility decisions of others by weakening the influence of traditional institutions opposing contraceptive use. Or some couples may

imitate others. Dasgupta (1993) considers some of these possibilities.

5. *Implications for Policy*

The difficulty in assessing and quantifying environmental externalities, which are potentially very large for couples in industrialized nations, renders all the estimates even more uncertain. The numbers presented above are too shaky to guide policy decisions. Nevertheless it is useful to consider what they would entail if taken at face value, setting aside the question of environmental externalities.

Once the direction and magnitude of childbearing externalities have been identified, there is a clear policy implication: governments can, in principle, improve the welfare of their populations by inducing couples to have more children where externalities are positive, and fewer children where they are negative. One means to achieve this would be by internalizing the externality, such that through taxes or subsidies the couple is brought to face the full social cost or benefit of the incremental child when making its fertility decision. For example, in industrialized countries, the size of the retirement pension could be linked to each couple's fertility level; or a couple could receive a bonus of US$170,000 (say) on the birth of each child. In developing nations, each couple might be compelled to make a lump sum payment at the time of the birth of each child. In reality such policies appear to be neither practical nor desirable. They would adversely affect the well-being of children and probably fall disproportionately on the poorer segment of society in developing nations.

Nor would such policies necessarily be desirable from a theoretical perspective, because the measured externality depends sensitively on the institutional context and on public sector resource allocation. If a country has a positive externality to childbearing due to a large military establishment, the externality could be reduced by reducing the share of military expenditures in the national budget. If a country has a positive externality to childbearing due to a generous unfunded public pension program, the externality could be eliminated by switching to a funded system. If there is a large negative childbearing externality in a developing nation due to public education, note should be taken that fertility is falling rapidly in most parts of the world where it is not already low, and that transfers may soon be flowing upward across the age distribution. The existence of negative childbearing externalities reflects the ability of individual couples to pass on some of the childrearing costs to society at large. While not unimportant, this is probably not a major reason why fertility is high. More likely high fertility results in part from obstacles faced by couples in obtaining contraceptives, and in part from the absence of institutions which provide superior substitutes for many of the services that children provide – from insurance against sickness, unemployment, or death of spouse, to financial institutions and reliable pension programs. It would most likely be a mistake to attempt to fine-tune policy to these measured externalities rather than to make more efficient the institutional context within which childbearing decisions are made.

See also: Family Size Preferences; Family Theory and the Realities of Childbearing Behavior; Family Theory: Economics of Childbearing; Family Theory: Feminist–Economist Critique; Family Theory: Role of Changing Values; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Gender and Reproductive Health; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Personality Disorders

Bibliography

Blandy R 1974 The welfare analysis of fertility reduction. *The Economic Journal* **84**(333): 109–29

Dasgupta P 1993 *An Inquiry into Well-Being and Destitution*. Clarendon Press, Oxford, UK

Demeny P 1972 The economics of population control. In: *National Academy of Sciences, Rapid Population Growth: Consequences and Policy Implications*. Johns Hopkins University Press, Baltimore, MD

Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8

Lee R D 1991 Evaluating externalities to childbearing in developing countries: The case of India. In: *Consequences of Rapid Population Growth in Developing Countries*. Taylor and Francis, London, pp. 297–342

Lee R D, Miller T 1990 Population growth, externalities to childbearing, and fertility policy in the Third World. Proceedings of the World Bank Annual Conference on Development Economics, 1990. Supplement to *The World Bank Economic Review* and to *The World Bank Research Observer* pp. 275–304

Lee R D, Miller T 1997 The life time fiscal impacts of immigrants and their descendants. Project on the economic demography of interage income reallocation, demography, UC Berkeley. Draft of Chapter 7 for *The New Americans, a report of the National Academy of Sciences Panel on Economic and Demographic Consequences of Immigration*. National Academy Press pp. 297–362

Nerlove M, Razin A, Sadka E 1987 *Household and Economy: Welfare Economics of Endogenous Fertility*. Academic Press, London

Ng Y 1986 The welfare economics of population control. *Population and Development Review* **12**(2): 247–66

Roush W 1997 Putting a price tag on nature's bounty. *Science* **276**: 1029

Willis R 1987 Externalities and population. In: Johnson D G, Lee R D (eds.) *Population Growth and Economic Development: Issues and Evidence*. Wisconsin University Press, Milwaukee, WI

R. D. Lee

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Childhood and Adolescence: Developmental Assets

Developmental assets represent a theoretical construct first articulated in 1990 (Benson 1990). Based on a synthesis of scientific studies in pertinent fields, developmental assets identify a series of social and psychological strengths which function to enhance health outcomes for children and adolescents. The purposes of ongoing research pertaining to developmental assets are to develop new lines of scientific inquiry on the sources and consequences of strength-building approaches in child and adolescent development and to attempt to provide a conceptual roadmap to guide the design and implementation of community-wide initiatives that are aimed at promoting healthy development among children and adolescents. The conceptual, research, and application dimensions of developmental assets are described here.

1. *Developmental Assets in Social and Conceptual Context*

The framework of developmental assets weaves together into an *a priori* conceptual model, a set of developmental experiences, resources, and opportunities, each of which contributes to important health outcomes, conceived as both the reduction of health-compromising behaviors and the increase of positive or thriving outcomes such as school success. Though the framework is supported by scientific study, it was purposefully designed to fuel and guide community-based approaches to strengthen the natural and inherent socialization capacity of communities. Therefore, assets include the kinds of relationships, social experiences, social environments, and patterns of interaction known to promote health and over which a community has considerable control.

Developmental assets, then, represent a framework, grounded in scientific study, with the applied aim of reweaving the developmental infrastructure of a community by activating multiple sources of asset building. These include informal, nonprogrammatic relationships between adults and youth; traditional socializing systems such as families, neighborhoods, schools, congregations, and youth organizations; and the governmental, economic, and policy infrastructures which inform those socializing systems. The intent is to encourage the mobilization of asset-building efforts within many settings of a child's life and to increase those efforts for *all* children and adolescents within a community. The developmental assets framework, the theoretical underpinnings of the framework, and its partner concept of asset-building communities are discussed in depth in a series of recent publications (e.g., Benson 1997, Benson 1998, Benson et al. 1998).

1.1 *Connection to Other Areas of Scientific Inquiry*

The developmental asset framework is related to several other streams of scientific study, which also seek to identify positive developmental experiences and competencies known to enhance health and well-being among adolescents. Among these is the emerging exploration of protective factors in the fields of alcohol, tobacco, and pregnancy prevention. For example, Resnick and his colleagues (1997) demonstrated the importance of family and school connections in reducing multiple forms of health risk behaviors. Similarly, the study of resiliency has identified characteristics that enable some children to navigate through and around what often are debilitating environmental risks and experiences. This has contributed considerably to our understanding of the scope and nature of social and psychological strengths (Masten et al. 1990, Rutter 1985, Werner and Smith 1992). In addition, the more applied field of youth development champions the inclusion of this emerging body of knowledge about developmental strengths into policy, programs, and practice (Pittman and Cahill 1991). The developmental asset model builds on these areas of inquiry and includes a number of elements from them in the 40 core developmental assets.

The developmental asset framework also provides a complementary approach to the paradigm of deficit reduction. A deficit reduction model is focused on reducing threats, obstacles, and risks that interfere with healthy development. Among these are abuse; neighborhood violence; access to alcohol, other drugs, and firearms; poverty; and family dysfunction. While research has shown that these factors are related to a number of negative outcomes, efforts that focus mainly on controlling or reducing them represent incomplete approaches to health promotion. As Benson and colleagues (Benson 1997, Benson et al. 1998) have discussed, approaches which depend exclusively on deficit reduction may unintentionally expand the role of professionals, programs, and policy in child and adolescent health. This may be to the detriment of more informal and natural capacities that may be rooted in the community.

The asset model provides, then, an alternative and complementary set of 'benchmarks' or targets that can be added to the necessary and influential paradigm of risk reduction. Rather than focusing solely on problems or threats to be *reduced*, it accents relationships, experiences, resources, and opportunities to be *promoted*.

1.2 Community as a Context for Human Development

The developmental asset framework is also connected, both intellectually and strategically, to the fields of community change and community building. These areas have historically focused on the economic, service, and environmental infrastructures of a city, defining both the inherent capacity of local communities for promoting civic health (Kretzman and McKnight 1993, McKnight 1995) and identifying those core processes of engaged community which can inform the health of residents. Among these emerging constructs are social trust, personal efficacy, and social capital (Sampson et al. 1997). In addition, the conceptualization and definition of the developmental assets were informed by other community research efforts focusing on the concepts of social norms, civic engagement, indigenous leadership, and community capacity building (Benson 1997, Benson et al. 1998).

2. The Developmental Asset Framework

The original configuration of 30 developmental assets (e.g., Benson 1990) was expanded to 40 developmental assets in 1996, based on analysis of data gathered on 254,000 6th–12th-grade students, additional synthesis of child and adolescent research, and consultations with researchers and practitioners (Benson 1997). The framework's conceptual foundations are based on empirical studies of child and adolescent development, as well as applied studies in prevention, health promotion, and resiliency. The development of this conceptual foundation involved a research synthesis which focused on integrating developmental experiences that are widely known to inform three types of health outcomes among adolescents: (a) the prevention of high-risk behaviors (e.g., substance use, violence, sexual intercourse, school dropout); (b) the enhancement of thriving outcomes (e.g., school success, affirmation of diversity, prosocial behavior); and (c) resiliency or the capacity to overcome adversity. The assets were framed initially around adolescent development and they are assessed through a self-report survey (Leffert et al. 1998). The assets were extended downward conceptually to include young children (birth–age 10) which then encompasses a more lifespan context (Roehlkepartain and Leffert 2000).

The conceptualization of the asset framework identified naming developmental factors that were particularly robust in predicting health outcomes and for which there was evidence that they could be generalized across gender, race/ethnicity, and family income. In addition, the assets were conceived to reflect core developmental processes which include the relationships, social experiences, social environments, patterns of interaction, norms, and competencies over which a community of people has considerable control. That is, the assets are more about the primary processes of socialization than the equally important arenas of economy, services, and physical infrastructure of a city (Benson et al. 1998).

Because the developmental asset framework was designed not only to inform theory and research but also to have practical significance for the mobilization of communities, the 40 assets are placed in categories that have conceptual integrity and can be described easily to the residents of a community. As seen in Table 1, they are grouped into 20 external assets (i.e., environmental, contextual, and relational features of socializing systems) and 20 internal assets (i.e., skills, competencies, and commitments). The external assets include four categories: (a) support, (b) empowerment, (c) boundaries and expectations, and (d) constructive use of time. The internal assets are also placed in four categories: (a) commitment to learning, (b) positive values, (c) social competencies, and (d) positive identity. The scientific foundations for the eight categories and each of the 40 assets are described in more detail in Scales and Leffert (1999).

3. Measurement, Descriptive Data, and Prediction

Since 1996, numerous studies of 6th–12th-grade young people in public and private schools in the United States have been conducted using the *Search Institute Profiles of Student Life: Attitudes and Behaviors*, a self-report survey. This 156-item self-report survey measures the 40 developmental assets, developmental deficits (e.g., whether youth watch too much television or are victims of violence), thriving indicators (e.g., school success, physical health behaviors), and high-risk behaviors (e.g., alcohol and other substance use, antisocial behavior, school problems) (Leffert et al. 1998). The most recent aggregate sample is made up of 99,462 6th–12th-grade youth from public and alternative schools in 213 cities and towns in the United States who took the survey during the 1996–7 academic year. This sample has served as a focal point for several studies of the relation of assets to risk behaviors and thriving outcomes (Benson et al. 1998, Leffert et al. 1998, Scales et al. 2000).

3.1 Examples of Descriptive Data and the Additive Nature of Developmental Assets

The self-report survey is primarily used as a means of communicating aggregate data on a community's youth. A report, developed for each city or school district that uses the survey, often becomes a widely shared document and is used to frame a community-wide discussion and serves as a focal point to mobilize around raising healthy youth (Benson et al. 1998). A

Table 1
Forty developmental assets

External	
Support	1. *Family support*—Family life provides high levels of love and support
	2. *Positive family communication*—Young person and her or his parent(s) communicate positively, and young person is willing to seek advice and counsel from parents
	3. *Other adult relationships*—Young person receives support from three or more nonparent adults
	4. *Caring neighborhood*—Young person experiences caring neighbors
	5. *Caring school climate*—School provides a caring, encouraging environment
	6. *Parent involvement in schooling*—Parent(s) are actively involved in helping young person succeed in school
Empowerment	7. *Community values youth*—Young person perceives that community adults value youth
	8. *Youth as resources*—Young people are given useful roles in the community
	9. *Service to others*—Young person serves in the community one hour or more per week
	10. *Safety*—Young person feels safe at home, at school, and in the neighborhood
Boundaries and expectations	11. *Family boundaries*—Family has clear rules and consequences, and monitors the young person's whereabouts
	12. *School boundaries*—School provides clear rules and consequences
	13. *Neighborhood boundaries*—Neighbors take responsibility for monitoring young people's behavior
	14. *Adult role models*—Parent(s) and other adults model positive, responsible behavior
	15. *Positive peer influence*—Young person's best friends model positive, responsible behavior
	16. *High expectations*—Both parents and teachers encourage the young person to do well
Constructive use of time	17. *Creative activities*—Young person spends three or more hours per week in lessons or practice in music, theater, or other arts
	18. *Youth programs*—Young person spends three hours or more per week in sports, clubs, or organizations at school and/or in community organizations
	19. *Religious community*—Young person spends one or more hours per week in activities in a religious institution
	20. *Time at home*—Young person is out with friends 'with nothing special to do' two or fewer nights per week
Internal	
Commitment to learning	21. *Achievement motivation*—Young person is motivated to do well in school
	22. *School engagement*—Young person is actively engaged in learning
	23. *Homework*—Young person reports one or more hours of homework every school day
	24. *Bonding to school*—Young person cares about his or her school
	25. *Reading for pleasure*—Young person reads for pleasure three or more hours per week
Positive values	26. *Caring*—Young person places high value on helping other people
	27. *Equality and social justice*—Young person places high value on promoting equality and reducing hunger and poverty
	28. *Integrity*—Young person acts on convictions and stands up for her or his beliefs
	29. *Honesty*—Young person tells the truth even when it is not easy
	30. *Responsibility*—Young person accepts and takes personal responsibility
	31. *Restraint*—Young person believes it is important not to be sexually active or to use alcohol or other drugs
Social competencies	32. *Planning and decision making*—Young person knows how to plan ahead and make choices
	33. *Interpersonal competence*—Young person has empathy, sensitivity, and friendship skills
	34. *Cultural competence*—Young person has knowledge of and comfort with people of different cultural/racial/ethnic backgrounds
	35. *Resistance skills*—Young person can resist negative peer pressure and dangerous situations
	36. *Peaceful conflict resolution*—Young person seeks to resolve conflict non-violently.
Positive identity	37. *Personal power*—Young person feels he or she has control over 'things that happen to me'
	38. *Self-esteem*—Young person reports having high self-esteem
	39. *Sense of purpose*—Young person reports that 'my life has a purpose'
	40. *Positive view of personal future*—Young person is optimistic about her or his personal future

Source: 1997. Search Institute. Reproduced with permission.

Table 2
The relation of assets to patterns of high-risk behavior

High-risk behavior patterns		Percent with high-risk patterns			
Category	Definition	If 0–10 Assets	If 11–20 Assets	If 21–30 Assets	If 31–40 Assets
Alcohol	Has used alcohol three or more times in the past month or got drunk once or more in the past two weeks	53	30	11	3
Tobacco	Smokes one or more cigarettes every day or uses chewing tobacco frequently	45	21	6	1
Illicit drugs	Used illicit drugs three or more times in the past year	42	19	6	1
Sexual intercourse	Has had sexual intercourse three or more times in lifetime	33	21	10	3
Depression/ suicide	Is frequently depressed and/or has attempted suicide	40	25	13	4
Antisocial behavior	Has been involved in three or more incidents of shoplifting, trouble with police, or vandalism in the past year	52	23	7	1
Violence	Has engaged in three or more acts of fighting, hitting, injuring a person, carrying or using a weapon, or threatening physical harm in the past year	61	35	16	6
School problems	Has skipped school two or more days in the past month and/or has below a C average	43	19	7	2
Driving and alcohol	Has driven after drinking or ridden with a drinking driver three or more times in the past year	42	24	10	4
Gambling	Has gambled three or more times in the past year	34	23	13	6

Source: 1998. *Applied Developmental Science*. © Lawrence Erlbaum Associates, Inc., Hillsdale, NJ. Reproduced with permission.

dichotomous form of reporting the assets, whereby each asset is simplified into a single percentage of youth who have, or do not have, the asset, is utilized as an effective method for communicating the asset profile to diverse subgroups within a community. This also allows for a simple summation of the average number of youth assets in any given community. Based on this aggregate sample, youth report having, on average, 18 of the 40 developmental assets. As part of a standardized report, communities also receive a breakdown of the percentage of young people who report different levels of developmental assets. In the aggregate sample, 20 percent have 0–10 assets, 42 percent have 11–20, 30 percent have 21–30, and 8 percent have 31–40 (Benson et al. 1998). When these categories are combined in cross-tabulation with risk behaviors or thriving indicators, they become a significant source of information for community members. For example, Table 2 shows the percentage of 6th–12th grade youth in the aggregate sample who report that they engage in each of 10 risk behavior patterns by their level of developmental assets. The table also includes a definition of each risk behavior pattern. It is important to note that for each of the 10 risk behavior patterns, the percentage of students reporting the risk behavior declines as the level of assets rises. This same relationship between assets and risk behaviors has been observed across grade, gender, racial or ethnic group, and across all communities studied.

The opposite pattern between level of assets and positive or thriving behaviors has also been repeatedly demonstrated. For these assessments youth who report fewer assets are also less likely to report each of the thriving indicators including school success and the affirmation of diversity (Benson et al. 1998).

3.2 Grade and Gender Differences

Some variation has been observed across communities and in different subgroups of adolescents. Overall,

Table 3
Assets by grade within gender among an aggregate sample of youth surveyed in the 1996–7 school year

		Male		Female	
	All	6–8	9–12	6–8	9–12
1. Family support	64 percent	71 percent	59 percent	71 percent	60 percent
2. Positive family communication	26	30	20	35	23
3. Other adult relationships	41	37	40	44	43
4. Caring neighborhood	40	42	35	47	37
5. Caring school climate	25	26	19	34	23
6. Parent involvement in schooling	29	38	22	38	24
7. Community values youth	20	24	15	28	16
8. Youth as resources	25	29	21	31	21
9. Service to others	50	50	41	61	51
10. Safety	55	54	70	42	50
11. Family boundaries	43	43	38	48	44
12. School boundaries	46	55	35	61	41
13. Neighborhood boundaries	46	53	40	54	41
14. Adult role models	27	26	21	35	28
15. Positive peer influence	60	68	46	76	57
16. High expectations	41	50	35	50	35
17. Creative activities	19	15	13	27	21
18. Youth programs	59	59	58	59	58
19. Religious community	64	67	56	73	64
20. Time at home	50	53	45	56	48
21. Achievement motivation	63	59	54	71	71
22. School engagement	64	55	57	70	73
23. Homework	45	38	36	51	55
24. Bonding to school	51	48	45	59	54
25. Reading for pleasure	24	21	17	35	27
26. Caring	43	39	28	57	52
27. Equality and social justice	45	41	28	61	54
28. Integrity	64	53	58	65	75
29. Honesty	63	59	55	71	68
30. Responsibility	60	54	56	64	66
31. Restraint	42	53	25	67	36
32. Planning and decision making	29	24	25	32	33
33. Interpersonal competence	43	28	25	59	61
34. Cultural competence	35	31	24	46	42
35. Resistance skills	37	38	29	48	38
36. Peaceful conflict resolution	44	33	29	59	54
37. Personal power	45	40	48	41	49
38. Self-esteem	47	53	54	42	39
39. Sense of purpose	55	58	61	51	49
40. Positive view of personal future	70	70	70	70	71

$N = 99{,}462$.
Source: 1998. *Applied Developmental Science*. © Lawrence Erlbaum Associates, Inc., Hillsdale, NJ. Reproduced with permission.

females report experiencing more of the assets than males. With males and females across each of the asset levels, 45 percent of females report that they experience more than half of the 40 assets compared to 30 percent of males who report that they experience more than half of the assets.

When comparing the effect of grade and gender differences in each of the individual assets, at least small effects (about 0.20; Cohen 1988) are observed in about one-half of the assets, suggesting somewhat pervasive, but small, differences in the contextual experiences of boys and girls over the adolescent years (Leffert et al. 1998). In addition, a fairly consistent decline in the reports of the assets for both males and females across this age period are found. That is, young adolescents (i.e., 6th–8th graders) tend to report experiencing more of the assets than older adolescents (i.e., 9th–12th graders) (see Table 3).

3.3 Prediction of Risk Behaviors and Thriving Indicators

In reports to participating communities, analyses similar to that presented in Table 2 are included to demonstrate the relation of developmental assets to both the risk behaviors and thriving indicators. In other studies using aggregate samples, regression analyses are used to assess the extent to which the developmental assets are useful in predicting either a reduction in risk behaviors (Leffert et al. 1998) or a promotion of thriving indicators (Scales et al. 2000). Those analyses have shown that demographic variables accounted for a range of 5–14 percent of the total variance of each of the models constructed to examine risk behaviors. In each analysis, a set of the developmental assets contributed a significant amount over and above the influence of the demographic variables, accounting for a total of 21–41 percent of the variance explained in the reduction of each of the individual risk behavior patterns and for 66 percent of the variance in a composite index of risk behaviors.

Similarly, Scales and colleagues (Scales et al. 2000) examined the extent to which developmental assets predicted thriving behaviors and how it varied across different ethnic groups. The demographic variables accounted for a range of 1–8 percent of the total variance explained by each of the models and each model accounted for a total of 10–43 percent of the variance explained in the individual thriving indicators and from 47 to 54 percent of the variance explained in a thriving index.

4. Application

Developmental assets, then, have particular utility for predicting reduction in multiple forms of risk-taking and thriving behavior (e.g., Leffert et al. 1998, Scales et al. 2000). The descriptive portraits of assets from hundreds of American communities suggest that these developmental asset targets are normatively fragile. A set of comprehensive and interlocking strategies have been proposed to mobilize the inherent and natural asset-building capacity of community residents (Benson 1997, Benson et al. 1998) and community socializing systems, including families (Roehlkepartain and Leffert 2000), schools (Starkman et al. 1999), congregations (Roehlkepartain 1998), and youth-serving organizations (Nelson 1998). New conceptual and measurement efforts are underway to extend the asset framework to a series of developmental phases in the 0–20 age range.

Though the focus of this last section is on the influence of communities, it is noted here that the developmental assets construct can be applied in a wide variety of ways. For example, national youth-serving systems and their local affiliates utilize the framework for strategic planning and as a training tool for staff, boards, and volunteers. A growing number of foundations utilize the work for both frame-funding initiatives and to evaluate proposals. Professionals (e.g., social workers, counselors) utilize the framework to design interventions for individual children and adolescents.

4.1 Defining Asset-building Communities

Asset-building communities are geographies of place which maximize attentiveness to promoting developmental strengths for all children and adolescents (Benson 1997). The dynamics and processes by which communities mobilize their asset-building capacity are a relatively unexplored line of inquiry, both theoretically and empirically. An initial framework for understanding the asset-building capacity of communities provides a set of core principles (Benson 1997, Benson et al. 1998). Among these are the principles of developmental redundancy (the exposure to asset-building people and environments within multiple contexts), developmental depth (a focus on nurturing most or all assets in children and adolescents), and developmental breadth (extending, by purpose and design, the reach of asset-building energy to all children and adolescents).

In activating these core principles, five sources of asset-building potential are hypothesized to exist within all communities, each of which can be marshaled via a multiplicity of community mobilization strategies. These sources of potential asset-building influence include: (a) sustained relationships with adults, both within and beyond family; (b) peer group influence (when peers choose to activate their asset-building capacity); (c) socializing systems; (d) community-level social norms, ceremony, ritual, policy, and resource allocation; and (e) programs, including school- and community-based efforts to nurture and build skills and competencies.

In brief, asset-building communities are distinguished as relational and intergenerational places, with a critical mass of socializing institutions (e.g., families, schools, neighborhoods, youth organizations, religious communities) choosing to attend to the developmental needs of all children and adolescents. Developmental assets become a language of the common good, uniting sectors, citizens, and policy in the pursuit of shared targets for all children and adolescents. The commitment of a community and its people, institutions, and organizations is both long-term and inclusive.

Ultimately, rebuilding and strengthening the developmental infrastructure in a community are conceived less as a program implemented and managed by professionals and more as a mobilization of public will and capacity. A major target for this level of community engagement is the creation of a normative culture in which all residents are expected by virtue of

their membership in the community to promote the positive development of children and adolescents.

Within the context of American society, this vision requires considerable transformation in prevailing resident and socialization systems, norms, and operating principles. As argued in numerous publications defining this conceptual model of asset-building community, American cities are typically marked by age segregation, civic disengagement, social mistrust, a loss of personal efficacy, and the lack of collaboration across systems.

4.2 Asset-building Communities: The National Movement

The marshaling of community capacity to consistently and deeply attend to the development of children and adolescents is conceived less as the implementation of a program and more the awakening of latent human and institutional potential to build developmental strengths. A series of practical tools targeted at community residents and civic leaders provides conceptual and strategic counsel for mobilizing asset-building capacity. The first asset-building initiative began in St. Louis Park, Minnesota in 1995. In the following five years, more than 500 other American communities began to craft community-wide initiatives. Organized more as a social movement than as the replication of a program, communities are encouraged to tailor their initiatives to local realities and capacities and in response to the data from local asset profiles. Because these initiatives are complex, multi-sector 'experiments' in changing local culture, and because they occur in a variety of rural, suburban, and urban settings, there is increasing investment in learning from these communities about innovations and effective practices in mobilizing residents and systems, with 'feedback loops' emerging to inform both the theory of community change and the development of practical resources. Several longitudinal studies will add additional insight to this evolving knowledge about the influence of community on human development.

See also: Adulthood: Developmental Tasks and Critical Life Events; Biopsychology and Health; Childhood Health; Childhood Sexual Abuse and Risk for Adult Psychopathology; Community Organization and the Life Course; Developmental Sciences, History of; Divorce and Children's Social Development; Early Childhood: Socioemotional Risks; Environments for Education; Environments for Learning; Human Development and Health; Human Development, Bioecological Theory of; Human–Environment Relationships; Infancy and Childhood: Emotional Development; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development; Mental Health Programs: Children and Adolescents; Poverty and Child Development; Psychobiology of Stress and Early Child Development; Social Competence: Childhood and Adolescence

Bibliography

Benson P L 1990 *The Troubled Journey: A Portrait of 6th–12th Grade Youth*. Search Institute, Minneapolis, MN

Benson P L 1997 *All Kids Are Our Kids: What Communities Must Do To Raise Caring and Responsible Children and Adolescents*. Jossey-Bass, San Francisco, CA

Benson P L 1998 Mobilizing communities to promote developmental assets: A promising strategy for the prevention of high-risk behaviors. *Family Science Review* **11**: 220–38

Benson P L, Leffert N, Scales P C, Blyth D A 1998 Beyond the 'village' rhetoric: Creating healthy communities for children and adolescents. *Applied Developmental Science* **2**: 138–59

Cohen J 1988 *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ

Kretzman J P, McKnight J L 1993 *Building Communities From the Inside Out: A Path Toward Finding and Mobilizing a Community's Assets*. Center for Urban Affairs and Policy Research, Evanston, IL

Leffert N, Benson P L, Scales P C, Sharma A R, Drake D R, Blyth D A 1998 Developmental assets: Measurement and prediction of risk behaviors among adolescents. *Applied Developmental Science* **2**: 209–30

Masten A S, Best K M, Garmezy N 1990 Resilience and development: Contributions from the study of children who overcome adversity. *Development and Psychopathology* **2**: 425–44

McKnight J 1995 *The Careless Society: Community and its Counterfeits*. Basic Books, New York

Nelson L I 1998 *Helping Youth Thrive: How Youth Organizations Can—and Do—Build Developmental Assets*. Search Institute, Minneapolis, MN

Pittman K J, Cahill M 1991 *A New Vision: Promoting Youth Development*. Center for Youth Development and Policy Research, Academy for Educational Development, Washington, DC

Resnick M D, Bearman P S, Blum R W, Bauman K E, Harris K M, Jones J, Beurhring T, Sieving R E, Shew M, Ireland M, Bearinger L H, Udry J R, Tabor J 1997 Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association* **278**(10): 823–32

Roehlkepartain E 1998 *Building Assets in Congregations: A Practical Guide for Helping Youth Grow Up Healthy*. Search Institute, Minneapolis, MN

Roehlkepartain J L, Leffert N 2000 *What Young Children Need to Succeed: Working Together to Build Assets From Birth to Age 11*. Free Spirit Press, Minneapolis, MN

Rutter M 1985 Resilience in the face of adversity: Protective factors and resistance to psychiatric disorder. *British Journal of Psychiatry* **147**: 598–611

Sampson R J, Raudenbush S W, Earls F C 1997 Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* **277**: 918–24

Scales P C, Benson P L, Leffert N, Blyth D A 2000 Contribution of developmental assets to the prediction of thriving among adolescents. *Applied Developmental Science* **4**: 27–46

Scales P C, Leffert N 1999 *Developmental Assets: A Synthesis of the Scientific Research on Adolescent Development*. Search Institute, Minneapolis, MN
Starkman N, Scales P C, Roberts C 1999 *Great Places to Learn: How Asset-Building Schools Help Students Succeed*. Search Institute, Minneapolis, MN
Werner E E, Smith R S 1992 *Overcoming the Odds: High Risk Children from Birth to Adulthood*. Cornell University Press, Ithaca, NY

P. L. Benson and N. Leffert

Childhood: Anthropological Aspects

The anthropological study of childhood first documents and accounts for the variety of childhoods found around the world; second, uses the comparative ethnographic record to test hypotheses about human development; and, third, studies the mechanisms in child, family, and community life for the acquisition, internal transformations, sharing, and intergenerational transmission of culture.

A thought experiment will illustrate the anthropological point of view about childhood. Imagine a newborn, healthy infant. What is the most important thing that you could do to influence the life of that infant? Most respond by mentioning dyadic interaction with the baby: hold and touch the infant a lot; provide good nutrition and health care; provide stimulation to achieve school success; love the baby; give it wealth and social capital, and so forth. Anthropologists believe that the most important influence in human development is the *cultural setting within which the infant will grow up*. It is how, why, and by whom children are held, loved, fed, stimulated, punished, provided resources, and so forth, and how that varies so widely across human communities, that is the focus of inquiry. Shaping a whole person engaged in family and cultural community life is the 'purpose' of childhood development from an anthropological perspective. Childhood is a cultural project with goals, meanings, constant adaptation, and struggle, and anthropology provides the evidence for the startling and remarkable varieties of childhoods lived around the world. Biological, psychological, and cultural anthropologists collaborate in the study of childhood, since biology, mind, and culture are all required to understand childhood.

The study of childhood and the process of children acquiring culture 'was almost entirely neglected by anthropologists until after 1925' (Whiting 1968). Although much progress has been made, anthropology does not yet provide a single unified theory of why and how childhoods vary around the world, or of childhood acquisition of culture. Rather the field offers rich, multivariate hypotheses and data on childhood (Super and Harkness 1997).

1. *The Stages of Childhood*

Five stages of human growth and development are common to *Homo sapiens*: infancy, childhood, juvenility, adolescence, and adulthood (Bogin 1999). Margaret Mead described lap children (infants, aged 0–1), knee children (toddlers, 2–3), yard children (preschool, 4–5), and community children (juveniles in middle childhood, 6–12). Anthropologists analyze the cultural meaning of the very idea of 'stages,' since stages are used to account for children's behavior ('he's crying but it is OK, because he's still a toddler'), as well as to assure and define normal and appropriate development ('she is eight, and so old enough to start helping run our household'). Human cultures weave wonderful variations, meanings, and stories around pan-human maturational stages of childhood. The Beng of Ivory Coast for example, believe that young children are still partly in yet another 'stage,' a cultural world called *wrugbe*, where ancestors share life with prebirth children who are ambivalent about leaving that world. This helps explain for Beng why infants cry or are sickly: they want to return to *wrugbe*.

2. *Conceptions of Childhood in Anthropology*

There are a variety of perspectives on childhood in anthropology. In one view, children are socialized into a set of norms and customs that they learn and then perpetuate. In this view, children are small adults in the making, ready receptors of traditions, shaped by parents and community adults to insure continuity in cultural and moral education, competence for survival in the ecology of the community, respect for tradition, appropriate behavior and respect for elders in demeanor and gender roles.

Second, children's personalities and minds are understood as reflections of the cultural themes as well as the anxieties children grow up with (such as in the work in Bali of Bateson and Mead 1942). The focus is on the semiotics and communication of cultural meanings to children, on how these cultural patterns are absorbed and internalized, in turn reproducing the meanings as well as neurotic obsessions of their parents' cultures.

Third, the psychocultural, or personality integration model (Whiting and Whiting 1975), begins with the climate, history, and ecology of a community, which shapes child-care practices, which in turn produce psychological effects on children, effects produced by direct social learning as well as by psychodynamic processes shaping personality and defenses in children. These children become adults who then project into myths, rituals, art, and other forms (including in turn their own practices as parents) the learned patterns as well as intrapsychic conflicts produced in childhood and shared by others in their community. Children and adults alike have universal needs of the self—

hunger for recognition, reward, and material and bodily satisfaction—to which cultures respond through the cultural careers made available to children in a community (Goldschmidt 1990). Culture inevitably thwarts these hungers, leading to intrapsychic and cultural conflicts. Melford Spiro used psychodynamic and sociocultural approaches to understand the ideological, political, and ecological reasons for and consequences of the care of children by designated community caretakers, or *metapelets*, in socialist-inspired agricultural collective groups in Israel (Spiro 1975).

Fourth, anthropologists study the 'developmental niche' of childhood: everyday physical/social settings, cultural customs of care, and the psychology of the caretakers as shaped by their cultural models of parenthood that direct behavior (the goals, meanings, and rationales for parenting and being a child) (Harkness and Super 1996). Parenting of children also is shaped by the organic hardware given by our common mammalian heritage, and by socioeconomic conditions in the community. Children experience culture as it is practiced within their family's daily routine of cultural life. Cultural routines consist of activities children engage in (mealtimes, bedtimes, family visits, chores, going to church, school, play, etc.) Activities are the primary mechanisms bringing culture to and into the mind of the child. Activities consist of goals and values; tasks of an activity; the scripts for how to engage in that activity; the people present and participating in the activity; and the motives and feelings of those involved (Weisner 1996).

Finally, some anthropologists view childhood itself as a cultural construction shaped by forces within as well as outside a single cultural community. The very idea of what a child or parent is, in this view, is more the outcome of processes of power in an increasingly global political economy, in which children as well as parents, are 'constructed' or 'positioned' by these agents of power (Stephens 1995).

3. *Some Cultural Influences on Children's Development*

3.1 *Cultural Scale and Complexity*

Children in more complex societies (with occupational specialization, an extensive market economy, a nucleated settlement pattern, centralized and hierarchical political and legal system, and a centralized religious priesthood) are more likely to seek help and assistance from others, to try and dominate or control others, and to be more egoistic. Children in less complex societies are more likely to show nurturance toward other children (to offer assistance and respond to their requests), be more responsible, make more responsible suggestions to others, and do more tasks required for family and community survival. Mothers who have heavy subsistence workloads are more likely to expect responsible work from children and use stricter discipline. Children living in extended, joint, or expanded households and family systems are more often involved in directive, aggressive interactions, while children living in smaller nuclear families are more often engaged in sociable and intimate interactions with parents and others, and fathers are more involved with children. Of course children and adults everywhere are nurturant, seek help, or are aggressive. These patterns only reflect the modal tendencies of communities, not a rigid uniformity within them (Whiting and Edwards 1988).

3.2 *Gender Differences*

Gender differences in children's development are recognized and shaped by all cultures (Ember 1981). Of five kinds of interpersonal behavior in children aged from three to 11—nurturance, dependency, prosocial dominance, egoistic dominance, and sociability—girls on average were more likely than boys to be nurturing toward others, while boys were more likely to be egoistically dominant and aggressive than girls. Play styles and types vary by gender (girls are more likely to do work-play and to do so nearer their homes, for instance). Women and girls do most 'mothering' of children well into the juvenile period in most cultures, so girls experience care by their own sex, while boys do not, leading to differences in early gender identification, and psychosocial and self-development. Peer groups have a tendency to segregate by gender, and children prefer same-sex children to interact with. Cultures with more mixed-age and mixed-gender groups around children are likely to have less sex-segregated roles for children. Individual differences among boys and girls, even within communities where there are strong overall gender differences, are usually substantial.

Father roles are recognized in all societies. Fathers seldom are involved in direct care of infants and young children but fathers do have complementary nurturing and affiliative roles, and are more involved in economic, protective, and didactic child training. In a study of 80 preindustrial societies, fathers were more proximate and involved with young children in monogamous, nuclear-family, and nonpatrilocal situations, and wherever mothers make relatively large contributions to family subsistence. Father involvement is related to sociocultural evolution: foraging societies report more father participation in childcare, while horticultural, agrarian/peasant, and pastoral tend to have less. There is an upswing in contemporary societies in encouraging paternal care. The cultural beliefs about gender (how women's as well as men's roles are defined by parenting), as well as the ability of fathers to provide consistently for their children,

influence father involvements. Poverty and uncertain economic life, or migration and dislocation, can drastically change father as well as mother involvement in patterns of childcare.

3.3 Emotional Development

Emotional development in childhood is influenced by cultural expectations at each developmental stage about the kinds of demeanor expected. A child should show that he/she is a certain kind of cultural person with an appropriate self and identity. Cultural management of emotion relies on what Robert Levy (1973) called 'redundant cultural control.' Tahitians (Levy 1973), for example, as well as many Pacific Island, Asian and other cultures, expect children to be calm, gentle, and quiet in demeanor (except for an extended period of adolescence and youth called *taure'are'a* in Tahiti, in which adventures, autonomy, rebellion, and aggressiveness are culturally expected and common). Redundant community management of 'gentleness' includes many beliefs and practices: children are somewhat distanced from their mothers and fathers after infancy and live with peers; socialization networks are diffuse, meaning that affect towards others is diffused; severe anger is 'strongly discouraged' while mild transient episodes are tolerated; threats are common while actual aggression towards children is not; accidents are reinterpreted as punishment by spirits for aggression and this is widely believed to be true; there can be magical retaliation for serious anger; and it is generally shameful to show lack of control. A *culture complex* of many interrelated beliefs and practices of this kind is a strong sign that some emotional pattern or competence in children is of adaptive and moral importance to a society.

3.4 Basic Trust and Attachment

Basic trust and attachment are fundamental in childhood in all cultures. Anthropological studies show that a wide range of family and parenting practices can produce close affiliation and trusting attachments in children. Successful attachment does not depend on only one kind of maternal care in nuclear families, nor a specific kind of infant and toddler behavioral style. Although the individual child is named and recognized everywhere, individualism and egoistic autonomy as goals are not at all universal; rather, sociocentric and interdependent self-development are common ideals (Shweder and Bourne 1991). Chisholm (1999) proposes an evolutionary developmental hypothesis regarding trust and attachment. Environments varied during the long course of evolution. Less threatening, more favorable material and social conditions led to greater investment in fewer children, and so encouraged closer attachments to one or a few caregivers. Unfavorable conditions encouraged what are called 'insecure' or avoidant/ambivalent infant and child attachments, since in threatening, insecure conditions that was the more adaptive, successful parental and child response likely to increase the chances of children reaching reproductive age. Most cultures provide multiple caretakers to children, not a single person, and care is 'socially distributed.' Indeed, living an entire childhood exclusively in one's natal home may well be the exception around the world. Older siblings and cousins are widely used as caretakers. Extended families in village and agrarian-based societies have high levels of multiple care of children. In India and elsewhere in Southeast Asia, for instance, there is an intense, 'relational,' childhood experience with several 'maternal' figures, a pattern found widely across cultures (Weisner and Gallimore 1977).

3.5 Developmental Goals

Anthropology does not assume that competencies valued in Western communities (verbal skills, cognitive abilities, or signs of egocentric autonomy of the self, for instance), are necessarily meaningful child developmental outcomes elsewhere, although all cultures are concerned over some version of good communication, mental ability, and self- and personhood. LeVine et al. (1994) contrast pedagogical goals (cognitive and social stimulation to prepare children for literacy and schools, as well as for an individualistic and autonomous self- and personhood away from their natal home) and pediatric goals (concern for survival, health, and physical growth of infants, and subsequent responsible engagement in family subsistence and continuity), comparing families in Boston, USA and the Gusii of western Kenya. Many parents and cultures have mixed goals, and are ambivalent about the constantly changing requirements for childhood. Anthropology has a unique point of view regarding the goals for a good childhood: the production of cultural well-being in children. Well-being is more than physical health or the attainment of skills and competence, or of successful subsequent reproduction, important as these are. Well-being is the ability of a child to engage and participate in the activities deemed desirable by a cultural community, and the psychological experiences that go along with that participation.

4. The Acquisition of Culture

The roles and settings in which children acquire culture matter for when and how children learn. Children are apprentices to more experienced community members in doing important tasks, and this apprenticeship situation is a powerful learning experience for children. Play and work blend in childhood learning. Imaginative, fantasy, toy, physical, and motoric play (including organized sports with rules) varies according to whether adults encourage it, whether it is

considered 'beneficial' for children by adults because it enhances desired competencies or societies' developmental goals (such as cognitive and school-like activities in many contemporary cultures), and whether children's sheer inventiveness, creativity, and exuberance take over. The Kpelle of Liberia have children playing on 'the mother-ground,' or open public spaces where children can observe, lurk nearby, and imitate adults going about their activities in an agrarian village community. Formal schooling and the 'outside' world of employment and the nation-state contrast sharply with this mother-ground of childhood. Anthropological studies of schooling find striking differences in the culture of classrooms around the world, including different teaching practices and student expectations. Cultures vary in the 'moral and cultural curriculum' accompanying literacy and other training, classroom and peer norms, gender, and class circumstances mirrored in school practices, and daily routines of school (Tobin et al. 1989). Participation in ceremonies and rituals at times of baptism, birthdays, naming ceremonies, puberty, and marriage also are powerful influences on children's acquisition of cultural knowledge. Such ceremonies crystallize cultural beliefs and practices; they intensify emotionally, politically, and socially salient key concerns that parents and communities have about childhood, and elevate the goals the community shares for children and parents (Turner, 1967).

Multiple cultural and mental processes are involved in culture acquisition. However, the relative importance of different mental and cultural mechanisms for emotional, social, or cognitive learning is currently not well understood in anthropology. Evolved tendencies of the mind prepare children to understand the world in certain ways. For example, children in widely disparate cultural communities seem to share understandings about what living things are like and how they behave and think (Hirschfeld and Gelman 1994). Psychodynamic processes transform emotionally salient cultural information. Stories and narratives embed cultural knowledge, shape recall, and organize cultural knowledge into sequences with shared local meaning. Sociolinguistic studies of child language acquisition show wide variations in how and when parents talk to their children, and view language learning as embedded in interactional routines shaped by cultural practices, with children as active learners (Schieffelin and Ochs 1986). Cultural knowledge is available to and used by children in the form of cultural models and schemas for how to comprehend and act in the world. Neither children nor parents generally 'know that they know' most cultural knowledge, and it is not usually in conscious awareness, even as they act in accord with their culture's beliefs. Cultural beliefs and practices have powerful 'directive force' in guiding child behavior and child socialization in part because of this shared, implicit, everyday understanding put into action (D'Andrade 1995).

5. *Anthropological Methods and the Study of Children*

Anthropological methods for the study of children include ethnography and participant observation (Weisner 1996). These methods fit with the anthropological concept of childhoods lived in cultural pathways in naturalistic settings. Systematic observational procedures, field guides for comparative studies, and special procedures for sampling children's activities and time use enhance ethnography (Munroe and Munroe 1994). Anthropologists also use assessments standard in child development for comparing physical growth, and the cognitive and socioemotional life of children, often revising these to insure that culturally appropriate procedures and meaningful outcomes are being measured. Film and video records of childhood are invaluable for comparative studies of cultural activities, emotional expression, holding patterns, or gaze and attention.

6. *Anthropology and the Study of Childhood in the Twenty-first Century*

Anthropology has always been concerned with the experience and the cultural worlds of minorities, the poor and non-literate, and of those, including children, who are so often unable to give voice to and represent their own world. Life histories and autobiographical accounts have provided rich data, as in the classic *Nisa: The Life and Words of a !Kung Woman* (Shostak 1981). Scheper-Hughes (1992) describes infants and young children in deeply impoverished political–economic circumstances in urban northeastern Brazil, circumstances leading to high infant and child mortality, and anger and despair among parents (and anthropologists). Anthropological studies of African-American families and economically downwardly mobile families in the USA demonstrate how some (but not all) can rely on extended kin in their struggles with poverty. Anthropological studies of childhood disability and deviance find greater acceptance and social integration of children with physical and cognitive disabilities in many communities, as long as children are able to live as sufficiently cultural persons in their communities and are not violent or dangerous to others (Ingstad and Whyte 1995). Anthropologists are concerned with children at risk around the globe, including, for example, children under stress from academic examinations in Japan and Korea, immigrant children in Europe and elsewhere, street children, or children facing change in East Africa. Child sexual and physical abuse around the world is now a recognized concern for anthropologists. Cultural beliefs and practices regarding appropriate discipline and treatment of children clearly do vary widely, and Western notions of abuse are not universal. However,

repeated and unchecked physical aggression, or intrafamilial sexual relations between close kin and children are nowhere defined as normative and acceptable (Korbin 1981). Anthropologists are concerned with children's rights, recognizing their vulnerable status and the lack of provision of basic protections for children (*Cultural Survival Quarterly*). World youth cultures are growing in importance due to the influence of the Internet and mass communications around the world. These are all topics for the anthropology of childhood in the twenty-first century. However, the comparative study of powerful local and regional cultural differences in parenting, childhood, and family life across populations around the world will continue to provide enduring scientific questions for anthropology.

See also: Adolescent Development, Theories of; Childhood Health; Children and the Law; Gender-related Development; Infancy and Childhood: Emotional Development; Life Course in History; Life Course: Sociological Aspects; Trust, Sociology of; Youth Culture, Anthropology of; Youth Culture, Sociology of

Bibliography

Bateson G, Mead M 1942 *Balinese Character*. Special Publications of the New York Academy of Sciences **2**. New York Academy of Sciences, New York

Bogin B 1999 *Patterns of Human Growth*, 2nd edn. Cambridge University Press, New York

Chisholm J 1999 *Death, Hope, and Sex. Steps to an Evolutionary Ecology of Mind and Morality*. Cambridge University Press, Cambridge, UK

Cultural Survival Quarterly. World Report on the Rights of Indigenous Peoples and Ethnic Minorities. Cambridge, MA (http://www.cs.org)

D'Andrade R 1995 *The Development of Cognitive Anthropology*. Cambridge University Press, New York

Ember C M 1981 A cross-cultural perspective on sex differences. In: Munroe R H, Munroe R L, Whiting B B (eds.) *Handbook of Cross-cultural Human Development*. Garland, New York

Goldschmidt W 1990 *The Human Career*. Blackwell, Cambridge, MA, pp. 531–80

Harkness S, Super C M (eds.) 1996 *Parents' Cultural Belief Systems: Their Origins, Expressions, and Consequences*. Guilford Press, New York

Hirschfeld L A, Gelman S A 1994 *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, Cambridge, UK

Ingstad B, Whyte S R 1995 *Disability and Culture*. University of California Press, Berkeley, CA

Korbin J E (ed.) 1981 *Child Abuse and Neglect: Cross-cultural Perspectives*. University of California Press, Berkeley, CA

LeVine R A 1988 Human parental care: Universal goals, cultural strategies, individual behavior. In: LeVine R A, Miller P M, West M M (eds.) *Parental Behavior in Diverse Societies. New Directions for Child Development* **40**. Jossey-Bass, San Francisco, pp. 3–12

LeVine R A, Dixon S, LeVine S, Richman A, Leiderman P H, Keefer C H, Brazelton T B 1994 *Child Care and Culture. Lessons from Africa*. Cambridge University Press, Cambridge, UK

Levy R 1973 *Tahitians. Mind and Experience in the Society Islands*. University of Chicago Press, Chicago

Munroe R L, Munroe R H 1994 Behavior across cultures: Results from observational studies. In: Lonner W J, Malpass R (eds.) *Psychology and Culture*. Allyn and Bacon, Boston, MA, pp. 107–11

Scheper-Hughes N 1992 *Death Without Weeping. The Violence of Everyday Life in Brazil*. University of California Press, Berkeley, CA

Schieffelin, B B, Ochs E (eds.) 1986 *Language Socialization Across Cultures*. Cambridge University Press, New York

Schlegel A, Barry H 1991 *Adolescence. An Anthropological Inquiry*. Free Press, New York

Shostak M 1981 *Nisa: The Life and Words of a !Kung Woman*. Harvard University Press, Cambridge, MA

Shweder R A, Bourne E J 1991 Does the concept of the person vary cross-culturally? In: *Thinking Through Cultures Expeditions in Cultural Psychology* Harvard University Press Cambridge, MA, pp. 113–55

Spiro M E 1975 *Children of the Kibbutz*, 2nd edn. Harvard University Press, Cambridge, MA

Stephens S (ed.) 1995 *Children and the Politics of Culture*. Princeton University Press, Princeton, NJ

Super C M, Harkness S 1997 The cultural structuring of child development. In: Berry J, Dasen P R, Saraswathi T S (eds.) *Handbook of Cross-cultural Psychology*, 2nd edn. Allyn and Bacon, Boston, Vol. 2, pp. 3–39

Tobin J, Wu D, Davidson D 1989 *Preschool in Three Cultures: Japan, China, and the United States*. Yale University Press, New Haven, CT

Turner V W 1967 *The Forest of Symbols*. Cornell University Press, Ithaca, NY

Weisner T S 1996 Why ethnography should be the most important method in the study of human development. In: Jessor R, Colby A, Shweder R (eds.) *Ethnography and Human Development. Context and Meaning in Social Inquiry*. University of Chicago Press, Chicago, pp. 305–24

Weisner T S, Gallimore R 1977 My brother's keeper: Child and sibling caretaking. *Current Anthropology* **18**: 169–90

Whiting J 1968 Socialization: Anthropological aspects. *International Encyclopedia of the Social Sciences* **14**: 545–51

Whiting B B, Edwards C P 1988 *Children of Different Worlds: The Formation of Social Behavior*. Harvard University Press, Cambridge, MA

Whiting J W M, Whiting B B 1975 *Children of Six Cultures: A Psychocultural Analysis*. Harvard University Press, Cambridge, MA

T. S. Weisner

Childhood Cancer: Psychological Aspects

Substantial progress in medical anticancer treatment (chemotherapy, radiotherapy, surgery, bone-marrow transplantation) has improved dramatically the long-term survival rates of children and adolescents with the diagnosis of cancer (malignant tumors, lymphomas, and leukemias). Some decades ago, the essential threat for the family was to face the death and loss of

the child. Since more than two-thirds of the young patients survive, cancer in children and adolescents no longer is regarded as unevitably fatal but as a chronic disease which can result in complete cure, death' or survival with lasting neurological, orthopedical, or neuropsychological impairments.

At the end of the twentieth century, there is a consensus about the necessity of an honest explanation of the life-threatening nature of the disease to the child. The statistical survival rates, however, do not indicate the prognosis in an individual case, turning the former threat of certainty of death to the threat of uncertainty of the very personal fate. 'Damocles syndrome' is the term Koocher and O'Malley (1981) have coined to characterize the psychological situation of the family that has terminated oncological therapy but cannot be sure definitely to have defeated the disease.

1. Psychological Adaptation to Childhood Cancer as a Balance of Stress and Resources

The occurence of childhood cancer is not associated with any known predisposing social or psychological feature in the child or in the family. Its presence can be regarded a randomly occuring highly stressful life event for 'normal' families. As a consequence, psychological research and intervention is not concerned with the detection of suspected psychological causes of cancer but with an understanding of the stress impact imposed by cancer diagnosis on the child's and family's psychological functioning and well-being.

The risk of the child or the family to master or to fail the turmoils of the cancer experience may be compared to a balance with one scale pan containing all the stressors resulting from the disease and the other scale pan containing all the resources, competencies, and sources of social support to handle the stress. In some vulnerable families, the strong stressors and demands clearly outweigh only weak resources, thus increasing the probability of emotional suffering and behavioral disturbance in the child and the risk of marital discord or parental depression. Other families are characterized by resilience, i.e., the available resources are strong enough to outweigh the threatening disease impact, thus giving rise to a good adjustment and a personal sense of mastery. Actually, the large majority of patients succeed in mastering the plethora of disease and treatment-related stressors and challenges (Noll et al. 1999).

2. Enhancement of Effective Coping with Childhood Cancer by Family Counseling

Understanding adaptation to childhood cancer as a psychological balance between burden and resources may serve as a basis for effective psychological counselling strategies. Therefore, the therapeutical task on the way to enhancement of adaptation is a twofold one: the first one is to decrease stress impact and the second one is to increase resource availablity. In order to develop concrete counseling strategies for intervention, a differentiated knowledge is required of the most common and threatening sources of stress, on the one hand, and of the most powerful resources that can be mobilized, on the other. Selected typical stressors for the family of the cancer-sick child are presented in Table 1. They cover practical, emotional, social, and existential domains. As a counterpart, Table 2 shows important resources that can buffer and diminish cancer stress impact, improve adaptation, stabilize quality of life, and protect individual family members and the whole family structure from decompensation. Research has highlighted the central role of family communication behavior. For instance, the intensity of interparental exchange about disease issues is connected closely to an improved parent–child contact and a high disease related information level in

Table 1
Typical stressors, demands, and challenges frequently occurring during the course of childhood cancer

Organization of treatment (meetings with physicans, insurances, etc.)
Financial expenditures
Gathering information on cancer and its treatment
Repeated hospitalizations including the alternative to stay at home with the siblings or be in the hospital with the ill child
Treatment-related distress according to painful and frightening medical procedures (e.g., bone marrow aspirations) and treatment side-effects (e.g., nausea and vomiting due to chemotherapy)
Reduced time for recreation and holiday
Role conflict (e.g., father between ongoing job demands and family need for support)
Re-examinination of future perspectives (job perspectives of the mother; schooling of the child)
Sharing responsibility for treatment decisions (e.g., amputation in the case of bone sarkoma)
Subconscious violation, disappointment, frustration of the hope for a normal and healthy child
Religious doubts
Uncertainty of final outcome; fear of relapse, fear of loosing the child

Source: Chesler and Barbarin (1987), Noeker and Petermann (1990), Noeker et al. (1990)

Table 2
Protective family resources that enhance an effective adaptation to childhood cancer

Communication competencies
Attributing positive meanings to the situation
Commitment to the family as a unit
Engaging in active, flexible coping efforts
Balancing illness demands with other family needs
Maintaining social integration
Developing collaborative relationships with professionals
Ability to ask for and accept support
Clear religious or philosophical belief systems
Parenting style demonstrating consistent rules and expectations

Source: Koocher and O'Malley 1981, Kupst et al. 1995, Noeker and Petermann 1990, Patterson 1991

the child and the siblings. A fair and respectful sharing of the new burdens and tasks is associated with a high family coherence and stability (Noeker et al. 1990).

3. *Psychological Counseling Strategies and Interventions Across Different Phases of Childhood Cancer Treatment*

Based on the knowledge of typical stress and resource factors in families with childhood cancer, psychological counseling programs have been developed (Chesler and Barbarin 1987, Noeker et al. 1990). The counseling interventions are related closely to the specific stressors that characterize certain phases of treatment course. Counseling sessions may take place at the clinic or at home. Basically, all family members should join them; however, also sessions with family subsystems may be indicated, e.g., when mother and father want to find out an agreement on consistent parenting behavior. In the following, selected topics and strategies of a psychological counseling program are delineated according to the major phases of anticancer treatment process.

3.1 *Supporting the Initial Coping with Cancer Diagnosis*

After being told the diagnosis of childhood cancer, most parents show strong affective reactions like shock, sadness, anger, despair. The primary task of the counselor here is to show nonjudging acceptance of these emotions and to verbalize and clarify them. Ideosyncratic parental lay theories on why exactly their own child has got cancer may interfere with the medical treatment concept. When exploring these subjective disease and treatment concepts special attention has to be paid to frequent parental conceptions of feeling personally responsible and guilty for cancer manifestation (e.g., via genetics, bad food, psychological stress). Psychological clarification and integration of the strong emotional reactions and as well as a stepwise restructuring of inappropriate disease concepts improve the parents' cognitive capacity for understanding and accepting the oncological treatment concept.

The next major counseling task is to initiate an active, mutually respectful, and supportive coping attitude and behavior within the family unit including:

(a) clarification and expression what kind of support every person in the family needs most from other family members for stabilizing outer and inner equilibrium;

(b) encouragement of open communication and expression of personal feelings, concerns, sorrows, and anxieties among the family members; and

(c) creation of a shared commitment to regard the cancer disease as a common challenge for the entire family.

3.2 *Phase of Intensive Oncological Treatment*

After the family has overcome the initial shock of cancer diagnosis psychological counseling should assess primary vulnerability and resource factors in the child or in the family. Factors of vulnerability that were already present before diagnosis (e.g., behavioral or developmental disorders in the child, intense sibling rivalry, marital discord, inconsistent parenting styles, socioeconomic disadvantages) may interfere with a competent adaptation to demands of disease and treatment. Their assessment improves the chances for early and preventive intervention. At least as important as the assessment of risks and deficits, however, is the exploration of strenghts and competencies. If a counselor turns the family's attention to positive experiences of crises in the past that were resolved successfully, problem-solving strategies may be reactivated that have proven specific effectivity in this individual family. In addition, psychological counseling in this phase includes:

(a) Encouraging active participation in the management of the illness;

(b) Initiating clear and practical responsibilities for the day-to-day management of additional chores and tasks; making sure that agreements are perceived as a fair sharing of the burden by every family member;

(c) Parenting issues (e.g., ways to show affection without spoiling and pity);

(d) Mobilization and acceptance of concrete social support from neighborhood or classmates; and

(e) Facilitating communication between family and healthcare team.

3.3 *Positive Course: Long-term Remission of Disease and Process of Reintegration*

After termination of intensive oncological therapy, anxieties stay alive in patient and family concerning a return of the disease which would require a completely

new start of therapy with, however, decreased chances for cure. Families suffer from intrusive thoughts and emotions concerning cancer relapse. Cognitive-behavioral intervention strategies can support the patient and family members to observe these recurrent thoughts consciously, to accept them as a natural reaction to their unnormal situation, and at the same time, to not to get too much involved every time anxieties arise. In addition, counselor and family may cautiously anticipate and go through possible responses and options for the case of an actually occuring relapse.

Termination of therapy also requires encouragment of the child to give up their sick role and corresponding privelegies and to promote the social reintegration in previous roles, groups, and obligations.

3.4 Negative Disease Course: Recurrent Relapses, End-stage of Disease, Death, and Dying

The diagnosis of cancer recurrency requires a similar accepting and verbalizing counseling behavior concerning the intense emotional reactions as in the former diagnosis shock. This process may include a clarification of ambivalent tendencies towards continuation or discontinuation of oncological treatment since the parents wish to maintain any chance for cure, on the one hand, and do not want to make the child suffer from a stressful but senseless therapy, on the other. With the worsening of treatment course, the threat of loss of the child is gradually intensified leading to the following counseling topics:

(a) Preparing parents for communication with the child on death and dying. Finding individually appropriate images integrating truth and hope;

(b) Searching for phantasies on a good way to spend the remaining time together; development of other important goals besides survival like e.g., freedom from pain; and

(c) Enhancing imagery on a suitable way of parting in dignity.

If the child actually dies, it is important to say that the therapy failed and not the child or the family. The family's engagement shown for the child should be acknowledged. Individually suitable mourning rituals may be conceived that combine keeping related to the deceased child and, at the same time, letting them go. Family members may be encouraged to take their time for mourning and grief according to their very personal needs not according to expectations of their social network (see *Bereavement*).

Summarized, many counseling interventions focus on the enhancement of open and clear family communication. The mobilization of this key resource for effective coping with the burden of childhood cancer may refer to the intrafamilial exchange of personal needs and concerns, the practical organization of day-to-day routines, or the subtle communication between parents and child about their chances of survival. Effective extrafamilial communication may refer to establishing cooperation and agreements with the staff or to mobilizing and accepting support from the extended social network. If such competencies are not only applied but even refined during the treatment process, the 'cancer experience' may even stimulate processes of individual maturity and family coherence.

4. From a 'Protective' to an Open Approach

There follow some guidelines on how to inform children with cancer about their diagnosis. Some decades ago, a so-called 'protective' approach was preferred which consisted of not telling the child the life-theratening nature of the disease. In the meantime, not only changes in ethical considerations but also the progress in medical treatment options have led to a broad consensus that there is no responsible alternative to telling the truth even to preschool children. Today, medical progress allows the combination of the bad news of cancer diagnosis with the good news of reasonable hope for cure. Therefore, telling children the truth includes two realistic key messages: first, that death has to be expected if no treatment is performed, and second, that even treatment does not necessarily guarantee survival. Information given to the child should be appropriate to their cognitive and emotional age and maturity. In small children, showing concrete treatment procedures and instruments and offering plausible explanations for sensory experiences like vomiting during chemotherapy is more effective than abstract explanations about malignant cells. An approach that avoids telling the truth implies serious medical and psychological 'side-effects' to be considered:

(a) A child will not accept the enormously distressing treatment procedures without the awareness of the serious consequences in case of therapy refusal. Thus, true information is necessary for compliance with treatment, and compliance with treatment is necessary for medical treatment success.

(b) Children who lack plausible information on their condition tend to develop other explanations for the occurrence of their disease (Eiser 1993). A frequent one is the conception that the illness represents a punishment for own misbehavior.

(c) Trust in the parents is one of the most important psychological resources for the child to protect them from feelings of loneliness and depression. A child uncovering the parents' well-meant lies will not only have to struggle with the adversities of disease and treatment but, in addition, with social alienation.

See also: Biopsychology and Health; Cancer: Psychosocial Aspects; Cancer Screening; Childhood Health; Chronic Illness, Psychosocial Coping with; Illness: Dyadic and Collective Coping; Pain, Health Psychology of; Well-being and Health: Proactive Coping

Bibliography

Chesler M A, Barbarin O 1987 *Childhood Cancer and the Family. Meeting the Challenge of Stress and Support*. Brunner/Mazel, New York

Eiser C 1993 *Growing up with a Chronic Disease: The Impact on Children and their Families*. Kingsley, London

Koocher G P, O'Malley J E (eds.) 1981 *The Damocles Syndrome: Psychosocial Consequences of Surviving Childhood Cancer*. McGraw-Hill, New York

Kupst M J, Natta M B, Richardson C C, Schulman J L, Lavigne J V, Das L 1995 Family coping with pediatric leukemia: 10 years after treatment. *Journal of Pediatric Psychology* **20**: 601–17

Noeker M, Petermann F 1990 Treatment-related anxieties in children and adolescents with cancer. *Anxiety Research* **3**: 101–11

Noeker M, Petermann F, Bode U 1990 Family counselling in childhood cancer: Conceptualization and empirical results. In: Schmidt L R, Schwenkmezger P, Weinman J, Maes S (eds.) *Theoretical and Applied Aspects of Health Psychology*, Harwood Academic Publishers, Chur, pp. 241–53

Noll R B, Gartstein M A, Vanatta K, Correl J, Bukowski W M, Davies W H 1999 Social, emotional, and behavioral functioning of children with cancer. *Pediatrics* **103**: 71–8

Patterson J M 1991 Family resilience to the challenge of a child's disability. *Pediatric Annals* **20**: 491–99

F. Petermann and M. Noeker

Childhood Depression

Once thought to be rare or virtually nonexistent, depression in young people has been reliably documented. In this article, the meaning of depression, its clinical manifestation, and the recent history of studies in depression are discussed. Key issues in the study of depression include prevalence and course of depressive disorders in young people, gender and age differences, and comorbidity with other disorders. Topics that are the focus of current research include those familial, cognitive, and biological factors that may contribute to or maintain depression. Effective treatments for childhood or adolescent depression have begun to be identified, but many questions about optimal single or combined treatments for the disorder remain to be answered.

1. *The Meaning of the Term 'Depression' in Young People*

The term 'depression' in children or adolescents can refer to three different, increasingly restrictive definitions. At the least restrictive level, 'depression' refers to a negative or low mood, such as sadness. In medical terms, if the low mood persists beyond an expectable duration of time, this is a single *symptom*. It may occur in isolation or in conjunction with other symptoms. At a more restrictive level, 'depression' refers to a particular set of symptoms that frequently occur together. Such a set of symptoms is designated a *syndrome*. For instance, children with low mood often simultaneously experience boredom, low energy, and social withdrawal, among other possible symptoms. Such a syndrome can be assessed at one point in time by the use of symptom review checklists. At the most restrictive level, 'depression' refers to a psychiatric *disorder*, characterized by a significant and persistent change in the child's functioning. Depressive disorders have, in addition to a number of defining symptoms, a certain duration and course, a level of severity that causes considerable distress, and an impact on the daily functioning of the young person. Assessment of depressive disorders requires a comprehensive clinical interview, possibly accompanied by symptom checklists or rating scales.

Depression as a single symptom will not be a focus of this article. Instead the focus will be primarily on depression as a disorder, and secondarily on depression as a syndrome, since syndromal depression has been the subject of several important longitudinal studies among youths.

Depression is one of the psychiatric disorders in which disturbed mood is the core defining characteristic. Diagnoses of depression (unipolar depression) in the American Psychiatric Association (1994) nomenclature include Major Depressive Disorder and Dysthymic Disorder, both of which refer to conditions in which mood is dysphoric or low. In depression, children may report feeling sad, unhappy, bored or uninterested in usual activities, angry, or irritable, or may appear sad or tearful. By contrast, bipolar mood disorders, or manic-depressive disorders, are conditions in which mood fluctuates episodically between such dysphoric states and unusually elevated, irritable, or expansive mood states.

2. *Recognition and Recent History*

Recognition that depression occurs with considerable frequency in children and adolescents has been a relatively recent development. In the 1950s and 1960s, when psychoanalytic theory provided the primary conceptual model for psychiatric diagnosis and psychological assessment, clinicians rarely diagnosed depression in young people, partly because the development of superego and ego ideal functions, considered necessary to generate and maintain depression, was incomplete in youngsters. Child clinicians were also faced with the reality that their most common referral problems involved behavioral disorders or school performance problems, rather than mood problems. By the 1970s, there was recognition that depression might be 'underlying' such problems, but it was not until the emphasis on phenomenological or symptom-focused diagnosis that depression began to be assessed systematically in young people. Lewinsohn

et al. (1993) have proposed that the study of affective disorders in children and adolescents really began in the 1970s when several sets of investigators, including Carlson, Cytryn, Kovacs, Poznanski, Puig-Antich, Rutter, and their colleagues demonstrated that such disorders do occur and can be reliably assessed in young people. This development corresponded to the shift in psychiatric assessment toward systematic review of presenting symptoms, and away from nondirective, inferential assessment based on play observations or patient narrative reporting.

Since the advent of symptom-driven diagnosis in US psychiatry in 1980, the same diagnostic criteria used to diagnose depressive disorders in adults have been applied to children and adolescents. At present, the diagnosis of Major Depressive Disorder (MDD) requires at least one episode in which the child has had five or more of the following symptoms, including one of the first two, for a minimum of two weeks: (a) depressed or irritable mood; (b) markedly diminished interest or pleasure in activities; (c) weight or appetite loss or gain; (d) insomnia or hypersomnia; (e) psychomotor agitation or retardation; (f) fatigue or loss of energy; (g) feelings of worthlessness or excessive guilt; (h) decreased ability to think, concentrate, or make decisions; (i) recurrent thoughts of death or suicide or a suicide attempt or plan. The diagnosis of Dysthymic Disorder (DD) is given if depressed or irritable mood is present most days for a year or more, and if mood disturbance is accompanied by two or more of six key symptoms: poor appetite or excessive eating, insomnia or hypersomnia, low energy or fatigue, low self-esteem, poor concentration or difficulty making decisions, and feelings of hopelessness. MDD and DD are not mutually exclusive, and some children or adolescents present with a long-standing DD, upon which an episode of MDD has been superimposed: a condition referred to as 'double depression.'

Parallel to the application of adult criteria to childhood depression, there have been numerous studies in the psychopathology of depression, and in its treatment, which have applied concepts and research models from the adult field and extended them to children. To give just two examples, both cognitive factors known to characterize adult depression (e.g., distortions in thinking) and cognitive therapy, which is known to be effective for adult depression, have been studied in young people. Although the use of adult diagnostic criteria for child and adolescent depression is now conventional in research and clinical practice, there continues to be controversy regarding the developmental sensitivity and adequacy of this approach. In particular, the school of thought known as developmental psychopathology (Cichetti et al. 1994) is characterized by an emphasis on understanding the multiple contributions of developmental sciences toward understanding both normal development and disorders such as depression. Merely applying the same criteria and research concepts from adults to children does not adequately account for the interactive contributions of cognitive development, family processes, school environment, and biological development to the development of mood disorders. Developmental psychopathologists seek to go beyond merely establishing that depressed children differ from nondepressed children on a set of symptoms and associated features, and to determine the conditions and processes that contribute to these differences.

3. *Prevalence and Gender Differences*

Prevalence of depression in children and adolescents varies across studies, in part depending on whether a syndrome (elevated scores on a continuous scale) or a diagnosed disorder (based on parent and/or child interviews) serves as the measure of depression. Fleming and Offord's (1990) review of studies showed that MDD in preadolescent children occurred in as low as 0.4 percent and as high as 2.5 percent of the samples. Upper estimates for adolescents were higher, with a range from 0.4 percent to 6.4 percent. Prevalence of DD in children ranged from 0.6 percent to 1.7 percent, and in adolescents from 1.6 percent to 8.0 percent. The subsequent Oregon Adolescent Depression Project (Lewinsohn et al. 1993) found a point prevalence of 2.57 percent for MDD and 0.53 percent for DD, and lifetime prevalence of 18.48 percent for MDD and 3.22 percent for DD in US high school students (mean age = 16.6 years).

Depression is associated with age and with gender (Birmaher et al. 1996a). Rates are higher in adolescents than in children. Phenomenology also differs to some extent by age, with adolescents with MDD more likely than children to have anhedonia, hypersomnia, weight change, or lethal suicide attempts. Among depressed children there is equal gender representation, but in adolescents the ratio is about two females to one male, similar to the pattern among adults. The reasons for gender differences in adolescent depression are the focus of considerable current research. Sociocultural pressures on girls, biological changes associated with puberty, and sex differences in cognitive coping mechanisms have been proposed as possible explanations.

4. *Course of Depressive Disorders in Childhood*

Kovacs and her colleagues were the first to study the course of depressive disorders in prepubertal children. Their sample consisted of children who had been referred for clinical services, and was not an epidemiological sample. The average (mean) duration of the index episode of MDD in their sample was 32 weeks and half had recovered by nine months (median duration). For DD, however, the median duration of

the episode was four years. Both MDD and DD children demonstrated a high likelihood of having a second depressive episode within a nine-year follow-up period, with the risk higher among DD than among MDD children (Kovacs 1996). By contrast, children diagnosed with an Adjustment Disorder with depressed mood (one or a few depressive symptoms in reaction to a stressor) were not at risk of developing MDD during follow up.

Data from the Oregon Adolescent Depression Project (Lewinsohn et al. 1993), with an epidemiological sample, indicated that there is an extraordinarily variable duration of MDD in teenagers. Mean episode duration was 26 weeks, but median was eight weeks, with a range of 2 to 520 weeks. Seventy-five percent of the adolescents recovered by 24 weeks. Overall, about 33 percent of the adolescents who recovered from their initial episode of MDD had a second episode within four years. Adolescents with onset of MDD before age 15.5 years had longer episodes and shorter times between recovery and relapse than did adolescents with later onset.

Using syndromal measures of depression assessed on rating scales or clinician-completed symptom checklists, studies both in the USA and in the UK have demonstrated that depression during the developmental period confers an increased risk of depression during adulthood. Harrington and his colleagues (Harrington et al. 1990) have shown that depressed young people followed up an average of 18 years into adulthood have nearly a fourfold increased risk of adult depression when compared to a control group matched for nondepressive childhood symptoms.

5. *Comorbidity and the Risk of Developing Bipolar Disorder*

Studies from New Zealand, Puerto Rico, and the continental USA reviewed by Angold and Costello (1993) have shown that depression (either MDD or DD) in young people is very often accompanied by other disorders. The most common comorbid conditions in these studies were oppositional or conduct disorders and the various anxiety disorders. Rates of conduct or oppositional disorders in depressed children or adolescents ranged widely across studies, from 20 percent to 80 percent. Most studies showed rates of anxiety disorders between 30 percent and 50 percent, but some showed rates exceeding 70 percent. In the Oregon project, Lewinsohn and colleagues (1993) found that the most frequent comorbid diagnoses for depressed adolescents were anxiety disorder (21 percent) and substance use disorder (20 percent), with 12.4 percent of depressed teenagers having a disruptive behavior disorder.

Birmaher and colleagues (Birmaher et al. 1996a), in their review of the comorbidity literature, found that, except for substance use disorders, most of the other comorbid conditions developed before the MDD. However, conduct disorder was sometimes found to develop after MDD and to persist after resolution of the depression. Comorbid MDD and DD, in particular, have been found to predict longer depressive episodes, more suicidality, and worse social adjustment. The presence of comorbid anxiety disorders appears to raise the risk of suicidality and of substance abuse, and to be associated with poorer response to psychotherapy. There has been particular interest in comorbid depression and disruptive behavior disorders, perhaps because these represent a combination of two distinct types of disorders, internalizing and externalizing, that would not ordinarily be expected to co-occur. Birmaher and colleagues' (1996a) review indicates that these young people are at increased risk for suicide attempts and adult criminality, and have poorer response to acute treatment, but also have more positive responses to placebo treatment and have fewer depressive recurrences.

Bipolar disorder includes distinct periods of depression and of mania. Because bipolar disorder is well recognized as distinct from unipolar depression, with a different course and different treatment requirements, it is important to assess the risk of developing bipolar disorder in depressed young people. Follow-up studies of clinically referred youths have varied in duration of follow-up period and in definition of bipolar disorder. They suggest that about 20 percent of adolescents with MDD go on to develop bipolar disorder, but the range of estimates to date varies widely. Factors that predict bipolar outcome include psychotic symptoms during the depressive episode, family history of bipolar disorder, and hypomanic reactions to antidepressant medication. There is conflicting evidence regarding the predictive utility of acute vs. prolonged onset of depression in predicting bipolar outcome. Both acute onset of severe depression in hospitalized adolescents and early onset of DD in (nonhospitalized) children have been associated with later bipolar outcome, suggesting there may be more than one path to a bipolar outcome.

6. *Current Research Emphases: Psychosocial Correlates, Biology, and Treatment*

Psychosocial correlates, biological correlates, and treatment efficacy and effectiveness are three areas of current research emphasis in childhood and adolescent depression. Correlates include an array of factors that are associated with and may make a causal contribution to depression or maintaining a depressive episode. Some psychosocial correlates, such as family factors and cognitive factors, can become targets of treatment in psychosocial intervention.

Compared to controls, the families of depressed young people are characterized by higher levels of parent–child and/or marital conflict, poor parent–

child communication, and more distant, less affectionate parent–child relationships (Birmaher et al. 1996a). The challenge for researchers is to determine whether these factors are specific to depression, and if so, whether they are causal.

A number of cognitive factors has been associated with depression in youths. These include cognitive distortions that emphasize negative interpretations of events, a tendency to attribute negative events to enduring and internal causes but positive events to transitory and external causes, low self-esteem, and low estimates of personal control and competence. Although such factors are present while youngsters are depressed, it is not clear whether they represent pre-existing risk factors or state-dependent correlates of the depressive episode. An important research question is to delineate the processes through which children acquire such cognitive characteristics.

Biological correlates of depression include family-genetic factors and markers of a depressed state. Children of depressed parents are more likely to develop depression than are children of nondepressed parents. Studies in adults show a significant genetic component to the transmission of mood disorders, and family interaction studies suggest that depressed parents have deficits in parenting behaviors that raise the risk of poor adaptation in their children. Both genetics and experience, therefore, are likely involved in cross-generational transmission of depression.

A number of potential biological markers of MDD have been investigated, in the desire to clarify the biological basis of the disorder. Among these are secretion of growth hormone after pharmacological challenges, abnormalities in functioning of the hypothalamic–pituitary–adrenal axis, and abnormal sleep EEG patterns. Results to date do not suggest any single marker sufficiently sensitive to or specific to childhood or adolescent MDD that can be used for diagnostic purposes.

Effective treatment is a major current focus of research. Cognitive behavior therapy (CBT) involves working with the young person to understand and to modify thoughts and behaviors that are likely contributing to depression (see *Cognitive Therapy*). Interpersonal psychotherapy (IPT) involves working with the young person to understand the impact of relationship or role conflicts on depression and to modify interpersonal patterns (see *Interpersonal Psychotherapy*). There is considerable evidence to support the efficacy of CBT both for childhood and for adolescent depression, when compared to control conditions such as a waiting list (Birmaher et al. 1996b, Reinecke et al. 1998). Most of the child study subjects have been children with a depressive syndrome, and some of the interventions have been school-based. More of the adolescent studies have involved teenagers with a depressive disorder and have been clinic-based. There have been very few studies to date comparing CBT to other active treatments. IPT has also proven more effective than clinical monitoring for depressed adolescents, both in terms of symptom reduction and in terms of improved social functioning (Mufson et al. 1999).

Studies of older tricyclic antidepressant medications did not demonstrate efficacy in children or adolescents. A recent study by Emslie and his colleagues (Emslie et al. 1997) using fluoxetine, a selective serotonin reuptake inhibitor, did show a significantly better outcome for those on medication than for those receiving a placebo. As is true in the psychosocial treatment arena, research on medication treatment has been limited to acute or short-term treatment.

Numerous critical questions remain to be investigated. These include the relative efficacy of various psychosocial, medication, and combined treatments, the long-term efficacy of such treatments, optimal duration of treatment, and effectiveness of university- or laboratory-based interventions in the broader clinical world. It is not yet clear whether individuals who fail to respond to a first treatment will do better with an alternative. Finally, a major challenge to treatment research is to identify strategies and methods to address the comorbid conditions that so often accompany depressive disorders in young people.

See also: Anxiety Disorder in Children; Behavior Therapy with Children; Child and Adolescent Psychiatry, Principles of; Childhood Health; Depression; Depression, Clinical Psychology of; Depression, Hopelessness, Optimism, and Health; Early Childhood: Socioemotional Risks; Infancy and Childhood: Emotional Development

Bibliography

American Psychiatric Association (APA) 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. APA, Washington, DC

Angold A, Costello E 1993 Depressive comorbidity in children and adolescents. *American Journal of Psychiatry* **150**: 1779–91

Birmaher B, Ryan N, Williamson D, Brent D, Kaufman J 1996b Childhood and adolescent depression (II). *Journal of the American Academy of Child and Adolescent Psychiatry* **35**: 1575–83

Birmaher B, Ryan N, Williamson D, Brent D, Kaufman J, Dahl R, Perel J, Nelson B 1996a Childhood and adolescent depression (I). *Journal of the American Academy of Child and Adolescent Psychiatry* **35**: 1427–39

Cicchetti D, Rogosch F, Toth S 1994 A developmental psychopathology perspective on depression in children and adolescents. In: Reynolds W, Johnston H (eds.) *Handbook of Depression in Children and Adolescents*. Plenum, New York pp. 123–41

Emslie G, Rush J, Weinberg W, Kowatch R, Hughes R, Carroll W, Carmody T, Rintelmann J 1997 A double-blind, randomized, placebo-controlled trial of fluoxetine in children and adolescents with depression. *Archives of General Psychiatry* **54**: 1031–7

Fleming J E, Offord D R 1990 Epidemiology of childhood depressive disorders—a critical review. *Journal of the American Academy of Child and Adolescent Psychiatry* **29**: 571–80
Harrington R, Fudge H, Rutter M, Pickles A, Hill J 1990 Adult outcomes of childhood and adolescent depression. *Archives of General Psychiatry* **47**: 465–73
Kovacs M 1996 The course of childhood-onset depressive disorders. *Psychiatric Annals* **26**: 326–30
Lewinsohn P, Hops H, Roberts R, Seeley J, Andrews J 1993 Adolescent psychopathology. *Journal of Abnormal Psychology* **102**: 133–44
Mufson L, Weissman M, Moreau D, Garfinkel R 1999 Efficacy of interpersonal psychotherapy for depressed adolescents. *Archives of General Psychiatry* **57**: 573–9
Reinecke M A, Ryan N, DuBois D 1998 Cognitive–behavioral therapy of depression and depressive symptoms during adolescence. *Journal of the American Academy of Child and Adolescent Psychiatry* **37**: 26–34
Reynolds W, Johnston H (eds.) 1994 *Handbook of Depression in Children and Adolescents*. Plenum, New York

J. F. Curry

Childhood Health

Child health psychology is an interdisciplinary field concerned with the physical, cognitive, social, and emotional functioning and development as they relate to health and illness issues in children, adolescents, and their families. It has emerged as a separate discipline, in recognition of the unique ways in which children are affected by illness compared with adults.

There are several reasons for establishing child health psychology as a separate discipline from that of adults. First, children's understanding of health, and the causes and implications of illness, differ from those of adults. Second, children's health is important both for understanding the current health of the child and in terms of implications for adult health. Children who adopt unhealthy lifestyles tend to become adults with unhealthy lifestyles. Third, diseases that affect children may have different implications for daily life compared with conditions that affect adults. Thus, children with diabetes need insulin on a daily basis, whereas many adults may be treated by diet or medication alone. Untreated, cancer in children is rapidly fatal, but may be more chronic in adults. Fourth, the impact of childhood illness is not restricted to the individual but affects the whole family. Parents, and doctors, make decisions on behalf of children, often with little clear understanding of the child's preferences.

Child health psychology is concerned with the relationship between psychological and physical well-being. The scope of child health psychology is generally considered to include the following:

(a) psychosocial contributions to outcomes in pediatric conditions;

(b) assessment and treatment of behavioral and emotional consequences of disease;

(c) role of psychology in health care settings;

(d) promotion of health and health-related behaviors;

(e) prevention of illness and injury among children and adolescents; and

(f) training.

In reviewing the achievements of child health psychology, this initial conceptualization has been followed, though it is immediately apparent that some areas have been the focus of much greater research activity than others.

1. Psychosocial Contributions to Outcomes in Pediatric Conditions

It has been argued that chronic illness challenges the child's normal development by limiting opportunities, restricting play and activities, the attainment of autonomy, and potentially compromising family and peer relationships. These effects may differ specifically as a function of the child's age. For infants, chronic illness is most likely to affect parent–child relationships, restrict mobility, or limit opportunities to socialize with peers. Separation from parents is a key issue. For older children, the impact will have a more direct effect in terms of reduced schooling, compromised peer relationships, more time with adults, concerns about body image, and awareness of vulnerability and possible death. Cancer in adolescence may extend a period of dependency on parents and reduce opportunities to establish close interpersonal relationships with the opposite sex.

Clinically, the study of how development proceeds despite such adversity is justified in terms of increasing our understanding of normal development, offering insights where problems occur and improving clinical services. The identification of protective factors, such as parenting, is also important. In some cases, family variables may be better indicators of outcome than illness predictors. Thus, Carlson-Green et al. (1995) studied 63 children who had been treated for a brain tumor. The best predictors of children's behavior problems were family and demographic variables, while the best predictors of achievement were illness and demographic variables. The implications are that outcome following traumatic illness and treatment may be very much influenced by positive family variables. Identification of factors that moderate outcomes is of central interest in child health psychology.

2. Assessment and Treatment of Behavioral and Emotional Consequences of Disease

Measurement of psychosocial adjustment has been central to research in this area. In that a central task of childhood is to develop toward an autonomous, healthy, and well-functioning adult, adjustment has

been defined in developmental-normative terms. 'Good adjustment, then, is reflected as behavior that is age-appropriate, normative and healthy, and that follows a trajectory toward positive adult functioning. Maladjustment is mainly evidenced in behavior that is inappropriate for the particular age, especially when this behavior is qualitatively pathological or clinical in nature' (Wallander and Thompson 1995 pp. 125–6).

While the basic question concerning how illness and treatment affects the child's normal development is a very real one, the issue of measurement is complex. Standardized measures that capture the kind of experiences described so graphically by parents in interview studies are not available.

Measurement of adjustment has often been based on parents' reports and most commonly is assessed using the Child Behavior Checklist (Achenbach 1991). This measure has come under considerable criticism, not least because it lacks sensitivity for work with sick children (Perrin et al. 1991). Many of the items assess somatic symptoms. As a consequence, children with cancer (or any physical illness) inevitably have higher scores than healthy children (see *Childhood Cancer: Psychological Aspects*).

Improvements in the management of a number of life threatening conditions, for example, cancer and cystic fibrosis, have led to increased survival. However, this has been achieved at the cost of often daily, frequently painful, and certainly intrusive treatment regimens, which may continue for many years. Survival alone is therefore no longer perceived to be an adequate outcome measure. A central concern must be the extent to which treatment compromises quality of life (QoL). This change in emphasis has necessitated the development of alternative means of assessing treatment outcomes. Clinicians and researchers have therefore turned to measures of QoL to provide a more comprehensive account of individual patient experience as well as information on which to base decisions about service provision. Acceptance of this broader perspective on outcomes has led to a recognition of the need to identify methods of medical management, which impose fewer restrictions on quality of life.

Quality of life is a difficult concept to define and an even more difficult one to measure. Critical to most definitions is the notion that an individual's perception of their QoL is unique, and it therefore follows that efforts must be made to elicit information directly from the individual. Proxy ratings, as provided by caregivers, are useful, but not a substitute for an individual's own information.

Only some 3 percent of trials in pediatric oncology include a measure of QoL. There are few examples of their use in intervention studies (Kazak et al. 1996). To a large extent, this can be attributed to reticence on the part of clinicians to use the measures. Objections are generally based on perceptions of poor psychometric properties. In addition, many measures lack face validity; they do not appear to tap issues of relevance to the children under study. Many measures are criticized as too long, too repetitive, or simply inappropriate.

It would be appropriate to conclude this section with some discussion of the increasing use of qualitative measures. These are preferred by some, while others acknowledge the need for qualitative measures in combination with the more conventional quantitative measures. Qualitative methods can be used in the development of quantitative measures, ensuring that the latter include issues of importance to the target population. Qualitative methods may also be the method of choice in situations involving sensitive issues or for work with very young or handicapped children (Eiser and Twamley 1999).

3. *Intervention Programs*

Interventions have been reported to facilitate return to school following long illness. Varni et al. (1993), for example, reported a social skills program to teach children returning to school following diagnosis with cancer how to deal with other children's questions. A number of methods were used (e.g. role-play) to prepare children for teasing and allow them to develop the skills to know what to say and save face.

Other programs are aimed at improving self-care. Holzheimer et al. (1995), for example, used a teaching video, instruction booklet, and opportunities to rehearse the desired skills, and reported improved use of an inhaler in preschool children with asthma.

Efforts to treat procedure-related pain have been reported most frequently in pediatric oncology (see *Childhood Cancer: Psychological Aspects*). Children with cancer undergo regular painful procedures (e.g., lumbar punctures and bone marrow aspirations). The earliest reports by Jay et al. (1986) described interventions based on cognitive behavior therapy, including breathing exercises and use of imagery. Filmed modeling has also been used. Although the early reports were often based on a small number of case studies, more recent work has involved randomized clinical trials. Kazak et al. (1996) reported that a combined physiological and psychological intervention was more successful than a physiological intervention alone.

Interventions for disease-related pain for children remain poorly developed, despite the very high levels of pain experienced by children with conditions such as cancer, arthritis, hemophilia, and sickle cell disease. Interventions need to target both disease-related pain and the child's perceptions of pain. Walco et al. (1999) conclude that more work is needed which documents the emergence of pain experiences and coping over time. Interventions need to be based on a variety of approaches, and target the pain experience within a social and family context.

Schools based programs for children with headaches may be particularly successful. Larsson and Carlsson (1996) randomized 26 children aged 10–15 years with chronic tension type headache to a nurse administered relaxation training intervention or no treatment control. The five-week program involved twice weekly sessions, each lasting approximately 20 minutes. Headache activity was reduced in children in the intervention group at post-treatment and six-month follow-up compared with the no treatment controls.

4. *The Promotion of Health and Health-related Behaviors*

Drug and alcohol use among teenagers and adolescents usually occurs in group situations, and the use of such substances has been found to be strongly motivated by self-presentational desires such as the need for social approval and peer acceptance. Family factors, including parental modeling (often parents themselves are heavy users of the substance), and the family–child relationship, are key risk factors for subsequent substance use. In addition, peer relationships appear to be strongly related to the extent of alcohol and illicit drug use. Both types of substance abuse seem to be more common among those who report academic difficulty or poor adjustment at school. Other reasons for the illicit use of alcohol and drugs include the wish to convey autonomy or rebelliousness, and the reduction of anxiety in intergroup settings. Some may use alcohol or drugs if they feel unable to cope with the pressures associated with growing up; as a sort of temporary escape from school problems, social anxieties, or home difficulties.

5. *Issues Relating to Training*

The nature of much of the work with children involves collaboration with other professionals: clinicians, teachers, and social services. Issues of training particularly need to address the potential obstacles to successful collaboration. There is inevitably some friction between psychologists and pediatricians in the way in which they approach research and clinical issues concerned with chronically sick children. While both share some ideals about what should be done to help children with chronic conditions and their families, their training and experience does not necessarily mean that they share the same research agenda.

6. *Future Directions*

Increasing recognition of the closeness between physical and psychological health has contributed to an increasing visibility of psychological factors in pediatric medicine. So too have changes in treatment. In reality, caring for the child with chronic illness is not always about achieving a cure, but should be about enhancing the QoL. As such, there is pressure to include measures of QoL as an integral part of the evaluation of clinical trials. The development of disease-specific QoL scales means that health outcomes can now be considered from a more holistic point of view. QoL measures also have a role in evaluating innovative treatments. Children undergoing bone marrow transplants necessarily experience isolation, pain, and lengthy hospitalizations, and assessments of their QoL during treatment and after are essential.

Given the increasing costs of health care, it is critical that treatments are evaluated in terms of both physical and psychological well-being. Thus, evaluations of the success or otherwise of growth hormone therapy must not be based on height alone, but consideration also needs to be given to the child's psychological well-being.

There remains scope for considerable improvements in measures available. Current work rarely draws on expertise in developmental psychology to direct the format or content of measures. There is an emerging literature documenting children's language, memory, and emotional development, much of which could be used to guide new measures. It is also important that measures are based on a theoretical framework. In the past, theory has been sparse or embedded in a model of maladjustment. It is likely that theories of normal child development will prove more useful in the future.

Despite the criticisms that can be made, there is no doubt that the establishment of a child health psychology in itself has done much to increase the profile of work with children. The difficulties involved may be considerable, but these also contribute to its attractiveness. That children have a unique perspective on life, despite the hardships of a serious illness, is without debate. From such a perspective, and for many people, work with children is therefore not only humbling, but intrinsically more satisfying, than work with other groups.

See also: Adolescent Health and Health Behaviors; Child Care and Child Development; Childhood Depression

Bibliography

Achenbach T M 1991 *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Department of Psychiatry, University of Vermont, Burlington, VT

Bradlyn A S, Ritchey A K, Harris C V, Moore A K, O'Brien R T, Parsons S K, Pattersan K, Pollock B H 1995 Quality of life research in pediatric oncology: Research methods and barriers. *Cancer* **78**: 1333–9

Carlson-Green B, Morris R D, Krawiecki N 1995 Family and illness predictors of outcome in pediatric brain tumors. *Journal of Pediatric Psychology* **20**: 769–84

Eiser C, Cotter I, Oades P, Seamark D, Smith R 1999 Health-related quality of life measures for children. *International Journal of Cancer* **512**: 87–90
Eiser C, Twamley S 1999 Talking to children about health and illness. In: Murray M, Chamberlain K (eds.) *Qualitative Health Psychology: Theories and Methods*. Sage, London
Holzheimer L, Mohay H, Masters I B 1998 Educating young children about asthma: Comparing the effectiveness of a developmentally appropriate asthma education videotape and picture book. *Child: Care, Health and Development* **24**: 85–99
Jay S M, Elliott C, Varni J W 1986 Acute and chronic pain in adults and children with cancer. *Journal of Consulting and Clinical Psychology* **54**: 601–7
Juniper E F, Guyatt G H, Feeny D H, Ferrie P J, Griffith L E, Townsend M 1996 Measuring quality of life in children with asthma. *Quality of Life Research* **5**: 35–46
Kazak A E, Penati B, Boyer B A, Himelstein B, Brophy P, Waibel K, Blackall G F, Daller R, Johnson K 1996 A randomized controlled prospective outcome study of a psychological and intervention pharmaceutical protocol for procedural distress in pediatric leukemia. *Journal of Pediatric Psychology* **21**: 615–32
Landgraf J M, Abetz L, Ware J E 1996 *Child Health Questionnaire (CHQ): A Users Manual*. The Health Institute, New England Medical Centre
Larsson B, Carlsson J 1996 A school-based, nurse administered relaxation training for children with chronic tension-type headache. *Jouranl of Pediatric Psychology* **21**: 603–14
Perrin E C, Stein R E K, Drotar D 1991 Cautions in using the Child Behavior Checklist: Observations based on research about children with a chronic illness. *Journal of Pediatric Psychology* **16**: 411–21
Peterson L, Oliver K K, Brazeal T J, Bull C A 1996 A developmental exploration of expectations for and beliefs about preventing bicycle collision injuries. *Journal of Pediatric Psychology* **20**: 13–22
Varni J W, Katz E R, Colegrove R, Dolgin M 1993 The impact of social skills training on the adjustment of children with newly diagnosed cancer. *Journal of Pediatric Psychology* **18**: 751–67
Walco G A, Sterling C M, Conte P M, Engel R G 1999 Empirically supported treatments in pediatric psychology: Disease related pain. *Journal of Pediatric Psychology* **24**: 155–67
Wallander J L, Thompson R J 1995 Psychosocial adjustment of children with chronic physical conditions. In: Roberts M C (ed.) *Handbook of Pediatric Psychology*. Guilford Press, New York, Vol. 2

C. Eiser

Childhood Sexual Abuse and Risk for Adult Psychopathology

1. *Overview of the Issue*

Numerous studies on the prevalence, characteristics, and consequences of childhood sexual abuse (CSA) have been conducted since the 1970s, but deriving conclusions about this literature has been hampered by the wide variability in definitions and methods used across studies. Nonetheless, enough research has accumulated from both clinical and general population studies to support the claim that CSA is significantly *correlated* with increased risk for various forms of adult psychopathology. There has been disagreement, however, about whether CSA *causes* adult psychopathology. Since the 1990s, a few studies have been conducted that allow for causal interpretations about the effects of CSA on adult psychopathology.

The purpose of this article is to describe the historical and scientific development of CSA research, from early clinical studies to current studies of the general population. Definitional and methodological issues in CSA research will be discussed, and a review of recent studies that have used powerful methodologies for understanding cause-and-effect relationships will be presented. Finally, some general conclusions about the effects of CSA on adult psychopathology will be provided, and future directions for this research will be suggested.

2. *The Historical Development of Research on Childhood Sexual Abuse*

The evolution of CSA research, as discussed below, has followed a three-phase course that roughly parallels the progression of research found in all psychiatric epidemiological research: (a) pioneering publications and numerous clinical studies; (b) studies using non-clinical (general population) samples to estimate prevalence and correlates of CSA; and (c) studies using scientifically rigorous methods that allow for cause-and-effect conclusions about CSA and adult psychopathology.

2.1 *Phase 1 Research: Pioneering Publications and Clinical Research*

Modern awareness of the prevalence, characteristics, and possible consequences of CSA in the USA can be traced to the late 1970s and early 1980s. During this first phase of modern CSA research, a few pioneering publications, most notably Finkelhor's *Sexually Victimized Children* in 1979 and Russell's *Sexual Exploitation*: *Rape*, *Child Sexual Abuse*, *Sexual Harassment* in 1984 provided data that indicated CSA occurred frequently to children and adolescents in the USA. Following these works, public attention to CSA was further stimulated by several highly publicized media accounts of CSA, and clinical research increased dramatically. Results from early clinical studies were mostly of women, and indicated a high percentage of women reporting CSA, with many women indicating that the abuse involved family members and/or serious abuse (i.e., intercourse CSA). These studies also found that a history of CSA was associated with elevated rates of several psychological problems, including substance abuse, anxiety dis-

orders, depression, self-injurious behaviors, and interpersonal functioning deficits. As a result of this increased attention to CSA, dramatic changes in the reporting and verification of CSA occurred in the USA: from 1976 to 1993, there was a 25-fold increase in verified cases of CSA. Although much of the original research on CSA in the 1970s and 1980s was centered on the USA, other countries, including Canada, the UK, and several European nations began or increased research on CSA.

2.2 Phase 2 Research: General Population Studies of Prevalence and Correlates of CSA

The second phase of CSA research in the early 1980s addressed a methodological limitation of earlier studies i.e., that most early studies were based on samples of children who were identified as being abused (forensic samples) or samples of adults who were in treatment (clinical samples). These studies were limited because results from forensic and clinical samples may differ significantly from results found among persons who did not report the abuse or who were not in treatment. Numerous nonclinical studies used college samples, but again, women and men in college who report a history of CSA may not generalize to the adult general population for many reasons, including that college students may over-represent socioeconomically advantaged persons and under-represent psychologically impaired persons. Some researchers have also argued that first-year college students may have very recently experienced CSA, and long-term consequences have not yet occurred.

Forensic, clinical, and college samples studies are quite important for developing a base of research to build upon, but the most scientifically rigorous sampling design is one that uses a random sample from the general population. Random sampling indicates that every person in an identified community or even an entire country has an equal chance of being selected for the study (although random samples may exclude certain age groups or persons living in institutions).

In CSA research, the basic methodology of a general population study is to ask adult participants if they experienced any sexually abusive acts during childhood (the adult retrospective recall method). Researchers also often ask participants to report the characteristics of the abuse (e.g., who was the abuser, the ages of the abuser and participant), any known effects of the abuse, and/or any current psychological problems they may be experiencing.

Over 20 adult retrospective studies with general population samples were conducted in North America in the last two decades of the twentieth century, and prevalence rates varied dramatically across studies, with about 2–62 percent in women and 3–16 percent in men. Sources of this variability include differences in (a) definitions of CSA; (b) geographical region of sample; (c) methods of data collection (telephone, mailed, or face-to-face interviews); (d) response rates; and (e) the number of questions used to ask about CSA. Definitions of CSA have varied across all studies to date, primarily in four dimensions: (a) the ages used to delimit childhood (most use 18 or 16); (b) whether or not the respondent is asked to self-define the experience as 'abusive' (e.g., asking if the experience was unwanted, the result of force or trickery, and/or was considered abusive); (c) whether or not differences in age between the respondent and the perpetrator are used to define CSA; and (d) the types of sexual activities asked about (e.g., contact only acts vs. both contact and noncontact acts).

Different definitions of CSA and the number of screening questions used most likely contribute to most of the variability across studies. Definitions that use older age cutoffs, and include both contact and noncontact acts produce higher prevalence rates. In addition, surveys that use two or more screening questions for CSA typically result in higher prevalence rates compared to surveys using only a single question. For a thorough review of how different definitions and methods influence prevalence rates of CSA, see Finkelhor (1994). In this report, Finkelhor reviews the bulk of studies to date, and concludes that about 20 percent of North American women and about 5–10 percent of North American men report a history of CSA experiences that include both contact and noncontact acts. The 20 percent estimate for women is consistent with a recent nationally representative study of women in the USA that used multiple screening items for CSA and face-to-face interviews (Vogeltanz et al. 1999). Prevalence estimates from other countries, including Australia, the UK, The Netherlands, New Zealand, Spain, Sweden, and Switzerland, appear to be fairly consistent with the North American estimates, suggesting that CSA occurs at fairly consistent rates in European and English-speaking countries.

Some of these earlier general population studies measured possible consequences of CSA, but most varied significantly in the types of psychological problems assessed and the instruments used. For example, many studies used unstandardized questions about the occurrence of numerous psychological symptoms, others used standardized questionnaires and psychological symptom checklists, and a very few used structured clinical interviews that allowed the researcher to make a clinical diagnosis for several psychological disorders. Although results from these studies varied, the majority, like earlier clinical studies, reported significant relationships between adult women's reports of experiencing CSA and their greater likelihood of experiencing current symptoms of substance abuse, depression, anxiety, sexual functioning problems, eating problems, and self-injurious behaviors. A large limitation of these studies, as with earlier clinical studies, was the relative lack of inclusion of men. Researchers preliminarily concluded that men

with a history of CSA were also reporting elevated rates of various psychological problems.

In addition to reporting on the simple (bivariate) relationships between CSA and adult psychological problems in the general population, a few researchers looked at whether some characteristics of CSA might be more predictive of adult psychological problems. This preliminary data suggested that frequent and/or severe abuse, and abuse by a biological parent have been modestly, but consistently linked to a greater likelihood of problems in adulthood.

By the 1990s, several comprehensive reviews of CSA research had been published, with most authors claiming that CSA was correlated with numerous psychological problems, symptoms, or disorders. However, the majority of studies on CSA did not use methodologies that could determine if CSA *caused* adult problems, and several researchers challenged this causal assumption. The basic argument was that many persons who report a history of CSA also report that their childhood family environments were pathologic, including such problems as conflicts with parent(s), physical or emotional abuse, lack of parental warmth or caring, living with only one biological parent, and parental psychopathology. Therefore, it may be these family environment factors are the causes of adult psychopathology as well as a risk factor for the occurrence of CSA. For an example of a prominent debate on this issue, see Nash et al. (1998). As a result of these challenges, it became incumbent upon CSA researchers to address this issue in future research.

To address cause-and-effect issues in epidemiological research, researchers must use either prospective designs or rigorous multivariate studies (see a description below on these methods). This final phase of research is complex and very costly, and only a few studies to date have been rigorous enough to allow for cause-and-effect interpretations. A summary of this work is described below.

2.3 Phase 3 Research: Studies Examining the Causal Status of CSA on Adult Psychopathology

To date, only a few studies have been able to test adequately for a causal relationship between CSA and adult psychopathology. Although these studies vary in their definitions of CSA and measures of psychopathology, they are similar in three important ways. First, all the studies had large general population random samples of adults, or in one case, a nationally representative random sample of 10- to 16-year-olds. Second, all of the studies used sophisticated methodologies that allowed for some level of causal interpretation. These methods were of two types: a prospective design, in which researchers collect multiple measures from individuals over time, therefore capturing a more accurate account of what happened before and after the occurrence of CSA; or cross-sectional multivariate designs, in which data about early family environment, the occurrence of CSA, and adult psychopathology are gathered at one point in time. Two twin studies using a cross-sectional multivariate design are also included here because the twin study provides unique information about possible genetic contributions to risk for CSA and highly shared environments. Although cross-sectional multivariate designs cannot definitively test for causal relationships, rigorous multivariate methodologies provide a high degree of confidence about causality. Finally, all of the studies used structured diagnostic interviews allowing for clinical diagnosis of psychological disorders. Although standardized questionnaires may also provide an important way of measuring levels of psychological symptoms or distress in the general population, a diagnosis provides a clear statement that the individuals being interviewed have experienced significant impairment in their lives. Diagnoses also provide information about lifetime occurrence of disorders, whereas symptom checklists typically assess only for current levels of functioning.

The studies that meet the above criteria come from the USA, Australia, and New Zealand. In the first of two general population prospective studies, Fergusson and colleagues (1996) made yearly evaluations of a New Zealand sample of over 1,000 children from birth to age 18. At age 18, CSA before age 16 and diagnosable psychological disorders were measured. Results were that CSA significantly predicted major depression, anxiety disorder, conduct disorder, substance use disorder, and suicidal behaviors after controlling for prospectively measured family environment factors. The highest risk for having a psychological disorder was found among individuals reporting CSA intercourse experiences.

Boney-McCoy and Finkelhor (1996) used a prospective design in which they interviewed a nationally representative sample of children, aged 10–16, at two times, approximately 15 months apart. Results were that sexual assaults occurring to the children after the Time 1 interview predicted Time 2 diagnoses of major depression after controlling for previous lifetime occurrence of depression, the quality of the parent–child relationship, parental education, race, and whether the child lived with both parents. However, the strength of the sexual assault–major depression association was weaker after controlling for family environment factors.

Using a large random New Zealand community sample of women, Mullen and his colleagues (1993) reported that CSA was independently related to several diagnosed psychological disorders, after controlling for inadequate parenting, parental divorce, and early physical abuse. This study also found that the severity of abuse was the strongest predictor of later problems.

Finally, two twin studies also supported the independent effects of CSA on adult psychological

disorders, controlling for different aspects of family environment. Dinwiddie et al. (2000) conducted diagnostic interviews and measured CSA in 2,700 Australian twin pairs (male and female monozygotic and dizygotic pairs). A single question assessed for 'forced' CSA, and diagnoses were obtained for substance abuse, major depression, anxiety disorders, and conduct disorder. The only family environment factor measured was parental psychopathology. CSA significantly predicted all diagnoses in women and all but social phobia in men, after controlling for parental alcoholism and depression. The authors reported that when one twin reported CSA and the other twin did not (discordant for CSA), rates of disorders were not significantly different. However, further analyses revealed much higher rates of disorders among twins who were both abused (concordant), indicating that CSA may have specific effects on the development of disorders.

In the most recent and most sophisticated multivariate twin study to date, 1,411 adult female twins from the USA were given structured clinical interviews to determine lifetime diagnoses of major depression, generalized anxiety disorder, panic disorder, bulimia nervosa, alcohol dependence, drug dependence, and the presence of two or more of the disorders (comorbidity). CSA was measured using multiple screening questions to determine both contact and noncontact abuse. Several family environment factors were measured, but, uniquely, the researchers interviewed the respondents' parents, and made clinical diagnoses of disorders and obtained parent ratings of various family environment factors. Results were that self-reported CSA of any kind was significantly associated with all disorders, except bulimia nervosa, after controlling for parental (or twin) reported family environment factors and history of parental psychopathology. In all disorders (including bulimia nervosa), intercourse CSA was more strongly related to psychological disorders than any other CSA form (genital, nongenital, or any CSA). As in the previous twin study, the cotwin control analyses indicated that there are specific effects of CSA on adult psychopathology, but that shared environmental factors clearly contribute to both risk of CSA and development of future disorders.

3. *Conclusions*

The studies reviewed used the most scientifically rigorous methods to date for testing the causal relation between CSA and adult psychopathology. The studies also had the benefit of all using the same type of measurement for adult psychopathology, a clinical interview that used psychiatric diagnostic criteria for determining the presence of a disorder. The studies support four main conclusions. (a) The statistical relationship between CSA and adult psychopathology remains significant after controlling for certain aspects of negative family environment, although the strength of the relationship attenuates from slightly to considerably, depending on the psychological disorder being predicted. (b) The reported strength or size of the independent relationships between CSA and risk for adult psychopathology were modest, indicating that (i) family environment serves as both a risk factor for CSA and a mediator for the effects of CSA on adult psychopathology; and (ii) many individuals who report a history of CSA do not develop adult problems. (c) More severe forms of CSA i.e., intercourse CSA, lead to a greater risk of developing psychological disorders. (d) Men were under-represented in these studies, but when studied, also had similar risks to women for developing psychological disorders as a consequence of CSA history.

4. *Future Directions*

Research on the prevalence, characteristics, and correlates of CSA has progressed enormously in the last two decades, and findings from each phase of research have informed the next. Despite this progress, there are still some core problems that should be corrected, including the need for standard definitions of CSA, more information about what measurement strategies result in the most accurate information about CSA, and the inclusion of more men in studies. Findings from current research indicate the need for future studies of CSA that assess for a wide range of family environment factors, because it is now certain that the effects of CSA on adult psychological problems are influenced by the familial context in which children live. There is no consensus, however, on which aspects of family environment most influence responses to traumatic events such as CSA, and this research is needed. Although there is some data suggesting that more severe forms of CSA lead to greater risk for adult problems, there is still much to be learned about how characteristics of the abuse may influence later problems, and how abuse characteristics may interact with family environment factors. For example, there is no information about how family environment factors may interact with CSA differently, depending on whether or not the abuser lives in the home with the child. Finally, more information is needed about what happens after CSA, including measurement of subsequent psychological problems, coping methods, and determining what factors seem to protect individuals from developing long-term psychological problems. In order to answer these questions, CSA researchers must continue to use prospective and multivariate cross-sectional studies, but there continues to be a great need for clinical and forensic studies that can inform the direction of more costly epidemiological studies.

See also: Anxiety Disorder in Children; Behavior Therapy with Children; Child Abuse; Child and

Adolescent Psychiatry, Principles of; Child Care and Child Development; Childhood Depression; Children and the Law; Children, Rights of: Cultural Concerns; Sex Offenders, Clinical Psychology of; Violence and Effects on Children

Bibliography

Boney-McCoy S, Finkelhor D 1996 Is youth victimization related to trauma and depression after controlling for prior symptoms and family relationships? A longitudinal study. *Journal of Consulting and Clinical Psychology* **64**: 1406–16

Conte J R 1994 Child sexual abuse: Awareness and backlash. *Future of Children* **4**: 224–32

Dinwiddie S, Heath A C, Dunne M P, Bucholz K K, Madden P A F, Slutske W S, Bierut L J, Statham D B, Martin N G 2000 Early sexual abuse and lifetime psychopathology: A co-twin control study. *Psychological Medicine* **30**: 41–52

Finkelhor D 1979 *Sexually Victimized Children*. Free Press, New York

Fergusson D M, Horwood L J, Lynskey M T 1996 Childhood sexual abuse and psychiatric disorder in young adulthood, II: Psychiatric outcomes of childhood sexual abuse. *Journal of the American Academy of Child and Adolescent Psychiatry* **35**: 1365–74

Finkelhor D 1994 Current information on the scope and nature of child sexual abuse. *Future of Children* **4**: 31–53

Kendler K S, Bulik C M, Silberg J, Hettema J M, Myers J, Prescott C A 2000 Childhood sexual abuse and psychiatric and substance use disorders in women. *Archives of General Psychiatry* **57**: 953–9

Mullen P E, Martin J L, Anderson J C, Romans S E, Herbison G P 1993 Childhood sexual abuse and mental health in adult life. *British Journal of Psychiatry* **163**: 721–32

Nash M R, Neimeyer R A, Hulsey T L, Lambert W 1998 Psychopathology associated with sexual abuse: The importance of complementary designs and common ground. *Journal of Consulting and Clinical Psychology* **66**: 568–71

Rind B, Tromovitch P, Bauserman R 1998 A meta-analytic examination of assumed properties of child sexual abuse using college samples. *Psychological Bulletin* **124**: 22–53

Russell D E H 1984 *Sexual Exploitation: Rape, Child Sexual Abuse, Sexual Harassment*. Sage, Beverly Hills, CA

Vogeltanz N D, Wilsnack S C, Harris T R, Wilsnack R W, Wonderlich S A, Kristjanson A F 1999 Prevalence and risk factors for childhood sexual abuse in women: National survey findings. *Child Abuse & Neglect* **23**: 579–92

N. D. Vogeltanz-Holm

Children and the Law

The ways in which the law has positioned children have varied from society to society and over time. In part this is due to differences and changes in family life, in the material world that children inhabit, and in the way we think about children. The legal standing of children today in the West would be unrecognizable to the eighteenth century observer, who would have seen children as the property of the father. The law relating to children reflects changes in women's legal standing as well as global diversity and changing social and cultural attitudes and practices. With the rapid development of international law in the second half of the twentieth century global variations have become more visible and contested. These variations express some of the key social and political tensions raised by the changing significance of childhood at the beginning of the twenty-first century. So despite the emergence of international laws on children, the ways in which children's welfare and rights are secured differ.

1. *A Brief History of Children and the Law*

Under the Roman civil law doctrine of *patria potestas* the father had unlimited control over his children. This doctrine shaped Western legal systems' vesting of parental rights in the father to the exclusion of the mother until the eighteenth and nineteenth centuries. Within Anglo-American jurisprudence, the common law treated the father as the legitimate child's natural guardian. Legislative reform throughout the nineteenth century encroached upon the rights of the father, giving mothers limited rights to custody of their children in specific circumstances. By the early twentieth century the welfare of the child had become the guiding principle in disputes over children on divorce. However it was only much later in the twentieth century that the rights of mothers and fathers within marriage in relation to their children were equalized.

With the rise in divorce and parental separation in the last decades of the twentieth century and the questioning of earlier norms by the growing feminist movements (Bridgeman and Monk 2000) legislatures and policy makers considered how the welfare of those children affected could be properly safeguarded and promoted. Developments have included the emergence of alternative dispute resolution, the delegalization of family disputes and the adoption of 'no-fault' divorce (Eekelaar and Katz 1984). (See *Family Law*.)

1.1 *Legitimacy*

While legitimate children were seen as the property of the father, illegitimate children were seen as the property of no one. Such a child was *nullius filius* and had no legal relationship with his or her parents. The early bastardy laws were aimed at preventing illegitimate children from becoming a charge on the community—and attempted to do so by punishing the unmarried mother and the reputed father, and charg-

ing either the mother or both for the relief of the child. In the twentieth century, the status of the illegitimate child has changed substantially as has the language concerning legitimacy. Article 25(2) of the Universal Declaration of Human Rights 1948 provides that 'all children, whether born in or out of wedlock, shall enjoy the same social protection' and Article 2(2) of the United Nations Convention on the Rights of the Child 1989 (UNCRC) requires states to take all 'appropriate measures to ensure that the child is protected against all forms of discrimination or punishment on the basis of the status … of the child's parents.' Some countries have abolished the concept of illegitimacy altogether, others have permitted legitimation by subsequent marriage and have legislated to mitigate the adverse legal consequences by extending to illegitimate children the same rights accorded to legitimate children. The parental status of the father has been increasingly recognized in some countries; others, such as those in South Asia continue to recognize only the nonmarital child's legal relationship with the mother (Goonesekere 1998).

1.2 Child Welfare and Protection

Much of the history of children and the law has been dominated by campaigns to protect children and to advance their welfare. The traditional account of the history of childhood is a story of rescue. The nineteenth century witnessed the development of agencies that were dedicated to rescuing children from abuse and neglect within the family. Societies for the prevention of cruelty to children were established in many Western cities, and by the early twentieth century there were over 34 such societies in the USA and fifteen elsewhere. By the late twentieth century, child rescue had become an international as well as national cause (Kent 1995, Van Bueren 1998).

Then, as today, there was debate surrounding the proper role of the state regarding intervention into family life. The ideology of family privacy and the idea that children were the property of their parents combined to render legislatures reluctant to encroach too far into family life. However children of poor and working class families were more visible and more intensively policed than those from wealthier and more powerful sections of the community. The 'discovery' of 'the battered child' in the middle of the twentieth century, and of child sexual abuse in the final decades, gave rise to significant changes in legal practice, such as revisions in the rules regarding the admissibility of children's testimony.

In contrast to the legislative reluctance to intervene in any substantial way with the parent–child relationship, there was considerable legislative activity in the nineteenth and twentieth centuries to protect children from exploitation and abuse outside the home and to better advance their welfare. In the public sphere in industrializing countries, laws were introduced to restrict and finally to prohibit the employment of children in various occupations (e.g., in factories and coal mines). Initially such legislation restricted the hours worked; with the advent of compulsory education laws children were further withdrawn from the labor force. Concerns about child health resulted in measures to reduce infant mortality rates and to improve diet through welfare provision such as school meals. Such measures benefited all children by improving public hygiene, but many focused on the urban poor.

1.3 Institutional Developments

The development of mass education for children in the nineteenth and twentieth centuries has resulted in the school becoming central to the experience of childhood. The law relating to children and education was controversial given that disputes over education raised issues around parental rights, religious freedom, and social justice.

Another institutional expression of the legal and social changes surrounding childhood was the growth of the juvenile court movement in the late nineteenth and early twentieth centuries (see *Juvenile Justice: International Law Perspectives*). Between 1899 and 1908, juvenile courts had been set up in states in the USA, Australia, and England (Mack 1909). Separate penal institutions for juvenile offenders had already been established with reformatories for juvenile offenders being built in the USA and England by the mid-nineteenth century. The rationale for these developments was that children were harmed by exposure to adult depravity, especially within the penal system, that children were more victims than villains, and that with timely and appropriate intervention the child could be rescued from a life of crime and immorality. In a number of jurisdictions there was no distinction made between the child who offended and the neglected child—both were seen as the victim of the socioeconomic environment. The dominance of welfare considerations in the juvenile courts, however, resulted in the neglect of procedural rights for children. Procedural justice for children in the criminal and civil justice systems became an issue in the 1960s and 1970s in a number of jurisdictions. For example in the USA the Supreme Court decision in Re Gault (1965) established that children were entitled to the protection of the Constitution.

Later the idea of a family court emerged in response to the demand that decisions taken in relation to children needed to be informed by specialist knowledge about their needs and welfare. This in turn has led to the emergence of specialist judges in some jurisdictions.

Despite significant historical variations in the position of children under the law, by the end of the twentieth century in civil proceedings the child's welfare was treated as the primary if not paramount

consideration. The language used in many legal systems is no longer that of parental rights alone but rather of parental responsibility; rights that parents do enjoy must be exercised for the benefit of the child. The child is no longer seen as the property of parents.

2. *The Welfare of Children*

Today most legal systems consider the child's welfare to be a primary consideration in the resolution of disputes involving children—a position reinforced by the near universal signing and ratification of the UNCRC.

The law relating to children is increasingly expressed in terms of advancing their welfare though there are exceptions (see Sect. 2.2). The centrality of the welfare of the child is expressed in domestic and international law. Principle 1 of the UN Declaration of the Rights of the Child 1959 provides that in the enactment of laws for the special protection of children 'the best interests of the child shall be the paramount consideration.' This set the international benchmark for assessing legislation affecting the welfare of children until the UNCRC of 1989. The UNCRC is based on the '3 Ps': rights to protection, provision, and participation. Article 3 of the UNCRC provides that: 'In all actions concerning children, whether undertaken by public or private social welfare institutions, courts of law, administrative authorities or legislative bodies, the best interests of the child shall be a primary consideration.' Article 3 is wider than Principle 1 of the UN Declaration insofar as it states that the child's welfare shall be the primary consideration of 'public or private social welfare institutions.'

While the welfare principle is the primary concern in judicial proceedings affecting children the ability of legislation directly to safeguard and promote the welfare of children is limited by the material and ideological circumstances of their upbringing. The condition of children in the world today varies enormously; this can be seen in different mortality rates, child labor, access to education, malnutrition, the prevalence of child prostitution, and the presence of child soldiers. These issues give rise to intense debate nationally and internationally. For example, while Article 32 of the UNCRC requires states parties to the UNCRC to recognize the right of the child to be protected from economic exploitation and from performing any work that is likely to be hazardous or to interfere with the child's education' child labor, including bonded labor, is still widespread throughout South Asia and South America. The International Labour Organization Convention No.138 links the child labor issue with education and provides that the minimum age for admission to employment must not be lower than the compulsory school age and in any event should not be lower than 15.

The debates surrounding the welfare rights of children are intense because they raise questions regarding state power, whose definition of welfare prevails, and respect for the child's identity and community. These issues are contested at both the national and global level (Alston 1994).

2.1 *Children and Families*

The conventional account of the different ways in which children are treated by the law has two divergent strands—one that emphasizes the child's place within a particular family (belonging to a specific religious, ethnic, and cultural community) and one that offers a more individualized approach, recognizing the distinctiveness of interests even within the family. Domestic and international law sees the child's family as being the natural and proper environment for the child's upbringing; even though the concept of the family is undergoing change, ranging from increased cohabitation to gay marriage and adoption (Cornell 1998). There is considerable social value in ensuring that, within limits, parents should be free to bring up their children in a way that is consistent with their own values and beliefs. However, there is variation in how legal systems regulate family life, and in particular how they allocate powers and duties to parents and determine the circumstances in which family privacy can be overridden.

2.1.1 Child abuse. While child abuse is not confined to the family, much of the debate about the legal framework focuses on this setting. The abuse of children in 'public care' (while regularly plagued by scandal) tends to generate discussion about the accountability of welfare bureaucracies and the quality of care provided by the corporate parent. Abuse within the home gives rise to special problems for legal systems. Not only do childrearing practices vary, but some groups (e.g., the urban poor) are subject to more coercive forms of intervention than others. In some jurisdictions, for example, the grounds for intervening into family life in order to protect children are drawn in broad terms (e.g., statutes identify 'immorality' as constituting neglect and thus permit the removal of the child), while in others there is a minimum harm threshold that has to be satisfied if any intervention is to be justified. Jurisdictions vary in the degree to which they adjudicate between the competing values of parents' right to raise their children according to their own values, on the one hand, and the child's right to be protected, on the other (Schwartz-Kenney et al. 2001).

2.1.2 Relationship breakdown. The rise in relationship breakdown in the West has led to a number of significant developments.

There has been an increase in disputes over the custody of children when relationships end. Many jurisdictions have moved towards no-fault divorce, and where private ordering is unsuccessful disputes over children are increasingly resolved via mediation. Where such disputes are not resolved through mediation, court adjudication tends to focus on the welfare of the child. The enforcement of orders made in relation to custody and access, or contact, is problematic; the courts in some jurisdictions (such as the USA) at times threaten to imprison the custodial parent if they continue to refuse to allow the noncustodial parent access.

Custody disputes occasionally involve child abductions. Most states have provisions dealing with child abduction nationally and have entered into bilateral and regional treaties to deter the practice (e.g., the Inter-American Convention on the International Return of Children 1989). At the international level, the Hague Convention on the Civil Aspects of International Child Abduction (1980), which has been signed by over 50 countries including most in North and South America and Europe, focuses on the enforcement of rights of custody between contracting states. The Convention applies where a child has been wrongfully removed or retained in breach of rights of custody. The principle behind the Convention is that children who have been abducted should be returned immediately to the country from which they have been abducted and that any dispute as to custody should be resolved within that jurisdiction. Article 13 of the Hague Convention provides defenses to an order for the immediate return of the child including where the child objects and has attained an age and maturity where it is appropriate to take account of his or her views. This recognition of the autonomy rights of the child reflects how far the law has moved away from the idea that children are the property of their parents.

2.2 *Juvenile Crime*

While legal doctrine, procedure, and philosophy regarding the treatment of juvenile offenders vary across nations, international law at the end of the twentieth century has begun to lay down minimum rules for the administration of juvenile justice. In 1985 the UN General Assembly adopted the UN Standard Minimum Rules for the Administration of Juvenile Justice (the 'Beijing Rules'). The Beijing Rules constitute a benchmark against which national juvenile justice systems can be measured covering the investigation, prosecution, and punishment of juvenile crime.

Today the trend is towards a balancing of welfare and punishment considerations. The discrediting of aspects of the rehabilitative ideal in some countries in the 1960s and 1970s has led to a pragmatism and managerialism in juvenile justice. Increasingly juvenile offenders are segregated from adult offenders and diversion and reparation schemes have developed to reduce the risk of contamination from the formal justice system and to underscore the offender's responsibility to the victim and the wider community. However in some countries (e.g., some states in the USA) there are moves to treat juvenile offenders as if they were adults, including exposing them to the possibility of capital punishment for offences committed whilst a minor (Strater 1995).

3. *The Modern Children's Rights Movement*

It is now accepted that children have rights (Freeman 1997). The meaning and realization of such rights is contested locally and globally. Ideas about children's rights are linked with the child's own community and culture.

The UNCRC signals that children have civil, political, and social rights. These 'human rights' are not just concerned with the welfare of children, though many of the UNCRC's provisions are concerned with their protection (e.g., from abuse) and the adequate provisioning of childhood. The distinctive feature of the modern children's rights movement is that alongside the traditional concern with the welfare and protection rights of children it aims to promote their 'liberty rights.' The rights contained in Articles 12 to 16 of the UNCRC, are indispensable to developing respect for the children in society and to fostering their participation in the community (De Winter 1997). The UNCRC has also acted as a catalyst for the further development of children's ombudsmen (Flekkoy 1991)

However the meaning of the right to participate is contested (see Ncube 1998), and legal systems are cautious about seeing the child as a legal subject. While in a number of jurisdictions children are able to instruct lawyers to represent them in private as well as public law proceedings, invariably the court's view of that child's welfare will prevail over the child's wishes.

In short, changing views of children are effecting legal change. Children are now acknowledged as having a voice in law in an increasing number of jurisdictions. Familial relations are undergoing change pursuant to a trend towards the democratization of family life (Beck 1997). Changing views about children and human rights are bringing about a movement to end the corporal punishment of children. The UNCRC is giving rise to increased tension between the local and global positioning of the child. Debates on protecting children from traditional childrearing practices and the labeling of these as 'abusive' has given rise to critiques of universalist conceptions of rights. Yet states themselves are subject to scrutiny as to how they are meeting their obligations under the UNCRC. While the way in which the law treats children continues to be contested, the emergent language of rights concerning social and intergenerational rela-

tions threatens to subvert many established ways of treating children and to open up spaces in which children can acquire their own voice.

See also: Adoption and Foster Care: United States; Child Abuse; Children, Rights of: Cultural Concerns; Dissolution of Family in Western Nations: Cultural Concerns; Divorce and Children's Social Development; Divorce, Sociology of; Family as Institution; Family Law; Family Processes; Family Theory: Role of Changing Values; Nontraditional Families and Child Development; Poverty and Child Development; Regulation: Family and Gender; Street Children: Cultural Concerns

Bibliography

Alston P (ed.) 1994 *The Best Interests of the Child: Reconciling Culture and Human Rights*. Clarendon Press, Oxford, UK

Beck U 1997 Democratization of the family. *Childhood: A Global Journal of Child Research* **4**(2): pp. 151–68

Bridgeman J, Monk D (eds.) 2000 *Feminist Perspectives on Child Law*. Cavendish, London

Cornell D 1998 *At the Heart of Freedom: Feminism, Sex and Equality*. Princeton University Press, Princeton, NJ

De Winter M 1997 *Children as Fellow Citizens*. Radcliffe Medical Press, Oxford, UK

Eekelaar J, Katz S N (eds.) 1984 *The Resolution of Family Conflict: Comparative Legal Perspectives*. Butterworths, Toronto, Canada

Flekkoy M 1991 *A Voice for Children*. Jessica Kingsley, London

Freeman M D A 1997 *The Moral Status of the Child*. Martinus Nijhoff, The Hague, The Netherlands

Goonesekere S 1998 *Children, Law and Justice*. Sage, New Delhi, India

Kent G 1995 *Children in the International Political Economy*. St Martin's Press, New York

Mack J 1909–10 The juvenile court. *Harvard Law Review* **23**: 104–22

Ncube W (ed.) 1998 *Law, Culture, Tradition and Children's Rights in Eastern and Southern Africa*. Dartmouth, Aldershot, UK

Schwartz-Kenney B, McCauley M, Epstein M (eds.) 2001 *Child Abuse: A Global View*. Greenwood Press, Westport, CT

Strater S 1995 The juvenile death penalty: In the best interests of the child? *Loyola University Chicago Law Journal* **26**: 147–82

Van Bueren G (ed.) 1998 *Childhood Abused: Protecting Children against Torture, Cruel, Inhuman and Degrading Treatment and Punishment* (Programme on International Rights of the Child) Ashgate, Aldershot, UK

J. Roche

Children, Rights of: Cultural Concerns

The last 100 years have witnessed enormous progress in the recognition of individual human rights throughout the world. In addition, various identity groups, including women, racial, ethnic and religious minorities, the disabled, and gay males and lesbians all have claimed, and received to varying degrees, rights to equal treatment under law, nondiscrimination in employment in both the public and private sectors, and fuller integration into the economic and political life of their countries.

It is perhaps not surprising, therefore, that a worldwide movement to recognize children as a distinct subgroup with claims of rights also occurred. In 1924, the League of Nations adopted the first Declaration of the Rights of Children. In 1989, the United Nations General Assembly adopted the UN Convention on the Rights of the Child (UN Convention 1989), which has been called the 'most comprehensive and detailed of all … international human rights instruments' (Alston 1994, p.1). The Convention has been ratified by 192 of the 194 UN members. In the United States conservative political groups, concerned that the CRC undermines parental authority and gives government too much power over families, have contributed to blocking ratification. Somalia lacks an internationally recognized government, precluding ratification.

Despite the seemingly widespread acceptance of the concept of children's rights, the idea that children, individually or as a class, should have *rights* (claims enforceable against others) is actually revolutionary and problematic both in philosophical and practical terms. Determining what rights to provide children requires that societies confront fundamental value issues quite different from those entailed in recognizing other groups' claims.

First, children generally are not autonomous; they are dependent. Ideally, they are part of families; their life chances are greatly affected by the care they receive from their families. Parents need authority to fulfill these responsibilities; some highly respected child development specialists argue that giving children rights may be incompatible with sound child development (Goldstein et al. 1973). In addition, giving children rights may conflict with parental rights and privacy. However, parents do not always act in their children's best interests and the community as a whole has an interest in the development of future generations. Children have their own views and desires. Therefore, it must be decided how responsibility for the well-being and upbringing of children should be allocated between parents, families, the child, and the community as a whole.

Second, intellectual and emotional capacities develop gradually, for both biological and social reasons. These limits in capacity may conflict with claims that children should have the same liberties and privileges as adults (Purdy 1992). A theory of children's rights must address the question, what entitles persons to be full participants in a country's civil and political life?

This article reviews the historical development of children's rights. It then describes the different cat-

egories of claims made under the rubric of children's rights and discusses value issues that need to be resolved as governments, courts, and citizens seek to define the children's status in each society. It concludes by examining the challenges countries face implementing any such rights.

1. The Historical Context

For much of history children were viewed largely as nonpersons, the property and responsibility of their parents, who had a right to control their upbringing, even their very existence (Eekelaar 1986). Children (usually defined as persons under 21 or 18 years of age) also could not participate in the political or civic life of their countries.

In the mid-1800s, some countries began assuming state responsibility for promoting and protecting children's well-being. Initially, this entailed providing free public education and passage of laws protecting children from severe physical abuse by their parents. Over the next 100 years, Western industrialized countries adopted more laws affecting the status of children, including laws limiting child labor in settings outside the family, passage of compulsory school attendance laws, and greater state efforts to protect children from parental maltreatment. The idea that childhood was a period for gradual maturation and lesser responsibility also was reflected in the creation of the juvenile court, which was directed to treat, not punish, children who violated the law (Hawes 1991).

While these policies were advocated in the name of children's rights, and certainly contributed to their well-being, these were not the types of civil rights that other groups were demanding and receiving. The focus was on their protection, not their autonomy or fuller integration into the economic and political life of their countries. Moreover, except where parental behavior was seen as inimical to the social order, parents continued to have virtually total authority over their children's upbringing. In the United States, where the movement towards child protection was most evident, the US Supreme Court ruled early on that parents had a constitutional right to control their children's upbringing (Meyer v. Nebraska 1923).

In fact, these laws often were supported primarily as a means of protecting adult interests. For example, child labor laws and compulsory schooling were championed by newly emerging labor unions, which viewed the availability of child labor as a barrier to higher wages for adults (Woodhouse 1992). Proponents of the juvenile court promoted it as a needed intervention to protect society from children who were not being adequately supervised by their parents, often poor, immigrant parents (Schlossman 1977).

Since the 1960s, countries have continued expanding the protection of children's well-being. The establishment of a right to education for virtually all children throughout the world, increased state efforts to guarantee children a minimal level of material well-being, and greatly expanding state intervention into family privacy through child abuse laws have all benefited children. Moreover, during this period, in many countries children were recognized as rights-holders with respect to freedom of speech and other civil liberties and were given a greater degree of legal autonomy over the decisions affecting their lives.

Yet the scope of applicable rights remains highly contested. At the extreme, some commentators have argued for 'children's liberation,' with the complete elimination of distinctions between children and adults with respect to all political, social, and familial rights (children, as reflecting of their status, usually have not been the demanders of rights, in contrast to other groups). For example, they argued that children should have the right to vote, to decide whether to go to school, and to a share of the family income to spend as they choose (Farson 1974).

Few proponents of children's rights go this far. Most advocates propose limiting, but not eliminating, parental autonomy in controlling their children's lives and giving children increasing autonomy as they mature, in recognition that capacities develop over time. Yet even the more limited types of claims of rights raise profoundly difficult issues.

2. What is Meant by Children's Rights

The term children's rights is used to encompasses a variety of different claims (Campbell 1992). Distinguishing these categories reveals the tensions in the general concept and helps identify the issues related to the implementation of various types of claims.

2.1 Basic Human Rights

The first claim is that children are entitled to basic human rights, including the right to life itself, the right not to be 'owned' by another, the right to freedom from tobrture or other cruel punishment, and the right to be free from discrimination based on race, color, sex, religion, or national, ethnic, or social origin.

Such rights are widely accepted for adults, despite continuing problems with implementation, especially with respect to nondiscrimination based on gender, race, and ethnicity. Since these rights are not tied to competence or capacity, there is no reason for denying them to children on this basis. The critical step is accepting that children are *persons*, not chattels or incomplete human beings. The adoption of this premise by virtually all countries reflects a fundamental shift in the conception of children.

Still, application of these rights to children is not entirely unproblematic, at least with respect to the relationship of parents and children. Few countries

still consider children to be the property of their parents. However, many people agree with Professor Charles Fried that 'the right to form one's child's values, one's child's life plan … are extensions of the basic right not to be interfered with in doing these things for oneself' (Fried 1978). Yet, such a right constitutes a form of 'ownership' and may conflict with the child's right to an 'open future' and with society's interest in how future adults are socialized (Feinberg 1980).

There is no easy resolution to this conflict. Unless the state were to intervene in family life in ways that would be deemed unacceptable in most countries, parents will continue to make critical decisions for children, regardless of the child's wishes. As noted, respected child development experts assert that parents need autonomy to successfully raise their children.

Moreover, deference to parental or family authority may be demanded by a society's political, cultural, or religious traditions. In some countries, supporting parental autonomy is seen as critical to supporting political and cultural diversity. In others, the proper role of parents may be seen as teaching children to accept national cultural values and traditions, not as helping children develop into 'autonomous' adults (see generally Alston 1994). Many countries have reserved the right to apply the CRC in light of their own cultural or religious laws, although there is evidence that the CRC is influencing these values (*International Journal of Children's Rights* 1995). Whether children's rights are universal or culture specific likely will remain one of the major debates in implementing the CRC.

In fact, given that children's values and perspectives are malleable and constructed, someone will shape their values and personhood. Parents, families, neighborhoods, peers, government institutions such as schools and general cultural norms all play a part. Thus, the issue is what are the appropriate roles of the family and the state in making critical decisions, such as those regarding schooling, religion, where the child shall live, not whether children should have the right to control their own upbringing (Coons et al. 1991). This question goes to the heart of how a society defines the relationship between the state and the individual, adults and children alike.

Children also are denied some basic civil rights available to adults because societies have an interest in how they are socialized. At a minimum, children are compelled to go to school, thus restricting their liberty rights. Finally, there is still substantial debate over the status of the fetus with respect to abortion and maternal behavior during pregnancy. Because no consensus could be reached on this issue, the CRC leaves the question to each country.

Despite these problems, the recognition that children are persons, with rights, along with the acceptance of the basic premise of the CRC that in all decisions affecting children their 'best interests' shall be a primary consideration, constitutes a great advance that clearly benefits the vast majority of children.

2.2 Welfare Rights

A second category of claimed rights is that government should guarantee children a basic level of well-being; these are often called welfare rights. The CRC provides that children should have a right to 'a standard of living adequate for the child's physical, mental, spiritual, moral and social development' (ART 27), 'the enjoyment of the highest attainable standard of health …' (ART 24), and to education (ART 28). The CRC goes beyond these basic goods, and seeks to afford children a broader set of welfare rights, including the right 'to rest and leisure, to engage in play and recreational activities appropriate to the age of the child and to participate freely in cultural life and the arts' (ART 31).

Since the 1950s, governments have assumed increasing responsibility for guaranteeing that children have adequate food, nutrition, shelter, and health care. In many countries, children benefit from general social welfare policies, which provide all citizens with universal health insurance and various forms of income support. Most industrialized countries also have partially socialized the cost of caring for children through children's allowances, paid parental leave after the birth of a child, and public provision of day care. Many less economically developed countries are moving in the same direction, despite limited financial resources.

The laggard in this regard, given its capacity, is the United States. In the late 1900s child poverty actually grew in the United States. Resistance in the US reflects, in part, the belief of many political figures that childhood poverty is caused by 'irresponsible' adult behavior, such as having children out of wedlock, and that it is impossible to provide goods to children without also giving them to adults, thereby 'rewarding' the adults' action. In addition, many politicians support the premise of libertarian theorists like Robert Nozick that there is no moral justification for the redistribution of wealth (Nozick 1974).

Even in countries that accept a welfare state, the level of support that a child should receive remains largely a political issue, not a legal right enforceable by courts. No matter how wealthy a country, tradeoffs between levels of socially provided health care, housing, education, and other welfare goods are now seen as inevitable. Courts, appropriately, are reluctant to get too deeply into these decisions. Since the scope of the claimed right is so vague, courts lack the legitimacy and authority to order elected governments to provide people with specific levels of well-being. In addition, some theorists contend that, in the political process,

framing political debates in terms of moral obligations that the citizens of a country pledge to provide to each other may prove more productive thn trying to establish rights (O'Neill 1988, cf. Freeman 1992).

Another critical question is whether welfare rights should encompass the right to 'equal opportunity.' Opportunity focuses on each child's options in adulthood. Social scientists have struggled to identify the goods that must be provided to children during childhood in order to equalize their opportunities as adults. Much of the focus is on education. But opportunity is also highly influenced by the quality of parenting a child receives and the social and cultural environment in which the child is raised. A right to equal opportunity is far more difficult to guarantee than a right to basic necessities and may require restricting parental rights (Fishkin 1983).

2.3 Protection Rights

A third category of claimed rights for children is protection from inadequate care from their caretakers, especially their parents. Since giving children more protection changes their relationship with their parents and the state's relationship with parents and families, this right raises fundamental issues quite different from welfare rights.

The degree of justifiable state intrusion against a parent's will remains highly debated. Should the state treat parents as 'stewards' or 'trustees' of their children, and seek to ensure that parents provide children a high level of care, or should the state be the protector 'of last resort,' intervening only if parental care is clearly harmful?

Most countries have decided that coercive intervention should be limited to situations where the actual or threatened harm to the child's physical or emotional well-being is substantial. This approach reflects the judgement that unwanted intervention is a drastic step from the child's, as well as the parent's, perspective. It also is based on evidence that the care provided to children removed from their families sometimes is as bad or worse than the parental care (Wald 1982).

However, a few countries have adopted laws that seek to provide greater protection for children. Five countries have banned any corporal punishment of children; while these laws do not carry any significant sanctions for violation, they reflect a different conception of the rights of children and the role of the state. In recent years, evidence that early child-rearing can substantially affect a child's school readiness and performance has led to increased calls for more extensive monitoring of how each child is being reared.

In the future, governments will continue to struggle with the appropriate grounds for protective intervention. The critical task, however, will be to develop and connect children to social programs that enhance all children's well-being and future opportunities without resorting to coercive interventions. By the time protection rights must be invoked, the child has suffered substantial harm and the capacity of the state to rectify the situation is often limited (US Advisory Board 1991).

2.4 Liberty and Participation Rights

The fourth category of rights does focus on autonomy. In virtually all countries, persons under 18 lack the right to vote and to marry, they are restricted in making decisions about whether to work, and in their freedoms of expression and other civil rights. In addition, in many countries children accused of criminal behavior generally are denied some due process rights available to adults in that country, for example the right to a jury trial; these limitations have been justified as necessary to promote the rehabilitative ideals of the juvenile justice system.

In 1967, the US Supreme Court determined that children are entitled to at least some of the rights guaranteed by the free speech provision of the US Constitution (Tinker v. Des Moines Independent Community School District 1969). Since then the liberty rights of minors have been gradually increased in the US and many other countries. The CRC provides that: '(the) child shall have the right to freedom of expression; this right shall include the freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of the child's choice' (ART 13); '(s)tates parties shall respect the right of the child to freedom of thought, conscience, and religion' (ART 14); '(s)tates parties recognize the rights of the child to freedom of association and to freedom of peaceful assembly' (ART 15). Notably, the CRC does not provide adult status in other areas, such as the right to vote, marry, work, or to be from constraints imposed only on children, such as the obligation to attend school, or to be at home by a certain hour (curfew laws).

Courts, legislatures, and the drafters of the CRC have not arrived at any consistent theory justifying granting children some rights and not others. Nor have they resolved the tension between claims for special protection and for full participation. If adolescents are seen as having the maturity and capacity to make decisions about their health care, to choose what they read or see, or to vote, does it follow that they should be held fully responsible for their antisocial behavior?

The critical issues relate both to capacities and to questions about developmental needs. For some rights, such as voting or marrying, it is generally accepted that some minimum level of competence is

necessary to the proper exercise of the right. Children often are denied other rights, such as unlimited access to books or movies, on the assumption that early exposure to certain materials or ideas may be harmful to them.

An overriding question is whether it makes sense to assume that the same age line should govern access to all political and civil rights. Franklin Zimring has proposed treating children as possessing a 'learner's permit,' whereby they are given increasing autonomy and responsibility over time (Zimring 1982). Different rights might be available at different ages, based on best guesses about average capacity. Age lines are inherently arbitrary; not all people attain capacities at the same age. But individualized decisions about whether someone is old enough to vote, marry, or leave school are likely, in practice, to be even more arbitrary, as is having a single age trigger all rights.

2.5 Autonomy Within the Family

The issues regarding autonomy rights are even more difficult with respect to the right of children to act independently of their parents' wishes. The extreme view, that children of any age should be free to make their own decisions, has not received acceptance. In many countries giving children any autonomy within the family would be incomprehensible. In other countries, there are debates about whether children should have a right to make certain major decisions, such as whether to have an abortion or receive treatment for substance abuse, without their parents' permission or knowledge. Advocates of giving children these rights contend that many children will be reluctant to seek out needed services if their parents are informed. Children have also been given some rights to determine important aspects of their life, such as which parent to live with in the case of divorce. The case for giving children rights is strongest with respect to decisions where parents may not be counted upon to generally act in their children's best interests.

There is no easy resolution to the tension between parental rights and children's rights. It is unrealistic to assume that children can have total liberty rights, enforceable against their parent's opposition. Asking courts to enforce most such rights would be a waste of court resources and the orders would be unenforceable as a practical matter. But giving children some rights is feasible, for example, access to drug treatment or abortion without parental permission. The question is whether the benefits of giving children autonomy with respect to the decision outweigh the harms of limiting parental oversight of their children. These rights will have to reflect the legal and social structure of each country and should not be thought of as universal children's rights.

3. The Future

The situation of children throughout much of the world has been improved substantially during the twentieth century. Concepts of rights likely contributed to these advances. The recognition that children are independent human beings with claims to special attention and resources is a major advance with respect to universal recognition of human rights (Freeman 1992).

However, further development of children's rights will require resolving the value issues often masked by rights talk. Despite the unprecedented rapidity of the ratification of the CRC, the fact that the contours of many of the proposed rights are not very clearly defined will make enforcement and monitoring by the international bodies problematic. In fact, the very vagueness of the obligations may account for the rapid ratification and almost total support of the CRC, despite the fact that many of the proposed rights seemingly conflict with deeply held cultural and religious values in many countries. Countries expect to interpret the provisions in light of their customs and values.

With respect to welfare rights, those concerned with helping children must recognize that realization of most of the proposed rights requires policies that help families, not just children. Children's well-being generally is inseparable from that of the parents. Greater provision of welfare rights should lessen the occasions that protection rights will need to be invoked. However, in establishing welfare rights countries need to ensure that when adults seek to act on behalf of children, their own needs do not take precedence.

With respect all general social policies, countries must think through the implications of the fact that children, who may constitute a third to half of the population, cannot vote. The poorest children are further disenfranchised by the general powerlessness of their parents. Some countries are making significant efforts to include the voices of children in discussions of issues related to them and to more general public policy issues. The recognition of children's right to speak for themselves could significantly alter the place of children in the state (Melton and Limber 1992).

In the poorest countries, it will be very difficult to promote any form of children's rights unless general poverty is alleviated. Many countries continue to deal with problems of street children, child prostitution, child labor, and infanticide. The advancement of children's rights in these countries will be intimately linked to women's rights, since it is unlikely that children will have economic security and greater liberty if their mothers lack these rights. It will also require wealthier countries to demonstrate concern for children beyond national borders.

Children's rights advocates sought to make the 1900s the century of the child. They were successful in many respects. Future advances will require finding

adequate balances for the tensions that have been identified, as well as eliminating the negative impacts on children of discrimination based on race, class, and gender. For most children, these later factors may be greater barriers to their rights and well-being than age.

Bibliography

Alston P 1994 The best interests principle: towards a reconciliation of culture and human rights. In: Alston P (ed.) *The Best Interests of the Child*. Clarendon Press, Oxford, UK

Armstrong M, Chuulu M S, Himonga C, Letuka P, Mokobi K, Ncube W, Nhlapo T, Rwezania B, Vilakazi P 1995 Towards a cultural understanding of the interplay between children's and women's rights: an Eastern and Southern African perspective. *International Journal of Children's Rights* **3**: 333–68

Campbell T 1992 The rights of the minor: as person, as child, as juvenile, as future adult. In: Alston P, Parker S, Seymour J (eds.) *Children, Rights and the Law*. Clarendon Press, Oxford, UK

Convention of the Rights of the Child 1989 G.A. Res 44/25 U.N., GAOR, 44th sess., Annex, Supp. No 49 at 167 U.N. Doc A/44/49

Coons J, Mnookin R, Sugarman S 1991 Puzzling over children's rights. *Brigham Young University Law Review* 307–50

Eekelaar J 1986 The emergence of children's rights. *Oxford Journal of Legal Studies* **6**: 161–82

Farson R 1974 *Birthrights*. Macmillan, New York

Feinberg J 1980 The children's right to an open future. In: Aiken W, LaFolette H (eds.) *Whose Child? Children's Rights, Parental Authority, and State Power*. Rowman and Littlefield, Tottowa, NJ

Fishkin J S 1983 *Justice, Equal Opportunity, and The Family*. Yale University Press, New Haven, CT

Freeman M 1992 Taking children's rights more seriously. In: Alston P, Parker S, Seymour J (eds.) *Children, Rights and the Law*. Clarendon Press, Oxford, UK

Fried C 1978 *Right and Wrong*. Harvard University Press, Cambridge, MA

Goldstein J, Freud A, Solnit A J 1973 *Beyond the Best Interests of the Child*. Free Press, New York

Hawes J 1991 *The Children's Rights Movement: A History of Advocacy and Protection*. Twayne Publishers, Boston

International Journal of Children's Rights 1995 Special Issue: Multiculturalism and The Rights of the Child **3**: 1–144

Melton G, Limber S 1992 What children's rights mean to children: children's own views. In: Freeman M, Vreeman R (eds.) *The Ideologies of Children's Rights*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Meyer v. Nebraska 1923 *United States Supreme Court Reports* **262**: 390–403

Nozick R 1974 *Anarchy, State and Utopia*. Basil Blackwell, Oxford, UK

O'Neill O 1988 Children's rights and children's lives. *Ethics* **98**: 445–63

Purdy L 1992 *In Their Best Interest?* Cornell University Press, Ithaca, NY

Schlossman S L 1977 *Love and the American Delinquent*. University of Chicago Press, Chicago

Tinker v. Des Moines Independent Community School District 1969 *United States Supreme Court Reports* **393**: 503–26

United States Advisory Board on Child Abuse and Neglect 1991 *Creating Caring Communities*. US Department of Health and Human Services, Washington, DC

Wald M 1982 State intervention on behalf of endangered children. *International Journal of Child Abuse and Neglect* **6**: 3–45

Woodhouse B 1992 'Who Owns the Child?': *Meyer* and *Pierce* and the child as property. *William and Mary Law Review* **33**: 995–1122

Zimring F 1982 *The Changing Legal World of Adolescence*. Free Press, New York/London

M. S. Wald

Children, Value of

In demography, the term 'value of children' most often refers to the benefits parents receive from having and rearing children. Benefits may accrue from the children themselves, from the experience of rearing them, or from the responses of kin, community, and society at large. Children also entail costs for parents and the 'value of children' sometimes refers to their net value (benefits less costs). Benefits and costs of children are shaped by the economic conditions of life, by forms of social organization, and by cultural beliefs and practices. The net value of children underlies parents' desires for children; childbearing desires, in combination with the ability to achieve them determine whether or not individuals or couples have children and how many children they have.

1. Theoretical Development

The economic value of children is a key component of fertility variation and change. In agricultural economies and during early periods of industrialization, parents and kin make the heavy investment of time and money in young children in order to reap the rewards of children's labor from adolescence onward (Schultz 1973). The old-age security value of children is particularly important in contexts where no public provisions exist for elder care (Cain 1985). For large numbers of people throughout the world, the economic value of children continues to be a primary benefit of parenthood.

For most parents in industrial societies, however, children provide very little or no economic value. Children's labor does not produce subsistence for the family nor do children often support their parents in old age. Children are extremely costly to raise and often require economic support well into the young adult years as they complete their formal education. The location and organization of paid employment

produces high opportunity costs of childrearing for parental employment. Fertility declines in the twentieth century have been attributed in large part to declines in the economic value and increases in the economic costs of children to their parents (Fawcett 1983).

Of course, scholars have long recognized that the value of children extended beyond their economic benefits. Hoffman and Hoffman (1973) identified the following psychological needs fulfilled by parenthood and the experience of childrearing: (a) adult status and social identity; (b) expansion of the self (tie to a larger entity, 'immortality'); (c) morality (including religion, altruism, good of the group); (d) primary group ties, affiliation; (e) stimulation, novelty, fun; (f) creativity, accomplishment, competence; (g) power, influence, effectance; and (h) social comparison, competition. The influence of such benefits on fertility would depend on the availability of alternative mechanisms for fulfilling the identified needs. They also recognized that childrearing carried with it potential psychological costs such as stress and worry.

Recent theoretical arguments reject lists of psychological benefits as sufficient to explain why people continue to have children when they become economic liabilities. One alternative theory claims that the reduction of uncertainty is the primary goal of parenthood (Friedman et al. 1994). Those with limited access to other means for uncertainty reduction, such as stable careers and marriages, will want to become parents. Few of the hypotheses derived from this theory stand up to empirical scrutiny and those that do can be explained by the economic and social opportunity costs of children (Lehrer et al. 1996).

Building on the work of cultural anthropologists and sociologists, Schoen et al. (1997) proposed the theory that children produce social capital for parents. Social capital consists of social relationships and the social resources they provide. At birth, children strengthen parents' ties to kin. Through schooling and other activities, they link parents with community resources. As adolescents and young adults, children bring new information, ideas, and social relationships to parental households. And most parents eventually obtain in-laws and grandchildren as a consequence of parenthood. Parents may also, of course, incur loss of social capital because they have less time to maintain friendships, relationships with co-workers, or connections to social and political organizations.

While the distinctions among economic, psychological, and social benefits and costs of children make some sense (and also fit disciplinary boundaries), it is important to recognize their interrelationships. Economic subsistence comes first in a list of parental needs, so the economic value of children may be sufficient to stimulate parenthood, whether or not children have any social or psychological value. Social and psychological benefits may be viewed as extras that make the task of childrearing less burdensome, or at least outweigh the social and psychological costs of parenthood. On the other hand, the kin and community ties produced by children may be indirect sources of economic benefits. Several of the benefits identified as psychological also have social components: primary group ties include the parents' relationship, itself a form of social capital; expansion of self includes ties to larger social groups, including kin, community, and society.

2. *Measuring the Value of Children*

The economic value and costs of children could be measured indirectly by estimating the value of children's labor and transfers to parents over the life course, expenditures related to childrearing, and the value of parental time that might otherwise be spent on income-producing activities (e.g., Rosenzweig 1978). Similar estimates of social or psychological values have not been attempted, and it could be argued that a shared metric for economic, social, and psychological values does not exist. Even if it were possible to measure the true net value of children, parents' perceptions of those values are what enter into fertility decisions.

The term 'value of children' is most closely associated with a set of surveys conducted in the mid-1970s in the United States, the Philippines, South Korea, Taiwan, Indonesia, and Thailand (Fawcett 1983). Although based on the psychological values of children discussed above, the surveys also contained information on the perceived economic and social benefits and costs of children. Respondents were asked open-ended questions about reasons for having or not having children and were asked to rate the importance of several lists of child benefits and costs as reasons to have or not have children or their next child. Analyses of these data remain the primary source of current knowledge about variation in the value of children.

Measures of the benefits and costs of children are also the basis for subjective-expected-utility or expectancy-value models of fertility decisions (e.g., Davidson and Jaccard 1979). Respondents are asked to rate the value of possible outcomes of having or not having a child and also the subjective probability that having or not having a child will produce the outcome. The product of those two responses comprises the importance of a given outcome as a reason to have or not have a child. Almost all of the research based on these models has focused on the total expected utility of having a child, rather than on the relative importance of specific child benefits and costs.

More recently, several European Fertility and Family Surveys and the US National Survey of Families and Households used structured importance ratings to measure the value of children to parents. In these surveys, the rating scales are more extensive than the three-point scale used for the earlier VOC surveys, but

the questions do not cover all of the theoretical benefits and costs discussed above. Analyses of these data to date have combined ratings of social and psychological benefits of children into a single scale (e.g., Schoen et al. 1997).

3. Variations in the Value of Children

3.1 The Value of First and Later-born Children

First, second and higher-order births are associated with distinct benefits and costs (Bulatao 1981, Fawcett 1983). The first child confers the status of parenthood, so that benefits associated with parenthood *per se* can be acquired by having only one child. Adult status, relationship stability, parent-child interaction, and kin connections are all cited as primary reasons for becoming a parent. The first child is also associated with the greatest increase in opportunity costs, that is, constraints on parental time and energy.

The most important value of a second child is to provide a sibling for the first (Bulatao 1981). Second, children may further strengthen partnership and kin ties and provide additional opportunities for rewarding parent-child interactions. The value of higher-order births is predominantly economic—each child contributes additional labor or economic security for parents. Restrictions on parents' time are also associated with higher birth orders, but at a diminishing rate. Financial costs of children become more salient to parents at the birth of fourth and higher-order siblings.

3.2 Gender and the Value of Children

Overall, men and women perceive the values and costs associated with children in much the same way. The few differences that are observed are consistent with traditional gender roles (Fawcett 1983). Men are on average more concerned than women about the financial costs of childrearing and about having sons to continue the family name. The latter difference is particularly pronounced in patrilineal societies. Women place greater importance than do men on the work and strain of raising children, the opportunity costs of children for other activities, and the benefits of children for the marital relationship; the last difference is also larger in patrilineal than in bilineal kinship systems.

Differences in the perceived benefits of daughters and sons are also related to differences in the roles and behaviors of men and women (Fawcett 1983). Sons are valued more than daughters for kinship ties, that is, to continue the family name, and for financial assistance. These values are especially pronounced in patrilineal family systems and where men have greater access than women to economic opportunities. Daughters are valued more than sons for household and childcare help and companionship. In patrilineal societies, the values associated with sons are produced for the most part in adulthood, those associated with daughters during childhood, consistent with the practice of daughters' leaving the parental home upon marriage. Many of the benefits of children, particularly social and psychological benefits, do not appear to differ for sons and daughters.

3.3 Socioeconomic Variation in the Value of Children

As noted above, the economic value of children is associated with agricultural and household economic production and the absence of social insurance for elderly parents. Although children also entail high economic costs in such contexts, they are necessary for survival. Industrialization and urbanization reduce the economic value of child labor and increase the costs of rearing children to be economically independent. Economic development eventually leads to the development of social insurance that reduces further the economic value of children for support in old age. As a result, children's net economic value is perceived to be lower in industrialized wealthy countries than in poorer countries with a greater dependence on agricultural and household production (Fawcett 1983).

Psychological benefits and costs of children are reported to be more important in industrial and postindustrial settings than in agricultural settings (Fawcett 1983). This difference may arise because of the priority of economic survival over psychological wellbeing, that is, children may provide psychological benefits (and costs) for parents in all settings, but these components of child value become salient only when children become irrelevant to economic survival. On the other hand, the increasing complexity and impersonal character of daily life in industrialized urban societies may produce a greater need for the love and companionship and stimulation of children. At the same time, psychological costs of childrearing may increase because kin and community take less responsibility for the supervision and care of children.

Because the social capital value of children has only recently been introduced into theoretical discussions of the value of children, it is difficult to know how such values might vary under different economic conditions. In agricultural settings there may be a stronger association between economic and social ties so that the latter are not distinguishable from the former. Only in industrial and urban societies may social capital be sufficiently separable from financial capital to identify it as a separate source of child value.

Socioeconomic variation in the value of children is also evidenced across families within societies (Fawcett 1983). A consistent finding from VOC surveys was that urban respondents place a lower economic value and a higher emotional value on children than do respondents living in rural areas. Similarly, education is inversely associated with children's economic value and directly associated with their emotional value as well as with perceived restrictions or opportunity costs of parenthood. Direct financial costs and childcare stresses do not vary substantially across countries or across individuals in different economic circumstances. Increasing economic status is associated with desire for increased child 'quality' which means greater financial and time/energy investments in each child. Thus, the perceived cost of childrearing, other than opportunity costs, remains essentially the same across socioeconomic levels.

3.4 Culture and the Value of Children

Broad cultural values may also serve as sources of specific or general values of children. Religious institutions and beliefs may support the value of children for social and psychological benefits. For example, Catholicism is viewed as a support for large-family values in the Philippines, Confucianism for the high value of sons to carry on the family name in some Asian countries. The relative values of daughters and sons are associated with broad cultural values on gender equality (Fawcett 1983).

Lesthaeghe (1983) argued that ideational change is an independent force underlying current low fertility in Western countries. He identifies the two most salient features of this change as secularization and individuation. Secularization allows more latitude to individual morality, individuation stresses the importance of personal self-actualization. Using national surveys of social and family values, Lesthaeghe and coworkers (e.g., Lesthaeghe and Meekers 1986) distinguished two dimensions of family values—a nonconformity dimension linked to partner relationships and nonmarital childbearing; and the 'meaning of parenthood,' including beliefs that children are necessary for 'fulfillment' and for marital success. Measures of secularism and individuation were strongly associated with nonconforming family values, but only weakly associated with the meaning of parenthood.

4. The Value of Children and Fertility

Studies using ratings such as those in the Value of Children Surveys have generally found that high perceived economic benefits of children are associated with a large family size. The importance of psychological benefits and restrictions on parental activities are associated with a small family size (Fawcett 1983). The relationship between perceived child values and fertility is not, however, as strong as some scholars had hoped, and the financial and time/effort costs of children are not associated with family size.

Relatively moderate associations between values of children and completed fertility should not be surprising. Specific values are associated with particular numbers of children, not consistently with large or small numbers. Child values can influence only desired fertility, so that the relationship between values and outcomes depends on the degree of fertility control. In addition, measures of the perceived value of children have often been relatively crude (e.g., 3-point response scale). When a particular parity progression is specified, when contraception is pervasive, and when precise measures of the expected value (net of cost) of children are generated, very high correlations are observed with birth intentions and eventual births (e.g., Davidson and Jaccard 1979).

See also: Demographic Transition, Second; Family Size Preferences; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Intergenerational Relations; Family Theory: Role of Changing Values; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Reproductive Rights in Affluent Nations

Bibliography

Bulatao R A 1981 Values and disvalues of children in successive childbearing decisions. *Demography* **18**: 1–25

Cain M 1985 Fertility as an adjustment to risk. In: Rossi A S (ed.) *Gender and the Life Course*. Aldine, New York

Davidson A R, Jaccard J J 1975 Population psychology: A new look at an old problem. *Journal of Personality and Social Psychology* **31**: 1073–82

Fawcett J T 1983 Perceptions of the value of children: Satisfactions and costs. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries, Vol. 2: Supply and Demand for Children*. Academic Press, New York

Friedman D, Hechter M, Kanazawa S 1994 A theory of the value of children. *Demography* **31**: 375–401

Hoffman L W, Hoffman M L 1973 The value of children to parents. In: Fawcett J T (ed.) *Psychological Perspectives on Population*. Basic Books, New York

Lehrer E L, Grossbard-Shechtman S, Leasure J W 1996 'A theory of the value of children'—Comment. *Demography* **33**: 133–9

Lesthaeghe R 1983 A century of demographic and cultural change in Western Europe: An exploration of underlying dimensions. *Population and Development Review* **9**: 411–35

Lesthaeghe R, Meekers D 1986 Value changes and the dimensions of familism in the European Community. *European Journal of Population* **2**: 225–68

Rosenzweig M R 1978 The value of children's time, family size and non-household child activities in a developing country: Evidence from household data. In: Simon J L (ed.) *Research in Population Economics: An Annual Compilation of Research, Vol. 1.*

Schoen R, Kim Y J, Nathanson C A, Fields J, Astone N M 1997 Why do Americans want children? *Population and Development Review* **23**: 333–58

Schultz T 1973 The value of children: An economic perspective. *Journal of Political Economy* **81**: S2–S13

E. Thomson

Children's Play: Educational Function

Until recently, many writers have considered children's play to be a trivial and inconsequential activity; they have also disagreed on its definition. Today, however, social scientists and educators appear voracious in their study of children's play. To many researchers, play is viewed as a generative force in children's social, emotional, and cognitive development (see Rubin et al. 1983). The developmental and educational significance of play in childhood is discussed herein.

1. *Classical Theories of Play*

Contemporary research on the topic of children's play draws heavily from early theoretical accounts about the functional significance of the phenomenon. The 'surplus energy' theory characterized play as 'blowing off steam'. Schiller, an eighteenth-century philosopher defined play as the aimless expenditure of exuberant energy. He wrote that work satisfied the primary needs of the human species. Once these needs were met, the superfluous energy that remained resulted in play. Because children were not responsible for their own survival, they were thought to have a total energy 'surplus,' which was depleted through play. Schiller raised a number of contemporary issues in his writings. First, he considered play a symbolic activity through which the participant could transform and transcend reality and thereby gain new symbolic representations of the world. Second, he distinguished between forms of play—material superfluidity resulted in physical play, esthetic superfluidity culminated in dramatic–symbolic play. Collectively, these notions reappear in the later writings of Piaget, Vygotsky, and Bühler.

According to the relaxation theorists such as Lazarus, the purpose of play was to restore energy expended in work. Thus, labor was viewed as energy consuming, resulting in an energy deficit. This deficit could be replenished through rest or sleep, or by engaging in play.

Subsequently recapitulation theorists posited that cultural epochs were repeated sequentially in children's play; the animal stage (children's climbing and swinging); the savage stage (hunting, tag, hide-and-go-seek); the nomad stage (keeping pets); the agricultural–patriarchal stage (playing with dolls, digging in sand); and the tribal stage (team games). They also believed that play served as an outlet for the catharsis or release of unnecessary, primitive racial instincts, thereby preparing individuals for the intellectually advanced activities of the modern era. For example, Patrick suggested that contemporary occupations required abstract reasoning, concentrated attention, and coordinated eye-hand activities, all of which were presumed to be relatively recent evolutionary acquisitions. Because this worked tapped recently acquired skills, it was more taxing than physical labor. As such, relief from fatigue could be gained through play, or the practice of 'racially old' activities (e.g., hunting, fishing).

According to pre-exercise or practice theorists such as Groos, the period of childhood existed so that the organism could play. Humankind's relatively long period of immaturity was considered necessary to allow for children to practice the instinctively-based complex skills that would be essential for survival in adulthood. Thus, the adaptive function of play was to prepare children to perfect skills that they would require in adulthood.

Of further interest, Groos noted that children's play comprised 'don't have to' activities—while playing, children are more interested in the processes rather than the products of the behavior. In this regard, recent speculations concerned with effectance motivation can be traced back to Groos' writings.

Importantly, Groos noted that children's play changed with development. First, there was experimental play, which included sensory and motor practice activities. Such play evolved into constructive play and the practice of higher mental powers. The purpose of such activity was to aid in the development of self-control. Second, there was socionomic play, which included fighting and chasing, as well as imitative, social and family games (dramatic play). This form of play was thought to aid in the development of personal relationships.

Despite obvious limitations, not the least of which are limited supportive, empirical bases, these early theories continue to have an impact on how many people think about play today. For example, parents and teachers are often heard to express reservations about poor weather conditions that might keep children indoors all day. The traditionally held belief is that such restriction constrains the expenditure of surplus energy.

2. *Modern Theories of Play*

Several common themes run across modern twentieth-century interpretations of children's play. These include the belief that: (a) children need to play in order

to express themselves or to relieve themselves of anxieties and fears; and (b) play both causes and reflects developments in social, cognitive, and linguistic prowess.

2.1 Psychoanalytic Theory

Freud (1961) believed that play provided children with important avenues for the expression of wish fulfillment and the mastery of traumatic events. He argued that play allowed the child to transcend the rigid sanctions of reality thereby serving as a safe context within which the child could vent socially unacceptable impulses. Freud addressed the mastery aspect of play through the repetition compulsion, a psychic mechanism that allows individuals to cope with traumatic events through a compulsive repetition of components of the disturbing events. Children were thought to use play to become active masters of situations in which they were once passive victims. For example, if her mother, for not tidying up her room, scolded a young child, the child may re-enact the scene numerous times with a doll, casting herself in the role of angry mother.

More recently, psychoanalytic theorists have expanded on Freud's conceptualizations of play. For example, with regard to wish fulfillment, Peller noted that children's choices of roles are often based on feelings of love, admiration, fear, or anger for a particular person. Such roles allow children to fulfill the wish to be like certain others. With regard to mastery, Erikson believed play served to allow the child to integrate biological and social spheres of functioning. Through play, children create model situations in which aspects of the past are re-lived, the present represented and renewed, and the future anticipated.

2.2 Piaget

Piaget (1962) suggested that play represented the purest form of assimilation. In assimilation, children incorporate events, objects, or situations into existing ways of thinking. Thus, as 'pure assimilation,' play was not considered an avenue to cognitive growth, but rather as a reflection of the child's present level of cognitive development.

Borrowing largely from Karl Bühler and Spencer, Piaget described three stages in the development of play. Practice play first appeared in infancy, consisting of sensorimotor actions (e.g., clapping hands). Through this 'functional exercise,' Piaget believed that children acquired and honed the basic motor skills inherent in their everyday activities. Symbolic play, appearing around the second year, required an implied representation of absent objects (e.g., pretending to bake a cake while in the sand box). In contrast with practice play, where actions were exercised and elaborated for their functional value, symbolic play allowed the exercise of actions for their representational value. Games-with-rules was the last structural category to develop; this type of play activity necessarily incorporated social coordination and a basic understanding of social relationships.

2.3 Vygotsky

Like Piaget, Vygotsky (1967) framed play within a larger psychological theory of children's cognition. Vygotsky argued that children used symbolic play as an essential link in the association of abstract meanings and their associated concrete objects. Symbolic play was useful in allowing children to conceive of meanings independently of the objects that they may represent. Thus, unlike Piaget, Vygotsky argued that play was not a reflection of egocentrism, but rather a symbolic process that brings into being the mediating role of signs. For example, when children first begin to use words, the word is perceived as a property of the object rather than as a sign denoting the object. The child grasps the external structure of the word–object relation earlier than the internal structure of the sign-referent relationship. The fusion of word with object follows the more general fusion of action and object. In infancy, the child relates to things as 'objects of action.' For higher mental processes to develop, things must become 'objects of thought' and practical actions must become mental representations (e.g., volitional choice). Play precipitates this emancipation of meaning from object to action. The central event responsible for the emancipation is the use of one object (e.g., a stick) to substitute for another (a real space ship), or the use of one action (a jump) to denote another (a space launch). To Vygotsky, movement in the field of meaning is the predominant feature of play.

2.4 Other Modern Theories

Drive modulation theorists proposed that excessively high and excessively low levels of stimulation are aversive; play is used as a means of modulating the arousal associated with this aversion. For example, when confronted with a novel object, specific exploration allowed the child to explore its features and relieve arousal through increasing familiarity with the object. Following specific exploration, an optimal level of arousal is sought through diverse exploration, or stimulus-seeking activity. This latter form of exploration, or play, increases stimulation when we are 'bored' and continues until arousal reaches an optimal level. Thus, play is viewed as a stimulus-producing activity that is generated by low levels of arousal.

Recent theoretical accounts have emphasized the role of play, particularly pretense, in the development of children's theory of mind (e.g., Lillard 1998), including the provision of opportunities to build and expand mental representations. Further, educational

thinkers have posited that creativity and flexibility is promoted by children's play (e.g., Bruner 1972). Accordingly, play allows the exploration of new combinations of behaviors and ideas within a psychologically safe milieu. Through play, children develop behavioral 'prototypes' that may be used subsequently in more 'serious' contexts. For example, a young child may button and unbutton her doll's dress many times, and thereafter incorporate her accomplishments from this play session when dressing herself. As such, the means become more important than the ends, and since accomplishing goals is not important in play, children are free to experiment with new and unusual combinations of behavior.

Finally, linguists have proposed ways in which play may help children perfect newly acquired language skills and increase conscious awareness of linguistic rules. Play provides a superior context within which children may gain valuable language practice as they experiment with the meaning, the structure, and the function of language (Davidson 1996). Play conversations also work to improve communication skills. These skills, in turn, are important components of many developmental acquisitions attained during childhood, particularly narrative representation, social cognition, intersubjectivity, and fantasy play. Goncu (1993), for example, has suggested that the improvisational processes typical of social pretend play are critical to the development of intersubjectivity (i.e., the development of a mutual understanding between play participants). These processes prepare children for an ever increasingly complex social life within which a variety of interactional contexts exist that range from the more ritualized and structured to the more improvisational. Within each of these interactional contexts, however, are elements of both improvisation and social structure (Sawyer 1997). Thus, among the unique properties of children's peer play are its framed and improvisational nature, each which the child must master.

3. *Issues in Defining Play*

It is one thing to think about *why* play exists in the human repertoire; it is something else altogether to define it. Following from Rubin et al. (1983), the following characteristics, when taken together, define play.

(a) Play is not governed by appetitive drives, compliance with social demands, or by inducements external to the behavior itself; instead play is intrinsically motivated.

(b) Play is spontaneous, free from external sanctions, and its goals are self-imposed.

(c) Play asks 'What can I do with this object or person?' This question differentiates play from exploration, which asks 'What is this object (or person) and what I do with it (him or her)?'

(d) Play is not a serious rendition of an activity or a behavior it resembles; instead it consists of activities that can be labeled as pretense

(e) Play is free from externally imposed rules—this characteristic distinguishes play from games-with-rules.

(f) Play involves active engagement—this distinguishes play from daydreaming, lounging, and aimless loafing.

4. *Developmental Progressions in Children's Play*

Between the periods of infancy and middle childhood, children's play undergoes an evolution in both form and content. These progressions are reviewed briefly below.

4.1 *Infant and Toddler Play*

By the end of the first year, infants begin to demonstrate rapid growth in representational thinking. Decontextualized behavior, which is first demonstrated in the first year, involves the 'out of context' production of familiar behavior. For example, the infant may close her eyes, put her head on a pillow, and lie in a curled position at a time of day (e.g., mid-morning), and in a context (e.g., playground) that is detached from the situational context when and where sleeping or napping occurs. By the middle of the second year, the toddler coordinates the use of several objects in his or her demonstrations of decontextualized behavior (e.g., a 'Teddy Bear' is fed from an empty cup).

This latter use of objects in pretense captures the essence of the second developmental component of play, self-other relationships. When pretense appears at about 12 months, it is centered around the child's own body (e.g., the child feeds herself). Roughly between 15 and 21 months, play becomes other-referenced; however, the 'other' is typically an inanimate object, as in the 'Teddy Bear' example noted above. Moreover, during this period, when others are involved in pretense activities, they are passive recipients of the child's behavior. Beyond 20 months, and increasingly so up to about 30 months, the child gains the ability to 'step out' of the play situation and to manipulate the 'other' as if it were an active agent (e.g., the Teddy Bear 'feeds' a doll with a plastic spoon). The developmental significance of these accomplishments should not be easily underestimated. Advances in maturity of play reflect the young child's increasing ability to symbolically represent things, actions, roles, and relationships.

A third component of play is the use of substitute objects. The ability to identify one object with another (e.g., a stick is used as a laser gun) is paradigmatic of symbolic representation. The fourth component of

play is the coordination and sequencing of pretense. Between the ages of 12 and 20 months, toddlers' pretend acts become increasingly coordinated into meaningful sequences. At first, the child produces a single pretend gesture (drinking from a plastic cup); later, the child relates, in succession, the same act to the self and then to others (drinks from the cup, feeds the Teddy Bear from the cup). Subsequently, in a multi-scheme combination, the young child is able to coordinate different sequential acts (pours tea, feeds the Teddy Bear, puts bear to sleep). By the end of the second year, children indicate verbally that these coordinated sequences are planned prior to execution (child self-verbalizes sequence of pretense behavior prior to acting).

4.2 *The Play of Preschoolers and Elementary School-age Children*

The above noted constituents of play are mastered prior to or near the child's second birthday. These elements of play become increasingly shared with others as children mature. But why is shared or social pretense important? As noted above, there are several functions of sociodramatic play. Such play creates a context for mastering the communication of meaning. It provides opportunities for children to learn to control and compromise; these opportunities arise during discussions and negotiations concerning pretend roles and scripts and the rules guiding the pretend episodes. Also, social pretense allows for a 'safe' context within which children can explore and discuss issues of intimacy and trust.

By 36 months, children are generally able to communicate pretend scripts to adults and peers; by five years, they can discuss, assign, and enact play themes while continuing to add novel components. By the middle years of childhood, social pretend becomes a venue for self-disclosure and the sharing of confidences especially among close friends.

5. *Correlates and Outcomes of Play*

5.1 *Play and Cognitive Development*

Children (three- to five-year-olds) who engage frequently in sociodramatic and constructive play tend to perform better on tests of intelligence than their age-mates who are more inclined to play in a sensorimotor fashion. Interestingly, children who frequently play in a constructive fashion (e.g., building things; constructing puzzles) are likely to be proficient at solving convergent problems (problems with a single solution). Those who frequently play in a dramatic fashion are likely to be proficient at solving divergent problems (problems with multiple solutions).

5.2 *Play and Social Development*

Because successful participation in social pretense requires many of the skills theorized to be associated with the achievement of competent peer relationships, this type of play is viewed as a marker of social competence from toddlerhood to the middle and late childhood years. Preschoolers who frequently engage in sociodramatic play are more socially skilled than their age-mates who infrequently engage in such activity. Moreover, results from various training studies indicate that instruction in sociodramatic play is associated with increases in cooperation, social participation, and role-taking skills (see Rubin et al. 1983 for a review).

5.3 *Play and Language Development*

The mechanisms by which play may aid in the development of linguistic competencies are straightforward. Children frequently play with the different forms and rules of language. This play may take the form of repeating strings of nonsense syllables (phonology), substituting words of the same grammatical category (syntax), or intentionally distorting meaning through nonsense and jokes (semantics). As a result, language play may help children perfect newly acquired language skills and increase conscious awareness of linguistic rules, as well as provide a superior context in which the child may gain valuable language practice.

Generally speaking, particular phases in the development of symbolic play and language tend to co-occur. For example, sociodramatic play appears to be an important factor in the development of oral language development and vocabulary, story production, story comprehension, communication of meaning, and the early development of literacy (Davidson 1996, Shore 1995). Indeed, it has been reported that training children to engage in pretense with others improves their language skills, literacy development, and mathematical thinking (see Fromberg and Bergen 1998 for relevant reviews).

6. *Summary and Conclusions*

It is clear that play is a developmental phenomenon of significant proportion. Not only does it seem to provide a window into the child's cognitive and socio-emotional being, but it also appears to be a propelling force for the development of cognitive, language, and socio–emotional skills. Thus, play should be considered an informal, enjoyable, and relatively stress-free means of providing children with intellectual and social stimulation.

See also: Cognitive Development in Childhood and Adolescence; Infant and Child Development, Theories of; Personality Development in Childhood; Play and Development in Children; Play, Anthropology of; Social Cognition in Childhood; Socialization and Education: Theoretical Perspectives

Bibliography

Bruner J S 1972 The nature and uses of immaturity. *American Psychologist* **27**: 687–708

Davidson J I F 1996 *Emergent Literacy and Dramatic Play in Early Education*. Delmar, Albany, NY

Freud S 1961 *Beyond the Pleasure Principle*. Norton, New York

Fromberg D P, Bergen D 1998 *Play from Birth to Twelve and Beyond: Contexts, Perspectives and Meanings*. Garland, New York

Goncu A 1993 Development of intersubjectivity in social pretend play. *Human Development* **36**: 185–98

Howes C 1992 *The Collaborative Construction of Pretend*. State University of New York Press, New York

Lillard A S 1998 Playing with a theory of mind. In: Saracho O N, Spodek B (eds.) *Multiple Perspectives on Play in Early Childhood Education*. State University of New York Press, Albany, NY, pp. 11–33

Piaget J 1962 *Play, Dreams, and Imitation in Childhood*. Norton, New York

Rubin K H, Fein G G, Vandenberg B 1983 Play. In: Mussen P H (ed.) *Handbook of Child Psychology: Socialization, Personality and Social Development*. Wiley, New York, Vol. 4, pp. 693–774

Sawyer R K 1997 *Pretend Play as Improvisation. Conversations in the Preschool Classroom*. Lawrence Erlbaum Associates, Mahwah, NJ

Shore C 1995 *Individual Differences in Language Development*. Sage, Thousand Oaks, CA

Vygotsky L S 1967 Play and its role in the mental development of the child. *Soviet Psychology* **12**: 62–76

K. H. Rubin

China: Sociocultural Aspects

China as a political entity must be distinguished from China conceived in cultural terms. Within the political boundaries of China are many groups that are not generally considered to be ethnically or culturally Chinese (Han), although they are Chinese citizens. By the same token there are significant populations generally considered to be ethnically Chinese living outside China's political boundaries—'overseas Chinese' (*huaqiao*). This article discusses China's nonHan peoples briefly, but focuses primarily on ethnic Chinese or Han peoples who constitute the vast majority of China's population. Topics addressed are ethnicity and identity, scale and complexity, local communities, the local impact of imperial and state institutions, family and kinship, religion and ritual, and gender. Throughout, attention is drawn to the conceptual challenges China poses for the discipline of anthropology.

1. Ethnicity and Identity

China's government recognizes officially 56 *minzu* or 'nationalities,' based on an evolutionary scale of material progress derived from Lewis Henry Morgan and Friedrich Engels. Some groups, notably Tibetans and Uighurs of Western China, and Mongols in the north, have historical claims to large geographical areas, and aspirations for political autonomy. Other groups (for example the 'Miao,' 'Zhuang,' and 'Yi') have been dispersed among Han and other *minzu* over many provinces in South, Central, and Southwestern China. Official recognition as a nationality confers certain legal privileges (most famously, exemption from the 'one-child' policy of population control), but also implies backwardness and legitimates paternalistic treatment from the government. In this respect, China's current government continues a self-consciously 'civilizing' mission with respect to its non-Han populations that dates to imperial times (Harrell 1995).

China's 'Han' peoples are by no means culturally homogeneous, but scholars are divided as to the degree to which Han Chinese can be said to share a common culture. For example, only about half of Han peoples are native speakers of 'Mandarin' or 'the common language' (*putonghua*). The rest speak one of about a dozen major 'dialects' (related languages, but often mutually unintelligible). Broadly speaking, the provinces of Southern and Southeastern China manifest the greatest linguistic diversity among Han Chinese. This diversity reflects the fact these areas were incorporated into the Chinese empire somewhat later than North and Central China. In addition, the preservation of regional cultures in the South and Southeast results from the fact that these areas have been relatively less frequently beset by dramatic population displacements caused by dynastic crises, droughts, floods, and other disasters than have areas in the North and West.

Mitigating regional and linguistic diversity is the fact that Han Chinese of all provinces share a common written language, elite traditions, and a long history of political unification under a succession of dynasties and into the era of nationalist and, now, communist governments. Moreover, although there is no clear consensus as to the precise defining characteristics of Chinese culture, many Chinese communities are marked by similar family institutions, popular religious customs, class relations, forms of corporate association, and broadly Confucian values emphasizing, most importantly, filial piety. Yet the issue of

China's cultural unity, or lack thereof, has long vexed anthropologists, and is increasingly becoming one of political moment in the Sinocentric world. Among China's national minorities, for example, there are movements among Tibetans and Uighurs for independence. By the same token, regional and dialect-based communalism is strong among Han Chinese in the Southeast.

At the time of writing, the potentially most volatile locus of separatist sentiment is Taiwan. Separated politically from the mainland during 50 years of Japanese colonial rule (1895–1945) and subsequently governed by an exile nationalist government, Taiwan's majority Han population—including speakers of the Southern Min dialect (*minnanhua*) shared with southern Fujian province, and of Hakka, another dialect widely dispersed in Southern China—increasingly asserts its political and, in some cases, cultural separation from China. These assertions are anathema to China's government, which considers Taiwan a renegade province. Given such circumstances, the issue of cultural unity and diversity has a political significance that extends well beyond the concerns of anthropologists.

2. *Scale and Complexity*

China's vast scale and complexity pose particular problems for anthropology because the discipline's crucial organizing concepts—culture and society—have developed historically in the study of small-scale societies. Anthropology's traditional claim to illuminate social-cum-cultural systems holistically—most explicitly in the functionalist traditions of Bronislaw Malinowski and A. R. Radcliffe-Brown—is more easily accomplished in ethnographic descriptions in which overarching institutions and consciousness of common identity do not reach much above the village or tribal level. Given such limits, the interpenetration of kinship, political, religious, and economic institutions is described and analyzed more easily than in China, where economic, cultural, religious, and political ties link metropolitan centers to distant rural locales. An additional complicating consideration is the historical depth of supra-local social organization: China's history of economic, political, and cultural integration at local, regional, and empire-/nationwide levels has exerted and continues to exert a strong influence on local life and institutions. China's significance for the discipline of anthropology lies most of all in the challenge this complexity poses for adapting and modifying conceptual and theoretical tools developed in analyses of societies of much more limited scale.

To this end, the path-breaking work of G. William Skinner, growing out of 'central-place theory,' is particularly significant (Skinner 1964–5, Skinner 1977). Skinner has developed a spatial-cum-temporal model of market-, town-, and city-centered regions in China in late imperial times that form an eight-tiered nested hierarchy culminating in nine 'macro-regions.' Ascendance in the central-place hierarchy (for example, from villages to market towns, to higher-level central places) is characterized by increasing specialization of economic function, social complexity, and cultural sophistication. By the same token, macro-regions are divided into more densely populated and economically concentrated cores and more sparsely populated, and less productive, peripheries. These distinctions are reflected in class characteristics, kinship and religious organization, and other social differences.

Interpenetrating the hierarchy of economic local and regional systems is a hierarchy of political administration. Although there exists a rough articulation of administrative and economic regional organization (for example, many district capitals are also important economic centers), these correspondences are far from perfect (for example, consumers often have access to more than one higher-level economic center, whereas every administrative district is discrete, with a single capital at the next higher level). The ramifications of regional analysis for more conventional styles of ethnographic work are fundamental. Put simply, Skinner's work obviates imagining China as being crudely dividable into distinctions such as 'rural' and 'urban,' or 'elite' and 'folk': From the vantage point of village-based ethnographic studies, it is necessary to situate a locale with reference to its place in both administrative and economic systems if one is to comprehend its place with reference to China as a whole. This caveat applies across the board of social relations and culture—kinship, religious and ritual organization, social class, gender relations, and demographic processes. The potential implications of regional analysis for Chinese anthropology are well recognized, although yet to be thoroughly assimilated. The existence of China's essentially 'global' scale of civilization antedating the emergence of the Western-centered 'capitalist world system' also poses a fundamental conceptual challenge to the recent growth of academic interest in processes of 'globalization.' At the very least, China should provide a comparative reference point to discussions that all too often imagine 'globalization' to be unique to Western history and/or capitalism.

3. *Local Communities*

Historically, China's population has been overwhelmingly rural (until recently about 85 percent could broadly be characterized as rural), although some of China's cities were probably the world's largest during the European Middle Ages. But, as noted above, 'rural' is a designation too crude to capture the cultural

and social distinctions between farming villages located in close proximity to large urban centers (with consequent access to urban markets and culture), and those located in distant peripheries. In brief, more centrally located rural locales tend to conform to more orthodox forms of social organization (for example, with leadership in the hands of Confucian literati), and more peripheral ones tend to be more variable and less orthodox (leadership often exercised by local 'strongmen'). Villages in many parts of China are 'nucleated' (dwellings grouped close together), although some areas (the Sichuan basin, for example) lack nucleated villages, with the population being dispersed. Typically, groups of villages are linked to a market town economically and culturally, and this may also serve as a focus for important local ritual activities, higher-level kin-based groups (e.g., lineage corporations), and other voluntary associations. Most classic ethnographic studies have focused on villages with some references to villagers' participation in market-town-focused activities.

The general picture that emerges from village-based ethnographies is one of considerable variation, but some significant commonalities are also discernible. Single-surname villages and strong corporate patrilineages seem to be relatively more common in the more productive and prosperous regional cores and in the South. Such villages are more likely to find ritual solidarity in lineage-based activities. In contrast, temples to local gods are the more frequent ritual focus of multi-surname villages. The contrast should not be overdrawn, however. Ancestor worship at domestic altars, although discouraged during the high tides of communist reform, seems to have been nearly ubiquitous among Han Chinese, and may be making a comeback in areas where it was repressed during the 1960s and 1970s. By the same token, some deities (Guanyin, for example) are worshipped nearly universally, and most locales produced some local gods closely associated with local lore and history.

Although some village-based ethnographies have a penchant for treating the village as a self-contained community, Skinner argues forcefully that the standard-marketing community was until recently the most important unit of rural social life (Skinner 1964–5). In frequent, sometimes daily, trips to markets, farmers had access to the cultural amenities of the town, interacted with traders, landlords, and other local dignitaries, and found occasion to worship and attend festivals at temples. Higher-level lineage corporations often built lineage halls in market towns, and voluntary associations with market-system-wide memberships also concentrated their activities there. Improved transport and economic development have to some degree eclipsed the traditional 'standard market towns,' thus expanding the horizons of China's rural population, and (at least in some areas) resulted in the emergence of what Skinner terms 'modern trading centers.' During the Maoist era, many of the functions of 'standard-marketing communities' were assumed by communes (often territorially isomorphic with them), but in the reform era, local markets are re-emerging rapidly (Skinner 1985).

4. *The Local Impact of Imperial and State Institutions*

In imperial times, the impact of the state on local life was mediated largely by local elites; imperial administration extended to the county level, but the maintenance of social stability depended on local leaders and institutions (Ch'u 1962). The imperial state managed to maintain its hegemony largely because local elites were thoroughly committed to the Confucian ideology upon which the legitimacy of both the state and literati leadership rested. This commitment, in turn, was sustained in part by the imperial examination system. Aspirants in the highly competitive examinations were required to master Confucian classics. Success came in the form of academic degrees and appointment to administrative office. The symbolic and material rewards for successful candidates redounded to the benefit of their kin and communities. Consequently, local wealth and institutional effort were invested in the production of successful examination candidates (for example, lineage corporations often established schools for its sons), on the one hand, and successful candidates were able to use their official positions to return wealth and prestige to their communities. Although the prospects for advancement through the theoretically meritocratic examination system were distant or unrealistic for all but the relatively wealthy (Ho 1962), its values exercised a strong hold on popular imagination in all of China's social classes.

Beyond the pervasive impact of the imperial examination system on China's class system and consciousness, state power affected localities primarily through taxation and social control. In these functions, too, county magistrates often relied on the mediating services of local elites, with whom they shared a similar class background and Confucian education.

The mediating role of local elites and institutions in integrating localities and the state began to unravel in late imperial times, and reached a crisis during the Republican era, when government policy tended to undermine local authority (Duara 1988). The momentum during the post-1949 era was to establish greater central control by binding local cadres and officials more firmly to the policies of the center, often undermining local leadership and institutions. The period of economic reforms beginning in the late 1970s has seen a general loosening of central control over individuals (as, for example, in the initiation of the famous 'household responsibility system') and, to some degree, over local governments. As a result, some of the forms of traditional communalism (for

example, territorial cults worshipping local gods) are re-emerging in some parts of China (Dean 1993, Jing 1996).

5. *Family and Kinship*

Patrilineal descent, virilocal residence, and equal inheritance among sons characterize the standard model of Han Chinese kinship. It is a standard model less in a statistical sense than in the fact that it serves as an ideal type; even families whose own organization does not conform are likely to acknowledge this ideal type as how things ought to be. In principle, a bride should leave her natal household and take up residence in the natal household of her groom—including his parents, unmarried siblings, married brothers, and their wives and children. Prior to 1949, 'five generations living in a single household' was widely held to be both an admirable and enviable achievement. But it was an achievement also recognized to be difficult to attain. Domestic strife, economic misfortune, mortality, or infertility could intervene to prevent the achievement of such a 'grand family.'

The characteristic ethos of Chinese family life is one of the dimensions of Chinese culture that has been examined most closely by ethnographers. The picture that emerges of the Chinese domestic cycle includes tensions between daughters-in-law and their mothers-in-law, stemming in part from their divergent interests with respect to maintaining a large, unified family structure. Daughters-in-law often desire to escape the authority of their mothers-in-law, and agitate for a family division; mothers-in-law exercise their often strong influence over their sons in the interest of keeping the extended family united (Fei 1939, Wolf 1968, 1972, Cohen 1976). By the same token, sons owe filial allegiance to their fathers, but they also owe their children as good a start in life as they can provide. These circumstances can set brothers' interests against each other (Freedman 1966). In the end, competing interests and domestic tensions can result in agreement to divide an estate. For many rural families, the most tangible and immediate manifestation of division was the setting up of separate cooking stoves. Members of the formerly united family continued to reside in the same building, but the separate stoves indicate separate budgets and, symbolically, a parting of ways.

As one might expect, by no means all families were of ideal-typical form. We have already noted some of the factors that might prevent achievement of the ideal of a 'five-generation' household. In addition, we now know that both uxorilocal marriages and 'little daughter-in-law' marriages were quite common in some parts of China (Wolf and Huang 1980). Uxorilocal marriages seem to have occurred most commonly as a strategy in families lacking a male heir. In such circumstances, a son-in-law might agree to marry into his wife's family. Because a married-in son-in-law was looked down upon (having at least partially abandoned his filial commitments to his own patriline), they typically were men of relatively lower socioeconomic status and background. Because such marriages self-consciously deviated from the ideal type, arrangements for inheritance, descent of children, and so forth would be stipulated in a formal marriage contract. (For example, children might be divided between the patrilines of both the husband and the wife; the groom might agree to change his own surname and become his father-in-law's adopted son; the groom might agree only to support the father-in-law in his dotage.)

'Little daughter-in-law' marriage involves adopting an infant or girl with the intention that she marry her foster brother when she comes of age. No longer widely practiced (they are now illegal), such marriages were common in some locales through the first half of the twentieth century. They were said to save the expense and trouble of a wedding, but Wolf and Huang contend that their real advantage lay most of all in the fact that mothers-in-law were able thereby to develop more amiable relations with their daughters-in-law, having raised them essentially as daughters. Wolf and Huang also argue that improvement of mother/daughter-in-law relations were won at the expense of conjugal ones; divorce rates and infidelity were higher among 'minor marriages' in northern Taiwan, and fertility was lower (Wolf and Huang 1980).

Contracts typically drawn at family division might stipulate that some portion of the estate remain undivided in the form of a corporation in memory of an honored ancestor (often the father of the sons dividing the estate). The income of the corporation typically might be shared among the descendents, used to build an ancestral hall, or to fund annual banquets in memory of shared ancestors. Corporations formed in this fashion are often referred to as lineages (Freedman 1958, 1966). However, lineage corporations could also be formed when a group of patrilineally related men decided to invest in shares and to purchase some common property. By the same token, unrelated parties might form similar corporations to help put the maintenance of a temple or any other enterprise on a stable financial footing. In short, although many lineages in China were organized as shareholding corporations with both ritual and (in some cases) entrepreneurial goals, similar corporations were also founded on bases other than kinship (Sangren 1984).

Lineage corporations, as noted above, tended to be more numerous in prosperous, core areas, especially in the Southeast. However, even in the absence of lineage corporations, patrilineal kinship ties were considered to be important. Consequently, some ambiguity attends to Chinese terms for 'lineage' (*congzu*, *zu*) because they might refer to a formal corporation or merely to those related to one patrilineally.

The nature and quality of affinal ties seems to have varied considerably by region and social class. Women sought to maintain ties with their natal families, and in some areas such ties were important in the development of the social connectedness (*guanxi*) so important to Chinese social relations. However, in terms of customary law, women retained few rights in their natal lines. Except for gifts bestowed upon them at marriage, women did not inherit. Moreover, after her death a women could be worshipped as an ancestor only in the patriline of her sons.

6. *Religion and Ritual*

Considered as philosophical or liturgical traditions, Confucianism, Daoism, and Buddhism have exercised an important historical influence on Chinese religion. However, these influences are not clearly distinguishable at the level of popular belief and practice. For example, because of its emphasis on 'filial piety,' ancestor worship is often considered to be in some sense 'Confucian.' However, ideas having to do with the afterlife of the soul, the nature of supernatural spirits, geomancy (the operation of unseen forces in the landscape), and communication with supernatural powers all play an important role in ancestor worship, just as they do in the worship of the local deities (often termed 'Daoist') central to territorial cults, and in the popular worship of Buddhist deities (Guanyin, for example). In other words, Buddhism, Daoism, and Confucianism are relatively distinguishable only in the contexts of monastic Buddhist institutions, the texts and practices of the ordained Daoist priesthood, and (arguably) the self-consciously Confucian writings of the official elite.

'Popular,' 'folk,' or 'local' religion is, as C. K. Yang argued influentially in his definitive treatise (Yang 1961), 'diffused' throughout the institutions of social life. Ancestor worship, for example, can be considered the religious dimension of domestic life. Similarly, territorial cults give form to local communities at levels ranging from neighborhood 'Place God' (*Tudi Gong*) shrines, village-level temples, market-based temples, upward to the City God temples found in administrative capitals (Sangren 1987). City Gods also played a role in the official rites of the imperial state. District magistrates were not only governors; they also officiated as priests in the state cult which culminated in the sacrifices to heaven performed by the emperor himself (Feuchtwang 1977).

Fundamental to popular religion is a belief in the power of supernatural spirits (*ling*); from the point of view of the majority of worshippers, it matters less whether a god or goddess is Buddhist or Daoist, heterodox or orthodox (from the point of view of the Daoist priesthood or Buddhist clergy), than whether it has a reputation for answering prayers and performing miracles. Individuals pray to gods for blessings for themselves and their families; community leaders organize ritual celebrations to the same gods in the hope of ensuring prosperity on behalf of the community. Indeed, such celebrations constitute one of the main entertainments of local life, punctuating the annual calendar.

The majority of Chinese gods are viewed as deified historical personages; their accumulated legends of posthumous divine intervention on behalf of individuals and communities play a crucial role in the dissemination of their cults. The cults of some gods are of strictly local provenance, while the cults of others (such as those of Guanyin, the 'Goddess of Mercy,' and Guandi, the 'God of War') are popular throughout China. In imperial times, successful local cults might grow to the point where they received imperial recognition, with the emperor claiming the authority to promote and demote such deities. Pilgrimages from local temples to distant centers tied local communities into wider ritual spheres and provided welcome opportunities to travel and see the world outside an individual's own locale.

Chinese polytheism is widely conceived in terms of a celestial bureaucracy that roughly mirrored the imperial state; many gods are viewed as supernatural governors of their assigned districts. Alongside such celestial officials, however, Chinese people worshipped a variety of mother goddesses, tricksterish imps, and even demonic figures.

At death the soul is imagined to journey through the underworld where it is judged and (if found guilty) punished for its deeds. One of the obligations of descendants is to perform rituals on behalf of the deceased to bribe these underworld officials and their demonic henchmen, thereby winning the soul's release. Both Daoist priests and Buddhist clergy may be employed to perform such funerary rites. Lonely ghosts (souls of those who die in tragic circumstances or who have no descendents to worship them as ancestors) are pitied and, in some cases, feared for the mischief they may bring down upon the living. Communities commonly propitiate such spirits during the seventh lunar month.

Communication with unseen powers is established by a variety of techniques. Individual worshippers can often cast 'moon blocks' or draw lots at temple altars in an attempt to discern a deity's response to their queries. 'Spirit writing' via an apparatus believed to be possessed by a god or spirit is another common technique (Jordan and Overmeyer 1986). Spirit-mediums possessed by gods speak directly in the gods' voices (Elliott 1955), and revelation through such mediums has played an important role in the production of hagiographies (Seaman 1987, Kleeman 1994).

There has been a significant growth of Western academic interest in Chinese popular religion in recent years. One of the reasons for this growth is the close

association between local social organization and collective ritual activity. Moreover, as Paul Katz argues, China's closest analog to Western 'civic society' or a 'public sphere' is most likely to be found in local temples and their rituals (Katz 1995).

7. Gender

The feminist movement in academia has inspired a large body of research and writing about women in China. In anthropology, much of this work has focused on domestic life. Among the most important conclusions is that, despite the ideological emphasis on patriarchy, women exercise a good deal of practical influence (Wolf 1972). Much of this influence is linked to mothers' close emotional relations with their children, especially their sons. Whereas Confucian ideology emphasizes the fundamental importance of father–son ties, 'filial piety' manifested in popular myth, entertainment, and ritual more often emphasizes children's affection for their mothers. Some analysts (e.g., Martin 1988) believe that women's views on life differ so radically from those of men that a distinctive female ideology exists, one that often contradicts the male, or official, ideology. Alternatively, differences in men's and women's views can be conceived as generated within the family system considered holistically as a productive process.

One of the problems confronting sinological anthropologists at the time of writing is to assess, on the one hand, the degree to which changes in gender ideology (the government now officially advocates gender equality) have improved women's lives—very little, according to some (Wolf 1985)—and, on the other hand, the degree to which changes in family organization (for example, those consequent on the 'one-child family' policy) have altered gender ideology.

See also: Area and International Studies: Development in Southeast Asia; East Asia, Religions of; East Asian Studies: Culture; East Asian Studies: Gender; East Asian Studies: Politics; East Asian Studies: Society; Historiography and Historical Thought: East Asia; International Migration by Ethnic Chinese; Kinship in Anthropology; Nationalism, Historical Aspects of: East Asia

Bibliography

Ch'u T Y 1962 *Local Government in China Under the Ch'ing*. Harvard University Press, Cambridge, MA

Cohen M L 1976 *House United, House Divided: The Chinese Family in Taiwan*. Columbia University Press, New York, NY

Dean K 1993 *Taoist Ritual and Popular Cults of Southeast China*. Princeton University Press, Princeton, NJ

Duara P 1988 *Culture, Power, and the State: Rural North China, 1900–1942*. Stanford University Press, Stanford, CA

Elliott A J A 1955 *Chinese Spirit-medium Cults in Singapore*. Department of Anthropology, London School of Economics and Political Science, London

Fei H T 1939 *Peasant Life in China: A Field Study of Country Life in the Yangtze Valley*. Kegan Paul, Trench, Trubner, London

Feuchtwang S 1977 School-temple and city god. In: Skinner G W (ed.) *The City in Late Imperial China*. Stanford University Press, Stanford, CA

Freedman M 1958 *Lineage Organization in Southeastern China*. Athlone, London

Freedman M 1966 *Chinese Lineage and Society: Fukien and Kwangtung*. Athlone, London

Harrell S 1995 Introduction. In: Harrel S (ed.) *Cultural Encounters on China's Ethnic Frontiers*. University of Washington Press, Seattle, WA

Ho P T 1962 *The Ladder of Success in Imperial China: Aspects of Social Mobility, 1368–1911*. Columbia University Press, New York

Jing J 1996 *The Temple of Memories: History, Power, and Morality in a Chinese Village*. Stanford University Press, Stanford, CA

Jordan D K, Overmyer D L 1986 *The Flying Phoenix: Aspects of Chinese Sectarianism in Taiwan*. Princeton University Press, Princeton, NJ

Katz P R 1995 *Demon Hordes and Burning Boats: The Cult of Marshal Wen in Late Imperial China*. State University of New York Press, Albany, NY

Kleeman T F 1994 *A God's Own Tale: The Book of Transformations of Wenchang, the Divine Lord of Zitong*. State University of New York Press, Albany, NY

Martin E 1988 Gender and ideological differences in representations of life and death. In: Watson J L, Rawski E S (eds.) *Death Ritual in Late Imperial and Modern China*. University of California Press, Berkeley, CA

Sangren P S 1984 Traditional Chinese corporations: Beyond kinship. *Journal of Asian Studies* **43**: 391–415

Sangren P S 1987 *History and Magical Power in a Chinese Community*. Stanford University Press, Stanford, CA

Seaman G 1987 *Journey to the North: An Ethnohistorical Analysis and Annotated Translation of the Chinese Folk Novel Pei-yu Chi*. University of California Press, Berkeley, CA

Skinner G W 1964–5 Marketing and social structure in rural China, Parts I, II, and III. *Journal of Asian Studies* **24**: 3–43, 195–228, 363–99

Skinner G W 1977 Cities and the hierarchy of local systems. In: Skinner G W (ed.) *The City in Late Imperial China*. Stanford University Press, Stanford, CA

Skinner G W 1985 Rural marketing in China: repression and revival. *The China Quarterly* (September): 393–413

Wolf M 1968 *The House of Lim: A Study of a Chinese Farm Family*. Appleton-Century-Crofts, New York

Wolf M 1972 *Women and Family in Rural Taiwan*. Stanford University Press, Stanford, CA

Wolf M 1985 *Revolution Postponed: Women in Contemporary China*. Stanford University Press, Stanford, CA

Wolf A P, Huang C S 1980 *Marriage and Adoption in China, 1845–1945*. Stanford University Press, Stanford, CA

Yang C K 1961 *Religion and Ritual in Chinese Society: A Study of Contemporary Social Functions of Religion and Some of Their Historical Factors*. University of California Press, Berkeley, CA

P. S. Sangren

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Chinese Law

China's modern legal institutions reflect powerful political, economic, and social forces that have struggled to shape them since the latter part of the nineteenth century. For thousands of years, before the last dynasty fell, cultural values in the world's oldest continuous empire were inhospitable to ideals of 'rule of law' that evolved slowly in the West. In the Republic (1911–49) formal legal institutions were only superficially borrowed from abroad. From 1949 to 1979 the Maoist party-state reduced law to the merest tool of totalitarian politics. Mao's successors launched economic reforms in 1979 that have led to the creation of the most significant legal institutions in Chinese history, but at the end of the twentieth century traditional values endured, the ideology and the apparatus of the party-state remained in place, governmental institutions were disorderly, and the economy in flux. The new institutions can grow in strength and legitimacy only if they receive sustained and powerful political support from China's leadership and greater recognition in the legal culture of China's officials and its populace.

1. *The Chinese Legal Tradition*

Contemporary Chinese institutions should be viewed in light of profound differences between Chinese and Western legal history. Imperial China blended law and morality in contrast to Western Europe, where secular and sacred authority were separated early. The dominant cult and philosophy of Confucianism emphasized governance by men who acquired moral authority by emulating ancient sages in setting virtuous examples of benevolence and social rightness for their subjects to follow. Law was regarded as a set of inferior norms that supplemented more basic principles, especially rules of propriety (*li*) that differentiated individuals according to their status as determined by age and rank in family and society. Confucianism was briefly rivaled by the early philosophical school of Legalism, which stressed the need for harsh penalties using positive law (*fa*) to deter wrongdoing, but both schools shared a vision of society in which proper behavior derived from an individual's status in the hierarchies in which he or she lived.

Law was first codified in the Qin Dynasty (third century BC) and recodified and augmented by a complex body of regulations in subsequent dynasties, notably in the Tang and Ming. Principally penal, it unambiguously reinforced ideas of hierarchy and subordination and was addressed to officials, not to the populace. It was enforced by local county magistrates without legal training or expertise, as part of their general duties to govern on behalf of the emperor. Specialized institutions for adjudication like the centralized judicial systems that developed over centuries in the West were absent, although there were unofficial legal specialists at provincial and central levels. The outcomes of cases had to be substantively correct according to both law and Confucian morality. The concerns for procedural justice and uniformity of results that have come to mark Anglo-American law were absent.

The Chinese state exercised its rule mostly in an indirect manner, through local elites—landowners, family heads, and village elders—which enforced local customs. The concept of personal rights did not develop because the basic units of society were not individuals but rather the collectivities of family, clan, village, and guild. Economic transactions arose and were enforced largely in the context of custom-governed relationships. The official philosophy exerted strong social pressure in favor of mediation and compromise. Litigation before the magistrates was time-consuming, degrading, and costly. Civil disputes were common, nonetheless, but most were settled extra-judicially. Unlike the West, where lawyers emerged, any tendency for legal specialists to act as intermediaries between individuals and the state was actively discouraged, although in the late Qing men who facilitated litigation, albeit tarred as 'litigation tricksters,' did flourish.

From the mid-nineteenth century until 1949, when the People's Republic of China (PRC) was established, sporadic and inconsistent attempts were made to transplant foreign legal institutions. These failed to take root since they were often too complex as well as being irrelevant to Chinese conditions. China established its first professional bar during the Republican period, but lawyers' training and qualifications were uneven and their standards of professional behavior low. Judges were both few and poorly educated, and judicial professionalism and independence were weakened by corruption and favoritism. The authoritarian Nationalist Party undercut the spirit of the new legal reforms.

2. *The People's Republic of China*

2.1 *Maoism, 1949–79*

Under Mao Zedong the Chinese Communist Party (CCP) mounted extensive programs of economic reconstruction and social change, relying on both previous experience in ruling large areas before 1949 and on Soviet models, which entwined state institutions with the Party. Law was used as a political tool, along with mass organizations and propaganda media, to mobilize the populace to carry out policies. The criminal process was declared an instrument to exercise 'dictatorship' over members of the former

exploiting classes, and sanctions frequently varied according to political 'campaigns' and policies of the moment. Systematized codes of criminal law or criminal procedure were lacking, and the courts merely formalized findings of guilt by the police and the procuracy, which was a prosecutorial agency. The police could impose sentences of as long as four years without any judicial involvement.

The collectivization of almost all private property left little scope for noncriminal law. In the planned economy, disputes between state-owned enterprises were resolved informally through flexible, highly pragmatic attempts to adjust problems without fixing legal blame. Disputes among individuals were usually dealt with through mediation. Mediation committees that formed part of a totalitarian control apparatus which penetrated deeply into Chinese society were charged with attaining politically correct results that would benefit socialist construction and strengthen the party-state's control over 'bad elements.' Politicization was not total, of course; traditional attitudes among mediators and the populace persisted.

Mao and other leaders, determined to speed continued revolution and social change, refused to regularize law and administration. The Cultural Revolution (1966–76), further reduced the relevance of legal institutions that had already been politicized. After the overthrow of the 'Gang of Four' in 1976, the lawlessness of the Cultural Revolution moved their successors, led by Deng Xiaoping, to advocate adoption of orderly legal institutions. An era of reform began, and since then law has risen to greater prominence than ever before in Chinese history.

2.2 Reform Since 1979

Legal reform has been driven by the economic reforms that began in 1979 and have unfolded irregularly but irreversibly. A growing and increasingly differentiated non-state sector has been created and modernization has been aimed at constructing a 'socialist market economy,' including a legal system. China's departure from previous Maoist disregard for formal legislation has led to one of the greatest outpourings of legislation in history.

The structure of the Chinese state has been defined in a Constitution and in 'organic' laws dealing with key state institutions such as the courts and central and local legislative bodies; the General Principles of Civil Law, a partial civil code intended to mature into a comprehensive one; 'basic' laws such as codes of criminal law and criminal procedure; and enactments by the central government, subnational units, and by central ministries and their local branches. These define newly recognized economic relationships and participants in expanding markets and address regulatory problems generated by economic reform. Much legislation, however, has been incomplete and *ad hoc* as new economic policies appeared, reflecting the difficulties that the leadership has had in defining the direction of economic reform. There is a continuing struggle between concepts of law as a framework for economic activity by autonomous actors and as an administrative instrument.

Extensive legislation also signals China's participation in a global economic community. Direct foreign investment has been addressed by legislation on various investment vehicles and their operation, as well as on such matters as intellectual property, labor, customs, foreign exchange, bank lending and guaranties, and export and import licenses. A taxation system has been established. China has acceded to an extensive range of international agreements, such as the UN Convention on the International Sale of Goods, whose rules have become part of Chinese law.

2.2.1 Building a judicial system. The courts, formerly scorned as 'rightist' institutions at the end of the 1950s and as 'bourgeois' during the Cultural Revolution, have been rebuilt in a four-level hierarchy. The number of civil and economic disputes brought to the courts rose from 2.4 million cases in 1990 to 5.7 million in 1999, while the number of disputes brought to mediation committees declined from 7.4 million in 1990 to 5.1 million in 1999. Growing reliance on contracts and the increase in litigation suggests increasing acceptance of concepts of law-based rights.

The Chinese judicial system presents many problems at the beginning of the twenty-first century. Judges are poorly trained, and most still lack a complete legal education despite efforts to raise their educational qualifications. Over half of the cases brought to the courts are resolved through judicial mediation rather than adjudication of competing claims and rights. Judges often prefer to resolve cases by mediation to avoid reversal by a higher court; lower courts, fearful of being reversed, sometimes request instructions from a higher court before they issue a judgment, thereby rendering meaningless the right of unsuccessful parties to appeal. The finality of judgments is impaired by legislation permitting noncriminal decisions to be reopened within two years after they become effective. Sometimes, too, higher courts reviewing the quality of the work of lower courts reopen decisions even though they have already taken legal effect. The role of judges has been defined only ambiguously, and adjudication has not been significantly differentiated from decision making by administrative agencies in the course of implementing policies.

The independence, powers, and effectiveness of the courts have been constrained by the requirement that they follow CCP policies. In 2000, over 70 percent of

judges were members of the CCP and the principal affairs of the courts, including personnel matters, were directed by Party organizations. Judges are appointed and their salaries paid by the local governments in the jurisdictions in which they serve, leading to 'local protectionism' that frequently influences the outcomes of litigation. In addition, *guanxi* (relationships), corruption, and bribery are often employed to influence outcomes that on many occasions pervert justice.

2.2.2 The legal profession. The bar was formally reestablished in 1980 following a twenty-year hiatus after a brief experiment with a Soviet-style bar was ended, and legal education was revived. There were over 120,000 lawyers in 2000, but the educational level of many older lawyers is low and legal education remains highly formalistic. By 2000 China had over 8,000 law firms, most of which were state-run, but the number of 'cooperative' firms was growing. The sudden expansion of the legal profession created enormous temptations for lawyers, judges, and officials to engage in bribery and other corrupt practices. The state continued to regulate and scrutinize lawyers' activities, and a major unresolved contradiction existed between the concept of a professional bar and CCP opposition to autonomous organizations and professions.

2.2.3 The criminal process. The criminal process continues to be a tool for the politicized administration of law, as when political leaders focus it on activities deemed to 'endanger the security of the state,' or on other particular types of criminal activity. The formal rationality of the criminal process has been slowly increased in a criminal code and code of criminal procedure, but the extensive power of the police and the CCP over the criminal process have been only ineffectually restrained, and the police-administered system of sanctions begun under Mao remained in place in 2000.

2.2.4 Administrative law. The Chinese leadership began in the 1990s to address the need to create legal institutions that might curb bureaucratic arbitrariness. A series of laws gave affected persons or organizations the right to sue agencies that have acted unlawfully, defined the wide assortment of punishments that may be imposed by administrative agencies, and recognized situations in which governmental agencies may be liable for injurious consequences of their acts. However, the jurisdiction of the courts and their power to restrain arbitrariness remained limited. Chinese laws and administrative rules have generally given agencies very broad discretion, while judicial control of administrative action has been limited by a reluctance to allow courts to review the validity of general rules issued by administrative agencies or to decide that they had improperly used their discretion. The courts were at best only at the same level of authority as the other institutions of the state apparatus; their limited reach reflected the subordination of law to the bureaucracy.

2.3 The Future

The development, shape and meaningfulness of Chinese legal institutions will depend upon critical factors that lie outside the law.

2.3.1 Policy. As long as the CCP rules China, its policy will have to overcome its ambivalence toward the role—and rule—of law. The Chinese Constitution, amended in 1999 to declare that 'The People's Republic of China shall be governed according to law and shall be built into a socialist country based on the rule of law,' also affirmed 'the leadership of the Chinese Communist Party, Marxism-Leninism and Mao Zedong Thought, and "Deng Xiaoping Theory"' as 'guiding principles.' The leadership has tended to equate law with discipline and to treat it as an instrument to maintain the dominant role of the CCP in Chinese society. At the same time, it has also recognized that law can further rationalize decision making and implementation of policy, while increasing legitimacy at home and abroad.

2.3.2 Structural problems. At the beginning of the twenty-first century institutions for law making heavily reflect the imprint of pre-reform doctrine and practice. Three principal law-making agencies share the central government's legislative power under the Constitution adopted in 1982—the National People's Congress, its Standing Committee and the State Council—but their respective powers were only very generally defined and subject in practice to informal negotiations. The State Council, at the head of the executive branch of the central government, together with the ministries, commissions, and bureaus that are subordinate to it, possesses broad power to generate rules superior to all local enactments.

Although subnational units and more than 20 functional bureaucracies of the central government issue regulations, distinctions among the rules that they issue, and between rule making and implementation, are blurred. No effective mechanism has existed to measure legal norms for consistency with higher-level norms, and the Constitution is not justiciable. Chinese administrative agencies have exercised the power both to issue and interpret their own rules, and to require the courts to enforce them. The lower courts are formally denied power to interpret laws, although in practice the Supreme People's Court has asserted a

strong role in the interpretation and clarification of laws. These problems are aggravated by the frequently provisional nature and tentative style of legislation.

Although the dramatic Chinese economic reform at the end of the twentieth century was greatly promoted by increasing the power of local governments, it also facilitated extensive interpenetration of government and business. It was carried out without carefully defining property rights, and local governments took advantage of the uncertainty to form alliances with private enterprises. Local government involvement in enterprises varied from disguised ownership to acceptance of payoffs and bribes. Overall, in the year 2000, marketization was often not synonymous with privatization, and some Western scholars perceived the emergence of corporatist relationships that closely linked non-governmental actors to local governments. Also, the growth of local power increased local deviations from central state policies, and undermined uniformity in the application of legal rules and the reach of the central government generally.

Legal development remains tied to economic reform; its vigor will in turn not only on the strength of the economy and the consolidation of reforms already accomplished, but also on whether solutions are fashioned to deal with basic problems that have been difficult to surmount. The state sector of the economy has been governed by rules and practices to which legal rules are essentially irrelevant. Although reform has long been a goal of the leadership, it has been unable to chart a clear course between privatization and continued dedication to state ownership, or to create mechanisms to deal with the large-scale unemployment and associated social distress that further industrial reform would generate. The establishment of enforceable rules on the governance of formerly state-owned firms that have been converted remains problematic. Financial system reform began in the last years of the twentieth century but was incomplete and faced serious obstacles; and both creation and regulation of capital markets were still in early stages.

2.3.3 The crisis of values. Reform dramatically relaxed state control over the lives of many Chinese in noticeable ways, but it also created severe social dislocations including income disparity and an impoverished 'floating population' of as many as 100 million peasants who have flocked to China's cities seeking employment. Discontent has risen, among peasants angry at their exploitation by local cadres and among unemployed workers at state-owned enterprises that have closed down. Crime, violent and otherwise, has risen, provoking widespread concern about social order.

The profound political and economic changes of the last two decades of the twentieth century unsettled both traditional and Communist values. The Party's ideology became hollow and its legitimacy increasingly questioned, while the opening of China to the rest of the world exposed the Chinese people to new values and ideas, including Western concepts of legality. The weakening of the totalitarian grip on individual lives and the continuing flux of economic reform have fostered the re-emergence of an emphasis on personal relationships and clientelism. Corruption has grown despite continued efforts by the leadership to check and punish its many manifestations, and has aroused alienation and cynicism among many Chinese.

2.3.4 Chinese legal culture: continuity and change. Chinese legal culture continues to reflect competing currents. Traditional values remain strong. Many Chinese remain unwilling to take their disputes to courts, choosing to rely on personal relationships or to defer to authority. In the courts, concern for procedural justice is weak. Bureaucrats continue to want to enjoy broad discretion. At the same time, the extensive social and economic changes sparked by reform have promoted consciousness of legal rights and willingness to use legal processes to assert such rights. Lawsuits against government agencies are increasing, although they remain relatively small in number, and peasant and worker protests often invoke published laws and policies to resist official behavior that they consider to be unjust. Some Chinese legal scholars, officials and intellectuals have called for a legal system with a national and autonomous judiciary that applies standards of procedural fairness. Some economic actors in the non-state sector desire stronger protection of their transactions by rules enforced meaningfully and consistently by the power of the Chinese state. Despite the resistance of the CCP to the growth of civil society, the continuing development of non-state economic activity, the strength of communal traditions and the tenacity of some nongovernmental organizations in Chinese society could combine to advance the development of legal consciousness.

3. *Conclusion: Perspectives*

It should not be assumed that legal development will lead to Western-type institutions or to liberal democracy. The domain considered 'legal' and the boundaries between it and other areas of Chinese state, society, and economy will not necessarily converge with Western concepts, and rights may remain 'soft.' Nor should Western observers overstate the supposed virtues for China of Western concepts and institutions, themselves imperfect and under question.

Twenty years of reform efforts began a journey toward greater legality, and further efforts are underway at the beginning of the twenty-first century to advance judicial reform, add coherence to Chinese law making, and develop administrative law further. The

accession of China to the World Trade Organization would impose on China international obligations to increase the transparency of government and the reach of legal institutions. The General Agreement on Tariffs and Trade and other agreements that are implemented by the WTO require all member nations to adopt and implement their laws relating to trade in a manner consistent with the rule of law as it is understood in the West, and China will have to adjust its legal institutions to comply with WTO standards. The deepening of legal reform most depends, however, on political reform. Legal institutions will be hobbled by political constraints as long as any Chinese leadership, Communist or post-Communist, maintains an instrumental view of law, remains ambivalent about the rule of law, and inhibits the growth of an active civil society. For law to grow more meaningful, leadership policy and official ideology must enlarge the domain of the law, end or greatly dilute one-party domination, and remedy institutional weaknesses in the structure of the Chinese state. Even the strongest political commitment will require considerable time to implement further reform, overcome the serious limits on state capacity that make the governance of China difficult under any circumstances, and inspire popular confidence in institutions. The processes of institutional change that have begun can only work slowly at best.

See also: China: Sociocultural Aspects; East Asian Studies: Economics; East Asian Studies: Politics; East Asian Studies: Society; Globalization: Legal Aspects; Law and Society: Sociolegal Studies; Law as an Instrument of Social Change; Legal Culture and Legal Consciousness; Mediation, Arbitration, and Alternative Dispute Resolution (ADR)

Bibliography

Bodde D, Morris C 1967 *Law in Imperial China*. Harvard University Press, Cambridge, MA

Chen J 1999 *Chinese Law: Toward an Understanding of Chinese Law, Its Nature and Development*. Kluwer Law International, The Hague, The Netherlands

Corne P H 1996 *Foreign Investment in China: The Administrative Legal System*. Hong Kong University Press, Hong Kong

He W 1995 Tongguo sifa shixian shehui zhengyi: dui zhongguo faguan xianzhuang de yige toushi [The realization of social justice through judicature: a look at the current situation of Chinese judges]. In: Xia Y (ed.) *Zou xiang quanli de shidai: Zhongguo gongmin quanli fazhan yanjiu* [*Toward a Time of Rights: A Perspective of the Civil Rights Development in China*]. China University of Politics and Law Press, Beijing, People's Republic of China, pp. 209–84

Huang P C 1996 *Civil Justice in China: Representation and Practice in the Qing*. Stanford University Press, Stanford, CA

Keller P 1994 Sources of order in Chinese law. *American Journal of Comparative Law* **42**: 711–59

Lawyers Committee for Human Rights 1996 *Opening to Reform?: An Analysis of China's Revised Criminal Procedure Law*. Lawyers Committee for Human Rights, New York

Lawyers Committee for Human Rights 1998a *Wrongs and Rights: A Human Rights Analysis of China's Revised Criminal Law*. Lawyers Committee for Human Rights, New York

Lawyers Committee for Human Rights 1998b *Lawyers in China: Obstacles to Independence and the Defense of Rights*. Lawyers Committee for Human Rights, New York

Lubman S (ed.) 1997 *China's Legal Reforms*. Oxford University Press, Oxford, UK

Lubman S 1999 *Bird in a Cage: Legal Reform in China After Mao*. Stanford University Press, Stanford, CA

Turner K G, Feinerman J V, Guy R K (eds.) 2000 *The Limits of the Rule of Law in China*. University of Washington, Seattle, DC

Van der Sprenkel S 1962 *Legal Institutions in Manchu China: A Sociological Analysis*. Athlone, London

Xia Y (ed.) 1995 *Zouxiang quanli de shidai: Zhongguo gongmin quanli fazhan yanjiu* [*Toward a Time of Rights: A Perspective of the Civil Rights Development in China*]. China University of Politics and Law Press, Beijing, People's Republic of China

S. B. Lubman

Chinese Revolutions: Twentieth Century

In the 1800s, a leading Sinologist claimed the Chinese were the 'most rebellious' but 'least revolutionary' of peoples and Karl Marx, along with other Western social theorists, argued that peasants would never be the driving force in radical movements for change. The twentieth century would prove both the Sinologist and the theorists wrong—many times over. Nearly every decade of it saw a revolutionary event of one kind or another break out in China. In many of these, peasants played central roles.

1. Key Events

The first quarter of the twentieth century witnessed a series of inter-related insurrections that toppled the Qing Dynasty (1644–1911). Known collectively as the 1911 Revolution, these paved the way for the establishment of the Republic of China (ROC) on January 1, 1912, which was also when a revolutionary activist, Sun Yat-sen (1866–1925), was inaugurated as this new country's first President. A path breaking student-led mass struggle also took place in the century's first quarter. This was the May 4th Movement of 1919, which ended with the dismissal from office of three high-ranking officials. In addition, two revolutionary organizations were formed in this period: the Nationalist Party or Guomingdang (GMD), which was led initially by Sun and then after his death by Chiang Kai-shek (1887–1975); and the Chinese Communist Party (CCP) of Mao Zedong (1893–1976) and Deng Xiaoping (1904–1997).

The second quarter of the century began with the GMD and CCP working together in a United Front

(1924–1927) orchestrated by Sun and his Soviet advisor Borodin. The alliance's goal was to defeat warlordism and imperialism and thus get the *Geming* (a Chinese term for revolution whose shades of meaning are returned to below) back on track. In 1927, however, Chiang turned on his erstwhile allies, who had just helped him reunify the country through the military campaigns and mass movements known as the Northern Expedition. He launched a White Terror against members of the CCP and left-wing rivals within the GMD. Seven years later, after persistent GMD extermination campaigns, the CCP set out on its epic Long March (1934–1935) to safety in isolated northern base areas. The two organizations, each tightly disciplined and Leninist in structure, allied again from 1937 to 1945 in a Second United Front against Japan. The 1925–1950 period concluded, however, with a Civil War (1945–1949). This ended with the GMD's retreat to Taiwan and Mao's proclamation, on October 1, 1949, of the establishment of a new People's Republic of China (PRC)—something that this son of a relatively well-to-do farmer was only able to do because his party had won so much peasant support. The CCP then launched a series of radical initiatives, including a partially successful effort to gain adherence to a new Marriage Law (1950) that gave women equal rights in family matters and strove to minimize the power of patriarchal kin-groups.

The century's third quarter began with Mao taking the lead in further campaigns, some of which caused enormous misery. The most notorious included the 'Anti-Rightist Campaign' (a Red Terror directed largely against intellectuals) and the 'Great Leap Forward' (a bizarre experiment in irrational utopianism that contributed to a famine of horrendous proportions). Later in this quarter-century, Mao, angered by his post-Great Leap marginalization within the CCP hierarchy, rallied young loyalists (Red Guards) to challenge the Party bureaucracy. Thus began the Cultural Revolution (1966–1976). Mao's charge, soon echoed by the Gang of Four (a clique that included Jiang Qing, his wife), was that high officials had lost their revolutionary zeal and become corrupt.

The Cultural Revolution spiraled into a chaotic war of all against all that continued into the first year of the century's final quarter. This period also witnessed the abortive revolution of 1989, during which giant demonstrations took place in Beijing and other cities. In the ROC (as Taiwan was re-baptized in 1949), a major breakthrough came with the lifting of a decades old policy of martial law. This laid the groundwork for an East Asian variation of the 'Velvet Revolutions' of Central Europe: the nonviolent transfer of power from the GMD to an opposition party early in the year 2000.

By the time the 1900s ended, so many aspects of Chinese politics and society had been transformed or challenged by insurgents that it now seems strange that the first edition of this Encyclopedia included an entry on 'The Chinese Problem' but not 'The Chinese Revolution.' The rise of the CCP and the Cultural Revolution still lay in the future, but there were clear signs that the twentieth century would be a revolutionary one for China. One indication of just *how* revolutionary is that, by the 1990s, many Chinese were living lives that would have been unrecognizable to their great-grandparents or even their parents. Moreover, the words 'China' and 'revolution' were by then inextricably linked in the minds of many people.

This linkage was particularly strong in the PRC where the term *Geming* continues to have a sacred patriotic meaning. This is of strategic significance to the CCP, which is still struggling, as this entry is being written, to ride out an ongoing legitimacy crisis. The regime continues to find it useful to play the revolutionary legacy card periodically. It did so, for example, in 1999 when American bombs mistakenly killed three Chinese journalists in Belgrade. The victims were quickly dubbed 'revolutionary martyrs' and the official media linked these new deaths to those of patriotic participants in hallowed historic events of the 1910s–1950s.

In addition, even though Westerners often view the words 'reform' and 'revolution' as contrastive, in China *gaige* (reform) is often presented as a method for carrying forward the *Geming*. The policy of opening the PRC to international trade and moving away from a command economy that Deng and his successors have followed since 1979 is typically presented in English language works as a pragmatic 'Reform Program' and a step away from state socialism. Its PRC proponents, however, describe it as an effort to 'build socialism with Chinese characteristics' and an effort to keep the Revolution from ossifying.

It is not just in the PRC that the words 'China' and 'revolution' go together naturally. Most of the few Chinese names and faces recognized by foreigners are those of insurgents—from Mao to the anonymous 'man-who-stopped-the-tanks' in 1989. And many ROC residents grew up learning history from schoolbooks that treated the story of the 1900s as an unending revolutionary struggle, which encountered a terrible setback in 1949 but did not die.

2. *Categories and Definitions*

This brings us to a question too rarely addressed: did China experience one Revolution or several? Many official PRC histories (as well as some Marxist ones in other languages and standard GMD accounts) refer to a single ongoing Revolution with many stages. Western social theorists (such as Theda Skocpol) and comparative historians (such as Crane Brinton) sometimes prefer to think of a single great event that began in 1911 and ended in 1949. They then fit this almost 40-year long event into a Great Social Revolution model

that is also used to interpret more compact phenomena such as the French upheavals of 1789–1799. Many non-Marxist Sinologists, meanwhile, speak of three relatively discrete revolutions: a Republican one in 1911, a National one in the 1920s, and a Communist one.

2.1 The Term 'Geming'

Complicating these questions is the term *Geming*, the meaning of which at times overlaps with, at others diverges from its Western equivalents. *Geming* literally means 'stripping the mandate,' that is, the successful carrying out of an upheaval that demonstrates, through its ability to overturn a ruling house and establish a new dynasty, that heaven now sides with a new regime. It was only from the 1890s on, in part because of new senses the characters had acquired in Japan, that *Geming* began to be linked to the creation of a new political system, not just completion of a dynastic cycle. *Geming*'s original linkage to dynastic cycles makes it curiously like its English and French counterparts, in which visions of return and of forward movement alike can be found. In 1789, the Janus-faced aspect of revolution was signaled by the use of classical goddesses to portray new political virtues. In the 1910s, it was signaled by anti-Qing leaders alternating between presenting themselves as avenging the legacy of the Ming (1368–1644), the ethnically Chinese ruling group the Manchus had displaced, and as struggling to create a new system.

If there are overlaps between the terms 'revolution' and *Geming*, however, there is also a difference: it is never clear whether the sense of the Chinese word is singular or plural, lower or upper case, whether a text is referring to The Revolution or one of many. This is more than merely a linguistic issue: to compare events, we need to know when they began and ended. With France, for example, there is a consensus that the revolution began in 1789 and lasted about a decade, though other supplemental revolutions occurred later (1830, 1848, and 1871). With China, there are more options, including the three alluded to above.

2.2 Problems with Standard Categorization Schemes

There are, however, limitations to each. The continuous Revolution vision focuses too tightly on the achievements of a particular organization. It distorts the shape of twentieth-century history by downplaying struggles not guided by the particular Leninist party—the CCP or GMD—being celebrated.

The 1911–1949 vision is problematic for use in comparison—even though comparativists champion it. Skocpol, for example, does not explain why, if the 'Great Social Revolutions' are essentially alike in structural terms, the Chinese one took so much longer to play out. Brinton's 'anatomy of revolution' model becomes problematic because 'terrors' occurred at different points in the Chinese *Geming*. In other revolutions, he claims, this 'fever' hit the patient just once, then broke.

A tripartite schema, finally, leaves too little room for radical developments that do not move in lockstep with the rise of governments. 1911, the late 1920s, and 1949 were key turning points in the political story but not necessarily the social and cultural ones.

3. Overlapping Revolutions

A more satisfying approach divides China's recent past, for heuristic purposes, into four overlapping revolutions, each following its own timetable: one political, one cultural/intellectual, one diplomatic, and one socioeconomic. Let us consider each briefly.

3.1 Politics

The political revolution has the clearest starting date: October 10, 1911. GMD official historians present the events that followed the Wuchang Uprising of that day (really a military mutiny) as adhering to a master plan crafted by Sun Yat-sen and carried out by members of his Revolutionary Alliance, the organization that evolved into the Nationalist Party. But many participants had no contact with the Alliance and knew little of Sun or his ideas. Only some were committed to republican ideals. Others were motivated by distrust of the Qing 'barbarians,' anger at particular officials, or a desire to be on the winning side of a Mandate shift.

Still, the 1911 Revolution extinguished not just a dynasty but a system, hence it deserves to be thought of as politically revolutionary. The National Assembly and other republican institutions quickly proved ineffectual. But the complete failure of attempts by Yuan and later another warlord to found new dynasties demonstrated that the rules of political life had changed.

3.2 Cultural and Intellectual Change

Attempts by Chinese revolutionaries to transform traditional belief structures and patterns of behavior began after the Opium War (1839–1842). The military defeat suffered then by China led some within the dominant scholar-official class to begin questioning longstanding assumptions about the inferiority of foreign cultures. The first generations of intellectuals to rethink these issues, while interested in specific things the West had to offer (*yong* or techniques),

were seldom revolutionary in their approach to questions of fundamentals (*ti* or essences). Their belief in the superiority of traditional moral codes remained unshaken; they merely sought efficacious ways to combine Confucian *ti* with Western gunboats and other sorts of foreign *yong*.

The *ti*/*yong* distinction remained at the heart of the debate for decades, until the iconoclastic New Culture Movement (1915–1923) led by intellectuals who had gone abroad, usually to Japan, or been exposed to foreign ideas at one of China's recently formed Western-style schools. They disagreed about many things but shared three convictions. China's weakness was due to the enduring power of entrenched 'Confucian' and 'feudal' beliefs (such as the veneration of age) and practices (such as arranged marriages). Chinese should welcome the best the West had to offer: ideas and methods of inquiry included. And intellectuals needed to educate and mobilize others.

Much New Culture Movement energy was directed toward the publication of new periodicals, which were filled with articles on Social Darwinism, Pragmatism, Anarchism, Marxism, and other imported ideologies. These periodicals also attempted to serve another purpose: their articles, written in a 'plain speech' as opposed to a 'classical' style, were supposed to inform and inspire the masses. This literary move to break down the barriers between intellectuals and non-intellectuals was reinforced by public speaking campaigns on issues ranging from medicine to international affairs.

An important new turn came with the May 4th Movement, when student activists with ties to the New Culture Movement actively appealed to members of other classes to join them in protesting terms of the Treaty of Versailles that transferred control of former German territories in China to Japan. After sympathy strikes by workers and merchants, the students achieved their main domestic goals: the dismissal of three corrupt officials and the release of all those arrested for protesting. But their inability to stop the Treaty of Versailles from taking effect illustrated the need for a diplomatic revolution.

3.3 Diplomatic Change

No diplomatic revolution could take place until the rise of a strong central regime committed to overturning the unequal treaties forced upon the Qing by foreign powers. The Northern Expedition brought this about. After reunifying the nation, however, Chiang claimed that the spread of communism was more dangerous than the continuation of imperialism. Unlike some of his warlord predecessors, Chiang was an outspoken critic of the unequal treaties, but he insisted that China could only regain its place in the world once its house was in order.

World War II precipitated a sea change, as China's position as one of the Allies gave its leader new leverage in pushing for an end to unequal relations with the West. And, in 1943, the Allies renounced all claims to special privileges within Shanghai and other 'treaty-port' cities. This decision and Japan's withdrawal in 1945 made China freer of foreign influence than it had been for a century.

CCP historians argue, however, that the diplomatic revolution was still not complete—and they have a point. The GMD regime remained economically and diplomatically dependent on the USA. If credit for starting the diplomatic revolution rightly goes to Chiang, it remains true that China did not fully regain its status as an independent country until after 1949.

3.4 Socioeconomic Change

The socioeconomic revolution also began long before the Communists came to power. Most notably, the late Qing and Republican (1912–1949) eras saw dramatic shifts in the types of people in control of the government and witnessed the rise within the social hierarchy of merchants. But the development of new alignments near the top, though important, did not have much effect upon the lives of the vast majority of Chinese who continued to reside in villages and work small plots of land. For them, a socioeconomic revolution did not take place until the CCP's land reform campaigns. This socioeconomic revolution (like the others) did not hit all parts of the country at the same time but rather played itself out according to regionally distinctive timetables. By the early 1950s, however, it had affected most of China's villages.

3.5 Unfinished Revolutions

If China's twentieth century is envisioned as a time of multiple revolutions, we need to keep in mind that most were incomplete—and that struggles for change in other spheres were pursued sporadically but never fully carried through. One important unfinished *Geming*, linked to but not quite the same as any of those just described, was the revolution for women. Aspects of sexual politics changed throughout the 1900s, but the dream of complete equality promoted by revolutionaries in various decades never became a reality. Much the same can be said for the democratic revolution: calls for democratization were heard often but oligarchy and one-party authoritarianism tended to prevail.

Ironically, though China's twentieth century was filled with transformations, its final two decades were strikingly similar to its first three. In Taiwan, moves to institute a system of free elections recall experiments made around 1911. On the mainland, there were May

4th era parallels for the calls in 1989 for an end to corruption. And in the automatic weapon fire that accompanied the massacres, there were echoes of killings of the 1920s that were justified as necessary efforts to maintain order and/or protect the *Geming*.

4. Lessons From the Chinese Case

China's twentieth-century experience suggests the need for social scientists to pay more attention in the future to a few factors. One, already mentioned, is the revolutionary potential of villagers. Another is scale. Revolutions have often been treated as purely 'national' events. Recent work on China shows that local dynamics are often crucial—even in upheavals driven by nationalism. So, too, are transnational flows of people (key participants in 1911 were overseas Chinese) and ideas. This suggests that, in future, models used for comparing revolutions will need to make more room not just for revolutions of widely varying lengths but also revolutions that were influenced by local, national, and transnational impulses and actors.

However, China's recent history reinforces some important insights in the comparative and theoretical literature on revolutions. It shows that Alexis de Tocqueville was right to insist that nothing does more to inspire revolutionary activism than an inefficient authoritarian regime's introduction of a reform program that offers too little and comes too late. The 1911 Revolution (which was preceded by unsuccessful Qing reforms) and even, in a way, the mainland's abortive revolution of 1989 and the recent fall of the GMD in Taiwan illustrate the power of Tocqueville's insight. These cases also illustrates the value of stressing, as Marx and Skocpol among others have, the revolutionary potential of groups at the fringe of the governing elite who have rising expectations and are frustrated by being kept away from real power. In addition, Chinese events of the third quarter of the twentieth century in particular underscore the validity of the claim made by both Tocqueville and Skocpol, again among others, that revolutions, while attempting to do other things, often end up strengthening the power of central governments. Finally, the Chinese case shows just how much truth there is in the tragic maxim that revolutions all too frequently and all too regrettably end up devouring their own children.

See also: China: Sociocultural Aspects; Communism; Communist Parties; East Asian Studies: Economics; East Asian Studies: Politics; East Asian Studies: Society; Maoism; Revolutions, Sociology of; Revolutions, Theories of

Bibliography

Bergère M C 1998 [English translation by Lloyd J] *Sun Yat-sen*. Stanford University Press, Stanford, CA

Bianco L 1971 [English translation by Bell M] *Origins of the Chinese Revolution, 1915–1949*. Stanford University Press, Stanford, CA

Brinton C 1965 *Anatomy of Revolution*. Vintage, New York

Dirlik A 1989 *The Origins of Chinese Communism*. Oxford University Press, New York

Duara P 1988 *Culture, Power and the State: Rural North China, 1900–1942*. Stanford University Press, Stanford, CA

Fairbank J K, Goldman M 1998 *China: A New History*, enlarged edn. Belknap Press of Harvard University Press, Cambridge, MA

Fitzgerald J 1996 *Awakening China: Politics, Culture, and Class in the Nationalist Revolution*. Stanford University Press, Stanford, CA

Friedman E, Pickowicz P G, Selden M 1991 *Chinese Village, Socialist State*. Yale University Press, New Haven, CT

Goldstone J A (ed.) 1994 *Revolutions: Theoretical, Comparative, and Historical Studies*, 2nd edn. Harcourt, Fort Worth, CA

Hartford K, Goldstein S (eds.) 1987 *Single Sparks: China's Rural Revolutions*. Sharpe, Armonk, New York

Ono K 1989 [English translation by Fogel J (ed.)] *Chinese Women in a Century of Revolution, 1850–1950*. Stanford University Press, Stanford, CA

Perry E J, Selden M (eds.) 2000 *Chinese Society: Change, Conflict and Resistance*. Routledge, London

Skocpol T 1979 *States and Social Revolutions*. Cambridge University Press, Cambridge, UK

Spence J D 1982 *The Gate of Heavenly Peace*. Penguin Books, New York

Tang T 2000 Interpreting the Revolution in China: Macrohistory and micromechanisms. *Modern China* **26**(2): 205–38

Wakeman F Jr. 1975 *The Fall of Imperial China*. Free Press, New York

Wang Z 1999 *Women in the Chinese Enlightenment: Oral and Textual Histories*. University of California Press, Berkeley, CA

Wong R B 1997 *China Transformed: Historical Change and the Limits of the European Experience*. Cornell University Press, Ithaca, NY

Wright M C (eds.) 1968 *China in Revolution: The First Phase, 1900–1913*. Yale University Press, New Haven, CT

Yue D, Wakeman C 1985 *To The Storm: The Odyssey of a Chinese Revolutionary Woman*. University of California Press, Berkeley, CA

J. N. Wasserstrom

Christian Liturgy

The Greek word *leitourgos* means, literally, 'work of the people' and was associated in ancient Greece both with the payment of civic dues and the performance of ritual duties. In the Christian context, by contrast, it had at first a more specific connotation, concerning the performance of the Eucharistic action—although this apparent narrowing of reference in fact indicated that this action was itself the supreme collective obligation and source of collective unity. Only much later did the term come to denote the entirety of Christian ritual practice. At first, in the seventeenth century, it was used as a neutral term, covering both the Catholic 'Mass' and Protestant 'Communion';

later, Catholic writers such as Dom Odo Casel worried about the degeneration in meaning if *cultus* tended to substitute 'liturgy' as a word describing the whole outward and inward *opus* of Christian piety.

To comprehend the history of the latter, therefore, one must attend also to the evolution in meaning of other terms: 'mystery,' 'cult,' 'rite,' and 'sacrament.' In the antique world there existed, broadly speaking, a contrast between public 'liturgies,' connected to the regular order of the city and its upholding, on the one hand, and private 'mysteries,' 'cults,' or 'rites,' sometimes connected with a certain dissension from civic order, on the other. Public liturgy included animal sacrifice, and was concerned primarily with a symbolic apportioning of different parts of the beast to gods and different classes of men. It was at once a mimesis and a reinstitution of civic order. Private cults, by contrast, also involved deviant sacrifice. Sometimes this was a deviation upwards, as with the Pythagoreans, who modified or refused animal sacrifice, and associated less bloody offerings not with a feeding of the gods but with the transit of their own souls to a higher realm. Alternatively, there were deviations downwards, as with the Eleusinian mysteries, which were essentially older, alien, more agrarian rites involving chthonic gods, reinterpreted in an urban context. Here initiates identified themselves with the perpetual dying and rising to life again of a god, who had been originally a god of fertility. By participating in the god's own self-salvation, the *mystes* hoped to regenerate their lives.

Christianity fused together public liturgy and private rite, with momentous consequences. From early times, it interpreted worship, *latreia*, as meaning, after the incarnation, the entire offering of the whole person to God in charity, which included charity toward one's neighbor. In this way, the whole span of human life was reconceived as 'liturgical,' since the new 'city' was an eternal city which also embraced true human life in time. On the other hand, the language of 'mystery' and 'initiation' was also embraced. In the case of St. Paul's use of the term *mysterion*, it is true, the background is very unlikely to be that of the pagan mystery cult, as was once thought. Instead, the background is Jewish apocalyptic: thus Paul speaks of a primordial 'mystery' now disclosed to us in Christ, a mystery anticipated in the Jewish Passover, which involves a passage through destruction to renewal. Nevertheless, this meant that, at its heart, Christianity involved the mystery of the death and rising again of God, a mystery that was made present again in the rituals of baptism and the Eucharist. Later, Patristic authors expanded this notion in terms that owed something to pagan notions of initiation into secret knowledge: the Syrian fathers spoke of *Raza*, an originally Persian term (*Raz*) denoting secrets of state within the imperial court; one can note here that in this more oriental context, the 'secret' and the 'public' were already identified before Christianity, although neither was yet democratically available to all. This 'orientalism' was also present more in Rome than in Greece, for, in the case of Rome, the plebs were admitted or 'initiated' only gradually into the rites of *connubium* (sacred marriage). The Greek fathers spoke of *mysterion* in ways that fused Jewish apocalyptic expectation and exaltation with a Greek sense of participating in a hidden drama that yields understanding. Nevertheless, the association with mystery religions was viewed typologically, and distance as well as proximity was emphasized. One can note in particular the extent to which the pagan mysteries' involvement of the reslaying of a god was exaggerated in the interests equally of resemblance with and contrast to the (voluntary) death of Christ. Unlike the pagan mysteries, the mystery of Christ involved a once and for all death and an unshakeable resurrection which saved not a god but mortals.

It remains the case, however, that now, in a more 'oriental' (and also more Roman) mode, the most public emerged from the most secret, a rite into which one first had to be initiated as a catechumen. Unlike the ancient Orient, furthermore, (although pagan Rome had already evolved in this direction), *all* could potentially be initiated into the *Raza*, or inner court secret. This was to break with the ancient Greek and Roman association of the aristocratic with the eternal and transcendent reserved for a few, on the one hand, and the democratic, associated with the immanent, open, 'positive,' unmysterious, and available to the many, on the other. Plato, in the *Laws*, had already begun to de-eroticize the transcendent, or, alternatively, to eroticize the democratic, since one must express this both ways around. Now, however, not just in a theoretical text, but in actual practice, the secret was publicized, and, equally, the most public—the Eucharist which engenders the *ecclesia*, the corporate identity—was rendered secret (and permanently mysterious, even for the catechized). Once, either some were to ascend, or all were to remain on an immanent plane; now all were to ascend, continuously.

In accordance with this new conjoining of liturgy and cult, the sacrifice of the Eucharist was public, yet involved no unequal apportionings. All now ingested all that was offered, and yet in eating the totality, the elements and those whom they fed were offered in their entirety back to God. In addition, all 'sacrificial economy' was broken with; nothing was any longer expected from God in return for one's giving. The new sacrifice was one of pure gratitude to the God who, in any case, gives. As with the Pythagoreans, sacrifice involves mainly *human* elevation. This elevation, however, also now includes humans' own free giving in charity to others. In this fashion, elevation no longer abandons the city. Indeed, for Augustine in the *City of God*, it is the whole city which is offered, which is elevated—since the city exists through elevation.

Because of this new, nonsacrificial economy, Augustine refused most of the pagan terms for ritual

or worship, espousing only the Greek term *latreia* (CD VI 3). Nevertheless, he spoke of a true *Cultus* as involving a participation in the one true offering made by Christ (CD VII 30). This was an 'inner' *cultus*, not in the sense that outer signs are inefficacious, but in the sense that they must be truly intended. Indeed, participation in the mystery of Christ through the Eucharist and the annual liturgical calendar is now so intense that christians inner life must be understood also in cultic terms: their heart is not mainly 'ltheins,' but rather an altar upon which they sacrifice to God (CD X 3; Ep. 140 *Ad Honoratum*, 18, 45).

Thomas Aquinas reiterated and expounded Augustine's understanding of *latreia* and *cultus*. He stresses that cult does not have God as its *object* but as its *end*, since the aim is not to please God, but to be united with Him, and this is not brought about *through* the work of worship; rather, God brings it about Himself by meeting humans in liturgical acts of worship (S.T. III, Q. 81 a 5; Q. 24 a 5).

Up until the twelfth and thirteenth centuries, this understanding of cult was in the most part preserved. It worked against any notion that the essence of liturgy is a matter of 'correct procedure,' or that the liturgical sphere is a special domain standing over the practical and theoretical aspects of life. Even in Gratian's *Decretals*, the aim of the canonist is to allow proper scope to the force of local custom, so long as this is not inconsistent with the custom and understanding of the church in general, following Augustine of Canterbury's precept that 'place does not approve a custom, custom approves a place' (*Decretals*, 12. C 10).

Nonetheless, the later Middle Ages witnessed an increased notion of juridical thinking in the liturgical sphere, resulting in some separation of private piety from public ritual: in this way, the Christian logic of 'the public secret' started to come undone. The Reformation then protested against a mechanistic formalism which tended to suggest that the following of certain procedures secured salvation. In response, the Council of Trent sought diligently to reconnect the outer and the inner, and to reassert the notion of a *sacramentum* as an outward sign of an inner reality whose exteriority could not be dispensed with, since the fullness of this reality was only to be eschatologically disclosed, and then would transcend the inward/outward contrast. However, the implementation of the Tridentine decrees involved more and more a dry insistence upon outward observance, as sacramental practice came to be viewed as a series of instituted motions with formally consistent entailments ensuing automatically. Such an outlook at once harmonized with Enlightenment rationalism, and was itself part of what the Enlightenment rejected.

By contrast, the spirit of the most ancient tradition was renewed and rethought by Cardinal Bérulle, who insisted that the whole of Christian life, outer and inward, was a participation in, and, in a sense, a re-enactment of the incarnation: only the divine man who utterly gave Himself showed, by giving beyond humanity, true humanity. Ultimately, the French School which he founded helped to sustain some unbroken tradition of Christian liturgy as encompassing all of Christian life, since it was a sharing in the mystery of Christ. The centrality of 'mystery,' and so of the notion of a public secret (or a secret publicness) was emphatically renewed in the nineteenth century by Dom Odo Casel who (though he overstressed the influence of Greek mystery religion) was a decisive influence on Catholic liturgical renewal in the twentieth century.

In many ways, Vatican II restored the theological centrality of liturgical mystery, although it is questionable whether its practical liturgical recommendations were entirely in keeping with this understanding. It misinterpreted a particular ancient local practice at Rome as indicating that originally the priest stood behind the altar to celebrate the liturgy, whereas the design of even the most ancient basilicas suggests this could not have been the case. Most scholars now reject this conclusion, but the adopted new practice of the priest facing the congregation tends to reduce both the sense of approaching eschatological mystery, and of an equal approach by the entire public, both priest and people. Likewise, Vatican II mistook documents of broad direction for liturgical enactments, such as that of Hippolytus and Justin Martyr, as indications of original 'simple' liturgies, organized more formally and without supposedly 'messy' repetitions. In this way, another dimension of public mystery was lost in practice: the endlessly 'stammering' recommencement of an approach to an altar where one can truly worship—an approach that cannot be completed in time. Thereby both eschatological and apophatic aspects were diminished in practice, even though the reformers had stressed these in theory. Equally, the new Latin liturgy and its vernacular translations tended to lose metaphoric richness, typological resonance, and a sense of language as an epiphanic vehicle.

For these reasons, efforts in the tradition of Bérulle and Casel to restore the centrality of Christological mystery and public secrecy to Christian practice remain only partially accomplished.

See also: Pilgrimage; Ritual; Sacrifice; Symbolism in Anthropology; Symbolism (Religious) and Icon

Bibliography

Aquinas T 1964–81 *Summa Theologia*. Eyre and Spottiswood, London

Augustine of Hippo 1984 *Civitas Dei* Bettenson H. Penguin, Harmondworth, UK

Beard M, North J, Price S 1998 *Religions of Rome*. Cambridge University Press, Cambridge, UK, 2 Vols.

Bouyer L 1968 *Life and Liturgy*. Sheed and Ward, London

Bouyer L 1986 *Mysterion*: *Du Mystère à la Mystique*. Aubier, Paris
Bremmer J 1994 *Greek Religion*, Greece and Rome new surveys. Oxford University Press, Oxford, UK
Bremmond H 1932 *Histoire Litteraire du Sentiment Religieuse en France t. 9: La Vie Chrétienne sous L'Ancien Régime*. Aubier, Paris
Bugnini A 1990 *The Reform of the Liturgy* 1948–1975 (trans. M J O'Connwell). liturgical Press, Collegeville, MN
Burkert W 1985 *Greek Religion. Archaic and Classical* [trans. Raffan J]. Blackwell, Oxford, UK
Burkert W 1987 *Ancient Mystery Cults*. Harvard University Press, Cambridge, MA
Casel O 1942 *Das Christliche Kultmysterium*. Munich, Ratisbonne
de Certeau M 1975 *L'Écriture de l'histoire*. P.U.F, Paris
Congar Y 1947 Le Christ, image de Dieu invisible. *La Maison-Dieu: Revue de Pastorale Liturgique* **59**: 132–61
Congar Y 1967 L'Ecclesia ou communicante Chrétienne, sujet integral de l'action liturgique. In: Jossua J-P, Congar Y (eds.) *La Liturgie après Vatican II*. Aubier, Paris, pp. 246–82
Dalmais I H 1990 *Raza* et sacrament. In: de Clerck P, Palazzo E (eds.) *Rituels: Mélanges offerts au Père Gy*. Editions du Cerf, Paris, pp. 173–82
Duval A 1985 *De Sacraments au Concile de Trente*. Paris
Elich T 1991 Using liturgical texts in the Middle Ages. In: Austin G (ed.) *Fountain of Life*. Pastoral Press, Washington, DC, pp. 69–83
Gratian 1995 *The Treatise on Laws* [trans. Thompson A]. Catholic University Press of America, Washington, DC
Harrison T 2001 *Greek Religion: Belief and Experience*. Duckworth, London
Lyonnet S 1967 La Nature du culte dans le N.T. *La Liturgie après Vatican II* Aubier, Paris
Milbank J 1998 The politics of time: Community, gift and liturgy. *Telos* **113** (Fall): 41–69
Mohrmann C 1965 *Sacramentum* dans les plus anciens textes Chrétiens. In: (eds.) *Études sur le Latin des Chrétiens*. , Rome
Pickstock C 1998 *After Writing: On the Liturgical Consummation of Philosophy*. Blackwell, Oxford, UK
Price S 1999 *Religions of the Ancient Greeks*. Cambridge University Press, Cambridge, UK
Riedwig C 1987 *Mysterienterminologie bei Platon*, *Philon et Klemens von Alexandrien*. Berlin
de Roten P 1992 Le vocabulaire mystagogique de saint Jean Chrysostome. In: Triacca A M, Pistoia A (eds.) *Mystagogie: Pensée Liturgique d'aujourdhui et liturgie ancienne*. Edizioni Liturgiche, Rome
Rotureau G 1944 *Le Cardinal de Bérulle*, *Opuscules de Pieté*. Aubier, Paris
Zaidman L B, Pantel P S 1992 *Religion in the Ancient Greek City* [trans. Cartledge P]. Cambridge University Press, Cambridge, UK

C. Pickstock

Christian Parties: European

Christian (or Christian democratic) parties have been among the most successful political movements in Europe. Together with their Social democratic counterparts, they have dominated European politics. They have been leading members of governmental coalitions in such countries as Italy, Germany, Austria, Belgium, and The Netherlands; they currently form (together with their conservative allies) the largest group in the European Parliament.

Christian democratic parties are not just parties that use this label (some do not). As a matter of fact, Christian democratic parties are hardly Christian. They are secular parties operating in highly secularized societies. Today they seem almost indistinguishable from conservative or liberal parties. However, they have a distinct history that both accounts for their particular nature and helps explain their major contribution to politics. This contribution can be articulated around two paradoxical outcomes: although they were formed initially to challenge the emerging European liberal democratic order, these parties eventually became pillars of secularism, liberalism, and democracy in Europe. Christian democratic politicians also pioneered the process of European integration, and although they fought hard against socialism, they ended up building vast welfare states. Their evolution is a stunning illustration of how democratization can be the contingent outcome of political strategies rather than result directly from the dissemination of democratic ideas and principles.

1. Origins

Contemporary Christian democratic parties evolved from the confessional parties that were created in the second part of the nineteenth century and were an expression of political Catholicism. These Catholic parties grew out of the largely antiliberal and ultramontane mass Catholic movement that challenged the ascendancy of liberalism in Europe from a 'fundamentalist' and theocratic perspective (as codified in the 1864 papal encyclical *Syllabus of Errors*). Indeed, Christian democracy was a concept coined in opposition to liberal democracy. Though spearheaded by the Catholic Church which feared for the loss of its privileges, especially in the field of education, Catholic movements won their independence from the church through their transformation into Catholic parties. The Catholic Church resisted this process, which robbed it from its monopolistic control over its flock, as much as it could; but it failed to thwart this development because democracy provided Catholic activists with a powerful source of power and legitimation. Although initially strongly opposed to democracy on ideological grounds, these activists quickly realized that their interests lay in its consolidation and further expansion (Kalyvas 1996).

The process through which confessional parties were formed had two important, though contradictory, consequences: first, it turned religion into the foundational element of confessional parties, the core of their identity; yet religion proved more of a hindrance than an advantage; second, the religious

appeal produced a highly heterogeneous sort of party, composed of interest groups which had been united only by their adherence to the message of religious defense; yet this social heterogeneity increased the salience of class within these parties.

2. *Religion*

Confessional parties, albeit friendly to religion, wanted to disassociate themselves from too close a relationship with the church. Religion restricted their appeal as well as their autonomy. Likewise, the church could only protect its universalistic identity by moving away from these parties. However, the confessional character of these parties could not be shed because religion had become the cement that kept their heterogeneous social basis together. This quandary was solved in an ingenious yet momentous way. Confessional parties redefined religion into a nebulous humanitarian and moral concept that allowed them to be simultaneously Christian and secular. Vague formulations such as 'religious inspiration,' or 'values of Christian civilization' remain the only references to religion one finds in the official discourse of these parties. This led to a situation whereby it is perfectly possible to be simultaneously a Christian democrat and an atheist. In fact, this is not even perceived as a contradiction.

Hence, Christian democratic parties contributed in a fundamental way to the 'desacralization' and secularization of their countries' politics. Paradoxically then, the politicization of religion contributed to the secularization of politics. In a perverse fashion, Christianity was drained of its religious content even while being legitimated as a political identity—and this feat was accomplished by its proponents rather than its opponents. The secularization of confessional parties was, thus, endogenous to these parties and took place well before World War II, rather than being a delayed adaptation to external societal developments as often thought. Besides consolidating democracy, this development further enhanced the position of Christian democratic parties by laying the path, after World War II, for interdominationalism, thus turning Christian democracy into a dominant party in confessionally mixed societies such as Germany.

3. *Class*

The social basis of confessional parties was made of numerous and often-conflicting interests: social heterogeneity was these parties' hallmark from the outset. In this sense, Christian democratic parties were catch-all parties *avant la lettre* (van Kersbergen 1994). This heterogeneity was the direct result of their ideological profile that emphasized religion at the expense of class. However, external 'nonclassism' produced internal 'classism.' Powerful Catholic Workers' and Peasants' associations had to be incorporated into the new parties which eventually adopted a peculiar confederate structure based on organizations defined in terms of class (*standen* or *lager*). The ensuing conflicts gave rise to intensely accommodationist and compromising practices that were necessary for ensuring the parties' unity and cohesion.

Hence, mediation between these assertive and divergent interest groups became imperative. As a result, Christian democratic parties became particularly skilled in the exercise of the politics of mediation (van Kersbergen 1994), something their opponents have derided as opportunism and a belief 'that the ends justify the means.' The principle of subsidiarity (higher authorities, such as the state, should intervene only where individuals or smaller communities are not competent), a central concept in the process of European integration, can also be traced back to these developments.

Herewith (rather than in papal encyclicals, such as the 1893 encyclical *Rerum Novarum*) lies the source of Christian democracy's strong social component that distinguishes it from mainstream conservative or liberal parties. Indeed, a number of studies have found that Christian democratic strength is positively associated with high levels of welfare expenditure and high levels of unionization. Van Kersbergen (1995) has identified a distinctively Christian democratic core of social policies, which he calls social capitalism. This policy core differs significantly and systematically from both the liberal and social democratic conceptions of social citizenship. The Christian democratic welfare state is as large, in terms of expenditures and size, as its Social democratic counterpart; but it is quite different. It privileges families rather than individuals, cash benefits rather than social services, and seeks to preserve rather than subvert labor market outcomes. Following the end of World War II, social capitalism (along with political anticommunism), rather than religion, provided the foundation on which Christian democratic parties stood—and from which they ruled. Social capitalism was, thus, not only the outcome, but also the means of Christian democratic mobilization.

In the course of the 1990s, Christian democratic parties entered into a protracted and deep crisis. They have experienced steep electoral decline (Austria, Belgium, The Netherlands) or even total collapse (Italy); they have been implicated in major financial scandals (Belgium, Italy, Germany, Austria); and they have failed to expand, as expected, in Eastern Europe. Although the end of the Cold War can be partially credited for some of these developments, this crisis goes far deeper. The imperatives of global economic competition have undermined the social capitalist arrangements that had guaranteed both the parties' cohesion and their electoral appeal. In other words, this is a crisis of Christian democracy rather than a

crisis of Christian democratic parties. Christian democratic parties will have to reinvent themselves if they want to remain a relevant force in European politics. Yet if they look back, they will find that their often forgotten history holds a vast repertoire of frequently unintended but nonetheless ingenious practices of adaptation and reinvention.

See also: Church and State: Political Science Aspects; Consumption, History of; Party Systems; Religion and Economic Life; Religion: Mobilization and Power; Religion, Sociology of; Religious Stratification; Socialization: Political; Western European Studies: Religion; Western European Studies: Society

Bibliography

Buchanan T, Conway M 1996 *Political Catholicism in Europe, 1918–1965*. Clarendon Press, Oxford, UK

Irving R E M 1979 *The Christian Democratic Parties of Western Europe*. Allen and Unwin, London

Kalyvas S N 1996 *The Rise of Christian Democracy in Europe*. Cornell University Press, Ithaca and London

Kalyvas S N 1998 From pulpit to party: Party formation and the Christian Democratic phenomenon. *Comparative Politics* **30**: 293–312

van Kersbergen K 1994 The distinctiveness of Christian Democracy. In: Hanley D (ed.) *Christian Democracy in Europe: A Comparative Perspective*. Pinter, London and New York

van Kersbergen K 1995 *Social Capitalism: A Study of Christian Democracy and the Welfare State*. Routledge, London and New York

S. N. Kalyvas

Christianity: Evangelical, Revivalist, and Pentecostal

While 'globalization' as a concept arose as a fiscal planning tool in the 1950s, its extension as an explanatory concept for the sort of territory previously covered by such frameworks as 'world systems theory' dates from the 1970s. At the time of writing, Roland Robertson and Peter Beyer are the key writers in the area as it relates to religion. Indeed, Robertson's pioneering framework arises out of the conundrum which religion presents for standard modernization theory. The rise of Islamic and other forms of religious fundamentalism denies the linear rationalist assumptions of modernization theory and such offspring as secularization theory, which suggest that regional difference and religious conviction should be passing away in the face of rational technical systems and the homogenization of culture. Robertson has since then further articulated the theory through the concept of 'globalization,' which allows for simultaneous global homogenization of culture and local reassertions of difference. Robertson's theory that cultural forces move faster than political and economic forces has received considerable support from historians and analysts of Christian missions, which have noted that the indigenization of Christianity has historically been faster and deeper in Asia and Africa than the movement of formal missionary and colonial structures. In South America, by contrast, there is a body of opinion which states that the combination of state apparatus and church extension actually retarded the extension of Christianity. It is clear, then, that globalization theory is vital for the understanding of the expansion of Christianity generally, just as the study of the expansion of Christianity is vital for understanding the processes of globalization.

While a slow starter to get into missions, evangelical Christianity began to articulate an increasingly sophisticated global vision from the late 1600s, with the rise of the Pietist movement. As Lewis has noted, Pietism restored a sense of personal call to Protestant orthodoxy which, through agencies such as the University of Halle and Moravian missions, expressed itself in a growing internationalism. The correspondence of Zinzendorf, the migration of Moravians, and missionary outreach to Russia, Greenland, and India, created networks and examples of successful missions which informed the expanding networks of new settler societies and expanding capitalism. It was on a trip to the New World that John Wesley, for instance, first ran into Moravians, who, through the influence of Peter Böhler, provided the synthesizing principle to Wesleyan eclectic theology and practice which in turn made it a vibrant missionary force. It was likewise among Scots-Irish and Dutch settlers in the New World that religious revival produced the combination of Protestant theology and religious experience which has become known as 'Evangelical Christianity.' Definitions differ, usually with personal affiliation, but a standard definition of Evangelicalism offered by David Bebbington locates the movement in a quadrilateral of theology and praxis involving activism, cross-centeredness, Bible-centeredness, and conversionism. The fact that nothing like the French Revolution occurred in England helped legitimize the movement as a source of social and cultural stability. The inner energies of the movement, combining a universalizing crucicentrism and biblicism, with conversion-oriented activism, meant that the tens of thousands of converts to the Evangelical Awakening on both sides of the Atlantic provided a vast pool of willing support to the evangelical program to change globalizing British society. The leaders of the movement—in particular William Wilberforce, Lord Shaftesbury, the Clapham sect, and Selina, Countess of Huntingdon—were able to draw upon these networks of support in a dynamic program of social activism over the century 1750–1850. This activism

arguably provided important impulse force to the processes of globalization, just as globalizing Western economies provided important resources and opportunities for evangelical expansion and change.

Protestants from Luther's time were faced with an expanding globe in tension with an increasingly outmoded ecclesiology. While, during Luther's own lifetime, Diaz rounded the Cape of Good Hope (1488), Columbus was driven west in pursuit of a millennial vision to land on the shores of America (1492), and other parts of the world were opened to Western influence by da Gama, Cabot, Cortes, Magellan, and Pizarro, Protestants were combining with increasingly nationalistic ruling classes (the nobility in Germany, bourgeoisie in Switzerland and The Netherlands) to reinstitute the 'Christendom' ideal within the national churches of rising nation states. While Catholic empires ruled the seas, Protestant missions were restricted to Europe. While the theology became increasingly arid and formalized, the impact of these churches was to reinforce the nexus between nationalism and capitalism. Capitalism expanded during the period through royal chartered companies, combining national expansion and capitalist enterprise. Halle Pietism specifically used this link to expand into Russia, just as Puritans used the chartered company to establish themselves in Massachusetts Bay. The CMS first fought with and later used the East India Company in Asia, while the Claphamites moved into West Africa on the tails of colonization companies. While Weber's 'spirit of Protestantism' has been revised and attacked in some quarters, it is nonetheless a useful heuristic when viewing the interaction of shipping, chartered, and colonization companies on the one hand, and missionary agencies on the other.

Protestantism provided an 'inner fascination with the world,' and evangelicalism adopted the same fascination and extended it. There has been considerable work done, for instance, on the link between enlightenment and Evangelicalism, indicating that Evangelicalism, through its commercial and missionary emphases, was a major carrier of rationalizing Western concepts of time and space into the rest of the world during the late nineteenth and early twentieth centuries. Not only did Wesley proceed on the assumption of an 'inner enlightenment,' using both the vocabulary and technologies of enlightenment to further the evangelical cause (he was, for instance, an assiduous publisher), but Puritan and evangelical input into the sciences from the sixteenth to the nineteenth centuries was immense. Clergy were not only involved in agencies such as the Royal Society, but were key suppliers of information to enlightenment projects such as the social sciences. (Malcolm Prentis, for example, has demonstrated the link between clerical collection and the construction of an international culture of research in anthropology.)

The logical place to look for the globalizing influence of evangelicalism is through the agency of missions. The marginality of missions at the beginning of the nineteenth century demonstrates the close association of evangelical expansion with Western expansion. By 1800, North America was subject to European settlement, but only on the eastern half and the western rim of the continent, and though there was no real challenge to Hispanic hegemony in South America, Christianization had not really pierced deeply below the surface of the indigenous and emerging plantation cultures. The impact of missions in the Indic, wider Asian, and Muslim worlds was negligible. Obviously, then, developments in the nineteenth century were critical for the development of global Christianity in our own period. Europe at the time was weak and divided religiously. The industrial revolution encouraged both concentration of the sort of wealth which could capture foreign markets and provided the scientific curiosity and the speed of travel to take advantage of new opportunities. While most European expansion in the eighteenth century had been mercantilist in nature—with small colonies of isolated Europeans trading with Asian nations in sealed cantons such as Shanghai, Macau, and Hong Kong—the nineteenth century saw a shift toward extensive settlement and active domination of local politics. At least in part, this was because it was now technically possible, with advances in communication and military hardware, to dominate other less advanced cultures. So Britain absorbed India, Burma, and Ceylon into a new Empire, France annexed Indo-China, and The Netherlands finalized its annexation of Indonesia. Other European powers founded colonies, such as the German colonies in PNG, and annexed parts of Africa, such as Togoland, Southwestern Africa, and the like. At least for the British, Stanley argues that there was a consistent imperial impulse to British society, but no plan of conquest, most acquisitions having been gained in order to protect trading rights. Formal control was never exercised when the much cheaper form of informal control could be used, and it was the increasing need for formal control, driven by the intensified competition of European powers for global resources, which drove Britain to acquire and then defend India as the keystone in its defense of the East.

In terms of the other global religious cultures, evangelicalism also faced weakened opposition. One advantage of the time was that Islam was in retreat—in 1821 Greece was retaken, and from the 1830s European powers began dividing up Islamic parts of Northern Africa such as Morocco, Algeria, and Tunisia. Roman Catholicism at the beginning of the nineteenth century was under siege from the forces of modernization. With the eclipse of Portugal and Spain as imperial powers, increasingly the push for missions in the Catholic world came from France, Belgium, and (significantly for Australia) Ireland. While the Orthodox churches did not pursue much in the way of missions, despite the spread of the Russian Empire

during this period, the Protestants, energized by the Evangelical revival, certainly did, both within and outside the English speaking world. Stanley suggests that, while there were agencies such as the Society for the Propagation of the Gospel with a missionary element to them, before the Evangelical awakening there was no consistent acceptance of the burden of foreign missions by either Anglican or Dissenting Protestants. Europe was the same. In 1810, for instance, a revival known as the Reveil spread out of Geneva and affected the French Protestant world deeply, leading to considerable missionary endeavor overseas. Norwegian revivals, most notably commenced by Hans Nielsen Hauge, a lay reader preaching in a way which was not legalized until 20 years after his death, energized Scandinavian missions to North America and later to South America. German missionaries, in the tradition of Halle and the Moravians, traveled all over the world. The interaction of these revival movements and missions can be seen in their common adherence to interdenominational agencies such as the Evangelical Alliance, which was founded in 1847, and held conferences in France, Belgium, Switzerland, Germany, Canada, Sweden, and the United States.

The model provided by the Alliance of a voluntary society was to become the classic form for Protestant missions. Many were formed at the prompting of chaplains and clergy travelling with army, trading company or political representatives. The first of them was the Baptist Missionary Society (1792), inspired by William Carey, who in turn had been inspired to mission by the accounts of the travels of such explorers as Captain James Cook in the Pacific. It was followed by the London Missionary Society (1795), the British and Foreign Bible Society (1804), the Wesleyan Methodist Missionary Society (1813), and many others. Because the Church of Scotland refused the requests of its Evangelical party to form their own missionary agency, most Scots siphoned off into other agencies, such as the LMS. In 1810, preceded by many small bodies dedicated to evangelizing First Nations' peoples, the first American foreign agency was founded in the American Board of Commissioners for Foreign Missions.

The extension of evangelical missions had a number of major interactions with the growth of global society. The first was that, with increased need for personnel, the missions started their own Bible training institutions. These in turn legitimized the long-standing 'old dissenting' academies, supported religious emancipation and social reform, and acted to break down the exclusivism of the established church-linked universities. European states, at slower or faster rates, responded to this broadening of the religious franchise (among other secularizing tendencies) by stepping back from the official sanctioning of one form of religious truth or another. By the end of the century, most European states had become generally 'Christian' countries without confessional attachments, and many, in their colonies (particularly in Canada and Australia) were actively encouraging liberal democratic states. Not only did missions change the education of the increasingly vital professional class at home, but in exporting the Western educational model, also exported Western professional values into the nation states which began to arise in Asia, Africa, and the Pacific from the late nineteenth century. In exporting values such as professional disinterestedness, the importance of technical expertise, and basic human rights, a common culture was established which facilitated not only the spread of global society, but also (through preservation and re-skilling of Fourth World peoples) the means for local opposition to global homogenization. The specter of evangelical African bishops nay-saying advanced liberal proposals at a Lambeth conference, or Aboriginal evangelicals controlling the balance of power in the national assemblies of the United churches of Canada and Australia, have demonstrated the interactivity between evangelical missions and mediation of global culture.

The establishment of comity arrangements between evangelical missions in foreign countries highlights the second major interaction. The movement of religious traditions outside their national boundaries produced a relativization of their content and form, creating movements toward ecumenism based on 'common gospel' assumptions, diminishing the importance of liturgical and cultural tradition, and homogenizing the culture of international Christianity. This was imported back into the West as de-denominalization, and into Third World countries as 'alliance Christianity,' enabling Christianity to be described as a unitary object in contact with other world religions. This sense of relativization of form sparked concerns about relativization of truth, concerns expressed most markedly in the fundamentalism/modernism debates of the 1920s onwards.

Fundamentalism self-consciously identifies itself as a countermodernist movement, particularly in terms of theological modernism. Evangelicalism maintains something of this thrust, at the end of the twentieth century increasingly identifying itself as a counter-postmodernist movement. On another level, however, these debates can be seen as localized resistance to homogenizing forces at the global level, resistance seen most markedly in the general opposition of national evangelicals to agencies such as the World Council of Churches. In some cases, this meant the localization of evangelicalism into extreme nationalist forms. In most cases, however, especially after the rise of the National Association of Evangelicals in the USA and the Billy Graham Organization, the centrality of missions and a self-identity partly defined by contention for the faith caused evangelicals after World War II to take the same globalizing path that the WCC had already taken. This is reflected in the formation of agencies

such as the Lausanne Committee, the World Evangelical Fellowship, and worldwide denominational fellowships such as the World Alliance of Reformed Churches. Particularly after World War II, these agencies provided the basis for expansion of English and, with increasing hegemony, American evangelicalism onto the world stage. By the 1990s, it was clear that denominations had become almost irrelevant to the self-definition of most evangelicals, and that localized megachurches were more important for agenda setting.

Likewise, evangelical missions had shifted from being denominationally and agency based to increasing integration with global movements in migration, occupation, and tourism. Short-term mission, and mission originating from countries in the 'South,' contribute nearly as much to the total Christian missionary workforce as do traditional, First World, agency based missions, and is likely soon to outstrip it.

Within these tendencies in evangelicalism in the nineteenth century arose the issue of 'revivalism.' Revivalism is a technical accommodation between the experience of revival, as formalized during and after the First Great Awakening by Calvinists such as Jonathan Edwards, and the utilitarianism implicit in early nineteenth century postenlightenment thought. The emphasis of Edwards and Wesley in defending the use of 'means' for the furtherance of religious revival were codified and extended during frontier revivals in New York State and the expanding American frontier. Not least among the codifiers was Charles Grandison Finney, whose *Lectures on Revival* became the great 'how to' book of the nineteenth century. Revivalism as it expanded from the USA to England, and then around the world (by way of church and migrant diasporas) developed an array of techniques to maintain the association between Christian mass evangelism and spiritual response. It became the tool of choice for many evangelicals as Western societies seemed to be slipping away from their commitments to national religious frameworks, a set of ideas crystallized in the premillennial theology of archetypal evangelistic movements such as the Brethren, Baptists, and Methodists. As a technical accommodation to utilitarianism, revivalism was a natural client and promoter of globalization, which was very largely fueled by the technical innovation of high modernist capitalism. So revivalism quickly spread by the development of Methodist itineracy into the big campaign evangelism of D. L. Moody, Alexander Somerville, Gypsy Smith, and the like. Such itineracy became global, as shipping, then automobiles, and then airplanes, brought other parts of the globe within reach. As communications technologies also improved—through telegraphed newspaper reports, international postage, radioed accounts, and finally television coverage (culminating in the global campaign by satellite of Billy Graham in 1996)—evangelicalism spread with revivalism to all parts of the globe. It is important to note in this regard the high degree of evangelical participation in modern technological enterprises, from Back to the Bible Radio, to the massive media empire of the Universal Church of the Kingdom of God in Brazil. This has been seen as a paradox by some commentators—it is less obviously so when revivalism's accommodation to technological utilitarianism is taken into account. Though revivalism is not limited to evangelicalism (Catholic revivalism, for instance, or the adoption of revivalistic techniques by Hindu and Buddhist promoters in South and Southeast Asia), and not all evangelicalism is revivalistic, it is clear that the two are blood relations and natural partners in many cultures around the world. At the same time, the fact that revivalism is a distillation of 'revival tradition' into a technological form which can thereby cross cultural boundaries makes it eminently transportable. For this reason, Edith Blumhofer has called the major global form of modern revivalism, pentecostalism, 'world evangelicalism,' because of its facility with communications technologies. The ability of pentecostalism to 'cast' itself as a communal form of thought through the abstracting processes of print media and CITs makes it a prime globalizing force. On the one hand, this makes for global expansion of evangelicalism through its revivalist and Pentecostal forms. On the other hand, the same process homogenizes its distinctives and causes it to adapt new local distinctives as it moves across cultures. The lack of a single center (unlike either Islam or Catholicism) is thus a great facilitator of growth, but is also the prime cause of internal dissension and fragmentation as various local evangelical cultures seek to adopt or invent normative centers which enable local communities and protect them from the effects of fragmentation. Neo-fundamentalism, Calvinist renewal movements in the Southern Baptist Convention, the Toronto and Brownsville revivals—each of these may be seen as the creation of new centers, which feed from the energies of globalization, and attempt to create islands of order amidst the chaos of global society.

In particular, evangelicalism as it stands at the beginning of the twenty-first century is clearly largely a two-thirds world phenomenon. As Piggin, Reed, and others have shown, revival is the inculturation of Christianity into local cultures. It is not too much to say, therefore, that the theology of revival, particularly as evidenced in Pentecostal revivalism since the 1950s, is a mainstream legitimating structure for the indigenization of Protestant evangelicalism. This is particularly evident in revivalistic outbreaks such as the East Africa Revival in the 1930s, the blossoming of Pentecostalism in Brazil and later in Argentina, and the Aboriginal revival commencing at Galiwin'ku in Australia's Northern Territory from 1978. In each case, evangelicalism struck new roots and found new accommodations with existing religious cultures which have in turn been exported into the global religious network.

While the East Africa revival had comparatively little experience of charismatic manifestations, Brazil and the Aboriginal revival have been highly charismatic in their combination of local and general characteristics. As Freston notes, the massive expansion of evangelicalism in that showcase of evangelical inculturation, Brazil, is likewise largely Pentecostal in nature. As he noted in 1998, a 'survey of evangelical institutions in Greater Rio de Janeiro discovered that, of the 52 largest denominations, 37 were of Brazilian origin, virtually all Pentecostal. While only 61 percent of all evangelical churches were Pentecostal, 91 percent of those founded in the previous three years were' (Freston 1998, p. 74). Of the 450 million 'evangelicals' and 750 million 'Great Commission Christians' estimated by David Barrett to exist in 1990, some 350 million could be counted within the Pentecostal or charismatic camp. It is clear that at the end of the twentieth century, evangelicalism's expanding edge was clearly Pentecostal/charismatic in nature, leaving the national traditions of the First World heartlands significant challenges of reinvention in their struggle to survive the disappearance of their defining ethnic and national boundaries.

See also: Globalization and World Culture; Globalization, Anthropology of; Historiography and Historical Thought: Christian Tradition; Religion: Evolution and Development; Religion, Sociology of

Bibliography

Bebbington D W 1992 *Evangelicalism in Modern Britain: A History from the 1730s to the 1980s*. Baker Book House, Grand Rapids, MI

Beyer P 1994 *Religion and Globalization*. Sage Publications, Thousand Oaks, CA

Blumhofer E L 1993 *Restoring the Faith: the Assemblies of God, Pentecostalism, and American Culture*. University of Illinois Press, Urbana, IL

Blumhofer E L, Spittler R P, Wacker G A (eds.) 1999 *Pentecostal Currents in American Protestantism*. University of Illinois Press, Urbana, IL

Carwardine R 1978 *Transatlantic Revivalism: Popular Evangelicalism in Britain and America, 1790–1865*. Greenwood Press, Westport, CT

Featherstone M, Lash S, Robertson R (eds.) 1995 *Global Modernities*. Sage Publications, Thousand Oaks, CA

Freston P 1998 Latin American perspectives. In: Hutchinson M, Kalu O (eds.) *A Global Faith: Essays on Evangelicalism and Globalization*. CSAC, Sydney

Marsden G M 1980 *Fundamentalism and American Culture: the Shaping of Twentieth Century Evangelicalism, 1870–1925*. Oxford University Press, New York

McGrath A 1995 *Evangelicalism and the Future of Christianity*. InterVarsity Press, Downers Grove, IL

Noll M A, Bebbington D W, Rawlyk G A (eds.) 1994 *Evangelicalism: Comparative Studies of Popular Protestantism in North America, the British Isles, and Beyond, 1700–1990*. Oxford University Press, New York

Poewe K (ed.) 1994 *Charismatic Christianity as a Global Culture*. University of South Carolina Press, Columbia, SC

Pollak-Eltz A, Salas Y (eds.) 1998 *El pentecostalismo en América Latina entre tradición y globalización*. Ediciones Abya-Yala, Quito, Ecuador

Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London

Robertson R, Garrett W R (eds.) 1991 *Religion and Global Order*. Paragon House, New York

Smith T L 1980 *Revivalism and Social Reform: American Protestantism on the Eve of the Civil War*. Johns Hopkins University Press, Baltimore, MD

Synan V 1997 *The Holiness-Pentecostal Tradition: Charismatic Movements in the Twentieth Century*, 2nd edn. W. B. Eerdmans, Grand Rapids, MI

Thomas G M 1989 *Revivalism and Cultural Change: Christianity, Nation Building, and the Market in Nineteenth-Century United States*. University of Chicago Press, Chicago

M. Hutchinson

Christianity in Anglo-America and Australasia

Historically, the major religious traditions for Anglo-America and Australasia are similar, even though the social contexts in which they are located vary considerably. That they all function, more or less, as cultural establishments within their countries provides a substantial basis for looking at them as a unified religious bloc. The countries making up this large bloc extending across continents—that is, Canada, the United States, Australia, and New Zealand—share origins as English colonies and, thus, are relatively young as nations and bound by Anglo legacies of language and culture. While the countries themselves are relatively young, in another sense the Protestant cultural establishments within them are increasingly viewed by many people as old and faltering—yet another reason for looking at them as of one piece. For all these reasons, comparative analysis of religious trends and dynamics is both possible and desirable, despite some obvious differences in size, polity, and demographics for the various countries.

1. Historical Considerations

From the outset, we should not underestimate the extent to which Anglo religious life continues to be fundamentally shaped by two historical realities, first, a Protestant background, and thus a theological heritage of resistance to religious authority and, second, a common experience and trajectory in the encounter of religion and modernity in the West. Obviously the two are closely related. The Protestant

Reformation unleashed religious energies previously contained in medieval Catholicism and loosened the hold of ascriptive loyalties, thereby giving rise to greater autonomy of the individual in matters of faith and morals. Greater religious choice was accompanied by an elaboration of religious forms and styles, or a further working out generally of what is often referred to as the 'voluntary principle' within Protestantism. This principle is most pronounced in the United States, but elsewhere with Anglo influence, even in England itself, religious establishments have functioned in the modern period in an environment of considerable individual freedom. Protestantism has both given shape to, and been shaped by, high levels of individual freedom. In fact, three quite distinct historical waves of Protestant dissent legitimizing greater individual autonomy may be identified: first, the Calvinist movements; second, the Methodist revivals; and more recently, the Pentecostals. As David Martin observes (1978), the three religious waves correspond to shifting, ever-widening spheres of social influence. Unlike the Calvinists who were limited mainly to social elites, the Methodists empowered working-class and middle-income populations, and the Pentecostals reached still lower social strata, and did so–and still does–in places and contexts often beyond the reach of the other two. For all three, the emphasis on the individual, on education and hard work, and on taking charge of one's life and making the most of it resulted in upward mobility, and with that, a re-focusing of theology and moral principles giving greater weight to the role of an individual's choice and conscience.

The encounter with modernity in the West led to a wide array of consequences, but most notably, to a loosening of the binding character of tradition and memory and an increased awareness of religious pluralism. Both of these mesh well with the mounting importance of individual autonomy and reliance upon personal choice as articulated religiously. With all this came foundational shifts, or what Peter Berger (1979) describes as a 'loss of ontological certainty.' Broadly speaking, the challenges to religious authority and increased pluralism ushered in an era of greater negotiation in matters of faith and commitment in the face of a widening array of choices. As the term 'secularization' is often used, it is much too encompassing and glosses over the nuances of religion's encounter with modernity to adequately describe what happens in this situation, but suffice it to say that the latter as a historical and increasingly global process, creates ruptures in shared religious views and forces upon ordinary people a posture of cognitive bargaining. Organized religion's monopoly over religious and spiritual questions is easily undermined. Religion's presence in the public arena undergoes a qualitative shift. For our purposes here, what is important is not just the parallels between Protestant influence and religion's encounter with the Enlightenment but the fact that, as Berger and many other commentators point out, Protestantism has confronted the modern situation longer and to a greater extent probably than any other religious tradition. That of course is rapidly changing in the contemporary world where no religious tradition can insulate itself from the pluralizing and individualizing trends. Yet for this very reason, the mainline Protestant experience is important and indeed paradigmatic of what other religions are now facing. Indeed, the fate of Protestantism in the modern West is what people of other faith traditions are quick to notice and worry about as perhaps their own. What they see are deep internal strains within the tradition and disputes between liberal progressives, on the one hand, and conservative traditionalists, on the other, over moral and religious principles and how to forge responsible religious styles in the contemporary world.

Having said all this, it is not necessarily the case that 'Protestant' translates to 'Anglo' and certainly not that Anglo faith traditions can claim any superiority in their responses to modern life. The spiritual malaise that has currently fallen upon the historic, so-called 'mainline' Protestant institutions throughout much of Anglo-America and Australasia augurs against any such claim of superiority. Nor is there any necessary presumption that other faith traditions must inevitably follow the Protestant trajectory in its confrontation with modernity; there are multiple passages and courses of development. Even to argue that the Protestant experience, given its strong Anglo connections, is paradigmatic for other religious traditions is to risk the charge that this is still another, albeit subtle claim to hegemony—alongside a history of racism, sexism, and capitalist exploitation. But we need not draw this latter inference, especially considering the visible, widespread examples of 'Protestantization' within other faith traditions, presumably the choice of conscious believers from within those traditions. Rather, we should look to the various countries described in this article as case studies in which to describe Anglo religious trends, and to examine how these old Protestant establishments view themselves and extent to which they still define and shape religious life within their environments. Writing some 30 years ago, sociologist Charles H. Anderson 1970 observed that 'The decline of the white Protestant majority and of white Protestant hegemony in twentieth century America has encouraged the growth of self-conscious Protestant community' (1970, p. 3). It is an observation about religious identity which applies, in varying degrees, to Anglo-Saxon populations in all the daughter societies of England, and thus a good reason why we should take a close look at these societies.

2. *United States*

We look first at the United States. It is the prime example of a country with a strong Protestant legacy where the religious norms of individualism are now

widely diffused and where the historic, so-called mainline Protestant denominations are mired in a deep spiritual malaise. Some would argue that the United States is the society that has undergone the greatest impact with modernity, that in many respects its beliefs and values are highly secular, or more precisely, that religion in this country is secularized 'from within,' and thus a model of sorts of what to expect in the modern world. Yet, as is commonly observed, by almost any standard of comparison the United States remains distinctive among modern nations given its high levels of church-going and religious membership. Despite significant shifts in religious styles, levels of religious participation for the country as a whole have not greatly changed over the past four decades. Protestant values encouraging the responsible practice of faith on the part of ordinary believers, and not just by religious elites or by politicians seeking votes, remains deeply ingrained in its public life.

It is widely accepted that the religious distinctiveness of the United States is explained by 'supply-side' thinking, that is, by a history of separation of church and state (the heritage of 'voluntarism') and a culture encouraging innovative, competitive religious leaders to gather followers around them and to organize new churches. Innovation in recruitment methods, in developing new organizations, and in the framing of religious messages coincides with a culture that stresses rational choices on the part of individuals. The fact that so many Americans 'switch' religious affiliations and move in and out of religious organizations frequently underscores a high level of choice and social accommodation. Indeed, there is much to support the rational-choice perspective on individuals and their religious affiliations and styles. In religion, as with the mass marketing of goods and services of all kinds, Americans respond to skillful packaging and presentation of the product. 'Selling God' is taken for granted in an economically-driven culture as the selling of anything else, despite the fact that preachers, priests, and rabbis often resist any notion that theirs is a message that in any way would or should be sold.

But there are other considerations involved in accounting for religious trends and dynamics in the United States. An important example is the downward spiral of the 'oldline' Protestant churches, some of them with histories dating to colonial times. It is much too massive a restructuring of American religious life to explain simply by the rise of televangelism or the use of any other innovative techniques for recruiting. Episcopalians, Presbyterians, the United Church of Christ (formerly, the Congregationalists), the Disciples of Christ, and the United Methodists all began to lose members in the mid-1960s and have continued to do so in the intervening years. The declines appear to be slowing down and may possibly be bottoming out, but the fact is that this switch in fate for these culturally established churches—made up largely of white, middle-class, Anglo types—came so suddenly and involved religious organizations so strikingly different in theological heritage and polity could only signal something deeply troubling in the culture at large. To begin with, the low and declining birth rates (well below the replacement level) finally caught up with these churches. The birth rates were already lower than for other traditions, but after the birth-control pill became available to the public the rates dropped even further—indeed, so low that many of these churches in the late 1960s and 1970s had so few teenagers they could not provide very effective programming. A 'generation gap' emerged as large numbers of the post-World War II generation effectively dropped out of active involvement. In fact, the best predictor of declines in these churches were the diminished church school enrollments 10 years earlier: fewer children were brought up in the churches, beginning in the early 1960s, which later translated into significant declines in worship attendance and financial contributions. The decade of the 1960s was itself an important turning point in a broader cultural sense. John F. Kennedy was elected the first Roman Catholic president, an event significant symbolically to a Protestant sensibility and its fears of losing power and influence. The antiestablishment ethos of that period brought on by the struggles over civil rights, the Vietnam War, and gender role and family changes likewise made for a distinct change of mood and outlook, working against those religious institutions that had long been closely identified with the mainstream values.

It is sometimes said that these old Anglo institutions have depleted their theological resources, that the problem is one of 'tired blood.' Certainly the institutions do suffer from a loss of energy and direction and for a long time have ridden on inherited cultural capital. The shift in demographics was itself not inconsequential: the size of the Protestant majority steadily decreased in the twentieth century, as did the ethnic consciousness and identity underlying many of the Protestant communities. Anglo populations and cultures generally feel the squeeze resulting from growing Hispanic and non-Christian faiths. Yet despite the erosions of consciousness and institutions, in a very real sense these churches triumphed to a degree within the culture: many of their goals arising out of the Social Gospel movement were achieved and, more to a theological point, the Protestant principle about believing for yourself and taking responsibility for your actions long inculcated within these traditions now emerged in, admittedly, a far more radical and liberated form. In the 1960s and 1970s, young people—many of them nominally Protestant—turned inward and explored religious alternatives; they sought after experiential faith in a more direct, intimate sense than they had usually found in the dry rituals of the established churches; they felt free to create their own personal collages of belief drawing from a variety of

sources and traditions, and worried less about theological consistency and more about the deeper meanings of whatever beliefs they held. Words like 'soul' and 'spiritual' which in the 1950s had all but disappeared in public discourse now returned with excitement. In this and so many other ways, signs pointed to a spiritual ferment which in its early phases that was not so much opposed to religion as alienated from existing bourgeois religious forms, and increasingly so for many college-educated, middle-class youth.

With the declines of the older religious institutions came other developments that point to major restructuring of American religion. One is the deep cleavage between liberals and conservatives, the latter having been rejuvenated in its crusades to restore a Christian society along the lines as defined by evangelicals and fundamentalists. Since the 1970s the latter have grown often picking up dropouts and dissidents from the oldline Protestant churches, and not without subtle appeals often—in the most conservative sectors—to reclaim an older, Anglo-based moral and religious order. While the thesis of a 'culture war' is easily overstated, there is much tension and occasionally overt conflict over unresolved issues like abortion, homosexuality, and prayer in public schools. Such disputes arise out of seriously conflicting, indeed incommensurate, notions of moral and religious authority—whether out of humanistic conceptions of the self and metaphorical views about truth in search of personal truth and happiness or literal interpretations of Biblical authority and a monarchical view of a transcendent God who commands people to obedience.

A second and related development is the rise of a full-blown spiritual quest culture that now permeates much of the nation's population—among mainline Protestants who find their own rituals and worship services in need of rejuvenation, among many 'seeker-minded' evangelicals who know very little about Christian tradition but are eager to adapt faith to their needs and concerns, and among secularists who turn to New Age religions (or more broadly, the 'New Spirituality') in search of inner truth and enlightenment (Roof 1999). This quest culture drives some Protestants to rediscover and reclaim their heritages; for the great majority, it seems to raise their levels of spiritual sensitivity more than their actual commitments to institutions. At present, there is considerable exploration not just of what was once called 'alternative religions' but of psychological teachings and inspirational literature, often of a rather generic quality; much influence of popular television programs and films that address spiritual themes directly, and many new entrepreneurs operating in the religious marketplace offering their versions of spiritual wisdom, holistic thinking, and mind-cure. In fact, many people who are interested in the New Spirituality have no difficulties acknowledging a Protestant past, which means that large numbers of Americans today combine old religious identities and new spiritual sensitivities with great ease, and in ways that appear to be personally rejuvenating but which to their grandparents would no doubt be incomprehensible.

3. *Canada*

Similar patterns of post-World War II Protestant decline are evident in Canada, despite the fact that the religious situation is very different in that country. Canada is historically less pluralistic, though that is now rapidly changing, and much more shaped historically by the presence of two large religious constituencies, Catholic and Protestant, each vying with one another for power and influence. Catholics have long been numerous in Lower Canada and the Protestants, especially Anglicans and Scotch Presbyterians, are sizable in Upper Canada. For the country as a whole, Protestants for a long time numbered more than Roman Catholics. But as of the national census in 1991, Roman Catholics, benefitting from a higher birth rate, edged above them with roughly 12,500,000 members, concentrated largely in Quebec. The United Church of Christ (a merger in 1925 of Congregationalists, Methodists, and some Presbyterians) is the next largest at approximately 2,000,000 members, concentrated in Ontario and Prairies; the Anglicans (plus the Orthodox) have substantial numbers, around one million. Baptists, Presbyterians, Lutherans, Pentecostals, and assorted other conservative Protestants are the next largest groups, most of them below a million members each. Birth rates for Mormons, the Salvation Army, and for conservative Protestant groups generally are higher than for the more established Protestant denominations.

But the fate of Protestantism in Canada differs in some important respects compared with the United States. Mainline Protestants have suffered membership declines there, losing both to the conservative churches and to Roman Catholic parishes. Proportionately they have lost more to Catholicism than is probably the case in the United States where just the opposite pattern appears to be more prominent—that is, somewhat greater losses to the religious conservatives. Statistically this is what one would expect: the larger the competing population, the greater the likelihood of mainline Protestants switching in, as a result of marriage or by religious choice. The conservative religious presence in Canada is neither as well-institutionalized nor as publicly visible as in its neighbor to the South. Sociologist Reginald W. Bibby (1993) argues in fact that the alleged recent growth of conservative Protestantism in Canada is largely a misperception, more really a 'circulation of the saints,' or movement from one evangelical church to another, than a case really of active and successful recruitment

from other churches. Protestants on the whole, and conservative Protestants in particular, tend to be converted over and over again—as a result of social and geographic mobility, marriage, divorce, friendship patterns, preference for religious leaders, and congregational activities. Bibby's research suggests some slight increase in the recruitment of outsiders into the conservative fold over the past two decades, but even with these increases the actual number of converts remains small.

Overall, levels of religious participation in Canada are moderate—higher than those of most European countries but lower than found in the United States. Conservative Protestants have higher levels of weekly worship attendance than other Protestants and Roman Catholics, but because of their relatively small size proportionately and tendency to circulate among themselves their impact is not as widely felt as in the United States. The liberal–conservative cleavage within Protestantism is certainly evident, but the 'culture-war' infrastructure in Canada is weaker than in the United States, and therefore less of structural feature within the society. The conservative-moralist flavor of evangelical Protestantism never had sway over Canadians in the way historically it did, and to some extent still does, in the United States. Of greater significance is the influence of the New Spirituality which functions less as a separate enclave than to permeate religious communities of all kinds, and especially the more moderate-to-liberal communities. Again, to quote Bibby, Canadians are very much into 'religion a la carte.' Or, as he explains, 'New Age religion seems to follow the pattern of most new entries into the Canadian religion market, offering consumers optional item that can be added to more conventional religious beliefs and practices' (Bibby 1993, p. 52). It is not clear whether this is more true in Canada than in the United States; in both countries great numbers of people interested in New Age and metaphysical thought continue to identify, to a considerable extent, with an inherited faith tradition—as Protestant or Roman Catholic. This represents a major convergence for the two countries in the adaptation of religion to modernity. In both of these highly individualized settings, religion is not so much abandoned in some strict sense as it is privatized and modified to fit into people's life-situations at any given time.

4. *Australasia*

The very term 'Australasia' signals a confrontation and mixing of cultures, of East and West, a term quite fitting to both of our remaining countries, Australia and New Zealand. Both countries have a heritage of European settlements and Anglo dominance over aboriginal cultures, with the practices and structures of that heritage reflected in their social institutions. Both countries know the tensions arising out of an expanding religious pluralism, brought on largely by successive waves of new migrants. Both are often described as increasingly secular societies, especially in political and civic life.

Yet there are significant differences. Australia has seen fewer new religions or even new versions of older religions than has New Zealand. Religious history for the former is more a telling of stories about a European past unlike for the latter where such stories are shaped more by local conditions and cultural and religious admixtures. Even Anglican church history differs. In Australia, the Anglican Church began as a convict church served by chaplains for whites, whereas in New Zealand, the Anglican Church was at first, and for a long time afterwards, a Maori aboriginal church served by missionaries. It was not until the mid-1800s that the Pakeha, or white Anglicans, began to outnumber the Maoris. The Maori Ratana and Ringatu faiths are two examples of New Zealand's popular mixing of religious and cultural themes, combining aboriginal and Christian elements. This difference between the two countries need not be exaggerated, but it does register in religious and spiritual styles.

Australia's religious profile today is as follows: Roman Catholics, 27 percent; Anglicans, 22 percent; Presbyterians, Methodists, and Uniting Church, 13 percent; other Christian, 11 percent; other Religions, 3 percent; no religion or not stated, 23 percent. The Catholic ascendency over Anglicans emerged about 50 years ago. The old, well-established religious communities—the Anglicans, the Uniting Church, the Presbyterians, and Lutherans—have all suffered declines in absolute numbers during the past two decades. The Christian or Christian-derived groups growing include the Pentecostals, Oriental Christian, Jehovah's Witnesses, and Mormons. Non-Christian groups rapidly increasing but still relatively small proportionately are Buddhists, Hindus, and Muslims, representing the latest of migrating populations into Australia. The greatest change in religious identification since World War II, and particularly among those who were born in this period, is the increase in people having 'no religion' or choosing not to state a religion in the national census. Since the 1970s this population has doubled. Likewise, religious attendance has declined in most of the churches while, somewhat paradoxically, stated interest in spiritual well-being appears to have increased.

In New Zealand, the Anglicans remain the largest single constituency considering that both the Pakeha and the Maoris are included. Roman Catholics are the next largest population, followed by Presbyterians with a continuing, strong Scottish heritage. About 50 percent of the population identify with one of the major Christian churches; by contrast, the evangelicals and Pentecostals appear not to have had as much success in recruiting in New Zealand as they have in Australia. Recent migrant religious communities are

growing but are infinitesimal in size compared with all others. There have been substantial declines in religious affiliation and congregational involvement since World War II, especially for major churches, and mostly as a result of trends among younger whites. A third of the population currently report having 'no religion' or object to stating what it is. That the white and non-white populations are on somewhat different religious trajectories is apparent, evident most in the Anglo trends toward reduced organizational involvement; for the Maoris with their folk traditions and close spiritual attachments to the environment, organizational participation was never held up as a religious norm in quite the same way it was for Anglos.

Both Australia and New Zealand have experienced enormous changes in the past half-century, as they evolved from being made up predominately of villages and tribes to becoming more urban and multicultural in character. Religion has lost much of its power to integrate life experiences and has become one among many life-worlds in which people may or may not participate. Anglo religious traditions have suffered in the process. There, as elsewhere in modern societies, we observe greater eclecticism and pragmatism. However, compared with Canada, and even more so with the United States, two things stand out making the religious responses to modernity in Australasia different: one is that fundamentalism seems to attract fewer people and offers less of a viable alternative, and the second is there is less movement generally from one faith tradition to another, the more common movement being simply to leave organized religion in favor of 'no religion.'

See also: American Studies: Religion; Catholicism; Christianity in Asia; Christianity in Central and South America; Christianity Origins: Primitive and 'Western' History; Church and State: Political Science Aspects; Nationalism: General; Religion and Politics: United States; Religion: Definition and Explanation; Religion: Nationalism and Identity; Secularization

Bibliography

Anderson C H 1970 *White Protestant Americans: From National Origins to Religious Group*. Prentice-Hall, Englewood Cliffs, NJ

Berger P L 1979 *The Heretical Imperative: Contemporary Possibilities of Religious Affirmation*. Anchor Press, Garden City, NY

Bibby R W 1993 *Unknown Gods: The Ongoing Story of Religion in Canada*. Stoddart, Toronto, Canada

Bouma G D 1992 *Religion: Meaning, Transcendence and Community in Australia*. Longman Cheshire, Melbourne, Australia.

Martin D 1978 *The Dilemmas of Contemporary Religion*. St. Martin's Press, New York.

Roof W C 1999 *Spiritual Marketplace: Baby Boomers and the Remaking of American Religion*. Princeton University Press, Princeton, NJ

Wuthnow R 1988 *The Restructuring of American Religion: Society and Faith Since World War II*. Princeton University Press, Princeton NJ

W. C. Roof

Christianity in Asia

Christianity in Asia has an ancient history though it remains a minority religion. There are approximately 85 million Christians in Asia but more than 40 percent of them live in just two countries, the Philippines and South Korea. The minority status of Christianity in Asia gives it a distinctive sociological character in relation to Christianity in other regions of the world. For many Asian theologians this distinctive sociological context has meant that the particular focus of their articulation of the Christ of Asia is that of dialogue with the dominant religious and cultural traditions of Asia. In the postcolonial era a strong element of Christian social militancy was observable in many Asian countries. This militancy was associated with the ecumenical Christian Council of Asia, and especially its sponsorship of Urban Rural Mission in a number of countries. The Catholic Federation of Asian Bishops Conference is also an organization which has confronted frequently social and political controversies in particular Asian countries. Postcolonial Asian Christian social protest has focused on various concerns including the continuing social power of caste in India, workers' rights in South Korea and Southeast Asia, and democratic reform of authoritarian political regimes which characterise government in much of the region. A more recent form of militancy in Asian Christianity is associated with the rise of charismatic/Pentecostal and fundamentalist/evangelical groups which have arisen within and beyond the denominational boundaries and institutions established during the colonial era. Adherents of the new charismatic and evangelical forms of Asian church regard Christian worship, spirituality, and personal evangelism, rather than social transformation in the wider society, as the essential vocation of Asian Christians. This more transcendental and salvific emphasis represents a clear rejection of the social protest of earlier forms of postcolonial Asian Christianity. The new militancy is more focused on claims to spiritual power for the new forms of Asian Christian spiritual and congregational life, and the rights of all Asian Christians to freedom of worship, and to proselytise their neighbors. Just as the earlier ecumenical and socially radical militancy was the occasion for formal State resistance in South Korea in the 1970s, and in Singapore, Malaysia, and

Hong Kong in the 1980s, so Asian Christians now find their claimed freedoms to worship and evangelize occasion increased social scrutiny and resistance amongst majority religious and cultural groups and their political leaders.

1. South Asia

There is evidence of Christian activity in South India and Sri Lanka from the sixth century. Kerala is the region where most activity was concentrated, and Christians are still strong there, comprising 20 percent of the population. Christians also comprise a significant proportion of the population of Tamil Nadu. Portuguese colonisers saw a connection between trade and religion and, with the aid of the Jesuit order, encouraged the majority of the population of their settlement in Goa to convert to Christianity by the end of the sixteenth century. Under British rule the East India Company resisted the deployment of missionaries, judging that their presence would exacerbate existing religious tensions between Hindus and Muslims. But in the nineteenth century missionaries were allowed into most of British India. Converts were few and most were from marginal tribal groups, or from untouchables who have much to gain by abandoning Hinduism for Christianity. Since the 1980s a vocal Dalit (untouchable) movement has emerged which is radically anticaste and draws on Latin American-style liberation theology for its inspiration. With the rise of religious politics in India since the Premiership of Mrs Gandhi, religious conflict has become endemic in many parts of India and this has had an increasing impact on Christians. In Sri Lanka Christians are most prominent in the West and make up 30 percent of Colombo's population.

2. Christianity in China

Christianity in China goes back to the migration of Nestorian Christians from Iraq to China along the trade routes in the seventh century CE. The Jesuits arrived in China in the sixteenth century under the leadership of Matteo Ricci who experimented with a radical indigenisation of Christianity to Chinese culture. His efforts at cultural translation were controversial in Rome and the Pope insisted on a reversion to Roman rituals and traditions in 1744. After the opium wars and the Nanking treaty of 1842, Catholic and Protestant missionaries settled in many parts of China. The association between Christianity and Imperialism was strong and conversions were few until the failure of the Boxer rebellion at the end of the nineteenth century when many, especially young, Chinese turned to Christianity as a potential source of social and spiritual restoration. With the advent of Communist government in 1949 all foreign missionaries were removed from China and churches closed. Ironically the exclusion of foreign missionaries allowed Chinese Christianity to flourish in a way it had not before and the Christian presence in China today is much greater than it was at the time of the Communist revolution. Most of the officially recognized churches in China, whether Catholic or Protestant in origin, are known as 'three-self' churches, which means that they are not dependent on extra-Chinese ecclesiastical authority, nor do they rely on funds from overseas. There are also a large number of unofficial churches and congregations in China whose leaders try to circumvent controls on officially recognized churches, and in particular controls on proselytism. Evangelistic and worship meetings are held in secret, in apartments, or in open spaces away from city or town centres. The mass migration of many tens of millions of rural migrants into towns and cities in China since the 1980s, the largest contemporary movement of people on the planet, has occasioned conversions to new religions, including Christianity, as migrants seek to find a substitute for village-folk religion. Churches in urban areas, and especially in southern China, are currently growing very fast, fuelled by the spiritual and ritual vacuum occasioned by the Cultural Revolution and the enforced break-up of traditional Chinese religions, and by intellectual and popular dissatisfaction with secular Communist ideology.

3. Hong Kong, Taiwan, Japan, and Korea

One fifth of the population of Hong Kong are Christian and just under half are Catholics. The principal Protestant and Catholic traditions are represented and there are also a large number of independent and Pentecostal churches which have become established since the 1970s. The majority of Christians in Taiwan are Presbyterian, reflecting a missionary drive in the nineteenth century. Many new Christian groups emerged after the emigration of supporters of the government of Chian Kai-sheck to Taiwan from mainland China. The Jesuit missionary Francis Xavier introduced Christianity to Japan in 1549, after his initial mission to India who converted the Japanese court to Christianity. More than 300,000 Japanese followed their rulers and converted by the end of that century. However, Christianity was later proscribed, though small pockets of Christian activity remained. In the twentieth century Japan has been characterized by remarkable religious innovation, and many New Religious Movements have emerged, particularly since the Second World War. Parallel developments have occurred in Japanese Christianity. New indigenous styles of Christianity have emerged, and Pentecostal groups have also proliferated.

The Catholic Church has roots in Korea as far back as the sixteenth century but Protestantism has taken

root in Korea in a manner it has nowhere else in Asia. Forty-one percent of the population are now Christian, and the majority belong to indigenous Korean and Protestant churches. The beginning of the twentieth century saw a particularly rapid influx of Koreans into the Protestant churches, and many Protestant churches have also experienced considerable growth since the Second World War. Seoul has some of the largest church buildings, and the largest congregations, anywhere in the world. Explanations for the strength of Protestantism in Korea are various. One theory is that there is a uniquely individualist strain to Korean Confucianism and that Protestantism individualism, with its more progressive conception of the role of the individual in the fast changing world of the twentieth century, proved particularly attractive to Koreans who were dissatisfied with Confucianism.

4. *Southeast Asia*

As a consequence of more than three centuries of settlement by the Spanish, and after the Spanish American war, by the United States, Christianity is the principal religion of the Philippines. Virtually the whole country was converted to Roman Catholicism under Spanish rule, though a Muslim minority remained, primarily in Mindanao. The Roman Catholic Church, however, was organized and run exclusively by foreign clergy and resentment at this resulted in an internal revolution whose outcome was the Philippine Independent Church which broke away from Catholicism after 1860. Protestant missionaries arrived with the Americans, including Episcopalians who set out to convert the mountain peoples in North Luzon as they had not been reached by the Catholics.

Roman Catholic missions were active in Vietnam, Cambodia, and Laos in the seventeenth and eighteenth centuries, and established churches in all three countries. Cambodia now has very few Christians indeed. After Year Zero Christians were totally purged from Cambodia. In Laos the Catholic Church has had most converts amongst tribal peoples on the border with Thailand but remains an insignificant presence elsewhere in the country. In Vietnam Christians, the great majority of whom are Catholic, constitute 7 percent of the population. The church was indigenized much more effectively in the North, whereas in the South local leadership was discouraged and there are far fewer adherents. Catholics have been close to political leadership in both North and South Vietnam and the Catholic Church in Rome and in Vietnam was a strong supporter of moves for peace during the war, and of subsequent North–South reconciliation efforts after the war.

Uniquely in Southeast Asia, Thailand has never been colonized by a foreign power and Thai Buddhism is particularly resistant to foreign, including Christian, influence. Strong Chinese communities in the large cities, and especially in Bangkok, are the main locus of Christian adherents. Myanmar Buddhists are also highly resistant to Christianity. Christians are most numerous amongst the Karen tribal group in the forests of the Northwest who are not Buddhist, and do not regard themselves as Burmese.

Christianity was introduced into the Indonesian archipelago by the Dutch in the sixteenth century. Christians are most numerous in Northern Sumatra where the non-Muslim Batak people converted to Christianity in large numbers. There are also churches in most areas of Java, and on some of the other islands. Indonesian Christianity has been subject to the State religious doctrine of Panchasilla which requires that all religious groups express loyalty to the state and tolerance of other religions. Indonesian Christians were active in the struggle for independence from the Dutch after the Second World War and their partnership in the birth of the new independent nation may partly explain why, almost uniquely in the Muslim world, there have in the past been a number of religious conversions. However, a peaceful inter-religious climate characterized by neighborliness and dialogue between the religions recently has turned to conflict as the Suharto and Habibi governments have stirred up religious conflict as a way of maintaining a hold on the votes of the predominantly Muslim population.

Christians in Malaysia number around 7 percent of the population, and 15 percent in Singapore. Dating back to the Portugese settlement of Malacca in the fifteenth century, Malaysian Christianity has taken up a range of influences including Portugese Catholic, Dutch Presbyterian, American Methodist, Anglican, and Pentecostal. Christians are confined to Chinese and Indian minorities on the whole, though tribal peoples in East Malaysia have converted to Christianity in large numbers. The fastest growing forms of Christianity in both Singapore and Malaysia today are Pentecostal or Charismatic. This modern style of Christianity has spawned many independent churches and has also been very influential in mainstream Catholic and Protestant congregations. Both countries have experienced a surge in religious interest and affiliation in recent years, sparked in part by Islamic resurgence among the Malays. This is reflected in a range of religious innovations both within Islam, Buddhism, and Hinduism, as well as in Christianity.

See also: Catholicism; Christianity: Evangelical, Revivalist, and Pentecostal; East Asia, Religions of; India, Religions of; Religion: Evolution and Development; Religion, Sociology of

Bibliography

Ackerman S E, Lee R 1988 *Heaven in Transition*. University of Hawaii Press, Honolulu, HI

Barrett D B (ed.) 1982 *World Christian Encyclopedia: A Comparative Study of Churches and Religions in the Modern World AD 1900–2000*. Oxford University Press, Nairobi, Zimbabwe
Digan P 1984 *Churches in Contestation: Asian Christian Social Protest*. Orbis Books, Maryknoll, NY
von der Mehden F R 1986 *Religion and Modernization in Southeast Asia*. Syracuse University Press, Syracuse, NY
Neill S C 1985 *History of Christianity in India, 1707–1858*. Cambridge University Press, Cambridge, UK
Palmer S J 1967 *Korea and Christianity: The Problem of Identification with Tradition*. Hollym Corp, Seoul, South Korea

M. Northcott

Christianity in Central and South America

Throughout the pre-Hispanic, colonial, and contemporary periods, the history of religion in Central and South America has been inextricably linked to political structures, rituals that ordered everyday life, and movements of rebellion, with Catholicism having played a central role for almost four centuries. After the break up of the monopoly exercised by the Catholic church and the rapid growth of Evangelical churches there is growth, but also fragmentation; creativity as well as uncertainty; individualism, but also the constitution of new communities; freedom, along with the authoritarianism of the leaders of the base communities or of the Pentecostal churches; political quietism side by side with right-wing activism. There are even churches that reversing the traditional direction have established branches outside Latin America.

1. *Religion Before the Conquest*

As is generally the case in the preindustrial world, in which ideological claims are advanced in mythological terms and reinforced through ritual means, in the area now known as Latin America the social order was religiously legitimized. A proper discussion of the various ways in which this religious legitimization took place would require considering the range of societies found in Meso and South America, from kinship-based groups, found for instance in the Amazonian area, to chiefdoms, to states. Since such an examination cannot be undertaken here, it may suffice to say that societies organized as chiefdoms tend to be theocratic, their political life being heavily ritualized; this also applies to early states, such as the Inca and the Aztec, in which the rulers were sacralized and were considered as the pivot of their society. Thus, in the Andean world the Inca dynasty, originally based in the area of Cuzco in southeastern Peru, claimed descent from Inti, the Sun.

More generally, religious legitimation meant that social distinctions, obligation to work, and access to resources were regarded as being regulated by an order that encompassed the social and the physical world. In order to maintain this order it was necessary to structure time and space, as well as the relations between age groups and the sexes. The organization of time was accomplished at several levels: at the individual level, this involved rites of passage; at the macrosocial levels, on the other hand, the structuring required the use of liturgical calendars which were related to the seasons and, above all, to agricultural work. Similarly, space was structured by establishing mythologically based distinctions among neighborhoods and, at a more extended level, by dividing a territory by means of shrines along imaginary lines, an example of which is the *ceque* system centered in Cuzco, the Inca capital.

2. *Conquest, Resistance, Accommodation*

The fact that the political, religious, and economic spheres were less differentiated in the Aztec and Inca states than in sixteenth-century Western Europe may have contributed to the speed of the victory of the Spanish and Portuguese adventurers. The 'modernity' of the conquerors should not be exaggerated, however, for the Iberian discovery and conquest of America took place immediately after the defeat of the Muslim kingdom of Granada, and the expulsion of the Jews who had refused to convert to Christianity. It can be said, therefore, that the conquest itself took place as a kind of crusade at a time of politico-religious triumphalism. Moreover, attention should be paid to the fact that Columbus seems to have endowed his enterprise with a millenarian aura, hoping that a portion of the wealth found in the Indies would be used in the conquest of Jerusalem. This millenarian component can also be found at work in the self-understanding of some of the members of the religious orders involved in the process of conversion of the vanquished.

In any event, while the crusading spirit validated the conquest, the transcendentalization of religion and the concomitant separation between the roles of priests and those of soldiers and administrators allowed some clerics to denounce the abuses of conquerors, without, however, condemning the conquest itself. During the colonization of Latin America, therefore, Christianity fulfilled the double role traditionally accomplished by religion: that of validating a social order, beginning with the justification of the conquest itself, while also providing the justification for judging that same order.

Religions' double function can also be seen at work among the indigenous population, for whom Christianity served as a vehicle of accommodation as well as of protest. The first aspect involved what is

generally known as syncretism, that is, the amalgamation of Christianity and indigenous religions. Thus, the Christian god incorporated elements of Andean divinities such as Wirakocha and Pachacamac, while the Virgin Mary, as the Virgin of Guadalupe, subsumed the Aztec goddess Tonantzin, 'Our Mother'; Quetzalcóatl, on the other hand, was assimilated to the apostle Thomas. The fact that the physical appearance of some of the inhabitants of the Christian pantheon was said to resemble that of the Indian population, rather than that of the conquerors, indicates on the one hand the extent to which the subjugated population had assimilated the new religion, and on the other, the intimate connection between religion and ethnic identification.

In some cases, conquered elites sought the protection of some of the most bellicose inhabitants of the Christian pantheon, expecting to profit from the power these supernatural beings had shown on the Spanish side; thus, the apostle Santiago, patron of Spain, became syncretized with the Andean Illapa (Lightning), and already in the sixteenth century the Virgin of Copacabana was enlisted in intradynastic struggles. Christianity also played a role in the indigenous rebellions against Spanish domination. This can be seen already on the sixteenth-century Taki Onqoy, and more clearly in the late eighteenth-century Tupac Amaru rebellion, in which the rebels regarded themselves as better Christians than the Spaniards.

3. *Religion in the New Republics*

The accommodation between indigenous religions and Christianity should not obscure the fact that from the beginning of the conquest, in order to forestall challenges from within, Spanish authorities sought to destroy the indigenous religions, undertaking 'extirpation of idolatries' campaigns; similarly, in order to protect Catholic orthodoxy from external threats, Spanish authorities tried to isolate the colonies from Protestantism, Judaism, and liberalism: Christianity—the Baroque Christianity of the Counter Reformation—served as one of the bastions of Spanish power during the three centuries of colonial rule. The role played by religion in the wars of independence of the early nineteenth century was ambiguous. Whereas the lower clergy tended to side with those who wanted to achieve independence from Spain, the prelates, many of whom had been born in the Iberian peninsula, were generally in favor of maintaining the colonial regime. Once independence was achieved in the nineteenth century, the Creole elites kept in place the church's prerogatives, while at the same time seeking to control it, just as the kings of Spain had done. After the governments of the new republics and the Vatican signed concordats regulating the status of the church, Roman Catholicism generally enjoyed a privileged position, one that was enshrined in the republics' constitutions.

The situation, however, was not always as favorable as clerical groups would have wanted. As in Spain itself, the very hegemony of the church contributed to an atmosphere of anticlericalism which led to attempts to curtail that institution's privileges; this anticlerical attitude was intensified by the influence of liberal and positivist currents. The most radical measures against Roman Catholicism were taken in Mexico, a country that re-established relations with the Vatican only in 1992. In general, however, Latin American governments have regarded the church as an ideological ally, one which played a crucial role in the legitimization of the social order.

4. *Religion and Everyday Life*

Encompassing domestic devotions and public anticlericalism, individual rites of passage and official liturgies, Catholicism has been an omnipresent reality in the life of Latin Americans. But, as one would expect, Catholicism has been lived differently by men and by women, by members of upper and lower classes, in urban and rural milieus. As in southern Europe, in Latin America everyday religion in its private and public forms has been the domain of women; thus, except for special occasions such as Lent and Christmas, women have constituted the majority of the faithful in Sunday services. Similarly, it is the mothers who traditionally have been in charge of introducing children to the rituals and the moral injunctions of Catholicism. It is in the process of religious socialization that social class plays an important role, as children of the middle and upper classes are generally subject to religious indoctrination in schools directed by members of religious orders, whereas those who attend state schools are subject to minimal religious education.

One of the consequences of this religious indoctrination is that Catholic revival movements such as Catholic Action, or political parties linked to Catholicism, such as Christian Democracy, involve mainly middle and upper class persons. In large cities, public religious activity is usually restricted to Sunday masses and to occasional processions, in some of which participate government dignitaries. The legitimizing role of Catholicism as the state religion can be seen at work in the solemn liturgies with which church and state commemorate independence day, and even more so in the consecration of entire countries to Jesus or the Virgin Mary.

4.1 *Popular Religion*

The counterpart of official Catholicism is represented by what is known as 'popular religion'—that is, those symbolic practices that seem to want to escape, and even to oppose, the control of religious elites. Ambivalence is at the core of popular religion, since in

order to express opposition it is necessary to make use, against the grain, of the symbolic resources of the institutions which one opposes. Popular religion is part and parcel of a complex system of assimilation and rejection, through which subordinate groups sometimes negotiate and others struggle over access to cultural and in the last instance material goods. Latin American popular religion comprises the syncretic processes referred to above, in which elements of the symbolic world of the Spanish conquerors were incorporated into that of the Andean, Mesoamerican, and other conquered groups, as well as contemporary ritual practices in urban and rural milieus, such as processions and pilgrimages, which represent the counterparts of the liturgical functions choreographed by the official Church.

Despite the current academic infatuation with symbolic resistance, however, it must be said that the parodic, carnivalesque, rituals of popular religion generally have not constituted serious threats to the hegemony of Catholicism or of the state. In fact, the expenditures required by the ritual calendar of indigenous communities instead of leveling economic disparities may increase them.

4.2 *Millenarian Movements*

In general, in order for popular religious practices to turn into insurrections it has been necessary that the state attempt to suppress in a violent manner groups that have sought to isolate themselves from a changing society in order to live according to the teachings of the Gospel under the guidance of a charismatic leader. This happened, among other places, in northeastern Brazil from 1893 to 1897, in the millenarian movement of Canudos; it happened again in southern Brazil from 1912 to 1916 in the Contestado rebellion. Both the Canudos and Contestado movements constituted reactions against the secular, positivistic values of the Brazilian republic—and in the case of the Contestado rebellion it is clear that it was triggered by the dislocations brought about by capitalism. Like millenarian and messianic movements in general, those that have taken place in Latin America have involved marginal regions in transitional periods, and have been led by marginal, literate individuals, such as Antonio Conselheiro, leader of Canudos, and Miguel Lucena Boaventura, known as José Maria, leader of Contestado.

5. *Liberation Theology*

At the elite level, the privileged status of Catholicism in Latin America has not gone unchallenged either. In the nineteenth century, anticlerical groups, influenced by positivism, had already decried the church's influence; similarly, in the twentieth century politically progressive groups opposed the alliance between oligarchic governments and conservative church. It was, however, during the early 1960s, prompted first by the triumph of the Cuban revolution and then by the Second Vatican Council, that the traditional equation between church and conservative politics was broken. The Cuban revolution, combined with the Council's attempts to re-examine the relation between the church and the world, forced some members of the Latin American clergy to question the role the church had played in the maintenance of an unjust social order. That questioning led to the emergence of a 'prophetic' understanding of Christianity known as the Theology of Liberation. Influenced by the social sciences as much as by traditional methods of theological exegesis, the liberation theologians removed concepts such as sin and salvation from the mere personal and spiritual realms to one that took into account the structures of the societies within which sins were committed and salvation was sought. This theological approach was attacked by conservative Catholics, clerical and lay alike, who accused its proponents of forsaking Christianity and falling prey to Marxism. It cannot be emphasized enough, however, that despite its apparent threat to Catholicism, the theology of liberation constituted an attempt to save Latin America's dominant religion by enlisting it in the task of reforming society. That ultimately, attacked by the local churches and neutralized by Rome, the theology of liberation gave up its distinctive use of sociological categories and retreated into traditional spirituality, was to be expected, given that the liberation theologians were, after all, theologians.

6. *Alternatives to Catholicism*

But Rome's victory was a hollow one, for the presuppositions shared by both conservative and revolutionary clerics—that Latin America is and will remain Roman Catholic—proved to be illusory. As the conflicts between Rome and the radical theologians were taking place, a far more important development was unfolding, namely the spread of Protestantism—or, to be more correct, of Evangelical forms of Christianity. Nevertheless, despite the fact that the triumph of Protestantism may have demonstrated the ultimate irrelevance of the confrontation between Rome and the theology of liberation, there is a strong affinity between this theology, which could be considered as a Protestant Catholicism, and the evangelical churches. In both cases we find a puritan faith centered upon the Word, suspicious of the rituals of official Catholicism as well as of those that constitute popular religion. It is as if both movements away from a sacramental, hierarchical, agrarian-based vision of the world had developed as a response to the

weakening of the agrarian economy with its rituals and mythologies in order to come to terms with a predominantly urban world ruled by the market.

But, instead of the 'disenchantment' of the world, to use Weber's term, we find two religiously ambiguous, half-enchanted formations, the Evangelical and the Catholic, whose task it is to mediate between the two worlds. In both versions of Christianity there is still the attempt to preserve, if only though the Word, the presence of the transcendent, while at the same time deritualizing everyday activities. The causes and consequences of this deritualization have been explained in terms of a rationalization of economic activities, involving, for instance, the avoidance of the burden represented by wasteful ritual expenditures, and in more general terms turning away from the pressure exercised by community and tradition. The result is a world of ever widening economic disparities which must be confronted in an instrumental manner by the individual as individual, rather than as a member of a community. But, rather than repressing emotion, individualism and instrumental reason seem to exacerbate it. Thus, as in the tense coexistence of Pietism and Enlightenment found in eighteenth-century northern Europe, we find that the strongest forms of Latin American non-Catholic Christianity are Pentecostal; this has led to the emergence of charismatic forms of Catholicism—one more example of the proliferation of religious offerings in the religious market that emerged after the break up of monopolistic Catholicism.

In this context reference should be made to the growth of New Age religions among the middle and upper classes, as well as of religions of African origin among groups which are not of African descent, and in countries, such as Argentina, which do not have a significant black population.

While the early spread of Protestantism was caused by the activities of North American missionaries, the subsequent growth of evangelical churches is now largely fueled by local ferment. In some countries, such as Puerto Rico, El Salvador, Guatemala, Chile, and Brazil, evangelicals constitute a significant proportion of the population, whereas in others, such as Venezuela and Colombia, that presence is much smaller. Considering, however, the continuous growth of non-Catholic forms of Christianity, it can be said without reservation that the equation between Latin American religion and Catholicism, presupposed, among others, by the advocates of liberation theology, is a thing of the past. It is true that the Catholic church can still exercise pressure upon governments and legislatures on matters related to the control of female sexuality, but even in these cases, despite the fact that abortion is still illegal in Latin America, birth control campaigns are not uncommon, as the birth rate has decreased in the last several decades, a fact which shows that the influence of the church on the everyday life of ordinary people has diminished substantially.

6.1 Conservative Trends in Catholicism

A development that will likely contribute to the further loss of popularity of the Catholic church among the lower classes, while perhaps increasing it among high-income groups, is the growing importance of conservative Catholic groups such as the Opus Dei, a secretive organization supported by the Vatican. But even in the case of the Opus Dei, one can see its promotion as an attempt on the part of the church to accommodate itself to a changing world. In effect, despite its traditionalism and authoritarianism, the values promoted by this organization are consonant with the demands of a capitalist order, and in more general terms with the demands of modernity.

See also: Catholicism; Christianity: Evangelical, Revivalist, and Pentecostal; Latin American Studies: Religion; Latin American Studies: Society; Religion, Sociology of

Bibliography

Annis S 1987 *God and Production in a Guatemalan Town*. University of Texas Press, Austin, TX

Bastian J-P 1997 *La mutación religiosa de América Latina. Para una sociología del cambio social en la modernidad periférica*. Fondo de Cultura Económica, Mexico City

Burga M 1988 *Nacimiento de una utopía. Muerte y resurrección de los incas*. Instituto de Apoyo Agrario, Lima, Peru

Diacon T A 1991 *Millenarian Vision, Capitalist Reality: Brazil's Contestado Rebellion, 1912–1916*. Duke University Press, Durham, NC

Duviols P 1971 *La lutte contre les religions autochtones dans le Pérou colonial (L'extirpation de l'idolatrie entre 1532 et 1660)*. Institut Francais d'Études andines, Lima [1977 *La destrucción de las religiones andinas (durante la conquista y la colonia)*. UNAM, Mexico City]

Flores Galindo A 1988 *Buscando un Inca: identidad y utopía en los Andes*. Editorial Horizonte, Lima, Peru

Garrard-Burnett V, Stoll D (eds.) 1993 *Rethinking Protestantism in Latin America*. Temple University Press, Philadelphia, PA

Krickeberg W, Trimborn H, Müller W, Zerries O 1961 *Die Religionen des alten Amerika*. Kohlhammer, Stuttgart [1968 *Pre-Columbian American Religions*. Weidenfeld & Nicolson, London/Holt, Rinehart and Winston, New York]

Lafaye J 1984 *Mesías, cruzadas, utopías. El judeo-cristianismo en las sociedades iberoamericanas*. Fondo de Cultura Económica, Mexico City

Levine R 1992 *Vale of Tears: Revisiting the Canudos Massacre in Northeastern Brazil, 1893–1897*. University of California Press, Berkeley

Martin D 1990 *Tongues of Fire: The Explosion of Protestantism in Latin America*. Blackwell, Oxford-Cambridge, UK

Mecham J L 1966 *Church and State in Latin America*. University of North Carolina Press, Chapel Hill, NC

Prien H-J 1978 *Die Geschichte des Christentums in Lateinamerika*. Vandenhoeck & Ruprecht, Göttingen

Stoll D 1990 *Is Latin America turning Protestant? The Politics of Evangelical Growth*. University of California Press, Berkeley

G. Benavides

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Christianity: Liberal

'Liberal' is the designation given to a broad trajectory in Christianity by sympathisers and critics alike. In some of its manifestations it is also referred to as 'modernism.' Both terms draw attention to a defining characteristic: a dissatisfaction with earlier forms of religion and a concern to replace them with less restrictive alternatives more open to the spirit of the age. Whilst it has rarely led to the establishment of new churches, liberalism has had a significant impact within all the mainline Christian denominations in the West.

Understood in this broad sense, liberal Christianity can be seen to date back even before the Renaissance to the fifteenth century, when the term 'modern' was first invoked in relation to reforming movements in theology and spirituality like the '*Via Moderna*' and the '*Devotio Moderna*.' Since then Western Christianity has been challenged repeatedly by individuals and movements who have championed the freedom of the individual Christian over against the authority of institutional religion. Yet it is in the modern period that liberal Christianity has become most prominent and has played its most important role in shaping modernity itself.

By the end of the twentieth century something of a consensus had developed amongst sociologists of religion that liberal Christianity was in inexorable decline. This consensus coincided with a period when theological liberalism had fallen out of fashion. It will be challenged in what follows on three grounds: (a) it overlooks the internal variety of liberal Christianity, (b) it ignores evidence of its continuing vitality on the ground, and (c) it fails to recognize that liberal Christianity remains well adapted to many of the socio-economic formations and cultural trends of the modern world.

1. *Varieties of Liberalism*

1.1 *Rationalist*

One of the most important ways in which liberal Christianity is implicated with the rise of modernity is through its central role in the Enlightenment of the mid-seventeenth to mid-eighteenth centuries. In their different ways such figures as the Deists, Thomas Jefferson, Tom Paine, René Descartes, John Locke, and Immanuel Kant were all concerned with the liberal reform of religion. They helped shape a liberal Christianity which was characterized by (a) hostility to 'traditional' religion, conceived as superstitious, heteronomous, and divisive, (b) confidence in human reason and the primacy of the sovereign individual, (c) activist optimism about the possibilities of human and social improvement, (d) belief in the harmonious unity of all true ('natural') religion, and (e) high valuation of freedom. Rationalist liberalism tended to be primarily an intellectual movement, though it was often related closely to political radicalism. At the institutional level it gave rise to the Unitarian and Universalist churches.

1.2 *Romantic*

Whilst many sociologists identify liberal Christianity with its rationalist forms, there are other important varieties. By the late eighteenth century, for example, a Romantic liberalism had begun to exercise an important cultural and religious influence. In America this was best represented by Transcendentalism, and in Germany by the theology of Friedrich Schleiermacher. Romantic liberalism inherited rationalist liberalism's belief that the individual rather than the institution was the locus of true religion ('my mind is my church,' as Tom Paine had put it), but stressed the authority of feeling, imagination, experience, and self-consciousness rather than reason. This stress on the importance of individual experience of God has been carried through into the twentieth century in the work of theologians like Rudolph Bultmann and Paul Tillich.

1.3 *Ethical and Social*

Kant had located religious authority in moral reason. This 'ethicisation' of the Christian religion was taken further in Germany by theologians like Albrecht Ritschl, Wilhelm Hermann, and Adolf von Harnack, and in America by theologians like Horace Bushnell and Walter Rauschenbusch. The latter, the leader of the Social Gospel movement, was responsible for interpreting Christianity as a force for social reform in industrial society.

1.4 *Liberation and Feminist*

In the twentieth century the influence of liberal Christianity continues to be felt in new reform movements within the churches and theology, most notably in some important varieties of Liberation theology (a product of Latin America) and Feminist Theology (a product of North America and Western Europe). In both we find a characteristic emphasis on human liberation and the authority of experience, combined with a thoroughgoing criticism of existing structures of power and domination in both the churches and wider society, and a bias towards the oppressed and marginal.

2. The Development of Modern Liberalism

Liberalism has had the greatest impact within mainline Protestant churches, and has been particularly influential in countries where Protestantism has been dominant. Most commentators consider its heyday to have been from the mid-nineteenth century through the 1920s. Liberalism flourished because it was able to meet the new challenges posed by the forces of modernisation. Its critical stance towards Christian tradition enabled it to assimilate the rise of historical method and its application to the Bible. Equally, its long-held belief in progress and the capacity of human reason to fathom the mysteries of God and the world helped it embrace the discoveries of modern science, including evolution. What is more, its libertarianism and individualism enabled it to support the interests of the new middle classes and to play a legitimating role in relation to democracy and the modern state. Far from being threatened by rapid modernisation in the nineteenth century, liberal Christianity tended to view itself as the religious and moral engine of social progress. Nowhere was this truer than in the USA. In Catholic countries the picture was very different. There, led by an increasingly defensive Rome, the church explicitly repudiated the errors of 'modernism' (which included political and religious liberalism), and allied itself with the forces of reaction. In 1907 Pope Pius X condemned those Catholic 'modernists' who had embraced aspects of theological liberalism, and introduced an anti-Modernist oath for the clergy which effectively brought to an end all attempts to develop a liberal Catholicism. It was not until the Second Vatican Council of 1962–5 that a moderate liberalism gained official sanction. Since then there has been an explosion of liberal thought in the Catholic church, despite Pope John Paul II's attempts to curb its influence. The work of theologians like Hans Küng, as well as the rise of Liberation Theology and Feminist Theology has been particularly notable.

The twentieth century has brought unprecedented challenges for Liberal Christianity. These include the rise of conservative evangelical and fundamentalist Christianity and, at the theological level, the rise of Neo-orthodoxy. At the same time many elements of the liberal creed—including belief in progress, reason, 'humanity,' and the ideal of religious unity—have been shaken both inside and outside the churches. This together with shrinking attendance in the mainline Protestant churches most influenced by liberalism has led many commentators to conclude that it is in terminal decline. What this conclusion overlooks, however, is that the mainline denominations still account for the vast majority of Christians (especially if one includes the Catholic church), and that many within these denominations (including the Catholic churches) continue to embrace some version of liberal Christianity (see below). What is more, liberalism remains extremely well adapted to societies that show no real sign of diminishing their commitment to freedom, equality, democracy, and a liberal individualism reinforced by the institutions of a free-market economy.

3. Liberalism on the Ground

Dean Kelley's book *Why the Conservative Churches are Growing* (1972) captured a new mood in the sociology of religion. Where secularisation theory had previously encouraged the conclusion that the more traditional and antimodern forms of religion would be those which would decline most rapidly, the growth of conservative and sectarian bodies alongside evidence of the decline of more liberal denominations led to a change of mind after the 1960s. As noted below, sociological theory was quickly mobilized to explain this change, and to reinforce the prediction of further liberal decline.

Much of the case for liberal decline focuses on the American example. It is clear from census data in the US, like the General Social Survey (GSS), that the three denominations it classifies as 'liberal' (Presbyterian, Episcopal, and United Church of Christ) did indeed decline after 1970. So too did the three denominations it labels as 'moderate' (United Methodist, Lutheran, and Disciples of Christ), together with the Roman Catholic church. By contrast more conservative denominations like the Southern Baptist Convention and the Seventh-day Adventists grew in numbers. There is some evidence, however, that numbers in the liberal denominations may now be stabilising (Roof and McKinney 1987), and it may therefore be premature to pronounce on the relative success or failure of liberal and conservative Christianity in modern times.

It is also necessary to exercise caution when relying on data relating to levels of denominational attendance and affiliation alone in assessing the state of liberal Christianity in the churches. The main problem with this method is that it is blind to complex patterns of belief and commitment within churches and denominations. For example, most mainline churches today—both Catholic and Protestant—contain both liberals and conservatives within their congregations, and it is increasingly hard to label whole denominations 'liberal' or 'conservative.' What is more, there is some important evidence that even more conservative forms of Catholic and evangelical Christianity are increasingly permeable to the influence of liberalism. Thus James Davison Hunter's extensive research among college-age Evangelicals in American uncovered evidence of an increasing liberalisation of their belief, values, and practice (Hunter 1987).

The best way to assess the relative strength of liberalism in the contemporary churches would seem to be through intensive and widespread congregational

study. The most important work of this kind is that undertaken by Ammerman (1997) who studied 23 representative congregations across the USA and surveyed almost 2,000 individuals. Her discovery was that respondents fell into three categories: liberal or 'Golden Rule' Christians (51 percent), evangelicals (29 percent), and social activists (19 percent). On the basis of her research Ammerman challenges the assumption that religious liberalism is a spent force, and that liberal religiosity is a paler reflection of conservative Christianity. The Golden Rule Christianity she describes is characterized by an emphasis on the primacy of good deeds motivated by love, care, and compassion, and by a belief in the importance of religious tolerance.

It may be that Ammerman has discovered a fifth variety of Christian liberalism—one which might be labelled 'relational,' and whose significance on the ground in the second part of the twentieth century has been largely overlooked.

4. *Sociological Interpretations*

At least three clusters of explanations have been offered by sociologists of religion to account for the rise of liberal Christianity in modern times. It has been explained as (a) an accommodation or even a capitulation to modernity (Peter Berger), (b) a natural outgrowth of Protestantism and, in particular, of the latter's emphasis on personal subjective conviction (Ernst Troeltsch), and (c) a means by which the clergy, their social status undermined by modernity, have attempted to protest and regain a social role (Jeffrey Hadden). Woodhead and Heelas (2000) have also drawn attention to liberal Christianity's compatibility with modern socio-economic formations and wider cultural trends such as the turn to the self.

Sociologists have also developed theories to account for liberal Christianity's apparent decline. Kelley explained this by drawing a contrast with conservative religion. Where the latter was 'strict' and 'challenging,' the latter was the opposite. As such, he argued, it was unable to generate or sustain commitment, consensus or strong community. Peter Berger offered a more rigorous version of this explanation by arguing that plausibility is a function of unanimity. The strong, unified communities that characterize conservative religion are better able to sustain plausibility than are the more diffuse and less disciplined communities of liberalism. Meanwhile sociologists like Talcott Parsons and David Martin have also argued that liberalism is a victim of its own success: because its beliefs and values are so close to those of the wider culture it is no longer able to sustain a distinctive identity nor to hold or attract adherents.

Much of this theoretical work depends on a contrast drawn between liberal and conservative Christianity. It is also important to note another boundary: that between liberal Christianity and radical or 'alternative' forms of spirituality. For whilst the fate of liberal Christianity is bound up with that of conservative Christianity, it is also bound up with that of new forms of religiosity like the New Age. In some ways the latter seems to represent an intensification of key liberal themes like individualism and freedom, but without liberal Christianity's continuing commitment to some form of institutional church. It remains to be seen whether the apparent growth of such religiosity will in the end have the effect of strengthening or of weakening liberalism.

See also: American Studies: Religion; Christian Liturgy; Civil Religion; Feminist Theology; Liberalism; Liberalism: Historical Aspects; Protestantism and Gender; Rationalism; Rationality in Society; Reformation and Confessionalization; Religion and Politics: United States; Religion: Evolution and Development; Religion, Sociology of; Religiosity: Modern

Bibliography

Ammerman N T 1997 *Congregation and Community*. Rutgers University Press, New Brunswick, NJ

Hunter J D 1987 *Evangelicalism. The Coming Generation*. University of Chicago Press, Chicago

Hutchison W R (ed.) 1968 *American Protestant Thought in the Liberal Era*. University Press of America, Lanham, MD

Hutchison W R 1976 *The Modernist Impulse in American Protestantism*. Harvard University Press, Cambridge, MA

Kelley D M 1972 *Why Conservative Churches are Growing*. Harper & Row, New York

Michaelsen R S, Roof W C (eds.) 1986 *Liberal Protestantism: Realities and Possibilities*. Pilgrim Press, New York

Miller D E 1981 *The Case for Liberal Christianity*, 1st edn. Harper & Row, San Francisco

Reardon B M (ed.) 1968 *Liberal Protestantism*. Stanford University Press, Stanford, CA

Roof W C 1978 *Community and Commitment. Religious Plausibility in a Liberal Protestant Church*. Elsevier, New York

Roof W C, McKinney W 1987 *American Mainline Religion. Its Changing Shape and Future*. Rutgers University Press, New Brunswick, NJ

Woodhead L J P, Heelas P L 2000 *Religion in Modern Times*. Blackwell, Malden, MA

L. Woodhead

Christianity Origins: Primitive and 'Western' History

This article considers Christianity in the first 300 years of its existence, before it achieved a close alliance with the Roman state. It pays particular attention to the social forms of early Christianity and their relation to wider society.

1. *The Jesus Movement*

The movement that centered around Jesus of Nazareth in his own lifetime seems an unlikely candidate for the eventual transformation of Western society. It appears to have been one of many such movements in early first century Palestine led by a charismatic Jewish teacher appealing to a poor and mainly rural audience (Theissen 1978). Jesus's message, which is only comprehensible within a framework of contemporary Jewish beliefs and expectations, centered round the proclamation of the imminent reign of God. From this expectation arose the urgent and sovereign demands to repent and believe in the gospel. All other concerns were secondary—including on occasion those of Jewish law and custom. Yet this message was presented as 'good news': what Jesus offered his followers was the most intimate relationship with a God not of wrath and judgment but of love and mercy—a 'father' who cares with loving tenderness for each one of his children. This God places no barriers on relationship with him—all are welcomed into his kingdom irrespective of their moral, social, or religious status.

2. *The Pauline Revolution*

Whilst Jesus' mission was primarily to the Jews, the unrestricted address of his message gave it a universalist momentum which would make possible the later spread of Christianity beyond Israel. As a universal religion, Christianity provided an alternative to the national or civic religions, both Roman and Jewish, of the time. It appears to have been the apostle Paul, whose letters are preserved in the New Testament, who played the decisive role in drawing out these universalist implications and giving theological justification to a 'mission to the gentiles.'

A Jew as well as a Roman citizen, Paul was converted by a vision of the risen Jesus. The faith whose spokesman he subsequently became was centered on this risen, cosmic Christ rather than on the historical Jesus of Nazareth. There were important sociological implications in this shift. As Troeltsch (1931) noted, Jesus's original message was 'individualistic' in the sense that it was focused on intimate relation between the individual and God. Whilst its universalist and egalitarian message fostered a broad sense of community between all those called to love God and neighbor, it neither fostered new communities nor made any attempt to influence wider society. It remained a reforming faction within Judaism (Sanders 1985, Elliott 1995).

Paul's reinterpretation of Christianity altered these dynamics of the early Jesus movement very significantly. As Schweitzer (1931) argued, Paul developed a 'Christ-mysticism' in which the believer is incorporated through faith into the 'body of Christ.' Though this mysticism also has an individualist emphasis, incorporation into the body of Christ is corporate—those who have faith are united not only with Christ but with one another. This corporate Christ-mysticism undergirded the development from the mid-first century of a church (ekklesia), which took the form of local communities linked together in a 'catholic' (universal) alliance by their common possession of the Spirit of Christ.

3. *The Emergence of Catholic Christianity*

Until at least the second century these early Christian communities were charismatic communities (communities of the Spirit), in which social authority was not institutionalized, but conferred by the Spirit. There seems to have been no clear hierarchy of authority, with different functions (such as apostle, teacher, prophet, and miracle-worker) being regarded as mutually constitutive of the body of Christ. In a development that Troeltsch (1931) categorizes as the emergence of 'catholic' and 'sacramental' Christianity, however, the spirit gradually became institutionalized in the sacraments, particularly those of baptism and the eucharist. Here the presence of Spirit is, as it were, guaranteed. The sacraments are material signs of the freely given grace of God and the efficacious tokens of salvation.

The development of a sacramental Christianity allowed for the development of stable and enduring communities not based on the unpredictable outpourings of the spirit. It went hand in hand with the emergence of a clergy whose authority was bound up with their exclusive authorisation to handle and distribute the sacraments. Their status was not based on personal charisma, superior religious achievement, or inheritance. Rather, they were the authorised representatives of the wider Christian community. Early documents defending a sacramental priesthood reveal that this development was not uncontroversial. On the one hand it made possible catholicity and order. On the other it led to exclusions, most notably the exclusion of women from positions of authority in the church.

The development of catholic Christianity also involved the definition and maintenance of uniformity in belief and liturgical practice. This achievement was also a difficult and remarkable one given that 'early Christianity' was never as unified as that title implies. Despite the idealized backward glance of a later era (such as that of the fourth-century church historian, Eusebius), Christianity came into being as a diverse set of largely autonomous communities spread around the Mediterranean basin and in Syria and Asia Minor. Many were centered around a particular apostle and a particular gospel (whether in oral or written form), and developed distinctive forms of belief and practice.

If we compare the four gospels contained in the New Testament (probably the products of such communities), we get some idea of the range of beliefs they held and of their very different understandings of Jesus.

In the face of this diversity, the establishment of an authorised scriptural, creedal, and rhetorical tradition was as important as that of a universal sacramental priesthood (Cameron 1991). By the second century we find early representatives of catholic Christianity listing the documents which should be treated by Christians as authoritative, and which would eventually come to form the New Testament. These were then bound up with the first authoritative Christian scripture, the Jewish Bible or 'Old Testament'. The formation of this scriptural 'canon' went hand in hand with the development of a 'canon of faith.' Both were later debated and defined by the councils and creeds which would become such a distinctive feature of Christianity. Together authorized scripture and doctrine came to define the boundaries of 'orthodoxy.' Again, this process involved exclusions, including that of a large number of gospels, lives, and acts of Jesus, the apostles and saints which are now classified as 'apocryphal,' together with a large body of philosophical–theological literature influenced by Christian, Jewish, Platonic, and Persian sources, which is often classified together as 'gnostic.'

Against the spiritualizing tendencies of the gnostics (a tendency which took further the Pauline spiritualization of Christ), the emerging catholic church developed an emphasis which may be characterised as 'materialist.' The authority of the clergy, e.g., was said to rest on an 'apostolic succession' which consisted of a historical and physical continuity established through the 'laying on of hands' by Christ and the apostles down to the present generation. Likewise, the church was the visible community of men and women gathered together to receive these sacraments rather than an invisible body of the elect, and the authorized means of salvation were the visible and tangible sacraments. In many cases too Christian hope continued to be focused on a physical resurrection, rather than on the release of an immaterial soul from the body. A more hostile attitude to the body and material life would, however, become a feature of some of the asceticism and monasticism that developed within Christian circles from the end of the third century onwards.

Despite this materialist emphasis, however, early Christianity was not involved in any direct attempt to reform the society within which it found itself. Jesus had directed his followers' energies to 'the one thing needful'—love of God and neighbor—rather than to social reform, and this emphasis continued in early catholic Christianity. To the extent that Jesus commanded his followers to love all, including the Roman soldier and the tax collector, it could even be argued that the Jesus movement had a broader social reach than the Pauline and post-Pauline communities whose energies were focused on love of 'the brethren.' Their duty was to 'build up the body of Christ' rather than to change 'the world'—the latter being a category which derived from this mentality.

The result, as Troeltsch (1931) argued, was that the early Christian communities did not develop a 'social teaching.' The Christian response to social problems such as poverty was to advocate individual acts of charity rather than social reform. Ownership of property was neither abolished nor condemned, but possessions were to be used to help the Christian community. Similarly, in relation to class and social position, the early church initiated a revolution within its own walls—slave and free, male, and female were equal before Christ and in relation to salvation—which left wider patterns of social inequality (including slavery and the position of women) virtually untouched. The state, even when persecuting Christians, was regarded by most early Christians as the wielder of a proper and God-given authority that should call forth respect and obedience rather than attempts at reform.

4. *Alliance of Church and State*

Despite its failure to develop a social teaching, it is clear that early Christianity had a significant impact on its wider social context. It appears to have initiated an inner revolution within the Roman Empire whose effect was felt in a number of ways—not least through the new educational and welfare opportunities it offered, and through the model of an inclusive society which it provided. Whilst it is impossible to reconstruct the nature and extent of the growth of Christianity in the first three centuries of its existence, it is estimated that by the beginning of the fourth century it may have accounted for up to 10 percent of the population of the Empire. Its success appears to have been due to its ability to form 'a compact, even massive, constellation of commitments' (Brown 1997). Morality, philosophy, and ritual, which formally had formed separate spheres of activity in the 'pagan' world, were brought together by the church, and fused into a universal religion.

It was these new characteristics and potencies which eventually enabled Christianity to serve as a unifying and legitimating force for an empire which had once persecuted it. Constantine formalized the process whereby church and state grew into alliance with one another after AD 312, and church leaders rapidly exploited the new opportunities that this opened. In this way a decisive alteration in Christianity's relation to the social order took place, one which would have the most far-reaching consequences not only for the evolution of the church, but for the social and political ordering of Christianity's territories in the East as well as in what—under Christian influence—would eventually become Western Europe.

See also: Classical Archaeology; Historiography and Historical Thought: Christian Tradition; Historiography and Historical Thought: Islamic Tradition; Judaism; Near Middle East/North African Studies: Religion

Bibliography

Brown P 1997 *The Rise of Western Christendom. Triumph and Diversity AD 200–1000*. Blackwell, Malden, MA and Oxford, UK

Cameron A 1991 *Christianity and the Rhetoric of Empire. The Development of Christian Discourse*. Berkeley, CA, Los Angeles, Oxford, UK

Elliott J H 1995 The Jewish Messianic movement: from faction to sect. In: Esler P F (ed.) *Modelling Early Christianity: Social-Scientific Studies of the New Testament in its Context*. Routledge, London and New York

Hazlett I 1991 ed. *Early Christianity. Origins and Evolution to AD 600*. SPCK, London

Meeks W A 1983 *The First Urban Christians: The Social World of the Apostle Paul*. Yale University Press, New Haven, CT

Sanders E P 1985 *Jesus and Judaism*. Fortress Press, Philadelphia

Schweitzer A 1931 *The Mysticism of Paul the Apostle*. A & C Black, London

Theissen G 1978 *Sociology of Early Palestinian Christianity*. 1st American edn. Fortress Press, Philadelphia

Troeltsch E 1931 *The Social Teachings of the Christian Churches* (trans. Wyon O). George Allen and Unwin, London, MacMillan, New York, Vol. 1

L. Woodhead

Chronic Illness, Psychosocial Coping with

1. Background

Improvements in health care technologies and treatments have resulted in increased life expectancies and improved disease management for individuals with chronic illnesses. To a great degree, quality of life for many individuals with these illnesses may be determined by the ways they deal with the illness. Thus, identifying effective and ineffective ways of coping with these diseases may lead to the development of more efficacious interventions for these individuals.

Since 1980 there has been a substantial amount of research devoted to understanding the relation between coping with chronic illnesses and psychological adaptation. Although there have been some consistent findings regarding coping and its impact on psychological outcomes, particularly in the area of coping with pain, the enthusiasm for the empirical study of coping in general has dampened significantly over the course of the past several years. Indeed, recent reviews of coping research have harshly criticized the literature, particularly assessment methodologies (see Coyne and Racioppo 2000). Thus, much of the initial promise for coping research to enhance clinical practice has not been realized.

2. Historical Perspective and Current Concepts of Coping

The psychological study of coping dates back to Sigmund Freud (1896/1966), who put forth the concept of defense mechanisms, defined as mental operations that kept painful thoughts and feelings out of awareness. The next major shift in the study of coping was brought about as a result of cognitive theories. The focus on intrapsychic processes that intervene between events and responses to events increased with the introduction of other cognitive theories such as Beck (1976). According to cognitive theories, cognitive coping mediated between stressful events and psychological and physical responses to stressful events. It was hypothesized that, by examining individual coping differences, a greater understanding of why people react differently to the same events would be achieved.

Research on stress and coping exploded with the work of Lazarus and Folkman (1984), who put forth the transactional stress and coping paradigm. According to Lazarus, coping refers to cognitive and behavioral efforts to manage disruptive events that tax the person's ability to adjust (Lazarus 1981, p. 2). Chronic illness can pose a number of life stressors including loss of physical and social functioning, alterations in body image, managing difficult and complex medical regimens, and chronic pain. According to Lazarus and Folkman (1984) coping responses are a dynamic series of transactions between the individual and the environment, the purpose of which is to regulate internal states and/or alter person-environment relations. The theory postulates that stressful emotions and coping are due to cognitions associated with the way a person appraises or perceives his or her relationship with the environment. There are several components of the coping process. First, appraisals of the harm or loss posed by the stressor (Lazarus 1981) are thought to be important determinants of coping. Second, appraisal of the degree of controllability of the stressor is a determinant of coping strategies selected. A third component is the person's evaluation of the outcome of their coping efforts and their expectations for future success in coping with the stressor. These evaluative judgements will lead to changes in the types of coping employed, as

well as play a role in determining psychological adaptation. Two main dimensions of coping are proposed, problem-focused and emotion-focused coping. Problem-focused coping is efforts aimed at altering the problematic situation. These coping efforts include information seeking and problem solving. Emotion-focused coping are efforts aimed at managing emotional responses to stressors. Such coping efforts include cognitive reappraisal of the stressor and minimizing the problem.

How the elements of coping unfold over time is a key theoretical issue involved in studies of coping processes. Despite the fact that the theory is dynamic in nature, most of the research utilizing the stress and coping paradigm put forth by Lazarus (1981) has relied on retrospective assessments of coping and has been cross-sectional. However, a team of researchers, including Affleck, Tennen, and Keefe (e.g., Keefe et al. 1997) has utilized a daily diary approach to assessing coping with pain, a methodology that can examine the proposed dynamic nature of coping.

Towards the end of the twentieth century there has been also been an expansion in theoretical perspectives on cognitive coping. The literature on cognitive processing of traumatic life events has provided a new direction for coping research and broadened theoretical perspectives on cognitive methods of coping with chronic illness. According to cognitive processing theory, traumatic events can challenge people's core assumptions about them and their world (Janoff-Bulman 1992). The unpredictable nature of many chronic illnesses, as well as the numerous social and occupational losses, causes many individuals to question beliefs they hold about themselves. For example, the diagnosis of chronic obstructive pulmonary disease (COPD) can challenge a person's core beliefs about personal invulnerability. To the extent that a chronic illness challenges core beliefs, integrating the illness experience into their pre-existing beliefs should promote psychological adjustment. Cognitive processing has been used as the phrase to define cognitive activities that help people view undesirable events in personally meaningful ways and find ways of understanding the negative aspects of the experience, and ultimately reach a state of acceptance. Attempts to find meaning or benefit in a negative experience are ways patients may be able to accept the losses they experience. Focusing on the positive implications of the illness or finding personal significance of a situation are two ways of finding meaning in the illness. When considering meaning-making coping, one must distinguish coping activities that help individuals to find redeeming features in an event from the successful outcome of these attempts. For example, people who have a serious illness may report that as a result they have found a new appreciation for life or that they place greater value on relationships. Patients may also develop an explanation for the illness that is more benign (e.g., attributing it to God's will). While cognitive processing theory constructs have been applied to adjustment to losses such as bereavement (e.g., Davis et al. 1998), these constructs have received relatively little attention from researchers examining coping with chronic illness.

Another coping process that falls under the rubric of cognitive coping is social comparison. Social comparison (SC) is a common cognitive process whereby individuals compare themselves to others in order to obtain information about them (Gibbons and Gerrard 1991). According to SC theory, health problems increase uncertainty; uncertainty increases the desire for information and creates the need for comparison. Studies of coping with chronic illness have included social comparison as a focus. A certain type of SC, downward comparison, has been the focus of empirical study among patients with chronic illnesses such as rheumatoid arthritis (RA) (Tennen and Affleck 1997). Wills (1981) has suggested that people experiencing a loss can experience an improvement in mood if they learn about others who are worse off. Indeed, there is evidence to suggest that SC increases as a result of experiencing health problems (Kulik and Mahler 1997). One proposed mechanism for SC is that downward comparison impacts cognitive appraisal by reducing perceived threat. When another person's situation appears significantly worse, then the appraisal of one's own illness may be reduced (Aspinwall and Taylor 1993).

3. *Assessment of Coping*

3.1 *General Coping Checklists*

Folkman and Lazarus' Ways of Coping Checklist (WOC, Folkman and Lazarus 1980) has been one of the most widely used instruments to assess coping efforts. This instrument contains two major subscales, problem-focused and emotion-focused coping, as well as a number of subscales including wishful thinking, cognitive restructuring, information seeking, seeking support, self-blame, and minimization. Instructions typically ask the individual to rate how he or she manages the stressor (Manne and Zautra 1989).

Another measure that has been used is the Coping Strategies Inventory (CSI, Tobin et al. 1989). The CSI distinguishes two dimensions of coping, engagement/disengagement strategies and focusing on the problem/focusing on emotions about the stressor. Problem-focused engagement is composed of problem-solving and cognitive restructuring; problem-focused disengagement is composed of problem avoidance and wishful thinking. Emotion-focused engagement is composed of social support and expressed emotion; emotion-focused disengagement is composed of social withdrawal and self-criticism.

Measuring meaning-making coping and other methods of cognitive processing has been done utilizing existing measures. Some aspects of meaning-making coping can be assessed using the cognitive reappraisal subscales of the COPE (Carver et al. 1989) and the Ways of Coping Checklist (Lazarus and Folkman 1984). Other means of measuring the process of meaning-making involve using measures of cognitive processing. For example, the Impact of Events scale (Horowitz et al. 1979) measures attempts to integrate a traumatic event with current schemas. Other studies have utilized questions tailored specifically for their population.

3.2 Illness-specific Checklists

The majority of illness-specific coping instruments have been designed to assess coping with pain associated with chronic illnesses such as rheumatoid arthritis (RA) and osteoarthritis (OA). Two instruments, the Vanderbilt Pain Management Inventory (VPMI) and the Coping Strategies Questionnaire (CSQ) have been the most widely used instruments. Both measures assess the degree to which patients employ a variety of cognitive and behavioral mechanisms to reduce the impact of painful episodes. Brown and Nicassio (1987) developed the VPMI to assess cognitive and behavioral pain-coping strategies. The 18-item VPMI has two subdimensions, active and passive pain coping. The CSQ comprises seven subscales measuring distinct coping strategies. Factor analyses of the CSQ in both RA and OA samples provide evidence for a two-factor solution, Coping Attempts and Pain Control and Rational Thinking (Keefe et al. 1987).

The Coping with Rheumatic Stressors (CORS, Lankveld et al. 1994) was specifically designed to measure stressor-specific coping in RA. This measure is unique in that it measures coping separately with three stressors, pain, limitations, and dependence. The three coping with pain scales are comforting cognitions, decreasing activity, and diverting attention. The three coping with limitation scales are optimism, pacing, and creative solution seeking. The two coping with dependence scales are making an effort to accept dependence and showing consideration.

3.3 Daily Diary Instruments

Only one instrument, the Daily Coping Inventory (Stone and Neal 1984), has been developed to assess daily coping. This inventory has been adapted for chronic pain coping by Affleck et al. (1992). Patients are asked whether or not they utilize each of seven categories of coping: (a) pain reduction attempt; (b) relaxation; (c) distraction; (d) redefinition; (e) vent emotions; (f) seek spiritual comfort; and (g) seek emotional support. These coping categories have been reduced using factor analyses to two factors, labeled emotion-focused and problem-focused coping (Affleck et al. 1999).

4. Studies Using the Stress and Coping Paradigm

4.1 Cross-sectional Studies

Early studies of coping using the stress and coping paradigm were cross-sectional and utilized retrospective checklists such as the WOC. The earliest studies divided coping into the general categories of problem- and emotion-focused strategies, and focused mostly on psychological outcomes, rather than pain and functional status outcomes.

Later studies have investigated specific types of coping. For example, Felton et al. (1984) examined two types of coping, wish-fulfilling fantasy, and information seeking, using a revision of the Ways of Coping Checklist. Wish-fulfilling fantasy was a more consistent predictor of psychological adjustment than information seeking. While information seeking was associated with higher levels of positive affect, its effects on negative affect were modest, accounting for only 4 percent of the variance. In a second study, Felton and Revenson (1984) examined coping of patients with arthritis, cancer, diabetes, and hypertension. Wish-fulfilling fantasy, emotional expression, and self-blame were associated with poorer adjustment, while threat minimization was associated with better adjustment. Scharloo et al. (1998) conducted a cross-sectional study of individuals with COPD, RA, or psoriasis. Unlike the majority of studies, illness-related variables such as time since diagnosis and the severity of the patient's medical condition were entered first into the equation predicting role and social functioning. Overall, coping was not strongly related to social and role functioning. Among patients with COPD, passive coping predicted poorer physical functioning. Among patients with RA, higher levels of passive coping predicted poorer social functioning.

Very few studies have examined coping with other chronic illnesses. Several studies have investigated the association between coping and distress among individuals with MS. Pakenham et al. (1997) categorized coping as either emotion– or problem-focused, and found that emotion-focused coping was related to poorer adjustment, while problem-focused coping was associated with better adjustment. In contrast, Wineman and Durand (1994) found that emotion- and problem-focused coping were unrelated to distress. Mohr et al. (1997) found that problem solving and cognitive reframing strategies are associated with lower levels of depression, whereas avoidant strategies are associated with higher levels of depression.

As previously noted, most studies have used instructions that ask participants how they coped with the illness in general, rather than asking participants how they coped with specific stressors associated with their illness. Van Lankveld et al. (1994) assessed how RA patients cope with the most important stressors associated with RA (pain, functional limitation, and dependence). When coping with pain was considered, patients with similar degrees of pain who used comforting cognitions and diverted their attention from the pain reported higher well-being. Limiting one's activity was associated with lower well-being. When coping with functional limitation was examined, patients who used pacing of their activity reported lower levels of well-being, and use of optimism was associated with higher well-being after functional capacity was controlled for in the equation. Finally, when coping with dependence was examined, only showing consideration was associated with higher well-being after functional capacity was controlled for in the equation.

4.2 *Longitudinal Studies*

Unfortunately, there have been relatively few studies that have employed longitudinal designs. Overall, passive coping strategies such as avoidance, wishful thinking, and withdrawal, as well as self-blame, have been shown to be associated with poorer psychological adjustment (e.g., Scharloo et al. 1999), and problem-focused coping efforts such as information seeking have been found to be associated with better adjustment (e.g., Pakenham 1999).

5. *Studies of Coping with Chronic Pain*

The majority of these studies have utilized longitudinal designs. For example, Brown and Nicassio (1987) studied pain-coping strategies among RA patients and found that patients who engaged in more passive coping when experiencing more pain became more depressed six months later than patients who engaged in these strategies less frequently. Keefe et al. (1989) conducted a six-month longitudinal study of the relationship between catastrophizing and depression in RA patients. Those patients who reported high levels of catastrophizing had greater pain, disability, and depression six months later. Other investigators (Parker et al. 1989) have reported similar findings. Overall, studies have suggested that self-blame, wishful thinking, praying, catastrophizing, and restricting activities are associated with more distress, while information seeking, cognitive restructuring, and active planning are associated with less distress.

As previously mentioned, several recent studies have employed prospective daily study designs in which participants complete a 30-day diary for reporting each day's pain, mood, and pain-coping strategies using the Daily Coping Inventory (Stone and Neale 1984). These studies, which have been conducted with RA and OA patients, have shown that emotion-focused strategies, such as attempting to redefine pain to make it more bearable, and expressing distressing emotions about the pain, predict increases in negative mood the day after the diary report. The daily design is a promising new method of evaluating the link between coping strategies and mood. More importantly, these studies can elucidate coping processes over time. For example, Tennen et al. (2000) found that the two functions of coping, problem- and emotion-focused, evolve in response to the outcome of the coping efforts. An increase in pain from one day to the next increased the likelihood that emotion-focused coping would follow problem-focused coping. It appeared that, when efforts to directly influence pain were not successful, participants tried to alter their cognitions and adjust rather than influence the pain.

6. *Challenges to the Study of Coping with Chronic Illness*

Recently, the general literature on coping has received a great deal of criticism from researchers (e.g., Coyne and Racioppo 2000). The main concern voiced in reviews regards the gap between the elegant, process-oriented stress and coping theory and the cross-sectional, retrospective methodologies that have been used to evaluate the theory. Although the theory postulates causal relations among stress, coping, and adaptation, the correlational nature of most empirical work has been unsuitable to test causal relations. In addition, retrospective methods require people to recall how they coped with an experience, and thus are likely to be influenced by both systematic and non-systematic sources of recall error. Coping efforts as well as psychological outcomes such as distress are best measured close to when they occur. Recent studies have used an approach that addresses these concerns. These studies have employed a microanalytic, process-oriented approach using daily diary assessments (e.g., Affleck et al. 1999). These time-intensive study designs allow for the tracking of changes in coping and distress close to their real-time occurrence and moments of change, are less subject to recall error, and capture coping processes as they unfold over time. The daily assessment approach also can evaluate how coping changes as the individual learns more about what coping responses are effective in reducing distress and/or altering the stressor. These advances may help investigators to more fully examine whether the methods used to cope with stressors encountered in

the day-to-day experience of living with a chronic disease predict long-term adaptation. Unfortunately, this approach has only been utilized among individuals with arthritis and has not been applied to individuals dealing with other chronic illnesses.

Another key problem with coping checklists that has been noted in a number of reviews of the coping with chronic illness literature is the instructional format. The typical instructions used (e.g., 'How do you cope with RA?') are so general that it is not clear what aspect of the stressor the participant is referring to when answering questions. Thus, the source of the stress may differ across study participants. There are problems even when the participant is allowed to define the stressor prior to rating the coping strategies used. The self-defined stressor may differ across participants, and thus the analyses will be conducted with different stressors being rated.

A third assessment problem regards the definition of coping. While Lazarus and Folkman (1984) regard only effortful, conscious strategies as coping, other investigators have argued that 'automatic' coping methods also fall under the definition of coping (Wills 1997). Indeed, some coping responses may not be perceived by the individual as choices, but rather automatic responses to stressful events. For example, wishful thinking or other types of avoidant types of coping such as sleeping or alcohol use may be categorized by researchers as a coping strategy, but not categorized as such by the individual completing the questionnaire because the individual did not engage in this as an effortful coping strategy. A related and interesting issue regards the categorization of unconscious defense mechanisms. Cramer (2000), in a recent review of defense mechanisms, distinguishes between defenses that are not conscious and unintentional and coping processes that are conscious and intentional. However, there has been an interest in repressive coping, suggesting that some researchers regard defensive strategies such as denial and repression under the rubric of coping. More clarity and consistency between investigators in the definition of coping, particularly when unintentional strategies are being evaluated, would provide more clarity for research.

A fourth assessment issue regards the distinction between 'problem-focused' and 'emotion focused' coping efforts. While researchers may categorize a particular coping strategy as problem-focused coping, the participant's intention may not be to alter the situation, but rather to manage an emotional reaction. For example, people may seek information about an illness as a way of coping with anxiety and to alter their appraisal of a situation, rather than to engineer a change in the situation. The lack of an association between emotion-focused coping and psychological outcomes may, in part, be due to a categorization strategy that does not account for the intention of the coping. Studies utilizing these two categories to distinguish coping dimensions may wish to evaluate coping intention.

There are a number of additional methodological and conceptual challenges that are specifically relevant to studies of coping with chronic illness. First, relatively few studies control for disease severity in statistical analyses. Extreme pain or disability can result in both more coping attempts and more distress. Studies that do not take into account these variables may conclude mistakenly that more coping is associated with more distress. In addition, little attention has been paid to the effects of progressive impairment on the selection of coping strategies, and in the perceived effectiveness of those strategies. Chronic progressive illnesses may be expected to increase feelings of hopelessness. For example, Revenson and Felton (1989) studied changes in coping and adjustment over a six-month period and found that lower acceptance, more wishful thinking, and more negative affect accompanied increases in disability.

Another issue is the lack of longitudinal studies. Clearly, longitudinal studies would help the literature in a number of ways. First, this type of design might help clarify whether coping influences distress or whether coping is merely a symptom of distress, a criticism frequently raised in critiques of coping (e.g., Coyne and Racioppo 2000). Second, longitudinal studies may clarify the role of personality factors in coping. While some investigators suggest that personality factors play a limited role in predicting coping, other investigators argue that coping is a personality process that reflects dispositional differences during stressful events.

Although the lack of progress in the area of coping is frequently attributed to methods of assessment and design, the relatively narrow focus on distress outcomes may also account for some of the problem, particularly when coping with chronic illness is being evaluated. Chronic illness does not ultimately lead to psychological distress for the majority of patients. Indeed, many individuals report psychological growth in the face of chronic illness, and are able to find personal significance in terms of changes in views of themselves, their relationships with others, and a changed philosophy of life (Tennen et al. 1992). While positive affect is included as an adaptational outcome in some studies (e.g., Bendtson and Hornquist 1991), the majority of studies do not include positive outcomes. Positive affect will be a particularly important outcome to evaluate when positive coping processes such as cognitive reappraisal and finding meaning in the experience are examined, as these types of coping may play a stronger role in generating and maintaining positive mood than in lowering negative mood.

Finally, relatively few studies have focused solely on coping and distress and have not taken into account potential moderators such as level of pain, appraisals of controllability, gender, and personality. A careful evaluation of potential moderators will provide both

researchers and clinicians with information about which circumstances particular coping strategies are most effective.

7. Conclusions

As Lazarus points out in his commentary in *American Psychologist*, 'A premise that occurs again and again … is that for quite a few years research has disappointed many who had high hopes it would achieve both fundamental and practical knowledge about the coping process and its adaptational consequences. I am now heartened by positive signs that there is a growing number of sophisticated, resourceful, and vigorous researchers who are dedicated to the study of coping' (Lazarus 2000). It is clear that, despite the multiple methodological problems that this area of research has faced in the past, a heightened awareness of these limitations has led to the application of sophisticated methods that might assist this field in fulfilling the high hopes for this field of research. If investigators in the field of coping with chronic illnesses can adapt daily-diary methods to their populations, focus on specific stressors related to the illness when instructing participants to answer coping questions, include coping appraisals and the perceived efficacy of coping efforts, and carefully delineate illness-related, contextual, and dispositional moderators, the findings may lead to the development of effective interventions for clinicians hoping to improve the quality of life for these individuals.

See also: Chronic Illness: Quality of Life; Chronic Pain: Models and Treatment Approaches; Coping across the Lifespan; Coping Assessment; Coronary Heart Disease (CHD), Coping with; Illness Behavior and Care Seeking; Illness: Dyadic and Collective Coping; Pain, Health Psychology of; Pain, Management of; Rheumatoid Arthritis: Psychosocial Aspects; Social Support and Health; Stress and Coping Theories; Well-being and Health: Proactive Coping

Bibliography

Affleck G, Tennen H, Keefe F J, Lefebre J C, Kashikar-Zuck S, Wright K, Starr K, Caldwell D 1999 Everyday life with osteoarthritis or rheumatoid arthritis: independent effects of disease and gender on daily pain, mood, and coping. *Pain* **83**: 601–9

Affleck G, Urrows S, Tennen H, Higgins P 1992 Daily coping with pain from rheumatoid arthritis-patterns and correlates. *Pain* **51**: 221–9

Aspinwall L G, Taylor S E 1993 Effects of social comparison direction, threat and self-esteem on affect, self-evaluation and expected success. *Journal of Personality and Social Psychology* **64**: 708–22

Beck A T 1976 *Cognitive therapy and the emotional disorders*. International Universities Press, New York

Bendtsen P, Hornquist J O 1994 Rheumatoid arthritis, coping and well-being-cross-comparisons and correlational analyies. *Scandinavian Journal of Social Medicine* **22**: 97–106

Brown G, Nicassio P 1987 Development of a questionnaire for the assessment of active and passive coping strategies in chronic pain patients. *Pain* **3**: 53–84

Carver C S, Scheier M F, Weintraub J K 1989 Assessing coping strategies: A theoretically-based approach. *Journal of Personality and Social Psychology* **56**: 267–83

Coyne J C, Racioppo M W 2000 Never the twain shall meet? Closing the gap between coping research and clinical intervention research. *American Psychologist* **55**: 655–64

Cramer P 2000 Defense mechanisms in psychology. *Psychological Review* **3**: 357–70

Davis C G, Nolen-Hoeksema S, Larson J 1998 Making sense of loss and benefiting from the experience: Two constructs of meaning. *Journal of Personality and Social Psychology* **75**: 561–74

Felton B, Revenson T 1984 Coping with chronic illness: A study of illness controllability and the influence of coping strategies on psychological adjustment. *Journal of Consulting and Clinical Psychology* **52**: 343–53

Felton B J, Revenson T A, Hinrichsen G 1984 Stress and coping in the explanation of psychological adjustment among chronically ill adults. *Social Science and Medicine* **10**: 889–98

Folkman S, Lazarus R S 1980 An analysis of coping in a middle-aged community sample. *Journal of Health and Social Behavior* **21**: 219–39

Freud S 1966 Further remarks on the neuro-psychoses of defense. In: Strachey J (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London (Original work published 1896), Vol. 3, pp. 141–58

Gibbons F X, Gerrard M 1991 Downward comparison and coping with threat. In: Suls J, Wills T A (eds.) *Social Comparison: Contemporary Theory and Research*. Erlbaum, Hillsdale, NJ, pp. 317–46

Horowitz M, Wilner N, Alvarez W 1979 Impact of event scale: A measure of subjective distress. *Psychosomatic Medicine* **41**: 209–18

Janoff-Bulman R 1992 *Shattered Assumptions: Towards a New Psychology of Trauma*. Free Press, New York

Keefe F J, Affleck G, Lefebvre J, Starr K, Caldwell D, Tennen H 1997 Pain coping strategies and coping efficacy in rheumatoid arthritis: a daily process analysis. *Pain* **69**: 35–42

Keefe F J, Brown G K, Wallston K A, Caldwell D S 1989 Coping with rheumatoid arthritis pain: Catastrophizing as a maladaptive strategy. *Pain* **37**: 51–6

Keefe F J, Caldwell D S, Queen K T, Gil K, Martinez S, Crisson J 1987 Pain coping strategies in osteoarthritis patients. *Journal of Consulting and Clinical Psychology* **55**: 208–12

Kulik J, Mahler H 1997 Social comparison, affiliation, and coping with acute medical threats. In: Buunk B P, Gibbons F X (eds.) *Health, Coping and Well-Being: Perspectives from Social Comparison Theory*. Erlbaum, Mahwah, NJ, pp. 227–61

Lazarus R S 1981 The stress and coping paradigm. In: Eisdorfer C, Cohen D, Kleinman A, Maxim P (eds.) *Models for Clinical Psychopathology*. Spectrum, New York, pp. 177–214

Lazarus R S 2000 Toward better research on stress and coping. *American Psychologist* **55**: 665–73

Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York
Manne S L, Zautra A J 1989 Spouse criticism and support: Their association with coping and psychological adjustment among women with rheumatoid arthritis. *Journal of Personality and Social Psychology* **56**: 608–17
Mohr D C, Goodkin D E, Gatto N, Van Der Wende J 1997 Depression, coping, and level of neurological impairment in multiple sclerosis. *Multiple Sclerosis* **3**: 254–8
Pakenham K I 1999 Adjustment to multiple sclerosis: Application of a stress and coping model. *Health Psychology* **18**: 383–92
Pakenham K I, Stewart C A, Rogers A 1997 The role of coping in adjustment to multiple sclerosis-related adaptive demands. *Psycholog Health and Medicine* **2**: 197–211
Parker J, Smarr K, Buescher K, Phillips L R, Frank R, Beck N C 1989 Pain control and rational thinking: Implications for rheumatoid arthritis. *Arthritis and Rheumatism* **32**: 984–90
Revenson T, Felton B 1989 Disability and coping as predictors of psychological adjustment to rheumatoid arthritis. *Journal of Consulting and Clinical Psychology* **57**: 344–8
Scharloo M, Kaptein A A, Weinman J A, Hazes J M W, Breedveld F, Anderson S K, Walker S E, Rooijmans H G M 1999 Predicting functional status in patients with rheumatoid arthritis. *Journal of Rheumatology* **26**: 1686–93
Scharloo M, Kaptein A A, Weinman J A, Hazes J M W, Willems L L N A, Bergman W, Rooijmans H G M 1998 Illness perceptions, coping and functioning in patients with rheumatoid arthritis, chronic obstructive pulmonary disease and psoriasis. *Journal of Psychosomatic Research* **44**: 573–85
Stone A A, Neale J M 1984 A new measure of daily coping-development and preliminary results. *Journal of Personality and Social Psychology* **46**: 892–906
Tennen H, Affleck G 1997 Social comparisons and occupational stress: The identification-contrast model In: Buunk B P, Gibbons F X (eds.) *Health, Coping and Well-Being: Perspectives from Social Comparison Theory*. Erlbaum, Mahwah, NJ, pp. 262–98
Tennen H, Affleck G, Armeli S, Carney M 2000 A daily process approach to coping: Linking theory, research and practice. *American Psychologist* **55**: 626–36
Tennen H, Affleck G, Urrows S, Higgins P, Mendola R 1992 Perceiving control, construing benefits and daily processes in rheumatoid arthritis. *Canadian Journal of Behavioral Science* **24**: 186–203
Tobin D L, Holroyd K A, Reynolds R V, Wigal J K 1989 The hierarchical factor structure of the Coping Strategies Inventory. *Cognitive Therapy and Research* **13**: 343–61
van Lankveld W, van de Bosch P V, van Putte L, Naring G, van der Staak C 1994 Disease-specific stressors in rheumatoid arthritis: Coping and well-being. *British Journal of Rheumatology* **33**: 1067–73
Wills T A 1981 Downward comparison principles in social psychology. *Psychological Bulletin* **90**: 245–71
Wills T 1997 Modes and families of coping. In: Buunk B, Gibbons F X (eds.) *Health, Coping and Well-being: Perspectives from Social Comparison Theory*. Erlbaum, Mahwah, NJ, pp. 167–93
Wineman N M, Durand E J, McCulloch B J 1994 Examination of the factor structure of the ways of coping questionnaire with clinical populations. *Nursing Research* **43**: 266–73

S. Manne

Chronic Illness: Quality of Life

1. *Background*

Increased longevity and the development of sophisticated health care technologies and treatments mean that people today live with chronic health conditions over extended periods of their lives. Quality of life (QOL) has become an important goal of treatment and marker of success in health care interventions in chronic illness generally. In many disorders, e.g., osteoarthritis, health interventions will have little impact on mortality statistics but great potential for reducing disability and increasing QOL.

QOL research first developed in cancer settings where the balance of quality and duration of life became a key concern in decisions to use novel treatments with very serious side effects and only partial efficacy. However, over the past 20 years there has been a burgeoning of research activity in every major chronic illness category. This is evident for different diseases at various structural levels. For instance, the US research funding agency, the National Heart Lung and Blood Institute, require almost all clinical trials and many epidemiological studies they fund to have a QOL component. In rheumatology, there is an international, professionally endorsed cooperative called OMERACT (Outcome Measures in Rheumatoid Arthritis Clinical Trials). They seek to improve QOL outcome measurement through data-driven interactive consensus. In cancer, the European Organisation for Research on the Treatment of Cancer (EORTC) has been established by interested professionals. They have developed a core QOL measure and disease-specific modules for various types of cancer. To the extent that QOL assessment reflects consideration of aspects of these diseases beyond the biomedical, QOL assessment reflects the increasingly biopsychosocial perspective in modern health care.

2. *Uses of Quality of Life Assessment in Chronic Illness*

QOL research in chronic illness has policy, treatment evaluation, descriptive, and individual clinical uses. There is now a vast literature on, and compendia of, instruments for this very wide set of circumstances and health conditions (e.g., Bowling 1997). At a policy level, initiatives to develop QOL assessment are concerned with means to identify those components of the healthcare system which are most effective in achieving good health outcomes. The Medical Outcomes Study is the major example of this approach. As part of a two-year prospective study to identify the association of reimbursement aspects of a health funding system with health outcomes in the US, a health-related QOL (HRQOL) was devised. The most widely used version, the short-form 36-item ques-

tionnaire (SF-36) was developed as a generic HRQOL instrument which could be used with the general population as well as patient groups (Stewart et al. 1989). The SF-36 has eight dimensions: physical functioning; role limitations because of physical health problems; bodily pain; social functioning; general mental health; role limitations because of emotional problems; vitality, energy, or fatigue; and general health perception. Items on each dimension are weighted so that each scale adds to 100 points, with higher scores indicating better health. Thus health profiles across the eight dimensions can be easily compared in graphic form across different disease or treatment groups. For SF-36, individuals rate their own health status. A different approach has been tried at the general population level in the State of Oregon, USA, to assist the public in prioritizing medical treatments for inclusion in publicly funded health care. Here, people were asked to rate the hypothetical impact of various levels of symptoms and restrictions as a way to create a hierarchy of levels of challenge to QOL. However, it has not proven possible to date to develop a system whereby the public are willing to withhold certain treatments from patients on the basis of lesser clinical or QOL benefit.

Much of the work using QOL instruments in chronic illness has been to evaluate the QOL impact of differing types of treatment, e.g., comparing medications. In this formulation, QOL is seen as a dependent variable. Treatment comparisons are usually made on a combination of generic and disease-specific QOL instruments (see *Quality of Life: Assessment in Health Settings*) to cover the range of dimensions in which one intervention may differ from another in impact. In priority terms, QOL is seen on a continuum. It is positioned after efficacy and safety of the treatment and before or after cost-effectiveness, depending on the costs and condition under consideration. One merit of this approach has been to illustrate the sometimes very differing perspectives of professionals and patients. For instance, innovative limb-sparing treatment for cancer was found to result in poorer QOL for patients than the traditional amputation option (Sugarbaker et al. 1982). This finding was subsequently used to refine treatment such that initial difficulties with the limb-sparing approach could be removed and it was then subsequently offered as the better QOL option.

QOL research has provided a large descriptive body of research outlining the experience of various health conditions. This work is of benefit to professionals, patients, and the public understanding the typical HRQOL challenges in different health conditions; e.g., the SF-36 has been used to provide a HRQOL profile of nine common chronic medical conditions (Stewart et al. 1989). From this we can see that cardiac conditions such as myocardial infarction and congestive heart failure have a greater overall impact on HRQOL than do others such as diabetes.

A final use of QOL assessment is to facilitate decision making for the individual patient. While some instruments such as the Dartmouth Primary Care Cooperative Information Project (COOP) charts have been developed with clinical settings and ease of administration and interpretation as a priority, the evidence is that there is little formal QOL assessment in this format at present. This is so despite significant professional support for the broad concept of QOL assessment in clinical practice.

3. *Differing Perspectives on Quality of Life: Needs Versus Wants*

Quality of life has been generically defined as the capacity of an individual to achieve their life plans or the difference, at a particular point in time, between the hopes and expectations of an individual and their present situation. In Calman's (1984) definition, a small difference or 'gap' is commensurate with a good QOL. He proposes that the gap is widened by experiences such as illness and that it can be reduced by improving the present situation or by reducing the person's hopes and expectations to a more 'realistic' level. The World Health Organization (WHO) definition acknowledges possible cultural, as well as individual, differences in the understanding of the concept of QOL (see *Quality of Life: Assessment in Health Settings*). The WHOQOL instrument is based on a set of five domains: physical function; emotional function; social relationships; levels of independence; and environment. While core domains have been demonstrated to be relevant across cultures, the individual facets and items to assess these domains are developed within cultures. Thus while the domain of independence will feature in both British and Indian versions of the instrument, the particular items to best reflect that domain may differ. The WHOQOL instrument also illustrates the multifaceted nature of QOL; working capacity and vitality/fatigue might relate to work life from the physical function domain while working environment and transport would relate to work from the environment domain. QOL has also been defined more narrowly in a health-related context (HRQOL) (see *Quality of Life: Assessment in Health Settings*). HRQOL provides more focus on those specific aspects of life function likely to be influenced by illness although there is debate about how to differentiate health-related and non-health-related aspects of QOL.

Underlying most definitions of QOL, although usually not articulated, is one of two philosophical perspectives—the 'needs' or the 'wants' approach to QOL (Browne et al. 1997). The needs approach views QOL as the extent to which universal human needs are met. Many HRQOL instruments conform to the needs approach: the themes, questions, and question weightings used are standard and established independent of

the person being assessed. The wants approach allows for a definition of QOL using themes, questions, or question weightings determined by the individual being evaluated. The methodologies used to achieve the wants approach are very varied and include content analysis and repertory grid (see *Quality of Life: Assessment in Health Settings* regarding individual QOL methods). These measures typically do not limit the focus of enquiry to health and it is with these assessments (which usually focus more on perception than function) that there are findings of high levels of QOL in groups with major chronic illness as discussed in the next section.

4. Challenges to QOL Assessment in Chronic Illness

One major challenge relates to the comparison of professional and patient perspectives on patient QOL. As reviewed by Sprangers and Aaronson (1992), professionals are poor judges of patient QOL. Moreover, professionals consistently rate patient QOL as poorer than do patients. Where family members are involved in ratings, their views are intermediate. There is little understanding of the processes that facilitate patients to develop or maintain a reasonable QOL in the face of what others define more negatively.

A second complementary challenge arises from comparative QOL assessment where instrument use allows comparison with the general population. Many of these studies find little difference in QOL between 'healthy' individuals and those with serious health conditions; both groups reporting high QOL (Allison et al. 1997). Such study comparisons differ depending on the wants versus needs approach taken to assessing QOL as discussed in the previous section. Interesting work in the cancer area shows how patients with serious cancers, when compared with orthopedic patients and matched healthy controls, focus away from health issues in order to maintain a good level of life satisfaction (Kreitler et al. 1993).

One way to address both of the challenges outlined is to see QOL as having a trait or homeostatic dimension such that individuals 'reset' their evaluative framework in adversity to recreate an appropriate QOL level for themselves. Diener (2000) discusses psychological perspectives on adaptation and goal setting relevant to this issue. The practical challenge when developing this understanding is how to use the knowledge to promote psychological growth in the context of a chronic health condition.

5. Conclusions and Outlook

To date, social and behavioral scientists have made major contributions to QOL assessment in chronic illness on a scale of cooperation with the medical sciences that is probably unique. They have done so in ways that reflect the needs/wants dichotomy of approaches. Many have provided psychometric skills within a largely biomedical framework to address challenges such as increasing responsiveness to change in evaluative research instruments (e.g., McGee et al. 1999) or in developing markers to identify clinical as distinct from statistical significance with HRQOL instruments (e.g., Juniper et al. 1994). The aim of the latter is to be able to use HRQOL scores in much the same manner as one would use blood pressure or body temperature readings to determine what course of action to take and when to do so. Others have been concerned with methods of documenting the salient and dynamic features of QOL for individuals; this has involved both qualitative and quantitative methods from a more phenomenological perspective. The challenge for the coming decades is to find some reconciliation between these two perspectives. In this way QOL assessment can be better used to direct services at individual and societal levels while expanding our understanding of how QOL is developed, maintained and altered as necessary through the course of chronic illness and its treatment.

See also: Chronic Illness, Psychosocial Coping with; Chronic Pain: Models and Treatment Approaches; Coping across the Lifespan; Quality of Life: Assessment in Health Settings; Stress and Coping Theories; Welfare: Philosophical Aspects; Well-being (Subjective), Psychology of

Bibliography

Allison P J, Locker D, Feine J S 1997 Quality of life: a dynamic construct. *Social Science Med.* **45**: 221–30

Bowling A 1997 *Measuring Health: A Review of Quality of Life Scales*, 2nd edn. Open University Press, Birmingham, UK

Browne J P, McGee H M, O'Boyle C A 1997 Conceptual approaches to the assessment of quality of life. *Psychology and Health* **12**: 737–51

Calman K C 1984 Quality of life in cancer patients—an hypothesis. *Journal of Medical Ethics* **10**: 124–7

Diener E 2000 Subjective well-being. The science of happiness and a proposal for a national index. *American Psychologist* **55**: 34–43

Juniper E F, Guyatt G H, Willan A, Griffith L E 1994 Determining a minimal important change in a disease-specific quality of life questionnaire. *Journal of Clinical Epidemiology* **47**: 81–7

Kreitler S, Chaitchuk S, Rapport Y, Kreitler H, Algor R 1993 Life satisfaction and health in cancer patients, orthopaedic patients and healthy individuals. *Social Science Med.* **36**: 547–56

McGee H M, Hevey D, Horgan J H 1999 Psychosocial outcome assessments for use in cardiac rehabilitation service evaluation: a 10-year systematic review. *Social Science and Medicine*. **48**: 1373–93

Sprangers M A G, Aaronson N K 1992 The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *Journal of Clinical Epidemiology* **45**: 743–60

Stewart A L, Greenfield S, Hays R D, Wells R D, Rogers W H, Berry S D, McGlynn E A, Ware J E 1989 Functional status and well-being of patients with chronic conditions: Results from the Medical Outcomes Study. *JAMA—Journal of the American Medical Association* **262**: 907–13

Sugarbaker P H, Barofsky I, Rosenberg S A, Gianola P J 1982 Quality of life assessment of patients in extremity sarcoma clinical trials. *Surgery* **91**: 17–23

H. M. McGee

Chronic Pain: Models and Treatment Approaches

1. Introduction

Pain is one of the most common reasons for seeking medical care (Hart et al. 1995, Gureje et al. 1998). For many, pain is an adaptive response warning of the potential for bodily harm. This signal helps to prevent further injury and tissue damage. However, research as well as common experience suggests that pain is much more than a mere barometer of the amount of tissue damage. Despite the universal experience of pain, it is a difficult phenomenon to define precisely. We have all seen instances where two people report radically different amounts of pain despite an equivalent injury. Clinicians have also noted that soldiers report little or no pain after a combat injury until after they are removed from the battlefield. The wide variations in how people communicate their pain further complicate matters. Some may express pain by crying or groaning, whereas others are more stoic and handle their pain by gritting their teeth and suffering silently.

It is important to distinguish pain from nociception. Nociception is a physiological process by which transduction of sensory information is activated in specialized nerve endings that convey information about tissue damage (e.g., location, descriptive features) to the central nervous system. Pain, on the other hand, is an integrated perceptual process. The International Association for the Study of Pain (IASP) recognizes the complexity of pain and treats it as a phenomenological experience unique to each person. According to the IASP definition, pain is 'an unpleasant sensory and emotional experience normally associated with tissue damage or described in terms of such damage' (Merskey 1986). In addition to this global definition, many commonly used pain terms and related phenomena have been described (Turk and Okifuji 2001). For example, pain may be acute and a response to an identifiable trauma or disease, or it may be episodic and occur intermittently as in the case of migraine headache. Pain may also be relatively constant, extending for years. Unlike those with acute or recurrent pain, people with chronic pain may suffer 24 hours a day, 365 days a year. Table 1 includes a list of four classes of pain based upon timing and clinical implications.

In this article an overview is provided of the current understanding of pain and its management. First the conceptualizations that have dominated views of pain historically are reviewed. The current conceptualizations that integrate multifactorial aspects of pain are described. The article also focuses on psychosocial factors that have been shown to contribute to the subjective experience of pain as well as adaptation and response to treatment. Following the discussion of model of pain, a historical overview of the most common perspectives and treatments of pain is presented. The importance of a multidisciplinary approach to understanding and treating complex chronic pain syndromes is stressed. Finally, the psychosocial stra-

Table 1
Four types of pain

Transient Pain: Pain elicited by activation of nociceptive process but no significant injury is present and pain stops as soon as the activation ends. This type of pain is ubiquitous in everyday life and rarely is a reason to seek heath care. EXAMPLES: pain associated with injection for immunization
Acute Pain: Pain elicited by nociceptive activation at the site of injury or disease. This type of pain is often a reason to seek medical care. EXAMPLES: post-surgical pain, pain associated with muscle sprain
Recurrent Pain: Episodic or intermittent pain. Each pain episode is relatively short lasting but recurs across an extended period of time. EXAMPLES: migraine headaches, tic douloureux, sickle cell crisis
Chronic Pain: Usually elicited by an injury but lasting for a long period of time after the original damage has been healed. At the chronic stage, no objective pathology may be present that can explain the presence of pain. EXAMPLES: chronic low back pain, fibromyalgia

Source: Turk & Okifuji 2001

tegies that incorporate behavioral and cognitive–behavioral perspectives with the goals of improving pain and functioning of people suffering from pain are emphasized. It is not meant to be suggested, however, that pain is purely psychological or that physical factors are unimportant. Rather emphasis is put on psychosocial factors, as they have previously been less integrated into conventional thinking about chronic pain. The psychological approaches described are almost always employed within the multidisciplinary treatment program (Flor et al. 1992).

2. *Historical Conceptualizations of Pain*

Over the centuries, many prominent scholars have attempted to uncover the mechanisms underlying pain. Perhaps among the most influential was the seventeenth century French philosopher, René Descartes. Descartes conceptualized pain as the result of representation of a painful stimulus traveling along the nervous system starting at the periphery (the flame depicted in Figure 1 that will travel along the nerve from the periphery to the spinal cord eventually reaching the brain where it acknowledged as 'pain'). Descartes's approach formed the basis for the mind–body dualism that plays a prominent role in the understanding and treatment of pain.

Figure 1
L'homme de René Descartes (reproduced by permission of René Descartes Paris: Charles Angot, 1664)

2.1 *Biomedical (pain = physical pathology) Model*

The traditional biomedical view that continues to dominate is reductionistic. This perspective assumes that every report of pain must be associated with a specific physical cause (e.g., pain caused by the flame in Figure 1). As a consequence, the extent of pain reported should be directly proportional to the amount of tissue damage. The expectation is that once the physical cause has been identified, appropriate treatment will follow and positive outcomes will be attained. Treatment typically focuses on removing the putative cause of the pain, or by chemically or surgically disrupting the pain pathways (cutting or blocking the transmission of signals from the periphery to the brain). According to this model, once the cause of the pain is removed or the pain pathways are blocked, pain should be eliminated.

There are, however, several perplexing features of pain that do not fit neatly into the biomedical model with its suggestion of a one-to-one relationship between tissue pathology and symptoms. Most notably, not all pain accompanies observable pathology. For example, in over 85 percent of the cases of back pain, one of the most common pain conditions in Western countries, the cause of the pain is unknown (Deyo 1986). Conversely, physical pathology does not necessarily cause pain. Diagnostic imaging studies identify significant pathology in up to 35 percent of people who are pain-free (Jensen et al. 1994). Thus, a paradox exists. On one hand there are people without significant pathology who report severe pain and on the other hand there are people with objectively determined physical pathology who do not report the presence of any pain.

2.2 *Psychogenic ('It's All in Your Head') Model*

According to this model, pain, in the absence of physical pathology, is a result of the patient's inherent personality traits or psychological problems. The psychogenic view is posed as the flip side of the coin from the physical or biomedical model. From this perspective, if the report of pain occurs in the absence of, or is disproportionate to, objective physical pathology, then, *ipso facto*, the pain reports must have a psychological etiology. Thus, a dichotomy is posed, pain is either somatagenic or psychogenic.

2.3 *Motivational (Secondary Gain) Model*

A related conceptualization to the psychogenic one described specifically focuses upon motivation. From this perspective, reports of pain in the absence of physical pathology are attributed to the desire of the person complaining to obtain some benefit such as

attention, time off from undesirable activities, or financial (disability) compensation (i.e., secondary gain). Unlike the psychogenic model, in the motivational model, the assumption is that the person is consciously attempting to acquire a desirable outcome based on the complaint of pain. Thus, the complaint of pain in the absence of pathology is taken to be fraudulent. Simply put, the person who reports pain in the absence of objective pathology is lying in order to obtain a desired outcome.

2.4 *Operant (Social Reinforcement) Model*

The operant or social reinforcement model is based upon the behavioral learning paradigm in which the probability of exhibiting a specific behavior depends on the consequences that follow the behavior. Behaviors are most likely to recur when they are reinforced by desirable consequences (e.g., attention, sympathy), by removal of something aversive (e.g., participating in undesirable tasks, work activity), or avoidance of some undesirable consequence. Fordyce (1976) suggested that certain behaviors associated with pain (e.g., limping, grimacing) could be reinforced to the point where the behaviors recur without the original nociceptive stimuli. Pain behaviors that are consistent with acute pain and that may be reflexive responses to noxious stimulation, may become chronic pain behaviors even when there is no physiological basis for exhibiting those behaviors. Thus, in this model, chronic pain is considered a manifestation of learned behaviors that need to be extinguished, whereas well behaviors, such as activity, need to be increased. The laws of operant learning are viewed as playing a central role in the maintenance of chronic pain behaviors. The role of other factors, such as physiology, emotions, the persons' attitudes and beliefs, play very little role in the operant model.

3. *Multidimensional Conceptualization of Pain*

The unidimensional models described fail to explain individual and situational variation in pain experiences. The article will now focus on more contemporary models of pain with multifactorial bases.

3.1 *Gate Control Theory*

Melzack and Wall (1965) developed the gate control theory that integrated physical and psychological factors regarding pain experience. Briefly, the gate control theory proposes that a mechanism in the dorsal horn of the spinal cord acts as a 'gate' that can inhibit or facilitate transmission of nerve impulses from the periphery to the brain. Melzack and Wall postulated that not only injury sites at the periphery influence the gating mechanism but also by people's psychological states. They suggested that the reticular formation in the brain functions as a central biasing mechanism inhibiting the transmission of pain signals. Psychological factors that affect the reticular formation may modulate the pain experience.

Melzack and Wall emphasized the modulation of inputs in the dorsal horn of the spinal cord and on the dynamic role of the brain in pain processes and perception. As a result, the gate control theory integrates psychological variables such as past experience, attention, and other cognitive activities into current research on and therapy for pain. Prior to this formulation, psychological processes largely were dismissed as merely reactions to pain. This new model suggested that severing nerves, the putative pain pathways, was inadequate as a host of other factors modulated the input. Perhaps the major contribution of the gate control theory was that it highlighted the central nervous system as playing an essential role in the perception and interpretation of nociceptive stimuli associated with the experience of pain.

The physiological details of the gate control model have been challenged and it has been suggested that the model is incomplete. As additional knowledge has been gathered since the original formulation of the gate control model, specific points have been disputed and physiological details of the model have been refined. The conceptual aspects of the gate control theory have, however, proved remarkably resilient and flexible in the face of accumulating scientific evidence. The gate control model still provides a powerful summary of the phenomena observed in the spinal cord and brain, and has the capacity to explain many of the most puzzling problems encountered in the clinic. The gate control theory has had enormous heuristic value in stimulating further research in the basic science of pain mechanisms and in spurring new approaches to treatment.

3.2 *Cognitive–Behavioral Model*

The focus of the cognitive–behavioral model is on the person's own subjective perspectives (e.g., attitudes, believes, expectations) and feelings about their plight (Turk et al. 1983). The model assumes that although nociceptive stimuli precede pain, how the person perceives the nociceptive event forms a total pain experience by interacting with the sensory event. For example, negative and pessimistic views by people about their pain condition (e.g., 'It's hopeless and it will never get better.') and their capabilities for managing pain and stress (e.g., 'There is absolutely nothing I can do about my pain.') are likely to exacerbate their emotional distress and sense of disability. Similarly, if one views pain as inevitable (e.g., 'I was injured, that is why I hurt.'), attention to sensory events may become pronounced, and as a

consequence, relatively subtle sensory information may be interpreted as being painful.

The cognitive–behavioral model also acknowledges the effects of physiological factors and the environment on behaviors and of behaviors on thoughts and feelings. Reporting of symptoms to family and health care providers is influenced by how people view their pain problem and, of course, physical pathology.

The basic assumption of the cognitive–behavioral model is that people are not passive entities who simply react reflexively to nociception or social reinforcement, but actively engaging themselves in defining the experience. Based upon past learning and medical history, people develop subjective representations of illness and symptoms—'schemas.' These schema becomes the filters through which people process new sensory stimulus (Cioffi 1991).

Beliefs about the meaning of pain and one's ability to function despite pain are important aspects of pain schema. When confronted with pain, people draw causal, covariational, and consequential inferences about their symptoms based upon their own schematic references. For example, if the schema includes a strong belief that all physical activities must be ceased when experiencing pain and that pain is an acceptable justification for neglecting domestic and occupational responsibilities, poor adaptation and coping are likely to result (Williams and Thorn 1989).

4. *Treatments*

As has been discussed previously, pain is a complex phenomenon and diverse in its nature and degrees of associated impairment. It is therefore important to know the objectives of treatments for pain. Many pain states are expected to remit over time with little intervention required. Some types of acute pain may require surgical or pharmacological therapy to treat underlying pathology or to block noxious sensory information. A more comprehensive plan is required for treating people with chronic pain. Many people with chronic pain live with their pain for years, and the adverse effects of pain may be generalized across all aspects of their lives, familial, social, occupational, as well as physical. Given the multilayered problems associated with pain, treatment goals become more global. Rather than treating pain, it becomes necessary to treat people with chronic pain.

4.1 *Historical Approaches to Treatment*

The earliest approach to treating pain was aligned closely with unimodal views; particularly those based on the biomedical model. Pain was thought of as reflecting a physical problem. Early treatment approaches focused on the idea that something had to be removed or signals had to be interrupted to bring about relief from pain. The earliest reference to medication is referred to in an ancient Egyptian text dating back to 1550 BC where the god Isis recommended the use of opium to relieve Ra's headache (Bonica 1953). Ancient pain treatments have included crocodile dung, oils derived from ants, earthworms and spiders, spermatic fluid from frogs, and moss scraped from the skull of a victim of a violent death. A quick surf of the Internet will reveal that many equally esoteric preparations are endorsed readily even today.

The most common contemporary treatment approaches continue to be based primarily on the biomedical model of pain involving pharmacotherapy (with new drugs touted on an almost weekly basis) or surgery performed along nearly every site of the nervous system including the periphery (sympathectomy), the spinal cord (percutaneous cordotomy, destructive neural blockade), and the brain (thalamatomy, prefrontal lobotomy). Although current methods are more sophisticated than the ancient ones, the basic principles remain the same—alteration of an alleged physical cause should result in symptomatic improvement. Some modalities are quite helpful and a great many people suffering from acute and cancer pain have benefited from those advances in pain medicine and surgery. However, surgery and pharmacotherapy are not panaceas and do not result in adequate control of pain in all cases.

In contrast, some primitive treatment methods focused on the cause of pain being purely psychogenic. The assumption was that if no physical cause could be detected then the pain must be caused by emotional disturbance that required treatment and once psychological problems were resolved the pain would resolve. Such a psychogenic bias can be seen in the current Diagnostic and Statistical Manual (American Psychiatric Association 1994) for mental disorder. Two types of pain disorders are listed as mental disorders: Pain Disorder Associated with Psychological Factors and Pain Disorder Associated with Both Psychological Factors and a general medical condition. In both cases, pain is the primary complaint, with psychological factors being considered important in the onset, severity, exacerbation, or maintenance of the pain. A somewhat presumptuous assumption underlying these disorders is that the 'appropriate' physical etiology of chronic pain must be identifiable (e.g., free from psychological factor) or patients are suffering from a psychiatric disorder. The assumption ignores the current consensus that pain is not a pure sensory experience but is inherently a biopsychosocial phenomenon resulting in an intense, subjective experience of discomfort.

4.2 *Comprehensive Approaches to the Treatment of the Person with Chronic Pain*

The recognition of pain, particularly chronic pain, as a major healthcare issue, has resulted in the development

of growing number of specialists and specialized facilities offering treatments for pain over the last few decades. In the United States alone, over 3300 pain specialists and treatment facilities have been identified (Marketdata Enterprises 1995).

Comprehensive, multidisciplinary treatments involving teams of professionals (physicians, psychologists, physical therapists, occupational therapists, vocational therapists, and nurses) have been developed to treatment patients with the most recalcitrant problems. Flor et al. (1992) summarized the major characteristics of these patients as those with seven years of pain history and at least one failed surgery. These patients are likely to be unemployed, receiving disability payments, and present an array of psychosocial and motivational issues. In addition to the suffering associated with chronic pain, these patients present socioeconomic difficulties for society due to loss of productivity and costs associated with disability payments and medical care.

Focus will be on the pain management approaches that have been developed based upon the psychological models described earlier. Although these interventions can be used separately, with few exceptions (e.g., biofeedback for tension-type headache) unimodal approaches are less effective than comprehensive rehabilitation programs. Typically, the psychological approaches are part of more comprehensive rehabilitation program in which psychologists work together with therapist from other disciplines (e.g., physicians and physical therapists).

The multidisciplinary rehabilitation programs are philosophically distinct from unimodal medical and surgical treatments. The first and most critical distinction is that the major goals of the multidisciplinary treatment programs extend beyond pain relief. Rather, they focus on physical and psychological functioning, as well.

Overall, rehabilitation programs that incorporate a significant psychosocial component are most effective in improving patient functioning, returning patients to work, reducing use of analgesics, reducing health care utilization, and reducing disability costs (Okifuji et al. 1998). 'Cure' of pain (e.g., total relief from pain) is, unfortunately, rarely attained. Reduction of pain is usually only moderate but comparable to that observed with usual medical and surgical approaches that are more invasive and have increased likelihood of iatrogenic consequences.

4.2.1 Treatment based on the operant model. As noted previously, the operant model of pain focuses upon pain-related behaviors believed to be maintained by social reinforcement. Such behaviors are considered to be maladaptive and thus need to be eliminated. As discussed, reinforcers are the consequences of behaviors that increase the likelihood that behaviors will be repeated. Thus, treatment involves eliminating the rewarding consequences of pain behaviors and increasing the positive responses for activity and other well behaviors. In these programs, moaning, grimacing, and other pain behaviors are ignored. Usual activities of daily living and functional exercises are prescribed and positively reinforced. Medications are made time-contingent, rather than prescribed on an 'as needed' (prn) basis that is believed to reinforce positively the pain behaviors. Simultaneous with the attempts to extinguish pain behaviors, operant treatments are designed to help patients acquire a set of new, more adaptive, behaviors. Quota-based exercise programs with gradually increasing functional activities form the core of operant treatment.

Although operant factors may play an important role in the maintenance of disability, the model has been criticized at two levels. First, it fails to integrate factors other than reinforcement, such as sensory, emotional, and cognitive factors in the overall pain experience. Second, the assumption that pain behaviors are acquired, maintained, and extinguished solely through environmental reinforcement contingencies is questioned. For example, physical signs, patients' self-efficacy beliefs, and depression are also reportedly related to pain behaviors (Buckelew et al. 1953).

Pain behaviors have been considered to be maladaptive manifestations of pain, guided by an incentive for attention or avoidance of physical activity. However, pain behaviors may be functional if indeed the behaviors protect patients from further injury or exacerbation of pain. Determination of whether an overt pain behavior in a given patient is functional or maladaptive needs to be based upon careful assessment of various factors associated with his or her pain conditions not just the reinforcement contingencies. Despite the criticisms, operant-based treatments generally are successful in reducing overt pain behaviors, increasing well-behaviors, and decreasing analgesic medication.

4.2.2 Treatments based on the cognitive–behavioral model. As noted earlier, the cognitive–behavioral model of pain acknowledges the importance of cognitive variables interacting with sensory, affective, behavioral ones to establish, maintain, and exacerbate the pain experience. The nature of specific techniques may vary from program to program, however, the primary goals of the cognitive–behavioral approach are relatively uniform (i.e., enhancement of patients' sense of control over their symptoms, increased use of adaptive skills to cope with pain and stress). Cognitive–behavioral therapy is designed to assist patients to identify, evaluate, and correct maladaptive conceptualizations and dysfunctional

Table 2
Common examples of cognitive errors

Overgeneralization: Extrapolation from the occurrence of a specific event or situation to a large range of possible situations. 'This coping strategy didn't work; nothing will ever work for me'
Catastrophyzing: Focusing exclusively on the worst possibility regardless of its likelihood of occurrence. 'My back pain means my body is degenerating and falling apart'
All-or-none Thinking: Considering only the extreme 'best' or 'worst' interpretation of a situation without regard to the full range of alternatives. 'If pain is not completely gone, I cannot do anything right'
Jumping to Conclusions: Accepting an arbitrary interpretation without a rational evaluation of its likelihood. 'The doctor didn't return my call today. He thinks I am a hopeless case'
Selective Attention: Selectively attending to negative aspects of a situation while ignoring any positive factors. 'Physical exercises only make my pain worse'
Negative Predictions: Assuming the worst. 'I know I will not get better even with all these therapies and everyone will dislike me for that'
Mind Reading: Make arbitrary assumptions about others without finding out what others are thinking. 'My husband does not talk to me about my pain because he doesn't care about me'

beliefs about themselves and their predicament. Additionally, patients are taught to recognize the connections linking cognition, affect, and behavior along with their joint consequences. Patients are encouraged to become aware of and to monitor the impact those maladaptive thoughts may have in the maintenance and exacerbation of maladaptive behaviors (see Table 2 for common maladaptive thoughts).

The cognitive–behavioral approach consists of four interrelated phases. These include (a) reconceptualization, (b) acquisition of coping skills and self-management strategies, (c) skill consolidation, and (d) generalization and maintenance. The first phase, reconceptualization, uses cognitive restructuring, a method that encourages people to identify and change maladaptive thoughts and feelings that are associated with the experience of pain. The crucial element in successful treatment is bringing about a shift in the patient's thought processes, away from well-established, habitual, and automatic but maladaptive thoughts toward more hopeful and rational ones. Cognitive restructuring helps foster the reconceptualization by helping patients to become aware of the role thoughts and emotions play in potentiating and maintaining stress and pain. The aim of this phase is to combat the sense of demoralization that many with chronic pain experience. Generally, the process of cognitive restructuring begins with a presentation of a situation or event that provoked a pain-related response. The situation is dissected to identify key thoughts and feelings that precede, accompany, and follow an episode or exacerbation of ongoing pain and pain-related problems. Then patients are encouraged to challenge the legitimacy of the thoughts—was it true? Was it reasonable? Was it the only way to respond? What alternatives are available? and so on. Patients are instructed to gather evidence for or against their own maladaptive automatic thoughts. Alternatives are discussed, with the suggestion that different ways of thinking can affect mood, behavior (e.g., reducing physical activity leading to greater disability) and even physiological activity (e.g., increase muscle tension and thereby exacerbating pain).

The second phase is the acquisition of self-management skills. A wide variety of techniques have been shown to be effective for reducing suffering and disability. Some of these strategies are self-regulatory skills (e.g., relaxation, controlled breathing, and attention diversion) that allow pain sufferers to regulate their own physiological responses that may be involved in the maintenance and exacerbation of pain. Other self-management strategies include stress-reduction skills (e.g., problem-solving, behavioral rehearsals) that allow people with chronic pain to effectively manage the stress-inducing thoughts, behaviors, and emotions that trigger pain, emotional distress, and other maladaptive responses. Instead of being a passive recipient of a medical intervention (e.g., medication, anesthetic nerve block), patients now learn to use self-management strategies to play an active role in managing the myriad of problems created by the presence of persistent pain. Research has suggested that there is no one specific coping skill that best manages pain and disability (Fernandez and Turk 1992). It is recommended generally that chronic pain patients should be taught various types of coping skills to help them acquire a range of options.

Phase 3 is skill consolidation. During the skill-consolidation phase, patients practice and rehearse the skills that they have learned during the acquisition phase and apply them outside the hospital or clinic. Practice may start with the mental rehearsal, during which patients imagine using the skills in different situations. Therapists can make use of role-playing, in which patients rehearse learned skills in situations that mirror their home environments. Therapists may start with a relatively easy examples and then introduce scenarios that are progressively more realistic. The importance of skill consolidation through home-practice cannot be overstated. When patients practice skills at home, it is useful for them to record their experiences including any difficulties that arise. Once problems associated with using the newly acquired skills are identified, these become targets for further discussion.

The final stage is phase 4; the preparation for generalization and maintenance. To maximize the likelihood of maintenance and generalization of treatment gains, therapists focus upon the cognitive activity of patients as they confront problems throughout treatment (e.g., failure to achieve goals, plateaus in progress, recurrent stress). These circumstances are used as opportunities to assist patients to learn how to handle setbacks and lapses because they are probably inevitable parts of life and will occur after the termination of the treatment. In the final phase of treatment, discussion focuses on ways of predicting and dealing with symptoms and related problems following treatment termination. Patients are encouraged to anticipate future problems, stress, and symptom-exacerbating events during treatment and to plan how to respond and cope with these problems.

Since self-initiating pain management is a key factor in pain rehabilitation. Some type of cognitive–behavioral approach generally is included in a multidisciplinary pain program.

Cognitive–behavioral approaches have been demonstrated to be effective with a wide range of debilitating pain syndromes including low back pain (Lanes et al. 1995), arthritis (Parker et al. 1995), and fibromyalgia (Turk et al. 1998). Moreover, these methods have been shown to be effective with children (Walco et al. 1992) and geriatric samples (Calfas et al. 1992).

5. *Summary*

This chapter reviewed the definition of pain and the historical context within which pain has been conceptualized, made a distinction between nociception (a sensory process) and pain (a perception), and described the most common conceptualizations based the unidimensional and multidimensional models of pain. It was noted that treatment strategies have closely followed the conceptualizations of pain and have included unidimensional modalities (i.e. biomedical, psychogenic) and progressed to multidimensional modalities.

Although pain generally is considered a physical phenomenon, pain involves various cognitive, affective, and behavioral features. These psychological factors are important not only in determining the perception of pain, but also defining disability and patients' general well being. It should be clear from the review that pain has three main components: physical, psychosocial, and behavioral, that interact to define the unique pain experience. Because the pain experience is subjective and idiosyncratic, it cannot be understood without evaluating how patients perceive and appraise their conditions. A complete clinical picture involves consideration of how patients view their plight. By understanding the phenomenology of chronic pain and disability, effective treatment can be planned to alleviate persistent and debilitating pain, improve physical and psychological functioning, thereby reducing the disability that accompanies chronic pain.

See also: Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Pain, Health Psychology of; Pain, Management of; Pain, Neural Basis of.

Bibliography

American Psychiatric Association, 1994, Diagnostic and Statistical Manual of Mental Disorders: DSM-IV, American Psychiatric Association, Washington, DC

Anonymous 1995 Chronic pain management programs: a market analysis, Marketdata Enterprises, Tampa, FL

Bonica J J 1993 *The Management of Pain*. Lea & Febiger, Philadelphia

Buckelew S P, Murray S E, Hewett J E, Johnson J, Huyser B 1953 Self-efficacy, pain, and physical activity among fibromyalgia subjects. *Arthritis Care Research* **8**: 43–50

Calfas K, Kaplan R, Ingram R 1992 One-year evaluation of cognitive–behavioral intervention in osteoarthritis. *Arthritis Care Research* **5**: 202–9

Cioffi D 1991 Beyond attentional strategies: cognitive–perceptual model of somatic interpretation. *Psychological Bulletin* **109**: 25–41

Deyo R A 1986 Early diagnostic evaluation of low back pain. *Journal of General Internal Medicine* **1**: 328–38

Fernandez E, Turk D C 1992 Sensory and affective components of pain: separation and synthesis. *Psychological Bulletin* **112**: 205–17

Flor H, Fydrich T, Turk D C 1992 Efficacy of multidisciplinary pain treatment centers: a meta-analytic review. *Pain* **49**: 221–30

Fordyce W E 1976 *Behavioral Methods in Chronic Pain and Illness*. Mosby, St. Louis

Gureje O, Von Korff M, Simon G E, Gater R 1998 Persistent pain and well-being: a World Health Organization study in primary care. *JAMA* **280**: 147–51

Hart L G, Deyo R A, Cherkin D C 1995 Physician office visits for low back pain. Frequency, clinical evaluation, and treatment patterns from a U.S. national survey. *Spine* **20**: 11–19

Jensen M C, Brant-Zawadzki M N, Obuchowski N, Modic M, Malkasian D, Ross J 1994 Magnetic resonance imaging of the lumbar spine in people without back pain. *New England Journal of Medicine* **331**: 69–73

Lanes T C, Gauron E F, Spratt K F, Wernimont T, Found E, Weinstein J 1995 Long-term follow-up of patients with chronic back pain treated in a multidisciplinary rehabilitation program. *Spine* **20**: 801–6

Melzack R, Wall P D 1965 Pain mechanisms: a new theory. *Science* **150**: 971–9

Merskey H 1986 Classification of chronic pain. Descriptions of chronic pain syndromes and definitions of pain terms. *Pain –Supplement* **3**: S1–226

Okifuji A, Turk D, Kalauokalani D 1998 Clinical outcome and economic evaluation of multidisciplinary pain centers. In: Block A R, Kremer E F, Fernandez E (eds.) *Handbook of Pain Syndromes: Biopsychosocial Perspectives*, Erlbaum, Mahwah, NJ

Parker J C, Smarr K L, Buckelew S P, Stucky-Ropp R, Hewett J, Johnson J, Wright G, Irvin W, Walker S 1995 Effects of stress management on clinical outcomes in rheumatoid arthritis. *Arthritisis and Rheumatism* **38**: 1807–18

Turk D C, Meichenbaum D, Genest M 1983 *Pain and Behavioral Medicine: A Cognitive–Behavioral Perspective*. Guilford, New York

Turk D C, Okifuji A 2001 Pain terms and taxonomies of pain. In: Loeser J D, Butler S H, Chapman C R, Turk D C (eds.) *Management of Pain*. 3rd edn. Lippincott Williams & Wilkins, Philadelphia

Turk D C, Okifuji A, Sinclair J D, Starz T 1998 Interdisciplinary treatment for fibromyalgia syndrome: clinical and statistical significance, *Arthritis Care and Research* **11**: 186–95

Walco G A, Varni J W, Ilowite N T 1992 Cognitive–behavioral pain management in children with juvenile rheumatoid arthritis. *Pediatrics* **89**: 1075–79

Williams D A, Thorn B E 1989 An empirical assessment of pain beliefs. *Pain* **36**: 351–58

D. C. Turk

Chronology, Stratigraphy, and Dating Methods in Archaeology

1. Introduction: Estimating Age in the Archaeological Record

Of the twenty or so dating methods employed in twenty-first-century archaeology, the US Congressional Office of Technology Assessment (OTA) report on *Technologies for Prehistoric and Historic Preservation* (US Congress, OTA 1986) listed only seven and highlighted two radiocarbon and archaeomagnetic techniques; see Taylor 2000, pp. 75–60. In many regions of the world, radiocarbon, dendrochronology, obsidian hydration, and archaeomagnetism are the most common physical dating methods used, presumably in that order, but often together as a group providing redundant checks on one of the most important goals of archaeology—providing a framework of time, a chronology, upon which a rational reconstruction of the past can be built. Regardless of an archaeologist's theoretical or political perspectives, the first goal is to order regional prehistoric or historic sites, a single site, or a portion of the site into meaningful cultural and chronological segments. Chronology is a primary goal, but should not be confused with the intent of reconstructing the past, regardless of what an archaeologist's definition of the ultimate purpose may be.

Table 1
Datable materials and appropriate dating methods in common use in early twenty-first century

	Magnetism	Hydration	Amino Acid	E.S.R.	Luminescence	Fission Track	Uranium Series	K-Ar; $^{40}Ar/^{39}Ar$	Radiocarbon	Dendrochronology
obsidian	4	4				4		4		
other volcanic rocks	4			4	4	4	4	4		
glass						4				
unburned sediment	4				4					
burned sedimentary stone				4	4					
slag					4	4			4	
pottery, baked earth	4			4	4				4	
shell			4	4			4		4	
tooth enamel			4				4		4	
bone, antler, ivory, teeth			4				4		4	
wood/charcoal									4	4
plant fragments									4	

During the formative cultural historic perspective in archaeology in the first half of the twentieth century, constructing regional and site-oriented chronologies became the primary goal of archaeology, particularly in North America and the UK (Trigger 1989). A dissatisfaction with this rather unilineal explanation of the past stirred archaeologists to look beyond simply constructing chronologies and to begin to understand process and social relationships in the past. Dating methods and the construction of cultural chronologies, however, have remained the basement upon which archaeologists build their understanding of social process through time.

While it would be optimal to discuss all the potential dating methods available to archaeologists, this is not possible in a short synopsis. The most commonly used

========== Numerical Age ==========

======= Calibrated Age =======

======== Relative Age ========

= Correlated Age =

TYPE OF METHOD

Sidereal	Isotopic	Radiogenic	Chemical and Biological	Geomorphic	Correlation
Historical records	*^{14}C	Fission-track	Amino acid racemization	Soil-profile development	Lithostratigraphy
*Dendrochronology	K-Ar and ^{39}Ar-^{40}Ar		*Obsidian and tephra hydration	Rock and mineral weathering	Tephrochronology
Varve chronology	Uranium-series	Thermoluminescence	Lichenometry	Progressive landscape modification	Paleomagnetism
	Uranium-trend	Electron-spin resonance	Soil chemistry	*Rate of deposition	*Archaeomagnetism
	Cosmogenic isotopes (^{210}Pb, ^{10}Be, ^{36}Cl, etc.)		Rock varnish chemistry	Rate of deformation	*Fossils and artifacts
				*Geomorphic position	Stable isotopes
					Orbital variations
					Tectites and microtectites

Figure 1
Commonly used dating methods in archaeology. Those marked with (*) are discussed in some detail in this entry
Source: adapted from Waters 1992

methods, and those showing promise, will be covered here (see Table 1 and Fig. 1). Refer to the publications in the bibliography for more detailed treatments.

2. *Archaeological Stratigraphy*

Stratigraphic relations have always been the primary method to infer the relative age of artifacts within a site. Stratigraphy is defined here as the study of the spatial and temporal relationships between the sediments and the soil (Waters 1992, pp. 60–1). Much of this section's discussion can be more robustly understood by referring to the entry *Geoarchaeology*; an awareness of the geological processes of archaeological site formation is crucial to an understanding of dating archeological sediments. Indeed, the methods used by Quaternary geologists to date sediments include most of the methods used by archaeologists and discussed here (see Fig. 1). Stratigraphic studies of archaeological sites are designed to define objectively and categorize the sediments and soils, the contact units between them, and the amount of time they represent, as well as their relationship to the surrounding sediment history.

Archaeological stratigraphy is based on the geological concepts of the law of superposition, which states that older sediments are emplaced at a lower level than more recent sediments. Therefore, sites, features, and artifacts residing in lower levels are, by definition, older than those in upper levels. This conceptual framework in archaeology appears to have begun with John Frere in 1797 attempting to understand the relationship between stone axes in a sedimentary sequence in England: 'The manner in which…[the hand axes] lie would lead to the persuasion that it was a place of their manufacture and not of their accidental deposit…It may be suggested that the different strata were formed by inundations happening at distant periods' (Frere in Rapp and Hill 1998, p. 5).

This idea that features or artifacts found within a site in the same stratigraphic level are contemporaneous forms the foundation upon which archaeological stratigraphy is based. In early twenty-first century archaeology, the various relative and absolute dating methods such as radiocarbon or archaeomagnetic dating are used to verify this assumption and produce a site chronology (see Fig. 2). Fig. 2 shows a simple stratigraphic profile of a single excavation unit with the relative positions of natural and cultural features, and the position of radiocarbon dates recovered from the various levels. Note that the radiocarbon dates indicate that, in general, the stratigraphy is intact in that the oldest dates are at the bottom and the latest dates at the top, except for one date (CAMS 43177) at 3690 ± 50 BP which seems to be out of sequence and suggests 'small-scale' disturbance probably caused by the digging of the pit structure during the later ceramic period (Shackley et al. 2000).

Time-sensitive artifacts, such as previously dated pottery or projectile point types, can be used as *index fossils* within the stratigraphic column to create a relative chronology within the site. The stratigraphic

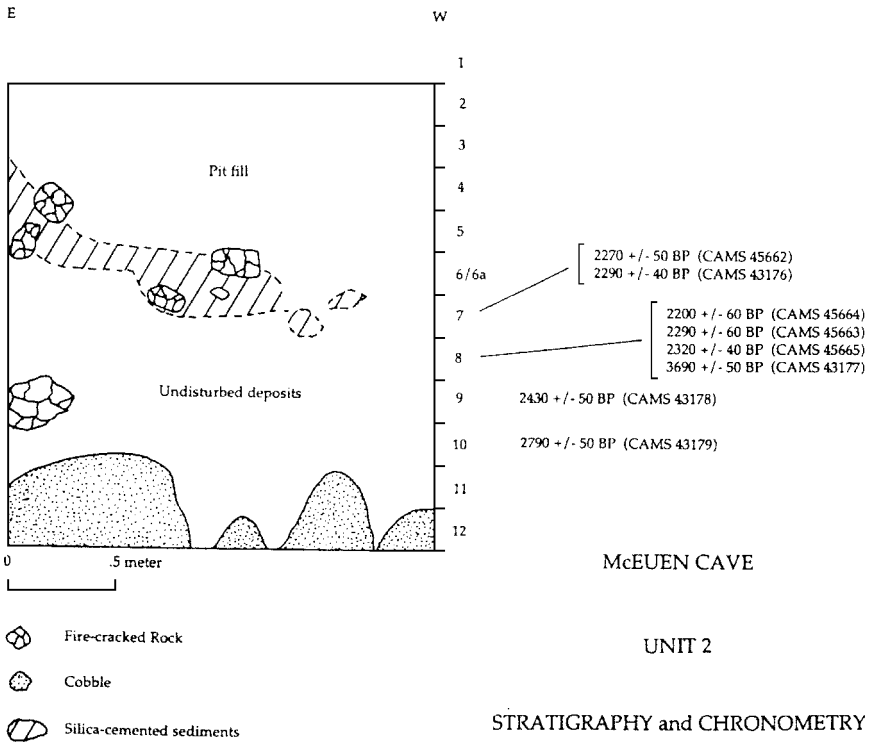

Figure 2
Simple stratigraphic wall profile of a unit at McEuen Cave (AZ W:13:6 ASM) showing position of bedrock, undisturbed strata, pit fill structure, and the position of recovered radiocarbon dates. Source: Shackley et al. 2000

example discussed above also used index fossils, in this case projectile point types, to add another chronological piece of information. In Level 8, where the early radiocarbon date as well as some more recent radiocarbon dates were recovered, Middle Archaic as well as Late Archaic projectile point forms were recovered; both confirming the time range suggested by the radiocarbon date ranges (Fig. 2, Shackley et al. 2000). In this instance the stratigraphy, radiocarbon dates, and index fossils all contributed to an understanding of the site chronology and the integrity of the archaeological deposit.

2.1 Constructing Stratigraphic Chronologies: Competing Ideas

It is important to note, however, that there is considerable recent controversy over the proper method to employ in the interpretation of archaeological stratigraphy (Farrand 1984, Harris et al. 1993, Stein 1987, Waters 1992). Archaeological stratigraphic codes similar to geological stratigraphic codes and methodologies have been proposed but not generally adopted, since many geoarchaeologists are convinced that the existing geological code is adequate for archaeological research (see Stein 1987). More recently, Harris has proposed an 'archaeological tool' to understand archaeological stratigraphy called the 'Harris Matrix,' essentially a simple postexcavation way in which to understand the relationships between stratigraphic units in a single diagram as reflected on paper, or more recently in digital form (see Harris et al. 1993). Farrand (1984) and others, mainly geologists, have questioned Harris's assumption that archaeological stratigraphy is primarily culturally conditioned rather than geologically determined. In a Harris Matrix analysis the relationship between stratigraphic units is based mainly on index fossils, for example the relationship between artifacts and features, rather than geological features or context—essentially emphasizing content rather than structure. Despite criticism from many geoarchaeologists, the Harris Matrix analysis has become increasingly popular with archaeologists working in varieties of geological and cultural contexts worldwide, particularly in Europe. The growing popularity of the Harris Matrix approach will probably continue, particularly

in small site settings where a larger geological view is difficult to obtain or not necessarily relevant.

3. Sidereal Dating Methods

Sidereal dating methods include historical records, glacial varve dating, and dendrochronology. While only dendrochronology will be discussed here, glacial varve dating is quite useful in the upper latitudes, and in some cases may be quite precise. Historical records are a basic dating tool in historic and ethnohistoric archaeology, in addition to the social information they may contain.

3.1 Dendrochronology

The science that uses annual tree rings for dating past events and reconstructing past environmental conditions has undergone explosive growth in recent decades (Dean 1997). While dendrochronology as an archaeological dating tool enjoyed tremendous expansion in the 1990s, particularly in western North America, Europe, Siberia, and the eastern Mediterranean, it is used wherever appropriate trees occur and a sequence has been established. Most importantly, directly dated tree rings have been instrumental in the calibration of the radiocarbon timescale discussed below. This process indicated that the radiocarbon chronology underestimates the true ages of materials older than 2,000 years and that ^{14}C dates must be corrected.

The technique is based on the concept that each year (four seasons) a tree will accumulate one growth ring, and that that ring's attributes reflect the specific climatic regime of that year. Since each year is essentially a unique climatic record, the attributes of the tree ring (thickness of inner and outer bands) are correspondingly unique. In trees, particularly fast-growing conifers, the progression of rings from pith to circumference presents an 'unalterable temporal order, and the production of but one ring per year provides the incremental regularity necessary to establish a fixed [and absolute] time scale' (Dean 1997, p. 34). More than 180 tree and shrub species worldwide possess the attributes required for successful dendrochronological studies: visible and unambiguous ring definition, production of a set number of rings (generally only one) per year, mainly climate-controlled growth, and the presence of useable morphological features that allow for ring comparison. Cross-dating, matching patterns of ring variation among trees, is the necessary principle of dendrochronology. A sequence within a region is derived from overlap between cut trees as well as archaeological specimens, sometimes with great time depth into the thousands of years. While this process is apparently simple, the application necessarily requires precision. At the turn of the twenty-first century, image analysis technology was being used to develop electronic workstations, including image databases in laptop computers, perform routine comparisons between the database and archaeological wood samples similar to techniques used in obsidian hydration analysis.

Today, in a number of regions of the world (upper altitudes in the North American Southwest, Western Europe, the Eastern Mediterranean) dendrochronology forms the backbone of chronometric dating. When available, dendrochronology provides an absolute chronology that anchors other dating techniques.

4. Isotopic Dating Methods

Perhaps the one area of dating that has seen the greatest advances recently are techniques based on radioactive decay: ^{14}C, K-Ar, and $^{40}Ar/^{39}Ar$. Advances in accelerator mass spectrometry and laser fusion have propelled these techniques into the forefront of the arsenal of methods used to deal with time in the archaeological record. Importantly, ^{14}C is generally restricted to dating the last 50,000 years and $^{40}Ar/^{39}Ar$, previously restricted to time periods near and over one million years, are now nearly overlapping with $^{40}Ar/^{39}Ar$ using the laser fusion method and yielding younger and younger ages. These methods do have limitations, but the organic and mineral restrictions are being transcended almost daily.

4.1 Radiocarbon Dating (^{14}C)

Given that the greatest level of human activity occurred after the evolution of modern Homo sapiens, in the last 45,000 years or so, radiocarbon dating has become the most commonly relied upon dating method in archaeology (Taylor in Taylor and Aitken 1997). Radiocarbon dating, now in its fifth decade of general use, is a primary tool used by archaeologists and Quaternary geologists to date the past.

4.1.1 The radiocarbon method. There are three principal isotopes of carbon which occur naturally: ^{12}C, ^{13}C (both stable), and ^{14}C (unstable or radioactive). The radiocarbon method is based on the rate of decay of the radioactive or unstable carbon isotope 14 (^{14}C), which is formed in the upper atmosphere through the effect of cosmic ray neutrons upon nitrogen 14. The reaction is: $^{14}N + n \Rightarrow {}^{14}C + p$ (where n is a neutron and p is a proton).

The ^{14}C formed is rapidly oxidized to $^{14}CO_2$ and enters the Earth's plant and animal lifeways through

Table 2
Recent pre-3000 BP (RCY) maize and cucurbit AMS ^{14}C dates from the American Southwest showing reporting conventions

Site	Site type/ context	Lab number	Radiocarbon age BP (1σ)	Dendrocalibrated 2σ age ranges BP [1]
McEuen Cave, AZ	rock shelter	CAMS 43177	3690 ± 50	4151 (4056--3934) 3835
Los Posos, AZ [2]	overbank alluv.	CAMS 34923	4050 ± 50	4792 (4517--4442) 4411
Three Fir Shelter, AZ	rock shelter	BETA 26275	3610 ± 170	4417 (3885) 3457
Tornillo Shelter, NM	rock shelter	GX 12720	3225 ± 240	4076 (3439--3408) 2784
Cerro Juanaqueña, CHIH[3]	trincheras	INSTARR 3985	3310 ± 60	3685 (3548--3479) 3364
Lukachukai, AZ	pit house	AA 9317	3445 ± 45	3829(3684--3543)3508
Lukachukai, AZ	pit house	AA 9321	3050 ± 50	3379(3315--3212)3004
Lukachukai, AZ	pit house	AA 9319	3135 ± 45	3443(3375--3357)3209
Bat Cave, NM	rock shelter	A 4167	3010 ± 150	3550 (3205--3083) 2778
Bat Cave, NM	rock shelter	A 4189	3060 ± 110	3473 (3317--3214) 2894

Source: Shackley *et al.* 2000
1 The 2σ dendrocalibrated minimums (intercepts) and maximums are based on a bidecadal correction in calendar years BP, rounded to the nearest decade, from Stuiver and Reimer (1993, version 3.0.3c, Method A). The maize dates are not corrected for C4 carbon uptake and can actually be up to 240 years older than reported here assuming –10 0/00. 2 This date and context is somewhat suspect, but was associated with a similar charcoal date at Los Pozos in the Tucson Basin (Freeman 1997, see also Gregory 1999). 3 This date on a *Cucurbita sp.* is probably from a wild species. However, three specimens of maize were dated to over 2800 RCY BP (Hard and Roney 1998, p. 1661).

photosynthesis and the food chain. Plants and animals which utilize carbon in biological food chains take up ^{14}C during their lifetimes. They exist in equilibrium with the ^{14}C concentration of the atmosphere; that is, the number of ^{14}C atoms and nonradioactive carbon atoms stays approximately the same over time. As soon as a plant or animal dies, they cease the metabolic function of carbon uptake; there is no replenishment of radioactive carbon, only decay.

Libby, Anderson, and Arnold (Taylor and Aitken 1997) first discovered that this decay occurs at a constant rate. They found that after 5,568 years half the ^{14}C in the original sample will have decayed, and that after another 5,568 years half of that remaining material will have decayed, and so on. The half-life ($t_{1/2}$) is the name given to this value, which Libby measured at 5568 ± 30 years. This became known as the Libby half-life. After 10 half-lives, there is a very small amount of radioactive carbon present in a sample. At about 50,000 to 60,000 years, then, the limit of the technique is reached (beyond this time, other radiometric techniques must be used for dating, such as $^{40}Ar/^{39}Ar$, as discussed above). By measuring the ^{14}C concentration or residual radioactivity of a sample whose age is not known, it is possible to obtain the number of decay events per gram of carbon. By comparing this with modern levels of activity (1890 wood corrected for decay to AD 1950) and using the measured half-life, it becomes possible to calculate a date for the death of the sample.

It follows from this that any material which is composed of carbon may be dated. Herein lies the true advantage of the radiocarbon method—it is able to be uniformly applied throughout the world. A list of the tremendous quantity of organic material that can be dated by radiocarbon is available at the University of

Waikato's radiocarbon Web site, a standard for understanding the technique: http://c14.sci.waikato.ac.nz/webinfo/int.html.

4.1.2 Calibration and radiocarbon dating. As mentioned above, there are a number of effects that can cause errors in the measurement of radiocarbon dates. For example, shell, in constant contact with more recent atmospheric carbon, will generally yield young dates; conversely a shell artifact deposited in older limestone sediments will obtain a much older date than its actual death. These reservoir effects are in part mitigated by the use of various calibration algorithms, such as the CALIB (2001) and OxCal (2001) programs, both available on-line.

When a radiocarbon lab returns a date from a sample such as 5568 BP, it does not mean that it dates to 3619 BC, because the true half-life of radiocarbon is 5730 years, and, more importantly, the proportion of radiocarbon in the atmosphere has varied through time, as discussed above. So the calibration utilities are written to allow for differential in the absorption of ^{14}C by different materials (i.e., marine shell versus wood charcoal), and to allow for different atmospheric effects. Using the CALIB 4.2 calibration, a radiocarbon assay of 5568 BP with a 1 standard deviation of 55 years on wood charcoal yields a date of:

68.3 (1 Σ) cal 4454–4416 BC
probability distribution 0.409
4408–4354 BC 0.591

Note that there are two dates with ranges of a number of years. The range includes the one standard deviation, and the two dates are due to multiple intercepts on the calibration curve. We can be 68.3 percent certain that the dates fall either from 4454 to 4416 BC or from 4408 to 4354 BC. There are multiple possible dates because the radiocarbon date of 5568 BP intercepts the calibration curve at more than one point. Due to the perturbations in the absorption of radiocarbon over the millennia, the variance is sometimes full of 'wiggles' on the curve, so placement can occur at two or more points.

Radiocarbon dates are reported in a standardized method, as shown in Table 2. The convention calls for reporting the provenience of the sample, the laboratory number, the radiocarbon age, and then the calibrated (in this case dendrocalibrated) age at one or two standard deviations (note that most of these dates yielded multiple intercepts). The calibration method used and any further contextual information should also be supplied.

4.1.3 Accelerator mass spectrometry (AMS) and radiocarbon. From the inception of radiocarbon dating, ^{14}C ages of samples were calculated by decay counting in mainly scintillation counters. This requires a relatively large sample, depending on the amount of carbon remaining in that sample. By the late 1970s a number of researchers discovered that when accelerating sample atoms in the form of ions to much higher energies in particle accelerators, a much smaller sample was required to derive confident dates—in most cases only milligrams instead of tens of grams for scintillation counting. Both cyclotron and tandem accelerator mass spectrometers have been used to accomplish this, with tandem accelerators becoming the most popular. One additional advantage of acceleration is that the 'stripping process' disassociates all molecular species with the result that carbon isotopes can be isolated, and contamination minimized. AMS ^{14}C dating theoretically may push the time frame back to 100,000, effectively overlapping $^{40}Ar/^{39}Ar$ laser fusion dating (Taylor and Aitken 1997).

5. Radiation Dating Methods

As with most of the other types of dating methods, radiation dating methods have undergone tremendous advances in the last decade, although many of these methods remain somewhat controversial in their applications. Electron spin resonance (ESR) and thermoluminescence (both based on the accumulation of trapped electrons in minerals), and fission track dating (produced when alpha particles are created by spontaneous fission of ^{238}U leaving a damage trail), have all had their detractors, but have recently gained recognition as techniques that are useful in dating materials not readily possible with the more accepted technology, for example glass, tooth enamel, and pigments (see Table 1). Archaeomagnetic dating has gained wide acceptance in many regions of the world and deserves some discussion.

5.1 Archaeomagnetic Dating

Archaeomagnetic and paleomagnetic dating both rely on the phenomenon of the frequent and predictable shifts in the Earth's magnetic polarity in space and time. The premise for the method states that the Earth's magnetic poles wander (show secular variation) or flip (reverse direction), and that these variations provide a temporal fingerprint that can be detected in rock and sediments. Iron dipoles in minerals in soft sediments align themselves parallel with the Earth's magnetic field at the time of deposition until the sediment solidifies, for example as baked clay in a hearth. This creates detrital remnant magnetism whose direction and intensity precisely reflects the Earth's magnetic field at the time of deposition; and by matching their magnetic orientation to a master record, it is possible to derive a relatively accurate date. This master curve has been

derived for the last 10,000 years. Archaeomagnetism is useful in dating material, as appropriate, when other useful dating material, appropriate for other techniques is absent.

6. Chemical, Temperature, and Water Affected Dating Methods

Some of the more controversial dating methods are those that are based upon changes in chemicals, temperature, and/or water for calibration. Amino acid racemization held great promise for dating organic materials, but is not generally reliable, while obsidian hydration dating, which was also promising, has become equally problematic.

6.1 Obsidian Hydration Dating

Obsidian, a quenched rhyolite glass is common along plate boundaries and volcanic arcs where crustal remelting has occurred. A glassy brittle rock with remarkable cutting properties, it has been used throughout human history in the production of stone tools (Shackley 1998). Obsidian is a noncrystalline glass, and therefore a disordered substance, and moves toward an ordered state by crystallization or, more precisely, perlitization (Friedman et al. 1997, Stevenson et al. 1998). This is accomplished by the absorption of water and a breaking of silicon and aluminum bonds, ultimately devitrifying the glass and forming perlite. As the process proceeds, an absorption front and hydration rim forms that theoretically occurs at a regular, linear rate through time. Measuring this rim with a petrographic microscope in microns theoretically yields an absolute date. On the surface, this would seem to be a remarkable method, resolving a number of dating issues in archaeological contexts in which organic datable material does not occur, but obsidian artifacts are abundant. More recently, the intervening negative effects of variable temperature and humidity through time have been found to influence the rate at which glass will hydrate. Due to this a number of researchers have rejected the method *in toto* (see Morgenstein et al. 1999, Ridding 1996). Still, a number of researchers have continued to utilize the method, and it has a number of devoted followers, even in the face of discrepancies between radiocarbon dates and obsidian hydration dates in the same stratigraphic position. Stevenson, and others have attempted to derive intrinsic hydration models based on the premise that each single obsidian nodule contains a different proportion of water than any other, and this will produce ages compatible with other dating methods (see Stevenson et al. 1998). Most do agree that obsidian hydration is a useful relative dating method which can be used to determine the extent of mixing in a stratigraphic column; given the model, one expects to see larger rim measurements at the bottom of a site than at the top. If this is not the case, then stratigraphic mixing is indicated. Obsidian hydration holds great promise, but most archaeologists are hesitant to include it in the arsenal of absolute dating methods.

See also: Geoarchaeology

Bibliography

CALIB Radicarbon Calibration Software (2001) http://depts.washington.edu/qil/calib/

Carver M O H 1990 Digging for data: Archaeological approaches to data definition, acquisition and analysis. In: Francovich R, Manacorda D (eds.) *Lo scavo archaeologica: della diagnosi all'edizione*. Edizioni all'Insegna del Giglio, Firenze, Italy, pp. 45–120

Colcutt S N 1987 Archaeostratigraphy: A geoarchaeologist's viewpoint. *Stratigraphica Archaeologica* **2**: 11–18

Dean J S 1997 Dendrochronology. In: Taylor R E, Aitken M J (eds.) *Chronometric Dating in Archaeology*. Kluwer/Plenum Publishers, New York, pp. 31–64

Farrand W R 1984 Stratigraphic classification: Living within the law. *Quarterly Review of Archaeology* **5**: 1–5

Freeman A K L 1997 Middle to Late Holocene stream dynamics of the Santa Cruz River, Tucson, Arizona: Implications for human settlement, the transition to agriculture and archaeological site preservation. Ph.D. thesis, University of Arizona, Tucson, AZ

Friedman I, Trembour F W, Hughes R E 1997 Obsidian hydration dating. In: Taylor R E, Aitken M J (eds.) *Chronometric Dating in Archaeology*. Kluwer/Plenum Publishers, New York, pp. 297–322

Göksu H Y, Oberhofer M, Regulla D 1991 *Scientific Dating Methods*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Gregory D A, Adams J L (eds.) 1999 *Excavations in the Santa Cruz River Floodplain: The Middle Archaic Component at Los Pozos*. Center for Desert Archaeology, Tucson, AZ

Hard R J, Roney J R 1998 A massive terraced village complex in Chihuahua, Mexico, 3000 years before present. *Science* **279**: 1661–4

Harris E C, Brown M R III, Brown G J (eds.) 1993 *Practices of Archaeological Stratigraphy*. Academic Press, London

Herz N, Garrison E G 1998 *Geological Methods for Archaeology*. Oxford University Press, Oxford, UK

Morgenstein M E, Wicket C L, Barkatt A 1999 Considerations of hydration-rind dating of glass artefacts: Alteration morphologies and experimental evidence of hydrogeochemical soil-zone pore water content. *Journal of Archaeological Science* **26**: 1193–210

OxCal Radiocarbon Calibration Software (2001) http://units.ox.ac.uk/departments/rlaha/orau/06_frm.htm

Rapp G Jr., Hill C L 1998 *Geoarchaeology: The Earth Science Approach to Archaeological Interpretation*. Yale University Press, New Haven, CT

Ridding R 1996 Where in world does obsidian hydration work? *American Antiquity* **61**: 136–48

Shackley M S (ed.) 1998 *Archaeological Obsidian Studies: Method and Theory*. Kluwer Academic/Plenum Publishing, New York

Shackley M S, Huckell B B, Huckell L W 2000 Late Preceramic farmer/foragers at the foot of the Mogollon Rim: The McEuen Cave Archaeological Project Testing Report (AZ

W:13:6 ASM). Report prepared for the Bureau of Land Management, Safford Area, AZ, US Department of Interior
Stein J K 1987 Deposits for archaeologists. *Advances in Archaeological Method and Theory* **11**: 337–95
Stevenson C M, Mazer J J, Scheetz B E 1998 Laboratory obsidian hydration rates: Theory, method, and application. In: Shackley M S (ed.) *Archaeological Obsidian Studies: Method and Theory*. Kluwer Academic/Plenum Press, New York, pp. 181–204
Stuiver M, Reimer P J 1993 Extended ^{14}C data base and revised CALIB 3.0 ^{14}C age calibration program. *Radiocarbon* **35**: 215–30
Taylor R E 2000 Science-based dating methods in historic preservation. In: Williamson R A, Nickens P R (eds.) *Science and Technology in Historic Preservation*. Kluwer Academic/Plenum Press Publishers, New York, pp. 75–96
Taylor R E, Aitken M J (eds.) 1997 *Chronometric Dating in Archaeology*. Kluwer Academic/Plenum Publishers, New York
Trigger B G 1989 *A History of Archaeological Thought*. Cambridge University Press, Cambridge, UK
US Congress, Office of Technology Assessment 1986 *Technologies for Prehistoric and Historic Preservation*, OTA-E-319. US Government Printing Office, Washington, DC
Waters M R 1992 *Principles of Geoarchaeology: A North American Perspective*. University of Arizona Press, Tucson, AZ

S. Shackley

Church and State: Political Science Aspects

The umbrella term Church and State is conventionally used to refer to a range of topics within political science which concern the relationships between religious organizations, institutions, and authorities on the one hand and the polity on the other. Across the world religion has from the earliest times been centrally concerned with the shaping and authoritative allocation of values with which political science in its broadest signification is concerned. In the West churches in their proper sense represent only one type of religious collectivity—albeit historically the most important—whose relationship with the state has been the subject of intense and prolonged controversy; sects, denominations, and cults, each with their distinctive characteristics and drives, have also given rise in their orientation to political authorities, and vice versa, to difficult issues of recognition, regulation, and control. Within the other world religions the term church has no analogue, although the Buddhist *sangha* is often thought to share some features. In the other monotheistic religions, such as Islam and Judaism, and in Hinduism the location of religious authority and the degree of its independence from political authority suggest quite different patterns from those typical of church–state relations in the West. The establishment since 1948 through various international charters and conventions on human rights of foundational claims to religious freedom presents the challenge of institutionalizing ostensibly universal claims in this field within culturally diverse contexts.

1. *Church and State in the Perspective of Political Development*

In the developmentalist, or modernization, perspective traditional religiopolitical systems were typically understood to combine religious and political functions within a single structure of authority. In certain, particularly oriental, cultures this took the form of institutions of divine kingship or theocratic rule, while in Christian cultures, where the church was from early on conceived as a distinct institution, rulers usually claimed and often successfully asserted over against the church some species of sacral authority. From the time of Constantine, the emergent system of church establishment shifted from epoch to epoch between imperial authority over the church and Papal or Patriarchal authority over the emperors and other temporal authorities, a tension which was only partially resolved into contrasting patterns prevalent in different parts of the Christian world at the times of the East–West Schism in the eleventh century and the Reformation/Counter-Reformation in the sixteenth. In the high Middle Ages the Popes attempted to assert theocratic authority over the kings, princes, and emperors of Western Europe, while in the Orthodox East the Caesaropapist pattern of harmony between a dominant emperor and a subservient patriarch predominated. After the Reformation the authority of secular princes vis-à-vis the church was promoted both in the Protestant north and in the Roman Catholic south as the contest was prosecuted by and between the sponsors of the respective confessions.

Within Islam the institutional differentiation between religious and secular authority did not develop as Medina under Muhammad was considered to provide an unimprovable model for the life of the community of believers. The Prophet himself recognized the sovereignty of God alone; accordingly, it has not been open to his successors (the caliphs) or, more recently, the people and their representatives to claim sovereign authority in their own right. Within this restrictive view it has been the duty of all who exercise authority on behalf of the community faithfully to implement the rules encapsulated in the *Qur'an* and the traditions of the Prophet as interpreted by legal scholars, the jurisconsults. Similarly, within Judaism, the authority of divine law, as interpreted by rabbinical scholars, was held to be supreme within the *diaspora*. The recent reassertion of theocratic claims such as these has helped to undermine the plausibility of the developmentalist perspective with its implication of a unilinear secularization process.

2. Religion and the Rise of the Modern State in the West

The rise of the modern state with its characteristic claim to exclusive authority within a particular territory can be seen as part cause, part consequence of the Renaissance and Reformation crises in Western Europe which divided it between a Protestant north and a Roman Catholic south. It was the religious wars of the sixteenth and seventeenth centuries, brought to an end by the Peace of Westphalia of 1648, which institutionalized, over the protests of the Pope, the right of sovereign authorities to decide *inter alia* on the religious constitution of their territories. In northern Europe, typically, Protestant national monarchies reinforced their developing authority (and their resource base) by taking the church over and introducing more or less Erastian (i.e., unilateral secular) patterns of control; in the south the typical pattern which resulted was an alliance between royal absolutism and national Catholic hierarchies until the French Revolution led to a similar assault on the church's property and independent authority. The subsequent decline of dynastic rule throughout Europe in the nineteenth century and the emergence of nationalist and liberal democratic mass politics saw a further weakening of church authority north and south, east and west. Institutional secularization was furthermore accompanied by a decline in levels of orthodox religious belief and practice as urbanization and industrialization transformed social and economic structures, although revivalism of various sorts meant that the trend was not unilinear. In the USA the first amendment to the Constitution introduced for the first time a relatively thoroughgoing separation of church and state, which barred Congress from making any law 'respecting an establishment of religion or prohibiting the free exercise thereof.' A feature of the pattern which emerged in this context has been that, far from following the general secularising trend seen in Europe' a 'churching of America' occurred as the proportion of members and attenders in the population tended to increase. Few other political systems, even among the liberal democracies, have, however, followed the example of constitutionally mandated disestablishment and even where they have, as for example in the case of France in 1905, the principle of separation has not been upheld with the judicial rigor characteristic of the American case in recent decades.

3. Religious Cleavages and the Development of Party Systems

With the development of mass electoral politics in the nineteenth century the emergent party systems tended to reflect divergent systems of religious—in addition to social, economic, and other—cleavages; indeed, the relative prominence of religious or religion-related cleavages during the formative period up to the 1920s accounts for the survival, by inertia, of patterns of confessional politics which otherwise might well have disappeared. In northern Europe the existence of state–church systems had led over time to the emergence of movements of religious protest and dissent, which typically aligned themselves with other reforming elements in left or liberal movements. In the Counter-Reformation south on the other hand, where religious dissent had been more or less successfully repressed by the historic alliance of throne and altar, reforming and revolutionary movements tended to be militantly secularist, with the effect that the connection between religion and the political right was consolidated. A third pattern developed in the band of mixed-religion territories, which spanned Europe from Ireland to Transylvania, where the liberalization of political systems, when it occurred, ushered in patterns of confessional politics which, in nonmajoritarian settings, took the form of consociationalism.

Beyond Europe, developing party systems tended only marginally or incidentally to be affected by religious cleavages as such; religious connections were usually outgrowths of ethnic or community identities, whether on the part of immigrant populations as in the USA, South Africa, or Australia, or on the part of anticolonial movements committed to seizing independence on behalf of indigenous populations. In the latter case independence movements varied in terms of the religious structure of the populations affected, with the formerly united Indian subcontinent divided between Hindu, Islamic, and (in the case of Sri Lanka) Buddhist leaderships heading up parties which, on independence, became dominant in their respective systems. In more recent decades the often secular leaderships of the movements which successfully claimed independence have been challenged by movements of religious insurgency, driven in part by disappointment with the fruits of independence. Some authors have even identified in these developments a global revival of the influence of religion in politics.

4. The Political Resurgence of the Religious Factor

The late twentieth century has certainly seen a recrudescence of the religious factor in politics which would have surprised (and dismayed) the more naive modernization theorists. In 1979–80 the religious-led Iranian revolution, the prominence in the United States Presidential election of an emergent new Christian right, the overthrow of Nicaraguan dictator Somoza, and the emergence after the visit of Pope John Paul II to his home country Poland of the church-supported Solidarity, indicated that the phenomenon was not restricted to any one corner of the world. Each case could be seen as involving reactions to what were regarded by the 'fundamen-

talist' militants and activists involved as trends toward secularization or obstacles of entrenched corruption. Prominent among the issues involved were respect for religious authority, the maintenance of religion-related rules governing morality in all its guises (including in the West abortion, pornography, and contraception), and the re-introduction as normative for the political community of core religious values whether Christian, Islamic, Judaic, Buddhist, or Hindu. The liberal rule that the state should be neutral as between different religions and between the religious and nonreligious was most often regarded as either illusory or misguided: illusory, because the all-embracing nature of religious belief- and value-systems left no neutral space to be legitimately marked out, or misguided, because political authority in particular ought not to be exempted from the authoritative purview of religious values.

5. *The Contemporary Challenge to State Religious Neutrality*

In few, if any, cases has the ideal of state religious neutrality been realized, as the concept itself has progressively become problematized by significant voices claiming that the term has often been made to cover for the artificial and prejudicial exclusion from public debate of religious claims. Until the 1950s in the USA 'mainline' Protestant religion remained as a sort of informal establishment, recognized, albeit without the confessional label, in certain Supreme Court judgments, but since the early 1960s the separation rule has been applied to exclude religious symbol and activities from public institutions such as schools. In a number of cases of continuing church establishment, evenhandedness has been approached by the extension of the list of recognized religions and other 'communities of belief' (as in Belgium, where the state pays the salaries and pensions of religious officials of six recognized confessions) or the effective diminution of the privileges of establishment (as in the UK). Since the end of the Cold War official hostility to religion as such (amounting in the case of Albania to the attempted suppression of religion for over 20 years after 1967) has greatly diminished, although it has not disappeared; thus, China and North Korea, for example, continue to exercise heavy-handed control of religious bodies and activities. Within the world of Islam meantime there has been a significant growth in the number of states, which are either organized as theocratic regimes dedicated to the implementation under clerical leadership of religious law, the *shari'a* (e.g. Iran and Sudan), or which harbor important movements that aim to introduce such a regime (e.g. Pakistan and Algeria).

In the West in particular, the increasing religious diversity of populations occasioned by the emergence of new religious movements, both independently and within existing traditions, and the growth of immigrant communities with distinct confessional profiles has given rise in recent decades to a range of new problems, or the re-emergence of old ones. Thus in connection with Scientology and Transcendental Meditation, questions of recognition have arisen in several countries: are they to be regarded as religions and so accorded attendant tax or other advantages, or not? More established religious traditions, in particular Islam, have called for their values to receive the protection of the law, so that what is judged to be blasphemy should be punishable through the civil courts. In the field of education, historically one of the most contentious in church–state relations, several issues continue to present themselves: should independent schools provided for the education of children of particular religious communities receive recognition, tax advantages, or direct state aid? What provision, if any, should public schools make for collective worship or religious instruction and how should any such arrangements take account of the new religious pluralism? Should public schools allow the display of distinctive religious symbols either on the walls, as in the case of crucifixes, traditional in Catholic communities, or on the pupils themselves, as in the case of Islamic girls' headscarves? Many of these issues continue to agitate public debate episodically but none so much as the traditionally religion-related issues, such as abortion, where religious conservatives of the most varied religious persuasion continue to prosecute a bitter struggle against what is typically seen as an outgrowth of secular humanism.

See also: Civil Religion; Political Culture; Political Sociology; Reformation and Confessionalization; Religion and Politics: United States; Religion, History of; Religion: Mobilization and Power; Religion: Nationalism and Identity; Religion: Peace, War, and Violence; Religious Nationalism and the Secular State: Cultural Concerns

Bibliography

Boyle K, Sheen J (eds.) 1997 *Freedom of Religion and Belief. A World Report*. Routledge, London

Monsma S V, Soper J C 1997 *The Challenge of Pluralism. Church and State in Five Democracies*. Rowman Littlefield, Lanhom, MD

Robbers G (ed.) 1996 *State and Church in the European Union*. Nomos

Robbins T, Robertson R (eds.) 1987 *Church–State Relations: Tensions and Transitions*. Transaction Books, New Brunswick, NJ

Weber P 1990 *Understanding the First Amendment Religion Clauses*. Greenwood Press

Witte J 1999 *Church and State: The American Constitutional Experiment*. Westview Press, Greenwood, CT

J. T. S. Madeley

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Cingulate Cortex

The cingulate cortex is located in the medial wall of the cerebral hemispheres and has extensive reciprocal connections with various limbic structures as well as motor and premotor areas. In studies with rodents and studies of brain activation in humans, the cingulate cortex is implicated in the processes of selective attention and response selection. This article explores the essential areas of convergence of the two bodies of literature, and offers a common mnemonic/associative interpretation of cingulate cortical function.

1. *Anatomy of the Cingulate Cortex*

The cingulate cortex is a part of the limbic cortex, a term referring to the cortical areas that receive axonal fibers from neurons in the anterior group of thalamic nuclei. By modern convention, cingulate cortex constitutes Brodman's areas 24 and 29 in small animals such as rabbits and rats, and also an additional area, 23, in primates and humans. Brodman's areas 24 and 29 are often referred to, respectively, as anterior and posterior cingulate cortex. A cytoarchitectural map illustrating these areas is shown in Fig. 1.

Both the anterior and posterior cingulate cortices receive afferent fibers from the anterior medial (AM), midline, and intralaminar thalamic nuclei. However, the anterior cingulate cortex is also innervated by the medial dorsal (MD) and parafasicular thalamic nuclei, while the posterior cingulate cortex receives projections from the remaining members of the anterior group of thalamic nuclei: the anterior ventral (AV), anterior dorsal (AD), and lateral dorsal (LD) thalamic nuclei. In addition, neurons of the lateral dorsal and ventral anterior thalamic nuclei project to the posterior cingulate cortex.

Cingulate cortical neurons are richly innervated by pontine and midbrain fibers that fairly uniformly distribute norepinephrine and serotonin but restrictively distribute dopamine to the anterior cingulate cortex. Many additional afferent systems project to the cingulate cortex, including fibers from the visual cortex, hippocampus, subiculum, entorhinal cortex, and amygdala.

Cingulate cortical neurons send efferent fibers to most of the aforementioned thalamic areas, the subiculum, entorhinal cortex, pons, and many areas of the striatal motor system including the caudate nucleus, nucleus accumbens, and zona incerta. Cingulate cortical neurons have also been found to project to multiple areas of the motor and premotor cortex, suggesting that numerous parallel pathways exist whereby cingulate neurons can modulate motor output systems of the brain. In primates there exist direct reciprocal projections of cingulate neurons to lateral

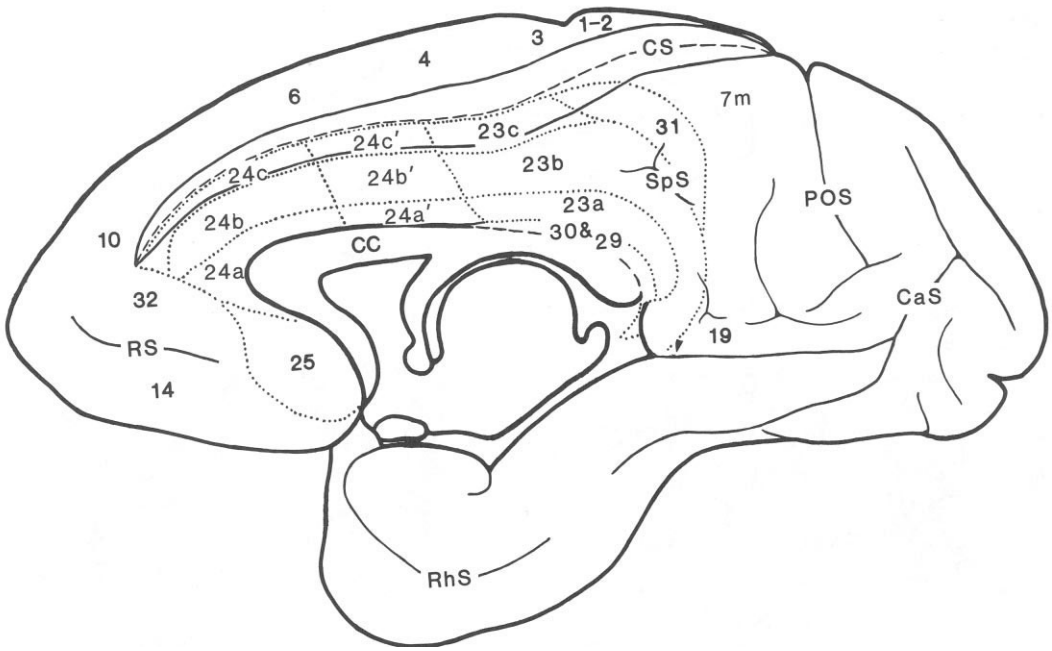

Figure 1
Cytoarchitectural map of the limbic cortex in the rhesus monkey based on Brodman's divisions

prefrontal and parietal cortex, areas believed to be critical for higher-order perceptual and mnemonic functions (Goldman-Rakic 1988).

2. The Role of the Cingulate Cortex in Attention

2.1 Associative Attention: Cingulate Cortical Neuronal Activity

Research in behavioral neuroscience with rabbits and rats indicates that the cingulate cortex mediates selective attention, or attention focused on particular stimuli. The stimuli that are selectively processed by cingulate cortical neurons are associatively significant stimuli, i.e., stimuli that signal important events such as reward or aversion, and call for action on the part of the subject. Since the selective processing of the significant stimuli is a learned, associative process, an apt characterization of the cingulate cortex is that it mediates associative attention to significant stimuli.

Support for the hypothesis that the cingulate cortex mediates associative attention to significant stimuli is shown by results of numerous studies on the activity of cingulate cortical neurons during Pavlovian and instrumental conditioning in animals. For example, extensive research has documented the responses of cingulate cortical neurons during discriminative instrumental learning in rabbits (Gabriel 1993, 2001). In these studies, rabbits occupying a large rotating wheel apparatus learned to step in response to a tone cue (CS+) to prevent a foot-shock delivered five seconds later, and they learned to ignore a different tone (CS−) not followed by shock. Neuronal activity in multiple areas of the cingulate cortex exhibited the development, during training, of massive discriminative neuronal activity, defined as significantly greater firing frequencies in response to the CS+ than to the CS− (see Fig. 2). Discriminative activity also developed during acquisition of a discriminative approach response, in which rabbits learned to make oral contact with a drinking spout to obtain a water reward following the CS+ and to inhibit contact following the CS−, which did not predict a water reward (Freeman et al. 1996).

Discriminative neuronal activity in the cingulate cortex has also been reported during classical Pavlovian conditioning of heart rate and eyeblink responses in rabbits (Powell et al. 1990). In addition, studies in rats have demonstrated the occurrence of neuronal responses in the anterior and posterior cingulate cortex that are specific to stimuli that predict reinforcement during appetitive conditioning (Segal and Olds 1972, Takenouchi et al. 1999). The activity of cingulate cortical neurons in all these studies may be viewed as a neuronal code for the associative significance of cues since the neurons developed selective responses to stimuli that predicted the occurrence of significant events.

Cingulate cortical neurons have also been shown to exhibit salience compensation, a phenomenon that is supportive of a cingulate cortical involvement in mediating associative attention (Gabriel 1993, Gabriel and Taylor 1998). When rabbits are trained with nonsalient cues such as a brief duration (200 ms) CS+ and CS−, greater brief-latency cingulate cortical discriminative neuronal responses are observed than when training is carried out with more enduring (e.g., 500 ms) CSs. The enhanced neuronal encoding of the brief stimuli, or salience compensation, has been interpreted as an attentional process that amplifies the neural representation of nonsalient yet associatively significant stimuli in order to maximize the resources available for processing those cues.

The importance of the attentional processing in the cingulate cortex is demonstrated by studies showing that bilateral, combined lesions of the anterior and posterior cingulate cortices severely impair discriminative avoidance and approach learning in rabbits (Gabriel 1993). In addition, adult rabbits exposed to cocaine *in utero*, an exposure which induced morphological and biochemical abnormalities in the anterior cingulate cortex, exhibited attenuated anterior cingulate discriminative neuronal activity and learning deficits when nonsalient CSs were used during discriminative avoidance learning and Pavlovian conditioning of eyeblink responses (Gabriel and Taylor 1998, Romano and Harvey 1998). The results of these studies illustrate that deficits due to cingulate cortical damage emerge when a high demand is placed on attentional processing.

2.2 Executive Attention: Anterior Cingulate Cortex

The application of neuroimaging and electrical recording techniques, such as positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and high-density electroencephalography (EEG), has led to a large volume of data that implicate the cingulate cortex in cognitive processing. In convergence with results from animal studies, many cognitive experiments in humans have supported the idea that the anterior cingulate cortex is involved in processes subserving attention. These studies show that the anterior cingulate cortex is engaged during tasks in which routine or automatic processing is insufficient, as when novel or conflict-laden situations are encountered. This type of attentional processing has been termed *executive attention* (Posner and DiGirolamo 1998). Situations likely to require executive attention are: a) planning and decision making; b) error detection; c) novel and early stages of learning; d) difficult and threatening situations; and e) overcoming habitual behavior.

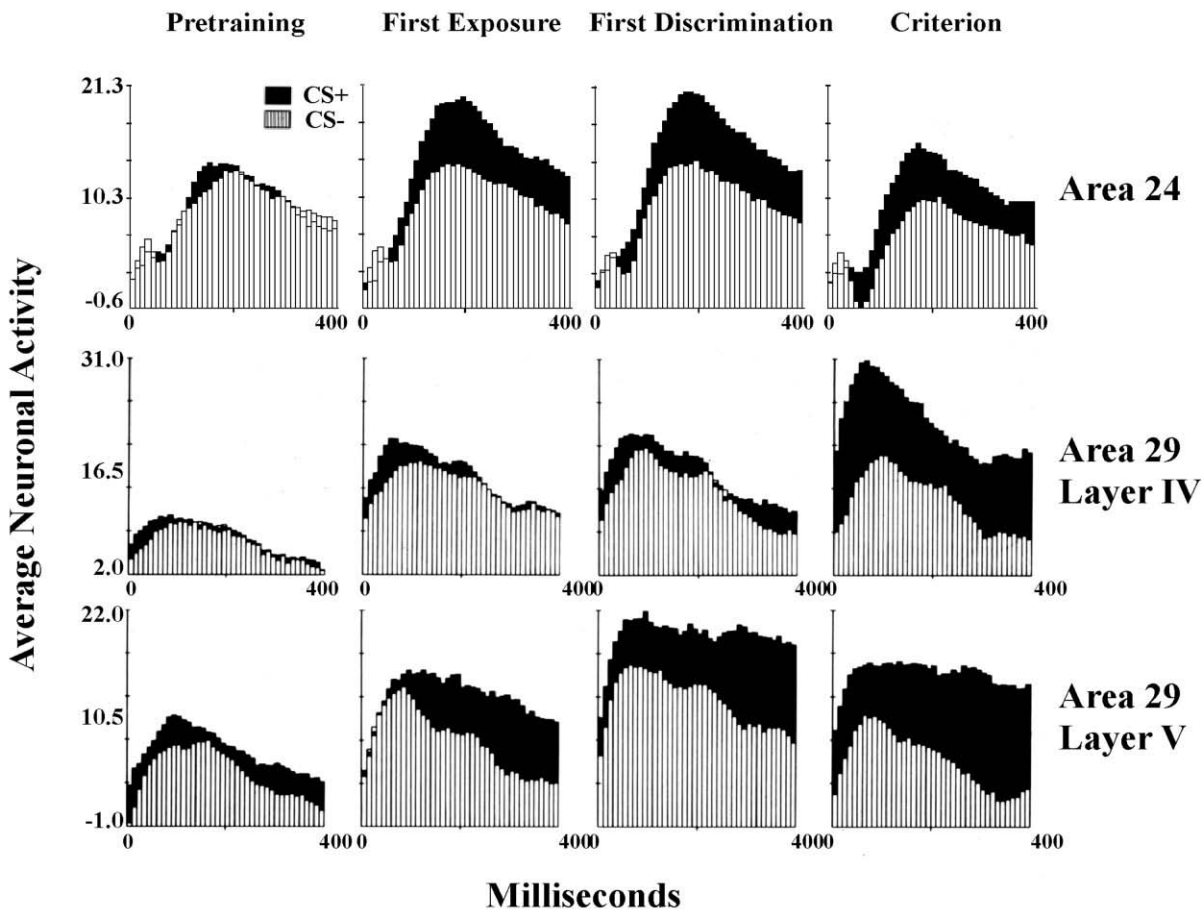

Figure 2
Average anterior (Area 24) and posterior (Area 29, cellular layers IV & V) cingulate cortical integrated unit activity elicited by CS+ and CS− in rabbits during pretraining, first exposure session, first significant behavioral discrimination, and criterial behavioral discrimination in a discriminative avoidance task. The neuronal activity for Area 24 is plotted in the form of standard scores normalized with respect to the pre-CS baseline for 40 consecutive intervals following CS onset. Area 29 data are plotted starting 100 ms after tone onset

The role of the anterior cingulate cortex in executive attention is supported by numerous imaging studies that have shown activation in the anterior cingulate cortex during tasks that engender conflict (Posner and DiGirolamo 1998). An example would be a task that requires the selection of a particular response from multiple competing responses. Most of these studies employ a subtraction technique whereby the brain activation found in a neutral or control condition is subtracted from the activation produced by an experimental condition. For example, activation has been found in the anterior cingulate cortex during the generate-uses task, which requires subjects to state uses commonly associated with visually or acoustically presented words (e.g., generating the response 'drive' to the stimulus word 'car'). The subtracted control condition for this task involves merely reading and pronouncing the words. It is argued that executive attention (and thus anterior cingulate cortical activation) is brought into play as a result of the conflict created by the multiple uses that are potentially relevant to a given stimulus. The activation in the anterior cingulate cortex in this task declines as the subjects are repeatedly exposed to the same words and the generation of uses becomes more routine and less dependent on executive control.

Additional evidence for the contribution of the anterior cingulate cortex to executive attention comes from the results of multiple experiments using the Stroop task, a task that requires subjects to name the

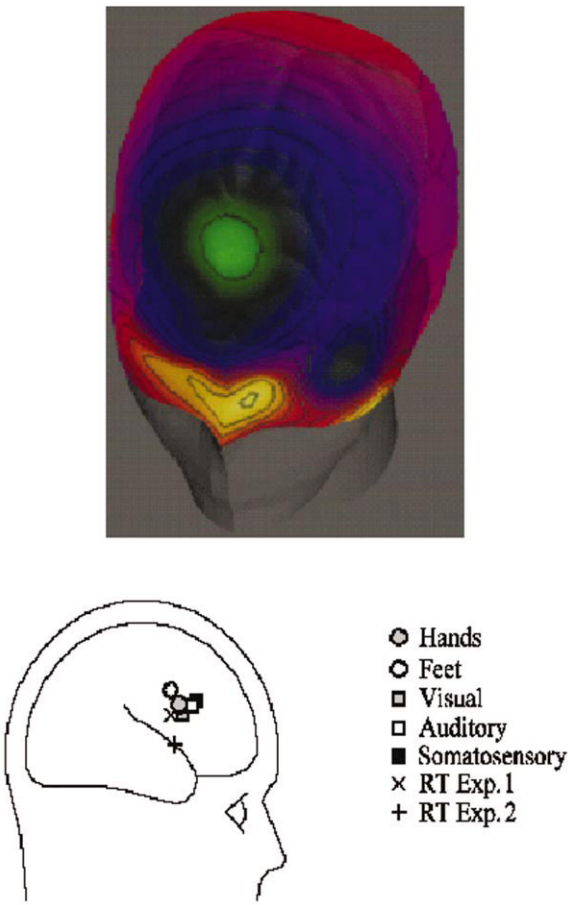

Figure 3
A 3-D and sagittal view of the human brain illustrating the source of error-related negativity (ERN) found after brain electric source analysis (BESA) of event-related brain potentials. The bottom/right figure summarizes the results of several ERN studies that have demonstrated that the source of the ERN is not affected by response modality (subjects responding with their feet or hands) or error feedback modality (visual, auditory, somatosensory). Also shown is the ERN source for two reaction time experiments, one involving a decision of whether a number was 'smaller than/larger than' (RT Exp.1) and another involving a classification of words into semantic categories (RT Exp. 2)

ink color of visually presented words in a congruent condition (e.g., the word 'red' printed in red ink), an incongruent condition (e.g., the word 'green' printed in red ink), or a neutral condition (e.g., the word 'door' printed in red ink) (Posner and DiGirolamo 1998). The anterior cingulate cortex has been found active in both the incongruent and congruent conditions when compared to the neutral condition. It has been suggested that both the congruent and the incongruent conditions involve conflict because subjects must respond to the ink color while inhibiting a response to the word's meaning. Some studies have found more activation in the incongruent condition than in the congruent condition, in line with the expectation that the incongruent condition creates more conflict and thus recruits more executive attention in the anterior cingulate cortex.

Studies employing high-density scalp recordings of EEG have pointed to an involvement of the anterior cingulate cortex in error detection, another aspect of executive attention (Dahaene et al. 1994, Holroyd et al. 1998, Gehring et al. 1993). These experiments have demonstrated a marked electrical negativity at mid frontal regions of the scalp. The negativity peaks about 100 milliseconds after subjects make an incorrect response, such as an erroneous key press in a reaction time task. Brain electrical source analyses (BESA) carried out independently by separate investigators consistently localize the error-related negativity (ERN) to the anterior cingulate cortex. These results support the role of the anterior cingulate cortex in executive attention invoked during error-related processing (see Fig. 3).

3. Movement-related Processing: Response Selection by the Cingulate Cortex

Considering the ample connections of the cingulate cortex with motor and premotor cortex as well as areas of the striatal motor system, it is not surprising that the cingulate cortex has been linked to processes such as response selection. Several studies have documented the existence of a topographic organization of the cingulate cortex with respect to particular response modalities. For example, different areas of the anterior cingulate cortex are active depending on whether subjects perform in tasks involving oculomotor, manual, or spoken responses (Paus et al. 1993).

A case study of patient D.L., who sustained a circumscribed right hemisphere lesion of the anterior cingulate cortex after surgery to remove a tumor, adds further support for cingulate cortical involvement in movement-related processing (Turken and Swick 1999). Interestingly, D.L. exhibited entirely normal performance in Stroop-like and divided attention tasks when responses were orally reported; however, D.L. showed a dramatic deficit in the same tasks when manual responses were required. These results were interpreted as favoring the idea that command signals are sent to motor output areas through the anterior cingulate cortex. The authors characterize the role of the anterior cingulate cortex as confirming the appropriateness of the selected response, thus to facilitate

correct responding while suppressing incorrect responding.

Premotor neuronal activity in cingulate cortex has been demonstrated in several studies of the single-unit correlates of learning and performance. In one such study, approximately half of all single neurons in anterior and posterior cingulate cortex in rabbits exhibited premotor firing ramps that consisted of progressive increases in firing frequency preceding the onset of the behavioral (locomotory) avoidance responses (Kubota et al. 1996). Also, neuronal firing in ventral portions of the anterior cingulate cortex was correlated with the onset of licking behavior during appetititve conditioning of rats (Takenouchi et al. 1999). Thus, cingulate cortical neurons become active preceding the initiation of learned motor responses.

The involvement of the anterior cingulate cortex in response selection does not negate the role of the cingulate cortex in associative and executive attention. Appropriate response selection for a given situation can only occur if attention is devoted to the significant associative stimuli and if conflict among competing motor responses is resolved.

4. *Emotion and the Affective Dimension of Pain*

Traditionally, the cingulate cortex has been viewed as part of a brain circuit that is involved in the experience and expression of emotion (Papez 1937, Maclean 1975). Although more recent evidence has suggested an important role for the cingulate cortex in processes such as attention and response selection, the role of the cingulate cortex in emotion remains unchallenged.

Activations of the anterior cingulate cortex, in particular, have been found to accompany the experience of emotion in numerous neuroimaging studies. For example, when cerebral blood flow (CBF) was measured using PET while subjects viewed emotional film clips and recalled emotional situations, the anterior cingulate cortex was the only structure to exhibit CBF changes correlated with subjects' scores on the Levels of Emotional Awareness Scale (LEAS), a test that measures the capacity to perceive and differentiate complex emotions in oneself and others (Lane et al. 1998). The results suggested that individual differences in emotional awareness could be related to the degree of activation in the anterior cingulate cortex.

Evidence has also been presented in support of a role for the anterior cingulate cortex in mediating the emotional response to pain in humans (Price 2000). Pain is thought to involve two components, a sensory component and an affective component. The affective component reflects the unpleasantness associated with pain and its long-term consequences. The anterior cingulate cortex receives direct input from spinal pain pathways and other input from areas (e.g., the prefrontal cortex) that are involved in cognitive aspects of pain processing, such as evaluating the immediate threat of the pain and its potential interference with daily activities. The anterior cingulate cortex is thus positioned to integrate these two types of pain-related information in order to select appropriate coping responses such as escape or avoidance.

5. *Learning and Memory*

5.1 *Distinct Roles of the Anterior and Posterior Cingulate Cortex*

Compelling evidence suggests an important role for the cingulate cortex in the mediation of learning and memory processes. Available data indicate that the contributions of the anterior and posterior cingulate cortices are functionally distinct. A contribution of the anterior and posterior cingulate cortices to early and late stages of learning, respectively, has been documented in discriminative avoidance learning in rabbits as well as in conditioned visual discrimination in rats (Gabriel 1993, Bussey et al. 1997). In rabbits, discriminative neuronal activity (see Sect. 2.1) in the anterior cingulate cortex develops after fewer training trials than in the posterior cingulate cortex. The observations of neuronal activity coincide nicely with restricted lesion studies showing that lesions confined to the anterior cingulate cortex result in a deficit of behavioral performance in the early stages of learning, whereas lesions confined to the posterior cingulate cortex result in a loss of performance at later stages of learning.

5.2 *Context-specific Retrieval Patterns and Spatial Processing: Posterior Cingulate Cortex and the Anterior Thalamus*

Although the evidence clearly indicates an involvement of the cingulate cortex in the learning-based coding of associatively significant stimuli, further evidence indicates that this coding subserves the retrieval of learned, context-appropriate responses. Evidence in support of this idea is shown by the existence of unique topographical distributions of CS+-related neuronal activity in different cell layers of the posterior cingulate cortex and in various anterior thalamic nuclei in rabbits, during discriminative instrumental learning. Some layers are activated maximally by the CS+ in the initial stages of training, others in intermediate training stages, and others as the rabbits attain asymptotic discriminative performance. The distribution of the activations changed not only across time (the stage in training) but also with respect to the spatial context. For instance, the same set of cues elicited different patterns of activation depending on whether the rabbits were engaged in a

moderately learned discriminative avoidance task or (in a separate training apparatus) a well-learned discriminative approach task. These context-specific patterns of CS+-elicited activity could be associated with the learned responses that are appropriate to a given situation or context. Thus, when a given context specific pattern is elicited on cue presentation, the learned response is retrieved. This mechanism could subserve pattern separation, i.e., the ability to defeat proactive and retroactive interference when multiple similar cues are associated with different memories or responses, as when one tries to recall the names of several recently met individuals. The retrieval hypothesis is consistent with the finding that the brief-latency cue-elicited context-specific patterns in the posterior cingulate cortex are followed by premotor firing, i.e., firing which precedes the onset of the learned behavioral response (see Sect. 2.3).

Additional evidence suggests that the context-specific patterns in the posterior cingulate cortex depend on the integrity of hippocampal connections, which may supply information concerning the operative spatial context to the cingulate cortex. Fornix lesions which disconnect the hippocampal formation from the anterior thalamus disrupt the training-stage-related patterns in the posterior cingulate cortex and these lesions impair concurrent performance in two different discriminative learning tasks that employ very similar cues (Smith et al. 2000).

Posterior cingulate cortical neurons also have functional properties that are similar to those found in neurons of the hippocampus and parietal cortex during spatial processing. For instance, rodent hippocampal neurons are selectively active when subjects occupy a particular location in space, while other neurons code information about directional heading, independently of spatial location or ongoing behavior. Direction-coding neurons have also been documented in the posterior cingulate cortex and related thalamic nuclei, and these neurons, together with those of the hippocampal formation, are thought to contribute to the sense of direction and place in spatial learning situations. Interestingly, in primates, hippocampal neurons are selectively active when the subject is viewing (rather than occupying) a particular space in the environment. The posterior cingulate cortex of primates contains neurons that discharge with eye movement and eye position, a tuning property also found among neurons in the frontal and parietal cortical areas (Olson et al. 1996). The eye direction-coding neurons in primate cingulate cortex are hypothesized to participate in the spatial interpretation of retinal images.

6. Concluding Comment

Studies carried out by behavioral neuroscientists using rats and rabbits as subjects have shown that the cingulate cortex is a critical substrate of learned responses to predictive stimuli. Cingulate cortical neurons in these animals code associatively significant stimuli and exhibit context-specific topographic patterns that could mediate cued retrieval of context-appropriate learned behavior. These functions occur as a result of intimate interactions of the hippocampal and cingulothalamic brain regions. Studies of cognitive neuroscientists concerning brain activation during cognitive task performance in human subjects have yielded results that are fundamentally in agreement with the studies with rats and rabbits. For example, there is clear agreement that the anterior cingulate cortex subserves an attentional role as its neurons are recruited in situations of high cognitive conflict, e.g., when stimuli acquire new meanings at the outset of learning, or when a decision among multiple-response alternatives must be reached. The involvement in behavioral learning and the associative and memory-bearing characteristics of cingulate cortical neuronal activity have also led behavioral neuroscientists to speak of cingulate cortical attention as associative in character, i.e., a learned form of attention. Cognitive neuroscientists have, on the other hand, discussed the cingulate cortex as involved in attentional processes without reference to memory. Given the findings of behavioral neuroscience and the very close neuroanatomical association of the cingulate cortex with other structures (e.g., the hippocampus) that are acknowledged components of the brain's memory system, it is very likely that early in the twenty-first century there will be an even greater convergence of behavioral and cognitive neuroscience upon a common mnemonic interpretation of cingulate cortical function.

See also: Attention, Neural Basis of; Emotion, Neural Basis of; Learning and Memory, Neural Basis of; Neural Representations of Direction (Head Direction Cells); Pain, Neural Basis of

Bibliography

Bussey T J, Muir J L, Everitt B J, Robbins T W 1997 Triple dissociation of anterior cingulate, posterior cingulate, and medial frontal cortices on visual discrimination tasks using a touchscreen testing procedure for the rat. *Behavioral Neuroscience* **111**(5): 920–36

Dahaene S, Posner M I, Tucker D M 1994 Localization of a neural system for error detection and compensation. *Psychological Science* **5**: 303–5

Freeman J H, Cuppernell C, Flannery K, Gabriel M 1996 Limbic thalamic, cingulated cortical and hippocampal neuronal corellates of discriminative approach learning in rabbits. *Behavioral Brain Research* **80**: 123–36

Gabriel M 1993 Discriminative avoidance learning: A model system. In: Vogt B A, Gabriel M (eds.) *Neurobiology of Cingulate Cortex and Limbic Thalamus*. Birkhauser, Toronto, Canada

Gabriel M, Talk A 2001 A tale of two paradigms: Lessions learned from parallel studies of discriminative instrumental learning and classical eyeblink conditioning. In: Steinnetz A,

Gluck M, Solomon P (eds.) *Model Systems and the Neurobiology of Associative Learning: A Festschift in Honor of Richard F. Thompson*. Laurence Erlbaum Associates, NJ
Gabriel M, Taylor C 1998 Prenatal exposure to cocaine impairs neuronal encoding of attention and discriminative learning. In: Harvey J A, Kosofsky B E (eds.) *Cocaine: Effects on the Developing Brain*. New York Academy of Sciences, New York
Gehring W J, Gross B, Coles M G H, Meyer D E, Donchin E 1993 A neural system for error detection and compensation. *Psychological Science*. **4**: 385–90
Goldman-Rakic P S 1988 Topography of cognition: Parallel distributed networks in primate association cortex. *Annual Review of Neuroscience* **11**: 137–56
Holroyd C B, Dien J, Coles M G 1998 Error-related scalp potentials elicited by hand and foot movement: Evidence for an output-independent error-processing system in humans. *Neuroscience Letters* **242**: 65–8
Kubota Y, Wolske M, Poremba A, Gabriel M 1996 Stimulus-related and movement-related single-unit activity in rabbit cingulate cortex and limbic thalamus during performance of discriminative avoidance behavior. *Brain Research* **72**: 22–38
Lane R D, Reiman E M, Axelrod B, Yun L S, Holmes A, Schwartz G E 1998 Neural correlates of levels of emotional awareness: Evidence of an interaction between emotion and attention in the anterior cingulate cortex. *Journal of Cognitive Neuroscience* **10**(4): 525–35
MacLean P D 1975 Sensory and perceptive factors in emotional functions of the triune brain. In: Levi L (ed.) *Emotions: Their Parameters and Measurement*. Raven Press, New York
Olson C R, Musil S Y, Goldberg M E 1996 Single neurons in posterior cingulate cortex of behaving macaque: Eye movement signals. *Journal of Neurophysiology* **76**(5): 3285–300
Papez J W 1937 A proposed mechanism of emotion. *Archives of Neurology and Psychiatry* **38**: 725–43
Paus T, Petrides M, Evans A C, Meyer E 1993 Role of the human anterior cingulate cortex in the control of oculomotor, manual, and speech responses. *Journal of Neurophysiology* **70**: 453–69
Posner M I, DiGirolamo G J 1998 Executive attention: Conflict, target detection, and cognitive control. In: Parasuraman R (ed.) *The Attentive Brain*. MIT Press, Cambridge, MA
Powell D A, Buchanan S L, Gibbs C M 1990 Role of the prefrontal-thalamic axis in classical conditioning. In: Uylings H B M, Van Eden C G, De Bruin J P C, Corner M A, Feenstra M G P (eds.) *The Prefrontal Cortex: Its Structure, Function and Pathology* (vol. 85). Elsevier Science, Amsterdam
Price D D 2000 Psychological and neural mechanisms of the affective dimension of pain. *Science* **288**(5472): 1769–72
Romano A G, Harvey J A 1998 Prenatal cocaine exposure: Long-term deficits in learning and motor performance. In: Harvey J A, Kosofsky B E (eds.) *Cocaine: Effects on the Developing Brain*. New York Academy of Sciences, New York
Segal M, Olds J 1972 Behavior of units in hippocampal circuit of the rat during learning. *Journal of Neurophysiology* **35**(5): 680–90
Smith D M, Patel J, Gabriel M 2000 Hippocampal-cingulo-thalamic interactions supporting concurrent discriminative approach and avoidance learning in rabbits. *Society for Neuroscience Abstracts* **26**: 198
Takenouchi K, Nishijo H, Uwano T, Tamura R, Takigawa M, Ono T 1999 Emotional and behavioral correlates of the anterior cingulate cortex during associative learning in rats. *Neuroscience* **93**(4): 1271–87
Taylor C, Freeman J H, Holt W, Gabriel M 1999 Impairment of cingulothalamic learning-related neuronal coding in rabbits exposed to cocaine *in utero*: General and sex-specific effects. *Behavioral Neuroscience* **113**: 62–77
Turken A U, Swick D 1999 Response selection in the human anterior cingulate cortex. *Nature Neuroscience* **2**(10): 920–4

L. Burhans, A. Talk, and M. Gabriel

Circadian Rhythms

This article reviews the nature of biological rhythms in mammals, how they develop, and their eventual deterioration. It discusses the suprachiasmatic nuclei (SCN), the major pacemaker in the brain, and recent advances in understanding the molecular mechanisms of this unique timing system. It also discusses the adaptive value of having an endogenous clock in the brain and its implications for administering therapeutic agents.

1. *What Are Circadian Rhythms?*

All our behavioral, physiological, and endocrinological functions are controlled by an endogenous clock that measures time in approximately 24-hour intervals. The rhythms the clock generates are known as circadian rhythms (from the Latin, *circa*, about, and *dies*, day). The 'about' is very important. Living systems do not merely respond to cyclic changes in their environment in some phase-locked fashion. Leaves of plants that open during the day and close at night still open and close when the plants are kept in constant darkness. Humans and other animals maintain a period of about 24 hours in bodily functions when they are in environments without time cues. Noise, temperature, and social stimuli can also synchronize, or entrain, the clock, but the strongest entraining signal is the light/dark (LD) cycle.

2. *What Are Their Properties?*

Figure 1 illustrates two important properties of circadian rhythms in the wheel-running activity of a nocturnal mammal, the white-footed mouse (*Peromyscus leucopus*), that has been placed in a 24-hour LD photoperiod consisting of one hour of light and 23 hours of darkness. It takes the mouse about a week for its activity to become entrained by the LD cycle, after which it begins running immediately after the lights go off. It runs for about 10 hours and then becomes inactive. After 60 days the animal is released into constant darkness. There its activity free-runs with a period of 23.6 hours, that is, it begins to run 24 minutes earlier each day.

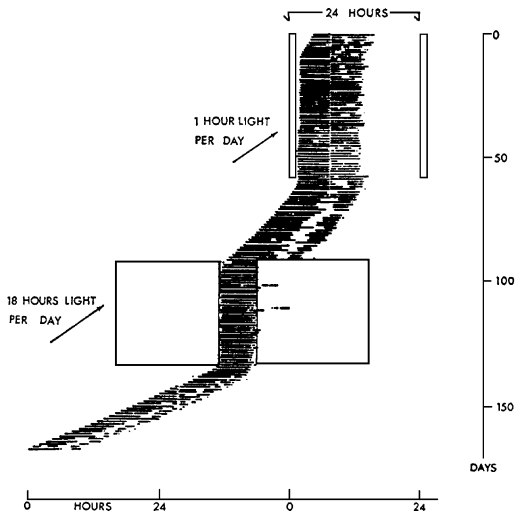

Figure 1
Entrained and freerunning rhythms of wheel-running activity in *Peromyscus leucopus*. The animal uses a running wheel during its nightly activity period which lasts about 10 hours and is recorded as a heavy black lime. Successive days are plotted beneath each other. On the 10th day of the experiment the animal had locked on to (was entrained by) the daily light pulse (one hour duration). On the 60th day, the light was discontinued. The rhythm persisted with a circadian period of 23.6 hours. On the 90th day, it was again entrained by a 24-hour LD cycle in which the light pulse lasted 18 hours. On the 140th day the light was discontinued, and the rhythm again free ran, but with a period of 23.0 hours (from Pittendrigh 1974)

After two weeks in the dark, the mouse is put into a 24-hour LD cycle consisting of 18 hours of light and six hours of darkness. It immediately becomes entrained to the new photoperiod and its activity is tightly confined to the dark period. When it is released back into constant dark, two aspects of its activity are different from the previous stay in the same condition. First, the amount of its activity is lessened. Second, the period of its free-running rhythm is shorter—23 instead of 23.6 hours (Pittendrigh 1974). This indicates that the clock remembers the previous LD cycle and this influences both the amplitude and period of the free-running rhythm. A 'memory' of the preceding daylength is even encoded in the electrophysiological activity of an isolated slice of brain containing the SCN (Mrugala et al. 2000).

Rhythms have different phase relations to the LD cycle. Figure 2 shows entrained rhythms in several physiological parameters in a human subject in a 16:8 LD cycle. Body temperature is high during the day and falls at night, when the subject is asleep. Cortisol is highest at the end of the dark period. Growth hormone

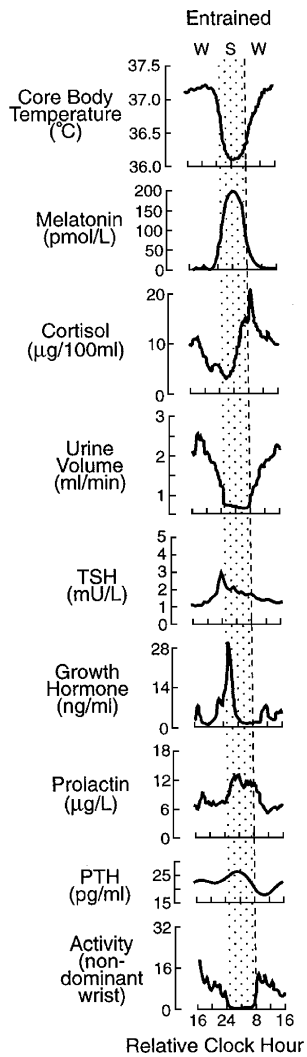

Figure 2
Entrained rhythms of colonic temperature, plasma levels of cortisol, urine volume, and plasma levels of thyroid stimulating hormone, growth hormone, prolactin and parathyroid hormone, and activity in a human monitored in a 16:8 LD cycle. W = waking; S = sleeping (from Czeisler and Khalsa 2000)

is low during the day and shows a sharp peak at the beginning of the dark period, during sleep. Immune system components also have circadian frequencies, and this can be very important in determining both responses to antigens and the timing of treatments such as chemotherapy in cancer. Ideally, one would want to administer a therapeutic agent at a time when it would produce the fewest unwanted side effects and yet be most effective in attacking the dividing cell.

3. Why Do We Need Circadian Rhythms?

There is a distinct advantage in timing various behaviors and metabolic processes to the appropriate time of day. Organisms have to sleep, and they generally do so either in the day or at night. One would want potential mates and food to be available when one is active. But a simple hourglass mechanism, timing precise 24-hour periods, would be inadequate because of seasonal changes in day length. A circadian clock, sensitive to external conditions, can be reset each day, making the clock conform to the environment. Furthermore, an internal clock allows organisms not only to respond to changes in the environment but also to anticipate them. Lizards gain heat by basking in the sun. When they are in their burrows, still with low body temperatures, they crawl to the opening of the burrow and stick their heads above the desert floor before the sun comes up. This way they are in a position to gain enough heat to become active as soon as possible. The early bird *does* catch the worm, and its internal clock wakes it up before dawn. The body temperature of humans falls in the evening, while we are still awake, and begins to rise in the early morning, while we are still asleep. Even in the same entrained environment, people have different phase relations with the LD cycle; some are larks and some are owls.

4. What Controls Circadian Rhythms?

Almost all known plants and animals exhibit circadian rhythms. For instance, the microscopic single-celled aquatic plant, *Gonyaulax polyedra*, is phosphorescent, lighting up at night, and dimming in the day. In 1958, Hastings and Sweeney demonstrated that *Gonyaulax* showed peaks and troughs of luminescence even in constant darkness, but the peak time of luminescence shifted a little later each day. If the plant was exposed to brief pulses of light, the peak of luminescence could be shifted to almost any time of day, depending on when the light pulse was given. Thus, light reset the *Gonyaulax* clock.

In 1972, Moore and Lenn placed a radioactive label into the eyes of rats and found that there was a direct pathway from the retina to two tiny nuclei in the hypothalamus. These nuclei lie behind the eyes right above the location in the brain where fibers from the left and right retinas cross, the optic chiasm, and therefore the nuclei are named the SCN. In that same year, Stephan and Zucker showed that lesions of the SCN could permanently eliminate or weaken circadian patterns of behavior (for a review, see Rusak and Zucker 1979). In a series of papers beginning in 1987, Lehman and his colleagues demonstrated that locomotor activity rhythmicity can be restored in hamsters with SCN lesions by transplants of fetal SCN tissue. There is now no doubt that, for most rhythms, the SCN is the main clock in the mammalian brain. Each morning, light from the eye sends electrical signals to the SCN and resets it. The SCN, in turn, synchronizes the rest of the brain and sets the pace for all daily activity patterns, just as a conductor synchronizes all the instruments in an orchestra.

The analogy of the SCN to an orchestra conductor is inadequate, however, because the cells of the SCN do not act as a single multioscillator unit. Welsh et al. (1995) demonstrated that SCN cells grown in culture oscillate at different rates: this means that each single SCN neuron functions as an independent circadian clock.

5. Are Clocks Similar Throughout the Animal Kingdom?

Clocks are being dissected rapidly through genetic analysis. Clock genes have been identified in cyanobacteria (blue–green algae), the fungus *Neurospora*, and the fruit fly *Drosophila*. In 1988, Ralph and Menaker (1988) identified a *tau* mutation in Syrian hamsters. Hamsters homozygous for the mutation had free-running activity rhythms of about 20 hours, compared to the wild type hamster's rhythm of about 24 hours. In 1994, Takahashi's laboratory discovered the first mouse circadian mutant, *Clock*, which was arrhythmic (Vitaterna et al. 1994). The *Drosophila* clock is proving highly amenable to molecular analysis and there are close connections to the mouse mutation. Many clock genes are highly conserved: some of the same molecules are present in fruit flies and mice, and it is possible that common molecular clock mechanisms from bacteria to humans will be found in the not-too-distant future (see Young 2000, for a review).

6. How Do Circadian Rhythms Develop and Age?

Using 14^{C}-labeled deoxyglucose to monitor metabolic activity in the SCN, Reppert and Schwartz (1986) found that there was a distinct day–night oscillation of metabolic activity in both rat and primate SCN. But fetuses do not process light. They found in rats that the maternal circadian system coordinated the timing of the fetal rhythms. It takes about a month after birth for strong circadian rhythms to develop. In most, but not all individuals, the robustness of the rhythms declines with age. Thus, rats that normally sleep mostly in the light, and eat and drink mainly in the dark, distribute all these activities about equally in the light and the dark when they get old. The mean amount of sleep, food eaten, and water drunk, etc., stays the same: only the pattern changes. These changes are correlated with aberrant SCN firing patterns in SCN slices taken from old rats. This

implies that aging could either disrupt coupling between SCN pacemaker cells or their output, or cause deterioration of the pacemaking properties of the individual cells (Satinoff et al. 1993). A better understanding of clock mechanisms should lead to more efficacious clinical treatments for alleviating disorders in the elderly, such as sleep disturbances, that are associated with changes in the clock itself. But aside from pathologies in aging, a more complete understanding of clocks and how they are entrained by light will be useful in helping to ameliorate the discomfort and disability that comes from conditions as disparate as jet lag, sleep disorders in shift workers and blind people, and depression and manic-depressive disorder.

See also: Childhood Health; Sleep and Health; Sleep Disorders: Psychiatric Aspects; Sleep Disorders: Psychological Aspects; Sleep: Neural Systems; Suprachiasmatic Nucleus

Bibliography

Czeisler C, Khalsa S 2000 The human circadian timing system and sleep-wake regulation. In: Kryger M, Roth T, Dement W (eds.) *Principles and Practice of Sleep Medicine*. W.B. Saunders, Philadelphia, PA

Hastings J W, Sweeney B M 1958 A persistent diurnal rhythm of luminescence in *Gonyaulax polyedra*. *Biological Bulletin* **115**: 440–58

Lehman M, Silver R, Gladstone W, Kahn R, Gibson M, Bittman E 1987 Circadian rhythmicity restored by neural transplant. Immunocytochemical characterization of the graft and its integration with the host brain. *Journal of Neuroscience* **7**: 1626–38

Moore R Y, Lenn N J 1972 A retinohypothalamic projection in the rat. *Journal of Comparative Neurology* **146**: 1–14

Mrugala M, Zlomanczuk P, Jagota A, Schwartz W 2000 Rhythmic multiunit neural activity in slices of hamster suprachiasmatic nucleus reflect prior photoperiod. *American Journal of Physiology-Regulatory Integrative and Comparative Physiology* **278**: R987–94

Pittendrigh C 1974 Circadian oscillations in cells and the circadian organization of multicellular systems. In: Schmitt F, Worden F (eds.) *The Neurosciences: Third Study Program*. MIT Press, Cambridge, MA

Ralph M, Menaker M 1988 A mutation of the circadian system in golden hamsters. *Science* **241**: 1225–7

Reppert S, Schwartz W 1986 Maternal suprachiasmatic nuclei are necessary for maternal coordination of the developing circadian system. *Journal of Neuroscience* **9**: 2724–29

Rusak B, Zucker I 1979 Neural regulation of circadian rhythms. *Physiological Reviews* **59**: 449–526

Satinoff E, Li H, Liu C, McArthur A, Medanic M, Tcheng T, Gillette M 1993 Do the suprachiasmatic nuclei oscillate in old rats as they do in young ones? *American Journal of Physiology-Regulatory Integrative and Comparative Physiology* **265**: R1216–22

Vitaterna M H, King D P, Chang A M, Kornhauser J M, Lowrey P L, McDonald J D, Dove W F, Pinto L H, Turek F W, Takahashi J S 1994 *Science* **264**: 719–25

Welsh D, Logothetis D, Meister M, Reppert S 1995 Individual neurons dissociated from rat suprachiasmatic nucleus express independently phased circadian firing rhythms. *Neuron* **14**: 697–706

Young M 2000 Circadian rhythms. Marking time for a kingdom. *Science* **288**: 451–3

E. Satinoff

Cities: Capital, Global, and World

1. The Global City: Introducing a Concept and its History

Each phase in the long history of the world economy raises specific questions about the particular conditions that make it possible. One of the key properties of the current phase is the ascendance of information technologies and the associated increase in the mobility and liquidity of capital. There have long been cross-border economic processes—flows of capital, labor, goods, raw materials. Over the last century, these took place largely within the interstate system, where the key articulators were national states and colonial empires. The international economic system was ensconced largely in this interstate system. This has changed rather dramatically during the 1990s as a result of privatization, deregulation, the opening up of national economies to foreign firms, and the growing participation of national economic actors in global markets.

It is in this context that we see a rescaling of what are the strategic territories that articulate the new system. With the partial unbundling or at least weakening of the national as a spatial unit due to privatization and deregulation and the associated strengthening of globalization, come conditions for the ascendance of other spatial units or scales. Among these are the subnational, notably cities and regions; cross-border regions encompassing two or more subnational entities; and supranational entities, such as global electronic markets and free-trade blocs. The dynamics and processes that get terrritorialized at these diverse scales can in principle be regional, national, or global.

We can locate the emergence of global cities in this context and against this range of instantiations of strategic scales and spatial units. In the case of global cities, the dynamics and processes that get territorialized are global. This article examines first some of the key theoretical and empirical elements of the global city model, followed by a brief history of the evolution of the literature on cities in the global economy. Section three is a more in-depth discussion of the organizing hypotheses of the global city model.

Sections four and five discuss two specific features: the question of place in a global economy and the question of city-to-city networks in domains other than the economic.

2. Elements in a New Conceptual Architecture

The globalization of economic activity entails a new type of organizational structure. To capture this theoretically and empirically requires, correspondingly, a new type of conceptual architecture. Constructs such as the global city and the global-city region are, in my reading, important elements in this new conceptual architecture. Arrighi's (1994) analysis is of interest here in that it posits the recurrence of certain organizational patterns in different phases of the capitalist world economy, but at gradually higher orders of complexity and expanded scope, and timed to follow or precede particular configurations of the world economy.

There are today several closely linked terms used to capture a particular intersection between global processes and cities (see Stren 1996 and Savitch 1996 for overviews). The most common is world cities, a term attributed to Goethe and then relaunched in the work of Peter Hall (1966) and more recently respecified by John Friedmann (Friedmann and Goetz 1982). Other related terms are 'supervilles' (Braudel 1984), informational city (Castells 1989). Choosing how to name a configuration has its own substantive rationality. The decision to formulate the term 'global city' (Sassen 1984) stemmed out of a recognition of the specificity of the current period. The term world city has precisely the opposite attribute: it refers to a type of city which we have seen over the centuries (Braudel 1984, Hall 1966, King 1990), and most probably also in much earlier periods in Asia than in the West. In this regard it could be said that most of today's major global cities are also world cities, but that there may well be some global cities that are not world cities in the full, rich sense of that term. This is partly an empirical question; further, as the global economy expands and incorporates additional cities into various cross-border networks, it is quite possible that the answer to that particular question will vary over time. Thus the fact that Miami has developed global city functions beginning in the late 1980s does not make it a world city in that older sense of the term (Nijman 1996). (See generally Abu-Lughod 1999, Short and Kim 1999.)

3. The Elements of a New Theoretical Framework

By the early 1980s a number of scholars had begun to study cities in the context of globalization (Walton 1982, Ross and Trachte 1983, Sassen 1982, Rodriguez and Feagin 1986). But it is one article in particular, 'The World City Hypothesis' by Friedmann and Goetz (1982) on which attention centered. This article took a variety of elements that were emerging in the research literature on cities, on the global economy, on immigration, and a number of other subjects, and sought to formalize these into several propositions about the role of cities in the global economy. The key elements in this framework were the emergence of several cities as basing points for global capital, a hierarchy (albeit a shifting one) of such cities, and the social and political consequences for these cities of being such basing points.

With the books by Castells (1989), King (1990), and Sassen (1991/2001), what had been a hypothesis in the early 1980s became a full-fledged theorization and empirical specification. These three books add important and distinct propositions to the general framework: Castells' proposition that globalization as constituted today has engendered a space of flows that reconfigures economic and political power; King's enlargement of the frame of reference to show that the highest levels of internationalization had taken place in the cities of colonial empires rather than in the center of the world economy; Sassen's proposition that it is not simply a matter of global coordination but one of the production of global control capacities and that an examination along these lines allows us to understand the role of global cities as production sites.

It is important to distinguish what is different about this literature from a broader, earlier literature on world cities prominently represented by the work of Peter Hall already in the 1960s, and a new literature on megacities especially focused on Latin America and Asia. These literatures do not have the fact of globalization and the centrality of crossborder networks connecting cities as crucial variables. The earlier literature on world cities is closer to the notion of capitals of empires: one city at the top of the power hierarchy.

In the current literature on global cities the determining factor is a cross-border, global network of cities that function as strategic sites for the management and specialized servicing of global economic operations. There is no such entity as a single global city, as there is with the capital of an empire; by definition, the global city is part of a network of cities. Similarly, an older literature focused on past world cities, as in the work of Braudel (1984), and earlier studies of major centers of world commerce and banking, as well as more recent work on urban hierarchies in the world system (Chase-Dunn 1984), are to be differentiated from the current literature if we historicize the world economy and specify what is distinct today. Finally, we need to distinguish between a narrowly specified literature on global and world cities today and various literatures that directly or indirectly contribute to our under-

standing of these cities, notably the research on producer services.

By the mid-1990s the subject had clearly emerged as a rather large field for research among scholars in many different disciplines and countries. We can see this in the variety of authors and themes in several state-of-the-art collections, notably by Fainstein et al. (1993), Knox and Taylor (1995), Noller et al. (1995), Lo and Yeung (1996), and several others that elaborate, critique, expand the empirical base, and generally advance this theoretical and methodological project. We can also see it in several new important books that set the stage for highly focused research on particular variables, notably Meyer (1991), Thrift and Leyshon (1994), Keil (1993), Eade (1996) among others. We also see the creation of several book series by various publishers in different countries: the series on World Cities edited by Knox for Belhaven Press, the series edited by Milton Santos and his colleagues in Sao Paulo for Hucitec, the series edited by Martin Wentz for Campus Verlag, are just some.

It is not only the growth of the research literature but also the growth of a body of critical responses and analyses that signals the strength and vigor of this field of inquiry. There is only space here for the briefest mention, a sort of guide to criticisms: Logan and Swanstrom's (1990) critique of the excessive weight given to global structural processes in comparing internal vs. external factors that shape a city's economic development; Hammet's (1996) critique of Sassen's proposition that globalization has contributed to social and economic polarization in global cities; Markusen and Gwiasda's (1994) critique of the notion that New York is at the top of the US urban hierarchy and how a comparison with Washington shows that the latter has a higher level of specialization than New York in many advanced specialized services, notably in legal services; critiques of the literature for its neglect of grassroots transnationalism and the new kinds of politics and identitiy formation that this entails; Beauregard's (1991) critique of the explanatory variables for changes in the built environment and the real estate industry; Simon's (1995) critique of the neglect of the periphery, notably Africa; the debate in *Urban Affairs* (March 1998) on the concept of the global city and a similar one in *Urban Studies* (Summer 2001); the special issue on 'Segregations Urbaines' of *Societes Contemporaines* in 1995; the special issue of *Urban Studies* in 1996, and many others.

There are two types of scholarly literature that intersect with this body of research on cities and the global economy, and indeed often invoke or use it to develop their arguments. They are on the one hand a literature of anthropological and cultural studies on transnationality, globalization, and identity formation (Holston 1996, Low 1999). The other is the scholarship by regional economic geographers on the global economy, who have also focused on cities (e.g., Moulaert and Scott 1997, Gravestijn et al. 1999). In the last few years there has been a new interest in this subject by geographers (Veltz 1996, Scott 2001, Storper 1997).

In terms of method, a number of strategies have been developed. Even where there are data on intercity flows, it will take a lot of work to constitute the requisite data sets. In this regard, an ambitious initiative by the National Academy of Sciences of the US examines how we can construct better data sets at the scale of the city (NAS 2000).

Among the quantitative methodological and data formation strategies are the efforts by Smith and Timberlake (2002) and by Taylor et al. (2002). Smith and Timberlake (2002) conceptualize urban areas as central nodes in multiplex networks of economic, social, demographic and information flows. They use the methodological logic of network analysis, using particularly two measures: one of these is structural or relational equivalence between actors (i.e., cities) in a network; the second measure is centrality. Both of these measures relate to a number of propositions developed in the literature on cities in the global economy.

Taylor et al. (2002) have developed a new and pioneering data set that makes it possible to map the global networks of offices of the leading firms worldwide in several specialized corporate services, such as accounting, law, advertising, and finance. These networks of offices can be used to classify cities in terms of their participation in cross-border networks. The data can be analyzed using a variety of hypotheses and statistical as well as other methods.

There are several other efforts, but space dictates singling out just a few. David Meyer (1991) has developed ways of analyzing international networks through which a variety of exchanges of capital take place. Castells (1989) and Sassen (1991/2001) have developed several techniques of analysis which range from methods to understand the place of cities in global markets to expanding the representation of the global. In *The Informational City* and *The Global City* the authors sought to establish rather broadly what is the array of data sets that can be brought into an analysis of this subject—from international flows of capital and information to very localized social effects. This was an effort to resist the simplification in mainstream accounts which emphasize the global dispersal of activities and telecommunications and exclude most social issues.

Techniques for data analysis traditionally used by economic geographers can also be helpful. For instance Wheeler's (1986) examination of the dispersion of higher-order financial services throughout the US urban hierarchy—which he found had proceeded at a much slower rate than the dispersion of headquarters of other large corporations—can also be used for cross-border hierarchies. Wheeler found that corporations tend to proceed up the urban hierarchy for their advanced service and banking needs.

Elliott (1999) developed a test for the socioeconomic polarization hypothesis in global cities.

4. *The Global City Model: Organizing Hypotheses*

There are seven core hypotheses that organize the data and the theorization of the global city model. There follows a brief discussion of each as a way of producing a more precise representation of the model. (See Sassen 1991/2001.)

First, the geographic dispersal of economic activities that marks globalization, along with the simultaneous integration of such geographically dispersed activities, is a key factor feeding the growth and importance of central corporate functions. The more dispersed a firm's operations across different countries, the more complex and strategic its central functions—that is, the work of managing, coordinating, servicing, financing a firm's network of operations.

Second, these central functions become so complex that increasingly the headquarters of large global firms outsource them: they buy a share of their central functions from highly specialized service firms: accounting, legal, public relations, programming, telecommunications, and other such services. Thus while even ten years ago the key site for the production of these central headquarter functions was the headquarters of a firm, today there is a second key site: the specialized service firms contracted by headquarters to produce some of these central functions or components of them. This is especially the case with firms involved in global markets and non-routine operations. But increasingly the headquarters of all large firms are buying more of such inputs rather than producing them in-house.

Third, those specialized service firms engaged in the most complex and globalized markets are subject to agglomeration economies. The complexity of the services they need to produce, the uncertainty of the markets they are involved with either directly or through the headquarters for which they are producing the services, and the growing importance of speed in all these transactions, is a mix of conditions that constitutes a new agglomeration dynamic. The mix of firms, talents, expertise from a broad range of specialized fields makes a certain type of urban environment function as an information center. Being in a city becomes synonymous with being in an extremely intense and dense information loop. This is a type of information loop that as of now still cannot be replicated fully in electronic space, and has as one of its value-added features the fact of unforeseen and unplanned mixes of information, expertise, and talent; these can produce a higher order of information. This does not hold for routinized activities which are not as subject to uncertainty and non-standardized forms of complexity. Global cities are, in this regard, production sites for the leading information industries of our time.

A fourth hypothesis, derived from the preceding one, is that the more headquarters outsource their most complex, unstandardized functions particularly those subject to uncertain and changing markets and to speed, the freer they are to opt for any location because the more the work actually done in the headquarters is not subject to agglomeration economies. This further underlines that the key sector specifying the distinctive production advantages of global cities is the highly specialized and networked services sector. In developing this hypothesis I was responding to a very common notion that the number of headquarters is what specifies a global city. Empirically it may still be the case in many countries that the leading business center is also the leading concentration of headquarters, but this may well be because there is an absence of alternative locational options. But in countries with a well developed infrastructure outside the leading business center, there are likely to be multiple locational options for such headquarters.

Fifth, these specialized service firms need to provide a global service which has meant a global network of affiliates or some other form of partnership, and as a result we have seen a strengthening of cross-border city-to-city transactions and networks. At the limit this may well be the beginning of the formation of transnational urban systems. The growth of global markets for finance and specialized services, the need for transnational servicing networks due to sharp increases in international investment, the reduced role of the government in the regulation of international economic activity, and the corresponding ascendance of other institutional arenas, notably global markets and corporate headquarters—all these point to the existence of a series of transnational networks of cities. One implication of this, and a related hypothesis for research is that the economic fortunes of these cities become increasingly disconnected from their broader hinterlands or even their national economies. We can see here the formation, at least incipient, of transnational urban systems. To a large extent it seems to me that the major business centers in the world today draw their importance from these transnational networks.

A sixth hypothesis, is that the growing numbers of high level professionals and high-profit making specialized service firms have the effect of raising the degree of spatial and socioeconomic inequality evident in these cities. The strategic role of these specialized services as inputs raises the value of top level professionals and their numbers. Further, the fact that talent can matter enormously for the quality of these strategic outputs and, given the importance of speed, proven talent is an added value, the structure of rewards is likely to experience rapid increases. Types

of activities and of workers lacking these attributes, notably in manufacturing and industrial services, are likely to get caught in the opposite cycle.

A seventh hypothesis, is that one result of the dynamics decribed in hypothesis six, is the growing informalization of a range of economic activities which find their effective demand in these cities yet have profit rates that do not allow them to compete for various resources with the high-profit making firms at the top of the system. Informalizing part or all production and distribution activities, including of services, is one way of surviving under these conditions.

The first four hypotheses qualify what has emerged as a dominant discourse on globalization, technology, and cities, which posits the end of cities as important economic units or scales. There is a tendency in that account to take the existence of a global economic system as a given, a function of the power of transnational corporations and global communications. According to the global city model, the capabilities for global operation, coordination, and control contained in the new information technologies and in the power of transnational corporations need to be produced. By focusing on the production of these capabilities we add a neglected dimension to the familiar issue of the power of large corporations and the capacity of the new technologies to neutralize distance and place. A focus on the production of these capabilities shifts the emphasis to the practices that constitute what we call economic globalization and global control.

A focus on practices draws the categories of place and work process into the analysis of economic globalization. These are two categories easily overlooked in accounts centered on the hypermobility of capital and the power of transnationals. Developing categories such as place and work process does not negate the centrality of hypermobility and power. Rather, it brings to the fore the fact that many of the resources necessary for global economic activities are not hypermobile and are, indeed, deeply embedded in place, notably places such as global cities, global-city regions, and export processing zones.

This entails a whole infrastructure of activities, firms, and jobs necessary to run the advanced corporate economy. These industries are typically conceptualized in terms of the hypermobility of their outputs and the high levels of expertise of their professionals rather than in terms of the production or work process involved and the requisite infrastructure of facilities and non-expert jobs that are also part of these industries. This in turn brings with it an emphasis on economic and spatial polarization because of the disproportionate concentration of very high and very low income jobs in the city compared with what would be the case at a larger scale such as the region or the country. A focus on regions, in contrast will lead to an emphasis on broad urbanization patterns, a more encompassing economic base, more middle sectors of both households and firms. Emphasizing place, infrastructure, and non-expert jobs matters precisely because so much of the focus has been on the neutralization of geography and place made possible by the new technologies.

Dealing with place brings with it the problem of boundaries. These are at least of two sorts, the boundary of the territorial scale as such and the boundary of the spread of globalization in the organizational structure of industries, institutional orders, places, and so on. In the case of the global city it is possible to opt for an analytic strategy that emphasizes core dynamics rather than the unit of the city as a container—the latter being one that requires territorial boundary specification. Emphasizing core dynamics and their spatialization (in both actual and digital space) does not completely solve the boundary problem, but it does allow for a fairly clear trade-off between emphasizing the core or center of these dynamics and their spread institutionally and spatially.

Finally, the detailed examination of three particular cities (Sassen 1991/2001) brought to the fore the extent to which these cities collaborate through their very specific advantages rather than simply competing with each other. In focusing on global finance in the 1980s and 1990s it becomes clear that the growth of the major centers was partly derived from the growing network of financial centers. In looking at the broader network it also becomes clear to what extent it was and remains characterized by a pronounced hierarchy among the growing number of centers that make up the network.

The growth of networked cross-border dynamics among global cities includes a broad range of domains—political, cultural, social, criminal. There are cross-border transactions among immigrant communities and communities of origin and a greater intensity in the use of these networks once they become established, including for economic activities that had been unlikely until now. We also see greater cross-border networks for cultural purposes, as in the growth of international markets for art and a transnational class of art curators; and for non-formal political purposes, as in the growth of transnational networks of activists around environmental causes, human rights, and so on. These are largely city-to-city cross-border networks, or, at least, it appears at this time to be simpler to capture the existence and modalities of these networks at the city level. The same can be said for the new cross-border criminal networks.

In brief, recapturing the geography of places represented by the network of global cities allows us to recapture people, workers, communities, and more specifically, the many different work cultures, besides the corporate culture, involved in the work of globalization. It also brings with it an enormous research agenda, one that goes beyond the by now familiar

focus on cross-border flows of goods, capital and information.

Further, by emphasizing the fact that global processes are at least partly embedded in national territories, such a focus introduces new variables in current conceptions about economic globalization and the shrinking regulatory role of the state. That is to say, the space economy for major new transnational economic processes diverges in significant ways from the duality global/national presupposed in much analysis of the global economy. The duality national vs. global suggests two mutually exclusive spaces—where one begins the other ends. One of the outcomes of a global city analysis is that it makes evident that the global materializes by necessity in specific places and institutional arrangements a good number of which, if not most, are located in national territories.

The two final sections examine two particular aspects that illustrate some of these issues concerning place in a global economy and city-to-city networks in realms other than the economic.

5. *New Forms of Centrality*

Several of the organizing hypotheses in the global city model concern the conditions for the continuity of centrality in advanced economic systems in the face of major new organizational forms and technologies that maximize the possibility for geographic dispersal. Historically, centrality has largely been embedded in the central city. Have the new technologies and organizational forms altered the spatial correlates of centrality?

Today there is no longer a simple straightforward relation between centrality and such geographic entities as the downtown, or the central business district. In the past, and up to quite recently in fact, the center was synonymous with the downtown or the CBD. The spatial correlate of the center can assume several geographic forms. It can be the CBD, as it still is largely in New York City, or it can extend into a metropolitan area in the form of a grid of nodes of intense business activity, as we see for instance in Frankfurt and Zurich. The center has been profoundly altered by telecommunications and the growth of a global economy, both inextricably linked; they have contributed to a new geography of centrality (and marginality). Simplifying we can identify four forms assumed by centrality today.

First, while centrality can assume multiple spatial correlates, the CBD in major international business centers remains a strategic site for the leading industries. But it is one profoundly reconfigured by technological and economic change (Graham and Marvin 1996). Further, there are often sharp differences in the patterns assumed by this reconfiguring of the central city in different parts of the world, notably the United States and Western Europe (Veltz 1996, Kunzmann 1994).

In the United States, major cities such as New York and Chicago have large centers that have been rebuilt many times, given the brutal neglect suffered by much urban infrastructure and the imposed obsolescence so characteristic of US cities. This neglect and accelerated obsolescence produce vast spaces for rebuilding the center according to the requirements of whatever regime of urban accumulation or pattern of spatial organization of the urban economy prevails at a given time. In Europe, urban centers are far more protected and they rarely contain significant stretches of abandoned space; the expansion of workplaces and the need for intelligent buildings necessarily will have to take place partly outside the old centers. One of the most extreme cases is the complex of La Defense, the massive, state-of-the-art office complex developed right outside Paris to avoid harming the built environment inside the city. This is an explicit instance of government policy and planning aimed at addressing the growing demand for central office space of prime quality. Yet another variant of this expansion of the 'center' onto hitherto peripheral land can be seen in London's Docklands. Similar projects for recentralizing peripheral areas were launched in several major cities in Europe, North America, and Japan during the 1980s. (See Marcuse and van Kempen 2000.)

Second, the center can extend into a metropolitan area in the form of a grid of nodes of intense business activity. One might ask whether a spatial organization characterized by dense strategic nodes spread over a broader region does in fact constitute a new form of organizing the territory of the 'center,' rather than, as in the more conventional view, an instance of suburbanization or geographic dispersal. Insofar as these various nodes are articulated through digital networks, they represent a new geographic correlate of the most advanced type of 'center.' This is a partly deterritorialized space of centrality. Indeed much of the actual geographic territory within which these nodes exist falls outside the new grid of digital networks, and is in that sense partly peripheralized.

This regional grid of nodes represents a reconstitution of the concept of region. Far from neutralizing geography the regional grid is likely to be embedded in conventional forms of communication infrastructure, notably rapid rail and highways connecting to airports. Ironically perhaps, conventional infrastructure is likely to maximize the economic benefits derived from telematics. This is an important issue that has been lost somewhat in discussions about the neutralization of geography through telecommunications.

Third, we are seeing the formation of a transterritorial 'center' constituted, partly in digital space, via intense economic transactions in the network of global cities. These networks of major international business centers constitute new geographies of centrality. The

most powerful of these new geographies of centrality at the global level binds the major international financial and business centers: New York, London, Tokyo, Paris, Frankfurt, Zurich, Amsterdam, Los Angeles, Sydney, Hong Kong, among others. But this geography now also includes cities such as Bangkok, Seoul, Taipei, Sao Paulo, Mexico City. The intensity of transactions among these cities, particularly through the financial markets, trade in services, and investment, has increased sharply, and so have the orders of magnitude involved. At the same time, there has been a sharpening inequality in the concentration of strategic resources and activities between each of these cities and others in the same country, a condition that further underlines the extent to which this is a cross-border space of centrality.

The pronounced orientation to the world markets evident in such cities raises questions about the articulation with their nation-states, their regions, and the larger economic and social structure in such cities. Cities have typically been deeply embedded in the economies of their region, indeed often reflecting the characteristics of the latter; and they still do. But cities that are strategic sites in the global economy tend, in part, to disconnect from their region. This conflicts with a key proposition in traditional scholarship about urban systems, namely, that these systems promote the territorial integration of regional and national economies.

In the case of a complex landscape such as Europe's we see several geographies of centrality, one global, others continental and regional. A central urban hierarchy connects major cities, many of which in turn play central roles in the wider global system of cities: Paris, London, Frankfurt, Amsterdam, Zurich. These cities are also part of a wider network of European financial/cultural/service capitals, some with only one, others with several of these functions, which articulate the European region and are somewhat less oriented to the global economy than Paris, Frankfurt, or London. And then there are several geographies of marginality: the East–West divide and the North–South divide across Europe as well as newer divisions. In Eastern Europe, certain cities and regions, notably Budapest, are rather attractive for purposes of investment, both European and non-European, while others will increasingly fall behind, notably in Rumania, Yugoslavia, and Albania. We see a similar differentiation in the south of Europe: Madrid, Barcelona, and Milan are gaining in the new European hierarchy; Naples, Rome, and Marseilles are not quite.

Fourth, new forms of centrality are being constituted in electronically generated spaces. For instance, strategic components of the financial industry operate in such spaces. The relation between digital and actual space is complex and varies among different types of economic sectors. But it is increasingly becoming evident that the highly complex configurations for economic activity located in digital space contain points of coordination and centralization.

6. *The Global City as a Nexus For New Politico-cultural Alignments*

The incorporation of cities into a new cross-border geography of centrality also signals the emergence of a parallel political geography. Major cities have emerged as strategic sites not only for global capital, but also for the transnationalization of labor and the formation of translocal communities and identities (Smith 1997). In this regard cities are sites for new types of political operations and for a whole range of new 'cultural' and subjective operations (Watson and Bridges 1999, Allen et al. 1999). The centrality of place in a context of global processes makes possible a transnational economic and political opening for the formation of new claims and hence for the constitution of entitlements, notably rights to place. At the limit, this could be an opening for new forms of 'citizenship' (Isin 2000, Holston 1996).

The emphasis on the transnational and hypermobile character of capital has contributed to a sense of powerlessness among local actors, a sense of the futility of resistance. But an analysis that emphasizes place suggests that the new global grid of strategic sites is a terrain for politics and engagement. The loss of power at the national level produces the possibility for new forms of power and politics at the subnational level. Further, insofar as the national as container of social process and power is cracked (Brenner 1998, Taylor 1995) it opens up possibilities for a geography of politics that links subnational spaces across borders. Cities are foremost in this new geography. One question this engenders is how and whether we are seeing the formation of a new type of transnational politics that localizes in these cities.

Immigration, for instance, is one major process through which a new transnational political economy and translocal household strategies are being constituted (Portes 1997, Skeldon 1997). It is one largely embedded in major cities insofar as most immigrants, certainly in the developed world, whether in the US, Japan, or Western Europe, are concentrated in major cities. It is, in many regards, one of the constitutive processes of globalization today, even though not recognized or represented as such in mainstream accounts of the global economy. This configuration contains unifying capacities across national boundaries and sharpening conflicts within cities. Global capital and the new immigrant workforce are two major instances of transnationalized actors that have unifying properties across borders, and thus internally to each, and find themselves in contestation with each other inside cities. Researching and theorizing these issues will require approaches that diverge from the more traditional studies of political elites, local party

politics, neighborhood associations, immigrant communities, and so on through which the political landscape of cities and metropolitan regions has been conceptualized in urban studies.

One way of thinking about the political implications of this strategic transnational space anchored in global cities is in terms of the formation of new claims on that space. The city has indeed emerged as a site for new claims: by global capital which uses the city as an 'organizational commodity,' but also by disadvantaged sectors of the urban population, frequently as internationalized a presence in large cities as capital. The 'de-nationalizing' of urban space and the formation of new claims by transnational actors, raise the question 'Whose city is it?'

This is a space that is both place-centered in that it is embedded in particular and strategic locations; and it is transterritorial because it connects sites that are not geographically proximate yet are intensely connected to each other. If we consider that large cities concentrate both the leading sectors of global capital and a growing share of disadvantaged populations—immigrants, many of the disadvantaged women, people of color generally, and, in the megacities of developing countries, masses of shanty dwellers—then we can see that cities have become a strategic terrain for a whole series of conflicts and contradictions (Allen et al. 1999, Tardanico and Lungo 1995). We can then think of cities also as one of the sites for the contradictions of the globalization of capital, even though the city cannot be reduced to this dynamic.

7. *Conclusion*

An examination of globalization through the concept of the global city introduces a strong emphasis on strategic components of the global economy rather than the broader and more diffuse homogenizing dynamics we associate with the globalization of consumer markets. As a consequence, this also brings an emphasis on questions of power and inequality. And it brings an emphasis on the actual work of managing, servicing, and financing a global economy. Second, a focus on the city in studying globalization will tend to bring to the fore the growing inequalities between highly provisioned and profoundly disadvantaged sectors and spaces of the city, and hence such a focus introduces yet another formulation of questions of power and inequality.

Third, the concept of the global city brings a strong emphasis on the networked economy because of the nature of the industries that tend to be located there: finance and specialized services, the new multimedia sectors, and telecommunications services. These industries are characterized by cross-border networks and specialized divisions of functions among cities rather than international competition *per se*. In the case of global finance and the leading specialized services catering to global firms and markets—law, accounting, credit rating, telecommunications—it is clear that we are dealing with a cross-border system, one that is embedded in a series of cities, each possibly part of a different country. It is a *de facto* global system.

Fifth, a focus on networked cross-border dynamics among global cities also allows us to capture more readily the growing intensity of such transactions in other domains—political, cultural, social, criminal.

Global cities around the world are the terrain where a multiplicity of globalization processes assume concrete, localized forms. These localized forms are, in good part, what globalization is about. Recovering place means recovering the multiplicity of presences in this landscape. The large city of today has emerged as a strategic site for a whole range of new types of operations—political, economic, 'cultural,' subjective. It is one of the nexus where the formation of new claims, by both the powerful and the disadvantaged, materializes and assumes concrete forms.

See also: Cultural Studies: Cultural Concerns; Culture, Sociology of; Globalization and World Culture; Globalization, Anthropology of; Globalization: Geographical Aspects; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Information Society; Information Society, Geography of; Information Technology; International Communication: History; Internet: Psychological Perspectives; Science and Technology: Internationalization

Bibliography

Abu-Lughod J L 1999 *New York, Los Angeles, Chicago: America's Global Cities*. University of Minnesota Press, MN

Allen J, Massey D, Pryke M (eds.) 1999 *Unsettling Cities*. Routledge, London

Arrighi G 1994 *The Long Twentieth Century. Money, Power, and the Origins of Our Times*. Verso, London

Beauregard R 1991 Capital restructuring and the new built environment of global cities: New York and Los Angeles. *International Journal of Urban and Regional Research* **15**(1): 90–105

Braudel F 1984 *The Perspective of The World*. Collins, London, Vol. 3

Brenner N 1998 Global cities, global states: Global city formation and state territorial restructuring in contemporary Europe. *Review of International Political Economy* **5**(1): 1–37

Castells M 1989 *The Informational City*. Blackwell, London

Chase-Dunn C 1984 Urbanization in the world system: New directions for research. In: Smith M P (ed.) *Cities in Transformation*. Sage, Beverly Hills, CA

Eade J (ed.) 1996 *Living the Global City: Globalization as a Local Process*. Routledge, London

Elliott J R 1999 Putting 'global cities' in their place: Urban hierarchy and low-income employment during the post-war era. *Urban Geography* **20**(2): 95–115

Fainstein S, Gordon I, Harloe M 1993 *Divided City: Economic*

Restructuring and Social Change in London and New York. Blackwell, New York
Friedmann J, Goetz W 1982 World city formation: An agenda for research and action. *International Journal of Urban and Regional Research* **6**: 309–44
Graham S, Marvin S 1996 *Telecommunications and the City: Electronic Spaces, Urban Places*. Routledge, London
Gravesteijn S G E, van Griensven S, de Smidt M C (eds.) 1998 Timing global cities. *Nederlandse Geografische Studies* **241**
Hall P 1966 *The World Cities*. McGraw-Hill, New York
Hamnett C 1996 Why Sassen is wrong: A response to Burgers. *Urban Studies* **33**(1): 107–10
Holston J (ed.) 1996 Cities and citizenship. *Public Culture* **8**(2)(Winter)
Isin E (ed.) 2000 *Democracy, Citizenship and the Global City*. Routledge, London
Keil R 1993 *Weltstadt- Stadt der Welt: Internationalisierung und lokale Politik in Los Angeles*. Westfaelisches Dampfboot, Munster, Germany
King A D 1990 *Global Cities: Post-Imperialism and the Internationalization of London*. Routledge, London
Knox P L, Taylor P J (eds.) 1995 *World Cities in a World-System*. Cambridge University Press, Cambridge, UK
Kunzmann K R 1994 Berlin im Zentrum europäischer Städtnetze. In: Werner S (ed.) *Hauptstadt Berlin. Band 1: Nationale Hauptstadt Europaeische Metropole*. Berlin Verlag, Berlin, pp. 233–46
Lo F, Yeung Y (eds.) 1996 *Emerging World Cities in Pacific Asia*. United Nations University, Tokyo
Logan J R, Swanstrom T (eds.) 1990 *Beyond the City Limits: Urban Policy and Economic Restructuring in Comparative Perspective*. Temple University Press, Philadelphia
Low S M 1999 Theorizing the city. In: Low S M (ed.) *Theorizing the City*. Rutgers University Press, New Brunswick, NJ, pp. 1–33
Markusen A, Gwiasda V 1994 Multipolarity and the layering of functions in the world cities: New York city's struggle to stay on top. *International Journal of Urban and Regional Research* **18**: 167–93
Machimura T 1998 Symbolic use of globalization in urban politics in Tokyo. *International Journal of Urban and Regional Research* **22**(2): 183–94
Marcuse P, van Kempen R 2000 *Globalizing Cities. A New Spatial Order*. Blackwell, Oxford, UK
Meyer D R 1991 Change in the world system of metropolises: The role of business intermediaries. *Urban Geography* **12**(5): 393–416
Moulaert F, Scott A J 1997 *Cities, Enterprises and Society on the Eve of the 21st Century*. Pinter, New York
National Academy of Sciences (NAS) 2000 *Panel on Urban Data Sets*. Committee on Population, NAS, Washington, DC
Nijman J 1996 Breaking the rules: Miami in the urban hierarchy. *Urban Geography* **17**(1): 5–22
Noller P, Prigge W, Ronneberger K (eds.) 1994 *Stadt-Welt*. Campus Verlag, Frankfurt, Germany
Portes A (ed.) 1995 *The Economic Sociology of Immigration*. The Russell Sage Foundation, New York
Rodriguez N P, Feagin J R 1986 Urban specialization in the world system. *Urban Affairs Quarterly* **22**(2): 187–220
Ross R, Trachte K 1983 Global cities and global classes: The peripheralization of labor in New York City. *Review* **6**(3): 393–431
Santos M, Souze M A A, Silveira M L (eds.) 1994 *Territorio Globalizacao e Fragmentacao*. Editorial Hucitec, Sao Paulo, Brazil
Sassen S 1982 Recomposition and peripheralization at the core. *Contemporary Marxism* **5**: 88–100
Sassen S 1984 The new labor demand in global cities. In: Smith M P (ed.) *Cities in Transformation*. Sage, CA
Sassen S 1991/2000 *The Global City: New York, London, Tokyo*. Princeton University Press, NJ
Savitch H V 1996 Cities in a global era: A new paradigm for the next millenium. In: Cohen M A, Blair A, Ruble J, Tulchin S, Garland A M (eds.) *Preparing for the Urban Future. Global Pressures and Local Forces*. Woodrow Wilson Center Press, Washington, DC, pp. 39–65
Scott A J 2001 *Global City-Regions*. Oxford University Press, Oxford, UK
Short J R, Kim Y 1999 *Globalization and the City*. Longman, Essex
Simon D 1995 The world city hypothesis: Reflections from the periphery. In: Knox P L, Taylor P J (eds.) 1995 *World Cities in a World-System*. Cambridge University Press, Cambridge, UK, pp. 132–55
Skeldon R 1997 Hong Kong: Colonial city to global city to provincial city? *Cities* **14**(5): 265–71
Smith R C 1997 Transnational migration, assimilation and political community. In: Crahan M, Vourvoulias-Bush A (eds.) *The City and the World*. Council of Foreign Relations, NY
Smith D, Timberlake M 2002 Cross-border air traffic patterned networks. In: Sassen S (ed.) *Global Networks/Linked Cities*. Routledge, London
Smith M P, Feagin J R (eds.) 1995 *The Bubbling Cauldron: Race, Ethnicity, and The Urban Crisis*. University of Minnesota Press, Minneapolis
Storper M 1997 *The Regional World: Territorial Development in a Global Economy*. Guilford Press, New York
Stren R 1996 The studies of cities: Popular perceptions, academic disciplines, and emerging agendas. In: Cohen M A, Blair A, Ruble J, Tulchin S, Garland A M (eds.) *Preparing for the Urban Future. Global Pressures and Local Forces*. Woodrow Wilson Center Press, Washington, DC, pp. 392–420
Tardanico R, Lungo M 1995 Local dimensions of global restructuring in urban Costa Rica. *International Journal of Urban and Regional Research* **19**(2): 223–49
Taylor P J 1995 World cities and territorial states: The rise and fall of their mutuality. In: Knox P L, Taylor P J (eds.) 1995 *World Cities in a World-System*. Cambridge University Press, Cambridge, UK, pp. 48–62
Taylor P J, Beaverstock J V, Walker D R F 2002 Introducing GaWC: Researching world city network formation. In: Sassen S (ed.) *Global Networks/Linked Cities*. Routledge, London
Thrift N, Leyshon A 1994 A phantom state? The de-traditionalization of money, the international financial system and international financial centres. *Political Geography* **13**(4): 299–327
Veltz P 1996 *Mondialisation Villes et Territoires: L'Economie d'Archipel*. Presses Universitaires de France, Paris
Walton J 1982 The international economy and peripheral urbanization. In: Norman I, Fainstein S (eds.) *Urban Policy under Capitalism*. Sage, CA, pp. 119–35
Watson S, Bridges G (eds.) 1999 *Spaces of Culture*. Sage, London
Wheeler J O 1986 Corporate spatial links with financial institutions: The role of the metropolitan hierarchy. *Annals of the Association of American Geographers* **76**(2): 262–74

S. Sassen

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Cities: Cultural Types

Urban forms and structures vary from one region of the world to the other. For better understanding of such variations the notion of culture realm is helpful.

1. *The Intermediate Level of Looking at Cities*

From the global point of view cities are local concentrations of population sharing certain features such as site and location factors, form (streets, houses), functions, and land uses. These features are common to all cities around the globe.

At the opposite end of the scale there is the individual city. From the ideographic point of view each city is unique as to its historical development and design.

At an intermediate level and mainly on cosmographic, religious, mercantile, or military grounds, the cities in a specific region of the world have developed certain peculiar traits common only to them and distinguishing them from the cities of other regions. These traits derived from ideas and thoughts of their inhabitants about their way of life and the desire to shape their settlements accordingly. Cities are, to a certain degree, a mirror of the intentions of their founders and of all successive generations over the centuries.

This aspect has been somewhat neglected, even to the present day. Prior to the 1950s, professional geographers were occupied mainly with location and growth factors, and urban functions. During the first quarter of the twentieth century a few German geographers made urban morphology their research topic, and it was in this period that a few cultural–genetic studies appeared. The Austrian Oberhummer participated in the Transcontinental Excursion of 1912 across the United States and published a comparative study on American and German cities in a memorial volume (Oberhummer 1915). Fleure (1920) wrote an article on various types of European cities for the 1920 *Geographical Review*, while in 1928 Passarge convened a symposium on urban issues in various countries and two years later published the results in his book *Stadtlandschaften der Erde* (Townscapes of the World) (Passarge 1930).

Interest in cross-cultural research arose again after 1950 when urban land uses and urban structure became major research topics. Simultaneously, the British geographer Smailes (1955) drew the attention of the English-speaking scientific community to urban morphology with his paper on townscapes. These attempts were, however, very soon superimposed by the so-called 'quantitative revolution' and the rise of *urban social geography*.

This is why the cultural–genetic approach was given rather little attention in standard textbooks. Of course, a few cross-cultural studies appeared, by Scargill (1979), Brunn and Williams (1983), and Agnew et al. (1984). The models of some cultural types of cities are discussed in Ehlers (1992). But separate chapters devoted to cultural types of cities are only found in the urban geography textbooks by Beaujeu-Garnier and Chabot (1963), Hofmeister (1999), and Rugg (1972). A vast amount of material on cultural types of cities will, however, be available in the near future when the series 'Urbanization of the Earth' started by Tietze in 1977 is completed.

It is postulated that urbanization processes are the same all over the world. However, they encounter different cultures and consequently generate different urban forms and structures in various parts of the world.

These regional differences may be seen against the background of culture realms. In the present author's opinion one may distinguish between twelve culture realms in the world and twelve cultural types of cities, respectively. These are the European, Russian, Chinese, Japanese, Southeast Asian, Indian, Oriental (Middle Eastern), Central African, South African, Australian, Anglo-American, and Latin American types.

We should, however, be aware of the fact that the notion of culture realm implies a rather high degree of generalization. As we attempt to go further into details we shall soon realize, for example, that Anglo-American cities are far from alike. The mere existence of the two political entities of the United States and Canada has had some bearing upon urban developments, so that US cities will look somewhat different from Canadian cities (see Goldberg and Mercer 1986). In an attempt to do justice to such variations Holzner et al. (1967) derived 34 sub-types of cities.

Even within the confines of the Dominion of Canada, cities are different to a certain degree as we compare the cities in the predominantly French-settled Province of Quebec with the cities of the Province of Ontario so that we may at least distinguish between a French-Canadian type and an Anglo-Canadian type of city (see Hecht 1977). A further breakdown might lead to even more regional subtypes.

2. *Cultural Types of Cities in their Culture Realms*

The following paragraphs will restrict themselves to these twelve cultural types of cities corresponding to the twelve great culture realms of the world, and to a few selected cultural traits.

2.1 *The European City*

Europe's cities originated from processes that started in the eighth century with the consolidation of political

power and economic development. The ruler's power, be it clerical or secular, found its manifestation in cathedrals, monasteries, and castles. To the present day, churches and castles have remained focal points of European cities.

Usually the early European town was a dual settlement: besides the ruler's court, merchants' quarters developed as the second point of origin. The town was a closed entity, separated visually from its hinterland by a wall, its residents living under special jurisdiction and enjoying the right of trading. The market place and town hall became the focal point of the city. The heritage of these essentials of early European urbanism are still obvious in present-day townscapes.

There are, however, regional variations. For example, Southern Europe's towns were usually built on hills corresponding to the acropolis of Athens, with the effect of many streets actually being stairway. The hilltop location gave them a certain amount of safety not only against enemies but also against the torrential floods of the seasonally inundating Mediterranean rivers. The numerous squares were integrated intensively into the daily lives of their residents.

In contrast to the Continent, British towns show a less compact building fabric due to much earlier dismantling of their fortifications and a strong desire of people for single family dwellings, these traits making them look more like Anglo-American than European cities.

2.2 The Russian City

Many old Russian towns developed on the western bluffs of rivers with the kremlin as their core while the settlement was less compact on the eastern bank. The second core was the *posad* or merchants' quarters. Small suburban settlements called *slobody* developed as living quarters of other population groups, while still farther away often fortified monasteries were founded. The kremlin eventually lost its strategic function and mutated to become the city's administrative and cultural center.

In the colonial period grid-pattern quarters were added to many historic towns. Especially in the Islamic towns of Central Asia, the modern Russian quarters made for a conspicuous contrast to the cul-de-sac layout of the historic town center.

During the Soviet era, socialist principles were applied to town planning. The concept of microrayon corresponded with the American neighborhood principle inasmuch as entities of 6,000 to 25,000 residents with basic service functions were conceived. Three to four microrayons were called a district, and there were two more levels in the intra-urban hierarchy.

The post-Soviet changes of the 1990s brought foreign investments, mainly into hotels and office buildings, the privatization of *kolkhoz* markets and some residential quarters, and a somewhat greater mixture of social status groups.

2.3 The Chinese City

The widespread regular grid pattern supposedly originated in military camps on the dangerous northern margin of the later Chinese Empire. Geomantics played a decisive part in the layout of towns. The rectangular system was related to yin, representing the Earth, in contrast to the circular yang, representing Heaven. The ruler's palace always faced south. A fixed number of streets ran north–south, crossing the east–west streets at right angles. However, some streets were a little displaced in order to prevent evil spirits from getting through.

During the nineteenth century many towns along the east coast and the rivers expanded by so-called 'concessions,' i.e., foreign missions, industrial, port and trading facilities, and living quarters established by Europeans and Americans.

In the present People's Republic there is a hierarchical structure of urban residents with a certain number of households forming a residents' group, several groups a residents' committee, and several committees a town. Residents are usually assigned dwelling units and social services by production complexes, these *danweis* being highly autonomous entities. The planning of great modern industrial parks since 1958 gave rise to a number of satellite cities that are supposed to relieve the metropolises from uncontrolled expansion as well as eventually abolishing the urban–rural dichotomy.

2.4 The Japanese City

The rectangular *cho*-pattern copied from China and the observing of geomantic rules made for similarities to the Chinese cities. Around 1950 one half of Japan's cities originated from the castle town or *joka-machi* of the fifteenth and sixteenth centuries. The castle district is located in a strategic hillside position, protected by a system of walls and ditches and divided into the palace area of the *daimyo* in the center and the quarters of the higher and lower *samurai* (warriors). Beyond the wall there were the quarters of the craftsmen, merchants, and priests. During the reforms of 1868 the *daimyo* lost their privileges. The castle district mutated to become a civic center in those castle towns that were able to acquire manufacturing and administrative functions. The other half of Japan's cities are port cities, market places, shrines or spas.

During the twentieth century many small shopping streets grew to the size of large shopping centers, although without any societal and cultural functions. Huge underground centers at the railroad stations

became keen competitors for the traditional shopping streets of the temple districts.

A number of walled industrial parks with dormitories for single workers and multi-storey tenements for married workers' families called *danchi* were established. The third trend was land reclamation beyond the former coastlines for large modern industrial plants with their own port facilities and infrastructure of dwelling units, supermarkets, and social institutions.

2.5 The Southeast Asian City

Geomantic and anthropomorphic rules were observed in their layout. The port was considered the 'heart' of the city, while the seat of the clerical power on the south side was its 'soul.' The north was identified with the body's head and was supposed to accommodate the representative functions. The east was compared with the right or working hand and assigned to the craftsmen, the west with brain work and mental culture, and assigned to the officials and guardians of the palace.

Since the Age of Discoveries, European urban patterns and buildings have been introduced, Jakarta's Dutch origin becoming obvious by its *grachten* and Manila's Spanish origin by its *plaza mayor*. As in many colonial cities there was an intermediate status group between the indigenous people and the colonial elite. In Southeast Asia this group is made up of the Chinese, who dominate trade and commerce.

A conspicuous element of many towns is their fishing villages where boat people, often referred to as 'sea gypsies,' spend most of their lifetime on boats or sampans.

Great percentages of Indonesia's urban population are found in settlements referred to as 'kampungs,' these settlements having a functional mixture of dwellings, crafts, and trade, and originating in villages still lacking most urban amenities. Some authors consider them squatter settlements. In recent times many kampungs have been integrated into the building fabric of the growing metropolises, while only those at the urban fringe have maintained their village-like appearance.

2.6 The Indian City

Hindu urban culture is based on meditation images and symbols such as mandalas with a regular street pattern arranged around temples devoted to one of the deities. The regular Indian grid is centered on an intersection of two major axes with a temple nearby, while the palace is integrated into the wall. Ritual and secular uses of streets have been highly interwoven.

The caste system played a decisive role in Hindu culture. The four major castes originated in various parts of the deity Vishnu's body, specific quarters of the city being allocated to each of them. Priests used to live together with the ruler and the members of his court in the central temple and palace district or in the northern sector of town. The officials privileged by their position as supervisors of the ruler's water and land resources lived in the eastern sector, the merchants in the south, and the peasants in the west.

Over the centuries more than 3,000 subcastes have developed due to secessions, political quarrels, and the growing division of labor. These *jatis*, characterized by hereditary professions and high rates of endogamy have made for a very low degree of mobility and a high degree of segregation.

In North India the irregular Indian grid influenced by Islamic culture is characterized partly by winding streets of changing width and crooked byroads.

A completely new element was added to these towns by the Anglo-Indian stations of the British comprising the 'cantonment' or military quarters, the 'civil lines' or quarters of the officials of the Indian Civil Service, private entrepreneurs, and the 'railway colony.' A number of shops were lined up along the Mall while the so-called *sadr bazaar* served those troops recruited from the indigenous people for whom the Hindu town was off limits. Even a new house type was created: the bungalow, with its rectangular shape, pyramid-like roof thatched or covered with bricks, one storey high with porches on each side, and put on stilts for better ventilation and protection against floods.

2.7 The Oriental or Islamic City

There has been a long-lasting controversy as to whether the cities from Morocco to Pakistan should be termed Oriental or Islamic. Adherents to the latter opinion consider the Friday mosque and the bazaar the two dominating elements, while their opponents argue that the suq, the caravansary, and the *hammam* or bathhouse can be traced back to the Roman colonnades, basilica, and thermae of pre-Islamic times, respectively.

One of the most conspicuous traits of the medina or historic town center is its cul-de-sac pattern, while the few thoroughfares are exceptions. One explanation for this phenomenon is the juxtaposition of clans of different origins as to their ethnicity, religion and language, and their desire to live segregated from other residents. Second, in contrast to the Roman carriages, camels and mules were used for transportation. The third reason is the legal status of the cul-de-sacs: they are not public spaces, but rather are owned by their neighbors.

Much real estate of the medina is in *waqfs*, i.e., foundations donated by rulers, officials, or merchants for religious and beneficial purposes or the use of their children, and administered by trustees. Neither these nor the users are interested in investments, so that

these foundations prevent urban renewal, leading to the decay of buildings and the emigration of the wealthier people.

The bazaar is structured in a way that the most valuable goods such as jewelry, books, or perfumes are traded in the most centrally located and covered lanes. Spices, shoes, and carpets are found some distance from the center, while pottery, leather goods, auto parts, and other manufactured and bulky goods are found in a still more peripheral location. In recent years this pattern has been weakened. Simple stalls have been replaced by shops with shop windows, and the whole bazaar has faced competition from the development of a central business district (CBD) and modern suburban shopping centers.

2.8 The Central African City

There has been a controversy as to whether the Yoruba towns of southern Nigeria were urban settlements in the strict sense of the word. The Hausa developed a special kind of segregation, since no strangers were allowed into their town, those people founding settlements of their own outside the gates: *sabon-gari* which in Hausa means 'new town.' The British colonial administration adopted the *sabon-gari* system for ethnic and hygienic reasons.

Although the numbers of European soldiers, administrators, and business managers remained small, their settlements were spacious in contrast to the crowded areas of the indigenous people. There were intermediate groups between them and the European elite: the Levantines in West Africa and the Indians in East Africa. They were mainly in trade, transport, and lower ranks of the administration. The Indians especially had their own temples, schools, clubs, and bazaars with *dukas*, i.e., open shops with verandahs closed with wooden shutters at night. In recent times modern shopping centers have developed beside the duka-style bazaars. A slow transformation has occurred from a society composed of the European upper class, the Indian middle class, and the African lower class toward a society structured along socioeconomic lines.

In many towns the indigenous population is divided into a dominating tribal group and several minority groups from different tribes, each of them stressing their own tribal traditions, a phenomenon referred to as 'retribalization' or 'peasantization.'

2.9 The South African City

The early towns of the Cape Province and the towns in Orange Free State and Transvaal were founded by the Afrikaners, and all others by the British. Within a few decades they went full circle from the colonial town through the apartheid town, and a transition period to the post-apartheid town. As early as 1913 the Native Land Act prohibited Africans from acquiring land outside their reservations, or later homelands. They lived either as servants in their employers' households or were forced to stay in 'locations' and hostels.

The central city was reserved for white citizens and, as an exception, for some colored groups, by the Group Areas Act of 1948. Even white people lived more or less segregated as to their Afrikaner, British, or Portuguese origin. The Natives Resettlement Act of 1954 initiated comprehensive relocation. Often railroad tracks, highways, or canals served as barriers between the quarters of racial groups, while each racial quarter was assigned a considerable number of places of employment in order to keep commuting through other racial quarters to a minimum.

However, complete segregation was never achieved. Since 1986, non-white enterprises have been admitted to so-called 'free trade areas' while 'grey areas' have been legalized, and even 'free settlement areas' open to all racial groups have been established. The repeal of the racial laws in 1991 brought about considerable mobility and squatting.

2.10 The Latin American City

The Spaniards founded their overseas towns around the *plaza mayor* with the *cabildo*, cathedral, and courthouse. They had a central–peripheral social gradient with the upper-class people living in the *cuadras* (blocks) nearest to the plaza.

This social structure has persisted to the present partly because centrally located residences are still considered a privilege. Part of the city center mutated into the CBD, with some storeys being added to old patio houses and *edificios* or high-rise office buildings being constructed. Behind the fashionable *paseos*, overcrowded multi-storey tenements or *conventillos* had developed, which received many rural migrants who eventually further migrated to the shantytowns at the urban fringe.

Since the 1930s a sectoral pattern has been superimposed on the traditional zonal structure by the development of manufacturing plants and worker's quarters along railroad lines, and by the centrifugal expansion of upper-class quarters.

At the urban fringe a more or less high percentage of people live in shantytowns called *villas miserias* in Argentina or *favelas* in Brazil. Most of their occupations are in the so-called 'informal sector' of the economy. As a measure of relief, *poblaciones* or public housing estates in combination with site-and-service projects have been built by the governments of various South American states.

2.11 The Anglo-American City

While the grid pattern had already been introduced to colonial North America the grid was applied

rigorously to most towns beyond the Appalachian Mountains after the release of the Land Ordinance in 1785.

The United States was the pacemaker in skyscraper construction. Competition from huge suburban shopping centers and the loss of purchasing power of the clientele of downtown shops brought decay to the downtown areas of US cities. Since the 1950s, revitalization programs have been carried out with malls and shopping *gallerías* often in combination with skyways, downtown motels, and modern cultural centers with concert halls, convention centers, and exhibition halls. Historical foundations have purchased and resold many historic buildings by means of revolving funds.

Since in large areas the founding of towns and the construction of railroads occurred simultaneously, railroad tracks often cut right through the town center with level crossings, while large manufacturing areas developed right outside the CBD along the tracks. The inner residential areas decayed, because most of the building stock was made of timber and prone to swift degradation, and the improved-value system of taxation was no incentive for investments, so that tax delinquency and vacancy rates increased rapidly. Moreover, since the 1830s low-status groups of immigrants concentrated in these inner quarters where they were close to their fellow countrymen and to manufacturing jobs. In recent years many such areas have been upgraded by 'urban homesteading' and gentrification.

For many decades suburbia has been dominated by single-family homes and duplexes. Especially, young families moved to the suburbs seeking better schools for their children. Other reasons were investments in high-quality real estate and the chance to live close to neighbors sharing the same lifestyle ('lifestyle suburbs'). The metropolises became completely fragmented by dozens of huge shopping centers, these functioning as catalyzers for modern office and business parks, and the growth of 'urban villages' termed 'edge cities' by Garreau (1991).

2.12 The Australian City

In general, Australia's cities very much resemble the Anglo-American type. There are, however, certain differences. The CBD has never experienced such a high concentration of skyscrapers with all their disadvantages. The inner suburbs of the nineteenth century have not experienced a decay similar to that of US cities due to more favorable conditions: the long rows of terrace houses are rather solid buildings made of brick or natural stone (with the exception of the Queenslander house made of timber and set on stilts), the site-value system of taxation, the much smaller concentration of non-British minority groups, most of these immigrants only arriving after 1950 in a country with a booming economy and high wages enabling them to become home owners in some suburban areas within a few years, and the high degree of gentrification.

Urban sprawl has been even greater than in the United States and made for approximately 70 percent home ownership. Despite this and a high rate of car ownership there are amazingly few urban expressways (with the exception of Perth). A widely accepted planning concept is the development of a limited number of growth corridors along railroad lines and major highways with designated district centers around suburban railroad stations.

See also: Cities: Internal Structure; Cultural Diversity, Human Development, and Education; Cultural Geography; Cultural Landscape in Environmental Studies; Cultural Landscape in Geography; Cultural Resource Management (CRM): Conservation of Cultural Heritage; Ecology, Cultural; Urban Activity Patterns; Urban Growth Models; Urban History

Bibliography

Agnew J A, Mercer J, Sopher D E (eds.) 1984 *The City in Cultural Context*. Allen & Unwin, New York

Beaujeu-Garnier J, Chabot G 1963 *Traité de géographie urbaine*. A. Colin, Paris

Brunn S D, Williams J F (eds.) 1983 *Cities of the World. World Regional Urban Development*. Harper & Row, New York

Ehlers E (ed.) 1992 Modelling the city: Cross-cultural perspectives. *Colloquium Geographicum* **22**, Bonn

Fleure H J 1920 Some types of cities in temperate Europe. *Geographical Review* **10**: 357–74

Garreau J 1991 *Edge City. Life on the New Frontier*. Doubleday, New York

Goldberg M A, Mercer J 1986 *The Myth of the North American City*. University of British Columbia Press Vancouver, BC

Hecht A 1977 *Die anglo- und frankokanadische Stadt. Trierer Geographische Studien Sonderheft 1*. Trier, Germany, pp. 87–11

Hofmeister B 1999 *Stadtgeographie*, 7th edn. G. Westermann Brunswick, Germany

Holzner D Dommisse E J Mueller J E 1967 Toward a theory of cultural–genetic city classification. *Annals of the Association of American Geographers*: 367–81

Oberhummer E 1915 *Amerikanische und europäische Städte. Memorial Volume of the Transcontinental Excursion of 1912*. New York

Passarge S (ed.) 1930 *Stadtlandschaften der Erde*. Breslau, Germany

Rugg D S 1972 *Spatial Foundations of Urbanism*. W C Brown Co, Dubuque, IO

Scargill D I 1979 *The form of cities*. St. Martin's Press, New York

Smailes A E 1955 Some reflections on the geographical description and analysis of townscapes. *Transactions of the Institute of British Geographers*: 99–115

Tietze W (ed.) 1977 *Urbanization of the Earth/Urbanisierung der Erde*. Berlin

B. Hofmeister

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Cities, Images of

Community appearance matters to people. This article reviews scientific findings on the visual features of cities that convey a strong and desirable image to people who experience them. The article defines the concepts, reviews the research, and discusses methodological questions and future research and use of the findings.

1. City Form: A Scientific Approach

City form is shaped by and affects many people. Research shows that appearance is central in human responses to their surroundings (cf. Nasar 1994, 1997). City form 'should be guided by a "visual" plan: a set of recommendations and controls' for its appearance (Lynch 1960). US legislation and the courts grant governments the authority to control appearances (Mandelker 1996) and most American cities do so (Lightner 1993). To work, appearance controls must consider people's image of places. This article centers on the two key aspects of the image: imageability and linkability. Imageability refers to the probability that an environment will evoke a strong image from observers (Lynch 1960); and linkability refers to the probability that an environment will evoke a strong favorable response from observers (Nasar 1997).

For most of its history, urban design—the practice of shaping urban form—has followed a philosophical approach. Theorists speculated on what ought to be, but did not arrive at or test their speculations scientifically. Lynch (1960) suggested and tested a scientific approach. He assumed that people would more likely know, and so use, an environment that was easy to read or legible. This made legibility a valid purpose for research and design.

Lynch described the environmental image as having three parts: *identity, structure*, and *meaning*. This means humans recognize or identify objects (identity), they see a recognizable pattern of relationships between objects (structure), and they draw emotional value (or have feelings) about the objects and structure (meaning). The meaning of a place may take a denotative or a connotative form. Denotative meanings are the same as identity; they refer to judgments of what the place is. Connotative meanings refer to inferences about the quality and character of the place or its users. People often think of such connotative meanings as a question of aesthetics. This article avoids the term 'aesthetics,' because of its connection to art, where a statement may take priority over pleasure, and because many people view aesthetics as something one cannot quantify. Aesthetics also has its roots in philosophy and normative theory. Philosophers question the degree to which aesthetic experience arises from psychical distance or phenomenological involvement; and designers offer top-down speculations about the correct aesthetic for the design of places.

Fechner (1876) introduced the bottom-up approach to the study of aesthetics (meanings). He studied human evaluations of attributes of simple stimuli, such as rectangles and polygons. Almost a hundred years later, Berlyne (1971) revived this scientific approach, called empirical aesthetics; and at a time when psychologists sought real-world relevance, Wohlwill (1976) extended the ideas and approach of Berlyne to the study of real places. Although evaluative responses, such as preference, may vary across individuals (cf. Little 1987), the theories of Berlyne (1971) and Wohlwill (1976) held that certain kinds of physical features would likely have hedonic value. More recently, some theories suggest an evolutionary basis for certain shared environmental preferences (Kaplan and Kaplan 1989, Orians 1986, Ulrich 1993). Theories differ in their interpretation of the importance and interdependency of the individual and the environment in response. Research, however, confirms that beauty rests more in the features of the place evaluated than in the head of the evaluator (Stamps 1995).

Although Lynch recognized the importance of meaning, he felt one could not easily manipulate it through changes in urban form. He accepted the conventional wisdom that preference is highly variable across individuals. He judged meaning as impractical to study, and concentrated on form—identity and structure—separate from meaning.

2. Components of City Imageability

Lynch interviewed residents in three cities to see what they recalled about their cities; he found strong consensus across respondents. They converged on five elements that enhance a city's imageability: landmarks, paths, districts, edges, and nodes. *Landmarks* are visible reference points, such as towers or mountains. *Paths* are channels for movement, such as streets or walkways. *Districts* are large sections of a city that have some recognizable, common perceived identity distinguishing them from other areas. *Edges* are barriers, such as shorelines, rivers, or railroad cuts. *Nodes* are focal points of intensive human activity. Research confirms the stability of these elements for many populations and cities around the world (cf. Nasar 1997). Although the images and prominence of elements may vary across populations and places (Rapoport 1977), the correct arrangement of the elements can heighten the imageability of a city. Research that grew from Lynch's seminal work has yielded much information about mental maps, distance perception, and wayfinding (cf. Evans 1980).

To shape urban form, imageability is not enough. One must consider people's evaluation of the city, the meanings they see, or their evaluative image.

3. Components of the Evaluative Image of the City

Communities might not need a scientific understanding of the bases for evaluative responses if the designers, design review boards, and other experts who shape places produced designs that pleased people. Regrettably, research shows that they often do not (cf. Nasar 1994, 1997).

To find the evaluative image, one must consider both the evaluative responses important to people and the features of the environment that people notice and evaluate. Research has found three important aspects of human evaluative response to places (Russell and Snodgrass 1989). Preference is a purely evaluative dimension. Mixes of pleasure and arousal produce excitement and relaxation. Exciting places feel more pleasant and arousing than boring ones; and relaxing places feel more pleasant but less arousing than stressful ones.

Evaluative response to places may arise from two formal and symbolic variables (cf. Kaplan and Kaplan 1989, Nasar 1994). Formal variables have to do with the structure of form and include such things as shape, proportion, scale, complexity, incongruity, novelty, and order. Symbolic or content variables have to do with the connotative meanings associated with the forms. Several kinds of theories discuss the relationship between these variables and response. One set of theories view preference as dependent on arousal (Berlyne 1971, Mandler 1984, Wohlwill 1976). Of the many variables these theories cite as affecting arousal, complexity and novelty (atypicality) have garnered the most research attention. In theory, complexity and novelty increase arousal, interest, and excitement; but preference has an inverted U-shaped relationship to arousal. Preference would increase with increases in complexity or novelty up to a point, after which increases in complexity or novelty would produce a downturn in preference. Another theory offers an evolutionary model in which human survival depended on preference for involvement and making sense, and as a result, humans now prefer places that offer involvement and either make sense or promise to make sense (Kaplan and Kaplan 1989). This theory posits complexity and mystery (the promise of new information ahead, as in a deflected vista) as creating involvement; and it posits coherence and legibility as helping people make sense of things. People should like a mix of complexity, mystery, coherence, and legibility.

Research shows seven environmental features as prominent in human perception and evaluation of places: naturalness, order, complexity, novelty (atypicality), upkeep, openness, and historical significance (cf. Nasar 1994, 1997). People recognize variation from natural (vegetation) to human-made. Research shows that humans prefer vegetation, that preference increases with the addition of vegetation, decreases with the increase in human-made elements, and that people dislike obtrusive signs, utility poles, overhead wires, and billboards, traffic, and intense land uses. Commuters drive out of their way to use a parkway rather than a less natural expressway; and research suggests that exposure to vegetation may have restorative or healing effects.

Research shows that people notice and prefer order. Preference for order has emerged for many kinds of urban settings and for various ordering variables, including legibility, coherence, identifiability, clarity, compatibility, and congruity. People also prefer well-kept to dilapidated areas. Dilapidation and disorder such as vandalism, boarded up buildings, and litter, which researchers refer to as physical incivilities, also contribute to a perception of the breakdown of social controls, fear of crime, and crime.

Complexity relates to the number of different elements and the distinctiveness between those elements in a scene. Research shows that people notice variations in complexity, and that interest, excitement, and viewing time increase with complexity, but that preference tends to be highest for moderate levels of complexity. Though some research points to contradictory findings, those findings suffer from method biases. Research shows that novelty and atypicality also increase excitement and interest. People prefer moderate to low levels of novelty or atypicality. Though some studies show contradictory results, the discrepancies arise from flaws in measuring novelty and familiarity.

People readily notice changes in spaciousness. Preference increases with openness, but people also like some spatial definition. People also like mystery (in the form of deflected vistas), but for uncertain conditions such as urban areas deflected vistas and uncertainty about information ahead heightens fear.

Places may have historical significance or just look historical. In either case, they evoke favorable response. People also prefer popular styles to the high-style designs. The preference for historical significance and certain popular styles over high styles may arise from connotative meanings associated with them or from the mix of complexity and order in them.

Naturalness, upkeep, and historical significance appear to be symbolic variables, while the others appear to be formal variables, but each one may work for formal or symbolic reasons. In addition, people may like some of these variables for their contribution to order or for their associations with status. Naturalness, upkeep, open views, order (compatibility), and historical significance enhance order, but these same features may look like feature that wealthier

persons can afford. People notice status, make accurate judgments of social status from environmental cues, and prefer high-status to low-status areas.

For integrative reviews on perception and preference of these features, see Kaplan and Kaplan (1989), Nasar (1988a, 1988b, 1989, 1994, 1997) and Wohlwill (1976); for vegetation see Kaplan (1995), Ulrich (1983, 1993), and Wohlwill (1983); for disorder incivilities, see Taylor (1989); for atypicality, see Mandler (1984), Purcell and Nasar (1992), and Whitfield (1993); for fear and mystery, see Nasar and Jones (1997).

4. Methodological Issues

Visual quality research makes choices in selecting respondents, environmental stimuli, measures of environmental features, and measures of evaluative response. The choices involve trade-offs between what is practical, what allows experimental control, and what will generalize to the real situation. For a full review of the methodological issues see Nasar (1997, 1999). One thing that differentiates visual quality research from other social science inquiries is the need to get response to the appearance of places (Marans and Stokols 1993). For this, many studies use color photographs and slides. Research shows that responses to such stimuli, even though they lack movement and sound, accurately reflect on-site response (Stamps 1990).

5. Future Directions

Future research needs to better define the linkage between the judged and actual physical attributes. It should examine movement through environments. It should apply scientific methods to historical data to examine longitudinal aspects of evaluative response. It should use meta analysis to integrate findings of previous studies statistically. It should supplement verbal responses with psychophysiological measures and observation of behavior.

To derive specific guidelines for special situations, one can also use visual quality programming. This involves applied research to develop information for a visual quality plan or guidelines. Nasar (1997, 1999) demonstrates examples of this for various applications. This can lead to improvements in the visual quality of communities for residents and visitors.

See also: Memory for Meaning and Surface Memory; Mental Imagery, Psychology of; Mental Representations, Psychology of; Multi-attribute Decision Making in Urban Studies; Space: Linguistic Expression; Spatial Analysis in Geography; Spatial Cognition

Bibliography

Berlyne D E 1971 *Aesthetics and Psychobiology*. Appleton-Century-Crofts, New York

Evans G 1980 Environmental cognition. *Psychological Bulletin* **88**: 259–87

Fechner G 1876 *Vorschule der Asthetik*. Breitopf and Hartel, Leipzig, Germany

Kaplan R, Kaplan S 1989 *The Experience of Nature: A Psychological Perspective*. Cambridge University Press, New York

Kaplan S 1995 The restorative benefits of nature: Towards an integrative framework. *Journal of Environmental Psychology* **15**: 169–82

Lightner B 1993 *A Survey of Design Review Practice in Local Government*. (*Monograph*) *MEMO: Planning Advisory Service (PAS)*. American Planning Association, Chicago

Little B R 1987 Personality and the environment. In: Stokols D, Altman I (eds.) *Handbook of Environmental Psychology, Vol. 1*. Wiley, New York, pp. 205–44

Lynch K 1960 *The Image of the City*. MIT Press, Cambridge, MA

Mandelker D R 1998 *Land Use Law*. Mische, Charlottesville, VA

Mandler J M 1984 *Stories, Scripts, and Scenes: Aspects of Schema Theory*. Erlbaum, Hillsdale, NJ

Marans R W, Stokols D 1993 *Environmental Simulation: Research and Policy Issues*. Plenum, New York

Nasar J L 1988a *Environmental Aesthetics: Theory, Research, and Applications*. Cambridge University Press, New York

Nasar J L 1988b Perception and evaluation of residential streetscenes. In: Nasar J L (ed.) *Environmental Aesthetics: Theory, Research, and Applications*. Cambridge University Press, New York, pp. 275–89

Nasar J L 1989 Perception, cognition and evaluation of urban places. In: Altman I, Zube E (eds.) *Human Behavior and Environment: Public Spaces*, Plenum, New York, pp. 31–56

Nasar J L 1994 Urban design aesthetics: The evaluative quality of building exteriors. *Environment and Behavior* **26**: 377–401

Nasar J L 1997 *The Evaluative Image of the City*. Sage, Thousand Oaks, CA

Nasar J L 1999 *Design by Competition: Making Design Competition Work*. Cambridge University Press, New York

Nasar J L, Jones K 1997 Landscapes of fear and stress. *Environment and Behavior* **29**: 291–323

Orians G H 1986 An ecological and evolutionary approach to landscape aesthetics. In: Penning-Rowsell E C, Lownthal D (eds.) *Landscape Means and Values*. Allen and Unwin, London, pp. 3–25

Purcell A T, Nasar J L 1992 Experiencing other peoples houses: A model of similarities and differences in environmental experience. *Journal of Environmental Psychology* **12**: 199–211

Rapoport A 1977 *Human Aspects of Urban Form*. Pergamon Press, Oxford, UK

Russell J A, Snodgrass J 1989 Emotion and environment. In: Stokols D, Altman I (eds.) *Handbook of Environmental Psychology, Vol. 1*. Wiley, New York, pp. 245–80

Stamps A E 1990 Use of photographs to simulate environments. A meta-analysis. *Perceptual and Motor Skills* **71**: 907–13

Stamps A E 1995 Stimulus and respondent factors in environmental preference. *Perceptual and Motor Skills* **80**: 668–70

Taylor R B 1989 Toward an environmental psychology of disorder: Delinquency, crime and fear of crime. In: Stokols D, Altman I (eds.) *Handbook of Environmental Psychology, Vol. 2*. Wiley, New York, pp. 951–86

Ulrich R S 1983 Aesthetic and affective response to natural environment. In: Altman I, Wohlwill J F (eds.) *Behavior and the Natural Environment: Human Behavior and Environment, Advances in Theory and Research, Vol. 6*. Plenum, New York, pp. 85–125

Ulrich R S 1993 Biophilia and the conservation ethic. In: Kellert S R, Wilson E O (eds.) *The Biophilia Hypothesis*. Island Press, Washington, DC, pp. 73–137

Whitfield T W A 1983 Predicting preference for everyday objects: An experimental confrontation between two theories of aesthetic behavior. *Journal of Environmental Psychology* **3**: 221–37

Wohlwill J F 1976 Environmental aesthetics: The environment as a source of affect. In: Altman I, Wohlwill J F (eds.) *Human Behavior and the Environment: Advances in Theory and Research, Vol. 1*. Plenum, New York, pp. 37–86

Wohlwill J F 1983 The concept of nature: A psychologist's view. In: Altman I, Wohlwill J F (eds.) *Behavior and the Natural Environment*. Plenum, New York, pp. 5–37

J. L. Nasar

Cities, Internal Organization of

Economic activities, land uses, and socioeconomic status of population seldom distribute evenly or randomly in an urban area. They typically differentiate into internally homogeneous clusters. It is difficult not to conceive that this differentiation is governed by some underlying principles of spatial organization. And urban form affects the economic efficiency, social equity, environmental quality, and sense of place. Therefore, understanding theories of urban spatial organization helps advance knowledge and shape better futures for urban areas.

Contemporary urban development has long spread beyond the political boundaries of a city. A typical urban landscape in an advanced economy is a contiguous urbanized area that consists of multiple cities and their suburbs. Hence, urban spatial organization must be analyzed in the context of an urban region.

Most studies in urban spatial organization have been conducted in North America where an advanced capitalist market economy prevails. Their generalizations may not be completely applicable to cities that have evolved in different economic modes, such as those where the public sector dictates urban development or communal institutions control property rights. Furthermore, North American urban places are different from the rapidly growing mega-cities (having populations of more than 15 million people) in Asia and Latin America.

1. *Urbanization and Suburbanization*

During the nineteenth century, North America rapidly industrialized and urbanized. For example, between 1820 and 1920, the number of US cities with populations of 5,000 persons or more exploded from 39 to 1,467. During this period, these cities were also under transformation. Originally evolved from a focal point of employment, these cities were relatively small and compact because contemporary transportation systems limited their expansion. Technological advances led to successive transportation systems: streetcar, suburban rail, and subway. Each new system helped to push the urbanized area further.

Similarly, in the early twentieth century, the use of trucks, automobiles, and telephones vastly expanded the urban horizon and dispersed wholesale and manufacturing uses to the suburbs. Meanwhile, new building technologies such as elevators and skyscrapers helped rebuild city centers into central business districts (CBDs)—containing blocks having concentrations of high-rise buildings of offices and retail.

While the urban share of the US population first reached 50 percent in 1920, the most striking change that would sweep America, suburbanization, was delayed by the Great Depression and World War Two. In the postwar period, three interrelated processes restructured the urban landscape. High rates of family formation and the 'baby boom' (a surge of births between 1946 and 1964) created extreme housing shortages that the existing urban fabric could not accommodate. Concurrently, huge federal and state expressway-building programs opened up new residential suburbs while urban renewal efforts accelerated the outward movement of blacks from segregated ghettoes into nearby predominantly white neighborhoods. Throughout the 1950s and the 1960s, these last two processes resulted in the so-called 'white flight' or 'flight-from-blight' phenomena (Mieszkowski and Mills 1993). These processes accelerated an already existing decentralizing trend and led to further dispersal of economic activities. As suburban centers captured jobs, housing, and stores, the central cities lost their dominance in metropolitan population and employment. The result was the growth of sprawling urban regions. These trends continued through the late twentieth century as urbanized areas gradually become polycentric.

By 2000, the US demographic character was firmly established. For example, 80 percent of its population was classified as urban, up from 56 percent fifty years earlier. Furthermore, suburbanization gained momentum. In 2000, the share of the urban population living in the suburbs was 60 percent, a complete reversal from the 40 percent of 1950. Suburbanization is not unique to the US; other nations of advanced economies are undergoing similar changes (Ingram 1998).

2. *Patterns of Spatial Organization*

Evaluating urban spatial organization encompasses several patterns, including land use differentiation, population distribution, household characteristics,

income and racial segregation, employment distribution, changing building densities, and the rise of subcenters and polycentric urban forms.

Urban land uses are highly differentiated. Land uses can be compatible or incompatible with one another. Compatible uses generate mutual benefits or positive externalities. Incompatible uses create harmful consequences or negative externalities. Compatible uses usually coexist with one another while incompatible uses show rigid segregation. It is debatable whether public land use control or voluntary market forces contribute to this segregation.

Theoretically, land uses display a number of geometric forms: concentric rings, sectors or wedges, multiple nuclei or small clusters, or a linear formation from a narrow strip to an arc or a corridor. The Burgess model (1925) posits that land uses fall into concentric zones extending outward from the CBD. Two alternative formulations complementing this schema are the sectoral model, arguing that wedges of similar activities radiate from the CBD along transportation corridors, and the multi-nuclei model asserting that secondary CBDs and suburban economic centers emerge to accommodate second-order activities. It appears that these three forms of land-use arrangements can coexist in a single urban area. For example, concentric rings and corridor-type of uses are evident in a polycentric Los Angeles.

The way that population is distributed displays some regularities as well. Urban population densities tend to be higher around the CBD and lower at the edges. Alonso (1964) developed a monocentric model to reflect this pattern and calculated a density decay function, $D(\mu) = D.exp(-b\mu)$, where the average density, D, starts at the fringe of the CBD and declines at a rate, b, per unit distance, μ, away from the center. This function flattens as time changes, i.e., densities at the fringe will marginally increase or the limit of a city continues to expand. Over time, population densities have decreased around the core, reflecting the depopulation of most of the central cities since the 1950s. The monocentric model is limited because its approximation of densities is specific to a specific scale of observation. The gross population density, a commonly used measure, tends to underrepresent the net density in the periphery because it factors unpopulated and nonresidential areas into the analysis. Comparing neighborhood by neighborhood, the net residential density in the outer suburbs may not be too different from that in the inner suburbs.

The spatial distribution of household characteristics is not uniform. Households of small size, female-headed, or with no children are more concentrated in the core. Families and larger size households are more dispersed in the periphery where larger lots and detached housing are more abundant. The differences in preference also reflect the impact of family cycles in housing choice. Families tend to be more sensitive to house size and school quality.

Income segregation is the most noticeable feature of American urban spatial organization. Broadly speaking, higher income groups reside in the suburbs while the low-income groups stay in the core. This general pattern is skewed by the high concentration of poverty in the core and also interrupted by other irregularities. Small high-income precincts are commonly found in the urban core (e.g., Coral Cables in Miami, Georgetown in Washington DC, and Pacific Heights in San Francisco). Similarly, not all suburban households are high-income because many suburban development projects target the middle class. Unlike population density, changes of income level rarely move smoothly away from the CBD. Neighborhoods with large income differences are often separated by such physical barriers as rivers, highways, or even manmade barriers.

Employment distribution exhibits more dispersal than residential distribution. Manufacturing activities started to move out from the center in the 1900s as trucks replaced rail freight movement. By mid century, as air transportation become important, manufacturing and distribution centers gravitated toward airports. Expressway expansion has provided ubiquitous access within urban areas and diminished the locational advantages of their cores. Today, manufacturing and wholesale activities tend to cluster around freeway interchanges in the suburbs. Lower-order retailing activities have also moved to the suburbs as purchasing power has shifted in that direction. Suburban malls were first developed in the late 1950s and within decades modern and bigger malls were built in the newer suburbs, leaving CBD and old suburban retail centers to falter. Despite that the average CBD's share of metropolitan employment has fallen to less than ten percent, CBDs have not completely lost their role because most of them still represent the biggest single concentration of specialized services and government activities.

Building and development densities also exhibit a distance decay relationship with the CBD but their gradients are much steeper and often disrupted by surges of density in secondary CBDs and suburban centers. In general, high-rise development and multi-family housing cluster in the primary and the secondary centers more than in other part of the urban region.

The rise of subcenters and polycentric urban form is a direct result of continuous urban expansion and shifts in population and employment distribution. As population and employment disperse, they gravitate to suburban centers and transform the old monocentric urban structures into polycentric ones. These centers have a propensity to cluster or distribute loosely in a sectoral form. These centers evolve from older settlement points or they create their own set of activities and population concentration. Garreau (1991) characterized the rise of these centers as 'Edge Cities,' recently developed places with a sizeable office

and retail space attracting large numbers of commuters. Today, these centers are increasingly self-contained and making suburbs independent of the cores of their metropolitan regions. Recent studies reveal that office and professional service activities are continuing to scatter. As a result, more than half of the commuting traffic is among these centers.

3. *Searching for the Principles of Spatial Organization*

Explaining the organizing principle(s) of spatial organization requires sophisticated understanding of the social, economic, technological, cultural, and other phenomena influencing urban form and structure. Spatial organization is derived from observable patterns specific to the scale of observation. Also, the paradigm adopted by a researcher generally shapes the findings. This section identifies and discusses common determinants of urban spatial organization.

3.1 *Factors that Influence Urban Form and Structure*

Second to topography, the urban morphology, the built environment and the associated property rights, is the most important constraint to urban development. The 'urban capital stock,' i.e., lots, buildings, streets, road grids, rail tracks, and expressways are relatively inelastic to changes. Since modifications of urban morphology involve very costly land assembly, older areas in the core are likely to be left derelict. Similarly, streets and freeways usually form impasses of a land use or boundaries to a neighborhood. Infrastructure, such as transportation facilities and transshipment centers and their ancillary uses resist changes. Their obsolescence or closure causes dereliction and blight. Though building structure is malleable and conversion of uses is possible, these transitions are lengthy and sometimes costly, especially in the absence of market demand and public intervention. In contrast, development in the fringe is less risky, subject to less uncertainty and cheaper. As real estate activities tends to be cyclical, construction occurs in waves. Investment converges on certain locations for a short duration and then wanes. Sometimes, these activities create an impressive mark on urban form—power centers, corridors of strip malls, and sprawling gated communities. Successive urban expansions have reflected the impetuous and speculative nature of the development process.

Technological changes greatly affect spatial organization. The prime factor behind persistent decentralization has been advances in transportation technology. For example, expressway systems have radically extended urbanized areas. Contemporary sprawl typically hops along newly developed transportation corridors. Similarly, the development of remote and difficult terrains is possible because of technological changes to such infrastructure systems as water supply, power grid, and sewerage and drainage. Finally, the advent of modern telecommunications incorporating personal computers, facsimiles, and cellular telephones is producing new locational requirements for firms (Graham and Marvin 1996). While some speculate that the revolution in information technology will cause the 'death of distance' and the reintegration of home and workplace, evidence indicates that it merely expands choices of human interaction and does not reduce commuting (Kotkin 2000). Its impact on urban structure is still not clear, as it has generated both centralization and decentralization of economic activities.

Income, wealth, and preferences have a great influence on population distribution. Income and wealth affect consumer preferences and the ability to live in specific locations. Research on housing choice has consistently shown the importance of such attractive attributes as neighborhood amenities, school quality, personal safety, and quality of life. As the spatial distribution of these attractive attributes is highly differentiated, higher income groups tend to outbid lower income groups in all localities in the price competition for these attributes. Therefore, high-income group clusters are not confined to the suburbs. Both 'middle-class flight' and 'gentrification' reflect how higher income groups are sensitive to the perceived changes of neighborhood quality. Furthermore, preferences in choosing neighbors along racial or ethnic lines are translated into such discriminatory practices as selective real estate information, fiscal zoning and other restrictive land use regulations. These practices challenge the argument that income and racial segregation is voluntary. Evidence that minority inner-city residents have greater difficulties in dispersing to the suburbs in comparison to other groups of similar income level indicates that racial preference has prevented smooth and market-driven neighborhood transition.

Economic restructuring destabilizes the old order of industrial location. Traditionally, manufacturing firms were sensitive to transportation costs reflected in the distance toward market or raw material. However, in today's postindustrial era, minimization of transportation costs has become less important as their share in total production costs has declined. Such factors as labor costs and quality, business climate, economies of scale, and quality of life are new considerations. In addition, the agglomeration effects where inter-industry and inter-firm linkages allow sharing of innovations and access to large labor pools and networks of suppliers and purchasers are important. The postindustrial era has also witnessed the employment dispersal that, together with residential

decentralization, has developed into a multi-point and multi-directional commuting pattern. Under the new economy, knowledge-based industries require the new set of locational considerations discussed above (Wheeler et al. 2000). Finally, under globalization, urban regions have developed more external links and expanded their hinterlands. While at the same time, the economic functions of inner-city neighborhoods have changed. In some instances, the increase flow of capital, information, products, and people has given rise to specialized districts such as ethnic quarters for immigrants and clusters of import and export firms. In others, especially where in economically isolated neighborhoods, high rates of unemployment have led to severe distress.

Although the private sector leads urban development, the public sector plays an important role as the major undertaker of infrastructure projects and provider of such services as public schools, law and order, waste collection and disposal, parks and recreation, and land use regulation. Infrastructure steers future development and the level of public services directly affects the quality of life. In addition, local tax levels and the efficiency and attitude of local government shape the business climate that most private businesses rely on in making investment decisions. American urban areas commonly contain a large number of small municipalities. This fragmentation has resulted in great variations in the quality of services and business climate. Labeled as the Tiebout effect, these variations induce businesses and residents to move among places, leading localities to compete with one another to attract investment and middle-class residents. This competition is often unequal because some municipalities have fewer resources, and suffer from eroding tax bases and the desertion of the middle class. Such unequal competition reinforces segregated and fragmented spatial organization.

3.2 Interpretations of Spatial Organization

The study of the spatial organization of urban places can take many approaches. Each casts a different image and each adopts a particular set of methods and paradigms. Here are the major approaches.

The ecological approach developed at the University of Chicago was the earliest systematic study of urban form (Park 1925). It viewed the city as a composite of multiple ecological colonies segregated by income, class, and ethnicity. Its organization principle was that the succession and invasion of one colony into another yielded a specific urban form of concentric zones. It suggested a distribution of colonies where the transients, low-income groups, and recent immigrants clustered around the CBD. The working class and more stable immigrant groups were further out. Single-family dwellings came next and commuters occupied the outmost urban zone. This approach stimulated generations of urban studies. It highlighted the competition and conflict over space among socioeconomic groups. It treated spatial form as a manifestation of a dynamic process of commanding space.

The utility maximization model has been the most dominant approach used to analyze spatial organization in the latter half of the twentieth century. Based on Alonso's seminal monocentric model that assumed that firms and individuals made tradeoffs between transportation costs and locational rent, it stimulated numerous studies focusing on land use and transportation modeling (Ottensmann 1975). The subsequent work commonly adopted neoclassical economic principles as the underlying organizing theme and treated an urban area as one aggregate unit of optimization where each agent settled in the location that maximized its utility and profit. This work was generally mathematical and followed a set of deductive methods that required vigorous formulation of premises. Although the resulting findings followed strong reasoning, they were contingent on the assumptions stated in the model. This work has been employed for the simulation of outcomes of one variable when other variables were under careful control.

An emerging interdisciplinary field, urban morphology, examines the physical form of cities, stresses the historical effects of the built environment and property rights, and studies the interaction of elements of the physical structure at various scales (Vance 1990). A related approach, space syntax, correlates human activities and attributes with physical configurations and linkages of buildings and street blocks. Architects and urban designers have adopted these techniques to examine the evolution of building forms and styles, street layouts, and the massing of buildings throughout history (Kostof 1992). They have extended beyond principles of design to interpretations of historical and institutional constraints and behavioral interactions with space.

The systems approach attempts to understand spatial organization in its entirety (Bourne 1982). It views the city as a human body where its internal environment interplays with its external environment. It tries to understand spatial organization holistically by studying the form, interrelationships, behavior, and evolution of activities of an urbanized area. It considers urban structure an undefined reflection of the historical and organizational principles of society, which, in turn, are the products of current and previous operating rules of culture, technology, economy, and social behavior. Since this approach refuses to isolate these elements, it has had to devise an overarching analytical framework to incorporate various organizing principles. The difficulties in developing paradigmatic coherence have prevented this attempt to flourish.

The most recent effort to understand spatial organization, the postmodern approach, has been adopted across such disciplines as geography, sociology, anthropology, cultural studies, and urban studies (Dear 2000, Soja 1989). Sometimes containing neo-Marxist doctrines, it examines urban development in the context of the rapid transformation of society under globalization and economic restructuring. It is sensitive to the worldviews held by groups differentiated by gender, class, ethnicity, and race. While it does not specifically examine spatial organization *per se*, it provides insights onto how space is perceived, battled over, and controlled. It legitimizes the use of qualitative methods to study the complex urban process. It considers the city not as one economic unit or a production center but a place with multiple functions. It also stresses the interaction of countervailing internal, regional, and global forces in polarizing and tearing apart the traditional connections of urban spaces.

See also: Spatial Interaction Model; Urban Sprawl; Spatial Organization Models; Spatial Pattern, Analysis of; Urban History/Towns and Cities in History; Urban Geography; Urban Sociology; Urban Studies, General; Income Distribution; Wealth Distribution

Bibliography

Alonso W 1964 *Location and Land Use*. Harvard University Press, Cambridge, MA

Anas A, Arnott R, Small K A 1998 Urban spatial structure. *Journal of Economic Literature* **36**: 1426–64

Berry B J L 1965 Internal structure of the city. *Law and Contemporary Problems* **30**(1): 111–19

Bourne L S (ed.) 1982 *Internal Structure of the City*. Oxford University Press, New York

Burgess E W 1925 The growth of the city. In: Park R E, Burgess E, McKenzie R (eds.) *The City*. Chicago University Press, Chicago

Dear J M 2000 *The Postmodern Urban Condition*. Blackwell, Malden, MA

Garreau J 1991 *Edge City: Life on the New Frontier*. Doubleday, New York

Graham S, Marvin S 1996 *Telecommunications and the City: Electronic Spaces, Urban Places*. Routledge, London

Harris R, Lewis R 1998 Constructing a fault(y) zone: Misrepresentations of American cities and suburbs, 1900–1950. *Annals of the Association of American Geographers* **88**(4): 622–39

Ingram G K 1998 Patterns of metropolitan development: What have we learned? *Urban Studies* **35**(7): 1019–35

Kostof S 1992 *The City Assembled: The Elements of Urban Form Through History*. Bulfinch, Boston

Kotkin J 2000 *The New Geography: How the Digital Revolution Is Reshaping the American Landscape*. Random House, New York

Mieszkowski P, Mills E S 1993 The causes of metropolitan suburbanization *Journal of Economic Perspectives* **7**(3): 135–47

Ottensmann J R 1975 *The Changing Spatial Structure of American Cities*. Lexington Books, Lexington, MA

Park R E, Burgess E, McKenzie R (eds.) 1925 *The City*. University of Chicago Press, Chicago

Queen S A, Thomas L F 1939 *The City: A Study of Urbanism in the United States*. McGraw-Hill, New York

Soja E 1989 *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. Verso, New York

Vance J E Jr. 1990 *The Continuing City: Urban Morphology in Western Civilization*. Johns Hopkins University Press, Baltimore

Webber M M, Dyckman J W, Foley D L (eds.) 1964 *Explorations into Urban Structure*. University of Pennsylvania Press, Philadelphia

Wheeler J O, Aoyama Y, Warf B (eds.) 2000 *Cities in the Telecommunications Age: The Fracturing of Geographies*. Routledge, New York

Yeates M 1998 *The North American City*, 5th edn. Harper and Row, New York

S. Wong

Cities: Internal Structure

Urban research has always focused on big cities. Big cities are concentrates of the cultural and economic potential of mankind. They are centers of innovation as well as hotbeds of social and ecological problems. Moreover, they have long outgrown the limits of perception. Not only do planners plead for a return to human dimensions, but all those engaged in urban research seek to delimit manageable segments, either by subject matter or by spatial criteria.

1. General Principles

1.1 Accessibility and Transport Technology

The city is a centered system based on the twofold premise that the city center is the engine of development and the place easiest to access. Accessibility of both the center and the entire city area is determined by its opening up for development and by transport technology (see Fig. 1).

A city of pedestrians and coaches tended to be circular. When streetcars came into use in Europe, they started mostly at the town gates. This made for star-shaped growth, with the interstitial areas lagging behind in development. Where public transport is predominant, the city center remains the easiest place to access. The *métropole concentrée* of Paris is an example. The absolute predominance of private transport reduces access to the center; under a liberal political system and with land abundantly available, a

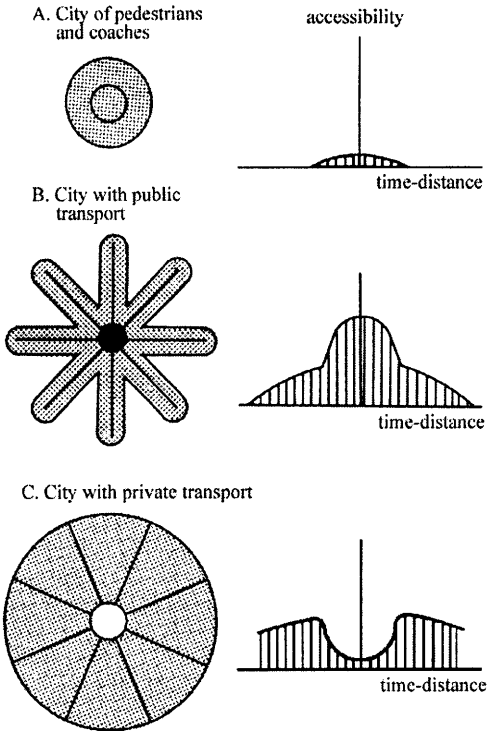

Figure 1
The accessibility of the city center

city or metropolitan region designed for car traffic emerges in several steps. Los Angeles offers a prototype of this.

1.2 Distance Decay

There are two theoretical approaches for analyzing gradients from the center to the periphery:

1.2.1 Social gradients. The central peripheral organization of urban society may be described by means of centrifugal and centripetal social gradients. These gradients permit the following diagnosis regarding urban growth. Wherever the city center is also the social center and a centripetal social gradient is in evidence, the demand for space of the upper classes extends outward from the center. This causes residential and social upgrading of the adjacent middle-class districts. In the middle zone, former lower-class quarters are turned into middle-class districts. The Founders Period growth of Berlin, Budapest, Copenhagen, Paris, Prague, and Vienna was of this kind. This type of upgrading must not be confused with

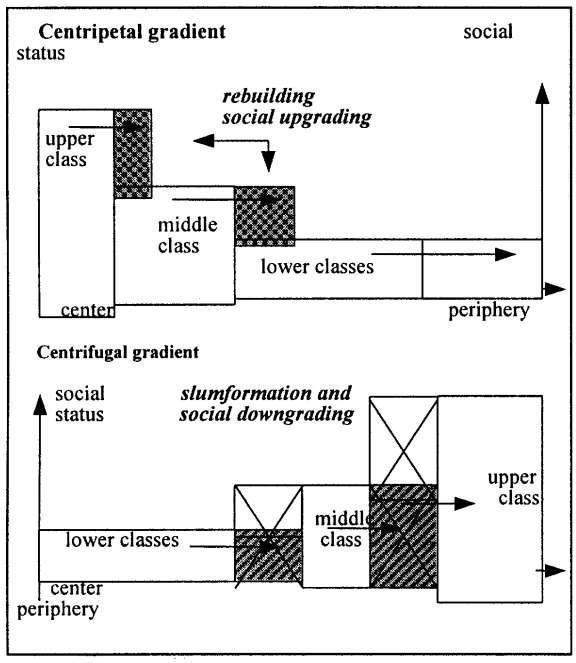

Figure 2
Centrifugal and centripetal social gradients
(*Stadtgeographie* (*Urban Geography*) 1998, p. 110)

the recent phenomenon of gentrification in North American cities. Here, middle and upper income urbanites move centripetally into rundown central districts (see Fig. 2).

Wherever a centrifugal social gradient predominates, as in US cities, a filtering down takes place: with the deterioration of the building fabric in the center, lower-class people move outward into adjacent middle and upper-class districts. This social downgrading of high quality living quarters is hard to stop and presented a major problem for British town planners. Having accomplished the redevelopment of rundown early industrial terrace houses next to the old town centers, they were faced with the daunting task of redeveloping devastated mid- and late nineteenth century terrace house districts (while the necessary renovation of early twentieth century council housing was partly effected by privatization under Thatcher).

1.2.2 Land values gradient. According to the theory of urban land markets (Alonso 1964), urban land use mirrors transport cost as well as the rent of land. Regularly, in socio-economically intact city centers, there is a center-periphery gradient with several consecutive zones of use outward from the central business district (CBD). Replacement of historical city-models by the new model of suburbia together with the re-

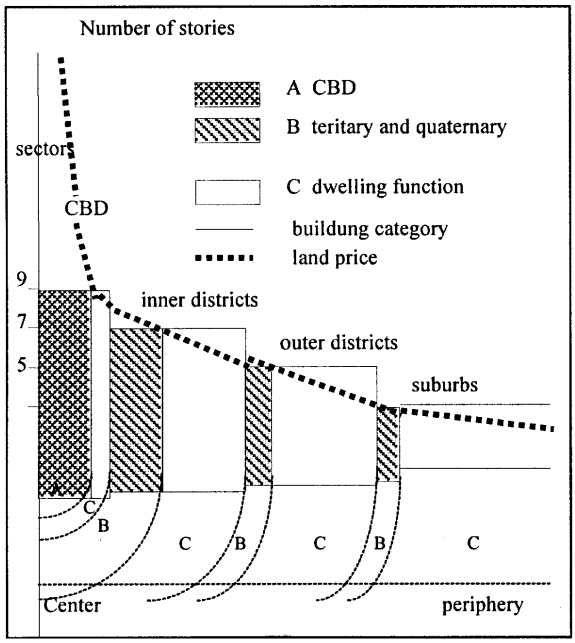

Figure 3
Land values gradient and zoning

duction in accessibility have made for an abandonment of central areas, resulting in 'craters' of land prices and visible decay of inner cities in the USA and parts of Western Europe. In the centers, the gradient now falls in the opposite direction. The Alonso model excludes restrictions on planning land use and vertical development. With zoning laws, building categories, and other legal regulations, the gradient of real estate prices is altered. Each category or zone is divided in two, with the inner zone obviously better suited for business purposes and office space than the outer zone that is used for dwelling as it is more profitable (see Fig. 3).

1.3 Hierarchical Structures

Hierarchical order is one of the basic types of systemic organization. When organizing the physical space of urban areas, a hierarchical order has been attempted wherever planned development was feasible and land in abundant supply. Examples run from 17C Sicily (Granmichele) via Howard's New Town idea (1902), with city center and garden suburbs in a cluster city, to major planning projects for urban peripheries in Europe at present. The hierarchy is constructed from the bottom upwards. The basic units are electoral wards, which in many European countries and in North America double up as statistical units: their dimensions are derived from the traditional 'pedestrian ideology' of 1 km (i.e., 15 minutes walking distance) as standard size. Comprehensive hierarchical urban structures have been realized only in the former socialist countries of Eastern Europe, above all the USSR. In the rest of Europe, hierarchical organization of urban space is the exception rather than the rule. However, it is frequently found in traditional Oriental cities. There, a spatial hierarchy of family, clan, and local quarters is based on family, religious, and ethnic affiliations. The hierarchical organization of living quarters and sub-centers is paralleled in the hierarchy of semi-private, semi-public, and public links as well as markets.

2. The Impact of Political Systems

2.1 Social-ecological City Models

In their paper, *The Nature of Cities* (1945), Harris and Ullman tried to model internal patterns within cities. The triad of models presented makes use of diverse development phenomena:

(a) invasion and succession of social groups in Burgess' zonal model,

(b) transport-induced ribbon-developments along traffic lines in Hoyt's sector model, and

(c) collective preferences as to the allocation of workplaces in the Harris-Ullman multiple-nuclei model.

However, three important underlying assumptions have not yet been made explicit:

(a) the political system of liberalism,

(b) the historical 'one-dimensionality' of urban development, and

(c) the concept of the city center as CBD.

The models mentioned represent ideal types of North American cities of the interwar period (see Fig. 4).

2.2 Historical-political Systems and the Concept of the City Center

Changes in the political system alter the conceptions of the city and urban society. The function of the city center is changed. European urban development may be defined as a succession of four types of political systems and the respective types of cities. Their concepts of the city center were markedly different: (a) in the burghers' city of medieval, feudal, territorial state the marketplace was the social center; (b) which shifted to the ruler's residence in the residence city of the absolutist state. Thus, a social gradient falling from the center toward the periphery is the general rule in pre-industrial cities. (c) In the age of liberalism, Great Britain created the prototype of the industrial city. A social gradient rising from the center toward the periphery predominated, a development later paralleled in North American cities. (d) Again, Great

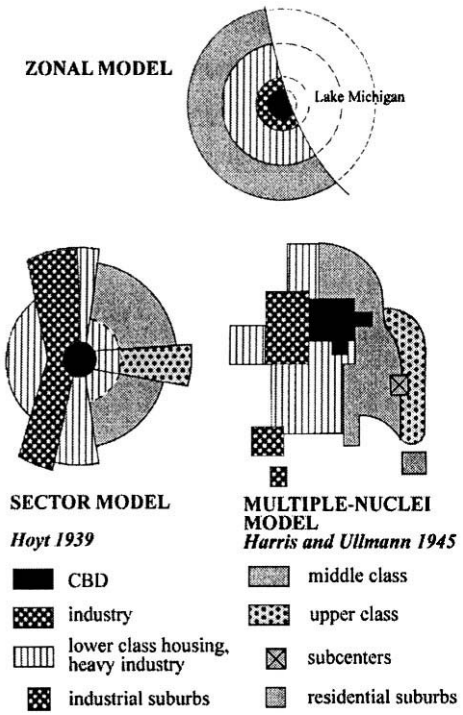

Figure 4
Social ecological models of US cities during the inter-war-period

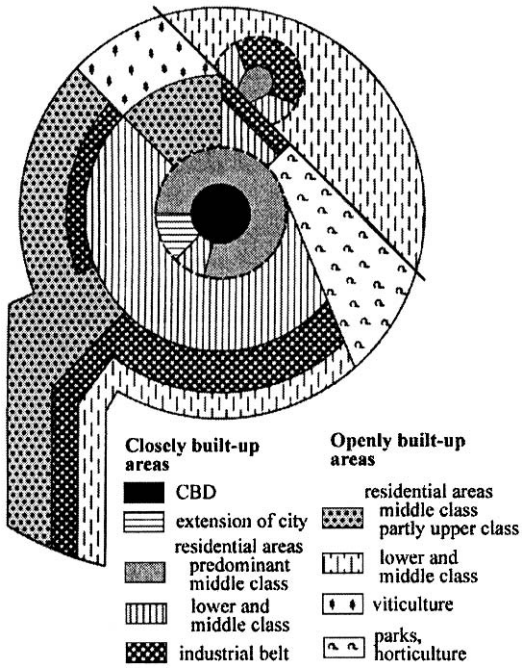

Figure 5
The social ecological model of a continental European city—Vienna (*Geoforum* 1970, **4**: 61)

Britain set the rules for the New Town idea, the attempt at structuring the amorphous masses of big industrialized cities on a human scale, with the city divided into parts with different functions. From the outset, spatial segregation of inhabitants was barred from the design of New Towns—a fact that is still influential in urban planning today. European city development is most complex; various superpositions and, to different degrees, the persistence of historical structures caused the diversification of socioeconomic patterns and different social gradients.

Vienna in the 1960s offers the model for the traditional continental European city:

(a) Social status is still highest in the core and declines toward the periphery. Depending on site preferences individual districts deviate from this rule.

(b) The CBD has maintained some residential functions and is not surrounded by slums, but by middle and upper class residential districts, encircled by lower class quarters.

(c) The wide belt of multistory blocks is followed on its outside by Founders Period industrial quarters, followed in their turn by loosely built-up districts.

(d) The fringe zone is not defined by an extensive speculation area of 'vacant land' as in North America, but by sectors of intensive agricultural land use (in keeping with the von Thünen model) such as truck farming and viticulture, as well as of allotment gardens and weekend homes (see Fig. 5).

2.3 The Impact of State Socialism

Under state socialism, municipal governments were the local planning authorities, responsible for the mass of multistory housing that was unprofitable because of 'social' low rents. They lost all chances of capital accumulation from real estate property, and became dependent on financial endowments from the central government and on central planning. The consequences were:

The big cities' increasing demand for space was met by extensive incorporations.

Public construction (by state, municipal and other collective institutions) took absolute precedence over private construction.

Architectural design followed totalitarian principles, stressing elements of over wide avenues and huge squares, both imposing and strategically important.

Massive anti-segregation strategies were pursued.

City centers were usually put under monument protection.

City enlargement was effected via New Cities in the shape of hierarchically structured giant multistory blocks.

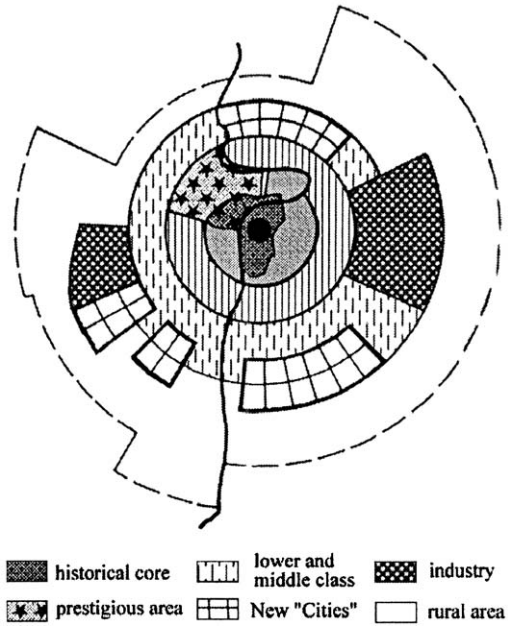

Figure 6
The model of a socialist city: Prague (*Vienna. Bridge between Cultures*, 1993, p. 102)

In compliance with the Charter of Athens, industrial and dwelling zones were strictly separated.

Commuting requirements were taken care of by subway construction.

Forced industrialization created extensive industrial zones with plants located close to railway lines and super-highways.

Extensive leisure zones were created on the fringes, with both collective recreation facilities and private second homes (see Fig. 6).

2.4 City Growth and Political Systems

Cities are growing systems, with increasing populations, and/or with increasing demand for space for diverse urban functions (housing, work, education, leisure, traffic, etc.) in a setting of economic growth and technical innovation.

The diverging political-economic effects of central planning and of a free enterprise economy cause significant differences in the physical growth of cities in the respective systems. Cities grow in two directions: laterally and vertically.

2.4.1 The third dimension. Vertical growth means high-rise construction. Up to the 1970s, when some cities changed their laws, European cities could not expand vertically because of strict zoning laws, but were forced to grow laterally. This necessitated the conversion of dwelling space into office space, e.g., in Paris and Vienna. High-rise construction in European cities started relatively late (see Fig. 7). Its location within the city—frequently subject to special permission—is different from that in North America. There, the vertical structure of urban skylines shows that land prices peak in the center while in European cities 'monument protection' bars high-rises from the centers. Thus, the new landmarks of banks, insurance firms, corporation headquarters, and hotels keep a polite distance from the old landmarks of churches, town halls, and castles.

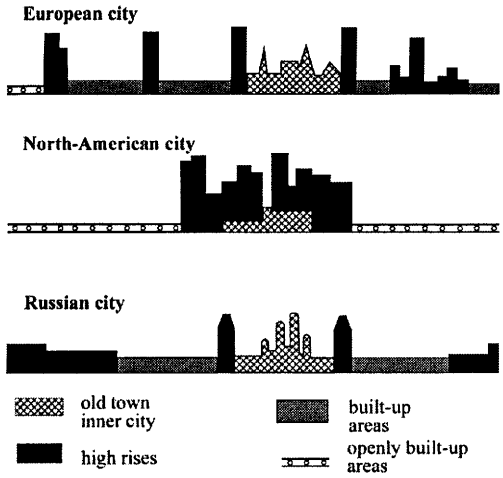

Figure 7
Skylines of European, North-American, and Russian million-cities (*Stadtgeographie* (*Urban Geography*) 1998, p. 198)

For the sake of access to the various supply and disposal mains, high-rises are preferably located along urban 'scars': at the interfaces of traditional zones where former boundaries still show in open space or low physical objects. Frequently, new high-rises accent not only the edge of traditional inner cities but, centripetally, also major points of access to older outer cities and suburbs. High-rises also mark the front of growth of the CBD, busy commuter train stations, as well as 'satellite' districts. They are also instrumental in slum clearing.

2.4.2 Urban fringe development. Development of urban fringes differs materially according to whether it takes place under private capitalism, in welfare states or, in retrospect, under state capitalism.

(a) Under US private capitalism, cities grow as profits from land speculation and rising land prices are

invested in land development and in technical improvements. This starts an upward spiral of development and rising prices. North American cities show two wide zones of speculation. The inner one, around the CBD, is marked by decay and slum areas at present. Far more impressive, at least because of its extent, is the peripheral zone of vacant land around core cities and suburbs, which, even in the 1950s amounted to 20 to 60 percent of the core areas (Bartholomew 1955). Anglo-American urban geography textbooks completely ignore those huge tracts of vacant land—implying that non-utilization of these spaces is taken for granted rather than considered a problem.

(b) In continental Europe, too, urban fringes attracted speculation. During World War I, they were occupied by so-called 'emergency gardening plots,' later turned into allotment gardens. Frequently, these became temporary settlements, partly forerunners of a second-home periphery. All over post-World-War I continental Europe, spontaneous settlements typically marked city fringes—a consequence of fundamental political changes. The absence of government checks on land use made for temporary settlements, among them the pavilions of 'chaotic urbanization' in France as well as the often illegal occupation of land around big central European cities (Belgrade, Budapest, Bukarest, Sofia, Vienna, and Warsaw). This shows that the succession states were less powerful than the Austro-Hungarian monarchy in repelling indigent illegal immigrants. Those postwar squatters were comparable to today's squatter settlements on the fringes of Third World cities beyond the reach of state or municipal authorities.

(c) As to planning and regulating the growth of agglomerations, Great Britain set the standard in the early twentieth century, with the two important concepts of the New Town (referred to above) and the Green Belt. The creation of a Green Belt presupposes government control of land use, replacing the market mechanism of the liberal age. Originating in London, Green Belts have also become constituent features of zoning plans in other cities of the former British Empire. The urban development plan of Ottawa shows a Green Belt several miles wide. Persistent urban sprawl caused a characteristic overspill. In the USA, too, Green Belt concepts were introduced into various urban development plans, but because of the enormous extent of suburbanization they have not been realized along the edges of core cities. At present, there is little interest in creating public green spaces and leisure areas there.

(d) The opposite is true for the former socialist countries where extensive public recreation areas and large private second-home districts exist. (Neither of these important elements of the periphery of urban regions is to be found in the USA.) The public recreation area of Moscow goes back to a Green Belt idea already incorporated in the city development plan of 1935. It is 20 to 40 km wide and includes major sports and cultural facilities. In most of the big cities of the former USSR, the Green Belt was an integral part of development planning.

3. *On the Nature of Cities: the Immanent Question*

With current suburbanization and counter-urbanization tendencies, the model of the city as a centered system is becoming obsolete. Scenarios have been developed which anticipate the existence of cities as 'non-places.' But there were also new models developed for the American 'urbanlike system.' During the period 1950–2000, this rapidly growing new system of suburbia created a sort of extensive—though not ubiquitous—network, in part destroying former central place hierarchies, and in part developing the surroundings of metropolitan areas. It also frequently stopped the process of restructuring central cities by means of downtown redevelopment and gentrification. Core cities of mega-metropolises such as Chicago got another chance of redevelopment as new (air) traffic junctions, centers of business, finance, and the quaternary public sector, as well as cultural centers. The dichotomy of the ideal types of 'urbanlike' structures and new mega structures seems to correspond to postindustrial America's abandoning and simply 'forgetting' the areas of urban desertification of earlier industrial city development.

See also: Cities: Cultural Types; Cities, Internal Organization of; Development and Urbanization; Ecology, Political; Fertility Transition: Economic Explanations; Spatial Interaction; Sustainable Development; Sustainable Transportation; Transportation Geography; Urban Activity Patterns; Urban Geography; Urban Growth Models; Urban Life and Health; Urban Poverty in Neighborhoods; Urban System in Geography

Bibliography

Alonso W 1964 *Location and Land Use. Towards a General Theory of Land Rent*. Cambridge, MA

Bartholomew H 1955 *Land Uses in American Cities*. Cambridge, MA

Carter H 1997 *The Study of Urban Geography*, 4th edn. Arnold, London

Heinritz G, Lichtenberger L (eds.) 1986 *The Take-off of Suburbia and the Crisis of the Central City*. Erdkundliches Wissen 76, Steiner, Wiesbaden, Germany

Herbert D T, Thomas C J 1997 *Cities in Space: City as Place*, 3rd edn. Fulton, London

Knox P L 1995 *Urban Social Geography: An Introduction*, 3rd edn. Longman, Singapore

Lichtenberger E 1993 *Wien-Prag*. Metropolenforschung, Böhlau, Vienna
Lichtenberger E 1994 The future of the European city in the West and the East. *European Review* **3**(2): 182–93
Lichtenberger E 1998 *Stadtgeographie* (*Urban Geography*), 3rd edn. Teubner, Stuttgart, Germany. Italian translation, Geografia dello spazio urbano, 1st edn., Unicopli, Milano, Italy
Lichtenberger L 1970 The Nature of European Urbanism. *Geoforum* **4**: 45–62
Lichtenberger L 1976 The Changing Nature of European Urbanization. *Urban Affairs Annual Reviews* **11**: 81–107
Lichtenberger L 1992 Political Systems and City Development in Western Societies. A Hermeneutic Approach. *Colloquium Geographicum* **22**: 24–40
Lichtenberger L 1993 *Vienna. Bridge between Cultures*. Wiley
Short J R 1996 *The Urban Order: An Introduction to Cities, Culture and Power*. Oxford, MA

E. Lichtenberger

Cities: Post-socialist

1. Issues of Definition

Socialist governments were formed at different times in different parts of the world: in Europe (USSR, eastern Europe), Asia (China, Cambodia, Vietnam), Africa (16 claimed to be socialist, but only Angola, Ethiopia, Mozambique qualified as Afro-Marxist regimes), and the Americas (Cuba). The term socialist is confined to a short historical period: the Soviet Union apart, it refers to the 40-year period after 1948. For the countries in the 'South' the period is even shorter, in the case of Cambodia from 1975 until 1989. It is doubtful that we can talk of a 'generic socialism,' beyond the fact that each country was characterized by the abolition of private ownership and the concentration of economic resources in the hands of the state and by the monopolization of political power in the hands of a 'vanguard party.' (In 1994 still only 15 percent of the population of Ethiopia lived in towns.) The term post-socialist is used to describe urban areas in societies which, until the breaching of the Berlin Wall in 1989, had been known as 'socialist.' At present no government, with the possible exception of China and Cuba, describes itself as 'socialist,' therefore socialist cities no longer exist.

1.1 Socialist Cities?

At a formal level, if it is accepted that these were 'real existing socialist societies,' then their cities were by definition socialist cities—the position adopted in the Soviet Union in 1931. The more difficult question is whether or not the socialist city is qualitatively different from the city in capitalist society. Despite dissenters, the consensus is that a distinction existed (French and Hamilton 1979, Pensley 1998, Smith 1996, Szelenyi 1996). Arguably an *a fortiori* case exists for believing that these characteristics might be found in Russia. As the first socialist country it had the longest period in which to experiment and put its ideological principles into practice. Additionally, it compelled its first satellites in eastern Europe to adopt its model of industrial and urban development. From this point of view it would be possible to gain insights into the essence of the socialist and post-socialist city by studying the experience of the former socialist countries in Europe.

Addis Ababa is a post-socialist city as are Berlin, Beijing, Budapest, Phnom Penh, and Moscow. The capital city in both post-socialist and socialist societies is the distillery of the society's contradictions and conflicts; it contains in hypertrophied form what is found to a much lesser extent in all other cities. The transformation of societies from being socialist to post-socialist is not a unilinear process (Grabher and Stark 1997). Neither does the transition from socialist to post-socialist city follow a clear pattern (e.g., Harloe 1996 Phe and Nishimura 1992). However, generalizations can be made about these cities in terms of their external, representational, and material form, their economic functions, and the social relationships amongst their populations.

2. Urban Form

2.1 The Symbolic, Representational Environment

The aesthetic of the socialist city in the Soviet Union before 1939 represented a triumphant proletariat in a hostile but not yet belligerent world. After 1945 that aesthetic had to include the victory in war against fascism. Thus, to the mausoleum and ubiquitous monuments to the heroes of the Great October Socialist Revolution, in particular Lenin, Marx, and Engels, were added grave, granite memorials to those who died in the titanic struggle against national socialism.

The aesthetic and symbolic representation of the post-socialist city is much more difficult to identify. Just as Soviet leaders, especially, but not only, Stalin and Khrushchev, demolished churches and did much to remove all vestiges of the tsarist past, post-socialist leaders are rebuilding cathedrals—the most classic being the Cathedral of Christ the Saviour near the Kremlin in Moscow—and removing statues to the now fallen heroes of socialism. China, too, destroyed much of the civilizational grandeur of the Imperial City in Beijing, while Cuba allowed Havana to wither (Segre et al. 1997).

Berlin, Budapest and Moscow encapsulate the characteristics of the post-socialist city. During

the *belle époque* their architectures expressed their countries' imperial status and cosmopolitan air. Subsequently, architecture came to symbolize inter-war authoritarianism, the defeat of fascism, a divided world—epitomized in the divided Berlin. In contrast to the socialist city, the post-socialist city has no state decreed prototype. The socialist city, as conceived during the 1920s and 1930s based on planning and reason, would, using the most up-to-date construction materials and building technologies to erect multi-storey blocks of flats, be the acme of modernism. From 1949 this was the blueprint for all socialist societies.

Now coarse granite blocks that characterised government buildings of the socialist city have been replaced by leafy atriums encased in glass, symbolically representing transparency, the current metaphor for democracy. The architectural preferences of the *nouveau riche* group tend to fall into two categories: either a sentimental veneration of the past and national styles, or an international orientation expressed in a pastiche and rococo post-modernism.

Changing architectural styles and usages of town squares are sound barometers of historical periods. The memorial and monument, which became popular at the turn of the twentieth century, embellished and blemished the streets, squares, and parks of towns and villages after the end of World War I, especially in the newly created, states. Ljubljana European even before the war had begun to substitute a Slovene for its German façade (Jezernik 1998). This process reached its apogee in the socialist city where a baroque Marian column, for example, was 'relocated' in a less visible place, while obelisks were erected to the Red Army and marble memorials were built to partisan leaders.

In many post-socialist cities, these heroes of 40 years have been, if not destroyed, decapitated or mutilated, then banished from public view, in some instances to museums so that they may be reminders of the images of dictators and criminals. These acts highlight the transience not just of heroes, but also of the myths about the inhabitants' past and who they wish or think themselves to be. The erection and removal of monuments are potent symbolic acts, which create and sustain myths and then denigrate and depose them. In the post-socialist city new buildings are themselves monuments, but less to people or events than to the new-found ideology and to the power of money. The architecture of new government buildings represents visions of the meaning and exercise of political power.

The post-socialist city celebrates three myths. The first is a ritualistic pageantry of the past, so easily portrayed in a coat of arms and the insignia of local governance that are deemed to dignify the present. Second, in keeping with contemporary Western planning theory and practice, the squares are pedestrianized and garlanded with new lamps, benches, and cafes. They are again public spaces for everyday events, not just for formal celebrations of socialist power. This is the myth of the harmonious, integrated community where everyone interacts as a free person; where the exotic and bizarre in appearance are silently applauded for their non-conformity as though these ostensible forms are a defence for everyone against socialist conformity. The third myth is that this square, which is really a circus, a place for entertainment and the consumption of material goods and pleasure, is a symbol of Europeanization or Westernization. The civic leaders of the post-socialist city bask in the sunshine of their enthusiasm for what its citizens regard as an abandonment of grayness.

The post-socialist city is cosmopolitan in theory and localist in spirit, and latently nationalist in practice. It is an arena where people learn the laws governing contracts, and the rules protecting and regulating private property. At the same time it massages into life an enemy, which for inhabitants of the socialist city was defined in terms of 'class' and property relations, a mode of analysis no longer acceptable. In the post-socialist city 'the enemy' is the alien, the stranger, the immigrant.

Berlin is distinctive from other cities in many ways, but its singular claim to uniqueness is that its post-socialist incarnation, in combining its socialist and capitalist faces in one city, should be the most visible healing of the painful rupture experienced by European nations. Ironically, however, this particular post-socialist city reveals in concentrated form the contradictions that are endemic in many of the most populous capitalist cities and which the socialist city was to have overcome. A skyline of cranes tells of a booming economy, while unemployment is at a high level; a thriving cultural and political boundary-testing artistic scene co-exists with a growing intolerance of immigrants.

2.2 *The Material Environment*

2.2.1 The economy. The keystone of the new global consensus amongst governments is the privatization of public assets and of government functions and the establishment of new property rights. This policy, which is already influencing the structure and functioning of post-socialist cities, is hampered or aided by deeply ingrained regional, cultural traditions.

The economic dimension of cities most distinguishes the socialist from the post-socialist city. The socialist city was unique in that economic wealth, which the privileged enjoyed, could not be conspicuous. Second, the post-socialist city is not just shaped by the wealth of an indigenous bourgeoisie, but also by foreign capital invested in real estate. This critical difference between the socialist and post-socialist city is already visible in three processes; the gentrification of parts of

the central city, the creation of a central business district and more luxurious shopping centers, and an influx and visible presence of foreign migrants.

Cities are shaped by movements of capital. In the main, the protective walls which surrounded socialist cities have been pulled down. Previously, in theory but also to a certain extent in practice, cities had planned economic relations with one another as part of a regional, national, and international division of labor. Post-socialist cities are in competition with each other and increasingly subjected to global competition. One aspect of globalization is the attraction of tourists. Only a small proportion of post-socialist cities have a real potential to make this a source of local economic development. More importantly, globalization accelerates attempts to commodify a city's history and culture, which is achievable largely through its sanitization and its conceptual Warholization and material McDonaldization. Globalization has also helped to accelerate the growth of the criminal economy, which is shaping the urban morphology, most visibly (especially in capital cities) in private financial institutions and shopping malls. A number of post-socialist cities, mainly, but not only, capitals (Moscow, Almaty, Berlin, Sofia, Shanghai, Phnom Penh) have become centers for the consumption and transhipment of heroin and other drugs, for trade in illegal immigrants and prostitutes, and for smuggling.

Organized crime and corruption are pervasive but particularly evident in the largest cities. In some countries there is virtually no sphere of revenue-generating activity (including charities) which has not been visited by extortionists ('insurance brokers'). In the space of a decade gangs of extortionists have evolved into 'enforcement partners.' Before signing business contracts, companies acquire information about each other's enforcement partners. Only if the enforcement partners have recognized each other and given mutual guarantees will the contract with all its formal juridical and business attributes be signed.

Whereas the socialist city was characterized by a manufacturing profile with a grossly underdeveloped service sector, the post-socialist city is dominated, at both extremes of the class spectrum, by trading and the provision of services. Both the indigenous and foreign élites have generated a demand for casinos, restaurants, night clubs, and other leisure facilities. This stimulates the local economy, creates new forms of labor market stratification and, with the commodification of land, speculation in real estate. In Cambodia the overnight transfer of state housing into private ownership in 1989 presented local government officials with a golden opportunity, of which they availed themselves, to sell off public land.

2.2.2 Land reform. Land nationalization was central to the twentieth century socialist (agrarian) revolutions. Guided by an idea of social justice it addressed issues of exploitative landlord–peasant relationships, rural poverty and land hunger and the contrast between the housing standards of rich and poor city dwellers. The state allocated land to publicly owned enterprises and institutions, free of charge and for use in perpetuity. No value was imputed to the location of the land. Since in most cases land was not in short supply, organizations applying to the state for plots asked for and received more than they required. The bundle of *de jure* use rights enjoyed by individuals and juridical entities varied from country to country, but in no case were they able to dispose of land, even of that which was surplus to their requirements. Yet, at the same time, land users behaved as though they were powerful, private owners, with the result that detailed building and zoning regulations were frequently neglected or ignored.

The land reforms that have accompanied change at the end of the century have been guided by no such ideal of justice, but are regarded as an obligatory component in the process of privatization and the establishment of a market economy. While one of the defining features of the socialist city was the state expropriation and nationalization of land, the defining feature of the post-socialist society is the returning of land and property to private ownership. This includes the return to the former owners all forms of property, including agricultural land and urban real estate. This process of restitution constitutes one of the thorniest of problems faced by governments and societies. Countries have approached both the general policy of privatization and the specificity of restitution differently (Strong et al. 1996)

In the first case, Belarus, Kyrgyzstan, Uzbekistan, and China are among the firmly against land privatization, while in Russia a minority of the population favors the unrestricted buying and selling of land. And, even in countries where it is legal this has not necessarily meant the formation of a functioning land market. These policies are coming under pressure from external sources, with the European Union in 1999 declaring its opposition to requests from prospective membership candidates (such as the Czech Republic and Hungary) to continue their restrictions on foreign ownership of land. Through the selling of restituted buildings to banks and private companies and the conversion of more substantial houses into luxury flats, shops, and restaurants, restitution plays a distinct role in the stratification and shaping of post-socialist cities.

2.2.3 Housing. The interplay of a variety of factors, primarily the ideological imperative to develop industry rapidly, bequeathed highly polluted and degraded environments to post-socialist cities. The use of less

polluting fuels and the introduction of further legislation enforcing air quality control in the 1980s and then the dramatic fall in industrial output in the 1990s combined to reduce stationary sources of air pollution. On the other land, air pollution has increased as a result of the rapid rise in private car ownership and the poor quality of fuel available (Shahgedanova et al. 1999). Thus, there has been a compensatory rise in vehicular pollution, especially carbon monoxide discharge, which is most marked in the economically successful post-socialist cities. Thus, while the sources of pollution are changing, the general level remains the same.

Although the housing stock in socialist societies varied considerably in terms of ownership and physical structure, socialist housing policy rested on two principles: that no-one should draw an unearned income from renting space and that households should not enjoy the use of housing space above a certain, administratively defined norm. Larger properties belonging to more affluent citizens were taken into public ownership and then subdivided and leased at low rents. In most countries state agencies commissioned and financed new building, which they then allocated.

In 1988 the governments of the two largest socialist systems, China and the USSR, introduced legislation designed to transform housing policy. They signalled that rents in the public sector would have to rise and encouraged sitting tenants to buy their homes. In doing so they prepared the ground for the post-socialist city, one of whose defining features is the absolute right of the owner-occupier to dispose of property. The system of targeted housing allowances and the gradual introduction of market-level (or at least cost-covering) rents that is being introduced, is intended over time to make owner-occupation more appealing. This is unlikely to happen until there have been substantial increases in real incomes. The availability of construction materials on the market has meant an expansion in self-build. This not only augments the supply of housing but can also be used to generate income through (sub-)letting. Generally, however, social housing will, in one form or another, play an important role for the foreseeable future (Wang and Murie 1999).

Such has been the importance attached by some governments to the commercial housing sector as a powerful motor to drive the economy and urban development that, in China for instance, many of the 'private' developers are either outright public organizations or government–private partnerships. In many post-socialist societies, the decline in the role and authority of the central state has been accomplished by a growth in local government bureaucracies. But the parvenu princelings in the cities and provinces, who now have greater tax-raising powers, nonetheless lack the financial capacity to fulfil the functions devolved to them, which has meant a much-reduced expenditure on the technical and social infrastructure.

2.2.4 Real estate markets. One of the most defining and visible features of the post-socialist city has been the property boom and the proliferation of real estate agents, unknown in the socialist city. Requiring a minimum of knowledge and finance to be established, real estate agencies have proliferated in most post-socialist cities (and provided another lucrative source of income for organized crime). The immediate force driving the urban economy and responsible for transforming the class structure has been the buying, selling, renting, and construction of real estate. Its development is hampered by the fact that, after a decade of scandals associated with pyramid and other 'banking' schemes, individuals are wary of investing savings in private banks so that a mortgaging system is either totally absent or only in its infancy.

While much of the apparatus of town planning and building, with its institutional rhythms, careers, and entrenched interests, remains intact and ensures that the movement away from high-rise construction, based on the use of prefabricated units and the mass production of standardized parts, will take time, corruptible politicians and impotent and impoverished officials oversee the making of visual urban chaos.

Currently, a challenge is being mounted (from within the European Union) against the decade-long hegemony of the USA in the debate over the role of markets in housing and land (as expressed in Stryuk 1996 (Urban Institute, Washington), Bertaud and Rehaud 1997 (World Bank).

3. *Social Relations in the Post-socialist City*

3.1 *Population*

The war-ravaged societies, principally in the 'South,' have created demographically imbalanced cities. In Phnom Penh, 29 percent of households are headed by women. In Moscow, too, because of higher male mortality rates, women predominate. The average age is considerably higher in the cities of some countries, while in others the urban population is much younger. One visible sign of the gender imbalance is the greater presence on the streets of women engaging in petty trading or begging.

Radical transformations of society are accompanied by population movements and new residential configurations. The Soviet Union used an internal passport (*propiska*) system to regulate population flows into metropolitan and other large cities and to control their demographic and ethnic composition. After 1945 this method of control was exported as part of its 'administration package' to countries which adopted the Marxist model of development. Some post-socialist cities, notably Moscow, illegally retain it as an instrument of social control. Urban residents' associations in Ethiopia continue to regulate movement into their territories.

3.2 The Wealthy

Post-socialist societies have created a class of *nouveau riche*, who differ from élites under the previous system by an ostentatious display of their wealth status, especially through the universal symbols of home and private car, both of which make their mark on the shape of, and life in, the post-socialist city. Accommodation for élites under socialism was distinguished by the following features. First, élites lived in the city not out of town and had access to (but did not own) a country villa. Second, they did not live in socially exclusive enclaves, neither, third, did they live close to foreigners. This is in direct contrast to élite districts in post-socialist cities. Members of the new, privileged indigenous élite live in close proximity to foreigners often on small estates consisting of low-rise, detached housing surrounded by walls or fences covered by closed circuit television.

Views differ on the extent to which social segregation could be found in socialist cities (Szelenyi 1996; Prawelska-Skrzypek 1988; Dangschat 1987, Hamilton 1993, Weclawowicz 1992). The market economy of the post-socialist city intensifies any existing tendencies to segregation through the processes of gentrification and suburbanization. In the first case, a location in or near the city center, in buildings erected for prosperous families prior to the country's socialist revolution, is attractive to a variety of social groups. Alongside the modernization of old buildings is the construction of prestige housing and of retail and office space.

The arrival of international organizations has created an additional demand for higher quality housing to rent and purchase. The subsequent rise in the price of housing sometimes leads to the displacement of the local population and a reduction in density as pre-socialist residential buildings in Phnom Penh and Prague changed from multifamily to single-family occupation.

Suburbanization was described and denigrated as an anarchic, capitalist-driven sprawl associated with low-rise, single-family dwellings. Now, from Berlin to Moscow to Almaty to Beijing the flight of the rich to suburban and ex-urban settlements, usually along the main arterial routes, has begun in earnest. This development has been greatly assisted by the growth in car ownership. Both processes put pressure on protective green belts and invade parkland within cities.

3.3 Marginal Groups

In socialist cities apart from gypsies against whom sanctions were not strictly applied or were ineffectual, begging was unknown. Work was available for the able-bodied: for the elderly there were pensions which could be supplemented by other legal activities and the state provided for other 'vulnerable' groups, such as orphans and handicapped people. In the 1960s existing anti-parasite legislation was strengthened to deal with begging and vagrancy, described as anti-social(ist) behavior. As soon as the socialist city transmogrified into the post-socialist city begging, street children, and homelessness appeared as features of the urban landscape (Andrusz 1998, Lugalla and Mbwambo 1999). However, it is doubtful that in cities of less developed countries (such as Dar-es-Salaam and Phnom Penh) where—as South, post-socialist cities, public hygiene, especially the disposal of sewage and solid waste generally, and the provision of clean drinking and washing water remain the key issues—such groups are a wholly new phenomenon, although their numbers have increased. In the 'South,' the vast majority of street children come from the countryside, have inhabited the streets for a longer period, and have received no education at all. Their families of origin are much larger and very frequently polygamous and the mothers illiterate.

Socialist cities had their quotas of materially deprived and vulnerable groups, alcoholics, vagrants, and homeless people. But, unlike in post-socialist cities,they were rarely seen or talked about and did not form pressure groups or demonstrate.

Since, ideologically, such groups could not exist, neither did the organizations to help them; when they did appear the agents of social control dealt with them under the appropriate anti-parasite legislation. Today the visible presence of large numbers of dispossessed people has led to the establishment of a statutory and non governmental framework to address the situation of these individuals. However, even where statutory regulations exist, they are frequently not implemented. The institutional and ideological legacy has ensured that the law enforcement agencies continue to pursue a policy of harassment, intimidation, and abuse of position towards members of these groups. With the help of intergovernmental aid agencies and Western philanthropy, indigenous voluntary organizations are advocating that a more humane and positive approach be taken. But, in societies where most people are, or perceive themselves to be, suffering a deterioration in their standard of living, there is little public sympathy for beggars and special pleading indigents.

Another visible feature of the post-socialist city is that of the completed, but uninhabited luxury housing development, where supply has exceeded demand both because of the high asking price and because either the purchaser does not have the right to acquire the freehold or the period of the lease is too short. At the other end of the housing spectrum are the equally new phenomena of squatting and homelessness. The transition from socialist to post-socialist was quintessentially represented for a brief moment during *perestroika* when part of the space between the Hotel Rossiya and Red Square in Moscow was squatted by homeless people. The unplanned use of land reaches it apogee with squatter settlements in and around Asian and African post-socialist cities. In 1994, up to 15

percent of the population in Phnom Penh was living in such settlements.

Uninhabited, privately constructed, luxury housing on the one hand, and homelessness and squatting on the other, are features of post-socialist cities that were unknown in socialist cities.

4. Conclusion

The post-socialist city is a transparent city. Prostitution and poverty existed in St. Petersburg, Budapest, Beijing, and Havana. The ideology of socialism drew down a curtain on the urban stage concealing and reducing, but not eradicating social phenomena which the system decreed should not exist. It also banished the prosperous and powerful behind grey urban facades or to sylvan retreats. One of the imperatives of capitalism is that the homeless should be seen, disadvantaged groups should be allowed to express and demonstratively broadcast the inequities of their status, and the rich should parade their wealth.

See also: Cities: Cultural Types; Democratic Transitions; Development and Urbanization; Economic Transformation: From Central Planning to Market Economy; Socialism; Socialist Societies: Anthropological Aspects; Transition, Economics of; Urban Activity Patterns; Urban Geography; Urban Growth Models; Urban History; Urban Places, Planning for the Use of: Design Guide; Urban Planning: Central City Revitalization; Urban Studies: Overview

Bibliography

Andrusz G 1998 Housing and class structuration in Russia. In: van Weesep J (ed.) *Proceedings of the Conference Transformation Processes in Easter Europe. Housing and Labour Markers*. The Netherlands Organisation for Scientific Research, The Hague, Part 1 pp. 27–41

Andrusz G, Harloe M, Szelenyi I (eds.) 1996 *Cities after Socialism. Urban and Regional Change and Conflict in Post-Socialist Societies*. Blackwell, Oxford, UK

Bertaud A, Renaud B 1997 Socialist cites without land markets. *Journal of Urban Economics* **41**: 137–51

Dangschat J 1987 Sociospatial disparities in a 'socialist' city: the case of Warsaw at the end of the 1970s. *International Journal of Urban and Regional Research* **11**(1): 37–96

French R A, Hamiliton R E I (eds.) 1979 *The Socialist City. Spatial Structure and Urban Policy*. John Wiley, New York

Grabher G, Stark D (eds.) 1997 *Restructuring Networks in Post-socialism, Legacies, Linkages and Localities*. Oxford University Press, Oxford, UK

Hamiliton E 1993 Social areas under state socialism: the case of Moscow. In: Solomon S (ed.) *Beyond Sovietology: Essays on Politics and History*. M E Sharpe, New York

Harloe M 1996 Cities in the transition. In: Andrusz G, Harloe M, Szelenyi I (eds.) *Cities After Socialism. Urban and Regional Change and Conflict in Post-Socialist Societies*. Blackwell, Oxford, UK

Jezernik B 1998 Monuments in the winds of change. *International Journal of Urban and Regional Research* **22**(4): 582–8

Lugalla L, Mbwambo J 1999 Street children and street life in urban Tanzania: the culture of surviving and its implications for children's health. *International Journal of Urban and Regional Research* **23**(2): 329–44

Pensley D S 1998 The socialist city? A critical analysis of Neubaugebiet Hellersdorf. *Journal of Urban History* **5**: 563–602

Phe H H, Nishimura Y 1992 Housing in Hanoi. *HABITAT INTL* **15**(1/2): 101–26

Prawelska-Skrzypek G 1988 Social differentiation in old central city neighbourhoods in Poland. *Area* **20**(3): 221–32

Segre R, Coyula M, Scarpaci L 1997 *Havana. Two Faces of the Antillean Metropolis*. John Wiley, Chichester, UK

Shahgedanova M, Burt T P, Davies T D 1999 Carbon monoxide and nitrogen oxides pollution in Moscow. *Water, Air and Soil Pollution* **112**: 107–31

Smith D M 1996 The socialist city. In: Andrusz G, Harloe M, Szelenyi I (eds.) *Cities after Socialism. Urban and Regional Change and Conflict in Post-Socialist Societies*. Blackwell, Oxford, UK, pp. 70–99

Strong A, Reiner T, Szyrmer J 1996 *Transitions in Land and Housing. Bulgaria, The Czech and Poland*. St. Martin's Press, New York

Stryuk R (ed.) 1996 *Economic Restructuring of the Former Soviet Block. The Case of Housing*. Urban Institute Press, Washington, DC

Szelenyi I 1996 Cities under Socialism—and: After. In: Andrusz G, Harloe M, Szelenyi I (eds.) *Cities after Socialism. Urban and Regional Change and Conflict in Post-Socialist Societies*. Blackwell, Oxford, UK, pp. 286–317

Wang Y P, Murie A 1999 Commercial housing development in urban China. *Urban Studies* **36**(9): 1475–94

Weclawowicz G 1992 The socio-spatial structure of the socialist cities of East-Central Europe. In: Lando F. (ed.) *Urban and Rural Geography*. Cafoscarina, Venice, pp. 129–40

Wu V 1998 The Pudong development zone and China's economic reforms. *Planning Perspectives* **13**: 133–65

Wu W P 1999 Shanghai. *Cities* **16**(3): 207–16

G. Andrusz

Citizen Participation

'Citizen participation' refers generally to citizen involvement in public decision making. In planning and related fields, the term also has a specialized meaning, designating efforts to facilitate participation of citizens who would normally be unable or disinclined to take part.

1. The Emergence of an Ambiguous Definition

Alexis de Tocqueville, the chronicler of American habits in the mid-nineteenth century, remarked on widespread popular participation in the country's civic life. Americans seemed ready to form groups to address any problem. When mass immigration swelled cities at century's end, citizen activists spearheaded wide-ranging progressive reforms in sanitation, land use, and government organization. American city

planning was a direct outgrowth of this citizen movement. Planning commissions were one means of institutionalizing public influence over decisions.

Nevertheless, the 1960s heard an explosion of talk about 'citizen participation' in planning, Community Action and the War on Poverty, and Model Cities. Part of the explanation is that citizens turned away from public life in the quiescent 1950s, but the primary reason is that the Johnson administration policy agenda focused on citizens who had never participated much in public decision making: the poor and blacks. Community Action would involve them in unprecedented ways.

The traditional approach to citizen participation in planning was represented by the federal Housing Act of 1954, which called for citizen advisory committees in urban renewal and other local projects. A 1966 Department of Housing and Urban Development Program Guide suggested categories to be represented, including business groups, civic clubs, churches, schools, government agencies, social service organizations, the mass media, neighborhood organizations, and 'ethnic or racial groups.' The examples and custom made it likely that, with the possible exceptions of neighborhood and racial representatives, most advisory committee members would come from an educational and economic elite.

The Economic Opportunity Act of 1964, the hallmark of the War on Poverty, declared a new approach: 'The term 'community action program' means a program ... which is developed, conducted, and administered with the maximum feasible participation of residents of the areas and members of the groups served.' Two years later, Model Cities began with a consonant call for 'widespread citizen participation.' One sign of a new departure was the Office of Economic Opportunity's investment in community organizing as part of Community Action and its technical assistance to community organizations in Model Cities. The War on Poverty aimed to activate the urban poor, particularly blacks.

'Citizen' participation and 'community' action or participation differ in an important way. A citizen is an individual, whereas a community is a collectivity. Citizens are members of the state, with certain rights, but they do not necessarily have intervening attachments. Communities are groups with membership norms (ranging from residence in a territory to conformity to rules of behavior), social relations, and loyalties. Participation of individual citizens is a quantitative matter: it can be counted, and more is arguably better than less.

In contrast, participation of communities is a qualitative matter. Communities do not take action en masse, but, rather, individuals act in the name of communities. This situation introduces questions of representation: What is a community, who are its members, and to what degree do participants represent members' interests? Partly a matter of calculation, this is largely a matter of judgment, which community members can render and about which many have strong convictions. Thus federal and local officials who encouraged or coped with citizen participation faced community activists who challenged them and forced consideration of what would be representative participation of citizens who are community members.

In addition, community activists, stimulated by the language of the War on Poverty, raised issues of power: could participation of residents in developing, conducting, and administering programs mean anything other than community control? And, if so, what should residents have control over, and how much control was significant? Answers to these questions were complicated by ambiguities in definitions of community and representation: who would have to exercise power over what in order for a community to be represented in making decisions?

In the early 1970s, the Nixon Administration narrowed the meaning of citizen participation in two respects. 'Citizens' became individuals again, rather than community members. 'Participation' less often meant activism or power than opportunities to speak at formal hearings conducted by others, who would make decisions. Over the years, federal planning, housing, community development, and environmental legislation has come to include provisions for citizen participation, and local practice has followed suit. In general, the balance of these practices has taken the narrower view of citizen participation.

Planning in Europe and Australia has adopted language of citizen participation, with variations in practice similar to those in the United States. United Nations and World Bank policies speak of 'community participation' in aid projects. In this context 'community' refers to the locus or target of intervention more than an expectation that an entire community participate in planning and implementation. Practice encounters similar issues about representation and power to American and European planning.

2. *What 'Citizen Participation' Could Include But Often Does Not; A Broader Definition*

Because 'citizen participation' originally referred to involving people in government-initiated programs, common usage often does not include autonomous citizen activities, even though they are acts of involvement in the society and may aim to influence government. This broader 'citizen participation' includes three types of activity commonly associated with 'community organization.'

'Locality development,' or 'community development,' involves the organization of residents of an area to create the capacity to improve their situation. Projects may address various needs, such as health, housing, education, and infrastructure. Activities may

include popular education to understand local conditions.

'Social planning' involves citizens in working with technical data to solve substantive problems. These activities aim to prepare plans for programs to address needs, which may be in many areas. Participants might focus on a single issue, or a community might develop a multifaceted plan.

'Social action' involves the organization of a population to press actors in the larger society, usually institutions and often government, to change policies or practices, redistribute economic or social resources, or yield power over these things. Disadvantaged populations may favor this approach, but many issue-oriented groups (such as the environmental movement and consumer groups) may choose it. Activities include mass mobilization and advocacy.

It is reasonable to consider these activities part of a broad definition of 'citizen participation' as efforts by citizens (including, but not only, the poor and minorities) to influence the policies and practices of government, basic social institutions, or their own neighborhood or community. However, in this context the conventional emphasis on efforts to involve citizens in government-initiated activity produces theoretical ambiguity. In addition, those who hold the narrower view may find conflicts in practice when encountering citizens or communities who hold the broader view.

3. *Arguments for Citizen Participation; Practical Benefits*

Different views of citizen participation are associated with different arguments for it and claims regarding its benefits. A basic distinction can be drawn between a focus on the benefits to government and other service providers and a focus on the benefits to citizens and communities. The first view emphasizes the ways citizens contribute knowledge and, especially, authority, such that program operators can be more confident citizens will use and benefit from services. The second view emphasizes the ways citizens get more power and economic and social resources that promote their development. The first view is more likely to be associated with the narrower view of citizen participation, and the second with the broader view, but the two may overlap and the correspondence is imprecise. Four types of benefits can be identified.

The first emphasizes *individual development*. Citizen participation has been seen as therapy for the alienated and powerless; involvement in political activity may contribute to self-esteem, a sense of potency, and growth. Inherently, participation in public life makes someone a citizen. Collaboration helps individuals become part of networks that provide social supports. Instrumentally, participation can bring power, knowledge, the ability to solve problems, and improvements in living conditions, including services.

The latter benefits involve sharing in outcomes of collective action, in *community development*. When members of a community articulate and promote their interests with one another, they can resolve conflicts and discover shared community interests. They will develop relationships they can use in acting together on behalf of those interests. As a result, they may gain knowledge, power, resources, and relationships that enable them to influence institutions and other actors and make decisions that change conditions and solve problems. This may mean having more power over societal resources, becoming more self-sufficient, or both. In general, a community whose members deliberate and make decisions together is likely to have a sense of vitality and potency and attract members' loyalty.

Organizational development is a third benefit of citizen participation. When citizens contribute knowledge about their needs, service providers are more likely to design and deliver services that meet those needs, solve problems, and are used. When citizens participate in planning programs, they are more likely to develop relationships with organizations, consider programs legitimate, and use or accept them. In general, voluntary organizations, including community associations, need participation because citizens provide three resources essential to organizational operation and development: work, money, and legitimacy.

Finally, citizen participation contributes to *societal development*. Active citizenship in a democratic framework can be considered inherently good for society. When activists openly discuss their interests, they can resolve conflicts among private interests and discover common, public interests. They can build relationships and learn together about shared conditions. As in communities, citizen participation offers society the possibility of articulating and solving problems knowledgeably and legitimately, with the result that collective actions, including programs, serve more people well.

Some proponents of citizen participation portray it as a means to various ends, whereas others, particularly those concerned with societal development, present it as an end in itself. It can be both, and the delineation of means and ends varies with focus on individuals, communities, organizations, or society.

4. *Specific Purposes of Participation Activities; The Issue of Power*

The discussion of benefits from citizen participation identifies many possible purposes for activities in which citizens participate in public life. These can be classified as follows:

(a) Communicating information (including perceptions, beliefs, opinions, hopes, expectations, and intentions);

(b) Developing relationships (creating new ones and strengthening existing ones);

(c) Developing the capacity to act and organizing action (including organizing coalitions, planning, strategizing, and creating and exercising power);

(d) Preserving or changing selves and/or others (including policies, practices, conditions, and relationships).

These categories may be seen hierarchically, in that communicating information builds relationships, which in turn develop a capacity to act, which may effect change in some or all parties. However, in practice and in the abstract, the categories overlap: for example, communication may contribute to developing relationships, the creation of which is change and which have the capacity to bring about other change, such as further communication.

In short, citizen participation can enable citizens and others who participate with them to learn and can empower them to act together in ways that would have been impossible otherwise. However, when citizens or those they interact with believe they have conflicting interests, they may participate with other purposes. They may seek to prevent the communication of information, the development of relationships, action by others, and any change in their own practices. Under these conditions, when government officials or funders, for example, organize citizen participation, they may involve citizens but limit their influence.

Sherry Arnstein (1969) analyzed this central question of citizens' power in terms of 'a ladder of citizen participation,' ranging from low to high power:

Nonparticipation

(a) Manipulation: placing citizens on powerless advisory committees for the purpose of 'educating' them or gaining their support for others' positions;

(b) Therapy: trying to change citizens' beliefs, attitudes, or values as a diversion from analyzing and changing social conditions;

Degrees of tokenism

(c) Informing: giving information to citizens but not requesting or listening to their points-of-view;

(d) Consultation: asking citizens' opinions but not making a commitment to act on them;

(e) Placation: placing a few citizens on committees or boards that have limited influence over policies and practices;

Degrees of citizen power

(f) Partnership: sharing decision making power with citizens;

(g) Delegated power: delegating some decision making power to citizens;

(h) Citizen control: citizens have considerable independent power over decisions.

This hierarchy focuses on power over overt decisions, such as policies or programs that address problems. This is the most familiar of three 'levels' of power. Beneath it are decisions about the issues or problems that go onto the agenda for action. It is possible for citizens to have considerable power in selecting from among alternatives on the table without having influence in determining the array of possibilities. Still one level lower are decisions, frequently tacit, about which social conditions are defined as issues or problems and even considered for the agenda. Citizens may have influence in choosing among recognized issues for a group or organization's agenda but have little power to draw public attention to certain conditions that trouble them greatly. The purposes of citizen participation may pertain to each of these levels of power.

5. *Means of Citizen Participation*

The means of citizen participation are related to its purposes. Some means serve single purposes, but most may serve several purposes. Means can be grouped in five general categories.

Groups, Organizations, Committees, Boards, Councils, and Other Institutions are entities that offer arenas where members and others can deliberate, make decisions, and otherwise act. Examples are community organizations and planning councils, agencies that deliver services or administer programs, planning commissions and other boards that guide or regulate public action, citizen review boards, and task forces. All are concerned about and intervene in public life. They vary in the degree to which they involve citizens in activities characterized by purposes high on Arnstein's ladder.

These entities conduct *meetings* where participants can communicate information, develop relationships, develop the capacity to act, and act in ways that preserve or change people or things. Meetings may be formal or informal, regularly scheduled or *ad hoc*, generally oriented toward organizational affairs or focused on specific issues or problems. Meetings may have single or multiple purposes, such as discussion, planning, or decision making. Some organizations restrict citizens to discussion, whereas others may involve them in planning or decision making as well.

These entities can sponsor *inquiries*, events, or services for one-directional communication of information. Information centers allow citizens to ask for information. Activities in which citizens provide information include public hearings, surveys, focus groups, various structured group activities, referenda, and elections. Organizations vary in their interest in giving information to citizens or receiving it from them.

These entities can sponsor such other *action* as organizing people and developing and implementing interventions. Examples include place- or issue-based organizing, housing development, school reform, and

health care provision. Line departments in state and local government, community organizations, and community development corporations typically sponsor these activities. They vary in the extent to which they require or can accommodate lay citizens, rather than experts. Organizations vary in efforts to involve citizens in these activities.

These entities or others can sponsor *technical assistance* to increase citizens' ability to join the entities, participate in meetings, make effective inquiries, and take part in action. Technical assistance includes classes or workshops, as well as *ad hoc* or ongoing consultation or mentoring. 'Advocate planners' work with or represent citizen groups in planning. Topics for assistance include means of participating in organizations such as using parliamentary procedure, engaging in public discussion, making decisions, and exercising leadership; technical and analytic tasks such as budgeting, scheduling, and grant-writing; managerial tasks such as running and organization and planning; and knowledge about substantive fields.

6. *Participation and Representation*

6.1 *Meanings of Representation*

When 'citizen participation' refers to individual citizens, each person can be said to represent him- or herself, and participation and representation are equivalent. Both can be measured quantitatively, in terms of how many people or what percentage of members of some unit, such as a geographic neighborhood, participate.

In contrast, when 'citizen participation' refers to a collectivity, such as a community or a racial or cultural group, representation is more complicated and, while related to participation, separate from it. Representation might be measured as a ratio of participants to total members, but to do so would not take into account the characteristics that define the identity of the group, distinguish subgroups from one another, and demarcate various interests. One common shortcut is to identify certain demographic characteristics of the whole—such as race, sex, income, and age—and seek participants who resemble the overall population in these respects. However, this approach assumes that all blacks, women, poor people, or elderly are similar and that what matters about and to them is a simple product of their race, sex, income, and/or age.

There is no single correct way to define or measure representation. Rather, it is useful to consider in a particular instance what aspects of the people involved matter and how they might be represented. For example, in a neighborhood with bad housing, groups may form around distinct positions on improving conditions; their views may be unrelated to race or income, and associated mainly with whether households own or rent and whether they have children. Representation could be defined primarily in terms of positions on issues and secondarily in terms of housing tenure and family composition. Determination of what interests, positions, or characteristics should be represented is consequential because it affects not only which individuals participate, but also which positions are given voice.

6.2 *Differences in Ability to Participate*

Nevertheless, even if there is agreement about how many and which participants might represent an overall population, some groups can much more easily participate than others. Citizen participation depends on motive, opportunity, and means. The three are related.

A common motive, particularly with regard to participation in activities sponsored by government or other large institutions, is the possibility of influencing conditions, for the benefit of a private or a public interest. A strong motive, that often leads citizens to create or activate an organization, is the desire to resist an action seen as a threat, such as urban renewal, highway construction, or massive development.

When participation opportunities arise, individuals may take part for the pleasure of interacting with others—either specific persons or a group in general. At the same time, they may assess whether these are opportunities to exercise power or accomplish meaningful ends. Government activities, for example, may be linked to formal power but lie beyond local citizens' influence; a community organization may offer individuals considerable opportunity for neighborhood influence but have little control over important outside forces.

Citizens vary in their ability to take advantage of these opportunities. Time is a basic resource and constraint. Those who work full time, parents, and, in particular, single parents have little time for meetings. Middle-class citizens, especially those who have a flexible schedule, do not work full-time, or are supported by a spouse, and particularly those without children at home, are most likely to have time to participate in public life.

Thus citizens' motivations rest on a calculation of the returns on an investment of time. Different citizens are likely to figure the possibilities for influence differently, based on their experience, relationships, skills, and confidence. The means at their disposal affect their ability to take advantage of opportunities. Participation is aided by a cognitive style that allows seeing how the particulars of a moment fit into a big picture—geographically, socially, and temporally. Planning depends on imagining a medium- to long-term future and having confidence in one's ability to affect it. Substantive knowledge matters, both for making informed decisions and for having the confidence to take part. Procedural expertise is crucial,

including knowing how to organize people, run a productive meeting, resolve conflicts, and direct discussion toward agreement. Social relations are an important source of work, influence, and knowledge—about issues and participation opportunities.

The conditions of middle-class life are particularly likely to give citizens these means and motivate them to see and take advantage of opportunities for participation. Extended formal education, professional work, intricate social networks, and general success in influencing the world prepare middle-class citizens to participate in ways that elude many who are poorer.

These conditions encouraged the creation of remedial 'citizen participation' activities. Extensive, effective participation of low-income citizens depends on overcoming these obstacles. Time is a difficult constraint to change, but linking civic events with activities in which people already participate is important, as is child care. Training and experience can give people skills and confidence to organize and take an active part in public affairs. Crucially, low-income citizens, as any citizens, are more likely to participate when they see real opportunities for power.

7. *Evaluating Citizen Participation Activities*

Does citizen participation work? Do some kinds of citizen participation work better than others? How does narrowly defined 'citizen participation' compare with the rest of broadly defined citizen participation? Is citizen participation worth the effort—for citizens, for the government, for anyone else?

These are good, crucial questions. Yet they are difficult to answer. First, terms must be defined. Second, little of the relevant empirical evidence is collected in writing. Hence available data offer only specimens of answers from a larger undetermined universe of possibilities. Third, in any case, the most precise answers are contingent: it depends.

7.1 *Defining the Questions*

Whether citizen participation 'works' depends on expectations. Most generally, citizen participation is a strategy for solving a problem. Commonly, problems are defined substantively: for example, the poor are unemployed, low-income city residents lack affordable food shopping, or public health programs do not reach low-income children. Alternatively, some problems may be defined procedurally, and many substantive problems, in fact, have procedural roots: for example, the city health department, local hospitals, the state Medicaid division, and low-income health advocates cannot agree on an approach for taking care of indigent children.

One might ask whether citizen participation in planning for such problems contributed to better solutions than would have been likely otherwise. However, because there cannot be a contrasting case where everything was the same except for citizen participation, it is impossible to answer that question. Moreover, the purpose, or at least the effect, of citizen participation may be to redefine the problem. In particular, in the case of procedural problems defined by the exclusion of citizens, citizen involvement in itself may provide a solution. Hence it is reasonable simply to study events to assess the influence of particular actions by specific people.

Still, citizens may participate in activities that solve problems without themselves having influence over the results. Tensions between the narrow and broad notions of 'citizen participation,' as well as conflicts over substantive and procedural issues, will lead others to try to limit the power of certain citizens. To evaluate the effects of citizen participation, it is necessary to examine the roles and influence of particular groups of citizens at each level of power.

Finally, citizens may participate powerfully in solving problems, but those who take part may not represent a larger group in whose name they speak. An elite may deliberately try to corner power for themselves. Alternatively, a small group of activists who have the skills, time, and connections to participate in wide-ranging activities may, with the most generous of intentions, take leadership, work closely together, succeed in advancing projects, develop proprietary feelings for them, and continue along the easy path of standing for a larger group. They may succeed in solving problems for members of their larger community, who can be said to be represented by the outcomes, but not in the process.

Thus the question of whether citizen participation works, complexly combines issues of effectiveness in solving problems, power, and representation. Still, answers to the question do not affect the normative position that citizen participation in public affairs is a democratic right.

7.2 *The Haphazardly Collected Empirical Data*

Few citizens record their participatory experiences, probably because few have both the interest and time. Perhaps those who consider themselves successful are most likely to document their activities or draw the attention of observers who will do so. For these reasons there is no way of knowing what universe reported cases represent.

Moreover, it is unclear what should be considered a sufficient account of a case. While many case studies focus on specific 'citizen participation' activities, their context shapes their effects. Hence it would be informative to see the entirety of influences on citizen participation in a case, but it may be uncertain at what distance from the 'main action' to draw boundaries around the case. Within those boundaries there is the question of what the unit of analysis should be—what should be taken as the unit of citizen participation that

is to be measured for its effects. It might be an incident, an individual or collective actor, a relationship, a strategy, a tactic, another action, or a combination of these things. Further, it is unclear what characteristics of these elements might matter. For example, when might citizens' age, education, family composition, housing tenure, income, length of residence, occupation, place of residence, race, religion, or sex, be pertinent to their participation in public affairs and the effects of their activities? And how would the context of the case affect the influence of the unit of citizen participation?

These questions are conceptual but also empirical, dependent on data collection and analysis. Little available material is robust enough to address all these questions, and, as a result, it is difficult to compare the effects of citizen participation in different contexts. These questions should direct the recording of cases. For now, the available data should be interpreted as complexly as possible. They offer specimens—examples of what citizen participation may accomplish—which could become the basis for generalizations.

One can find examples of success and failure. There are instances where organized citizens planned and developed programs. There are instances where organized citizens stopped projects they opposed. In these cases, citizens have benefited from such resources as organization, money, effort, skill, knowledge, planning, strategy, alliances, time, and commitment. These findings are consistent with political science research.

See also: Citizenship and Public Policy; Civil Society, Concept and History of; Community Power Structure; Community, Social Contexts of; Community Sociology; Neighborhood Revitalization and Community Development; Participation: Political; Political Representation; Poverty Policy; Poverty, Sociology of; Urban Poverty in Neighborhoods

Bibliography

Arnstein S R 1969 A ladder of citizen participation. *Journal of American Institute of Planners* **35**: 216–24

Baum H S 1997 *The Organization of Hope*. SUNY Press, Albany, NY

Checkoway B 1995 Six strategies of community change. *Community Development Journal* **30**: 2–20

Checkoway B, Pothukuchi K, Finn J 1995 Youth participation in community planning: What are the benefits? *Journal of Planning Education and Research* **14**: 134–9

Fisher R 1984 *Let the People Decide*. Twayne, Boston

Forester J 1989 *Planning in the Face of Power*. University of California Press, Berkeley, CA

Frieden B J, Kaplan M 1975 *The Politics of Neglect*. MIT Press, Cambridge, MA

Glass J J 1979 Citizen participation in planning: the relationship between objectives and techniques. *Journal of the American Planning Association* **45**: 180–9

Heskin A D 1991 *The Struggle for Community*. Westview Press, Boulder, CO

Hinsdale M A, Lewis H M, Waller S M 1995 *It Comes from the People*. Temple University Press, Philadelphia

Marris P, Rein M 1982 *Dilemmas of Social Reform*, 2nd edn. University of Chicago Press, Chicago

Medoff P, Sklar H 1994 *Streets of Hope*. South End Press, Boston

Midgley J, Hall A, Hardiman M, Narine D 1986 *Community Participation, Social Development and the State*. Methuen, London

Mogulof M B 1970 *Citizen Participation: A Review and Commentary on Federal Policies and Practices*. Urban Institute, Washington, DC

Moynihan D P 1970 *Maximum Feasible Misunderstanding*. Free Press, New York

Narayan D, Ebbe K 1997 *Design of Social Funds; Participation, Demand Orientation, and Local Organizational Capacity*. World Bank Discussion Paper No. 375. World Bank, Washington, DC

Oakley P, Marsden D 1984 *Approaches to Participation in Rural Development*. International Labour Office, Geneva, Switzerland

Rosener J B 1975 A cafeteria of techniques and critiques. *Public Management* **57**: 16–19

Rothman J, Tropman J E 1987 Models of community organization and macro practice perspectives: Their mixing and phasing. In: Cox F M, Erlich J L, Rothman J, Tropman J E (eds.) *Strategies of Community Organization*, 4th edn. Peacock Publishers, Itasca, IL

Spiegel H B C 1968 *Citizen Participation in Urban Development;* Vol. I *Concepts and Issues*. NTL Institute for Applied Behavioral Science, Washington, DC

Stoecker R 1984 *Defending Community*. Temple University Press, Philadelphia, PA

H. S. Baum

Citizenship and Public Policy

The conceptualization of the relationships between citizenship and public policy depends critically on how these terms are defined, and in turn on the sociopolitical context in which they were developed and are located. In the social sciences, these concepts are the products of different histories and significance within Europe and North America. In the European context, public policy is directly connected with social citizenship in the historical context of the emergence of the welfare state and industrial capitalism. Social citizenship can only be understood in the broader historical context of social class relations, the modern capitalist economy and the nation state (Turner 1986). For radical sociologists, citizenship and public policy were state strategies to secure the political compliance of the urban working class (Mann 1987). In the USA, citizenship was originally defined in political terms by reference to the individual rights that followed from the War of Independence and the framing of the Constitution. American citizenship was essentially about nation-building, that is, about the creation of a 'people' in relation to the aspiration of political leaders

(Shklar 1991). In the formation of the American people, political leaders created diverse 'civic ideals' (Smith 1997) that blended liberal, democratic republican, and inegalitarian ascriptive elements. In the late twentieth century, following the civil rights movements and intensive waves of immigration, social citizenship is related to government policies towards naturalization, ethnic integration, and multiculturalism. It is difficult to think of American citizenship without thinking of the 'American dilemma' of race relations (Myrdal 1944). Because the historical development of the welfare states in Europe and America is fundamentally different, the terms of debate have different meanings, functions, and significance. However, with globalization there is some convergence between these different traditions as nation states confront cultural hybridity, a global labor market, and the partial erosion of national sovereignty.

Conventional forms of citizenship were associated with the modernization of society and with the development of the administrative framework of the modern states. In the nineteenth century, for example in the public policies of Bismarck, there emerged a close relationship between nationalism, social insurance, and state formation. The formation of national identity, political integration, and citizenship were aspects of the modernization of state administrations, and citizenship came to form a basis for the convergence of national and masculine identity (Nelson 1998). Nineteenth-century nationalism meant that under the orchestration of the nation state, public policy was also cultural policy. However, with globalization, the sovereignty of the state has been compromised, and the modern debate about policy and citizenship has to be set within the political constraints of a global economy.

The concept of citizenship in British social theory has passed through several stages, from the idealism of T. H. Green to the welfare theories of T. H. Marshall and Richard Titmuss. Marshall's *Citizenship and Social Class and Other Essays* (1950) and *Social Policy in the Twentieth Century* (1967) have dominated recent debate about social citizenship. The principal question in this tradition has concerned how far a comprehensive welfare policy can be effectively implemented in a capitalist system without either destroying the profitability of capitalist enterprises through excessive taxation or compromising the principles of redistributive justice (Loney et al. 1983). Skeptics of the Marshallian approach have claimed that welfare benefits contributed more to the wellbeing of the middle class than the working class. The contemporary British welfare debate has concentrated on the question of European union, the effects of a centralized bureaucracy, and the possibility of the implementing a social wage.

These policy debates can be contrasted with the American tradition. One might argue controversially that, while the European policy debate has been about the possibility of equality in a capitalist economy, the American legacy, drawing on the work of both Alexis de Tocqueville (1966) and John Dewey (1963), has been concerned with electoral democracy and political access. The Tocquevillian position has been to emphasize the contribution of churches, voluntary associations, and community initiative in the delivery of public policy. However, the distinctively American contribution to public policy has been from pragmatism which, from John Dewey to Richard Rorty (1998), has placed its policy aspirations in the role of education for access and participation (Diggins 1994). It is also important to recognise that philanthropy has played a much larger role in policy development and delivery in America than in Europe (Wuthnow 1996).

In general, citizenship establishes the broad legal and social parameters within which public policy is set, but public policy creates the administrative and legislative framework within which citizens can effectively enjoy their rights. Historically, institutionally, and analytically, citizenship and public policy are interconnected and interdependent. Citizenship is a collection of rights and obligations that gives individuals a formal juridicial identity. Social citizenship involves social membership, a distribution of rewards, the formation of identities, and a set of virtues relating to obligation and responsibility. It is constituted by social institutions such as the jury system, parliament, and welfare states. Whereas the history of political institutions identifies the origins of citizenship with the Greek *polis*, social citizenship, as the outcome of the American and French Revolutions, is an essentially modern conception. It is analytically a product of modern liberal theory, specifically de Tocqueville's theory of democratic association.

The social rights of social citizenship are very different in their consequences in contrast with legal and political rights. Social rights require the provision of social services and transfer payments, and as a result involve the state in public expenditure. They also require some administrative structure to deliver these services, and hence further involve the state in fiscal management. Recognition of this fiscal role of the state gave rise to the famous division of welfare by Titmuss (1958) into social welfare (education, health, and social services); fiscal welfare (allowances and relief from taxation); and occupational welfare (benefits received by employees through employment). This provision of social rights clearly illustrates the tensions and contradictions between the state and the market. Radical sociologists like Jürgen Habermas (1976) predicted a 'legitimation crisis,' because the growth of state expenditure in response to electoral pressure would necessarily undermine capitalist profitability. One can argue therefore that the neoliberal policies of the governments of Ronald Reagan and Margaret Thatcher involved 'rolling back the state,' the privatization of welfare, and a return to third-sector involvement through charities, philanthropic in-

stitutions, and the voluntary associations. These strategies sought to restore profitability through deregulation, subsidiarity, and community initiatives, but also emphasized the importance of individual and family responsibility for welfare delivery. Conservative critics of bureaucratic welfare argued that welfare benefits undermined the family through payments to single parents, who as a result had no necessary incentive to work and clear incentives not to marry. In the late twentieth century, so-called 'third-sector strategies,' which encouraged local initiative, community development, and voluntary sector delivery, became fashionable, not only in North American and Europe but also in Australia and New Zealand (Brown et al. 2000).

The notion of 'public policy' also requires some preliminary clarification. It is a general term to describe the efforts of governments to coordinate the provision of a variety of governmental services and utilities. Public policy expresses the political intentions and choices of government, and creates the framework within which social planning takes place. 'Public policy' is often referred to as 'social policy,' typically when a broader philosophical dimension is involved. Social policy is often distinguished from public policy by its broad welfare dimension; it attempts to regulate the provision of five social services: housing, education, health, social security, and personal social services. The connection between policy and citizenship was overtly expressed by Marshall, who claimed that social policy includes the general policy of governments with regard to action having a direct and explicit impact on the welfare of citizens by providing them with services or income. Social Administration is the science of the provision of such services through social policy, and in Britain it was established as a discipline at the London School of Economics in 1912. In Europe in the 1840s, the growth of social medicine expressed the idea that the health of individuals should involve the state in the policing of society. Public policy and social citizenship have to be understood as aspects of the growth of state administration designed to bring about social order in a world of expanding capitalism.

For detailed treatment of these and related fields see *Health Policy*; *Immigration: Public Policy*; *Industrial Policy*; *Policy History: Regimes*; *Poverty Policy*; *Public Health*; *Social Insurance: Legal Aspects*.

1. *Citizenship and the Rise of Public Policy*

Both citizenship and public policy are modern institutions. Neither social citizenship nor public policy could exist without the nineteenth-century expansion of public administration and law. In pre-modern times, it is clear that the church responded to the satisfaction of need through principles of reciprocity and redistribution. Ecclesiastical responses to poverty and need in the Middle Ages were located within a theological debate about economic exchange in which low wages were interpreted through a moral discourse on justice. The church developed various institutional responses to poverty and created the modern framework of charity (Troeltsch 1912). Before the development of the capitalist market system, the poor could rely on custom and communal organizations for subsistence. In the medieval period, guilds and fraternal associations in the towns provided welfare benefits to members and their dependants. These guilds often evolved into mutual societies and the principles of mutuality often survived the growth of capitalism. In traditional societies, it was the church, the patriarchal family, or the landlord who determined the conditions of survival, rather than the labor contract.

We can see the origins of public policy as a social struggle over the conditions by which markets function. In his classic study of *The Three Worlds of Welfare Capitalism*, Esping-Andersen (1990, p. 35) claimed that 'the mainsprings of modern social policy lie in the process by which both human needs and labor power became commodities and, hence, our wellbeing came to depend on our relation to the cash nexus.' However, the impact of the market is modified or moderated by the growth of social citizenship which provides, through public policy initiatives, a variety of safety nets to protect the individual from the full vagaries of unemployment, sickness, and disability. With the decline of traditional patterns of social reciprocity, citizenship expressed through public policy reduces the contingencies of dependence on the market.

The historical controversy in the social sciences has been around the capacity of public policy and citizenship to bring about an effective redistribution of resources. There are broadly two views about the historical role of welfare states. Radical critics of capitalism (O'Connor 1973, Piven and Cloward 1972) have argued that public policy aimed at creating social citizenship is primarily concerned with legitimating governments and protecting them from political instability arising from social conflicts, specifically from a revolutionary working class. Liberal theories of welfare (Lipset 1960) argue that welfare policies reduce the material causes of class conflict, incorporate the working class and enhance democratic access to the state. For sociologists like Frank Parkin (1979), social welfare transforms class conflict into status competition. In broad terms, one can conclude that the production of inflationary pressure of welfare benefits from public policy initiatives has been regarded by the state and its elites as more tolerable than civil conflict and revolutionary protest. Whether or not public policies to institutionalize social citizenship have a real effect on social equality will depend on what type of welfare state (if any) is created by government actions,

and how these policies have emerged historically in relation to different forms of capitalism.

In recent scholarship, the radical or Marxist paradigms have been less influential, and debate about public policy and government agency has been influenced by the work of Michel Foucault. Social scientists have in particular recognized the importance of Foucault's concept of 'governmentality' as a paradigm for understanding the microprocesses of administration and control within which self regulation and social regulation are united (Foucault 1991).The concept of 'governmentality' provides an integrating theme that is concerned with the sociopolitical practices or technologies by which the self is constructed. 'Governance' or 'governmentality' refers to the administrative structures of the state , the patterns of self-government of individuals and the regulatory principles of modern society. Foucault argued that governmentality has become the common foundation of all forms of modern political rationality, that is, the administrative systems of the state have been extended in order to maximize the state's productive control over demographic processes. This extension of administrative rationality was first concerned with demographic processes of birth, morbidity, and death, and later with the psychological health of the population. 'Governmentality' can be seen as an administrative rationality that produces the modern self as a consequence of social services. This perspective has been valuable in understanding the growth, for example, of social gerontology as a science, a component of public policy, and as the basis for new professions to discipline the elderly (Katz 1996).

Foucault's historical inquiries gave rise to a distinctive notion of power, in which he emphasized the importance of its local or micro manifestations, the role of professional knowledge and expertise in the legitimation of such power relationships, and the productive rather than negative characteristics of the effects of power. His approach can be contrasted usefully with the concept of power in traditional Marxist sociology, where power is visible in terms of the police and army, concentrated in the state, and ultimately explained by the ownership of the economic means of production. In the Marxist perspective, power is typically negative and signifies a system of institutions that contain, prohibit, and control. Foucault's view of power is more subtle, with an emphasis on the importance of knowledge and information in modern means of surveillance.

'Governmentality' is the generic term for these power relations. It was defined as 'the ensemble formed by the institutions, procedures, analyses and reflections, the calculations and tactics, that allow the exercise of this very specific albeit complex form of power, which has as its target populations'(Foucault 1991, p. 102). The importance of this definition is that the power of the state in the modern period has been less concerned with sovereignty over things (land and wealth) and more concerned with maximizing the productive power of administration over population and reproduction. Furthermore, Foucault interpreted the exercise of administrative power in productive terms, that is enhancing population potential through, for example, state support for the family. For example, the state's involvement in—and regulation of—reproductive technology is an important example of governmentality in which the desire of couples to reproduce is enhanced through the state's support of new technologies. The existence of a demand for fertility is supported by a profamilial ideology that regards the normal household as a reproductive social space. These Foucauldian perspectives have been useful in providing an historical understanding of the relationships between family, state, and public policy. Feminist criticism of neoliberal policy to protect the family has noted that these policy initiatives implicitly or explicitly ignore the presence of married women in the labor force (Wilson 1977).

2. *The Dimensions of Citizenship*

In historical terms, social citizenship and public policy have been shaped by the relationship between state and market. Public policy attempts to promote a set of general social conditions through which effective entitlement to social resources can be sustained. Welfare has, in practice, never been an unconditional right; entitlements have, in reality, been tied to contribution. The entitlement to benefits in liberal welfare systems have typically been through work, war, and reproduction.

Work was fundamental to the conception of citizenship in the welfare state as described in W. H. Beveridge's *Social Insurance and Allied Services* (1942) and *Full Employment in a Free Society* (1944). Individuals could achieve effective entitlements through the production of goods and services, namely through gainful employment, which was essential for the provision of adequate pensions and superannuation. These entitlements also typically included work care, insurance cover, retirement benefits, and healthcare.

Citizenship for male workers characteristically evolved out of economic conflicts over conditions of employment, remuneration, and retirement .Service to the state through warfare generates a range of entitlements for the soldier–citizen. Wartime service typically leads to special pension rights, health provisions, housing, and education for returning servicemen and their families. War service has been important, as we have seen, in the development of the evolution of social security entitlements (Titmuss 1962).

Finally, people achieve entitlements through the the formation of households and families, which become the mechanisms for the reproduction of society through the birth, maintenance, and socialization of children. These services increasingly include care for

the aging and elderly as generational obligations continue to be satisfied through the private sphere (Finch 1989). These services to society through the family provide entitlements to both men and women as parents, that is as reproducers of the nation. These familial entitlements become the basis of family security systems, various forms of support for mothers, and health and educational provision for children. Although the sexual activity of adults in wedlock is regarded in law as a private activity, the state and church have clearly taken a profound interest in the conditions for and consequences of lawful (and more particularly unlawful) sexual activity. Following Foucault, we can argue that heterosexual reproduction has been a principal feature of the regulatory activity of the modern state. It is evident that the values and norms of a household constituted by a married heterosexual couple provides the dominant ideal of British social life , despite the fact that 49 percent of live births in 1998 occurred outside marriage and that among the majority of one-family households 30 percent had no children. In fact, the moral force of the idea of marriage and domesticity is so compulsive in contemporary society that in America a number of states are considering legislation that would enable gay couples to form 'civil unions,' entitling them to about 300 rights and benefits currently available under state law to married heterosexual couples.

3. *Regulatory Regimes and Public Policy*

This liberal pattern of public policy for social citizenship has been eroded because the three foundations of effective entitlement have been transformed by economic, military, and social changes. It may sound perverse to suggest that in contemporary Britain the decline of economic participation has brought about an erosion of citizenship, when participation in the labor force has been rising continuously since the early 1990s. Increasing economic activity has been especially important for women; between 1971 and 1999, the proportion of the adult female population being economically active increased from 56 to 72 percent. However, high levels of economic participation mask a real change in the nature of the economy and obscure a transition from old to new welfare regimes. The new economic regime is based on monetary stability, fiscal control, and a reduction in government regulation of the economy. In this new economic environment, one version of the 'Third Way strategy' involves not protecting individuals from the uncertainties of the market that had dominated welfare strategies between 1930 and 1970, but helping people to participate successfully in the market through education (lifelong learning schemes), flexible employment (family-friendly employment strategies), and tax incentives (Myles and Quadagno 2000). However, while increasing rates of economic activity have been a positive aspect of economic liberalization, much of this increase in economic participation has required the casualization of the labor force. While the number of men in part-time employment doubled between 1984 and 1999, radical changes in the labor market (job sharing, casualization, flexibility, downsizing, and new management strategies) have disrupted work as a career. While for employers functional and numerical flexibility has broken down rigidities in the workplace, these strategies have compromised job security (Abercrombie and Warde 2000, p. 81).

Sociological studies of social class suggest that, while levels of unemployment have been falling in association with the long American economic boom of the 1990s, the contemporary class structure has new components—an 'underclass' of the permanently unemployable (typically single-parent welfare claimants), a declining middle class associated with the decline of middle management, and the 'working poor' whose skill levels do not permit upward mobility (Sennett 1998). There is some academic consensus that features of the class structure do not encourage active citizenship through economic entitlement, but these changes in the nature of employment are perhaps insignificant when compared to the graying of the population and the social problem of retirement. It is clear that the stereotype of the elderly as a dependent and passive population in disengagement theory is false, but it is also true that the aging of the population has important implications for the shape of the working population and for employment as a basis for entitlement. Intergenerational conflict in the struggle over resources is likely to become an important element in social divisions in this century.

Public policy and social citizenship were aspects of modern state administration that developed in the nineteenth century. We can interpret this administrative complex as a response to social class conflicts in the development of modern capitalism. Social rights were granted as political mechanisms to secure an acceptable level of social solidarity. Walter Bagehot (1963) in his *The English Constitution* warned in 1867 of the dangers of working class combination, which would result in the 'supremacy of ignorance,' and encouraged the 'higher classes' to exercise wisdom and foresight. An extension of social rights was a prudent response to the dangers of combination among the 'lower classes' and social citizenship has remained a valuable public policy option to avoid civil war. However, social citizenship and capitalism have remained in a state of permanent tensions—as Marshall recognized in his concept of 'hyphernated society' in *The Right to Welfare and Other Essays* (Marshall 1981). There is a permanent tension between the liberal rights of individual freedom in a capitalist marketplace and principles of equity and justice that require state interventions to protect social rights.

Public policy and welfare systems oscillate between different regulatory environments. A regulatory regime may be defined as 'a historically specific

configuration of policies and institutions which structures the relationship between social interests, the state, and economic actors in multiple sectors of the economy' (Eisner 1993, p. 1). These regimes fluctuate between interventionist welfare systems and deregulated privatized systems. For example, in retrospect, we can see that American public policy passed through different regulatory regimes in the Progressive Era, the New Deal, postwar reconstruction, and the contemporary period. Following the depression, President Roosevelt and his advisors introduced a legislative programme aimed at social and economic recovery. The National Industrial Recovery Act was characteristic of the new policy regime, but the Progressive Era still depended on market forces. The New Deal and postwar reconstruction required an alliance between the state and capital to conduct a global war and then to achieve an economic recovery. In wartime Britain, the social policies of John Maynard Keynes, as expressed in *The General Theory of Interest, Employment and Money* (1936), were adopted to finance the war. Social Keynesianism also formed the basis for postwar reconstruction where the policy of supporting public works, such as direct investment in infrastructure projects, promoted the recovery of employment. These public policies were in direct opposition to the 'Treasury view,' which advocated fiscal constraint and limited government intervention. In the late twentieth century, there was a departure from interventionist policies; environmental controls and workers' safety are not primary objectives, and were seen to conflict with corporate profitability. In both the USA and UK, there has been a similar history involving a rapid departure from the principles of the New Deal, Social Keynesianism, and the postwar consensus in favour of policies that do not assume, for example, full employment or universal criteria in the provision of pensions.

4. Conclusion: Public Policy and Globalization

This historical pattern of policy options changed significantly in the last two decades of the twentieth century with the emergence of a global economy. Although it would be an exaggeration to claim that globalization has resulted in the decline of national sovereignty (Hirst and Thompson 1996), it is the case that global economic processes constrain the capacity of national governments to make independent decisions about national public policy. For example, volatile financial markets responding to global information systems can undercut national public policy through the collapse of local currencies. The Asian financial crisis of the late 1990s was a dramatic illustration of how global uncertainty can destabilize currencies and prevent governments from adhering to public policies that have inflationary consequences. Volatile markets, the fragmentation of public policy by constant restructuring of government agencies and the economy, an emphasis on individual responsibility and subsidiarity, and the recommodification of services through neoliberalism have encouraged some social scientists to argue that public policy has become postmodern (Petersen et al. 1999). A new policy environment may involve the globalization of delivery through global enterprises that provide services for governments under 'outsourcing' arrangements. The growth of private prisons, managed by global enterprises, to replace or supplement state prison services would be one example. Social experiments in policy delivery in a global context may not be postmodern, but they will certainly be increasingly translocal. As a consequence, the conventional mixture of state, market, and voluntary sector as the framework of public policy will change radically to reflect the changing circumstances of social citizenship in a global economy.

See also: Civil Society, Concept and History of; Civil Society/Public Sphere, History of the Concept; Citizenship, Historical Development of; Citizenship: Political; Citizenship: Sociological Aspects; State and Society; Welfare State

Bibliography

Abercrombie N, Ward A 2000 *Contemporary British Society*. Polity Press, Cambridge, UK
Bagehot W 1963 *The English Constitution*. Collins, London
Barbalet J M 1988 *Citizenship*. Open University Press, Milton Keynes, UK
Beveridge W H 1942 *Social Insurance and Allied Services*. HMSO, London
Beveridge W H 1944 *Full Employment in a Free Society*. Allen & Unwin, London
Brown K, Kenny S, Turner B S 2000 *Rhetorics of Welfare*. Macmillan, Basingstoke, UK
de Tocqueville A 1966 *Democracy in America*. Doubleday, New York
Dewey J 1963 *Freedom and Culture*. Capricorn, New York
Diggins J P 1994 *The Promise of Pragmatism. Modernism and the Crisis of Knowledge and Authority*. University of Chicago Press, Chicago
Eisner M A 1993 *Regulatory Politics in Transition*. The John Hopkins University Press, Baltimore
Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Polity Press, Cambridge, UK
Finch J 1989 *Family Obligations and Social Change*. Polity Press, Cambridge, UK
Foucault M 1991 'Governmentality'. In: Burchell G, Gordon C, Miller P (eds.) *The Foucault Effect. Studies in Governmentality*. Harvester, London, pp. 87–104
Habermas J 1976 *Legitimation Crisis*. Heinemann, London
Hirst P, Thompson G 1996 *Globalization in Question*. Polity Press, Cambridge, UK
Katz S 1996 *Disciplining Old Age. The Formation of Gerontological Knowledge*. University of Virginia Press, Charlottesville, WV
Keynes J M 1936 *The General Theory of Interest, Employment and Money*. Macmillan, London

Lipset S M 1960 *Political Man. The Social Bases of Politics*. Doubleday, Garden City, NY
Loney M, Bosell D, Clarke J (eds.) 1983 *Social Policy and Social Welfare*. Open University Press, Milton Keynes, UK
Mann M 1987 'Ruling class strategies and citizenship'. *Sociology* **21**(3): 339–54
Marshall T H 1950 *Citizenship and Social Class and Other Essays*. Cambridge University Press, Cambridge, UK
Marshall T H 1967 *Social Policy in the Twentieth Century*. Hutchinson, London
Marshall T H 1981 *The Right to Welfare and Other Essays*. Heinemann, London
Myles J, Quadagno J 2000 'Envisioning a Third Way.' *Contemporary Sociology* **29**(1): 156–67
Myrdal G 1944 *The American Dilemma*. Harper, New York
Nelson D D 1998 *National Manhood. Capitalist Citizenship and the Imagined Fraternity of White Men*. Duke University Press, Durham, NC
O'Connor J 1973 *The Fiscal Crisis of the State*. St Martin's Press, New York
Parkin F 1979 *Marxism and Class Theory. A Bourgeois Critique*. Tavistock, London
Petersen A, Barns I, Dudley A, Harris P 1999 *Poststructuralism, Citizenship and Social Policy*. Routledge, London
Piven F, Cloward R A 1972 *Regulating the Poor*. Tavistock, London
Rorty R 1997 *Achieving Our Country. Leftist Thought in Twentieth-Century America*. Harvard University Press, Cambridge, MA
Sennett R 1998 *The Corrosion of Character. The Personal Consequences of Work in the New Capitalism*. W W Norton, New York
Shklar J N 1991 *American Citizenship. The Quest for Inclusion*. Harvard University Press, New York
Smith R 1997 *Civic Ideals. Conflicting Visions of Citizenship in US History*. Yale University Press, New Haven, CT
Titmuss R 1958 *Essays on the Welfare State*. Allen & Unwin, London
Titmuss R 1962 Income Distribution and Social Change. *A Case Study in Criticism*. Allen and Union, London
Titmuss R 1968 *Commitment to Welfare*. Allen & Unwin, London
Troeltsch E 1912 *The Social Teaching of the Christian Churches*. University of Chicago Press, Chicago
Turner B S (ed.) 1993 *Citizenship and Social Theory*. Sage, London
Turner B S 1986 *Citizenship and Capitalism. The Debate over Reformism*. Allen & Unwin, London
Wilson E 1977 *Women and the Welfare State*. Tavistock, London
Wuthnow R 1996 *Poor Richard's Principle. Recovering the American Dream through the Moral Dimension of Work, Business and Money*. Princeton University Press, Princeton, NJ

B. S. Turner

Citizenship, Historical Development of

Citizenship means membership in a political community. As *membership*, citizenship confers the status of equality among all citizens with respect to the rights and duties that the status implies. Citizenship also signifies a form of active *behavior* towards the community, which constitutes the good and responsible citizen. These two basic meanings of citizenship apply to all of the historical phases that the formation of citizenship as subject and concept has undergone. The politico-legal status and the ideal of civic virtue constitute the two aspects of a historical concept that has taken on a variety of further meanings and functions over time.

It has shifted between membership in ancient communities and legal membership in the modern state, between a concrete legal status and a wide-ranging concept within political theory. It has been used to describe substantive rights and obligations as well as to sketch a normative ideal of politics. While the theoretical claims of citizenship often tend to be universal, their meaning varies according to historical and national context. The concept of citizenship is rooted in social ethics and institutionalized by law. As such, it is an object of political theory. To analyze the historical development of citizenship means to combine the development of norms of conduct regarding good civic behavior with the legal institutionalization of citizen status and its theoretical conceptualization. It also means tracing the trajectory from a status of the city-state community to a key concept of modern democracy and global political theory.

1. *Citizenship as a Historical Subject*

Every human group has developed institutions by which to define its members and procedures for making new members. The Greeks, however, were the first society to combine the legal provisions of membership with a political theory of membership virtues and institutions in order to perpetuate their idea of citizenship. The legal status of citizenship established equality among Athenian citizens in terms of their rights and obligations. The status of equal membership marked privilege vis-à-vis nonmembers. The overall legal status of Athenian citizenship was defined by sharp boundaries, which distinguished Athenians from foreigners, resident aliens, and slaves. The minority of full citizens faced a majority of non-privileged noncitizens whose rights were severely circumscribed. Apart from the legal framework, the concept of Athenian citizenship included a set of citizen values, behaviors, and communal attitudes. 'Passive' legal citizenship was not necessarily, but ideally and practically, linked to democracy by the civic virtues of an 'active' citizen, which permitted him to 'share in the polis,' a politically active and autonomous community (Manville 1990). Greek city-state citizenship and its dual construction as legal framework and civic ideal was conceptualized in the political philosophy of Aristotle as a model of 'ruling and being ruled in turn.' It became the model of citizenship as such.

Roman citizenship was more complex, expansive, and legalistic than its Greek counterpart. The history

of Roman citizenship over eight centuries, from the end of the monarchy to the decline of the Roman Empire, reveals the stages of its development from an instrument for defining the city-state community to one for the legal integration of an extensive world empire. The Twelve Tables established the quality of Roman citizenship as a legal benefit and attractive political and social status. In the fourth century BC, with the territorial expansion of Roman rule over Latium, Roman citizenship was conferred upon annexed enemies and (Italian) allies alike as an instrument for promoting loyalty and political integration. By the first century BC, citizenship of the Roman Republic had reached its greatest 'density': the highest level of participatory rights in the government of the republic and equality before the law accompanied a golden age of civic education and virtues. This construction lost its balance in the Roman Empire. The largest territorial expansion of Roman citizenship, its conferral upon nearly all men within the confines of the Empire except for slaves, meant the erosion of its quality as a privilege and a proof of superior morality. With the decline of the Empire, the diminution in standards of citizenship corresponded to the degeneration of civic educational standards. Although the concept of citizenship did not necessarily depend on democracy, its development was most favored by the political conditions of a strongly participatory, republican order. A centralized, differentiated legal order within the city or state had, at least, proved to be the indispensable precondition for the development of citizenship. With the loss of a central, universal legal structure at the end of the Roman Empire, citizenship became obsolete as a political concept.

Christianity rejected the model of political order to which ancient philosophy, especially Aristotle and Cicero, had contributed. It developed a complete alternative system of social and moral values, which helped to establish a new institutional political order. The rise of Christian corporatism and its unlimited commitment to the Kingdom of God produced a dichotomous organization of the body politic. A dual system arose of metaphysically based allegiance to the Church and personal allegiance to the monarch. A system of multifaceted loyalty replaced the concentration of the citizen's loyalty upon the state. The notion of citizenship was transferred to a spiritual 'City of God' (Augustine). Thus, the institutional point of reference for the ancient model of citizenship disappeared. In the Middle Ages, the term 'citizenship' (French: *citoyenneté*) did not describe an Englishman's (or Frenchman's) relationship with his country. In keeping with its immediate etymological origins ('city'), it referred solely to the rights and duties of a free city- or town-dweller, i.e., to a local, urban community.

The lack of a centralizing and nationalizing state power in parts of medieval Europe left room for the growth of a strong, urban citizenship, particularly in the city-states of northern and central Italy. Along with the weakness of the monarchy, a certain continuity of Roman law and civic responsibility formed the backdrop for the upsurge of a politically active municipal life with a high participation of the citizenry in decision-making. A vigorous philosophical revival of the Aristotelian concept of citizenship paved the way for a resurgence of the classical concept of citizenship during the Renaissance. The founding of universities, which promoted the development of Roman law, a vital communal life, and a rich theoretical literature on a rational and egalitarian order (particularly the work of Niccolo Macchiavelli) made the city-state world of Upper Italy a birthplace of modern citizenship.

The theory and political order of state sovereignty in European absolutism with its concentration of central state power and its unambiguous claim to loyalty prepared the ground for a national rather than local concept of modern citizenship. The personal and territorial delimitation of sovereign states and the ordering of their international relations increased the problem of defining membership in and allegiance to the state. The call for (individual) rights of religious freedom and self-government, strengthened by religious opponents of the state, and revolutionary efforts at founding government upon popular sovereignty, for example in England, were the predecessors to active citizenship. The institutional foundations for such participatory claims and rights remained narrow, however. They were often confined to the level of local government or socially privileged groups. The contrast between a basically oligarchic and a democratic political citizenship (Heater 1990) became evident.

The eighteenth century brought the breakthrough of citizenship on a central level of political theory and institutions. The word 'citizen'—and its French equivalent *citoyen*—became a key concept in the legitimation of the political struggle against the feudal *ancien régime* (Gosewinkel 2001), as well as in the struggle for equality before the law, freedom from religious discrimination and arbitrary arrest, the extension of political rights and popular democracy. The decisive step from a political and educational program to a legal guarantee was taken by American constitutionalism, however. The revolution of the American colonies against the ties and obligations of subjecthood to the English Crown was accompanied by a decisive change in legal terminology. Within the decade before the enactment of the Federal Constitution (1787), the term 'citizen' came to replace previously related words such as 'subject' and 'inhabitant' in constitutional texts. The republican concepts of 'citizen' and 'citizenship' marked a political and terminological break with the feudal age. From this point onward citizen and citizenship became core terms of a state order based on democracy and constitutionalism. The American constitution intro-

duced a revolutionary new type of legal instrument that firmly rooted any state power in the law, while simultaneously subjecting it to legal control. The constitutionalization of state power represented a fundamental development in early modern Western history: the subjection of social and political relationships to legal norms (*Verrechtlichung*). With the constitutionalization of membership status in a republican state order, the concept of citizenship gained new importance. It unified the claims to equality and political participation inherent in the classical concept of city-state citizenship, transferred it to the state level and gave it a new quality of legal efficacy. The constitutionalization of citizenship with its classical connotation of civil idealism and civic virtue represented a strong and permanent challenge to inequality and exclusion from rights in political practice. At the same time, the shift in scale from the small local community to the level of the abstract state risked attenuation in the practices and understandings of citizenship. The need for a new, expansive legitimation of citizenship to the state arose.

The French Revolution made citizenship (*citoyenneté*) one of its key concepts (Waldinger et al. 1993). This new citizenship combined four traits that were to become essential for the development of citizenship throughout the nineteenth and twentieth centuries: its *egalitarian*, antifeudal impetus, its confirmation as a key concept of the *legal constitution*, its association with extensive *individual rights*, and finally its *nationalization*. The modern concept of citizenship arose together with the concept of the nation-state (Bendix 1964) and became one of its central legal institutions. The age of revolution and constitutionalism in the Western world was also an age of an increasing delimitation of national citizenries. Extended citizenship law came to define membership in the nation-state as well as the rules of naturalisation policy.Whether they were based on an inclusionary, territorial model (USA, France) or an exclusionary, descent-based model (Central and Eastern Europe), citizenship and citizenship law became key instruments for defining national identity and controlling migration in a modern world characterized by increasing transnational mobility (Brubaker 1992). The central function of citizenship in defining membership in the sense of nationality was supplemented by a second main function: that of conferring upon the citizen individual rights vis-à-vis the state. In the nineteenth and early twentieth century, Western constitutional states extended the range of civil, political, and, increasingly, social rights which were reserved mainly for their own citizens. The nationalization of citizenship as a membership status corresponded to a nationalization of citizens' rights. Citizenship became an institution for distributing 'life opportunities' in a world of nation-states.

As the importance of citizenship as a legal entitlement increased, its extension to new members became more and more contested within the national citizenry. The struggle for inclusion, for example in American constitutional law, dominates the history of citizenship to this day. The gradual extension of equal civil rights to groups resident on American soil who had been denied citizenship rights and subjected to discrimination on the basis of ethnic or national origin, religious beliefs, social status, or gender determined the direction that citizenship was to take. The claims of discriminated groups to equality became the motor for full inclusion in the community defined by citizenship.

With the decline of liberal democracy, and the rise of radical nationalism, racism, and totalitarian dictatorship in the constitutional states of Europe and Asia, the period between the two world wars represented an interruption in the development of modern citizenship. The dominance of ascriptive national, ethnic, or racial criteria in admission to citizenship, the splintering of the citizenry through hierarchical classes of rights, the withdrawal of civil rights and massive expatriation of millions, destroyed the core of equality within the concept of citizenship. The rise of an army of stateless people deprived of rights and protection (Arendt 1951) revealed how dependent citizenship was on the liberal legal structures of the nation-state.

Developments after the Second World War were characterized by a dual tendency. On the one hand, the restoration of liberal democracy saw a reconstruction of citizenship. This was reinforced and extended to the global arena after 1989 with the end of ideological block confrontations and the adoption of democratic and constitutional patterns by most of the formerly communist states. On the other hand, there is also evidence of a tendency towards a certain 'devaluation' of citizenship. The weakening of nation-state structures, the trend towards transnational political unions, global standards and guarantees of civil rights as human rights have all diminished the importance of (national) citizenship for the conferral of individual rights. While formal citizenship is still crucial on the level of the right to full participation in the political arena, this no longer applies to economic and social rights. Citizenship as national membership status is of decreasing importance for the exercise of these increasingly relevant rights (Soysal 1994).

The historical development of citizenship from its beginning in the Greek city-states has been characterized by a multiple process of expansion. Citizenship expanded from a membership status in a local community to a central membership in the territorial nation-state. Citizenship as entitlement to individual rights was transferred from the level of the nation-state to that of supranational communities. From its inception in the Greek city-communities, the concept and ideal of citizenship has spread all over the world to states based on the principle of constitutional democracy. The substantive program of citizenship as a set of individual rights has expanded from political and

civil to encompass social and economic rights, and ultimately cultural and environmental rights as well.

2. *The Historical Conceptualization of Citizenship*

Two main features mark the conceptualization of citizenship. The first consists of the twofold structure of a legally defined status of membership and rights on the one hand and a philosophical and normative ideal of 'good' civic behavior on the other. The second feature depends on the basic structure of equality within citizenship. Its conceptual development may be interpreted as a claim for recognition as equal and a struggle for inclusion. In the classical age of citizenship, its inclusive function consisted of either defining a ruling oligarchy with precisely delineated privileges or expanding a low standard of rights to an extensive group for the purposes of political integration. The linkage between a substantively high standard of citizenship rights and its extension to a broad group of persons was the result of a change in politics and theory that began in the early modern period. Between the sixteenth and eighteenth centuries, the concept of citizenship lost its coherence and republican ethos (Riesenberg 1992). The classical idea of citizenship survived as a historical concept in education and political literature, however. It corresponded doubly with the era's need for a break with old categories and concepts. First, the vision of an 'active' citizen fit well with the notion of a modern, visionary, and enterprising 'polytechnic man' as outlined in the political theories of Thomas Hobbes and John Locke. Second, the concept of citizenship was enlarged by bringing the terms 'citizen' and 'subject of the monarch' closer together, and indeed mingling them. The era's most elaborate conceptualization of citizenship, undertaken by Jean Bodin, transferred citizenship from the local to the state level. According to Bodin, the citizen was a superior subject of the sovereign monarch. Citizenship denoted a direct, personal link between the individual and his sovereign. This conceptual transfer opened the way for a revolutionary, egalitarian view: from the perspective of the sovereign all citizens were equal *qua* subjection to the monarch's sovereign authority. It was not until the eighteenth century that this kind of passive equality was transformed by the subject-citizen into a new, revolutionary call for active citizenship, i.e., for participation in sovereignty.

The concept that paved the way for this fundamental change of perspective was the revolutionary idea of 'civil society.' It emerged in the later seventeenth and eighteenth centuries as a result of a crisis both in the reality and idea of social order. In an age marked by an unprecedented commercialization of human relations, the growth of market economies and political revolution, the regulatory idea of the social order was no longer derived from a deity external to the temporal world, but from within it and its ethical values. The vision of an all-encompassing 'society' of human beings, equal, autonomous, and rational by nature, who defined their own political order, and challenged the traditional norms of monarchical order, ecclesiastical claims, and social hierarchy. Beginning with John Locke and the philosophers of the Scottish Enlightenment, and passing on to Kant, Hegel, and eventually Marx, the idea of 'civil society' (in German: '*bürgerliche Gesellschaft*') laid the groundwork for a new idea of citizenship. Although not always developed systematically, the concept of the citizen and citizenship constituted the normative core of civil society (Haltern 1985). Out of Locke's view of civil society as a system of individual rights and duties, Kant's definition of the public arena in civil society as a sphere of juridical equality among citizens, and Hegel's (Riedel 1972) and Marx's vision of civil society as a field of conflicting particular interests, elements of a new understanding of citizenship emerged. It now referred to the central state, not the local community, and was based on individual autonomy, not obedience and subjection. It was legally constituted and based on the principle of legal equality, not oligarchic privilege. Finally, it was inclusive with regard to the principles of property and achievement. This new citizenship represented an ideal and became a key concept in the language of American and European revolutions at the end of the eighteenth century. It was not, however, the product of political revolution. Its achievement was the result not only of the conceptualization of 'civil society,' but also of its realization as a new formation of social and political order throughout the Western world. This process marked such a major *caesura* in the historical development of citizenship that it may be considered the rise of a 'second citizenship' (Riesenberg 1992).

In contrast to the universal and egalitarian concept of 'civil society,' the development of citizenship in the nineteenth century ran up against three fundamental obstacles to its realization in political practice. The first obstacle was nationality. Citizenship was conceptualized as membership in the nation-state. While the rights conferred upon citizens by the liberal constitutions of the nineteenth century grew in number and substance, they were increasingly confined to nationals. The universalist pathos of human rights, which was still inherent in the revolutionary declarations of the American and French Revolutions (e.g., The Declaration of the Rights of Man and Citizen), was increasingly reduced to national entitlements and interpreted in this light by legal scholars. The pre- and transnational origins and impetus of 'civil society' gave way to a nationally restricted concept of constitutional citizenship, which had lost its universalist ethos of civic values.

The second obstacle was gender inequality. Citizenship was conceptualized in a seemingly gender-neutral manner, although both membership status and the civic ideal of an 'active citizen' were based on rights and activities that were largely reserved for men. The construction of the state as a 'male state' was a challenge to the principle of equality by law which was proclaimed by the American and French revolutions. From its beginning, the revolutionary principle of citizenship did not imply equal rights for women. The traditional concept of citizenship of the nineteenth and first half of the twentieth century was deeply gendered. Thus it became a central issue of deconstruction in feminist critiques (Pateman 1988).

The third obstacle to democratic citizenship was criticized by Karl Marx in the 1840s. He demonstrated that the provision of equality in civil and political citizenship rights was merely a formal, legal one: it left untouched the practical inequalities in people's abilities to exercise the rights or legal capacities that constituted citizen status. Moreover, even the exercise of the rights of membership could not influence the basic conditions of class inequality because of the purely formal nature of these rights. Marx interpreted bourgeois citizenship as an instrument of class rule, and thus as an institution of industrial society, which was particularly driven by class. Citizenship as a phenomenon of the modern industrializing world thus corresponded to the interest of the new social sciences. It was sociologists, beginning with Max Weber, who analyzed the origins and growing importance of citizenship in the modern social order. Because citizenship was imbedded in the 'Western' model of modernization and industrialization, it came to be identified as a specifically Western concept of political and social order.

During the first half of the twentieth century, though, intellectual interest in the concept of citizenship was mainly historical. As a political ideal it suffered both from Marx's cutting critique and the crisis of the liberal-democratic idea between the world wars. Thus, only the restoration of a democratic political order after the Second World War and the significant success of the welfare states established at that time paved the way for the revival of citizenship as a key concept in both social analysis and politics. It was against this background that the 1950 article on 'Citizenship and social class' by the English sociologist T. H. Marshall attained its path-breaking influence over the entire debate on citizenship in the second half of the twentieth century. Marshall took up the interpretation of citizenship as a set of rights and analyzed its historical development as a successive creation of civil, political, and social rights. Marshall particularly stressed two conclusions that were crucial to the revaluation of citizenship as both an analytical instrument and a political ideal: his emphasis on the gradual expansion of rights and the ultimate undermining of class conflict by the achievement of social rights weakened the Marxist critique. At the same time, Marshall asserted that by challenging existing inequities and calling for their abolition as an individual right, equality (as the conceptual core of citizenship) was a dynamic principle that continued to exercise an emancipatory and inclusive influence.

By doing so, Marshall's concept of citizenship attained the force of a universal claim to equality that was applicable to a virtually unlimited range of subject matters. Despite continuing Marxist skepticism, social citizenship was interpreted as a call for the stabilization and expansion of social welfare rights (Turner 1986). The lack of civil and political rights strengthened the call for citizenship as a key concept of opposition to dictatorships in the Soviet block. The idea of civil society was rediscovered in Central and Eastern Europe as a basis for citizenship rights and translated into a political program for restructuring post-Soviet societies after 1989.

This expansive conceptualization has had a dual effect. After the downfall of the Soviet block and the global expansion of democratic constitutionalism, citizenship was adopted globally as a constitutive element of the democratic order (Thompson 1970). New states in Asia, Africa, and South America used citizenship to integrate into the society of world democracy. At the same time, social conflict and cultural change within the states with citizenship traditions were often conceptualized as calls for equality and recognition expressed in terms of citizenship. Massive transnational migration and attempts to take account of 'multiculturalism' (Kymlicka 1995), the enforcement of gender equality and sexual rights (Lister 1997), the debates within liberalism and communitarianism surrounding the foundation and values of a 'good' civic life—all of these issues have been interpreted as struggles over citizenship rights. Ultimately, citizenship has become increasingly detached from its association with the nation-state and conceptualized instead as 'transnational citizenship' (Bauböck 1994), culminating in the vision of 'world citizenship' (Heater 1996).

The global success and expansion of the term 'citizenship' has not, however, necessarily stabilized the underlying concept. As a political ideal, citizenship has been transplanted to states that lack the historical foundation of a 'civil society.' It may, therefore, prove inadequate for the social and cultural traditions of non-Western states. Historically, citizenship's efficacy as a claim for rights was always based on political sovereignty—be it monarchical or popular. This was a prerequisite both for addressing individual claims and enforcing the corresponding civil duties. With the pluralist expansion of citizenship rights in a situation of declining nation-state sovereignty and a nascent world state order, citizenship as it currently exists may lack the requisite unifying force. As Roman history demonstrates, the global expansion of citizenship does not necessarily lead to its global achievement.

See also: Citizen Participation; Citizenship: Political; Citizenship: Sociological Aspects; Civic Culture; Civil Society, Concept and History of; Civil Society/Public Sphere, History of the Concept; Democracy; Democracy, History of; Democratic Theory; Freedom/Liberty: Impact on the Social Sciences; Freedom: Political; French Revolution, The; Human Rights, History of; Liberalism: Historical Aspects; Marshall, Thomas Humphrey (1893–1981); Public Sphere: Nineteenth- and Twentieth-century History; Rights; State, History of

Bibliography

Arendt H 1951 *The Origins of Totalitarianism*. Harcourt Brace Jovanovich, New York

Barbalet J M 0000 *Citizenship. Rights, Struggle and Class Inequality*. University of Minnesota Press, Minneapolis, MN

Bauböck R 1994 *Transnational Citizenship. Membership and Rights in International Migration*. Edward Elgar Publishing, Aldershot, UK

Beiner R 1995 *Theorizing Citizenship*. State University of New York Press, Albany, NY

Bendix R 1964 *Nation-Building and Citizenship*. University of California Press, Berkeley, Los Angeles, London

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Gosewinkel D 2001 Citizenship, subjecthood, nationality: concepts of belonging in the age of modern nation states. In: Eder K, Giesen B (eds.) *European Citizenship Between National Legacies and Postnational Projects*. Oxford University Press, Oxford, UK, pp. 17–35

Hallern U 1985 *Bürgerliche Gesellschaft*. Wissenschaftliches Buchgesellschaft, Darmstadt

Heater D 1990 *Citizenship. The Civic Ideal in World History, Politics and Education*. Longman, London

Heater D 1996 *World Citizenship and Government. Cosmopolitan Idea in the History of Western Political Thought*. Macmillan Press, Basingstoke, UK

Hufton O H 1992 *Women and the Limits of Citizenship in the French Revolution*. University of Toronto Press, Toronto

Kymlicka W 1995 *Multicultural Citizenship. A Liberal Theory of Minority Rights*. Clarendon Press, Oxford, UK

Kymlicka W, Norman W 1995 Return of the citizen: A survey of recent works on citizenship theory. In: Beiner R (ed.) *Theorizing Citizenship*. State University of New York Press, Albany, NY

Lister R 1997 *Citizenship. Feminist Perspectives*. Macmillan Press, Basingstoke, UK

Manville P B 1990 *The Origins of Citizenship in Ancient Athens*. Princeton University Press, Princeton, NJ

Marshall T H 1950 *Citizenship and Social Class—and other Essays*. Cambridge University Press, Cambridge, UK

Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA

Riedel M 1972 Artikel Bürger, Staatsbürger, Bürgertum. In: Brunner O, Conze W, Koselleck R (eds.) *Geschichtliche Grundbegriffe*. Ernst Klett Verlag, Stuttgart, Vol. 1, pp. 672–725

Riesenberg P 1992 *Citizenship in the Western Tradition*. The University of Carolina Press, Chapel Hill, NC

Soysal Y 1994 *Limits of Citizenship*. University of Chicago Press, Chicago

Thompson D 1970 *The Democratic Citizen*. Cambridge University Press, New York

Turner B S 1986 *Citizenship and Capitalism: The Debate over Reformism*. Allen and Unwin, London

Waldinger R, Dawson P, Woloch I 1993 *The French Revolution and the Meaning of Citizenship*. Greenwood Press, Westport, CT

Walzer M 1970 The problem of citizenship. In: Walzer *Obligations*. Harvard Unversity Press, Cambridge, MA

D. Gosewinkel

Citizenship: Political

It may seem superfluous to subtitle a discussion of citizenship as 'political.' The term is at core ineradicably political: its oldest, most basic, and most prevalent meaning is a certain sort of membership in a political community. There are nonetheless good reasons to underline just how deeply political citizenship is, as this essay strives to do. The chief reason is that, precisely because citizenship is so profoundly political a term, there are recurrent pressures to depoliticize both its meaning and its accompanying practice. These pressures, and the understandings of citizenship they propagate, are best seen as confirmations of the political character of citizenship, not as alternative, apolitical conceptions.

1. Four Meanings of Citizenship

Perhaps the most familiar meaning of citizenship is in fact the seminal one. In both ancient and modern republics and democracies, a citizen has been a person with political rights to participate in processes of popular self-governance. Yet we also commonly speak of citizenship as a more purely legal status. Citizens are people who are recognized legally as members of a particular political community and who, therefore, possess some basic rights to be protected by that community's government, whether or not those rights include rights of political participation. During the last century, moreover, many have come to use citizen as a way of referring to those who belong to almost any human association, whether a political community or some other group. I can be said metaphorically to be a citizen of my neighborhood, my fitness club, and my university as well as my broader political community. Finally, we often use citizenship to signify not just membership but certain standards of proper conduct, implying that only 'good' citizens are truly citizens in the full meaning of the term.

The word citizen derives from the Latin *civis*, meaning a member of an ancient city-state, preeminently the Roman republic; but *civis* was a Latin rendering of the Greek term *polites*, a member of a Greek *polis*. Innumerable scholars have told how a renowned resident of the Athenian *polis*, Aristotle, defined a *polites* or citizen as someone who rules and is ruled in turn, making citizenship conceptually inseparable from political governance (Aristotle 1968, p. 1275a23). Aristotle doubtless pleased many Athenians by arguing that such a status and activity, properly performed, represented the highest form of life available to most men. Yet this life was not available to Aristotle himself; he was not an Athenian citizen but a metic, a resident alien. He also suggested that the philosophic life he could and did pursue was in the end the highest of all; and the fact that he chose to pursue this life in a city that denied him citizenship may call into question how valuable he really thought citizenship to be.

His fulsome praise of the citizenship his Athenian hosts had created thus suggests how greatly definitions of citizenship have always been shaped by the political structures of power within which they have been offered. It clearly would not have been prudent for Aristotle to denigrate Athenian citizenship. The fact that neither Aristotle nor most of other residents of Athens, including aliens, women and slaves, were eligible for citizenship also underlines how citizenship originated not only as a way of structuring membership, but also as a way of distributing power within a particular political regime. That distribution disempowered far more Athenians than it enfranchised. Even so, the ideal of citizenship as self-governance that Athens and Aristotle established has often served since as an inspiration and instrument for political efforts to achieve greater inclusion and engagement in political life.

As such, this ancient idea of citizenship has often seemed politically threatening to many rulers, who have abolished or redefined the category. It was for this sort of political reason—because the regimes that had created citizenship succumbed to conquest by Alexander the Great's monarchical empire—that ancient Greek citizenship disappeared. And it was for a similar political reason—because the Roman republic gave way to imperial rule generated from within—that Roman citizenship came to have a different meaning than the one Aristotle articulated. In principle, Roman citizenship always carried with it the right to sit in the popular legislative assembly that had been the hallmark of Athenian citizenship. But as participation in that assembly became increasingly meaningless as well as impractical for most imperial inhabitants, Roman citizenship became essentially a legal status comparable to modern nationality (Pocock 1995). It provided rights to legal protection by Roman soldiers and judges in return for allegiance to Rome. That status was less 'political' in the sense that it no longer evoked recurrent engagement in practices of self-governance, but it was quite plainly a politically crafted status that represented a new distribution of power of enormous political significance.

Citizenship was then eclipsed in the West by the various feudal and religious statuses of the medieval Christian world, but it did not vanish entirely. 'Burghers' or the *bourgeoisie* were citizens of municipalities that often had some special if restricted rights of self-governance within feudal hierarchies. Such burghers remained, however, fundamentally subjects of some ruling prince or lord, with their citizenship chiefly providing legal rights of protection in the manner of Roman imperial citizenship. In contrast, during the Renaissance some Italian cities achieved both independence and a meaningful measure of popular self-governance. They invoked ancient 'republican' ideals of participatory citizenship to define and defend their regimes. Their experiences in turn fed into the antimonarchical revolutions that created the first modern republics, including the short-lived seventeenth century English Commonwealth and late eighteenth-century French Republic, as well as the still-enduring United States (Pocock 1975).

In complex fashion, those revolutions inaugurated transformations 'from subjectship to citizenship' across much of the globe that are still ongoing today, when most of the world's governments proclaim themselves to be 'republics' of some sort populated by citizens. It is in that context that we have come to use citizen ubiquitously for almost every kind of membership in every kind of organization, and to equate genuine citizenship with being a good, contributing member of those organizations. These pervasive popularizations of the term reflect, however, political developments that have in some respects diminished the significance of citizenship even as the term has spread.

2. *The Politics of Apolitical Citizenship*

Men created the early modern republics in an international realm that had been organized by the 1648 Treaty of Westphalia into a system of mutual recognition among overwhelmingly monarchical nation–states. In gaining acceptance within that system, the new republics defined their citizens as having the same international status as national monarchical subjects. For international purposes, these citizens, too, were simply persons who owed allegiance to and could claim protection from particular governments. Thus, Westphalian international law treated modern republican citizenship as akin to the legalistic, protection-oriented version of Roman citizenship.

Furthermore, Americans especially forged their republic amid racial and gender hierarchies that few leaders sought to challenge. Hence they felt compelled

to argue that, though free blacks and women might be citizens, citizenship did not in fact inherently entail rights of political participation. It guaranteed, once again, only more limited rights to certain judicial and executive protections (Smith 1997). For long stretches of time, then, both international and domestic politics worked to strengthen legalistic as opposed to the more participatory conceptions of citizenship despite the rise of modern republicanism.

Yet even though courts made the narrower, protection-centered view of citizenship legally authoritative, the notion that genuine citizenship involved rights of political participation remained a resonant rhetorical tool of legislative and constitutional reformers. Eventually both domestic protest movements and international pressures, including the need for broad support in wartime, converged to work for the expansion of the franchise to all adult citizens in the US and most of the western world. In America, blacks won both citizenship and voting rights after the Civil War, even though most came to be effectively disfranchised in the 'Jim Crow' segregation era; and women gained the franchise after World War I. In both cases, arguments appealing to their public service, especially in wartime, and to the idea that true citizenship must include the franchise, played key roles in their successes (Foner 1988, Flexner 1973).

In Britain and to some degree in other western European nations that had been politically configured essentially by feudal and industrial class systems, modern citizenship was wrought out via somewhat different struggles. As Marshall famously argued, first middle and then working class political pressures resulted in the expansion of civil rights of property and protection, then in near-universal rights of political participation, and finally and incompletely, in 'social rights' for all national citizens that included income, housing, medical, and educational guarantees (Marshall 1950). But even as the franchise has broadened in the US, Europe, and elsewhere, the notion of citizenship as active participation in meaningful self-governance has become more remote to many modern citizens. In part the logistics of large-scale modern societies make effective democratic participation very difficult. In part the economic and cultural developments that have led to a focus on 'social citizenship' make political activism seem less important. Engagement in one's social and economic organizations can appear more pressing. Perhaps, then, the term citizenship has become common in such contexts because it is there that people find the memberships that mean the most and in which they can most actively participate. If so, then the inevitable corollary is that citizenship understood as political self-governance has become quite secondary to many modern citizens. Yet this, too, is a signal political development. It is probable that like their predecessors in other regimes, many who wield power in modern republics are content when those they govern think of citizenship chiefly in terms of subnational, often nongovernmental associations, and in terms of the 'good citizen's' civic service rather than vigorous political participation. Certainly few policies within modern republics do much to enhance the feasibility and potency of such participation.

3. *The Prospect of Postnational Citizenships*

Though some scholars and democratic activists lament this current circumstance, others stress that the heightened transnational economic, transportation, and communication systems that we call globalization are in any case making traditional notions of national citizenship obsolete (Soysal 1994). Regional associations, international legal institutions, and transnational economic, cultural, and political organizations are all said to be more likely to shape humanity's future than existing national regimes. Hence membership in such bodies will represent the most important forms of citizenship in the twenty-first century.

That such globalizing trends exist is undeniable, though often national governmental actors remain major players even in transnational or international organizations and institutions. Despite advances in communication and transportation, moreover, meaningful participation in the governance of such populous and geographically far-flung entities seems even more chimerical for most people, deepening the eclipse of citizenship's oldest meaning. Thus there is a real prospect that the idea of citizenship increasingly will be severed not only from engagement in traditional forms of self-governance, but even from membership in some titularly sovereign political community. It may become a term for membership and participation in a wide variety of human groups, often simultaneously.

There are, however, reasons to doubt this scenario. History suggests that the leaders of political communities rarely give up power willingly. Therefore, it is not surprising that efforts to resist globalizing trends and reinvigorate loyalties to existing nations and regimes are also visible players in modern 'citizenship politics,' particularly in regard to immigration policies. A truly all-encompassing global government, moreover, still seems a fantasy, so that memberships in particular political communities are likely to remain important features of human life, even if those communities come to be constituted in new ways. Under at least some conditions, moreover, many people may feel great concern over the decline in forms of citizenship through which they can exercise some genuine control over their collective lives. The fact that political and social reform movements have often gained wide support by insisting that citizenship means sharing in governance shows that such feelings can be politically powerful fuel driving quite important changes.

Thus, we cannot rule out the possibility that older notions of participatory citizenship may continue to play a role in the recrafting of political institutions and communities that the twenty-first century will inevitably see. But whatever forms of citizenship result, they will almost certainly be the products of political contests that result in distributions of powers and memberships to some people and not others, distributions that will convey to them certain rights and protections, and not others. Hence, citizenship will remain what it has always been, a fundamentally political status through which human beings partly order both their individual and their collective lives.

See also: Citizen Participation; Citizenship and Public Policy; Citizenship, Historical Development of; Citizenship: Sociological Aspects; Civic Culture; Participation: Political; Political Culture

Bibliography

Aristotle 1968 *The Politics of Aristotle*. Clarendon Press, Oxford

Flexner E 1973 *Century of Struggle: The Woman's Rights Movement in the United States*. Atheneum, New York

Foner E 1988 *Reconstruction: America's Unfinished Revolution, 1863–1877*. Harper & Row, New York

Marshall T H 1950 *Citizenship and Social Class and Other Essays*. Cambridge University Press, Cambridge, UK

Pocock J G A 1975 *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition*. Princeton University Press, Princeton, NJ

Pocock J G A 1995 The ideal of citizenship since classical times. In: Beiner R S (ed.) *Theorizing Citizenship*. State University Press of New York, Albany, NY

Smith R M 1997 *Civic Ideals: Conflicting Visions of Citizenship in U.S. History*. Yale University Press, New Haven, CT

Soysal Y N 1994 *Limits of Citizenship: Migrants and Post-national Membership in Europe*. University of Chicago Press, Chicago

R. M. Smith

Citizenship: Sociological Aspects

Classical sociological theory was especially interested in the social mechanisms ensuring social solidarity, and in the sources of conflicts that challenge the social structure. It underlined roles, functions or dysfunctions, the networks of sociability which bring the actors together, etc. In general, the founders of the discipline were not very concerned with politics and hardly dealt with the question of citizenship, which marks the integration of actors within their nation. Consequently, until very recently, the question of citizenship had its origins within political theory, having abandoned its social dimensions for a long period.

From Aristotle to Locke and Rousseau, many philosophers have reflected on the nature of the social link, the commitment of individuals within the public space, the formation of the social contract and general willingness. People are conceived here as rational beings: By their entrance into the political community, they have access to the status of citizen which alone gives meaning to its own history. In the contemporary era, from Sheldon Wolin (1993) to Carole Pateman (1970), this reflection of political philosophy on the foundations of a democracy of citizens is illustrated by an immense literature, of which Benjamin Barber (1984) and his 'strong democracy' or Selya Benhabib (1996) are good representatives.

The revival of the political theory on citizenship has, however, found its origins: Hannah Arendt sought within the Greek *polis* the foundations of a *via activa* which would give life to citizens who were indifferent to the social, determined only by their reason, to enter into public space. After Arendt, it was Jürgen Habermas who took it upon himself to seek the origin of modern public space as a place of deliberation and discussion. According to him, public opinion formed by all citizens was born in the seventeenth and eighteenth centuries, in France and in England. Viewing capitalism as a system that alienates actors and destroys their exchanges, he only found reason within contemporary 'communicational behavior'; the use of modern information techniques would therefore give life to a democracy of citizens capable of communicating among themselves. Arendt and Habermas played an immense role in the revival of a political theory on citizenship capable of leading to a community which is undiluted by citizens who are little concerned by their particular historical identity, or their cultural identity, either social or even less biological.

A reflection upon the sociological foundations of citizenship should consequently find food for its debates elsewhere. Thus, it can be claimed that during the nineteenth century, de Tocqueville was one of the few thinkers to extend the eighteenth century tradition by giving it a more sociological context. He clearly poses the question of the commitment of ordinary citizens. Tocqueville asks himself questions about how to avoid the isolation and apathy of individuals, which favors, in France, for example, an indifference propitious to all forms of authoritarianism; in his opinion, the forms of local self-government established by US democracy limit these dangers. For Tocqueville, public space is doomed to silence and to state domination if the associations and social groups that bring individuals together with specific identical interests do not intervene within the elaboration of public politics. Citizens find themselves, from the beginning, plunged into the social. Whereas Marx thrusts aside the coming

of citizenship for the birth of a society that has rejected capitalism and sees only alienation in present society, thus silencing the purely political role of determined citizens by their place in production relationships, Tocqueville was already interested in the sociological aspects of citizenship.

Strangely enough, between Tocqueville and T. H. Marshall, the predominance of the social was so important that it was necessary to wait until the end of World War II for the concept of citizenship to return to the heart of the debates: However, this time, it was not the perspective of Rousseau that prevailed but more that of Tocqueville. During his famous conference in 1949, Marshall retraced the evolution which led from legal citizenship, created during the eighteenth century with the obtaining of civil rights, to political citizenship, obtained during the nineteenth century with the exercising of political rights, and then finally to social citizenship which is granted to all, in the twentieth century, with the triumph of the Welfare State and social rights (minimum salary, health provision, etc.) (Marshall 1977).

For Marshall, the triumph of capitalism does not prevent the implementation of citizenship, this time, full and entire: From then onwards, for the first time, citizenship openly bears an essential sociological dimension, as social redistribution considers the diversity of social situations this time beyond the common role of the citizen. If the civil and political dimensions concerned all citizens in their actual state, independently from their specific social identity, the economical dimension aims at correcting the inequalities amongst citizens. Marshall himself does not underline this modification on the principle of citizenship that thrusts it into the social domain, and hardly interferes with its universalistic dimension. He does not consider either—and was later criticized for this—the persistence of so many social inequalities which remain, even in the age of the Welfare State, and occasionally even worsen, actually threatening the citizenship of the most deprived. In this way, Ralph Dahrendorf stresses 'all those who are rejected by citizenship': 'the noncitizens' who are the immigrants; 'those who are no longer entirely citizens,' i.e., the elderly; and finally 'those who are not yet citizens,' i.e., youth. The erosion of citizenship questions the image of the 'good citizen' equipped with all the necessary attributes for entering into public space. The explosion of the 'underclass' breaks the image of a citizenship which is full and entire, which all people would benefit from in an identical manner in the developed world (Heissler 1994).

From then onwards, a fairly important turnaround of perspective occurred. It no longer involved considering, in a traditional manner, that 'in the Nation State each citizen stands in a direct relation to the sovereign authority of the country' (Bendic 1977). It did not reflect on the conditions of admission to citizenship which separate the 'insiders' from the 'outsiders' (Gunsteren 1988). Neither did it extend this type of reflection which uses citizenship as its foundations for the territory of the nation state by dealing with the case of a postnational citizenship which would take place, for example, in the new public European space where all citizens who have become 'cosmopolitan' would benefit from formal identical rights based upon an intangible constitutional principle, that of 'constitutional patriotism' dear to Habermas, hoping for the emergence of collective mobilizations destined to accentuate the democratic dimension (Cesarini and Fulbrook 1996, Delanty 1997). Instead, the systematic research of elements of a 'differentiated citizenship' was undertaken, taking into account the multiple sociological dimensions peculiar to each citizen, whether it be economical, cultural, or a matter of genus.

For Kymlicka, 'the members of certain groups are incorporated into the political community not only as individuals but also through the group. I have sometimes described these rights as forms of differentiated citizenship' (Kymlicka 1995, p. 174). If many sociologists dealt with the consequences of these socio-economical inequalities on political participation and the exercise of the profession of the citizen, it is even more the dimensions, diagrammatically coming under culture, that increasingly attract attention. In a context of growing crisis within the nation state that greatly effects citizenship in Weber's sense, or even in Marshall's or Bendix's sense, it was suggested that the theory of integration, for example that of Marshall, 'does not necessarily work for culturally distinct immigrants or for various other groups which have been historically excluded from full participation in the national culture—such as blacks, women, religious minorities, gays and lesbians. Some members of these groups still feel excluded from the "common culture," despite possessing the common rights of citizenship' (Kymlicka 1995, p. 180).

Stemming from such a perspective, citizenship finds itself this time plunged into the social with the risk of losing its original meaning: As a result, public space becomes diversified to an infinite extent as citizens preserve their identity there and democracy itself changes its meaning. As Amy Gutmann asks 'What does it mean for citizens with different cultural identities, often based on ethnicity, race, gender or religion, to recognize ourselves as equals in the way we are treated in politics?' (Guttman 1992, p. 3). The democracy of citizens in fact finds itself greatly modified in the same way as the political game, the strategy of parties and of pressure groups who use such identitarian groups as the basis for their actions. Affirmative action, a policy openly destined to smooth out socioidentitarian inequalities by privileging the members of deprived 'ethnic groups,' is the clear outcome which implies a differential management of citizenship and sets down an infinite number of problems of justice and of equality. From the moment

that we consider that individuals possess a 'thick self,' it becomes difficult to claim the recovering of these differences by a 'veil of ignorance' (Rawls [1971] 1987), differences which are deep-rooted within these community memberships considered from now on as being essential.

A considerable literature grew up in the area of multiculturalism, which constantly further reduces the importance of the classical theory of citizenship. By moving onward from the fact that it is unfair to ignore the identity of citizens, the tendency is to reinstate their culture with the risk of slipping toward a deep relativism. In this way, the rediscovery of the identity shared by citizens relegitimizes the particular national cultures, the 'ethnic' feelings and also the nationalistic ideologies (Birnbaum 1996) which consider these neglected ethnic groups to be the foundations of their action in aid of the new citizens who are members of these particular homogenous cultural groups.

Plunged into a growing communitarism, citizenship therefore leads to a nationalistic revival. On an internal level of society, he justifies the 'tribalization' of society into many specific homogenous groups separate from each other: with the image of the working class of yesteryear which constituted a countersociety of which they claimed to be the active citizens, immigrants are supposed to conserve their culture, their language, and their own customs by benefiting from the right to vote, of local citizenship when they have not been naturalized. Women or homosexuals, like all social and cultural minorities, are also invited to join together in a particular manner.

The first among the feminist critics severely emphasized the great indifference of the classical theoreticians of citizenship toward the feminine gender: the French Revolution, which was geared towards the universalism of citizens, did, however, force women to return to the only private space (Landes 1988, Hunt 1992). In a more general way, the woman citizen demands consideration for her body and her own values in the exercise of this role within the public space (Young 1990). In this way, it is in fact the classical theories of deliberative democracy and also the models of the integration of the social system, for example, in their systematic Parsonian presentation, which find themselves questioned (Birnbaum 1996). To take into consideration the sociological variables of citizenship is, therefore, in one way or another, to give an advantage to the 'thick self' to the detriment of the 'thin' self upon which the classical theories of citizenship were formerly built (Walzer 1994, Kymlicka and Norman 1994).

See also: Citizenship and Public Policy; Citizenship, Historical Development of; Citizenship: Political; Democracy; Democracy, History of; Marshall, Thomas Humphrey (1893–1981); Nations and Nation-states in History; State Formation; State, History of; State, Sociology of the; Tocqueville, Alexis de (1805–59); Women's Suffrage

Bibliography

Barber B 1989 *Strong Democracy*. University of California Press, Berkeley, CA

Bendic R 1977 *Nation-building and Citizenship*. University of California Press, Berkeley, CA

Benhabib S (ed.) 1996 *Democracy and Difference*. Princeton University Press, Princeton, NJ

Birnbaum P 1996a From multiculturalism to nationalism. *Political Theory* **24**(1)

Birnbaum P 1996b Sur la citoyenneté. *L'Année sociologique* **46**(1)

Cesarini D, Fulbrook M (eds.) 1996 *Citizenship, Nationality and Migration in Europe*. Routledge, London

Delanty G 1997 Models of citizenship: Defining European identity and citizenship. *Citizenship Studies* **1**(3)

Gunsteren H 1988 Admission to citizenship. *Ethics* July

Guttman A 1992 Introduction. In: Taylor C (ed.) *Multiculturalism and the Politics of Recognition*. Princeton University Press, Princeton, NJ

Heissler B 1994 A comparative perspective on the underclass: Questions of urban poverty, race and citizenship. In: Turner B, Hamilton P (eds.) *Citizenship, Critical Concepts*. Routledge, London

Hunt L 1992 *The Family Romance of the French Revolution*. University of California Press, Berkeley, CA

Kymlicka W 1995 *Multicultural Citizenship*. Clarendon Press, Oxford, UK

Kymlicka W, Norman W 1994 Return of the citizen: A survey of recent work on Citizenship Theory. *Ethics* January

Landes J 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca, NY

Marshall T H 1977 Citizenship and social class. In: Marshall T H (ed.) *Class, Citizenship and Social Development*. Chicago University Press, Chicago

Pateman C 1970 *Participation and Democratic Theory*. Cambridge University Press, Cambridge, UK

Rawls J [1971] 1987 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Shklar J 1991 *American Citizenship*. Harvard University Press, Cambridge, MA

Young I M 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

Walzer M 1994 *Thick and Thin. Moral Argument at Home and* ..... University of Notre Dame Press, Notre Dame, IN

Wolin S 1993 Democracy, difference and re-cognition. *Political Theory* **21**(3)

P. Birnbaum

Civic Culture

Civic Culture theory asserts that democracy is stable or consolidated when specifically democratic attitudes and practices combine and function in equilibrium

with certain non-democratic ones. It was formulated and tested in empirical research in the late 1950s and early 1960s, years still reverberating with the memories of the rise of Communism and Fascism, the collapse of democracy and the catastrophes of World War II. It drew from a contemporary social science literature similarly influenced by the interwar history of the stalemated Third French Republic, the deeply flawed Weimar Germany, the Austrian and Spanish civil wars, and the breakdown of the Fourth French Republic.

It could point to the long tradition in political theory of 'mixed government,' from Plato and Aristotle through to Montesquieu, which supported this anti-populism and prudentialism. Among the post World War II social science influences were works of Joseph Schumpeter, Paul Lazarsfeld and Bernard Berelson, Edward Shils, Robert A. Dahl, and Harry Eckstein, among others.

Schumpeter rejected the 'classic democratic' assumption of the necessity of an informed, activist, rational public for a genuine democracy, and proposed in its place, an 'elites competing for votes' theory (Schumpeter 1947, Chaps. 21 and 22). This minimalist theory could be reconciled with more realistic assumptions of a relatively ignorant and indifferent demos.

Paul Lazarsfeld and Bernard Berelson, theorizing from their 'panel' voting studies of the 1940s similarly saw democracy as associated with a set of cultural and social conditions having the effect of limiting the intensity of social conflict. These included relative economic and social stability, a pluralistic social organization, a basic value consensus. and what we would now call a 'civil society.' They described a democratic equilibrium as involving mixes of involvement and indifference, stability and flexibility, consensus and cleavage (Berelson et al. 1954).

Edward Shils, on the democratic prospects of the new nations, emphasized the importance of a 'widely dispersed civility.' By this he meant a moderate sense of nationality, a degree of interest in public affairs, a consensus on values, institutions, and practices, and a recognition of individual rights and obligations. He wrote ' … These qualities should not be intense, and they need not be either equally or universally shared.' (Shils 1960, pp. 387ff.)

Robert Dahl's theory of polyarchy, elaborated in 1956 in a full-length contrast with populistic and Madisonian democracy, belongs among these early social science influences on the civic culture. His early characterization of the American political system as providing ' … a high probability that any active and legitimate group will make itself heard effectively at some stage in the process of decision … ' (Dahl 1956, p. 145; see also Dahl 1970; Dahl 1989) reflected the minimalist mood that Dahl shared with the generation that emerged out of the great depression and World War II. At the time that *The Civic Culture* was being written, Dahl's concept of polyarchy already had introduced an empirically grounded and quantitative set of concepts into democratic theorizing. Democracy was not an essence. In its full sense it did not exist, and probably could not exist. Hence the concept *polyarchy* was developed to refer to real political entities which attained measured performance levels on specified empirical dimensions.

Harry Eckstein was the first to emphasize the 'mixed' or paradoxical side of democracy, recognizing the necessity for a democracy not only to represent and formulate the will of the public, but to govern it authoritatively. In his *Theory of Stable Democracy* he describes some of the qualities which enable democracies to reconcile responsible authority and democratic responsiveness. Such reconciliation is facilitated by balances among contrasting qualities; participant behavior is balanced by deference to authority, dogmatism by pragmatism, and the like. Institutionally he attributes democratic stability to the degree to which social authority patterns coincide or are 'congruent' with political ones.

Civic culture theory had the advantage of being formulated in the context of a major empirical investigation informed by this historical experience, and benefiting from this prior research and scholarship. Its results were reported in book form in 1963, in paperback form in 1965; and were reprinted in 1989 (Almond and Verba, 1963, 1965, 1989) and remains in print as of this publication. It was widely reviewed in social science periodicals and in 1980 a retrospective volume was published with critiques of the theory and the findings (Almond and Verba 1980).

The data were made available in the Interuniversity Consortium for Political and Social Research at the University of Michigan, and have been utilized in many secondary studies.

Four of the five countries that it investigated were chosen because they exemplified democratic stability and instability in the first half of the twentieth century—the United Kingdom and the United States exemplifying stability on the one hand—Germany and Italy exemplifying democratic instability and breakdown on the other. The fifth, Mexico, was a target of opportunity, selected with the thought that it might provide some insight into problems of democratization outside the North American-European area. The method used in the study, of combining structured and open-ended questions administered to probability samples of national populations, provided a richer set of data specifically responsive to questions arising out of this historical experience and body of speculative theory.

The conception of stable democratic political culture as a 'mixed' political culture received a fuller elaboration than it had been given in the earlier work of Berelson and Lazarsfeld and Eckstein. The mix of democratic political culture was based on political role theory. People in stable democracies were both citizens

and subjects, and they needed to accommodate their non-political, private, and parochial roles. A thriving, stable democracy consists not only of voters, demonstrators, petition signers, and politician button-holers; but of taxpayers, jurors, and military conscripts; as well as parents, mates, work-persons, voluntary association-members, vacationers, and private, self-involved individuals.

It is this mix of roles—participant as well as subject, non-political as well as political—that democratic citizens of a stable democracy must balance and accommodate; and which their institutions must choreograph in a process of converting demands and supports into outputs and outcomes. To shift metaphors, civic culture theory was an equilibrium theory in which political buyers and sellers reach prices at which the political market is cleared. Civic culture theory specified what conditions had to be present in order to clear these markets.

These mixes and balances were located empirically in the British and American cases in the late 1950s and early 1960s—this combination of influentialism and deferentialism, involvement and indifference, conflictual and consensual attitudes, principled and instrumental ones, The relative absence of these balances in the German and Italian cases was noticeable. There was more deferentialism and less participationism in the British case than in the American. A 'reserve of influence' was also noted in the American and British cases, based on the finding that Americans and the British acknowledged the obligation to participate far more frequently than they reported actually participating, This discrepancy between performance and obligation could be viewed as a kind of 'default' mode, a reserve supply of participatory energy available for crises. The civic culture would run cool normally and at a moderate speed, but it had a reserve of influence to draw upon in the twists and turns of democratic politics, as the concerns and interests of different groups of voters were engaged.

By the time the retrospective volume, *The Civic Culture Revisited* was published in 1980 it was evident that British and American civic culture were in trouble (Almond and Verba 1980, Chaps. 5 and 6).

The balance of consensus and conflict had moved toward conflict; pride in nation and confidence in government were down. Participationism had declined. In contrast Germany showed dramatic gains in social trustfulness, confidence in government, and civic competence (Almond and Verba 1980, Chap. 7). In Italy political alienation and extreme partisan antagonism continued largely unchanged. In Mexico the political culture of ambivalent belief in the legitimacy of the democratic revolution, and the corruption of politicians and office-holders, also still survived (Almond and Verba 1980, Chap. 9).

That patterns of political culture would change in response to changes in economy, demography, politics and public policy, communications technology, and popular education should not have occasioned surprise. That the exemplars of the civic culture of the 1950s—Britain and the United States—should be showing signs of wear and tear in the 1970s and 1980s; and that the problem child of democracy—Germany—was showing strong signs of an emerging civic culture, were not causes for rejecting civic culture theory. The question was whether the changes observed in the two decades after the civic culture study, were in a direction which sustained or disproved the theory. The evidence was moderately supportive of the theory. Thus, for example, the withdrawal of Johnson from the 1968 presidential race, despite the fact that political tradition would have legitimized another term of office, and the resignation of Nixon from the presidency in the 1970s were clear evidences of American instability; and the political disorders of the 1960s and 1970s were clear evidences of cultural disequilibria of one kind or another—conflict had undermined consensus, the legitimacy of government had declined, the modes of participation had radicalized. In contrast Germany had had several decades of experience of effective political leadership and remarkable economic growth appearing to produce a moderating *koalition-fähig* partisanship, growing popular trust in government, civic obligation, and the like.

The place of civic culture theory in the contemporary theory of democracy is in some doubt. In the continuing theoretic polemic about the nature of democracy anything settling for less than perfection partakes of sin. It was precisely to avoid this commingling of the sacred and the secular that led Robert Dahl to invent the concept of polyarchy and to place the concept of *democracy* somewhat, but not completely, off limits. However Dahl's Polyarchy III (which is the closest to ultimate democracy that he gets) is achieved through increasing the depth and extent of attentive publics within the larger mass public, corresponding to the significant issues confronting the polity and policy elites. Modern information technology makes it possible that this gap between policy elites and the mass public might be significantly reduced (Dahl 1989, pp. 338 ff.).

The state of the polemic regarding the competence, rationality, and potential effectiveness of mass publics in contemporary polyarchies is well argued in a symposium held during the 1990s. The merits of the various options—elitism, inventive utilization of information technology, or reducing the scope of politics—are among the issues debated (Friedman 1999).

Civic Culture theory might have enriched this discussion somewhat by affirming the legitimacy of other than political claims upon humankind. If one views the full range of role demands on the time and resources of humans, how do we choose among them? How do these choices interact? What are the tradeoffs,

synergies, and opportunity costs? How do we weight the claims of the civic world against the demands of profession, family, edification, and pleasure?

See also: Democracy; Democracy: Normative Theory; Democratic Theory; International Relations, History of; Lazarsfeld, Paul Felix (1901–76); Participation: Political; Schumpeter, Joseph A (1883–1950); Trust, Sociology of

Bibliography

Almond G A, Verba S 1963 *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton University Press Princeton, NJ. Reprinted in 1965 by Boston, Little Brown & Co., and in 1989 by Sage Publications, Newbury Park, CA

Almond G A, Verba S (eds.) 1980 *The Civic Culture Revisited*. Little Brown & Co., Boston. Reprinted in 1989 Sage Publications Newbury Park, CA

Berelson B, Lazarsfeld P, McPhee W 1954 *Voting: A Study of Opinion Formation in a Political Campaign*. University of Chicago Press, Chicago, Chap.14

Dahl R A 1956 *A Preface to Democratic Theory*. University of Chicago Press, Chicago, p. 145

Dahl R A 1970 *Polyarchy: Participation and Opposition*. New Haven, CT

Dahl R A 1989 *Democracy and Its Critics*. Yale University Press New Haven, CT, pp. 338 ff

Friedman J (ed.) 1999 Special Issue: Public ignorance and democratic theory. *Critical Review; An Interdisciplinary Journal of Polilitics and Society* **12**(4)

Schumpeter J A 1947 *Capitalism, Socialism, and Democracy*. Harper and Brothers, New York, Chap. 21 and 22

Shils E 1960 *Political Development in The New States*. Mouton, The Hague, The Netherlands, pp. 387 ff.

G. A. Almond

Civil Law

1. Structural Aspects

In the comparative study of law, legal systems are often seen as belonging to 'legal families.' These may be religious-based legal systems, they may be geographically defined, or they may be distinct because of their structure and methodology. The principal, or characteristic difference between the families of the common law and of the civil law lies in the last of these. This is true despite the fact that both legal families have their roots in the Corpus Iuris Civilis of Justinian (AD 534) and that the common law has preserved some of these traditions to this day, while the civil law has moved closer to the common law in some areas.

With respect to both families, it is dangerous to generalize; there can indeed be marked differences between legal systems belonging to the same 'family.' Thus, the use of codes (for example, the Uniform Commercial Code), the existence of a large body of statutory law, and the adoption of the influential *Restatements* may make the legal system of the United States appear closer to the civil law than to its English roots (see Farnsworth 1996, p. 227). Such a conclusion, however, would misconceive the role and function of statutory law and of the methodology in its application in the civil law, on the one hand, and in common law legal systems, on the other hand.

Codes and statutes consist of rules in all legal systems. These may be very specific or they may be stated with some degree of generality. In some areas, common law codes and other legal texts display extraordinary detail, attempting to anticipate every eventuality. The common law judge, as a consequence, will begin with the factual circumstances of the case and after examining, comparing, and weighing the factual elements of the case, will attempt to find a rule that fits it. This rule may be one of statutory law or derive from prior decisional law. The civilian judge sees the case as a problem to be solved within the legal *structure* of the legal system. First, the problem at hand must be fitted into a legal category. Next, subcategories and sub-subcategories must be identified until 'legal rule' or 'concept' and the problem at hand match. Civilians proceed by *deductive reasoning*, while the common law approach employs an *inductive methodology*.

Categorization, as a structural characteristic of the civil law, also results in the drawing of sharp distinctions between different areas of the law. Private law deals with legal problems that arise between natural persons or between natural and legal persons (such as corporations), while public law (for instance, constitutional law, administrative law, but also criminal law) addresses the relationship between citizen and state. In some civil law systems, this results in the establishment of special courts to deal with these different areas of the law. In Germany, for instance, there are 'ordinary' courts with competence for private law and criminal law, and separate court systems with competence for questions of administrative law, labor law, and social law, each system has its own 'Supreme Court.' Since many cases will obviously involve 'mixed questions' (e.g., of labor and private contract or tort law), there may be contradictory rules of law, with no unifying 'Supreme Court.' In France, the *Conseil d'Etat* stands beside the *Cour de Cassation*, with exclusive competence in certain public fields of law. The exclusivity of the *Conseil's* competence avoids some of the German problems. Germany and Italy have Constitutional Courts, other civil law countries do not. While the divisions are more pronounced in Germany than elsewhere, categorization of fields of law and the establishment of specialized courts unavoidably leads to a high degree of specialization within the legal profession, both among lawyers as

well as judges. These lines are much more fluid in common law.

The central role of the *legal rule* in civil law legal systems also explains why precedent (that is, the binding effect of a prior court decision on a different, but similar subsequent case) is quite different from what it is in the common law. The civilian judge applies the law but is not bound (obviously with some exceptions) by earlier decisions of a higher court. In common law, in contrast, it is the decision of the highest court that ultimately is the law and therefore binds inferior courts. Subsequent discussion will return to these points.

2. History

Roman law had no comprehensive codification before Justinian's Corpus Iuris Civilis. Law, such as it was, was divided into *ius civile* and *ius gentium*. The former applied among Roman citizens, the latter applied to legal relationships among others (Romans and foreigners, foreigners among themselves, and slaves). Judicial functions were exercised by *praetores*. The *praetor prereginus* administered the *ius gentium* and conceived and developed legal concepts unknown to the strict Roman *ius civilis*. In a different context, of course, the English chancellor—at the beginning of the equity-jurisprudence—performed equally creative functions. Similar ideas—for equity and judicial creativity—underlie the 'general clauses' of the civil law (see *vide infra*).

With population growth and increasing urbanization, a new profession—that of the jurisconsult—arose. They were legal advisors who prepared written opinions for cases. As the number of opinions grew, principles could be derived from them which could be taught to students and serve as a basis for advice to judges. The development of principles fosters more abstract ways of thinking. Categories are conceived and problems are classified for assignment to various categories. Under Emperor Justinian, opinions, decisions, and other materials were gathered and the Corpus Iuris Civilis was prepared.

The Corpus Iuris stands at the beginning of what Roman law means today. This Code evolved in subsequent times as a result of the work of the Glossators and Commentators. All taken together constituted the 'Ius Commune,' the common law of Europe before, in subsequent centuries, countries and areas began to grow in different ways and today's 'different families' began to take shape. In Europe, England gave rise to the common law (although, especially in the early period, with many Roman law elements). The Scandinavian countries, France, and Germany had a leading role in developing civil law. Once again, 'civil law' is not homogeneous. There are differences among civil law countries, and these differences are also reflected in the legal systems of those countries that modeled their codes or statutes after earlier rules of others. Two codifications in particular replaced the earlier 'ius commune' and proved very influential: the French *code civil* of 1804 and the German *Bürgerliches Gesetzbuch* of 1896 (see also the Austrian *Allgemeine Bürgerliche Gesetzbuch* of 1811). The French Code influenced the development of legal systems not only in many European (especially the Latin and Eastern); countries, it also spread to the Near East, Central and South America, and even to parts of North America (e.g., Louisiana). The German Code affected the law of Eastern and Southern Europe (e.g., Hungary, the Czech Republic, and Slovakia, Yugoslavia, the Baltic States, and Greece) as well as of Japan and China (see Zweigert and Kötz 1998, 159, 160). Roman Law still exists in Southern Africa as Roman-Dutch law, an admixture of Roman law and old Dutch customary law which has interacted with English common law (see Zweigert and Kötz 1998). Codification extended, in these and in other countries, not only to private law and private-law relationships but encompassed criminal law, commercial law, and both civil (private) and criminal procedure.

3. Sources of Law

3.1 Primary Sources

Civil law systems draw a sharp distinction between primary and secondary sources. Primary sources are enacted law, custom, and 'general principles of law.' Of these, the main source is the enacted (statutory) law; it predominates in civil law systems.

A code in a civil law system consists of general principles, arranged in order of importance. At the beginning there may be general rules regulating basic problems that need to be addressed before the particular problem can be analyzed. For example: if a plaintiff seeks damages for breach of contract, preliminary analysis must determine whether the contract was validly concluded. Provisions dealing with invalidity and avoidance of contracts usually are found in the general part of a civil code. Such a general part may be followed by particular parts dealing with individual fields of law, such as torts, contracts, property, or the law of succession. The main or basic codes are supplemented in increasing number by special statutes or codes of limited coverage with which the legal system reacts to new societal problems, for instance, in areas such as consumer protection, telecommunication, and new media.

Custom is also a primary source of law, but tends to be less important in practice because it is often difficult to prove its pervasive observance in society. Customs are nonwritten rules, developed and observed over years and now part of social and economic thinking.

'General principles of the law' are what the term expresses: basic principles of the legal system which are pervasive of it and derive from norms of positive law. Civil law judges resort to 'general principles of the law' as guidelines in the interpretation of statutory norms both for the purpose of defining their interrelation and for the purpose of their application. This is of particular importance when dealing with statutory norms that are rather abstract in their formulation. It is tempting to consider this process to be not very different from the case law methodology of the common law. There is an important difference, however. The common lawyer derives the appropriate interpretation by reliance on precedent. The civilian judge is not so restricted but derives the interpretation considered to be appropriate from the structure of the legal system and the general principles of law that pervade it; nor will the decision in the present case have a necessary effect on later cases. This is not to say that later cases may not reach the same conclusion: at the point of what French lawyers call *jurisprudence constante* and German lawyers *ständige Rechtsprechung*, such decisional law may itself be regarded as having risen to the position of 'general principles of law.'

3.2 Secondary Sources

Secondary sources consist of case law and the legal literature. The legal literature consists of monographs and contributions to the legal periodical literature as well as commentaries. The last are particularly important in civil law countries that follow or are close to the Germanic legal tradition. Commentaries are detailed annotations of each provision of a particular code, consisting of an analysis that brings together all case law dealing with this provision, opinions of others as expressed in the periodical literature, and the commentator's own evaluation and summary.

Case law by itself, as already mentioned in other contexts, does not have the same central importance in a civil law system as it does in a common law system. It is indeed a secondary source, for the judge is bound only by the enacted law, except in the few cases in which—as discussed—the decisional law has reached the level of *jurisprudence constante* or in systems (for instance, Germany) in which a 'constitutional court' has power to bind other courts. To repeat, however, such cases are rare.

Judges will read commentaries and the legal literature in general, just as lawyers do. The process of 'law finding' and its application, however, does not restrict the court to these sources.

4. Developing the Law

If 'law' is a norm in the form of a statutory codification or a 'general principle' derived from such statutory norms, how can a judge decide a case for which no norm or general principles exist and, in the absence of legislative action, how can law develop further? (For comparative discussion see Capalli 1998, p. 87, Adriaansen 1998, p. 107, Baudenbacher 1999, p. 333).

Legislative action is of course the classic instrument for legal change. But the process, from drafting a bill until its ultimate passage and entry into force, can be long. A commission for the revision of the German law of obligations, for instance, has worked on this project for more than a decade. The final report of the commission was published in 1992. Nevertheless, it was not until 2001 that a bill was introduced in Parliament. Thus, problems may need to be addressed that existing legal norms do not cover. The Swiss Civil Code is unique in its candid grant of discretion to the judge to fill the gap. It provides in Art. 1, § 2: 'if the Code does not furnish an applicable provision, the judge shall decide in accordance with customary law, and failing that, according to rule which he would establish as legislator.' The assumption must be, of course, that even with such a grant of authority, the judge will try to fashion a result that conforms to the general structure and tradition of the legal system. The result will not contravene existing conceptions of what the law should be, but rather will fill the gap, and thereby contribute to the evolution of the law. One way of doing this is to work by analogy, that is, to compare the present problem with other problems that similarly require a weighing of the interests of the parties in the dispute. The analogy then extends the balance struck with respect to other problems of the present case. An example is the treatment of leasing in German law. The German Civil Code contains provisions on leases and for sales, but not for leasing. Since leasing is regarded as possessing elements from both fields of law, case law and literature now derive rules from both fields and apply them to the phenomenon of leasing.

The Swiss provision just quoted confers discretion. Yet it was suggested that no court will exercise unfettered discretion but will seek to follow the structure and to implement the values of the particular legal system. This has very much been the experience in other civil law legal systems, for instance, the German. The German Civil Code, as do others, contains provisions of enormous breadth and, consequently, lack of specificity. These are 'general clauses,' of which § 242 of the German Civil Code is perhaps the best example. It requires parties to a contract to perform their obligations in 'good faith.' Section 242 has become the source of an extensive body of case law and of legal concepts not addressed specifically in the civil code. When the economic upheavals of the 1930s threatened parties to contracts with economic ruin, the German Supreme Court invoked Section 242 to develop a doctrine of 'frustration of contract' (*Wegfall der Geschäftsgrundlage*). This doctrine, interestingly never accepted to that extent in the United States (despite the greater freedom of common law courts to

fashion law through decisions), permits courts in appropriate cases to adjust the obligations of the parties. The French *Conseil d'Etat*—but not the *Cour de Cassation*—developed a similar remedy, the doctrine of *imprévision* (see De Laubadère 1984, § 637). Good faith, so the German court states, requires the protection of both parties: cancellation of the contract (by analogizing the situation to one of impossibility) would be unfair to one party; enforcing it in full (against the background of totally different and unexpected circumstances) would be similarly unfair to the other party.

The obligations of parties to a contract begin with its conclusion. However, expectations will have been created earlier and their disappointment may cause damage. In certain circumstances, such a situation could be dealt with under the law of *Delikt* (tort), but sometimes the preconditions for a remedy in tort will be lacking. Again, the extension of an existing provision may provide relief. Section 276 of the German Civil Code provides for liability for damage caused intentionally and negligently. For liability to arise, there must be some relationship, some duty owed to the other party. Such a relationship, so states the case law, is created by contract negotiations. Parties thus have a *precontractual* duty of care to each other and may have a remedy for its breach. This is the doctrine of *culpa in contrahendo*.

The civilian judge, despite the different theoretical structure and focus of the legal system, thus can and does display the same creativity as do common law judges, with the principal—perhaps only—distinction that the decision in the individual case does not 'make law.' That new concepts can become part of the law, through repeated practice, is shown by the two examples just given. At that point, sources and methodology may differ, the practical result no longer does.

The basic conceptual difference between common law and civil law finds reflection in still another area: the role of a judge and the conduct of a case. In common law jurisdictions, particularly the United States, litigation is 'fact driven.' The facts need to be established and the applicable precedent must then be found. With such an emphasis on the facts of a case, the role of the lawyers is a particularly active one. The judge functions as a neutral arbiter; judges of courts of appeal ultimately decide questions of law. In a civil law court, the judge starts with the rule of law, as outlined, and searches for the facts that he or she needs for the further categorization of the case. The judge's role thus is much more active. Facts are elicited by the judge; lay judges may participate as members of the court (particularly in commercial matters), but there is never a lay jury in private law matters. The judge also ascertains the applicable law and, if the applicable rule of law happens to be foreign law because of the international aspect of the case, may even have to ascertain the content of the foreign rule of law *ex officio*. As a result, decisions of civil law appellate courts will read quite differently from those of appellate courts of common law countries. Attention to factual detail will be very slight (this is particularly true of high French courts) and the emphasis will be on legal norms, rather than the other way round.

5. *Legal Education and the Legal Profession*

In the United States of the eighteenth and nineteenth centuries, the young person could become a lawyer by 'reading law in the chambers of a lawyer,' in other words by serving as a clerk. In the civil law countries of Continental Europe, legal education has been the monopoly of the universities for centuries. With 'doctrine'—not case law—the heart of the legal system, law faculties not only served as the only way of entry to the profession, but its faculty members equaled in stature the position of the high judges in common law systems (see Dawson 1978). (For the practice of '*Aktenversendung*' in Germany in the earlier centuries, see Rheinstein 1938.) German appellate and supreme courts routinely sent case files to the prominent law faculties, virtually for final decision on doctrinal grounds. The influence of law professors on the development of legal doctrine thus has a long tradition and continues to this day.) The course of study reflected, at least until recent times, both the strict categorization of the legal system into specific fields of law and the high level of abstraction that characterizes the deductive analytic method of the civil law. Practical training of the young lawyer was left to a period of time required to be spent in an official training program (such as the *Referendariat* in Germany) or as a junior lawyer under the direction of a licensed lawyer, combined with additional training (such as in France). Full admission to the practice of law then occurs with the completion of these additional requirements. The mindset of the young lawyer had been formed, of course, by the time the university studies were completed, nor would it be changed by the acquisition of practical skills.

In more recent times, law curricula increasingly reflect the offering or requirement of election of interdisciplinary subjects. There is also increasing instruction in the use of new technologies, for instance, how to access and use legal databases. However, to the extent that these offerings represent 'skills training,' they may not be expected to change the structure of the legal system.

6. *Outlook*

Civil law systems differ among themselves, as noted initially. This also holds true for the course of study and practical training required to become a lawyer. As a result, a lawyer—as in the case of other professionals—may face impediments in attempting to es-

tablish a practice in another country or even to render occasional services there. In the European Union, the 1988 so-called 'Diploma Directive' was one of the first steps to implement the European Community's freedom-of-establishment and free-supply-of-services provisions. The 1998 'Establishment Directive' requires foreign lawyers who want to work permanently to pass an aptitude test or to have three years of practical experience in the state where they want to work. The Council of Bars and Law Societies of the European Community (CCBE) is an organization which includes the national associations of lawyers of the member states of the EC. It is the author of the 'Code of Conduct for Lawyers of the European Community.'

In the European Union, civil law meets common law in a shared organizational structure. Efforts to facilitate the transborder practice of law have been accompanied by many legislative acts by the Community that go beyond regulatory economic action but also affect private law. Legal scholars are documenting the doctrinal commonalities in various fields of law—under such titles as 'European-wide tort law' (see Von Bar 1996), or 'Property law in Europe' (Von Bar is currently working on this project)—and there has been discussion concerning the feasibility of a European Civil Code (see Kötz 1996, Lando 1999). If there is, then, some movement toward a new *ius commune* in Europe—which, however, will ultimately extend to a significant additional number of countries as a result of the enlargement of the European Union—many other civil law countries around the world will not participate directly in these developments. But the 'new European law' does not only seek to bridge differences among the civil law countries and between them and the common law; it will also effect reforms and innovations within the civil law. These may well serve as models for other civil law countries, for instance, for those of South America that grow together through similar processes of economic cooperation and integration.

See also: Common Law; Law: Overview; Legal Education; Legal Reasoning and Argumentation; Legal Scholarship; Lex Mercatoria; Supreme Courts

Bibliography

Adriaansen R 1998 Open forum: At the edges of law: Civil law v. common law. A response to Professor Richard B Capalli. *Temple International and Comparative Law Journal* **12**: 107

Baudenbacher C 1999 Some remarks on the method of civil law. *Texas International Law Journal* **34**: 333

Capalli R B 1998 Open forum: At the point of decision: The common law's advantage over the civil law. *Temple International and Comparative Law Journal* **12**: 87

Dawson J P 1978 reprint 1986 *Oracles of the Law*. University of Michigan Law School, Ann Arbor

De Laubadère A 1984 *Moderne & Devolvé, 1 Traité des contrats administratifs*, § 637, 2nd edn. L.G.D.J., Paris

Farnsworth E A 1996 A common lawyer's view of his civilian colleagues. *Louisiana Law Review* **57**: 227

Kötz H 1996 *Europäisches Vertragsrecht*. Mohr, Tübingen, Germany

Lando O (ed.) 1999 *Principles of European Contract Law Parts I and II*. Kluwer International, Dordrecht, The Netherlands

Rheinstein M 1938 Law faculties and law schools: A comparison of legal education in the United States and Germany. *Wisconsin Law Review* **5**: 42

Von Bar C 1996 *Gemeineuropäisches Deliktsrecht*. Beck Publishing Company, Munich, Germany

Zweigert Kötz 1987 *An Introduction to Comparative Law* 3rd edn. Clarendon Press, Oxford, pp. 109–15, 154, 231

P. Hay

Civil Liberties and Human Rights

'Civil liberties' and 'human rights' are closely related terms that embrace the basic freedoms and claims to which individuals are entitled, either as citizens of a particular state or by virtue of being human. The very ideas of civil liberty and human rights presuppose two intertwined convictions: that individuals or groups in civil society have moral status independent of the organized power of society (e.g., the state), and that this power must respect the rights that accrue to this status. Though they arose as a response to political and normative claims in the West, at the beginning of the twenty-first century civil liberties and human rights are at least officially endorsed by virtually all countries and the community of international law. The content and scope of these concepts are contested on philosophical and political grounds. Accordingly, the concepts provide a good method of examining the ways in which important philosophical and legal concepts interact with political and historical forces.

1. Basic Historical Background

The idea of a moral status independent of the state is tied in multifarious ways to notions of higher (or natural) or universal law, democracy, individual conscience, and limited government. These notions have echoes in history as far back as ancient Athens and Rome, and in medieval Christian thought. The modern turn toward individual 'rights' (as opposed to natural 'law') was the product of a complex historical process that included: Renaissance and humanist emphases on human achievement and creativity; the Protestant Reformation's stress on individual religious conscience and religious pluralism; the Enlightenment's belief in the power of reason and the individual,

the growth of markets, and—most importantly—the rise of democracy (see *Law and Democracy*). The specifically liberal tradition of limited government and natural rights arose in the political and intellectual history of several European countries, notably England, Scotland, France, and the Netherlands, and in the United States, in the seventeenth and eighteenth centuries.

Social contract theory in the seventeenth and eighteenth centuries envisioned 'social contracts' between the government and the citizenry based on the consent of the governed and the protection of natural rights. Building on the theory of Thomas Hobbes, John Locke (1988) maintained in his influential *Second Treatise on Government* (a work defending the Glorious Revolution in England in 1688, and the 1689 English Bill of Rights) that government's primary purpose is the protection of the rights of individuals found in the state of nature, specifically rights to life, liberty, and property. The American Declaration of Independence (1776) and the French Declaration of Man and the Rights of Citizens (1789) carried these ideas further, declaring the primacy of civil and political liberties.

2. *Basic Aspects of Civil Liberty and Human Rights*

2.1 *Definitions*

Though the distinction between 'positive' and 'negative' rights (first articulated by philosopher Isaiah Berlin (1969) is often blurred or overstated, civil 'liberties' are often considered negative rights, in that they serve as shields that protect the liberty and rights of individuals and members of civil society from state oppression. They represent claims against state action. Classic civil liberties include freedom of speech, the press, and assembly, freedom of religion, due process and fairness in legal proceedings (especially criminal process), privacy, and freedom from illegitimate discrimination (see *Censorship and Secrecy: Legal Perspectives*; *Discrimination*).

Civil liberties should be distinguished from civil rights. Civil 'rights' are often construed as more 'positive' rights, in that they entail the state bestowing a power to do something affirmative, or taking action to protect fundamental interests or claims against private (nongovernmental) actions. For example, the right to use privately owned public accommodations or facilities, or the right not to be discriminated against in private employment, can be construed as civil rights. More broadly defined positive rights may include such claims as the right to a job, to obtain adequate housing, and to share in more equal distribution of resources. More aggressive state action is needed to effectuate such rights (see *Civil Rights*; *Fundamental Rights and Constitutional Guarantees*).

Civil liberties and rights are generally claims tied to citizenship; in particular, legal orders. 'Human rights' are more universal in nature; they exist simply because one is a human being. These include the civil liberties discussed above, as well as freedom from torture, slavery, and degrading treatment; freedom of the family; and basic self-determination. Debate swirls around whether such rights include basic economic, social, and cultural rights and needs, or broader collective goods.

Though human rights claims can be derived from specific domestic or international legal sources, their claims are distinctively moral. As political theorist Jack Donnelley (1989) remarks, 'Human rights claims are essentially extralegal; their principle aim is to challenge or change existing institutions, practices or norms, especially legal institutions.' Accordingly, Donnelley emphasizes the 'possession paradox': Having a right is most important when 'enjoyment of the object of the right is threatened or denied.' Thus, human rights claims typically arise when a particular claim is not afforded legal protection by a particular country, such as homosexual conduct in some states in the United States, or religious conscience in China.

The distinct concept of 'human' rights arose in the aftermath of World War II, with the widespread condemnation of the atrocities the Nazis committed against Jews and other minorities. After having declined in the wake of skepticism and new political movements in the nineteenth and twentieth centuries (e.g., utilitarianism, emotivism, nationalism, and Marxism), notions associated with natural rights and natural law enjoyed a revival in the aftermath of the war, as democratic theorists regained respect for more objective moral principles that provide standards by which to evaluate the practices of states. Such theorists as Leo Strauss (1950) and Edward Purcell (1973) have written about a postwar 'crisis' in democratic theory along these lines. The term 'human' rights avoided the intellectual and political baggage associated with 'natural' law and rights, while at the same time pointing to universally-held moral principles. In the unprecedented Nuremberg trials held after the war, Allied prosecutors convicted Nazi leaders of crimes against peace and humanity. Also, in the wake of World War II, the new United Nations made human rights an important part of its agenda, and Japan and West Germany, under the aegis of occupying forces, adopted constitutions that protect basic civil liberties and rights.

2.2 *Relationship to the State*

The concept of individual or natural rights is historically and pragmatically related—both positively and negatively—to the emergence of the modern nation states from feudalism between the thirteenth and seventeenth centuries. In 1648, the Peace of Westphalia, which ended the murderous Thirty Years'

War in Europe between Catholic and Protestant states, constituted the first formal international recognition of the nation state's autonomy from religious authority. It also established the first official tolerance of religious pluralism, a crucial move in the rise of civil liberty and human rights. Yet the Westphalian model of international law left no room for the international enforcement of individual rights, as its main objective was the recognition of the principle of territorial sovereignty (domestic jurisdiction) of strong states.

Yet the rise of strong nation states made individual rights more important than they had been in the past, spawning new theories about the obligations of states to citizens. Indeed, another paradox (which also involves the endemic jurisprudential debate between legal positivism and forms of legal analysis based on natural law) concerns the relationship between rights claims and their enforcement or recognition (see *Natural Law*). Though many theorists persuasively contend that rights claims exist independently of legal protection, rights claims (positive and negative) have to be recognized and enforced by those in power in order to be effective. The Nuremberg trials present a classic example of this fact. (Some have called the verdicts 'victors justice.') As James Madison wrote, following the logic of Hobbes and Locke, liberal freedom can exist only when the state is strong enough to protect its citizens, but also limited enough so as not to oppress them. Writing in the aftermath of World War II, Hannah Arendt (1951) chillingly portrayed how vulnerable stateless people are to abuse of their humanity. Legal theorists Stephen Holmes (1995) puts the matter succinctly: 'Weak-state pluralism is a recipe not for liberalism, but for a proliferation of rival and coercive mafias, clans, gangs, and cults … Liberal government … is meant to solve the problem of anarchy *and* the problem of tyranny within a single and coherent system of rules (pp. 270–1).

3. *Some Basic Issues*

Important questions have been raised about the content and intellectual foundation of human rights and civil liberties. What is the scope of rights? Are such rights derived from political or legal agreement, or are they postulates of theological or philosophical inquiry? Do we grasp them by intuition or reason? Which claims are fundamental, and which less fundamental, and how can we make this determination? Should the list of fundamental claims include only basic political and civil liberties, or should it also include social and economic rights? Are rights claims culturally determined or relative, as the American Anthropological Association officially declared in 1947, or are there general principles that make certain claims universal?

Some thinkers, such as Christian Bay (1982), maintain that human rights stem from human 'needs,' which include shelter, food, and livelihood; other emphasize human beings' distinctive moral nature, stressing human dignity, self-respect, and citizenship. The debate concerns those who define human nature in largely materialistic or naturalistic terms, and those who define human nature in terms of such qualities as rationality, moral capacity, or spirituality.

Writers such as Henry Shue (1980), distinguish basic from less basic rights. A right is basic if its enjoyment is 'essential to the enjoyment of all rights.' These rights include physical security, economic security or subsistence, and liberty to participate in the economic and political life of the community. Still others, such as Donnelley, argue that this list is insufficient because a fully developed life requires more opportunities and attributes than these minimums. However, history shows that 'negative' civil liberties are necessary to protect us from the state, so these should always be on the short list of basic rights. If we decide to include more rights as basic, we must do so without sacrificing basic civil liberties. And we must understand that the longer the list of basic rights, the greater the potential for conflict among rights and social policies designed to promote them.

Theorists such as French jurist Karel Vasak (1982) posit 'generations' of rights based on historical development. The first generation consists of political and civil liberties, while the second generation embraces egalitarian social and economic rights. The so-called third generation rights involve humanity as a whole, including cultural self-determination, environmental health, solidarity, and peace.

The founding movements and documents in the rise of liberal democracy accentuated civil and political liberties. Yet the rise of socialism, Marxism, and the working class in the nineteenth century spawned the advocacy of social and economic rights in addition to (or instead of) civil and political rights. In the twenty-first century, such rights are found in the constitutions or fundamental laws of communist (and former communist) states and many developing or Third World States (see *Socialist Law*; *Postcolonial Law*). Though developed liberal states are mainly dedicated to political and civil rights, social and economic rights often comprise parts of their social and legislative policy. In the late 1990s, such internationalist groups as the Lawyers Committee for Human Rights and the Fair Labor Association contend that corporations' use of factories in developing countries has made the protection of economic and social rights in those countries a primary concern.

Human Rights covenants in the United Nations reflect these debates (see *International Law and Treaties*). In 1946, the Economic and Social Council of the UN established the Commission on Human Rights, which led to the Universal Declaration of Human Rights in 1948, a foundational document that has achieved the status of customary international law. In 1966, the International Covenant on Civil and

Political Rights (ICCPR) and the International Covenant in Economic, Social, and Cultural Rights (ICESCR) were signed by most states, taking effect in 1976 (see *Torture: Legal Perspectives*). The ICCPR protects such basic civil liberties as freedom from arbitrary punishment, forced servitude, and unfair criminal process; freedom of thought, conscience, and religion; personal liberty and security; freedom of the family; freedom to participate in fair elections; and equal suffrage. An Optional Protocol of the ICCPR commits ratifying states to allow a special committee of experts to examine claims by individuals against them.

The ICESCR covers such rights as the right to work under good conditions; the right to an adequate standard of living, and the right to social security, food, clothing, shelter, and basic health. The ICESCR is less stringent in its wording than the ICCPR. Signatories established no rank ordering of these rights, and enforcement is more a matter of exposure and persuasion than force.

4. *Civil Liberties in Practice*

The protection of civil liberties varies (in legal provisions and applications of these provisions) in different countries due to cultural and political factors (see *Fundamental Rights and Constitutional Guarantees*; *Constitutionalism, Comparative*). We can cite just a few of many examples. For instance, free speech doctrine and practice in the United States protect the advocacy of illegal action, including racist rhetoric, that falls short of directly triggering a disturbance of the peace or inciting violence. Canada, Israel, Germany, France and many other countries, on the other hand, prohibit speech that advocates racism (racist rhetoric), regardless of the likelihood of illegal action. In Germany it is illegal to belong to the Nazi party or to wear a Nazi uniform, while courts in the United States have expressly protected such actions.

Though virtually all countries have basic legal protections for criminal suspects, standards vary widely, especially when we look at practice rather than the letter of the law. In Russia, where legal institutions are poorly developed, preventive detention and criminal procedure rights of criminal suspects are poorly enforced despite legal protections to the contrary; France does not recognize the privilege against self-incrimination. In the area of religion, the United States maintains an exceptionally strict separation of church and state, while such democracies as Ireland, Italy, and Germany allow more accommodation between state and religion. In India, 'personal laws' linked to the major religions (Hindu, Muslim, Christian, and Parsi) are distinguished from normal civil law, thereby accommodating culturally-based discrimination against women in such areas as marriage, divorce, and inheritance. And communist China has persecuted such religious groups as Christians and the Falun Gong because their views are allegedly contrary to state ideology.

Social scientists and legal scholars cite several factors that influence the extent to which countries will support civil liberties and rights on a sustainable basis. Commitment to civil liberty has historically been accompanied by social pluralism, legal institutions based on rule of law, and the differentiation of the state from civil society (see *Rule of Law*; *Law and Development*). More specific explanations include such factors as the existence of a bill of rights and judicial independence, judicial leadership and control of cases dockets, and a culture of rights consciousness that encourages citizens to think in terms of rights (see *Judicial Review in Law*; *Legal Culture and Legal Consciousness*; *Rights: Legal Aspects*).

More recent explanations point to the presence of political and social movements that engender legal change (e.g., the Civil Rights Movement in the United States; the freedom movement of South Africa led by Nelson Mandela). Recently Charles Epp (1998) has pinpointed sustained pressure exerted by a 'support structure for legal mobilization,' which consists of rights-advocacy organizations and lawyers, and sufficient funding from private and (especially) public sources. Rights revolutions succeeded in the United States and Canada in recent decades because of the presence of these factors, while India's rights movement was thwarted despite a favorable Supreme Court because such factors were absent (see *Law as an Instrument of Social Change*).

5. *Human Rights in International Practice*

Until recently, the international system remained committed to the Westphalian model's strong presumption in favor of the 'domestic jurisdiction' of states. But this situation slowly began to change with the growth of consciousness of human rights, the democratic ethic, and globalization (see *Globalization: Legal Aspects*). The most important events before the end of World War II and the Nuremberg trials include: the abolition of slavery in the British empire in the 1830s and 1840s, culminating in the League of Nations' Slavery Convention of 1926; the policy of 'humanitarian invention' by Western states to protect Christian citizens abroad in the nineteenth century; and several conventions and treaties protecting the rights of soldiers in war promulgated in nineteenth and twentieth centuries.

In the wake of World War II and the humiliating failures of the League of Nations, the world community established the United Nations, which promulgated the UN Charter. Whereas the Westphalian model is premised on the freedom of states over their

domestic jurisdictions, the 'UN Charter' or 'new international law' model embraces the Kantian model of international relations and law, which emphasizes universal peace and human dignity. In reality, the models coexist in the contemporary world, posing sometimes vexing questions about where to draw the line between state sovereignty and international human rights norms. Yet the conferees who established the Charter rejected a proposal to authorize intervention to protect rights, and a clause in the Charter expressly prohibits intervention in 'matters which are essentially within the domestic jurisdiction of states.'

In recent decades, many international treaties and forums have been established under the aegis of the UN or other regional and international organizations to promote recognition of human rights (see *International Law and Treaties*). States have signed treaties in conventions against torture, genocide, racial and gender discrimination, and treaties protecting refugees and children. Again, social and political movements have been indispensable to the promotion of human rights logic and practice. In addition to hundreds of regional and intergovernmental organizations ('IO's, such as the Organization of American States, the Organization of African Unity, etc.), such nongovernmental organizations (NGOs) as Amnesty Inter national, Human Rights Watch, and the International Committee of the Red Cross have played major roles in raising awareness, linking international organizations, and even bringing cases for enforcement in relevant jurisdictions.

States have drafted regional agreements to protect rights on all continents except Asia. In Europe, the European Convention of Human Rights and Fundamental Freedoms (based on the ICCPR) formed the European Court of Human Rights, which takes cases after they have been heard by the relevant domestic courts (see *European Union Law*). Member states have agreed to accept all of the court's rulings, leading, for example, to changes in Britain's law of criminal procedure (most notably in the area of pretrial detention) and changes in several states' laws concerning the rights of children born out of wedlock.

Not surprisingly, politics has affected the application of the ICCPR and ICESCR covenants. During the Cold War era, Western countries championed the ICCPR, while communist countries supported the ICESCR. Third World countries have advocated rights of self-determination, cultural rights, and collective rights concerning resources (debates over law of the seas, etc.). Cultural relativism remains an issue. In 1993, Asian countries challenged claims about the universality of political and civil rights during the UN World Conference on Human Rights, arguing that such rights can be counterproductive, even dangerous, in the context of economic underdevelopment, fragmented nationalism, and fragile state institutions. Theorists such as Donnelley counter by pointing out that these arguments ignore the actual plights of persecuted minorities in these countries, thereby serving the interests of entrenched elites or tyrants.

Enforcement of human rights at the international level has remained problematic because of the continued normative and prudential reluctance to intervene in the domestic jurisdiction of states. Hobbes and Locke would have predicted such a result in the absence of an international sovereign with sufficient power to enforce protections of rights. As a result, the main support for human rights has been in the form of moral persuasion (far from meaningless, if not always efficacious), exposure of violations through research and publication, and the deployment of such measures as economic sanctions. UN organizations have investigated several countries, including Chile, Rwanda, Somalia, Zaire, several Latin American countries, Iran, Iraq, and South Africa.

6. *Victories and Defeats: The Continuing Dilemma of Rights and the State*

The decline of Cold War politics in the UN Security Council has enabled the UN to be somewhat more aggressive, sponsoring peacekeeping and actual interventions to protect human rights in Somalia, Iraq, and Bosnia in the early 1990s. Yet the efforts in Somalia and Iraq have proved short-lived, and Bosnia's fate was quite uncertain as of 1999. And in 1994, the world stood by while massive genocide took place in Rwanda. In 1999, Cuba, China, and Sudan championed the norm of the territorial sovereignty of states in order to shield their abuses of human rights from international intervention, even though these states are themselves members of the United Nations Human Rights Commission.

To be sure, Yugoslav President Slobodon Milosevic was defeated in his attempt to take over Kosovo in 1999, yet this victory was won by the military commitment and might of the North Atlantic Treaty Organization under the leadership of the United States and Britain; victory came only after Serbian forces had already carried out massive ethnic cleansing. During the war, the war-crimes tribunal in the Hague indicted Milosovic for war crimes; earlier (1995) it had indicted Serbian leader Radovan Karadzic and his top military officer, General Ratko Mladic. Yet no political body is in charge of bringing such indictments, and the United States and leading European countries have not managed to arrest these Serbian leaders, preferring instead to target lesser figures as of this writing.

After Kosovo, Czech President Vaclav Havel wrote in 'Kosovo and the End of the Nation-state' that the nation state would end in the next century, giving away to an international community governed 'by universal or global respect for human rights, by universal civic equality and the rule of law, and by a global civil society'. Writer Leon Wieseltier (1999) replied that no oppressed soul had ever been saved by

the forces of 'global civil society.' Kosovo was delivered from Milosovic by the willful acts of allied nation states. Though the nation state is a source of evil, 'it is also the nation-state from which we may demand rescue from such evils. The ethical content of a particular sovereignty is what finally matters.'

See also: Censorship and Secrecy: Legal Perspectives; Civil Rights; Civil Rights Movement, The; Constitutionalism, Comparative; Discrimination; European Union Law; Freedom/Liberty: Impact on the Social Sciences; Freedom: Political; Fundamental Rights and Constitutional Guarantees; Globalization: Legal Aspects; Human Rights, Anthropology of; Human Rights, History of; Human Rights in Intercultural Discourse: Cultural Concerns; Human Rights: Political Aspects; International Law and Treaties; Judicial Review in Law; Law and Democracy; Law and Development; Law as an Instrument of Social Change; Legal Culture and Legal Consciousness; Legal Positivism; Natural Law; Rights: Legal Aspects; Rule of Law; Socialist Law; Torture: Legal Perspectives

Bibliography

Arendt H 1951 *The Origins of Totalitarianism*. Harcourt, New York

Bay C 1982 Self respect as a human right: Thoughts on the dialectic of wants and needs in the struggle for human community. *Human Rights Quarterly* **4**: 53–75

Berline I 1969 Two concepts of liberty. In *Four Essays on Liberty*. Oxford University Press, London

Cassese A 1986 *International Law in a Divided World*. Clarendon Press, Oxford

Conot R E 1983 *Justice at Nuremberg*. Harper & Row, New York

Donnelley J 1989 *Universal Human Rights in Theory and Practice*. Cornell University Press, Ithaca, NY

Epp C R 1998 *The Rights Revolution: Lawyers, Activitists, and Supreme Courts in Comparative Perspective*. University of Chicago Press, Chicago, IL

Henkin L 1990 *The Age of Rights*. Columbia University Press, New York

Holmes S 1995 *Passions and Constraints: On the Theory of Liberal Democracy*. University of Chicago Press, Chicago, IL

Lawson E (ed.) 1991 *Encyclopedia of Human Rights*. Taylor and Francis, New York

Locke J 1998 *Two Treatises of Government* [ed. Laslett P]. Cambridge University Press, Cambridge, UK

Purcell E A Jr 1973 *The Crisis of Democratic Theory: Scientific Naturalism and the Problem of Value*. University Press of Kentucky, Lexington, KY

Shue H 1980 *Basic Rights: Subsistance, Affluence, and US Foreign Policy*. Princeton University Press, Princeton, NJ

Steiner H J, Alston P (eds.) 1996 *International Human Rights in Context: Law, Politics, Morals*. Clarendon Press, Oxford, UK

Strauss L 1950 *Natural Right and History*. University of Chicago Press, Chicago, IL

Vasak K (ed.) 1982 *The International Dimensions of Human Rights*. Greenwood Press, Westport, CT

Wieseltier L 1999 Winning ugly. *The New Republic* June 28: 27–33

D. A. Downs

Civil Religion

In a seminal thesis published in the Winter, 1967, issue of *Daedalus*, and in later revisions of his argument, Bellah 1967 claimed to have discerned two kinds of civil religion in the USA. One was fairly traditional: a composite of biblical themes that were compatible with the natural law tradition mediated by the church. This legacy saw in the history of the US a version of God's election of ancient Israel. That is why Bellah referred to it as a 'special civil religion' that was compounded of prophetic warnings and commands to a chosen nation burdened with particular rights and responsibilities. The other civil religion was utilitarian rather than traditional, and it was based primarily on the thought, interests, and experiences of the American people themselves. Describing it as 'the lowest common denominator of church religions,' Bellah argued that it paid more attention to the interests than to the responsibilities of the people. Interpreted in terms of the social contract rather than of the covenant, it owed far more to John Locke than to the Bible (Bellah 1976a, p. 57).

Bellah alternated between arguing that the civil religion was vital and enduring and issuing warnings that the civil religion was in a precarious state. Bellah at first seemed to be sure that the civil religion was alive and well: far from dead, it would show its vitality during the American Bicentennial of 1976 (Bellah 1973). However, in reflecting on that celebration and also on the latest Presidential campaign addresses, Bellah was quite clear that Americans had largely forgotten about a past that they had never clearly understood.

Bellah himself alternated between thinking that his article coincided with the debut of the civil religion on the American scene and with its decline. On the one hand, he suggested that the civil religion had only come into being through his publication of the 1967 article and appeared confident in its continued existence (Bellah 1973). On the other hand, disappointed in the results of the Bicentennial, he regretted that it was only an 'empty and broken shell' (Bellah 1975).

Some religious leaders have condemned American civil religion as authoritarian, dangerous, and idolatrous. They are joined in this criticism by some leading politicians who see a civil religion as idolatrous. Some critics, therefore, argued that Bellah was engaged in an attempt to revitalize the nation itself by infusing its political institutions with religious meaning. Bellah merely wished, through exhortation, and

admonition to recall the nation to its higher purpose (Crouter 1990). To others it was quite clear that Bellah was trying to resuscitate Protestant beliefs and public influence at a time when both seemed to be losing credibility and support. At the very least he was trying to invoke the legacy of the Protestant establishment of the late nineteenth century. 'Bellah introduced theological principles that he presumed overarched the state and the religions it protected' (Hammond et al. 1994 pp. 8–9). These criticisms persisted, despite Bellah's disclaimer that the civil religion was merely another way of talking about a world view (Hammond et al. 1994, p. 2).

Bellah's claim to have identified a civil religion that endures regardless of those who believe in it or who can verify its existence has elicited criticism or comments from those who see his work not as sociology but as political theology or ideology. Indeed, some noted that his interpretation of American religious and political ideals found its justification in theological propositions (Hammond et al. 1994, pp. 8–9). Perhaps in response to these critics, Bellah has argued that the civil religion needed no help from himself or from anyone else; it could endure on its own terms. On one occasion Bellah insisted that the civil religion had never been a majority viewpoint but continued to exist enshrined in certain texts, particularly Lincoln's Gettysburg address. It, therefore, did not matter how many people believed in the civil religion or whether evidence could be found for it through the use of questionnaires. The civil religion was there, Bellah (1976b, pp. 153–4) argued: a matter of 'faith in certain abstract propositions which derive ultimately from God. If the "larger society" does not conform to them, so much the worse for it' (Bellah 1976b, pp. 153–4).

Bellah himself has pointed out that his case for the American civil religion is quite in keeping with a Durkheimian approach to social life. Societies do express their identity and define themselves in religious terms; indeed any enduring form of social life may well become serious about its foundations, standards, boundaries, and destiny. The sacred is a pervasive aspect of social order. No wonder, then, that for Bellah (1989), while the concept of the civil religion may be dispensable, it nonetheless points to an enduring problem concerning the relation of the political to the religious aspects of any society. (Marty 1974).

To sum up Bellah's many and quite variant readings of the civil religion and of his own arguments, it is helpful to distinguish two sets of axioms. These are shared among sociologists who study religion and are not idiosyncratic to Bellah alone, but Bellah does provide an example of how they may operate in the thought of a single sociologist. On the one hand, Bellah identified himself as working within a set of assumptions that he would attribute to Durkheim. On the other hand, Bellah would also acknowledge the validity of a Weberian viewpoint, that thinks of charisma as being somewhat ephemeral or evanescent and always distorted by any attempt to make it part of a routine or rational social order. Religion, on this view, is disruptive to institutions and is, therefore, most evident during times of crisis or chaos. Bellah would therefore not be surprised that that some sociologists, therefore, have found in the civil religion an episodic phenomenon that is most visible in times of crisis and is only gradually rooted in enduring institutions or ways of life (Marty 1974).

The debate on civil religion thus reflected a wide range of assumptions about the relation of religion to complex societies. Bellah, in keeping with his Durkheimian assumptions, saw the US as largely individualistic and utilitarian as a result of a break with the Biblical tradition. In his view the Constitution owed more to Locke and to notions of interest than to covenantal theology; for Bellah, that break represented a considerable decline in the moral vision of the nation.

On the other hand, also in keeping with a Durkheimian interest in a religion of humanity, Bellah and his associate, Philip Hammond, saw more universal possibilities for the civil religion. All societies express their political unity in civil religious terms. Their critics, however, accused them of expressing an American notion of manifest destiny under the guise of the civil religion (Weddle 1983).

Despite the protest by Hammond et al. (1994, p. 2) that Bellah did not have in mind an 'idolatrous worship' of the American nation–state, Bellah argued that American culture has within it the potential of becoming the basis for a global civil religion. It was a point in keeping with his Durkheimian interest in a religion of humanity, and others have found some support for his thesis. There is some evidence, for instance, that the US space program transformed astronauts from American celebrities into representatives of a global civil religion (Wilson 1984). Furthermore, the US is apparently not peculiar in its use of national heroes as exponents of a civil religion; similar processes appear to be at work in Yugoslavia (Flere 1994).

Subsequent studies have suggested that religion may still be engaged with the political system, but far less at the national than at the local level. Rather than influencing a broad range of social values, religion is now more likely to be engaged in interest-group and single-issue politics. Instead of a steady pull on the direction of social change, religion therefore increasingly exerts a temporary, however intense, influence during spurts of social mobilization on the part of particular communities and constituencies (Demerath and Williams 1992).

Some have argued that the civil religion is no longer a national conscience but a set of partial ideologies. It is, therefore, merely 'a confusion of tongues speaking from different traditions and offering different visions of what America can and should be' (Wuthnow 1998).

That is perhaps why Bellah's more universal claim for the civil religion has aroused another set of criticisms to the effect that the civil religion, at least as Bellah conceived it, ignored the presence and claims of minorities and smacked of 'cultural imperialism' (Moseley 1994, p. 18).

Beyond the context of the US, however, scholars have found the notion of a civil religion to be a particularly suggestive concept. Despite—or because of—Bellah's 'broad and diffuse use of the term,' civil religion, and the 'theoretical instability' of Bellah's model, there has been a proliferation of studies of civil religion in a wide range of national contexts (Crouter 1990, p. 161).

Drawing on Dobbelaere's (1986) discussion of the civil religion, it would be possible to distinguish four conditions under which religion would have different relationships to a national political system. Where religion is still an enduring and vital institution, and where it is still central to the nation–state, one would expect to find such traditional forms of civil religion as Islam and Eastern Orthodoxy. However, where traditional religion has been eroded or transformed, one might expect to find more secularized cultural systems like Soviet Marxism or Nazism. More ephemeral or episodic forms of religion might persist and remain central to a nation–state; consider the notion of a civil religion that is episodic and not widely known but still central to the history, traditions, and identity of the US.

Dynamic relationships between the center and the periphery, however, will change the meaning and location of civil religious beliefs and symbols. In Japan, since the nineteenth century, Shinto has moved from the center, where it was the civil religion of an aristocratic and military elite, to the villages and clans, where it has strengthened national resistance to Western influence. Whereas under the Meiji, Shinto had been central and enduring as a national civil religion, under occupation by the US Shinto was no less enduring, however marginal it had become to the public ideology of the nation (Takayama 1988, p. 328).

Although Japan's national culture was transformed into one that was largely secular and democratic, Shinto remained vital to the ability of Japan to recover from the war and to rebuild its economy while preserving a sense of continuity with the past. However, as Shinto has been adopted by the corporations and the nation–state as a means of mobilizing the loyalty of workers and citizens, it has become more secular and arguably less significant to the mobilization and motivation of Japanese citizens and workers (Dobbelaere 1986).

The beliefs and symbols of the civil religion may therefore become the source of legitimacy for the nation–state or the target of cultural opposition. On the one hand, an ideal of an ethnically pure nation–state may legitimate the most brutal forms of ethnic cleansing. As in the case of Yugoslavia, it would be a mistake to underestimate 'the power of a system of reified, prescriptive culture to disrupt the (contradictory) patterns of social life' (Hayden 1996, p.784). Similarly, in Chile a form of the civil religion reinforced by the church and articulated by the Pinochet regime sought to give religious legitimacy to a repressive military elite and regime (Christi and Dawson 1996). Such attempts, however, associate civil religion with a regime rather than with a nation–state as a whole and, therefore, place it in a more marginal location. The same observation applies to attempts by the dominant regime in Malaysia to use Islamic beliefs and practices to reinforce social discipline and to legitimate the political and economic goals of the state (Regan 1976, p. 103).

As nations become increasingly secularized, the civil religion, to the extent that it survives, will become marginal to the culture and politics of the nation–state and put in only episodic appearances during periods of social mobilization around specific issues. For instance, in Celtic heroes and symbols were central to the resistance of France, Belgium, and the Netherlands against Germany in the nineteenth century. Recently, however, France and Spain have met with resistance from Celtic communal groups on their own peripheries. The symbols and identity of the Celtic periphery are thus being co-opted by the European Union in an attempt to assert a Pan-European communal culture (Dietler 1994).

As the cases of Northern Ireland and Wales demonstrate, there is no necessary connection between strong 'nationalist doctrines' on the periphery and an equally strong 'nationalist politics' (Beuilly 1985, p. 74). In the US, marginal groups that wished to define themselves in opposition to the national center that seemed to them to be insufficiently legitimate have established their own alternative versions of the civil religion: the Seventh Day Adventists being one example (Bull 1989, p. 181).

Some have, therefore, argued that the civil religion is an essential arena for the contest among opposing communal and interest groups and between the center and the periphery, (Willaime 1993, p. 573). There is an intimate connection between civil religious protest on the periphery and oppositional politics toward the center. Indeed, it would appear that civil religion comes into being as a way of arbitrating the protest of local or communal groups against the state. Mexican–Americans, for instance, have joined elements of Catholic liturgy with traditional folk celebrations to mobilize and discipline farm-workers for an agricultural union movement in the United States. The movement used traditional religious symbols that had been central to Mexican culture for the purpose of opposing a dominant class and its institutions in the US (Bennett 1988). Indeed, highly sectarian forms of the civil religion have been developed in order to protest the secularization of the civil religion at the

political center; the Unification Church would be a prime example of this authoritarian tendency (Robbins et al. 1976). Similarly, representatives of Native American communities have attacked a politicized civil religion at the center as being inimical to the Indian traditions (Deloria 1992).

As Bryan Wilson and others have argued, there are strong secularizing tendencies in the Christian faith, and these have been deployed on various occasions against civil religion. In Sri Lanka the role of the King has been crucial in maintaining social and cosmic harmony; even the British took the role of the King in the most important ritual of Sri Lankan civil religion, until protest from missionaries forced them to cease their involvement. As a result, the ceremony has been degraded into a festival of Sri Lankan cultural arts, and the society as a whole has lacked the means of legitimating the center to the periphery, (Seneviratne 1984).

There is some evidence that the decay or removal of the civil religion from public discourse in some countries has created space for a secular civil society to emerge. In Norway, civil religious symbols have been used by conservative and Christian elites to legitimate the monarchy, but there is a general tendency to keep religion on the margins of public life; even during times of crisis civil religious symbols may be notable by their absence from public discourse. As they remain on the periphery, however, the symbols of civil religion may be deeply, if not widely held. In Norway, although public discourse is secular and political legitimacy is derived from the legal system, the Christian communities of the south-west still maintain a hope that Norway may become a Christian nation (Furseth 1994 pp. 46, 50–1). Although Canada lacks a civil religion, religious symbols and beliefs remain vital to regional and local communities (Reimer 1995).

On the other hand, some have argued that it is the very secularization of the center that raises fears that the society as a whole may disintegrate; these fears are the source of demands for a revitalized civil religion (Willaime 1993, p. 571). As Martin (1997, pp. 28–35) has pointed out, for a society to survive it must have continuity, to achieve which it must maintain a certain identity. That identity, furthermore, always implies a difference between the society in question and all others. By having a cultural as well as political center, a society forecloses other possible bases for integration. The full range of possibilities, and the uncertainty of making choices among alternative identities, is not open to any society that wishes to maintain its identity, and therefore also its difference from other societies, over time.

Once this is understood, Martin (1997, pp. 29–35) argues, it is no longer problematical that religion may become one of the markers of identity or difference. Whether religion becomes an exclusive marker or becomes associated with other forms of the sacred in a particular society depends on a wide range of circumstances that are both historical and geopolitical. Thus, it is not unusual that traditional forms should be co-opted by chauvinism or become a source of peripheral resistance to the center.

In the course of trying to clarify the civil religion thesis, various proponents have argued that there are a wide range of types based either on the content of the ideology or its social constituency. Martin, however, has pointed out the free-floating nature of the sacred and its contingent relation to religion and politics. Rather than developing complex typologies to encompass these relations, what is needed, therefore, is a series of statements of the 'If … then …' variety that stipulate the conditions under which civil religion may be more or less central, marginal, traditional, secularized, popularized, or politicized. It is appropriate to ask to what extent religion is 'locked into the core processes of cohesion, power, and control' and to investigate the extent to which religion and its dominant institutions maintain their traditional 'relation to territory and history, to national belonging and death' (Martin 1997, p. 104). One must also ask whether the religious community in question has a voluntaristic or ethnoreligious base and whether its symbolic options are more generic or particular. The answers to these questions will then help one interpret each community's or nation's construction of 'the world' as relatively open or closed, hostile or indifferent (Martin 1997 p. 56).

Further work, therefore, remains to be done on civil religion as an expression or outgrowth of conflict within and between civilizations. On this level, it remains to interpret civil religions as cultural developments of civilizations translated from their centers to new peripheries. Following Martin, it would be possible to view the American civil religion as a continuation and residue of the English civil war over a century earlier, just as conflicts between the North American center and its Central or Latin American periphery are continuations of the struggle between Northern and Southern Europe. On the basis of Martin's argument one could interpret the American civil religion as the response of one periphery to the English center: a response composed partly of a tradition of establishment and partly of voluntaristic and ethnoreligious dissent (Martin 1997, p. 57). Similarly, one could investigate Japanese Shinto and Sri Lankan civil religion as the development of Buddhism on Asian peripheries.

If arguments concerning the civil religion are to contribute to mainstream social scientific discussions of nationalism however, further work remains to be done on civil religions as the result of the 'civilizing process' as religion is transplanted from the center to periphery. Relatively few scholars interested in civil religion have focussed on the work of Anderson (1983), who has traced the interaction of Western Christianity with a wide range of societies in both hemispheres. In Anderson's view, Roman Christian-

ity, carried by the Church throughout Europe, had integrated a wide range of local cultures and religions within a common civilization and by means of a *lingua franca* (Latin). In developing local vernaculars into which the Bible and liturgy were then translated, however, the Church succeeded in encouraging indigenous elites to develop a national culture resistant to the imperial center. These smaller and more cohesive entities thus represented a limited and secular reduction of the religious civilization that created them. Thus civil religions, by this argument, are secularized remnants of a trans-national religious civilization.

See also: Citizenship: Sociological Aspects; Civic Culture; Religion and Politics: United States; Religion: Mobilization and Power; Religion: Nationalism and Identity; Religion: Peace, War, and Violence

Bibliography

Anderson B 1983 *Imagined Communities, Reflections on the Origin and Spread of Nationalism*. Verso, London

Bellah R N 1967 Civil religion in America. *Daedalus*, *Journal of the American Academy of Arts and Sciences* **96**: 1000–21

Bellah R N 1973 American civil religion in the 1970s. *Anglican Theological Review , Supplemental Series* **1**: 8–20

Bellah R N 1975 *The Broken Covenant*. Seabury, New York, pp. 142–158

Bellah R N 1976a The revolution and the civil religion. In: Bauer J C (ed.) *Religion and the American Revolution*. Fortress Press, Philadelphia

Bellah R N 1976b Response to the panel on civil religion. *Sociological Analysis* **37**(2): 153–9

Bellah R N 1978 Religion and legitimation in the American republic. *Society* **15**(4): 16–23

Bellah R N 1980 The five religions of modern Italy. In: Bellah R, Hammond P (eds.) *Varieties of Civil Religion*. Harper & Row San Francisco

Bellah R N 1989 Comment. *Sociological Analysis* **50**(2): 129–46

Bennett S 1988 Civil religion in a new context: The Mexican–American faith of Cesar Chavez. In: Benavides G, Daly M W (eds.) *Religion and Political Power*. State University of New York Press, Albany, NY

Beuilly J 1985 Reflections on nationalism. *Philosophy of the Social Sciences* **15**: 65–75

Bull M 1989 The Seventh-day Adventists: Heretics of American civil religion. *Sociological Analysis* **50**(2): 177–87

Christi M, Dawson L 1996 Civil religion in comparative perspective: Chile under Pinochet (1973–1989). *Social Compass* **43**(3): 319–38

Crouter R 1990 Beyond Bellah: American civil religion and the Australian experience. *The Australian Journal of Politics and History* **36**(2): 155–65

David D H 1998 Editorial: Civil religion as judicial doctrine. *Journal of Church and State* **40**(1): 7–24

Deloria V 1992 Secularism, civil religion, and the religious freedom of American Indians. *American Indian Culture and Research Journal* **16**(2): 9–20

Demerath N J, Williams R H 1992 *A Bridging of Faiths: Religion and Politics in a New England City*. Princeton University Press, Princeton, NJ

Dietler M 1994 'Our ancestors the Gauls': Archeology, ethnic nationalism, and the manipulation of Celtic identity in modern Europe. *American Anthropologist* **96**(3): 584–605

Dobbelaere K 1986 Civil religion and the integration of society: A theoretical reflection and an application. *Japanese Journal of Religious Studies* **13**(2–3): 127–45

Flere S 1994 Le Développement de la sociologie de la religion en Yougoslavie après la deuxième guerre modiale (Jusqu'a son démembrement). *Social Compass* **41**(3): 367–77

Furseth I 1994 Civil religion in a low key: The case of Norway. *Acta Sociologica* **37**: 39–54

Hammond P E, Porterfield A, Moseley J G, Sarna J D 1994 Forum: American civil religion revisited *Religion and American Culture. A Journal of Interpretation* **1**: 1–7

Hayden R M 1996 Imagined communities and real victims: Self-determination and ethnic cleansing in Yugoslavia. *American Ethnologist* **23**(4): 783–801

Martin D 1990 *Tongues of Fire, The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK

Martin D 1997 *Does Christianity Cause War?* Clarendon Press, Oxford, UK

Marty M 1974 Two kinds of civil religion. In: Richey R, Jones D (eds.) *American Civil Religion* 1st edn. Harper & Row, New York

Moseley J 1994 In: Hammond P E, Porterfield A, Moseley J G, Sarna J D (eds.) *Forum: American Civil Religion Revisited. Religion and American Culture, A Journal of Interpretation* **Winter 4**(1): 13–18.

Neuhaus R J 1986 From civil religion to public philosophy. In: Rouner L S (ed.) *Civil Religion and Political Theology*. University of Notre Dame Press, Notre Dame, IN, pp. 98–110

Regan D 1976 Islam, intellectuals and civil religion in Malaysia. *Sociological Analysis* **37**(2): 95–110

Reimer S H 1995 A look at cultural effects on religiosity: A comparison between the United States and Canada. *Journal for the Scientific Study of Religion* **34**(4): 445–57

Richey R, Jones D (eds.) 1974 *American Civil Religion*. Harper & Row, New York

Robbins T, Anthony D, Doucas M, Curtis T 1976 The last civil religion: Reverend Moon and the Unification Church. *Sociological Analysis* **37**(2): 111–26

Seneviratne H L 1984 Continuity of civil religion in Sri Lanka. *Religion* **14**: 1–14

Takayama K P 1988 Revitalization movement of modern Japanese civil religion. *Sociological Analysis* **48**(4): 328

Weddle D L 1983 Review of varieties of civil religion. *Journal of the American Academy of Religion*. **LI**(3): 198–9

Willaime J 1993 La religion civile à la française et ses métamorphoses. *Social Compass* **40**(4): 571–80

Wilson C R 1984 American heavens: Apollo and the civil religion. *Journal of Church and State* **26**(2): 209–26

Wuthnow R 1998 *The Restructuring of American Religion: Society and Faith since World War II*. Princeton University Press, Princeton, NJ, p. 244

R. K. Fenn

Civil Rights

Civil rights legally protect individuals or groups from certain forms of oppression. While civil rights are commonly associated with the 1960s movement in the USA to establish equality for people of African

descent and—more generally—with the US Bill of Rights, by the end of the twentieth century their reach and recognition was global. In modern political, academic, and public usage, civil rights embody, and provide legal support for, basic concepts of human dignity and respect for individuals in their diverse cultures and ways. Recognition and enforcement of civil rights, or some assortment of the most fundamental civil rights, is widely understood as a necessary element of freedom, democracy, and equality.

Civil rights are usually established in a written constitution, statute, or treaty. However, they have been based on unwritten constitutions and have occasionally arisen from the pronouncements of monarchs and conquerors, perhaps most famously Napoleon. These sources of civil rights protections are not self-executing: some countries have strongly worded civil rights provisions in written constitutions that are not enforced; the meaning and interpretation of civil rights provisions are usually controversial; and despite the prevalence of civil-rights aspirations and rhetoric, no country has as yet found a reliable method for systematic, consistent protection of civil rights.

1. *Origins*

The origins of some civil rights go back almost a thousand years to the earliest limitations on governments, which at least indirectly protected individuals. Civil rights as such emerged in the seventeenth and eighteenth centuries, as social development and political and philosophical thought emphasized the individual and, later, based national sovereignty and legitimacy on democratic forms of government. However, systematic legal enforcement, and even widespread aspirational recognition, of civil rights did not occur until after World War II. They quickly achieved global acceptance in the last half of the twentieth century; the laws of almost all countries, and several widely adopted international pacts, at least purport to protect some array of fundamental rights of individuals.

Approached historically and contextually, the various civil rights can be traced to practices by governments or by powerful individuals or institutions that came to be viewed as oppressive. Thus, the Fourth Amendment to the US Constitution limiting police searches was a response to the dreaded house-to-house general searches by British authorities in the colonial period. However, such origins are often obscured, because civil rights—like other basic legal precepts—are usually analyzed and clothed with the rhetoric of foundational principles and tend to take on, in law as well as in politics and fiction, an abstract, timeless quality that transcends history, experience, and human agency. This has occurred in the USA with freedom of speech, which has acquired a major place in the national origin story, although its most significant features developed in the mid-twentieth century by common forms of social change (Kairys 1982, Levy 1985).

2. *Scope*

The most common civil rights are: prohibition of discrimination based on race, ethnicity, religion, and gender; the right to personal security, including protections for persons accused or suspected of crimes; the right to vote and to participate in democratic political processes; and freedom of expression, association, and religion. Privacy, which in the US system encompasses personal security and autonomy, as well as control over personal information, is of growing concern throughout much of the world.

Personal security was the first, and remains the primary, civil right. It was a major focus of leading Anglo–American legal authorities, from William Blackstone (1769/1992) and St. George Tucker (1803) to Thomas Cooley (1874). The pathbreaking civil rights text by Thomas Emerson and David Haber (1952) considers the right to personal security first and describes its significance in the introduction: 'In a society based upon human dignity and the development of the individual personality, clearly all members are entitled to security of the person—protection from bodily harm, involuntary servitude, and the fear of physical restraint.' Although the rights of criminal defendants in the USA are commonly thought to originate in 1960s Supreme Court rulings, the US Bill of Rights has as its major subject and most numerous protections a series of limits on the government's power to punish individuals criminally.

International bodies, treaties, and courts—and many countries—prefer the term 'human rights,' which usually encompasses a broader array of rights and places different priorities on some particular rights than does US civil rights law (see *Law: Defense of Insanity*). The United Nations Charter, the *Universal Declaration of Human Rights* (1948), and many other human-rights treaties and covenants protect social, economic, and cultural rights as well as political and civil rights. These rights place affirmative obligations on governments to provide all their people with minimal nutrition, healthcare, housing, and education. Eleanor Roosevelt, criticizing the US reluctance to include economic and social rights, put it best: 'You can't talk civil rights to people who are hungry.'

Internationally, the content of human rights also often exceeds specific US civil rights protections, including, for example, prohibition of the execution of children and mentally disturbed adults and protections against domestic abuse. Other nations may also place

different relative importance on various rights, such as rejecting absolutist free-speech rights and emphasizing equality rights. While the USA has adopted most of the human rights treaties and covenants (although often with specific reservations), US courts have not consistently enforced them. The USA is also unusual among the nations of the world in its renewed embrace of capital punishment and in the retrenchment of procedural and fair trial rights to expedite executions.

In most countries, civil rights protect individuals from their governments, but protection for groups and from other individuals have also been recognized. Prohibitions of slavery around the world limit individuals, as well as governments. Equal rights for women have included limits on individually inflicted violence against women and recognition of the rights of women as a group.

3. *Universal, Absolute, and Conflicting Rights*

Civil rights lose their meaning and importance if they are not extended to all people. However, the universality of civil rights, both within a particular country and across national borders, and the common notion that extension of civil rights to those previously denied them has no affect on others, are problematic.

Typically, recognition of civil rights disturbs a status quo in which privileges and hierarchies were entrenched. If previously silenced speakers or voters gain the rights to speak and vote, the effectiveness of the voices and votes of the previously privileged is diminished. Abolishing racial discrimination in employment or college admissions diminishes the prospects of the previously privileged. Abolishing racial discrimination in public accommodations diminishes the freedom of association of those who wish to exclude racially. Recognition of the rights of people who have been oppressed involves a determination that the interests, expectations, tastes, and sometimes civil rights of others constituted an illegitimate or disproportionate privilege which came only at the expense of that oppression, and to which they should not have been, and no longer are, entitled. The failure to acknowledge the interconnectedness of rights, interests, privileges, and expectations has probably made acceptance of civil rights advances more difficult than it has to be. This has been a substantial problem in the USA, where interconnectedness is only acknowledged for affirmative action and where civil rights tend to be articulated in absolutist terms.

Absolutist conceptions of civil rights usually rest on the same thinking, memorably critiqued by Justice Oliver Wendell Holmes' famous example of shouting 'fire' in a crowded theater. Words have effects, otherwise we might not be so fond of them, and in some circumstances those effects outweigh the value of speech. Such occasions are rare, but they do exist—for example, where one person's speech is interfering with or drowning out another's or all others. Nevertheless, the US Supreme Court has protected electoral campaign finances as speech in absolutist terms that make even minimal regulation difficult or impossible (Buckley vs. Valeo 1976). The Court characterized money as speech and refused to recognize as significant or legitimate countervailing concerns the importance of fair elections to society and other individuals or the speech rights of those whose voices cannot be heard in money-driven elections.

Absolutist conceptions of this sort, though often articulated by staunch advocates of individual rights, are reminiscent of the 'Lochner era' in the USA, in which the civil right to due process was used by the Supreme Court to invalidate a series of economic reforms aimed at protecting working and poor people. The combination of an absolutist formulation and a civil right based on unrestrained free enterprise yielded one of the most notorious eras in US constitutional law. These rights that people around the world seek and rely on for protection from powerful individuals, groups, and institutions, often at considerable personal sacrifice, can become another instrument for the powerful.

There is also a fundamental sense in which all civil rights protect individuals by limiting others—either all others collectively, because civil rights prohibit certain actions by majorities and governments, or individually, in the situations in which civil rights protect against private as well as governmental actions. Thus, civil rights both bestow and limit freedom. They limit the ability of even democratically constituted governments, for example, to discriminate against a racial minority or to ban an opposition political party, by placing the individual's rights above collective power. Civil rights are, in this sense, both a limit on and a necessary feature of democracy.

These tensions permeate heated controversies about the universality of civil rights across national borders. The conventions on the rights of women referred to above have been the subject of an unusual array of national reservations, usually on the ground that they clash with cultures and religions, which also can claim civil rights protection. For example, female genital mutilation violates several of the most fundamental and generally accepted civil and human rights, but it is also central to some cultures and religions.

4. *Formulation and Enforcement: the US Model*

The US tradition of extensive judicial power, which includes judicial innovation beyond the specific language of a constitution, statute, or treaty, is gaining some acceptance internationally. This expanded acceptance is in part due to its reputed tendency to

protect civil rights, although the role of courts in civil rights matters (and more generally) is controversial in the USA. Many other countries have tended to rely on international human rights treaties or legislation. Courts lack the democratic legitimacy of legislatures, while their very distance from democratic processes could enhance the potential for protection of individual rights against even mobilized majorities. Protection of individuals and groups from majoritarian oppression is the hallmark of meaningful civil rights and provides the best measure of a particular society's civil rights record. A closer look at the history and record of a particular country—the USA, because of its historical role and wide recognition as a leader and model for civil rights—demonstrates both the potential and the difficulty of sustained, systematic protection of civil rights.

The role of courts in the civil rights victories of the 1960s and throughout US history is often exaggerated. The integration of public schools (Brown vs. Bd. of Educ. 1954) was a milestone, but most aspects of equality for African Americans in the twentieth century were established by the US Congress. In the 1960s these included banning racial discrimination in voting, housing, employment, and public accommodations. The major civil rights advances of the nineteenth century were achieved by constitutional amendment and legislation.

The history of US courts reveals them to be more often an obstacle to civil rights than a protector of them. For example, the US Supreme Court ruled in 1857 that African Americans have 'no rights which the white man is bound to respect' and are not fully citizens or human beings (Dred Scott vs. Sanford), which destroyed the Missouri Compromise on slavery and was a contributing cause of the Civil War; approved segregation even after the Civil War amendments to the Constitution (Plessy vs. Ferguson 1896); negated the 'privileges and immunities' clause of the Fourteenth Amendment (Slaughterhouse Cases 1873), which applied civil rights protections against the states; refused to protect even the most basic free speech rights until the 1930s (Davis vs. Massachusetts 1897; Kairys 1982); and approved the imprisonment of all persons of Japanese ancestry on the west coast during World War II without any charges or proof of individual guilt (Korematsu vs. US 1944). Generally, throughout US history, oppressed minorities and individuals facing repression by government or organized majorities—including in modern times the repressive McCarthyism of the 1950s and the approval of state criminalization of gay sexual activity continuing into the new millennium—have not been significantly protected by the courts.

There are only two periods in US history characterized by systematic or sustained judicial protection of civil rights, from about 1937 to 1944 and from about 1961 to 1973. Both these periods occurred in the context of mass popular support and successful legislative initiatives for civil rights. During these periods, the Supreme Court established: a system of limited government and protection of individual rights available to people of ordinary means probably unrivaled in world history, including strenuous protection of speech, association, and privacy; procedural and substantive rights of persons accused or suspected of crime; the rights of women; voting and participatory rights; and prohibitions of discrimination (Kairys 1993, 1998; W. Va. Bd. of Educ. vs. Barnette 1943).

The main judicial vehicle for protecting civil rights in these two periods was the 'strict scrutiny' standard of judicial review, by which any government action that infringed or restricted these rights was strictly scrutinized and presumptively invalid, surviving only if supported by a 'compelling' government interest that could not be furthered by 'less restrictive' means. While these decisions seldom articulated absolute rights (although absolutist rhetorical flourishes were common), in practice strict scrutiny meant near-certain invalidity.

Starting in the mid-1970s, as the national political mood and the composition of the justices changed, the Supreme Court repudiated or significantly diluted almost all of the major civil rights advances of the post-World-War-II period. This retrenchment was led in the political arena by President Ronald Reagan and a conservative movement that emphasized limiting government and espoused the sometimes extreme distrust of government evident in many periods of US history (Wills 1999).

The thrust was to empower the people by limiting the courts—in the popular phrase, 'judicial restraint'—which, paradoxically, strengthens government and weakens individual rights. Thus, this conservative judicial restraint trend reviled the 1965 case establishing a civil right to privacy (Griswold vs. Conn.), a decision that protected the individual and the people generally from a state law that banned all use of birth control. A judicially based system of protection of civil (or any other) rights requires the courts to intervene when government is infringing on protected rights, and conservative justices have not hesitated to intervene to protect civil rights they value (e.g., Buckley vs. Valeo 1976; Shaw vs. Reno 1993). Historically, the pattern in the USA has been advocacy of judicial restraint by those whose values and approaches are being rejected by the courts. Perhaps the most vehement and successful advocate of judicial restraint was liberal President Franklin Roosevelt, whose reforms were struck down by a Supreme Court intent on imposing unrestricted laissez-faire economics (Kairys 1998).

In this retrenchment, the rights to exercise non-mainstream religion and against establishment of religion were substantially undercut, including the refusal to protect a Native American religious ceremony that pre-dates Christianity (Empl. Div. vs. Smith 1990). Speech rights available to people of

ordinary means were retrenched (for example, the public areas and circumstances available for protest were substantially narrowed), while speech rights available to corporations and wealthy individuals were enhanced, including recognition of corporate free speech rights. The equality rights of minorities were undercut by adoption of virtually insurmountable burdens of proof, resulting in denials of relief for some egregious practices left over from segregation. The equality rights of the white majority have been enhanced, resulting in successful challenges to electoral reapportionment and affirmative action, even where they are good faith remedial efforts to end discrimination against African Americans and other minorities. As a result, almost all the winning parties in racial equality cases decided by the Supreme Court in the last two decades of the twentieth century were white. The US attempt to end centuries of discrimination that included slavery and forced segregation was short and mostly limited to erasing de jure discrimination; the mass of African Americans live in the same segregated poverty they endured under de jure segregation (Kairys 1993, 1994, 1996; Memphis vs. Greene 1981, Richmond vs. Croson 1989, Shaw vs. Reno, 1993).

This retreat was accomplished by a range of judicial means. In some areas, such as exercise of religion and some aspects of privacy, the Court withdrew strict scrutiny protection. In others, such as equality, establishment of religion, and some aspects of free speech, a range of new rules made it near impossible to prove a violation—such as the requirement for proving that government not only violated a civil right but did so purposely, with the specific motivation to violate the right.

None of these new rules is required by the language of the US Constitution or by rules of legal reasoning or analysis; nor is the strict scrutiny standard that was used to protect civil rights. This is characteristic of the judicial model, which in the USA yielded cyclic protection of civil rights in some (not all) periods in which support for those rights had strong, sustained, and politically mobilized support.

By the end of the twentieth century, the lead in civil rights passed from the USA to other nations and to international human rights tribunals and agreements. The Constitutional Court of South Africa led a worldwide trend to ban or limit the death penalty (State vs. Makwanyane 1995); the European Court of Human Rights, unlike the Supreme Court of the USA (Bowers vs. Hardwick 1986), protected the rights of gay men and lesbians (Dudgeon vs. United Kingdom 1981); and the rights of racial minorities and women were advanced principally by international human rights agreements. Progress in this direction has come mostly as a result of international human rights agreements and the growing sense they represent that oppression and repression, though they have hardly ceased, are no longer internationally acceptable.

5. *Prospects for Civil Rights*

Since civil rights can encompass a variety of possible limits on government to protect individuals and groups (and some limits on other individuals and affirmative obligations of governments), the details of their substance, context, and history are necessary for an understanding of their social meaning and importance. In this sense, it is hard to say that any of us is 'for civil rights,' and perhaps harder to say we are 'against civil rights' (which might mean government without limits), without specifying whose rights, which rights, and the context. For some, government is most oppressive when it limits gun ownership or attempts to overcome the effects of past discrimination; for others, it is most oppressive when it limits exercise of religion, abortion rights, or the rights of historically oppressed minorities. In the post-World-War-II period, the USA set an international standard for protection of minorities and the enforcement of a range of civil rights that protect the individual from oppression by government and powerful majorities; but US history also demonstrates the fragility of civil rights, and the use of civil rights to insulate the powerful from popular reforms aimed at protecting working and poor people.

Civil rights have played an important role in the advancement of human dignity, respect for the ways of diverse cultures and individuals, and democratic forms of government. In a relatively short time in the last half of the twentieth century, racism moved from acceptable or trivial to wrong, and a deep sense of civil rights and fair play became an international intellectual and moral standard. However, this has not stopped—by any reasonable measure—racial, ethnic, religious, and national oppression or violence, or the routine denial of the range of basic civil rights. There is a substantial and troubling divide between our words, aspirations, and moral precepts on the one hand, and our actions on the other. We may be a species whose gift of reason and empathy has outpaced its habits of survival, which still reside—as if helpless to listen to our own pleas—in deepseated (perhaps significantly genetic) tribalism. If this is so, the struggle for civil rights and human dignity will be long, indeed, but all the more compelling.

No single source or enforcement mechanism for civil rights has proved consistently superior to others. Establishment and enforcement seem most effective when accomplished by the most popular and participatory means available, and popular understanding of the importance and history of civil rights is probably the most significant factor in their continued vitality.

See also: Bill of Rights; Civil Liberties and Human Rights; Civil Rights Movement, The; Discrimination; Discrimination: Racial; Fundamental Rights and Constitutional Guarantees; Human Rights, Anthropology of; Human Rights, History of; Human Rights in Intercultural Discourse: Cultural Concerns; Human

Rights: Political Aspects; Race and the Law; Rights; Rights: Legal Aspects

Cases:
Bowers vs. Hardwick, 478 US 186 (1986).
Brown vs. Board of Education, 347 US 483 (1954).
Buckley vs. Valeo, 424 US 1 (1976).
Davis vs. Massachusetts, 167 US 43 (1897).
Dred Scot vs. Sanford, 60 US (19 How.) 393 (1857).
Dudgeon vs. United Kingdom, 45 European Ct. of Human Rights (ser.A), 4 E.H.R.R. 149 (1981).
Employment Division vs. Smith, 494 US 872 (1990).
Griswold vs. Connecticut, 381 US 479 (1965).
Korematsu vs. United States, 323 US 214 (1944).
Lochner vs. New York, 198 US 45 (1905).
Miller vs. Johnson, 115 S.Ct. 2475 (1995).
Memphis vs. Greene, 451 US 100 (1981).
Plessy vs. Ferguson, 163 US 537 (1896).
Richmond vs. Croson, 488 US 469 (1989).
Shaw vs. Reno, 509 US 630 (1993).
Slaughterhouse Cases, 83 US (16 Wall.) 36 (1873).
State vs. Makwanyane, 1995(3) South Africa Law Reports 391 (1995)
West Virginia State Board of Education vs. Barnette, 319 U.S. 624 (1943).

Bibliography

Blackstone W 1992 *Commentaries on the Laws of England*. Hein, New York
Convention of Discrimination Against Women 1992 Recommendation No. 19, Violence Against Women
Cooley T 1874 *Constitutional Limitations*. Little, Brown, Boston
Copelon R 1998 The indivisible framework of international human rights: Bringing it home. In: Kairys D (ed.) *The Politics of Law*, 3rd edn. Basic Books, New York, Chap. 9, pp. 216–39
Cover R 1975 *Justice Accused*. Yale University Press, New Haven, CT
Emerson T, Haber D 1952 *Political and Civil Rights in the United States*. Dennis, Buffalo, NY
Henkin L 1990 *The Age of Rights*. Columbia University Press, New York
Henkin L, Neuman G, Orentlicher D, Leebron D 1999 *Human Rights*. Foundation Press, New York
Janis M, Kay R 1990 *European Human Rights*. University of Connecticut Law School Foundation Press, Hartford, CT
Kairys D (ed.) 1982 *The Politics of Law*. Pantheon Books, New York
Kairys D 1993 *With Liberty and Justice for Some*. New Press, New York
Kairys D 1994 Race trilogy. *Temple Law Review* **71**: 1–12
Kairys D 1996 Unexplainable on grounds other than race. *American University Law Review* **45**: 729–49
Kairys D (ed.) 1998 *The Politics of Law*, 3rd edn. Basic Books, New York
Levy L 1985 *Emergence of a Free Press*. Oxford University Press, New York
Locke J 1771 *Two Treatises on Civil Government*. Whiston, London
Paine T 1992 *The Rights of Man*. Hackett, Indianapolis, IN
Rousseau J–J 1901 *On the Social Contract*. Tudor, New York
Tribe L 1988 *American Constitutional Law*, 2nd edn. Foundation Press, New York
Tucker St. G 1969 *Blackstone's Commentaries, with Notes of Reference to the Constitution and Laws of the Federal Government of the United States and of the Commonwealth of Virginia*. Rothman, South Hackensack, NJ
Universal Declaration of Human Rights. 1948 G.A. Res. 217A, U.N. Doc. A/810
Wills G 1999 *A Necessary Evil, A History of American Distrust of Government*. Simon and Schuster, New York
Winant H 2001 *The World is a Ghetto*. Basic Books, New York

D. Kairys

Civil Rights Movement, The

The modern American Civil Rights Movement was one of the pivotal freedom struggles of the twentieth century. This article will discuss the historic oppression of African Americans and how the Civil Rights Movement fought to liberate that population. It will also analyze the origins of that movement as well as its national and international achievements.

1. Black Oppression

Even as late as the middle of the twentieth century, millions of Black citizens in the USA were socially oppressed, economically exploited, and disenfranchised politically. These conditions of subjugation endured in a nation viewed as the world's leading democracy. This view of the USA was projected nationally and internationally by White US leaders. The US image of democracy contrasted sharply with the actual treatment of African Americans. Unlike European immigrants, Black Africans were forcibly transported to America as slaves. As captives of the institution of slavery for over two centuries, Blacks supplied the free slave labor that assisted the USA in becoming an economic superpower. Slavery denied Blacks the basic rights that constitute the foundation of a democracy. Indeed, during slavery, African Americans were officially defined as chattel, not human beings.

The American Civil War (1861–5) was the force that overthrew the slave regime. With the triumph of Union forces in 1865, Black equality became a real possibility. For a brief period following the Civil War it appeared that the former slaves would be granted their democratic rights. During the Reconstruction period, Blacks gained expanded citizenship rights including freedom of movement, restricted male access to the franchise, and access to employment. This was a promising beginning but it would not endure for long.

By the turn of the twentieth century the Reconstruction period had come to an end. In the early

1900s, a formal system of racial segregation known as Jim Crow was firmly established. The new form of racial oppression required that Blacks and Whites be segregated on the basis of race. Thus, legally the two races were not allowed to attend the same movie theaters, drink from the same water fountain, sit on the same side of a courtroom, or be sworn in with the same Bible. Blacks and whites were prohibited from occupying the same space on a public bus or train. In short, Blacks were denied equal access to public accommodations. Moreover, Blacks were excluded from the political process. Toward the close of the nineteenth century the emerging Jim Crow regime received backing from the highest court of the nation. Thus an 1896 Supreme Court Ruling, *Plessy vs. Ferguson*, declared that racial segregation was constitutional. It ruled that it was constitutionally legal for the two races to use separate-but-equal facilities. With this ruling the Jim Crow regime became national in scope although it was more rigorously enforced in the South.

By the middle of the twentieth century the majority of Blacks were disenfranchised. In the South, Blacks held no significant political offices and they were constant victims of terror and violence including lynchings. In the labor market Blacks were restricted to low-paying undesirable jobs. As a result, economic exploitation of Blacks was widespread. Beyond material subjugation, African Americans experienced daily personal humiliation because racial segregation marginalized them as a people and labeled them as an inferior race. Human dignity was stripped from African Americans: simple titles of respect such as 'Mr.' or 'Mrs.' were withheld, and even White youngsters held symbolic authority over all Blacks, however elderly or eminent.

2. *The Movement's Origin*

African Americans consistently attacked the Jim Crow regime. Black protests began during slavery and remained evident during the Jim Crow period. At times, resistance was collective and public, while at other times it remained covert and limited in scope. By the 1950s African Americans had produced a long, rich tradition of social protest. The modern Civil Rights Movement drew upon this tradition and embedded itself deeply within the historic struggle for Black liberation.

The Civil Rights Movement took deep root in the south in the mid-1950s. Thus, this movement emerged where Black oppression was most intense and where racial segregation was firmly entrenched. Given the scope of oppression and the power of the White opposition, the birth of a powerful resistance movement was unanticipated, especially by Whites who thought that segregation had become part of the natural order.

The movement took off during this period for several reasons. First, by the 1950s large numbers of African Americans had migrated to southern cities where they developed tightly knit urban communities and dense, effective communication networks. These resources made it possible for mass mobilization to occur. Second, by the 1950s the National Association for the Advancement of Colored People (NAACP) had won successful legal rulings against the Jim Crow system, especially in the 1954 Supreme Court decision, *Brown vs. Board of Education*, which reversed the 1896 *Plessy vs. Ferguson* decision. The Brown ruling stated that separate schooling based on race was unconstitutional. The implications of this ruling went far beyond this case for it delegitimized the entire system of racial segregation and encouraged struggles to implement the court orders.

Additionally, changing international relations were important to the birth of the movement. By the middle of the twentieth century African and Third World countries were gaining independence through anticolonial struggles. African Americans identified with those struggles, which intensified their own thirst for freedom. Moreover, the context of decolonization and Cold War rivalry rendered the US Federal Government susceptible to pressure from an oppressed Black population because the USA sought to persuade these new Third World nations to model themselves after US democracy, not the Soviet alternative. However, US racial oppression stood as a barrier to harmonious relations between the USA and new Third World nations. The proliferation of television and communication satellites made it possible for Black oppression to become visible worldwide. America's participation in the two world wars and the Korean War also rendered it vulnerable to Black protest because in those wars the USA championed egalitarian values. In this context, Black soldiers were radicalized while fighting for democracy on distant shores. Thus, the federal government came under increased and sustained international and domestic pressure to support efforts to overthrow institutionalized racial segregation.

2.1 *Strategies, Tactics, and Goals*

These factors created the fertile soil from which the Civil Rights Movement emerged. By employing the strategy of mass, nonviolent, direct action, this movement mobilized widespread social protest. For such a strategy to succeed, White communities, businesses, and institutions had to be disrupted and prevented from doing business as usual. To this end, civil rights leaders and organizers mobilized thousands of African Americans to confront the Jim Crow regime through social protest.

The Civil Rights Movement succeeded in mobilizing massive nonviolent social protest. Innovative tactics

included economic boycotts (beginning with the year-long boycott of a bus company in Montgomery, Alabama, sparked by the arrest of Rosa Parks, in December 1955 and led by Martin Luther King Jr.); sit-in demonstrations intensified in February 1960 by Black college students at a lunch counter in Greensboro, North Carolina; dramatic confrontations in the streets of Birmingham, Alabama in 1963; and mass marches (including a massive mobilization of Whites and Blacks in the August 1963 March on Washington, which culminated in King's 'I have a dream' speech, and protest marches led by King that met with police violence in Selma, Alabama, in January 1965).

The goal of these protests was to overthrow racial segregation and empower African Americans by seizing the franchise. Southern officials utilized their institutional power and the resistance of the larger White community in an intense, and often violent, effort to defeat the movement and to maintain legally enforced racial segregation. Movement participants—many of them women, children, and college students—were often beaten and brutalized by southern law enforcement officials, and thousands were arrested and jailed for their protest activities. Some leaders and participants—such as Medgar Evers, of the Mississippi NAACP in 1963, and three civil rights workers in Mississippi in 1964—were murdered.

Nevertheless, the widespread and highly visible confrontations in the streets, which contrasted the brutality and the inhumanity of the White segregationists with the dignity and resolve of Black protestors, made the cause of Black equality the major issue in the USA for over a decade during the 1950s and 1960s. The nation and its leaders were forced to decide publicly whether to grant African Americans their citizenship rights or to side with White segregationists who advocated racial superiority and the undemocratic subjugation of African Americans.

3. *National Achievements*

The movement could not be dismissed. Eloquent leaders and their massive followings sustained the pressure on local elites and the federal government. Countless heroic figures inspired and organized a massive following. Among them were Rosa Parks, a dignified older woman and local NAACP activist, who sparked the Montgomery bus boycott when she defied an order to move to the back; Martin Luther King, Jr., who emerged from the Montgomery bus boycott and the Southern Christian Leadership Conference (SCLC) to assume a position of preeminent moral leadership and national influence; James Forman, executive secretary of the more militant Student Nonviolent Coordinating Committee (SNCC), who was to challenge SCLC's and King's strategy; SNCC's leader, Stokely Carmichael (now Kwame Toure), who introduced the slogan 'Black power'; James Farmer and Floyd McKissick who led the Congress of Racial Inequality (CORE); and Fannie Lou Hamer, who went to a voter registration meeting run by SNCC, was arrested at the courthouse in Indianola, Mississippi, when she tried to register to vote, and then was viciously beaten in prison. Hamer personified the thousands of women who played important roles in organizing and leading the movement.

These leading figures and thousands of movement participants articulated Black suffering and the democratic aspirations of African Americans of every generation and circumstance. In addition, thousands of Whites (students, ministers, lawyers, and other civil rights workers) were inspired to join the Movement. They participated in lunch-counter sit-ins, mass demonstrations, and campaigns such as the 1964 Mississippi Freedom Summer Project, a campaign that involved hundreds of volunteers in voter registration drives and the creation of 'freedom schools.' Some, like Andrew Goodman and Michael Schwerner, who were involved in the Freedom Summer campaign, and Viola Liuzzo, a Michigan homemaker shot by Klansmen after a rally in support of the march from Selma to Montgomery, lost their lives in the Movement and, in turn, helped inspire others.

Differences regarding ideology, leadership style, and goals emerged within the movement as the SCLC, NAACP, SNCC, CORE, and other organizations reached different judgments about the value of nonviolent action, racial integration and separatism, the role of Whites in the movement, and the influence of Black nationalism. Despite these controversies, the moral challenge and the widespread social disruption caused by the economic boycotts, the marches, the sit-ins, and other forms of nonviolent direct action, coupled with the international pressure, created an impasse in the nation that had to be resolved.

As a result, the Civil Rights Movement achieved important legislative victories in Congress. The landmark Civil Rights Act of 1964 outlawed discrimination in public accommodations on the basis of race, color, religion, or national origin and it identified the legal measures to be used to achieve racial integration. Moreover, it barred discrimination in employment practices on grounds of race, color, religion, national origin, or sex. The passage of the 1965 Voting Rights Act was another major achievement, for it suspended the use of literacy tests, authorized the attorney general to challenge the constitutionality of poll taxes, and introduced procedures that provided for the appointment of examiners to ensure that all restrictions on Black voter registration be ended. In short, the Voting Rights Act enfranchised the southern Black population, making it possible for a historic Black elected political class to emerge.

3.1 *International Achievements*

The significance of the Civil Rights Movement extends far beyond its historic overthrow of the Jim Crow

regime. This Movement has affected US politics in fundamental ways. It demonstrated to the oppressed Black community how such protest could be successful, and it made social protest respectable. The Civil Rights Movement also proved that social protest is capable of generating significant change. Hence, the movement broadened the scope of US politics and inspired diverse movements for citizenship rights and social justice in the USA and abroad. Before the Movement, many groups in the USA—women, Hispanics, Native Americans, farm workers, the physically disabled, gays and lesbians, etc.—were oppressed but unaware of how to resist or galvanize support. The Civil Rights Movement provided a model of successful social protest and produced a host of new tactics and social change organizations. Moreover, this Movement had an influence on freedom struggles around the world. Participants in freedom struggles in Africa, Eastern Europe, the Middle East, Latin America, and China have made it clear that they were deeply influenced by the US Civil Rights Movement.

4. *Continuing Challenges*

For all its success and influences, however, the Civil Rights Movement did not solve all of America's racial problems. At the start of the twenty-first century, African Americans and many other non-White groups are still at the bottom of the social and economic order. These current conditions are exacerbated by a social climate in which the disadvantaged are often blamed for their own predicament. Thanks to the Civil Rights Movement, a relatively large Black middle class has emerged. Yet, over a third of the Black community remains trapped in poverty. Black Americans are disproportionately the victims of police brutality and housed in the rapidly growing prison industry. Thus, during the infancy of the new millennium, some Blacks are the recipients of middle-class comforts while millions of others are victims of poverty and degradation. Poverty and inequality are also widespread outside the Black community. It may be that protest remains the only viable means to achieve greater empowerment. If this is the case, the Civil Rights Movement has left a rich legacy to inspire and inform future struggles. The sparking of a renewed interest in the academic study of social protest is an unnoticed legacy of the Civil Rights Movement. Through such studies it is possible that new mechanisms of social change will come to light.

See also: Civil Liberties and Human Rights; Civil Rights; Conflict Sociology; Integration: Social; Minorities; Nonviolence: Protective Factors; Race and Gender Intersections; Race Identity; Racial Relations; Racism, History of; Racism, Sociology of; Slavery as Social Institution; Slaves/Slavery, History of; Social Movements: Resource Mobilization Theory; Social Movements, Sociology of

Bibliography

Branch T 1988 *Parting the Waters: America in The King Years 1954–63*. Simon and Schuster, New York

Carson C 1981 *In Struggle: SNCC and the Black Awakening of the 1960's*. Cambridge University Press, Cambridge, UK

Garrow D 1986 *Bearing the Cross: Martin Luther King, Jr., and the Southern Christian Leadership Conference*. William Morrow, New York

Klinker P A, Smith R M 1999 *The Unsteady March: The Rise and Decline of Racial Equality in America*. University of Chicago Press, Chicago

Layton A S 2000 *International Politics and Civil Rights Policies in the United States 1941–1960*. Cambridge University Press, Cambridge, UK

McAdam D 1988 *Freedom Summer*. Oxford University Press, New York

Morris A D 1984 *Origins of the Civil Rights Movement: Black Communities Organizing for Change*. Free Press, New York

Morris A D 1999 A retrospective on the Civil Rights Movement: Political and intellectual landmarks. *Annual Review of Sociology*

Robnett B 1997 *How Long? How Long?* Oxford University Press, New York

A. Morris

Civil Service

The civil service is the generic name given in English to the administrative apparatus of the state. The term was first introduced in the British administration in India and then in the UK (1854), and has become almost universally synonymous with civilian (i.e., nonmilitary and nonjudicial) administrators employed by central governments. The civil service is also referred to by different terms (e.g., 'public bureaucracy'), and it differs from other bureaucracies by virtue of its public missions, however defined. It has a collective justification, namely, doing something otherwise unattainable for a certain pubic end, and therefore, civil servants have been entrusted with the task of being the guardians, or at least the interpreters, of the common good. The modern civil service is above all a 'state institution' and this article will focus on the changing relationship between the state and its specialized bureaucratic institution—the civil service.

1. *History*

There were bureaucratic administrators long before they were known as 'civil servants.' Historical antecedents can be found mainly in centralized States and empires where public bureaucracies were developed to serve the rulers, or the dynasty. Despite their subservience to the emperors and kings, however, these bureaucracies became 'professional,' that is, they manifested some conception of themselves as servants

of the 'state,' the polity, or even the community (Eisenstadt 1965). Such embryonic civil service institutions existed in different forms first in the Egypt and China empires, and later, for instance, in the Byzantine, Sassanid, Abbasid, and Ottoman empires. Sculptures of the ancient Egyptian scribes date back to the Old Kingdom (ca. 2600 BC). They belonged to a body of officials who served the Kingdom, as did their Assyrian and Babylonian counterparts. Even more professional were the ancient Chinese functionaries employed in the rather developed court machinery. These officials—whose rank was conspicuously exhibited by the shape and color of the buttons worn on their caps—became much later known in the West as 'mandarins.' The term is often used pejoratively to indicate the aloofness of civil servants, the official jargon and the sheltered life of their long careers.

Broadly speaking, the history of public administration testifies that there has always been a need to link state (or ruler) goals, with the official expertise of a centralized bureaucracy. These bureaucratic entities were engaged in three main types of activities: tech nical services such as land registration or water allocation; social and political regulation such as administrating justice or collecting taxes; and above all—managing war-related affairs (Finer 1997, Vol. 1, pp. 59–72).

The European patterns of the civil service have emerged gradually from the Greek and especially the Roman law and administrative traditions, of, for example, public construction, food supply regulation, and population census. They were later influenced by the medieval feudal system of financial and judicial administration. Yet, the most strenuous collective effort has always been related to mobilizing money, soldiers, and equipment for the military. Accordingly, throughout the history of most countries, organizational practices, public and private, have been influenced directly by the most readily available example—the military model. This model was further developed to encompass the colonial expansion of some European countries, which required new organizational expertise, both in content and in scope. The influence of the military is most readily seen through such administrative concepts as 'chain of command,' or 'line and staff,' so much so that recent attempts to debureaucratize the civil service are essentially an attempt to break away from the dominance of the military bureaucratic model.

2. *The Modern Civil Service*

The emergence of the modern civil service is connected directly to the crystallization of the European style state, first in France and Prussia and later in all the 200 or so states that populate the globe in the early twenty-first century. As an institution the civil service dates back to the middle seventeenth century with the inauguration of competence entrance examinations in Prussia under Friedrich Wilhelm I, analogous to the old Chinese practice. Equally important, the German Cameralists were the first modern scholars–practitioners to study public administration and to offer a set of principles for the 'management of the state' (Gross 1964, Vol. I, p. 108). The next stages in the development of the civil service were diffused, but a few milestones can be presented, all of them related to the attempt to create an institutional instrument with acquired experience in running state affairs. The dissimilar features should also be noted as each civil service is also a product of its particular environment, and the human composition of each of them reflects recurring attempts of social groups to enter the ruling elites of their respective states.

2.1 *The Napoleonic Reforms Early Nineteenth Century*

The Napoleonic Reforms converted the previous French royal service into a State public service. France, and before that Prussia, created the continental model of a professional, tenured and in-house trained civil servant. More than in other countries, continental civil servants embody the state, and the continuity of state affairs, and operate within a specific legal framework of administrative law (Crozier 1964, Suleiman 1974).

2.2 *The Northcote–Trevelyan Report 1854*

This report introduced competitive examinations into the British civil service and uniform methods of recruitment and promotion across departments. It eventually succeeded in eliminating patronage. Making it a professional life career attracted the educated to enter the senior administrative class, which was given the task of shaping public policy and advising politically elected ministers. Below them were the executive and clerical classes, which were given managerial and routine responsibilities. The UK was the first to establish a civil service commission to oversee the entire operation of the civil service. Politically neutral and very secretive, the British senior home civil service acquired a reputation of integrity and impartiality, and of being guardians of crown affairs, as distinct from the governments of the day.

2.3 *The Pendelton Act 1883*

The Pendelton Act in the USA marked the beginning of federal civil service reform, aimed at abolishing the previous 'spoils system' in the central government.

Patronage in the federal government was closely related to electoral turnovers in the Presidency and Congress. Unlike the UK, the American system has never achieved the same degree of uniformity, political neutrality, or continuity. In the USA, however, the civil service has been much more open, with recruitment to the senior positions, less dependent on social class and on elite education systems. Consequently, the civil service has never had the aura of 'officialism' as in Europe, and has not acquired the prestige and the consequent status of being a permanent agent of the federal state.

2.4 Nondemocratic Regimes

In nondemocratic regimes the position of the civil service institution is related directly to the scope of the state's monopoly of the different spheres of life. The stronger the state, the more dominant is the public bureaucracy—operating sometimes from the ruler's palace, the military barracks, or the party's headquarters. In such regimes, the civil service hardly exists as a differentiated professional institution. In the Soviet Union, for example, the lines between state and party bureaucracies, and indeed between political and administrative decisions, were practically nonexistent. In this respect, the most troubling phenomenon appeared with the total submission of the highly developed German civil service (including the military and the judicial) to Nazi leaders. This raises the question: Can there be safeguards against turning professional competence—and loyalty—into a sharp tool in the hands of evil political masters?

2.5 New States

In the New States the civil service is a mixture of colonial heritage and unique local elements. In India, and in some other former British colonies, the civil service is a strong institution, and it helps to keep the country going, despite recurring crises and accusations that it impedes development. In many other new countries, particularly where the military took over, the civil service hardly functions. It cannot maintain law and order, let alone carry the task of providing social services. The weaker the state grows, the more helpless is the civil service institution, and the weaker the civil service, the less the ability to sustain stability and encourage positive changes.

3. The Rational Model

Max Weber (1864–1920), one of the most prominent students of bureaucracy, delineated three 'ideal-types' of legitimate authority—the legal-rational, the traditional, and the charismatic, and suggested that the first one requires the existence of a 'rational' administrative staff. Strongly influenced by the Prussian example, Weber drew a list of the characteristics of a modern bureaucracy and the conditions that contributed to its emergence and growth, particularly in capitalist market economies (Weber 1947). His ideas reflected, and to some extent strengthened, the heavy reliance—both in continental Europe and in the USA—of civil service institutions upon administrative law: legal codes, rules, regulations, and precedents. To this very day the notable exception is the UK which (unlike Canada, Australia and its other former dominions) does not have a civil service act of Parliament and does not require legislation to change its public administration system.

The rational ('Weberian') model of the civil service as a permanent institution also includes the following common features, which are assumed to exist in the public bureaucracies of the developed states: the centralization of authority; formal modes of operation; a central commission or agency in charge of civil service affairs; established procedures for recruitment, tenure, promotion, compensation, job evaluation, discipline, and the right to associate and strike; and arrangements for securing political neutrality. In addition, civil servants must adhere to a special code of conduct which specifies their obligations, and constitutes the ethos (or perhaps the myth) of impersonality, neutrality, and guardianship of state secrets.

The important point about the rational model of the civil service is that there is a connection between the development of the modern state and this new form of bureaucratic organization. Indeed, in the twentieth century the civil service and the entire public sector expanded rapidly in response to growing roles of the state, particularly with the growth of the welfare state.

Governments and their civil service apparatus have been asked to provide more services and to find answers to new problems such as environmental degradation. But at the same time, the 'administrative' or 'bureaucratic' state has also come under attack, in terms of the size, structure, and functions of the civil service. New post-Weberian questions have been floated: to whom should the civil service be responsive and accountable: to the state, the law, the government of the day, the multicultured conglomeration of sectors and groups, or perhaps to the individual customers of its services?

4. What do Civil Servants Do?

Civil servants are in charge of a rich menu of government activities ranging from artificially 'seeding clouds' in the sky to induce rainfall, to operating schools, or administrating programs for the increasing number of aged people. In between, civil servants, as in ancient times, continue to collect taxes and perform state functions that cannot be trusted to, or done by,

social (voluntary) or private market institutions. These various activities can be grouped under the following headings: shaping and implementing public policy decisions; providing services to individuals, groups and organizations; and administrating regulatory schemes in areas such as aviation, drugs, and election financing.

The main structural entities for carrying out these activities could be divided roughly into three (a) the regular civil service departments and agencies—responsible for the generic governmental functions; (b) the special statutory agencies, authorities, commissions, etc.—responsible for specific tasks removed from the regular civil service, (e.g., the USA Securities and Exchange Commission); and (c) government corporations entrusted with running utilities and other commercial enterprises regarded as natural monopolies or in some other way related to national interests (e.g., postal services, public broadcasting, electricity companies, and regional development projects). The combinations of the various types of activities and the different structures reflect the myriad and ever-changing scope of civil service responsibilities.

It is impossible to prescribe the right number of civil servants per capita, given the differences in their functions, and in the number of intra-state governmental levels in different states. In economically developed states, the tendency is to cut down the scope of civil service activities (and consequently its size too) through privatization, outsourcing, etc. In new states the pressure still exists on the civil service to do much more for their hard-pressed societies.

5. *Studying the Civil Service*

Woodrow Wilson (1887) was among the pioneers who attempted to create 'a practical science of administration,' aimed at improving not only the personnel, but also the organization and methods of government offices. He saw the field of administration as a field of business which should be outside the sphere of politics. Hence his famous dictum: separate policy-making from administrative execution. But there were other scholarly currents as well. From the 'scientific management' movement in the USA (ca. 1900s) came the notion of 'efficiency,' aimed at increasing workers' productivity through detailed time and motion studies, standardization of tools, and careful attention to training. From the 'human relations' school (ca. 1930s) came the message that in all organizations, individual and group motivations are the most important factors. A little later, and more directly applied to public administration, attempts were made to develop prescriptive principles. For instance, establish one top executive; fit people to organizational structures; ensure unity of command; use specialized staff; maintain homogeneity in organizational subdivisions; delegate authority; match responsibility with authority; and limit span of control (Gulick and Urwick 1937).

A change came with Herbert Simon's (1946) attack on these principles which he regarded as being no more than 'proverbs of administration.' Simon's concept of 'bounded rationality' that governs the life of all organizations, was an important step away from prescriptions towards observations, and from practical advice on how to run organizations, to laying the foundation of public administration theory and of a comparative perspective. First came the formal studies of the civil service as an institution within the executive branch, or in the context of administrative law. Later, mainly as a result of the impact of studies in political science, the context as well as the horizon were expanded, away from the previous formalism. For example: power, group theory, and communication were introduced into the vocabulary of public administration research. More recently, the related area of public policy has emerged with new emphasis on economic models and game theory, as well as on the redefinition of values, public goods, and the role of politics in decision-making (Majone 1989). New foci were introduced such as implementation, social regulation, and comparative policy analysis in areas such as welfare, health, and environment (Goodin and Klindermann 1996, pp. 551–641).

The most challenging and inadequently researched new area has remained the role of the civil service institution in democracies. This is a practical issue as well: how to reform the civil service within a new democratic framework in the era of state weakening (Silberman 1993).

6. *Blurring the Boundaries of the Civil Service*

The modern version of the state is changing rapidly, and if there is going to be a 'skeleton state,' there will be a skeleton civil service as well. This process has been gradually occurring both because of internal changes such as the weakening of pubic trust in political institutions, and under the impact of globalization. In the early 2000s there are states still struggling to build a British-type civil service (which does not exist in the UK anymore); other states in which few changes have taken place in their public bureaucracies; and economically-rich states in which the role of the civil service is visibly contracting. There are also new supra-state bureaucracies, not only in international organizations such as the European Union, but also in inter-state agencies and in joint public management of regional projects. These developments have shattered many of the old features of the classical Weberian-type civil service.

Examples:

(a) Civil service monopoly on official information is no longer feasible and government secrecy has yielded

to freedom of information acts enacted in most democratic states (Galnoor 1989).

(b) With modern information technology and multichannel communication, civil servants' anonymity (and perhaps discretion too) has been gradually disappearing.

(c) The previous civil service—interest groups network, has been replaced by a much more intricate and uncontrollable system of organization in the mushrooming 'third sector' (nongovernmental, nonprofit) of public interest groups, NGO's, and philanthropic funds.

(d) The civil service is under great pressure to be more 'representative' in its membership, or at least to assist in achieving equal opportunity through 'affirmative action' and other measures.

(e) Direct accountability of the civil service to the public is demolishing the last walls of insulation. One example is the spread of the Swedish Ombudsman institution. Another is the authority given to special investigation commissions to probe the decision-making discretion of public officials.

These developments were followed by important changes in some of the traditional features of the civil service. For example, in many states the civil service is no longer a life career and new opportunities for entrance to (and exit from) senior positions have been introduced. The emphasis on management skills has downgraded the policy-advice roles of civil servants. The results-oriented business approach has diverted attention to the quality of service aspect, including the idea that civil servants should be rewarded according to their standing in customers' satisfaction surveys.

In states where New Public Management (NPM) reforms were was introduced, a new post-public bureaucracy model is slowly emerging. It requires questioning many assumptions about the civil service—both in scholarship and in practice (Peters and Wright 1996). The most important change in this respect is in the blurring of the old boundaries between the public and private sectors, and between politics and administration.

The first blurred boundary was pointed out long ago by Merriam (1944): the civil service has constituencies, and in many countries corporations exert a great deal of influence on public policy. However, with increased regulation, the private sector became even more involved, and hence the phenomenon of excessive clientelism (Nadel and Rourke 1975). Moreover, as part of the process of 'reshaping the state,' the size, structure, and functions of the civil service are under attack. Independent public and private entities and agencies are performing what used to be civil service tasks such as postal services. Other functions have been 'privatized,' 'marketed,' or 'outsourced' to the private sector: running prisons, computer services, or even recruitment to the civil service itself.

As for the politics–administration dichotomy, for many years it disappeared from public administration textbooks. Everybody knew that civil servants exercise state power both formally (law application) and informally (policy shaping). However, starting in the 1980s the distinction returned as part of NPM. The new textbooks dealing with the civil service are now called 'introduction to public sector management,' and the civil service is conceived to be market-oriented, customer-driven, deregulated, decentralized, etc. Thus the old Wilsonian dichotomy has returned through the back door: an institutional separation within the civil service between the 'departments' in charge of policy and the 'performance-based organizations' or 'executive agencies' responsible for delivering services to customers.

See also: Administration in Organizations; Administrative Law; Bounded Rationality; Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Decision-making Systems: Personal and Collective; Delegation of Power: Agency Theory; Executive Branch, Government; Governments; Public Administration: Organizational Aspects; Public Administration, Politics of; Public Bureaucracies; Public Management, New; Weber, Max (1864–1920)

Bibliography

Aberbach J D, Putnam R D, Rockman B A 1981 *Bureaucrats and Politicians in Western Democracies*. Harvard University Press, Cambridge, MA

Barnard C 1938. *The Functions of the Executive*. Harvard University Press, Cambridge, MA

Crozier M 1964 *The Bureaucratic Phenomenon*. University of Chicago Press, Chicago

Dror Y 1971 *Design for Policy Sciences*. Elsevier, New York

Eisenstadt S N 1965 *Essays in Comparative Institution*. Wiley, New York

Finer S E 1997 *The History of Government From the Earliest Times*. Oxford University Press, Oxford, UK

Heady F 1979 *Public Administration: A Comparative Perspective*, 2nd edn. Marcel Dekker, New York

Hood C 1999 *Regulation Inside Government*. Oxford University Press, Oxford, UK

Galnoor I 1989 Government Secrecy. In: *International Encyclopedia of Communications*. Oxford University Press, Oxford, UK, pp. 34–7

Goodin R E, Klingemann H-D (eds.) 1996 Public policy and administration. In *A New Handbook of Political Science*. Oxford University Press, Oxford, UK, pp. 551–641

Gross B M 1964 *The Managing of Organizations*. The Free Press, New York

Gulick L, Urwick L (eds.) 1937 *Papers on the Science of Administration*. Institute of Public Administration, New York

Majone G 1989 *Evidence, Argument and Persuasion in the Policy Process*. Yale University Press, New Haven, CT

Merriam C E 1944 *Public and Private Government*. Yale University Press, New Haven, CT

Nadel M V, Rourke F E 1975 Bureaucracies. In: Greenstein F I, Polsby N W (eds.) *Handbook of Political Science*. Addison-Wesley, Reading, MA, **5**: 373–440

O'Donnell G 1973 *Modernization and Bureaucratic Authoritarianism: Studies in South American Politics*. University of California Press, Berkeley, CA
Peters B G, Wright V 1996 Public policy and administration: Old and new. In: Goodin R E, Klingemann H-D (eds.) *A New Handbook of Political Science*. Oxford University Press, Oxford, UK, pp. 628–41
Pressman J L, Wildavsky A 1984 *Implementation*, 3rd edn. University of California Press, Berkeley, CA
Silberman B 1993 *Cages of Reason: The Rise of the Rational State in France, Japan, the United States and Great Britain*. University of Chicago Press, Chicago
Simon H A 1946 The proverbs of administration. *Public Administration Review* **6**: 53–67
Simon H A 1957 *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*, 2nd edn. Macmillan, New York
Suleiman E N 1974 *Politics, Power and Bureaucracy in France*. Princeton University Press, Princeton, NJ
Weber M 1947 *The Theory of Social and Economic Organization*. Oxford University Press, New York
Wildavsky A 1987 *Speaking Truth to Power*. Transaction, New Brunswick, NJ
Wilson J O 1989 *Bureaucracy*. Basic Books, New York
Wilson W 1887 The study of administration, *Political Science Quarterly* **2**: 197–222

I. Galnoor

Civil Society, Concept and History of

Historically, the idea of civil society takes two very different forms. In the first, civil society is 'political community' (*societas civilis* or *koinonia politike*) encompassing a state undifferentiated from society (Ellis 2000). Here, civil society is coterminous with the state: that is power relations ordered through law and institutions with the objective of ensuring social harmony. In the second, civil society is a self-regulating, self-governing body outside and often in opposition to the state, represented both as the nexus of societal associations expected to generate civility, social cohesion and morality, and as the site of reciprocal economic relations among individuals engaged in market exchange activity.

1. Koinonia Politike *and its Historical Modulations*

In Greek and Roman political thought the notion of civil society as political community was not limited to the legal category of citizenship. Central to Aristotle's (384–22 BC) conception of political community was the recognition that people lived in different social spheres and their status varied in terms of property, skills, and abilities. The art of politics was the use of laws and institutions to organize activities within these different spheres with the objective of attaining a harmonious or 'just' social environment. Aristotle regarded the citizen as one who shared in the administration of justice and held office to this end (Aristotle 1965). In this sense, citizenship in the Athenian city-state was as much a moral as a legal category (Ehrenberg 1999). The household was not excluded from Aristotle's moral scheme. Household subsistence production was 'natural,' while production for commercial exchange and profit was 'unnatural' and subversive of the moral order. The exercise of justice presupposed constraints on commercial activity.

For Cicero (106–43 BC) and the Roman lawyers, civil society (*societas civilis*) was the equivalent of *res publica* (commonwealth), or 'an assemblage of people in large numbers associated in an agreement with respect to justice and a partnership for the common good' (Cicero 1988). Cicero viewed justice as rooted in man's natural 'social spirit' which was informed by reason and induced individuals to forego a measure of self-interest in the interest of common good (Ehrenberg 1999). Cicero was writing when Rome had ceased to be a harmonious 'commonwealth' run by its citizens and their subsistence producing households, but had become an imperial city in which landowners, financiers, and merchants coexisted uneasily with a vast and hungry population of peasants and slaves. Cicero saw the realization of justice as the 'even balancing of rights, duties and functions' within the state (Cicero 1988). *Societas civilis* thus represented groups and individuals united by laws and institutions, which organized their activities and sought to achieve a flexible equilibrium among them.

Until the end of the eighteenth century, the various societies of western Eurasia generally conformed to the Aristotelian model of a politically constituted moral community. In western Europe, however, the idea of *koinonia politike* virtually disappeared during the Middle Ages when extreme political fragmentation led to the Church assuming a central role in government and social life.

On the other hand, in the Byzantine Empire, the idea of a politically constituted community persisted and the Church was subordinated to the political and moral authority of the emperor. The Ottoman Empire succeeded to this mode of political discourse, but added ideas of bureaucratic and hierarchical organization derived from the political traditions of the agrarian empires farther east, notably Persia. This synthesis, which predated the Ottomans, was formalized in the ninth and tenth centuries by Islamic philosophers well-versed in the works of Plato and Aristotle, as well as by the bureaucratic elites of the Islamic states who were educated in the Persian tradition. It was resisted by some sections of the Islamic establishment on the grounds of the equality of all before God. The ensuing power struggles did not, however, result in an autonomous 'Islamic Church,' but rather the religious establishment was subsumed within the political hierarchy.

In the Ottoman Empire, *koinonia politike* was epitomized by the figure of the just ruler, whose ability

to establish 'good order,' or *shariah* in the Islamic sense, predicated upon an absence of social strife, constituted justice. Exercise of justice involved laws and institutions accommodating different interest groups in order to sustain the conditions for household subsistence production and to achieve equitable distribution of resources among power holders. Thus laws and institutions represented negotiated settlements between the ruler and different interest groups.

In western Europe, the idea of community created by God and comprising all mankind (i.e., all Christians) was a challenge to the notion of differentiated community constituted through politics to achieve social harmony and justice. St Augustine's *City of God* (*De Civitate Dei*, 413–26), where men were equal before God and united in the universal church, stood in contrast to the Greek and Roman *polis* of differentiated spheres of life. The Church and the temporal rulers responsible for its government derived their right to rule from God and stood outside the community.

However, struggles between the Church and temporal rulers, coinciding with the rise of monarchies, resulted in the division of social activities into spiritual and temporal spheres governed, respectively, by the Church and the kings. The concept of *koinonia politike* found a new foothold in the institution of monarchy. Between the fourteenth and seventeenth centuries, European monarchs, endowed with divine right and with absolute powers, nevertheless engaged in constant negotiations with different interest groups. Political society rested on the exchange of entitlements to different groups and individuals in return for their obedience and/or services they rendered.

2. Koinonia Politike *Transformed*

In the sixteenth and the seventeenth centuries, following a period of religiously-driven civil wars, the idea of a Christian *koinonia politike* was replaced by the concept of the overriding sovereignty of the secular ruler who subordinated religious claims and restored peace. This was the Hobbesian moment in which the modern state rose to define, limit, and enforce moral alternatives to God's order (Maier 1987). Competition among European monarchies, warfare, civil war, and peasant uprisings all contributed to centralization of political power. By the end of the seventeenth century, this was reflected in the institution of the sovereign monarch as a site of bureaucratic administration and regulation. Commercial expansion in the seventeenth and eighteenth centuries meant that economic activity became a primary target of state regulation. The administrative, regulatory state was represented in discourses on economic administration that combined cameralism, mercantilism, political arithmetic, and bullionism.

Thus political or civil society was no longer the politically constituted community characterized by a diffusion of political power. Instead, political society referred to the sphere of absolute sovereignty. Social harmony, which remained the legitimizing objective of rule, was the product of bureaucratic regulations and no longer the outcome of reciprocal exchanges or negotiated settlements between the ruler and the different sectors of society.

Attempts to formalize the sovereign state also had the effect of pointing to the existence of a sphere outside the political domain. In the latter part of the eighteenth century, a new conception of civil society as the site of self-regulation developed, referring to voluntary associations freed from the corporate grip of the Church and urban institutions and to the sphere of economic activity . This concept of a self-regulatory, self-governing society was often in opposition to the regulatory, political domain of the state.

The dichotomous conception of state and civil society, was, however, preceded by the distinction in Natural Law theory between *status civilis* and *status naturalis* (Tribe 1988). The latter was the sphere of discord which the English philosopher Thomas Hobbes (1588–1674) described as *bellum omnium contra omnes*—barter society in which individuals contracted with and against one another (Hobbes 1949), and to which his countryman John Locke (1632–1704) assigned the contentious process of the formation of private property. For Hobbes, the state of nature ended when it submitted to the 'civil union' 'called state or civil society.' For the Prussian Cameralists of the same period, organization of economic activity in political society, or its regulation, ensured the security and prosperity of the commonwealth, including providing the population with subsistence and productive employment. For the English mercantilist, James Steuart (1712–90), self-interest had to be restrained and directed by the 'reason of state' if it were not to damage public good, which he defined in terms of England's increasing commercial success.

Concentration of political power in administrative monarchies had resulted in a decline in the importance of the corporate bodies and deliberative institutions that had earlier served to mediate power. Groups that were excluded from the political process, as well as new bureaucratic elites and commercial classes, sought inclusion through societal associations and political institutions that could provide them with a voice against absolutism. In France, Britain, and Germany respectively, Montesquieu (1689–1755), Ferguson (1723–1816), and Kant (1724–1804) proposed conceptions of civil society focused on delineation of associational spaces, of environments for societal negotiations, of civility, and of publicity (Jacob 1991).

Responding to the despotism of the eighteenth-century French ancient regime, Montesquieu saw in civil society (*l'état civil*) a context for the societal negotiation of the absolute power of the monarch that was not a domain separate from the monarchy. His concept of division of powers addressed a situation in

which the monarchy had severed itself from the network of power relations within the political society and had appropriated for itself a separate sphere of power (*l'état politique*). It sought to reintroduce the monarch into the *l'état civil* by means of institutions that would check the absolute authority of the ruler and balance it against the authority of the landed aristocracy, their advocates in the judiciary, and commercial interests (Richter 1998).

Kant first employed the notion of civil society, or *bürgerliche Gesellschaft*, in the sense of political society inseparable from the Prussian absolutism of Frederick the Great (1712–1786), which he considered indispensable for social stability. Secondly, *bürgerliche Gesellschaft* referred to the public sphere or a domain of literate citizenry that was separate from the arena of political power and action. Kant regarded the political arena as the reserve of the state, or the ruler. *Bürgerliche Gesellschaft* referred to a sphere 'beyond the political order,' 'beyond the particularistic concerns of political action' where practical issues of governance could be debated on the basis of universal principles of reason. For Kant, the critical practice of exposing actual state policies to the light of universal reason could act to restrain the absolute power of the ruler as well as legitimizing his power (Ellis 2000). Kant's *bürgerliche Gesellschaft* was composed of individuals from the Prussian bureaucratic and bourgeois elites, educated and trained in state schools and administrative offices, as well as members of social clubs and associations, who could by dint of reason rise beyond the trappings of class or official status.

The thinkers of the Scottish Enlightenment who addressed the issue of civil society did not limit themselves to the issue of restraints to centralized state authority. Adam Ferguson pointed to the corrosion of civic spirit in political society, where the successful commercial classes became servile to the administrative state, which provided them with a 'rule of law' but deprived them of their traditional rights (Keane 1988, Ferguson 1995). He conceived of civil society as networks of self-governing and self-regulating voluntary associations, such as self-help groups and 'friendly' or charitable societies, which had expanded rapidly in the eighteenth century and, in Britain, played an important part in poverty relief efforts. Ferguson pointed to the potential of voluntary associations for engendering civility beyond the special interests of state administration and the commercial classes. The central question facing Ferguson and other eighteenth century European thinkers was how society increasingly differentiated administratively and economically, could remain integrated and harmonious. For Ferguson, as for his compatriot, David Hume (1711–1776), civility was the basis for social cohesion, and he saw it as rooted in sociability or moral and emotional communication among persons 'that fostered social bonds and friendships and cultivated manners and moral tastes' (Trentmann 2000). Ferguson did not place civil society in opposition to the state; rather, civil society constituted a protective shield from the uncertainties of social and political life.

3. *Civil Society as a Separate Domain*

For Adam Smith (1723–1790), also a thinker of the Scottish Enlightenment, civilized society consisted in self-regulating and interdependent networks of economic relations among individuals and groups, originating in the decisions of individuals competing in markets for goods, labor and capital. Economic activities of self-interested individuals were guided by the universal, natural laws of supply and demand. For Smith, the civilizing of economic society and the harmonizing of individual interests presupposed a mode of sociability and reciprocal sympathies among individuals (Trentmann 2000).

Smith proposed a separation between the civilized society of economic activity and the political sphere of the state and insisted on the liberation of labor, capital (including land), and goods from the network of relations of political society. The doctrine of *laissez faire*, *laissez passez*, first introduced by the eighteenth-century French Physiocrats and taken up by Smith, was a plea for the removal of regulations that privileged the landed classes through restrictions on the grain trade, merchants through grants of monopolies, and the 'poor' through grain subsidies provided by the Poor Laws. Smith was an advocate for the new industrial order and opposed to regulations that represented interventions by the state in the economic process, including the free exchange of labor, capital (including land), and of goods.

Smith argued that if individuals, unimpeded by privilege and state regulation, could act in accordance with their self-interest, markets could be trusted to allocate resources equitably in the form of wages, rents, and profits. This, for Smith, was the key to progress, economic growth, and national prosperity (Smith 1974). On the other hand, he was not entirely indifferent to the dislocations generated by markets and he assigned the civil state the duties of administering justice and providing security, including the protection of private property and correction of severe inequalities created by market forces.

Smith's civilized society, like Ferguson's civil society, Kant's *bürgerliche Gesellschaft*, and Montesquieu's *l'état civil*, represented positive images of civil society, which, in the course of encounters with hitherto unknown regions of Asia and the Americas during the seventeenth and eighteenth centuries, became part of Europe's definition of itself as the domain of the 'civilized' (Kocka 2000). The discourse of civilized Europe created its opposite in images of an uncivilized non-Europe, notably the East, which was perceived as the domain of despotism *pace* Montesquieu, of chafing regulations over economic

activity *pace* Smith, and of the absence of private property *pace* Marx. Following the last decade of the eighteenth century, with the onset of revolutionary upheavals against absolutist regimes, the terms *bürgerliche Gesellschaft* and civil society were increasingly equated with bourgeois society and acquired a negative connotation as 'the reign of dissoluteness, misery and physical and ethical corruption' (Hegel 1967, Kocka 2000). The European idea of itself as 'civilized' was effectively distanced from understandings of civil society as bourgeois society . No matter how 'barbaric' Europe became in Europe, it would remain 'civilized' vis-à-vis non-Europe, its optimistic image of itself having been fixed in a vocabulary of domination.

Hegel (1770–1831) responded to this situation with the idea of the universal state that was to end the separation between the political and the economic. He subscribed to the political economists' notion of civil society as a nexus of economic interests and relationships that continuously reproduce themselves. He did not, however, share Smith's utopian faith in the mechanistic operations of the market. Revolutionary upheavals had revealed a discrepancy between models of an orderly economy subject to universal laws of nature and the chaotic reality of contemporary society. Conceptualizations of civil society as an autonomous domain capable of generating order and progress through its own dynamics now appeared dubious.

Hegel's civil society was not the civilized society of Smith, but bourgeois society which manifested extremes of wealth and poverty that threatened to destroy the productivity of individuals pursuing their self-interest (Reidel 1984). The destructive potential of civil society, he felt, could be restrained through redirecting individual self-interest by administrative means, including justice, police, and moral measures.

Marx (1818–1883) subscribed to Hegel's critique of political economy and his conception of civil society as an arena of self-interest and divisiveness with a potential for self-destruction (Bottomore 1983). But unlike Hegel, Marx was not concerned with the containment of bourgeois society's destructive potential. Instead he focused on the revolutionary transformation of economic relations in civil society, which he believed was possible through the mobilization of conflicts between different interest groups. For Marx, the state, distinct from civil society or *bürgerliche Gesellschaft*, is largely limited to formalistic and negative activities since in his view, civil society both preceded and determined the state. Civil society required no regulation but was ruled through the contingencies of class struggle.

4. *Civil Society and the Liberal State*

Political and academic debates of the nineteenth century addressed the question of how to achieve social stability in market societies where the logic of economic activity overrode moral and political concerns regarding equity. Civil, or liberal, society was perceived as a domain shaped by or reformed through the practices of a central state representing the public interest. This constitutes a major departure from eighteenth century thinking. On the one hand, civil society, its actors, its activities, and the economic relations that characterized it, were viewed as inseparable from their legal and administrative formulations. This resulted in the creation of legal entities including trade unions, corporations, and family, and of voluntary and charitable associations, which, while autonomous from the state, remained within the bounds of its administrative–legal vision (Neocleous 1996). On the other hand, the state, 'objectified' in its administrative and legal practices, was understood to stand apart from civil society and to mediate divergent interests and needs.

From the perspective of the English utilitarians, most notably Jeremy Bentham (1748–1832), and of the German social economist Lorenz von Stein (1815–1890), civil society, or the market economy of the political economists, needed to be actively constituted. It was not enough simply to remove the obstacles of obsolete privilege and restrictive policies of mercantilism of the *ancien régime*. For Bentham, the new economic order required positive state intervention, and government was inseparable from an 'art of directing the national industry to purposes to which it may be directed with greatest advantage.' Von Stein, agreeing with Hegel's perception of civil society as a site of conflict and oppression, identified the 'social problem' as the main obstacle to economic progress. Progress required a market economy, or civil society; however, the injustices and inequalities it generated must be ameliorated through the state's administrative activity. The creation of social citizenship and the active participation of the citizen in the state's decision making were central to this process (Pasquino 1981). Thus, civil society represented an outcome of collective struggles and clashes among divergent interests that were mediated through the state's administrative practices, so that for von Stein, as for Bentham, the state was not the *Rechtsstaat* ('rule of law' state) that stood outside civil society, but the *Sozialstaat* (social state) representing a process whereby society was continuously formed and reformed.

Not all observers of nineteenth century Europe saw the centralized administrative state as the liberator of civil society. Conservative Romantics rejected the notion of the state or politics shaping society and assigned self-policing power to society and church. Alexis de Tocqueville (1805–1859) in his *De la democratie en Amerique* (1835–1840) saw the real danger to modern society in the new despotism of the all-pervasive state administrations rather than in class conflicts. Anticipating his fellow Frenchman, Michel Foucault (1926–84), de Tocqueville pointed to the

administrative suffocation of civil society as evidenced in the state's monopoly of public education, health care, and social services to the poor and unemployed that subjected all aspects of citizens' lives to state scrutiny (Keane 1988). De Tocqueville stressed the importance of voluntary associations in placing checks on administrative despotism by providing the services that people expected from government, thus pre-empting government intervention in civil society. Recalling the ideas of the Scottish Enlightenment, de Tocqueville saw voluntary associations as generating and expressing community civic values.

Civil society all but disappeared from political and academic debates in the late nineteenth century and the period following World War I, a time of slackening economic growth, rising working-class activism, imperial rivalries, and wars, which witnessed the displacement of political power from state administrations to major organized groups in society. The European welfare states, heirs to the liberal states, increasingly became sites for centralized and bureaucratic bargaining among political parties, labor unions, and business cartels vying for economic and political power and blurring the distinction between state and civil society, public and private (Maier 1987).

Antonio Gramsci (1891–1937), leader of the Italian Communist Party, represented a notable exception to the lack of interest in the notion of civil society in the post-World War I era. He reformulated the Marxist–Hegelian understanding of civil society in terms of the corporatist Zeitgeist of Italy in the 1920s. For Gramsci, civil society was not merely the sphere of individual needs but also of organizations where the hegemony of the ruling class and consent to that rule was negotiated. In this sense, civil society comprised not only all material (economic) but also political and cultural relations. While Marx insisted on separation between state and society, Gramsci held the two were interrelated. Hegemony, which was basic to Gramsci's notion of civil society, presupposed 'interpretations' of the economic structure; that is, the political and cultural mediation of different interests (Bobbio 1988). In that sense, Gramsci had a kindred spirit in Hegel rather than Marx.

5. *Late Twentieth Century Understandings of Civil Society*

After World War II, the European political and economic order focused politics on the state with welfare states subsuming civil society to state regulation. As in the late nineteenth century and the post-World War I era, economic growth, distribution, and security represented the core themes of politics. However, now these themes were generalized to regions outside Western Europe, including Eastern Europe and post-colonial societies in the so-called Third World. The different regions varied greatly in terms of how growth was attained and sustained, and the efficiency of realizing their economic and social goals. They also varied in terms of their political institutions and modes of resolution of social and political conflict. From the late 1940s to the mid-1970s sustained economic growth in the capitalist economies of Western Europe and North America made challenges to economic redistribution less urgent, and the internationalization of the security problem (e.g., the formation of NATO) led to a consensus regarding security. Capital transfers from a booming United States economy, including military and economic aid, helped to sustain the developing states of the Third World, while the statist socialist regimes relied on the ideological and political disciplining of their populations to generate surpluses centrally distributed by the state.

During the 1950s and 1960s, some social scientists in Western Europe and the United States predicted that under conditions of long-term growth, political problems could be largely transformed into noncontroversial administrative routines resolvable by experts. Socialist regimes perfected this functionalist perception of politics where government excluded all competing forms of political organization, including political parties, trade unions, and collective bargaining. However, this perspective precluded discussion of civil society. Socialist ideologues identified civil society with bourgeois society and held it was superseded by the advent of socialism.

In the mid-1970s, slowing economic growth in the advanced capitalist societies resulted in constraints on the ability of the state to tax business and distribute assistance to the working class in the form of social welfare benefits. The fiscal crisis was further aggravated by rising costs of welfare provisions, and called into question the validity of ideas that subsumed society to state regulation. Since the 1970s, social and political debates have largely focused on the critique of social statism and the differentiation of civil society from the state.

From a neo-conservative stance, over-politicization of the state and the fusion of political and non-political spheres of social life has eroded political authority, weakened governability, and destroyed the autonomy and authority of non-political spheres, including family, religion, and the market. Neo-conservatives proposed the boundaries between state and civil society be redrawn. Civil society, in opposition to the state, represented the depoliticized sphere of market activity and all else, including religion and family. It should be reintegrated through a cultural model that cohered the exchange activities of self-interested individuals. This is a model of the self-regulating civil society of non-profit voluntary associations reminiscent of the eighteenth-century

Scottish Enlightenment. During the 1980s and 1990s, social science paradigms based on individual self-interest and rational expectations, presented models for social programming that would render a politicized civil society unnecessary. In the neoconservative view, the state is a lean structure characterized by effective authoritarian forms of action, approximated by its Thatcherite interpretation in 1980s Britain.

For proponents of the voluntary social movements paradigm, statism had disastrous environmental, political, and social consequences. They proposed an understanding of civil society as an autonomous politicized sphere independent of regulation and constraint by bureaucratic political institutions. Social movements such as the environmentalist, anti-nuclear, and women's and gay movements of the 1970s, were politically motivated by concerns about the environment and quality of life, and issues of equity, authenticity, and participation. Their agendas required 'spontaneous' organization and not 'officially' sanctioned institutions of political parties and trade unions. Here, civil society referred to an intermediate institutional space between private (personal) and public (the object of official political institutions and actors).

Jurgen Habermas' (1929) idea of public sphere parallels this understanding of civil society in the context of ideas and communications; it includes the organization of civic or bourgeois opinion as represented by associations and media. Public sphere is a space where communication about collective values takes place. In advanced capitalist societies where political and economic domains merge, the state became a major actor in the market economy and thus no longer able to advance the common good. This, for Habermas, signals the legitimation crisis of advanced capitalist societies and raises the issue of the formulation of a public discourse outside of the spheres of market economy and the welfare state and in doing so, (in the spirit of Kant) subjects both spheres to the critical scrutiny of communicative rationality.

In the mid-1970s, the crises in socialist states, initially Poland, provoked Central and Eastern European intellectuals to address the issue of the politicization of civil society. After failed revolutions in Hungary (1956) and Czechoslovakia (1968), the new evolutionism of Adam Michnik and Jacek Kuron proposed the bottom-up construction of a civil society to repulse state intrusions into social life. For other dissident socialist intellectuals in the region, most notably Vaclav Havel, civil society represented the domain of anti-politics; it was a vision of society not simply independent of the state but opposed to it.

During the 1980s and 1990s in many parts of the world, including the former socialist states of Central and Eastern Europe, the expansion of the market system, including privatization programs, led to two divergent perceptions of civil society. The first represented civil society in opposition to global capitalism and to the state as the implementor of market reforms. The second emphasizes the centrality in civil society of religion, family, and voluntary associations for generating moral, economic, and cognitive norms.

See also: Civic Culture; Civil Society/Public Sphere, History of the Concept; Communication and Democracy; Democracy: Normative Theory; Democratic Theory; Hegel, Georg Wilhelm Friedrich (1770–1831); Hobbes, Thomas (1588–1679); Kant, Immanuel (1724–1804); Liberalism: Historical Aspects; Montesquieu, Charles, the Second Baron of (1689–1755); Participation: Political; Smith, Adam (1723–90)

Bibliography

Aristotle 1965 In: Barker E (ed.) *The Politics*. Oxford University Press, New York

Bobbio N 1988 Gramsci and the concept of civil society. In: Keane J (ed.) *Civil Society and the State: New European Perspectives*. Verso, London

Bottomore T (ed.) 1983 *Civil Society. A Dictionary of Marxist Thought*. Blackwell, Oxford, UK

Calhoun C (ed.) 1992 *Habermas and the Public Sphere*. MIT Press, Cambridge, MA

Cicero 1988 *The Republic*. Cambridge University Press, Cambridge, UK

Cohen J L, Arato A 1992 *Civil Society and Political Theory*. MIT Press, Cambridge, MA

Donzelot J 1984 *L'Invention du social*. Fayard, Paris

Ehrenberg John 1999 *Civil Society: The Critical History of an Idea*. New York University Press, New York

Ellis E 2000 Immanuel Kant's two theories of civil society. In: Trentmann F (ed.) *Paradoxes of Civil Society: New Perspectives on Modern German and British History*. Bergham, New York

Gierke O 1958 *Political Theories of the Middle Age*. Beacon, Boston

Ferguson A 1995 *An Essay on the History of Civil Society*. Transaction, New Brunswick, NJ

Havel V 1988 Anti-political politics. In: Keane J (ed.) *Civil Society and the State: New European Perspectives*. Verso, London

Hegel G W 1967 (Knox T M, English translation) *The Philosophy of Right*. Oxford University Press, London

Hobbes T 1949 (Lamprecht S P, English translation) *De Cive or The Citizen*. Greenwood, New York

Jacob C M 1991 The enlightenment redefined: The formation of modern civil society. *Social Research* **58**(2): 475–95

Kant I 1963 What is enlightenment? Beck L W (ed.) *On History*. The Bobbs-Merrill Company, Indianapolis, IN

Keane J (ed.) 1988 *Civil Society and the State: New European Perspectives*. Verso, London

Kocka J 2000 Zivilgesellschaft als historisches Problem und Versprechen. In: Hildermeier M, Kocka J, Conrad C (eds.) *Europäische Zivilgesellschaft in Ost und West: Begriff, Geschichte, Chancen*. Campus Verlag, Frankfurt/Main, Germany

Maier C (ed.) 1987 *Changing Boundaries of the Political: Essays on the Evolving Balance between State and Society, Public and Private in Europe*. Cambridge University Press, Cambridge, UK

Neocleous M 1996 *Administering Civil Society: Towards a Theory of State Power*. MacMillan, London
Offe C 1987 Challenging the boundaries of institutional politics: Social movements since the 1960s. In: Maier C (ed.) *Changing Boundaries of the Political: Essays on the Evolving Balance Between State and Society, Public and Private in Europe*. Cambridge University Press, Cambridge, UK
Pasquino P 1981 Introduction to Lorenz von Stein. *Economy and Society* **10**(1): 1–6
Reidel M 1984 *Between Tradition and Evolution: The Hegelian Transformation of Political Philosophy*. Cambridge University Press, Cambridge, UK
Richter M 1998 Montesquieu and the concept of civil society. *The European Legacy* **3**(66): 33–41
Rosanvallon P 1988 The decline of social visibility. In: Keane J (ed.) *Civil Society and the State: New European Perspectives*. Verso, London
Smith A 1974 *The Wealth of Nations*. Penguin, Harmondsworth, UK
Trentmann F (ed.) 2000 *Paradoxes of Civil Society: New Perspectives on Modern German and British History*. Bergham, New York
Tribe K 1988 *Governing Economy: The Reformation of German Economic Discourse, 1750–1840*. Cambridge University Press, Cambridge, UK

H. Islamoglu

Civil Society/Public Sphere: History of the Concept

The closely related concepts of civil society and public sphere developed in the early modern era to refer to capacities for social self-organization and influence over the state. Civil society usually refers to the institutions and relationships that organize social life at a level between the state and the family. Public sphere is one of several linked terms (including 'public space,' simply 'public,' and the German *Öffentlichkeit*, or publicness) that denote an institutional setting distinguished by openness of communication and a focus on the public good rather than simply compromises among private goods. Located in civil society, communication in the public sphere may address the state or may seek to influence civil society and even private life directly. Key questions concern the extent to which it will be guided by critical reason, and how boundaries between public and private are mediated.

1. *Civil Society and Self-organization*

The distinction of 'civil society' from the state took its modern form in the seventeenth and eighteenth centuries. Prior to this separation, political and social realms were seldom clearly distinguished. When they were, the social was exemplified by the family and often subordinated as the realm of necessity or mere reproduction to the broader public character and possibilities for active creation that lay in the state. The Greek conception of the *polis*, for example, usually referred to both, but when a distinction was made, it clearly favored the state.

Roman law contributed the idea of *civitas* and a stronger sense of relations among persons that were neither narrowly familial nor specifically about constituting the political society through the state. Medieval political and legal theory developed this theme, especially in relation to the freedoms claimed by medieval cities but also in relation to the Church. Some strands juxtaposed the notion of legitimacy ascending from 'the people' to the eventually dominant idea of divine right of kings, with its notion of legitimacy descending from God. Also influential was the distinction of civil from criminal law (in which the former governs relations formed voluntarily among individuals and the latter the claims of the whole society against malefactors). Nonetheless, it was only in the course of early modern reflection on the sources of social order that civil society came to be seen as a distinct sphere.

A crucial step in this process was the 'affirmation of ordinary life' (Taylor 1989). Whereas the Greek philosophers had treated the private realm—including economic activity—as clearly inferior to the public realm associated with affairs of state, many moderns placed a new positive value on family and economic pursuits. They argued that both privacy and civil society needed to be defended against encroachments by the state. In this context, it was also possible to conceive of a public sphere that was not coterminous with the state but rather located in civil society and based on its voluntary relations. In this communicative space citizens could address each other openly, and in ways that both established common notions of the public good and influenced the state.

Social contract and natural law theories—especially as joined in the work of John Locke—contributed to this shift by suggesting ways in which the creation of society conceptually preceded the creation of government. From this it was only a short step to say that the legitimacy of government depended on its serving the needs of civil society (or of 'the people'). Thomas Paine and other advocates of freedom from unjust rule advanced an image of the freedoms of Englishmen which was influential not only in England and America, but in France, notably in Montesqueiu's account of the 'spirit' of laws which combined an appreciation of English division of powers with an older tradition of republican (and aristocratic) virtue. From Rousseau through Tocqueville, Comte, and Durkheim, this French tradition developed an ever-stronger account of the autonomy of the social (resisting not only the claims of the state but the Cartesian postulate of the primacy of the individual subject).

A crucial innovation was to understand society as at least potentially self-organizing rather than organized only by rulers. If there was a single pivotal intellectual source for this, it lay with the Scottish moralists. In

Adam Smith's (1776) notion of the invisible hand, the market exemplified this self-organizing capacity but did not exhaust it. In his *Essay on the History of Civil Society* (1767), Adam Ferguson presented human history as a series of social transformations leading to modern society. This prompted Hegel (1821) to treat civil society as a field in which the universal and particular contended; their reconciliation depended on the state. The idea of civil society also shaped classical political economy and ideas of social evolution, and informed Marx's account of the stages of historical development as combinations of productive capacity and (conflict-ridden) social relations. Marx also challenged the notion that markets were neutrally self-organizing, emphasizing the role of historical accumulations of power.

Though the actual analyses differed, what had been established was the notion of society as a distinct object of analysis, not reducible to either state or individual. People formed society impersonally as actors in markets, more personally as parties to contracts. The idea of civil society hearkened back to the sort of social life that emerged among the free citizens of medieval cities because this was largely self-regulated—as distinct from direct rule by ecclesiastical or military authorities. It also suggested 'civility' in interpersonal relations. This meant not just good manners, but a normative order facilitating amicable or at least reliable and nonthreatening relationships among strangers and in general all those who were not bound together by deep private relations like kinship. Equally important, the idea of civil society included—in some versions—the notion that communication among members might be the basis for self-conscious decisions about how to pursue the common good. This notion is basic to the modern idea of public sphere.

2. *The Idea of a Public Sphere*

Rousseau (1762) famously sought to understand how social unity could result from free will rather than external constraint. This depended, he argued, on transcending the particular wills of many people with a general will that was universal. Kant admired Rousseau's pursuit of unity in freedom as distinct from mere social instinct (as in Aristotle's notion of a political animal) or imposition of divine authority. He relied implicitly on the idea of a collective conversation through which individual citizens reach common understandings. Likewise the development of representative institutions in eighteenth century England informed and anchored a public discourse directed at bringing the will and wisdom of citizens to bear on affairs of state. Finally, the idea of the people as acting subject came to the fore in the American and French revolutions. The idea of a public sphere anchored democratic and republican thought in the capacity of citizens in civil society to achieve unity and freedom through their discourse with each other.

Kant, like many eighteenth-century philosophers, lacked a strong notion of the social. This Hegel (1821) supplied, rejecting social contract theory because even in Rousseau's notion of a general will it suggested that the union achieved in the state depended not on its own absolute universality but on a development out of individual wills. Nationalism also shaped ideas of society and political community in holistic ways well matched to unitary states (Calhoun 1999). Marx's (1843, 1927) critique of politics based on bourgeois individual rights further challenged the adequacy of civil society as a realm of freedom and unity. Where Hegel thought that the state in itself might overcome the tension between necessity and freedom and the clash of particular wills, Marx held that only a transformation of material conditions including the abolition of private property could make this possible. As a result, theories stressing stronger ideas of the social were apt to offer weaker notions of public life. The Marxist tradition denigrated 'mere democracy' as an inadequate means of achieving either freedom or unity.

The ideas of public sphere and civil society developed primarily in liberal theory. These were not always seen in the manner of Hegel as merely 'educative' on the way to a more perfect latter unity. Nor was political unity necessarily left to the workings of an invisible hand or other unchosen system, but freedom was treated commonly as a matter of individual rather than collective action. This accompanied the rise of relatively asocial understandings of the market (Polanyi 1944). In addition, the emerging notion of the public sphere was not clearly distinct from other usages of 'public.' State activity, for example, was sometimes described as public without regard to its relationship to democracy or its openness to the gaze or participation of citizens. This usage survives in reference to state-owned firms as 'public' regardless of the kind of state or the specifics of their operation.

More important was the overlapping concept of 'public opinion' (see *Public Opinion: Political Aspects*). The dominant eighteenth-century usage emphasized open expression and debate, contrasting free public opinion to absolutist repression. At the same time, it generally treated public opinion as a consensus formed on the basis of reasoned judgment. 'Opinion' was something less than knowledge, but especially where it had been tested in public discourse, it was not simply sentiment and it gained truth-value from reflexive examination. Various euphemisms like 'informed opinion' and 'responsible opinion,' however, reflected both a bias in favor of the opinions of elites and an anxiety about the possibly disruptive opinions of the masses. During the nineteenth century, this anxiety came increasingly to the fore. Tocqueville (1840) and Mill (1859), thus, both contrasted public opinion to reasoned knowledge; Mill especially worried about 'collective mediocrity' in which the opinion of debased

masses would triumph over scientific reason. While advocates of the public sphere saw rational-critical discourse producing unity, critics saw mass opinion reflecting psychosocial pressures for conformity. Implicitly, they associated reason with individuals rather than any collective process. The distinction between 'public' and 'crowd' or 'mass' was lost in such views (Splichal 2000). Early positivist research into public opinion approached it as explicable on the basis of social psychology rather than as a species of reasoned argument. Toennies (1922) sought a way to discern when each approach ought to apply.

Conversely, in the late nineteenth and twentieth centuries, a new field of public opinion research developed that approached public opinion as an aggregation of individual opinions. The shift was based largely on the development of empirical polling methods. It brought a renewal of attention to differences within public opinion, and thus to the distinction between public and crowd (Blumer 1948, Key 1961). It also focused attention on patterns of communication among members of the public rather than the more generalized notions of imitation or emotional contagion. New media—first newspapers, and then broadcast—figured prominently in efforts to understand public communication. While Lippman (1960) and a variety of social psychologists worried that the new media would produce the descent to a lowest common denominator of public opinion that liberals had long feared, Dewey (1927) and other pragmatists defended the capacity for reason in large-scale communication. In this, they hearkened back to the eighteenth-century hopes of Kant and Rousseau.

Even before the apotheosis of the opinion poll, Cooley (1909) had argued emphatically that public opinion ought to be conceived as 'no mere aggregate of individual opinions, but a genuine social product, a result of communication and reciprocal influence.' A key question was whether this communication and reciprocal influence amounted to the exercise of reason. Peirce (1878) had argued that among scientists the formation of consensus on the basis of openness and debate was the best guarantee of truth. Could this view be extended into less specialized domains of public discourse? This has been an enduring focus for Jurgen Habermas, the most influential theorist of the public sphere.

3. *Habermas*

In the context of some cynicism about democratic institutions, Habermas (1962) set out to show the unrealized potential of the public sphere as a category of bourgeois society. He challenged most directly the tendencies in Marxism and critical theory to belittle democratic institutions—and also the collapsing of public into state characteristic not only of Hegel but of actually existing socialism. Habermas celebrated the emancipatory potential of a collective discourse about the nature of the public good and the directions of state action. This could be free insofar as it was rational—based on the success of argument and critique rather than the force of either status or coercion—and could achieve unity by disregarding particular interests—like particular statuses—in favor of the general good. The best version of the public sphere was based on 'a kind of social intercourse that, far from presupposing the equality of status, disregarded status altogether.' It worked by a 'mutual willingness to accept the given roles and simultaneously to suspend their reality' (Habermas 1962, p. 131).

The basic question guiding Habermas' exploration of the public sphere was: to what extent can the wills or opinions guiding political action be formed on the basis of rational-critical discourse? This is a salient issue primarily where economic and other differences give actors discordant identities and conflicting interests. For the most part, Habermas took it as given that the crucial differences among actors were those of class and largely political-economic status; in any case, he treated them as rooted in private life and brought from there to the public. He focused on how the nature, organization, and opportunities for discourse on politically significant topics might be structured so that class and status inequalities were not an insuperable barrier to political participation. The first issue, of course, was access to the discourse. This was not so simple as the mere willingness to listen to another's speech, but also involved matters like the distribution of the sorts of education that empowered speakers to present recognizably 'good' arguments. Beyond this, there was the importance of an ideological commitment to setting aside status differences in the temporary egalitarianism of an intellectual argument.

The public sphere joined civil society to the state by focusing on a notion of public good as distinct from private interest. It was however clearly rooted in civil society and indeed in the distinctive kind of privacy it allowed and valued.

The bourgeois public sphere may be conceived above all as the sphere of private people coming together as a public; they soon claimed the public sphere regulated from above against the public authorities themselves, to engage them in a debate over the general rules governing relations in the basically privatized but publicly relevant sphere of commodity exchange and social labor. The medium of this political confrontation was peculiar and without historical precedent: people's public use of their reason (Habermas 1962, p. 27).

This public use of reason depended on civil society. Businesses from newspapers to coffee shops, for example, provided settings for public debate. Social institutions (like private property) empowered individuals to participate independently in the public sphere;

forms of private life (notably that of the family) prepared individuals to act as autonomous, rational-critical subjects in the public sphere. But the eighteenth-century public sphere was also distinguished by its normative emphases on openness and rational political discourse. Habermas' concern focused on the way later social change brought these two dimensions into conflict with each other.

The idea of publicness as openness underwrote a progressive expansion of access to the public sphere. Property and other qualifications were eliminated and more and more people participated. The result was a decline in the quality of rational-critical discourse. As Habermas later summed up:

> Kant still counted on the transparency of a surveyable public sphere shaped by literary means and open to arguments and which is sustained by a public composed of a relatively small stratum of educated citizens. He could not foresee the structural transformation of this bourgeois public sphere into a semantically degenerated public sphere dominated by the electronic mass media and pervaded by images and virtual realities (Habermas 1998, p. 176).

While Habermas' account of the continuing value of the category of public sphere evoked by the eighteenth-century ideal set him apart from Horkheimer and Adorno (1944) and their pessimistic turn in critical theory, he largely incorporated their critique of 'mass society' as 'administered society' into his survey of twentieth-century developments and with it many of the fears of nineteenth-century liberals. He held that the public sphere was transformed not only by simple increase of numbers but by the success of various new powers at re-establishing in new form the power to 'manage' public opinion or steer it from above. Public relations agents and public opinion polls replaced rational-critical debate; electronic media allowed openness but not the give and take conversation of the eighteenth-century coffee houses. At the same time, rising corporate power and state penetration of civil society undermined the distinction of public and private, producing a 'refeudalization' of society.

4. *Arendt*

Hannah Arendt also focused on the problem of collapsing distinctions between public and private. Arendt emphasized the capacity of action in public to create the world that citizens share in common. The term 'public,' she wrote, 'signifies two closely interrelated but not altogether identical phenomena: It means, first, that everything that appears in public can be seen and heard by everybody and has the widest possible publicity. … Second, the term "public" signifies the world itself, in so far as it is common to all of us and distinguished from our privately owned place in it' (Arendt 1958, pp. 50, 52). Public action, moreover, is the crucial terrain of the humanly created as distinct from the natural world, of appearance and memory, and of talk and recognition. Such action both requires and helps to constitute public spaces—spaces held in common among people within which they may present themselves in speech and recognize others. Public action is thus a realm of freedom from the necessity—notably of material reproduction—that dominates private life.

Arendt's usual term, 'public space,' leaves the 'shape' of public life more open than the phrase public sphere. Public action can create institutions, as in the founding of the American Republic, but as action it is unpredictable. Its publicness comes from its performance in a space between people, a space of appearances, but it is in the nature of public action to be always forming and reforming that space and arguably the people themselves. This conceptualization offers clear advantages for thinking about the place of plurality in the public sphere. As Arendt wrote of America, 'since the country is too big for all of us to come together and determine our fate, we need a number of public spaces within it' (1972, p. 232).

Arendt saw this plurality threatened not just by mass conformity but by the reduction of public concerns to material matters. A focus on sex as much as on the economy threatens the public–private distinction. It not only intrudes on intimacy and private life but impoverishes public discourse. Arendt (1951) saw this problem as basic to totalitarianism, which could allow citizens neither privacy nor free public discourse. Totalitarianism is distinguished from mere tyranny by the fact that it works directly on private life as well as limiting public life. This is not just a matter of contrasting intentions, but of distinctively modern capacity. Modern sociological conditions offer rulers the possibility to reach deeply into the family in particular and personal life in general, to engineer human life in ways never before imagined.

This potential for collapsing the public and private realms is linked to Arendt's unusually negative view of civil society. 'Society,' she writes, is 'that curious and somewhat hybrid realm which the modern age interjected between the older and more genuine realms of the public or political on one side and the private on the other' (1990, p. 122). Civil society is first and foremost a realm of freedom *from* politics. But public freedom is freedom *in* politics. This calls for action that creates new forms of life, rather than merely attempting to advance interests or accommodate to existing conditions. This distinguishes Arendt's view, and republicanism generally, from much liberal thought: 'Thus it has become almost axiomatic even in political theory to understand by political freedom not a political phenomenon, but on the contrary, the more or less free range of nonpolitical activities which a given body politic will permit and guarantee to those who constitute it' (1990, p. 30).

The founding of the United States was a favorite example of such action for Arendt. The American

Founders imagined and created a new kind of society, a new set of institutions. This relied on citizens' public commitments to each other rather than assumptions about human nature or mere external application of law. The Founders 'knew that whatever men might be in their singularity, they could bind themselves into a community which, even though it was composed of "sinners," need not necessarily reflect this "sinful" side of human nature' (1990, p. 174). Arendt's vision of public life as central to a moral community shares much with a republican tradition that deplores the modern decline of the public sphere—generally associated with the rise of particular interests at the expense of concern for the general good, the deterioration of rational public discourse about public affairs, or outright disengagement of citizens from politics (see *Public Sphere: Nineteenth- and Twentieth-century History*). Republican accounts of the public sphere place a strong emphasis on the moral obligations of the good citizen; recent scholarship has often questioned whether citizens lived up to significantly higher standards in earlier eras (Schudson 1998).

5. *Differentiation in the Public Sphere and Civil Society*

Habermas' account of the public sphere has been enduringly influential (see Calhoun 1992). Its delayed translation into English in 1989 ironically contributed to an invigorating new reading shaped by both the fall of communism and widespread projects of privatization in the West. Critics within communist societies had revived the notion of civil society (as distinct from simply 'society') in order to speak of the realm outside state control and its relative absence in communist societies. Likewise, transitions away from right-wing dictatorships were often treated in terms of a 'return of civil society' (Perez-Diaz 1993). In the US, the idea of civil society was linked not only to democracy but to reliance on voluntary organizations and philanthropy (Powell and Clemens 1998, Putnam 2000).

What civil society signifies in contemporary political analysis is the organization of social life on the basis of interpersonal relationships, group formation, and systems of exchange linking people beyond the range of intimate family relations and without reliance on direction by the government. As a number of scholars of Africa have noted, it incorporates an unfortunate understanding of family privacy that underestimates the positive and supraprivate social roles that African kin organizations can play (see essays in Harbeson et al. 1994). Even more basically, references to civil society often fail to distinguish adequately between systemic capitalist economic organization and much more voluntary creation of social organization through the formation of civic associations, interest groups, and the like—a distinction Habermas has sought to stress. This has sometimes been a source of confusion in use of the public sphere concept to analyze distinctive institutional developments in diverse political and cultural settings (Calhoun 1993).

Civil society has been important to defenders of free market economics because it suggests the virtues of an economy in which participants' choices are regulated by their interests rather than their official statuses. In principle, such an economy is able to effectively produce and circulate goods on the basis of prices rather than government direction. Civil society has been equally important to advocates of democracy because it signifies the capacity of citizens to create amongst themselves the associations necessary to bring new issues to the public agenda, to defend both civil and human rights, and to provide for an effective collective voice in the political process. This involves both a free press and political mobilization on the basis of parties and interest groups (see Cohen and Arato 1992 for the most detailed review; also Chandhoke 1995, Seligman 1992, Alexander 1998, Keane 1999). Habermas (1992, p. 367) summarizes the recent usage: 'civil society is composed of those more or less spontaneously emergent associations, organizations, and movements that, attuned to how societal problems resonate in the private life spheres, distill and transmit such reactions in amplified form to the public sphere. The core of civil society comprises a network of associations that institutionalizes problem-solving discourses on questions of general interest inside the framework of organized public spheres.' Habermas' work more generally, however, reveals this to be a minimally theorized as well as optimistic usage. It highlights one aspect of civil society but does not make clear the most basic issue.

While part of the heritage of the idea of civil society has been the effort to organize society through public discourse, an equally influential part has been the claim to privacy, the right to be left alone, the opportunity to enter into social relations free from governance by the state or even the public. The idea of business corporations as autonomous creatures of private contract and private property thus reflects the heritage of civil society arguments as much as the idea of a public sphere in which citizens joined in rational-critical argument to determine the nature of their lives together. Civil society refers to the domains in which social life is self-organizing, that is, in which it is not subject to direction by the state. But this self-organization can be a matter of system function or of conscious collective choice through the public sphere (Calhoun 2001).

Habermas' account of the public sphere drew a variety of important critical responses. One of the first focused on the extent to which he focused on the bourgeois public sphere and correspondingly neglected nonbourgeois public life and failed to clarify some of the conditions built into the bourgeois ideal. Negt and Kluge (1972) responded with an account of the

proletarian public sphere. Clearly, workers have at many points built their own institutions, media, and networks of communication, and entered into contention with bourgeois elites and other groups over the collective good. But if this is a discursive competition—that is, if workers and bourgeois argue over what constitutes the collective good rather than only fighting about it—then this implies an encompassing public sphere, albeit an internally differentiated one.

Nancy Fraser (1992) has influentially emphasized the importance of 'subaltern counterpublics' such as those framed by race, class, or gender. Some publics—even very partial ones—may claim to represent the whole; others oppose dominant discursive patterns and still others are neutral. Not all publics that are distinguished from the putative whole are subaltern. As Michael Warner (2001) has suggested, the deployment of claims on an unmarked public as *the* public sphere is also a strategy, generally a strategy of the powerful. Yet, it is important to keep in mind both that the existence of counterpublics as such presupposes a mutual engagement in some larger public sphere and that the segmentation of a distinct public from the unmarked larger public may be a result of exclusion, not choice. Feminist scholars especially have drawn attention to both the gender biases within family life that disempower women and the historically strong gender division between public and private realms on which male political freedom has generally rested (Elshtain 1993, Young 2000).

6. Conclusion

Theories of civil society focus on the capacity for self-organization of social relations, outside the control of the state and usually beyond the realm of family. The basic question posed by theories of the public sphere is to what extent collective discourse can determine the conditions of this social life. Contemporary research on civil society and the public sphere turns on the breadth of political participation, the extent to which capitalist markets limit other dimensions of self-organization in civil society, the existence of multiple or overlapping public spheres, the impact of new communications media, and the quality of rational-critical discourse and its relationship to culture-forming activities. These issues also inform discussions about international civil society and its public sphere.

The concepts of civil society and public sphere took on their primary modern dimensions in the late eighteenth and early nineteenth centuries in Western Europe and to a lesser extent the United States. They have become important in a variety of other settings, including in conceptualizing social autonomy in relationship to communist and authoritarian states. They inform democratic projects as well as academic research in a variety of settings and are in turn themselves informed by cultural creativity and social action.

See also: Citizenship and Public Policy; Civil Society, Concept and History of; Democracy; Individual/Society: History of the Concept; Public Good, The: Cultural Concerns; Public Sphere: Nineteenth- and Twentieth-century History; State, History of

Bibliography

Alexander J C 1998 *Real Civil Societies: Dilemmas of Institutionalization*. Sage, Thousand Oaks, CA

Arendt H 1951 *The Origins of Totalitarianism*. Harcourt Brace, New York

Arendt H 1958 *The Human Condition*. University of Chicago Press, Chicago

Arendt H 1972 *Crises of the Republic*. Harcourt Brace Jovanovich, New York

Arendt H 1990 *On Revolution*. Penguin, New York

Blumer H 1948 Public opinion and public opinion polling. *American Sociological Review* **13**: 542–54

Calhoun C (ed.) 1992 *Habermas and the Public Sphere*. MIT Press, Cambridge, MA

Calhoun C 1993 Civil society and public sphere. *Public Culture* **5**: 267–80

Calhoun C 1999 Nationalism, political community, and the representation of society: Or, why feeling at home is not a substitute for public space. *European Journal of Social Theory* **2**(2): 217–31

Calhoun C 2001 Constitutional patriotism and the public sphere: Interests, identity, and solidarity in the integration of Europe. In: De Greiff P, Cronin P (eds.) *Transnational Politics*. MIT Press, Cambridge, MA

Chandhoke N 1995 *State and Civil Society: Explorations in Political Theory*. Sage, New Delhi

Cohen J, Arato A 1992 *The Political Theory of Civil Society*. MIT Press, Cambridge, MA

Cooley C H 1909 *Social Organization: A Study of the Larger Mind*. Scribner, New York

Dewey J 1927 *The Public and its Problems*. Ohio State University Press, Columbus, OH

Elshtain J B 1993 *Private Man, Public Woman*. Princeton University Press, Princeton, NJ

Ferguson A 1767 *Essay on the History of Civil Society*. Transaction Publishers, New Brunswick, NJ

Fraser N 1992 Rethinking the public sphere: a contribution to the critique of actually existing democracy. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA, pp. 109–42

Habermas J 1962/1991 *The Structural Transformation of the Bourgeois Public Sphere: An Inquiry into a Category of Bourgeois Society* [trans. Burger T]. MIT Press, Cambridge, MA

Habermas J 1992 *Between Facts and Norms*. MIT Press, Cambridge, MA

Habermas J 1998 In: Cronin C, De Grieff P (eds.) *The Inclusion of the Other*. MIT Press, Cambridge, MA

Harbeson J W, Rothchild D, Chazan N (eds.) 1994 *Civil Society and the State*. Lynne Rienner, Boulder, CO

Horkheimer M, Adorno T W 1944/1972 *Dialectic of Enlightenment*. Herder and Herder, New York

Hegel G W F 1821 *The Philosophy of Right* [trans. Knox T M]

Keane J 1999 *Civil Society*. Stanford University Press, Stanford, CA

Key V O 1961 *Public Opinion and American Democracy*. Knopf, New York

Lippman W 1960 *Public Opinion*. Macmillan, New York
Marx K 1843/1975 On the Jewish question. *Marx–Engels Collected Works*. Lawrence and Wishart, London, Vol. 3, pp. 146–74
Marx K 1927/1975 Critique of Hegel's philosophy of law. *Marx–Engels Collected Work*. Lawrence and Wishart, London, Vol. 3, pp. 1–129
Mill J S 1859 *On Liberty*. Penguin, London
Negt O, Kluge A 1972/1993 *The Public Sphere and Experience*. University of Minnesota Press, Minneapolis, MN
Peirce C S 1878/1992 *The Essential Peirce: Selected Philosophical Writings, 1867–1893*. Indiana University Press, Bloomington, IN
Pérez-Díaz V M 1993 *The Return of Civil Society: The Emergence of Democratic Spain*. Harvard University Press, Cambridge, MA
Polanyi K 1944 *The Great Transformation: the Political and Economics Origins of Our Time*. Beacon, Boston
Powell W, Clemens E (eds.) 1998 *Private Action and the Public Good*. Yale University Press, New Haven, CT
Putnam R 2000 *Bowling Alone*. Simon and Schuster, New York
Rasmussen T 2000 *Social Theory and Communication Technology*. Ashgate, London
Rousseau J-J 1762 *The Social Contract*. Everyman Paperback Classics, London
Schudson M 1998 *The Good Citizen: A History of American Civic Life*. Free Press, New York
Seligman A 1992 *Civil Society*. Free Press, New York
Sennet R 1977 *The Fall of Public Man*. Knopf, New York
Smith A 1776 *On the Wealth of Nations*. Penguin, Harmondsworth, UK
Splichal S 2000 Defining public opinion in history. In: Hardt H, Splichal S (eds.) *Ferdinand Toennies on Public Opinion*. Rowman and Littlefield, London, pp. 11–48
Taylor C 1989 *Sources of the Self*. Harvard University Press, Cambridge, MA
Tocqueville A de 1840/1844/1961 *Democracy in America*. Schocken, New York, Vol. 2
Toennies F 1922/2000 *Kritik der öffentlichen Meinung*; selections translated. In: Hardt H, Splichal S (eds.) *Ferdinand Toennies on Public Opinion*. Rowman and Littlefield, London, pp. 117–210
Warner M 2001 *Public and Counterpublics*. Zone Books, Cambridge, MA
Young I M 2000 *Inclusion and Democracy*. Oxford University Press, Oxford, UK

C. Calhoun

Civilization, Concept and History of

The concept of civilization is inextricably connected with the conditions of its emergence, most notably with the rise of historical consciousness in Europe in the eighteenth and nineteenth centuries, as well as the globalisation of this form of historical understanding and correlative forms of intellectual practice. The concept is complex and imprecise in its definition, but ubiquitous in its uses, and inextricably imbricated with other categories by which historical materials are organized, such as culture, nation, and race. Apart from designating certain morphological features of human society, particularly with reference to urbanism and urbanity, civilization has been a schema for historical categorization and for the organization of historical materials. Here it has generally taken two forms, the universalist evolutionist, and the romantic particularist. The latter was tending to regain, in the ascendant context of identity politics, a certain hegemonic primacy worldwide at the close of the twentieth century. In all, the concept of civilization forms a crucial chapter in the conceptual, social, and political history of history; it, or its equivalents are presupposed, implicitly or explicitly, in the construal and writing of almost all histories.

1. Pre-History

1.1 The Past Continuous

The mental and social conditions for speaking about civilization in a manner recognizable in the year 2000 were not available before the middle of the eighteenth century. Hitherto, in Europe as elsewhere, large-scale and long-term historical phenomena, which later came to be designated as civilizations, had been categorized in a static manner that precluded the consciousness of directional or vectorial historicity as distinct from the mere register of vicarious change.

Hitherto, the succession of large-scale historical phenomena, such as Romanity or Islam, had been regarded (a) typologically, most specifically in the salvation-historical perspective of monotheistic religious discourse, in which successive events are taken for prefigurations and accomplishments of each other; (b) in terms of the regnal succession of world-empires; (c) in the genre of regnal succession, which started with the Babylonian king-lists and the earliest stages of Chinese historical writing, and culminated in medieval Arabic historical writing. Not even the schema of state cycles evolved by the celebrated Ibn Khaldun (d. 1406), where civilization (*'umrân*) was quasi-sociologically identified with various organizational forms of human habitation and sociality, could meaningfully escape from this finite repertoire of possible historical conceptions.

In the perspective of typology, the continuity of historical phenomena was expressed in the repetition of prophecies successively reaffirming divine intent and inaugurating a final form of order whose *telos* would be the end of time. Thus the Jewish prophets repeat each other and are all figures for Abraham; Jesus is at once the repetition and termination of this unique cycle of terrestrial time and is prefigured in Jewish prophecies; Muhammad is the final accomplishment and the consummation of earlier prophetic revelations, prefigured in Jewish and Christian scriptures; his era inaugurates the consummation of time with the Apocalypse. The structure of time in the Talmud, in the Christian writings of Eusebius (d. 339),

of Augustine's Spanish pupil Orosius (fl. 418), of Bernard de Clairvaux (1153), as in Muslim writings such as Muhammad's biography by Ibn Ishâq (d. after 761) or the universal histories of Tabarî (d. 923) and Ibn Kathîr (d. 1373), is homologous.

All but Jewish typology is independent of ethnic origin or geographical location, and construes historically significant units as religious communities of ecumenical description. This yields the second mode of organizing historical phenomena, the regnal. Thus, the regnal categorization of long-term historical phenomena of broad extent was expressed in terms of the succession of four ecumenical world-empires, succeeding one another as the central actors in world history: the Assyrio-Babylonian, the Median-Persian, the Alexandrian-Macedonian, and the Roman. This perspective was shared by the Book of Daniel and by ecclesiastical works, especially Syrian and Byzantine apocalypses influenced by it, albeit with minor variations, as in Orosius' substitution of the Carthaginian world-empire for the Median-Persian. Muslim caliphs considered their own ecumenical world empire to be the fifth and final order of world history, a conception shared by Muslim apocalypticism and, with many complications and nuances, by universal histories written in Arabic, all of which regarded dynastic succession as both prophetic inheritance and as the renewal of ecumenical imperial ambition, transferred from one line to another.

Analogously, medieval Christian polities, Byzantine as well as Frankish, subscribed to the same theory of *translatio imperii* by regarding themselves as being in a direct line of typological continuity with Rome, variously through Byzantium, the 'New Rome,' or through the Holy Roman Empire. In both, Romanity was the worldly cement of Christianity. This was a conception developed by Eusebius for his contemporary overlord Constantine, and was to remain effective until the dawn of modern times. In all cases, the past was understood to have been completed at its inception, with subsequent polities re-enacting the foundational event.

Finally, mention must be made of the disjunction between these meta-historical and transcendental realms of typological continuity, and the all-too-human chaos of particular histories. No movement or qualitative change is discernible in the context of these, only the predictable succession of wars, rapine, pestilence, and occasionally of praiseworthy acts, without connection with an order of reality that might transcend the events themselves and render sense unto them.

1.2 The Past Estranged

Two roughly contemporary events heralded a new conception of history that made the modern notion of civilization conceivable. Both were specifically European, but their conceptual consequences were globalized in the course of the nineteenth and twentieth centuries. The first was Humanism, particularly in Italy, which for the first time broke the spell of Roman continuity by construing the immediate past as an age of darkness, and by confining Roman grandeur to the republican and early imperial ages. Thus Petrarch's (d. 1374) project of classicism counterposed to living tradition, Flavio Biondo's (d. 1463) anticipation of division of history into the classical, the medieval, and the modern, and Lorenzo Valla's (d. 1457) refutation of medieval documentary forgeries such as the *Donation Constantini*, based on an argument from anachronism: Together, these laid the ground for a view of history as the domain of change rather than of repetition.

Needless to say, the notion of *translatio imperii* was no longer tenable in this context. It is at this point that Humanism converged with the other event foundational of the modern historical consciousness, namely, the Reformation. Criticism of the Church by Wycliff (d. 1384) and Luther, and the historiographic expression of this anti-Traditionalist fundamentalism in John Foxe's *Acts and Monuments* (1563), were in crucial ways congruent in their conception of the past with Humanism. The identification of the Pope with the Antichrist, the designation of the greater part of the history of Christianity as a history of falsehood, and the devalorization of the immediate past as abiding Tradition and its construal as degeneration, also led to the rejection of the notion of *translatio imperii*, and the substitution of notions of Reformation and renovation to that of transfer in continuity.

Thus the ground was prepared for notions of rise, decline, and fall—most notably the decline and fall of Rome and the reclamation of Roman republicanist models in a spirit of revivalism—that were finally to mature in the eighteenth century, with Gibbon and Montesquieu among others, spurred along with the development and ultimately, in the late eighteenth and during the nineteenth centuries, the institutional transformation of philology and antiquarianism into history as a topic of research detached from rhetoric. This was far beyond the late flowering of medievalizing typology with Bossuet (d. 1704), and it made the past tangible in its having-been (*Vergangenheit*), most graphically represented in the establishment of museums during the eighteenth century in many European capital cities.

With Voltaire and other eighteenth-century figures like Volney and Chardin, another notion crucial for speaking of civilization was developed. The notion of qualitative societal and cultural difference (*les moeurs*)—quite apart from the dynastic and the religious—was now available in the eighteenth century, as Europe was accounting for her differences from the Ottomans, the Persians, the Chinese, and tribal peoples in the Americas.

Whereas previously the notion of historical senescence may have been used in a tragical and rhetorical sense, the new notion of decadence required the correlative notion of progress and amelioration. These notions are the very conditions of possibility for conceiving of civilization, as the accomplishment of a continuous line of historical development in which origins and beginnings are transcended rather than repeated.

2. *Word and Concept*

2.1 *Civilization and Culture*

The terms civilization and culture are intimately related in their reference, and in many instances are used almost interchangeably, according to national and linguistic conventions. Both are terms of ancient vintage which underwent a gradual lexical expansion until, in the eighteenth century, they came to designate meanings that are recognizable in 2001.

In the course of the seventeenth and eighteenth centuries, the meaning of the term culture expanded, in English, French, and German (as *Kultur*) from a medieval sense indicating the cultivation of land and of the religious cult, figuratively to denote the maintenance and cultivation of arts and letters. This figurative sense was further extended in the course of the eighteenth century to encompass the non-material life of human societies in a very broad sense encompassing the refinement of manners no less than intellectual and artistic accomplishments associated with the Enlightenment—the former sense still persists in terms such as *haute couture* and *Kulturbeutel* (vanity case).

The term was used both locally and in contrast to societies adjudged still living in a state of nature, though in German the accent had been on an aesthetic of the lofty and the sublime as distinct from the crassly material, in association with a correlative emphasis on cultivation, *Bildung*, both individual and collective. In this way, the term was opened up to impregnation by the emergent notion of progress, the progress of individual societies as of humanity in general regarded both as a process natural to human society and as a principle of normative ranking among societies.

Not dissimilarly, and in imitation of 'culture,' in the eighteenth century the term civilization underwent—most especially in France and somewhat later in England—a figurative expansion in its lexical reference from the Latin *civilis*, life under reputable forms of government, to the broader designation of order, civil, and governmental. This order befitted developed societies that might regard themselves as civilized in contradistinction to other, barbarian or savage, yet to be civilized societies. The contrastive connotations of 'civilization' were far more accentuated than those of 'culture,' which was more commonly used in Germany, a land then with little or no experience of the world outside Europe. 'Culture' also came to be used in France and more saliently in England, as in Germany, decidedly to signal social distance and social distinctions within particular countries.

In all cases, these two terms increasingly came to be associated with a developmental perspective on history: not only the linear and cumulative course traversed by historical phenomena in time, but also of languages, geological layers, plant and animal species, and human societies generally (and later, races), towards greater differentiation, complexity and accomplishment. Correlatively, the meanings conveyed by these terms were implied by other terms or by none at all, as for example with Rousseau and Voltaire.

The nineteenth century witnessed a complicated relationship between 'culture' and 'civilization' whose fields of connotation and denotation shade into each other in a manner that has not helped the distinctive clarity and definition of either. The crucial player here was Germany, where *Kultur* took a decidedly romantic-nationalist turn in the early nineteenth century, dwelling on national uniqueness and individuality, and buttressed by the emergence of *Kulturwissenschaft* as a discipline and by studies of folklore.

The further developments of this politico-cultural impulse towards the end of the nineteenth century led not only to the profuse discourses on decadence, *Entartung*, ('disnaturation') but to the extension to France of this particularitic understanding of culture—and of civilization—under the influence of de Joseph de Maistre and the Catholic Counter-Enlightenment. As a result of Franco-German conflicts and of the severe stresses within France, a battle was waged between the advocates of 'culture,' upholders of national particularism, and of 'civilization,' champions of Enlightenment universalism accused by their detractors of crass materialism, a battle that reached its apogee during the first World War in the polemics between Romain Rolland and Thomas Mann.

Yet 'civilization' itself had been increasingly more receptive to particularism and nationalism, most specifically in historical writing. Books on the civilization of France, Germany, Europe, Italy, and England, emerged from the second quarter of the nineteenth century, and in certain ways the very currency of the term made it open to divergent uses. Towards the end of that century, the term came to be used under the influence of the German notion of *Kultur*, a term often rendered as 'civilization' in French translations of German works. One additional but decisive factor was English anthropology, with the appearance in 1871 of Edward Tylor's *Primitive Culture* (1958), and the subsequent predominance of the term 'culture' to designate a condition, on a ladder leading from savagery to barbarism and finally to civilization, in Anglo-Saxon anthropology. With the discovery and mystique of classical Greece towards

the end of the eighteenth century, of the unity of the 'West' from Greece, through Rome, on to the Romano-Germanic peoples in the Middle Ages, culminating in modern European civilization, was to become the *locus classicus* of this notion of civilization.

In the latter part of the nineteenth and throughout the twentieth century, both the terms and conceptions outlined were subsequently taken over and became crucial instruments of historical categorization in India, China, and the Arab countries and elsewhere. In Arabic *thaqafa*, an equivalent to the German *Bildung*, came to stand for culture, normally taken to designate intellectual and artistic life in a manner strongly elitist in character, and *hadara* was used for civilization, taken as a more general concept indicating the entire life of society including material life.

2.2 Continuities: Relativism

Linear developmentalism had, in general, underlain the entire body of diverse discourses on civilization and of its sister concept of culture. This developmentalism bifurcated along lines that may be characterized broadly as German and French in original inspiration.

Of the two, the former had been the more intimately associated with romantic national and, later, with civilizational particularism, producing a natural history of human groups regarded in analogy to organic species. These human groups were thus conceived as self-subsistent, continuous over time, largely impermeable, essentially intransitive, and according to many representatives of this view, almost congenitally given to conflict and war.

Originating in conservative reactions to the Enlightenment, with a strong anti-Gallic political impulse, this theory of historical and social development was associated with figures such as Johann Gottfried Herder in Germany and Edmund Burke in England, though it did have a strong representation in France among royalist, Catholic, and other anti-revolutionary (1789, 1848, 1871) currents represented by figures such as de Maistre, Gobineau, and Gustave Le Bon. The principal conceptual feature of this anti-mechanistic concept of history was insistence on individuality in the histories of different nations, races, and civilizations—terms often conflated in various combinations, in analogy with biological organisms, and perhaps best captured in the capacious semantic field of the German word *Volk*. In this sense, one may speak of continuity with pre-Enlightenment concepts of a social organism modeled upon the integrated somatic unity of the human body, as had been previously thought in medieval Arabic historico-political writings and in medieval European conceptions from the time of John of Salisbury (d. 1180). The fear of decline and decadence, conceived as a breakdown of a natural order, was the specific point at which continuity with medieval organismic concepts of the historico-political order made itself evident and conceptually formative. This was at a time when the idea of progress—and in contrast to it—had become a genuinely historical category, involving a consequential conception of change, of evolutionism, of the temporality specific to events (*Verzeitlichung*), and of a distanciation correct between human and natural histories.

The course of a particular history was seen to reside in a number of essential features, which Herder termed *Kräfte*, resulting in a history which, increasingly elevated and evolutive in the course of time as it may be, was still governed by principles which were, in essence, changeless, principles which imparted individuality upon these intransitive histories.

Whereas the Enlightenment provided, and the nineteenth century elaborated, a notion of genetic development along an axis of cumulative time to which civilizations and other historical masses are subject, the organismic, particularist notion conceived a civilization as bound to a self-enclosure inherent in its origins. Although this did not necessarily lead to an historical cyclism, it constituted its conceptual condition of possibility, and facilitated culturalist notions of nationalism that spoke in terms of 'revival.'

It was a cyclical notion of the history of cultural and civilizational circles, *Kulturkreise* according to Ernst Troeltsch (1920) which was made conceivable by the manifold discourses on decadence, malfunction, and historical pathology, notions which implicitly involve a measure against a more consummate state of organic health and well-being implicit in the foundations of each. It was the systematic elaboration of the decadence/normalcy structure that led to great schemas of world-history, divided into intransitive civilizations, of Oswald Spengler's *Untergang des Abendlandes* (Spengler 1922) and of the Spenglerian heresy represented by Arnold Toynbee's *A Study of History* (Toynbee 1934–1959).

Described by Claude Chaunu (1984 as 'a samsara of historical forms,' these vitalist theories of the rise and inevitable decline of civilizations constitute a naturalistic morphology of historical becoming. From this perspective, civilizations were seen as historical phenomena that are perpetually in conflict with one another, each endowed with a particular ethos or animating principle. Such were Spengler's (1922) Magian culture (the Perso-Islamic) and Promethean culture (the European). Such were also Toynbee's (1934–59) Syriac and other civilizations, though with this latter author the inner definition of civilizations was less clearly predetermined, and founded on a firmer and far more scrupulous empirical foundation than with Spengler. Nevertheless, Toynbee does characterize civilizations in terms of particularistic impulses, such as the aestheticism of Greek civilization, the religious spirit of the Indian, and the mechanistic ethos of the West. Each of these is an integrated pattern of daily life, on

attitude towards the holy, a style of jurisprudence, a manner of government, an artistic style, and much more.

With both authors, the historical phenomena respectively designated as 'cultures' and 'societies,' obey an iron law of rise and decline, of glory and sensecent atrophy, the terminal phases of which Spengler, in keeping with German usage of the day, derisively termed 'civilization.' It is noteworthy that this morphology of historical masses, be they called civilizations, cultures or societies, that this monocausal description in terms of basic traits such as the Promethean or the aesthetic, was congruent with certain developments in anthropology, particularly in American anthropology in the first half of the twentieth century (Franz Boas, Alfred Kroeber, Ruth Benedict, and Edward Sapir) which identified separate societies according to self-consistent and intransitive personality profiles they ostensibly gave rise to. This had a decisive influence upon the introduction of organismic thinking into the human sciences in general. After a long period of disrepute, the 'culturalized' notions of human collectivities, of self-enclosure, of continuity, have come back to center stage at the close of the twentieth century, correlatively with the politics of identity worldwide.

The organismic and vitalist notion of civilization was, and still is, extremely effective in the writing of history and in the late twentieth century has had a certain political salience in terms of Samuel Huntington's 'war of civilizations' and the mirror-image riposte in terms of the 'dialogue of civilizations.' It denies the possibility of a general human history, which it regards, in the words of Ernst Troeltsch (1920), as being 'violently monistic,' and construes the task of a history of civilizations as separate histories of Europe, China, India, Islam, Byzantium, Russia, Latin, and Protestant Europe, and others, in various possible permutations and successions, as separate cultural and moral spheres which are merely contiguous in space. Correlatively, this constituted the conceptual armature of certain forms of violently particularistic history that was mirrored or endogenously paralleled, among others in India for instance (Savarkar 1969), or in the writings of radical Muslim ideologues.

2.3 Novelties: Universalism

The deficit in historicity evident in romantic and vitalist historism, with its emphasis on organic continuity, was to a great extent made up in the more consistently evolutionist accounts of historicism. It was this conception of history, at once evaluative, evolutive, and vectorial, which bore the burden of the universalist notion of civilization. Whereas in historism the bearers of civilization—or rather of 'culture'— were particular peoples or individual historical itineraries, such as the West or Islam, the whole of humanity partook of the development of civilization which, in the historicist perspective was universal and continuous.

Several versions of this were in evidence, of which two are of particular note due to their wide conceptual incidence and to the social and political influence they exercised through worldwide social movements, both revolutionary and gradualist, inspired by the Enlightenment. According to this conception of a universal civilization, human societies pass through a uniform series of progressive developments which result in intellectual and moral elevation. They also result in superior social, economic, and political levels of development, marked by a higher and wider order of rationality, social differentiation, and control over nature whose instrument is science.

The more consistently universalist of these versions—exemplified, among others, by Jean Marquis de Condorcet, Auguste Comte, and Herbert Spencer, and in the evolutionist anthropology of Tylor (1958)—saw the whole of humanity as being predisposed to this upward movement. Nevertheless, according to this view some peoples may still subsist in a condition that others (Europe) had already surpassed, being still captive to superstition, to a weak organization of society and the economy, and to undeveloped political institutes (despotism, as in the case of Orientals, or otherwise various forms of acephalic organization, as is the case with primitive societies). The only caveat here is that many theorists of the evolution of human societies did not see human improvement as uniform in developed societies themselves, but that developed societies were not only internally differentiated in the levels of accomplishment attained by different social groups, but were also liable to fall far below the moral ideals that development makes possible. Nevertheless, much writing on the history of universal civilization, especially in countries influenced by Marxism, such as the Soviet Union or the People's Republic of China, saw the historical itineraries they led as crucial and exemplary steps in the development of humanity at large. In this way, the histories of Russia and China come to recuperate and tap world history at large by acting as the vanguard and exemplar of its future consummation.

Correlatively, the other conception of universalism, and this is one of profound conceptual and political importance, recognized the past contributions of various civilizations—the Mesopotamian, Egyptian, and Islamic, and in some instances the Indian and the Chinese as well—to the course of human civilization, which eventually lodged itself in Europe, defined as the abode of Romano–Germanic history, or even in small parts of Europe, such as France or Prussia. This writing of universal history, much in evidence in the plethoric literature of universal histories both popular and learned, was expressed in what is perhaps its most

accomplished general formulation by Georg Hegel (1956) in the nineteenth century, and by Karl Jaspers (1949) in the twentieth—the latter is little-read today, but he is one who nevertheless captures this notion with special clarity.

Not infrequently, this second type of universalist history is allied to an important element derived from the romantic theory of the history of civilizations treated in Sect. 2.2, namely the presumption of very long-term individual continuity. Thus the point is habitually made, with varying shades of emphasis and nuance, that universal civilization made Europe its eventual home because some abiding characteristics possessed *ab initio* by 'the West,' such as rationality, the spirit of freedom, vigor and dynamism. This changeless West is counterposed to an eternal East, such as Mesopotamia, Egypt, and Islam which, their relative erstwhile merits apart, constitute, in a categorical degradation, a mere prehistory to fully developed civilization. East and West—not to speak of Islam, the name of a religion transmuted into an atopic location—are metageographical notions which do not allow, except with anachronistic violence, for projection into the antique and late antique worlds. However, this does not disturb the ideological coherence of this notion of civilization.

3. *Beyond Totality*

Writing about civilization according to the manners outlined contained both large-scale abstraction and a great deal of precise empirical historiography. Under the influence of Marxism, acknowledged as such as well as implicit, historical scholarship, and most specifically social and economic history in the traditions of Max Weber and of the Annales school, produced specifications regarding material and other aspects of civilization that allayed, to a considerable degree, the rhetorical force of thinking civilizations in terms of purely moral and ideological continuities, betokening exclusive socio-historical groups. The notion of a Judæo-Christian civilization, as distinct from textual typology within the Bible, is an excellent illustration of this, having been born and expanded in specific circumstances following the second World War.

In this way, civilizations for contemporary historical writing have come to comprise the total historical conditions that exist in a specific place and time: functional and organizational forms of the state, the *longue durée* of demographic, agricultural, economic, social, urban, ecological, climatic, and other forces underwritten by geographical structures and relations, and non-material culture, such as arts, letters, cognitive structures, and religions. The possibilities of a total history of a civilization is, it must be stressed, an expansion of a previous and more limited form of the cultural history of a particular epoch, exemplified by Jakob Burckhardt's studies of Constantine and Renaissance Italy (Burckhardt 1955, Vols. 1 and 3).

As a result, since about 1950 it has become possible historically to specify the material elements that constitute the history of a civilization, however its temporal and geographical boundaries may be defined. Correlatively, it has become possible to conceive the specific differences between histories—China and Europe for instance, as in the work of Jacques Gernet (1988)—beyond a discourse on immobility and other immanent characteristics ascribed to this history or that, and to think of specificity in proper historical terms, such as the relative weight of various elements of the rural economy, the relation between state and the economy, the impact of metallurgy, and much more.

In this context, civilizational continuities came to be reconsidered in terms of historically determinant factors of a predominantly geographical nature, not so much in the spirit of geographical determinism as described by the German school of Friedrich Ratzel (1895), but with a greater degree of temporal specification, mediated by Lucien Febvre's (1925) consideration of historical geography and culminating in Fernand Braudel's study of the Mediterranean (1972–73). By the same token, it has become possible squarely to face the nominalist caution required in thinking about civilizations, and to think of their constitution, specification, and collapse in terms of the concrete historical investigation of demography, economy, and society without recourse to the metaphysical rhetoric of decline (Tainter 1988).

These specifications apart, it remains true that the construal of civilizational intransitivity, with Braudel as with others, still needs to resort to a new redaction of the rhetoric of permanence, most particularly with regard to non-material culture, now underscored and almost overdetermined by considerations of relief, soil, water supply, and means of transport—all of which are undeniable factors, albeit ones that modern technology and economy, most poignantly the post-modern economy, have rendered questionable.

Nevertheless, recent historical research has made it concretely possible to tap the genial formulations made by Marcel Mauss in 1930 (Febvre et al 1930) concerning the categorization of historical masses: of civilizations as a 'hyper-social systems of social systems,' as trans-societal and extra-national units of historical perception and categorization. These are conceived in opposition to specific social phenomena, and this conception of civilization valorises the distinction between civilization, society, and culture, freeing the first of the deterministic and totalizing rhetorical glosses of metahistory, and making possible a veritable history of civilizations. Civilization may thus be considered as at once a particular instance of historical becoming, and a specific ideological redaction of the past whose relation to historical reality can be questioned and rendered historical. In this context,

the historian may also be able to valorise the extremely expansive *longue durée* implied by such theories as Dumézil's (1958) Indo-European tri-functionalism without recourse to organismic and totalizing figures of particularity and continuity (Le Goff 1965). An historian may be similarly able to valorise recent studies that stress the occurrence and communicability of recurrent phenomena of the imaginary order across vast spaces, cultures, histories, and times, as instanced by Carlo Ginzburg's study of the European witch-hunts (Ginzburg 1990). Finally, given the accent on complexity, one might be able to take the precise and nuanced study of levels and modes of socio-economic, political, institutional, ideational, and other instances of complexity—rather than criteria of simple continuity—as crucial to the delimitation of historical phenomena that one designates as 'civilizations.'

See also: Civilizational Analysis, History of; Civilizations; Cultural Landscape in Geography; Enlightenment; Global History: Universal and World; History: Overview; Modernization and Modernity in History; Societies, Types of; Society/People: History of the Concept; State and Society; State, History of; Time, Chronology, and Periodization in History

Bibliography

Auerbach E 1984 Figura. In: Auerbach E *Scenes from the Drama of European Literature*. University of Minnesota Press, MN

Bénéton P 1975 *Histoire des Mots: Culture et Civilisation*. Fondation Nationale des Sciences Politiques, Paris

Braudel F 1972–73 *The Mediterranean and the Mediterranean World in the Age of Philip II*. Reynolds S (trans). 2 Vols. Collins, London

Braudel F 1993 *Grammaire des Civilisations*. Flammarion, Paris

Burckhardt J 1955 *Renaissance Italy*. Gesemmete Werke. Vols 1 and 3. Schwab, Basle, Switzerland

Chaunu P 1981 *Histoire et Décadence*. Librairie academique Perrin, Paris

Collingwood R G 1946 *The Idea of History*. Clarendon Press, Oxford, UK

Dumézil G 1958 *L'idéologie Tripartie des Indo-Européens*. Collection Latours, Bruxelles, Belgium

Enciclopedia Einaudi 1977–84. G. Einaudi, Torino (s.v. 'Cultura materiale', 'Civilità')

Encyclopædia Universalis 1968 Encyclopædia Universalis France, Paris (s.v. 'Civilization', 'Culture et civilization')

Encyclopedia of Religion and Ethics 1980 T and T Clark, Edinburgh, UK (s.v. 'Civilisation')

Febvre L 1925 *A Geographical Introduction to History*. Kegan Paul, Trench, Trubner & Co., London

Febvre L, Mauss M, Tonnelat E, Nicophoro A 1930 *Civilisation. Le Mot et l'Idée*. Renaissance du Livre, Paris (*Première Semaine internationale de Synthèse*, Fascicule 2)

Fehl E (ed.) 1971 *Chinese and World History*. The Chinese University of Hong Kong, Hong Kong

Gernet J 1988 *A History of Chinese Civilization*. Forster J R (trans.). Cambridge University Press, Cambridge, UK

Ginzburg C 1990 *Ecstasies: Deciphering the Witches' Sabbath*. Hutchinson, London

Hegel G W F 1956 *Philosophy of History*. Sibree G (trans.). Dover, New York

Herder J G von 1968 *Reflections on the Philosophy of History of Mankind*. Abridged by Manuel F, University of Chicago Press, Chicago and London

Herder J G von 1969 *JG Herder on Social and Political Culture*. F M Barnard (ed. and trans.). Cambridge University Press, Cambridge, UK

Ibn Khaldun 1958 *The Muqaddimah: An Introduction to History*. Rosenthal F (trans.). Pantheon Books, New York

Jaspers K 1949 *Vom Ursprung und Ziel der Geschichte*. Artemis Verlag, Zürich, Switzerland

Kemp A 1991 *The Estrangement of the Past. A Study in the Origin of Modern Historical Consciousness*. Oxford University Press, Oxford and New York

Kosellek R, Widmer P (eds.) 1980 *Niedergang. Studien zu einem geschichtlichen Thema*. Klett-Cotta, Stuttgart, Germany

Kroeber A L 1963 *An Anthropologist Looks at History*. University of California Press, Berkeley and Los Angeles

Kroeber A L, Kluckhohn C 1952 *Culture: A Critical Review of Concepts and Definitions*. The Peabody Museum, Cambridge, MA

Le Goff J 1965 *La Civilisation au l'occident médieval*. Arthaud, Paris

Lewis M W, Wigen K E 1997 *The Myth of Continents. A Critique of Metageography*. University of California Press, Berkeley and Los Angeles

Marrou H I 1938 Culture, civilization, décadence. *Revue de Synthèse*, **December**: 133

Ratzel F 1895 *Anthropogeographische Beitrage*. Duncker und Humblot, Leipzig, Germany

Rüsen J, Gottlob M, Mittag A (eds.) 1998 *Die Vielfalt der Kulturen. Erinnerung, Geschichte, Identität, 4*. Suhrkamp, Frankfurt am Main, Germany

Savarkar V D 1969 *Hindutva: Who is a Hindu?* 5th edn. Veer Savarkar Parakashan, Bombay, India

Schlanger J E 1971 *Les Métaphores de l'Organisme*. Vrin, Paris

Spengler O 1922 *Der Untergang des Abendlandes*. 2 Vols. W. Braumüller, Vienna and Leipzig

Tainter J A (ed.) 1988 *The Collapse of Complex Societies*. Cambridge University Press, Cambridge, UK

Toynbee A 1934–1959 *A Study of History*. 12 Vols. Oxford University Press, London

Troeltsch E 1920 Der Aufbau der europäischen Kulturgeschichte. *Schmollers Jahrbuch für Gesetzgebung, Verwaltung, und Volkswirtschaft im Deutschen Reiche* **44**(3): 1–48

Tylor E B 1958 *Primitive Culture*. 2 Vols. Harper, New York

Zureiq K 1969 *Nahnu wa't-tarikh*. Dar al- Ilm lil-Malayin, Beirut, Lebanon

A. Al-Azmeh

Civilizational Analysis, History of

The term 'civilizational analysis' is used here to describe a whole cluster of traditions, rather than a specific theoretical perspective. The shared theme is a plurality of fundamental and comprehensive socio-cultural patterns, seen as sufficiently different from

each other to justify the idea of civilizations in the plural, and in contradistinction to civilization in the singular. This is the conceptual framework preferred by some of the most seminal theorists in the field. This is predominant in contemporary debates, although authors who opted for a different terminology must be included in the present survey.

1. *Origins and Directions of Civilizational Analysis*

1.1 *General Characteristics*

In the context of a broader history of social thought, the civilizational approach emerges as a counterpoint and a potential corrective to mainstream modes of theorizing. It is only in the most recent phase that one can speak of progress towards it becoming a fully-fledged alternative. The most obvious contrast has to do with visions of history: civilizational lines of interpretation stress the plurality of historical trajectories and contest the claims of general evolutionistic theories. For the same reason, the emphasis on original and mutually irreducible cultural configurations runs counter to functionalist postulates of universal constraints or imperatives. This need not mean an outright rejection of any common frame of reference, but its explanatory scope is at best limited.

Civilizational analysis deals with units of larger dimensions and longer duration than the single societies that they encompass, and this focus of interest leads it to question the self-contained image of society (and the underlying fixation on the nation state) that has been central to the sociological tradition. The large-scale and long-term patterns in question can be conceptualized in various ways, and different approaches may be reflected in more or less developed typologies of civilizational formations. In some cases, the organic metaphors commonly linked to functionalist images of society reappear on the civilizational level, but they should not be mistaken for a defining characteristic of civilizational thought.

If civilizational analysis is defined in these broad terms, it does not emphasize any particular type of intercivilizational relations at the expense of others. The historical experience to be analyzed includes closures, encounters, and conflicts. In different situations, some of these patterns of interaction become more salient. Different theoretical approaches may entail correspondingly selective views of the intercivilizational field. For example, debates on this topic have been marked by a disproportionate emphasis on civilizational conflict.

Finally, the civilizational perspective may serve to highlight macrohistorical continuities as well as major ruptures. Overall, contemporary historians have been more sensitive to civilizations as phenomena of the *longue durée* (Braudel is the prime example), but comparative sociologists (such as S. N. Eisenstadt) have been no less interested in the civilizational dynamics set in motion by major cultural breakthroughs.

1.2 *Early Developments and Nineteenth-century Reversals*

The explicit idea of civilizations in the plural seems to have grown out of the same historical developments as that of civilization in the singular (Starobinski 1983), although it took longer for the plural to be codified for official usage. Both are eighteenth-century responses to internal transformations of the West as well as to encounters with non-Western societies and traditions. However, the pluralistic view was much less clearly articulated and became more marginal as the new phase of Western expansion from the early nineteenth century onwards seemed to herald a triumph of civilization in the singular.

In retrospect, elements or anticipations of civilizational analysis can be found in the writings of earlier authors. Giambattista Vico's *New Science* (definitive Italian edition, 1744; for an English translation, see Vico 1968) is perhaps the most frequently invoked example. Those who regard Vico as a pioneer stress his interest in 'the *differences* as well as the similarities in the patterns and paces of social and cultural processes and structures *across histories*' (Nelson 1976, p. 875). But it can be objected that the latent civilizational perspective remains subordinate to an exclusive focus on European cycles of rise and decline, seen against the background of Greek, Roman, and Jewish sources. The most systematic recent interpretation of Vico's work (Lilla 1993) suggests that the whole comparative inquiry is only a sideline to a traditionalist critique of modernity.

Another ambiguous precursor is Montesquieu, whose *Spirit of the Laws* (first published, 1748; for an English translation, see Montesquieu 1989) linked the analysis of legal and political regimes to 'customs' and drew on new knowledge of Asian civilizations (especially China). But the project as a whole is still centered on the traditional problematic of political philosophy and its implications for a reformist response to absolutism.

Eighteenth-century encounters with Asia were widely and variously reflected in European thought and literature. The attitude of the educated public was less prejudiced than it later became, and the 'intuition about the equal value of cultures,' which Charles Taylor (1990) has identified as a persistent but often subordinate theme in Western thought, was more in evidence. A shift to markedly more Eurocentric approaches took place around 1800 (Osterhammel 1998). The potential openings to civilizational analysis, which we can attribute to the Enlightenment in

retrospect, were thus blocked by contrary trends before they could translate into more lasting results.

If some progress was made towards the articulation of a more pluralistic approach, it was more directly linked to the concept of culture than to the ideas of civilizations in the plural. In Herder's *Ideas for a Philosophy of the History of Mankind* (first published, 1784–91; for an English translation, see Herder 1968), there is no reference to cultures in the plural, and culture in the singular remains closely associated with progress and enlightenment. But the emphasis on the singularity of each people gives a pluralizing twist to the model. Cultural particularities are, however, thematized in a way which proved much more important to the history of nationalist thought than for the civilizational approach to human diversity.

The mainstream of nineteenth-century social thought was uncongenial to civilizational analysis. This was the case for the theories drawing on the legacy of German idealism as well as the positivistic conceptions of general social evolution, reinforced by (but not derived from) the Darwinian discovery of biological evolution. The Marxist tradition may be seen as a meeting ground for ideas from these two sources. On both sides, a unilinear vision of history and a universalistic model of development set strict limits to the recognition of cultural plurality.

Contrary to some recent suggestions, it does not seem justified to include Hegel among the pioneers of civilizational analysis. His plurality of collective spirits is only a derivative aspect of the progress of the world spirit throughout universal history, although he undeniably made a certain effort to grasp the individuality of the cultures (including China and India) that he consigned to lower levels of the unfolding design. Similarly, Marx's scattered comments on the Asiatic world and its deviation from the Western pattern of development do not go beyond a general contrast between progress and stagnation (explained in terms of fundamental economic structures). But his well-known analysis of India under British rule reflects some interest in the characteristics of a markedly alien civilization.

2. *The Classics of Civilizational Analysis: Sociologists and Metahistorians*

Major contributions to civilizational analysis—on theoretical as well as substantive levels—were made during the most formative period of modern sociological thought (roughly between 1890 and 1920). But the new perspectives were not developed to the same degree as the ideas that became more central to sociological theory. In the absence of systematic sociological inquiry, the comparative study of civilizations was pursued along very different lines by writers who defied the conventional academic division of labor.

2.1 *Durkheim, Mauss, Weber: The Sociological Discovery of Civilizations*

A clearly defined sociological concept of civilizations in the plural appears for the first time in a short text by Durkheim and Mauss, little noticed at the time and long neglected by the most influential interpreters of the Durkheimian tradition, but more recently rediscovered by theorists who have tried to reactivate the civilizational perspective. The *Note on the notion of civilization* was first published in 1913 (for an English translation with a commentary by Benjamin Nelson, see Durkheim and Mauss 1971) and is obviously related to arguments developed in other writings during the last phase of Durkheim's work, such as the *Elementary Forms of the Religious Life*. The common theme is a perceived need to go beyond the concept of society formulated in earlier works (now seen as too close to the idea of a given system of structures and functions) and to explore various ways of doing so.

The particular insight provided by the notion of civilizations in the plural has to do with large-scale units and long-term processes which encompass multiple societies. Civilizations 'reach beyond the national territory' and 'develop over periods of time that exceed the history of a single society'; they constitute 'a moral milieu encompassing a certain number of nations' or 'a plurality of interrelated political bodies acting upon one another' (Durkheim and Mauss 1971, pp. 810–11). The emphasis is on the cultural unity of civilizational complexes ('moral milieu' is obviously to be understood in a broad sense), and political plurality appears as a normal rather than a problematic condition. But there is no *a priori* classification of unifying and fragmenting factors: Durkheim and Mauss call for a comparative study of the civilizational potential inherent in various categories of social phenomena: 'the unequal coefficient of expansion and internationalization (Durkheim and Mauss 1971, p. 811). Although the connection is never made explicit, this variety can also be seen as a matter of the forms and degrees of social creativity, and thus related to a theme which figures prominently in other texts.

Mauss returned to the problematic of civilizations in a later debate with Lucien Febvre and others. A specific civilization is, as he put it, a 'family of societies' or a 'hyper-social system of social systems' (Mauss 1930, p. 89; the term 'system' is evidently not used in a very rigid sense). He proposed a more advanced conceptual framework than in the earlier text: the elements of civilizations—the unequally unified ideas, practices, and products characteristic of a civilizational complex—are distinguished from their forms, i.e., the patterns that grow out of complex combinations of such elements. In addition, the comparative study of civilizations would deal with the characteristics of the areas or regions over which they expand, as well as the interconnections of the societies that belong to them. With regard to the last point,

Mauss adds a passing but potentially far-reaching comment: societies may 'singularize' themselves and enhance their individual features against a broader civilizational background. The varying outcomes and possible implications of such processes are also a matter for comparative analysis.

Mauss went on to draw an important theoretical conclusion. The plurality of civilizations is the most striking case of a characteristic common to all forms of social life: their arbitrariness, or—in other words—the formative role of collective choices embodied in more or less coherent patterns (Mauss 1930, p. 97). Once again, the civilizational perspective brings out the theme of social creativity.

Mauss seemed to have thought of this interpretive model as equally applicable to 'primitive' and 'civilized' societies (if anything, the anthropologists had been quicker to recognize civilizational phenomena than the sociologists). However, it is clearly more attuned to the latter case (the reference to 'families of societies' with shared cultural horizons and historical traditions is easier to understand in that context). The idea of civilizations in the plural is thus reconnected to civilization in the singular, and the prime objects of comparative analysis would be the advanced civilizational complexes (Hochkulturen) whose dynamics have shaped the course of world history. Neither Durkheim nor Mauss made any significant moves in that direction, and later attempts linked to the Durkheim school (especially Marcel Granet's work on China) did not do much to concretize the theoretical project outlined above.

The most important substantive contribution of classical sociology is to be found in Max Weber's analyses of the contrasts between patterns of development in Western and non-Western civilizational settings. But the absence of explicit conceptual foundations and the lack of a clearly defined research program made it difficult to distinguish the civilizational perspective from more narrowly focused parts of Weber's work. The idea of culture as a distinctive way of lending meaning and significance to the world—outlined in Weber's earlier methodological writings—is never put to systematic or comparative use. Although the major civilizational studies variously refer to 'cultural centers,' 'cultural areas,' and 'cultural worlds,' cultural patterns and contexts are never theorized as such.

The resultant ambiguity of Weber's argument is reflected in later controversies around his work. Forms of rationality and dynamics of rationalization were the main foci of his comparative analyses, and he repeatedly stressed that they were always embedded in specific settings. But when it comes to details, the constitutive frameworks—from overall civilizational configurations to specific sociocultural spheres that crystallize within them in varying ways—are overshadowed by a seemingly uniform—but unequally developed—rationality in progress. Those who set out to reconstruct Weber's problematic have therefore been tempted to ground it in more or less differentiated universal models of rationality, rather than to start with a plurality of cultural patterns and their formative imprints on civilizational complexes.

The civilizational aspect of Weber's project neither was given from the outset nor equally present in all parts of his work. His original and most abiding concern was the combination of multiple and successive developments which had set the Western trajectory apart from all others. The interest in other civilizations developed in this connection and did not lead to totalizing interpretations of non-Western traditions in their own terms or attempts to reorient the comparative strategy from their points of view. None of Weber's comparative studies aspires to do what Louis Dumont later proposed to do through a new interpretation of India. On the other hand, closer consideration of non-Western cases raises questions which go beyond the initial issues, reveals new interconnections and opens up new perspectives on the Western experience.

This shift towards a more comprehensive framework was still in progress in Weber's last writings, but the results varied—both in kind and in degree—from case to case. Weber's most extensive work on the ancient world focuses on socioeconomic structures. The analysis of ancient Judaism is primarily concerned with a religious breakthrough and its long-term rationalizing potential. The studies of China and India (Weber 1920–21, pp. 1–2) deal most extensively with complex and long-term interrelations of cultural traditions and institutional contexts, and are therefore closest to the model of civilizational analysis as defined by Durkheim and Mauss.

2.2 The Other Tradition: from Spengler to Toynbee and Beyond

As shown above, the sociological classics left civilizational analysis in a very inconclusive and fragmentary state: there was no connection between the Durkheimian sketch of a theoretical framework and the Weberian exploration of historical testing grounds, and neither of the two overtures was followed by further work. For several decades after Weber's death, the idea of a comparative study of civilizations was mainly associated with metahistorical projects of the kind exemplified by Oswald Spengler's *Decline of the West* (1923) (for an English translation, see Spengler 1926–28) and Arnold Toynbee's *Study of History* (Toynbee 1934–61). Among later attempts to deal with questions raised by Spengler and Toynbee, Franz Borkenau's posthumously published fragments (Borkenau 1981) deserve special mention.

In general terms, this tradition should be given credit for preserving the insight that the dynamics of

world history involve units of greater size and longer duration than the single (or artificially singled-out) societies more familiar to mainstream scholarship, and that the patterns of formation, flowering and decline of such macro units call for closer examination. Moreover, analyses in this vein have sometimes thrown new light on inter-civilizational relations, despite a tendency to stress the separate and self-contained history of each civilizational domain. Spengler's notion of 'pseudomorphosis' is a case in point. It refers to the impact of dominant cultures in latent decline on those emerging within their orbit; the original example was the transformation of the Near East in the shadow of the Roman Empire, culminating in the rise of Islam, but plausible attempts have been made to generalize the concept.

On the other hand, the Spengler–Toynbee tradition has been beset by major problems. Apart from a general tendency to indulge in speculation far beyond the limits of historical evidence, more specific weaknesses are inherent in the overall pattern. Spengler, Toynbee, and those who followed their lead were —notwithstanding differences in emphasis—inclined to exaggerate the cultural or societal closure of civilizational units. Nevertheless, at the same time, they took a cross-civilizational identity of developmental patterns for granted: a uniform, more or less consistently cyclical model was applied across otherwise rigid boundaries.

When it came to concrete analysis and demarcation of civilizational domains, this line of thought led to a dilemma. If the self-contained units in question are defined based on unique cultural features, claims to cross-cultural understanding and theorizing are thereby undermined. To avoid this self-defeating turn, exemplified by Spengler's work, Toynbee shifted the focus towards civilizations as 'societies,' distinct and durable frameworks of interaction, but found it very difficult to specify the characteristics of a civilizational society. At an advanced stage of his project, he sought to defuse the problem by relegating civilizations to the prehistory of universal religions.

3. *The Renaissance of Civilizational Analysis*

Since the 1970s, several developments have led to a revival of interest in civilizational analysis. The plurality of civilizations has resurfaced as a key theme in the work of prominent historians (Braudel 1994). Projects drawing on anthropological and sociological traditions have moved towards a civilizational per-s-pective; here Louis Dumont's work on India is of particular importance (Dumont 1967). Occasional attempts have been made to bring the problematic legacy of Spengler and Toynbee into the orbit of a historical sociology of civilizations. Various approaches to a shared problematic are represented in the International Society for the Comparative Study of Civilizations, which publishes *Comparative Civilizations Review*. However, when it comes to innovative theoretical conceptions, two projects seem to deserve a somewhat more detailed account.

3.1 *Benjamin Nelson: Orientations and Encounters*

Benjamin Nelson's programmatic outline of a civilizational theory, presented in a series of essays (Nelson 1981) but never developed in a systematic fashion, is inseparable from a reinterpretation of Max Weber's work. As Nelson argued (in the 1960s and 1970s, when more restrictive readings of Weber dominated the field), the agenda most succinctly summarized in the 'Author's Introduction' to the *Protestant Ethic*, centered on the comparative study of civilizational patterns and their historical trajectories. At the same time, Nelson went beyond Weber in thematizing the cultural cores of civilizational complexes. The different clusters of cultural orientations, also equated with 'structures of consciousness,' gave rise to correspondingly different frameworks for the rationality of conduct, social coexistence, and reflexive thought. Nelson took a particular interest in the rationalizing efforts which led to the overcoming of traditional dualisms, such as those of the religious and the mundane life or the insider and the alien, and this enabled him to reinterpret the triangular comparison of China, India, and the West in more focused terms than Weber had done.

But the emphasis on breakthroughs to more inclusive and interactive forms of social life did not lead Nelson to neglect the other side of civilizational dynamics: some of his essays stress the productive potential of internal conflicts at the level of the most basic cultural premises (he refers to them as 'civil wars within the structures of consciousness'). His favorite example was the highly articulate tension between faith and reason in the course of the eleventh- and twelfth-century transformation of Western Christendom.

This crucial phase of European history, much more important for Nelson's genealogy of modernity than it had been for Weber, also exemplifies the decisive role of intercivilizational encounters: the interaction with the Byzantine and Islamic worlds had an epoch making impact on Western ways of life and thought, not least through an unprecedented revival of interest in classical sources. Nelson discussed other encounters, such as the contacts between China and the West, and although his treatment of this problem was in some ways inconclusive, he did more than anybody else to integrate it into the domain of civilizational theory.

3.2 *S. N. Eisenstadt: Breakthroughs and Dynamics*

In contemporary social theory, S. N. Eisenstadt's work stands out as the most sustained exploration of

civilizational themes. His interest in this field grew out of several research projects, but they converged in a critique of the conventional distinction between tradition and modernity. The diversity of modern societies could not be explained without reference to the ongoing formative role of traditions, and the most important factors of that kind have to do with enduring civilizational legacies. A comparative study of empires, designed to distinguish their complex social and political structures from stereotypes of traditional society, raised questions about civilizational backgrounds and their influence on imperial formations. Most importantly, a comparative analysis of modern revolutions (Eisenstadt 1978) took an explicitly civilizational turn: the 'great revolutions' that had come to be seen as paradigms of radical change were based on a more fundamental cultural shift which opened the constitutive visions of social order to dissent, protest, and innovation. The revolutionary cultural foundations of modernity mark it as a new civilization.

After discovering the civilizational dimension from these different angles, Eisenstadt's next step was a closer examination of the historical cases that seemed to have brought it to the fore in the most revealing way. The epoch already described by some earlier authors as an 'Axial Age,' covering a few centuries around the middle of the last millennium BC, was—as Eisenstadt argued—the prime example of a civilizational breakthrough. In some major cultural centers (ancient Greece, ancient Israel, India, and China), radical changes to cultural ontologies gave rise to new images of social order: ideas of a fundamental contrast between transcendental and mundane realities, unknown to more archaic cultures, translated into orderbuilding visions and strategies.

Nevertheless, the expanded scope for imagination, in the interpretive as well as the institutional domain, was also conducive to higher levels of interpretive conflict and ideological rivalry; the elites and coalitions that mobilized the new cultural resources had to confront more or less structured currents of heterodoxy and dissent. Eisenstadt's analyses have highlighted the variety and dynamism of social formations that develop within this framework, although the Axial constellation—defined in the most general sense—seems to have a uniform structure (for a theoretical discussion, accompanied by case studies of major civilizations, see Eisenstadt 1986). Among the offshoots of the Axial transformation, European civilization—shaped by recurrent combinations of diverse sources—became the main center of another civilizational mutation: the transition to modernity. But other civilizational backgrounds left their traces on the specific patterns of modernity that emerged within their orbit.

Both the analysis of Axial civilizations and the general interpretive framework growing out of it are still in progress. The results so far achieved suggest that this is a particularly promising line of inquiry.

4. *Alternative Views*

Among the few constructive responses to the metahistorical tradition discussed in Sect. 2.2, Jaroslav Krejci's work is noteworthy for its scope and ambition (Krejci 1982, 1993). The starting point is a critique of inconsistencies and loose ends in Toynbee's theory. In particular, the incomplete analysis of creative elites as civilization builders and the unclear status of universal religions as supracivilizational units are singled out for reconsideration from a more sociological angle. As Krejci sees it, 'protagonist groups and changing relationships between them (such as the shifting balance between Brahmins and kshatryas in India) play a key role in the construction of civilizations.' But they operate through specific core institutions (empires, states, churches, or ideological communities of various kinds) and are inspired by distinctive—more or less overtly religious—visions of the human condition. Religious traditions are thus reintegrated into the civilizational frame of reference, and their formative role is analyzed in terms of interpretive patterns that lend meaning to life and death.

Krejci distinguishes various 'paradigms of the human predicament,' ranging from the theocentric invented in ancient Mesopotamia to the utilitarian version of the anthropocentric in the modern West. This key to civilizational theory provides an alternative to the more common typologies of social formations, based on the division of labor, and to the a-theoretical use of geographic or historical criteria to demarcate civilizations. However, once the defining anthropological premises have been identified, geographical and historical perspectives can be given their due; it becomes possible to distinguish civilizational areas and sequences.

Arguments developed by Johann P. Arnason (1988, 2001) draw more directly on classical sources. The ideas put forward by Durkheim and Mauss on one side and Weber on the other are seen as incomplete insights to be synthesized, but this would require a closer connection between the Weberian theme of world interpretation and the Durkheimian notion of collective representations. The idea of imaginary significations—introduced by Cornelius Castoriadis—appears as the most suitable basis for such a rapprochement.

Imaginary significations are horizons of meaning, irreducible to experiential foundations as well as to functional constraints or rational principles. Specific clusters of such significations are at the core of different social worlds and structure their relations to nonsocial domains of reality as well as their internal forms of differentiation and integration. On this view, civilizational patterns can be analyzed as the most comprehensive and distinctive constellations of imaginary significations. In that capacity, they give rise to specific ways of being in the world and corresponding types of relationships between the main spheres of

social life. These constitutive frameworks make it possible for civilizations to encompass groups of societies and maintain their identity throughout successive historical phases (for Durkheim and Mauss, integrative capacities manifested in space and time were the defining feature of civilizations). At the most visible level, civilizational formations take the shape of regional configurations, central to the agenda of comparative history.

To speak of civilizations in this sense is not to prejudge the levels of coherence, unity, and consistency. The approach just outlined can allow for significant variations in all these respects. In particular, it may be argued that some civilizations are more markedly characterized by conflicting cultural orientations than others (among the major non-Western traditions, interpretations of India have laid more emphasis on this theme than those of China). Comparative analyses of such differences—and other related ones—would be the most effective antidote against the identitarian and overintegrated models which continue to obstruct the progress of civilizational studies.

See also: Civilization, Concept and History of; Civilizations; Primitive Society; States and Civilizations, Archaeology of

Bibliography

Arnason J P 1988 Social theory and the concept of civilization. *Thesis Eleven* **20**: 87–105
Arnason J P *Civilization and Difference*. Sage, London
Borkenau F 1981 *End and Beginning: On the Generations of Cultures and the Origins of the West*. Columbia University Press, New York
Braudel F 1994 *A History of Civilizations*. Penguin, Harmondsworth, UK
Dumont L 1967 *Homo hierarchicus: essai sur le système des castes*. Gallimard, Paris
Durkheim E, Mauss M 1971 Note on the notion of civilization. *Social Research* **38**(4): 808–13
Eisenstadt S N 1978 *Revolution and the Transformation of Societies: A Comparative Study of Civilizations*. Free Press, New York
Eisenstadt S N (ed.) 1986 *The Origins and Diversity of Axial Civilizations*. SUNY Press, Albany, NY
Herder J G 1968 *Reflections on the Philosophy of the History of Mankind*. Chicago University Press, Chicago
Krejci J 1982 Civilization and religion. *Religion* **12**: 29–47
Krejci J 1993 *The Human Predicament: Its Changing Image: A Study in Comparative Religion and History*. St. Martin's Press, New York
Lilla M 1993 *G.B. Vico: The Making of an Anti-modern*. Harvard University Press, Cambridge, MA
Mauss M 1930 Les civilisations: Elements et formes. In: Febvre L et al. (eds.) *Civilisation: Le mot et l'idee*. La Renaissance du Livre, Paris
Montesquieu C 1989 *The Spirit of the Laws*. Cambridge University Press, Cambridge, UK
Nelson B 1976 Vico and comparative historical civilizational sociology. *Social Research* **43**(4): 874–81
Nelson B 1981 *On the Roads to Modernity: Conscience, Science and Civilizations*. Rowman and Littlefield, Totowa
Osterhammel J 1998 *Die Entzauberung Asiens: Europa und die asiatischen Reiche im 18*. Beck Verlag, München, Germany
Spengler O 1926–28 *The Decline of the West*. Knopf, New York, Vols. 1–2
Starobinski J 1983 Le mot 'civilisation.'*Le temps de la reflexion* **4**: 13–52
Taylor C 1990 Comparison, history, truth. In: Reynolds F, Tracey D (eds.) *Myth and Philosophy*. SUNY Press, Albany, NY
Toynbee A J 1934–61 *A Study of History*. Oxford University Press, Oxford, UK, Vols. 1–12
Vico G B 1968 *The New Science of Giambattista Vico*. Cornell University Press, Ithaca, NY
Weber M 1920–21 *Gesammelte Aufsätze zur Religionssoziologie*, Bd. 1–3. Mohr Verlag, Tübingen, Germany

J. P. Arnason

Civilizations

1. Introduction

The term 'civilization' has been used in modern social science and historical literature in several different ways. One such way, developed above all in Germany from about the end of the nineteenth century through the period up to World War II and perhaps best represented by scholars like Alfred Weber (and taken over to a certain extent in the English-speaking world by R. M. McIver), designated 'civilization'—as distinct from 'society' and above all from 'culture'—as encompassing above all the technological, material factors and to some extent organizational aspects of social life as against the deeper, more 'spiritual' cultural and aesthetic ones.

Another designation of the term, made famous by Norbert Elias in his *Über den Prozess der Zivilisation* (1939) focused on the 'socializing' process through which the image of the civilized person, as constructed in the courtly and also early bourgeois society in Europe, was promulgated and institutionalized. This designation of civilization was related to an earlier one rooted in the French Enlightenment, in which civilization was seen as the opposite of barbarism. However, in later works by Elias's followers, for instance Goudsblom, this view of civilization was extended to cover many other societies and historical periods, going back even to the impact of the presumably first domestication of fire.

The third and most extensive designation of civilization was promulgated by scholars such as Max Weber,

Emile Durkheim, Oswald Spengler, Pitirim Sorokin, Arnold Toynbee, A.L. Kroeber, Carroll Quigley, Cristopher Dawson, Fernand Braudel, William H. McNeill, Adda Bozeman, or Immanuel Wallerstein, and lately very forcefully by Samuel Huntington. However great the differences in perspective, methodology, focus, and concepts that pervade the works of these scholars, they share the use of the term civilizations as distinct societal-cultural units which share some very important, above all cultural, characteristics. Here we shall use the term civilization in a way very close to, but also distinct from, such a designation.

Civilization as combination of ontological or cosmological visions, of conceptions of trans-mundane and mundane reality, with the definition, construction, and regulation of the major arenas of social life and interaction

The central analytical core of the term civilization as employed here—as distinct from such social formations as political regimes, different forms of political economy or collectivities like 'tribes,' ethnic groups or nations, and from religion or cultural traditions—is the combination of ontological or cosmological visions, of visions of trans-mundane and mundane reality, with the definition, construction, and regulation of the major arenas of social life and interaction.

The central core of civilizations is the symbolic and institutional inter-relation between the formulation, promulgation, articulation, and continuous reinterpretation of the basic ontological visions prevalent in a society, its basic ideological premises and core symbols on the one hand, and the definition and regulation of major arenas of institutional life on the other. Such definitions and regulations construct the broad contours, boundaries, and meanings of the major institutional formations and their legitimization, and greatly influence their organization and dynamics.

The impact of such ontological visions and premises on institutional formation is effected through the processes of interaction and control that develop in a society. Such processes of control—and the opposition to them—are not limited to the exercise of power in the 'narrow' political sense. Rather, they are activated by major elites in a society. The most important such elite groups are the political, the cultural, and the economic ones and those which construct the solidarity and collective images of the major groups, all of which have different cultural visions and represent different interests.

The structure of such elite groups is closely related, on the one hand, to the basic cultural orientations prevalent in a society; that is, different types of elite groups bear different orientations or visions. On the other hand, and in connection with the types of cultural orientations and their respective transformation into basic premises of the social order, these elite groups tend to exercise different modes of control over the allocation of basic resources.

Such combination of ontological visions and of structuration of institutional formations and collective identities constitutes an inherent component of the formation of any society, and is always closely interwoven with the more organizational aspect of any institutional formation—political, economic, or family and kinship.

The very implementation or institutionalization of such premises and the concomitant formation of institutional patterns through processes of control, symbolic and organizational alike, also generate tendencies to conflict and change. The crystallization of these potentialities of change usually takes place through the activities of secondary elite groups who attempt to mobilize various groups and resources to change aspects of the social order.

The full development of the distinct ideological and institutional dimensions, and of some awareness of their distinctiveness, has emerged in some very specific historical settings—namely, the so-called Axial Civilizations—even if some very important kernels thereof can be identified in some archaic civilizations such as those of ancient Egypt, Assyria, or Mesoamerica.

2. *Axial Age Civilizations: the Reconstruction of the World and the Crystallization of Distinct Civilizational Complexes*

By Axial Age civilizations (to use Karl Jaspers' nomenclature), we mean those civilizations that crystallized during the thousand years from 500 BC to the first century of the Christian era, within which new types of ontological visions, of conceptions of a basic tension between the transcendental and mundane orders, emerged and were institutionalized in many parts of the world. Examples of this include ancient Israel; later in Second-Commonwealth Judaism and Christianity; Ancient Greece; possibly Zoroastrianism in Iran; early imperial China; Hinduism and Buddhism; and, beyond the Axial Age proper, Islam.

The crystallization of these civilizations constitutes a series of some of the greatest revolutionary breakthroughs in human history, which have shaped contours of human history in the last two to three millennia. The central aspect of these breakthroughs was the emergence and institutionalization of new ontological metaphysical conceptions of a chasm between the transcendental and mundane orders.

The development and institutionalization of these ontological conceptions entailed the perception of the given mundane order as incomplete, inferior—oftentimes as evil and polluted. It gave rise in all these civilizations to attempts to reconstruct the mundane

world, from the human personality to the sociopolitical and economic order, according to the appropriate 'higher' transcendental vision.

The revolutionary conceptions, which first developed among small groups of autonomous, relatively unattached 'intellectuals' (a new social element at the time), were ultimately transformed into the basic 'hegemonic' premises of their respective civilizations, and were subsequently institutionalized. That is, they became the predominant orientations of both the ruling elites and of many secondary elites, fully embodied in the centers or subcenters of their respective societies.

One of the most important manifestations of such attempts in all these civilizations was the strong tendency to construct societal centers to serve as the major autonomous and symbolically distinct embodiments of respective ontological visions, as the major loci of the charismatic dimension of human existence. But at the same time the 'givenness' of the centers could not necessarily be taken for granted. The construction and characteristics of the center tended to become central issues under the gaze of the increasing reflexivity which focused above all on the relations between the transcendental and mundane orders. The political dimension of such reflexivity was rooted in the transformed conceptions of the political arena and of the accountability of rulers. The political order as one of the central loci of the mundane order had to be restructured according to the precepts of the transcendental visions. The rulers were usually held responsible for organizing the political order according to such precepts.

At the same time the nature of rulers became greatly transformed. The king-god, embodiment of the cosmic and earthly order alike, disappeared, and a secular even if often semisacral ruler appeared. Thus there emerged the conception of the accountability of rulers and community to a higher authority, God, Divine Law, or a metaphysical vision. Accordingly, the possibility of calling a ruler to judgement appeared. One such dramatic appearance of this conception occurred in ancient Israel, in the priestly and prophetic pronunciations. 'Secular' conceptions of such accountability to the community and its laws appeared in both the northern shores of the eastern Mediterranean, in ancient Greece, as well as in the Chinese conception of the Mandate of Heaven.

Concomitantly with the emergence of these conceptions of accountability of rulers, autonomous spheres of law began to develop as somewhat distinct from purely customary law. Such developments could also entail some beginnings of a conception of rights even if the scope of these spheres of law and rights varied greatly. Of special importance from the point of view of our analysis is the fact that one of the most important manifestations of the attempts to reconstruct the social order was the strong tendency to define certain collectivities and institutional arenas as most appropriate for the implementation of transcendental visions.

3. *Autonomous Elites as Bearers of Civilizational Visions: Change, Protest, and Heterodoxies*

The development of new ontological metaphysical conceptions in the Axial civilization was closely connected with the emergence of a new type of elite, carriers of models of cultural and social order. These were often autonomous intellectuals, such as the ancient Israelite prophets and priests, and later on the Jewish sages, the Greek philosophers and sophists, the Chinese literati, the Hindu Brahmins, the Buddhist Sangha, the Islamic Ulema. Initial small nuclei of such groups of cultural elites developed the new ontologies, the new transcendental visions and conceptions, and were of crucial importance in the construction of the new 'civilizational' collectivities.

The new type of elites differed greatly from the ritual, magical, and sacral specialists in the pre-Axial Age civilizations. They were recruited and legitimized according to autonomous criteria, and were organized in settings distinct from those of the basic ascriptive political units of the society. The acquired a potentially countrywide and also trans-country status of their own. They also tended to become potentially independent of other categories of elites, social groups, and sectors.

At the same time a far-reaching transformation of other elites, such as political elites, took place. All these elites saw themselves not only as performing specific technical activities—be they those of scribes, ritual specialists, and the like—but also as potentially autonomous carriers of a distinct order related to the prevalent transcendental vision. They saw themselves as the autonomous articulators of the new order, and rival elites as both accountable to them and as essentially inferior. Moreover, each of these groups of elites was not homogeneous, and within each of them a multiplicity of secondary influentials developed.

These new groups became transformed into relatively autonomous partners in the major ruling coalitions. They also constituted the most active elements in the movements of protest and processes of change that developed in these societies and which evinced some very distinct characteristics at both symbolic and organizational levels.

First, there was a growing symbolic articulation and ideologization of the perennial themes of protest found in any human society, such as rebellion against the constraints of division of labor, authority, and hierarchy, and of the structuring of time dimension, the quest for solidarity and equality, and for overcoming human mortality.

Second, utopian orientations were incorporated into the rituals of rebellion and the double image of society. It was this incorporation that generated

alternative conceptions of social order and new ways of bridging the distance between the existing and the 'true' resolution of the transcendental tension.

Third, new types of protest movements appeared. The most important were intellectual heterodoxies, sects, or movements which upheld the different conceptions of the resolution of the tension between the transcendental and the mundane order, and of the proper way to institutionalize such concepts. Since then, continuous confrontation between orthodoxy on the one hand, and schism and heterodoxy on the other, has been a crucial component in the history of mankind.

Fourth, and closely related to the former, was the possibility of the development of autonomous political movements and ideologies usually oriented against an existing political but possibly also religious center.

All these developments ushered into the arena of human history the possibility of the conscious ordering of society, and also the continuous tension that this possibility generated. The new dynamics of civilization transformed group conflicts into potential class and ideological conflicts, cult conflicts into struggles between the orthodox and the heterodox. Conflicts between tribes and societies could become missionary crusades. The zeal for reorganization informed by each civilization's transcendental vision made the entire world at least potentially subject to cultural-political reconstruction.

4. *The Expansion of Axial Civilizations*

Concomitantly with the institutionalization of Axial civilizations, a new type of intersocietal and intercivilizational world history emerged. To be sure, political and economic interconnection have existed between societies throughout human history. Some conceptions of a universal kingdom emerged in many post-Axial civilizations, like that of Genghis Khan, and many cultural interconnections developed between them, but only with the institutionalization of Axial civilizations did a more distinctive ideological and reflexive mode of expansion develop, with potentially strong semimissionary orientations.

It was indeed in close connection with the Axial civilizations' tendency to expansion that there developed new 'civilizational' collectivities, distinct from political and from 'primordial' ones, yet impinging on them, continuously challenging them, and provoking continual reconstruction of their respective collective identities. Such processes were effected by the interaction between the new autonomous cultural elites and the various carriers of solidarity and political elites of the different continually reconstructed 'local' and political communities.

In the continuous encounter of Axial civilizations with non-Axial or pre-Axial civilizations it was usually the Axial that came out victorious, without however necessarily obliterating many of the symbolic and institutional features of the non-Axial. The most important case of an encounter of non-Axial with Axial civilization in which the former absorbed the latter has been Japan.

5. *The Multiplicity of Axial Civilizations and World Histories*

The general tendency to reconstruct the world and to expand was common to all the post-Axial age civilizations. But the concrete implementation varied greatly. There emerged a multiplicity of different, divergent, yet mutually impinging world civilizations, each attempting to reconstruct the world in its own mode, and either to absorb the others or to segregate itself from them.

Two sets of conditions were of special importance in shaping these different modes of institutional creativity and expansion. One such set are variations in the basic cultural orientations. The other is the concrete structure of the social arenas in which these institutional tendencies can be played out.

Among the different cultural orientations the most important have been differences in the very definition of the tension between the transcendental and mundane orders and the modes of resolving this tension. There is the distinction between the definition of this tension in relatively secular terms (as in Confucianism and classical Chinese belief systems and, in a somewhat different way, in the Greek and Roman worlds) and those cases in which the tension was conceived in terms of a religious hiatus (as in the great monotheistic religions and Hinduism and Buddhism).

A second distinction is that between the monotheistic religions in which there was a concept of God standing outside the Universe and potentially guiding it, and those systems, like Hinduism and Buddhism, in which the transcendental, cosmic system was conceived in impersonal, almost metaphysical terms, and in a state of continuous existential tension with the mundane system. The 'secular' conception of this tension was connected, as in China and to some degree in the ancient world, with an almost wholly this-worldly conception of salvation.

A third major distinction refers to the focus of the resolution of the transcendental tensions. Here the contrast is between purely this-worldly, purely other-worldly, and mixed this- and other-worldly conceptions of salvation. The metaphysical nondeistic conception of this tension, as in Hinduism and Buddhism, tends towards an other-worldly conception of salvation, while the great monotheistic religions emphasize different combinations of this- and other-worldly conceptions of the transcendental vision.

Another set of cultural orientations which influenced the expansion of the various Axial civilizations was the extent to which the access to their centers and

major attributes of the sacred within them was open to all members of the community or was mediated by specific institutions.

In addition, there are differences in the way in which relations between the attributes of cosmic and social order of civilizational collectivities and those of the major primordial ascriptive collectivities are conceived—the extent to which there is a disjunction between the two, to which these respective attributes are mutually relevant, each serving as a referent of the other.

But the concrete working out of all such tendencies depends on the second set of conditions—namely the arenas for their concretization. These conditions included, first, the respective concrete economic political-ecological settings, whether they were small or great societies, whether they were societies with continuous compact boundaries, or with cross-cutting and flexible ones. Second was the specific historical experience of these civilizations—especially in terms of mutual penetration, conquest, or colonization.

6. *Internal Transformation of the Axial Civilization: Secondary Breakthroughs and the Crystallization of Modern Civilization*

One of the most important aspects of the dynamics of Axial civilizations was the possibility of development within them of internal transformation, of what has been designated as secondary breakthroughs, the most important illustrations of which have been Second Temple Judaism and Christianity; later Islam, Buddhism, and to a lesser extent Neo-Confucianism, all of which developed out of heterodox potentialities inherent in the respective 'original' Axial civilizations.

But the most dramatic transformation from within one of the Axial civilizations has probably been that of modernity as it first emerged in Western Europe and as it expanded—encompassing most parts of the world, giving rise to development of multiple, continually changing modernities.

The cultural and political program of modernity constituted in many ways a sectarian heterodox breakthrough in the West and Central European Christian Axial civilization. Such transformation took place through the Reformation and in the Great Revolutions, in which there developed very strong emphasis on the bringing together of the City of God and the City of Man. It was in these revolutions that sectarian activities were taken out from marginal or segregated sectors of society and became interwoven not only with rebellions, popular uprisings, movements of protest but also with the political struggle at the center, and were transposed into the general political movements and the centers thereof.

It was above all in the French Revolution that the fully secular transformation of the sectarian antinomian orientation with strong Gnostic components took place. It was epitomized in the Jacobin orientations which became a central component of the modern political program—to reappear yet again forcefully, as Alain Besançon has shown, in the Russian Revolution, and later in the Chinese and Vietnamese revolutions.

The strong sectarian roots of modernity and of the tensions between totalistic Jacobin and pluralistic orientations which developed in Europe find very strong resonance in the utopian sectarian traditions of the Axial civilizations. It is also the religious roots of the modern political program that explain the specific modern characteristics of what may be seen as the most antimodern contemporary movements—namely the various fundamentalist movements which, contrary to the view which defines them as traditional, constitute a new type of Jacobin movement constructing tradition as a totalistic ideology.

7. *The Cultural and Political Program of Modernity: Premises and Antinomies*

The cultural and political program of modernity, as it crystallized first in Western Europe from around the seventeenth century, was rooted in the premises of the European civilization and European historical experience and bore these imprints—but at the same time it was presented and perceived as being of universal validity and bearing.

The radical innovation of this cultural program as it developed in Europe lay first in the 'naturalization' of man, society, and nature; second in the promulgation of the autonomy and potential supremacy of reason in the exploration and even shaping of the world; and third the emphasis on the autonomy of man, of his reason and/or will.

In connection with these orientations there took place far-reaching transformations of the symbolism and structure of modern political centers, as compared with their predecessors in Europe or with the centers of other civilizations. The crux of this transformation was first the charismatization of the political centers as the bearers of the transcendental vision of the cultural program of modernity; second the development of continual tendencies to permeation of the peripheries by the centers and of the impingement of the peripheries on the centers, of the concomitant blurring of the distinctions between center and periphery; and third was the incorporation of themes of protest as basic, legitimate components of the premises of these centers. These themes became central components of the project of emancipation—a project which sought to combine equality and freedom, justice and autonomy, solidarity and identity of modern political discourse and practice. The program also entailed a distinctive mode of the construction of the boundaries of collective identities. Such identities were not taken as preordained by some transcendental authority, but

continually constructed and continually problematized, becoming also foci of political struggle by national and ethnic movements.

The civilization of modernity as it developed in the West was from its very beginning beset by internal contradictions, giving rise to critical discourse which focused on the tensions and contradictions between its premises, and between these premises and institutional development. The most important such tensions in this program were first that between totalizing and more pluralistic conceptions of its major components—of the very conception of reason and its place in human life and society, and of the construction of nature, of human society and its history; second, between reflexivity and active construction of nature and society; third, those between different evaluations of major dimensions of human experience; and fourth between control and autonomy.

These basic tensions, contradictions, and antinomies inherent in the cultural program of modernity were continually played and worked out in major institutional arenas.

8. *Continually Changing Multiple Modernities*

It was out of the conjunction of these cultural orientations with the development of market, commercial, and industrial economies; with the crystallization of a new political and state order; and with military and imperialist expansion, that the civilization of modernity emerged. Its crystallization and expansion were not unlike those of the expansion of all historical civilizations. What was new was first that the great technological advances and the dynamics of modern economic and political forces made this expansion, the changes and developments attendant on them and their impact on the societies to which it expanded much more intensive. The expansion, through the use of military, political, and economic forces, of modern civilization which took place first in Europe and then beyond it continually combined economic, political, and ideological aspects and forces, and its impact on the societies to which it expanded was much more intense than in most historical cases. It spawned a tendency—rather new and practically unique in the history of mankind—to the development of universal, worldwide institutional, cultural, and ideological frameworks and systems. Yet all of these frameworks were multicentered and heterogeneous, each generating its own dynamics.

Of special importance in this context was the relative place of the non-Western societies in the various—economic, political, ideological—international systems which differed greatly from that of the Western ones. It was not only that it was Western societies which were the 'originators' of this new civilization. Beyond this and above all was the fact that the expansion of these systems, especially insofar as it took place through colonialization and imperialist expansion, gave to the Western institutions the hegemonic place in these systems. Yet it was in the nature of these international systems that they generated a dynamics which gave rise both to political and ideological challenges to existing hegemonies, as well as to continual shifts in the loci of hegemony within Europe, from Europe to the United States, then also to Japan and East Asia.

But it was not only the economic, military-political, and ideological expansion of modernity from the West throughout the world that was important in this process. Of no lesser significance was the fact that this expansion has given rise to continual confrontation between the cultural and institutional premises of Western modernity and those of other civilizations. Thus, while the spread or expansion of modernity has indeed taken place throughout most of the world, it did not give rise to just one civilization, one pattern of ideological and institutional response, but to at least several basic variants—and to continual refracting thereof.

Consequently, multiple modernities have emerged. These civilizations, which share many common components and which continually constitute mutual reference points, have been continually unfolding, giving rise to new problematiques and reinterpretations of the basic premises of modernity. All these attest to the growing diversification of the visions and understanding of modernity, of the basic cultural agendas of different sectors of modern societies—far beyond the hegemonic vision of modernity that was prevalent before. The fundamentalist—and the new communal-national—movements constitute one of such new developments, in the unfolding of the potentialities and antinomies of modernity.

Such developments may indeed also give rise to highly confrontational stances—especially to the West—but these stances are promulgated in changing modern idioms, and they may entail a continual transformation of the cultural programs of modernity. At the same time the new diversity was closely connected—perhaps paradoxically—with the development of new multiple common reference points, and with a globalization of cultural networks and channels of communication far beyond what existed before.

See also: Civilization, Concept and History of; Civilizational Analysis, History of; Cultural History; Cultural Psychology; Elias, Norbert (1897–1990); Elites: Sociological Aspects; Hegemony: Cultural; National Character; Societies, Types of

Bibliography

Breuer B 1994 Kulturen der Achsenzeit. Leistung und Grenzen eines geschichtsphilosophischen Konzepts. *Saeculu* **45**: 1–33

Durkheim E, Mauss M 1971 Note on the notion of civilization. *Social Researc* **38**: 808–13
Elias E 1939 *Über den Prozess der Zivilisation*. Haus zum Falken, Basel
Eisenstadt S N 1973 *Tradition, Change and Modernity*. John Wiley & Sons, New York
Eisenstadt S N 1982 The axial age: the emergence of transcendental visions and rise of clerics. *European Sociology* **23**(2): 294–314
Eisenstadt S N (ed.) 1986 *The Origins and Diversity of Axial Age Civilizations*. State University of New York Press, Albany, New York
Eisenstadt S N 1982 Heterodoxies and dynamics of civilizations. *Diogene* **120**: 3–25
Eisenstadt S N, Achlama R 1992 *Kulturen der Achsenzeit II, ihre institutionelle und kulturelle Dynamik*. Suhrkamp, Frankfurt am Main, 3 Vols.
Eisenstadt S N 1996 *Japanese Civilization: A Comparative View*. University of Chicago Press, Chicago
Eisenstadt S N 1999 *Fundamentalism, Sectarianism and Revolution. The Jacobin Dimension of Modernity*. Cambridge University Press, Cambridge, UK
Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon and Schuster, New York
Kroeber A L, Kluckhohn C 1952 *Culture: A Critical Review of Concepts and Definitions*. The Museum, Cambridge, MA
MacIver R M 1931 *Society: Its Structure and Changes*. R. Long & R. R. Smith, Inc. New York
Melko M 1969 *The Nature of Civilizations*. Porter Sargent, Boston
Nelson B 1981 *On the Roads to Modernity*. Rowman and Littlefield, Totowa, NY
Ogburn W F 1922 *Social Change: With Respect to Culture and Original Nature*. Huebsch, New York
Schluchter W 1979 *Die Entwicklung des okzidentalen Rationalismus—Eine Analyse von Max Webers Gesellschaftsgeschichte*. Mohr, Tubingen
Schluchter W 1981 (1985) *The Rise of Western Rationalism, Max Weber's Developmental History*. University of California Press, Berkeley
Schluchter W 1989 *Rationalism, Religion and Domination. A Weberian Perspective*. University of California Press, Berkeley
Weber A 1921 Prinzipielles zur Kultursoziologie. *Archiv für Sozialwissenschaft und Sozialpoliti* **XL**(VII): 1–49, J-C-B. Mohr, Tubingen
Weber A 1931 Kultursoziologie. In: Enke F (ed.) *Handwörterbuch der Soziologie*. Alfred Vierkandt, Stuttgart, pp. 284–94
Weber M 1922–3 *Gesammelte Aufsätze zur Religionssoziologie*. Mohr, Tübingen

S. N. Eisenstadt

Clan

The modern, now widely accepted, definition of the clan, as a group of persons who believe themselves to be related by unilineal descent but who are unable to trace genealogical connections linking all members of the group, has emerged out of a complex intellectual history stretching back to the Enlightenment. Many social theorists of the eighteenth and nineteenth centuries who were concerned to understand the origins of the state and democracy favored evolutionist explanations. Some viewed the family as the original form of society which, over the course of human history, had aggregated into progressively larger and more complex kinship-based groupings of clans and tribes, which were transformed eventually into the territorially-based political formation of the state. Drawing on the work of Barthold Niebuhr (1828), George Grote (1851) and other historians of Greece and Rome, social theorists such as Sir Henry Maine, Numa Fustel de Coulanges, and Lewis Henry Morgan used the ancient Greek and Roman kinship groups, the *genos* and the *gens*, as models upon which they based their understanding of early stages of social evolution.

Maine's (1861) analysis of the evolution of legal systems, using the example of the Roman *gens* and the concept of agnation—that is, kinship traced through exclusively male links—posited a primitive stage of social development in which membership in patriarchal kinship groups defined a person's social status. Maine emphasized that the *gens* was a corporate group which had a legal personality that endured beyond the lives of its individual members, an insight which has been very influential in subsequent work on clans and other forms of descent grouping.

In contrast to Maine's theory of primitive patriarchy, John Ferguson McLennan (1865) and Lewis Henry Morgan (1877) argued that the early stage of social evolution was characterized by group marriage, which meant that paternity was uncertain, and kinship was traced in the maternal line. For some authors favoring the theory of primitive matriliny, the term 'clan' was applied exclusively to matrilineal descent groups, while *gens* referred to patrilineal groups, a terminological distinction that has fallen out of use in the modern literature. Morgan, whose own research among the Iroquois in New York State had been influenced by Grote's analysis of the ancient Greek *genos*, saw descent group organization, which he termed the 'gentile' system, as characteristic of much of early human social history. McLennan, basing his reasoning on contemporary reports of female infanticide in India, claimed that primitive societies were obliged, in their struggles for survival, to engage in this population-limiting practice, which in turn led kin groups to procure wives by capture from outside their own group. McLennan termed this marriage outside the group 'exogamy,' which came to be seen, by many nineteenth and early twentieth century social theorists, as another characteristic attribute of clanship.

McLennan was also influential in developing the theory of totemism, which built on E. B. Tylor's notion that primitive religion was based on animism, the worship of inanimate objects or fetishes which were believed to be the abodes of spirits. Along with his follower W. Robertson Smith and Smith's student Sir

James Fraser, McLennan defined totemism as a system that originally combined animistic religious beliefs with matrilineal descent and exogamy. According to this view, through their ignorance of paternity caused by group marriage and their belief in animism, primitive societies posited links to original animal or plant ancestors. These natural species, which were treated as sacred emblems or 'totems' of different clans, were worshipped periodically by clan members in totemic rituals.

Emile Durkheim (1912), in his work on the social origins of religion which drew extensively on early ethnographic reports of aboriginal Australian societies, reproduced many of these classic arguments concerning totemism, particularly emphasizing the concept of a sacred social solidarity based on a belief in a shared substance between clan members and their ancestral totemic species. Claude Lévi-Strauss' (1962) subsequent general critique of theories of totemism, which emphasized the lack of a necessary coincidence between totemic species as emblems, exogamous clans, and religious sacrifice, served to unhitch the clanship concept from the pseudo-historical hypotheses of nineteenth century evolutionism while, at the same time, helping to clarify the classificatory logic operative in cultural categories such as the clan.

The work of the British structural-functionalist A. R. Radcliffe-Brown (1950) and his followers E. E. Evans-Pritchard, Meyer Fortes (Fortes and Evans-Pritchard 1940), and others between the two world wars, did much to put the study of unilineal descent systems on a firmer ethnographic footing, while refining the conceptual framework for their study. The model of the segmentary lineage, consisting of a nested set of increasingly inclusive corporate unilineal descent groups linked by a comprehensive genealogical structure, is the most well-known outgrowth of their work, with the term 'clan' often being used by these authors to label a particular segmentary level within a lineage system. Although Radcliffe-Brown's (1950, pp. 39–40) definition of the clan, given in the first sentence of this article, has come to be widely used, this application of the term to characterize one of the levels in a segmentary lineage system is not always appropriate. It reflects a continuing influence of earlier understandings of totemism, in which descent group exogamy and totemic taboos were taken to be defining attributes of clanship.

Although a definitional contrast between clan and lineage is thus still not always made clearly in the more modern literature, there is analytical value in maintaining a definite distinction, since the organizational implications of the two forms of descent grouping are different. Members of a lineage know, or claim to know, the genealogical connections interlinking all members of the group, and these links, viewed in terms of generation and relative birth order, often provide a basis for calculating relations of seniority or degrees of relatedness between individuals and segments within a lineage. Lineage genealogy can also be the basis for internal segmentation of groups, according to the segmentary lineage model. In contrast, lacking such a comprehensive internal genealogical armature, relations of clanship are categorical in character and typically nonhierarchical within the clan. Persons are members of a clan because of being the offspring of their fathers or mothers, with the terms 'patriclan' and 'matriclan' often being used to indicate a clan's mode of recruitment of its members. Although ascending descent links beyond the grandparental or great-grandparental generation are normally of little or no organizational significance, clan members in many societies do nonetheless recognize a founding clan ancestor, who is often of mythical or nonhuman status. Thus, when clans do segment according to putative genealogical links, this is typically 'from the top', with reference to the founding ancestor and his or her children, producing a set of sub-clan categories that also lack comprehensive internal genealogical structures.

Terminological confusion is also possible in the case of the descent group known as the conical clan, or ranked lineage, which combines characteristic features of both lineages and clans. Such descent groups are differentiated internally into a high ranked, lineage-like, chiefly or noble descent line, and a lower ranked and internally undifferentiated clan-like category of commoners (Kirchhoff 1959, Friedman 1979). Chiefly rank in a conical clan typically is based on relative birth order in present and ancestral generations, so that senior sons or daughters of senior ranking ancestors keep careful track of their pedigrees to validate their noble status. Junior offspring of junior ancestors, on the other hand, have little motivation in remembering their genealogies, and their affiliation to the group is more categorical in character. Numerous ethnographers have drawn attention to the structural potential of conical clanship, particularly as a transitional formation between uncentralized and centralized political systems. Conical clan structures are relatively common in Polynesian societies, such as the Maori, and in south-east and central Asian societies, such as the Mongols.

The clans of the Scottish Highlands also displayed such internal ranking (Dodgshon 1998). The Scottish clan chief and his close relatives enjoyed high status due to their close patrilineal connection to the founding clan ancestor. The commoner members of the clan, on the other hand, could not necessarily demonstrate such genealogical links, but were bound to their clan chief by diverse social ties including land tenancy or marriage alliance, as well as real or fictive kinship expressed by a common surname.

Systems of clanship occur in many parts of the world, and their diversity of form can be understood in relation to several key variables. Of prime importance is their capacity for organized collective action as corporate groups, which frequently is limited or absent

due to the spatial dispersal of clan memberships. Particularly in territorially extensive societies with large populations, a clan's members often reside in many different locations, and there are limited possibilities for the full membership of a clan to meet as a total group. In such circumstances, clans may act collectively on only a few occasions, such as during annual rituals like the aboriginal Australian *corroboree* ceremonies, which clans celebrated to ensure the continued fertility of their totemic species. In other cases of spatially dispersed clan membership, a clan may never meet as a totality, but simply act as a named category from which localized sets of clan members may be mobilized to pursue collective action if necessary or advantageous.

In addition to their variable capacities for organizing group action, numerous authors have also drawn attention to the potential of clanship systems to serve a social networking function. For example, in West Africa among speakers of the Mande family of languages (Jackson 1974), a limited set of clan names is found throughout this large zone, and a clan member making a long journey can feel confident of finding members of his or her clan who will provide assistance at the distant destination.

Even in cases where the set of clan names changes from one society to an adjacent one, conventional equivalences are often established between clan names to facilitate the extension of such networks of mutual aid. Similar clan networks have also been noted among North American Amerindian and aboriginal Australian societies.

Although ethnographic analysts of clanship frequently have conflated it with lineage structure, as mentioned above, the logic of clanship is often more a matter of 'sub-ethnicity' rather than 'super-lineage.' Many cultures' understandings of clanship, which are often rooted in myth, posit a primordial clan identity which is seen as immutable, much like many cultures' conceptualizations of ethnicity. It is this theme that Lévi-Strauss (1962) explored in his work on totemism, in which the classificatory logic of clans as social species is seen as homologous to cultural classifications of animal and plant species. In such cases, the total number of clans in a society may be quite small and be viewed as unchanging, with the various clanship categories standing in long-term interrelationships of marriage alliance, ritual cooperation or other modes of mutual exchange or solidarity. In societies where the total number of clan categories is only two, these groupings are conventionally referred to as 'moieties.'

In contrast, some societies have systems of proliferating clans in which it is evident to the analyst, from the presence of large numbers of clans with relatively small memberships in the context of overall population growth, that clans are undergoing segmentation, although this is not acknowledged by clan members themselves. Such systems of proliferating clanship tend to be found in politically uncentralized societies with high rates of spatial mobility, where local groups frequently split as a mode of dispute settlement. In such cases it may be posited that, following an episode of intraclan dispute which has been resolved by group fission and spatial displacement, the members of the resultant new segments have an active interest in denying knowledge of former genealogical interlinkage. Such systems of proliferating clanship, characterized by many small clans that are widely dispersed spatially, are not conducive to the maintenance of the patterns of long-term interclan alliance mentioned above.

In some cases, the term 'clan' has also been applied to territorial groups which are recruited both on the basis of unilineal descent and long-term co-residence. Such a usage is common in New Guinea. Here, inmigrant strangers hailing from other clans may initially be welcomed in order to strengthen a localized clan grouping.

Over time, cultural theories of personal identity may posit, for example, that through continued consumption of foodstuffs cultivated on the clan's land, the physical nature of such incomers becomes transformed into that of true clansmen, thus converting coresidence into shared unilineal descent. In effect, as argued by the American kinship specialist, George P. Murdock (1949), such a 'compromise kin group' reflects the difficulty of constituting a large-scale, residentially unified and collectively functional kinship grouping on the basis of unilineal descent alone.

Finally, it should be recognized that, in common parlance, the term 'clan' is often used in a metaphorical way to refer to any group of persons who act toward each other in a particularly close and mutually supportive way. Thus, criminal organizations such as the Mafia may be referred to as 'clans,' in recognition of the ideology of kinlike solidarity that binds their members together.

See also: Chiefdoms, Archaeology of; Kinship in Anthropology; Matrifocality; Tribe

Bibliography

Dodgshon R 1998 *From Chiefs to Landlords*. Edinburgh University Press, Edinburgh, UK

Durkheim E 1912 *Les Formes Elémentaires de la Vie Religieuse*. Alcan, Paris [1915 *The Elementary Forms of the Religious Life*. Allen & Unwin, London]

Fortes M, Evans-Pritchard E E (eds.) 1940 *African Political Systems*. Oxford University Press for the International African Institute, London

Friedman J 1979 *System, Structure and Contradiction in the Evolution of 'Asiatic' Social Formations*. National Museum of Denmark, Copenhagen, Denmark

Fustel de Coulanges N 1876 *La Cité Antique*. Hachette, Paris [1980 *The Ancient City*. Johns Hopkins University Press, Baltimore]

Grote G 1851 *A History of Greece*, 3rd edn. John Murray, London

Jackson M 1974 The structure and significance of Kuranko clanship. *Africa* **44**(4): 397–415
Kirchhoff P 1959 The principles of clanship in human society. In: Fried M (ed.) *Readings in Anthropology*. Crowell, New York, Vol. II
Kuper A 1988 *The Invention of Primitive Society*. Routledge, London
Lévi-Strauss C 1962 *Le Totemisme Aujourd'hui*. Presses Universitaires de France, Paris [1969 *Totemism*. Penguin, Harmondsworth, UK]
Maine H 1861 *Ancient Law*. John Murray, London
McLennan J F 1865 *Primitive Marriage*. Adam & Charles Black, Edinburgh, UK
Morgan L H 1877 *Ancient Society*. Henry Holt, New York
Murdock G P 1949 *Social Structure*. Macmillan, New York
Niebuhr B 1828 *The History of Rome*. Taylor, Cambridge, UK
Radcliffe-Brown A R 1950 Introduction. In: Radcliffe-Brown A R, Forde D (eds.) *African Systems of Kinship and Marriage*. Oxford University Press for the International African Institute, London

P. Burnham

Class Actions: Legal

A class action is a procedure whereby one or more claimants, called the class representative, may bring suit (or, more rarely, be designated as a defendant class) to obtain a remedy responsive to the legal interest of all members of the class (Lindblom 1996). The remedy may be an injunction, for example prohibiting racial discrimination in public schools, or compensatory damages and, in some instances, punitive damages, for example in claims of mass consumer fraud. The suit may seek both an injunction and monetary redress for or against a class. The procedure facilitates assertion of similar claims on behalf of a large number of allegedly injured parties, including claims that could not, as a practical matter, otherwise be asserted on account of the cost of litigation. A suit against a defendant class typically would seek imposition of an identical remedy against all of them, for example, against stockholders or creditors of a corporation. By the same token, because the procedure aggregates claims, it can impose very heavy liability and hence can become a weapon of 'legal blackmail.' The procedure is governed by Rule 23 of the Federal Rules of Civil Procedure in federal courts and by similar rules in state courts. Class suits are very controversial but they are an important component in the American use of litigation to address public issues of compensatory and distributive justice.

1. History

English courts recognized litigation on behalf of groups from no later than the fifteenth century and perhaps earlier. The concept of a class suit evolved in the Court of Chancery in the sixteenth and eighteenth centuries, in such cases as rights to fish in a river, rights of creditors against an insolvent debtor, rights of shareholders to enforce duties of corporate directors and obligations of parishioners to pay church titles (Yeazell 1987). By the end of the eighteenth century the basic formula for a class suit had been established, in these terms: 'Plaintiff, suing on behalf of himself and all others similar situated, alleges as follows: …'

In the United States the class suit procedure was elaborated by Justice Joseph Story in his treatise on Equity and was recognized in the federal courts, notably in the case of *Smith v. Swormstedt*, (57 US [16 How] 288 [1853]). In the nineteenth century the class suit evolved in state court procedure, particularly in litigation by city taxpayers complaining about improper municipal expenditures and in proceedings to reorganize insurance companies that had overextended themselves. The 'taxpayers' suit' has since evolved into a standard procedure for obtaining judicial review of action by municipal and state government. The insurance reorganization proceedings have since evolved, in one direction into bankruptcy procedure and in another offshoot into procedures for reorganization of insurance companies and banks (Hazard et al. 1998). The class suit procedure was used only infrequently until the litigation involving desegregation of the public schools in the 1960s and 1970s, in which the procedure was a standard technique.

The class suit procedure was given greater status and more precise definition in the Federal Rules of Civil Procedure, adopted in 1938. Federal Rule 23 permitted class suits for injunctions, damages and multiple claims against a limited fund. In 1966 Federal Rule 23 was elaborately amended to provide clearer criteria for when class suits maybe maintained and greater protection for class members, particularly those who were not the designated representative parties.

The validity of class suit procedure has been recurrently questioned in terms of due process. The issue is essentially whether, and in what circumstances, a judgment in a class suit can validly determine the rights of members of class who do not actively participate in the litigation. The courts have upheld the concept of the class suit but have given ambiguous pronouncements about its effect on absent class members. Leading Supreme Court cases addressing that issue include *Hansberry v. Lee* (311 US 32 [1938]) disapproving a class suit where the representative has a conflict of interest; *Phillips Petroleum Co. vs. Shutts* (472 US 797, [1985]) approving a class suit in state court determining rights of residents of other states; and *Amchem Products, Inc. vs. Windsor* (521 US 591, [1997]), and *Ortiz vs. Fibreboard Corp.* (527 US 815 [1999]) disapproving class suit settlements covering potential claimants who could not be given notice of the suit.

2. Requirements for a Class Suit

The requirements for a class suit in federal court are set forth in Federal Rule 23. Essentially similar requirements prevail in state court procedures.

The class suit complaint must be set forth, first, that plaintiff is a party injured in the manner detailed in the complaint's subsequent allegations and, second, that the plaintiff maintains the action on behalf of all members of the described class, for example, all Black children seeking admission to the defendant public school or all consumers who borrowed money from the defendant lending institution. The first of these allegations establishes 'standing' to sue and the second asserts plaintiff's assumption of a representative capacity in the litigation. The remedy must be sought for all members of the class. The complaint must describe the wrong in accordance with the usual pleading rules. It must also allege that the members of the group are too numerous to sue separately; that there are common questions presented in the claims for the class; and that the representative party will be an adequate representative of the entire class. It must also be shown that common relief for all class members is necessary or at least convenient and that the procedure will be efficient as a means of resolving the many claims involved.

The terminology in Federal Rule 23(b) differentiates three types of class suits: 23(b)(1), involving a course of conduct by defendant that affects all class members; 23(b)(2), in which an injunction is sought against conduct affecting all members of the class; 23(b)(3), in which similar injury has allegedly been inflicted on all class members, for which damages are sought. These distinctions are not mutually exclusive. For example, a proper class suit can seek both an injunction against future wrong and damages for past conduct. However, other provisions of Rule 23 impose different procedural requirements on the basis of this typology, particularly a requirement that all members of a '(b)(3)'class be given individual notice of the pendency of the suit. This discrepancy as well as redundancy in other terminology in Rule 23 have resulted in much confusion in administrating class suits. There is a large and complex procedural jurisprudence concerning these requirements (Wright et al. 1986).

However, there are two essential questions. The first is whether, all things considered, the litigation may be maintained on a group basis, or may not be so maintained. This is the 'certification' issue, so-called because the court must certify the suit as a proper class proceeding if it is to proceed as such. The second critical question is whether individual notice to all class members will be required.

3. Certification

A class suit may be maintained as such only with the approval of the court, that is, certification that the suit is proper. An initiative by the plaintiff to proceed with a class suit is ordinarily necessary, except in unusual situations where a defendant class is established. However, a plaintiff initiative is not sufficient without the court's approval. Usually the question of certification is a major preliminary issue, strenuously contested with legal and factual argument. A great deal depends on resolution of this issue. If certification is denied, the lawsuit lapses into a claim by one or a few parties, rather than on behalf of a large group. Reduction of the size of the case to the claims of the individual representatives usually makes further pursuit of the litigation by the plaintiff unattractive or impractical. On the other hand, if certification is granted, the lawsuit may immediately assume major proportion for the defendant, often creating a compulsion to reach expensive settlement.

Despite extensive procedural jurisprudence, and intensive professional and academic debate, the issue of certification in a specific case remains relatively open ended and hence very much a matter of judicial judgment and discretion (Newberg and Conte 1992). For many years the question of certification was held to be an interlocutory determination and hence not subject to appellate review until final resolution after determination of the merits of the class members claims. See *Eisen v. Carlisle & Jacquelin* (417 US 156 [1974]). This made the certification issue in the trial court all the more crucial. Denial of certification was termed the 'death knell' of a class suit, while grant of certification usually resulted in defendant feeling obliged to negotiate a settlement with the class. In 1999, Rule 23 was amended to permit immediate appeal of a grant or denial of certification. Immediate appeal of the trials courts' certification decision has been permitted under class suit procedure in most states.

4. Initiative to Prosecute Class Suits

A class suit can be initiated by a pre-existing group, such as a labor union or trade association; by a political action organization wishing to press a legal contention, such the NAACP's litigation to end school desegregation; or by a group recognizing itself to have a common interest in making claims, such as disaffected shareholders of a corporation. Initiation of such efforts requires assistance of a lawyer willing to prosecute the case. In modern practice such an initiative is often by lawyers specializing in class suit litigation. Because class suits typically involve large stakes so far as defendant is concerned, most class suits are strongly contested. That prospect in turn requires that the organizer, whether an action group or the lawyer, have financial resources and staying power to sustain costly and protracted litigation. Legal grievances that can be framed as class suits hence

ordinarily result in escalation of a dispute into litigation of major proportion.

The initiator of a class suit, whether a claimant action group or a lawyer, defines the grievant group by the description of the class in the complaint. The class is defined in terms of common characteristics and usually a specified time interval, for example, 'female employees of defendant corporation in the period January, 1998 through December, 1999.' The time interval refers to the period in which the alleged injuries occurred and is called the 'class period.' A larger class may be subdivided into subclasses having different specific characteristics, either by the plaintiff or by order of the court.

Some class suits are almost entirely the result of lawyer initiative. A lawyer envisions that a wrong has been committed against many people, locates someone fitting the description to serve as class representative, and then manages the lawsuit. The lawyer's incentive in such a case includes the prospect of large fees upon obtaining a judgment or, much more likely, a settlement. Some analysts differentiate between 'cause' class suits, such those asserting civil rights, and 'money' class suits, where large damages are sought. 'Money' class suits are typified by cases where inventive lawyers frame class damage claims out of transactions too small to be worth individual prosecution but affecting thousands of alleged victims. A classic illustration is a California case seeking restitution from a taxi company that allegedly fixed its meters to record fares at a rate higher than legally permitted. See *Daar vs. Yellow Cab Co.* (67 Cal. 2d 695 [1967]): Claims on behalf of corporate stockholders, alleging misrepresentation in the corporation's financial projections, are a common type of modern class suits. Many class suits involve both politically significantly issues and large financial stakes. In any event the class permits isolated individual legal grievances to be amalgamated into large scale claims.

5. *Uses and Abuses of Class Actions*

The typical class suit defendant is a business corporation or a government bureau. The class suit procedure is an important device for commutative and distributive justice on behalf of individuals who have suffered legal wrong at the hands of powerful actors in modern mass society. At the same time, the class suit procedure is a menacing device by which self-constituted protagonists, chiefly lawyers, can exploit relatively minor legal mistakes to reap large recoveries from 'target' defendants. Officials of such a defendant typically consider that they cannot afford the risk to their organization's continuity of a big judgment—the risk, in common parlance, of 'betting the ranch.'

A class suit also presents opportunity for the plaintiff's lawyer to exploit the class, often through connivance of the defendant. A class suit settlement typically calls for defendant to pay the plaintiff's lawyer a substantial fee. A defendant is typically indifferent whether settlement money goes to the lawyer or to the class members. A settlement could pay could pay $1 million to the lawyer and $9 million to the class, or $2 million to the lawyer and $6 million to the class. The terms of a settlement injunction can be similarly manipulated. The class members typically are dispersed and not organized, hence in a weak position to contest the terms of settlement (Coffee 2000).

6. *Use and Abuse of Class Suits*

Most class suits are resolved by settlement, typically after a period of intensive discovery and extensive motion practice. Hence, although class suit litigation typically is expensive, there are few cases where an adjudication has determined the merits. Most of the controversy over class suits is founded on dispute about the fairness of settlements.

Control against blackmail of a defendant or exploitation of absent class members is afforded through exercise of responsibility by class counsel and by court supervision. Most class counsel are faithful to their responsibilities, notwithstanding that class suit defendants typically assert that they are being oppressed. But some class counsel have been flagrantly exploitive. The quality of court supervision varies greatly. Some judges conduct searching inquiries into the basis and terms of a proposed settlement, particularly the attorney fees. Some judges make only a perfunctory review. Close supervision is in any event difficult because there is no reference point as to the merits of the claim and both sides are at high risk in a trial on the merits—plaintiff through possible loss of its investment in the litigation, defendant through possible loss on the merits (Hensler et al. 2000).

Both the opportunity to achieve justice for individuals suffering common wrongs and the danger to defendants of 'betting the ranch' could be improved if clearer definition could be drawn of claims appropriate for class suits. Efforts in this direction have been without much success. Congress adopted special provisions concerning securities class suits in federal court but their effect has been avoided by bringing suits in state courts. Proposed legislation would extend federal court jurisdiction to class suits involving interstate transactions, but encountered intense criticism that they would stifle proper class suits and inappropriately burden the federal courts. Revisions in Rule 23, in addition to making certification orders immediately appealable, have been proposed but probably would have modest effects.

Much of the controversy over class suits implicates other aspects of American civil procedure, including broad discovery, the right of jury trial of issues of fact, and the American rule concerning court costs (Yeazell

1987). Class suits typically involve extensive discovery which, under broad American discovery practice, can extend to thousands of documents and dozens of witness depositions. The right of jury trial imposes inherent risks to a 'deep pocket' defendant and, perhaps more important, precludes decisions by the judge of issues of fact that might narrow the controversy or resolve it altogether. The American cost rule, whereby the winner of litigation generally is not entitled to recover its litigation costs, means that a class suit plaintiff has no 'down side' risk of becoming liable for a defendant's costs. Moreover, many class suits claims are based on statutes that provide recovery of litigation costs to winning plaintiffs but not to defendants. A commonly voiced criticism of these statutes is that they contemplated individual suits, where litigation costs would make enforcement of rights practically impossible, whereas the class suit is another approach to enforcement of individual rights through a 'private attorney general.'

See also: Legal Process and Social Science: United States; Liability: Legal; Litigation; Parties: Litigants and Claimants; Procedure: Legal Aspects; Rules in the Legal Process

Bibliography

Coffee J 2000 Class action accountability: Reconciling exist, voice, and loyalty in representative litigation. *Columbia Law Review* **100**: 370

Hazard G, Gedid J, Sowle S 1998 An historical analysis of the binding effects of class suits. *University of Pennsylvania Law Review* **146**: 1849

Hensler D R, Pace N M, Dombey-Moore B, Giddens B, Gross J, Moller E K 2000 *Class Action Dilemmas: Pursuing Public Goals for Private Gain*. RAND, Santa Monica, CA

Lindblom P H 1996 *Group Actions and the Role of the Courts: A European Perspective*. Kluwer, The Hague, Netherlands

Newberg H, Conte A 1992 *Newberg on Class Actions*, 3rd edn. McGraw-Hill, New York

Vos W de 1996 Reflection on the introduction of a class action in South Africa. *Tydskrif vir die Suid-Afrikaanse Reg* **4**: 639–57

Wright C, Miller A, Kane M 1986 *Federal Practice and Procedure*. West, St Paul, MN, Vols. 7A–7C

Yeazell S 1987 *From Medieval Group Litigation to the Modern Class Action*. Yale University Press, Haven, CT

G. C. Hazard, Jr.

Class and Law

1. Basic Marxist Assumptions

The basic assumptions in a Marxist explanation of the origin, the development, and the function of law include the following: Productive labor is the basis and thus the focus of social ascriptions, i.e., of a collectively sharped pattern of social identity (not gender, ethnicity, and the like). On a certain level of productivity a surplus can be produced. The mode of production and appropriation of a surplus changes historically. Classes evolve out of unequal property relations to the means of production. Owners and non-owners of the means of production form the dominating vs. the dominated, exploited classes. (In non-Marxist approaches, e.g., in Max Weber, classes are defined not only according to their market position or possession but also according to their relation to the mode of production and appropriation; Weber 1964, p. 688.) Classes are not only structural, anonymous entities. Classes operate as collective actors vested with class consciousness (*Klasse für sich*) and class interests. The interests of the economically dominating class are backed by the state as an instrument of suppression. (In contrast, Max Weber suggests that the legal order has an impact on the power order of a society (Weber 1964, p. 678), which is determined by, among other things, the economic *Klassenlage*, insofar as the legal order protects the free disposition of the owners of the means of production (Weber 1964, p. 682).

At the same time, law disguises class domination behind a veil of ideological justification. The legal ideology of the formal equality of autonomous market subjects legitimates both the economic and the political order. Instead of mere force, a system of alleged universal legitimized domination, and thus the step from might to right, is institutionalized. Therefore, not only class interests but also general (class-independent) conditions of a society are secured by state and law. By its intellectual sublimation and doctrinal systematization, law gains a relative autonomy. Law is no longer the pure expression of the economic basis or state power. There exists a feedback of legal regulations on the economic basis including the class structure.

According to the basic assumptions of Marxist historical materialism, the systematic relationship between the class basis and the legal superstructure evolves in historical stages: In ancient 'primitive,' 'classless' societies, which do not produce a surplus, there is no need for state and law. In hierarchical societies that are economically based on slavery or a feudal system, state and law play a dominant role. In the tradition of Marxist legal theory, this characterization was denied by Eugen Pašukanis (1924). According to his approach, law exists only in capitalist (market) societies with antagonistic interests of the owners of commodities. Preceding social formations are marked by lawless domination. The stage of the dictatorship of the proletariat is, according to Marx/Engels and Lenin, signified by the use of law as an instrument of overt state domination, even of terror. After that, law is used as an instrument to build up and to secure a socialist society. The future of state and law in a communist society was an open question. The answers range from the notion of the withering away

of law and state, to a 'state of the whole people' in which law still serves administrative purposes.

In the following systematic presentation the concept of law is used either in the sense of legislative acts, or the activities of the legal staff, or the legally relevant activities, opinions, and beliefs of the citizens in general.

2. Class Legislation

In Marx's base–superstructure scheme (Marx 1859, pp. 8–9), law (as an element of the superstructure) is conceived of as an expression of the economic base. In the writings of Marx and Engels, the instrumental view comes in via the notion of *Rückwirkung* (feedback) of the superstructure onto the economic base. This instrumental view was dominant from the time that the communists gained power; for example, in the writings of Stalin. In the instrumentalist view, law is conceived of as an instrument used by the ruling class(es) in order to realize a socialist order. The theoretical background of Marxist legal theory can be characterized by this oscillation between law as an expression of the given class structure and law as an instrument of class domination.

But what are the concrete mechanisms by which a link between social classes and legislation, between the dominant class(es) and the legislative majority, can be achieved (as soon as there are distinct political bodies)? What is the logic of 'class legislation' (Mill 1861, Chap. 10)?

On a primitive level, members of the dominating class(es) are *in persona* members of the legislative body. With growing differentiation 'class representation' becomes important: parliamentary representatives are to a large extent nominated or elected by members of the dominant class; the legislative power of the dominant class can be secured by special arrangements (e.g., Prussian *Dreiklassenwahlrecht* (1849–1918) by which the representatives were elected according to the amount of taxes paid by groups of voters). This becomes a problem with the introduction of equal, universal suffrage. Class influence then can be maintained by a combination of economic and political positions, by regulations concerning access to Members of Parliament and government, by lobbying and corruption. Alternatively, input-orientation–output-orientation becomes more relevant; i.e., via threats and sanctions, capitalists can use their economic power (reduction, relocation, export of jobs, and investments, flight from taxation, etc.) to make the legislator issue certain regulations. The legislator has to take into account the consequences threatened by the employers. The subtlest form—often found in the critique of the bourgeois ideology of formal law—is the use of general, universal norms, the application of which confirms economically privileged positions. The construct of formal equality in civil law among allegedly free autonomous persons veils substantial inequality in market positions (e.g., in the labor market between employers and workers, in the housing market between landlords and tenants). But this normative pattern is relativized by special, protective regulations in these fields, when the weaker position of one group is acknowledged and becomes legally relevant.

3. Class Justice

It is a truism that courts cannot act neutrally in a class society and in a class state. If class justice is the justice of the class state that is organized according to the interests of the dominating class, how is it organized? What mechanisms ensure that the courts will act in the interest of the ruling class? To which class do judges belong? How are the interests of the dominating class transferred into the courts? (See *Judges*.)

The easiest example is, of course, where members of the economically and politically dominating class themselves have jurisdiction. The *Landesherr*, for example, is in charge of the administration of justice in his region. But with the professionalization of the judiciary, and a rudimentary separation of powers, the problem arises as to how the interests of the dominating class can be promoted by persons who are not themselves members of this class but rather act as class representatives. A number of factors help to promote the interests of the dominant class.

The first factor that promotes the interests of the dominant class is socialization. Judges tend to come from a similar background (Griffith 1981). Family background, and especially the occupation of the father, helps to explain the class bias in the administration of justice. And the high costs of education, in particular the costs of academic legal education, make the legal professions accessible only to members of families with a high income.

The second factor promoting the interests of the dominant class is the process of professionalization. That all judges share a common academic background and training is likely to have an impact on the attitude of law students, reinforcing a *status quo* orientation. Furthermore, judges must adhere to a set of professional standards; if they do not, they may suffer political consequences, or be subjected to professional discipline. Thus, in almost every state—be it a dictatorship, an apartheid regime, or a democracy—organizational means are successfully employed to secure the conformity of the legal profession to the political regime, and to the dominant political ideology.

Illuminating examples of the operation of these mechanisms can be found in Karl Liebknecht's critique of the class justice during the German Reich and in Prussia before the First World War: 'If we look

at where our judges come from it is obvious out of which milieu, from which point of view they regularly will decide. Naturally the judges will be recruited only from the possessing classes simply because of the high costs of education, because of the low income at the beginning of their career, by which the need for an adequate way of life, that still exists, cannot be satisfied.' (Liebknecht 1910, pp. 26–7) During this period the criterion of political reliability was overtly and restrictively used by the court administration to regulate the access to the judiciary. There are also brief remarks on the English system by Max Weber (1964, p. 1049) on the recruitment of judges from the advocacy who served the capitalistic interests of their clients (For England see Griffith (1981) and the recent analysis of Abel (1998): the access to the judiciary is politically controlled by the Lord Chancellor and there still exist strong financial barriers.) (See *Judicial Selection*.)

Although according to Marxist theory, state and law are always 'class state' and 'class law,' in socialist countries the term 'class justice' was not used as a self-description of their courts, but solely in a pejorative sense against 'bourgeois justice', etc.

Is there still—as it was evident for Karl Liebknecht in cases where workers were involved (as an accused or, in civil cases, with an employer on the other side)—a class bias in the administration of justice? What about civil cases in which the parties come both from the same class, i.e., from the class of the owners of the means of production, or from the proletariat—and here possibly as blue-collar workers versus white-collar workers? What if conflicting parties avoid court litigation? And in penal cases, how do we compare legally similar cases in which the judgment differs solely because of the different class affiliation of the accused? The appropriate field of study of class justice seems to be labor courts with employers and employees as parties, i.e., where class conflict is transformed into legal conflict (cf. Rottleuthner 1984).

There is substantial evidence documenting discrimination in penal courts with a high selectivity against the poor, from police arrest, through state prosecution, to disparities in sentencing and imprisonment. Assessing the extent to which capital punishment (see *Death Penalty*) is discriminatory in the US is methodologically complicated because of the correlation of crimes rates with social status, race and even gender (see *Gender and the Law*). Further, statistical correlations between race or gender and the law are hard to interpret because it is still an open question how these factors become operative within legal argumentation. The overrepresentation of lower class people in the penal system could be explained to a certain degree by legal standards, because there are correlations between legally relevant variables like observability of a crime, recidivism, confession, way of life, and sociologically relevant variables like (lower) class (see *Crime and Class*). The use of purely external variables that do not take into account legal argumentation raises several problems. The 'background approach' (i.e., the use of variables like father's occupation, religion, ethnicity, party, and other group affiliations) in judicial research did not lead to clear results. Does there still exist a homogeneous working class culture with distinct socialization patterns? Does legal education and role expectations on the job neutralize a possible impact of the social background? What about lay judges who have almost no legal training? When a jury is selected, lawyers consider the composition of race and gender (see *Juries*). There is strong evidence that neither the background or education of judges, nor their more recently achieved attributes like group or party affiliations, have an impact on the outcome of judicial decisions (particularly for labor courts; cf. Rottleuthner 1984). The focus of research into the outcomes of litigation has therefore shifted from the personality of judges to features of the parties to a court conflict (cf. Galanter 1974). Success in a civil court litigation can be explained by the experience of repeat players—in general bigger firms which tend to come out ahead of private one-shotters.

The focus of sociolegal research has furthermore shifted away from the courts to what happens out of court: to different legal needs (of the rich and poor), to unequal access to courts and other legal or extralegal remedies (see *Justice, Access to: Legal Representation of the Poor*), in general to the selectivity of legal procedures (see *Law, Mobilization of*). The problem of access to courts has been adressed by Max Weber (1964, p. 719) in the case of the English system: monetary barriers for the poor lead to denial of the administration of justice.

4. *Class Consciousness and Law*

There are no empirical studies that contrast the attitudes of workers (or of the proletariat) to legal norms and legal institutions with those of employers (or of capitalists). Empirical research into knowledge and opinion about law uses simple indicators of social status (preferably income or occupation), but no class variables in a strict sense. Georg Lukács (1920) dealt theoretically with the attititude of the proletariat towards legality and illegality and towards the bourgeois law and state in general. In order to establish successfully a proletarian state, the proletariat has to acquire a sober, purely tactical attitude toward law and state. The state has to be seen solely as an element of power, as an empirical entity without any normative obligatory force. Legality or illegality are not matters of principle but of utility. This instrumental attitude towards law and state can also be found among the members of Communist parties. After their experiences with bourgeois legislation and class justice, they, having gained power, used law as an instrument of

suppression and of strengthening the socialist state. The doctrine of the withering away of law and state was discarded; and law was not understood as a limit on state power.

5. *Abiding Marxian Concepts*

The replacement of the Marxian concept of class with constructs like status and milieu, and with indicators such as income, education, and occupation can itself be explained by substantial social changes. Labor is no longer the focus of social ascriptions. The forms of property or ownership (of the means of production) have changed and have become objectively blurred by the development of big companies (see *Property: Legal Aspects*). Other forms of social inequality became important like gender, race, ethnicity, nationality, age, etc. Of course, hierarchical stratification still exists: the haves and have-nots (the distinction between *Besitzenden* and the *Besitzlosen* constituted the *Klassenlage* in Weber 1964, p. 679), rich and poor, the legally included and excluded, that share or do not share the benefits and compensations of the welfare state, the politically dominating and dominated, etc. Instead of property, the notion of access becomes more important. Scholars are interested in the access to the means of production, but also the access to money, to the labor market, to information, knowledge, education (cultural capital), access to medical care, to other social benefits and rewards, to legal services and remedies, to leisure time options, to political power, etc. It is important to look at correlations and congruence between these varieties of access. There are no longer collectively conscious actors like classes. Some authors deny the existence of collectivities altogether under the heading of individualization; others speak of social movements (Eder 1993). Finally, the role of law has changed. Law is no longer conceived of as a general structure of society, whether based on social contractual consensus or on class domination or an anonymous economic base. Law serves as a political instrument, among others, used in order to achieve certain goals. Law, in particular welfare and labor law, is applied in order to compensate for the shortcomings of formal equality. Law is constitutive of state and social activities (see *Law as Constitutive*); and law can work to limit political power.

Basic Marxian concepts are used today only metaphorically. Everything, including culture and law, becomes a matter of production. The notion of capital is converted into social capital and cultural capital (Bourdieu 1986). Class becomes an empty shell signifying every form of inequality. Lenski (1966), for example, speaks of various class systems according to different principles of stratification. There is not only a 'power class,' but also a 'sexual' class system, (on class in general, see Milner 1999). Marxist legal theory could be understood as an economic analysis of law. But what today is labelled as 'economic analysis of law,' the individualistic, utilitarian approach of the Chicago school, would have become the target of Marx' sarcasm.

Class has been used as an explanatory variable, namely as a background variable, in order to explain the behavior of legislators or judges, or as a variable attributed to the accused or to the parties involved in a legal conflict. But it has regularly been used in a non-Marxian sense; in empirical research, indicators other than the property relationship to the means of production are often used. Thus the concept of class has been dissolved into notions like status, stratum, milieu, etc., and indicators like income, occupation, and level of education have taken the place of the theoretical construct of class. In sociology, as in sociolegal studies, class has become a vague notion relevant to social stratification or to social movements.

The critical heritage implied in analyses of class legislation, class justice, and legal class consciousness lies in the disclosure of inequality that is veiled behind formally equal legal rules. Marxian and Marxist analysis like Critical Legal Studies (see *Critical Legal Studies*) teach us about the 'dialectic' of formal and material equality. Formal equality in law means that there are legal rules framed in neutral language with no overt discrimination. But the application of these rules has a discriminatory impact on those who do not, in practice, have the options that are imputed to allegedly autonomous persons. The strategy of argumentation is fundamental for the discourse of inequality in general (see *Discrimination*.) insofar as it holds not only for class discrimination but also for unequal treatment in the cases of gender, race, color, ethnicity, nationality, etc. Also the attempts to overcome the contradiction between formal and material inequality, legally and/or politically, are similar in the various domains. Should one adopt compensatory legal measures, such as protective regulations? Should one opt for positive discrimination? (See *Affirmative Action: Comparative Policies and Controversies*; *Affirmative Action: Empirical Work on its Effectiveness*.) Or should one dispose of law generally because of its class character as Marx did in his critique of the Gotha program: law is by necessity a law of inequality (*Recht der Ungleichheit*) that cannot cope with the variety of individuals. (A similar argument can be found in radical feminism against the intrinsic male character of law. See *Feminist Legal Theory*.)

A negative heritage of the Marxist juxtaposition of class and law lies in the conception of law either as the expression of a basic anonymous class structure or as an instrument of the dominant class(es) as conscious actors. What has been systematically neglected is the ubiquitous constitutive role of law (already in the fundamental property relationships) and the functioning of law as a possible limit to political power; in short, the intrinsic value of legal normativity for a social order.

See also: Affirmative Action: Comparative Policies and Controversies; Class Consciousness; Class: Social; Crime and Class; Gender, Class, Race, and Ethnicity, Social Construction of; Justice, Access to: Legal Representation of the Poor; Networks and Linkages: Cultural Aspects; Social Class and Gender

Bibliography

Abel R L 1998 *The Legal Profession in England and Wales*. Blackwell, London

Bourdieu P 1986 Forms of capital. In: Richardson J G (ed.) *Handbook of Theory and Research for the Sociology of Education*. Greenwood, Westport, CT, pp. 241–58

Eder K 1993 *The New Politics of Class. Social Movements and Cultural Dynamics in Advanced Societies*. Sage, London

Galanter M 1974 Why the 'haves' come out ahead: Speculations on the limits of legal change. *Law Society Review* **9**: 95–160

Griffith J A G 1981 *The Politics of the Judiciary*, 2nd edn. Fontana, London

Institut gosudarstva i prava AN SSSR (ed.) 1970 *Marksistko-leninskaja obščaja teorija gosudarstva i prava* (Marxist–Leninist general theory of state and law). Moscow

Lenski G 1966 *Power and Priviledge: A theory of Social Stratification*. Mcgraw-Hill, New York

Liebknecht K 1910 Gegen die preußische Klassenjustiz (Against the Prussian class justice). In: *Gesammelte Reden und Schriften*, Vol. III. 1960 Dietz, Berlin, pp. 3–55

Lukács G 1920 1971 Legality and illegality. In: *History and Class Consciousness: studies in Marxist Dialectics*. MIT Press, Cambridge, MA

Marx K 1859 Zur Kritik der politischen Ökonomie. Vorwort (On the critique of political economy. Preface). In: *Marx-Engels-Werke*, Dietz, Berlin 1975, Vol. 13, pp. 7–11

Mill J S 1861 1962 *Considerations on Representative Government*. Reprint Regnery, Chicago

Milner A 1999 *Class*. Sage, London

Pašukanis E 1924 1980 General theory of law and Marxism. In: Beirne P, Sharlet R (eds.) *Selected Writings on Marxism and Law*. Academic Press, London

Rottleuthner H (ed.) 1984 *Rechtssoziologische Studien zur Arbeitsgerichtsbarkeit* (Socio-legal studies on labor courts). Nomos Verlagsgesellschaft, Baden-Baden

Weber M 1964 *Wirtschaft und Gesellschaft* (Economy and society). Kiepenheuer and Witsch, Berlin

H. Rottleuthner

Class Consciousness

1. A Key Concept of Marxist Theory

In Marx's work the concept of class consciousness is based on his theory of social classes and the distinction he makes between existing social classes and politically active ones. It is not sufficient for a large number of individuals merely to be living under similar social conditions for them to constitute a class. In *The Eighteenth Brumaire of Louis Bonaparte* (1852), Marx describes the living conditions of peasant families. He sees them as having few connections with each other and even less with the rest of society. For this reason, these families do not form a social class. Only groups of individuals engaged in common activities, mainly relations of production and exchange, constitute a social class.

In the struggle that necessarily opposes the owners of the means of production and the workers, who possess only their labor power, all the conditions are fulfilled for the proletariat to constitute a true social class. The workers are all equally dependent upon their employers. They are forced to sell their labor and must also resist the stronghold of capital whose demands continually threaten their livelihood. This situation of dependence and resistance, this class struggle, which accompanies the beginnings of capitalist relations, becomes particularly intense due to the system's growing contradictions. As observed by Marx from 1845 onward, this is the context in which the issue of class consciousness arises. In other words, it is the situation through which workers become conscious of their shared socioeconomic conditions, of their fundamentally antagonistic relationship with the capitalists, and hence of the need for political struggle. This awareness signals a new change within the working class itself, leading it from the condition of class 'in itself' to one of class 'for itself.'

In *The Poverty of Philosophy* (1847), Marx analyzes the economic transformation of the working class and the ensuing changes in its subjective position. Since workers have been brought together, 'agglomerated,' in the same workplaces, they experience similar labor conditions. They face the same demands from their employers, and are therefore led to organize strikes together and to form 'coalitions,' which persist beyond the period of strike. Such struggles and communal activities bring the entire working class above the level of local conflicts and heighten its awareness of its political role. This rise in consciousness represents more than simply the awareness of a particular situation. For Marx, proletarian consciousness is simultaneously the discovery by the laborers of their extreme alienation and of their need to overcome such alienation through a form of action aimed at destroying the capitalist mode of production. Class consciousness is considered to be the *sine qua non* of social revolution.

Such a concept, which is both an element of Marxist philosophy of history and a theory of social change, comes from two major sources: one theoretical, the other empirical. Hegel, in *The Phenomenology of Mind* (1807), describes 'self-consciousness' as a moment in the mind's evolution by means of which the subject reaches a heightened level of awareness and a new potential for action. Yet it is essentially the changes in working conditions in the UK and France that

inspired Marx to formulate such a concept. As E. P. Thompson (1963) shows in his study, *The Making of the English Working Class*, the years 1820–40 witnessed a decline in the traditional organization of the various trades, a re-enforcement of the process of proletarianization of the English working class, and the beginnings of chartism. In France, during the 1840s an increasing number of scientific and journalistic writings emphasized the unification of the working class, who thereby transcended divisions between the trades and traditional organizations. Newspapers appeared, written by workers, in which they asserted their need to express their own claims and their hopes for a new form of social organization. Some of these reiterated the appeal formulated by Saint-Simon as early as 1820, calling for class consciousness as the fundamental condition necessary for political activity by the agents of production.

2. *A Political Debate*

In 1865, Proudhon, one of the founders of anarchism, developed a theory similar to that of Marx. In his work, *De la capacité politique des classes ouvrières* (On the political potential of the working classes), he poses the question of the conditions necessary for a social class to unite in political action. He formulates three conditions: the achieving of self-consciousness, the formulation of a political program, and the capacity to carry out such a program. He believes that the working class has fulfilled the first two conditions since 1848, but that it has yet to reach the third condition. He emphasizes, even more than does Marx, the dynamics of the social movement and also the problems it faces as it attempts to achieve its goals.

This theory of class consciousness led to divergent and far-reaching interpretations. One could indeed conclude from it that the function of a labor party should be only to encourage class consciousness, without aiming to control or direct it. On the contrary, Lenin, in his 1902 publication *What is to be Done*, states that the class consciousness of the workers has led spontaneously within the context of capitalism to purely economic claims. He considers that it is not the working class, but rather intellectuals of bourgeois origin, who produced the theory of social revolution. He therefore concludes that the political leadership of the movement must not be handed over to representatives of the working class (who tend towards reformism and anarchism), but rather to professional revolutionaries, united by strong party discipline. These sociopolitical arguments provoked critical reactions on the part of Marxist theoreticians (Rosa Luxembourg, Georg Lukacs, Antonio Gramsci), who feared the consequences of replacing class consciousness by an authoritarian party.

3. *Four Sociological Criteria*

After World War II, certain works suggested that such a theory should be made sociologically relevant. To this end Mann (1973) distinguished four principles which together would constitute class consciousness: (a) identity—the definition of oneself as a member of the working class; (b) opposition—the designation of an opposing class; (c) totality—the vision of society as a whole; (d) alternative—an alternative vision of society based on different principles of organization and aimed at replacing the established order. These four conditions give rise to many different questions which numerous studies have attempted to answer.

Many studies lean toward a positive answer to the first principle, or condition, concerning the awareness of an identity. Richard Hoggart's work (1957) asks whether it is pertinent to emphasize the originality of working-class culture. Using the perspective of cultural anthropology, the study undertaken by Hoggart in England in the 1950s illustrates the uniqueness of the attitudes, the habits and the representations of the lower classes. Hoggart underlines the permanence of a strong identity among members of this milieu and their heightened awareness of the social distance existing between themselves and other social classes. The opposition between 'us' and 'them' lies at the heart of their entire social representation.

The world of the 'others' is a vague congregation of employers of the private sector and civil servants. However, such a feeling of membership does not necessarily lead to the consciousness that there exists an inevitable conflict. Moreover, class consciousness is limited essentially to those who work in industry. One might add that studies which adopt the perspective of social stratification bring nuances to this notion of identity. The works by W. L. Warner, published in the USA in 1949, suggest that the social structure differentiates between multiple levels of social status. Three of them—the lower-middle class, the upper-lower class, and the lower-lower class—together make up the working class. To this is added the distinction between ethnic groups, characterized by differing feelings of membership.

The principle of the opposition to, and the designation of, an adversarial class may be regarded both from the perspective of representation and from that of action. Many studies indicate that it is not among the most disadvantaged workers that one finds the most entrenched adversarial stances. Material difficulties, unemployment, insecurity, and lack of qualification generate a consciousness of the opposition between 'us' and 'them,' but not a polarization against an adversarial class. It is mainly in professions such as mining and among dock workers, where attachment to one's job and loyalty between members of the working group are very strong, that one finds the most adamant expressions of general opposition to authority (Touraine 1966). In industry, among skilled

workers, attitudes are less radical. In their research, J. H. Goldthorpe and his collaborators (1969) formulated the general hypothesis that prosperous 'affluent' workers, who have a more functional relationship to their work, seek to obtain job stability and a higher salary, and are more interested in consumption than conflict. Patterns of strike participation confirm the main points of this thesis. In industry, strikes have wide participation, while in the service sector (banking, administration), where the organizing unions are concerned more with obtaining satisfaction for particular claims (raises, better working conditions, shorter working hours), they tend to result in negotiations.

The theory of class consciousness presupposes that the working class, due to its subordinate position, will eventually come to have a global image of society, as opposed to those who own wealth and whose horizons are limited by their personal interests (Lukács 1971). Studies of the 'images' which workers hold of the social system help to shed light on such assumptions (Bulmer 1975). These studies show that workers' representations are not as homogenous as Marxist theory might lead one to assume. Depending on whether the people questioned interpret the social system in terms of power, of prestige, or of property, they are more or less likely to perceive the social system in terms of conflict. The most adversarial perception of society is expressed by those whose perceptions are linked to power. Yet such a representation is not the most frequent one. W. L. Warner's studies of US workers already suggest that differences exist depending on the type of workplace, the degree of urbanization, and the level of qualification. The attitude of the 'affluent workers,' because they tend toward representations related to possession, suggest the generalization of an image of society in which the majority is made up of a middle class, subdivided according to wealth, income, and consumption habits.

These studies cast doubt upon the argument that claims that the working class is driven by a will for political change and by a capacity to support an alternative program. During the 1960s, these observations led to the proposal that such a political potential for change was no longer to be found among unqualified workers, but instead in a 'new working class' made up of technicians and more qualified workers from cutting-edge industries such as computers, communications, and petrochemical industries (Mallet 1963). This idea, which places emphasis on new skills, has not been confirmed. It assumes that these highly qualified technicians and workers will call upon their unions to adopt radical political stances. Yet the low level of union membership among workers in general, and in particular among such technicians, tends to indicate that they have different types of demands. They ask their unions to negotiate to safeguard their own interests, while expecting political parties to defend national interests. In this sense they reproduce the traditional divisions found in public opinion rather than following the model of political awareness supposedly characteristic of the working class.

See also: Bourgeoisie/Middle Classes, History of; Class: Social; Labor Movements, History of; Marx, Karl (1818–89); Mobility: Social; Poverty, Culture of; Social Class and Gender; Social Mobility, History of; Social Movements, Sociology of; Social Stratification; Working Classes, History of

Bibliography

Bulmer M (ed.) 1975 *Working Class Images of Society*. Routledge & Kegan Paul, London

Dahrendorf R 1959 *Class and Class Conflicts in Industrial Society*. Routledge & Kegan Paul, London

Goldthorpe J H, Lockwood D, Bechhofer F, Plat J 1969 *The Affluent Worker in the Class Structure*. Cambridge University Press, Cambridge, UK

Hoggart R 1957 *The Uses of Literacy. Aspects of Working-Class Life with Special References to Publications and Entertainments*. Chatto & Windus, London

Lockwood D 1958 *The Blackcoated Worker. A Study in Class Consciousness*. George Allen and Unwin, London

Lukács G 1971 [trans. Livingstone R] *History and Class Consciousness: Studies in Marxist Dialectics*. Merlin Press, London

Mallet S 1963 *La nouvelle classe ouvrière*. Editions du Seuil, Paris

Mann M 1973 *Consciousness and Action among the Western Working Class*. Macmillan, London

Thompson E P 1963 *The Making of the English Working Class*. Gollancz, London

Touraine A 1966 *La conscience ouvrière*. Editions du Seuil, Paris

Warner W L, Meeker M, Eells K 1960 *Social Class in America. A Manual of Procedure for the Measurement of Social Status*. Harper & Row, New York

P. Ansart

Class: Social

Class is a key concept in sociological theory, but its precise meaning and definition is highly contested. It is also a core explanatory variable in much empirical research, yet there is enormous diversity in the ways in which it is operationalized and measured. Most sociologists agree that social class refers to how people make a living, and that there are relatively stable patterns of inequality between different classes—but there the consensus ends. Indeed, some sociologists now argue that the concept of social class has outlived its usefulness and should be abandoned altogether.

1. Marx's Class Theory

The concept of 'class' was central to Marx's theory of how societies are constituted and how they change,

although Marx himself never produced a definitive statement of how class was to be conceptualized. The treatment of class in his different works is not always consistent, but five basic themes emerge.

First, although different classes generally have different levels of income and different life-styles, it is not their income or life-style *per se* that distinguishes them. The crucial determinant of class is ownership (or nonownership) of productive property. Because people either own, or do not own, the basic means of production in a society, it follows that there can be only two principal classes in any mode of production. Under capitalism, these are the bourgeoisie and the proletariat.

Second, class is not a 'position' which we occupy in society, but a relationship which one group in society has to another. This relationship is one of exploitation, for those who own the means of production add to their wealth by using the labor power of those who own nothing. Because class relations are always exploitative, they are always antagonistic. Class struggle is an inherent feature of all class societies, and this perpetual battle between classes is the motor driving human history.

Third, class is more than just a concept developed by social scientists to help them make sense of the world. Classes exist, and they have real effects. Even where the members of a given class do not recognize their class identity (a situation which Engels labeled 'false consciousness'), class is still an objective reality shaping their lives.

Fourth, class struggle is a feature of every society since the development of settled agriculture. Even where divisions seem to reflect factors other than property ownership (e.g., in caste systems where status at birth counts more than wealth), the 'real' force structuring social relations is still class. Class relations shape every aspect of social life. Politics, law, art, and philosophy are the 'superstructural' expressions of a more 'basic' relation between owners and nonowners of the means of production. Class relations impose their stamp on every aspect of life.

Fifth, Marx claimed that the class relation between bourgeoisie and proletariat was becoming increasingly sharp. Capitalist societies were polarizing between a small number of large capitalists and a large number of propertyless proletarians. He predicted that the condition of the proletarians would become increasingly miserable, and that sooner or later, there would be a revolution which would replace the capitalist system with a socialist one in which there would be no classes because all productive property would be owned in common.

2. *The Marxist Tradition of Class Analysis*

Later Marxists have wrestled with some major problems left unresolved by Marx's analysis.

2.1 *How Many Classes?*

One of the most intractable problems arises out of Marx's insistence that class is structured around the relationship between owners and nonowners of the means of production. Logically, this implies that there are only ever two classes, yet in *Capital* he identifies *three* classes in modern capitalism (landowners, wage laborers, and capitalists), and, analyzing the 1848 events in Paris, he finds as many as *nine* (proletarians, financiers, industrialists, the middle class, the petty bourgeoisie, the lumpenproletariat, intellectuals, the clergy, and the peasantry).

Later Marxists have sought to resolve this confusion in two ways. First, they have distinguished 'classes' and 'class fractions.' Owners of capital, for example, are a single class, but different 'fractions' of this class squabble when their interests diverge. Thus, landowners seek to extract high rents, but this conflicts with the interests of industrialists who want to keep their overhead costs down. Industrialists and landowners share a common class interest in the maintenance of the capitalist system, but their different interests in respect of profits and rents creates different fractions within the bourgeoisie.

Second, Marxist theorists have distinguished between a pure 'mode of production' (where there are only ever two classes) and actual 'social formations' (which may contain elements of more than one mode of production, and thus more than two classes). For example, France and the UK are structured around a capitalist mode of production (with two classes, bourgeoisie and proletariat), but they also contain elements of an earlier, feudal, mode of production (the surviving elements of a peasant class in France, and an aristocratic class in the UK).

2.2 *Theorizing the Middle Class*

A common feature of modern capitalism in all parts of the world is the growth of managerial, administrative, and professional occupations. This 'middle class' cannot easily be explained as either a 'fraction' of one of the two 'main' classes, or a remnant of an earlier mode of production. Braverman (1974) suggested that large sections of it are 'really' part of the proletariat because they are increasingly subject to typically 'proletarian' conditions of labor such as job insecurity and deskilling, but this interpretation has been challenged, and higher up the hierarchy it is clear that managers and professionals are in quite a different situation from manual workers as regards remuneration, security of employment, career prospects, and workplace autonomy.

Recognition of these differences has resulted in various attempts to develop Marx's theory to take account of the distinctiveness of the 'new middle class.' Carchedi (1975) theorized it as employees who sim-

ultaneously contribute to the functions of 'collective worker' (e.g., by coordinating the labor process) and of capital (e.g., by controlling their fellow workers). Poulantzas (1975) distinguished it on 'political' (supervisory functions) and 'ideological' (mental labor), as well as 'economic' (nonproductive activity) criteria. And Wright (1985), while emphasizing the division between owners and nonowners of capital assets as fundamental, differentiated nonowners according to their ability to 'exploit' skills and organizational assets (a class of 'expert managers,' for example, exploits both skills and organizational assets, while proletarians can exploit neither).

All of these formulations have been influential. Their weakness, however, is their complexity (Wright's schema, for example, generates a total of 12 different classes). The more Marxist theorists have tried to develop a descriptively adequate account of the contemporary class structure, the more they have had to sacrifice the theoretical sharpness of Marx's original approach.

3. *Weber's Concept of Class*

Weber contradicts almost every element in Marx's approach. Where Marx insists that classes arise out of the organization of production, Weber treats them as distributional categories. For him, 'economic classes' can be identified wherever individuals share a common *market situation*, either in property markets (where those who are 'positively privileged' can live off revenues from assets) or in labor markets (where those who are 'positively privileged' can command high salaries in return for their scarce skills and qualifications). 'Social classes' are simply clusters of economic classes between which inter- or intragenerational mobility is 'easy and typical.'

In this approach, social classes are not defined in relation to each other, as they are for Marx, but are mapped positionally in a hierarchy according to their market capacity. Weber therefore has no problem identifying a 'middle class'; indeed, logically, there are two middle classes (a petty bourgeoisie of small property owners with a relatively weak labor-market position, and an intelligentsia who own few assets but who command high returns in the labor market because of their education and training). They sit between an upper class (positively privileged in property and skills) and a working class (negatively privileged on both, and therefore relatively poorly remunerated).

'Class' is not 'real' for Weber as it is for Marx—it is simply an analytical construct, a label to refer to clusters of individuals. Sometimes people express a common class identity and act accordingly—but a failure to think and act in class terms does not constitute a 'false consciousness' as it does for Engels.

It is also clear that Weber does not see class as pervasive and enduring in the way that Marx does. For him, class (market power) is only one dimension of power in society. In the modern period, it is often the most important dimension, but in other periods, the 'social power' of status groups (such as the old European aristocracy or the high castes in India), or the 'political power' of parties, mobilizing to turn state authority to their own advantage, has outweighed that of classes. Even today, status and political power can cut across class lines. There is no assumption in Weber, as there is in Marx, that the dominant economic classes are also the dominant social and political force in any society.

4. *The Weberian Tradition*

4.1 *Class Location and Social Action*

Weber's approach to class analysis was essentially concerned with classification—social classes are ideal types. This has made things easier for later theorists, for unlike Marx's approach, Weber's work can readily be adapted and developed to take account of changed conditions (such as the growth and increased complexity of the middle classes).

What is lacking in Weber's approach, however, is a causal theory. For Marx, classes *act*—they drive history through struggle. For Weber, however, classes are simply categories into which people can be classified, and the role played by these categories in the explanation of social phenomena is left open. Class may help explain some of the things that people do, but it may be irrelevant to others—the usefulness of the concept is left to empirical research to determine.

This is both a strength and a weakness of the Weberian legacy. Its strength is that it has enabled the development of class typologies which appear both valid (they adequately capture some of the key sources of differentiation in contemporary societies) and reliable (they predict fairly accurately different patterns of social behavior and attitudes). Its weakness is that Weberian class analysis has often been limited to mapping differences between classes rather than explaining them. It generates useful depictions of the 'class structure' but seems to lack a theory of 'class action' (why and how a person's location in the structure influences social action).

4.2 *Work, Market, and Status Situations*

One influential attempt to link people's class position to their attitudes and behavior was David Lockwood's (1958) study of clerical workers in Britain. He defined class position by people's 'work situation' (i.e., the

degree of autonomy they enjoy and the authority invested in them or exercised over them) and 'status situation' (i.e., their occupational prestige) as well as their 'market situation' (i.e., their income, job security, and career prospects), and he showed that on all three criteria, clerical employees differed radically from manual workers. These differences were then reflected in their behaviors and attitudes, such as willingness to join a trade union.

Lockwood and his colleagues utilized much the same approach in a later study of the British working class (Goldthorpe et al. 1969). The research refuted the *embourgeoisement* thesis—the idea that affluence was blurring the boundary between the working class and the middle class—by identifying clear differences in the work, market and status situations of well-paid manual workers as compared with white-collar workers. These differences were then held to explain the different patterns of consciousness found in each group. Lockwood (1966) also analyzed differences *within* the working class, arguing that different work and 'community' situations could account for the different 'images of society' characteristic of traditional proletarian workers (e.g., miners), traditional deferential workers (e.g., farm laborers), and privatized workers (such as those employed in modern, high-wage industries).

Since then, John Goldthorpe (1987) has used differences in the 'work situation' (authority relations) and 'market situation' (income, security, and prospects) of different occupational groups to develop a new, 11-category model of the class structure. Developed initially as a framework for the study of social mobility in Britain, this typology became the basis for a major international study of mobility rates (Erikson and Goldthorpe 1992), and it has also been used to investigate issues such as the class basis of voting behavior (Heath et al. 1985). In comparison with other schema, it has been found to have a stronger predictive power and clearer internal consistency (Marshall et al. 1988, Breen and Rottman 1995), and it has won widespread acceptance in much of Western sociology.

4.3 Class Structuration and Social Closure

Lockwood's concern with linking people's experience, at work and in the locality, with their sense of class identity was later elaborated in Giddens's theory of class structuration. He explicitly addressed the problem of how a shared economic location becomes important for collective social action.

Weber's answer to this question emphasized patterns of social mobility and closure—social classes are clusters of individuals occupying market situations between which mobility is readily possible. Giddens (1973) refers to this as the 'mediate' structuration of social classes. He argues that market situations in modern capitalism typically cluster into three categories based around ownership of property, possession of qualifications, and possession of manual labor power, and that movement between these three is typically limited. These are the three social classes of the modern period, but Giddens then goes on to identify the factors ('proximate structuration') which keep them apart and help promote distinctive class identities. Here he echoes Lockwood's work by emphasizing the importance of differences in the workplace (e.g., in authority relations) and in the locality (e.g., residential segregation). He argues that proximate structuration always leads to some degree of 'class awareness,' though not necessarily to 'class consciousness' in a Marxist sense.

The problem of linking the 'class position' that people occupy to their values, beliefs, and actions was tackled very differently by Frank Parkin (1979), who resurrected Weber's neglected concept of 'social closure.' Closure refers to the way groups try to improve or maintain their privileges by restricting the access of others, and Parkin identified two main strategies. 'Exclusion' operates downwards and involves protection of privileges by dominant groups; 'usurpation' operates upwards and represents the attempt by subordinate groups to claim more privileges for themselves.

Exclusion typically seeks to defend privileges associated with property rights or qualifications (e.g., professional closure). Parkin defines the dominant class as those whose resources (revenues or salaries/fees) derive mainly from the exercise of exclusion. Usurpation, by contrast, is typified by forms of solidaristic action such as trade-union organization, and it is the defining feature of the subordinate class. Those in white-collar occupations, who frequently claim rewards on the basis of their individual qualifications (exclusion) and through membership of the organized labor movement (usurpation), constitute an 'intermediate class' between these two.

Defining classes by their action (mode of closure), rather than by their location in a structure of positions, Parkin claims to have sidestepped the recurring problem of how to link class positions to behavior. In his view, the 'class structure' derives from collective forms of action, rather than the other way around, and there is no structure of positions independent of the way people act.

5. Measurement of Class

Theoretical debates over whether and how to divide the population into discrete class categories are reflected in the different ways the class concept has been operationalized in empirical research.

5.1 Class as a System of Categories

Class has often been operationalized in quite crude and untheorized ways. Many studies have simply

distinguished 'manual' and 'nonmanual' occupational groups, and some have used the marketing industry's sixfold system of classification based on spending power, but neither of these approaches corresponds to any sociological concept of class. In the UK, research has often used the government's five (later six) class schema, but the criteria underpinning this system shifted over time from occupational prestige to skill levels creating confusion over what the classes designate.

In recent years, there has been a growing consensus around the use of systems of classification based on the Goldthorpe schema, and even the UK government's system of classification has now been revised to bring it more into line with the neo-Weberian emphasis on work and market situation (Rose and O'Reilly 1997). However, there is still intense disagreement among researchers over whether it makes sense to measure class in terms of *categories* at all, irrespective of how they are derived.

5.2 Class as a Continuous Scale

In the USA, there is a long tradition of measuring class differences on a continuous scale rather than a categorical system of classification. The best known example is the 96-point scale developed by Blau and Duncan (1967) for their study of social mobility. They justified using a continuous scale rather than discrete categories on the grounds that there are no clear cutoff points between occupations on any of the pertinent criteria that might differentiate them. In other words, classes shade off into one another.

Their approach attracted widespread criticism. Despite reporting high correlations (> 0.9) between occupational prestige and income, and education levels, they were accused of measuring status differences rather than differences in market power, and they were attacked for their 'conservative' assumption that there is a consensus over the prevailing system of occupational rewards (Horan 1978). Partly in order to counter such criticisms, Stewart et al. (1980) developed a scale which ranks occupations according to the 'social distance' that separates them, and which does not therefore depend on any assumption of value consensus over the worth of different positions. In practice, this scale looks very similar to those based on prestige rankings and correlates highly with them, which suggests that the problem of 'ideological contamination' of occupational prestige scales has probably been exaggerated.

The advantage of continuous scales over categorical schema (such as Goldthorpe's) is that they avoid the problem of drawing artificial class 'boundaries' (Kelley 1990). There is, however, no reason why we should not use both in empirical research.

6. The Future of Class Analysis

Controversy continues today, not only over how to measure class, but also over whether the concept remains useful.

6.1 Women and Class

Feminists have criticized the failure of class analysis adequately to encompass the position of women. Like students, retired people, and the unemployed, married women who do not have full-time jobs are usually classified according to the occupation of their husbands, and even women who do have jobs may be allocated to their male partner's social class where that is 'higher' than theirs. Goldthorpe defends this on the grounds that the 'life chances' of a lower-class woman married to a higher-class man are shaped more by his market situation than by hers, but this issue still generates considerable disagreement. Very different pictures of the class structure emerge depending on whether women are allocated to their own, or to their husband's, class location.

6.2 Nonclass Identities

Feminists have also criticized class analysis for the assumption that 'class location' is more important than gender in determining life chances and influencing values and behavior. Their argument is complemented by those (notably in the USA) who emphasize the primary importance of race and ethnicity rather than class. There has also been some debate over whether people's 'consumption location' (e.g., as home owners or renters) outweighs their class location, and postmodernists insist that clear class divisions have now fragmented into a multitude of different interests and identities.

Class analysts have tended to respond to such criticisms by appealing to evidence that class identity is still primary (Marshall et al. 1988). They point out that class effects may be crosscut by gender, race, or other identities without class itself losing its significance.

6.3 The Relevance of Class

Some critics (Pahl 1989, Clark and Lipset 1991) argue that class is a concept that has outlived its usefulness. People no longer think of themselves in class terms, movement between class locations is common, boundaries between classes have blurred, and class analysis has failed to explain the causal link between class locations and social outcomes.

Responding to this, Breen and Rottman (1995) accept that class is a weak source of identity for many

people, but they argue that people's 'objective' class location is still crucial in explaining many aspects of their lives. Recent work on class differences in health and morbidity (Wilkinson 1996) are one striking example. It is probably fair to conclude that class does still correlate significantly with many of the phenomena studied in social science, but we often lack a clear explanation of the mechanism which translates 'class position' into social outcomes.

See also: Class Consciousness; Consumption, Sociology of; Equality of Opportunity; Income Distribution; Inequality; Inequality: Comparative Aspects; Marx, Karl (1818–89); Mobility: Social; Social Class and Gender; Social Mobility, History of; Status and Role: Structural Aspects; Underclass; Wealth Distribution; Weber, Max (1864–1920)

Bibliography

Blau P M, Duncan O D 1967 *The American Occupational Structure*. Wiley, New York

Braverman H 1974 *Labor and Monopoly Capital*. Monthly Review Press, New York

Breen R, Rottman D B 1995 *Class Stratification: Comparative Perspective*. Harvester Wheatsheaf, New York

Carchedi G 1975 Economic identification of the new middle class. *Economy and Society* **4**: 1–86

Clark T N, Lipset S M 1991 Are social classes dying? *International Sociology* **6**: 397–410

Erikson R, Goldthorpe J H 1992 *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Clarendon Press, Oxford, UK

Giddens A 1973 *The Class Structure of the Advanced Societies*. Hutchinson, London

Goldthorpe J H 1987 *Social Mobility and Class Structure in Modern Britain*, 2nd edn. Clarendon Press, Oxford, UK

Goldthorpe J, Lockwood D, Becchoffer F, Platt J 1969 *The Affluent Worker in the Class Structure*. Cambridge University Press, London

Heath A, Jowell R, Curtice J 1985 *How Britain Votes* 1st edn. Pergamon Press, Oxford, UK

Horan P M 1978 Is status attainment research a theoretical? *American Sociological Review* **43**: 534–41

Kelley J 1990 The failure of a paradigm: Log-linear models of social mobility. In: Clark J, Modgil C, Modgil S (eds.) *John H Goldthorpe: Consensus and Controversy*. Falmer Press, London

Lockwood D 1958 *The Blackcoated Worker*. Allen and Unwin, London

Lockwood D 1966 Sources of variation in working class images of society. *Sociological Review* **14**: 249–67

Marshall G, Newby H, Rose D, Vogler C 1988 *Social Class in Modern Britain*. Unwin Hyman, London

Marx K, Engels F 1968 *Collected Works in One Volume*. Lawrence and Wishart, London

Pahl R E 1989 Is the emperor naked? *International Journal of Urban and Regional Research* **13**: 709–20

Parkin F 1979 *Marxism and Class Theory: A Bourgeois Critique*. Columbia University Press, New York

Poulantzas N 1975 *Classes in Contemporary Capitalism*. New Left Books, London

Rose D, O'Reilly K 1997 *Constructing Classes: Towards a New Social Classification for the UK*. ESRC and ONS, Swindon, UK

Stewart A, Prandy K, Blackburn R M 1980 *Social Stratification and Occupations*. MacMillan, London

Weber M 1968 *Economy and Society*. Bedminster Press, New York

Wilkinson R 1996 *Unhealthy Societies: The Afflictions of Inequality*. Routledge, London

Wright E 1985 *Classes*. Verso, London

P. Saunders

Classical Archaeology

The notion of 'classical archaeology' is of relatively recent origins in the vocabulary of the humanities. Its common usage dates to the second half of the nineteenth century, when archaeology, a newly formed university discipline, sought its place within the broader intellectual framework of the sciences of antiquity, the *Altertumswissenschaft* of German scholarship. For knowledge of antiquity to be scientific, it had to encompass all the creations of human genius. Thus, just as philology had emancipated itself from theology, so was archaeology to free itself from philology in order to constitute, together with the latter and with ancient history, the third pillar of *Altertumswissenschaft*. This disciplinary triangle effectively constitutes the foundations of the modern conception of the classical past—that is, the past of the Greco-Roman world.

1. The Definition of Classical Archaeology

Such a definition of classicism is quite evidently arbitrary. It rests on a very specific and thoroughly European experience of antiquity, in which Greco-Roman literature and its concepts are overwhelmingly privileged. In specifying the chronological and geographical boundaries of the classical world, nineteenth century historians and archaeologists have confirmed a spatio–temporal distinction already initiated during antiquity, and notably in the Alexandrian golden age of the third and second centuries BC. Following this definition, the classical world effectively begins with the dispersal of the Greeks in the Mediterranean during the eighth century BC, and ends with the fall of the occidental Roman empire in 476 AD. In terms of its geographical extension, the classical world reaches as far as the farthest expansion of the Roman empire, at the end of the first century AD.

As can be seen, this mode of historical representation clearly focuses on the Mediterranean as the zone of contact between Europe, Africa and Asia. But where does this Mediterranean space actually ends? Is

it somewhere in the vast European plains? On the shores of the Black Sea? At the foot of the mountain ranges bordering Anatolia or Africa? According to their readings of both the classical sources and the surrounding landscapes, experts have drawn different boundaries to the classical world. Given this diversity of opinions, German archaeologists have proposed the notion of *Randkulturen* to designate all those civilizations known to have been in contact with the classical world without fully merging with it: Scythians, Parthians, Germans, Phoenicians, Egyptians, and many others.

Chronology is not much better established. When the Archaic epoch was first identified, archaeologists knew nothing of the Minoan world and the palace civilizations of Greek antiquity, civilizations which later on would end being called preclassical. At the other end of the scale, the date commonly chosen to signal the conclusion of Greco-Roman history—the fall of the occidental Roman empire in 476 AD—proves to be equally problematic: it takes no account of the persistence of the oriental Roman empire until the fall of Constantinopolis in 1453.

2. *The History of Classical Archaeology*

With its uncertain chronological and geographical boundaries (a problem admittedly shared with protohistory), classical archaeology stands somewhat apart from other, more global, strands of archaeology. It is both the oldest among these archaeologies, and the one most influenced by tradition: the ancient Greek word *archaiologia* retains within it all its meaning as a discourse on antiquity. It was, after all, in fifth century Greece that the historical genre first appeared, the reasoned discourse on the past to which the Greeks gave the name of *historia* (enquiry). And since Varron in the first century BC, the title of antiquary has been given to the man who seeks to interpret and classify ancient objects or monuments. In Greece as well as Rome, these antiquaries contributed to elucidate the past; they assembled series of objects and monuments, collected inscriptions, and then put them in order and sought to understand them.

In this respect, the collapse of the Occidental Roman Empire was not without consequence, for with it a whole social class of knowledge- producers and users came to disappear. Secular men of letters were gradually replaced by clerics, whose function was to educate the masses and the elite in the Christian faith. To be sure, there were among these clerics some antiquarians who undertook to collect and investigate the monuments of antiquity. But one has to wait until the end of the Middle Ages to find an antiquarian movement comparable to that of Greco-Roman times.

It was in Renaissance Italy that a new passion for antiquity emanated which was also a means for reinventing the contemporary world. Collections of Greek and Roman manuscripts arose passions, ancient coins were coveted, monuments begun to be excavated for the sculptures and architectural remains they may have contained. The past that was sought after and revived was that of Greece and Rome, if only because a reference to this world was seen as a means to do away with a present seen as barbarous, illiterate and indeed 'gothic.'

For Renaissance scholars, antiquity was quintessentially Greco-Roman: the literary works, sculptures and architecture of these periods were seen as unsurpassable productions, to be studied and emulated by the best contemporary artists. This classical focus dominated scholarly and artistic production until the end of the eighteenth century, and it explains that 'barbarian' antiquities received only marginal interest. The Renaissance has thus contributed to conflating the knowledge of the past with the investigation and understanding of the Greco-Roman heritage—the part of universal culture believed to be 'classic' by virtue of being the veritable pedestal of humanist scholarship. There were some scholars who manifested considerable interest for the oriental world, Egypt or even Mesopotamia, and others, particularly in Scandinavia, Britain and Germany, who paid attention to local antiquities, but the model for the organization of knowledge derived directly from the prestigious Greco-Roman tradition.

Renaissance scholars and their Enlightenment successors assembled together coins, inscriptions, sculptures, and sometimes even ceramics or other objects of daily life. These collections were displayed in 'cabinets of curiosity' and eventually came to adorn the first public museums, such as the Venetian collections and the Ashmolean Museum in Oxford (from 1683). However, since no careful observations were made regarding the place of discovery and the mode of manufacture of these collected works and monuments, it was only with the greatest difficulty that scholars were able to classify them in terms of their dates and provenance. In fact, it was only the antiquarians of the eighteenth century who were first able to lay the grounds of a critical analysis of archaeological monuments.

These advances occurred at different paces, and they were due both to discoveries in the field and to new attitudes towards finds and their study. In this respect, the fortuitous discovery of the buried cities of Pompeii and Herculaneum at the beginning of the eighteenth century did much to awaken the interest of the elite in the Greco-Roman world. What was being unearthed there by the excavating teams of the King of Naples were no longer usual ruins, but entire cities, covered by volcanic eruptions and thus miraculously preserved from the ravages of destruction. Closely guarded by the Neapolitan government, the exploration of Pompeii gave rise to an unprecedented craze for the painting, architecture and implements of daily

life of the time. Given the context of their discoveries, these items could also for the first time be attributed a secure chronological position. Naples thus played the role of an antiquarian capital, and gradually came to compete with Rome in the esteem of travelers and connoisseurs.

Without interpretation, however, discovery is nothing. In the second part of the eighteenth century, two men revolutionized the understanding of antiquity. The first of these was Johannes Winckelmann (1717–68), a German scholar from a small Prussian town who achieved tremendous influence with his work. His *History of Art in Antiquity*, published in Leipzig in 1764, offered to amateurs of Greco-Roman art the first ever systematic and chronological treatise on the subject. Aided by a thorough familiarity with textual sources, Winckelmann depicted and commented on antique objects scattered throughout European collections. Drawing on his personal study of the most important Roman collections, he could propose to his readers a remarkable history of plastic forms in antiquity. The style of this work, the quality of its descriptions, and the philosophical spirit which animated it, all explain its unprecedented success: with it, the occident could at last discover the sources of Renaissance art in the classical world. Winckelmann, otherwise an enemy of the aristocracy and determined critic of the ancient regime, received his highest acclaim from the courts of kings and princes, among those who sought antiquity for the aesthetic pleasure it procured them.

One of the most notable aristocrats in the court of France, the Comte de Caylus (1692–1765), was instrumental in drawing the attention of his contemporaries to another aspect of the study of the past. For Caylus, what justified the quest for antiquity is not the aesthetic, but the technical achievements of the past. He therefore developed an approach to the study of past techniques in which can be recognized our current concerns with artifact morphology and archaeometry. Each in his way, Winckelmann for the appreciation of art and Caylus for the understanding of techniques, laid down the bases for a history of the *oeuvres* of antiquity, a research programme developed and followed throughout the nineteenth century.

3. *The Rules of Classical Archaeology*

The exploration of antiquity thus gradually transformed itself into a new discipline of archaeology. Under this framework, nineteenth century scientists elaborated and imposed new rules for the extraction, analysis and publication of evidence from the past. Archaeology distanced itself from collections and from the social milieu of collecting, and advocated instead the necessity of observing finds in the field, and of recognizing them as coherent wholes open to rigorous analysis. While the antiquarian channeled his passion for the past to the collection of objects, the archaeologist aimed to valorise these objects and monuments in the context of their discovery. This approach found its most systematic expression in Germany with the research programme of E. Gerhard (1795–1867), who effectively fought for an autonomous archaeology, free from collectors, philologists and artists. Archaeologists must take the place of connoisseurs with their various intermediaries, and themselves seek for the evidence on the ground.

They must also distance themselves from philologists, and challenge the primacy of written over nonwritten sources. Finally, they must take leave of aesthetic considerations, and study ancient productions in all their material and technical dimensions. Building on advances in historical studies, this positivist program also drew on developments in geology and natural history to forge for itself its specific scientific instruments.

Three complementary pillars came then to form then backbone of archaeology. One, *stratigraphy*, consists of observing the conditions of deposition of objects and monuments in the ground. The second, *typology*, seeks to identity in the objects themselves some evidence regarding the time and place of their manufacture. The third, *technology*, deals with the modes of production, the raw materials, the procedures of fabrication. It was in fact only by the end of the nineteenth century, when these three strands were mastered and marshaled together, that archaeology became a fully-fledged discipline able to undertake the reliable and methodical investigation of the past. The body of doctrines and institutions which accompanied this development contributed to giving archaeology, and specifically to classical archaeology, its distinctive identity. In this respect E. Gerhard created a novel kind of institution which prefigures in many ways those of today: the *Instituto di corrispondenza archeologica*, established in 1829 in Rome through the activities of aristocrats and scholars from across Europe, took it as its aim to collect, excavate and publish as thoroughly as possible all the antiquities of the Mediterranean world.

By drawing archaeologists to the field, by elaborating a modern strategy of publication based on detailed sections and descriptions, the *Instituto* virtually launched the new discipline that was embodied in this dedicated research institution located in Rome, one of the capitals of the classical world.

Until the middle of the nineteenth century, the exploration of the Greco-Roman world had been a matter of individual enterprise with occasional royal support. From then on, archaeology would become what the discoverer of Greek Asia Minor T. Wiegand (1864–1936) has called a *Grosswissenschaft*, a state-sponsored science. The leading European powers established then various institutes or schools in Rome and Athens, and these encouraged archaeological

expeditions throughout the main urban sites of the Mediterranean. The enormous success encountered by the private initiatives of the German businessman Heinrich Schliemann (1822–90) at Mycenae and Troy fueled a veritable scramble for new excavating territories among the countries of Europe and even America. This competition contributed notably to the growth of collections in the main European and American museums. The rise of such museums as the British Museum, the Louvre, the Berlin museums and even the Metropolitan Museum in New York, all catering for an ever-increasing public, corresponded with the ascendancy of a classical archaeology closely entwined with the economic developments of capitalism and colonialism. The conquest of the past was effectively an instrument of foreign policy in the hands of all main powers, beginning with the core area of the Greco-Roman world and reaching towards the Orient, Africa, and the globe as whole.

Archaeological schools, museums and, of course, universities were all part of this expansion, from which they have benefited. Whereas in the mid nineteenth century only Germany was endowed with a network of university Chairs in classical archaeology, by 1914 all European nations as well as the United States had established university curricula in classical archaeology. This three-pronged movement, involving the development of museum collections, the systematic excavation of sites, and the setting of an academic theoretical and educational framework, effectively gave to classical archaeology its modern appearance.

4. *The Paths of Classical Archaeology*

Set out during the second half of the nineteenth century, this disciplinary model witnessed during the first half of the twentieth century a considerable expansion and consolidation in both qualitative and quantitative terms. Alongside a marked increase in the number of excavations carried out, the main archaeological institutions followed the German model by systematically publishing catalogues of finds and collections according to specific descriptive protocols. This increased and refined considerably the existing knowledge on artistic and artisanal productions: sculptures, architecture, ceramics were all researched, and so were numismatics and glyptics. Thus classical archaeology forged itself a vast corpus of systematic reference, bearing equally on chronology, typology and technology.

Some criticisms and contradictions emerged, however, during the second half of the twentieth century. The implicit supremacy of Greco-Roman culture came under challenge; the comparative study of cultures raised questions over the Hellenocentrism characteristic of most Classical archaeologists, while advances in the study of mentalities promoted a critique of Eurocentrism. The Greek and Roman worlds are now understood as melting-pots rather than *hochkulturen*. Under the influence of economic history, scholarly interest has shifted from a history of art in the narrow sense towards a wider history of productions. The better chronological grasp allowed by typological studies has served to advance studies on colonization, on long distance contacts and trade. Studies of the relations between centre and periphery, between Greeks, Roman and 'barbarians' have benefited from developments in the field of protohistory. It may indeed be said that the research program of classical archaeology has undergone a wide-reaching redefinition during the 1960s.

No longer considered as the elder of the sciences of antiquity, it is now one discipline among others, sharing resources and addressing issues similar to other strands of archaeology. In this context, such topics as the history of the landscape, or that of the movements of populations and the exchanges of goods and ideas, all open up new perspectives which contribute to extricate classical archaeology from its somewhat marginal position.

At the same time, advances in the iconography and the sociology of representations make it possible to cast a new light and reach a more satisfying understanding of otherwise well-known bodies of evidence. Thus, building on the strength of its centuries-long traditions as a science dedicated to Ancient art, Classical archaeology is developing now into to a human science in the full sense of the term.

See also: Christianity Origins: Primitive and 'Western' History; Chronology, Stratigraphy, and Dating Methods in Archaeology; Historiography and Historical Thought: Classical Period (Especially Greece and Rome)

Bibliography

Bérard C, Vernant J-P 1989 *A City of Images. Iconography and Society in Ancient Greece*. Princeton University Press, Princeton, NJ

Bordein A, Hölscher T, Zanker P (eds.) 2000 *Klassiche Archaeologie, eine Einführung*. Dieter Reimer Verlag, Berlin, Germany

Elsner J A 1955 *Art and the Roman Viewer, the Transformation of Art from the Pagan World to Christianity*. Cambridge University Press, Cambridge, UK

Goldhill S, Osborne R (eds.) 1994 *Art and Text in Ancient Greek Culture*. Cambridge University Press, Cambridge, UK

Marchand S L 1996 *Down from Olympus. Archaeology and Philhellenism in Germany*. Princeton University Press, Princeton, NJ

Morris I (ed.) 1994 *Classical Greece, Ancient Histories and Modern Archaeologies*. Cambridge University Press, Cambridge, UK

Parslow C C 1998 *Rediscovering Antiquity, Karl Weber and the Excavations of Herculaneum, Pompeii and Stabiae*. Cambridge University Press, Cambridge, UK

Schnapp A 1997 *The Discovery of the Past*. Harry Abrams, New York
Thomson de Grummond N (ed.) 1996 *An Encyclopedia of the History of Classical Archaeology*. Greenwood Press, Westport, CT

A. Schnapp

Classical Conditioning and Clinical Psychology

Classical conditioning occurs when neutral stimuli are associated with a psychologically significant event. The main result is that the stimuli come to evoke a set of responses or emotions that may contribute to many clinical disorders, including (but not limited to) anxiety disorders and drug dependence. Research on conditioning has uncovered many surprising details about the underlying learning process, as well as methods for eliminating emotional and behavioral problems.

1. *Historical Background*

Classical conditioning was first studied systematically at the turn of the twentieth century by the Russian physiologist, Ivan Pavlov (Pavlov 1927). In the usual description of his best-known experiment, Pavlov rang a bell and then gave a dog some food. After a few pairings of bell and food, the dog began to salivate to the bell, and thus anticipated the presentation of food. The classical conditioning phenomenon quickly attracted psychologists who applied it to clinical issues. Before most of Pavlov's work was available in English, John Watson showed that human emotions are also influenced by classical conditioning (Watson and Rayner 1920). Watson showed an infant boy a stimulus (which happened to be a laboratory rat) and then made a frightening noise. After a few pairings of the rat and the noise, the child became afraid whenever the rat was presented. This fear generalized to a rabbit, a dog, and a fur coat. Watson saw conditioning as a means by which emotions could be elicited by an expanding range of cues.

The application of conditioning to clinical issues became central to the behavior therapy movement that began in the 1950s and 1960s (e.g., Wolpe 1958, see also *Behavior Therapy: Psychological Perspectives*). The idea was that psychiatric disorders could be understood and treated using scientifically established principles of learning. The view that anxiety disorders result from classical conditioning was later criticized (e.g., Rachman 1977). However, many of the criticisms were directed at an obsolete view of conditioning, the science of which continued to advance after behavior therapy began (see below).

2. *Behavioral Consequences of Classical Conditioning*

The events in Pavlov's experiment are often described using terms designed to make the experiment applicable to any situation. The food is the 'unconditional stimulus,' or US, because it unconditionally elicits salivation before the experiment begins. The bell is known as the 'conditional stimulus,' or CS, because it only elicits the salivary response conditional on the bell–food pairings. The new response to the bell is correspondingly called the 'conditional response' (CR), while the natural response to the food itself is the 'unconditional response' (UR).

Culture has created the impression that conditioning is a rigid affair in which a fixed event comes to elicit a fixed response. In fact, conditioning is more complex and dynamic than that. For example, signals for food may evoke a large set of responses that prepare the organism to digest food: They can elicit secretion of gastric acid, pancreatic enzymes, and insulin in addition to the famous salivary response. The CS can also elicit approach behavior, an increase in body temperature, and a state of arousal and excitement. When a signal for food is presented to a quiescent, food-replete animal, the animal may get up and eat more food. Signals for food evoke a whole 'behavior system' that is functionally organized to deal with the meal (Timberlake 1994).

Classical conditioning is also involved in other aspects of eating. Through conditioning, humans may learn to like or dislike different foods. In infrahuman animals, flavors associated with nutrients (sugars, starches, calories, proteins, or fats) come to be preferred. Flavors associated with sweet tastes are also preferred, while flavors associated with bitter tastes are avoided. At least as important, flavors associated with illness become disliked, as illustrated by the person who gets sick drinking tequila and consequently learns to hate the flavor. The fact that flavor CSs can be associated with such a range of biological consequences (USs) is important for omnivorous animals that need to learn about new foods. And it has clinical implications. For example, chemotherapy can make cancer patients sick, and can therefore cause the conditioning of an aversion to a food that was eaten recently (or to the clinic itself). And conditioning can enable external cues to trigger food consumption and craving, a potential influence on overeating and obesity.

Classical conditioning also occurs when we ingest drugs. Whenever a drug is taken, it constitutes a US, and it may be associated with potential CSs that are present at the time (rooms, odors, injection rituals, etc.). CSs that are associated with drug USs can have

an interesting property: They often elicit a conditioned response that seems opposite to the unconditional effect of the drug (Siegel 1989). For example, although morphine causes a rat to feel less pain, a CS associated with morphine elicits an opposite increase, not a decrease, in pain sensitivity. Similarly, although alcohol can cause a drop in body temperature, a conditioned response to a CS associated with alcohol is typically an increase in body temperature. In these cases, the conditioned response is said to be 'compensatory' because it counteracts the drug effect. Compensatory responses are another example of how classical conditioning helps organisms prepare for a biologically significant US.

Compensatory conditioned responses have implications for drug abuse. First, they can cause drug tolerance, in which repeated administration of a drug reduces its effectiveness. As a drug and a CS are repeatedly paired, the compensatory response to the CS becomes stronger and more effective at counteracting the effect of the drug. The drug therefore has less impact. One implication is that tolerance will be lost if the drug is taken without being signaled by the usual CS. Consistent with this idea, administering a drug in a new environment can cause a loss of drug tolerance and make drug overdose more likely (see Siegel 1989). A second implication stems from the fact that compensatory responses may be unpleasant. A CS associated with an opiate may elicit several compensatory responses—it may cause the drug user to be more sensitive to pain, undergo a change in body temperature, and perhaps become hyperactive (the opposite of another unconditional morphine effect). The unpleasantness of these responses may motivate the user to take the drug again to get rid of them. Compensatory responses may often resemble withdrawal effects (Siegel 1989). The idea is that the urge to take drugs may be strongest in the presence of CSs that have been associated with the drug. The hypothesis is consistent with self-reports of abusers who, after a period of abstinence, are tempted to take the drug again when they are re-exposed to drug-associated cues.

Classical conditioning is also involved in anxiety disorders, as Watson originally envisioned. We now know that CSs associated with frightening USs can elicit a whole system of conditioned fear responses, again broadly designed to help the organism cope. In animals, cues associated with frightening events elicit a set of natural defensive reactions that have evolved to prevent attack by predators. They also elicit changes in respiration, heart rate, and blood pressure, and even a (compensatory) decrease in sensitivity to pain. Brief CSs that occur close to the US in time can also elicit adaptively timed protective reflexes. For example, the rabbit blinks to a brief signal that predicts a mild electric shock near the eye. The same CS, when lengthened in duration and paired with the same US, elicits mainly fear responses. And fear elicited by a CS may potentiate the conditioned eyeblink response elicited by another CS or a startle response to a sudden noise. Once again, CSs do not merely elicit a simple reflex, but evoke a complex and interactive set of responses.

Classical fear conditioning can contribute to phobias (where specific objects may be associated with a traumatic US) as well as other anxiety disorders. For example, in panic disorder, people who have unexpected panic attacks can become anxious about having another one (see *Panic Disorder*). In this case, the panic attack (the US or UR) may condition anxiety to the external situation in which it occurs (e.g., a crowded bus) and also internal ('interoceptive') CSs created by early symptoms of the attack (e.g., dizziness or a sudden pounding of the heart). These CSs may then come to evoke anxiety or panic responses. Panic disorder may begin because external cues associated with panic can arouse anxiety, which may then exacerbate the next unconditional panic attack and/or panic response elicited by an interoceptive CS (Bouton et al. 2001). Interestingly, the emotional reactions elicited by CSs may not require conscious awareness for their occurrence or development. Indeed, fear conditioning may be independent of conscious awareness (e.g., LeDoux 1996).

In addition to eliciting conditioned responses, CSs also motivate ongoing behavior. For example, presenting a CS that elicits anxiety can increase the vigor of instrumental or operant behaviors that have been learned to avoid or escape the frightening US. Thus, an individual with panic disorder will be more likely to express avoidance in the presence of anxiety cues. Similar effects may occur with CSs that predict other USs (such as drugs or food)—as already mentioned, a drug-associated CS may motivate the drug abuser to take more drugs. The potential influence of classical conditioning on behavior is thus extensive and ubiquitous.

3. *The Learning Process in Classical Conditioning*

Modern research has revealed some important details about the learning process that underlies classical conditioning (Rescorla 1988). For example, conditioning is not an inevitable consequence of pairing a CS with a US. Such pairings will not cause conditioning if there is a second CS present that already predicts the US (Kamin 1969). This sort of finding ('blocking') suggests that a CS must provide new information about the US if learning is to occur. Many theorists now suppose that conditioning is determined by the discrepancy between (a) the US predicted by all CSs present on a trial and (b) the US that actually happens on the trial (Rescorla and Wagner 1972). One implication is that, depending on the size and direction of the discrepancy, the pairing of a CS and traumatic

US, for example, can cause an increase in fear conditioning, no change in conditioning, or even a decrease in conditioning.

The latter implication is interesting. A new CS can acquire a negative value if the US is smaller than that which other CSs present predict. Casually speaking, the new CS predicts 'less US than expected.' Such negative signals are called 'conditioned inhibitors' because they inhibit performance elicited by other CSs. They are clinically relevant because they may hold pathological CRs like anxiety at bay. A loss of the inhibition would allow the anxiety response to emerge.

Classical conditioning is most robust if the CS and US are intense or salient. It is also best if the CS and US are novel. For example, in 'latent inhibition,' repeated exposure to the CS alone before conditioning can diminish its ability to elicit responding when it is paired with the US. In the 'US pre-exposure effect,' repeated exposure to the US before conditioning can likewise decrease the conditioning that later occurs when a CS and the US are paired. One idea is that the CS and the US must be 'surprising' at the time of their pairing for learning to occur. Thus, the effects of pairing a CS with trauma or drug USs may depend in subtle ways on the individual's prior experience with the CS and US.

There are important variants of classical conditioning. In 'sensory preconditioning,' two stimuli (A and B) are first paired, and then one of them (A) is later paired with the US. Stimulus A evokes conditioned responding, of course, but so does stimulus B—indirectly, through its association with A. One implication is that exposure to a potent US like a panic attack may influence our reactions to stimuli that have never been paired with the US directly; the sudden anxiety response to stimulus B might seem spontaneous and mysterious. A related finding is 'second-order conditioning.' Here, A is paired with a US first and then subsequently paired with stimulus B. Once again, both A and B will evoke responding. Sensory preconditioning and second-order conditioning increase the range of stimuli that can control conditioned responses.

Emotional responses can also be conditioned through observation. For example, a monkey that merely observes another monkey being frightened by a snake can learn to be afraid of the snake itself (Mineka 1992). The observer learns to associate the snake (CS) with its own emotional reaction (US/UR) to the other monkey being afraid. Although monkeys readily learn to fear snakes, they are less likely to associate other salient cues (such as colorful flowers) with fear in the same way. This is an example of 'preparedness' in classical conditioning—some stimuli are especially effective signals for some USs because evolution has made them that way. (Another example is the fact that tastes are easily associated with illness but not shock, whereas auditory and visual cues are easily associated with shock but not illness.) Preparedness may explain why human phobias tend to be for certain objects (snakes or spiders) and not others (knives or electric sockets) that may as often be paired with pain or trauma.

4. *'Unlearning' in Classical Conditioning*

Once one accepts a role for conditioning in behavioral and emotional disorders, the question becomes how to eliminate it. Pavlov studied 'extinction': conditioned responding decreases if the CS is presented repeatedly without the US after conditioning. Extinction is the basis of many therapies that reduce pathological conditioned responding through repeated exposure to the CS. Another elimination procedure is 'counterconditioning,' in which the CS is paired with a very different US/UR. Counterconditioning was the inspiration for 'systematic desensitization,' a behavior therapy technique in which frightening CSs are deliberately associated with relaxation during therapy (Wolpe 1958).

Although extinction and counterconditioning reduce unwanted conditioned responses, they do not destroy the original learning, which remains in the brain, ready to return to behavior under the right circumstances. For example, conditioned responses that have been eliminated by extinction or counterconditioning can recover if time passes before the CS is presented again ('spontaneous recovery'). Conditioned responses can also recover if the patient returns to the original context of conditioning (the general situation, mood, or state in which conditioning occurred), or if the current context is associated with the US (Bouton 2000). All of these phenomena are potential mechanisms for relapse. Techniques that may minimize relapse include conducting therapy in the contexts where the disorder is a problem, conducting therapy in multiple contexts, or providing the client with retrieval cues or retrieval strategies that help recall therapy (Bouton 2000). In the long run, contemporary research on extinction and counterconditioning may suggest ways to optimize their clinical effectiveness.

5. *Challenges*

The idea that classical conditioning is a basis of behavior disorders, particularly anxiety disorders, has not gone unchallenged. However, most challenges have been directed at early versions of conditioning theory that did not recognize factors such as information value, latent inhibition, preparedness, or context. For instance, Rachman (1977) noted that London air raids during World War II did not cause an increase in anxiety disorders despite their emotional impact. We now know that air raids might not have caused conditioning if the potential CSs were familiar

rather than novel, if the raids were experienced in the presence of safety cues that inhibited fear (e.g., relatives, bomb shelters), or if the raids were signaled by other cues (e.g., sirens) that could have 'blocked' conditioning of potential CSs (Kamin 1969). Another criticism was that fears in the general population are disproportionately directed toward things like snakes, a fact that is consistent with the preparedness principle (above). As a final example, critics of conditioning explanations of panic disorder have asked why a CS like a pounding heart does not elicit panic in the context (say) of athletic exercise, or why extinction exposure to the CS without panic during exercise does not eliminate its ability to cause panic in other situations. The answer may be that the loss of responding that occurs in extinction is especially specific to the context in which it is learned (Bouton et al. 2001). Thus, although fear may extinguish in the context of exercise, that extinction will not abolish the CS's ability to elicit fear in other contexts, such as a crowded bus or a shopping mall. The fact that conditioned responses generalize more across contexts than extinction does may be a reason why many disorders are so persistent.

6. *Future Directions*

Basic research on classical conditioning will continue to investigate the circumstances that allow (and prevent) conditioning to occur, how conditioning is represented in memory and the brain, and how it ultimately influences cognitions, emotions, and behavior. Research will also address how extinction and other behavior-elimination procedures work and can be improved. That research may eventually explain why conditioning processes like extinction seem more context-specific than conditioning itself. And it will also eventually provide a more complete account of the factors that determine the nature and form of conditioned responses, and how the many responses that CSs can evoke can also interact and interrelate.

As our understanding of classical conditioning continues to deepen and expand, so will our insight into its possible role in various clinical disorders. Understanding the role of conditioning in causing a disorder, however, will probably also require prospective studies in which clinical investigators observe conditioning trials as they naturally occur in the world and then measure their effects as the disorder actually develops. It will also benefit from a better appreciation of how conditioning processes interact with biological factors (such as genetically linked vulnerabilities) that may also play a role. And it will benefit from a better understanding of how conditioning interacts with cognitive factors such as thoughts, beliefs, and awareness. Although biological and cognitive factors are sometimes viewed as alternatives to simple learning processes like classical and operant conditioning, a complete appreciation of any disorder will probably require a more integrative perspective on how these factors work in combination.

In the meantime, classical conditioning remains a surprisingly rich scientific phenomenon that can be expected to come into play whenever people experience significant emotional and biological events and can associate them with other events in their world.

See also: Autonomic Classical and Operant Conditioning; Behavior Therapy: Psychological Perspectives; Classical Conditioning, Neural Basis of; Clinical Psychology: Animal Models; Fear Conditioning; Operant Conditioning and Clinical Psychology; Panic Disorder; Pavlov, Ivan Petrovich (1849–1936); Watson, John Broadus (1878–1958)

Bibliography

Bouton M E 2000 A learning theory perspective on lapse, relapse, and the maintenance of behavior change. *Health Psychology* **19** (Suppl.): 57–63

Bouton M E, Mineka S, Barlow D H 2001 A modern learning theory perspective on the etiology of panic disorder. *Psychological Review* **108**: 4–32

Kamin L J 1969 Predictability, surprise, attention, and conditioning. In: Campbell B A, Church R M (eds.) *Punishment and Aversive Behavior*. Appleton-Century-Crofts, New York, pp. 279–96

LeDoux J 1996 *The Emotional Brain*. Simon and Schuster, New York

Mineka S 1992 Evolutionary memories, emotional processing and the emotional disorders. *The Psychology of Learning and Motivation* **28**: 161–206

Pavlov I P 1927 *Conditioned Reflexes*. Oxford University Press, Oxford, UK

Rachman S 1977 The conditioning theory of fear-acquisition: A critical examination. *Behavior Research and Therapy* **15**: 375–87

Rescorla R A 1988 Pavlovian conditioning: It's not what you think it is. *American Psychologist* **43**: 151–60

Rescorla R A, Wagner A R 1972 A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black A H, Prokasy W F (eds.) *Classical Conditioning II*. Appleton-Century-Crofts, New York, pp. 64–99

Siegel S 1989 Pharmacological conditioning and drug effects. In: Goudie A J, Emmett-Oglesby M W (eds.) *Psychoactive Drugs*. Humana Press, Clifton, NY pp. 115–80

Timberlake W 1994 Behavior systems, associationism, and Pavlovian conditioning. *Psychonomic Bulletin and Review* **1**: 405–20

Watson J B, Rayner R 1920 Conditioned emotional reactions. *Journal of Experimental Psychology* **3**: 1–14

Wolpe J 1958 *Psychotherapy by Reciprocal Inhibition*. Stanford University Press, Stanford, CA

M. E. Bouton

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Classical Conditioning, Neural Basis of

Classical conditioning, a phenomenon described by Pavlov around 1900, is an elementary form of associative learning that is considered to be an essential building block for complex learning. This article will present some essential characteristics of classical conditioning that permit this learning to be a model system *par excellence* for understanding the neurobiology of how the brain encodes, stores and retrieves memory.

1. Introduction and Historical Background

Classical or Pavlovian conditioning is the simplest form of associative learning by which animals, including humans, learn relations among events in the world so that their future behaviors are better adapted to their environments (Rescorla 1988). Generally, classical conditioning ensues when an initially neutral stimulus (conditional stimulus, CS) is paired in close temporal proximity with a biologically significant stimulus (unconditional stimulus, US) that elicits an unlearned, reflexive behavior (unconditional response, UR). Through CS–US association formation, the animal learns to exhibit a learned behavior (conditional response, CR) to the CS that generally (a) resembles the UR (but not always), (b) precedes the US in time, and (c) reaches a maximum at about the time of US onset. A typical classical conditioning arrangement is represented in Fig. 1. [It should be noted that the contingent (informational), rather than contiguous (temporal), relationship between CS and US is essential in classical conditioning. For detailed treatment of this see Kamin 1968, Rescorla, 1968, Wagner et al. 1968.]

Classical conditioning was first characterized by Ivan P. Pavlov, a Russian physiologist. [The English translation of Pavlov's book *Conditioned Reflexes* was first published in 1927. Classical conditioning was also independently discovered by an American Psychologist, Edwin B. Twitmyer.] Having already conducted prominent work on the digestive system, for which he received a Nobel Prize in 1904, Pavlov employed a salivary-conditioning procedure in dogs to systematically characterize some of the fundamental principles of classical conditioning. In brief, Pavlov's dogs were presented with a discrete CS (e.g., the beat of a metronome) just prior to the delivery of a US (e.g., meat powder). Initially, the subjects did not respond to the CS but salivated profusely (UR) to the US. With repeated CS–US pairings, Pavlov's dogs exhibited salivation (CR) to the CS that both preceded US onset and occurred in the absence of the US.

2. Behavioral Principles of Classical Conditioning

Since Pavlov, various types of classical conditioning procedures have been developed, ranging from a potently fast (one-trial) taste aversion conditioning

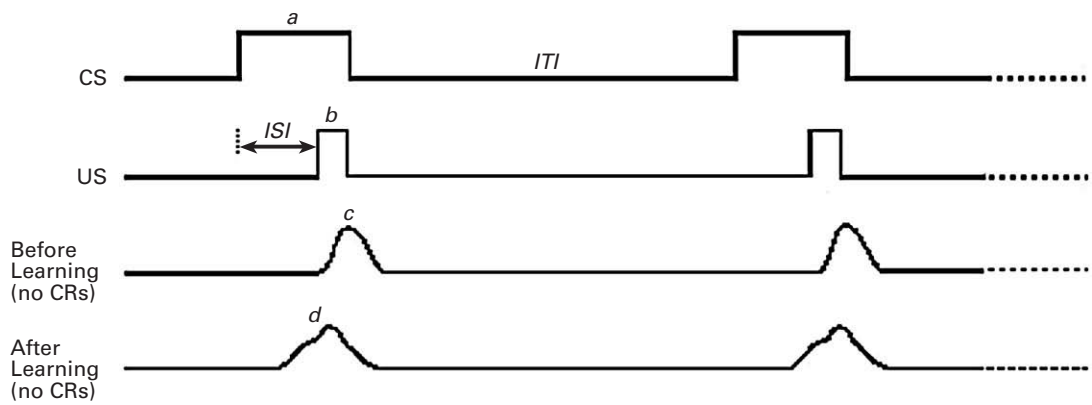

Figure 1
A typical classical conditioning procedure requires that the CS and the US are paired in close temporal proximity. Presentation of CS and US may be scheduled according to two different temporal arrangements: (i) In delay conditioning, CS onset procedes US onset, such that the two stimuli overlap and terminate together; (ii) In trace conditioning, the CS and the US are separated by some 'empty' interval in which neither stimuli is present. The top two traces indicate presentation of the CS and the US in a delay conditioning arrangement. The bottom two traces depict behavioral responses before learning and after learning. (*a*), CS duration; (*b*), US duration; (*c*), UR; (*d*), CR + UR; *ISI* (Inter-stimulus Interval), the time between the CS onset and the US onset; *ITI* (Inter-trial Interval), the time between the termination of one CS–US pairing and the beginning of the next CS–US pairing

(e.g., Garcia et al. 1974) to a relatively slow and incremental eyeblink conditioning (e.g., Gormezano et al. 1983), and a wide variety of organisms have been used, ranging from invertebrates (e.g., *Aplysia*) to primates (e.g., humans). These different types of classical conditioning, however, all share certain common factors that influence the formation of CS–US associations which can be classified into three general categories: *contiguity* (or temporal) constraints, *sensory* constraints, and *contingency* (or informational) constraints (Table 1). Clearly, these constraints must be considered when employing classical conditioning in learning and memory research.

In addition to simple CS–US pairings (also called first order conditioning), there are many other training protocols within classical conditioning that can serve as effective tools for investigating the theoretical and biological mechanisms of learning and memory. Some of these are presented in Fig. 2.

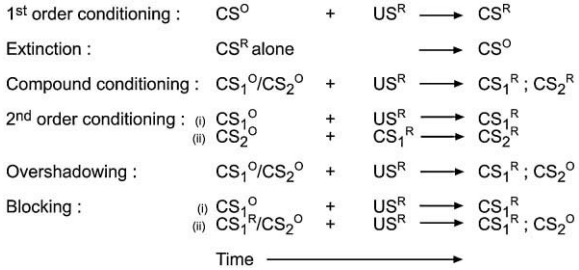

Figure 2
Various classical conditioning phenomena. CS^O = no or small magnitude CR to CS; CS^R = CR to CS; US^R = UR to US; CS_1/CS_2 = simultaneous presentations of the two CSs; in Overshadowing, CS_1 is more salient than CS_2; 2nd order conditioning is often elusory and transient when it does appear

3. *Classical Conditioning as a Model System for Studying the Neurobiology of Learning and Memory*

Current views recognize that, in mammals, there are multiple forms or aspects of learning and memory (e.g., habituation, sensitization, classical conditioning, priming, procedural learning, episodic, semantic) that are subserved by different structures in the central nervous system (e.g., reflex pathways, cerebellum, amygdala, striatum, diencephalon, medial temporal lobe, neocortex) (Squire 1987). Considerable progress has been made in the understanding of these different learning and memory systems through the use of various experimental techniques (e.g., lesions, reversible inactivation, drug administration, neural recordings, genetic manipulations, brain imaging). The success of this work is due in large part to the use of classical conditioning as a model system, since it provides an important advantage over other, more complex, forms of learning in that the stimuli involved

Table 1
Behavioral principles of classical conditioning

Contiguity constraints	Sensory constraints	Informational constraints
ISI function. Conditioning occurs when the CS and the US are paired within a well-defined temporal window. The optimum conditionable ISI can range anywhere from a fraction of a second in eyeblink conditioning to hours in taste aversion conditioning.	**Repetition.** The magnitude of conditioning is proportional to to the number of times the CS–US pairings are presented.	**Contingency.** Organisms learn not only about the temporal relationship between stimuli but also about the predictive relationship between the stimuli. In general, if the CS is positively correlated with the US, excitatory conditioning will occur. In contrast, if the CS is negatively correlated with the US, then inhibitory conditioning will follow.
Timing. The peak onset of the CR generally corresponds to the time at which the US is presented. This temporal process allows the organism to anticipate and respond adaptively to the occurrence of the US.	**Familiarity.** The more familiar the CS and/or the US are to the organism, the slower the rate of conditioning (CS pre-exposure and US pre-exposure effects). The asymptote of conditioning, however, remains the same.	**Genetic predisposition.** Conditioning is robust when the CS and the US are relevant to each other. Some CS–US associations do not readily occur.
ITI function. For a fixed ISI, conditioning appears to be more efficient with longer ITIs. As the ITI lengthens, the possibility of the ISI and ITI interacting and thus interfering with the CS–US temporal processing decreases.	**Intensity.** Generally, the more intense (or salient) the CS and/or the US, the stronger the conditioning. The use of a more intense US will result in a faster rate as well as a higher asymptotic conditioning. The use of a more salient CS will also result in a faster rate of conditioning but the asymptote of conditioning will be the same.	

(CS and US) are well defined and can be precisely controlled, and the behavioral output is discrete and may be accurately assessed.

3.1 Localization of Brain Substrates of Classical Conditioning: Rationale

The identification of the locus of learning and memory storage in the brain is a prerequisite to an understanding of the neurochemical, cellular and molecular mechanisms by which organisms acquire and retain information. For many years the task of localizing memory storage (the *engram*) has been the paramount challenge facing investigators of memory mechanisms in mammalian systems. Classical conditioning is especially attractive as a model system for use in these types of investigations since only two stimuli are involved, and thus the learning or association of CS and US must occur at the brain site(s) where the two pieces of information converge. For instance, in classical eyeblink conditioning, where animals (such as mice, rats, rabbits or humans) learn to exhibit eyeblink responses to a CS (e.g., tone) that has been paired with a US (e.g., air-puff to the eye), one can trace the pathways from the peripheral sensory receptors in the ear (for CS) and around the eye (for US) to the brain and examine those brain regions where the CS and US pathways converge. It is only in the past 20 years, however, that technology has permitted this type of analysis, and only then at the gross structural level.

The convergence of CS and US information, however, is not a sufficient condition for identifying the locus of learning. In order for a particular region of the brain to be considered a viable candidate as a learning and memory site, it must demonstrate the following criteria:

(a) permanent lesions of the putative site prior to CS–US training should completely and permanently abolish the acquisition of the CR;

(b) permanent lesions made after training should completely abolish the expression of the CR, and the CR should not be reacquired with further CS–US pairings;

(c) reversible inactivation during CS–US training should block the development of the CR such that when the structure is activated again the CR should develop comparably to that of a naïve animal (i.e., no evidence of *savings*);

(d) reversible inactivation following training should temporarily impair the expression of the CR;

(e) learning-related neural activities should occur that correspond with and immediately precede the behavioral CR;

(f) electrical stimulation of the CS and the US input pathways to the putative site should effectively substitute for the peripheral stimuli and support conditioning; and

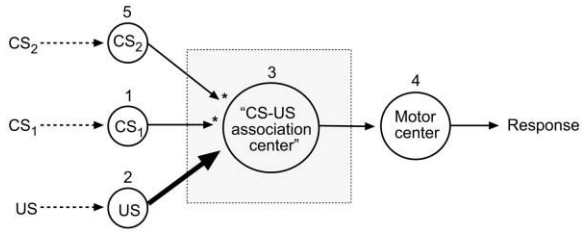

Figure 3
A highly schematized diagram of a hypothetical learning site in classical conditioning. The shaded box represents the locus of conditioning. * denotes modifiable connections underlying learning

(g) electrical stimulation of the putative site should evoke the CR.

If there is a single locus of learning that supports classical conditioning, then these seven criteria must be demonstrable within that specific learning site; whereas if there are multiple structures that encode CS–US association, then these seven criteria must be satisfied collectively by these structures. Figure 3 illustrates how these different criteria can be applied to a putative locus of CS–US association.

In Fig. 3, the afferent CS and US information is relayed via *hypothetical* structures 1 (and 5 for different CSs) and 2 (for US), respectively, to the learning site 3. The outputs (efferents) from structure 3 activate the motor center 4 that in turn controls the CR. Permanent lesions of structures 1, 2, 3 or 4 prior to conditioning will block the acquisition and/or expression of the CR. Thus, the permanent lesion technique (e.g., electrical, chemical, radio-frequency, aspiration, ischemia) is limited in that it does not allow for dissociation of the site of learning from the sites of input or from motor centers. In contrast, the reversible inactivation (pharmacological, cooling) technique, which temporarily inactivates the neurons within a structure, is not so constrained. For example, reversible inactivation of structures 1, 2 or 3 during conditioning will block acquisition of the CR to CS_1. In subsequent CS–US training when the inactivation has been reversed, the animal should learn as though it is naïve. By contrast, reversible inactivation of structure 4 at the time of CS–US training will block the expression of the CR, but once the inactivation is removed, the animal will immediately exhibit CRs to the CS because the CS–US association center was not affected during conditioning.

Thus structure 4 must be efferent to the site of learning. Structures 1, 2 and 3 can be further dissociated by examining the effects of reversible inactivation to CSs of different sensory modalities. Whereas inactivation of structure 2 and 3 will block conditioning to all CS modalities (e.g., CS_1 and CS_2), inactivation of structure 1 will block conditioning to CS_1 but not CS_2. Thus structure 1 must be afferent to

the site of learning. Finally, structures 2 and 3 can be dissociated by examining the reversible inactivation effects following conditioning: once the animal has acquired the CR, inactivating structure 3 should abolish the expression of CR, whereas inactivation of structure 2 should not interfere with the expression of the CR because the CS inputs to the CS–US association center remain intact. Interestingly, with continued CS–US training, inactivation of structure 2 will lead to extinction of the CR because the CS–US association center will, as a result, receive only the CS information (just as in CS-alone extinction training). If structures 1 and 2 relay information about the CS and the US, then electrical stimulation of the CS and US pathways should also support conditioning. Lastly, recordings from structure 3 (the site of learning) should reveal learning-related changes in neural activity that model the behavioral CR, whereas recordings from CS and US input structures should show stimulus-evoked neural activities.

The neural circuitry of the brain is almost infinitely complex and thus no single experimental technique is in itself sufficient to identify the site of learning. It is also important to note that different types of classical conditioning (e.g., eyeblink versus fear conditioning) are subserved by different neural circuits. Through the utilization of various techniques, however, the structures and mechanisms involved in this type of learning can be reasonably delineated and systematically analyzed for its validity.

3.2 Putative Cellular Mechanisms of Classical Conditioning

Once the learning and memory storage site has been identified, a logical next step is to determine what neural changes take place that allow a CS that previously did not elicit a CR (before conditioning) to now evoke a CR (after conditioning). It is generally assumed that an initially weak connection between a CS relaying structure and a CS–US association structure becomes strengthened as a function of CS–US paired training, such that the CS becomes able to effectively activate the CS–US association structure, which in turn activates the CR pathway. It is further hypothesized that changes in synaptic efficacy (that is, the efficacy with which a neuron is able to communicate with another neuron) underlie this type of strengthening of the CS relay-to-CS–US association connection (e.g., Hebb's postulate (Hebb 1949)). Two forms of experimentally-induced synaptic plasticity, long-term potentiation (LTP) and long-term depression (LTD), have received close scrutiny as the most promising cellular mnemonic mechanisms (Bliss and Collingridge 1993; Ito 1989). LTP and LTD refer to sustained increase and decrease of synaptic transmission, respectively, following different stimulation patterns of afferent fibers. In brief, LTP is characterized by its rapid inducibility and longevity (lasting of the order of hours *in vitro* to weeks *in vivo*), as well as its being strengthened by repetition and demonstrating specificity and associativity. LTD displays similar characteristics desirable of an information storage mechanism. Both LTP and LTD have been demonstrated in various brain structures, including those that are hypothesized to be critical for learning and memory.

One can easily imagine how LTP and LTD might be applicable to classical conditioning. For example, suppose that the CS pathway to the site of learning (pathway between steps 1–3 in Fig. 3) is initially weak and that, as a result, the CS alone cannot sufficiently activate the CS–US association center to produce a CR. Following CS–US pairings, if this pathway is strengthened via LTP (or LTP-like changes), then the CS alone will be able to activate the CS–US association center to elicit a CR. Similarly LTD or LTD-like changes may support conditioning by weakening the inputs to a structure that normally inhibits the CS–US association center (as is postulated to occur in eyeblink conditioning circuit). It is likely that combinations of LTP- and LTD-like forms of synaptic plasticity (and perhaps other unknown cellular changes), involving a network of synapses, are important in classical conditioning.

3.3 Neuronal Substrates of Eyeblink Conditioning

Classical eyeblink conditioning in rabbits has been used extensively to investigate brain mechanisms underlying learning and memory. Converging lines of evidence from lesion, recording, stimulation, reversible inactivation, and brain imaging studies indicate that the cerebellum mediates the formation of the CS–US association for eyeblink conditioning (Thompson 1990). In brief, selective lesions of the cerebellum (i.e., the interpositus nucleus) block the acquisition and retention of eyeblink CRs; the lesion is limited to the CR since the UR (reflexive eyeblink) to the US is not affected. (An important feature of eyeblink conditioning is that the CR and the UR can be dissociated and, thus, effects of various manipulations on memory versus performance can be carefully addressed.) Correspondingly, recording studies indicate that cells in specific regions of the cerebellum undergo plastic changes during eyeblink conditioning; for example, cells in the interpositus nucleus increase their activity (postulated to occur via LTP-like changes), while Purkinje cells in the cortex, which send inhibitory projections to the interpositus nucleus, decrease their activity (postulated to occur via LTD-like changes).

The involvement of the cerebellum in eyeblink conditioning is also evidenced by stimulation studies which show that direct stimulation of the two major afferents to the cerebellum, the mossy fibers from

the pontine nucleus and the climbing fibers from the inferior olive, can substitute for the peripheral CS and US, respectively. Since limited lesions of the pontine nucleus (i.e., the lateral region) abolish CRs to a tone CS but not to a light CS, and lesions of the inferior olive in animals that already acquired CRs result in behavioral extinction with continued CS–US-paired training, it is not likely that these afferent structures are the site of learning and memory storage. Reversible inactivation studies further support the conclusion that the cerebellum, and not its efferent structures, is the locus of CS–US association. Moreover, recent human brain-imaging studies reveal eyeblink conditioning-related activity changes in the cerebellum. Collectively, these findings strongly suggest that the cerebellum is essential for eyeblink conditioning. Although the cerebellum seems to be critical, the relative importance of the cerebellar cortex and the interpositus nucleus in supporting eyeblink conditioning is not clear and has been disputed (see Kim and Thompson 1997 for detailed treatment of this topics and a putative eyeblink conditioning circuit). The fact that the critical CS–US association in eyeblink conditioning occurs within the cerebellum, however, permits research at the molecular level of analysis.

3.4 Neuronal Substrates of Fear Conditioning

In fear conditioning, CSs such as tones, lights or experimental chambers are typically paired with aversive US such as electric shock. Following CS–US pairings, the CS can elicit numerous fear CRs, such as an increase in blood pressure, reduction in pain sensitivity (analgesia), fear-potentiated startle, and/or defensive freezing. Several lines of evidence point to the amygdala, one of the principle structures of the limbic system that seems to be situated such that it has access to both sensory inputs and response outputs, as a critical neural substrate for this type of emotional learning (LeDoux 1996). In brief, the critical role of the amygdala in fear conditioning is supported by observations that:

(a) amygdalar lesions (permanent and reversible) abolish various fear CRs as well as innate fear responses;

(b) selective lesions of structures afferent to the amygdala affect conditioning to specific CSs (e.g., medial geniculate nucleus of the thalamus for tones, and the hippocampus for contexts);

(c) selective lesions of structures efferent to the amygdala abolish specific CRs (e.g., the lateral hypothalamus for blood pressure CR, and the ventral region of the periaqueductal gray matter for freezing CR);

(d) recording studies reveal that neurons in the amygdala respond to both CS and US and undergo plastic changes during fear conditioning (e.g., LTP-like changes);

(e) electrical stimulation of the amygdala elicits fear responses; and

(f) drugs that block LTP also prevent fear conditioning.

However, there is also evidence that the amygdala may not necessarily be involved in fear learning and memory, and that other brain structures (e.g., insular cortex) may mediate fear conditioning (McGaugh et al. 1996). Thus, additional studies are required to firmly establish the role of the amygdala in fear conditioning.

4. Conclusion

Although classical conditioning is well understood at a behavioral level and the neuroanatomical circuits that underlie it are beginning to be unveiled, there is much left to learn. However, this most basic form of associative learning offers a useful means to investigate the synaptic and molecular mechanisms underlying learning and memory and may help reveal common biological mechanisms shared by all learning and memory systems.

See also: Autonomic Classical and Operant Conditioning; Behavioral Assessment; Cardiovascular Conditioning: Neural Substrates; Conditioning and Habit Formation, Psychology of; Eyelid Classical Conditioning; Fear Conditioning; Learning and Memory, Neural Basis of; Long-term Depression (Hippocampus); Long-term Potentiation and Depression (Cortex); Long-term Potentiation (Hippocampus); Memory, Consolidation of; Pavlov, Ivan Petrovich (1849–1936)

Bibliography

Garcia J, Hankins W G, Rusiniak K W 1974 Behavioral regulation of the milieu interne in man and rat. *Science* **185**: 824–31

Gormezano I, Kehoe E J, Marshall B S 1983 Twenty years of classical conditioning research with the rabbit. In: Sprague J M, Epstein A N (eds.) *Progress in Psychobiology and Physiological Psychology*. Academic Press, New York

Hebb D O 1949 *The Organization of Behavior: A Neuropsychological Theory*. John Wiley and Sons, New York

Ito M 1989 Long-term depression. *Annual Review of Neuroscience* **12**: 85–102

Kamin L J 1968 Attention-like processes in classical conditioning. In: Jones M R (ed.) *Miami Symposium on the Prediction of Behavior: Aversive Stimulation*. University of Miami Press, Miami, FL, pp. 9–32

Kim J J, Thompson R F 1997 Cerebellar circuits and synaptic mechanisms involved in classical eyeblink conditioning. *Trends in Neurosciences* **20**: 177–181

LeDoux J 1996 *The Emotional Brain*. Simon ans Schuster, New York

McGaugh J L, Cahill L, Roozendaal B 1996 Involvement of the amygdala in memory storage: Interaction with other brain systems. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 13508–14
Pavlov I P 1927 *Conditioned Reflexes*. Oxford University Press, London
Rescorla R A 1968 Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology* **66**: 1–3
Rescorla R A 1988 Behavioral studies of Pavlovian conditioning. *Annual Review of Neuroscience* **11**: 329–52
Squire L R 1987 *Memory and Brain*. Oxford University Press, NY
Thompson R F 1990 Neural mechanisms of classical conditioning in mammals. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **329**: 161–70
Wagner A R, Logan F A, Haberlandt K, Price T 1968 Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology* **76**: 171–80

J. J. Kim

Classical Mechanics and Motor Control

In order to control the execution of limb movements, the central nervous system must solve complex problems of mechanics. A substantial body of evidence supports the view that, in solving these problems the nervous system develops internal representations of the mechanics of the body coupled with its environment. Thus, through the process of motor learning (see *Motor Control*) the brain becomes implicitly an 'expert' in classical mechanics. After discussing the problems of kinematics and dynamics associated with the control of movement, this article introduces the formal definitions and the empirical evidence that constitute the underpinnings of the theory of internal representations in motor control.

1. Dynamics

According to the laws of Newtonian physics, if one wants to impress a motion upon an object with mass m, one must apply a force, F, that is directly proportional to the desired acceleration, a. This is Newton's equation:

$$F = ma$$

A desired motion may be expressed as a sequence of positions, x, that one wishes the object to occupy at subsequent instants of time, t. Such a sequence is called a trajectory and is mathematically represented as a function, $x = x(t)$. To use Newton's equation for deriving the needed time-sequence of forces, one must calculate the first temporal derivative of the trajectory, the velocity, and then the second temporal derivative, the acceleration. Finally, one obtains the desired force from this acceleration. The above calculation is an example of an inverse dynamic problem. The direct dynamic problem is that of computing the trajectory resulting from the application of a force, $F(t)$. Direct problems are a common challenge for physicists who are concerned, for example, with predicting the motion of a comet from the known pattern of gravitational forces. Unlike physicists, the brain deals most often with inverse problems: we routinely recognize objects and people from their visual images—an 'inverse optical problem'—and we find out effortlessly how to distribute the forces exerted by several muscles to move our limb in the desired way: an inverse dynamics problem.

One of the central questions in motor control (see *Motor Control*) is how the central nervous system may solve the inverse dynamics problem and generate the motor commands that guide our limbs (Hollerbach and Flash 1982). In the biological context, however, the inverse dynamic problem assumes a somewhat more complex form than the one described above. A system of second-order nonlinear differential equations is generally considered to be an adequate representation for the passive dynamics of a limb. A compact expression for such a system is:

$$D(q, \dot{q}, \ddot{q}) = \tau \tag{1}$$

where q, $\dot{q}$, and $\ddot{q}$ represent the limb configuration vector—for example the vector of joint angles—and its first and second time derivatives. D is a non-linear vector, valued mapping from the current state and its rate of change to the vector of joint torques, τ. In practice, the expression for D may have a few terms for a two-joint planar arm, or it may take several pages for more realistic models of the arm's multi-joint geometry. The inverse dynamics approach to the control of multi-joint limbs consists in solving explicitly for a torque trajectory, τ(t), given a desired trajectory of the limb, $q_D(t)$. This is done by replacing $q_D(t)$ for the variable q on the left side of Eqn. (1):

$$\tau(t) = D(q_D(t), \dot{q}_D(t), \ddot{q}_D(t)) \tag{2}$$

2. Kinematics, Statics, and Coordinate Systems

A significant computational challenge comes from the need to perform changes of representation—or, more technically, coordinate transformations—between the description of a task and the specification of the body motions. Tasks, such as 'hitting a ball with a racket,' are described most efficiently and parsimoniously with respect to some fixed reference points in the environment. In this example, the racket is the site at

which one interacts with the environment. Borrowing some terminology from robotics, such a site is called an 'endpoint.' The position of the racket is fully determined by six coordinates. These coordinates may be measured with respect to three orthogonal axes originating, for example, from the shoulder. Then, a position in endpoint coordinates may be specified as a point $p = (x, y, z, \theta_X, \theta_Y, \theta_Z)$. The coordinates, x, y, and z determine a translation with respect to the orthogonal axes. The angular coordinates, θ_X, θ_Y, and θ_Z determine an orientation with respect to the same axes. Consistent with this notation, a force in endpoint coordinates is a vector with three linear and three angular components, $F = (F_X, F_Y, F_Z, \tau_X, \tau_Y, \tau_Z)$.

A different way of describing the position of an arm is to provide the set of joint angles that define the orientation of each skeletal segment either with respect to fixed axes in space or with respect to the neighboring segments. Joint angles are a particular instance of generalized coordinates. According to the standard definitions of analytical mechanics, generalized coordinates are independent variables, which are suitable for describing the dynamics of a system (see for example Goldstein 1980).

Once we have defined a set of generalized coordinates we may also define a set of corresponding generalized forces. For example, if we use joint angles as generalized coordinates, the corresponding generalized forces are the torques measured at each joint. The dynamics of any mechanical system with N generalized coordinates are described by N second order differential equations relating the generalized coordinate to their first and second time derivatives and to the generalized forces. The dynamics Eqn. (1) is an example of formulation in generalized coordinates.

Movements are executed by the central nervous system activating of a multitude of muscles. Muscle coordinates afford the most direct representation for the motor output of the central nervous system. A position in this coordinate system is described by a collection of muscle lengths, $l = (l_1, l_2, \ldots, l_M)$. Accordingly, a force is a collection of muscle tensions, $f = (f_1, f_2, \ldots, f_M)$.

Both the transformations from generalized coordinates to endpoint coordinates, and from generalized coordinates to actuator coordinates are nonlinear mappings. In the case of the arm, the transformation from joint to endpoint coordinates is a nonlinear function:

$$p = L(q) \tag{3}$$

The transformation from joint to muscle coordinates is another nonlinear mapping:

$$l = M(q) \tag{4}$$

Some experimental studies (see Flash and Hogan 1985, Morasso 1981) have suggested that actions are planned by the brain in terms of endpoint coordinates. For example, one can mentally formulate (and execute) commands such as 'move the hand 10 cm to the right' without being concerned with the set of muscle commands that are involved with this action. However, once one has decided a plan of action one must somehow choose which muscles to activate and in what temporal order. In carrying out this task the brain must faces the challenges associated with kinematic redundancy: the imbalance between the number of joints that may participate in a movement, the number of degrees of freedom of the hand, and the number of independently controlled muscles acting upon the joints. There are typically fewer hand coordinates than joint angles and fewer joint angles than muscles. Such imbalance renders both transformations (3) and (4) non-invertible.

3. *Internal Models of Limb Dynamics*

The ability to generate a variety of complex behaviors cannot be attained by just storing somewhere the control signals for each action and recalling these signals when subsequently needed. Simple considerations about the geometrical space of meaningful behaviors are sufficient to establish that this approach would be inadequate (see Bizzi and Mussa-Ivaldi 1998). To achieve its typical competence, the motor system must take advantage of experience for going beyond experience itself, by constructing internal representations of the controlled dynamics. These representations allow us to generate new behaviors and to handle situations that have not yet been encountered (see *Motor Control Models: Learning and Performance*). A vivid illustration of how explicit representations of dynamics, also called internal models, may facilitate motor learning is offered by work of Atkeson and Schaal (1997, Schaal 1999) who studied the task of balancing an inverted pendulum on the hand of a robotic arm. They found that robots learn to carry out this task successfully when they can build an internal model of the dynamics associated with the balancing act. Such a model may be constructed using data derived from the observation of humans engaging competently in the same task.

The term 'internal model' refers to two distinct mathematical transformations: (a) the transformation from a motor command to the consequent behavior; and, (b) the transformation from a desired behavior to the corresponding motor command (see Kawato and Wolpert 1998). A model of the first kind is called a 'forward model.' Forward models provide the control system with the means not only to predict the outcome of a command, but also to estimate the current state in the presence of feedback delay. A representation of the mapping from planned actions to motor commands is called an 'inverse model.' Strong experimental evidence for the biological and behavioral relevance of internal models has been offered by experiments that involved the adaptation of arm movements to a

perturbing force field generated by an instrumented manipulandum (Sabes et al. 1998, Shadmehr and Mussa-Ivaldi 1994). The major findings of these studies are as follows: (a) when exposed to a complex but deterministic field of velocity-dependent forces, arm movements are first distorted and, after repeated practice, the initial kinematics are recovered; (b) if, after adaptation, the field is suddenly removed, after effects are clearly visible as mirror images of the initial perturbations; (c) adaptation is achieved by the motor system through the formation of a local map that associates the states (positions and velocities) visited during the training period with the corresponding forces; and, (d) after adaptation this map—that is the internal model of the field—undergoes a process of consolidation (see Brashers-Krug et al. 1996).

4. *The Neurobiological 'Building Blocks' of Internal Models*

Once it has been established that the motor system creates internal representations of complex multi-joint dynamics, it remains to determine how these representations may come about. As pointed out by David Marr (1982), any mathematical transformation may be carried out in different ways depending upon which building blocks or 'primitives' are employed (see *Computational Neuroscience*). Electrophysiological studies involving the stimulation of muscles and of the spinal cord in frogs (Bizzi et al. 1991) indicated: (a) that the stimulation of a site in the lumbar spinal cord results in the activation of multiple muscles acting on the leg on the same side of the stimulation; (b) that concomitant or 'synergistic' muscle recruitment generates a field of viscoelastic forces over a broad region of the leg workspace; and, (c) that the simultaneous activation of multiple spinal sites leads to the vectorial summation of the corresponding force fields. These and similar findings suggest that motor commands reaching the spinal cord from higher brain centers are not directed at controlling the forces of individual muscles or single joint torques. Instead, the descending motor commands modulate the viscoelastic force fields produced by specific sets of muscles. These force fields have influence over broad regions of the limb state space as each active muscle within a synergy contribute a significant force over a large range of positions and velocities.

From a mathematical standpoint, the force fields generated by neural modules in the spinal cord are nonlinear functions of limb position and velocity and of time: $\varphi_i(q, \dot{q}, t)$. Consistent with the finding of vector summation, the net force field induced by a pattern of K motor commands may be represented as a linear combination:

$$\sum_{i=1}^{K} u_i \varphi_i(q, \dot{q}, t) \tag{5}$$

In this expression, each spinal field is a force that depends upon the state of motion of the limb, $(q, \dot{q})$ and upon time, t, in a fixed stereotyped way. The descending commands, $(u_1, u_2, \ldots, u_K)$, act as coefficients that modulate the degree with which each spinal field participates in the combination. These commands can just select the modules by determining how much each one contributes to the net control policy. The linear combination (5) generates the torque that drives the limb inertia. Substituting it for $\tau(t)$ in Eqn. (1) one obtains:

$$D(q, \dot{q}, \ddot{q}) = \sum_{i=1}^{K} u_i \varphi_i(q, \dot{q}, t) \tag{6}$$

Least squares approximation can efficiently determine an optimal set of tuning coefficients given a desired trajectory, $q_D(t)$:

$$u_i = \sum_{j=1}^{K} [\Phi]_{i,j}^{-1} \Lambda_j \tag{7}$$

with

$$\begin{cases} \Phi_{l,m} = \int \varphi_l(q_D(t), \dot{q}_D(t), t) \bullet \varphi_m(q_D(t), \dot{q}_D(t), t)\, dt \\ \Lambda_j = \int \varphi_j(q_D(t), \dot{q}_D(t), t) \bullet D(q_D(t), \dot{q}_D(t), \ddot{q}(t))\, dt \end{cases} \tag{8}$$

The symbol $\bullet$ indicates the ordinary inner product.

While spinal force fields offer a practical way to generate movement, they also provide the central nervous system with a movement's representation (see *Neural Representations of Intended Movement in Motor Cortex*). This representation is geometrically similar to the representation of space by a set of Cartesian coordinates. In the latter case, we may take three directions—represented by three independent vectors—and then project any point in space along these directions. As a result, an arbitrary point in space is represented by three numbers, the coordinates x, y, and z. The movements of a limb can be considered as 'points' in an abstract geometrical space. In this abstract geometrical space, the force fields produced by a set of modules play a role equivalent to that of the Cartesian axes and the selection parameters that generate a particular movement may be regarded as generalized projections of this movement along the module's fields.

4.1 *A Computational Approach to Motor Adaptation*

If the dynamics change while the modules remain unchanged, then the representation of the movement must change accordingly. This is shown by the following argument. Suppose that a trajectory, $q(t)$, is

represented by a selection vector $c = (c_1, c_2, \ldots, c_K)$ for a limb with the dynamics of Eqn. (6). Now, suppose that the limb dynamics are suddenly modified by an additional load, $E(q, \dot{q}, \ddot{q})$. Leaving the representation and the fields unchanged we have now a new differential equation

$$D(q, \dot{q}, \ddot{q}) + E(q, \dot{q}, \ddot{q}) = \sum_{m=1}^{K} c_m \cdot \phi_m(q, \dot{q}, t) \quad (9)$$

whose solution is a trajectory $\tilde{q}(t)$, generally different from the original $q(t)$. The original set of coefficients, c_m, now generates the trajectory $\tilde{q}(t)$ and can, accordingly, be considered as its representation within the modified environment, $D+E$. The old trajectory is recovered by changing the selection coefficients to a new set, $c' = c + e$ with

$$E(q(t), \dot{q}(t), \ddot{q}(t)) = \sum_{m=1}^{K} e_m \cdot \phi_m(q(t), \dot{q}(t), t) \quad (10)$$

With these new coefficients, the new dynamics become equivalent to the old dynamics along the original trajectory.

The modified coefficients c' offer a new representation of the old movement $q(t)$ in the altered dynamics. This procedure for forming a new representation and for recovering the original movement is consistent with the empirical observation of aftereffects in load adaptation (see Shadmehr and Mussa-Ivaldi 1994). If the load is removed after the new representation is formed, the dynamics become

$$D(q, \dot{q}, \ddot{q}) = \sum_{m=1}^{K} (c_m + e_m) \cdot \phi_m(q, \dot{q}, t) \quad (11)$$

that can be rewritten as

$$D(q, \dot{q}, \ddot{q}) - \sum_{m=1}^{K} e_m \cdot \phi_m(q, \dot{q}, t) = \sum_{m=1}^{K} c_m \cdot \phi_m(q, \dot{q}, t) \quad (12)$$

Therefore, removing the load with the new representation corresponds approximately to applying the opposite load with the old representation.

Is it necessary for the motor system to modify a movement's representation each time the limb dynamics changes? Or is there a way for restoring the previously existing representations? From a computational point of view, whenever a dynamical change becomes permanent—as when we undergo growth or damage—it would seem convenient for the central nervous system to have the ability to restore the previously learned motor skills (that is, the previously learned movement representations) without need to relearn them all. It is possible for the adaptive system to restore, at least partially, the motor representations that preexist a change in dynamics by modifying the modules and their force fields. A specific modification is obtained when we may express the coefficients $e = (e_1, e_2, \ldots, e_K)$ as a linear transformation of the original coefficients $c = (c_1, c_2, \ldots, c_K)$:

$$e = Wc$$

This transformation is a coordinate transformation of the selection vector and may be implemented by a linear associative network (see *Neural Networks: Biological Models and Applications*). With a minimum of algebra one sees that

$$\begin{aligned} D(q, \dot{q}, \ddot{q}) + E(q, \dot{q}, \ddot{q}) &= \sum_{m=1}^{K} c'_m \cdot \phi_m(q, \dot{q}, t) \\ &= \sum_{m=1}^{K} (c_m + e_m) \cdot \phi_m(q, \dot{q}, t) \\ &= \sum_{m=1}^{K} c_m \cdot \bar{\phi}_m(q, \dot{q}, t) \end{aligned} \quad (13)$$

where the old fields ϕ_m have been replaced by the new fields

$$\bar{\phi}_m = \sum_{l=1}^{K} (\delta_{l,m} + W_{l,m}) \cdot \phi_l \quad \delta_{l,m} = \begin{cases} 1 & \text{if } l = m \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

This is, again, a coordinate transformation of the original fields that may be implemented by a neural network intervening between the descending commands and the original fields. By means of such coordinate transformations, one obtains the important result that the movement representation—that is the selection vector c—can be maintained invariant after a change in limb dynamics.

In conclusion, the system of motor primitives induced by independent modules within the spinal cord—as well as within higher structures of the nervous system—provides us with an alphabet for representing the mechanics of the body and for modifying this representation as required by changes in limb and environmental dynamics.

See also: Motor Control; Motor Control Models: Learning and Performance; Motor Cortex; Motor Skills, Psychology of; Neural Representations of Intended Movement in Motor Cortex

Bibliography

Atkeson C, Schaal S 1997 Robot learning from demonstration. In: Fisher D (ed.) *Machine Learning: Proceedings of the Fourteenth International Conference* (*ICML 97*). Morgan Kaufman, San Francisco

Bizzi E, Mussa-Ivaldi F A 1998 The acquisition of motor behavior. *Daedalus* **127**: 217–32

Bizzi E, Mussa-Ivaldi F A, Giszter S 1991 Computations underlying the execution of movement: A biological perspective. *Science* **253**: 287–91

Brashers-Krug T, Shadmehr R, Bizzi E 1996 Consolidation in human motor memory. *Nature* **382**: 252–5

Flash T, Hogan N 1985 The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience* **5**: 1688–703
Goldstein H 1980 *Classical Mechanics*. Addison–Wesley, Reading, MA
Hollerbach J M, Flash T 1982 Dynamic interactions between limb segments during planar arm movement. *Biological Cybernetics* **44**: 67–77
Kawato M, Wolpert D 1998 Internal models for motor control. *Novartis Foundation Symposium* **218**: 291–307
Marr D 1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W H Freeman and Co, San Francisco, CA
Morasso P 1981 Spatial control of arm movements. *Experimental Brain Research* **42**: 223–7
Sabes P N, Jordan M I, Wolpert D M 1998 The role of inertial sensitivity in motor planning. *Journal of Neuroscience* **18**: 5948–57
Schaal S 1999 Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences* **3**: 233–42
Shadmehr R, Mussa-Ivaldi F A 1994 Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience* **14**: 3208–24

S. Mussa-Ivaldi

Classical (Psychometric) Test Theory

1. Introduction

One of the most striking and challenging phenomena in the Social Sciences is the unreliability of its measurements: Measuring the same attribute twice often yields two different results. If the same measurement instrument is applied twice, such a difference may sometimes be due to a change in the measured attribute itself. Sometimes these changes in the measured attribute are due to the mere fact of measuring. For example, people learn when solving tasks and they change their attitude when they reflect on statements in an attitude questionnaire. In other cases the change of the measured attribute is due to developmental phenomena, or it might be due to learning between occasions of measurement. However, if change of the attribute can be excluded two different results in measuring the same attribute can be explained only by 'measurement error.'

Classical (Psychometric) Test Theory (CTT) aims at studying the reliability of a (real-valued) test score variable (measurement, test) that maps a crucial aspect of qualitative or quantitative observations into the set of real numbers. Aside from determining the reliability of a test score variable itself, CTT allows answering questions such as:

(a) How do two random variables correlate once the measurement error is filtered out (correction for attenuation)?

(b) How dependable is a measurement in characterizing an attribute of an individual unit, i.e., which is the confidence interval for the true score of that individual with respect to the measurement considered?

(c) How reliable is an aggregated measurement consisting of the average (or sum) of several measurements of the same unit or object (Spearman–Brown formula for test length)?

(d) How reliable is a difference, e.g., between a pretest and post-test?

2. Basic Concepts of Classical Test Theory

2.1 Primitives

In the framework of CTT, each measurement (test score) is considered being a value of a random variable Y consisting of two components: a 'true score' and an 'error score.' Two levels, or more precisely, two random experiments may be distinguished: (a) sampling an observational unit (e.g., a person) and (b) sampling a score within a given unit. Within a given unit, the true score is a parameter, i.e., a given but unknown number characterizing the attribute of the unit, whereas the error is a random variable with an unknown distribution. The true score of the unit is defined to be the expectation of this intraindividual distribution.

Taking the across units perspective, i.e., joining the two random experiments, the true score is itself considered to be a value of a random variable (the 'true score variable'). The 'error variable' is again a random variable, the distribution of which is a mixture of the individual units' error distributions. Most theorems of CTT (e.g., Lord and Novick, 1968) are formulated from this across units' perspective allowing talking about the correlation of true scores with other variables, for instance.

More formally, CTT refers to a (joint) random experiment of (a) sampling an observational unit u (such as a person) from a set Ω_U of units (called the population), and (b) registering one or more observations out of a set Ω_O of possible observations. The set of possible outcomes of the random experiment is the set product: $\Omega = \Omega_U \times \Omega_O$. The elements of Ω_O, the observations, might be qualitative (such as 'answering in category a of item 1 and in category b of item 2'), quantitative (such as reaction time and alcohol concentration in the blood), or consisting of both qualitative and quantitative components. In Psychology, the measurements are often defined by test scoring rules prescribing how the observations are transformed into test scores. (Hence, these measurement are also often called 'tests' or 'test score variables.') These scoring rules may just consist of summing initial scores of items (defining a psychological scale) or

Table 1
Basic Concepts of Classical Test Theory

Primitives	
The set of possible events of the random experiment	$\Omega = \Omega_U \times \Omega_O$
Test Score Variables	$Y_i : \Omega \to \mathbb{R}$
Projection	$U : \Omega \to \Omega_U$
Definition of the Theoretical Variables	
True Score Variable	$\tau_i := E(Y_i \mid U)$
Measurement Error Variable	$\varepsilon_i := Y_i - \tau_i$

might be more sophisticated representations of observable attributes of the units. CTT does not prescribe the definition of the test score variables. It just additively decomposes them into true score variables and error variables. Substantive theory and empirical validation studies are necessary in order to decide whether or not a given test score variable is meaningful. CTT only helps disentangling the variances of its true score and error components.

Referring to the joint random experiment described above the mapping $U: \Omega \to \Omega_U$, $U(\omega) = u$, (the unit or person projection) may be considered a qualitative random variable having a joint distribution with the test scores variables Y_i. Most theorems of CTT deal with two or more test score variables (tests) Y_i and the relationship between their true score and error components. (The index i refer to one of several tests considered.)

2.2 The core concepts: True score and error variables

Using the primitives introduced above, the true score variable $\tau_i := E(Y_i|U)$ is defined by the conditional expectation of the test Y_i given the variable U. The values of the 'true score variable' τ_i are the conditional expected values $E(Y_i|U = u)$ of Y_i given the unit u. They are also called the 'true scores' of the unit u with respect to Y_i. Hence, these true scores are the expected values of the intraindividual distributions of the Y_i. The 'measurement error variables' ε_i are simply defined by the difference $\varepsilon_i := Y_i - \tau_i$. Table 1 summarizes the primitives and definitions of the basic concepts of CTT.

3. Properties of True Score and Error Variables

Once the true score variables and error variables are defined a number of properties (see Table 2) can be derived, some of which are known as the 'axioms of CTT.' However, since the work done by Novick (1966) and Zimmerman (1975, 1976) it is well known that all these properties already follow from the definition of true score and error variables. They are not new and independent assumptions as has been originally proposed (e.g., Gulliksen 1950). All equations in Table 2 are no assumptions. They are inherent properties of true scores and errors. Hence, trying to test or falsify these properties empirically would be meaningless in just the same way, as it is meaningless to test whether or not a bachelor is really unmarried. The property of being unmarried is an inherent part or logical consequence of the concept of a bachelor.

Only one of the 'axioms of CTT' does not follow from the definition of true score and error variables: 'uncorrelatedness of errors variables' among each other. Hence, uncorrelatedness of errors has another epistemological status as the properties displayed in Table 2 (the other 'axioms'). Uncorrelatedness of errors is certainly a desirable and useful property; but it might be wrong in specific empirical applications (e.g., Zimmerman and Williams 1977). In fact it is an assumption and it plays a crucial rule in defining models of CTT.

Equation 1 of Table 2 is a simple rearrangement of the definition of the error variable. Equation. 2 shows that the variance of a test score variable, too, has two additive components: the 'variance of the true score variable' and the 'variance of the error variable.' This second property follows from Eqn. 3 according to which a true score variable is uncorrelated with a measurement error variable, even if they pertain to different test score variables Y_i and Y_j. Equation 4 states that the expected value of an error variable is zero, whereas Eqn. 5 implies that the expected value of an error variable is zero within each individual observational unit u. Finally, according to Eqn. 6 the conditional expectation of an error variable is also zero for each mapping of U. This basically means that the expected value of an error variable is zero in each subpopulation of observational units.

3.1 Additional Concepts: Reliability, Unconditional and Conditional Error Variances

Although the true score and error variables defined above are the core concepts of CTT, in empirical applications, the true scores can only be estimated. What is also possible, is to estimate the 'variances' of the true score and error variables in a random sample (consisting of repeating many times the random experiment described earlier). The variance $Var(\varepsilon_i)$ of the measurement error may be considered a gross parameter representing the degree of unreliability. A normed parameter of unreliability is $Var(\varepsilon_i)/Var(Y_i)$, the proportion of the variance of Y_i due to measurement error. Its counterpart is $1 - Var(\varepsilon_i)/Var(Y_i)$, i.e.,

$$Rel(Y_i) := Var(\tau_i)/Var(Y_i) \tag{1}$$

the 'reliability' of Y_i. This coefficient varies between zero and one. In fact, most theorems and most empirical research deal with this 'coefficient of reliability.' The reliability coefficient is a convenient information about the dependability of the measurement 'in one single number.'

Table 2
Properties of True Score and Error Variables Implied by Their Definition

Decomposition of the Variables	$Y_i = \tau_i + \varepsilon_i$	(1)
Decomposition of the Variances	$Var(Y_i) = Var(\tau_i) + Var(\varepsilon_i)$	(2)
Other Properties of True Score and	$Cov(\tau_i, \varepsilon_j) = 0$	(3)
Error Variables implied by their definition	$E(\varepsilon_i) = 0$	(4)
	$E(\varepsilon_i \mid U) = 0$	(5)
	for each (measurable) mapping of U: $E[\varepsilon_i \mid f(U)] = 0$	(6)

In early papers on CTT, reliability of a test has been defined by its correlation with itself (e.g., Thurstone 1931, p. 3). However, this definition is only metaphoric, because a variable always correlates perfectly with itself. What is meant is to define reliability by the correlation of 'parallel tests' (see below). The assumptions defining parallel tests in fact imply that the correlation between two test score variables is the reliability. Note that the definition of 'reliability' via Eqn. (1) does not rest on any assumption other than $0 < Var(Y_i) < \infty$.

'Reliability' is useful to compare different instruments to each other if they are applied in the same population. Used in this way, reliability in fact helps evaluating the quality of measurement instruments. However, it may not be useful under all circumstances to infer the dependability of measures of an individual unit. For the latter purpose one might rather look at the 'conditional error variance' $Var(\varepsilon_i \mid U = u)$ given a specific observational unit u or at the 'conditional error variances' $Var(\varepsilon_i \mid \tau_i = t)$ given the subpopulation with true score $\tau_i = t$.

4. Models of Classical Test Theory

The definitions of true score and error variables have to be supplemented by assumptions defining a model if the theoretical parameters such as the reliability are to be computed by empirically estimable parameters such as the means, variances, covariance, or correlation of the test score variables. Table 3 displays the most important of these assumptions and the most important models defined by combining some of these assumptions.

The assumption (a_1) to (a_3) specify in different ways the assumption that two tests Y_i and Y_j measure the same attribute. Such an assumption is crucial for inferring the degree of reliability from the discrepancy between two measurements of the same attribute of the same person. Perfect identity or 'τ-equivalence' of the two true score variables is assumed with (a_1). With (a_2) this assumption is relaxed: the two true score variables may differ by an additive constant. Two balances, for instance, will follow this assumption if one of them yields a weight that is always one pound larger than the weight indicated by the other balance, irrespective of the object to be weighed. According to Assumption (a_3), the two tests measure the same attribute in the sense that their true score variables are linear functions of each other.

The other two assumptions deal with properties of the measurement errors. With (b) one assumes measurement errors pertaining to different test score variables to be uncorrelated. In (c) 'equal error variances are assumed,' i.e., these tests are assumed to measure equally well.

4.1 Parallel Tests

4.1.1 Definition. The most simple and convenient set of assumptions is the model of 'parallel tests.' Two tests Y_i and Y_j are defined to be parallel if they are τ-equivalent, if their error variables are uncorrelated, and if they have identical error variances. Note that Assumption (a_1) implies that there is a uniquely defined latent variable being identical to each of the true score variables. Hence, one may drop the

Table 3
Assumptions and Some Models of CTT

Assumption used to define some models of CTT	
(a_1) τ-equivalence	$\tau_i = \tau_j$,
(a_2) essemtial τ-equivalence	$\tau_i = \tau_j + \lambda_{ij}$, $\lambda_{ij} \in \mathbb{R}$,
(a_3) τ-congenerity	$\tau_i = \lambda_{ij0} + \lambda_{ij1}\tau_j$, λ_{ij0}, $\lambda_{ij1} \in \mathbb{R}$, $\lambda_{ij1} > 0$
(b) uncorrelated errors	$Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$
(c) equal error variances	$Var(\varepsilon_i) = Var(\varepsilon_j)$.
Models defined by combining these assumptions	
Parallel tests are defined by Assumptions (a_1), (b) and (c).	
Essentially τ-equivalent tests are defined by Assumptions (a_2) and (b).	
Congeneric tests are defined by Assumptions (a_3) and (b).	

Note: The equations refer to each pair of tests Y_i and Y_j of a set of tests $Y_1, \ldots, Y_m$, their true score variables, and their error variables, respectively.

Table 4
The Model of Parallel Tests

Definition	Assumptions (a_1), (b) and (c) of Table 3
Identification	$E(\eta) = E(Y_i)$ $Var(\eta) = Cov(Y_i, Y_j), i \neq j$ $Var(\varepsilon_i) = Var(Y_i) - Cov(Y_i, Y_j), i \neq j$ $Rel(Y_i) = Corr(Y_i, Y_j), i \neq j$
Testability	
in the total population	$E(Y_i) = \mu$ $Var(Y_i) = \sigma_Y^2$ $Cov(Y_i, Y_j) = \sigma_\eta^2$
within each subpopulation	$s\ E^{(s)}(Y_i) = \mu_s$

Note: The indices i and j refer to tests and the superscripts s to a subpopulation. The equations are true for each test Y_i of a set of parallel tests $Y_1, \ldots, Y_m$ or for each pair of two such tests, their true score variables, and their error variables, respectively.

index i and denote this latent variable by η. The assumption of τ-*equivalence* may equivalently be written $Y_i = \eta + \varepsilon_i$, where $\varepsilon_i := Y_i - E(Y_i|U)$.

4.1.2 Identification. For parallel tests the theoretical parameters may be computed from the parameters characterizing the distribution of at least two test score variables, i.e., the theoretical parameters are identified in this model if $m \geqslant 2$. According to Table 4 the expected value of η is equal to the expected value of each of the tests, whereas the variance of η can be computed from the covariance of two different tests. The variance $Var(\varepsilon_i)$ of the measurement error variables may be computed by the difference $Var(Y_i) - Cov(Y_i, Y_j)$, $i \neq j$. Finally, the reliability $Rel(Y_i)$ is equal to the correlation $Corr(Y_i, Y_j)$ of two different test score variables.

4.1.3 Testability. The model of parallel tests implies several consequences that may be tested empirically. First, all parallel tests Y_i have equal expectations $E(Y_i)$, equal variances $Var(Y_i)$, and equal covariances $Cov(Y_i, Y_j)$ in the total population. Second, parallel tests also have equal expectations within each subpopulation (see Table 4).

Note that these hypotheses may be tested separately and/or simultaneously as a single multidimensional hypothesis in the framework of 'simultaneous equation models' via AMOS (Arbuckle 1997), EQS (Bentler 1995), LISREL 8 (Jöereskog and Söerbom 1998), MPLUS (Muthén and Muthén 1998), MX (Neale 1997), RAMONA (Browne and Mels 1998), SEPATH (Steiger 1995), and others. Such a simultaneous test may even include the hypotheses about the parameters in several subpopulations (see Table 4). What is not implied by the assumptions of parallel tests is the equality of the variances and the covariances of the test score variables in subpopulations.

For parallel tests $Y_1 + \cdots + Y_m$ as defined in Table 4, the reliability of the sum score $S := Y_1 + \cdots + Y_m$ may be computed by the 'Spearman-Brown formula for lengthened tests:'

$$Rel(S) = Rel(S/m) = \frac{m \cdot Rel(Y_i)}{1 + (m-1) \cdot Rel(Y_i)}$$

Using this formula the reliability of an aggregated measurement consisting of the sum (or average) of m parallel measurements of the same unit can be computed. For $m = 2$, each with $Rel(Y_i) = 0.80$, for instance

$$Rel(S) = 2 \cdot 0.80/1 + (2-1) \cdot 0.80 \approx 0.89$$

The Spearman–Brown formula may also be used to answer the opposite question. Suppose there is a test being the sum of m parallel tests and this test has reliability $Rel(S)$. What would be the reliability $Rel(Y_i)$ of the m parallel tests? For example, if $m = 2$, what would be the reliability of a test half?

4.2 Essentially τ-equivalent Tests

4.2.1 Definition. The model of essentially τ-equivalent tests is less restrictive than the model of parallel tests. Two tests Y_i and Y_j are defined to be 'essentially τ-equivalent' if their true score variables differ only by an additive constant (Assumption a_2 in Table 3) and if their error variables are uncorrelated (Assumption b in Table 3). Assumption (a_2) implies that there is a latent variable η that is a translation of each of the true score variables, i.e.,

$$\eta = \tau_i + \lambda_i,\ \lambda_i \in \mathbb{R}, \text{ such that } Y_i = \eta + \lambda_i + \varepsilon_i$$
$$\text{where } \varepsilon_i := Y_i - E(Y_i|U) \text{ and } \lambda_i \in \mathbb{R}$$

Also note that the latent variable η is uniquely defined up to a translation. Hence, it is necessary to fix the scale of the latent variable η. This can be done by fixing one of the coefficients λ_i (e.g., $\lambda_1 = 0$) or by fixing the expected value of η [e.g., $E(\eta) = 0$].

Table 5
The Model of Essentially τ-Equivalent Tests

Definition	Assumptions (a_2) and (b) of Table 3
Fixing the scale of η	$E(\eta) = 0$
Identification	$Var(\eta) = Cov(Y_i, Y_j), i \neq j$
	$Var(\varepsilon_i) = Var(Y_i) - Cov(Y_i, Y_j), i \neq j$
	$Rel(Y_i) = Cov(Y_i, Y_j)/Var(Y_i), i \neq j$

Table 5 summarizes the assumptions defining the model and the consequences for identification and testability. In this model, the reliability cannot be identified any more by the correlation between two tests. Instead the reliability is identified by

$$Rel(Y_i) = Cov(Y_i, Y_j)/Var(Y_i), \quad i \neq j.$$

Furthermore, the expected values of different tests are not identical any more within each subpopulation. Instead, the differences between the expected values $E^{(s)}(Y_i) - E^{(s)}(Y_j)$ of two essentially τ-equivalent tests Y_i and Y_j are the same in each and every subpopulation. All other properties are the same as in the model of parallel tests. Again, all these hypotheses may be tested via structural equation modeling.

For essentially τ-equivalent tests $Y_1, \ldots, Y_m$, the reliability of the sum score $S := Y_1 + \cdots + Y_m$ may be computed by the 'Cronbach's coefficient α:'

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{i=1}^{m} Var(Y_i)}{Var(S)}\right). \tag{2}$$

This coefficient is a lower bound for the reliability of S if only uncorrelated errors are assumed.

4.3 τ-Congeneric Tests

4.3.1 Definition. The model of τ-congeneric tests is defined by the Assumptions (a_3) and (b) in Table 3. Hence, two tests Y_i and Y_j are called τ-congeneric if their true score variables are positive linear functions of each other and if their error variables are uncorrelated. Assumption a_3 implies that there is a latent variable η such that each true score variable is a positive linear function of it, i.e.,

$$\tau_i = \lambda_{i0} + \lambda_{i1}\eta, \lambda_{i0}, \lambda_{i1} \in \mathbb{R}, \lambda_{i1} > 0,$$

or equivalently:

$$Y_i = \lambda_{i0} + \lambda_{i1}\eta + \varepsilon_i,$$

where $\varepsilon_i := Y_i - E(Y_i|U)$.

The latent variable η is uniquely defined up to positive linear functions. Hence, in this model, too, it is necessary to fix the scale of η. This can be done by fixing a pair of the coefficients (e.g., $\lambda_{i0} = 0$ and $\lambda_{i1} = 1$) or by fixing the expected value and the variance of η [e.g., $E(\eta) = 0$ and $Var(\eta) = 1$].

Table 6 summarizes the assumptions defining the model and the consequences for identification and testability assuming $E(\eta) = 0$ and $Var(\eta) = 1$. Other ways of fixing the scale of η would imply different formula. As can be seen from the formula in Table 6, the model of τ-congeneric variables and all its parameters are identified if there are at least three different tests for which Assumptions a_3 and b hold. The covariance structure in the total population implied by the model may be tested empirically if there are at least four test score variables. Only in this case the model has fewer theoretical parameters determining the covariance matrix of the test score variables than there elements in this covariance matrix. The implications for the mean structure are testable already for three test score variables provided the means of the test score variables are available in at least four subpopulations.

4.4 Other Models of CTT

The models treated previously are not the only ones that can be used to determine the theoretical para-

Table 6
The Model of τ-Congeneric Tests

Definition	Assumptions (a_3) amd (b) of Table 3
Fixing the scale of η	$E(\eta) = 0$ and $Var(\eta) = 1$
Identification	$\lambda_{i1} = \sqrt{\frac{Cov(Y_i, Y_j)\, Cov(Y_i, Y_k)}{Cov(Y_j, Y_k)}}, i \neq j, i \neq k, j \neq k$
	$Var(\varepsilon_i) = Var(Y_i) - \lambda_{i1}^2$
	$Rel(Y_i) = \lambda_{i1}^2 / Var(Y_i)$
Testability in the total population	$\frac{Cov(Y_i, Y_k)}{Cov(Y_j, Y_k)} = \frac{Cov(Y_i, Y_l)}{Cov(Y_j, Y_l)}, i \neq k, i \neq l, j \neq k, j \neq l$
between subpopulations	$\frac{E^{(1)}(Y_i) - E^{(2)}(Y_i)}{E^{(1)}(Y_j) - E^{(2)}(Y_j)} = \frac{E^{(3)}(Y_i) - E^{(4)}(Y_i)}{E^{(3)}(Y_j) - E^{(4)}(Y_j)}$

Note: The indices i and j refer to tests and the superscripts to one of four subpopulations.

meters of CTT such as reliability, true score variance, and error variance. In fact, the models dealt with are limited to unidimensional models. However, true score variables may also be decomposed into several latent variables. 'Confirmatory factor analysis' provides a powerful methodology to construct, estimate, and test models with multidimensional decompositions of true score variables. Note, however, that not each factor model is based on CTT. For instance, there are one-factor models that are not models of τ-congeneric variables in terms of CTT. A model with one common factor and several specific but uncorrelated factors is a counter example. The common factor is not necessarily a linear function of the true score variables and the specific factors are not necessarily the measurement error variables as defined in CTT.

4.5 Some Practical Issues

Once a measurement or test score has been obtained for a specific individual, one might want to know how dependable that individual measurement is. If the reliability of the measurement is known and if one assumes a normal distribution of the measurement errors which is homogeneous for all individuals, the 95 percent-confidence interval for the true score of that individual with respect to the measurement Y_i can be computed by:

$$Y_i \pm 1.96 \cdot \sqrt{Var(Y_i) \cdot (1 - Rel(Y_i))}. \tag{3}$$

Another result deals with the correlation between two true score variables. If the reliabilities for two test score variables Y_1 and Y_2 are known, and assuming uncorrelated measurement errors, one may compute

$$Corr(\tau_1, \tau_2) = \frac{Corr(Y_1, Y_2)}{\sqrt{Rel(Y_1)} \cdot \sqrt{Rel(Y_2)}}. \tag{4}$$

This eqn. is known as the 'correction for attenuation.'

Another important issue deals with the 'reliability of a difference variable,' for example, a difference between a pretest Y_1 and a posttest Y_2. Assuming equal true score and error variances between pre- and posttest implies identical reliabilities, i.e., $Rel(Y_1) = Rel(Y_2) = Rel(Y)$. If additionally uncorrelated measurement errors are assumed, the reliability $Rel(Y_1 - Y_2) := Var(E(Y_1 - Y_2|U)/Var(Y_1 - Y_2)$ of the difference $Y_1 - Y_2$ may be computed by:

$$Rel(Y_1 - Y_2) = \frac{Rel(Y) - Corr(Y_1, Y_2)}{1 - Corr(Y_1, Y_2)} \tag{5}$$

According to this formula, the reliability of a difference between pre- and posttest is always smaller than the reliability of the pre- and posttest, provided the assumptions mentioned above hold. In the extreme case in which there is no differential change, i.e., $\tau_2 = \tau_1 + \text{constant}$, the reliability coefficient $Rel(Y_1 - Y_2)$ will be zero. Obviously, this does not mean that the change is not dependable. It only means that there is no variance in the change, since each individual changes by the same amount. This phenomenon has lead to much confusion about the usefulness of measuring change (e.g., Cronbach and Furby 1970, Harris 1963, Rogosa 1995). Most of these problems are now solved by structural equation modeling (allowing the include latent change variables such as in growth curve models (e.g., McArdle and Epstein 1987, Willet and Sayer 1996) or, more directly, in true change models (Steyer, Eid, and Schwenkmezger 1997, Steyer, Partchev, and Shanahan, 2000). Models of this kind are no longer hampered by reliability problems and allow the explanation of inter-individual differences in intraindividual change.

5. Discussion

It should be noted that CTT refers to the population level, i.e., to the random experiment of sampling a single observational unit and assessing some of its behavior (see Table 1). CTT does not refer to sampling models that consist of repeating this random experiment many times. Hence, no questions of statistical estimation and hypothesis testing are dealt with. Of course, the population models of CTT have to be supplemented by sampling models when it comes to applying statistical analyses, e.g., via structural equation modeling.

Aside from this more technical aspect, what are the limitations of CTT? First, CTT and its models are not really adequate for modeling answers to individual items in a questionnaire. This purpose is more adequately met by models of item response theory (IRT) which specify how the probability of answering in a specific category of an item depends on the attribute to be measured, i.e., on the value of a latent variable.

A second limitation of CTT is the exclusive focus on measurement errors. 'Generalizability theory' presented by Cronbach et al. (1972) (see also Shavelson and Webb 1991) generalized CTT to include other factors determining test scores.

Inspired by Generalizability Theory Tack (1980), Steyer et al. (1989) presented a generalization of CTT, called 'Latent State-Trait Theory,' which explicitly takes into account the situation factor, introduced formal definitions of states and traits, and presented models allowing to disentangle person, as well as situation and/or interaction effects and from measurement error. More recent presentations are Steyer et al. (1992) as well as Steyer, et al. (1999). Eid (1995, 1996) extended this approach to the normal ogive model for analyses on the item level.

The parameters of CTT are often said to be 'population dependent,' i.e., meaningful only with respect to a given population. This is true for the

variance of the true score variable and the reliability coefficient. The reliability (coefficient) of an intelligence test is different in the population of students than the general population. This is a simple consequence of the restriction of the (true score) variance of intelligence in the population of the students. However, such a restriction neither exists for the true score estimates neither of individual persons nor for the item parameters λ_i of the model of 'essentially τ-equivalent tests,' for instance. Proponents of IRT models have often forwarded the population dependence critique. They contrast it with the 'population independence' of the person and the item parameters of IRT models. However, 'population independence' also holds for the person and the item parameters of the model of essentially τ-equivalent tests, for instance.

In applications of CTT it is often assumed that the error variances are the same for each individual, irrespective of the true score of that individual. This assumption may indeed be wrong in many applications. In IRT models no such assumption is made. However, it is possible to assume different error variances for different (categories of) persons in CTT models as well. In this case, the unconditional error variance and the reliability coefficient are not the best available information for inferring the dependability of individual true score estimates. In this case one should seek to obtain estimates of conditional measurement error variances for specific classes of persons. It is to be expected that persons with high true scores have a higher measurement error variance than those with medium true scores and that those with low true scores have a higher error variance again. (This would be due to 'floor and ceiling effects.') Other patterns of the error variance depending on the size of the true score may occur as well.

Such phenomena do not mean that 'true scores' and 'error scores' would be correlated; only the 'error variances' would depend on the true scores. None of the properties listed in Table 2 would be violated. As mentioned before, the properties listed in Table 2 cannot be wrong in empirical applications. What could be wrong, however, is the true score interpretation of the latent variable in a concrete structural equation model. Misinterpretations of this sort can be most effectively prevented by empirical tests of the hypotheses listed in the testability sections of Tables 4 to 6.

The most challenging critique of many applications of CTT is that they are based on rather arbitrarily defined test score variables. If these test score variables are not well chosen any model based on them is also not well founded. Are there really good reasons to base models on sum scores across items in questionnaires? Why take the sum of the items as test score variables Y_i and not another way of aggregation such as a weighted sum, or a product, or a sum of logarithms? And why aggregate and not look at the items themselves?

There is no doubt that IRT models are more informative than CTT models if samples are big enough to allow their application, if the items obey the laws defining the models, and if detailed information about the items (and even about the categories of 'polytomous items,' such as in 'ratings scales') is sought. In most applications, the decision how to define the test score variables Y_i on which models of CTT are built is arbitrary, to some degree. It should be noted, however, that arbitrariness in the choice of the test score variables cannot be avoided altogether. Even if models are based on the item level, such as in IRT models, one may ask 'Why these items and not other ones?' Whether or not a good choice has been made will only prove in model tests and in validation studies. This is true for models of CTT as well as for models of alternative theories of psychometric tests.

See also: Dimensionality of Tests: Methodology; Factor Analysis and Latent Structure: IRT and Rasch Models; Generalizability Theory; Psychometrics; Reliability: Measurement; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Arbuckle J L 1997 *Amos User's Guide: Version 3.6*. SPSS, Chicago, IL

Bentler P M 1995 *EQS. Structural Equation Program Manual*. Multivariate Software, Encino

Browne M W, Mels G 1998 Path analysis (RAMONA). In: SYSTAT 8.0 -Statistics. SPSS, Inc., Chicago

Cronbach L J, Furby L 1970 How should we measure 'change'—or should we? *Psychological Bulletin* **74**: 68–80

Cronbach L J, Gleser G C, Nanda H, Rajaratnam N 1972 *The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles*. Wiley, New York

du Toit M, du Toit S 2001 Interactive LISREL: User's Guide. Scientific Software International, Chicago

Eid M 1995 *Modelle der Messung von Personen in Situationen* [Models of measuring persons in situations]. Psychologie Verlags Union, Weinheim, Germany

Eid M 1996 Longitudinal confirmatory factor analysis for polytomous item responses: model definition and model selection on the basis of stochastic measurement theory [online]. *http://www.ppm.ipn.uni-kiel.de/mpr/issue1/art4/eid.pdf*: 1999-7-5

Fischer G H, Molenaar I W 1995 *Rasch Models*. Springer, New York

Gulliksen H 1950 *Theory of Mental Tests*. Wiley, New York

Jöreskog K G, Sörbom D 1998 *LISREL 8. Users Reference Guide*. Scientific Software, Chicago

Lord F M, Novick M R 1968 *Statistical Theories of Mental Test Scores*. Addison Wesley, Reading, MA

Muthén L, Muthén B 1998 *Mplus User's Guide*. Muthén & Muthén, Los Angeles

Neale M C 1997 MX: Statistical Modeling, 4th edn. Department of Psychiatry, Richmond, VA

Novick M R 1966 The axioms and principal results of classical test theory. *Journal of Mathematical Psychology* **3**: 1–18

Rogosa D 1995 Myths and methods: 'Myths about longitudinal

research' plus supplemental questions. In: Gottman J M (ed.) *The Analysis of Change*. Lawrence Erlbaum, Mahwah, NJ
Shavelson R J, Webb N M 1991 *Generalizability Theory. A Primer*. Sage, Newbury Park
Steiger J H 1995 Structural equation modeling. In: STATISTICA 5—Statistics II. StatSoft Inc., Tulsa, OK
Steyer R, Ferring D, Schmitt M J 1992 States and traits in psychological assessment. *European Journal of Psychological Assessment* **8**: 79–98
Steyer R, Majcen A-M, Schwenkmezger P, Buchner A 1989 A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. *Anxiety Research* **1**: 281–99
Steyer R, Partchev I, Shanahan M 2000 Modeling true intra-individual change in structural equation models: the case of poverty and children's psychosocial adjustment. In: Little T D, Schnabel K U, Baumert J (eds.) *Modeling Longitudinal and Multiple-Group Data: Practical Issues, Applied Approaches, and Specific Examples*. Erlbaum, Hillsdale, NJ, pp. 109–26
Steyer R, Schmitt M, Eid M 1999 Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*
Tack W H 1980 Zur Theorie psychometrischer Verfahren: Formalisierung der Erfassung von Situationsabhängigkeit und Veränderung [On the theory of psychometric procedures: formalizing the assessment of situational dependency and change]. *Zeitschrift für Differentielle und Diagnostische Psychologie* **1**: 87–106
Thurstone L L 1931 *The Reliability and Validity of Tests*. Edwards Brothers, Ann Arbor
Zimmerman D W 1975 Probability spaces, hilbert spaces, and the axioms of test theory. *Psychometrika* **40**: 395–412
Zimmerman D W 1976 Test theory with minimal assumptions. *Educational and Psychological Measurement* **36**: 85–96
Zimmerman D W, Williams R H 1977 The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology* **16**: 135–52

R. Steyer

Classification and Typology (Archaeological Systematics)

Classification is the initial means through which we impose a degree of order on the enormously diverse remains of the human past. As such, it is probably the single most basic analytical procedure employed by the archaeologist. Excavation yields an enormous diversity of materials that are not self-labeling; they must be endowed with identity and meaning by the excavator or the analyst. This is done in the first instance through classification.

1. Classification and Typology

Archaeologists often use the terms classification and typology interchangeably, but in this article a distinction will be made. A classification is any set of formal categories into which a particular field of data is partitioned, while a typology is a particular type of rigorous classification, in which a field of data is divided up into categories that are all defined according to the same set of criteria, and that are mutually exclusive. As will be shown later, most archaeological classifications of artifacts are typologies, while most classifications of cultures are not.

1.1 Archaeological Classification and Culture

The basic organizing concept for most prehistorians, as for most other anthropologists, is the concept of culture, but it is somewhat differently defined in the two cases. The cultural anthropologist conceives of the world as divided into a set of distinct peoples—tribes, nations, or ethnic groups—each of which has its own unique set of behavior patterns and beliefs, very often including its own language, which together constitute a culture. The prehistorian thinks of the ancient world as similarly partitioned, but the various long-vanished peoples can now be recognized only by the distinct kinds of artifact types they left behind. In place of forgotten languages and behavior patterns, every artifact type is treated as tantamount to a deliberate cultural expression—a culture trait. An archaeologically defined 'culture' is then a unique combination of artifact, house, and burial types, which are assumed, because of their cultural commonality, to be the remains left by a distinct, self-recognizing people. Those commonalities are recognized above all through processes of classification.

1.2 Kinds of Archaeological Classification

Obviously, any of the different kinds of material remains that archaeologists find can be classified, and there are in fact many different kinds of archaeological classifications and typologies. In the broadest sense, all of them fall into two categories, which may be called analytic and synthetic. Analytic classifications are classifications of one particular kind of object, in which all of the regularly recurring variants are recognized, defined, and named. The things most often classified are those that show a high degree of culturally patterned variability, including various kinds of stone tools and weapons; pottery; beads and other ornaments; house types; and grave types. Classifications of these things are usually typologies; that is, they partition the entire field of variability into a comprehensive set of mutually exclusive categories, because they are very commonly used for sorting and counting the objects found.

Artifact typologies can be made in a wide variety of ways, depending on what criteria of identity are considered important. This in turn will depend on the purpose for which the classification is made. Among the many kinds of artifact classifications it is possible to recognize purely morphological typologies, based on the overall form of objects; stylistic typologies,

which specially emphasize stylistic features; functional classifications, in which objects are classified according to their presumed use; 'emic' classifications, in which objects are classified according to criteria believed to have been important to the makers; and distributional typologies, in which objects are classified according to their distribution in time and space.

In addition to the analytic classifications of particular object types, there are also synthetic classifications, in which recurring combinations of different artifact, house, and grave types are taken together to define 'cultures.' These classifications are quite different from artifact classifications, in that they are not typologies. That is, they are not used to divide up material into discrete, mutually exclusive units. The boundaries between units are not always sharp, and the criteria of identity are not always uniform. Some 'cultures' have been identified primarily on the basis of pottery types, others by stone tool types, and still others by house types. Archaeological 'cultures' are above all historical constructs; they are the prehistorian's basic way of mapping the prehistoric world, by dividing it into units of study which can be thought of as equivalent to peoples.

Culture classifications generally have a chronological as well as a spatial dimension. That is, the classification includes cultures that existed in different areas, but also that existed in different periods of time in the same area. Very often a generalized regional culture, like Anasazi, is divided into a sequence of developmental stages, which in the case of Anasazi are designated as Pueblo I, II, III, IV, and V. Like biological classifications, then, culture classifications often have a genetic component, when later cultures are recognized as 'descended from' earlier ones.

2. *Historical Background*

Although the excavation of ancient sites, for antiquarian purposes, had its beginnings in the Renaissance, the scientific investigation of prehistory began only in the nineteenth century, above all in Scandinavia. It was Danish archaeologists who developed the first 'culture classification'—the division of all European prehistory into Stone, Bronze, and Iron Ages—and they also developed the specific artifact typologies on which the 'Three-Age System' was based. In the later nineteenth century, especially in France, prehistorians carried the same basic approach much further, dividing up the Stone Age into a whole succession of phases, or cultures, defined by distinctive tool types. In this work, artifact classification was conceived above all as an aid to dating; that is, to the placing of prehistoric remains in their proper chronological order.

In the Americas there was a general belief that prehistoric Indian remains were not more than two or three millennia old, and consequently there was not much interest in sorting out their chronology. American prehistorians were much more struck by the spatial than by the chronological variability of the indigenous cultures, as they began to recognize the very wide diversity of pottery types, tool types, and house types that had been used in different parts of the continent. Beginning in the early twentieth century, they set about defining and naming a whole panoply of localized cultures and subcultures on the basis of these variable traits. As their work intensified and their methods improved, however, they also became aware of temporal differences among the remains they studied, again based on typological features. In 1927 A. V. Kidder and his associates proposed the 'Pecos Chronology,' in which prehistoric and historic Southwestern remains were assigned to seven developmental phases, designated as Basket Maker II and III, and Pueblo I–V, each with its defining typological characteristics. Within a decade, similar chronological schemes had been devised in many other parts of North and Middle America.

Once they were drawn into the classificatory enterprise, prehistorians both in the Old and the New Worlds devoted much of their energies to the development of what have been called 'time–space grids,' which were to become the basic map of prehistory. The prehistoric world was divided into a set of cultures, and most cultures were further divided into developmental phases, strictly on typological grounds. To a very large extent, the schemes that were developed remain in use down to the present day.

By the middle of the twentieth century the 'time–space' grids were mostly in place, at least over North America and Europe, and archaeologists began making classifications for new purposes. As a result at that time of a strong influence from functionalist anthropology, it was argued that classifications should emphasize what the objects were used for, or what they meant to the makers and users, rather than simply what was useful to the archaeologist for purposes of identity and dating. The functionalist paradigm lasted for about a generation, and then was replaced by what may be called the nomothetic paradigm. It was argued, in the 1960s and 1970s, that scientific archaeology should devote itself not to historical issues of cultural development but to the testing of general, causal hypothesis about culture processes, and classifications should be developed that would aid in that process. In reality very few of them ever were; the nomothetic paradigm, as applied to classification, was much more a lofty ideal than a practical reality.

A subsequent revolution, or at least an anticipated revolution, came about with the introduction of computers. The practical problem in making typologies was always that of limiting the attributes to be considered to a finite number, without at the same time introducing the bias of human judgment. It was believed, however, that if all the possible attributes of a group of objects were fed into a computer, the

machine itself could determine, on a purely quantitative basis, which were and were not important. Thus was born the concept of 'numerical taxonomy,' in which computers would designate types, and differentiate them from other types, purely on the basis of the numbers of shared traits, regardless of what the traits were. The goal was 'automatic classification,' in which human judgment would be altogether eliminated.

After two decades of experimentation, however, this goal was found to be illusory. Unless there was some preselection of attributes to be coded, based on human judgmental decisions, the classifications produced by computers were far too cumbersome and particularized to have any practical utility. Every partitioning scheme that was tried produced hundreds of types. Moreover, the 'types' they produced could not be shown to have meaning with reference to any specific purpose. As a result, most of the typologies that are in use today are still those that were developed in the earlier part of the century, in the heyday of 'time–space partitioning.'

3. *Artifact Classifications and Types*

Basic to all artifact classifications is the concept of 'type.' Whatever kind of material is being classified—pottery or stone tools, for example—it is partitioned into a set of mutually exclusive categories that are usually called types. The type concept is actually a good deal more complex than it at first appears, and it has been the subject of various controversies that will be considered later. A type consists in the first instance of a body of objects having common features that set them apart from other objects. However, the type concept also includes our ideas about the things and what they have in common, and the words and sometimes the pictures that we use to describe them. Every type, in short, has members (actual objects), a description, a definition, and a name.

It is important to notice that these things may be modified independently of one another; we may refine our notions about what defines a particular type, based on the finding of additional material, but we may also find better ways of defining and describing the type, even if no new material is found. We may find that some characteristics are important that we had formerly ignored. Useful type concepts are always and necessarily mutable: they evolve continually as more material is found, but also as we develop new ideas about what is and is not important.

Within any typological system, the types must have two characteristics: identity and meaning. A type which cannot be recognized by any objective measure has obviously no practical utility. It would be possible, in theory, to conceive of a type including all of the pottery made at Pueblo Bonito between 1100 and 1125 CE, and such a type would have enormous interpretive utility if it could be recognized. In fact, it cannot: there is no way of differentiating the pottery made at Pueblo Bonito from that made at other nearby sites. On the other hand, it is also possible to designate types that are readily recognizable—for example, all vessels having scratch marks on one side—but that have no evident significance for any purpose.

3.1 *The Criterion of Identity*

The basic criterion of identity for artifact types have been designated as variables, and attributes. To make the distinction in the simplest terms, 'color' is a variable, while 'red' is one of the attributes of the color variable. Artifact types are never designated on the basis of *all* their visible attributes. To do so would result in a typology in which every single object was a separate type, since no two things are ever absolutely identical. Rather, certain variables are selected out as a basis for the differentiation of types, while others are ignored. To cite one example, color is usually treated as a significant variable in the case of pottery types, because it is something produced deliberately by the vessel makers, whereas it is nearly always ignored in classifications of stone tools, because it is an accidental property of the lithic material selected. Some qualities are ignored simply because they do not vary: they are common to all of the types in a typology.

Every variable has a specified set of attributes, and these also are selected in accordance with the needs of the typologist. How much distinction is made between attributes of the same variable—colors, for example—will depend partly on their identifiability, but also on how much hair-splitting is necessary for the typologist's purposes. Pottery vessels made in the prehistoric American Southwest may exhibit a very wide variety of surface colors, but typologists have generally been content to assign them to five color categories: white wares, yellow wares, buff wares, orange wares, and red wares.

3.2 *The Question of Purpose*

Above all, it is the typologist's purpose that determines which variables and which attributes are selected in making a classification. Artifact classifications can in practice be made for a very wide variety of purposes, and classifications that yield meaningful results for one purpose may not do so for another.

The various purposes that may be served by artifact classifications can for convenience be characterized as basic, ancillary, and instrumental. Basic purposes are served when we classify objects in such a way as to learn or to express something important about the objects themselves. Pottery vessels, for example, may be classified on the basis of their constituent clays and tempers, which indicate where they were made, or they may be classified on the basis of vessel shapes, which may indicate what they were used for, or they may be

classified on the basis of decorative designs, which will say something about the cultural preferences of the makers and users.

Objects, however, may also be classified for ancillary purposes: not because we want to learn or say something about the material itself, but because we want to use it as a guide to other understandings. Pottery types and certain stone tool types have long been treated as 'index fossils' or horizon markers, to identify a particular culture or a particular period in time. Their presence in a site may enable us to date that site within a century or even a generation, or to say that it was inhabited by a particular people and not by another, contemporary people. Some rather elaborate and highly particularized pottery classifications have been developed primarily as an aid to dating sites. Classifications made for this purpose will place special emphasis on whatever features show the most recognizable variability in time and space, whether or not any functional significance can be attached to them. Other kinds of ancillary classifications have been developed in order to yield information about manufacturing technologies, about resource acquisition, and for other purposes.

Some classifications are also made purely in the interest of practical convenience; for example, economy of description. Editors will usually not allow the archaeologist an unlimited number of pages in which to describe a mass of finds, such as beads or cutting tools. For economy of space they must be described in groups rather than individually. Some classifications are also made for the same reason that library books are classified: there must be a coherent way of dividing up the material into groups, for purposes of storage.

3.3 Frequency Seriation

When artifact types are used as a basis for the dating of sites, this is often done through the technique of frequency seriation. We recognize that cultures do not evolve in time through a series of instantaneous leaps, in which old artifact types are suddenly and totally replaced by new ones. Rather, there is a gradual process of transformation and replacement, in which some new types are becoming increasingly common at the same time that older ones are becoming less common. Consequently, sites may be assigned to a particular developmental phase, such as Pueblo II or Pueblo III, not on the basis of types absolutely present or absent, but on the percentages of particular types present or absent.

Obviously, frequency seriation requires quantification: the actual counting of the numbers of each artifact type present. It is for this reason that artifact typologies must be different from other kinds of classifications, such as culture classifications. A typology is a sorting and counting system, and as a result it must have a degree of rigor not necessarily found in other classifications. The complete typology must be a comprehensive set of categories (types), such that there is one and only one type for each object found. The types must all be defined on the basis of the same set of criteria, and they must be mutually exclusive.

By way of summation, it may be said that a typology is a conceptual system made by partitioning a specified field of entities into a comprehensive set of mutually exclusive categories (types), according to a uniform set of criteria dictated by the purposes of the typologist. Within any typology, each type is a category created by typologists, into which they can place discrete objects having specific identifying characteristics, to distinguish them from objects having other characteristics, in a way that is meaningful to the purposes of the typology.

4. Problems and Controversies: the 'Typological Debate'

Everyone recognizes that archaeological types are not self-labeled; it is the classifiers who give them names and definition. There has nevertheless been a very long-running debate over whether our types are 'natural' or 'artificial.' Are we merely 'finding the joints in nature,' as one proponent has it, or are we imposing our own artificial order on nature? In reality, both things are true: nearly all artifact types are partly natural and partly artificial. They are natural in that the differences between one object and another have objective reality; they were not created by us. On the other hand it is we, the typologists, who decide which distinguishing characteristics we will focus on, and which we will ignore, in making a typology. There may often be varying degrees of 'naturalness' between types in the same typology. Some types will stand out very sharply in a great many respects, while we may decide to differentiate other types only because of minor stylistic differences that are nevertheless important for dating purposes.

A related question is whether types should be created by object clustering or by attribute clustering. Should we begin our typology by dividing up a collection of objects into groups that look intuitively similar to one another, or should we first decide which variables and attributes will be important, and then decide that all unique combinations of those attributes will automatically constitute a type? Again, the practical reality lies between the two positions. Virtually all useful typologies develop dialectically through a feedback between object clustering and attribute clustering. We begin, necessarily, with a collection of objects, and make some initial observations about what seem to be the most obvious differences, on the basis of which we divide them into types. As more material accumulates, however, our ideas about what is and is not important change, and we may add some new criteria of differentiation while eliminating others.

Often we will find that we have split hairs too finely in the differentiation of some types, and not finely enough in other cases.

Types, we say, must be defined by a combination of 'internal cohesion and external isolation.' They must have features that are common to all of their members, but they must also lack features that are possessed by the members of other types. Archaeologists, however, have differed in their emphasis on one or another of these characteristics. Some have argued that types must be defined by central tendencies, without a strict definition of their boundaries; other have insisted that if typologies are to be used as sorting systems, every type must have clear boundaries. There is no one correct solution to this problem; it will depend to a considerable extent on the purpose for which the typology is to be used. The sharper the type boundaries, the more useful is the typology for sorting purposes. It must be recognized, however, that in practice there are very few artifact types that do not exhibit some fuzziness at the boundaries. The sorter will often, like a baseball umpire, have to make purely arbitrary decisions in borderline cases.

Although the theoretical literature on archaeological classification is voluminous, much of it bears little relation to what really happens in practice. There are two reasons for this disjunction. First, most of the literature refers to closed classifications, intended to classify only material already in hand. Such classifications can be as rigidly formal and immutable as the classifier wishes. In practice, however, the vast majority of artifact classifications are open systems, intended for the processing of future finds as well as for material already in hand. Such systems must necessarily by mutable: capable of continual adjustment as more material comes to hand.

Second, too many authors have ignored the fact that types must have not only identity, but also meaning relevant to some specific purpose or purposes. As we have seen above, a great many legitimate purposes may be served by archaeological classifications, and the nature of the classifications will vary accordingly.

See also: Archaeology and Philosophy of Science; Ceramics in Archaeology; Classification: Conceptions in the Social Sciences; Culture as Explanation: Cultural Concerns

Bibliography

Adams W Y, Adams E W 1991 *Archaeological Typology And Practical Reality*. Cambridge University Press, Cambridge, UK

Cormack R M 1971 A review of classification. *Journal of the Royal Statistical Society*, Series A, **134**: 321–53

Doran J E, Hodson F R 1975 *Mathematics and Computers in Archaeology*. Harvard University Press, Cambridge, MA

Dunnell R C 1971 *Systematics in Prehistory*. Free Press, New York

Dunnell R C 1986 Methodological issues in Americanist artifact classification. *Advances in Archaeological Method and Theory* **9**: 149–207

Ford J A 1954 The type concept revisited. *American Anthropologist* **56**: 42–54

Gardin J-C 1962 *Archaeological Constructs*. Cambridge University Press, Cambridge, UK

Kaplan A 1984 Philosophy of science in anthropology. *Annual Review of Anthropology* **13**: 25–39

Klejn L 1982 *Archaeological Typology*, trans. Dole P. BAR International Series **153**

Krieger A D 1944 The typological concept. *American Antiquity* **9**: 271–88

Marquardt W H 1978 Advances in archaeological seriation. *Advances in Archaeological Method and Theory* **1**: 257–314

McKern W C 1939 The Midwestern Taxonomic Method as an aid to archaeological culture study. *American Antiquity* **4**: 301–13

Rouse I 1960 The classification of artifacts in archaeology. *American Antiquity* **25**: 313–23

Rouse I 1967 Seriation in archaeology. In: Riley C L, Taylor W W (eds.) *American Historical Anthropology*. Southern Illinois University Press, Carbondale, IL pp. 153–96

Sokal R R, Sneath P H A 1963 *Principles of Numerical Taxonomy*. W. H. Freeman & Co, San Francisco, LA

Spaulding A G 1960 Statistical description and comparison in artifact assemblages. In: Heizer R F, Cook S F (eds.) *The Application of Quantitative Methods in Archaeology*. Viking Fund Publications in Anthropology **28**: 60–83

Whallon R, Brown J A (eds.) 1982 *Essays on Archaeological Typology*. Center for American Archaeology Press, Evanston, IL

W. Y. Adams

Classification: Conceptions in the Social Sciences

Classification is the assignment of objects to classes. For example, an educational researcher might want to establish a taxonomy of teaching styles that covers all possible approaches to teaching. A psychologist studying personality might be interested in whether children can be grouped into categories according to their patterns, or profiles, of personality traits. A sociologist might be interested in whether certain combinations of characteristics of urban areas (average socioeconomic status, crime rate, building types, etc.) occur much more often than other combinations. A biologist might want to study whether animals showing a particular phenotype have specific combinations, or patterns, of genetic codes. In all these cases, objects (teachers, children, urban areas, animals) are classified based on their patterns of some observable characteristics (teaching behaviors, personality traits, city characteristics, genes).

The task of classifying objects poses several problems. The objects to be classified, the properties based on which they are classified, and the way of assessing

similarities among objects have to be specified. The aim is to identify individual classes, to decide how many classes are warranted, and to establish procedures for identifying which class each object should be assigned to. Eventually, the success of the classification system must be evaluated. There are problems associated with each of these steps, and different solutions have been suggested for each problem.

In a sense, this paper recreates the process of establishing and testing a classificatory system. First, we define the topic more precisely, clarify the terminology, and provide some examples. Then, we present an outline of the first steps in a classification process and discuss one of the most debated issues: What is the concept of a class? After this, we deal briefly with procedures for assigning objects to classes and with the—frequently neglected—question of how to evaluate the resulting classification system.

1. *Definition of the Topic*

1.1 *Terminology*

Depending on the research tradition, the objects to be classified into a system are called elements, cases, units, exemplars, specimens or items. They are the sources or 'carriers' of properties, characteristics or variables. These properties may be dichotomous or polytomous, qualitative or quantitative. A property can only be useful in a classification, if it varies within the set of objects, that is, if at least two different values (categories, states, labels) on the respective property occur in the sample. When more than one property is used to characterize an object, the object can be described as a vector of values, a profile, a set of symptoms, or a pattern of features.

Sometimes, the data do not consist of objects and their properties, but of measures of relations between objects, such as their similarity, likeness, or belonging together. For example, a researcher might ask participants to rate the similarity between different politicians. The similarity ratings can then be used to classify politicians into groups. Based on this classification, the researcher can then study on which features people's perceptions of similarities between politicians are based.

The crucial assumption underlying classification is that objects are elements of a class, of a set, of a partition or—in biology—of a taxon. In other terminologies, the terms 'category' or 'cluster' are also used. Classification is the process of finding classes and of assigning entities to these classes. The end-product of this order-creating process, however, is often also referred to as 'classification.' To stress this distinction, the term 'classification system' can be used for the end-product, although in clinical psychology and biology the word 'taxonomy' is more common. Identification is the assignment of a specific case or object to (usually only) one of the classes.

1.2 *Limits of Discussion*

In studying classification in the behavioral and social sciences, we have to distinguish between two questions. (a) What theory and empirical evidence are available about how people classify objects? (b) What theories and methods are used to create classification systems? The first question will not be dealt with here (see *Concept Learning and Representation: Models*). The second question refers to the formal and empirical procedures used for defining classes and the rules that have evolved for assigning cases to classes. This second question is what concerns us here; it will be discussed from a conceptual rather than a statistical point of view. For more statistically-oriented discussions of classification methods see, for example, *Statistical Clustering*; *Mixture Models in Statistics*; *Configurational Analysis*.

One commercially 'booming' application of classification that this paper does not refer to is in biometric authentification. The aim of these methods is to use combinations of characteristics of individuals to uniquely identify each individual. Thus, conceptually, the task is to assign one of very many cases into a class, when there might be as many cases as classes. In this process, an observed pattern of features, or even an observed set of such patterns, is matched against a stored list of patterns. For example, DNA is used in this way in forensic criminology, as are records of the past *modus operandi* of individual criminals in detective work. This is, however a fairly unique example of classification, because the number of classes is intended to be equal to the number of cases, which is rarely the case in the social and behavioral sciences.

1.3 *Purposes of Classification*

The fundamental purpose of classification is to find structure. Typically, a large number of objects is reduced to a much smaller set of classes without too much loss of information about the objects. The data thus summarized allow objects to be identified, at least in part, through the class to which they belong. Specifying the boundaries describing a class has several advantages. One is that limits to generalization can be established, and another is that it becomes possible to generate predictions about how different classes are composed and how class membership relates to other variables.

2. *Some Examples of Classifications*

The most well-known examples of classifications are from the natural sciences, rather than the social sciences. A well known, still used, and expanding classification is Mendelejew's *Table of Elements*. It can be viewed as a prototype of all taxonomies in that it satisfies the following evaluative criteria: (a) Theoretical foundation: A theory determines the classes and

their order. (b) Objectivity: The elements can be observed and classified by anybody familiar with the table of elements. (c) Completeness: All elements find a unique place in the system, and the system implies a list of all possible elements. (d) Simplicity: Only a small amount of information is used to establish the system and identify an object. (e) Predictions: The values of variables not used for classification can be predicted (number of electrons and atomic weight), as well as the existence of relations and of objects hitherto unobserved. Thus, the validity of the classification system itself becomes testable.

Another successful classification system is biological taxonomy. Indeed, most attempts to formalize classification have some intellectual roots in this tradition (Sokal and Sneath 1963). The result of such classification is frequently depicted as a 'phylogenetic tree,' today often the result of comparative genomics. In biological taxonomy, however, theory is not so strong as to warrant completeness, as in the *Table of Elements* (e.g., how should one deal with archaebacteria?). Moreover, the identification of a specimen requires information from morphology and sometimes from behavioral observation. In addition, the system abounds with nested criteria. And, compared with physics, predictions of future developments or of 'missing links' in biological taxonomy are vague. However, the classes of the phylogenetic system are still useful because, at the very least, they indicate boundaries to generalization.

In the behavioral and social sciences, hundreds of classifications are published every year. Noteworthy examples are Bloom's taxonomy of educational objectives (Krathwohl et al. 1964), as well as the *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) and *ICD* (*International Classification of Diseases*) classification systems used in psychology and psychiatry. None of these systems have been formally derived, however. Instead, they were generated based on 'experience.' The resulting classes are so heterogeneous that they acknowledge many exceptions. Also, a phenomenon called 'comorbidity' shows that these classification systems are not optimal yet. It refers to the simultaneous existence of two or more disturbances in the same patient. If comorbidity is the rule rather than the exception, then the classification system loses plausibility and practicability.

3. *Preparing the Basis of a Classification*

3.1 *Selecting the Cases*

In the beginning of the process of developing a classification, two main questions arise. (a) Which elements are to be differentiated in a classification? One searches for a (complete, if possible) list of cases to be classified. This list is called the 'extension' of a classification. (b) Which properties characterize the cases? The list of these properties is called the 'intension' of the classification. The answers to these questions already determine, in part, the results of the classification. To quote Hartigan (1982, p. 2): 'Clearly, the selection of variables to measure will determine the final classification. Some informal classification is necessary before data collection: deciding what to call an object, deciding how to classify measurements from different objects as being of the same variable, deciding that a variable on different objects has the same value.' Sometimes, no well-defined population of objects is available from which to sample, and a preliminary selection has to be made intuitively. In such cases, future applications of the classification system may result in more or different classes from those originally obtained.

3.2 *Specifying the Properties*

The question of feature selection (selection of the properties on which the classification will be based) arises at two points in the process of classifying. First, as mentioned above, it arises at the very beginning of the process. The second opportunity to select properties comes when an established procedure is tested for identification. This problem is very similar to one in regression analysis: Which variables should be retained because they discriminate best, between the classes (see Pankhurst 1991). Even if computational problems do not play a role, use of too many properties can still be problematic, if measuring these variables is expensive or dangerous. In both instances of selecting properties, reliability is a very important issue. With decreasing reliability of the measurement of the properties, the identification of classes becomes more difficult.

Another important question is whether the values of the properties should be transformed before searching for classes. The results of most classification procedures will be influenced by transformations. If differences in variability between the variables are of substantive importance, no transformations that equate variability across variables should be used. The use of transformations is also called '*a priori* weighting.' '*A posteriori* weighting' refers to cases in which different variables are given different emphasis in the identification process.

Especially in routine applications, a good strategy for selecting properties to be retained in the final classification might be to find a minimal set of variables sufficient to discriminate between all the classes. Relative to the set of all variables, the minimal set may not be unique. If this is the case, one will often prefer a set with few practical problems, and replace properties and take other aspects, such as minimization of costs, into account (see Pankhurst 1991 for an algorithm).

3.3 Determining the Similarity

After the objects to be classified and their relevant properties have been selected, the similarity between objects is determined. Similarity is a key concept in classification. As was mentioned earlier, there are two basic ways to obtain similarity measures: The researcher can either collect similarity judgments from participants, or derive similarity measures from the empirical co-occurrence of properties. Methods for obtaining similarity judgments in the context of the first approach—similarity 'in the eyes of the participants'—are discussed in the context of data theory (Coombs 1964). 'Proximity' measures can also be derived from confusion or generalization data, association probabilities, substitutability ratings, sorting procedures, and so on. The second case—'similarity in the mind of researchers'—amounts to comparing the feature patterns of objects and describing the similarity between objects using *similarity coefficients*. A large, and still growing, number of these coefficients exist, and monographs on classification devote a lot of space to them. Therefore, the choice of one particular coefficient should be explicitly justified. An analysis of the metric properties of coefficients is given by Gower and Legendre (1986). To choose a coefficient, one may refer to their axiomatic foundation (Baulieau 1999).

Another important distinction in the selection of similarity coefficients refers to 'negative matching,' i.e., deciding whether to include observations stating that two objects agree that a property is absent rather than present that is, whether similarity between two objects in the absence, rather than the presence, of a property should be included in the similarity measure. Jaccard's (1908) coefficient excludes negative matchings.

Although most classification methods make use of similarity information, clustering models exist that do not refer to similarity. Another aspect that might be taken into account is the concept of similarity used: Why only use pairwise co-occurrences, and not higher order contingencies (Daws 1996)?

4. Establishing a Classification

After a measure of similarity has been selected, the next step is the actual classification of the objects based on the similarities between them. Formally (e.g., Biggs 1999), classes can be thought equivalent to (a) partitioning a set into subsets, (b) classifying a set of objects, and (c) distributing a set of objects into a set of 'boxes.' These various perspectives differ markedly in their implications for classification. For example, in most mathematical conceptualizations, an element is classified into exactly one class. Some clustering procedures, however, allow for residual elements, which are not considered clusterable.

Depending on the approach to classification that a researcher has chosen, certain considerations and precautions are necessary. For example, similarity judgment data may not fulfill some necessary assumptions: Generally, related objects are to be located in the same class. The relationship xRx when 'x is related to x' has the properties of reflexivity: xRx, symmetry: $xRy \Rightarrow yRx$, and transitivity: xRy and $yRz \Rightarrow xRz$. These are the properties of an equivalence relation. If the empirical data are similarity judgments, they do not necessarily fulfill this relation. Some properties of this relation can be tested statistically. Another important consideration applies if the objects are classified by using rules referring to features. In such cases, these rules need to be free of contradictions (for a test, see Feger 1994).

Only a few substantive theories in the behavioral and social sciences allow one to deduce the number and kind of classes needed to describe a given range of phenomena. Therefore in many cases inductive procedures have to be used to generate classes. For this, one needs a concept of what constitutes a class. Many researchers apply inductive classification methods without ever considering explicitly the class concept that their method implies. The following part of the paper gives a brief discussion of the class concepts implied in frequently used methods for finding classes. The list is not complete, and the 'cluster analysis proper' dominates all other approaches, because of the frequency of its use.

Before discussing class concepts in detail, one more general distinction needs to be made. If classes are defined by properties of objects, two levels of definition can be distinguished. A general definition specifies the relationship between the properties and the classes. Specific definitions provide detailed translations of the general definition into formal operations for assigning the objects to the classes. Obviously there can be many different specific definitions. General definitions can be ordered by the kind and amount of variability they allow among objects within the class. There are two general positions with respect to within-class variability. The 'monothetic' position (Sutcliffe 1993) assumes that a class is defined by one or a few necessary properties. The 'polythetic' counter-position (Gyllenberg and Koski 1996) assumes that some properties of a specified total set, not necessarily the same for every object, are sufficient. According to this position, a property is shared by most, but not necessarily all objects of a given class. Proponents of the monothetic camp tend to stress that some properties are more important than others, and that these properties should be used to establish the classification. The opposite position assumes equal importance of all properties. As a third type of general definition, one may add definitions referring to a 'prototype,' that is, the most typical example of a class or a hypothetical mean object. In this last case, 'closeness' or similarity decides about class membership, and the prototype may be defined with or without allowing for variation in its properties.

Given properties as the base for a classification, the actual observations often are represented as a data matrix containing, for example, the objects as the columns and their properties as the rows. The cells of the matrix contain either the values '0' or '1' to indicate the absence or presence of properties, or they contain frequencies, durations, intensities, or symbols (in the case of qualitative polytomous items) that indicate the type or degree of the respective property in the respective object. As this enumeration shows, the procedure can accommodate data of all scale types. The goal now is to find a 'feature by classes' matrix, called—corresponding to its purpose—the reference or identification matrix, or simply 'a classification.'

4.1 Concepts of Classes

Cluster analysis proper. When authors (e.g., Everitt 1993) illustrate the concept of a cluster, they often use two-dimensional graphs to show the clusters as clouds of points (representing the objects). The clouds can have various forms; generally there are 'gaps' between the clusters that contain no data points, so that the clusters are isolated from one another. While such explanations of the cluster concept seem intriguing as they invoke classical 'gestalt' concepts, it is important to remember that the properties of (good) figures are defined by several 'laws,' not just one or two axioms or rules, as in cluster concepts.

Helpful as visualizations are, the more general definition of a cluster does not refer to any particular conception of space, be it dimensional, metric or Euclidean. Set theory defines a cluster as the maximal subset of elements for which proximities within this subset are larger than between any elements of the subset and elements not contained within it. As was discussed above, proximities are information about the extent to which objects 'belong together,' and could be expressed in many different ways, for example, as similarities, distances, ranks, or binary information about set membership. More than one subset may exist; subsets may be disjointed or overlapping; and they may or may not be hierarchically ordered. Given this very broad conceptualization, social scientists have access to a large number of clustering procedures. The large number of options reveals that no 'one and only' definition of a cluster can be found. Presumably, the availability of so many approaches is one reason for the paucity of comparative studies on methods of clustering.

Clustering procedures can be classified as 'leading to a structure that is either hierarchical or non-hierarchical.' The most frequently applied classification procedures are hierarchical, disjointed, and provide exactly one class for each object. The best known hierarchical procedures are agglomerative, that is, in a series of partitions they successively and with increasing dissimilarity, fuse the objects into classes. Each step provides a set of classes, from which the researcher has to make his choice. Once a fusion is made, it is irrevocable, so the early fusions should be very reliable. *Additive clustering* (Shepard and Arabie 1979) is a hierarchical partitioning allowing membership of objects in any number of classes. Here, the classes might be interpreted as properties (Lee 1999).

The cluster concept treated thus far is based on similarity as formally represented either in a space or by set theory. A close relative is prototype theory, popular in cognitive research. A prototype can be defined as a vector of values of selected properties; usually a list of cases as exemplars of this prototype is also available. One fundamental assumption of the prototype-oriented approach can be formulated as follows: If there is high similarity among a set of patterns, these patterns are also similar to an—observed or inferred—prototypical pattern. An inferred pattern could, for example, be the vector of mean values. This pattern has high or maximal similarity to every other pattern. The idea of inferring the prototypical pattern from the data forms a bridge to the similarity-based conception. But the researcher has to be more active in abstracting and defining a specific instance as the prototype.

Contingencies of higher order than similarities between the properties are exploited in some other generalizations of the concept of similarity-based clustering, such as Configural Frequency Analysis (Krauth and Lienert 1973) and Pattern-Analytic Clustering (McQuitty 1987). For example, Configural Frequency Analysis identifies combinations of properties that occur more often than expected from some specified base model.

A recent trend, increasing in strength, is to use mixture models for clustering. The original purpose of these methods was to base classification on a model that allows for inference-statistical treatment. But they have since found wider purposes. The basic idea of mixture models can be illustrated using the following example: Assume that a sample of measurements of body height is drawn from a human population. While it is known that all the cases are male or female, gender is not recorded for individual respondents. It is, however, possible, based on the distribution of heights in the total sample, to estimate the coefficients of the separate height distributions for men and women. This is done by interpreting each measurement as a sum of weighted height measurements for women and for men. These weights are the probabilities for each measurement to be from a man and from a woman. 'Thus the density function of height has been expressed as a superposition of two conditional density functions; it is known as a finite mixture density.' (Everitt 1993, p. 110).

Mixture models are based on a 'space' concept rather than a 'similarity' concept; clusters are regions of relative point densities in this space. The assumptions for mixture models are comparable with those of

the general linear model: cardinal scale level and multivariate normal (or similar) distributions of the data. A comparatively common mixture model for categorical data is latent class analysis (De Soete 1993).

To conclude this classification of class concepts, one further concept needs to be mentioned. This conception, models for block structure, is close to the raw data matrix and the Aristotelian tradition. A block is a maximal rectangular submatrix combining some objects and some properties with the same (or similar) values in the cells of the data matrix. The scale level of the values is not fixed; and the similarity concept is not invoked in the analytical procedure. In a block, the set of partially similar objects corresponds to the extension of a concept or class. The set of partially similar properties corresponds to the intension. The symmetry in the definitions of intension and extension is fully exploited and preserved (see Feger and De Boeck 1993).

4.2 *Evaluation of a Clustering Result*

Although model evaluation is only a part of the overall evaluation of a classification (see Sect. 5), it is an important one. As Dunn and Everitt (1982, p. 94) state: 'Since clustering techniques will generate a set of clusters even when applied to random, unclustered data, the question of validating and evaluating becomes of great importance.' Jain and Dubes (1988) classify the criteria of validation as follows:

> *External criteria* measure performance by matching a clustering structure to a priori information.... *Internal criteria* assess the fit between the structure and the data, using only the data themselves.... *Relative criteria* decide which of two structures is better in some sense, such as being more stable or appropriate for the data.

Considerable progress has been made in internal statistical cluster evaluation. Statistical procedures exist for testing the existence of 'natural' clusters, for testing the adequacy of computed classifications, and for the determination of a suitable number of clusters (see, e.g., Bock 1996). A very plausible way to evaluate any solution, independent of the clustering approach used, is to reproduce, or 'derive,' from the solution all information that the solution gives about raw data that would fit with the solution, and then to compare this information with the actual raw data.

4.3 *Procedures to Assign Cases to Classes*

Procedures to assign single cases to classes are needed for two purposes. One purpose is to assign newly observed cases to the classes of an already existing classification. The other purpose is to evaluate a classification by taking 'old' cases from the original sample on which the classification was based, and checking which class they would be assigned to. In both cases, the question is: Into which class should the case be placed? In practice, experts (e.g., physicians) are often consulted for the answer this question. In other cases, numerical procedures ('automatic classification') are used. Here, the properties may be used, either sequentially, as in a diagnostic key, or simultaneously by some type of matching between the case and the existing classes (see Dunn and Everitt 1982, especially on diagnostic keys). Quite often, as identification with certainty is impossible 'either because too many characters are variable within taxa or because all assessments of character states are subject to error, probabilistic identification methods are often used' (Dunn and Everitt 1982, p. 112). Of the probabilistic procedures, the Bayes approach (see *Decision Theory: Bayesian*) and discriminatory analysis (see *Multivariate Analysis: Classification and Discrimination*) are especially well known.

Other placement rules can be used if they are transparent, unambiguous, and do not lead to contradictions. For example, the principle of 'nearest neighbor' computes the distances of a new pattern to all existing classes. It assigns the new case to the class to which the distance is shortest (for details and other rules, see Looney 1997). Rules may also include options such as rejecting a case as 'not classifiable' or postponing a decision until more information is available. Most rules currently applied are compensatory, but rules could also be disjunctive or conjunctive, requiring at least one value to reach a high amount, or all values to surpass a given minimum (Coombs 1964).

Different rules lead to different results, especially if the classes vary in their *a priori* probability, if the distributions and covariances of the variables are very different, and if the number of observations is small. The single most important criterion for evaluating an assignment procedure is the number of correct classifications. But this 'apparent error rate' is optimistically biased, because it does not take into account the probability of correct assignments by chance. If base rates of class membership are known, the predictions have to perform better than the base rate (Pires and Branco 1997).

5. *Evaluating a Classification*

While every step of the classification process can be evaluated (Milligan 1996), two stages have received special attention: the evaluation of class definitions and of the identification procedure. Both have been mentioned previously. There also exist procedures to evaluate the overall performance of a classification system. The main method is 'cross validation,' using a new sample of data comparable to the old one, or splitting the original sample randomly into two halves and using one half to evaluate the classification obtained in the other half.

Another way of evaluating the results of a classification process is by comparing the results of using different classification procedures. Usually the researcher has several choices in the classification process, and is not forced by theory to select just one option. Examples include multiple options about the selection of cases and variables, of similarity coefficients, of clustering models and of identification rules. With computers, it is easy to try several combinations of such choices. Confidence that the classification captures substantial information in the data grows with the amount of agreement in the results from different combinations of choice options (for the evaluation and comparison of solutions, see Everitt 1993). To aid the interpretation of the resulting class structure, Milligan (1996, p. 346) suggests deliberately adding 'ideal types,' that is, characteristic patterns constructed by the researcher, to the data, and to assess what clusters these patterns are assigned to.

Can a classification be wrong? In most cases, a classification is just a systematic description, and as such, may, or may not, be useful. New observations may require changes in the classification. But if there exists a theory about the definitions of the classes, and the theory is strong enough to allow for specific predictions, then these predictions can be falsified and/or lead to revisions of the classification system.

6. Conclusions

More in the past than in the present, opinions fundamentally critical of the possibility of classification in the social sciences have been expressed. For example, Galt and Smith (1976, p. 58) stated: 'Because they usually lack measurable dimensions, social entities are difficult to classify, and any given system of classification will inevitably be arbitrary.' Indeed, numerical classification definitely requires measurement, or more generally, the interpretation of observations as variables. Indeed, every variable is itself a classification, defined as a set of disjointed, exclusive and together sufficient classes, and the 'categories' of variables are referred to as their values. Without variables in this formal sense, one might consider a heuristic equivalent in abstraction from, and ordering of, observations: the 'ideal type,' as introduced by Max Weber.

Variables used to establish a classification may be discrete or continuous. For a classification to be justified, the frequency or density distributions of the properties should not be equal distributions, but show one or several peaks. Then, when the joint distributions of more than one variable are considered, some combinations of values may be more frequent than other combinations, perhaps even more frequent than would be expected based on the marginal distributions. This is one of the fundamental phenomena enabling the formal definition of classes.

While classification is a 'process,' the temporary result is a 'structure.' Dynamic aspects, such as the development of a property and other trends and changes, might be included in the definitions of variables. This does not, however, make a formal classification a process model. In this sense, classification is static. It temporarily fixes the—sometimes turbulent—streams of information. Changing a classification means the (re-)interpretation of some substantive area.

See also: Statistical Clustering; Multivariate Analysis: Classification and Discrimination; Mixture Models, in Statistics; Measurement Theory: Conjoint; Person-centered Research; Configurational Analysis

Bibliography

Arabie P, Hubert L-J, De Soete G (eds.) 1996 *Clustering and Classification*. World Scientific, Singapore

Baulieau F B 1999 Two variant axiom systems for presence/absence based dissimilarity coefficients. *Journal of Classification* **14**: 159–70

Biggs N L 1999 *Discrete Mathematics*. rev. edn. Clarendon Press, Oxford, UK

Bock H-H 1996 Probability models and hypothesis testing in partitioning cluster analysis. In: Arabie P, Hubert L J, De Soete G (eds.) *Clustering and Classification*. World Scientific Singapore , pp. 377–453

Coombs C H 1964 *A Theory of Data*. Wiley, New York

Daws J T 1996 The analysis of free-sorting data: Beyond pairwise cooccurrences. *Journal of Classification* **13**: 57–80

De Soete G 1993 Using latent class analysis in categorization research. In: van Mechelen I, Hampton J, Michalski R S, Theuns P (eds.) *Categories and Concepts*. Academic Press, London, pp. 309–30

Dunn G, Everitt B S 1982 *An Introduction to Mathematical Taxonomy*. Cambridge University Press, Cambridge, UK

Everitt B S 1993 *Cluster Analysis*. 3rd edn. Edward Arnold, London

Feger H 1994 *Structure Analysis of Co-occurrence Data*. Shaker, Aachen, Germany

Feger H, De Boeck P 1993 Categories and concepts: Introduction to data analysis. In: van Mechelen I, Hampton J, Michalski R S, Theuns P (eds.) *Categories and Concepts*. Academic Press, London, pp. 203–23

Galt A H, Smith L J 1976 *Models and the Study of Social Change*. Wiley, New York

Gower J C, Legendre P 1986 Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* **3**: 5–48

Gyllenberg M, Koski T 1996 Numerical taxonomy and the principle of maximum entropy. *Journal of Classification* **13**: 213–29

Hartigan J A 1982 Classification. In: Kotz S, Johnson N L, Read C B (eds.) *Encyclopedia of Statistical Sciences*. Wiley, New York, Vol. 2, pp. 1–10

Jaccard P 1908 Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise de Science Naturelle* **44**: 223–70

Jain A K, Dubes R C 1988 *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ

Krathwohl D R, Bloom B S, Masia B B 1964 *Taxonomy of Educational Objectives*. Longman, London

Krauth J, Lienert G A 1973 *KFA—Die Konfigurationsfrequenzanalyse*. Alber, Freiburg, Germany
Lee M D 1999 An extraction and regularization approach to additive clustering. *Journal of Classification* **16**: 255–81
Looney C G 1997 *Pattern Recognition Using Neural Networks*. Oxford University Press, New York
McQuitty L L 1987 *Pattern-Analytic Clustering: Theory, Method, Research and Configural Findings*. University Press of America, New York
Milligan G W 1996 Clustering validation: Results and implication for applied analyses. In: Arabie P, Hubert L J, De Soete G (eds.) *Clustering and Classification*. World Scientific, Singapore, pp. 341–75
Pankhurst R J 1991 *Practical Taxonomic Computing*. Cambridge University Press, Cambridge, UK
Pires A M, Branco J A 1997 Comparison of multinomial classification rules. *Journal of Classification* **14**: 137–45
Shepard R N, Arabie P 1979 Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review* **86**: 87–123
Sokal R R, Sneath P H A 1963 *Principles of Numerical Taxonomy*. Freeman, San Francisco
Sutcliffe J P 1993 Concept, class, and category in the tradition of Aristotle. In: van Mechelen I, Hampton J, Michalski R S, Theuns P (eds.) *Categories and Concepts*. Academic Press, London, pp. 35–65

H. Feger

Classifiers, Linguistics of

Classifiers are overt morphemes that constitute morphosyntactic systems which are semantically motivated and subject to discourse-pragmatic conditions of use. Classifier systems are not found in Indo–European languages. They are in essence secondary linguistic systems characterized, on the one hand, by their clear lexical origin and persistent semantic motivation and, on the other, by their functioning as morphosyntactic systems. The better-known systems are the numeral classifier systems of Asian or Amerindian languages, illustrated in Table 1.

Classifier studies became of interest to general linguists in the 1970s, following proposals to capture the universal semantic properties of classifier systems. Adams and Conklin (1973), Denny (1976), Allan (1977) have provided the framework for many of the subsequent descriptions and discussions, and can be considered as classics of the field. Adams and Conklin were the first to wade through much comparative data and claim the existence of some universal semantic properties, based primarily on data from Asian numeral classifier systems. They established the primacy of three basic shapes, which are semantically combinations of one of the major dimensional outlines of objects (1D, 2D, 3D) with a secondary characteristic of consistency and/or size. This combination is directly inherited from the most common lexical sources of a basic set of classifiers, which are the primary elements of the physical world being handled for the survival of human communities (Table 2).

Table 1
Examples of numeral classifier systems

Japanese	
enpitsu ni-hon	hon ni-satsu
pencil 2-CL(1D)	book 2-CL(bound volume)
'two pencils'	'two books'
Tzotzil	
j-p'ej alaxa	j-ch'ix kantela
1-CL(3D) orange	1-CL(1D) candle
'one orange'	'one candle'

Table 2
Lexical sources of basic classifiers

Classifiers	Lexical origin
1D: long-rigid	tree/trunk
2D: flat-flexible	leaf
3D: round	fruit

Denny (1976) is the work of a psychologist handling data secondhand. He offers the appealing proposal that the semantic traits of classifiers may be organized into three kinds, those of 'social, physical and functional interaction,' assuming that what classifiers are 'good for' is to signal how humans interact with the world. Under social interaction he places interaction with animate entities of our world, principally fellow human beings, classified by sex, social rank, or other categorization schema, as well as other entities such as divinities and other powers specific to a culture. In the physical interaction realm, objects of the world are classified along certain parameters linked to their nature as manipulable and manipulated objects, principally the parameter of shape. Finally, in the functional interaction realm, entities of the world are classified by the use to which they are put, such as items of clothing, hunting or fishing, transportation, for instance.

Allan (1977) is a first typological of study of so-called classifiers, based on a broad data base of fifty classifier languages. Although the reliability of the data is variable and different types of nominal classification systems are lumped together under the label of 'classifiers,' there is still remarkable overlap between his seven 'categories of classification' (material, shape, consistency, size, location, arrangement, and quanta) and Denny's three. Two of Allan's original statements are of particular interest for later discussions on the nature and purpose of classifier categorization; they are the fact of a total absence of color classifiers, and the constraint that the characteristics denoted by the categories of classification be perceivable by more than one of the senses alone, such as sight and touch,

Table 3
Specific, unique, and repeaters in Jakaltek-Popti'

Type of classifier	CL	Class members
specific:	no'	ALL animals, except dog, AND all products of animals (*no' hos* 'egg,' *no' lech* 'milk,' leather shoes, wool blankets etc …)
unique:	metx'	ONLY *metx' tx'i* 'dog'
repeater + specific:	tx'otx'	*tx'otx' tx'otx'* 'dirt, ground' AND all objects made of clay (*tx'otx' xih* 'pottery jug' etc …)
repeater + unique	atz'am	ONLY *atz'am atz'am* 'salt'

where sight means primarily perception of shape. It is worth noting here that to arrive at the kind of statements of universals found in the above mentioned publications meant wading through vast amounts of data from fieldwork often containing large sets of classifiers, and interpreting their semantics through approximate translations, in order to identify those universal characteristics.

Besides varying as to the semantics of the individual classifying elements, systems of classifiers vary greatly as to the number and the specificity of the classes around which the systems seem to be organized. The classes headed by classifiers can vary from very simple to very complex; they can be small or large, homogeneous or extremely heterogeneous. Homogeneous classes are those with transparent semantic motivation, while heterogeneous classes are usually considered to be composed of a core set of prototype elements to which others have been added through various means of extension. Therefore, within the literature on classifier systems one finds different labels to indicate the nature of the classes themselves. One talks, for instance, of specific, general, unique and repeater classifiers. *Specific classifiers* are the most common type. The classes they head are built around prototypical exemplars, with incorporation of other elements by any number of types of extensions of the class. One of the most notorious examples in the literature of a specific classifier heading a very heterogeneous class is the case of the Japanese numeral classifier *hon*, used prototypically for long, thin objects.

General classifiers, as their label indicates, are largely desemanticized and head large heterogeneous classes with no distinct semantic motivation. Large Asian numeral classifier systems are known to have general classifiers. At the opposite end, *unique classifiers* head classes of just one element. One finds in the literature examples of unique classifiers for certain animals, for instance, such as the elephant or the tiger, or even the dog, generally interpreted *a posteriori* as highlighting some cultural item of particular significance. Finally, the term *repeater* refers to classifiers that are homophonous with a noun while being either unique or specific (Table 3). The existence of repeaters is what makes for the openendedness of some classifier systems.

The semantic studies of the 1970s had a tendency to overlook the existence of different types of systems, often lumping together classifier systems with other systems. Towards the end of the twentieth century, attention has been given to the fact that classifier systems are one type of nominal classification system among several others, as argued in Grinevald (2000), in contrast to the position taken by Aikhenvald (1999). The non-lumping position argues that classifier systems are intermediate systems in a continuum of nominal classification systems that range from lexical to morphosyntactic systems. At the lexical end they may be distinguished from two types of systems, the measure terms and class terms systems, with which they are often either confused or consciously lumped. At the grammatical end, they are widely considered as distinct from the gender systems and the noun class systems, which are both essentially grammaticalized concordial systems.

All languages have lexical sets of measure terms, the expression 'measure terms' being used here as a cover term for what are strictly speaking measures and for types of arrangements. Examples of English measure terms include actual measure terms such as a glass of water, a pound of sugar, a slice of bread, a sheet of paper, and arrangements such as a pile of books, a group of children, a line of cars. Class terms are sets of lexical items used in lexicogenesis; they participate in compounding processes of word formation that are functionally equivalent to derivational processes. The English class terms '-berry' (as in strawberry, blueberry, boysenberry, goodberry, loganberry), '-tree' (as in apple tree, banana tree, cherry tree), or even '-man' (as in mailman, policeman, garbage man) are the functional equivalent of French derivational suffixes such as '-*ier*' (as in *pommier* 'apple tree,' *bananier* 'banana tree,' *cerisier* 'cherry tree') or '-*ier*/-*eur*' (as in *facteur* 'mailman,' *policier* 'policeman,' *éboueur* 'garbage man').

Table 4
Class markers in Tswana (Bantu)

a.	**le**-kau	**le**	**le**-leele	**le**	**le**-ntsho	**le**	**le**	opelang	**le**-le
	5-boy	5	5-tall	5	5-black	5	5	sing	5-DEM
	'this tall black boy who is singing'								
b.	**Le**-kau	**le** lapile;	Ke	**le**	thusitse				
	5-boy	5-is tired	I	5	have helped				
	'the boy is tired, I have helped him'								

At the grammatical end of the continuum of nominal classification systems are the gender systems found in Indo–European languages and the noun class systems of Bantu languages. Note in the examples inTable 4 the ubiquitous presence of noun class markers (here of class 5) on nouns, adjectives, and demonstratives, and in the verb as pronominal clitics. The different degrees of grammaticalization of gender/noun class systems and classifier systems are generally assessed according to the set of criteria listed in Table 5.

Beyond distinguishing classifier systems from other systems of nominal classification it is important to acknowledge the existence of several subsystems of classifiers, which are usually identified and labeled primarily by their morphosyntactic locus. The best known and documented types are the ones found as elements of the noun phrase itself, such as the numeral classifiers (numeral + CL) used in quantifying expressions; the noun classifiers (CL noun) so-called for appearing with a bare noun, not linked to the expression of quantification or possession; and the genitival or possessive classifiers (poss + CL) that are part of possessive constructions. The verb forms are the locus of two possible systems of classification: the verbal classifiers (verb-CL) which belong to the systems of nominal classification and the lesser known verb classifiers, which actually classify types of verbs rather than nominal arguments.

Noun classifiers; Jakaltek-Popti' (Craig 1986, p. 264)
xil naj xuwan no7 lab'a
saw CL John CL snake
'(man) John saw the (animal) snake'

Numeral classifiers; Ponapean (Rehg 1981, p. 130)
pwihk riemen 'two pigs'
pig 2 + CL: animate
tuhke rioapwoat 'two trees'
tree 2 + CL: long

Genitive classifiers; Ponapean (Rehg 1981, p. 184)
kene-i mwenge 'my(edible) food'
CL-GEN.1 food
were-i pwoht 'my(transport) boat'
CL-GEN.1 boat

Verbal Classifiers; Cayuga (Mithun 1986, pp. 386–8)
a. ohon'atatke: ak-hon'at-a:k
it-potato-rotten past.I-CL-eat 'I (potato)ate a rotten potato'
b. so:wa:s akh-nahskw-ae'
dog I-CL:domestic.animal-have
'I have a (pet) dog'
c. skitu ake'-treh-tae'
skidoo I-CL: vehicle-have
'I have a car'
Gunwinggu (Oates 1964 in Mithun 1986, p. 389)
d. gugu ga-bo:-mangan
water it-CL: liquid-fall
'water is falling'

The major argument to prove the existence of different types of classifiers is the co-occurrence of several independent systems in the same language. These are systems with different inventories of classifier morphemes, different semantics and different morphosyntactic loci, such as the coexisting numeral and possessive classifier systems of Micronesian languages like Ponapean, for instance. Although much remains to be done in terms of the study of the semantics of classifiers, preliminary exploration of a correlation between the major morphosyntactic types of classifiers known and their semantic profiles point to the following alignment (Grinevald 2000).

Shape seems to be the dominant semantic parameter in numeral classifier systems, while function is the major semantic parameter of genitival classifier systems, and material the major one of noun classifier systems, in the following pattern:

(a) numeral classifiers = physical categories:
two-ROUND oranges;
three-LONG RIGID pencils;
four-FLAT FLEXIBLE blankets
(b) genitive classifiers = functional categories
my-EDIBLE food;
his-DRINKABLE potion;
their-TRANSPORT canoe
(c) noun classifiers = material/essence categories
an ANIMAL deer;
the ROCK cave;
MAN musician

The claim that there exist different types of classifiers raises two questions about the function of classifiers: one is the unavoidable one about the common function of classifiers in general in the languages that avail themselves of such systems. The other arises from the identification of different types of classifier systems and concerns the distinct functions that those different

Table 5
Criteria for distinguishing noun classes and classifiers

	Noun classes	Classifiers
a.	classify all nouns	don't classify all nouns
b.	in a small number of classes	in large(r) number
c.	closed system	open system
d.	fused with other grammatical categories (number, case …)	not fused
e.	can be marked on N	not marked on N itself
f.	in concord/agreement pattern	not part of concord systems
g.	N assigned to one class	can be to assigned to several classes
h.	no speaker variation	possible speaker variation
i.	no register variation	possible formal vs informal use

Source: Dixon 1982

types of classifier fulfill, in view of their different morphosyntactic loci and semantic profiles. It has to be noted that when the issue of the function of classifier systems has been addressed in the literature, it has generally been, admittedly or not, from the perspective of numeral classifiers only, without regard for the variety of classifier types. In this context, (numeral) classifiers have been seen as markers of individuation or unitizing that operate in languages in which the semantics of nouns is taken to be essentially equivalent to that of mass or concept nouns of Indo–European languages.

One proposal dealing with the distinct functions of the various types of classifier systems has been an analysis of noun phrases as layered structures parallel to verbal layered structures in which different types of operators are found. In this framework numeral classifiers were considered to be quantification operators with semantics that appealed to the handling of items to be counted, hence primarily shape and physical characteristics. It was further argued that possessive classifiers were localizing operators, and their semantics linked to the function of the items appropriated, while the noun classifiers were quality operators and as such appealed to the material or essence of the items that appeared as arguments of discourse (Grinevald 2000). Much remains to be done to document the variety of classifier systems well enough to be able to address this issue comprehensively.

One of the major challenges of classifier studies is that the essentially intermediate nature of classifier systems, as secondary linguistic systems halfway between lexicon and grammar, means great variability of the systems. Acknowledging the need to take into account the inherent dynamics of classifier systems makes the descriptive task both more onerous and more productive if comparative and typological work is to proceed properly. The need is felt to include a number of dynamic variables to handle the description of specific classifier systems, by attending to their place in the overall grammar of the language within which they develop and are used. These variables include:

(a) their degree of grammaticalization: within each subtype of classifier system, one can identify systems at different stages of grammaticalization. For instance, incipient systems of noun classifiers can be found on the Australian continent next to well-established ones, meanwhile the numeral classifiers of the Chibchan languages of Central America are much more grammaticalized than those of Asian languages.

(b) the age of the system: some systems are very old, e.g., the Chinese system of numeral classifiers, while others can be argued to be only several centuries old, like the q'anjob'alan-Mayan noun classifiers.

(c) the productivity of a classifier system must be considered too, independently of its age. For instance the Thai numeral classifier system, which is very old, is also very productive: it is open and adapting to the language of modern life, while the noun classifier system of Jakaltek-Popti', which is not very old, seems frozen and unable to cope with the classification of modern imports and products.

(d) the particular classifier system needs to be assessed in the context of the common phenomenon of areal spread of such systems. The spread can operate either through the actual borrowing of a system, morphology included, as was the case with the expansion of the original Chinese numeral classifier system into its surrounding regions, or through the borrowing of the idea and motivation for the development of such systems, as seems to have taken place between the q'anjob'alan languages of Guatemala and their neighboring Mamean languages.

While early propositions of matching classifier systems with morphological types of languages were not enlightening, there is indeed a tendency for different types of nominal classification systems in general, and of different types of classifiers in particular, to distribute themselves in clusters around the world. For instance, gender systems are a widespread phenomenon in Indo–European languages, while noun class systems were originally mostly known from

Bantu languages. In addition, they have been described for languages of Australia and Papua New Guinea, and are perhaps more widespread than previously recognized in Amazonia.

Turning our attention to classifier subtypes, numeral classifiers are best known for their presence in South East Asian languages, but have also been identified in America, in particular Mesoamerica. Noun classifiers appear to be a rare type, mostly identified in Mesoamerica and Australia, while possessive classifiers are the hallmark of Micronesian languages, although they are also found in various parts of America. As for verbal classifiers, they have been documented for North American languages and for signed languages, though it is sometimes difficult to establish how segmentable and identifiable the actual classifier morphemes are in verbal predicates.

Classifier studies raise difficult methodological issues. There is the first-degree challenge of the fieldwork to be done to produce descriptions of these systems. Fieldwork faces the problem of the ethnocentrism of the semantic analysis. This will continue as long as much of this work is done through translation, and is still largely carried out by linguists who are native speakers of languages where the phenomenon does not exist. (Work done on South East Asian languages is to some extent an exception to this problem.) For instance, it is often difficult to say whether the semantics of a classifier is one of the strictly physical characteristic of shape or one of function, since certain shapes naturally lend themselves to certain functions. Are objects hollowed or made to be concave to be considered for their shape, as round hollow objects, or for their function, as recipients and containers? The basic issue is, of course, whether the right questions are being asked of these systems in the first place. There is the further challenge of the fundamental lexico-grammatical nature of such systems, with their common open-endedness and subtle discourse functioning that require extensive studies all too rare in the descriptive tradition. And there is the second-degree challenge facing linguists working on secondhand data of uncertain reliability and common incompleteness, particularly in terms of the pervasive dynamics and complex internal typology of such systems, of the kind introduced in this article.

A number of publications are shaping the field of classifier studies in the context of the wider discipline of nominal classification. Craig (1986) meant to begin to confront various approaches to the study of nominal classification systems and included articles on the acquisition, historical development, discourse function, and semantics of classifier systems. Senft (2000) is a more recent collection (based on a 1993 working conference) where the issues of typology, grammaticalization, and function of classifiers are further elaborated. Aikhenvald (1999) is a substantial monograph which attests to the importance of Amazonian data, for the richness of its classification systems and the challenges they pose to the proposed typology, and which underlines in general the extreme complexity of the inter-relation of systems in many languages. Sands (1995) is a useful survey of nominal classification systems in Australia which reveals two interesting phenomena: one is the parallel development, in different languages, of concordial noun class systems and noun classifier systems out of the same lexical material of generic nouns; and the other is the documentation of the various stages of the evolution of noun classifier systems from the discourse sensitive use of generic nouns and through the increasingly frequent collocation of generics and nouns in classifier constructions. Bisang (1996, 1999) provides overviews of the grammaticalization dynamics through which the classifier systems of East and South East Asian languages arose, within the areal typological frame called for by the language contact situation of the region.

In terms of an agenda for the development of classifier studies in the twenty-first century, the work is proceeding on two fronts. On the linguistic fieldwork front, there is still an enormous need for more comprehensive descriptions. And, given the fact that many systems worth investigating are to be found in the languages of Amazonia, Australia, and Papua New Guinea, the challenging nature of this fieldwork has to be kept in mind. Work on nominal classification processes in signed languages is also under way; a collective volume on that topic in a cross-linguistic perspective is scheduled to appear following a working conference in 2000 (Emmorey in press).

A better understanding of the general phenomenon of classifiers should also emerge from the confrontation of what is now known of nominal classification systems with the much less known phenomenon of verb classification. This phenomenon has been described for some Australian languages, but has not come fully to the attention of linguists interested in nominal classification. However, the identification of a similar phenomenon in South American languages such as the Barbacoan languages of Ecuador and its ongoing description should facilitate further exploration of the parallels between the organization of nominal and verbal linguistic expressions and the similarity in the function of their operators, in particular that of their respective classifier systems.

On the more theoretical front, various debates are open and in need of further consideration. They include taking a position on the *a priori* lumping or not lumping together of most nominal classification systems. Lumping means subsuming, under the label of classifiers, cases of class terms, measure terms, noun classes, as well as classifiers. This position is defensible in terms of how the new data collected in the field, particularly data on languages until recently never described, appear startling if not overwhelming because of overlapping layers of elements of supposedly various types of systems. Alternatively one can opt to

tease apart different types of classification systems using as a reference clear cases of classifier systems with the characteristics given above in Table 5.

To handle the cases of data overlaps one can proceed with a number of tools from a functional-typological approach to the study of language. One is the notion of prototype, with its accompanying concept of fuzzy boundaries; this allows for some systems to be intermediate between two systems, such as noun class and classifier. Another is the notion of the grammaticalization dynamics, inter- and intra-types of nominal classification systems, allowing in particular for variations along this parameter within the same type, which can change an analysis of multiple classifier systems to one of incipient noun class system for instance. A third is the notion that several systems may indeed co-exist, the original system and another one evolved in part from it, resulting in homophonous morphemes belonging to various systems, such as class terms and classifiers.

Another major issue due for more debate in the twenty-first century is the issue about how much the linguistic phenomenon of classifiers is linked to categorization of referents in the world linguistic classification of nouns. This has been best articulated by Lucy (1992), while Foley (1997, Chap. 12) provides a good summary of recent experimental work on the cognitive impact of classifiers in categorization. It is clear that before constructing experimental studies that use classifiers in the search for the nature of the links that hold between language and cognition, an assessment of the degree of grammaticalization of the systems is needed. Ongoing discussions of so-called classifiers in signed languages underline also their fundamental discourse functions of referent identification and referent tracking, and the likelihood that the term 'classifiers,' now well established in the literature, both in its narrower and wider scope, may very well be a misnomer which delays a better grasp of their role in language.

See also: First Language Acquisition: Cross-linguistic; Foreign Language Teaching and Learning; Language Acquisition; Language Development, Neural Basis of

Bibliography

Adams K L, Conklin N F 1973 Towards a theory of natural classification. *Papers from the Regional Meeting of the Chicago Linguistic Society* **9**: 1–10

Aikhenvald A 1999 *Classifiers. A Typology of Noun Categorization Devices*. Clarendon Press, Oxford, UK

Allan K 1977 Classifiers. *Language* **53**: 285–310

Berlin B 1965 *Tzeltal Numeral Classifiers*. Mouton, The Hague

Bisang W 1996 Areal typology and grammaticalization: Processes of grammaticalization based on nouns and verbs in East and Mainland South East Asian languages. *Studies in Language* **20–3**: 519–97

Bisang W 1999 Classifiers in East and Southeast Asian languages: Counting and beyond. In: Gvozdanovic J (ed.) *Numeral Types and Changes Worldwide*. Mouton de Gruyter, Berlin

Craig C G (ed.) 1986 *Noun Classes and Categorization*. John Benjamins, Amsterdam

de León L 1988 Noun and numeral classifiers in Mixtec and Tzotzil: A referential view. Ph.D thesis, University of Sussex, UK

Denny J P 1976 What are noun classifiers good for? *Papers from the Regional Meeting of the Chicago Linguistic Society* **12**: 122–32

Dixon R M W 1982 Noun classifiers and noun classes. In: Dixon R M W (ed.) *Where Have all the Adjectives Gone? And Other Essays on Semantics and Syntax*. Mouton, The Hague, pp. 211–33

Dixon R M W 1986 Noun classes and noun classification in typological perspective. In: Craig C G (ed.) *Noun Classification and Categorization*. John Benjamins, Amsterdam, pp. 105–12

Emmorey K (ed.) in press *Perspectives on Classifier Constraints in Sign Languages*. Erlbaum, Mahwah, NJ

Foley W 1997 *Anthropological Linguistics: An Introduction*. Blackwell, Oxford, UK

Greenberg J H 1978 How does a language acquire gender markers? In: Greenberg J H (ed.) *Universals of Human Language, Vol. III*. Stanford University Press, Stanford, CA, pp. 47–82

Grinevald C 2000 A morphosyntactic typology of classifiers. In: Senft (ed.) *Nominal Classification*. Cambridge University Press, Cambridge, UK, pp. 50–92

Lucy J 1992 *Grammatical Categories and Cognition*. Cambridge University Press, Cambridge, UK

Matsumoto Y 1993 Japanese numeral classifiers: A study of semantic categories and lexical organization. *Linguistics* **31**: 667–713

Mithun M 1986 The convergence of noun classification systems. In: Craig C G (ed.) *Noun Classification and Categorization*. John Benjamins, Amsterdam, pp. 379–97

Oates L 1964 A tentative description of the Gunwinggu language. *Oceania Linguistic Monographs* **10**, University of Sydney, Australia

Rehg K 1981 *Ponapean Reference Grammar*. University Press of H7awaii, Honolulu, HI

Sands K 1995 Nominal classification in Australia. *Anthropological Linguistics* **37**: 247–346

Senft G 2000 *Systems of Nominal Classification*. Cambridge University Press, Cambridge, UK

C. Grinevald

Classroom Assessment

Around the end of the 1980s, the traditional conception of teachers' classroom assessment roles began to change. Previously, teachers' classroom assessment responsibilities were narrowly focused on summative decisions narrowly related to grouping, grading, and selecting students. In the late 1980s, close examination of what teachers do and what decisions they are called on to make in their classrooms increased substantially

views of teachers' classroom lives, responsibilities, and assessments (Jackson 1990, Wittrock 1986). A synthesis of this research identified three generalizations that provided a useful perspective of the realities of the classroom teachers' classroom.

First, classrooms are both academic and social environments that teachers must master and understand to successfully instruct and interact with their students. Many teacher decisions are dependent on the social and academic knowledge they acquire about their students. Second, classrooms are busy, interactive, and ad hoc settings that call on the teacher to make many and varied decisions. Third, although many nonteachers view classrooms as a unified whole, teachers know that such uniformity is illusionary. In classrooms, teachers continually deal with a range of individual student concerns and issues. Understanding the implications of these three classroom realities produces a broad domain for assessments.

Given the richness and complexity of classrooms, a more realistic description of classroom assessments is the process of collecting, synthesizing, and interpreting information to aid teachers in making classroom decisions. While the overriding purpose is to make decisions, classroom assessment involves many different decisions and contexts. In particular, classroom assessments focus on learning about students at the start of school, planning and delivering instruction to students, and formally assessing student learning. Note that each of these three assessment focuses is dependent on the collection, synthesis, and interpretation of assessment. These three focuses are used to structure the discussion of classroom assessments below.

1. *Classroom Assessment at the Start of School*

The beginning days of school are important for both teachers and students. In the first few weeks of school the teacher and students must get to know and understand each other so that they can be organized into a classroom learning community. The activities in the early few days of school set the stage for how well students will behave, attend, and learn during the school year (Airasian 2001, Stiggins 1997). In the first early days of school, teachers have their antenna up, observing, listening, mentally recording, and assessing their perceptions of the students. In order to know how to group, teach, motivate, manage, accommodate, and reward students, the teacher must learn their particular characteristics.

Many forms of planned and unplanned, formal and informal, sources of information contribute to teachers' perceptions of their students, e.g., 'on the fly' observations, hearsay from the school grapevine, and prior teachers' comments, as well as information such as school records, formal assessment results, and performance in the classrooms. Two increasingly important areas teachers want to know about are student disabilities and medication. Note that although cognitive information is important to all teachers, the classroom society requires that affective and psychomotor student characteristics also must be identified.

Integrating the pieces of formal and informal information gathered in the first two or three weeks, the teacher forms a description or perception of each student and the class as a whole. This provides the teacher with the kind of nitty-gritty information needed to make the classroom function effectively (Good and Brophy 1997). Because teachers cannot spend a great deal of time assessing their students at the start of school, the validity and reliability of the initial student assessments are important. Two major validity concerns during start-of-school assessments are labeling students based on stereotypes and treating students' cultural or language differences as if they were student deficits (Oakes and Lipton 1999). The main concern of reliability is that teachers obtain sufficient and recurring information before labeling students. These important initial teacher assessments are often overlooked as an important and influential form of classroom assessment.

2. *Classroom Assessment in Planning and Delivering Instruction*

Teacher classroom decisions about planning and delivering instruction encompass a variety of issues. There are, for example, many considerations that teachers must recognize and assess in order to successfully plan lessons for their students. Student characteristics vary with student readiness, attention span, prior subject knowledge, attitude toward school, disabilities, and other characteristics that must be considered when planning instruction. Similarly, school and classroom resources range from textbooks to copying machines to sophisticated laboratory apparatus, and can hamper or enhance lesson planning. Time is a critical factor in planing, as every teacher knows. Also, teacher characteristics such as subject matter knowledge, physical limits, and preferred teaching style influence planning. A significant amount of assessment is involved in decision making for valid and viable lesson plans (Wragg 1997).

Once relevant information about the student, the teacher, and the instructional resources are identified, the teacher's task is to synthesize and decide how to construct a set of instructional plans containing educational objectives, instructional materials, teaching strategies, and assessment procedures (Airasian 2001). Different objectives call for different forms of instruction and assessment, and teachers must be able to teach students in more than one way. Objectives indicate the outcomes of student learning. Higher level objectives include cognitive processes such as application, analysis, synthesis, and evaluation. Lower

level objectives emphasize rote memorization. Because educational objectives are developed before instruction begins, teachers often must make a decision to adapt objectives, materials, and instructional strategies to suit student readiness and needs. Suitable strategies to accommodate students with disabilities also must be planned. The decisions teachers make to match instruction to objectives and make appropriate instruction for students with disabilities improve the validity of their instruction and assessment.

During teaching, the teacher is concerned with decisions and assessments to determine how well the instruction is progressing. Planning and instructing assessments are integrally related; the processes constantly cycles from planning to delivering to revising to planning and so on. There is a logical, continuous, and natural link between the two processes.

Oral questioning is the most common form of instructional assessment because it best fits the flow of instruction (Airasian 2001). During instruction, teachers ask questions for many reasons, to reinforce important points, to maintain students' attention, to assess student learning, and to promote deeper processing of important information. Teachers use both convergent questions that have a single correct answer and divergent questions that have more than one appropriate answers. Lower level questions tap recall and memorization while higher level questions tap processes more complex than recall. Classroom questioning strategies can be improved by asking questions related to important objectives, avoiding overly general questions, distributing questions among many students, allowing sufficient 'wait time' before calling on students, stating questions clearly to avoid confusion, probing student responses with follow-up questions such as 'why' or 'explain your answer,' and remembering that oral questioning is a social process in which student answers should be treated with respect, regardless of the quality of the answer.

3. *Assessments of Formal Learning*

3.1 *General Aspects of Formal Assessments*

Formal assessment is the culmination of planning and delivering instruction. It focuses on the extent to which students have learned from instruction. There is an important difference between good teaching and effective teaching. Good teaching refers to what teachers do during planning and delivering instruction. Effective teaching refers to whether students have learned from their instruction. Formal assessments are concerned with the effectiveness of learning from instruction. Formal assessments are also called summative assessments and commonly include tests, projects, term papers, lab reports, portfolios, performances, products, and final examinations. These are assessments that can have important consequences for students and therefore are taken seriously by students, parents, and teachers (Black 1998).

A fair and valid formal assessment includes information and skills similar to those presented in instruction. While the type of assessment strategy chosen to assess students depends on the nature of instruction, all types should represent the objectives and instruction presented. Factors such as the age of the students, the subject matter assessed, and the length of time for testing, all impact the length of formal assessments.

Obtaining fair and valid formal assessments involves alignment among objectives, instruction, and assessments, providing students with good instruction, and selecting appropriate strategies to assess learning. Formal assessments gather valid and reliable samples of student performance and use them to make generalizations about general student learning. The most important preparation for formal assessment is a good teacher. Students should also be familiar with the assessment item formats and be given a review session prior to the assessment.

If these factors are not met, invalid assessment results can occur. Other practices that diminish assessment validity are: failure to develop assessments based on objectives and instruction, failure to assess all the important objectives taught, failure to select item types that prevent students from showing their full performance, including topics or objectives not taught, including too few items to obtain adequate assessment reliability, and using tests to punish students. Further, the success of formal assessment can be undone if the test questions are faulty or confusing. Poorly constructed or unclear assessment questions do not provide students a fair chance to show what they have learned from instruction, and consequently, diminish assessment validity.

3.2 *Types of Formal Classroom Assessments*

There are many types of test items that are used in classrooms assessments. Selection items include multiple-choice, true–false, and matching items to which students respond by selecting an answer from a set of presented items. Supply items include short-answer, completion, and essay items to which students are required to create or supply their own answers. Selection items can cover many items in a short time and can be scored quickly. Supply items can be constructed quickly and permit students to provide their own constructed answers. Selection items are difficult to construct and encourage guessing, while supply items are difficult to score and cover smaller samples of instruction (Gronlund 1998).

Common guidelines for writing and critiquing test items include: (a) assess important objectives; (b) state items clearly, describing the students' task; (c) avoid

ambiguous and confusing wording and sentence structure—students should have a clear understanding of what is expected of them; (d) use vocabulary appropriate to the students being assessed; (e) write selection items that have one correct answer; (f) provide information about the nature and form of the desired response, particularly for essay questions; (g) avoid clues to correct answers; and (h) review items before assessing students.

In assembling and preparing items for a formal assessment the following suggestions should be applied: (a) group items of the same type; (b) place selection items first and supply items last; (c) provide directions for each type of test item; and (d) diminish assessment anxiety by giving advanced notice of the assessment, providing a review session before assessment, and most of all, providing students with good instruction. Many students experience anxiety before and during testing. While it is difficult to eliminate test anxiety, these strategies can lower it. Plan accommodations for students with disabilities. Two types of accommodations should be addressed, one for test administration (e.g., having directions read to students, giving extra time) and one for the test itself (e.g., divide the test into small section, provide a sample of each test item, arrange student items from concrete to abstract).

Unfortunately, cheating on classroom assessments is a fairly common occurrence. Forms of cheating range from looking at another's paper, bringing crib sheets into class, to other illicit strategies (Cizek 1999). No matter how or why it is done, cheating is dishonest and unacceptable. When cheaters state or imply that the work they have turned is their own, they are lying, and should be penalized. Useful strategies to discourage cheating on classroom assessments include spreading seating arrangement, careful proctoring, and movement around the classroom during testing.

Ultimately, all formal classroom assessments will be scored, usually by the classroom teacher. Scoring selection items is straightforward, efficient, and objective. Each student's score is compared to a scoring key and an overall score is obtained. Selection items are typically scored objectively, that is, two or more independent scorers would agree on a student's score. Supply and especially essay items tend to be more difficult and time-consuming to score. Because student responses to supply items are more lengthy and varied than those of selection items, the former are more likely to be subjectively scored. That is, scores of two or more independent scorers do not agree on the same or similar student score. Many factors influence essay subjectivity, including handwriting, spelling, neatness, and teacher fatigue. These factors are not central to the essay, but their presence influences the teacher's perception of students' essays and can influence the objectivity of essay scoring.

Two common essay scoring approaches are holistic and analytic. Holistic scoring provides a single score to describe the essay's quality. Analytic scoring breaks the essay down into component parts such as organization, spelling, accuracy, and grammar and gives each component an individual score. Holistic scoring is most used in grading students, while analytic scoring is most used to correct and improve initial drafts of written responses. To ensure objectivity in essay scoring, the following steps should be followed. Define what constitutes a good essay answer before it is administrated. Tell students whether handwriting, spelling, grammar, and punctuation will count in scoring the essay. If possible, score students essays anonymously. If multiple essays are in the assessment, score all students' answers to the first essay question before moving on to score the second essay item, and so on. Reread some of the essays a second time to determine the reliability of scoring.

In addition to selection and short-answer items, there are other important types of items that are important in classroom assessments. The most prominent of these item types is performance assessment, also referred to as authentic or alternative assessments (Mehrens et al. 1998). Performance assessments allow students to demonstrate what they know or can do in a real situation. Examples of performance assessments are essays, pronouncing a foreign word, setting up laboratory equipment, catching a ball, reciting a poem, identifying unknown chemicals, generalizing experimental data, working in cooperative groups, obeying school rules, and painting a picture. All of these performances require more than memorization and a one or two word response.

All performance assessments are developed in four steps: (a) identifying the purpose of the performance assessment; (b) stating the observable aspects of the performance, also called performance criteria; (c) selecting a suitable setting to carry out the performance assessment; and (d) scoring the quality of the performance. The key aspect of assessing performance assessments is the identification of the criteria that define a good performance. Performances are normally broken down into specific, observable criteria that can be individually assessed. Criteria should be specific and unambiguous. For example, stating 'information is presented in a logical sequence' is better than stating the more ambiguous 'has organization,' and 'can be heard in all parts of the room' is better than 'speaks correctly.' Statements of clear performance criteria are important for both holistic and analytic scoring approaches.

Multiple approaches to scoring students' performance assessments are available, and all are based on performance criteria. Checklists, rating scales, and scoring rubrics are most commonly used to assess performance assessments. A checklist is a written list of performance criteria that the teacher uses to judge student performance on each of the criteria. Checklists allow only 'yes' or 'no' judgments of each criterion. A rating scale is a written list of performance criteria that

permits the teacher more than two choices (e.g., good, fair, poor or excellent, good, fair, poor) to judge student performance of each criterion. A scoring rubric summarizes the overall performance on the criteria into holistic descriptions representing different levels of a student's overall performance. Rubrics describe performance in a summative way, while checklists and rating scales provide specific diagnostic information about each criterion in a formative way (Airasian 2001, Goodrich 1997).

Another important addition to performance assessment is the portfolio. A portfolio is a carefully selected collection of a student's performances that show accomplishments and improvements over time. Portfolios allow students and teachers to revisit and reflect prior work. Like any performance assessment, performance criteria are defined to identify and judge each of the individual pieces and the overall portfolio. As in all classroom assessments, the criteria should be aligned to the teacher's objectives (Arter and Spandel 1992). The purpose of performance assessment is the same as all formal classroom assessments, to determine how well students have learned from the instruction they were provided. To improve the validity and reliability of performance assessments, teachers should select performance criteria that are appropriate for their students, observe and record student performance while it is being performed rather than at some later date, judge student performance in terms of the performance criteria not the personal characteristics of the students and, if possible, observe a student's performance more than once.

Managing and scoring portfolios is a time-consuming activity, and teachers who attempt portfolio assessment are advised to start a portfolio with a single topic with a limited number of entries in the portfolio.

4. *Grading*

Grading is the formal process of judging the quality of a student's performance. Grades are always based on teacher judgment. However, the helping relationship that teachers have with their students can make it difficult to judge them in a completely objective manner. Further, since there is no uniformly accepted teacher grading strategy, teachers must find a grading approach that they feel is fair to themselves and to the students (Brookhart 1998, Frisbie and Waltman 1992).

All grading approaches are based on comparisons. The most common grading comparisons are norm-referenced and criterion-referenced grading. Norm-referenced grades are determined by comparing how a given student performed compared to the performance of other test takers. Norm-referenced grading is also called grading on the bell curve. Criterion-referenced grades are determined by comparing how a student performed in comparison to pre-established standards. In norm-referenced grading not all students can attain high scores, while in criterion-referenced grading they can if they all reach the standard. Grading students based on a comparison of student performance to the teacher's estimate of the student's ability is not recommended because estimating ability is difficult to do accurately. Similarly, grading based on student improvement over time is also not recommended. In general, regardless of the grading approach selected, it is strongly advised that grades be based mainly on students' academic performance.

5. *Assessment of Ethical Responsibilities*

Thus far discussion has focused on the technical aspects of classroom assessment. However, it is important to recognize that teachers' assessments have short-term and long-term consequences for students, thus requiring that teachers have an ethical responsibility to make decisions that are the most valid and reliable as possible. A number of groups in the USA have set standards for teachers' ethical performance (American Federation of Teachers et al. 1990 and National Education Association 1992–3). Among teachers' ethical responsibilities are: to provide students access to varying points of view; not to expose students to embarrassment or ridicule; not to exclude, deny, or grant advantages on the basis of students' race, color, creed, gender, national origin, religion, culture, sexual orientation, or disability; and not to label students with stereotypes.

This article has indicated that classrooms are complex environments that call upon teachers to make many and varied decisions. The bases for these decisions derive from a wide range of formal and informal assessment information. Although it is not expected that every teacher assessment decision will always be correct, it is expected that they can provide defensible assessment evidence to support classroom decisions. This should be expected in a context in which teachers' actions have important consequences for students.

See also: Classroom Climate; Educational Assessment: Major Developments; Instructional Design; Instructional Psychology; Performance Evaluation in Work Settings; Program Evaluation; Teacher Behavior and Student Outcomes; Teacher Expertise; Teaching and Learning in the Classroom; Test Administration: Methodology

Bibliography

Airasian P W 2001 *Classroom Assessment: Concepts and Applications*. McGraw-Hill, Boston

American Federation of Teachers, National Council on Measurement in Education, National Education Association 1990 Standards for teacher competence in educational assessment

of students. *Educational Measurement: Issues and Practice* **9**(4): 30–2
Arter J, Spandel V 1992 Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice* **11**: 36–44
Black P J 1998 *Testing: Friend or Foe? Theory and Practice of Assessment and Testing*. Falmer Press, London
Brookhart S M 1998 *Teaching about Grading and Communicating Results*. School of Education, Duquesne University, Pittsburgh, PA
Cizek G J 1999 *Cheating on Tests: How to Do it, Detect it, and Prevent it*. Erlbaum Associates, Mahwah, NJ
Frisbie D A, Waltman K K 1992 Developing a personal grading plan. *Educational Measurement: Issues and Practice* **11**(3): 35–42
Good T L, Brophy J E 1997 *Looking in Classrooms*, 7th ed. Longman, New York
Goodrich H 1997 Understanding rubrics. *Educational Leadership* **54**(4): 14–17
Gronlund N E 1998 *Assessment of Student Achievement*. Allyn & Bacon, Boston
Jackson P W 1990 *Life in Classrooms*. Teachers College Press, New York
Mehrens W A, Popham W J, Ryan J M 1998 How to prepare students for performance assessments. *Educational Measurement: Issues and Practice* **17**(1): 18–22
National Education Association 1992–3 Ethical Standards for Teachers' Relations with Pupils. In: *NEA Handbook*. National Education Association, Washington, DC, pp. 366–7
Oakes J, Lipton M 1999 *Teaching to Change the World*, 1st ed. McGraw-Hill, Boston
Stiggins R J 1997 *Student-centered Classroom Assessment*, 2nd edn. Merrill, Uppersaddle River, NJ
Wittrock M C (ed.) 1986 *Handbook of Research on Teaching*. Macmillan, London
Wragg T 1997 *Assessment and Learning*. Routledge, London

P. W. Airasian

Classroom Climate

In the 30 years since the pioneering use of classroom climate assessments in an evaluation of Harvard Project Physics (Walberg and Anderson 1968), the field has undergone remarkable growth, diversification, and internationalization. Literature reviews (Fraser 1994, 1998) place these developments into historical perspective and show that classroom climate assessments have been used as a source of dependent and independent variables in a variety of research applications spanning many countries. The assessment of classroom climate and research applications has involved a variety of quantitative and qualitative methods, and an important accomplishment within the field has been the productive combination of quantitative and qualitative research methods (Tobin and Fraser 1998).

A historical look at the field of classroom climate over the past few decades shows that a striking feature is the availability of a variety of economical, valid, and widely applicable questionnaires that have been developed and used for assessing students' perceptions of classroom climate. This article makes some of these valuable instruments readily available by describing some major questionnaires and their past application in various lines of research.

Although using students' and teachers' perceptions to study classroom climate forms the focus of this article, this method can be contrasted with the external observer's direct observation and systematic coding of classroom communication and events and the techniques of naturalistic inquiry, ethnography, case study, or interpretive research. In the method considered in detail in this article, defining the classroom climate in terms of the shared perceptions of the students and teachers has the dual advantage of characterizing the setting through the eyes of the participants themselves, and capturing data which the observer could miss or consider unimportant.

1. *Instruments for Assessing Classroom Climate*

Historically, the development of classroom climate instruments commenced three decades ago with the appearance of *Learning Environment Inventory* (LEI) and *Classroom Environment Scale* (CES). The LEI was developed in conjunction with evaluation and research related to Harvard Project Physics (Walberg and Anderson 1968). The respondent expresses degree of agreement with each of 105 statements (seven per scale) using the four response alternatives of strongly disagree, disagree, agree, and strongly agree. The names of some of the scales are cohesiveness, speed, difficulty, goal direction, and disorganization. The CES (Moos and Trickett 1987) grew from a comprehensive program of research involving perceptual measures of a variety of human environments including psychiatric hospitals, prisons, university residences and work milieus (Moos 1974). The final published version contains nine scales with ten items of true–false response format in each scale. Scales include involvement, teacher support, task orientation, and innovation.

Three more contemporary classroom climate instruments are described below: *Science Laboratory Environment Inventory* (SLEI); *Constructivist Learning Environment Survey* (CLES); and *What Is Happening In This Class* (WIHIC) questionnaire.

1.1 *Science Laboratory Environment Inventory*

Because of the importance of laboratory settings in science education, an instrument specifically suited to assessing the climate of science laboratory classes at the senior high school or higher education levels was developed (McRobbie and Fraser 1993). The SLEI

has five scales (student cohesiveness, open-endedness, investigation, rule clarity, and material environment), each with seven items. The five response alternatives are almost never, seldom, sometimes, often, and very often. Typical items are 'I use the theory from my regular science class sessions during laboratory activities' (integration) and 'we know the results that we are supposed to get before we commence a laboratory activity' (open-endedness).

1.2 Constructivist Learning Environment Survey

The CLES (Taylor et al. 1997) was developed to assist researchers and teachers to assess the degree to which a particular classroom's climate is consistent with a constructivist epistemology, and to assist teachers to reflect on their epistemological assumptions and reshape their teaching practice. The CLES has 36 items with five response alternatives ranging from almost never to almost always. The scales are personal relevance, uncertainty, critical voice, shared control, and student negotiation. Typical items are 'I help the teacher to decide what activities I do' (shared control) and 'other students ask me to explain my ideas' (student negotiation).

1.3 What Is Happening In This Class (WIHIC) Questionnaire

The WIHIC questionnaire brings parsimony to the field of classroom climate by combining modified versions of the most salient scales from a wide range of existing questionnaires with additional scales that accommodate contemporary educational concerns (e.g., equity and constructivism). Whereas an Australian sample of 1,081 students in 50 classes responded to the original English version, a Taiwanese sample of 1,879 students in 50 classes responded to a Chinese version that had undergone careful procedures of translation and back translation (Aldridge et al. 1999). This led to a final form of the WIHIC containing the seven eight-item scales of student cohesiveness, teacher support, involvement, investigation, task orientation, cooperation, and equity.

1.4 Different Forms of Questionnaires

The instruments discussed above have not only a form to measure perceptions of 'actual' or experienced classroom climate, but also another form to measure perceptions of 'preferred' or ideal classroom climate. The preferred forms are concerned with goals and value orientations and measure perceptions of the classroom climate ideally liked or preferred. For example, an item in the actual form such as 'there *is* a clear set of rules for students to follow' would be changed in the preferred form to 'there *would be* a clear set of rules for students to follow.'

Tobin and Fraser (1998) point out that there is potentially a problem with nearly all existing classroom climate instruments when they are used to identify differences between subgroups within a classroom (e.g., males and females) or in the construction of case studies of individual students. The problem is that items elicit an individual student's perceptions of the class as a whole, as distinct from a student's perceptions of his/her own role within the classroom. For example, items in the traditional class form might seek students' opinions about whether 'the work of the class is difficult' or whether 'the teacher is friendly towards the class.' In contrast, a personal form of the same items would seek opinions about whether 'I find the work of the class difficult' or whether 'the teacher is friendly towards me.' For these reasons, most of the questionnaires discussed above have a personal form.

Comprehensive statistics supporting the validity and reliability of the above questionnaires are provided in Fraser (1998).

2. Research Involving Classroom Climate Instruments

2.1 Associations Between Student Outcomes and Classroom Climate

The strongest tradition in past classroom climate research has involved investigation of associations between students' cognitive and affective learning outcomes and their perceptions of psychosocial characteristics of their classrooms. Fraser's (1994) tabulation of 40 past studies shows that associations between outcome measures and classroom climate perceptions have been replicated for a variety of cognitive and affective outcome measures, a variety of classroom climate instruments, and a variety of samples (ranging across numerous countries and grade levels). Using the SLEI, associations with students' cognitive and affective outcomes were found for a sample of approximately 80 senior high school chemistry classes in Australia (Fraser and McRobbie 1995 and Fraser 1993), 489 senior high school biology students in Australia (Fisher et al. 1997) and 1592 grade 10 chemistry students in Singapore (Wong and Fraser 1996).

2.2 Evaluation of Educational Innovations

Classroom climate instruments can be used as a source of process criteria in the evaluation of educational innovations. An evaluation of the Australian Science Education Project (ASEP) revealed that, in comparison with a control group, ASEP students perceived

their classrooms as being more satisfying and individualized and having a better material environment (Fraser 1979). The significance of this evaluation is that classroom climate variables differentiated revealingly between curricula, even when various outcome measures showed negligible differences. Recently, the incorporation of a classroom climate instrument within an evaluation of the use of a computerized database revealed that students perceived that their classes became more inquiry oriented during the use of the innovation (Maor and Fraser 1996). In an evaluation of an urban systemic reform initiative in the USA, use of the CLES painted a disappointing picture in terms of a lack of success in achieving constructivist oriented reform of science education (Dryden and Fraser 1996).

2.3 Differences Between Student and Teacher Perceptions of Actual and Preferred Climate

An investigation of differences between students and teachers in their perceptions of the same actual classroom climate and of differences between the actual climate and that preferred by students or teachers was reported by Fisher and Fraser (1983) for a sample of 116 classes for the comparisons of student actual with student preferred scores, and a subsample of 56 of the teachers of these classes for contrasting teachers' and students' scores. Students preferred a more positive classroom climate than was actually present for all five climate dimensions. Also, teachers perceived a more positive classroom climate than did their students in the same classrooms on four of the dimensions. These results replicate patterns emerging in other studies in other countries (Fraser 1998).

2.4 Teachers' Attempts to Improve Classroom Climates

Feedback information based on student or teacher perceptions has been employed in a five-step procedure as a basis for reflection upon, discussion of, and systematic attempts to improve classroom climate (Yarrow et al. 1997). First, all students in the class respond to the preferred form of a classroom climate instrument, while the actual form is administered in the same time slot about a week later (assessment). Second, the teacher is provided with feedback information derived from student responses in the form of profiles representing the class means of students' actual and preferred climate scores (feedback). These profiles permit identification of the changes in classroom climate needed to reduce major differences between the nature of the actual climate and that preferred by students. Third, the teacher engages in private reflection and informal discussion about the profiles in order to provide a basis for a decision about whether an attempt would be made to change the climate in terms of some of the dimensions (reflection and discussion). The main criteria used for selection of dimensions for change are, first, that there should be a sizeable actual-preferred difference on that variable and, second, that the teacher should feel concerned about this difference and want to make an effort to reduce it. Fourth, the teacher introduces an intervention of about two months' duration in an attempt to change the classroom climate (intervention). For example, strategies used to enhance the dimension of teacher support could involve the teacher moving around the class more to mix with students, providing assistance to students, and talking with them more than previously. Fifth, the student actual form of the scales is re-administered at the end of the intervention to see whether students perceive their classroom climate differently from before (reassessment).

Yarrow et al. (1997) reported a study in which 117 preservice education teachers were introduced to the field of classroom climate through being involved in action research aimed at improving their university teacher education classes and their 117 primary school classes during teaching practice. Improvements in classroom climate were observed, and the preservice teachers generally valued both the inclusion of the topic of classroom climate in their preservice programs, and the opportunity to be involved in action research aimed at improving classroom climate.

2.5 Combining Quantitative and Qualitative Methods

Significant progress has been made towards the desirable goal of combining quantitative and qualitative methods within the same study in research on classroom climates (Tobin and Fraser 1998). Fraser's (1999) multilevel study of classroom climate incorporated a teacher-researcher perspective as well as the perspective of six university-based researchers. The research commenced with an interpretive study of a Grade 10 teacher's classroom at a school which provided a challenging classroom learning climate in that many students were from working-class backgrounds, some were experiencing problems at home, and others had English as a second language. Qualitative methods involved several of the researchers visiting this class each time it met over five weeks, using student diaries, and interviewing the teacher-researcher, students, school administrators, and parents. A video camera recorded activities for later analysis. Field notes were written during and soon after each observation, and team meetings took place three times weekly. The qualitative component of the study was complemented by a quantitative component involving the use of a questionnaire which linked three levels: the class in which the interpretive study was undertaken; selected classes from within the school; and classes distributed throughout the same State.

This enabled a judgment to be made about whether this teacher was typical of other teachers at the same school, and whether the school was typical of other schools within the State. Some of the features identified as salient in this teacher's classroom climate were peer pressure and an emphasis on laboratory activities.

2.6 Cross-national Studies

Educational research which crosses national boundaries offers much promise for generating new insights for at least two reasons (Aldridge et al. 1999). First, there usually is greater variation in variables of interest (e.g., teaching methods, student attitudes) in a sample drawn from multiple countries than from a one-country sample. Second, the taken-for-granted familiar educational practices, beliefs, and attitudes in one country can be exposed, made 'strange,' and questioned when research involves two countries. In a recent cross-national study, six Australian and seven Taiwanese researchers worked together on a study of classroom climate. The WIHIC was administered to 50 junior high school science classes in Taiwan (1,879 students) and 50 classes in Australia (1,081 students) (Aldridge et al. 1999). An English version of the questionnaire was translated into Chinese, followed by an independent back translation of the Chinese version into English, again by team members who were not involved in the original translation. Qualitative data, involving interviews with teachers and students and classroom observations, were collected to complement the quantitative information and to clarify reasons for patterns and differences in the means in each country.

Data from the questionnaires guided the collection of qualitative data. Student responses to individual items were used to form an interview schedule to clarify whether items had been interpreted consistently by students and to help to explain differences in questionnaire scale means between countries. Classrooms were selected for observation on the basis of the questionnaire data, and specific scales formed the focus for observations in these classrooms. The qualitative data provided valuable insights into the perceptions of students in each of the countries, helped to explain some of the differences in the means between countries, and highlighted the need for caution when interpreting differences between the questionnaire results from two countries with cultural differences.

2.7 Transition from Primary to High School

There is considerable interest in the effects on early adolescents of the transition from primary school to the larger, less personal climate of the junior high school at this time of life. Midgley et al. (1991) reported a deterioration in the classroom climate when students moved from generally smaller primary schools to larger, departmentally organized lower secondary schools, perhaps because of less positive student relations with teachers and reduced student opportunities for decision making in the classroom. Ferguson and Fraser's (1998) study of 1,040 students from 47 feeder primary schools and 16 linked high schools in Australia also indicated that students perceived their high school classroom climates less favorably than their primary school classroom climates, but the transition experience was different for boys and girls and for different school size 'pathways.'

3. Conclusion

The major purpose of this article has been to make this exciting research tradition involving classroom climate more accessible to wider audiences by portraying several widely applicable instruments for assessing perceptions of classroom climate and by describing several major lines of previous research.

See also: Classroom Assessment; Educational Assessment: Major Developments; Environments for Learning; Group Processes in the Classroom; School as a Social System; Teacher Behavior and Student Outcomes; Teaching and Learning in the Classroom

Bibliography

Aldridge J M, Fraser B J, Huang T-C I 1999 Investigating classroom environments in Taiwan and Australia with multiple research methods. *Journal of Educational Research* **93**: 48–62

Ferguson P D, Fraser B J 1998 Changes in learning environment during the transition from primary to secondary school. *Learning Environments Research* **1**: 369–83

Fisher D L, Fraser B J 1983 A comparison of actual and preferred classroom environment as perceived by science teachers and students. *Journal of Research in Science Teaching* **20**: 55–61

Fraser B J 1979 Evaluation of a science-based curriculum. In: Walberg H J (ed.) *Educational Environments and Effects: Evaluation, Policy, and Productivity*. McCutchan, Berkeley, CA, pp. 218–34

Fraser B J 1994 Research on classroom and school climate. In: Gabel D (ed.) *Handbook of Research on Science Teaching and Learning*. Macmillan, New York, pp. 493–541

Fraser B J 1998 Science learning environments: assessment, effects and determinants. In: Fraser B J, Tobin K G (eds.) *International Handbook of Science Education*. Kluwer, Dordrecht, The Netherlands, pp. 527–64

Fraser B 1999 'Grain sizes' in learning environment research: combining qualitative and quantitative methods. In: Waxman H, Walberg H (eds.) *New Directions for Research on Teaching*. McCutchan, Berkeley, CA, pp. 285–96

Maor D, Fraser B J 1996 Use of classroom environment perceptions in evaluating inquiry-based computer assisted learning. *International Journal of Science Education* **18**: 401–21
Midgley C, Eccles J S, Feldlaufer H 1991 Classroom environment and the transition to junior high school. In: Fraser B J, Walberg H J (eds.) *Educational Environments: Evaluation, Antecedents and Consequences*. Pergamon, London, pp. 113–39
Moos R H 1974 *The Social Climate Scales: An Overview*. Consulting Psychologists Press, Palo Alto, CA
Moos R H, Trickett E J 1987 *Classroom Environment Scale manual*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA
Taylor P C, Fraser B J, Fisher D L 1997 Monitoring constructivist classroom learning environments. *International Journal of Educational Research* **27**: 293–302
Tobin K, Fraser B J 1998 Qualitative and quantitative landscapes of classroom learning environments. In: Fraser B J, Tobin K G (eds.) *International Handbook of Science Education*. Kluwer, Dordrecht, The Netherlands, pp. 623–40
Walberg H J, Anderson G J 1968 Classroom climate and individual learning. *Journal of Educational Psychology* **59**: 414–19
Yarrow A, Millwater J, Fraser B J 1997 Improving university and primary school classroom environments through preservice teachers' action research. *International Journal of Practical Experiences in Professional Education* **1**(1): 68–93
Dryden M,, Fraser B J 1996 *Evaluating Urban Systemic Reform Using Classroom Learning Environment Instruments*. Paper presented at the annual meeting of the American Educational Research Association, New York
Fisher D, Henderson D, Fraser B 1997 Laboratory environments and student outcomes in senior high school biology. *American Biology Teacher* **59**: 214–19
Fraser B J, McRobbie C J 1995 Science laboratory classroom environments at schools and universities: A cross-national study. *Educational Research and Evaluation* **1**: 289–317
McRobbie C J, Fraser B J 1993 Associations between student outcomes and psychosocial scince environment. *Journal of Educational Research* **87**: 78–85
Wong W L F, Fraser B J 1996 Environment-attitude associations in the chemistry laboratory classroom. *Research in Science and Technological Education* **14**: 91–102

B. J. Fraser

Cleavages: Political

'Political cleavages' are political divisions among citizens rooted in the structure of a given social system. However, although cleavages are political divisions, not all political divisions among citizens spring from structural cleavages. For one to talk of 'cleavages' such divisions must be permanent and noncontingent. They must orient people's behavior and sense of belonging stably and constantly. Political cleavages are the partisan expression of an underlying division among the members of a given society (whether national, subnational, or supranational).

1. *The Lipset–Rokkan Model*

The concept of 'cleavage' has been current in the social sciences for some time, although it was given full development only in the 1960s by Seymour Martin Lipset and Stein Rokkan. Both of them political sociologists by training, Lipset and Rokkan (1967) sought to redefine and specify the 'social bases of politics.' Writing when structural-functionalism was at its height—and, therefore, influenced by the Parsonian theory which assigned to the political parties the function of encapsulating social conflicts and stabilizing the social system—they set out to explain the persistence of party systems in the European democracies. In the 1960s, in fact, those systems still displayed features similar to those that had been institutionalized at the beginning of the century. Not surprisingly, their explanation was called the theory of the 'freezing' of the European party systems.

Their method was primarily historical–sociological in so far as it connected existing political divisions in the European countries with the principal cleavages that had opened up in the course of their development, from the birth of the nation-state in the sixteenth century to its full democratic maturation in the twentieth. The specific political cleavages that gave rise to the modern party systems accordingly were seen to be the result of two great historical processes: the one that had bred national revolutions (and, therefore, the formation of the modern European nation-states), and the one that had engendered the industrial revolution (and, therefore, the formation of modern European capitalist systems).

National revolutions had created two structural divisions: (a) between the center and the periphery, or between the groups and areas that sought to impose a single public authority on a given territory and the groups and areas which asserted their traditional autonomy against such centralizing pressures; (b) between the lay state and the church, or between groups which sought to separate temporal from religious authority and groups intent on preserving the intimate connection between them. The industrial revolution in its turn created two further structural divisions: (a) between agriculture and industry, or between groups and areas whose survival depended on traditional activities and groups and areas which endeavored to remove traditional constraints in order to foster the growth of new activities and production methods; (b) between capital and labor, or between the groups that dominated the new industrial structure and the workers, whose only possession was their capacity to perform labor.

In Europe, only the parties that reflected these cleavages were able to survive, that is, reproduce themselves electorally and institutionally. The institutionalized interaction among these parties gave rise to the modern party systems which, in individual European countries, and in forms that differed from

one country to another, still conserved in the mid-twentieth century the cleavages that had arisen in previous ones.

2. Subsequent Debate

The Lipset–Rokkan model heavily influenced the debate conducted during the 1960s on the political parties. The discussion started from the premise that political parties were necessary to make democracy safe (i.e., stable), as Schattschneider (1948) had already argued. However, the model was not endorsed universally, at least in its entirety. In a study of a small Scandinavian democracy, Eckstein (1966) pointed out the existence of multiple political divisions, identifying ones due to specific disagreements on particular public policies, others due to cultural divergences on interpretations of political life, and yet others arising from segmental cleavages caused by objective social differences. Again in 1966, Daalder examined the small democracies of continental Europe and pointed out the existence of political divisions due to factors (for instance, the nature of the political regime or the concept of nationality) other than those envisaged by the Lipset–Rokkan model.

But it was Sartori (1969) who challenged most radically the Lipset–Rokkan model, by reversing its causal logic. For Sartori, it was not social divisions that encouraged the birth of parties; rather, it was the parties that gave visibility and identity to a particular structure of social divisions. In short, Sartori argued, political sociology (and political science) should take the place of sociology of politics if partisan politics in the European democracies were to be understood properly. Lipset (1970) himself acknowledged the ability of parties to exacerbate politically a cleavage that might socially be in decline. Nonetheless, he reiterated that a social basis was necessary for a party to exist. Thus, while for Lipset and Rokkan social cleavages were necessary, though not sufficient, for the formation of parties and of party systems, for Sartori they were neither necessary nor sufficient because politics can only be conducted independently of other social spheres. This autonomy of the parties from society had already been shown by Kirchheimer (1966) in his celebrated study in which he investigated the transition from the 'party of social integration' to the 'catch-all party,' that is, a party able to represent diverse classes and social groups electorally.

From the 1970s onwards, partly due to the development of more sophisticated techniques of social research, the debate moved in a more microempirical and less macrohistorical direction. The decade saw numerous studies of electoral behavior, although their results were equivocal. While early studies like Rose (1974) showed the relative decline of politics based on social cleavages (or 'cleavage politics,' as it came to be called), the magnitude and implications of this decline were given various interpretations by scholars. The 1992 study by Franklin suggested that the decline of cleavage politics was ineluctable, those of Inglehart (1977), Dalton et al. (1984) and subsequent studies until seemingly showed that cleavage politics were evolving in a new direction so that 'cultural' cleavages were now taking the place of fading social cleavages and reorienting electoral and political behavior.

For these authors, the new structure of divisions might indeed have a 'social basis,' but it was manifest in a clash of values: between industrial values (in favor of the quantitative growth of affluence) on the one hand, and postindustrial ones (which gave priority to the quality of life and the protection of the environment) on the other. Associated with each side were socioeconomic groups and geographical areas, but the clash involved distinct (and opposed) cultural conceptions and lifestyles. Of course, there was no lack of criticism of this approach—especially by Bartolini and Mair (1990)—given that it emptied Lipset and Rokkan's original concept of cleavage of much of its meaning. For this reason, Bartolini and Mair proposed the following redefinition of the notion: (a) empirically, a cleavage must be definable in terms of social structure; (b) normatively, a cleavage is a system of values which gives a sense of collective identity to a social group; (c) behaviorally, a cleavage is manifest in the interaction among political actors. Thus redefined, the concept of cleavage is broader in its compass and becomes a means to order social relations.

3. The Freezing of Cleavages

Sociologists and economists also joined the debate. Goldthorpe (1996), for example, found that traditional social divisions were still conditioning political allegiances and electoral choices at the end of the twentieth century. Other studies appeared which, although they extended the concept of social cleavage, continued to frame it in structural terms. Lijphart (1977), in his study of the small consociative democracies of continental Europe, and then in his analyses of the established democracies (Lijphart 1999), showed that ethnic divisions performed the same function in structuring identity and behavior as did the other social divisions of the Lipset–Rokkan model. These divisions, too, sprang from the long historical process that had led to the formation of the nation-state. Thereafter, they had continued to predominate despite the divisions created by the process of industrialization. In the nation-states, the divisions between agriculture and industry, and between capital and labor were absorbed by more basic ethnic-linguistic cleavages. According to Lijphart, the diverse nature of these cleavages lay at the origin of the two principal models of democracy (what he called 'consensual' and 'majoritarian') that developed in the West after the World War II.

The model of consensual democracy based on the inclusion in the executive of all the country's main ethnic groups proved highly effective (in stabilizing democracy). It was accordingly used by authors (starting from Sartori and his studies of party systems in the 1970s) to investigate the workings of national societies connoted by identity divisions, albeit based on ideology rather than ethnicity or language. The reference here is to the postwar European democracies distinguished by the presence of powerful communist parties. Even these democracies were consensual in nature, although their operation was sustained, not by inclusive coalitions in the executive (access to which was barred to communist parties, owing to the geopolitical cleavages created by the Cold War), but by consensual practices in parliament. However, while these ideological cleavages proved unstable with the passage of time, this was not the case of ethnic ones. It seemed, indeed, that the model of consensual democracy had ended up by 'freezing' ethnic allegiances, though managing to cushion their impact.

The reasons why party systems were frozen in the postwar European democracies were expressly investigated by Mair and Bartolini (1990). These two authors examined three different hypotheses with regard to the freezing process. First, it may involve the freezing of social cleavages, that is, the stabilization of the social structure from which the parties draw legitimation for their political action. Second, the freezing may be due to the institutionalization of the political parties, albeit accompanied by the fading of the social divisions that had prompted their formation (here by 'institutionalization' is meant the parties' ability to stand as the only practicable electoral choices). Third, the freezing may be due to the stabilization of the party system as such, or put otherwise, the institutionalization of the system of interactions among the main political actors. Mair and Bartolini seem to suggest that the third of these hypotheses is the most plausible, given that both the hypothesis of the freezing of social cleavages and that of the freezing of political parties must admit to so many exceptions that they are not falsifiable. In short, for both authors a distinction must be drawn between the freezing of party systems and the freezing of individual parties.

The freezing hypothesis has also been discussed in terms of voting behavior. Several surveys have sought—using different indicators—to collect reliable data on the stability and instability of voting choices. Many scholars, from Pederson (1983) to Maguire (1983) and especially Bartolini (2000), have shown that rates of aggregate electoral volatility were relatively low until the 1980s: which corroborated Lipset and Rokkan's original contention that continuity rather than change was the distinguishing feature of partisan politics in Europe. These studies came in for criticism, of course, mainly on the grounds that electoral stability does not necessarily coincide with stable interaction among the political parties, and that stability may conceal processes of dealignment and realignment sufficient to gainsay the logic of the Lipset–Rokkan model.

4. *Between Europe and America*

The theory of cleavages has been developed on the basis of the experiences of the Western European countries, with no reference to the other great model of democracy: that of the United States. And yet it was precisely in the United States that modern political parties and party systems were invented. Indeed, the myth of American exceptionalism has been fostered by this European neglect; a neglect motivated by the belief that American society, unlike those of European countries, is based on cross-cutting cleavages which—as Lipset maintained as early as 1963—are unable to produce stable divisions among citizens. This absence of cleavage politics, the argument ran, gave rise to the depolarization of partisan conflict in the United States that underpinned the stability of 'American democracy.' In short, the more cleavages multiply and interweave, the more numerous the divisions among citizens become, and the safer democracy grows.

And yet, as Bensel (1987) showed, the situation in the United States was not so clear-cut, for that country, too, displayed, and still does, a stable political cleavage; sectional rather than social and cultural, although it has latterly acquired these features as well. This is the political cleavage between states and regional areas expressed in two radically different conceptions of the balance of powers to be struck between the states and the center of the federation. And it should not be forgotten that this fracture provoked one of the most violent and bloody civil wars of the modern age. It is around this cleavage that the various party systems that have arisen since the foundation of American republic have structured themselves.

In the light of the postnational experience of Europe at the end of the twentieth century, the case of the United States appears less exceptional than it did in the past. This is because the process of European integration has generated a sectional divide among geo-economic areas of the continent which cuts across the traditional (in Europe) party-political axis ranging from right to left. And in this case, too, the new contraposition has taken the form of a different interpretation of the balance of powers that should be established between the European and national institutions. Can European integration be regarded as a further historical cleavage—in addition to those singled out by the Lipset–Rokkan model—destined to produce another political structural cleavage? If so, the cleavage theory might be updated, this time bridging the European and American experiences.

See also: Conflict/Consensus; Conflict Sociology; Ethnic Conflict, Geography of; Ethnic Conflicts; Party Systems; Pluralism; Political Geography; Political Sociology; Race Relations in the United States, Politics of

Bibliography

Bartolini S 2000 *The Class Cleavage: The Political Mobilization of the European Left, 1860–1980*. Cambridge University Press, Cambridge, UK

Bartolini S, Mair P 1990 *Identity, Competition and Electoral Availability*. Cambridge University Press, Cambridge, UK

Bensel R F 1987 *Sectionalism and American Political Development: 1880–1980*. University of Wisconsin Press, Madison, WI

Daalder H 1966 Parties, elites and political development(s) in Western Europe. In: LaPalombara J, Weiner M (eds.) *Political Parties and Political Development*. Princeton University Press, Princeton, NJ

Dalton R J, Flanagan C S, Beck P A 1984 *Electoral Change in Advanced Industrial Democracies: Realignment or Dealignment?* Princeton University Press, Princeton, NJ

Eckstein H 1966 *Division and Cohesion in Democracy: A Study of Norway*. Princeton University Press, Princeton, NJ

Franklin M T, Mackie T, Valen H (eds.) 1992 *Electoral Change: Responses to Social and Attitudinal Structures in Western Countries*. Cambridge University Press, Cambridge, UK

Goldthorpe J H 1996 Class and politics in advanced industrial societies. In: Lee D J, Turner B S (eds.) *Conflict about Class: Debating Inequality in Late Industrialism*. Longman, London

Inglehart R 1977 *The Silent Revolution. Changing Values and Political Styles among Western Publics*. Princeton University Press, Princeton, NJ

Kirchheimer O 1966 The transformation of the Western European Party systems. In: LaPalombara J, Weiner M (eds.) *Political Parties and Political Development*. Princeton University Press, Princeton, NJ

Knutsen O 1988 The impact of structural and ideological party cleavages in Western European democracies: A comparative empirical analysis. *British Journal of Political Science* **18**: 323–52

Lijphart A 1977 *Democracy in Plural Societies: A Comparative Exploration*. Yale University Press, New Haven, CT

Lijphart A 1999 *Patterns of Democracy*. Yale University Press, New Haven, CT

Lipset S M 1963 *Political Man: The Social Bases of Politics*. Johns Hopkins University Press, Baltimore, MD

Lipset S M 1970 *Revolution and Counterrevolution*. Anchor Books, New York

Lipset S M, Rokkan S 1967 Cleavage structures, party systems and voter alignments: An introduction. In: Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments*. Free Press, New York

Maguire M 1983 Is there still persistence? Electoral change in Western Europe, 1948–1979. In: Daalder H, Mair P (eds.) *Western European Party Systems Continuity and Change*. Sage, Beverly Hills, CA

Pederson M N 1983 Changing patterns of electoral volatility in European party systems, 1948–1977. In: Daalder H, Mair P (eds.) *Western European Party Systems. Continuity and Change*. Sage, Beverly Hills, CA

Rose R 1974 *Electoral Behaviour: A Comparative Handbook*. Free Press, New York

Sartori G 1969 From sociology of politics to political sociology. In: Lipset S M (ed.) *Politics and the Social Sciences*. Oxford University Press, New York

Sartori G 1976 *Parties and Party Systems*. Cambridge University Press, New York

Schattschneider E E 1948 *The Struggle for Party Government*. University of Maryland, College Park, MD

S. Fabbrini

Climate Change and Health

This article outlines the potential impacts on human health of climate change due to the accumulation of greenhouse gases in the earth's atmosphere. It describes the range of potential mechanisms by which health could be affected and the difficulties of estimating the magnitude of such effects. It concludes with a brief discussion of how, not withstanding the need to prevent climate change as far as possible, humankind will have to adapt to changing climate if the adverse effects are to be minimized.

1. Background

1.1 Climate Change

Human activities, particularly the burning of fossil fuels, but also changes in land use, are leading to the accumulation of greenhouse gases such as carbon dioxide and methane in the earth's atmosphere. The resulting increase in 'radiative forcing' is leading to warming of the earth's surface. The United Nations set up the Intergovernmental Panel on Climate Change (IPCC)—a multidisciplinary body of scientific advisers, which in its third assessment report forecast an increase in the average global temperature of 1.4–5.8 °C between 1990 and 2100 (IPCC 2001). There are a number of sources of uncertainty in projections of future climate, including changes in greenhouse gas emissions and concentrations, the sensitivity of climate to greenhouse gases, and the impact of modulating processes, such as the short-term cooling effects of aerosols as a result of industrial emissions. However, it does appear likely that the rate of climate change over the twenty-first century will be far greater than any natural changes in world climate over the past 10,000 years. There has been substantial warming since 1856, when records began, with particularly rapid increases in temperatures since about 1980. The warmest year on record was 1998, partly as a result of the marked El Niño which occurred over 1997–8. In its second assessment report the IPCC concluded that the balance of evidence suggested that the impact of human

Table 1
Mediating processes and direct and indirect potential effects on health of changes in temperature and weather

Mediating process	Health outcome
Direct effects	
Exposure to thermal extremes	Changed rates of illness and death related to heat and cold
Changed frequency or intensity of other extreme weather events	Deaths, injuries, pyschological disorders; damage to public health infrastructure
Indirect effects	
Disturbances of ecological systems: Effect on range and activity of vectors and infective parasites	Changes in geographical ranges and incidence of vector borne disease
Changed local ecology of water borne and food borne infective agents	Changed incidence of diarrheal and other infectious diseases
Changed food productivity (especially crops) through changes in climate and associated pests and diseases	Malnutrition and hunger, and consequent impairment of child growth and development
Sea level rise with population displacement and damage to infrastructure	Increased risk of infectious disease psychological disorders
Biological impact of air pollution changes (including pollens and spores)	Asthma and allergies; other acute and chronic respiratory disorders and deaths
Social, economic, and demographic dislocation through effects on economy, infrastructure, and resource supply	Wide range of public health consequences: mental health and nutritional impairment, infectious diseases, civil strife

Source: McMichael and Haines 1997

activities on global climate was now discernible (IPCC 1996). The subsequent IPCC report pointed to 'new and stronger evidence that most of the warming over the last 30 years was attributable to human activities.'

1.2 Other Global Environmental Changes

Climate change is not occurring in isolation and there are a range of other global changes—stratospheric ozone depletion, loss of biodiversity, changes in land use patterns and depletion of aquifers, all of which may also have effects on human health and society. There are linkages between climate change and some of these other phenomena, for example, the rise in the temperature of the lower atmosphere may increase stratospheric ozone depletion. Deforestation leads to loss of biodiversity, particularly when it involves tropical forests, and also results in a release of substantial amounts of carbon dioxide into the atmosphere. Population growth in developing countries and unsustainable patterns of consumption in industrialized nations are increasing the strain on earth's life support systems and the demand for energy (McMichael and Powles 1999).

2. Potential Impacts on Health

2.1 Range of Effects

Climate change is likely to have substantial effects on human health through a range of pathways (McMichael and Haines 1997, McMichael et al. 1996) (Table 1). The potential effects of climate change on health are sometimes divided into direct and indirect to separate those impacts where the chain of causation is short, such as increased deaths during heatwaves, and those that are mediated through a longer causal chain. The latter include changes in ecosystems which can, for example, effect the distribution of insect vectors of disease. Most of the anticipated effects are likely to be adverse, although some, such as possible reductions in excess winter death rates due to warmer winters in cool-temperate countries, could be beneficial. Quantification of potential impacts is complicated by the many uncertainties involved and any estimates should be taken as indicative.

2.2 Approaches to Assessing Potential Impacts

There are a number of approaches to assessing the potential impacts of climate change on health. These

include the study of historical analogues that simulate certain aspects of future climate change. One example is the study of the effects of the El Niño/Southern Oscillation (ENSO), a large irregularly occurring atmosphere–ocean system which results in relatively short-term climate changes over the Pacific region every 2–7 years. The warm event (El Niño) is followed frequently by a cold event (La Niña). The ENSO is also linked by distant connections (teleconnections) to climatic anomalies elsewhere in the world, particularly in countries bordering the Pacific and Indian oceans. Integrated mathematical modeling is also being used increasingly to estimate the future impact on health of climate change. In order to undertake such modeling each component of the sequence of climate, environmental and social change, is represented mathematically.

2.3 Direct Effects of Heat and Cold

The link between heatwaves and increased death rates has been described in many parts of the world. Excess mortality especially is experienced by the elderly, in particular by those who live in disadvantaged areas without adequate air conditioning. Some of the increase in deaths is due to mortality displacement, i.e., short-term shift in the time of death of those who would have died anyway in the near future. The threshold at which increased death rates occur depends on population acclimatization and is, therefore, higher in those cities where the populations are used to high temperatures. The impact of the first heatwave on mortality in a given summer is often greater than the impact of subsequent heatwaves, probably because a disproportionate number of susceptible people die during the first heatwave. Several studies have quantified the impact of climate change on heat-related mortality. For example, one study (Kalkstein and Greene 1997) estimated an annual excess mortality attributable to climate change (assuming acclimatization) of between 500 and 1,000 for New York and 100 and 250 for Detroit by the year 2050.

There is controversy over the degree to which increases in summer mortality will be outweighed by decreases in winter mortality. Although death rates are higher in winter than in summer in the temperate countries, the relationship may not be directly due to low temperatures and increased viral infections may be partly responsible. Some countries with very cold winters, for example, Russia, seem to have low excess winter mortality, probably because of the effective adaptation of the population to cold winters by the use of warm winter clothing and adequate indoor heating (Donaldson et al. 1998). The UK has a particularly high winter excess mortality and this may be due, at least in part, to fuel poverty. Within Europe larger increases in winter mortality may occur with decreasing temperature in warmer locations, e.g., Athens, than in colder locations (Eurowinter Group 1997) perhaps because populations in countries with generally mild winters fail to wear suitable clothing or their housing is not adapted to low temperatures.

2.4 Studies of the Effects of El Niño/Southern Oscillation

ENSO can affect rainfall, leading to either droughts or floods in parts of the world, as well as causing increases in temperature and changes in the frequency, intensity, and geographical distribution of extreme weather events such as storms.

The ENSO cycle has been associated with substantial changes in the incidence of malaria in countries such as Pakistan, Sri Lanka, Colombia, and Venezuela (reviewed by Kovats et al. 1999). The incidence of dengue fever (a viral disease carried by mosquitoes) is affected by the ENSO cycle in some Pacific Islands (Hales et al. 1996). There are large increases in the numbers of people affected by natural disasters at a global level in El Niño years, and the year following (Bouma et al. 1997). Other health impacts of El Niño include increases in respiratory disorders due to very high levels of air pollution as a result of forest fires that occurred, for example, both in Indonesia and Brazil in association with the 1997/98 event. The ENSO phenomenon is clearly not strictly an analogue for global climate change but does demonstrate that some diseases and health outcomes are sensitive to changes in climate. Recently, there have been suggestions that the frequency of El Niño may increase in the future as a result of climate change (Timmermann et al. 1999). Increasingly, forecasting is being used to give early warning of El Niño in order to improve preparedness and reduce the adverse effects.

2.5 Mathematical Modeling of Malaria

Mathematical modeling has been applied to the assessment of likely changes in the geographical range of vector-borne diseases such as malaria. One estimate, for example, suggests that approximately 45 percent of the world's population live in zones of potential malaria transmission as defined by current climatic circumstances, and this would increase to around 60 percent towards the end of the next century assuming other relevant factors remain constant (Martens et al. 1995). Highly aggregated models such as the one used in this example are, of necessity, unable to take into account complexity of future changes. Nevertheless, they give a broad indication of the potential magnitude and direction of change and are continually being refined. They suggest that changes in distribution of malaria are likely to occur particularly at the edges of the current distribution, including, for example, mountainous regions in the tropics and subtropics. Estimation of numbers of excess cases and deaths in

the twenty-first century as a result of climate change is hampered by our lack of knowledge, for example, about the potential advances in the development of an effective vaccine for malaria, the distribution of impregnated bed nets to reduce transmission, and the trends in the development of resistance of the parasites to drugs used in treatment.

A number of empirical studies in Zimbabwe, Rwanda, and Ethiopia have examined how climate variability influences the distribution of malaria. They have indicated that highland malaria can respond to climatic variability, but whether changes in the altitudinal range of malaria which have apparently been observed in a number of sites are due to global climate change, is currently a matter of scientific debate. Only long-term monitoring of climate, vector populations, and the incidence of malaria, as well as potential confounding factors, such as changes in vector control programs and forest cover can finally resolve the controversy.

2.6 Other Vector-borne Diseases

Other vector-borne diseases which may be affected include those carried by ticks such as tick-borne encephalitis (inflammation of the brain) and Lyme disease. The former occurs widely in Central and Eastern Europe and in Scandinavia and the latter occurs in both Europe and the North Eastern US. Other factors which may effect tick-borne diseases include the pattern of forest cover that can influence the distribution of animal hosts on which the ticks can feed and changing patterns of leisure activities which may influence exposure to bites by infected ticks. There are several clinical types of Leishmaniasis, which are transmitted by sandflies in Asia, the Americas, Southern Europe, and Africa. Sandflies are sensitive to changes in temperature and, for example, it was estimated that a 3 °C increase in temperature could increase both the geographic and seasonal distribution of one important species in Southwest Asia (Cross and Hyams 1996). The tsetse fly that transmits sleeping sickness (human African trypanosomiasis) is also climate sensitive. In Latin America, the distribution of Chagas' disease, which is transmitted by the triatomine bug and causes long-term damage to the heart and to the muscle of the gastrointestinal tract, could be affected.

2.7 Extreme Events, Malnutrition, and Sea Level Rise

On average, every year around 120,000 people were killed by natural disasters between 1972 and 1996, with around 60 percent of the deaths occurring in Africa. Over the same period on average nearly 140 million people were affected by such disasters annually, most of these were living in Asia (International Federation of Red Cross and Red Crescent Societies 1998). Drought, famine, and flood are the main categories of disaster responsible for the majority of people affected.

Floods may cause a range of impacts on health including deaths and injuries from trauma or drowning, increased incidence of diarrheal disease and sometimes leptospirosis caused by exposure to infected rats' urine in floodwaters. Malnutrition may increase following flooding in some countries where food security is a problem. The impacts on mental health may be substantial and in some cases long-lasting. An increase in suicides was reported from Poland following floods in 1997 and an increase in behavioral disorders amongst children has been reported. Some parts of the world may experience increased rainfall due to climate change that could lead to larger floods (IPCC 1998).

Climate change could exacerbate periodic and long-term shortages of water especially in the arid and semi-arid parts of the world (IPCC 1998). Droughts tend to effect health particularly by causing a reduction in the availability of food. There may also be an increase in diarrheal diseases because water is short and there may not be sufficient for hygienic purposes. Severe drought may not invariably result in famine or major food shortages. For example, a severe drought in southern Africa in 1992 resulted in crop failure rate approaching 80 percent in some of the most affected areas but famine was averted because of regional cooperation and external assistance which provided grain shipments (Noji 1997). Unfortunately, international assistance to support humanitarian relief has fallen overall, for example, it declined 17 percent in real terms between 1992 and 1996, whereas emergency aid has tended to increase. After remaining steady at under half the United Nations target of 0.7 percent of gross national product (GNP) for more than 20 years, aid as a share of donor's wealth fell to 0.25 percent in 1996, its lowest level ever (International Federation of Red Cross and Red Crescent Societies 1998). If this trend continues, populations in the twenty-first century may be not only more vulnerable to climatic disaster but less likely to receive effective assistance.

There have been many studies to assess potential changes in food production globally and regionally under conditions of climate change. In general, it appears likely that agricultural yields may increase in the twenty-first century in middle to high latitudes depending on crop type, growing season, and changes in temperature and seasonality of precipitation. However, there are concerns that yields may decrease in parts of the tropics and subtropics particularly where dryland, nonirrigated agriculture predominates (IPCC 1998). This could lead to increased hunger, particularly in Africa.

Sea level rise caused by climate change may result in displacement of some populations particularly those living on deltas and low lying islands as well as leading

to salination of fresh water and increased vulnerability to extreme events.

2.8 Air Pollution

The weather has a substantial influence on the ambient concentrations of air pollutants, for example, high-pressure systems often create a temperature inversion which traps pollutants in the boundary layer near the earth's surface. Because of increases in anticyclonic conditions in summer in some parts of the world, climate change may increase concentrations of some pollutants. Ozone formation and destruction occurs by means of a complex series of photochemical processes and concentrations in the troposphere may be higher under climate change depending on the emission of precursors. Any increase in forest fires could have substantial effects on human health because of the formation of 'haze' with high concentrations of fine particulates (see earlier discussion on El Niño). Concentrations of aeroallergens (pollen, etc.) could be affected by both temperature and precipitation but other factors are also involved such as changes in land use and farming practices (Emberlin 1994).

3. Adaptation and Vulnerability

The reductions in greenhouse gas emissions resulting from the Kyoto protocol of the UN Framework Convention on Climate Change are likely to have little effect on the projected rises in temperature within the first half of the twenty-first century (Parry et al. 1998). Thus, reducing vulnerability to climate change is an important goal for public health in the twenty-first century. There are a number of factors which influence vulnerability, notably poverty with its associated lack of resources and infrastructure. Although historically the majority of greenhouse gas emissions have come from the industrialized nations, vulnerability to climate change is probably greater in developing countries.

Adaptation may be autonomous, indicating a natural or spontaneous response by individuals, or purposeful, typically by governments or other institutions in response to projected climate change. The latter might include strengthening existing disease surveillance systems for potentially climate sensitive diseases, improving vector control programs, and enhancing disaster preparedness plans.

4. Conclusions

Whilst adaptation to climate change is necessary this does not negate the importance of strategies to mitigate climate change, particularly by reducing fossil fuel use (Haines and McMichael 1997). This could have near term benefits by reducing deaths and other adverse effects on health of air pollution (Working Group on Public Health and Fossil Fuel Combustion 1997). The provision of 'clean energy' is an important contribution to improving health, particularly in developing countries (Haines and Kammen 2000).

See also: Desertification; Globalization and Health; Globalization: Geographical Aspects; Health Policy

Bibliography

Bouma M J, Kovats R S, Goubet S A, Cox J S, Haines A 1997 Global assessment of El Niño's disaster burden. *Lancet* **350**: 1435–38

Emberlin J 1994 The effects of patterns in climate and pollen abundance on allergy: 1994. *Allergy* **49**: 15–20

Cross E R, Hyams K C 1996 The potential effect of global warming on the geographic and seasonal distribution of *Phlebotamus papatasi* in Southwest Asia. *Environmental Health Perspectives* **104**: 724–27

Donaldson G C, Tchernjavskii V E, Ermakov S P, Bucher K, Keatinge W R 1998 Winter mortality and cold stress in Yekaterinberg, Russia: Interview survey. *British Medical Journal* **316**: 514–18

The Eurowinter Group 1997 Cold exposure and winter mortality from Ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. *Lancet* **349**: 1341–46

Haines A, McMichael A J 1997 Climate change and health: implications for research, monitoring, and policy. *British Medical Journal* **315**: 870–4

Haines A, Kammen D 2000 Sustainable energy and health. *Global Change and Human Health* **1**: 78–87

Hales S, Weinstein P, Woodward A 1996 Dengue fever epidemics in the South Pacific driven by El Nino southern oscillation? *Lancet* **348**: 1664–5

International Federation of Red Cross and Red CrescentSocieties 1998 *World Disaster Report 1998*. Oxford University Press, New York

Intergovernmental Panel on Climate Change (WGI) Houghton J T, Meira Filho L G, Callander B A, Harris N, Kattenberg A, Maskell K (eds.) 1996 *Climate change, 1995—the science of climate change: Contribution of Working Group 1 to the second assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, New York

Intergovernmental Panel on Climate Change, Watson R T, Zinyowera M C, Moss R H, Dokken D J (eds.) 1998 *The regional impacts of climate change, an assessment of vulnerability*. Cambridge University Press, New York

Intergovernmental Panel on Climate Change Working Group 1 2001 *Third assessment report*. Intergovernmental Panel on Climate Change, www.ipcc.ch

Kalkstein L S, Greene J S 1997 An evaluation of climate/mortality relationships in large US cities and the possible impacts of climate change. *Environmental Health Perspectives* **105**: 84–93

Kovats R S, Bouma M J, Haines A 1999 El Niño and Health, WHO Task Force on Climate and Health, WHO/SDE/PHE/99.4, Geneva

Martens W J M, Jetten T H, Rotmans J, Niessen L W 1995 Climate-change and vector-borne diseases: A global modeling perspective. *Global Environment Change* **5**: 195–209
McMichael A J, Haines A 1997 Global climate change: The potential effects on health. *British Medical Journal* **315**: 805–9
McMichael A J, Haines A, Sloof R, Kovats R S 1996 Climate Change and Human Health WHO, EHG 96/7, Geneva
McMichael A J, Powles J W 1999 Human numbers, environment, sustainability and health. *British Medical Journal* **319**: 977–80
Noji E K (ed.) 1997 *The Public Health Consequences of Disaster*. Oxford University Press, New York
Parry M, Arnell N, Hulme M, Nicholls R, Livermore M 1998 Adapting to the inevitable. *Nature* **395**: 741
Timmermann A, Oberhuber J, Bacher A, Esch M, Latif M, Roeckner E 1999 Increased El Nino frequency in a climate model forced by future greenhouse warming. *Nature* **3989**: 694–7
Working Group on Public Health and Fossil-Fuel Combustion 1997 Short-term improvements in public health from global-climate policies on fossil-fuel combustion: An interim report. Lancet **350**: 1341–8

A. Haines

Climate Change, Economics of

There are two sides to the economics of climate change. The first recognizes the economic costs and potential benefits that can be attributed to the physical and natural impacts of a changing climate—warmer temperatures, changes in precipitation patterns, rising sea level, and so on. These economic impacts include the cost of adapting to change in addition to the economic consequences that remain after such adaptation is effected. They also include the benefits that climate change might bring that would otherwise not have been forthcoming. The second side recognizes the economic costs that would be attributed to policies designed to mitigate climate change. These costs, too, include the cost of adapting to policy in addition to the residual consequences that persist in the wake of this adaptation.

Estimates of the economic impacts on both sides are highly uncertain, given our inability to understand fully the science of climate change and to look many decades into the future with any clarity. Estimates of both are also evolving continuously over time, so any estimate must be read with a sense of what was known at the time that it was published. The coverage offered here will provide some insight into this evolution even as it reports on the latest results that were available at the turn of the twentieth century.

1. The Economic Cost of Climate Change Impacts

Costs associated with the impacts of climate change are generally judged in terms of economic damage that would be avoided if the change did not occur; benefits are similarly estimated in terms of fortuitous impact that otherwise would not have happened. The Intergovernmental Panel on Climate Change (IPCC) reported preliminary estimates of the annual economic impact of a doubling of concentrations of greenhouse gases ($\sim$ 550 ppmv) and an associated 2.5 °C increase in global mean temperature (Houghton et al. 1996). The estimates reported by the IPCC for the United States, for example, ranged from a low of $55.5 billion (1990$) offered by Nordhaus (1991) to a high of $139.2 billion (1990$) authored by Titus (1992). The low and high estimates were calculated to be 1.0 percent and 2.5 percent of anticipated gross domestic product in 2065, respectively. The year 2065 was chosen as a benchmark because that was when the specified doubling of concentrations was anticipated to occur.

All of the estimates reported in 1995 by the IPCC were dominated by declines in agricultural production. Agriculture is, of course, a sector whose current practices would likely be threatened by higher temperatures and less precipitation. Most of the estimates for agriculture and other sectors were, however, the result of vulnerability studies that paid little attention to the ability of humans and their institutions to reduce economic damage and expand economic opportunity by adapting—that is, by changing practices so that they became less vulnerable to the new climate or so that they could take greater advantage of its appearance. Moreover, most of these early studies relied on relatively primitive methods of tracking the different regional consequences of a 2.5 °C warming. Global climate modelers have long expected that some regions would see temperatures increase by more than 2.5 °C while others might actually get cooler. Some areas would get wetter while others get drier. Seas would rise in some places and actually fall more slowly elsewhere (where the coastline is actually rising at present). Finally, few of the early studies were able to consider the effects of changes in humidity, frequency of extreme temperature events, or any of the other more subtle physical ramifications of global warming.

More recent cost estimates have begun to overcome these shortcomings. Table 1 presents regional cost estimates of market impacts published by Mendelsohn et al. (2000) for a 2 °C warming and a 50 cm increase in sea level. Notice that the overall annual effect on world economic activity, a 0.3 percent reduction, is much smaller than in the earlier studies. Effects on agriculture still dominate, but the regional distribution of impacts is striking. Some regions, notably North America and Europe, are now seen to benefit from warming whereas others such as Africa are severely harmed. These results were based on a statistical approach that looked carefully at how various regions cope with their current climates to see how other regions might respond if their climates changed.

Table 2 gives estimates from Tol (1998) for a 1 °C warming in even geographical greater detail. Only

Table 1
Regional market impacts for a 2 °C warming (billions 1990$)

Region	Agriculture	Forest	Coast	Energy	Tourism	Total	%GDP
Africa	−11	−6	0	−2	−3	−22	−0.8
Asia	−14	−1	−5	−9	−8	−37	−0.1
Latin America	−7	−5	0	−4	−5	−22	−0.4
Europe	34	16	−4	4	31	82	0.2
North America	21	14	0	0	13	47	0.2
Oceana	−3	−1	0	−1	−1	−7	−0.6
OECD	26	16	−7	2	45	82	0.15
Non-OECD	6	−1	−3	14	−18	−40	−0.1

Source: Mendelsohn et al. (1999, Table 1)

Table 2
Annual impact for a 1 °C warming (billions 1990$)

Region	Estimate	Standard deviation	Percent of GDP	Standard deviation
American OECD	175	107	3.4	2.1
European OECD	203	118	3.7	2.2
Pacific OECD	32	35	1.0	1.1
Central Europe & former Soviet Union	57	108	2.0	3.8
Middle East	4	8	1.1	2.2
Latin America	−1	5	−0.1	0.6
Southern and southeast Asia	−14	9	−1.7	1.1
Centrally Planned Asia	9	22	2.1	5.0
Africa	17	9	4.1	2.2

Source: Tol (1998, Table 7)

Latin America and southern and southeast Asia suffer losses in his work, and many regions (including Africa) benefit substantially. Modest warming might, it would seem, be a good thing. Indeed, Mendelsohn et al.'s work also supports this suggestion. Tol's results do, however, offer a warning about leaping to that conclusion too quickly. Two columns in Table 2 report standard deviations for his estimates. The standard deviation is a measure of uncertainty that indicates roughly a 66 percent likelihood that the true impact of a modest 1 °C warming would lie within a range of plus or minus the standard deviation from the recorded figure. For example, then, Tol suggests roughly a 66 percent likelihood that the annual impact of a 1 °C warming on the Pacific members of the OECD would lie between a 0.1 percent loss (= 1.0–1.1) in gross domestic product (GDP) and a 2.1 percent gain in GDP. Given current understanding of climate and economic systems, therefore, there is roughly a 17 percent chance that GDP would climb by more 2.1 percent and a 17 percent chance that GDP would fall by more than 0.1 percent of GDP. There is also a 17 percent chance that the economic damage suffered by southern and southeast Asia in the wake of a 1 °C warming would be larger than 2.8 percent of GDP. The size of the uncertainty within which impact estimates must be contemplated is still enormous, and the distributional ramifications of this uncertainty can be quite unsettling.

The role of adaptation and learning how to incorporate physical impacts other than temperature is clearly demonstrated in Table 3. Regional estimates offered by four other scholars plus Tol are depicted there for a 2.5 °C temperature increase. Some include adaptation—switching crops, adjusting planting dates, adding or eliminating irrigation, adjusting fertilizer practices, and so on; others do not. Some include the fertilizing effect on plant productivity of higher carbon dioxide concentrations in the atmosphere, while others do not. Notice that carbon dioxide fertilization can turn damages into benefits and that adaptation can reduce damage and increase benefits. Indeed, all of Tol's estimates with adaptation represent gains. Be warned, though, that he still reports enormous ranges of uncertainty surrounding them.

Finally, all of these results envision smooth if not predictable climate change. The real concern on the impacts side could, however, be the potentially exaggerated effects of sudden, surprising and perhaps irreversible consequences of warming. Economic systems never cope well with sudden changes in their environment, even if the changes are ultimately beneficial. As a result, the estimates quoted above could be dwarfed if the physical impacts of climate change are

Table 3
Economic impacts on agriculture of a 2.5°C warming (percent of gross agricultural product)

	Kane et al.	Tsigas et al.		Darwin et al.	Reilly et al.			Tol	
Fertilization?	No	No	Yes	No	No	Yes	Yes	Yes	Yes
Adaptation?	No	No	No	Yes	No	No	Yes	No	Yes
American OECD	0.03	−0.31	0.05	0.10	0.03	0.00	0.00	−0.25	1.30
European OECD	−0.52	−0.73	0.14	−0.41	−0.34	−0.06	−0.02	0.55	2.09
Pacific OECD	−2.08	−1.38	−0.06	0.31	−0.31	−0.04	−0.01	−0.15	0.80
Central Europe/former Soviet Union	−0.02	−1.48	−0.07	0.14	−0.18	−0.25	−0.18	0.94	2.65
Middle East	−0.01	−1.48	−0.07	0.14	−0.18	−0.25	−0.18	−0.44	0.58
Latin America	0.05	−2.18	−0.47	0.10	−0.22	−0.15	−0.16	−0.76	0.55
Southern and southeast Asia	−0.08	−2.26	−0.32	−0.04	−0.91	−0.17	−0.13	−0.66	0.63
Centrally Planned Asia	3.84	−3.97	0.28	0.11	−10.09	0.04	0.53	1.73	3.10
Africa	−0.01	−1.48	−0.07	0.14	−1.18	−0.25	−0.18	−0.23	0.47

Source: Tol (1998, Tables 1 and 2). He references Kane et al. (1992), Tsigas et al. (1996), Darwin et al. (1996), and Reilly et al. (1996)

not smooth. Yohe and Schlesinger (1998) computed the economic cost of sea level rise on the developed coastline of the United States with and without sufficient foresight for markets to response to the threat of inundation. The difference between the two, one estimate under the best of circumstances of the extra cost of surprise, was as large as 100 percent even for sea-level rise trajectories in the middle of its own range of uncertainty.

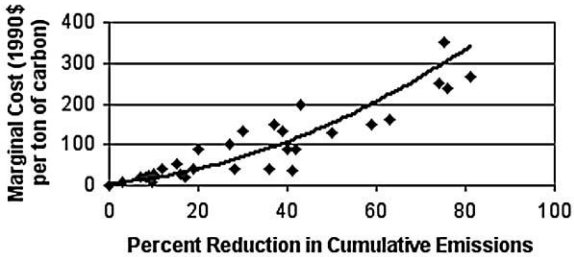

Figure 1
The marginal cost of reducing emissions of carbon dioxide

2. *The Economic Cost of Mitigating Climate Change*

The scale and pace of climate change can be influenced by policy interventions that slow the emission of greenhouse gases. Many researchers have investigated the cost of this sort of climate change mitigation. They have, in particular, focused their attention on energy consumption and the resulting emission of carbon dioxide. Carbon dioxide is a product of burning fossil fuel, and its emission varies from fuel to fuel. Burning coal, for example, emits 25 percent more carbon per unit energy than burning oil, and burning oil emits 43 percent more than natural gas. Burning hydrogen emits no carbon. Hydroelectric, wind, solar, and nuclear power are similarly carbon-free sources of energy. Mitigation simply involves substituting carbon-free sources of energy for carbon-based fuels *and* low-carbon fossil fuels such as natural gas for high-carbon fuels such as coal. Our ability to effect and to sustain this sort of substitution over the very long run depends upon the availability of new technology and the supply of low-carbon and carbon-free sources of energy.

The most effective means of conveying the cost of mitigation is to track the economic impact of reducing cumulative global emissions through the year 2100 from 'baseline' levels that would have been anticipated in the absence of any policy intervention. Fig. 1 displays reductions in cumulative emissions from various baselines through the year 2100 against the estimated tax (marginal cost) that would have to be imposed per ton of carbon to achieve those reductions. The points portrayed there indicate selected estimates published by various researchers through the middle of 1998, and the curve summarizes these cost data as a function of percentage emissions reduction. Fig. 1 shows clearly that the marginal cost of emissions reduction increases at an increasing rate even though the taxes estimated for any particular reduction in emissions are disperse. This dispersion is a reflection of the uncertainty with which these costs can be computed—uncertainty caused by assumptions about technology, supplies, and the intensity with which future economic activity would employ energy along the baseline.

2.1 *Interventions that Limit Atmospheric Concentrations*

The Framework Convention on Climate Change (FCCC) committed the globe in 1992 to holding

concentrations of greenhouse gases below levels that would prevent 'dangerous anthropogenic interference with the climate system.' The precise concentration target that corresponds with this imperative has not yet been identified, so many researchers have investigated the economic cost of reducing emissions over the next 100 years so that concentrations do not exceed a range of thresholds. Manne and Richels (1997) estimate that the cost of achieving the most popular threshold, 550 ppmv, could be as high as $3.5 trillion (1990$) or as low as $650 billion (1990$). Estimates from other researchers were comparable; but they differed from one another for the same reasons as mentioned above. Costs for lower thresholds such as 450 ppmv are much higher. However, some possible low-emissions baselines achieve stable concentrations around 750 ppmv without any intervention of any kind.

Cost estimates for meeting concentration thresholds depend critically on the timing and location of each unit of emissions reduction. Wigley et al. (1996) have called the 'where' and 'when' flexibility components of cost. Their 'WRE' results emphasize that costs would be minimized if each ton of emissions reduction were taken from the least costly source regardless of where it is located. Their results also required that emissions should be reduced at any point in time only if the present value of the associated cost is in line with all other reductions at all other times. Compared with emissions reductions suggested by the IPCC (see Houghton et al. 1992), the combined effect of exploiting both types of flexibility allowed the 'WRE' path to cut the cost of achieving the same 550 ppmv threshold from the same middle emissions baseline by more than 80 percent.

The 'WRE' results are controversial because 'when' flexibility implies that early reductions in emissions would be smaller than they would under the IPCC proposal. As a result, the near-term pace of climate change would be larger. Meanwhile, 'where' flexibility has served as an anchor for a wide range of proposals that would allow countries to trade permits to emit greenhouse gases. The idea here is that an emerging market for permits would work to ensure that least cost sources of reductions were always exploited.

It is difficult to contemplate near-term mitigation in the absence of any knowledge about the appropriate concentration target and without any real understanding about whether future baseline emissions will be relatively high or low. Yohe and Wallace (1996) looked as this problem as one of hedging. Policy directed at a low threshold along a high-emissions path would be far more vigorous in the near-term than a policy directed at a high threshold along a low-emissions path. Either would be in error, though, if either the presumed target or the presumed emissions path turned out to be incorrect. As a result, adopting either would impose extra cost on the global economy. Yohe and Wallace reported that least-cost hedging would support focusing on a middle concentration target such as 550 ppm and assuming that emissions would otherwise track slightly higher than commonly accepted 'best-guess' baseline. This sort of hedging would increase the costs of meeting a concentration target, but only modestly under the assumption of maximum geographical and intertemporal flexibility.

2.2 *The Kyoto Emissions Reduction Protocol*

The Third Conference of the Parties of the FCCC agreed in 1997 through the Kyoto Protocol to impose a set of greenhouse gas emissions targets for the world's developed countries (the so-called Annex I or Annex A countries). The targets were different for different countries, but their combined effect would reduce total emissions from Annex I countries by nearly 6 percent relative to their 1990 levels and almost 20 percent from their 1999 levels by 2012. Non-Annex I countries were exempted by the Protocol from any emissions reduction.

Subsequent research has raised a large number of issues in regard to achieving any eventual FCCC concentration limit. Some are technical and deal with accounting procedures for counting emissions reductions. Others are more fundamental. First among these is the observation that full compliance by Annex I by 2012 with fixed total Annex I emissions thereafter will not stabilize concentrations at any level for most baselines. It follows that non-Annex I countries will eventually have to accept limits on their emissions, as well, if the FCCC objective of stable concentrations is to be achieved. Indeed, a 550 ppmv threshold would not be achieved along many baselines even if Annex I eliminated all dependence on fossil fuel by the middle of the twenty-first century.

Second, Annex I compliance with their Kyoto targets by 2012 does not conform with the cost-minimizing pattern of maximal intertemporal flexibility for most baselines and most concentration targets. Passing through the Kyoto benchmark increases the cost of meeting any threshold above 500 ppmv along all but the most energy intensive baselines.

Finally, negotiations about how to arrange for geographic flexibility within the implementation of the Kyoto Protocol are critical. The equity and cost implications of allowing flexibility with Annex I and/or the ability for Annex I countries to be credited for reductions that they underwrite in non-Annex I countries are enormous. McKibben and Wilcoxen (1999) have argued that changes in the so-called 'terms of trade' caused by massive transfers of wealth to non-Annex I countries in exchange for emissions reduction credits could actually lower their economic welfare. In addition, Manne (1999) observed that the partial global coverage of the Protocol could lead to significant 'leakage' so that global emissions might not fall as far as expected. Why? Because restricted emissions

in Annex I would cause the prices of fossil fuels to fall and thereby increase emissions across the developing world.

2.3 Optimal Emissions Reductions

Nordhaus (1991) was the first researcher to weigh the present value of the benefits of mitigation policy against the present value of their costs to compute an economically optimal policy trajectory over the long term. His results were based on the anticipation that impacts would be smooth and amount to roughly 1 percent of gross world product if the global mean temperature rose by 2.5 °C. Corroborated by his own subsequent work and by others, they support modest early intervention followed by smooth and gradual tightening of emissions restrictions. Indeed, most optimality exercises propose carbon taxes between $10 and $20 around 2000; and most see those taxes increasing over time at $\sim$ 3 percent per year. None of the results achieve stable atmospheric concentrations along most baselines, and none come close to the restrictions imposed on 2012 emissions by the Kyoto Protocol.

3. Synthesis as the Future Unfolds

Synthesizing the economics of climate change and climate policy is an evolving process that will continue well into the twenty-first century. As researchers learn more about both, the costs associated with both could easily fall. However, they may not, particularly on the climate side of the calculus. The possibility of sudden and, as yet, unanticipated impacts could change the picture dramatically.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Climate Change and Health; Climate Impacts; Climate Policy: International; Desertification; Ecological Economics; Economic Geography; Environmental Adaptation and Adjustments; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Planning; Environmental Policy; Environmental Vulnerability; Food Security; Globalization: Geographical Aspects

Bibliography

Darwin R F, Tsigas M, Lewandrowski J, Raneses A 1996 *World Agriculture and Climate Change*. US Department of Agriculture Report 703. US Department of Agriculture, Washington, DC

Houghton J T, Callander B A, Varney S K (eds.) 1992 *Climate Change 1992. The Supplementary Report to the IPCC Scientific Assessment*. Cambridge University Press, Cambridge, UK

Houghton J T, Meira Filho L G, Callander B A, Harris N, Kattenberg A, Maskell K (eds.) 1996 *Climate Change 1995: The Science of Climate Change*. Cambridge University Press, Cambridge, UK

Kane S, Reilly J M, Tobey J 1992 An empirical study of the economic effects of climate change on world agriculture. *Climatic Change* **21**: 17–35

Manne A 1999 International carbon agreements, trade and leakage. In: *Proceedings of the IEA/EMF/IIASA Energy Modeling Meeting*. Energy Modeling Forum, Stanford, CA

Manne A, Richels R 1997 On stabilizing CO_2 concentrations—cost-effective emission reduction strategies. In: Cameron O K, Fukuwator K, Morita T (eds.) *Proceedings of the IPCC Asia–Pacific Workshop on integrated Assessment Models*. Center for Global Environmental Research, Ibaraki, Japan, pp. 439–59

McKibben W, Wilcoxen P 1999 The theoretical and empirical structure of the G-Cubed model. *Economic Modelling* **16**: 123–48

Mendelsohn R, Morrison W, Schlesinger M E, Andronova N G 2000 Country-specific market impacts of climate change. *Climatic Change* **45**: 553–69

Nordhaus W D 1991 To slow or not to slow: the economics of the greenhouse effect. *Economic Journal* **101**: 920–37

Nordhaus W D 1994 *Managing the Global Commons: The Economics of Climate Change*. MIT Press, Cambridge, MA

Reilly J M, Baethgen W, Chege F E, van de Geijn S C, Lin E, Iglesias A, Kenny G, Patterson D, Rogasik J, Roetter R, Rosenzweig C, Sombroek W, Westbrook J 1996 Agriculture in a changing climate: impacts and adaptation. In: Watson R T, Zinyowera M C, Moss R H (eds.) *Climate Change 1995: Impacts, Adaptations and Mitigation of Climate Change: Scientific–Technical Analysis—Contributions of Working Group II to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK, pp. 427–68

Titus J G 1992 The cost of climate change to the United States. In: Majumdar S K, Kalkstein L S, Yarnal B, Miller E W, Rosenfeld L M (eds.) *Global Climate Change: Implications, Challenges and Mitigation Measures*. Pennsylvania Academy of Science, Easton, PA, pp. 217–35

Tol R S J 1998 *New Estimates of the Damage Costs of Climate Change. Part I: Benchmark Estimates*. Institute for Environmental Studies, Vrije Universiteit, Amsterdam

Tsigas M E, Frisvold G B, Kuhn B 1996 Global climate change in agriculture. In: Hertel T W (ed.) *Global Trade Analysis: Modeling and Applications*. Cambridge University Press, Cambridge, UK, pp. 3–15

Wigley T, Richels R, Edmonds J A 1996 Economic and environmental choices in the stabilization of atmospheric CO_2 concentrations. *Nature* **379**: 240–43

Yohe G, Schlesinger M E 1998 Sea-level change: the expected economic cost of protection and abandonment in the United States. *Climatic Change* **38**: 447–72

Yohe G, Wallace R 1996 Near-term mitigation policy for global change under uncertainty: minimizing the expected cost of meeting unknown concentration thresholds. *Environmental Modeling and Assessment* **2**: 47–57

G. W. Yohe

Copyright © 2001 Elsevier Science Ltd. 1999
All rights reserved.

 ISBN: 0-08-043076-7

Climate, History of

1. Introduction

Climatologists and climate historians have assembled robust evidence that the world's climate has changed significantly over the past millennium. However, for most historians climate still is an unacknowledged constant. There remain serious intellectual and practical obstacles to understanding and using the new evidence that is now becoming available (Richards 2001). The first section of this article reviews the discussion on the issue. The second section examines the evidence and the approaches used for reconstructing past weather and climate. The third section reviews the main trends of climate variability over the last millennium. In the last section the assessment of climate impacts on premodern societies is considered.

2. The Discussion on 'Climate and History'

Every society develops philosophical and mythical interpretations about the role of the natural environment in human affairs. Enlightenment thinkers concluded that cultures were determined or strongly shaped by climate. Economists and geographers (e.g. Stanley Jevons, Eduard Brückner) assumed that economic life was affected by climatic cycles. Environmental determinism was carried to an extreme by the geographer Ellsworth Huntington in the early twentieth century (Fleming 1998). Sociologist Emile Durkheim summarily rejected efforts to link human performance to climate changes. He postulated that social issues could be explained solely by social factors (Glaeser 1994). The discussion on the social significance of climatic variations was resumed by historians of the French Annales School (e.g., Emmanuel Le Roy Ladurie) after World War II. Ladurie suggested that historical climate should first be reconstructed for its own sake without considering its potential significance for human history. This issue should only be addressed in a second step based on reliable reconstructions of past climate. However, he was skeptical in this respect, postulating that 'the narrowness of the range of secular temperature variations and the autonomy of the human phenomena which coincide with them in time make it impossible for the present to claim that there is any casual link between them' (Le Roy Ladurie 1971, p. 293). The English meteorologist Hubert Lamb (1988), who took an active interest in history, became Le Roy Ladurie's most prominent opponent. Lamb was convinced that weather and climate had affected human affairs in the past and that humankind would do well to examine some of the lessons provided by nature.

From the early 1990s the framework of the discussion changed. On one hand the issue of the increased greenhouse effect was put on the agenda stimulating efforts towards reconstructing past climates. In this respect historical climatology cooperated with scientific disciplines. Remarkable progress was made from mostly isolated attempts at reconstructing local climate histories, to successful attempts at boiling down regional evidence into quasi-homogeneous highly correlated monthly time series of temperature and precipitation indices on the supraregional scale (Pfister et al. 1999). In the late 1990s historical climatology was moving at the center of the controversial detection debate about anthropogenic climate change, because documentary data is the only evidence for assessing the frequency and clustering of rare but socioeconomically significant disasters such as intense storms, severe floods, and droughts.

On the other hand the mainstream of historians turned away from structural history and the '*longue durée*' in favor of discourse analysis. As a consequence the incentive to investigate the impacts of the climatic variability being detected by climate historians declined. In the context of the ongoing debate the issue of the social perception of (reconstructed) climate change may become attractive for historians of ideas.

3. The Reconstruction of Weather and Climate from Natural and Manmade Archives

The global climate of the last millennium is reconstructed using evidence from both natural and manmade archives. Data from natural archives (e.g., tree ring or ice core data) is essential for those periods of history and those regions of the world for which documentary evidence is sparse or nonexistent, such as precolonial North America. However, most reconstructions from natural archives cannot be broken down into sufficiently short units of time (e.g., months or seasons) and into specific parameters (temperature and precipitation) (Bradley 1999) that would be needed for conclusive investigations into the human dimension of climatic change.

Data from manmade archives (i.e., documentary evidence) is well researched in Europe and East Asia. Investigations have hardly begun in Latin America. The evidence for Africa is spotty. In the Islamic world the possibly abundant evidence remains to be explored.

Documentary evidence is classified into descriptive and proxy data. Descriptive data includes chroniclers' narratives of weather patterns characteristic of a particular region. Chinese historians have drawn upon observations found in local gazetteers maintained by local gentry in nearly every district. In order to portray more objectively the character of extreme events, chroniclers referred to the duration of snow cover or the freezing of bodies of water, to the development of crops, and to high and low water levels. In Europe daily weather observations were promoted by the rise of planetary astronomy from the late fifteenth century.

Regular instrumental measurements of weather began in the late seventeenth century. From 1860 national meteorological networks came into being.

Evidence providing an indirect measure of climate is mostly drawn from administrative records. These may yield long, continuous, and quasi-homogeneous series of climate related data that reflect the beginning of agricultural activities (e.g., the vine harvest), agricultural production (e.g., yield of vineyards), or the time of freezing and opening up of waterways (Pfister et al. 1999).

Records of rogations (i.e., standardized religious ceremonies to put an end to a meteorological stress situation) are a promising source for the Spanish-speaking world. In Spain rogations were recorded in the account books of both the municipalities and the church (Martin and Barriendos 1995).

Manmade archives are interpreted by historical climatology, which serves as an interface between climatology and history. It is directed towards three objectives (Pfister et al. 1999):

(a) reconstructing weather and climate as well as natural disasters prior to the creation of meteorological networks;

(b) investigating the vulnerability of past societies to climatic extremes and natural disasters;

(c) exploring past discourses on and social representations of climate.

Usually the evidence available for a given month or season is converted to ordinal indices for temperature and precipitation. The computing of transfer functions with instrumental series allows temperature and precipitation to be assessed for the pre-instrumental period. Series of indices were also included in statistical models to reconstruct monthly mean air pressure at sea level for the eastern North Atlantic-European region (25°W to 30°E, and 35°N to 70°N) back to 1659 (Luterbacher et al. 2000).

4. *Climatic Trends and Anomalies Over the Last Millennium*

4.1 *Three Main Phases*

Palaeoclimatologists and climate historians describe three main phases of climatic change over the past millennium: A 'Medieval Warm Period' to 1300 (Hughes and Diaz 1994); a subsequent cool phase lasting to the late nineteenth century that is labeled 'Little Ice Age' because glaciers in most regions of the globe were expanding during that time (Bradley and Jones 1996). The twentieth century is the warmest period of the millennium, partly as a consequence of the increased greenhouse effect. However, such generalizations on the global level mask a broad array of contrasting regional and local trends. Moreover, in order to investigate human vulnerability to climatic stress, the perspective of 'ages' needs to be broken down to regional monthly or seasonal temperature and precipitation patterns. So far, this level of detail is only available for Europe and China.

4.2 *Central Europe*

After a cold phase in the twelfth century, winters were prevailingly warm from 1180 to 1300. From 1300 to 1900 the winter half-year was colder than today. This is related to more frequent and sustained advection of cold, dry continental air-masses from the (north) east. Severe winters were frequent from 1306 to 1328, 1430 to 1490, 1565 to 1615, 1655 to 1710, 1755 to 1860, and 1880 to 1895. From 1365 to 1400, 1520 to 1560, and from 1610 to 1650 moderate winters prevailed. Springs were extremely cold in the 1690s and in the 1740s.

Summers do not show distinct long-term characteristics. Those in the thirteenth century were prevailingly warm and dry. In the fourteenth century clusters of cold and wet summers occurred repeatedly (e.g., in the 1310s and 1340s). From 1380 to 1430 and again from 1530 to 1565 the summer half-year was as warm as today. Over the last third of the sixteenth century cold spells and long rains in midsummer expanded at the expense of warm anticyclonic weather. This tendency culminated in the 1590s (Pfister et al. 1999). Summers at the beginning and end of the seventeenth century were prevailingly cool while those from 1630 to 1687 were moderate. In the 1700s several warm decades (the 1720s, the 1730s, and the 1780s) stand out in England and in Central Europe, whereas the first half of the nineteenth century, particularly the 1810s, was markedly cooler (Bradley and Jones 1996).

4.3 *Russia*

Winters became more severe at the end of the sixteenth century, in particular from 1620 to 1680 and in the first half of the nineteenth century.

In the summer half-year droughts were frequent from 1201 to 1230, 1351 to 1380, and 1411 to 1440. A period of comparatively warm conditions in all seasons stands out during the first half of the sixteenth century. Subsequently, cold spells occurred more often from 1590 to 1620 and from 1690 to 1740 with a peak in the 1730s. Droughts occurred frequently from 1640 to 1659 and from 1680 to 1699. The six decades from 1770 to 1830 were warm, and droughts were frequent from 1801 to 1860. Summers from 1890 to 1920 were by far the coldest in the last 500 years. This included an unusually large number of extreme dry and wet seasons (Bradley and Jones 1996).

4.4 *China*

In South China the thirteenth century was the warmest of the last millennium. Three cold periods—1470 to

1520, 1620 to 1740, and 1840 to 1890—are identified, the 1650s being by far the coldest decade. Rainfall during the seventeenth century was extremely variable. Temperatures during the eighteenth century, unlike in Europe, rarely climbed to twentieth century levels, but precipitation conditions were more favorable. Climate variability increased markedly throughout the nineteenth century to a maximum in the early twentieth century. In North China two cold periods—1500 to 1690 and 1800 to 1860—stand out over the last six centuries. Considering all seasons, the period from 1650 to 1670 was the coldest, but the summer half-year was almost equally cold from 1580 to 1600 (Wang 1991).

4.5 *The Mediterranean*

After a cold twelfth century the period 1200 to 1400 was very warm in the southwest. Annual precipitation in Morocco was generally lower from the sixteenth to nineteenth centuries (Bradley and Jones 1996). In Catalonia (northeastern Spain) dry spells in the winter half-year were frequent in the mid-sixteenth century, but almost absent from 1580 to 1620. Numerous autumnal floods were reported from 1580 to 1630, from 1770 to 1800, and again from 1840 to 1870 (Martin and Barriendos 1995).

4.6 *Latin America*

In both Spanish and Portuguese America there seems to have been a trend to greater aridity in the 1700s compared to the 1600s. Dendroclimatic evidence for the Santiago de Chile area indicates higher than average rainfall from 1450 to 1600 whereas droughts became frequent over the subsequent centuries (e.g., 1637 to 1640, 1770 to 1773, 1790s, 1810s). In the Buenos Aires region (Argentina) the 1700s were drier than the previous century. Prolonged droughts are recorded in the 1690s, the 1710s, the 1750s, and 1771 to 1774 (Claxton 1993).

4.7 *ENSO*

The El-Niño Southern Oscillation (ENSO) is the result of a cyclic warming and cooling of the ocean surface in the central and eastern Pacific that strongly affects rainfall in the areas around the Pacific and the Indian Ocean. Archival data suggests that ENSO episodes from 1600 to 1900 had more intense and global effect than those of the twentieth century. For example, the worst droughts in the colonial history of India (mid 1590s, 1629 to 1633, 1685 to 1688, 1788 to 1793, 1877 to 1878) are related to ENSO connected failures of the monsoon. For the last two events the global dimension of these episodes is demonstrated (Grove and Chappell 2000).

5. *The Historical Significance of Climatic Change*

The issue of whether climatic change has had a significant impact on history is controversial. It should not be overlooked that both 'climate' and 'history' are blanket terms located on such a high level of abstraction that relationships between them cannot be investigated according to the rules of scientific methodology. In order to become more meaningful, the issue needs to be broken down to lower scales of analysis e.g., by focusing on specific human activities and/or needs in relation to a given set of climatic variables. Regarding preindustrial societies this concerns primarily the availability of biomass (e.g., food, fodder) and energy (e.g., wind, water-power) but also processes of population dynamics (e.g., patterns of disease and epizootics, as well as fertility of men and livestock), and transport and communications as well as military and naval operations. Undoubtedly, beneficial climatic effects tend to enlarge the scope of human action, whereas climatic shocks restrict it or even lead to emergency situations. Which climatic constellations matter for energy availability and population dynamics depend on the environmental, cultural, and historical context.

Models of climatic effects on society are often framed as a chain of causation. Climatic patterns have a first order or biophysical impact on agricultural production or on the outbreak of diseases or epizootics. These may have second order effects on prices of food or raw materials, which may then ramify into the wider economy and society (third order impacts). The farther we move away from first order impacts, the greater the complexity of the factors masking the climatic effect. It is also plain that it is easier to investigate the effects of short-term impacts. In dealing with the effects of multidecadal climate variations we have to account for modifications in the economic, institutional, and environmental setting so great as to vitiate any attempt at strict comparison or measurement (Kates et al. 1985). Most climatic impacts were related to food scarcity or famines.

Crises were triggered by a slump in overall agricultural production. This could be a consequence of climatic shocks or warfare. In Central Europe severe climate induced crises (e.g., 1569–74, 1627–29, 1692–94, 1769–72, 1816–17, 1853–55) were connected to a cumulation of unfavorable weather patterns, which made the traditional risk minimizing strategies ineffective (Richards 2001). Rainfall is the limiting factor in the subtropical and tropical zones; in higher latitudes it is summer warmth. Connections between climatic anomalies and diseases are complex. Some diseases (e.g., cholera) are climate related whereas others (e.g., bubonic plague) are not (Rotberg and Rabb 1983). The theory of pre-industrial trade cycles considers the harvest the critical determinant influencing urban income and rural employment levels. A sharp rise in food prices promoted widespread un-

employment, begging, and vagrancy that further propagated infectious diseases and increased crisis mortality (Post 1985).

Crises represented a major challenge for political and social systems. Rather than investigating changes in average values, historical climatology should focus on changes in the frequency and severity of extremes. The evidence is growing that exogenous shocks (including natural disasters) have a tendency to cluster rather than being randomly distributed along the time axis, as is often believed. This allows us to distinguish between periods of high and low climatic stress.

An example is provided by sixteenth-century Europe: there was a sudden increase in the number of cold anomalies after 1565. Over the subsequent decades climate became more significant for food prices than population levels and increases in the money supply. The case is even more obvious for wine production which as a consequence of an almost uninterrupted series of cold summers nearly collapsed from 1585 to 1600 across a large region ranging from Switzerland to Hungary. The slump of vine production had far-reaching consequences for major social groups depending on vine growing. Many peasant communities suffered such a large collective damage from the effects of continuous crop failures that they pressed the authorities to permit witch hunts. Thousands of witches were burnt as scapegoats of climatic change (Behringer 1999).

Based on the new reconstructions that are becoming available, the significance of climatic variability needs to be reassessed in many contexts of economic, social, and environmental history without including deterministic overtones.

See also: Climate Change and Health; Climate Change, Economics of; Climate Impacts; Desertification; Environmental Determinism; Irrigation Societies; Water Resources

Bibliography

Behninger W 1999 Climatic change and witch hunting: The impact of the little ice age an mentalities. In: Pfister C, Brazdil R, Glazer R (eds.) *Climatic Variability in Sixteenth-century Europe and its Social Dimension*. Kluwer, Dordrecht, The Netherlands

Bradley R S 1999 *Palaeoclimatology*. Academic Press, San Diego, CA

Bradley R S, Jones P D (eds.) 1996 *Climate Since AD 1500*. Routledge, London

Claxton R H 1993 The record of drought and its impact in colonial Spanish America. In: Herr R (ed.) *Themes in Rural History of the Western World*. Iowa State University Press, Ames, IA

Dupâquier J 1989 Demographic crises and subsistence crises in France, 1650–1725. In: Walter J, Schofield R (eds.) *Famine, Disease and the Social Order in Early Modern Society*. Cambridge University Press, Cambridge, UK

Fleming J R 1998 *Historical Perspectives on Climate Change*. Oxford University Press, Oxford, UK

Glaeser B 1994 Soziologie der Umwelt. In: Ernste H (ed.) *Pathways to Human Ecology*. Lang, Bern, Switzerland, pp. 115–32

Grove R, Chappell J (eds.) 2000 *El Niño, History and Crisis*. White Horse Press, Knapwell, UK

Hughes M K, Diaz H F 1994 *The Medieval Warm Period*. Kluwer, Dordrecht, The Netherlands

Kates R W, Ausubel J H, Berberian M (eds.) 1985 *Climate Impact Assessment, Studies of the Interaction of Climate and Society*. Wiley, Chichester, UK

Lamb H H 1988 *Weather, Climate & Human Affairs*. Routledge, London

Le Roy Ladurie E 1971 *Times of Feast, Times of Famine: A History of Climate Since the Year 1000*. Doubleday, Garden City, NY

Luterbacher J, Rickli R et al. 2000 Monthly mean pressure reconstruction for the late Maunder minimum period (AD 1675–1715). *Journal of Climatology* **20**: 1049–66

Martin V J, Barriendos V M 1995 The use of rogation ceremony records in climatic reconstruction. *Climatic Change* **30**: 201–21

Pfister C, Brázdil R, Glaser R (eds.) 1999 *Climatic Variability in Sixteenth Century Europe and its Social Dimension*. Kluwer, Dordrecht, The Netherlands

Post J D 1985 *Food Shortage, Climatic Variability, and Epidemic Disease in Preindustrial Europe*. Cornell University Press, Ithaca, NY

Richards J F 2001 *The Unending Frontier: Environmental History in the Early Modern World*. University of California Press, Berkeley, CA

Rotberg R I, Rabb T K (eds.) 1983 *Hunger and History*. Cambridge University Press, Cambridge, UK

Wang S W 1991 Reconstruction of temperature series of North China from 1380s to 1980s. *Science in China* B **34**: 751–9

Wigley T M L, Ingram M J, Farmer G (eds.) 1981 *Climate and History*. Cambridge University Press, Cambridge, UK

C. Pfister

Climate Impacts

1. Introduction

While there is scientific consensus that increased atmospheric concentrations of greenhouse gases will likely raise global temperatures, with associated increases in global precipitation and sea level, there is no consensus on how fast and how much the climate may change, on how regional climates may change, or on how climate variability may change. Climate change impact assessment for a country or region consists of a set of tasks beginning with problem definition and leading through sector analysis to analysis of adaptation methods and response policies.

A broad understanding of the potential future with climate change demands multifaceted analyses, involving study of both biophysical and socioeconomic processes. A wide range of methods for climate change impact analysis has been developed, from simple

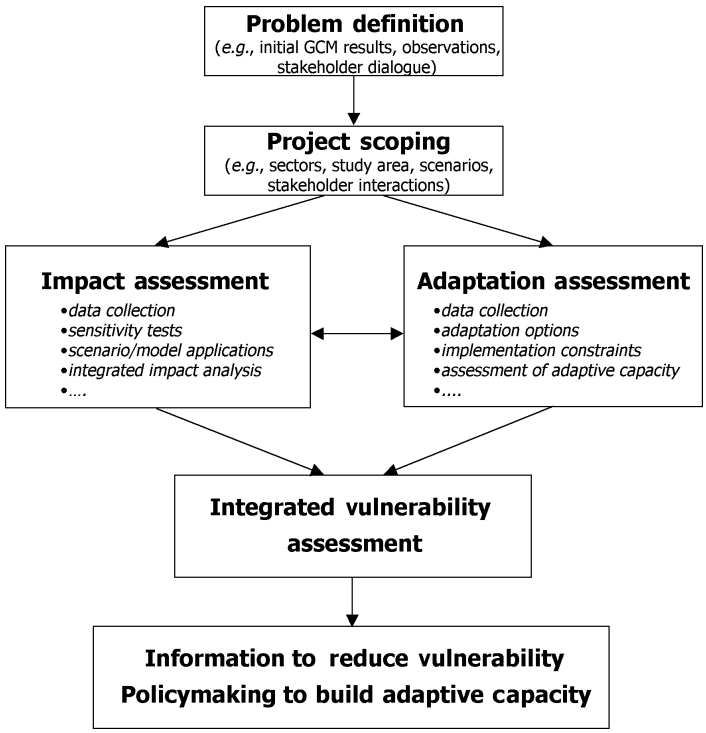

Figure 1
Integrated impacts, adaption, and vulnerability framework
Source: Rosenzweig and Iglesias 2000

regression models to complex integrated systems models. Techniques are becoming ever more complex as more interacting systems and the propagation of uncertainties are included in the analysis. The challenge is to simulate the biophysical and socioeconomic aspects of a system (such as agriculture, human health, urban areas) in a framework appropriate to regional, national, international, and global scales. Spatial analyses and first-order biophysical impacts are important, as well as assessment of vulnerability in the socioeconomic welfare of the different components of the system. Thus, biophysical scientists and social scientists must work together to provide realistic assessments of how climate change might affect a system in the future (see Fig. 1).

Methodological issues to be resolved include how to generalize from the enormous heterogeneity of exposure units and systems and how to address spatial scales and units of analysis from field to region to nation and beyond. Models must be continually tested, calibrated, and validated, and improved for their use to be well-founded. The inclusion of the transient nature of climate change and its associated uncertainties in the modeling techniques is particularly important.

2. *Approach*

There are several approaches that serve as foundations to climate change impact studies. One approach is based on climate change scenarios, that is, projections of what future climate variables (and the characteristics of future impacts) may be like. Equilibrium climate change scenarios have been most often used in this approach, but recent more realistic studies and projections incorporate dynamic or 'transient' climate change scenarios. The climate change scenarios approach sometimes includes the study of responses of the system to past climatic variations, in order to allow comparison with future projections.

Another approach is threshold-based, and attempts to define the limits of sensitivity of a system as it is currently configured to changes in climatic variables. The first approach addresses the question, 'What will the system be like in a given future changed climate?' while the threshold approach asks, 'What type, magnitude, and rate of climate change would seriously perturb the system as we know it?' This approach most often applies transient scenarios of climate change. Both of the approaches construct a chain of causality from the biophysical responses at a small scale to

socioeconomic effects at the regional, national, and international levels.

Several different techniques from the field of economics have been used in climate change impact analysis. One technique is the utilization of economic data to estimate the value of climate to the exposure unit (i.e., farmers) implicitly through regression equations. Linear programming models of the national sector (i.e., agriculture) are also used, as well as linked national and regional models.

Analysis of adaptive responses to climate change is an important part of climate change impacts research. The biophysical approaches described above allow the explicit examination of exposure unit adaptations, while the economic approach deals with adaptation implicitly.

3. *Climate Change Scenarios*

Climate change scenarios are defined as plausible combinations of climatic conditions that may be used to test possible impacts and to evaluate responses to them. Scenarios may be used to determine how vulnerable a sector is to climate change, to identify thresholds at which impacts become negative or severe, and compare the relative vulnerability among sectors in the same region or among similar sectors in different regions.

It is still difficult, if not impossible, to associate probabilities with any particular scenario of climate change, due to uncertainties in future emissions of radiatively active trace gases and in the response of the climate system to those emissions. Thus, impact studies based on climate change scenarios do not make actual predictions; rather, they are useful in defining for critical biophysical and socioeconomic systems directions of change, relative magnitudes of change, and potential critical thresholds of climate-sensitive processes. By conducting climate change impact analyses, researchers and resource managers are conducting 'practice' exercises, which help to engender flexibility in the systems' responses to potentially changing conditions in the future.

3.1 *Arbitrary Scenarios*

The simplest scenario is the application of prescriptive changes, such as a 2 °C increase in temperature and/or a 10 percent decrease in precipitation, to observed climate. Tests with such simple changes can help to identify the sensitivities of systems to changes in different variables. One can isolate the effects of one climate variable, for example, temperature, while holding other variables constant. However, such tests do not offer a consistent set of climate variables, since evaporation, precipitation, wind, and other variables are all likely to change with change in temperature. Arbitrary scenarios do, however, provide a set of responses to which other types of scenarios may be compared.

3.2 *Historical Analogs*

Another type of climate change scenario is based on the historical record. Observations from cool or warm, wet or dry historical periods are used to construct scenarios for use in modeling studies of climate change impacts. Such periods are also useful for the insights provided by studying the responses of any given system to periods of climatic extremes. The Dust Bowl of the 1930s in the Southern Great Plains is a well-known example (see e.g., Warrick 1984), but past freeze events, aquifer depletion, and lake-level changes have also been used to study societal responses to regional climate change (Glantz 1988).

A difficulty with either of these scenario approaches as proxies for the global warming currently predicted for increasing CO_2 and other trace gases is that the patterns of climate warming may be different depending on the nature of the atmospheric forcing mechanisms.

3.3 *GCM-based Scenarios*

Climate change scenarios are also derived from global climate model (GCM) experiments with specified forcing mechanisms (e.g., 1 percent annual increase in greenhouse gas concentrations in the atmosphere). Current GCM model experiments are conducted to produce transient climate projections. The advantages of GCM scenarios are their internal consistency and global extent. GCMs estimate how regional and global climates may change in response to increased concentrations of trace gases. Thus, regional and global climate responses are internally consistent. The climate variables are also physically consistent, as heat, moisture, and energy processes are calculated from a consistent set of equations representing physical processes.

At present, GCMs represent current climate at global and zonal (latitudinal) scales, but do not do particularly well at simulating regional current climate. Differences in climate projections among GCMs increase as scale decreases from the global to the regional and gridbox levels. GCM simulation of current temperature regimes is better than simulation of current hydrological regimes. A range of GCM scenarios should be included in the design of impact studies in order to incorporate a range of climate sensitivities to greenhouse gas forcing, and it is very important to consider GCM regional climate change projections as examples of possible future climates, rather than actual predictions.

Because GCM simulation of current climates is often inaccurate, direct projections of GCM-generated

future climates is seldom used. Changes in climate variables in the perturbed simulations relative to the control run are often applied to historical observed weather data to create the climate change scenarios used in impact studies. Absolute model biases are omitted by using the relative changes. Thirty years of current climate data are often used to develop the baseline climate scenario to which the GCM changes are applied. A 30-year period is considered long enough to represent 'normal' climate variability. Recent periods, e.g., 1951–80, or 1961–90, are often selected, representing current climate and having accurate data most easily available. The latter period contains some of the warmest years on record that may have been caused by the enhanced greenhouse effect.

The use of GCM transient scenarios (i.e., time dependent) in climate change impact studies is growing, since they provide a much more realistic picture of the projected warming from current conditions to some point in the future.

4. *Integrated Global Change Scenarios*

Climate is not the only factor that will be changing as the twenty-first century unfolds. Population growth and changing economic and technological conditions are likely to affect world society and the environment even more than changes in climate. It is important to take such changes into account in climate change impact analyses: first, because climate change will occur not in the present but in the future, and second, because such changes may affect the sensitivity of a system or sector to climate. However, predicting population growth rates and future economic conditions is equally if not more uncertain than predicting the future climate. Therefore, future scenarios need to be designed carefully to address a range of possible conditions. One approach is to contrast 'optimistic' and 'pessimistic' views of the future. In the optimistic scenario, population growth rates are low, economic growth rates and incomes rise, environmental pollution decreases, and land degradation abates. In more pessimistic scenarios, population growth rates are high, economic growth rates and incomes are low, environmental pollution increases, and land degradation accelerates. A scenario of no change (i.e., present conditions) should also be included. The differential effects of climate change on current conditions, and on these two alternative scenarios of the future may then be evaluated.

In order to place possible changes in climate in the context of potential socioeconomic changes, estimates of population, economic growth, and technological change are needed. These estimates will also affect future rates of CO_2 and other greenhouse gas emissions. Economic projections beyond the next 10 to 20 years generally are unreliable. Furthermore, socioeconomic factors are not unrelated, since changes in population are likely to affect national and per-capita income. Recent IPCC scenarios include estimates of population and economic growth rates for a set of possible futures (IPCC 2000).

Socioeconomic factors that are often considered in future scenarios include population, income, productivity, and technology levels. Environmental factors may include stratospheric and tropospheric ozone levels and changes in land use. Institutions and legal structures may change as well, but these evolutions are very hard to predict. The World Bank (1994) and the United Nations (1999) have published population estimates by country through 2100 for a range of scenarios. The World Bank (1993) has published estimates of changes in income. Various economic models are used to project such productivity factors as gross domestic product (GDP) into the future. Population and economic growth may bring increases in urbanization, expansion of agriculture and mining of natural resources, and accelerating rates of deforestation, habitat fragmentation, desertification, and water and air pollution (FAO 1993, Dregne and Chou 1992).

4.1 *CO_2 and Greenhouse Gas Emission Scenarios*

CO_2 and greenhouse gas emission scenarios are needed, especially for agriculture because of the need to estimate crop responses to the CO_2 fertilization effect, as well as projections of sea-level rise. Climate modelers need estimates of future levels of atmospheric CO_2 and other trace gases in order to prepare transient scenarios of future climate. Crop and forest modelers also need such estimates in order to take fertilization effects into account in their impact analyses. Global emissions of CO_2, the most important greenhouse gas, depend primarily on fossil fuel use in three major sectors—electrical generation, industry, and transportation. A growing world economy consists of growth in industrial production, consumption of goods, and travel and concomitant increases in energy use. Deforestation also contributes to CO_2 emissions and is linked to economic growth as land is converted from natural ecosystems to agriculture and other uses.

The IPCC (2000) and others estimate growth rates of world carbon emissions from fossil fuels and deforestation in order to calculate atmospheric CO_2 levels for climate projections. Such calculations are also important in international negotiations that consider limiting CO_2 emissions. Since only a portion (about one-half) of the carbon added to the atmosphere remains, carbon cycle models are used to translate carbon emissions into atmospheric levels of CO_2. Models that include the effects of CO_2 fertilization, feedback from stratospheric ozone depletion, and the radiative effects of sulfate aerosols have been

combined to project radiative forcing of climate, changes in global-mean temperature, and sea level (Wigley and Raper 1992). Recent projections have tended to reduce the projected rates of warming and sea-level rise, but they are still four to five times the rates observed over the twentieth century.

5. *Modeling Techniques: A Case Study of the Agricultural Sector*

Modeling techniques of several kinds are used to study potential impacts and responses of agriculture to changing climate and atmospheric composition. The agricultural sector is chosen to illustrate the range of modeling techniques, because agriculture is a key socioeconomic sector for development in many regions, agricultural land use is a primary driver of land-use change, and agriculture is a sector vulnerable to global environmental change. Choice of technique depends on the sphere of analysis considered and the research questions posed.

5.1 *Biophysical Modeling*

5.1.1 Crop suitability. Spatial analysis consists of identification of critical environmental limits (primarily climate, soil and water resources) of specific crops or agricultural systems, applications of climate change scenarios, and calculation of resulting spatial shifts in crop or agricultural regions. This agroclimatic method provides an approximation of possible changes in crop areas from a biological perspective, but does not address potential changes in either yield or production.

5.1.2 Potential production. Potential production may be estimated from climatic variables or indices such as length of growing season, precipitation, evapotranspiration, solar radiation, and temperature. A prime example of this technique is found in the Agro-Ecological Zone Project of the FAO (FAO 1978). The FAO Agro-Ecological Zone modeling technique simulates both crop zonation and potential production (Leemans and Solomon 1993, Cramer and Solomon 1993, Fischer et al. 2001).

5.1.3 Statistical regression models. Multiple regression models have been developed from the statistical relationships between historical crop yields and climatic variables in specific locations (i.e., Waggoner 1983). The use of regression models is limited by their lack of explanatory power, since the techniques rely on statistical coefficients rather than on descriptions of the underlying biophysical relationships.

5.1.4 Dynamic crop models. Dynamic crop growth models formulate the principal physiological, morphological, and physical processes involving the transfers of energy and mass within the crop and between the crop and its environment. Such models have been developed for most of the major crops, with the aim of predicting their responses to specified climatic, edaphic, and management factors governing production. Dynamic models capable of simulating the response of crops to climatic variables may be used in conjunction with GCM climate change scenarios to explore the consequences of increased atmospheric CO_2 and climate change on yields and phenology or to determine thresholds of crop growth sensitivity to changing climate variables. They are also useful for testing possible adaptations to climate change, such as altered planting dates, irrigation scheduling, or crop variety (see Rosenzweig and Iglesias (1994) for applications of crop models to climate change impact evaluation).

5.2 *Economic Techniques*

Economic measures are an important component of the information that policymakers need to evaluate the climate change issue. Economic analyses are concerned with the reciprocal relations between physical and biological changes on the one hand and the economic responses of individuals and institutions on the other. Once crop yield impacts are estimated, it is useful to translate such biophysical responses into economic measures of human welfare. While biophysical analyses focus primarily on the production of agricultural crops, economic analyses consider both producers and consumers of agricultural goods. Economic measures of interest include the responses of input and output market prices to yield changes and the responses in terms of inputs and outputs that affected individuals make to minimize losses or maximize gains, based on the changes in production and consumption opportunities and in price. If climate change causes substantial changes in outputs, price and quality changes can result, which, in turn, can lead to further market-induced output changes. Even if prices remain constant, accurate indications of output changes are needed if production practices and types of outputs may change.

Previous work on the economics of environmental stresses on agriculture has resulted in a number of general findings (Rosenzweig and Hillel 1998). Important points are that both producers and consumers are included in the domain, that economic activities constitute a type of societal adaptation to environmental stresses, leading in the most part to mitigation of negative effects, and that environmental stresses have differential effects on the comparative advantage of regions and countries.

Economic models calculate estimates of the potential impacts of climate change on measurable

economic quantities, including production, consumption, income, gross domestic product (GDP), and employment. It is important to remember, however, that these may be only partial indicators of social welfare. Different social systems, households, and individuals may not be represented in models that are based on producer and consumer theory. Furthermore, many of the economic models do not account for climate-change induced alterations in land availability and water for irrigation; these nonmarket aspects of a changing climate may be critical.

As a starting point, the gathering of available information about production, consumption, and policies provides a framework for determining the existence and possible magnitude of economic vulnerability in the agricultural sector (US Country Studies Program 1994). Microeconomic farm-level models are designed to simulate the decision-making process of a representative farmer in regard to methods of production and allocation of capital, labor, and land and infrastructure. Such models are based on the goal of maximizing economic returns to inputs. Some farm-level models include a range of farmer behavior in regard to risk, for example risk-averse or risk-neutral.

Macroeconomic equilibrium models of the agricultural sector include price-responsive behavior for both consumers and producers. Equations for these relationships are developed based on economic principles that consumers will maximize the utility of their food-buying and that producers (farmers) will minimize their costs of production. Such models usually are calibrated for a given reference year; for climate change purposes, the models solve for the reference year given perturbations in crop production and water supply and demand for irrigation derived from biophysical techniques (see e.g., Adams et al. 1990). Population growth and improvements in technology are set exogenously (i.e., not computed dynamically in the model). Model results include equilibrium prices and quantities.

General equilibrium economic models are useful because they measure the potential magnitude of climate change impacts on the economic welfare of both producers and consumers of agricultural goods. They do not, however, provide a detailed picture of how the economy will respond over time. These models may overestimate the adjustment of the agricultural economy to climate change. Results of changes in production and prices from agricultural sectoral models can then be used in general equilibrium models of the larger economy.

Regression models have been developed that test for statistical relationships between climate variables and economic indicators such as farm values. Some recent studies utilize these methods known as the 'Ricardian' approach (e.g., Mendelsohn et al. 1994, Polsky and Easterling 2001). The behavior of consumers is not included in this approach and world food prices, domestic farm output prices, and thus farm revenues that are dependent on changes in agricultural production inside and outside of the US are assumed to be held constant.

6. *Integrating across Sectors*

Integrated studies link the biophysical and economic realms, and ideally may extend to interactions both within and across sectors such as agriculture and its competing demands for water by irrigators or urban users, or shifting patterns of land use between agricultural and forest (or other natural) ecosystems. This is a more realistic, but more complicated approach, because individual biophysical and socioeconomic sectors will not be affected by climate change in isolation. For example, agricultural responses will be sensitive not only to changes in crop yields, but also to alterations in water supplies, demand for water from other sectors, and to the inundation and salinization of arable land by rising seas. The following are some examples of integration in agricultural impact studies:

(a) Parry et al. (1988) report on integrated agricultural sector studies in high-latitude regions in Canada, Iceland, Finland, USSR, and Japan, that involved teams of meteorologists, agronomists, and economists. The general conclusions of the studies were that warmer temperatures may aid crop production by lengthening the growing season at high latitudes, but that potential for higher evapotranspiration and drought conditions may counteract the positive effects and may even be detrimental to productivity.

(b) Adams et al. (1990) conducted an integrated study for the US, linking models from atmospheric science, plant science, and agricultural economics. While the outcomes for US agriculture in the study depended on the severity of climate change and the compensating effects of carbon dioxide on crop yields, the simulations suggest that irrigated acreage will expand and that regional patterns of US agriculture will shift with predicted global warming. With the more severe climate change scenario tested, the movement of US production into export markets was reduced substantially.

(c) The Missouri, Iowa, Nebraska, and Kansas (MINK) study integrated potential biophysical and economic effects of climate change on agriculture and other sectors (Rosenberg 1993). The study incorporated the physiological effects of CO_2 and adaptation by farmers to the climatic conditions of the 1930s. Even with the relatively mild warming (1.1 °C) of the 1930s and with farmer adaptation and CO_2 effects taken into account, regional production declined by 3.3 percent. Given the estimate of 2.5 °C warming for doubled CO_2 conditions, the results of the MINK study imply agricultural losses of about 10 percent (Cline 1992).

(d) Strzepek et al. (1996) linked climate change impacts in Egypt on agriculture, water, and the coastal zone in an economic model. This integrated study demonstrates that the sectors directly affected by climate change need to be analyzed in concert with the other sectors of the economy in sufficient detail so that feedback can be part of the analysis. Egypt was found to be highly vulnerable to the warming as well as to changes in precipitation and river runoff that are forecast to accompany greenhouse-gas-induced climate change.

In its fullest sense, integrated assessment attempts to close the loop by linking the greenhouse gas emissions caused by human activities, the climatic consequences of the emissions, the impacts of the climate changes on important systems, and the feedback of the impacts back to the generation of greenhouse gas emissions. Modeling frameworks have been devised to integrate the causes, impacts, feedbacks, and policy implications of global climate change (Nordhaus 1992, Manne et al. 1993, Hulme and Raper 1993, Alcamo et al. 1993, Edmonds et al. 1993). An example of a feedback in such models is the pathway leading from energy consumption, to greenhouse gas emissions, to climate warming, to changes in demand for energy (e.g., decreases in demand for energy for heating and increases in demand for energy for air conditioning), and thus back to changes in energy consumption. The models may be used, for example, to explore the effects of policies limiting greenhouse gas emissions, the ensuing reduction in global warming, and the alteration of potential climate change impacts, for example, on agriculture.

7. *Thresholds, Risk, and Surprises*

The identification of thresholds in climate change impacts research involves the analysis of the effects of different levels of climate forcing on a system or activity and the identification of possible discontinuities in response. The determination of critical levels of climate change for any given system may be separated into biophysical and socioeconomic realms. In the biophysical realm, although the thermal regimes and responses of managed and unmanaged ecosystems and water resource availability are complex, critical temperatures (minimum, optimum, and maximum) have been defined for many individual processes. In the socioeconomic realm, defining critical levels of warming is more challenging, due, at least in part, to the interplay of supply, demand and prices, and to the adaptability of the system. Here, determining critical levels of warming involves defining relative impacts on actors from diverse geographic and social groups.

For example, global effects of climate change in agriculture measured with current economic valuation techniques generally are predicted to be small to moderate. This occurs because the economic system is, in general, effective in fostering adaptation to the projected biophysical changes. However, the global perspective masks differences in levels of effects, regionally and socially. Studies done to date concur that there will be significant change in global agricultural patterns. All regions are likely to be affected, but large differences occur among regions. While changes in global production with climate change may be small, the potential remains for regional vulnerability to food deficits due to distributional problems of getting food to specific regions and groups of people. For subsistence farmers and people lacking entitlement to food, lower yields may result not only in measurable economic losses, but possibly malnutrition and starvation. Several studies have addressed vulnerability to food deficits explicitly and found potential increases (e.g., Rosenzweig and Parry 1994, Fischer et al. 2001).

Risk can be evaluated when the probability of occurrence of an event is known, but in impact evaluation, the associated probabilities to a particular scenario are generally not known. Therefore, the inclusion of uncertainty (i.e., when the event is known but the probabilities that it will occur are not known) into climate change impact methods is very important and recent studies are now beginning to include explicit methods to deal with it. Earlier studies have often used 'best estimate' scenarios that represent the mid-point of predictions. The inclusion of a range of scenarios representing upper and lower bounds of the predicted effects is more realistic and allows for the propagation of uncertainty throughout a model system. Further, probability distributions of different events may be defined, with contrasts between low probability catastrophic events (surprises) and higher probability gradual changes in climate trends.

One 'surprise' (i.e., when reality departs qualitatively from expectations) may lead to another in a cascade, since subsystems are connected. Complex systems and chaos theory provide conceptual and analytical tools for anticipating and preparing for surprises. Identification of potential surprises and communication of them to the public and policymakers should allow improvements in environmental and societal resilience to surprise. Surprises related to global climate change may be either scientific or societal in nature. The anticipation of surprises in the science of global climate change may be encouraged by efforts to integrate across disciplines, to support a multiplicity of research approaches, and to focus on outlier outcomes and unconventional views. Beyond the anticipation of scientific surprise, it seems worthwhile to increase the resilience and adaptability of social structures, so that the sensitivity to impacts of unexpected or uncertain perturbations is decreased. Such societal preparedness might include the diversification of economic, productive, and technological systems; the establishment of disaster, coping, and entitlement systems; and the creation of adaptive management systems capable of learning from surprises.

See also: Climate Change and Health; Climate Change, Economics of; Climate Policy: International; Environmental Adaptation and Adjustments; Environmental Challenges in Organizations; Environmental Change and State Response; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; Environmental Sciences; Environmental Surprise; Environmentalism: Preservation and Conservation; Global Environmental Change: Human Dimensions; Integrative Assessment in Environmental Studies; International and Transboundary Accords, Environmental; Land Use and Cover Change

Bibliography

Adams R M, Rosenzweig C, Pearl R M, Ritchie J T, McCarl B A, Glyer J D, Curry R B, Jones J W, Boote K J, Allen L H Jr 1990 Global climate change and US agriculture. *Nature* **345**: 219–24

Alcamo J, Kreileman G J J, Krol M, Zuidema G 1993 Modelling the global society-biosphere-climate system: Part 1: Model description and testing. *Water, Air and Soil Pollution* **76**: 1–35

Cline W R 1992 *The Economics of Global Warming*. Institute for International Economics, Washington, DC, p. 399

Cramer W P, Solomon A M 1993 Climatic classification and future global redistribution of agricultural land. *Climate Research* **3**: 97–110

Dregne H E, Chou N -T 1992 Global desertification, dimension, and costs. In: Dregne H E (ed.) *Degradation and Restoration of Arid Lands*. Texas Tech University, Lubbock, TX

Edmonds J A, Pitcher H M, Rosenberg N J, Wigley T M L 1993 Design for the Global Change Assessment Model GCAM. Paper presented at the International Workshop on Integrated Assessment of Mitigation, Impacts and Adaptation to Climate Change, 13–15 October 1993. International Institute for Applied Systems Analysis, Laxenburg, Austria, p. 7

Fischer G, Shah M, van Velthuizen H, Nachtergaele F O 2001 *Executive Summary Report: Global Agro-ecological Assessment for Agricultive in the 21st Century*. IIASA, Laxenburg, Austria

Food and Agriculture Organization of the United Nations 1978 *Report on the Agro-Ecological Zones Project*. Vol. 1. Methodology and Results for Africa. FAO, Rome, p. 158

Food and Agriculture Organization of the United Nations 1993 *Agriculture: Towards 2010*. United Nations, Rome

Glantz M H (ed.) 1988 *Societal Responses to Regional Climatic Change*. Westview Press, Boulder, p. 428

Hulme M, Raper S 1993 An integrated framework to address climate change (ESCAPE) and further developments of the global and regional climate modules (MAGICC). Paper presented at the International Workshop on Integrated Assessment of Mitigation, Impacts and Adaptation to Climate Change, 13–15 October 1993. International Institute for Applied Systems Analysis, Laxenburg, Austria, p. 14

IPCC 1995 *IPCC Guidelines for Assessing Impacts and Adaptations under Changing Climate*. Carter T R, Parry M L, Harasawa H, Nishioka S (eds.), University College London and Center for Global Environmental Research, London and Tsukuba, p. 59

IPCC 2001 *Special Report on Emissions Scenarios*. Nakicenovic N, Swart R (eds.) Cambridge University Press, Cambridge, UK

Leemans R, Solomon A M 1993 Modeling the potential change in yield and distribution of the earth's crops under a warmed climate. *Climate Research* **3**: 79–96

Manne A, Mendelsohn R, Richels R 1993 MERGE—A Model for Evaluating Regional and Global Effects of GHG Reduction Policies. Paper presented at the International Workshop on Integrated Assessment of Mitigation, Impacts and Adaptation to Climate Change, 13–15 October 1993. International Institute for Applied Systems Analysis, Laxenburg, Austria, p. 14

Mendelsohn R, Nordhaus W D, Shaw D 1994 The impact of global warming on agriculture: A Ricardian analysis. *The American Economic Review* **84**(4): 753–71

Nordhaus W D 1992 The DICE model: Background and structure of a dynamic integrated climate economy model of the economics of global warming. Cowles Foundation Discussion Paper No. 1009. New Haven, CT

Parry M L, Carter T R, Konijn N T (eds.) 1988 *The Impact of Climatic Variations on Agriculture. Volume 1: Assessments in Cool Temperate and Cold Regions*. Kluwer, Dordrecht, The Netherlands, p. 876

Polsky C, Easterling W E III 2001 Adaptation to climate variabilily and change in the US Great Plains: A multi-scale analysis of Ricardian climate sensitivities. *Agriculture, Ecosystems and Environment* **85**: 133–44

Rosenberg N J (ed.) 1993 *Towards an Integrated Impact Assessment of Climate Change: The MINK Study*. Kluwer, Dordrecht, The Netherlands

Rosenzweig C, Hillel D 1998 *Climate Change and the Global Harvest: Potential Impacts of the Greenhouse Effect on Agriculture*. Oxford University Press, New York, p. 324

Rosenzweig C, Iglesias A (eds.) 1994 *Implications of Climate Change for International Agriculture: Crop Modeling Study*. US Environmental Protection Agency, Washington, DC

Rosenzweig C, Parry M L 1994 Potential impact of climate change on world food supply. *Nature* **367**: 133–8

Strzepek K M, Onyeji S L, Saleh M, Yates D N 1996 An assessment of integrated climate change impacts on Egypt. In: Strzepek K M, Smith J B (eds.) *As Climate Changes: International Impacts and Implications*. Cambridge University Press, Cambridge, UK, pp. 180–200

United Nations 1999 Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat (1999). *Long-range World Population Projections: Based on the 1998 Revision* (ESA/P/WP.153)

US Country Studies Program 1994 *Guidance for Vulnerability and Adaptation Assessments*. Washington, DC

Waggoner P E 1983 Agriculture and a climate changed by more carbon dioxide. In: *Changing Climate*. National Academy of Sciences Press, Washington, DC, pp. 383–418

Warrick R A 1984 The possible impacts on wheat production of a recurrence of the 1930s drought in the U.S. Great Plains. *Climatic Change* **6**: 5–26

Wigley T M L, Raper S C B 1992 Implications for climate and sea level of revised IPCC emissions scenarios. *Nature* **357**: 293–300

World Bank 1993 *Income Projections*. Washington, DC

World Bank 1994 World Population Projections 1994–95: Estimates and Projections with Related Demographic Statistics. Eduard Bos, My T. Vu, Ernest Massiah, and Rodolfo A. Bulatao. 532 pages. Published for the World Bank by The Johns Hopkins University Press, Baltimore, MD

C. Rosenzweig

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Climate Policy: International

As early as the closing years of the nineteenth century natural scientists postulated that human activities, particularly burning fossil fuels, could lead to climate change, but it was only late in the twentieth century that the international community began to develop public policies to deal with climate change. The lag resulted from the necessity of developing scientific understanding and public awareness and the immense difficulty of crafting and implementing policies that would be effective. The climate is a classic common good. Policies to deal with it must overcome all of the obstacles to providing common goods.

1. History

Svante Arrhenius, a Swedish chemist and Nobel laureate, first put forth the theory that rising concentrations of carbon dioxide would lead to global warming in 1896. According to his calculations a doubling of carbon dioxide (CO_2) in the atmosphere would result in an increase of 5° Celsius in global average surface temperatures (Arrhenius 1896). Arrhenius's forecast was based on the concept of a greenhouse effect, which was first advanced by Jean-Baptiste Fourier in 1827. Fourier argued that the sun's rays would enter the atmosphere, but not all of them would escape, thus warming the atmosphere. John Tyndall developed this idea by suggesting that particular atmospheric gases—water vapor and carbon dioxide—were responsible for the greenhouse effect. Arrhenius's contribution was to suggest that the strength of the greenhouse effect would increase. His forecast, however, was largely ignored until the second half of the twentieth century. The dominant view was that the climate was essentially constant except for short-term fluctuations.

Starting in the 1950s, scientists, nongovernmental organizations (NGOs), and eventually governments began to pay greater attention to the possibility of climate change. Scientific advances were important. Roger Revelle and Hans Suess of the Scripps Institute of Oceanography developed doubts about the ability of the oceans to absorb the amount of carbon dioxide that was being emitted (Revelle and Suess 1957). The belief that the oceans could absorb all of the carbon dioxide was a principal factor shaping the view that the climate was constant. The increasing availability of data was also important. In 1957 a station to measure atmospheric carbon dioxide was established at the Mauna Loa observatory in Hawaii. By the 1960s this station was reporting a steady increase in carbon dioxide concentrations in the atmosphere. In 1963 the US-based Conservation Foundation sponsored a meeting on climate issues. The report forecast that a doubling of carbon dioxide in the atmosphere would produce a 3.8 °C temperature increase. In 1964–5, Roger Revelle chaired the Environmental Pollution Panel of the US President's Science Advisory Committee. The panel concluded that projected increases in carbon dioxide concentrations in the atmosphere could produce changes in the climate (USA/PSAC 1965).

Deepening concern with climate change stimulated research. Fourier, Tyndall, and Arrhenius had provided the broad framework for this research. Models could be built using the framework. Testing the models would require collecting and assembling data. Lewis Frye Richardson, a British meteorologist, developed the first numerical weather prediction system in the early years of the twentieth century (Richardson 1922). Although at the time the calculations required were too complex for his system to have any practical use, when computers were developed his system could be applied. It provided the basis for the development of General Circulation Models (GCMs) which simulate the global circulation of the atmosphere. GCMs are basic instruments for weather and climate prediction. To study the climate, GCMs must be coupled with other models, particularly of ocean circulation.

The Mauna Loa observatory continued to provide data documenting an increase of CO_2 concentrations in the atmosphere. The World Meteorological Organization (WMO) established the World Weather Watch (WWW) in 1968. WWW promotes standardized observations and the exchange of data. In 1967 WMO and the International Council of Scientific Unions (ICSU) launched the Global Atmospheric Research Program (GARP). GARP was an international research program that focused on the transient behavior of the atmosphere and the factors that determine statistical properties of the general circulation of the atmosphere. GARP culminated in a 12 months global experiment that began on 1 December 1978, in which the earth's atmosphere and weather were observed and measured.

As scientific knowledge of the climate system increased, scientists issued firmer warnings about the possibility and dangers of climate change. In 1977 a US National Academy of Sciences panel that Roger Revelle chaired concluded that if the use of fossil fuels continued to increase at present rates, average global surface temperature could rise by about 6 °C over the coming 200 years, with potentially ominous consequences for agriculture and fishing (USA/NAS 1977). Public and governmental concern mounted during the 1970s. The possibility of climate change and the effects of this change were discussed at a series of UN-sponsored large-scale conferences on the human environment (1972), food (1974), water (1977), and desertification (1977).

WMO initiated a program on the climate in 1974 and, together with the United Nations Environment Program (UNEP) and the International Council of Scientific Unions, convened the First World Climate Conference in Geneva, Switzerland in February 1979.

The First World Climate Conference brought together experts who focused on the scientific aspects of climate change. Later in 1979, on the basis of the conference's recommendation, WMO established the World Climate Programme (WCP). The WCP consists of the World Climate Research Program (which succeeded GARP), the World Climate Data and Monitoring Program, the World Climate Applications and Services Program, and the World Climate Impact Assessment and Response Strategies Program. The WCP fosters international research, coordinates data and monitoring activity, and facilitates access to information.

Stimulated by a sequence of abnormally hot years, concern about climate change grew stronger during the 1980s. Pressure grew for the adoption of international policies to mitigate climate change. WMO and UNEP sponsored a workshop of climate scientists on Developing Policies for Responding to Climatic Change' in Villach, Austria and Bellagio, Italy in the fall of 1987. In their report on the workshop the scientists concluded that as a consequence of the emission of greenhouse gases (GHG) average global surface temperatures would increase at a rate of 0.3 °C per decade. There would also be changes in precipitation and soil moisture and sea level rise. They called for a treaty to reduce GHG emissions (WMO 1987). In June 1988, the government of Canada organized an expert conference in Toronto on 'The Changing Atmosphere: Implications for Global Security.' The conference called for a reduction in deforestation and a 20 percent cut back in CO_2 emissions from 1988 levels by 2005 with the eventual aim of a 50 percent cut back (WMO et al. 1988). Representatives in many national legislative bodies and other governmental officials called for action.

The Toronto conference brought out sharp differences between the USA and many other countries. These differences have surfaced regularly. Because the USA had the largest greenhouse gas emissions, and because of its place in the global economy and the strength of its scientific establishment, these differences have had a profound impact on the negotiations about the climate change regime.

In 1978 the National Climate Program Act (PL 95-367) established the United States National Climate Program Office in the National Oceanic and Atmospheric Administration which coordinated research efforts on climate change throughout the federal government and US involvement in international programs. US national research efforts were substantial and the USA played an important role in international research efforts. The USA, however, was concerned about the costs of cutting GHG emissions (Brenton 1994, pp. 167–9, Rowlands 1995, pp. 74–6) and sought to organize a mechanism that would have broad legitimacy for obtaining objective scientific advice. The USA was an important leader in the decision in 1988 by the World Meteorological Organization and the United Nations Environment Program to create the Intergovernmental Panel on Climate Change (IPCC).

IPCC is open to all members of WMO and UNEP. Its mandate is to assess scientific, technical, and socioeconomic information relevant to understanding the risk of human-induced climate change. It has three working groups and a task force. Working Group I assesses the scientific aspects of the climate system and climate change. Working Group II addresses the vulnerability of socioeconomic systems to climate change and options for adapting to it. Working Group III assesses options for mitigating climate change. The Task Force on National Greenhouse Inventories oversees the National Greenhouse Gas Inventories Program. The IPPC does not carry out research. It bases its assessments mainly on published and peer reviewed scientific literature. Governments appoint members of IPCC.

The Intergovernmental Panel on Climate Change issued its first scientific assessment in 1990 (Houghton et al. 1990). The Second World Climate Conference was held later that year. The IPCC report and the Second World Climate Conference laid the scientific foundation for negotiations toward a treaty on climate change. The IPCC stated in its report that it was confident of the existence of a natural greenhouse effect and that emissions resulting from human activities were substantially increasing the atmospheric concentrations of greenhouse gases. It specifically mentioned carbon dioxide, methane, chlorofluorocarbons, and nitrous oxide. Carbon dioxide, chlorofluorocarbons, and nitrous oxide are removed from the atmosphere only slowly. Their sources and sinks in the atmosphere, biosphere, and oceans determine their atmospheric lifetimes, which can last from decades to centuries. The IPCC concluded that global mean surface air temperature had increased by 0.3 °C to 0.6 °C over the past 100 years and the size of this increase was consistent with the predictions of climate models.

Greenhouse gas emissions are essentially a product of the number of people times the level of development. The energy intensity of development and the carbon intensity of energy use modify the effects of the level of development. Under IPCC's business-as-usual scenario, greenhouse gas emissions would increase following historically based trajectories. The IPCC predicted that the consequence of the increased concentration of long-lived gases in the atmosphere would be an increase in the global mean surface temperature of about 0.3° Celsius per decade. This would result in an increase of about 1 °C by 2025.

The IPCC report and the Second World Climate Conference, where the report was discussed and its conclusions reaffirmed, provoked concern and also made it obvious how complicated dealing with climate change would be. Governments responded with what in the history of international diplomacy must be

regarded as alacrity. The first session of the Intergovernmental Negotiating Committee for a Framework Convention on Climate Change was held in February 1991.

The goal was to have a treaty ready for signature at the United Nations Conference on Environment and Development, which was held in Rio de Janeiro in June 1992. The Rio Conference was an immense gathering (Brenton 1994, pp. 223–35). Representatives of 178 governments attended including 117 heads of state or government. More than 1,400 NGOs were represented and there were more than 35,000 accredited representatives. The deadline set by the date of the Rio Conference and the publicity associated with the conference put considerable pressure on governments.

Despite this pressure, the USA insisted that it would not sign a treaty that contained binding emissions limitations. Many NGOs and several governments wanted the treaty to include emission limitations. The US administration led by President George H. W. Bush did not believe that at that time the USA could fulfill a commitment to limit emissions, and it doubted the capacity of other countries to do so. Eventually the US position was accepted (Soroos 1997, pp.191–200), and in June 1992 154 states and the European Union (EU)—in legal terms the European Economic Community—signed the United Nations Framework Convention on Climate Change (UNFCCC). The treaty came into effect in March 1994 after 50 states had become parties. As of mid-2000, there were 184 parties to the convention.

2. *The United Nations Framework Convention on Climate Change*

The UNFCCC is one of the most far-reaching treaties ever negotiated. Since greenhouse gas emissions are the result of so many aspects of modern life, becoming a party to the treaty commits states to subjecting almost all aspects of their economic activities to some form of international scrutiny and supervision. The legal instruments that will be negotiated under the UNFCCC will have profound impacts on national economies.

The objective of the treaty is '… stabilization of greenhouse gas concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system' (Article 2). Greenhouse gases are defined as 'gaseous constituents of the atmosphere, both natural and anthropogenic, that absorb and re-emit infrared radiation.' In its first scientific assessment IPCC concluded that during the decade from 1980 to 1990 carbon dioxide had contributed 55 percent of the greenhouse gases to the change in radiative forcing, chlorofluorocarbons 24 percent, methane 15 percent, and nitrous oxide, 6 percent. Carbon dioxide emissions result primarily from fossil fuel combustion and land use and land use changes. Agriculture contributes a large share of methane and nitrous oxide emissions.

The production and use of chlorofluorocarbons were controlled under the 1985 Vienna Convention for the Protection of the Ozone Layer and the 1987 Montreal Protocol on Substances that deplete the Ozone Layer and subsequent amendments. These accords phase out the production and use of chlorofluorocarbons.

The first principle of the UNFCCC is that 'The Parties should protect the climate system for the benefit of present and future generations of humankind, on the basis of equity and in accordance with their common but differentiated responsibilities and respective capabilities' (Article 3). Other principles include exhorting the parties to give special consideration for developing countries, take 'precautionary measures,' promote sustainable development, and promote a 'supportive and open international economic system.'

All parties to the treaty undertake commitments (Article 4) to meet the treaty's objective. Though these commitments are stated in general terms in the framework convention, they are to be made more specific through additional legal instruments negotiated later. The UNFCCC defines various categories of countries and establishes differential responsibilities for them. There are three categories of parties to the treaty: developed countries, developed countries with special financial responsibilities, and developing countries.

Developed countries, or Annex I countries, are exhorted to take 'immediate action' to limit greenhouse gas emissions. Annex I includes 38 states, of which 13 were Eastern Europe states in transition to democracy and market economies, and the European Union (in formal legal terms the European Community). Article 4, paragraph 2, explicitly requires Annex I countries 'to adopt national policies and take corresponding measures on the mitigation of climate change' by limiting their anthropogenic emissions of greenhouse gases. The Article further requires them to report on the steps that they take with the aim of 'returning individually or jointly to their 1990 levels these anthropogenic emissions of carbon dioxide and other greenhouse gases not controlled by the Montreal Protocol.'

The countries that are listed in Annex II of the UNFCCC are required 'to provide new and additional financial resources to meet the full agreed costs incurred by developing country Parties in complying with their obligations' to produce national inventories of their 'emissions by sources and removals by sinks of all greenhouse gases not controlled by the Montreal Protocol.' They also have responsibilities for providing financial assistance to the developing countries for other agreed tasks (Article 4, paragraph 3). Annex

II includes Annex I countries except those in transition to democracy and market economies.

The requirements for developing countries are more modest. They are specifically required to submit their inventories to the UNFCCC secretariat. Beyond that they are exhorted to adopt policies and take measures to mitigate climate change and adapt to it.

3. *The Kyoto Protocol*

The process of negotiating legal instruments under the UNFCCC was launched at the first Conference of the Parties (COP-1) in Berlin in 1995. The IPCC published its second assessment report in 1996 (Houghton et al. 1996, Watson et al. 1996, Bruce et al. 1996). The report confirmed that climate change was occurring because of human actions and analyzed the impacts of climate change and the measures that could be taken to adapt to and mitigate climate change. The report provided a strong stimulus for further action.

An initial agreement to establish legally binding quantified emissions limitations and reduction commitments was reached at COP-3 in Kyoto, Japan on 10 December 1997. Article 3, paragraph 1 of the Kyoto Protocol would require Annex I countries collectively to reduce their overall greenhouse gas emissions 'by at least five percent below their 1990 levels in the commitment period 2008 to 2012.' Specifically the Kyoto Protocol required the USA to reduce its greenhouse gas emissions to 93 percent of the 1990 levels, Japan to 94 percent, and the European Economic Community to 92 percent. Norway, on the other hand, is allowed to increase its emissions to 101 percent of the 1990 level, Australia to 108 percent, and Iceland to 110 percent. The European Community's obligation applied to the Community as a whole. Under arrangements made within the EU some countries, especially Germany and the UK, which had the largest emissions, agreed to reduce their emissions more than 8 percent below the 1990 levels, which would allow others, such as Greece and Portugal, to increase their emissions above the 1990 levels. Greece would be allowed to increase its emissions by 25 percent and Portugal by 27 percent.

The Kyoto Protocol covered six greenhouse gases that were not covered by the Montreal Protocol: carbon dioxide, methane, nitrous oxide, hydrofluorocarbons, perfluorcarbons, and sulphur hexaflouride. The limitations apply to 'net changes in greenhouse gas emissions by sources and removals by sinks resulting from direct human-induced land-use changes and forestry activities, limited to afforestation, reforestation and deforestation since 1990, measured as verifiable changes in carbon stocks in each commitment period' (Article 3, paragraph 3). The emissions from the other five greenhouse gases that were covered in the protocol were measured in terms of their carbon dioxide equivalents calculated in terms of global warming potential. The methodologies for measuring emissions and calculating global warming potential were those developed by the IPCC.

The Kyoto Protocol included three flexible mechanisms. One was emission trading. Annex I countries would be able to trade emission allowances. For instance, if a party had difficulty meeting its required limitations, it would be able to purchase emission allowances from another party that had emissions that were lower than its limitation. Joint Implementation among Annex I countries was another flexible mechanism. If a party engaged in a project to increase the sinks on another party's territory through reforestation the two parties could share the credit for the increased sinks. The third flexible mechanism was the Clean Development Mechanism. Annex I countries could engage in projects in non-Annex I countries that reduced prospective emissions and share the credit for this. The US government and many economists argued that these flexible mechanisms would make it possible to limit emissions at the least possible cost. The US administration of President William J. Clinton argued that the USA could only meet its commitment in the Kyoto Protocol if it were allowed to use flexible mechanisms.

How well these flexible mechanisms would work was unknown. Emission trading mechanisms were included in US domestic legislation and the 1979 Convention on Long Range Transboundary Air Pollution. Various pilot projects have been undertaken. Whatever the fate of the Kyoto Protocol, flexible mechanisms in some form will likely be part of the climate change regime. Flexible mechanisms will put heavy demands on the capacity of UNFCCC organs and the parties to the convention.

The Kyoto Protocol did not include requirements that specify policies and measures that parties must adopt, although it could have under the terms of the UNFCCC. It did not require that countries impose a carbon tax or ensure that appliances or automobiles meet efficiency standards. Many European states argued that any effort to limit greenhouse gas emissions should require uniform policies and measures. Some US economists also made this argument. The US government preferred that international accords state obligations and that parties should be free to adopt whatever policies and measures they choose to meet their obligations. Policies and measures—whether they are uniform or country-specific—will inevitably be part of the climate change regime. They will also place heavy demands on the capacity of countries that are parties to the UNFCCC.

Of the Annex I countries in the late 1990s only Austria, Germany, Luxembourg, the UK, and the Eastern European countries in transition to democracy and market economies had emissions below 1990 levels (Grubb et al. 1999, p. 82). The UK's emissions fell because of its transition from coal to natural gas

and Germany's because of the collapse of East German industries. Industries had also collapsed in the countries in transition.

As of January 2000 84 countries had signed the Kyoto Protocol and 22 had ratified it. None of the Annex 1 countries had ratified the Kyoto Protocol. Prior to COP III the US Senate adopted a resolution by a vote of 95 to 0 stating that the USA would not agree to a treaty limiting emissions unless limitations also applied to developing countries. After COP III adopted the Kyoto Protocol, Senators from both parties stated that the protocol was not ratifiable. Senators and representatives of industry and labor argued that unless developing countries were included in the emission limitation requirement, developing countries would have an economic advantage. Some argued that industry would move from developed to developing countries.

Without US ratification, it would be difficult for the Kyoto Protocol to come into effect. To come into effect 55 states, which together accounted for 55 percent of the 1990 carbon dioxide emissions of Annex I countries, must have become parties to the protocol (Article 25). US emissions accounted for more than a third of the 1990 totals for Annex I countries.

Several factors contributed to the differences between the USA and other countries and to the US reluctance to accept the Kyoto Protocol. Under the US legal system, individual citizens or NGOs can sue the government to force it to comply with an international treaty to which the USA is a party. The USA does not ratify treaties unless the government is confident that it can comply. Other states have different legal systems where suing the government is much more difficult. For other states becoming a party to a treaty is frequently a statement of intention to try to comply. The size and climate of the USA are factors in the US propensity to consume energy. More compact countries with milder climates have more modest energy requirements. Unlike many developed countries, the US population was growing in the late twentieth century and was projected to continue to grow, leading to increased demands for energy and other products that would increase greenhouse gas emissions. At least partly because the USA was a petroleum producer, the US public was addicted to inexpensive energy. Efforts to increase energy prices regularly produced political outcries.

If the Kyoto Protocol did not come into effect, a new agreement would have to be negotiated. If the protocol did enter into force, it would require that additional limitations should be negotiated for subsequent periods—that is, those after 2008–2012—and that these negotiations should begin no later than 2005 (Article 3, paragraph 9). Eventually a global bargain will have to be struck involving all parties, both developed and developing countries. The process of negotiating instruments to implement the UNFCCC will be long-lasting.

Bringing developing countries into the climate change regime will be essential but extraordinarily difficult to achieve. Developing countries accounted for slightly more than 30 percent of greenhouse gas emissions in 1990, and their emissions were growing rapidly. As of 2000 developing countries accounted for just fewer than 80 percent of the world's population. Future population growth was projected to occur primarily in developing countries. Developing country governments desire rapid economic growth. Population and economic growth will increase developing country greenhouse gas emissions. Developing countries will become the dominant source of greenhouse gas emissions in the twenty-first century.

Developing countries are also extremely vulnerable to climate change. The small island states could be submerged by sea-level rise. Sea-level rise could threaten coastal zones where developing country populations are often concentrated. Countries that are heavily dependent on agricultural production could suffer greatly from climate change. Petroleum-producing countries could see their incomes cut if petroleum consumption were to decrease.

Promoting economic growth is arguably one of the most effective steps developing countries could take to deal with climate change. The more developed it is, the more resources a country has to devote to limiting greenhouse gas emissions. Developed countries also can adapt more easily to climate change. Developing countries would never agree to limit their economic growth. To become effective the climate change regime will have to find ways to promote sustainable development.

4. *Issues*

The climate is a classic common good. Human action anywhere affects the climate and humans everywhere are affected by the climate. Mitigating climate change will require the combined efforts of governments and—in response to government policies—of individuals throughout the world. Article 4 of the United Nations Framework Convention on Climate Change requires that these efforts be based on equity. This acknowledges the common good character of the climate. Parties to the UNFCCC will only act if they feel that the burdens of dealing with climate change are borne equitably. The treaty embodies one concept of equity. The special requirements placed on Annex I and Annex II countries recognize their relative wealth and the fact that because of their historical lead in industrializing they have contributed more to current concentrations of greenhouse gases in the atmosphere. The resolution adopted by the US Senate reflects another concept of equity, the necessity of all states taking at least some action particularly in view of the growing share of greenhouse gases emitted by non-Annex I countries. These different concepts of equity

will have to be reconciled for effective action to be taken.

Knowledge is a second issue. Decades of work have greatly improved the natural science of climate change. Climate change models have become more sophisticated and the amount of data has increased. There are roughly two dozen large-scale models that are used in this work. They produce somewhat different results, particularly with respect to forecasts of regional impacts, but the range of differences with respect to the global average mean surface temperature is not different from the range that has existed since the beginning of the twentieth century. Even though work on the economic and social dimensions of climate change started later substantial progress has been made. Sophisticated models have been developed to estimate future emissions of greenhouse gases and the costs of mitigation strategies (Nordhaus 1994, Watson et al. 1996, pp. 263–396).

Estimates produced by these models vary because they are sensitive to assumptions about appropriate model structure and demographic and economic growth and the availability of demand (energy efficiency) and supply-side (alternative sources of supply) energy options. Research has produced better understanding of the mechanisms that promote compliance with international accords. Despite this progress in the natural and social sciences much remained to be done, especially concerning the economic and social dimensions. Efforts to adapt to and mitigate climate change will require modifying human behavior. Understanding how to promote appropriate modifications in behavior is a crucial issue for research.

See also: Climate Change and Health; Climate Change, Economics of; Climate, History of; Climate Impacts; Tropospheric Ozone: Agricultural Implications; United Nations: Political Aspects

Bibliography

Arrhenius S 1896 On the influence of carbonic acid in the air upon the temperature on the ground. *Philosophical Magazine* **41**: 237–76

Brenton T 1994 *The Greening of Machiavelli: The Evolution of International Environmental Politics*. Earthscan, London

Bruce J P, Lee H, Haites E F (eds.) 1996 *Climate Change 1995: Economic and Social Dimensions of Climate Change: Contribution of Working Group III to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK

Grubb M, Vrolijk C, Black D 1999 *The Kyoto Protocol: A Guide and Assessment*. Royal Institute of International Affairs, London

Houghton J T, Jenkins G J, Ephraums J J 1990 *Climate Change: The IPPC Scientific Assessment, Report Prepared for IPCC by Working Group I*. Cambridge University Press, Cambridge, UK

Houghton J T, Meira Filho L G, Callander B A, Harris N, Kattenberg A, Maskell K 1996 *Climate Change 1995: The Science of Climate Change: Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK

Nordhaus W D 1994 *Managing the Global Commons: The Economics of Climate Change*. MIT Press, Cambridge, MA

Revelle R, Suess H E 1957 Carbon dioxide exchange between atmosphere and ocean and the question of an increase in atmospheric CO_2 during the past decade. *Tellus* **9**: 18–27

Richardson L F 1922 *Weather Prediction by Numerical Process*. Cambridge University Press, Cambridge, UK

Rowlands I H 1995 *The Politics of Global Atmospheric Change*. Manchester University Press, Manchester, UK

Soroos M 1997 *The Endangered Atmosphere: Preserving a Global Commons*. University of South Carolina, Columbia, SC

United States of America, National Academy of Sciences (USA/NAS) 1977 *Climate, Climatic Change, and Water Supply*. National Academy Press, Washington, DC

United States of America, President's Science Advisory Council (USA/ PSAC) 1965 *Restoring the Quality of Our Environment: Report of the Environmental Pollution Panel*. The White House, Washington, DC

Watson R T, Zinyowera M C, Moss R H 1996 *Climate Change 1995: Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analyses: Contribution of Working Group II to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK

World Meteorological Organization (WMO) 1987 *Developing Policies for Responding to Climatic Change: A Summary of the Discussion and Recommendations on the Workshop held in Villach, 28 September–2 October 1987 and Bellagio, 9–13 November 1987*, (WMO/TD 225). WMO, Geneva

World Meteorological Organization (WMO), Environment Canada, United Nations Environment Program 1988 *The Changing Atmosphere: Implications for Global Security, Toronto Canada, 27–30 June 1988, Conference Proceedings*. WMO, Geneva

H. K. Jacobson

Clinical Assessment: Interview Methods

The single most common method of assessment in both clinical practice and research is an interview, whereby the clinician speaks directly to a person to obtain the clinical assessment (Widiger and Saylor 1998). Additional methods of assessment, such as self-report inventories, projective instruments, or laboratory tests, are often used to supplement or inform a clinical interview, but only under quite special circumstances would a clinician rely solely upon one of these other techniques, whereas clinicians and researchers will often rely solely upon a clinical interview.

1. Advantages of a Clinical Interview

Many of the advantages of a clinical interview are somewhat obvious, but worth noting for the record

nevertheless (Groth-Marnat 1997). First, clinical interviews are substantially more flexible than alternative methods. Interviewers can alter the focus, depth, or even the style of an interview to be optimally responsive to the particular demands, interests, or needs of the respondent or the assessment. Response sets (intentional, habitual, or unconscious tendencies to provide false or misleading responses) can affect the validity of a clinical interview (Rogers 1995), but interviewers can themselves be sensitive and responsive to symptom exaggeration, distortion, or denial during the course of an interview. Interviewers may notice if a respondent is being excessively acquiescent or defensive, if the mood state of a respondent is contributing to excessive self-denigration, or if the responses are inconsistent across the interview. The interviewer can then alter immediately the format, style, or scoring of the interview to make adjustments for the response sets, can conduct follow-up queries to assess for the presence of problematic response sets, or even discuss a response set directly with the respondent in order to decrease its effects.

A disadvantage of other methods of clinical assessment is that they will routinely cover domains of functioning that will not be particularly relevant or necessary for an issue or patient at hand and yet, at the same time, fail to cover in adequate depth the domain of functioning of most interest or relevance to the patient and clinician. For example, most omnibus self-report inventories attempt to cover virtually all domains of psychopathology but must then provide an inadequate assessment for any one of them. An interviewer has the unique advantage of being able virtually to abandon a focus of inquiry during the course of an assessment to spend more time and effort on a particular line of investigation.

Finally, the clinician conducting the interview is usually the person who will ultimately provide the clinical report, and seeing or hearing for oneself is usually much more compelling than being told by something or someone else. The clinician is able to see and experience firsthand the person's behaviors, feelings, statements, and manner of relatedness. The presentation of the patient's psychopathology in his or her speech, affect, and behavior within the clinician's office can provide a powerfully vivid and compelling portrayal.

2. *Limitations of Unstructured Clinical Interviews*

Many of the benefits and advantages of a clinical interview, however, fail to be realized in routine clinical practice, as the freedom and authority provided by a clinical interview does not come without substantial responsibilities, costs, and limitations. The reliability and validity of a clinical interview depend substantially upon the conscientiousness, skills, and talents of the clinician. Seeing for oneself can be very compelling, but it can be equally illusory and deceiving. Many clinicians may place excessive faith, or at least have excessive confidence, in their own perceptions and judgments, despite the fact that studies have shown repeatedly that unstructured clinical assessments often obtain poor agreement across different interviewers (Dawes 1995, Garb 1998) (see *Clinical Psychology: Validity of Judgment*; *Clinical versus Actuarial Prediction*). Two clinicians relying upon their own skills, talents, and abilities will often provide different conclusions regarding the same patient. At least one of them will be wrong, but both will believe it is the other clinician. The instrument of the unstructured clinical interview is for the most part the clinician, and there are perhaps few clinicians who truly recognize or adequately appreciate their own limitations, deficits, and flaws.

Many studies have documented that unstructured clinical interviews tend to be unreliable and are highly susceptible to primacy effects, halo effects, false expectations, misleading assumptions, and confirmatory biases (Dawes 1994, Garb 1998, Widiger and Saylor 1998), and this research appears to have had only a marginal effect upon the beliefs or behavior of most individual practitioners (e.g., Westen 1997). The lack of an adequate impact of this research is perhaps due in part to the ability of persons to believe that the research is primarily relevant to persons other than themselves. An advantage of other methods of assessment is that the research indicating, for example, systematic errors within a laboratory instrument, will clearly be relevant to almost any administration of that instrument. Research indicating the systematic errors of a sample of clinicians might not be applicable to a clinician who did not actually participate in that particular study. Clinicians can then argue and believe that the decision-making research is not really applicable to them because they are in fact adequately sensitive and responsive to the issues, errors, biases, or concerns identified in this research.

A major source for the failure of unstructured clinical interviews to provide reliable or valid assessments is the failure to conduct systematic or comprehensive assessments. An innovation of the American Psychiatric Association's (APA) *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*; APA 1994) is the provision of relatively specific and explicit diagnostic criteria for each mental disorder (see *Mental and Behavioral Disorders, Diagnosis and Classification of*). Reliable and valid clinical diagnoses are now readily obtained as long as the interviewer does indeed comprehensively assess every diagnostic criterion in a systematic manner (Nathan and Langenbucher 1999). Many studies, however, have indicated that clinicians will often reach a conclusion after determining the presence of only a small subset of the diagnostic criteria set, and will fail to assess for the presence of additional symptomatology of other possible disorders (Garb 1998, Widiger and Saylor 1998). Zimmerman and Mattia (1999), for example, com-

pared the clinical diagnoses provided for 500 patients who were assessed with unstructured clinical interviews with the diagnoses provided by a semistructured interview that systematically assessed for the presence of the diagnostic criteria for most of the commonly occurring Axis I mental disorders (i.e., mental disorders other than personality disorders or mental retardation). More than 90 percent of the patients receiving the unstructured clinical interview were provided with only one diagnosis, whereas more than a third of the patients assessed with the semistructured interview were discovered to have met the diagnostic criteria for at least three different mental disorders. Comorbidity among mental disorders has substantial significance and importance to clinical treatment, yet it appears to be grossly underrecognized in general clinical practice (Zimmerman and Mattia 1999).

A variety of studies have also indicated that clinicians relying upon unstructured clinical interviews routinely fail to assess for the presence of the specified diagnostic criteria (Widiger and Sanderson 1995). One of the more compelling demonstrations of this failure was provided by Morey and Ochua (1989). Morey and Ochua provided 291 clinicians with the 166 *DSM-III* (APA 1994) personality disorder diagnostic criteria and asked them to indicate which *DSM-III* personality disorder(s) were present in one of their patients and to indicate which of the 166 diagnostic criteria were present. Kappa for the agreement between their diagnoses and the diagnoses that would be given based upon the diagnostic criteria they indicated to be present was very poor, ranging from 0.11 (schizoid) to only 0.58 (borderline). In other words, their clinical diagnoses agreed poorly with their own assessments of the diagnostic criteria for each of the personality disorders. Comparable results have since been reported in many subsequent studies (Widiger and Saylor 1998).

Among the more consistently documented errors in clinical practice are gender and racial biases in the application of diagnostic criteria (Garb 1998). *DSM-IV* diagnostic criteria sets will contain a degree of gender and racial bias (Hartung and Widiger 1998). However, racial and gender biases that have been documented empirically have been due in large part to a failure of clinicians to adhere to the specified diagnostic criteria set for a respective mental disorder (Whaley 1997, Widiger 1998). When clinicians are compelled to follow closely the criteria set for a mood, psychotic, or personality disorder, gender and racial biases are significantly less likely to occur.

3. *Advantages of Semistructured Clinical Interviews*

Limitations of unstructured clinical interviews can be addressed in part through the administration of more structured interview schedules in which a set of specified questions must be administered, the interpretation and scoring of the responses to which are guided by an accompanying manual. These interview schedules can vary substantially in the extent to which the questions are open-ended, observations of the respondent are included, and the interviewer is allowed to conduct follow-up queries. A fully structured interview would be essentially equivalent to a verbally administered self-report inventory (Widiger and Saylor 1998). Most interview schedules, however, are more accurately described as being semistructured, as they will include subtle and indirect questioning, will require follow-up queries, and will include open-ended questions, the responses to which will require professional judgment and expertise for interpretation and scoring. The only difference between some semistructured interviews and a skilled clinician is that the inclusion of a semistructured interview documents explicitly the obtainment of a reliable, replicable, systematic, objective, and comprehensive assessment of all of the relevant symptomatology.

Semistructured interviews are the preferred method for obtaining clinical assessments in research, but are perhaps rarely used in general clinical practice. Clinicians perceive semistructured interviews as being constraining, impractical, or superficial (Westen 1997). Semistructured interviews are indeed constraining, as they are a means by which to ensure that the findings are reliable, replicable, systematic, comprehensive, and objective by constraining the clinician from failing to assess all of the necessary criteria in a minimally adequate manner (Segal 1997). Most semistructured interviews, however, allow and do in fact encourage clinicians to have a significant impact through the administration of follow-up queries and the reliance upon their professional judgment for the scoring of the responses.

One of the major impediments to the implementation of a semistructured interview in general clinical practice is the amount of time that is required for their complete administration. Researchers will often pay both interviewers and patients for two or three hours of interviewing; no such funding luxury is available in general clinical practice. However, the amount of time required for the administration of a semistructured interview can be reduced substantially by first administering a screening questionnaire to narrow the line of inquiry. Screening questionnaires with a high false positive rate (i.e., err in the direction of identifying too much rather than too little psychopathology) are also useful in alerting the clinician to domains of functioning that might have been otherwise neglected.

Many clinicians will also perceive some of the required questioning to be simplistic or superficial. However, it is important to appreciate that a substantial amount of research has informed the development of a particular line of questioning. Semistructured interviews can in fact be an excellent source

for discovering new and effective methods of inquiry. Semistructured interviews, however, will not be as effective as an unstructured interview in establishing rapport. Most clients will appreciate the comprehensive and thorough nature of a semistructured interview, but if the establishment of rapport is a major clinical issue, then a lengthy semistructured interview will at times be problematic.

4. Recommendations for Future Research

A clinical and scientific limitation of many semistructured interviews is the absence of data normally obtained through the course of the development and validation of a psychometric instrument. For example, semistructured interview reliability data are often simply confined to an agreement with respect to the scoring of a previous or concurrent administration of the interview. The poor reliability obtained in general clinical practice is due to inconsistent, incomplete, or idiosyncratic interviewing. It is unclear if some of the semistructured interviews have actually resolved this problem given the absence of studies on the interrater (or test–retest) reliability of independent administrations of the interview (Rogers 1995, Segal 1997, Widiger and Saylor 1998).

There are a variety of different interview schedules to assess the same domains of psychopathology. An advantage of this diversity is the availability of different options to choose from. However, current research suggests that these different interview schedules are providing different findings and it is unclear if the failure to replicate findings across studies is due to idiosyncratic administration of interviews, differences in setting, or differences in the interview schedules. One suggestion has been to confine future research to just one interview schedule (Regier et al. 1998). This confinement would contribute to the obtainment of more uniform results, but at the cost of the failure to appreciate the extent to which the results in fact reflect unique aspects of a particular interview schedule. What is needed are studies comparing directly the concurrent and predictive validity of alternative interview schedules within the same patient sample.

Normative data are also lacking for many of the semistructured interviews. The test manuals that accompany the publication of a semistructured interview are often surprisingly weak in their coverage of reliability and validity data. Diagnoses obtained through the administration of a semistructured interview are used as the criterion by which the validity of other instruments is evaluated, but semistructured interview schedules may rely too heavily for their own derivation on simply face validity. In defense of the validity of semistructured interviews, the most compelling published research concerning the etiology, course, pathology, and treatment responsivity of various mental disorders has relied substantially on the administration of a semistructured interview. The results of this extensive research provide considerable support for the construct validity of the respective semistructured interviews that were used in these studies. In addition, detailed summaries of the reliability and validity of alternative interview schedules are provided in a number of published papers and texts (e.g., Rogers 1995, Segal 1997, Widiger and Sanderson 1995).

5. Conclusions

In sum, the many advantages of semistructured interviews clearly outweigh their limitations and disadvantages. Many are now being used in general clinical practice when the results of the clinical assessment might be subsequently questioned or reviewed (e.g., custody, disability, and forensic assessments). A highly talented clinician can provide a more valid assessment than a semistructured interview, but it is risky to assume that one is indeed that talented clinician. It would at least seem desirable for a talented and insightful clinician to be fully informed by a systematic and comprehensive assessment. Semistructured interviews are used routinely in general clinical research and perhaps will eventually be used routinely in general clinical practice. Individually administered intelligence tests are comparable to a fully structured clinical interview, particularly an assessment of verbal intelligence that involves a series of specified questions, the responses to which are scored according to a test manual. Very few clinicians would attempt to diagnose mental retardation in the absence of the administration of one of these structured interviews. Perhaps in the future no clinician will attempt to diagnose an anxiety, mood, psychotic, dissociative, personality, or other mental disorder without at least considering the results obtained by the administration of a respective semistructured interview.

See also: Clinical Psychology: Validity of Judgment; Clinical versus Actuarial Prediction; Minnesota Multiphasic Personality Inventory (MMPI)

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association, Washington DC

Dawes R M 1994 *House of Cards: Psychology and Psychotherapy Built on Myth*. Free Press, New York

Garb H N 1998 *Studying the Clinician. Judgment Research and Psychological Assessment*. American Psychological Association, Washington, DC

Groth-Marnat G 1997 *Handbook of Psychological Assessment*, 3rd edn. Wiley, New York

Hartung C M, Widiger T A 1998 Gender differences in the diagnosis of mental disorders: Conclusions and controversies of *DSM-IV*. *Psychological Bulletin* **123**: 260–78
Morey L C, Ochua E S 1989 An investigation of adherence to diagnostic criteria: Clinical diagnosis of the *DSM-III* personality disorders. *Journal of Personality Disorders* **3**: 180–92
Nathan P, Langenbucher J W 1999 Psychopathology: Description and classification. *Annual Review of Psychology* **50**: 79–107
Regier D A, Kaelber C T, Rae D S, Farmer M E, Knauper B, Kessler R C, Norquist G S 1998 Limitations of diagnostic criteria and assessment instruments for mental disorders. Implications for research and policy. *Archives of General Psychiatry* **55**: 109–15
Rogers R 1995 *Diagnostic and Structured Interviewing. A Handbook for Psychologists*. Psychological Assessment Resources, Odessa, FL
Segal D L 1997 Structured interviewing and *DSM* classification. In: Turner S M, Hersen M (eds.) *Adult Psychopathology and Diagnosis*. Wiley, New York, pp. 24–57
Westen D 1997 Divergences between clinical and research methods for assessing personality disorders: Implications for research and the evolution of Axis II. *American Journal of Psychiatry* **154**: 895–903
Whaley A L 1997 Ethnicity/race, paranoia, and psychiatric diagnoses: Clinician bias versus sociocultural differences. *Journal of Psychopathology and Behavioral Assessment* **19**: 1–20
Widiger T A 1998 Sex biases in the diagnosis of personality disorders. *Journal of Personality Disorders* **12**: 95–118
Widiger T A, Sanderson C J 1995 Assessing personality disorders. In: Butcher J N (ed.) *Clinical Personality Assessment. Practical Approaches*. Oxford University Press, New York, pp. 380–94
Widiger T A, Saylor K I 1998 Personality assessment. In: Bellack A S, Hersen M (eds.) *Comprehensive Clinical Psychology*. Pergamon, New York, pp. 145–67
Zimmerman M, Mattia J I 1999 Psychiatric diagnosis in clinical practice: Is comorbidity being missed? *Comprehensive Psychiatry* **40**: 182–91

T. A. Widiger

Clinical Psychology: Animal Models

Over the years, considerable debate has surrounded the question of whether psychopathology and its treatment can be studied meaningfully in animals. One side has argued that psychopathological syndromes such as anxiety and depression are uniquely human and cannot be experienced in animals. Another side argues that there are both naturally occurring and experimentally induced psychopathological states in animals that closely parallel those seen in humans. This side also argues that there is so much to be learned from the systematic study in controlled settings of emotional or otherwise disturbed behavior in animals that this more than offsets any problems created by potential species differences. Proponents of this position bring animals into laboratory settings and study them under controlled experimental conditions to help us better understand various aspects of human disorders. The goal in these attempts is to develop an *animal model* of a disorder or its treatment.

1. Historical Background

Although the study of emotions such as fear and sadness in animals dates back at least to Darwin (1872), the experimental study in animals of the neurotic extremes of emotional states and other aspects of psychopathology did not begin until some years later (e.g., Pavlov 1927). Shortly after that time in the United States, where the methods of Pavlov and Thorndike to study learning were enthusiastically embraced, interest in Pavlov's so-called 'neuroses of the experiment' spread during the 1930s and 1940s. Indeed, a number of well-known laboratories were established to study what came to be called *experimental neurosis* (e.g., Liddell, Gantt, Masserman, and N. R. F. Maier). At the time, this work was reasonably influential, partly because other extant models of what causes psychopathology were generally very primitive. By contrast, ideas for studying human psychopathologies through developing animal models seemed remarkably advanced. Indeed, the success of various experimental manipulations in producing disordered behavior and emotions in several different species, was clearly influential in establishing the foundations of behavioral approaches to the etiology and treatment of anxious and depressive disorders.

Unfortunately, this early work on experimental neurosis was fairly unsystematic. Investigators explored the effects of experimental variants on traditional learning paradigms (often discovered accidentally) that seemed to produce disturbed behavior in their animals. However, the next steps were not taken. One would have been to manipulate systematically various aspects of the procedures to determine what the critical (causal) features were. Investigators also needed to demonstrate (but did not) compelling phenotypic (symptomatic) and/or functional similarities between the 'neurotic symptoms' seen in animals and human patients. Instead, somewhat superficial, and often anthropomorphic, assertions of similarity were made. As a consequence of these failures, the study of animal models for psychopathological disorders fell into relative obscurity for several decades (see Mineka and Kihlstrom 1978).

2. Contemporary Use of Animal Models

A resurgence of interest started about 1970, when a number of investigators began to make persuasive arguments that animal models of some disorders could

be very useful if certain criteria are adhered to in developing the model. For example, Seligman (1975) and McKinney (1974) both argued that animal models can be useful if one attempts to document similarities and parallels in the symptoms, etiology, therapy, and prevention of the animal and human syndromes. Obviously, at the outset not all of the parallels will be possible to detail because little may be known about some of these factors (e.g., prevention) for either the animal or the human disorder. Nevertheless, herein lies one of the special advantages of developing an animal model. Initially some compelling similarities must be drawn between the human disorder and the animal model. Then, however, the animal model can be used to test hypotheses about other possible parallels (e.g., prevention) that often cannot easily be tested experimentally with humans. Some of the work developing such full-fledged animal models has been quite successful, as discussed below.

Alternatively, others have argued that requiring adherence to all these criteria may be unnecessarily restrictive, given that for most disorders no such complete models have yet been discovered. This is partly because of significant limitations on the range of human symptoms that can be modeled in animals—especially if expression of that symptom is mediated by higher cortical structures in the brain not shared by most species. There are, however, many very interesting and important 'mini-models' which help illuminate different aspects of the symptomatology, *or* etiology, *or* prevention, *or* treatment of these disorders. Mini-models are emotional, behavioral, cognitive, and/or physiological phenomena studied in animals (or humans) that may clarify some of the most prominent features of the origins or treatment of a disorder. The behaviors and emotional responses are manipulated experimentally through either behavioral or physiological experimental manipulations. Although any given mini-model may illuminate only a subset of prominent features of a disorder, in nature such factors would operate in interaction with other factors in the etiology or treatment of a human disorder.

3. *Animal Models in Psychopharmacology and Behavioral Pharmacology*

Nowhere has the use of animal models been more prominent than in the fields of psychopharmacology and behavioral pharmacology, where researchers develop new medications to treat mental disorders and try to understand how the medications work to ameliorate symptoms. Nearly all medications are initially tested on animals before being approved for use by humans—to determine both their effectiveness and their safety. However, before this can be done, researchers must first develop and validate an animal model of the human disorder (or often just a subset of symptoms of a disorder) before they can test their medications. For example, determining if a new medication serves as an *anxiolytic* (anxiety-reducing) medication, requires knowing how to produce strong symptoms of anxiety in animals (as well as knowing that the measure of anxiety is functionally if not phenotypically, related to human anxiety). The same, of course, applies to treatments of other disorders.

Staying with the anxiety example, at least 30 different animal models have been used to test the effectiveness of anxiolytic drugs. One of the most common uses a conflict/punishment procedure. Rats are first trained to press a bar to obtain food reinforcement on an occasional basis. Later they can still obtain food on some trials but now they also receive an unpleasant electric shock following the food. Not surprisingly, this punishment procedure puts hungry rats in a state of conflict: anxiety about punishment now conflicts with hunger. Typically, rates of responding for the food are diminished substantially unless an effective anxiolytic medication is given (e.g., diazepam from the benzodiazepine category). Having been validated as a model in numerous studies with medications known to reduce anxiety in humans, the model is then used to test potential new anxiolytic compounds.

Researchers must beware, however, that this method produces both false positives and false negatives. Sometimes a medication that seems to work in the animal model, is later shown not to work in humans. Alternatively, a medication that does not seem effective with animal models may nonetheless be effective in humans. In such cases further work is needed to determine the source of the discrepancy. To illustrate, a relatively new anxiolytic compound known as buspirone (not from the benzodiazepine category) did not initially seem effective using the conflict/punishment procedure in rats. However, buspirone operates through different physiological mechanisms, and has a different time course of action, than do traditional benzodiazepine compounds. Animal models incorporating this knowledge do show buspirone to be effective. Thus researchers must beware, when an established animal model does not demonstrate effectiveness of a novel compound, that the new compound should not be dismissed without further work. Such work is needed to avoid the risk of prematurely screening out potentially effective novel medications (e.g., Rodgers 1997).

Another use of pharmacological models involves testing theories of the physiological underpinnings of different disorders, rather than their treatment. Here researchers use various pharmacological agents to induce physiological and behavioral states in animals that resemble those seen in a human disorder. Caution is also needed here. Observing that a drug induces a state in animals resembling that in humans neither justifies the conclusion that the same system is nor-

mally involved in the human case, nor speaks to the issue of whether there may be a number of alternative routes to the human disorder (e.g., Weiss and Uhde 1990).

4. Animal Models of Anxiety and Anxiety Disorders

4.1 Specific and Social Phobias

Historically, research on animal models of anxiety—especially fears and phobias—began before models of other disorders. The role of classical conditioning (see *Classical Conditioning and Clinical Psychology*) in the etiology of specific phobias, first proposed by Watson and Rayner (1920) has been the subject of some controversy since about 1970. Some theorists ask how classical conditioning can play an important role both: (a) when some people recall no traumatic conditioning experiences; and (b) and when others who can recall traumatic experiences are not fearful/phobic. They have also wondered why some objects and situations are much more likely to become the objects of fears and phobias than others (e.g., phobias for snakes, spiders, and heights are far more common than for cars, guns, or knives, which may also be associated with trauma).

However, contemporary research with animal models illustrates that Pavlov's and Watson's core ideas about the role of classical conditioning were sound, but need to be expanded to incorporate the broader knowledge now available about the complexities of conditioning (see also *Classical Conditioning and Clinical Psychology*). For example, a primate model showed that conditioning sometimes occurs observationally or vicariously—that is, watching someone behave fearfully with, for example, a snake, may be sufficient to induce a fear of snakes in the observer (without any direct trauma occurring). Moreover, the role of conditioning (vicarious or direct) must be considered in light of various vulnerability and invulnerability factors that influence the outcome of a traumatic conditioning experience (cf. Mineka and Zinbarg 1996). For example, animal research by Pavlov and others laid the foundation for showing how individual differences in personality/temperament (such as levels of trait anxiety) affect conditioning and the likelihood of acquiring fears and phobias. Animal work also illustrates a wide range of experiential differences across individuals that strongly affect the outcome of direct or indirect conditioning experiences (and therefore why many *without* phobias will have such histories that involved putative conditioning events). For example, having extensive previous neutral or positive experiences with a potentially phobic object (e.g., a dog) can prevent the acquisition of dog phobia if the individual is later bitten by a dog. Moreover, a primate model showed that being reared with a strong sense of mastery and control over one's environment makes one less susceptible to the effects of the stressors that later may be involved in conditioning incidents.

Finally, there are evolutionarily based predispositions to acquire fears and phobias of certain objects or situations (e.g., snakes, water) that once posed a threat to our early ancestors more readily than other objects and situations not present in our early evolutionary history (e.g., guns, knives). Thus animal models have shown us that personality, experiential, and evolutionary factors may serve as diatheses or vulnerability factors for the development of phobias in certain individuals, given appropriate experiential input (cf., Mineka and Zinbarg 1996).

In social phobias, people have strong and persistent fears of various types of social interaction where they fear they may be evaluated and judged unfavorably. Again, traumatic conditioning experiences (direct or vicarious) are thought to play an important role, but with the same caveat as for specific phobias. That is, animal models demonstrate that personality, experiential, and evolutionary variables determine a person's level of vulnerability/invulnerability to developing social phobias. Considerable knowledge in this area stems from animal models of social anxiety (such as occur following defeats in physical fighting, when the animal typically becomes afraid of all dominant conspecifics rather than simply the one involved in the defeat) (cf. Mineka and Zinbarg 1996).

4.2 Panic Disorder with Agoraphobia

Individuals with panic disorder have recurrent unexpected panic attacks, usually associated with persistent anxiety about having another attack (anticipatory anxiety). Many also develop some degree of agoraphobic avoidance, learning to avoid situations in which they fear panicking. A full-fledged animal model of panic disorder remains elusive, although some pharmacological agents that provoke panic attacks in humans with the disorder, do seem to produce a panic-like state in some primates. Nevertheless, animal mini-models of panic and anxiety together have proved important in the development of a new theory of the origins of panic disorder that is largely based on contemporary principles of learning studied in animals (Bouton et al. 2001). The essence of this complex theory is that the occurrence of panic attacks sets the stage for the conditioning of anxiety to both internal and external cues that preceded a panic attack, thus explaining the development of both anticipatory anxiety and anxiety leading to agoraphobic avoidance. In addition, internal cues associated with the beginning of an attack can become conditioned to elicit panic attacks themselves. For example, a few heart palpitations that often occur early during an attack

could come to serve as conditional stimuli that trigger full-blown panic attacks (see *Classical Conditioning and Clinical Psychology*).

4.3 Post-traumatic Stress Disorder (PTSD)

This disorder develops in some individuals following exposure to a traumatic event in which the person experienced or witnessed events involving actual or threatened death to themselves or others. Symptoms include persistent re-experiencing of the event (e.g., through nightmares and flashbacks), persistent avoidance of stimuli associated with the trauma, and arousal symptoms such as difficulty concentrating and exaggerated startle responses. Recently, animal models of PTSD have been the focus of much research, providing useful insights into the nature of this disorder as well as its etiology and treatment. Animal models involve exposure to unpredictable and/or uncontrollable stress which initiates the emotional, behavioral, and physiological symptoms resembling those seen in human PTSD. The intense physical stressors used in animal studies resemble those in some forms of human traumatization associated with PTSD, including torture, abuse, and assault. This has in turn drawn attention to the role that perceptions of uncontrollability/unpredictablity play in the development of human PTSD symptoms and has led to studies testing these ideas.

This animal model has also shown the powerful role that various vulnerability and invulnerability factors play in determining who is more or less likely to develop PTSD, given exposure to the same trauma. For example, the animal model has revealed that prior exposure to uncontrollable stressors prior to the relevant trauma sensitizes the animal (or human), making it *more* likely to develop PTSD-like symptoms than animals without prior exposure to the uncontrollable stressors. Conversely, prior exposure to controllable stressors before the relevant trauma leads to an immunization effect, making it *less* likely the animal will develop PTSD-like symptoms. These hypotheses are beginning to be corroborated in human research. Finally, the idea that perceptions of uncontrollability and unpredictability mediate many aspects of PTSD has led to the formulation of hypotheses about the importance of re-instilling a sense of control and predictability as part of treatment. Such hypotheses are currently being tested (see Mineka and Zinbarg 1996).

5. Animal Models of Depression

Clinically significant depression is a surprisingly common condition. There are emotional, motivational, cognitive and somatic symptoms. Some of these can be modeled better in animals than others (the cognitive ones are especially difficult). Since the 1960s, various animal models of depression have produced useful insights into human depression, ranging from new ideas concerning etiology, prevention, and treatment. The models have differed in several ways, perhaps most notably in the methods used for causing 'depression' in the animals.

Some of the earliest and most striking work with primates used a social separation paradigm. Following on earlier work demonstrating that human infants undergoing prolonged separations from their mothers showed a biphasic response to the separation, Harlow and colleagues in the 1960s began to study this phenomenon in infant rhesus monkeys separated from their mothers. Both the monkey and human infants typically go through an initial state of intense agitation and distress (the protest phase—often seen as a prototype of anxiety), followed several days later (if the separation persists) by a phase of despair/depression, characterized by social withdrawal and rejection. Although some have argued this is at best a model of infant depression, others have argued that infant depression is in fact a prototype for adult human depression, with most of the prominent symptoms being functionally quite similar (except the cognitive ones).

By studying social separation in monkeys one can manipulate experimentally numerous variables, both before and during the separation, to test hypotheses about factors promoting minimal versus exaggerated responses to separation—something that obviously cannot be done in human infants. Research using animal models made it clear, for example, that having a sibling or alternate caregiver present during the separation from mother can attenuate (but usually not eliminate) the response to separation. But similar research with different species of monkeys, in which separated infants automatically get adopted by 'aunts,' also showed that reducing behavioral signs of distress is not tantamount to reducing physiological signs of distress and arousal which can remain high (e.g., Coe et al. 1985). Another line of primate work showed the importance of preseparation experiences in determining the outcome of any given separation. For example, infant monkeys of one species (bonnet macaques) whose mothers are relatively permissive, allowing the infants considerable freedom to interact with other adults, cope reasonably well with separations. By contrast, infants of another species (pigtail macaques) whose mothers are quite possessive and restrictive of their infant's freedom, do not cope as well with separations (e.g., Kaufman 1973). Observational follow-up work with human infants has often corroborated the hypotheses developed based on experimental animal models which are better able to pinpoint causal factors (see Mineka and Zinbarg 1991).

Another influential animal model of depression derives from the learned helplessness phenomenon and theory (e.g., Seligman 1975). In the late 1960s, Seligman and his colleagues Maier and Overmier noted that laboratory dogs initially exposed to uncontrollable shocks later showed major deficits in learning to control shock in different situations; indeed, they mostly seemed to accept the shock passively rather than trying to escape it. Yet animals first exposed to equal amounts of controllable shock showed no such deficits. Learned helplessness theory proposed that exposure to uncontrollable events leads one to learn that responses are ineffective in bringing relief, i.e., one is helpless to control important outcomes. The expectation of helplessness leads to: (a) cognitive deficits (difficulty learning to control shock), (b) motivational deficits (reduced incentive to try responding in other situations because of a belief that responses will be ineffective); (c) emotional changes (e.g., feelings of sadness, depression, and anxiety); and (d) physiological changes that occur with uncontrollable but not controllable stress. Seligman (1975) later proposed that the primary symptoms of depression resembled these primary changes seen with learned helplessness quite strongly.

Seligman also proposed etiological similarities. A large percentage of humans experiencing clinical depression have had one or more significant life stressors in the recent past (such as a major loss or unemployment, etc.). His argument was that all such precipitants could be seen as inducing a sense of lack or loss of control over important aspects of one's environment. Thus perceptions of helplessness may be the proximal cause of such cases of depression. Finally, Seligman also developed the corollary hypothesis that re-instilling a sense of control may be a core ingredient in the most effective treatments for depression (see Peterson et al. 1993).

This animal model of depression led to an enormous amount of animal and human research, much of it continuing today. Interestingly, although originally developed as a model of depression, its relevance has expanded to include a role in theories of several anxiety disorders (most notably PTSD, see above). For example, the initial emotional state during and following uncontrollable stress is one of intense and diffuse anxiety, which may lapse later into a depressive state. This idea is highly consistent with neurochemical results, indicating a shift from an intense aroused anxious state to a later depressed state, sometimes called conservation-withdrawal (e.g., Woodson et al. 1998). Interestingly, as noted above, anxious symptoms also precede depressive symptoms during primate and human infant separations, as well as following major losses in adults (Bowlby 1980). In addition, far more people who suffer depression at some point in their lives have suffered from an anxiety disorder first than the reverse (cf. Maser and Cloninger 1990). Thus both animal models of depression discussed here illustrate important features of the sequential relationship between anxious and depressed symptoms that also occurs in humans.

Other animal models of depression have also provided somewhat different but also partially overlapping information about both the symptom picture, and etiological factors in depression. The learned helplessness model has also generated hypotheses about effective treatment and prevention. Pharmaceutical companies make ample use of the most practical animal models to test their antidepressant compounds. Although recent etiological theories of human depression have generally now come to incorporate features that cannot be modeled in animals (such as feelings of hopelessness about the future), many useful insights from the original models remain generally intact.

6. *Other Animal Models*

Animal models of numerous other disorders have also been studied, including psychopathy, schizophrenia, and addictions. For example, regarding psychopathy, Newman (1997) and his colleagues used the well-known animal syndrome that stems from dysfunction in septo-hippocampal areas of the brain to model various features of psychopathic behavior—most notably the failure to inhibit inappropriate behaviors and inability to delay gratification. Fowles and Missel (1994) also summarized an important series of findings using animal and human mini-models, indicating the centrality of deficits in passive avoidance learning (learning what not to do to avoid punishment) to many important features of psychopathy. Both lines of work have led to new insights on psychopathy.

In schizophrenia, one early example of the use of animal models came when researchers first attempted to discovered how traditional antipsychotic medications work to reduce schizophrenic symptoms in humans. Animal research helped to show the important role that the neurotransmitter dopamine plays in the effects of these medications, and also helped lead to the development of the dopamine hypothesis of schizophrenia (including its etiology). This very influential theory, now known to be overly simplistic, was developed and tested in good part with animal mini-models.

Finally, the use of animal models to study of addictions has proved very important for understanding how addictive substances induce brain and behavioral changes. Using alcohol as an example, initially this work necessitated developing methods through which animals became addicted to alcohol (or other drug of interest). Subsequently, animal research with alcohol has revealed the numerous different neurotransmitter pathways in the brain with which alcohol interacts, leading to numerous alterations in brain function and behavior. Animal research has also facilitated understanding of brain mechanisms involv-

ed in maintaining the motivation and desire to drink. In addition, animal research focusing on environmental determinants of alcohol use, such as availability of alternative reinforcers, stress, and ease of access, etc., has provided important information. Some of this work has also focussed on genetic determinants of alcohol preference and alcohol use, and on how genetic factors interact with environmental factors (National Institute on Alcohol Abuse and Alcoholism 1997).

7. Conclusions

As indicated by this necessarily selective review of historical and contemporary research on animal models, such research has made important contributions to our understanding of various forms of human psychopathology and its treatment (see also *Classical Conditioning and Clinical Psychology*). Some of this research has attempted to develop full-fledged models by uncovering parallels between symptoms, etiology, treatment, and prevention in the animal model and the human disorder (e.g., the learned helplessness model of depression). However, an even greater amount of work exemplifies the mini-model approach in which only a subset of cardinal features of a human disorder are studied. An important advantage of studying animal models is that through experimental manipulation one can better determine what the critical causal features are than is generally possible in human research. With this background of success, further research with animal models certainly will continue to provide important new insights and information about many areas of human psychopathology and its treatment.

See also: Animal Rights in Research and Research Application; Anxiety and Anxiety Disorders; Anxiety Disorder in Children; Childhood Depression; Depression; Depression, Clinical Psychology of; Genes and Behavior: Animal Models; Panic Disorder; Pavlov, Ivan Petrovich (1849–1936); Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

Bibliography

Bouton M, Mineka S, Barlow D H 2001 A modern learning theory perspective on the etiology of panic disorder. *Psychology Review*, **108**: 4–32

Bowlby J 1980 *Attachment and Loss, III: Loss, Sadness and Depression*. Basic Books, New York

Coe C, Wiener S, Rosenberg L, Levine S 1985 Endocrine and immune responses to separation and maternal loss in nonhuman primates. In: Reite M, Field T (eds.) *The Psychobiology of Attachment and Separation*. Academic Press, New York

Darwin C R 1872 *Expression of Emotion in Man and Animals*. John Murray, London

Fowles D C, Missel K A 1994 Electrodermal hyporeactivity, motivation, and psychopathy: Theoretical issues. In: Fowles D C, Sutker P, Goodman S H (eds.) *Progress in Experimental Personality and Psychopathology Research*. Springer, New York

Kaufman I C 1973 Mother–infant separation in monkeys: An experimental model. In: Scott J P, Senay B (eds.) *Separation and Depression: Clinical and Research Aspects*. AAAS, Washington, DC, pp. 33–52

Maser J, Cloninger C (eds.) 1990 *Comorbidity in Anxiety and Mood Disorders*. American Psychiatric Press, Washington, DC

McKinney W T 1974 Animal models in psychiatry. *Perspectives in Biology and Medicine* **17**: 529–41

Mineka S, Kihlstrom J F 1978 Unpredictable and uncontrollable aversive events. *Journal of Abnormal Psychology* **87**: 256–71

Mineka S, Zinbarg R 1991 Animal models of psychopathology. In: Walker C E (ed.) *Clinical Psychology: Historical and Research Foundations*. Plenum Press, New York, pp. 51–86

Mineka S, Zinbarg R 1996 Conditioning and ethological models of anxiety disorders: Stress-in-dynamic-context anxiety models. In: Hope D (ed.) *Perspectives on Anxiety, Panic, and Fear*. 43rd Annual Nebraska Symposium on Motivation. University of Nebraska Press, Lincoln, NE, pp. 135–211

Newman J P 1997 Conceptual models of the nervous system: Implications for antisocial behavior. In: Stoff D M, Breiling J, Maser J D (eds.) *Handbook of Antisocial Behavior*. Wiley, New York, pp. 324–35

National Institute on Alcohol Abuse and Alcoholism (NIAAA) 1997 Ninth special report to the US Congress on *Alcohol and Health*. NIH Publication, Washington No. 97–4017

Pavlov I P 1927 *Conditioned Reflexes*. Oxford University Press, London

Peterson C, Maier S F, Seligman M E P 1993 *Learned Helplessness: A Theory for the Age of Personal Control*. Oxford University Press, New York

Rodgers R J 1997 Animal models of 'anxiety': Where next? *Behavioral Pharmacology* **8**: 477–96

Seligman M E P 1975 *Helplessness: On Depression, Development, and Death*. Freeman, San Francisco

Watson J B, Rayner R 1920 Conditioned emotional reactions. *Journal of Experimental Psychology* **3**: 1–14

Weiss S R, Uhde T W 1990 Animal models of anxiety. In: Ballenger J (ed.) *Neurobiology of Panic Disorder*, Wiley-Liss, New York pp. 3–27

Woodson J C, Minor T R, Job R F S 1998 Inhibition of adenosine deaminase by erythro-9-(2-hydroxy-3-nonyl) adenine (EHNA) mimics the effect of inescapable shock on escape learning in rats. *Behavioral Neuroscience* **112**: 399–409

S. Mineka

Clinical Psychology in Europe, History of

1. Introduction

Europe is characterized by its social, linguistic, and cultural diversity (e.g., Drenth et al. 1990, Poortinga 1996), which is rooted in very different histories and traditions of the 48 or so countries. Even the geo-

graphical or political extent of Europe is a matter for some debate, though it is assumed that the growth of the European Union (currently 15 states, all of them in western Europe) will clarify this in future years. Although psychology as a discipline originated with Wundt's laboratory in Leipzig, Germany, in 1879, the profession of clinical psychology was considerably slower to develop in Europe than in the USA (Sexton and Misiak 1976, Eysenck 1990, Sexton and Hogan 1992).

The major influence for the development of clinical psychology in many western European countries was the rebuilding of Europe after World War II in the 1940s and 1950s and the development of mental health facilities by European states, as the challenge of mental problems and mental health presented itself to health authorities (Lunt 2000). In eastern European countries clinical psychology developed much later as these countries emerged in the 1980s from political regimes which were critical of the discipline (e.g., Pawlik 1996). It is only relatively recently that these countries have begun to develop a profession of clinical psychology.

2. *The Field of Clinical Psychology*

Although there are differences between countries, it is possible to provide a general definition of the field which would be agreeable to all European countries; clinical psychologists may be defined as psychologists who

> apply psychology in a clinical context, usually a hospital, medical or community setting, with people (patients or staff) who consider themselves to be in need of a psychological perspective on their lives. In practice, the majority of clinical psychologists contribute to the assessment and treatment of people who see themselves as having psychological problems, such as those with mental health difficulties, but they also work with the handicapped, families, those with learning difficulties, and, more widely, with staff and organisations (Llewelyn 1994).

There are some differences in the practice of clinical psychology, from countries where the majority are employed by the state such as the United Kingdom where most clinical psychologists work for the National Health Service (NHS) and there is relatively little private practice, to countries such as Switzerland and Germany where many clinical psychologists work in private practice, with charges reimbursed by medical insurance companies. Furthermore, there are differences in the dominant activity of clinical psychologists: in many European countries the main activity is psychotherapy, whereas in others there is a broader role which includes assessment and other forms of intervention, and also a role in training and consulting with other staff. While the dominant paradigm informing clinical psychology in the United Kingdom is cognitive behavior therapy, other European countries such as France and Italy historically have been more influenced by psychoanalytic ideas. Although all European countries have been influenced by the USA, this influence has been perhaps most striking in the Scandinavian countries which embraced the 'scientist practitioner' approach to practice from an early stage.

2.1 *Europe*

As mentioned, the political face of Europe is developing and changing rapidly (see Lunt 1998). Many of the central and eastern European countries have applied for membership of the European Union, and are working towards criteria which will enable them to join. This will mean an expanded and changed Europe in the not too distant future, bringing together the early members of the European Union and nonmember states; many of the latter in Eastern Europe and formerly belonging to the Soviet Union. The countries of Europe constitute a very diverse group that may nevertheless be grouped into broad regions with some commonality in their organization and practice of clinical psychology. For these purposes, at a very broad level it is possible to identify a Nordic group of countries with considerable commonalities in their practice (the Scandinavian countries), Great Britain which has much in common with education and training in the USA and other parts of the English-speaking world, the German speaking countries, a southern European group, and an eastern European group. Of course, within these regions, the countries are characterized by their individuality and their diversity, and the different histories that have had a profound influence on the development of clinical psychology.

3. *The Early Years of Clinical Psychology*

There are a number of models in the field of clinical psychology, a field which emerged in European countries substantially after World War II (see above). It emerged as a recognizable profession at different times in different countries. In the United Kingdom the two World Wars provided a significant impetus to the emergence of this profession, initially through the need to develop psychological tests to recruit suitable personnel, and later in 1948 with the formation of the NHS that provided a considerable impetus to the development of this new profession. In these early days, the role of the clinical psychologist was largely one of laboratory technicians administering psychometric and other tests, usually for medical practitioners (Eysenck 1990). However, in the United Kingdom, the development of behavior therapy, in particular under the influence of Hans Eysenck who

established clinical psychology as a profession in England in 1949, led to clinical psychologists developing a therapeutic role, and by the 1960s they had become clinical practitioners in their own right.

In the Nordic countries, also, clinical psychology emerged substantially after World War II, in the 1940s and early 1950s; the Norwegian Psychological Association was founded in 1934, the Danish Psychological Association in 1947, the Swedish Psychological Association in 1955, and the Finnish Union of Psychologists in 1957. The formation of these Psychological Associations reflected the emergence in these countries of a profession of psychology, mainly clinical psychology, which progressed through similar broad major phases as the UK; that is, a phase focusing on diagnostic examination and testing of individual patients, a phase focusing on therapeutic work involving mainly psychotherapy, and a phase focusing on indirect work with other professional groups through techniques such as consultation and training.

In Spain, on the other hand, clinical psychology has had a shorter history (Belloch and Olabarria 1994) and is said to have emerged in the 1970s along with fundamental changes in Spanish society. Similarly in Italy, clinical psychology as a professional application emerged substantially in the 1960s and there remain some tensions over the autonomy and role of clinical psychologists and their relationship with medical practitioners, especially psychiatrists. The Association of Greek Psychologists, founded in 1963, reflects the existence of professional psychology in Greece, although at that time and until very recently, all clinical psychologists received their training overseas, and the profession was very much dominated by the medical profession.

As mentioned above, clinical psychology emerged much later in Eastern European countries, and it was not until the 1980s that psychotherapy was fully recognized as a profession (Pawlik 1996). Until recently, nearly all psychologists were employed in state institutions, though in recent years there has been a growth in demand which has resulted in a growth in private practice.

In all European countries, clinical psychologists represent the largest group of psychologists, and experienced a rapid growth in their number between the 1960s and the 1990s with the expansion of mental health provision and the growing awareness of the contribution of clinical psychology to a wide and diverse range of areas of work (EFPPA 1997). Indeed, the 31 Member Associations of EFPPA (see below) represent 150,000 psychologists, the majority of whom are clinical psychologists.

Since the 1990s, in many European countries the emergence of the specialty of health psychology, with a focus on prevention rather than treatment, and the promotion of health rather than a more therapeutic or curative function, has led to attempts to define and differentiate a new field of psychological activity within the mental health field. This has been supported by the WHO commitment to Health for All by the Year 2000; EFPPA has had three Task Forces, focusing respectively on clinical psychology, health psychology, and psychotherapy, which have drawn up a model defining overlap and separate areas of activity within the health field, with corresponding commonalities and differences in education and training (EFPPA 1997).

4. *Training of Clinical Psychologists*

Education and training of clinical psychologists varies in European countries though, again, there are major regional groupings. Across Europe, as in other regions of the world, the education and training period has increased, with moves in some European countries for doctorate training for clinical psychologists, and more demanding requirements in all countries for the internship period. In the United Kingdom, there is a strong commitment to a scientist–practitioner model: 'the clinical psychologist is first and foremost an "applied scientist" or "scientist–practitioner" who seeks to use scientific knowledge to a beneficial end' (Marzillier and Hall 1990). This commitment also characterizes the Nordic countries which have been influenced substantially by the USA, and also by the UK. These countries would sign up to the definition that:

> clinical psychologists share several common attributes. They are psychologists because they have been trained to use the guidelines of knowledge of psychology in their professional work. They are clinicians because they attempt to understand people in their natural complexity and in their continuous adaptive transformations … they are scientists because they utilize the scientific methods to achieve objectivity and precision in their professional work. Finally, they are professionals because they render important human services by helping individuals, social groups, and communities to solve psychological problems and improve the quality of life (Kendall and Norton-Ford 1982, p. 4).

This model is also espoused by Spanish clinical psychology where Belloch and Olabarria (1994) state that clinical psychology training is 'very similar to that proposed in the 1940s, in the famous Boulder Conference, organised by the APA.' In France, where there is a strong clinical psychoanalytic tradition, there is less of a commitment to the 'Boulder' model, and more of a philosophical or hermeneutic tradition in relation to education and practice.

However, since 1957 when the Treaty of Rome provided the foundation for the European Community (later Union), there have been requirements on individual countries ('states') to provide procedures for the mutual recognition of psychologists' qualifications across national boundaries (see McPherson 1988). Wider moves within the European Union such

as the Bologna Agreement, which was signed by 29 Ministers in 1999 and commits them to greater convergence in terms of university degree structures, mean that even within Europe there is likely to be greater similarity in terms of structures of education and training for clinical psychologists.

5. *Professionalization of Clinical Psychology*

The period since the 1950s has seen a greater professionalization of clinical psychology in all European countries, with the development of codes of ethics (Lindsay 1996) and increased regulation and laws for clinical psychologists (Lunt 2000) across European countries. These political and professional developments have been supported by EFPPA which has brought together clinical psychologists from all over Europe to work on the professional aspects of practice at a European level and to support individual European countries seeking to develop their ethical codes, laws protecting the title of psychologist, and education and training in clinical psychology.

6. *Organization of Clinical Psychology*

It is also possible to trace the history of clinical psychology through its organization in Europe (Gilgen and Gilgen 1987, Lunt 1998). Many European countries founded scientific societies for psychologists in the early twentieth century whose purpose was to foster research and psychological science. At the time of World War II, separate professional associations to meet the needs of professional psychologists, often mainly clinical psychologists, were founded in a number of European countries. These associations had, as a focus, issues concerning professional practice and emerged in some countries as Trade Unions, negotiating salaries and terms and conditions of work for clinical psychologists as well as wider professional issues such as regulation, legislation, and ethical codes. In 1981, at a time when the provisions of the European Community demanded that European member states encourage mobility of professionals across Europe, the European Federation of Professional Psychologists Associations was formed to bring together professional associations in Europe and to collaborate on matters of common professional concern.

EFPPA currently has 31 member associations representing all the countries of the EU, all other countries in western Europe, and a growing number of member associations from central and eastern Europe. The federation provides a unique opportunity for comparison between the practices of different European countries and a forum for discussion and debate of important issues. The formation of EFPPA in 1981, when matters of mobility and mutual recognition were becoming more pressing, was due to a realization by psychologists and psychology associations of member states that a federation would provide a professionally and politically useful way to move forward and to begin to develop common policies in this area. Clinical psychologists, in particular, were faced with the growing prospects of mobility between countries in the European Union, and the implications of the Treaty of Rome that provided the foundation for the European Community in 1957.

As a federation of Professional Psychology Associations representing around 150,000 professional psychologists in Europe, EFPPA spends most of its efforts on clinical psychology and clinical psychologists, also the largest group of psychologists within Europe (as in the rest of the world). In many countries psychology has become one of, if not the, most popular subjects to study at university. The majority of students studying psychology aspire to become clinical psychologists, and for this reason, in many European countries, there is an oversupply of qualified practitioners. Many countries now operate a so-called *numerus clausus*, either at the start of the psychology study, or during the study. This controls the numbers in training. In the United Kingdom, where specialist training in clinical psychology is funded by the NHS—where the vast majority of clinical psychologists work— the number of 'trainee' posts is strictly limited and is planned according to staffing needs in the different regions of the country. In other countries, where there is a tradition of predominantly private practice in clinical psychology, there are large numbers of qualified psychologists unable to find work. In all European countries, the ratio of female to male students is between 6:1 and 3:1, leaving the profession in danger of becoming an almost feminized profession in the future (Schorr and Saari 1995). There has also been some difficulty in many countries in recruiting students from the range of ethnic groups represented in Europe's increasingly multiethnic population. This clearly has implications for the clinical treatment of different client groups.

7. *The Future*

As new fields of practice in the health field evolve, in particular health psychology, there are pressures on clinical psychology. For example, in some countries it has been said that there may no longer be a field of clinical psychology, since there are strong moves towards a broader field of health psychology and a greater focus on preventive work. These newer areas, such as health psychology, neuropsychology, and forensic psychology, are leading to greater specialization within and outside clinical psychology. Although European countries differ in the extent and nature of their specialisms within the health field, there is an increasing trend for specialization and demands for higher qualifications. In one respect, psychologists

working within the health system could be said to be becoming more generic, while on the other hand there are increasing specializations within this field of work.

8. Summary

The brief 50-year history of clinical psychology in Europe has seen an enormous increase in numbers both of students and of practitioners, such that the majority of psychologists are now clinical psychologists. This rapid professionalization has been accompanied by higher qualifications, greater regulation, the development of ethical codes, and all the characteristics of traditional professions. However, as the number of clinical psychologists increases, and the question of mobility across Europe becomes more pressing, there will be increasing attempts to develop more common frameworks and standards for education and practice; the challenge will be to achieve a balance between allowing individual countries their own autonomy which reflects their differing history and culture ('subsidiarity' as it is called), and developing more common agreed frameworks of practice which reflect a possible future 'federalization' of Europe.

See also: Clinical Psychology in North America, History of; Freud, Sigmund (1856–1939); Psychiatry, History of; Psychoanalysis, History of; Psychoanalysis in Clinical Psychology; Psychotherapy: Ethical Issues; Psychotherapy, History of: Psychiatric Aspects; Training in Clinical Psychology in the United States: Practitioner Model; Training in Clinical Psychology in the United States: Scientist–Practitioner Model

Bibliography

Belloch A, Olabarria B 1994 Clinical psychology: Current status and future prospects. *Applied Psychology: An International Review* **43**(2): 193–211

Drenth P J D, Sergeant J A, Takens R J 1990 *European Perspectives in Psychology*, vol. 2. Wiley, Chichester, UK

EFPPA 1997 Report to the General Assembly of EFPPA of Clinical Psychology Task Force. Available from EFPPA secretariat

Eysenck H 1990 Clinical psychology in Europe and in the US: Development and future. In: Drenth P J D, Sergeant J A, Takens R J (eds.) *European Perspectives in Psychology*. Wiley, Chichester, UK, Vol 2

Gilgen A R, Gilgen C K (eds.) 1987 *International Handbook of Psychology*. Greenwood Press, New York

Kendall P C, Norton-Ford J D 1982 *Clinical Psychology. Scientific and Professional Dimensions*. Wiley, New York

Llewelyn S P 1994 Assessment and therapy in clinical psychology. In: Spurgeon P, Davies R, Chapman A (eds.) *Elements of Applied Psychology*. Harwood Academic Publishers, Chur, Switzerland

Lindsay G 1996 Developing an ethical psychological practice. In: Georgas J, Manthouli M, Besevegis E, Kokkevi A (eds.) *Contemporary Psychology in Europe*. Hogrefe & Huber, Göttingen

Lunt I 1998 Psychology in Europe: challenges and opportunities. *The European Psychologist* **3**(2): 93–101

Lunt I 2000 Psychology as a profession. In: Pawlik K, Rosenzweig M (eds.) *The International Handbook of Psychology*. Sage, London

Marzillier J, Hall J 1990 *What is Clinical Psychology?*, 2nd edn. Oxford University Press, Oxford, UK

McPherson F 1988 Psychologists and the EEC. *The Psychologist* **9**: 353–5

Pawlik J 1996 The situation of the psychologists in Eastern European countries today. In: Georgas J, Manthouli M, Besevegis E, Kokkevi A (eds.) *Contemporary Psychology in Europe*. Hogrefe & Huber, Göttingen

Poortinga Y 1996 Cultural diversity in Europe: Extrapolations from cross-cultural research for professional psychology. In: Georgas J, Manthouli M, Besevegis E, Kokkevi A (eds.) *Contemporary Psychology in Europe*. Hogrefe & Huber, Göttingen

Pilgrim D, Treacher A 1992 *Clinical Psychology Observed*. Routledge, London

Schorr A, Saari S (eds.) 1995 *Psychology in Europe*. Hogrefe, Göttingen

Sexton V S, Hogan J D (eds.) 1992 *International Psychology. Views from around the World*. University of Nebraska Press, Lincoln & London

Sexton V S, Misiak H (eds.) 1976 *Psychology around the World*. Brooks/Cole, Monterey, CA

I. Lunt

Clinical Psychology in North America, History of

Clinical psychology is concerned primarily with the study of psychopathology and with its diagnosis and treatment. It shares this domain with several other mental health disciplines, including psychiatry, social work, nursing, and various types of counseling. Compared to these other disciplines, clinical psychology is distinctive for its training in research and for its expertise in psychometrics and the behavior therapies. North America played an important role in the emergence of clinical psychology. The field usually dates its origin from the founding of the first psychology clinic in 1896 by Lightner Witmer (1867–1956) at the University of Pennsylvania (Routh 1996). Clinical psychology in the English-speaking parts of Canada developed in a pattern similar to that seen in the US (with doctoral training required for independent practice) but somewhat later in time. The field developed in French-speaking Canada and in Mexico in a way more resembling that of European countries, with master's or licenciate-level training required for independent practice. The North American Free Trade Agreement (NAFTA) now exerts pressure on all three countries and their states and provinces to

coordinate these differences to a greater degree, to permit more freedom of movement to qualified clinical psychologists.

1. The Prehistory of the Mental Health Field

The need for humans to deal with the problems now called mental illness did not emerge suddenly a century ago. It seems reasonable to assume that such problems have existed in some form in every society through all the millenia of human experience. The ancient literatures of India, Egypt, China, Greece, and Rome contain descriptions of disturbed behavior, often interpreted in religious terms as some type of retribution by magical or divine forces. Legal systems as they developed in all of these civilizations necessarily included provisions for seeing to the management of the affairs and property of persons who were temporarily or permanently unable to manage for themselves (Routh 1998).

2. Greek Ideas Concerning Psychopathology

Western concepts of psychopathology have their roots in those of the ancient Greeks, including the writings attributed to the physician Hippocrates (460–377 BC). These Hippocratic writings include terms such as melancholia, mania, paranoia, and dementia, with meanings not all that different from their present ones, albeit with different explanations. For example, in Greek, the word 'melancholia' simply means 'black bile.' In the Hippocratic theory of the humors, a person suffering from severe mental depression had an overabundance of black bile, a substance thought to be produced by the spleen. Within this system, one aspect of treatment quite reasonably aimed to reduce the amount of black bile by administering a purgative such as hellebore. Another disorder was that of 'phrenitis,' literally meaning an inflammation of the mind. This referred to mental disturbance accompanied by fever, and the approach taken was simply to wait for the fever to abate. Though some of these Hippocratic ideas may seem strange to us now, it is worth reflecting why some of them lasted well into the eighteenth century and beyond.

3. Emergence of Psychiatry

Although according to Herodotus (ca. 484–425 BC), medical specialties existed even in ancient Greece, the one we now call psychiatry did not emerge until the late eighteenth century in Europe. The ancients did not conceptualize mental disorders as a separate category but regarded them as being illnesses like any other. When psychiatry did emerge, with the work of such pioneers as Philippe Pinel (1745–1826), Benjamin Rush (1745–1823), and Vincenzo Chiarugi (1759–1820), it was associated with the development of mental asylums or hospitals as a separate locations for the care of those with mental derangements. The theory of 'moral treatment' that was typical of that time tried to minimize the use of coercive methods such as chaining patients to restrain them and instead insisted that they be treated with kindness and courtesy. It was often found that even some very disturbed patients responded positively to such a regimen.

4. Modern Psychology and the Study of Psychopathology

Long before a formal discipline of psychology existed, people in every society still no doubt reflected upon human experience and behavior. As was the case of Hippocrates in relation to medicine and psychiatry, the influence of ancient Greek philosophers such as Plato (427–347 BC) and Aristotle (384–322 BC) upon our present psychological concepts was pervasive. There is conventional agreement that psychology emerged as a formal academic discipline only in the mid-nineteenth century in Europe. Wilhelm Wundt (1832–1920) of the University of Leipzig is usually named as the founder of the field, and 1879, the year in which he set up his psychology laboratory there, is celebrated as the key event in its origin. Wundt's work and those of the other early psychologists often focused on sensory processes, reaction time, and memory. It is also noteworthy that the study of psychopathology was a possible topic of psychological study even in those days. The eminent psychiatrist Emil Kraepelin (1856–1926) was influenced by Wundt's writings and was interested in psychology. His main motive for becoming a psychiatrist was that this was the only way he could see to make a living while doing psychological research. He later actually studied under Wundt, who encouraged him to keep on with his work combining psychology and psychiatry. Kraepelin set up psychology laboratories at his psychiatric clinics in both Heidelberg and Munich.

Many of the pioneers in psychology in both Europe and the US were trained as medical doctors. In France these included Theodule Ribot (1839–1916), who wrote about diseases of memory and about personality disorders. An important French colleague was Pierre Janet (1859–1947), who studied anxiety, hysteria, and obsessions and developed concepts of dissociation that continue to be influential today. In the US the leading pioneer in psychology, William James (1842–1910) was originally trained in medicine but wrote a psychology textbook that proved to be the most influential of all. In 1896, James gave his Lowell lectures on exceptional mental states, much influenced by the work of Janet. Boston neurologist Morton Prince also became interested in Janet's writings and published a description of a woman with multiple

personalities who had been his patient. Prince established the *Journal of Abnormal Psychology* in 1906 and later gave it to the American Psychological Association. Subsequently, in 1926, he established the Harvard Psychological Clinic, which was a research facility rather than one delivering mental health services. The most influential medically trained student of psychology of this time was no doubt Sigmund Freud (1856–1939). Breuer and Freud's book *Studies in Hysteria*, was published in 1895 and Freud's book on the interpretation of dreams in 1900. The first international psychoanalytic meeting was held in Salzburg in 1908. In 1909, Freud came to the US for the first and only time.

5. *Lightner Witmer and Clinical Psychology*

As the above paragraphs make clear, Witmer was hardly the first to suggest that psychologists study psychopathology. Instead, his main contribution was to go beyond that to advocate that psychologists try to help people as well as study them. Witmer had been an undergraduate at the University of Pennsylvania and then for a time, before going to Leipzig to obtain his Ph.D. under Wundt, served as a school teacher. He had as a student a young man with marked difficulty in reading and was able to help the youngster succeed in school and go on to attend college. This turned out to be a formative experience for Witmer. After Witmer had obtained his Ph.D. and returned to his *alma mater* as a psychology professor, a school teacher named Margaret Maguire asked his advice about one of her pupils with a spelling problem. Witmer reasoned that if psychology was of any practical use, it should be able to be of help in a case of this kind. Thus was the psychological clinic and the field of clinical psychology launched.

Witmer's clinic worked more with children than with adults and tended to concentrate on academic difficulties such as reading, spelling, or general backwardness in school as opposed to emotional or behavioral problems. His historic forebears are thus not Hippocrates and Pinel but rather eighteenth and nineteenth century French physicians and special educators such as Jacob Pereire (1715–1780) (who taught deaf-mutes to speak), J. M. G. Itard (1775–1838) (who worked with the 'wild boy' of Aveyron), and Edouard Seguin (1812–1880) (a physician who devised a 'physiological method' of sensory and motor training in an attempt to remediate mental retardation). Witmer used existing laboratory procedures including the Seguin formboard and sensory/motor procedures adapted from Wundt to evaluate the children referred to him and often tried to teach them simple tasks as a part of his diagnostic efforts. In his treatment activities he often collaborated with school teachers, as well as with physicians, thus serving as more of a consultant than doing anything resembling present-day psychotherapy. As a matter of fact, he was little influenced by the activities of the Boston School of psychotherapy that was contemporary with his work, nor later by Freud and his psychoanalytic movement.

Witmer is not remembered for any noteworthy scientific discoveries but rather for his persistence in enacting this new role of the clinical psychologist. At the University of Pennsylvania's Ph.D. program, he essentially trained most of the first generation of clinical psychologists. He maintained his clinic as a service and training facility and in 1907 began a journal, the *Psychological Clinic*, to publicize these activities (Witmer 1907).

6. *The Binet Test*

There is still no consensus among psychologists as to precisely how to interpret its findings. Still, Alfred Binet's 'metric scale' of intelligence (Binet and Simon 1905) may be the most noteworthy piece of technology developed by psychology in its first century. Certainly it had a major impact on the new field of clinical psychology. In fact, before World War II, probably the most characteristic activity of the typical clinical psychologist was the administration of the Binet test and other similar measures (Routh 1994). This was true despite the fact that Lightner Witmer, the founder of the field, was quite critical of the Binet test and used it only as one part of his extensive battery of laboratory procedures.

In the light of how influential his test was, it is interesting to note that Alfred Binet himself was not particularly identified with the field of clinical psychology. Originally trained as a lawyer, Binet became part of the circle around the influential neurologist Jean Charcot at the Salpetriere in Paris. Much of his psychology was self-taught, through extensive reading at the *Bibliothèque Nationale*. In France, Binet became known as one of the founders of the entire field of psychology (often characterized worldwide as 'experimental psychology') and edited the influential journal, *Annēe psychologique*. As is well known, Binet's successful attempts to devise an intelligence test departed from the conventional approach of using relatively 'pure' sensory and motor tasks to use complex worksamples of the kinds of things schoolchildren might be expected to know and do.

For some reason, Binet's new test did not create as much of a stir in his homeland as it did in the US. Psychologist Henry Goddard, who directed the psychology laboratory at the Vineland Training School, in New Jersey, had the Binet test translated and soon confirmed its impressive validity in identifying persons with mental retardation. The use of the Binet spread like wildfire among the early clinical psychologists in the US, beginning with those employed in the field of mental retardation. Goddard founded the first psy-

chology internship in 1908 at Vineland, NJ. Goddard went on to become the first professor of clinical psychology at Ohio State University, like the University of Pennsylvania an important early training center in the field (Routh 1994).

Lewis M. Terman (1877–1956) at Stanford University developed a standardized version of Binet's test, collected normative data for it, and introduced certain refinements such as the ratio IQ score (originally suggested by Wilhelm Stern of Hamburg). The 1916 Stanford–Binet, as it was called, dominated this field for many years. At about the same time, 1915, Robert Yerkes pointed out the unsuitability of the concept of mental age and of this testing format for use with adults and introduced his own 'point scale' as a substitute for it. Yerkes and his colleagues were also responsible for the development of group intelligence tests, the Army Alpha (for those who could read) and Army Beta (for the illiterate), used for mass testing of military recruits during World War I. Another wartime development was Robert S. Woodworth's Personal Data Sheet published in 1917. This was one of the first rationally developed self-report questionnaires intended to detect neurotic tendencies (Routh 1994).

7. *The Child Guidance Center Movement*

The development of child guidance centers was another factor that influenced early clinical psychologists in the direction of working with children more than with adults. The first child guidance clinic was the Institute of Juvenile Research, established in 1909 by physician William Healy in conjunction with the juvenile court of Chicago. The idea behind such facilities was that careful clinical study of children engaging in antisocial activities could assist in guiding them away from crime. Healy was joined at first by clinical psychologist Grace Fernald and subsequently by her replacement, Augusta Bronner (Healy and Bronner 1926). The child guidance clinic was the origin of the 'clinical team' of psychiatrist, psychologist, and social worker that later spread to other settings. The typical pattern was that the psychiatrist saw the child, the social worker saw the family, and the psychologist did the testing. The child guidance movement was supported by the Harkness family's philanthropy in the form of the Commonwealth Fund and replicated in many US cities and abroad.

8. *The First Clinical Psychology Organization*

In 1917, in Pittsburgh, a group of eight clinical psychologists organized themselves into what they called the American Association of Clinical Psychologists (AACP) and invited 48 colleagues to join them (Routh 1994). They were led by J. E. W. Wallin (1876–1969) and Leta Hollingworth (1886–1939). Incidentally, Hollingworth was the first to suggest in 1918 that a person trained in clinical psychology receive a distinctive type of degree, the doctor of psychology. The new organization was viewed by many as divisive and was soon incorporated into the American Psychological Association as its Clinical section. An effort by the same group to introduce procedures for certifying qualified clinical psychologists failed, however.

9. *Psychometric Developments*

The interwar years were a fertile time for the emergence of various new psychometric procedures, many of which continue to be in use today. For example, in 1921, the Swiss psychiatrist Herman Rorschach (1884–1922) published his well-known inkblot test. It was brought to the US by a child psychiatrist who taught it to a clinical psychology graduate student at Columbia named Samuel Beck (1896–1976). Beck then proceeded to do his dissertation on this new test and eventually to develop his own system for administering and scoring it. Psychologist Bruno Klopfer (1900–1971), a disciple of Carl Jung (1875–1961), also introduced the Rorschach to the US and developed a separate system for administering and scoring it (see *Projective Methods in Psychology*). In 1936 the Thematic Apperception Test was introduced by Henry A. Murray (1893–1988) of the Harvard Psychological Clinic, and a colleague. Also in 1935 Edgar A. Doll (1889–1969), introduced the Vineland Social Maturity Scale, an interview-based method involving informants familiar with the person, for assessing the social competence of individuals suspected of mental retardation. David Wechsler (1896–1981) published the original version of his Wechsler–Bellevue intelligence test for adults. This was but the first of many Wechsler tests of intelligence and memory. It introduced the use of the deviation IQ, a standard score comparing the individual to age-matched normative subjects. In 1943, psychologist Starke R. Hathaway and psychiatrist J. C. McKinley introduced the first edition of the Minnesota Multiphasic Personality Inventory (MMPI). The MMPI had novel 'validity' indicators, and its measures of psychopathology were empirically keyed to psychiatrically defined groups (see *Minnesota Multiphasic Personality Inventory (MMPI)*).

10. *Organizational Activities*

In 1937 a new organization known as the American Association of Applied Psychology (AAAP) split off from the American Psychological Association (APA) to provide a home for various professionally oriented groups including the clinical psychologists, who at this same time dissolved the Clinical Section of the APA.

The AAAP began to publish the *Journal of Consulting Psychology*, which subsequently developed into a high-prestige clinical psychology journal (Routh 1994).

It was also at this time that some preliminary developments in psychology began in other parts of North America. In 1937, for example, the first psychology curriculum was devised at UNAM, the National Autonomous University of Mexico, in Mexico City. In 1939, the Canadian Psychological Association was founded. It had 38 members to begin with, and it has been estimated that there were only 53 psychologists in all of Canada at the time. Needless to say, there were few prewar developments in Canada or Mexico specifically relating to clinical psychology.

11. The Post-World War II Boom in US Clinical Psychology

Clinical psychology expanded so greatly in the US after World War II that many brief historical accounts of the field even consider its development to have begun at that time. The war effort tended to draw everyone into it, either on the battlefield or on the home front. Many psychologists whose interests prior to the war had been strictly in research and in the academic side of the field found themselves assigned to carry out psychological testing or to help medical staff in treating psychiatric casualties. After the war, it was clear that the Veterans Administration (VA) would have to be vastly expanded to deal with the need for residential care, psychotherapy, or at least vocational counseling of some of those returning from military service.

In 1945, the VA and the newly established National Institute of Mental Health in the US came to the APA to ask it to establish a system for accrediting training programs in clinical psychology. The government intended to pour millions of dollars into training such individuals and needed to know which programs were competent to carry this out. In response, APA created a system of accreditation, and for the first time, it began to be possible to say who was a well trained clinical psychologist and who was not. David Shakow (1901–1981) was the architect of the 1949 conference held in Boulder, Colorado, which ratified what has come to be called the 'scientist–practitioner' model of training clinical psychologists (Raimy 1950) (see *Training in Clinical Psychology in the United States: Scientist–Practitioner Model*).

Postwar clinical psychologists continued in their role as mental testers, but gave more emphasis to the assessment of personality and psychopathology, not just cognitive status. This was the heyday of projective tests, and at least for a time the Rorschach inkblot was an appropriate symbol for the clinical practitioner of psychology. A well known exemplar of the clinical psychologist as projective tester in this era was David Rapaport (1911–1960), chief psychologist at the Menninger Clinic in Kansas. A two volume set of books published at this time by Rapaport and coworkers (Rapaport et al. 1945, 1946) established the Rorschach, the TAT, and the Wechsler test to be a 'full test battery' for almost a half century to come.

The clinical psychologists of this era were also eager to become full-fledged psychotherapists as well as mental testers. Their route to therapeutic training was blocked to some extent by the American Psychoanalytic Association's 1938 policy decision (contrary to Freud's own views) that only psychiatrists were to be trained in psychoanalysis. Clinical psychologists thus became very ingenious in devising ways of becoming therapists. The best known of them was perhaps Carl Rogers (1902–1987), whose original therapy supervisor was the social worker Jessie Taft. Taft, in turn, had received her training from Otto Rank, a psychologist from Vienna who had received orthodox psychoanalytic training there and had been a close colleague of Freud's. Rogers was successfully assertive in other ways. At one point in his career he was director of a child guidance clinic when such administrative positions were supposed to be held only by physicians. Rogers also was determined to combine his psychology training with his role as a therapist. He was among the first to produce recordings of actual psychotherapy sessions, and was a pioneer in doing controlled research on the outcome of psychotherapy. Rogerian therapy ('client centered' or 'person centered,' therapy, as it was later called) is still practiced and studied both in North America and elsewhere (Routh 1994, 1998) (see Person-centered Therapy).

In 1945, the first state law certifying psychologists for independent practice was passed by Connecticut. By 1977, all states in the US had passed such certification or licensing laws regulating the use of the title, 'psychologist' or the practice of psychology (Routh 1994).

12. The Behavior Therapy Movement

Some psychologists were of the opinion that clinical psychologists in their professional activities should not simply try to duplicate the activities of psychiatrists. In 1913, John Watson had boldly proclaimed a behavioristic approach to psychology, which was widely influential at least in academic psychology in the US. In a famous paper reporting research carried out under Watson's supervision, Jones (1924) described the case of 'Peter,' whose fear of rabbits she desensitized. Not even Jones herself realized the wider implications of this study at the time, but in the light of events many years later some considered her to have been 'the mother of behavior therapy' (see *Behavior Therapy: Psychological Perspectives*).

The behavior therapy movement progressed not only by following its own agenda but by attacking its

opponents. Psychologist Eysenck (1952) in England thus skeptically reviewed the evidence for the effectiveness of psychotherapy. It was not enough, he noted, to show simply that psychotherapy patients improved. One also needed to consider the rate of spontaneous improvement of patients who did not receive psychotherapy.

In 1962 in Charlottesville, Virginia, a behavior therapy conference was held, sponsored by psychiatrist Joseph Wolpe and psychologists Andrew Salter and Leo Reyna. At the time Wolpe had just published a book on his success in treating patients with phobias using the behavioral method of systematic desensitization. Soon afterward, Lang and Lazovik (1963) published the first controlled study of desensitization, in treating snake phobia. Soon the behavior therapy movement was in full swing, with its own organizations, journals, and many adherents. Sidney Bijou and his colleagues established behavioral treatments based on the research of B. F. Skinner. These came to be known as applied behavior analysis and were especially influential in work with the behavior disorders of children and of those with mental retardation (see *Behavior Analysis, Applied*).

13. Canadian Clinical Psychology

It was in the 1960s that clinical psychology finally came of age in Canada. It was and is the largest applied specialty in psychology numerically in that country, as it is in the rest of the world. In 1965, the Couchiching Conference basically endorsed the 'Boulder model' of scientist–practitioner training for clinical psychology. Some Canadian doctoral programs such as the one at McGill even sought accreditation by the American Psychological Association. There was even something of a boom north of the border. It is said that by 1966, more than half the doctoral psychologists in Canada were either American born or trained in the US. In 1983, the Canadian Psychological Association established its own program of accreditation. The first CPA-accredited doctoral programs in clinical psychology were those at McGill, Concordia, and Simon Fraser Universities. The success of doctoral training in clinical psychology tells only part of the story there, however. In Quebec and some other eastern provinces, the master's degree was accepted as the entry level of training for independent clinical practice. By 1996, Canada had 88 graduate training programs in professional psychology (including clinical): 57 doctoral and 31 terminal masters.

Canadian clinical psychology is noted for particular strength in the area of neuropsychology, which built on Canadian strengths in the neurosciences, including the work of neurosurgeon Wilder Penfield at the Montreal Neurological Institute. In academic psychology, the research and writings of Donald Hebb concerning the CNS (conceptual nervous system) were influential. It was at McGill University that Olds and Milner published their famous 1954 paper on the reinforcement of an animal's behavior by electrical stimulation of its brain. On the clinical side, Ronald Melzack elaborated his theory of gating mechanisms influencing the experience of pain. Brenda Milner explored the role of the hippocampus in semantic memory, including work with her famous patient, 'H.M.' This man developed permanent memory deficits after surgery inadvertently destroyed his hippocampus bilaterally. Doreen Kimura documented the left cerebral hemisphere advantage in dichotic listening.

14. Clinical Psychology in Mexico

In 1997, the Division of Clinical Psychology of the APA held its midwinter board meeting in Mexico City, hosted by Juan Jose Sanchez-Sosa, the director of the school of psychology at UNAM, the National Autonomous University of Mexico. Sanchez-Sosa provided his US colleagues with a tour of this school, itself as large as many an entire college campus in the US. The school offers both Master's and Ph.D. degrees in psychology, but as in much of Europe, these degrees are intended for those headed for academic and research careers. Those who intend to practice psychology, including clinical psychology, in Mexico need only a 'licentiate' or diploma to do so, which is awarded after 6 years of what to persons trained in the US seems to be undergraduate training. But students in such a program spend essentially full time on psychology, without the need for a broad liberal arts distribution of courses. This includes a significant amount of practicum experience. The psychology clinic, one of the practicum facilities used at UNAM, features a variety of clinical activities, including psychological testing, psychodynamic therapy, behavior therapy, group therapy, and even biofeedback. Since there is no certification or licensing system beyond the licentiate degree itself, it is difficult to be sure how many of these graduates are practicing clinical psychology in a way parallel to what would be seen in the US or Canada.

15. Independent Practice of Clinical Psychology

Although Lightner Witmer in his clinical work often collaborated with school teachers, physicians, or others, psychologists working in his clinic were never supervised by members of any other profession. This tradition of independent work has continued within the field of clinical psychology, somewhat in contrast with social work and nursing. The post-World War II expansion of the field in the US was primarily in the

public sector, typically VA hospitals, but also child guidance centers and eventually community mental health centers. The large government training grants supporting clinical psychology programs at the time presupposed that the graduates would go to such public sector jobs or teach in colleges and universities. In the 1980s, the Reagan administration made most such training grants a thing of the past (Routh 1994).

David Mitchell, a Ph.D. student of Lightner Witmer, was one of the first individuals to make his living primarily in the private practice of psychology. Eventually, he was joined by many others. After all, the state and provincial laws that developed to regulate psychology after 1945 specified what qualifications were necessary to offer one's services to the public as a psychologist. Many psychologists trained in Boulder-model Ph.D. programs did no research after graduation, and eventually the idea of training psychologists as practitioners rather than scientist–practitioners emerged. Beginning with the University of Illinois in 1966, a number of programs began to offer the doctor of psychology (Psy.D.) degree rather than the Ph.D. The conference in Vail, Colorado in 1973 officially legitimized such practitioner training for the first time. In fact, a number of nonuniversity-affiliated schools of professional psychology sprang up beginning in the 1970s, many of them offering Psy.D. degrees. Often these programs were supported only by student tuition, and many students assumed substantial loans to finance their education (Routh 1994). No such private school of professional psychology has emerged in Canada (nor in Mexico) (see *Training in Clinical Psychology in the United States: Practitioner Model*).

As psychologists in private practice emerged in larger numbers, they also became more active politically. The first practitioner became president of the APA in 1977, and within 20 years the first Psy.D. was elected to this position. The practitioners began to dominate both the APA and the Canadian Psychological Association to a greater and greater extent. In response, many academic psychologists retreated to form national organizations of their own that were more research-oriented. Thus, the American Psychological Society was founded in 1988. Similarly, in 1989, academic and research psychologists in Canada founded the Canadian Society for Brain, Behaviour, and Cognitive Science.

Practicing psychologists both in the US and Canada battled psychiatrists for their share of the mental health 'market.' Thus, they fought to obtain hospital privileges. They supported 'freedom of choice' legislation to become eligible as health providers reimbursable by health insurance companies and Health Maintenance Organizations (HMOs). In 1988, a law suit by clinical psychologist Bryant Welch and others forced the American Psychoanalytic Association to begin to admit psychologists for training at its local institutes (Routh 1994). Most recently, a number of practicing psychologists in the US, led by Patrick DeLeon, have been trying to obtain the right to prescribe medications for mental health conditions, so far without much success.

16. The Continued Commitment of Clinical Psychology to Research

Well before the founding of Witmer's clinic, there was a strong interest on the part of psychology in research on psychopathology, including its diagnosis and treatment. This continues to be the case. In fact, psychologists are far more likely than psychiatrists or those in other fields to be principal investigators on research grants from the National Institute of Mental Health. Such research is international in scope and is a collaborative interdisciplinary enterprise. The turf battles that characterize the marketplace of practice are much less typical in the research arena, where clinical psychologists often cooperate smoothly with experimental psychologists and statisticians as well as with medical colleagues. In research, it is neither possible nor necessary to draw any bright line to show the boundaries between these fields.

The diversity of such research is so great that it would be impossible to cover it in an article as brief as this one. Instead, a few examples must suffice. In 1954, psychologist Evelyn Hooker received her first NIMH grant to study homosexuality. Her work, much of it using the types of projective testing that were typical of clinical psychological work at the time, suggested that homosexuals might be essentially normal psychologically. This research was related to the later decisions to delete homosexuality as a pathological category from the Diagnostic and Statistical Manual of the APA.

In the 1950s, psychologist Leonard Eron began collecting peer-rating data on aggressive behavior in 8-year-old children. The subsequent longitudinal research he and his colleagues did helped establish the fact that aggressive behavior is highly stable well into adulthood. Psychologist Gerald R. Patterson studied aggressive behavior in children using direct behavioral observations, relating it to coercive processes in the parent–child dyad and doing controlled intervention studies showing how it could be reduced.

In the late 1950s, psychologist C. Keith Conners devised a simple teacher rating scale for the assessment of varied types of disordered behavior in school children. Together with child psychiatrist Leon Eisenberg he helped carry out the first controlled studies of the effects of stimulant medications on children's disruptive behavior (Conners and Eisenberg 1963). Such research formed an important basis of present concepts of Attention Deficit Hyperactivity Disorder (ADHD), the most commonly diagnosed type of child psychopathology and one still typically treated with stimulant medications. By the 1970s, Thomas Achenbach and his colleagues had begun to develop the use of parent, teacher, and self-ratings for

child behavior into the most widely used forms of assessment by mental health workers.

In 1962, Meehl published a classic paper on 'schizotaxia, schizotypy, and schizophrenia,' elaborating his concepts as to how genetic factors might be involved in the development of this disorder (Meehl 1962). In 1973, psychologist Holzman and co-workers announced their discovery of smooth pursuit eye-movement difficulties in patients with schizophrenia (Holzman et al. 1973). This neurological symptom proved to be an important 'trait' marker in the first degree relatives of schizophrenics as well, whether or not they manifested any overt psychopathology. In Denmark, psychologist Sarnoff A. Mednick carried out a series of studies using the excellent public registers that characterize that country to do longitudinal, epidemiological studies of schizophrenia, implementing the type of research that Meehl had only been able to imagine.

Problems of the dysregulation of affect and emotion are the most ancient in the field of psychopathology, having been with us since Hippocrates. Beginning in the 1960s, psychologist Richard Lazarus elaborated his concepts of stress, appraisal, and coping and demonstrated experimentally how his subjects' use of different coping strategies could dampen or heighten their physiological stress (Lazarus 1966). Beginning in the 1970s, Charles Spielberger and his colleagues began the development and validation of measures of state as well as trait anxiety and later of state and trait anger as well. In the 1970s, Martin E. P. Seligman and his colleagues showed how his studies of learned helplessness in dogs could be used to reconceptualize human depression.

In conclusion, clinical psychology has both international and important North American roots. In its first century, it has developed into a viable science and profession, and there is good reason to suppose that its trajectory will continue in the twenty-first century.

See also: Clinical Psychology in Europe, History of; Psychiatry, History of; Psychoanalysis, History of; Psychotherapy, History of: Psychiatric Aspects; Training in Clinical Psychology in the United States: Practitioner Model; Training in Clinical Psychology in the United States: Scientist–Practitioner Model

Bibliography

Binet A, Simon T 1905 A new method for the diagnosis of intellectual level of abnormal persons. *Annee Psychologique* **11**: 191–244

Conners C K, Eisenberg L 1963 The effects of methylphenidate on symptomatology and learning in disturbed children. *American Journal of Psychiatry* **120**: 458–64

Eysenck H J 1952 The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology* **16**: 319–24

Healy W, Bronner A F 1926 *Delinquents and Criminals: Their Making and Unmaking*. Macmillan, New York

Holzman P S, Proctor L R, Hughes D W 1973 Eye tracking patterns in schizophrenia. *Science* **181**: 179–81

Jones M C 1924 The elimination of children's fears. *Journal of Experimental Psychology* **7**: 382–90

Kraepelin E 1987 *Memoirs*. Springer, Berlin

Lang P J, Lazovik A D 1963 Experimental desensitization of a phobia. *Journal of Abnormal and Social Psychology* **66**: 519–25

Lazarus R S 1966 *Psychological Stress and the Coping Process*. McGraw-Hill, New York

Meehl P E 1962 Schizotaxia, schizotypy, and schizophrenia. *American Psychologist* **17**: 827–38

Raimy V C (ed.) 1950 *Training in Clinical Psychology*. Prentice-Hall, New York

Rapaport D, Gill M M, Shafer R 1945 *Diagnostic Psychological Testing*, Vol. 1. Yearbook, Chicago

Rapaport D, Gill M M, Shafer R 1946 *Diagnostic Psychological Testing*, Vol. 2. Yearbook, Chicago

Routh D K 1994 *Clinical Psychology since 1917: Science Practice and Organization*. Plenum, New York

Routh D K 1996 Lightner Witmer and their first 100 years of clinical psychology. *American Psychologist* **51**: 244–7

Routh D K 1998 Hippocrates meets Democritus: A history of psychiatry and clinical psychology. In: Bellack A S, Hersen M (eds.) *Comprehensive Clinical Psychology*, Vol. 1. Pergamon, New York, pp. 1–48

Witmer L 1907 Clinical psychology. *Psychological Clinic* **1**: 1–9

D. K. Routh

Clinical Psychology: Manual-based Treatment

The practice of psychotherapy has seen dramatic and sweeping changes since the 1950s. One of the most dramatic changes impacting the delivery and dissemination of specific psychotherapeutic service has been the development of detailed explanatory manuals for complex psychological treatments. More recently client 'workbooks' are available that guide the client through therapeutic direction. The main function of these treatment manuals is to, 'outline the procedures, techniques, and strategies which comprise an acceptable implementation of a given [psychotherapeutic approach]' (Luborsky and DeRubeis 1984, p. 7) and to make the process of psychotherapy more accessible to clients and clinicians alike. Typically these manuals describe session-by-session strategies for specific problems such as depression or phobias for therapists to follow.

Many changes in the field have led to the impetus for specific treatment manuals to be developed including the demand for 'good and better research,' in which the therapeutic intervention is specified for all to examine, and the burgeoning of the hegemonic managed behavioral healthcare system. In response to these many demands, the number of manuals has proliferated. Although the psychotherapeutic com-

munity has positively received the majority of manuals, some front-line clinicians raise objections to this approach, stating, among other objections, that manuals damage the therapeutic relationship, and that clinical innovation is restricted (Addis et al. 1996). Despite these obstacles, treatment manuals appear entrenched in the forefront of clinical psychology.

1. *Developmental History of Treatment Manuals*

The initial impetus for the development of treatment manuals came from psychotherapy researchers who in the early 1960s began to test broadly the effectiveness of specific treatments in controlled outcome studies. Looking to demonstrate successfully that psychological interventions could withstand rigorous scientific investigation, similar to that of existing pharmacological treatments (Luborsky and DeRubeis 1984), scientist–practitioners realized that they needed treatment tools that would allow for systematic replication and comparison. Wilson (1996) more specifically pointed out that treatment manuals sought to eliminate any 'error' associated with 'clinical judgment' or intuition that might cause one therapist to behave in a substantially different manner from another. Thus, to study the effectiveness of these therapies, treatments were condensed into manuals that could then be reviewed and used across studies. Many researchers hoped that by utilizing treatment manuals, presented in this fashion, psychological interventions would be able to withstand the methodological constraints of research protocols. More specifically, 'treatment manuals help support the internal validity of a given study by ensuring that a specific set of treatment procedures exists, that procedures are identifiable, and that they can be repeated in other investigations' (Dobson and Shaw 1988, p. 675). This is in contrast to the conduct of treatment outcome research prior to manualization, in which specific therapeutic techniques were often not explained and thus could not be compared to other treatments or be replicated by other investigative groups. As a consequence, the use of a treatment manual is currently a prerequisite to receive federal funding for psychotherapy research in many countries.

Another push to develop specific treatment manuals came from the development of the Agency for Health Care Policy and Research (AHCPR) in the United States in 1989. The sole purpose of this Agency was to facilitate identification of the effectiveness of specific strategies for specific disorders, with the aim of increasing the quality and reducing the cost of healthcare (Barlow 1996). One major mechanism of accomplishing this goal was the creation of clinical practice guidelines that explicitly articulate the optimal strategies for assessing and treating a variety of psychological problems (see *Psychotherapy: Clinical Practice Guidelines*). Treatments recommended in these clinical practice guidelines are typically based on two specific factors: (a) efficacy; or internal validity of the specific treatment, the determination of which is based on the results of a systematic evaluation of the intervention in a controlled setting, and (b) effectiveness; or clinical utility of the treatment, which is based on the feasibility, generalizability, and cost-effectiveness of the intervention actually being delivered in a local setting. Based on these equally important and rigorous bases of evidence, the development of treatment manuals that could produce the necessary evidence was encouraged. As a result, manual-based treatments have been incorporated as one of the major components of evidence-based service delivery (Strosahl 1998).

2. *Manualized Treatment in Behavioral Health Care Settings*

One of the final steps in the progression of manual-based treatments is their incorporation into managed care treatment settings. Strosahl (1998) points out that these manuals are especially appealing to these settings because they are essentially an easily discernable roadmap for the most appropriate way to implement clinical practice guidelines. Within managed care organizations, where psychotherapy must demonstrate: (a) its overall effectiveness, (b) may be limited to a certain number of sessions, and (c) increasingly is delivered by practitioners with less than doctoral degrees, treatment manuals are embraced for their ability to facilitate the delivery of empirically supported treatments at lower costs. Treatment manuals increasingly are also adopted because they aid in the training of master's level clinicians.

3. *Specific Treatment Manuals*

The earliest treatment manuals were based on behavioral treatment techniques (Dobson and Shaw 1988). These treatments were the logical outgrowth of behavior therapy's de-emphasis of therapist variables in favor of specific procedures (Parloff 1998). Nonbehavioral varieties of treatment manuals are currently available, although behavioral and cognitive behavioral techniques tend to predominate. The majority of treatment manuals do not adhere to more traditional psychotherapy models, which tend to emphasize a more individualized theory-based understanding of underlying patient problems (Wilson 1996). Rather, most manuals target the specific diagnostic categories specified by *DSM-IV*. For example, specific manuals exist for a wide variety of anxiety disorders including panic disorder, generalized anxiety disorder, obsessive-compulsive disorder, post-traumatic stress disorder, and social phobia. Most of

these manuals utilize cognitive and behavioral techniques with an emphasis on identifying and challenging maladaptive cognitions and eliminating behaviors that serve to increase and maintain anxiety. Other manuals with a cognitive behavioral emphasis include treatments for bulimia nervosa, weight reduction, stress reduction, depression addictive behaviors, and sexual dysfunction.

As noted above, other forms of therapy have been manualized into specific books such as interpersonal psychotherapy for depression (IPT), which stresses the alleviation of destructive interpersonal patterns that are maintaining depressive mood, and Dialectical Behavior Therapy (DBT) for borderline personality disorder which draws on cognitive-behavioral techniques and eastern philosophy. More recently, important new treatments for bipolar disorder and schizophrenia, featuring family systems therapy directed at emotional lability in these families, have been developed. All of these treatments have received empirical support and are also incorporated into many aspects of clinical practice guidelines.

4. *Pros and Cons of Treatment Manuals*

One distinct advantage of treatment manuals is that they have been shown to be effective in controlled treatment outcome studies (Barlow and Hofmann 1997, Nathan and Gorman 1998). Thus, when practitioners implement techniques they can do so with the confidence that they are delivering services that have a high probability of success. Furthermore, manuals guide clinicians to use strategies that are spelled out and are clearly discernable.

Wilson (1996) points out that treatment manuals can reduce the errors that might be associated with unrestrained clinical judgment. 'What is typically overlooked, however, and what research on human judgement so clearly documents, is that individualized clinical judgement can just as easily produce worse results than standardized treatments, by introducing errors and inappropriate strategies that are not part of manual-based treatments' (Wilson 1996, p. 302). Manuals, however, outline specific techniques to be used in each session, which have been created by systematic investigation, rather than relying upon clinical judgment to choose a session topic.

Due to their structure, treatment manuals may also facilitate a more highly focused and efficient therapy. With a limited number of sessions and specific goals and strategies outlined for each of these sessions, Wilson (1998) suggests that this more focused approach may actually lead to a more active engagement in therapy for patients.

Treatment manuals also make it easier to train and supervise therapists in specific clinical techniques and strategies (Calhoun et al. 1998). By providing nascent clinicians and supervisors with specific guidelines and techniques to monitor, treatment manuals streamline the learning process and delineate a specific set of therapeutic skills to be learned. This in turn may lead to a greater aptitude and ability for learning therapeutic techniques in general, and a larger armentarium of clinical skills.

Treatment manuals, as structured interventions, can also help clinicians deliver therapy in a brief and in many cases non-traditional format (Craske et al. 1995). For example, even a very brief and unstructured intervention for patients presenting to an emergency room with panic attacks may be effective if delivered early enough to less severe patients (Swinson et al. 1992). Considering the high prevalence rates of clinical and subclinical mental disorders in primary care settings (Fifer et al. 1994) such interventions are clearly cost effective.

Finally, utilization of treatment manuals can promote innovation in the delivery of clinical services. That is, when highly specified treatments are delivered, and assuming appropriate outcomes measures are collected, clinical administrators can determine when a specific treatment is working and for whom. In the case of failures either individually or systematically, it will then be clear that innovations to the treatment program are needed and should be incorporated and evaluated.

Despite these many advantages, a sizeable segment of practicing clinicians object to the notion of treatment manuals, pointing out the many disadvantages in the delivery of what has been called derisively, robot-like therapy (Parloff 1998). For example, dissenters have argued that randomized controlled trials, utilized to validate most manualized treatments, exclude patients with co-morbid diagnoses, among other constraints, and that the results of these types of research studies inherently have limited applicability to 'real-world' settings (Bologna et al. 1998). These same individuals note that highly structured, manualized psychological interventions used in clinical trials cannot generalize to practice settings where psychotherapy is often applied more flexibly and adapted to meet the needs of the patients, who may present with multiple problems (Barlow 1996). This set of objections is based on the fundamental idea that the scientific method, as it is known, is incapable of proving the effectiveness or ineffectiveness of psychotherapy, and that some alternative methodological approach to psychotherapy research is preferable. These are important criticisms that cannot be ignored. But Wilson (1998) outlines how there are significant and important differences between the use of treatment manuals in therapy protocols and their use in clinical practice, with the latter context allowing for greater flexibility. Furthermore, the beginning of much needed research in this area suggests generalizability of manualized treatments to front line clinical settings despite high co-morbidity and a diverse population base (Barlow et al. 1999).

Other clinicians bemoan that structured treatment manuals will compromise the integrity of the therapeutic relationship and thus interfere with the therapy process (Garfield 1996). For example many of these clinicians fear that manualized therapy may reduce the therapy process to mechanized and robotic delivery of techniques, thereby eliminating the importance of the therapist as a clinical tool. Further, innovation and the development of unique and creative strategies will be stagnated by reliance on prescribed treatment strategies. Parloff (1998) summarized these sentiments by stating that many saw the reliance on actuarial decision making over honed clinical judgment as 'perverse.' Many clinicians also feel hindered by having to adhere to a specific set of goals for a given session, and prefer to rely on their own judgment to guide them. However, Wilson (1996) notes, 'if clinical artistry is taken to connote such necessary therapeutic elements as developing a therapeutic relationship and engaging patients in the change process, then treatment manuals do not obfuscate it—rather they demand it' (p. 305). It can also be assumed that even though the therapist utilizing a treatment manual is encouraged to deliver the techniques outlined, within this delivery they are also encouraged to rely on their own personal skills to convey the therapy to the patient (Dobson and Shaw 1988). Finally, the data indicate that therapists engaged in manualized treatments in clinical practice can form strong alliances (Addis et al. 1996). Simply because the therapist has to spend a segment of time conveying specific therapeutic techniques does not mean that they are not building alliances with their patients. Furthermore, a great deal of time is allotted in manual-based therapies to emphasize alliance building. Examples include such strategies as identifying client and therapist expectations for treatment and eliciting client feedback (Addis et al. 1996).

Criticisms of treatment manuals also encompass the belief that manuals may undermine successful case formulation. That is, many clinicians argue that manual-based treatments will be less effective because these approaches assume that all individuals with a given disorder are uniform and have the same symptoms for the same reasons (Wilson 1996). As such, these techniques will 'miss the boat' for a certain subset of patients and will attempt to treat them with techniques that are not applicable. The fact that manuals do not allow for clinical exploration that might lead to the discovery of what works for patients with dissimilar etiology is a commonly expressed concern that may lead to diminished treatment innovation. In fact, treatment manuals do emphasize the current symptom picture over past development processes and idiosyncratic etiology. But research has yet to suggest that therapeutic attention to past developmental processes contributes in a substantial way to the amelioration of the disorder.

Many of these criticisms of manualized treatment have some validity and will be grist for the research mill. Research on the effectiveness or external validity of therapeutic interventions will be the ground on which these competing ideas will be evaluated and tested.

5. *The Future of Structured Treatment Manuals*

The science and practice of psychotherapy has undergone comprehensive shifts in the 1990s, none arguably more central than the proliferation of semistructured treatment manuals. In this era of treatment guidelines, best practice algorithms, and behavioral healthcare management organizations, treatment manuals increasingly are being embraced. Despite objections, and lingering doubts, treatment manuals have many merits, with demonstrated efficacy the principal attribute. As psychotherapeutic services become more integrated with other forms of healthcare in the twenty-first century, treatment manuals with demonstrated outcomes may become the preferred method of nondrug interventions, and 'formularies' of effective psychological treatments may begin to appear in mental health services or managed behavioral healthcare organizations. In this context, research will progress to address the reasonable concerns that have arisen regarding the implementation of these strategies.

See also: Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Clinical Trials; Psychotherapy: Case Study; Psychotherapy: Clinical Practice Guidelines; Psychotherapy: Ethical Issues; Psychotherapy Process Research

Bibliography

Addis M E, Wade W A, Hatgis C 1999 Barriers to dissemination of evidence based practices: Addressing practitioners' concerns about manual based psychotherapies. *Clinical Psychology: Science and Practice* **6**(4): 430–41

Barlow D H 1996 Health-care policy, psychotherapy research, and the future of psychotherapy. *American Psychologist* **51**(10): 1050–58

Barlow D H, Hofmann S G 1997 Efficacy and dissemination of psychological treatments. In: Clark D M, Fairburn C G (eds.) *Science and Practice of Cognitive Behaviour Therapy*. Oxford University Press, Oxford, UK, pp. 95–117

Barlow D H, Levitt J T, Bufka L F 1999 The dissemination of empirically supported treatments: A view to the future. *Behaviour Research and Therapy* **37**: 5147–62

Calhoun K S, Moras K, Pilkonis P A, Rehm L P 1998 Empirically supported treatments: Implications for training. *Journal of Consulting and Clinical Psychology* **66**: 151–62

Craske M G, Maidenberg E, Bystrisky A 1995 Brief cognitive-behavioral versus non-directive therapy for panic disorder. *Journal of Behaviour Therapy and Experimental Psychiatry* **26**: 113–20

Dobson K S, Shaw B F 1988 The use of treatment manuals in cognitive therapy: Experience and issues. *Journal of Consulting and Clinical Psychology* **56**(5): 673–80
Fifer S K, Mathias S D, Patrick D L, Mathias S D, Mazonson P D, et al. 1994 Untreated anxiety among adult primary care patients in a health maintenance organization. *Archives of General Psychiatry* **5**: 740–50
Garfield S L 1996 Some problems associated with 'validated' forms of psychotherapy. *Clinical Psychology: Science and Practice* **3**(3): 218–29
Luborsky L, DeRubeis R J 1984 The use of psychotherapy treatment manuals: A small revolution in psychotherapy research style. *Clinical Psychology Review* **4**: 5–14
Nathan P E, Gorman J M 1998 *A Guide to Treatments that Work*. Oxford University Press, New York
Parloff M B 1998 Is psychotherapy more than manual labor? *Clinical Psychology: Science and Practice* **5**(3): 376–81
Strosahl K 1998 The dissemination of manual-based psychotherapies in managed care: Promises, problems, and prospects. *Clinical Psychology: Science and Practice* **5**(3): 382–6
Swinson R P, Soulios C, Cox B J, Kuch K 1992 Brief treatment of emergency room patients with panic attacks. *American Journal of Psychiatry* **149**: 944–6
Wilson G T 1996 Manual-based treatments: The clinical application of research findings. *Behavior Research and Therapy* **34**(4): 295–314
Wilson G T 1998 Manual-based treatment and clinical practice. *Clinical Psychology: Science and Practice* **5**(3): 363–75

D. H. Barlow and K. A. I. Greene

Clinical Psychology: Validity of Judgment

A large body of research exists on the validity of judgments made by mental health professionals (Garb 1998). Most of the studies have been conducted by clinical psychologists, counseling psychologists, and psychiatrists, but important studies have also been conducted by neuropsychologists, social workers, and sociologists. The studies describe how well mental health professionals perform on a range of tasks, e.g., how well they make diagnoses and treatment decisions.

Research on mental health professionals should be of interest to the general public. Many of the studies can help us to understand important social issues, e.g., the occurrence of race bias, gender bias, and other types of biases. Also, a large number of the studies bear on questions that are important for our justice system. These questions include: should mental health professionals be allowed to testify as expert witnesses? (see *Expert Witness and the Legal System: Psychological Aspects*). Are mental health professionals able to make accurate predictions of violence? Are they able to make accurate decisions regarding child abuse and domestic violence? Do they make appropriate judgments when petitioning to have individuals committed to psychiatric hospitals? Finally, the empirical findings should also be of particular interest to individuals who devise health care policy and decide what services to reimburse. For example, there is a new and intense controversy over the validity of judgments made by clinicians who use the Rorschach Inkblot Test (Wood et al. 1996) (see *Projective Methods in Psychology*). This controversy is well known within the field of clinical psychology, but is not yet well known by people in other professions.

There is another group that should be interested in the results from studies on mental health professionals. Consumers of mental health services should be especially interested in the results of studies on clinicians. If one includes family members of consumers of mental health services and potential consumers of mental health services, then one can conclude that virtually everyone should be interested in this research.

Overall, a huge number of studies (over 1,000) have been published on the validity of clinical judgments and most of the studies have been well designed. Yet, they are not well known outside of mental health fields. In this article, highlights of the research will be described. Topics include: (a) assessment of personality and psychopathology, (b) diagnosis, (c) case formulation, (d) prediction of behavior, and (e) treatment decisions. In general, mental health professionals are able to make reliable and valid judgments for some tasks, but not for others (Garb 1998).

1. Assessment of Personality and Psychopathology

Mental health professionals almost always evaluate a client's personality traits and psychiatric symptoms. Some clinicians also evaluate a client's defense mechanisms. Personality traits can include characteristics like narcissism and dependence, while psychiatric symptoms can include things like hallucinations or panic attacks. A defense mechanism, as defined by psychoanalytic theory, is an unconscious strategy that protects the ego from anxiety. For example, a client may push impulses and thoughts that are unacceptable to the ego into the unconscious.

Results on reliability and validity vary for the tasks of describing psychiatric symptoms, personality traits, and defense mechanisms. Mental health professionals are often good at describing psychiatric symptoms. This should not be surprising. Clients are often able to report if they have had hallucinations, panic attacks, or other symptoms. On the other hand, inter-rater reliability varies widely for describing personality traits, and it is poor for describing defense mechanisms. Perhaps this is because these tasks require clinicians to make more inferences. Given the poor results for describing defense mechanisms, it is important to point out that many clinicians do not perform this task. Psychodynamic therapists regularly evaluate clients' defense mechanisms, but other clinicians (e.g.,

cognitive behavior therapists) rarely concern themselves with this task (see *Psychoanalysis in Clinical Psychology*).

2. *Diagnosis*

Diagnostic classification systems have been constructed to help clinicians make diagnoses. The most commonly used classification system in the United States is the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (1994, generally referred to as *DSM-IV*). This classification system contains specific and explicit criteria for making diagnoses (see *Mental and Behavioral Disorders, Diagnosis and Classification of*).

Clinicians' diagnoses are reliable and moderately valid, but only when they attend to diagnostic criteria. Unfortunately, there is evidence that a significant number of clinicians do not adhere to criteria when making diagnoses. That is, many clinicians may think that they are making diagnoses according to the *DSM-IV* criteria, but they do not refer to the criteria when making a diagnosis, and examination of their diag noses reveals that they are not made in accordance with the *DSM-IV* criteria. This can lead to different types of problems including race bias, gender bias, age bias, and the underdiagnosis or overdiagnosis of some mental disorders. These problems are described below.

The most widely replicated finding for race bias involves the differential diagnosis of schizophrenia and psychotic affective disorders. African-Americans and Puerto Rican Hispanics with bipolar affective disorder (formerly called manic depression) are more likely than Whites with bipolar affective disorder to be misdiagnosed as having schizophrenia. For this reason, Black patients and Puerto Rican Hispanic patients are more likely than White patients to be overmedicated with neuroleptic medications, and their depressive symptoms are more likely to be untreated.

The most widely replicated finding for gender bias involves the differential diagnosis of histrionic personality disorder and antisocial personality disorder. When different groups of mental health professionals have been given identical case histories except for the designation of gender, clinicians have been more likely to diagnose women as having a histrionic personality disorder and men as having an antisocial personality disorder. Histrionic personality disorder is characterized by overly dramatic, attention seeking behaviors (e.g., uncomfortable when not the center of attention), and antisocial personality disorder is characterized by antisocial behaviors (e.g., habitual lying, having no regard for others, showing no remorse after hurting others) (see *Personality Disorders*).

The most widely replicated finding for age bias involves the differential diagnosis of organic impairment and depressive disorder. Compared to young and middle-aged patients, elderly patients are more likely to be diagnosed as having organic impairment and they are less likely to be diagnosed as having a depressive disorder, even when all of the clients are described by the same case history except for the designation of age. Of course, someone diagnosed as having organic impairment will be less likely to receive psychotherapy and antidepressant medicine.

It should be noted that even when clinicians attend to diagnostic criteria and apply them the same way for different groups of patients (e.g., for African-American and White patients), diagnoses can be biased (Widiger 1998). For example, diagnoses can be biased because diagnostic criteria, not the cognitive processes of clinicians, are biased. Diagnostic criteria are said to be biased if they are more valid for one group than for another (e.g., if diagnostic criteria for a particular disorder are more valid for males than for females). In general, little is known about whether diagnostic criteria are biased.

Research has also described other types of errors. Mental health professionals disagree strongly over whether dissociative identity disorder (formerly called multiple personality disorder) is overdiagnosed or underdiagnosed. There is also a controversy over whether attention-deficit/hyperactivity disorder (ADHD) is overdiagnosed. Diagnoses of ADHD have doubled in frequency in recent years (see *Attention-deficit/Hyperactivity Disorder (ADHD)*), while diagnoses of dissociative identity disorder have increased 10-fold. Finally, research suggests that clinicians underdiagnose mental disorders in the mentally retarded, they also underdiagnose mental disorders (e.g., major depression) in terminally ill patients, they frequently underdiagnose personality disorders, they underdiagnose substance abuse in psychiatric patients, and they underdiagnose mental disorders in individuals admitted to substance abuse treatment programs.

3. *Case Formulations*

The most difficult task for mental health professionals involves making causal judgments. When making causal judgments, clinicians try to explain the causes of their clients' behaviors and symptoms. Inter-rater reliability and validity for this task is often poor. This has been true when psychodynamic clinicians described clients using psychoanalytic theory and when behavior therapists conducted functional analyses to understand the relations that exist between causal variables and behavior problems. One study on reliability will be described in detail. In this study (DeWitt et al. 1983), case formulations were made by two teams of psychodynamically trained clinicians after viewing videotapes of intake evaluations that lasted from 60 to 90 minutes. Both teams were composed of three clinicians. Descriptions were made by consensus. Each team was to, 'Define the basic

neurotic conflict(s) that lie at the core of the patients' difficulties. Include the kind of stress to which the patient is vulnerable' (p. 1124). Each team wrote formulations for 18 adults who sought psychotherapy for pathological grief reactions after the death of a parent. With regard to the results, agreement between the two teams was poor. Typically, they mentioned different symptoms, emphasized different conflictual areas, and formulated different cause–effect explanations.

4. *Prediction of Behavior*

With regard to predicting behavior, mental health professionals have been able to make reliable and moderately valid judgments. For the prediction of suicide and the prediction of violence, inter-rater reliability has ranged from fair to excellent. However, validity has been poor for the prediction of suicidal behavior (suicidal behavior refers to suicide gestures, suicide attempts, and suicide completions). For example, in one study (Janofsky et al. 1988), mental health professionals on an acute psychiatry in-patient unit interviewed patients and then rated the likelihood of suicidal behavior in the next seven days. They were not able to predict at a level better than chance. In contrast to the prediction of suicidal behavior, predictions of violence have been more accurate than chance. In fact, though the long-term prediction of violence is commonly believed to be a more difficult task than the short-term prediction of violence, both short- and long-term predictions of violence have been moderately valid. Finally, the validity of clinicians' prognostic ratings has rarely been studied, but there is reason to believe that clinicians' prognostic ratings for patients with schizophrenia may be too pessimistic (patients with schizophrenia may do better than many clinicians expect).

5. *Treatment Decisions*

Perhaps the most important task is to make treatment decisions. Unfortunately, many problems exist with the treatment decisions made by clinicians. Problems that will be described include: poor inter-rater reliability, nonconformance with ethical and legal issues, and the use of controversial techniques.

Inter-rater reliability has been poor for several judgment tasks. Two examples will be given. In one study (Felton and Nelson 1984), six clinical psychologists, all trained in behavioral assessment, were asked to formulate specific treatment plans for three clients. When psychologists conducted interviews, treatment plans were in agreement only 59 percent of the time. When psychologists conducted interviews and also used questionnaires and role-playing sessions to collect additional assessment information, treatment plans were in agreement only 62 percent of the time. Inter-rater reliability has also been poor when psychiatrists have made decisions about whether a patient with major depressive disorder should receive psychotropic medicine, electroconvulsive treatment, and/or psychotherapy (e.g., Keller et al. 1986). Differences in the amount and type of treatment could not be explained by variation in the clinical characteristics of patients. The best predictor of treatment was medical center, indicating that the type of treatment severely depressed patients receive depends largely on which hospital they go to.

For some decision tasks, many mental health professionals make decisions that are not in agreement with legal and ethical principles. For example, several studies have reported that many psychiatrists do not make appropriate judgments about committing a patient to a hospital. For example, in one study (Bagby et al. 1991), 26 percent of the individuals depicted as meeting the criteria for involuntary hospitalization were not recommended for commitment. At the same time, 20 percent of those who did not meet the legal standard for commitment were recommended for involuntary hospitalization.

Another example will also illustrate how many mental health professionals make decisions that are not in agreement with legal principles. Mental health professionals are mandated by law to report child abuse, but a widely replicated finding in clinical judgment research is that large numbers of clinicians do not report child abuse (e.g., Brosig and Kalichman 1992). This may sometimes occur because they are unfamiliar with mandatory reporting laws, and it may sometimes occur because they believe it will interfere with treatment.

Some mental health professionals use treatment interventions that are controversial. For example, one controversy involves the recovery of lost memories. Some clinicians will inform their clients that they believe their emotional problems are due to having been abused. They may say this even when the clients have no memory of having been abused. They may believe the client has been abused because of some of the client's symptoms, e.g., low self-esteem or sexual dysfunction. Clinicians may then use a variety of techniques to help clients 'remember' having been abused. For example, they may tell the clients that they were abused and repeatedly ask them to remember the events. Unfortunately, these interventions may lead a client to 'remember' an episode of abuse that never occurred.

6. *Discussion*

Though biases and errors sometimes occur, it is important to note that clinicians' judgments are frequently reliable and valid. For example, diagnoses are moderately useful when they are made by clinicians

who attend to the *DSM-IV* criteria. When made appropriately, diagnoses can inform us about the nature, course, outcome, and recommended treatment for patients. It is also important to note that many clinicians use treatment interventions that have been supported by empirical research. Empirically validated treatment interventions used by psychologists are listed in the Winter 1995 issue of *The Clinical Psychologist* and in subsequent issues of the journal.

See also: Clinical versus Actuarial Prediction

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders DSM-IV*, 4th edn. American Psychiatric Association, Washington, DC

Bagby R M, Thompson J S, Dickens S E, Nohara M 1991 Decision-making in psychiatric civil commitment: an experimental analysis. *American Journal of Psychiatry* **148**: 28–33

Brosig C L, Kalichman S C 1992 Clinicians' reporting of suspected child abuse: a review of the empirical literature. *Clinical Psychology Review* **12**: 155–68

DeWitt K N, Kaltreider N B, Weiss D S, Horowitz M J 1983 Judging change in psychotherapy: reliability of clinical formulations. *Archives of General Psychiatry* **40**: 1121–8

Felton J L, Nelson R O 1984 Inter-assessor agreement on hypothesized controlling variables and treatment proposals. *Behavioral Assessment* **6**: 199–208

Garb H N 1998 *Studying the Clinician: Judgment Research and Psychological Assessment*. American Psychological Association, Washington, DC

Janofsky J S, Spears S, Neubauer D N 1988 Psychiatrists' accuracy in predicting violent behavior on an inpatient unit. *Hospital and Community Psychiatry* **39**: 1090–4

Keller M B, Lavori P W, Klerman G L, Andreasen N C, Endicott J, Coryell W, Fawcett J, Rice J P, Hirschfeld R M A 1986 Low levels and lack of predictors of somatotherapy and psychotherapy received by depressed patients. *Archives of General Psychiatry* **43**: 458–66

Widiger T A 1998 Invited essay: Sex biases in the diagnosis of personality disorders. *Journal of Personality Disorders* **12**: 95–118

Wood J M, Nezworski M T, Stejskal W J 1996 The comprehensive system for the Rorschach: A critical examination. *Psychological Science* **7**: 3–10

H. N. Garb

Clinical Treatment Outcome Research: Control and Comparison Groups

The essential paradigm for determining if a treatment is effective is to compare it with something else. However, the issue of what that 'something else'—a control or comparison group—should be in any given clinical trial is quite complicated and has enormous theoretical and practical implications. This article will briefly review the role of control and comparison conditions in clinical treatment outcome research and describe advantages and disadvantages of several different types of control and comparison designs. For readers interested in more detailed review of this important area, the classic texts by Kazdin (1980, 1998) are excellent sources.

1. Background

The randomized controlled trial is the *sine qua non* of treatment efficacy research (see *Psychological Treatments: Randomized Controlled Clinical Trials*). A given treatment cannot be considered 'empirically validated' (see Chambless and Hollon 1998) unless trials purporting to establish its effectiveness have employed at least two fundamental methodological features: (a) a credible control condition, and (b) random assignment to the experimental and control conditions. These critical features offer enormous advantages in evaluating a treatment's efficacy. For example, random assignment of participants to the experimental vs. control condition is essential in ruling out alternate explanations of findings, including regression to the mean, maturation, history, effects of repeated testing, and selection basis (Kazdin 1998). The selection of an appropriate comparison or control group also is essential in determining, all other things being equal, if a treatment is effective for a given disorder, and whether it has benefits relative to other treatments for the disorder, as well as how it may achieve its effects and what its 'active ingredients' may be.

2. What Should Control Groups Control for?

The historical paradigm for control groups in behavioral research is the placebo control for pharmacotherapy efficacy trials. Placebo (*I shall please*) controls provide a means of evaluating a pharmacotherapy's efficacy compared with those 'placebo' or 'nonspecific' effects which may occur simply because participants believe they are receiving a credible treatment (Frank 1961). Placebo effects in pharmacotherapy trials, which may be associated with substantial and durable improvement, may occur because of participants' expectations of improvement, demand characteristics, the instillation of hope, the provision of support and education, and other nonspecific effects that are associated with simply being in treatment. Thus, because (a) the pill (or, sometime, injection) in placebo control conditions is medically inert, and (b) nonspecific elements are assumed to be comparable in the 'active medication' and pill

placebo conditions, differences in outcome are assumed to be due to the unique 'active ingredients' of the experimental medication.

However, the issue of what constitutes an appropriate 'placebo' condition (and even if the term 'placebo' is appropriate for control conditions in behavioral research) in psychotherapy efficacy research is much more complex and has been the subject of much controversy for many years (see Critelli and Neumann 1984, O'Leary and Borkovec 1978, Parloff 1986, Wilkins 1986). Soon after Rosenthal and Frank (1956) declared that the only adequate control condition in psychotherapy research was a placebo group, several problems with this strategy were raised. Defining a 'psychologically inert' approach that is an appropriate 'placebo' control for behavioral research raised a number of conceptual, practical, and ethical concerns (O'Leary and Borkovec 1978).

Conceptually, defining an appropriate nonspecific control condition for psychotherapy is difficult because many of the 'nonspecific' elements that are common to many psychotherapies (e.g., formation of a therapeutic alliance, provision of support and education, and the instillation of hope) are not inert, in that they are associated with legitimate psychological processes of change and have been shown to have powerful effects on psychological conditions (O'Leary and Borkovec 1978). Moreover, Parloff (1986) noted that referring to these 'common elements' or 'nonspecific elements' as placebos suggests that these powerful mechanisms of change are somehow spurious.

Practically, it can be difficult to conceive of a psychologically inert approach (e.g., an approach that has no theoretically supported rationale for influencing the behavior in question) that is sufficiently credible to participants (O'Leary and Borkovec 1978). This is particularly difficult in the case of lengthy trials where it is essential to retain participants in those conditions over periods of several weeks or months. Furthermore, there has been little agreement on what constitutes an acceptable 'placebo'; thus, there is no single, widely recognized or standardized placebo condition. This makes comparison of effect sizes (a standardized estimate of the magnitude of the effect of the experimental condition relative to the control) across studies very complex (Parloff 1986). This can pose problems, for example, in meta-analyses aimed at comparing the effect sizes for different types of behavioral approaches for a given disorder.

Ethically, for many treatment-seeking populations it is highly questionable whether providing a placebo—a treatment that is expected to have no effect on the presenting problem—can be justified. Furthermore, many placebo conditions, in presenting a credible rationale but no interventions that are expected to improve the participant's problem, are also inherently deceptive (Kazdin 1980, O'Leary and Borkovec 1978).

Given these issues, a variety of different control conditions have been proposed for psychotherapy research. Several of the most common will be described below.

3. *Types of Control Groups*

3.1 *No-treatment Controls*

This strategy compares the experimental approach to no treatment; addressing the question 'Does this treatment work better than no treatment at all?' In this paradigm, participants are assessed, randomized to treatment condition, and reassessed after completion of treatment (or after an equivalent length of time for the no-treatment control group).

3.1.1 Advantages. No-treatment controls are generally seen as the 'minimal' or basic standard for evaluating the effectiveness of an intervention (Chambless and Hollon 1998). No treatment control conditions are sometimes referred to as assessment-only controls, as they control for the effects of the study assessments and the passage of time. Thus, they are useful in evaluating conditions that have a high likelihood of improving without intervention (e.g., spontaneous remission) or when the natural history of a disorder is not well established.

3.1.2 Disadvantages. There are a number of problems with no-treatment control conditions. No-treatment controls do not control for the effects of participant expectancies, 'common elements' or nonspecific effects, or time spent in treatment. Ethical and practical issues arise when no-treatment controls are used with populations seeking treatment for a particular condition, and particularly when the condition is life-threatening or associated with significant medical, psychosocial, or other problems. Furthermore, withholding treatment from individuals who request it raises ethical issues, particularly for those disorders where effective alternative treatments have been demonstrated to exist. No-treatment controls may also raise practical issues, because individuals assigned to the no-treatment condition may seek alternative therapies or become demoralized and drop out of the study, particularly for more lengthy trials. This can lead to problems associated with confounding of conditions or differential attrition.

Some of the practical and ethical concerns associated with no-treatment control conditions can be avoided in designs where the experimental treatment is 'added on' to a standard treatment program and the control group receives the standard treatment only. Thus, all participants would, at minimum, receive the standard treatment. This design addresses the question, 'Does the addition of Intervention X improve

outcomes over standard treatment for this disorder' and avoids many ethical and practical problems. However, this strategy still does not control for time spent or nonspecific effects.

3.2 Wait List/Delayed Treatment Controls

This strategy is similar to the no-treatment control approach, but differs in that some form of treatment is offered to those assigned to the no-treatment control condition after completion of study assessments. Thus, after completing all past treatment assessments, participants originally assigned to the delayed-treatment condition may be given the option of no further treatment or receiving the experimental treatment or an alternative approach.

3.2.1 Advantages. Like the no-treatment control condition, this strategy is one that addresses the question, 'Is Treatment X effective compared with no treatment?' and can help clarify the effects of the experimental intervention with respect to the natural history of the disorder. Provided that participants in the delayed treatment control are adequately monitored and the disorder of interest is one where the likelihood of clinical exacerbation and negative consequences are minimal, this strategy avoids some of the ethical problems associated with withholding treatment from individuals who need or request it. Some prospective participants who might not accept randomization to a no-treatment control condition may be more likely to accept randomization to a delayed treatment control.

3.2.2 Disadvantages. Like the no-treatment control strategy, delayed treatment controls do not control for differences in nonspecific effects of treatment, including expectancies, demand characteristics, attention, and so forth. Moreover, evaluation of long-term effects of treatments is difficult with delayed treatment controls, as provision of treatment to the participants in the delayed-treatment condition may obfuscate determination of long-term effectiveness of the treatment in follow-up evaluations. This can, of course, be avoided by waiting to provide treatment to the delayed-treatment condition until after the completion of all follow-up assessments, but this may in turn result in the same practical and ethical problems associated with no-treatment controls.

3.3 Attention/Discussion/Minimal Treatment Controls

This strategy most closely resembles what has been referred to as a psychotherapy 'placebo.' In this paradigm, the participants are assigned to a condition that provides some common elements of psychotherapy (e.g., time spent, presentation of a credible rationale for the approach, receiving support and attention), but does not offer the hypothesized 'active ingredients' of the experimental intervention. The specific nature of this type of control group varies widely from trial to trial, but may include interventions such as having the participants meet with a clinician who provides education and support, 'discussion control' groups where several participants meet to provide mutual support or are presented with education materials, 'bibliotherapy' conditions where participants receive written reading or other educational materials such as videotapes, and several more.

3.3.1 Advantages. Attention-control conditions, by controlling for time spent and providing some common elements of psychotherapy, are typically seen as more stringent control than no-treatment or delayed-treatment control conditions. If well-conceived, carefully implemented and monitored attention controls can provide a test of the hypothesized active ingredients of the experimental condition and may allow for some evaluation of mechanisms of action of the experimental condition, which is not usually feasible with no-treatment or delayed-treatment controls.

3.3.2 Disadvantages. As noted above, there have been several problems with attention/discussion control conditions. First, it is unclear whether many of these approaches actually adequately control for nonspecific elements of the experimental intervention (Basham 1986). Second, given that a key ingredient of such approaches is the presentation of a convincing rationale, but at the same time the intervention '…should have no currently supported theoretical reason why the placebo would influence the behavior under question' (O'Leary and Borkovec 1978, p. 821), a truly convincing rationale for these conditions is not so easily formulated. If participants are not convinced or do not see themselves as deriving adequate benefit from the approach, they may be more likely to drop out of the condition, leading to problems of differential attrition in the control vs. experimental condition. Furthermore, since the attention/discussion is intended to be 'inert' and not directly affect the disorder or problem, ethical questions regarding the withholding of treatment may be raised for those conditions where effective alternative approaches exist. Finally, as noted above, because there has been little consensus on what such controls should consist of (either within or across different disorders), attention control conditions vary widely from study to study, making cross-study comparisons difficult.

3.4 Active Treatment or Comparative Controls

In this strategy, the experimental intervention is compared with another 'active' intervention, ideally one of demonstrated efficacy for the disorder. Thus, a comparative design directly compares two treatment conditions without conceptualizing either as being a formal control group (Basham 1986). Rather than addressing the issue of 'Does this treatment work?' this strategy addresses questions such as 'Is Treatment A more effective than Treatment B' and 'Does Treatment A offer advantages over standard treatment for this disorder?' It should be noted that for many disorders, the comparison condition may be a psychotherapy or a pharmacotherapy.

3.4.1 Advantages. When well conceived and executed, comparison controls do control for many artifacts and demand characteristics, including time spent in treatment, the provision of a credible rationale, and other common elements or nonspecific effects (although it should be noted that equivalence of nonspecific elements across conditions must be supported by process evaluations). This strategy has the additional advantage of permitting more thorough evaluation of treatment process and mechanisms of action, and it can help identify what treatment is best for a given disorder, as well as what type of treatment is best for what type of individual (Basham 1986, Kazdin 1986, O'Leary and Borkovec 1978). Furthermore, by comparing the experimental intervention to a treatment of demonstrated efficacy for the disorder (e.g., a 'standard' or 'reference' condition), this approach avoids deception and many of the ethical problems discussed earlier.

3.4.2 Disadvantages. Comparison controls, by controlling for nonspecific elements and essentially pitting the hypothesized 'active ingredients' of one approach against those of another, are highly stringent tests of novel interventions. Thus, such studies often lead to small effect sizes and findings of no significant differences between interventions (Luborsky et al. 1975), which in turn may lead to problems in interpretation of the study outcomes. For example, in a trial comparing two active treatments where no significant difference in effectiveness is found, interpretation of the 'absolute' effectiveness of the treatments is problematic. That is, although the treatments compared were not significantly different, it is often hard to gauge the magnitude of the effect of either treatment on the symptoms targeted (Nathan 1997). Kazdin (1986) has also pointed out that comparative designs may overemphasize elements that differentiate the treatments compared (e.g., use of specific techniques) over other important variables that may affect outcome (e.g., characteristics of therapists or participants). Finally, for those disorders where a number of viable comparison approaches are available, but no one clear standard or 'reference' condition that is widely acknowledged as the most effective, choice of the comparison condition is challenging. Kazdin (1986) has provided thoughtful recommendations for the selection and implementation of appropriate active treatment control conditions.

The relative stringency of comparison controls relative to attention/discussion, delayed-treatment, or no-treatment controls has practical implications as well. For example, within a given field, if one type of intervention (Treatment A) routinely uses no-treatment controls in efficacy studies, while another type of intervention (Treatment B) uses discussion control conditions, while trials evaluating Treatment C routinely use active treatment controls, the estimated effect size of Treatment A is likely to be larger than that for Treatment B, which is in turn likely to be larger than for Treatment C, even though all three approaches may be comparably effective. Progressively smaller effect sizes for different types of control conditions (no treatment, placebo, active treatment) has been suggested in several meta-analyses (Basham 1986).

3.5 Dismantling Controls

Also referred to as component control condition, this strategy essentially requires specification of a given treatment or treatment 'package' into its component elements, and then evaluating the effectiveness of the treatment with or without one specific element or technique of interest. For example, a cognitive behavioral approach might be evaluated with and without a relaxation training component.

3.5.1 Advantages. By altering only a single element of a treatment, component control strategies offer an elegant strategy for evaluating well-defined techniques and components. Component control conditions can address issues of particular theoretical or clinical relevance, by isolating the most effective 'ingredients' of complex or multifaceted treatment packages. Furthermore, because the interventions compared differ only with respect to a single element, many of the practical, ethical, and theoretical issues associated with placebo and no-treatment controls can be avoided (O'Leary and Borkovec 1978).

3.5.2 Disadvantages. Before treatments or treatment packages can be dismantled, they must first be demonstrated to be effective. Thus, for a given disorder, there may only be a small number of approaches with sufficient empirical support that

would justify their 'dissection' in a dismantling study. Similarly, not all approaches are easily parsed into discrete components. Furthermore, to justify a full dismantling trial, a compelling rationale for the component to be evaluated must be made, e.g., the component might be '…very risky, costly, or theoretically interesting' (Strayhorn 1987).

3.6 Pill Placebo Conditions

Citing the problems noted above with placebo conditions for psychotherapy research, Klein (1997) and others have argued for wider use of pill placebo controls for psychotherapy trials. In this strategy, the experimental psychotherapy would be compared with a pill placebo condition. This strategy was used in the landmark NIMH Treatment of Depression Collaborative Research Program (Elkin et al. 1985), which compared cognitive therapy and interpersonal psychotherapy to a placebo/clinical management condition (with imipramine/clinical management as a reference condition).

3.6.1 Advantages. The use of pill placebo conditions in psychotherapy research has a number of potential advantages. First, when carefully implemented, pill placebos provide control for participant expectations and many demand characteristics. Klein (1997) also points out that this approach enables researchers to evaluate psychotherapy effects in the context of a pharmacotherapy-responsive population. This approach avoids problems with 'psychotherapy placebo' conditions, as it compares the effects of a given psychotherapy with those nonspecific effects of a pill placebo. This strategy may also allow clearer evaluation of treatment processes and mediation effects (Crits-Christoph 1997).

3.6.2 Disadvantages. Pill placebo controls require the availability of a widely recognized effective pharmacotherapy for the disorder in question, as the rationale for the pill placebo control requires that participants expect to improve by taking the pill. Thus, a major drawback of this approach is that for many conditions for which psychotherapies may be appropriate, no such reference pharmacotherapy exists. Moreover, because most pharmacotherapies are delivered in conjunction with a minimal supportive psychotherapy condition to foster compliance and enhance retention (conditions which closely resemble placebo psychotherapies), the problems associated with conceiving and implementing minimal or placebo psychotherapies may not be avoided entirely (Carroll 1997).

4. Summary

Just as there is no one 'perfect' clinical trial, there is no one perfect control group. The particular control group selected does determine, however, the nature of the questions that can be addressed by the trial (e.g., 'Does this treatment work' to 'Does this treatment work for the reasons we believe it does' to 'Does this treatment have specific benefits compared with other available treatments?').

Selection of an optimal control condition for a given study thus depends on many factors, including the level of development of the treatment being evaluated, the availability of alternative treatments for the disorder, the severity or natural history of the disorder, and many others (Crits-Christoph 1997). Thus, early in the development of a field, where questions remain about a disorder's natural course, and available treatments are scarce, no-treatment or delayed treatment controls may be appropriate choices. However, as a field becomes more sophisticated and alternative treatments are identified, research questions also become more fully developed and require the use of control groups that can address progressively more complex questions.

See also: Censuses: Demographic Issues; Internal Validity; Psychological Treatments: Randomized Controlled Clinical Trials; Psychological Treatments, Empirically Supported

Bibliography

Basham R B 1986 Scientific and practical advantages of comparative design in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology* **54**: 88–94

Chambless D L, Hollon S D 1998 Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology* **66**: 7–18

Carroll K M 1997 Manual guided psychosocial treatment: A new virtual requirement for pharmacoltherapy trials? *Archives of General Psychiatry*. **54**: 923–8

Critelli J W, Neumann K F 1984 The placebo: Conceptual analysis of a construct in transition. *American Psychologist* **39**: 32–9

Crits-Christoph P 1997 Control groups in psychotherapy research revisited. *Treatment*, *1*, *Comment 1*, posted September 22

Elkin I, Parloff M B, Hadley S W, Autry J H 1985 NIMH treatment of depression collaborative research program: Background and research plan. *Archives of General Psychiatry* **42**: 305–16

Frank J D 1961 *Persuasion and Healing*. Johns Hopkins University Press, Baltimore, MD

Kazdin A E 1980 *Research Design in Clinical Psychology*. Harper and Row, New York

Kazdin A E 1986 Comparative outcome studies of psychotherapy: Methodological issues and strategies. *Journal of Consulting and Clinical Psychology* **54**: 95–105

Klein D F 1997 Control groups in pharmacotherapy and psychotherapy evaluations. *Treatment*, *1*, *Article 1*, posted September 22

Luborsky L, Singer B, Luborsky L 1975 Comparative studies of psychotherapies: Is it true that 'everyone has won and all must have prizes'?*Archives of General Psychiatry* **32**: 995–1007
Nathan P E 1997 Would a pill placebo have redeemed Project MATCH? *Treatment*, 1, *Comment* 3, posted September 22
O'Leary K D, Borkovec T D 1978 Conceptual, methodological, and ethical problems of placebo groups in psychotherapy research. *American Psychologist* **33**: 821–30
Rosenthal D, Frank J D 1956 Psychotherapy and the placebo effect. *Psychological Bulletin* **53**: 294–302
Strayhorn J M 1987 Control groups for psychosocial intervention outcome studies. *American Journal of Psychiatry* **144**: 275–82
Wilkins W 1986 Placebo problems in psychotherapy research: Social-psychological alternative to chemotherapy concepts. *American Psychologist* **241**: 551–6

K. M. Carroll

Clinical versus Actuarial Prediction

Paul Meehl's book *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (Meehl 1954) concluded that the prediction of numerical criterion variables of psychological interest (e.g., faculty ratings of graduate students who had just obtained a Ph.D.) from numerical predictor variables (e.g., scores on the Graduate Record Examination, grade point averages, ratings of letters or recommendation) is better done by a proper linear model than by the clinical intuition of people presumably skilled in such prediction. The point of this article is to review summaries and conclusions subsequent to Meehl's original one and to present evidence that even what can be termed 'improper' linear models (Dawes 1979) often yield predictions superior to human intuition.

1. *Type of Statistical Models*

A proper linear model is one in which the weights given to the predictor variables are chosen in such a way as to optimize the relationship between the prediction and the criterion. Simple regression analysis, where the predictor variables are weighted in order to maximize the correlation between the subsequent weighted composite and the actual criterion, is the most common example. Discriminant function analysis is another example; weights are given to the predictor variables in such a way that the resulting linear composites maximize the discrepancy between two or more groups. Ridge regression analysis, another example, attempts to assign weights in such a way that the linear composites correlate maximally with the criterion of interest in a new set of data.

2. *Review of the Empirical Findings*

Meehl was concerned primarily with the statistical vs. clinical methods for *integrating* information; thus, he compared instances in which both types of prediction had been made on the basis of exactly the same data. (He also insisted that the accuracy of the statistical model should not be checked on the same data on which it was derived—or that the sample size be so large that it will not appear superior owing to capitalizing on chance fluctuations.) Twelve years later, Jack Sawyer (1966) published a review of about 45 studies; again, in none was clinical prediction superior. Unlike Meehl, Sawyer also included studies in which the clinician had access to more information than that used in a statistical model—for example, interviews of people about whom the predictions were made, or interviews by experts who had access to the statistical model information prior to the interview. Such interviews did not improve the clinical predictions. In fact, the predictions were better when the opinions of the interviewers were ignored.

A prototypical study by Carroll et al. (1988) supports Sawyer's conclusion. A Pennsylvania parole board considered about 25 percent of the 743 parolees to be failures within one year of being released, for reasons such as being recommitted to prison, absconding, being apprehended on a criminal charge, or committing a technical parole violation. A parole board interviewer's ratings had predicted none of these outcomes; the largest correlation was only 0.06. In contrast, a three-variable model based on the type of offense that had led to imprisonment, the number of past convictions, and the number of noncriminal violations of prison rules did have a modest predictability, correlating about 0.22, a result consistent with earlier findings that actuarial predictions based on prior record predict with a correlation of about 0.30 across a large number of settings. When parolees were convicted of new offenses, the seriousness of their crimes was correlated 0.27 with the interviewers' ratings of assaultive potential, but a simple dichotomous evaluation of past heroin use correlated 0.46. All parole board interviewers had access to all this statistical information, but did worse.

None of these correlations is particularly high; first, the sample is highly select, being limited to those who have been convicted of a crime; second, not all the parolees who committed crimes were caught, and third, these types of behaviors are not as predictable as we believe they are or would like them to be. The difference in the effectiveness of actuarial vs. clinical prediction is, however, clear.

Moreover, this difference is consistent with comparisons of actuarial vs. clinical methods for predicting violence (see, e.g., Werner et al. 1984, Monahan 1997).

An important qualification: the best prediction in general is that neither violence nor criminal behavior will be repeated. Although the general 'base rate' prediction is that people will *not* repeat problems, judges—professional and nonprofessional alike—have a bias to believe that repetition is common. The studies show that judgments about who is more likely than whom to repeat are much better made on an actuarial than on a clinical basis.

After his book had been out about 30 years, Meehl (1986) was able to conclude. 'There is no controversy in social science which shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one.' Since that time, even more evidence has been accumulating in favor of Meehl's generalization and practical conclusion; 110 studies that Dawes et al. (1989) reviewed favored it. Subsequently, Grove and Meehl (1996) published a meta-analysis involving even more studies. Their conclusion (Grove and Meehl (1996 p. 293) was, 'The *clinical method* relies on human judgment that is based on information contemplation and, sometimes, discussion with others (e.g., case conferences). The *mechanical method* involves a formal, algorithmic, objective procedure (e.g., equation) to reach the decision. Empirical comparison of the accuracy of the two methods (136 studies over a wide range of predictors) shows that the mechanical method is almost invariably equal to or superior to the clinical method.'

There are four logical relationships possible between the information set on which a clinician or expert makes a prediction and information set on which a formal model (e.g., equation) is based. These sets may be identical, which was the original requirement for a comparison in Meehl's 1954 book. Or one information set may be a subset of the other; in the studies reviewed by Sawyer and in almost all subsequent studies having this structure, the set on which the model is based is the subset of the set available through the clinician (who, for example, is allowed to supplement a sparse information set with an interview). In the field of psychology, the model prediction has always been superior. There are, however, exceptions in the field of medicine, in the situations where the clinical physician has access to more information than is used by the model (*not* in situations where the inputs are identical). Even in these, however, the models may be modified on the basis of interviewing the clinicians themselves to 'distill' what they are responding to, so that once again the statistical prediction becomes superior. For example, a predictive system termed APACHE-II did not predict as well as physicians who would survive in emergency wards, yet the modified model termed APACHE-III did in fact predict better than physicians who would survive the first 24 hours (Knaus et al. 1991). Another possible relationship is that the information sets are overlapping, which has not been studied much.

3. *Explanation for the Empirical Findings*

Why the consistent results? The answer involves an understanding of which factors are favorable to the linear model and which disfavor the intuitive prediction. Whatever additional factors there may be that disfavor the model or favor the clinician are outweighed by the former types of factors in almost all contexts.

First, consider the statistical prediction. Each instance to be predicted is characterized in terms of its aspects that allow its location in a category or along a dimension. These categories and locations have been found in the past to be *in general* predictive. Aspects that have no such predictive power are automatically ignored. The model involves the weighting of the predictive aspects. Moreover, these aspects often are diverse, e.g., an undergraduate grade point average, a number of prison violations, an instance of violence that led to hospitalization (even a rating routinely made on the first day of hospitalization). The statistical model automatically makes the predictive variables comparable, by assigning weights to integrate them.

Now, consider the drawbacks of attempting to make an intuitive integration of the information. Suppose even that we are not interested in making an optimal prediction, just in integrating information from diverse and incomparable dimensions. Suppose, for example, we were deciding between two jobs where the important considerations are pay and enjoyment of the activities involved. It may be very easy, knowing our preferences for different types of activities, to judge which job will be more enjoyable. But now job A pays more than job B, but is less enjoyable. Which should be chosen? We must weight the two dimensions—at least implicitly—if we are to make a choice. What psychologists have found (e.g., Svenson 1992, Langer 1994) is that in such conflicting dimensions situations people generally search for reasons to dismiss one or another of the dimensions as 'not really that important.' Then, there is no conflict between dimensions, and all that must be done is to make an ordinal judgment, again of the form 'more is better.'

Now, consider the additional complications of trying to decide on an intuitive basis which of two instances is more predictive of something. While we might at least have some insight—explicit or implicit—into how we assess differences and how much to care about them, we often have less insight into how to assess differences and to weight them in order to predict. (Feedback observing *all* outcomes without being affected by our judgment—e.g., of success or failure—is necessary, but not sufficient; see Einhorn and Hogarth (1978).)

As noted, statistical integration has the advantage that weighting obviates the problem of conflicting dimensions. The question then arises of whether the weighting system need be optimal in order to maintain the advantage. The answer—based on mathematical

considerations, simulations, and empirical investigation—is no. The robust result is that so long as the dimensions are weighted in the correct direction, *ad hoc* weighting (e.g., 'intuitive weighting,' unit weighting, or even weights chosen according to some random sampling scheme) yields results that are close enough to those provided by optimal weighting that the resulting linear composites still outperform clinical intuition, especially when the predictive variables tend to be positively correlated with each other. Composites based on such nonoptimal weights have been termed 'improper linear models' (Dawes 1979). In fact, not only may such models yield predictions similar enough to optimal models that they outperform clinical judgment, but they may even outperform optimal models on cross-validation, because they are not subject to the 'overfitting' problem that plagues many optimal models. For example, the 'robustness' of unit-weighted models on cross-validation has been noted as far back as the 1930s by Wilks (1938). The mathematical rational for this robustness of unit weighting has been provided by Wainer (1976) and Wainer and Thissen (1976).

In fact, improper models such as unit-weighted models may be particularly advantageous in situations where models developed in one context are to be applied in a slightly different context—which many of us (Dawes 1997) believe to be the norm in social science, as opposed to applying a model developed on one sample of a particular population to another sample drawn from this exact same population. ('Cross-validation' actually refers to such a subsequent application to a sample from the exact same population; a far more descriptive term would be simply 'validation,' where what is commonly termed a 'validation' sample should be termed a 'development' sample. The point here is that when we move across contexts that vary to some—perhaps known perhaps unknown—degree, even the estimate based on what is standardly termed 'cross-validation' may be overly optimistic.)

An empirical overview of the success of improper models in general is provided by Dawes (1979). Perhaps the most famous example of an improper model is that provided by Goldberg (1965) to predict a diagnosis of psychosis vs. neurosis by using MMPI profiles. The unit-weighted composite obtained by adding together three scaled scores indicating psychosis (*L*, *Pa*, *Sc*) and subtracting two indicating neurosis (*Hy*, *Pt*) not only 'outperformed all diagnosticians' (Goldberg 1965, p. 24) but was more stable across subsamples than were nonlinear complex scores—and it was not for want of trying enough of the latter.

Later, Dawes and Corrigan (1974) demonstrated that improper weights (in fact two weighting systems chosen on random bases except for the direction of the weights) not only outperformed clinical judgment in the MMPI diagnosis problem and several others, but did as well as the linear composites based on the diagnostic experts' judgments. Unit weighting was even better.

4. Conclusions and Recommendations

Thus, 'the whole trick is to know what variables to look at and then to know how to add' (Dawes and Corrigan 1974, p. 105). Of course, there are some contexts where configural, multiplicitive, or even more complicated models are more appropriate than are simple additive models; such models may, for example, be found in the area of predator–prey population dynamics. But the decision making discussed here involves a prediction of important human outcomes where—although the human experience may be complex—what can be best distilled from it are simple predictive variables where more (or less) is better (e.g., test scores, indicators of past performance, past criminal or psychiatric record). Not only do we find such simple monotone relationships in our studies, but we search for such relationships and tend to code our social world in terms of variables capturing monotone relationships to what is important to us. (Occasionally, a predictor variable has a single-peaked relationship to the criterion, as when moderate aggression is more desirable in a business person; such a variable is easily transformed into a monotone variable by evaluating distance from the ideal.) When there are strong scientific reasons for hypothesizing a complex model, naturally they should not be ignored. In the absence of such theory or evidence, however, the claim that it is possible to construct a valid nonadditive model 'in our head'—particularly one representing some sort of valid, ineffable intuition—is pure hubris.

In contrast, an understanding of the research reviewed in this article leads to 'an awareness of the modest results that are often achieved by even the best methods, [an awareness which] can help to counter unrealistic faith in our predictive powers and our understanding of human behavior. It may well be worth exchanging inflated beliefs for an unsettling sobriety, if the result is openness to new approaches and variables that ultimately increase our explanatory and predictive powers' (Dawes et al. 1989, p. 1673).

For the most recent survey and analysis, see Swets et al. (2000).

See also: Clinical Psychology: Validity of Judgment

Bibliography

Carroll J S, Werner R L, Coates D, Galegher J, Alibrio J J 1988 Evaluation, diagnosis, and prediction in parole decision making. *Law and Society Review* **17**: 199–228

Dawes R M 1979 The robust beauty of improper linear models in decision making. *American Psychologist* **34**: 571–82

Dawes R M 1997 Qualitative consistency masquerading as quantitative fit. In: Dall Chiara M L, Doets K, Mundici D

(eds.) *Structures and Norms in Science*. Kluwer, Dordrecht, The Netherlands, pp. 387–94
Dawes R M, Corrigan B 1974 Linear models in decision making. *Psychological Bulletin* **81**: 95–106
Dawes R M, Faust D, Meehl P E 1989 Clinical versus actuarial judgment. *Science* **243**: 1668–74
Einhorn H J, Hogarth R M 1978 Confidence in judgment: persistence in the illusion of validity. *Psychology Review* **85**: 395–416
Goldberg L R 1965 Diagnosticians vs. diagnostic signs: the diagnosis vs. neurosis from the MMPI. *Psychological Monographs* **79**: 1–28
Grove W M, Meehl P E 1996 Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical–statistical controversy. *Psychology, Public Policy, and Law* **2**: 293–323
Knaus W A, Wagner D P, Lynn J 1991 Short-term mortality predictions for critically ill hospitalized adults: science and ethics. *Science* **254**: 389–94
Langer E 1994 The illusion of calculated decisions. In: Schank R, Langer E (eds.) *Beliefs, Reasoning and Decision Making: Psycho-Logic in Honor of Bob Abelson*. L. Erlbaum, Hillsdale, NJ, pp. 33–53
Meehl P E 1954 *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press, Minneapolis, MN
Meehl P E 1986 Causes and effects of my disturbing little book. *Journal of Personality Assessment* **50**: 370–5
Monahan J 1997 Clinical and actuarial predictions of violence. In: Faigman D, Kaye D, Saks M, Sanders J (eds.) *Modern Scientific Evidence: the Law and Science of Expert Testimony*. West, St. Paul, MN, Vol. 1, pp. 300–18
Sawyer J 1966 Measurement and prediction, clinical and statistical. *Psychological Bulletin* **66**: 178–200
Svenson O 1992 Differentiation and consolidation theory of human decision making: a frame of reference for the study of pre- and post-decision processes. *Acta Psychologica* **80**: 143–68
Swets J A, Dawes R M, Monahan J 2000 Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest* **1**: 1–26
Wainer H 1976 Estimating coefficients in linear models: it do not make no nevermind. *Psychological Bulletin* **83**: 213–7
Wainer H, Thissen D 1976 Three steps toward robust regression. *Psychometrika* **41**: 9–34
Werner P D, Rose T L, Yesavage J A, Seeman K 1984 Psychiatrists' judgments of dangerousness in patients on an acute care unit. *American Journal of Psychiatry* **141**: 263–6
Wilks S S 1938 Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika* **8**: 20–6

R. M. Dawes

Clitics, Linguistics of

Clitics straddle the boundary between words and affixes because they have some properties of each. Therefore clitics pose a quandary for the longstanding division of labor in grammar between syntax and morphology. Clitics can be found in nearly every language; however, they may be hard to recognize, because they may be disguised as contracted forms, for instance. Clitics also exhibit some unique properties that have been especially difficult to account for in grammar. For that reason, they help linguists to reformulate hypotheses about the grammatical structure that underlines language (Nevis 2000).

1. *Illustration of Clitics*

Grammatical analysis offers a basic distinction between words and affixes: words are independent elements used for phrase and sentence formation (see *Word, Linguistics of*), whereas affixes are word-building units that attach to roots and stems (see *Morphology in Linguistics*). Problematic for this division are two sorts of objects: those that form phrases but are not fully independent, and those that help to build words but exhibit a loose attachment that typical affixes do not. Both kinds have been labeled *clitics*, a term derived from Ancient Greek *enklitikón* 'leaning on a previous (word),' which in various guises has been used in grammatical description for two millennia.

As an illustration of clitics, consider English contracted forms, such as *'ve* in *you've*. It lacks the independence of *have*, but otherwise acts like a full verb by helping to form sentences and by showing agreement with the subject. If *'ve* is a verb, then it is a verb pronounced without a vowel (since the *e* is silent here)—can a word in English consist of a single pronounced consonant? While most contractions in English and in other languages qualify as clitics, not all do.

A different sort of example is the English possessive *'s* (see *Possession (Linguistic)*, and Jespersen 1922). Unlike other affixes in English, it attaches to an entire phrase, as in *the King of Morocco's death*, where the *'s* is affixed to the phrase *the King of Morocco* rather than just to the word *Morocco* (that is, this is not Morocco's death). If *'s* is to be treated as an affix, then grammarians will have to adopt an expanded view of affixation that includes such phrasal affixes.

The study of clitics is important because their position on the edge of the word/affix distinction sheds light on the complexities involved in determining how languages are constructed. The passage in Sect. 3 below will survey some of the unique properties attributed to clitics that have forced linguists to accommodate new facts into their theories of grammar.

2. *Types of Clitics*

Clitics have traditionally been cited according to the way they attach: they may be *enclitics* (attaching leftward onto a host), *proclitics* (attaching rightward), or *endoclitics* (attaching inside the host). In Spanish

me 'me' illustrates both the enclitic and the proclitic options: when *me* follows its host word, as in *díga-me* 'tell me,' it functions as an enclitic, but when it precedes the host as in *no me díga* 'don't tell me' it is a proclitic. Most scholars assume that clitics attach in a very loose fashion to their hosts, and therefore must occur outside of regular affixes. Following that reasoning, proclitics precede any and all prefixes, and enclitics follow all suffixes. Endoclitics are so rare—if they exist at all—that specialists have examined any putative cases closely to see whether the positing of endoclitics can be avoided altogether.

In technical descriptions of languages, a category 'clitic' remained commonplace throughout the twentieth century, but seldom had it been incorporated into grammatical theories until the 1970s, when the phenomenon was surveyed by Zwicky (1977). Zwicky's early investigations established three types of clitics: simple clitics, special clitics, and bound words. In later work, bound words and special clitics merged, producing two categories: simple and special clitics.

Simple clitics can be substituted for full words. The simple clitic is thus considered a reduced version of an independent word, sharing its meaning and manifesting a similar pronounciation. An example is the pronoun *them* in a sentence like *we like them*, which is often reduced in casual speech to *we like 'em*. The clitic form *'em* is relatable to full form *them* and the clitic's distribution in a sentence is a subset of the distribution of the full word, with the contraction occuring in most places the nonclitic is found, but when the pronoun occurs under emphasis or in isolation, clitic *'em* is not permitted and the full form is used. For example, in response to the question *Who do you like better, us or them?*, the full word *them* is an acceptable answer, but the contraction *'em* is not. Although simple clitics might be related to their free forms, they are not mere reductions due to fast speech: in this regard note that the English simple clitics like *'em* are used in slow speech as well.

Special clitics, on the other hand, either lack a freeform equivalent or show some special syntax (to be discussed in the next section). The English possessive *'s* lacks a nonclitic alternative, therefore it constitutes a special clitic. As illustrated below, the Tagalog pronouns are restricted to the second position of the sentence, so they constitute special clitics in Zwicky's scheme.

3. *Unique Properties of Clitics*

Accommodating clitics in theories of grammar has proven challenging. Initial investigations into clitics focused on their status in between words and affixes and pondered whether the notion clitic should be viewed as a special unit of grammar (distinct from the affix and the word) or as a sort of deviant affix or deviant word. Although early views opted for separate treatment, the debate persevered through the 1990s, with some scholars arguing for affixal status and others for word status. Positing the clitic as a unit distinct from the word and the affix has led scholars to attribute to it a special set of properties (Zwicky 1977, Spencer 1991, Katamba 1993). Usually the main thrust of this approach is to present a scheme locating and attaching the clitic in the appropriate place in a phrase and attaching the clitic to one of two neighbors.

Unique characteristics such as clitic doubling and special syntax have been claimed to be shared with neither affixes nor words. Such special properties help highlight differences between clitics on the one hand and words or affixes on the other, or to enrich linguists' understanding of the syntactic patterns of pronouns and other word classes. Some unique properties are particular to clitic pronouns, and in the last quarter of the twentieth century many scholars have devoted their efforts to the study of these aspects of clitic pronouns. In particular, the clitic pronouns in Spanish and French have attracted considerable attention because the pronouns may occur in places where regular nouns may not. French and Spanish direct objects follow the verb if they are noun phrases (in French *tu as vu le chien* 'you have seen the dog') whereas they precede the verb if they are pronouns (*tu l'as vu* 'you have seen it,' literally *you it have seen*). There are other differences as well (van Riemsdijk 1999).

Clitic doubling occurs when a pronoun appears in the same clause as the noun phrase to which it refers. For example, in certain South American Spanish varieties, the proclitic pronoun *lo* can be found together with the direct object:

Lo vimos a Juan
Him saw to Juan
'We saw John'

Normally pronouns replace noun phrases rather than repeat them, so this behavior is noteworthy.

Another pattern unique to clitics is *Wackernagel's Law*, which refers to the restriction of elements to a slot immediately after the first unit of a phrase or clause; that initial unit can be a phrase or word (Steele 1977, Nevis et al. 1994). An example of Wackernagel clitics comes from Tagalog. In Tagalog, pronouns and certain adverbs must occur after the first word or phrase of the sentence. Usually this is the verb, but it can also be the negative *hindi* or some other word. The following sentences offer several clitic pronouns and particle clitics (*na* 'already,' *lamang* 'only,' and politeness marker *po*; such 'particle clitics' function as adverbs in Tagalog), which are highlighted to indicate the second position in the sentence.

Nakita ***ko*** ***na*** ***siya***
Seen I already her/him
I have already seen her/him'

Hindi ***ka*** ***niya*** *kapatid*
Not you his/her sibling
'You aren't his/her sibling'

Tatlo ***lamang*** ***po*** ***sila***
Three only POLITE them
'There are only three of them'

The Tagalog clitics must remain in the second slot in the sentence, no matter what comes before them (verb, negative, or numeral).

4. *Clitics vs. Leaners*

Studies on clitic elements must establish that the objects under investigation exhibit mixed status between affixes and words. Diagnostic tests can be used to identify properties of clitics, words and affixes (Zwicky 1977). Because the phonological attachment of a clitic to its host cannot be assumed and must be proved, it is important to distinguish clitics from *leaners* (also called *quasiclitics* or *semiclitics*), which are simply unstressed words. Although clitics also usually lack stress, they nevertheless show clear evidence of being part of a word's pronunciation. By contrast, a leaner is merely unstressed and does not become part of another word's pronunciation.

For example, the contraction of English *is* to *'s* entails a change in pronunciation according to the last sound of the word it attaches to. Accordingly it is pronounced *z* at the end of *dog* but *s* after *cat*:

The dog's barking.
The cat's purring.

This is the behavior of the plural suffixes as well, which has two pronunciations, *z* and *s*, under the same conditions as the contracted form of *is*:

The dogs are barking.
The cats are purring.

Thus, contracted *'s*, which is a clitic, and plural *s*, which is a suffix, share a pattern of pronunciation that depends crucially on the last sound of a word. This fact of pronunciation distinguishes clitic *'s* from a word that is merely unstressed, as demonstrated by the pronunciation of the word *who* when used to introduce a relative clause (e.g., *the one who got elected*)—it is usually pronounced with a reduced vowel in casual conversation (that is, more like *huh* and not with full vowel *u*). Here unstressed *who* is considered a leaner rather than a clitic because it does not attach phonologically to a neighboring word as does contracted *'s*.

It is not always easy to decide what is a clitic and what is not. When speakers contract *want to* to *wanna*, investigators determine whether *wanna* can be derived in a principled way from *want* plus *to* using regular rules of grammar. It it can, then *wanna* may contain a clitic form of *to*. However, if additional rules are needed to handle this case in an idiosyncratic way, then the derivation is not considered valid and *wanna* might instead be considered a single indivisible word rather than a composite of *want* and *to*. In English, *wanna* is not just a matter of fast speech reduction because it can convey a difference in meaning from the sequence *want* plus *to*: a sentence like *Tung Chee-hwa I want to succeed* offers two meanings (*I want Tung Chee-hwa to succeed* and *I want to succeed Tung Chee-hwa*), whereas the sentence *Tung Chee-hwa I wanna succeed* is restricted for most speakers to one meaning (*I want to succeed Tung Chee-hwa*) and resists the other interpretation.

5. *Clitics as Derived Words and Phrasal Affixes*

Many of the scholars researching clitic doubling, and to a certain extent Wackernagel's Law, seem to assume that clitics are wordlike in nature. A number of studies treat clitics as derived words. But even if most clitics are analyzable as derived words, there is nevertheless a small residue of clitics that are better handled as affixes. English possessive *'s* is an example of one such clitic: it is best treated as an affix attaching not to any individual word but to a group of words insofar as it offers no wordlike characteristics beyond phrasal affiliation and clearly patterns with the other inflectional affixes of English in morphology and pronunciation. To the chagrin of theoreticians, both kinds of clitics appear to be required to handle such puzzles as the wordlike contracted auxiliaries of English (e.g., *'ve*) and its phrasal affix possessive.

Clitics can sometimes cluster into longer sequences. These sequences exhibit looser order than affix chains but stricter order than word order is generally capable of. In Tagalog, for example, all pronoun clitics consisting of a single syllable precede all particle clitics, which in turn precede all pronoun clitics consisting of two syllables (e.g., *ko* 'I' precedes *na* 'already,' and *na* precedes *siya* 'her/him.'

Nakita ***ko*** ***na*** ***siya***
Seen I already her/him
'I have already seen her/him'

But among the two-syllable pronouns order is not restricted; *nila* 'they' and *ako* 'me' may occur in either order:

Hindi ***nila*** ***ako*** *nakita* = *Hindi* ***ako nila*** *nakita*
Not they me seen
'They didn't see me'

Similarly, order need not be limited among the particle clitics; *lamang* 'only' and the politeness marker *po* occur in two orders as well.

Tatlo ***lamang po*** *sila* = *Tatlo* ***po lamang*** *sila*
Three only POLITE them
'There are only three of them'

Surprisingly, such clitic sequences have been used to argue for the more wordlike standing of clitics, for the affixal nature of clitics, and for the unique status of clitics. Clearly more work needs to be done in this area.

6. *Clitics in Historical Studies*

In the 1980s and 1990s the focus was on accommodating clitics into current theories of grammar, but previously to this most investigations of clitics had scrutinized historical change and the development of

affixes from separate words. Affixes can arise from several sources historically; one such source is a former word. A word can become less stressed, begin to lean on a neighboring word, and eventually turn into a fully dependent element. This is known as the *agglutination cycle* (see *Grammaticalization*). For instance, the Modern English adjective suffix *ly* developed from a Germanic word that also yielded Modern English *like*. One hypothesis is that full words first cliticize before becoming affixes, as demonstrated in the nearly extinct Balto-Finnic language Livonian, which has a comitative suffix *-ks* meaning 'with' (e.g., *suu-ks* 'with the mouth') that came from a former word *kansa* 'with' (that is, this word would originally have been the phrase **suun kansa*, literally 'mouth with'), which cliticized into *-kas* before becoming suffix *-ks*. However, it is by no means evident that cliticization is a necessary stage in the development of affixes from words.

Another question is whether the agglutination cycle operates in one direction only. Normally a word weakens into a clitic, and then the clitic fuses as an affix into the host. For example, German *dem* 'the' has fused with *zu* 'to' to form *zum* 'to the.' So if the agglutination cycle is limited to one direction of development (that is, it is unidirectional), then affix sequencing might be a good indication of an earlier word order. Since the affixes would have been separate words previously, they could provide a tool for linguistic rconstruction of word order patterns in earlier stages of a language (see *Linguistics: Comparative Method*).

But exceptions to this scenario have been noted. An affix can loosen and can become a clitic, and a clitic can likewise become a separate word. The modern English possessive is a case in point because it is a clitic derived from a regular suffix in old English. Older English exhibited attachment of the possessive to the head noun in a phrase (as is expected of an affix) rather than to the end of the phrase (as a phrasal-affix type clitic):

Earlier: The King's crown of England
Now: The King of England's crown

An illustration of a clitic turning into a word can be found in the Sami languages (formerly called Lappish), where the original abessive ending developed first into a clitic and then in some varieties of Sami into a separate word. The abessive expresses absence (and is translated as 'without'). In Northern Sami, the abessive is an enclitic on a preceding noun (e.g., *airoj-taga* 'without oars'), whereas in the Enontekiö subdialect it has become an independent adverb, standing even without a preceding noun:

Mun báhcen taga
I go without
'I do without'

Clitics may play a role in word order change. When verbal clitics are attracted to the second position, the Wackernagel slot, that pattern may be generalized to other, nonclitic verbs. If a language had originally placed verbs at the end of a sentence, generalizing this second position pattern would cause verbs to migrate to second position. That is, a subject–object–verb pattern will become subject–verb–object (see *Word Order*). Although the details of this proposal may not be widely accepted, it demonstrates that the role of clitics should not be ignored in the study of word order change.

7. Conclusion

Most, if not all, languages offer at least one clitic element. However clitics themselves do not seem to form a homogeneous class, as illustrated in English by the rather wordlike contracted auxiliary *'ve* and the more affixal possessive *'s*. Clitics cross-cut most word classes (see also *Word Classes and Parts of Speech*) and may appear as adverbs, pronouns, prepositions, auxiliary verbs, and the like. Apparently, the only word classes inaccessible to clitics are nouns and verbs, as well as adjectives (and this is because putative 'clitic nouns' and 'clitic verbs' would be treated separately as *incorporations*; see *Linguistics: Incorporation*).

Apart from the historical studies, much of the literature on clitics in the first three-quarters of the twentieth century was pretheoretical in nature insofar as investigators struggled to understand the fundamentals of what a clitic is. Subsequent debates were often couched in particular theoretical frameworks or focused exclusively on certain kinds of clitics (especially pronominal clitics). Although numerous generalizations have been established, no real consensus had resulted in a single theory of clitics by the end of the twentieth century—probably because clitic phenomena do not constitute a uniform phenomenon. To supplement such argumentation, future research on clitics will have to incorporate evidence from areas such as psycholinguistic experimentation (see also *Psycholinguistics: Overview*) and language disorders as well as acquisition of language (see also *First Language Acquisition: Cross-linguistic*), and will have to better integrate studies on the historical evolution of clitics. But above all, the study of clitics will continue to assist scholars in defining and delimiting related areas such as Phonology, Morphology, and Syntax).

See also: First Language Acquisition: Cross-linguistic; Grammaticalization; Linguistics: Comparative Method; Linguistics: Incorporation; Morphological Case in Linguistics; Morphology in Linguistics; Phonology; Possession (Linguistic); Syntax; Valency and Argument Structure in Syntax; Word Classes and Parts of Speech; Word, Linguistics of; Word Order

Bibliography

Bauer L 1988 *Introducing Linguistic Morphology*. Edinburgh University Press, Edinburgh, UK

Carstairs A 1981 *Notes on Affixes, Clitics, and Paradigms*. Indiana University Linguistics Club, Bloomington, IN
Jespersen O 1922 *Language, Its Nature, Development and Origin*. Allen and Unwin, London
Katamba F 1993 *Morphology*. St Martin's Press, New York
Nevis J 2000 Clitics. In: Booij G, Lehmann C, Mugdan J, Skopeteas S (eds.) *Morphology: An International Handbook on Inflection and Word Formation*. de Gruyter, Berlin, Article 41
Nevis J A, Joseph B D, Wanner D, Zwicky A 1994 *Clitics: A Comprehensive Bibliography 1892–1991*. Benjamin, Amsterdam
van Riemsdijk H (ed.) 1999 *Clitics in the Languages of Europe*. de Gruyter, Berlin
Spencer A 1991 *Morphological Theory: An Introduction to Word Structure in Generative Grammar*. Blackwell, Oxford, UK
Steele S 1977 On the count of one. In: Juilland A (ed.) *Studies Presented to Joseph Greenberg*. Anma Libri, Saratoga, CA
Zwicky A 1977 *On Clitics*. Distributed by the Indiana University Linguistics Club, Bloomington, IN

J. A. Nevis

Closed and Open Systems: Organizational

Closed and open systems refer to whether organizational entities, such as groups and organizations, are viewed as relatively closed or open to their environment. Such a perspective has a profound influence on how organizations are described and studied. When treated as closed systems, organizations are unaffected by their environment, and attention is directed inward to internal structures and behaviors. As open systems, organizations are interdependent with their environment, and focus is outward to how such interaction is managed.

Knowledge of closed and open systems derives from the broad framework of general systems theory (GST) which seeks to explain the structure and behavior of complex wholes called systems (e.g., von Bertalanffy 1956, Miller 1978). This broad metatheory is based on related research from the physical, biological, and social sciences, and seeks to discover general laws which apply to all levels of systems from single cells to societies.

This article shows how GST applies to organizations. It describes the systemic properties of organizations as closed and open systems, and explains how they are structured and managed.

1. Organizations as Systems

Organizational scholars use GST to describe the general properties of organizational systems. This includes defining their constituent parts and how they are inter related to form a system. It also involves identifying different levels of organizational systems and explaining how they interact with each other.

1.1 Organizational Parts, Relationships, and Wholes

A key premise of GST has to do with the definition of a system and how it forms an organized whole. A system is composed of parts and relationships among them. The system provides the framework or organizing principle for structuring the parts and relationships into an organized whole. This makes systems capable of behaving in ways that are greater than merely the sum of the behaviors of their parts, thus leading to the common adage: 'the whole is greater than the sum of its parts.' Organizational scholars have expended considerable effort in identifying the constituent members or subunits of organizational systems and examining relationships among them. They have sought to discover the organizing principals through which the parts are arranged into a coherent whole. For example, job design researchers have identified different elements of jobs; they have shown how they can be combined to affect employee motivation. Group dynamics scholars have spent considerable time addressing issues of group membership and member interaction. They have discovered different ways of structuring groups to perform tasks that members could not do working alone. They have shown how, under certain conditions, groups can outperform individuals, thus leading to various high-performing group designs, such as self-managing teams, quality circles, and cross-functional teams. Similarly, organization theorists have identified the different components of organizations and examined relations among them. They have found different ways to organize the components and relationships for competitive advantage.

1.2 Levels of Organizational Systems

A second principle of GST has to do with the multilevel nature of systems. Systems exist at different levels; the levels exhibit a hierarchical ordering, with a higher level of system being composed of systems at lower levels. For example, societies are composed of organizations; organizations are composed of groups; groups are composed of individuals; and so on. Higher-level systems provide constraints and opportunities for how a system organizes its parts, and the nature of those parts affects the system's organizing possibilities. Thus, to describe a system at a particular level and explain its behavior, it is necessary to look both upward to the higher level system within which it is embedded and downward to its constituent parts.

This multilevel perspective has led scholars to identify different levels of organizational systems, and to focus on understanding them and how they interact with each other. Considerable attention has been directed at specifying appropriate levels of analysis, both for conceptualizing about organizational systems and for aggregating and disaggregating data that apply

to different levels. For example, scholars have developed analytical guides for aggregating data from individuals to devise measures of group and organization functioning. As researchers have developed more comprehensive theories and more powerful analytical methods, they have made finer distinctions among levels of organizational systems, particularly above the organization level. This has led to at least five levels of organizational systems:

(a) individual member, role, or job;
(b) group;
(c) organization;
(d) population of organizations and/or alliance among organizations; and
(e) community of populations and/or community of alliances.

2. Organizations as Closed Systems

Closed systems do not interact with the environment, and consequently their behavior depends largely on the internal dynamics of their parts. They seek to maintain a steady state or equilibrium among their parts while performing goal-directed behaviors. Because the environment is inconsequential for goal achievement, system behavior is highly specified and maximally controlled within the system.

Until the late 1960s, organizational scholars tended to employ a closed-system perspective for studying organizations. Organizational environments were seen as relatively simple and predictable, and thus were not problematic or significant for how organizations behaved. Attention was directed at the internal dynamics of groups and organizations, particularly at how member behavior was controlled to achieve specific objectives and goals. This led to extensive knowledge of organizational control mechanisms and search for the best way to structure organizational systems. For example, at the individual level, job design researchers discovered how to analyze and maximally specify the most efficient way to perform various tasks. At the group level, researchers showed how group structures and processes could contribute to member conformity to group norms. At the organization level, scholars studied a variety of devices for controlling member behavior, such as managerial hierarchy, rules/procedures, and functional design.

3. Organizations as Open Systems

In the late 1960s, organizational scholars began to broaden their focus to external forces affecting organizational systems. This open systems view was fueled by growing applications of GST to the social sciences and by realization that the behavior of organizational systems could not be adequately explained without examining environmental relationships and their effects on the system (e.g., Aldrich 1979, Buckley 1967, Katz and Kahn 1966, Lawrence and Lorsch 1967, Scott 1981, Thompson 1967). It has led to considerable research about organizational environments, their dynamics and effects, and how organizational systems interact with them. An open systems perspective draws attention to how organizations exchange information and resources with their environment, and how the two mutually influence each other. Moreover, it provides a number of useful concepts for understanding how organizations maintain functional autonomy while influencing and adapting to external forces.

3.1 Critical Functions

To survive and prosper, open systems need to perform at least four critical functions:

(a) transformation of inputs of energy and information to produce useful outputs;
(b) transaction with the environment to obtain needed inputs and to dispose of outputs;
(c) regulation of system behavior to achieve stable performance; and
(d) adaptation to changing conditions.

Because these different functions often place conflicting demands and tension on open systems, system viability depends on maintaining a dynamic balance among them. Organizational researchers have devoted considerable time to identifying and explaining how these four functions operate and contribute to organizational survival and effectiveness. This has led to knowledge about how organizations and groups produce products and services through operating and developing different technologies; how they protect their technologies from external disruptions while acquiring raw materials and marketing finished products; how they regulate themselves for stable performance while initiating and implementing innovation and change. This research defines a key role of management in organizational systems as sustaining a dynamic balance among these four functions; one that allows the organization sufficient stability to operate rationally yet requisite flexibility to adapt to changing conditions.

3.2 Information and Resource Flows

As open systems, organizations seek to sustain a cycle of activities aimed at taking in inputs of information and resources from the environment, transforming them into outputs of goods and services, and exporting them back to the environment. This cycle enables organizations to replenish themselves continually, so long as the environment provides sufficient inputs and the organization delivers valued outputs. Considerable research has gone into understanding how organi-

zations manage these information and resource flows. One perspective focuses on how organizations process information to learn how to improve themselves and to relate to their environments. Organizations must be capable of learning from their experiences and of disseminating such knowledge widely if they are to change themselves to respond to emerging conditions. Consequently, research has been directed at organizational learning and knowledge management, particularlyatdiscoveringmechanismstoenhancelearning capability, such as shared databases, groupware, and decision-support systems. Another view concentrates on how organizations compete for resources through managing key resource dependencies. Attention has been directed at how organizations gain access to resources without becoming overly dependent on those who supply them. Still another perspective focuses on how organizations gain legitimacy from environmental institutions, so they can continue to function with external support. Researchers have sought to identify the institutional demands of powerful resource providers, and how organizations respond to them.

3.3 Boundaries

In managing information and resource flows, organizations, like all open systems, seek to establish boundaries around their activities. These organizational boundaries must be sufficiently permeable to permit necessary environmental exchange, yet afford the organization adequate protection from external demands to allow for rational operation. Organizational scholars have devoted considerable attention to understanding the dual nature of organizational boundaries. They have studied various boundary-spanning roles that relate the organization to its environment, such as sales, public relations, and purchasing. They have examined how organizational members perceive and make sense out of environmental input, and how organizational boundaries vary in sensitivity to external influences. Researchers have also identified different strategies for protecting transformation processes from external disruptions while remaining responsive to suppliers and customers.

3.4 Self-regulation

Viewed as open systems, organizations use information about how they are performing to modify future behaviors. Referred to as cybernetics, this information feedback enables organizations to be self-regulating (e.g., Ashby 1956). They can adjust their behavior to respond to deviations in expected performance. To be effective however, organizations must have sufficient diversity of responses to match the variety of disturbances encountered. Thus, as technologies and environments become more complex and uncertain, organizations seek to become more flexible and nimble. They can enhance their adaptive capabilities through such innovations as downsizing, employee empowerment, and re-engineering.

Extensive research has been devoted to understanding how organizations regulate themselves. Using modern information technology, organizations have developed a variety of methods for setting goals, obtaining information on goal achievement, and making necessary changes. For example, database technologies enable organizations to store large batches of information and to logically connect events, actions, and outcomes. Employees can access and revise this information through on-line, common databases.

3.5 Equifinality

As open systems, organizations display the property of equifinality. They can achieve objectives with varying inputs and in different ways. Consequently, there is no one best way to design and manage organizations, but there are a variety of ways to achieve effective performance. This contrasts sharply with a closed system perspective which seeks the one best way to structure and control organizations.

Organizational researchers have devoted considerable attention to identifying different choices for designing and managing organizations. This has led to a virtual revolution in new organizational designs aimed primarily at making organizations leaner, more flexible, and more responsive to human resources. For example, traditional bureaucratic structures that emphasize efficiency and control have increasingly been supplanted with designs that emphasize flexibility and innovation, such as matrix organizations, horizontal organizations, network organizations, and virtual organizations. Moreover, researchers have developed contingency theories which specify under what technological, environmental, and human conditions different organizational designs are likely to be most successful. For example, these newer, more innovative designs are best suited to situations where technologies are complex, environments are unpredictable, and people have high growth needs.

4. Promising Trends

Organizational researchers have recently applied systems concepts to understand how organizations can adapt to rapidly changing, unpredictable environments. They have borrowed heavily from complexity theory which seeks to explain the behaviors and changes that can occur when the parts of complex systems interact (e.g., Holland 1995, Brown and Eisenhardt 1998).

Such systems tend to be self-organizing in response to environmental feedback; they can change in nonlinear and dynamic ways and can invent entirely new responses to external forces. These concepts provide a dynamic, change-oriented perspective on organizations. They help to explain how organizations can restructure themselves continually to keep pace with fast-changing environments.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Organization: Overview; Organizational Climate; Organizational Decision Making; Organizations and the Law; Organizations, Sociology of

Bibliography

Aldrich H E 1979 *Organizations and Environments*. Prentice-Hall, Englewood Cliffs, NJ

Ashby W R 1956 *An Introduction to Cybernetics*. Wiley, New York

Brown S L, Eisenhardt K M 1998 *Competing on the Edge*. Harvard Business School Press, Boston

Buckley W 1967 *Sociology and Modern Systems Theory*. Prentice-Hall, Englewood Cliffs, NJ

Holland J H 1995 *Hidden Order: How Adaptation Builds Complexity*. Addison-Wesley, Reading, MA

Katz D, Kahn, R L 1966 *The Social Psychology of Organizations*. Wiley, New York

Lawrence P R, Lorsch J W 1967 *Organization and Environment*. Harvard University Press, Boston

Miller J G 1978 *Living Systems*. McGraw-Hill, New York

Scott W R 1981 *Organizations: Rational, Natural, and Open Systems*. Prentice-Hall, Englewood Cliffs, NJ

Thompson J O 1967 *Organizations in Action*. McGraw-Hill, New York

von Bertalanffy L 1956 General systems theory. *General Systems* **1**: 1–10

T. G. Cummings

Co-constructivism in Educational Theory and Practice

Ever since Piaget's dynamically Kantian epistemology, it has been widely accepted as a pervasive assumption that learning is a constructive process. In contrast to the epistemological assumption of empiricism that what we know is a direct reflection of ontological reality, learning is considered as an active construction of knowledge. Learners, as they strive to make sense of their world, do not passively receive stimulus information matching independent physical structures, but genuinely interpret their experience by (re)organizing their mental structures in increasingly sophisticated ways, while interacting with the physical and symbolic environment. According to Piaget and most of his successors in cognitive, developmental, and educational psychology, this process of adaptive and viable reality construction is enabled and constrained both by biologically grounded structures (the strength and scope of which, however, are not yet well known) and by the already existing preknowledge (concepts, operative schemas, and structures) of the individual.

Even though the constructivist assumption makes some traditional problems in both psychology and education easier to solve, it also raises some new ones. An important problem is how we can think of achieving intersubjectivity. How can individuals who personally construct their knowledge independently of each other come to the same or similar cognitive structures? How can we *share* a knowledge of our culture if people are conceived of as being solo learners, and socially isolated Robinson Crusoe figures?

One striking answer, which at the same time challenges traditional (Western) epistemological constructivism, stems from symbolic interactionist (Mead 1934) and sociocultural theory (Vygotsky 1962). It claims that learning is fundamentally a *social activity*. Learning and enculturation are not bounded by the individual brain or mind but are intrinsically social endeavors, embedded in a society and reflecting its knowledge, perspectives, and beliefs. People construct their knowledge, not only from direct personal experience, but also from being told by others and by being shaped through social experience and interaction. The basis of personal development and enculturation, thus, is not the socially isolated construction of knowledge, but its co-construction in a social and cultural space. Or, as Bruner puts it: 'Most learning in most settings is a communal activity, a sharing of the culture. It is not just that the child must make his knowledge his own, but that he must make it his own in a community of those who share his sense of belonging to a culture' (Bruner 1986, p. 86). Knowledge, from this perspective, is no longer seen as solely residing in the head of each individual, but as being distributed across individuals whose joint interactions and negotiations determine decisions and the solution of problems.

1. Concept and Process of Co-construction

No precise and widely accepted definition of the concept and process of co-construction can be found in psychological or educational literature. What has been provided is very diverse and depends on the theoretical context in which it is embedded. Differences can be found with regard to at least three aspects:

(a) the social type of discourse eligible to be called co-constructive: mother–child dialog, peer interaction, teacher–student interaction, learning in teams, computer-supported collaborative work;

(b) the psychopedagogical processes involved in productive co-constructive activity: productive dialog such as exploratory talk and collective argumentation, collaborative negotiation after sociocognitive conflict or as a process of reciprocal sense-making, joint construction of a shared understanding, elaboration on mutual knowledge and ideas, giving and receiving help, tutoring and scaffolding;

(c) the expected outcomes of collaboration: taken-as-shared individual vs. socially shared cognitions; convergence and intersubjectivity; academic task fulfillment, student motivation, and conceptual development; effects on skills in listening, discussion, disputation, and argumentation.

Common to most theoretical contexts of co-constructivism is the implication of some kind of collaborative activity and, through joint patterns of awareness, of seeking some sort of convergence, synthesis, intersubjectivity, or shared understanding, with language as the central mediator. Theorists, moreover, largely converge in the adopted methodology of microgenetic analysis that has been used to examine the inherently fragile processes of co-construction.

1.1 (Neo-)Piagetian Perspective

In a (neo-)Piagetian framework, true dialog becomes possible and facilitates the individual cognitive construction of operatory structures when children are able to take other persons' points of view into consideration and when they are able to resolve sociocognitive conflict. Although regarded (by the early Piaget) as a developmental factor, social interaction—specifically, peer interaction—remains more of a catalyst for individual cognitive development. According to studies carried out by co-workers of Piaget (e.g., Doise and Mugny 1984, Perret-Clermont et al. 1991), social factors, such as the need to deal with conflicting perspectives, can have a productive impact on cognitive behavior. For example, in a Piagetian conservation task, pupils more easily progressed to a subsequent level of development after having been confronted by contradictory judgments given by an adult or another child.

1.2 (Neo-)Vygotskyan context

In Vygotsky's cultural-historical view of development as a process of meaningful appropriation of culture, the interactive foundation of the cognitive is at the core of the developmental process. In contrast to Piaget's view, however, 'the constructivist principle of the higher mental functions lies outside the individual—in psychological tools (such as 'language') and interpersonal relations' (Kozulin 1998, p. 15). According to Vygotsky's claim that interpersonal interactions on a social plane serve as prototypes for intrapersonal processes, i.e., for functions to be internalized, co-construction can be seen as (asymmetrical) adult–child interaction, or interaction between a child and a more capable peer, in the 'zone of proximal development.' 'What a child can do today in co-operation, tomorrow he will be able to do on his own' (Vygotsky 1962, p. 87). The quality and development of higher order thinking is prepared by the co-constructive patterns and distinctive properties of social interaction. Meaningful new learning emerges by embedding mental functions (like logical argumentation, proof, reflection, or problem solving) into specific forms of goal-directed interaction and dialog, where more knowledgeable individuals tailor a task in such a way that a child can successfully coperform it. The acquisition of a new concept or mental function becomes progressively more skillful as the child learns to respond in gradually more sophisticated and personally more meaningful ways to the co-constructive, sense-mediating context of adult regulations, and eventually takes over responsibility for his or her own learning.

1.3 Perspective of Situated and Socially Shared Cognition

Situated learning theory views human cognition as being embedded in and inseparable from specific sociocultural contexts. The goal of learning is to enter a community of practice and its culture, i.e., to learn, like an apprentice, to use tools as practitioners use them (Brown 1989). As a process, learning takes place through the interaction and transaction between people and their environments. Co-construction, from a situated cognition perspective, can be seen as having two or more individuals collaboratively construct a shared understanding, or a solution to a problem, which neither partner entirely and necessarily possesses beforehand (Chi 1996). In a widely quoted definition proposed by Roschelle and Teasley (1995): 'Collaboration is a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem' (p. 70).

At the heart of this concept of co-construction are two coexisting activities: collaboratively solving the problem, and constructing and maintaining a *joint problem space*. Both activities require constant negotiations and recreations of meaning, i.e., trying to find out what can reasonably be said about the task in hand, and occur in structured forms of conversation and discourse utilizing language and physical actions as their most important mediators and resources. With the use of symbolic tools, it becomes possible for the conversants to express and objectify meanings, to

compare and change them deliberately, to exchange and renegotiate them with others, and to reflect on the organization of judgments and arguments (see van Oers 1996). However, as observational studies show, co-constructive learning is hardly a homogeneous but an inherently fragile process in the service of convergence and mutual intelligibility. The achievement of a shared conceptual structure cannot be reliably predicted, nor does the iterative construction of a joint problem space through cycles of displaying, confirming and repairing occur by simply putting two students together. As Roschelle and Teasley (1995, p. 94) remark:

> Students' engagement with the activity sometimes diverged and later converged. Shared understanding was sometimes unproblematic and but oftentimes troublesome. The introduction of successful ideas was sometimes asymmetric, although it succeeded only through coordinated action. These results point to the conclusion that collaboration does not just happen because individuals are co-present: individuals must make a conscious, continued effort to coordinate their language and activity with respect to shared knowledge.

1.4 *Context of Discourse Linguistics: Grounding*

From the perspective of communication or conversation analysis, co-constructive or collaborative learning requires individuals to establish, maintain, and update some degree of mutual understanding. The basic process by which this is accomplished between individuals is called *grounding* (Clark and Brennan 1991). Grounding as a basic form of collaboration means the moment-by-moment coordination and synchronization of the content-specific as well as the procedural aspects and steps of co-constructive activity. There is no need, however, to fully ground every aspect of an utterance. Clark and Brennan (1991, p. 148) frame a pragmatic criterion for grounding: The conversants 'mutually believe that they have understood what [they] meant well enough for current purposes.' Thus, the techniques that are used for grounding are shaped by the goal and the medium of communication. That is, the criterion of grounding and the techniques exploited for its maintenance dramatically change according to the purpose of communication (e.g., planning a party, swapping gossip, or gaining deep understanding) and the constraints of its medium (copresence and visibility in face-to-face communication; sequentiality and reviewability in letter communication, e-mail, or computer-supported collaborative work).

1.5 *Pedagogical Context of Tutoring*

Another aspect of concern for the social nature of learning—and for a crucial way in which it is supported by culture—is instructional dialog or conversation. This term refers to a discursive activity in classrooms that permits the co-construction of meaning between teachers and students, tutors and tutees, the more and the less experienced. Consistent with Vygotsky's theory of the constructive role played by adults in children's acquisition of knowledge, the teacher's goal of assistance can be seen as trying to get the students to share his or her understanding and knowledge. However, because of the asymmetrical distribution of knowledge between teachers and students, understanding might be expected to be less jointly constructed in instructional conversation than it is observed to be in peer-cooperative dialog. Actions that tutors or teachers can take in order to elicit responses, including some co-constructive behavior from a tutee, are, for example, described in literature on reciprocal teaching and on cognitive apprenticeship (Collins et al. 1989). They can be subsumed under two broad categories: (a) modeling, scaffolding, and fading as content-specific ways of providing hints, strategies, and situational forms of coaching and guidance that are tailored to the needs of individual students; and (b) prompting as a more content-neutral invitation by the tutor to elicit elaborations, reflections, and self-explanations from students (Chi 1996).

2. *Pedagogical Facilitation of Co-construction*

The question of how best to support co-constructive learning is concerned with the design of effective collaborative learning environments. Much empirical work has addressed the conditions under which productive collaborative interaction is most likely to occur, and a whole range of possible ways to enhance its quality has been provided. Among the input characteristics that exert a complex influence upon the quality of interaction are: the preparation of the students for collaborative learning (including training for cooperation and discourse prior to the collaborative learning event), the establishment of a culture of dialog and of problem-based learning, group characteristics (composition, size, ability and sex), the goal and incentive structure of the task, and the structuring of group interaction (see, for a review, Webb and Palincsar 1996).

2.1 *Importance of Dialog*

Probably the most important single feature of a culture of collaborative learning is dialog as opposed to, e.g., solo learning and teacher monologs. Emphasis on joint learning and instructional conversation among peers, and between teachers and students, is associated with the internal mediating processes that are essential for an understanding of how co-construction through discourse operates and influences outcomes. The pedagogical cultivation of processes such as negotiation of meaning, reciprocal sense-making, revising one's cognitions in situations of sociocognitive conflict, precise verbalization of reasoning and knowl-

edge, listening to others' lines of argumentation, tuning one's own information to that of a partner, giving and receiving help, or modeling cognitive and metacognitive activities to be internalized by the participating individuals should, thus, be placed at the core of instructional design.

2.2 Support Structures for Collaborative Thinking and Problem Solving

A means of improving the quality of collaborative thinking is explicit process-related and task-related support structures in the learning environment (Pauli and Reusser 2000). *Process-related support structures* refer to the structuring of the interaction through the implementation of scripts for collaboration, such as 'reciprocal teaching', 'scripted cooperation', or 'prompting' for questions and elaborations. These techniques have in common the fact that a set of cognitive and metacognitive strategies which have to be used in a prescribed way is provided. A complementary way of supporting collaborative learning is to provide students with *task-specific support structures* and help. The main goal of task-related assistance, including more or less explicit instructions, domain-specific formats of task representation, and the modeling of strategies, is to scaffold students' domain knowledge construction, understanding, and skill acquisition. What is not yet clear, however, is how much structuring of the interaction is actually beneficial. Ideally, the quality and quantity of guidance and help has to be adjusted to the learners' subjective needs. As Cohen (1994) has pointed out, overstructuring interaction may be counterproductive and have detrimental effects, if it 'prevent[s] students from thinking for themselves and thus gaining the benefits of the interaction' (p. 22).

One promising possibility for making environments more supportive for collaborative learning is to enrich learning situations with technology. Well-designed *computer-based cognitive tools* provide users with both process-related and task-related instruments of thought and communication. As mediational resources and cognitive tools for the representation, negotiation, and modeling of concepts and activities, educational software has the potential—by making conceptual structures and processes visible, accessible, and manipulable on a computer screen—to facilitate processes of sharing understanding and of achieving convergence and intersubjectivity (Reusser 1993).

2.3 Structuring the Role of Teachers

The role of teachers in the co-constructive activities of learners can be described within the didactic framework of 'cognitive apprenticeship' (Collins et al. 1989). According to the ethnographic model in which practices and principles of traditional craftsmanship are applied to cognitive learning activities, teachers, experts, or more capable peers provide guidance and support to learners as they participate as apprentices in authentic and task-related, structured social interactions. As opposed to a transmissionist view of instruction, teachers should provide aid in the intellectual development of students in ways that leave room for negotiation and joint expansion of meaning: (a) as scaffolds and role models for the behavior that students are expected to engage in; (bi) as active participants in learning groups aiming at shaping the group's dialog; (c) as monitors of co-constructive norms in social interactions in which negotiation of taken-as-shared meaning is essential (Webb and Palincsar 1996); (d) as advocates of content-specific standards and of the achievement of convergence and intersubjectivity in understanding and problem solving.

Associated with this shift in the pedagogical orientation of teachers is a shift in the role of learners and the organization of classrooms. In the wake of a view that sees learning essentially as sociocultural interaction, classrooms should develop from aggregations of solo learners to communities engaged in co-constructive learning. That is, individuals should become acculturated members of a culture and community through collaboration and negotiation. Or, as Bruner (1986, p. 123) has put it: culture as 'a *forum* for negotiating and renegotiating meaning and for explicating action ... is constantly in process of being recreated as it is interpreted and renegotiated by its members.'

See also: Cooperative Learning in Schools; Piaget, Jean (1896–1980); Piaget's Theory of Human Development and Education; Situated Cognition: Contemporary Developments; Situated Cognition: Origins; Vygotskij, Lev Semenovic (1896–1934); Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Brown J S, Collins A, Duguid P 1989 Situated cognition and the culture of learning. *Education Researcher* **18**(1): 32–42

Bruner J S 1986 *Actual Minds, Possible Worlds*. Harvard University Press, Cambridge, MA

Chi M T H 1996 Constructing self-explananations and scaffolded explanations in tutoring. *Applied Cognitive Psychology* **10**: 10–49

Clark H H, Brennan S E 1991 Grounding in communication. In: Resnick L B, Levine J, Teasley S D (eds.) *Perspectives on Socially Shared Cognition*. American Psychological Association, Washington, DC pp. 127–49

Cohen E G 1994 Restructuring the classrooms: Conditions for productive small groups. *Review of Educational Research* **64**: 1–35

Collins A, Brown J S, Newman S E 1989 Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In: Resnick L B (ed.) *Knowing, Learning, and In-*

struction: Essays in the Honor of Robert Glaser. Erlbaum, Hillsdale, NJ
Doise W, Mugny G 1984 *The Social Development of the Intellect*. Pergamon Press, Oxford
Kozulin A 1998 *Psychological Tools. A Sociocultural Approach to Education*. Harvard University Press, Cambridge, MA
Mead G H 1934 *Mind, Self, and Society*. University of Chicago Press, Chicago
Pauli C, Reusser K 2000 Cultivating students' argumentation and reasoning in solving mathematical text problems through the use of a computer tool: a video-based analysis of dialogues. *Research Report*. University of Zurich Institute of Education (http://www.didac.unizh.ch)
Perret-Clermont A-N, Perret J-F, Bell N 1991 The social construction of meaning and cognitive activity in elementary school children. In: Resnick L B, Levine L, Teasley S D (eds.) *Perspectives on Socially Shared Cognition*. American Psychological Association, Washington, DC pp. 41–62
Reusser K 1993 Tutoring systems and pedagogical theory: Representational tools for understanding, planning, and reflection in problem-solving. In: Lajoie S P, Derry S (eds.) *Computers as Cognitive Tools*. Erlbaum, Hillsdale, NJ pp. 143–78
Roschelle J, Teasley S D 1995 The construction of shared knowledge in collaborative problem solving. In: O'Malley C E (ed.) *Computer-supported Collaborative Learning*. Springer, Berlin pp. 69–97
van Oers B 1996 Learning mathematics as meaningful activity. In: Steffe L P, Nesher P, Cobb P, Goldin G A, Greer B (eds.) *Theories of Mathematical Meaning*. Erlbaum, Mahwah, NJ pp. 91–113
Vygotsky L S 1962 *Thought and Language*. Harvard University Press, Cambridge, MA
Webb N M, Palincsar A S 1996 Group processes in the classroom. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Macmillan, New York pp. 841–73

K. Reusser

Code switching: Linguistic

Code-switching (CS) refers to the mixing, by bilinguals (or multilinguals), of two or more languages in discourse, often with no change of interlocutor or topic. Such mixing may take place at any level of linguistic structure, but its occurrence within the confines of a single sentence, constituent, or even word, has attracted most linguistic attention. This article surveys the linguistic treatment of such *intrasentential* switching.

In combining languages intrasententially, various problems of incompatibility may arise. The most obvious derive from word order differences: under what conditions, if any, can the boundary between constituents ordered differently in two languages host a switch? Other potential combinatorial difficulties involve mismatches in grammatical categories, subcategorization patterns, morphology, and idiomatic expressions. Systematic examination of the spontaneous speech of bilinguals resident in a wide range of communities suggests, however, that speakers generally manage to circumvent these difficulties. CS tends not to produce utterances that contain monolingually ungrammatical sentence fragments. Discovery of the mechanisms enabling such 'grammatical' CS is the major goal of current research. Central questions include locating permissible switch sites and ascertaining the nature (hierarchical or linear, variable or categorical) of the constraints on switching.

1. Background

Though CS is apparently a hallmark of bilingual communities world-wide, it has only begun to attract serious scholarly attention in the last few decades. Researchers first dismissed intrasentential code-switching as random and deviant (e.g., Weinreich 1953/1968) but are now unanimous in the conviction that it is grammatically constrained. The basis for this conviction is the empirical observation that bilinguals tend to switch intrasententially at certain (morpho) syntactic boundaries and not at others. Early efforts to explain these preferences proceeded by proscribing certain switch sites, for example, between pronominal subjects and verbs or between conjunctions and their conjuncts. However, these particular sites were soon reported to figure among the regular CS patterns of some bilingual communities.

The first more general account of the distribution of CS stemmed from the observation that CS is favored at the kinds of syntactic boundaries which occur in both languages. The *equivalence constraint* (Poplack 1980) states that switched sentences are made up of concatenated fragments of alternating languages, each of which is grammatical in the language of its provenance (see also Muysken 2000). The boundary between adjacent fragments occurs between two constituents that are ordered in the same way in both languages, ensuring the linear coherence of sentence structure without omitting or duplicating lexical content.

That general principles, rather than atomistic constraints, govern CS is now widely accepted, though there is little consensus as to what they are or how they should be represented. Much current research assumes unquestioningly that the mechanisms for language switching follow directly from general principles of (monolingual) grammar. Theories based on this assumption tend to appeal to such abstract grammatical properties as inter-constituent relationships (e.g., government, case assignment) and/or language-specific features of lexical categories (i.e., subcatgorization of grammatical arguments, inherent morphological features).

Since Klavans's (1985) proposal that CS was constrained by structural relations, the formal linguistic theories successively in vogue have each been extended to encompass the data of CS. Di Sciullo et al. (1986), for example, identified the relevant relations as C-

command and government: CS cannot occur where a government relation holds. Replacement of the function of government in standard theory by the notion of feature agreement led to a parallel focus on feature matching in CS studies. The *Functional Head Constraint* (Belazi et al. 1994) adds language choice to the features instantiated in functional and lexical categories, prohibiting CS where a mismatch occurs. A more recent Minimalist proposal (MacSwan 1999) restricts CS at structural sites showing cross-language differences in monolingual features.

This distinction between lexical and functional categories is not new to CS research. It is a hallmark of theories invoking the complement structure of individual lexical items to characterize permissible CS sites (e.g., Joshi 1985) and its sequel, the *Null Theory of CS* (Santorini and Mahootian 1995; see also Bentahila and Davies's *Subcategorisation Constraint* 1983). Perhaps the most detailed model involving the contrast between lexical properties and functional (or 'system') morphemes is the *Matrix Language Frame* model (Azuma 1993, Myers-Scotton 1993). Here, structural constraints on CS result from a complex interaction between a dominant matrix language and the prohibition against embedding 'system' morphemes from the 'embedded' language in matrix language structure.

The assumption that bilingual syntax can be explained by general principles inferred from the study of monolingual grammar has not yet been substantiated. While formal theories of grammar may account well for monolingual language structure, including that of the monolingual fragments in CS discourse, there is no evidence to suggest that the *juxtaposition* of two languages can be explained in the same way. Bilingual communities exhibit widely different patterns of adapting monolingual resources in their code-mixing strategies, and these are not predictable through purely linguistic considerations. The equivalence constraint, as formalized by Sankoff (1998), is a production-based explanation of the facts of CS, which incorporates the notions of structural hierarchy and linear order, and accounts for a number of empirical observations in addition to the equivalent word order characterizing most actual switch sites. These include the well-formedness of the monolingual fragments, the conservation of constituent structure, and the essential unpredictability of CS at any potential CS site. The mechanisms of monolingual and bilingual grammars are not assumed *a priori* to be identical.

2. *Evaluating CS Theories*

There has been remarkably little cross-fertilization among CS theories; indeed, each has been greeted with a host of counter-examples. Testing the fit of competing models against the data of CS should be a straightforward matter since they often make competing predictions. But their disparate assumptions, goals, and domains of application have hindered such efforts. Assessment of the descriptive adequacy of a theory of CS requires that at least two methodological issues be resolved. One involves classification of other-language phenomena, the other, confronting the predictions of the theory with the data of actual bilingual behavior.

It is uncontroversial that CS differs from the other major manifestation of language contact: *lexical borrowing*. Despite etymological identity with the donor language, established loanwords assume the morphological, syntactic, and often, phonological, identity of the recipient language. They tend to be recurrent in the speech of the individual and widespread across the community. The stock of established loanwords is available to monolingual speakers of the recipient language, who access them normally along with the remainder of the recipient-language lexicon. Loanwords further differ from CS in that there is no involvement of the morphology, syntax, or phonology of the lexifier language.

Recent research has shown that borrowing is actually much more productive than implied above (see the papers in Poplack and Meechan 1998). In particular, the social characteristics of recurrence and diffusion are not always satisfied. This results in what has been called, after Weinreich (1953/1968), *nonce borrowing* (Sankoff et al. 1990). Like its established counterpart, the nonce loan tends to involve lone lexical items, generally major-class content words, and to assume the morphological, syntactic, and often, phonological identity of the recipient language. Like CS, on the other hand, nonce borrowing is neither recurrent nor widespread, and necessarily requires a certain level of bilingual competence. Distinguishing nonce borrowings from single-word CS is conceptually easy but methodologically difficult, especially when they surface bare, giving no apparent indication of language membership.

The classification of lone items is at the heart of a fundamental disagreement among CS researchers over (a) whether the distinction between CS and borrowing should be formally recognized in a theory of CS, (b) whether these and other manifestations of language contact can be unambiguously identified in bilingual discourse, and (c) criteria for determining whether a given item was switched or borrowed. Researchers who consider lone other-language items to be CS tend to posit an *asymmetrical* relationship, in which one language dominates and other-language items are inserted (e.g., Joshi 1985, Myers-Scotton 1993). Where the class of CS is (in the first instance) limited to unambiguous *multiword* fragments, both languages are postulated to play a role (Belazi et al. 1994, Sankoff 1998, Woolford 1983). Muysken (2000) admits the possibility of both strategies.

The appropriateness of data is also relevant to evaluating CS theories. The literature on CS largely is

characterized by the 'rule-and-exception' paradigm. Despite the onslaught of counter-examples provoked by successive CS theories, very few have in fact been tested systematically against the data of spontaneous bilingual usage. Instead, both the theories and tests of their applicability tend to be based on isolated examples, drawn from judgments, informant elicitation, and linguist introspection. The relation between such data and actual usage is not known; nor do they permit us to distinguish between the recurrent and systematic patterns of everyday interaction and examples which may be judged 'acceptable' in some sense, but which rarely or never occur.

The equivalence constraint has been verified as a general tendency in Spanish-English (Poplack 1980); Finnish-English (Poplack et al. 1987), Arabic-French (Naït M'Barek and Sankoff 1988), Tamil-English (Sankoff et al. 1990), Fongbe-French and Wolof-French (Meechan and Poplack 1995), Igbo-English (Eze 1998), French-English (Turpin 1998) and Ukrainian-English (Budzhak-Jones 1998) bilingual communities. But most of the voluminous literature on CS, especially of the 'insertional' type, is based on data which represents, properly speaking, lexical borrowing. As only the grammar and word order of the recipient language are pertinent to borrowing, attempts to understand the structure of CS based on a mixture of borrowing and true CS (e.g., Myers-Scotton 1993 and many others) appear unwieldy or descriptively inadequate.

3. *Identifying the Results of Language Contact*

Insofar as CS and borrowing are based on some principled combination of elements of the monolingual (i.e., *unmixed*) vernaculars of the bilingual community, it is important to have as explicit an idea as possible of the nature of these vernaculars before concluding that a code-mixed element is behaving like one or the other. The analysis of code-mixing as a discourse mode requires access to the grammars of the contact languages as they are spoken, and spoken language is characterized by structural variability. In confronting, rather than evading this variability, Sankoff et al. (1990) and Poplack and Meechan (1998) developed a method to compare bilingual structures with the unmixed source languages of the same speakers. Making use of the framework of linguistic variation theory (Labov 1969), the inherent variability of such forms is used to determine their status. If the rate and distribution of, for example, case-marking of the contentious lone other-language items show quantitative parallels to those of their counterparts in the (unmixed) recipient language, while at the same time differing from relevant patterns in the donor language, the lone other-language items are inferred to be *borrowed*, since only the grammar of the recipient language is operative. If they pattern with their counterparts in the (unmixed) donor language, while at the same time differing from the patterning in the unmixed recipient language, the lone other-language items must result from CS.

Quantitative analysis of language mixing phenomena in typologically distinct language pairs shows that lone other-language items, especially major-class content words, are by far the most important component of mixed discourse. These lone items show the same fine details of quantitative conditioning of phonological, morphological, and syntactic variability as dictionary-attested loanwords, both of which in turn parallel their unmixed counterparts in the recipient language (Poplack and Meechan 1998). This tendency is apparent regardless of the linguistic properties of the language pair. This is evidence that most lone items are borrowed, even if only for the nonce, despite the lack, in some cases, of dictionary attestation or diffusion within the community.

4. *Future Directions*

Lack of consensus characterizing the discipline is related to a number of methodological problems. Foremost among them is failure to distinguish code-switching from other types of language mixture, which, despite similarities in surface manifestation, are fundamentally different mechanisms for combining languages. The current state of knowledge suggests that borrowing, nonce or established, is the major manifestation of language contact in most bilingual communities. Its linguistic structure is well accounted for in the traditional language contact literature (Weinreich 1953/1968). Intrasentential CS involving multiword fragments of two or more languages is also attested in some communities. Achievement of consensus on an empirically verifiable characterization of the rules for juxtaposing these fragments within the sentence remains an important goal for CS research. Fit between theories and data could be improved by a broader empirical base. This would permit researchers to situate bilingual behavior with respect to the monolingual vernaculars implicated in language mixing, account for the disparate CS strategies that have evolved in different bilingual communities, and distinguish among incommensurable manifestations of bilingual language contact.

See also: Bilingual Education: International Perspectives; Bilingualism and Multilingualism; Bilingualism: Cognitive Aspects; First Language Acquisition: Cross-linguistic; Second Language Acquisition; Sociolinguistics

Bibliography

Azuma S 1993 The frame-content hypothesis in speech production: Evidence from intrasentential code switching. *Linguistics* **31**: 1071–93

Belazi H M, Rubin E J, Toribio A J 1994 Code-switching and X-bar theory: The functional head constraint. *Linguistic Inquiry* **25**(2): 221–37
Bentahila A, Davies E E 1983 The syntax of Arabic-French code-switching. *Lingua* **59**(4): 301–30
Budzhak-Jones S 1998 Against word-internal code-switching: Evidence from Ukrainian-English bilingualism. *International Journal of Bilingualism* **2**(2): 161–82
di Sciullo A-M, Muysken P, Singh R 1986 Government and code-mixing. *Journal of Linguistics* **22**(1): 1–24
Eze E 1998 Lending credence to a borrowing analysis: Lone English-origin incorporations in Igbo discourse. *International Journal of Bilingualism* **2**(2): 183–201
Joshi A K 1985 Processing of sentences with intrasentential code-switching. In: Dowty D R, Karttunen L, Zwicky A M (eds.) *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*. Cambridge University Press, Cambridge, UK, pp. 190–204
Klavans J L 1985 The syntax of code-switching: Spanish and English. In: King L D, Maley C A (eds.) *Selected Papers from the XIIIth Linguistic Symposium on Romance Languages*. Benjamins, Amsterdam, pp. 213–21
Labov W 1969 Contraction, deletion, and inherent variability of the English copula. *Language* **45**(4): 715–62
MacSwan J 1999 *A Minimalist Approach to Intrasentential Code Switching*. Garland Press, New York
Meechan M, Poplack S 1995 Orphan categories in bilingual discourse: Adjectivization strategies in Wolof-French and Fongbe-French bilingual discourse. *Language Variation and Change* **7**(2): 169–94
Muysken P C 2000 *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press, Cambridge, MA
Myers-Scotton C 1993 *Duelling Languages*. Clarendon Press, Oxford, UK
Naït M'Barek M, Sankoff D 1988 Le discours mixte arabe/français: des emprunts ou des alternances de langue? *Revue Canadienne de Linguistique* **33**(2): 143–54
Poplack S 1980 Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. *Linguistics* **18**(7/8): 581–618
Poplack S, Meechan M (eds.) 1998 *Instant Loans, Easy Conditions: The Productivity of Bilingual Borrowing; Special Issue of the International Journal of Bilingualism*. Kingston Press, London
Poplack S, Wheeler S, Westwood A 1987 Distinguishing language contact phenomena: Evidence from Finnish-English bilingualism. In: Lilius P, Saari M (eds.) *The Nordic Languages and Modern Linguistics*. University of Helsinki Press, Helsinki, pp. 33–56
Sankoff D 1998 A formal production-based explanation of the facts of code-switching. *Bilingualism: Language and Cognition* **1**(1): 39–50
Sankoff D, Poplack S, Vanniarajan S 1990 The case of the nonce loan in Tamil. *Language Variation and Change* **2**(1): 71–101
Santorini B, Mahootian S 1995 Code-switching and the syntactic status of adnominal adjectives. *Lingua* **96**: 1–27
Turpin D 1998 "Le français c'est le last frontier": The status of English-origin nouns in Acadian French. *International Journal of Bilingualism* **2**(2): 221–33
Weinreich U 1953/1968 *Languages in contact*. Mouton, The Hague
Woolford E 1983 Bilingual code-switching and syntactic theory. *Linguistic Inquiry* **14**(3): 519–36

S. Poplack

Coeducation and Single-sex Schooling

Coeducation or single-sex schooling does not only refer to an organizational form of separating or mixing girls and boys, but also refers to such questions as: should there be different goals, curricula, rights, and outcomes for the two genders? Equality is not perceived in all countries and not always an important principle for men and women. For a long time, education was oriented to prepare girls and boys for their different spheres in adult life. Today, equality is the goal, but it is not at all clear whether separation or coeducation is the better way to reach this goal.

1. *A Brief History of Coeducation*

The history of educational systems shows, for most countries, that especially for education beyond elementary levels, single-sex schools have been the preferred form, although coeducation was not unusual. Financial restrictions in mass education forced coeducational schooling, but for ideological reasons, separation was preferred by the authorities. While this is mainly true for Europe, in the United States, soon after establishing public schools, coeducation became the norm. David Tyack and Elisabeth Hansot describe the beginning as 'smuggling in the girls' which led to the adoption of coeducation by gradually moving from 'why to why not' (Tyack and Hansot 1992, p. 47). Even in secondary education, only 12 cities out of 628 reported that they had single-sex high schools at the end of the nineteenth century. The European secondary education was for boys only—especially in Germany; girls received secondary education only on a private basis and were not allowed to attend higher education at all. Prussian universities enrolled women, but not before 1908.

Democratic movements in all countries valued equality of education and therefore pledged for coeducation. The debates emphasized the question of difference or similarity of women and men as human beings. While the assumption of similarity historically was associated with the preference for coeducation, those who postulated differences were divided in advocates and opponents to coeducation. The advocates thought coeducation would ensure that girls and boys themselves would make sure they behaved like girls and boys. The opponents feared that girls and boys together would encourage the loss of their engendered behavior.

In most European countries, the years between 1960 and 1980 brought big educational reforms, and coeducation was one of those (Wilson 1991, p. 203)—although it was more of a side-effect of the reforms, and astonishingly, after the heated discussions in the first half of the twentieth century, a change nearly without debate. Efforts to establish 'schools for all'

would not go along with separation. Poland has had equal opportunity programs since 1965; Sweden since the late 1960s, and most other European countries since the beginning 1970s, except for Greece (mid-1970s) and Spain (1980s). Today all European countries have coeducation with the exception of Ireland, and to a lesser degree, in England and Wales. Some countries still have single-sex classes in some subjects, such as Physical Education and Crafts.

In the US, with its long tradition of coeducation, Title IX of the Educational Amendments in 1972 was the major legal tool in implementing equal education by stating that discrimination on the basis of sex is illegal in any educational program receiving federal funding. This already made clear that coeducation does not necessarily guarantee equality and the absence of sexism.

2. *New Women's Movement and Coeducation*

'Sexism' was the new term to mark inequality for women and men in society. Evaluations of the educational reforms showed differences in attainment between the countries, between social groups and between gender. A close look at processes in the educational system reveals successes as well as failures: women participate much more in higher education but there is under-representation of young women in some educational fields, especially in vocational education and in science. At the same time, women's self-confidence remains at lower levels than men's. The new Women's movement criticized the outcomes of coeducation, and showed that it put women at a disadvantage. Beginning in the 1980s, the critics started a new debate on coeducation. It provided a new research area, although one has to admit that the basis of the literature, even in the late 1990s was still rather thin; the reports are mostly 'anecdotal' (AAUW 1998).

3. *Research about Coeducation and Single-sex Schooling*

Compilations of research show the advantages and disadvantages of coeducation. The Organization for Economic Co-operation and Development (OECD) published in 1986 a report on 'Girls and Women in Education,' and the American Association of University Women (AAUW) released their report: 'How Schools Shortchange Girls,' in 1992. There are at least two fields in which research found relevant gender inequalities:

(a) Coeducation retains the gender-hierarchical division of the world. It strengthens gender-specific interests and seldom encourages thinking and behavior against the traditional gender-stereotypes. The school curriculum serves as a 'hidden curriculum' to ensure those processes.

(b) Classroom interaction processes show unbalanced communication structures. Dominance and forced attention for boys are tolerated, self-effacement by and disregard of girls remain unnoticed.

The central reason for this outcome lies in gender stereotypes and in the gender-division of labor: both are still to such a degree 'normal' that the 'masculine dominance' still remains effective, unquestioned by most people.

The hidden curriculum was a major theme in the OECD report, describing three forms of sex bias:

(a) Picture books, reading schemes and children's literature are characterized by the lack of representation of women and female roles.

(b) The curriculum and textbooks in school tend to be either male-oriented or female-oriented rather than 'bicultural.'

(c) The presentation of men and women in the curriculum shows a picture of the social world that is 'more sexist than reality' (OECD 1986, p. 75).

These results did force many of the political institutions responsible for the educational system to set up new criteria for school books and material. One can admit that there have been some changes, although the situation still has not changed radically.

Quite a lot of mostly qualitative studies deal with the interaction process. Maggie Wilson resumés their results:

> Although teachers often deny that differential treatment of boys and girls exists in classrooms, case studies from Belgium, Spain, Sweden, Greece, West Germany and England and Wales chart the great amount of attention which teachers give young boys, albeit in the form of both praise and censure, and their overtly expressed preference for teaching boys, especially at the upper levels of school and in certain subject areas, such as in Science and Mathematics (Wilson 1991, p. 213).

Research in the USA is very much in line with these findings (Riordan 1990).

Florence Howe, and others, called on the 'myth of coeducation.' She wrote: 'One of the central ideas of coeducation provides a central myth: that if women are admitted to men's education and treated exactly as men are, then all problems of sexual equity will be solved' (Howe 1984, p. X). The valuing of admittance as the only criteria for equality made us forget that there were other qualities in school education that proved to be unequal, e.g., the hidden curriculum and the interaction processes.

Although the AAUW report did not recommend single-sex schooling, it mentioned that research studies would indicate that girls often learn and perform better in same-sex work groups than in mixed-sex groupings (AAUW 1992, p. 130). This, together with many other publications, supported the remaining

women colleges in the USA and also helped to keep, and even to start private girls' schools. In Europe, especially in Germany, campaigns to enroll young women in science and technology subjects experimented with the separation of gender.

At the end of 1997, the AAUW organized a round table entitled 'Separated by Sex—a critical look at single-sex education for girls,' in order to resume the outcomes of the debate about coeducation and single-sex education since their 1992 report. The main results are consistent with research from other countries and could therefore be state-of-the-art for coeducation today:

(a) There is neither a significant correlation between the self-concept of students, nor the gender-stereotyping and coeducational or single-sex schooling.

(b) Students in single-sex schools rarely believe that mathematics and science are specifically male subjects, while students in coeducation schools believe this more often.

(c) These beliefs do not lead to differences in capacities between the students from the two types of schooling.

(d) Better results were found in the US single-sex schools only for 'at-risk-students,' especially Spanish-American girls from low socioeconomic families. The results are very small however, and they are due to the academic orientation of these schools.

(e) Sexism could be found everywhere, not the separation of gender or coeducation but mainly the awareness of teachers is responsible for non-sexist environments.

(f) The majority of students wish to attend co-educational schools.

Altogether, the report called for more complexity in the research design as well as in the theorizing of gender. Both could help to deal with educational questions of gender equity more adequately than just criticizing coeducation and valuing separation.

4. Perspectives

Since the beginning of the new debate on coeducation, working with girls was added to working with boys about masculinity (Connell 1995), and, slowly but surely, is changing the coeducational classroom.

The research reports, as well as the practical experiences in different schools, show that single-sex education can help to deal with some of the problems of gender. More especially, those courses dealing with gender stereotypes; physical education programs that let girls and boys have new untypical experiences, can help them gain more self-respect and a better understanding of gender-processes. But it cannot be done just by separating girls and boys, it requires awareness and cautious acting!

Coeducational settings do need more sensitive reflections about what is going on in classrooms. Barrie Thorne analyzed the 'gender play' in schools and she found: 'Gender boundaries are episodic and ambigious, and the notion of "borderwork" should be coupled with a parallel term—such as "neutralization"—for processes through which girls and boys (and adults who enter into their social relations) neutralize or undermine a sense of gender as division and opposition' (Thorne 1993, p. 84).

Further research, gender studies in programs for teaching credentials as well as teacher training for more awareness of interaction processes dealing with gender questions, would help to make education gender-sensitive and valuable for both girls and boys.

See also: Education and Gender: Historical Perspectives; Education (Higher) and Gender; Education (Primary and Secondary Schools) and Gender; Gender and School Learning: Mathematics and Science; Gender Differences in Personality and Social Behavior; Mathematical Education; Sex-role Development and Education

Bibliography

American Association of University Women Educational Foundation 1992 *How Schools Shortchange Girls. A Study of Major Findings on Girls and Education*. Marlowe, New York

American Association of University Women Educational Foundation 1998 *Separated by Sex—a Critical Look at Single-sex Education for Girls*. Washington, DC

Connell R W 1995 *Masculinities*. University of California Press, Berkeley, CA

Diller A, Houston B, Morgan K P, Ayim M 1996 *The Gender Question in Education*. Westview Press, Boulder, CO

Eder D, Evans C C, Parker S 1995 *School Talk*. Rutgers University Press, New Brunswick, NJ

Faulstich-Wieland H 1991 *Koedukation—enttäuschte Hoffnungen?* [*Coeducation—disappointed hopes?*] Wissenschaftliche Buchgesellschft, Darmstadt, Germany

Faulstich-Wieland H, Horstkemper M 1995 *'Trennt uns bitte, bitte nicht!'. Koedukation aus Mädchen- und Jungensicht*. [*'Please don't separate us!' Coeducation as Girls and Boys see it*]. Leske and Budrich, Opladen, Germany

Howe F 1984 *Myths of Coeducation*. Indiana University Press, Bloomington, IN

OECD 1986 *Girls and Women in Education. A Cross-national Study of Sex Inequalities in Upbringing and in Schools and Colleges*. Paris

Riordan C 1990 *Girls and Boys in School. Together or Separate?* Teachers College, Columbia University, New York

Thorne B 1993 *Gender Play*. Rutgers University Press, New Brunswick, NJ

Tyack D, Hansot E 1992 *Learning Together. A History of Coeducation in American Public Schools*. Russell Sage Foundation, New York

Wilson M (ed.) 1991 *Girls and Young Women in Education. A European Perspective*. Pergamon Press, Oxford, UK

H. Faulstich-Wieland

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Cognition, Distributed

Like all other branches of cognitive science, distributed cognition seeks to understand the organization of cognitive systems. Like most of cognitive science, it takes cognitive processes to be those that are involved in memory, decision making, inference, reasoning, learning, and so on. Also following mainstream cognitive science, it characterizes cognitive processes in terms of the propagation and transformation of representations.

What distinguishes distributed cognition from other approaches is the commitment to two related theoretical principles. The first concerns the boundaries of the unit of analysis for cognition. While boundaries are often a matter of tradition in a field, there are some general rules one can follow. Bateson (1972) says one should bound the unit so that things are not left inexplicable. This usually means putting boundaries on units where the traffic is low. The second principle concerns the range of mechanisms that may be assumed to participate in cognitive processes. While mainstream cognitive science looks for cognitive events in the manipulation of symbols (Newell et al. 1989), or more recently, patterns of activation across arrays of processing units (Rumelhart et al. 1986, McClelland et al. 1986) inside individual actors, distributed cognition looks for a broader class of cognitive events and does not expect all such events to be encompassed by the skin or skull of an individual.

When one applies these principles to the observation of human activity 'in the wild,' at least three interesting kinds of distribution of cognitive process become apparent: cognitive processes may be distributed across the members of a social group, cognitive processes may be distributed in the sense that the operation of the cognitive system involves coordination between internal and external (material or environmental) structure, and processes may be distributed through time in such a way that the products of earlier events can transform the nature of later events. The effects of these kinds of distribution of process are extremely important to an understanding of human cognition.

The roots of distributed cognition are deep, but the field came into being under its current name in the mid-1980s. In 1978, Vygotsky's *Mind in Society* was published in English. Minsky published his *Society of Mind* in 1985. At the same time, Parallel Distributed Processing was making a comeback as a model of cognition (Rumelhart et al. 1986). The nearly perfect mirror symmetry of the titles of Vygotsky's and Minsky's books suggests that something special might be happening in systems of distributed processing, whether the processors are neurons, connectionist nodes, areas of a brain, whole persons, groups of persons, or groups of groups of persons.

1. *Mind in Society*

For many people, distributed cognition means cognitive processes that are distributed across the members of a social group (Salomon 1993). The fundamental question here is how the cognitive processes we normally associate with an individual mind can be implemented in a group of individuals? A wide range of disciplines in the social sciences has explored this question.

Treating memory as a socially distributed cognitive function has a long history in sociology and anthropology. Durkheim, and his students, especially Halbwachs (1925), maintained that memory could not even be coherently discussed as a property of an isolated individual. Roberts (1964) proposed that social organization could be read as a sort of architecture of cognition at the community level. He characterized the cognitive properties of a society (its memory capacity and ability to manage and retrieve information) by looking at what information there is, where it is located, and how it can move in a society. Schwartz (1978) proposed a distributional model of culture that emphasized the distribution of beliefs across the members of a society. Romney et al. (1986) created quantitative models of the patterns of cultural consensus. The identification of patterns raised the question of why such patterns form. Sperber (1985) introduced the idea of an epidemiology of representations. He suggested an analogy in which anthropology is to psychology as epidemiology is to pathology. In the same way that epidemiology addresses the distribution of pathogens in a population, anthropology should treat questions about the distribution of representations in a community. A similar set of developments followed from Dawkins' (1976) discussion of 'memes' as the cultural analog of genes. These ideas have now coalesced in the field of memetics (Blackmore 1999). March and Simon (1958) argued that organizations can be understood as adaptive cognitive systems. Juries are an important class of distributed problem solving organization and they have been intensely studied by social psychologists (Hastie 1993). Of course, in social psychology there is a vast literature on small-group decision making, some of which discusses the properties of aggregates.

Scientific communities have received special attention because the work of science is fundamentally cognitive and distributed. The phenomena that have been explored include how the organization of communication media in a scientific community affect the kinds of things the community can learn (Thagard 1993), how conditions external to the individual scientists can affect their individual choices in ways that lead to different high-level structures to emerge (Kitcher 1990), how the distribution of cognitive activity within social networks and between people and inscriptions accounts for much of the work of science (Latour 1987), and how scientific facts are

created by communities in a process that simply could not be fit into the mind of an individual (Fleck 1935).

Economists have been interested in the tension between what is individually rational and what is rational at the aggregate level. This theme has been explored in game theory under the rubric of the Prisoner's Dilemma, the paradox of the commons, and other cases where individual rationality and group rationality diverge (Von Neumann and Morgenstern 1964).

Anthropologists and sociologists studying knowledge and memory, social psychologists studying small-group problem solving and jury decision making, organizational scientists studying organizational learning, philosophers of science studying discovery processes, and economists and political scientists exploring the relations of individual and group rationality, all have taken stances that lead them to a consideration of the cognitive properties of societies of individuals. There is ample evidence that the cognitive properties of a group can differ from the cognitive properties of the members of the group.

2. *The Society of Mind*

The work described above looks for mind-like properties in social groups. This is the *Mind in Society* reading. The metaphor can be run the other way as is done in Minsky's *Society of Mind* (1985). Rather than using the language of mind to describe what is happening in a social group, the language of social groups can be used to describe what is happening in a mind.

Minsky argued that to explain intelligence we need to consider a large system of experts or agencies that can be assembled together in various configurations to get things done. Minsky also allowed that a high-level agency itself could be composed of low-level agencies. With Papert (Minsky and Papert 1988), he argued that the low-level agencies (the ones that take on 'toy-sized problems') could be implemented as distributed computations in connectionist nets. Minsky said, '… each brain contains hundreds of different types of machines, interconnected in specific ways which predestine that brain to become a large, diverse society of partially specialized agencies' (1988). What this means of course is that the cognition of an individual is distributed cognition too.

A problem that remained unsolved by Minsky's work was 'how such systems could develop managers for deciding, in different circumstances, which of those diverse procedures to use' (1988). That is, how can the relations among the agencies get organized to perform new functional skills? To solve this problem, Minsky and Papert invoked biological maturation. An alternative way to approach this problem is to note that each 'society of mind' resides and develops in a community of similar societies of mind. This means, of course, that both what's in the mind, and what the mind is in are societies. Getting internal agencies into coordination with external structure can provide the organization of the relations between the internal agencies that is required to perform the new functional skill.

Vygotsky developed this idea of the social origins of individual psychological functions in S*ociety of Mind* (Vygotsky 1978, Wertsch 1985). Vygotsky argued that every high-level cognitive function appears twice: first as an interpsychological process and only later as an intrapsychological process. The new functional system inside the child is brought into existence in the interaction of the child with others (typically adults) and with artifacts. As a consequence of the experience of interactions with others, the child eventually may become able to create the functional system in the absence of the others. This could be seen in Minsky's terms as a mechanism for the propagation of a functional skill from one society of mind to another. From the perspective of distributed cognition, this sort of individual learning is seen as the propagation of a particular sort of pattern through a community. Cultural practices assemble agencies into working assemblages and put the assemblages to work. Some of these assemblages may be entirely contained in an individual, and some may span several individuals and material artifacts. The patterns of activity that are repeatedly created in cultural practices may lead to the consolidation of functional assemblages, the atrophy of agencies that are rarely used, and the hypertrophy of agencies that are frequently employed. The result can be individual learning or organizational learning, or both.

2.1 *Interaction as a Source of Novel Structure*

An important property of aggregate systems is that they may give rise to forms of organization that cannot develop in the component parts. Freyd (1983) argued that some of the features of language that are identified as linguistic universals could arise out of the necessity of sharing the linguistic code. For instance, the reason that linguistic categories tend to approximate discrete structures may have little to do with the organization of the brain, and everything to do with the problem of pushing a complex representation through a very narrow channel. As Minsky and Papert point out, symbols can be expected to arise where there are bottlenecks in communication. That means we should look for the origins of symbols at places where the information 'traffic' is relatively low—or at the boundaries of our various units of analysis.

The phenomena related to the social distribution of cognition are most often investigated using ethnographic methods. In some cases, however, simulation models may be used to test hypotheses about the behavior of such distributed systems. For example, Hutchins and Hazlehurst (1995) explored Freyd's

ideas in a series of simulation models in which individuals (modeled by connectionist networks) interact with one another. They developed a robust procedure in which a shared lexicon emerges from the interactions of individuals. Hazlehurst and Hutchins (1998) demonstrated the emergence of reduced conventional sequences of lexical items—which they take to be the beginnings of syntax. These conventional sequences arise only in the condition of negotiated learning where the representing structures must simultaneously come to accurately represent the world and be shared among individuals, that is, be able to pass the communication bottleneck between individuals. Representations that are learned inside an individual, without the requirement of sharing them with others, come to represent the world, but do not show the reduced conventional code aspects that are the hallmarks of language and syntax.

By simultaneously considering the society of mind and mind in society, the distributed cognition approach provides a new place to look for the origins of complexity. Phenomena that are not predictable from the organization of any individual taken in isolation may arise in the interactions among individuals. Once having developed in this larger system, they may become elements of cultural practices and thereby become available for appropriation by individuals. This sort of scheme may be a partial solution to the paradox of how simple systems can lead to more complex ones.

3. *The Material Environment*

A second major thread in the fabric of distributed cognition concerns the role of the material environment in cognitive activity. Again, the question of where to bound the unit of analysis arises. The potential of the material environment to support memory is very widely recognized. But the environment can be more than a memory. Cognitive activity is sometimes situated in the material world in such a way that the environment is a computational medium.

Cognitive artifacts are the *Things that Make Us Smart* in the title of Norman's (1993) book. The notion that cognitive artifacts amplify the cognition of the artifact user is fairly commonplace. If one focuses on the products of cognitive activity, cognitive artifacts do seem to amplify human abilities. A calculator seems to amplify one's ability to do arithmetic, writing down something one wants to remember seems to amplify one's memory. Cole and Griffin (1980) point out that this is not quite correct. When I remember something by writing it down and reading it later, my memory has not been amplified. Rather, I am using a different set of functional skills to do the memory task. Cognitive artifacts are involved in a process of organizing functional skills into cognitive functional systems.

Consider an example from the world of ship navigation (Hutchins 1995). Navigators frequently face the problem of computing the ship's speed from distance traveled over a given period of time. If a ship travels 1,500 yards in 3 minutes, what is the speed of the ship in knots? There are many ways to solve this problem. Most readers of this article would probably attempt to use a paper and pencil plus their knowledge of algebra to solve it. That procedure is effective, but not nearly as efficient as the 'so-called' 3-minute rule. An experienced navigator need only see the problem stated to see that the answer is 15 knots. The speed in knots equals the number of hundreds of yards covered in 3 minutes. The use of this rule is a case of situated seeing. The rule itself is an internal cognitive artifact. But suppose the ship covered 4,000 yards in 7 minutes? For that problem a material artifact called the three-scale nomogram is more appropriate. A nomogram has three logarithmic scales: one each for distance, time, and speed. If the values of any two variables in a distance/rate/time problem are known, the other can be determined by laying a straight edge on the nomogram so that it touches the known values. The straight edge will touch the third scale at the value of the answer. It is clear that cognitive work is being done, but it is also clear that the processes inside the person are not, by themselves, sufficient to accomplish the computation. A larger unit of analysis must be considered. The skills of scale reading and interpolation are coordinated with the manipulation of objects to establish a particular state of coordination between the straight edge and the nomogram. This is a very different set of agencies than was involved in doing the problem via algebra and paper and pencil. In fact, the skills that are needed to use the nomogram are the things that humans are good at: pattern matching, manipulation of objects in the world, and mental simulation of simple dynamics (Rumelhart et al. 1986).

A computation was performed via the manipulation of a straight edge and nomogram. The nomogram was designed in such a way that the errors that were possible in algebra are impossible when using the nomogram. It is essential to distinguish the cognitive properties required to manipulate the artifact from the computation that is achieved via the manipulation of the artifact. This is a key point, and the failure to see it clearly has been the source of many difficulties in cognitive science.

4. *Distributing Cognition in Time*

Simon (1998) offered a parable as a way of emphasizing the importance of the environment for cognition. He argued that, as we watch the complicated movements of an ant on a beach, we might be tempted to attribute to the ant some complicated program for constructing the path taken. In fact, Simon says, that trajectory tells us more about the beach than about the

ant. Similarly, in watching people thinking in everyday settings, we may be learning as much about their environment for thinking as about what is inside them. The environments of human thinking are not 'natural' environments. They are artificial through and through. They develop over time. The crystallization of partial solutions to frequently encountered problems in artifacts such as the 3-minute rule and the nomogram is a ubiquitous strategy for the stabilization of knowledge and practice. Humans create their cognitive powers in part by creating the environments in which they exercise those powers.

5. Conclusion

It does not seem possible to account for the cognitive accomplishments of our species by reference to what is inside our heads alone. One must also consider the cognitive roles of the social and material world. But, how shall we understand the relationships of the social and the material to cognitive processes that take place inside individual human actors? This is the problem that distributed cognition attempts to solve.

According to Gardner (1985), a more or less explicit decision was made in cognitive science to leave culture, context, history, and emotion out of the early work. These were recognized as important phenomena, but their inclusion made the problem of understanding cognition too complex. The 'Classical' vision of cognition that emerged was built from the inside out, starting with the idea that the mind is a central logic engine. From that starting point, it followed that memory could be seen as retrieval from a stored symbolic database, that problem solving is a form of logical inference, that the environment is a problem domain, and that the body is an input device (Clark 1996). Attempts to reintegrate culture, context, and history into this model of cognition have proved very frustrating. The distributed cognition perspective aspires to rebuild cognitive science from the outside in, beginning with the social and material organization of cognitive activity.

See also: Cognitive Science: Overview; Situated Cognition: Origins

Bibliography

Bateson G 1972 *Steps to an Ecology of Mind*. Balantine Books, Location

Blackmore S 1999 *The Meme Machine*. Oxford University Press, Oxford

Clark A 1996 *Being There: Putting Brain, Body and World Together Again*. MIT Press, Cambridge, MA

Cole M, Griffin M 1980 Cultural amplifiers reconsidered. In: Olson D (ed.) *The Social Foundations of Language and Thought*. Norton, London

Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford

Fleck L 1979 *The Genesis and Development of a Scientific Fact*. University of Chicago Press, Chicago

Freyd J 1983 Shareability: the social psychology of epistemology. *Cognitive Science* **7**: 191–220

Gardner H 1985 *The Mind's New Science*. Basic Books

Halbwachs M 1925 *Les Cadres Sociaux de la Mémoire*. Libraire Felix Alcan, Paris

Hastie R 1993 *Inside the Juror: The Psychology of Juror Decision Making*. Cambridge University Press, Cambridge, UK

Hazlehurst B, Hutchins E 1998 The emergence of propositions from the coordination of talk and action in a shared word. *Language and Cognitive Process* (special issue on Connections Approaches to Language Development edited by Kim Plunkett). **2** & **3** April/May

Hutchins E 1995 *Cognition in the Wild*. MIT Press, Cambridge, MA

Hutchins E, Hazlehurst B 1995 How to invent a lexicon: the development of shared symbols in interaction. In: Gilbert N, Conte R (eds.) *Artificial Societies: The Computer Simulation of Social Life*. UCL Press, London

Kitcher P 1990 The distribution of cognitive labor. *The Journal of Philosophy* **87**(1): 5–22

Latour B 1987 *Science in Action*. Harvard University Press, Cambridge, MA

March J, Simon H 1958 *Organizations*. Wiley, London

McClelland J, Rumerlhart D, PDP research group 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol 2: Psychological and Biological Models*. MIT Press, Cambridge, MA

Minsky M 1985 *Society of Mind*. Simon and Schuster, Hemel Hempstead, UK

Minsky M, Papert S 1988 *Perceptrons*. MIT Press, Cambridge, MA

Newell A, Rosenbloom P, Laird J 1989 Symbolic architectures for cognition. In: Posner M (ed.) *Foundations of Cognitive Science*. MIT Press, Cambridge, MA

Norman D 1993 *Things that Make Us Smart*. Addison Wesley, Location

Roberts J 1964 The self-management of cultures. In: Goodenough V (ed.) *Explorations in Cultural Anthropology: Essays in honor of George Peter Murdock*. McGraw-Hill

Romney A K, Weller S, Batchelder W 1986 Culture consensus: A theory of culture and informant accuracy. *American Anthropologist* **88**(2): 313–38

Rumelhart D, McClelland J, PDP research group 1986 *Parallel Distributed Processing: explorations in the Microstructure of Cognition. Vol 1: Foundations*. MIT Press, Cambridge, MA

Rumelhart D, Smolensky P, McClelland J, Hinton C 1986 Schemata and sequential thought processes in PDP models. In: McClelland J, Rumelhart D, PDP research group (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol 2: Psychological and Biological Models*. MIT Press, Cambridge, MA

Salomon G 1993 *Distributed Cognitions*. MIT Press, Cambridge, MA

Simon H 1998 *The Sciences of the Artificial* 3rd edn. MIT Press, Cambridge, MA

Schwartz T 1978 The size and shape of a culture. In: Barth F (ed.) *Scale and Social Organization*. Universitetesforlaget

Sperber D 1985 Anthropology and psychology: towards an epidemiology of representations. *Man* **20**: 73–89

Thagard P 1993 Societies of minds: Science as distributed computing. *Studies in History and Philosophy of Science* **24**: 49–67
Von Neumann J, Morgenstern O 1964 *Theory of Games and Economic Behavior*. Wiley, Chichester
Vygotsky L 1978 *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA
Wertsch J 1985 *Vygotsky and the Social Formation of Mind*. Harvard University Press, Cambridge, MA

E. Hutchins

Cognitive Aging

Cognitive aging is concerned with age-related changes in adulthood in the basic processes of learning and memory, as well as the complex higher order processes of language and intellectual competence or executive functioning. Although most of the literature has been concerned with explaining the mechanism of cognitive decline, there is also a substantial interest in issues such as compensation and the role of external support, including collaborative problem solving.

1. *Definition of Cognitive Aging*

There have been two distinct traditions in the study of cognitive aging. The first grew out of experimental child psychology while the second derives from psychometric roots.

1.1 *Experimental Study of Memory Functions and Language*

The concern in this literature is to explicate possible causal variables that would explain why many adults suffer memory loss and decline in the complex manipulation of language variables such as text processing. The typical approach here is to design experiments testing for the effects of single variables in carefully controlled laboratory settings requiring only limited numbers of subjects. Because there is often little interest in individual differences, or population parameters, study participants are typically drawn from convenience samples (McKay and Abrams 1996).

1.2 *Descriptive Study of Adult Intellectual Development*

Descriptive studies of adult intellectual development often stem from the longitudinal follow up of samples first assessed in childhood or adolescence. Other such studies may represent carefully stratified samples from defined populations, first assessed at a particular life stage, whether in early adulthood or in early old age. Although descriptive studies often begin as cross-sectional inquiries, they are most frequently conducted as longitudinal analyses since the interest is often in individual differences in intraindividual change, or in the elucidation of typologies of individuals who follow different growth trajectories. These are frequently large-sample studies, and the use of correlational or quasi-experimental approaches is typical (Baltes et al. 1999, Schaie 1996b).

2. *Methodological Issues*

2.1 *Age-comparative vs. Age Change Designs*

Much of the experimental cognitive aging literature is based on age-comparative studies, which typically contrast a group of young adults (typically college students) with convenience samples of community-dwelling older adults in their sixties and seventies. It should be recognized that such comparisons are fraught by the problem that it is often unreasonable to assume that the two age groups can be adequately matched for other status variables that might provide a rival explanation for any observed age difference on the dependent variable. This creates particular problems for identifying the mechanisms that may be implicated in age-related decline from young adulthood into old age. Age-comparative designs are also inadequate in explaining individual differences in age changes. The latter can only be investigated by means of longitudinal paradigms (Schaie 1965).

2.2 *The Role of Response Speed*

A number of theorists have argued that changes in the central nervous system are the primary common cause for the observed age-related declines in cognitive performance. In fact, there have been many published analyses that show a substantial reduction in age differences, if some measure or measures of reaction time or perceptual speed is partialled out of the relation between a given cognitive process and chronological age (Salthouse 1999). This issue is of particular concern because it is not clear whether the observed average increase in reaction time (generally assumed to be of the magnitude of approximately 1.6 from the early twenties to the late sixties), while reliably demonstrable in the laboratory, is of significance in many or most tasks of daily living.

3. *Basic Findings from the Experimental Literature on Cognitive Aging*

Most of this literature is cross-sectional in nature and usually consists of a comparison of convenience samples of young adults (often sophomore psychology

students) and of community-dwelling older adults (often participants in adult education programs). The major findings regarding age differences in cognitive performance include the following.

3.1 Memory

It is currently thought that older persons are at a disadvantage in retrieving information from memory when the information to be retrieved is complex and when there are few cues or other environmental support. Hence, age differences are far greater in recall than in recognition of information. It is also thought that the magnitude of age difference in memory is far greater when a task involves effortful processing than when automatic processing is involved. Hence, greater age differences have been found for explicit than implicit memory. Older persons are also thought to have greater difficulty in integrating the context of information they are trying to remember. It is also thought that working memory capacity (i.e., the information kept in immediately accessible store) becomes reduced with increasing age. On the other hand there is little evidence for age differences in long-term storage. Memory deficits occurring with age include nonverbal tasks such as memory for spatial location, memory for faces, and for actions and activities. Studies of prospective memory (i.e., remembering something to be done in the future) suggest that older people do well in remembering simple and event-based tasks, but are at a disadvantage when tasks become complex or are time-based. In sum, it appears that age differences are known to increase in magnitude as a function of the processing requirement of a given task (Salthouse 1999, Smith 1996).

3.2 Language

Age-related differences in language behavior are closely related to the processes of encoding and retrieving verbal materials discussed above. But in addition there appear to be greater age differences in textual tasks that involve recent connections than in those that involve recollection of older connections.

Language production also seems to be adversely affected in older persons under intense time pressure. The interesting tip-of-the-tongue phenomenon involving word-finding difficulty, however, seems to be more likely with infrequently used words. Significant age differences have also been found in language planning, that is, in planning what one intends to say and how to say it during language production. Hence, older persons are more likely to engage in hesitations, false starts, and repetitions. Age-linked deficits in story recall is thought to be more of a general deficit in connection formation than in specific communication ability. Older persons tend to benefit from textual material that provides priming of association because it contains learned semantically linked information (McKay and Abrams 1996).

4. Basic Findings from the Descriptive Literature on Age Changes in Intellectual Competence

Changes in intellectual competence over the adult life span have been studied primarily with either the Wechsler Intelligence Scale or with ability batteries derived from the Thurstonian Primary Mental Ability framework.

Distinctions are made between fluid abilities thought to be innate and crystallized abilities which involve the utilization of culturally acquired knowledge (Cattell 1963). More recently further distinctions have been introduced between the mechanics (or basic processes) of intellectual competence and the pragmatics that involve cultural mediation (Baltes et al. 1984).

What has been found in most longitudinal studies is that the adult life course of mental abilities is not uniform. The so-called fluid abilities (sometimes defined as cognitive mechanics or primitives) peak in early midlife and begin to decline in the early sixties. By contrast, crystallized abilities, or the pragmatics of intellectual competence that represent abilities acquired in a given cultural context (particularly verbal abilities), do not usually peak until the fifties are reached and begin to show significant decline only in the seventies and often show only minimal decline even in the eighties (Schaie 1996a). However, recent work in advanced old age suggests increasing convergence and steeper decline for both aspects of intellectual competence, probably caused by the increasing decline of sensory and central nervous system functions (Baltes and Lindenberger 1997, Baltes and Mayer 1999).

However, at any particular time, a cross-sectional snapshot may yield very different ability profiles because of the fact that subsequent population cohorts reach different asymptotes in midlife. For example, there has been a positive linear cohort trend in the twentieth century for inductive reasoning, the basic component of most problem-solving tasks, while there has been a negative trend in numeric skills. The magnitude of cohort differences in abilities since the 1950s has been comparable to the average age changes observed from young adulthood into the seventies. Hence, many older persons may appear to have declined markedly in comparison to young peers, even though the age difference may be primarily due to what might be called the obsolescence of earlier cohorts (Schaie 1996b).

Studies of individual differences suggest that while most persons have declined on some aspect of intellectual functioning from their own peak as the sixties are reached, that specific patterns of decline

may well depend on complex patterns of individual life experience. Most healthy community-dwelling persons are able to maintain a high level of function until advanced old age (but see Baltes and Mayer 1999 for the consequences of sensory dysfunctions). Because most tasks of daily living represent complex combinations of basic cognitive processes, many individuals can maintain their abilities above the minimally necessary threshold level for independent functioning by often rather complex compensatory processes (cf. Baltes et al. 1984, Baltes et al. 1999).

5. *Can Cognitive Aging be Slowed or Reversed?*

Once the course of adult intellectual development had been described and a number of antecedents of individual differences had been identified, it then became useful for researchers to think about ways in which normal intellectual aging might be slowed or reversed.

In a number of laboratories (primarily in the USA and in Germany) cognitive training programs have been developed that have been applied in the laboratory, and more recently in cooperative multisite intervention trials. In contrast to training young children, where it can be assumed that new skills are conveyed, older adults are likely to have access to the skills being trained, but through disuse have lost their proficiency. Information from longitudinal studies is therefore particularly useful in distinguishing individuals who have declined from those who have remained stable. In the former, training is directed towards remediation of loss, while in the latter the enhancement of previous levels of functioning are sought with the intention of compensating for possibly cohort-based disadvantage of older persons.

Results from such cognitive interventions allow the conclusion that cognitive decline in old age, for many older persons, is likely to be a function of disuse rather than of the deterioration of the physiological or neural substrates of cognitive behavior. For example, a brief five-hour training program for persons over 65 resulted in average training gains of about one half SD on the abilities of spatial orientation and inductive reasoning. Of those for whom significant decrement could be documented over a 14-year period, roughly 40 percent were returned to the level at which they had functioned when first studied. The analyses of structural relationships among the ability measures prior to and after training further allow the conclusion that training does not result in qualitative changes in ability structures, and is thus highly specific to the targeted abilities. A seven-year follow up further demonstrated that those subjects who showed significant decline at initial training do remain at substantial advantage over untrained comparison groups (Willis and Schaie 1994). It should be noted, however, that while cognitive training may improve performance in the elderly and may function to reduce effects of age decrement, such training will also be effective in enhancing the performance of young adults such that age differences tend to remain robust (cf. Baltes and Kliegl 1992).

6. *Other Related Topics in Cognitive Aging*

Much of the work on cognitive aging in the past has been concerned with age-related development in the mechanics and basic processes of cognition. It should be recognized that current attention in the study of cognitive aging is turning to the discovery of how these basic processes operate within more complex domains. Of particular interest here are the study of wisdom (e.g., Baltes and Staudinger 1993, Sternberg 1990), the application of the basic processes to social cognition (e.g., Staudinger 1999), the development of expert systems (e.g. Charness and Bosman 1990), and in everyday problem solving (Willis 1996). The extensive literature on these topics is beyond the scope of this article.

See also: Aging and Health in Old Age; Aging, Theories of; Aging Mind: Facets and Levels of Analysis; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Differential Aging; Ecology of Aging; Lifespan Theories of Cognitive Development; Memory and Aging, Cognitive Psychology of; Memory and Aging, Neural Basis of; Old Age and Centenarians; Social Cognition and Aging; Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

Bibliography

Baltes P B, Dittmann-Kohli F, Dixon R A 1984 New perspectives on the development of intelligence in adulthood: Toward a dual process conception and a model of selective optimization with compensation. In: Baltes P B, Brim O G Jr (eds.) *Lifespan Development and Behavior*. Academic Press, New York, Vol. 6, pp. 33–76

Baltes P B, Kliegl R 1992 Further testing of limits of cognitive plasticity: Negative age differences in a mnemonic skill are robust. *Developmental Psychology* **28**: 121–5

Baltes P B, Lindenberger U 1997 Emergence of a powerful connection between sensory and cognitive functions across the adult life span: A new window at the study of cognitive aging. *Psychology and Aging* **12**: 12–21

Baltes P B, Mayer K U (eds.) 1999 *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, Cambridge, UK

Baltes P B, Staudinger U M 1993 The search for a psychology of wisdom. *Current Directions in Psychological Science* **2**: 75–80

Baltes P B, Staudinger U M, Lindenberger U 1999 Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology* **50**: 471–507

Cattell R B 1963 Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology* **54**: 1–22

Charness N, Bosman E A 1990 Expertise and aging: Life in the lab. In: Hess T H (ed.) *Aging and Cognition: Knowledge*

Organization and Utilization. Elsevier, Amsterdam, pp. 343–85
McKay D G, Abrams L 1996 Language, memory, and aging: Distributed deficits and the structure of new-versus-old connections. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 251–65
Salthouse T 1999 Theories of cognition. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 196–208
Schaie K W 1965 A general model for the study of developmental problems. *Psychological Bulletin* **64**: 92–107
Schaie K W 1996a *Intellectual Development in Adulthood, The Seattle Longitudinal Study*. Cambridge University Press, New York
Schaie K W 1996b Intellectual functioning and aging. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 266–86
Schaie K W, Willis S L 1999 Theories of everyday competence. In: Bengtson V L, Schaie K W (eds.) *Handbook of Theories of Aging*. Springer, New York, pp. 174–95
Smith A D 1996 Memory. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 236–50
Staudinger U M 1999 Social cognition and psychological approach to an art of life. In: Blanchard-Fields F, Hess T B (eds.) *Social Cognition, Adult Development and aging*. Academic Press, San Diego, CA, pp. 343–75
Sternberg R J 1990 Wisdom and its relation to intelligence and creativity. In: Sternberg R J (ed.) *Wisdom: Its Nature, Origins, and Development*. Academic Press, New York, pp. 142–9
Willis S L 1996 Everyday problem solving. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 287–307
Willis S L, Schaie K W 1994 Cognitive training in the normal elderly. In: Forette F, Christen Y, Boller F (eds.) *Plasticité cérébrale et stimulation cognitive*. Fondation Nationale de Gérontologie, Paris, pp. 91–113

K. W. Schaie

Cognitive and Interpersonal Therapy: Psychiatric Aspects

Cognitive and interpersonal therapies are each structured psychological treatments. They differ considerably in their rationale and therapeutic procedures, but both have been subject to research since their inception. In recent years, each has been demonstrated to offer effective symptomatic help for a range of specific mental disorders. This has facilitated their adoption in publicly or insurance funded health services, where major changes have been taking place in the pattern of therapeutic provision and in the training of mental health professionals. In English-speaking countries, where psychiatrists trained in psychological treatments were, even 10 years ago, more likely to offer psychoanalytic psychotherapy than any other kind, provision of cognitive or interpersonal therapy is increasingly common. Cognitive therapy is also provided frequently by clinical psychologists. This article will provide a brief overview of their respective rationales, methods, and current applications. It concludes with some speculations on future developments.

1. Cognitive Psychotherapy

1.1 Overview

All forms of cognitive therapy (CT) work from the premise that common mental disorders are consequent to and/or maintained by faulty thinking (rather than the reverse). Cognitive therapists, therefore, set out to identify cognitions associated with a target problem (such as depressed mood or hypochondriacal behavior) then use this analysis as the basis for an explicit therapeutic plan. Treatments addressing these cognitions can adopt a number of techniques, depending upon the problem addressed, the formulation of an individual case, and the specific form of cognitive psychotherapy favored by individual practitioners.

1.2 Rationale

Although cognitive therapy has only gained currency since 1970, its basic aims are not new. Attempts to restore mental health by arguing sufferers out of their false beliefs underpinned much 'moral therapy' in eighteenth-century asylums. As a modern movement, CT emerged from behavior therapy (qv). This had its theoretical rationale in learning theory (qv), derived from experimental manipulation of contingent responses in laboratory animals. Behavioral treatments had sought behavioral change through functional analysis of target symptoms. Behavioral theory paid little attention to the 'black box' of the mind and any mediating role it played between environmental stimulus and behavioural response. Cognitive therapy developed in reaction to this denial of the importance of thought, but was helped by clinical evidence that thinking could actively obstruct the progress of behavioral treatment unless it was explicitly attended to. The behavioral and cognitive approaches to psychological treatment have retained many common features, including emphasis on explicit formulation, and an empirical and collaborative approach. Behavioral and cognitive techniques may be combined within a treatment, and the close relationship between the two is reflected in the designation of 'cognitive-behavior therapy' or 'CBT' for much work that remains essentially cognitive.

Cognitive therapists have used a variety of models to account for how cognitive processes contribute m psychopathology. While these can be impressive in their orderliness and ingenuity, and can be of great heuristic value in practice, they nearly always derive from clinical experience. At the same time, independent support for such models has been sought from experimental psychology. Few cognitive findings

in emotional disorders remain robust when stringently tested, although a good deal of evidence exists that selective preference for negative memories is common in people with depressed mood, and irrational expectations of future danger is more common in people prone to anxiety (Williams et al. 1990).

The development of cognitive therapy has gone through a succession of stages of increasing theoretical complexity. These have accompanied a tendency, shared with other maturing psychotherapies, to follow early successes with relatively straightforward cases by attempts to deal with those that are more resistant to simple measures. While the theoretical ramifications of this are beyond the scope of the present article, it is helpful to distinguish between two kinds of cognition that have been seen as pathogenic—'surface' and 'deep.' Surface cognitions are available to introspection, transient, and situation specific. Deep cognitions are less easy to access, enduring and more global in scope.

The most influential package of models informing therapeutic practice derive from Beck and co-workers (Beck 1989). As applied to common disorders in which the main features are anxiety symptoms or depression, cognitive changes lead to evident differences in appraisal of a 'cognitive triad' of three interlinked areas. These are someone's automatic attitudes concerning their world, themselves, and what is likely to happen. Beck terms surface cognitions of these kinds 'automatic thoughts.' While their content is colored by the particulars of an individual's experience and personal values, some themes have been demonstrated as common to people having a particular kind of emotional disorder. For instance, when anxiety is marked, automatic thoughts about future danger are common; with depression, personal helplessness, and perceived inability to change their personal circumstances; with hostility, thoughts that the world is a bad or very unreliable place. Beck identifies two other kinds of cognitive pathology that accompany these developments. One is a set of ways in which cognitive processes frequently are distorted. These ensure that appraisals based on automatic thinking are likely to be confirmed. Examples include making of arbitrary or personal inferences, selective abstraction and overgeneralization from experience, and all or nothing ('dichotomous') thinking. At the level of deep cognitions, Beck recognizes the presence of 'schemas' as constellations of assumptions and beliefs which, being more latent but enduring cognitions that can be reactivated by events, are associated with predisposition to emotional disorders. The main functional relationships between these elements are summarized in Fig. 1.

While a good deal of reasoning in cognitive therapy represents extensions or amendments to this framework, a significant conceptual departure came with revision of the concept of the 'schema.' Used to refer to a variety of kinds of deep cognition, the term's

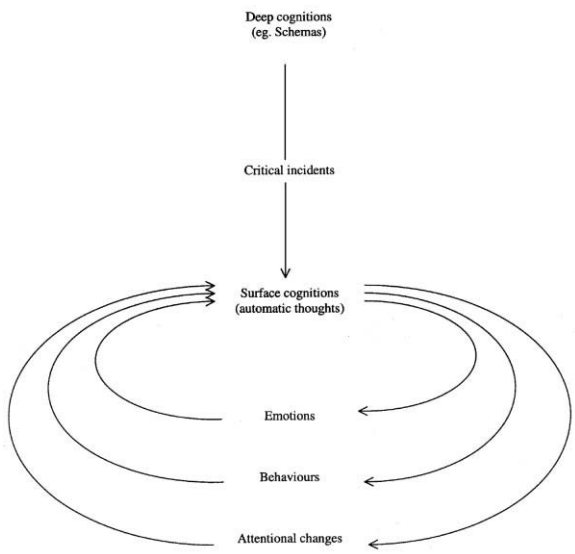

Figure 1
A map of problem cognitions

adoption by Young (1994) to refer to 'early maladaptive schemas' (EMS) has had important practical implications. An EMS is not only latent and enduring, but presumed to result from experiences early in life. It is rigid in restricting the scope of thinking and likely to be associated with very strong affect if challenged. Because of this emotional valence, EMSs can appear to be quasi-autonomous, acting to preserve themselves (schema maintenance). There can be a preference for executive functions that either fail to activate the schema (schema avoidance) or that disguise it behind displays of contrary traits (schema compensation). EMSs resistance to change means they cannot be inferred straightforwardly from behavior or simple enquiries.

1.3 Methods

In general, therapeutic strategies aim to counteract cognitive distortions by teaching patients skills by which they can recognize and revise problematic cognitions, as well as working to challenge specific current cognitions and to establish a more adaptive way of thinking. Essentially, this will involve techniques for the identification of surface and deep cognitions and their sequelae that are not otherwise amenable to simple introspection, and for the modification of beliefs and cognitive distortions so encountered. Although this account will concentrate on interventions directed at cognitions, both of these tasks are carried out in practice through the prescription of behavioral experiments and exercises. For instance, when working with people subject to panic

attacks who fear they may die from a heart attack during a panic episode, the assumption that physical symptoms such as dizziness, breathlessness, and awareness of accelerated heartbeat indicate collapse and death are imminent is addressed by a behavioral procedure. In order to break the link between these experiences and expectations, a patient agrees to hyperventilate under controlled conditions until these sensations are induced. The strength of their beliefs concerning the consequences of the hyperventilation exercise would be rated and recorded at its outset and its conclusion, as part of a therapeutic examination of the validity of the expectations in the light of a disconfirmatory experience. This example illustrates the general principle that, even when using a behavioural manipulation, a cognitive therapist would insist it subserves a cognitive goal, and that attention is paid to its cognitive impact throughout.

Measures to identify automatic thoughts include analysis of situations in which a target problem occurs. This is unlikely to be successful through generalized recall: specific instances need to be examined in detail. These might be recounted in the session; lived through between sessions after a patient is instructed on how to maintain a detailed and contemporaneous record of associated negative thoughts; or reconstructed as a patient has a shift of affect in a session. Other techniques such as role-plays or induction of imagery may be employed in sessions to stimulate automatic thoughts. Measures intended to facilitate reattribution expose not only the irrationality of automatic thoughts but their incompatibility with experience. The importance of active testing in modifying them, alongside regular and explicit reappraisal of the validity of the thoughts, has already been referred to. Patients and therapists are likely to collaborate in drawing up a set of alternative explanations whose fit with the facts of experience can then also be tested.

As experience of cognitive techniques has developed, there has been a progressive shift of interest away from factors leading simply to inaccurate appraisal of situations, to those by which faulty cognitions, and problems associated with them, are maintained (Salkovskis 1991). These may be automatic thoughts associated with plans of action that perpetuate a problem by protecting a faulty cognition from challenge (safety behaviors) or attentional shifts that have a similar impact through distraction. Interventions that address these maintaining factors directly can allow cognitive techniques to be effective where attention to errors in appraisal alone would not.

Deep cognitions in the form of schemas are not only more difficult to identify than surface cognitions, but harder to attempt to change therapeutically. Identification needs a more probing and hypothetical approach in the course of guided discovery by use of techniques such as the 'vertical arrow.' This is a reiterative questioning of patients' suppositions about the implications of a negative idea being the case, until a fundamental and general belief about themselves emerges. Attempts to change schemas would normally only be made when there had been some progress with emotional symptoms. As the previous description of the kinds of concealment and resistance associated with EMSs implies, schema-focused work is likely to be more complicated and prolonged than work with automatic thoughts. Because schemas may embody templates for how someone relates to others, and be associated with manifest relationship problems, therapeutic work with them is likely to resemble psychodynamic psychotherapy (qv) more closely than other forms of cognitive therapy (Safran and Segal 1990).

1.4 Applications

Unlike psychoanalytic psychotherapies, cognitive psychotherapy has developed through its application to conditions in which symptoms or problematic habits are associated with specific patterns of thinking. Its growth since 1970 has been accompanied by progressive developments in psychiatric nosology. The diagnostic and statistical manuals of the American Psychiatric Association chart diagnostic developments which have involved the progressive refinement of diagnoses based on anxiety, depression, psychosis, and distortions of personality, with the effective invention of categories to encompass the so-called somatoform, dissociative, adjustment, and eating disorders (APA 1968, 1980, 1994). These have facilitated understanding of common cognitive patterns associated with these and development of specific treatment techniques. Use of cognitive therapy in depression and anxiety disorders are best established, while diagnostic subclassifications that have clarified the characteristics of bipolar depressive disorder and panic disorder have been followed by specific cognitive techniques for their management. The category of personality disorders has not only been refined through this period, but, with the advent of multiaxial classification, been designated as an independent axis for summary clinical descripton (axis II). Given that the basic difference between disorders of personality from the symptom focused categories of axis I lies in their early onset and relative stability, models of their cognitive pathology have emphasized 'deep' cognitions over surface ones, and more recently the early maladaptive schema model of Young. Efforts to link specific personality types with consistent schema formations continue. A further development of particular significance to psychiatrists has been the application of cognitive therapy to schizophrenia. These have included treatments for specific symptoms (delusions and hallucinations) as well as measures to live with the impact of illness and enhance coping capacities (Chadwick et al. 1996).

The vast majority of cognitive therapy is provided as individual therapy. However, group treatments

have been pioneered for specific disorders including depressive, anxiety, and eating disorders. Beck has advocated its use with couples experiencing conflict.

More recently, attempts have been made to differentiate between psychological therapies with reference to the strength of independent evidence concerning their clinical effectiveness. An empirically supported treatment is one which is clearly defined and, for a given clinical problem, is consistently more efficacious than placebo treatments on the evidence of controlled clinical trials. On this basis, cognitive therapies have emerged as empirically supported in adults for anxiety disorders (including panic disorder, generalized anxiety disorder, and social phobia); unipolar depression; anorexia nervosa and bulemia nervosa (Roth and Fonagy 1996).

1.5 Comment

The remarkable growth of cognitive therapy has been facilitated by a number of external developments: the growth of clinical psychology as a profession whose members receive training in cognitive theory and techniques within their core curriculum; compatibility between cognitive models of psychopathology and classifications adopted for diagnosis, care management, and organization of the evidence base; and an intellectual Zeitgeist that favors explanations based on information processing or neurocognitive paradigms. The relative rapidity of the clinical effects of CT has invited comparisons of their efficacy with those of pharmacological treatments, while their longer term impact is not always clear. They enjoy a high level of acceptability that reflects their collaborative and transparent style. However, the emphasis on monitoring, rating, and homework exercises can be seen as excessively demanding by some patients, and the relative lack of attention to the historical origins of difficulties is not to all patients' taste. Practical methods are relatively easily learned and taught, although growing evidence concerning the added value of intensive training and supervision for clinical outcomes suggests this ease is deceptive. Important developments in the theoretical base and practice of cognitive therapies continue to be made. Some of these, in indicating the importance of formative experience, of the relationship with the therapist and patterns of relating, and of increasingly inaccessible ('unconscious') cognitive structures underlying behavior, are narrowing the gap between cognitive and psychodynamic models of psychotherapy in practice.

2. Interpersonal Therapy

2.1 Overview

Interpersonal therapy (IPT) is an increasingly common model of brief psychotherapy that was developed as a specific treatment for depression in the 1970s. Findings from research into the interpersonal precipitants of depression were used in designing an intervention that could bring symptomatic relief through improvements in interpersonal functioning (Klerman et al. 1984). Different therapeutic strategies are used, depending upon which of four basic kinds of interpersonal problem is paramount in an individual case. IPT is now being adapted to other disorders and settings (Klerman and Weissman 1993).

2.2 Rationale

IPTs basic rationale is that depressed mood is usually secondary to deterioration in interpersonal relationships, and can be reversed by deliberate attention to the quality of current relationships. Historically, it was developed in deliberate contrast to psychodynamic therapies, in which the importance of interpersonal relationships within the therapy itself as well as in a patient's life are paramount. A key difference has been that psychodynamic psychotherapies not only pay considerable attention to the formative influence of past relationships, but they focus on (maladaptive) patterns of relating that are seen as characteristic of a person. Interpersonal therapy pays attention to current relationships for their own sake. It does not conceptualise underlying patterns, although it would expect someone's style of relating to improve through reinforcement of positive changes achieved in the relationships targeted in therapy. Although the 'interpersonal psychiatry' of Sullivan (1953) is frequently quoted as an antecedent of IPT, its focus on in-session interactions and on enduring patterns of relating is inconsistent with this new therapy, despite their mutual emphasis on the importance of good interpersonal relationships for personal mental health.

IPT has used theoretical developments in a number of fields, from attachment theory to life events research, to highlight the association of onset of depression with loss through grief or major life changes, conflict, and isolation. Its model identifies whether a person's most pressing need is resolution of grief, role transitions, interpersonal disputes, or interpersonal deficits.

The model was designed to be researchable from the outset, its method being summarized in a manual (Klerman et al. 1984) to promote consistent application in practice. Studies have not only addressed its efficacy for specific conditions (see below) but the impact of process on outcome. An important literature has therefore also developed concerning the impact of training on the practice of psychotherapy (Rounsaville et al. 1988).

2.3 Method

As a time-limited therapy, IPT was pioneered over 12 sessions per treatment, now often 16. A typical

treatment is subdivided into initial, treatment, and termination phases. During the initial phase, the patient is educated to see their difficulties as the consequence of having a depressive illness, and to allow themselves to occupy a sick role. (This means they should not feel responsible for this state of affairs, and allow others to take on some of their normal duties so they can concentrate on recovery.) A detailed inventory is drawn up with the therapist summarizing all current relationships, however insignificant. This not only provides a map of potential areas of difficulty, but of sources of potential support and opportunities for the development of relationships in future. Detailed questioning in the first phase allows a treatment focus to be identified which reflect four distinct forms of interpersonal need: grief (where the loss of a significant other through death has not been overcome); role disputes (where conflict in a key relationship, perhaps in the form of an impasse rather than overt fighting, cannot be resolved); role transitions (where adaptation to a different situation, commonly following loss events, is required). In some residual cases, where a patient's interpersonal situation is particularly impoverished through an inability to establish relationships, a fourth category of interpersonal deficits applies.

Different therapeutic techniques are likely to be required in each of these instances, the model being sufficiently flexible to accommodate these. Examples would be assisted mourning with grief, and attention to communication in interpersonal disputes. In all cases, there is careful attention to affect and considerable emphasis on its successful expression throughout treatment.

Therapy concludes with explicit attention to termination, both in anticipating and working through loss of the therapy and in planning for continuing progress along the lines tried out during the therapy.

2.4 Applications

IPTs use in the treatment of major depression was highlighted by a large randomized controlled trial sponsored by the US National Institute of Mental Health (Elkin et al. 1989). This remains one of the largest comparative trials of psychological and pharmacological treatments ever conducted, providing information on the relative benefits from two psychotherapies (CT and IPT), imipramine and a drug placebo. The results were encouraging for IPT, showing it to be as effective as CT in relieving symptoms overall, while having the lowest attrition, and significantly better results than CT among the most severely depressed patients.

Since its original application in studies of clinical depression, IPT's use has been broadened to incorporate depressed populations with special needs; patients with distinct mental disorders and adaptations of the therapeutic process to fit different working contexts (cf. Klerman and Weissman 1990). Depression in adolescents and the elderly, as well as the chronic low grade depression known as dysthymic disorder, have all been shown to benefit from treatment (Markowitz 1998). Modifications of therapeutic technique can be involved, for instance greater involvement of significant others in the treatment process with adolescents, and use of less frequent maintenance sessions after the phase of regular sessions in dysthymia. Adjustment of content is likely to be involved with use of IPT with other disorders. A great deal of exploratory work in the field of substance misuse has so far failed to show significant benefits. An area of greater promise has been in the treatment of bulimia nervosa, where lasting clinical improvements comparable to those from CT have been achieved (Agras et al. 2000).

The principal adaptations to the model to suit different contexts have been its shortening to six brief sessions for use with subclinical populations in primary care settings (cf. Klerman and Weissman 1990). Attempts to develop a group model of IPT for treatment of social phobia are promising but at an early stage.

3. Conclusion

Although cognitive and interpersonal therapeutic packages are relatively new, each has elaborated principles of good psychiatric management—respectively, the importance of a patient's attitude or social relationships to their health and recovery—that are widely recognized. Their rapid growth reflects their proven efficacy for specific disorders, the relative ease with which they can be learned and disseminated, and the promise of brief psychological treatment for clinical conditions that account for a large proportion of psychiatric practice. Both are widely used alongside physical treatments such as psychotropic medication. Developments are likely to include continuing influence of basic research in psychology and the neurosciences on the refinement of therapeutic models and methods, and increasing rapprochement with other therapeutic models as integrative models of treatment are developed for more complex and treatment-resistant conditions.

Bibliography

Agras W S, Walsh B T, Fairburn C G, Wilson G T, Kraemer H C 2000 A multicenter comparison of cognitive-behavioural therapy and interpersonal theapy for bulimia nervosa. *Archives of General Psychiatry* **57**: 459–66

American Psychiatric Association 1968 *Diagnostic and Statistical Manual of Mental Disorders*, 2nd edn. APA, New York

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. APA, New York

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. APA, New York
Beck A T 1989 *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York
Chadwick P D J, Birchwood M J, Trower P 1996 *Cognitive Therapy for Hallucinations and Delusions*. Wiley, Chichester, UK
Elkin I, Shea M T, Watkins J T, Imber S D, Sotsky S M, Collins J F, Glass D R, Pilkonis P A, Leber W R, Docherty J P, Fiester S J, Paloh M B 1989 National Institute of Mental Health Treatment of Depression Collaborative Research Programme: General effectiveness of treatment. *Archives of General Psychiatry* **46**: 971–83
Klerman G L, Weissman M M, Rounsaville B J, Chevron E S 1984 *Psychotherapy for Depression*. Basic Books, New York
Klerman G L, Weissman M M (eds.) 1990 *New Applications of Interpersonal Psychotherapy*. American Psychiatric Press, New York
Markowitz J C 1998 *Interpersonal Psychotherapy for Dysthymic Disorder*. American Psychiatric Press, New York
Roth A, Fongagy P 1996 *What Works for Whom?* Guilford Press, New York
Rounsaville B J, O'Malley S, Foley S, Weissman M M 1988 Role of manual-guided training in the conduct and efficacy of interpersonal psychotherapy for depression. *Journal of Consulting and Clinical Psychology* **56**: 681–8
Safran J D, Segal Z V 1990 *Interpersonal Process in Cognitive Therapy*. Basic Books, New York
Salkovskis P M 1991 The importance of behaviour in the maintenance of anxiety and panic: A cognitive account. *Behavioural Psychotherapy* **19**: 6–19
Sullivan H S 1953 *The Interpersonal Theory of Psychiatry*. Norton, New York
Williams J M G, Watts F N, MacLeod L, Matthews A 1990 *Cognitive Psychology and the Emotional Disorders*. Tavistock, London
Young J E 1994 *Cognitive Therapy for Personality Disorders: A Schema Focused Cognitive Therapy*. Professional Resource Exchange, Sarasota, FL

C. J. Mace

Cognitive Anthropology

Cognitive anthropology attempts to link anthropology with the cognitive sciences. Culture, as one of anthropology's central objects of research, has an effect in two respects: on the one hand, in a material sense in the form of cultural phenomena, on the other hand, mentally, in the form of cultural contents. Cultural contents are based on mental representations. While cultural phenomena are public and thus easy to document ethnographically, cultural contents are not directly accessible since mental representations cannot be observed. What actually happens inside people's heads is the object of research of the cognitive sciences. Cognitive sciences see the human experience of reality and human thinking as acts of processing information. In addition to the study of perception, mental representation, and memory, the objective is to make transparent those mental faculties that enable all humans to internalize the social and cultural characteristics of the society into which they are born. But these are also central topics of anthropological research.

1. *From Ethnoscience to the Cognitive Sciences*

The year of birth of the 'cognitive revolution' in which, apart from social anthropology, various human sciences participated, to some extent independently of each other, is considered to be 1956. That year, at a conference on information theory at the Massachusetts Institute of Technology, Newell and Simon presented a paper on computer programs, Miller presented his famous treatise *The Magical Number Seven*, and the 28-year-old Chomsky read excerpts from his thesis *Three Models of Language*. In the same year, Bruner's book *A Study of Thinking* was published and two anthropologists, Goodenough and Lounsbury, published the first two programmatic articles on cognitive anthropology, *Componential Analysis and the Study of Meaning* and *A Semantic Analysis of the Pawnee Kinship Usage*.

With cognitive anthropology or, more precisely, with the ethnoscience phase of cognitive anthropology, a new field of research came to the fore. Its goal was to describe other cultures in their own conceptualization, that is, in emic terms or from the inside. Different cultures categorize the world differently and apply a different type of logic in dealing with their environment. This difference should be recorded as it reveals different cognitive worlds. The underlying question was an old one: What does the 'order out of chaos' look like? More precisely, the basic questions were 'How do other cultures label the things in their environment, and how are these labels related to each other?' This according to the assumption that with the help of cultural phenomena the cultural contents, and thus the mental representations, could also be documented.

1.1 *Three Premises*

One can argue that three premises formed the background to this ambitious program:

1.1.1 Premise 1: Culture is common (shared) knowledge. A highly significant definition of culture was coined by Goodenough: 'A society's culture consists of whatever it is one has to know or believe in order to operate in a manner acceptable to its members, and to do so in any role that they accept for any one of themselves. … It is the forms of things that people have in mind, their models for perceiving, relating, and otherwise interpreting them. … Culture does not exist of things, people, behaviour, or emotions, but

in the forms or organizations of the things in the minds of people' (Goodenough 1957, pp. 167–8).

Seen in this way, culture is a mental phenomenon.

1.1.2 Premise 2: Knowledge has the form of a cultural grammar. Anthropologists must, on the basis of the statements made by their informants, inductively discover this abstract and shared knowledge as a systematic mental representation. In principle, they can, in order to do so, limit themselves to one person, as applies when learning a foreign language, where at first sight it suffices to have one speaker at one's disposal. The knowledge system of a culture is understood to be a conceptual model which embraces the organizational principles of the culture and the behavior of its members. The model is, so to speak, a cultural grammar.

1.1.3 Premise 3: Language is the best means of access to mental phenomena. With the reduction of chaos, certain phenomena and characteristics are selected from the environment as being significant, named, and given a classificatory meaning. The main (but not the only) proof of the existence of a category is its label.

The equation of culture and knowledge proved to be very fruitful. In the 1960s, ethnoscience experienced rapid success. Innumerable studies were published—which was to be expected—about terminologically densely structured individual fields, such as those on kinship, colors, ethnozoology, ethnobotany, or illness. One succumbed, as it were, to the great theoretical temptation to reduce complex and ostensibly heterogeneous things to a few rules (inclusion, exclusion, and intersection) and to present them as elegant models (taxonomy, paradigm, see below) which were looked upon as mental representations of individuals.

At the beginning of the 1970s, however, only a very few studies appeared, and Keesing (1972, p. 229) was justified in starting a paper with the sentence: 'Tell me, whatever happened to ethnoscience?' Here lies the irony: when, at the end of the 1950s, ethnoscience took over this model from linguistics and it became the focal point in the 1960s, it had already been swept aside in linguistics itself by Chomsky's new generative linguistics.

The result was an opening up and turning towards modern trends in neighboring disciplines. Computer science proved to have particularly strong influence; when the first computer programs appeared which played chess, the question arose, 'If computers can have programs, why can't people, too?' In endeavoring to reproduce human cognitive representations in the model, cognitive anthropology adopted the 'information-processing approach.' In so doing, the assumption (at that time) that cognitive processes follow the same pattern universally, whether in humans, animals, or the machine (in other words, that the software is the same everywhere, irrespective of the hardware in which it is processed) was, from a philosophical point of view, highly explosive.

1.2 Revised Premises

As a consequence, the three premises of ethnoscience were reconsidered and extended; the emphasis turned to where knowledge is sought and how it is represented.

1.2.1 Revised Premise 1: Turning towards the individual. Attention at the end of the twentieth century was no longer focused on the collective knowledge system as the whole of a culture (as représentation collective in the sense of Durkheim), supposed to be recorded as the ideal type, but on the scattered, variable knowledges acquired, stored (memorized), and applied by individuals in their everyday life. The focus shifted, because it was recognized that inferences cannot be directly drawn from cultural phenomena and linguistic material in order to elaborate individual cognitive processes or representations.

1.2.2 Revised Premise 2: Operationalization instead of categorization. As soon as knowledge is no longer defined as an isolated, static system (that is, simply as grammar) but as something which is evident (verbally or nonverbally) in everyday use by individuals, it becomes clear that many categories and semantic fields have no fixed boundaries, and cannot be defined in the classical sense (fuzzy sets). They are now also grouped according to what a person can do in daily life ('taskonomy' instead of taxonomy) or else according to prototypes (best example from a category). It is everyday cognition that a housewife needs when she shops, a milkman when he distributes dairy products to his customers according to a certain pattern, a Yakan in the Philippines when he wants to enter a house correctly.

1.2.3 Revised Premise 3: Turning away from language as the only instrument to code knowledge. Knowledge is also expressed by means of actions or emotions. Habitual actions in particular can be very 'eloquent' in the sense of tacit knowledge. In 1977, the computer specialist Schank and the sociopsychologist Abelson introduced the significant term 'script' to describe stereotyped sequences of actions in certain situations. Thus, although language remains one of the focal points, it is treated differently;

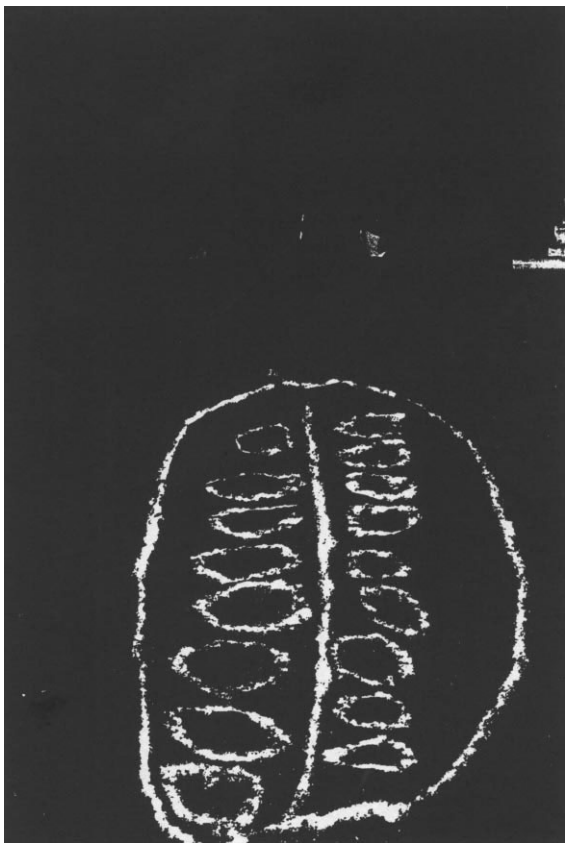

Figure 1
Ndanda, an old Yupno man (Papua New Guinea), draws a picture of the world.

Figure 2
Gwarane, an old Yupno woman, sorts objects (sorting task) following a hot/cold schema, after having first verbally given a taxonomic order which, however, plays no role in her everyday life.

no longer as a lexicon, but in everyday use as discourse from which inferences must be drawn as to the intended message (Hutchins 1980). Beyond this, the (controversial) idea is that the structure of knowledge (as stored in the mind as representation) is not necessarily language-like: 'Knowledge organized for efficiency in day-to-day practice is not only nonlinguistic, but also not language-like in that it does not take a sentential logical form' (Bloch 1991, pp. 189–190). In addition, the question arises whether language should not be ignored more frequently because certain kinds of knowledge cannot be externalized linguistically, or only with great difficulty (Wassmann 1993) (Fig. 1 and Fig. 2).

If the individual acting in his or her daily life now arouses interest, this is a consequence of a paradigmatic change: cognitive anthropology now considers itself more clearly as part of the cognitive sciences, which, inter alia, leads to the word 'cognition' being better understood. Cognition is no longer an expression of a culture as a whole and abstracted from linguistic material, but understood as the mental activity of individuals who actively apply knowledge in different contexts, in that they think, generalize, draw inferences, perceive, recognize and categorize; analyze, combine, assess possibilities, solve problems and make decisions; classify, differentiate and choose; remember and master new situations. These activities are performed individually or between individuals but, nevertheless, take place somehow within the broad framework of the 'Culture.'

2. *The Representation of Knowledge*

The most important models to represent knowledge in cognitive anthropology are briefly described below.

2.1 *Taxonomy*

The inner order of a semantic field (e.g., 'kinship,' 'colors') depends on a small number of ordering principles which structure the lexical components

Table 1
Taxonomy

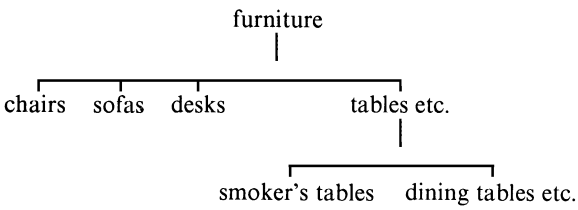

Table 2
Paradigm

		Sex	
		a1	a2
Generation	b1	grandfather	grandmother
	b2	father	mother
	b3	Ego	
	b4	son	daughter
	b5	grandson	granddaughter

(lexemes) of the field (e.g., 'mother,' 'red'). The ordering principles are inclusion, exclusion (contrast), and intersection. A taxonomy is the description of a semantic field; it lists the categories (lexemes) and shows how they are connected with each other, i.e., according to the principles of inclusion and exclusion in hierarchic order. Categories at the same level exclude each other (exclusion) while categories at the lower level are included in the categories at the higher level (inclusion). The grouping in Table 1 shows that 'dining tables' are different from 'smokers' tables' but that both are a kind of table. And also that chairs are distinguished from tables. It does not say what are the distinguishing characteristics.

2.2 *Paradigm*

If a semantic field is structured according to the principle of intersection, it is a paradigm. Hierarchy, inclusion, and exclusion are missing; instead the distinctive features (or criterial attributes) are stated which distinguish the different categories. In order to construct such a model, a componential analysis is made (Tyler 1969). A complete lexicon of a semantic field, e.g., 'blood relationship' is established and then those characteristics (components) are looked for according to which of the lexemes differ, e.g., 'male' and 'female' in the dimension 'sex,' as well as the distance from Ego in the dimension 'generation.' Every single lexeme is now defined by a bundle of components. In the graphic representation in Table 2 these components intersect at the defined lexeme.

Taxonomy and paradigm are the classic models of ethnoscience. They are strongly idealized, consistently emic and only to a small degree 'cognitive.' They are understood as the 'mind' of a whole culture and (as we know today) they are not really instruments of thinking.

Table 3
Prototypes

Prototype	Category
chair	furniture
car	vehicle
orange	fruit
gun etc.	weapon etc.

2.3 *Prototype*

Within a category of objects, that object is a prototype (for the whole category) which is thought to be the best example or the clearest case. Thus, for example, in the category 'furniture' the 'chair' is a better example than, for instance, 'radio.' Of course one could define 'furniture' as a category of objects having certain semantic characteristics or attributes in common (and whatever does not have these characteristics does not belong). However, this (checklist) definition does not allow any grading such as 'chair' being a better example, more of a prototype of 'furniture' than, for instance, 'radio' (although both belong to 'furniture'). It may happen that there are no criterial attributes common to all parts (i.e., no semantic field, no class according to ethnoscience), but only a large number of characteristics which may 'match' some but by no means all. As a consequence, they have only a 'family resemblance'-structure (a term Rosch adopted from Wittgenstein (Rosch and Merris 1975)). Following Lévi-Strauss, prototypes are particularly good to think (Table 3).

2.4 *Script, Schema and Cultural Model*

The information theory approach forces the anthropologist to be explicit. Exactly that has to be made explicit which normally remains implicit. This additional information is called 'script.' It is the tacit knowledge enabling us to also understand incomplete descriptions and suggestions: we automatically add what is missing by an inference process. Every situation requires specific knowledge and accordingly there are scripts for 'eating in a restaurant,' 'playing football,' 'attending a birthday party.' But not only our actions are based on scripts, but our language as well, as exemplified briefly in the following; a story told with all the details would be tedious.

I am in New York and somebody asks me the way to Coney Island; I tell him to take the 'N'-train to the terminus. This instruction only makes sense 'if this

improperly specified algorithm can be filled out with a great deal of knowledge about how to walk, pay for the subway, get in the train and so on' (Schank and Abelson 1977, p. 20).

If the typical characteristics of a situation are grasped, hence the stereotypical, the standard-like is stressed but raised to a higher level of abstraction; we can talk about schemata and cultural models (which partly replace the older term of the folk model). All the knowledge we acquire, remember, and communicate about this world is neither a simple reflection of this world nor does it consist of a series of categories (as ethnoscience assumed), but it is organized into different situation-relevant, prototypical, simplified sequences of events. We basically think in simplified worlds: '… cultural models are composed of prototypical event sequences set in simplified worlds' (Quinn and Holland 1987, p. 32). What is more, these models are probabilistic and partial; they are actual frames we can use to react to new situations as well. They are world-proposing yet cannot be directly observed, since they are not presented but merely represented by the behavior of the people. They are models *of* the mind and *in* the mind. The organizing principle behind these models seems to be metonymy, a part of the whole, the prototypical, conspicuous part is passed off as the whole, i.e., a whole is represented by one of its parts (cf. the frame theory of Minsky, and story grammar by Rumelhart).

2.5 Mental Images

Knowledge may also be represented by inner images. The image schema consists of schematized, simplified images. These images are able to make comprehensible and imaginable physical objects or logical relationships difficult to conceptualize. The organizing principle behind them appears to be the metaphor; through analogy, information from the physical world is introduced into the nonphysical world, for example, rage can be imagined as a hot liquid in a container, evaporation as rising molecules springing out of the water like popcorn, electricity as a crowd of people (in front of the gate of a stadium at a sports event).

3. How Deep?

The structure of the representations of knowledge as presented here under Sects. 2.3, 2.4, and 2.5 allows us to answer three questions central to cognitive anthropology (Quinn and Holland 1987, pp. 3–4).

(a) It is able to explain the apparent systematicity of cultural contents (knowledge) by pointing to a number of general-purpose models which can repeatedly be integrated into other concrete models which are special-purpose orientated.

(b) Mastering the enormous amount of knowledge every one of us has is possible because we only store what is prototypical, reduced, but are able to actualize it on demand (in concrete situations), i.e., supplement it with details (instantiation).

(c) It is possible to interpret new experiences because these models not only represent knowledge but also allow us to draw conclusions from them to new situations.

When answering these three questions, modern cognitive anthropology faces three more general problem areas.

3.1 Knowledge and Knowing

Giddens writes 'The vast bulk of the "stocks of knowledge" … incorporated in encounters is not directly accessible to the consciousness of the actors. Most of such knowledge if practical in character, it is inherent in the capability to 'go on' within the routines of social life' (Giddens 1984, p. 4). Imagine having to describe to somebody how to ride a bicycle; doing so, one has to question the traditional conception of knowledge. It seems to be advantageous to distinguish between knowledge (what is known, as an abstract pool of information, declarative and verbalized) and knowing (how to do something in practice, implicit and hidden, primarily accessed through performance)—the focus being on the latter. We may even reorient our analysis and reverse the process—by starting with knowing and seeing how knowledge is constituted from it (Borofsky 1994).

3.2 Language and Cognition

Does language shape our thinking? This question seems to be receiving attention once more (Gumperz and Levinson 1996). Thus spatial orientation and communicating it certainly belongs to the cognitive and linguistic basic equipment of all people and societies. However, there are different linguistic systems of orientation, and, for us, very basic spatial categories such as 'left,' 'right,' 'in front' or 'behind' cannot be taken for granted. Many languages do not know these terms and instead use a geocentric system based, for example, on the cardinal points and use this not only in navigation but also in everyday life (the glass is not placed to the left of the plate but to the east, for instance). However, these differences of a linguistic and cultural kind (probably) also influence the (cognitive) perception of spatial relationships as well as their storage in memory (Wassmann and Dasen 1998). Here cognitive anthropology contradicts the prevailing school of thought of cognitive linguistics in which the worldwide diversity of languages is only seen as a cultural phenomenon and hence one of surface.

3.3 Universals?

Cognitive processes such as categorizing, classification, memorizing, and perception are the founda-

tions of the contents of knowledge and structure these. In principle they are thought to be universal but can be applied in different ways.

'We found evidence of differences across cultural groups, differences in habitual strategies for classifying and for solving problems, differences in cognitive style, and differences in rates of progression through developmental stages … These differences, however, are in performance rather than in competence. They are differences in the way basic cognitive processes are applied to particular contexts, rather than in the presence or absence of the processes. Despite these differences, then, there is an underlying universality of cognitive processes' (Segall et al. 1990, p. 184).

It seems, however, that the deep structure might be influenced by culture in a more lasting way than has been assumed. In the cognitive sciences, the fact that a major part of the mental representations and processes studied might be of a cultural nature receives little attention. Frequently universality is simply postulated without taking into consideration the possibility of a cultural cogeneration. To sensitize cognitive scientists to the ambitious question of the range of cultural variability is a task for which no discipline is better suited than anthropology. A promising field of work for both disciplines could be the PDP-models (Parallel Distributed Processing), also called 'connectionistic models' or neuronal networks, which Rumelhart and McClelland (1986) developed as computer models. These are able to construct models of the functioning of schemata: not as loading in a set of instructions, but as gradually building up associative links among repeated or salient aspects of experience. This vagueness and dependency on empirical knowledge from our everyday life, hence of cultural phenomena, make these models attractive to anthropologists as well (Shore 1996, Strauss and Quinn 1997).

See also: Cognitive Modeling: Research Logic in Cognitive Science; Cognitive Science: History; Cognitive Science: Overview; Conceptual Blending; Cultural Studies of Science; Grammatical Relations; Mental Models, Psychology of; Mental Representation of Persons, Psychology of

Bibliography

Bloch M 1991 Language, anthropology and cognitive science. *Man* **26**: 183–98

Borofsky R 1994 On the knowledge and knowing of cultural activities. In: Borofsky R (ed.) *Assessing Cultural Anthropology*. McGraw-Hill, New York, pp. 331–46

Giddens A 1984 *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, Berkeley, CA

Goodenough W 1957 Cultural anthropology and linguistics. In: Garvin P L (ed.) *Reports of the Seventh Annual Round Table Meeting on Linguistics and Language Study*. Georgetown University Press, Washington, DC, pp. 167–73

Gumperz J, Levinson S C (eds.) 1996 *Rethinking Linguistic Relativity*. Cambridge University Press, Cambridge, UK

Hutchins E 1980 *Culture and Inference*. Harvard University Press, Cambridge, MA

Keesing R M 1972 Paradigms lost: The new ethnography and the new linguistics. *Southwestern Journal of Anthropology* **28**: 299–332

Keesing R M 1987 Models, 'folk' and 'cultural': Paradigms regained? In: Holland D, Quinn N (eds.) *Cultural Models in Language and Thought*. Cambridge University Press, Cambridge, UK, pp. 368–93

Quinn N, Holland D 1987 Culture and cognition. In: Holland D, Quinn N (eds.) *Cultural Models in Language and Thought*. Cambridge University Press, Cambridge, UK

Rosch E, Merris C 1975 Family resemblances: Studies in internal structure of categories. *Cognitive Psychology* **7**: 573–605

Rumelhart D E, McClelland J L (eds.) 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA

Schank R C, Abelson R P 1977 *Scripts, Plans, Goals and Understanding: An Enquiry into Human Knowledge Structures*. Erlbaum, Hillsdale, NJ

Segall M H, Dasen P, Berry J, Poortinga Y 1990 *Human Behavior in Global Perspective: An Introduction to Cross-cultural Psychology*. Allyn and Bacon, Boston

Shore B 1996 *Culture in Mind: Cognition, Culture, and the Problem of Meaning*. Oxford University Press, New York

Strauss C, Quinn N 1997 *A Cognitive Theory of Cultural Meaning*. Cambridge University Press, Cambridge, UK

Tyler S (ed.) 1969 *Cognitive Anthropology: Readings*. Holt, Rinehart, and Winston, New York

Wassmann J 1993 *Das Ideal des Leicht Gebeugten Menschen. Eine Ethno-Kognitive Analyse der Yupno von Papua Neuguinea*. Reimer Verlag, Berlin

Wassmann J, Dasen P R 1998 Balinese spatial orientation. Some empirical evidence of moderate linguistic relativity. *Journal of the Royal Anthropological Institute* **44**: 689–711

J. Wassmann

Cognitive Archaeology

Cognitive archaeology is an approach to research explicitly emphasizing the central role of cognition and mental phenomena in explanations of the past. It rejects behaviorism, with its emphasis on stimulus–response relationships, and the overriding concern with environmental adaptation which underlies much processual archaeology. The theoretical underpinnings of cognitive archaeology derive instead from post-positivist philosophy of science, cultural anthropology, linguistics, psychology and, more broadly, the cognitive neurosciences. Primary research topics include the origin of art, belief, language, tool use, and the human mind; and the reconstruction of prehistoric religion and ideology. Cognitive archaeologists look to a broad range of empirical data to investigate these interests, including stone tools, settlement patterns, ceramics, and art and iconography. Ethnohistorical

data are also commonly employed as a starting point, as many cognitive archaeologists are advocates of the direct-historical approach. While concerns with art and belief and an emphasis on individual cognition and motivation ally cognitive archaeologists with recent postmodernist concerns, typically they maintain a commitment to scientific method, including the testing of hypotheses and the development of scientific knowledge.

1. *Definitions*

Cognitive archaeology has been defined in various ways. According to Renfrew (1982) it is 'the archaeology of the mind.' Flannery and Marcus (1993, 261) define it as 'a study of all those aspects of ancient culture that are the product of the human mind … cosmology … religion … ideology … iconography … and all other forms of human intellectual and symbolic behaviour that survive in the archaeological record.'

These definitions reflect a rejection of a long-held archaeological belief that cognitive phenomena or mental products cannot be observed in the archaeological record, let alone reconstructed from it. Cognitive archaeologists, in contrast, commit a significant part of their effort in reconstructing the past to just this area of concern. While they do not deny the relevance of culture-history, technology, adaptation, subsistence, trade, and other traditional archaeological topics, cognitive archaeologists argue that a holistic interpretation or explanation of the past requires more than these traditional topics alone can provide.

2. *Philosophical and Theoretical Foundations*

The development of cognitive archaeology as an intellectual trend has involved the rejection of certain key tenets of scientific archaeology as practiced by Anglophone researchers (variously called new, processual or settlement-subsistence archaeology) in the latter half of the twentieth century. One of these tenets is logical positivism as a philosophy of science, with its emphasis on explanatory cover laws, the hard distinction between theory and fact, and a belief in unequivocal singular tests for hypotheses. Another is behaviorism which implicitly serves as the explanatory paradigm for much processual archaeology (Peebles 1992). This emphasizes the importance of stimulus-response relationships in explanation, especially concerning adaptation to the environment, and denies the relevance of mind, intellect, and cognition in human action, along with their byproducts such as belief, ritual, and art. To the processual archaeologist these are epiphenomenal, meaning derivative in origin and secondary in importance, and therefore analytically irrelevant.

Cognitive archaeology has instead adopted ongoing developments in the philosophy of science, including especially post-positivist approaches. Commonly these maintain a commitment to scientific knowledge, albeit acknowledging that over time science only increasingly approximates the 'real' truth, with scientific method based on 'inference to the best hypothesis' rather than singular critical tests (Kelley and Hanen 1988). Implicit in much cognitive archaeology moreover is *embodied scientific realism*. This accepts the existence of a world independent of our understanding of it, as well as the fact that we can have stable knowledge of this world. But it also holds that knowledge is relative to our bodies, minds, and interactions with the environment. This conclusion derives from empirical results of cognitive neurosciences studies which show that certain universal aspects of human conceptual development result from the embodied nature of the mind (Lakoff and Johnson 1999).

These philosophical commitments resolve two fundamental even if largely unrecognized conceptual contradictions in processual archaeology. First, by adopting embodied scientific realism, cognitive archaeology supports a model of humankind and its development that is fully reconcilable with evolutionary principles. Processual archaeology, even when explicitly claiming to be evolutionarily based, instead implicitly invokes a pre-Darwinian model of disembodied reason and mind due to its disavowal of any real relevance for cognitive phenomena. Because cognitive capabilities (which we all share) are ignored, they must be taken as a given; they 'just are' in processual archaeological theory. Ultimately this position then must resort to creationism and divine intervention to explain the totality of the human condition, which must include our ability to think and reason. Second, by insisting that mental phenomena are irrelevant but also conceding that the goal of archaeological science is to create knowledge (a cognitive construct), processual archaeology further requires a difference *in kind* between prehistoric peoples, whose thinking is putatively irrelevant and epiphenomenal, and contemporary westerners, including archaeologists whose professional purpose is to create knowledge. Such a distinction in kind between different peoples precludes the kinds of law-like explanations that processual archaeology has sought as its goal. Cognitive archaeology gives equal weight to the importance of mental phenomena in the lives of both prehistoric and contemporary humans.

Other theoretical aspects of cognitive archaeology derive from its relationship with cultural anthropology. These partly reflect the definition of culture as a cognitive worldview, a system of beliefs and meanings, thereby helping to situate aspects of cognitive archaeological research as a kind of anthropological archaeology in the interpretation of prehistoric culture. Similarly, anthropological studies of the nature

of traditional thought and belief systems (e.g., Sperber 1982) have dispelled the notion that these are necessarily irrational and therefore somehow inherently inaccessible to rationalist science. As anthropologist Robin Horton (1982) has instead shown, all cultures share a core of cognitive rationality that involves the development of theory, based on deductive, inductive, and analogical reasoning, used in the explanation, prediction, and control of events. The analysis and interpretation of culture and its expression in symbolism, meaning, and worldview need not then necessarily reduce to particularistic and empathic statements based on some kind of privileged access to the feelings of people in the past, but instead can be based upon empirically grounded generalizations that stand up to empirical analysis.

Perhaps most importantly, linguistic, psychological, and neuropsychological models of the human mind—brain and how it operates have been instrumental in cognitive archaeological studies. And while some of the topics that cognitive archaeologists include in their analyses, such as ritual, belief, symbolism, and meaning, are the same as those that are highlighted in humanist and postmodernist studies and approaches (called 'post-processualism' in archaeology), the models, techniques, and methods that the cognitive archaeologists employ to understand these aspects of the past are quite different from those that postmodernists might allow, as the scientific models used by cognitive archaeologists themselves imply.

3. *Analytical Approaches: Linguistic and Psychological Models*

A measure of the versatility of cognitive archaeology is seen in the fact that many of its applications have occurred at opposite ends of the archaeological spectrum: Paleolithic or Stone Age archaeology, which involves studies of the earliest human societies (even including in some cases the archaeology of prehuman hominids); and late prehistoric, protohistoric and historical archaeology, where direct cultural and linguistic links tie the archaeological past to recent peoples studied by ethnographers and historians.

Perhaps the first example of an in-depth cognitive archaeological study was provided by Paleolithic archaeologist André Leroi-Gourhan (1967) in his analysis of western European cave paintings dating from about 35,000 to 10,000 years ago. A similar approach was used by historical archaeologist James Deetz (1977) in his study of Euro-American Colonial artifacts. In both cases analysis was based on a structuralist model of mind. Originating in the linguistic theories of Ferdinand de Saussure and Roman Jacobson, structuralism posits a human mind that organizes empirical phenomena and concepts in terms of binary oppositions or dualities: black versus white; male versus female; good versus bad; etc. Structuralist analysis then concerns not so much the dual objects themselves but the relationships between the pairs. Both Leroi-Gourhan and Deetz showed that such a relational structure could be identified in their respective Paleolithic and Colonial data, thereby significantly amplifying their abilities to make inferences about the symbolic meanings of their archaeological remains.

Psychological models of the human brain–mind have figured prominently in other cognitive archaeological studies. Paleolithic archaeologist Thomas Wynn (1989), for example, has used psychologist Jean Piaget's theories on the development of thought in children and what these imply about spatial abilities to infer kinds of intelligence and levels of cognitive development in *Homo habilis* and *Homo erectus*, based on analyses of the kinds of stone tools that these ancient hominids created. Psychologist William Noble and archaeologist Iain Davidson (e.g., 1996) have collaborated in a series of studies, bringing together archaeological evidence on early symbol-making with theories of perception and communication in order to chart the evolution of the human mind and the appearance of language. Steven Mithen (1996) has invoked neural models of brain modularity along with cognitive scientist Howard Gardner's theory of multiple intelligences in an effort to explain the 'Upper Paleolithic Revolution': the seemingly sudden appearance of art, symbolism, and presumably belief roughly 50,000 years ago, literally tens of thousands of years after the earliest skeletal evidence for anatomically modern humans. In these cases psychological models have not been used to help explain the archaeological record: rather, the archaeological record has been employed in order to make inferences, informed by psychological theories, about prehuman and human intellectual and cognitive evolution.

4. *Analytical Approaches: Neuropsychological Models*

A subset of the psychological models used in cognitive archaeological studies involves *neuropsychology*: models based on theories of the mind–brain and the nervous system and how the two interact. These have been particularly influential in analyses of hunter-gatherer rock art tied to shamanistic religions where trance or altered states of consciousness (ASC) were considered key kinds of religious experiences. Because all humans share the same neural architecture, the mental, visual, emotional, and bodily reactions of ASC fall within a predictable range, regardless of cause, culture, or time period (Hobson 1994). This fact creates a kind of 'neuropsychological bridge' promoting analytical access to aspects of the ancient mind that involved ASC experiences.

Using the results of clinical studies, Lewis-Williams and Dowson (1988) constructed a model of the effects

of ASC on mental imagery—the 'visions' of a shaman's trance. Their model has three components: three progressive stages of ASC; the principles by which mental imagery is perceived at each stage; and the seven most commonly perceived entoptic phenomena, the phosphenes or geometric light images generated in the brain and visual system during the first stage of an ASC. After confirming the validity of their model using corpora of rock art known from ethnographic evidence to have been produced by shamans to depict visionary imagery, these archaeologists then used it to test whether European Paleolithic rock art, which lacks any ethnographic record, was also shamanistic in origin.

Central to their analysis is the use of *analogy*, but in this case analogy based on timeless and unchanging determining structures (human neuropsychology and its effects on ASC imagery), not analogy based on formal similarities. Note too that their concern was the *origin* of Upper Paleolithic art, not its *meaning*. While neuropsychology predicts human reactions to ASC, it does not tell us how such experiences will be interpreted or understood by different peoples or cultures and at different times. Subsequent neuropsychological studies of rock art have also considered the origins of universal aspects of shamanic symbolism, such as 'mystical flight' or 'death and rebirth,' as metaphoric expressions of the bodily and emotional hallucinations of ASC which were used verbally and graphically to describe what otherwise is a largely ineffable experience (Whitley 2000).

5. *The Direct Historical Approach*

Another analytical strategy, called the direct historical approach, involves the development of interpretive models of aspects of traditional cultures, like religion or ideology, based on ethnographic and ethnohistoric evidence. If derived from directly relevant cultures, these models can be cautiously applied to the prehistoric past to both chart continuity and identify change in cognitive systems. Marcus and Flannery (1994), for example, used sixteenth and seventeenth century Spanish documents to reconstruct Zapotec Indian religion in southern Mexico. Comparisons with archaeological evidence of religion and ritual indicated that the ethnohistorically described religious system first appeared between 200 BC and AD 100. Thomas Huffman (1996) has likewise used African Nguni ethnography concerning settlement structure, religion, and ideology to interpret the site of Great Zimbabwe, dating from the thirteenth to the fifteenth centuries.

Though sometimes criticized for projecting ethnography on to the prehistoric past, careful applications of the direct historical approach need do no such thing. Indeed, Huffman has shown that they can result in a rewriting of aspects of the ethnographic record when the archaeological evidence contributes subsidiary information not previously documented in writing. Moreover, the applicability of the direct historical approach is enhanced by the fact that religious and belief systems tend to be very conservative and slow to change, as other recent archaeological analyses have shown (e.g., Taçon et al., 1996).

6. *Future Trends*

Two circumstances favor the continued development and growth of cognitive archaeology. The first is its scientific robustness, based on the fact that many practitioners successfully incorporate different methodologies and kinds of data in their analyses and interpretations. Rock art researchers, for example, have combined the direct historical approach with neuropsychological and physical models, creating convergent scientific theories (e.g., Whitley et al., 1999). These are advantaged because the use of different kinds of data tested with distinct methodologies greatly enhances the degree of confirmation of their hypotheses and ensures that they do not follow from any one set of methodological assumptions. The results are scientific interpretations and explanations that are as empirically grounded and as well-tested as any that archaeology can offer.

Equally importantly, cognitive archaeology increasingly borrows from the cognitive neurosciences, an area of research that has advanced dramatically in the 1990s and which promises to influence greatly, if not change, all disciplines concerned with human behavior. Archaeologists have not yet exploited the range of information that is relevant to their research and that is currently available in the cognitive neurosciences, nor can we yet predict what future advances in this field will imply for archaeology. But what is certain is that, as cognitive neuroscientists improve our understanding of how the contemporary human mind-brain operates, this will contribute further to our understanding of cognitive aspects of the prehistoric past.

See also: Archaeology and Philosophy of Science; Archaeology and the History of Languages; Cognitive Anthropology; Cognitive Neuroscience; Cognitive Science: History; Cognitive Science: Overview; Structuralism

Bibliography

Deetz J J F 1977 *In Small Things Forgotten: The Archaeology of Early American Life*. Anchor Press-Doubleday, New York

Flannery K V and Marcus J1993 Cognitive archaeology. *Cambridge Archaeological Journal* **3**: 260–7

Hobson J A 1994 *The Chemistry of Conscious States: Toward a Unified Model of the Brain and the Mind*. Little, Brown and Company, Boston

Horton R 1982 Tradition and modernity revisited. In: *Rationality and Relativism*, edited by Hollis M and Lukes S, pp. 201–60, MIT Press, Cambridge, MA
Huffman T 1996 *Snakes & Crocodiles: Power and Symbolism in Ancient Zimbabwe*. Witwatersrand University Press, Johannesburg, South Africa
Kelley J H and Hanen M P 1988 *Archaeology and the Methodology of Science*. University of New Mexico Press, Albuquerque, NM
Lakoff G and Johnson M 1999 *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, New York
Leroi-Gourhan A 1967 *Treasures of Prehistoric Art*. Abrams, New York
Lewis-Williams J D and Dowson T A 1988 The signs of all times: entoptic phenomena in Upper Paleolithic art. *Current Anthropology* **29**: 201–45
Marcus J and Flannery K V 1994 Ancient Zapotec ritual and religion: an application of the direct historical approach. In: *The Ancient Mind: Elements of Cognitive Archaeology*, edited by Renfrew C and Zubrow E B W, Cambridge University Press, Cambridge, UK, pp. 55–74
Mithen S 1996 *The Prehistory of the Mind: A Search for the Origins of Art, Religion and Science*. Thames and Hudson, London
Noble W and Davidson I 1996 *Human Evolution, Language and Mind: A Psychological and Archaeological Inquiry*. Cambridge University Press, Cambridge, UK
Peebles C S 1992 Rooting out latent behaviorism in prehistory In: Gardin J-C, Peebles C S (eds.) *Representations in Archaeology*, Indiana University Press, Bloomington, IN, pp. 357–84
Renfrew C 1982 *Towards an Archaeology of Mind*. Cambridge University Press, Cambridge, UK
Sperber D 1982 Apparently irrational thoughts. In: *Rationality and Relativism*, Hollis M, Lukes S (eds.) MIT Press, Cambridge, MA, pp. 149–80
Taçon P S C,Wilson M and Chippindale C 1996 Birth of the Rainbow Serpent in Arnhem: land rock art and oral history. *Archaeology in Oceania* **31**: 103–24
Whitley D S 2000 *The Art of the Shaman: The Rock Art of California*. University of Utah Press, Salt Lake City, UT
Whitley D S, Dorn R I, Simon J M Rechtman R and Whitley T K 1999 Sally's rockshelter and the archaeology of the vision quest. *Cambridge Archaeological Journal* **9**: 221–247
Wynn T 1989 *The Evolution of Spatial Competence*. University of Illinois Press, Urbana, IL

D. S. Whitley

Cognitive Control (Executive Functions): Role of Prefrontal Cortex

Control refers to the ability to direct mental function and behavior in accord with an internally represented set of intentions. This is manifest in higher cognitive function in many forms: the ability to direct attention to a specific stimulus within a large array of other competing, and perhaps more salient stimuli (e.g., finding the face you are looking for within a crowd), maintaining a new and important piece of information in mind against distraction (e.g., remembering a telephone number you just got from directory assistance until you dial, while in a phone booth at a noisy airport terminal), overcoming a compelling but undesirable behavior (e.g., suppressing the urge to scratch a mosquito bite) or pursuing a complex but unfamiliar behavior (e.g., playing a new piano piece) and, perhaps most importantly, the ability to respond flexibly and productively in novel circumstances (e.g., mentally explore the consequences of a complex sequence of moves in a game of chess).

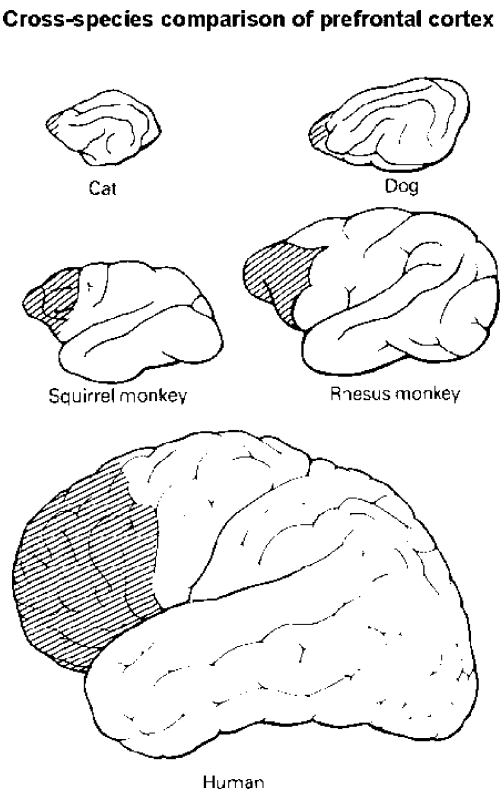

Figure 1
Relative size of prefrontal cortex (grayed region) in five species (http://www.psychol.ucl.ac.uk/kate.jeffery/C567/Lecture13_Frontal/overheads/img003.jpg)

The distinction between controlled and automatic processing is one of the central concepts within modern cognitive psychology. Controlled processing is considered to be effortful, and to rely on a limited capacity system, while automatic processing is assumed to occur independently of this system (Posner and Snyder 1975, Shiffrin and Schneider 1977). This concurs with common experiences, such as the ability to carry on a conversation while driving a car (an automatic process) but not while conducting multi-

digit arithmetic in one's head (a process that relies on control). Many contemporary theories posit that there is actually a continuum between controlled and automatic processing (for a review, see Cohen et al. 1990). Nevertheless, virtually all theorists acknowledge the need for some mechanism, or set of mechanisms, responsible for the coordination of processing in a flexible fashion—particularly in novel or demanding tasks. This idea figures centrally in Baddeley's classic theory of working memory (Baddeley 1986), which postulates two critical components: a storage component responsible for the active maintenance of information in a short-term store, and a executive control component responsible for the manipulation and coordinated use of this information. For example, in a multi-digit multiplication problem, the storage component maintains the intermediate products while the executive carries out the arithmetic operations.

The postulation of a central executive closely paralleled theorizing regarding the nature of frontal lobe function (for a review, see Shallice 1982), based on the clinical observation that patients with frontal lesions exhibit a 'dysexecutive syndrome.' The frontal lobes are the area of the brain most highly expanded in humans relative to other species (see Fig. 1). Damage to this area is associated with impairments in characteristically human cognitive functions that are directly dependent upon control (such as planning, or the ability to respond adaptively in novel circumstances), and disturbances of this structure have been implicated consistently in neuropsychiatric diseases such as schizophrenia that also appear to be uniquely human. Indeed, the earliest neurologists, neuroscientists and neuropsychologists recognized the importance of the frontal lobes in the control of behavior. Perhaps this observation was made most dramatically in the classic case of Phineas Gage, the unfortunate railroad foreman who suffered damage to his prefrontal cortex (PFC) when a 1-1/2″-diameter rod penetrated his skull in a construction accident. Originally someone who was considered to be thoughtful, responsible, and of sound judgment, following the accident he was described as 'capricious … and unable to settle on any of the plans he devised for future action' (Harlow 1848). Such changes have been observed repeatedly in patients with damage to the frontal cortex (for a review, see Stuss and Benson 1986). Furthermore, patients with frontal damage perform poorly in tasks that require even the simplest forms of cognitive control, such as the Wisconsin Card Sort Task (WCST) (Milner 1964), and, most recently, brain imaging studies consistently have revealed increased activity of the prefrontal cortex while subjects are performing tasks that demand cognitive control (for a review, see Miller and Cohen 2001). Although this accumulation of findings has led most investigators to assume that the prefrontal cortex plays a critical role in cognitive control, they have provided little insight into the specific contribution that it makes, or the mechanisms by which it operates. However, more detailed neurobiological studies have begun to shed light on this question.

Using single unit recording techniques, Fuster and Alexander (1971) and Niki (1974) reported the remarkable finding that some neurons in the prefrontal cortex continue to fire during the delay between a stimulus and a contingent response, even after the stimulus has disappeared. Goldman-Rakic and her colleagues followed up on this finding and, in a series of elegant studies, demonstrated that the firing of such neurons was stimulus-selective, and was critical for performance in delayed-response tasks (Goldman-Rakic 1990).

These findings strongly implicate neuronal activity in PFC as the site of temporary storage of information in working memory. Recent neuroimaging studies of humans have provided strong convergent support for this idea, demonstrating sustained activity in PFC during the performance of simple working memory tasks (for a review, see Smith and Jonides 1999). However, this discovery seemed to pose an interesting puzzle. Most of the neuropsychological data suggested that the PFC housed the central executive which, at least within Baddeley's influential theory of working memory, was clearly distinguished from the storage component thought by most to be housed primarily in more posterior structures. Computational models of cognitive control have offered a different perspective on this problem that may reconcile the role of the PFC in control and storage.

1. *Computational Models*

Shallice (1982) presented the first account of the role of PFC in cognitive control within an explicitly computational framework. He proposed that the PFC housed a supervisory attentional system (SAS)—a mechanism by which PFC coordinates complex cognitive processes. Shallice's theory was described in terms of a production system architecture. This has appeal, as it relates the executive functions of frontal cortex to the well characterized mechanisms of other production system theories, which include the active representation of goal states to coordinate the sequences of production firings involved in complex behaviors (e.g., Anderson 1983). One feature of goal representations is that they must be maintained actively throughout the course of a sequence of actions to direct behavior effectively. This coincides with the observation that PFC appears to be specialized for the active maintenance of task-relevant information. Thus, it is possible that PFC is specialized for the maintenance of a *particular type* of information, as a means of executing control over performance.

While Shallice's theory was not implemented originally as a functioning model, Kimberg and Farah (1993) proposed a model of frontal function—also

using a production system architecture—that simulated performance in a variety of tasks considered to rely on frontal lobe function (including the WCST), and illustrated that damage to this component of the model produced impairments similar to those observed in patients with frontal damage. However, while such models have produced insights into the functional role of PFC in cognitive control, their components do not have a transparent mapping on to specific neurobiological mechanisms. Neurobiological plausibility is not a requirement, *per se*, of a theory that seeks to explain the cognitive functions of PFC. Nevertheless, the question of how goal directed behavior arises from the firing of millions of neurons presents a mystery in its own right. Furthermore, understanding how different forms of cognitive impairment arise from different forms of damage to PFC (e.g., the effects of stroke or injury vs. schizophrenia) provides an important motivation for understanding how cognitive control may arise from the specific neurobiological mechanisms housed within PFC.

Recently, investigators have begun to use neural network models to better understand how PFC may carry out its functions. Such models (also known as connectionist, or parallel distributing processing models) simulate the behavioral performance of human subjects (or animals) in cognitive tasks using neurally-plausible processing mechanisms (e.g., the spread of activity among simple processing units along weighted connections; see Rumelhart and McClelland 1986). The goal of this effort is to identify principles of neural function that are most relevant to behavior. Using this approach, Dehaene and Changeux (1989), Levine and Prueitt (1989), and Cohen and Servan-Schreiber (1992) have all described models of prefrontal function, and used these to simulate the performance of normal and frontally-damaged patients in tasks that are sensitive to PFC damage, such as the WCST and others. All of these models simulate PFC function as the activation of a set of units that represent the 'rules' of the task; that is, units whose activation leads to a task-relevant response, even when this may not be the one most strongly associated with the stimulus. However, in most models, the PFC units themselves are not responsible for generating the response directly. Rather, they influence the activity of other units whose responsibility this is. This is clearly illustrated by a model of the Stroop task developed by Cohen et al. (1990).

In the Stroop task (Stroop 1935), subjects respond to words either by reading them, or by naming the color in which they are displayed. In the critical condition, the word itself conflicts with the color in which it is displayed (e.g., the word GREEN displayed in red). Subjects have no trouble reading such words (e.g., saying 'green'). However, when they are asked to name the color (e.g., say 'red'), they are significantly slower, and sometimes even make errors. This reflects the highly practiced, and therefore prepotent tendency

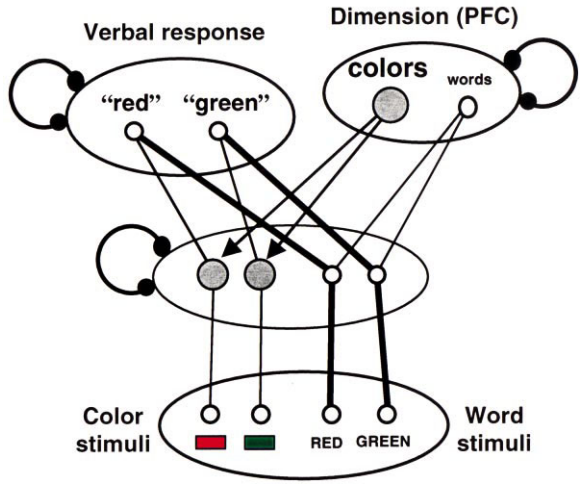

Figure 2
Model of the Stroop Task (adapted from Cohen et al. 1990), showing the pathways for word reading, color naming, and top-down control. Connections between units in different layers are excitatory and within layers are inhibitory (closed circles). Shown with color dimension unit active, biasing color-naming pathway

to read written words, which interferes with the ability to name the color of a Stroop conflict stimulus. The ability to name the color in the face of such interference (that is, to produce the weaker, but task-relevant response) is a simple but clear example of cognitive control.

Cohen et al. (1990) constructed a model of this task, as shown in Fig. 2. The connections among the units in this model defined two processing pathways, one for word reading and another for color naming. The connections in the word-reading pathway were stronger, capturing the assumption that this was the more practiced task. Because of these stronger connections, information flowing along the word pathway interfered with color naming, simulating the interference effects observed when human subjects perform this task. Indeed, the model's ability to produce a response to the color in the face of such interference required the addition of a set of units (labeled 'Dimension' in Fig. 2), which provided additional activation of the units in the color-naming pathway. This additional activation biased processing in favor of this pathway, allowing it to compete more effectively with, and prevail over, activity flowing along the stronger word pathway. This biasing effect corresponds precisely to the role of top-down control in neurophysiological models of attention (such as the Biased Competition Model of Desimone and Duncan 1995), and has been proposed as a mechanism by which PFC exerts control over processing (Cohen and Servan-Schreiber 1992, Miller and Cohen 2001).

Models such as this have been used to simulate the performance of both normal subjects and patients with frontal damage in a wide range of tasks that tap cognitive functions commonly associated with PFC function, such as working memory, attention, behavioral inhibition, planning, and problem solving.

2. Guided Activation as a Method of Control

The Stroop model brings several features of PFC function into focus. First, it emphasizes the view that the role of the PFC in control is modulatory, guiding the flow of activity along pathways in other parts of the brain that are responsible for task performance. For example, activating the color unit does not in itself transmit information about a particular response (red or green). Rather, it simply insures that activity flowing along the color-naming pathway will have a greater influence over the response than activity flowing along the word pathway. In this way, representations within the PFC can function as intentions, rules, or goals (comparable to those in a production system architecture), by setting up the appropriate relationship between a stimulus (or category of stimuli) and an associated response (or set of responses) through the proper guidance of activity along pathways in other parts of the brain. Recent neurophysiological findings regarding the firing properties of neurons in PFC provide strong support for this view (for a review, see Miller and Cohen 2001). Recent neuroimaging studies also provide convergent support, indicating that PFC activity occurs when behavior relies upon explicit knowledge about rules or arbitrarily determined conjunctions of stimulus features (for a review, see Miller and Cohen 2001). Note that this function is not necessarily restricted to mappings from stimuli to responses, but applies equally well to mappings involving internal states (e.g., thoughts, memories, emotions, etc.). Note also that this function is not necessarily unique to PFC. There may also be more local forms of control, responsive to regionally specific needs for the biasing of processing. However, the wide range of anatomic connections that the PFC shares with virtually all associative areas of the brain places it in a strategic position to guide the flow of activity between and among these other regions (for a review, see Miller and Cohen 2001)—a position well suited to its presumed role in the control of higher cognitive processes.

The emphasis of this guided activation model of PFC on its modulation of other brain areas actually responsible for task execution is consistent with the classic pattern of neuropsychological deficits associated with frontal lobe damage: The individual elements of a complex behavior are usually left intact, but the subject is not able to coordinate them in a task-appropriate way (for example, a patient who, when preparing coffee, first stirred and then added cream; Shallice 1982). The guided activation model also captures another critical feature of theories about the role of PFC in executive function: the importance of sustained activity as a critical component of control. For a representation to have a biasing influence, it must be activated. For this influence to endure (e.g., over the course of performing a task), it's activity must be sustained. For example, to continue color naming, the activity of the color unit must be maintained, lest the word begin to dominate processing. Similarly, an increase in the demand for control requires greater or more enduring activity of the corresponding units in PFC. This concurs with accumulating evidence from the neuroimaging literature, that tasks thought to rely on controlled processing consistently engage the PFC (for a review, see Smith and Jonides 1999). It is also consistent with both behavioral and neuroimaging findings regarding the effects of practice on automaticity and the involvement of PFC. Increased practice on a task should strengthen its underlying pathway, reducing its reliance on control. Indeed, consistent practice on a task reduces the amount of PFC activity observed, and PFC damage can impair new learning but spare performance on well-practiced tasks (for a review, see Miller and Cohen 2001).

Finally, this approach helps to unify the role that PFC plays in the variety of cognitive functions with which it has been associated, such as selective attention, behavioral inhibition, and executive function in working memory. These can all be seen as varying reflections, in behavior, of the operation of a single underlying mechanism of cognitive control: the biasing effects of representations in PFC on processing in pathways responsible for task performance. For example, selective attention and behavioral inhibition may be viewed as two sides of the same coin: Attention is the effect of biasing competition in favor of task-relevant information, and inhibition is the consequence that this has for the irrelevant information (cf. the Biased Competition Model of Desimone and Duncan 1995). This assumes that inhibition occurs due to local competition between conflicting representations (e.g., between the two responses in the Stroop model), rather than centrally by the PFC. The 'binding' function of selective attention can also be explained by such a mechanism, by assuming that PFC representations can select the desired combination of stimulus features (over other competing combinations) to be mapped on to a response.

3. Outstanding Questions

The guided activation model provides an integrative, and mechanistically explicit framework for considering the role of PFC in cognitive control. At the same time, it brings into focus a number of important

and unanswered questions. For example, despite the longstanding observation of delay-period activity for PFC neurons, and the centrality of this property in neural network models of PFC, little is known about the actual mechanisms by which neuronal activity is sustained. This could reflect a cellular property of PFC neurons (e.g., bistability), or a circuit-level phenomenon (e.g., recirculation of activity within the PFC, or between PFC and other structures). Several models have proposed the latter, assuming that representations are maintained in PFC as attractor states (for a review, see O'Reilly et al. 1999). However, neurobiological studies are needed to confirm this hypothesis. A closely related question concerns the mechanisms by which patterns of activity are updated in PFC. These must be able to satisfy two conflicting demands: On the one hand, they must be responsive to relevant changes in the environment; and on the other, they must be resistant to updating by irrelevant changes. Neurophysiological studies suggest that PFC representations are selectively responsive to task-relevant stimuli and robust to interference from distractors (for a review, see Miller and Cohen 2001). Not surprisingly, two hallmarks of damage to PFC are perseveration (failure to update) and distractibility (inappropriate updating). These observations suggest the operation of mechanisms that insure the appropriate updating of PFC activity in response to behavioral demands.

Recent modeling work has suggested that brainstem dopaminergic systems and the basal ganglia may play an important role both in updating representations in PFC and learning how and when to do so (for a review, see Miller and Cohen 2001). Other work has suggested that the anterior cingulate cortex—a midline structure within the frontal lobes—may play an important role in monitoring task performance, and identifying the need to allocate control (Botvinick et al. 2001). Such theories make important predictions about the neural mechanisms underlying cognitive control that serve as valuable challenges to future neurobiological research in this area.

An equally fundamental question concerns the nature of representations in PFC and how they arise. For example, the Stroop model assumes there are units that represent each of the two dimensions of the stimulus (color and word), with the appropriate connections to the corresponding pathways. However, are there such units in the PFC for every possible combination of stimulus and response of which tasks may be composed? This seems unlikely. Yet, there must be sufficient representational richness to support the flexibility in behavior that the PFC seems to afford. What principles define this set of representations, how they are organized, and how they are learned are important questions at both the computational and neurobiological levels.

A better understanding of the mechanisms underlying active maintenance may also provide insight into one of the most striking and perplexing properties of cognitive control: its severely limited capacity. This has long been recognized in cognitive psychology (Posner and Snyder 1975, Shiffrin and Schneider 1977), and is painfully apparent to anyone who has tried to talk on the phone and read email at the same time. The resource limitation of cognitive control has played an explanatory role in many important models of human cognition. However, to date, no theory has provided an explanation of the limitation *itself*. This is a *sine qua non* of cognitive control, and therefore provides an important benchmark for any theory that seeks to explain its underlying mechanisms.

Finally, it is important to recognize that the PFC is certainly not the only brain structure involved in cognitive control. As noted above, mechanisms similar to those within PFC may operate locally in other parts of the brain. Furthermore, there are certainly other types of mechanism critical to cognitive control. For example, the mechanisms responsible for keeping an immediate goal in mind (e.g., working on a book chapter) are not likely to be the same ones responsible for realizing long-term goals (e.g., getting the book published). While the former may be guided by representations actively maintained in PFC, the latter almost certainly engage mechanisms of long-term storage. Some suggestions have been made about how the PFC may interact with the hippocampus to orchestrate the storage and retrieval of goal representations at appropriate times (O'Reilly et al. 1999), however, this remains another area in need of further research.

4. *Conclusion*

One of the great mysteries of the brain is how purposeful, goal-directed behavior emerges from the millions of relatively simple processing units that are its basic computational elements. Behavioral, neuropsychological, and neurobiological data converge on the idea that the frontal lobes play a critical role in cognitive control. Neural network models have begun to suggest how this may be carried out, as sustained patterns of activity within PFC modulate, or 'guide' the flow of activity along pathways in other parts of the brain responsible for performing a task. However, many fundamental questions remain to be addressed. The human brain is arguably the most complex device in the known universe, and its capacity for higher cognitive function continues to be one of its deepest mysteries. Unraveling this mystery stands as one of the most exciting challenges in science, and the rapid development of sophisticated new empirical methods (such as functional brain imaging) and theoretical tools (such as neural network modeling) offer hope that this challenge can be met.

See also: Dysexecutive Syndromes; Prefrontal Cortex; Prefrontal Cortex Development and Development of Cognitive Function

Bibliography

Anderson J R 1983 *The Architecture of Cognition*. Harvard University Press, Cambridge, MA

Baddeley A 1986 *Working Memory*. Clarendon Press, Oxford, UK

Botvinick M M, Braver T S, Carter C S, Barch D M, Cohen J D 2001 Conflict monitoring and cognitive control. *Psychological Review*

Cohen J D, Dunbar K, McClelland J L 1990 On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychology Review* **97**: 332–61

Cohen J D, Servan-Schreiber D 1992 Context, cortex and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychology Review* **99**: 45–77

Dehaene S, Changeux J P 1989 A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience* **1**: 244–61

Desimone R, Duncan J 1995 Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* **18**: 193–222

Fuster J M, Alexander G E 1971 Neuron activity related to short-term memory. *Science* **173**: 652–4

Goldman-Rakic P S 1990 Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. *Progress in Brain Research* **85**: 325–35

Harlow J M 1848 Passage of an iron rod through the head. *Boston Medical and Surgical Journal* **39**: 389–93

Kimberg D Y, Farah M J 1993 A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex organized behavior. *Journal of Experimental Psychology* **122**: 411–28

Levine D S, Prueitt P S 1989 Modeling some effects of frontal lobe damage—novelty and perseveration. *Neural Networks* **2**: 103–16

Miller E K, Cohen J D 2001 An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* **24**: 167–202

Milner B 1964 Some effects of frontal lobectomy in man. In: Warren J M, Akert K (eds.) *The Frontal Granual Cortex and Behavior*. McGraw-Hill, New York, pp. 313–31

Niki H 1974 Prefrontal unit activity during delayed alternation in the monkey. 1. Relation to direction of response. *Brain Research* **68**: 185–96

O'Reilly R C, Braver T S, Cohen J D 1999 A biologically-based computational model of working memory. In: Miyake A, Shah P (eds.) *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge University Press, New York

Posner M I, Snyder C R R 1975 Attention and cognitive control. In: Solso R L (ed.) *Information Processing and Cognition*. Erlbaum, Hillsdale, NJ

Rumelhart D E, McClelland J L 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA

Shallice T 1982 Specific impairments of planning. *Philos. Trans. R. Soc. London Ser. B* **298**: 199–209

Shiffrin R M, Schneider W 1977 Controlled and automatic human information processing: II. Perceptual learning automaticity, attending and a general theory. *Psychology Review* **84**: 127–90

Smith E E, Jonides J 1999 Storage and executive processes in the frontal lobes. *Science* **283**: 1657–61

Stroop J R 1935 Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* **18**: 643–62

Stuss D T, Benson D F 1986 *The Frontal Lobes*. Raven, New York

J. D. Cohen

Cognitive Development: Child Education

By the nature of their subject matter, education and cognitive development are closely intertwined. The goal of education, to produce knowledgeable problem solvers, who can apply their knowledge flexibly in real world contexts, and who have the skills and motivation to acquire effectively new knowledge and understanding, is what models of cognition and cognitive development attempt to describe and explain—that is, how we learn, remember and know, and how these processes change and are affected by the interplay of environment and person over development.

This article first presents a brief historical overview and summary of the assumptions underlying psychological and especially cognitive developmental theories and research as they pertain to education, and then considers a set of current issues that highlight the link between the two and the potential for dynamic interaction. 'Education' is clearly a broad term that can encompass learning throughout the lifespan. In this article the central focus will be on education defined as formal instruction during the school years.

1. Historical Overview

Historically, the impact of understanding of cognitive development on educational practice can be traced to the influences of psychology in general. The great movements of functionalism and behaviorism that shaped psychology in the first half of the twentieth century also affected models of education and educational practice. The subsequent 'cognitive revolution' affected education in a variety of ways. Theory and research on cognitive skill development (e.g., strategy acquisition and use) and on conceptual structure (e.g., concept organization and change; logicomathematical reasoning) led to a concern with facilitating reasoning and problem-solving strategies, and with facilitating concept acquisition and retrieval, especially through active, hands-on problem solving.

Although educational practice has been influenced at the most general level by ideas about cognition and cognitive development, a dynamic interplay between cognitive developmental research and educational practice has been neither widespread nor systematic in the typical classroom. This is not surprising, as noted by Olson and Bruner (1996, p. 9): 'Theoretical knowl-

edge of how cognition develops continues to grow but just how to relate this knowledge to the practical contexts … to educate … remains almost as mysterious as when such efforts first began.' This is beginning to change as new tools emerge. Over the past decades, cognitive science research has produced detailed analyses of tasks and learners that allow more sophisticated models of the knowledge that a learner brings to the educational setting, and of the component skills necessary for successful performance within specified domains such as mathematics and physics. In addition, there are powerful models of early and late conceptual skill development, and of reasoning processes that emerge within and outside of formal education. These models carry the promise of better specification of the design and implementation processes for successful learning environments.

1.1 Specific Influences

The influence of cognitive developmental theory and research on education has changed as models of cognition, learning, and education have altered. During the heyday of behaviorism, especially in the USA, the explanatory mechanisms underlying both development and education were based on elementary principles of associative learning; educational strategies were based on behavior analysis, association, and reinforcement contingencies; and educational outcome was assessed in quantitative terms. At the beginning of the twenty-first century, although general learning principles inform educational practice (e.g., the relative value of internal vs. external reinforcement, the effectiveness of spaced vs. massed trials, learning and forgetting curves, curricula organized as an accumulation of simple units) a strict learning model approach has largely been restricted to behavioral modification programs and some programmed instruction, most notably for foreign language learning.

The most general result of the cognitive revolution was a new focus on the active, problem-solving child and classroom. Although an oversimplification, two influences were particularly important. One influence is from the broad class of structuralist models, most notably Jean Piaget's genetic epistemology. Most of these models have been based on the assumption that the child is an active participant in the acquisition and organization of knowledge—that is, that there is a dynamic interplay between the knower's current level of understanding and the information and problems presented by the environment. This perspective led to a focus on understanding the child's view of events, on describing the nature and organization of concepts and reasoning structures, and on matching instruction to developmental level. A second influence, arising from the broad array of information-processing models of cognitive functioning, led to a focus on specifying task structure and problem solving strategies, and on the information processes underlying memory, reasoning, and learning.

2. Cognitive Development

Examples of the specific influences of cognitive development theories and research can be captured by answers to the questions: who is the learner, what is being learned, and how does learning proceed?

2.1 Who Is the Learner?

Probably one of the most revolutionary ideas of scholarly study and research on childhood and education undertaken in the twentieth century was that neither education nor cognitive development can be understood without taking account of the active child/learner. Although this view informed ideas about education from an early point (cf. White 1992), it was strengthened by widespread interest in the theories of Jean Piaget and their adaptation to educational issues, especially mathematics and science education. The central idea is that the child actively assimilates new information on the basis of general reasoning structures, which themselves change in regular ways over the course of development. Recent research on the cognitive abilities of the very young child and infant have complemented this idea to suggest that knowledge and knowledge acquisition in some domains (such as language, arithmetic, and the elementary natural and social worlds) is precocious. One conclusion from this work is that the child brings to the formal educational context a set of robust ideas (or 'theories') about the physical, social, and natural worlds.

The goal of information-processing/cognitive science models is to develop descriptions of learners, tasks, processes, and performance that will allow a detailed specification of the componential skills and processes underlying cognitive activities. From this more information-processing perspective, the learner is seen as a consumer who gathers, stores, organizes, and uses information in problem solving, and who, with development and instruction, becomes increasingly adept at directing, monitoring, and manipulating these skills strategically across a variety of content areas.

2.2 What Is Being Learned?

There is of course no single 'cognitive developmental' answer to the question of what is learned, but rather a family of answers that have in common a focus on knowledge and mental representation as the products of experience and/or education.

The gist of most structuralist viewpoints is that cognitive development consists of qualitative changes in the ways that concepts (i.e., knowledge about the mathematical, material, and social worlds) are organized. These changes enable the child to reason in increasingly complex ways. The description of conceptual structure varies: according to Piagetian models, conceptual structure is described in terms of general and universal logical-mathematical relations; other models refer to causal or semantic structures that may vary across content domains.

The answer to 'what is learned' from an information-processing perspective on cognitive development is skill-based. A large number of cognitive developmental studies have documented changes in the speed and control of processing skills (perception, attention, and memory), and associated increases in cognitive performance, as well as changes in conceptual organization and rule use. Catalyzed by a seminal paper by Flavell (1979) on metacognition and cognitive monitoring, cognitive developmental research focused on the acquisition of cognitive processing strategies, and metacognitive skills and knowledge, as important mechanisms of developmental change in cognitive performance. Teaching these skills to improve general performance across content area (e.g., self-monitoring, strategy retrieval and use) has informed research and practice, as has teaching these skills within a variety of content areas, including reading, writing, mathematics, and science.

2.3 *How Does Learning Proceed?: Implications for Education*

The central tenet of constructivist models is that cognitive change/learning is a reciprocal process of interpreting information according to current conceptual structure and adapting that structure to task demands. Learning does not proceed by recording received information directly or passively. Learning proceeds at best according to constructivist models, when the learner can experiment actively to discover, invent or infer the solutions to problems. This approach is reflected in active, hands-on classrooms, where a host of strategies attempt to foster active engagement: the discovery method, activity centers, and inquiry-based instruction that focuses on acquiring principles, not just facts (e.g., Hedegaard and Lompscher 1999).

From a cognitive science/information-processing perspective, learning proceeds by acquiring declarative and procedural knowledge within content domains, domain-specific and domain-general rules for operating with that content, strategies and skills for interpreting and attending to new information, and an ability to monitor, access, and control these skills intentionally, including learning how to learn (cf. Bruer 1993, Bransford et al. 1999).

3. *Contemporary Issues for Cognitive Development and Education*

The following section discusses four complementary areas in cognitive developmental research that have a strong potential to affect educational models and practice: (a) the metaphor of the child as 'universal novice', (b) conceptual change models—the role of naive theories and 'misconceptions', (c) individual differences, (d) cognitive development and education as social phenomena.

3.1 *The Development of Novice-to-Expert*

The discovery that experts and novices, regardless of age, differ with respect not only to the sheer amount of knowledge they possess, but also with respect to the organization of that knowledge, offers a metaphor for the processes of education in which universal novices become educated experts. Many studies have demonstrated that prior knowledge predicts learning outcomes, and that the knowledge base of novices in any particular domain differs from that of experts in systematic ways (Bransford et al. 1999). This approach has the potential to make a powerful impact on educational practice because it suggests that the acquisition of domain-specific information, not just domain-general skills, is necessary for deep knowledge acquisition. Expert content knowledge can compensate for other skills that predict performance, such as age, aptitude or metacognitive skills. The expert–novice distinction and attempts to design curricula to move novices to become experts has become a catchword of general education in math and science, and in specialized curricula such as medicine.

3.2 *Intuitive Knowledge and Misconceptions*

Studies of reasoning, concept formation, and conceptual understanding in young children have demonstrated convincingly that there is probably no period during development when the child is a 'blank slate.' Rather, even young children show rich knowledge about a variety of concept domains, including number, biological kinds, physics, social phenomena, and the like. Although this knowledge can provide a strong initial base for formal instruction, some early knowledge may clash with information presented in formal learning contexts for mathematical, literary, and scientific domains, because the concepts to be learned in these domains do not match everyday experience. For example, children's difficulties in performing mathematical operations on fractions or negative numbers arise in part from inappropriately generalizing knowledge about natural, whole numbers (Gelman 1994). Similarly, informal knowledge about biological kinds, movement through space, physical causality, and even cosmology, can clash with learning

about formal, complex systems in school (e.g., acceleration, force, gravity, and biology). Most clearly demonstrated in math and science, the study of 'misconceptions' has shown the power of children's implicit and everyday concepts. It underscores the importance of basing instruction on a good diagnosis of current understanding, and of devising ways to promote conceptual change through active experimentation and confrontation in rich everyday contexts. There is general agreement that 'a logical extension of the view that new knowledge must be constructed from existing knowledge is that teachers need to pay attention to the incomplete understandings, the false beliefs, and the naive renditions of concepts that learners bring with them to a given subject' (Bransford et al. 1999, p. 10).

3.3 *Individual Differences*

In its attempts to explain mental growth and change, mainstream cognitive developmental research has focused more on phenomena that are believed to characterize development universally, and less on individual differences. Nonetheless, a number of traditional and emerging areas suggest that systematic differences not only in cognitive style and cognitive strategies, but also in more basic conceptual structure, may provide a means to tailor educational practice more closely to children's specific learning needs.

For example, well documented sex differences in spatial skills that may be tied to differences in representation mode or organization, not just to experiential differences, may be exploited to develop alternative methods of mathematics instruction. Analogously, descriptions of multiple forms of intelligence (cf. Gardner 1993) may offer to the teacher different approaches to a topic and different modes of presenting key concepts (Bransford et al. 1999).

Detailed studies tracing the processes of conceptual change, strategy acquisition and discovery, and the like, also illustrate the large range of individual differences in developmental rate, style, and pattern. Microgenetic studies, i.e., investigations that follow the emergence, development, and consolidation of cognitive skills at an intensive individual level over a period of time, have been applied to a variety of content domains such as language, mathematical skills and problem solving, scientific reasoning, and memory and concept development (cf. Siegler and Crowley 1991, Weinert and Schneider 1999). These studies have illustrated the large variability in skill learning and use, and have shown that average developmental functions do not characterize developmental change at the individual level. Insight into the conditions that facilitate problem solving and that help consolidate newly formed competencies may inform the development of more individualized learning assessment and curricula.

3.4 *Cognition in Context*

Researchers have recently returned to classic questions concerning the role of culture in cognition, the importance of context and motivation in explaining and understanding cognition in everyday contexts, and the influences of formal and informal learning contexts on cognitive development. Two phenomena have heightened such interest: national differences in cognitive performance, especially in mathematics and science; and research findings showing large discrepancies between cognitive performance in formal educational settings and informal everyday contexts. Both of these perspectives have motivated new research on the types and effects of formal and informal education and have amply illustrated the effects of schooling on a variety of cognitive tasks tapping mathematics, logic, classification, and memory strategies (cf. Rogoff and Chavajay 1995).

A cognitive model that has been highly influential in education is based on the theories of Lev Vygotsky. Vygotsky (1962) characterized cognition as the internalization of external and culturally transmitted structure, rules, and principles that are mediated by language. According to this model, development proceeds most effectively when there is adequate environmental support within the 'zone of proximal development,' a construct to indicate the difference between a child's actual and potential performance. The zone of proximal development is usually measured as the difference between tasks a child can solve working independently, and those a child can solve with assistance from adults, instructors or other competent models. This approach underlies 'reciprocal education' and 'reciprocal teaching', in which the learner acquires strategies from expert models in social settings. The educational goal is to develop supporting social contexts in which a 'community of learners' collaborates in fostering learning outcomes (Brown and Campione 1994).

4. *Conclusions*

As noted above, mainstream cognitive developmental research and theory have influenced educational practice at only the most general levels. As the relatively new field of multidisciplinary cognitive science has become established and institutionalized, its methods and results are being tested in school contexts (cf. Bruer 1993, Bransford et al. 1999). The fields of cognitive development and education are ripe for forging collaborations that allow the science of learning to inform the practice of education in classroom contexts.

See also: Cognitive Development in Childhood and Adolescence; Cognitive Development: Learning and Instruction; Instructional Technology: Cognitive Science Perspectives; Piaget, Jean (1896–1980);

Piaget's Theory of Human Development and Education; Situated Cognition: Contemporary Developments; Situated Learning: Out of School and in the Classroom; Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Bransford J, Brown A L, Cocking R (eds.) 1999 *How People Learn: Brain, Mind, Experience and School*. National Academy Press, Washington, DC

Brown A L, Campione J C 1994 Guided discovery in a community of learners. In: McGilly K (ed.) *Classroom Lessons: Integrating Cognitive Theory and Classroom Practice*. MIT Press, Cambridge, MA

Bruer J T 1993 *Schools for Thought*. MIT Press, Cambridge, MA

Flavell J H 1979 Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* **43**(10): 906–11

Gardner H 1993 *Frames of Mind*. Basic Books, New York

Gelman R 1994 Constructivism and supporting environments. *Human Development* **6**: 55–82

Hedegaard M, Lompscher J (eds.) 1999 *Learning Activity and Development*. Aarhus University Press, Aarhus, Denmark

Olson D, Bruner J 1996 Folk psychology and folk pedagogy. In: Olson D, Torrance N (eds.) *The Handbook of Education and Human Development*. Blackwell, Cambridge, MA pp. 9–27

Rogoff B, Chavajay P 1995 What's become of research on the cultural basis of cognitive development? *American Psychologist* **50**(10): 859–77

Siegler R S, Crowley K 1991 The microgenetic method: A direct means for studying cognitive development. *American Psychologist* **46**(6): 606–20

Vygotsky L S 1962 *Thought and Language*. MIT Press, Cambridge, MA

Weinert F, Schneider W (eds.) 1999 *Individual Development from 3 to 12: Findings from the Munich Longitudinal Study*. Cambridge University Press, New York

White S H 1992 G. Stanley Hall: From philosophy to developmental psychology. *Developmental Psychology* **28**(1): 25–34

M. Bullock

Cognitive Development in Childhood and Adolescence

In a typical textbook on cognitive development, one would find chapters on representation, memory, language, conceptual development, reasoning, problem solving, and strategy development and use. Currently, no single theory unites the study of all of these areas of cognition. Indeed, little communication exists among them. Not unlike the Indian parable in which six blind men give different answers to 'what is an elephant,' researchers in each area suggest that something different is 'what develops' in cognitive development in childhood and adolescence. In this article the domain of causal reasoning is used to describe general trends in cognitive development. After reviewing age-related developments in causal reasoning, processes that have been implicated as mechanisms of change in causal reasoning specifically and cognitive development more generally (including analogical reasoning, attention, working memory, selection and use of strategies, and domain-specific knowledge) are described. The article closes with a brief review of postnatal neural developments thought to underlie developmental changes in these processes.

1. Causal Reasoning

Whereas no one theory unites the numerous disparate domains of cognitive development, some cognitive operations are common across them. Causal reasoning provides an illustrative example of how several areas of cognition develop in concert to support a higher-level skill. Reasoning can be defined as goal-directed activity that often relies on the use of inferences to reach a conclusion (DeLoache et al. 1998). Causal relations are 'the cement of the universe' (Mackie 1980) in that they provide systematic links between precursors and consequences. Causal reasoning thus allows for identification of the cause–effect relations observed in the world. Inhelder and Piaget (1958, 1964), founders of scholarship in cognitive development, began the tradition of research on causal reasoning in the developing child and set the stage for much of the research done to date. Motivated by the desire to understand how children and adolescents think and reason about events in the world, researchers have examined infants' perceptions of causality, young children's ability to perform causally linked actions in sequential order so as to achieve a goal, and older children's abilities to infer causal mechanisms and to test for cause–effect relations (i.e., to reason scientifically).

For more detailed discussion of the above topics and for additional references see *Transfer of Learning, Cognitive Psychology of*; *Early Concept Learning in Children*; *Infant Development: Physical and Social Cognition*; *Piaget's Theory of Child Development*; *Scientific Concepts: Development in Children*; and *Problem Solving (Everyday), Psychology of*.

2. Early Developments in Causal Reasoning

2.1 Perception of Causality in Infancy

Within a few months of birth, infants demonstrate sensitivity to causes and their consequents. Oakes and Cohen (1990) showed 7- and 10-month-old infants different events in which one toy traveled across a screen and either (a) made contact with another toy, which immediately moved; (b) made contact with another toy, which moved only after a brief delay; or

(c) stopped before contacting the other toy, which, after a brief delay, moved. Only the first event contained the causal elements of spatial and temporal contiguity. The researchers found that the 10-month-olds were sensitive to the causal structure of the events as demonstrated by longer looking to the events that violated causal principles. Leslie (1984) found that infants as young as six and a half months are sensitive to causal structure when simpler stimuli (e.g., colored blocks) are used.

2.2 Enacting Causal Sequences to Achieve a Goal

By the second half of the first year of life infants not only appear to recognize basic conditions of causality, they also use them to guide their own actions to reach desired goals. In Willatts (1984), nine-month-old infants were presented with a cloth within their reach on which rested a desired toy just outside their reach. The infants needed to remove a barrier in order to pull the cloth and thereby obtain the toy. Compared to infants in a control condition in which the toy rested just off the cloth, infants in the 'causal' condition more often removed the barrier and pulled the cloth. By 12 months of age infants are able to solve a more complex version of this means–ends task: they successfully navigate multiple steps including removing a barrier to reach a cloth, pulling the cloth to reach a string, and reeling in a toy attached to the string. By 18 months, in age-appropriate versions of means–ends tasks, children use a variety of strategies and monitor the effectiveness of their strategies for reaching their goals (see Willatts 1990 for a review).

2.3 Planning a Path from Cause to Effect

Shortly after children show appreciation of existing causal connections, they begin to create their own paths from causes to effects. In Bauer et al. (1999), 21- and 27-month-old children were required to plan a course of action to achieve an effect. The children were shown either the initial state or the goal state of a problem that required a three-step solution. Each of the steps was necessary, but not sufficient, to achieve the goal. For example, for the problem of 'making a rattle' (Step 1: put a block in a cup; Step 2: cover the cup and block with a second cup; Step 3: shake the cups to make a rattling sound), an experimenter modeled either the initial step of putting the block in the cup or the goal step of shaking the rattle. In both cases, the experimenter verbally provided the children with the goal of the activity (i.e., 'make a rattle'). Note that the children were given the same amount of information (one step of the three-step solution) in each of the two conditions. When the children were shown the initial step of the causal sequence, neither the 21- nor the 27- month-olds solved the problem on more than 8 percent of the trials. In contrast, when provided with the goal step of the solution, 23 percent of the 21-month-olds and 50 percent of the 27-month-olds were able to plan a path to the goal. This study not only demonstrates the primacy of the goal state in aiding children's production of cause–effect sequences, but also shows rapid developments in planning abilities.

2.4 Identifying Causal Relations in the Preschool Years

Whereas in the toddler years, children show facility with planning a course of action to achieve a specific effect, in the preschool years, they reason as effectively about effects, causes, and the steps that unite them. Gelman et al. (1980) trained three- and four-year-old children to read three-picture 'stories' that depicted the initial state of an object, a causal agent, and the object in a transformed state. For example, children were shown an intact coffee cup (initial state), a hammer (causal agent), and a broken coffee cup (transformed state). The children then were presented with a series of stories in which one of the three pictures was missing. They were asked to complete the story by selecting one of three choice cards. The researchers found that both three- and four-year-olds were able to infer the causal agent as well as the initial and transformed states. In related research it has been shown that children understand a number and variety of causal relations, including melting, cutting, and burning (Goswami and Brown 1989).

Just as important as the ability to infer causes, consequences, and causal agents is the ability to recognize that not all cause–effect relations are deterministic. In the real world, causality tends to be probabilistic. For example, if an object is tremendously heavy, pulling the cloth on which it rests might not be sufficient to retrieve it. Kalish (1998) tested three- and five-year-olds and adults' understanding of probabilistic causal relations in the domain of illness. Whereas adults recognize that causes of illness are probabilistic (e.g., coming into contact with a sick person does not inevitably result in getting sick), children treat them as deterministic. This research suggests that the development of sensitivity to probabilistic causality is more protracted, relative to understanding of causal events with definite outcomes. Thus, lack of sensitivity to probabilistic outcomes represents a limitation of the causal understanding of preschool-age children.

3. Later Developments in Causal Reasoning

In addition to demonstrated competencies in reasoning about cause–effect relations, preschool-age children also show early manifestations of scientific reasoning. The goal of scientific reasoning is to test

hypotheses in order to identify cause–effect relations. Children as young as four and five years of age will search for the causal mechanism that resulted in an effect even when they do not see it (Bullock 1984). Therefore, even very young children are sensitive to the necessity of a cause. Nevertheless, just as they initially overgeneralize and treat all cause–effect relations as deterministic (i.e., failing to recognize the probabilistic nature of some causal relations), as shown below, in the preschool and early school years, children also overgeneralize one of the strongest cues to causality, namely, contiguity.

Schlottman (1999) presented five-, seven-, and nine-year-old children and adults with a mystery box that could contain one of two mechanisms. Both mechanisms caused a bell to ring when a ball was dropped in one of two holes in the box. One causal mechanism was 'slow' in ringing the bell because the ball had to travel down a runway in order to ring the bell at the opposite end of the box. The other causal mechanism was 'fast' because the ball dropped onto a lever that acted like a seesaw and immediately rang the bell. In the task, a ball was dropped into one of two holes in the box. Seconds later, the second ball was dropped and the bell immediately rang. Participants were asked to identify which of the two balls had caused the bell to ring. When they did not have knowledge of which of the two mechanisms was in the box, participants of all ages attributed causality to the contiguous event. That is, they selected the ball that was dropped immediately before the bell rang. Even after they were informed of which mechanism was in the box, five- and seven-year-olds continued to select the contiguous event, regardless of mechanism. In contrast, when appropriate, the nine-year-olds and adults ignored contiguity and made decisions based on the properties of the mechanism. This research makes clear both the profound influence that prior knowledge and experience have on causal and scientific reasoning and the great strides in treatment of data made by children in the early elementary school years.

Even beyond the early elementary school years, individuals' beliefs influence the ways in which they design experiments and the hypotheses that they test. Schauble (1996) asked fifth and sixth graders and noncollege adults to design experiments to identify the variables that affect the speed of a boat down a canal and the extension of a spring into water when a weight is attached. Both adults and children entered the experiment with beliefs about the objects and the relations about which they were to reason. In general, adults' beliefs were more appropriate than children's beliefs. For example, only 30 percent of the adults compared with 80 percent of the children expressed the belief that 'big things weigh more than small things.' *A priori* beliefs about causal variables influenced the experiments that both the children and the adults designed. Specifically, adults used their experimentation trials to understand the variables for which they did not hold prior beliefs. In contrast, children used their experiments to confirm the beliefs they held prior to investigation. This difference was partially responsible for the overall higher performance of adults in identifying the causal variables.

Schauble's (1996) research also makes clear that children and adults differ in the systematicity with which they approach the experimental space. First, although the children and the adults conducted the same total number of experiments, the children often inadvertently duplicated their experiments (even though they were provided with data cards to record their experimental manipulations). Second, within an experiment, the children were less systematic in the conduct of trials. They were less likely to control variables across two trials and often changed two variables at once in their attempts to test causal hypotheses. Because they based their causal inferences on confounded tests, children showed lower levels of performance relative to adults. These patterns indicate that in addition to domain-specific knowledge (as assessed by prior beliefs), domain-general experimentation strategies influence children's abilities to design experiments and to determine causal mechanisms.

Scientists are required not only to design experiments to test hypotheses, but are also required to draw conclusions on the basis of data obtained by others. In drawing conclusions, scientists must attend to a number of features, including the presence or absence of co-variation, the availability of a plausible causal mechanism, the size of the sample, and the sampling method used. Koslowski et al. (1989) presented sixth and ninth graders and college students with a series of story problems, each of which described the way evidence was gathered (direct intervention or correlation), the sample size (large or small), the causal mechanism (present or absent), and the results (co-variation or no co-variation). Participants were then asked to judge the extent to which they believed the proposed cause was responsible for an observed effect. Whereas all participants were sensitive to the co-variation of cause and effect, developmental changes were apparent in sensitivity to each of the other types of information. Sixth graders continued to give high ratings for the proposed cause when co-variation was present even when there was no causal mechanism provided and when a small sample size was used. Ninth graders provided high ratings for their confidence in the proposed cause with either a small or a large sample size, but only when a causal mechanism was present. College students showed even more refined scientific reasoning. They were less confident in the proposed cause when a causal mechanism was absent (even when co-variation was present) and when a small sample size was used. These findings suggest a developmental progression in scientific reasoning in which co-variation is primary, followed by a sensitivity to the presence of causal mechanism, with a later developing sensitivity to sample size. Notice that

absent from the list of features to which college students were sensitive is the nature of the evidence: Koslowski et al.'s participants did not discriminate direct intervention studies from correlational approaches. Thus although, with development, participants evidenced greater awareness of the types of information that scientists use in their evaluation of a potential cause and effect relation, the additional distinction of sampling method is needed to truly 'think like a scientist.'

4. *Mechanisms of Cognitive Change*

From infancy through adolescence, children's understanding of causality undergoes tremendous refinements. Developments in several cognitive domains support these advances in causal reasoning. The abilities include analogical reasoning, attention, working memory, selection and use of strategies, and domain-specific knowledge. Developments in each domain are described in turn.

Analogical reasoning refers to the process by which knowledge is transferred from one domain to another. It is appropriate when the domains share fundamental or 'deep structure' similarities. Analogical reasoning plays an important role in causal and scientific thinking because it allows for extension of knowledge from a well-known or better known situation to another less well-known domain. Children as young as 16 to 20 months of age show evidence of generalization of knowledge from one domain to another (Bauer and Dow 1994). Goswami and Brown (1989) demonstrated that preschoolers can use their understanding of causal relations to complete analogies wherein the higher order relation between the elements is the causal mechanism (e.g., chocolate is to melted chocolate as a snowman is to a melted snowman). Analogical reasoning also permits transfer of strategies from one task to another. For example, Chen and Klahr (1999) found that fourth graders transferred the strategy of control of variables even after a seven-month period. With development there are changes in the efficiency and proficiency with which analogical transfer is employed (see DeLoache et al. 1998 for a review).

Attention also shows tremendous developments from infancy to adolescence. There are documented developmental differences in the ability to sustain attention on a task, to switch attentional focus between tasks, and within a task, to focus on relevant versus irrelevant features (e.g., Ruff and Lawson 1990). In research on causal reasoning, the tasks presented to young children typically require focus on a limited number of features. The absence of irrelevant features makes it easier for young children to focus on those that are related to the causal structure. As tasks and problems become more complicated and more variables and features are involved, it is necessary to differentiate what is relevant from what is irrelevant, and to ignore 'distractions.' A comparison of Willatts' (1984) research on means–ends problem solving by nine-month-olds with Schlottman's (1999) research on the use of contiguity to identify causal mechanism by five to nine-year-olds makes clear that the older children are required to attend to many more features to produce or identify the cause–effect relation. In addition, with development, people become better able to attend selectively to the variables at hand. Selective attention may promote the control of variables when designing experiments and assist children in ignoring deceptive surface-feature similarities (Chen and Siegler 2000).

Age-related changes in working memory also play a large part in the development of causal reasoning abilities. Working memory refers to the ability to hold information in mind and simultaneously process it. A major model of working memory proposes that it is comprised of short-term information stores or 'buffers,' access to and use of which is controlled by a central executive function that also retrieves information from long-term memory and controls action planning and goal-directed behaviors (Baddeley 1986). Over the course of development, there are age-related changes in each of these functions. Whether developmental changes are the product of increases in available memory capacity (Halford et al. 1994) or of increasingly efficient use of available capacity (Case 1992) is under debate. Nevertheless, what is clear is that with development, there are increases in the length of time over which information can be held or maintained and the facility and speed with which it can be processed (see Gathercole 1998 for a review). For example, absent sustained, focused attention or rehearsal, auditory information remains persistent in short-term memory for 12 seconds in adults, 10 seconds in 10–12 year olds, and for only eight seconds in six–seven-year-olds (Keller and Cowan 1994). As reviewed in Swanson (1996), there also are age-related increases in working memory span, or the number of items that can be recalled in the context of a concurrent processing task. As demonstrated in Schauble's (1996) research, each of these elements plays a supportive role in causal reasoning. That is, compared with adults, children were more likely to lose sight of their plans for experimentation and were less systematic in their execution of plans. In each case, the presumed source of lower performance was difficulty remembering what already had been tested and what had been found.

Also implicated in developments in causal reasoning are changes in the selection and use of problem-solving strategies. For example, as demonstrated in Schauble (1996), relative to children, adults are more likely to use a control of variables strategy when testing hypotheses. Developmental changes in the selection and use of more and less sophisticated and effective strategies can be characterized as a series of overlapping 'waves' of approaches or solutions to

problems (Siegler 1996). Multiple strategies exist in the repertoire simultaneously. They compete with one another both over time and within a given problem resulting in waves of use as some strategies gradually decline and the frequency of others gradually increases. The implication is that strategies such as control of variables likely exist in the repertoires of both children and adults, alongside less advanced strategies of experimentation. Consistent with this suggestion, it has been shown that second graders can be trained to use a control of variables strategy (Chen and Klahr 1999). Nevertheless, the presence of the strategy in the child's repertoire does not guarantee that it will be used when appropriate or that it will be executed successfully when it is deployed.

Finally, analogical reasoning, attention, working memory, and selection and use of strategies all are influenced by the individual's familiarity with the domain. For example, Chi (1977) documented higher levels of reasoning by children in domains in which they had expertise (e.g., chess) than in domains in which they were relative novices. Goswami and Brown (1989) found that analogical reasoning was facilitated by knowledge of the causal mechanisms involved in the transformations that children were asked to judge. Similarly, Schauble (1996) found that both children's and adults' beliefs about causality influenced the hypotheses that they tested. Thus, the influence of domain knowledge is consistent throughout the life span. The facilitative effects of knowledge also extend to the component processes that support problem solving. Specifically, working memory span is greater for familiar words than for nonwords (Hulme et al. 1991). In addition, greater experience within a domain allows participants to guide their attention to the relevant attributes of the problem to be solved, and to disregard those features that are irrelevant to the solution (Goswami and Brown 1989).

5. *Neurological Changes Related to Cognitive Development*

Although the precise linkages are not clear, it is widely assumed that cognitive developmental changes in processes such as attention and working memory are related to developments in the neural substrates that subserve them. The prefrontal cortex supports planning and working memory and undergoes a protracted course of development (see Nelson et al. 2000 for review). For example, although adult levels of metabolic activity are approximated in the first year of life, development continues into adolescence (Chugani 1994), and frontal pruning of synapses (Huttenlocher 1994) and myelination (Jernigan et al. 1991) continue into adolescence. Thatcher (1992) demonstrated that the coherence of EEG patterns between frontal and posterior lobes increases throughout middle childhood. Case (1992) has shown that the patterns of coherence in EEG mirror the patterns in the development of attention over the same time period (see Case 1998, for a review).

Other neurological developments include brainwide changes such as myelination that results in faster transmission of impulses between neurons. There also are age-related increases in white matter density that are consistent with greater myelination of fiber tracts throughout adolescence (Paus et al. 1999). Increased speed of processing benefits working memory and attention.

6. *Summary*

The refinements that occur through brain development produce changes in the cognitive processes that support causal understanding. With development, children accrue a greater knowledge base, process information more rapidly, attain greater attentional and strategic resources, and improve in the efficiency with which they utilize resources. These component processes support a range of behaviors that are united by the appreciation of causal mechanisms, from behaviors as basic as infants' perception of causality to adolescents' abilities to design and evaluate experiments to test causal relations. Although we observe qualitative differences in these behaviors through the course of development, what underlies these advances are continuous changes in the brain and the resulting cognitive component processes that develop to support them.

Bibliography

Baddeley A D 1986 *Working Memory*. Oxford University Press, Oxford, UK

Bauer P J, Dow G A A 1994 Episodic memory in 16- and 20-month-old children: Specifics are generalized, but not forgotten. *Developmental Psychology* **30**: 403–17

Bauer P J, Schwade J A, Wewerka S S, Delaney K 1999 Planning ahead: Goal-directed problem solving by 2-year-olds. *Developmental Psychology* **35**: 1321–37

Bullock M 1984 Preschool children's understanding of causal connections. *British Journal of Developmental Psychology* **2**: 139–48

Case R 1992 The role of the frontal lobes in the regulation of cognitive development. *Brain and Cognition* **20**: 51–73

Case R 1998 The development of conceptual structures. In: Kuhn D, Siegler R S, Damon W (eds.) *Handbook of Child Psychology*, 5th edn. John Wiley, New York, Vol. 2, pp. 745–800

Chen Z, Klahr D 1999 All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development* **70**: 1098–120

Chen Z, Siegler R S 2000 Across the great divide: Bridging the gap between understanding of toddlers' and older children's thinking. *Monographs of the Society for Research in Child Development*, Vol. 65.

Chi M T H 1977 Age differences in memory span. *Journal of Experimental Child Psychology* **23**: 266–81

Chugani H T 1994 Development of regional brain glucose metabolism in relation to behavior and plasticity. In: Dawson G, Fischer K (eds.) *Human Behavior and the Developing Brain*. Guilford Press, New York, pp. 153–75
DeLoache J S, Miller K F, Pierroutsakos S L 1998 Reasoning and problem solving. In: Kuhn D, Siegler R S, Damon W (eds.) *Handbook of Child Psychology*, 5th edn. Wiley, New York, Vol. 2, pp. 801–50
Gathercole S E 1998 The development of memory. *Journal of Child Psychology and Psychiatry* **39**: 3–27
Gelman R, Bullock M, Meck E 1980 Preschoolers' understanding of simple object transformations. *Child Development* **51**: 691–9
Goswami U, Brown A L 1989 Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition* **35**: 69–95
Halford G S, Maybery M T, O'Hare A W, Grant P 1994 The development of memory and processing capacity. *Child Development* **65**: 1338–56
Hulme C, Maughan S, Brown G D A 1991 Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language* **30**: 685–701
Huttenlocher P R 1994 Synaptogenesis, synapse elimination, and neural plasticity in human cerebral cortex. In: Nelson CA (ed.) *Minnesota Symposium on Child Psychology: Vol. 27—Threats to Optimal Development: Integrating Biological, Psychological, and Social Risk Factors*. Erlbaum, Hillsdale, NJ, pp. 35–54
Inhelder B, Piaget J 1958 *The growth of Logical Thinking From Childhood to Adolescence*. Basic Books, New York
Inhelder B, Piaget J 1964 *The Early Growth of Logic in the Child*. Harper & Row, New York
Jernigan T L, Trauner D A, Hesselink J R, Tallal P A 1991 Maturation of human cerebrum observed in vivo during adolescence. *Brain* **114**: 2037–49
Kalish C W 1998 Young children's predictions of illness: Failure to recognize probabilistic causation. *Developmental Psychology* **34**: 1046–58
Keller T A, Cowan N 1994 Developmental increase in the duration of memory for tone pitch. *Developmental Psychology* **30**: 855–63
Koslowski B, Okagaki L, Lorenz C, Umbach D 1989 When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development* **60**: 1316–27
Leslie A M 1984 Spatiotemporal continuity and the perception of causality in infants. *Perception* **13**: 287–305
Mackie J L 1980 *The Cement of the Universe: A Study of Causation*. Clarendon Press, Oxford, UK
Nelson C A, Monk C S, Lin J, Carver J L, Thomas K M, Truwit C L 2000 Functional neuroanatomy of spatial working memory in children. *Developmental Psychology* **36**: 109–16
Oakes L M, Cohen L B 1990 Infant perception of a causal event. *Cognitive Development* **5**: 193–207
Paus T, Zidgenbos A, Worsley K, Collins D L, Blumenthal J, Giedd J N, Rapoport J L, Evans A C 1999 Structural maturation of neural pathways in children and adolescents: in vivo study. *Science* **283**: 1908–11
Ruff H A, Lawson K R 1990 Development of sustained, focused attention in young children during free play. *Developmental Psychology* **26**: 85–93
Schauble L 1996 The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology* **32**: 102–19
Schlottman A 1999 Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology* **35**: 303–17
Siegler R S 1996 *Emerging Minds: The Process of Change in Children's Thinking*. Oxford University Press, New York
Swanson H L 1996 Individual and age-related differences in children's working memory. *Memory and Cognition* **24**: 70–82
Thatcher R W 1992 Cyclic cortical reorganization during early childhood. *Brain and Cognition* **20**: 24–50
Willatts P 1984 The Stage-IV infants' solution of problems requiring the use of supports. *Infant Behavior and Development* **7**: 125–34
Willatts P 1990 Development of problem-solving strategies in infancy. In: Bjorklund D F (ed.) *Children and Strategies: Contemporary Views of Cognitive Development*. Erlbaum, Hillsdale, NJ, pp. 23–66

P. J. Bauer and M. M. Burch

Cognitive Development in Infancy: Neural Mechanisms

1. Background

Until the past decade the study of cognitive development in human infants has been conducted relatively independently of any consideration of the brain. This relative neglect of biological factors in the study of behavioral development is surprising since the origins of developmental psychology can be traced to biologists. Darwin was one of the first to take a scientific approach to human behavioral development, and to speculate on the relations between phylogenetic and ontogenetic change. Piaget, who was originally trained as a biologist, used then-current theories of embryological development to generate his accounts of human cognitive development. McGraw and Gesell tried to integrate brain development with what was known of behavioral development. While they focused on motor development, they also extended their conclusions to mental and social development (Gesell 1928, McGraw 1943). While both these authors developed sophisticated informal theories that attempted to capture non-linear and dynamic approaches to development, their efforts to relate brain development to behavioral change remained very speculative due to the paucity of knowledge at the time.

From the 1960s to the late 1980s biological approaches to human behavioral development were neglected for a variety of reasons, including the widely held belief among cognitive psychologists during that period that the 'software' of the mind is best studied without reference to the 'hardware' of the brain. However, the recent expansion of knowledge at the end of the twentieth century on brain development

makes the task of relating it to behavioral changes considerably more viable than previously. In addition, new molecular and cellular methods, along with theories based on artificial neural networks, have led to great advances in our understanding of how primate brains are constructed during ontogeny. These advances, along with those in functional neuroimaging, have led to the recent emergence of the interdisciplinary science of developmental cognitive neuroscience (see Johnson 1997).

What benefits can accrue from taking a developmental cognitive neuroscience approach to infants? First, considering evidence from brain development may help constrain, or even change, the type of cognitive theories that we consider. Second, being able to relate brain to cognitive development will potentially allow a more complete explanation not only of normal development, but also of developmental disorders resulting from genetic abnormality, and the long-term effects of early brain damage.

2. *Human Postnatal Brain Development*

A number of lines of evidence indicate that there are substantive changes during postnatal development of the human brain. Perhaps most obviously, the volume of the brain quadruples between birth and adulthood. This increase comes from a number of sources such as more extensive fiber bundles, and nerve fibers becoming myelinated. In addition, there is a dramatic increase in size and complexity of the dendritic tree of many neurons. Less apparent with standard microscopy, but evident with electron microscopy, is a corresponding increase in the density of synapses.

Huttenlocher (1990) and colleagues have reported a steady increase in the density of synapses in several regions of the human cerebral cortex. For example, in parts of the visual cortex, the generation of synapses (synaptogenesis) begins around the time of birth and reaches a peak at around 150 percent of adult levels toward the end of the first year. In the frontal cortex (the anterior portion of cortex, considered by most investigators to be critical for many higher cognitive abilities), the peak of synaptic density occurs later, at around 24 months of age (see Goldman-Rakic et al. 1997). Although there is variation in the timetable, in all regions of cortex studied so far, synaptogenesis begins around the time of birth and increases to a peak level well above that observed in adults.

Somewhat surprisingly, regressive events are commonly observed during the development of nerve cells and their connections in the brain. Due to the paucity of human data, the regional timetable for this decrease is still unclear and there is controversy about whether or not it shows differences between regions. Nevertheless, in humans, most neocortical regions and pathways appear to undergo this 'rise and fall' in synaptic density, with the density stabilizing to adult levels at different ages during later childhood. The postnatal rise-and-fall developmental sequence can also be seen in other measures of brain physiology and anatomy. For example, using PET, Chugani et al. (1987) observed an adult-like distribution of resting brain activity within and across brain regions by the end of the first year. However, the overall level of glucose uptake reaches a peak during early childhood, which is much higher than that observed in adults. The rates return to adult levels after about 9 years of age for some cortical regions. The extent to which these changes relate to those in synaptic density is currently the topic of further investigation.

A controversial issue in developmental neuroscience concerns the extent to which the differentiation of the cerebral neocortex into areas or regions with particular cognitive, perceptual, or motor functions can be shaped by postnatal interactions with the external world. This issue reflects the debate in cognitive development about whether infants are born with domain-specific 'modules' for particular cognitive functions such as language, or whether the formation of such modules is an activity-dependent process (see Elman et al. 1996, Karmiloff-Smith 1992). Since around 1900 neuropsychology has taught us that the majority of normal adults tend to have similar functions within approximately the same regions of cortex. However, we cannot necessarily infer from this that this pattern of differentiation is intrinsically prespecified (the product of genetic and molecular interactions), because most humans share very similar pre- and postnatal environments. In developmental neurobiology this issue has emerged as a debate about the relative importance of neural activity for cortical differentiation, as opposed to intrinsic molecular and genetic specification of cortical areas. Supporting the importance of the latter processes, Rakic (1988) proposed that the differentiation of the cortex into areas is due to a protomap. The hypothesized protomap either involves prespecification of the tissue that gives rise to the cortex during prenatal life or the presence of intrinsic molecular markers specific to particular areas of cortex. An alternative viewpoint, advanced by O'Leary (1989) among others, is that genetic and molecular factors build an initially undifferentiated 'protocortex,' and that this is then subsequently divided into specialized areas as a result of neural activity. This activity within neural circuits need not necessarily be the result of input from the external world, but may result from intrinsic, spontaneous patterns of firing within sensory organs or subcortical structures that feed into the cortex, or from activity within the cortex itself (e.g., Katz and Shatz 1996).

Although the neurobiological evidence is complex, and probably differs between species and regions of the cortex, overall it tends to support the importance of neural activity-dependent processes (see Johnson 1997 for a review). With several notable exceptions, it seems

likely that activity-dependent processes contribute to the differentiation of functional areas of the cortex, especially those involved in higher cognitive functions in humans. During prenatal life, this neural activity may be largely a spontaneous intrinsic process, while in postnatal life it is likely also to be influenced by sensory and motor experience.

3. *Methods for Studying Human Postnatal Brain Development*

Part of the reason for the recently renewed interest in relating brain development to cognitive change comes from advances in methodology which allow hypotheses to be generated and tested more readily than previously (see also Nelson and Bloom 1997). One set of tools relates to brain imaging. Some of these imaging methods, such as positron emission tomography (PET), are of limited utility for studying transitions in behavioral development in normal infants and children due to their invasive nature and their relatively coarse temporal resolution. However, two other methods may prove more useful.

Since the 1960s, scalp recorded event-related potentials have been used to assess brain function in infants and children for several decades. These recordings can either be of the spontaneous natural rhythms of the brain (EEG), or the electrical activity evoked by the presentation of a stimulus (ERP). Recent developments of the ERP method allow the relatively quick and easy installation of large numbers of sensors, thus making the method easier to use and also improving spatial resolution. Functional MRI allows the noninvasive measurement of cerebral blood flow with fine spatial resolution and temporal resolution on the order of seconds. Although this technique has been applied to children (Casey et al. 1997), the distracting noise and vibration, and the presently unknown possible effects of high magnetic fields on the developing brain, make its usefulness for healthy children under 4 or 5 years of age unclear. However, there has been at least one MRI study of infants initially scanned for clinical reasons (Tzourio et al. 1992), and the advent of 'open' scanners in which the mother can hold the infant may increase possibilities further.

Apart from brain imaging, the neural basis of cognitive development in infants can be examined by administering behavioral 'marker tasks' to infants who have suffered perinatal brain damage or developmental disorders of genetic origin. These marker tasks are adapted from tasks previously linked to a brain region or pathway in adult primates and humans by cognitive neuroscience studies. By testing infants or children with versions of such a task at different ages, the researcher can use the success or otherwise of individuals as indicating the functional development of the relevant regions of the brain. Finally, there is a continuing need for the neuroanatomical study of postmortem tissue. For a variety of reasons such studies are difficult to conduct.

4. *Relating Brain to Cognitive Development*

A number of different approaches have been taken to relating brain to cognitive development. These different approaches depend on very different sets of assumptions about development.

4.1 *Maturational Models*

The most common approach to developmental cognitive neuroscience is based on a maturational framework, in which it is assumed that as particular brain regions mature they allow or enable new cognitive functions to come on line. By this view postnatal brain development is assumed to be heavily governed by genetic and molecular factors, and relatively (though not completely) independent of experience. In brief, postnatal brain development is seen as a necessary, but not sufficient, cause of change in cognitive abilities. Two areas in which this approach has been applied concern the transition from subcortical to cortical control over visually guided behavior, and the later onset of frontal and prefrontal cortex control.

In one of the first specific attempts to relate changes in behavior to brain development in infants, Bronson (1974) presented evidence that the subcortical retinocollicular visual pathway primarily controls visually guided action in the newborn human infant. He also showed that it is only by around 3 months of age that visually-guided behavior switches to cortical pathways. More recent research indicates that there is probably some, albeit limited, cortical activity in newborns, and that the onset of cortical control over behavior is a gradual, rather than all-or-none, transition. Johnson (1990) updated Bronson's thesis to incorporate several different cortical pathways now known to underlie visually guided action in adult primates. The logic underlying this model was that changes in visually guided behavior of infants over the first months of life could be attributed to the graded onset of each of several different cortical pathways. Further, which pathways were active could be predicted from the developmental neuroanatomy of the primary visual cortex at that age, since this structure was the gateway to most of these pathways. While this model had reasonable success in accounting for the sequence changes in behavior observed, in the past few years studies involving ERPs, and studies of infants with focal cortical damage, show that frontal cortical regions are active earlier than more posterior regions, a sequence not predicted by the original Johnson (1990) model.

Another prominent maturational model has concerned the onset of prefrontal cortex functioning. In

terms of structural neuroanatomy, this part of the cortex shows the most prolonged development of any region of the human brain, with changes in synaptic density detectable even into the teenage years (Huttenlocher 1990). Diamond (1991) has argued that the maturation of prefrontal cortex during the period 6–12 months accounts for a number of transitions observed in the behavior of infants in object permanence and object retrieval tasks. In one such task infants younger than 8 months often fail to accurately retrieve a hidden object after a short delay period if the object's location is changed from one where it was previously successfully retrieved. The basis for Diamond's claims come from the observations that (a) monkeys with lesions to the dorsolateral prefrontal cortex (DLPC) show the same patterns of impairment as young human and monkey infants, and (b) there are neurochemical and neuranatomical changes in the human DLPC at around the age they begin to perform successfully. Diamond (1991) has speculated that the DLPC is critical for performance when (a) information has to be retained or related over time or space, and (b) a prepotent response has to be inhibited. She argues that prior to the maturation of the DLPC, infants do not successfully perform tasks that require both of these abilities.

Further evidence linking success in the object permanence task to frontal cortex maturation in the human infant comes from two sources. The first of these is a series of EEG studies with normal human infants (e.g., Bell and Fox 1992), in which increases in frontal EEG responses correlate with the ability to respond successfully over longer delays in delayed response tasks. The second source is work on cognitive deficits in children with a neurochemical deficit in the prefrontal cortex resulting from Phenylketonuria (PKU). Even when treated, this inborn error of metabolism can have the specific consequence of reducing the levels of a neurotransmitter, dopamine, in the dorsolateral prefrontal cortex. These reductions in dopamine levels in the dorsolateral prefrontal cortex, result in these infants and children being impaired on tasks thought to involve parts of the prefrontal cortex, such as the object permanence task and an object retrieval task, and being relatively normal in tasks thought to depend on other regions of the cortex (Diamond et al. 1997, Welsh et al. 1990).

4.2 Selectionist Models

As discussed earlier, during the postnatal development of the cortex there is a rise and fall in synaptic density. This observation has led to a number of 'selectionist' theories being advanced in which the essential notion is that there is experience related sculpting of neural connectivity. For example, Changuex (1985) proposes that molecular and genetic processes specify the initial overproduction of synaptic contacts. These initial connections are labile, but either stabilize or regress, depending on the activity in the postsynaptic cell, a process referred to as 'selective stabilization.' Changeux and Dehaene (1989) suggested that this model could be used to bridge from the brain to cognitive and behavioral levels, and that the same process of 'Darwinian' change could occur. Perhaps the best example of this type of change at the behavioral level comes from the work on phonemic discrimination in infants, showing that while they can initially discriminate a very large range of phonetic boundaries used in speech, including those not found in their native language, this ability becomes restricted to those boundaries important for their native language around 12 months of age.

Selectionist models have recently been criticized for the assumption that the initial stage of overproduction is not sensitive to experience (Quartz and Sejnowski 1997), and for focusing too heavily on only one aspect of neural development (Purves 1994). It is also important to remember that neuroanatomical measures of synaptic density are a static measure of dynamic processes. Since there is constant turnover of synapses, it is unlikely that there are clearly distinct phases of growth and pruning. Rather, both stages are likely to be simultaneously occurring within cortical regions.

4.3 Activity-dependent Models

A number of factors suggest that the field needs to move beyond the maturational framework. First, increasing evidence from developmental neuroscience suggests that neuronal activity itself plays a vital role in prenatal brain development, and it would seem reasonable to suggest that the same processes may extend into postnatal life. Second, there is evidence from neuroimaging and the study of infants with focal brain damage to suggest that there are dynamic changes in the timing and pattern of cortical activation in infants relative to adults. These dynamic changes take a number of forms, including changes in the overall spatial extent in cortex-activated ones (localization), changes in the extent to which the activation of a cortical region is stimulus-specific (specialization), and changes in the temporal stage of cortical processing at which specialization can be observed (see Johnson 2000).

Event-related potential experiments with infants have indicated that both for word learning (Neville 1991) and face processing (de Haan et al. submitted) there is increasing localization of processing with age/experience of a stimulus class. That is, more widespread scalp leads show a difference between the words or faces that control stimuli in younger infants than in older ones. In the example of face processing, both the left and the right ventral visual pathways are differentially activated by faces in early infancy, but in

many (but not all) adults this further localizes only to the right ventral pathway (de Haan et al. 1998). In the example of word recognition, differences are initially found over widespread cortical areas, but narrow to left temporal leads after further experience with this class of stimulus (Neville 1991). Johnson (2000) presented an 'interactive specialization' framework within which changes in localization are a direct consequence of increases in specialization within and between cortical pathways. He suggests that this framework also provides a way of thinking about the fact that the same behavior can be mediated by different patterns of cortical activation in infants from those observed in adults.

Activity-dependent models can also be extended more broadly to the view that, at a given stage in postnatal development, the human infant may actually seek out the sensory input it needs to enable the further specialization of it its own brain. In other words, the infant is not a passive absorber of experience, but rather an active and selective seeker of it. Thus, the infant changes its 'effective environment' during development. One example of this comes from the development of face processing, where it has been argued that primitive tendencies for the newborn to orient to face-like stimuli ensures that developing cortical circuitry is preferentially exposed to that class of stimulus (see Johnson 1997).

See also: Brain Development, Ontogenetic Neurobiology of; Functional Brain Imaging; Infant Development: Physical and Social Cognition; Neural Plasticity; Prefrontal Cortex; Prefrontal Cortex Development and Development of Cognitive Function; Prenatal and Infant Development: Overview; Sensitive Periods in Development, Neural Basis of

Bibliography

Bell M A, Fox N A 1992 The relations between frontal brain electrical activity and cognitive development during infancy. *Child Development* **63**: 1142–63

Bronson G 1974 The postnatal growth of visual capacity. *Child Development* **45**: 873–90

Casey B J, Trainor R J, Orendi J L, Schubert A B, Nystrom L E, Giedd J N, Xavier Castellanos J L et al. 1997 A developmental functional MRI study of prefrontal activation during performance of a go-no-go task. *Journal of Cognitive Neuroscience* **9**: 835–47

Changeux J P 1985 *Neuronal Man: The Biology of Mind*. Pantheon Books, New York

Changeux J P, Dehaene S 1989 Neuronal models of cognitive functions. *Cognition* **33**: 63–109

Chugani H T, Phelps M E, Mazziotta J C 1987 Positron emission tomography study of human brain functional development. *Annals of Neurology* **22**: 487–97

de Haan M, Oliver A, Johnson M H 1998 Electrophysiological correlates of face processing by adults and 6-month-old infants. *Journal of Cognitive Neuroscience* **36** Supp. S

de Haan M, Pascauis O, Johnson M H (submitted) Spatial and temporal characteristics of cortical activation in adults and infants viewing faces

Diamond A 1991 Neuropsychological insights into the meaning of object concept development. In: Carey S, Gelman R (eds.) *The Epigenesis of Mind: Essays on Biology and Cognition*. Erlbaum, Hillsdale, NJ, pp. 67–110

Diamond A, Hurwitz W, Lee E Y, Bockes T, Grover W, Minarcik C 1997 Cognitive deficits on frontal cortex tasks in children with early-treated PKU: Results of two years of longitudinal study. *Monographs of the Society for Research in Child development, Monographs* No. **252**: 1–207

Elman J L, Bates E, Johnson M H, Karmiloff-Smith A, Parisi D, Plunkett K 1996 *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press, Cambridge, MA

Gesell A L 1928 *Infancy and Human Growth*. Macmillan, New York

Goldman-Rakic P S, Bourgeois J, Rakic P 1997 Synaptic substrate of cognitive development: Life-span analysis of synaptogenesis in the prefrontal cortex of the nonhuman primate. In: Krasnegor N A, Reid Lyon G, Goldman-Rakic P S (eds.) *Development of the Prefrontal Cortex. Evolution, Neurobiology and Behavior*. Paul H Brookes, Baltimore, MD, pp. 27–48

Huttenlocher P R 1990 Morphometric study of human cerebral cortex development. *Neuropsychologia* **28**: 517–27

Johnson M H 1990 Cortical maturation and the development of visual attention in early infancy. *Journal of Cognitive Neuroscience* **2**: 81–95

Johnson M H 1997 *Developmental Cognitive Neuroscience: An Introduction*. Blackwell, Oxford, UK

Johnson M H, de Haan M 2001 Developing cortical specialization for visual-cognitive function: The case of face recognition. In: McClelland J L, Siegler R S (eds.) *Mechanisms of Cognitive Development: Behavioral & Neural Perspectives*. Lawrence Erlbaum Associates, Mahwah, NJ

Johnson M H 2000 Functional brain development in infants: elements of an interactive specialization framework. *Child Development* **71**(1): 75–81

Karmiloff-Smith A 1992 *Beyond Modularity: A developmental Perspective on Cognitive Science*. MIT Press/Bradford Books, Cambridge, MA

Katz L C, Shatz C J 1996 Synaptic activity and the construction of cortical circuits. *Science* **274**: 1133–8

McGraw M B 1943 *The Neuromuscular Maturation of the Human Infant*. Columbia University Press, New York

Nelson C A, Bloom F E 1997 Child Development and Neuroscience. *Child Development* **68**: 970–87

Neville H J 1991 Neurobiology of cognitive and language processing: Effects of early experience. In: Gibson K R, Petersen A C (eds.) *Brain Maturation and Cognitive Development: Comparative and cross-cultural Perspectives*. Adaline de Gruyter Press, Hawthorne, NY, pp. 355–80

O'Leary D D 1989 Do cortical areas emerge from a protocortex? *Trends in Neuroscience* **12**: 400–6

Purves D 1994 *Neural Activity and the Growth of the Brain*. Academia Nazionale Dei Lincei, Cambridge University Press, Cambridge, UK

Quartz S R, Sejnowski T J 1997 A neural basis of cognitive development: A constructivist manifesto. *Behavioural and Brain Sciences* **20**: 537–96

Rakic P 1988 Specification of cerebral cortical areas. *Science* **241**: 170–6

Tzourio N, de Schonen S, Mazoyer B, Bore A, Pietrzyk U, Bruck B, Aujard Y, Deruelle C 1992 Regional cerebral blood flow in

two-month old alert infants. *Society for Neuroscience Abstracts* **18**: 1121
Welsh M C, Pennington B F, Ozonoff S, Rouse B, McCabe E R 1990 Neuropsychology of early-treated phenylketonuria: Specific executive function deficits. *Child Development* **61**: 1697–713

M. H. Johnson

Cognitive Development: Learning and Instruction

Scientific conceptualizations of the relations between cognitive development and learning on the one side, and learning and instruction on the other side, have always been controversial in psychology and continue to be so. Theoretical standpoints depend on whether they take a universal or differential perspective, whether they are dominated by evolutionary-genetic or environment-oriented approaches, and whether they are biased ideologically toward an optimistic or pessimistic view on education. For these reasons, this entry will (a) present the universal relations between cognitive development, learning, and instruction; (b) analyze differential aspects of these three concepts; (c) describe theoretical approaches to the relation between cognitive development and learning; and (d) present theoretical conceptualizations of the relation between learning and instruction. The final section will contain some conclusions for educational practice.

1. Cognitive Development, Learning, and Instruction from a Universal Perspective

Inspired by evolutionary ideas in general, and the epistemological approach of Jean Piaget (1970) in particular, most classic theories on childhood cognitive development can be characterized by four basic principles. Their theoretical models are:

(a) universal (valid for all human beings),

(b) general (valid for all cognitive phenomena),

(c) structural (valid for all basic changes in the cognitive systems and functions that internally determine the acquisition of knowledge and skills),

(d) naturalistic-descriptive (stating that developmental changes are caused by the species-specific nature of human beings and may be influenced by environmental variables but not produced by them).

A typical example is Flavell's (1970) theoretical assumption 'that cognitive changes during childhood have a specific set of formal "morphogenetic" properties, that presumably stem from the biological-maturational growth process underlying these changes. Thus, childhood cognitive modifications are largely inevitable, momentous, directional, uniform, and irreversible' (p. 247).

In such a general conception, learning processes and their external conditions take a subordinate role. Instructions may not only promote natural development but also disrupt or impair it, and although cognitive development is also a condition and outcome of education, it is, nonetheless, a particularly important goal when the interest is in maintaining the dynamics and continuity of the developmental process for as long as possible in order to reach the highest possible level of cognitive development (Kohlberg and Mayer 1972).

At the beginning of the twenty-first century, many scientists are skeptical about such purist theories of cognitive development. This applies not only to the role of learning and of external learning opportunities but also to the impact of instruction and schools on cognitive development. For example, Geary (1995) discriminates between primary and secondary biological abilities. He views primary abilities as innate mental dispositions that enable even infants to learn from experience under almost any sociocultural conditions through the assistance of a strong intrinsic motivation. This is how they acquire domain-specific competencies (e.g., a mother tongue; elementary numerical skills; physical, biological, psychological, and social knowledge).

In comparison, secondary skills vary greatly between cultures, subcultural groups, and cohorts. The learning processes needed to acquire secondary competencies (scientific knowledge, higher mathematical skills, metacognitive skills) are cumulative, proceed at a relatively slow pace, require individual effort, and generally call for extrinsic motivation and didactic support. The level of secondary abilities that may be achieved collectively, and is actually achieved individually, depends on the state of cultural development and the availability of schools or equivalent institutions.

Hence, Geary's (1995) model assumes that early cognitive development depends on a large number and variety of necessary, autonomous, and automatically effective learning processes, and that although instructions (models, feedback, reinforcement, indications, availability of learning opportunities, etc.) take an important role, they are not decisive, because the acquisition of primary abilities is universal, genetically predetermined, and intrinsically motivated. In contrast, cognitive development in middle childhood, adolescence, and adulthood is influenced far more strongly by a person's cultural, familial, and individual situation. Instructionally guided and promoted learning is decisive for the course, the contents, and the level of cognitive development. However, there are large interindividual differences—even under similar cultural conditions—in the speed, efficacy, and quality of learning. Such individual differences in cognitive learning and, thus, in cognitive development, cannot be explained completely as consequences of prior learning.

2. Cognitive Development, Learning, and Instruction from a Differential Perspective

Each attained state of cognitive development determines learning, and the outcomes of cumulative learning processes influence the course of cognitive development. This general rule is moderated by the strength and stability of individual differences in intellectual abilities.

Ever since Binet and Simon constructed the first intelligence test in 1905, psychometrically oriented research has assumed that interindividual differences in abilities are relatively stable from middle childhood onward, and that these differences in the speed and quality of cognitive development permit long-term predictions. In a theoretically sophisticated but skeptical review of the available empirical findings, Wohlwill (1980) concluded, 'A reasonable coherent picture of the stability of IQ, over different portions of the age span, and some of the variables affecting it, has emerged from … research' (p. 401). How can we explain this high stability in individual IQ differences? The most plausible answer is to assume the existence of cognitive and social mechanisms with the same effect as the Matthew Principle of 'For unto everyone that hath shall be given.'

Despite numerous theoretical and methodological controversies, the available research on twins leads to the conclusion that one half or slightly more of the variance in IQ in the industrialized nations is determined genetically (which naturally also means that almost one half of this variance is not determined genetically). In contrast to much prejudiced belief, genetic factors and environmental conditions do not impact independently on the individual's cognitive development, but covary to a major degree. As a rule, biological parents are also the most important actors in the young child's social environment. Persons already react differentially in accordance with the infant's genotype, and children become actively involved in selecting preferential segments of the environment at an early age. These covariations help to strengthen and stabilize interindividual differences. Moreover, more intelligent children profit to a larger extent from the same learning opportunities and instructional aids than their less intelligent peers. Those who are more successful as a result of a cumulation of favorable factors develop a higher degree of self-confidence that encourages them to make a greater effort to solve difficult problems. Finally, young persons with greater abilities and higher achievements generally get a better education, which leads to greater career opportunities that are also associated in the long term with more cognitive stimulation and challenge.

Scientists and educators differ greatly in their response to the empirical evidence of stable differences in cognitive development. In the 1960s and 1970s, people thought that an egalitarian leveling out of individual differences in cognitive abilities and achievements was possible and could be achieved through intervention. The concept of mastery learning was viewed as the key for this. The idea was to grant less gifted and unsuccessful students a temporary increase in necessary learning time, while simultaneously adapting instruction to match their learning aptitudes as closely as possible. The educational and, to some extent, ideological aspirations invested in this model of learning have never paid off. Students with low abilities and poor prior knowledge differ greatly in what they can learn during a similar amount of learning time. As a result, large differences in the amount of learning time are required if equal achievements are to be attained in a heterogeneous population (Slavin 1987).

The opposing standpoint is well-characterized by Herrnstein and Murray's (1994) book *The Bell Curve*. According to their research in the USA, the interplay of differences in genetic endowment and sociocultural conditions lead, at a very early stage, to stable interindividual differences in intellectual abilities, motivational tendencies, and patterns of social behavior. The attempts to increase equality through compensatory education in recent decades have led to no long-lasting reduction in intellectual differences. As a result, the authors, strongly recommend discontinuing interventions for disadvantaged children and investing the available financial resources in educating those students whose high mental potential means that they will be responsible for generating the largest part of the gross national product in the years to come.

Both radical positions give a very one-sided interpretation of the current state of research. They overlook the fact that all children have to learn all competencies, and that all students profit from good instruction, while, at the same time, major differences in cognitive abilities and performance cannot be leveled out through intensive developmental interventions and academic instruction.

3. Cognitive Development and Learning: Which Comes First?

Psychology has a long tradition of assuming that learning is the only mechanism for explaining cognitive development. With the expert approach, this assumption has even survived the shift from behaviorist theories to cognitive models. Indeed, the expert approach has even be applied to cognitive development (Carey 1984, Sternberg 1998). For example, Carey (1984) has postulated: 'Children differ from adults only in accumulation of knowledge' (p. 37). As a result: 'Children know less than adults. Children are novices in almost every domain in which adults are experts' (p. 64).

Gagné (1968) had adopted a similar position in this criticism of the structuralist stage theories of cognitive development (e.g., Piaget 1970):

In an oversimplified way, it may be said that the stage of intellectual development depends upon what the learner knows already and how much he has yet to learn order to achieve some particular goals. Stages of development are not related to age, except in the sense that learning takes time. They are not related to logical structures, except in the sense that the combining of prior capabilities into new ones caries its own inherent logic Gagné 1968 (p. 189).

Does learning really determine all of cognitive development during childhood? There is still not enough empirical evidence to answer this question. Nonetheless, the confirmation of domain-specific knowledge in infancy, of the acquisition of demanding (linguistic or numerical) competencies during early childhood without any recognizable signs of a prior acquisition of the necessary cognitive preconditions, and the interindividual similarly in cognitive development despite different learning opportunities, supports the idea that innate predispositions must also be involved in not only specific but also unspecific cognitive learning processes.

As children grow older, the externally controlled acquisition (through learning) of competencies seems to become more important for cognitive development than internally guided learning processes. For example, the information-processing approach has revealed large intraindividual differences in solving structurally similar tasks that nonetheless differ in content. The influence of learned declarative as well as procedural knowledge on the development of cognitive competencies has been confirmed convincingly. Studies on how novices develop into experts in specific contexts have also shown what outstanding levels of performance can be achieved through long-term, deliberate practice (Siegler 1991).

Even though relations between innate and acquired competencies, between ontological and logical restrictions to learning, between domain-specific and domain-unspecific learning opportunities, between explicit and implicit learning modes, between neurobiological and behavioral indicators of learning, are currently a field of intensive research, new methodological paradigms, and numerous theoretical speculations, they have still not been subjected to solid and consistent theory formulation (Richardson 1998).

4. *Learning and Instruction: Are Learners Really Their Own Best Instructors?*

The role and function of instruction for learning, particularly for academic learning, were already controversial topics in the philosophical tradition of educational science, and this has not changed during the more than 100-year-old history of educational psychology and the scientific study of teaching methods. Some examples of what are, in part, very different conceptions of teaching methods are non-directive instruction as an aid toward self-generated insight in the student (the Socratic method); the adaptation of learning goals, learning conditions, and learning methods to the mental state of the individual learner; the application of psychological laws to initiate learning and make it successful (experimental teaching methods at the beginning of the twentieth century); open but stimulating learning environments to arouse students mentally (the educational reform movement of the twentieth century); the planning, organization, structuring, and evaluation of learning in the classroom through the teacher (direct instruction); or the moderation of the activities of independent and responsible student learners (self-regulated learning approach).

The theoretical (and sometimes ideological) controversies over the role and function of instruction for learning are particularly relevant at the start of the twenty-first century. In a comparison of psychological research on learning and educational research on teaching. Shulman (1982) even believed that he could recognize a potential paradox: 'Although the research on learning has taught us the importance of the active, transforming role of the learner, the research on teaching continues to demonstrate the importance of direct instruction, an approach which seems to suggest a passive view of the learner' (p. 97).

The apparantly paradoxical pattern of results from experimental and developmental research on the one hand, and research on classroom instruction on the other hand, is not at all contradictory when one goes beyond the negative stereotype associated with direct instruction and considers the features and operationalizations used in many models of direct instruction. Direct instruction can be characterized by the following points: (a) The teachers classroom management is effective and the rate of student's interruptive behavior is very low; (b) the teacher maintains a strong academic focus and uses instructional time intensively to initiate and facilitate active, constructive, and goal-directed learning activities; (c) the teacher ensures that as many students as possible achieve successful learning processes by carefully chosing appropriate tasks, clearly presenting subject-matter information, continuously diagnosing each student's learning progress and learning difficulties, and providing effective help through remedial instruction.

Many studies have shown that instruction in which the teacher actively supports the learning process of the active and constructive working students is more effective than an educational strategy in which the teacher's only role is to provide for external conditions that made individual learning possible (Weinert and Helmke 1995).

Nonetheless, self-regulated learning has become a broad field of psychological research in recent years as

well as a powerful movement for educational reform (Boekaerts 1999). This reveals a broad consensus that self-directed learning is one of the most important goals of education. The only controversy is over whether this goal can be attained exclusively through self-directed learning activities by students, regardless of whether or not these learners possess appropriate learning strategies or how far the contents of learning are close to the individual learner's experience.

5. *Conclusions*

Although it has been possible to solve or clarify many scientific problems regarding the relation between cognitive development, learning, and instruction during the last century, psychology and educational science are still far from possessing satisfactory theoretical models. Nonetheless, the current state of knowledge reveals a few interesting perspectives and conclusions for future psychological research and current educational practice:

(a) At the present time, neurobiological and psychological research are still unable to work in concert, particularly with babies and young children. There is a need for improvement in the forms of interdisciplinary cooperation and a speeding up of knowledge transfer.

(b) A very important question seems to be what do we have to teach in order to ensure that certain competencies will be in any way acquired and lead to the emergence of the desired special patterns of cognitive development.

(c) As well as examining universal changes in development, there is a need for studies on interindividual differences in cognitive structures and processes to improve the understanding of learning effects and the effectiveness of instruction. Which deficient competencies can be compensated or substituted through learning and through instruction, and which trade-offs can be anticipated?

(d) As well as studying representative samples of children at different stages of development, basic research needs to pay more attention to the developmental course of both gifted and mentally retarded children. Which qualitative differences in ability and development can be found? What consequences do they have for the possibility, speed, and quality of learning? How can instruction be adapted in line with individual differences in intelligence, talent, and learning?

(e) It has proved to be unacceptable theoretically and also dysfunctional in terms of research strategy to study cognitive development in middle age and old age in an (inverted) analogy to childhood development. Developmental processes in the early and late sections of the lifespan differ in fundamental terms.

(f) No universal, unequivocal, and concrete recommendations for educational practice can be derived from the available general theories on development, learning, and instruction.

At the beginning of the twenty-first century, it therefore seems wise for school administrators and teachers not to follow radically one-sided scientific recommendations, but to apply variable combinations of instruction methods to meet different educational goals and comply with the different learning preconditions in their students.

See also: Childhood and Adolescence: Developmental Assets; Cognitive Development: Child Education; Cognitive Development in Childhood and Adolescence; Cognitive Psychology: Overview; Cognitive Styles and Learning Styles; Education: Phenomena, Concepts, and Theories; Educational Learning Theory; Instructional Design; Instructional Psychology; Intelligence, Prior Knowledge, and Learning; Learning and Instruction: Social-cognitive Perspectives; Learning Theories and Educational Paradigms; Piaget's Theory of Human Development and Education; School Achievement: Cognitive and Motivational Determinants

Bibliography

Binet A, Simon R 1905 Application des méthodes nouvelles au diagnostic du niveau intellectuelle chez des enfants normaux et anormaux d'hospice et d'école primaire. *Année Psychologique* **11**: 245–336

Boekaerts M 1999 (ed.) Self-regulated learning: Where we are today. *International Journal of Educational Research* 31: 443–551

Carey S 1984 Cognitive development. The descriptive problem. In: Gazzaniga M S (ed.) *Handbook of Cognitive Neuroscience*. Freeman, New York, pp. 37–66

Flavell J H 1970 Cognitive change in adulthood. In: Goulet R, Baltes P. B. (eds.) *Life-span Developmental Psychology: Research and Theory*. Academic Press, New York, pp. 247–53

Gagné R M 1968 Contributions of learning to human development. *Psychological Review* **75**: 177–91

Geary D C 1995 Reflections of evolution and culture in children's cognition. *American Psychologist* **50**: 24–36

Herrnstein R J, Murray C 1994 *The Bell Curve*. The Free Press, New York

Kohlberg L, Mayer R 1972 Development as the aim of education. *Harvard Educational Review* **49**: 294–303

Piaget J 1970 Piaget's theory. In: Mussen P H (ed.) *Carmichael's Manual of Child Psychology*. Wiley, New York, Vol. 1, pp. 703–32

Richardson K 1998 *Models of Cognitive Development*. Psychology Press, Hove, UK

Shulman L S 1982 Educational psychology returns to school. In: Kraut A G (ed.) *The G Standley Hall Lecture Series*. American Psychological Association, Washington, DC, Vol. 2, pp. 77–117

Siegler R S 1991 *Children's Thinking*, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ

Slavin R E 1987 Mastery learning reconsidered. *Review of Educational Research* **57**: 175–213

Sternberg R S 1998 Abilities are forms of developing expertise. *Educational Researcher* **27**: 11–20

Weinert F E, Helmke A 1995 Learning from wise mother nature or big brother instructor: The wrong choice as seen from an educational perspective. *Educational Psychologist* **30**: 135–42
Wohlwill J F 1980 Cognitive development in childhood. In: Brim O G, Jr., Kagan J (eds.) *Constancy and Change in Human Development*. Harvard University Press, Cambridge, MA, pp. 359–444

F. E. Weinert

Cognitive Dissonance

1. *Foundations of Dissonance Theory*

The theory of cognitive dissonance is elegantly simple: it states that inconsistency between two cognitions creates an aversive state akin to hunger or thirst that gives rise to a motivation to reduce the inconsistency. According to Leon Festinger (1957), cognitions are elements of knowledge that people have about their behavior, their attitudes, and their environment. As such, a set of cognitions can be unrelated, consonant, or dissonant with each other. Two cognitions are said to be dissonant when one follows from the obverse of the other. The resultant motivation to reduce dissonance is directly proportional to the magnitude and importance of the discrepant cognitions, and inversely proportional to the magnitude and importance of the consistent cognitions. This tension is typically reduced by changing one of the cognitions, or adding new cognitions until mental 'consonance' is achieved. Festinger's original formulation proved to be one of the most robust, influential, and controversial theories in the history of social psychology. Although a number of challenges and revisions have been suggested, the basic behavioral observation remains uncontested and continues to stimulate fresh research.

Application of this theory has yielded many surprising and nonintuitive predictions. For example, conventional wisdom suggests that behavior follows from attitudes; dissonance theory, however, identifies conditions under which just the opposite occurs. An early and often replicated experiment illustrates the power and counterintuitiveness of the theory. In what is now known as the induced compliance effect, Festinger and Carlsmith (1959) asked individuals to perform 30 minutes of a mind-numbingly tedious activity, and then to persuade a waiting participant that the activity was in fact quite interesting. This situation created cognitive dissonance in most individuals—they believed that the task was boring, yet inexplicably found themselves arguing quite the opposite. Half of the participants were given a ready excuse for telling this lie—they were paid $20 to do so—while the other half, paid only $1, had no such excuse. Those with a clear justification for their odd behavior experienced no dissonance and, as one would expect, later reported that the task was rather boring. The other half, however, given insufficient justification for their behavior, experienced dissonance between the knowledge that the task was boring and the reality that they were misleading a fellow participant into believing the opposite. Rather than endure the aversive experience of believing one thing but saying another, these individuals changed their opinion and convinced themselves that the task was actually interesting. In other words, their attitude was shaped by their behavior.

Subsequent studies have confirmed the basic theory of cognitive dissonance and demonstrated its far-reaching impact. For example, cognitive dissonance explains the increased commitment so frequently observed following a severe initiation into a group. The theory also explains why, when faced with a choice among several desirable options, we observe the tendency to highlight positive aspects of the chosen option and negative aspects of the rejected alternatives after (and only after) the choice has been made. In the course of such studies, we have learned much about the boundary conditions associated with the theory and have identified anomalies not easily explained by the original theory. Since the 1960s, a number of theoretical revisions have sought to subsume these limitations under a unifying theory. This article summarizes briefly the leading reformulations of dissonance theory and speculates on future directions.

2. *An Early Theoretical Challenge: Self-perception Theory*

Dissonance did not come quietly into psychology, nor did controversy begin over the finer points of the underlying process. Rather, the theory challenged the reigning theoretical paradigm of behaviorism of the 1950s by questioning the sovereign utility of basic learning theory. Rather than increased rewards leading to more positive attitudes, Festinger and Carlsmith (1959) had shown quite the opposite: participants who received a smaller reward for counterattitudinal advocacy developed more positive attitudinal responses. The majority of early criticisms, therefore, focused not on the details of the dissonance theory but rather on its fundamental legitimacy. During parts of the next two decades, one of the enduring intellectual feuds in social psychology debated whether dissonance phenomena were the result of complex cognitive processes in the mind of the participant, or whether they were merely the result of complex cognitive processes in the mind of the experimenter. Daryl Bem (1972) took the position that one could derive similar predictions through far more parsimonious behavioral processes. He argued that participants in dissonance experiments were not experiencing negative psychological tension due to inconsistency, but rather were simply inferring their attitudes from their behavior and the situation in

which it occurred. In essence, he suggested that people view their own behavior as though they were outside observers, and infer their underlying attitude from an analysis of their behavior. In support of this position, Bem replicated Festinger and Carlsmith (1959) and demonstrated that independent observers, aware of the monetary inducements, attributed attitudes to the participants that were nearly identical to the participants' actual attitudes. The critical test between these theories revolved around the search for physiological arousal. Festinger was quite clear that inconsistent cognitions created an aversive motivational state that could presumably be measured. Bem's self-perception theory, by contrast, predicted neither psychological nor physiological tension. However, Zanna and Cooper (1974) found indirect evidence for physiological arousal by showing that if arousal were misattributed to an irrelevant source, the effects of dissonance disappeared. The debate over dissonance vs. self-perception was finally laid to rest by a series of experiments that identified the precise conditions under which each was operative. Fazio et al. (1977) found that small discrepancies between attitude and behavior (defined as those within a person's latitude of acceptance) tended to elicit self-perception processes, but larger discrepancies (those that fall in the person's latitude of rejection) were more likely to generate dissonance processes. Croyle and Cooper (1983) added direct support for Festinger's original position by showing that engaging in counterattitudinal advocacy, at least outside people's latitude of acceptance, is marked by measurable increases in people's skin conductance responses. Ultimately, the importance of self-perception theory lay in its contributions to identifying boundary conditions of dissonance processes and in provoking research that established physiological arousal as one of the hallmarks of dissonance.

3. *Introduction of the Self*

It may be that all cognitive inconsistencies are not psychologically equivalent. According to some theorists, inconsistencies that implicate aspects of the self maintain a privileged position. This suspicion led Elliot Aronson (1968) and Claude Steele (1988) to consider the role of self-concept in the dissonance process. Aronson concurred with Festinger's view that dissonance was caused by inconsistencies, but argued that it was a particular inconsistency that mattered most in arousing dissonance, i.e., the discrepancy between a person's general expectations for the self and his or her actual behavior. In other words, the arousal due to dissonance came about when a person's belief that he or she was a good and rational individual was called into question by behavior that was neither good nor rational. Aronson predicted that dissonance arousal would be more frequent and more powerful among those with high self-esteem—that is, among those whose past history had led them to believe that their high internal standards of behavior were likely to be achieved. By contrast, Aronson predicted that those with low self-esteem, who were accustomed to behaving less competently, would not be surprised or discomfited to find themselves once again behaving in an incompetent manner.

To illustrate his point, Aronson (1968) argued that the dissonance aroused in Festinger and Carlsmith's (1959) experiment was not due to inconsistency between the thoughts 'I believe the task was dull' and 'I told someone the task was interesting.' Instead, Aronson proposed that dissonance was aroused by inconsistency between cognitions about the self (e.g., 'I am a decent and truthful human being') and cognitions about the behavior (e.g., 'I have misled a person … (and) conned him into believing something that just isn't true,' p. 24). Aronson concluded: 'at the very heart of dissonance theory, where it makes its clearest and neatest prediction, we are not dealing with just any two cognitions; rather we are usually dealing with the self-concept and cognitions about some behavior. If dissonance exists it is because the individual's behavior is inconsistent with his self-concept' (1968, p. 23).

Claude Steele and his colleagues (1988) suggested a different interpretation of the role of the self in creating dissonance. Like Aronson, Steele viewed inconsistent cognitions as a threat to the self; but unlike Aronson's self-consistency model, he suggested that the primary function of dissonance reduction was not to rescue the specific self-cognitions threatened by a behavioral outcome, but instead to restore the completeness of the overarching self-system. The difference, then, was not about the origin of dissonance arousal, but rather about the purpose and mechanism underlying dissonance reduction. Whereas Aronson focused on individual susceptibility to dissonance arousal, Steele's self-affirmation theory focused on individual resilience to dissonance. Interestingly, from Aronson's point of view, it is those individuals who have a positive self-regard that are most likely to experience dissonance arousal. Steele, by contrast, asserted that it is those same individuals—wrapped in their armor of self-resources—who feel immune from the need to reduce dissonance. Based on the data provided by Aronson and Steele, it is probably fair to say that those with high self-esteem are both more susceptible to dissonance arousal and more resistant to its effects because they can focus on their many other strengths. Holding all else constant, high self-esteem serves both as a catalyst that generates dissonance and as a buffer that mitigates the need to reduce dissonance.

4. *A New Look at Dissonance*

In 1984, Cooper and Fazio provided a comprehensive review of the dissonance literature and challenged the

dominant assumption that dissonance was driven by a need for psychological consistency. According to their 'New Look' model, dissonance is aroused when people perceive that their behavior has been responsible for bringing about consequences that are unwanted or aversive. If there are no such consequences, then inconsistent behavior will not produce the state of dissonance. For example, Cooper and Worchel (1970) replicated Festinger and Carlsmith's (1959) study with a condition in which the waiting participant was not convinced by the subject's lie. In this condition, the aversive consequence of misleading a fellow participant was removed along with all evidence of the dissonance process. Cooper and Fazio concluded from this and many other studies that responsibility for an aversive event rather than cognitive inconsistency plays the vital role in producing cognitive dissonance.

5. *Future Directions—Self-standards Model of Dissonance*

Each of these reformulations of dissonance theory differs with respect to one major issue: what is the role of self-concept in dissonance processes? Is it a problem, a benefit, or completely irrelevant to the arousal and reduction of cognitive dissonance? There are data in favor of each position—data that are not easily reconcilable. However, a recent synthesis discussed by Cooper (1999) and Stone (1999) suggests that dissonance is caused by a discrepancy between the outcome of a behavioral act and the standard to which it is compared. According to the self-standards model, sometimes the standard that people use to measure their behavioral outcomes are personal and idiosyncratic. In such cases, people's views of themselves will play a crucial role. At other times, the assessment of an act is based on broad, normative standards that are shared in the culture. At these times, the self will not play a role in the dissonance process.

In summary, it is still useful to think of dissonance as involving inconsistency among cognitive elements and to conclude that inconsistency produces motivation for change. It is also fair to conclude that the future of dissonance theory will include a role for behavioral consequences, an assessment of the self, and an analysis of the contextual variables that make different standards the basis of judgment for behavioral outcomes. The evolution of cognitive dissonance calls for an integration that will likely include insights from the currently dominant perspectives.

See also: Attitudes and Behavior; Motivation and Actions, Psychology of; Motivation: History of the Concept; Self-monitoring, Psychology of; Social Comparison, Psychology of

Bibliography

Aronson E 1968 Dissonance theory: Progress and problems. In: Ableson R P, Aronson E, McGuire W J, Newcomb T M, Rosenberg M J, Tannenbaum P H (eds.) *Theories of Cognitive Consistency: A Sourcebook*. Rand McNally, Chicago

Bem D J 1972 Self-perception theory. *Advances in Experimental Psychology* **6**: 1–62

Cooper J 1999 Unwanted consequences and the self: In search of the motivation for dissonance reduction. In: Harmon-Jones E, Mills J (eds.) *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology*. American Psychological Association, Washington, DC, pp. 149–73

Cooper J, Fazio R H 1984 A new look at dissonance theory. *Advances in Experimental Psychology* **17**: 229–62

Cooper J, Worchel S 1970 Role of undesired consequences in arousing dissonance. *Journal of Personality and Social Psychology* **16**: 199–206

Croyle R, Cooper J 1983 Dissonance arousal: Physiological evidence. *Journal of Personality and Social Psychology* **45**: 782–91

Fazio R H, Zanna M P, Cooper J 1977 Dissonance and self-perception: An integrative view of each theory's proper domain of application. *Journal of Experimental Psychology* **13**: 464–79

Festinger L 1957 *A Theory of Cognitive Dissonance*. Row, Peterson, Evanston, IL

Festinger L, Carlsmith J M 1959 Cognitive consequences of forced compliance. *Journal of Abnormal Social Psychology* **58**: 203–10

Steele C M 1988 The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Psychology* **12**: 261–302

Stone J 1999 What exactly have I done? The role of self-attribute accessibility in dissonance. In: Harmon-Jones E, Mills J (eds.) *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology*. American Psychological Association, Washington, DC, pp. 175–200

Zanna M P, Cooper J 1974 Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology* **29**: 703–9

J. Cooper and K. M. Carlsmith

Cognitive Functions (Normal) and Neuropsychological Deficits, Models of

The field of cognitive neuropsychology centers around two coupled goals: to use patterns of cognitive deficits in brain-damaged patients to inform theories and models of how cognitive processes are carried out by the brain, and to apply existing models to explain the specific deficits of individual patients in order to design more effective strategies for remediating these deficits. The roots of this effort can be traced back to the pioneering work of Broca, Wernicke, and Lichtheim in the mid- to late nineteenth century. These neurologists attempted to decompose complex cognitive func-

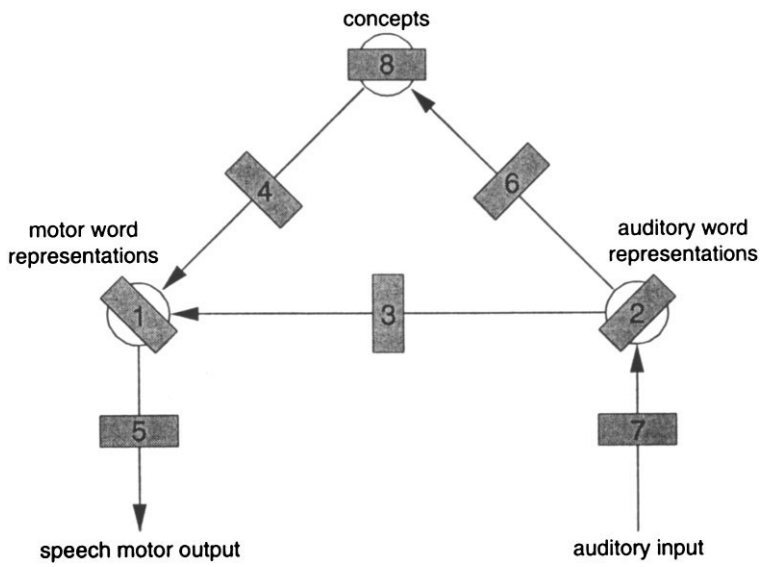

Figure 1
Lichtheim's (1885) model of aphasia. Grey bars indicate lesions to processing centers (circles) or pathways between them (arrows), giving rise to the following classic aphasia syndromes (in modern terminology): (a) Broca's aphasia; (b) Wernicke's aphasia; (c) conduction aphasia; (d) transcortical motor aphasia; (e) dysarthria; (f) transcortical sensory aphasia; (g) pure word deafness; (h) anomic aphasia

tions, such as language, into the joint operation of multiple functional 'centers' with specific patterns of connectivity between them. Damage either to the centers themselves or to the pathways between them were thought to give rise to distinct patterns of cognitive deficits. Indeed, the traditional scheme used today to categorize patterns of language impairment into distinct clinical syndromes—such as Broca's aphasia, Wernicke's aphasia, transcortical sensory or motor aphasia, etc.—derives from the Wernicke–Lichtheim model of the organization of the language system developed in the late nineteenth century (see Fig. 1).

1. *Box-and-arrow Modeling*

Although the form of explanation offered by these so-called 'diagram makers' was criticized roundly in the early twentieth century, it is echoed in modern-day cognitive neuropsychology in the form of box-and-arrow information processing models (see Fig. 2). While the functions ascribed to the centers or components are far more specific than in the nineteenth century, the same underlying explanatory logic is applied—the patterns of performance of brain-damaged patients are explained by positing one or more 'lesions' to the so-called functional architecture of the cognitive system. Typically, the predictions of the model, both in normal operation and under damage, have consisted of verbal descriptions based on fairly general notions about how the various modules would operate and interact. While these types of predictions may suffice for capturing the more general characteristics of normal and impaired cognitive functioning, they become increasingly unreliable as the model is elaborated to account for more detailed phenomena.

Two recent trends have increased the usefulness of box-and-arrow theorizing within cognitive neuropsychology. First, with improvements in techniques for structural lesion localization in patients and for functional brain imaging in both patients and normal subjects, there has been a more concerted effort to situate the components and pathways in specific brain regions. This is important because information on neuroanatomic localization places strong constraints on how components participate in various tasks and on how the system must be damaged to account for the performance of specific patients.

The second, and perhaps more important trend has been the development of working computer simulations of cognitive models that can both reproduce the characteristics of normal performance and can exhibit the appropriate deficits when damaged in a manner analogous to brain damage. Computational modeling makes it possible to demonstrate the sufficiency of the underlying theory in accounting for the phenomena by making the behavior of a detailed cognitive model explicit. A working simulation guarantees that the underlying theory is neither vague nor internally inconsistent, and the behavior of the

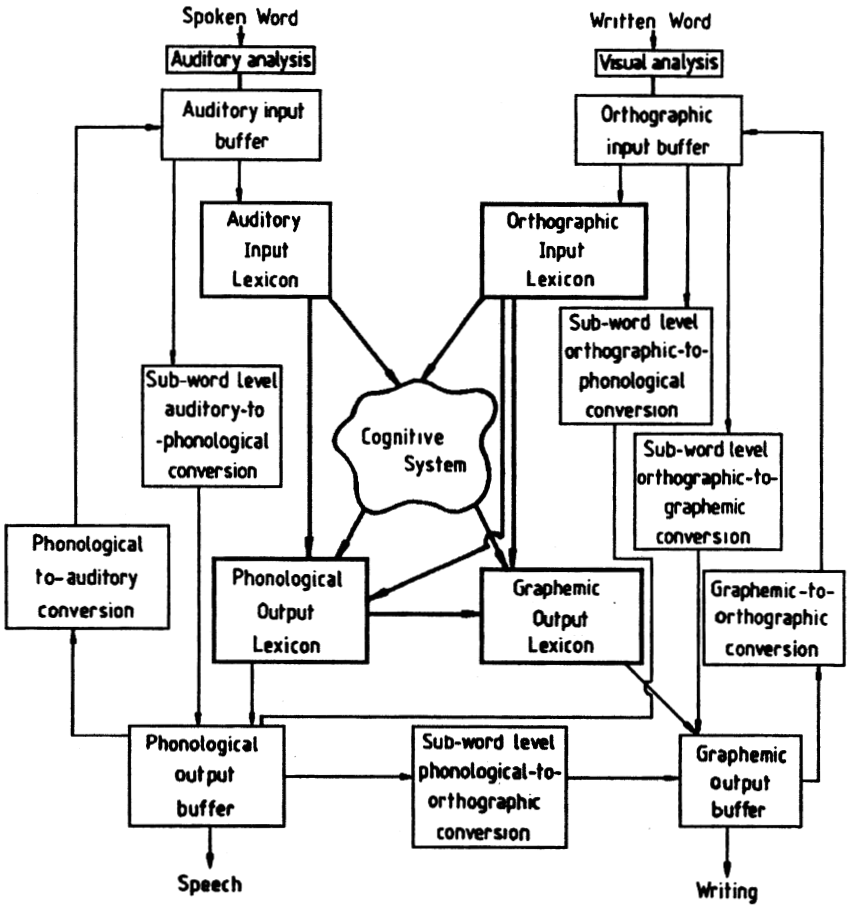

Figure 2
A box-and-arrow model of written and spoken word processing (from Howard and Franklin 1988)

simulation can be used to generate specific predictions of the theory.

One example of computational modeling based on box-and-arrow theorizing is the work of Coltheart and co-workers (Coltheart et al. 1993) in simulating a dual-route model of word reading. In the model, one pathway from print to sound applies grapheme-phoneme correspondence rules (e.g., B at the beginning of a word is pronounced /b/), while the other uses memorized whole-word correspondences. (These are the two pathways in Fig. 2 from the 'Orthographic input buffer' to the 'Phonological output buffer' that bypass the 'Cognitive System.') The rule route is effective for regular words (e.g., MINT) and for pronounceable but meaningless pseudowords (e.g., RINT); however, the lexical route is needed to pronounce exception words (e.g., YACHT, PINT) whose pronunciations violate the rules. In Coltheart and co-workers' implementation, the rule route consists of a collection of template-matching rules that operate left-to-right on the input string and generate single phonemes at fixed intervals. The lexical route is a version of a highly influential model of word recognition developed by Rumelhart and McClelland (1982), known as the Interactive Activation model, which contains a separate processing unit for each word in the vocabulary. Damage to the lexical route yields a reading pattern in which exception words are regularized (e.g., PINT read to rhyme with MINT), analogous to patients with acquired surface dyslexia (Patterson et al. 1985). Conversely, damage to the rule route causes impaired reading of pseudowords relative to words, corresponding to acquired phonological dyslexia (Beauvois and Derouesné 1979).

2. *Connectionist Modeling*

Although substantial progress has been made within the framework of box-and-arrow theorizing, many

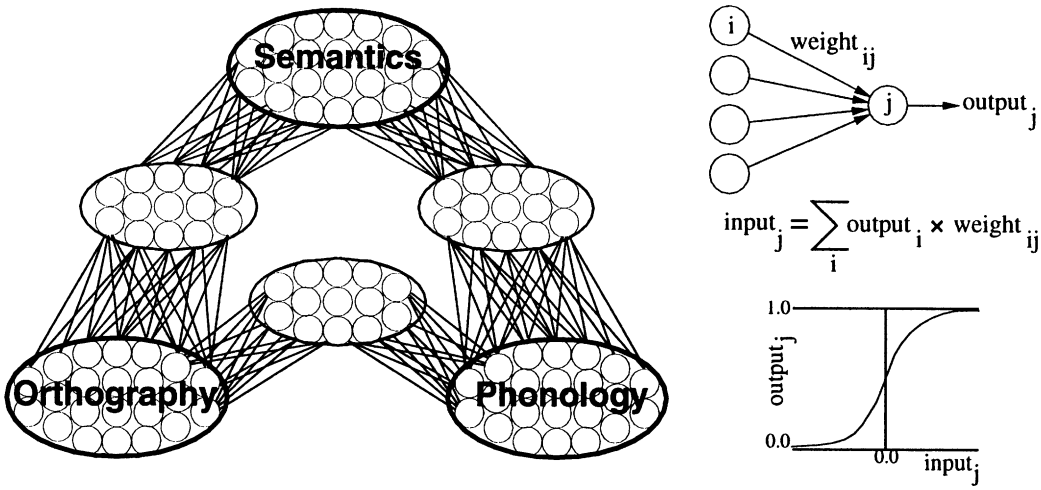

Figure 3
A distributed connectionist framework for lexical processing. The output of each unit is a smooth, nonlinear function of the summed, weighted input from other units (adapted from Plaut 1997)

researchers have come to believe that, in order to capture the full range of cognitive and neuropsychological phenomena, a formalism is needed that is based more closely on the style of computation employed by the brain. One such formalism that is widely used in connectionist modeling (see, e.g., McClelland et al. 1986, McLeod et al. 1998, Rumelhart et al. 1986).

In connectionist models—sometimes called neural networks or parallel distributed processing systems—cognitive processes take the form of cooperative and competitive interactions among large numbers of simple, neuron-like processing units (Fig. 3). Typically, each unit has a real-valued activity level, roughly analogous to the firing rate of a neuron. Unit interactions are governed by weighted connections that encode the long-term knowledge of the system and are learned gradually through experience. The activity of some of the units encodes the input to the system; the resulting activity of other units encodes the system's response to that input. The patterns of activity of the remaining units constitute learned, internal representations that mediate between inputs and outputs. Units and connections generally are not considered to be in one-to-one correspondence with actual neurons and synapses. Rather, connectionist systems attempt to capture the essential computational properties of the vast ensembles of real neuronal elements found in the brain through simulations of smaller networks of units. In this way, the approach is distinct from computational neuroscience (Sejnowski et al. 1988) which aims to model the detailed neurophysiology of relatively small groups of neurons. Although the connectionist approach uses physiological data to guide the search for underlying principles, it tends to focus more on overall system function or behavior, attempting to determine what principles of brain-style computation give rise to the cognitive phenomena observed in human behavior.

The simplest type of connectionist system is a feedforward network, in which information flows unidirectionally from input units to output units, typically via one or more layers of hidden units (so called because they are not visible to the environment). Such networks are useful in many contexts but have a limited ability to process time-varying information. In such contexts, recurrent networks, that permit any pattern of interconnection among the units, are more appropriate. In one common type of recurrent network, termed an attractor network, unit activities gradually settle to a stable pattern in response to a fixed input. Recurrent networks can also learn to process sequences of inputs and/or to produce sequences of outputs. For example, in a simple recurrent network (Elman 1990), the internal representation generated for each element in a sequence is made available as an additional input to provide context for processing subsequent elements. Critically, the internal representations themselves adapt so as to encode this context information effectively, enabling the system to learn to represent and retain relevant information at multiple time scales.

An issue of central relevance in the study of cognition is the nature of the underlying representation of information. Connectionist models divide roughly into two classes in this regard. In localist models, such as the Interactive Activation model mentioned earlier, each unit corresponds to a distinct,

familiar entity such as a letter, word, concept, or proposition. By contrast, in distributed models, such entities are encoded not by individual units but by alternative patterns of activity over the same group of units, so that each unit participates in representing many entities. Both localist and distributed models are 'connectionist' in the sense that the system's knowledge is encoded in terms of weights on connections between units.

Because localist models specify the form and content of representations, they tend to de-emphasize the role of learning. With a distributed model, by contrast, there is greater emphasis on the ability of the system to learn effective internal representations. Thus, instead of attempting to stipulate the specific form and content of the knowledge required for performance in a domain, the approach instead stipulates the tasks the system must perform, including the nature of the relevant information in the environment, but then leaves it up to learning to develop the necessary internal representations and processes.

Learning in a connectionist system involves modifying the values of weights on connections between units in response to feedback on the behavior of the network. A variety of specific learning procedures are employed in connectionist research; most that have been applied to cognitive domains, such as back-propagation (Rumelhart et al. 1986) take the form of error correction: change each weight in a way that reduces the discrepancy between the correct response to each input and the one actually generated by the system. Although it is unlikely that the brain implements back-propagation in any direct sense, there are more biologically plausible procedures that are computationally equivalent (see, e.g., O'Reilly 1996).

In an early application of error-correcting learning, Rumelhart and McClelland (1986) showed that a single network could learn to generate the past-tense forms of both regular verbs (e.g., BAKE ⇒ 'baked') and irregular verbs (e.g.,TAKE ⇒ 'took'), thereby obviating the need for dual rule-based and exception mechanisms (Pinker 1999), analogous to those in the dual-route reading models mentioned earlier. Although aspects of the approach were criticized strongly (Pinker and Prince 1988), many of the specific limitations of the model have been addressed in subsequent simulation work (see, e.g., MacWhinney and Leinbach 1991, Plunkett and Marchman 1993, 1996). Of particular interest is recent work by Joanisse and Seidenberg (1999) showing that damage either to phonological or to semantic representations within single processing system can account for the observation of selective impairments in performance on regular vs. irregular verbs following Parkinson's disease vs. Alzheimer's disease, respectively (Ullman et al. 1997).

A similar line of progress has taken place in the domain of English word reading. An early connectionist model (Seidenberg and McClelland 1989) provided a good account of word reading but was poor at pronouncing word-like pseudowords (e.g., MAVE, Besner et al. 1990). A more recent series of simulations (Plaut et al. 1996) showed that the limitations of this preliminary model stemmed from the model's use of poorly structured orthographic and phonological representations. By contrast, networks with more appropriate representations were able to learn to pronounce both regular and exception words, and yet also pronounce pseudowords as well as skilled readers. Moreover, damage to semantic representations in such networks gave rise to surface dyslexia, in which patients produce regularization errors to exception words. In closely related work (Hinton and Shallice 1991, Plaut and Shallice 1993), damage to the pathway between orthography and phonology, combined with secondary damage to the semantic pathway, yielded the complementary pattern of deep dyslexia—often viewed as a severe form of phonological dyslexia—in which patients are extremely poor at pronouncing pseudowords and make semantic errors in reading words aloud (e.g., misreading RIVER as 'ocean'; see Coltheart et al. 1980). With the additional of an attentional mechanism, fully recurrent networks have also been used to account for the interaction of both perceptual and lexical/semantic factors in the reading errors of neglect dyslexic patients (Mozer and Behrmann 1990). In fact, such networks have been applied to a wide range of neuropsychological phenomena, including selective impairments in face recognition (Farah et al. 1993), visual object recognition (Humphreys et al. 1992), spatial attention (Cohen et al. 1994), semantic memory (Farah and McClelland 1991), anomia and aphasia (Dell et al. 1997), spelling (Brown and Ellis 1994), and executive control (Cohen and Servan-Schreiber 1992).

One of the main attractions of distributed connectionist models is their ability to discover the structure implicit in ensembles of events and experiences. Accomplishing this, however, requires making only very small changes in response to each input so that the resulting weight values reflect the long-term experience of the system. Attempts to teach such networks the idiosyncratic properties of specific events one after the other do not generally succeed since the changes made in learning each new case produce 'catastrophic interference' with what was stored previously in the weights (McCloskey and Cohen 1989). McClelland et al. (1995) observed, however, that catastrophic interference does not occur if continued training of old knowledge is interleaved with the training of new knowledge. They proposed that the brain employs two complementary learning systems: a cortical system for gradual learning using highly overlapping distributed representations, and a subcortical, hippocampal-based system for rapid learning using much sparser, less-overlapping representations. On their account, stored instances in the hippocampus provide the training input for past experience that

must be inter-leaved with ongoing experience to prevent interference in cortex. The argument was that learning in cortex and in distributed networks are similarly constrained, so that the strengths and limitations of structure-sensitive learning in networks explained *why* the brain employs two complementary learning systems in hippocampus and neocortex.

Although fully recurrent networks are capable of learning to exhibit complex temporal behavior, for reasons of efficiency it is more common to apply simple recurrent networks in temporal domains. For example, Elman (1991) demonstrated that a simple recurrent network could learn the structure of an English-like grammar, involving number agreement and variable verb argument structure across multiple levels of embedding, by repeatedly attempting to predict the next word in processing sentences. St. John and McClelland (1990) also showed how such networks can learn to develop a representation of sentence meaning by attempting to answer queries about thematic role assignments throughout the course of processing a sentence. There have, however, been relatively few attempts at applying simple recurrent networks to neuropsychological phenomena.

3. Future Directions

In many ways, the application of computational modeling to understanding normal and impaired cognition is still in its infancy. Only a small fraction of the relevant behavioral phenomena have been addressed in any detail by existing models. Certainly considerable fruitful work remains in applying existing methods to a broader range of empirical issues. Even so, it seems clear that existing computational frameworks have a number of limitations which hamper their broader application. This is particularly true with regards to the application of connectionist networks to complex temporal domains, such as language, reasoning, and problem solving. While there have been some promising initial steps in these areas, substantial development of the computational methodology itself is likely to be necessary before satisfactory models will be possible.

4. Summary

Researchers interested in human cognitive processes have long used computer simulations to try to identify the principles of cognition. The strategy has been to build computational models that embody a set of principles and then examine how well the models capture human performance in cognitive tasks. A number of formalisms have been used to model cognitive processing in normal individuals. Those based on general principles of neural computation—including connectionist or neural-network models—have proven most effective at capturing the effects of brain damage on cognition. Considerable work remains, however, in extending such models to address more complex temporal phenomena.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Neural Networks: Neurocomputation; Cognitive Neuropsychology, Methodology of; Computational Neuroscience; Connectionist Approaches; Connectionist Models of Language Processing; Language Acquisition; Language Development, Neural Basis of; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Neural Networks: Biological Models and Applications; Neuropsychological Functioning, Assessment of; Syntactic Aspects of Language, Neural Basis of; Word Recognition, Cognitive Psychology of

Bibliography

Beauvois M-F, Derouesné J 1979 Phonological alexia: Three dissociations. *Journal of Neurology, Neurosurgery, and Psychiatry* **42**: 1115–24

Besner D, Twilley L, McCann R S, Seergobin K 1990 On the connection between connectionism and data: Are a few words necessary? *Psychological Review* **97**(3): 432–46

Brown G D A, Ellis N C (eds.) 1994 *Handbook of Normal and Disturbed Spelling*. Wiley, New York

Cohen J D, Romero R D, Servan-Schreiber D, Farah M J 1994 Mechanisms of spatial attention: The relation of macrostructure to microstructure in parietal neglect. *Journal of Cognitive Neuroscience* **6**(4): 377–87

Cohen J D, Servan-Schreiber D 1992 Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review* **99**(1): 45–77

Coltheart M, Curtis B, Atkins P, Haller M 1993 Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review* **100**(4): 589–608

Coltheart M, Patterson K, Marshall J C (eds.) 1980 *Deep Dyslexia*. Routledge and Kegan Paul, London

Dell G S, Schwartz M F, Martin N, Saffran E M, Gagnon D A 1997 Lexical access in normal and aphasic speakers. *Psychological Review* **104**: 801–38

Elman J L 1990 Finding structure in time. *Cognitive Science* **14**(2): 179–211

Elman J L 1991 Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* **7**: 195–225

Farah M J, McClelland J L 1991 A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General* **120**(4): 339–57

Farah M J, O'Reilly R C, Vecera S P 1993 Dissociated overt and covert recognition as an emergent property of a lesioned neural network. *Psychological Review* **100**(4): 571–88

Hinton G E, Shallice T 1991 Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review* **98**(1): 74–95

Howard D, Franklin S 1988 *Missing the Meaning?* MIT Press, Cambridge, MA

Humphreys G W, Freeman T, Müller H J 1992 Lesioning a connectionist model of visual search: Selective effects on distractor grouping. *Canadian Journal of Psychology* **46**: 417–60

Joanisse M F, Seidenberg M S 1999 Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Science, USA* **96**: 7592–7

Lichtheim L 1885 On aphasia. *Brain* **7**: 433–84

MacWhinney B, Leinbach J 1991 Implementations are not conceptualizations: Revising the verb learning model. *Cognition* **40**: 121–53

McClelland J L, McNaughton B L, O'Reilly R C 1995 Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–57

McClelland J L, Rumelhart D E, PDP Research Group (eds.) 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*. MIT Press, Cambridge, MA

McCloskey M, Cohen N J 1989 Catastrophic interference in connectionist networks: The sequential learning problem. In: Bower G H (ed.) *The Psychology of Learning and Motivation*. Academic Press, New York, Vol. 24, pp. 109–65

McLeod P, Plunkett K, Rolls E T 1998 *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford University Press, Oxford, UK

Mozer M C, Behrmann M 1990 On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience* **2**(2): 96–123

O'Reilly R C 1996 Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation* **8**(5): 895–938

Patterson K, Coltheart M, Marshall J C (eds.) 1985 *Surface Dyslexia*. Erlbaum, Hillsdale, NJ

Pinker S 1999 *Words and Rules: The Ingredients of Language*. Basic Books, New York

Pinker S, Prince A 1988 On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**: 73–193

Plaut D C 1997 Structure and function in the lexical system: Insights from distributed models of naming and lexical decision. *Language and Cognitive Processes* **12**: 767–808

Plaut D C, McClelland J L, Seidenberg M S, Patterson K 1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115

Plaut D C, Shallice T 1993 Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology* **10**(5): 377–500

Plunkett K, Marchman V A 1993 From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition* **48**(1): 21–69

Plunkett K, Marchman V A 1996 Learning from a connectionist model of the acquisition of the English past tense. *Cognition* **61**(3): 299–308

Rumelhart D E, Hinton G E, Williams R J 1986 Learning representations by back-propagating errors. *Nature* **323**(9): 533–6

Rumelhart D E, McClelland J L 1982 An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review* **89**: 60–94

Rumelhart D E, McClelland J L 1986 On learning the past tenses of English verbs. In: McClelland J L, Rumelhart D E, PDP Research Group (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*. MIT Press, Cambridge, MA, pp. 216–71

Rumelhart D E, McClelland J L, PDP Research Group (eds.) 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. MIT Press, Cambridge, MA

Seidenberg M S, McClelland J L 1989 A distributed, developmental model of word recognition and naming. *Psychological Review* **96**: 523–68

Sejnowski T J, Koch C, Churchland P S 1988 Computational neuroscience. *Science* **241**: 1299–1306

St. John M F, McClelland J L 1990 Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence* **46**: 217–57

Ullman M T, Corkin S, Coppola M, Hicock G, Growdon J H, Koroshetz W J, Pinker S 1997 A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience* **9**: 266–76

D. C. Plaut

Cognitive Maps

1. Introduction

A cognitive map is a representative expression of an individual's cognitive map knowledge, where cognitive map knowledge is an individual's knowledge about the spatial and environmental relations of geographic space. For example, a sketch map drawn to show the route between two locations is a cognitive map—a representative expression of the drawer's knowledge of the route between the two locations. Because a cognitive map represents/demonstrates an individual's geographic knowledge, geographers, psychologists, and others use them as the principle means by which to assess how we learn, process, and store geographic information gained from primary (e.g., walking through an area) and secondary (e.g., reading a map) sources. An understanding of how we undertake these mental tasks is thought to be important because it reveals the fundamental cognitive processes and structures that underlie spatial decision and choice making, and thus spatial behavior—why we choose certain routes, places to visit, locations to live, and so on (see *Spatial Cognition*; *Behavioral Geography*).

2. Cognitive Mapping

The combined process by which we learn, store, and use information relating to the geographic world is

Table 1
Lynch's classification

Category	*Description*
Paths	Paths are the channels along which an individual moves. They may include streets, walkways, railways
Edges	Edges are the linear elements not considered as paths. They are the boundaries between two phases, linear breaks in continuity such as shores or walls
Districts	Districts are the medium-to-large scale sections to the city, conceived as having a two-dimensional extent, which the observer mentally enters, and which have some common identifiable character
Nodes	Nodes are points, the strategic spots in the city into which an observer can enter, and which are the intensive foci to and from which (s)he is traveling. They may be primarily junctions, transportation changeovers, a crossing, or convergence of paths
Landmarks	Landmarks are another type of point-reference. They are usually a physical object such as a building, sign, store, or mountain

known as cognitive mapping, and this term is used to define the field of study which investigates this process. Cognitive mapping as a field of enquiry is relatively young. Whilst there are a handful of studies which predate 1960, the vast majority of research has occurred after this date and the publication of Kevin Lynch's seminal work 'The Image of the City'. Ever since Lynch's ground-breaking book, cognitive mapping research has been a multidisciplinary endeavor, undertaken by geographers, psychologists, anthropologists, computer, and information scientists. However, whilst a vast amount of research has been conducted as yet there is only limited consensus as to the fundamental processes of cognitive mapping. As such, there are a number of theories that seek to explain how we construct, process, and store cognitive map knowledge and how this knowledge is used to make spatial decisions and choices. These theories generally have been formulated on the basis of evidence provided by cognitive maps (see Downs and Stea 1977, Gärling and Golledge 1993, Portugali 1996, Golledge and Stimson 1997).

3. *Cognitive Maps*

Until very recently the bases and processes of human cognitive mapping was exclusively measured and assessed through cognitive (sketch) maps and other external forms of knowledge representation (e.g., estimating distances). More recently these methods of data generation have been supplemented by qualitative and neurological approaches. Moreover, there has been a transference of experimental setting from the laboratory to the natural environment.

Whilst a cognitive map is any 'map' which represents an individual's knowledge of an area, it generally takes the form of a sketch map drawn on a sheet of paper (it could, however, be a drawing in sand or a map constructed out of natural material). In experimental conditions, subjects are given a sheet of paper and are asked to draw a map of a certain location, area, or route between locations. The scale of the geographic realm to be drawn can vary substantially from the global (e.g., draw a map of the world) to the local (e.g., draw a map of your neighborhood). Variants on this simple sketch mapping exercise include providing respondents with a small portion of the map to provide a scale and reference, and teaching subjects a sketch map language where specific symbols are used to denote particular features. By aggregating together the cognitive maps of several individuals it is possible to determine their shared level of knowledge and which elements of an environment are most salient. This is the technique pioneered by Lynch. He analyzed individuals maps by classifying their elements into five different classes (see Table 1) which he then used to produce a composite map where the symbol size/shading density is proportional to the number of times an element appeared on the individual maps. Using this technique he aggregated together the sketch maps of residents in Boston, Jersey City, and Los Angeles to create composite cognitive maps of these cities (see Fig.1)

The analysis used by Lynch is a content classification. A number of other classification schemes have been used to analyze cognitive maps. For example, there have been classifications that assess map style, structure, and accuracy. In these cases the focus moves beyond what elements an individual draws to assessing the relationship between the elements and their relativity to the real world. In addition, the accuracy of the spatial relations portrayed can be analyzed statistically using spatial statistics. For example, in many studies bidimensional regression has been used to compare the geometry of the cognitive map to a cartographic map. Bidimensional regression is a two-dimensional equivalent of linear regression that quantifiably assesses scale, rotation, and translation differences between the actual and estimated pattern of responses.

Using the technique of sketch mapping to generate data about a person's cognitive map knowledge and

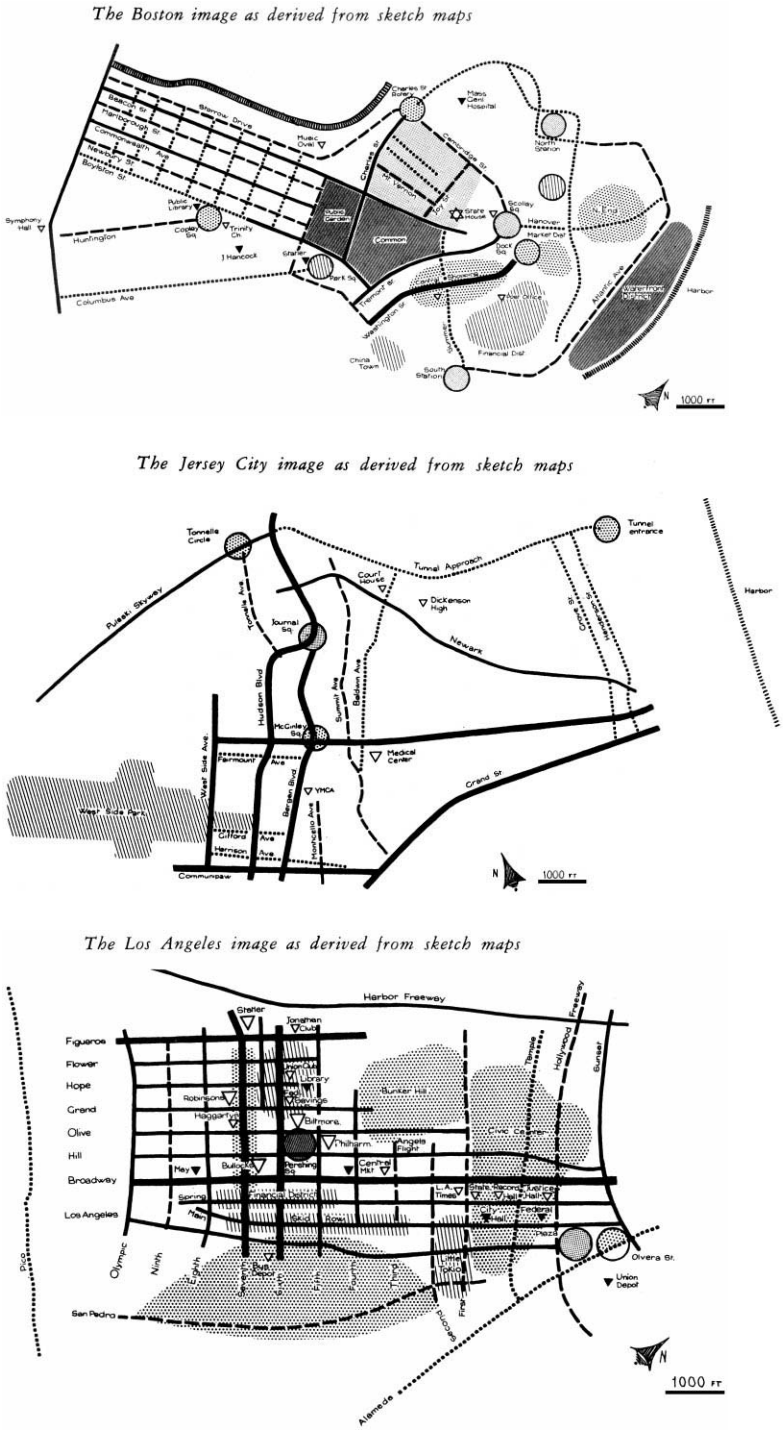

Figure 1
Cognitive maps of Boston, Jersey City and Los Angeles

their ability to use this knowledge is not without criticism. For example, a number of researchers have argued that sketch maps have a number of qualities that make them unreliable and inaccurate measures of spatial knowledge (note not geographic knowledge): they are dependent upon drawing abilities and familiarity with cartographic conventions; they suffer from associational dependence where later additions to the sketch will be influenced by the first few elements that are drawn; their content and style are influenced by the size of paper used for sketching; they are difficult to subjectively score and code; and they often show less information than the respondent knows. As a consequence of these criticisms it is becoming less common for researchers to use sketch mapping as an analytical tool to assess individual and collective cognitive mapping. Instead, researchers are turning to a range of other techniques.

4. *Other Ways to Measure Cognitive Map Knowledge*

4.1 *Other Action Measures*

Other forms of action (e.g., drawing) measures can be divided into those that, like sketch mapping, are two-dimensional in nature and those that are unidimensional. Other two-dimensional measures include completion tasks and recognition tasks. Completion tasks require an individual to complete a task that has already been started for them. For example, spatial cued response tests require subjects to place locations in relation to locations that are preplaced. A highly cued version of this are cloze procedure tests that require a subject to 'fill in' a missing space (an aspatial example of which would be, 'a dog barks but a cat ------?' Recognition tasks measure how successful subjects are at identifying spatial relationships. Iconic tests require the respondent to identify correctly features on a map or aerial photograph. Configuration tests require a subject to identify correctly which configuration, out of several, displays the correct spatial relations. Verifiable statement tests require subjects to identify whether a textual description of a spatial relationship is true or false. Spatial cued response data are often analyzed like sketch maps using bidimensional regression.

Cloze procedure and recognition tests are analyzed by constructing an accuracy score that reveals as a percentage the number of correct placements or recognitions.

Unidimensional tests seek to uncover one-dimensional aspects of cognitive map knowledge such as distance and direction. These dimensions are thought to be representative of spatial knowledge in general, but are particularly useful for measuring levels of route (procedural) knowledge. Distance tasks are used to assess an individual's knowledge of the distance between locations. In his review, Montello (1991) identifies five groups of tests designed to measure cognitive distance estimates: ratio scaling, interval and ordinal scaling, mapping, reproduction, and route choice. Ratio scaling adapts traditional psychophysical scaling techniques to a distance context, with subjects estimating the distance to a location as a ratio of some other known distance, such as an arbitrary scale or the length of a ruler. Interval and ordinal scaling are similar to ratio scaling but differ in their level of measurement: paired comparison requires a respondent to decide which one of a pair of distances is longer; ranking requires a respondent to rank various distances in order along the dimension of length; rating requires a respondent to assign the distance between places to a set of predetermined classes that represent relative length; partition scales require a respondent to assign distances to classes of equal-appearing intervals of length. Mapping is the measurement of distances from a sketch map for comparison with the actual distances. Reproduction requires a respondent to provide distance estimates at the scale of the estimated distance. Route choice consists of inferring judgments of cognitive distance from the choice of route an individual makes when asked to take the shortest route between two locations.

Like sketch map data, distance data can be analyzed individually or aggregated beforehand to provide data about a group. A common method to analyze ratio scaling, mapping, and reproduction data is to regress the cognitive distance estimates onto the objective distance values, observing the relationship between the two. Another common strategy, and one also used with interval or ordinal distance data, is to analyze the data using multidimensional scaling techniques (MDS). MDS techniques explore the latent structure of a set of distance estimates by assessing the dimensionality of the data. They do this by constructing a two-dimensional space from one-dimensional data using a series of algorithms. In essence, they construct a 'map' showing the relationship between a number of objects. This 'map' can then be compared to an actual map using techniques such as bidimensional regression.

Direction tasks assess an individual's knowledge of the direction between two locations. The most common direction task is pointing. Pointing involves standing at, or imagining being at, a location and pointing to another location. An alternative technique involves drawing on a compass the direction to a location from another. Direction estimates have been analyzed by comparing the estimates to the actual directions, often through a simple subtraction process. In other cases, a technique of projective convergence has been used to construct a 'map' from estimates by calculating where estimates to the same location but from different sites intersect.

4.2 Qualitative Approaches

In recent years there has been an increase in the use of qualitative methodologies to investigate cognitive map knowledge. In some cases, this has involved a scientific approach and in others an interpretative approach. The scientific approach continues the tradition described above, but rather than externally representing their knowledge through actions (e.g., drawing, pointing) individuals are required to describe verbally routes or layouts in experimental conditions. Like cognitive maps these data can be analyzed for content, style, structure, and accuracy. The interpretative approach, however, is less structured in terms of data collection. It posits that talking to and observing individuals as they interact with an environment reveals information concerning spatial behavior. Such an approach might seek to gain a spatial understanding of an area by adopting a strategy of in-depth interviews, discussing the reasoning behind spatial decision making.

4.3 Neurological Approaches

In contrast to the approaches above, which seek to understand the process of cognitive mapping by examining external measures (action or verbal), neurological approaches measure neural activity within the brain. A set of brain-scan techniques exist, differentiated by their temporal and spatial resolution. Magnetic Electroencephalography (MEG) has a high temporal resolution (0–300 ms) but low spatial resolution (providing only a general indication of neural activity). Positron Emission Tomography (PET) is the converse, with functional Magnetic Resonance Imagery (fMRI) providing a middle ground for both parameters. To reveal information about spatial cognition, scans are taken as the individual undertakes a series of spatial, problem-solving tasks. As such, neurological approaches seek to explain spatial thought and spatial behavior by identifying its neural, physiological bases rather than its psychological basis. To date there has been very little work that has sought to marry the findings of neurological and psychological research to provide a comprehensive, physiological and psychological model of cognitive mapping.

4.4 Naturalistic Settings

Many studies, particularly from psychology, which seek to understand cognitive map knowledge are conducted in controlled laboratory settings. For example, respondents may be required to learn the layout of objects in a experimental laboratory, and then to map the location of objects, or estimate distance and directions between objects. The laboratory is attractive to the researcher because they can control/monitor all the variables that may affect spatial learning and processing. Some researchers, however, question the ecological validity of this approach given that cognitive map knowledge concerns the geographic environment and it is this environment in which spatial behavior occurs. For these reasons they suggest that testing should occur in the natural environment, and increasingly this is becoming more common.

5. Summary

Cognitive maps are representative expressions of spatial knowledge. They are part of a wider set of analytical measures which seek to determine how we learn, store, and process knowledge of the geographic environment. These measures are important because they reveal fundamental aspects of cognition and reveal the cognitive processes that underlie spatial decision and choice making.

See also: Hippocampus and Related Structures; Knowledge (Explicit and Implicit): Philosophical Aspects; Mental Maps, Psychology of; Spatial Thinking in the Social Sciences, History of

Bibliography

Downs R M, Stea D 1977 *Maps in Minds: Reflections on Cognitive Mapping*. Harper & Row, New York

Gärling T, Golledge R G (eds.) 1993 *Behavior and Environment: Psychological and Geographical Approaches*. North Holland, Amsterdam

Golledge R G, Stimson R J 1997 *Spatial Behavior: A Geographic Perspective*. Guildford Press, New York

Lynch K 1960 *The Image of the City*. Technology Press, Cambridge, MA

Montello D R 1991 The measurement of cognitive distance: methods and construct validity. *Journal of Environmental Psychology* **11**: 101–22

Portugali J (ed.) 1996 *The Construction of Cognitive Maps*. Kluwer, Dordrecht

R. Kitchin

Cognitive Modeling: Research Logic in Cognitive Science

Cognitive science is a genuinely interdisciplinary field, which owes its existence to the insight that, in different disciplines, interesting research was based on the common assumption that cognition could be regarded as computation (see *Artificial Intelligence in Cognitive Science*; *Cognitive Science: Overview*). It follows that if cognition is computation, theories of cognition should be specified in terms of representations and the computational steps performed on them. Thus, cogn-

itive modeling follows naturally from the basic tenet of cognitive science.

Cognition has been addressed by philosophy for at least 2,500 years, by psychology since well into the nineteenth century, and by artificial intelligence since the mid-twentieth century—anyone would be ill-advised to mistake cognitive science as the only science of cognition. In fact, it overlaps considerably with cognitive psychology and parts of several other disciplines. Cognitive modeling is a unifying methodology for the whole field of cognitive science.

Cognitive modeling combines research methods of vastly different origin. The first group consists of techniques of formal analysis of tasks and systems, usually from philosophy, logic, theoretical linguistics, mathematics, physics, and the foundations of computer science. The second group consists of the empirical methods used predominantly in experimental psychology and in neuroscience, which are used to test models for cognitive adequacy. Finally, the third group of methods are the programming techniques developed in artificial intelligence, which are used to build working computer models. As a whole, the methodology combines formal and empirical analysis with constructive synthesis.

To identify cognitive modeling with computer simulation would be wrong for two reasons: first, we would ignore that building the computer model is just one, albeit essential, part of the methodology. Second, a computer simulation may be successful if it produces the same kinds of results, such as a commercial chess program, but a cognitive model (e.g., of a human cognitive function) must arrive demonstrably at the same results through the same kinds of computations (see also *Social Simulation: Computational Approaches*).

1. Cognitive Modeling as Second-order Model Construction

1.1 Epistemological Perspective

Cognition comprises sophisticated means of a system's adaptation to its environment, notably planning, which in turn draws on anticipation of the results of actions. Anticipation rests on learning and, in its most advanced forms, on episodic memory. Planning and decision rely on mental representations of the system's environment (world model) and the system as an actor in it (the system's self-model). The construction of these models is constrained by the interaction between system and environment. In the case of organisms, the necessities of evolution have ensured that internal world models are sufficiently realistic to be adaptive and ensure the survival of the species. In technical cognitive systems (e.g., autonomous robots), the need also arises to represent important features of the environment. Cognition therefore implies modeling the environment, the system itself, and other systems (as in discourse models used in communication).

Science in general aims at constructing models. Cognitive science attempts to model cognition in biological as well as technical systems. And since model building (in the sense discussed above) is an essential part of cognition, scientific models of cognition are about how cognitive systems construct models of their environment and of themselves. In short, cognitive modeling amounts to second-order modeling (and it is important to keep both levels well apart).

1.2 General Characteristics of Models

A model is a mapping from an empirical domain (a set of elements and certain relations defined between them) to another one, the model domain (often a numerical one, as in measurement). Modeling is constrained on both sides. The empirical domain, as viewed for modeling, is a highly reduced abstraction, and the model domain (a formal system) often includes relations that are not relevant to the model at all.

Scientific models are abstractions in the sense that the empirical domain grasped by the model comprises only part of the objects (elements) and their relations. Which ones to select depends on the epistemological interest, i.e., on the theoretical perspective as well as on intended applications. For instance, a typical driver's map of some region ignores geology, climate, biology, etc., and concentrates exclusively on roads, abstracting even there from most of the details. In cognitive modeling, we usually focus on certain aspects of mental representations (e.g., of a memory trace), and some relation or relations, which may be as different as association (linking two elements) and entailment (the semantic relation of logical implication between two statements).

On the model side, we must be careful to define which of the many known relations in the formal system being used to model the empirical domain is part of the model. This is well known from psychological scaling, where only a few of the relations known to hold among real numbers (if those have been chosen as the model domain) are valid model characteristics. For example, only the relation $\geqslant$ may be valid for an ordinal scale, while differences and ratios between numbers are meaningless. Likewise, the computer programs typically chosen in cognitive modeling exhibit a wealth of parameters and other details of data structures and control flow, some of which are certainly irrelevant as an aspect of the model. With computer programs, however, it is much more difficult to analyze just which relations are valid parts of the model (see Sect. 5.2).

2. Model Domains for Cognitive Modeling

The traditional framework of cognitive modeling has been defined by Marr (1982) and Newell (1982). They distinguish an abstract level of cognitive theory (the knowledge level) from the level of description typically chosen in cognitive modeling (symbol level, or algorithmic level). The level below that (implementation level) is considered as irrelevant to cognitive modeling. This does not exclude the construction of computational models for specific implementations, however. Such models have been used with much success as mediators between psychological and neurobiological analyses of brain functioning.

2.1 Symbol-processing Approaches

According to Newell and Simon (see *Artificial Intelligence: Connectionist and Symbolic Approaches*; *Cognitive Science: Overview*; *Problem Solving and Reasoning, Psychology of*), cognitive processes are symbol transformations on arbitrary complex symbol structures (i.e., mental representations). Accordingly, the classical approach to cognitive modeling aims to construct programs that manipulate symbol structures of compositional semantics by means of algorithms taken from artificial intelligence (e.g., heuristic search). This approach adheres closely to the Turing Machine model of computation.

This approach provides considerable degrees of freedom regarding how to go about constructing models. In practice, most of the work has made use of a production system architecture, or of a declarative knowledge base coupled with an inference algorithm. Therefore, the *if–then* rules of production systems and logical statements (e.g., the Horn clauses used in Prolog) are among the most widely used formalisms for cognitive modeling. Semantic networks, or frames, with inheritance (the *is a* relation) are another well-known example of this approach.

2.2 Connectionist Approaches

These approaches are different with respect to the algorithmic level. Simple elements or 'nodes' (which may be regarded as abstract neurons, see *Artificial Intelligence: Connectionist and Symbolic Approaches*; *Connectionist Approaches*) are connected in a more or less pre-specified way, the connectionist network's architecture. Each element's output is a function of its inputs integrated over time, and is passed on to other nodes that are connected with it. Two groups of connectionist models can be distinguished according to the semantics of representation employed: parallel distributed processing (PDP) and localist networks. In the latter each node is a representation of something (e.g., a concept), whereas in PDP it is the vector of activation values taken over a number of nodes that has representative character. This aspect of PDP models has been highlighted as pertaining to a 'subsymbolic' level by Smolensky (1988), who also stresses that artificial neural networks define a computational architecture that is nearer to symbol processing than to biological neural networks.

2.3 Nonlinear Dynamics and Other Approaches

Connectionist models, relying on differential equations rather than logic, paved the way to simulations of nonlinear dynamic systems (imported from physics) as models of cognition (see also *Self-organizing Dynamical Systems*).

Purely descriptive mathematical models have also been used in cognitive science, of course, but they do not take the form of an implemented computer program, and hence cannot be considered to be at the heart of cognitive modeling, but rather to be part of the formal analyses typically executed to arrive at sound specifications for cognitive models (see *Mathematical Models in Philosophy of Science*).

3. Cognitive Modeling and Cognitive Architectures

3.1 Unified and Modular Theories of Cognition

The process of cognitive modeling makes use of computational architectures, as we have seen for symbol processing and connectionist frameworks. As a special case, the framework may be a general theory about the architecture of the human mind, usually called a cognitive architecture.

Relying on a general cognitive architecture is to assume that all cognitive processes instantiate the same principle (e.g., firing a production rule). The present state of cognitive science casts doubt on the reasonableness of this assumption. The functional organization of the human brain is such that cognitive functions may be highly specialized, neuroanatomically focused, not open to introspection, and working in parallel with other cognitive processes (Kolb and Wishaw 1990).

3.2 General Cognitive Architectures

All these architectures take the form of a production system. This computational architecture, developed in the 1960s, comprises knowledge in the form of production rules (*if–then* rules), contained in a permanent memory, plus a working memory of unlimited capacity and a rule interpreter for control. Several lines of development have led to a number of systems that claim to be both a unified theory of human

cognition and a program development environment for cognitive modeling, among them SOAR (Laird et al. 1987, Newell 1990), ACT, known best in the versions ACT* (Adaptive Control of Thought; Anderson 1983) and its revision, ACT-R (Atomic Components of Thought; Anderson and Lebiere 1998), and others such as CAPS, EPIC, or PRODIGY. Apart from being production systems at heart, all these architectures differ markedly. For instance, SOAR relies on productions only, whereas ACT* also has a declarative memory (a spreading activation network of unlimited capacity). Both address learning, but differently (see Johnson (1998) for a more extensive comparison of ACT and SOAR). ACT-R (following EPIC there) now also includes a number of perceptual and motor components. This means that at least some of these architectures have grown beyond a uniform approach.

The advantages of modeling within a cognitive architecture are obvious. Much of the work of modeling has already been done, and programmers can use special predefined functions. At least the architectures of long-standing tradition (ACT and SOAR) boast a number of successful empirical tests and applications (see also *Cognitive Theory: ACT*; *Cognitive Theory: SOAR*. Drawbacks are that the modeler is confined to a specific architecture and its means, and that these architectures have grown so big that it becomes difficult to assess an architecture's relevance to some specific small-scale model.

3.3 Alternative Computational Frameworks

Cognitive modeling may be based on almost any computational model. These may be specialized (tailored) processing architectures of artificial intelligence, such as the blackboard architecture (Hayes-Roth 1985), case-based reasoning (see *Problem Solving and Reasoning: Case-based*), or architectures for autonomous social agents (see *Autonomous Agents*). Beyond symbol processing, we find connectionist models of the PDP (see McClelland (1999) for an overview) or the localist orientation (Page 2000). Arbitrary combinations of different approaches (hybrid systems) may also be employed (see also *Artificial Intelligence: Connectionist and Symbolic Approaches*).

It may be advantageous to base one's model of a specific cognitive function on a computational framework that is ideally suited for the task, and do without the massive overhead of the big cognitive architectures. The usual drawback is that one cannot build on the long work of others. COGENT (Cooper and Fox 1998) promises to ease the work of individual model building by providing, in modern object-oriented programming style, a toolkit for cognitive modeling with a user-friendly graphical interface for development.

At present, it is difficult to recommend a particular computational framework for cognitive modeling. What works best, or what is easier to develop, depends heavily on one's own experience, and on the cognitive function one wants to model.

4. Cognitive Modeling Produces Theories

The computer programs resulting from cognitive modeling have the status of a well-formulated theory about cognition. They have the advantage of being explicit (no computer would execute a 'magic' command) and fully specified (for the same reason), which means that theories cannot focus on certain issues while leaving others in the dark—something that is just too easily done on paper. As soon as the program has been implemented successfully on a computer, and produces the expected output, this is proof of the logical consistency of the theory and positive evidence (proof being impossible) for its adequacy.

4.1 Generative Theories as Compared with Scientific Laws

The most important characteristic of cognitive modeling is that it results in generative theories: computer programs that not only explain, but actually *produce*, the cognitive phenomena in question. This is very different from 'natural laws' stating, for example, that in our universe, nothing can travel with a speed beyond that of light.

Models in psychology and other behavioral and social sciences usually take the form of a system of equations for computing the value of some variables given the value of some other variables. These models, however, do not give a detailed explanation (as in an algorithm, i.e., a step-by-step computation without any 'magical' operations) of how the resulting values are arrived at. Being generative is the main advantage of cognitive models.

4.2 Validation Strategies in Cognitive Modeling

Cognitive modeling is more than just constructing a computer program that produces data more or less indistinguishable from 'real data' as gathered in psychological experiments. As a methodology, it includes testing and comparing models. This is where the pitfalls of cognitive modeling lie: they do not render themselves easily to the falsification strategy (Popper 1968) usually recommended in science.

First, anyone who has constructed a model that seems to work well is reluctant to focus on its weaknesses (but this is typical of all science: no one should misinterpret Popper as obliging individual scientists to falsify their own theories; it is sufficient

that there are other scientists around to do that.) Second, the programs that are the result of cognitive modeling are often highly complex, and their relation to empirical data cannot been tested exhaustively, but only for specific cases. Third, there is a fundamental problem: it is always the case that many models (each consisting of a representational and an operational, or process, part) may fit the same data.

Criteria for the empirical assessment of models are: good fit to data in essential aspects (e.g., difficulty of tasks is reflected in computational effort in the model; the model produces preferences, or errors, of the kind found in human subjects, etc.), and formal qualities (e.g., the fewer parameters on which the model depends, the better: a variant of Occam's razor). Typically, some empirical studies give rise to the model, and further empirical studies are used to test it.

Formal analyses of the tasks used in the domain (e.g., analyzing the syntactic structure and complexity of sentences in computational psycholinguistics) can and should be used not only to find hints as to how a model could be constructed, but also to exclude classes of models.

4.3 An Evaluation of Cognitive Modeling

Cognitive modeling is the research methodology which follows from the basic tenet of cognitive science that the essential aspect of cognition is that it is a computational process. Its main assets are that it produces theories which are explicit, complete, consistent, and generative. Formal analyses of the domain as well as empirical studies are required both as a prerequisite for the construction of models, and as the means of validating these models, e.g., to test their generalizability. Cognitive science since the 1970s has demonstrated that this is an extremely successful research strategy.

For further information, the reader is referred to Scarborough and Sternberg (1999) for an excellent collection of approaches to cognitive modeling.

See also: Computational Approaches to Model Evaluation; Connectionist Models of Concept Learning; Connectionist Models of Development; Connectionist Models of Language Processing; Knowledge Representation; Network Models of Tasks; Neural Networks: Biological Models and Applications; Scientific Discovery, Computational Models of

Bibliography

Anderson J R 1983 *The Architecture of Cognition*. Harvard University Press, Cambridge, MA

Anderson J R, Lebiere C 1998 *Atomic Components of Thought*. Lawrence Erlbaum Associates, Hillsdale, NJ

Cooper R, Fox J 1998 COGENT: A visual design environment for cognitive modeling. *Behavior Research Methods, Instruments and Computers* **30**: 553–64

Hayes-Roth B 1985 A blackboard architecture for control. *Artificial Intelligence* **26**: 251–321

Johnson T R 1998 A comparison between ACT-R and SOAR. In: Schmid U, Krems J F, Wysotzki F (eds.) *Mind Modelling: a Cognitive Science Approach to Reasoning, Learning and Discovery*. Pabst, Berlin, pp. 17–37

Kolb B, Wishaw I Q 1990 *Fundamentals of Human Neuropsychology*. Freeman, San Francisco

Laird J E, Newell A, Rosenbloom P S 1987 SOAR: An architecture for general intelligence. *Artificial Intelligence* **33**: 1–64

Marr D 1982 *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco

McClelland J L 1999 Cognitive modeling: Connectionist. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 137–44

Newell A 1982 The knowledge level. *Artificial Intelligence* **18**: 87–127

Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

Page M 2000 Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences* **23**: 443–512

Popper K R 1968 *The Logic of Scientific Discovery*, 5th edn. Hutchinson, London

Scarborough D, Sternberg S (eds.) 1999 *Methods, Models, and Conceptual Issues*. MIT Press, Cambridge, MA

Smolensky P 1988 On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1–74

G. Strube

Cognitive Neuropsychology, Methodology of

Cognitive neuropsychology is the study of what one can learn about the organization of the cognitive system from observing the behavior of neurological patients. Its utility derives from one aspect of the organization of the cerebral cortex, namely the relative localization of function, such that different regions of the cortex are differentially involved in different types of cognitive process. Thus patients can have disorders that are relatively specific to a single cognitive process, and so investigation of their impairments can provide critical information on the nature of the process.

One still controversial example is that of the so-called 'category-specific disorders' in which the patient loses the ability to identify one particular class of item. Thus certain patients with herpes simplex encephalitis can virtually entirely lose the ability to identify animals, plants, and foods, but show a far better preservation of their knowledge of human artifacts (Warrington and Shallice 1984). This is a loss of knowledge and not of early perceptual processes or the lower levels of language. By contrast other patients can be considerably better at identifying the 'animate categories' than human artifacts (Warrington and McCarthy 1983). This double dissociation was orig-

inally interpreted in terms of the semantic representations of sensory quality knowledge being stored separately from knowledge about function, the latter being the more critical for identifying artifacts. Such an approach has been elaborated into quite complex models (e.g., Farah and McClelland 1991). Alternative ways of modeling the double dissociation have, however, been produced (e.g., Gonnerman et al. 1997, but see Garrard et al. 1998). Moreover the original results have led to extensive functional-imaging investigations of the regions activated when different types of knowledge are concerned (e.g., Moore and Price 1999).

Very different methodologies have been used when patients are studied experimentally for what they can tell about the organization of cognitive functions. In the early twentieth century, when the nineteenth-century method of rather casual clinical descriptions was rejected for a more scientific approach, patients began to be studied as groups. Initially this took the form of a series of case histories (e.g., Head 1926), but in the 1950s and 1960s it became normal to average behavior over patients with particular characteristics. Groups could be defined by their gross lesion site (e.g., Milner 1958), or in terms of their global syndrome (e.g., in language fluent vs. nonfluent aphasia deriving from the work of Goodglass et al. 1964).

This approach has remained the standard one for those approaching neuropsychology with either a physiological psychology or clinical background. However in the 1960s and 1970s researchers more influenced by cognitive psychology began to use single-case studies increasingly because they were held to capture more satisfactorily from a theoretical perspective the essence of selective disorders corresponding to impairments of single cognitive processes. Indeed, in the 1980s it became popular amongst this group of researchers to argue that only single-case studies could be properly said to produce behavioral findings from neurological patients from which one could draw theoretical conclusions about the organization of the normal cognitive system (Ellis 1987). The position was best articulated by Caramazza (1986), whose argument was essentially that the behavior of the average of a group of subjects was not necessarily equivalent to that of any possible patient in the group. The argument will be discussed in more detail below.

One may now distinguish a variety of different types of experimental methods in regular use in cognitive neuropsychology. However I will begin with the simplest and arguably the most powerful method.

1. *The Single-case Study*

In this type of study one selects for study a patient whose disorder appears to be in conflict with the implications of an existing theory, or to suggest a novel theoretical possibility. An example of the first type of situation is the study of patient KF (Shallice and Warrington 1970), where it was shown that the patient performed quite normally on a variety of tests of auditory-verbal long-term memory but was grossly impaired on measures of what would now be called the 'phonological buffer' (Baddeley 1986), such as the recency component of the (auditory-verbal) free recall serial position curve. This was in conflict with memory theory of the time, which held that the laying down of information in the long-term (secondary) memory system required that it be retained initially in a short-term (primary) memory system. The study showed that measures of auditory-verbal long-term memory could be normal even if measures of auditory-verbal short-term memory were grossly impaired, and was held to be in conflict with a serial organization of the short- and long-term verbal memory systems.

An example of the second type is the analysis of the writing disorder of patient LB (Caramazza et al. 1987). In this study it was shown that semantic variables did not influence LB's rate of spelling errors, and also that lexical variables had little effect, but that performance was strongly dependent on word length. Moreover the errors could nearly always be interpreted in terms of single or double operations on single letters, such as substitutions, deletions, insertions, and transpositions. The pattern was held to support the existence of a graphemic buffer which held abstract letter-level information while words were being written.

Studies of these sorts should involve either two or three stages following the stage of patient selection. *First, one needs to carry out standard clinical (baseline) tests*. The single-case study methodology differs from that in most other sciences in that replication of any given result is not generally possible for other researchers in the field. If a highly counterintuitive result is obtained with apparently major theoretical consequences, such as the initial observations of category-specific semantic memory dissociations (e.g., Nielson 1946) or the observations of object-based neglect described in patient NG (Caramazza and Hillis 1990), then it is not open to other workers in the field to attempt to replicate it. They generally do not have access to the patient. Therefore it is necessary that as rich a description of the patient's general cognitive characteristics are given so as to provide the raw material for other researchers in the field to produce alternative explanations of any novel experimental findings. This is best provided by giving the patient's performance on standard clinical tests in the particular cognitive domain.

Second, the investigation will involve tests used to examine particular theoretical possibilities for the patient's difficulties. These can be obtained from the experimental psychological or neuropsychological literature or be specifically constructed. It is normally necessary that control data from normal subjects also be provided. Again because of the general impossi-

bility of replicating investigations in findings in single-case studies, all theoretical inferences drawn from the patient's performance should be based on the results of at least two experiments, or on a patient-internal replication of a single experiment. It is inappropriate to draw theoretical inferences from a single unreplicated experimental finding.

The third and optional stage of the procedure is a detailed analysis of the patient's performance in certain tests, rather than just using overall scores. This can include the effects of relevant stimulus variables, of different error types, of consistency of performance on the same items across replications of the same material, and so on.

The theoretical utility of the single-case study needs to be considered in the context of two different types of model. The first are models of the global cognitive architecture in a given domain in which the components are functional subsystems (Posner 1978) that are anatomically isolable and so can be individually damaged by neurological disease. Such models can be characterized as having a *modular functional architecture* where the term 'modular' is used with a less specific referent than in Fodor's (1983) sense. For such models, when a single subsystem is damaged, one will obtain a dissociation between impaired performance on tasks which require the subsystem and intact performance on tasks which do not require it. A standard move in cognitive neuropsychology is to treat the existence of a set of dissociations in a patient between tasks Y_1 to Y_i and task N_1 to N_j as evidence in favor of models where the pattern can arise from damage to one or more subsystems or connections, and against models where no such combination can produce the observed *overall pattern of performance*. This however depends on the following assumptions which are derived in part from those of Caramazza (1986) (see Shallice 1988 for a much more extensive discussion):

(a) Cognitive subsystems are qualitatively and quantitatively similar across individuals, at least by comparison with the effects of neurological disease.

(b) The task procedure, used by the patient, in the sense of the cognitive schemas that are controlling the overall operation of the collection of functional subsystems involved, does not differ markedly from that in use by normal subjects. This implies that the patient is not using a strategy unusual in normals, as, for instance, in letter-by-letter reading, which involves sounding out each letter in turn and then attempting to read by combining them into a word, as is found in the syndrome pure alexia.

(c) Reorganization of function following the lesion has had no more than a secondary effect in task performance. Clearly if the patient's cognitive system is qualitatively different from a normal system in its organization of subsystems, it will not be possible to make inferences from the patient's performance to the organization of the normal system.

(d) The tasks N_1 to N_j, on which performance is impaired, are not more demanding of some critical cognitive resource than tasks Y_1 to Y_i where performance is intact.

The most complex of these is the last, the presupposition that for most behavioral purposes the effects of damage to a subsystem can be ordered on a single dimension—resource. It has been argued by Glymour (1994) that the introduction of this concept makes cognitive neuropsychology a theoretically intractable enterprise, as far as the determination of the modular functional architecture using the observed overall pattern of performance of patients is concerned. Glymour argues that the degrees of freedom open to possible explanations would be greater than the constraints imposed by the observations of task performance. However, that analyses based on the resource concept can in practice be used to differentiate between alternative models has been shown in the concrete case of phonological output buffer disorder by Shallice et al. (2000).

The introduction of the resource concept allows one to infer that if a complementary set of dissociations is observed in a second patient producing a *double dissociation*, in at least one of the two cases, the pattern of Y_1 to Y_i contrasting with N_1 to N_j cannot be explained merely in terms of differential resource demands on a single subsystem being made across the set of tasks (see Shallice 1988, Chap. 10 for extended discussion). It should be noted, however, that these arguments are restricted to modular cognitive architectures. The existence of a double dissociation does not entail that the underlying system is modular; other possibilities include a continuum of processing space, within which different regions are damaged, overlapping processing regions, and systems in which different levels or aspects of operation of a subsystem can be separately damaged, as for instance on-line processing and the capacity to learn (see Shallice 1988, Chap. 11). However one alternative which does seem to be ruled out is an equipotential distributed system (Bullanaria and Chater 1995, but see Plaut 1995).

Inferences in neuropsychology from impaired behavior to mechanism are not, however, restricted to inferences based on the overall pattern of performance and the global organization of a modular functional architecture (see McCloskey 2000). The second type of model is those specifying how particular subsystems operate. More detailed aspects of the patient's performance can give more specific information about the operation of a processing system. Thus, in attentional dyslexia, patients show migrations of letters from one word to another (e.g., 'flip snag' to 'snip flag') (Saffran and Coslett 1996). This is compatible with damage to a mechanism which specifies over which parts of the visual field the results of letter-form level analysis are admitted for higher-level processing (see Mozer 1991 for more detailed discussion of the theoretical implications of the disorder).

2. The Multiple-case Study and the Functional Syndrome

Any individual patient's performance can be influenced by a variety of factors which reduce its relevance for inferences to normal processing. The patient may have an idiosyncratic functional or anatomical organization of their cognitive system, use an unusual strategy in some tests, or the disorder may even be influenced by psychiatric factors, so-called 'functional overlay.' One pragmatic way of partially guarding against these possibilities is to report two or more patients showing the same pattern of symptoms.

Second, if a dispute arises about the interpretation of empirical aspects of the report of an individual case, there is no way it can be resolved, other than by further investigation of the patient's difficulties, or reanalysis of existing findings. In practice these approaches are only rarely undertaken. Therefore, the recourse tends to be to find alternative arguments on the basic theoretical position held, in which case the original findings become redundant. (An example of an unresolvable theoretical question in the *individual patient* is arguments of whether evidence for syllabic structure in graphemic buffer disorder patient LB represents the effects of orthographic or phonological based processes; see Caramazza and Miceli 1990, Jonsdottir et al. 1996).

One way of making the neuropsychological database more solid is to establish *functional syndromes*, or characteristics held in common by all of a group of patients. However, a combination of characteristics can arise because of the anatomical proximity of *different* functional systems, producing so-called associated deficits. The concept of functional syndrome has therefore often been rejected metatheoretically on the grounds that a single counterexample can lead to the collapse of the putative functional syndrome. This is not necessarily the case. The functional syndrome may *fractionate* with both subvarieties being theoretically interpretable; one example is the two varieties of so-called letter-by-letter readers who circumvent the word-reading difficulty induced by their pure alexia. One group uses the names of the letter forms and knowledge of the spelling; the other uses a sounding-out strategy (see Patterson and Kay 1982). Moreover, the overall performance of a single patient can also derive from multiple impairments. This is not therefore an argument against the functional syndrome perspective, given that one restricts functional syndromes to ones in which all patients behave in the same fashion, and treats the subsyndromes resulting from fractionation of an original functional syndrome as each requiring theoretical explanation. The functional syndrome is therefore a concept which has the utility of producing a more solid database on which to assess alternative theories (see, for example, Shallice et al. 2000 for a concrete example of the approach).

Functional syndromes are of relevance for theories other than modular ones. Connectionist models can predict functional syndromes with complex characteristics too (see, for example, Mozer and Behrmann 1990). However, a different type of multiple single-case study is more usually relevant to this type of model. This is the use of a database for confronting theory of all patients showing deficits in a particular processing domain; their behavior is then individually fitted, but using different parameter values within a common underlying model. The prototypic example of this approach is the study of Dell et al. (1997) on naming errors in aphasics (but see Ruml et al. in press for detailed criticism).

3. The Anatomically Based Group Study

A method which is very frequently used is to group patients according to the location of lesion and average the results of each group. A potential problem with this approach is that patients in a given group may vary qualitatively as well as quantitatively. The reason is simple. The defining anatomical region will typically involve subsystems other than that which is the critical theoretical focus of the study. Thus Caramazza's (1986) argument on the functional heterogeneity of the group becomes relevant, in particular that the pattern of performance of the average of the group may not correspond to that of any individual representative of the group.

This possibility should not, however, be thought to invalidate the approach completely. While it presents a theoretical danger, its potency depends upon the nature of the averaged data and the nature of this theory. Thus, for certain types of theories the same problem occurs when the findings of normal subjects are confronted with theory. Second, for certain types of evidence, in particular double dissociations, it requires the conditions, which are *a priori* highly unlikely, for the pattern to arise as an average artifact and *not* be present in individual complementary pairs of subjects within the contrasting groups (see Shallice 1988, Chap. 9). Thus the existence of the double dissociation in the group data virtually always entails it occurring in specific pairs of individual subjects. This means that the situation when making inferences from the *overall pattern of performance* of groups to theory is virtually equivalent to that when making inferences from individual subjects. I know of no actual case where theoretical inferences from an anatomically based group study have been criticized as being vitiated by an averaging artifact.

A further key issue in anatomically based group studies is how to define the relevant region for patient inclusion. Typically, regions are defined *a priori*, such as lesions being confined to or involving a particular lobe. However, approaches are being developed in which determination of group membership is indirectly influenced by statistical procedures for analyz-

ing functional-imaging results, such as statistical parametric mapping (SPM) (Friston 1997). This has resulted in the development of more complex procedures for determining the specification of the anatomical regions most appropriate for most sharply differentiating good and poor performance on a given measure. For instance, Stuss et al. (1998) have used classification and regression test (CART) procedure (Breiman et al. 1984) to assign patients to subgroups.

4. The Functionally Based Group Study Approach

An alternative methodology is to constitute groups according to the functional characteristics, such as when studies contrast Broca's aphasics and Wernicke's aphasics, or amnesics and normal controls. In many respects the methodological issues raised are similar to those of the previous approach. However, a critical problem relates to the criteria used for group definition, and hence the possibility of replication. In studies contrasting amnesic and control groups, for instance, this has often been a critical issue (see Butters and Cermak 1974). However, in aphasia research it has been a key issue since the 1980s. Thus there has been a very long-standing and as yet unresolved controversy on whether the comprehension performance of Broca's aphasics mimics their production problems, in the sense of showing a dissociation between comprehension of active and passive sentences. Some early cases seemed to indicate that in certain patients it did not (see Berndt et al. 1996).

However, Grodzinsky et al. (1999) reject this meta-analysis on the grounds that it applied too loose criteria for patient selection. Some patients included, it was argued, were not Broca's aphasics. Caramazza et al. (2000) have pointed out that Berndt et al.'s findings are still obtained when the more restrictive criteria, endorsed by Grodzinsky et al., are applied.

To be satisfactory, this approach must be based on effective functional syndromes that do not fractionate and on good operational criteria for diagnosing them; Broca's and Wernicke's aphasia probably do not pass the test. The danger of the clinically based group study approach is that if the functional syndrome which is used to define group membership does fractionate, then the critique of Caramazza (1986) will be potentially highly relevant, and only very simple characteristics of the averaged findings will prove inferentially relevant.

See also: Brain Damage: Neuropsychological Rehabilitation; Case Study: Logic; Case Study: Methods and Analysis; Case-oriented Research; Cognitive Functions (Normal) and Neuropsychological Deficits, Models of; Neuropsychological Functioning, Assessment of

Bibliography

Baddeley A D 1986 *Working Memory*. Clarendon Press, Oxford, UK

Berndt R S, Mitchum C C, Haedinges A W 1996 Comprehension of reversible sentences in agrammatism: A meta-analysis. *Cognition* **58**: 289–308

Brieman L, Friedman J H, Olshen R A, Store C J 1984 *Classification and Regression Trees*. Wadworth, Belmont, CA

Bullinaria J A, Chater N 1995 Connectionist modelling: Implications for cognitive neuropsychology. *Language and Cognitive Processes* **10**: 227–64

Butters N, Cermak L S 1974 Some comments on Warrington and Baddeley's report of normal short-term memory in amnesic patients. *Neuropsychologia* **12**: 283–5

Caramazza A 1986 On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition* **5**: 41–66

Caramazza A, Hillis A E 1990 Spatial representation of words in the brain implied by studies of a unilateral neglect patient. *Nature* **346**: 267–9

Caramazza A, Miceli G 1990 The structure of graphemic representations. *Cognition* **27**: 243–97

Caramazza A, Cappa E F, Ray A, Berndt R S 2001 Agrammatic Broca's aphasia is not associated with a single pattern of comprehension performance. *Brain and Language* **76**: 158–84

Caramazza A, Miceli G, Villa G, Romani C 1987 The role of the graphemic buffer in spelling: Evidence from a case of acquired dysgraphia. *Cognition* **26**: 59–85

Dell G S, Schwartz M F, Martin N, Saffran E M, Gognon D A 1997 Lexical access in normal and aphasic speech. *Psychological Review* **104**: 801–38

Ellis A W 1987 Imitations of modelarity, or, the modularity of mind: Doing cognitive neuropsychology without syndromes. In: Coltheart M, Sartori G, Job R (eds.) *The Cognitive Neuropsychology of Language*. Erlbaum, London

Farah M J, McClelland J L 1991 A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology General* **120**: 339–57

Fodor J A 1983 *The Modularity of Mind*. MIT Press, Cambridge, MA

Friston K 1997 Analysing brain images: Principles and overview. In: Frackowiak R S J et al. (eds.) *Human Brain Function*. Academic Press, San Diego, CA, pp. 25–42

Garrard P, Patterson K, Watson P C, Hodges J R 1998 Category specific semantic loss in dementia of Alzheimer's type. Functional-anatomical correlations from cross-sectional analyses. *Brain* **121**: 633–46

Glymour C 1994 On the methods of cognitive neuropsychology. *British Journal of Philosophy of Science* **45**: 815–35

Gonnerman L, Anderson E S, Devlin J T, Kempler D, Seidenberg M S 1997 Double dissociation of semantic categories in Alzheimer's disease. *Brain and Language* **57**: 254–79

Goodglass H, Quadfasel F A, Timberlake W H 1964 Phrase length and the type and severity of aphasia. *Cortex* **1**: 133–53

Grodzinsky Y, Pinango M M, Zurif E, Drai D 1999 The critical role of group studies in aphasia: Comprehension regularities in Broca's aphasia. *Brain and Language* **67**: 134–47

Head H 1926 *Aphasia and Kindred Disorders of Speech*. Cambridge University Press, Cambridge, UK

Jonsdottir M, Shallice T, Wise R 1996 Language-specific processes in graphemic buffer disorder. *Cognition* **59**: 169–97

McCloskey M 2001 The future of cognitive neuropsychology. In: Rapp B (ed.) *The Handbook of Cognitive Neuropsychology*. Psychology Press, Philadelphia, pp. 593–610
Milner B 1958 Psychological deficits produced by temporal-lobe excision. *Research Publications, Association for Research in Nervous and Mental Disease* **36**: 244–57
Moore C J, Price C J 1999 A functional neuroimaging study of the variables that generate category-specific object processing differences. *Brain* **122**: 943–62
Mozer M C 1991 *The Perception of Multiple Objects*. MIT Press, Cambridge, MA
Mozer M C, Behrmann M 1990 On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience* **2**: 96–123
Nielson J M 1946 *Agnosia, Apraxia, Aphasia: Their Value in Cerebral Localization*. Hoeber, New York
Patterson K E, Kay J 1982 Letter-by-letter reading: Psychological descriptions of a neurological syndrome. *Quarterly Journal of Experimental Psychology, Series A* **34**: 411–41
Plaut D C 1995 Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Psychology* **17**: 291–321
Plaut D C, Shallice T 1993 Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology* **10**: 377–500
Posner M I 1978 *Chronometric Explorations of Mind*. Erlbaum, Hillsdale, NJ
Ruml W, Caramazza A 2000 An evaluation of a computational model of lexical access: Comment on Dell et al. 1997 *Psychological Review* **107**: 609–34
Saffran E M, Coslett B 1996 Attentional dyslexia in Alzheimer's disease: A case study. *Cognitive Neuropsychlogy* **13**: 205–28
Shallice T 1988 *From Neuropsychology to Mental Structure*. Cambridge University Press, Cambridge, UK
Shallice T, Warrington E K 1970 Independent functioning of the verbal memory stores: A neuropsychological study. *Quarterly Journal of Experimental Psychology* **22**: 261–73
Shallice T, Rumiati R I, Zadini A 2000 The selective impairment of the phonological output buffer. *Cognitive Neuropsychology* **17**: 517–46
Stuss D T, Benson D F 1984 Neuropsychological studies of the frontal lobes. *Psychological Bulletin* **95**: 3–28
Stuss D T, Alexander M P, Hamer L, Palumbo C, Dempster R, Binns M, Levine B, Izuakawa D 1998 The effects of focal anterior and posterior brain lesions on verbal fluency. *Journal of the International Neuropsychological Society* **4**: 265–78
Warrington E K, McCarthy R 1983 Category specific access dysphasia. *Brain* **106**: 859–78
Warrington E K, Shallice T 1984 Category specific semantic memory impairment. *Brain* **107**: 829–54

T. Shallice

Cognitive Neuroscience

The discipline has emerged in the 1990s at the interface between the neural sciences and the cognitive and computational sciences. On one side, it grows out of the traditions of cognitive psychology and neuropsychology, which use behavioral experiments to uncover the processes and mechanisms lying behind human cognitive functions, and of computational approaches within cognitive psychology, which rely on computational models to develop explicit mechanistic accounts of these functions. On the other side, it grows out of the traditions of behavioral, functional, and systems neuroscience, which use neurophysiological and neuroanatomical methods to explore the mechanisms underlying complex functions. It draws on findings and principles of cellular and molecular neuroscience. It joins these approaches with the use of new functional brain imaging methods, such as functional magnetic imaging (fMRI), positron emission tomography (PET), as well as other methods including electroencephalography (EEG) and magnetoencephalography (MEG), and with a growing research tradition in computational neuroscience.

1. *The Microstructure of Cognition*

1.1 *Patterns of Activity Arising in Ensembles of Simple Elements*

A starting point for cognitive neuroscience is the idea that a cognitive or mental state consists of a pattern of activity distributed over many neurons. For example, the experience an individual has when holding, sniffing, and viewing a rose is a complex pattern of neural activity, distributed over many brain regions, including the participation of neurons in visual, somatosensory, and olfactory, and possibly extending to language areas participating in representing the sound of the word 'rose' and/or other areas where activity represents the content of an associated memory that may be evoked by the experience.

These patterns of activation arise from excitatory and inhibitory interactions among the participating neurons, mediated by connections called synapses. The inputs neurons receive cause them to 'fire' or emit impulses called spikes or action potentials, which travel down their axons to synaptic terminals where they cause the release of chemicals that then have excitatory or inhibitory influences on the neurons on the other side of the synapse. The combined effect of the incoming signals to each neuron, together with its recent history, determines whether it will fire at a particular moment. Figure 1 indicates something of the fundamental circuitry involved, though it should be noted that only one out of 100 of the neurons in the tiny region shown (about 3×3 mm) are indicated. While the computations performed by individual neurons should not be underestimated (see *Neurons and Dendrites: Integration of Information*), it seems likely that what gives the system its power and complexity is the number of neurons involved (most estimates place the number in the human brain between 10 and 100 billion) and of the density of

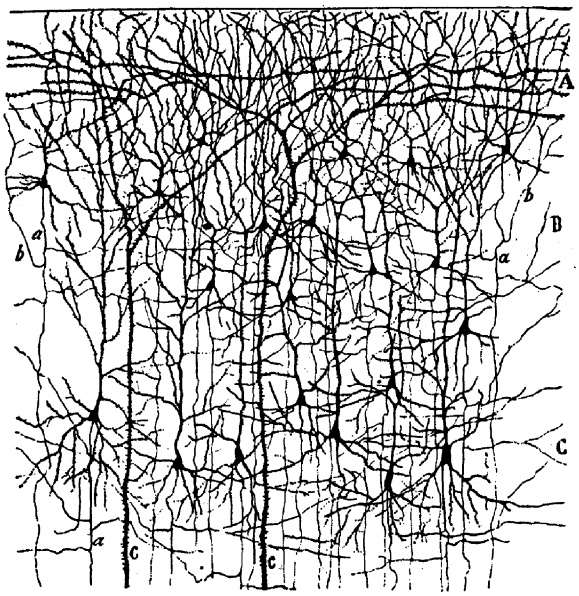

Figure 1
An early camera lucida drawing of the circuitry of the neocortex, based on the Golgi stain method, which impregnates just one out of every 100 cortical neurons. The diagram depicts the rich dendritic branching structure of the individual neurons present, whose cell bodies appear as small, pyramid-shaped blobs. The dendrites (and the little spines visible on the surfaces of some of the larger dendrites) are the structures on which the neurons receive most of their inputs from other neurons

connections among them (typical cortical neurons receive between 10,000 and 100,000 individual synapses from other neurons).

1.2 Distributed Representations

A great deal of research has concerned the nature of the active representations the brain uses for objects of perception or cognition, such as the rose discussed above. There is now a great deal of support for the view that the brain's representations typically consist of patterns of activity involving fairly large ensembles of neurons. Individual neurons are often described as 'detectors' for particular stimulus or situational features or conjunctions of features (e.g., the 'edge detectors' introduced by Hubel and Weisel 1962 in their seminal studies in visual cortex), but most such neurons are fairly broadly tuned, so that they will also be partially activated by a wide range of stimuli overlapping in one way or another with the optimal stimulus, and thus will participate at least partially in the representation of many different inputs.

A prime example of distributed representation is the representation of the direction of arm movements in the motor cortex. It appears that the representation of a particular direction of reaching is a pattern over a large population of neurons, each of which responds maximally to a particular preferred direction, but responds to a lesser degree to neighboring directions, and thus participates partially in the representation of many different directions of reaching (Georgopoulos et al. 1986). There are other types of distributed representations used in the brain, in which a neuron can participate in two different representations, without there being a clear shared feature or other similarity between the situations that cause the neuron to fire. For example, in the hippocampus, individual neurons participate in distributed representations of the animals' location in external space and other aspects of the current behavioral situation. Interestingly, the same neurons may participate in different ways in the representation of different environments, or even of two distinct representations of the same environment when the animal in performing different tasks (Markus et al. 1995).

1.3 Knowledge and Learning in the Strengths of Connections

The particular pattern of activation that arises in experiencing an input (or in reconstructing a memory or formulating an imagined experience) is determined by the connections among the neurons. A key issue, then, is to understand the processes that lead to the formation of the specific excitatory and inhibitory connections that shape the processes of perception, cognition, and action. Generally, it is thought that largely activity independent processes establish an initial skeleton framework of connectivity early in development, for example, causing connections to form between neurons in the retina of the eye and other neurons in the lateral geniculate nucleus, a way station for visual information on the way to the cortex. Then, activity-dependent processes selectively refine and stabilize some of the connections, and perhaps cause new ones to form, while other connections are pruned away.

Activity-dependent processes continue throughout life, at least in many parts of the brain, and appear to provide the basis of both explicit and implicit learning. They have been the subject of intense scrutiny in neuroscience. Donald Hebb, the mid-twentieth century neuropsychologist, proposed that if one neuron participates in firing another, the connection from the first to the second will be strengthened (Hebb 1949). Hebb's idea has been encapsulated in the phrase 'cells that fire together wire together.' While there is no direct proof that this is a principle basis of learning in the brain, the idea has received a great deal of experimental support in experiments that have been

carried out in slices of brain tissue (see *Neural Plasticity*). It should be understood that there may also be plasticity at the level of the whole neuron (in some specialized brain areas, neurons are continually created and incorporated into circuits while others are continually being lost). There is likely also to be some plasticity at the level of the branches of axons and/or dendrites, which provide the scaffolding underlying the formation and loss of synaptic connections.

2. *System-level Organization: The Macrostructure of Cognition in the Brain*

2.1 *Specialization of Brain Regions*

A central and important fact about the organization of cognition in the brain is that individual brain regions are specialized. The cerebral cortex can be partitioned conceptually into primary, secondary, and tertiary cortical zones (Luria 1966). According to this conception, the primary areas contain neurons whose responses can be largely characterized as reflecting relatively simple, local properties of inputs or outputs within a given modality, such as the presence of an oriented line segment at a particular position on the retina of the eye, the presence of acoustic energy in a particular frequency band, or the presence of a tactile stimulus at a particular point on the skin surface. Corresponding motor areas contain neurons whose responses may correspond to the activation of specific muscles or elementary movement elements. Secondary areas contain neurons whose responses represent higher-order stimulus attributes within a given modality, such as conjunctions of features, and the representations in these areas may be relatively invariant over some lower-level properties, such as position of the stimulus containing the feature on the sensory surface (Tanaka 1996). Tertiary areas are responsible for representations that transcend individual modalities, such as representations of the current task context, or representation of one's location in extra-personal space, or representation of semantic content. It should be noted that this picture is only a very crude approximation, and many so-called primary areas appear to participate in the representation of the global structure of a stimulus or response situation, and many areas that are treated as modality specific can be modulated by influences from other modalities (see below). It should also be noted that structures outside the neocortex also play very important roles in cognitive functions. Among these are the diffuse neuromodulatory systems that regulate behavioral/cognitive states such as alertness, wakefulness, and mood; and other systems in the thalamus, lymbic system, and cerebellum.

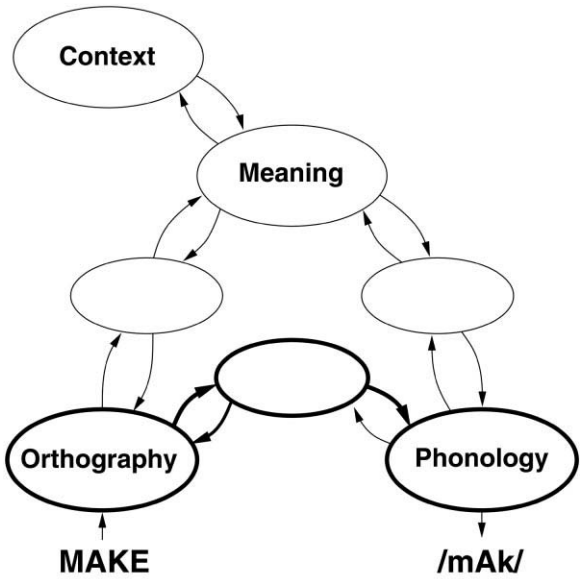

Figure 2
The interactive, distributed framework for modeling individual word reading of Seidenberg and McClelland 1989. All of the relevant processing pathways are assumed to be bidirectional

2.2 *Modular vs. Interactive Approaches to the Organization of Function*

The above provides only the starting place for the formulation of an understanding of how cognitive processes arise from neural activity. There are two contrasting views: (a) The modular approach, championed by David Marr for vision and Noam Chomsky for language, and systematized as a general approach by Fodor (1983), holds that the brain consists of many separate modules that are informationally encapsulated in that their operation is informed only by a very limited range of constraining sources of information. The modular view also holds that the principles of function are specific to each domain, and that distinct and individualized mechanisms are used to subserve each distinct function. For example, the initial assignment of the basic grammatical structure to a sentence is thought to be based only on the syntactic classification of words and their order and is thought to be governed by the operation of a system of structure sensitive rules. The module that carries out this assignment is thought to be structured specifically so that it will acquire and implement structure-sensitive rules, and to contrast in the principles that it employs internally with other modules that carry out other tasks, including other aspects of language processing, such as the assignment of meanings to the words in a sentence. In Fodor's view, there are many specialized modules (corresponding approximately to

primary and secondary cortical areas and their subcortical inputs and outputs). These are complemented by a general-purpose cognitive system that is completely open-ended in the computations that it can undertake and in the range of informational sources that it can take into consideration.

The alternative, interactive approach, has its seeds in the ideas of Luria (1966), and has been championed by Mesulam (2000) and by Rumelhart et al. (1986), and overlaps with the ideas of Damasio (1989). On this view, cognitive outcomes such as the assignment of an interpretation to a sentence arise from mutual, bidirectional interactions among neurons in populations representing different types of information. An example of a system addressing the representations and interactions involved in reading individual words aloud is shown in Fig. 2. As the figure suggests, the formation of the sound of a word from a visual input specifying its spelling arises from an interactive process involving orthographic (i.e., letter identity), semantic, phonological, and contextual information. Both the modular and the interactive view are consistent with the idea that neurons in the brain are organized into populations specialized for representing different types of information. Where they differ is in the extent and the role of bidirectional interactions among participating brain areas.

2.3 *Evidence of Interactive Processes in the Brain*

The debate between modular and interactive approaches is a long-standing one, and can be seen as the modern legacy of a history of diverse views on the localization of functions within the brain (Luria 1966). While the debate is likely to continue to evolve with additional empirical evidence, it may be worth considering a few elements of evidence that support the idea that processing may be interactive. One relevant anatomical point is the fact that connectivity within and between brain areas is generally reciprocal: when there are connections from region A to region B, there are nearly always return connections.

While there is no consensus on the function of reciprocal connections, there is some evidence that they subserve distributed, interactive computations, at least in particular cases. For example, there is evidence that interactive processes influence the activation of individual neurons in primary visual cortex (Area V1). Traditionally, individual neurons in this area have been seen as encoding the presence of segments of oriented edges at particular positions in a visual display. Recent evidence suggests, however, that primary visual cortex participates in a distributed and interactive process that contributes to the representation of global stimulus properties such as figure-ground organization. The firing of neurons in primary visual cortex is strongly affected by temporary inactivation of corresponding portions of secondary visual cortex, suggesting that reciprocal interactions between these areas shape neuronal responses in V1 (Hup et al. 2001). Although the initial response of neurons in V1 is determined by appropriate oriented line segments at a specific location, by about 80 ms their firing is heavily dependent on the global structure of the display (Lee et al. 1998). Furthermore, neurons in V1 respond to illusory contours that fall in their receptive field. The response occurs at a lag of about 80 ms, suggesting an indirect source, perhaps arising from feedback from higher cortical areas (Lee and Nguyen 2001). There is also considerable evidence of between modality interactions. For example, activity in auditory processing areas associated with speech perception is enhanced by visible speech (Callan et al. 2001). There are many corresponding examples of cross-modal influences in single neuron recording studies in animals.

3. *Methods and Approaches in Cognitive Neuroscience*

Cognitive neuroscience is a highly interdisciplinary endeavor, and draws on a wide range of research methods and approaches, each with its own history and underlying theoretical frame of reference. One important challenge for the field is to find ways of integrating the insights gained from the different methods to allow the field as a whole to converge on a common theoretical framework. Here the predominant research approaches are briefly described, and some of the prospects for integration are considered.

3.1 *Lesion and Behavior Approaches (Cognitive Neuropsychology and Behavioral Neuroscience)*

These research approaches within the field have the oldest historical roots, based as they often are in the assessment of the effects of naturally-occurring brain damage on cognitive function. A seminal case study was the report by Broca (1861) of a man with a severe disturbance of language arising from a large brain lesion in the posterior portion of the left frontal lobe. Since Broca's day, neurologists and neuropsychologists have investigated the effects of accidental or therapeutic brain lesions in humans, and many key insights have arisen from these studies (see *Agnosia*; *Amnesia*; *Aphasia*; *Dyslexia (Acquired) and Agraphia*). The subdiscipline of cognitive neuropsychology has arisen specifically around the study of the effects of brain lesions (see *Cognitive Neuropsychology, Methodology of*). A complementary has grown up around the use of brain lesions in animals carried out with specific experimental intent. This tradition is relevant to human cognitive neuroscience in view of the very close homology between many structures in the human brain and corresponding structures in the primate and

rodent brains. This work has obvious advantages in that lesions can be carefully targeted to particular brain areas to test specific hypotheses (see *Lesion and Behavior Approaches in Neuroscience*). Many key insights have emerged from this work, including the discovery of complementary processing streams in the visual system (Ungerleider and Mishkin 1982; see *Neural Plasticity*). However, the approach is not without its pitfalls, since a lesion may have unintended and unobserved effects in other brain regions; the refinement and extension of experimental lesion techniques is ongoing.

3.2 Neuronal Recording Studies

Studies relying on microelectrodes to record from neurons in the brains of behaving animals can allow researchers to study the representations that the brain uses to encode information, and the evolution of these representations over time. Several fundamental observations have, some of which have been discussed above. These studies indicate, among other things, that the brain relies on distributed representations, that neurons participate dynamically and interactively in the construction of representations of external inputs, and that the representational significance of the firing of a particular neuron can vary as a function of context. Neuronal recording studies have had a profound impact on our understanding of the nature of representations of extrapersonal space. There are neurons in the brain that encode the location of objects in extrapersonal space simultaneously in relation to many different parts of the body, including the limbs and the head (Duhamel et al. 1991, Graziano and Gross 1993) and other neurons that encode the locations of objects in relation to other objects (Olson and Gettner 1995). Furthermore, recordings from neurons in parietal cortex suggest that when we move our eyes from one location to another, we update our internal representations of the locations of important objects in space, based on where we anticipate they will be after the upcoming eye movement (Duhamel et al. 1992).

An important recent development is the ability to record from up to 100 individual neurons at a time (see *Perception and Cognition, Single-/Multi-neuronal Recording Studies of*). A key finding that has come out of this work is the confirmation in studies with rodents that the simultaneous and successive patterns of activity acquired during behavior may be reactivated in the brain during subsequent sleep (Wilson and McNaughton 1994). Such methods are in their infancy, but their potential to shed light on the moment-by-moment relations between activations of different neurons and between distributed brain representations and specific inputs and outputs makes them essential to the future of cognitive neuroscience.

3.3 Functional Brain Imaging

Cognitive neuroscience has arisen as a separate discipline in tandem with the emergence of functional brain imaging methods (PET and fMRI) as major tools for the analysis of human cognition, and it may be that the prospect of visualizing specifically human cognitive activity has been a major catalyst. First used to analyze cognitive functions by the St. Louis group (Petersen et al. 1988; see *Functional Brain Imaging*), these methods are now coming into widespread use. While these methods currently have low temporal and spatial resolution compared to neuronal recording studies, they still provide our best opportunity to explore the neural mechanisms underlying distinctly human cognitive functions.

To date the observations arising from functional imaging studies have tended to corroborate findings from other methods, and/or to explore commonalities and differences in human and animal brain organization. As one recent example, it has now been possible to visualize the alternating strips in visual cortex reflecting what are known as ocular dominance stripes. Beyond corroboration, a great deal of new information has also been provided by functional brain imaging studies. For example, in a fairly early PET study, investigators found that an area of the cerebellum became active when subjects were required to generate the action that goes along with a concrete object (e.g., the word HAMMER requires a response such as 'pound'). Subsequent investigation of an individual with damage to this region of the cerebellum indicated that the patient had considerable difficulty with the generation task, confirming the importance of this area in the task.

Brain imaging studies, like lesion studies, have often been used to try to determine the loci in the brain associated with particular cognitive functions. However, in addition to this, brain imaging has begun to reveal a great deal about the plasticity of the brain, since patterns of brain activation can change dramatically with practice (Karni et al. 1998). Imaging is also being used in search of distributed networks in the brain that contribute to particular cognitive functions. For example, Just et al. (1996) have shown that as sentences become more complex, there is an increase in neural activity in an ensemble of brain regions, including Broca's and Wernicke's area on both the left and to a lesser degree the right side of the brain. As another example, investigators have begun to use covariation in neural activity in different brain regions in an effort to determine which brain regions are influencing each other's activation (Maguire et al. 2000) in different task situations.

Imaging methods (including magneto and electroencephalography, as well as fMRI and PET) are likely to improve dramatically over time, allowing far higher spatial and temporal resolution. The potential for this to bring us closer to the goal of understanding the

details of information processing in the human brain will be discussed below.

3.4 Computational and Mathematical Modeling Approaches

While investigations relying on lesion and behavior approaches, neuronal recording studies, and functional brain imaging have provided and will continue to provide the empirical evidence on which to build our understanding of the basis of cognitive functions in the brain, these approaches, even when used in a convergent way, may still fail to provide a complete understanding of how cognitive functions emerge from underlying neural activity. This may require the use of additional tools provided by mathematical modeling and computer simulation. These approaches allow researchers to formulate possible accounts of specific processes in the form of explicit models that can be analyzed mathematically or simulated using computers to determine whether they can account for all of the relevant neural and behavioral evidence.

Three examples of cases in which computational models have already led to new thinking will be briefly considered. First, a number of computational modeling studies have shown that many aspects of the receptive field properties of neurons and their spatial organization in the brain can arise through the operation of very simple activity-dependent processes shaped by experience and a few rather simple additional constraints (Linsker 1986, Miller et al. 1989; see *Neural Development: Mechanisms and Models*). Second, models may aid in the understanding of the pattern of deficits seen in patients with brain lesions. Certain patients with an acquired dyslexic syndrome known as deep dyslexia make a striking form or error known as semantic errors; for example the patient may misread APRICOT as 'peach.' In addition, all such patients also make visual errors, for example, reading SYMPATHY and 'symphony.' Early, noncomputational accounts postulated that there must be two separate lesions, one affecting visual processing and the other affecting semantic processing. However, computational models of the reading process (Hinton and Shallice 1991; see *Cognitive Functions (Normal) and Neuropsychological Deficits, Models of*) have shown that a single lesion affecting either the visual or the semantic part of an interactive neural network will lead to errors of both types. Thus, the coexistence of these errors may be an intrinsic property of the underlying processing architecture rather than a reflection of multiple distinct lesions. A third area where computational models have shed considerable light is in the interpretation of the receptive field properties of individual neurons (Zipser and Andersen 1988). While initial interpretations were based on verbally describable features such as oriented bars or edges, such properties are not always apparent, and even when they are, a more detailed characterization may be possible in computational terms (Pouget et al. 1999). A further area of fertile research is in the use of computational models to explain and catalog the ways in which neuronal activation changes dynamically in the course of task performance (Moody et al. 1998).

4. Open Issues in Cognitive Neuroscience

Cognitive neuroscience is young, and there is a great deal of work to be done. No aspect of cognition is fully understood, and in general, the more abstract or advanced the cognitive function, the less is known about its neural basis. A few of the most important and interesting issues that remain to be addressed are considered briefly here.

4.1 How Does the Brain Learn?

There is a great deal known about the basic mechanisms of synaptic plasticity, but typically these are studied in highly reduced preparations such as brain slices. The basic processes that are studied in slices surely play a role in the shaping of neural connections in the whole, living brain, but they are also undoubtedly modulated by processes that are usually eliminated in slices. We know that attention and engagement in processing is essential for learning, and there is good reason to believe that learning is gated by various neuromodulatory mechanisms in the brain, but the details of the modulation and gating processes are only beginning to be explored.

4.2 What Makes an Experience Conscious?

Although some considerable progress has been made in characterizing the concomittants of consciousness (see *Consciousness, Neural Basis of*), there is no overall understanding of exactly what it is about the activity of the brain that gives it the attribute of consciousness. It appears likely that consciousness will not be localizable; although it may be highly dependent on specific brain structures (e.g., those that regulate sleep vs. wakefulness, etc.), it may well depend on the intact functioning of many interacting parts of the brain. Exactly why or how consciousness arises from these interactions is not at all understood.

4.3 What is the Basis for the Unique Cognitive Capacities of the Human Brain, Relative to that of Other, Simpler Organisms?

The issue of what sets humans apart from other organisms remains one of the central unresolved questions. The similarity of the human genome to that of closely related species can be taken in different

ways. It can suggest to some that a very small number of specific faculties have been added which differentiate the human from, say, the chimpanzee; or it could suggest that rather than new faculties, the human brain really differs only in the expansion and extension of structures already present to a degree in other organisms. The idea that the highest cognitive functions are emergent functions rather than localizable or locally encoded in genes remains an attractive, though elusive possibility.

5. *The Future of Cognitive Neuroscience*

Nobel laureate Eric Kandel has suggested that cognitive neuroscience will increasingly assume center stage in the neurosciences in the twenty-first century, and it has begun to make dramatic inroads into the field of cognitive psychology, where many leading investigators have redirected their research to exploit ideas and methods from neuroscience. Future research in cognitive neuroscience will address the general issues raised above as well as many other topics. What makes the future of the field so exciting is the prospect of further development of a number of important contributing methodologies. Breakthroughs in functional brain imaging and other related methods are likely to provide far greater spatial and temporal resolution of brain activity. Another, very important area of methodological advance is the ability to create genetically altered brains especially in small mammals and invertebrates, and thereby to explore the consequences of these alterations for function (see *Memory: Genetic Approaches*). These methods have already reached the point where it is possible to allow an organism to develop normally, and then induce a region-specific gene knockout, thereby providing the opportunity to investigate, for example, the effect of the alternation of synaptic plasticity in a specific part of the brain. Breakthroughs should be expected in many other areas of cognitive neuroscience as well, including neuronal recording, functional imaging, and computational modeling approaches. Together, these methods will lead to a deeper understanding of how the highest capabilities of the human mind arise from the underlying physical and chemical processes in the brain.

See also: Animal Cognition; Brain, Evolution of; Cerebral Cortex: Organization and Function; Cognitive Control (Executive Functions): Role of Prefrontal Cortex; Cognitive Neuropsychology, Methodology of; Cognitive Psychology: History; Cognitive Psychology: Overview; Cognitive Science: History; Cognitive Science: Overview; Cognitive Science: Philosophical Aspects; Comparative Neuroscience; Computational Neuroscience; Evolutionary Social Psychology; Human Cognition, Evolution of

Bibliography

Broca P 1861 Remarques sur le siege de la faculte de la parole articulee, suives d'une observation d'aphemie (perte de parole). *Bulletin de la Societe d'Anatomie* **36**: 330–57

Callan D E, Callan A M, Kroos C, Vatikiotis-Bateson E 2000 Multimodal contribution to speech perception. *Cognitive Brain Research*

Damasio A R 1989 Time-locked multiregional retroactivation: A system-level proposal for the neural substrates of recall and recognition. *Cognition* **33**: 25–62

Duhamel J-R, Colby C L, Goldberg M E 1991 Congruent representations of visual and somatosensory space in single neurons of monkey ventral intraparietal cortex (area VIP). In: Paillard J (ed.) *Brain and Space*. Oxford University Press, Oxford, UK, pp. 223–6

Duhamel J-R, Colby C L, Goldberg M E 1992 The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**: 90–2

Fodor J A 1983 *Modularity of Mind: An Essay on Faculty Psychology*. MIT Press, Cambridge, MA

Georgopoulos A P, Schwartz A B, Kettner R E 1986 Neuronal population encoding of movement direction. *Science* **233**: 1416–9

Graziano M S A, Gross C G 1993 A bimodal map of space: Somatosensory receptive fields in the macaque putamen with corresponding visual receptive fields. *Experimental Brain Research* **97**: 96–109

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Hinton G E, Shallice T 1991 Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review* **98**(1): 74–95

Hubel D H, Weisel T 1962 Receptive fields, binocular orientation and functional architecture in the cat's visual cortex. *Journal of Physiology* **166**: 106–54

Hup J M, James A C, Girard P, Lomber S G, Payne B R, Bullier J 2001 Feedback connections act on the early part of the responses in monkey visual cortex. *Journal of Neurophysiology* **85**: 134–45

Just M A, Carpenter P A, Keller T A, Eddy W F, Thulborn K R 1996 Brain activation modulated by sentence comprehension. *Science* **274**(5284): 114–6

Karni A, Meyer G, Rey-Hipolito C, Jezzard P, Adams M M, Turner R, Ungerleider L G 1998 The acquisition of skilled motor performance: Fast and slow experience-driven changes in primary motor cortex. *Proceedings of the National Academy of Science USA* **95**(3): 861–8

Lee T S, Mumford D, Romero R, Lamme V A F 1998 The role of primary visual cortex in higher level vision. *Vision Research* **38**: 2429–54

Lee T S, Nguyen M 2001 Dynamics of subjective contour formation in early visual cortex. *Proceedings of the National Academy of Science USA*

Linsker R 1986 From basic network principles to neural architecture, I: Emergence of orientation columns. *Proceedings of the National Academy of Sciences USA* **83**: 7508–12

Luria A R 1966 *Higher Cortical Functions in Man*. Basic Books, New York

Maguire E A, Mummery C J, Buchel C 2000 Patterns of hippocampal-cortical interaction dissociate temporal lobe memory subsystems. *Hippocampus* **10**(4): 475–82

Markus E J, Qin Y, Leonard B, Skaggs W E, McNaughton B L, Barnes C A 1995 Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *Journal of Neuroscience* **15**: 7079–94

Mesulam M M 2000 *Principles of Behavioral and Cognitive Neurology*. Oxford University Press, New York
Moody S L, Wise S P, di Pellegrino G, Zipser D 1998 A model that accounts for activity in primate frontal cortex during a delayed matching-to-sample task. *Journal of Neuroscience* **18**(1): 399–410
Olson C R, Gettner N 1995 Object-centered direction selectivity in the supplementary eye field of the macaque monkey. *Science* **269**: 985–8
Petersen S E, Fox P T, Posner M I, Mintun M, Raichle M E 1988 Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* **331**: 585–9
Pouget A, Deneve S, Sejnowski T J 1999 Frames of reference in hemineglect: A computational approach. *Progress in Brain Research* **121**: 81–97
Rumelhart D E, McClelland J L, the PDP Research Group 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. MIT Press, Cambridge, MA
Seidenberg M S, McClelland J L 1989 A distributed, developmental model of word recognition and naming. *Psychological Review* **96**: 523–68
Tanaka K 1996 Inferotemporal cortex and object vision. *Annual Review Neuroscience* **19**: 109–39
Ungerleider L G, Mishkin M 1982 Two cortical visual systems. In: Ingle D J, Goodale M A, Mansfield R J W (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA
Wilson M A, McNaughton B L 1994 Reactivation of hippocampal ensemble memories during sleep. *Science* **265**: 676–9
Zipser D, Andersen R A 1988 A back propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature* **331**: 679–84

J. L. McClelland

Cognitive Psychology: History

Since the beginning of experimental psychology in the nineteenth century, there had been interest in the study of higher mental processes. But something discontinuous happened in the late 1950s, something so dramatic that it is now referred to as the 'cognitive revolution,' and the view of mental processes that it spawned is called 'cognitive psychology.' What happened was that American psychologists rejected behaviorism and adopted a model of mind based on the computer. The brief history that follows (adapted in part from Hilgard (1987) and Kessel and Bevan (1985)) chronicles mainstream cognitive psychology from the onset of the cognitive revolution to the beginning of the twenty-first century.

1. Beginnings

From roughly the 1920s through the 1950s, American psychology was dominated by behaviorism. Behaviorism was concerned primarily with the learning of associations, particularly in nonhuman species, and it constrained theorizing to stimulus–response notions. The overthrow of behaviorism came not so much from ideas within psychology as from three research approaches external to the field.

1.1 Communications Research and the Information Processing Approach

During World War II, new concepts and theories were developed about signal processing and communication, and these ideas had a profound impact on psychologists active during the war years. One important work was Shannon's 1948 paper about *Information Theory*. It proposed that information was communicated by sending a signal through a sequence of stages or transformations. This suggested that human perception and memory might be conceptualized in a similar way: sensory information enters the receptors, then is fed into perceptual analyzers, whose outputs in turn are input to memory systems. This was the start of the 'information processing' approach—the idea that cognition could be understood as a flow of information within the organism, an idea that continues to dominate cognitive psychology.

Perhaps the first major theoretical effort in information processing psychology was Donald Broadbent's *Perception and Communication* (Broadbent 1958). According to Broadbent's model, information output from the perceptual system encountered a filter, which passed only information to which people were attending. Although this notion of an all-or-none filter would prove too strong (Treisman 1960), it offered a mechanistic account of selective attention, a concept that had been banished during behaviorism. Information that passed Broadbent's filter then moved on to a 'limited capacity decision channel,' a system that has some of the properties of short-term memory, and from there on to long-term memory. This last part of Broadbent's model—the transfer of information from short- to long-term memory—became the salient point of the dual-memory models developed in the 1970s.

Another aspect of Information theory that attracted psychologist's interest was a quantitative measure of information in terms of 'bits' (roughly, the logarithm to the base 2 of the number of possible alternatives). In a still widely cited paper, George Miller (1956) showed that the limits of short-term memory had little to do with bits. But along the way, Miller's and others' interest in the technical aspects of information theory and related work had fostered mathematical psychology, a subfield that was being fueled by other sources as well (e.g., Estes and Burke 1953, Luce 1959, Garner 1962). Over the years, mathematical psychology has frequently joined forces with the information

processing approach to provide precise claims about memory, attention, and related processes.

1.2 The Computer Modeling Approach

Technical developments during World War II also led to the development of digital computers. Questions soon arose about the comparability of computer and human intelligence (Turing 1950). By 1957, Alan Newell, J. C. Shaw, and Herb Simon had designed a computer program that could solve difficult logic problems, a domain previously thought to be the unique province of humans. Newell and Simon soon followed with programs that displayed general problem-solving skills much like those of humans, and argued that these programs offered detailed models of human problem solving (a classic summary is contained in Newell and Simon (1972)). This work would also help establish the field of artificial intelligence.

Early on, cross-talk developed between the computer modeling and information-processing approaches, which crystallized in the 1960 book *Plans and the Structure of Behavior* (Miller et al. 1960). The book showed that information-processing psychology could use the theoretical language of computer modeling even if it did not actually lead to computer programs. With the 'bit' having failed as a psychological unit, information processing badly needed a rigorous but rich means to represent psychological information (without such representations, what exactly was being processed in the information processing approach?). Computer modeling supplied powerful ideas about representations (as data structures), as well as about processes that operate on these structures. The resultant idea of human information processing as sequences of computational processes operating on mental representations remains the cornerstone of modern cognitive psychology (see e.g., Fodor 1975).

1.3 The Generative Linguistics Approach

A third external influence that lead to the rise of modern cognitive psychology was the development of generative grammar in linguistics by Noam Chomsky. Two of Chomsky's publications in the late 1950s had a profound effect on the nascent cognitive psychology. The first was his 1957 book *Syntactic Structures* (Chomsky 1957). It focused on the mental structures needed to represent the kind of linguistic knowledge that any competent speaker of a language must have. Chomsky argued that associations *per se*, and even phrase structure grammars, could not fully represent our knowledge of syntax (how words are organized into phrases and sentences). What had to be added was a component capable of transforming one syntactic structure into another. These proposals about transformational grammar would change the intellectual landscape of linguistics, and usher in a new psycholinguistics.

Chomsky's second publication (1959) was a review of *Verbal Behavior*, a book about language learning by the then most respected behaviorist alive, B. F. Skinner (Skinner 1957). Chomsky's review is arguably one of the most significant documents in the history of cognitive psychology. It aimed not merely to devastate Skinner's proposals about language, but to undermine behaviorism as a serious scientific approach to psychology. To some extent, it succeeded on both counts.

1.4 An Approach Intrinsic to Psychology

At least one source of modern cognitive psychology came from within the field. This approach had its roots in Gestalt psychology, and maintained its focus on the higher mental processes. A signal event in this tradition was the 1956 book *A Study of Thinking*, by Bruner, Goodnow, and Austin (Bruner et al. 1956). The work investigated how people learn new concepts and categories, and it emphasized strategies of learning rather than just associative relations. The proposals fit perfectly with the information-processing approach—indeed, they were information processing proposals—and offered still another reason to break from behaviorism.

By the early 1960s all was in place. Behaviorism was on the wane in academic departments all over America (it had never really taken strong root in Europe). Psychologists interested in the information-processing approach were moving into academia, and Harvard University went so far as to establish a Center for Cognitive Studies directed by Jerome Bruner and George Miller. The new view in psychology was information processing. It likened mind to a computer, and emphasized the representations and processes needed to give rise to activities ranging from pattern recognition, attention, categorization, memory, reasoning, decision making, problem solving, and language.

2. The Growth of Cognitive Psychology

The 1960s brought progress in many of the above-mentioned topic areas, some of which are highlighted below.

2.1 Pattern Recognition

One of the first areas to benefit from the cognitive revolution was pattern recognition, the study of how people perceive and recognize objects. The cognitive approach provided a general two-stage view of object recognition: (a) describing the input object in terms of

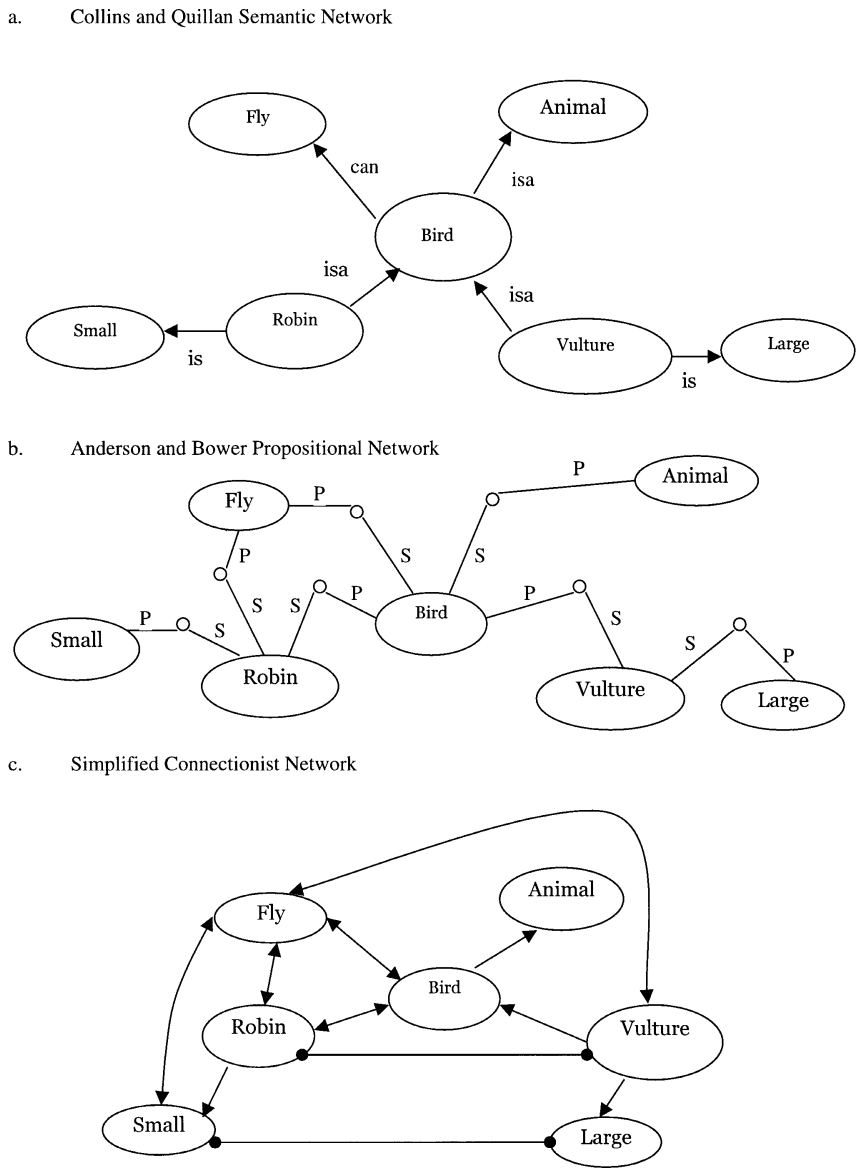

Figure 1
(a) Part of a Collins and Quillian (1969) semantic network. Circles designate concepts and lines (arrows) between circles designate relations between concepts. There are two kinds of relations: subset–superset ('Robin is a bird') and property (e.g., 'Robins can fly'). The network is strictly hierarchical, as properties are stored only at the highest level at which they apply. (b) Part of an Anderson and Bower (1973) propositional network. Circles represent concepts and lines between them labeled relations. All propositions have a subject–predicate structure, and the network is not strictly hierarchical. (c) Part of a simplified connectionist network. Circles represent concepts, or parts of concepts, lines with arrowheads depict excitatory connections, and lines with filled circles designate inhibitory connections; typically numbers are put on the lines indicate the strength of the connections. The network is not strictly hierarchical, and is more interconnected than the preceding networks

relatively primitive features (e.g., 'it has two diagonal lines and one horizontal line connecting them'); and (b) matching this object description to stored object descriptions in visual memory, and selecting the best match as the identity of the input object ('this description best matches the letter A'). This two-stage

view was not entirely new to psychology, but expressing it in information-processing terms allowed one to connect empirical studies of object perception to computer models of the process. The psychologist Ulrich Neisser (1964) used a computer model of pattern recognition (Selfridge 1959) to direct his empirical studies and provided dramatic evidence that an object could be matched to multiple visual memories in parallel.

Other research indicated that the processing underlying object perception could persist after the stimulus was removed. For this to happen, there had to be a visual memory of the stimulus. Evidence for such an 'iconic' memory was supplied by Sperling in classic experiments in 1960 (Sperling 1960). Evidence for a comparable brief auditory memory was soon provided as well (e.g., Crowder and Morton 1969). Much of the work on object recognition and sensory memories was integrated in Neisser's influential 1967 book *Cognitive Psychology* (Neisser 1967). The book served as the first comprehensive statement of existing research in cognitive psychology, and it gave the new field its name.

2.2 Memory Models and Findings

Broadbent's model of attention and memory stimulated the formulation of rival models in the 1960s. These models assumed that short-term memory (STM) and long-term memory (LTM) were qualitatively different structures, with information first entering STM and then being transferred to LTM (e.g., Waugh and Norman 1965). The Atkinson and Shiffrin (1968) model proved particularly influential. With its emphases on information flowing between memory stores, control processes regulating that flow, and mathematical descriptions of these processes, the model was a quintessential example of the information-processing approach. The model was related to various findings about memory. For example, when people have to recall a long list of words they do best on the first words presented, a 'primacy' effect, and on the last few words presented, a 'recency' effect. Various experiments indicated that the recency effect reflected retrieval from STM, whereas the primacy effect reflected enhanced retrieval from LTM due to greater rehearsal for the first items presented (e.g., Murdock 1962, Glanzer and Cunitz 1966). At the time these results were seen as very supportive of dual-memory models (although alternative interpretations would soon be proposed—particularly by Craik and Lockhart 1972).

Progress during this period also involved empirically determining the characteristics of encoding, storage, and retrieval processes in STM and LTM. The results indicated that verbal material was encoded and stored in a phonologic code for STM, but a more meaning-based code for LTM (Conrad 1964, Kintsch and Buschke 1969). Other classic studies demonstrated that forgetting in STM reflected a loss of information from storage due to either decay or interference (e.g., Wickelgren 1965), whereas some apparent losses of information in LTM often reflected a temporary failure in retrieval, (Tulving and Pearlstone 1966).To a large extent, these findings have held up during over 30 years of research, although many of the findings would now be seen as more limited in scope (e.g., the findings about STM are now seen as reflecting only one component of working memory, e.g., Baddeley (1986), and the findings about LTM are seen as characterizing only one of several LTM systems, e.g., Schacter (1987)).

One of the most important innovations of 1960s research was the emphasis on reaction time as a dependent measure. Because the focus was on the flow of information, it made sense to characterize various processes by their temporal extent. In a seminal paper in 1966, Saul Sternberg reported (Sternberg 1966) that the time to retrieve an item from STM increased linearly with the number of items in store, suggesting that retrieval was based on a rapid scan of STM. Sternberg (1969) gave latency measures another boost when he developed the 'additive factors' method, which, given assumptions about serial processing, allowed one to attribute changes in reaction times to specific processing stages involved in the task (e.g., a decrease in the perceptibility of information affected the encoding of information into STM but not its storage and retrieval). These advances in 'mental chronometry' quickly spread to areas other than memory (e.g., Fitts and Posner 1967, see also Schneider and Shiffrin 1977).

2.3 The New Psycholinguistics

Beginning in the early 1960s there was great interest in determining the psychological reality of Chomsky's theories of language (these theories had been formulated with ideal listeners and speakers in mind). Some of these linguistically inspired experiments presented sentences in perception and memory paradigms, and showed that sentences deemed more syntactically complex by transformational grammar were harder to perceive or store (Miller 1962). Subtler experiments tried to show that syntactic units, like phrases, functioned as units in perception, STM, and LTM (Fodor et al. (1974) is the classic review). While many of these results are no longer seen as critical, this research effort created a new subfield of cognitive psychology, a psycholinguistics that demanded sophistication in modern linguistic theory.

Not all psycholinguistic studies focused on syntax. Some dealt with semantics, particularly the representation of the meanings of words, and a few of these studies made use of the newly developed mental chronometry. One experiment that proved seminal was reported by Collins and Quillian (1969). Partici-

pants were asked simple questions about the meaning of a word, such as 'Is a robin a bird,' and 'Is a robin an animal?'; the greater the categorical difference between the two terms in a question, the longer it took to answer. These results were taken to support a model of semantic knowledge in which meanings were organized in a hierarchical network, e.g., the concept 'robin' is directly connected to the concept 'bird,' which in turn is directly connected to the concept 'animal,' and information can flow from 'robin' to 'animal' only by going through 'bird' (see the top of Fig. 1). Models like this were to proliferate in the next stage of cognitive psychology.

3. *The Rise of Cognitive Science*

3.1 *Memory and Language*

Early in the 1970s the fields of memory and language began to intersect. In 1973 John Anderson and Gordon Bower published *Human Associative Memory* (Anderson and Bower 1973), which presented a model of memory for linguistic materials. The model combined information processing with recent developments in linguistics and artificial intelligence (AI), thereby linking the three major research directions that led to the cognitive revolution. The model used networks similar to that considered above to represent semantic knowledge, and used memory-search processes to interrogate these networks (see the middle of Fig. 1). The Anderson and Bower book was quickly followed by other large-scale theoretical efforts that combined information processing, modern linguistics, and computer models. These efforts included Kintsch (1974), which focused on memory for paragraphs rather than sentences, and Norman, Rumelhart, and the LNR Research Group (1975), Anderson (1976), and Schank and Abelson (1977), which took a more computer-science perspective and focused on stories and other large linguistic units.

As psychologists became aware of related developments in linguistics and artificial intelligence, so researchers in the latter disciplines become aware of pertinent work in psychology. Thus evolved the interdisciplinary movement called 'cognitive science.' In addition to psychology, AI, and linguistics, the fields of cultural anthropology and philosophy of mind also became involved. The movement eventuated in numerous interdisciplinary collaborations (e.g., Rumelhart et al. 1986), as well as in individual psychologists becoming more interdisciplinary.

3.2 *Representational Issues*

In the 1970s and early 1980s, cognitive science was much concerned with issues about mental representations. Whereas the memory-for-language models described earlier had assumed representations that were language-like, or propositional, other researchers argued that representations could also be imaginal, like a visual image. Shepard and Cooper (1972) provided evidence that people could mentally rotate their representations of objects, and Kosslyn (1980) surveyed numerous phenomena that further implicated visual imagery. In keeping with the interdisciplinariness of cognitive science, AI researchers and philosophers entered the debate about propositional versus imaginal representations (e.g., Block 1981, Pylyshyn 1981). In addition to questions about the modality of representations, there were concerns about the structure of representations. While it had long been assumed that propositional representations of objects were like definitions, researchers now proposed the representations were *prototypes* of the objects, fitting some examples better than others (Tversky 1977, Mervis and Rosch 1981, Smith and Medin 1981). Again the issues sparked interest in disciplines other than psychology (e.g., Lakoff 1987).

The cognitive science movement affected most areas of cognitive psychology, ranging from object recognition (Marr 1982) to reasoning (e.g., Johnson-Laird 1983) to expertise in problem solving (e.g., Chase and Simon 1973). The movement continues to be influential and increasingly focuses on computational models of cognition. What has changed since its inception in the 1970s is the kind of computational model in favor.

4. *Newer Directions: Connectionism and Cognitive Neuroscience*

4.1 *Connectionist Modeling*

The computer models that dominated cognitive psychology from its inception used complex symbols as representations, and processed these representations in a rule-based fashion (for example, in a model of object recognition, the representation for a frog might consist of a conjunction of complex properties, and the rule for recognition might look something like 'If it's green, small, and croaks, it's a frog'). Starting in the early 1980s, an alternative kind of cognitive model started to attract interest, namely 'connectionist' (or 'parallel distributed processing') models. These proposals have the form of neural networks, consisting of nodes (representations) that are densely interconnected, with the connections varying in strength (see the bottom of Fig. 1).

In 1981 Hinton and Anderson published a book surveying then existent connectionist models (Hinton and Anderson 1981), and in the same year McClelland and Rumelhart (1981) presented a connectionist model of word recognition that explained a wide variety of experimental results. The floodgates had been opened,

and connectionist models of perception, memory, and language proliferated, to the point where they now dominate computational approaches to cognition. Why the great appeal? One frequently cited reason is *neurological plausibility*: the models are clearly closer to brain function than are traditional rule-based models. A second reason is that connectionist models permit *parallel constraint satisfaction*: different sources of activation can converge simultaneously on the same representations or response. A third reason is that connectionist models manifest *graceful degradation*: when the model is damaged, performance degrades slowly, much as is found in human neurological disorders (Rumelhart et al. 1986).

From a historical viewpoint, there is an ironic aspect about the ascendancy of connectionist models. Such models return to the pure associationism that characterized behaviorism. While connectionist models hardly fit with all behaviorist dictums—their representations are not restricted to stimuli and responses, and they routinely assume massively parallel processing—still their reliance on associations runs counter to Chomsky's arguments that sheer associationism cannot explain language (Chomsky 1957, 1959). This issue has formed part of the basis for critiques of connectionist models (e.g., Fodor and Pylyshyn 1988). Newell (1990), one of the founders of traditional computational models, suggested a plausible resolution: lower-level cognitive processes like object recognition may be well modeled by connectionist models, but higher-level cognitive processes like reasoning and language may require traditional symbolic modeling.

4.2 Cognitive Neuroscience

The other major new direction in cognitive psychology is the growing interest in the neural bases of cognition, a movement referred to as 'cognitive neuroscience.' There had been little interest in biological work in the research that brought about the cognitive revolution. That early work was as much concerned with fighting behaviorism as it was with advancing cognitive psychology, and consequently much of the research focused on higher-level processes and was completely removed from anything going on in the neurobiology of its day. Subsequent generations of cognitive psychologists solidified their commitments to a purely cognitive level of analyses, by arguing that the distinction between cognitive and neural levels of analyses was analogous to that between computer software and hardware, and that cognitive psychology (and cognitive science) was concerned primarily with the software. Since the early 1990s, views about the importance of neural analyses have changed dramatically. There is a growing consensus that the standard information processing analyses of cognition can be substantially enlightened by knowing how cognition is implemented in the brain.

While many factors may have been responsible for the change in view, three will be mentioned here. First, the rise of connectionist models that were loosely inspired by brain function had the side effect of increasing interest in what was known in detail about brain function. Indeed, these two recent movements have become increasingly intertwined as connectionist models have increasingly incorporated findings from cognitive neuroscience. Second, in the 1970s and 1980s there were breakthroughs in systems-level neuroscience that had implications for mainstream cognitive psychology. As one example, research with neurological patients as well as with nonhuman species established that structures in the medial temporal lobe were essential for the formation of memories that could serve as the basis for recall ('explicit memory'), but not for other kinds of 'implicit memories' (e.g., Schacter 1987, Squire 1992). As another example, research with nonhuman primates showed that two distinct systems were involved in the early stages of visual perception: a 'where' system that is responsible for spatial localization, and a 'what' system that is responsible for recognition of the object (e.g., Ungerleider and Mishkin 1982). Both of these discoveries had major implications for the cognitive level of analyses, e.g., computational models of object recognition or memory would do well to divide the labor into separate subsystems.

The third factor responsible for the rise of cognitive neuroscience is methodological: the development of neuroimaging techniques that produce maps of neural activity while the brain is performing some cognitive task. One major technique is positron emission tomography (PET). In 1988, Michael Posner, Marcus Raichle, and colleagues used PET in a groundbreaking experiment to localize neurally different subprocesses of reading—PET images obtained while participants looked at words showed activated regions in the posterior temporal cortex; images obtained while participants read the words revealed additional regions activated in the motor cortex; and images obtained while participants generated constrained associations to the words revealed still additional regions in frontal areas (Posner et al. 1988). PET analyses of object recognition, memory, and language soon followed. A more recent technique is functional magnetic resonance imaging (fMRI). It is now being used to study virtually every domain of human cognition.

5. Conclusion

This article has given short shrift to important contributions that tend to fall off the mainstream of cognitive psychology. One such case is the work done by Dan Kahneman and Amos Tversky (e.g., Kahneman and Tversky 1973, Tversky and Kahneman 1983) on the use of heuristics in decision making, which can result

in deviations from rational behavior. Another example is the cognitively inspired study of memory and language deficits in neurological patients (Shallice (1988) provides a review). There are other cases like these which deserve a prominent place in a fuller history of cognitive psychology.

Bibliography

Anderson J R 1976 *Language, Memory, and Thought*. Erlbaum, Hillsdale, NJ

Anderson J R, Bower G H 1973 *Human Associative Memory*. Winston, Washington, DC

Atkinson R C, Shiffrin R M 1968 Human memory: a proposed system and its control processes. In: Spence K, Spence J (eds.) *The Psychology of Learning and Motivation*. Academic Press, San Diego, CA, Vol. 2

Baddeley A D 1986 *Working Memory*. Oxford University Press, Oxford, UK

Block N J 1981 *Imagery*. MIT Press, Cambridge, MA

Broadbent D E 1958 *Perception and Communication*. Pergamon, New York

Bruner J S, Goodnow J J, Austin G A 1956 *A Study of Thinking*. Wiley, New York

Chase W G, Simon H A 1973 Perception in chess. *Cognitive Psychology* **4**: 55–81

Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands

Chomsky N 1959 Review of B F Skinner's *Verbal Behavior*. *Language* **35**: 26–58

Collins A M, Quillian R M 1969 Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* **8**: 240–7

Conrad R 1964 Acoustic confusions in immediate memory. *British Journal of Psychology* **55**: 75–84

Craik F I, Lockhart R S 1972 Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior* **11**: 671–84

Crowder R G, Morton J 1969 Precategorical acoustic storage (PAS). *Perception and Psychophysics*. **5**: 365–73

Estes W K, Burke C J 1953 A theory of stimulus variability in learning. *Psychological Review* **60**: 276–86

Fitts P M, Posner M I 1967 *Human Performance*. Brooks/Cole, Belmont, CA

Fodor J A 1975 *The Language of Thought*. Crowell, New York

Fodor J A, Bever T G, Garrett M F 1974 *The Psychology of Language: an Introduction to Psycholinguistics and Generative Grammar*. McGraw-Hill, New York

Fodor J A, Pylyshyn Z W 1988 Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71

Garner W R 1962 *Uncertainty and Structure as Psychological Concepts*. Wiley, New York

Glanzer M, Cunitz A R 1966 Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior* **5**: 351–60

Hilgard E R 1987 *Psychology in America: a Historical Survey*. Harcourt, Brace Jovanovich, San Diego, CA

Hinton G E, Anderson J A 1981 *Parallel Models of Associative Memory*. Erlbaum, Hillsdale, NJ

Johnson-Laird P N 1983 *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press, Cambridge, MA

Kahneman D, Tversky A 1973 Psychology of prediction. *Psychological Review* **80**: 237–51

Kessel F S, Bevan W 1985 Notes toward a history of cognitive psychology. In: Buxton C E (ed.) *Points of View in the Modern History of Psychology*. Academic Press, New York

Kintsch W 1974 *The Representation of Meaning in Memory*. Erlbaum, Hillsdale, NJ

Kintsch W, Buschke H 1969 Homophones and synonyms in short-term memory. *Journal of Experimental Psychology* **80**: 403–7

Kosslyn S M 1980 *Image and Mind*. Harvard University Press, Cambridge, MA

Lakoff G 1987 *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, Chicago

Luce R D 1959 *Individual Choice Behavior: a Theoretical Analysis*. Wiley, New York

Marr D 1982 *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco

McClelland J L, Rumelhart D E 1981 An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review* **88**: 375–407

Miller G A 1956 The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* **63**: 81–97

Miller G A et al. 1960 *Plans and the Structure of Behavior*. Holt, New York

Miller G A 1962 Some psychological studies of grammar. *American Psychologist* **17**: 748–62

Murdock Jr B B 1962 The serial position effect in the recall. *Journal of Experimental Psychology* **64**: 482–8

Neisser U 1964 Visual Search. *Scientific American* **210**: 94–102

Neisser U 1967 *Cognitive Psychology*. Appleton-Century-Crofts, New York

Neisser U 1968 The processes of vision. *Scientific American* **219**: 204–14

Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

Newell A, Shaw J C, Simon H A 1958 Elements of a theory of human problem solving. *Psychological Review* **65**: 151–66

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Norman D A, Rumelhart D E, and the LNR Research Group1975 *Explorations in Cognition*. Freeman, San Francisco

Posner M I, Petersen S E, Fox P T, Raichle M E 1988 Localization of cognitive operations in the human brain. *Science* **240**: 1627–31

Pylyshyn Z W 1981 The imagery debate: analogue media versus tacit knowledge. *Psychological Review* **88**: 16–45

Rosch E, Mervis C B 1981 Categorization of natural objects. *Annual Review of Psychology* **32**: 89–115

Rumelhart D E, Hinton G E, Williams R J 1986 Learning representations by back-propagating errors. *Nature* **323**: 533–6

Rumelhart D E, McClelland J L, PDP Research Group 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA

Schacter D L 1987 Implicit memory: history and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **13**: 501–18

Schank R C, Abelson R P 1977 *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. Erlbaum, Hillsdale, NJ

Schneider W, Shiffrin R M 1977 Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* **84**: 1–66
Selfridge O G 1959 Pandemonium: A Paradigm for Learning. The Mechanization of Thought Processes. Her Majesty's Stationery Office, London
Selfridge O G, Neisser U 1960 Pattern recognition by machine. *Scientific American* **203**: 60–8
Shallice T 1988 *From Neuropsychology to Mental Structure*. Cambridge University Press, Cambridge, UK
Shannon C E 1948 A mathematical theory of communication. *Bells Systems Technical Journal* **27**: 379–423, 623–56
Shannon C E, Weaver W 1949 *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL
Shepard R N, Cooper L A 1972 *Mental Images and their Transformations*. MIT Press, Cambridge, MA
Skinner B F 1957 *Verbal Behavior*. Appleton-Century-Crofts, New York
Smith E E, Medin D L 1981 *Categories and Concepts*. Harvard University Press, Cambridge, MA
Sperling G 1960 The information available in brief visual presentations. *Psychological Monographs* **77**(3, no. 478)
Squire L R 1992 Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* **99**(2): 195–231
Sternberg S 1966 High-speed scanning in human memory. *Science* **153**: 652–4
Sternberg S 1969 The discovery of processing stages: extensions of Donders' method. *Acta Psychologica* **30**: 276–315
Treisman A M 1960 Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology* **12**: 242–8
Tulving E, Pearlstone Z 1966 Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior* **5**: 381–91
Turing A M 1950 Computing machinery and intelligence. *Mind* **59**: 433–60
Tversky A 1977 Features of similarity. *Psychological Review* **84**: 327–52
Tversky A, Kahneman D 1983 Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review* **90**: 293–315
Ungerleider L G, Mishkin M 1982 Two cortical visual systems. In: Engle D J, Goodale M A, Mansfield R J (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA, pp. 549–86
Waugh N C, Norman D A 1965 Primary memory. *Psychological Review* **72**: 89–104
Wickelgren W A 1965 Acoustic similarity and intrusion errors in short-term memory. *Journal of Experimental Psychology* **70**: 102–8

E. E. Smith

Cognitive Psychology: Overview

1. Meanings of the Term 'Cognitive Psychology'

Cognitive Psychology has at least three different meanings. First, the term refers to 'a simple *collection of topic areas*,' that is, of behaviorally observable or theoretically proposed phenomena that are studied within the boundaries of the field of Cognitive Psychology. Second, the term alludes to the fact that cognitive psychologists attempt to explain intelligent human behavior by reference to a *cognitive system* that intervenes between environmental input and behavior. The second meaning of Cognitive Psychology thus refers to a set of assumptions governing the operations of the proposed cognitive system. Third, Cognitive Psychology means a particular *methodological* approach to studying, that is, to empirically addressing potential explanations of human behavior. The two latter meanings of Cognitive Psychology are discussed in some depth below, after a very brief consideration of the scope of modern Cognitive Psychology and its historical roots.

2. The Scope of Cognitive Psychology

At present, Cognitive Psychology is a broad field concerned with many different topic areas, such as, for instance, human memory, perception, attention, pattern recognition, consciousness, neuroscience, representation of knowledge, cognitive development, language, and thinking. The common denominator of these phenomena appears to be that all of the phenomena reflect the operation of 'intelligence' in one way or another, at least if intelligence is broadly defined as *skill of an individual to act purposefully, think rationally, and interact efficiently with the environment*. Thus, at a general level, Cognitive Psychology is concerned with explaining the structure and mental operations of intelligence as well as its behavioral manifestations.

3. Historical Roots of Cognitive Psychology

3.1 The Term Cognitive Psychology

The term Cognitive Psychology (latin: *cognoscere*; greek: *gignoskein* = to know, perceive) is rather young. Although we do find the related term 'cognition' mentioned occasionally in the psychologies of the late nineteenth and early twentieth century (e.g., James, Spence, Wundt) where it denoted the basic elements of consciousness and their combinations, the present meanings of the term cognitive psychology owe little to the early theoretical and philosophical considerations of the human mind. Rather, the current modern meanings of the term owe much more to (a) the fact that the study of cognition emerged in opposition to the prevailing behavioristic view in the 1940s and 1950s that was trying to explain human behavior primarily in terms of its antecedent environmental conditions, and to (b) the availability of both new theoretical concepts (e.g., information theory [Shannon], cybernetics [Wiener], systems theory

[von Bertalanffy]) and practical computing machines that offered new insights into the potential nature of the mental device intervening between the outside world and human behavior.

Accordingly, the history of Cognitive Psychology can be parsed into four distinct periods: philosophical, early experimental, the cognitive revolution, and modern cognitive psychology. Early philosophy (ancient Egyptians, Greek philosophers, British empiricists) provided a context for understanding the mind and its processes (i.e., associations), and identified many of the major theoretical issues that were later studied empirically (e.g., how does perception work? how are concepts represented?). Early experimental work began, at the latest, during the middle part of the nineteenth century, when Fechner empirically studied the relation between stimulus properties (e.g., weight) and accompanying internal sensations, and when, in 1879, Wilhelm Wundt founded the first department of psychology in Leipzig, Germany. The early experimental phase of Cognitive Psychology was in full swing when, during the early 1900s, Donders and Cattell conducted perception experiments on imageless thought, and Frederic Bartlett investigated memory from a naturalistic viewpoint.

4. *Modern Cognitive Psychology*

In the mid-1950s and early 1960s, Cognitive Psychology experienced a renaissance. *Cognitive Psychology*, a textbook that systematized the re-emerged science, was written by Neisser and was published in the United States (1967). Neisser's book was central to the solidification of Cognitive Psychology as it gave a label to the field and defined the topical areas. Neisser used the computer metaphor to capture the selection, storage, reception, and manipulation of information in the human cognitive system. In 1966, Hilgard and Boweradded a chapter to their book *Theories of Learning* in which the idea of using computer programs to serve as models of theories of cognition was developed.

The 1970s saw the emergence of professional journals devoted to Cognitive Psychology, such as *Cognitive Psychology*, *Cognition*, *Memory & Cognition*, and a series of symposia volumes, including the *Loyola Symposium on Cognition* and the *Carnegie–Mellon Series*. Cognitive laboratories were built, symposia and conferences appeared at both national and international levels, courses in Cognitive Psychology were added to the curricula, textbooks written on the topic, and professors of Cognitive Psychology hired.

In the 1980s and 1990s, serious efforts to discover the neural components that are linked to specific cognitive constructs began, and the field underwent transformations due to major changes in computer technology and brain science. As a result, Cognitive Psychology converged with computer science and neuroscience to create a new discipline called 'Cognitive Science.'

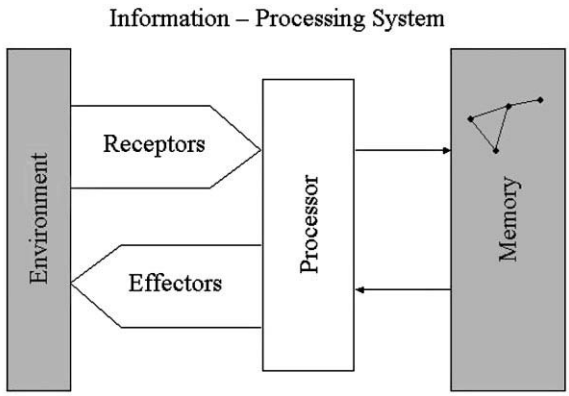

Figure 1
A basic information-processing system

Even more recently, with the advent of new ways to see the brain at work (e.g., functional magnetic resonance imaging fMRI, positron emission tomography PET, electroencephalogram EEG), cognitive psychologists have expanded their operations to neuroscience, in the hope of being able to empirically localize the components of the brain that are involved in specific operations of the cognitive system.

5. *The Properties of the Cognitive System: The Computer Metaphor*

Currently, the dominant metaphor underlying theoretical and empirical research in Cognitive Psychology is the computer metaphor. According to the computer metaphor, the cognitive system of humans, that is, the device intervening between environmental input and behavior, can be understood best in analogy to an information-processing framework. A basic information-processing system (see Fig. 1) contains two basic components, a memory and a processor, that interact with each other. In addition, the processor interacts with the environment through receptors and effectors. Newell and Simon argue that any physical-symbol system, such as an information-processing system, has the necessary and sufficient means to generate intelligent action (physical symbol systems hypothesis).

In laymen's terms, the information-processing framework formally described by Newell and Simon (1972, pp. 20–21) can be said to be based on seven basic ideas (Lachman et al. 1979): (a) humans are viewed as autonomous, intentional beings who interact with the external world; (b) the cognitive system is a general-purpose, symbol-processing system; (c) there exists a fundamental distinction between processes and data (i.e., memories). Data are acted on by processes that

manipulate and transform data; (d) cognitive processes take time, such that predictions about response times can be made if it is assumed that processes occur in sequence and have specifiable complexity; (e) the cognitive system is a limited-capacity processor that has both structural and resource limitations; (f) the cognitive system depends on, but is not entirely constrained by, a neurological substrate; (g) the goal of psychological research is to specify the processes and representations underlying intelligent performance on cognitive tasks.

6. *Current Themes of the Computer Metaphor*

The idea that a human cognitive system can be viewed as an information-processing device has had a dramatic impact on both theoretical and empirical research on the functioning of the human mind. A few select current themes in Cognitive Psychology reflecting this approach are the following.

6.1 *Data-driven or Conceptually Driven Processes?*

A data-driven mental process is one that relies almost exclusively on the 'data,' that is, on the stimulus information being presented in the environment. Whereas data-driven processes are assisted very little by already known information, conceptually driven processes are those that rely heavily on such information. Thus, a conceptually driven process uses the information already in memory, and whatever expectations are present in the situation, to perform a task; data-driven processes use only the stimulus information.

The distinction between data-driven and conceptually driven processes has been studied intensely in, among others, the area of pattern recognition. Models of perception attempt to explain, in large part, how patterns are recognized. Early models assumed that this process was primarily data-driven. However, the results of more recent research suggests that pattern recognition is also influenced by top-down conceptual processes.

6.2 *Attention*

Attention is often assumed to be a critical mental resource that is necessary for the operation of any mental process. Most theories that discuss attention assume that it is a limited mental resource and that the amount of attention that is available determines how many separate processes can be simultaneously performed.

One of the perhaps most interesting problems surrounding the role of attention has been concerned with the question of whether *attentional selection*

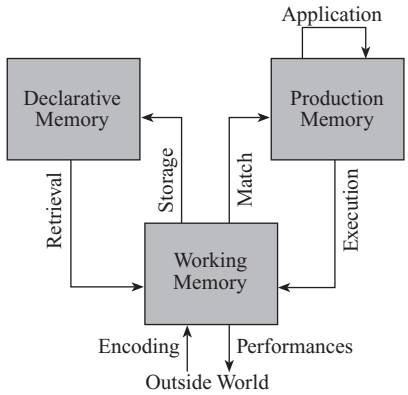

Figure 2
The basic architecture of ACT* (Anderson 1983)

occurs early or late within the cognitive system. Much of the early research on this topic focused on the extent to which unattended stimuli are processed. Early selection models hold that selection occurs at a relatively early level in the cognitive system, that is, before meaning has been extracted. Initial support for this notion came from dichotic listening tasks in which listeners had to verbally repeat information presented to one ear while different (or the same) information was simultaneously presented to the other ear. Results suggested that little of the information presented to the unattended ear was noticed. However, it was soon realized that attentional selection was not an all-or-none phenomenon. Thus, at least under some circumstances unattended information seems to reach higher levels of the cognitive system.

The topic of attentional selection has received rather widespread interest not only in studies with normal adults but also with neuropsychological samples because in some patient populations (e.g., attention-deficit disorder, schizophrenia), the attentional selection system appears to be impaired.

6.3 *Separate or Unitary Memory Systems?*

The debate over memory types has a long history in Cognitive Psychology. For example, in the 1960s Atkinson and Shiffrin introduced an information-processing model containing sensory, short-term, and long-term memory stores. Craik and Lockheart, a few years later, advanced a unitary view of memory. More recently, distinctions have been made between declarative/explicit (intentionally recollecting earlier experiences) and procedural/implicit (nonintentional influences from earlier exposure) memory systems, that, in turn, have been challenged by the argument that many apparent dissociations can be accommodated when one considers the match between encoding operations and retrieval operations (transfer-appropriate process-

ing). At present, it is unclear how many distinct memory systems exist in the human cognitive system. However, recent studies with amnesic patients and brain-imaging studies seem to suggest that memory may not be unitary.

6.4 The Nature of the Cognitive Architecture

Cognitive architectures specify the permanent properties of the human cognitive system, akin to the hardware of a modern computer. Recent proposals sketch the human cognitive architecture in a much more fine-grained and detailed manner than was apparent in Newell and Simon's earlier proposal of a basic information-processing device.

Fig. 2 depicts the basics of a cognitive architecture, that has been proposed by Anderson (1983), called ACT* (*a*daptive *c*ontrol of *t*hought). In Anderson's architecture, the existence of a long-term declarative memory for basic facts that are connected to each other in a semantic net is assumed. In addition, Anderson proposes a second long-term, procedural, memory that consists of productions. Each production has a set of conditions that test elements of working memory and a set of actions that create new structures in working memory. Strengths are associated with each long-term memory element (both network nodes and productions) as a function of its use. Working memory itself is activation based and contains the activated portion of declarative memory plus declarative structures generated by production firings and perception.

Activation spreads automatically (as a function of node strength) through working memory and from there to other connected nodes in declarative memory. Activation, along with production strength, determines how fast the matching of production proceeds. Selection of productions to fire is a competitive process between productions matching the same data.

New productions are created by compiling the effects of a sequence of production firings and retrievals from declarative memory. Whenever a new element is created in working memory, there is a fixed probability that it will be stored in declarative memory.

Cognitive architectures like ACT* are important not only in their own rights, that is, because they are theories of the structural and processing components of the human cognitive system, but also because they set important constraints for more specific theories that address more local characteristics of the cognitive system.

7. Recent Challenges to the Computer Metaphor

In recent years, an increasing number of theorists have come to reject the view that the human cognitive system operates like a computer. Two new metaphors have been proposed. First, some theorists have argued that the human cognitive system might be better understood in terms of a *brain metaphor*, assuming that cognitive systems consist of elementary, neuron-like units that are connected and produce behavior as a whole. Second, at least some areas within Cognitive Psychology have adopted an *ecological*, or *context metaphor*, arguing that cognitive systems need to be understood in terms of organism-environment relations.

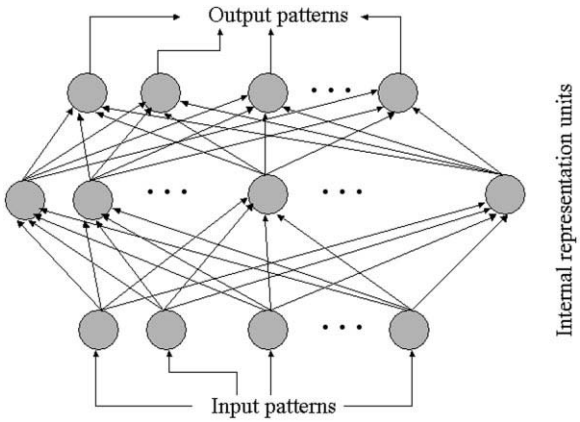

Figure 3
A basic connectionist model

7.1 The Brain Metaphor

According to the brain metaphor, human cognition is best understood in terms of the properties of the brain. The brain metaphor, and more specifically, so-called neural-like, connectionist networks as computational implementations of how our brain might work, have become highly popular in recent years and have seriously challenged the premier status of the computer metaphor when it comes to theorizing about the nature of human cognition.

Connectionist networks, neural networks, or parallel distributed processing models as they are variously called, differ from theories based on the computer metaphor in various respects. For example, in theories adhering to the computer metaphor, all processes assumed to underlie human behavior need to be explicitly described. Connectionist networks, on the other hand, can to some extent 'program' themselves in that they can learn to produce specific outputs when certain inputs are given to them. Furthermore, connectionist theorists often reject the use of explicit rules and symbols and use distributed representations, in which concepts are characterized as patterns of activation in a network.

Current connectionist networks typically have the following characteristics:

(a) the network consists of elementary or neuron-like units or nodes that are connected to each other such that a single unit has many links to other units;

(b) units affect other units by exciting or inhibiting them;

(c) the units usually takes the weighted sum of all input links and produces a single output to another unit if the weighted sum exceeds some threshold value;

(d) the network as a whole is characterized by the properties of its units, by the manner in which the units are connected to each other, and by the algorithms or rules used to change the strength of connections among units;

(e) networks can have different structures of layers; they can have a layer of input units, intermediate layers (of so-called 'hidden units'), and a layer of output units;

(f) a representation of a concept is stored in a distributed manner by a pattern of activation throughout the network;

(g) the same network can store many different patterns without them necessarily interfering with each other;

(h) one algorithm or rule used in networks to permit learning to occur is known as backward propagation of errors.

How do individual units act when activation impinges on them? Any given unit can be connected to several other units (see Fig. 3). Each of these other units can send an excitatory or an inhibitory signal to the first unit. This unit generally takes a weighted sum of all these inputs. If the sum exceeds some threshold, then the unit produces an output that may feed into other units.

This type of network can model cognitive behavior without recourse to the kinds of explicit rules found in the domain of the computer metaphor. The networks do so by associating various inputs with certain outputs, and by storing patterns of activation in the network. The networks typically make use of several layers to deal with complex behavior. One layer consists of input units that encode a stimulus as a pattern of activation. Another layer is an output layer that produces some response, again as a pattern of activation. When the network has learned to produce a particular response at the output layer following the presentation of a particular stimulus at the input layer, it can exhibit behavior that very much looks like a rule being applied.

One of the most critical aspects of connectionist networks is the learning rule or algorithm used to form patterns of activation. One algorithm that has been used to permit connectionist networks to learn is called backward propagation. At the beginning of a learning episode, the network is set up with random connection weights among the units. During the early stages of learning, when the input pattern has been presented, the output units often produce a response that is not the required output pattern. Backword propagation compares this imperfect pattern with the known required response, noticing the differences. It then back-propagates activation through the network such that the units are adjusted in such a way that they will tend to produce the required pattern on the next learning cycle. This process is repeated with a particular stimulus pattern until the network produces the required response pattern.

Networks have been used to produce very interesting results. For example, Sejnowski and Rosenberg produced a connectionist network called NETtalk that takes an English text as its input and produces reasonable English speech as its output. Thus, the network appears to have learned the 'rules of English pronunciation' but has done so without requiring explicit rules that combine and encode sounds in various ways.

7.2 *The Ecological Metaphor*

Ecological psychology focuses specifically on the interdependencies of humans and their environments, which typically are studied under real-world conditions rather than in the laboratory. The approach has given rise to a variety of very different approaches and lines of research, only two of which will be briefly considered. One approach is concerned with explaining and understanding perception, and can be traced to E. Brunswick and more recently, J. J. Gibson; the other approach is concerned with social behavior and appears to go back to K. Lewin and R. Barker. Although the two lines of research share the ecological perspective of examining functional adaptations of organisms to their environment, they are concerned with different issues and employ different methods. The two theoretical lines are much less explicit about the actual properties of the assumed cognitive system that intervenes between perception and action than are the computer and brain metaphors discussed above, and both have much less influence in contemporary Cognitive Psychology as do their two contenders. Nevertheless, they represent a succinctly different understanding of how the cognitive system might function and belong, at least in part, to the realm of Cognitive Psychology.

7.2.1 Gibson's ecological psychology. Following classical ecological theory, Gibson regards organism and environment to be an inseparable pair. A critical feature of this conception is that environment is not defined independently of organisms, nor are organisms defined independently of environments.

Gibson considers the first task of ecological psychology to be an adequate description of the environment. Environment consists of a medium, substances, and surfaces separating substance from medium. In a successful adaptation, organisms need to

perceive which aspects of surface, substance, and medium persist and which aspects change in regard to specific environmental events.

The ecological approach to visual perception assumes that senses represent evolved adaptations to an organism's environment. These adaptations develop in relation to environmental factors contributing to an organism's survival. Evolutionary success requires sensory systems that directly and accurately depict the environment. The key stimulus features that contribute to an organism's survival which Gibson termed 'affordances' are invariant. Affordances differ according to situation and species and are perceived directly from the pattern of stimulation arising from them. They do not change as the needs of observers change; affordances have both objective and subjective properties, becoming a fact of the environment and a fact of behavior.

Gibson's approach is sometimes referred to as the theory of direct perception, and has received empirical support from, among others, research by E. J. Gibson, studying the development of perceived invariance in infancy. E. J. Gibson's research, for instance, has demonstrated that one of the properties of perceptual learning and development does appear to be the increasing ability to extract information about the permanent properties of objects.

7.2.2 Barker's ecological psychology. According to R. Barker, behavior should be studied without outside manipulation or imposition of structure. Rather than contrive artificial settings, Barker advocated the study of behavioral settings that already exist, using methodology that exerts minimal influence upon the situation.

To collect systematic records of behavior in natural contexts, Barker argued for the establishment of so-called field stations, established organizational units that were to continue over time and whose staff included both continuing and visiting researchers. Barker and co-workers established the Midwest Psychological Field Station in Oskaloosa, Kansas, where for a period of 24 years detailed systematic records were kept of community life. Observers were stationed throughout the town, and recorded everyday activities of children. Barker concluded that the behavior of a child could often be predicted more accurately from knowing the situation the child was in, than from knowing individual characteristics of the child.

Barker's conception of ecological psychology rests on several assumptions, (a) human behavior must be studied at a level that recognizes the complexity of systems of relations linking individuals and groups with their social and physical environments; (b) environment–behavioral systems have properties that develop over long periods of time; (c) change in one part of the system is likely to affect other parts of the system; and (d) the challenge of ecological psychology is to obtain sufficient understanding to be able to predict and control the effects of planned and unplanned interventions.

Barker's conception of ecological psychology has been extended and refined in recent years (e.g., J. Barker, K. Fox, A. Wicker). For example, Fox has linked it to social accounting theory, showing how data from large-scale inventories of behavior settings can reveal changes in the quality of life in communities.

From the brief desciption of ecological psychology above, it should be clear that despite the fact that the ecological metaphor has more than some intuitive appeal to it, both its level of precision and its scope are far smaller than those of the computer and the brain metaphors. Consequently, the ecological metaphor plays a very minor role in current Cognitive Psychology.

8. *Research Methods in Cognitive Psychology*

Cognitive psychologists rely heavily on the experimental method, in which independent variables are manipulated and dependent variables are measured to provide insights into the specifics of the underlying cognitive system. To statistically evaluate the results from experiments, Cognitive Psychology relies on standard hypothesis testing, along with inferential statistics (e.g., analyses of variance).

The research methods Cognitive Psychology utilize depend, in part, on the area of study and consist primarily of chronometric methods, memory methods, cross-population studies, case studies, measures of brain activity, and computational modeling.

8.1 *Chronometric Methods*

Beginning with early work by Donders (1868–1969), cognitive psychologists have used reaction times to measure the speed of mental operations. Donders developed the so-called subtractive method. For example, task A might be assumed to require Process 1, whereas task B might require Processes 1 and 2. Donders assumed that cognitive operations are independent of each other and are processed in a strictly serial manner. Thus, the duration of Process 2 can be estimated by subtracting the response time for task A from the reaction time for task B.

To circumvent some of the restrictive assumptions of the subtractive method, Saul Sternberg introduced an additive factors logic. According to this logic, if a task contains distinct processes, then there should be variables that selectively influence the speed of each process. Thus, if two variables influence different processes, their effects should be statistically additive. By contrast, if two variables affect the same process, their effects should statistically interact.

More recently, researchers have empirically identified for so-called cascaded systems in which neither the

assumptions of Donders nor of Sternberg hold because mental processes occur simultaneously at different information-processing levels. New measures have been developed that combine reaction time measurement with the measurement of other properties of the cognitive system (e.g., speed-accuracy tradeoff functions, eye-tracking methods).

8.2 Memory Methods

One of the first to experimentally study human memory was Hermann Ebbinghaus who developed the savings technique to assess retention of nonsense syllables. Retention was measured in terms of the number of trials necessary to relearn a list of syllables relative to the number of trials necessary to learn the list for the first time.

More recently, researchers have begun to distinguish between three different aspects of memory, encoding, retention, and retrieval, and have developed methods to study the three aspects in isolation. For example, one way of investigating encoding processes is to manipulate humans' expectancies by way of intentional vs. incidental learning instructions. By contrast, retrieval processes are often studied in one of two general ways. On an explicit memory test, participants are presented a list of materials and at some later point in time are given a test in which they are asked to retrieve the earlier presented material. Retrieval is measured in terms of recall, recognition, or cued recall. On implicit memory tests, participants are not directly asked to recollect an earlier episode, but rather, are asked to engage in a task where performance might benefit from earlier exposure to the stimulus items. Interestingly, recent research has demonstrated that some neuropsychological populations (e.g., amnesic patients) can be unimpaired on implicit memory tests, but can show considerable impairment in explicit memory tests.

8.3 Case Studies

Although relatively rarely used in Cognitive Psychology, single case studies can provide vital information on how the cognitive system may be structured and which specific processes might be necessary to complete specific tasks. The classic case of HM, who as a consequence of an earlier epilepsy operation, acquired severe memory loss on explicit tasks though not implicit tasks, might serve as a striking example arguing for the dissociation of different memory structures. Case studies can provide rather convincing constraints for cognitive psychologists' understanding of the architecture of human cognition.

8.4 Cross-population Studies

Cognitive Psychology relies heavily on college students as their research participants. Recently, however, there has been an increasing interest in comparing both the structure and the processes of the cognitive system across distinct populations. For example, studies of cognition from early childhood to older adulthood attempt to trace developmental changes in specific mental operations, such as speed of processing and memory. In addition, studies with special clinical populations are conducted in order to understand breakdowns in mental functioning, such as they occur in Alzheimer's disease, schizophrenia, or amnesia.

8.5 Measures of Brain Activity

In recent years a variety of possibilities to measure some correlates of mental activity in the brain (e.g., evoked potentials, positron emission tomography [PET], functional magnetic resonance imaging [fMRI]) have become available. Evoked potentials measure the electrical activity of systems of neurons; PET measures blood flow. Because the measures differ widely in their invasiveness and in terms of their temporal and spatial resolutions, a combination of these methods, together with behavioral measurements (e.g., reaction time, accuracy) appears to be an extremely promising candidate for increasing understanding of the interaction between Neuropsychology and Cognitive Psychology.

8.6 Computational Modeling

Since the early research of Newell and Simon on the General Problem Solver, theoretical assumptions concerning both the structure of, and the processing within, the cognitive system have been tested by implementing the assumptions in running computer programs. Recently, the modeling has been of two different varieties, connectionist versus symbolic modeling (see above). For example, in order to model the processes underlying human language learning, the symbolic modeling approach assumes that humans acquire a set of rules that specify how language constituents can be combined within a language and can be specified in a running program. Alternatively, the cognitive system might acquire the 'rules of language' without directly specifying these rules as symbols at all, that is, within a distributed representational system (connectionist theory). At present, the controversy surrounding these two alternative modeling accounts is far from settled, and, importantly, reflects a fundamental issue regarding the nature of the human cognitive system that was addressed in more detail above (computer metaphor vs. brain metaphor).

9. The Future of Cognitive Psychology

If the past is a good predictor of the future, then the future of Cognitive Psychology is difficult to predict.

Most likely, however, neither the theoretical scope nor the empirical methods of the field are going to change dramatically. If there will be disagreement among scientists, it will concern the nature of the mental system intervening between environmental input and behavior. The most desirable future scenario is perhaps one in which the three main metaphors (i.e., computer, brain, ecological) will be integrated into a coherent one. Because the three metaphors deal with distinct levels of the human mind, this scenario is perhaps not an unlikely, remote theoretical possibility.

See also: Brain Development, Ontogenetic Neurobiology of; Cognitive Neuroscience; Cognitive Psychology: History; Cognitive Science: History; Cognitive Science: Overview; Cognitive Science: Philosophical Aspects; Experimentation in Psychology, History of; Human Cognition, Evolution of; Information Processing Architectures: Fundamental Issues; Intelligence, Evolution of; Intelligence: Historical and Conceptual Perspectives; Mathematical Psychology; Problem Solving and Reasoning, Psychology of; Psychology: Historical and Cultural Perspectives; Psychology: Overview

Bibliography

Anderson J R 1983 *The Architecture of Cognition*. Harvard University Press, Cambridge, MA

Anderson J R 1995 *Cognitive Psychology and its Implications*. Freeman, New York

Broadbent D E 1958 *Perception and Communication*. Pergamon, New York

Chomsky N 1959 Review of Skinner's verbal behavior. *Language* **35**: 26–58

Fodor J A 1983 *The Modularity of Mind*. MIT Press/Bradford Books, Cambridge, MA

Fodor J A, Pylyshyn Z W 1988 Connectionism and cognitive architecture: A critical analysis. *Cognition* **28**: 3–71

Gardner H 1985 *The Mind's New Science: A History of the Cognitive Revolution*. Basic Books, New York:

Gibson J J 1979 *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston

Lachman R, Lachman J L, Butterfield E C 1979 *Cognitive Psychology and Information Processing*. Erlbaum, Hillsdale, NJ

McClelland J L, Rumelhart D E (eds.) 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2. MIT Press/Bradford Books, Cambridge, MA

Miller G A, Galanter E, Pribram K H 1960 *Plans and the Structure of Behavior*. Holt, Rinehart & Winston, New York

Minsky M 1985 *The Society of Mind*. Simon and Schuster, New York

Neisser U 1967 *Cognitive Psychology*. Appleton-Century-Crofts, New York

Newell A 1980 Physical symbol systems. *Cognitive Science* **4**: 135–83

Newell A 1991 *Unified Theories of Cognition*. Cambridge University Press, Cambridge, MA

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Pylyshyn Z W 1984 *Computation and Cognition*. MIT Press, Cambridge, MA

Rumelhart D E, McClelland J L (eds.) 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1. MIT Press/Bradford Books, Cambridge, MA

Simon H A 1969 *The Sciences of the Artificial*. MIT Press, Cambridge, MA

P. A. Frensch

Cognitive Science: History

The roots of cognitive science extend back far in intellectual history, but its genesis as a collaborative endeavor of psychology, computer science, neuroscience, linguistics, and related fields lies in the 1950s. Its first major institutions (a journal and society) were established in the late 1970s. This history describes relevant developments within each field and traces collaboration between the fields in the last half of the twentieth century.

A key contributor to the emergence of cognitive science, psychologist George Miller, dates its birth to September 11, 1956, the second day of a Symposium on Information Theory at MIT. Computer scientists Allen Newell and Herbert Simon, linguist Noam Chomsky, and Miller himself presented work that would turn each of their fields in a more cognitive direction. Miller left the symposium 'with a strong conviction, more intuitive than rational, that human experimental psychology, theoretical linguistics, and the computer simulation of cognitive processes were all pieces from a larger whole, and that the future would see a progressive elaboration and coordination of their shared concerns' (Miller 1979).

This early conference illustrates an enduring feature of cognitive science—it is not a discipline in its own right, but a multidisciplinary endeavor. Although a few departments of cognitive science have been created at universities in subsequent decades, most of its practitioners are educated and spend their careers in departments of the contributing disciplines. The relative prominence of these disciplines has varied over the years. Computer science and psychology have played a strong role throughout. Neuroscience initially was strong, but in the years immediately following the 1956 conference its role declined as that of linguistics dramatically increased. By the 1970s, such disciplines as philosophy, sociology, and anthropology were making distinctive contributions. Recently, with the emergence of cognitive neuroscience, neuroscience has once again become a central contributor.

1. Intellectual Ancestors of Cognitive Science

1.1 Artificial Intelligence

One of the central inspirations for cognitive science was the development of computational models of cognitive performance, which bring together two ideas. First, conceiving of thought as computation was an offshoot of the development of modern logic. In his 1854 book, *The Laws of Thought*, the British mathematician George Boole demonstrated that formal operations performed on sets corresponded to logical operators (and, or, not) applied to propositions; Boole proposed that these could serve as laws of thought. Second, conceiving of computers as devices for computation can be traced back to Charles Babbage's plans in the 1840s for an 'analytical engine' and his collaboration with Lady Lovelace (Ada Augusta Byron) in developing ideas for programming the device. These ideas gained new life in the 1930s and 1940s with the development of automata theory (especially the Turing machine), cybernetics (centered on Norbert Weiner's feedback loops), designs for implementing Boolean operations via electric on/off switches (Claude Shannon), and information theory (also Shannon). Implementation became possible with the invention of electrical circuits, vacuum tubes, and transistors and was put on a fast track by World War II (ENIAC was completed at the University of Pennsylvania in 1946). By the mid-1950s, Newell and Simon (at RAND and then Carnegie-Mellon) produced the first functioning program for reasoning, a theorem-prover called Logic Theorist, and the first list-processing language, IPL. Meanwhile, John McCarthy and Marvin Minsky at MIT were developing a broad based agenda for the field they named artificial intelligence (AI) and more specialized endeavors also got underway (e.g., machine translation of languages and chess-playing programs, neither truly successful until the 1990s).

1.2 Psychology

During the same period psychology began emerging from a long domination by behaviorism, especially in North America. Behaviorism had the lasting impact of focusing experimental psychology on explaining behavior and relying on behavior as its primary source of evidence. Radical behaviorists, such as B. F. Skinner, actively opposed positing internal processes and focused on what was observable: describing how behavioral responses changed with contingencies of reinforcement. Other behaviorists, such as Clark Hull, were willing to posit variables intervening between stimulus and response, such as drive, but emphasized doing so in the context of developing a mathematico-deductive theory accounting for behavior. Edward Tolman, an atypical behaviorist, went so far as to propose that rats navigate their environments by constructing cognitive maps. Most psychologists, however, took learning rather than cognition as their domain of concern. The sentiment was well captured by George Mandler, looking back on his graduate student days at Yale: '… cognition was a dirty word for us … because cognitive psychologists were seen as fuzzy, hand-waving, imprecise people who never really did anything that was testable' (quoted in Baars 1986).

Although behaviorism cast a broad shadow in the US, alternatives which later influenced cognitive science thrived elsewhere: Continental Europe (especially Jean Piaget's genetic epistemology), the UK (e.g., Sir Frederic Bartlett's appeal to schemas to explain memory distortions and Donald Broadbent's analyses of memory and attention), Germany and Austria (Gestalt psychology), and the Soviet Union (Lev Vygotsky and Alexander Luria). Within the US, psychophysics (and to some extent developmental, social, and clinical psychology) functioned largely outside behaviorism's influence. Several psychologists who later pioneered a more cognitive approach, including Miller, Ulric Neisser, and Donald Norman, received their training in S. S. Stevens's Psychoacoustic Laboratory at Harvard. Miller earned his Ph.D. in 1946 for research on optimal signals for spot jamming of speech. Just 10 years later, he was talking about the structure of internal information processing systems, such as the 'seven plus or minus two' limitation in such domains as short-term memory. In 1960, Miller, Eugene Galanter, and Karl Pribram broke new ground in their *Plans and the Structure of Behavior*. In the same time frame, Miller's Harvard colleague Jerome Bruner was pioneering several strands of cognitive research: he showed that internal mental states influence perception and, arguing that categories are central to thought, Bruner, Goodnow, and Austin traced how people acquire them in their 1956 book, *A Study of Thinking*.

1.3 Neuroscience

Research into the brain was long thought to be relevant to understanding mental processes. One line of research focused on deficits stemming from brain lesions, such as Broca's classic nineteenth century work lining articulate speech to what is now called Broca's area. Although a holistic tradition in early twentieth century brain research temporarily turned investigators away from localization studies, Norman Geschwind and others gave new life to this approach in the 1950s. As well, improvements in electrophysiological techniques, including brain stimulation, single cell recording, and EEG recording, provided additional clues. As evidenced by a 1948 conference, 'Cerebral Mechanisms in Behavior,' there was eager engagement at the time between neurophysiologists, biologically oriented psychologists, and computer scientists.

One of the fruitful products of this engagement in the 1940s to 1960s was the development of neural networks, a kind of computational modeling pioneered by neurophysiologist Warren McCulloch and logician Walter Pitts. Donald Hebb proposed to build cell-assemblies by strengthening connections between neurons that fired simultaneously, a technique still in use. Oliver Selfridge had layers of units competing in parallel to recognize patterns in his *Pandemonium* simulation. Frank Rosenblatt built layered networks that learned through error correction (*Perceptrons*). Neural networks lost influence due to a devastating critique of Perceptrons by Minsky and Seymour Papert in 1969, but were revived when more effective techniques became available in the 'new connectionism' of the 1980s and beyond.

1.4 Linguistics

Linguistics began to move towards a central role in the emerging interdisciplinary discussion of mind and brain around the time of the 1956 MIT conference. In the early decades of the twentieth century, linguistics had changed its emphasis from reconstructing the history of languages to studying the structure of languages. Structuralist linguists such as Franz Boas, Edward Sapir, and the positivist Leonard Bloomfield focused on lower-level structural units (phonemes and morphemes). In the 1950s, post-Bloomfieldian Zellig Harris turned his attention to syntax and introduced the idea of transformations that normalized complex sentences by relating them to simpler kernel sentences. This idea launched a revolution in linguistics when it was further developed by Harris's student Noam Chomsky at the University of Pennsylvania and then MIT. In his 1957 *Syntactic Structures*, Chomsky proposed the idea of a grammar as a generative system—a set of rules that would generate all and only members of the infinite set of grammatically well-formed sentences of a human language—and argued that finite state and phrase structure grammars, though generative, were inadequate. A series of transformations was needed to obtain an appropriate surface structure from an initial deep structure created by means of phrase structure rules.

2. The Maturation of Cognitive Science

If the 1956 symposium represented the birth of cognitive science, it had a lot of maturing to do before it solidified into a major recognizable area of scientific inquiry. It did not even obtain its name and institutional identity until the mid- to late 1970s. But in the intervening two decades, interaction and collaboration between computer science, psychology, and linguistics developed and began to bear fruit.

2.1 Artificial Intelligence

Though based in computer science, much artificial intelligence (AI) research was directed towards accounting for the kinds of behavior studied by psychologists. Newell and Simon soon went beyond their initial Logic Theorist program to a General Problem Solver they used in less formal domains, such as solving Tower of Hanoi problems. They developed such concepts as subgoals, heuristics, and satisficing and introduced the production system framework, which employs rules that operate on the contents of working memory when their antecedent conditions are satisfied. Their collaborative work culminated in their 1972 book *Human Problem Solving*, but each went on to develop further systems, such as SOAR and extensions of EPAM. At MIT McCarthy developed a list processing language (LISP), which became a standard tool of AI. Students working with him and Minsky wrote LISP programs to perform such tasks as retrieving semantic information (Raphael's SIR) and solving algebra word problems (Bobrow's STUDENT) and geometric analogies (Evans's ANALOGY). A 1968 book reporting this work also included a seminal chapter by Ross Quillian introducing semantic networks. At Stanford, a team headed by Charles Rosen built a computer-controlled robot named Shakey that could reason backwards from goals and take appropriate actions with boxes that were found in its environment. Working with a simulated box world rather than a physical one, Terry Winograd's SHRDLU at MIT offered innovations in data structures and planning and had the most successful natural language interface of the early 1970s. Around the same time William Woods developed Augmented Transition Network (ATN) grammars at Harvard and BBN. As the 1970s progressed, AI researchers recognized the limitations of reasoning with only atomized information. Some proposed larger-scale knowledge structures, such as Roger Schank's scripts and MOPs and Minsky's frames. Also, considerable progress was made in such specialized areas as expert systems, speech understanding programs, and computational linguistics.

2.2 Psycholinguistics

One of the most fruitful collaborations was between psychology and linguistics. Modern psycholinguistics had already begun to emerge in the early 1950s, especially in the context of a summer seminar sponsored by the Social Science Research Council in 1953. One of the aims of this interaction between post-Hullians and post-Bloomfieldians was to investigate the psychological reality of linguistic constructs such as phoneme. Though many of their empirical strategies still thrive in some form, the field was impacted by the echos of Chomsky's revolution in linguistics. In his 1959 review of B. F. Skinner's *Verbal Behavior*,

Chomsky emphasized not only that linguistic behavior does not consist in reproduction of acquired responses but is creative in the sense that there is no bound to the novel but grammatically well-formed sentences one might produce or hear. Focusing on the poverty of the stimulus argument (that the input to children is inadequate to induce a language), Chomsky also later argued for an innate language capacity (Universal Grammar). A long-standing conflict developed between Chomskian developmental psycholinguists and those taking a cognitive interactionist perceptive (e.g., Elizabeth Bates). Meanwhile, Chomkian inspired psychologists showed that sentences with more transformations in their derivation were more difficult to process. Later, changes in linguistic theory led to a more nuanced psycholinguistics, including some Chomskian approaches.

2.3 Psychology

Information processing psychology drew explicitly or implicitly on computational ideas from information theory and AI. Some of the first glimmers of an information processing approach to psychology appeared in the work of Miller and Bruner, who established a Center for Cognitive Studies at Harvard in 1960. Research in the center focused on a host of topics including conceptual organization, language processing and development, visual imagery, memory, and attention. Eleanor Rosch, a Bruner student, began work that led in the 1970s to a view of categories that emphasized prototypes, fuzzy boundaries, and the primacy of basic-level categories. And in 1967 Ulric Neisser, one of the Center's many research fellows and visitors, published *Cognitive Psychology*. This book introduced and synthesized the newly burgeoning work on information processing, particularly emphasizing attention and pattern recognition, and it quickly became the bible for a new generation of students.

Although the Center closed in 1970, information-processing approaches to psychology had already begun to spread to other universities. Stanford University, for example, built from its existing strength in mathematical psychology (including the work of William K. Estes) quickly to emerge as a premier center for information processing. In 1968 Richard Atkinson and Richard Shiffrin developed a model that integrated previous work on control processes, sensory memory (Sperling), short-term memory (Peterson and Peterson), and the distinction between short-term and long-term memory (William James; Waugh and Norman). Roger Shepard did elegant work in mathematical psychology (e.g., he pioneered nonparametric multidimensional scaling), but is best known for his research on mental imagery and mental rotation with such students as Lynn Cooper and Jacqueline Metzler. For example, they demonstrated that when subjects had to decide whether a comparison stimulus was a rotation or a mirror image of a geometrical form, their reaction times increased linearly with the degree of rotation. This suggested that subjects mentally rotated the comparison stimulus—an attention-grabbing claim at a time when mentalism was still suspect in many quarters. A third researcher, Gordon Bower, moved from mathematical models of learning towards more cognitively oriented work on the nature of mental representations. One of his students, John Anderson, worked with Bower on a very influential semantic network model (HAM), that was described in their 1973 book, *Human Associative Memory*. Later Anderson combined it with a production system component in ACT* and its predecessors.

Another exemplar is University of California, San Diego, where the trio of Peter Lindsay, Donald Norman, and David Rumelhart created a collaborative research group (LNR) in a new department and institution. In 1975 they published *Explorations in Cognition*, which ended with what may have been the first published use of the term cognitive science: 'The concerted efforts of a number of people from … linguistics, artificial intelligence, and psychology may be creating a new field: cognitive science' (Norman et al. 1975). (A second candidate for first use is a 1975 book by Bobrow and Collins, *Representation and Understanding: Studies in Cognitive Science*.) In addition to work on varied topics including memory, word recognition, problem solving, imagery, and analogy, the group implemented an ambitious memory model, MEMOD. It featured active structural networks, which used a common semantic network format to represent both data and process (e.g., declarative and procedural knowledge). A decade later yet another collaborative group coalesced at UCSD around Rumelhart and James McClelland; known as the PDP (parallel distributed processing) group, it helped bring neural networks (connectionist models) back to center stage.

Sensing the potential to catalyze cognitive science programs at these and other universities, the Alfred P. Sloan Foundation launched an initiative that eventually provided $17.4 million over 10 years to such institutions as MIT and the University of California, Berkeley. During 1982–84 another foundation, The System Development Foundation, contributed $26 million for computational linguistics and speech, with its largest support going to the Center for the Study of Language and Information at Stanford.

During the same period, linguist-turned-computer-scientist Roger Schank, psychologist Allan Collins, and computer scientist Eugene Charniak began a new journal called *Cognitive Science*. Describing a converging view of natural and artificial intelligence in his introduction to the first issue in 1977, Collins wrote: 'This view has recently begun to produce a spate of books and conferences, which are the first trappings of an emerging discipline. This discipline might have been called applied epistemology or intelligence the-

ory, but someone on high declared it should be cognitive science and so it shall. In starting the journal we are just adding another trapping in the formation of a new discipline.' Drawing upon some of the early Sloan money, Donald Norman and his colleagues at UCSD organized the La Jolla Conference on Cognitive Science. But as planning proceeded the idea of a new society germinated, and the conference (held August 13–16, 1979) became the first annual meeting of the Cognitive Science Society.

3. Developing New Identities

By 1980 cognitive science had developed an institutional profile and was the focus of serious funding initiatives. It also had an identity, one that emphasized computational modeling of cognitive and linguistic processes, but also incorporated linguistics and psycholinguistics. The subsequent two decades have seen major efforts to revise this initial identity. Three contributing factors, each involving perspectives from disciplines not central to the cognitive science of the 1970s, deserve brief mention. First, the development of cognitive science gave philosophers, long interested in issues surrounding the mind, a chance to address such issues in the context of ongoing empirical research. Jerry Fodor argued that cognitive processes are autonomous from the neural substrate and capable of being realized in multiple ways; this view, congenial to the cognitive science of the 1970s and 1980s, was subsequently challenged by Patricia and Paul Churchland, who have emphasized the co-evolution of cognitive science and neuroscience.

Second, mavericks in a variety of disciplines found cognitive science's nearly exclusive focus on processes within the head a limitation. Philosopher Hubert Dreyfus challenged the attempt to analyze cognition as formal computational processes. Neisser integrated information processing with the ecological psychology of James J. and Eleanor J. Gibson. Edward Hutchens and many others began to focus on how embodiment and the situatedness of agents contribute to their cognitive performance.

Finally, although study of the brain largely disappeared from cognitive science in the 1960s and 1970s, partly because research on brain processes seemed too remote to contribute to understanding cognitive operations, the late 1980s and 1990s saw the emergence of cognitive neuroscience. Michael Posner, Marcus Raichle, and Steven Petersen collaborated at Washington University to show how images from PET could be used to link brain processes to cognitive processes. More recently greater spatial resolution has been gained using fMRI and greater temporal resolution using EEG-based methods such as ERP. Whether cognitive science can successfully incorporate these pilgrimages out into the world and down into the brain, or whether it will ultimately fractionate, is a question still unanswered.

See also: Artificial Intelligence in Cognitive Science; Behaviorism, History of; Cognitive Neuroscience; Cognitive Psychology: History; Cognitive Science: Overview; Cognitive Science: Philosophical Aspects; Psychology: Historical and Cultural Perspectives; Neuroscience, Philosophy of

Bibliography

Baars B J 1986 *The Cognitive Revolution in Psychology*. Guilford Press, New York

Bechtel W, Abrahamsen A, Graham G 1998 The life of cognitive science. In: Bechtel W, Graham G (eds.) *A Companion to Cognitive Science*. Blackwell, Malden, MA

Gardner H 1985 *The Mind's New Science*. Basic Books, New York

Hirst W (ed.) 1988 *The Making of Cognitive Science*. Cambridge University Press, Cambridge, UK

Miller G A 1979 A Very Personal History (Occasional paper no. 1). Center for Cognitive Science, Cambridge, MA

Norman D A, Rumelhart D E, and the LNR Research Group 1975 *Explorations in Cognition*. Freeman, San Francisco, CA

W. Bechtel, A. Abrahamsen, and G. Graham

Cognitive Science: Overview

Cognitive science (CS) is a young discipline that emerged from a research program started in 1975. It partially overlaps with its mother disciplines: psychology, artificial intelligence, linguistics, philosophy, anthropology, and the neurosciences. By no means the only discipline dedicated to the study of cognition, cognitive science is unique in its basic tenet that cognitive processes are computations, a perspective which allows for direct comparison of natural and artificial intelligence, and emphasizes a methodology that integrates formal and empirical analyses with computational synthesis. Computer simulations as generative theories of cognition (see *Cognitive Modeling: Research Logic in Cognitive Science*) have therefore become the hallmark of CS methodology.

Today, CS is an internationally established field. The dominant tradition of the early years, close to artificial intelligence and its symbol-processing architectures, has been enriched by alternative computational architectures (e.g., artificial neural networks) and by the recognition that natural, especially human cognition, rests on biological as well as on social and cultural foundations. CS studies cognitive systems, which may be organisms, machines, or any combination of these acting in an environment that may be open and dynamically changing. Cognition in CS denotes a class of advanced control mechanisms that allow for sophisticated adaptation to changing

needs (e.g., learning and planning) through computations operating on mental representations. Cognition typically coexists with simpler regulatory mechanisms, like reflexes. CS recognizes that cognition in biological systems is implemented in brain processes, but emphasizes the importance of analyses at the functional level, with cognitive neuroscience relating both domains. Applications of cognitive science may be found in human–computer interaction and in the design of software and information systems, as well as in human factors engineering, health care, and, most notably, in education.

1. Cognition and the Cognitive Science Approach

Although reflection on the mind dates back at least to Plato, the term 'cognition,' etymologically based on ancient Greek *gignoskein* and Latin *cognoscere*, is relatively recent. It surfaces in nineteenth-century psychology, which exclusively dealt with the phenomenology of consciousness, e.g., in Spencer's characterization of the interrelations of human feelings. At about the same time, the triad of thinking, feeling, and willing of eighteenth century *Vermögenspsychologie* became the well-known taxonomy of the mind, dividing it into cognition, emotion, and volition.

We all have an intuitive understanding of what 'cognition' refers to, and there is common agreement that thinking, memory, and language, and 'the use or handling of knowledge' (Gregory 1987, p. 149) are correctly subsumed under that term. On the other side, it is difficult to define the term strictly. Minimalist approaches (e.g., Searle 1990) would like to reserve its use for the contents of consciousness, whereas a maximum approach has been taken by Maturana and Varela (1980, p. 13), where they claim that 'living as a process is a process of cognition' (cf., Boden 2000, for a critique). The leading opinion, however, will consider unconscious processes as well as conscious ones (Neisser 1967, Norman 1981) without attributing cognitive abilities to every living organism (e.g., a tree) and, indeed, without confining the use of the term to biological systems. Still, 'cognition' continues to be a rather ill-defined term, which even as ambitious a project as *The MIT Encyclopedia of the Cognitive Sciences* (Wilson and Keil 1999) has not dared to treat in an article of its own.

The hallmark of the CS approach to cognition is to identify cognitive processes with computation: cognition is information processing. Not every computation is cognition, however, which means that these computational processes must be characterized (Newell and Simon 1976), and further restrictions must be named, such as referential content or 'intentionality' (Fodor 1975), or system complexity (Smolensky 1988). The rest of this section will sketch the way towards a computational theory of mind.

1.1 Formal Approaches in Philosophy, Computing, and Neuroscience

The philosophical tradition on which CS draws may be traced from the invention of number systems to medieval algebra (Raimundus Lullus) and on to Descartes and Leibniz, and from the Aristotelian syllogisms to Frege's seminal work on formal logic. Apart from being a history of the development of formal systems, this philosophical tradition can be seen as an analysis of thinking and reasoning aimed at separating content and form of argumentation. The syllogisms of Aristotle, which continued to constitute the core of logic, virtually unchanged, over more than 2,000 years, are the first milestone: In analyzing an argument such as 'All human beings will die. Socrates is a human being. Therefore, Socrates will die' as $p \rightarrow q$, $p \therefore q$(*modus ponens*) in the Aristotelian tradition (combined with the formal advancements of Lullus and Descartes), the specific content of the argument is separated from the form of reasoning, and it becomes clear that the latter is sufficient for warranting a true conclusion, provided that the input to this logical vehicle consists of premises that are true. An important generalization about reasoning had been found. It was Leibniz in the late-seventeenth century who advocated the use of formal reasoning in the hope that all fruitless discussions might be ended just by formalizing the arguments and computing the conclusions—hopelessly optimistic from our view (there would be endless debate on how to formalize the premises), but instrumental for an account of thinking as symbol processing.

Frege's (1879/1967) reformulation of logic laid the foundations of modern logical semantics and marks the beginning of the modern tradition that led to CS. The theory of symbol processing is formalized as a general theory of computation in the works of Gödel, Turing, Church, and Post between 1931 and 1943, soon to become the foundation of computer science and, especially, artificial intelligence. Logical positivism (as epitomized in Wittgenstein's *Tractatus*, written during World War I) and logical semantics (foremost, Tarski's work in the 1920s) constitute the philosophical legacy on which CS could draw.

The invention of computing machinery in the 1940s (Zuse, with Babbage as an isolated forerunner in the nineteenth century) was instrumental in promoting a computational perspective to cognition. A major step was the founding of artificial intelligence (AI) in 1956. The work done in AI, notably on human problem solving by Newell and Simon, may well be regarded as CS at a time when the term was not around yet.

The symbol processing tradition was complemented in the neurosciences by the invention of the formal neuron (an abstraction from biological neurons; McCulloch and Pitts 1943), as well as by analog approaches to computing and self-regulation and general systems theory (cybernetics; Wiener 1948).

The detection of representational functions of single neurons in the visual system (receptive fields; Hubel and Wiesel 1962) paved the way for a new, scientific approach to the concept of mental representation, while the development of artificial neural networks, notably the perceptron (Rosenblatt 1958), showed a possible way to combine these new discoveries with the idea of computation.

1.2 Other Disciplines: Linguistics and Anthropology

Linguistics made a huge step forward when researchers such as Jakobson and Troubetzkoy succeeded in discovering the common abstract structure behind the phonemes utilized in different languages. Linguistic structuralism, later to be applied to syntax by Harris and Chomsky, became another driving force in the development of CS, as well as structuralism developed in ethnology by Levi-Strauss, and in cognitive anthropology. In addition, analysis of formal languages (Chomsky 1959a) provided the link to computer science.

1.3 The Rise of Cognitive Psychology

Nineteenth-century psychology has only recently been rediscovered, but failed to contribute to CS's early development because it had been all but extinguished by behaviorism from about 1915 to 1960. Only the later behaviorists considered internal variables (notably Hull), or even memory and mental representation (Tolman). It was information theory (developed by Shannon 1948), or rather its insufficiency to explain central psychological phenomena such as the memory span, which brought G. A. Miller in the mid-1950s to reconsider human information processing. His collaboration with Noam Chomsky, and especially the latter's poignant critique of behaviorist approaches to language (Chomsky 1959b), served to reorient psychological research towards issues of internal storage and processing, towards a psychology that no longer ignored mind and consciousness, yet was careful to stay within the limits of scientific rigor (Miller et al. 1960). Cognitive psychology underwent a rapid development and in 1967, Ulric Neisser wrote the first textbook of the new field, coining its name. Gardner (1985) names further influential researchers in psychology, among them Bruner (notably his work on strategies) and Jean Piaget. Although Piaget's numerous monographs on cognitive development did not become available in English before the 1960s, he is certainly a forerunner of CS who always insisted on the importance of formal principles for explaining cognitive development.

1.4 The Origin of Cognitive Science

A state-of-the-art report on CS by the Alfred P. Sloan Foundation concludes that 'What has brought the field into existence is a common research objective: to discover the representational and computational capacities of the mind and their structural and functional representation in the brain' (Sloan Foundation 1978, p. 6). But it would not be false to say that the Sloan Foundation acted as midwife at the birth of CS. Its committee diagnosed convergent approaches visible across disciplines and went one step beyond that diagnosis to unite what still were very different approaches. Institutionalization followed soon: a journal first (*Cognitive Science*, in 1977), then, two years later, a society and a yearly conference and, still later, doctoral programs and research projects, all evolving into a flourishing new field.

2. The Basic Tenet: Cognition as Computation

The modern idea of computation was formulated in the 1930s. It owes much to Hilbert's program for the complete axiomization of mathematics, limited by Gödel's (1931) proof that not all theorems about a formal system are provable with the means provided by that very same system. Some people have tried to use this result against the idea that cognition could be computation, but did not realize that there may well be truths about the human mind that it fails to arrive at, let alone prove them formally.

The best-known definitions of computation rely on recursive functions (Church 1941), logical productions (Post 1943), or the abstract machine designed by Turing (1936). All these approaches have been proved to be equivalent in scope, and especially the Turing machine may be considered the direct forerunner of today's computers. The common core is the idea of a formal system that uses symbols, i.e., variables and operators combined to form symbolic expressions that are manipulated according to fixed rules and to the internal state of the system.

2.1 Physical Symbol Systems

When they received the Turing Award for their ground-breaking work in AI, Newell and Simon expanded the theory of symbol processing and coined the Physical Symbol Systems Hypothesis (PSSH): 'A physical symbol system has the necessary and sufficient means for intelligent action' (Newell and Simon 1976, p. 117).

A physical symbol system is a formal system. Like all formal systems, it has an 'alphabet' of (at least two) arbitrarily defined symbols, as well as operators to create and transform symbol structures (symbolic expressions) of arbitrary complexity from the elemen-

tary symbols of the alphabet according to syntactic rules. This system is 'physical' in the sense that it has been implemented in a suitable way. One such way is the encoding of the symbols as levels of voltages in an array of transistors, and the operations by hard-wired connections between transistors; thus it is done in semiconductor chips. A different way would be to encode the symbols as 'spikes' (action potentials) of neurons. Other ways are possible, but those two already show how the theory can be applied to organisms as well as to technical systems. The point is that a system like that needs the physical implementation in order to function in the real world and become more than just an idea. The kind of implementation is arbitrary, however, because the system is functioning according to its symbolic expressions and syntactic rules, completely independent of its implementation.

Symbols are arbitrary signs, but they designate objects or processes, including processes in the system itself. Their semantics is defined either by reference to an object (in the sense that depending on the respective symbolic expression, the system exerts an influence on the object or is influenced by it), or by the symbolic expression being executable as a kind of program.

Physical symbol systems may have a lot of symbol structures, which means that they need to have a symbol store, or memory, which is unlimited according to Newell and Simon. In fact, physical symbol systems are Turing-equivalent computing devices and, therefore, the PSSH is equivalent to the notion of cognition being computation. The PSSH cannot be proven, nor can it be refuted formally. It gained plausibility, however, through empirical studies of human problem solving and its simulation (Newell and Simon 1972). The PSSH has integrated AI into computer science through the common reference to the theory of automata, a platform which gives a foundation to CS as well. The distinction between functional and implementation levels, elaborated by Newell (1982) and Marr (1982), enabled CS as a science of biological as well as technical cognitive systems.

2.2 *Philosophical Foundations: Functionalism and the Computational Theory of Mind*

Mental states have been analyzed as 'intentional attitudes' in the philosophy of mind, consisting of a propositional content (e.g., $P =$ *the sun is shining*) and an attitude that characterizes one's own relation to that proposition (e.g., *I wish that P would become true*). Fodor (1975) developed this approach further, arriving at a 'language of thought' that treats the propositional content as data and the intentional relation as an algorithmic one.

If we accept these as the elements of a 'language of thought,' then the question arises of how mental states relate to brain states: a well-known problem in philosophy. Following Putnam (1960), Fodor and others conceptualize the relation between brain and mental states as being parallel to the relation between a computer (i.e., the hardware) and a program running on that computer: the mind as the software of the brain. This approach is known as the computational theory of mind. It fits well with the PSSH, and it soon became the dominant framework in CS. However, it addresses (potentially) conscious thought only, ignoring lower cognitive processes.

2.3 *Achievements and Drawbacks of the Classical Symbol-processing Approach*

From the twenty-first-century perspective, the PSSH and the computational theory of mind together constitute the classical period of CS, spanning the decade from 1975 to 1985. Within that period, cognitive modeling (see *Cognitive Modeling: Research Logic in Cognitive Science*) emerges as CS's characteristic methodology. Its applications include the following fields.

2.3.1 Problem solving. The former general model of problem solving as heuristic search (Newell and Simon 1972) was enlarged by recognizing the importance of domain-specific knowledge, which became the foundation of an AI technology (knowledge-based systems or 'expert systems'; Hayes-Roth et al. 1983, Buchanan and Shortliffe 1984) and inspired much psychological research on expertise (e.g., Ericsson and Smith 1991).

2.3.2 Cognitive architectures. Rule-based architectures evolved into ambitious models of human cognition in general, comprising memory, problem solving, learning, and some natural language processing. The best-known of these systems are SOAR (Laird et al. 1987, Newell 1990) and the impressive series of ACT, ACT*, and ACT-R frameworks, all developed by John Anderson (Anderson 1976, 1983, Anderson and Lebière 1998).

2.3.3 Natural language processing. From the late 1950s on, theoretical linguistics has been dominated by Noam Chomsky. His theories (notably Chomsky 1981) are framed as theories of human linguistic competence and have inspired CS research on human parsing (Frazier 1987, Mitchell 1994) and on language acquisition (Pinker 1984).

2.3.4 Computers and education. Knowledge-based systems have been built for purposes of instruction, so-called 'intelligent tutorial systems' (Psotka et al.

1988). Although their cost-efficiency relation turned out to be not well suited for general education, they have been used with success for the training of specialists.

The main drawback of research during the classical period of CS was that it excluded many important issues that could well have been covered by assuming mental representation and cognitive algorithms. The original objective for AI, stated by Simon (1955)—to address problems whose solution by a human being would lead us to attribute intelligence to that person—led CS and AI largely to ignore problems that do not seem to require intelligence in people. However, these problems turned out to be the real 'tough nuts,' e.g., navigation and other skilled action.

The computer technology of that period was still mainframe oriented, and interactive and graphic technologies were scarcely developed. This state of the art did not encourage researchers to model real-time agent–environment interaction, although there were some exceptions, e.g., Winograd (1972).

The 'methodological solipsism,' as advocated by Fodor (1980) in connection with his theory of mental representation, and the dominance of logic-based approaches in AI (especially during the 1980s) made CS researchers believe that all interesting aspects of cognition happened within a single symbol system.

In Fodor's 'language of thought,' content is defined as a truthful representation of (a part of?) the world, as in Tarski semantics and similar approaches, which unduly constrains mental representation and ignores its constructive nature.

To summarize, CS research in fact fell short of the scope even its original symbol-processing framework provided.

3. *Alternative Computational Frameworks*

From a very abstract viewpoint, a Turing machine is all one ever needs to compute. However, different architectures or virtual machines may make some computations easy and others difficult. Classical CS had adopted symbol-processing frameworks. There is a large gap, however, between a functional specification and the way in which a brain is built. The new connectionist movement, which surfaced in the mid-1980s, attempted to bridge the gap with a framework that was all but forgotten at that time.

3.1 *Artificial Neural Networks: Connectionism*

McCulloch and Pitts (1943) presented the biological neuron as an abstract computing device. Hebb (1949) added a lot of hypotheses, most of which turned out to be correct in the meantime (e.g., that enduring changes in the transmission efficiency of certain synapses are the neurophysiological basis of memory). Rosenblatt (1958) built a simple artificial neural network (ANN) that could be used for pattern recognition, the 'perceptron.' Soon, connectionism (the name adopted for this line of research) flourished. Its boom came to an end, however, when perceptrons ran into trouble with certain distinctions, and their limitations were mathematically proven (Minsky and Papert 1969). From then on, only a few researchers, mostly in biology and biophysics, continued the connectionist tradition.

The sudden rebirth of connectionism started with the discovery of an architectural change—introduction of 'hidden layers' in ANNs—that overcame the limitations of the perceptron (Rumelhart, McClelland and the PDP group 1986). After a period of heated dispute, connectionism has now been integrated into mainstream CS. Numerous ANN architectures have been developed, and hybrid connectionist–symbolic systems have also been constructed.

3.2 *Distributed Representations as Subsymbols*

At the heart of the dispute over connectionism was the issue of mental computation. Symbolic and 'localist' connectionist architectures (see Page 2000 for an overview) maintain that variables (or the nodes of an ANN) can be interpreted as being meaningful. In 'parallel distributed processing' (PDP), however, entities are represented by patterns, usually by a 'feature vector' containing the activation values of formal neurons. Smolensky (1988) claimed that only the subsymbolic approach (referring to the elements of a feature vector) grasps the essence of cognition, whereas symbol processing approaches were confined to a mere approximation. To the contrary, Fodor and Pylyshyn (1988) argued that connectionism either was just an implementation of symbol processing (hence banal), or inadequate for modeling productivity and systematicity in natural language. The debate was never resolved, but it is generally recognized that suitability for cognitive modeling is more important than an abstract decision about which framework is 'better,' and that, in fact, both frameworks share important characteristics, being computational as well as representational.

3.3 *Beyond Connectionism: Nonlinear Dynamics*

Since the early 1990s, another computational framework has been claimed to be useful for cognitive modeling: the theory of nonlinear dynamic systems, initially known as 'chaos theory' (Port and van Gelder 1995). Fine-grained analyses of movement (e.g., in phonetics) and of developmental changes are the intended area of application. Although this renders a more precise picture of how cognitive processes are implemented (like the neural models used in biology

are more detailed than connectionist accounts), it could well be that these characteristics are less important on the functional level (Eliasmith 2000).

4. The 'New' Cognitive Science: Interacting Cognitive Systems

The technical cognitive systems of classical AI and CS were systems that were *only* cognitive at their functional level. This is never the case in biological systems: animals live, move and eat, and reproduce. All this basic behavior is made possible without cognition in the sense of thoughtful decision (in lower animals, at least). Built-in physiological regulations, reflexes, instincts and species-specific behaviors serve to achieve the necessary adaptation. Learning in its most primitive form (classical conditioning) has been demonstrated in some worms and, of course, higher species; categorization of stimuli (the forerunner of concepts) in birds and mammals; episodic memory at least in some mammals, especially in apes; full-fledged language in humans only.

Cognition has evolved because of the adaptive value of learning (and the cultural tradition that is based upon it) and thinking (in mental simulation of an act and its consequences). But cognition came late; it has to coexist with those basic regulatory mechanisms mentioned above. In the human species, most of an individual's knowledge, and even specialized cognitive processes, e.g., reading and writing, have to be learned and trained within the culture that provides it. Sensorimotor processes (e.g., driving in a city) rely heavily on a system's interaction with its environment and with special tools (like a car). Also, we are embedded in a social world: many cognitive processes, foremost language, are acquired through social interaction. All this cannot be modeled adequately by a single symbol-processing system that scarcely (if at all) interacts with its environment and with other cognitive systems. In CS, this paradigm shift occurred gradually, starting in the mid-1980s. Having extended its scope since its classical period, modern CS appears as the science of cognitive systems in a fuller sense, as described above.

4.1 Situated Cognition

Using the methodology of field studies, Suchman (1987), Hutchins (1995a, 1995b) and others could demonstrate how much people rely on cues and representations provided by the environment. While traditional accounts of planning rely on mental representations and processes exclusively, human planning can be shown to rely on maps, road signs, guidance by other people, and other external sources. The importance of external representations in problem solving has been repeatedly demonstrated (Zhang 1997).

As is so often the case, 'situated cognition' was introduced as an alternative to the 'old' paradigm of CS. Vera and Simon (1993), however, insisted that CS from its very start (e.g., Newell and Simon 1976) does not exclude situatedness, but rather emphasizes its necessity. Indeed, to realize that cognition is situated can also solve the problem of symbol grounding: mental representations arise and get their meaning in the context of acting in the world (a solution already envisioned by Newell 1980). Situated cognition, it seems, highlights a formerly neglected aspect of CS.

Situatedness also includes the body, not only in the basic sense that cognition coexists with other, more basic processes in organisms. Rather, mental representations are in many ways influenced by the cognitive representation of the body, as Lakoff and Johnson (1980) have shown for metaphors.

4.2 The Social Aspects of Cognition

Culture, and social action, shape a lot of our thoughts and cognitive skills: 'Sociality and culture are made possible by cognitive capacities, contribute to the ontogenetic and phylogenetic development of these capacities, and provide specific inputs to cognitive processes' (Sperber and Hirschfeld 1999, p. cxi). By focusing on a single individual concerned with a single task only, CS had followed the model provided by experimental psychology, its most prominent source of empirical data. The recent change of view in favor of social and cultural factors was mainly due to work in an applied CS area where social interaction is central: education. Lave (1988) gives numerous demonstrations of cultural influences on the solution of apparently context-free tasks in mathematics. More generally, culture-oriented theories of development and learning (beyond sociobiology as in Lumsden and Wilson 1981) have been advanced by Cole (1991) and Tomasello et al. (1993).

Because of the sheer amount of detail provided in cultures, especially in modern industrial ones with libraries and a high degree of specialization in work, it is very difficult to integrate cultural aspects in cognitive theories. The problem is eased when analyses focus on narrowly defined tasks.

A related field is research on groups (e.g., in computer-supported cooperative work (CSCW); Olson and Olson 1997) and on teams of experts (Hinsz et al. 1997). Here, as well as in knowledge engineering (Strube et al. 1996), the task itself comprises the exchange of knowledge and, at least to a certain degree, the development of 'shared knowledge' or 'shared mental models' (Lewis and Sycara 1993).

4.3 A New Paradigm: Autonomous Social Agents and Mixed Groups

According to the dominant paradigm in CS, cognitive systems are conceived as autonomous social agents,

situated in a complex dynamic environment. This view bears no accidental resemblance to the shift in AI instigated by new architectures in robotics (Brooks 1991) and the development of intelligent agents (see Wooldrige and Jennings 1995 for an overview).

Agent approaches emphasize the complex action control needed for an agent (robot or organism) that pursues its own goals, which are always many, and potentially conflicting (Maes 1990). Because of restricted resources, agents in the real world cannot be fully rational; however, 'bounded rationality' (Simon 1955) has been claimed a universal principle of human thought (Gigerenzer et al. 1999).

Distributed AI (Bond and Gasser 1988) studies cooperation and competition among agents. The biennial RoboCup contest (Kitano et al. 1997) epitomizes this line of research, having robot or simulated agent teams playing soccer games against each other. The relevance for CS lies in the development of integrated architectures that comprise cognitive and more primitive (so-called 'reactive') regulation, as well as social interaction. This fits well with the current emphasis on situated cognition and with applied problems of CS, e.g., in the field of office automation and cooperation in mixed groups that comprise human workers as well as technical systems, as in the case of air traffic (Hutchins 1995b).

5. *Achievements and Present State*

CS in the twenty-first century has grown from an innovative interdisciplinary field to an academic discipline of its own, albeit in a still early stage of institutionalization. CS has been evolving differently, however, in different parts of the world. It has a solid infrastructure of departments, graduate programs, etc., in the UK and North America, in France, and in some other countries, but remains still an early stage of institutionalization in Germany, for instance, and still less in many other different countries. In the USA, cognitive neuroscience has split away from cognitive science, at least organizationally. In other countries like France, however, it would be impossible to imagine CS without brain research and neuroscience. The necessity and importance of real-world applications of CS in education, industry, and other fields is generally recognized, but CS still lacks a well-developed professional profile. On the other hand, CS has an impressive record of success in research, an infrastructure of international and national academic societies (foremost the Cognitive Science Society, founded in 1979), international and national conferences, and dedicated CS journals (*Cognitive Science*, founded in 1977, and *Cognitive Science Quarterly*, founded in 2000). An analysis of publications in *Cognitive Science* in 1977–95 found evidence not only for a dominance of psychologists and computer scientists among the CS community, but also for CS 'as a discipline of its own [...] becoming increasingly more common' (Schunn et al. 1998, p. 117). CS, as it seems, is still but steadily evolving.

See also: Cognitive Psychology: History; Cognitive Psychology: Overview; Cognitive Science: History; Cognitive Science: Philosophical Aspects

Bibliography

Anderson J R 1976 *Language, Memory, and Thought*. Erlbaum, Hillsdale, NJ

Anderson J R 1983 *The Architecture of Cognition*. Harvard University Press, Cambridge, MA

Anderson J R, Lebière C 1998 *Atomic Components of Thought*. Erlbaum, Hillsdale, NJ

Boden M A 2000 Autopoiesis and life. *Cognitive Science Quarterly* **1**: 135–46

Bond A H, Gasser L (eds.) 1988 *Readings in Distributed Artificial Intelligence*. Kaufmann, San Mateo, CA

Brooks R A 1991 Intelligence without representation. *Artificial Intelligence* **47**: 139–59

Buchanan B G, Shortliffe E H 1984 *Rule-based Expert Systems*. Addison-Wesley, Reading, MA

Chomsky N 1959a On certain formal properties of grammars. *Information and Control* **2**: 137–67

Chomsky N 1959b Verbal behavior—Skinner B F. *Language* **35**: 26–58

Chomsky, N 1981 *Lectures on Government and Binding*. Foris, Dordrecht, The Netherlands

Church A 1941 *The Calculi of Lambda-Conversion*. Princeton University Press, Princeton, NJ

Cole M 1991 A cultural theory of development: what does it imply about the application of scientific research? Special issue: culture and learning. *Learning and Instruction* **1**: 187–200

Eliasmith C 2000 Is the brain analog or digital? The solution and its consequences for cognitive science. *Cognitive Science Quarterly* **1**: pp. 147–70

Ericsson K A, Smith J (eds.) 1991 *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press, Cambridge, UK

Fodor J A 1975 *The Language of Thought*. Crowell, New York

Fodor J A 1980 Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* **3**: 63–73

Fodor J A, Pylyshyn Z W 1988 Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**: 3–71

Frazier L 1987 Sentence processing: a tutorial review. In: Coltheart M (ed.) *The Psychology of Reading. Attention and Performance*. Erlbaum, Hove, UK, Vol. 12, pp. 559–86

Frege G 1879/1967 Begriffsschrift: a formula language, modeled upon that of arithmetic, for pure thought. In: van Heihenoort J (ed.) *From Frege to Gödel: a Source Book on Mathematical Logic, 1879–1931*. Harvard University Press, Cambridge, MA pp. 5–82

Gardner H 1985 *The Mind's New Science: a History of the Cognitive Revolution*. Basic Books, New York

Gigerenzer G, Todd P M, ABC Research Group 1999 *Simple Heuristics That Make Us Smart*. Oxford University Press, New York

Gödel K 1931 Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. *Monatshefte für Mathematik und Physik* **38**: 173–98
Gregory R L (ed.) 1987 *The Oxford Companion to the Mind*. Oxford University Press, Oxford, UK
Hayes-Roth F, Waterman D A, Lenat D B 1983 *Building Expert Systems*. Addison-Wesley, Reading, MA
Hebb D O 1949 *The Organization of Behavior*. Wiley, New York
Hinsz V B, Tindale R S, Vollrath D A 1997 The emerging conceptualization of groups as information processors. *Psychological Bulletin* **121**: 43–64
Hubel D H, Wiesel T N 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* **160**: 106–54
Hutchins E 1995a *Cognition in the Wild*. MIT Press, Cambridge, MA
Hutchins E 1995b How a cockpit remembers its speeds. *Cognitive Science* **19**: 265–88
Kitano H, Asada M, Kuniyoshi Y, Noda I, Osawa E, Matsubara H 1997 RoboCup: a challenge problem for AI. *AI Magazine* **18**: 73–85
Laird J E, Newell A, Rosenbloom P S 1987 SOAR: an architecture for general intelligence. *Artificial Intelligence* **33**: 1–64
Lakoff G, Johnson M 1980 *Metaphors We Live By*. Chicago University Press, Chicago
Lave J 1988 *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge University Press, Cambridge, UK
Lewis C M, Sycara K P 1993 Reaching informed agreement in multispecialist cooperation. *Group Decision and Negotiation* **2**: 279–99
Lumsden C J, Wilson E O 1981 *Genes, Minds and Culture*. Harvard University Press, Cambridge, MA
Maes P (ed.) 1990 *Designing Autonomous Agents. Theory and Practice from Biology to Engineering and Back*. MIT Press, Cambridge, MA
Marr D 1982 *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco
Maturana H R, Varela F J 1980 *Autopoiesis and Cognition: the Realisation of the Living*. Reidel, Dordrecht, The Netherlands
McCulloch W S, Pitts W 1943 A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–33
Miller G A, Galanter E, Pribram K H 1960 *Plans and the Structure of Behavior*. Holt, New York
Minsky M, Papert S 1969 *Perceptrons*. MIT Press, Cambridge, MA
Mitchell D C 1994 Sentence parsing. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA, pp. 375–409
Neisser U 1967 *Cognitive Psychology*. Prentice-Hall, Englewood Cliffs, NJ
Newell A 1980 Physical symbol systems. *Cognitive Science* **4**: 135–83
Newell A 1982 The knowledge level. *Artificial Intelligence* **18**: 87–127
Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA
Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ
Newell A, Simon H A 1976 Computer science as empirical enquiry: symbols and search. *Communications of the ACM* **19**: 113–26
Norman D A 1981 Categorization of action slips. *Psychological Review* **88**: 1–15
Olson G M, Olson J S 1997 Research on computer supported cooperative work. In: Helander M, Landauer T K, Prabhu P V (eds.) *Handbook of Human–Computer Interaction*, 2nd rev. edn. Elsevier, Amsterdam, pp. 1433–56
Page M 2000 Connectionist modelling in psychology: a localist manifesto. *Behavioral and Brain Sciences* **23**
Pinker S 1984 *Language Learnability and Language Development*. Harvard University Press, Cambridge, MA
Port R, van Gelder T J 1995 *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, MA
Post E 1943 Formal reductions of the general combinatorial decision problem. *American Journal of Mathematics* **65**: 197–268
Psotka J, Massey D L, Mutter S A (eds.) 1988 *Intelligent Tutoring Systems: Lessons Learned*. Erlbaum, Hillsdale, NJ
Putnam H 1960 Minds and machines. In: Hook S (ed.) *Dimensions of Mind*. New York University Press, New York, pp. 138–64
Rosenblatt F 1958 The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* **65**: 386–408
Rumelhart D E, McClelland J L & the PDP Group 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA
Schunn C D, Crowley K, Okada T 1998 The growth of multidisciplinarity in the Cognitive Science Society. *Cognitive Science* **22**: 107–30
Searle J R 1990 Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences* **13**: 585–95
Shannon C 1948 A mathematical theory of communication. *Bell System Technical Journal* **27**: 379–423 , 623–656
Simon H A 1955 A behavioral model of rational choice. *Quarterly Journal of Economics* **69**: 99–118
Sloan Foundation 1978 *Cognitive Science 1978. Report of the State of the Art Committee*. Alfred P. Sloan Foundation, New York
Smolensky P 1988 On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1–23
Sperber D, Hirschfeld L 1999 Culture, cognition and evolution. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. cxi–cxxxii
Strube G, Janetzko D, Knauff M 1996 Cooperative construction of expert knowledge: the case of knowledge engineering. In: Baltes P B, Staudinger U M (eds.) *Interactive Minds*. Cambridge University Press, Cambridge, UK, pp. 366–93
Suchman L A 1987 *Plans and Situated Actions: the Problem of Human–Machine Communication*. Cambridge University Press, New York
Tomasello M, Kruger A C, Ratner H H 1993 Cultural learning. *Behavioral and Brain Sciences* **16**: 495–511
Turing A 1936 On computable numbers, with an application to the Entscheidungs problem. *Proceedings of the London Mathematical Society (Series 2)* **42**: 230–65; **43**: 544–546 (addendum)
Vera A H, Simon H A 1993 Situated action: a symbolic interpretation. *Cognitive Science* **17**: 7–48
Wiener N 1948 *Cybernetics*. Wiley, New York
Wilson R A, Keil F C (eds.) 1999 *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA
Winograd T 1972 Understanding natural language *Cognitive Psychology* **3**(1)

Wooldridge M, Jennings N R 1995 Intelligent agents: theory and practice. *Knowledge Engineering Review* **10**: 115–52
Zhang J J 1997 The nature of external representations in problem solving. *Cognitive Science* **21**: 179–217

G. Strube

Cognitive Science: Philosophical Aspects

Cognitive science emerged as a distinct field in the middle 1950s influenced by two important developments. First was the construction of digital computers and their capacity to perform operations that apparently require intelligent thinking. Second was Noam Chomsky's idea that linguistic capacities involve in some sense *knowledge* of and conforming to unconscious grammatical rules and that at least some of this knowledge is innate. The first suggested that the mind's operations could be understood on the model of a computer implementing an internally represented program (Turing 1950). The second that mental capacities are best describable in *intentional* terms like 'knowledge,' 'belief,' and 'following a rule.' Both of these developments rejected behaviorism (Skinner 1953). In its extreme forms behaviorism endorsed the view that human (and other animal) mental capacities are best understood in terms of causal connections between stimuli and responses and that learning is best understood as the change of such connections under reinforcement. Behaviorists tended to reject questions about the internal structures and processes that mediated stimulus and responses. They viewed intentional concepts as meaningless and/or unscientific.

Some mental states are attributed to a person, e.g. beliefs, memories, perceptions, desires, while others are properly attributed only to subpersonal parts or faculties, e.g. the language faculty and the visual system. Both kinds of mental states and processes involving them are *intentional*. 'Intentionality' refers to the fact that mental states *represent*. For example, the thought that New York is tropical represents New York as being tropical. It has long been observed that a thought can represent what doesn't exist (e.g., thoughts about Santa Claus) and misrepresent what does exist. Intentional mental states possess *semantic* properties. They refer and are evaluable as true or false. The thought that New York is tropical refers to New York and is false. Semantic features are shared with natural language expressions and by other kinds of representations (e.g., maps and pictures). For example, *the belief* that New York is tropical and 'New York is tropical' possess the same semantic content. It is widely held that the semantic features of nonmental items are ultimately derived from intentional mental states. Many theories in cognitive science assume that mental processes consist of mental states that are connected by virtue of their semantic contents. Such processes include ones that are 'rational' in that they tend to produce true beliefs or promote survival. Mental information processing consists of sequences of rationally related mental states.

Most cognitive scientists assume that psychological states are part of the natural physical order. However it is usually thought that cognitive science will, like other special sciences, employ its own taxonomy of states, and in particular one that differs from that of neurophysiology. The 'received' view is that psychological states are certain kinds of functional states. A functional state or property is one that is individuated in terms of a causal role. For example, what makes something a carburetor is its causal role in taking gas and air as inputs and producing a mixture of the two as output. In the case of psychological states the causal role involves causal connections to other psychological states, to stimuli, and to behavior. An important feature of functionalist accounts is that physically different kinds of structures can *realize* a given functional state by satisfying its functional specification. Thus carburetors can be made out of metal or plastic and minds out of brains or computers.

There are three big issues in the philosophy of mind. How is rational thinking possible? How can the mind represent the world? What is consciousness? Cognitive science has much to say about the first and has given new twists to the second and third. Below are discussions of these and related issues.

(a) Is cognitive science possible? There are philosophical traditions which claim that there can be no science of the mind. The main contemporary objection revolves around the idea that intentionality and rationality are *normative* categories and this disqualifies them from being the subject of scientific laws and causal explanations. Ruminations along these lines can be found in Wittgenstein (1953) and Ryle (1949) and later taken up by philosophers as diverse as Kripke (1982), Davidson (1980), McDowell (1994), and Quine (1960). Davidson's arguments have been especially influential among philosophers. He (1969) argues that in attributing beliefs, thoughts, desires, and other propositional attitudes to one another we are engaged in a project of *interpretation*. Further, he claims that interpretation is guided by a holistic 'principle of charity.' This principle dictates that when attributing mental states we *ought*, other things being equal, to maximize the subject's rationality. Attributions guided by charity are holistic since whether or not a belief or preference, etc. is rational depends on other beliefs, etc. The normativity of *rationality* makes intentional states normative as well. Davidson claims that *interpretation* is so different from the way physical properties are assigned as to make it impossible for there to be any *strict* laws connecting intentional psychological states with physical states or with each

other. But while Davidson's (and other such) arguments have been influential it is far from clear that they are sound or even whether if sound they really would undermine the employment of intentional vocabulary in scientific laws and explanations. In any case, cognitive scientists persist in employing intentional concepts in formulating explanations and theories. For example, there is a lively area of cognitive science concerning how people engage in everyday logical and statistical inference that proposes hypotheses concerning causes of errors, selection of conclusions, and so forth (Johnson-Laird 1983). The mental states and processes studied are, of course, intentional ones. The issues of whether the mind and what aspects of mentality can be scientifically studied and whether there are laws involving intentional states are more likely to be resolved by the success (or lack thereof) of the theories produced in the cognitive sciences than by philosophical argument.

Most cognitive scientists think that a science of the mind like other sciences contains laws, which explain psychological capacities and processes. The question arises of how such laws are related to biological laws. According to Fodor (1975), psychological laws are neurophysiologically implemented. In this view psychology is autonomous in that its taxonomy and laws are specific to it but every instance of a psychological law is also an instance of a more basic law or causal mechanism. If this is right then various kinds of mechanisms can implement psychological laws so psychology is not restricted to humans but can apply to Martians and computers. Cognitive scientists disagree about exactly how much can be learned about psychology by studying neurophysiology. 'Top down' theorists (Fodor 1981) think that very little beyond crude neural geography of mental capacities can be learned from neural sciences. At the other extreme, 'bottom up' theorists think quite a lot can be learned about the mind and perhaps even that neurophysiology should replace cognitive psychology.

(b) What is the status of folk psychology? The mental states people attribute to one another (and sometimes animals) include belief, knowledge, memory, desire, perception, and emotions. Normal humans are capable of attributing mental states with reliability and employing them in explaining other mental states and behaviors. Such explanations seem to conform to general principles like 'if a person wants q and believes that doing A is the only way to get q then unless she has a reason not to do A she will intend to do A' and also to somewhat more specific principles like 'if a person sees a friend coming then unless she has some special reason not to she will greet him.' The collection of such principles has come to be known as 'folk psychology.' One issue concerning folk psychology is whether it is approximately true. Fodor (1981, 1987) argues that it is and provides the appropriate starting place of a deeper theory of the mind. In contrast Churchland (1995) argues that folk psychological explanations are often mistaken or vacuous and predicts that a developed science of the mind will abandon it in much the same way that physics abandons folk physics. At the current stage of development many cognitive theories do employ folk psychological concepts (or refinements of them) and produce theories that sometimes explain folk psychological regularities.

(c) Are there unconscious and subpersonal mental states? Folk psychology recognizes that there are unconscious mental states and processes. For example, most of a person's memories are not present to consciousness and the mental processes a person engages in when driving a car might well be unconscious. Psychoanalytic theory, which to some extent has been appropriated by folk psychology, recognizes unconscious thoughts and desires that are not normally accessible to consciousness, at least not without therapeutic assistance. Theories in cognitive science go quite a bit further, often positing intentional states that are in principle inaccessible to consciousness and are properly attributed only to parts of the mind. For example, some psycholinguistic theories posit a language module that *cognizes* grammatical rules formulated in terms of concepts that are unknown to the person but in some way guide comprehension and production of her speech.

Some functional structures engage in mental processes relatively independently of others. Such structures are said to be *cognitively encapsulated modules*. The mental processing engaged in by many modules is not accessible to consciousness although the end product of the processing may be. For example, there is evidence that the mind contains a 'face detecting' module that is able to determine whether or not a person has seen a particular face previously. The picture of the mind that emerges from this is in which it consists of many modules each dedicated to a particular task (e.g., face recognition, speech production, character trait attribution) and a general reasoner which receives information from the modules and whose operations are (partly) consciously accessible. There are disagreements concerning the extent to which cognitive capacities are modularized. Some (e.g., Pinker 1995) seem to think that the mind is massively modularized, while others (Fodor 1983) ascribe much more importance to general reasoning capacities.

(d) What are propositional attitudes? Belief, knowledge, desire, and other folk psychological mental states are said to be *propositional attitudes*. The reason is the widely held view that the 'that' clauses used for attributing such states refer to propositions. There are various views about the nature of propositions but they all have in common that propositions are the part of the meaning of a sentence that is or determines the conditions under which the sentence is true. The question arises of exactly what it is to have a propositional attitude; for example, the belief that

New York is tropical. This question is often seen (e.g., Fodor 1981) as having two parts: (i) what is it to have a belief? (ii) what is it to have a belief that expresses a particular proposition? The most widely accepted answer to the first question is that a belief is a functional state, i.e., it has a certain causal role. One approach to answering the second question is that a belief expresses the proposition that p in virtue of involving a *representation* whose content is that of the intentional state. So to believe that New York is tropical is to be in a functional state (believing) that involves a representation that has the intentional content *that New York is tropical*. This provides a nice explanation of why beliefs are truth evaluable, are about things, can be involved in inferences, etc. It is because these are features of the representations that they contain. The view that the mind contains representations is usually called 'the representational theory of mind' or RTM. It should be cautioned that RTM is restricted to *explicit* beliefs. Implicit beliefs (e.g., your belief you had prior to reading this sentence that no giraffe is bigger than the empire state building) are dispositions to form explicit beliefs under suitable circumstances. This account of propositional attitudes is the beginning of an account of how mental states can have intentional content. It has its dissenters. Some philosophers (Dennett 1981) suggest that there may be no internal representations with the same content as that clause we use in attributing a belief, but it is still appropriate to attribute the belief because of the person's behavioral (including linguistic behavior) dispositions that themselves may involve states that represent at the subpersonal level.

(e) What is the nature of mental representations? Cognitive science is up to its neck (and beyond) in representations. Some representations are involved in propositional attitudes, others in relatively high-level unconscious cognitive systems like the language faculty, and others lower level systems that may implement the higher level systems. There are various views about the nature of these representations. One widely held view (Fodor 1975) is that most mental representations, at least those involved in propositional attitudes, belong to an internal language of thought, 'mentalese.' Mentalese contains basic expressions—predicates, names, connectives, etc. and rules of combination. The rules allow for the construction of a potential infinity of complex expressions from a finite basic vocabulary. There are important differences between a natural language like English and mentalese. The most important difference (as will be further discussed in item (g)) is that mentalese syntax is logical form. Whereas English can be used for communication, mentalese is used for thinking. Understanding a natural language can be accounted for in terms of processes that 'translate' natural language into mentalese. But understanding mentalese cannot be understood in the same way on the pain of regress. Rather mentalese is not literally 'understood' but rather is the language in which thinking and other mental processes take place. Whereas natural languages must be learned, many proponents of mentalese think that in some sense it is innate.

(f) What is thinking? Perception and rational thinking involve mental states that are semantically related. For example, a person sees Macguire swing at a ball and the ball go over the centerfield fence. He thinks that it is a home run number 62 and then that Macguire has broken Maris' home run record. On the representational theory of mind the process running from the visual perception to the thought that the home run record has been broken consists of many representations, e.g., a visual representation of shapes and colors, a perception that the ball is going over the fence, etc. These are not arbitrarily related but are related to each other by virtue of their *semantic* features. These semantic features involve relations to things external to the mind, e.g., to Macguire, the property of being a home run, and so on. How does the mind 'know' to go from a representation, which refers one thing to a representation that refers to a related thing if it has access only to the representations and not to their references? A closely related question is how can the mind engage in reasoning, which leads it from some true representations (say about light striking the retinas) to other true representations (say about the scene in front of the eyes). The computational theory of mind suggests answers to this question. It says that mental processes like thinking and perceiving are *computations* on mental representations. A computer is a device which is able to follow a program for manipulating representations on the basis of their syntax. So any relation which can be reduced to or encoded in syntactical relations can, in principle, be computed. Logical relations, e.g. logical implication, can be reduced to syntactical relations of sentences whose syntactic forms are logical forms (as is the case for Mentalese) computation can then account for logical inference. By encoding semantic features in syntax the computer can manipulate representations in ways that respect their semantics. Researchers in computer science have shown how many tasks that apparently require intelligence can be educed to computations.

According to the computational theory of mind (CTM), the mind is a computer and mental processes consist of computations on Mentalese representations. In this way the mind is able to manipulate representations that are semantically related. Of course there is a big difference between a computer and a mind. The computer is programmed by a human programmer who supplies interpretations for the symbols it manipulates. If the CTM is correct the mind that programs what the mind implements is a product of the structure of the brain and that, presumably, is at least partly dependent on evolution. The interpretation of the symbols that the mind manipulates is not provided by a programmer. Exactly what determines the semantic

features of mental representations is discussed in item (i).

Not every cognitive scientist is impressed by CTM as an account of mental processes. Some go along with the idea that mental processes involve a kind of computation but conceive of computation very differently from the TM account. They suggest that the mind has a connectionist *architecture*. I discuss this view in the next section. The most famous philosophical objection to CTM is due to John Searle (1980). Suppose that understanding a language, e.g. Chinese, involves following a program. Searle argues that this cannot be correct since it is possible for a person who knows no Chinese but is able to follow program instructions just like a computer implements the program. The program is written so that questions in Chinese are input and answers in Chinese are output. Searle observes that although the man implements the program he clearly has no clue as to the meanings of the Chinese symbols. There have been many replies to this objection. Perhaps the most convincing from proponents of CTM is that implementing a program is necessary but not sufficient for language understanding. The language must also be translated into the person's Mentalese. Of course this still leaves open the question of what it is for a symbol of Mentalese to represent, which will be discussed in item (i).

A different worry about CTM is whether it can account for certain kinds of reasoning, specifically inductive inference. Inductive reasoning involves considering various hypotheses and coming up with the one that is best supported by the evidence. We employ it in producing explanations, identifying causes, and so forth. For example, Sherlock Holmes solved a case when he realized that the fact that no dog barked was evidence that the murderer was known to the dogs. The worry, which is sometimes called 'the frame problem' (Pylyshyn 1987), is that no computational program can realistically perform this kind of task. In inductive reasoning almost any bit of information may be relevant. We seem to have the ability to survey a great deal of what we know and come up with what is evidentially relevant. But a program that operated by having to survey a vast number of representations evaluating them for relevance would seem to be completely impractical. There are too many computations to be performed. One response to this is to think that CTM may provide good accounts of the information processing that occurs in mental modules but that it is not very good at accounting for the mental capacities of a general reasoner. Thus those who like CTM are attracted to the view that the mind is massively modular.

(g) Classical or connectionist architecture? In the classical account of computation discussed in item (f), the architecture of the mind is that of a classical computer (or system of such computers) and mental processes are operations on linguistic-like representations by virtue of their syntactic forms. There is an alternative account of cognitive structure and computation that has been developed called 'connectionist architecture.' Roughly, a connectionist system consists of a network of *nodes* joined together in a pattern of connections. Each node can be activated (or activated to a certain degree) and can receive signals and send signals to certain connected nodes. Whether or not a signal travels via a connection from node A to node N depends on the weight of the connection. Some nodes are activated by external stimuli (input nodes) and others send signals outside the network. At any time the state of the network is determined by the weights of the connections and the activation of the nodes. Signals that activate the input nodes are propagated throughout the system to the output nodes. Thus connectionist systems can be thought of as computing certain outputs given certain inputs. Further, the connectionist network can be 'trained' by altering the connection weights depending on whether or not a given output is 'appropriate' for a given input. Connectionist networks have been constructed that do a number of tasks e.g. recognize letters of the alphabet, 'read' text, recognize faces, and so forth by such training.

Connectionist cognitive architecture is apparently quite different from a classical computer. There is no 'executive' that is following a program. Computations are not performed on sentences but on the totality of connections among nodes. Although the state of a connectionist system can be thought of as a representation—say as representing that the cat is on the mat—unlike sentences of mentalese it needn't contain any parts that correspond to 'the cat,' 'is on,' and 'the mat.' Proponents of connectionism think that it provides a model of mental states and processes that is more plausible than classical accounts. One reason is that some find it difficult to believe in mentalese. Another is that it seems natural for connectionist systems to implement vague concepts since a connectionist system can be trained to respond in a graded manner. The holism of connectionist representation is also appealing to some and there are suggestions that its holistic features may help with the frame problem mentioned in item (f). Finally, connectionist networks are reminiscent of assemblies of neurons and so strike some as biologically realistic. Proponents of classicism point out that connectionism is more than reminiscent of behaviorism. Like behaviorism it is an *associacionist* psychology that construes mental processes in terms of establishing and modifying associations. Although it is in a sense holistic it is far from clear how that will help accounting for inductive inference. In fact critics of connectionism set up a dilemma (Fodor and Pylyshyn 1988, Fodor and McLaughlin 1990): either the connectionist architecture implements a classical architecture, in which case it is not really an alternative to classicism, or it fails to account for essential features of thought. These features are *productivity* and *syst-*

ematicity. Productivity involves the fact that once a thinker has basic concepts she is able to produce a potential infinity of novel thoughts involving those concepts. Systematicity is the feature that any thinker who can think a thought can think related thoughts that apparently have the same components. For example, if one can think Jack loves Jill then one can also think Jill loves Jack. These features are easily accounted for by classicism since the thoughts correspond to syntactically structured representations. But a connectionist system can be capable of being in a state that represents that Jack loves Jill without being capable of being in a state that represent Jill loves Jack since it need not have parts that correspond to *Jack*, *loves*, *Jill*. If the connectionist system does have such parts then it is merely implementing a classical system.

(h) What are concepts? The *concept* plays an important role in the cognitive sciences. Thoughts (beliefs, memories, desires, etc.) are composed of concepts and so what mental processes a thinker can engage in depends on what concepts she possesses. Developmental psychology is interested in how people acquire concepts and whether some concepts are innate. There are various views about the nature of concepts. Advocates of RTM think of concept tokens as representations but there is a wide diversity of views on what makes a particular representation a particular concept, say the concept *horse*. One view is that a concept is something like a definition. For example, the concept *horse* may be the definition *is a large land mammal that has been domesticated for riding*. To possess the concept *horse* is to know the definition. This view has come under much deserved criticism. One problem is that not all concepts can have definitions without circularity. More serious even is that most words (and the concepts they are associated with) do not seem to have definitions at all. There are horses that are not large and large animals that have been domesticated for riding (e.g., elephants) that are not horses. Another view is that concepts are prototypes. A prototype consists of a core exemplar—a representation of something that is a paradigm example of the concept—and then a similarity metric that determines how close something is to the paradigm. For example, the concept *bird* consists of the representation *robin* and a metric that makes eagles pretty good birds and penguins pretty bad ones. But while there is evidence that thinkers do judge instances of a concept as better or worse examples, the account faces some of the same difficulties faced by the definition account. A somewhat more general approach considers the inferences that a thinker is disposed to make concerning thoughts containing a concept as individuating the concept. Such 'conceptual role' theories of concepts face a dilemma. Either all of the inferences involving the concept are individuative of it (holism) or only some are (molecularism). If the first then, as our beliefs change so do our concepts and it is unlikely that two people ever share the same concept. If the second then the question arises of what makes some inferences concept constituting. There are arguments in the philosophical literature (Quine 1960, Fodor and LePore 1992) that there is no principled distinction between the two but also some proposals (Peacocke 1992) for how to make the distinction. Finally, there is the view that concepts are expressions in mentalese that are individuated by their syntax and by their reference (Fodor 1998). This view allows for thinkers with very different beliefs to share the same concept. But it also allows for the bizarre possibility of someone possessing the concept horse while believing that horses are edible fruits.

Cognitive theories that posit innate knowledge are also committed to the innateness of the concepts that constitute the knowledge. Some cognitive scientists go much further and claim that many of our concepts are innate. One reason for this is the difficulty in accounting for how concepts can be learned. As Fodor (1980) observed, they cannot be learned by testing various hypotheses about them since the formulation of the hypotheses already requires possessing the concept. At one time Fodor thought that this line of argument showed that even the concept *carburetor* is innate. Fodor has since moderated his view but there is no consensus concerning how concepts are acquired.

(i) How does the mind represent the world? What makes a component of a mental state a representation is that it possesses semantic properties, e.g., it refers, has truth-value, and so on. But what exactly determines that a given representation possesses a certain semantic property. The Cartesian tradition generally thought that intentionality is a distinct and basic feature of mental substance. But most philosophers of cognitive science who think that there are mental representations think that whatever determines semantic features has to be within the realm of natural science. On the view that was once widely held in philosophy that concepts are images the answer to this question is that resemblance makes for representation. But even cognitive scientists who posit mental images do not think that these literally resemble their references. The two views that are currently most widely advocated are *informational semantics* and *teleological semantics* or some combination of the two Millikan 1984, Fodor 1987, 1990, Loewer 1998). Simplified informational semantics says that the fact that a certain state carries certain information under certain circumstances or is reliably caused by certain properties under certain circumstances determines its semantic properties. *Teleological semantics* says that semantic properties of a representation are determined by its biological function. A simplified combined view is that the function of carrying certain information under certain circumstances determines representations of semantic properties. For example, it is not implausible that there is a certain system of a frog's brain with the function of being in a particular state R

if and only if a fly is nearby when the circumstances are normal (e.g., in a pond, good light, etc.). If the circumstances are normal then an occurrence of R in the frog's brain carries the information (to other parts of the frog's brain) that a fly is present. This kind of account fits very nicely with the view that the mind is a kind of information processor. But whether it can be developed so as to provide a plausible account of the semantics of the mental representations involved in human thought is a big and very open question.

See also: Cognitive Modeling: Research Logic in Cognitive Science; Cognitive Neuroscience; Cognitive Psychology: Overview; Cognitive Science: History; Cognitive Science: Overview; Consciousness and Sensation: Philosophical Aspects; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Knowledge (Explicit and Implicit): Philosophical Aspects; Marr, David (1945–80); Reference and Representation: Philosophical Aspects

Bibliography

Block N 1995 On a confusion about a function of consciousness. *Behavioral and Brain Sciences* **18**: 227–47

Chomsky N 1954 *Syntactic Structures*. Mouton, The Hague

Chomsky N 1959 A Review of Skinner's Verbal Behavior. *Language* **35**: 26–58

Churchland P M 1995 *The Engine of Reason, The Seat of the Soul*. MIT Press, Cambridge, MA

Davidson D 1980 *Essays on Actions and Events*. Oxford University Press, Oxford, UK

Dennett D 1981 *Brainstorms*. MIT Press, Cambridge, MA

Dennett D 1994 *Consciousness Explained*. Little Brown, New York

Descartes R 1641/1970 Meditations on first philosophy. In: ES Haldane, Ross GRT (trans.) *The Philosophical Works of Descartes*, Cambridge University Press, Cambridge, UK, Vol. 1, pp. 131–200

Fodor J A 1975 *The Language of Thought*. Crowell, New York

Fodor J A 1981 *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. MIT Press, Cambridge, MA

Fodor J A 1983 *The Modularity of Mind*. MIT Press, Cambridge, MA

Fodor J A 1987 *Psychosemantics*. MIT Press, Cambridge, MA

Fodor J A 1990 *A Theory of Content: And Other Essays*. MIT Press, Cambridge, MA

Fodor J A 1998 *Concepts: Where Cognitive Science Went Wrong*. University Press, Oxford, UK

Fodor J A, LePore E 1992 *Holism: a Shopper's Guide*. Blackwell, Oxford, UK

Fodor J A, McLaughlin B 1990 Connectionism and the problem of systematicity. *Cognition* **35**(2): 185–204

Fodor J A, Pylyshyn Z 1988 Connectionism and cognitive architecture: a critical analysis. In: Pinker S, Mehler J (eds.) *Connections and Symbols*. MIT Press, Cambridge, MA

Johnson-Laird P 1983 *Mental Models: Towards a Cognitive Science of language, Inference, and Consciousness*. Cambridge University Press, Cambridge, UK

Kossylyn S M 1980 *Image and Mind*. Harvard University Press, Cambridge, MA

Kripke S A 1982 *Wittgenstein on Rules and Private Language*. Harvard University Press, Cambridge, MA

Loewer B 1998 Guide to naturalizing semantics. In: Hale B, Wright C (eds.) *The Companion to the Philosophy of Language*. Blackwell, Oxford, UK

Loewer B, Georges, Rey 1991 *Meaning in Mind*. Blackwell, Oxford, UK

Marr D 1982 *Vision*. Freeman and Co.

Millikan R 1984 *Language, Thought, and Other Biological Categories*. MIT Press, Cambridge, MA

McDowell J 1994 *Mind and World*. Harvard University Press, Cambridge, MA

Nagel T 1979 *Mortal Questions*. Cambridge University Press, Cambridge, UK

Peacocke C 1992 *A Study of Concepts*. MIT Press, Cambridge, MA

Pinker S 1997 *How the Mind Works*. Norton, New York

Pylyshyn Z W 1984 *Computation and Cognition*. MIT Press, Cambridge, MA

Pylyshyn Z W 1987 *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Ablex, Norwood, NJ

Rey G 1997 *Contemporary Philosophy of Mind*. Cambridge, MA

Ryle G 1949 *The Concept of Mind*. Hutchenson's University Library, London

Quine W 1960 *Word and Object*. Technology Press of MIT, Cambridge, MA

Searle J 1980 Minds, brains, and programs with commentaries. *The Behavioral and Brain Sciences* **3**: 417–57

Searle J R 1992 *Rediscovery of the Mind*. MIT Press, Cambridge, MA

Skinner B P 1953 *The Science of Human Behavior*. Macmillan, New York

Smolenski P 1988 On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1–74

Turing A 1950 Computing machinery and intelligence. *Mind* **59**: 433–60

Wittgenstein L 1953 *Philosophical Investigations* [trans. Anscombe GEM]. Macmillan, New York

B. Loewer

Cognitive Styles and Learning Styles

1. Style Differences in Cognition and Learning

The chorus line of a popular song once claimed that … *it's not what you do but the way that you do it*. This idea—making the *how* as important as the *what*—is intriguing. Furthermore, as the song insists, the character of an individual is invariably woven into how a task is completed. This can be seen in any number of human endeavors, for example, sport, art, handwriting, thinking, learning, even conversation.

In any performance then, a sense of person, as well as context, combines to produce a typical pattern, hallmark, leit-motif, signature, or style. The very same idea underlies the suggestion that individuals possess a personal way of thinking (cognitive style) or learning (learning style). The following discussion will introduce the construct of style differences in cognition and learning and consider its significance for lifelong learning in both the world of education and workplace.

2. *The Style Construct in Psychology*

The term *construct* refers to a psychological idea or notion. Examples of constructs are intelligence, personality, or self-concept. A *style construct* appears in a number of academic disciplines—in psychology it has been used in a number of different areas such as personality, cognition, communication, motivation, perception, learning, and behavior.

The theory of style, unfortunately, has been characterized by a tendency for researchers to (a) work in isolation; (b) develop their own instruments for the assessment of style; and (c) the creation of independent style labels with little reference to the field. A widespread use of the term 'style' has led to a number of different definitions and terminology. Consequently, those interested in validity or verifiability, and an accepted nomenclature for a theory of style, have faced considerable difficulty. The idea of style in educational psychology, nonetheless, is recognized as a key construct of individual differences in human performance.

3. *Cognitive Styles*

While Allport (1937), in work which developed the idea of 'lifestyles,' was probably the first researcher to deliberately use the 'style' construct in association with cognition, the following key areas of psychology contributed to an emerging field of cognitive style.

3.1 *Perception*

Experimental work—reflecting an emphasis on the 'regularities' of information-processing which were derived from the German gestalt school of perceptual psychology—led to an early development of the 'style construct' of field dependence–independence (Witkin et al. 1971).

Individuals were found to rely upon the surrounding 'field' or 'context' to a greater or lesser extent, when reorienting an object relative to the vertical. This was subsequently found to correlate with competence in disembedding shapes from a field and experimental participants were found to either rely heavily on the field for orientation or shape discrimination (field-dependent) or little or not at all (field-independent).

3.2 *Cognitive Controls and Cognitive Processes*

A second significant influence in the development of cognitive style was the study of cognitive processes related to individual adaptation to the environment, exemplified by the work of Gardner and co-workers at the Messinger Clinic in the USA. This work was shaped, originally, by psychoanalytic theories of ego psychology—which was typified by studies focusing upon variables in ego adaptation to the environment.

This led to the identification of several cognitive processes including perceptual attitudes, cognitive attitudes, and cognitive controls. Further work related to this area led to several stylistic labels and models, supporting the general notion of a cognitive style (see Messick 1976).

3.3 *Mental Imagery*

A third key influence in the development of cognitive style reflected work looking at mental representation. Early in the scientific study of psychology, attention was given to the notion that some people have a predominantly verbal way of representing information in thought, while others are more visual or imaginal. Paivio (1971) further developed this notion with a dual coding measurement of mental imagery. Riding and Taylor (1976) identified, as fundamental to the construct of cognitive style, the verbal-imagery dimension of cognitive style.

3.4 *Personality Constructs*

A fourth and separate influence on the field of cognitive style involved researchers utilizing personality-based constructs to develop a model of learning style (Myers 1978). The most influential model was the Myers–Briggs Type Indicator, developed from Jung's typology of personality constructs and 'psychoanalytic ego psychology' (Jung 1923). It offered an alternative model of cognitive style in contrast to those flowing from cognitive psychology.

4. *The Further Development of Cognitive Styles*

A contemporary resurgence of interest in style differences over the last three decades has resulted in three distinct developments, all involving the generation of new models of cognitive style or learning styles.

The first development involved researchers independently constructing new models of individual difference in various aspects of cognitive functioning. This approach tended to conceptualize styles as the discovery of new psychological phenomena, for example, styles of thinking, intuition, creativity, decision-making, and motivation.

Examples of some of this work generating additional labels of cognitive style include a model of perceptual style (Gregorc 1982); the Adaptor–Innovator cognitive style of decision-making (Kirton 1994); and the Assimilator–Explorer cognitive style of creativity (Kaufmann 1989). Another more elaborate model of mental self-government presented by Sternberg (1996, 1997) represented a theory of style derived from notions of government. According to this theory, people can be understood in terms of mental government, that is, processes of function, form, level, scope, and learning, which impact upon thinking and learning.

The second development reflected work aimed at a synthesis of theory and a consensus in the understanding of cognitive style. This approach was characterized by a focus upon the construct validity of cognitive style and its application. The work of Curry (1987) and Rayner and Riding (1997), for example, both attempted to synthesize or integrate existing theory of cognitive and learning styles.

The work of Riding over 20 years involved a reclassification of cognitive style models (see Riding and Rayner 1998). The structure of cognitive style was defined as two-dimensional, comprising the Wholist–Analytic style dimension, relating principally to cognitive organization and the Verbal–Imagery style dimension, relating principally to mental representation.

An individual's cognitive style was defined as a person's tendency to process information wholistically or analytically, that is, either as a whole piece or piecemeal, while at the same time mentally representing information using imagery or language. While each dimension was thought to be independent, they were conceptualized as continua and it was not suggested that an individual could only use one or the other way of thinking.

The further development of a computer-based assessment for cognitive style analysis (CSA) by Riding reflected a deliberate attempt to integrate both the Wholist–Analytic and Verbal–Imagery dimensions of cognitive style (see Riding 1991). An extensive number of empirical studies over a number of years were conducted using the CSA at the University of Birmingham and provided evidence to support this model of cognitive style (see Riding and Rayner 1998).

5. *Learning Styles*

A third more widespread development in contemporary work on style differences in cognition and learning led to the generation of yet more labels, which were described as models of learning styles. This learning-centered tradition of 'style' is arguably distinguished by four major features:

(a) the intention of developing new concepts of learning to reduce reliance upon tests of intelligence or ability;

(b) a focus on the learning process and achievement;

(c) a primary interest in the effect of individual differences upon pedagogy;

(d) a parallel construction of new assessment instruments and models of learning style.

The primary concern for educationists working within this learning-centered tradition lay with the process of learning and its context. It focused on individual differences in the process of learning rather than within the individual learner. Models of learning styles in this tradition included the following four groups.

5.1 *Models Focusing on the Learning Process—Based on Experiential Learning*

These derived from a theory of experiential learning, and the most influential example was the work of Kolb (1976), who described learning style as the individual's preferred method for assimilating information, in an active learning cycle. Kolb constructed a two-dimensional model comprising perception (concrete/abstract thinking) and processing (active/reflective information processing) as fundamental aspects of an experiential learning cycle.

5.2 *Models Focusing on the Learning Process—Based on Orientation to Study*

These derived from a theory of information processing and learning processes, and the most influential example was developed by Entwistle called the 'Approaches to Study Inventory.' Entwistle (1981) found that approaches to study often reflected either a surface or deep engagement with the study task. This was later extended to include four key orientations to study: meaning, reproducing, achieving, and holistic. The model was further refined as an integrated conception of the learning process, which described a series of actions linked to specific learning strategies identified in his original model.

5.3 *Models Focusing on Instructional-preference*

These set out to measure a range of environmental or instructional factors affecting an individual's learning behavior. A leading example of this type was the Learning Styles Inventory (LSI) developed by Dunn et al. (1989). The learning style elements identified in the LSI were: *environmental stimulus* (light, temperature); *emotional stimulus* (persistence, motivation); *sociological stimulus* (peers, adults); *physical stimulus* (perceptual strengths, time of day—morning vs. afternoon); and *psychological stimulus* (global/analytic, impulsive/reflective).

5.4 Models Focusing on Cognitive Skills, and Learning Strategy Development

The fourth group of learning style labels focused on an individual's developing cognitive ability and repertoire of cognitive skills or ability to learn, together with related behavioral characteristics, which were understood to comprise an individual's learning profile. Learning style was typically perceived as a multimodal construct and understood to describe a range of intellectual functioning relating to the learning activity.

An example of this type of model was the Learning Styles Profile developed by the North American Association of Secondary School Principals (Keefe 1988). This style construct described 24 key elements in learning style, grouped into three categories: cognitive skills—relating to aspects of information processing; perceptual responses—encompassing perceptual responses to data; study and instructional preference—referring to motivational and environmental elements affecting learning preferences.

6. Cognitive Style or Learning Styles

The models in the learning-centered tradition shared several limitations. First, they reflected a construct that by definition was not stable—it was grounded in process and therefore susceptible to rapid change. Second, they did not describe a developmental rationale for the concept of learning style nor easily correspond to other models of assessment, thereby suggesting a problem for conceptual validity. Third, they attracted peer criticism for lacking psychometric rigor and a systematically developed theory supported by empirical evidence (see Grigerenko and Sternberg 1995).

The learning-centered tradition, however, reflected a continuing need for a theory of individual differences which could be applied to the learning context. It also reinforced previous work in the area of cognitive style and pointed to the potential for profiling the personal learning style of an individual (see Rayner 2000).

7. Implications of Style for Lifelong Learning

At the 1997 Seventh International Conference on Thinking, Howard Gardner argued that in the not too distant future people will look back to the end of this millennium, and laugh at the 'uniform school.' They will be greatly amused, he suggested, by the idea that educationists actually believed they could teach the same things to all children at the same time and in the same way. To believe that the *uniform school* can provide efficient or effective education, he concluded, was to endorse educational failure!

The extent to which an awareness of learning style or the self as a learner is currently considered and managed within the educational context raises key questions for the design of instruction and pedagogy, including a consideration of:

(a) assessment-based learning;
(b) differentiation in the curriculum;
(c) learning method routines;
(d) professional development.

Each of these approaches, if adopted with an eye to considering the benefit of a pedagogy which is style-friendly, encourages interactive learning, builds upon the principles of individual difference, and adopts the idea of developing a learning expertise within the learner, will provide a foundation for lifelong learning (see Rayner 2000).

Implicit in all of this work, and equally relevant to the classroom as to the workplace, is the notion of the *matching hypothesis*. The full value and significance to the professional of cognitive and learning styles rest ultimately with the belief that if it is possible to make a better match between person and environment, then performance will improve and achievement will be enhanced. Moreover, formal education might be made more effective by matching style to materials, to presentation, mode and structure, through nurturing strategy development to maximize style effectiveness.

The delivery of a curriculum, albeit for schooling or workplace training, will improve with an increasing depth of differentiation and a match between individual differences with targeted learning or activity. Such an approach will build upon personal strengths, sow seeds of success, and reap the benefits of learning enhancement.

The final word in offering an overview of style differences in thinking, learning, and behavior is left to Sternberg (1996, p. 363). He succinctly stated that in the world of learning and education, '*styles matter*'! This view mirrors that of many workers in the field who remain interested in knowing more about those individual differences which affect human performance.

See also: Mental Imagery, Psychology of; Mental Representations, Psychology of; Metacognitive Development: Educational Implications; Self-regulated Learning

Bibliography

Allport G W 1937 *Personality: A Psychological Interpretation*. H. Holt and Co., New York

Curry L 1987 *Integrating Concepts of Cognitive or Learning Style: A Review with Attention to Psychometric Standards*. Canadian College of Health Service Executives, Ottawa, ON, Canada

Dunn R, Dunn K, Price G E 1989 *Learning Styles Inventory*. Price Systems, Lawrence, KS

Entwistle N J 1981 *Styles of Learning and Teaching: An Integrated Outline of Educational Psychology for Students, Teachers and Lecturers*. Wiley, Chichester, UK

Gregorc A R 1982 *Style Delineator*. Gabriel Systems, Maynard, MA
Grigerenko E L, Sternberg R J 1995 Thinking styles. In: Saklofske D H, Zeidner M (eds.) *International Handbook of Personality and Intelligence*. Plenum, New York
Jung C G 1923 *Psychological Types*. Harcourt Brace, New York
Keefe J W 1988 *Profiling and Utilising Learning Style*. National Association of Secondary School Principals, Reston, VA
Kaufmann G 1989 *The Assimilator-Explorer Inventory*. University of Bergen, Bergen, Norway
Kirton J W (ed.) 1994 *Adaptors and Innovators*, 2nd edn. Routledge, London
Kolb D A 1976 *Learning Style Inventory: Technical Manual*. Prentice-Hall, Englewood Cliffs, NJ
Messick S 1976 *Individuality in Learning*, 1st edn. Jossey-Bass, San Francisco
Myers I 1978 *Myers–Briggs Type Indicator*. Consulting Psychologists Press, Palo Alto, CA
Paivio A 1971 Styles and strategies of learning. *British Journal of Educational Psychology* **46**: 128–48
Rayner S 2000 Reconstructing style differences in thinking and learning: profiling learning performance. In: Riding R, Rayner S (eds.) *International Perspectives on Individual Differences*. Ablex, Stanford, CT, Vol. 1
Rayner S, Riding R 1997 Towards a categorisation of cognitive styles and learning styles. *Educational Psychology* **17**: 5–28
Riding R J 1991 *Cognitive Styles Analysis*. Learning and Training Technology, Birmingham, UK
Riding R J, Rayner S G 1998 *Cognitive Styles and Learning Strategies*. David Fulton Publishers, London
Riding R J, Taylor E M 1976 Imagery performance and prose comprehension in 7 year old children. *Educational Studies* **2**: 21–7
Sternberg R J 1996 Styles of Thinking. In: Baltes P B, Staudinger U M (eds.) *Interactive Minds*. Cambridge University Press, Cambridge, UK
Sternberg R J 1997 *Thinking Styles*. Cambridge University Press, Cambridge, UK
Witkin H A, Oltman P, Raskin E, Karp S 1971 *A Manual for Embedded Figures Test*. Consulting Psychologists Press, Palo Alto, CA

S. G. Rayner

Cognitive Theory: ACT

ACT is a theory of human cognition that posits particular ways of representing knowledge and the mechanisms by which such knowledge is acquired and used (Anderson 1983, 1993, Anderson and Lebiere 1998). The theory is implemented as a computer simulation system that generates the theory's predictions, thereby facilitating quantitative comparisons with experimental data.

1. *What's in a Name?*

The etymology of the acronym ACT is the subject of some debate. The original definition, Adaptive Control of Thought, has not been consistently used, and several books written about the theory suggest alternative candidates, e.g., *The Adaptive Character of Thought* (Anderson 1990) and *The Atomic Components of Thought* (Anderson and Lebiere 1998). However, these publications postdate ACT's modest beginnings by more than a decade, making the simpler moniker A Cognitive Theory seem the most parsimonious answer.

2. *Architectures and Models*

Among computational systems designed to model cognition, there is a critical distinction between a cognitive *architecture* and a cognitive *model* built within an architecture. A cognitive architecture defines a specific way of representing knowledge and a fixed set of mechanisms for processing knowledge. A cognitive model, on the other hand, specifies the knowledge that is required to perform a particular task. Any one architecture supports a wide variety of models, all of which use the same mechanisms to capture behavior in different tasks, just as the brain presumably employs a common set of mechanisms across a variety of tasks.

The ACT theory is a cognitive architecture. It provides mechanisms for the retrieval and learning of facts (declarative knowledge) and the selection, application, and learning of skills (procedural knowledge). All cognitive models built within ACT share these mechanisms; what differs across ACT models is the task-specific knowledge (i.e., the facts and skills themselves) input to the system. ACT models have successfully fit behavioral data across a wide variety of tasks, including arithmetic, navigation, categorization, and game playing (see Table 1). The success of these ACT models across such varied domains provides support not only for the models themselves but for the explanatory power of the ACT architecture.

This article focuses on the current version of the ACT theory, called ACT-R (Anderson and Lebiere 1998). Nevertheless, a historical sketch of the theory's evolution and how it relates to contemporary research are presented as well as a brief discussion of future research issues.

3. *Basic Features of the Theory*

The four main tenets of the ACT theory are as follows:

(a) the ability to perform a complex task can be decomposed into separate pieces of knowledge;

(b) these pieces of knowledge are learned through experience, i.e., using a piece of knowledge is akin to practice;

(c) at any given time, the current focus of attention (also called the goal) influences what knowledge is used; and

(d) there are two types of knowledge, *declarative* (for facts) and *procedural* (for skills), that have distinct representations and learning mechanisms.

Table 1
A list of tasks and phenomena modeled with the ACT theory

1	Visual search including menu search
2	Subitizing
3	Dual tasking including Psychological Refractory Period (PRP)
4	Similarity judgments
5	Category learning
6	List learning experiments
7	Paired-associate learning
8	The fan effect
9	Individual differences in working memory
10	Cognitive arithmetic
11	Implicit learning (e.g., sequence learning)
12	Probability matching experiments
13	Hierarchical problem solving tasks
14	Strategy selection
15	Analogical problem solving
16	Dynamic problem solving tasks including military command and control
17	Learning of mathematical skills including interacting with Intelligent Tutoring Systems (ITSs)
18	Development of expertise
19	Scientific experimentation
20	Game playing
21	Metaphor comprehension
22	Learning of syntactic cues
23	Syntactic complexity effects and ambiguity effects
24	Dyad communication

Declarative knowledge is represented as nodes in an associative network. The more activated a given node, the more easily the corresponding fact can be accessed. Each node has an associated base-level activation B that increases each time the fact corresponding to that node is accessed (learning) and decays with time (forgetting) according to the equation

$$B = \ln\left(\sum_j (t - t_j)^{-d}\right) \tag{1}$$

where t is the current time, t_j is the time of the jth use of the node, and d is a global decay rate. The quantity B offers a summary description of a node's past history of use and hence a reasonable estimate of its likelihood to be needed in the future.

Each link between two nodes in the network has a continuous-valued quantity S that measures the strength of association between those two nodes. The current focus of attention works by selecting a subset of nodes in the network to be attended. Each attended node gets a share of a limited amount of attentional activation W. (Because attentional activation is limited, the more nodes in the focus, the less attentional activation for each.) This activation then propagates from each attended node to related nodes in the network in proportion to the corresponding link strengths. This attentional activation produces context effects because only those nodes related to the current focus of attention receive extra activation.

The *total* activation at node i is the sum of its base-level activation B_i and the attentional activation it receives from nodes in the focus of attention:

$$A_i = B_i + \sum_j W_j S_{ji} \tag{2}$$

where the sum is over attended nodes j. Retrieval of facts is determined by the total activation of the corresponding nodes. Specifically, the node that is retrieved is the one whose total activation plus some added noise is highest, given that this sum is above a global threshold. This added noise represents stochasticity in the system. The time it takes to recall this node decreases exponentially as a function of its (noisy) activation which, combined with Eqn. (1), produces a power-law speedup with practice (see *Learning Curve, The*).

Procedural knowledge is represented by a set of condition–action pairs called production rules (e.g., IF the goal is to add two numbers a and b, and the fact that c is the sum of a and b can be recalled, THEN say c is the answer). Each production has several associated quantities that reflect how useful it has been in past applications. These quantities are learned by experience and used to compute an estimate of the expected gain from applying each rule. For example, the cost associated with a particular production rule is the weighted average of the rule's past costs of application (measured in units of time) and a prior estimate of cost. When several rules are candidates at the same time, the one whose expected gain plus some added noise is highest gets selected. In this way, production rules that have been more useful (i.e., more successful and less costly) in the past are more likely to be selected. In addition, each piece of procedural knowledge is strengthened with each application, and that strength decays with time. A production rule's speed of application increases exponentially with strength.

The ACT theory also posits more complicated mechanisms by which production rules and declarative nodes are initially created. In both cases, the goal plays an important role in the form that new knowledge takes. These separate mechanisms highlight ACT's distinction between acquiring new pieces of knowledge and refining the continuous quantities associated with each piece of knowledge. Moreover, these two modes of knowledge representation—symbolic (nodes, production rules) and sub-symbolic (activations, costs, strengths)—make ACT a *hybrid* system. This

distinguishes it from wholly symbolic (e.g., Soar) and wholly sub-symbolic (e.g., connectionist) systems.

4. Historical Development

ACT has been under development since the 1970s. It began as a theory of semantic memory and now encompasses learning, memory, problem solving, attention, perception, and action. The following provides a brief sketch of its development and places each version of the theory in its historical context. Until the 1970s, mathematical models were the typical formalism used to describe and predict cognitive psychological phenomena. A disadvantage of mathematical models, however, is that they are limited in the complexity of processes they can describe. Thus, to provide a mechanistic account of complex cognitive processes, computational models began to be developed. The first well-defined version of the ACT theory, called ACTE (Anderson 1976), was one such model. It introduced the distinction between declarative (i.e., factual) and procedural (i.e., skill-based) memory and the notion of declarative activation.

By the 1980s, however, some researchers were questioning whether the development of computational models (and the theories they implemented) was sufficiently constrained to produce reasonable models of the 'true' underlying representations and processes (e.g., Anderson 1978, Newell 1973). One approach that reduces this problem involves developing models within a cognitive architecture, where a fixed set of representations and mechansims are used to test a variety of models. That is, models are constrained to work within the strictures of the architecture. The next version of the ACT theory, ACT* (Anderson 1983) was a cognitive architecture, like others developed around this time (see also *Cognitive Theory: SOAR*; Newell 1990). ACT* extended its predecessor by specifying an activation calculus for declarative knowledge and a new mechanism for acquiring procedural knowledge.

Besides using ACT* to model a variety of task domains—from language processing to paired-associate learning—this version of the theory was applied to the practical problem of improving computer-aided instruction. So-called 'intelligent' tutoring systems were built based on ACT* cognitive models of algebra problem solving, geometry theorem proving, and computer programming (e.g., Anderson et al. 1989, 1990). Because these models could solve the required problems in each domain, they enabled the corresponding tutoring systems to follow students' problem-solving, give feedback when students made a mistake, and offer hints when students were confused. Moreover, because these systems tracked the steps students were taking and hence the knowledge they were using, specific predictions of the theory could be generated and tested in this scaled-up, real-world learning context.

The latest incarnation of the ACT theory, ACT-R (Anderson 1993, Anderson and Lebiere 1998), was developed in the 1990s, inspired by a rational analysis of cognition (Anderson 1990). Among other things, ACT-R includes a refined activation calculus and a more plausible mechanism for acquiring procedural knowledge. These changes were designed to reflect the way human cognition adapts to the structure of the environment. In addition, ACT-R has been put to the toughest challenges in testing its fidelity to empirical data: ACT-R models have been able to fit fine-grained, multivariate data simultaneously across several experiments (e.g., Anderson and Matessa 1997), and they have captured patterns of performance in individual subjects across tasks (e.g., Lovett et al. 2000).

In sum, the ACT theory has evolved into its current form by virtue of the guiding force of several kinds of constraints. Throughout ACT's development, experimental data have been used to test the veridicality of the theory. In some cases, this empirical constraint has invoked a reevaluation of some aspect of the theory (e.g., single-trial learning of production rules). More generally, even when the theory's predictions have been met, refinements were made so that more detailed, fine-grained datasets could be modeled. Besides these empirical constraints on the theory there are theoretical constraints imposed top-down from the architectural status of ACT's claims. That is, an ACT model may need to be designed in a certain way so that the knowledge it specifies is sufficient to perform the given task when ACT mechanisms are applied. Finally, based on the rational analysis of cognition (Anderson 1990), constraints have been imposed on the theory so that it includes the kind of processing that is necessary and sufficient to meet the demands of the environment.

5. Future Issues

The ACT theory is still under active development. An extension has been added (called ACT-R/PM, Byrne and Anderson 1998) that incorporates perception and motor modules (e.g., eyes, ears, and hands). This extension enables the system to model interaction with the environment. In addition, ACT models have been developed for a variety of new domains, including complex, dynamic tasks such as air-traffic control. In some cases, the data sets being modeled even include eye-movement protocols. Yet another area of development involves exploring the relationship between the algorithmic level of description of the ACT-R computer simulation system (cf. Marr 1982) and a corresponding neural level implementation.

See also: Cognitive Psychology: History; Cognitive Psychology: Overview; Cognitive Theory: SOAR; Knowledge Spaces; Knowledge Representation; Logics for Knowledge Representation; Mathematical

Psychology; Problem Solving and Reasoning, Psychology of;

Bibliography

Anderson J R 1976 *Language, Memory, and Thought*. Erblaum, Hillsdale, NJ

Anderson J R 1978 Arguments concerning representations for mental imagery. *Psychological Review* **85**: 249–77

Anderson J R 1983 *The Architecture of Cognition*. Harvard University Press, Cambridge, MA

Anderson J R 1990 *The Adaptive Character of Thought*. Erlbaum, Hillsdale, NJ

Anderson J R 1993 *Rules of the Mind*. Erlbaum, Hillsdale, NJ

Anderson J R, Lebiere C 1998 *The Atomic Components of Thought*. Erlbaum, Mahwah, NJ

Anderson J R, Boyle C F, Corbett A T, Lewis M W 1990 Cognitive modeling and intelligent tutoring. *Artificial Intelligence* **42**: 7–49

Anderson J R, Conrad F G, Corbett A T 1989 Skill acquisition and the LISP tutor. *Cognitive Science* **13**: 467–505

Anderson J R, Matessa M 1997 A production system theory of serial memory. *Psychological Review* **104**: 728–48

Byrne M D, Anderson J R 1998 Perception and action. In: Anderson J R, Lebiere C (eds.) *The Atomic Components of Thought*. Erlbaum, Mahwah, NJ

Lovett M C, Daily L Z, Reder L M 2000 A source activation theory of working memory: Cross-task prediction of performance in ACT-R. *Cognitive Systems Research* **1**: 99–118

Marr D 1982 *Vision*. Freeman, San Francisco

Newell A 1973 You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In: Chase W C (ed.) *Visual Information Processing*. Academic Press, New York

Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

M. C. Lovett

Cognitive Theory: SOAR

SOAR is a computational theory of human cognition that takes the form of a general cognitive architecture (Laird et al. 1987, Newell 1990, Rosenbloom et al. 1992). SOAR (not an acronym) is a major exemplar of the architectural approach to cognition, which attempts the unification of a range of cognitive phenomena with a single set of mechanisms, and addresses a number of significant methodological and theoretical issues common to all computational cognitive theories (Anderson and Lebiere 1998, Newell 1990, Pylyshyn 1984). SOAR is also characterized by a set of specific theoretical commitments shaped primarily by attempting to satisfy the functional requirements for supporting human-level intelligence, manifest in soar's parallel existence as a state-of-the art artificial intelligence system (Laird et al. 1987). This focus on functionality, and its attendant theoretical commitments, is what makes soar both distinctive and controversial in cognitive psychology. SOAR represents the last major work of Allen Newell, one of the founders of modern cognitive science and artificial intelligence, and a pioneer in the development of architectures as a class of cognitive theory.

1. *Multiple Constraints on Mind and Computational Theories of Cognition*

Newell (1980a, 1990) described the human mind as a solution to a set of functional constraints (e.g., exhibit adaptive (goal-oriented) behavior, use language, operate with a body of many degrees of freedom) and a set of constraints on construction (a neural system, grown by embryological processes, arising through evolution). The structure of SOAR is shaped primarily by three of the functional constraints: (a) exhibiting flexible, goal-driven behavior, (b) learning continuously from experience, and (c) exhibiting real-time cognition (elementary cognitive behavior must be evident within about a second).

The emergence of computational models of cognition in information processing psychology (and artificial intelligence) represented a significant theoretical advance by providing the first proposals for physical systems that could, in principle, satisfy the functional constraints of exhibiting intelligence (Newell et al. 1958, Newell and Simon 1972). However, they raised a set of difficult methodological and theoretical issues that cognitive science still grapples with today. Among these issues are: (a) the problem of irrelevant specification (in a complex computer program, which of the myriad aspects of the program carry theoretical content, and which are irrelevant implementation details) (Reitman 1965); (b) the problem of too many degrees of freedom (an unconstrained computer program can be modified to fit any data pattern); and (c) the problem of identifiability (any sufficiently general proposal for processing schemes or representations can mimic the input/output characteristics of any other general processing or representation scheme (Anderson 1978, Pylyshyn 1973).

2. *SOAR as a Confluence of Five Major Technical Ideas in Cognitive Science*

SOAR can be seen as a confluence of five major technical ideas in cognitive science, which, taken together, are intended to address the three functional constraints summarized above, as well as the fundamental methodological issues concerning computational models.

2.1 *Physical Symbol Systems*

SOAR is a physical symbol system. The physical symbol system hypothesis asserts that physical symbol systems

are the only class of systems that can in principle satisfy the constraint of supporting intelligent, flexible behavior. Physical symbol systems are a reformulation of Turing universal computation (Church 1936, Turing 1936) that identifies symbol processing as a key feature of intelligent computation. The requirement is that the system be capable of manipulating and composing symbols and symbol structures—physical patterns with associated processes that give the patterns the power to denote either external entities or other internal symbol structures (Newell 1980a, 1990, Simon 1996). The key to the universality of Turing machines and physical symbol systems is their programmability: content can be added to the systems (in the form of programs) to change their behavior, yielding indefinitely many response functions.

2.2 Cognitive Architectures

SOAR is a cognitive architecture. A cognitive architecture is a theory about the fixed computational structure of cognition (Anderson and Lebiere 1998, Newell 1990, Pylyshyn 1984). Computational systems that are programmable must have some kind of fixed structure that processes the variable content: a set of primitive processes, memories, and control structures. The theoretical status of this underlying structure has not always been clear in cognitive models. For example, when a cognitive model is programmed in Lisp, the theorist intends to make some theoretical claims about the program (e.g., that the steps of the program corresponds in some way to the cognitive steps of the human performing the task), but probably intends to make no theoretical claims about Lisp as the architecture that executes the program (e.g., the fact that unused memory structure are reclaimed via a garbage collection process is theoretically irrelevant).

A cognitive architecture explicitly specifies a fixed set of processes, memories, and control structures that are capable of encoding content and executing programs. Cognitive models for specific tasks can be developed in such architectures by programming them. The theoretical status of various parts of a programmed implementation is now considerably clarified: what counts is the structure of the architecture (not its particular implementation), and the cognitive model's program, which makes a set of specific commitments about the form and content of knowledge used in a specific task. Thus, implemented cognitive architectures go a long way toward solving the irrelevant specification problem.

Cognitive architectures, especially those with temporal mappings and integrated learning mechanisms, can also address the degrees of freedom problem and identifiability problem in four ways. First, to the extent that architectures have a constrained temporal mapping, the space of possible programs that yield both the required functionality and temporal profile is considerably reduced. Second, to the extent that architectures have learning components that can acquire new knowledge (e.g., about a specific task), the form of that knowledge is no longer freely under control of the theorist. Third, to the extent that architectures are programmable (and are also constrained by a temporal mapping or learning mechanism), they permit a single set of processing assumptions to be applied to a diverse range of tasks, constraining that theory by a broader range of data. Fourth, to the extent that cognitive architectures are comprehensive and include some perceptual and motor components, they can be used to provide closed-loop models of complete tasks, so that no explanatory power need be ascribed to anything external to the model.

2.3 Production Systems

All long-term memory in SOAR is held in the form of productions (Anderson 1993, Newell 1973). Each production is a condition-action pair. The conditions form access paths and the actions form the memory contents. Productions continuously match against a declarative working memory that contains the momentary task context, and matching productions put their contents (actions) back into the working memory. Productions are the lowest level of elementary memory access available in SOAR, and Newell's (1990) temporal mapping onto human cognition places them approximately at the 10 ms level. This mapping provides strong constraints on the shape of cognitive models built in SOAR that must operate in real time.

SOAR's productions form a recognition memory. Such recognition memories have a number of features that make them attractive as models of human memory: they are associational in nature (access is via the contents of working memory); they are fine-grained and independent (which makes them a good match for continuous, incremental learning mechanisms); they are dynamic (a production system by itself defines a computationally complete system that can yield behavior; other processes are not needed to access or execute the memory structures); and they are cognitively impenetrable (their contents and structure may not be arbitrarily searched over, examined, or modified, but only accessed via automatic association). All of these properties place them in sharp contrast to memories in digital computers, which are static structures (not processes), freely addressable by location.

3. Search in Problem Spaces Supported by a Two-level Automatic/Deliberate Control Structure

SOAR achieves all cognition by search in problem spaces, and architecturally supports this by a flexible,

two-level recognize–decide–act control structure. Problem spaces are based in part on the idea that search in combinatoric spaces is the fundamental process for attainment of difficult tasks. The nature of such search is seen most easily in tasks like chess that have a well-defined set of operators and states. A search space consists of a set of (generated) representational states and operators that transition between states.

Problem spaces as realized in SOAR extend the standard notion of search in an important direction: problems spaces are taken to be the fundamental way that humans accomplish all cognitive tasks, including routine (i.e., well-practiced) tasks. SOAR is, therefore, one realization of the problem-space hypothesis (Newell 1980b), which asserts that all deliberate cognitive activity occurs in problem spaces. The key to this move lies in the role of knowledge in problem spaces: problem spaces freely admit of any amount of knowledge for guiding search, executing operators, or formulating the space initially in response to a task. Because SOAR provides a set of mechanisms (described next) that support this kind of knowledge use, behavior in SOAR spans the well-known continuum between knowledge-intensive processing (little search) and knowledge-lean processing (much search) (Newell 1990).

Supporting knowledge-driven search places strong functional demands on the architecture's control structure: at any step in the problem-solving process—selecting the next operator, generating the next state, etc.—any relevant knowledge must be brought to bear. There are two parts to the solution to this problem: the mechanisms for appropriate indexing of the knowledge, and the mechanisms for retrieving and applying the relevant knowledge during search. The indexing concerns learning, discussed below.

For retrieving and applying the knowledge during search, SOAR relies on a two-level control structure that separates the automatic access of knowledge via the productions from the deliberate level of problem solving. Each cognitive step is accomplished by a recognize–decide–act cycle. In the recognize phase, all productions that match the current state fire, producing new content in the working memory. Part of this retrieved content is about what the system should do next—the possible operators to try in the current state, the relative desirability of these operators (e.g., operator A is better than operator B), and so on. Next, in the decide phase, a fixed (domain independent) decision procedure sorts out these preferences in working memory to determine if they converge on a consistent decision. In the event that this processing clearly determines the next step, the decision procedure places in working memory an assertion about what that step should be. In the act phase, that step is taken (by additional production rule firings): the move to the next state in internal problem space search, or the release of motor intentions in external interaction. If it is not clear what to do next (e.g., several operators have been proposed, but no knowledge is evoked to prefer one option to another, or there are conflicts in the retrieved knowledge), an *impasse* has arisen, and the decision procedure records in working memory the type of the *impasse*, and sets a subgoal of resolving that *impasse*. In this way, SOAR's problem solving gives rise automatically to a cascade of subgoals whenever the knowledge delivered by the recognition memory is insufficient for the current task.

The critical feature of this control structure is its run-time, least-commitment nature: each local decision in the problem space is made at execution time by assembling whatever relevant bits of knowledge can be retrieved (by automatic match) at that moment. Decisions are not fixed in advance, and there are no architectural barriers to the kinds of knowledge that can be brought to bear on the decisions.

3.1 Continuous, Impasse-driven Learning

SOAR continuously acquires new knowledge in its long-term memory through an experience-based learning mechanism called chunking (Laird et al. 1987, Rosenbloom and Newell 1983). This mechanism generates new productions in the long-term memory by preserving the results of problem solving that occurred in response to impasses. The conditions of the new production consist of aspects of the working memory state just before the impasse, and the actions of the production consist of the new knowledge that resolved the *impasse* (e.g., an assertion that one of the proposed operators is to be preferred to the other in the current situation). Upon encountering a similar situation in the future, the production will automatically match and retrieve the knowledge that permits SOAR to avoid the *impasse*. Thus, chunking is a mechanism that converts problem solving into recognition memory, continuously moving SOAR from knowledge-lean to knowledge-rich processing.

Chunking in SOAR has two important functional properties. First, it begins to provide a solution to the knowledge-indexing problem raised earlier. The system assembles its own indices out of the contents of working memory in a way that is directly aimed at making the knowledge retrievable when it is relevant to the immediate demands of the task at hand. Second, learning permeates all aspects of cognition in SOAR. Chunking applies to all kinds of impasses, so any problem space function is open to learning improvements: problem-space formulation, operator generation, operator selection, and so on.

4. Major Architectural Implications and Specific Domains of Application

SOAR can be used as a theory in multiple ways (Newell 1990). Qualitative predictions can be drawn from SOAR

as a verbal theory, without actually running detailed computer simulations. These qualitative predictions can be both domain-general (cutting across all varieties of cognitive behavior) and domain-specific. The theory can be also be applied to specific domains by developing detailed computational models of a task; this involves programming SOAR by adding domain-specific production rules to its long-term memory, and generating behavioral traces.

4.1 Domain-independent Predictions

A principal prediction of a theory of human cognition is that humans are intelligent; the only way to clearly make that prediction is to demonstrate it operationally. SOAR makes this prediction only to the extent that the system has been demonstrated to exhibit intelligent behavior. As a state of the art AI system that has been applied to difficult tasks (ranging from algorithm design to scheduling problems), SOAR makes the prediction to a greater degree than other psychological theories.

SOAR makes a number of general predictions related to long-term memory and skill (Newell 1990). These include the prediction that procedural skill transfer is essentially by identical elements, and will usually be highly specific (Singley and Anderson 1989, Thorndike 1903); the bias of *Einstellung* will occur—the preservation of learned skill when it is no longer useful (Luchins 1942); the encoding specificity principle (Tulving 1983) holds; and recall will generally take place by a generate-and-recognize process (Kintsch 1970). The best known of SOAR's general predictions is the power law of practice, which relates the time to do a task to the number of times the task has been performed (Newell and Rosenbloom 1981, Snoddy 1926).

4.2 Domain-specific Predictions

SOAR models have been constructed across a range of task domains, and the behavior of the models has been compared with human data on those tasks. One area that has received considerable attention is human-computer interaction (HCI). Some of the successes in this area, such as a detailed model of transcription typing (John 1988), are a result of SOAR inheriting the results of the GOMS theory (Goal, Operators, Methods, and Selection rules), a theory developed in HCI to predict the time it takes expert users to do routine tasks (Card et al. 1983). (GOMS can be seen at one level as a specialization of SOAR, missing features such as learning and impassing.) Other SOAR HCI models depend crucially on SOAR's real-time interruptability (a function of the two-level control structure) and SOAR's learning mechanism. SOAR models have been developed of real-time interaction and learning in video games (John et al. 1994), novice-to-expert transitions in computer menu navigation (Howes and Young 1997), and a programmer's interaction with a text editor (Altmann and John 1999), among others.

SOAR models have also been developed of problem solving (Newell 1990), sentence processing (Lewis 2000), concept acquisition (Miller and Laird 1996), and interaction with educational microworlds (Miller et al. 1999). In all SOAR models (as with any cognitive model), the explanatory power is shared to varying degrees by both the content posited by the theorist for the particular task and the architectural mechanisms. For example, in the sentence processing model, SOAR's control structure and learning mechanism, coupled with the real-time constraint, lead directly to a theory of ambiguity resolution that yields a novel explanation of apparent modularity effects and their malleability (Lewis 1996a, Newell 1990), but the architecture provides little apparent constraint on the choice of grammatical theory, which also plays a role in the empirical predictions (Lewis 1996b). Similarly, the general theory of episodic indexing of attention events embodied in the text editor model depends critically on SOAR's continuous chunking mechanism (Altmann and John 1999), while the specific behavioral traces are a function, in part, of task strategies that could be accommodated by alternative architectures.

5. Critiques of SOAR, and Future Directions

Critiques of SOAR fall into three major classes: critiques of specific models built within SOAR, critiques of the architecture itself, and critiques of the general methodological approach of building comprehensive architectural theories. For example, specific empirical critiques have been made of SOAR models of the Sternberg memory search task (Lewandowsky 1992) and immediate reaction tasks (Cooper and Shallice 1995). The theoretical challenge is understanding the extent to which the empirical problems can be resolved within the existing architecture, or whether they point back to problems in the architecture itself (Newell 1992b). (The fact that the latter is a real possibility demonstrates that the architectural approach has made some headway on the identifiability and degrees of freedom problems.)

At the architectural level, nearly every major assumption of SOAR has been challenged in the literature (see the multiple book review in BBS for a range of assessments; Newell 1992a). Many of these architectural-level criticisms have been aimed at the *uniformity* assumptions in SOAR (all tasks as problem spaces, all long-term memory as productions, all

learning as chunking), which appear at first to run strikingly against the prevailing mode of theorizing in both cognitive psychology and cognitive neuroscience, which emphasizes functional specialization and distinctions over computational generality. The evaluation of SOAR in light of these concerns is not always transparent, however. For example, the analysis of SOAR's implications for modularity (particularly in language processing) revealed that SOAR is not only consonant with, but even predicts, many of Fodor's diagnostics of modular systems (Lewis 1996a, Newell 1990).

Finally, the general approach to cognitive theory that SOAR embraces has come under sharp criticism (most notably by Cooper and Shallice 1995) for not living up to the promise of addressing the methodological concerns identified above, and for not yielding theories with deep empirical coverage that clearly gain their explanatory power from general architectural mechanisms. To the extent that these critiques depend on practice with the SOAR theory specifically, their implications for the broader approach are insecure. Other architectural theories (e.g., ACT (Anderson and Lebiere 1998) and EPIC (Meyer et al. 1995)) exist in the field, and each has adopted somewhat different ways of dealing with these methodological issues that may or may not make them suspect to the same criticisms.

The evolution of SOAR as a theory, and its broader role in cognitive science, is likely to proceed along two fronts. First, SOAR will remain an important source of ideas for developing theories of complex cognition, even for those theorists who do not embrace the architecture whole cloth, or reject the architectural methodology. A harbinger of this can be seen in cognitive neuroscience: as researchers begin to tackle the problem of understanding the nature of 'executive' processes and their realization in the brain, models like SOAR can provide concrete proposals for a set of functionally sufficient mechanisms for the control of deliberate cognition; (see the recent volume on working memory and executive control for evidence of such interaction by Miyake and Shah 1999). Second, SOAR will continue to evolve as a unified set of mechanisms itself, informed in part by the continued application of SOAR to difficult AI problems, and in part by the continued construction and empirical evaluation of detailed models of cognitive tasks that focus on unique aspects of the architecture.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence in Cognitive Science; Artificial Intelligence: Search; Cognitive Theory: ACT; Deductive Reasoning Systems; Expert Systems in Medicine; Intelligence: History of the Concept; Production Systems in Cognitive Psychology; Scientific Discovery, Computational Models of

Bibliography

Altmann E M, John B E 1999 Episodic indexing: A model of memory for attention events. *Cognitive Science* **23**(2): 117–56

Anderson J R 1978 Arguments concerning representations for mental imagery. *Psychological Review* **85**(4): 249–77

Anderson J R 1993 *Rules of the Mind*. Erlbaum, Hillsdale, NJ

Anderson J, Lebiere C 1998 *Atomic Components of Thought*. Erlbaum, Hillsdale, NJ

Card S K, Moran T P, Newell A 1983 *The Psychology of Human–Computer Interaction*. Erlbaum, Hillsdale, NJ

Church A 1936 An unsolvable problem of elementary number theory. *The American Journal of Mathematics* **58**: 345–63

Cooper R, Shallice T 1995 SOAR and the case for unified theories of cognition. *Cognition* **55**(2): 115–49

Howes A, Young R M 1997 The role of cognitive architecture in modelling the user: SOAR's learning mechanism. *Human–Computer Interaction* **12**: 311–43

John B E 1988 *Contributions To Engineering Models of Human–Computer Interaction*. Carnegie Mellon University, Pittsburgh, PA

John B E, Vera A H, Newell A 1994 Toward real-time GOMS: A model of expert behavior in a highly interactive task. *Behavior and Information Technology* **13**: 255–67

Kintsch W 1970 Models for free recall and recognition. In: Norman D A (ed.) *Models of Human Memory*. Academic Press, New York

Laird J E, Newell A, Rosenbloom P S 1987 SOAR: An architecture for general intelligence. *Artificial Intelligence* **33**: 1–64

Lewandowsky S 1992 Unified cognitive theory: Having one's apple pie and eating it. *Behavioral and Brain Sciences* **15**(3): 449–50

Lewis R L 1996a Architecture matters: What SOAR has to say about modularity. In: Steier D M, Mitchell T M (eds.) *Mind Matters: Contributions to Cognitive and Computer Science in Honor of Allen Newell*. Lawrence Erlbaum Associates, Mahwah, NJ

Lewis R L 1996b Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research* **25**(1): 93–115

Lewis R L 2000 Specifying architectures for language processing: Process, control, and memory in parsing and interpretation. In: Crocker M W, Pickering M, Clifton C Jr (eds.) *Architectures and Mechanisms for Language Processing*. Cambridge University Press, Cambridge, UK

Luchins A S 1942 Mechanization in problem solving. *Psychological Monographs* **54**(6): no. 28

Meyer D E, Kieras D E, Lauber E, Schumacher E H, Glass J, Zurbriggen E, Gmeindl L, Apfelblat D 1995 *Adaptive Executive Control: Flexible Human Multiple-task Performance Without Pervasive Immutable Response-selection Bottlenecks*. University of Michigan, Ann Arbor, MI

Miller C S, Laird J E 1996 Accounting for graded performance within a discrete search framework. *Cognitive Science* **20**: 499–537

Miller C S, Lehman J F, Koedinger K R 1999 Goals and learning in microworlds. *Cognitive Science* **23**(3): 305–36

Miyake A, Shah P (eds.) 1999 *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge University Press, Cambridge, UK

Newell A 1973 Production systems: Models of control structures. In: Chase W G (ed.) *Visual Information Processing*. Academic Press, New York

Newell A 1980a Physical symbol systems. *Cognitive Science* **4**: 135–83

Newell A 1980b Reasoning, problem solving and decision processes: The problem space as a fundamental category. In: Nickerson R (ed.) *Attention and Performance VIII*. Erlbaum, Hillsdale, NJ
Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA
Newell A 1992a Precis of unified theories of cognition. *Behavioral and Brain Sciences* **15**(3): 425–92
Newell A 1992b SOAR as a unified theory of cognition: Issues and explanations. *Behavioral and Brain Sciences* **15**(3): 464–92
Newell A, Rosenbloom P 1981 Mechanisms of skill acquisition and the law of practice. In: Anderson J R (ed.) *Cognitive Skills and Their Acquisition*. Erlbaum, Hillsdale, NJ
Newell A, Shaw J C, Simon H A 1958 Elements of a theory of human problem solving. *Psychological Review* **65**: 151–66
Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ
Pylyshyn Z W 1973 What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin* **80**(1): 1–24
Pylyshyn Z W 1984 *Computation and Cognition*. Bradford/MIT Press, Cambridge, MA
Reitman W 1965 *Cognition and Thought*. Wiley, New York
Rosenbloom P S, Laird J E, Newell A (eds.) 1992 *The SOAR Papers: Research on Integrated Intelligence*. MIT Press, Cambridge, MA
Rosenbloom P S, Newell A 1983 The chunking of goal hierarchies: A generalized model of practice. In: Michalski R S, Carbonell J, Mitchell T (eds.) *Machine Learning: An Artificial Intelligence Approach II*. Morgan Kaufman, Los Altos, CA
Simon H A 1996 The patterned matter that is mind. In: D a M Steier T (ed.) *Mind Matters: Contributions to Cognitive and Computer Science in Honor of Allen Newell*. Erlbaum, Hillsdale, NJ
Singley M K, Anderson J R 1989 *The Transfer of Cognitive Skill*. Harvard University Press, Cambridge, MA
Snoddy G S 1926 Learning and stability. *Journal of Applied Psychology* **20**: 1–36
Thorndike E L 1903 *Educational Psychology*. Lemke and Buechner, New York
Tulving E 1983 *Elements of Episodic Memory*. Oxford University Press, New York
Turing A M 1936 On computable numbers, with an application to the Entscheidungsproblem. Paper presented at the Proceedings of the London Mathematics Society

R. L. Lewis

Cognitive Therapy

1. Definition

Cognitive therapy is one of a large number of psychotherapy approaches that was developed in the latter part of the twentieth century. It is associated principally with its originator, Dr. Aaron Beck. The treatment model falls within a broader class of treatments, which are referred to as the cognitive-behavioral therapies (Dobson, 2001). The cognitive-behavioral therapies share certain theoretical underpinnings, notably: (a) that cognitive events affect behavior (the mediational hypothesis); (b) that cognitive events can be assessed and systematically modified (the accessibility hypothesis); and (c) that cognitive change can be employed to cause therapeutic changes in behavior or adaptive functioning. In this regard, cognitive therapy is similar to other schools of cognitive-behavioral psychotherapy, such as rational-emotive psychotherapy.

Cognitive therapy distinguishes itself from other cognitive-behavioral therapies by the particular organization of its theoretical constructs. The model proposes that the manner in which an individual views, appraises, or perceives events around himself/herself is what dictates their subsequent emotional responses and behavioral choices. Although the content of situation-specific appraisals will vary as a function of the person's activities, the model also posits that these appraisals may be accurate or distorted, positive or negative. For example, an individual reacts to what another individual said in an interpersonal interaction, as well as to what he or she thinks about those statements.

The extent to which the individual sees situations accurately, or conversely may be distorting, is in part a function of the individual's core beliefs (also referred to as underlying assumptions, or schemas). These core beliefs are hypothesized to be intrapsychic phenomena that emerge over the person's lifetime, based on their experiences. Once established, these beliefs not only increase the likelihood of certain cognitive reactions to life events, but also then influence proactively the way in which an individual chooses to spend their time, career, partner choices, etc., which also tend to reinforce these beliefs. Thus, over time cognitive beliefs become the basic factor in what may become an increasingly negative and closed feedback loop (see Fig. 1).

Cognitive therapy is a systematic treatment, which is founded on the cognitive model of distress. The principal emphases in therapy are assisting the patient to identify, evaluate, and modify the potentially faulty information processing they engage in, as well as the underlying beliefs or schemas that drive that information processing. This process typically begins with an assessment of adaptive functioning and behavioral patterns that might either enhance or interfere with the process of treatment. In the case of depression, the patient may be encouraged to increase their activities to alleviate negative mood, and to ensure that the patient is at least engaged in their life sufficiently to generate negative information processing that can then be the focus of treatment. In the case of marital distress, negative interactional patterns (e.g., fighting, spousal abuse) will be assessed and modified before interventions aimed at relationship beliefs will be attempted.

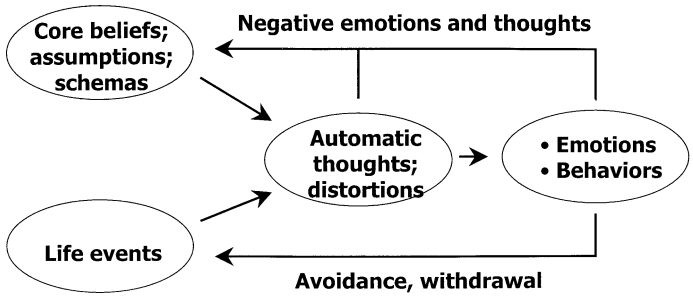

Figure 1
The cognitive therapy model of emotional distress

Once the patient is able to engage in a review of their negative thinking, a number of methods can be employed to assess this thinking. Strategies include counts of particular thoughts, questioning in the therapy session, thoughts about imagined situations, or the use of a written thought record (J. Beck 1995). Dependent on the pattern of dysfunctional thinking that is identified, various therapeutic techniques can be employed to change these patterns. For example, if a patient recurrently believes they 'cannot' be assertive with certain people, the therapist and patient can collaborate to set up circumstances in which they can test this idea empirically. In doing so, adaptive thinking and various behavioral competencies can be enhanced.

At a certain stage of treatment, it is typically the case that a patient's situation-specific automatic thoughts are accurate and adaptive, and that they are functioning more adaptively than when they first arrived for treatment. At this stage of treatment, the focus will move to the more general beliefs that prompted the patient's problems in the first instance. This work includes a search for the common themes underlying specific thoughts. These themes can take the form of beliefs about how 'the world' operates, or schemas that have developed about the self. Techniques that can be used include inductive questioning, a review of the negative thinking patterns seen by the therapist and patient, examination of historical patterns of thought and behavior, and a review of life events related to the problem. Once identified, dysfunctional beliefs can be challenged systematically through behavioral experiments, bibliotherapy, open discussion with key people in the patient's life, and other 'assumptive techniques' (J. Beck 1995).

In summary, cognitive therapy is a systematic and progressive form of treatment that typically includes assessment and modification of adaptive functioning, situation-specific automatic thinking, and more engrained, long-term beliefs and self-schemas. The cognitive content of various forms of disorders is increasingly understood, and it includes themes of danger in anxiety (Beck and Emery 1985), loss in depression (Beck et al. 1979, Clark et al. 1999), and transgression in anger (Beck 1999). Further, the cognitive processes associated with various disorders (e.g., magnification of perceived danger in anxiety) are also increasingly understood.

2. *Intellectual Context*

Beck's original interest was in understanding and treating depression. He began his departure from his original psychoanalytic training through a series of studies of the dreams and daytime thoughts of depressed patients, in which he discovered that the content of these thoughts had stereotypical content. Further, he hypothesized that these depressed patients engaged in negative distortions of their world in order to obtain these thoughts. Based on these observations, he developed a treatment in which the negative cognitions were systematically tested, in order to undermine cognitive distortions and the negative thinking seen in depression (Beck et al. 1979).

The early formulation of cognitive therapy was put to its first empirical test in an outcome study in the late 1970s. Based on a developed treatment manual, cognitive therapy was contrasted with antidepressant medication. This study revealed that cognitive therapy had equal outcome to medication in the short term and superior outcome at follow-up. As a consequence of the original success of cognitive therapy, a number of outcome studies were conducted in the area of depression in subsequent years.

The first meta-analysis of these studies (Dobson 1989) used a standard outcome measure, the Beck Depression Inventory, and compared cognitive therapy to other forms of treatment. This analysis revealed that cognitive therapy patients ended treatment 0.5 standard deviations less depressed than patients receiving other treatments. Further, compared to wait list or placebo conditions, the effect size for cognitive therapy indicated that it was over 2 standard deviations superior to no treatment. Although the results

of the above meta-analysis were criticized due to questions about the selection process for including studies, the conclusion that cognitive therapy is at least equal to pharmacotherapy has not been challenged seriously. Further, a more recent meta-analysis (Gloaguen et al. 1998) has confirmed the results found earlier by Dobson (1989).

More specifically, Gloaguen et al. (1998) found similar effect sizes to the earlier analysis, although the overall magnitude of these effects was somewhat attenuated. Notably, in an examination of long-term effects, it was also reported that relapse rates from cognitive therapy were approximately one-half of those observed for drug therapy for depression. Most recently, DeRubeis et al. (1999) conducted a 'mega-analysis' of treatment for depression, in which they collapsed the data from four outcome studies and analyzed the effects of cognitive therapy for depression, relative to drug therapy. They concluded that these two forms of therapy were equally effective, even for severely depressed patients. As such, it is an established fact that cognitive therapy is an effective treatment for clinical depression.

Notwithstanding the demonstrated efficacy of cognitive therapy for depression, a number of questions continue to require theoretical and empirical attention. The model underlying cognitive therapy states that cognitive distortions and situation-specific negative thoughts emerge from an interaction between core beliefs (also referred to as cognitive schemas, or underlying assumptions) and the life events that impinge on these beliefs. Once the core beliefs are activated and negative cognitions are produced, the emotional and behavioral consequences of these thoughts naturally flow (see Fig. 1). This mediational model, while intuitive and having lead to a successful treatment technology, has yet to be validated through research. In fact, based on a comprehensive examination of the research that has tested the various assumptions of cognitive therapy (Clark et al. 1999), the conclusion is that cognitive mediation remains to be proven.

Another challenge for the cognitive therapy of depression model comes from a component analysis of this treatment. Jacobson et al. (1996) conducted a randomized clinical trial in which depressed patients received either only the behavioral interventions associated with cognitive therapy, behavioral interventions and those aimed at situation-specific distortions, or the complete cognitive therapy program. Contrary to predictions, all three treatment conditions had equal outcomes, both in the short term and at up to two years follow-up. This study suggests that the cognitive interventions that are the hallmark features of the treatment may not, in fact, contribute to outcome more than the behavioral components of the treatment. If replicated, these results suggest that the active ingredients of cognitive therapy should be reconceptualized.

Further issues exist with regard to cognitive therapy of depression. These include the role of patient characteristics that potentially effect treatment outcome, the role of life stress in depression, the recent focus on schema-focused cognitive therapy (Clark et al. 1999), the risk of relapse relative to other treatments, and assessment issues relative to the adherence and competence of therapists in providing cognitive therapy.

3. *Changes in Focus or Emphasis Over Time*

Branching out from the early work on depression (Beck et al. 1979), Beck and his colleagues extended the cognitive model to other clinical conditions. Beck and Emery published a treatment manual for anxiety disorders in 1985, which was followed by works dedicated to marital disorder (Beck 1988), personality disorders (Beck et al. 1990), substance use and abuse (Beck et al. 1993), and—most recently—anger and aggression (Beck 1999). Other works on cognitive therapy have appeared for bipolar disorder, as well as numerous chapters in various sources (see J. Beck 1995 for a review).

Throughout the development of cognitive therapy, some features have remained constant. First, while the *content* of cognition related to various forms of disorders necessarily varies, there has been a consistent emphasis on the *process* of thinking in cognitive therapy that permits many of the treatment techniques to be applied across disorders. For example, a common intervention in cognitive therapy is assessing the reality basis, or veridicality, of certain perceptions. This intervention can be used successfully in many forms of anxiety disorders, depression, marital distress, and anger-related problems.

A second consistent emphasis in cognitive therapy has been a focus on treatment efficacy. From the outset of the development of cognitive therapy, Beck and colleagues have maintained that efficacy studies are required to establish the clinical utility of cognitive therapy. No doubt it is in part due to this emphasis that many of the psychological treatments now being recognized as empirically supported are variants of cognitive therapy.

4. *Methodological Issues or Problems*

Two primary sets of methodological issues have constrained the development of cognitive therapy. The first of these is related to the methodological issues inherent in clinical research. The randomized clinical trial methodology, which has become the standard in the development of psychotherapy, makes a number of requirements on investigators. These include the need for a well-defined independent variable, which in the context of psychotherapy means a treatment

manual. Unfortunately, a consequence of this requirement is that fully developed treatment manuals are often developed before the research that is needed to substantiate them has been conducted. A second requirement of clinical trials is that the subject group be clearly specified. In psychotherapy research, this requirement has often been translated into using diagnostically related groups, ideally with few or no complicating co-morbid conditions. The relative homogeneity of research participants, while providing a good test of the intervention, unfortunately leads to what may be relatively poor generalizability of research findings. The requisite training for research trials has also been controversial. A fourth area of controversy surrounds the measurement and evaluation of outcome. Typically, the outcomes of a given treatment can be assessed in a number of ways (diagnostically, at the symptom level, or using specialized assessment tools), and the analysis of the outcomes can also be handled in several ways. As a consequence of these various strategies to conduct clinical trials, the scientific status of various treatments can be challenged. Given that most psychological treatments are developed in universities with publicly funded grants, and further given the necessarily limited set of questions that any one treatment study can address, it is not surprising that considerable efficacy and effectiveness research in the area of cognitive therapy is needed.

The second major methodological issue for cognitive therapy is the measurement of mechanisms of change. The cognitive model of distress makes a series of assumptions about the nature of cognitive processes involved in the development and treatment of various disorders (see Fig. 1). The measurement of cognition has been a difficult task, however—particularly as some of the constructs in the model are hypothetically latent until activated (Ingram et al. 1998). Even in the area of depression, where the greatest concentration of work has taken place to assess these constructs, some of the hypothesized mechanisms of change have yet to be established (Clark et al. 1999, Ingram et al. 1998). Research from the level of both theory and therapy are required to further evaluate the relationship between cognitive change and other parameters of clinical change in cognitive therapy.

5. *Probable Future Directions of Theory And Research*

Despite the success of cognitive therapy, and its phenomenal growth over the past two decades, there is no doubt that considerable development remains (Dobson and Khatri 2000). Some of the directions for this work that have been identified in the literature include:

(a) A need for continued emphasis on efficacy research, particularly as cognitive therapy is applied to new and innovative areas.

(b) A need for effectiveness research that assesses such issues as the utility of cognitive therapy relative to other treatments, the clinical acceptability of treatment to patients, and other issues that affect how practical the treatment is to apply in varied clinical contexts.

(c) Continued work on the measurement of cognition (Ingram et al. 1998) and the mechanisms of change in cognitive therapy is clearly warranted. Cognitive therapy is a complex, multi-component treatment, the mechanisms of which are only beginning to be studied. The results of this research will contribute both to the theory and application of the treatment.

(d) Theoretical therapy development and efficacy work are needed to address the issue of whether cognitive therapy is an exhaustive theory that can integrate other models (Dobson and Khatri, 2000), or whether it might be integrated optimally into other treatment models.

(e) There is a need for research that assesses the success and failure of cognitive therapy for patients of diverse backgrounds. This 'aptitude by treatment' research will help the theory of cognitive therapy to develop, as well as to ensure the most appropriate treatment is provided to different patient populations.

(f) Predicated on the assumption that cognitive therapy continues to enjoy strong clinical outcomes and popularity in the treatment community, further research is needed to understand the optimal method to disseminate this treatment approach. Tied to this development are issues related to how to best measure therapist adherence and competence in cognitive therapy.

See also: Behavior Psychotherapy: Rational and Emotive; Behavior Therapy: Psychiatric Aspects; Behavior Therapy: Psychological Perspectives; Cognitive and Interpersonal Therapy: Psychiatric Aspects; Depression, Clinical Psychology of

Bibliography

Beck A T 1988 *Love is Never Enough*. Harper and Row, New York

Beck A T 1999 *Prisoners of Hate: The Cognitive Bases of Anger, Hostility and Violence*. Harper Collins, New York

Beck A T, Emery G 1985 *Anxiety Disorders and Phobias: A Cognitive Perspective*. Basic Books, New York

Beck A T, Freeman A et al. 1990 *Cognitive Therapy of Personality Disorders*. Guilford Press, New York

Beck A T, Rush A J, Shaw B F, Emery G 1979 *Cognitive Therapy of Depression*. Guilford Press, New York

Beck A T, Wright F D, Newman C, Liese B S 1993 *Cognitive Therapy of Substance Abuse*. Guilford Press, New York

Beck J 1995 *Cognitive Therapy: Basics and Beyond*. Guilford Press, New York

Clark D A, Beck A, Alford B A 1999 *Cognitive Theory and Therapy of Depression*. Wiley, New York

DeRubeis R J, Gelfand L A, Tang T Z, Simons A D 1999 Medications vs. cognitive behavioral therapy for severely depressed outpatients: A mega-analysis of four randomized comparisons. *The American Journal of Psychiatry* **156**: 1007–13

Dobson K S 1989 A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology* **57**: 414–9

Dobson K S (ed.) 2001 *Handbook of Cognitive-behavioral Therapies*, 2nd edn. Guilford Press, New York

Dobson K S, Khatri N 2000 Cognitive therapy: Looking backward, looking forward. *Journal of Clinical Psychology* **56**: 907–23

Gloaguen V, Cottraux J, Cucherat M, Blackburn I 1998 A meta-analysis of the effects of cognitive therapy in depressed patients. *Journal of Affective Disorders* **49**: 59–72

Ingram R, Miranda J, Segal Z V 1998 *Cognitive Vulnerability to Depression*. Guilford Press, New York

Jacobson N S, Dobson K S, Truax P, Addis M, Koerner K, Gollan J, Gortner E, Prince S 1996 A component analysis of cognitive behavioral treatment for depression. *Journal of Consulting and Clinical Psychology* **64**: 295–304

K. S. Dobson

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Coh

Cohort Analysis

A cohort is a set of individuals entering a system at the same time. Individuals in a cohort are presumed to have similarities due to shared experiences that differentiate them from other cohorts. Cohort analysis seeks to explain an outcome through exploitation of differences between cohorts, as well as differences across two other temporal dimensions: 'age' (time since system entry) and 'period' (times when an outcome is measured). This article exposits difficulties inherent to cohort analysis, indicates promising directions, and provides context.

1. Cohorts

Cohorts as analytic entities appear in the social sciences, the life sciences, epidemiology, and elsewhere. A cohort can be a set of people, automobiles, trees, whales, buildings; the possibilities are endless. System entry can refer to birth—a person is born, or to any dated event—a machine is assembled on a particular date. A set of individuals who begin serving a prison term at the same point in time might also define a cohort for certain purposes, in which case 'birth' refers to initiation into a particular role system and 'age' becomes duration (time since system entry). The breadth of the time interval that defines membership in a particular cohort depends on analytic considerations and the nature of the phenomenon under study.

The main difficulties inherent to cohort analysis can be illustrated with data from the NORC General Social Survey, a national sample survey conducted annually or biennially in the United States. Respondents in this repeated cross-sectional survey are queried about their emotional wellbeing. Each cell in Table 1 shows the percentages, organized by age and survey year, of those who identified themselves as being 'very happy.' Each row shows how happiness levels change across survey year (period) for people within a given age group. Each column reveals age variation in any survey year. The diagonals permit tracking of a single birth cohort over time. Those aged 20–29 in 1973 were born between 1943 and 1953, as were those aged 30–39 in 1983, and those aged 40–49 in 1993. Ten-year age categories lead here to cohorts operationalized as individuals born in contiguous 10-year intervals.

Nominal attempts to extract information from Table 1 would summarize percent 'very happy' by rows, columns, and diagonals. The most immediately visible pattern is the association between age and happiness. In all three periods the percentages of respondents who identify themselves as 'very happy' are smaller for those under age 40. This reading of the table ignores the possibility that observed variation is due to birth cohort, with younger cohorts less likely to report being happy. Although visual inspection of the diagonals does not suggest a consistent association between decade of birth and eventual happiness, the absence of an observable relationship does not confirm that no such relationship exists. A potential association between birth cohort and adult happiness may have been obscured by the effects of age and period. Finally, there appears to be little systematic variation between years in percent 'very happy.' In this instance, however, a potential association between period and

Table 1
Percent 'very happy' by age and period

Age	Period		
	1973	*1983*	*1993*
20–29	29%	30%	28%
	(347)	(372)	(278)
30–39	36%	28%	29%
	(294)	(354)	(381)
40–49	40%	31%	30%
	(247)	(228)	(329)
50–59	41%	29%	38%
	(253)	(212)	(205)
60–69	38%	37%	33%
	(192)	(201)	(166)
70–79	38%	38%	33%
	(117)	(123)	(155)
$N = 4454$			

Source: General Social Survey, 1973–1993
Note: Numbers in parentheses are base *N*s for the percentages

current happiness may have been obscured by age and cohort. A further, linked difficulty is that the data structure is imbalanced. Although periods are represented over all ages, and ages are represented over all periods, the observed age spans of cohorts necessarily differ.

This imbalance cannot be extirpated from analyses that take into account age, period, and cohort simultaneously. Comparable points hold for other data structures supporting the analysis of multiple cohorts.

Table 1 shows that data structures that allow for measurements on multiple cohorts necessarily measure age and period. Furthermore, knowledge of placement on any two of age, period, and cohort determines placement on the third. This dependency can be expressed as:

$$\text{Cohort} = \text{Period} - \text{Age}$$

which raises the questions of whether and how all three of age, period, and cohort can be included in cohort models. The linear dependency between age, period, and cohort, also known as the cohort analysis identification problem (see *Statistical Identification and Estimability*), is the point of departure for all modern discussions of techniques of cohort analysis. The identification problem is present irrespective of data structure.

2. Data Structures

Three commonly seen data structures engender cohort analysis (for a fuller discussion of data structures, see Fienberg and Mason 1985). Excluded from this discussion, however, would be the *single cross section*. Reducing Table 1 to a single column indicates the data structure of a single cross section. With this design, differences on an outcome variable by age can be interpreted either as age or cohort differences. The data structure provides no basis for choosing because there is but a single period. Interpretation of age differences on an outcome as being attributable to factors thought to vary with age or duration requires the assumption that there are no cohort or period differences, or that they are known. Parallel assumptions are necessary for interpreting the age differences as attributable to cohort. Thus, the single cross section does not permit cohort analysis.

Some cross-sectional surveys elicit retrospective data on birth, marital, or other kinds of histories. Figure 1(a) structures this design, a *single cross section with retrospective data*, as an upper-triangular age by period array with cohorts defined by diagonals.

The retrospective data structure provides information on ages, cohorts and periods, and introduces longitudinal information on individuals (see *Longitudinal Data*). When the longitudinal data in this design are used without regard to age, period, and cohort, the analyst is implicitly, and possibly inadvertently, assuming that at least one of age, period, and cohort is not essential. The same point holds for *panel study* data structures (Fig. 1(b)). A panel study begins with a single cross section, and is followed by one or more panels on the same individuals (units of analysis). This design is intended for the creation of longitudinal data. Figure 1(b) structures the panel study as a lower-triangular age by period array with cohorts on the diagonal. It is triangular because, in the simplest case, panel designs do not replenish the data structure with the addition of new cohorts after the initial cross section. A panel study could also be designed to include new cohorts at successive waves of data collection; the age, period, cohort dependency would remain.

In the *replicated cross-section* design or time series of cross sections (see *Longitudinal Data*; *Event History Analysis in Continous Time*) illustrated by Table 1,

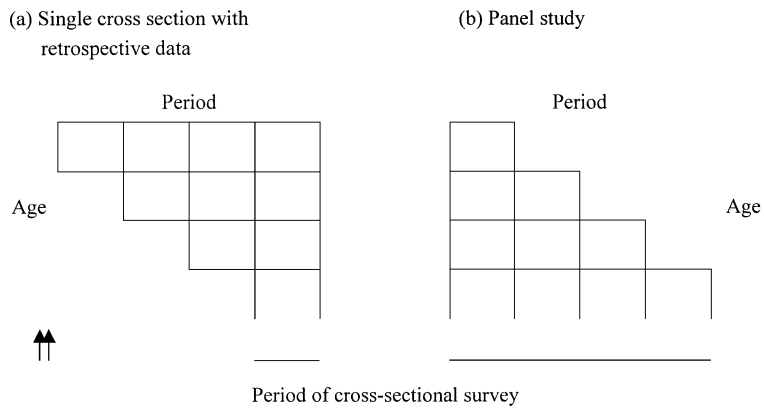

Figure 1
Retrospective and prospective data structures that engender age, period, and cohort

typically the same individuals are not tracked from one period to the next. However, each cross section can have a retrospective component, and thus this design can be longitudinal. Like the other multiple cohort designs, this structure permits cohort analysis, although the class of models it supports is less rich when retrospective information is unavailable.

3. Cohort Models

Cohort models may be fixed or random effect (see *Hierarchical Models: Random and Fixed Effects*); terms for age, period, and cohort may enter the model as discrete or continuous; one or more of the age, period, and cohort dimensions may be included in the model via an explicit, substantive measure of that dimension; interactions are possible. These are the most prominent possibilities in the literature on cohort analysis.

3.1 Fixed Effect: Discrete Age, Period, and Cohort

Assume an $I \times J$ age by period array (Table 1 is a 6×3 illustration), with age groups and period intervals of identical widths. The $K = I+J-1$ diagonals of the array correspond to cohorts. The basic fixed-effect model treats a parameter (θ_{ijk}) associated with a response variable as a linear function of discrete age, period, and cohort. Using dummy coding for age, period, and cohort, let:

$$\theta_{ijk} = \beta_0 + \sum_{i=2}^{I} \beta_i A_i + \sum_{j=2}^{J} \gamma_j P_j + \sum_{k=2}^{K} \delta_k C_k \qquad (1)$$

where the A_i, P_j, and C_k ($k = i-j+J$) are dummies for ages, periods, and cohorts, respectively. This is a fixed-effect model because inference is conditional on the ages, periods, and cohorts represented by a particular dataset. Although Eqn. (1) manifests usual normalization restrictions (omission of one dummy from each classification), this is insufficient to break the linear dependency between age, period, and cohort. Omission of all terms in one of the age, period, or cohort classifications eliminates the dependency. This is a satisfactory strategy if prior theory and information suggest that age, or period, or cohort is superfluous. On the other hand, if all three dimensions are deemed indispensable to the analysis, the dependency must be eliminated by one or more further restrictions on coefficients if the fixed-effect, discrete model is to be employed. Problems with this approach include collinearity among terms on the right hand side of Eqn. (1) remaining after additional coefficient restrictions have been introduced, and coefficient bias. Moreover, restrictions that might appear to be innocuously different, such as equating the coefficients of different subsets of adjacent categories, can lead to quite different sets of age, period, and cohort effects, with the various models all fitting the data equally well, or nearly so. One response to the problems of estimating Eqn. (1) is to massively overidentify one or more dimensions. For example, a priori knowledge may suggest that period can be represented by several, rather than many, categories. The data *cannot*, however, be relied on to contain the information on which to base overidentifying restrictions (Fienberg and Mason 1985, Glenn 1989, Heckman and Robb 1985, Kupper et al. 1983, Mason and Smith 1985, Wilmoth 1990).

3.2 Fixed Effect: Continuous Time

Let age (A), period (P), and cohort (C) be measured in continuous time, with $C = P-A$. Then the continuous time equivalent to Eqn. (1) is:

$$\theta_{APC} = f_A(A) + f_P(P) + f_C(C) \qquad (2)$$

where $f_A(A)$ is an $(I-1)$th-order polynomial in A, $f_P(P)$ is a $(J-1)$th-order polynomial in P, and $f_C(C)$ is a $(K-1)$th-order polynomial in C. Because of the $C = P-A$ linear dependency, the coefficients of the linear terms for A, P, and C are not estimable. As in the discrete case, at least one linear restriction must still be imposed and in any event the discrete variable approach is generally preferable.

3.3 Random Effect, Discrete Age, Period, Cohort

It is possible to view the modeling of age, period, and cohort effects from a Bayesian, hierarchical perspective (Nakamura 1986; see *Bayesian Statistics*; *Hierarchical Models: Random and Fixed Effects*). In this approach, it is convenient to characterize age, period, and cohort effects through first differences, as in

$$\theta_{ijk} = \beta_0^* + \sum_{i=2}^{I} \beta_i^* A_i^* + \sum_{j=2}^{J} \gamma_j^* P_j^* + \sum_{k=2}^{K} \delta_k^* C_k^* \qquad (3)$$

where $A_i^* = A_i - A_{i-1}$, $P_j^* = P_j - P_{j-1}$, and $C_k^* = C_k - C_{k-1}$. The approach assumes that the β_i^*, γ_j^*, and δ_k^* are separately distributed, and that they are random or exchangeable (Nakamura 1986, Sasaki and Suzuki 1987, 1989, Glenn 1989, Miller and Nakamura 1996). The exchangeability assumption requires that within the age dimension, within the period dimension, and within the cohort dimension, all permutations of the first-difference coefficients must be equally acceptable. This assumption makes possible the determination of age, period, and cohort effects without resorting to restrictions on coefficients. To the extent that the assumption fails, and it can when there are shocks

(e.g., war, famine, plague, a stock market crash) in the process being modeled, the random effects approach is not a panacea.

3.4 Substantive Measurement of Age, Period, or Cohort

When one or more of the age, period, and cohort dimensions, whether represented as discrete or continuous, is replaced by a variable chosen to measure the underlying process thought to be captured by age, or period, or cohort, the linear dependency is almost always broken. Attention is then appropriately focused on the theoretical and substantive merits of the specification. Models that include age, period, and cohort should be thought of as starting points, given their inferiority relative to models that are able to test ideas of how and why a cohort, or period, or age mechanism affects an outcome. Concretely, suppose the cohort dimension is held to reflect the impact of measured variable X, then Eqn. (1) might change to:

$$\theta_{ijk} = \beta_0 + \sum_{i=2}^{I} \beta_i A_i + \sum_{j=2}^{J} \gamma_j P_j + \delta X \tag{4}$$

where X is either constant over ages within a cohort, or varies by age within a cohort. Relative cohort size is an example of a variable that could be defined at the birth of each cohort, or allowed to vary as cohorts 'age' through the life cycle. Measured variables can also be employed in extensions and revisions of Eqns. (2)–(3). For example, in Eqn. (2) the polynomial in C can be omitted upon inclusion of some function of X (X itself; a polynomial in X; interactions between X and age or period).

In the extension of the random-effects approach that includes measured variables, cohort analysis becomes a specific case of random effects multilevel analysis (see *Statistical Analysis: Multilevel Methods*). This development can be expected in the course of the continued deployment of Bayesian statistical solutions, because the use of measured variables can enhance the validity of the exchangeability assumption. In the random-effects approach, substantive variables can be written into the model in the following way, for one or all of age, period, and cohort:

$$\begin{aligned} \beta_i^* &= \lambda_{0A} + \lambda_{1A} X_A + \tau_{iA} \\ \gamma_j^* &= \lambda_{0P} + \lambda_{1P} X_P + \tau_{jP} \\ \delta_k^* &= \lambda_{0C} + \lambda_{1C} X_C + \tau_{kC} \end{aligned} \tag{5}$$

where, for example, X_C is a measured variable for cohort, and for simplicity only one measured variable per dimension has been included, and only as a linear term. In this extension it is the τ_{iA}, τ_{jP}, and τ_{kC} that are assumed to be exchangeable within the age, period, and cohort dimensions, and this is likely to be more defensible—because substantively plausible underlying, measured variables should covary with the phenomena that produce shocks. Moreover, the exchangeability assumption becomes comparable to the assumption of a random-error term in a fixed effects approach with measured variables.

4. Interactions

The need for interactions often arises, either for substantive reasons (Converse 1976), or for technical, adjustment purposes (Mason and Smith 1985). It is possible to include interactions in both fixed- and random-effect models, regardless of the presence of measured variables. Fienberg and Mason (1985) elaborate readily implemented strategies for doing so in the fixed-effect framework, using either discrete or continuous age, period, and cohort. Not all interactions are estimable, and hence they cannot be added into the model at will.

5. Conclusions

Panel studies, cross-sectional studies with retrospective information, and replicated cross sections (including age by period arrays created from process-generated data) engender the analysis of a response variable as a function of age, period, and cohort as well as other factors. Such analyses must contend with the linear dependency between age, period, and cohort membership. The use of one or more measured variables held to underlie at least one of age, period, or cohort can break the linear dependency. So too can application of credible prior information, whether expressed as constraints on coefficients in fixed-effect models, or as exchangeability assumptions in random-effects or hierarchical models. Measured variables can, of course, be incorporated into both fixed-effect and random-effect models. This strategy is to be preferred, since it makes it possible to test ideas about substantive processes in the most direct way. Models that include age, period, and cohort can also include interactions between these dimensions, though not all such terms have estimable coefficients. Models that do not explicitly consider all three of age, period, and cohort, and yet are based on data structures that permit their inclusion, rest on the implicit assumption that age, or period, or cohort is irrelevant. This assumption should and can be assessed.

6. Further Reading

Fienberg and Mason (1985) discuss the formalization of the identification problem; identifiability of nonlinear components; identifiability of certain inter-

actions beyond those implicit in the simultaneous inclusion of age, period, and cohort; polynomial models; and other topics. Kupper et al. (1983) and Kupper et al. (1985) explore in depth several issues raised by the identification problem for the fixed effect discrete case, and take the stance (Kupper et al. 1985) that age-period-cohort models are no more informative than exploratory graphical displays. Ploch and Hastings (1994) illustrate the use of smoothed perspective plots. Robertson and Boyle (1998b) provide an overview of different strategies for graphical display of age-period-cohort data.

In Nakamura's (1986) Bayesian formulation, identifying linear restrictions are replaced by the assumption of exchangeability of first differences in age, period, and cohort effects (although Nakamura chooses to emphasize the closely related assumption of 'smoothness'). Hodges' (1998, Sect. 2) development of hierarchical models as linear models contributes to an understanding of how Nakamura's specification overcomes the singularity of the fixed effects model matrix. Berzuini and Clayton's (1994) Bayesian formulation, which is focused on second differences of effects, does not solve the identification problem. The social sciences literature in which Nakamura's specification is used or discussed (Sasaki and Suzuki 1987, 1989, Glenn 1989, Miller and Nakamura 1996) fails to make clear that it is not the Bayesian approach *per se* that provides an alternative to linear restrictions on coefficients, but rather Nakamura's particular formulation. Nakamura's (1986) computational approach has been supplanted by the use of Markov Chain Monte Carlo methods (The BUGS Project 2000; see *Monte Carlo Methods and Bayesian Computation: Overview*; *Markov Chain Monte Carlo Methods*).

Robertson and Boyle (1998a), and Robertson et al. (1999) review the largely disciplinary-specific epidemiological literature on the methodology of cohort analysis, which has focused primarily on additive models, and conclude that only the nonlinear components of age, period, and cohort can be used reliably. Holford et al. (1994) employ substantive reasoning about cell malformation in carcinogenesis, develop specifications in which age is inherently nonlinear (e.g., logarithmic) and thus eliminate the identification problem through choice of functional form. Mason and Smith's (1985) extended study of tuberculosis mortality combines substantive reasoning based on prior information and expectations, uses one body of data to guide modeling of another, includes an interaction term, employs a substantive, measured variable, and concludes that potential interactions require at least as much attention as the identification problem itself.

Hobcraft et al. (1985) focus on theoretical reasons for the use of age, period, and cohort in different areas within demography. Ní Brolcháin (1992) takes aim at the relevance of the cohort dimension for understanding temporal variation in human fertility. Discussions of this kind can help cohort analysts become more substantively grounded. Research using measured variables in place of accounting categories (e.g., cohort size instead of a cohort classification) is not, however, in need of such assistance (Easterlin 1980, Ahlburg and Shapiro 1984, Welch 1979). Blossfeld (1986) provides a persuasive example of the use of massive overidentification within a single dimension.

See also: Bayesian Statistics; Hierarchical Models: Random and Fixed Effects; Longitudinal Data; Longitudinal Data: Event–History Analysis in Discrete Time; Markov Chain Monte Carlo Methods; Statistical Analysis: Multilevel Methods; Statistical Identification and Estimability

Bibliography

Ahlburg D A, Shapiro M O 1984 Socioeconomic ramifications of changing cohort size: An analysis of U.S. postwar suicide rates by age and sex. *Demography* **21**: 97–108

Becker H A (ed.) 1992 *Dynamics of Cohort and Generations Research*. Thesis Publishers, Amsterdam

Berzuini C, Clayton D 1994 Bayesian analysis of survival on multiple time scales. *Statistics in Medicine* **13**: 823–38

Blossfeld H-P 1986 Career opportunities in the Federal Republic of Germany: A dynamic approach to the study of life-course, cohort, and period effects. *Eur. Sociol. Rev.* **2**: 208–25

Converse P E 1976 *The Dynamics of Party Support: Cohort Analyzing Party Identification*. Sage, Beverly Hills, CA

Easterlin R A 1980 *Birth and Fortune: The Impact of Numbers on Personal Welfare*. Basic Books, New York

Elder G H Jr 1999 [1974] *Children of the Great Depression: Social Change in Life Experience*, 25th Anniversary edn. University of Chicago Press, Chicago

Fienberg S E, Mason W M 1985 Specification and implementation of age, period, and cohort models. In: Mason W M, Fienberg S E (eds.) *Cohort Analysis in Social Research: Beyond the Identification Problem*. Springer-Verlag, New York, pp. 44–88

Glenn N D 1989 A caution about mechanical solutions to the identification problem in cohort analysis: Comment on Sasaki and Suzuki. *American Journal of Sociology* **95**: 754–61

Heckman J, Robb R 1985 Using longitudinal data to estimate age, period and cohort effects in earnings equations. In: Mason W M, Fienberg S E (eds.) *Cohort Analysis in Social Research: Beyond the Identification Problem*. Springer Verlag, New York, pp. 137–50

Hobcraft J, Mencken J, Preston S 1985 Age, period, and cohort effects in demography: A review. In: Mason W M, Fienberg S E (eds.) *Cohort Analysis in Social Research: Beyond the Identification Problem*. Springer Verlag, New York, pp. 89–135

Hodges J S 1998 Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). *Journal of the Royal Statistical Society: Series B* **60**: 497–536

Holford T R, Zhang Z, McKay L A 1994 Estimating age, period and cohort effects using the multistage model for cancer. *Statistics in Medicine* **13**: 23–41

Kupper L L, Janis J M, Karmous A, Greenberg B G 1985 Statistical age-period-cohort analysis: A review and critique. *Journal of Chronic Disease* **38**: 811–30

Kupper L L, Janis J M, Salama I A, Yoshizawa C N, Greenberg E 1983 Age-period-cohort analysis: An illustration of the problems in assessing interaction in one observation per cell data. *Commun. Statist.–Theor. Meth.* **12**: 2779–807
Mason W M, Fienberg S E (eds.) 1985 *Cohort Analysis in Social Research: Beyond the Identification Problem*. Springer Verlag, New York
Mason W M, Smith H L 1985 Age-period-cohort analysis and the study of deaths from pulmonary tuberculosis. In: Mason W M, Fienberg S E (eds.) *Cohort Analysis in Social Research: Beyond the Identification Problem*. Springer Verlag, New York, pp. 151–227
Miller A S, Nakamura T 1996 On the stability of church attendance patterns during a time of demographic change: 1965–1988. *Journal for the Scientific Study of Religion* **35**: 275–84
Nakamura T 1986 Bayesian cohort models for general cohort table analyses. *Ann. Inst. Statist. Math.* **38**: 353–70
Ní Bhrolcháin M 1992 Period paramount? A critique of the cohort approach in fertility. *Pop. Dev. Rev.* **18**: 599–629
Ploch D R, Hastings D W 1994 Graphic presentations of church attendance using general social survey data. *Journal for the Scientific Study of Religion* **33**: 16–33
Robertson C, Boyle P 1998a Age-period-cohort models of chronic disease rates. I: Modeling approach. *Statistics in Medicine* **17**: 1305–23
Robertson C, Boyle P 1998b Age-period-cohort models of chronic disease rates. II: Graphical approaches. *Statistics in Medicine* **17**: 1325–39
Robertson C, Gandini S, Boyle P 1999 Age-period-cohort models: A comparative study of available methodologies. *Journal of Clinical Epidemiology* **52**: 569–83
Ryder N B 1965 The cohort as a concept in the study of social change. *Am. Sociol. Rev.* **30**: 843–61
Sasaki M, Suzuki T 1987 Changes in religious commitment in the United States, Holland, and Japan. *American Journal of Sociology* **92**: 1055–76
Sasaki M, Suzuki T 1989 A caution about the data to be used for cohort analysis: Reply to Glenn. *American Journal of Sociology* **95**: 761–5
The BUGS Project 2000 *http://www.mrc-bsu.cam.ac.uk/bugs/*
Welch F 1979 The effects of cohort size on earnings: The baby boom babies financial bust. *Journal of Politics and Economics* **87**: 565–97
Wilmoth J R 1990 Variation in vital rates by age, period, and cohort. *Sociological Methodology* **20**: 295–335

W. M. Mason and N. H. Wolfinger

Cold War, The

The term is widely accepted in historical writings. It refers to the epoch between 1947–8 and 1989–90. The nature of the Cold War, its causes and effects, and the reasons for its long duration are, however, controversial. Its meaning would exclude Asia because of the outbreak of wars there; the wars in Korea, Vietnam, and Afghanistan made a difference to how the United States (US) and the Soviet Union (USSR), the two superpowers, waged their global contest in Europe. The situation in Europe resembled neither peace nor war; *East–West Conflict: Confrontation and Détente* expresses the ambiguity, but this label is less eye-catching.

1. *The Concept and its Salience*

The persistent central features of the *Cold War* are the global contest between the US and the USSR, the dependence of the allies on the security guarantee of their respective superpower, and bipolarity. The latter was reinforced in the core area, the arms race, due to the widening gap between the extensive and highly diversified weapons systems of the superpowers and the arsenals of Britain, France, and the Peoples' Republic of China (PRCh), the other three established nuclear powers. This explains that American and Russian writings hold to *Cold War* as a concept fitting the entire epoch. It allows for distinguishing between moments of imminent war—Cuba, October 1962, and Yom Kippur war, October 1973—times of high tensions and war outside Europe—Korea, 1950–53; Vietnam, 1964–75; Afghanistan, 1979–88—long intervals of *détente*, and even moments of concerted crisis management (June 1967 Near East war) or collaboration (during the Laos crisis and Vietnam war in the 1960s and ending wars in Africa in the late 1980s).

This version pays less attention to the fact that throughout the four decades not all parties to the Cold War were involved on the same issue-at-stake and at the same time. France joined Britain (UK) and the US in 1948 in founding the West German state as the Western allies' response to Stalin's anchoring of Poland, Bulgaria, Romania, Hungary, and Czechoslovakia firmly into the Soviet security-zone, and maintained staunch opposition to proposals for a neutralized, but nationally rearmed united Germany (Hitchcock 1998). Since the mid-1960s, France became the spokesman for a Europe less dependent on the US and discussing terms of settlement with 'Russia.' The Federal Republic of Germany (FRG) converted *Cold War* into *cold peace* with the ratification of *Ost- und Deutschlandverträge* (1970–3). Both governments cooperated in defending *détente* in and for Europe; by implication, this aspiration for a European peace-zone rejected the American concept of the indivisibility of the global contest between East and West. The PRCh, the only principal ally the USSR ever had (1950–58/9), made its peace with the US in 1972 on Pejing's term of 'One China.' While the PRCh had provoked the US during the 1950s to continuous nuclear sabre-rattling, it turned its antihegemonic posture throughout the 1970s and 1980s against the USSR.

1.1 *The Components of the Term*

The definition refers to four aspects of the conflict: (a) the ideological antagonism between the Western concept of freedom of choice for the people of the domestic

regime and external alignments of their state and the Soviet-imposed monopoly of the 'workers' and land-labor party and subjugation to the community of socialist states; (b) the geostrategic struggle for bases of power-projection; (c) the domestic political contest, but almost exclusively within Western countries, about commitments to the military alliance or equidistance to both of the superpowers; (d) the dispute about where to draw the line between permissible and noncompatible elements in the policy mix of market-oriented and government-controlled economies. Of these four components, the last became least important for refueling the confrontation; the convergence of the mixed economies rather promoted the idea that *Osthandel*, credit facilities, and technology transfer might enhance not only 'liberalization' in the East European economies, but also generate transformation of the political system.

1.2 The Ideological Antagonism

The ideological antagonism was the first impulse for *confrontation*, but it was also subject to changing threat perceptions and shifts of emphasis in the balance between *confrontation* and *détente* in the overall relationship. The wider notion of *East–West Conflict* posits the Cold War as a distinctive period into the ideological struggle, originated in 1917–18, between the Wilsonian Impulse and Lenin's urge for peoples' democracy as the basis for securing peace (Link 1980). From this perspective, Stalin and Truman likewise took the lead in splitting the world along 'alternative ways of life.' Western leaders argued that the West had to learn its lessons from the vain hope to appease an expansionist regime. 'No more Munichs' informed the confrontation stance of *containment* as the West's answer to the external threat. The concomitant stance at the home-front is based on the thesis that authoritarian regimes abuse state-power against large parts of their own population; therefore Communists must not be given a chance to occupy crucial positions of government such as Interior, Public Transport, or Justice.

Conversely, Communist ideology maintained that *capitalism* on the one hand inspires the state to prosecute the working-class and thus breeds *fascism*, and on the other hand is inherently expansionist and therefore knows no boundary to its domination. The belief, however, that capitalist nations are bound to fight for defeating rivals, made Stalin expect that Britain would resist the US' aspiration to become heir to the British Empire. The *new feature* of the post-World War II international system is that the US and the UK were competitive partners in founding the International Organizations destined to develop and monitor rules of conduct for international trade, currency exchange, and development aid (Gardner 1969). The UK and France, but also Italy, expected the up-to-then reluctant US government to come to the rescue and infuse capital into the European economics as the only strategy which could render support to the parties willing to exclude Communists from governments (Lundestad 1986). The formation of Western-oriented governing coalitions ensured that the European allies, as well as Japan, cooperated with the US in the evolution of the International Economy, to which the USSR and the PRCh were not negotiating parties after the outbreak of the Cold War. Most allies did not comply, however, with the American demand to restrict 'trading with the communist enemy,' except in conspicious moments of Soviet war activities.

The ideological component became a wasted asset after its overexploitation during the first phase (1947–53). Khrushchev's condemnation of Stalin's death toll in February 1956 and the NATO allies' perception that the Kremlin was unlikely to launch a general war on European territory (1956/7) put a break on the momentum of ideology as a driving force in the East–West conflict. But it was on the Kremlin to end the system conflict. In 1987 Gorbachev rescinded the Breshnev doctrine and made it known that hardliners in Eastern European governments should not reckon with Moscow calling on the Red Army to back up unpopular communist regimes (Adomeit 1998).

1.3 The Long Duration of the Cold War

For explaining the longevity of the Cold War, the account therefore has to focus on the other two factors: the structures which emerged with the consolidation and reaffirmation of the divisions of Europe and of Germany, and the pecularities of the military–geostrategic balance between the two sides.

1.3.1 The divisions of Europe and Germany. The key feature is the hinge between the two *divisions*. The stark contrast between Stalin's refusal to consider concessions on Poland, Bulgaria, and Romania in exchange for American economic assistance and his demand to have a say on questions concerning primarily intra-West relations, e.g., the status of the Ruhr area or of Norway, provoked the US and the UK to tie their zones firmly to Western institutions (Deighton 1990). The implementation of this basic strategy was, however, complicated by clashes of interest between the US, France, and Britain. These tensions prevented them from advancing as far as the USSR had with incorporating its German state, the GDR, into the Soviet Empire. Hence they could not, as wanted, negotiate from a position of strength, but expected to be asked to give away on what 'the West' did not yet have, whereas the USSR would persist in its refusal to put its reign over Eastern Europe on the agenda. The stalemate was compounded, when the reimposition of the Ulbricht regime in 1953 revealed that the Kremlin considered the GDR as its west-side lever for control over Poland and Czechoslovakia.

The structural impediment to a negotiated settlement disappeared when Moscow tacitly bowed to Germany's entry into NATO (1955) and when the 'West' acknowledged the fact that the USSR was strong enough to prevail in Eastern Europe (1956). German *Ostpolitik* put the final stamp to the 'normativity of the facts.' On this platform, the so-called Helsinki process, the Conference on Security and Cooperation in Europe (CSCE), lingered on throughout the 1973–89 period. Neither side was ready to proceed towards peace-making.

The Cold War might have ended earlier if Gorbachev's predecessors had responded to the US view that Germany's double-bind into a US-centered NATO and European integrated structures would prevent Germany from positing a threat to the security of the USSR (Schmidt 1993–5). Instead, Moscow's attempts to push the divisiveness of nuclear diplomacy within NATO rather than wait for the outcome of such conflicts provoked the 'Atlanticists' to close ranks and reiterate the standard thesis that the West could prevent the USSR from winning the Cold War if its members resisted the temptation to court Moscow and make separate deals with the USSR. The Soviet leadership's 'stupidity' is said to have saved the US or NATO to get rearmament projects—such as NSC-68 or the 1979 dual-track decision—through.

1.3.2 The peculiarities of the military–geostrategic balance. The key feature is the fundamental asymmetry between the nature of America's and Russia's predominance in their respective sphere. This asymmetry is reinforced by the imbalance of military forces between NATO-Europe and the USSR's forward deployed forces (Kugler 1993).

The advantage of the US in its global contest with the USSR was that all other principal powers, including Germany and Japan, were allied to the US. This helped in the build-up of the international economy, but not necessarily in defense. Consequently, US diplomacy was absorbed as much in intra-West crisis management as in the context of East–West relations. The allies wanted to be assured that there were no long-term security risks involved for them in America's option for sponsoring the resurgence of the enemies and occupiers of World War II as strongholds of the West. Having imposed limits on Germany's and Japan's military status, the US could not reckon with a defense contribution for some years to come; both were prohibited to engage in 'out-of-area' defense activities (Schmidt and Doran 1996).

By contrast, the USSR gained a formidable ally in the PRCh, who was keen to test the credibility of the American commitment to South Korea and Taiwan. Although the USSR copied the West's community building by setting up COMECON (1949) and the Warsaw Treaty Organisation 1955, the Kremlin relied for exerting influence *de facto* on the penetration of the bloc partners' central agencies: the State Party, Secret Service, and top military echelon. The USSR also made sure through logistic measures that the Soviet military could operate from its allies' territory independently, and if need be without the consent of the incumbent governments (Wolfe 1970).

The military imbalance between NATO-Europe and the USSR existed throughout 1947 until the mid-1980s. The US found no takers—except the FRG from 1961 onwards—for its concept of balanced collective forces in the sense that the US were the sole provider of strategic forces, that is, weaponry projected to hitting targets located in the USSR, whereas the allies would recruit the ground forces and tactical air forces required for the defense of the region against infiltration or surprise attacks. Britain and France preferred to duplicate the US strategic role, albeit on a much minor scale, and took the relaxation of tensions in Europe and their entanglement in the legacies of their imperial rule as the rationale for reducing their force assignments to NATO.

2. *The Intersection of Economic and Strategic Decisions: The European Structure*

The failure of the 'Big Four' to cooperate on a German Peace Treaty is identified as a crucial turning point towards the Cold War (Leffler 1992). Stalin did not want to give the impression that the US, thanks to the atomic bomb and its economic wealth, could impose its will on the USSR; he therefore pressed his claims on Iran, Turkey, during the Peace Conferences with Germany's war-time allies and in the Allied Control Council. This in turn was taken in Washington as a signal that hope of Russian cooperation must be abandoned. Truman and Byrnes had been reluctant to confront the USSR when the Cold War started over the fate of Eastern European nations (1944–7). Now, in view of the havoc the 1946–7 crisis wreaked on the economies of Britain and France, the US government conceived that Western Europe could not save itself (Hogan 1992).

From there on, the 'West' displayed a dynamic of its own. The first act was that the economic recovery of Europe was said to require the inclusion of (West) Germany; the price for getting France to change its German policy was the assurance that the US and the UK would back up the French quest for security against a resurgent Germany. The second act followed: Britain and France pleaded with Washington that the Marshall Plan initiative would not suffice to attain stability without an American security guarantee. Against the background of the Berlin blockade, Truman in October 1948 ordered that the US provide the major counterbalance to the 'ever-present threat of the Soviet military power.' In September 1948, the Brussels Treaty Organization (BTO) had already resolved that the defense of Western Europe should be

as far to the east (of West Germany) as possible. Because this was beyond the means available to the BTO and not yet compatible with the military strategy of American and British defense planning, it was only a question of developments in the Cold War that the third issue be placed on the agenda: German rearmament. The logic behind the build-up of conventional forces sounded compelling: why do the old allies expect the US to provide ground and air forces and the Supreme Allied Commander to NATO's integrated forces with a view to realize the concept of forward defense when they deny the indispensibility of a sensible German defense contribution, on whose territory the French in particular wanted to stop the enemy's offensive? At that time, however, all the US military could offer was an air offensive to deter the USSR; they did not yet want to rely on 'the bomb' (Ross), and the USSR had acquired by then (Fall 1949) an atomic capability.

The outbreak of the Korean War generated the fear of a similar war-by-proxy in divided Germany. The US did not only reverse its stance on what countries were of absolute importance to US security by now declaring Korea the test case of what became the domino syndrome, that no US ally should fall prey to a Communist invader, but also designated the National Security Paper NSC-68 as the platform for a massive conventional rearmament of the US, adequate force deployment abroad and military aid to upgrade the defense capabilities of its European and Asian allies (May 1993).

The issue of German rearmament was divisive, but the US compromised with France on the understanding that German forces be integrated into a European Defense Community (EDC), whereas the EDC would delegate strategic planning and fixing force requirements to NATO. The project became the victim of French domestic politics; the US and the UK had to comply with the French request that no German soldier be officially recruited until France had ratified the EDC and German Contractual Agreements. The demise of the EDC meant that Washington had to forego the hope that US force deployment would be a temporary stop-gap until the European allies, and especially the FRG, were able to do more towards meeting NATO's minimum force requirements.

The parallel staged fourth act (1952–4) produced the *definitive structure of the Cold War in Europe*: the NATO, to which the FRG was finally admitted, had made itself dependent on the credibility of the US (and UK) strategic deterrent. This provoked the issue how the USSR would react to the logic that NATO's new member would have to be equipped with the same weaponry as the allied forces into which the German divisions were to be incorporated. Chancellor Adenauer made it clear that the FRG did not want to provoke Moscow unnecessarily by dislocating medium-range ballistic missiles on German soil, whose later replacements might be able to hit targets in Russia, but he and his successors insisted on having at least a crucial German vote in all decisions on the use of nonstrategic weapons in NATO's custody in case deterrence to all types of war failed (Heuser 1998, Trachtenberg 1999).

This final act of incorporating the two German states into alliances dominated by the nuclear-strategic superpowers had a far-reaching effect: plastering the 'front-line' states with short-range atomic weapons was the best insurance premium that the Germans would not launch war from their soil (Ullman 1991). As long as the US and USSR stationed troops there and resolved to maintain control over the warheads, the danger of accidental war could be excluded. The parallel to the military factor written into the strategic landscape is the interest of the Western nuclear powers in confirming the *status quo* and discuss on that territorial basis the questions of putting ceilings on rearmament. Separately, the two German states relaunched their rivalry and demanded (1955–73) allies and Third World countries alike to subscribe to the *Alleinvertretungsanspruch* of the Western democratic or peoples' democratic German republic.

3. *The History of the Concept, Major Developments, and Empirical Results*

The invention of the term is accredited to Walter Lippmann who took issue with G. F. Kennan's 'X'-article 'Sources of Soviet Conduct.' Kennan's long telegram of February 22, 1946 molded the agenda of America's containment policy. His arguments were selectively used by top officials in Washington to assert that the 'Soviets' will develop all means and methods ranging from threat of military aggression via propaganda warfare to clandestine activities to a degree without precedent in history. Therefore, the US had to strengthen its executive branch, introduce a national security policy, establish a professional intelligence agency, and develop an unassailable military–industrial base (Leffler 1992, Paterson 1979).

The opposite Communist view emphasizes three aspects: the economic aggressiveness of the long-standing hegemonic US project to attain a 'one-world market' and abuse of the open-door doctrine for penetrating other nations' economy; the manipulation of anti-communism as a club to suppress claims for social justice and equality of rights within the American and other Western capitalist systems; and America's acquisition of bases bordering on the USSR for purposes of encirclement and monitoring inside the USSR. *In geostrategic terms the US became a direct neighbor of the USSR*, whereas the USSR required long-distance bombers (it introduced such jets by 1955) and Inter-Continental Ballistic Missiles (ICBM) to show the US its vulnerability. The strong engagement of the US in the post-1990 contests about Caucasian and Middle East oil and gas concessions

and pipeline routes indicates that the US–USSR rivalry did not end with the Cold War.

The 'Revisionists' (LaFeber 1991, Paterson 1979) disclaim that Stalin pursued an offensive strategy aiming at extending Soviet rule; they discern rather the US' responsibility for causing the Cold War. The postrevisionist approach accounts for the Soviet aspect of the fundamental change in the Cold War since about 1954–5, and explains the US interest in *détente*, i.e., the coinciding interest in *Long Peace* (Gaddis 1987). Moscow settled the Austrian question, tolerated the FRG's entry into NATO, and opted for reconciliation with Tito, the only East European leader who had got away with breaking ranks with Stalin. The cumulation point was the declaration of the doctrine of *peaceful co-existence*. The Western allies' match was tacit satisfaction with consolidating the *status quo* after they had absorbed the FRG. Stability in Europe turned the UK's, US's, and France's attention to the Third World. In reaction to Communist China's revolutionary foreign policy, the Soviet leadership in 1960 declared wars of liberation from colonial rule legitimate and thus distinguished peaceful co-existence in the developed world from just wars in the 'southern' part of the globe. In this sense, the East–West conflict was exported to the Third World.

The US resolved on noninterference in the internal affairs of the Soviet bloc and relaxation of tensions. The US was somewhat dependent on Russia's resistance to become the air and nuclear strategic arm of Pejing's violent anti-American activities in the Far East. The emerging picture is that the US, in parallel with the Berlin and Cuba crises (1958–62), wanted to concert the superpowers' activities in China's '*hinterland*' (former 'Indochina') in the sense that both exert pressure on the parties to a conflict amenable to their respective influence and through such agreement attain the neutralization of the conflict area, including the instalment of all-party or power-sharing coalition governments, however unstable (Nelson 1995).

3.1 European Perspectives on the Cold War

A different strand in the history of the concept is the perception of the impact of the other Western powers, especially Britain and France, on how the struggle was waged (Greenwood 2000, Bozo 1991). The contribution is twofold: (a) the *Cold War* is viewed as a new stage in great power rivalry. These studies stress the Europeans' interest in devoting resources to restoring or preserving their assets and commitments overseas (Kent 1993, Bossuat 1992). (b) The second contribution takes a different direction: it challenges the assumption that Western Europe by 1947–8 was in such a critical state that the leading nations had to 'invite' the US to reconstruct and protect Western Europe. Examining the economic potential of the European nations and how their governments used the financial means provided under the Marshall Plan, Milward (1984) presents the thesis of the rescue of the nation-state. His complex argument also draws on the observation that at that time the West did not reckon with an immediate Soviet threat. The implication is that an unwarranted haste was imposed on a process which in any case depended on what the European countries did to and for another, formally, US assistance was tied to the Uniting of (Western) Europe and to parallel bilateral agreements between the US and the recipient country.

3.2 Competitive Cooperation Between the Superpowers

By the end of the 1960s, researchers discerned the convergence of five developments which demonstrated the continuing relevance of the 'competitive cooperation' between the superpowers for the changing structures of the *Cold War*: (a) Soviet policy shifted towards accepting the US's and Canada's presence in Germany; (b) German *Ostpolitik* presumed the engagement of the US in NATO; (c) de Gaulle realized in the context of the Prague 1968 crisis that Breshnev's Russia was no party to his vision that dissolving Cold War structures in the West, i.e., France's disengagement from NATO's military organization, might induce 'Russia' to allow for more evolution of national communist and then independent states in Eastern Europe; (d) the US embroilment in the Vietnam war caused a change in the intellectual climate; the abuse of the USSR being the cause of every evil became obsolete; instead the US became the villain in the piece; (e) new developments in arms technology—e.g., high-precision weaponry as a substitute for atomic weapons; MIRV-technique; antiballistic missiles—were beyond Britain's and France's capability to follow on; this reinforced NATO Europe's self-elected dependency on the US with respect to security and defense (Hanrieder 1989).

Renewed US pressures on its allies to extend their conventional military capabilities and share the burdens for the 'Defense of the West' more evenly, the re-escalation of the strategic arms race between the superpowers under the aegis of limitation treaties, and the evidence presented by German *Ostpolitik* that negotiations with Russia generate tolerable results, combined to raise the basic questions: what are the costs of the ongoing *Cold War*? on what terms could the conflict be ended and converted into politico-economic competition? (Garthoff 1985).

3.3 From the Second Cold War to the End of the Cold War

The USSR, after getting Germany's pledge to observe the invulnerability of the territorial *status quo* in Europe, revived the global contest with the US, which

was immersed in domestic turbulences *and* disputes with all its allies about oil and the dollar. The Kremlin demonstrated its self-confidence by expanding Soviet naval forces and establishing bases, however short-lived, stretching from Vietnam via Mozambique, Somalia, and Angola to Central America. In contrast to the 1950–3 period, the US catchword 'arch of crises' did not resonate well with its allies. The latter wanted to develop *détente*, notwithstanding the fear, expressed by German Chancellor Helmut Schmidt in 1977, that the Soviets' new equipment (Backfire bombers; SS-20) made NATO Europe hostage to the Kremlin's whip of the will. This raised the question whether the USSR was about to win the Cold War or whether the Kremlin was overstretching the country's resources and would thus expose the USSR sooner rather than later to the need for radical changes in her system. The wishful thinking that the USSR would fall victim to its inability of continuous adaptation worked out after three decades of aspiring military strategic and political parity with the US. Some authors argue that Reagan's defence build-up deliberately forced the Soviet leadership to acknowledge the failure of its domestic system and hence the USSR's inability to persist in the contest and sustain bipolarity. (Adomeit 1998, Gaddis 1987, Wells' article in Hogan 1992). The way the global contest ended invited Americans to believe that they had won the Cold War.

4. *Methodological Problems*

The immersion of the concept *Cold War* in the perpetual clash of interests between East and West and within each 'bloc' subjects the interpretator on the one hand to the political climate of his own times, and may thus reload or de-emphasize the contentions of the past; on the other hand, access to newly available records reveals new insights into previous phases of the Cold War and thus demands rethinking the past but in a different way. Getting the balance right between these two operations is difficult when knowledge about one party to the conflict is the result of generations of research, whereas the more recent presentations of findings from Soviet, GRD, Czech, or Polish sources are both less systematic or structured and more exposed to be taken instantly as evidence for one interpretation e.g., of Stalin or Krushchev or another. The task to study the records comprehensively, but also be prepared to modify one's assessments in response to newly available empirical evidence is prone to collide with the other obligation, namely to explain what were the basic causes of a conflict and which causes—meaning in politics: sins of commission and omission—are accountable for what developments.

A second main problem is the exact phase-by-phase and overall intersection between the interpretation of the collaboration leading to the evolution of 'Western' structures in the economic sphere and the assessment of the concomitant events in the politico-strategic (military) sphere, where the USSR appeared to be the winner, but then ended the Cold War on western terms.

5. *Future Direction of Research*

'East' and 'West' waged the struggle by all means and methods except 'hot' war in the area stretching from Vancouver to Wladivostok. Hence, research must attend to different subject areas and to the expertise developed in the many scholarly communities. Some areas of research, e.g., diplomatic or intellectual history and biography, are more established than others. The impact of *intelligence* on policy-makers is a relatively new area of systematic research. International cooperation projects have done much to promote *nuclear history* and *strategic studies*, but the history of the military alliances and of national defense organizations and policies depend still on the authorities' grant of access as well as clearance and permission for publishing inspected material. The studied ambiguity of the political leaderships and the top military echelon about the worth of strategic nuclear weapons in case deterrence should fail and about the use of short-range atomic weapons deserves further study in order to know the implications of nuclear weaponry on the conduct of Cold War diplomacy and governmental guidance to their military.

Future research should be more systematic in the sense of *extending* the conceptualization beyond the national and bilateral focus to the regional context and *intensifying* the approach by addressing the fundamental questions about the changing nature of the struggle, its costs, the persistence or recreation of patterns of conflict, and above all the question whether the Cold War structure affected all other bilateral or intraregional conflicts or whether re-emerging older conflicts pervaded the East–West conflict, so that the parties to such conflicts used the *Cold War* for the purpose of engaging wealthy allies on their side.

See also: Arms Control; Communism, History of; Contemporary History; Diplomacy; Eastern European Studies: History; Eastern European Studies: Politics; International Relations, History of; Military and Politics; National Security Studies and War Potential of Nations; Peacemaking in History; Revolutions of 1989–90 in Eastern Central Europe; Second World War, The; Soviet Studies: Politics; War, Sociology of; Warfare in History

Bibliography

Adomeit H 1998 *Imperial Overstretch: Germany in Soviet Policy From Stalin to Gorbachev*. Nomos Verlagsgesellschaft, Baden-Baden, Germany

Bossuat G 1992 *La France, l'aide américaine et la construction européenne, 1944–1954*. Impr. nationale, Paris
Bozo F 1991 *La France and l'OTAN. De la guerre froide au nouvel ordre européen*. Masson, Paris
Deighton A 1990 *The Impossible Peace: Britain, the Division of Germany and the Origins of the Cold War*. Clarendon Press, Oxford, UK
Gaddis J L 1987 *The Long Peace: Inquiries into the History of the Cold War*. Oxford University Press, New York
Gardner R N 1969 *Sterling–Dollar Diplomacy. The Origins and the Prospects of Our International Economic Order* New expanded edn, McGraw-Hill, New York
Garthoff R L 1985 *Détente and Confrontation: American–Soviet Relations from Nixon to Reagan*. Brookings Institution, Washington, DC
Greenwood S 2000 *Britain and the Cold War 1945–91*. Macmillan, Houndsmill, UK
Hanrieder W F 1989 *Germany, America, Europe: Forty Years of German Foreign Policy*. Yale University Press, New Haven, CT
Heuser B 1998 *NATO, Britain, France and the FRG. Nuclear Strategies and Forces in Europe, 1949–2000*. Macmillan, Basingstoke, UK
Hitchcock W I 1998 *France Restored. Cold War Diplomacy and the Quest for Leadership in Europe, 1944–1954*. University of North Carolina Press, Chapel Hill, NC
Hogan M (ed.) 1992 *The End of the Cold War*. Cambridge University Press, Cambridge, UK
Kent J 1993 *British Imperial Strategy and the Origins of the Cold War, 1944–1949*. Leicester University Press, Leicester, UK
Kugler R L 1993 *Commitment to Purpose. How Alliance Partnership Won the Cold War*. RAND, Santa Monica, CA
LaFeber W 1991 *America, Russia, and the Cold War, 1945–1990*, 6th edn. McGraw-Hill, New York
Leffler M P 1992 *A Preponderance of Power: National Security, the Truman Administration, and the Cold War*. Stanford University Press, Stanford, CA
Link W 1980 *Der Ost-West-Konflikt. Die Organisation der internationalen Beziehungen im 20. Jahrhundert Kohlhommor*. Stuttgart, Germany
Lundestad G 1986 *East, West, North, South: Major Developments in International Politics, 1945–1986*. Oxford University Press, New York
May E R (ed.) 1993 *American Cold War Strategy: Interpreting NSC 68*. Bedford Books, Boston
Milward A S 1984 *The Reconstruction of Western Europe, 1945–1951*. Methuen, London
Nelson K L 1995 *The Making of Détente: Soviet–American Relations in the Shadow of Vietnam*. Johns Hopkins University Press, Baltimore, MD
Paterson T G 1979 *On Every Front: The Making of the Cold War*. Norton, New York
Schmidt G (ed) 1993/1995 *Ost-West-Beziehungen: Konfrontation und Détente 1945–1989,* 3 Vols. Brockmeyer Bochum, Germany
Schmidt G, Doran C F (eds.) 1996 *Amerikas Option für Deutschland und Japan. Die Position und Rolle Deutschlands und Japans in regionalen und internationalen Strukturen. Die 1950er und die 1990er Jahre im Vergleich*. Brockmeyer, Bochum, Germany
Trachtenberg M 1999 *A Constructed Peace. The Making of the European Settlement 1945–1963*. Princeton University Press, Princeton, NJ
Ullman R H 1991 *Securing Europe*. Adamantine Press, Twickenham, UK
Wolfe T W 1970 *Soviet Power and Europe, 1945–1970*. Johns Hopkins University Press, Baltimore, MD

G. Schmidt

Coleman, James Samuel (1926–95)

James S. Coleman was born on May 12, 1926 in Bedford, Indiana. He died on March 25, 1995 in Chicago, Illinois. Coleman was among the most important American sociologists of his generation. By importantly influencing both intellectual and policy debate, Coleman was unique among sociologists. By making important contributions to a large range of areas of scholarly concern, including mathematical sociology, sociology of education, and social theory, Coleman was unique among social scientists.

1. Life

James Samuel Coleman was born into a family of teachers, with roots in the landed gentry of the South on his father's side, and in various places in the Mid-West on his mother's side. His grandfather, Samuel, was a minister. The family moved frequently in Coleman's youth between places in Ohio, Arkansas, and Kentucky to settle, finally, in Louisville, Kentucky, where Coleman graduated from Manual High School in 1941. The origin is important for his basic sociological interests and positions. The marginality from moving around between southern and northern American cultures and the diversity of his origins instilled a strong curiosity about social relations. The teacher occupations of both parents created a strong interest in schools and education. Samuel, the minister grandfather, and the mother were important for Coleman's preoccupation with moral issues that profoundly influenced his work.

Coleman's choice of sociology as a vocation came quite late. He graduated from Purdue University in 1949, with a degree in Chemical Engineering, and his first job was as a chemist with Eastman Kodak. He had almost no undergraduate education in any social science. Nonetheless, in 1951 he began graduate study in sociology at Columbia University. Coleman's dual attraction to science and moral engagement made sociology an impeccable choice, or so it would seem in 1951. He had found industry frustrating and a likely eventual career in management unappealing. He wanted to devote his life to discovery and concluded it could only be about people, their relationships, and their social organization.

Columbia's sociology department gave Coleman four intense years and three important teachers: Paul

Lazarsfeld, Robert Merton, and Seymour Martin Lipset. Coleman is usually regarded as Paul Lazarsfeld's student. This is not quite correct. Lazarsfeld was not his dissertation advisor; it was Lipset. Lazarsfeld was not the teacher who had the most profound intellectual influence on Coleman; it was Merton. Lazarsfeld did involve and use Coleman for the development of mathematical and statistical tools for social analysis, and these activities created the point of departure for some of Coleman's most important later work. However, there is an important difference between Coleman's *Introduction to Mathematical Sociology* (1964) and Lazarsfeld's branch of mathematical thinking in the social sciences. Coleman's main objective with the use of mathematics was the development of theoretical insights and conceptual development. Lazarsfeld's major contributions were to the codification of research procedures, that is, methodology. Coleman has made important contributions to methods, but his most remarkable quality as a sociologist was his ability to develop sociological ideas and sustain them with empirical evidence. This is much closer to Merton's style of theorizing about empirical matters, though Merton often relied on evidence produced by others.

Coleman obtained his first faculty position in 1956 as an Assistant Professor at the University of Chicago. In 1959 he moved to Johns Hopkins University to create his own sociology department. He developed a small organization with an intellectual intensity and excitement that was truly remarkable. It was perhaps unsustainable and, for a variety of reasons, Coleman went back to the University of Chicago in 1973, now as a University Professor, and stayed there until his death in 1995.

2. *Scholarship*

The body of scholarly work is large. It includes 30 books and monographs and more than 300 articles and chapters in books. These contributions have profoundly influenced and, in several cases, defined the agenda for important subfields of sociology: sociological theory, sociology of education, sociology of the family, mass-communication, social stratification, political sociology, mathematical sociology, and sociological methods. The contributions include three major pieces of policy-oriented research on schools and educational opportunities that are the major examples of 'sociology making a difference' in the 1960s, 1970s, and 1980s. The scholarly work covers a phenomenal range of approaches to social research. There is work about social systems (Coleman 1961) and about individual behavior (Coleman 1990b). There is basic research as well as applied. There is quantitative (Coleman et al. 1966, Coleman 1987) as well as qualitative analysis (Coleman 1961). There are contributions to economics (Coleman 1966), political theory (Coleman 1982), moral philosophy (Coleman 1974), statistics and probability theory (Coleman 1981), and education (Coleman 1990a).

It is difficult to characterize James Coleman's contributions to the discipline because there are so many components. It includes two major and two very different paradigms of what sociology is about. One is a Durkheimian paradigm: it sees the task as studying how social structure creates and constrains individual action and causal social processes. The other paradigm takes departure in individual purposeful action for developing properties of social systems and structures. The former project governs most of Coleman's empirical work on educational processes and social processes in educational institutions. The latter project, the development of a sociological, rational action theory, moves, so to speak, in the opposite direction. It is theory aimed at understanding social systems themselves, their development, and properties, beginning with a theory of individual action. It occupied most of Coleman's attention in the last years of his life.

Almost all of Coleman's empirical research was about growing up, about schools, and learning and educational opportunity. His first study of schools was a study of adolescent subcultures based on surveys of students in ten high schools in northern Illinois. The report on this research was published in 1961 as *The Adolescent Society* (1961). In the preface, Coleman notes two reasons for undertaking this study. One was his wish to study how schools might be more effective; he states that the study had its origin in 'a deep concern I have had, since my own high school days, with high schools and with…ways to make an adolescent experience with learning more profitable…' (1961, p. viii). The second concern was an interest in different types of social systems and the value systems they reflect. These two interests, in creating effective schools and other social organizations and in the values and cultures of social systems, would guide Coleman's research throughout his life.

The wish to study the effects of high schools on learning did not receive much attention in *The Adolescent Society*, possibly because the empirical evidence Coleman obtained about these effects, and their relationship to status systems, is quite difficult to interpret. *The Adolescent Society* is primarily a study of adolescent subcultures. The policy prescriptions Coleman drew from his analysis were about how to change these subcultures to encourage learning and academic achievement. This would be done by changing the structure of competition so that students would see academic achievement as a collective good, like sports.

The interest in the effects of schools and school social organizations on learning became the main preoccupation of Coleman's educational research for almost four decades. The link between the two concerns was always clear, and the interest in the social systems created in schools re-emerges as the

dominant concern again in his latest research. All of his educational research sees the effectiveness of schools related to the social systems created in schools and in the family and community that form the context for the school social systems. Coleman was very much a sociologist of schools.

Coleman performed three major studies of schools and educational processes after *The Adolescent Society*. The first so-called 'Coleman Report' (for a while referred to by many as the Coleman–Campbell report) was a report on the massive *Equality of Educational Opportunity* survey (Coleman et al. 1966). It is probably the largest social science survey research ever undertaken with some 600,000 students surveyed. The Report's main finding was that the instructional facilities and resources, including per pupil expenditures, have small and comparatively insignificant effects on academic achievement relative to the influence of family background. This result created considerable controversy, a controversy that continues today. A very large number of studies have focused on the issue since then. The best of this research has replicated the main conclusion: there is very little relationship between the amount of expenditure on schools and achievement outcomes, net of the family backgrounds of students. Next to family background, the family background of a student's peers, as measured for example by the racial composition of schools, seems most relevant, especially for minorities. This conclusion had an important consequence for American schools as it was used to justify racial desegregation in public schools. The second Coleman Report focused on this effect on policy: the use of busing to integrate schools and reap some of the benefits of desegregation on the academic achievement of minorities. The Report concluded that the policy had been ineffective. This conclusion was widely interpreted as denying the benefits of desegregation documented in the First Report and was strongly attacked by many, including many of Coleman's fellow sociologists. The interpretation was not correct. Coleman did not doubt the relevance of social systems created by student body composition for schools' educational climates. He doubted the long-term benefits of busing as a remedy for segregation because it encouraged white flight from desegregated public schools.

The allegation that Coleman had reversed his position would be repeated with the publications from the third major piece of educational research published in the early 1980s. In a series of publications, for example *Public and Private High Schools* (Coleman and Hoffer 1987), Coleman and co-authors concluded that Roman Catholic parochial schools produced more learning and superior educational outcomes than public schools because of superior social resources, or social capital created by the involvement of parents with schools and their personnel. These conclusions resulted in another set of published controversies, now about the existence of school effects—the very effects that were said to have been denied by Coleman in the first Report. Now, the tables had been reversed. The critiques now denied the existence of school effects, in particular the apparent superiority of Catholic schools.

Coleman strongly believed in the use of mathematics as a language for the development of social theory. His *Introduction to Mathematical Sociology* (1964) remains unsurpassed as a text on how to use mathematics, in particular stochastic process models, to mirror social processes of diffusion and influence in groups. It is an enormously rich collection of tools for the analysis of social processes. Coleman remained committed to the use of mathematics for the development of theory. Over time, he changed the type of theory focused upon, from modeling the mechanism of social processes in his early work, to modeling the outcomes of purposive action in his later work. Throughout, Coleman's contributions provide perhaps the most important demonstration of the power of mathematics as a tool for social theory in contemporary sociology.

Already in the 1960s, Coleman published several articles on a model for collective decisions where actors trade control over events and resources to maximize their interests. This work came about as a result of Coleman's interest in developing games for learning and for social theory, where players engage in transactions designed to mirror important learning tasks or to simulate basic rules of social systems. One of these games was called the legislative game, and was designed to simulate the transactions going on in a legislative body with vote trading. Coleman formalized this game in a simple model that provided equilibrium outcomes providing measures of the power of actors and the value of events. These ideas became the basis for what became his major preoccupation in the 1970s and 1980s: creating a rational action basis for social theory where the emergence of macro social structures are clearly linked to micro foundations.

The outstanding result of many years of effort became *Foundations of Social Theory* (1990). It is a major book in ambition, achievement, and size. It provides theory and theoretical tools for the analysis of existing society and the creation of better societies. The book moves from the analysis of elementary action and social relations of trust and authority to the development of norms and more complex authority systems or corporate actors, such as large corporations, and then on to the analysis of the major institutions of modern society. A major concern for Coleman was the distribution of power in society and the problem of holding corporate actors responsible for their actions. *Foundations of Social Theory* is an unfamiliar contribution to a discipline where theory often seems to be an exegeses about the theory created by those who are safely dead, or abstract conceptual development with little if any empirical reference. This

major book aims to shape the discipline by providing a theory and a mathematical structure that may have extraordinary potential for research. The realization of this potential depends not only on the quality of the ideas, but also on the discipline's ability to retool. Some concepts from the book have become prominent in sociology, for example Coleman's ideas about the micro–macro link and about social capital. However, this is only a part of this enormously rich contribution. Coleman knew that an extraordinary educational effort was needed, and he devoted much of his attention to this effort in his last years, for example, by founding and editing the journal *Rationality and Society*.

Coleman's theoretical work created a landmark in social theory. His concern for developing a mathematical language for the analysis of empirical processes created a rich set of tools that eventually may change the way sociologists work. His empirical work often includes imaginative interpretation and brilliant insights. Though his emphasis increasingly was on the project that resulted in *Foundations of Social Theory*, Coleman repeatedly returned to empirical research on causal processes with individual actions and lives as the outcome. The synthesis was accomplished in some of Coleman's latest empirical work, on schools, family, and community.

3. *Impact*

There was no lack of recognition of Coleman's scholarly contributions by the bodies that confer the highest prestige upon scientists. Coleman was elected to the American Academy of Arts and Sciences in 1966, the National Academy of Education in 1966, the American Philosophical Society in 1970, the National Academy of Sciences in 1972, and the Royal Swedish Academy of Sciences in 1984. He was a Fellow at the Center for Advanced Study in the Behavioral Sciences, a Guggenheim Fellow, and Fellow at the Wissenschaftskolleg zu Berlin. He received numerous honorary degrees from universities in the USA and Europe.

Despite all the recognition, American sociology had, and still has, considerable ambivalence toward Coleman. His international reputation, particularly in Germany, the Netherlands, and Scandinavia, and his reputation outside of sociology seemed often more solid than his reputation in American sociology. He became President of the American Sociological Association (ASA) from 1991 to 1992, but he never held another elected office in the ASA. His election to the Presidency in 1990 was the result of a write-in campaign and not of a nomination by the Association. Indeed, the leadership of the ASA, including the ASA President, tried to censor him in the mid-1970s for producing subversive sociology, because his work on the impact of busing on white flight threatened traditional liberal beliefs. It was an ignominious act that almost succeeded.

The ambivalence of the profession toward Coleman had two main sources. One was Coleman's unwillingness to specialize in one of the usual three main roles for sociologists: as theorist, methodologist, or researcher. This contradicts the implicit theory most of us have that one cannot be outstanding in all three roles. Coleman was outstanding in all three, but specialists in theory, methods, or research had difficulties recognizing his achievements. The second, and probably major, source of ambivalence was Coleman's use of research to draw policy inferences. Coleman stated what his research meant for policy and was not hesitant to do so when it contradicted conventional wisdom. He loved controversy. Each of the three Coleman Reports stated a conclusion that infuriated many: that school resources have little impact on academic achievement compared to the family resources of a child; that busing to achieve racial integration speeds up the process of white flight from our central cities; that schools organized as many private Roman Catholic schools produce more learning and less inequality in learning than schools organized like the typical public school. In every instance, an army of researchers tried to find fault with the evidence for these conclusions and largely failed. They concentrated on specific statistical issues. This was a mistake when confronting Jim Coleman. He anticipated criticisms by demonstrating the main finding in several ways. Moreover, his powerful intuitions about the mechanisms that create observed outcomes produced coherent arguments that Coleman effectively integrated with empirical analysis, while the opposition failed to formulate an alternative argument that could be sustained with evidence.

Sociology is profoundly influenced by James Coleman. So are many individual lives and careers. None of his students and followers moved with him through all the projects, but remain in one or the other of his many areas of interests. Those who had the privilege to know him were always profoundly touched by his excitement of translating into empirical analysis theoretical ideas about how social structure affects individuals. Coleman's enormous mental (and physical) energy never ceased to amaze. This energy and creativity were sustained by his certitude about the importance of the project of making sociology a better tool for a better society.

Appraisals of Coleman's scholarly and policy contributions may be found in Sørensen and Spilerman (1993) and Clark (1996). The latter volume, published after Coleman's death, includes a complete bibliography and several autobiographical fragments.

See also: Adolescence, Sociology of; Affirmative Action: Empirical Work on its Effectiveness; Educational Sociology; Family and Schooling; Family as Institution; Public Policy Schools; Racial Relations; Rational Choice Theory: Cultural Concerns; Rational

Choice Theory in Sociology; Sociology, History of; Theory: Sociological

Bibliography

Clark J (ed.) 1996 *James S. Coleman*. Falmer Press, London
Coleman J S 1961 *The Adolescent Society*. Free Press, Glencoe, IL
Coleman J S 1964 *Introduction to Mathematical Sociology*. Free Press, Glencoe, IL
Coleman J S 1966 The possibility of social welfare function. *American Economic Review* **56**(5): 1105–12
Coleman J S 1974 Inequality, sociology, and moral philosophy. *American Journal of Sociology* **80**: 739–64
Coleman J S 1981 *Longitudinal Data Analysis*. Basic Books, New York
Coleman J S 1982 Re-contracting, trustworthiness, and the stability of vote exchanges. *Public Choice* **40**: 89–94
Coleman J S 1990a *Equality and Achievement in Education*. Westview Press, Boulder, CO
Coleman J S 1990b *Foundations of Social Theory*. Belknap Press, Cambridge, MA
Coleman J S, Campbell E Q, Hobson C J, McPartland J, Mood A, Weinfeld F D, York R L 1966 *Equality of Educational Opportunity*. Government Printing Office, Washington, DC
Coleman J S, Hoffer T 1987 *Public and Private High Schools: The Impact on Communities*. Basic Books, New York
Sørensen A B, Spilerman S (eds.) 1993 *Social Theory and Social Policy: Essays in Honor of James S. Coleman*. Praeger, Westport, CT

A. B. Sørensen

Collective Behavior, Sociology of

1. Introduction

The sociology of collective behavior begins simultaneously with the inception of modern sociology. Gustave Le Bon published *The Crowd in 1895*, two years before Emile Durkheim's *Suicide*, and two years after Durkheim's *The Division of Labor in Society*. Le Bon's descriptions and analyses of crowd behavior in the wake of the French Revolution and Paris Commune stimulated further study of crowds and, more generally, collective behavior.

Collective behavior is noninstitutionalized, unconventional group activity such as panics, crazes, mass delusions, incited crowds, riots, and reform or revolutionary movements. A sociological approach to collective behavior focuses on social conditions such as political structures and shared beliefs as these conditions influence patterns of collective behavior.

Le Bon approached the study of crowds skeptically. He argued that individuals can lose their self-identity in crowds and can commit acts they would not do alone, including physical aggression. He felt the anonymity of crowds permitted passions to surface which could be translated into violent behavior. These ideas were parallel to the distinction made by Gabriel Tarde between a public that rationally considers political issues and a crowd which acts hastily without rational consideration of the matters at hand. Yet Le Bon also felt that crowds were progressive in that they challenged existing social arrangements and often led to social change that would not have occurred without them (Wood and Jackson 1982).

Even pre-dating Le Bon in 1856 was Alexis de Tocqueville who, in *The Old Regime and the French Revolution*, argued that the French Revolution resulted from social changes occurring over several centuries, culminating in the Revolution of 1789, aided by—using a current term—the widespread sense of relative deprivation among the bourgeoisie (French middle class) and sans-culottes (French masses). He posed a paradox, and then resolved it, by indicating that the country less constrained by classical feudalism, France, was where the Revolution occurred instead of the more constrained England or Germany, for example. This occurred because France, by gradually dismantling feudalism, had improved the political and economic circumstances of the bourgeoisie and the *sans-culottes* over the centuries, but these people still felt deprived because it was clear the Old Regime would permit only so much advancement for them. This sense of deprivation relative to what they felt was their just station in life—that their expectations were significantly different from their experience–motivated participation in this monumental movement that toppled the Old Regime and became the exemplar for many subsequent revolutions.

2. Smelser's Theory of Collective Behavior

The modern, post-World War II, sociology of collective behavior is initiated by Neil J. Smelser's *Theory of Collective Behavior* (Smelser 1963). This book introduced a thoroughly sociological approach to the investigation of collective behavior. Smelser's theory contrasted with the previously dominant social psychological approach pioneered by scholars such as Robert Park and E. W. Burgess, Herbert Blumer, and Ralph Turner and Lewis Killian, who focused on topics such as crowd interaction, types of collective behavior, and emergent norms in undefined social situations.

Smelser, by stressing such variables as structural conduciveness and structural strain, along with the other variables in his theory—growth and spread of generalized beliefs, precipitating factors, mobilization of participants for action, and weakening of social control—highlighted the importance of sociological conditions as determinants of collective behavior.

In focusing on these six sociological factors, he moved the field of collective behavior away from

social psychological theories stressing single variables such as anonymity of crowds, economic deprivation, alienation, and strong leadership (Wood and Jackson 1982).

Smelser showed that these single variables had a place in understanding the development of collective behavior, but that they had to be seen in combination with other relevant variables in order to increase the likelihood of collective behavior actually occurring. Indeed, Smelser argued that the six variables in his system constituted necessary, or required, conditions for the occurrence of collective behavior; and when combined at the same time and place, the six variables constituted a sufficient, or fully predictive, condition for the occurrence of some type of collective behavior. Thus, when the six conditions occurred simultaneously, Smelser predicted that a craze, panic, riot, reform, or revolutionary movement would occur.

Smelser furthermore argued that different aspects of the six conditions, and their particular combinations, could help predict which of the five types of collective behavior mentioned above would be the one to occur (e.g., a reform or revolutionary movement). If, for example, there was considerable economic strain in society, the spread of radical political beliefs, precipitating factors such as the arrest of dissidents, attempts by movement leaders to mobilize politically large numbers of people, and the weakening effectiveness of agents of social control such as the police and army, Smelser's theory would then ask, in what kind of society did all this occur? If the society was democratic, Smelser argues that this kind of society would be structurally conducive to the development of a reform, or norm-oriented, movement such as the Civil Rights Movement in the USA. If instead the society was authoritarian, Smelser argues that this kind of society would be structurally conducive to the development of a revolutionary, or value-oriented, movement such as the Russian Revolution.

Smelser's theory attracted considerable attention from the start, and it generated a large empirical literature within a decade of its publication, much of which supported the theory (Marx and Wood 1975).

3. *Subsequent Approaches*

3.1 *Resource Mobilization*

Smelser's theory also attracted critics, some of whom developed their own approaches that form the basis of current collective behavior analysis. These other approaches will now be examined for their contributions to the sociology of collective behavior, many of which owe at least indirect inspiration from *Theory of Collective Behavior*, as is indicated by several themes such as conditions of conduciveness and strain in the seminal collection of articles covering the last decade of social movement research organized by McAdam and Snow (1997). As Klandermans (1997, p. 201) states, 'One need not concur with Smelser … to appreciate the usefulness of these categories and assess maybe with some surprise how akin they are to other more recent conceptualizations.'

The first well-known approach signaling not only a critique of, but also an attempted alternative to, *Theory of Collective Behavior*, was resource mobilization theory pioneered by Mayer Zald and John McCarthy. While retaining an interest in the origin of collective behavior and social movements—a key question for Smelser—the resource mobilization approach began to shift the basic question addressed. Resource mobilization, and other subsequent approaches, became particularly interested in social movement success or failure, and the conditions—such as availability of resources like money and media attention—that are associated with movement success or failure.

3.2 *Political Process Model*

Doug McAdam pioneered another important approach to understanding the success—as well as development—of social movements by focusing on the political process involved in such movement development and success. In his first book on the topic, *Political Process and the Development of Black Insurgency* 1930–1970, McAdam (1982) outlined, and documented the importance of, four conditions that he showed to be associated with the rise of militant black politics and black political achievements in the USA in this crucial 40-year period. These four conditions were as follows: (a) 'the degree of organizational "readiness" within the minority community'; (b) 'the level of "insurgent consciousness" within the movement's mass base'; (c) the 'structure of political opportunities'; and (d) the 'level of social control' (McAdam 1982, pp. 40, 53). Using historical examples to show variations in these four conditions, McAdam does a masterful job of explaining the ensuing black political movements of mid-twentieth century America.

In focusing on the four conditions, McAdam has both drawn upon, and extended, the scope of Smelser's six conditions of social movement development. There is overlap between McAdam's variables and Smelser's with, for example, organizational readiness being an aspect of mobilization of participants for action, and insurgent consciousness being an aspect of generalized beliefs. Yet McAdam also extends Smelser's model by using the four conditions to help explain the successes of black protest in the USA as well as understand its initial development in this historical period. Finally, McAdam, his associates, and others have further extended and applied the political process model to yet other instances of social movement activity (McAdam and Snow 1997).

3.3 Political Opportunity Model

One of the conditions in McAdam's model is 'the structure of political opportunities,' a concept elaborately developed by Sidney Tarrow. In a series of books and monographs, Tarrow (1991, 1998) has utilized the political opportunity model to understand the circumstances whereby contentious social movements can achieve success. Using a wide variety of case studies, often ranging from the social movements of the 1960s to the movements, conflicts, and changes in Eastern Europe in the 1990s, Tarrow shows how a variety of political structures have influenced the success, or lack of success, of these social movements.

Drawing on the seminal work of Michael Lipsky and others, Tarrow (1991, p. 32) defines political opportunity structure as 'political situations in which states become vulnerable to collective action and in which ordinary people amass the resources to overcome their disorganization and learn where and how to use their resources.' This definition clearly incorporates an important element of resource mobilization theory, namely the significance for movement success of masses of people obtaining and utilizing resources such as money, media attention, and organizational connections. But it also puts these resources in the special political context of governments being particularly susceptible to mass political pressure.

Regarding several scholars' use of political opportunity theory, such as J. Craig Jenkins and Charles Perrow, Tarrow (1991, pp. 34–6) points to five variables often used by these scholars to identify conditions associated with governments being especially susceptible to the influence of social movements: (a) the degree of openness or closure of the polity, with open polities more favorable to social movements; (b) the stability or instability of political alignments, with less stable alignments more favorable to social movements; (c) the presence or absence of allies or support groups, with more allies and support groups favorable to social movements; (d) divisions or lack of divisions within the elite, with elite divisions favoring social movements; and (e) tolerance or intolerance of protest by the elite, with tolerance of protest favoring social movements. These and other elements of political opportunity theory are woven throughout various complex analyses by Tarrow and others to show how shifting political circumstances significantly affect the fortunes of social movements, often resulting in cycles of protest for the movements.

3.4 Recruitment: Social Network Analysis

One of the more informative later approaches to social movements is social network analysis that helps explain why people with similar ideologies are differentially recruited—that is, why some of them join social movements and others do not. As discussed, Smelser (1963) was the first modern collective behavior analyst to note that single variables such as ideology, by themselves, did not have considerable explanatory power. Snow, Zurcher, and Ekland-Olson (in McAdam and Snow 1997) demonstrated that those who belong to social networks are considerably more likely to act upon their ideology by joining a social movement than those without network affiliations. This goes a long way toward explaining why empirical studies since the 1970s have shown low to moderate, as compared to strong, relations between ideological commitment and movement participation (Wood and Jackson 1982). That is, the relationship between ideology and movement participation is influenced by the extent of social network involvement, so that taking network involvement into account strengthens the relationship between ideology and movement participation.

In fact, this latter social network finding was preceded, and influenced by, the earlier finding that membership in secondary organizations—such as business, trade union, church, or even recreational associations—facilitated recruitment to social movements (Marx and Wood 1975, Wood and Jackson 1982). In contrast, mass society theory, as elaborated by William Kornhauser, had previously argued that the absence of such memberships facilitated recruitment. Other discussions of social network analysis have included topics such as the types of networks needed to facilitate recruitment, the role of conflicting group memberships in recruitment, and the size of networks for recruitment (McAdam and Snow 1997). Thus, with its organizational focus, social network analysis provides valuable insights into issues ranging from how political ideology gets translated into social movement participation to the influence of organizational size on movement recruitment.

3.5 The New Social Psychology

Bert Klandermans and his associates have developed a new social psychological approach to social movements, aptly summarized in his book, *The Social Psychology of Protest* (Klandermans 1997). This new social psychology analyzes recruitment to social movements, as did an earlier social psychological approach, but also includes such topics as framing of social movements, identity politics, new social movements, the transformation of discontent into movement participation, and the relation of social psychology to social movement organizations and interorganizational analysis. Whereas an earlier social psychology of social movements emphasized the influence of single variables such as relative deprivation or ideology on movement participation (Wood and Jackson 1982), or focused on emergent norms influencing such participation (Turner and Killian 1972), the new social psychology sees potentially sympathetic individuals responding to a variety of

social and cultural influences that may, or may not, lead to actual movement participation. Among these influences are the social networks and organizations discussed previously, which are seen to promote participation. The new approach also discusses circumstances, such as a deepening political consciousness, favoring—or inhibiting—sustained participation, as discussed by Molly Andrews in *Lifetimes of Commitment*.

Among the influences more recently discussed is the manner in which a movement is framed, which is its presentation to the public—especially by the media—an approach significantly inspired by Ralph Turner's important article, 'The public perception of protest' (Turner 1969). Is a movement for female reproductive rights portrayed as a movement to enhance the social, economic, and political rights of women, or is it presented as a movement to deprive unborn children the right to live? Is a movement for better hours, wages, and working conditions portrayed as a just cause for exploited working people, or is it an excuse for 'union bosses' to expand their power? The new social psychology goes into detail explaining how movements and their oppositions attempt to frame the movements in ways to encourage or discourage actual participation by those with initial interest in the movements.

One of the more controversial emphases discussed by the new social psychology—and which has its own literature by scholars such as Alberto Melucci—focuses on the role of personal identity in understanding the development of recent movements and why certain people are drawn to these movements. In an informative book entitled *Ideology and the New Social Movements*, Alan Scott (1990) indicates why social movements such as the student movement of the 1960s, the modern women's movement, and ethnic movements such as the black movement in the USA should be called new social movements. He argues that these movements are new because, first, they are not the traditional labor movement that arose from conflicts and contradictions between groups of labor and capital involved in the means of economic production, and second, because they are instead based on—and arise from—issues of personal identity such as one's gender, status as a student, and racial or ethnic background. This type of position draws loosely on Orrin Klapp's imaginative, though less empirically oriented book, *Collective Search for Identity*.

Whereas Scott argues that these considerations of identity are paramount in forming the new social movements—and figure in any success of the movements—others such as Todd Gitlin (1995) have argued that excessive involvement with personal identity issues such as these can detract from more general or universal themes such as the fight for decent conditions of work, fair allocation of income in a society, and improvement of lifestyles for all instead of just for particularistic group interests. Still other arguments have focused on the fact that a key basis for organizing women and ethnic minorities, especially, has been the need to improve their status economically, occupationally, educationally, legally, and in terms of lifestyles and life opportunities. Responding to higher education budget cuts in the 1990s, students from varying backgrounds rallied together, along with faculty from varying backgrounds, to fight these cuts and preserve educational and career opportunities (Wood 1998). Thus, identity politics may well be fairly important in organizing given modern social movements, but the continuing issues of economics, education, legal rights, lifestyles, and life opportunities remain as important to the new movements as to the labor movement.

3.6 Historical–Comparative Analyses of Revolutions

There is a body of highly scholarly, detailed historical and comparative analyses of revolutions. Barrington Moore, George Rudé, E. P. Thompson, Charles Tilly, Theda Skocpol, Eric Hobsbawm, and Jack Goldstone are among the major contributors to the understanding of such classical social and political upheavals as the French Revolution, Russian Revolution, English Revolution, and Chinese Revolution. These authors combine analyses of conditions common to several revolutions, such as the weakening of state authority, with analyses of conditions unique to some upheavals but not others, such as differing cultural ideology.

Skocpol (1979), for example, analyzes social revolutions that entail not only political change at the top of society, but also large-scale social–economic change throughout the society. She argues that weakening of the state and military, peasant rebellions, and alienated marginal elites combined to bring about social revolutions in France, Russia, and China. Goldstone (1991) also looks at similar conditions of revolutions, as well as demographic conditions, but adds in an ideological dimension of why social revolutions were more likely to occur in the West with its Judeo-Christian ideology that emphasized large-scale transformations of society. Thus, for Goldstone, social revolutions instead of palace coups or military conquests, for example, were seen as the secular application of Western ideology as compared with non-Western ideologies such as Hinduism or Buddhism that were associated with different political results. This body of literature is complex, but highly informative for deeper understandings of the processes whereby revolutions develop.

3.7 Analyses of Other Forms of Collective Behavior: Riots, Panics, Crazes, and Mass Delusions

The sociology of collective behavior has been interested in longer-term, more organized social move-

ments, as well as shorter-term, less organized, and less political forms of noninstitutionalized group behavior. A. C. Kerckhoff and K. W. Back's interesting study of mass delusion, *June Bug*, analyzed perceptions in an American city that its inhabitants were being bitten each summer by a June Bug. The two researchers were in this city, determined that no such insect existed, and then utilized Smelser's (1963) theory of collective behavior to understand the development of this unusual craze.

In the late twentieth century, various cults have been the focus of collective behavior analysis, with discussions including Jim Jones' People's Temple cult that committed mass suicide, the Heavens Gate cult that acted similarly while anticipating transportation to another planet, and the San Diego-based cult focusing on a goddess, with a public access television show, who also promised her followers transportation to an extraterrestrial land (Tumminia and Kirkpatrick 1995). These studies have focused on the circumstances—social and personal—that would lead people to join such cults, as well as circumstances that keep them involved, especially in terms of members' susceptibility to suggestions and new influences due to their isolation from familiar surroundings and social networks.

Riots are a type of crowd behavior studied since Le Bon's *The Crowd*. For many years riots—that is, violent crowd activities—were seen as irrational, spontaneous, and unorganized group activities. This picture has changed over the years, with riots now more often seen as outcomes of oppression and deprivation, having some rational connection to anticipated outcomes, and having at least some organization (McPhail 1989). Indeed, Gary Marx felt compelled to add a correction to the increasingly ideological characterization of riots by discussing 'issueless riots,' such as crowd violence after athletic victories. Nonetheless riots, along with social movements, are typically seen as responses to difficulties that groups experience that are at least understandable, or even as rationally calculated responses to these difficulties (Tarrow 1991, 1998).

4. Conclusion

The sociology of collective behavior is one of the most dynamic and innovative fields of study in sociology. Indeed, Alain Touraine has stated that the study of social movements is sociology. Collective behavior has always had the ability to inspire and appall. Those sympathetic to the French Revolution remain in awe of the social class changes brought about by it, whereas those in opposition remain dismayed by the violence connected with it. Even a harsh critic of collective behavior like Le Bon felt that progress would not occur without the crowds he nonetheless disliked. As such, collective behavior will continue inspiring and appalling large numbers of people since it persists in dealing with the most electrifying of human dramas.

See also: Action Theory: Psychological; Action, Collective; Crowds in History; Deprivation: Relative; Identity: Social; Ideology: Political Aspects; Ideology, Sociology of; Network Analysis; Networks: Social; Panic, Sociology of; Revolutions, Sociology of; Revolutions, Theories of; Social Movements: Resource Mobilization Theory; Social Movements, Sociology of

Bibliography

Gitlin T 1995 *The Twilight of Common Dreams: Why America is Wracked by Culture Wars*. 1st edn. Metropolitan Books, New York

Goldstone J A 1991 Ideology, cultural frameworks, and the process of revolution. *Theory and Society* **20**: 405–53

Klandermans B 1997 *The Social Psychology of Protest*. Blackwell, Cambridge, MA

Marx G T, Wood J L 1975 Strands of theory and research in collective behavior. *Annual Review of Sociology* **1**: 363–428

McAdam D 1982 *Political Process and the Development of Black Insurgency 1930–1970*. University of Chicago Press, Chicago

McAdam D, Snow D A 1997 *Social Movements: Readings on Their Emergence, Mobilization, and Dynamics*. Roxbury, Los Angeles

McPhail C 1989 Blumer's theory of collective behavior: the development of a non-symbolic interaction explanation. *Sociological Quarterly* **30**: 401–23

Scott A 1990 *Ideology and the New Social Movements*. Unwin Hyman, London

Skocpol T 1979 *States and Social Revolutions*. Cambridge University Press, Cambridge, UK

Smelser N J 1963 *Theory of Collective Behavior*. Free Press of Glencoe, New York

Tarrow S G 1991 *Struggle, Politics, and Reform: Collective Action, Social Movements, and Cycles of Protest*. Cornell University/Western Societies Program, Occasional Paper No. 21, 2nd edn. Cornell University Press, Ithaca, NY

Tarrow S G 1998 *Power in Movement*, 2nd edn. Cambridge University Press, New York

Tumminia D, Kirkpatrick R G 1995 Unarius: emergent aspects of an American flying saucer group. In: Lewis J R (ed.) *The Gods Have Landed: New Religions from Other Worlds*. State University of New York, Albany, NY, Chap. 4 pp. 85–104

Turner R 1969 The public perception of protest. *American Sociological Review* **34**: 815–31

Turner R H, Killian L M 1972 *Collective Behavior*, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ

Wood J L 1998 With a little help from our friends: student activism and the crisis at San Diego State University. In: DeGroot G J (ed.) *Student Protest: The Sixties and After*. Addison Wesley Longman, New York, Chap. 19 pp. 264–79

Wood J L, Jackson M 1982 *Social Movements: Development, Participation, and Dynamics*. Wadsworth, Belmont, CA

J. L. Wood

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Collective Beliefs: Sociological Explanation

The sociological tradition considers the explanation of collective beliefs as an essential task. In his famous *Elementary forms of religious life* (1912, 1995), Durkheim presents this explanatory project as the center of gravity of the sociological discipline. Although of decisive importance, this project remains confronted with certain difficulties. First, the extreme diversity of the collective beliefs. The members of modern and traditional collectivities develop many and varied forms of thought: ideological, religious, magic, scientific, etc. Impressed by this diversity, sociologists have sometimes multiplied the explanatory categories and by doing so have forgotten that the vocation of any explanatory theory is to establish the simplicity behind the apparent complexity.

Another difficulty is the permanent and undoubtedly irreducible *tension* on which the sociological explanation of collective beliefs rests. The sociologist must maintain a sufficient proximity to the observed beliefs as to be able to identify their specific qualities, while remaining sufficiently distant to preserve the autonomy of the sociological interrogation. The history of the various ideological drifts of the sociological analysis shows that this balance is not always easily accessible.

A short detour by sociological tradition will initially remind us of the diversity of the sociological explanations of the collective beliefs. On a second occasion, we will identify some of the general criteria used by sociologists to evaluate the respective merits of these various explanations. Finally, we will be interested in the reasons why many contemporary sociologists consider the generalization of the principle of 'symmetry' as significant progress for the sociological explanation of collective beliefs.

1. *The Diversity of Sociological Explanations*

In a classical article intended to provide a general theoretical framework for the sociology of knowledge, R. K. Merton (1945) summarizes the nature of the sociological explanation in the following way: to establish the correlations between collective beliefs, conceived as dependent variables, and 'the other existential factors of society and culture' conceived as independent variables. If such a formula makes it possible to unify many sociological studies devoted to the explanation of collective beliefs, this unity remains however purely 'formal.' The diversity of the sociological explanatory modes must be considered on at least two levels: on the one hand the nature of these 'existential factors' mentioned by Merton, on the other the nature of the 'relation' which links these factors to the collective beliefs.

Some examples may be useful. Ideology constitutes a traditional object of the sociology of collective beliefs. These sociopolitical belief systems play a central role in the legitimization of the social order of modern societies. The sociological explanation of ideologies generally consists of replacing them in the singular interaction system in which they appear. There is not, however, a single way to conceive the nature of this system. The Marxist tradition, for example, identifies this system with a complex set of 'social interests.' This explanatory scheme initially suggested by Marx (1852) is primarily of a 'utilitarist' nature. If a social group believes in the value of such or such a politico-social organization, it is not because of its intrinsic value, but because this precise type of organization reinforces directly or indirectly its social domination, and consequently its material interests. The social utility of the ideological belief overrides its truth or falseness.

If this utilitarist approach of collective beliefs has been presented by K. Mannheim (1929, 1991) as a fundamental stage for the emergence of the sociology of knowledge, it is not the only possible one. The paretian study of religious beliefs, for example, rests on a very different conception of the existential factors mentioned by Merton. In his famous *Traité de sociologie générale*, Pareto (1916, 1968) analyzes the social diffusion of the religions as the consequence of social 'feelings.' He affirms the importance to clearly separate the respective influence of 'derivations'—ideas, theories, theologies, etc.—and 'residues'—deep-rooted feelings: 'the social value of the religions,' Pareto writes, 'depends very little on derivations, enormously on the residues. In several religions, there is a significant group of residues (…) which correspond to feelings of discipline, submission, hierarchy' (§1854). In a way, this explanatory scheme developed by Pareto is very different from the one proposed by Marx for the ideologies: if a social group massively adopts a religion, it is not because of its direct or indirect social 'utility' but because it somehow manages to satisfy the dominant 'passions' of this group. In another way, however, these two explanatory schemes are quite similar. Marx for the ideologies and Pareto for the religions both explicitly consider that the link between the dependent and independent variables has to be conceived as a 'causal' relation. The adoption of a belief by a social group appears in both cases as the mechanical consequence of 'forces'—interest or passions—which dominate the conscience of its members. These forces remain out of the control of the social actors.

This causal approach has sometimes been used to explain the existence of magic beliefs. Lévy-Bruhl (1922, 1960) sees the persistence of magic beliefs in traditional societies as the mechanical consequence of a specific mental structure: the 'primitive mentality.' This mentality, Lévy-Bruhl suggests, prevents the members of these communities from perceiving the

objective difference between verbal similarity and real similarity and, in a more general way, the difference between the relations between the words and the relations between the things. However, works of Durkheim and Weber show that this approach of magic beliefs is far from being the most fruitful. In their respective analyses of the magic beliefs, they both identify the 'existential' factor to the immediate environment of the social actors. They both also conceptualize the relation between this environment and the collective belief as 'rational' or more precisely as *subjectively rational*. 'The acts motivated by magic,' Weber writes (1922, 1979), 'are acts at least relatively rational (…): they follow the rules of the experience even if they are not necessarily acts according to the means and ends.' The same point is emphasized by Durkheim (1912, 1995), in particular, when he compares the rationality of traditional 'rites' and the rationality of modern 'techniques': 'the rites which [the primitive] use to ensure the fertility of the ground (…) are not, to him, more irrational than are, to us, the technical processes used by our agronomists (…). The powers linked to these rites do not appear particularly mysterious. For those who believe in them, those forces are not more unintelligible than are gravity or electricity to a contemporary physicist.' The explanatory strategy proposed by Weber and Durkheim thus exists mainly to identify the role of the collective belief in the adaptation process of the social actor to his immediate environment, and thus to reconstruct the 'meaning' of the belief for this actor.

The sociologists have also paid great attention to scientific beliefs. Sorokin (1937), for example, attempts to demonstrate 'that what a given society regards as true or false, scientific or unscientific (…) is fundamentally conditioned by the nature of the dominant culture.' He analyzes the relation between the social 'credibility' of scientific representations of reality and the evolution of cultural values. The Sorokinian 'existential' factor is different from the factors previously observed in Marx, Pareto, Weber, or Durkheim's works: it consists mainly of a 'cultural framework' evolving in a cyclic way. The 'relation' between dependent and independent variables remains however similar to the relation theorized by Durkheim and Weber. The social groups, Sorokin says, select their scientific beliefs according to a general principle of 'logical dependence' or 'logical consistency.' Merton (1938, 1970) shares with Sorokin the will to establish the social conditions of the scientific development. However, he emphasizes, the 'functional' nature of the relation which linked the natural sciences of the seventeenth century to their socio-historical contexts. Scientific knowledge, Merton observes, is obviously developed on the basis of cognitive constraints but this development also integrates, in variable proportions, the influence of social factors. The statistical analysis applied to technological inventions shows in particular that a great number of these inventions are intended to bring a solution to maritime transport, mining industry, or military technology problems. The production of the scientific beliefs may be partially interpreted as an attempt of the scientific community to satisfy an explicit or diffuse social demand.

2. *Three Elementary Conditions of Sociological Explanations*

As shown by these classical examples, sociologists refer in the course of their studies to multiple existential factors: interest (Marx), passion (Pareto), structure of interaction (Durkheim, Weber), cultural framework (Sorokin), socioeconomic development (Merton). They diverge moreover in their way of representing the nature of the relation between these factors and the studied collective beliefs. Three general approaches of this relation are regularly observable: causal (Marx, Pareto), functional (Merton), and rational (Durkheim, Weber, Sorokin). It also happens sometimes that sociologists try to evaluate the objective range of their own explanations. The question they then have to face can be summarized in the following way: which conditions do sociologist need to respect in order to produce a solid explanation of collective beliefs? Three elementary conditions deserve to be briefly underlined.

First condition: the sociological explanation of collective belief must be more than the translation of the phenomenon to explain a new idiom

The sociologists of beliefs generally start from the identification of an enigmatic phenomenon. Why does the industrial bourgeoisie of the nineteenth century believe in the virtues of parliamentary monarchy (Marx)? Why do Zunis Indians adopt a representation of the world in seven categories (Durkheim)? Why are corpuscular and ondulatory theories of light widely accepted during the nineteenth century (Sorokin)? etc. But it sometimes happens that their answers are only apparent. The methodological analysis of the theory of 'primitive mentality' shows thus that if the theory of Lévy-Bruhl does not convince anymore it is essentially because it does not explain anything. The key concepts of this theory, Boudon underlines (1990, 1994), 'only paraphrase the beliefs. They explain the confusion of the primitives between verbal associations and causal relations by the primitive's tendency to make this confusion, tendency whose reality is guaranteed by the fact that they do make it indeed.' A similar point is made by Merton (1945) concerning Sorokin's theory. 'It seems tautological,' Merton writes, 'to say, as Sorokin does, that in a sensualist society and culture, the sensualist system of the truth based on the testimony of the sensory organs must prevail. Because he had already defined the sensualist mentality as one

which conceives reality as being only what is perceived with the sensory organs.' More recently, S. Cole (1996) pointed out that the analysis of the production of scientific beliefs by certain constructivist sociologists of sciences is primarily of circular nature. These sociologists generally claim 'that it is impossible to separate the technical from the social; that all science is inherently social. This turns their entire argument into a tautology. If science is inherently social, this means that the technical aspects of scientific discoveries by necessity must be determined socially. This is indeed the question we are examining.'

Second condition: the sociological explanation of collective beliefs must avoid ambiguous and/or occult concepts

Sociologists use a great number of conceptual categories. Many proposals produced on the basis of the combination of these categories make clear and univocal sense, some however remain more ambiguous. For example, certain sociological explanations rest on an ambiguous concept of 'social,' or more precisely a confusion between the 'social' and 'historical' dimensions of a collective belief. To show, however, that a belief appears in a particular historical context is not the same as to demonstrate its social character. L. Laudan (1977) is right to observe that 'the fact that certain assertions are contextual, that one adhered to it only to a certain time and in a certain place establishes neither that they are necessarily of social origin, nor even that they are opened to the sociological analysis.' The sociological use of the concept of 'interest' constitutes another well-known example of ambiguity. This concept is sometimes used by sociologists of science in order to describe the origin of the prestige or credibility of scientific beliefs. In an underdetermined situation, contemporary sociologists of science say, the scientists choose between two concurrent theories the one which appears conform to their 'interests.' However, how do we have to conceive the exact nature of these interests? Are they material interests, symbolic interests, technical, or cognitive interests, etc.? As long as the concept of interest remains ambiguous, the sociologist cannot legitimately claim to explain the observed beliefs. In addition, following a Weberian principle, many sociologists closely associate the solidity of an explanation to the manner in which these explanations avoid using occult categories, i.e. invisible 'forces'—as the Paretian concept of 'residue' or the 'primitive mentality' of Lévy-Bruhl—which determines the behaviors of the social actors. The methodological principle is simple and easily acceptable: the sociologist seeking to account for the origin of a collective belief may refer to the existence of an occult cause only insofar as he cannot explain this belief more easily, in particular by identifying his meaning. The obvious difference between the explanations of magic beliefs proposed on the one hand by Lévy-Bruhl and on the other by Durkheim and Weber seems to be directly linked to the application of this general principle by the latter.

Third condition: the sociological explanation of collective belief must be compatible with observable facts

This third condition can be interpreted in various ways. The first and simplest one consists of assuming that explanatory theories of scientific beliefs are not different from other scientific theories. They must not only dispose of various internal qualities (absence of circularity, precision, etc.) but also be able to be confronted with precise facts, and further still to resist this empirical test. These facts might be elementary observations. The observation, for example, of the persistence of the magic beliefs in Western societies seems hardly compatible with the assumption of a causal link between the magic beliefs and a 'primitive mentality' (Lévy-Bruhl). In the same way the observation of the developments of new religious movements in large cities enters openly in contradiction with the hypothesis that the contemporary process of urbanization is essentially a vector of disintegration of collective belief (Borhek and Curtis 1975). These facts can also be more complex products of the interpretative reconstruction work done by the sociologist or anthropologist. When R. Horton (1993) argues for example against the tenants of symbolist theory of magic, he points out that this theory, for which magic beliefs are mainly the expression of a symbolic 'desire,' enters into flagrant contradiction with what the members of the traditional societies say about their own rituals: '(…) in denying the paramount importance of explanation, prediction and control as guiding aims of traditional African religious life, ''symbolists'' (…) are committing the cardinal interpretative sin of flouting the actor's point of view.' Another way to interpret this third condition consists of associating the solidity of an explanatory theory with its capacity to give an account of other beliefs than the one for which it was initially elaborated. Is it for example possible to apply the explanatory scheme imagined by Durkheim and Weber for magic beliefs to other types of beliefs, such as the ideological ones? The fact that this transposition is not only possible but brings an objective gain of intelligibility for the comprehension of the ideological phenomenon (Boudon 1986, 1989) demonstrates the real value of their general conception of the sociological explanation.

3. *The Symmetry Principle*

Sociologists frequently add to these three general conditions a final one: a condition known as the 'principle of symmetry.' This principle implies that the explanation of the collective beliefs should be equally applied to phenomena of various natures. This last condition is not reducible to the conditions previously mentioned (the third one in particular) insofar as, to identify the nature of this plurality, it introduces a

criterion of truth and falseness, rationality and irrationality. A sociological theory of collective beliefs is said to be symmetrical if it explains *in the same terms* true beliefs and false beliefs, rational beliefs and irrational beliefs. This principle, implicitly used by Sorokin (1937), is directly opposed to the 'a-rationality principle' according to which the sociological explanation of the collective beliefs cannot be indifferent to the content of these beliefs.

The interest of the principle of symmetry is double. From a methodological point of view, it requires the suspension of all evaluation related to the potential validity or invalidity of the analyzed collective belief. It thus represents a guarantee of neutrality. From the point of view of the empirical research, it makes it possible to follow in unified and thus simplified terms the manner in which the collective beliefs obtain a social recognition, or inversely an absence of social recognition.

The sociological use of this principle followed, however, two paths of unequal importance. The first, chosen by some sociologists of science, exists mainly to explain the scientific beliefs in terms of explanatory categories previously devoted to false beliefs. Linked to an exclusively causal approach of scientific belief production, this first use of the principle of symmetry has greatly contributed to the diffusion of a relativistic representation of science (Dubois 2001). There is, however, another way to apply this principle. It consists of following the major intuition of the theory proposed by Weber and Durkheim: to seek behind all collective beliefs, true or false, socially validated or socially marginal, the influence of 'reasons,' objective and/or subjective. Thus as R. Horton (1993) is reconstructing the limited rationality of certain African rituals, he defines the major nature of its explanatory project as follows: elaborating 'an explanatory framework which would deal "symmetrically" with the traditional and the modern, the prescientific and the scientific.'

See also: Action, Theories of Social; Belief, Anthropology of; Cognitive Dissonance; Collective Memory, Anthropology of; Culture, Sociology of; Durkheim, Emile (1858–1917); Ideology, Sociology of; Knowledge, Sociology of; Mannheim, Karl (1893–1947); Rational Choice Theory in Sociology; Social Movements, Sociology of; Theory: Sociological; Tocqueville, Alexis de (1805–59); Weber, Max (1864–1920)

Bibliography

Borhek J T, Curtis R F 1975 *A Sociology of Belief*. Wiley, New York

Boudon R 1986/1989 *The Analysis of Ideology*. University of Chicago Press, Chicago

Boudon R 1990/1994 *The Art of Self-Persuasion: The Social Explanation of False Beliefs*. Polity Press, London

Cole S 1996 Voodoo sociology: recent developments in the sociology of science. *Annals of the New York Academy of Sciences* **775**: 274–87

Dubois M 2001 *La nouvelle sociologie des sciences*. Presses Universitaires de France, Paris

Durkheim E 1912/1995 *The Elementary Forms of Religious Life*. Free Press, New York

Horton R 1993 *Patterns of Thought in African and the West*. Cambridge University Press, Cambridge, UK

Laudan L 1977 *Progress and its Problems*. University of California Press, Berkeley, CA

Lévy-Bruhl L 1922/1960 *La mentalité primitive*. Presses Universitaires de France, Paris

Mannheim K 1929/1991 *Ideology and Utopia*. Routledge & Kegan Paul, London

Marx K 1852 The eighteenth brumaire of Louis Bonaparte. In: Marx K, Engels F (eds.) *Collected Works*. Lawrence & Wishart, London, Vol. 11

Merton R K 1938/1970 *Science, Technology and Society in Seventeenth Century England*. Fertig, New York

Merton R K 1945/1971 Sociology of knowledge. In: Gurritch G, Moore W E (eds.) *Twentieth Century Sociology*. Books for Libraries Press, Freeport, New York

Pareto V 1916/1968 *Traité de sociologie générale*. Droz, Genève, Switzerland

Sorokin P 1937 *Social and Cultural Dynamics, Fluctuation of Systems of Truth, Ethics, and Law*. American Book Company, New York, Vol. 2

Weber M 1922/1979 *Economy and Society*. University of California Press, Berkeley, CA

M. Dubois

Collective Identity and Expressive Forms

If key concepts or expressions can be identified that function to capture the animating spirit of different epochs, then certainly one candidate concept for the latter quarter of the twentieth century is the concept of collective identity. Indeed, it is a concept that came of age in the latter part of the past century, as reflected in the outpouring of scholarly work invoking the concept directly or referring to it indirectly through the linkage of various collectivities and their identity interests via such concepts as identity politics, identity projects, contested identities, insurgent identities, and identity movements. This article provides an analytic overview of scholarly work on the concept by considering, in order, its conceptualization, its various empirical manifestations, the analytic approaches informing its discussion and analysis, and several unresolved theoretical and empirical issues.

1. Conceptualization

The concept of collective identity, just as the base concept of identity, is rooted in the observation that interaction between two or more sets of actors mini-

mally requires that they be situated or placed as social objects. To do so is to announce or impute identities. Hence, interaction among individuals and groups, as social objects, is contingent on the reciprocal attribution and avowal of identities. This character of identity is highlighted in Stone's (1962) conceptualization of identity as the 'coincidence of placements and announcements.' This process holds for both individuals and collectivities, and it probably has always been a characteristic feature of human interaction, whether the interaction was among early preliterate humans or among those in the modern social world. To note this is not to ignore the sociological truism that the issue of identity becomes more problematic and unsettled as societies become more structurally differentiated, fragmented, and culturally pluralistic (Castells 1997, Giddens 1991). But historical variation in the extent to which matters of identity are problematic does not undermine the double-edged observation that the reciprocal imputation and avowal of identities is a necessary condition for social interaction and that identities are thus rooted in the requisite conditions for social interaction.

Delineating the interactional roots of identities does not explain what is distinctive about collective identity, as there are at least three conceptually distinct types of identity: personal, social, and collective. Although they often overlap, one cannot be inferred from the other. Hence the necessity of distinguishing among them.

Social identities are the identities attributed or imputed to others in an attempt to situate them in social space. They are grounded typically in established social roles, such as 'teacher' and 'mother,' or in broader and more inclusive social categories, such as gender categories or ethnic and national categories, and thus are often referred to as 'role identities' (Stryker 1980) and 'categorical identities' (Calhoun 1997). Whatever their specific sociocultural base, social identities are fundamental to social interaction in that they provide points of orientation to 'alter' or 'other' as a social object.

Personal identities are the attributes and meanings attributed to oneself by the actor; they are self-designations and self-attributions regarded as personally distinctive. They are especially likely to be asserted during the course of interaction when other-imputed social identities are regarded as contradictory, as when individuals are cast into social roles or categories that are insulting and demeaning (Snow and Anderson 1987). Thus, personal identities may derive from role incumbency or category-based memberships, but they are not necessarily comparable since the relative salience of social roles or category membership with respect to personal identity can be quite variable.

Just as social and personal identities are different yet typically overlapping and interacting constructs, such is the relationship between collective and social and personal identities. Although there is no consensual definition of collective identity, discussions of the concept invariably suggest that its essence resides in a shared sense of 'one-ness' or 'we-ness' anchored in real or imagined shared attributes and experiences among those who comprise the collectivity and in relation or contrast to one or more actual or imagined sets of 'others.' Embedded within the shared sense of 'we' is a corresponding sense of 'collective agency.'

This latter sense, which is the action component of collective identity, not only suggests the possibility of collective action in pursuit of common interests but even invites such action. Thus, it can be argued that collective identity is constituted by a shared and interactive sense of 'we-ness' and 'collective agency.' This double-edged sense can be culled from classic sociological constructs such as Durkheim's 'collective conscience' and Marx's 'class consciousness,' but is reflected even more clearly in most conceptual discussions of collective identity, although the agentic dimension is sometimes implied rather than directly articulated (e.g., Castells 1997, Cerulo 1997, Eisenstadt and Giesen 1995, Jasper and Polletta 2001, Jensen 1995, Levitas 1995, Melucci 1989, 1995).

A common theme running throughout a segment of the literature is the insistence that collective identity is, at its core, a process rather than a property of social actors. Such work acknowledges that collective identity is 'an interactive and shared definition' that is evocative of 'a sense of we,' but then highlights the process through which social actors recognize themselves as a collectivity, contending that this process is more vital to conceptualizing collective identity than any resultant product or property (e.g., Melucci 1989, pp. 34, 218, passim). Few scholars would take exception with the importance of the process through which collective identities develop, but it is both questionable and unnecessary to contend that the process is more fundamental than the product to understanding the character and functionality of collective identity. Not only is the product or 'shared we' generative of a sense of agency that can be a powerful impetus to collective action, but it functions, as well, as the orientational identity for other actors within the field of action. More concretely, it is the constructed social object to which the movement's protagonists, adversaries, and audience(s) respond (Hunt et al. 1994), and which, in turn, may have implications for the operation of its organizational carrier, affecting the availability and character of allies, resources, and even tactical possibilities (Jensen 1995). The initial projected collective identity may be short-lived and transient, subject to modification and even transformation during the course of ongoing collective (inter)action, but the set of properties that makes up the initial collective identity, as well as whatever subsequent ones emerge, constitute objects of orientation and interaction for other collectivities within the field of action.

If it is acknowledged that there is something of substance to collective identities, how are they distinguished from social and personal identities? Several factors appear to be at work. First, collective identities may or may not be embedded in existing social identities, since they are often emergent and evolving rather than firmly rooted in prior social categories. This is often the case with the collective identities that emerge in the course of dynamic social protest events (for illuminating examples, see Walder's research on the Beijing Red Guard Movement (2000), and Calhoun's account of the Chinese student movement of 1989).

Second, the collective, shared 'sense of we' is animating and mobilizing cognitively, emotionally, and sometimes even morally. The shared perceptions and feelings of a common cause, threat, or fate that constitute the shared 'sense of we' motivate people to act together in the name of, or for the sake of, the interests of the collectivity, thus generating the previously mentioned sense of collective agency. That potential inheres within social identities, but they typically function more like orientational markers as the routines of everyday life are negotiated. When they are activated or infused affectively and morally, it is arguable that they have been transformed into collective identities. Third, the emergence and operation of collective identities means that other social identities have subsided in relevance and salience for the time being. In other words, collective identities, when they are operative, generally have claims over—not so much normatively as cognitively and emotionally—other identities in terms of the object of orientation and character of corresponding action. Examples abound, as observed frequently in the case of many protest gatherings, gripping fads, joyous and celebratory sports crowds, and the concerted campaigns and actions associated with social movement activism. Fourth, while collective identities and personal identities are obviously different, they are still very much interconnected in the sense that collective identities are predicated, in part, on constituents' embracement of the relevant collective identity as a highly salient part of their personal identity and sense of self (Gamson 1991). Finally, while the attribution or avowal of all identities is interactionally contingent, collective identities tend to be more fluid, tentative, and transient than either categorically based social identities or even personal identities.

2. *Empirical Manifestations*

Empirical manifestations of collective identity can vary in a number of significant ways. One important axis of variation is the size of the collectivity and the corresponding scope of its claims. If the essence of collective identity resides in a sense of 'we-ness' associated with real or imagined attributes in contrast to some set of others, then it follows that collective identities can surface among almost any grouping or aggregation in a variety of contexts, ranging from relatively small cliques and gangs to sports fans and celebrity devotees, to laborers and occupational groupings, to neighborhoods and communities, to even broader categories such as sexual and gender categories, religions, ethnic groups, and nations.

The preponderance of empirical research on collective identity has focused on the last, more inclusive set of categories—sexuality and gender, religion, ethnicity, and nationality. Illustrative is Taylor and Whittier's (1992) research on lesbian identity and lesbian social movements; Nagel's (1996) work on the resurgence of collective identity among American Indians; Cornell and Hartmann's (1998) overview of the construction of ethnic and racial identities in the modern world; and Anderson's (1991) and Calhoun's work on nationalism, which the latter defines, in part, as one 'way of constructing collective identities' (1997, p. 29). An additional characteristic of research on collective identity is its association with the study of social movements, no doubt because such mobilizations tend to be both generative of and dependent on collective identities (e.g., Gamson 1991, Hunt et al. 1994, Jasper and Polletta 2001, Melucci 1989, Snow and McAdam 2000).

Although collective identities can congeal in various aggregations and contexts, they appear not to do so on a continuous basis historically. Instead, their emergence and vitality appear to be associated with conditions of sociocultural change or challenge, socioeconomic and political exclusion, and political breakdown and renewal, thus suggesting that they cluster historically in social space. The latter part of the twentieth century has generally been regarded as one such period of collective identity effervescence and clustering, with some scholars characterizing this period in terms of identity crises and collective searches for identity (e.g., Castells 1997, Gergen 1991, Giddens 1991, Klapp 1969).

In *The Power of Identity*, Castells captures both this characterization and the kinds of conditions thought to be associated with the various manifestations of collective identity during this period:

> Along with the technological revolution, the transformation of capitalism, and the demise of statism, we have experienced, in the last quarter of the century, the widespread surge of powerful expressions of collective identity that challenge globalization and cosmopolitanism on behalf of cultural singularity and people's control over their lives and environment (Castells 1997, p. 2).

3. *Analytic Approaches*

To note that expressions of collective identities cluster historically according to the conjunction of various social conditions does not specify the character or

content of the emergent collective identities. This issue has been addressed from the vantage point of three contrasting perspectives: primordialism, social structuralism, and social constructionism.

Both the primordialist and structuralist views can be construed as variants of an overarching essentialist perspective which posits that a collectivity's identity basically flows naturally from some underlying set of characteristics, often reduced to a single determinative attribute regarded as the collectivity's 'defining essence.' From the primordialist point of view, the defining characteristic is typically an ascriptive attribute, such as race, gender, or sexual orientation, or sometimes a deep, underlying psychological or personality disposition. From a structuralist perspective, the critical characteristic is typically a kind of master social category implying structural commonality, such as social class, ethnicity, or nationality; a set of relational ties or networks suggesting structural connectedness; or a mixture of both. Individuals who are similarly situated structurally, such that they are incumbents of similar roles, work in similar enterprises, are linked to the same social networks, or members of the same social class, religion, or ethnic group, are presumed to have a shared collective identity or at least be candidates for such.

The constructionist perspective, in general, rejects both the primodialist and structuralist variants of the essentialist argument, seeing the presumed link between identities and their ascriptive or structural moorings as being more indeterminate than postulated. Instead, attention is shifted to the construction and maintenance of collective identities. Collective identities are seen as invented, created, reconstituted, or cobbled together rather than being biologically preordained or structurally or culturally determined.

Assessment of the three perspectives in terms of their relative analytic utility for explaining the character and content of collective identities reveals considerable support for the constructionist thesis. This may be due in part to the currents of fashion, influenced by the winds of multiculturalism, postmodernism, and identity politics, but it is due, more importantly, to other factors. One is that the hypothesized link between identities and the primordial attributes or structural categories in which they are presumably anchored is too mechanistic. In its hard version, it is contradicted by the sociological observation that people are often members of the same categories or groups in different ways and with varying degrees of commitment and identification, thus suggesting that inferring correspondence between personal, social, and collective identities solely on the basis of primordial or structural categories is empirically suspect. Additionally, the claims of primordialist and structuralist arguments do not fare well when confronted with the observation that people generally have multiple identities (e.g., family, work, leisure, gender, ethnic, religious, political, and national) that are differentially invoked or avowed depending on their relative salience and their situational pervasiveness.

Salience refers to the relative importance of an identity in relation to other identities (Stryker 1980); pervasiveness or comprehensiveness refers to the situational relevance or reach of any particular identity and the corresponding degree to which it organizes social life, including collective action (Cornell and Hartmann 1998, Snow and McAdam 2000). Given the fact that increasing numbers of individuals live in a world in which they are the carriers of multiple and often conflicting identities, what determines any particular identity's relative salience and pervasiveness, and thus the influence of its claims, vis-à-vis others? Clearly such matters are not determined solely by an identity's primordial roots or structural footing. Finally, much of the empirical evidence is consistent with the constructionist argument. Two highly evocative examples include Trevor-Roper's (1983) account of the retrospective invention of the distinctive Highland culture and tradition so redolently associated with all of Scotland, and James' parallel conclusion, based on extensive archeological and archival research, regarding the origins of the modern Celts:

> …the idea of a race, nation or ethnic group called Celts in Ancient Britain and Ireland is indeed a modern invention. It is an eighteenth- and nineteenth-century reification of a people that never existed, a factoid…assembled from fragments of evidence drawn from a wide range of societies across space and time, This reification served the interests of a range of cultural expectations, aspirations, and political agendas—and still does (James 1999, p. 136).

Such conclusions should not be read as unequivocal refutations of the primordialist and structuralist arguments, since constructed identities are not fabricated whole cloth but typically knit together by drawing on threads of past and current cultural materials and traditions, structural arrangements, and even primordial attributes. These materials and attributes constitute the kinds of stuff from which collective identities—particularly ethnic, religious and national ones—are fashioned, and thereby function, in varying degrees, to constrain the construction process. Interpretative constraint also may be exercised by the institutional contexts and relations of power in which contestants are embedded (Castells 1997, Jensen 1995). Additionally, analyses of the relationship between collective identity and participation in social movements repeatedly point to the experience of collective action itself as a fertile seedbed for the generation of collective identities (Calhoun 1991, Fantasia 1988, Melucci 1989, Walder 2000). Thus, while collective identities are undeniably constructed, they rarely are constructed *carte blanche*; rather, they typically are forged not only with the

materials suggested by the primordialist and structuralist perspectives, but with and through the experience of collective action itself.

4. Theoretical and Empirical Issues

Relevant to a thoroughgoing understanding of collective identity, whatever its empirical locus, are several theoretical and empirical issues that require more careful consideration than often accorded.

4.1 Identity Work (the Expression of Collective Identities)

Fundamental to understanding collective identity, particularly from a constructionist standpoint, are the processes through which it is created, expressed, sustained, and modified. These processes have been conceptualized as variants of 'identity work,' which encompasses the range of activities people engage in, both individually and collectively, to signify and express who they are and what they stand for in relation or contrast to some set of others (Schwalbe and Mason-Schrock 1996, Snow and Anderson 1987, Snow and McAdam 2000). At its core is the generation, invocation, and maintenance of symbolic resources used to bound and distinguish the collectivity both internally and externally by accenting commonalities and differences (Eisenstadt and Giesen 1995, Schwalbe and Mason-Schrock 1996, Taylor and Whittier 1992). Symbolic resources include the interpretive frameworks (or frames), avowed and imputed names, and dramaturgical codes of expression and demeanor (e.g., particularistic styles of storytelling, dress, adornment, and music) that are generated and employed during the course of a collectivity's efforts to distinguish itself from one or more other collectivities. Concrete examples include the various forms of identity talk such as 'atrocity tales' and 'war stories' that group members repeatedly tell each other, prospective adherents, and the media (Hunt and Benford 1994); particular songs and styles of music that invite participation and that are politically and emotionally evocative, such as 'We Shall Overcome' (Eyerman and Jamison 1998); key words and slogans that function in a similar fashion, such as 'Liberte, Fraternite, and Egalite' and 'Workers of the World Unite'; and systems of gestures and signs, such as the raised clinched fist and the peace sign, that function similarly to the tradition of heraldry (Pastoureau 1997).

These and other symbolic resources function as boundary markers of collective differentiation, distinguishing insiders from outsiders, or protagonists from antagonists, in a fashion that heightens awareness of in-group commonalities and connections and out-group differences. Together they congeal into a kind of 'semiotic bricolage' (Schwalbe and Mason-Schrock 1996) that gives symbolic substance to the claimed distinctive 'we,' and it is largely through this *bricolage* that collective identity is expressed and known publicly. While the boundary making and maintenance functions of these symbolic resources, or *bricolage*, is widely acknowledged, what accounts for the differential resonance or carrying power of different symbolic markers is less well understood.

4.2 The Problem of Identity Correspondence

A not uncommon problem with analyses of collective identity is the tendency to reify the collective identity, and thus take for granted the link between the individuals that make up the collectivity and the shared, overarching identity. This gloss is particularly troublesome in light of the observation that people typically have multiple identities that vary in salience and pervasiveness. Thus, how is any particular collective identity reconciled with other identities adherents possess? How, in other words, do the shared cognitions and feelings indicative of a collective identity move center stage at the individual level? Such questions allude to what has been referred to as the problem of 'identity correspondence'—the alignment or linkage of personal and collective identities (Snow and McAdam 2000).

One general answer is that this alignment 'is accomplished by enlarging the personal identities of a constituency to include the relevant collective identities as part of their definition of self' (Gamson 1991, p. 41). But what are the processes through which prospective adherents come to embrace the relevant collective identity, such that personal and collective identity are correspondent or congruous? Two broad processes have been suggested: identity convergence and identity construction.

Identity convergence refers to the union of personal and collective identities when both are congruent, such that an extant collectivity provides a venue for an individual to act in accordance with her or his personal identity. The analytic problem is not one of identity construction or transformation, but one of linkage or bridging and the identification of the mechanisms that facilitate the convergence. A number of such mechanisms have been identified. One operates at the organizational level, entailing the occasional appropriation of existing solidary networks by movement organizations (Snow and McAdam 2000); the other mechanisms are variants of rational choice processes.

One is based on a 'tipping' or 'threshold' model, which posits that collective identities are assumed by individuals when the perceived actions of others reach a point that suggests that the payoffs for adopting, or at least acting in accordance with, the collective identity outweigh doing otherwise. Illustrative is the contention that such tipping points played a critical role in explaining language and identity shifts among

Russian-speaking immigrants 'beached' in four of the republics of the former Soviet Union (Laitin 1998). A related mechanism is the existence of intergenerational investments in personal identities that may have implications for the embracement of future collective identities (Laitin 1998). A third rational choice explanation holds that the collective identities associated with social movements—and, by implication, with other collectivities as well—can be regarded as 'selective incentives' for those who seek to express and affirm their personal identities (Friedman and McAdam 1992). The explanatory utility of these arguments, as well as the network appropriation thesis, is contingent on two underlying assumptions: that there is pre-established congruence among some number of personal identities and the proffered or available collective identities; and that collective identities are constituted by the aggregation or convergence of parallel personal identities. While the first assumption is empirically tenable, the second is questionable from the vantage point of many scholars of collective identity (e.g. Jasper and Polletta 2001, Melucci 1989, 1995).

In the absence of correspondence between personal identities and collective identities, some variety of identity work is necessary in order to facilitate their alignment. This alignment can vary significantly, ranging from the elevation of the salience of a particular identity to a fairly dramatic change in one's sense of self. Four identity construction processes have been identified that capture this variation: identity amplification, identity consolidation, identity extension, and identity transformation (Snow and McAdam 2000).

Identity amplification affects a change in an individual's identity salience hierarchy, such that an existing but lower-order identity becomes sufficiently salient to ensure engagement in collective action, as in the many cases in which in the identity of woman was elevated and expanded in conjunction with the Women's Movement; identity consolidation involves the adoption of an identity that is a blend of two prior but seemingly incompatible identities, as in the case of the union of environmentalists and labor activists and 'Jews for Jesus'; identity extension entails the expansion of the situational pervasiveness of an individual's personal identity so that its reach is congruent with the collectivity's, as when individuals come to see themselves as representatives for a specific cause that transcends other role obligations and identities; and identity transformation involves a dramatic change in identity, such that individuals now see themselves as remarkably different than before, as often occurs in the case of conversion to a new group or movement.

The mechanisms or processes underlying these various forms of identity construction include framing processes in which identities are announced or renounced, embraced or rejected, and modified or reframed in the course of various interactions with adherents, antagonists, and bystander audiences (Benford and Snow 2000, Hunt et al. 1994); engagement in collective action, as when direct observation or experience functions as a demonstration event that gives rise to a situationally specific collective identity or affirms collective claims and thus helps to render salient, and perhaps pervasive, what was previously a secondary or marginal personal or social identity (Melucci 1989, Calhoun 1991, Walder 2000); or some combination of both framing and actual engagement.

Given the variety of ways in which identity correspondence can be affected, the question arises as to whether the relevance of the convergence and construction processes varies by type of collectivity. In the case of social movements, for example, it has been hypothesized that movements that are culturally different, greedy in terms of the cognitive and behavioral demands, and/or politically radical are likely to rely more on identity construction than convergence processes (Snow and McAdam 2000). Whether this is the case is an empirical question, but it does caution against presuming that what accounts for identity correspondence and shifts in one context necessarily holds for another.

Just as there may be variation between types of collectivities and the processes of identity convergence and identity construction, so it is reasonable to ask whether these processes might vary in importance at different points in the life course of a social movement or ethnic or nationalist mobilization? Rather than assuming that a particular process, such as the tipping process or identity amplification, operates routinely with respect to the emergence of a collective identity, might not these processes be more relevant at particular junctures in the career of a movement's collective identity? Preliminary consideration of such questions suggests that network appropriation, rational choice, and constructionist explanations, rather than being mutually exclusive and antithetical, may interact and combine in interesting ways in explaining the emergence, institutionalization, and diffusion of collective identities across different contexts (Snow and McAdam 2000).

4.3 Dimensions, Layers, and Types of Collective Identity

Although there is an extensive literature on collective identity, with considerable discussion regarding its conceptualization and sources, this literature has been relatively mute regarding variation in its form. The concept is most often invoked as if it is an invariant, uniform collective phenomenon. This is not the case, however, as collective identities can be multidimensional and be multilayered within a specific locus, and they may also vary by type. The multidimensionality of collective identity is indicated by reference to its

cognitive, emotional, and moral dimensions (Jasper and Polletta 2001, Melucci 1989). The relative importance of each of these dimensions to the vitality and motivational force of a collective identity has not been elaborated, however. Presumably the presence of each of these dimensions yields a more robust and vital collective identity. Clearly a collective identity in which the boundaries between 'us' and 'them' are unambiguously drawn, in which there is strong feeling about those differences, and in which there is a sense of moral virtue associated with both the perceptions and feelings, should be a more potent collective identity than one in which either the emotional or moral dimensions are weakly developed.

Similarly, several analyses have noted that collective identity can be multilayered, such that there can be variation in its locus. Three such layers have been noted with respect to social movements (Gamson 1991, Stoecker 1995). They include, beginning with the broadest and potentially most inclusive layer: the social movement community or solidary group, which can be thought of as the constituent layer, as in the case of black Americans in relation to the civil rights movement; the social movement layer as in the case of the civil rights movement; and the organizational layer, as represented by the Southern Christian Leadership Conference (SCLC), the Congress for Racial Equality (CORE), and the Student Nonviolent Coordinating Committee (SNCC) in the context of the civil rights movement. In principle, each successive layer may be embedded in the larger more inclusive layer, giving rise to a generalized, cohesive collective identity at the community or national level. But clearly the existence of a collective identity at one level does not automatically generalize to or incorporate another level. Thus, collective identities can be built around the organizational carriers of a movement, as in the case of SCLC and SNCC, without necessarily representing the broader movement, which indicates the potential for identity conflicts at the collective level and the potential for schism and factionalization. Such observations suggest the need for more careful consideration of the often multilayered character of collective identities and of greater specification of the ways in which they can interact and combine, and with what consequences. Also, these observations call for caution in generalizing about the scope and functioning of collective identities, particularly with respect to broader social categories such as ethnicities and nationalities.

Finally, it is reasonable to wonder if collective identities vary by type. At the most general level, Hunt et al. (1994) distinguish among protagonist, antagonist, and audience or bystander identities, arguing that even though protagonist or oppositional identities have received most of the scholarly attention, each type or field of identity is fundamental to understanding the interactive dynamics underlying the emergence, character, and functioning of collective identities. Noting as well that collective identities arise and operate within an interactive context 'marked by power relationships,' Castells distinguishes among legitimizing, resistance, and project collective identities (1997, pp. 7–10). Legitimizing identities are associated with dominant institutions or the state, whereas both resistance and project identities represent two basic forms of the antagonist identity, the former generated by devalued and stigmatized collectivities, and thus constituting the seedbed for identity movements and politics, and the latter associated with movement beyond resistance to the construction not only of alternative identities but also a new system that valorizes rather than defiles the new identity. The important issue is not whether such typologies of collective identities are exhaustive, but the emphasis on their contextually embedded and interactional character and their different consequences.

See also: Collective Beliefs: Sociological Explanation; Discourse and Identity; Identity and Identification: Philosophical Aspects; Identity Movements; Identity: Social

Bibliography

Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Benford R D, Snow D A 2000 Framing processes and social movements: an overview and assessment. *Annual Review of Sociology* **26**: 611–39

Calhoun C 1991 The problem of identity in collective action. In: Huber J (ed.) *Macro–Micro Linkages in Sociology*. Sage Publications, Newbury Park, CA

Calhoun C 1997 *Nationalism*. University of Minnesota Press, Minneapolis, MN

Castells M 1997 *The Power of Identity*. Blackwell Publishers, Oxford, UK

Cerulo K A 1997 Identity construction: new issues, new directions. *Annual Review of Sociology* **23**: 385–409

Cornell S, Hartmann D 1998 *Ethnicity and Race: Making Identities in a Changing World*. Pine Forge Press, Thousand Oaks, CA

Eisenstadt S N, Giesen B 1995 The construction of collective identity. *Archives Europeanes de Sociologie* **36**: 72–102

Eyerman R, Jamison A 1998 *Music and Social Movements: Mobilizing Traditions in the Twentieth Century*. Cambridge University Press, Cambridge, UK

Fantasia R 1988 *Cultures of Solidarity: Consciousness, Action, and Contemporary American Workers*. University of California Press, Berkeley, CA

Friedman D, McAdam D 1992 Collective identity and activism: networks, choices, and the life of a social movement. In: Morris A D, Mueller C M (eds.) *Frontiers in Social Movement Theory*. Yale University Press, New Haven, CT

Gamson W A 1991 Commitment and agency in social movements. *Sociological Forum* **6**: 27–50

Gergen K 1991 *The Saturated Self. Dilemmas of Identity in Contemporary Life*. Basic Books, New York

Giddens A 1991 *Modernity and Self-identity: Self and Society in the Late Modern Age*. Stanford University Press, Stanford, CA
Hunt S A, Benford R D 1994 Identity talk in the peace and justice movement. *Journal of Contemporary Ethnography* **22**: 488–517
Hunt S A, Benford R D, Snow D A 1994 Identity fields: Framing processes and the social construction of movement identities. In: Larana E, Johnston H, Gusfield J R (eds.) *New Social Movements: From Ideology to Identity*. Temple University Press, Philadelphia, PA
Jasper J M, Polletta F 2001 Collective identity and social movements. *Annual Review of Sociology* **27**: 283–305
James S 1999 *The Atlantic Celts: Ancient People or Modern Invention?* British Museum Press, London
Jensen J 1995 What's in a name? Nationalist movements and public discourse. In: Johnston H, Klandermans B (eds.) *Social Movements and Culture*. University of Minnesota Press, Minneapolis, MN
Klapp O E 1969 *Collective Search for Identity*. Holt, Rinehart and Winston, New York
Laitin D D 1998 *Identity in Formation: The Russian-speaking Populations in the Near Abroad*. Cornell University Press, Ithaca, NY
Levitas R 1995 We: Problems in identity, solidarity and difference. *History of the Human Sciences* **8**: 89–105
Melucci A 1989 *Nomads of the Present: Social Movement and Identity Needs in Contemporary Society*. Temple University Press, Philadelphia, PA
Melucci A 1995 The process of collective identity. In: Johnston H, Klandermans B (eds.) *Social Movements and Culture*. University of Minnesota Press, Minneapolis, MN
Nagel J 1996 *American Indian Ethnic Renewal: Red Power and the Resurgence of Identity and Culture*. Oxford University Press, New York
Pastoureau M 1997 *Heraldry: Its Origins and Meanings*. Thames and Hudson Ltd., London
Schwalbe M L, Mason-Schrock D 1996 Identity work as group process. *Advances in Group Processes* **13**: 113–47
Snow D A, Anderson L 1987 Identity work among the homeless: The verbal construction and avowal of personal identities. *American Journal of Sociology* **92**: 1336–71
Snow D A, McAdam D 2000 Identity work processes in the context of social movements: Clarifying the identity/movement nexus. In: Stryker S, Owens T, White R W (eds.) *Self, Identity and Social Movements*. University of Minneapolis Press, Minneapolis, MN
Stoecker R 1995 Community, movement, organization: The problem of identity convergence in collective action. *Sociological Quarterly* **36**: 111–30
Stone G P 1962 Appearance and the self. In: Rose A (ed.) *Human Behavior and Social Processes*. Houghton Mifflin, Boston
Stryker S 1980 *Symbolic Interactionism: A Structural Version*. Benjamin/Cummings, Menlo Park, CA
Taylor V, Whittier N E 1992 Collective identity in social movement communities: Lesbian feminist mobilization. In: Morris A D, Mueller C M (eds.) *Frontiers in Social Movement Theory*. Yale University Press, New Haven, CT
Trevor-Roper H 1983 The invention of tradition: The Highland tradition of Scotland. In: Hobsbawm E, Ranger T (eds.) *The Invention of Tradition*. Cambridge University Press, Cambridge/London
Walder A 2000 Identities and interests in the Beijing Red Guard Movement: Structural explanations reconsidered. Unpublished Manuscript

D. A. Snow

Collective Memory, Anthropology of

Focused on shared, collective, or social memory practices, this article examines the linkages between memory and history with attention to the fields of power, in which the struggle for domination over remembrance and tradition, the manipulation of retrievable historical consciousness and collective forgetting, takes place.

1. *Definition of Concepts: Memory and History*

The term *history* is often used to describe representations of the past that appear in written or narrative form, a primary medium through which states, elites, or dominant descent groups confiscate linear time and proclaim official chronologies as master narratives. The writing of history is 'a colonization' of time 'by the discourse of power' (de Certeau 1988). The term *memory*, by contrast, is conventionally applied to those oral, visual, ritual, and bodily practices through which a community's collective remembrance of the past is produced or sustained (Connerton 1989). It includes the vast complex of unofficial, noninstitutionalized knowledge not yet sedimented into formal traditions and which represent the 'collective consciousness' of whole groups (Halbwachs 1980), forming a counterweight to the knowledge that is privatized and monopolized by certain elites for the defense of established interests. While memory may be a moving reservoir of history, furnishing the 'raw material' for representing the past, it is not the same as history (Le Goff 1992). Individual remembrance, collective memory, and narrative history interact in highly complicated ways, shaping each other as different versions of the past are constructed and reconstructed, modified, and invented.

2. *The Politics of Time: History and Anthropology*

The theories, models, and methods that were developed by anthropologists to study non-European worlds tended toward a 'rejection of historical research of any kind' (Evans-Pritchard 1962), at least within the limits of the discipline. Having as its central theme 'the unlettered and forgotten peoples,' the anthropological endeavor was inclined to disregard the existence of indigenous histories, which included, in particular, the trauma of the colonial encounter. But if

making history is a social practice that produces peoples, bodies, and places, as de Certeau (1988) suggests, then the very erasure of such a universe of time and being must be understood as a figuration of power. The effacement of historicity is a political operation.

Nineteenth-century anthropology was deeply implicated in the projects of Western modernity: the construction of nation-states and the creation of colonial systems with which Europeans came to dominate other worlds. During this process of global expansion, and the establishment of modern hegemonic polities, the past was increasingly mobilized as a symbolic discourse 'in the definition of, and the marking of, the boundaries of states and nations' (Cohn 1981, p. 228). Anthropologists of this period contributed to the colonial undertaking by a reliance on analytic tools that used Western notions of time as an index of unequal evolutionary development. The prerogative to know and control the past, by defining it as 'history' (an order of chronological or linear time), was an entitlement of civilization. The social worlds outside of Europe came to be marked as timeless: a closed symbolic universe inhabited by 'the people without history' (Wolf 1982). Displaced into the sphere of alterity, colonial subjects were incarcerated by the temporal logic of an expansive European world system.

The past as history became a way to construct schemes of classification which differentiated Europeans from others. Anthropologists began to delineate stages of human development via the comparative method: by identifying phenomena of a general type (taboo, totemism, polygamy), their goal was to map out a global history of social institutions. In this endeavor, however, analytic units were 'taken out of the flux of time' (Evans-Pritchard 1962, p. 175): social forms, stripped of their unique features and context-bound meanings, escaped temporality, appearing static and unchanging. The comparative method, in its attempt to achieve conceptual stability as a sociological proposition, was explicitly anti-historical in emphasis and thereby served to obscure the effects of Western contact and domination.

By the early twentieth century, anthropologists had decisively abandoned the comparative paradigm. But in spite of the shift to new methods of analysis, the discipline's aversion to historical research intensified. Scholars in Britain, Germany, and the United States asserted that the customs of native peoples were to be found not in colonial archives, missionary reports, or dubious travel accounts but *in the field*: reliable data about non-Western societies demanded first-hand observation. This was accomplished at first through brief expeditions, and later by long-term residence of the anthropologist in native communities. British scholars, in their interpretation of the ethnographic venture, reasserted the enlightenment program: anthropology was to become 'a natural science of society, based on a descriptive analysis of the structure and functioning of social forms' (Cohn 1981, p. 232). American scholars, by contrast, adopted a romantic idealist stance: anthropology emerged as a hermeneutic science, focused on the descriptive reconstruction and symbolic capture of entire cultural systems. Although the nation-bound projects of anthropology began to diverge, the ethnographic field method was adopted as a common research standard, thereby imparting and codifying the now conventional disciplinary gaze: the observers' eyes were trained to *see* society only at a particular point in time. Ethnography, in its obsession with reifying the authenticity of peoples, fabricated a closed temporal frame: a *present present* or a *past present*, as Evans-Pritchard (1962) termed it in a critical commentary on the discipline's 'breach with history,' illuminating the distortions which flow from the imposition of *a priori* temporal limits upon 'traditional societies.'

But it was precisely this aversion to history that led anthropologists to discover 'memory' as a valuable ethnographic tool. Evans-Pritchard (1962), for instance, suggested that the past was always 'incapsulated in a context of present thought': the past, *as memory*, was embedded in material, ritual, and narrative practices and therefore 'part of the social life which the anthropologist can directly observe' (pp. 177–8). The pursuit of collective memory was to constitute a major change in ethnographic practice: the 'life story' of local worlds—of bodies, magic, and markets—could be rendered intelligible through the oral archive. But in the colonizing imagination of Western anthropologists, the work of memory was consigned to the uneventful register of *structural*, *mythical*, and *sacred* time: an order of signs without history. Peoples' means of remembering traumatic events (famine, migration, conquest) 'were always seen to reinforce the system in place, never to transform it' (Comaroff and Comaroff 1992, p. 21). In a shift from historical past to ethnographic present, anthropologists erected 'counterfeit signposts' of a primitive world of enduring traditions—an analytic fiction. The anthropology of collective memory, in consequence, began as a kind of anti-history.

3. *The Shape of Memory: Modes of Historical Consciousness*

In their ethnographic explorations, anthropologists encountered indigenous memory forms that did not seem to fit the European idea of historical time: an objective, unmediated, linear account of past events. By a focus on myths, ritual, oratory, and body practices, they uncovered the narrative and performative possibilities of *a history of the present*: the social sense of time whereby people attempted 'to grasp the world as both a synchronic and a diachronic totality' (Lévi-Strauss 1966, p. 263). In the realms of memory,

past and present—although theoretically distinct—were conjoined to link, from generation to generation, the living with the dead.

Among Australian aborigines, for instance, *historical* or *commemorative rites* 'recreate the sacred and beneficial atmosphere of mythical times—the 'dream age'—mirroring its protagonists and their great deeds,' and transport 'the past into the present'; *rites of mourning* 'correspond to an inverse procedure: instead of charging living men with the personification of remote ancestors, these rites assure the conversion of men who are no longer living into ancestors,' thereby bringing 'the present into the past' (Lévi-Strauss 1966, pp. 236–7).

These images of time are different, perhaps, but not 'other.' In the rites performed, past and present are propelled into a social space—the body. The act of remembering takes corporeal form: memory is embodied (Connerton 1989). But this transposition of temporal worlds, which connects the living and dead to the ancestors via the mimetic power of embodiment, is not an abrogation of time. It is a generative reconstitution of society: a practice of empowerment, a political engagement with the matrix of life (Stoller 1995).

Other anthropological studies suggest that collective memory may be systematized on the basis of *relative temporalities*, that is, a flexible system of polymorphous reference points. Among certain peoples of the Ivory Coast, like the Guéré, 'the consciousness of a historical past has developed alongside a multiplicity of other times' (Le Goff 1992, pp. 8–9): *mythical*, *genealogical*, *lived*, and *projected time*. Such a register of memory forms belongs to the aftermath of the colonial project. Under the impact of prolonged social trauma, the disparate logics of time and memory chart the precarious matrices of life: a synchrony of struggle, violence, and perseverance.

4. *Memory-sites: The Localization of Remembrance*

The anthropological analysis of collective memory moves toward a renunciation of linear temporality in favor of multiple kinds of time as experienced at the levels where the individual takes root in the social: language, demography, economics, politics. Memory production is in this sense not dependent on a single individual's experience of the past. A remembrance, a social understanding of events that is represented as memory, can be constructed by sharing with others 'sets of images that have been passed down to them through the *media of memory*—through paintings, architecture, monuments, ritual, storytelling, poetry, music, photos, and film' (Watson 1994, p. 8). In every society, we can identify an array of memory-sites or places of commemorative record and practice where remembrance anchors the past: *topographical* places (archives, libraries, museums); *monumental* places (cemeteries, architectural edifices); *symbolic* places (commemorative rites, pilgrimages, emblems); *functional* places (manuals, autobiographies, associations); and places of *power* (states, elites, milieux) which 'constitute their historical archives in relation to the different uses they make of memory' (Le Goff 1992, pp. 95–6, Nora 1986). These memory-sites furnish a series of locations where knowledges of the past are conveyed and sustained by a circulation of signs that calls attention to its own logic of inclusion, exclusion, and selective incompleteness.

5. *Unsanctioned Archives: Memory and Countermemory*

The anthropology of collective remembrance is not only concerned with affirmation and heritage, but with processes of contestation, the aftermath of violence, and the emergence of countermemories (Stoller 1995). A *social history of remembering* must take account of 'the mechanisms that make unsanctioned remembrance possible' under repressive political regimes: in asking how people remember what is meant to be forgotten, 'matters of transmission'—the mechanics of shared memory and hidden histories—take center stage (Watson 1994). Here, interpersonal memories, commemorations, theater, drama, folklore, secret, and oppositional histories are 'the venues within which alternative remembrances'—unauthorized and unapproved memories of the past—can be located and analyzed (Watson 1994, p. 2). Collective memory practices are not only linked to sites of domination that attempt to legitimate a given social order, but also to unsanctioned sites of struggle, opposition, and transformation.

6. *The Traffic in Memory: Global Hegemonies and Local Struggle*

In all societies, there exist distinct moments when new representations of the past are forged, contested, and put to cultural and ideological use. Battles waged over ethnic origin or national identity may be linked to the constitution of radically new kinds of memory-archives (Hobsbawm and Ranger 1983). The creation of nations, ethnicities, and subjects inevitably alters the shape of the past. 'Memory' is therefore not a generic term of analysis, but itself an object appropriated, transformed, and politicized. Or, put differently, memory can be nationalized, medicalized, aestheticized, gendered, bought, and sold.

Historical consciousness shaped by nation-building, scientific ideology, and global capitalism plays a significant role in the formation of social identities. Detailing the politics of remembrance in postsocialist China, Schein (2000) shows how state-orchestrated

reforms, staged against the chaos of the Cultural Revolution and implemented during a period of increasing exchange and fascination with the West, produced a popular appetite for exotic others: commodified images of minority peoples, like the Miao, were fashioned by an acute nostalgia for 'ancient' traditions and those feminized cultural essences on which to recraft a distinctive Chinese nationness. The result was a boom in memory recuperation among China's urbanites and Miao elites, who simultaneously played to a Western valorization and high-spending tourist consumption of dwindling pasts. But such a newly forged cultural attachment to memory artifacts, the 'fossil bed' phantasms of folk heritage, also entails transformation. In other words, the work of symbolic restoration 'may conceal very new projects' (Le Goff 1992, p. 9). The recruitment of the past for revolutionary or political ends, as in the case of China, is entangled in a twofold venture: the forging of a distinctive ethnonational imaginary and the renegotiation or recalibration of *place* in the global order.

The pursuit, rescue, and celebration of collective memory is always socially motivated and has, thus, to be understood in positional terms. The global context—the theater of transnational consumer capitalism—is, according to Friedman (1994), particularly relevant in understanding how local memory forms are affected by the formative impact of world systemic processes. In the case of modern Greece, as Friedman shows, national selfhood was crafted by internalizing the ways in which European elites, in constructing their own 'civilized' origins, mythologized classical antiquity as the cradle of Western civilization. Greek nationalists found their past in the institutional memory of the West, thereby accelerating incorporation into an expanding European world system. Such acts of cosmological repositioning are symptomatic of a transglobal system that fuses socioeconomic transformation and collective memory to 'the reconfiguration of the map of the world's peoples' (Friedman 1994, p. 123).

Yet even in the global arena, there exist other possibilities of memory production. In the Hawaiian case, Friedman's instructive contrast, indigenous communities attempted to extricate themselves from Western dominance by projecting a value system, produced in the modern context, onto an aboriginal past. The result was a body of memories composed of mythic history, a work of folklore and folklorization 'rife with simulacra of tropical fantasies.' Social and economic marginalization, in the aftermath of World War II, gave rise to an intense period of Hawaiian cultural revival. Recuperation efforts, centered on the reconstitution of ethnic history by retrieving precolonial social forms, brought into play a *museology of memory* that was staged in opposition to the West: a precontact Hawaii stocked with all the items to be found in Western historical archives—libraries, museums, missionary records. The Hawaiian confiscation of these colonial memory-stores proceeded, as Friedman documents, by forging an indigenous past uncoupled from the larger world, a Western-imposed modernity, which had obliterated a population and absorbed its social history into the projects of global economic systems.

Anthropologists have firmly established the need to attend to these paradoxes of mobile cultural and symbolic forms—this *traffic in memory*—during periods of transnational crisis and restructuring (Appadurai 1996). In the context of global systems, where the practice of identity is brought on stage, the logic of time and the meaning of memory have been radically transposed. In the global arena, the production of historical consciousness becomes increasingly entangled with the commodity form: encoded by temporal longings and mythic value, it furnishes the common stock for a *memory market* saturated by cargo-type products of surplus and lack, bearing the deep imprint of the globally hegemonic mode in which people's ethnonational remembrances are manufactured, exchanged, and consumed (Taussig 1980). But memory consumption can also be a subordinate production, one that creatively and tactically resists the homogenizing tendency of centralized systems. This suggests, as Schein (2000) observes, an alternate sense for cultural production: 'the myriad ways in which people make use of received [memory] products presents a measure of possibility, not only for autonomy, but also for subversion of the dominant order' (p. 16). Put differently, the modalities of collective memory practice attest to the possibilities of *symbolic creativity*—even in the province of globalizing modernities.

See also: Collective Memory, Psychology of; Cultural History; Expressive Forms and the Evolution of Consciousness; History and Memory; Memory: Collaborative; Memory Retrieval

Bibliography

Appadurai A 1996 *Modernity at Large*. University of Minnesota Press, Minneapolis, MN

Cohn B S 1981 Toward a rapprochement. *Journal of Interdisciplinary History* **12** (2): 227–52

Comaroff J, Comaroff J 1992 *Ethnography and the Historical Imagination*. Westview Press, Boulder, CO

Connerton P 1989 *How Societies Remember*. Cambridge University Press, Cambridge, UK

de Certeau M 1988 *The Writing of History*. Columbia University Press, New York

Evans-Pritchard E E 1962 *Social Anthropology and Other Essays*. The Free Press of Glencoe, New York

Friedman J 1994 *Cultural Identity and Global Process*. Sage, London

Halbwachs M 1980 *The Collective Memory*. Harper & Row, New York

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK
Le Goff J 1992 *History and Memory*. Columbia University Press, New York
Lévi-Strauss C 1966 *The Savage Mind*. University of Chicago Press, Chicago
Nora P 1986 *Les lieux de mémoire*. Gallimard, Paris, 3 Vols.
Schein L 2000 *Minority Rules*. Duke University Press, Durham, NC
Stoller P 1995 *Embodying Colonial Memories*. Routledge, New York
Taussig M T 1980 *The Devil and Commodity Fetishism in South America*. University of North Carolina Press, Chapel Hill, NC
Watson R S (ed.) 1994 *Memory, History, and Opposition under State Socialism*. School of American Research Press, Santa Fe, NM
Wolf E R 1982 *Europe and the People Without History*. University of California Press, Berkeley, CA

U. Linke

Collective Memory, Psychology of

1. Introduction

Collective memory is a sociological concept, though shot through with psychological presumptions. Its most important theorist is Maurice Hawlbachs, a second-generation student of Emile Durkheim, who wrote *Les Cadres sociaux de la memoire* (*The Social Frameworks of Memory* 1992) and *La Memoire Collective* (*Collective Memory*) ([1950], 1992). Hawlbachs and those working in the collective memory tradition posit that not only do individuals remember, but collectivities as well. Just as there is great interest among psychologists, psychoanalysts, and neuroscientists to discover, describe, and specify the processes through which individuals remember—*how, what, when*, and *why* remembering occurs—so too do social scientists—sociologists, historians, anthropologists, political scientists—attend to the social processes through which collectivities—families, groups, nations—recover the past, conceptualize it through narrative structures, and memorialize it through memorials, museums, rites, and other commemorative forms.

Hawlbachs' essay was first published in English in 1980 with an introduction by Mary Douglas, herself a neo-Durkheimian. In her introduction, Douglas drew the link between Hawlbachs' concept and the research of the British psychologist, Sir Frederick Bartlett who, in *Remembering: A Study in Experimental and Social Psychology* (1932), contextualized memory by arguing that what is remembered of the past depends upon frameworks or schemas of understanding imposed upon it from the present. Since the publication of Hawlbachs' writings in English, the interest in collective memory has greatly accelerated the development of the themes defined by Hawlbachs, Bartlett and Douglas and their influence on disparate academic fields. In the last decades of the twentieth century, collective memory generated a large literature focusing on the relation between the past and the present, raising questions especially about how the present influences the reconstruction of the past and the ways in which historical events and circumstances impinge or intrude on the present. These issues have often been addressed through empirical studies of particular rites, festivals, commemorations, and memorials in different national and historical settings.

What follows is a discussion of the sources of convergence and divergence in the psychology of collective memory as expressed in the academic literature and as it impacts on these broader politics of memory. There is widespread agreement concerning both the reality of collective memory as well as its malleability, i.e., memory's susceptibility to revision, manipulation, etc. However, differences among scholars exist as well, focusing particularly on the question of the motives and interests that lie beneath the phenomenon and determine its form. A neo-Durkheimian framework will be characterized where the past is understood as a symbolic resource often employed to reduce tensions and strains presently confronting the collectivity; collective memory is understood as a complex social process—including instrumental and symbolic struggles by various members of the group over the definition of the past—to strengthen the ties that bind individual members together within the collectivity. Others adopt a neo-Freudian perspective where tensions and strains within a present day collectivity are largely understood as deriving from the unavoidable intrusion of a traumatic past. The past is the source, rather than a resource, of challenges confronting a present-day collectivity. Collective memory, here, is often seen as an impressive impediment to the establishment of a healthy, post-traumatic community.

2. Individual and Collective Memory: Convergence and Consensus

Students of collective memory, like researchers on individual memory studies, understand memory to be a constitutive feature of the present. In both, memory is distinguished from history; the former describes a process of reconstruction, occurring *ex post facto* and possibly bearing little relation to the historical past being remembered. Nonetheless, an understanding of oneself in relation to an 'as if it were true' past is believed to be an essential feature of individual life. Personal identity depends on locating oneself in a particular past in which the present is bounded and made meaningful through history. The person experiences him or herself in a continuous, com-

prehensible relation to a past; an orientation toward the present and the future is only coherent if in relation, in part, to an understood past. Collective memory similarly appreciates the temporal dimension of social life, where sociality depends on a past to support and give continuity and meaning to a present. No collectivity can function without a sense of its own traditions and its own continuity with the past (Shils 1983, Elias 1994). Where no such past exists—as in new collectivities, for example—traditions, as Eric Hobsbawm has argued, are 'invented.' 'The invention of tradition' (Hobsbawm 1983) is a social process in which memory is confabulated on behalf of a present requiring a history, for the purpose of creating a collective or national identity. But if collective memory emphasizes the diachronic relation between a present that requires a past, it also underscores the fundamentally social and synchronous character of memory formation. Collective memory concerns itself with dialogical, discursive processes occurring as a result of communication between various groups and institutions comprising the collectivity.

Just as the autobiographical memory of the individual is not identical to the history being recalled —always subject to reformulation, reconstitution, and distortion as a result of the temporal distance between remembering and what is remembered (Schachter 1995, Loftus and Ketcham 1994)—so, too, is collective (or historical) memory vulnerable to similar possibilities of confabulation, revision, and remodeling (Kammen 1995, Schudson 1995). Remembering, whether individually or collectively, always involves the representation after the fact of past events or experiences, never the return to the actual past. Freud himself described this parallel process between individual and collectivity when he wrote

> If we do not wish to go astray in our judgment of their historical reality, we must above all bear in mind that people's 'childhood memories' are only consolidated at a later period, usually at the age of puberty; and that this involves a complicated process of remodeling, analogous in every way to the process by which a nation constructs legends about its early history. (1913, p. 206n)

Particular memories, rather than being an enduringly stable feature of either individual or group life, are continually susceptible to instability and disruption. However, temporal distance is not the only explanation for the breach between history and memory. Interest and/or affect—present-day requirements of the individual or the collectivity—dictate the need for remembering. The imperatives of the present are decisive, it is believed, in the nature and character of remembering that occurs. As those present-day needs change, or are altered by circumstance, so too are memories altered. While there is probably not complete alteration, memory researchers are impressed by the capacity for substantial revision to occur.

3. *The Neo-Durkheimian Perspective: The Embeddedness of Collective Memory*

American sociology of collective memory has been dominated by a 'presentist' psychology, building upon, refining, and revising elements of Hawlbachs' insight concerning the present's contribution to the recovery of the past. Here, research has focused on the relation between memory as a mechanism of response and contemporary strains in the collectivity and other social practices similarly related to responding to tensions within the social fabric. The unifying idea behind this research rests on the importance of a common 'collective consciousness' to social solidarity and cohesion—as originally defined by Durkheim (1997)—and the important role that memory plays in constituting that consciousness. Memory, it is argued, is 'embedded' (Prager 1998) in 'mnemonic communities,' (Zerubavel 1996) meaning that it is constituted, elaborated, and altered in response to social processes and practices occurring in a collectivity at any given point in time. Armed with an uncontested understanding of the past, collective consciousness is strengthened because of a shared identity with the past; divergent memories, in contrast, express disunity and challenges to social solidarity.

3.1 *Rites and Sites of Remembrance*

Collective consciousness is heavily structured through historical referents and the practices of rituals, rites, and commemorations in part serve to strengthen a common sense of the collectivity through shared history. Shils and Young (1953, see also Cannadine 1983) writing about the British coronation ceremonies and Ozouf (1991) describing ritual forms of performative remembering of the French Revolution identify the social significance of these forms of political practices and, through careful readings of both the content and the form of commemoration, describe the symbolic process of historical recovery on behalf of present-day political and ideological needs and interests (Connerton 1989). While it has been noted by Bakhtin (1968) and others that commemorations, like other moments of liminality (Turner), have the potential to take on a life-of-their own (e.g., Davis (1977) by either anticipating the restructuring of the collective consciousness or promoting social division, collective memory research has largely concerned itself with the particular stabilizing effects of past events remembered (e.g., Olick 1999, Schwartz 1996, Wagner-Pacifici 1991, Zelizer 1992) and its contribution to the collective consciousness. Nora (1989) calls attention to *les lieux de memoire*, or sites of memory, highlighting the necessity in the modern world to locate special sites for memory to occur. Because of the discontinuity in modern societies between memory and history—in contrast to tra-

ditional societies where 'real environments of memory' exist—memory sites occupy sacred ground self-consciously constructed to mark the present's continued link with its past and to combat forgetting.

3.2 Frameworks of Remembering

Contemporary researchers have identified narrative, discursive, and other semantic features of collective memory to demonstrate the socially embedded, constructivist, and presentist quality of collective memory. In *Presenting the Past*, Prager (1998) describes how individual memory relies upon available cultural narratives to provide context and meaning to what otherwise would be relatively meaningless and random events in an individual's life (see also Schudson 1995). These narratives—for example, a tale of victimization or of physical or sexual abuse—speak to contemporary understandings of the ways in which the past informs or frames the present. Collective memory, too, relies on culturally provided narratives coding the past so as to make it meaningful and available to the present. The past can only be remembered discursively and rhetorically, as part of a dialogical process between social members or organizations collectively remembering (Middleton and Edwards 1990, Douglas 1986, Shotter 1990).

3.3 History as Constraint

'The past in the present versus the present in the past' (Schudson 1989) captures one important emendation to Hawlbach's presentism, arguing against a too strong unmooring of the present from the past. Seeking to explicate the ways in which history imposes certain constraints upon what can be remembered, researchers have sought to restrict an unbridled social constructivism seemingly encouraged by Hawlbachs and others. Collective memory, despite its role as a resource for legitimation, cannot completely override history. As Schudson argues, the past imposes itself on memory (a) through the strength of prior remembrances that cannot be contested, e.g., the institution of slavery in American history, but instead are part of an obdurate historical reality that cannot be ignored, and (b) because the process of historical reconstruction occurs through a 'politics of memory' in which various groups within the collectivity offer differing versions of the past, are debated and socially negotiated, and form 'the rhetorical structure of social organization' (see Schudson 1992). A third constraint that elevates history in relation to memory, according to Schudson, is the role of individual trauma, a topic that will be returned to shortly. In a similar vein, Schwartz (1996) argues against a radical presentism that fails to recognize the importance to the collectivity of continuity and coherence with the past. He suggests that the reinterpretation of the past *de novo* is inhibited by counteracting social imperatives to understand the present as continuous with and coherent to what has preceded it. Finally, Olick (1999) strikes a similar cautionary note by emphasizing the ways in which earlier commemorations of historical events themselves serve as constraining influences on current memory. He refers to the dialogue between current memorializations and prior ones as constituting 'genre memories,' in which the present is discursively steeped in these mediated pasts.

4. Neo-Freudianism: Trauma and the Embodiment of the Past in the Present

Neo-Durkheimian approaches to collective memory emphasize the continuity of the present to the past, underscoring the important role that remembering plays in the constitution of the present. A neo-Freudian perspective, in contrast, insists that collective memory is a social process in response to social ruptures, or discontinuities, that have occurred in the past that, because not fully assimilated in conscious experience, subsequently interfere with the smooth functioning of collective life. The title of Friedlander's autobiographical work, *When Memory Comes* (1991), an account of his discovery of his Jewish roots in pre-Holocaust Czechoslovakia, captures this idea that memory is an imperative, driven from the past, intruding upon and interfering with the present. Memory, as the title suggests, *will* come. Memory and Freud's 'return of the repressed' are complementary concepts: together they define the present as always susceptible to intrusions from the past. Collective memory expresses not merely social practices but rather processes driven by psychological law. Friedlander (1992) acknowledges the large debt that scholars of the Holocaust owe to Freud when they invoke his ideas to the remembering and understanding of this event. The study of the Holocaust, it should be noted, has emerged as the paradigmatic literature for those researchers studying the phenomenon of collective memory in various world settings and who approach it in relation to past social ruptures. The past, it is argued, persists as a kind of 'haunting' presence in the present, hovering but intangible.

The endurance of 'traumatic disruptions' occupy, for the neo-Freudians, comparable conceptual terrain to that of collective consciousness for the neo-Durkheimians. But in the latter, the past finds expression in the present: trauma dictates that the past survives *as if* it were the present. Caruth (1996) defines trauma as wounds inflicted on the individual too overwhelming to be fully absorbed and available to consciousness. The consequence is that trauma lives on unconsciously—latently—but ultimately finds a voice in the present. A traumatic moment becomes 'embodied' (Prager 1998) or engraved on the person

and, in time, intrudes upon later experiences in the present (Van Der Kolk and Van Der Hart 1996). Yet if this process describes the persistence of the past in individual experience, similar processes are also believed to shape collective experience, in which collective memory is characterized as the mediated process of remembering earlier collective or cultural trauma.

Moses and Monotheism (1939) serves as a foundational text, in which Freud describes how trauma transforms the present as repetition of a prior traumatic moment. He characterizes the Jewish murder of their leader, Moses, as a traumatic reliving of the earlier murder of the primal father and thus captures the way in which traumatic guilt, for a time buried, finds expression in an unconscious reenactment of the past in the present. As Caruth asserts about Freud's argument, traumatic history helps explain not the practices that help reinforce the collective consciousness, but the overwhelming traumatic history that shapes the culture. Collective memory articulates the relation of a particular present for a group to its traumatic history that helped constitute it; in *Moses and Monotheism*, Freud attempts to capture the historical origins of both a particular Jewish culture, and the Christian response to it. Building upon this theme but extending it, Asman (1997) asserts that an original traumatic assault on an early period of Egyptian monotheism, long before Moses, continued to serve as a 'stabilizer of memory,' repetitively creating a religious enemy, long after the 'real' memory no longer possessed any carriers. While this early brutal assault on the first monotheists did not survive in memory, its real impact nonetheless led to the Egyptian experience of the Jew as a symbol of danger and crystallized into Western anti-Semitism, which continues to this day to shape Western political culture. Here, trauma knows no closure—even after the events are long forgotten and the victims no longer survive, it shapes and molds subsequent history in the image of the traumatic past.

Profound social disruptions—like slavery and the Civil War, the Vietnam War, and political assassinations in the United States (Neal 1998), the dictatorships and the 'disappeared' in the Southern Cone of Latin America in the 1970s and 1980s (Hollander 1997), the European Holocaust, the rise and fall of communist regimes in Eastern Europe, ethnic conflicts in Central and Eastern Europe—have contributed to the prolific production of neo-Freudian scholarship on post-traumatic collective memory. Unlike neo-Durkheimian perspectives that tend to downplay the reality of historical experience and its impact upon the collectivity in favor of an appreciation of constructivism, a neo-Freudian approach takes seriously the materiality of terror and oppression and seeks to understand its enduring consequences after the fact on the collectivity. The lasting effects of the past on the present rely on the ideas of the persistence of unconscious guilt, transference, and projection. Despite the appeal, this perspective holds because it seriously confronts historical realities and its insistence that social institutions and individuals are irrevocably transformed as a result of these experiences, these concepts nonetheless seem inadequate analytic tools to provide a psychological justification for the survival of trauma long after its victims have disappeared. A more robust theory of the transmission of trauma across generations, or across eons, is necessary, if one is not to think that ultimately history, however traumatic, in time finds closure in spite of the enormity of the crime, the disruption, the terror.

5. *Conclusion*

Collective memory is of more than passing academic interest. Hannah Arendt (1951, p. 423) defines totalitarianism as a system that forbids grief and remembrance; it severs an individual's or a group's relation to a continuous past. This obliteration of the past, for Arendt, is synonymous with the death of moral man. Today, political elites throughout the world appreciate the constructed nature of a national memory and seek, through it, to shape their own political cultures. Similarly, there is often recognition of traumatic pasts that threaten to overshadow present experience and their need to proactively respond. As many nation-states in Eastern Europe, Southern Africa, and Latin America have attempted to recover or institutionalize democracy in the 1980s and 1990s, the challenge to incorporate the past into current politics—either by defining the polity in relation to the past or by insisting upon its transcendence—relies on both politics and psychology of collective memory. The Tribunal of Truth and Reconciliation, for example, in the Republic of South Africa, not only acknowledges collective memory as a political reality, but also endorses a psychology that recognizes the heuristic effects of remembering to overcome the legacy of apartheid and to promote national unity. Other nations, still cognizant of the importance of collective memory, rather than promoting reconciliation between past and present, have implemented a politics of 'forgetting,' taking the position that the present can serve to refashion the past. The psychology of collective memory has emerged as a central point of political and intellectual controversy in various democratic states, and the politics of memory a central feature of post-dictatorial states.

See also: Collective Beliefs: Sociological Explanation; Collective Identity and Expressive Forms; Collective Memory, Anthropology of; Commemorative Objects; Durkheim, Emile (1858–1917); Memory Retrieval; Psychoanalysis in Sociology; Psychoanalysis: Overview; Ritual; Social Trauma; Solidarity, Sociology of; Tradition, Anthropology of; Traditions in Sociology

Bibliography

Arendt H 1951 *The Origins of Totalitarianism*. Harcourt Brace, New York

Asman J 1997 *Moses the Egyptian, The Memory of Egypt in Western Monotheism*. Harvard University Press, Cambridge, MA

Bakhtin M 1968 *Rabelais and His World*. Harvard University Press, Cambridge, MA

Bartlett F C 1932 *Remembering*. Cambridge University Press, Cambridge, UK

Cannadine D 1983 The context, performance and meaning of ritual: The British monarchy and the 'invention of tradition,' c. 1820–1977. In: Hobsbawm E, Ranger T (eds.) *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Caruth C 1996 *Unclaimed Experience: Trauma, Narrative and History*. Johns Hopkins University Press, Baltimore, MD

Connerton P 1989 *How Societies Remember*. Cambridge University Press, Cambridge, UK

Davis N 1977 The abbeys of misrule. In: *Society and Culture in Early Modern France*. Stanford University Press, Stanford, CA

Douglas M 1986 *How Institutions Think*. Syracuse University Press, Syracuse, NY

Durkheim E 1997 *The Division of Labor in Society*. The Free Press, New York

Elias N 1994 *The Civilizing Process*. Blackwell Publishers, New York

Freud S 1913 *Notes on a Case of Obsessional Neuroses*, standard edn. Norton Press, New York, Vol. 10

Freud S 1939 *Moses and Monotheism*, standard edn. Norton Press, New York, Vol. 23

Friedlander S 1991 *When Memory Comes*. Noonday Press, New York

Friedlander S 1992 Trauma, transference, and 'working through' in writing the history of the Shoah. *History and Memory* **4**: 39–59

Hawlbachs M *On Collective Memory*. The University of Chicago Press, Chicago

Hobsbawm E 1983 On inventing traditions. In: Hobsbawm E, Ranger T (eds.) *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Hollander N 1997 *Love in a Time of Hate: Liberation Psychology in Latin America*. Rutgers University Press, New Brunswick, NJ

Kammen M 1995 Some patterns and meanings of memory distortion in American history. In: Schachter D (ed.) *Memory Distortion*. Harvard University Press, Cambridge, MA

Loftus E, Ketcham K 1994 *The Myth of Repressed Memory*. St. Martin's Press, New York

Middleton D, Edwards D 1990 Conversational remembering: A social psychological approach. In: Middleton D, Edwards D (eds.) *Collective Remembering*. Sage Publications, Newbury Park, CA, pp. 23–45

Neal A 1998 *National Trauma and Collective Memory*. M. E. Sharpe, Armonk, NY

Nora P 1989 Between memory and history: Les lieux de mémoire. *Representations* **26**: 7–25

Olick J 1999 Genre memories and memory genres: A dialogical analysis of May 8, 1945. *American Sociological Review* **64**: 391–402

Ozouf M 1991 *Festivals and the French Revolution*. Harvard University Press, Cambridge, MA

Prager J 1998 *Presenting the Past, Psychoanalysis and the Sociology of Misremembering*. Harvard University Press, Cambridge, MA

Schacter D 1995 Memory distortion: History and current status. In: Schacter D (ed.) *Memory Distortion*. Harvard University Press, Cambridge, MA

Schudson M 1989 The present in the past versus the past in the present. *Communication* **11**: 105–13

Schudson M 1992 *Watergate in American Memory: How We Remember, Forget and Reconstruct the Past*

Schudson M 1995 Dynamics of distortion in collective memory. In: Schacter D (ed.) *Memory Distortion*. Harvard University Press, Cambridge, MA

Schwartz B 1996 The expanding past. *Qualitative Sociology I* **9**: 275–82

Shils E 1983 *Tradition*. University of Chicago Press, Chicago

Shils E, Young M 1953 The meaning of the coronation. *Sociological Review* **1**: 5–23

Shotter J 1990 The social construction of remembering and forgetting. In: Middleton D, Edwards D (eds.) *Collective Remembering*. Sage Publications, Newbury Park, CA, pp. 120–38

Turner V 1975 *Drama, Fields and Metaphors: Symbolic Action in Human Society*. Cornell University Press, Ithaca, NY

Van Der Kolk B, Van Der Hart O 1995 The intrusive past: The flexibility of memory and the engraving of trauma. In: Caruth C (ed.) *Trauma, Explorations in Memory*. Johns Hopkins University Press, Baltimore, MD

Wagner-Pacifici R 1996 Memories in the making: Shapes of things that went. *Qualitative Sociology* **19**: 301–321

Zelizer B 1992 *Covering the Body: The Kennedy Assassination, the Media, and the Shaping of Collective Memory*. The University of Chicago Press, Chicago

Zerubavel E 1996 Social memories: Steps to a sociology of the past. *Qualitative Sociology* **19**: 283–300

J. Prager

Collectivism: Cultural Concerns

Collectivism is a cultural pattern found especially in East Asia, Latin America, and Africa. It is usually contrasted with individualism found in the West, e.g., in Western and Northern Europe, the USA, Canada, Australia, and New Zealand. Any typology is an oversimplification. Societies are not purely individualist or collectivist, but some mixture of the two. The construct is widely used by those interested in studying cultural differences.

1. Characteristics

Among the most important characteristics of this cultural pattern are the following:

(a) Individuals define themselves as aspects of a collective, interdependent with some ingroup, such as the family, tribe, co-workers, nation, or religious, political, ideological, economic, or aesthetic group.

(b) They give priority to the goals of that collective rather than to their personal goals.

(c) Their behavior is determined more often by the norms, roles, and goals of their collective than by their personal attitudes, perceived rights, or likes and dislikes.

(d) They stay in relationships even when the costs of staying in these relationships exceed the advantages of remaining.

These are defining attributes of collectivism. There are as many kinds of collectivism as there are collectivist cultures. To distinguish among collectivist cultures, additional attributes are necessary (Triandis 1995).

2. *History*

The term was first used in political philosophy in the eighteenth century in contrast to the individualism advocated by John Locke and Adam Smith. An early exponent was Jean Jacques Rousseau, who in his *Social Contract* (1762) advocated submission to the 'general will,' which was conceived as the common core of societal opinion. Georg Fredrick Hegel and Karl Marx also proposed collectivist philosophic systems.

The term was introduced into psychology by Hofstede (1980), in a study of the values of IBM employees in some 80 countries. He found that by summing the responses to value items of individuals from each country, and using the country as the unit of analysis, it was possible to identify a cluster of responses that reflected collectivism and individualism. Triandis (1995, in press) summarized the literature which used that construct in social psychology, and Kagitcibasi (1997) provided a critical evaluation of the construct.

Historical analysis suggested that collectivism was the typical cultural pattern in traditional societies, and individualism began emerging in England as early as in the thirteenth century. Traces of individualism have also been found in ancient Greece.

3. *Distribution*

Within any society there are individuals who behave like persons in collectivist cultures. They are called *allocentrics*, and contrast with those who behave like persons in individualist cultures who are called *idiocentrics*. It is assumed that there are both allocentrics and idiocentrics in every society, but their distributions are different, with more allocentrics found in collectivist cultures. The distinction between collectivism and allocentrism is important because in some cases a phenomenon observed at the cultural level is quite different from the phenomenon at the individual level of analysis.

Collectivism is maximal in relatively homogeneous societies, such as theocracies and monasteries, while individualism is maximal in heterogeneous societies that are very affluent. Thus, there will be few idiocentrics in monasteries and few allocentrics among Hollywood stars.

All individuals have access to cognitive systems that include both allocentric and idiocentric cognitions but they sample them with different probabilities, depending on the situation. For example, if the ingroup is under attack most individuals become allocentric. In the company of other allocentrics the norms for allocentric behavior become salient and individuals are more likely to sample allocentric cognitions. Some situations provide very clear norms about appropriate behavior (e.g., in a house of worship), while other situations do not (e.g., at a party). Individuals will be more allocentric in the former than in the latter situations. When the ingroup can supervise an individual's behavior norms are more likely to be observed, and the individual will be more allocentric. When individuals are asked to think for two minutes about what they have in common with their family and friends they are likely to sample allocentric cognitions. Situations are more important than the personality of individuals as determinants of allocentric behaviors.

Collectivism can appear in all or none of the domains of social life. For example, it can be found in politics, religion, aesthetics, social life, economics, or philosophy, as was the case in China during the Mao period, or in none of these domains, as among Hollywood stars.

4. *Antecedents*

4.1 *Homogeneity of the Collective*

An important antecedent is the homogeneity of the collective. If members of the collective do not agree concerning what are the group's norms of behavior, it is left up to the individual to decide which normative system to follow.

4.2 *Interdependence Imposed by the Situation*

Collectivism is higher when members of the collective are interdependent, for example when they engage in activities that must be coordinated and when what one does affects the success of others, or where cooperation is essential for success. When the goal is to execute large projects such as irrigation canals, or defensive walls, collectivism is more likely. In societies where people are financially interdependent, collectivism is high.

4.3 *Population Density*

In dense social environments many rules develop that are designed to reduce conflict and ensure the smooth functioning of the group. People make sure that the

rules are observed. When group norms are more important than attitudes as determinants of social behavior, collectivism is high. Urban environments often include members of different groups, and individuals can decide for themselves whether to observe the norms of one or another group, thus they become idiocentric. On the other hand, an urban setting may have neighborhoods of culturally homogeneous individuals, which increases the collectivism of that group.

4.4 Social Stability

When the collective is stable, it is more likely to develop agreements about norms, and to make sure that the norms are observed. By contrast, individuals who have migrated from their societies or have experienced social mobility are more likely to be idiocentric. When social change is slow, the norms can be observed more reliably than when social change is rapid. There is evidence that the older members of all societies are more allocentric than the younger members.

4.5 Low Economic Status

The lower social classes tend to be more conforming to social norms than members of the upper classes. The poor have fewer choices in life, and tend to do what norms specify. When resources are limited, one often depends on group members for assistance, especially in emergencies. These factors increase collectivism. On the other hand, those in positions of leadership tend to be idiocentric.

4.6 One Large Ingroup Rather than Many Ingroups

Those who only have one ingroup can channel all their energy into that group. Also, they cannot afford to develop poor relationships with members of that group, so they are more likely to observe its norms. Those who have many ingroups have superficial relationships with most of them, and if they do not have good relationships with one of these ingroups they drop it and spend more of their energy with some other ingroup. Intense involvement in relatively large ingroups (extended family, religious, or national ingroup) is more common among collectivists than individualists. Intensive involvement in individualist cultures is usually limited to first-degree family members.

4.7 Low Exposure to Other Cultures

People who only know one culture are unlikely to have more than one set of norms for social behavior. They tend to be maximally ethnocentric, authoritarian, and submissive to ingroup authorities. Those who are more educated, traveled, and have lived in more than one cultural group develop idiocentric tendencies.

4.8 Salient Common Fate

Common fate with members of the ingroup (e.g., the ingroup is under attack) increases collectivism. Time pressure for decisions has similar consequences. When there is time pressure an ingroup authority must make quick decisions that do not allow time for discussion, and that places most members of the ingroup in a condition of common fate, as well as requiring conformity to these decisions.

4.9 Collective Mobility is Advantageous

If individual upward social mobility is not possible, then collective mobility may be used. Thus, individuals invest their energy in promoting the status of their ingroup.

4.10 Large Families

In large families it is not practical to allow each child to follow idiosyncratic schedules, or have much privacy. Many rules are enforced, and that creates collectivism.

4.11 Strict Religious Upbringing

Most religions use a large number of norms that are reliably observed. That increases collectivism.

4.12 Little Exposure to Modern Mass Media

US-made television is widely available throughout the world. Content analyses show that the themes used are highly individualistic (emphasis on pleasure, doing what the individual wishes to do even if that is inconsistent with the wishes of authorities). Countries where people have little exposure to the US-made mass media are more collectivist.

5. Consequences

5.1 Socialization

Children are socialized to be obedient, reliable, follow traditions, do their duties, sacrifice for the ingroup, and be conforming. An important value is self-control.

For example, in a Chinese preschool children are taken together to the bathroom only three or four times a day. This procedure is extolled by Chinese teachers as providing orderliness and training in self-control. American educators find it objectionable because it does not allow for individual differences or personal choice, which is an important value among individualists.

5.2 Low Well-being

While members of all cultures are satisfied with their life, members of collectivist cultures are less satisfied than members of individualist cultures. Well-being tends to be correlated with individualism (Diener et al. 1995). Even when controlling statistically for income, this relationship persists.

5.3 Consequences for Cognition, Emotion, and Motivation

Markus and Kitayama (1991) have reviewed an extensive literature that shows that members of collectivist cultures differ from members of individualist cultures in their cognition, emotion, and motivation.

5.4 Social Behavior

Collectivists are extremely concerned with saving their face, and even attempt to save the face of the persons with whom they are interacting. Preserving harmony within the ingroup and being pleasant in relationships with ingroup members are often important values. On the other hand, collectivists do not care much about their relationships with outgroups. They usually behave quite differently toward ingroup and outgroup members, showing sacrifice and extreme cooperation with ingroup members, and suspicion and hostility toward outgroup members.

5.5 Receive Much Social Support if they Conform

Allocentrics report that they receive more social support and a better quality of social support than idiocentrics. However, those who do not conform to ingroup norms are rejected, and may even be killed (Triandis 1995).

5.6 Have Poor Skills in Entering New Groups

Members of collectivist cultures do not have good skills for entering new groups because they are not socialized to do this. They are comfortable as members of their ingroups, thus they do not need to enter others. On the other hand they have more intimate relationships with those with whom they interact than idiocentrics who tend to have superficial relationships (e.g., cocktail parties).

5.7 Are Easy to Mobilize for Confrontations with Outgroups

When ingroups are in conflict it is easier to arouse members of collectivist cultures to fight for the ingroup than it is to do so with members of individualist cultures. Ethnic cleansing is more often found in collectivist than in individualist cultures.

5.8 Describe the Self by Using Roles, Groups

When asked to complete 20 statements that begin with 'I am ...,' members of collectivist cultures provide sentence completions that have more social content (I am a brother, I am a member of the communist party) than members of individualist cultures.

5.9 Attributions

Allocentrics make external attributions in explaining events, while idiocentrics show the opposite pattern. Thus, idiocentrics make the fundamental attribution error (seeing internal rather than external causes for the behavior of other people, while actors see external causes more often than internal causes) more frequently than allocentrics. Also, allocentrics attribute their successes to the help they have received from others and their failures to personal shortcomings (e.g., I did not try hard enough).

5.10 Have the Ideal of Fitting-in Rather than Shining

Allocentrics prefer to fit in and be like most others in their ingroup rather than shining and sticking out. Idiocentrics tend to be high in self-enhancement (report that they are better than most other people on most desirable attributes). Allocentrics tend to be modest when they compare themselves to others.

5.11 Show Little Social Loafing

Allocentrics show little evidence of social loafing (the tendency to produce less when they work in a group than when working alone if their output is not identifiable). They usually produce as much as possible when they work with ingroup members, but they show social loafing if their co-workers are outgroup members.

5.12 Within the Ingroup Distribute Resources Equally or According to Need

When interacting with ingroup members, members of collectivist cultures tend to distribute resources equally or according to need. When interacting with outgroup members, however, they use the equity principle (to each according to contribution).

5.13 Recreation Occurs in Groups

Members of collectivist cultures seek recreation in groups to a greater extent than members of individualist cultures (e.g., bowling rather than skiing).

5.14 Communication

Members of collectivist cultures rely on the context (distance between and position of bodies, eye contact, gestures, tone of voice) more than on the content of communications. For example, rather than saying 'no' they may serve incongruous foods, to put across their negative message without causing the other person to lose face. When they describe a person they are likely to use context, e.g., 'he is intelligent in the market place,' rather than decontextualized statements such as 'he is intelligent.' Silence is used more than in individualist cultures. Compliments are not used as frequently as in individualist cultures. The languages of collectivist cultures do not require the use of 'I.' Collectivists use more action verbs (e.g., he asked for help) rather than state verbs (e.g., he is helpful), which suggests that they use communications that make greater use of context, are more concrete, and place less emphasis on internal attributes of the person.

5.15 Conflict Resolution

Members of collectivist cultures avoid conflict with ingroup members, and in the case of disagreement prefer silence to argument. They prefer mediation to confrontation.

5.16 Conformity

Members of collectivist cultures show more conformity in experiments that measure conformity.

5.17 Emotions

Members of collectivist cultures get angry when something unpleasant happens to ingroup members more often than when something unpleasant happens to themselves. They do not express negative emotions as often as members of individualist cultures.

5.18 Apologize More for Transgressions of the Ingroup

Members of collectivist cultures apologize more when an ingroup member commits a crime than is common in individualist cultures.

6. Globalization and Collectivism

In the process of globalization, is collectivism to be avoided and individualism to be promoted? This issue is especially relevant as societies become more affluent, because they tend to acquire some of the attributes of the West, including individualism. The tribe is rejected, and the individual becomes the center of social change. People in cities, in small apartments that do not have room for extended families, drift toward individualism.

Is 'economic development' necessarily associated with more individualism and less collectivism? My view is that the problem is much more complex than just stated. First, one can be both modern (use computers and fax machines) and collectivist. Second, societies are complex, i.e., they have many aspects. Some aspects might change and become more Western, while most aspects may stay the way they were. In general superficial attributes, dress, public behavior, and economic life might change while fundamental values may not. In any case, for collectivist societies, one might ask: is it desirable to 'promote' individualism, or is it better to stay with collectivism? As is the case with most complex issues, the answer is 'it depends.'

A major problem can be the extremity of collectivism and individualism. Extreme collectivism is stifling. One becomes a tool manipulated by others, unable to reach personal goals. Extreme individualism is isolating, and leads to ego-centered narcissism. Furthermore, extremely individualistic societies are often extremely competitive, and that results in remarkably high levels of stress. Escapes from stress may include drug abuse.

It is possible to list both positive and negative factors associated with each cultural pattern. In the case of individualism we find affluence, self-determination, and self-actualization. These factors are associated with feeling good about oneself, and being satisfied with life and happy. Also, in individualist cultures there is an emphasis on individual rights, and people are expected to shape the direction of societal change to fit their personal needs. Participation in democratic processes is emphasized. Personal freedom is an important value. In democracies the state usually acts so as to satisfy personal values.

On the other hand, Naroll (1983) has assembled much empirical information suggesting that the weak interpersonal bonds that are associated with individualism result in higher rates of civil disturbance,

riots, murder, delinquency, theft and fraud, drug and alcohol abuse, suicide, brawling, wife beating, child battering, and mental illness. The competition found in individualist cultures often results in some segments of the society remaining economically behind and becoming especially susceptible to social pathologies.

By contrast, collectivism is associated with social controls, including extreme punishment for crimes, prompt punishment, and active gossip. Social control is also achieved through emotional closeness, social support, appeals to memorable myths, and a plausible ideology that supports the ingroup's norms. People feel embedded in the social system, so they are less likely to try to change it, to commit crimes or suicide, to steal from the ingroup, or use drugs that isolate them from it. One of the best predictors of drug abuse is the combination of high pressure for achievement with low social support, a combination often found in individualist cultures.

On the negative side, collectivism is often found in societies where groups assume that the individual is relatively unimportant and must do what is best for the group. This is especially significant when the group is the state. People are expected to be 'patriotic,' i.e. sacrifice themselves for the good of the state. However, very often what is good for the state is in fact only good for those in power, who frequently take advantage of the masses to further their personal goals. As we have seen, leaders in all cultures are idiocentric, and the history of the world offers too many examples where they order their followers to do what is good for the leader rather than what is good for the followers. Also, collectivism makes it easier to convince the ingroup to fight outgroups, and wars are more frequent between collectivist countries than between individualist cultures. In a world with nuclear weapons, there may be much to be said in favor of reduced collectivism.

Note also that while the social pathology associated with individualism is formidable, it does not affect most individuals in a society. Pathology by definition is found among a small minority in any society. Thus, the majority can enjoy the benefits of affluence without the negative consequences of individualism.

In short, an evaluation of the desirability of collectivism or individualism depends on what criteria we wish to maximize. If we wish to emphasize the creativity that individualism allows, which often results in more affluence, the freedom and self-actualization of individuals, democratic processes, and the independence of individuals from the social pressures of groups, we may wish to see globalization leading to more individualism. But if we wish to emphasize the values of the ingroup, we may resist the weakening of collectivism associated with affluence. These values include religious beliefs, the social support that one can get from the ingroup in moments of crisis, and the security, both material and emotional, that the ingroup can provide.

In addition to the above considerations, there are issues of individual differences. For those who are intelligent, educated, easily changed, adaptable, an individualistic society can offer a great deal. Creativity requires freedom to pursue different courses of action, and in a society where one is supposed to consult others, and act according to well-established customs, creativity is not likely to flourish. On the other hand, for those who value especially family life, harmony, security, and a life of little stress, a collectivist society may be optimal.

See also: Action, Collective; Collective Behavior, Sociology of; Collective Memory, Psychology of; Individualism versus Collectivism: Philosophical Aspects

Bibliography

Diener E, Diener M, Diener C 1995 Factors predicting the subjective well-being of nations. *Journal of Personality and Social Psychology* **69**: 851–64

Hofstede G 1980 *Culture's Consequences*. Sage, Beverly Hills, CA

Kagitcibasi C 1997 Individualism and collectivism. In: Berry J W, Segall M H, Kagitcibasi C (eds.) *Handbook of Cross-cultural Psychology*, 2nd edn. Allyn & Bacon, Boston, pp. 1–50

Markus H, Kitayama S 1991 Culture and self: implications for cognition, emotion, and motivation. *Psychological Review* **98**: 224–53

Naroll R 1983 *The Moral Order*. Sage, Beverly Hills, CA

Triandis H C 1995 *Individualism and Collectivism*. Westview Press, Boulder, CO

Triandis H C in press Individualism and collectivism. In: Matsumoto D (ed.) *Handbook of Cultural Psychology*. Oxford University Press, New York

H. C. Triandis

Colonialism, Anthropology of

The story of the relationship between colonialism and anthropology (or more specifically, of the anthropology of colonialism) is not the story of a neatly unfolding self-consciousness. It is, rather, the story of successive contexts of questions that have shaped the intellectual space in which the problem of colonialism for anthropology has been formulated and engaged.

Anthropological self-consciousness of colonialism, that is, of colonialism *as a distinctive problem for anthropological knowledge and anthropological practice*, begins roughly in the 1970s. In the early years of that decade two edited volumes appeared dealing with anthropology's relation to the social and political world of which it was a part. One came out of the tradition of US cultural anthropology, *Reinventing*

Anthropology edited by Hymes (1972); and the other out of the tradition of British social anthropology, *Anthropology and the Colonial Encounter* edited by Asad (1973). Both collections perceived the discipline to be in serious crisis; and both were radical in the sense that they were explicitly concerned to subject that crisis to a fundamental self-scrutiny.

Hymes's volume urged that the social and political conditions in which anthropology had emerged as a professional academic discipline in the US at the turn of the twentieth century had, by the late 1960s, significantly altered. First, there was the overwhelming fact that with political decolonization there was a fundamental change in the conditions in which fieldwork could be conceived and undertaken. At the same time, the relation between the US government and the Third World (as the colonized or recently decolonized part of humanity came to be called from the 1950s) had given rise to troubling moral-political questions about the state of the anthropological enterprise. In particular, US involvement in attempted counter-insurgency in Latin America, and the role of anthropologists as cultural experts in the war in Indochina, served to pose the question of the relationship between 'objectivity' of ethnographic data and political interest in a very acute way (see Stocking 1991). What place would anthropology have in raising fundamental questions about injustice and inequality in the world it inhabited, and what was its role in trying to transform that world? Was anthropology to be employed as an instrument of Western domination or as a means of liberation? Moreover, the publication of Malinowski's field diaries in 1967 and the revelation of his barely veiled racism further focussed attention on the ethics of the relation between ethnographer and informant.

In this context, the challenge *Reinventing Anthropology* sought to pose was whether the discipline would be able to transform—to *reinvent*—itself in a way that reflected the demands for political accountability, ethical concern, and critical social commitment on the part of anthropologists. Two strains in US anthropology inspired the volume's contributions: one—reflected especially in the articles by Hymes and Stanley Diamond—was the radical Boasian tradition in the work of people like Edward Sapir, Ruth Benedict, and Paul Radin, who thought of anthropology as a reflexive critique of civilization. The other strain—reflected in Eric Wolf's contribution—was a post World War II Marxism (much influenced by the cultural-evolutionary materialism of Julian Steward) which was concerned to highlight the plight of peasant societies caught in the juggernaut of capitalist imperialism.

The intellectual tradition of British social anthropology has of course a very different historical context than US cultural anthropology, one far more intimately interconnected with the subordination, management and rule of colonial subjects. Not surprisingly, then, the concerns of Asad's *Anthropology and the Colonial Encounter* were somewhat differently inflected than those of Hymes's volume. To be sure, they shared a concern that since World War II fundamental changes had occurred in the world inhabited by social anthropology. The emergence of new nations, especially in Africa where much British social anthropological fieldwork was carried out—Sudan (1956), Ghana (1957), Nigeria (1960)—made it increasingly difficult to avoid the pertinence of a more historical approach that brought the larger political and economic system of colonialism into view. And in consequence it was increasingly difficult not to be critical (as indeed Edmund Leach was in a limited way) of the ahistorical functionalism and empiricism that characterized British anthropological theorizing.

In the middle 1960s, New Left critics had in fact launched an attack on British social anthropology. And part of the project of Asad's volume was to respond to these and other radical claims, not simply to affirm or deny them, but to more adequately specify and formulate them. There was a sense that perhaps, while undeniable, the relation between colonialism and social anthropology was more complicated than was allowed by the polemical tone of accusation and disavowal. In her contribution, for example, Wendy James argued that the very existence of social anthropology in the colonial period constituted a source of *potential* radical criticism of the colonial order, and she sought to provide a more nuanced account of Malinowski's own active social reformism during the interwar years. Stephan Feuchtwang, in his contribution, discussed the relation between the social and institutional conditions of British anthropology and the knowledges it produced. In an allied way, Asad examined 'objectifications' of colonial power produced in two authoritative domains of systematic knowledge—the functionalist anthropology of Africa and Islamic orientalism. Asad was concerned to understand the connections between these different images of rule (one largely of 'consent,' the other largely of 'repression'), their ideological roots, and conceptual consequences.

In subsequent decades, the anthropological preoccupation with colonialism has tended to be largely (if not solely) a phenomenon of the US academy. At any rate, in the 1980s there was a gradual shift away from the concern with the *role* of the discipline in the colonial/imperial enterprise and the politics of its practitioners, toward a concern with colonial forms of knowledge about non-European worlds. The register in which the problem of colonialism for anthropology was conceived was now less a political than an epistemological one. Or rather, linking power and knowledge, colonialism was now conceived as a problem of (in the language of the day) the politics of representation. This shift has many interconnected sources. I shall only be able to touch on a few of the more salient ones.

At one level, of course, the shift is connected to the decline of anti-imperialism and anticolonialism as going *politico-ideological* concerns. On the one hand, beginning with the end of the Vietnam War and accelerating through the years of Ronald Reagan's US Presidency there is a marked decline in the First World intelligentsia's interest in revolutionary change in the Third World. On the other hand, in the Third World itself there is a decline (indeed a rapid unraveling) of the experiments with radical social transformation in many of the independent countries of South Asia, Africa, the Caribbean, and Latin America. At another level this shift was connected to a number of *intellectual* transformations in the understanding and practice of the human sciences. Very broadly speaking, these may be characterized in terms of a general drift away from the determinism (Leslie White, Julian Steward) and functionalism (Radcliffe-Brown) of postwar anthropology and toward a social/cultural constructionism. Principal among these transformations was the so-called 'turn to meaning' in social and cultural analysis. In many ways, of course, the work of Geertz (1973) embodied—indeed, helped to constitute—this constructionist moment. In his work, culture came to be understood as a web of semiotic processes or relations, and cultural analysis as a practice less of uncovering determinate causes, or of explaining functions, as of deciphering meanings. Two features of this anthropology are of special importance here: one is that its criticism of objectivism brought about a renewed sensitivity to 'the native's point of view' and 'local knowledges'; and the other is that the critique of positivist social science led to a greater openness to interdisciplinarity, to literature and history in particular (Geertz's famous 'blurring of genres').

This work, in turn, set the stage for the emergence in the 1980s of an interest in the *textuality* of ethnographic description and analysis, in the senses in which these are artifacts of language—hence, in their way, *literary*—and therefore open to an investigation of their rhetorical and poetic structures (Clifford and Marcus 1986). How is ethnographic authority constructed? What are the linguistic devices that enable anthropological representations to produce their distinctive 'realistic' effects? Some of this work sought explicitly to reconceive the task of a critical anthropology (Marcus and Fischer 1986).

This preoccupation with modes of representation in social and cultural analysis was, in part at least, the context for the rise of what would come to be called colonial studies. An interdisciplinary field (drawing largely on literary critics, anthropologists, and historians), colonial studies would help to institutionalize the renewed scholarly focus on colonialism. The central inspiring text here was Said's *Orientalism* (1978). Borrowing in some measure from the middle work of Foucault (1979), this book thematized the idea of colonialism as a *discursive* formation. Said set out to demonstrate the ways in which, in the (especially French and British) colonial discourse of orientalism, the Orient was invented. The Orient, therefore, was less a geo-political location than a figure in the cultural-political imagination of the West. *Orientalism* opened up a rich vein of research and a critical demand to interrogate the West's authoritative discourses on the non-West—anthropology chief among them—and to unmask the ways in which they produced and reproduced their hegemonic knowledges. Interestingly Said himself (1988) would later express doubts about the value and future of a discipline so deeply implicated in the making of such hegemonic knowledge.

One anthropological work that drew together some of these themes—though in a way less pessimistic about the possibility of a revised anthropological enterprise—was Fabian's *Time and the Other* (1983). The subtitle of the work was significant: *How Anthropology Makes its Object*. It focussed attention precisely on the 'constructedness' of anthropological discourse, specifically on the uses of Time in anthropology's constitution of its object (the savage, the primitive, the native, the Other). Fabian sought to show how the existential copresence of the Other in the dialogical interaction of ethnographic fieldwork gets transformed into a temporal distancing—what he called a denial of coevalness or cotemporality—in the theoretical construction of authoritative anthropological knowledge. This 'conjuring trick' served perpetually to consign the native to a temporal zone outside of the Time of anthropology, which is to say the privileged Time of the West. It is easy to see that the lessons of Fabian's book are wider than the issue of temporality alone. What it demonstrated was that unless anthropology subjects its own discourse to critical scrutiny, unless it pays attention to the conceptual assumptions that go into the *making* of its objects, it is liable to reproduce colonial knowledge. Other work would later seek to extend and develop this inquiry into the formation of anthropological objects (see Scott 1994).

Another set of anthropological concerns with colonialism in the 1980s was more closely connected to the antinomies of Marxism. The defeat of the communist project in Western Europe after 1920 had given rise to a strain of Marxism (soon to be called Western Marxism) focussed on what had hitherto been dismissed as superstructural issues. Marginalized until the rise of the New Left in the 1960s, this strain became more prominent in the 1970s, especially in the North Atlantic academy. Antonio Gramsci's theorization of hegemony, Louis Althusser's reconceptualization of ideology as representation, and Raymond Williams's displacement of the base/superstructure model for the analysis of literature (and cultural production more generally) offered less reductionistic critical languages in which to situate social and cultural life in relation to capitalism. An interest in rethinking the problem of the articulation of modes of production, and of

precapitalist (and precolonial) formations was one direction of Marxist work that had actually begun earlier. But in another direction, folded into the anthropological concern with local knowledges and finite histories, this cultural Marxism offered the prospect of an antiuniversalistic account of hitherto invisible idioms of resistance to dominant power (including, of course, globalizing capital), and of hitherto unacknowledged forms of (often informal) agency in the quotidian making of histories. In their very titles, James Scott's *Weapons of the Weak* (1985), and Jean Comaroff's *Body of Power, Spirit of Resistance* (1985), offer exemplification of this interest in the anthropology of anticolonial resistance.

Throughout the 1980s history was becoming a site of profound scholarly preoccupation. Indeed within the discipline of history itself there was already a growing interest in the intersection of modes of cultural and historical representation. Work in feminist history, in particular, helped to reformulate relations between power, difference, and representation. But perhaps the most significant contribution to the argument for more than a nodding relationship between anthropology and history—and this too on the terrain of colonialism—has been the pioneering work of Cohn (1987). Cohn's essays were only brought together in 1987, under the title, *An Anthropologist among the Historians and Other Essays*, but from the early 1960s he had been urging the theorization of the intersection of history and anthropology. In a number of wide-ranging essays on British India—from concerns with the formation of historical, legal, and statistical modes of colonial knowledge, to an interest in ritualized representations of political authority—Cohn demonstrated that historical analysis of colonial rule was indispensable for anthropological understanding. This work inspired a number of innovative studies in the historical anthropology of colonial India (among them Dirks 1987). Significantly too, India was the conceptual and intellectual-political site of another important contribution to the debate about history for an anthropology of colonialism—the Subaltern Studies Collective's re-interrogation of liberal and nationalist accounts of colonial rule in India and the anti-colonial struggle against it (Guha 1984). Transdisciplinary in perspective, the work of the Collective resonated with the anthropological discontent with universalism and Eurocentrism and its interest in the agency of non-dominant idioms, knowledges, and practices.

In the 1990s one can discern another shift in the questions driving the anthropology of colonialism. These new questions turn less on the textuality of colonial representation or the agency of resistance than on the problematization of the relationship between colonialism and modernity, and the implications of this relationship for understanding the postcolonial present (Asad 1991, 1992, Scott 1997). The profoundly epochal sense attending the *fin de siecle*, the sense of a crisis of modernity (and for some, an emerging *post*modernity), has made for a growing preoccupation with a re-examination of the formation of the modern world and its moral, epistemic and political modes of legitimization. Much of this re-historicizing work has sought to criticize the progressivist story of modernization.

There are several dimensions to this anthropological preoccupation with colonialism and modernity. One of the earliest has been the concern with the problem of nation and nationalism. There had of course been an older multidisciplinary interest in nations and nationalisms, namely, the 'new nations' projects of the early 1960s. These studies were primarily concerned with inquiring into the modernizing and democratizing challenges facing the newly independent nation-states. However, two books more than any others reopened and significantly reshaped the debate about the nation: Gellner's *Nations and Nationalism* (1983), and Anderson's *Imagined Communities* (1983). Each was focussed differently, but they were both inspired by the problem of the relation between nation/nationalism and *modernity*. If the 'new nations' work tended teleologically to take nationhood as a normal developmental threshold, what Gellner and Anderson foregrounded was the idea that nations were not natural entities nor nationalisms primordial sentiments or ideologies. Both, they argued, were profoundly historical, and that historical character, moreover, owed most to the making of the *modern* world.

Gellner and Anderson would both be criticized for their emphasis on the modernity of nations and nationalisms, but it was precisely this connection that those interested in colonialism were concerned to explore. Anticolonial nationalisms and the making of postcolonial sovereignties clearly had a very complex—and as yet, perhaps, unresolved—relation to the modernizing project of late colonialism. In this regard the decisive work was Chatterjee's seminal book, *Nationalism and the Colonial World*, originally published in 1986 but becoming influential in debates from the early 1990s. A noticeable feature of this work was its interdisciplinarity, and especially the place of anthropology (as the authoritative discourse on Europe's Others) within it. While indebted to Said's *Orientalism* for its opening deconstructive move, Chatterjee undertook a sustained examination of anthropological assumptions (e.g., about rationality) undergirding the distinction between Western and non-Western nationalisms.

Connected to this focus on rethinking the relations between modernity and nation, has been the renewed attention given to the problem of modern colonial power generally, and the colonial state in particular. Especially important here has been the later work of Foucault on governmentality (1991) that raised the question of the place of the state in a larger and much transformed understanding of power. Questioning a

deeply influential tradition of social and political analysis owing much to both Marx and Weber, Foucault's concept of governmentality sought to think (conceptually) about a field of power in which the state was not the central analytical domain; and (historically) about a concrete moment in European history when political power becomes more dispersed, less dependent upon the institutions of coercion. This concept opened up the possibility of reconceiving the questions through which the problem of colonial power was formulated. Where modernization narratives (liberal or Marxist) had plotted a story of the progressive rationalization of the state (the more or less of power; the counter-positioning of power and freedom), what now emerged was the possibility of a story of colonial power told in terms of the political rationalities by which it was organized. The point here was to urge an historicization in which the emergence of a colonial rationality of government (i.e., a form of colonial power directed at transforming conditions so as to produce effects on colonial conduct) could come more sharply into view. And the significance of this is that this rationality is central to the making of our postcolonial worlds (see Scott 1999).

There have been critics, of course, who are skeptical about this thematization of governmentality in the anthropology of colonialism. The more sympathetic of them argue that neither the old master-narrative of progress and modernization nor the new Foucauldian revisionism are adequate because neither appreciates the inherent *doubleness* of colonial rule. The essential paradox of the colonial state, it is argued, is that it was always caught between *two* principal imperatives—ruling subjects and making citizens—and, therefore, always moved simultaneously in two directions: coercion and civilization. The crucial hinge on which this doubleness hung was the colonial language of law and legality (Comaroff 1998). In a related direction some critics have been concerned that the whole relation between metropole and colony has been ill-conceived (Cooper and Stoler 1997). The worry here is the assumption among many students of empire of a Manichaeism that sees the problem of colonialism as a binary—colonizer/colonized—one. These critics argue that, to the contrary, the conception of an encounter between an empty colony simply remade by Europe and a Europe that remains ever unmarked by its empire is untenable. They urge a view that places metropole and colony in a single (if internally differentiated) analytical field, and focuses precisely on the multiple tensions—inclusions/exclusions, hierarchies, reversals, etc.—through which Europe and its empire were (unevenly and asymmetrically) co-constructed.

Since its emergence as a distinctive theme in anthropology in the 1970s, the problematization of colonialism has been thought of as part of an orientation towards the present at least as much as, if not more than, the past. In this sense there has always been about it an animating ethos of criticism—whether criticism of our worlds or criticism of our disciplines. The anthropological problem about the colonial past, in other words, has been connected to raising critical questions about the postcolonial present: it is part of the making of a distinctively *postcolonial* anthropology. However this connection and its implications have been more often assumed than systematically thought out. Consequently, anthropology has been less concerned than it might, not merely about the ways in which understandings of the colonial past are conditioned by contemporary 'scenes of inquiry,' but more importantly about the ways in which a self-consciousness about alterations in the latter ought *strategically* to alter the kinds of questions we pursue about the nature of the former. The point, then, is that for anthropological criticism to have purchase on the present it is not enough to insist that the purview of colonial studies be enlarged (to encompass Europe, for instance, or gender, or the state), it is necessary to specify the domain of contemporary questions that are being addressed in so doing. This is the only way to gauge the extent to which our critical exercises are meeting a contemporary demand as opposed to simply expanding an already established cognitive field.

This is the challenge. It is an important challenge because the crisis upon which the formerly colonized world entered in the 1980s and 1990s (unlike those of the 1960s and 1970s) is a crisis not merely of policy orientations or development options or even ideological alternatives. It is a crisis of the legitimacy of the modern/modernizing assumptions that have undergirded the very making of these colonial and postcolonial societies and polities (assumptions, for example, about the relation between culture and politics; between nation and sovereignty; between individual rights and historical traditions, and so on). Our *conjuncture*, in short, is shifting. And therefore, if it is to more adequately theorize the relation between the problem–space in which it produces knowledge and the questions through which it does so, the anthropology of colonialism has to attend more systematically to the changing contours and character of this postcolonial present and its crisis.

See also: Colonialism: Political Aspects; Colonization and Colonialism, History of; Racism, History of

Bibliography

Anderson B 1983 *Imagined Communities*. Verso, New York

Asad T (ed.) 1973 *Anthropology and the Colonial Encounter*. Humanities Press, New York

Asad T 1991 From the history of colonial anthropology to the anthropology of western hegemony. In: Stocking Jr G W (ed.) *Colonial Situations: Essays in the Contextualization of Ethnographic Knowledge*. University of Wisconsin Press, Madison, WI

Asad T 1992 Conscripts of western civilization. In: Gailey C (ed.) *Dialectical Anthropology: Essays in Honor of Stanley Diamond*. University Press of Florida, Gainesville, FL
Chatterjee P 1986 *Nationalism and the Colonial World: A Derivative Discourse?* Zed, London
Clifford J, Marcus G E (eds.) 1986 *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA
Cohn B S 1987 *An Anthropologist among the Historians and Other Essays*. Oxford University Press, Delhi
Comaroff Jean 1985 *Body of Power, Spirit of Resistance*. University of Chicago Press, Chicago
Comaroff John 1998 Reflections on the colonial state, in South Africa and elsewhere: Factions, fragments, facts and fictions. *Social Identities* **4**(3): 321–61
Cooper F, Stoler A L (eds.) 1997 *Tensions of Empire: Colonial Cultures in a Bourgeois World*. University of California Press, Berkeley, CA
Dirks N B 1987 *The Hollow Crown: Ethnohistory of an Indian Kingdom*. Cambridge University Press, Cambridge, UK
Fabian J 1983 *Time and the Other: How Anthropology Makes its Objects*. Columbia University Press, New York
Foucault M 1979 *Discipline and Punish* (trans. Sheridan A). Vintage, New York
Foucault M 1991 Governmentality. In: Burchell G, Gordon C, Miller P (eds.) *The Foucault Effect: Studies in Governmentality*. University of Chicago Press, Chicago
Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York
Gellner E 1983 *Nations and Nationalism*. Blackwell, Oxford, UK
Guha R (ed.) 1984 *Subaltern Studies*. Oxford University Press, Delhi, Vol. 1
Hymes D (ed.) 1972 *Reinventing Anthropology*, 1st edn. Pantheon, New York
Marcus G E, Fischer M J 1986 *Anthropology as Cultural Critique*. University of Chicago Press, Chicago
Said E W 1978 *Orientalism*, 1st edn. Pantheon, New York
Said E W 1988 Representing the colonized. *Critical Inquiry* **15**: 205–25
Scott D 1994 *Formations of Ritual: Colonial and Anthropological Discourses on the Sinhala Yaktovil*. University of Minnesota Press, Minneapolis, MI
Scott D 1997 Colonialism. *International Social Science Journal* **154**(December): 517–26
Scott D 1999 *Refashioning Futures: Criticism after Postcoloniality*. Princeton University Press, Princeton, NJ
Scott J C 1985 *Weapons of the Weak: Everyday Forms of Peasant Resistance*. Yale University Press, New Haven, CT
Stocking Jr G W 1991 Colonial situations. In: Stocking Jr G W (ed.) *Colonial Situations: Essays in the Contextualization of Ethnographic Knowledge*. University of Wisconsin Press, Madison, WI

D. Scott

Colonialism: Political Aspects

Colonialism in the English language originally indicated a practice or idiom associated with British colonies (e.g., the phrase 'the place was going ahead' was described in 1887 as a 'colonialism') and was only recently applied to the relationships between metropolitan states and their colonial possessions. Its modern usage dates from the 1950s, although the relationships it embraces have a very long history indeed.

The word is derived from the Latin *colonia*, which was a synonym for the Greek ἀποικία, meaning a settlement of people from home, and it should, therefore, properly apply to the process of establishing and governing such settlements. It is now closely linked, and often interchangeable, with imperialism (e.g., see *Imperialism: Political Aspects*), but it is a particular form of the latter. The ideas that the term embraces have shifted in line with the shift in the idea of a colony.

1. The Classic Colony

The classical Greek colony was a new community, originating from an existing πόλις, and, apart from the Corinthian colonies, was independent, although tied to the mother country by sentimental and religious affinities. The high era of Greek colonization lasted from the middle of the eighth century BC until the early sixth century BC, by when Greeks had 'colonized' much of southern Italy, Sicily, and the Black Sea littoral. The motives were a mixture of land shortage and traded expansion (Boardman 1980).

Roman colonies generally had greater strategic purposes. They were settlements of Roman citizens, normally veterans, who garrisoned a new town in a hostile country and thus extended Rome's *imperium* or rule. Similar expansionist activities occurred in many other parts of the world, as with the Mogul and Ottoman Empires or the Vijayanagara Empire in southern India.

At the beginning of the sixteenth century, European powers extended their physical occupation of land to settle in the Americas, later in Australasia and parts of Africa. Colonialism involved settlement in areas noncontiguous to the motherland and should be carefully distinguished from conquest *per se* which involves physical incorporation into the parent state. Settlers left their homes for many reasons. Some fled religious and other forms of persecution; some were risk-takers, seeking opportunities to better themselves; some were sent abroad by families to resolve domestic difficulties or by the state to penal colonies.

Governing such distant possessions has always been problematic. Initially, the local, or native, populations were subdued forcibly or persuaded to sign agreements, and administrative authority passed to the new immigrants and their successors. Financial and military strength ensured metropolitan dominance at first. Over time, as the colonies grew in population and economic strength, their immigrant inhabitants sought a much greater degree of self-governance. Initially, the European powers opposed such aspirations and anticolonial wars took place that ultimately wrested much

of South America from Spanish and Portuguese control and part of North America from British control.

The American War of Independence, which gave birth to United States of America composed of 13 ex-colonies, persuaded British governments to treat more sympathetically the demands of their surviving colonies. Following Lord Durham's 1839 Report, the North America Act granted Canada considerable autonomy. In 1867, Canada became a Dominion, a term adopted by the major self-governing British colonies in 1907 and sought unsuccessfully by others, such as Southern Rhodesia. Genuine devolution of powers kept Australia and, after the Anglo-Boer War, South Africa within the British Empire. These countries still remained constitutionally subject to the British Parliament and it was not until the Statute of Westminster in 1930 that they acquired sovereignty.

The metropolitan states were largely successful in twentieth century in managing the transition of their possessions from an inferior 'colonial' status to an independent state attached emotionally to the motherland. The British failed in Ireland as did the French in Algeria. Australia, Canada, and New Zealand retained the monarch of the United Kingdom as their head of state, although vocal groups challenged this relationship in the latter part of the twentieth century. In South Africa, where the original Dutch immigrants (the Afrikaners) were dominant, this link was broken in 1961.

Colonies, in the sense of territories largely peopled, and certainly controlled, by descendants from the motherland, developed political cultures of their own, closely related to the specific circumstances that generated the initial settlements (Hartz 1964).

2. Modern Colonialism

Modern colonialism was not characterized by settlements but by external control. As a practice it dates, with a few exceptions (e.g., the Indian subcontinent), from the end of the nineteenth century, when the major European powers laid claim to, and attempted to rule over, enormous tracts of the globe—mostly in Africa and Asia—without establishing permanent settlements in them. In the twentieth century, Japan, Russia, and the USA joined the major European powers in such colonial activity. The term thus came to refer to a condition of unequal relations in which a strong 'colonial' state controls, and usually exploits, an alien and weaker people in the latter's 'homeland.' The senior administrators were nonresidents.

While the practice has a long pedigree, the word itself, with its pejorative overtones, only appeared in the aftermath of the World War II. President Harry S. Truman used the word 'imperialism' in 1949 to cover all forms of external control over indigenous inhabitants, but by 1955, when the first volume of his *Autobiography* was published, he used the term colonialism to describe such relations. At this time, colonialism was limited to the practices of the Western European powers, the expansionist policies of the Soviet Union and China still being subsumed under the umbrella of imperialism.

This was short-lived. The term 'colonial power,' referring to any state dominating another, was soon extended to the Soviet Union and the USA, and even in Pakistan, to India where Kashmir was concerned.

The 1919 *Treaty of Versailles* generated the League of Nations, which introduced a new type of colony, by granting to certain imperial powers (and Dominions like Australia and South Africa) mandates to rule over what had been the German and Italian empires. In theory, the recipient countries were obliged to ensure the paramountcy of 'native interests.' With the creation of the United Nations, mandated territories became trust territories and more attention was given, through its standing Trusteeship Council, to the responsibilities of the colonial powers. By the 1950s, 'native interests' (except in South West Africa where South Africa was the trustee) came to embrace the right to self-determination and independence.

The major colonial powers, Britain and France, organized their relations with their settler colonies, their modern colonies, and other states through different departments or offices (e.g., the Dominions Office or the India Office in Britain) to reflect their distinct relationships.

3. Colonialism in Practice

3.1 Administration

Ruling modern colonies presented difficult problems. Metropolitan states had neither had personnel nor the resources to administer them closely. Occasionally they relied upon chartered companies to act on their behalf. Some parts were intensively controlled; much was ignored altogether for many years; most were administered through the acquiescence of local leaders. This British principle of indirect rule, with roots in the Indian experience, was ideologically based (Lugard 1922) as well as pragmatically required. Other European powers emphasized direct rule. In fact, all powers practised both models as needs demanded.

The debate on the nature of colonial political rule continues. While the exercise of power was crude and excessive on occasions, recent scholarship has emphasized how local peoples sometimes took the initiative in crafting political strategies that limited colonial opportunities. After World War I, colonial administrators lacked the military support or the will to impose unwelcome policies on subjects accepted the paternalistic assumptions of trusteeship (Robinson 1965) but worked, in the British empire, towards economic and political development (Kirk-Greene

2000) and, in the French, towards assimilation. The idea that colonialism was unacceptable arose with the growth of democracy in the metropolitan states and the changing intellectual climate after World War II.

3.2 Anticolonialism

Throughout the modern colonial period some individuals and groups campaigned for independence, most notably in India, but the real pressure to end colonial rule came immediately after 1945. This reflected a number of seismic shifts in the context of international affairs. The arrival of internal democracy in the colonial states themselves challenged the undemocratic politics of the colonies; the battle against Nazism virtually destroyed the legitimacy of a racial hierarchy; both the Soviet Union and the USA, linked by the bipolar confrontation of the Cold War, yet representing quite different ideologies, shared an antipathy to overt colonial empires; the financial and military costs of retaining overseas possessions became politically prohibitive in the United Kingdom and the Netherlands.

Most colonies obtained a negotiated independence peacefully. Violence—both civil and military—was sometimes apparent, most especially in the Portuguese empire and where white settlers (classical colonists) were entrenched. The logic of colonialism worked itself out as the motives for occupation disappeared and the costs of continued rule escalated. The British colonies became independent seriatim. With their emphasis on retaining the initiative in the 'transfer of power,' officials sought the kind of educated professionals who were calling for independence (on India, see Mansergh and Moon 1970–83). In contrast, the French absorbed that elite into metropolitan organisations until the logic dictated a rapid and almost simultaneous decolonisation in 1961 (Mortimer 1969). Colonialism had effectively come to an end by the middle of the 1960s.

However, postmodernists such as Edward Said have argued that colonialism was neither a finite era nor a particular institutional structure of superordination and subordination. Both the departing colonialists and the triumphant nationalists failed to appreciate the enduring consequences of colonial rule. Cultures, economies, and institutions were so deeply embedded in the new states that their futures remained dependent upon the heritage of colonial overrule (Said 1978).

3.3 Neocolonialism

Political independence, however, was not matched by full local control. At the Bandung Conference of nonaligned states in April 1955, the Indonesian president, Ahmed Sukarno, claimed that colonialism 'has also its modern dress, in the form of economic control, intellectual control…by a small but alien community within a nation.' This perspective, similar to Said's, gave birth to the idea of 'neocolonialism' and fitted well with the growing strength of Marxist scholarship in the 1960s and 1970s.

The word itself probably first appeared in 1961 in the pages of the radical journals *The New Statesman* and *The New Left review*, but it was soon common usage in most European languages. Lord Home, when British Foreign Secretary, attempted at the United Nations in October 1961 to answer 'the charge of neocolonialism.' With the end of formal control over most of the empires, scholars and activists still perceived a decisive degree of control exercized by the metropolitan powers, essentially through economic, technological, and cultural dominance (Crozier 1964). Francophone scholars spoke of the Philippines as being '*entièrement néocolonisés par les Etats Unis*.'

Charges of neocolonialism were not limited to the great colonial powers but were extended to any state from the rich and developed world exerting influence over the policies of poor and less developed countries. Several forms have been identified. Political, economic, and cultural forces have been said so to constrain the choices available to such countries that they are determined by the interests of the stronger states. Further, it has been argued that the rapid decolonization was a calculated policy to move from overt colonial rule to indirect control through political superiority in the international community, economic dominance, and cultural influences. While the evidence for such a conscious policy is weak, the poorer excolonies were undoubtedly constrained by their political, economic, and cultural inheritances. Some states responded more robustly and effectively than others to their weaknesses.

3.4 Internal Colonialism

There has been a further recent refinement of the term. Whereas it had been used to describe a particular relationship between countries (originally metropolitan states and colonies, later sovereign states), it is sometimes used to describe broadly similar processes at work within a single state. Thus, particular groups, through their dominance of political and economic power, ensured that other groups are kept in long-term subservience. It has been used in the context of East Bengal in Pakistan and in South Africa, where the 'colonial' people were the white minority and the 'subject' people the black majority.

4. Conclusion

Colonialism is now normally used in a pejorative sense and is associated with crude exploitation. Few would deny the reality at times of oppression, economic exploitation and an unconcern for human and civil rights. The colonial powers' primary interests were

usually selfish and largely economic. But alongside was a genuine commitment to the principles of trusteeship and paternal development. Attempts to evaluate the costs and benefits of colonialism coincided with its formal ending (Perham 1962), but a growing revisionist literature, in part reflecting the reduced status of Marxist scholarship, is emphasizing the advantages for the modern state of the enlargement of scale, modern educational and economic practices, and the opportunities provided for some (but not all) by integration into the world economy. Nevertheless, it is equally important to acknowledge the relative short period of colonial rule. Davidson (1964) was right to observe: 'Looking back, one may see now that the colonial period was no more than a large episode in the onward movement of…life; in another sense, it was an unexampled means of revolutionary change.'

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Fourth World; Hegemony: Anthropological Aspects; Imperialism, History of; Imperialism: Political Aspects; Nationalism: General; Nationalism, Sociology of; Postcoloniality; Third World

Bibliography

Boardman J 1980 *The Greeks Overseas: Their Early Colonies and Trade*, 3rd edn. Thames and Hudson, London

Crozier B 1964 *Neo-colonialism*. Dufour Editions, Bodley Head, London

Davidson B 1964 *The African Past: Chronicles from Antiquity to Modern Times*. Little Brown, Boston

Hartz L 1964 *The Founding of New Societies*. Harcourt, Brace and World, New York

Kirk-Greene A H M 2000 *Britain's Imperial Administrators 1858–1966*. St. Martins, New York

Lugard F 1922 *The Dual Mandate in British Tropical Africa*. Edinburgh, UK

Mansergh N, Moon P (eds.) 1970–1983 *The Transfer of Power*. Her Majesty's Stationery Office, London

Mortimer E 1969 *France and the Africans*. Faber, London

Perham M 1962 *The Colonial Reckoning*, 1st edn. Knopf, New York

Robinson K 1965 *The Dilemmas of Trusteeship*. Oxford University Press, London

Said E W 1978 *Orientalism*. Pantheon, New York

R. Hodder-Williams

Colonization and Colonialism, History of

1. Introduction: Problems of Definition and the Semantic Field

'Colonization is [...] a phenomenon of colossal vagueness' (Osterhammel 1997, p. 4), because it covers large and rather different parts of the world and its history. Therefore, a convincing theory of colonization and colonialism in general or of European colonialism in particular does not exist and will probably never exist, because it is not feasible. But scholars have still to define the concepts they use, because an agreement upon their meaning makes communication possible. *Colonization* has to do with *migration*, because it describes the movement of people from one part of the world to another to establish a settlement, quite often an agrarian one. In this sense, the term has a neutral connotation. In contrast, *colonialism* has become a general invective against western policy, especially since the Bandung conference of recently decolonized Asian countries in 1955. In the nineteenth century, however, it was used more or less neutrally to characterize the condition of colonies and the (speech) habits of colonials (Fieldhouse 1981, p. 6).

We have no choice but to accept the change of meaning that *colonialism* has undergone, though we can try to neutralize political emotions. In this sense, colonialism can be defined as the control of one people by another, culturally different one, an unequal relationship which exploits differences of economic, political, and ideological development between the two (Reinhard 1996, p. 1). *People* instead of *nation* or *state* is used because basically no sophisticated political organization is necessary on either side. And the terms *difference* and *development* are used in a strictly descriptive sense and without any value judgement. They do not suggest that it is more desirable to have nuclear weapons instead of bow and arrow or that there exists a general and 'normal' path of human development with the West at the end. But differences of development are essential to distinguish colonial rule from *empire* in general. Roman rule over ancient Greece and Russian control of East Germany were imperial, not colonial.

Colonialism, as an unequal relationship between human groups, very often was the outcome of *imperialism*, defined as a political activity with the intention to establish colonialism. But if the meaning of *imperialism* is limited to expansive policy of the nineteenth and twentieth centuries, *colonial expansion* has to be used to designate earlier policies. After *decolonization*, colonialism as a form of political dependency has become a mere phenomenon of history. But economic and cultural domination by former colonial powers such as Britain and France on the one hand, and by the US and the US controlled world economy on the other, still exists. It has been labeled *neocolonialism*. In the 1960s and 1970s theories of structural *dependency* explained economic underdevelopment as a consequence of western economic domination and as self-reproducing without chance of escape because of western control of the world economy. Several success stories of former colonies have falsified these theories, but their lesson on informal control as an element of colonialism remains. *Semicolonies* such as China or the Ottoman Empire about 1900 were formally in-

dependent, but in reality under economic and therefore indirect political control of the West. Western colonies in the Americas, in South Africa, and Australia established a *secondary colonialism* in their parts of the world. Even in European countries unbalanced internal group relations as in the case of Britain's 'celtic fringe' were qualified as *internal colonialism* (Hind 1984).

2. *The History and Geography of Colonization and Colonialism*

The expansion of Europe between the fifteenth and twentieth centuries was one of the most momentous processes of history. World history in the sense of global history became possible only through that process, because no part of the world escaped the direct or indirect impact of Europe. In 1914 half of the land surface of the earth and one-third of its population were still under direct colonial rule. But this was only the last chapter of a long story, because colonization and colonialism are essential components of history in general, especially, if we do not omit continental expansion in Russia, the Americas, and elsewhere in favor of overseas colonialism. Colonization in the sense of expansion of settlement and agrarian land use could even be considered the quintessence of human history before industrialization. Chinese colonialism is perhaps the most remarkable case because of the relative continuity of Han-Chinese expansion starting from the lower Huang basin thousands of years ago and leading to the present penetration of Tibet, Mongolia, and Xinjiang. The ancient world of the Mediterranean was full of Phoenician, Greek, and Roman settlement colonies. Colonial phenomena of the European Middle Ages became predecessors and models of the later expansion of Europe. Spain and Portugal colonized the reconquered South of the Iberian peninsula and the Canaries, whereas England did the same in Wales and Ireland, Old World experiences to be used in the New World later. People from central and western Europe settled in the East, a process to be continued by Russia in Siberia until the Pacific Ocean was reached. The French who were excluded from the colonization of the East became protagonists of the crusades instead. The first European colonies established by the crusaders in the Middle East were a failure, but in that context Italians developed a system of maritime trade and factories combined with the colonial production of high value commodities such as sugar. Not by chance America was 'discovered' by a Genoese and 'baptized' after a Florentine.

Beginning in 1415 the Portuguese followed the coast of Africa until they found their way to India in 1498. To control the spice trade of the Indian Ocean they established a trade empire based on fortified ports, factories, and colonial rule over spice producers such as Ceylon and the Moluccas, extending from East Africa to China and Japan, with Goa in the center. During the European conflicts of the early seventeenth century, the Dutch East India Company, in contrast to Portuguese crown capitalism, a privileged private corporation of shareholders, conquered most of the Portuguese system, but established a new capital in Batavia on Java. Arriving at the same time, the French and English East India Companies created their bases on the coast of India, a comparative advantage when European demand changed in the seventeenth century and Indian textiles became the leading commodity. In the eighteenth century they were replaced by tea and coffee with the China trade as core business. But trade remained under strict control of the Chinese and Japanese empires and 'western barbarians' were only tolerated until the middle of the nineteenth century when China and Japan were 'opened' for free trade by western military aggression. On Java and in India, however, the Mataram and Mughal empires began to disintegrate in the eighteenth century. This led to the control of Java by the Dutch and the conquest of all India by the British until 1818–49. Most of it was the result of contingency, but besides the interest in profits the ambition of local military, the 'men on the spot,' was a constant factor.

Since 1492 Castilians competing with the Portuguese discovered and conquered their India in the West. Spaniards emigrated to live in the cities of a New World. But despite a tremendous loss of native population, mostly through infectious diseases imported by the conquerors, the labor force consisted mostly of Amerindians, especially when, after the discovery of silver mines in northern Mexico and highland Peru, America became the treasure house of the Spanish monarchy. A proto-bureaucratic system of colonial government was established instead of traditional feudal lordship desired by the conquerors. But Iberian America was to remain a racially stratified society. Metropolitan Spain did not profit much from American silver which, because of differential inflation rates and overambitious policy, flowed out of the country to feed the Asian trade of the Dutch and the English. Via Europe and directly via the Philippines much of the American silver ended up in India and China. A kind of world trade system originated, but probably with rather marginal economic impact on Europe and on Asia. Brazil became a Portuguese colony almost by chance, economically important only in the later sixteenth century when sugar cane was grown on its abundant fertile soils. Indian labor was scarce here but the Portuguese controlled the opposite coast and therefore were able to provide the sugar industry with African slaves. When the sugar cycle reached its end, gold and diamonds created a new boom in Brazil during the eighteenth century, but with similar negative consequences for Portugal as for Spain 200 years earlier.

In the seventeenth century the Dutch transferred the Brazilian plantation economy to the Caribbean where the English, French, and Dutch had conquered several islands from Spain. No Amerindian labor force was left there, but the Dutch had occupied some of the Portuguese strongholds on the West African coast with the British and the French to follow. Thus the West Indies could be provided with African slaves. The English were the leading slave traders of the eighteenth century when the sugar business reached its culmination point. Large parts of the Caribbean were converted into an agro-industrial complex with a completely artificial society of black slaves, white lords, and an increasing number of mulattos.

In the meantime, another type of artificial new world had originated in North America which in the long run became a white man's country, the first and most successful of several 'new Europes' created by about 60 million Europeans who emigrated between the sixteenth and the twentieth century. North American Indians were few in number and not able to defend the land white immigrants desired. Whereas French Canada with its small settler population could at least attempt their integration, British North America had no room for them. They were removed, marginalized, or wiped out. Around 1780, British North America had more than 2.6 million inhabitants, compared with about 8 million inhabitants of England and Wales.

In 1775–1823 Europe lost most of her American colonies. But Britain, which dominated the colonial scene for most of the nineteenth century, compensated for the loss with the conquest of India and with several new colonies, some of them acquired during the wars against France, like Canada from France and South Africa from the Netherlands, other new foundations in Australia and New Zealand which had been explored during the late eighteenth century. But they were accepted rather reluctantly, because in an age of British dominated free trade colonial rule was considered unnecessary and too expensive. Therefore, self-financed self-government of white settler-colonies was now welcome. Between 1840 and 1931 these British 'dominions' became quasi-independent.

This relaxed attitude was to change in the age of *imperialism* (see *Imperialism, History of*), when nervous competition of old and new powers under the ideological impact of nationalistic social Darwinism led to the division of Africa within a few decades. Most of what was left of Asia was also occupied or at least controlled economically. Besides Britain's notorious rival France and the old colonial powers Portugal and Spain new competitors entered the stage: Russia and the King of Belgium, Germany and Italy, the USA and Japan, the first non-western imperialist power, which a few decades earlier had been close to becoming a semicolony. After World War I (see *First World War, The*) the German colonial empire and the Ottoman Empire were divided among the victors. The formal qualification as mandates of the League of Nations did not make much difference. Neither was the League of Nations able to prevent the Japanese occupation of Manchuria in 1931–2 and the Italian conquest of Ethiopia in 1935. Ethiopia had defended herself successfully against Italian aggression in 1896, but as the last African country had to yield to colonialism—if only for a few years. Times were beginning to change.

Less spectacular than expansion overseas, and therefore often ignored by historians of colonialism, was the 'quiet' continental expansion of Russia and the USA, of Canada, Australia, and South Africa, of Brazil, Argentina, and Chile. This expansion was possible only at the expense of indigenous populations and through the exploitation of differences of development. This turned out a very effective variety of colonialism, because the result was most often a white man's country, another 'new Europe,' where the former population was marginalized or wiped out. Only in few cases such as South Africa, Russian Central Asia and Caucasia non-western groups were numerically and culturally strong enough for later successful decolonization.

3. *Typology of Colonization and Colonialism*

There was no master plan for colonial expansion. Quite often it happened as an unintentional consequence of some other action. Nevertheless, typical sequences of actions occurred again and again, for instance preventive occupation to keep out possible competitors. The Portuguese used this in sixteenth-century Brazil, as did most European powers in nineteenth-century Africa. Military intervention in Africa was sometimes not intended as permanent occupation, because it was much too expensive. But when African resistance made a retreat impossible without loss of national prestige, the temporary occupation became a permanent colony. If not a matter of prestige, colonial policy was always based on a kind of cost-benefit analysis. Colonies were expected to be profitable. At least they had to finance their own government. The famous British model of *indirect rule* did originate much less from political wisdom or respect for other civilizations than from the necessity to keep administration costs down. In reality national budgets only profited from colonial empires in exceptional cases, because infrastructural costs were considered the responsibility of the government, whereas colonial profits remained private.

The most common individual motives for participating in colonial activities were indeed the desire for extra profits and the improvement of social status. These were certainly not the only motives, but they were almost never absent. Of course, their character changed over time. The conquerors of British India were capital accumulating profit seekers, but like those of Spanish America still with the intention to invest in

social status and not in capitalist enterprise in the way of the nineteenth and twentieth centuries. The second most common motive, missionary zeal, also changed over time. Instead of spreading the gospel *and* European civilization, modern colonizers followed the urge of a civilizing mission to educate barbarians. In the age of imperialism, however, it became common conviction that inferior races could not be 'improved,' but were doomed to serve or to die. Thus colonial adventurers could feel entitled to pose without restraint as the master race.

From the fifteenth to the twentieth century colonies usually represented one of three basic types, sometimes as a mixture:

(a) *Trade and/or military bases* such as the stronghold system of the Portuguese trade empire or Britain's naval bases in the nineteenth century or the frontier forts of the US, Spanish ecclesiastical missions sometimes served a similar purpose.

(b) *Colonies of settlement*, probably the primitive type of colony, especially when combined with a kind of stronghold as in the case of the roman 'Colonia.' Larger settlement colonies could only be founded at the expense of indigenous people who were either removed (New England) or transformed into a dependent labor force for an 'improved' economy (Spanish America, Algeria) or replaced by workers of foreign origin (Africans in the Caribbean, East Indians in Guyana).

(c) *Colonies of exploitation* with only a small number of members of the ruling people present most of them only temporarily, to run business, administration, and defense. Parts of colonial Latin America resembled that type, but it became the dominating type during nineteenth and twentieth century colonialism with British India as the prototype and much of Africa to follow.

Colonies of exploitation could not be ruled without massive collaboration of indigenous people (Robinson 1972). Collaboration had already been a necessary precondition for the Spanish conquest. To consider such collaborators as traitors, though, is an anachronistic value judgement of modern nationalist historiography. People like the Tlazcaltecans quite rationally served their own interest regardless of a Mexican nation that did not yet exist. Recent research on the *colonial situation* from the perspective of people under colonial rule has corrected such simplifications. Non-European history no longer looks just like an inversion of the eurocentric pattern when the colonialist image of benevolent western heroes ruling inferior races for their own benefit is simply inverted in the anticolonial story of western rascals systematically abusing helpless non-western victims. To some extent, the image of the helpless native itself is a creation of latent racism. In reality, a broad range of options to cope quite successfully with colonial domination ranging between complete acceptance of western patterns of behavior and very subtle methods of boycott was available to indigenous people, to some extent even to African slaves (see *Slaves/Slavery, History of*). The traditional contrast between colonial lords as active initiators of colonial processes and colonial subjects who had no choice but to suffer passively or at best to react against the actions of their betters has been questioned. But to believe after decolonization that colonial rule was only a short episode without much impact on the colonies seems to be an anticolonialist oversimplification.

4. *Decolonization*

The term *decolonization* was created in 1932 (Albertini 1966, p. 28) but the process is much older. A first wave of decolonization, the independence movements of most American colonies 1775–1823, reduced European colonial empires drastically. But Britain at least knew how to compensate for these losses. On the other hand, some lessons of the first wave were learned and made the second wave, the gradual increase of self-government in Britain's white dominions 1840–1931, much more easy going. Between world wars, anticolonial movements emerged in several exploitation colonies, with British India in the lead. Japanese occupation of Asian colonies during World War II (see *Second World War, The*) fostered independence movements which were successful after the war in a third wave of decolonization in the 1940s and 1950s. At that time, Africa was apparently not yet ready for decolonization. Colonial powers once again tried instead to compensate for their losses in Asia by economic development of their possessions in Africa. However, under the growing political pressure of the superpowers, the UN, and world public opinion some African countries became independent in the 1950s, and most in what was the fourth wave of decolonization during the 1960s. A fifth wave began in the Portuguese colonies in 1974–5 and accelerated political change in South Africa. But the sixth wave of decolonization in the 1990s, this time internal, included not only the end of South Africa's white minority regime, but also quite unexpectedly the breakdown of Soviet rule in Central Asia and Caucasia and Israel's first arrangements with her Arabs.

Decolonization was a truly dialectical process, because colonial rule produced its own contradiction, that is politically frustrated new indigenous elites. Everywhere westernized groups and not the traditional authorities were in the lead of independence movements. But an independence movement or even an independence war was not enough. In most cases a triangular constellation of political factors was the essential precondition: besides an internal independence movement, favorable international conditions such as allies, anticolonial policies of the world powers and the UN, and the willingness of the colonial power to accept the unavoidable were also necessary.

5. The Consequences of Colonialism and the Postcolonial World

Colonial processes were certainly never a one-sided affair, but it is simply not true that they left no impact on former colonies. The postcolonial world is not the same as before, even in Europe. Several American plants, the potato in particular, became essential for Europe's population and economic growth. Colonial bullion and colonial profits contributed to the rise of Europe, if rather marginally and indirectly. They were certainly not the sufficient precondition of industrialization. Colonial problems influenced European power relations, but rarely dominated them. Europe assimilated new cultural elements from all over the world, but without changing her character. In the colonies, however, ecology, economy, politics, society, and culture often changed more thoroughly, sometimes even fundamentally.

Intentionally or unintentionally colonial rule has transformed ecological systems. After new plant foods had been introduced, American savannahs changed into corn fields suffering from soil erosion and new weeds. Domesticated animals formerly unknown in the Americas, in Australia, and New Zealand revolutionized working and eating habits, transportation systems and even the whole way of life of such groups as the Indians of the North American plains. Unknown infectious diseases killed millions who were replaced by a new population of immigrants and imported slaves.

The deliberate diffusion of European and especially American cultivated plants such as maize and manioc created a new agriculture and allowed an expansion of food production, a necessary precondition of modern population growth in the colonies. But colonial economies were first of all based on the systematic exploitation of mineral resources such as American silver, African diamonds, gold and other ore, and more recently of oil in the Middle East and other parts of the world. In addition, they specialized in the production of high value agrarian products such as sugar, coffee, and tea on large plantations. Mines and plantations more often than not ruthlessly exploited a colonial labor force of slaves, contract and forced laborers. Because, as a rule, colonies could not industrialize, their independent successor states remained dependent on the world market price of their raw materials, which only in the case of oil took a turn to their advantage.

The essential political heritage of colonialism is the modern state (see *State, History of*) with its legislative, administrative, judicial, military, and educational apparatuses. Every former colony has changed into a modern nation state, but strictly within the arbitrary and sometimes absurd colonial borderlines. Therefore, some are extremely poor and very few are powerful enough for an independent foreign policy. But in contrast to former periods of history weak states today have their existence guaranteed by the international community (Jackson 1990). Internally, only few of them can build upon the tradition of a historical community. Instead of a modern civil society their colonial heritage is ethnic fragmentation, to some extent artificially created by colonialism. Quite often they had no alternative to the language of the former colonial power as national language. Under these circumstances the modern state in general and western democracy in particular could not become a success and were replaced by the para-stateliness of self-organizing groups based on remarkable combinations of traditional and modern elements.

Everywhere colonialism has created new social groups living in western-style urban environments: workers and employers, school teachers and professionals, civil servants and professional soldiers. Women have found new roles and quite often emancipated themselves from traditional restrictions. All this is based on western science and technology, on western ideas and ideologies like rationalism and individualism, the rule of law and human rights, Christianity and socialism. Nevertheless, it is essential to understand that this state of things no longer implies cultural dependency, but rather the opposite: non-western people in the meantime have become owners of a cultural heritage of western origin. For example, the role of English as common language of the world stems from historic British and American dominance. Increasingly, though, the acceptance of 'new Englishes' all over the world indicates that the Anglo-Saxons have been expropriated of their language.

See also: Civil Society, Concept and History of; Colonialism, Anthropology of; Colonialism: Political Aspects; Environmental History; Ethnic Groups/Ethnicity: Historical Aspects; Frontiers in History; Global History: Universal and World; Imperialism, History of; Imperialism: Political Aspects; Migrations, Colonizations, and Diasporas in Archaeology; National Liberation Movements; Nations and Nation-states in History; Racism, History of; Slaves/Slavery, History of

Bibliography

Albertini R v 1966 *Dekolonisation. Die Diskussion über die Verwaltung und Zukunft der Kolonien 1919–1960*. Westdeutscher Verlag, Köln und Opladen, Germany [1971 *Decolonization: The Administration and Future of the Colonies, 1919–1960*. Doubleday, Garden City, NY]

Albertini R v 1987 *Europäische Kolonialherrschaft 1880–1940*, 3rd edn. Steiner, Stuttgart [1982 *European Colonial Rule, 1880–1940: The Impact of the West on India, Southeast Asia, and Africa*. Greenwood Press, Westport, CT]

Bray W (ed.) 1993 *The Meeting of Two Worlds. Europe and the Americas 1492–1650*. Oxford University Press, Oxford, UK

Fieldhouse D K 1981 *Colonialism 1870–1945. An Introduction*. Weidenfeld and Nicholson, London
Forsyth J 1992 *A History of the Peoples of Siberia, Russia's North Asian Colony, 1581–1990*. Cambridge University Press, Cambridge, UK
Hind R J 1984 The internal colonial concept. *Comparative Studies in Society and History* **26**: 543–68
Pluchon J, Bouche D (eds.) 1991 *Histoire de la colonisation française*. Fayard, Paris, 2 Vols.
Meyer J, Thobie J (eds.) 1990–1991 *Histoire de la France coloniale*. Armond Colin, Paris, 2 Vols.
Holland R F 1985 *European Decolonization, 1918–1981*. Macmillan, Basingstoke, UK
Jackson R H 1990 *Quasi-states. Sovereignty, International Relations, and the Third World*. Cambridge University Press, Cambridge, UK
Louis W M, Low A, Canny N, Marshall P, Porter A, Brown J M, Winks R W (eds.) 1998–1999 *The Oxford History of the British Empire*. Oxford University Press, Oxford, UK, 5 Vols.
Lockhart J S, Schwartz S B 1983 *Early Latin America. A History of Colonial Spanish America and Brazil*. Cambridge University Press, Cambridge, UK
Milner II C A, O'Connor C A, Sandweiss M A (eds.) 1994 *The Oxford History of the American West*. Oxford University Press, New York
Olson J S (ed.) 1991 *Historical Dictionary of European Imperialism*. Greenwood Press, New York
Osterhammel J 1995 *Kolonialismus. Geschichte—Formen—Folgen*. C. H. Beck, Munich [1997 *Colonialism. A Theoretical Overview*. M. Wiener, Princeton, NJ]
Reinhard W 1983–1990 *Geschichte der europäischen Expansion*. W. Kohlhammer, Stuttgart, 4 Vols.
Reinhard W 1996 *Kleine Geschichte des Kolonialismus*. Kröner, Stuttgart [1997 *Petite histoire du colonialisme*. Belin SUP, Paris]
Reinhard W (ed.) 1999 *Verstaatlichung der Welt? Europäische Staatsmodelle und außereuropäische Machtprozesse*. Oldenbourg, Munich
Robinson R E 1972 Non-European foundation of European Imperialism. In: Owen R, Sutcliffe B (ed.) *Studies in the Theory of Imperialism*. Longman, London
Scammell G V 1989 *The First Imperial Age. European Overseas Expansion c. 1400–1715*. Unwin Hyman, London

W. Reinhard

Color Classification and Symbolism

'Color classification' is an ambiguous concept. It may refer to 'classification *of* colors,' or 'classification *by* colors'. Most research on color classification has opted for the first approach, which is concerned with the variation of linguistic categorization of colors in different cultures. This kind of research has been undertaken by linguists, psychologists, and cognitive anthropologists. By contrast, the few anthropologists who have opted for the second approach have been concerned rather with symbolic anthropology, or have worked in a structuralist tradition of sociocultural anthropology. Research into color classification has a long history, dating back to the 1870s (Allen 1879, Gladstone 1877, Magnus 1877). However, it was not until Berlin and Kay published their seminal book on basic color terms (1969) that research in this field virtually exploded, especially within cognitive and linguistic anthropology. Over the last 25 years, a wealth of data on color categories has been meticulously collected in different parts of the world, mainly with the help of the 'Munsell color chips.' As a general rule, randomly chosen informants have been asked to define the colors on these chips, and to 'map' them. The findings have then been tested as 'hard data,' often using quantitative methods (e.g. MacLaury 1986, Kay et al. 1991). This 'hard-wire' approach to color classification research has been further strengthened by the general assumption that color perception is a matter of neurobiology, universality, and evolution.

Although the evolutionary paradigm had featured in research on color classification for almost 100 years, Berlin and Kay were the first scholars to present substantial evidence for a theory concerning the universality and evolution of basic color terms. In the original formulation of their theory, they defined a particular sequence of developmental 'stages' of basic color terms, and suggested that these stages would correspond to the stages of technological evolution. The few cultures in which only two basic color terms were used would represent Stage I on the evolutionary ladder, while the prevalence of three color terms would place a particular culture one step higher up (Stage II). Four basic color terms would imply Stage III, five terms Stage IV, and so on. Societies with the maximum number of color terms (11) would represent the highest level of technological evolution.

Although Berlin and Kay have modified the original formulation of their theory (Kay et al. 1991), the main features of the paradigm have been generally accepted by scholars in their own and related fields. Part of the reason for this is that Berlin and Kay backed up their theory with a wide range of cross-cultural linguistic data, which showed a remarkable regularity with respect to the sequence of basic color terms: if only two terms were found in a particular culture, they were invariably white and black. If three terms were recognized, these were always white, black, and red. The fourth basic color term to be introduced was either yellow or green, the fifth either green or yellow, the sixth blue, and so on. In spite of some revisions and later developments of the theory (see Hardin and Maffi 1997, MacLaury 1999), there is still a general consensus among anthropologists about the sequence of basic color terms—even though the hypothesis about a corresponding technological evolution has yet to be proved. It would, for example, be difficult to maintain the theory about technological evolution, given that many hunting/gathering societies recognize more basic color terms than agriculturalists. In Africa, even the most sophisticated Bantu-speaking agricul-

turalist and agropastoral societies have only three basic color terms, whereas their hunting and gathering neighbors may have five or more.

However, strictly cognitive and linguistic research on color classification has met with criticism on several other accounts. The main criticism has been leveled at the 'hard-wire' methods employed in this kind of work, and at the lack of 'emic' approaches to the issue of color classification. It has been argued that the 'meanings' of color terms cannot be evaluated adequately when the data have been collected under conditions reminiscent of a laboratory situation (Saunders 1999; cf. Tornay 1978), and that 'affective features of meaning' may contribute to Berlin and Kay's evolutionary sequence of color terms (Osgood et al. 1975, Taft 1997). The problem of 'meaning' is perhaps the most controversial issue associated with the concept of color classification. Though a few representatives of the 'hard-wire' approach have attempted to account for the 'meanings' of individual color terms (Kay and McDaniel 1978, MacLaury 1992), the focus has still been on the implications of the informants' task of 'mapping' color chips (Lucy 1997, p. 327). By way of contrast, Lucy directs attention to Conklin's pioneering work on color categories in the 1950s (Conklin 1955). Applying a structural approach, Conklin discovered that the four-color categories recognized by the Hanunóo corresponded to an underlying four-way classification of other visual stimuli as well.

Conklin's findings correspond to the ideas proposed by the Swiss-French school of semiology, as represented by Saussure and Derrida. According to them, the meanings of signs depend on their positions in a larger system of signs (Layton 1997). Meanings are thus anchored in the syntagmatic relations among signs, but if we also take into consideration the meanings which Barthes (1967) has referred to by the concept of 'the symbolic relation between the sign and the signified,' it may be possible to identify congruities of symbolic meanings when 'parallel' syntagmatic sequences of signs are compared. Such a 'parallel' structure of signifying, syntagmatic sequences of signs may be referred to as a paradigmatic structure of meaningful categories.

To identify such paradigmatic structures in cultural symbolism is the core of anthropological research on symbolic classification. This approach has characterized symbolic anthropology since the late 1960s. When colors have been included in this kind of symbolic analysis, the research has thus been concerned with classification *by* colors, rather than *of* colors. However, at the same time as this structurally oriented research developed, there was also a growing awareness among some anthropologists of the importance of contextualization in the study of symbolic meanings.

The pioneering research in the domain of symbolic classification by colors was carried out by Victor Turner, who in 1966 wrote an article on the symbolic meanings of white, black, and red among the matrilineal Ndembu of Zambia (Turner 1966). Not only did Turner attempt to trace the basic symbolic meanings of the three colors, he also analyzed their variable references in different ritual and social contexts, and thus became one of the first anthropologists to put contextual analysis of cultural symbols on the agenda.

Turner's article stirred up an enthusiastic interest in color classification, especially among anthropologists working in the structuralist tradition. No matter whether the anthropologists found two, three, or four colors to be key symbols in a particular culture, they usually attempted to identify a dualistic system of symbolic classification, of which the colors formed a part. Thus, for instance, most of the contributions to Needham's book on dual classification included references to colors as 'parallel' devices of classification, alongside the right/left division that was the main topic of the book. While the concept of 'paradigmatic structures of signs' was not mentioned, there was an insistence on the congruity between different levels of symbolic classification. In many articles referring to different areas of the world, colors were regarded as part of this dual classification.

Needham's book was cast in the kind of structuralist tradition that marked the 1970s, and which was inspired by Lévi-Strauss (1963). The basic assumption was that human beings everywhere tend to classify their world into binary oppositions. However, even if a dualistic approach could account for the choice of two opposing categories of classification in certain specific situations in any society, it cannot solve the enigma of triadic classification, which is typical of the ritual use of colors in many parts of the world.

In most African cultures south of the Sahara, triadic classification by colors appears to permeate the ritual practice of healers and diviners (Baumann 1935, Straube 1960, White 1961, Ngubane 1977, Jacobson-Widding 1979, Kirby 1999). This applies particularly to the Bantu-speaking societies, where only three basic color terms are recognized, namely white, black, and red. In a systematic investigation of the symbolic references of the three colors in the matrilineal belt of Central Africa, Jacobson-Widding (1979) found that the only situations in which all three colors were visually represented were in connection with certain rituals, when the ritual specialist manipulated the colors as symbols with particular, contextually dependent references. In profane contexts in which the official ideology of social and moral order was referred to as a matter of principle, on the other hand, white and black were the only colors demonstrated (cf. Jacobson-Widding 1991). In Central Africa, there is even a formal prohibition against showing the color red in such contexts, for instance in connection with lawsuits. This implies that white and black are contrasted as binary oppositions only in certain contexts, like those in which the dominant structure of order is

supposed to be manifested, and in which the rationality of a logic based on two opposite values is to be applied. In contrast, red is represented in situations where indeterminate values are dealt with, and where a simplistic two-value logic would fail to account for the truth.

The concrete references of the color red may vary according to the context. However, whatever tangible phenomenon it may refer to in any particular situation, most ritual experts in central and southern parts of Africa will employ this color when they wish to represent ambiguity. In particular, red is used to represent phenomena that seem to defy classification into any of the 'officially' recognized categories of determinate values that are represented by white and black. Such phenomena, or states of affairs, may consist of bewildering emotions, anomalies, or mysterious spirits. Or they may be 'twilight' situations like dusk and dawn, birth and death, or any state of 'betwixt and between'—such as the state of initiands in a bush camp, or the state of adolescent girls, whose bodies are painted red during a period of seclusion. In essence, the color red can be viewed as a symbol of the kind of ambiguity that may result from conceptual boundary crossings. That is, when a particular phenomenon, or state of affairs, is considered to blur the boundaries of the concepts on which a 'rational' order is based, it is depicted as 'red.' This implies that the color red is associated with irrational feelings, and with the unknown.

Considering the ways in which ritual experts in Africa tend to interpret and manipulate colors, we may regard the color triad as a device for philosophical categorizations. At the most general level of meanings, the color triad appears to classify different kinds of phenomena according to three basic values of truth and morality. Whereas the combination of white and black is supposed to represent clear-cut values as logical opposites (i.e. 'good' vs. 'bad,' 'true' vs. 'false,' 'us' vs. 'them,' 'sky' vs. 'earth,' and so on), red will be introduced when the logical boundary distinguishing A from B is impossible to define. The color red will thus represent a third, indeterminate value, which can be classed neither as A, nor as B (nor as non-A), but rather as C (cf. Leach 1975).

This implies that the two-value logic of the Aristotelian philosophy of knowledge is supplemented with a third term, which in the case of African ritual practice is regarded as an *a priori* neutral value, in relation to the determinate values of truth and morality. However, this 'three-value logic' is assumed to be valid only in ritual thought and praxis, and not in contexts where profane, 'rational' reasoning is supposed to be applied.

Since red is used to symbolize an unknown factor, which by definition is initially value-neutral, this color will be associated with unknown potentialities and acquire a dynamic quality. Accordingly, it is sometimes associated with powerful emotions, sometimes with 'magic' power—a magic power that many people may regard as dangerous, polluting, or taboo, whereas others may associate it with miracles and miraculous healing. It is this 'magic' power that is exploited by the ritual expert in Africa, whenever he or she manipulates the color red in any of the many possible combinations or oppositions with white or black. As a general rule, the ritual expert will begin any 'sacred,' ritual seance with a visual demonstration of all three colors, as if to show the cards he or she intends to play with.

If we understand the concept of 'magic power' in a broad sense, some of the African associations of the color red may also be discerned in sacred rituals outside Africa, even within the Christian church in Europe. When red is combined with white, it usually points to connotations of the 'sacred,' or the 'holy.' Correspondingly, we often find that the combination of white and black is resorted to in situations where the order of society, or its moral dogma, is supposed to be demonstrated—not only in Africa, but in Europe and other parts of the world as well. However, whether used to represent 'sacred' values in the church or 'profane' values in the courtroom, any combination of colors from the triadic set of symbols constituted by white, black, and red may have different concrete references in different situations, in both Europe and Africa. The 'meanings' of these colors as culturally instituted symbols are thus to be found on different levels of interpretation. Colors may be used as concrete signifiers with context-dependent meanings and, at the same time, as representations of more general values of 'truth' and 'power.'

See also: Color Terms, Linguistics of; Color Vision; Vision, Psychology of

Bibliography

Allen G 1879 *The Colour Sense: Its Origin and Development*. Trubner, London

Barthes R 1967 *Elements of Semiology*. Cape, London

Baumann H 1935 *Die Lunda. Bei Bauern und Jägern in Inner-Angola*. Reimer, Berlin

Berlin B, Kay P 1969 *Basic Color Terms. Their Universality and Evolution*. University of California Press, Berkeley

Conklin H 1955 Hanunóo color categories. *Southwestern Journal of Anthropology* **11**: 339–44

Gladstone W E 1877 The color sense. *Nineteenth Century* **2**: 366–88

Hardin C L, Maffi L (eds.) 1997 *Color Categories in Thought and Language*. Cambridge University Press, New York

Jacobson-Widding A 1979 *Red–White–Black as a Mode of Thought. A Study of Triadic Classification by Colours in the Ritual Symbolism and Cognitive Thought of the Peoples of the Lower Congo. Uppsala Studies in Cultural Anthropology 1*. Almqvist & Wiksell International, Uppsala, Sweden

Jacobson-Widding A 1991 The encounter in the water mirror. In: Jacobson-Widding A (ed.) *Body and Space. Symbolic Models of Unity and Division in African Cosmology and Experience. Uppsala Studies in Cultural Anthropology 16*. Almqvist & Wiksell International, Uppsala, Sweden

Kay P, Berlin B, Merrifield W R 1991 Biocultural implications of systems of color naming. *Journal of Linguistic Anthropology* **1**: 12–25
Kay P, McDaniel C 1978 The linguistic significance of the meanings of basic color terms. *Language* **54**: 610–46
Kirby J P 1997 White, red and black: Color classification and illness management in Northern Ghana. *Social and Medicine Science* **44**(2): 215–30
Layton R 1997 *An Introduction to Theory in Anthropology*. Cambridge University Press, London
Leach E 1975 *Culture and Communication. The Logic by which Symbols are Connected*. Cambridge University Press, Cambridge
Lévi-Strauss C 1963 *Structural Anthropology*. Basic Books, New York
Lucy J 1997 The linguistics of 'color.' In: Hardin C L, Maffi L (eds.) *Color Categories in Thought and Language*. Cambridge University Press, New York
MacLaury R E 1986 Color in Meso-America. A theory of composite categorization. Doctoral Dissertation, University of California, Berkeley
MacLaury R E 1992 From brightness to hue: An explanatory model of color-category evolution. *Current Anthropology* **33**: 137–86
MacLaury R E 1999 Basic color terms: Twenty-five years after. In: Borg A (ed.) *The Language of Color in the Mediterranean*. Almqvist & Wiksell International, Uppsala, Stockholm
Magnus H 1877 *Die geschichtliche Entwicklung des Farbensinnes*. Veit, Leipzig
Needham R (ed.) 1973 *Right & Left. Dual Symbolic Classification*. The University of Chicago Press, Chicago
Ngubane H 1977 *Body and Mind in Zulu Medicine*. Academic Press, London
Osgood C E, May W H, Miron M S 1975 *Cross-cultural Universals of Affective Meaning*. University of Illinois Press, Urbana, IL
Saunders B 2000 Revisiting 'Basic Color Terms.' *Journal of the Royal Anthropological Society* **6**: 81–99
Straube H 1960 *Gedanken zur Farbensymbolik in afrikanischen Eingeborenenkulturen*. Studium Generale 1, Göttingen
Taft C 1997 *Generality Aspects of Color Naming and Color Meaning*. Department of Psychology, Göteborg University
Tornay S (ed.) 1978 *Voir et nommer les couleurs*. Laboratoire d'Ethnologie et de Sociologie Comparative, Nanterre
Turner V 1966 Colour classification in Ndembu ritual. In: Banton M (ed.) *Anthropological Approaches to the Study of Religion*, ASA Monographs. Tavistock, London; Humanities Press, New York, Vol. 3, pp. 47–84
White C M N 1961 *Elements in Luvale Beliefs and Rituals*. Rhodes-Livingstone Papers 32

A. Jacobson-Widding

Color Terms, Linguistics of

The words that different languages use to name colors have furnished a prime focus for research on the doctrine of *linguistic relativity*. This doctrine holds that (a) different languages categorize the world very differently and (b) the nonlinguistic thought and behavior of different peoples are deeply influenced by these differences in linguistic categorization. Words for colors have been considered an ideal area of lexicon for testing this doctrine because there exist scientific color order systems, which provide a precise metalanguage for describing color sensations independent of the words of any natural language. Such a metalanguage furnishes the means to characterize precisely the sameness or difference in meaning of color words in different languages.

1. *Restricting the Domain*

It is clear, however, that every language has immense—probably infinite—resources for denoting color sensations. First, every language so far studied can productively create complex expressions which describe the color of an object by indicating similarity to another kind of object which is known to saliently display the color. An example from English is 'like the color of a male mallard duck's neck feathers.' (Although most languages do not have a word that straightforwardly translates the English word 'color,' the speaker of such a language who wishes to characterize, say, a kind of fish as mallard green can usually say that, for example, the fish 'looks like' the neck feathers of a male mallard, relying on context and background knowledge to provide the interpreter with color as the relevant dimension of visual similarity. Although relativist critiques of color naming research have maintained that absence of a word meaning 'color' in a language indicates that its speakers see and/or conceive of colors very differently from, say, speakers of European languages, the experiences of workers who have actually studied color vocabularies in the field do not provide significant support for this claim. See, e.g., Conklin 1955, p. 340, Turton 1980, p. 322.) Second, many language have words like 'blond' or 'bay,' which denote a color only as restricted to a class of objects (hair and horses, respectively). Third, many languages have morphological and syntactic processes which create complex color terms out of simple color terms. English examples include 'blue-green,' 'yellowish,' 'pale reddish purple,' and so on. Finally, some languages, English included, have many linguistically simple terms which denote subtypes or 'shades' of colors denoted by other simple terms. For example, 'scarlet,' 'crimson,' 'vermillion,' 'puce,' 'magenta,' 'burgundy,' and 'maroon' are among the more commonly named shades of 'red.' The linguistics of color terms has become a focus of research and a source of controversy in part because of the practical need to arrive at a manageable set of 'color terms' in each language to enable comparative study.

2. *Color Terms and Linguistic Relativity*

Before taking up in greater detail the ways researchers have attempted to delimit the linguistic domain of color, it is useful to consider briefly the role color

naming has played in the long-running debate over linguistic relativity vs. semantic universals (q.v.). In the second half of the nineteenth century it had become common knowledge among scholars that the major color words of different languages were not always perfectly translatable into one another. To some, this represented *prima facie* evidence of biologically based, perceptual differences among different groups of human beings, in fact the last traces of Darwinian evolution observable among humans. For example, William Gladstone, Homeric scholar and Prime Minister of Britain, observing differences between the color words of classical Greek and the modern European languages, wrote, '… the organ of color and its impressions were but partially developed among the Greeks of the heroic age' (1858, pp. 457–99).

Although this view was widespread in the late nineteenth century, it was not unanimous. For example, in order to study the discrimination and naming of colors independently, the German ophthalmologist Hugo Magnus circulated a questionnaire to missionaries and traders with ten standardized color samples and instructions for using them in the 1870s. These instructions contained an explicit warning that failure of a language to distinguish lexically between two colors did not necessarily imply that speakers of that language did not distinguish the two colors perceptually. Magnus received completed questionnaires on twenty-five African, fifteen Asian, three Australian, and two European languages. He concluded in part, 'As regards the range of the color sense of the primitive peoples tested with our questionnaire, it appears in general to remain within the same bounds as the color sense of the civilized nations. At least, we could not establish a complete lack of the perception of the so-called main colors as a special racial characteristic of any one of the tribes investigated for us. We consider red, yellow, green, and blue as the main representatives of the colors of long and short wavelength; among the tribes we tested not a one lacks the knowledge of any of these four colors' (Magnus 1880, p. 6, as trans. in Berlin and Kay 1969, p. 141). Magnus did find widespread *lexical* neutralization of green and blue, that is, a single word covering both these colors, as have all subsequent comparative studies of color lexicons.

While nineteenth century and early twentieth century researchers saw differences among vocabularies in an evolutionary perspective (biological and/or cultural), in the second quarter of the twentieth century the tide began to turn against evolutionary explanations of human variation and toward cultural and linguistic relativity. This was especially true in America, in the tradition of the anthropologists and linguists F. Boas, E. Sapir, and B. L. Whorf, although there were precursors in the views of the German romantics, who had tended to associate the *geist* of a people strongly with its language. According to Whorf, 'The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face. On the contrary the world is presented in a kaleidoscopic flux of impressions which have to be organized in our minds. This means, largely, by the linguistic system in our minds' (1956, p. 212). Throughout the 1930s and 40s this view gained increasing prominence and in the 1950s and 60s achieved virtual hegemony in American anthropology and linguistics, with considerable influence in Britain and elsewhere outside of the US. (Chomskian universalism and nativism were to become influential only at the end of this period.)

The parade example chosen by the mid-century American relativists to make the twin points of semantic arbitrariness and linguistic determinism was the lexicon of color. The most influential introductory linguistics textbook of the period, for example, stated: 'There is a continuous gradation of color from one end of the spectrum to the other. Yet an American describing it will list the hues as red, orange, yellow, green, blue, purple, or something of the kind. There is nothing inherent either in the spectrum or the human perception of it which would compel its division in this way' (Gleason 1961, p. 4). This sweeping conclusion appears to have been based on two small-scale empirical studies of color terminology, whose results have subsequently been interpreted differently, one on the Hanunóo language (Philippines) (Conklin 1995) and the other a comparison of the color terminologies of languages (Ray 1953). Despite the paucity of evidence, the idea that different languages divide up the perceptual color space in mutually incomparable ways was widespread at the time, doubtless because it fits so well with the prevailing doctrine of radical cultural and linguistic relativity.

3. *Color Terms as Linguistic Objects*

Against this background a study appeared in 1969 which reached exactly contrary conclusions. After examining the color term systems of ninety-eight languages, Berlin and Kay (1969) [B&K] proposed that (a) the differences in color lexicons of the languages of the world could be explained as different combinations of a few landmark colors (which were probably human perceptual universals) and (b) the differences in color lexicons of the world's languages could be arranged in an evolutionary progression. In order to carry out this study, B&K found it necessary to limit the domain of inquiry to the 'basic color terms' of the languages being compared. (They examined twenty languages experimentally, using standard color stimuli, and interpreted another seventy-eight reports of color terminologies from the literature based on their formulation of the results of these twenty.) As a preliminary to this research, Berlin and Kay offered what they called a definition of 'basic color term' and much of the subsequent research on the linguistics of

color terms has had to do with attacking, defending, revising and evaluating this definition.

The B&K study was not without direct precursors. Lenneberg and Roberts (1956) [L&R] compared the color terms of Zuni monolinguals with those of Zuni-English bilinguals and English monolinguals. After isolating the words whose denotation was to be studied, L&R had each informant indicate on a representative array of 230 color samples the denotation of each color word as well as the color sample which best exemplified the reference of that word. B&K used essentially the same method and the same stimulus array as L&R (except for adding black, white, and shades of gray) with the exception that the color words whose denotation was to be studied were selected differently by L&R and by B&K. L&R proceeded by getting speakers to list as many color words as they could think of without color stimuli present, which produced a list of fifty-two terms. They also elicited lists of color words by asking for names for twenty-four selected color stimuli, producing a list of 105 terms. The color terms for which L&R show mappings are those most frequently elicited by the second method. L&R show mappings for seven color terms for Zuni monolinguals, corresponding closely to English pink, red, brown, yellow-orange, green, blue, and purple. (Recall they had no white, black, or gray stimuli.) They show mappings for eight color terms for Zuni bilinguals, adding a term for orange, '°olechin-anne' ('orange.like' < Eng. 'orange' + 'anne' 'like') and retracting the area of yellow accordingly. (They show ten mappings for English color terms (pink, red, orange, brown, yellow, green, light green, blue, light blue, purple). No terms like greenish, blue-green, or crimson were mapped, but light green and light blue were apparently ruled in by naming frequency considerations. (L&R's text is not explicit on this point.) Clearly in L&R, as in Conklin's, Ray's, and other early investigations of color word systems some tacit, if hazy, notion of major color term or principal color term is at work. It was this tacit notion that B&K tried to operationalize in their 'definition' of basic color term.

4. *B&K's Definition of Basic Color Term*

B&K offered what they unfortunately called a definition of basic color term [bct]. Their intent was not to postulate and define a new theoretical entity but to operationalize, insofar as possible, a tacit concept that was already in general use. The guiding intuition of the B&K 'definition' of basic color term was that each language has a small set of simplex lexemes (or word *senses*, as in the case of English 'green [color]' and 'orange [color]') of pure color meaning (i.e., unlike words such as 'blond'), which collectively name all the color sensations. The bcts are the members of that set. However, B&K stated the definition in terms of eight criteria, four primary and four subsidiary, the latter to be used in cases where application of the former is not sufficient to decide the bcts of a language. The four primary criteria for a bct were that it (a) be monolexemic (its meaning not predictable from the meanings of its parts, eliminating 'greenish,' etc.), (b) not be a hyponym (eliminating, e.g., 'crimson'), (c) be predictable of any kind of object (eliminating, e.g., 'blond') and (d) be 'psychologically salient' (eliminating e.g., 'puce' and 'like the neck feathers of a male mallard duck'). Two telling criticisms that have been leveled against the fourth criterion are (i) that it is not a linguistic criterion, while the other three are and (ii) that 'psychological salience' is neither a conceptually limpid nor a readily measurable notion. The four subsidiary criteria, designed for use in *ceteris paribus* situations, were that a questionable term (i) have a morphological distribution similar to that of the established bcts, (ii) not be the name of a kind of object (which would eliminate English 'orange' if subsidiary criteria were allowed to override primary criteria), (iii) not be a recent borrowing from another language, and (iv) not be morphologically complex. The B&K definition of bct has led both to practical problems in determining the bcts of a language and to conceptual and theoretical problems regarding the semantics of color in natural languages.

5. *Practical Problems with the B&K Definition of Bct*

The main practical problem in applying the B&K criteria to given languages is that since language change is always gradual there are many transitional cases. For example, there are many languages with bcts for black, white, red, yellow, and 'grue,' i.e., a single word covering green and blue. There are also many languages with bcts for black, white, red, yellow, green, and blue. When a language changes from the first type to the second type, it is not the case that everyone goes to bed Sunday night speaking the 'grue' language and gets up Monday morning speaking the 'green' vs. 'blue' language. Many kinds of in-between situations arise which make it difficult to say which words of the language naming part or all of the green/blue region of color space are bcts. Moreover, there is virtually always widespread inter-speaker variation in such cases, complicating further the task of deciding the bcts of 'the language' (See *Sociolinguistics*). Second, the four primary criteria of B&K do not always correlate and the definition gives no clue as to how the different criteria should be weighted in such cases. Also, as mentioned, cultural salience is not a straightforwardly observable characteristic. Workers who have taken the B&K 'definition' as a recipe rather than a heuristic have complained of problems in application in particular cases. Usually, however, these problem cases have involved a single

term or a single area of the color space, indicating that the area is one undergoing change.

A third practical issue has arisen in connection with the first B&K criterion of non-predictability of meaning of a bct from the meaning of its parts. It has been noted that in several Oceanic and Australian languages many of the presumably basic color terms are in fact reduplicated forms of roots that denote objects notoriously exhibiting the corresponding color. For example, in Yélîdnye (Rossel Island, Melanesia [Non-Austronesian]), the word for white *kpaaîkpaapî* represents reduplication of the nominal root *kpaapî* denoting a pure white cockatoo species. A similar situation holds for the black and red words, the only other plausible candidates for bcts in this language (Levinson 2000). The question arises whether *Yélîdnye*, and similar languages, can be said to have any bcts at all. The answer is yes. The criterion of non-predictability of meaning requires not only that from knowing the rest of the *Yélîdnye* language—including the word *kpaapî* and the pattern of reduplication which yields a salient attribute of the denotatum of a noun from reduplication of the noun stem—one could as a *listener* guess the meaning of *kpaapîkpaaî*, but also that as a *speaker* one would know an advance that *kpaapîkpaaî* was the way to express the idea of white. The latter of course would not be possible without memorizing this particular fact, because not all attributive words are formed by reduplication and even if they were the meaning 'white' might be based on some other object or substance.

6. Theoretical Issues Concerning the B&K Definition of Bct

It has been argued, principally by John Lucy (e.g., Lucy 1997), that a set of bcts isolated by B&K or associated researchers doesn't correspond to any linguistically legitimate system because there is no insistence in the B&K tradition of research that the set of bcts for a given language be definable by strictly formal, distributional criteria. This critique of the bct concept and practice is based on the questionable assumption that all morpho-syntactic categories correspond to semantic categories and vice versa (see Kay 1999). Moreover, careful empirical studies have shown that application either of the explicit B&K criteria or of the underlying intuition of a small set of linguistically simplex lexemes jointly exhaustive of the color domain can isolate sets of basic color terms which obviously form a coherent semantic system and whose members belong to several different form classes (see, for example, Maffi 1990 on Somali). It has also frequently been found that morphological category, including native vs. borrowed origin, can correlate with the evolutionary sequence posited by B&K ({black, white} > {red} > {yellow, green} > {blue} > {other}). The Somali case is again illustrative: the words for black, white, and red are intransitive verbs, the words for yellow and green are denominal adjectives and the word for blue (*buluug*), also a basic color term, is a noun. To say that something is blue, one has to say that it 'has' blue (Maffi 1990). Such neat correlations between morphology and evolutionary status are, however, sporadic.

Recently, the question has arisen whether all languages do in fact have a small set of lexemes whose denotations are restricted to color sensations and which jointly name all the color sensations. That is, does every language have a full set of basic color terms? This idea has been dubbed the Emergence Hypothesis (EH). (The question was first raised in unpublished work by Luisa Maffi. See also Lyons 1995, Lyons 1999, Kay 1999, Levinson 2000, Kay and Maffi 2000.) Research on this hypothesis has just begun but it appears that a small fraction of the world's languages may lack complete lexical partitions of the color domain (Levinson 1999, Kay and Maffi 2000). Preceding, and in some cases overlapping temporally, with the evolution of basic color term systems there may exist an evolution toward basic color term systems.

7. Nonbasic Color Terms and Noncolor Meanings of Color Terms

The vast majority of studies of color terms, especially since the early 1970s, have restricted themselves to the basic or major color terms, considering nonmajor terms only with a view to excluding them from further consideration. The L&R study described above typifies this practice. There have, however, been studies covering all the lexicalized color expressions of a language, a classic example being André's (1949) study of Latin color terms and a more recent example being Bricker's (1999) study of productive and semi-productive morphological process in the formation of secondary color words in two Mayan languages. Bricker also provides a thorough analysis of the non-color aspects of meaning of Mayan non-basic color terms, often involving texture and other aspects of surface appearance. Color words frequently also display metaphorical associations to moral, emotional, supernatural, and other kinds of non-physical entities and beliefs. Such associations have been widely reported in the ethnographic literature.

See also: Color Classification and Symbolism; Color Vision; Color Vision Theory; Kinship Terminology

Bibliography

André J 1949 *Étude sur les termes de couleur dans la langue latine*. G. Klincksieck, Paris

Berlin B, Kay P 1969 *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley

Bricker V R 1999 The morphology of color and texture words in the Maya language of Yucatan. *Anthropological Linguistics* **41**(3): 283–307
Conklin H C 1955 Hanunóo color categories. *Southwestern Journal of Anthropology* **11**: 339–44
Gladstone W E 1858 *Studies on Homer and the Homeric Age*. Tubner, London
Gleason H A 1961 *An Introduction to Descriptive Linguistics*. Holt, Rinehart and Winston, New York
Lenneberg E H, Roberts J M 1956 The language of experience: A study in methodology. Memoir 13 of *International Journal of American Linguistics*
Levinson S C 2000 Yélîdnye and the theory of basic color terms. *Journal of Linguistic Anthropology* **10**(1): 1–53
Lucy J A 1997 The linguistics of 'color.' In: Hardin C L, Maffi L (eds.) *Color Categories in Thought and Language*. Cambridge University Press, Cambridge, UK
Lyons J A 1995 Colour in language. In: Lamb T, Bourriau J (eds.) *Colour: Art and Science*. Cambridge University Press, Cambridge, UK
Lyons J A 1999 The vocabulary of colour with particular reference to Ancient Greek and Classical Latin. In: Borg A (ed.) *The Language of Colour in the Mediterranean*. Almqvist and Wiksell International, Stockholm
Kay P 1999 The emergence of basic color lexicons hypothesis. In: Borg A (ed.) *The Language of Colour in the Mediterranean*. Almqvist and Wiksell International, Stockholm
Kay P, Maffi L 2000 Color appearance and the emergence and evolution of basic color lexicons. *American Anthropologist* **101**(4): 743–60
Maffi L 1990 Somali color term evolution: Grammatical and semantic evidence. *Anthropological Linguistics* **32**: 316–34
Magnus H 1880 *Die geschichtliche Entwicklung des Farbensinnes*. Viet, Leipzig
Ray V 1953 Human color perception and behavioral response. *Transactions of the New York Academy of Sciences* **16**(2): 98–104
Turton D 1980 There's no such beast: Cattle and colour naming among the Mursi. *Man* **15**: 320–38
Whorf B L 1956 Science and linguistics. In: Carroll J B (ed.) *Language, Thought and Reality*. MIT Press, Cambridge, MA

P. Kay

Color Vision

Color has long interested people from many different professions: artists, engineers, physicists, psychologists, and biologists, all with their own viewpoints and areas of primary interest. We consider here how the underlying neural organization of the eye and brain allows us to discriminate color and produces our various color percepts.

The light from every location in space varies in two physical dimensions: intensity (or the number of photons) and wavelength (or the energy levels of the photons). Color vision is the ability to discriminate changes in the spectral composition of a visual stimulus independently of its effective intensity, and vice versa. Because color vision apparently evolved independently several times, it is an ability that we share with many other species. However, we will restrict our discussion to the color vision of humans and other Old World primates.

1. Trichromatic Limitation

A visual scene may include millions of different local spectral distributions of light. A human observer with normal color vision, however, can visually match every one of these millions of different colored stimuli by appropriately combining and adjusting the intensities of just three different chromatic lights. The three variable-intensity lights that can be combined to match any color are known collectively as primaries. Despite what one may have been taught, there is no requirement that the primaries be red, blue, and yellow, or any other particular set, and they can be drawn from similar colors or from very different colors. The only restriction is that no two primaries can be combined to match the third. This trichromacy implies that at some processing stage, color information is limited by being transmitted through a three-channel pathway. It does not identify the point at which the three-channel bottleneck occurs. Many earlier theories assumed that trichromacy implies the existence of only three cone types, but recent evidence indicates that the limited number of different channels to the brain, not the number of receptors, limits our color vision to three dimensions. The trichromatic limitation to our color vision, the collapse of millions of physical dimensions into just three perceptual dimensions, should not be taken to mean that we are poor at discriminating colors. We make very fine discriminations within this three-dimensional space and can distinguish millions of different colors. However, the colors we see are not related to the physical spectral characteristics of objects in any simple fashion.

To gain any useful information from the retinal image is a much more difficult task than one might imagine; it requires multiple levels of neural computation involving billions of cells that occupy about a third of the whole cortex. An understanding of some aspects of this complex computational problem helps explain some otherwise-puzzling features of color vision.

2. Separating Objects from the Illuminant

One problem faced by the visual system is that most natural objects are visible not because they themselves emit light, but only indirectly because they reflect some proportion of whatever light falls on them. The light that reaches our eyes from a given object is, therefore, a product of the intensity and spectrum of the light source and the reflectance characteristics of the object's surface. Since we do not have any independent

information about the illuminant, it is not possible to determine the reflectance characteristics of any region, its color or its lightness, with certainty. The visual system can only guess at the solution. The main way it does so is by assuming that the illuminant is relatively constant whereas objects vary across space and time (with eye movements). It then estimates the intensity and spectral distribution of each region by comparing the activity evoked by the light from this region with the average neural activity (assumed to be primarily a function of the illuminant) over small neighboring retinal regions and across 100 msec or so of time. As a result, the perceived color of an object depends not only on the light coming from it but also on the light coming from nearby points in space, as well as the light to which the eye has been adapted recently. Among the consequences of this process are the well-known color and brightness contrast effects.

3. *Separating Intensity and Wavelength Information*

A second problem for the visual system is that a photoreceptor can only report how many photons it has absorbed, and the number of photons absorbed is determined both by the number of incident photons (the intensity of the light) and by their respective energy levels (the wavelength of the light). Thus, intensity and wavelength are completely confounded in the receptors. The information, for instance, that 100 photons were absorbed—the only thing that the receptor can report—does not allow one to determine either the intensity or the wavelength of the light in that region. To determine *either* the wavelength *or* the intensity requires neural comparisons of the outputs of neighboring receptors and of different cell types in the cortex.

4. *Conflict between Acuity and Color Vision*

A third difficulty results from the fact that the visual system, over at least part of the retina, needs to maintain as high spatial acuity as possible. To build a detailed spatial map requires treating each receptor within the high-acuity area as a separate, equivalent spatial unit. However, it requires a comparison of the outputs of at least two receptors that *differ* in their spectral sensitivities at each location to separate intensity and wavelength information. There is then an inherent conflict between the requirements for color and those for high spatial acuity.

We do not yet fully understand how the visual system handles these and other problems, but we do understand some of the processes that contribute to the solutions in the early stages of visual processing. We will describe briefly the sequence of steps, as we understand them to date, in the neural processing in the retina and visual cortex that lead to our ability to

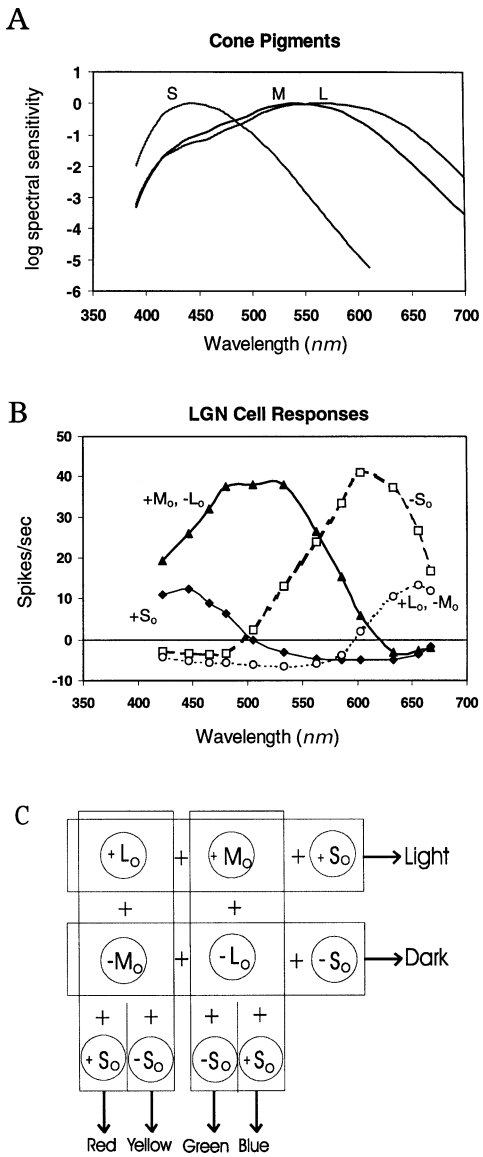

Figure 1
A: The spectral sensitivity of normal human cone photoreceptors. **B**: The average responses in spikes (impulses) per second of populations of P_c and K_c cells in macaque monkey lateral geniculate. These cells are spectrally-opponent, being excited by light of some wavelengths and inhibited (negative numbers) from the maintained rate by light of other wavelengths. **C**: A diagram of how the different spectrally-opponent geniculate cells may be combined in the striate cortex to produce the basic perceptual color systems

identify and characterize the color and lightness of objects.

5. Photopigments and Spectral Sensitivity

The initial stage in the visual response is the absorption of light by the photopigment molecules in the rod and cone receptors. Variations in the amount of light absorbed by the photopigments in a receptor trigger a complex series of reactions, leading ultimately to synaptic transmission of the information to the next neurons in the pathway to the brain. Each receptor type is differentially sensitive to light of different wavelengths. The basic shapes of the spectral sensitivity functions of the various cone photopigments are similar, but the spectral region to which they are most sensitive differs among the S, M, and L (for short-, medium-, and long-wavelength) cones (see Fig. 1A). The cone photopigments all have quite broad spectral sensitivity. The L and M pigments in particular absorb significantly at all visible wavelengths. The wavelength of the light is thus not given by which receptor is activated, since all are activated by lights of most visible wavelengths, but only by the relative activation of the three different receptor types.

The compromise made by the visual system between the requirements for spatial vision (all cones the same) and for color vision (cones differing in spectral sensitivity), can be seen in the fact that the spectral sensitivities of the L and M cones are very similar whereas the S cone curve is very different. Almost all our spatial vision is based on the L and M cones, with the S cones contributing just to color vision.

6. Color Vision Deficits

Almost all human observers appear to have the same S-cone photopigment, but some differ in their L and M cone pigments. About 7 percent of human males and perhaps 1 percent of females have so-called red-green color vision deficiencies. The most extreme of these color-defective individuals (very few are truly color-blind, although that term is often used) are missing either the L or the M cone pigment and therefore have only two cones types. They are thus dichromats, and can match any color using only two primaries. Protanopes are missing the L cone pigment, and deuteranopes the M cone pigment. With only the S cones and either the L or M cones, these dichromats can distinguish short from long wavelengths, but they confuse colors over the whole red-yellow-green portion of the spectrum, where the S cones are insensitive.

More common than dichromats are anomalous trichromats, in whom either the L or the M cone pigment is present but shifted in spectral sensitivity with respect to that in normal observers. They thus are trichromats, but they make different color matches than normals do. Finally, there are very rare individuals, tritanopes, who are missing the S cones and are, therefore, dichromats.

7. Genetic Basis of Color Vision Deficits

The prevalence of red–green color-vision defects and their higher incidence in males can be accounted for by the chromosomal location of the various photopigment genes. The S cone pigment gene is on chromosome 7, and tritanopia occurs equally frequently in both sexes. The M and the L cone photopigment genes, however, lie end-to-end on the X chromosome. Because they are very similar and are in adjacent positions on the same chromosome, the L and M photopigment genes are subject to interchanging parts. This results in a variety of L and M genes in the population. The extreme case is one in which the two genes end up being the same, thus specifying only an M or an L photopigment. Since males only inherit one copy of the X chromosome while females inherit two, a female would have to inherit anomalous genes from both parents to be color-defective. A male, on the other hand, only receives a single X chromosome, from his mother, and is correspondingly more likely to have anomalous or dichromatic color vision. An important corollary of this is that many females appear to have more than just two LM pigments, and thus more than three different cone photopigment types. Nonetheless, females with more than three cone types still have only a trichromatic color vision system. Presumably the two L cone types that such a woman might possess are treated as being identical by later neural processes and thus do not produce an additional chromatic dimension. This strongly suggests that the trichromatic limit to human color vision is the result of only three types of analyses in the retina, and of only three paths from there to the cortex, not, as was long supposed, because of the presence of only three cone types.

8. Receptor Function

The retina is part of the brain that just happens to lie some distance away, in the eye, and receptors are similar to neurons elsewhere in the nervous system. They respond to stimuli by either a depolarization or a hyperpolarization of their outer membranes, and they communicate by releasing synaptic chemicals in proportion to the amount of depolarization in the synaptic region. While ordinary neurons respond to external transmitter chemicals, however, photoreceptors respond to chemicals generated within the receptors themselves, triggered by the absorption of photons by their photopigments.

The photopigment molecules are in the outer segments of the photoreceptors. The receptor outer

segment membrane is maintained at an intermediate level of polarization by a balance between opposing sets of chemical interactions. The effect of increases and decreases in the amount of light absorbed by the photopigment molecules (due to a change in either the intensity or the wavelength of incident light) is to shift the polarization balance briefly in one direction or the other. This polarization change in the outer segment is conducted down the receptor to its synaptic region where it produces a change in the rate of synaptic chemical release. A decrement in light absorbed shifts the balance towards depolarization and leads to an increase in synaptic transmitter release to activate the next neurons in the chain to the brain; an increment in light absorption shifts the balance towards hyperpolarization and produces a decrease in transmitter release. As we move our eyes around in a scene, the amount of light absorbed by each receptor oscillates up and down, and the receptors quite faithfully transduce these oscillations into oscillating amounts of synaptic transmitter release. Since these variations in absorption could be due to either intensity differences or wavelength differences in the scene, however, it is left to the neural processing to sort out which is which and thus identify the color and brightness of each visual object.

9. *Retinal Connectivity*

There are three main types of retinal processing and three main paths from the eye to the cortex, which we will refer to as the $\boldsymbol{M}_c$, $\boldsymbol{P}_c$, and $\boldsymbol{K}_c$ paths. They are so called because they pass through the magno-, parvo-, and konio-cellular layers of the lateral geniculate nucleus, the relay station between the eye and the cortex. None of the cells in these three paths signals the amount of light in its region of the retina. Rather, each cell signals the differences between different distributions of light within its receptive field (RF), the region from which it receives its input. The cells do not signal the outputs of just one cone type, either, but rather they signal the sums and differences between the activity of different cone types.

The cells in the $\boldsymbol{M}_c$ path, in a spatially-opponent organization, signal changes in the amount of light absorbed by a small group of L and M cones relative to the amount absorbed by the L and M cones in a larger, overlapping area. Recall that the L and M cones absorb across the whole visible spectrum. Half the cells in this path fire if there is more light in the RF center than in the surround ($+\mathrm{LM}_c - \mathrm{LM}_s$); the other half fire if there is less light in the center than the surround ($-\mathrm{LM}_c + \mathrm{LM}_s$). With uniform illumination, center and surround responses will tend to cancel. This organization begins to emphasize information from objects and minimize that related to the illuminant, because if all the L and M cones in a region absorb more light, or if they all absorb less, it is probably due to a change in the illumination (as when the sun goes behind a cloud or emerges from one). The reflections from different parts of an object, on the other hand, will likely differentially activate one group of cones relative to their neighbors. Information about such intensity variations is what is captured and carried to the cortex by cells in the $\boldsymbol{M}_c$ pathway.

The cells in the $\boldsymbol{K}_c$ path, in a spectrally-opponent organization, signal changes in the amount of light stimulating one or more S cones relative to the amount stimulating the neighboring L and M cones ($+\mathrm{S}_c - \mathrm{LM}_s$). We shall refer to these as S-opponent cells, or $+\mathrm{S}_o$. Since the cone photopigments each have broad spectral sensitivity, one cannot identify the wavelength by which cone is activated. However, differences in the relative amount of activation of the different cone types is related to the spectral distribution of the light, and this information is captured by a spectrally-opponent organization. Short-wavelength light is absorbed relatively more by the S than by the L and M cones. It will therefore excite a cell in the $\boldsymbol{K}_c$ path. Long-wavelength light is absorbed more by the L and M cones than by the S cones, and will thus inhibit a $+\mathrm{S}_o$ cell in the $\boldsymbol{K}_c$ path.

Finally, cells in the $\boldsymbol{P}_c$ path (about 80 percent of the total) have both a spatially- and a spectrally-opponent organization. These cells, at least in the central retina, receive input from a single L or M cone in their RF center, and a small group of L and/or M cones in the surround. There are thus four varieties of these cells: $+\mathrm{L}_c - \mathrm{M}_s$ (or L_o), $-\mathrm{L}_o$, $+\mathrm{M}_o$, and $-\mathrm{M}_o$. The $+\mathrm{L}_o$ and $-\mathrm{M}_o$ cells have the same spectral response characteristics (both L−M), as do the $+\mathrm{M}_o$ and $-\mathrm{L}_o$ cells, (both +M−L), and are therefore combined in Fig. 1B. Since they encode the difference between the activity of L vs. M cones in spectrally-opponent organizations, they respond to color differences in the long-wavelength half of the spectrum. (It is the absence of these cells, because of a lack of either L or M cones, that makes a protanope or deuteranope unable to discriminate red from yellow from green.) However, these cells also signal changes in activation of a single L or M cone relative to that in the surrounding cones, in a spatially-opponent organization. A $+\mathrm{L}_o$ cell, for instance, fires to a large red patch, but also to a small light spot on a dark background. The $\boldsymbol{P}_c$ cells thus carry both color and intensity information. The neural processing in the retina begins but does not complete the process of separating color and intensity information.

10. *Cortical Processing*

The three main retino-cortical paths project separately to the striate cortex (also known as the first visual cortical area, V1), but within the cortex the paths are no longer discrete. Rather, V1 cells combine the outputs of the different cell types in various ways. Most of the early cortical processing involves building RFs to detect various spatial aspects of the pattern

(e.g., spatial frequency, orientation, motion, depth) within a local cortical region. With respect to color vision *per se*, the primary processing involves separating color and luminance information, and further separating changes due to the illuminant from those due to visual objects, by lateral interactions over large regions.

To separate luminance and color information, the outputs of $\boldsymbol{P}_c$ cells are combined in two different ways. When their outputs are summed in one way, the luminance components to their responses sum and the color components cancel. Summed in a different combination, the color components sum and the luminance components cancel. Consider a striate cortex cell that combines inputs from one or more $+L_o$ and $+M_o$ cells in a region. The cortical cell would respond to luminance variations but not to color variations, since the neurons providing its inputs both fire to luminance increments in the RF center and to decrements in the surround, but the color organizations of its inputs are opposite to each other (one being L-M and the other M-L). Combined with input from a $+S_o$ cell, this would produce a V1 cell that fires to white (light increments) and inhibits to black (light decrements) but does not respond to pure color variations. This is represented in the top row of Fig. 1C. However, a V1 cell receiving inputs from both $+L_o$ and $-M_o$ cells, or from both $+M_o$ and $-L_o$ cells (columns in Fig. 1C), would respond to color changes but not to luminance variations since their color responses would add, but their luminance RFs, which are opposite to each other, would cancel. This organization by itself would produce L-M color cells that would fire to so-called warm colors (red and yellow) and inhibit to cool colors (blue and green). M-L cells would fire to cool colors and inhibit to warm colors. As shown in Fig. 1C, the further addition of $+S_o$ or $-S_o$ cells can split these classes into separate red and yellow, and separate blue and green systems, respectively.

All of the primary visual information is passed through V1, but subsequent visual areas are partially specialized for the further analysis of various different functional aspects of vision. One later visual area (V4) is crucially involved with color perception. Individuals with localized V4 lesions can still discriminate objects on the basis of their color variations, but they report that the objects now appear to have no hue, as if viewed on a black-white television screen. There is also a report of one case with the reverse loss: a patient who could see colored but not black-white objects.

11. Color Appearance

The appearance of a color can be specified by values along just three perceptual dimensions known as hue, saturation and brightness. Hue refers to the characteristic described by such color names as red, yellow, green, and blue. Saturation refers to the extent to which the stimulus differs perceptually from a purely achromatic (i.e., white, gray, black) axis. The third dimension is brightness or lightness. That our perceptual space is three-dimensional reflects the basic trichromacy of vision.

A normal observer can describe the hue of any light (disregarding surface characteristics) by using one or more of only four color names (red, yellow, green, and blue). These so-called unique hues form two opponent pairs, red–green and blue–yellow. Red and green normally cannot be seen in the same place at the same time; if unique red and unique green lights are added in appropriate proportions, the colors cancel and one sees a neutral gray. Orange can be seen as a mixture of red and yellow, and purple as a mixture of red and blue, but there is no color seen as a red–green mixture (or as a blue–yellow mixture). This perceptual opponency is also reflected in color contrast. Red can induce the appearance of green into neighboring regions, and after staring at a red surface one sees a green after-image. The yellow–blue opponent pair produces similar effects. It was these perceptual characteristics of color that led Ewald Hering in the nineteenth century to propose that the various color systems were not independent but rather that color was processed in a spectrally opponent organization, an idea which has since been amply verified in the presence, discussed above, of spectrally-opponent cells in the path from receptors to the cortex.

See also: Color Vision Theory; Vision, Low-level Theory of; Vision, Psychology of; Visual Perception, Neural Basis of; Visual System in the Brain

Bibliography

De Valois R L, De Valois R L 1988 *Spatial Vision*. Oxford University Press, New York

Hurvich L M 1981 *Color Vision*. Sinauer Press, Sunderland, MA

Kaiser P K, Boynton R M 1996 *Human Color Vision*. Optical Society of America, Washington, DC

Neitz J, Neitz M 1998 Molecular genetics and the biological basis of color vision. In: Backhaus W G S, Kliegl R, Werner J S (eds.) *Color Vision*. Walter de Gruyter, Berlin, pp. 101–19

Spillmann L, Werner J S 1990 *Visual Perception: The Neurophysiological Foundations*. Academic Press, New York

K. K. De Valois and R. L. De Valois

Color Vision Theory

Color vision is the ability to distinguish and identify lights and objects on the basis of their spectral properties. This entry presents several key topics that underlie current theories of human color vision. These are trichromacy, color opponency, adaptation, and color constancy.

1. Introduction

Information about color is transformed as it flows from the stimulus through the initial stages of the human visual system. At each image location, the color stimulus is specified by the amount of power it contains at each wavelength. The classic color matching experiment shows that the normal human visual system is trichromatic: only three dimensions of spectral variation are coded by the visual system. The biological basis of normal trichromacy is that the retina contains three classes of cone photopigment. After the initial encoding of light by the cones, further processing occurs. Two aspects of this processing are particularly important. First, signals from three classes of cones are recombined to form a luminance and two color opponent channels. Second, there is adaptive signal regulation that keeps neural signals within their operating range and stabilizes the appearance of objects across changes of illumination.

2. Trichromacy

2.1 Color Matching

The physical property of light relevant for color vision is the spectral power distribution. A light's spectral power distribution specifies the amount of power it contains at each wavelength in the visible spectrum, often taken to lie roughly between 400 and 700 nm. In practice, spectral power distributions are measured at discrete sample wavelengths. Let the measured power values be denoted by $b_1, \ldots, b_{N_\lambda}$ where N_λ denotes the number of sample wavelengths. Then the vector

$$\mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_{N_\lambda} \end{bmatrix} \tag{1}$$

provides a compact representation of the spectral power distribution. Use of a vector representation for spectral quantities facilitates a variety of colorimetric computations (e.g., Brainard 1995). Wavelength sample spacings between 1 and 10 nm are typical.

Trichromacy is demonstrated by the basic color matching experiment (Wandell 1995, Brainard 1995). In this experiment, an observer views a bipartite field. One side of the field contains a test light. This light is experimentally controlled and can have an arbitrary spectral power distribution. On the other side of the field is the matching light. This consists of the weighted mixture of three primary lights. Each primary has a fixed relative spectral power distribution, but its overall intensity in the mixture can be controlled by the observer. The observer's task is to adjust the primary intensities until the mixture has the same color appearance as the test light. The primaries used in the experiment are chosen to be independent, so that no weighted mixture of any two produces a match to the third.

Because the matching light is constrained to be a weighted mixture of three primaries, it will not generally be possible for the observer to make the test and matching lights physically identical. For many test lights, however, the observer can adjust the matching light so that it appears identical to the test light even though the two differ physically. For some test lights, no choice of primary intensities will afford a match. In these cases one or more of the primaries can be mixed with the test light and primary intensities found so that the primary/test mixture matches the mixture of the remaining primaries. A useful descriptive convention for the color matching experiment is to assign a negative intensity to any primary that must be mixed with the test to make a match. Given this convention, any test light can be matched by a mixture of three independent primaries.

The color matching experiment is an empirical system. Given a test light described by a vector $\mathbf{b}$, the experiment returns a vector

$$\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix} \tag{2}$$

whose entries are the individual primary intensities. When the primaries are scaled by these intensities and mixed, a match to the test light is created. The vector $\mathbf{t}$ specifies what are called the tristimulus coordinates of the light $\mathbf{b}$. A theory of color matching should let us predict $\mathbf{t}$ for any test light $\mathbf{b}$, given the spectral power distributions of the primary lights.

As an empirical generalization, the color matching system is a linear system (e.g., Wyszecki and Stiles 1982, Brainard 1995, Wandell 1995). That is, if we have two test lights $\mathbf{b}_1$ and $\mathbf{b}_2$ with tristimulus coordinates $\mathbf{t}_1$ and $\mathbf{t}_2$, then any weighted mixture $(a_1\mathbf{b}_1 + a_2\mathbf{b}_2)$ of the two test lights has tristimulus coordinates given by the corresponding mixture $(a_1\mathbf{t}_1 + a_2\mathbf{t}_2)$. In these vector expressions, multiplication of a vector (e.g., $\mathbf{b}_1$) by a scalar (e.g., a_1) consists of multiplying each entry of the vector by the scalar, while addition of two vectors (e.g., $a_1\mathbf{b}_1$ and $a_2\mathbf{b}_2$) consists of adding the corresponding entries of the two vectors.

The linearity of color matching makes it possible to predict the match that will be made to any test light on the basis of a relatively small number of measurements. Consider the set of monochromatic lights with unit power. If N_λ wavelength samples are used in the underlying representation, then there are N_λ such lights and we can denote their spectral representations by $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{N_\lambda}$. Each of the $\mathbf{c}_i$ has a 1 as its ith entry

and zeros elsewhere. Note that any light $\mathbf{b}$ may be thought of as a weighted mixture of monochromatic lights, so that $\mathbf{b} = \sum_i b_i \mathbf{c}_i$ where b_i is the ith entry of $\mathbf{b}$. Let the vectors $\mathbf{t}_i$ specify the tristimulus coordinates of the monochromatic lights $\mathbf{c}_i$. The linearity of color matching then tells us that the tristimulus coordinates of any light $\mathbf{b}$ are given by $\mathbf{t} = \sum_i b_i \mathbf{t}_i$.

A set of tristimulus values $\mathbf{t}_i$ measured for monochromatic lights $\mathbf{c}_i$ is referred to as a set of color matching functions. Although these are often plotted as a function of wavelength, they do not represent the spectral power distributions of lights. The color matching functions may be specified by a single matrix

$$\mathbf{T} = \left[\mathbf{t}_1 \mathbf{t}_2 \mathbf{t}_3 \cdots \mathbf{t}_{N_\lambda} \right] \tag{3}$$

whose N_λ columns consist of the individual tristimulus coordinate vectors $\mathbf{t}_i$. This specification allows computation of tristimulus coordinates from spectral power distributions through simple matrix multiplication:

$$\mathbf{t} = \mathbf{T}\mathbf{b}. \tag{4}$$

Both tristimulus values and color matching functions are defined with respect to the primaries chosen for the underlying color matching experiment. The Commission Internationale de l'Eclairage (CIE) has standardized a system for color representation based on the ideas outlined above. The CIE system is widely used to specify color stimuli and many sources describe it in detail (e.g., Wyszecki and Stiles 1982, Brainard 1995, Kaiser and Boynton 1996).

The advantage of using tristimulus coordinates to describe color stimuli is that they provide a more compact and tractable description than a description in terms of wavelength. Tristimulus coordinates are compact precisely because they do not preserve physical differences that are invisible to the human visual system. The representational simplification afforded by tristimulus coordinates is extremely valuable for studying processing that occurs after the initial encoding of light. On the other hand, it is important to remember that the standard tristimulus representations (e.g., the CIE system) are based on matches made by a typical observer looking directly at a small stimulus at moderate to high light levels. These representations are not necessarily appropriate for applications involving some individual observers, nonhuman color vision, or color cameras (e.g., Wyszecki and Stiles 1982, Brainard 1995).

2.2 Biological Basis of Color Matching

The color matching experiment is agnostic about the biological mechanisms that underlie trichromacy. It is generally accepted, however, that trichromacy typically arises because color vision is mediated by three types of cone photoreceptor. Direct physiological measurements of individual primate cones support this hypothesis (see Wandell 1995, Rodieck 1998). First, the responses of individual cones depend only on the rate at which photopigment molecules are isomerized by the absorption of light quanta; once the intensity of two lights has been adjusted so that they produce the same isomerization rates, the cone response does not distinguish the two lights. This idea is referred to as the principle of univariance. Second, individual cones may be classified into one of three distinct types, each with a characteristic spectral sensitivity. The spectral sensitivity is proportional to the probability that light quanta of different wavelengths will isomerize a molecule of the cone's photopigment. The three types of cones are often referred to as the long- (L), middle- (M), and short- (S) wavelength-sensitive cones. If an observer has only three types of cones, each of which obeys the principle of univariance, two physically distinct lights that produce the same isomerization rates for all three classes of cones will be indistinguishable to the visual system. Quantitative comparison confirms that color matches set by a standard observer (defined as the average of matches set by many individual observers) are well predicted by the equations of isomerization rates in the L-, M-, and S-cones.

As described above, trichromacy occurs for most observers because their retinas contain cones with three classes of photopigments. Genetic considerations, however, indicate that some individuals have retinas containing four classes of cone photopigments (Sharpe et al. 1999). Either these individuals are tetrachromatic (mixture of four primaries required to match any light) or else their trichromacy is mediated by information lost after quantal absorption. In addition, some human observers are dichromatic (only two primaries must be mixed to make a match to any light.) Most cases of dichromacy occur because one photopigment is missing (Sharpe et al. 1999, Neitz and Neitz 2000).

An alternative to using tristimulus coordinates to represent the spectral properties of lights is to use cone coordinates. These are proportional to the isomerization rates of the three classes of cone photopigments. The three dimensional vector

$$\mathbf{q} = \begin{bmatrix} q_L \\ q_M \\ q_S \end{bmatrix} \tag{5}$$

specifies cone coordinates where q_L, q_M, and q_S denote the isomerization rates of the L-, M-, and S-cone photopigments respectively. It can be shown (e.g., Brainard 1995) that cone coordinates and tristimulus coordinates are related by a linear transformation, so that

$$\mathbf{q} = \mathbf{M}_{tq}\mathbf{t} \tag{6}$$

where $\mathbf{M}_{tq}$ is an appropriately chosen 3 by 3 matrix.

Computation of cone coordinates from light spectra requires estimates of the cone spectral sensitivities. For each cone class, these specify the isomerization rates produced by monochromatic lights of unit power. The sensitivities may be specified in matrix form as

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_L \\ \mathbf{s}_M \\ \mathbf{s}_S \end{bmatrix} \tag{7}$$

where each row of the matrix is a vector whose entries are the spectral sensitivities for one cone class at the sample wavelengths. Given $\mathbf{S}$, cone coordinates are computed from the spectral power distribution of a light as

$$\mathbf{q} = \mathbf{Sb} \tag{8}$$

Because Eqns. (4), (6), and (8) hold for any light spectrum $\mathbf{b}$, it follows that

$$\mathbf{S} = \mathbf{M}_{tq}\mathbf{T} \tag{9}$$

Current estimates of human cone spectral sensitivities are obtained from color matching data using Eqn. (9) together with a variety of considerations that put constraints on the matrix $\mathbf{M}_{tq}$ (Stockman and Sharpe 1999).

3. *Postabsorption Processing*

Color vision does not end with the absorption of light by cone photopigments. Rather, the signals that originate with the absorption of light are transformed as they propagate through neurons in the retina and cortex. Two ideas dominate models of this post-absorption processing. The first is color opponency: signals from different cone types are combined in an antagonistic fashion to produce the visual representation at a more central site. The second idea is adaptation: the relation between the cone coordinates of a light and its central visual representation is not fixed but depends instead on the context in which the light is viewed. Section 3.1 treats opponency, while Sect. 3.2 treats adaptation.

3.1 *Opponency*

Direct physiological measurements of the responses of neurons in the primate retina support the general idea of opponency (e.g., Dacey 2000). These measurements reveal, for example, that some retinal ganglion cells are excited by signals from L-cones and inhibited by signals from M-cones. One suggestion about why this occurs is that it is an effective way to code the cone signals for transmission down the optic nerve (see Wandell 1995).

A possible approach to understanding post-absorption processing is to keep the modeling close to the underlying anatomy and physiology and to characterize what happens to signals at each synapse in the neural chain between photoreceptors and some site in visual cortex. The difficulty is that it is not presently possible to cope with the complexity of actual neural processing. Thus many color theorists have attempted to step back from the details and develop more abstract descriptions of the effect of neural processing. Models of this sort are often called mechanistic models. These models generally specify a transformation between the quantal absorption rates $\mathbf{q}$ elicited by a stimulus and a corresponding visual representation $\mathbf{u}$ postulated to exist at some central site. The idea is to choose a transformation so that (a) the color appearance perceived at a location may be obtained directly from the central representation corresponding to that location and (b) the discriminability of two stimuli is predictable from the difference in their central representations.

Most mechanistic models assume that signals from the cones are combined additively to produce signals at three postreceptoral sites. Two of these sites carry opponent signals. These are often referred to as the red-green (RG) and blue-yellow (BY) signals. A third site carries a luminance (LUM) signal, which is not thought to be opponent. If we take

$$\mathbf{u} = \begin{bmatrix} u_{LUM} \\ u_{RG} \\ u_{BY} \end{bmatrix} \tag{10}$$

to be a three-dimensional vector with entries given by the LUM, RG, and BY signals, then the additive relation between cone coordinates $\mathbf{q}$ and the visual representation $\mathbf{u}$ may be expressed in matrix form:

$$\mathbf{u} = \mathbf{M}_o\mathbf{q} \tag{11}$$

Many (but not all) detailed models take LUM to be a weighted sum of L- and M-cone signals, RG to be a weighted difference between the L- and M-cone signals, and BY to be a weighted difference between the S-cone signal and a weighted sum of the L- and M-cone signals. In these models $\mathbf{M}_o$ would have the form

$$\mathbf{M}_o = \begin{bmatrix} m_{11} & m_{12} & 0 \\ m_{21} & -m_{22} & 0 \\ -m_{31} & -m_{32} & m_{33} \end{bmatrix} \tag{12}$$

where all of the m_{ij} are positive scalars representing how strongly one cone class contributes to the signal at one post-receptoral site.

Considerable effort has been devoted to establishing whether the linear form for the mapping between $\mathbf{q}$ and $\mathbf{u}$ is appropriate, and if so, what values should be

used for the m_{ij}. Several types of experimental evidence have been brought to bear on the question.

As an example, one line of research argues that four color perceptions, those of redness, greenness, blueness, and yellowness, have a special psychological status, in that any color experience may be intuitively described in terms of these four basic perceptions. Thus orange may be naturally described as reddish-yellow and aqua as greenish-blue. In addition, both introspection and color scaling experiments suggest that the percepts of redness and greenness are mutually exclusive so that both are not experienced simultaneously in response to the same stimulus, and similarly for blueness and yellowness (e.g., Hurvich and Jameson 1957, Abramov and Gordon 1994). Given these observations, it is natural to associate the RG signal with the amount of redness or greenness perceived in a light (redness if the signal is positive, greenness if it is negative, and neither red nor green if it is zero) and the BY signal with the amount of blueness or yellowness. Judgments of the four fundamental color perceptions, obtained either through direct scaling (e.g., Abramov and Gordon 1994) or through a hue cancellation procedure (e.g., Hurvich and Jameson 1957), are then used to deduce the appropriate values of the m_{ij} in the second and third rows of $\mathbf{M}_o$. When this framework is used, the entries for the first row of $\mathbf{M}_o$, corresponding to the LUM signal, are typically established through other means such as flicker photometry (e.g., Kaiser and Boynton 1996).

Other approaches to studying the opponent transformation include analyzing measurements of the detection and discrimination of stimuli (e.g., Wyszecki and Stiles 1982, Kaiser and Boynton 1996, Eskew, et al. 1999, Wandell 1999), and measurements of how the color appearance of lights is affected by the context in which they are viewed (e.g., Webster 1996). In part because of a lack of quantitative agreement in the conclusions drawn from different paradigms, there is currently not much consensus about the details of the transformation between $\mathbf{q}$ and $\mathbf{u}$. One of the major open issues in color theory remains how to extend the simple linear model described above so that it accounts for a wider range of results.

3.2 Adaptation

Figure 1 illustrates a case where the same light has a very different color appearance when seen in two different contexts. The figure shows two disk-annulus stimulus configurations. The central disk is the same in each configuration, but the appearance of the two disks is quite different. To explain this and other context effects, mechanistic models assume that at any given time and image location, the relation between the quantal absorption rates $\mathbf{q}$ and the visual representation $\mathbf{u}$ depends on the quantal absorption rates

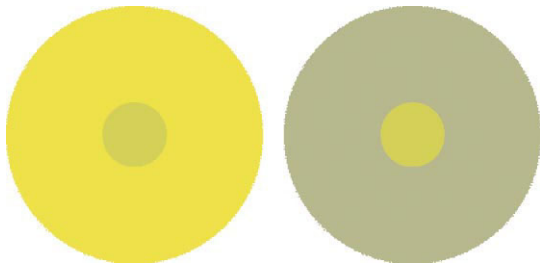

Figure 1
A color context effect. The figure illustrates the color context effect known as simultaneous contrast. The two central disks are physically the same but appear different. The difference in appearance is caused by the fact that each disk is seen in the context of a different annular surround. This figure is best viewed in color. A color version is available in the on-line version of the Encyclopedia

at other locations and at preceding times. To help fix ideas, it is useful to restrict attention to the disk-annulus configuration. For this configuration, the visual representation of the disk may be written as

$$\mathbf{u}_d = f(\mathbf{q}_d; \mathbf{q}_a, \varphi) \tag{13}$$

where $\mathbf{u}_d$ is the visual response to the disk, $\mathbf{q}_d$ and $\mathbf{q}_a$ are the cone coordinates of the disk and annulus respectively, and φ represents other contextual variables such as the size of the disk and annulus and any temporal variation in the stimulus. Clearly, f( ) must incorporate the sort of transformation described by the matrix $\mathbf{M}_o$ in Sect. 3.1 above.

As was the case with the discussion of opponency above, there is not wide agreement about how best to model adaptation. A reasonable point of departure is a cone-specific affine model. In this model, the visual representation $\mathbf{u}$ of a light is related to its cone coordinates $\mathbf{q}$ through an equation of the form

$$\mathbf{u} = \mathbf{M}_o(\mathbf{D}_1\mathbf{q} - \mathbf{q}_1) \tag{14}$$

where $\mathbf{M}_o$ is as in Eqn. (12) and

$$\mathbf{D}_1 = \begin{bmatrix} g_{L1} & 0 & 0 \\ 0 & g_{M1} & 0 \\ 0 & 0 & g_{S1} \end{bmatrix}, \quad \mathbf{q}_1 = \begin{bmatrix} q_{L1} \\ q_{M1} \\ q_{S1} \end{bmatrix} \tag{15}$$

In this formulation, the g's on the diagonals of $\mathbf{D}_1$ characterize multiplicative adaptation that occurs at a cone-specific site in visual processing, before signals from separate cone classes are combined. The entries of the vector $\mathbf{q}_1$ characterize subtractive adaptation. Equation (14) is written in a form that implies that the subtractive adaptation also occurs at a cone-specific

site. The entries of $\mathbf{D}_1$ and $\mathbf{q}_1$ depend on the cone coordinates $\mathbf{q}_a$ of the annulus as well as on spatial and temporal variables characterized by φ. Note that the cone-specific affine model is a generalization of the idea that the visual representation consists of a contrast code.

Asymmetric matching may be used to test the adaptation model of Eqn. (14). In an asymmetric matching experiment, an observer adjusts a match stimulus seen in one context so that it appears to have the same color as a test stimulus seen in another context. More concretely, consider Fig. 1. In the context of this figure, an asymmetric matching experiment could be conducted where the observer was asked to adjust the central disk on the right so that it matched the appearance of the central test disk on the left. Suppose such data are collected for a series of N test disks with cone coordinates $\mathbf{q}_{ti}$. Denote the cone coordinates of the matches by $\mathbf{q}_{mi}$. Within the mechanistic framework, the corresponding visual representations $\mathbf{u}_{ti}$ and $\mathbf{u}_{mi}$ should be equal. If Eqn. (14) provides a good description of performance then

$$\begin{aligned}\mathbf{M}_o(\mathbf{D}_{m1}\mathbf{q}_{mi}-\mathbf{q}_{m1}) &= \mathbf{M}_o(\mathbf{D}_{t1}\mathbf{q}_{ti}-\mathbf{q}_{t1})\\ &\Leftrightarrow \mathbf{q}_{mi} = \mathbf{D}_{m1}^{-1}(\mathbf{D}_{t1}\mathbf{q}_{ti}-\mathbf{q}_{t1}+\mathbf{q}_{m1})\\ &\Leftrightarrow \mathbf{q}_{mi} = \mathbf{D}_{tm}\mathbf{q}_{ti}-\mathbf{q}_{tm}\end{aligned} \tag{16}$$

where $\mathbf{D}_{tm} = \mathbf{D}_{m1}^{-1}\mathbf{D}_{t1}$ and $\mathbf{q}_{tm} = \mathbf{D}_{m1}^{-1}(\mathbf{q}_{t1}-\mathbf{q}_{m1})$. This prediction may be checked by finding the diagonal matrix $\mathbf{D}_{tm}$ and vector $\mathbf{q}_{tm}$ that provide the best fit to the data and evaluating the quality of the fit. Tests of this sort indicate that the cone specific affine model accounts for much of the variance in asymmetric matching data, both for the disk annulus configuration (Wandell 1995, 1999) and for more complex stimuli (Brainard and Wandell 1992). Nonetheless, there are clear instances for which Eqn. (16) does not give a complete account of asymmetric matching (e.g., Delahunt and Brainard 2000) and other color appearance data (e.g., Webster 1996, Mausfeld 1998, D'Zmura and Singer 1999).

The cone-specific affine model may also be tested against psychophysical data on the detection and discrimination of colored lights. Here again the model provides a reasonable point of departure but fails in detail (e.g., Eskew et al. 1999).

To extend the cone specific affine model, various theorists have suggested the need for adaptation at a second site (after signals from separate cone classes have been combined) and for the inclusion of nonlinearities in the relation between $\mathbf{q}$ and $\mathbf{u}$ (see references cited in the previous two paragraphs). An additional open question concerns how the entries of $\mathbf{D}_1$ and $\mathbf{q}_1$ are determined by the viewing context (e.g., Brainard and Wandell 1992, Delahunt and Brainard 2000).

4. *Color Constancy*

The discussion so far has focussed on how the visual system represents and processes the spectrum of light that enters the eye. This is natural, since light is the proximal stimulus that initiates color vision. On the other hand, we use color primarily to name objects. The spectrum of the light reflected to the eye from an object depends both on an intrinsic property of the object, its surface reflectance function, and on extrinsic factors, including the spectral power distribution of the illuminant and how the object is oriented relative to the observer.

Given that the light reflected to the eye varies with the illuminant and viewing geometry, how is it that color is a useful psychological property of objects? The answer is that the visual system processes the retinal image to stabilize the color appearance of objects across changes extrinsic to the object (e.g., changes in the spectrum of the illuminant). This stabilization process is called color constancy.

Color constancy is closely linked to the phenomenon of adaptation described above (Maloney 1999). Indeed, quantitative models of color constancy generally incorporate the same idea that underlies mechanistic models of visual processing: at some central site there is a visual representation $\mathbf{u}$ that correlates with color appearance. To stabilize this representation against changes in illumination, it is supposed that the relation between the quantal absorption rates $\mathbf{q}$ elicited by the light reflected from an object and the visual representation $\mathbf{u}$ depends on the scene in which the object is viewed. In the case of color constancy, the emphasis has been on how the visual system processes the retinal image so that the transformation between $\mathbf{q}$ and $\mathbf{u}$ has the effect of compensating for the variation in the light reflected to the eye caused by changes of illumination and viewing geometry. Psychophysical data on the color appearance of objects viewed under different illuminants are often well-modeled by transformations consistent with Eqn. (14) (e.g., Brainard and Wandell 1992).

The central theoretical question of color constancy is how the visual system can start with image data and factor it into an accurate representation of the surfaces and illuminants in the scene. This question has received extensive treatment, at least for simple scenes. A brief introduction to this literature on computational color constancy follows.

4.1 *Computational Color Constancy*

Consider a scene consisting of diffusely illuminated flat, matte surfaces. For such scenes, the spectrum $\mathbf{b}$ reflected to the eye from each surface is given by the wavelength-by-wavelength product of the spectral power distribution of the illuminant $\mathbf{e}$ and the surface reflectance function $\mathbf{s}$. The surface reflectance function specifies, at each sample wavelength, the fraction of

incident light reflected to the eye. The information about **b** coded by the visual system is its cone coordinates, which may be computed as

$$\mathbf{q} = \mathbf{S}\mathbf{b} = \mathbf{S}\ \mathrm{diag}(\mathbf{e})\mathbf{s} \tag{17}$$

where the function **diag**( ) returns a square diagonal matrix with its argument placed along the diagonal. Clearly **e** and **s** are not uniquely determined from knowledge of **q**: without additional constraints the color constancy problem is underdetermined. Fortunately the spectra of naturally occurring illuminants and surfaces are not arbitrary. Although the physical processes that constrain these spectra are not well understood, analyses of measurements of both illuminants and surfaces shows that their spectra are well described by small-dimensional linear models (see Brainard 1995, Maloney 1999).

Consider surface reflectances. It is possible to define three fixed basis functions so that naturally occurring surface reflectances are reasonably well approximated by a linear combination of these basis functions. Thus for any surface, we have

$$\mathbf{s} \approx w_{s1}\mathbf{b}_{s1} + w_{s2}\mathbf{b}_{s2} + w_{s3}\mathbf{b}_{s3} \tag{18}$$

where $\mathbf{b}_{s1}$, $\mathbf{b}_{s2}$, and $\mathbf{b}_{s3}$ are the spectra of the basis functions and w_{s1}, w_{s2}, and w_{s3} are scalar weights that provide the best approximation of **s** within the linear model. Eqn. (18) may be rewritten as

$$\mathbf{s} \approx \mathbf{B}_s\mathbf{w}_s \tag{19}$$

where the three columns of matrix $\mathbf{B}_s$ contain the basis functions and the vector w_s contains the scalar weights.

When the surface reflectance functions lie within a three-dimensional linear model Eqn. (17) may inverted, once an estimate $\hat{\mathbf{e}}$ of the illuminant has been obtained (see below for discussion of illuminant estimation.) Start by rewriting Eqn. (17) as:

$$\mathbf{q} = [\mathbf{S}\ \mathrm{diag}(\hat{\mathbf{e}})\mathbf{B}_s]\ \mathbf{w}_s = \mathbf{L}_{\hat{\mathbf{e}}}\mathbf{w}_s \tag{20}$$

where $\mathbf{L}_{\hat{\mathbf{e}}}$ is a three-by-three matrix that depends on the illuminant estimate. This matrix may be inverted using standard methods to yield an estimate of $\mathbf{w}_s$:

$$\hat{\mathbf{w}}_s = \mathbf{L}_{\hat{e}}^{-1}\ \mathbf{q} \tag{21}$$

The estimate may then be used together with Eqn. (19) to estimate the surface reflectance function.

Many computational color constancy algorithms assume a linear model constraint for surface reflectance functions. This reduces the constancy problem to finding an estimate of the illuminant to plug into Eqn. (20). For illustrative purposes, an algorithm due to Buchsbaum (1980) is described here. In Buchsbaum's algorithm, two additional assumptions are added. The first is that a three-dimensional linear model also describes illuminant spectral power distributions, so that

$$\mathbf{e} \approx \mathbf{B}_e\mathbf{w}_e \tag{22}$$

The second is that the spatial average of the surface reflectance functions ($\bar{\mathbf{s}}$) is the same in all scenes and known. These additional constraints imply that

$$\bar{\mathbf{q}} = [\mathbf{S}\ \mathrm{diag}(\bar{\mathbf{s}})\mathbf{B}_e]\hat{\mathbf{w}}_e = \mathbf{L}_{\bar{s}}\hat{\mathbf{w}}_e \tag{23}$$

where $\bar{\mathbf{q}}$ is the spatial average of the quantal absorption rates and $\mathbf{L}_{\bar{s}}$ is a known three-by-three matrix. Inverting Eqn. 23 yields an estimate for the illuminant $\hat{\mathbf{e}} = \mathbf{B}_e\mathbf{w}_{\hat{e}}$. This estimate is then used to provide the matrix $\mathbf{L}_{\hat{e}}^{-1}$ to be used in Eqn. (21).

Buchsbaum's algorithm shows how the addition of appropriate assumptions allows solution of the computational color constancy problem. The difficulty with Buchsbaum's algorithm is that its assumptions are too restrictive. In particular, it seems unlikely that the spatial average of surface reflectances is constant across scenes, nor do real scenes consist of diffusely illuminated flat, matte surfaces. Subsequent work has focused on ways to provide reasonable estimates of the illuminant and surface reflectances under other sets of assumptions (e.g., Maloney 1999).

4.2 Computational Color Constancy and Human Performance

How does the computational work relate to human performance? This question has not yet been resolved, but it seems appropriate to close with a few observations. First, the estimated linear model weights of Eqn. (21) may be associated with the mechanism responses **u** discussed in Sect. 2. In both types of theory, these quantities represent the visual response computed from the quantal absorption rates, and both are meant to allow direct prediction of appearance. In the mechanistic approach, one considers a series of transformations whose form is derived from experiments with simple stimulus configurations. In the computational approach, the form of the transformation is derived from consideration of the problem color vision is trying to solve. In both cases, however, the emphasis is on finding the appropriate parametric form of the transformation and on understanding how the parameters are set as a function of the image data. The connection between the two approaches is discussed in more detail by Maloney (1999).

The value of the computational approach to understanding human vision depends on how accurately the transformations it posits may be used to predict the appearance of stimuli measured in psychophysical experiments. There have been only a few empirical comparisons of this sort to date. These comparisons do, however, indicate that the computational approach shows promise for advancing our understand-

ing of human color vision (Bloj, Kersten, and Hurlbert 1999, Brainard, Kraft, and Longre 2001).

See also: Color Vision; Psychophysical Theory and Laws, History of; Psychophysics; Vision, Low-level Theory of; Vision, Psychology of; Visual Perception, Neural Basis of; Visual System in the Brain

Bibliography

Abramov I, Gordon J 1994 Color appearance: on seeing red—or yellow, or green, or blue. *Annual Review of Psychology* **45**: 451–85

Bloj M G, Kersten D, Hurlbert A C 1999 Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature* **402**: 877–9

Brainard D H 1995 Colorimetry. In: Bass M (ed.) *Handbook of Optics: Volume 1. Fundamentals, Techniques, and Design*. McGraw-Hill, New York, pp. 26.1–26.54

Brainard D H, Brunt W A, Speigle J M 1997 Color constancy in the nearly natural image. 1. Asymmetric matches. *Journal of the Optical Society of America A* **14**: 2091–110

Brainard D H, Kraft J M, Longère P 2001 Color constancy: developing empirical tests of computational models. In: Mausfeld R, Heyer D (eds.) *Colour Perception: From Light to Object*. Oxford University Press, Oxford, UK

Brainard D H, Wandell B A 1992 Asymmetric color-matching: How color appearance depends on the illuminant. *Journal of the Optical Society of America A* **9**(9): 1433–48

Buchsbaum G 1980 A spatial processor model for object colour perception. *Journal of the Franklin Institute* **310**: 1–26

D'Zmura M, Singer B 1999 Contrast gain control. In: Gegenfurtner K, Sharpe L T (eds.) *Color Vision: From Genes to Perception*. Cambridge University Press, Cambridge, UK, pp. 369–85

Dacey D M 2000 Parallel pathways for spectral coding in primate retina. *Annual Review of Neuroscience* **23**: 743–75

Delahunt P B, Brainard D H 2000 Control of chromatic adaptation: Signals from separate cone classes interact. *Vision Research* **40**: 2885–903

Eskew R T, McLellan J S, Giulianini F 1999 Chromatic detection and discrimination. In: Gegenfurtner K, Sharpe L T (eds.) *Color Vision: From Genes to Perception*. Cambridge University Press, Cambridge, UK, pp. 345–68

Hurvich L M, Jameson D 1957 An opponent-process theory of color vision. *Psychological Review* **64**(6): 384–404

Kaiser P K, Boynton R M 1996 *Human Color Vision*, 2nd edn. Optical Society of America, Washington, DC

Maloney L T 1999 Physics-based approaches to modeling surface color perception. In: Gegenfurtner K, Sharpe L T (eds.) *Color Vision: From Genes to Perception*. Cambridge University Press, Cambridge, UK, pp. 387–416

Mausfeld R 1998 Color perception: From Grassman codes to a dual code for object and illumination colors. In: Backhaus W G K, Kliegl R, Werner J S (eds.) *Color Vision—Perspectives from Different Disciplines*. Walter de Gruyter, Berlin, pp. 219–50

Neitz M, Neitz J 2000 Molecular genetics of color vision and color vision defects. *Archives of Ophthalmology* **118**: 691–700

Rodieck R W 1998 *The First Steps in Seeing*. Sinauer, Sunderland, MA

Sharpe L T, Stockman A, Jagle H, Nathans J 1999 Opsin genes, cone photopigments, color vision, and color blindness. In: Gegenfurtner K, Sharpe L T (eds.) *Color Vision: From Genes to Perception*. Cambridge University Press, Cambridge, UK, pp. 3–51

Stockman A, Sharpe L T 1999 Cone spectral sensitivities and color matching. In: Gegenfurtner K, Sharpe L T (eds.) *Color Vision: From Genes to Perception*. Cambridge University Press, Cambridge, UK, pp. 53–87

Wandell B A 1995 *Foundations of Vision*. Sinauer, Sunderland, MA

Wandell B A 1999 Computational neuroimaging: color representations and processing. In: Gazzaniga M (ed.) *The New Cognitive Neurosciences*, 2nd edn. MIT Press, Cambridge, MA, pp. 291–303

Webster M A 1996 Human colour perception and its adaptation. *Network: Computation in Neural Systems* **7**: 587–634

Wyszecki G, Stiles W S 1982 *Color Science—Concepts and Methods. Quantitative Data and Formulae*, 2nd edn. John Wiley, New York

D. H. Brainard

Combinatorial Data Analysis

Combinatorial data analysis (CDA) refers to a class of methods for the study of relevant data sets in which the arrangement of a collection of objects is the absolutely central concept. Characteristically, CDA is involved with either: (a) the identification of arrangements that are optimal for a specific representation of a given data set, and where such an exploratory process is typically carried out according to some specific loss or merit function that guides a combinatorial search over a domain of possible structures constructed from the constraints imposed by the particular representation selected; or (b) a confirmatory determination as to whether a specific object arrangement given *a priori* reflects the observed data, and where such a confirmatory process is typically operationalized by comparing the empirically observed degree of correspondence between some given data set and the specific structure conjectured for it, to a reference distribution constructed from the collection of all possible structures of the same form that could have been conjectured.

The boundaries of what CDA might encompass are somewhat open but generally we would exclude methods based on the postulation of strong stochastic models and their specific unknown parametric structures as underlying a given data set. Although CDA might use or empirically construct various weighting functions, the weights so obtained are not to be interpreted as parameter estimates in some presumed stochastic model viewed in turn as responsible for generating the data. Manifest data are emphasized solely, and the traditional concern for an assumed relationship between the data and a restrictively parameterized stochastic model is avoided. For

monograph-length reviews for both exploratory and confirmatory CDA, along with extensive bibliographic sources, see Hubert (1987), Arabie and Hubert (1992, 1996), Hubert et al. (in press).

Within each of the two sections to follow on confirmatory and exploratory combinatorial data analysis, we present several representative areas of application, from among many that could have been given, and hope that by induction, a reasonable characterization can be constructed for both confirmatory CDA (CCDA) and exploratory CDA (ECDA). At the least, CCDA includes most (if not all) of classical nonparametric statistics as a special case, and all extensions or generalizations that rely on randomization (or permutation) arguments (see, for example, Edgington 1995, for a variety of elegant experimental design applications). Similarly, and with not much of a stretch, most of scaling and seriation (multidimensional and otherwise), clustering (partitioning, hierarchical, and in the construction of additive trees and other network models), or more general representational structures (e.g., approximation by sums of (anti-)Robinson matrices—see Hubert and Arabie 1994) can be encompassed within ECDA.

1. *Confirmatory Combinatorial Data Analysis (CCDA)*

CCDA seeks a comparison of some given structure to a reference distribution constructed from all possible similar structures (and with typically the hope for the rejection of the null conjecture of a random assignment based on the extreme size of some index such as $I_{\text{obs}}(U, V)$ and its resulting small p-value). To introduce a general notational structure that can be used to present several ECDA exemplars, we first assume our interest is in some object set $S = \{O_1, O_2, \ldots, O_n\}$ (for example, S may contain n subjects), and we denote by $U = \{u_1, u_2, \ldots, u_n\}$ and $V = \{v_1, v_2, \ldots, v_n\}$ the two sequences of information we have on the objects. Thus, for object O_i, u_i and v_i might denote single numerical values on two variables, or possibly two sets of nominal scale labels, or even vectors of numerical information over a set of attributes (among other entities that would need to be explicated for each particular application). By various means, an index $I_{\text{obs}}(U, V)$ is obtained between the two sequences (where 'obs' denotes 'observed' and is used as a subscript to denote the particular one-to-one matching of u_i and v_i for $1 \leqslant i \leqslant n$). The relative size of $I_{\text{obs}}(U, V)$ is evaluated against all possible $n!$ indices that could be formed by permutating the u_i against the fixed v_i (for all i) or conversely (an index formed by some such permutation will be denoted simply by $I(U, V)$, and without a subscript of 'obs'). If $I_{\text{obs}}(U, V)$ is sufficiently extreme with respect to this reference distribution (i.e., if we have a small probability or p-value of obtaining an index as or more extreme than $I_{\text{obs}}(U, V)$ under the random matching of the entries within U and V), one could declare a significant association between the information provided by U and V for the n objects in S.

Although complete enumeration and subsequent tabling may be the most direct means for constructing a reference distribution for $I(U, V)$ (and against which the specific value of $I_{\text{obs}} \equiv I_{\text{obs}}(U, V)$ can be compared), in practice and because n may be so enormous for even moderately-sized values of n, various approximations must be used. For example, a distribution based only on a random sample could be considered close to what would be generated from complete enumeration (and used to obtain what can be referred to as a Monte Carlo p-value—see Hope 1968); or closed form formulae may be available for the first few moments of the complete distribution (such as the mean and variance), and an approximation based on these may be possible (see Mielke 1979, Hubert 1987, for extensive discussions of such approximations in a variety of CCDA cases).

1.1 *The Linear Assignment (LA) Model*

The LA model assumes a particular form for I_{obs} based on a given $n \times n$ matrix $\mathbf{R} = \{r_{ij}\}$, where r_{ij} denotes some measure of proximity derived or computed from the observed data values, u_i and v_j, i.e., $I_{\text{obs}}(U, V) = \sum_{i=1}^{n} r_{ii}$. The mean and variance of $I \equiv I(U, V)$ are known under the random assignment conjecture:

$$E(I) = \frac{1}{n}\sum_{i,j} r_{ij}$$

$$\text{Var}(I) = \frac{1}{n(n-1)}\left[\frac{1}{n}\left(\sum_{i,j} r_{ij}\right)^2 - \sum_j\left(\sum_i r_{ij}\right)^2 - \sum_i\left(\sum_j r_{ij}\right)^2 + n\sum_{i,j} r_{ij}^2\right] \quad (1)$$

Thus, assuming the appropriateness of a normal approximation (see Hubert 1987, pp. 8–12), $(I_{\text{obs}} - E(I))/\sqrt{V(I)}$ might be compared to a standard normal random variable to obtain the p-value that I_{obs} would engender.

We consider three applications of the LA model:

(a) If U and V denote two numerical sequences (e.g., rank orders or interval scales) for the objects in S, and we define the proximity

$$r_{ij} = \frac{(u_i - \bar{u})(v_j - \bar{v})}{\sqrt{\frac{1}{n}\sum_i (u_i - \bar{u})^2}\sqrt{\frac{1}{n}\sum_i (v_i - \bar{v})^2}} \quad (2)$$

where $\bar{u}$ and $\bar{v}$ are the means within U and V respectively, then I_{obs} reduces to the usual Pearson product-moment correlation coefficient. Thus, $E(I) =$

0 and $\mathrm{Var}(I) = 1/(n-1)$ provide the usual randomization moments, and in turn, a permutation test for a Pearson correlation is obtained as a special case of the LA model.

(b) If, however, the sequences U and V contain the consecutive integers from 1 to K that are used as nominal scale labels, and

$$r_{ij} = \begin{cases} 1 & \text{if } u_i = v_j \\ 0 & \text{if } u_i \neq v_j \end{cases} \tag{3}$$

then I_{obs} is the number of agreements between U and V in labels over the n objects in S. Here,

$$E(I) = n \sum_{k=1}^{K} \left[\left(\frac{n_k^{(u)}}{n} \right) \left(\frac{n_k^{(v)}}{n} \right) \right] \tag{4}$$

is the expected number of agreements under random assignment, where $n_k^{(u)}$ and $n_k^{(v)}$ denote the number of labels having the value k within the U and V sequences, respectively. Using $E(I)$ and $\mathrm{Var}(I)$ from its general formula in Eqn. (1) to standardize I_{obs}, a (normal) approximate p-value for the observed number of agreements could be given directly. It is more typical, however, merely to standardize I_{obs} for descriptive purposes based on the general index form of

$$\frac{I_{\text{obs}} - E(I)}{\max(I) - E(I)} \tag{5}$$

to produce Cohen's κ statistic (which for $\max(I) = n$ is bounded above by 1.0 and has expected value 0.0 under random assignment).

(c) As a final example of using the LA model, suppose U is a numerical sequence (e.g., rank order or interval scale) for the objects in S, where $u_1, \ldots, u_{n_1}$ refer to one group (G_1) of n_1 observations and $u_{n_1+1}, \ldots, u_n$ denote a second (G_2) of $n_2 (= n - n_1)$ observations; $V = \{v_i\}$ for $v_i = 1/n_1$, $1 \leqslant i \leqslant n_1$, and $v_i = -1/n_2$, $n_1 + 1 \leqslant i \leqslant n$; and the proximity $r_{ij} = u_i v_j$. Then, $I_{\text{obs}}(U, V) = \bar{u}_{G_1} - \bar{u}_{G_2}$, where $\bar{u}_{G_1}$ and $\bar{u}_{G_2}$ are the means within G_1 and G_2, respectively. Under random assignment, the $n!/n_1!n_2!$ possible allocations of n_1 scores to G_1 (and n_2 to G_2) are equally-likely, and $E(I) = 0$ and $\mathrm{Var}(I) = (n/((n-1)n_1 n_2)) \sum_i (u_i - \bar{u})^2$, where $\bar{u}$ is the mean for the entries in U. Thus, as the usual approximation for the randomization test for two independent samples, based on the difference between the two group means, $I_{\text{obs}}/\sqrt{\mathrm{Var}(I)}$ could be compared to the standard normal distribution to generate the approximate p-value.

1.2 The Quadratic Assignment (QA) Model

The QA model again assumes a specific form for the index $I_{\text{obs}}(U, V)$ but is now based on two given $n \times n$ matrices, $\mathbf{R}^{(u)} = \{r_{ij}^{(u)}\}$ and $\mathbf{R}^{(v)} = \{r_{ij}^{(v)}\}$, where $r_{ij}^{(u)}$ (or $r_{ij}^{(v)}$, respectively) denotes a measure of proximity between u_i and u_j (or v_i and v_j). The index I_{obs} is defined as the raw cross-product, $\sum_{i,j} r_{ij}^{(u)} r_{ij}^{(v)}$, and as before, the first several moments are available under the null conjecture of random assignment. The first moment is simple, $E(I) = (1/(n(n-1))) \sum_{i,j} r_{ij}^{(u)} \sum_{i,j} r_{ij}^{(v)}$; the variance and higher-moments are more involved and will not be reproduced here (see Mantel 1967, for $\mathrm{Var}(I)$ and the first application using the general form of the QA model in the literature).

(a) As a first example of the QA model, suppose U denotes a numerical sequence for the objects in S and V consists of the integers from 1 to K as labels to indicate membership in one of K groups (with n_k denoting the cardinality of group k, $1 \leqslant k \leqslant K$). Letting

$$r_{ij}^{(u)} = (u_i - u_j)^2 \quad \text{for } 1 \leqslant i, j \leqslant n$$
$$r_{ij}^{(v)} = \begin{cases} 1/(2n_k) & \text{if } v_i = v_j = k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

I_{obs} is then the common within-group sum-of-squares as in one-way ANOVA, and defines one of the common test-statistics used to evaluate the one-way ANOVA conjecture under random assignment. So, evaluating the size of I_{obs} by all $n!$ permutations of the u_i against the v_i for all i, provides the usual randomization test in the one-way ANOVA framework.

(b) As one immediate extension of this one-way ANOVA framework to assess what is called spatial autocorrelation (SA) (see Upton and Fingleton 1985, for an extensive review of assessing SA using the CCDA randomization approach), suppose $\mathbf{R}^{(u)}$ is still defined as above from the numerical sequence in U, but $r_{ij}^{(v)}$ is a distance measure between geographical locations for objects O_i and O_j. Then I_{obs} is an index of the degree to which objects that are close in value in U are also similar in geographical location (in fact, I_{obs} is proportional to what is called Geary's index of SA—see Upton and Fingleton 1985).

(c) As a final example of the QA model, assume both U and V contain nominal scale numerical labels for the objects in S, with U containing the integers from 1 to $K^{(u)}$ (with $n_k^{(u)}$ denoting the cardinality of group k, $1 \leqslant k \leqslant K^{(u)}$), and V containing the integers from 1 to $K^{(v)}$ (with $n_k^{(v)}$ denoting the number in group k, $1 \leqslant k \leqslant K^{(v)}$). If

$$r_{ij}^{(u)} = \begin{cases} +1 & \text{if } u_i = u_j \\ -1 & \text{if } u_i \neq u_j \end{cases}$$
$$r_{ij}^{(v)} = \begin{cases} +1 & \text{if } v_i = v_j \\ -1 & \text{if } v_i \neq v_j \end{cases} \tag{7}$$

then I_{obs} is twice the difference between the number of agreements (A) and disagreements (D) in the classifi-

cation of all object pairs by the two separate partitions represented by U and V. Explicitly, an agreement for an object pair is where the labels for the pair are identical in U and identical in V, or different in U and different in V; a disagreement occurs when the labels are the same in U (or V) but different in V (or U). What has now become known as Rand's Adjusted Index (using the general form for a descriptive statistic already noted for Cohen's κ, as discussed in detail by Hubert and Arabie 1985) can be given as

$$\frac{I_{\text{obs}}(U, V) - E(I)}{\max(I) - E(I)} \tag{8}$$

where $I_{\text{obs}}(U, V) = 2(A - D)$, $\max(I) = n(n-1)$, and

$$E(I) = \frac{1}{n(n-1)} \left[2 \sum_{k=1}^{K^{(u)}} (n_k^{(u)})^2 - (n+1)n \right] \times \left[2 \sum_{k=1}^{K^{(v)}} (n_k^{(v)})^2 - (n+1)n \right] \tag{9}$$

2. *Exploratory Combinatorial Data Analysis (ECDA)*

In contrast to CCDA, ECDA generally tries to optimize such an index and in the process to identify a structure that best represents a given data set. As the exemplars that follow will show, the data analysis tasks arising in ECDA usually can be given some type of combinatorial characterization that involves the identification of object groupings, partitions, or sequences of various types.

2.1 *Sequencing/Seriation/Unidimensional Scaling*

One of the most common tasks encountered in ECDA is the ordering of a set of objects (such as S) along a continuum so that the available data are well depicted by the identified ordering. Commonly, the given data are in the form of an $n \times n$ proximity matrix, say $\mathbf{Q} = \{q_{ij}\}$, whose entries we presently assume are non-negative and symmetric ($q_{ij} = q_{ji} \geqslant 0$ for $1 \leqslant i, j \leqslant n$, and without loss of generality, $q_{ii} \equiv 0$ for $1 \leqslant i \leqslant n$), and keyed as dissimilarities so that larger values correspond to the more dissimilar pairs of objects. The canonical ECDA task in this case is to find some reordering of the rows and simultaneously the columns of $\mathbf{Q}$ so that ideally the (so-called anti-Robinson) gradient structure or pattern is apparent in the reordered matrix, which would generally be reflected by an increase of entries moving away from the main diagonal within any row or within any column. More formally, we might ask to identify a permutation $\rho^*(\cdot)$ of the first n integers (so $\rho^*(i) = j$ for $1 \leqslant i, j \leqslant n$) that when used to relabel the rows/columns of $\mathbf{Q}$ as $\mathbf{Q}_{\rho^*} \equiv \{q_{\rho^*(i)\rho^*(j)}\}$, the following order conditions generally hold within $\mathbf{Q}_{\rho^*}$ (see Hubert and Arabie 1994):

$$\begin{aligned} q_{\rho^*(i)\rho^*(j)} &\leqslant q_{\rho^*(i)\rho^*(j+1)} \quad \text{for } 1 \leqslant i < j \leqslant n-1 \\ q_{\rho^*(i)\rho^*(j)} &\leqslant q_{\rho^*(i)\rho^*(j-1)} \quad \text{for } 2 \leqslant j < i \leqslant n \end{aligned} \tag{10}$$

There are a number of ways to operationalize the search for such a $\rho^*(\cdot)$:

(a) one of the best-known optimization problems in the Operations Research literature, the QA task (that we also met briefly in an earlier confirmatory context), can be used indirectly to find a 'good' $\rho^*(\cdot)$. Explicitly, given two $n \times n$ matrices, we might seek that row and column permutation, $\rho^*(\cdot)$, to maximize the QA cross-product index, $\sum_{i,j} q_{\rho^*(i)\rho^*(j)}(|i-j|)$, between the reordered matrix and a second target matrix $\{|i-j|\}$ defining an $n \times n$ matrix of distances between equally-spaced positions along a continuum. Conceptually, the n objects are placed at these equally-spaced positions, and (because the entries in $\mathbf{Q}$ and $\{|i-j|\}$ are fixed) the usual correlation coefficient is maximized between the distances along the continuum and the proximities between the objects placed at these positions. (For a review of computational methods for approaching the QA task in such a data analytic framework, see Hubert and Schultz 1976.)

(b) Alternatively, instead of equally spacing objects in S along a continuum, a natural generalization for identifying $\rho^*(\cdot)$ would be through unidimensional scaling. Here, we would attempt to estimate an appropriate set of locations, in addition to identifying an optimal permutation $\rho^*(\cdot)$, by minimizing, for example, the least-squares loss criterion of

$$\sum_{i,j} (q_{\rho^*(i)\rho^*(j)} - |x_i - x_j|)^2 \tag{11}$$

where $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$, and $\frac{1}{n}\sum_{i=1}^{n} x_i = 0$. So long as q_{ij} is strictly positive for all i and j when $i \neq j$, we may equivalently maximize $\sum_i (\sum_j q_{\rho^*(i)\rho^*(j)})^2$, and let the optimal coordinates be estimated as $\hat{x}_i = \frac{1}{n}(\sum_{k=1}^{i-1} q_{\rho^*(i)\rho^*(k)} - \sum_{k=i+1}^{n} q_{\rho^*(i)\rho^*(k)})$. (For suggested combinatorial optimization methods for the unidimensional scaling task, see Defays 1978, (for a branch-and-bound strategy) and Hubert and Arabie 1986 (for a dynamic programming approach).)

(c) Finally, and because satisfying the gradient conditions in Eqn. (10) might be considered only an indirect optimization goal in QA applied to unidimensional scaling, a more straightforward approach would directly attempt to optimize some index based on the violation/nonviolation of such conditions within $\mathbf{Q}_{\rho}^{*}$. This strategy has been suggested informally for decades (e.g., see Sibson 1971), and more recently using methods of global optimization based on dynamic programming in Hubert et al. (in press). Explicitly, we seek to maximize the difference between the

agreements and disagreements to the anti-Robinson gradient conditions in Eqn. (10) by maximizing

$$\sum_{1 \leqslant i < j \leqslant n-1} f(q_{\rho^*(i)\rho^*(j)}, q_{\rho^*(i)\rho^*(j+1)}) + \sum_{2 \leqslant j < i \leqslant n} f(q_{\rho^*(i)\rho^*(j)}, q_{\rho^*(i)\rho^*(j-1)})$$

where for $1 \leqslant i < j \leqslant n-1$,

$$f(q_{\rho^*(i)\rho^*(j)}, q_{\rho^*(i)\rho^*(j+1)}) = \begin{cases} +1 & \text{if } q_{\rho^*(i)\rho^*(j)} < q_{\rho^*(i)\rho^*(j+1)} \\ -1 & \text{if } q_{\rho^*(i)\rho^*(j)} > q_{\rho^*(i)\rho^*(j+1)} \\ 0 & \text{if } q_{\rho^*(i)\rho^*(j)} = q_{\rho^*(i)\rho^*(j+1)} \end{cases}$$

and for $2 \leqslant j < i \leqslant n$,

$$f(q_{\rho^*(i)\rho^*(j)}, q_{\rho^*(i)\rho^*(j-1)}) = \begin{cases} +1 & \text{if } q_{\rho^*(i)\rho^*(j)} < q_{\rho^*(i)\rho^*(j-1)} \\ -1 & \text{if } q_{\rho^*(i)\rho^*(j)} > q_{\rho^*(i)\rho^*(j-1)} \\ 0 & \text{if } q_{\rho^*(i)\rho^*(j)} = q_{\rho^*(i)\rho^*(j-1)} \end{cases}$$

2.2 Clustering/Partitioning/Hierarchical Clustering

Another very common task in ECDA is some variation on the theme of grouping objects in S based on a given proximity matrix $\mathbf{Q}$, where generally the groups formed should be internally cohesive and externally isolated or separated from the other groups. Depending on how the task is phrased, a number of variations are possible. One of the broadest interpretations of clustering would be to represent as well as possible a proximity matrix $\mathbf{Q}$ by the additive combination of the attributes defined by some to-be-identified collection of K possibly overlapping clusters or subsets of S. If for convenience, $\mathbf{Q}$ is keyed as a non-negative similarity matrix so large entries now denote the more similar pairs of objects, our objective would be to estimate a non-negative weight for each of the identified clusters and attempt a 'close' reconstruction of each proximity, say q_{ij}, by adding the weights for those subsets that include objects O_i and O_j (in addition to an overall estimated additive constant as well—a more detailed discussion of this additive clustering model can be found in Shepard and Arabie 1979). More mundane manifestations of clustering can be found in methods implemented in commercial software such as SPSS or SYSTAT, e.g., various strategies for identifying a single partition of an object set (such the k-means method) or complete partition hierarchies defined by collections of nested partitions that successively contain fewer and fewer classes (such as the well known agglomerative complete-link method).

Bibliography

Arabie P, Hubert L J 1992 Combinatorial data analysis. *Annual Review of Psychology* **43**: 169–203

Arabie P, Hubert L J 1996 An overview of combinatorial data analysis. In: Arabie P, Hubert L J, DeSoete G (eds.) *Clustering and Classification*. World Scientific, Singapore, pp. 5–63

Defays D 1978 A short note on a method of seriation. *British Journal of Mathematical and Statistical Psychology* **31**: 49–53

Edgington E S 1995 *Randomization Tests*, 3rd rev. edn. M. Dekker, New York

Hope A C A 1968 A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society*, *Series B* **30**: 582–98

Hubert L J 1987 *Assignment Methods in Combinatorial Data Analysis*. M. Dekker, New York

Hubert L J, Arabie P 1985 Comparing partitions. *Journal of Classification* **2**: 193–218

Hubert L J, Arabie P 1986 Unidimensional scaling and combinatorial optimization. In: de Leeuw J, Heiser W, Meulman J, Critchley F (eds.) *Multidimensional Data Analysis*. DSWO Press, Leiden, Germany, pp. 181–96

Hubert L J, Arabie P 1994 The analysis of proximity matrices through sums of matrices having (anti-)Robinson forms. *British Journal of Mathematical and Statistical Psychology* **47**: 1–40

Hubert L J, Schultz J 1976 Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology* **29**: 190–241

Hubert L J, Arabie P, Meulman J 1997 Linear and circular unidimensional scaling for symmetric proximity matrices. *British Journal of Mathematical and Statistical Psychology* **50**: 253–84

Hubert L, Arabie P, Meulman J (in press) *Combinatorial Data Analysis: Optimization by Dynamic Programming*. SIAM, Philadelphia

Mielke P W 1979 On asymptotic non-normality of null distributions of MRPP statistics. *Communications in Statistics: Theory and Methods* **A8**: 1541–50

Mantel N 1967 The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**: 209–20

Shepard R N, Arabie P 1979 Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review* **86**: 87–123

Sibson R 1971 Some thoughts on sequencing methods. In: Hodson F R, Kendall D G, Tăutu P (eds.) *Mathematics in the Archaeological and Historical Sciences*. Edinburgh University Press, Edinburgh, UK, pp. 263–6

Upton J G, Fingleton B 1985 *Spatial Data Analysis by Example: Point Pattern and Quantitative Data*. Wiley, New York

L. Hubert and P. Arabie

Commemorative Objects

Commemoration refers to the mobilization of symbols to awaken ideas and feelings about the past. The word commemoration derives from Latin (*com* (together) + *memorare* (to remember): to remember together). Remembering together occurs in every sphere of life—family, association, occupation, community, and society—but its dynamics are most visible in the

largest, societal, sphere. Commemoration is society's moral memory, distinguishing events its members believe to be deserving of remembrance from those deserving of being merely recorded. Commemoration promotes society's sense of itself by affirming its members' mutual affinity and identity.

1. Commemorative Symbolism

Commemoration is based on the premise that historical meanings can be represented by means of objects and rituals, including scripture, icons, music, monuments, shrines, naming conventions, and ritual observances. These 'cultural objects,' embodying the meaning of the past in physical form (Griswold 1987), cover the social landscape:

(a) Commemorative scripture includes biographies, eulogies, poems, plays, and commentaries. Defining the moral rather than factual aspects of the past, commemorative texts are hagiographic and thus distinguishable from analytic history and chronicling.

(b) Music, consisting of anthems, hymns, and inspirational songs, aligns melodic and harmonic structures and lyrics to historical events in such a way as to dramatize the nature and extent of their significance.

(c) Icons, signs resembling what they represent, commemorate the past by bringing it to visual presence through paintings, statues, prints, photographs, motion picture film, and videotape.

(d) Monuments, including obelisks, antique temples, and war memorials, are designed to elevate the public imagination through grand structures bearing dramatic reference to grand events.

(e) Shrines—including birthplaces, residences, places of death, headquarters—play a key role in commemoration because they bring the individual into direct contact with sacred objects. Shrines are sacred because they are, by definition, 'containers' of relics (*scrinium* (case, chest)) associated with great events or used by great people.

(f) Naming practices make the past ubiquitous by incorporating its memory into the identity of businesses, streets, cities, towns, counties and states, rivers, and mountains.

(g) Commemorative observances (periodic performance of anniversary, centennial, and holiday rites) have been analyzed more extensively than other mnemonic forms, but their functions are no different: to maintain the vitality of the past, to prevent it from being forgotten and to revive what is essential in the collective consciousness. Through commemorative observance:

> the group periodically renews the sentiment which it has of itself and of its unity; at the same time, individuals are strengthened in their social natures. The glorious souvenirs which are made to live again before their eyes, and with which they feel they have a kinship, give them a feeling of strength and confidence; a man is surer of his faith when he sees to how distant a past it goes back and what great things it has inspired (Durkheim [1915] 1965, p. 420).

Commemorative observances, together with scripture, music, icons, monuments, shrines, and naming practices, comprise sign systems representing past events, recommending them as being worthy of contemplation, affording individuals a common legacy and identity by making history tangible and accessible to all.

2. Commemorative Networks

The combination of commemorative symbols conveys meanings that differ from meanings conveyed by any one symbol. To activate the commemorative process through story, song, icon, ceremonial observance, or other symbolic forms, therefore, is to activate memories that include but extend beyond the persons and events commemorated. Commemoration works by constantly pairing symbolic objects with one another, and the significance of historical figures and events change as they are placed in commemorative networks composed of new objects absent in previous networks.

During the Progressive Era (1900–1920), for example, visual images of George Washington were paired with contemporary and historical figures, including Theodore Roosevelt and Woodrow Wilson, with whom he could have had no actual contact. Commemorators identified him as one who would have supported Progressive Era reforms. In Washington's own political values, in fact, they found progressivism's roots. They named him the first Progressive and contrasted him with selfish businessman of their own time. The pairing processes fanned out broadly not only within a network of people whose prestige enhanced and was enhanced by Washington's, but also to networks of events and structures (notably, political and economic reforms and relentless condemnation of plutocracy). In such mnemonic webs, existing beliefs about Washington were organized, their institutional roots deepened, and a new and autonomous collection of beliefs about him accumulated. Common features of people and events with which Washington was symbolically connected were generalized to Washington himself. In this way, Washington changed through the nineteenth century from a conservative symbol of the social and economic status quo in to a liberal symbol of social and economic reform.

Commemoration, as Washington's case suggests, is a structuring process that partially overrides the qualities of its objects and imposes upon them its own content. The key to this process is not the mere act of pairing or coupling one historical actor with another;

it is the attribution to one actor of qualities of other individuals located in the same network of memory. Commemoration is creative because it fills in what is lacking in its objects, whether historical figures or events, by attributing to them the qualities of other figures and events deemed similar.

3. *Commemoration, History, and Collective Memory*

Commemorative symbolism is distinguishable from the historiography on which it is based. History, as French sociologist Maurice Halbwachs ([1950] 1980) defined it, seeks an objective standpoint to assess the sequence, importance, mutual relations, causes, and consequences of past events. It is 'situated external to and above groups' and looks to the past independently of their immediate problems and concerns. Once established, Halbwachs believed, historical knowledge remains stable—its stream of facts and demarcations 'fixed once and for all.' In contrast, commemoration remains significant only so long as it remains relevant (pp. 80–1).

Traditional societies possess no history. History emerges as tradition ends, social memory fades, and commemorative symbolism loses credibility. Early Christian leaders, for example, read historical chronicles, but their followers were illiterate and had no conception of 'historical preoccupations such as we think of them. … Their memories were tied to rites of commemoration and adoration, to ceremonies, feasts, and processions' (Halbwachs [1941] 1992, p. 222).

Pierre Nora, a French historian, shares Halbwachs' belief in the ultimate opposition between history and commemoration. Analytic history, according to Nora, deconstructs the sacred past; the *lieux de mémoire* (sites of memory) reconstruct it. *Lieux de mémoire* are ubiquitous and comprise every type of commemorative sign: not only monuments and shrines but also flags and anthems, sanctuaries and ruins, statues and busts, portraits and history paintings, coins and medallions, holidays and rituals (Nora 1997).

Two paths lead to two different notions of the past. History disenchants the past; commemoration and its sites sanctify it. History makes the past an object of analysis; commemoration makes it an object of commitment. History is a system of 'referential symbols' representing known facts about past events and their sequence; commemoration is a system of 'condensation symbols' (Sapir 1930, pp. 492–3) expressing the moral sentiments these events inspire. History, like science, investigates the world by producing models of its permanence and change. Commemoration, like ideology, promotes commitment to the world by producing symbols of its values and aspirations (Geertz 1973, pp. 193–233). The contrast between history and commemoration is not entirely clear cut. History always reflects the ideals and sentiments that commemoration expresses; commemoration is always rooted in historical knowledge. Commemoration is intellectually compelling, therefore, when it symbolizes values whose past existence history documents; history is morally compelling when it documents events that can be credibly commemorated. Analytic history and commemorative symbolism are interdependent constituents of collective memory.

4. *Theories of Commemoration*

Multiculturalism, postmodernism, and hegemony theory, each prominent in the intellectual environment of the 1980s, shaped the agenda for recent debates over commemoration and history. Whether commemoration is a process of construction or retrieval, contested or shared, a mirror reflecting the concerns of society or a lamp of practical guidance and moral precedent, a system of social attachment or social control—on these and related issues two theories have been brought to bear.

4.1 *The Politics of Commemoration*

In the theory of the politics of memory, commemorative activities appear as contests among differently empowered communities. Producers of commemorative symbolism, ('reputational entrepreneurs' (Fine 1996)), manipulate the attitudes of their audience in the service of society's privileged strata (Gillis 1994). The 'invention of tradition' was as a mode of social control during Europe's democratic revolutions (Hobsbawm and Ranger 1983); state histories and counter histories performed a similar control function in many authoritarian nations, while 'official' and 'vernacular memory' in the United States supported the interests of the elite and the masses respectively (Bodnar 1992). The second strand of the politics of memory literature assumes that power is diffused rather than concentrated and that commemorative objects emerge out of a context of cross-cutting coalitions, networks, and enterprises rather than a ruling elite's interest in maintaining its privilege. The fate of artistic and presidential reputations, Holocaust memories, place-naming and monument-making, and the organization of museums exemplify commemoration's role in pluralistic networks of interest and power (Olick and Robbins 1998).

4.2 *Commemoration as a Cultural System*

The theory of the politics of memory conceives commemoration as a reflection of power distributions, assuming that different elements of the past become

more or less usable as resources of various groups fluctuate. Commemoration thus appears to be made and remade in the service of changing interests and needs. The theory of memory as a culture system, in contrast, conceives commemoration as a pattern of symbolic forms that articulate present experience and perpetuate present understandings of the world. Commemoration, in this theory, is less a construction of a manipulative interest group than a symbolic frame in terms of which people confer meaning on their experience and find direction for their conduct. Reputational entrepreneurs (Fine (1996)), thus, perform more like priests than like manipulators, seeking to reaffirm beliefs they share with their audience. Accordingly, commemoration is not only a model *of* society—a reflection of its needs, interests, fears, mentality, and aspirations—but also a model *for* society—a program defining its values and goals and a frame organizing and assessing the moral significance of its experience. Commemoration affects social reality by reflecting, shaping, and framing it.

The distinction between commemoration's mirroring and modeling functions is analytic, not empirical: both aspects are realized in every act of remembrance. Commemorative objects must reflect current problems before they can program ways to deal with them, for people cannot be oriented by a past in which they fail to see themselves. It is commemoration's relevance as a model, however, that renders it a useful mirror, for people have no reason to look for themselves in a past that does not already orient their lives.

The coexistence of commemoration's modeling and mirroring functions is evident in its use. When Abraham Lincoln ran for the presidency, for example, Republican organizers 'campaigned down' to the people by mirroring their preferences with symbols of railsplitting and the frontier; but when Lincoln assumed the presidency, his supporters, wishing to portray him in a more dignified light, modeled his image after George Washington. They produced portraits of Lincoln gazing at a bust of Washington or standing beneath or next to Washington's portrait in the presence of familiar symbols of state. Lincoln himself launched military offensives on Washington's birthday and regularly compared his situation with Washington's. When Lincoln died, eulogists keyed his memory to that of Washington; mourners placed portraits of Washington beside his coffin and purchased prints showing Washington welcoming him into Glory. People commemorated Lincoln, as they did other actors and events, in ways that affirmed their values and enriched their lives, enabling them to comprehend and cope with a complex world and to render their experience meaningful. Lincoln's image was in turn modified by these new modeling and mirroring connections. Pairings of Washington's and Lincoln's images, as Emile Durkheim ([1915] 1965) would have said, 'do not confine themselves to revealing the mental state with which they are associated: they aid in creating it' (p. 262). Commemorative structures thus conferred on Lincoln's images a meaning it would otherwise not have had.

5. *Negative Commemoration*

Traditional commemoration transmits positive images of the past. Negative commemoration, marking events weighing on the present generation as sources of shame and dishonor and articulating regret for harm inflicted, arose during the late twentieth-century in connection with the new recognition of minorities' entitlements. Every real community, as is now recognized, is a community of memory, one that retells stories about men and women who have exemplified its ideals and values. Such stories include not only tales of shared suffering that create strong bonds of identity but also tales of suffering wrongly inflicted. Defined by these premises are the new commemorative politics of regret, embodied in new American narratives revealing George Washington as slaveholder and Abraham Lincoln as colonizationist wishing to deport the slaves he emancipated, and new shrines, monuments, and museums marking the era of slavery, maltreatment of minorities, and unpopular war (Wagner-Pacifici and Schwartz 1991). As these new commemorative modes reflect the erosion of boundaries that once separated racial, religious, and ethnic groups and once justified aggression within and between nations, they affirm the state's recognition of the dignity of every citizen.

6. *Memory of Commemoration*

Commemorative symbols are neither direct portrayals of the events they commemorate nor products of the concerns of the generations creating them; they are parts of an ongoing process of meaning-making. Since commemorative symbols refer to the memory of their own genre, their meaning includes not only the event being commemorated but also how it has been commemorated elsewhere and in the past. In other words, commemorations are 'path dependent'—shaped by their development as well as their referents. Therefore, differences between commemorations are not just matters of the politics of the present; they reflect the forms and media available at different times. Earlier commemorations affect the availability of resources and forms required for present commemorations (Olick 1999). Conceived in light of their generic predecessors, war memorials can be critical as well as reaffirming. The Korean War Veterans Memorial, as originally conceived by a supervisory committee hostile to the Vietnam Veterans Memorial's antiwar inspiration, had to be uplifting, suggestive of honor, pride, and dignity, evoking admiration and gratitude rather than self-contempt and grief. It had to be utterly different from the Vietnam Memorial in design and message. The Memorial Committee re-

jected all designs failing to fulfill these requirements (Schwartz and Bayma 1999). Comparing commemorations in terms of the history of their genres helps to explain their differences and leads to deeper insights into the various ways society contemplates its past.

7. *From the Politics of Commemoration to the Semiotics of Commemoration*

Pulling together the threads of this summary, we see that knowledge of commemoration grows in spurts—a brace of writings in Europe during the first half of the twentieth century, then a period of quiet, then an eruption of theorizing and research during the last two decades of the century. Emile Durkheim was the first sociologist to study commemoration and commemorative rites systematically, but it was Maurice Halbwachs, Durkheim's student, who formulated the broader concept of collective memory and pioneered its empirical study in the 1920s and 1930s. Like other students of public knowledge and belief, Halbwachs was jaded by the intellectual climate resulting from the First World War's destructiveness. A branch of the sociology of knowledge, the sociology of memory took root in a society where common values had eroded, large groups of people had become alienated from their own traditions as well as one another, and distrust infected all human relations. Halbwachs systematized doubt about the customary frames for understanding the past and the customary symbols for commemorating it.

Still an object of distrust and struggle, commemoration is widely conceived as a mask concealing the interests of the powerful; yet, this conception has new meanings because it is part of a new cultural environment. The 1960s and 1970s were years of disillusionment resembling those through which Europe passed when Halbwachs was writing. It was an era in which 'conflict theories' of social order replaced 'consensus theories,' an era in which traditional accounts of the past, as of everything else, were repudiated. Three overlapping intellectual perspectives—multiculturalism, postmodernism, and hegemony theory—articulate the connection between the cultural revolution of the 1960s and 1970s and contemporary interest in the 'construction' of the past, and these three perspectives still animate the politics of commemoration. During the last decade of the century, however, a new research trend arose, not replacing but superimposing itself upon and insinuating itself into the theory of the politics of commemoration. The 'New Structural Memory,' as Kerwin Klein (2000) calls it, arose in the United States and Europe, emphasizing texts and symbols and de-emphasizing older notions of memory based on the beliefs, attachments, hatreds, and fears of individuals. Michael Schudson's *Watergate in American Memory* (1992) was one of the first efforts to define explicitly The New Structural Memory, suggesting that perception of the past is not, as has been regularly assumed, a property of individual minds but is rather social, located in and shaped by tangible and even countable things—books, records, holidays, statues, souvenirs—in short, a varied and everchanging assemblage of mnemonic objects. 'Structural memory' has, in fact, less to do with structures than with the devices by which commemoration addresses the problem of meaning. Social scientists at the turn of the twenty-first century focus more intensely on this problem—commemoration as a source of cognitive and moral orientation, inspiration, consolation, and purpose—than ever before. The semiotics of commemoration is taking its place beside the politics of commemoration.

8. *Conclusion*

The meaning of commemorative symbols derives from their relationship to other symbols and symbolic patterns. Historians, it is true, also combine previously discrete units of knowledge to produce new understandings; but they do so within a methodological framework that defines proper and improper uses of evidence. In many places (mainly but not only in non-Western nations), historians trained in critical historiography still accede to pressures to produce gratifying images of the past (Shils 1981, pp. 54–62), but even in these nations history and commemoration, like science and ideology, remain different lines of work (Geertz 1973, p. 231): the former combines evidence, however selective or biased, to justify belief; the latter combines symbols to sanctify belief.

Commemoration and history are different but not separate lines of work, which is why they must not be radically distinguished. Commemoration 'requires the support of groups delimited in time and space' while history remains relatively autonomous (Halbwachs [1950] 1980, p. 84); yet commemoration cannot always be dismissed as a distortion of history. Commemoration is based on some invention and much exaggeration, but it is also constrained by the historical record: materials of the 'invented' past include facts as well as biases and interests. In democratic societies, at least, where all commemorative interpretations are open to criticism, constructions of the past must be undertaken within the limits of reality—limits no one can ignore without cost (Schudson 1992, pp. 205–21). Since the commemorative process starts with historical reality and improves it by de-emphasizing the inessential and emphasizing what is necessary to round out the ideal, commemorative networks do not deny history; they valuate history; they identify and lift morally significant events out of the historical continuum and set them apart from and above ordinary ones.

If the relevance of the past should erode, however, what would be the fate of commemoration? The

relevance of the past has diminished during the late twentieth century on two levels. First, the individual's connection to his or her society and its past is weakened by the discrediting of its metanarratives—its great legends and myths. Since the key feature of metanarratives is embodied in a plot structure that interprets events in a way that can be widely understood and shared, people require them, collectively and individually, to make sense of experience. 'Of what great story is my life and death a part?' is the question they ask in their effort to define themselves and the meaning of their lives.

Second, as the stories that society tells about itself are discredited, the orienting past tense of the individual's life is negated. Thus, 'pure and unrelated presents in time' and momentous 'psychohistorical dislocations'—erosion of the bond that once linked men and women to vital symbols of their cultural tradition—are at the core of contemporary temporality (Jameson 1984, pp. 480–1, Lifton 1993, pp. 14–17). That diminished engagement with the past distinguishes the postmodern era is a reasonable claim; however, the adjective 'diminished' must be underscored, for 'diminishings' are more difficult to describe and comprehend than disappearances. The role of commemoration is an ambiguous role defined by a 'residual' pattern of reverence and a 'dominant' pattern of affectively neutral admiration for the past and an 'emergent' pattern of indifference (Williams 1987). At question is whether the residual pattern is destined to disappear or whether it remains because it is not residual at all, but rather because it represents a fundamental requirement of society.

As residual patterns of belief often reflect efforts to renew ties to a past that emergent patterns challenge, we witness in the United States a great surge of monument building, beginning in the 1980s—the very decade in which revisionist history (conceding only the sins of the American narrative to be grand) matured. Thus, the millions who continue to admire historical achievements, seek information about them, and commemorate them, include those driven by currents of reaction against the erosion of the past and of tradition.

See also: Collective Identity and Expressive Forms; Collective Memory, Anthropology of; Collective Memory, Psychology of; Durkheim, Emile (1858–1917); Hegemony: Cultural; Myths and Symbols: Organizational; Power in Society; Ritual; Ritual and Symbolism, Archaeology of; Symbolic Boundaries: Overview; Symbolism (Religious) and Icon

Bibliography

Bodnar J 1992 *Remaking America: Public Memory, Commemoration, and Patriotism in the Twentieth Century*. Princeton University Press, Princeton, NJ

Durkheim E [1915] 1965 *The Elementary Forms of the Religious Life*. Free Press, New York

Fine G 1996 Reputational entrepreneurs and the meaning of incompetence: Melting supporters, partisan warriors, and images of President Harding. *American Jouranl of Sociology* **101**: 1159–93

Geertz C 1973 Ideology as a cultural system. In: *Interpretation of Cultures*. Basic Books, New York, pp. 193–233

Gillis J R (ed.) 1994 *Commemorations: The Politics of National Identity*. Princeton University Press, Princeton, NJ

Griswold W 1987 A methodological framework for the sociology of culture. In: Clogg C (ed.) *Sociological Methodology*. American Sociological Association, Washington, DC, Vol. 17, pp. 1–35

Halbwachs M [1950] 1980 *Collective Memory*. Harper & Row, New York

Halbwachs M [1941] 1992 The legendary topography of the gospels in the Holy Land. In: Coser L A (ed.) *Maurice Halbwachs on Collective Memory*. University of Chicago Press, Chicago, pp. 193–235

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Jameson F R 1984 Postmodernism, or, the cultural logic of late capitalism. *New Left Review* **146**: 53–92

Klein K 2000 On the emergence of memory in historical discourse. *Representations* **Winter**(69): 127–50

Lifton R J 1993 *The Protean Self*. Basic Books, New York

Nora P 1996 *Realms of Memory*. Columbia University Press, New York

Olick J K 1999 Genre memories and memory genres: A dialogical analysis of May 8, 1945. Commemorations of the Federal Republic of Germany. *American Sociological Review* **69**: 381–402

Olick J K, Robbins J 1998 Social memory studies: From collective memory to the historical sociology of mnemonic practices. *Annual Review of Sociology* **24**: 105–40

Sapir E 1930 Symbols. In: Seligman E R (ed.) *Encyclopedia of the Social Sciences, XIV*. Macmillan, New York, pp. 492–5

Schudson M 1992 *Watergate in American Memory*. Basic Books, New York

Schwartz B, Bayma T 1999 Commemoration and the politics of recognition: The Korean War Veterans Memorial. *American Behavioral Scientist* **42**: 946–77

Shils E 1981 *Tradition*. University of Chicago Press, Chicago, IL

Wagner-Pacifici R, Schwartz B 1991 The Vietnam Veterans Memorial: Commemorating a difficult past. *American Journal of Sociology* **97**: 376–420

Williams R 1977 *Marxism and Literature*. Oxford University Press, New York

B. Schwartz

Commercial Banking

Commercial banks traditionally have performed three primary functions: they held time deposits (savings accounts), demand deposits (checking accounts), and they issued credit, primarily to firms but also to individuals. But commercial banks have played other roles as well, from the underwriting of securities to

providing a range of fee-based financial services to corporations. They have also been significant players in a range of national and international business communities, and their degree of power and influence in these communities has been the source of major controversy among scholars. More recently, a series of rapid changes have occurred in the commercial banking world that have raised questions about whether this institution will remain a distinct corporate form. In this article, research by social scientists on commercial banks is examined, focusing on their relations with nonfinancial corporations. Given the significant cross-national variation among banking systems and the legal environments within which they operate, emphasis will be primarily on the United States, although non-US examples will be referred to at various points.

1. *Financial Systems and Commercial Banks*

Zysman (1983) has suggested that there are two types of system into which most of the industrialized capitalist countries fall. In credit-based systems, such as those in France, Germany, and Japan, banks are the primary providers of long-term capital, and are permitted to own and vote stock in nonfinancial corporations. In capital market-based systems, as in the UK and the USA, securities issues (stocks and bonds), handled primarily by investment banks and insurance companies, are the principal sources of long-term debt, while commercial banks specialize in short-term debt. Gershenkron (1962) noted that the capital market-based systems are associated with relatively early industrialization while the credit-based systems are associated with later industrialization. Banks historically have played a particularly dominant role in the credit-based systems. But although there is no comparable period of bank influence in Britain, there have been periods of bank ascendance in the USA.

2. *The Rise of Commercial Banks: The United States, Germany, and Japan*

American commercial banks arose shortly after the Revolution, as businesses began to specialize and international trade expanded (Chandler 1977, pp. 28–31). The number of banks increased rapidly after 1790, to more than 200 by 1815. An important event in the development of American commercial banking occurred in 1832, when President Andrew Jackson refused to support a bill to recharter the Second Bank of the USA, a bank of national scope that had 22 branches located in several states. Three decades later, the National Bank Acts of 1863 and 1864 outlawed branch banking altogether (Roe 1994, pp. 54–9), and it was not until 1927 that American banks were allowed to branch even within states (Roe 1994, pp. 94–5). The resulting, relatively decentralized, commercial banking system is distinct from more centralized systems such as the one that developed in Germany in the late 1800s.

Despite antibranching laws, the leading US commercial banks became increasingly concentrated in the late nineteenth and early twentieth centuries. During this period, corresponding with socialist movements in Europe, and progressive and populist movements in the USA, a major chorus of analysts and critics emerged on both continents. In Europe, Rudolf Hilferding produced a Marxist analysis of what he called 'finance capital,' the domination of the German economy by a handful of leading banks. Bank power in both the USA and Germany emanated from the rapid growth and consolidation of industry, in which companies were increasingly dependent on the financing that the banks provided. The concentration of banking, and the alliances among the banks, made it difficult for nonfinancial corporations to leverage banks against one another, further increasing the banks' power. The American economy was dominated by a relatively small number of financiers who forged constellations of investment and commercial banks, leading industrial firms, and railroads. Louis Brandeis, in his classic *Other People's Money*, focused on investment banks as the centers of power during the period, but investment and commercial banks were not always easy to distinguish. Investment banks were permitted to hold deposits, and commercial banks were permitted to underwrite securities. Members of these alliances sat frequently on the boards of directors of multiple, nominally competing, financial institutions. George F. Baker, an ally of J. P. Morgan, sat on the boards of three banks and an insurance company.

Both German and Japanese banks developed later than those of the USA, beginning in the 1850s in Germany and the 1870s in Japan. By the beginning of the twentieth century, however, a handful of German and Japanese banks had become dominant actors in their respective economies, standing at the head of groups of companies, thus equaling or exceeding the power of American banks. The overwhelming power of the US banks began to ebb after 1912. The Clayton Act of 1914 outlawed director interlocks between competing firms, leading to a sharp decline in formal ties among the banks. The leading financiers of the period either died or retired, and although many transferred control to their sons or hand-picked successors, these second-generation leaders did not possess the power that their elders had enjoyed. Paradoxically, however, although US commercial banks suffered during the Great Depression, investment banks were more seriously affected.

A turning point was the Glass–Steagall Act of 1933, which forced commercial banks and investment banks to separate their activities. Commercial banks were

prohibited from underwriting securities, and investment banks were required to give up their deposits, which had been their primary source of capital. This set the stage for what some observers argue was a resurgence of bank power in the post-World War II period.

Although World War II severely disrupted both nations' economies, German and Japanese banks did not experience a similar decline in power. One reason for this is that, in Germany and Japan banks are permitted to hold equity in nonfinancial firms. Partly as a consequence of this, and partly because there was no equivalent to Glass–Steagall, banks in Germany and Japan have been considerably more involved than US banks in firm policy making. After World War II, the American occupational forces broke up the Japanese *zaibatsu* (business groups) but left the leading banks intact (Gerlach 1992). The German economy recovered quickly after the war and a small number of major banks remained dominant. Although most observers believe that German banks remain more powerful relative to industry than banks in the USA, there is now some controversy over how dominant their role is (Ziegler et al. 1985).

3. *Capital Dependence and Bank Power*

American commercial banking is stratified informally into national, regional, and local sectors. The national banks are led by a relatively small number of major money-center banks located primarily in New York but also in financial centers such as Chicago and San Francisco.

In recent years, as commercial banking has undergone major changes (described below), banks in areas such as Charlotte, North Carolina have emerged as among the largest. The national money-center banks are typically the leaders of major lending operations for the largest nonfinancial corporations. Regional and local banks, which make up the vast majority of US commercial banks, focus primarily on lending to mid-sized and small businesses.

Although the early American commercial banks were involved in long-term lending, US commercial banks generally have specialized in short-term notes of less than one year. Insurance companies have focused on long-term private debt, while investment banks have focused on the placement of long-term public bonds. A study by Go (1999) has shown that American commercial banks became more heavily involved in term loans of 1 to 10 years during the 1930s. Even prior to that, some large banks were involved in bonds as well as loans of up to 5 years. Given commercial banks' role as a source of corporate external financing, a number of scholars have assumed that their power over nonfinancial corporations was a function of the degree to which the latter were dependent on external financing. In their classic work, *The Modern Corporation and Private Property*, Berle and Means (1939) argued that the increasing profitability of large US corporations generally freed them from dependence on banks because firms were able to finance their investments with retained earnings. This view was accepted by the vast majority of social scientists into the early 1970s. The evidence for this argument was equivocal, however. A study by Lintner (1960) suggested that the use of external financing by American firms did not decline between 1925 and 1955. Stearns (1986) showed that the use of external debt financing by USA firms was at relatively low levels in the post-World War II period but that it began to increase around 1965, remaining relatively high into the early 1980s.

The evidence on the increasing use of external financing corresponded with a renewed focus on the power of commercial banks. This came from two sources. First, although unlike banks in Germany and Japan, US commercial banks had been prohibited from directly owning equity in other firms by the National Bank Act of 1863–4, in the post-World War II period they became increasingly active in managing corporate pension funds through their trust departments. As part of their fund management, banks invest in corporate stock, on which they have voting rights. A Congressional investigation in the 1960s led by Representative Wright Patman revealed individual bank trust department holdings of 5 percent or more in nearly 30 percent of the 500 largest USA industrial corporations. Despite some heavily publicized cases, it is unclear whether the banks actually used these holdings as leverage over corporate policy, as is believed to occur outside the USA, especially in Germany. Useem (1996), however, has suggested that institutional stockholders have become increasingly active in voicing their views to managers in recent years.

A second possible source of bank power was corporations' increased use of external financing. Mintz and Schwartz (1985) have argued that, given its value as a generalized resource, banks' control over capital has given them a broad 'hegemony' within the larger business community. Although they are rarely actively involved in company decision making, Mintz and Schwartz suggest that the banks are able to make broad policy decisions about the allocation of capital within the economy. As their evidentiary base, Mintz and Schwartz describe several examples of bank intervention in the affairs of nonfinancial firms. In addition, they show that the major commercial banks were the most central firms in the networks of director interlocks that span the leading US firms, a finding that has been replicated repeatedly across studies and over time (Mizruchi 1996). One could argue that the use of external financing does not ensure automatically that firms are dependent on banks. On the contrary, borrowing contains both tax benefits and, when interest rates are low, cost benefits. A study by

Mizruchi and Stearns (1994) shows, however, that even when controlling for the cost of capital, when retained earnings are high, borrowing tends to decline, and vice versa. This suggests, consistent with the arguments of a number of economists, organizational scholars, and sociologists, that firms will tend to avoid borrowing from banks if they have sufficient amounts of cash.

It is important to note that these same phenomena—reliance on external financing, ability to vote stock, and centrality in corporate interlock networks—have been used to suggest the power of financial institutions in the German and Japanese contexts. Gerlach (1992) argues that although Japanese firms became better able to finance investments with retained earnings in the postwar period, they remain heavily dependent on banks. Zysman (1983) suggests that German banks' control over virtually all sources of external financing, loans, bonds, and equity, along with their high direct equity holdings and willingness to use them gives them a high degree of leverage.

4. *Major Recent Changes in Commercial Banking*

Mintz and Schwartz's argument seemed to fit the German and Japanese cases quite well, but it was intended to apply to the USA. Although its applicability to the American case was always controversial, the argument was consistent with a considerable amount of evidence. Bank involvement in institutional stockholding had increased significantly since the 1950s, nonfinancial corporations' use of external financing had increased since the mid-1960s, and banks remained the most central firms in the corporate network from the early 1900s into the 1980s. Beginning in the early 1980s, however, a series of events occurred that altered the nature of commercial banking radically. Rapid changes in technology and the regulatory environment led corporations to reduce their reliance on banks for capital, as firms turned increasingly to direct borrowing through the issue of commercial paper. By 1994 the value of outstanding commercial paper in the USA equaled that of outstanding debt to commercial banks.

Meanwhile, the proliferation of alternative sources of savings and investment such as mutual, pension, and money market funds led individuals to reduce their deposits in commercial banks. In response to these losses, the largest commercial banks turned increasingly to high-risk lending, leading to several major bank failures and near-failures in the late 1980s and early 1990s. Both the number of commercial banks and the proportion of corporate debt acquired from commercial banks declined by one-third between 1979 and 1994 (Davis and Mizruchi 1999, p. 220). Large commercial banks responded to the loss of their traditional franchise by shifting their focus away from lending (especially domestic lending, which was no longer profitable) and toward financial services such as capital market services, foreign currency exchange, and derivatives. They also participated increasingly in securities underwriting, as legislation and lax enforcement slowly chipped away at the Glass–Steagall provisions. The largest commercial banks came increasingly to resemble investment banks. Bank profits, which had declined sharply in the late 1980s and early 1990s, rebounded by the mid-1990s. Both of these changes in bank strategies have had the effect of altering the social role of commercial banks within the American business community. As Davis and Mizruchi (1999) document, the largest banks reduced substantially the number of executives of major corporations appointed to their boards, thus losing their place at the center of the interlock network.

Two major developments of the late 1990s raise questions about the future of commercial banks as distinct entities. In April 1998, Citibank, the largest US commercial bank, announced a merger with Travelers Insurance, whose units included the investment firm of Salomon Smith Barney. This new megafinancial firm, called Citigroup, became the first corporation since Glass–Steagall openly to blend commercial and investment banking as well as insurance underwriting. In October 1999, the US Congress and the Clinton Administration reached an agreement and repealed the Glass–Steagall Act. The new legislation will allow financial corporations to offer a full range of services, from traditional lending to public stock offerings (*New York Times* October 23, 1999, A1, B4).

5. *The Future of Commercial Banking*

Given the tidal wave of changes occurring in the financial world at the the beginning of the twenty-first century, it is difficult to know whether commercial banks will continue to exist. Both the efficiency and control benefits of combining financial services under a single roof suggest that US commercial banks will move toward mergers with investment banks and other financial institutions, making them more like the 'universal banks' that exist in Germany and other countries. Of special interest to social scientists will be the extent to which these combinations lead to a resurgence in the power of US financial institutions. Regardless of the level of bank power from the beginning of the twentieth century to the early 1980s, it is undeniable that commercial banks became significantly less influential during the 1980s and 1990s. The effect of this anticipated integration of financial corporations on relations between financial and nonfinancial corporations will be an interesting process to observe.

See also: Corporate Finance: Financial Control; Economic History; Finance, Geography of; Financial

Institutions in Economic Development; Money and Finances, Psychology of; Pay and Compensation, Psychology of

Bibliography

Berle A A Jr, Means G C 1939 *The Modern Comporation and Private Property*. Macmillan, New York

Chandler A D Jr 1977 *The Visible Hand: The Managerial Revolution in American Business*. Harvard University Press, Cambridge, MA

Davis G F, Mizruchi M S 1999 The money center cannot hold: Commercial banks in the U.S. system of corporate governance. *Administrative Science Quarterly* **44**: 215–39

Gerlach M L 1992 *Alliance Capitalism: The Social Organization of Japanese Businessm, a book of essays*. University of California Press, Berkeley, CA

Gershenkron A 1962 *Economic Backwardness in Historical Perspective*. Harvard University Press, Cambridge, MA

Go T K 1999 *American Commercial Banks in Corporate Finance, 1929–1941: A study in Banking Concentration*. Garland, New York

Lintner J 1960 The financing of corporations. In: Mason E S (ed.) *The Corporation in Modern Society*. Harvard University Press, Cambridge, MA

Mintz B, Schwartz M 1985 *The Power Structure of American Business*. University of Chicago Press, Chicago

Mizruchi M S 1996 What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates. *Annual Review of Sociology* **22**: 271–98

Mizruchi M S, Stearns L B 1994 A longitudinal study of borrowing by large American corporations. *Administrative Science Quarterly* **39**: 118–40

Roe M J 1994 *Strong Managers, Weak Owners: The Political Roots of American Corporate Finance*. Princeton University Press, Princeton, NJ

Stearns L B 1986 Capital market effects on external control of corporations. *Theory and Society* **15**: 47–75

Useem M 1996 *Investor Capitalism: How Money Managers Are Changing the Face of Corporate America*. Basic Books/HarperCollins, New York

Ziegler R, Bender D, Biehler H 1985 Industry and banking in the German corporate network. In: Stokman F N, Ziegler R, Scott J (eds.) *Networks of Corporate Power: A Comparative Analysis of Ten Countries*. Policy Press, Cambridge, UK

Zysman J 1983 *Governments, Markets, and Growth: Financial Systems and the Politics of Industrial Change*. Cornell University Press, Ithaca, NY

M. S. Mizruchi

Commodity Chains

Commodity chains follow specific commodities as they are moved over space to connect production with consumption. Commodities are seen as originating in regions of production and moving through networks of economic agents, as with producers, distributors and transporters, to regions of consumption. The commodity chain concept has potential for demonstrating the effects of consumption on labor and the environment. Armed with greater geographical knowledge, organized around the commodity chain concept, consumers can make choices that more fully consider the effects of consumption on workers and the natural environment.

1. Conceptions of Commodity Chains

Since the middle 1970s there has been a resurgence of interest in space conceptualized not as a homogenous expanse of territory, but as channels, networks, nodes and chains of connection. In economic geography the main interest resides in investigating the spatial organization of business networks (Yeung 1994). Related to this interest, commodity chain analysis differs in that analysis focuses on the commodity, rather than industrial organizations or networks of inter-firm relations. The geographical literature contains numerous references to concepts similar to commodity chains. For example, Harvey (1989, p. 101) implies the notion of a commodity chain in arguing the importance of knowing the effects of consumption on the conditions of production: 'We can take our daily breakfast without a thought for the myriad of people who engaged in its production. All traces of exploitation are obliterated in the object.' However, the commodity chain concept only began to be used formally in the middle 1980s. Three versions of commodity chain analysis can be discerned: global commodity chains, systems of provision, and commodity circuits. These versions have recently been criticized as limited in scope and applicability, and an expanded commodity chain concept proposed.

1.1 Global Commodity Chains

The notion of commodity chains has been developed in the world systems approach in sociology with extensions in the literature of economic geography (Hopkins and Wallerstein 1986 and Gereffi 1994). As characterized by the world systems approach, a commodity chain refers generally to a network of labor and production processes whose end result is a finished commodity (Hopkins and Wallerstein 1986, p. 159). Gereffi and his co-workers see inter-organizational networks linking households, enterprises, and states to one another within the world economy (Gereffi et al. 1994, p. 2). Global commodity chains stress the organizational scope of transnational production arrangements, that is the linkages between economic agents, such as raw material suppliers, factories, traders, and retailers. In this view, commodity chains have three main dimensions: (a) an input-output structure, in the sense of a set of products

and services linked together in a sequence of value-adding economic activities; (b) territoriality, as with spatial dispersion or concentration of production and marketing networks; and (c) a governance structure, in terms of authority and power relationships that determine how financial, material, and human resources are allocated and flow within a chain.

Two distinct types of governance structure are distinguished: producer-driven and buyer-driven commodity chains. In producer-driven commodity chains, transnational corporations or other large integrated industrial enterprises control the production system. In buyer-driven commodity chains, large retailers, brand-named merchandisers, and trading companies set up decentralized production networks in exporting countries, typically in the Third World. The latter pattern of trade-led industrialization is common in labor intensive, consumer goods industries such as garments, footwear, toys, household goods, consumer electronics, and a wide range of handcrafted items, such as furniture or ornaments. International contracting is generally carried out by independent Third World factories making finished goods, rather than components or parts, under original equipment manufacturer arrangements, and with specifications supplied by the buyers and branded companies that design the goods (Gereffi 1994, p. 219–22). In these studies firms comprise the main unit of analysis and unequal distributions of wealth along the chain are explained as outcomes of firm competition and innovation.

1.2 Systems of Provision

A second approach to commodity chain analysis derives from Fine and Leopold's (1993, p. 600) rejection of horizontal analyses of consumption that stress commodification and aestheticization connected with retailing, and their favoring, instead, of the vertical dimension of the system of provision—the commodity-specific chain connecting production, distribution, marketing, and consumption, and the material culture surrounding these elements. Different systems of provision result from distinct relationships between material and cultural practices spanning the production, distribution, and consumption of goods (Fine and Leopold 1993). This perspective allows for a more dynamic consideration of consumer behavior in economic transformations within advanced capitalist economies.

Some accounts argue that production is now consumer-driven (Piore and Sabel 1984), while others emphasize technological change or regulatory institutions in producing economic change (Harvey 1989). It has been argued that both sets of literature inadequately conceptualize consumption, and that a system of provision approach allows for a more balanced treatment of the relationships between production and consumption (Leslie and Reimer 1999, p. 405). Fine and Leopold (1993, p. 28) also recognize the changing relationships between the symbolic meanings of goods and their physical contents. Thus it is important to trace not only commodity movements over space, but also the discourses and representations operating in systems of provision.

1.3 Commodity Circuits

A third approach to commodity chain analysis, tracing commodity circuits, combines work in the political economy tradition with ideas derived from post-structural social theory. The notion of a circuit of culture has been used in examining the ways geographical knowledge of commodity systems is shaped and reshaped and in showing how particular cultural artifacts move through the circuit. Unlike the notion of a chain, circuits have no beginning or end, but instead Cook and Crang (1996) envision dense webs of interactions between all sites, with commodities interrelating with other goods as they travel. Rather than privileging one site, such as production, in interpreting the meaning of a commodity, explanation lies in combining a number of different processes. An appropriate politics of consumption recognizes the fragmentary and contradictory nature of the knowledges through which commodity systems are imagined (Cook and Crang 1996). Rather than assuming that consumers have little awareness of the quality and geographical origins of commodities, this approach sees commodity circuits constructed and reconstructed through consumer knowledge. This view might be criticized for eventually abandoning the concept of a chain altogether and for neglecting ideas of consumer resistance and activism (Leslie and Reimer 1999).

2. Recent Tendencies in Commodity Chain Analysis

Although useful in terms of conceptualizing the connections between firms and actors across space in the production of final outputs, these versions of the commodity chain concept fail to connect the sign of the commodity at the consumption end with labor and the natural environment at the production end. In not following through at the consumption end, world systems theory misses a vital opportunity to connect commodity chains with advertising, popular culture, and the geography of consumption. By not tracing firms' activities back to the conditions of workers, economic geography reduces labor to being yet another input in the overall production process, divorcing economic analysis from the politics of production. By not rooting commodity production in the natural environment, all these approaches fail to connect with political ecology, an important emerging

area in critical theory (Peet and Watts 1996). Significantly, these instances of neglect occur at the most politically sensitive sites along commodity chains, preventing the concept from uniting, in a comprehensive politics, consumption, culture, labor, and nature.

2.1 Expanded Version of Commodity Chains

In a more complex model of commodity chains, the effects of the commodity at the consumption node, the social and natural conditions at the production node, and the complex intersections of commodity chains at various intermediating and terminal points, are added to a simple model of commodity movement. This expanded version integrates the generation of cultural (signifying and representational) effects at the consumption node with the social and natural conditions at the production end of the chain, the purpose being to show the effects of one on the other, bringing home to consumers the results of consumption. In contrast to earlier theories, this expanded formulation sees commodity movements (the vertical dimension) connecting nodes that are conceived as places (horizontal dimensions) located at different positions along commodity chains. This emphasis on place lends the analysis potential for bringing together several aspects of the production and consumption of commodities. In this vein, the commodity chain is posited as a way of tying together material and signified realities, consumption and production, and activities separated by space, providing a fuller interpretation of the material world and the real, human actors who live in it (Hartwick 1998).

2.2 Consumption Nodes

At the terminal nodal point of consumption, the radiating image of a commodity produces a halo effect of socially constructed meanings. Such meaning systems can be analyzed in a number of ways, including the aesthetics of representation and consumer taste. The preferred theoretical tradition is the semiotics of Peirce (1931–5). In such a materialist social semiotics, a commodity has two meanings: a first-order meaning at the level of function; and a second-order imputed meaning at the level of symbolic effect. Semiotic analysis tries to reveal the social prescription of the meaning of a commodity. In a postmodern sign economy, first-order meanings, based in physical needs, have been transformed into second-order representational meanings, emanating from social status and cultural allusion (Gottdiener 1995). However, for a complete explanation, the conditions of production must be included within social semiotic analysis. Stark contrasts in material and cultural conditions so intensify second-order symbolic meanings that they mutate into a third, specifically geographical, meaning-form. Images are usually added to commodities at, or near, the organizational centers of global trading systems, where the producers of sign value share the cultural norms of the consumers. Underdevelopment, unequal exchange, profit flows and power differences, vastly different conditions of life, must be hidden behind the image of the commodity to give the impression that consumption is effortless and has no negative side effects. In general, adding social semiotics to commodity chain analysis expands the horizontal analysis of commodities at the consumption end and creates awareness of vertical linkages to producers (Hartwick 1998).

2.3 Production Nodes

The conditions of production at the other terminus of the commodity chain include the social relations of production, and also relations between workers and families, hence the reproduction of the household. This, in turn, consists of social relations of reproduction (gender, class, ethnicity) and social relations with nature, including the worker-family's reproductive use of subsistence land and resources. The complete set of the conditions of existence of the producers, and the full range of natural conditions of production that enter into the materiality of the commodity, and pass thereby into the distant bodies of consumers, are explored.

2.4 Commodity Bundles

Commodity chains articulate one with another at various nodal points of the intersection of their pathways. At the production end, commodities of prime significance form the conditions of existence for secondary commodity chains. At other significant nodes, processing points in particular, chains converge into commodity bundle that make up imaged complexes entering the consumption node—the entire complex combines to signify status, wealth, and power through radiating interconnected significations. As Glennie and Thrift (1993, p. 605) note, any commodity chain is characterized by leakiness, in both material and symbolic ways, which influences other commodity chains—they catalog the intertwining between commodities, advertising of linked products, consumer reflexivity, and consumption across several commodity systems. Nodes of consumption are thus points of convergence for series of imaged complexes of commodity chains. In this last sense, each purchase in the market is seen as culminating in an entire complex of commodity production and representation, linking together several lines of waged workers, unwaged reproducers of waged workers, and natural resources.

3. Commodity Chains and the Politics of Consumption

Geographical analysis centered on commodity chains connects the generation of cultural (signifying) effects within consumption to the real and material (social and natural) conditions involved in production. Deconstruction, understood in a geographical and materialist sense, uncovers these hidden dimensions of the commodity sign. Moving to a geographical focus on place broadens commodity analysis giving it greater political potential as a critical tool. Using this device, a series of local worlds are linked by connecting-chains into global systems. Commodity chains enable a greater awareness of the geographical implications of the political economy of consumption. Informed by commodity chain analysis, consumers learn to deconstruct, or critically analyze, the disconnected messages that bombard them by reconnecting images with material realities. Also the focus on people-in-places, as key nodal points along a commodity chain, broadens economic analysis so that social, cultural, and natural relations can be incorporated. This gives commodity chain analysis greater political potential as a comprehensive political tool. Formulated in this way commodity chains are said to constitute an analytical basis for a transformative politics (Hartwick 2000).

See also: Development and the State; Development Theory in Geography; Economic Geography; Globalization: Political Aspects; International Business; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; International Trade: Geographic Aspects; Marxist Geography; Multinational Corporations; World Systems Theory; World Trade Organization

Bibliography

Cook I, Crang P 1996 The world on a plate: Culinary culture, displacement and geographical knowledges. *Journal of Material Culture* **1**: 131–53

Fine B, Leopold E 1993 *The World of Consumption*. Routledge, London

Gereffi G 1994 The international economy and economic development. In: Smelser N J, Swedberg R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ, pp. 206–233

Gereffi G, Korzeniewicz M, Korzeniewicz P 1994 Introduction: Global commodity chains. In: Gereffi G, Korzeniewicz M (eds.) *Commodity Chains and Global Capitalism*. Praeger Press, Westport, CT, pp. 1–4

Glennie P D, Thrift N J 1992 Modern consumption: Theorizing commodities and consumers. *Environment and Planning, D: Society and Space* **11**: 603–6

Gottidiener M 1995 *Postmodern Semiotics*: Material culture and the forms of postmodern life. Blackwell, Oxford, UK

Hartwick E 1998 Geographies of consumption: A commodity chain approach. *Environment and Planning, D* **16**: 423–37

Hartwick E 2000 Towards a geographical politics of consumption. *Environment and Planning, A* **32**: 1177–92

Harvey D 1989 *The Condition of Postmodernity: An enquiry into the origins of cultural change*. Blackwell, Oxford, UK

Hopkins T K, Wallerstein I 1986 Commodity chains in the world economy prior to 1800. *Review* **10**: 157–70

Leslie D, Reimer S 1999 Spatializing commodity chains. *Progress in Human Geography* **23**: 401–20

Peet, R, Watts M 1996 *Liberation Ecologies: Environment, Development, Social Movements*. Routledge, London

Peirce C 1931–5 *Collected Papers of Charles Sanders Peirce*. Harvard University Press, Cambridge, MA

Piore M, Sabel C 1984 *The Second Industrial Divide: Possibilities for Prosperity*. Basic Books, New York

Yeung H W 1994 Critical reviews of geographical perspectives on business organizations and organizations of production: Towards a network approach. *Progress in Human Geography* **18**: 460–90

E. R. Hartwick

Common Law

The term 'common law' first comes into use in England during the thirteenth century. It referred to the 'common' repertoire of norms and understandings that had begun to be recognized across the whole realm during the latter half of the previous century as central government was becoming consolidated. These common understandings, which were seen to contrast with surviving customs of a simply localized character, became institutionalized as *law* as they were invoked to underpin the decisions of judges in the royal courts. So this evolving 'common law' claimed to represent the immemorial understandings prevailing in the social world, and was at the same time put to use as an instrument of rule by the courts as they developed into a distinctive branch of government. This primary sense of the term 'common law' has survived and continues in use, particularly as a means of contrasting this source from others which the English courts now draw on concurrently—'*statute*' and '*equity*.' However, the term has now acquired a general secondary sense in identifying the broader legal domain formed through the diffusion of English law under the process of colonial expansion. This common law world is contrasted with that occupied by the codified civil law systems of continental Europe directly inspired by Roman law.

1. Historical Development in England

While the common law is seen to originate in the everyday practices and understandings of the social world, its growth from the twelfth century took place in the context of dual, linked processes: the progressive

consolidation of central government, and the gradual emergence of a specialist legal profession. This growth was realized through the discursive formulation of 'law' in the pleadings of lawyers and the judgments given in the Royal courts. It was associated from an early stage, and remains associated today, with one primary assumption and a distinctive method. The assumption was that the common law sought to realize and secure the practices and understandings of everyday life (*immemorial custom*). The strength of this assumption can be illustrated through numerous examples, including: the continuing invocation of the *reasonable man* in arriving at the appropriate standard of care in respect of claims arising from civil wrongs (*tort*); the care taken to ascertain normal business practice in the area concerned when judging commercial disputes; and the deference accorded to the customs of conveyancers by courts dealing with property disputes. The method distinctive to the common law lay in the procedure under which lawyers recovered the rules to underpin their arguments, and judges to justify their decisions, from the records of earlier judgments. This procedure—the *doctrine of precedent*—was made possible by the maintenance from an early stage of semi-official written records (the *Year Books*) and subsequently sustained by private series of reports before official law reporting became institutionalized. Consistency in decision-making was reinforced by the understanding that courts should, wherever possible, follow their own previous decisions and were obliged to follow those of courts superior to them in the judicial hierarchy (the doctrine of *stare decisis*). This procedure can be contrasted with the mode of decision associated with the continental civil law world under which the judge seeks to ground a decision in an authoritative, ultimately codified, text.

These working practices which fashioned the common law were facilitated by the close relationship prevailing from medieval times between the judges and the specialist groups which ultimately became the legal profession. Notwithstanding the strong historical link between adjudication and the consolidation of government, in England there has never been a career judiciary. The higher ranks of the judiciary have been recruited exclusively from the legal profession, with appointment to the 'bench' representing the ultimate career stage of the successful lawyer. While some elaborate ritual, including the conferment of knighthood in the case of the higher judiciary, marks the transition from barrister to judge, and some formal distance is subsequently maintained between judges and their former colleagues, judges remain sociologically very much part of the professional group from which they emerge. They continue to belong to long-established networks of information exchange and support—often crossing generations—begun in the great private schools, continued at university, and subsequently prevailing in the collegiate groups to which barristers belong (the Inns of Court). So the courts can be just as easily seen as the apex of the legal profession as a specialized branch of government.

The common law, developed in the King's courts through resort, to precedent, has traditionally been contrasted with two other sources of law. The first of these has been described by legal historians as arising directly out of the formal working practices of the King's courts. These were characterized from an early stage by procedures requiring anyone who sought redress to bring the claim within one of a limited number of elaborate written formulas (known as writs). If some mistake were made in the pleadings, the claimant would be denied his remedy. Ensuing direct appeals for help to the King, which such apparent injustices gave rise to, led to the growth of a parallel remedial jurisdiction presided over by the Lord Chancellor. This jurisdiction—known as equity had a distinct body of doctrines, rules, and practices, and was long administered in separate courts.

Both the common law and equity evolved alongside, and are contrasted with, statute law—rules promulgated by the sovereign. While progressively larger areas have become at least partially regulated by statute—notably in the fields of crime, commercial law, and property—a comprehensive legislative scheme of the kind introduced through the great continental codes of the eighteenth and nineteenth centuries has never been attempted in England. This more limited role and reach of legislation provides a further contrast with the civil law regimes of continental Europe.

The 're-discovery' of Roman law in Italy during the eleventh century is generally presented by historians as having a limited effect on the growth of the common law in England, but England was in no sense isolated from this learning. Some of the clerics who accompanied William I to England were distinguished as scholars of Roman law, so Justinian's great *Digest* was well known there from an early stage. Holdsworth (1909) also makes it clear that Bracton, a senior judge writing in the mid-thirteenth century, used Roman law to fill the gaps left by his extensive repertoire of case histories. While Roman law provided part of the intellectual climate within which a unified legal order evolved in England during the mediaeval period, its use as a direct source was limited. Even during the English Renaissance, the legal historian Maitland (1907) notes that the common law was still characterized by 'an amazingly continuous persistence of medieval doctrine.' He attributed this durability to the presence of the collegiate groups—the Inns of Court—within which English lawyers associated. He saw the medieval Inns as 'schools of law' which provided arenas of communication within which young lawyers were taught and their seniors formulated a coherent repertoire of legal doctrine. He wrote. 'What is distinctive of mediaeval England is not parliament, for we may everywhere see in Europe assemblies of estates … but the Inns of Court and the Year Books that were read therein, and we shall hardly

find their like elsewhere.' By the time of the great continental codifications of the eighteenth and nineteenth centuries, which were closely informed by Roman law, the evolving understandings and practices of English judges and lawyers were already part of a largely autonomous, differentiated legal order. By then the common law tradition was too well set to be disturbed.

2. *Geographical Diffusion of the Common Law*

While originating in England, the common law has undergone wide diffusion as English settlers carried it with them abroad and government at Westminster imposed it upon annexed or protected territories in the course of colonial expansion. Across India, North America, Australasia, and considerable areas of Africa and Asia, the common law remains today a significant component, linking together a range of independent legal orders. Following the now ancient precedent of the USA, most former colonies have retained significant elements of the common law, justifying the shared label of 'the common law world.' But the sheer diversity of contemporary arrangements within this loose association inevitably makes any generalization hazardous.

English settlers establishing colonies overseas took with them the general law of England as it stood at the moment of first settlement. By this means, the common law was established as the foundational law of the USA and of those regions of what became the British Commonwealth which were subject to settlement. In these cases, very little recognition was initially given to the existing laws and customs of indigenous peoples. In those further territories which were conquered, annexed, or declared protectorates as the British colonial empire subsequently expanded, the statutes of general application, common law, and doctrines of equity in force in England at the time were typically imposed by Order in Council from Westminster. In these latter cases, notably in India and Africa, early recognition was given to diverse local legal orders encapsulated in the course of colonial expansion. In the case of Africa, this recognition extended to indigenous dispute institutions. Elsewhere, notably in Canada and Australasia, recognition of the prior arrangements of indigenous peoples was postponed well into the twentieth century.

Once English law, including its major common law component, had been received in a colony, it remained in force until repealed or modified by the local colonial legislature or courts—or by the overarching authority represented by Parliament at Westminster and the Privy Council (the ultimate appellate tribunal of the Empire). These initial linkages between the metropolis and the colonial world have had variable duration. In the case of the American colonies they were severed early on by the Declaration of Independence. During the twentieth century they progressively disappeared in the case of self-governing Commonwealth countries; but some former colonial territories have chosen to retain a final appeal to the Privy Council, and a very few colonies still remain.

Beyond the impetus given by the immediate necessity of explicit constitutional provision, the 'fresh starts' generated by the colonial enterprise resulted in vigorous, eclectic localized growth for the common law. In these new local contexts, the energy to attempt rationalization, even towards codification, was perhaps greater than in the metropolis. Major initiatives included the codification of wide areas of private law and the criminal law undertaken in India; the Indian Penal Code was subsequently reused in modified form in several other colonial territories. In the USA, with the particular problem of its numerous private law jurisdictions, a general attempt at rationalization has been the American Law Institute's *Restatement of the Law*. This initiative, now in its third generation of restatements, has brought together judges, lawyers and professors in the preparation of an agreed statement of common law across the state jurisdictions. In the same context, the great American law schools have, on the whole, self-consciously attempted to present an overall view of the common law in their teaching and research. Elsewhere across the post-colonial world the superior courts in different jurisdictions have become increasingly attentive to each other's formulations, indicating explicitly in their judgments that they are working together on a shared overarching fabric which is in the widest sense a common law.

3. *Theorization of the Common Law*

Attempts to 'theorize' the common law, in the sense of drawing together rules and procedure for the purposes of exposition and commentary, were made from a very early stage. Holdsworth (1909) even, proposed that Glanvil's compilation, prepared sometime towards the end of the twelfth century, 'marks the true beginning of the common law.' Judge Bracton's great treatise, compiled on the basis of 2000 cases which he had collected from the plea rolls, followed in the mid-thirteenth century. Both Glanvil and Bracton were deeply involved in what was becoming specialized, legal administration and for the remainder of the medieval period the common law continued to reside in practice, receiving discursive formulation only in the judgments of the courts and in authoritative commentaries. All the great legal texts of the period were prepared by practitioners whose learning was acquired through their work and in the context of their specialist associations, the Inns of Court.

Just as native legal theory in the medieval period concentrated on parochial commentary, that of the

early modern period continued to focus largely on the immediate practicalities at hand. During the constitutional crisis early in the seventeenth century, Chief Justice Coke eloquently prioritized the relative authority of judges, King (he 'hath no prerogative but that which the common law allows him') and Parliament. But Coke's arguments were not distanced speculation; rather those of a practical man, the holder of high office seeking to justify his position in the urgent context of struggles for power.

Prior to the mid-eighteenth century the common law had never been the subject of exposition and research in the English universities. English university teaching of the common law can be said to begin with the lectures which William Blackstone, the first Vinerian Professor, delivered at Oxford in 1758, and which were subsequently published in his famous *Commentaries* (1765–69). The *Commentaries* took more the form of popular exposition than scholarly treatise; but the account he gave provided a framework that would remain readily recognizable even 200 years later. In this work the common law takes pride of place as the law by which 'proceedings and determinations in the king's ordinary courts of justice are guided and directed.' The validity of common law doctrines, 'depending on immemorial usage for their support,' is determined by the judges. In reaching judgment, 'it is an established rule to abide by former precedents where the same points come again in litigation.'

It was only in the early nineteenth century that a body of specifically legal theory, focused on the common law tradition yet distanced from practice, began to develop in England. This body of theory links law firmly to government and maintains a consistent focus on adjudication. The link between government and law is established at the outset in the work of John Austin, the first Professor of Jurisprudence at University College London. In *The Province of Jurisprudence Determined* (1832), Austin presents law as the command of a sovereign directed towards members of a society in a 'habit of obedience,' with failure to comply being met with 'sanctions.' That central association safely in place, English jurisprudence turned to focus—virtually exclusively over the next 175 years—on providing theoretical foundations for adjudication.

The first modern attempt to provide a general theory of adjudication was made by Henry Maine in *Ancient Law* (1861). In this work, taking a broad, comparative, historical sweep, he insisted on locating the core of law in adjudication and in the emergence of rules underpinning judicial decision making. He began by mapping out a broad transition from small, kin-based groups to larger, territorial units. The very origin of social life was identified in the steering role exercised within a group of kin by the senior male agnate. These old patriarchs made decisions on an *ad hoc* basis; no consistent rules underpinned the decisions they took, yet government was supposed to be by adjudication by the senior male, before whom all disputes were brought. In the form of society that followed, numbers of these small groups of agnates became clustered together under chiefs, but the (sometimes fictional) assumption of shared kinship remained the basic organizing principle. Then came the territorial stage, in which members identified themselves through their common occupation of a defined tract of land, rather than through kinship. Around the end of the second stage and the beginning of the third, 'law' developed as rulers began to pronounce the same judgments in similar situations, providing their decision-making with an underlying set of rules. Later in the development of territorially-based societies, the settlement of disputes fell in to the hands of a specialized elite, who alone had access to the principles to be followed in their resolution. As Maine wrote: 'What the juristical oligarchy now claims is to monopolize the *knowledge* of the laws, to have the exclusive possession of the principles by which quarrels are decided.' There followed the 'era of codes,' and so on; but we can leave the developmental process at this stage.

Several important features are clear from this summary. First, there is the focus on dispute resolution and the foregrounding of a particular mode of intervention. For Maine there were no structural changes in the dispute settlement process over the fundamental stages of societal development. From the senior male agnate onwards, disputes were resolved by decision, handed down by a third party; there was no suggestion of negotiatory modes of settlement giving way to processes of third-party adjudication. Second, the presence of a normative basis for decision-making was the key attribute of law for Maine, and the emergence of this feature heralded the transition from the prelegal to the legal world. Third, there was the later development of specialization as legal rules became separated off from other rules operating in society. So, for Maine, social life is the product of 'government,' and the key variables lie in the identity of the third parties responsible for adjudication and in the evolving criteria underpinning decisions.

After Maine, the perfection of theories of law as adjudication occupied the jurisprudence of the common law world through the rest of the nineteenth century and much of the twentieth. Much of this scholarship was generated in North America where thoroughgoing re-examination was made of the doctrines of precedent and *stare decisis*. The greater part of this work remained within the traditional commentary mode of legal scholarship, exemplified by Goodhart's now classical attempts to refine the doctrine of precedent (1931). More ambitiously, scholars like Frank and Llewellyn of the legal realist school drew on sociological perspectives in questioning vigorously whether these doctrines actually operated in the manner claimed in native legal theory. They indicated the relative freedom of judges to distinguish

previous decisions they did not want to follow, suggesting that rules were drawn from the repertoire to justify desired outcomes, as opposed to determining what those outcomes would be. This iconoclastic strand survived on to the end of the twentieth century in the work of the critical legal studies movement.

One notable attempt to escape from the grip of adjudication, and to think about law unconstrained by the institutional shapes of the modern state, was made by Hart (1961). Rejecting Austin's (1832) command theory as simplistic and in some respects inaccurate, Hart's elucidation of law lays central emphasis on rules. One essential category of these is made up of 'primary rules' imposing obligations. The distinguishing mark of such a rule is the seriousness of the pressure to confirm to it 'Rules are conceived and spoken of as imposing obligations when the general demand for conformity is insistent and the social pressure brought to bear upon those who deviate or threaten to deviate is great.'

While Hart concedes that a society might exist with primary rules only, they alone do not constitute a legal system. For the step from 'the pre-legal into the legal world' to be made, these primary rules must be supplemented with 'secondary rules' of three kinds: rules of recognition—making it possible to identify what the primary rules are; rules of change—providing for the alternation of primary rules when the need arises; and rules of adjudication—empowering certain people to determine whether, on a given occasion, the primary rules have been broken. Most systems will also contain other secondary rules authorizing the application of penalties where primary rules are violated; but this further refinement is not seen as essential to a legal system.

Through the device of these secondary rules, Hart (1961) appears to disassociate his model from any parochial institutional content and invest it with considerable value for cross-cultural purposes. However, the escape seems to be illusory, leaving Hart's account of law firmly grounded in the institutions of the state and state-sponsored adjudication. He himself recognizes that the three categories of secondary rules imply, and require, the presence of legislative and adjudicatory agencies. While Hart struggled to marginalize adjudication by identifying law as an affair of rules, subsequent exegeses of his work hastened to represent his core idea as a theory of adjudication. It is significant, too, that his most influential successor, Ronald Dworkin, has consistently refocused the task of jurisprudence on understanding and justifying judicial decision-making.

See also: Conventions and Norms: Philosophical Aspects; Folk, Indigenous, and Customary Law; Law: Change and Evolution; Law: History of its Relation to the Social Sciences; Legal Systems, Classification of; Norms; Tradition, Anthropology of

Bibliography

Austin J 1832 *The Province of Jurisprudence Determined*. J. Murray, London
Blackstone W 1765–69 *Commentaries*. Clarendon Press, Oxford, UK
Dworkin R 1977 *Taking Rights Seriously*. Duckworth, London
Frank J 1930 *Law and the Modern Mind*. Coward, New York
Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK
Holdsworth W 1909 *A History of English Law*. Methuen, London
Holmes O W 1881 *The Common Law*. Harvard, Cambridge, MA
Goodhart A L 1931 *Essays in Jurisprudence and the Common Law*. Cambridge University Press, Cambridge, UK
Llewellyn K N 1962 *Jurisprudence: Realism in Theory and Practice*. Chicago University Press, Chicago, IL
Maine H 1861 *Ancient Law*. J. Murray, London
Maitland F W 1907 *Select Essays in Anglo-American Legal History*. Cambridge University Press, Cambridge, UK

S. Roberts

Common Sense, Anthropology of

Common sense, taken for granted, obvious, self-evident, the only logical alternative: such labels place any challenge to a cultural orthodoxy at risk of annihilation by mockery. They are the terms of cultural power. Moreover, because they appeal to a universalizing rhetoric that brooks no possibility of exception within the known bounds of experience, they effectively disguise the agency that underlies that power. For this reason, they have even led critical analysts attempting to deconstruct the cultural specificities of common sense to treat the latter as established structures rather than as a work permanently, and unpredictably, in progress.

Thus, for example, in her justly famous explorations of 'self-evidence,' Mary Douglas (1966, 1975, pp. 276–318) presupposes the existence of a set of rules—a taxonomy—despite the ironic circumstance that even this assumption represents the common sense of the relatively stable, ordered world of bourgeois experience. Yet the genius of her approach should not be overlooked; indeed, it paves the way for this critique. Douglas shows how attributions of purity and pollution are defined in terms of failure to fit taxonomic schemes already in place; and the metaphor of place is central, for she defines 'dirt' as 'matter out of place' (1966, p. 48). In this reading, however, the conceptual cartography is already 'in place': the lineaments of the taxonomy are already known and accepted throughout the culture, often assumed in this approach to be a clearly bounded entity: 'Where there is dirt there is a

system' (Douglas 1966, p. 48). Such sharp delineation of boundaries, however, is itself a historically specific production of order created by agents exercising highly variable degrees of conscious volition, and commonly takes the form of the bureaucratic logic of nation states, for which the metonymic construction of the people as a social unit such as a kinship group serves to ground national identity in an apparently unchallengeable timelessness (Herzfeld 1992, 1997).

In such situations, the logic of ritual pollution and taboo may also infuse systems that claim the status of a universal rationality. While experience may show that democracies often fail to live up to their goals, for example, people often invent secular theodicies to explain failure—'it's the venality of those who have power, not the organizational structure itself'—rather than attempt to challenge the rationalizing rhetoric whereby those who manage the structures of power seek to perpetuate the status quo. Thus, rationality itself becomes a symbolic system, with its own hostility to the 'contamination' of alternative logics—as witness official hostility to, for example, New Age religious and medical practices (see Badone 1991, Danforth 1989). Here again, logical structure is mapped onto the distribution of power in such a way as to make seemingly deviant or alternative rationalities directly threatening to the bureaucrats and other power brokers of the state. Moreover the idea of a sharp mind–body, subject–object discrimination is so deeply embedded in the ontology of most educated Westerners and many others that in itself it comes—in the guise of an opposition between the sensible and the sensual—to constitute the very grounds of common sense in general.

Thus, Douglas's vision of common sense has the virtue of relativizing what might otherwise be taken as absolute and universal, but it can be extended much further through a conceptual shift of analytic emphasis from conceptual structure to the agency behind the deployment of rhetoric—to the uses, rather than simply the forms, of classification. This allows us to see why common sense so often appears to be static even when historical evidence shows it to have changed in content and logic: its authority appeals to assumptions of permanence but actually often rests on strategic opportunism. Since taxonomy constitutes the basis of knowledge, however, the realization that differing systems of classification both reflect and produce widely divergent understandings of how the world works was a crucial step in the process of relativizing the universalist intellectual claims that, in the West, find their popular expression in folk models of common sense.

Methodologically, it is those moments at which ethnographers are made to feel foolish that the locally dominant version of common sense makes itself felt. When, at last, I discovered that the shepherds among whom I was living in a Cretan mountain village were engaged in extensive, reciprocal animal theft and asked them why, they responded, as though it must have been obvious to anyone but an idiot, 'To make friends.' This baffling response provoked me to an entire ethnographic inquiry, in which I was able to show that, in an extremely demanding physical and social environment, agonistic relations that took the form of systematic, reciprocal animal theft provided en effective means of determining who would be both worthy to serve as 'friends' and tough enough to do so in a practically advantageous way.

Yet one of the criticisms made of the resulting study (Herzfeld 1985, cf. Gilsenan 1986, Loizos 1986) was that it overlooked both gender and (incipient) class inequality. In retrospect, I consider these critiques to be largely justified. A reflexive consideration of how I came to see the situation in the terms of the dominant group may now be of value in revealing the extent to which a structural view of social values can entail the uncritical acceptance of locally dominant models of common sense—a revealing convergence of anthropological generalization and local hegemony. My delight in 'discovering the logic of the system,' to use a phraseology popular in the structuralist era, obscured the clear evidence—which others were able to extract from my ethnographic descriptions—of inequalities, that such assertions of perfect reciprocity effectively masked by reproducing the image of an idealized past ('structural nostalgia,' Herzfeld 1997, pp. 109–38) in the less-than-ideal present (about which the villagers themselves had few illusions).

This example of how encounters between different notions of common sense produce ethnographic insight and can lead to a better understanding of the concealed operation of power is also useful for another reason. When we speak today of the exercise of power, we often—at least implicitly—mean a nation-state bureaucracy or a colonial state. But power is frequently diffused throughout a society in such a way as to be exercised at many analytically separable levels. The common sense of the village 'big men' was emphatically not that of the state, although its bearers called the latter into question by pointing out that the heroes of national independence had themselves been guerrillas known as 'thieves' (*kleftes*).

Notions of common sense infuse and legitimate claims to power at many levels. Such assessments of a culturally defined common sense as Bulmer's famous question, 'Why is the cassowary not a bird?' (1967), are thus not confined to the relatively 'harmless' zones of ritual and taboo. In a society in which it is said that harm will eventuate from any violation of taboo, the misuse of categories is never an insignificant matter. But in self-avowedly modern and rational societies, the issues take on a magnitude of scale that lends them particular force, although the logical principles remain fundamentally similar: virtually the same classificatory logic that we find in the taxonomic organization of cosmology also informs extremely violent practices of exclusion and even annihilation.

The theory of taboo enunciated by Mary Douglas in *Purity and Danger* (1966) has thus proved extremely useful in understanding the treatment of exiles and refugees (Malkki 1995) and of outsiders more generally (Herzfeld 1992). These expansions of her theory, in turn, have necessitated recognition of the kinds of *agency* involved in the activation of categories for specific social and political ends.

A similarly appreciative critique can be addressed to Clifford Geertz's approach to common sense (1983, pp. 73–93). Geertz represents common sense as a cultural *system*, utilizing a model through which he has exercised an enduring influence on the discipline. But if common sense is systematic, where does that quality lie? It must be in the rhetoric of those who invoke the idea of a unitary logic, rather than in the logic 'itself'—the cultural 'system.'

Akhil Gupta (1998) has shown with notable clarity how ideas of common sense reinforce hierarchy at several mutually encompassing levels. Drawing on his fieldwork in an agricultural village in western Uttar Pradesh, India, he shows how the rhetoric of science validates a particular view of common sense that supports the interests of the global economy at the expense of the villagers who are nonetheless expected to resign themselves to its authority.

The phrase 'common sense' implies universality, and in this respect shows that the concept originates in Enlightenment hopes of discovering a universal logic. Yet in the concept of sense, as in that of taste, the Cartesian separation of mind and body had not yet supervened in pre-Enlightenment times and does not seem particularly logical in many another culture (see Classen 1993, pp. 1–3).

Is common sense ideology? I would prefer to treat it, in Bourdieu's (1977) terms, as an example of *doxa*—that set of principles that both governs and is embedded in the culturally inculcated dispositions to which most daily routines are subject. It shares with *doxa* the quality of unreflexiveness: while it can be made the subject of explicit speech, for example, its authority depends on its immunity to questioning. In writing this entry, for example, I am questioning the Enlightenment presupposition of universality on which European notions of common sense are, at least implicitly, predicated.

But common sense is only revealed through practice. It is the invocation of common sense as a reason for doing things that gives it its ontological status in the world. A speaker who argues that something 'stands to reason' may wish to imply that what is invoked has universal status; but to the extent that this is the impression created, it simply testifies to the performative efficacy of the utterance. Thus, the task of the ethnographer is to document such utterances and to gauge their effect. In discussing the status of lying in rural Lebanon, for example, Michael Gilsenan (1976) demonstrates that calling the bluff of a liar by dint of another lie shows both what everyone knows (that lying in social life is inevitable) and that in the specific case a liar's failure to realize that he was being lied to showed him up as the less talented liar. Here common sense consists in realizing that everyone lies and not expecting to be believed oneself. Such a judgment would be considered a great deal less commonsensical in, for example, an academic debate (in which empirical reality probably approaches Grice's (1975) culture-specific maxims of conversation more closely than in virtually any other context (see also Briggs 1986, p. 83, for an assessment of these maxims and their relevance to fieldwork)). Especially in agonistic societies that find themselves advancing competing historical and cultural claims in the context of nation-state ideology, however, such notionally 'pure' commitments as the ideal of historical truth can prove more rhetorical and strategic than literal (see especially Shryock 1997). Moreover, the idea of a disinterested academic search for pure fact is itself riddled with rhetorical claims about both common sense and science, the latter often serving as the legitimating framework for a particular reading of the former. In other words, it is the local cultural rules that constitute the grounds for any appreciation of common sense; but it is the performance of social relations that provides the opportunity for the current understanding of common sense to be made manifest and graspable.

Finally, what of psychological inner states? According to the conventional anthropological wisdom, we cannot know the inner workings of informants' minds, still less generalize from these to the collective psychology of entire populations (e.g., Needham 1972, on 'belief'). Yet a degree of empathy seems to define one version of common sense, according to which the ethnographer usually manages to make precisely such intuitive judgments without difficulty and to the apparent benefit of the description overall (Leavitt 1996). To be sure, this may reflect the embedded common sense of the scholarly community rather than a careful analytical deduction: it implies an equivalence between empathy and knowledge. From another perspective, however, we may extrapolate the specific common sense of a given culture from informal judgments that reveal shared assumptions about what people think ('what they must think') under any given set of circumstances (e.g., Rosen 1995).

Implicit in the foregoing is a comparative project. If Leavitt is right in his demonstration that our own common sense allows us to accept what the logic of our profession prohibits, we see that, even internally, common sense does not necessarily possess the consistency or coherence it claims. More than that, however, contemplation of the limits beyond which we are not prepared to push our assumptions about shared human responses reveals significant cultural and historical differences about what constitutes common sense—what, in some contexts, constitutes the underlying ideology governing the use of language

itself. We may then prefer to adopt a posture in which the discipline of anthropology becomes, quite simply, the comparative study of common sense. In that case, however, the common senses in question are not unchanging systems but constitute symbolisms in use, and the questions anthropologists ask concern both the formal content of these symbolisms and the principles and contingencies governing their use. This methodology underlies the more serious attempts to produce a genuinely reflexive form of anthropology, one in which the cultural, social, and political presuppositions of ethnographic practice become part of the critical analysis of cultural relativism itself.

See also: Control: Social; Hegemony: Anthropological Aspects; Hegemony: Cultural; Ideology, Sociology of; Norms; Symbolic Boundaries: Overview; Symbolism in Anthropology; Taboo; Totemism

Bibliography

Badone E 1991 Ethnography, fiction, and the meanings of the past in Brittany. *American Ethnologist* **18**: 518–45

Bourdieu P 1977 *Outline of a Theory of Practice* (trans. Nice R). Cambridge University Press, Cambridge, UK

Briggs C L 1986 *Learning How To Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research*. Cambridge University Press, Cambridge, UK

Bulmer R 1967 Why is the cassowary not a bird? A problem of zoological taxonomy among the Karam of the New Guinea Highlands. *Man* **2**: 5–25

Classen C 1993 *Worlds of Sense: Exploring the Senses in History and Across Cultures*. Routledge, New York

Danforth L M 1989 *Firewalking and Religious Healing: The Anastenaria of Greece and the American Firewalking Movement*. Princeton University Press, Princeton, NJ

Douglas M 1966 *Purity and Danger: An Analysis of Concepts of Purity and Taboo*. Routledge & Kegan Paul, London

Douglas M 1975 *Implicit Meanings: Essays in Anthropology*. Routledge & Kegan Paul, London

Geertz C 1983 *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York

Gilsenan M 1976 Lying, honor, and contradiction. In: Kapferer B (ed.) *Transaction and Meaning: Directions in the Anthropology of Exchange and Symbolic Behavior* (ASA Essays, 1) ISHI, Philadelphia, PA, pp. 191–219

Gilsenan M 1986 Review of Herzfeld 1985. *American Ethnologist* **14**: 393–4

Grice H P 1975 Logic and conversation. In: Cole P, Morgan J L (eds.) *Syntax and Semantics 3: Speech Acts*. Academic Press, New York, pp. 41–58

Gupta A 1998 *Postcolonial Developments: Agriculture in the Making of Modern India*. Duke University Press, Durham, NC

Herzfeld M 1985 *The Poetics of Manhood: Contest and Identity in a Cretan Mountain Village*. Princeton University Press, Princeton, NJ

Herzfeld M 1992 *The Social Production of Indifference: Exploring the Symbolic Roots of Western Bureaucracy*. Berg, Oxford, UK

Herzfeld M 1997 *Cultural Intimacy: Social Poetics in the Nation–State*. Routledge, New York

Leavitt J 1996 Meaning and feeling in the anthropology of emotions. *American Ethnologist* **23**: 514–39

Loizos P 1986 Challenges and invitations. [Review of Herzfeld 1985.] *Times Literary Supplement*, 11 April

Malkki L H 1995 *Purity and Exile: Violence, Memory, and National Cosmology Among Hutu Refugees in Tanzania*. University of Chicago Press, Chicago

Needham R 1972 *Belief, Language, and Experience*. Basil Blackwell, Oxford, UK

Rosen L (ed.) 1995 *Other Intentions: Cultural Contexts and the Attribution of Inner States*. School of American Research, Santa Fe, NM

Shryock A 1997 *Nationalism and the Genealogical Imagination: Oral History and Textual Authority in Tribal Jordan*. University of California Press, Berkeley, CA

M. Herzfeld

Communalism

The concept of 'communalism' is based on that of 'community' which has to be understood in relation to the concepts of 'nation' and 'civil society.' It derives its meaning from arguments about the place in modern society of communities of belonging. According to liberal, political theory, communalism is based on primordial attachments and stands in the way of the formation of a modern, civil society, in which individuals act as equals on the basis of their allegiance to the state in which they enjoy the rights and duties of citizenship. In modernization theories communalism signifies a hybrid phase between traditional society and modern society, since it combines modern democracy with political mobilization of religious communities in a plural society. Such theories make a sharp distinction between secular nationalism on the one hand and communalism on the other. Inherent in all these theories is an opposition between religion, seen as part of traditional society, and secular modernity. The fact that religious communities form a strong political force in modern societies all over the world cannot be understood in terms of standard modernization and secularization theories.

Communalism then, signifies a politics of belonging that does not emphasize the nation in all its diversity, but the homogeneous, religious community. It resembles the current notion of communitarianism in that it refers to a political agenda of communicative interaction within a particular community on the basis of shared 'habits of the heart,' but rather than emphasizing the defense of a 'common way of life,' it is much more aggressively antagonistic towards other, similar communities. It should be clear that the term communalism comes up as an alternative to national-

ism, both in theories which explain the failure of the nation-state to emerge in certain parts of the world and in theories which explain the demise of the nation-state under conditions of globalization.

Like the use of the concept of 'tribalism' in the study of African politics, 'communalism' is primarily used in the analysis of Indian society and politics and is seldom encountered in theoretical arguments of a more general nature. It refers to the articulation of religious communities into mutually antagonistic, social, political, and economic groups in India. Again, like 'tribalism' in Africa, communalism in India has a history as colonial theory with great social and political consequences. It emerges in the context of the colonial modernization of Indian society. The Indian population was classified, counted, and measured in terms of community. Notions of caste and religion were crucial in the census operations which took place on a regular basis from the 1870s. British colonial policies were strongly influenced by essentialist assumptions about the nature of Hindus, Muslims, and Sikhs. In their system of governance the British sought alliances with the 'natural leaders' of communities and, ultimately, this created a system of representation in which membership of a community took precedence over place of residence. Recruitment for jobs both in civil administration and in the military was also organized around notions of race, caste, and religious community. The notion of 'martial races' was specifically applied to the Sikhs, a religious community in the Punjab, North India. The British recruited only 'pure' Sikhs who were following the strict norms of what the British considered to be the proper Sikh faith. In this way the Sikhs became a dominant section in the Indian military until today, and a particular 'orthodox' form of Sikhism was promoted with great social and political consequences for the region.

The extent to which language politics was influenced by colonial ideologies of communal difference was also important. The medium of communication was supposedly related to communal identity. A major issue was the question of whether Hindi or Urdu should be the language of instruction and the language of government besides English. In the debates about it, Hindi came to be conceptualized as the language of the Hindu community while Urdu came to be conceptualized as the language of the Muslim community. In the later part of the nineteenth century Hindu leaders increased their attacks on Urdu and Indo-Persian culture by emphasizing the communal basis of the differences in script and vocabulary between Hindi and Urdu. A new vernacular literature emerged which attacked the joint Hindu-Muslim culture, as expressed in Urdu or the spoken vernacular, Hindustani. The principal arenas for the Hindi-Urdu controversy were schools and entry exams for government service. After the Partition of 1947 Urdu became the national language of Pakistan and Hindi, besides English, the main administrative language of India. The identification of language with religious community was also the ground on which the Sikh community after Independence started to demand a special status for the language of their sacred scriptures.

The legal structure of both colonial and postcolonial India has been greatly influenced by the notion of communal difference. Colonial law was grafted on Hindu and Muslim legal traditions which were interpreted and codified in a joint effort of British legal scholars and native religious specialists. Ancient Sanskrit texts, for instance, became foundational for the creation of Hindu civil law. Personal law in India until the present day is based upon communal difference. The contradictions between the claim that India is a secular state and the existence of communal laws can be briefly illustrated by the Shah Banu conflict of 1985. Shah Banu had been divorced by her husband according to Muslim customary law in 1978, but obtained alimony on the basis of a general criminal law against vagrancy. Her husband appealed to the Supreme Court arguing that Muslim law did not require him to pay alimony, but the court rejected that argument in 1985 with the further comment that the government should draw up a uniform civil code. Muslim organizations quickly responded to this judgment with country-wide agitations, claiming that Muslim personal law was in danger. The then Prime Minister, Rajiv Gandhi (1944–1991), gave in to their demands and overturned the Supreme Court decision by making a law which would no longer require a Muslim husband to pay alimony. This in turn was taken up by Hindu communal organizations which objected against the government's pro-Muslim stance, as well as by secular feminists who saw the government action as male chauvinism.

Communalist ideologies emerged in India in a variety of forms during the twentieth century. In the 1920s, V D Savarkar (1883–1966) published a book which connected Indian territory with Hindu identity and thus formulated a communalist ideology that made Muslims, Christians, and others into outsiders, at best second-rate citizens. His thoughts continue to play a significant role in the ideology of the Rastriya Swayamsevak Sangh (RSS), the leading Hindu communalist organization founded in 1925. This organization and its many branches are active in trade unions, political parties, and religious organizations, but it primarily tries to train young men in a military fashion to 'defend the Hindu nation.' The RSS has often been at the forefront of political violence against Muslims and Christians, before and after Partition. It was the driving force behind actions to demolish a sixteenth-century mosque in the Hindu pilgrimage centre Ayodhya in 1992, an event which made the political party with which it is connected into the leading party of the country.

On the Muslim side it is the Muslim League and the Jama'at-i-Islami which are, historically, the major communalist organizations. The Muslim League's

demand for a separate homeland for Muslims in the early 1940s began the political process which led to the Partition in 1947. The poet Muhammad Iqbal had formulated the wish for separation in his writings, but his main influence had been on Muslim students both in India and abroad. It only received political significance in the 1940s when Jinnah (1875–1948), the leader of the Muslim League, took it up. The speed with which the separationist movement gained support demonstrates the extent to which the political and social system in India was conducive to the communalist articulation of religious difference in India, with immense political consequences until the twenty-first century. The complexity of Muslim communalism can be demonstrated by the fact that the founder of the most militant Muslim organization in South Asia, the Jama'at-i-Islami—and as such the counterpart of the Hindu RSS—Maulana Maududi (1903–1979), was in principle against the formation of Pakistan. Partition, however, forced Maududi to move to Pakistan, where his organization became a redoubtable political force in campaigns for an Islamic constitution and laws against unIslamic practices. His revolutionary ideology had a great influence on the Muslim brotherhoods in Egypt and the Sudan.

Another major communalist development in India is the emergence of the Khalistan movement among Sikhs. Sant Jarnail Singh Bhindranwale (1947–1984), a young religious leader, created a large movement, popular largely among the urban youth, demanding a separate state for the Sikhs, Khalistan. It was repressed by the Indian army in 1984. The movement can be partly explained in terms of an identity politics in which Sikhs try to maintain the symbolic boundaries which divide them and the surrounding Hindu majority. These symbolic boundaries, such as the wearing of uncut hair and turbans for men, are only maintained if they are supported by the fact that Sikhs derive material gains from their membership, as they clearly did in the colonial period. The political and economic aspects of membership are intrinsic to the religious ones and it is therefore not useful to understand the Khalistan movement as either a religious or a political movement. It is like the Muslim and Hindu movements, a communalist movement which articulates the religious community in a political and economic way.

In the twenty-first century era theories of the decline of the nation-state emphasize the possibility of transnational, virtual communities which depend on new communication technologies. The extent to which these communities articulate politics of communal belonging that is not territorially bound is striking. This form of transnational communalism enables a politics which simultaneously connects very diverse issues in very different politics. The best example of this has been the so-called Rushdie affair, which started with the banning of the novel *The Satanic Verses* by the Indian government in 1988 under pressure from Muslim communalist leaders. This was quickly followed up by demonstrations, organized by leaders of 'the' Muslim community (mostly Indian and Pakistani) in the UK, and then followed by a call for the execution of the author by the self-proclaimed leader of the Muslims, the Sh'ite cleric and leader of the Islamic Republic of Iran, Khomeiny. The use of the notion of community in these political actions is widely varied, but simultaneously it allows some sense of unity across the globe. It is strongly tied up with migration to the North by members of religious communtities and thus with minority politics and secularism in the countries of immigration. At the same time it is engaged in communalist projects in the countries of origin, where these communities might constitute the majority, as indeed Sri Lankan Buddhists, Indian Hindus, and Pakistani Muslims do. It is this kind of transnational communalism, strongly supported by new media, which will become more important in the era of globalization. This should not be understood as the defense of primordial attachments in the face of modernization and globalization, but as completely new forms of community and identity politics.

See also: Diaspora; South Asia: Sociocultural Aspects

Bibliography

Dumont L 1970 *Homo Hierarchicus. The Caste System and its Implications*. Weidenfeld and Nicholson, London

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Hansen T B 1999 *The Saffron Wave. Democracy and Hindu Nationalism in Modern India*. Princeton University Press, Princeton, NJ

Jaffrelot C 1996 *The Hindu Nationalist Movement*. Hurst, London

van der Veer P 1994 *Religious Nationalism. Hindus and Muslims in India*. University of California Press, Berkeley, CA

P. van der Veer

Communes, Anthropology of

For purposes of anthropological investigation, 'commune' may be taken to refer to relatively small communities, which have set themselves up, typically self-consciously, as alternative to the wider social order, and embrace collectivism in some substantial fashion. They tend to be small enough to operate principally by virtue of face-to-face relations, they are inclined to hold resources and property in common, and usually, they favor communal decision-making for political purposes. This article treats the study of

such communities by professional anthropologists or ethnographic researchers.

1. 'Alternative' Communities

The great bulk of research on communes may be found in the disciplines of history, sociology, and psychology. Anthropological studies are relatively few in number, and only one, small book treats communes as a general topic (Hostetler 1974b). Considering that communes have much in common with the sort of small-scale community that anthropologists have preferred to investigate, it is curious that these scholars have not shown more interest in them. Like village communities, communes lend themselves to the kind of intensive fieldwork diagnostic of ethnography, and are especially accessible to the anthropological aspiration to research communities in the whole.

Perhaps one reason for the relative lack of anthropological attention is the consideration that in the social science literature, communalism has been largely associated with Western social settings. Whereas anthropology has focused on 'other cultures,' most (but not all) communes are not *other* in the relevant sense. They characteristically present themselves as essentially *alternative* to the societal orders in which they arise. However, in the present connection, significantly, the otherness entailed by such alterity can be misleading. For, as inverse and, in part or whole, utopian reactions to the dominant social order, such communities share with the latter rootedness in the same body of social thought. In the West this body of thought revolves around the question of the relationship between society and the individual, especially when that relationship is apprehended as a dualism. Seeing the dominant social order as primarily individualistic, communal societies propose to resolve the dualism to the contrary, by creating social and cultural conditions that ensure the valuational primacy of the collective over the individual. Typically, the realization of the primacy is keyed to institutional mechanisms that move the individual to grasp his or her self-identity primarily in terms of the collective.

2. Anthropological and Ethnographic Studies

The most substantial ethnographic studies of communes pertain to four different communal social settings: North American Anabaptist settlements, the Israeli kibbutz, Japanese communes, and the Soviet kolkhoz. Typically these studies take their cue from the presumptively alternative social nature of their subject, and accordingly tend to feature the way in which collectivism affects 'normal' institutions and practices, and the capacity of the community to withstand the opposing pressures of the surrounding social order.

2.1 The Hutterian Brethren

Hostetler's work on the Hutterites is exemplarily anthropological, in its ethnographic comprehensiveness and depth of cultural knowledge. He studied this people over a period of several years, and his publications (see especially Hostetler (1974a) and Hostetler and Huntington (1996)) provide rich descriptions of the 500-year history of the Hutterites, their worldview and critical system of Christian belief and practice, the colonies' economic, political, and domestic organization, and the difficulties experienced 'today' by this people in sustaining their unique communal way of life. But the principal focus of his research is socialization. Given the long-term survival of this social order under presumptively adverse external societal and cultural conditions, Hostetler supposed as the key to this evident success the inculcation of cultural values. Accordingly, he set out to examine the Hutterite child-rearing system. He finds that the system of socialization is remarkably effective for reasons of its ideological certitude, utter consistency, and understanding attitude toward modest forms of individual deviancy.

The system is predicated on the Christian belief in the absolute authority of God and the presumption of the primacy of communal order as the key to ensure the submission of the will of the individual self to the will of God. Human nature is seen as dualistically divided between carnality and spirituality. The former quality is understood as in need of firm submission to the latter, a condition meant to be realized by means of communal existence. All the institutional practices of the Hutterites, including of course the education and training of children, are designed to function strictly on behalf of this spiritual principle of hierarchical authority. In result, Hostetler concludes, the Hutterite's sense of individual self is only weakly distinguished from his social identity and is predisposed to self-denial and obedience to authority. In view of this analysis, Hostetler (in line with Hutterite opinion, and perhaps tautologically) is inclined to explain defection as a product of aberrant circumstances in which individuality emerges as an unruly force. These circumstances are, he suggests, associated with undesirable competition among family lines and the failure of these families truly to subject their autonomy to the social whole.

In his substantial monograph on the Hutterites, John Bennett (1967), like Hostetler, is interested in how this order of communal life perdures and maintains its traditional integrity, given the contrary nature of the surrounding social environment. But instead of socialization, Bennett focuses on the agricultural economy of the colonies and the impact of technological development. In line with the general tenor of Hostetler's account, he argues that for reasons of their sure sense of identity, extremely efficient man-

agement, well-formed social order, and rational approach to economic matters, these communities are exceptionally adaptable. More particularly, because they are very open to technological change but not to consumeristic influence, they are enabled to thrive without corrupting their traditional way of life.

Zablocki's philosophically rich study of the Bruderhof (Zablocki 1971), another deeply religious Anabaptist community, less traditionalist than the Hutterites, also constitutes an important contribution to the anthropology of communes. Zablocki is a sociologist, but by virtue of his intensive fieldwork in a single settlement, and his attention to the quotidian course of social events, his study is patently ethnographical. He, too, was interested in the relative success of the Bruderhof. He addresses this issue in terms of 'commitment,' a notion pertaining to the fusion of self-interest with the needs of the community and given currency in Kanter's *Commitment and Community*, a sociological study (contemporary with Zablocki's) of communes in general (Kanter 1972).

In addressing the question of the manner in which the Bruderhof creates and sustains the commitment of its members, Zablocki takes, in addition to the standard functionalism of Hostetler's and Bennett's work, a strong psychodynamic turn. Unlike Hostetler, whose approach to the question of commitment typifies the 'culture and personality' school of interpretation, Zablocki draws on psychoanalysis in a way that bears directly on sociological dynamics, not personality. His main finding is that the Bruderhof is characterized by an institutional structure the critical function of which is to call on and employ the ordinarily repressed psychodynamic energy of collective behavior in order to sustain commitment. Hence, when the community suffers a social crisis that threatens commitment on the part of members (such crises are, he finds, inevitable), this structure of resolution—a joyful prayer meeting of the whole—is brought into play, revitalizing commitment. In effect, he depicts a community keyed to something like Rousseau's 'general will' (a concept of 'commitment' if ever there was one), the difference being that in the Bruderhof this will is regarded as accessible (it is thought to pre-exist its concrete determination) only by means of the realization of a spiritual rather than rational or political unity. By arguing that this unity is brought about through collective behavior and affective effervescence, Zablocki charges Durkheim's famous but vague sociological thesis about the transcendent force of society, with a determinate psychoanalytical dynamic.

Underlying Zablocki's concern for commitment is a general concern about the relation between community and freedom. This concern is of interest to him for reasons beyond the question of the Bruderhof, as he finds that American society in general suffers from a need for community. Hence, for him the examination of the Bruderhof carries lessons for the society at large, the most prominent of which is the importance of distinguishing a kind of freedom that is nonindividualistic.

2.2 *The Israeli Kibbutz*

Melford Spiro's (1970) first book on the kibbutz continues to serve as an anthropological introduction to this secular communal society. It presents the community's historical beginnings, ideological principles, and institutional practices. It was originally intended to serve as a preliminary chapter to the earlier book resulting from Spiro's basic research project on the kibbutz, namely a study of socialization and personality development (Spiro 1965). Like Hostetler on the Hutterites, Spiro was intrigued by this communal enterprise's goal of recreating man in society's image, making him fit to live in social harmony (the kibbutz ideology's hoped-for 'New Man'). And like Hostetler, Spiro approached this problem as a question of the causal relationship between culture and personality, presuming that the personality of the founders' children must take its shape from the social arrangements fashioned by the parents. In particular, seeing the goal of communalism as essentially incompatible with conflict and hostility, Spiro sought to determine the degree to which kibbutz socialization and its system of 'collective education' were successful in eliminating the 'aggression drive' in the children reared in this way. This is the research theme around which centers his psychologically expert study *Children of the Kibbutz* (the title is an expression of a kibbutz ideological shibboleth), the major product of his research on this community (Spiro 1965). As a result of his inquiry he finds what he already anticipated in the epilog of his first book on the kibbutz and is consistent with his penchant for basic psychoanalytical theory: that aggression, being a universal drive, can be curbed but not eliminated.

Spiro's psychoanalytical naturalism receives even greater emphasis in his later study of women in the kibbutz (Spiro 1979), a topic presenting itself in the light of the ideological aim to establish equality between the sexes. Here he concludes, unequivocally, that gender–role differentiation is due to 'precultural' factors and therefore not subject to cultural elimination. In fact, the direction of Spiro's explanatory exercises was signaled, even before the publication of his first volume on the kibbutz, in a provocative article (Spiro 1954), the thesis of which is that the kibbutz had managed by cultural design to do away with the family in favor of the undiluted integrity of the collective whole. A few years later, however, retracting this thesis and foreshadowing his naturalistic universalism, he concluded that the family does constitute a distinct feature of kibbutz social structure (Spiro 1960). The merits of his psychological orientation aside, Spiro's work on the kibbutz is astutely and invaluably attentive right from the start to the stream of social

problems that bear on the place of the individual and run steadily beneath the ideological and, even, empirically minded self-presentation of this communal society.

In *Two Kinds of Rationality*, Evens (1995) treats the question of liberalizing change in a kibbutz in relation to democratic procedure and generational conflict. His phenomenological emphasis on practice and tacit knowledge sharply distinguishes his approach in the present connection. It leads him to look beneath the ideology for, rather than biological or psychological causes, contradictory understandings that are so taken for granted they predispose their hosts, whose self-identity cannot be told apart from these understandings, to certain behaviors. He traces the understandings in question to Rousseau's contractarian theory of the relationship between the individual and society, and then to the sense of man and generation given in the Hebrew myth of Genesis. Like other students of communal societies, Evens is moved to use his research as a basis on which to develop a theory of human conduct; in his case, one keyed to an ontology of ambiguity and a considered notion of ethics. In an earlier essay (Evens 1975), he examines a case of stigmatization in the kibbutz, in order to bring into relief the deepest meanings of the most basic social position in this society, that of 'member' or *chaver kibbutz*. He found that these meanings betray implicit, unreflected ideological expectations of failure in the face of perfectibilism, and foster defensive but disturbing social acts.

Allison Bowes's (1989) *Kibbutz Goshen* not only serves as a sound ethnographic introduction to the kibbutz but also analyzes prominent social problems characterizing the community in which she did her research. Problems of hired labor, competition over work roles, the re-emergence of familism, the sexual division of labor, and conflict between the generations are all subjected to eclectic but useful and empirically substantive sociological analysis. More broadly, turning a remedial eye on the very idea of ideology, she contends that the definition of these apparently troubled social circumstances as problems indicative of failure derives in large part from a misguided tendency to see kibbutz ideology as a set of fixed and decisive principles rather than a pliable and adaptive system of values.

The problem of work roles is the focus of Israel Shepher's (1983) ethnography of a kibbutz (see also Shepher 1980). He is concerned with the threat to the practice of equality posed by the development of permanent work roles, given increasing specialization in the kibbutz economy and the concomitant demand for expert workers. He argues that in so far as this trend is comprehensively applicable to every member, the creation of a structure of inequality based on work roles is precluded, since everyone is equally indispensable to his or her place in the labor sphere. In *Women in the Kibbutz*, Tiger and Joseph Shepher (1975) take up the question of sex–role differentiation. Their aim is to use the example of the kibbutz to arrive at universal truth in respect of the place of women in society. Like Spiro, they propose a naturalism, but their argument is sociobiological rather than psychoanalytical, keyed to the concept of 'biogrammar' or the genetic transmission of behavioral characteristics.

2.3 *Japanese Communes*

David Plath's work on utopian communities is of special interest because it pertains to a non-Western social setting. In a brief discussion, representative of the usual analytical focus on the struggle of such communities to flourish in the midst of contrary social and cultural conditions, Plath (1966) examines the tactics used by four Japanese communal societies to adapt to the ambient milieu without compromising their reformatory goals. However, in his perceptive epilogue to *Sensei and His People* (Sugihara and Plath 1969), a first-person narrative of a Japanese commune authored by the wife of the commune's charismatic founder and leader, Plath identifies suggestive differences between this and Western communal ventures. He observes that, having evolved spontaneously, Shinkyo (the name of the commune) cannot be regarded as an 'intentional community,' and he brings into relief the fundamental way in which this community is continuous with, not simply reactionary to, the ambient society. The commune got its beginnings in reaction to the formalism of Tenrikyo, a nineteenth century Japanese religious movement related to Buddhism (Shinkyo's founder had served at one time as a Tenri priest), as well as to the 'bossism' of the local village community. But its defining emphases—moral responsibility, self-sacrifice, mutual reliance, social equality, and spiritual/bodily cleanliness—are manifestations of the traditional values of rural Japanese society. In view of the overtly complementary relationship between Shinkyo and the surrounding community, Plath's exercise suggests that this commune, although radical by the actual standards of the state of things against which it reacted, practices a nonabsolutist communalism. As a result, by contrast to Western communitarian ventures, the identification with the social whole is not so much secured institutionally as left—not simply in the end but for all practical purposes—to the individual member's quotidian discretion.

2.4 *The Soviet Kolkhoz*

Caroline Humphrey's (1983) substantial social anthropological study of the Soviet collective farm or kolkhoz may be mentioned here but not really considered. For one thing, as a direct function of the state, the *kolkhoz* is neither alternative in the relevant sense

nor self-generating. For another, although one can learn much about the social and politico-economic life of Soviet collectives from Humphrey's study, it is more directly concerned with problems of social change deriving from the Soviet Union's need to integrate into itself ethnic minorities (in this case, the Mongolian Buryat) than with collectivism itself. As a result, it is difficult to place Humphrey's book directly in the literature on communes, and, apart from her extensive use of materials published by ethnographers and analysts native to the society in which she worked, she herself neither draws on nor mentions that literature.

3. Communes and the Significance of Anthropological Research

Given the intensive nature of ethnographic fieldwork, it is not surprising that the anthropological study of communes tends to be distinguished by sharp attention to the discrepancies between communal practice and ideological design. In staying acutely mindful of this question, the anthropologist is dwelling on a theme at least implicit in the very operation of such communities. The ideological presumption of dualistic opposition between the individual and society defines the communal endeavor as dependent on self-sacrifice for its success, thus projecting the endeavor as a trial of transcendence in the face of material adversity. The adversity is typically apprehended as a matter of the pressures exerted by the external social environment. It is also associated, by communards and researchers alike, with a naturalistic presumption of one kind or another—especially self-interest and individualism. The research interests of socialization, adaptability, internal conflict, women's and work roles, familism, crisis management, commitment, etc., revolve round the ideological picture of these communities as social experiments under hostile conditions, and amount implicitly to attempts at measuring the success or failure of the communitarian enterprise.

As informative as the anthropological focus on the gap between theory and practice is, though, it falls short of the unique advantage anthropology has to offer, for it tends to confine the anthropologist to the communal endeavor's ideological social compass. That compass misleads, in part but significantly, by the presentation of itself as alternative or 'other': it is after all very much an integral function of the social universe against which its alternance is defined. In principle, however, even if working 'at home,' the anthropologist takes his or her direction from a more radical otherness, one that cannot be reduced to what is already known, at least not without disrupting or altering the latter. Thus, Plath's work in Japan may well disclose a communalism that is not dualist in the received (Western) sense of the term, and Evens's on the kibbutz, digging deconstructively beneath the ideology to a nondualist mythic core, begins to rethink the very idea of the difference between theory and practice, away from its implicit perfectibilism and towards a relativism that serves to reform 'ethics' as the very name of human existence.

Given the radical way in which they tend to betray a dualism of the individual and society, communal orders constitute exceptionally fertile ground for anthropological study. Because they thus magnify a fact that ordinarily gets taken for granted, their study helps explain, as Durkheim asserted of exceptional ('pathological') social phenomena, the norm. By the same token (the conspicuous disclosure of individual–society dualism), since the relation between the individual and society is also a predicative question for social science inquiry, the study of these societies critically orients self-examination into anthropology's own tacit foundations.

See also: Anthropology; Collective Beliefs: Sociological Explanation; Collectivism: Cultural Concerns; Communalism; Community, Expression of; Community, Social Contexts of; Community Sociology; Groups, Sociology of; Totemism

Bibliography

Bennett J W 1967 *Hutterian Brethren*. Stanford University Press, Stanford, CA

Bowes A M 1989 *Kibbutz Goshen*. Waveland Press, Prospect Heights, IL

Evens T M S 1975 Stigma, ostracism and expulsion in an Israeli kibbutz. In: Moore S F, Meyerhoff B G (eds.) *Symbol and Politics in Communal Ideology*. Cornell University Press, Ithaca, NY, pp. 166–209, (Revised 1980 as *Stigma and Morality in a Kibbutz*. In: Marx E (ed.) *A Composite Portrait of Israel*. Academic Press, London, pp. 179–210)

Evens T M S 1995 *Two Kinds of Rationality: Kibbutz Democracy and Generational Conflict*. University of Minnesota Press, Minneapolis, MN

Hostetler J A 1974a *Hutterite Society*. Johns Hopkins University Press, Baltimore, MD

Hostetler J A 1974b *Communitarian Societies*. Holt, Rinehart and Winston, New York

Hostetler J A, Huntington G E 1996 *The Hutterites in North America*. Harcourt Brace, Fort Worth, TX

Humphrey C 1983 *Karl Marx Collective: Economy, Society and Religion in a Siberian Collective Farm*. Cambridge University Press, Cambridge, UK

Kanter R M 1972 *Commitment and Community: Communes and Utopias in Sociological Perspective*. Harvard University Press, Cambridge, UK

Plath D 1966 The fate of utopia: adaptive tactics in four Japanese groups. *American Anthropologist* **68**: 1152–62

Shepher I 1980 Social boundaries of the kibbutz. In: Marx E (ed.) *A Composite Portrait of Israel*. Academic Press, London, pp. 137–177

Shepher I 1983 *The Kibbutz: An Anthropological Study*. Norwood Editions, Norwood, PA

Spiro M E 1954 Is the family universal? *American Anthropologist* **56**: 839–46

Spiro M E 1960 Addendum, 1958 (to Is the family universal?). In: Bell N W, Vogel E F (eds.) *A Modern Introduction to the Family*. Free Press, Glencoe, IL, pp. 72–75
Spiro M E 1965 *Children of the Kibbutz*. Schocken Books, New York
Spiro M E 1970 *Kibbutz: Venture in Utopia*, augmented edn. Schocken Books, New York
Spiro M E 1979 *Gender and Culture: Kibbutz Women Revisited*. Duke University Press, Durham, NC
Sugihara Y, Plath D 1969 *Sensei and His People: The Building of a Japanese Commune*. University of California Press, Berkeley, CA
Tiger L, Shepher J 1975 *Women in the Kibbutz*. Harcourt Brace Jovanovich, New York
Zablocki B 1971 *The Joyful Community*. Penguin Books, Baltimore

T. M. S. Evens

Communication and Democracy

Democracy can be defined in many different ways. This is one reason why so many different kinds of democratic states have existed. It can be an empirical scientific term, referring to the common people's participation in the political decision-making process, but it can also be a normative ideological concept referring to an ideal political system. In order to make historical and international comparisons possible, the term 'democracy' is used here in the empirical scientific sense, referring to the phenomenon by which all eligible members, directly or indirectly, participate in the decision-making process of the society to which they belong. Here the term democracy is not used for an ideology or a political goal.

The first half of this article reviews how the development of communication technologies, and the spread of both communication media and education among the masses, contributed to the evolution of democracy. The second half of this article deals with how modern communication media, especially mass media, function in a pluralistic democracy.

1. History

1.1 Democracy before the Mass Media Age

In major civilizations before the seventeenth century, the majority of the population was illiterate, and in early medieval times so were some European kings. The Vatican initiated a campaign in 527 AD to increase the number of literate priests. Even in late sixteenth century England, the illiteracy rate was as high as 80 percent among men and 95 percent among women (Parker 1980).

Under such constraints, 'democracy' as defined above was possible only in small communities or societies in which individual members could discuss the issues face to face. The 'direct democracy' of the ancient Greek city-states is the best known example of this. However, a long list can be made of other ancient and medieval civilizations with similar systems or practices: the first century German tribes (according to Tacitus), the Swiss cantons, the New England town meetings, and *yoriai* (meeting) in traditional Japanese villages.

While small-scale direct democracies at the city or village level existed, democracy at the empire, kingdom, or state level did not exist anywhere in the world before the seventeenth century. The main reason for this is that it was technically impossible.

1.2 Evolution of Democracy as a Result of Mass Printing

The illiteracy rate in Europe dropped rapidly in the seventeenth century. In England it declined to about 55 percent for men and 75 percent for women by the early eighteenth century (Parker 1980). This was true in some non-Western countries as well. In Japan, for example, the illiteracy rate was very high before the seventeenth century, even among the elite *samurai* (warrior) class. Peace during the early seventeenth through the nineteenth century, however, gradually 'civilized' ordinary people. As early as the late seventeenth century, Ihara Saikaku (a popular novelist, 1642–93) describes an illiterate *samurai* in one of his novels as 'sadly behind the times' (Dore 1965).

A natural consequence of an increasingly literate population was an ever increasingly open expression of political opinions and criticisms by middle-class intellectuals: low-ranking aristocrats, rich farmers, successful merchants, and so forth. The almost predictable response from the traditional political elite was repression, which was very severe and cruel in the sixteenth and seventeenth centuries but became gradually less so during the following two centuries. During this time, at least some constructive opinions and policy suggestions by middle-class intellectuals gradually came to be tolerated and incorporated into the policy-making considerations of the political elite.

This process started in England, with the reinstatement of a parliament in 1640 that demanded assurances against a resumption of the 'absolute monarchy' of Charles I, and spread to Western Europe and North America. Similarly, although later and slower, changes took place in some non-Western countries as well. For example, Japanese media historians point out that the control of publications by the central government became more relaxed and less widespread during the nineteenth century compared with the seventeenth century, due chiefly to the remarkable increase in the number of publications, which included private news

media such as the *kawaraban*, the first of which was established in 1615. By the early nineteenth century nationwide networks for the distribution of private primary school (*terakoya*) textbooks had been established, and political books banned in Tokyo flowed unhindered to distant and remote areas of the country through the textbook distribution pipeline (Konta 1977).

The increase of publications intensified the need for intellectuals and experts in the government bureaucracy who could understand and evaluate all of these published books and documents. This was especially noticeable in the fiscal, monetary, agricultural, and foreign policy areas as well as in the judicial system. Consequently, experts and scholars serving government came to be recruited based on their merits and competence rather than on their lineage. Specialization led to the creation of government agencies that dealt with defined and narrower policy areas. This division of responsibility within the new 'meritocratic' establishment meant the greater participation of lower ranking people in the political decision-making process.

As Cooley ([1909]1993, p. 75) stated, 'Printing means democracy, because it brings knowledge within the reach of the common people; and knowledge, in the long run, is sure to make good its claim to power.' This is equally applicable to the histories of both the East and the West.

1.3 Crisis of Democracy in the Mass Media Age

As discussed above, premodern undemocratic governments assumed that the masses should not participate in the political decision-making process; therefore they suppressed publications that criticized them and also any efforts at free expression on the part of their subjects. However, they did not attempt to operate their own mass media to influence the masses. This only began to happen with the emergence of new types of undemocratic governments that resulted from the development of mass media. Most of them took advantage of democratic tenets, such as free elections and freedom of expression, to seize power. Once they had seized power, however, they abolished or drastically restricted those fundamental democratic rights and became dictators. With a twisted logic, some of them persisted in calling their regimes democratic because the majority of the population once supported them. Fascist, Nazi, ultra-nationalist, militarist, communist, and radical socialist governments in the twentieth century all belong in this category.

When newspaper readerships were limited to well-educated middle-class intellectuals, they contributed to the evolution of democracy (as Cooley [1909]1993 points out); however, when newspaper circulation expanded further to include the general masses, newspapers sometimes changed to engines of mass agitation and exerted a dangerous influence on national policies.

In the past, nearly every modern democracy has had some experience of this. For example, it is well known that the sensationalism and jingoism of the William Randolph Hearst newspapers were responsible for the Spanish–American War of 1898. The British newspapers were largely held responsible for the unreasonably onerous reparations imposed on Germany at Versailles following the First World War (Nicolson 1969).

Chauvinistic, ethnocentric, and racist newspapers helped undemocratic political parties to seize power in Germany and Italy. In Japan, even though both the Emperor and the civilian government tried to curb Japan's military actions in China, many newspapers supported the military (to expand their circulation) and helped them take over the government.

When socialist or communist parties (hereafter 'socialist' includes communist) took power in Russia (1917), in China (1949), and many other countries, the majority of the masses were still very poor and illiterate. So when socialist governments emerged as the result of violent revolutions they were at first regarded as another variation of democracy, because the revolution enabled the traditionally repressed classes such as farmers and manual workers to participate, for the first time in their history, in the political decision-making process at the national level. When these socialist governments were first established, they were vulnerable and feared a counter-revolution by reactionary feudalistic forces. As they firmly believed that justice and truth based on Marxist socialist scientific theories were on their side, they did not permit any political parties other than their own, nor any antisocialist expressions, nor a free and open press, and certainly not free and secret elections. Socialist governments operated their own propaganda machines to educate the masses in an effort to catch up with the more economically and technologically advanced capitalist liberal democracies.

The socialist system that started off as one more variation of democracy, however, soon became more and more undemocratic as the power of the state came to be monopolized by a handful of the socialist party elite. These elites gradually evolved into special privileged classes such as the 'nomenclature' in Soviet Russia. In the most extreme cases, such as those in North Korea and Romania, the position of head of the state was completely privatized and was passed on, or an attempt was made to pass it on, to the head of state's son. The maintenance of a dictatorship by the socialist party elite depended on the strict control of mass communications, and widespread surveillance and coercion by the secret police.

Although these dictators managed to control the domestic mass media, it became harder to control the inflow of information from foreign countries as

communication and transportation technologies developed. Contact with foreigners through tourism, conferences, letters, the telephone, and other communication channels steadily increased in all socialist countries.

People trust the mass media only if their contents are congruent with their own experiences and observations. If the output of the mass media and the people's direct experiences and observations are incongruent over a long period, the credibility of the mass media is gradually lost. If the government operates the mass media, the credibility of the government is then jeopardized.

Based on the recognition of this reality, the Soviet Union leadership started the liberalization of information, called *glasnost*, in the late 1980s. The initial purpose of this policy was to allow free flows of information, a free press and freedom of speech within the socialist democratic system. Soon, however, it became obvious that a one-party dictatorship could not coexist with the policies of *glasnost*.

Wherever there is a free flow of information, freedom of the press and free speech, many divergent political ideologies and policies emerge and compete with each other for public support, thus making it impossible to maintain a one-party dictatorship. Some of these parties might be ethnocentric, racist, fascist, or even fanatical, and therefore undemocratic as the socialists have always claimed. However dangerous they might be, they do constitute a part of a modern 'pluralist democracy' that many countries in the world, including former fascist and socialist countries, have reached only by paying the high cost of social unrest, wars, and violent revolutions.

2. *How Do Communication Media Function in the Modern Pluralist Democracy?*

What is evident from a historical review of the relationship between the forms of communication (especially mass media) and the power elite is that in any society with a well-educated population criticisms of government policies and demands for participation in the political process are bound to be heard. There are only two ways to deal with this eventuality: either to suppress these criticisms and demands or to coexist with them. Historically, political leaders of almost all the countries in the world have chosen the former option at first. In countries where printing developed earlier than in other parts of the world, however, the latter choice, coexistence, was preferred, but only after much struggle and bitter experience. Thus, 'free criticism of the government,' which is practically the same as 'freedom of the press,' became the paramount prerequisite for a modern pluralist democracy.

In modern societies where communication technologies are highly developed and the freedom of the press is permitted, mass media have to compete not only with each other but also with all other forms of communication media and information flows. If a mass medium loses credibility with its consumers, it will lose their support. Under the capitalist market system this could lead to financial ruin, whereas under a one-party dictatorship this could mean the loss of the people's trust in the party and a subsequent toppling of the government.

As discussed elsewhere (see *Influence: Social*), the mass media appears to have some influence on individual attitudes, opinions, and behavior under certain conditions. However, even if it is proved that people are influenced by the mass media that they are regularly exposed to, the political implications for society as a whole are not quite clear because the mass media in pluralist democracies represent such a diversity of opinion. As long as the mass media remain diverse, their influence on actual policy and decision making is unclear. The political implications of mass media's influence on individual opinion become clearer only when the entire mass media become monolithic or united in a common cause. In what kinds of situations do mass media coalesce their opinions and have a marked influence on government policy and decision making?

2.1 *Agenda Building*

For the reasons mentioned above, the persuasive effects of a particular journalist or mass media company on political leaders are limited and may be marginal. However, when the mass media continually dwell on problems that the government is supposed to have solved but has not, due to negligence, then they can have a direct impact on government policies.

This phenomenon usually takes place in areas such as the environment, the rights of minority groups, and the handicapped. These issues tend to be ignored or forgotten partly because they do not always have powerful lobbying groups to promote them. Just by pointing out the existence of these problems through their investigations, the mass media sometimes succeed in mobilizing government agencies and politicians into creating a solution. This is the 'agenda-building function' of mass media's 'investigative reporting.'

2.2 *Enforcement of Social Norms*

It is said that while the police and prosecutors enforce laws, mass media enforce morals and ethics. There are many immoral or unethical acts committed for which people may not be arrested or prosecuted but which they do not want exposed. Money and sex scandals are typical examples. Lazarsfeld and Merton (1960) called

this social function of the mass media 'the enforcement of social norms.' The power of the mass media is so feared by political leaders just because celebrity scandals are more newsworthy than those of ordinary people.

Personal scandals do not always stay that way. When they are perceived as symptomatic of 'organizational corruption,' the subsequent revelations can have a direct impact on government policies. The introduction of competition into traditionally monopolistic areas, and drastic reforms of government organizations and systems, are often the result of widely reported exposés of such organizational corruption.

2.3 Mass Media–Public Opinion Congruence

Apart from basic social norms, public opinion is clear and decisive on certain issues. Taxes and war or military involvement are good examples. On these issues, the mass media tend to follow public opinion (partly because of their fear of losing circulation or ratings) and the mass media as a whole become less diverse as a result. Competition among relatively monolithic mass media further reinforces public opinion, and their combined demands and criticisms of the government become intensified. The term 'climate of opinion' has been used to explain this pressure. Under these circumstances, the government is often obliged to concede to the wishes of the mass media–public opinion alliance. This alliance is democratic in principle and, therefore, usually desirable. As described above, however, the possible danger of such an alliance in a mass society should not be ignored.

See also: Democracy; Democracy, History of; Democratic Theory; Electronic Democracy; Influence: Social; Literacy and Illiteracy, History of; Mass Media, Political Economy of; Mass Media, Representations in; Media Effects; Media, Uses of; News: General; Participation: Political; Political Advertising; Political Communication; Political Culture; Political Discourse; Printing as a Medium

Bibliography

Cooley C H [1909] 1993 *Social Organization*. Transaction, New Brunswick, NJ

Dore R P 1965 *Education in Tokugawa Japan*. University of California Press, Berkeley, CA

Konta Y 1977 *Edo no Hon'ya-san (Bookstores in Edo)*. Nihon Hoso Shuppan Kyokai, Tokyo

Lazarsfeld P F, Merton R 1960 Mass communication, popular taste and organized social action. In: Schramm W (ed.) *Mass Communications*, 2nd edn. University of Illinois Press, Urbana, IL, pp. 492–512

Nicolson H 1969 *Diplomacy*, 3rd edn. Oxford University Press, London

Parker G 1980 *Europe in Crisis, 1598–1648*. Harvester Press, Brighton, UK

Y. Ito

Communication and Social Psychology

Human communication is both social and cognitive because it is a process by which individuals exchange information and influence one another through a common system of symbols and signs. Compared with the study of language more generally, the study of the psychology of language is relatively new, and, when it did emerge as a field in the middle of the twentieth century, more attention was paid to understanding the psychology of grammar than to understanding language use or communication *per se*. Moreover, what attention was paid to understanding communication reflected mostly a cognitive perspective rather than a social perspective. The social psychological perspective on communication, then, is a very recent development within the study of language.

The social psychological perspective on communication complements the more purely cognitive perspective (Cherry 1957, Shannon and Weaver, 1949). The cognitive perspective treats communication skills much like language skills by identifying a set of procedures, mechanisms, and operations by which information is transmitted from some source to some recipient. The social psychology of communication emphasizes the fact that communication also involves social motives and has interpersonal consequences beyond information reception. From a social psychological perspective, communication is a social action that shapes reality, and not just a device for transmitting information from one location to another. Communication is action (see Searle 1969), and action is 'social' in so far as its subjective meaning and orientation takes other people into account. By the communicator and the audience taking each other into account, communication both creates new knowledge and constructs a shared reality (see Higgins 1992). The nature and functions of communication—what it is and what it does—reflect the fact that it is a social action.

From a social psychological perspective, communication is an interpersonal interaction, a game with interaction rules and social motives. The constraints on communication are not just the basic information to be transmitted or the information processing capacities of source and destination. There are, in addition, the social rules of appropriate communication, the social context of communication, and both what the communicator intends the message to do to

the audience and what the audience infers about the communicator's intentions. The social psychology of communication is also concerned with the social consequences of communication, including how communication influences audiences' and communicators' attitudes, impressions, attributions, and memory about the message topic. Since the 1950s, social psychological researchers have discovered several phenomena that reveal the different ways in which communication as a significant social action shapes the world in which we live.

1. *Taking the Audience Into Account*

Communicators do not simply encode some set of information and then transmit it to a destination. They choose how to formulate their message by taking the characteristics of the audience, including their audience's situation, into account. When another person is the topic of their message, they will encode the same information about that target person differently depending on what they think is the audience's attitude toward that person or what they think the audience already knows about that person, including what information they share in common with the audience. For example, people give different directions to a destination when talking to a 'local' than to a 'foreigner'. Communicators will even convey different information at the same time to more than one audience when they have a different relationship with each audience, such as using expressions that will have a different meaning to an in-group member than to an out-group member (see Fleming 1994). Communicators also take the audience's context into account, such as describing a target person differently depending on the context of other people from whom the audience must distinguish the target (e.g., Manis and Armstrong 1971; for a review, see Krauss and Fussell 1996).

When communicators take their audience into account in formulating their message, they are constructing a representation of the topic of the message that is shaped by their audience. This audience-shaped representation itself has effects beyond the original information to be transmitted. It can change communicators' own knowledge and attitudes, including their own attitude toward themselves (i.e., their self-esteem) if they spontaneously communicated a favorable image of themselves to others. It has been found that communicators' own evaluations of a target person, and what they remember about the target person's behaviors, will change over time to match the evaluative tone of the message they communicated about that person weeks before, a message that was originally designed to take into account the audience's positive or negative attitude toward the target person. Thus, what communicators end up saying about a target person in order to suit a specific social context ends up becoming what they believe about the target person after a period of time. This 'saying is believing' effect exemplifies how communication as a social action can shape people's world (see, for example, Higgins 1992).

2. *Taking the Communicator Into Account*

Audiences also do not simply decode information transmitted by some source. They believe that the message was a social action. Thus, they infer that the communicator chose to formulate the message in a particular way and that this choice is significant. They infer, for example, that communicators tailor their message to suit the attitude of the audience, and adjust accordingly their judgments about what the communicators truly believe (for reviews, see Higgins 1992, Krauss and Fussell 1996).

Messages are intentional acts, and inferring the communicator's intentions is critical to understanding the meaning of the message (see Grice 1975). A central assumption about how conversations should be conducted is that all information contributed by the participants during the conversation should be relevant to the goal of the conversation. When one participant in a conversation breaks this rule, it can create problems for the other participants. It has been suggested, for example, that the poor performance of participants in some decision-making studies could be because the experimenters in these studies broke this conversational rule of relevance by providing information to the participants when communicating to them that was not relevant to the participants' decision task. If the participants assume that the experimenter is following the rule of relevance, they will use the irrelevant information the experimenter has communicated to them and perform poorly (see Schwarz 1994). In order to understand the full meaning of a message, audiences must understand not only what information is actually transmitted in the message, but also infer what meaning the message conveys given a particular communicator and particular context. The meaning of a message is influenced by the communicator's characteristics because the audience uses these characteristics to draw inferences about which meaning of a word or phrase was intended by the communicator. For example, an audience would infer a different sport for the word 'football' depending on whether the communicator was known to be British or from the USA. By also taking the context of communication into account, audiences make different attributions about the communicator's intentions in producing a message, which changes the meaning of the message (see Hilton 1995).

Because messages are intentional acts, they provide information about the communicator and not just information about the message topic. Information about the communicator, in turn, can influence the

illocutionary force of a message—its perceived purpose and effectiveness as an action (see Searle 1969). The effectiveness of a persuasive message, for example, does not depend solely on the informativeness of its content, such as the quality of its arguments. It also depends on the characteristics of the communicator (e.g., perceived expertise, attractiveness) and the nature of the relationship between the communicator and the audience. Thus, both the meaning and the impact of a message are influenced by audiences' taking the characteristics of the communicator and the context of communication into account.

3. *Messages as Strategic Actions to Influence the Audience*

The significance of messages and the responses of audiences to them are not restricted to their symbolic meaning. As social actions they can be used strategically to imply causality and generalizability or to signal relationships. They are socially important as well as meaningful. Thus, they have significance both as signals and as symbols. Recent research has shown that interpersonal language has a number of inference-mediating properties. This research is based on a model of interpersonal language by which an objective and quantitative analysis of the language used is possible, such as how abstract or concrete the message is (see Semin and Fiedler 1988).

Subtle and relatively small differences in the level of abstraction of a message convey differential inferences to audiences. For instance, systematic variations in abstraction level of a message lead audiences to draw different inferences as to whether the behaviors of a person are influenced by situational factors or due to his or her personality makeup. Concrete representations of events lead people to think that the behaviors of an actor were due to situational constraints, whereas depicting the same event abstractly leads people to think that the behaviors were due to the personality makeup of the actor.

Applications of these findings in communication contexts have shown the role that strategic language use plays in influencing the types of inferences audiences make. One instance is the transmission and maintenance of stereotypes. For instance, it has been shown that behaviors that are consistent with stereotypic expectations are communicated at a higher level of linguistic abstraction compared with behaviors that are inconsistent with expectations. When audiences receive such messages, one finds that expectancy-consistent messages lead to the conclusion that these behaviors were due to enduring personality factors. In contrast, unexpected behaviors are depicted with concrete language, suggesting that such behavior is situationally constrained and thus unlikely to be repeated. This effect has been shown to be mediated by the level of abstraction of the messages. Diverse research findings demonstrate that subtle and implicit properties of language, such as abstraction, are strategically used, albeit unconsciously, in the composition of messages with a view to influence audience inferences. More importantly, these subtle linguistic cues in messages have been shown to influence the audience as intended but without their knowledge.

4. *Audience Responses to Messages as Signals of Relationships*

Just as communicators can influence their audience through strategic language use, audiences can respond to messages in ways that influence communicators. These responses can be verbal or nonverbal and they can be intended or unintended. One basic set of phenomena concerns the different ways in which audiences can influence communicators' functioning by signaling their relationship to them (see Giles and Coupland 1991). If the members of a social group, for example, perceive their social identity to be threatened by how members of a different social group communicate with them, they might respond by displaying behaviors that accentuate the difference between the two groups, such as exaggerating their distinct accent or using expressions that are distinctly associated with their own group.

Communicative acts and styles, such as speech rate, pausal phenomena, utterance length, facial expressions, gaze, and so on, are used when responding to others' communicative acts to accentuate differences (divergence) and reduce differences between groups (convergence). These are strategies by which individuals adapt to other people's communicative behaviors toward them. When communicators respond to different people whose social status and education vary, and whose power and closeness relationships to them vary, they will modify their use of phonology, median word frequency (an index of formality), and other language features in order to signal similarity with whoever is their current communicative partner. In communicative interactions, both the communicator and the audience use verbal and nonverbal actions to manage the discourse, to signal and control the relationship, and serve other interpersonal goals (see Giles and Coupland 1991).

Research on the social psychology of communication has increased our understanding of the role of communication as social actions that influence both partners in the social interaction. Both communicators and audiences take each other into account in producing and comprehending the message, and in so doing change the meaning and significance of the message beyond its content *per se*. Both communicators and audiences use communicative acts to influence their communicative partner. These role-taking and social influence processes have

consequences for communicators' and audiences' attitudes, attributions, and memories, and thus communicative acts shape people's understanding of the world in which they live.

Because communication is a social action that shapes understanding, it concerns a critical motivation–cognition interface. Variables reflecting the motivation–cognition interface have received relatively little attention, however. For example, do 'saying is believing' effects depend on communicators' being motivated to share reality with their audience? Are the consequences of message production on the communicator different depending on their motives for producing it, such as to transmit information, to make a good impression, or to entertain the audience? The social psychology of communication can make further contributions to understanding language use by studying more explicitly the role of social motives in communication. All self-regulation involves trade-offs, and thus communication must involve trade-offs among different motives. How the problems of trade-offs are solved and how the solutions influence the nature and consequences of communication are major questions for future research.

See also: Communication: Philosophical Aspects; Nonverbal Communication, Social Psychology of; Verbal Communication, Psychology of

Bibliography

Cherry C 1957 *On Human Communication*. Technology Press of Massachusetts Institute of Technology, Cambridge, MA

Fleming J H 1994 Multiple-audience problems, tactical communication, and social interaction: a relational-regulation perspective. *Advances in Experimental Social Psychology* **26**: 215–92

Giles H, Coupland N 1991 *Language: Contexts and Consequences*. Open University Press, Oxford

Grice H P 1975 Logic and conversation. In: Cole P, Morgan J L (eds.) *Syntax and Semantics: Speech Acts*. Seminar Press, New York, Vol. 3, pp. 365–72

Higgins E T 1992 Achieving 'shared reality' in the communication game: a social action that creates meaning. *Journal of Language and Social Psychology* **11**: 107–31

Hilton D J 1995 The social context of reasoning: conversational inference and rational judgment. *Psychological Bulletin* **118**: 248–71

Krauss R M, Fussell S R 1996 Social psychological models of interpersonal communication. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology: Handbook of Basic Principles*. Guilford Press, New York, pp. 655–701

Manis M, Armstrong G W 1971 Contrast effects in verbal output. *Journal of Experimental Social Psychology* **7**: 381–88

Schwarz N 1994 Judgment in a social context: biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology* **26**: 123–62

Searle J R 1969 *Speech Acts: An Essay In the Philosophy of Language*. Cambridge University Press, Cambridge, UK

Semin G R, Fiedler K 1988 The cognitive functions of linguistic categories in describing persons: social cognition and language. *Journal of Personality and Social Psychology* **54**: 558–68

Shannon C E, Weaver W 1949 *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL

E. T. Higgins and G. R. Semin

Communication and Transportation, History of

The history of communication refers to a broad and loosely integrated field of historical research which includes parts of the history of transport, traffic, communication, media and information techniques. It is a relatively young field and a new point of view within historical science. Some aspects have been partly studied by historians—economic historians in particular—for a long time, especially the role of new technologies in transportation and traffic, and the history of the railways, ships and automobiles, and telephone and telegraph. Since the 1970s new questions and combinations have arisen with respect to the revolution of communication (including computation, multimedia, and the Internet) and these have given an impetus to broader research.

1. Overlapping Concepts

The term 'communication' was first used in England within the sectors of transport, commerce, and trade immediately before the Industrial Revolution. It was the time of extension of the road system with turnpike roads, new attempts at canal building, and improvement of the British postal system. The improvement of roads, the regulation governing the use of rivers, the erecting of canals, and improvements of coaches and cars, boats, river- and ocean-going ships, were aimed at more transport capacity, the reduction of high costs of transportation and the widening of local and regional markets to national and international dimensions. The Industrial Revolution from the 1820s onwards completed the possibilities of transportation with the construction of railroads and steamships and communication was improved by telegraph and telephone.

Today these forms of facilitating the services of commerce, trade, and tourism—including the later inventions of automobiles, progress in the building of roads, and the airplane—are described under infrastructure and its technical equipment. There is therefore an overlap between infrastructure and communication, and the last term has become more

and more subsumed under the name of 'media.' Communication nowadays is only partly concerned with matters of transport and traffic, and there is danger of a one-sided contraction of the content of this term, because the media and media devices are only partly subsumed under infrastructure.

Economists, especially those in public finance, normally define communication and transport as a necessary prerequisite for development and growth, calling it 'infrastructure,' (Hirschman (1958), Tinbergen (1962), Jochimsen (1962, 1966)). Before one can produce anything for a broader market, one needs transport and communication. The term 'infrastructure' is a word that originally came from French railway-idiom and was adopted by NATO after World War II. Later economists used it as a new term. Good 'material infrastructure' includes the communication possibilities of roads, canals, railways, ships, and, of course, all media equipment. But 'infrastructure' includes not only 'material infrastructure,' it also refers to 'immaterial infrastructure,' or learning, education, schools at different levels, universities, research institutes etc. The role of communication in the diffusion of knowledge and information is evident. According to Kuznets (1959, 1965, 1966) the infrastructure was one of the most important prerequisites for the European Industrial Revolution. It guaranteed stable economic growth, the international and regional division of labour, and made an impact on specialization and on more, and higher economic, social and cultural integration.

Today 'communication' is used especially in connection with the media. The development of the media since 1970, culminating in the Internet, necessitated a new classification of all items of media techniques and their former use: from the human voice to messengers, to ballad singers or actors in a theatre; from writing, handwritten codices and letters to printing in books, pamphlets, newspapers and posters or typewriting; from pictures, drawings and paintings to photos, and since the late 1990s scanning possibilities; from messages given through drums, written on stones, papyrus, parchment and paper to transmission by telephone, telegraph, film, radio and TV, CD-Rom, DVD, the computer and the Internet. As a result, the discipline of journalism has widened its scope to all the mass media and even studies their history.

There is also an overlap with the term 'information,' which is especially stressed by economics and business sciences as an important outcome of communication. Since the 1960s 'information' has been seen as the 'new immaterial resource' in physiology, sociology, psychology, pedagogy, linguistics and information science, ethnology, political sciences and law, and, last but not least, in a bundle of technical sciences.

In economics, for example, it was F. A. Hayek (1968) who described markets as spontaneous networks of information over goods and services. In a competitive market the prices make it possible to have and to use more information than normal, and of course much easier, faster, and cheaper. Market prices in normal (competitive) situations provide very precise information about shortages and abundances, qualities, monopoly and so on. Business scientists study the organization of firms and the optimization of information within the business entities and stress information as an important factor of production.

As Pool (1963) points out, communication is a fundamental necessity for human social action. Therefore we can distinguish between: (a) interpersonal communication = information exchange within an individual person; (b) intrapersonal communication = information exchange between two or more persons; and (c) media-bound communication = exchange of information between media producers and media consumers (in mass consumption). Since the 1980s communication has come to mean not only information between human beings but also between computers.

The development of communication and transportation possibilities show in the very long run a dramatic widening of the abilities and capacities of human communication. From the stages of primarily personal experience of communication by ears and eyes it shifted to a secondary system which is technically arranged and represents 'artificial' communication (S. Krämer (1997)). Today this 'artificial' communication has reached a peak in the new media possibilities of TV, the PC, and the Internet. McLuhan (1964) points out that individuals recognize their environment primarily through the media, which increasingly use the new techniques of electronically transmitting and storing information. His diagnosis is that 'the medium is the message.'

According to N. Luhmann (1981) social systems can be defined as the possibilities of human beings (in their time-space related contexts) to take part in the experiences, procedures, and activities of the society. Their ability to participate is, however, dependent of the availability of communication and transport equipment and techniques. Religious and political barriers or facilitations, problems of economic and social inequalities and, of course, the complex of good or bad 'property rights' play a role.

The historical sciences, i.e., political, social, and economic history, as well as the history of technology, study the causes and consequences of the communication process from antiquity to modern times. As communication and transport are important prerequisites of development and globalization, one has to examine very thoroughly the rise of the West to see how it obtained leadership over the East. To a large extent this leadership originates from better and new infrastructure and the organization of transportation and the media. Every step of European development since the Middle Ages can be seen in connection with the big spurts and crucial points in the improvement of each new media and in infrastructural equipment.

The European case also seems to show that, with the improvement of media and transportation, an upgrading displacement of the state and old elite went hand in hand with an increasing participation of human beings in communication. N. Elias (1939) in his famous book on civilization has pointed out the strict limitations of monopoly power in western societies. This freed the way for more and more social groups and individuals to use communication and transportation techniques. From the *cursus publicus* of the Roman Empire, which was only open for the purposes of the state to the new postal system at the beginning of Early Modern Times; from the cavaliers' tours of the nobles and the commercial journeys of single merchants to mass tourism, from illiteracy to dense school systems; from the handwritten codices of the monks to Gutenberg's Bible and Jerry Cotton; from the transport of gold, silver, silk, and amber to the transport of bulk goods over the whole globe; from the limited diffusion of knowledge and messages by fire signals, letters, etc., to a worldwide web in a 'global village' with 'global information infrastructure.'

1.1 Phases of Development in Europe

Historians and media specialists think that within western culture one can find four revolutions and periods in communication and transport development: (a) the Middle Ages (especially from eighth–fourteenth century AD); (b) the Early Modern Times (fifteenth–seventeenth century AD); (c) the onset of and then the Industrial Revolution (eighteenth–first half of the nineteenth century AD); and (d) the Communication Revolution, beginning in the last decades of the twentieth century.

M. Weber (1896) in his famous article upon the decay of the Roman world sees one main cause of its decline in the communication and transportation complex. He argues that all ancient high civilizations from China, Mesopotamia, Egypt and the Greek world were river- and shore/coastal-oriented civilizations. The failure of the Roman Empire was to go deeper and deeper into the European continent. Though the quality of Roman road building was very high and a relay system, changing tired for fresh horses, enabled people to cover distances as far as 300 km a day, the ancient *cursus publicus* had not enough impact on common trade, commerce, and communication, as it was not open to the common man.

The Middle Ages and the Early Modern Times in Europe found new answers and, in the long run, more stable solutions in communication, and prepared the way for the great spurt of European civilization which, from the sixteenth century onwards, left behind the older and richer civilizations of Asia. The Middle Ages from the tenth and eleventh century AD onwards developed a feudalistic system with a weak State and government, but strengthened local and regional authorities. At this time dense networks of more or less autonomous cities were created in which trade and handicraft was defined as a primary privilege of the citizens, and new, dynamic dimensions of good 'property rights' were introduced to the manors. New communication and information techniques and systems were tested. We can see this in regional and interregional trade with new types of merchants and new commercial institutions (such as fairs and marketplaces, trade guilds, commercial law, special organizational forms of storing, borrowing and changing money—all these were the outcome of perpetual communication). The consequences were a very rapid diffusion of technical knowledge in mining, architecture, handicraft, and agriculture.

This new system was also open for inventions from the Orient, the Arabs and the Islamic world, and accepted new insights from China and India. Technical inventions which revolutionized communication and transportation were the horse collar (tenth century, used in China since 450 AD) and the stirrup. From 1200 onwards the diffusion of the use of Arab numbers began and made trade much more easier. The compass and the compass needle (China 1075) at about 1200 improved navigation. Though the quality of the techniques of Roman road building could not be attained, the opening up of roads to everybody was a big improvement on the *cursus publicus* in the ancient Roman world. In about 1380 cities, religious orders and guilds introduced the new system of sending messages regularly and for long distances by private postal institutions.

Predominantly face-to-face communication using primarily oral means of information shifted more and more to a written-media stage. The monasteries and the first universities with already-formed libraries had a great impact on learning and knowledge. From the midst of the fourteenth century onwards the improvement of shipping on the rivers by constructing canals and canal locks began. Measuring time took on new dimensions with the invention of more precise clocks with wheels at about 1330. From the midst of the fourteenth century we can find clocks on the towers of churches. In 1510 the first pocket watches appeared.

The Early Modern Times brought about the second revolution of communication in Europe. The adoption of Chinese navigation techniques was completed; the compass freed the way for discoveries and the exploration of the New World. Trade in mass goods began. This became easier when plantations began to produce 'fungible goods' (like tea, coffee, cotton, and tobacco), which enabled a new trade on commodity markets.

The production and the use of paper and the idea of the movable type for printing (Johannes Gutenberg 1397–1468)—also of Chinese origin—brought about the shift from a written to a printed culture. Books now could be printed at a relatively low price and in

large numbers. The woodcut print (firstly used in 1400) and the copperplate print (firstly used in 1446) now could be combined with this new printing technique. Pamphlets were printed and could be easily distributed. Political and religious controversies, teaching, and learning took on new dimensions. Newspapers were introduced. The monopoly of science and information was broken up. The book became important for the unlimited storage and diffusion of information. The pamphlet was used for political discourse and opened up possibilities for participation in political and religious debates, and the newspapers opened the way to inform a growing number of readers about recent events in society.

The sovereigns of these territories reorganized their postal system and widened communication possibilities. For example in 1486 'Kurfürstliche Botenanstalt in Brandenburg.' In 1490 the postal system in the German Reich was reorganized under Maximilian I, and in 1516 the Vienna–Brussels postline was introduced, while the Brussels–Rome postline introduced in 1522 of the Thurn & Taxis Vienna-Nurenberg. The new communication advantages cleared the way for a European expansion over the Atlantic and created a second West in the New World of America.

The third revolution began before the Industrial Revolution and influenced the period up to the present day in three stages. In the first stage the mercantile system introduced more regular traffic, a regular postal system, and supported canal- and road-building. Milestones in France included the 'Canal du Centre' between the Loire and the Saône (1555) and the 'Canal du Midi' between Southern France and the Atlantic (1666–1681). Soon after, many canals in German territories, in the Dutch Lands and in England were built. The introduction of railways from the 1820s and 1830s onwards, the intensive use of steamships from the 1850s and 1860s onwards, and the common use of telegraph, telephone and the overseas cable since the 1850s–1870s marked the second step, which reached its peak in a highly integrated world economy before World War I.

The automobile and new road building, especially in the form of highways; and the victory of the street over the railway and the airplane widened communication and transportation possibilities following the interwar period. This brought greater reductions in the regional differences in prices, greater allocations of production and resources, more division of labour, mass production for wide national and foreign markets, standardization, and the unification of languages, morals, and tastes. The third step began in about 1970 with the 'postindustrial phase' or 'age of information.' The multimedia revolution brought the technical inventions of ISDN, the mobile phone (cellular phone), the fax, the PC and the Internet (www, on-line and e-mail), and digital forms of TV, broadcasting and printing. The economic consequences were e-commerce, direct banking, more foreign investment in many countries, and a very sharp decrease in the share of costs for information and transport, globalization, digital traffic engineering, and teleworking. The socio-cultural consequences were teleteaching, networked scientific community, and telemedicine.

1.2 Historiographical Trends, Further Research, and New Fields

The historiography of former times seldom studied communication and transportation. Since the turn of the nineteenth century economic history reached an independent status and transportation problems were described under the aspect of commerce. In the period before World War I in nearly all industrialized countries the first technical museums were erected. They had a great impact on the study of the history of technology and innovations, especially railways, cars, automobiles, ships, airplanes, and information devices like the telephone and telegraph. The history of technology became a new subdiscipline within economic history. In the last few decades the science of journalism has widened to a science of media and examines all media, especially mass media, in a historical context. Today a modern history of communication has to combine all these fields in cooperation with other social and economic disciplines.

There are, of course, many possibilities for further historical research in the field, e.g., studies into the regional, national, and worldwide diffusion of communication and transportation techniques; the social consequences of media revolutions; the impact of the media and the transport revolution on political situations and military conflicts; the access to communication and information techniques; religious and political control and inhibition of communication (censorship); the effects of the transport revolution within economic development; the role of official and nonofficial statistics in information about economies; the consequences of changing communication and transportation in mass tourism, leisure, and hobbies. Other possibilities include the rise of special cultural forms, traditions and customs concerning, for example, letter writing, traveling by coach, railroad, automobile, ship and airplane, and studies of the intensification of communication in personally written, and in digital communication.

Even completely new fields can be developed under the aegis of communication. One example may be the invention of the European 'financial intermediation' in banking and stock exchange. It can be seen as an important new way of communication in the economic and business sphere of the seventeenth and eighteenth centuries.

How to finance firms and enterprises and how to finance states and their budgets—these are the important questions in every modern economy. In former

times people who had large savings and lent to the state or to private entrepreneurs could not get back the money before the end of the agreed term. So, normally, the only possibility was to hold large stockpiles of money without being able to lend it in order to remain liquid at any one time. This problem was shared by wealthy individuals and institutions, such as insurance companies, managements of orphans' properties, etc. No civilization before had managed these problems adequately until Europe founded new ways of communication to bring together lenders and borrowers. There are four points that stress this information and communication revolution in banking since the seventeenth century.

(a) Since the period between the fifteenth and seventeenth centuries, communication and information has improved through investment in new roads, cars, ships, canals and new media techniques. The unification and standardization of goods also brought improvement, especially through the creation of 'fungible' goods: where one item stands in quality for all the other pieces of this product. Markets could thus be transformed partly into markets with 'fungible' goods. The exchange in products for certain metals, and for certain crops, especially tropical plantation crops like cotton, tea, coffee, sugar, etc., was born. This facilitated economic communication enormously. As the qualities of these goods were defined *ex ante* and were well known, merchants could deal in great volumes of these goods and articles without bringing them physically to the markets. Transaction costs were thus dramatically reduced.

(b) In combination with stock exchanges the experiences with 'fungible' goods were transferred in the seventeenth century to the financial sector. Creating fungible bonds for the state and public authorities (instead of individual obligations for single lenders in former times which could not easily be transferred to other capitalists) pulled hidden private stockpiles of wealth into the public economy and brought the money into the hands of sovereigns and governments. This idea spread to fungible shares for enterprises. Now it became possible to erect big enterprises in the form of joint stock companies. The first joint stock companies were firms within the trading sector, followed by companies in infrastructure and communication, in the banking system and last of all in big industrial firms.

(c) The next point in this development was the improvement of the financial mediation of banks. A new type of bank developed which brought together people who had money but did not need it at the time and people who had no or too little money and wanted to make productive investments in business. The 'goldsmith bankers' of London were the first to succeed at this business, because they found out that the communicative role of a bank has to be the transformation of risk, of terms, and of size. Fungibility and the new communicative role of banks became the main motor for a 'financial revolution' as precondition of the Industrial Revolution.

(d) Since the beginning of the twenty-first century new steps in financial communication have been made, especially in the context of the PC and internet systems. New generations of financial products have arisen, including e-banking and worldwide speculations on stock exchanges.

See also: Advertising and Advertisements; Agenda-setting; Ethnocentrism; Film: History; Geographic Learning in Children; Narrative, Sociology of; Political Advertising; Public Sphere and the Media; Tactile Maps in Geography; Television: Industry; Visual Images in the Media.

Bibliography

Aldcroft D H, Freeman M J (eds.) 1983 *Transport in the Industrial Revolution*. Manchester, UK

Bagwell P S 1974 *The Transport Revolution from 1770*. London Historical Society, London

Beninger J R 1986 *The Control Revolution: Technological and Economic Origins of the Information Society*. Harvard University Press, Cambridge, MA

Birk A 1934 *Die Straße. Ihre verkehrs- und bautechnische Entwicklung im Rahmen der Menschheitsgeschichte*. Karlsbad

Brown D 1997 *Cybertrends. Chaos Power and Accountability in the Information Age*. Penguin Harmondsworth, Middlesex, UK

Dyos H J, Aldcroft D H 1971 *British Transport. An Economic Survey from the 17th Century to the 20th*. Leicester

Eisenstein E L 1983 *The Printing Press as an Agent of Change*. Cambridge, Vol. 2

Goody J 1987 *The Interface Between the Written and the Oral*. Cambridge University Press, Cambridge, UK

Habermas J 1981 *Theorie des kommunikativen Handelns*, 2nd edn. Vol 2

Lay M C 1995 *Die Geschichte der Straße. Vom Trampelpfad zur Autobahn*

Lepetit B 1984 *Chemins de terre et voies d'eau. Réseaux de transports et organisation de l'espace en France 1740–1840*

Mattelart A 1996 *The Invention of Communication*. University of Minnesota Press, Minneapolis, MN

McLuhan M 1964 *Understanding Media The Extensions of Man* McGraw Hill, New York

McLuhan M, Powers B A 1989 *The Global Village. Transformations in World Life and Media in the 21st Century*. New York, Oxford

Pool I, de S 1963 The role of communication in the process of modernization and technological change. In: *Industrialization and Society*. UNESCO, pp. 279–295

Rösener W (ed.) 2000 *Kommunikation in der ländlichen Gesellschaft vom Mittelalter bis zur Moderne*. Veröffentlichungen des Max Planck Instituts für Geschichte 156, Vandenhoeck & Ruprecht, Göttingen

Sösemann B (ed.) 1994 Beiträge zur Kommunikationsgeschichte, Vol. 1

Teuteberg H-J 1994 Entwicklung Methoden und Aufgaben der Verkehrsgeschichte. *Jahrbuch für Wirtschaftsgeschichte I*

Teuteberg H-J, Neutsch C 1998 (eds.) Vom Flügeltelegraphen zum Internet. Geschichte der modernen Telekommunikation. *Vierteljahrsschrift für Sozial- und Wirtschaftsgeschichte* **147**

Treue W 1965 *Achse Rad und Wagen. 5000 Jahre Kultur- und Technikgeschichte*. Bruckmann, München, Germany

Williams F 1982 *The Communications Revolution*. Sage, Beverly Hills, CA

F. Baltzarek

Communication: Electronic Networks and Publications

Historically, the basis for scholarly communication has included a variety of publications, such as scholarly journals, newsletters, conference proceedings, monograph series, and books. Face-to-face communication took place via conferences, workshops, colloquia, and other events. Purely audio communication was primarily restricted to telephone communication and radio broadcast. Video and film captured image, speech, and event for live broadcast and delayed replay.

However, electronic networks have expanded means of communication and publication, first via CD-ROM, local area networks, and since the early 1990s and later (depending on network development in the world) via global networks including the Internet. The Internet protocol enables digital communication via email, FTP (File Transfer Protocol), chat rooms for simultaneous, interactive electronic communication, and video conferencing for educational and consulting purposes. Publication can be instantaneous with the completion of a document, image, or data set by using hypertext formats (HTM, HTML), PDF, PostScript, or even text (ASCII) formats coupled with placement in the World Wide Web. Hypertext references allow linking between cited and related items. Electronic publishing (e-publishing) includes both self-publishing and refereed publishing in various forms directly on the Internet or in an electronic journal with access via the World Wide Web. Publishers of refereed and nonrefereed scholarly journals are increasingly providing digitized versions parallel to the print journals, and new 'digitally borne' journals available only in electronic form (for instance *Journal of Electronic Publishing* (http://www.press.umich.edu/jep/), *First Monday* (http://www.firstmonday.org), *Psycholoquy* (http://www.princeton.edu/~harnad/psyc.html), and many more; NEWJOUR (http://gort.ucsd.edu/newjour/NewJourWel.html) has listed all new electronic journals since 1995; Charles Bailey also provides the *Scholarly Electronic Publishing Bibliography* with periodic updates at http://info.lib.uh.edu/sepb/sepb.html; see also *A Field Guide to Sources on, about and on the Internet: Electronic Publishing* at http://www.cc.emory.edu/WHSCL/electronic.publishing.html) are originating from scientists, learned societies, and university presses.

Electronic publication has permeated all fields of traditional publication formats: e-books, e-journals, e-zines, electronic proceedings, newsletters, electronic theses, and especially digitally prepared encyclopedias (see Pang 1998, Miller 1998) offer the added multidimensional aspects of integrated active hyperlinks between related concepts, articles, references, citations, graphs, and tables, as well as home pages of authors and institutions. This augments the linear reading and information-seeking methods of the past, provided that the objects 'pointed' or linked to are available in digital form. The hypertext format contributes to more efficient, timely, and economical revision especially evident in encyclopedias, multimedial works, and complete websites built on the concept of combining information building blocks or content units to produce webpages by 'spontaneous generation' at user demand.

Similarly, a single work such as an article can become a 'dynamic document' which allows continuous additions, alterations, and comments also in the form of an attached discussion list. This procedure is changing the nature of the publishing paradigm both in the sense of a fixed 'date of publication' and in the finalized or 'closed' state of the information content in this work (examples of dynamic documents include the experimental journal *Living Reviews in Relativity* (http://www.livingreviews.org)).

The basis for digital connections are the electronic networks established within the framework of the Internet protocol (TCP/IP) and expanded to its present state in the World Wide Web. Electronic networks began in the 1950s and were developed by the Defense Advanced Research Projects Agency (DARPA) particularly for military defense purposes well into the 1980s, with some academic institutions in the USA included by merit of their involvement in defense research. The 'Internet' was then opened to include higher-education institutions in the USA at the end of the 1980s, as well as other US research libraries and schools. As of 1991, other institutions could also use the Internet, including non-US educational institutions and commercial enterprises.

This created new possibilities for global communication, scientific exchange, and building of international collaboratories, as well as enabling rapid dissemination of scientific results. The business world also recognized the potential of the Internet for e-commerce. Booksellers, publishers of all sorts, corporations providing educational programs, as well as other entrepreneurs, have developed business strategies which integrate the electronic communication and hypertext publication possibilities of the Internet. As a result of this development, certain aspects of free

access to information, as supported by libraries worldwide, have been commercialized, and new questions of licensing and ownership of intellectual property have been raised. Issues of electronic piracy, copyright, and protection of intellectual property have taken on new dimensions and demand new techniques to assure quality control, validity, authenticity, and ownership rights.

Similarly, the economics of scholarly publishing, sometimes referred to as a 'crisis in scholarly publishing,' have promoted the transition to electronic publication. In particular, the US Coalition for Networked Information (CNI) (http://www.cni.org), established in 1990 as a result of the opening of the Internet to general academic institutions, has dedicated itself since its inception to the analysis of the system of scholarly communication and its transformation through the effects of networked information transmission. CNI-initiated cooperation between university libraries and commercial publishers resulted in the first electronic journals project, TULIP (**T**he **U**niversity **Li**censing **P**rogram), in which pricing models and availability of electronic journals were tested (see Lynch 1995).

The nature and scope of the Internet as a distributed information network challenges traditional tasks associated with libraries, free access to information, and also the author–publisher relationship. Issues of 'access versus ownership' of scholarly articles through restrictions of licensing agreements could make publishers' servers the sole repositories of scholarly works. This also evokes a vehement controversy between certain segments of the academic community over 'who owns scholarly publishing' (Bacharach et al. 1998), Odlyzko (1995, 2001), Harnad (1995a), and others support the concept that the author and the author's institution, having supported the research and researcher, 'own' the information and have the right to archive it on an institutional e-print server, as well as to offer it cost-free to the scientific world via the institutional server, the author's homepage, or similar free-access mechanisms.

Thus, the original concept of 'open' access was synonymous with 'cost-free' access to information which is necessarily changing to access often with a price. Most publishers of electronic journals do not allow copies of the electronic files to be given to non-licensed users in the way that paper copies of non-subscribed journal articles were made available via interlibrary loan. Thus, certain aspects of scientific communication can no longer function without corresponding licensing agreements at the user's institution or immediate payment (pay-per-view) via credit card.

In addition, scholars were often frustrated by long delays from the time of submission to time of publication. The ensuing conflict of interests within the scientific community has given way to the extensive participation in preprint servers and to self-publishing or 'self-archiving' (see Harnad 1995b). Originally based on FTP-Archives (Gardner 1990) for scientific articles, the first preprint server, the so-called 'Ginsparg-Server,' originally developed at the Los Alamos National Laboratory in 1991 by Paul Ginsparg to allow researchers in high-energy physics immediate access to scientific articles without the delay of peer-review and publication schedules. Since then, other subject-oriented eprint servers patterned after this server have evolved (NCSTRL, REpec, CogPrints, PhysNet, MathNet). In 1999, representatives from these servers plus representatives from research libraries and digital library initiatives in the USA met to discuss a cooperative project to create a retrieval mechanism for all eprint servers. Originally called the Universal Preprint Service (UPS), this evolved into the Open Archives Initiative (OAI) (http://www.openarchives.org) which has determined a standard set of metadata and a harvesting protocol for search and retrieval of the metadata from all participating eprint servers, including not only self-archived articles, but also electronic theses (see also the National Digital Library for Electronic Dissertations and Theses (http.//www.ndltd.org)) and other formats of electronic scholarly information.

Another response to the development in scholarly publishing is the increased participation of university libraries and learned societies in alternative initiatives such as SPARC (Scholarly Publishing and Academic Resources Coalition (http://www.arl.org/SPARC/); see Johnson 1998, Rosenzweig 1999). Over 175 (at the beginning of 2000) research institutions, college and research libraries, consortia, and library associations from North America, Europe, Asia, and Australia in SPARC pool their resources to produce high-quality scholarly journals on a purely electronic basis. The intellectually high quality, but significantly cost-effective, scholarly journals created by this group challenge overpriced commercial competitor titles while preserving the academic tradition of quality through peer review.

Electronic publications have the potential of citation-linking from one electronic work to the full text referenced. This, however, has implications connected not only with standardization of technological markup, but also with cost and access to the full texts of the referenced works. 'Click-through,' or direct, access from a hyperlinked reference or concept in one electronic work in a publisher's or institutional server to another server where the targeted item is stored may be inhibited by licensing regulations or be subject to conditions negotiated (or not negotiated) by the responsible institution.

As a results of such restrictions for education and scientific research, the US Government supported development in 1999 of several subject-oriented e-print servers, PubMed, PubScience, and PubMedCentral (http://www.nih.gov/about/director/pubmedcentral/pubmedcentral.htm). These develop-

ments, in addition to those evolving out of the scientific communities, challenge prevailing publishers' policies for licensing scientific information while also providing mechanisms for free access to article information on a subject-based, as well as on a global and interdisciplinary, level.

To counteract this, about 25 publishers world-wide collaborated to develop the 'Cross Ref' service (http://www.crossref.org) which would allow reference-linking and click-through licensing agreements. This publishers' consortium has determined a standard set of bibliographic, content, and naming values (Digital Object Identifier, metadata), a common navigation tool (search engine), and a pricing scheme to facilitate direct access via references to full texts in other publishers' servers based on composite licensing agreements with all participating publishers or on a payment schemes ('pay-per-view,' 'pay-per-print') at point of use.

A major research study on the changing paradigms of scholarly publishing, the Open Journal Project, began in 1996 under the eLib Programme in Great Britain and has been expanded as of 2000 under the joint JISC–NSF International Digital Libraries Research Programme in the 'Open Citation Project' (http://opcit.eprints.org/). The 'Open Citation Project' will establish electronic reference links from 100,000 papers in the Los Alamos arXiv e-print server (http://arXiv.org) to the corresponding electronic full texts on other sites (publishers' servers as well as other noncommercial sites). In conjunction with the Computer Research Repository (CoRR), a considerable number of nonelectronic resources in the cited physics, mathematics, and computer science literature will be integrated, also demonstrating procedural aspects of connecting electronic and nonelectronic documents via citation-linking as an enhancement of the information value of electronic resources (see Hitchcock et al. 1998, 2000).

Electronic publication formats also have the potential for long-term archiving of scholarly publications combined with global access. In 1995, the Johns Hopkins University Press (JHUP) cooperated with the Milton S. Eisenhower Library with funding from the Andrew W. Mellon Foundation and the National Endowment for the Humanities (NEH) to digitize full-text scholarly journals from JHUP in the Project MUSE (http://muse.jhu.edu/muse.html). By 1999, 46 peer-reviewed JHUP journals in the humanities and the social sciences were available via a searchable database. This database will continue to be expanded to include all JHUP journals.

The JSTOR project (**J**ournal **Stor**age—The Scholarly Journal Archive), in which back volumes of journals in philosophy and the humanities have been digitized to facilitate electronic access and navigation, has developed a new, collaborative payment model (see Guthrie 1997a, 1997b). Member libraries pay an initial fee based on the size of the parent institution and number of students and faculty, as well as an annual fee for individual use. Thus, the consortium of library customers supplies the funds which allow continued maintenance of existing digitized journals, as well as seed-funding for systematic scanning, processing, and integration of further complete sets of other selected journals.

Standardized archiving procedures for scholarly publications in electronic form have not yet been achieved. The archival function of digitizing print volumes at key subject-oriented institutions can provide the respective scientific community with an electronic repository of materials which the publishers would not find financially feasible to digitize. Further use of digitization and development of standardized metadata for digital preservation are incorporated in the CEDARS Project in Great Britain (http://www.leeds.ac.uk/cedars/) which addresses strategic, methodological, and practical issues of digital preservation of information resources. Various projects in Australia, the USA, Germany, and other countries are also working on these issues.

The quality of content management and information provision through electronic publishing in any format is dependent on a variety of new concepts: standardized metadata, licensing terms and conditions governing access, citation-linking, subject-specific search engines and archival servers, intellectual design, and new forms of visualization of scientific knowledge. Metadata initiatives (for instance, the Text Encoding Initiative (TEI) (http://www-tei.uic.edu/orgs/tei/), the Government Information Locator Service (GILS) (http://www.gils.net/), the Encoding Archival Description (EAD) (http://lcweb.loc.gov/ead/), and the Dublin Core Metadata Initiative (DCMI) (http://lcweb.loc.gov/ead/) are some of the most notable; see also the IFLA Electronic Resources Bibliography for 'Digital libraries: Metadata resources' (http://www.ifla.org/II/metadata.htm)) parallel traditional library cataloging categories while accommodating new meta-information needs for resource discovery, search and retrieval, access and availability schemes for digital and non-digital items. Various projects, such as the Electronic Text Center of the University of Virginia or the Gutenberg Project (a world-wide collaboratory for digitizing literary classics and making them available in electronic format), were instrumental in promoting the transition from print to electronic texts and creating text-mark-up metadata schemes. Other metadata schemes implement structures for payment and access models (terms and conditions), and facilitate processing of remuneration and payment channeling of artist's honoraries or performance fees to the appropriate recipients, for instance in the fields of music and art (e.g., the Interoperability of Data in e-Commerce Systems).

Studies on the acceptance and use of electronic publications show that the natural and biomedical science communities demonstrate greater acceptance

and frequency of use of electronic-information resources (see The International Coalition of Library Consortia 1998, Rusch-Feja and Siebeky 1999). The social and behavioral sciences are less accepting of electronic resources due to often substantially longer articles and later integration of computers in these fields. There are also cognitive and behavioral reasons relating to qualitative assessment of use and acceptance of electronic information resources. Further research on digital presentation of scientific results, visualization techniques, and design aspects to heighten readability of electronic resources and render ease of interpretation of complex data sets, especially in the natural sciences, can be expected in the coming years. New developments in information retrieval and search precision, automated collection of metadata, and personalized information brokers in complex information portals will add value to electronic publications which surpasses that of print equivalents and will create new dimensions for electronic publishing and scholarly information in a networked environment.

See also: Computer Networking for Education; Computers and Society; Information Society; Publishing as Medium

Bibliography

Bacharach S, Berry R S, Blume M, Foerster T von, Fowler A, Ginsparg P, Heller S, Kestner N, Odlyzko A, Okerson A, Wigington R, Moffat A 1998 Who should own scientific papers? *Science* **281** (4 September): 1459–60

Bennett S 1998 Just-in-time scholarly monographs. *Journal of Electronic Publishing* **4** (1): *http://www.press.umich.edu/jep/04-01/bennett.html*

Electronic Literature Directory 1999–2000 (*http://directory.eliterature.org/*)

Electronic Literature Organization 2001 (http://www.eliterature.org/index2.html)

Fishwick F, Edwards L, Blagden J Economic implications of different models of publishing scholarly electronic journals for professional societies and other small or specialist publishers: Report to the Joint Information Systems Committee. Electronic Libraries Programme: *http://www.ukoln.ac.uk/services/elib/papers/tavistock/scholarly-journals/crane-lib.html*

Gardner W 1990 The electronic archive: Scientific publishing for the 1990s. *Psychological Science* **1**: 333–41

Ginsparg P 1994 First steps toward electronic research communication. *Computers in Physics* **8**: 390–6 (*http://xxx.lanl.gov/blurb/*)

Guthrie K M 1997a JSTOR: The development of a cost-driven, value-based pricing model. Scholarly Communication and Technology Conference, Emory University: *http://www.arl.org/scomm/ scat/guthries.html*

Guthrie K M 1997b JSTOR: From project to independent organization. *D-Lib Magazine* July/August: *http://www.dlib.org/dlib/july97/07guthrie.html*

Harnad S 1990 Scholarly skywriting and the prepublication continuum of scientific inquiry. *Psychological Science* **1**: 342–3 (also in *Current Contents* **45** (1991): 9–13)

Harnad S 1991 Post Gutenberg galaxy: The fourth revolution in the means of production of knowledge. *Public-access Computer Systems Review* **2**: 39–53 (also in *PACS Annual Review* **2** (1992)) and http://cogsci.soton.ac.uk/~harnad/papers/Harnad/harnad91.postgutenberg.html

Harnad S 1995a Electronic scholarly publication: Quo vadis? *Serials Review* **21**: 70–2 (also in *Managing Information* **2** (1995) and *http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad95.quo.vadis.html*)

Harnad S 1995b Publicly retrievable FTP archives for esoteric science and scholarship. A subversive proposal. In: Okerson A, O'Donnell J J (eds.) *Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing. An Internet Discussion about Scientific and Scholarly Journals and their Future*. Association of Research Libraries, Washington, DC

Harnad S 1999 The future of scholarly skywriting. In: Scammell A (ed.) *'i in the Sky: Visions of the Information Future'*. Aslib, London

Hitchcock S, Carr L, Hall W, Harris S 1998 Linking electronic journals: Lessons from the Open Journal Project. *D-Lib Magazine* (December) *http://www.dlib.org/december98/12hitchcock.html*

Hitchcock S, Carr L, Jiao Z, Bergmark D, Hall W, Lagoze C, Harnad S 2000 Developing services for open eprint archives: Globalisation, integration and the impact of links. *ACM 2000: Digital Libraries. Proceedings of the 5th ACM Conference on Digital Libraries, June 2–7, 2000*. ACM, San Antonio, TX, pp. 143–52

IFLA Electronic Resources: Digital libraries: Electronic journals and text archives (*http://www.ifla.org/II/etext.htm*)

IFLA Electronic Resources Bibliography: Digital Libraries: Metadata Resources (*http://www.ifla.org/II/metadata.htm*)

The International Coalition of Library Consortia (ICOLC) 1998 *Guidelines for statistical measures of usage of web-based indexed, abstracted, and full text resources*: *http//www.library.yale.edu/consortia/webstatspr.html*

The Interoperability of Data in e-Commerce Systems (INDECS) (*http://www.library.yale.edu/consortia/webstatspr.html*)

Johnson R K 1997 SPARC White Paper: *http://www.arl.org/sparc/resources/whitepaper.pdf*

Lynch C A 1995 The TULIP Project: Context, history, and perspective. *Library Hi Tech* **13**: 8–24

Miller S K 1998 Transforming print encyclopedias into successful electronic reference products. *Journal of Electronic Publishing* **3**: *http://www.press.umich.edu/jep/03-04/miller.html*

Odlyzko A M 1995 Tragic loss or good riddance? The impending demise of traditional scholarly journals. *International Journal of Human–Computer Studies* **42**: 71–122 *http://www.research.att.com/aemo/doc/complete.html*

Odlyzko A 2001 *The Rapid Evolution of Scholarly Communication*: *http://www.dtc.umn.edu/~odlyzko/doc/rapid.evolution.pdf*

Odlyzko A 1999 Competition and cooperation: Libraries and publishers in the transition to electronic scholarly journals. *Journal of Electronic Publishing* **4**: *http://www.press.umich.edu/jep/04-04/odlyzko0404.html* (also in Berry S, Moffat A (eds.) in press *The Transition from Paper: A Vision of Scientific Communication in 2020*. Springer, New York)

Pang A S-K 1998 The work of the encyclopedia in the age of electronic reproduction. *First Monday* **3**(9): *http://www.firstmonday.org/issues/issue3-9/pang/index.html*

Peek R P, Newby G B (eds.) 1996 *Scholarly Publishing: The Electronic Frontier*. MIT Press, Cambridge, MA

Peek R P, Pomerantz J P 1999 Electronic scholarly journal publication. In: Williams M E (ed.) *Annual Review of Information Science and Technology*, Vol. 33. Information Today, Medford, NJ, pp. 321–56

Rosenzweig M L 1999 Reclaiming what we own: Expanding competition in scholarly publishing. Plenary address to the Ninth Conference of the Association of College and Research Libraries, Detroit, Michigan, April 11: *http://www.arl.org/SPARC/rosenzweig.html*

Rusch-Feja D, Siebeky U 1999 Evaluation of usage and acceptance of electronic journals: Results of an electronic survey of Max Planck Society researchers including usage statistics from Elsevier, Springer and Academic Press. *D-Lib Magazine* October (full report: *http://www.dlib.org/dlib/october99/rusch-feja/10rusch-feja-fullreport.html*)

Wittenberg K 1998 CIAO: A new model for scholarly publishing. *Journal of Electronic Publishing* **3**

D. D. Rusch-Feja

Communication: Geographic Aspects

1. *Defining the Field*

Communications geography, which includes mass communications and telecommunications, considers the movements or flows, directions, and volumes of information between and among persons and areas. The communication may be within households and neighborhoods as well as firms, organizations, and states. While elements of information transfer have always been inferred components in human geography, for example, in the diffusion of religion or the diasporas spreading colonialism, it was during the 1960s and 1970s that the spatial components of communication began to receive serious attention. These decades, at least in the wealthy and industrialized worlds, experienced rapid urbanization and growth in tertiary sector employment and growth. They were associated with a postindustrial economy and society. The key employment categories in the services economy were retail, wholesale, and international trade, transportation, government, finance, health care, and education. Also leisure activities, telecommunications, printing and publishing, education, and the media experienced growth. For those social scientists studying these services, a refinement was needed; the term quaternary sector identified those in information and communications industries and services. Sociologist Daniel Bell in his 1973 *The Coming of the Post-Industrial Society* used this term in describing the economic/social transitions that were occurring in American society.

2. *An Emerging Field of Inquiry*

Those early geographers who developed this specialization came from backgrounds in economic, social, and urban geography. Some were interested in transportation networks, others in urban social behavior, and still others on the social impacts of Information and Communications Technology (ICT). During this same time geography was experiencing disciplinary shifts, including the emergence of social-behavioral geography, which utilized some concepts, models, and theories from psychology and sociology, and then humanistic geography, with close ties to history, literature, philosophy, and the arts. Human geographers in both spheres were interested in the ways people communicated, media producers, the diffusion of information, power issues associated with information production, how places were presented, and the ways information geographies might be mapped.

Some of the initial work by behavioral geographers applied the Shannon–Weaver model to geographical contexts. This model dealt with the flow of electronic information from its source to its destination. That flow could be person-to-person (face-to-face) or person-to-area (television reporter transmitting news to national or global audiences) or area-to-area (from a government to major trading partners). Many channels as well as barriers or filters could influence the transmission of information from its source to its final destination(s). The term gatekeeper was introduced and could be applied to a newspaper editor, owner of television stations, or government agency censoring reports to its population. The spatial dimensions of information flows, both interpersonal and corporate, were studied, as were the impacts of mass media. Geographers also mapped information and communication structures, including the newspaper sheds (circulation areas) and the media sheds for radio and television stations.

A major theoretical line of inquiry during the 1960s was the 'diffusion of innovations'; it was introduced to European and North American geographers by Swedish geographer Torsten Hägerstrand (1967) who used the concept to describe and predict the adoption rates of agricultural innovations. Rural sociologist Everett Rogers (1962), with his work on the social dimensions of diffusion, also influenced geographers' thinking. Diffusion research identified individuals as being innovators, early and late adopters, or laggards. Geographers examined the diffusion of crops, diseases, institutions, revolutions, and consumer purchases. Maps at varying scales depicted adoption rates. Some studies described patterns; others used sophisticated mathematical models to predict adoption rates.

One of the first publications in communications geography was *Human Geography in a Shrinking World* (Abler et al. 1975), which included discussions on space-adjusting technologies, tyrannies and human freedom, monocultures and minicultures in the media,

bureaucracy and alienation, and futuristic scenarios. Several authors discussed how these technological changes affected households, firms, and organizations. There was also interest in how to display 'shrinking world' concepts cartographically. Three additional concepts that formed the basis of discussion were time–space convergence, cost–space convergence, and human extensibility; the first referred to the rate places were 'moving closer together' in relative time, for example, with faster transnational phone calling capabilities, and the second, how lower costs reduced the friction of distance in placing calls or mailing items from the US east coast to the west coast. Over time the costs to communicate from place to place became more uniform, that is, distance ceased to be a barrier to communication; standardized postal rates in a country represent a good example of complete cost–space convergence. Human extensibility relates to the networking features of an individual, institution, university, government, or nongovernmental organization (NGO).

Communications geography continued its ascendancy during the 1970s and 1980s with contributions coming from geographers with backgrounds in the social sciences and humanities and with ties to rural and urban sociology, psychology, computer science, but also telecommunications, film studies, literature, architecture, the social impacts of technologies, and future studies. The role of mass media in promoting popular culture, news reporting, the place settings and images in film and television documentaries, state boosterism, travelers' writings, and landscape representation in the popular literature were common themes (Burgess and Gold 1985, Zonn 1990). The experiences of authors, painters, photographers, lyricists, and film directors were studied to communicate to readers the meanings of places and landscapes.

Cartography was also changing from a technique focusing on how best to display and represent data, whether pen-and-ink methods or automated mapping, to the construction and interpretation of maps, that is, their power, meaning, identity, and interpretation. Maps became more than displays of information, but products in which information is communicated with social meanings and where the author/cartographer (whether a scholar, company, or state) conveys different messages to readers. Context was important. The colors, languages, symbols, and projections are all elements of visual 'communication.' Maps were also designed for the visually and color impaired. Visualization became a concept associated with the presentation and representation of information, whether maps or graphs. Maps were but one subject of texts and discourses studied by social theorists in geography; these poststructuralists also investigated the constructions of wilderness paintings, landscape poetry, city boosterism, and foreign policy, all examples where power was important in communication (Barnes and Duncan 1992). Social theorists in geography benefited from the writings of linguists, philosophers, intellectual historians, and landscape architects.

Paralleling social theory contributions in communications geography in the early 1990s were case studies on specific countries, firms, communications networks, and the social impacts of ICT. A volume that highlighted these developments was *Collapsing Space and Time: An Introduction to the Geography of Information and Communication* (Brunn and Leinbach 1991), which contained chapters on telecommunications and regional development, ICT and religious broadcasting, banking, medical services, and distance learning, and the notion of the 'electronic cottage' or homeworking. What these authors and others writing in the early 1990s failed to grasp was the blossoming of GIS (Geographical Information Systems) and the rapid diffusion of the Internet and WWW.

3. *Current Research*

Speed, volumes, networking, and accessibility are the keys to understanding the geographies of the Internet. The net is a medium to send, receive, analyze, store, disseminate, manipulate, and map large amounts of print, sound, and visual data to, from, and about households, firms, organizations, institutions, and states. Developments are occurring in banking, real estate, education, health care, entertainment, and tourism, all of which are information based. The traditional information and communication services such as journalism, broadcasting, and telecommunications are also affected by these high tech developments. While the Internet and WWW have diffused to many parts of North America and Europe, large areas of the developing world remain unwired. An entirely new international lexicon has developed, including virtual community, cyberspace, cyberplace, chat-rooms, all terms relating to communication, place, and identity. Children and adults are learning these terms and how to use the technologies at the same time, sometimes elders learning from youth. Most materials (print or on-line) published about the Internet are in English as Internet communication is dominated by American English. Non-English users are almost forced to learn English if they want to communicate and interact in electronic space. Geographers keep apprised of developments in ICT not only by attending conferences and workshops, but also through GEOGCOMM, a listserv operated from the University of Kentucky, the IGU (International Geographical Union) Commission on the Geography of Information Society, and the journal *Netcom* (Networks and Communications Studies).

Much of the recent research by geographers addresses issues about cyberspace, electronic commerce, telecommunications, and social impacts. For example,

some geographers are investigating the geographies and geometries of virtual places, electronic scholarly communities, the virtual geographies of the military, children's games, how states are using the WWW to promote themselves to potential investors and tourists, and the uses of GIS by emerging democracies, NGOs, revolutionary and hate groups (Pickles 1995, *Geographical Review* 1997). Others are examining the meanings of electronic space, human rights in an electronic state, telecommunications and governance, global 24-hour banking, and ICTs in peripheral rural areas to promote grassroots economic development (Wilson and Corey 2000). Urban geographers are investigating how telecommunications are transforming urban spaces and systems, high tech industrial location in an information age, the diffusion of innovations, teleworking, gender and work, and the role of the state (Graham and Marvin 1996, Wheeler et al. 2000). Mapping communications in an electronic age and the meanings of accessibility in the information age are additional topics (Janelle and Hodge 2000, Dodge and Kitchen 2001). E-commerce has been studied with regard to financial services, recruiting, regulatory problems, and diffusion in developing countries (Leinbach and Brunn 2001).

The electronic worlds of the fax, email, chat-room, and WWW are changing the strategies for industrial location, how governments define and regulate public goods, how consumers purchase goods and services, and how individuals behave. High-speed lines move money (itself information) in split-second time around the world 24 hours a day. Time zones lose their meaning. Students, scholars, community organizers, and governments can request information and have it often the same hour or day. Companies, cities, and states can use the WWW to promote themselves almost anywhere (where there are connections) and to anyone (who has an operating machine). 'E' services, now existing in banking, health, entertainment, tourism, education, and religion, keep carving out niches.

4. Dilemmas and Future Directions

While advances in high tech communications are diffusing rapidly throughout the developed worlds, and rich nodes in the developing worlds, geographers and others remind us that low tech communications and geographies and 'digital divides' remain much in evidence for most of the world's population. Some of these recent innovations have not diffused globally, nor will they, as they may not be desired or welcomed by some societies and individuals. Some states attempt to block these ICT, which is difficult; others cannot afford to construct fiber optic networks or satellite receiving stations, and still others seek to construct filters to screen out undesirable information for their citizens, children, or politically dissident groups. What is emerging alongside rapid advances in communications technologies (toward more wireless communities and miniaturized devices) is a widening gap between the 'have and have-nots' within cities, countries, and the world. The long-term social, political, and spatial consequences of these 'ICT gaps' remain largely unknown, but they are receiving the attention of geographers and others.

A number of troublesome philosophical and methodological issues surface when conducting research on these topics. They include the availability, reliability, and cost of data from public and private sources, privacy issues when individual household or firm data are sought, and the appropriateness of Western models when constructing theories and describing or explaining results. One also needs to keep in mind the role of the state in gathering the data and publishing them in tabular, digital, or cartographic form. Transnational corporations are also major producers and consumers of information about individuals, groups, societies, and physical environments. Medical, employment, income, and family history data can always be manipulated for evil purposes by rouge states, organizations, or corporations.

Among the questions being raised about the human dimensions of technological change are not only what ICTs tell us about human life and society, but also children and adults, girls and boys, women and men, and those with learning and physical disabilities and impairments (color, visual, dexterity, hearing). These are populations and groups who exist in all societies that merit continued study by those with interests in cartography, visualization, GIS, social theory, E-commerce, geopolitics, grassroots democracy, and teaching second and third languages. We also would benefit from investigating what is happening to humans as we communicate in and outside of cyberspace. What happens to our identity as we become geolimnal selves? And what does this mean? Recent research in communications and information geography is beginning to addresses these practical and philosophical questions.

See also: Cartography; Communication and Democracy; Communication: Electronic Networks and Publications; Diffusion: Geographical Aspects; Evolution: Diffusion of Innovations; Information Society; Information Society, Geography of; Information Technology; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Political Communication; Space and Social Theory in Geography; Time–Space in Geography

Bibliography

Abler R, Janelle D, Philbrick P, Sommer J 1975 *Human Geography in a Shrinking World*. Duxbury Press, North Scituate, MA

Barnes T J, Duncan J S (eds.) 1992 *Writing Worlds: Discourse, Text & Metaphor in the Representation of Landscape*. Routledge, London
Bell D 1973 *The Coming Post-Industrial Society*. Basic Books, New York
Brunn S D, Leinbach T R (eds.) 1991 *Collapsing Space and Time: Geographic Aspects of Information and Communication*. HarperCollins Academic, New York
Burgess J, Gold J R (eds.) 1985 *Geography, The Media & Popular Culture*. St. Martin's, New York
Dodge M, Kitchen R 2001 *Mapping Cyberspace*. Routledge, London
1997 Introduction: Cyberspace and geographical space. *Geographical Review* **87**(2)
Graham S, Marvin S 1996 *Telecommunications and the City: Electronic Spaces, Urban Places*. Routledge, London
Hägerstrand T 1967 *Innovation Diffusion as a Spatial Process* [trans. from Swedish by Pred A, Haag A]. University of Chicago Press, Chicago
Janelle D G, Hodge D C (eds.) 2000 *Information, Place and Cyberspace: Issues of Accessibility*. Springer-Verlag, Berlin
Leinbach T R, Brunn S D (eds.) 2001 *Worlds of E-Commerce: Economic, Geographical and Social Dimensions*. Wiley, New York
Pickles J (ed.) 1995 *Ground Truth. The Social Implications of Geographic Information Systems*. Guilford Press, New York
Rogers E M 1962 *Diffusion of Innovations*. Free Press, New York
Wheeler J O, Aoyama Y, Warf B (eds.) 2000 *Cities in the Telecommunications: The Fracturing of Geographies*. Routledge, New York
Wilson M I, Corey K E (eds.) 2000 *Information Tectonics: Space, Place, and Technology in an Electronic Age*. Wiley, New York
Zonn L (ed.) 1990 *Place Images in Media: Portrayal, Experience and Meaning*. Rowman and Littlefield, Savage, MD

S. D. Brunn

Communication: Philosophical Aspects

S produces the sounds 'It's snowing' in the presence of *A*, and *A* instantaneously comes to know that it is snowing. *S* has *communicated to*, or *told*, *A* that it is snowing, and, as a result of *S*'s speech act, *A* came to know that it was snowing. Philosophical interest in communication turns on four interrelated questions. The first is about the logical structure of communication or, more specifically, about whether communication is a relation that holds among three things just in case the first communicated the second to the third. The second is about the explication, or analysis, of communication. The third is about the role of communication in the explication of linguistic meaning. The fourth is about how knowledge is acquired by communication. The questions are interrelated in that answers to any one question may constrain answers to one or more of the others.

1. *The Logical Structure of Communication*

The issue about the logical structure of communication can be explained in the following way. The verb 'to love' stands for a two-place relation, a relation that holds between two things just in case the first loves the second (temporal considerations are ignored for simplicity of exposition). This is indicated by the fact that if we take a sentence like

(a) Al loves Betty

and replace its singular referring expressions with variables, we obtain (what logicians call) the open sentence

(b) *x* loves *y*.

This indicates that the verb 'to love' stands for a two-place relation, a relation that holds between two things just in case the first loves the second. Similarly, we can say that giving is a three-place relation that holds among three things just in case the first gives the second to the third, because from a sentence such as

(c) Al gave that book to Betty

we obtain the three-place open sentence

(d) *x* gave *y* to *z*.

The issue concerning communication is whether we can likewise start with a sentence such as

(e) Sally communicated that it was snowing to Abe

and obtain the open sentence

(f) *x* communicated *y* to *z*.

If we can, then a further issue is about the sort of things that can occupy the *y* position. That is, what sort of *things* are the things we communicate, if communicating is represented by (f)? There is a philosophical issue about the logical structure of communication because there is disagreement about how to answer both questions.

The *prima facie* case for taking communication to be a three-place relation is simple: it is the obvious way to account for the fact that (e) entails

(g) There is something that Sally communicated to Abe—to wit, that it was snowing.

A possible reason to resist this first thought emerges when we ask what manner of *thing* is this thing, *that it was snowing*, which would be the entity Sally communicated to Abe if communication is a relation expressed by (f). For it is apt to appear that things communicated are what philosophers nowadays call *propositions*—mind- and language-independent entities that have truth conditions but are not in space and have no physical properties whatever. The trouble is that many think there are good reasons not to believe that such things exist (e.g., Quine 1970).

It must, however, quickly be said that the immediate source of concern here is not with communication *per se* but with *that-clauses* in sentences such as (e) and 'Sally believes that it is snowing'—that is, with expressions of the form 'that *S*,' where '*S*' holds the place of a sentence in the indicative mood. More generally, the concern is with what (for historical

reasons we need not go into) philosophers call the *problem of intentionality*. Do that-clauses refer, or purport to refer, to *entities* that are believed and communicated, or do they have some other semantic function? While this is a problem that can be stated without specific reference to communication, a complete account of communication presupposes its resolution. Until we understand the semantics of that-clauses, we will not know what communication is. But there is no universally accepted solution to the problem of intentionality, and for every position in logical space there are philosophers who occupy it. For about one-third of the twentieth century, the American philosopher Willard von Orman Quine had so effectively stigmatized propositions that most philosophers were too frightened to believe in them; but philosophy has recovered from Quine's influence, and most philosophers today feel that the case for propositions is stronger than any against them. At the same time, there is disagreement among propositionalists about the kind of propositions we believe and communicate. Here one finds two dominant views. The *Russellian view* (after the British philosopher and logician Bertrand Russell) holds that propositions are made up of the objects and properties our beliefs are about, while the *Fregean view* (after the German philosopher and logician Gottlob Frege) holds that propositions are made up of (what are nowadays called) *concepts*, or *modes of presentation*, of the objects and properties our beliefs are about. But even among Fregeans there is disagreement about the nature of concepts, and it is not clear whether the difference between the Russellian and Fregean views is in practice anything more than alternative styles of bookkeeping, since Russellians agree that concepts are crucial to the individuation of thoughts and acts of communication. In what follows I shall assume, which I believe to be correct (Schiffer in press), that communication is a relation holding among a communicator, a communicated proposition, and a person to whom that proposition is communicated, without prejudice as to whether those propositions are Russellian or Fregean.

2. *The Explication of Communication*

How can the communication relation be explicated, or analyzed? One question that falls under this rubric concerns what we can know *a priori* about the *concept* of the communication relation. The strongest view here holds that *a priori* conceptual analysis can arrive at a correct, informative, and noncircular completion of the schema

(h) Necessarily, for any agent x, proposition p, and agent y, x communicates p to y if, and only if, … .

Those influenced by the late British philosopher J. L. Austin, such as the American philosopher John Searle, hold that a completion is possible that appeals to rules that are constitutive of the practice of communicating, in the way that the rules that define chess are constitutive of that practice, but none of these theorists has, to my knowledge, made any serious attempt actually to complete the schema, and in any case there are persuasive reasons to think that the appeal to 'constitutive rules' is mistaken (Austin 1975, Searle 1969, Schiffer 1972).

The only serious attempt to provide a complete *a priori* analysis of communication has been by the late British philosopher H. P. Grice and by others who sought to improve on his attempts (Schiffer 1972, Bennett 1976, Grice 1989). Three things are distinctive of the Gricean program. First, it attempts to define communicating p wholly in terms of the communicator's audience-directed intentions, included in which is her intention to get her audience to believe p. Second, the means by which the audience is intended to arrive at his belief is by taking the communicator's perceptible act (e.g., her uttering certain sounds) as evidence that she performed that act intending to get her audience to believe p, and then to take her having that intention as evidence that she herself believes, or believes she knows, p, and then taking her believing (she knows) p as good evidence that p is true. So, when Abe came to know that it was snowing as a result of Sally's telling him that it was snowing, his knowing is due to his having adequate evidence—viz., the fact that Sally believed (she knew) that it was snowing—for the fact that it was snowing. Third, linguistic meaning—the meanings of natural language expressions—was then to be defined in terms of the now reduced notion of communication (or, more accurately, in terms of a reduced notion of speaker meaning that includes telling as a species of speaker meaning). In the end, one would have a reduction of the semantic to the psychological, which, we shall later notice, would set the stage for a reduction of psychological facts to physical or topic-neutral facts—facts, such as the fact that one event was caused by another, which do not entail but are consistent with there being no nontopic-neutral facts other than physical facts. (*Functional* facts are topic-neutral facts that pertain to the causal or counterfactual relations relating internal states to one another and to sensory input and behavioral output.)

Conceptual analysis of the kind practiced by Griceans is now generally regarded as *so* twentieth century. In addition to this general skepticism about definition-seeking conceptual analysis, there are specific problems with the best efforts in the Gricean tradition. It has proved very difficult to state nonsemantic conditions that are jointly sufficient for communicating that such-and-such; the complexity of the analyses inevitably proffered in the attempt to get sufficient conditions makes one doubt that ordinary speakers have the supersubtle intentions those accounts ascribe to them; and, related to these

problems, it seems that if we subtract from a speaker's intentions her intention to be telling her audience that such-and-such, we are not left with enough non-semantic propositional attitudes (i.e., thought states that have propositional content) in terms of which one could define communication (Schiffer 1987, pp. 242–9). Armchair reflection on the concept of communication can yield important necessary truths about communication, but it is doubtful if it can yield a definition, let alone a definition in wholly non-semantic terms.

Much more influential than attempts to define communication have been attempts to show that communication facts, like all other intentional facts, stand in a relation to physical facts which renders the intentional facts naturalistically respectable. But there is no universally accepted view about what is required to render the intentional a respectable part of the natural order. Some philosophers insist that intentional properties and relations be identical with physical properties and relations. Such an identity need not be discovered by *a priori* conceptual analysis; it could have as its model the *a posteriori* identification of, e.g., heat with molecular motion. Yet this insistence seems chimerical when the intentional relations are relations to propositions. For the idea of a *physical relation* to *propositions* is evidently incomprehensible (what on earth could be a physical relation to the proposition that $1^2 + 1^2 = 2$?). More promising is the idea that propositional-attitude relations like believing are functional/causal relations. If there are such identifications, it is doubtful that they can be discovered by *a priori* conceptual analysis, and the reasons for supposing that there are *a posteriori*, scientifically discoverable identifications of propositional-attitude relations with particular functional/causal relations are considerably less than compelling, typically based either on implausible views about what is required by the 'unity of science,' the way in which the various special sciences are related to one another and depend on fundamental physics, or on a naive conception of what is required for propositional-attitude properties to play the causal/explanatory roles that they evidently do play.

A weaker position, but one accepted by many, is that while intentional relations cannot be identified with physical or topic-neutral relations, nevertheless all intentional facts *supervene* on non-intentional facts in that there could not be two possible worlds that were identical as regards to their non-intentional facts but not as regards their intentional facts. This position faces the challenge of having to offer some sort of explanation of the supervenience relation that anchors the intentional in the non-intentional, rather than taking that supervenience simply as a brute, inexplicable fact (Schiffer in press).

Finally, there are those who are content to hold a neo-Cartesian dualism wherein the intentional does not supervene on the nonintentional but does play the causal-explanatory role it appears to play (Bealer in press). The trouble with this view, however, is that it is very difficult to see how the intentional could plausibly play its causal role without the sort of reliance on underlying physical facts which the dualism precludes. Those who hold the neo-Cartesian dualism will doubtless appeal to *causal relations* between intentional states and their underlying physical states, but the difficulty just alluded to is precisely that of making sense of there being psychophysical causal laws or mechanisms capable of sustaining the alleged causal relations.

3. *Communication and the Explication of Linguistic Meaning*

What is it for marks or sounds to have meaning in a given population? Obviously, it has something to do with how the members of that population use, or are apt to use, those marks or sounds. If what we are concerned with is meaning in a public language like English or French, then it is reasonable to think that the use of those marks or sounds in *communication* looms large in the answer. On this much all agree, and David Lewis (1983) has provided an elegant framework for posing the further questions about the explication of linguistic meaning and the role of communication in it. Think of a language as an abstract object that may or may not be used by anyone; more specifically, think of it as a mapping of sequences of marks or sounds onto those things that are sentence meanings—propositions, perhaps, or, more likely, abstract entities that determine a specific kind of speech act and propositional content. Then the problem is to say what relation must obtain between a population of persons and a language so conceived in order for that language to be a language of that population. It is likely that this question enjoys no neat answer in terms of a single set of necessary and sufficient conditions, but it is probably not unreasonable to suppose that sufficient conditions can be stated that in fact explain what it is by virtue of which particular languages are used by particular populations. Such sufficient conditions will include communicative speech acts, but, what other kinds of speech acts? Can a unified account of speaker meaning be provided which subsumes all relevant speech acts? If so, are the Griceans right in claiming that those speech acts can be explicated in wholly psychological terms, thereby yielding a reduction of all public-language semantic facts to psychological facts? How exactly will the relevant kinds of speech acts enter into the account of what makes a language a language of a given population? Will some notion of *convention* be needed here? If so, what sorts of conventions, and how is the relevant notion of convention itself to be explicated? There is little agreement on how these

questions should be answered, and, anyway, philosophers have lately lost interest in the questions. Of much more contemporary interest is the explication of meaning in systems of mental representation, and it is not plausible that that explication needs recourse to communication (Field 1978, Schiffer 1987).

4. *Knowledge by Testimony*

Knowledge by testimony is the knowledge gained via communication; it is knowing something because it was told to you. The liveliest issue at present about communication concerns the nature of this knowledge. It is a lively issue because a certain view that was more or less taken for granted has been challenged. In order to understand what is at issue, we need to see what the assumed view was, what objections have been brought against it, and what new view, or views, of knowledge by testimony has been proposed in its place.

If one knows *p*, then *p* is true and whatever justifies one in believing *p* is what elevates one's true belief to knowledge. Accordingly, disputes about the nature of knowledge by testimony are disputes about the kind of knowledge-providing *justification* for one's believing *p* when one knows *p* by virtue of having been told *p*. The long-standing view that has been challenged is that this justification is just the ordinary sort of empirical justification one has whenever one knows something because one has adequate empirical evidence for it. There is no established view about the exact nature of such knowledge, but part of the traditional view is the idea that there is a single set of topic-neutral epistemic principles that determine when one is justified in the truth of any given proposition to be evidence for the truth of any other given proposition. Examples of knowledge gained in this way would be one's knowing that Smith murdered Jones on the basis of such-and-such DNA and fingerprint evidence; one's knowing that one's cat had been on the kitchen table on the basis of seeing cat hairs on the table; one's knowing that a smoker had recently been in the room on the basis of seeing a smoldering cigarette in the ashtray; and so on. In the case of communication, the idea, applied to an example, is that when Abe comes to know that it is snowing because Sally told him that, his knowledge-providing justification is his taking Sally's utterance of 'It's snowing' to be adequate empirical evidence that it was snowing, and that his so taking it is based not on any special faculty or principles of testimony but on ordinary empirical knowledge together with general inductive and deductive procedures. The Gricean conception of communication, as ordinarily conceived, presupposes this evidential view of knowledge by testimony. On this view, the key ingredient in Abe's coming to know that it is snowing is that he takes Sally's utterance to be evidence that she believes she knows that it is snowing, and his taking her having that belief to be evidence that it is snowing. None of this is to say that Abe is consciously thinking thoughts, or making inferences, about what Sally thinks, any more than you are consciously thinking the thoughts, or making the inferences, involved in your inductive justification when you see the smoldering cigarette and instantaneously conclude that someone has recently been smoking in the room. On this conception, knowledge by testimony *reduces* to knowledge gained by other means—by perception, memory, demonstrative, and non-demonstrative inference—and no more requires special epistemological mechanisms or principles of its own than does, say, knowledge gained by police detection.

Philosophers who reject the traditional conception of knowledge by testimony do so for two reasons: first, they have reasons for thinking that it cannot account for all that we know by testimony, and second, they have at hand an alternative conception of knowledge by testimony. As regards the direct objections to the traditional conception, Coady (1992) argues that if we took away all of an agent's knowledge based on testimony that itself did not depend on other knowledge by testimony, there would not then be the wherewithal to provide adequate empirical justifications for beliefs based on testimony. Tyler Burge has objected that while in gaining knowledge by testimony we rely on a presumption of reliability, we 'need not engage in reasoning about the person's qualifications to be rational in accepting what he or she says, in the absence of grounds for doubt' (Burge 1993, p. 469). This is evidently the sort of consideration also at work behind Robert Audi's claim that knowledge by testimony is

> 'typically not inferential. Certainly when trusted friends speak to us on matters we have no reason to think are beyond their competence, we normally "just believe" what they tell us.' (Audi 1998, p. 132)

Those who reject the traditional view hold that knowledge by testimony, in the typical case, is immediate and non-inferential. Sometimes this is developed via an analogy with the way in which perceptual knowledge is thought to be immediate and non-inferential (Coady 1992). According to Burge (1993), we have an *a priori* entitlement to belief what we are told immediately without argument or evidence, unless there are stronger reasons not to do so.

The direct objections are not persuasive. Coady presupposes that if knowledge by testimony were inferential (in the sense in which taking one thing to be evidence for another is inferential), then the inferences involved would rely on general beliefs about the reliability of testimony. This is unwarranted. On the most plausible inferential account, the crucial thing on which interlocutors rely is their commonsense knowledge of what would explain a given person's having a certain belief. Abe, for example, comes to believe that

it is snowing because he comes to believe that Sally believes she knows that it is snowing, and, using his general knowledge about the sorts of ways one is likely to acquire such a belief, Abe believes that it is very unlikely that Sally would have her belief if it were not snowing. The status of the relied-upon folk psychological theory may be taken to be on all fours with the status of the commonsense physical theory we rely on in forming our beliefs about the movements of physical objects.

As for the Burge–Audi objection, if believing what someone says off the bat, or not needing to engage in reasoning about a person's qualifications, were a reason for taking knowledge by testimony to be nonevidential, we could easily be shown not to have most of the evidential knowledge that we have. What 'reasoning' is one actively and consciously engaged in when one immediately concludes from the sight of a wet sidewalk that it recently rained, or when one immediately believes that one's Vermeer has been stolen upon seeing that one's alarm system has been disabled and that one's back door is wide open? Whether knowledge is based on inference from evidence is not about the actual movement of thought, the considerations one actually ponders; it is about the structure of beliefs that sustain one's conclusion. What justifies our beliefs about the rain and the Vermeer, and what justifies our taking those examples to be examples of evidential knowledge, is that they involve inferences to the correct explanation. One takes the wet streets to be evidence that it rained because one believes that the streets are wet *because* it rained; one takes recent rain to be the correct explanation of why the streets are wet. Likewise, when one sees the disabled alarm system and open door, one immediately and unhesitatingly makes an inference to the explanation whose ingredients include the intentional disabling of the alarm by a thief whose motivation was to get the only thing of value in the house and who successfully made his escape, Vermeer in hand, via the back door.

Likewise, when Abe unhesitatingly accepts Sally's statement that it is snowing, he instantaneously makes an inference to an explanation of her utterance which entails that she believed she knew it was snowing and that part of the explanation of this is that it was snowing. Proponents of the non-evidential view will concede that Abe would not believe that it was snowing if he also thought either that Sally did not believe that or that her belief was not partly explained by the fact that it was snowing, but they will then claim that it is not necessary that Abe actually believe that Sally had her belief or that she had it because it was snowing. This is a mistake; Abe need not consciously ponder any thoughts, but he must have the beliefs in question, just as you must actually have the belief that the streets are wet because it rained when you see the streets and instantaneously form the belief that it rained, even though, so to say, that belief need never have consciously crossed your mind. Nor does it matter that Abe should form his belief while knowing next to nothing about Sally. For suppose you now learn, in whatever way you care to imagine, that I believe I know that my middle name is 'Richard.' Even if you only know me as the author of this article, you would almost certainly yourself believe that my middle name is 'Richard.' Why so? Because, as you know, it is extremely unlikely that a philosophy professor who is coherent and sane enough to write this article would be mistaken about his middle name. On the other hand, your knowing that I believe that the law of excluded middle is not determinately true or that so-and-so will win the next US presidential election probably would not significantly affect your beliefs about those propositions.

Philosophers who reject the evidential view may do so because they are mistaken about what is required to know something on the basis of an inference to the correct explanation. Two points seem relevant in this connection. First, in forming one belief a person may be relying on another belief even though the second is merely tacit, never brought to conscious thought. When you see the wet streets and actively believe that it rained during the night, you probably at no time entertained or actively believed that the streets are wet *because* it rained last night; yet you believe that all the same and would not believe that it rained last night if you did not. Likewise, Abe, in coming to believe that it snowed, probably does not actively believe that Sally would not believe that it snowed if it had not snowed, but he does have that belief and would not believe that it snowed if he did not. Second, a seven-year-old child may know that it rained on the basis of her perception of wet streets, and this because he/she knows that the streets are wet *because* it rained. But how does our seven-year-old know *that*? It is unlikely that he/she spent any time actually considering alternative explanations of the streets' wetness. The fact is, I submit, knowledge-providing justification is easy to come by in cases like this; we are relaxed in our ascriptions of knowledge in such cases, at least when nothing more is at stake. We count the child as knowing because we think he is right and because we do not think there is an alternative explanation of what made the streets wet that is a plausible competitor. Come to think of it, what is true of the child is true of most adults, or else we have very little evidential knowledge. We count someone as knowing, on the basis of his perception of wet streets, that it rained even though if asked to rule out alternative explanations of the streets' being wet he would not be able to articulate much of anything.

All that is required for knowledge by testimony to be evidential is that when we are justified in believing p because we were told p, then, roughly speaking, we are also justified in believing that the speaker would not have told us p unless she believed she knew p and that she would not have that belief if p were not the case. Moreover, as already indicated, to be justified in

thinking that someone would not have told us *p* if she had not believed (she knew) *p* need not require any special antecedent knowledge of the speaker's honesty or reliability. I am giving a talk on an unfamiliar campus and ask an unknown professorial-looking man where Blah-Blah Hall is. He points to a visible building; I head off in what I am confident is the right direction. But how did I know he was not lying, crazy, mistaken, etc.? By inference to the correct explanation; the assumption that he is expressing his belief, etc., is entailed by my tacit explanation of why he pointed and spoke as he did.

One who holds that Abe has non-evidential knowledge by testimony that it is snowing might concede that Abe has a justified true belief that Sally believes it is snowing in part because it is snowing, but yet insist that his justification is not good enough to yield knowledge. This is not very plausible. Abe knows that *if* it is snowing, *then* Sally believes it is snowing in part because it is snowing. So in order for him not to know the consequent of the conditional he knows, it would have to be possible—since it is agreed that Abe knows it is snowing—for one to know [*p* and if *p*, *q*] and not know *q*. And while there are indeed those who are willing to hold that this is possible (Dretske 1970, Nozick 1981), it remains extremely counterintuitive and is widely rejected.

I now know that the cup before me is red. I also know that its being red is the correct explanation of why I am having the visual experience I am now having. Yet my knowing that the cup is red may in this case be overdetermined, both direct *and* based on the evidence of my visual experience. In this case it is not unreasonable to suppose that, in addition to my evidential knowledge that the cup is red, I also have immediate and nonevidential knowledge that the cup is red. For it is not unreasonable to suppose that, given the way we are built as information processors, my having my visual experience would cause me to believe that the cup was red even if I did not have the conceptual wherewithal to form a belief about my sensory experience or its causes. This is evidently the moral to be drawn from the perceptual knowledge of certain animals and very young children. But knowledge by testimony seems disanalogous in an important respect. While very young children seem to believe what they are told even though they are conceptually unable to form beliefs about the beliefs of speakers or about the etiology of those beliefs, it is also the case that it is very difficult to see how there could be communication in public language by creatures who did not have a fairly rich commonsense propositional-attitude theory that allowed them reliably to ascribe beliefs to others and to form reliable hypotheses about the etiology of those beliefs. What we can conclude is that when an adult gains knowledge by testimony, it is because she knows that the speaker believes (she knows) the communicated proposition and that she has that belief in part because the communicated proposition is true. Whether anything like knowledge by testimony could occur in mature creatures who did not tacitly form such beliefs about the beliefs of speakers is perhaps to some extent an open question. If it is possible, then perhaps we should allow that knowledge by testimony is overdetermined; but whether or not it is overdetermined, the knowledge by testimony actually gained by mature persons is arguably evidential.

See also: Communicative Competence: Linguistic Aspects; Information Theory; Verstehen und Erklären, Philosophy of

Bibliography

Audi R 1998 *Epistemology: A Contemporary Introduction to the Theory of Knowledge*. Routledge, London
Austin J L 1975 *How to Do Things with Words*. Clarendon Press, Oxford, UK
Bealer G in press *The Integrity of Mind: A Non-Reductionist Account of Human Nature*. Oxford University Press, Oxford, UK
Bennett J 1976 *Linguistic Behaviour*. Cambridge University Press, Cambridge, UK
Burge T 1993 Content preservation. *Philosophical Review* **102**: 457–88
Coady C A J 1992 *Testimony: A Philosophical Study*. Oxford University Press, Oxford, UK
Dretske F 1970 Epistemic operators. *Journal of Philosophy* **67**: 1007–23
Field H 1978 Mental representation. *Erkenntnis* **13**: 9–61
Grice H P 1989 *Studies in the Ways of Words*. Harvard University Press, Cambridge, MA
Lewis D 1983 Languages and language. In: *Philosophical Papers*, Oxford University Press, Oxford, UK, Vol. 1
Nozick R 1981 *Philosophical Explanations*. Harvard University Press, Cambridge, MA
Quine W V 1970 *Philosophy of Logic*. Prentice-Hall, Englewood Cliffs, NJ
Schiffer S R 1972 *Meaning*. Clarendon, Oxford, UK
Schiffer S R 1987 *Remnants of Meaning*. MIT Press, Cambridge, MA
Schiffer S R in press *The Things We Mean*. Oxford University Press, Oxford, UK
Searle J R 1969 *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, London

S. Schiffer

Communication, Two-step Flow of

The concept of the 'two-step flow of communication' emerged from *The People's Choice*, the 1948 American voting study conducted by Paul Lazarsfeld and his colleagues at the Bureau of Applied Social Research of Columbia University. They observed that: 'ideas often flow from radio and print to opinion leaders and from

them to the less active sections of the population.' Thus, the flow of information and influence from the mass media to their audiences was suggested as taking place in two steps: from the media to the opinion leaders and from them to the public. This model certainly called into question most of the assumptions of the 'powerful media' notion by revealing the limits of media influence while highlighting the role played by personal influence and especially by certain individuals—the opinion leaders. These ideas caught the attention and imagination of a new generation of researchers who opened up a new theoretical and empirical vista.

1. *The Emergence of the Concept*

Beginning with *The People's Choice*, Lazarsfeld and his colleagues were occupied with a series of panel studies on the role of mass communication in the making of decisions in various areas (voting, consumption, public issues and others). The researchers focused on the impact of campaigns in the media and were surprised to find that interpersonal communication had a stronger impact on individual attitudes than the mass media alone. Moreover, they found that certain individuals were more central and influential in their groups, often acting as intermediaries between the mass media and the public. In comparison with the rest of the population, these individuals were found to be more exposed to the mass media. Thus, if personal influence was so important and if it was executed mainly by certain individuals who were more exposed to the mass media than the people they influenced, then it could be postulated that there was a two-step flow process from the media to the opinion leaders and from them to their followers. These intermediaries, called Opinion Leaders, could allow messages to pass through (backing them with their own personal authority), strengthen or weaken them, or block their passage entirely. This became the 'two-step flow of communication' model.

The most thorough and systematic examination of the two-step flow model and the notion of opinion leaders was the Decatur Study (after the community in which it was conducted). The study's procedures and findings were published in *Personal Influence* and represented one of the most impressive early documentations of the opinion leadership concept (Katz and Lazarsfeld 1955). The research focused on four areas of decision making in daily life: marketing, fashion, public affairs, and movie selection. By isolating specific changes of opinion and identifying the influences that allegedly produced them, the researchers located the individuals who were mentioned most frequently as the source of influence. As this was a pioneering study, Katz and Lazarsfeld applied various methods and measures to locate and identify the opinion leaders. Briefly, there are two ways to locate opinion leaders in specific areas of influence. One is through the testimony of those whom they have influenced. The other is through a self-designation procedure in which respondents themselves are asked about their influence on others. To verify the 'self-designating' answers, the researchers requested the confirmation of the persons mentioned as advice-seekers. Two-thirds confirmed the contact and the topic. Although the researchers were satisfied with this 'validation,' it is far from being an acceptable measure of validity and indeed attracted later criticism.

The early studies on personal influence and opinion leadership resulted in several attributes related to opinion leaders (Katz 1957). Let us briefly review the main ones: (a) Opinion leaders are found at every social level and in most areas of decision making, influence people from the same social level. (b) Opinion leaders are found in both of the sexes, all professions, all social classes, and all age groups. (c) Opinion leaders tend to be more involved in various social activities and social organizations and occupy central positions in their personal networks. (d) Opinion leaders are considered to be experts in their field, but this is informal recognition by close friends, relatives, co-workers, colleagues, and acquaintances. (e) Opinion leaders are more exposed to the mass media than nonleaders. (f) Opinion leaders are more interested, involved, and updated in the field in which they are influential. (g) Opinion leaders tend to be monomorphous, that is, they are usually experts in one area but rarely in various areas (i.e., polymorphous). (h) Opinion leaders manifest a specific communication behavior insofar as they are more involved in formal and informal personal communication than nonleaders. (i) Opinion leaders are usually well aware that they serve as sources of information and influence for others.

Since the introduction of the two-step flow model and the opinion leadership concept, numerous studies have sought to advance both the understanding and applicability of these ideas. Already by the early 1970s several hundred of these studies were available and the number has continued to grow steadily (for a review of almost 1,400 of these studies, see Weimann 1994). Thus, the 1960s and 1970s may be considered as the 'Golden Age' of the research on opinion leaders, yielding convincing empirical evidence of the opinion leaders concept in various areas: marketing and consumer behavior, fashion, politics and voting, family planning, science and scientific innovations, agriculture, and health care.

2. *Criticism of the Model*

The model's theoretical and practical appeal crossed the boundaries of disciplines, societies, and cultures. It

attracted researchers from medicine to politics, marketing to family planning, AIDS prevention to clothing fashion, and agriculture to birth control. However, the concept has been a subject of growing criticism. The weaknesses and shortcomings of the original conceptualization of opinion leadership, and especially that of the two-step flow model, were exposed by several studies which combined theoretical, ideological, and methodological aspects (for a review, see Katz 1987).

2.1 *Theoretical-ideological Criticism*

The media, according to the Columbia research tradition, were not as powerful as was assumed, once the public was found to be more active, less lonely, and less vulnerable. Defended by their selective consumption of the media and by the selective mediating function of the opinion leaders, people could fend off media influences and protect their attitudes, preferences, and values. The limited effects of the media as a central element in the opinion leadership and two-step flow concept became the main source for the theoretical-ideological criticism directed at this paradigm. Gitlin (1978), for example, rejected the two-step flow model and the functions related to the opinion leaders. The opinion leader is celebrated, he argued, because administrative researchers study proximate causes which are accessible to their patrons, the media organizations themselves, and the dominant values of the capitalist society. The ideological-theoretical bias of the two-step model, according to this criticism, is manifested in a set of assumptions: (a) *Commensurability of modes of influence*: the impact of the mass media was presumed to be comparable to personal influence, with the two 'forms of influence' regarded as functionally equivalent. (b) *Power as distinct occasions*: the measurement of influence by studying discrete incidents in a 'behaviorization of power.' (c) *Commensurability of buying and politics*: the areas of marketing, politics, fashion, and movie-going were perceived as comparable within a single theory and measurable by a single method. (d) '*Change*' *as the sole dependent variable*: influence was identified by the occurrence of 'attitude change,' disregarding the case of reinforcement as an effect or the prevention of change as an effect. The latent power of the media to prevent changes and maintain the status quo should not be regarded as a residual category, but as a major function and a powerful effect.

2.2 *Empirical-methodological Criticism*

The empirical-methodological weaknesses cited (see Weimann 1982) include the following aspects: (a) Ignoring the evidence of direct flow from the media to the public. (b) Ignoring horizontal flow among leaders or among followers. (c) Ignoring the more-than-two-step flow, or 'longer chains of influence.' (d) The use of a crude dichotomy between leaders and nonleaders instead of applying continuous measures of leadership. (e) Standardization of measurement for information flow and influence flow. (f) Standardization of measurement across domains, disregarding the variance across domains of decision making. (g) The use of short timeframes, far from reflecting the long-term, cumulative impact of media messages.

3. *The Paradigm's Fight for Survival*

The wave of criticism led to a decline in the popularity and attraction of the original concept and almost to its total collapse. However, the paradigm's 'fight for survival,' through a dynamic exchange of arguments, criticism, response, and debate, resulted in several modifications of the original model.

The first amendment to the two-step flow of communication was the introduction of multistep rather than two-step flow. This allowed the inclusion of other optional directions of flow, such as horizontal (among leaders, among followers) and vertical (leaders to followers, active to inactive); direct (linking two actors directly) and indirect (linking actors through a third party or a chain of intermediaries); and downward (leaders to followers, media to public) and upward (from opinion leaders to media, followers to influentials, marginals to centrals). Another important modification was the introduction of continuous measures of opinion leadership and their validation (e.g., Noelle-Neumann and Csikszentmihalyi 1992, Weimann 1991, 1994).

Finally, the most promising development is the merging of the model with modern network analysis (Weimann 1989). The developing theory and methodology of network analysis, combined with modern technologies of data processing, namely high-speed computers, special programs for sociometric data analysis, and such sophisticated mathematical procedures as matrix algebra and sociometric topology, has facilitated the study of large communities. It has also contributed to 'modernizing' the concepts of personal flow of media messages, opinion leadership, and flow of influence. Sheingold (1973) called for the use of social networks data as a 'resurrection' of the Columbia research agenda, and indeed much research on social networks in recent decades has focused on the role of influentials or 'centrals,' applying various structural measures taken from the network analysis procedure. These new developments have linked the two-step flow model with other theories and concepts, such as the Agenda-Setting process (e.g., is there a two-step flow of agenda setting?; see Weimann and Brosius 1994, Brosius and Weimann 1996) and have thus contributed to revitalizing some of the notions of the two-step flow model.

See also: Advertising: Effects; Advertising: General; Consumer Economics; Consumer Psychology; Consumption, Sociology of; Influence: Social; Lazarsfeld, Paul Felix (1901–76); Media Effects

Bibliography

Brosius H B, Weimann G 1996 Who sets the agenda? Agenda-setting as a two-step flow. *Communication Research* **23**: 561–80

Gitlin T 1978 Media sociology: The dominant paradigm. *Theory and Society* **6**: 205–53

Katz E 1957 The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly* **21**: 61–78

Katz E 1987 Communications-research since Lazarsfeld. *Public Opinion Quarterly* **51**: S25–S45

Katz E, Lazarsfeld P F 1955 *Personal Influence*. Free Press, Glencoe, IL

Lazarsfeld P F, Berelson B, Gaudet H 1944 *The People's Choice*. Duell, Sloan and Pearce, New York

Noelle-Neumann E, Csikszentmihalyi M 1992 *Personality Strength: A New Variable for Opinion-Attitude Research*, paper presented at the WAPOR Conference, St. Petersburg Beach, Florida

Rogers E M, Shoemaker F F 1971 *Communication of Innovation*, 2nd edn., Free Press, New York

Sheingold C A 1973 Social networks and voting: The resurrection of a research agenda. *American Sociological Review* **38**: 712–20

Weimann G 1982 On the importance of marginality: One more step into the two-step flow of communication. *American Sociological Review* **47**: 764–73

Weimann G 1989 Social networks and communication. In: Asante M K, Gudykunst W B (eds.) *Handbook of International and Intercultural Communication*. Sage, Newbury Park, CA

Weimann G 1991 The influentials: Back to the concept of opinion leaders? *Public Opinion Quarterly* **55**: 267–79

Weimann G 1994 *The Influentials: People who Influence People*. SUNY, Albany, NY

Weimann G, Brosius H B 1994 Is there a two-step flow of agenda-setting?*International Journal of Public Opinion Research* **6**: 323–41

G. Weimann

Communicative Competence: Linguistic Aspects

The term 'communicative competence' has come to be applied both to the knowledge that speakers have about how to use their linguistic repertoire, and to their degree of success in doing so. As such, the history of its evolution is entwined in the often heated debates about 'competence' and 'performance' in the emergence of interactional sociolinguistics and the ethnography of communication within sociolinguistics (see *Sociolinguistics*), and in the gradual definition of pragmatics in the context of linguistic and sociolinguistic theory (see *Pragmatics: Linguistic*). In addition, it has been used extensively in the applied linguistics field of language teaching and testing.

1. From Whence comes Communicative Competence?

In the mid-to-late 1960s, revolution was brewing in American linguistics. Noam Chomsky was busy sweeping away the structuralist approach to linguistics, which had resulted in catalogs of observations that overemphasized the diversity of languages, and, he felt, obscured the essential similarity between languages. After all, he reasoned, learning one's language (the singular 'language' is deliberately chosen here) is a natural human capacity; whatever allows young children to learn Navajo or French or Swahili, must, at root, be the same. There must be some underlying language capacity, common to all humans, which, when combined with experience of particular languages, results in a system of linguistic knowledge for those languages. Chomsky deliberately idealized the knowledge he was trying to describe, assuming for the sake of argument (and scientific advance) that all speakers of a language shared the same knowledge.

From these early speculations grew the notions of competence and performance (Chomsky 1965). Competence describes the knowledge that a speaker has of his or her language, the knowledge which underlies the production and comprehension of that language. Importantly, for Chomsky, competence is not, and has never been, an evaluative term. He uses it instead of the term 'knowledge' to avoid suggestions that knowledge of language is, or might be, conscious, when in fact knowledge of language is largely tacit and unconscious. So, his use of competence is not the ordinary use of the term, and must not be interpreted as contrastable with incompetent on some evaluative scale. Chomsky further argued that language behavior, actual utterances coming out of the mouths of speakers and being understood by hearers—performance—was but a poor reflection of underlying competence, beset as it could be by slips of the tongue, alcohol or novocaine induced slurring, and other physical impediments of various kinds. Thus, for Chomsky, the competence vs. performance distinction was very unevenly balanced, with all the richness and interest on the competence side, and just behavioral mishaps on the other.

This characterization incensed sociolinguist Dell Hymes, who was very interested in how young bilingual children choose which language to use and when, and how all speakers modify their language(s) depending upon the context in which they are using them. He had

been gradually evolving a way of looking at language in use which he called the 'ethnography of communication,' and which focused on the social meanings created and negotiated between people in communication. He emphasized the shortcomings of most of the sociolinguistics of the time which focused on statistical analyses of large populations, and argued for a focus on the blow-by-blow organization of interpersonal communication. Even though Chomsky was similarly rejecting the impersonal cataloging of linguistic facts, and was putting the focus on the task of the individual as a learner of any humanly possible language, Hymes felt that Chomsky's idealization of the language capacity as something possessed by a monolingual, and stripped of all the variation he observed in the behavior of even a single individual moving among different contexts of use made a mockery of what language was all about for him. Hymes was not prepared to follow Chomsky down a path which led to linguistics being seen as part of cognitive science; he was firmly committed to language as a social and behavioral phenomenon.

In reaction to Chomsky's coinage of the term competence, therefore, Hymes created the term 'communicative competence' which he intended to subsume and extend Chomsky's notion. The most celebrated quotation from Hymes is

> A child capable of any and all grammatical utterances, but not knowing which to use, not knowing even when to talk and when to stop, would be a cultural monstrosity. (Hymes 1967, p. 16)

He thus argues that the fundamental problem is

> to discover the underlying communicative competence that enables members of a community to use and interpret the use of language, and to provide a formal description that is a theory of that competence. (Hymes 1967, pp. 17, 18)

It is worth noting that the term was 'in the air' at the time. Not only was it beginning to be used in the field of second and foreign language teaching, but Campbell and Wales (1970) used it in connection with language acquisition. They are usually credited as having coined the term independently, although Hymes (1992) suggests that they might have indirectly got it from him by talking to the people he was talking to at the time.

2. *Hymes' Communicative Competence*

Hymes' notion of communicative competence was born out of a belief that acts of speech, such as promising or asserting or denying are produced and understood within the context of particular 'speech events,' such as conversations, lectures, or marriages, themselves occurring within particular speech situations within a speech community. The particular value that a speech act has is determined by the layers of context within which it is embedded. Importantly, the context includes not just objective factors such as who is speaking, where, and when, but subjective factors, such as what participants in a speech event believe themselves to be doing, and how they 'take' what is said to them. These speaker/hearer beliefs vary both individually and culturally, and cannot be ignored when trying to understand what a speaker means by an utterance. Cultures vary, for example, in how they view the loquacious individual, in the extent to which information is power, and to what status they accord the language of infants. Factors such as these significantly affect the microsociolinguistic world of interpersonal communication. Thus, for Hymes, it was fundamentally important that such factors be properly integrated into linguistics and that 'sociolinguistics' come to be synonymous with 'linguistics.' Since then, communicative competence has come to cover all aspects of language use, including text cohesion and coherence, rhetorical organization, speech act organization, register variation, dialect variation, and so on.

3. *Reactions and Counter-reactions*

The term communicative competence was accepted with alacrity by many other influential sociolinguists, including John Gumperz, Susan Ervin-Tripp, and Michael Halliday. The field of sociolinguistics sustained rapid advances as a result, seeing itself increasingly at loggerheads with formal linguists, which retained its focus on sentence-level grammatical issues. However, Chomsky was not unaffected by the firestorm. In his book *Rules and Representations* (1980) Chomsky refers to 'pragmatic competence,' and defines it as follows.

> Furthermore, the person who knows a language knows the conditions under which it is appropriate to use a sentence, knows what purposes can be furthered by appropriate use of a sentence under given social conditions. For purposes of inquiry and exposition, we may proceed to distinguish 'grammatical competence' from 'pragmatic competence,' restricting the first to knowledge of form and meaning and the second to conditions and manner of appropriate use, in conformity with various purposes. (Chomsky 1980, p. 224)

It is clear that, although he is not using the term communicative competence, he has many, if not most, of the key elements of it (see also *Pragmatics: Linguistic*). Moreover, he situates 'pragmatic competence' in relation to 'performance':

> Theories of grammatical and pragmatic competence must find their place in a theory of performance that takes into

> account the structure of memory, our mode of organizing experience, and so on. (Chomsky 1980, p. 225)

This quotation is important, because when Chomsky first proposed the 'competence' vs. 'performance' distinction, sociolinguists could see that their concerns did not fit into his restricted notion of competence, which meant that if they were anywhere, they must be part of what was at the time a very impoverished notion of 'performance' (slips of the tongue, etc., see above). So many writers, trying to find a place for communicative competence have seen it as part of a theory of 'performance,' and argued for it as such, despite the clear division Chomsky makes in the quotation above. The result is an ambiguity in the literature between communicative competence as knowledge and communicative competence as ability or skill for performance. This has meant that the term has sometimes been interpreted very cognitively, and at others very behaviorally, and at yet others, systematically ambiguously.

4. *Communicative Competence as Ability/Skill*

Despite the originally stated intention to make communicative competence match 'competence' as a formal description of what a person who knows a language knows, many writers have used the term in an evaluative sense in which various kinds of communicative competence are seen as being on continua from incompetence to competence. For example, in educational circles, communicative competence has often been used as one measure of a learner's success. The work of Basil Bernstein, for example, has usually been seen as suggesting that children from certain kinds of background are likely to be, in fact, communicatively incompetent in the terms defined by school success (Bernstein 1990). Children who cannot use the language in the ways expected by the school suffer from the effects of this incompetence. Similarly, in work by Cummins (1984) and others, bilingual children have been claimed to be communicatively incompetent to the point of being labeled semilingual because their ability to use each of their languages satisfies native speakers of neither language, who find them communicatively 'incompetent' in both languages.

In second language acquisition studies (see below) concerns with making second language learners competent users of the language have driven a range of studies under the umbrella of communicative competence. In all these sub-fields (and there are many others), the evaluative sense of communicative competence plays the main role. It is worth asking whether the ambiguity in the term's use is problematic or not. The answer is likely to depend more than anything on one's philosophical stance with respect to the nature of knowledge. Is knowledge something represented, something separate from behavior, or is knowledge simply the term we give to the capacity to act? This in turn depends on one's approach to the division between cognitive science on the one hand, and social and behavioral science on the other. For those who wish to see the two fields come together, the ambiguity of the term, straddling as it does the border between them is perhaps to be encouraged. For those, however, for whom cognitive science is a fundamentally different enterprise, in which behavior is the outward manifestation of knowledge (whether represented declaratively or through connectionist networking), the issue is more fundamental.

5. *Communicative Competence in Language Acquisition*

A range of different fields have used, and developed, the communicative competence notion. Two will be discussed briefly here: the sub-fields of first-language acquisition and of second-language acquisition.

In first-language acquisition, studies on very early communication prior to the emergence of real language (at a point where gestures and single 'words' are all the child has at his or her disposal) were particularly popular in the 1970s, and still continue in work as diverse as that trying to determine whether chimps can be said to have a communicative competence or that which aims to discover what kind of communicative system deaf children will invent if they are not presented with an acquirable language. Work with older children has focused, and continues to focus on how children develop the range of means to express speech acts, how they choose between styles of speech, how they switch between languages and dialects, and so on (cf. Romaine 1984, which also has an interesting discussion of the concept of communicative competence). There are also those interested in children who for one reason or another appear to have a relatively intact ability to communicate in the face of severe disorders to the linguistic system, and in those who have the reverse pathology, being capable of producing correctly structured utterances, but with severely disordered abilities to use language appropriately. These dissociations, if they truly exist (and current debates show divided opinion here), raise a range of epistemological questions about the status of linguistic competence in relation to pragmatic or communicative competence. General introductions to child language development, covering the issues listed above among many others can be found in Foster-Cohen (1990, 1999).

Considerable energy has been put into trying to model communicative competence correctly within the field of second language testing and second language acquisition. For language testers, interested in the evaluation of communicative competence (used

evaluatively, of course), it is vital that all the parts of the notion be laid out and defined, so that testers can be sure that they know what they are testing. Most discussions in this context make reference to Hymes' work, although the initial use of the term communicative competence in this context appears to have been independent of him. At first, communicative competence for language teachers meant simply an ability to get a message across in the target language, to be capable of 'spontaneous expression.' However, when Hymes' work became known, many language teachers adopted the more broadly sociolinguistic sense of the term, and argued that teaching the social meanings of second language expressions are as important as teaching the structures

Michael Canale and Merrill Swain are perhaps the best known for expanding the notion of communicative competence for those interested in second language acquisition in an article published in 1980. There they distinguish between 'grammatical competence' (syntax, morphology, lexis, semantics and morphology), 'sociolinguistic competence,' and 'strategic competence.' Sociolinguistic competence covers both sociocultural rules of how to use linguistic expressions appropriately, and rules of discourse, that is how to string utterances together appropriately to produce a contribution (written or spoken) that a native speaker would recognize as well-formed. Strategic competence covers both verbal and nonverbal communication strategies that allow a nonnative speaker to get over communication problems due to inadequacies in either grammatical or sociolinguistic competence, or due to performance limitations.

More recently, Lyle Bachmann (1990) has reemphasized the dynamic nature of communicative ability. The negotiation of meaning between participants in an interaction, emphasized in sociolinguistics by Michael Halliday from the early 1970s and in language teaching by Sandra Savignon from the early 1980s, picks up on the idea of the 'emergent properties of events' noted by Hymes when the communicative competences of two individuals come in contact. Bachmann's model overtly places performance mechanisms within communicative competence which he, of necessity, calls 'communicative language ability,' having given up the original competence vs. performance distinction with which Chomsky and Hymes began. Although designed for language teachers and testers, Bachmann's discussion lays out particularly clearly the range of knowledge and abilities that a communicatively competent language user must control.

6. *Communicative Competence and Pragmatic Competence Revisited*

It was noted above that Chomsky's definition of pragmatic competence bears strong resemblance to Hymes' definition of communicative competence. However, it must not be forgotten that the perspectives of the two are still quite different. Whereas Hymes' is interested in accounting for the diversity in language behavior, Chomsky is interested in the similarities across individuals, in discovering the universal properties which make us human. A similar mindset can be seen among all those who are interested in pragmatic competence as an account of human understanding of communication. Sperber and Wilson, in their account of pragmatics known as 'Relevance Theory' (Sperber and Wilson, 1986, 1995), focus on the inferencing procedures that individuals use to interpret utterances in context. They view their enterprise as explicating a cognitive approach to communication, and can as such be seen as providing a theory of communicative competence in the Chomskyan sense, but not in the Hymesian sense.

However, this theory is entirely compatible with the range of facts which a sociolinguistic account of communicative competence sees itself as needing to account for. Just as grammatical theory has shown its capacity for accounting for cross-linguistic, cross-dialectal, and idiosyncratic features of languages, so a good theory of pragmatics should be able to show how human communication works across similarly diverse individuals. Gumperz (1997) calls for such a general theory when he invokes the work of Jürgen Habermas on communicative competence. Habermas is interested in 'the universal conditions of possible understandings' between people, ultimately in the political and social realities created through communication; but Gumperz argues for a definition of communicative competence as

> the knowledge of linguistic and related communicative conventions that speakers must have to initiate and sustain conversational involvement … Studies of communicative competence, therefore, must deal with linguistic signs at a level of generality which transcends the bounds of linguists' grammatical systems and must concentrate on aspects of meaning or interpretation more general than that of sentence content. (Gumperz 1997, pp. 40, 41)

Sperber and Wilson would certainly agree. A concern for a general theory of human understanding, then, drives researchers from widely divergent backgrounds and training, and with significantly different assumptions.

See also: Competencies and Key Competencies: Educational Perspective; Communication: Philosophical Aspects; Communication and Social Psychology; Gender and Language: Cultural Concerns; Language and Society: Cultural Concerns; Nonverbal Communication, Social Psychology of; Psycholinguistics: Overview; Verbal Communication, Psychology of

Bibliography

Bachmann L F 1990 *Fundamental Considerations in Language Testing*. Oxford University Press, Oxford

Bernstein B 1990 *The Structure of Pedagogical Discourse. Class, Codes and Control: IV*. Routledge, London

Campbell R, Wales R 1970 The study of language acquisition. In: Lyons J (ed.) *New Horizons in Linguistics*. Penguin Books, London

Canale M, Swain M 1980 Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* **1**: 1–47

Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Chomsky N 1980 *Rules and Representations*. Columbia University Press, New York

Cummins J 1984 *Bilingualism and Special Education: Issues in Assessment and Pedagogy*. Multilingual Matters, Clevedon, UK

Foster S 1990 *The Communicative Competence of Young Children*. Longman, London

Foster-Cohen S 1999 *An Introduction of Child Language Development*. Longman, London

Gumperz J 1997 Communicative competence. In: Coupland N, Jaworska A (eds.) *Sociolinguistics: A Reader*. St. Martin's, NY

Hymes D 1967 Models of the interaction of language and social setting. *Journal of Social Issues* **23**: 8–28

Hymes D 1972 On communicative competence. In: Pride J, Holmes J (eds.) *Sociolinguistics*. Penguin Books, Harmondsworth

Hymes D 1992 The concept of communicative competence revisited. In: Pütz M (ed.) *Thirty Years of Linguistics Evolution: Studies in Honor of René Dirven*. Benjamins, Philadelphia, PA

Romaine S 1984 *The Language of Children and Adolescents: the Acquisition of Communicative Competence*. Blackwell, Oxford

Sperber D, Wilson D 1995 *Relevance Communication and Cognition*, 2nd end. Blackwell, Oxford

S. H. Foster-Cohen

Communism

Karl Marx famously began his *Manifesto of the Communist Party* (1848) with the words, 'A spectre is haunting Europe—the spectre of communism.' If that was an exaggeration at the time and for most of the nineteenth century, it became a significant reality in the twentieth century. The Bolshevik Revolution of 1917 in Russia saw the establishment of what came to be called by Communists 'the world's first socialist state.' For Communists themselves, 'Communism' meant both an international movement concerned with overthrowing capitalist systems and establishing a new social order and a society which would exist only in the future, 'communism' (with a lower case 'c'). It referred, in that second meaning, to the final goal of political and social development—the withering away of the state and the creation of a utopia in which the need for hierarchy and coercion would no longer exist. Instead, there would be a society of equals in which people would live harmoniously and collectively administer their affairs. Karl Marx (1818–1883) had, not surprisingly, very little to say about how this projected society of the future would work and only once distinguished between an initial and higher stage of communism. V. I. Lenin (1870–1924), who led the Bolshevik Revolution in Russia in 1917 and became the founder of the Soviet state, made a firmer distinction between what he called 'socialism' and 'communism.'

Soviet theorists continued to describe their system as 'socialist' and, as the years passed, the ultimate goal of 'communism' receded ever further into an uncertain future. However, the main instrument of rule within the Soviet Union, and subsequently within other 'socialist' countries, was the Communist Party. Most objective analysts of those systems have, with good reason, chosen to call them Communist rather than socialist. The term, 'socialist,' embraces a far wider range of social movements and governments than those which owed their allegiance to Marxism–Leninism (as the Soviet codification of the works of Marx and Lenin became known). Communist Parties were far from having a monopoly on the understanding of socialism and, indeed, in many countries—most notably, in Western Europe—socialist parties of a social democratic type competed throughout the twentieth century (generally with more success) with Communist Parties for the working-class vote.

While there were some important differences from one Communist country to another during the political heyday of Communism—a forty-year period between 1949 and 1989, when more Communist states existed than at any other time before or since—they had enough in common to constitute a distinct subset of the world's political systems. Five defining characteristics can be discerned to differentiate them from other totalitarian or authoritarian regimes and, much more fundamentally, from pluralist systems in which from time to time political parties calling themselves 'socialist' held governmental office. The first of these defining characteristics was the *monopoly of power of the Communist Party*, for which the euphemism employed in Communist states, especially after the death of Stalin, was 'the leading role of the party.' (In Stalin's time the term, 'the dictatorship of the proletariat,' was used to signify this phenomenon. In a peculiar conflation of social categories, it was taken as axiomatic that in the Soviet Union the Communist Party embodied the will of the proletariat.) The supreme authority and unchallengeable hegemony of the Communist Party was a given of Communist systems, although how much power was concentrated in the hands of a ruling group within that party and

how much in the hands of a single leader varied over time and from one Communist state to another.

The second defining characteristic of a Communist system was *democratic centralism*, a key concept embodied in the published rules of Communist Parties which, in practice, signified centralism to the virtual exclusion of democracy. Although often distinguished by Communist theorists from 'bureaucratic centralism' (which meant officials acting in a high-handed way, without any consultation with those lower in the party hierarchy), democratic centralism, in political reality, stood for a highly centralized, strictly disciplined party with extremely narrowly interpreted rights of intra-party debate. The slightest criticism of the party leadership from below and any horizontal, as distinct from vertical, links between party organizations were stigmatized as deviations from democratic centralism and harshly punished.

State, or at any rate non-capitalist, *ownership of the means of production* constituted the third defining feature of Communism. In Yugoslavia, a system of supposed self-management was developed which differed from the unadulterated state ownership of industry characteristic of the Soviet Union, most of Eastern Europe, and also of China until late in the twentieth century. Exceptions to state ownership were occasionally made in agriculture—Poland, for example, maintained an essentially private agriculture throughout most of its Communist period, in contrast with the collective and state farms to be found in the USSR and elsewhere—but not for industrial production.

The fourth defining characteristic of Communism was *the declared aim of building communism—a society in which the state would have withered away—as the ultimate, legitimizing goal*. This theoretical aspiration was not part of the operational code of a Communist system and in terms of everyday politics was far less important than the 'leading role of the party' or 'democratic centralism.' It was, nevertheless, of great ideological significance. In the early years of the Soviet Union, and among some Communist Party activists in the earliest stages of establishing a Communist system, the goal of building a classless society was of motivational and inspirational significance. Thus, if the first three defining characteristics of Communism are also the three key structural features of Communist systems, the goal of building communism was primarily of doctrinal importance. It had an important place in Leninist ideology and it was this goal-oriented vision which distinguished Communist systems, not least the Soviet one, from societies in which governments were formed by democratic socialist parties whose aspirations did not go beyond enhancing fairness and reducing social inequality. If in politics there is, in Oakeshott's (1962) words, 'neither starting-place nor appointed destination,' a party could hardly claim the right to rule on the grounds that it knew how to guide society to the ultimate goal. It was, however, because Communist theoreticians claimed that there was an appointed destination—that of communism—that they could justify the permanent exercise of the leading role of the Communist Party. It was, Lenin himself had argued, that party alone which possessed the theoretical insight and political experience to guide less advanced citizens to this harmonious, classless society.

A fifth defining feature of Communism was the *existence of, and sense of belonging to, an international Communist movement*. For the Soviet Union, leading that movement was a matter of profound ideological and practical significance. It helped to underpin the legitimacy of the Soviet state in the eyes of many of its inhabitants and was also an instrument of state power. Many foreign Communists, especially in the period before the Twentieth Congress of the Soviet Communist Party in 1956 when Nikita Khrushchev (1894–1971) revealed some of the crimes of Stalin, (most notably, the mass arrests and executions of communists between 1936 and 1938), believed that a genuinely new society was being built in Russia and were ready to follow the twists and turns of the 'general line' emanating from Moscow. The first cracks in the unity of the international Communist movement occurred when Tito's Yugoslavia broke with the Soviet Union. The Yugoslav Communists were denounced by the Cominform, the successor to the Comintern as the Soviet-dominated organization of international Communism, in 1948. Vastly more encouraging for adherents of Communism had been the coming to power of the Chinese Communists in 1949. Yet, notwithstanding significant contributions by Soviet engineers and other specialists to China's development in the 1950s, the relationship between the two most populous of Communist states was to turn sour by the end of that decade. The next major breach within the international Communist movement after that between Moscow and Belgrade turned out to be the Sino-Soviet dispute. There were important ideological differences in which the Chinese were critical of what they saw as Soviet 'revisionism' (of Marxism–Leninism), but also resentment on the part of China's Communists of the Soviet leadership's taking for granted that the USSR was 'more equal than others' within the international Communist movement. The overwhelming majority of Communist states took the Soviet side against Mao's China, as most of them were essentially client states of the USSR. However, international Communism could never be the same again after a country of the size and importance of China had broken with the first Communist state and had called into question the Soviet 'leading role' within that movement.

Nevertheless, there remained the 'Soviet bloc' in Eastern Europe. The Albanian and Yugoslav Communists had made their own revolutions and there were no Soviet troops in Czechoslovakia when the Czech Communists seized full power in February

1948. In most cases, however, the European Communist systems owed their existence to Soviet force of arms. These countries had been liberated from Nazi occupation only to have their freedom of political choice rapidly curtailed as local Communists were brought to power and subsequently kept there by Soviet willingness to support their rule by further armed intervention, if necessary. The Soviet invasion of Hungary in 1956 and of Czechoslovakia in 1968 put an end to political changes which were backed both by reformist Communists and by the majority of the populations of these countries. The crushing of the 'Prague Spring' defined the limits of reform for East Europeans for a further twenty years.

Soviet leaders and their counterparts in Eastern Europe in one sense recognized the fragility of their regimes inasmuch as they devoted vast resources to surveillance of dissident groups and to maintaining as rigorous a censorship as was possible. They could not, however, prevent ideas from crossing national boundaries, even by jamming foreign radio (as they did for most of the post-Second World War Communist era). Communist political elites, in some countries more than others, made pragmatic adjustments to doctrine, while all the time finding suitable quotations from Lenin by way of justification.

In response to the relative failure of Communist economic development—as the Soviet Union and the East European countries not only fell further behind the most advanced Western economies but failed to keep pace with the newly-industrialized countries of Asia in an era of technological revolution—the USSR, with the coming to power of a leader a generation younger than his immediate predecessors, Mikhail Gorbachev (1931–), embarked on a reformist course from 1985. Although the initial impetus to change was the declining economic growth rate, Gorbachev was soon giving priority to political reform over economic reform and radically revising Soviet foreign policy in ways which led to the end of the Cold War. By, however, removing the threat of armed intervention which had hung over the East European states for decades, he left the Communist regimes there to stand or fall on the basis of winning the support of their own peoples. This they manifestly failed to do, and in the course of 1989 virtually all the Central and East European countries which had been under Communist rule became independent and non-Communist. The unification of Germany and the end of the German Democratic Republic followed a year later.

The almost entirely peaceful ending of Communism in Eastern Europe had serious implications for the Soviet Union itself. It emboldened those who sought national independence in the most politically alienated of Soviet republics, especially the Baltic states, and the salience of the national question, on top of economic problems and the political tensions engendered within an increasingly divided Soviet elite, threatened not only the Soviet system but the survival of the Soviet state. Indeed, Gorbachev's initial aim of reforming the system had given way in 1988–89 to more far-reaching change from above, amounting to systemic transformation. The collapse in December 1991 of the Soviet state, as distinct from the dismantling of the system, was, however, very much an unintended consequence of radical reform. Two-way contagion between pluralistic reform in the Soviet Union and democratizing change in Eastern Europe and between the assertion of national independence in East-Central European states and in the Soviet Union's republics played an important part in producing the collapse of the USSR.

Personal political rivalries were also of significance. Boris Yeltsin (1931–) played the Russian card against the Union, partly in order to get Gorbachev out of the Kremlin and himself inside as the unrivalled President of Russia. Hard-liners within the Soviet leadership attempted to reverse the Gorbachev reforms, and especially his Union Treaty that would have devolved great powers to the republics, and mounted an attempted coup in August 1991. It began with the placing of Gorbachev under house arrest in his Crimean holiday home and ended a few days later with the arrest of the Chairman of the KGB, the Minister of Defence, the Vice-President and their collaborators. It also had the unintended consequence of hastening the breakup of the USSR which the putsch had been mounted to prevent.

In 1985 there were 23 Communist states in the world. A decade later there were at most five—China, Cuba, Laos, North Korea and Vietnam. Yet these states no longer had all the attributes of Communist systems outlined earlier. In particular, there was no international Communist movement worthy of the name, nor was there even much lip-service to the building of communism. Moreover, in the most important of these countries, China, great concessions had been made to the market, and allegiance to Marxism–Leninism was becoming increasingly tenuous. To the extent that these states retained, in their very different ways, enough features of Communism to be so defined, then at the beginning of the twenty-first century almost a quarter of the world's population lived under Communist rule. But the fact that China has by far the largest population of any country means that approximate figure disguises the extent and political significance of Communism's decline and fall. If idealists and professional revolutionaries could in the nineteenth century, and during much of the twentieth, believe that this was humankind's path to a more radiant future, such illusions at the start of a new millennium had been confined to what Marx called 'the dustbin of history.'

See also: Communist Parties; Dictatorship; Eastern European Studies: History; Eastern European Studies: Politics; Eastern European Studies: Society; Marx,

Karl (1818–89); Marxism/Leninism; Russian Revolution, The; Scientific Revolution: History and Sociology; Socialist Societies: Anthropological Aspects; Soviet Studies: Politics; Soviet Studies: Society; Totalitarianism

Bibliography

Brown A 1999 The study of totalitarianism and authoritarianism. In: Hayward J, Barry B, Brown A (eds.) *The British Study of Politics in the Twentieth Century*. Oxford University Press, Oxford, UK, pp. 345–394

Brown J F 1988 *Eastern Europe and Communist Rule*. Duke University Press, Durham, NC

Dallin A, Lapidus G W (eds.) 1995 *The Soviet System: From Crisis to Collapse*, revised edn. Westview, Boulder, CO

Harding N 1996 *Leninism*. Macmillan, London

Holmes L 1986 *Politics in the Communist World*. Clarendon Press, Oxford, UK

Linz J J 2000 *Totalitarian and Authoritarian Regimes*. Lynne Rienner, Boulder, CO

Marx K, Engels F 1848/1962 McLellan D (ed.) *Communist Manifesto*. Oxford University Press, Oxford, UK

Oakeshott M 1962 Political education. In: Oakeshott M (ed.) *Rationalism in Politics and Other Essays*. Methuen, London

Plamenatz J 1954 *German Marxism and Russian Communism*. Longmans, London

Polan A J 1984 *Lenin and the End of Politics*. Methuen, London

Simons W B, White S (eds.) 1984 *The Party Statutes of the Communist World*. Nijhoff, The Hague, Netherlands

A. Brown

Communism, History of

The notion of communism has been used in three different ways. First, since antiquity communism has been referred to ideas relating to advanced forms of socialism advocating the need to organize societies as communities owning means of production and consumption. Second, the term communism signifies political movements with 'bolshevism,' an extreme faction of Russian social democracy, as their protagonist since 1903. Third, communism denotes a political and socioeconomic system of state introduced for the first time in Russia in the year 1917. Following the three different interpretations of the word communism, researchers concentrated their work on:

(a) the idea of social and political thought,

(b) state political systems with particular emphasis on dictatorship (totalitarian ideas), and

(c) economic systems, especially nonmarket economies, planned economies and autarchies.

Research on communist political movements and systems of state has been complicated seriously due to changes in time and space since both have existed for almost 100 years and have manifested themselves in different cultural zones of the world, most intensively, however, in Europe and East Asia.

1. The Idea

The ideas of communism have surfaced since antiquity and have been present throughout the Middle Ages, the Renaissance, the Enlightenment, and the French Revolution. The proponents of communism resorted to moral or economic arguments advocating the need to organize societies as communities owning means of production and consumption. These ideas were adopted by some socialists in the nineteenth century. Marx and Engels published their '*Communist Manifesto*' (1848). To substantiate his ideas Engels drew on the then studies on primitive societies as representative of classless communities (the natural state of society was disrupted by the division into social classes). However, the notions of socialism and communism were sometimes used interchangeably in the nineteenth century. When social democracy developed in many countries, socialists hardly ever used the notion of communism, whereas anarchists resorted to it occasionally (Kropotkin's anarchocommunism).

Communist ideas have acquired a new meaning since 1918. They became equivalent to the ideas of Marxism–Leninism, that is, the interpretation of Marxism by Lenin and his sucessors. Endorsing the final objective, namely the creation of a community owning means of production and providing each of its participants with consumption 'according to their needs,' they put forward the recognition of the class struggle as a dominating principle of a social development. In addition, workers (i.e., the proletariat) were to carry out the mission of the reconstruction of the society. Conducting a socialist revolution headed by the *avant-garde* of the proletariat, that is, the party, was hailed to be a historical necessity. Moreover, the introduction of the proletariat dictatorship was advocated and hostile classes were to be liquidated. In the following decades communist ideas have been subject to dogmatizing, on the one hand, and, on the other, to current political pressures. Paramount importance was attributed to the 'classics,' namely to Marx, Engels, and Lenin, as well as Stalin during his rule. It was only the later 'classics' (i.e., Lenin and Stalin) who were believed to be entitled to correct and amend the views of the earlier 'classics.' The communist ideology was expanded to include new components such as an exemplary role of the USSR and 'the principle of exacerbating the class struggle parallel to the building of socialism' (Stalin), which was called into question after 1953. In the years to follow the idea of a peaceful coexistence of socialism and capitalism as well as that of a peaceful transition to socialism were added. Some differences have surfaced between communist ideas endorsed in the USSR and the Soviet

bloc, in Yugoslavia, where social and economic aspects of self-government were regarded as significant features of the communist ideology, and in China, where Maoism became a revolutionary ideology of the suppressed 'Third World' directed against capitalist metropolics (Harrison 1972) as well as of the 'rural guerrilla revolution' (Cohen 1964). Some West European countries endorsed the ideology of Eurocommunism, i.e. a transition to socialism, which complied with all rules of democracy and renounced the dictatorship of the proletariat. On the whole, the ideology gradually lost its momentum in the countries governed by communists, particularly in the Soviet bloc. A much greater significance was attached to this ideology by groups opposing the Soviet-style communism. In the late 1960s some milieus of 'the new left' in Europe and America formulated the ideas of communist spontaneity, which were in opposition to capitalism and bureaucratic communism.

Many researchers have treated communist ideas mostly as one of the interpretations of Marxism. Popper (1945) wrote about the division of Marxists into radical (communist) and moderate (social democratic) wings. Walicki (1995) claimed that Leninism and Stalinism 'were dominant forms of the Marxist ideology of the 20th century.' In his work on Marxism Kolakowski (1978) handled the ideas of Leninism as one of the interpretations of Marxism, yet in some areas he saw them as a departure from Marx's original views. An even clearer division of communist ideas from Marxism was drawn by Furet (1995). Although he did not deny that these ideas had originated from Marxism, Furet pointed to other sources of origin such as Russian populist ideas.

2. *The Political Movement*

When in 1903 the Russian Social Democracy split into two groups, the majority faction was called 'Bolsheviks' and the minority 'Mensheviks.' The Bolsheviks in 1912 formed a separate party, an organization of active revolutionaries. Although the principle of democratic centralism was introduced, in practice the system was centralist and devoid of any traces of democracy. Furthermore, violence was endorsed as a fundamental means of political activity. All these features could be accounted for by the situation in Russia, where the parliamentary system was in its infancy and despotism ruled unabated. The Bolsheviks came to power in Russia in 1917, whereas in 1918 they transformed themselves into a communist party. Since 1918 communist parties emerged also in other countries, mostly as a result of a split within socialist parties. In 1919 the Communist International (Comintern) was set up and it soon became centralized and dependent, as a matter of fact, on the Russian (or later Soviet) party. The terms 'communism' and 'bolshevism' were still used interchangeably and 'bolshevisation' came to denote the process of following suit and imitating the example of the Russian party.

Communist political movements manifested themselves most intensively in Europe and in China. They kept changing their political profiles adopting to the then current situation on the international arena (e.g., under the pressure of the 1929–1933 crisis), but, above all, complying with interests of the Soviet Union. In the 1930s this compliance was linked with Stalin assuming the role of the undisputed leader of the international communist movement. Opposition groups were excluded from the movement. One of the few groupings which remained active were the proponents of Leo Trotsky, a prominent party leader expelled from the Soviet Union. However, in the late 1930s, when faced with the threat of fascism, communist parties started to endorse the slogans of parliamentary democracy and entered into cooperation with socialists and even nonsocialist democratic parties (the 'People's Front').

When the Soviet Union entered into an agreement with Germany in 1939, communist parties all over the world declared their neutrality towards both parties in World War II. This policy was subject to a drastic change following the 1941 German invasion of the Soviet Union. Only then did communist parties come to regard the struggle against fascism as their priority and fostered the idea of forming the National Front. The idea, however, was rarely put into practice for several reasons including dependence of communists on the Soviet Union (although Comintern was dissolved in 1943) and their aspirations to play a dominant role in the resistance movement.

When World War II was drawing to an end communist parties took over power in the countries in which the Soviet army was present as well as in Yugoslavia and Albania. In the countries occupied by the armies of the Western allies, communists frequently entered into ruling coalitions. With the escalation of conflict between two political blocs and an upsurge of the 'cold war,' communist parties in the West went into opposition. The Soviet control over the international communist movement was strengthened anew in 1947 when the Information Bureau of Communist and Workers' Parties (Cominform) was established comprising the largest European parties. A year later the first serious conflict between parties came to the fore, when the Yugoslav party was condemned by Cominform. The victory of the communist party in China in 1949 (and later in Vietnam and Cuba) laid foundations for subsequent decentralization of the communist movement.

The process of decentralization took place gradually following the death of Stalin in 1953, who three years later was accused of great atrocities. The escalation of the crisis in 1956 brought about the dissolution of Cominform, and resulted in reforms in Poland and a revolution in Hungary. The 1960s witnessed a conflict between the Soviet and the Chinese parties

during which some other parties attempted to take advantage of the situation and gain independence. Yet another crisis, triggered by the intervention of the Soviet bloc in Czechoslovakia, resulted in the adoption of Eurocommunist ideas by some West European parties in the 1970s. At the same time, small communist groupings emerged in the West drawing on the experiences (or rather their own visions of the experiences) of China, Vietnam, Cuba, or Yugoslavia. Further crises in the international communist movement were linked to the Polish conflicts of 1980–1981. In the 1980s both ruling and opposition communist parties showed symptoms of decay. After the collapse of communist systems in Europe nearly all communist parties rejected their former ideological and organizational principles and transformed into socialist or social democratic parties.

In the countries ruled by communists a considerable number of works on the communist movement was published, yet their content was subject to control. The most important publications had even an official character (*The History of the Communist Party* 1938). Little changed in this respect in the communist historiography after the 'destalinization' although the party ceased to stress the official character of publications, relying heavily on the censorship and monitoring of publications. These publications rarely contained reliable information, featuring mainly interpretations devoid of any critical approach and containing hagiographic accents.

Many publications were written by representatives of the communist opposition, who embraced the fundamentals of communism yet criticized the political line of Stalin and his successors, as it was in the case of the Trotskyists. Some of these texts resembled political commentaries and contained different information not included in the official and semi-official communist literature. This trend comprised also publications based on extensive studies (e.g., Deutscher 1949, 1954, 1959, 1963).

Works by noncommunist authors fall into yet another category of literature relating to the history of the working-class movement. Although this literature contained fewer political commentaries, it demonstrated a strong socialist bias. From a left-wing socialist perspective, communism, in spite of all its shortcomings, was perceived as a proletariat movement close to socialist ideas and as a potential ally in the socialists' struggle against the bourgeoisie (e.g., Braunthal 1960). Right-wing socialists, on the other hand, saw it as a hostile force striving to exert influence over the same social class. In recent decades, however, the history of the workers' movement encompassing the history of the communist movement has become subject to more academic research.

Amongst authors who dealt with the communist movement there were some who came from within the party ranks as well as those who broke free from its legacy (Borkenau 1938, 1953). Others were emigrants from countries governed by communists, who drew on their life experience and language skills (Fejtö 1974). One should also consider here expert publications of academic institutions which dealt with communism as a sideline political movement within a larger scope of interests relating to the so-called Sovietology.

3. *The Political and Socioeconomic System*

Communists themselves are more prone to call it a socialist system. For the first time, such a system was introduced in Russia after 1917. The principles of it were outlined by Lenin before the victory of the revolution, yet his work on this subject was published only when the Bolsheviks came to power (*Gosudarstwo i rewoljucija* 1918). Two principles were of particular significance: (a) the state has to resort to violence in the political and social life, and (b) elements of a centralized bureaucracy, that is, record keeping and monitoring, play a key role in the organization of the economy.

In the Soviet political system the leading role was assigned to the centralized communist monoparty devoid of any traces of democracy within its ranks. All organizations were subordinated to the party and headed by its members. Moreover, the party exercised control over all publications and mass media directly or through state departments (e.g., censorship).

The state authorities at all levels reported to party organs. Initially the state authorities were organized in a hierarchical structure of councils with some chosen representatives of lower level sitting on the councils of a higher level. Following the establishment of the Soviet Union in 1922, the party and state authorities were to function according to the principles of a federal state. The 1936 constitution provided for general elections at all tiers of power. In practice, however, soon after communists came to power they introduced the principle of nominating from above, a principle that applied to both the party and state authorities. The approval of such nominations in public elections was a mere formality.

This political system was adopted after World War II in all communist countries. In the Soviet bloc, which included European communist countries (with the exception of Yugoslavia since 1948 and Albania since the 1960s) as well as Mongolia and North Korea (for a couple of years), there were initially informal methods of control from Moscow. System monitoring was exercised through the Cominform and Soviet advisers or directives sent to high-ranking officials. Since 1956 methods of control were restricted and formalized by means of political and military structures of the Warsaw Pact, as well as structures of the Council for Mutual Economic Assistance (COMECON). The remaining communist countries enjoyed a full or partial independence. Yugoslavia carried out some transformations in its political system, introducing

elements of a limited federalism and a local government.

Gradually the centralized systems began to crumble on international and state levels alike. Retaining its leading role, the USSR had to take into account more and more frequently the positions of the countries within the Soviet bloc. Following the crisis in Czechoslovakia and the invasion of the country by the Soviet bloc, the Brezhnev doctrine was proclaimed (1968) asserting the right of the communist community to intervene when any of its members demonstrates 'anti-socialist degeneration.' Several countries of the Soviet Block and China managed to hand down some power to regional levels.

A considerable autonomy was enjoyed by the police, and above all by security services. They gained a strong position in the period of widespread repression in the Soviet Union (collectivization and 'big cleansing' in the 1930s) as well as in the countries within the Soviet sphere of influence (in the late 1940s and early 1950s) and in China (during the cultural revolution). Furthermore, security services attempted to exert a still greater political influence also in other periods.

Another pivotal element of the system was the army. In the Soviet Union it owed its position to the role played in introducing the system and in World War II. Later it was a key means supporting the rivalry between superpowers and political blocs as well as the Soviet expansion in the countries of the Third World. Similarly, in China and Vietnam, the army was the co-creator of the system. In Poland the army won a particular importance in the face of the imposition of martial law in 1981. On the other hand, party officials feared an excessive degree of independence of the army, which had far-reaching repercussions in the Soviet Union during repressions directed towards army officials in the 1930s.

Since 1953 several communist countries, particularly those belonging to the Soviet bloc, witnessed recurring political crises. They were triggered by people's rising democratic aspirations influenced by examples emanating from Western democracies, by the lack of respect for national sovereignty and, to a large extent, by a growing economic crisis and a low standard of living. Another precipitating factor was the disillusionment as to the application of communist ideas in practice. Communist officials adopted the formula of 'real socialism' as a form of the existing system hindered by historical limitations on the road to the socialist and communist future.

The economy in communist countries was subordinated to politics. Nationalization was implemented in all spheres of economic life. Although agriculture and, to some degree, trade and small production were formally owned by cooperatives, they were supervised by the party and state administration. There were only negligible relics of private ownership with the exception of Poland and Yugoslavia where agriculture remained in private hands.

Centralized state planning functioned on the basis of 5-year plans, the allocation of raw materials and means of investments, regulated prices, and rationing commodities for sale. A top priority was the needs of the state, which resulted often from an international policy. A particular emphasis was placed on the development of heavy industry and reservoirs of raw materials, neglecting other branches of production. In the face of constant shortages and a crippled economy, many goods intended for production and consumption were subject to different forms of rationing. The state's monopoly on foreign trade along with slow technological progress, a limited exports potential and, consequently, import deficit led to a form of economic autarchy. Yugoslavia's economy demonstrated a slight divergence from this general pattern after the introduction of self-government in enterprises. Moreover, Yugoslavia fostered more extensively mechanisms of the market economy and managed to avoid autarchy. Nonetheless, it put a limit on foreign investments and failed to develop technological innovations.

Communist systems which were created almost exclusively in poorly developed or underdeveloped countries, demonstrated initially a high rate of economic growth. The rapid industrialization combined with urban development was based on extensive factors such as a cheap labor force. However, as these factors were gradually exhausted, the growth rate began to decline and the standard of living stabilized at a low level. Among other factors which aggravated the crisis were the growing costs of the arms race and political rivalry between superpowers and blocs.

From the late 1950s attempts were made to overcome the crisis. On the one hand, restraints of the arms race were accepted (the Helsinki Agreement of 1975, treaties on nonproliferation of nuclear and conventional weapons) and, on the other hand, planned economies were transformed slightly by introducing in them elements of a market economy and stimuli of a technological progress. In spite of a successful arms race reduction the costs of armaments and political expansion still exceeded the economic potential of communist countries, whereas the introduction of elements of market economy generated strong resistance of the bureaucracy in the state and the party administration.

The 1980s saw not only a growing opposition against the political system but also an ever stronger conviction of an upcoming defeat shared by communist officials. The need to conduct economic reforms called for a more extensive popular support, which was possible to achieve thanks to political reforms. Political reforms were pursued to a limited extent in the Soviet Union (economic *perestoyka*) accompanied by political *glasnost* – transparency) and in Poland (the 'round table' talks, an attempt to achieve a compromise with the opposition). Finally, the attempts to carry out limited reforms of the

economic system were doomed to failure. Coupled with a disintegrated political system, they led to the collapse of communism over the years 1989–1991.

In China, however, the authorities resorted to extreme means of repression against the opposition (the massacre at the Tian'anmen square, 1989). They were then successful in effecting its economic reforms encompassing devolution; privatization of agriculture, small- and medium-size industry, and trade; as well as the introduction of elements of market economy. The country managed to restore a fast rate of economic growth and development creating openings for foreign investments linked with the import of new technologies. The future will show the results of these reforms.

Communist researchers into the political and economic system concentrated their work mostly on its premises rather than on reality. A distinct character was noticeable only in the discussions before the 1930s which touched upon the 'Asiatic system of production.' These works tackled directly the prospects of building of the communist system in China, and, indirectly, they reflected some ideas on the political and economic system in Russia. Amongst basic characteristics of oriental societies they featured a dominating role of the state in the economy and, consequently, the durability of absolutist and bureaucratic tendencies. This concept was raised in discussions in the 1920s by Wittfogel. He broke away later with communism and came to regard the Soviet Union as a follower of Asian despotic systems (1962).

Later, the communist authors emphasized the class conditioning of the system. Many works presented a mystified picture of common interests of the state and society. They referred also to the notions of 'People's Democracy' or 'Socialist Democracy' (as a state of equal opportunities). Communist publications are of little academic value, yet sometimes they can communicate significant information.

In the noncommunist literature as a whole there were prominent publications which dealt with communist systems within a wider historical framework. In particular, they referred to the connections between the history of Russia and of the Soviet Union. The principle of a historical continuity was propagated (Kucharzewski 1948, Pipes 1990, 1993, Besançon 1980). This conception was often accompanied by a tendency to view communist systems as something superimposed by the USSR upon other countries.

The notion of totalitarianism was used initially in Italian fascism, later by researchers whose works dealt with German national socialism. In the wake of experiences and the atmosphere of the 'cold war', communist systems were regarded by some researchers as one of the forms of totalitarianism (Arendt 1951, Friedrich and Brzeziński 1956). Arendt saw the roots of totalitarian systems above all in the current demise of human values and the crisis of principles underlying the existence of societies and individuals. Friedrich and Brzeziński focused mainly on the techniques of wielding power and the totalitarian syndrome.

Concrete studies of the system were heavily influenced by the concept of totalitarianism (Fainsod 1953). The concept of communism as a form of totalitarian systems was undermined to some extent as a result of growing tendencies towards decentralization and curbing repression. Researchers started to concentrate more on the differences between communist countries as well as on the fact of how inconsistent their respective systems were (Brzeziński 1960). Another approach to communism stipulated that the system was an attempt to implement utopia invented by doctrinaires (Heller and Niekricz 1982, Malia 1994). There were also other interpretations of the communist system according to which it was a peculiar way to modernization concerning in particular the Third World (Schurmann 1966).

Some authors treated communism above all as a system of genocide. Political commentaries on the subject emerged before World War II. Afterwards, scientific studies were preceded by documentary literature (Herling-Grudziński 1951, Solzhenitsyn 1974). A different interpretation claimed that Stalinism was only a criminal distortion of the communist system (Medvedev 1971). In the 1960s, Conquest (1968) published monographs on various aspects of genocide in the Soviet Union. Most recently an international team has tried to apply a comprehensive approach to this problem (Courtois et al. 1997).

At present, communist systems exist exclusively in the Far East and in Cuba. These areas require further studies devoted to their specific character and ability to stand the test of time.

See also: China: Sociocultural Aspects; Cold War, The; Communism; Communist Parties; Dictatorship; Dictatorship in History; Eastern European Studies: History; Eastern European Studies: Politics; First World War, The; Genocide: Historical Aspects; Japan: Sociocultural Aspects; Marx, Karl (1818–89); Revolutions of 1989–90 in Eastern Central Europe; Russian Revolution, The; Socialism; Socialism: Historical Aspects; Socialist Societies: Anthropological Aspects; Soviet Studies: Economics; Soviet Studies: Politics; Soviet Studies: Society; Totalitarianism

Bibliography

Arendt H 1951 *The Origins of Totalitarianism*, 1st edn. Harcourt Brace, New York

Besançon A 1980 *Present sovietique et passé russe*. Librairic Générale Française, Paris

Borkenau F 1938 *The Communist International*. Faber & Faber, London

Borkenau F 1953 *European Communism*. Faber and Faber, London

Braunthal J 1960 *Geschichte der Internationale*. Verlag J.H.W. Dietz Nachf., Berlin and Bonn

Brzeziński Z 1960 *The Soviet Block—Unity and Conflict*. Harvard University Press, Cambridge, MA
Carrĕre d'Encausse H 1981a *Lenin: Revolution and Power*. Longman, London
Carrĕre d'Encausse H 1981b *Stalin: Order through Terror*. Longman, London
Cohen A A 1964 *The Communism of Mao-Tse-tung*. The University of Chicago Press, Chicago
Conquest R 1968 *The Great Terror*. McMillan, London
Courtois S, Werth N, Panné J-L, Paczkowski A, Bartosek K, Margolin J-L 1997 *Le livre noir du communisme. Crimes, terreur et répression*. Editions Robert Laffont, Paris
Deutscher J 1949 *Stalin. A Political Biography*. Oxford University Press, Oxford, UK
Deutscher J 1954 *The Prophet Armed: Trotsky 1879–1921*. Oxford University Press, Oxford, UK
Deutscher J 1959 *The Prophet Unarmed: Trotsky 1921–1929*. Oxford University Press, Oxford, UK
Deutscher J 1963 *The Prophet Outcast: Trotsky 1929–1940*. Oxford University Press, Oxford, UK
Fainsod M 1953 *How Russia is Ruled*. Harvard University Press, Cambridge, MA
Fejtö F 1974 *History of the People's Democracies. Eastern Europe since Stalin*. Penguin, Harmondsworth, UK
Friedrich C, Brzeziński Z K 1956 *Totalitarian Dictatorship and Autocracy*. Harvard University Press, Cambridge, MA
Furet F 1995 *Le passé d'une illusion: Essai sur I'idée communiste au XXe siècle*. Éditions Laffont, Paris
Harrison J P 1972 *The Long March to Power. A History of the Chinese Communist Party, 1921–1972*. Praeger, New York
Heller M, Niekricz A 1982 *Utopija u wlasti. Istorija Sowietskogo Sojuza z 1917 goda do naszych dnicj*. Overseas Publications Interchange Ltd., London
Herling-Grudziński G 1951 *A World Apart*. Heinemann, London
Kolakowski L 1978 *Main Currents of Marxism. Its Rise, Growth and Dissolution*, Vols. 1–3. Clarendon Press, Oxford, UK
Kucharzewski J 1948 *The Origins of Modern Russia*. Polish Institute of Art and Sciences, New York
Malia M 1994 *The Soviet Tragedy. A History of Socialism in Russia 1917–1991*. Free Press, New York
Medvedev R 1971 Let History Judge. Alfred A. Knopf, New York
Pipes R 1990 *The Russian Revolution*, 1st edn. Alfred A. Knopf, New York
Pipes R. 1993 *Russia under the Bolshevik Regime*, 1st edn. Alfred A. Knopf, New York
Popper K R 1945 *The Open Society and its Enemies*. Routledge and Sons, London
Schapiro L 1965 *The Goverment and Politics of the Soviet Union*. Random House, New York
Schurmann F 1966 *Ideology and Organization in Communist China*. University of California Press, Berkeley, CA
Solzhenitsyn A I 1974 *The Gulag Archipelago, 1918–1956: An Experiment in Literacy Investigation*. Harper & Row, New York
Walicki A 1995 *Marxism and the Leap to the Kingdom of Freedom. The Rise and Fall of the Communist Utopia*. Stanford University Press, Stanford, CA
Wittfogel K A 1962 *Dic orientalische Despotie. Eine vergleichende Untersuchung totaler Macht*. Kiepenhauer und Witsch, Köln

J. Holzer

Communist Parties

Originating after the Russian Revolution of November 1917, communist parties developed into a distinctive organizational form with a characteristic political ideology and an integral personal identity for their members. Modeled deliberately on the example of the Russian original, those communist parties that had remained faithful to the Soviet Union generally either declined or underwent radical transformation when the Soviet state collapsed in 1991.

In March 1918 Vladimir Lenin urged his victorious 'Bolshevik' ('majoritarian') faction of the Russian social democratic party to reconstitute itself under the name 'All-Russian Communist Party (bolsheviks),' known by its Russian initials as the RKP(b) and later renamed communist party of the Soviet Union. By this act of renaming, Lenin meant to signify that the former Bolsheviks would oppose replacement of Tsarist rule with parliamentary social democracy in Russia. They would stand instead for establishment of a 'dictatorship of the proletariat' as a deliberately coercive device for prohibiting political activity in pursuit of any goal other than communism, defined by Karl Marx as a society combining contribution according to ability with consumption according to need.

A year later Lenin convened his supporters together with foreign sympathizers resident in Moscow and a few delegates from other socialist parties who managed to make their way to the Russian capital, where they established the 'Communist International,' or Comintern. At its Second Congress in August 1920, this organization's leaders laid down the '21 conditions' which gave definition to the identity of communist parties. The three most important conditions obligated communist parties to practice democratic centralism, organizational exclusivity, and defense of the Soviet Union. 'Democratic centralism' came to mean that in a communist party, all local organizations, and all individual members must obligate themselves to comply with directives of a national directorate. This 'centralist' principle was to work in combination with the 'democratic' principle that the national directorate (called a 'Central Committee') must be elected by delegates from the local organizations and accept information from them about their local activities. Because the centralist principle included the Central Committee's right to nominate the sole candidate for obligatory election by the local organization ('committee') as its leader, in fact democratic centralism conferred on the Central Committee the right to select its own electors.

Organizational exclusivity obligated the communist party to separate itself from any socialist party, prohibiting joint membership in the two parties. Accompanying this requirement was a demand that even where communist parties could operate legally, a

communist party must maintain a clandestine organization for illegal revolutionary activities.

All members of communist parties were further required 'to render every possible assistance' (Daniels 1984) to the Soviet state, especially against the policies of their own governments. Lenin's ultimate successor Joseph Stalin would in 1927 transform this expectation into a definition of communist internationalism as the readiness 'unreservedly, unhesitatingly and without conditions … to defend the USSR' (Stern 1990). By imposing internal discipline enforced by the right of central leaders to remove local opponents, an organizational split with other Marxists, and a Russia-centered conception of the party's mission, the 21 conditions required foreign communist Parties to reprise the organizational history of the Soviet exemplar. As communist Parties emerged around the world, encouraged both by the success of the Soviet Party in establishing Russia's independence from foreign domination and by clandestine monetary subsidies from the Soviet comrades, they became identifiable by their adherence to a common political ideology known as Marxism–Leninism. Of course from the very beginning Marxism–Leninism existed in many variants. The 21 conditions were themselves an effort to enforce a minimal degree of uniformity on diverse conceptions of communist identity. Adherence to the ideas of 'Marx, Engels, Lenin, and Trotsky' characterized the Trotskyists who soon broke off in a 'Fourth International' whose adherents still exist (winning 5.3 percent of the vote in the 1995 French presidential election). 'Marxism–Leninism–Maoism' became the ideology of the Chinese Communist Party and of the splinter parties that broke off from national communist parties after the Chinese definitively split with the Soviets in 1963. Italian communists continued to be influenced by the ideas of Antonio Gramsci, whose independent conception of the reasons why the working class in industrial countries remained politically quiescent bore far more democratic implications than Lenin's own explanation of worker passivity. Until Stalin's death, the Soviet Party referred to its own ideology as 'Marxism–Leninism–Stalinism.'

Despite this diversity, communist political thought has generally shared certain core elements. Communists have classed their own political thinking as 'theoretical' in contrast to the 'ideological' programs of other political parties. They have affirmed a theoretical postulate identifying political parties with the economic interest of classes. This postulate defines 'class' as joining all persons who make their living in the same way and 'class interest' as the perpetuation of that way of making a living (as opposed, for example, to increasing the incomes of the individuals who compose a class). The same postulate identifies each communist party with a 'proletariat,' as the class of persons who earn their living by selling ('alienating') their labor to someone else. Communists have generally insisted that in the long run the proletariat can only avoid the fate of mass unemployment if it succeeds in overthrowing capitalism, i.e., the system of buying and selling labor as a commodity. In the short run, state ownership of industrial enterprises has tended to be regarded as a means of maintaining the proletariat against capitalist pressure. Communists are further distinguished from other Marxists by a view that agricultural smallholders are likely to be promising allies in shortening the road to the abolition of capitalism. Finally, both the long-run goal of abolishing capitalism and various short-run interests of the proletariat have been considered achievable only by a 'vanguard party,' one which relies on theoretical analysis to identify the interests of the proletariat rather than consulting the proletarians themselves. Where they engage in electoral competition, communists have seen their task as educating voters in the voters' true interests rather than responding to expression of interest by voters. Where they have acquired control of the state, communists have portrayed their task as preventing other parties from deceiving the proletariat by running their own independent candidates.

Because of the commitments to democratic centralism and to the theoretical self-understanding as a vanguard party, communists could only be 'cadres,' although in a sense distinct from the usage of this term introduced into political science by Maurice Duverger. By a 'cadre party' Duverger meant a nineteenth-century association of notables, who drew their personal identities from birth or achievement and saw the party as a device to help win contests for office (Duverger 1963). To a communist, 'cadre' took its meaning from Lenin's conception of the party member as a full-time professional revolutionary. A communist cadre, at least ideally, has derived his or her personal identity solely from self-transformation into an agent of the party. Democratic centralism provides that every party member may immediately be redeployed, regardless of his or her personal will, to any place or duty at the discretion of the central party authority. This total commitment gave an integral, all-or-nothing quality to the personal identity of a communist. As Stalin once told the Soviet Central Committee, 'Either you are nothing in the eyes of the party or you are a party member with full rights' (Arch Getty and Naumov 1999). Ken Jowitt has argued that this integral conception of personal identity accounts for communist parties' difficulties in attracting adherents in industrial societies, contrary to Marx's expectation that only these societies were ripe for communist ideas, because the complexity of industrial society encourages people to develop identities contextualized to varying situations. Conversely communist parties were more successful in attracting adherents in agrarian societies where less social diversity and more rigid subordination limited the fragmentation of selfhood (Jowitt 1992). Integral identification of the self with the party makes more comprehensible both the ex-

ceptional violence practiced by many communist parties, in whose eyes the victims were 'nothing,' and the extraordinary selflessness displayed by communists in such movements as the World War II Resistance against German occupiers or struggles for civil rights and economic justice in many countries. Where communist parties managed to gain control of the state, it is questionable whether they remained 'parties' at all. Communists became the rulers of the Soviet Union, eight countries in Eastern Europe, China and five other Asian countries, and Cuba. If by a 'party' we mean an organization that competes for political power by supplying candidates for executive or legislative office and by maintaining a coalition among its candidates after they win, then it is questionable whether ruling communist parties qualified. Of course a monopolist may be said to compete, so successfully as to exclude all rivals. But ruling communist parties performed three functions not usually associated with political parties, acting as territorial administrations, personnel offices for their governments, and ultimate policy makers. In the Soviet Union, the first secretary heading each local party committee was the ultimate authority over all government activity in the territory assigned to the committee. The highest ranking local government official was a subordinate of the first secretary, sitting as a member of the committee's bureau acting under the first secretary's direction and likely, at a later stage in his career, to be promoted into the position of first secretary. The first secretary was expected to exercise his plenipotentiary authority by recommending appointment or dismissal of all officials within the territory, with the exception of those officials whose appointment or dismissal could be recommended only by some higher party committee. The Political Bureau ('Politburo') of the Central Committee in turn 'recommended' policies to the appropriate government agencies, enforcing compliance with its policy recommendations by exercising its right to making binding recommendations about appointment or dismissal of the relevant officials. Where communist parties did not secure control of the state, they also did not practice these methods of rule, but at the same time their overt identification with the practice of complete penetration of the state hampered their electoral campaigns by stimulating suspicion among voters accustomed to distinguish party activities from conduct of government.

The collapse or transformation of many communist parties after the fall of the Soviet Union in 1991 resulted from their close identification with the Soviet State. Decline was not solely the result of the disappearance of monetary subsidies on which many parties had subsisted and others had drawn. From the beginning communists had conceived themselves as those Marxists loyal to the Soviet Union and ready to accept its authenticity as the revolutionary workers' state. Communist parties had repeatedly lost members or suffered organizational splits whenever events had dramatized the discrepancy between the emancipatory ideals of Marxism and the practices of the Soviet rulers. Such events as executions of senior Soviet officials in 1937–8, the 1939 Molotov–Ribbentrop pact with Nazi Germany, Nikita Khrushchev's 1956 denunciation of Stalin's crimes against the party and the crushing of the Hungarian uprising that resulted, the invasion of Czechoslovakia in 1968, and the suppression of the Polish labor movement Solidarnosc in 1980–1 had impelled waves of resignations from communist parties around the world. The reciprocal denunciations in 1963 by the Chinese and Soviet parties led to formation of parallel Maoist parties in most countries where communist parties existed. While the reforms known as *perestroika* championed by Mikhail S. Gorbachev after 1985 briefly encouraged some foreign communists loyal to the Soviet party, popular uprisings against ruling parties in Eastern Europe in the fall of 1989 and the defeat of the August 1991 coup against Gorbachev led many foreign communist parties to reconceive their identity. Lenin had said that the adoption of the name 'communist' was far more than a formal act; in East European states formerly ruled by communist parties, remnants of those organizations hurried to relabel themselves 'socialist' or 'democratic socialists,' while the independent-minded Italian Communist Party preferred 'Democratic Party of the Left.' Although the pattern of decline was mixed, with the independent Communist Party of Spain retaining its name and gaining adherents, the once powerful French Communist Party, noted for its loyalty to Moscow, found its support among voters dwindling, with former supporters deserting to the Socialists or even to the far-right National Front. Just as communist parties as a distinctive form of political organization had originated with the Soviet State and developed their identifying features by emulating Russian forms, so the collapse of the Soviet order ended the cohesiveness of communist parties as an identifiable phenomenon.

See also: Communism; Communism, History of; Dictatorship; Eastern European Studies: History; Eastern European Studies: Politics; Marx, Karl (1818–89); Marxism/Leninism; Party Systems; Planning, Politics of; Political Parties, History of; Socialism; Socialist Societies: Anthropological Aspects; Soviet Studies: Politics

Bibliography

Arch Getty J, Naumov O V 1999 *The Road to Terror: Stalin and the Self-Destruction of the Bolsheviks, 1932–1939*, trans. Sher B, Yale University Press, New Haven, CT, p. 233

Daniels R V (ed.) 1984 *A Documentary History of Communism, Vol. 2: 'Communism and the World'*. University Press of New England, Hanover, VT, p. 46

Duverger M 1963 *Political Parties: Their Organization and Activity in the Modern State*, 3rd edn. Wiley, New York, pp. 63–4
Jowitt K 1992 *New World Disorder: The Leninist Extinction* X. University of California, New Haven, CT, esp. pp. 67–87, 125
Stern G 1990 *The Rise and Decline of International Communism*. Edward Elgar, Brookfield, VT, p. 71

R. Anderson

Communitarianism: Political Theory

Characteristic key elements of communitarian theory are the following five points: a hermeneutic methodology, criticism of what was called atomistic individualism, an Aristotelian concept of the good in opposition to rights, the renaissance of the idea of community, and a renewed insight into the relevant political function of civic virtues. Communities can be defined as webs of social relations that encompass shared meanings and shared values. Families may qualify as minicommunities, villages often are, although not necessarily. Communities need not be concentrated geographically, like the scientific community. Especially traditional communities are not necessarily places of virtue, but they can be oppressive, monolithic, and authoritarian. So a main problem for communitarian authors is to find overarching criteria to avoid accepting automatically any commonality of shared self-understanding as a good community (Etzioni 1998, p. XIV).

In the 1980s, a group of political theoreticians—Charles Taylor, Michael Sandel, and Michael Walzer—challenged what they perceived as excessive individualism by the insight that individuality is based and dependent on social context. The radical liberal idea of an 'unencumbered self' was criticized as self-refuting, because it fostered antisocial behavior with the tendency to undermine its own social preconditions.

The academic and public impact of communitarianism has come about for two reasons: it seemed to give a theoretical answer to the most important political philosophy of the time, for example, the individualist liberalism of the renewed contract theory of John Rawls and others. The second reason is even more important: politically, communitarianism was a critical reaction against the 'liberalism of greed' of the Thatcher and Reagan years, so it became attractive to the New Democrats and New Labour as a rhetoric to win back the votes of the middle classes. In the early 1990s, Etzioni and others decided to take these ideas from the campus to the larger society. They founded the journal *The Responsive Community* and tried to establish communitarian ideas as a recognized third way of political thinking between liberalism and conservatism. Different from former communitarian ideas in the tradition of Ferdinand Tönnies (*Gemeinschaft und Gesellschaft*), Robert Nisbet, and Martin Buber, the new communitarians pointed out that a responsive community tries to avoid any authoritarianism and oppressiveness against the individual. The old communitarians privileged social bonds and traditional small communities over the individual. Communities in their definition were villages, small cities, some religious sects, and tribes. Asian communitarianism overemphasizes social harmony. The new American line of communitarian thinking concentrated much more on the question of balance between individual rights and social responsibilities. Individual autonomy and the common good are to be brought into a new equilibrium. It tried to add a sense of the social and communal preconditions of social integration to rights-centered liberalism and liberal neutrality.

The anthropologic key concept of modern communitarian thinking is Taylor's idea of humans as self-interpreting animals (Taylor 1985b). Modern identity is based on inescapable frameworks of interpretation, which provides the background language for strong evaluations of right and wrong, good and evil, higher and lower. The quality of this framework is decisive for the quality of the society. Sandel radicalized this idea and transformed it into an argument against Rawlsian liberalism: the original position as a metaphor for justice, where nobody knows about his or her later social position before a decision is taken, is not only practically impossible (Rawls would agree), but also theoretically undesirable, because this conception of justice would lead to a world inhabited by morally neutral people without any character. Persons without loyalties to their family, their community, nation, or people are in Sandel's view persons without character. MacIntyre emphasized the missing foundation of liberal theories of justice. Even more, modern concepts cannot provide any motivation to act morally, because they are not based on fully responsible persons, but only on character masks such as the manager, the therapist or the Kierkegaardian aesthete.

This anti-Rawlsian criticism was not conclusive, because his principles of justice were not meant to shape the whole person, but only their behavior in the political realm. Even more, Sandel did not take into account the possibility of a person taking a distant stand from their community. So the foundations of Sandel's (and, as traditional virtues are concerned, MacIntyre's) criticism seemed to be rooted in old-fashioned communitarianism without reflection of the fact that modern American society can be interpreted as the social basis for some kind of responsive individualism. Both Sandel and MacIntyre were uncomfortable with the label 'communitarian,' but their ideas belong to the present discussion of the concept.

A quite different kind of communitarian theory was developed by Walzer, who is also uncomfortable with

the term, but refers to the same set of methods and ideas. In his *Spheres of Justice. A Defense of Pluralism and Equality* (1983) his social criticism and his ideas for reform lean on the hermeneutic interpretation of commonly shared values. The well-known tension between shared values and social practice is sufficient to legitimize social reform. Walzer's method differs on the one side from liberal Kantian or Rawlsian social constructionism, and on the other from conservative belief in the objectivity of traditional values (social ontology). Justice is only possible if it takes into account the differentiation of spheres. Different modes of distribution have to be applied concerning membership and citizenship, political power, security and welfare, money, office, education, recognition, kinship, and love. There is no general theory of justice, but a special justice for each of these spheres. What is considered as just depends on commonly shared and publicly discussed opinions about the criteria of justice.

The liberal-communitarian debate of the 1980s and early 1990s seems to be over at present. Those who decided to stick to the label 'Communitarians' now call themselves liberal communitarians or communitarian liberals, such as Selznick (1992) and Etzioni (1996). The central idea is that, in a modern society, common values and virtues can only be developed by freely consenting individuals, not by the state or by pressure from religious movements of the fundamentalist denomination.

Selznick's 'communitarian liberalism' draws on the social philosophy of John Dewey. Morality is always community based. Character and integrity are relevant for the function of large organizations, including the state. A proper understanding of communities implies diversity and pluralism as well as social integration. It has to have a strong sense of a common good. Thus, Selznick combines sociological functionalism with the value-based traditions of American pragmatism. His interpretation of the term 'covenant' means a self-defining commitment, and an act of faith and resolve by the individual. It is prepolitical, foundational, and consensual. Liberal neutrality misses this aspect, and is based only on institutions for the purpose of social integration. Conservatism, on the other hand, tries to overload institutional processes with value orientations. The communitarian third way tries to restore values and virtues as functional prerequisites of the civil society.

Bellah and co-workers (1991) try to describe the state of American society and to make statements about its long-term viability through insights into its mores and the climate of opinion. They lean on Alexis de Tocqueville's *Democracy in America* (1835), which seems to be a key text for most communitarians. The results of the study are communicated in terms of eloquent sadness: the self-destruction of the liberation promises of modernity and the retreat from public engagement into the private sphere of the family and friends is stated by many of the interviews. Most of them articulated a profound yearning for the idealized small town, which can be interpreted as a yearning for meaning and coherence, for commitment and community. The social ecology is not intact if there are no communal answers to the extreme fragmentation of modern everyday life. The ubiquitous language of individualism, in which American self-understanding normally is articulated, limits the ways in which people think and cuts off the opportunity of a broader orientation (once called 'civic virtues,' Bellah et al. 1985, p. 290). Because of their option for more substantive ethical identities and more active political participation Bellah and co-workers were labeled as communitarians. They preferred to describe their position in a slightly different way to avoid the risk of being misunderstood, privileging only small face-to-face groups such as families, congregations, and neighborhoods, and of being opposed to the state and all the other larger structures that dominate life at the start of the twenty-first century. They insist that freedoms and a Good Society depends on institutions, and they are critical that many individualistic Americans have trouble understanding and accepting social and political institutions (Bellah et al. 1991).

Etzioni's main criticism of liberal political theory is the negligence of the social preconditions that enable individuals to maintain their psychological integrity, civility, and ability to reason. On the other hand, he distances himself from those communitarians who hold that as long as a community has agreed on a set of values, these are the ultimate criteria as to what is morally appropriate (Etzioni 1995, p. 17). The central problem is to establish criteria for the evaluation of the internal value system of different communities. The question is: how to avoid perspectivism and relativism, and to develop criteria to rule out authoritarian and oppressive community values. In the first place, internal democracy of the community is the criterion. But additional criteria are needed, because a majority in a specific community might violate individual and minority rights. The second criterion is a substantial process of consensus building within a community. Beyond any particular community, the values of a society as a whole are a framework within a higher order of legitimacy; that is, the application of the Constitution, especially the Bill of Rights. The third level is the cross-societal moral dialogues, in which the societal values have to be legitimized against others. The fourth criterion is the global community. But even the global community might act oppressively, as it did when women were excluded from voting worldwide until the end of World War I. So the last resort for the dissenting individual is the clearly communicable self-evidence of some core values.

All these criteria have in common that they are based on moral dialogues of different reach, not on ontological values or on artificially constructed value systems. The communitarian approach is the support

of concrete societal and, if necessary, cross-cultural moral dialogues. Of course, they have to be limited to a set of core values (Etzioni 1996, p. 237). Through these reflections Etzioni tries to overcome the restriction of traditional communitarian thinking to limited communities and to bring into discussion the notion of a global community. These are questions the world-wide discussion of communitarian ideas has made inevitable. Communitarian approaches have to solve the question: Which community? What kind of community? Who belongs and who does not? This will be important especially for further discussions on the problem of multiculturalism (cf. Taylor et al. 1992).

The discomfort with the label 'communitarian' seems to be a general characteristic of many authors in this field, with the exception of Etzioni. Communitarianism seems to be a permanent, accompanying phenomenon to liberalism, but can hardly stand on its own. The traditional social structures of community orientation have been dissolved by the processes of modernization, individualization, and fragmentation. Attempts to restore or invent dense forms of social integration as an answer to modernity such as communism, fascism (*Volksgemeinschaft*), or some forms of Islamic fundamentalism are discredited because of their atrocities and their contempt for individual rights. Thus, in modern, technology-based service economies, communitarianism is only acceptable as a reflexive admonition to care more about the common good, but not as a coherent and mobilizing social ideology. So the prediction follows that communitarian ideas in a moderate form of social self-reflection and social integration programs will be a permanent subdominant melody in the chorus of further modernizing societies, because it meets a fundamental human demand for social integration on a theoretical level.

See also: Civil Society/Public Sphere, History of the Concept; Communitarianism, Sociology of; Community/Society: History of the Concept; Community Sociology; Individual/Society: History of the Concept; Individualism versus Collectivism: Philosophical Aspects; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Mass Society: History of the Concept; Tocqueville, Alexis de (1805–59)

Bibliography

Bellah R N et al. 1985 *Habits of the Heart. Individualism and Commitment in American Life, 1st Perennial Lib. Edn.* Harper & Row, New York

Bellah R N et al. 1991 *The Good Society*, 1st edn. Alfred Knopf, New York

Etzioni A 1993 *The Spirit of Community. Rights, Responsibilities, and the Communitarian Agenda*, 1st edn. Crown, New York

Etzioni A (ed.) 1995 *New Communitarian Thinking. Persons, Virtues, Institutions, and Communities.* University Press of Virginia, Charlottesville, VA

Etzioni A 1996 *The New Golden Rule. Community and Morality in a Democratic Society*, 1st edn. Basic Books, New York

Etzioni A (ed.) 1998 *The Essential Communitarian Reader.* Rowman & Littlefield, Oxford, UK

MacIntyre A 1981/1984 *After Virtue. A Study in Moral Theory.* University of Notre Dame Press, Notre, Dame, IN

Reese-Schäfer W 1994 *Was istt Kommunitarismus?* Campus Verlag, Frankfurt

Sandel M J 1982 *Liberalism and the Limits of Justice.* Cambridge University Press, Cambridge, MA

Selznick P 1992 *The Moral Commonwealth. Social Theory and the Promise of Community.* University of California Press, Berkeley, CA

Taylor C 1985a Atomism. In: Taylor C (ed.) *Philosophy and the Human Sciences, Philosophical Papers 2.* Cambridge University Press, Cambridge, MA, pp. 187–211

Taylor C 1985b Self-interpreting animals. In: Taylor C (ed.) *Human Agency and Language. Philosophical Papers 1.* Cambridge University Press, Cambridge, UK, pp. 45–76

Taylor C 1989 *Sources of the Self. The Making of the Modern Identity.* Harvard University Press, Cambridge, MA

Taylor C, Gutmann A, Rockefeller S C, Walzer M, Wolf S 1992 *Multiculturalism and the Politics of Recognition.* Princeton University Press, Princeton, NJ

Walzer M 1983 *Spheres of Justice. A Defense of Pluralism and Equality.* Basic Books, New York

Walzer M 1990 The communitarian critique of liberalism. *Political Theory* **18**(1): 6–23

W. Reese-Schäfer

Communitarianism, Sociology of

Communitarianism is a social philosophy that favors social formulations of the good. It is often contrasted with liberalism, which assumes that the good should be determined by each individual. To the extent that social institutions and policies are required, these should be based on voluntary agreements among the individuals involved, expressing their preferences. In contrast, communitarians view institutions and policies as reflecting in part values passed from generation to generation. These values become part of the self through internalization, and are modified by persuasion, religious or political indoctrination, leadership, and moral dialogs.

1. A Brief History

The term 'communitarian' was first introduced in 1841, to mean 'of, pertaining to, or characteristic of a community or communistic system; communitive.' It

has been infrequently employed since then until the middle of the twentieth century.

Among early sociologists whose work is focused on communitarian issues (but did not draw on the term) are Ferdinand Tönnies, especially his comparison of the *Gemeinschaft* and *Gessellschaft*; Emile Durkheim's studies of socially integrating role of values and the relations between the society and the person; and George Herbert Mead's work on the self. Other early relevant sociological works are those of Robert E. Park, William Kornhauser, and Robert Nisbet.

In the 1980s communitarianism was largely advanced by Charles Taylor, Michael Sandel, and Michael Walzer. They criticized liberalism for its failure to realize that people are socially 'embedded,' overlooking the fact that people can have a strong attachment to their societies. They lamented liberalism's focus on the individualistic concept of self-interest.

Asian communitarians argue that to maintain social harmony, individual rights and political liberties must be curtailed. Some seek to rely heavily on the state to maintain social order (for instance, leaders and champions of the regime in Singapore and Malaysia), and some on strong social bonds and moral culture (as Japan does). Asian communitarians also hold that the West's notion of liberty actually amounts to 'anarchy'; that strong economic growth requires limiting political freedoms; and that the West uses its idea of legal and political rights to chastise other cultures that have inherent values of their own.

In 1990 a new school of communitarianism was founded. Among its leading scholars are William A. Galston (political theory), Mary Ann Glendon (law), Thomas Spragens Jr. (political science), Alan Ehrenhalt (writer), and sociologists Philip Selznick, Robert Bellah and his associates, and Amitai Etzioni, who wrote books that, in 1990, laid the foundations for responsive (democratic) communitarianism.

According to Selznick's (1992) *The Moral Commonwealth*, communitarianism seeks reconstruction of liberal perspectives to mitigate the excesses of individualism and rationalism, and to encourage an ethic of social responsibility. In community there is an irrepressible tension between exclusion and inclusion, and between civility and piety. Thus community is not a restful idea, a realm of peace and harmony. On the contrary, competing principles must be recognized and dealt with.

Responsive communitarians, a group founded by Amitai Etzioni, took communitarianism from a small academic discipline and introduced it into public life, as well as recast its academic content. Its tools were *The Responsive Communitarian Platform: Rights and Responsibilities* a joint manifesto summarizing the guiding principles of the group; an intellectual quarterly, *The Responsive Community*, whose editors include several social scientists; several books; position papers covering issues such as shoring up the family, organ donation, health care, and bicultural education; and numerous public conferences, op-eds, and a web site (www.gwu.edu/ ~ ccps).

2. *Basic Principles*

Responsive communitarianism assumes that societies have multiple and not wholly compatible needs, in contrast to philosophies built on one core principle, such as liberty for libertarianism. Responsive communitarianism assumes that a good society is based on a balance between liberty and social order, and between particularistic (communal) and society-wide values and bonds. This school stresses the responsibilities people have for their families, kin, communities, and societies—above and beyond the universal rights all individuals command, the focus of liberalism.

While a carefully crafted balance between liberty and social order defines a generic concept of the good society, communitarians point out that the historical-social conditions of specific societies determine the rather different ways a given society in a given era may need to change to attain the same balance. Thus, contemporary Japan requires much greater tolerance for individual rights, while in US society excessive individualism needs to be curbed.

Communitarians also pay much attention to the relationship between the self and the community. Political theorists depict the self as 'embedded', which implies that the self is constrained by the community. Responsive communitarians stress that individuals who are well integrated into communities are better able to reason and act in responsible ways than isolated individuals, but if social pressure to conform rises to high levels, it will undermine the individual self.

Communitarians pay special attention to social institutions. Several of these form the moral infrastructure of society: families, schools, communities, and the community of communities. Infants are born into families whose societal role is to introduce values and begin the development of the moral self. Schools' role is to further develop the moral self and to remedy moral development if it was neglected or distorted by the family.

Communitarians emphasize that children reared in well-functioning families and schools will still not be sufficiently equipped for membership in a good, communitarian society. This is a point ignored by those social philosophers who assume that once people have acquired virtue and are habituated, they will be adequately guided by their inner moral compass. In contrast, communitarians assume that commitments to moral values tend to deteriorate, unless these are continuously reinforced. A major societal role of communities is to reinforce these commitments of their members. This is achieved by the community's 'moral voice', the informal sanctioning of others, built

into a web of informal affect-laden relationships, that communities provide.

Studies indicate that Americans in the 1980s often avoided speaking up on moral issues. The work of Robert Bellah and his associates has been particularly influential here, demonstrating the rise of first expressive and then instrumental individualism, and their negative effects. Many Americans, and to some extent members of other Western societies, embraced the liberal ideology that what is morally sound is to be determined by each individual, and one should not judge the acts of others. Very large increases followed in crime, drug abuse, divorce, and loss of trust in public institutions and legitimate authority and fraying of social bonds.

Within this context responsive communitarians point out that if a society has communities whose social webs are intact, who share a moral culture, and whose members are willing to raise their moral voice, such a society can rest its social order largely on moral commitments rather than on the coercive state. That is, the moral voice can reduce the inevitable tension between liberty and social order and enhance both.

3. *Critiques and Responses*

Critics argue that the concept of community is difficult to define. In response, Etzioni defined community as an amalgam of a web of affect-laden relationships among a group of individuals, who also share a commitment to a set of shared values and mores, who have a shared history and identity—in short, a particular culture.

Other critics claimed that the population of contemporary societies is concentrated in cities and is geographically highly mobile, and thus bereft of community. Communitarians respond that communities persist in many city neighborhoods, often based on ethnic or religious ties. Moreover, communities need not be residential; members can be spread among nonmembers. For instance, homosexual groups often constitute communities even if they are not all neighbors. Also communities exist at work, among professionals and even in cyberspace.

Critics maintain that communities are authoritarian and oppressive. Communitarians respond that communities vary in this regard. Contemporary communities tend to be relatively democratic, given the relative ease in which people can relocate as well as shift their loyalties and psychic investments among various communities of which the same person may be a member, for instance, at home and at work.

Critics point out that communities tend to be homogenous and exclude people who are different. Communitarians respond that communities must respect laws that ban discrimination on racial and other grounds, but argue that given the considerable human benefit of community membership, a measure of self-segregation should be tolerated.

The fourth element of the needed institutional structure is the community of communities. Responsive communitarians have noted that communities need to be bound socially and morally into more encompassing entities, if violent conflict among them is to be avoided. Society should not be viewed as composed of millions of individuals, but as pluralism (of communities) within unity (the society). The existence of subcultures does not undermine societal unity as long as there is a core of shared values and institutions.

4. *Civic and Good*

Most recently, following the growing popularity of the concept of civic society, Etzioni argued that it is insufficient from a communitarian viewpoint. This is the case because civic society tends to be morally neutral on all matters other than the attributes citizens need to make them into effective members of a civic society, for instance, to be able to think critically. In contrast, a good society seeks to promote a core of substantive values, and thus views some voluntary associations and social activities as more virtuous than others.

In the same vein, communitarians argue that while everyone's right to free speech should be respected, some speech—seen from the community's viewpoint—is morally highly offensive and when children are exposed, damaging, For instance, the (legal) right to speak does not render verbal expressions of hate (morally) right.

While sociologists made numerous contributions to altered communitarian thinking, this philosophy challenged sociology to face issues raised by cross-cultural moral judgments. Sociologists tend to treat all values as conceptually equal; thus sociologists refer to racist Nazi beliefs and those of free societies by the same 'neutral' term, calling both 'values'. Communitarians use the term 'virtue' to indicate that some values have a high moral standing because they are compatible with the good society, while other values are not and hence are 'aberrant' rather than virtuous.

In the same vein, communitarians reject the claim of cultural relativism that all cultures command basically the same moral standing, and do not shy away from passing cross-cultural moral judgments. Thus, they view female circumcision, sex slaves, and *hudud* (chopping off the right hand of thieves) as violations of liberty and individual rights, and abandoning children, violating implicit contracts building into communal mutuality, or neglecting the environment, as evidence of a lack of commitment to social order and neglect of social responsibilities.

5. Public Voice

The discussion up to this point focused on responsive communitarianism as a social philosophy, its social science propositions and ethical implications. Responsive communitarians also have been playing a considerable public role. They are best understood as a new environmental movement, one concerned with the well-being of society rather than nature. Like environmentalism, communitarianism appeals to audiences across the political spectrum. The record shows that their influence extends from the moderate social democratic left (especially Tony Blair and Bill and Hillary Clinton) to the moderate Tory right (including public figures such as Kurt Biedenkopf in Germany and David Willetts in the UK). Green parties are also among those who often embrace communitarian concepts.

Communitarian terms have become part of the public vocabulary in the 1990s, especially references to assuming social responsibilities to match individual rights, while the term 'communitarianism' itself is used much less often. The number of articles about communitarian thinking in the popular press, increased twelvefold during the last decade of the twentieth century.

See also: Civic Culture; Civil Religion; Civil Society, Concept and History of; Communalism; Communitarianism: Political Theory; Individualism versus Collectivism: Philosophical Aspects; Liberalism; Republicanism: Impact on Social Thought; Republicanism: Philosophical Aspects

Bibliography

Bell D 1993 *Communitarianism and its Critics*. Clarendon Press, Oxford

Bellah R, Madsen R, Sullivan W M, Swidler A, Tipton S M 1986 *Habits of the Heart: Individualism and Commitment in American Life*. University of California Press, Berkley

Bellah R, Madsen R, Sullivan W M, Swidler A, Tipton S M 1991 *The Good Society*. Knopf, New York

Etzioni A 1993 *The Spirit of Community*. Touchstone, New York

Etzioni A (ed.) 1995 *New Communitarian Thinking*. University of Virginia Press, Charlottesville, VA

Etzioni A 1996 *The New Golden Rule*. Basic Books, New York

Etzioni A (ed.) 1998 *The Essential Communitarian Reader*. Roman and Littlefield, Lanham, MD

Multhall S, Swift A 1992 *Liberals and Communitarians*. Blackwell, Cambridge, MA

Rosenblum N L 1998 *Membership and Morals: The Personal Uses of Pluralism in America*. Princeton University Press, Princeton, NJ

Selznick P 1992 *The Moral Commonwealth*. University of California Press, Berkeley, CA

A. Etzioni

Communities of Practice

The term 'community of practice' is of relatively recent coinage, even though the phenomenon it refers to is age-old and social scientists have talked about it under various guises. In a nutshell, a community of practice is a group of people who share an interest in a domain of human endeavor and engage in a process of collective learning that creates bonds between them: a tribe, a garage band, a group of engineers working on similar problems.

1. Definition

Not everything called a community is a community of practice. A neighborhood for instance, is often called a community, but is usually not a community of practice. Three characteristics are crucial.

1.1 The Domain

Since a community of practice is focused on a domain of shared interest, it is not merely a club of friends or a network of connections between people. Membership therefore implies a minimum level of knowledge of that domain—a shared competence that distinguishes members from other people. (You could belong to the same network as someone and never know it.) The domain is not necessarily something recognized as 'expertise' outside the community. A youth gang may have developed all sorts of ways of dealing with their domain: surviving on the street and maintaining some kind of identity they can live with.

1.2 The Community

In pursuing their interest in their domain, members engage in joint activities and discussions, help each other, and share information. That is how they form a community around their domain and build relationships. Having the same job or the same title does not make for a community of practice unless members interact and learn together. The claims processors in a large insurance company or the students in American high schools may have much in common, but unless they interact, they do not form a community of practice. The Impressionists, for instance, used to meet in cafes and studios to discuss the style of painting they were inventing together. These interactions were essential to making them a community of practice even though they usually painted alone.

1.3 The Practice

A community of practice is not merely a community of interest—people who like certain kinds of movies, for

instance. Members of a community of practice develop a shared repertoire of resources: experiences, stories, tools, ways of addressing recurring problems—in short a shared practice. This takes interactions over time. A good conversation with a stranger on an airplane may give you all sorts of interesting insights, but it does not in itself make for a community of practice. The development of a shared practice may be more or less self-conscious. The 'windshield wipers' community of practice at an auto manufacturer makes a concerted effort to collect and document the tricks and lessons they have learned into a knowledge base. By contrast, nurses who meet regularly for lunch in a hospital cafeteria may not realize that their lunch discussions are one of their main sources of knowledge about how to care for patients, even though in the course of all these conversations, they have developed a set of stories and cases that become a shared repertoire for them to think about and discuss new cases.

1.4 A Familiar Experience

We all belong to communities of practice. They have been around for as long as human beings have learned together. At home, at work, at school, in our hobbies, we belong to several communities of practice at any given time. And the communities of practice to which we belong change over the course of our lives. In fact, communities of practice are everywhere.

Theoreticians often wonder whether the concept of community of practice is an analytical category, whether it exists only in the theoretician's mind, or whether it refers to actual social structures in the world. The answer is that it is both:

(a) The extent to which any social structure is a community of practice is not something that can be determined in the abstract. Is a family a community of practice? What about a group of workers? A sports team? An orchestra? A classroom? This is always an empirical question that can only be resolved by analyzing the way the group operates. In this sense, the concept is an analytical category.

(b) Yet, you can go into the world and actually see communities of practice at work. Moreover, these communities are not beyond the awareness of those who belong to them, even though participants may not use this language to describe their experience. Members can usually discuss what their communities of practice are about, who else belongs, and what competence is required to qualify as a member.

Communities of practice are a familiar experience, so familiar perhaps that it escapes our attention. Yet when it is given a name and brought into focus, it helps us understand the world better. In particular, it allows us to see the social world as structured by engagement in practice and the informal learning that comes with it, rather than more obvious formal structures such as institutional boundaries.

2. Theoretical Applications

Why have many social scientists found the concept of community of practice to be a useful unit of analysis? Well, if you want to understand broad issues such as culture, identity, and learning in terms of the processes by which people create systems of meanings, then it is useful to consider a unit of analysis where these processes involve a direct experience of engagement for participants.

Communities of practice break down the traditional dichotomy in social theory between perspectives that give primacy to structure (history, culture, myths, or class, of which moments of life are mere instantiations) and perspectives that give primacy to immediate experience (local interactions, of which broader structures are an emergent property). A community of practice is a mid-level unit of analysis that combines both elements. It is neither an abstract structure nor a passing experience. Unlike a culture, it is within the reach of individual participants. Members of a community of practice contribute to the development of the practice through direct engagement in their community. Unlike momentary interactions, however, a community of practice has an enduring character as a social structure. It can therefore accumulate collective cultural resources over time. It is an elementary structure that has all the characteristics of social life—the 'cell' of cultural production and reproduction.

Social scientists have used the concept of community of practice for a variety of analytical purposes, but the primary use of the concept has been in learning theory. The concept originated in studies of apprenticeship (Lave and Wenger 1991). People usually think of apprenticeship as a relationship between a student and a master, but studies of apprenticeship reveal a more complex set of social relationships through which learning takes place—with journeymen and more advanced apprentices. The term community of practice was coined to refer to the community that acts as a living curriculum for the apprentice. Then researchers started to see these communities everywhere, even when no formal apprenticeship system existed.

As a basis for a general social learning theory, however, the learning processes of a community of practice are not limited to training novices. The practice of a community is dynamic and involves learning on the part of everyone. The shared competence defined by the community is always in interplay with the experience of members. Sometimes, as in the case of apprenticeship, it is the competence of the community that pulls the experience of the novice until the novice has a full experience of competence. Sometimes, members bring new ideas and insights: it is their experience that pulls the competence of the community along. One can think of learning as a tension between competence and experience (Wenger

1998). Whenever either starts pulling the other, learning takes place. Learning so defined is a dynamic, two-way relationship between people and their communities. It combines personal transformation with the evolution of social structures.

While the concept of community of practice has been most widely used to analyze learning and the social organization of knowledge, it has also been useful for investigating other aspects of the social world: the construction of subcultures as a form of institutional resistance (Eckert 1989); the reproduction of social classes (Willis 1977); the formation of identities as trajectories through communities of practice and multimembership combining the simultaneous influence of multiple communities (Wenger 1998); local meanings and linguistic change through engagement in shared practice (Eckert and McConnell-Ginet 1992).

It is often useful to look at a social entity—a culture, a linguistic group, a 'social world' (Strauss 1978), an organization—as a constellation of communities of practice. Such social entities are best understood, not as a uniform group, but as a complex set of interconnected communities of practice, each with its own local 'mini-culture' as it were. Such a perspective makes it possible to understand local differences as well as the processes by which broad patterns are recreated in practice. The overarching entity then is the emerging property of interactions within and among local practices (Wenger 1998, Brown and Duguid 2000).

3. *Practical Applications*

Beyond social theory, the concept of community of practice has found a number of practical applications in business, organizational design, education, and civic life.

3.1 *Business Organizations*

The concept has been adopted most readily by people in business because of the increasing need to focus explicitly on knowledge (Wenger et al. 2002). Initial efforts had focused on information systems with disappointing results. Communities of practice provided a new approach, focused on the social structures that could best assume ownership for complex and dynamic knowledge with substantial tacit components. A number of characteristics make communities of practice a natural fit:

(a) Unlike training or research departments, they are not separate units. Rather they pervade the organization, since people belong to communities of practice at the same time as they belong to their business units or teams.

(b) Communities of practice address the informal and tacit aspects of knowledge creation and sharing, as well as the more explicit aspects.

(c) They allow a much closer connection between learning and doing, while still providing structures where learning can accumulate.

(d) In a time of globalization and disaggregation, they create connections among people across institutional boundaries and potentially across the globe.

From this perspective, the knowledge of an organization lives in a constellation of communities of practice each taking care of a specific aspect of the competence that the organization needs. However, the very characteristics that make communities of practice a good fit for stewarding knowledge—autonomy, practitioner-orientation, informality, crossing boundaries—are also characteristics that make them a challenge for traditional hierarchical organizations. How this challenge is going to affect these organizations remains to be seen.

3.2 *Education*

In business, focusing on communities of practice adds a layer of complexity to the organization—a kind of orthogonal structure focused on knowledge, while the core structure of the organization still focuses on business processes and results. But they do not imply a restructuring the whole system. Schools have been a bit slower at adopting the concept of communities of practice because sharing knowledge is already their main activity, and adopting communities of practice as a basic organizing principle implies a deeper rethinking of their structure. In educational circles, the hope is that communities of practice could bring the experience of schooling closer to everyday life along three dimensions.

(a) *Internally*. How to ground school learning experiences in practice through participation in communities around subject matters?

(b) *Externally*. How to connect the experience of students to actual practice through peripheral forms of participation in broader communities beyond the walls of the school?

(c) *Over the lifetime of students*. How to serve the lifelong learning needs of students by organizing communities of practice focused on topics of continuing interest to students beyond the schooling period?

From this perspective, the school is not the privileged locus of learning. It is not a self-contained, closed world in which students acquire knowledge to be applied outside, but a part of a broader learning system. The class is not the primary learning event. It is life itself that is the main learning event. Schools, classrooms, and training sessions still have a role to play in this vision, but they have to be in the service of the learning that happens in the world.

More generally, the concept of community of practice has promise in suggesting ways to organize

societies around issues and functions. The US government and the World Bank are experimenting with these approaches by connecting people across cities and countries with practice-based communities that complement place-based communities. New technologies such as the Internet have extended the reach of our interactions beyond the geographical limitations of traditional communities, but the increasing flow of information does not obviate the need for community. In fact, it expands the possibilities for community and calls for new kinds of communities based on shared practice.

See also: Action Theory: Psychological; Community, Expression of; Community, Social Contexts of; Community/Society: History of the Concept; Community Sociology; Cooperation and Competition, Psychology of; Cooperation: Sociological Aspects; Group Processes in Organizations; Interest Groups

Bibliography

Brown J, Duguid P 2000 *The Social Life of Information*. Harvard Business School Press, Harvard, MA

Eckert P 1989 *Jocks and Burnouts: Identity in the American High School*. Teacher's College Press, New York

Eckert P, McConnell-Ginet S 1992 Think practically and act locally: language and gender as community-based practice. *Annual Review of Anthropology* **21**: 461–90

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, New York

Strauss A 1978 A social-world perspective. *Studies in Symbolic Interaction* **1**: 119–28

Wenger E 1998 *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, New York

Wenger E, McDermott R, Snyder W 2002 *Cultivating Communities of Practice: A Guide to Managing Knowledge*. Harvard Business School Press, Harvard, MA

Willis P 1977 *Learning To Labour: How Working-Class Kids Get Working-Class Jobs*. Columbia University Press, New York

E. Wenger

Community Aesthetics

Just as politics was significant in the definition of modernity, aesthetics might well become the mark of postmodernity. Aesthetics is understood, of course, in its etymological sense: that we all 'feel.' Such aesthetics is the foundation of the community, the foundation of what once was called the postmodern 'tribe' (Maffesoli 1996a).

Aesthetics implies an impulse to live that cannot but offend fixed minds, which are only able to witness and analyze average thoughts and average ways of life. But aesthetics reminds us that our being is an event, or even advent. To return to the opposition between modernity and postmodernity, one can say that in the case of the former, history follows its course, whereas in the case of the latter, the event emerges. It intrudes. It forces its way and it does violence. Hence its irrevocably brutal, unexpected, and constantly astonishing characteristics. Here again, one is confronted by the difference between the tonality of drama or dialectic which postulate a possible solution or synthesis, and the tonality of the tragic that is intrinsically *aporic*.

The advent is singular. But its singularity is rooted in an archaic timeless substratum. The 'archaisms' in question are, of course, rethought according to the present, they are experienced in a specific way, but with the origins always kept in mind. Event–advent? For sure, what is qualitatively, intensely experienced, works to bring out what has always been present within our being, be it individual or collective. We can refer to Heidegger on this point, and to the attention he paid to a postmetaphysical thought, applying himself to bringing out the most 'simple' that serves as a substratum to human existence. But we can also refer to Leibnitz who, in his 'principle of indiscernible,' works to find a balance between absolute difference and repetitive identity (Vattimo 1990, p. 21). Between the two, the romanticism or the philosophy of life enhances the tragic aspect of the present, along with its demand, its impulse to live, and the sense of urgency it exudes.

Is this not characteristic of the astonishing contemporary attitudes, based on aesthetics, caring little or not at all about the consequences of their acts? Plural families, and successive and transient love affairs are proof of such behavior in the domain of emotion. Political changeability or ideological variations exemplify it as far as public life is concerned; the acceptance of the anarchical laws of production and, at the same time, the extraordinary suspicion towards them, attest to it in what can be called the economical disorder. An atmosphere of unconcern pervades the whole thing, not furthering the worry about the future but, on the contrary, an urge to live in the present according to a way of being that, during the ages, has developed progressively.

If we try to define this kind of aesthetic environment, we can compare it to an eternal paganism. A paganism that works to grab hold of life, grasps what it offers, anything that turns up. A pagan exuberance devoting itself to using the delights of the present, leading a daring, bold life; a life penetrated by the freshness of the fleeting, precarious, and, therefore, intense instant. In his analysis of Machiavelli's opposition to Christianity, Fichte speaks of his 'general impiety' Fichte 1981, p. 48). It seems that we can extrapolate what he says from his paganism. For it is clearly the essence of Christianity that we find again in the political sphere, in the economical conception of existence, or in the

quest for security that the various social institutions offer.

It is precisely this type of 'Christianity' against which the contemporary impiety rebels. The youthful aspect of its agitation, the 'freshness' of its rebellions, its exacerbated search for intense pleasure, multiform, in the present, all combine to make it view the 'ancient world' as its original homeland. 'Ancient world' should, of course, be understood in a metaphorical sense; that is to say, as anything that contravenes the various 'categorical imperatives' formulated by modern moralism, be they of a sexual, economical, or ideological nature. This return to the ancient—to the archaic—is a feature of postmodernity. As if, beyond a mere parenthesis, for better or worse, in the everyday life or in paroxysm, in a subdued way or, on the contrary, in destructive excesses, we are rediscovering the sublime quality of the beauty of the world. Only this beauty matters, so we may as well delight in it for what it is, even at the cost of submitting to its terrible, formidable laws that we have to accept anyway, which cannot but evoke the theme of *amor fati*, the important consequences of which can be estimated in a Nietzschean spirit.

It is as they wake up from the Promethean dream that more and more people choose to behave in a Stoic spirit. A generalized stoicism according to which that upon which we have no power becomes indifferent. And thus the *amor fati*, thanks to which our fate is not only predetermined but also accepted, even appreciated as such. That attitude generated a certain form of serenity that can seem paradoxical, but which is actually at the basis of various and numerous demonstrations of generosity, mutual aid, voluntary help, and humanitarian actions that stem from social life, and that even tend to multiply. For the acceptance of that which can go hand-in-hand with the concern to take part in that which is: not to bring it under control, but to accompany a given so that, possibly, one makes it give the best of itself. This is the aesthetic logic, described briefly. Thus, realization of the self or of the world no longer consists in a simple economical action, but blossoms in ecological interaction. That may be the way to switch from the Hegelian and Marxist control that is a feature of modernity, to what Bataille calls 'sovereignty,' a sovereignty working on structural reversibility and this is, more than anything else, the sign of pre- and postmodern eras.

If we apply it to the contemporary world, such 'proclivity' does not fail to secrete an undeniable wisdom. A wisdom that is nonactive without being passive, bringing out the tendency that, at a certain point, is inscribed in the reality in question. Thus, to use an idea (Maffesoli 1996b), moralism and necessity are replaced by a 'deontology' that takes situations earnestly and behaves consequently. Moralism rests on the injunction to be this or that. The individual has to submit themself to the plan that has been decreed *a priori*; society is also compelled to become what the intellectual, the politician, the expert thinks it has to be. Quite different is the deontology that adapts to a general tendency, that is heedful of the (spirit of the moment); in short that is in keeping with the opportunities of the present.

Only situations matter. Such immanentism does not imply indifference but, on the contrary, a constant conscience, an awareness of what is: the world, the other, the social. In short, we could talk about a 'co-presence' towards otherness under its various modulations. A variable intensity of 'co-presence,' but integrating the entirety of being, and no longer only this or that part or characteristic.

To reconsider a famous theme already brought up to date by Dodds, we can recall the role of the *daïmon* in the Greek tradition. Of course, Socrates put forth his own *daïmon*, and we can consider that it was a generalized belief that no one could ignore. But the interesting point we want to underline within the scope of this discussion is the tight link that exists between the *daïmon* and necessity, the role of which is so important within the science of ancient culture. In a word, we can say that a much greater number of things depend on necessity than the individual's own nature. It is that very fact that is expressed in various ways in tragedies: we are more acted upon than we really act ourselves. Fate is there—omnipotent, merciless—directing the individuals in spite of their own will in the direction of what has been written down. Here again, it is a matter of a certain form of predestination. To quote just one example out of many, the whole myth of Oedipus is based on such a 'necessity,' with the extreme consequences we know of.

Indeed, community aesthetics only encourages the rise of the impersonal. What is at stake in such a repetition of fate is the very negation of the philosophical foundation of the modern Western world: free will, the decision of the individual or social groups acting together to make History. The great fantasy of universality is the natural consequence of this. Opposed to that, the assertion or reassertion of cyclical systems invalidates such a free will (Dodds 1959, p. 51). The various mythical 'orients' that intrude into postmodernity take up with the impersonal powers again, and the inescapable aspect of action. Be it the various philosophies, or more simply the Buddhist, Hindu, or Taoist techniques, or African clairvoyance, directly confronting the telluric forces, or the Afro–Brazilian cults of possession, without forgetting the multiple New-Age practices, or quite simply the fascination astrology holds over people. All of this emphasizes the fact the individual is, at worst, only the plaything, and at best, the partner of forces that exceed him, and to which he knows he must adapt.

Expressions of this contemporary mythology can be seen in science fiction films, pop videos, sometimes even in advertisements. These expressions reveal the relativization of the free will by the supra-individual 'force.' Of course, strong minds laugh at this, but the

resonance is undeniable. It taps into the social imagination, ensures the success of folklore shows and historical reconstructions, inspires crowds to visit places of pilgrimage, and makes best sellers of rites-of-passage novels. We can call this the 'ethics of the aesthetic,' another way of asking the question posed by medieval alchemists who wondered about the *glutinum mundi*, the glue of the world that ensured that whatever the case, there was something rather than nothing, and that this something was coherent. The 'glue of the world' is, therefore, an impersonal force, a vital flux, in which everything participates in a mysterious attractive correspondence.

There are numerous poets, artists, and utopians who have celebrated such an attraction. It is possible to make a social anthropological reading of it. This is suggested through the term 'orgy,' that is to say, shared passion, social empathy. Or even, by distorting the expression of Durkheim, one could speak of an 'organic solidarity,' meaning that, whether willingly or not, everyone is, essentially, part of a whole that makes one what one is. In short, we only exist because the other, my close relation, or the Other, the social, gives me my existence. I am as I am because the other recognizes me as such. Such an assertion might seem shocking, but is this not, empirically, how, from the smallest to the largest, societies function. Such an 'effect of structure' is well described by Marie Douglas, in her book *How Institutions Think*, (Douglas 1992, p. 12). It enables us to understand that whoever does not submit to such recognition is rejected, stigmatized, or marginalized. Their exclusion is due to the fact that they don't have 'the smell of the clan,' or don't want to acquire it.

Thus, above and beyond individualism, whether theoretical or methodological, empirical social life is only the expression of successive sentiments of belonging. You are a member, part of, you integrate, participate or, in common parlance, 'you belong.' Even if during the good periods of modernity there was autonomy, distinctness, affirmation of an individual identity or class, today all that is only a deception, illusion, simulacrum. Sociology of the orgy (Maffesoli 1991), that is, order of the fusion, even the confusion that determines that everyone exists according to a principle of heteronomy.

With this in mind we can understand the return in force and resonance of the emblematic figures and other daily archetypes that exist in contemporary society. The phenomenon of 'fan' groups among the young generation is only one of the extreme forms of these multiple memberships, which they are not aware of. Thus one 'participates' magically with such and such a rock singer, sporting idol, religious or intellectual guru, or political leader. A participation that generates a quasi mystical communion, a common feeling of belonging. Gilbert Durand, commenting upon the major tragic figures, like Don Giovanni, emphasizes that they become pure 'objects.' More objects than subjects as they only exist in the mind of others, they become an 'ideal type' (Durand 1982, p. 207).

We can pursue this analysis by noting that these 'major abstractions,' these archetypes, tend to multiply, even to democratize themselves. Increasingly there are small major figures. Ultimately, each postmodern tribe has its emblematic figure just as each tribe, strictly speaking, possessed, and was possessed, by its totem. In both cases identity, free will, individual decision, or choice can be asserted or demanded, but they are dependent on the identities, decisions, and choices of the group. Note, moreover, that these archetypes regain force and vigor at the same time as the tragic atmosphere of the moment is affirmed. There is a correlation here that deserves consideration.

Instead of being opposed, irreducibly, instead of being exceeded, according to a dialectic and dramatic mechanism, in a soothing synthesis, freedom and necessity are, at certain moments, experienced in a 'contradictory' tension, which is what is called a conflicting harmony. This, of course, recalls the mystic tradition, or the philosophy of Hinduism, but, also, the process of individuation, as described by Jung (1980), in which the 'I' is used and experienced as an object of the Subject that encompasses it. This is an experience of the self that does not destroy the empirical individual, the 'I,' but which, on the contrary, raises it; that is to say, takes it up into a larger whole. Hence the intensity and jubilation of the tragic situation, of the Nietzschean *amor fati*: to be free in a necessity full of love (Aurigemma 1992, p. 250). In short, a form of dependence full of tranquillity in which the individual becomes fulfilled in a state of 'being more' in which he is revealed to himself.

We can see what these situations of 'being more' could be in contemporary society. Great gatherings, all kinds of crowds, multiple trances, mass sporting celebrations, musical excitements, religious or cultural effervescence. All these things raise the individual to a form of plenitude that has none of the colorlessness of economic or political functionality. In each one of these phenomena there is a sort of magic participation in the strange, strangeness, in a global nature that exceeds individual particularity. A global nature which is sacred, and in which each person receives communion. Is this an irony of the tragic, or 'ruse' of the collective imagination, returning this numinous dimension to the social circuit that modernity believed it had evacuated from social life? Re-enchantment of the world? Yes, in so far as one assists, undeniably, in the surpassing of simple utility, even instrumentality, whether this is individual or social.

In such a perspective, the world and the individual do not become, progressively, what they have to be according to a programmed finality, but, one could say, the 'befall' (*adviennent*) to what they are. The archetype, is as it were, only an aide in this revelation, something that serves as developer, used to bring out

what is already there. It is in this sense that there exists a strict relationship between the tragic dimension of the archetype and the accentuation of a cyclical conception of time.

This structural proximity between the archetypal procedure, the collective unconscious and the cycle stems from the fact that, to use a formula of Jung, all 'vital processes follow their own inner laws.' To take the example of the unconscious, you cannot make progress by forced marches, it comes 'in its own time.' Like a spring which trickles, gushes forth or dries up according to its own rhythm, the emergence of the unconscious flow can never be predicted with complete certainty.

We could also establish a connection between this and the fact that, like in alchemical thought and in the dynamic of the unconscious, there is something circular, better, spiralesque in the production of images. They do not obey the mechanic linearity of simple reason, but follow a series of convolutions, which particularly complicate interpretation. Indeed, there is a labyrinthine structure that exists in both the unconscious and the world of images. And if the former has been invested predominately by intellectual interpretation, this is not the case for the latter, which is still ignored, looked down upon, or marginalized by thinkers, at least by those that defend a strictly nationalist point of view.

It is certain that the convolutions or even, if we refer to a Jungian theory, the 'circumambulations,' describe the slow circular work that everyone makes in order to achieve, bit by bit, the realisation of what can be called 'greater being,' which is the work of a whole lifetime (Jung 1980, p. 145). The Tibetan *mandala* is a good illustration of this for the Oriental tradition, just as the myth of the quest for the holy grail expresses it in the Western tradition. In both cases there is repetition, cyclical movement and a tragic conception of life. The archetypal figures always proceed by recurrences, they always refer to a mythical time that is nondatable, the time of our stories and legends: 'at this time,' *l'illud tempus*.

This is striking in the case of mythical illustrations strictly speaking, of which there are many examples to be found in literature, cinema, theater, or song. But this phenomenon of timelessness, of cyclical or tragic accentuation is noticeable, also, in the daily or, *a fortiori*, show-like performance of contemporary stars. The *puer aeternus* that Michael Jackson, for example, represents; the depraved repentant that Madonna represents; or more prosaically, the divine rogue illustrated by the contemporary 'trader'; without forgetting the aggressive heroes of the sports world; all of these emerge from a re-enchantment of the world that has a strong repercussion in the collective unconscious. These figures don't create anything specific, they only repeat anthropologically rooted characters or ways of being. It is this cyclical aspect which means, moreover, that they are raised on a pedestal. And it is by communing in these redundant productions, by identifying in them, that everybody, at the end of a long initiation, which most of the time is nonconscious, exceeds oneself, 'has a ball,' in an aesthetic manner, in something surpassing confinement, or the shrinking, in the small individual 'I.'

Whether in empirical expressions of contemporary art or, and they are not necessarily very far removed, mythical contradictions, it is important to reveal a fundamental recurrence. Lévi-Strauss and Gilbert Durand have greatly insisted on this aspect: repetition, and the 'do-it-yourself' correlative to it, is at work in the great spiritual works of humanity (Durand 1980, p. 130). To a certain extent this repetitive aspect, whether a Nietzschean 'eternal return,' a writer's obsessive idea, a musician's typical musical phrase, a painter's 'touch,' the never-ending theoretical digression of a thinker, even the recognizable ritornello of a singer, all of these things underline the presence of the timeless in history, a sort of immobility in movement.

It is having in mind the redundancy of myth and the repetition in daily creations, without forgetting, of course, that which is at work in everyday life, that we can understand the part of intimate emotion secreted by the familiarity of phenomena, situations, ideas etc … which return with regularity. It is this that characterizes community aesthetics. '*L'habitus* as analyzed by Thomas of Aquinas insists on the structuring aspect of the established custom. The metaphor of the 'fold' that Deleuze offered to reflection is a way of bringing the resonance of habit up to date. All these things show that individual or collective improvement is not necessarily progress without end, as modern pedagogy has postulated, but can, at certain moments, be carried out with an awareness of what is recurrent: customs, myths, and rites, the habits of a given society. This is exactly what a premodern society offered, and it is possible that in postmodernity it has regained importance once more.

See also: Art: Anthropological Aspects; Art, Sociology of; Cultural Studies: Cultural Concerns; Film: Genres and Genre Theory; Postmodernism in Sociology; Postmodernism: Philosophical Aspects; Television: Genres.

Bibliography

Aurigemma L 1992 *Perspectives jungiennes*. Paris
Dodds E R 1959 *Les Grecs et l'irrationnel*. Paris
Douglas M 1992 *Comment pensent les institutions*
Dumont L 1991 *L'ideologie allemande*. Paris
Durand G 1982 *Le Retour des immortels*, *in Le Temps de la réflexion*
Durand G 1980 *L'ame tigrée*. Paris
Fichte J G 1981 *Machiavel et autres écrits philosophiques et politiques*. Payot, Paris
Giddens A 1987 *La Constitution de la société*. Paris
Jung C G 1980 *Mysterium conjunctionis*. Paris

Maffesoli M 1991 *The Shadow of Dionysus*. SUNY, New York
Maffesoli M 1996a *The Time of Tribes*. Sage
Maffesoli M 1996b *The Contemplation of the World*
Vattimo G 1990 *Ethique de l'interpretation*. Paris

M. Maffesoli

Community Economic Development

The terms community, economic, and development each has a clear meaning. Community usually suggests a defined geography with a set of people who hold common history, beliefs, or regard for the territory in which they reside. Economics refers to the workings of capital, i.e., money within institutions and among individuals to create or control wealth. Finally, development refers to a process by which institutions, organizations, or individuals gradually transform from their current state to something bigger or at least different from the original state. Each of these words when fused together takes the form of a new notion of local economic development. Community economic development is a branch or arena of local economic development with a different set of intents. Community economic development is several things in this context. It is a set of beliefs or a movement; an arena of practice; a set of tools or methods; and finally, a set of government programs. For these reasons, when the term community economic development is used it conveys a set of refined principles and practices. As a result for a catholic understanding of the field of community economic development each of the above notions needs to be discussed.

1. The Movement: People Based Economics

Community economic development focuses on distressed or economically disadvantaged communities as zones of action. That is, community economic development has a strong bias toward place. Depressed geographic and socially bound areas have been left behind by the normal processes of economic change (Goldsmith and Blakely 1992). As Goldsmith and Blakely 1992 show in their analysis of urban American disadvantaged populations reside in areas or regions of considerable wealth. But in many areas wealth is unequally distributed among the people creating conditions of observable poverty and neglect for those people and place that do not share the same economic advantages. These communities are deeply divided with: separate assets; separate opportunities; and are separate physically and socially isolated places within the larger national and international economy. Other scholars in the industrialized world have made similar observations. The basic thesis of community economic development is to address this situation, not entirely by the re-capture or redistribution of wealth from the larger society but through the creation of new wealth within these disadvantaged communities.

As Habersfeld, a leading community economic development practitioner (1981, p. 8) puts it clearly, 'residents of low income communities want development which promotes their interests.' Habersfeld describes these interests as the ability to build community capital by: (a) creating productive work for the residents of the distressed neighborhoods in their community or close by; (b) providing local goods and services for the community with local ownership wherever possible; (c) acquiring the same resources as other communities in the city or country with respect to governmental infrastructure that enhances the appearance and livability of the area; (d) generating new wealth through locally based and owned enterprises; (e) reducing the communities' dependence on external resources such as national and local government programs; and, (f) providing the residents with the experience and exposure to present themselves at the neighborhood level as equal participants in the larger city and regional politicoeconomic processes (Habersfeld 1981, p. 8).

Thus community economic development is based on a very strong and clear belief structures or a philosophy aimed at a people–location axis. This philosophical position is the basis for a movement that places the community's social development or social capital formation at the center of the economic process. In this sense it differs from other forms of economic development that focus almost exclusively on the wider economic transactions in the macroeconomy and assume that they will affect everyone within the economy. The stress on community or neighborhood as the target for economic development is also unusual because general economic theory does not recognize place or particular groups of people as central variables. In essence, community economic development is a movement based on the concept that the basic resources in poor communities can be enhanced to generate new opportunities and wealth within the community.

2. The Practice of Community Economic Development

The practice of economic development aims at developing, attracting, and retaining or capturing economic activity for a designated locality. Economic activity equates to increased tax base, jobs, and the resources to pay for civic goods ranging from roads to symphonies or museums. Economic development practice is devoted to attracting outside capital as the central means of improving the regional or local economy. In fact, incentives (discussed later) are offered as inducements for firms to move from existing locations to new places. This form of economic development assumes

that almost all forms of job and tax-generating activity is of benefit to the community that successfully induces it.

Community economic development does not subscribe to the above practices of moving or stealing firms from other locations. The reasons for this are simple. First, in most situations when a firm moves it takes with it skilled resources and leave behind unemployed workers in need of jobs. In essence, to produce wealth in one place, another is impoverished. This is not to say that firms should not move for economic reasons such as gaining access to better raw materials, human resources, or customers. However, the premise of community economic development is that net new wealth needs to be generated not merely transferring jobs from one place to another. Second, community economic development presumes that new economic activity needs to fit the human resources in the distress community (Blakely 1994). By fit, we mean that the jobs or the activity will absorb the local population with only minor training or access to locally available raw materials or markets.

Community economic development practice eschews the notion that all jobs are good jobs and that any economic activity is better than none. The reason for this is that too many low income communities have been victimized by corporatist activities that use their territory and resources and that do not benefit the locals. Countless examples can be cited where low income residents were promised jobs in new plants only to see the workforce imported by the new employers. Or in other instances, poor communities have been displaced or entirely destroyed by new shopping malls or plants designed to serve them. A classic case of this type of capital over community is the sad story of Poletown in Detroit which was razed for a new Chrysler autoplant that was supposed to employee several thousand local people (Wylie 1990). In the end of the plant never produced the anticipated jobs and the community was lost. There are many Poletown illustrations such as the London Docklands and other places around the world, where community life has been sacrificed for economic activity.

Community-based economic development practice rejects the notion that a community must be destroyed to be saved. Rather the goal of community economic development is to forge new economic activity out of the assets in the distressed community. Harvard Business Professor Michael Porter in a seminal article in the *Harvard Business Review* argues that inner city low income areas are indeed capable of generating their own new wealth (Porter, in Boston and Ross 1995). Porter argues that '… inner cities occupy what should be some of the most valuable locations in their respective regions, near congested high cost areas, major business centers, entertainment complexes, transportation infrastructure, and communication nodes. As a result, inner cities can offer a competitive edge … (Porter, in Boston, p. 305). Porter's thesis is consistent in many respect with other community economic development scholars prospective on distressed communities. However, not all of them share his enthusiasm for expanding the basic supplier industries such as food processing and local serving firms. Friendly community economic development scholars suggest that Porter's assessment is too glib. Professor David Sawicki, one of the most knowledgeable scholars in the field of economic development says, 'What the policy debate does not need is yet another model of urban redevelopment that raises unrealistic expectations about what can be accomplished and which directs potentially productive players (young, black, business school graduates) into private sector initiated activities without real roots in the inner city.' (Sawicki and Moody 1995, p. 75.)

The real issue is how the ghetto or any low income community possesses a sufficient resource base for any form of economic development. Michael Teitz, a leading policy scholar, who is sympathetic to locally based economic development suggests that neighborhoods or even larger communities within a city are not relevant units to focus on for development purposes. He argues that, '… neighborhoods by their very nature are problematic as a target for economic development strategy, insofar as the strategy is intended to generate economic activity and employment directly within their boundaries.' (Teitz 1989, p. 112.)

So clearly, the advocates of community based economic development practice have to provide tangible evidence that the locus of such development can be targeted at the distressed community level. My own assessment is that the notion of community targeting can be too drawn too finely. What every disadvantaged community must be able to do, is to connect itself with the growing sector of the larger regional economy. Community economic development practice then is to find ways to increase the value added productive capacity of the under-performing communities and their residents. This requires, as I suggest, in my book *Planning Local Economic Development* (Blakely, 1994) strategies that build long term institutional capacity with low income neighborhoods through instruments like Community Development Corporations to help guide the local community resources so that they can play a role in the wider economy. My perspective is shared by Avis Vidal (1992), the leading community development scholar, who suggests that these institutional vehicles build the human social capital that can link distressed areas by forging the necessary combination of political and economic links to the larger region.

3. *The Tools of Community Economic Development*

Economic development is almost always described by laypersons and even some professionals in term of the

tools. For example, tax reductions or low cost land incentives, a tool of economic development, is frequently called the economic development strategy for a city or region. Clearly, these tools are very visible. But the tool means very little without a more compete and well refined strategy. In community economic development the tools are aimed at dealing with the underlying pathology of the people and the place. There are three types of tools used to deal with community economic revitalization. These are: (a) locality development—improvements in the physical character and appearance of low income areas; (b) human capital formation—increasing the skill levels and capacity of low income individuals to gain employment or start businesses to serve their communities; and, (c) finally, social capital or capacity building—building new institutions and organizations that increase the political and economic institutional capacity of the community to develop and manage its own resources.

3.1 Locality Development Tools

Distressed communities are not attractive places. There are very observable conditions in disadvantaged areas in which the community lacks a viable economic base. The signs of this include boarded up storefronts, deteriorating public infrastructure such as streets and sidewalks, and visible criminal activity. These are the symptoms of the problems. But much like human disease, the symptoms must be arrested before a full effort can be launched to eradicate the problems. Another observable condition is idleness. Working age men and women are on the street during employment hours with no 'place' to go, or economic activity to attach them to the larger community. Again, this situation is symptomatic of closed factories or a lack of employment opportunities in the area.

Community economic development strategies attempt to deal with both the perceived, that is observable, as well as the real underlying situation. At one level, there are a host of community economic development tools aimed at dealing with the physical appearance of distressed areas. Some of these are well known, such as commercial facade improvement and community or neighborhood commercial revitalization projects. These programs use a variety of local initiatives aimed at bringing about physical revitalization of dilapidated building stock. Some projects do no more than provide paint and offer small loans for owner repairs and rehabilitation of their premises. Government agencies or services in the area attempt to re-do their facades to look less institutional, in character, thereby appearing safer, more lively and wholesome to passers by. More recently, new strategies called business or neighborhood improvement districts are being employed to reach the same ends. In these programs local merchants and sometimes residents agree on a small tax assessment as dedicated fund to do civic improvements and provide enhanced police services to the neighborhood. This tool is employed in communities of all economic strata. Improvement siatricts are now being touted as a new approach to ghetto physical revitalization. One of the major attributes of this approach is the degree of business and citizen involvement that it stimulates. This approach works where there is still a cadre of active local merchants with the resources to help get the activity started and with the energy to maintain it. In some communities the improvement district program is merged with community policing which incorporate the local residents as well as the merchants.

Physical based planning is a heavy component of many of the more recent tools like the enterprise and empowerment zones. The zone notion aims directly at targeting the place–people link in community economic development. No matter which name is used the basic notion is to provide government tax relief and other incentives to fix up or reclaim inner city or distressed are a land and buildings for re-use as business venues. Under most of these programs new businesses that employ local residents receive tax breaks or very low interest loans to use or expand an underutilized site. In addition to these tax incentives a firm locating in the targeted zone will receive special tax relief for every local worker hired.

There is some evidence that these strategies work under certain conditions. For example, New York City has seen a resurgence of many low income areas in dying neighborhoods using a combination of tough police actions to deter any forms of low level street crime and an aggressive clean up and fix up programs. The New York approach has been widely imitated across the United States in major cities with generally good results. However, some critics complain that the problems do not go away, but are merely shifted to even poorer neighborhoods in the city. Homeless advocates claim these programs have made the plight of the homeless even less tolerable since they target the homelessness and offer no alternatives for the vulnerable homeless populations.

3.2 Human Capital Formation

One of the most important companion tools for community economic development is employment development. Job creation is critical. As a result, a wide variety of approaches have been devised to improve the skill base of distress community residents. These approaches range from basic skills and English language training to job interviewing skills and even new business advice. Community-based job training programs run the gambit from those that provide specific training conducted on the employer's site to

pre-employment counseling and support services. (Fitzgerald 1999).

Job training does not presume that there are, or will be, enough jobs available or suitable for the residents of low income areas. One reason for this is the very high level of ex-offenders among the population. Another is that too many of the best jobs are too far away or the skills requirements too high for the local workforce. Nonetheless it can be argued that increasing the work effort in low income communities, especially moving low income people off welfare is as important to reduce the cycle of dependency. As Lawrence Mead, the architect of welfare to work argues, 'increasing the work effort of dependent populations leads to greater sense of personal worth, a reduction in idleness and crime.' (Mead 1989).

One means of combating the loss of jobs in a community is to help start more local high quality businesses that can serve the community. Low income areas have many of the same service needs as high income areas. Therefore, it makes sense to assist low income residents to start new locally serving businesses. Timothy Bates, a leading small business scholar, suggests that one road to economic revitalization is for more of the low income residents to form new businesses that they own and operate (Bates 1995). This notion is clearly incorporated in the numerous minority ownership programs of the Small Business Administration. Nonetheless, few successful American minorities own or operate businesses in the ghetto (Bates 1995). The notion is that the minority owned and operated business will tend to select staff of color from the community where the firm is located in much higher proportion than their Caucasian counterparts operating in the same locale. Still another reason for promoting minority business ownership in low incomes areas is to create wealth in the distressed community. Real wealth can translate into investments in the same area and new job formation for future generations.

3.3 Social Capital Development

Social skills are the root of economic capacity. It is clear that most economic gains are made through some form of collective action. This can be as simple as joining a club or organization composed of other successful people to forming a corporation. While this seems obvious, it has been overlooked in much recent scholarship. Low income communities are not only bereft of jobs and businesses, they seldom have any organizations. The clearest illustration of this is to examine the town entry signs in small towns across the United States. In well-to-do prosperous communities the city limit sign is accompanied by a larger welcome sign with the names of significant community business and professional organizations such as the Rotary, Lions, Kiwians, Future Farmers, and the Chambers of Commerce as well as similar organizations. These organizations are the bedrock for firm or professional entry into the community. In low income areas of a city or poor small towns there are few or no organizations, other than the church. Unfortunately, the social network necessary to develop and carry an idea to fruition is not present. As a result, one of the latest movements in community economic development is to assist distressed communities build internal networks that can work as systems of mutual assistance within the community.

Forming social capital is sometimes called capacity building among foundations and government agencies. It is a very old notion going at least back as far as the US Department of Agriculture Cooperative Assistance Programs, which started just prior to the Great Depression of the 1920s and 1930s. The basic idea is to help the low income community develop institutional skills similar to the advantaged communities. In most instances this entails the formation of a community development corporation or CDC. The CDC acts as an umbrella agency that organizes local business, assist neighborhoods in social and community services, and helps start community revitalization efforts. The role of CDCs in community economic development is well documented elsewhere. (Vidal 1992). The fundamental idea is to create the CDC, along with a network of other organizations to create a social base for collective action among the disenfranchised in low income residents. This is not an easy task. Lack of trust, envy, and a general sense of apathy must be overcome to trigger community social capital formation.

4. Government Community Economic Development Program

Government programs in community economic development go back to at least the early industrial era. Tenement and poverty area clearance and clean up schemes have been part of national, state, and local level to transform low income areas. The most well recognized community or neighborhood level programs were launched with Jane Adams at Hull House in Chicago at the turn of the twentieth century. The aim of Hull House, later called the Settlement House Programs, was to provide professional interventions in a community in a way that assisted the community leaders in helping themselves. As this idea was refined new programs emerged to accomplish many of the same goals. In England, Local Housing Councils took on similar roles with the low income workers.

Similarly, throughout Europe and particularly in Scandinavia programs were designed in the 1930s to eradicate the worst slum conditions. All of these programs have very similar components. Usually they

provide some basic services for the low income such as housing or home repairs. Additional services are organized by group or community workers to assist the residents form self help devices. These devices range from buying, farming, or business cooperatives to help poor communities meet basic needs for local control and ownership. Today cooperatives have grown into major institutions from this community economic development base. In the United States and throughout Europe, cooperatives operate banks, insurance companies, health services, and even telephone and electric power companies.

Another aspect of government programming has been the designation of special distressed neighborhood or community economic zones. The earliest of these in the United States were community action councils in designated poor neighborhoods in the 1960s. These community action councils were part of the War on Poverty. The concept was that targeting these areas for special government assistance would stimulate business relocations to them and thus increase local employment and capital formation. The most recent incarnation of these zones is called, e.g., enterprise or empowerment. The theory is that altering the tax breaks and employment incentives the area will become attractive for business relocations and expansions. Similar programs have been proposed and adopted in the United Kingdom, Australia, and several Asian countries including China with varying impacts. Green (1991) and several other scholars have questioned the validity of this approach. His work shows that special tax measures are seldom a reason for a firm to change location. In reality there are many factors, chief among them is the quality of nearby human resources (Blakely 1994).

Small and minority and low income business formation has been an important component of virtually all community economic development initiatives. More recently, government program efforts in this area have borrowed from the Bangladesh and the Basque Region of Spain experiences of group lending strategies called microlending. This approach blends both social capital and business formation into a single approach. The micro-loans programs provide small seed capital to a group of borrowers who may enter in different enterprises or may even use their small loan (US $100–US $5000) capital to buy clothes for work. The requirement is pledge and guarantee repayment as a group. The group component creates the social capital that is so frequently missing in other efforts. The group members learn to trust and rely on one another for other networks and to use one another's expertise as needed via bartering arrangements. This approach has had very good results in most of the places that it has been introduced. Another variant of this is the host of government programs that aim at community banking and other capital retention programs. Community Development Banks have been formed with the specific intent and mandate to help create new businesses in low income areas. These banks have government direct support through special underwriting or loans and some funding from traditional banks as part of a government requirement for licensing.

The efforts described above do not represent all government efforts in the field of community economic development. In some cases, government programs have community economic development as a secondary or lower consideration. Further, new government programs are spawned to meet new circumstances such as welfare reform. The range or type of programs described above represents a classical typology rather than an exhaustive list. This type of program is seen in most of the industrialized world in one or more of these forms.

5. *Summary*

Community economic development is a distinct movement. It is part of local economic development; however, it has three distinctly different goals. First, it is based on the concept that there is a link between the conditions of distressed places and those of the people in them. That is, it is important to deal with the social as well as the physical conditions of a place to successfully introduce new economic options and opportunities. Second, it adheres firmly to the notion that interventions in the community or territory must build new capacity or institutional vehicles that have political skills and economic ability to deal effectively with the wider region. Finally, effective community economic development rests on the building of internal–indigenous resources of the local people—through increasing their skills and their capacity to generate their own economic wealth.

See also: Community/Society: History of the Concept; Culture and Economic Development: Cultural Concerns; Development and Urbanization; Development, Economics of; Development: Rural Development Strategies; Development: Social-anthropological Aspects; Development: Socioeconomic Aspects; Local Economic Development; Neighborhood Revitalization and Community Development

Bibliography

Bates T 1995 Political economy of urban poverty in the 21st Century: How progress and public policy generate rising poverty. *The Review of Black Political Economy*. **24**: New Jersey: Transaction Press

Blakely E J 1994 *Planning Local Economic Development: Theory and Practice*. Sage Publications, Newbury Park

Blakely E, Small L 1995 Michael Porter: New Gilder of Ghettos. *Review of Black Political Economy*. **24**(2/3): 161–84

Boston T, Ross C 1995 Response to Michael Porter's model of inner city redevelopment. *The Review of Black Political Economy*. **24**(2/3): 303–36
Daniels B, Tilly C 1981 *Community Economic Development: Seven Guiding Principles Resources*. Congress for Community Economic Development, Washington, DC
Ferguson R, William Dickens W (eds.) 1999 *Urban Problems and Community Development*. Brookings Institutions Press, Washington, DC
Fitzgerald J 1999 Local-community initiatives in job training. *Economic Development Quarterly* **7**(2): 323–45
Goldsmith W, Blakely E J 1992 *Separate Societies: Poverty and Inequality in US Cities*. Temple University Press, Philadelphia
Green R 1991 *Enterprise Zones: New Directions In Economic Development*. Sage Publications, Thousand Oaks, CA
Habersfeld S 1981 Economic Planning in Distressed Communities: The Need to Take a Partisan Perspective. *Berkeley: Economic Development and Law Center Report*.
Hein B 1987 *Strategic Planning for Community Economic Development*. University Cooperative Extension, Ames, IA
Gurwitt R 1992 Neighborhoods and the urban Crisis. *Governing Magazine* **5**(13): 17–20
Malizia E 1985 *Local Economic Development Planning*. Praeger Publications, New York
Mead L 1989 The logic of workfare. *Annals of the American Academy of Political and Social Sciences* **501**: 543–59
Mier R 1993 *Social Justice and Local Economic Development Policy*. Sage Publications, Thousand Oaks, CA
Mier R, Moe K, Sherr I 1986 Strategic planning and the pursuit of reform, economic development equity. *Journal of the American Planning Association* **46**(4): 299–309
Medoff P, Sklar H 1994 *Streets of Hope: The Fall and Rise of an Urban Neighborhood*, South End Press, Boston
Nyden P, Figert A, Shibley M, Burrow D 1994 *Building Community: Social Science In Action*. Pine Forge Press, Thousand Oaks, CA
Sawicki D, Moody M *Deja-vu All Over Again: Porter's Model of Inner City Redevelopment*. Transaction Press, New Brunswick, NJ
Smith M 1988 *Power, Community and City*. Transaction Press, New Brunswick, NJ
Teitz M 1989 Neighborhood economics: Local communities and regional markets. *Economic Development Quarterly* **2**(2): 111
Vidal A 1992 Rebuilding Communities: *A National Study of Urban Community Development Corporations*. New School for Social Research, Milano Graduate School of Management and urban Policy, New York
Wylie J 1990 *Poletown: Community Betrayed*. University of Illinois Press, Champagne, IL

E. J. Blakely

Community Environmental Psychology

Research in community and environmental psychology emphasizes the powerful role of the everyday social settings where individuals live, work, and learn in shaping human development and behavior. This body of research also emphasizes the dynamic nature of the inter-relationship between individuals and social contexts. Social settings can be described in terms of three underlying sets of social climate dimensions: relationship dimensions, personal growth or goal-oriented dimensions, and system maintenance and change dimensions. Research findings on social settings may be organized according to four social ecological principles: environmental press, setting interdependencies, matching models, and a transactional view.

1. Historical Development

Psychological research on social settings has been guided by a *social ecological* perspective. The social ecological perspective emphasizes the powerful role of social context in shaping behavior and the dynamic nature of the inter-relationship between individuals and social contexts. The historical roots of research on social settings trace to Kurt Lewin's *field theory* (Cartwright 1951) developed during the 1930s and 1940s. Lewin argued that behavior is determined by the concrete situation that exists at the time the behavior occurs. During the 1950s, Barker's (1968) *Ecological psychology* presented a conceptual and observational framework for research on social settings. In contrast to psychology's traditional emphasis on person-based variables, Barker showed how the temporal and spatial characteristics of the everyday settings where individuals live, work, and learn mold participants' behavior.

Since the 1960s, both community and environmental psychology have pursued research on contextual determinants of behavior (Walsh et al. 1992). Community psychology focuses on the influence of the social environment, whereas environmental psychology focuses on the influence of the physical environment. The overlap of community and environmental psychology involves the study of social processes in physical settings—the psychology of social settings like the home, the workplace, and the school. Bronfenbrenner's (1989) *ecological* approach to human development has added a fuller understanding of how the characteristics of one type of setting, such as the family, are altered by other factors in participants' lives, such as occupational and educational settings.

Moos (1994) has developed integrated assessment procedures to measure the social climates of family, work, educational, and other social settings. Moos described social settings in terms of three underlying sets of social climate dimensions: relationship dimensions, personal growth or goal-oriented dimensions, and system maintenance and change dimensions (Table 1). Relationship dimensions assess the quality of personal relationships in a setting; that is, how involved people are, how much they help one

another, and how openly they express their feelings. Personal growth or goal orientation dimensions tap the directions in which an environment encourages personal change and development. System maintenance and change dimensions measure how orderly and organized a setting is, how clear it is in its expectations, how much control it maintains, and how responsive it is to change.

Research findings on social settings may be organized according to four social ecological principles: environmental press, setting interdependencies, matching models, and a transactional view. Environmental press and setting interdependencies describe how social settings affect behavior. Matching models and a transactional view describe the dynamic interplay between individuals and social settings.

Table 1
Social climate dimensions in family, work and educational settings

	Relationship dimensions	Personal growth dimensions	System maintenance and change dimensions
Family setting	Cohesion	Independence	Organization
	Expressiveness	Achievement	Control
	Conflict	Intellectual-Cultural	
		Recreational	
		Moral-Religious	
Work setting	Involvement	Autonomy	Clarity
	Co-worker cohesion	Task orientation	Managerial control
	Supervisor support	Work pressure	Innovation
			Physical Comfort
Educational setting	Involvement	Task orientation	Order and organization
	Affiliation	Competition	Rule clarity
	Teacher control		Teacher control
			Innovation

2. *The Psychology of Social Settings*

2.1 *Environmental 'Press'*

Traditional approaches to the prediction of behavior rely predominantly on demographic and person-centered variables. Yet, individuals are profoundly affected by the 'press' of the social settings in which they are embedded. For example, high family support, independence, social integration, and organization are associated with better psychological adjustment and more self-reliant and active coping. Family support may have especially protective psychological effects when family members face life crises and transitions (Pierce et al. 1996).

Child and adolescent development is fostered in families that encourage independence and provide modeling for instrumental and social skills, whereas it can be impeded in families that emphasize achievement in the context of conflict and accommodation to restrictive rules. For example, youth in more cohesive, expressive, well-organized, and socially oriented families are more likely to have higher self-confidence and social competence. Youth in families that value independence and achievement tend to be assertive and self-sufficient, whereas those in supportive and well-organized families tend to have a more even temperament. In contrast, a strong emphasis on achievement in the context of high family structure and a lack of cohesion is associated with depressed mood. Family conflict tends to engender behavior problems and impulsivity.

Four aspects of the work climate have been associated with employee distress and lack of mental and physical well-being: high job demands, insufficient opportunity to participate in decision-making, high supervisor control, and lack of clarity about the job and criteria for adequate performance (Karasek and Theorell 1990). In general, occupational distress is most likely when job demands are high and employees have little discretion in deciding how to meet them. When individuals are allowed to make decisions about their job, high work demands can be stimulating and can foster active problem solving and innovation.

The quality of interpersonal relationships at work can moderate these associations. Supportive relationships with co-workers and supervisors strengthen the salutary effects of autonomy and task orientation and moderate the adverse effects of highly demanding and constrained work settings. In general, co-worker cohesion and supervisor support promote better job attitudes and morale as well as higher work motivation and commitment. However, cohesion without task focus diminishes satisfaction and productivity, whereas the combination of high support and moderate structure promotes these outcomes.

Similarly, the most effective schools are supportive as well as task-oriented (Fraser and Walberg 1991). Educational settings need to emphasize student performance, but not at the expense of engagement and support. Supportive relationships with teachers and classmates and an emphasis on student participation in well-organized classrooms foster student morale, motivation to learn, and academic self-confidence. Students in task-oriented classes that set clear academic goals in the context of supportive relationships are especially likely to do well on standard achievement tests.

Learning environments characterized by competition and control with low support often lead to student anxiety and higher absence rates. Although significant achievement gains can occur in classes that emphasize task performance and competition and are lower in warmth, such classes are less successful in fostering student creativity and continuing motivation to learn. Students in flexible classes that encourage more individual initiative are more willing to work independently and have lower absenteeism.

2.2 Setting Interdependencies

The boundaries between social settings are permeable; the social settings where individuals live, work, and learn are inextricably related to one another (Parke and Kellam 1994). Sometimes the influence of one social setting can amplify that of another setting. For example, youngsters do better in classrooms with rules guiding interpersonal interaction that are similar to those they experience in their families (Booth and Dunn 1996). Youngsters from cohesive families obtain better grades in cohesive classes; those from structured families do better in structured classes; and those from *laissez-faire* families obtain better grades in *laissez-faire* classes. Similarly, stimulating home environments that are oriented toward learning independently help to predict more positive attitudes toward school and better academic achievement. Students in family and educational settings both of which are characterized by high support and structure tend to have the most positive academic self-concepts.

Likewise, the influence of the workplace is an important aspect in understanding family functioning in a broader social context (Eckenrode and Gore 1990). One pattern of work–family interaction involves positive carryover; personal gratification and positive experiences at work can enrich family life. Family members of individuals who enjoy their work tend to experience their families as more involving and supportive. Moreover, because parental work environments can influence the family climate, they may indirectly affect child functioning through their impact on family life. For example, fathers' positive work relationships have been associated with better family relationships and, in turn, fewer child adjustment problems.

At other times the influence of one setting can counteract that of another setting. Thus, family and school effects can be oppositional when there is discontinuity between the home and the school. For example, the academic underachievement and high dropout rate of children from families of low socioeconomic status may be due in part to a discontinuity between their home and school environments. Parents who are less educated are less likely to mirror teaching and learning processes that are similar to those that occur in school, such as praising and interacting with their child around mastering new knowledge.

Similarly, the work–family interface often involves negative carryover from the workplace to family life. Highly demanding, conflict-laden job situations often elicit a pattern of tense family interaction. For example, fathers' work stressors are related to fathers' depression and, in turn, to more family conflict and children's adjustment and physical health problems. Stressors in the workplace are especially likely to generate family distress among individuals who have more family-related demands. Thus, the association between work expectations, such as taking on extra duties and finishing job tasks by doing overtime, and perceived overload tends to be progressively stronger among single employees, those who are married, and those who have children.

3. Person–Setting Interaction

3.1 Matching Models

Considerable evidence documents the powerful connections between features of social settings and individuals' well-being, performance, and personal development. However, part of the influence of social settings depends on the personal characteristics of the individuals who experience them. Thus, according to *person–setting matching models*, adaptation is viewed in terms of the match or congruence between the environmental demands of a particular setting and an individuals' unique pattern of personal vulnerabilities and strengths. For example, individuals with few personal vulnerabilities generally adapt well except in settings with unusually high demands (Fig. 1). Conversely, those with many personal vulnerabilities typically adapt poorly except in settings with exceptionally low demands. Individuals with an intermediate level of personal vulnerabilities tend to adapt poorly in settings that tax areas of vulnerability, but adapt well in settings that emphasize areas of strength.

For example, moderate emphasis on system maintenance factors helps to promote ego control among individuals who need a well-structured setting. However, a strong emphasis on system maintenance factors, especially among developmentally mature and internally oriented persons, restricts individual growth and can foster passivity. Similarly, expressive relationships typically promote morale. However, highly independent or introverted persons who prefer fewer social bonds can feel hemmed in or overstimulated by interaction-oriented settings.

When individuals invest themselves in a setting, they are likely to be especially vulnerable to disruptions in that setting. For example, environmental demands that match an individuals' vulnerabilities in a highly salient setting are associated with more adverse health risks. Involvement in more settings leads to the development of a more complex self-system and provides protection against pressures and disappointments in any one setting.

More generally, social settings tend to maintain or accentuate individual preferences and attitudes that are congruent with dominant setting characteristics. The more intensive, committed, and socially integrated a setting is, the greater is its potential impact. Thus, when person–setting mismatch occurs, homogeneous settings exert the strongest pressure on incongruent individuals to change in the direction of the majority. A heterogeneous setting in contrast has

more diverse influences and provides participants with a wider choice of options.

3.2 A Transactional View

Thus far, we have described a passive process by which people find themselves in settings that mold their behavior. However, people are both influenced by and in turn influence social settings. Individuals select and build environmental niches that maintain and accentuate their dominant personal dispositions. Thus, the psychology of social settings adopts a *transactional view* that emphasizes the reciprocal, two-way interaction between the person and the environment (Stokols and Shumaker 1981).

In part, individuals create microenvironments that then 'reciprocate' by fostering personally preferred attitudes and behaviors. For example, people avoid some settings and, on the basis of their needs and dispositions, choose to participate in others. In turn, these chosen settings are likely to influence people to change in desired ways. Further, individuals' mood and behavior may alter the settings they function in, such as when a depressed individual's hopelessness and lack of enthusiasm leads to a reduction of support from family members and co-workers.

The Conceptual Level (CL) matching model integrates a transactional view with a person–setting matching model. According to the CL model, more mature individuals are better able to organize their environment. Those who are less mature need the stabilizing influence of a well-structured setting. In general, people who want to explore and shape their environment and who exhibit a strong need for independence profit more from less structured environments. For example, internally oriented individuals do better in more flexible environments, whereas externally oriented individuals tend to adjust better in well-structured settings.

Human–environment optimization describes the overall process whereby individuals and groups strive to achieve optimal environments that maximally fulfill their goals and needs. For example, organizational consultants and program evaluators use change-oriented interventions to institute organizational development programs in work and educational settings and clinicians apply such interventions to promote healthy family settings. More generally, through a long-term process of group and community participation and collaborative decision making, each individual can help to shape social values and the ensuing evolution of new settings.

4. Current Directions

Interest in the psychology of social settings has experienced a resurgence in contemporary research on

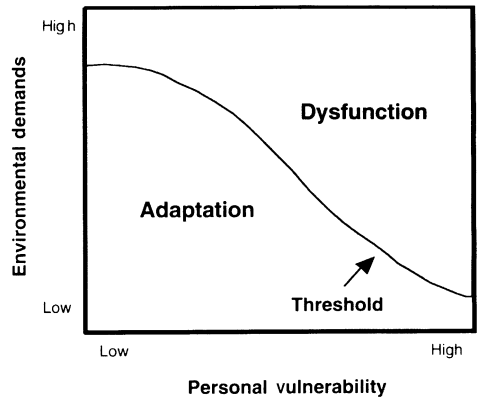

Figure 1
A matching model predicting psychological adjustment as a function of both environmental demands and personal vulnerability (adapted from Zubin and Spring 1977) (copyright 1977 American Psychological Association; reprinted by permission of the author)

psychological stress (Goldberger and Breznitz 1993). The social ecological principles that have guided the study of social settings are central to contemporary models of the stress and coping process. The social settings where individuals live, work, and learn can be significant sources of distress. Stressful settings, such as demanding work environments that allow little control, explain a wide range of behavioral and health outcomes. Moreover, linkages between settings can markedly increase stress responses, such as when a pileup of work demands exacerbates chronic family strains.

At the same time, an appreciation of the dynamic nature of the inter-relationship between people and social settings provides insight into some ways individuals can reduce or avoid distress. Social settings are sources of adaptive resources, such as social support from family members and colleagues at work, that can operate as stress resistance factors when stressors occur. In addition, *stress-diathesis* (vulnerability) models can help to educate individuals to seek settings where their adaptive capacities best match setting requirements. Finally, transactional models of coping (Lazarus and Folkman 1984) underscore how active, problem-focused coping strategies can help individuals to manage stressors successfully and even to grow psychologically through stressful encounters.

See also: Community Health; Community, Social Contexts of; Community Sociology; Environmental Psychology: Overview; Residential Environmental Psychology

Bibliography

Barker R G 1968 *Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior*. Stanford University Press, Stanford, CA

Booth A, Dunn J F 1996 *Family–School Links: How Do They Affect Educational Outcomes?* Lawrence Erlbaum, Mahwah, NJ

Bronfenbrenner U 1989 Ecological systems theory. *Annals of Child Development* **6**: 187–249

Cartwright D (ed.) 1951 *Field Theory in Social Science: Selected Theoretical Papers by Kurt Lewin*, 1st edn. Harper, New York

Eckenrode J, Gore S 1990 *Stress Between Work and Family*. Plenum Press, New York

Fraser B J, Walberg H J 1991 *Educational Environments: Evaluation, Antecedents, and Consequences*, 1st edn. Pergamon Press, Oxford, UK

Goldberger L, Breznitz S (eds.) 1993 *Handbook of Stress: Theoretical and Clinical Aspects*, 2nd edn. Maxwell MacMillan, New York

Karasek R, Theorell T 1990 *Healthy Work: Stress, Productivity, and the Reconstruction of Working Life*. Basic Books, New York

Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York

Moos R H 1994 *The Social Climate Scales: A User's Guide*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA

Parke R D, Kellam S G 1994 *Exploring Family Relationships With Other Social Contexts*. L. Erlbaum Associates, Hillsdale, NJ

Pierce G R, Sarason B R, Sarason I G (eds.) 1996 *Handbook of Social Support and the Family*. Plenum Press, New York

Stokols D A, Shumaker S A 1981 People and places: A transactional view of settings. In: Harvey J H (ed.) *Cognition, Social Behavior, and the Environment*. L. Erlbaum Associates, Hillsdale, NJ

Walsh W B, Craik K H, Price R H (eds.) 1992 *Person–Environment Psychology: Models and Perspectives*. L. Erlbaum Associates, Hillsdale, NJ

Zubin J, Spring B 1977 Vulnerability—a new view of schizophrenia. *Journal of Abnormal Psychology* **86**: 103–26

C. J. Holahan and R. H. Moos

Community, Expression of

Social identities are made up of people's overlapping, mutually reinforcing, competing, and conflicting memberships in a variety of communities. When people claim community membership they are simultaneously expressing their personal and social identities. The meanings of community membership, to both members and outsiders, take on a very intense personal weight as statements about who people are and how they should be treated. Communities are seen as the repositories of values, and the claim to community itself takes on a moral force in much contemporary discourse. In the process of expressing their identities and positions through community membership, and of expressing their judgements of others through *their* inclusion in or separation from community, people deploy markers of community such as locality, kinship, place of origin, nation, ethnicity, religion, age, gender, language, values, or resources as ways to express their position in a moral and social universe. Not all social categories or groups are communities, but when people use what they have in common with others to say who they are, they are expressing community.

Utopian communities have always been explicit and self-conscious expressions of a vision of human nature and potential. Descriptions from Plato to Rousseau and Huxley and practical attempts from the Diggers to the Spanish Anarchists and contemporary communes, have all been occasions for expressing a vision of what people are and should be, and for individuals to express their own potential and aspirations (Kanter 1972). The new communitarians within academia continue this tradition with their call for community formation as a counter to the fragmentation of postmodern life (Etzioni 1993).

1. Moral Expression and the Imagined Community

At least since the Enlightenment, there has been a concern among social philosophers with demographic, cultural, and economic change and the consequent demise of community. Durkheim, Marx, Tonnies, and Weber disagreed about the nature of the transition from community to mass society (*Gemeinschaft* to *Gesellschaft*), and they disagreed about the inevitability and desirability of rending traditional bonds and patterns to replace them with rational calculation and contract. But these classic social theorists all agreed that a profound transformation of social relationships was taking place before their eyes, replacing the ascribed statuses of communities regulated by custom with the statuses achieved by individuals.

In his preface to the second edition of *The Division of Labor*, Durkheim argued that as increasing division of labor expands the scale of society and diminishes the centrality of territorial social organization, new social groups must take the place of territorial units. 'A society composed of an infinite number of unorganized individuals … constitutes a veritable sociological monstrosity'—not only undesirable but impossible. There must be 'a whole series of secondary groups' that can 'drag' individuals into 'the general torrent of social life' (1964, p. 28) and that have 'a moral power capable of containing individual egos' (1964, p. 10). What must be added to Durkheim's formulation is the way that the moral power of communities is not only constraining of individuals but also expresses their identities, fears, and ambitions.

Anderson's (1983) influential concept of the 'imagined community,' with its paradigmatic representation in the newspaper reader who reads alone but is assured of a community of fellow readers, illuminates the ways that community membership expresses identity, to both oneself and to others. Imagined communities share features of reference groups in distinction to interaction groups or face-to-face communities. As Anderson's example of the newspaper reader suggests, developments in technology have undermined whatever geographical basis communities have ever had. Telephone, radio, television, and the Internet have made a commonplace of expressive communities without face-to-face interaction or geographic closeness and have made possible a huge variety of meaningful identities expressed through community with others. However, the impact of technology on community has not been determinative, but rather has been mediated by culture and by individuals.

Though technology has been important in the proliferation of these communities, it is important to remember that dispersed religious and political communities are ancient social forms. The medieval Christian church, the Jewish diaspora, and the community of scholars are familiar examples. Islam has, for over a millennium, been a community that both provides and expresses identity to millions. Mecca has been an orienting location for all practicing Muslims, and for many it has been a place of pilgrimage, but Mecca has been overwhelmingly a symbolic and expressive location. Mecca is the center of the Islamic world not because the Islamic community lives there in face-to-face interaction but because it is the symbolic focus of a community of faith.

When the poet Rupert Brooke wrote: 'If I should die, think only this of me:/That there's some corner of a foreign field/That is for ever England' or when the Psalmist sings of Jerusalem and members of three faiths claim Jerusalem as a sacred city, or when millions of Muslims face Mecca to pray, they are all addressing places that are physical localities, but also deeply elaborated symbols of national or religious communities.

2. *Community and Individualism*

One proposed alternative to physical locality as the basis for community is the ego-centered network. Wellman (1999), for example, argues that the researcher should begin with an individual and work outward, documenting the network of 'personal communities' and mapping the interconnections. Personal communities are 'loosely bounded, sparsely knit networks of specialized ties' (Wellman 1999, p. xiii). Where individuals act as agents constructing these networks, which vary from person to person, quantitative network analysis has become an effective, specialized way of studying them (Fischer 1982).

These personal communities, by definition, have neither the existence independent of a particular individual nor the moral force of the communities envisioned by Durkheim. Ego-centered networks are indisputably central elements of society, but they differ from expressive communities in important ways. In particular, looking at society as made up of networks or communities implies a different emphasis on the relationship between individuals and society (Gans 1988). While network analysis proceeds from individuals whose identity is not problematic, depictions of community stress the ways in which individual identities are constituted by membership and exclusion. It is the morally coercive or normative community expressing a vision of the common good that transcends individual interest that is mourned by those who feel that modern society, particularly in the United States, has lost a sense of community (Bellah et al. 1991).

Throughout history, people have expressed themselves through community. In the nineteenth century United States, a time of great economic and political change, working men and women practiced community with little regard to the ideological implications of their behavior. In their everyday lives, they transcended the tension between the group and the individual by acting in consort—to make bedding for each other at quilting parties, to care for a sick person in their midst while acutely aware they might later demand services in kind, and to create actively a community of accountability and mutuality (Hansen 1994). The gendered division of labor in these communities meant that women exercised a great deal of influence in an arena where power was not denied them, as it was in the public realm. Their actions became self-conscious only when they intentionally mobilized people to become political, as in the abolitionist and temperance movements, or religious, as when they felt God called them to a higher cause and they sought to convert others to Christianity. Through their everyday practices they constructed community and molded the expressive dimensions of their culture. The social sphere created through daily practices was both a source of secure identity and an expression of values capable of exerting discipline and excluding 'outsiders' or 'deviants.'

Seeing community as simultaneously expressive of identities and as something to be expressed reorients discussion away from an easy opposition between individualism and community. Symbolic communities connect people together through a common cause or identity, as in the African-American community (Bruggemann 1995), the community of people with disabilities, or the people listed in the *Social Register*. Symbolic communities, so far from requiring face-to-face interaction, frequently include the dead, as in lineage groups, ancestral lines, the community of martyrs, or communities of scholars who share intellectual 'ancestors.' Shared values or beliefs, ex-

pressed or manifested through ethnicity or nationality, or the symbolism of a shared history or a common cause, have also proven effective bases for community expression and as rhetorical levers in political actions. The symbolism invoked in using the term 'community' may be as much a rallying cry as a description of a social entity.

3. *Methodological Implications*

The opposition between 'community' and 'individualism' seen in much social theory has received a perverse reinforcement from the pervasive individualism of a great deal of social science research, especially in the United States. If 'street corner ethnography' does not capture the spatially dispersed communities of modern societies, research using random samples of isolated individuals also misses important elements of community. By starting with individuals and asking about their attributes, survey research necessarily misses the connections between individuals and the supraindividual social elements that are defined by the *nonindependence* of their members. Communities of orientation, reference, and memory have normative force and behavioral implications, but do not take obvious physical forms. They may, therefore, be invisible to both observers and participants. The same is true of the negotiated and directive, not entirely voluntary, communities facilitated by rapid transportation and easy communication in which people are embedded and that constitute such central features of modern social organization.

For many people in the United States, their high school graduating class provides an example of a community that is neither localized and face-to-face nor entirely voluntary but that still expresses important aspects of their identity. The public school has been the pre-eminent expressive institution of the secular democratic republic, transmitting the moral ideals and ideology of the society. The time children spend in high school is a period of intense face-to-face interaction with a community larger and more diverse than they have known before at an age characterized by identity formation and role confusion.

After the rite of passage of graduation, the face-to-face community becomes spatially dispersed and socially fragmented, but the graduating class endures as a reference group and expression of identity to self and to others. Interviews with Americans who have graduated from high school are studded with comparative references to former classmates who had or had not done well and to those who served as the concrete manifestation of 'people like me' (Townsend 2002). Classmates share histories and normative judgments, and make of themselves, and of each other, characters in morality tales they tell themselves. The meanings of belonging and exclusion are subject to revision and negotiation, but the community has an existence independent of any one of its members. Individuals may attempt to opt out, but the remaining members continue to hold them up to evaluation, and use them to symbolize their own sense of who they are and who they are not.

The salience of high school graduation as an expression of identity in later life is clearly class-specific. Professionals with advanced degrees, for whom high school graduation was a first step rather than a culmination, frequently dismiss the importance of an event which is for many others a significant achievement and the realization of an aspiration. For them, others who graduated from the same professional training institution may constitute a more important community of reference than their peers in high school. The specific features of expressive communities differ between situations and over time, but the social processes underlying them are general.

4. *Contradictions of Community—Gender, Race, and Unequal Access*

While the notion of community necessarily expresses what the members have in common, the concept and rhetoric of community is frequently used to divide, to exclude, and to justify differential treatment and access. British colonialists in Africa, for example, invented traditions, communities, and expressive forms for themselves in order to symbolize their own cohesion in spite of spatial dispersion and to lay claim to a higher social status than they would have been entitled to in Britain itself (Hobsbawm and Ranger 1983, Ngugi wa Thiongo 1981). At the same time, Britain's policy of indirect rule imposed tribal structure on complex indigenous social formations. The language of the tribal community was then used to control Africans' political and economic lives. In the period since the end of colonialism in Africa, ethnicity has replaced tribalism as the dominant expressive language about community, although Africans continue to inhabit a variety of identities. The language of ethnic communities is used to advance, and to conceal, a range of interests and identities based on generation, class, gender, nation, and continued external domination of the continent (Werbner and Ranger 1996).

The dynamic between an inclusive expression of identity and a concealment of internal divisions—'belonging' as a source of community solidarity and as a device to silence dissident expression—is played out in many social movements based in expressive communities. The Black Power and Women's Liberation movements in the 1960s and 1970s are powerful examples of the potential for mobilization on the basis of commonality and shared identity. They selected for people who shared political perspectives and ethnic and gender identities, but the inclusiveness that was their source of strength required that internal difference be obscured by expressions of unity.

Historically, women in the west were excluded from politics and the public sphere, though they found greater opportunities to exercise freedom and power in the social sphere, which was more egalitarian, but not free of gender hierarchies (Hansen 1994). White women mounted a campaign to gain access to the public sphere by focusing on women's suffrage. They called for a united womanhood, and appealed to a common degradation and a shared circumstance. As with activists 100 years later, the suffragists simultaneously excluded those who did not fit their conception of womanhood—women of color, working class women, immigrant women. In the second wave of feminism of the 1960s and 1970s, the similar rallying cry that 'sisterhood is global' was premised on a biased and exclusive identity. The empowerment achieved through mobilization and the increased sense of entitlement wrought discord among women because of the exclusivity that lay at the heart of the movement. Ironically, in reaction to the appeals to a universal womanhood that masked multiple and profound differences, many groups of women organized themselves as members of new communities of women with distinctive identities.

5. *Debate Over the 'Demise of Community'*

Currently, a vocal public and a broad spectrum of policy makers lament community demise. At the same time, social researchers are finding a renewal of civic participation and vitality in communities that appear invisible from a distance. The debate itself is an important expression of how Americans want to define themselves as a culture. Political conservatives worry about losing control over local government and school districts, which they see as repositories of community values. Political progressives worry about the depleted ranks of those wanting to and willing to mobilize for social change. Policy makers and grass roots activists worry about the low density of people in neighborhoods and what that means for the crime rate. With so many people working so many hours, who has the time or inclination to contribute toward the greater good of society, or at least to those beyond the household and kin network?

Observers concerned about the decline of community emphasize its moral centrality and the social values conveyed and expressed through community membership. Wilson (1996) argues that inner-city majority-Black neighborhoods suffer a crisis in community most acutely. With the removal of jobs that pay a living wage, the exit of middle- and working-class workers decreases the cultural and social capital those inner-city neighborhoods need to build and express a sense of community. An economically diverse neighborhood has citizens who model behavior that conforms to the values of the dominant culture and who work within the community to fight crime and cultural malaise and improve services, including public education.

Arguing that the decline of community is a major problem for Americans as a whole, Putnam (2000) provocatively characterizes the crisis in community life as a problem of 'bowling alone.' Putnam's analysis of civic participation, volunteering, sports leagues, and church-going shows a consistent decline over the past half-century. Putnam is also concerned about the loss of the social capital that is generated by invigorating community involvement. On the other hand, Sirianni and Friedland (2001) find an increase in citizen activism at the local level in battles over the environment, healthcare, and the media. They argue that looking only at engagement in 'traditional institutions of civic integration' such as the PTA, the NAACP, and the Elks, misses new forms of civic renewal that are dedicated to problem solving on the local and state levels, that are national in scope, and that express the new communities brought into being by new circumstances.

Concern about the loss of community is not new, but Wellman (1999) asserts that whenever researchers have set out to study communities, they have found them to exist and to be thriving. Perceptive observers have consistently found people expressing their identities through community membership, and asserting their places in communities, in situations where conventional wisdom has pointed to a lack of community—in locations such as suburbs, street corners, homeless shelters, and old people's homes where the physical situation was alleged to make community impossible, or among stigmatized groups who were alleged to lack the cultural resources to express community identity.

All participants in the debate acknowledge the importance of widespread economic change, capital mobility, international movements of people, urbanization, and the increasing entry of women into the paid labor force in transforming the shape of communities. Philipson (2002), for example, asks whether an increasing commitment to corporate workplaces might be colonizing not just workers' time, but their emotional loyalties, eroding independent community life by trying to meet employees' extracurricular needs while absorbing all their waking time. The question remains, however, whether community is declining or transforming.

6. *Continued Vitality of Communities*

The importance of expressive communities increases as locality recedes as the basis for social organization. Since community is essential for human existence, the task is to observe not the 'decline' but the *transformation* of the foundation of communities. In the global movements for the independence of East Timor or the end of apartheid in South Africa, for instance, we see several ways in which communities are both

expressed and expressive. In the face of exile, apartheid divisions into 'tribes' or 'peoples,' and forced prevention of assembly and movement, Black South Africans expressed themselves as a community. The expression of this community helped to forge a united movement. At the same time, international opponents of apartheid expressed moral outrage, and this expression became the basis for a global community. In South Africa and internationally, expression was vital to the creation of other aspects of community such as organizing efforts, congresses, publications, fund raising, and political action. The expression of community, of common interests and identities transcending differences and separations, was one of the most powerful rhetorical weapons of the movement.

Similar mutual reinforcements of symbolic expression and community action occur in a wide variety of situations where the basis for community is the expression of values and identities, and in which, once expressed, the community takes on a life of its own and becomes a potent symbol in its own right. Expressive communities, whether they are long standing or new, confront individuals as already existing, though amenable to change. By their allegiances to these communities, people take on and express a social identity.

See also: Communitarianism: Political Theory; Communitarianism, Sociology of; Community, Social Contexts of; Community Sociology; Individual/Society: History of the Concept; Individualism versus Collectivism: Philosophical Aspects; Moral Economy and Economy of Affection; Moral Sentiments in Society; Solidarity: History of the Concept; Solidarity, Sociology of

Bibliography

Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, New York

Bellah R N, Madsen R, Sullivan W, Swidler A, Tipton S M 1991 *The Good Society*. Knopf, New York

Bruggeman J 1995 Class, race, and symbolic community. *Critical Sociology* **21**: 71–88

Durkheim E 1902 *De la division du travail*, 2nd edn. Paris (1964 *The Division of Labor*. Free Press, New York)

Etzioni A 1993 *The Spirit of Community: Rights, Responsibilities, and the Communitarian Agenda*. Crown Publishers, New York

Fischer C S 1982 *To Dwell Among Friends: Personal Networks in Town and City*. University of Chicago Press, Chicago

Gans H J 1988 *Middle American Individualism: The Future of Liberal Democracy*. Free Press, New York

Hansen K V 1994 *A Very Social Time: Crafting Community in Antebellum New England*. University of California Press, Berkeley, CA

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Kanter R M 1972 *Commitment and Community: Communes and Utopias in Sociological Perspective*. Harvard University Press, Cambridge, MA

Ngugi wa Thiongo 1981 *Detained: A Writer's Prison Diary*. Heinemann, Nairobi, Kenya

Philipson I 2002 *Married to the Job*. New Press, New York

Putnam R 2000 *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, New York

Sirianni C, Friedland L 2001 *Civic Innovation in America: Community Empowerment, Public Policy, and the Movement for Civic Renewal*. University of California Press, Berkeley, CA

Townsend N 2002 *The Package Deal: Marriage, Work, and Fatherhood in Men's Lives*. Temple University Press, Philadelphia, PA

Wellman B (ed.) 1999 *Networks in the Global Village: Life in Contemporary Communities*. Westview Press, Boulder, CO

Werbner R, Ranger T (eds.) 1996 *Postcolonial Identities in Africa*. Zed, London

Wilson W J 1996 *When Work Disappears: The World of the New Urban Poor*. Knopf, New York

N. W. Townsend and K. V. Hansen

Community Genetics

Community genetics is a set of activities at the interphase of medical genetics and community medicine aiming to maximize the number of people benefiting from medical genetics while at the same time minimizing potential harm (Ten Kate 1999).

Clinical disciplines, including clinical genetics, are portrayed by the doctor waiting in his/her office for (a call from) the patient with his complaints and worries. Community medicine, including community genetics, is characterized by an anticipatory, proactive attitude and the large numbers of individuals that are approached. The health care worker, who is worried about the people in the community, leaves his/her office and searches for the people who are at risk, but do not know this yet, or do not know where to go with their worries.

While community medicine brings these characteristics to community genetics, medical genetics brings its expertise in genetics, but also its sensitiveness to ethical issues as respect for autonomy of the individual, beneficence, non-maleficence, and justice.

Although in the above definition the emphasis is on the application side of community genetics, the field has also a strong research side. Research is indispensable to find out what the community needs, and to evaluate the effects of the efforts to meet these needs.

The following activities belong to the field of community genetics.

1. Genetic Screening

Genetic screening is the application of a test on people for the systematic early detection or exclusion of a

hereditary disease or a heritable predisposition to a disease in the participant or in his or her children (see *Genetic Screening for Disease-related Characteristics*).

2. Genetic Education

Genetic illiteracy is widespread. How can people decide on participation in genetic screening programs, pre-pregnancy consultation, or clinical genetics services, if they do not understand the basic principles of genetics, and have misconceptions? There is thus a strong case for sharing basic genetic knowledge with the community (Jallinoja and Aro 2000). The general public should not be the only target, the professionals and policy makers also need to be considered. Apart from genetics, the education program should focus on the meaning of probabilities and on ethical aspects, in which respect for autonomy and solidarity should be stressed.

3. Pre-pregnancy Consultation

Many clients for clinical genetics services are referred when they are pregnant, with problems that better could have been solved before pregnancy. Evaluating a family history, diagnostic work-up of consultants and family members, and consultation of other experts may take more time than the ongoing pregnancy allows. During pregnancy there are less options available for the pregnant woman and her partner, and decisions have to be taken in a short time. The options are not restricted to prenatal diagnosis, termination of pregnancy or continuation, but also include the possibility of refraining from pregnancy at all, and sometimes also pre-implantation genetic diagnosis, artificial insemination by donor, or egg-cell donation. Moreover, some preventive measures have to start before pregnancy. Folic acid supplementation for the prevention of neural tube defects such as spina bifida or anencephaly, for example, should be started at least four weeks before conception and continue for eight weeks after conception. Controlling maternal diabetes and choosing medications, if needed, that are the least harmful to the expected child, should also start before conception. The same applies to correcting dietary insufficiencies and refraining from smoking, alcohol and drug consumption. Even if there is no indication for increased risk pre-pregnancy counseling would give opportunities for education about risks of advanced parental age, or over consumption of vitamin A, and about prevention of, e.g., toxoplasmosis and neural tube defects. Notwithstanding all these advantages, a systematic offer of pre-pregnancy counseling to all couples who wish to have a child has nowhere been effectuated on a satisfying scale (Schrander-Stumpel 1999).

4. Genetics in Primary Care

Clinical genetics centers are scarce and have few resources to meet such needs as pre-pregnancy consultation for the whole population, or to answer all questions people might have about genetics. Pre-pregnancy consultation should be readily available in the neighborhood of the couples considering pregnancy. The same applies to answering genetic questions that might arise outside the reproductive setting. In developed countries, primary care is very accessible and is an obvious first line for handling these tasks. This is especially the case in countries were primary care is the gatekeeper to facilities, such as clinical genetics (Hayflick and Eiff 1998). With the discovery of hereditary cancer genes like BRCA1 and BRCA2, the number of questions to primary care physicians because of a family history of cancer, have increased markedly. However, there is a problem. The developments in the field of genetics are fast, and there are only a limited number of routine situations. Education of professionals about genetics alone will not be sufficient, since its content will soon be outdated. A solution is to exploit the possibilities of the Internet for the support of primary care.

5. Genetics for Disadvantaged People

The majority of people are not living in developed countries. For these people, the education of primary care physicians is not a solution. Other approaches are needed for these people. An example of one other approach is the genetic nursing system that has been developed in South Africa, where some remote hospitals, which generally have a serious lack of medical doctors, have at least one local nurse who has received some training in genetics (Christianson et al. 2000). These nurses, who have substantial contact with the local population, can recognize frequent disorders and perform genetic counseling. More difficult cases are presented to a hospital-visiting clinical geneticist.

Disadvantaged populations are not only living in the third world, but are present in developed countries as well. Groups of immigrants with a culture and language that is different from the resident population do not easily turn to clinical genetics services. Belief systems and social systems, religious and ethical convictions, traditions and usual practice are possible impediments. Special attention is needed for immigrants to also profit from genetic information (Penchaszadeh and Puñales-Morejón 1998).

6. Access and Quality of Clinical Genetics Services

Although clinical genetics can be clearly distinguished from community genetics, the presence of clinical

facilities, their accessibility and their quality is a concern of community genetics (Harris and Reid 1997). Genetic education, genetic screening, prepregnancy consultation, genetics in primary care, and genetics for disadvantaged people all presuppose that there are answers to difficult questions and for genetic diagnosis which can be easily found.

7. *Monitoring*

Monitoring of the population with regard to genetic and congenital diseases, but also for unwanted potential side-effects of the use of medical genetics, e.g., insurance problems, is desirable. Monitoring can be done by continuous registration or periodic surveys. When timely detection of sudden changes in frequencies are of interest, continuous registration is preferable. Periodic surveys, however, in general are less expensive. A problem with registers of congenital anomalies is the relative rareness of many of the disorders. This limits the statistical power of individual registries. Power may increase by joining international networks, such as the European Registration of Congenital Anomalies (EUROCAT) or the International Clearinghouse of Birth Defect Registries (ICBDR) (Reefhuis et al. 1999).

8. *Conclusion*

Community genetics is a multidisciplinary field. It differs from clinical genetics by its community orientation. It is not population genetics or genetic epidemiology. Population genetics and genetic epidemiology serve to increase our understanding about the behavior of alleles in populations and the contribution of genes to disease. Both fields produce information that is extremely important for community genetics. Community genetics applies this knowledge, together with the contributions from clinical genetics, psychology, sociology, educational experts, law, ethics, patient organizations (Kent 1999), and the community.

The term 'community genetics' is a new one, first mentioned in literature in the 1990s (Modell and Kuliev 1998). Separate community genetic activities, however, have a longer history. The term is useful since it gives a clear label for a coherent field within the universe of health care. It helps to delineate clinical genetics and it conveys a positive answer to those who express fear for eugenic tendencies and genetization. This answer is: 'Yes, it is possible to apply genetics to the benefit of the community by serving the individuals in it, without intention to improve the gene pool or the health of future populations.' This is one of the reasons why the name community genetics should be preferred above the term 'public health genetics' (Khoury et al. 2000). The other reason is that classical outcome measures in public health are uptake, compliance, and reproductive choices. Community genetics prefers to measure outcome in terms of improvement of autonomy, empowerment of people, perception of benefit, decrease of harm, and increase in justice.

See also: Epidemiology, Social; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease-related Characteristics; Genetic Testing and Counseling: Educational and Professional Aspects; Mental Illness, Epidemiology of; Mental Illness, Genetics of; Public Health; Public Health as a Social Science; Schizophrenia and Bipolar Disorder: Genetic Aspects

Bibliography

Christianson A L, Venter P A, Modiba J H, Nelson M M 2000 Development of a primary health care clinical genetic service in rural South Africa—The Northern Province experience, 1990–6. *Community Genetics* **3**: 77–84

Harris R, Reid M 1997 Medical genetic services in 31 countries: An overview. *European Journal of Human Genetics* **5**(suppl. 2): 3–21

Hayflick S J, Eiff M P 1998 Role of primary care providers in the delivery of genetics services. *Community Genetics* **1**: 18–22

Jallinoja P, Aro A R 2000 Does knowledge make a difference? The association between knowledge about genes and attitudes toward gene tests. *Journal of Health Communication* **5**: 29–39

Kent A 1999 The role of voluntary consumer organisations in genetic services in the United Kingdom. *Community Genetics* **2**: 156–61

Khoury M J, Burke W, Thomson E J (eds.) 2000 Genetics and public health in the twenty-first century. Using genetic information to improve health and prevent disease. *Oxford Monographs on Medical Genetics*. Oxford University Press, Oxford, UK, Vol. 40

Modell B, Kuliev A 1998 The history of community genetics: the contribution of the haemoglobin disorders. *Community Genetics* **1**: 3–11

Penchaszadeh V B, Puñales-Morejón D 1998 Genetic services to the Latino population in the United States. *Community Genetics* **1**: 134–41

Reefhuis J, Zandwijken G R J, De Walle H E K, Cornel M C 1999 Birth defect and risk factor surveillance in the Northern and Southwestern Netherlands. *Community Genetics* **2**: 97–108

Schrander-Stumpel C 1999 Preconception care: challenge of the new millennium? *American Journal of Medical Genetics* **89**: 58–61

Ten Kate L P 1999 The concept of community genetics. *Community Genetics* **2**: 152

L. P. ten Kate

Community Health

Although the term 'community health' is used frequently in health-related textbooks, scientific articles, and publications of international organizations, pre-

cise definitions of 'community health' are hard to come by. The term is often not clearly distinguished from related terms such as 'community medicine,' 'community health care,' 'preventive medicine,' or even 'public health.' Nevertheless it is possible to identify essential elements in the general concept of community health which are common to most uses and which differentiate 'community health' from traditional biomedical and individual-based approaches to health. This article reviews the uses of the term 'community health' in the health field, discusses common themes implicit in its use, and summarizes some of the ways the concept has been reflected in recent research and practice.

1. *Multiple Uses of the Term 'Community Health'*

Perhaps because the term was originally adopted in the medical field, it has often been used to refer primarily to health services provided in the community (for example in community health centers or by means of health visitors) as opposed to more traditional institutional settings such as hospitals. Early on it was also used to refer to the provision of health services by the community (i.e., through services organized and funded by the community itself, rather than through the mechanisms of traditional private medical practice), especially services intended to prevent or control the spread of disease and attend to the 'medically indigent.'

Other uses of the term community health have been much broader. The term community health or community health care has also been used as a synonym for the community-oriented primary care movement pioneered by the work of Sydney Kark in South Africa in the 1940s and 1950s (Kark 1981). Community-oriented primary care involved the provision of health care in geographically defined communities through a coordinated system of community health centers. However, it was conceived as much broader than the simple provision of health services in or by the community. The approach included community participation and organization, a multidisciplinary and team practice, a broad approach to the determinants of ill health including social and cultural together with biological factors, and built in monitoring, evaluation, and research (Susser 1999).

Community health (or community health care) has also been used as a synonym for another imprecisely defined term, community medicine, which emerged in the 1960s in the United States as part of a general movement in academic centers towards 'community-based approaches' to health care. Community medicine was defined as the assessment of health needs and provision of health care to defined population groups as distinct from the prevailing system based on the individual patient. It implied the provision of both curative and preventive services as well as the use of epidemiologic techniques in assessing the needs of population groups, in setting priorities and in the assessment of the results achieved (Lathem 1976). The community medicine movement was associated with the creation of departments, centers, and programs in community medicine (or community health) in many universities. Although multidisciplinary teams were sometimes emphasized, community medicine was clearly conceived as a domain or subspecialty of the medical profession. In Great Britain, the term community medicine was officially established in the early 1970s as a distinct medical specialty and defined as a branch of medicine concerned with populations or groups rather than with individual patients, requiring special knowledge of epidemiology, organization and evaluation of medical care, and the medical aspects of health service administration.

The term community health (and community health promotion) has also been used to refer to the local, collaborative efforts of the public and private sectors in the prevention and treatment of disease. This perspective highlights the participation not only of the medical profession and allied health professions but also of schools, families, workplaces, and the public sector (Green and Ottoson 1999). Key characteristics of this use of the term are notions of local diagnosis and action as well as intersectoral and collaborative approaches between public and private partners. This conception distinguishes community health from traditional clinical practice on one hand (which focuses on the individual patient) and from population health (or public health) on the other hand (which focuses on the population as a whole).

Although sometimes distinguished from 'population health' because of its focus on local diagnosis and action, community health has also been used to refer to the health of populations generally. In this sense the term is synonymous with modern uses of the term public health. Originally 'public health' was sometimes used to refer to health services provided by governmental agencies or to government prevention campaigns (such as vaccinations) and environmental interventions (such as sanitation). More recently, however, public health has been used more broadly to refer to 'the combination of sciences, skills, and beliefs that is directed to the maintenance and improvement of the health of all the people through collective or social actions' (Last 1995).

2. *Common Themes Underlying the General Concept of 'Community Health'*

Despite the many different and often imprecise uses of 'community health' in the literature, it is possible to trace several underlying themes implicit in its use. The first concerns how health and disease are concep-

tualized: health is a state of complete physical, mental, and social wellbeing and not merely the absence of disease or infirmity (WHO 1978). In addition, health and disease are perceived as extremes of a continuum of health states, rather than as dichotomous categories. The consequences of this conceptualization of health and disease are that interventions to improve health should include not only the diagnosis and treatment of disease (usually emphasized in traditional medical practice) but also health promotion and disease prevention activities.

A second theme refers to the determinants and causes of ill health. Community health approaches take a broad view of the causes of ill health-including not only those emphasized by traditional biomedical approaches such as individual susceptibility, individual exposures to disease-causing agents, and health behaviors (e.g., smoking, drinking, diet, and exercise), but also characteristics of the physical and social environments of the places where people live and work. Health and disease are therefore no longer thought of as matters exclusively (or even predominantly) of individual characteristics or individual choice, but rather as conditioned by broader social and economic forces which affect the resources that individuals have and constrain the choices that they can make. For this reason community health approaches often emphasize intersectoral approaches to health promotion and disease prevention and treatment. Improving health requires not only traditional health care interventions (early detection, diagnosis and treatment) but also health education (e.g., through health promoters, the school system, or mass communication), interventions on the physical environment (e.g., sanitation, pollution reduction, improvement of recreational spaces), reorganization of work (reduction in exposure to noxious physical or chemical environments, improvement of the psychosocial work environment), and even economic and social policy interventions. Thus a community-health approach requires collaboration of the traditional health sector and public health agencies with many other sectors.

A third characteristic of community health which stems directly from the recognition of the multiple causes of disease is the emphasis on a multidisciplinary approach to the study of health problems and to their solution, bringing together not only traditional health practitioners (such as doctors and nurses) and biological scientists, but also public health practitioners, epidemiologists, health educators, and social scientists. Community health integrates personal health services with environmental, health policy, and social policy approaches, and draws on both quantitative and qualitative research.

A fourth underlying theme in community health emphasizes the importance of local determinants and local action. While recognizing society-wide determinants of health, the importance of local determinants, and the need for local needs assessment, local decision making, local priorities, and local action is stressed. The emphasis on local diagnosis and action is closely linked to the importance of local control and the need to recognize the specificities of communities in both assessing needs and designing and implementing interventions.

A fifth characteristic is the key importance of community participation, that is the active involvement of the community in all stages of the community health process, including identification of health needs, establishment of priorities, development and implementation of interventions, and evaluation of their effects. Community participation implies more than sporadic consultation with community leaders or participation of community members in the implementation of interventions. It requires direct involvement of the community, its representatives or its institutions in all aspects of the decision making process. However, the importance of community participation, and the extent to which true participation in the decision-making processes has actually existed in community health programs, has been the subject of much debate.

Finally, and perhaps most importantly, use of the term community implies recognition of the construct of community, or more generally of populations, as fundamental in understanding the causes of ill health. However community is defined, there is implicit recognition of the role of social interactions and social organization in shaping the attitudes of individuals and the choices that they make. The determinants of ill health can be conceptualized not only at the individual-level but also at the population, group, or community level. This notion is in stark contrast to the dominant paradigm in medicine where methodological individualism and the idea that disease itself and all its determinants are best conceptualized and measured at the individual-level is paramount. Thus, interventions can be targeted at communities themselves (as well as at community organizations or institutions), rather than focusing on individuals as is done in traditional medical practice. Another key idea that distinguishes the community health approach from that of traditional medical practice is that the object of study (and intervention) is not only the individual patient but also the community or population. Thus the practice of community health requires the ascertainment of health problems and needs not only in individuals but also in communities or populations.

3. *The Concept in Recent Research and Practice*

Many of the underlying themes of community health have appeared in health-related research and in health care and public health practice over the past few years. Three research-related areas in which elements of the community health paradigm have emerged are the investigation of community-level determinants of

health, the evaluation of community interventions (community intervention trials), and the notion of 'community-based' research. In public health practice, community health themes have formed the basis of the primary health care approach and have recently emerged in some aspects of the healthy cities and healthy municipalities movements.

3.1 Community-level Determinants of Health

The 1990s saw an increasing empirical investigation of community-level determinants of health and the extent to which community characteristics are related to health over and above the effects of individual-level factors. Research on community effects is part of a resurgence of interest in the group-level, macrolevel, or ecological determinants of health. It is related to a larger critique of the dominant paradigm in health research which implicitly assumes that all disease determinants are essentially individual-level attributes (Duncan et al. 1998, Diez-Roux 1998). Recent empirical investigations of community effects on health have focused on geographically-defined communities or neighborhoods. For example, a series of studies have now documented a relation between neighborhood socioeconomic context and health outcomes, including mortality, disease prevalence, and health behaviors (e.g., Haan et al. 1987, Davey Smith 1998, Robert 1999). These associations often persist after controlling for the socioeconomic characteristics of individuals, suggesting that characteristics of areas or neighborhoods *per se* may be important to health.

The recent development of the statistical technique of multilevel analysis, which allows the simultaneous examination of group-level and individual-level factors (Duncan et al. 1998, DiPrete and Forristal 1994), has greatly stimulated empirical research in this field. However, there has been relatively little conceptualization or empirical examination of the specific features of areas that may be relevant or of the mechanisms involved (MacIntyre et al. 1993). Potential intervening factors may include the physical characteristics and resources of neighborhoods as well as features of their social environments. In addition, research in this field still faces important conceptual and methodological challenges before firm inferences can be drawn regarding the presence and relative importance of neighborhood effects (particularly with regard to the role of individual-level social position). These challenges include better definition and characterization of communities or neighborhoods, specification of their relevant features, elucidation of the processes through which neighborhood factors may operate, and examination of the ways in which neighborhood factors may interact with individual-level characteristics.

3.2 Community Interventions

Closely linked to the notion of community determinants of health is the idea that interventions to improve health and prevent disease may be more effective if they target communities rather than (or in addition to) the individuals within them. Although the importance of community interventions such as improved sanitation has long been a basic tenet of public health, the idea that interventions to prevent chronic diseases such as cardiovascular disease should also focus on the community level only emerged in the last two decades. The rationale underlying the need for community interventions is that many health-related behaviors (such as smoking, sedentary lifestyle, and diet) related to today's prevalent chronic diseases are promoted and maintained by the physical and social environments of communities. Thus, by addressing community environments directly, and by modifying prevalent social norms, community interventions may be more effective than traditional individual-based one-on-one approaches (Syme 1986). In addition, by targeting the entire community rather than just individuals above an arbitrarily defined 'high risk' cut off, community interventions have the potential for shifting the entire distribution of risk and thus preventing a larger amount of disease (Rose 1985). The community intervention approach represents a radical departure from the traditional clinical approach which focuses only on individuals at highest risk and 'decontextualizes' risk behaviors by ignoring their broader social determinants.

Community intervention trials or their variants have been conducted in a variety of settings over the past 20 years including geographically defined communities, workplaces, and school (Sorensen et al. 1998). Examples of the types of interventions implemented as part of community intervention trials include health education activities, mass media campaigns, education of health professionals, screening, physical activity programs, and labeling in grocery stores and restaurants. Although some community intervention studies for the prevention of cardiovascular disease have found significant effects (Puska et al. 1985), the empirical evidence of program effects in these studies is generally modest (Koepsell et al. 1995, Fortmann et al. 1995). The presence of society-wide trends that make it difficult to detect program effects, the relative absence of interventions at the structural, macro or policy level, and the lack of significant community participation may account for the mixed results of some of these community trials.

3.3 Community Participation and Community-based Research

Many of the ideas underlying the concept of community health are also expressed in what has been

termed 'community-based research' (or 'community-centered research') (Israel et al. 1998). A key characteristic of this approach is the explicit recognition of the concept of community as a fundamental category in public health research, where community 'is characterized by a sense of identification and emotional connection to other members, common symbol systems, shared values and norms, mutual, although not necessarily equal, influence, common interests and commitment to meeting shared needs' (Israel et al. 1998). The notion of community includes not only geographically defined neighborhoods (which may or may not qualify as communities under the previous definition) but also, for example, geographically dispersed ethnic groups with a sense of common identity. Key elements of community-based research include: (a) building on strengths, resources and relations that exists within communities as part of health research; (b) collaborative partnership and the active participation of community members and representatives in all aspects of the research process; (c) the reciprocal transfer of knowledge, skills, capacity, and power; (d) the integration of knowledge and action, and (e) a broad approach to the determinants of health including not only the characteristics of individuals but also community characteristics and broader structural factors. Although often noted as an ideal approach to research in public health, true community participation has been infrequently implemented in health research, so its impact has yet to be determined.

3.4 The Primary Health Care and Healthy Cities/Communities/Municipalities Movements

Many of the constituent elements of 'community health' were reflected in the 'primary health care' approach formalized by the World Health Organization in 1978 (WHO 1978), which drew in part on community-oriented primary care (Kark 1981). As defined by WHO, primary health care extends beyond the provision of basic health services or first contact clinical care to environmental, economic, and social interventions which may have an impact on health, and emphasizes community participation and the need to eliminate inequalities both between and within countries. Primary health care was put forward as the primary strategy for achieving the goal of health for all, and was adopted by the health sectors of many nonindustrialized countries. Elements of community-oriented primary care were also adopted in some industrialized countries as illustrated for example, by the community health center program in the 1960s in the United States (Sardell 1988). Themes of the 'community health' paradigm have also recently emerged in the Healthy Cities, Healthy Communities, and Healthy Municipalities movements initiated in Canada in the 1980s, and subsequently adopted in the United States and by the Panamerican and World Health Organizations. Although the philosophical orientation, sponsors, and activities of these projects have varied substantially, they share the basic idea of challenging communities to develop projects to reduce inequalities in health and access to health care, and to develop healthy public policies at the local level through a multisectorial approach and increased community participation in health decision-making (Flynn 1996). They emphasize local government, decentralization, community involvement, and developing a broad range of strategies to address the social, environmental, and economic determinants of health. The evaluation of the effects of primary health care and related strategies on health outcomes is complex. Empirical studies are scarce, although there is some evidence that comprehensive and integrated approaches (such as those advocated by the primary health care movement) may be more effective and efficient in improving health overall than selective and vertical approaches (Taylor and Parker 1987).

4. Limitations and Implications of the Concept of Community Health

Ambiguities in exactly what 'community health' is intended to mean has led some to urge caution in the use of the term. For example, community health has often been associated with the organization and delivery of health care for the poor, despite the fact that it has been postulated as a framework for understanding the causes of ill health and a strategy for improving health in the population as a whole. The concept of 'community participation' has also been misinterpreted. For example, health education campaigns in the community, periodic gatherings with community members to inform them of ongoing activities, or even simply hiring community members for routine tasks at the health center has been taken as an indication of the 'participatory' nature of the program. There has been much debate around the concept of community participation itself, whether participation has been more than simply symbolic, and the extent to which it has been distorted, manipulated, and rendered rethorical (Ugalde 1985). The relationship between 'community health' and medical practice has also been contested. For example, despite formulations of the concept that emphasize the multiple determinants of health and the need for intersectoral approaches, community health has often been viewed as a branch or specialty of medicine, as clinical practice with a population or community base (where population or 'community' may refer simply to the group of patients that a particular medical practice sees).

The use of the term 'community' has also been problematic because 'community' is rarely precisely

defined, and has sometimes been used to refer to any group of persons. Moreover, the use of 'community health' in place of 'population health' or 'public health' has been criticized because the term 'community' conjures up ideas of consensus and equality (an idealized notion of community which may often not exist in reality), and in so doing may obfuscate the fact that population health (and public health) is strongly influenced by inequality (Popay 1996). It has also been argued that the use of 'community health' as opposed to 'public health' or related terms such as 'social medicine' has contributed to a desocialization and parochialization of the concept. The use of community medicine (or community health care) as a synonym of public health perpetuates the idea of public health as a subdivision of medicine rather than medicine as a subdivision of the broader public health (Anonymous 1985). Others have noted that by focusing on 'community' the concept emphasizes local action and shifts attention from the need to address larger scale social and structural determinants of health. The emphasis on decentralization in some formulations of the community health paradigm has not always been accompanied by true changes in decision-making power. In addition, it has often been associated with ideas of financial autonomy and self-sufficiency, sometimes with the intent of reducing government outlays, with potential consequences for the availability of resources at the local level and equity. In Latin America, the concept of 'collective health' which emerged in Brazil in the 1970s (*saúde colectiva* in Portuguese) has been used as an alternative to the terms 'public health' and 'community medicine,' reflecting a critique of the perceived biomedical reductionism and burocratization of the 'public health' and 'community medicine' approaches dominant in developed countries. The 'collective health' movement emphasized the social nature of the health-disease process, as well as the contribution of the social sciences (together with other disciplines) to knowledge and practice in the field (Silva Paim 1992).

5. Conclusion

Despite many ambiguities and reductionisms, the broadly defined philosophy of community health has been progressive in the health field because it recognizes the role of 'community' or more generally of interactions between persons in groups, social relations, and social organization in influencing people's health. Thus, it transcends the approach dominant in clinical medicine which reduces health and its determinants to a biological and individual phenomenon, and recognizes that each person is a member of multiple groups, communities, or populations (e.g., social classes, racial and ethnic groups, occupational groups, neighborhoods) and that characteristics of these groups are crucial in shaping that person's health over and above his or her individual characteristics. Moreover, it emphasizes the need for a multidisciplinary and intersectoral approach. Under the community health paradigm, the prevention of illness and death is no longer viewed purely as a matter of medical intervention, it is no longer a monopoly of doctors and health care providers. Health and disease are viewed not only as biological, but also as fundamentally social phenomena. Because of their recognition of the social and economic determinants of health, their concern with equity, and their strong emphasis on community participation, community empowerment, and community action, community health movements have been frequently linked to broader social movements, and to calls for social change and political action.

See also: Ecology and Health; Epidemiology, Social; Health Care Delivery Services; Health Education and Health Promotion; Health Interventions: Community-based; Public Health as a Social Science

Bibliography

Anonymous 1985 The distinction between public health and community/social/preventive medicine. *Journal of Public Health Policy* **6** (4): 435–9

Diez-Roux A V 1998 Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health* **88**: 216–22

DiPrete T A, Forristal J D 1994 Multilevel models: Methods and substance. *Annual Review of Sociology* **20**: 331–57

Duncan C, Jones K, Moon G 1998 Context, composition, and heterogeneity: Using multilevel models in health research. *Social Science & Medicine* **46**: 97–117

Flynn B C 1996 Healthy Cities: Toward worldwide health promotion. *Annual Review of Public Health* **17**: 299–309

Fortmann S P, Flora J A, Winkleby M A, Schooler C, Taylor C B, Faquhar J W 1995 Community intervention trials: Reflections on the Stanford Five-City project experience. *American Journal of Epidemiology* **142**: 576–86

Green L W, Ottoson J M 1999 *Community and Population Health*. McGraw Hill, Boston

Haan M, Kaplan G A, Camacho T 1987 Poverty and health: Prospective evidence from the Alameda County Study. *American Journal of Epidemiology* **125**: 989–98

Kark S L 1981 *The Practice of Community-oriented Primary Care*. Appleton-Century-Crofts, New York

Koepsell T D, Diehr P H, Cheadle A, Kristal A 1995 Invited Commentary Symposium on community intervention trials. *American Journal of Epidemiology* **142**: 594–9

Israel B A, Schulz A J, Parker E A, Becker A B 1998 Review of community based research: Assessing partnership approaches to improve public health. *Annual Review of Public Health* **19**: 173–202

Last J M (ed.) 1995 *A Dictionary of Epidemiology*, 3rd edn., Oxford University Press, New York

Lathem W 1976 Community medicine: Success or failure? *New England Journal of Medicine* **295**: 18–23

MacIntyre S, MacIver S, Sooman A 1993 Area, class and health: Should we be focusing on places or people? *Journal of Social Policy* **22**: 213–34
Puska P, Nissinen A, Tuomilehto J 1985 The community-based strategy to prevent coronary heart disease: Conclusions from ten years of the North Karelia Project. *Annual Review of Public Health* **6**: 147–93
Popay J 1996 Community health: The risks of confusion and complacency. *Journal of Epidemiology and Community Health* **50**: 6–14
Robert S A 1999 Socioeconomic position and health: The independent contribution of community socioeconomic context. *Annual Review of Sociology* **25**: 489–516
Rose G 1985 Sick individuals and sick populations. *International Journal of Epidemiology* **14**: 32–8
Sardell A 1988 *The U.S. Experiment in Social Medicine. The Community Health Center Program, 1965–86*. University of Pittsburgh Press, Pittsburgh, PA
Silva Paim J 1992 Collective health and the challenges of practice. In: *The Crisis of Public Health: Reflections for the Debate*. Scientific Publication No. 540. Panamerican Health Organization, Pan American Sanitary Bureau, Regional Office of the World Health Organization,Washington, DC
Smith G, Hart C, Watt G, Hole D, Hawthorne V 1998 Individual social class, area-based deprivation, cardiovascular disease risk factors, and mortality: The Renfrew and Paisley study. *Journal of Epidemiology and Community Health* **52**: 399–405
Sorensen G, Emmons K M, Hunt M K, Johnston D 1998 Implications of the results of community intervention trials. *Annual Review of Public Health* **19**: 379–416
Susser M 1999 Pioneering community-oriented primary care. *Bulletin of the World Health Organization* **77**(5): 436–8
Syme S L 1986 Strategies for health promotion. *Preventative Medicine* **15**: 492–507
Taylor C E, Parker R L 1987 Integrating PHC services: Evidence from Narangwal, India. *Health Policy and Planning* **2**: 150–61
Ugalde A 1985 Ideological dimensions of community participation in Latin American programs. *Social Science & Medicine* **21**: 41–53
World Health Organization 1978 Primary Health Care: A joint Report. *International Conference on Primary Health Care, Alma-Ata, Kazakh, S.S.R.*, World Health Organization, Geneva, Switzerland

A. V. Diez Roux

Community Organization and the Life Course

Communities and, more specifically, neighborhoods are one of several central social settings that affect human development. Approaches to neighborhood theory and measurement are first reviewed, followed by a discussion of neighborhood effects on children and adolescents and on adults and the elderly. The article concludes by highlighting new directions for research, policy, and practice.

1. *The Ecology of Human Development*

A fundamental proposition of developmental science is that social contexts affect human development. Most theories on social contexts seek to identify which attributes of a single context are developmentally positive or negative. Most theories emphasize the composition and structure of a setting. While this approach is useful, it is limited in that it says little about the processes that have a more direct impact on individual development. For theories to be maximally predictive and generalizable, they must also address processes. This is especially important because some effective or ineffective processes may be more common in settings with certain compositional or structural characteristics. Ultimately, attention must also be paid to the ways in which multiple contexts jointly affect development during particular life periods and over the life course as a whole. The approach advocated here is therefore in line with recent advances in ecological, life-span, and life-course frameworks (for a review, see Settersten 1999). These frameworks emphasize the multiple proximal and distal settings in which development takes place, the connections between these settings, the proximal and distal processes that occur within them, and how these settings and processes change over time.

2. *Theoretical Models of Neighborhood Effects*

Neighborhood effects on physical, psychological, and social outcomes may be tied to the composition and structure of neighborhoods, the social processes that occur within them, or specific combinations of composition, structure, and process.

Several theoretical models have guided most inquiry in this area, namely contagion models, models of social disorganization, models of collective socialization, competition models, models of relative deprivation, and institutional models (see also Furstenberg and Hughes 1997, Mayer and Jencks 1989). These and other models need not be contradictory; indeed, they may be complementary.

'Contagion' or 'epidemic' models suggest that those exposed to neighbors who engage in negative behaviors will themselves be more likely to engage in similar behaviors (e.g., Crane 1991). In these models, the assumption is that neighborhoods of low socioeconomic status (SES) are characterized by more problematic behaviors. Presumably, this also extends to attitudes, beliefs, and values. The focus of these models has almost exclusively been on the contagion of negative phenomena, not positive phenomena.

Models of 'social disorganization' suggest that neighborhoods with high levels of social problems become disorganized, which in turn results in deviant behavior at the individual level (e.g., Skogan 1990). Recent work on 'collective efficacy' has advanced

these models (e.g., Sampson et al. 1997). Social organization is bound explicitly to collective efficacy, which depends on neighbors trusting each other, sharing values, jointly monitoring and supervising youth, and intervening on behalf of the common good. As such, these also become models of 'collective socialization' in that the entire community takes responsibility for socializing its residents.

Models of 'relative deprivation' or 'social comparison' are models in which the presence of advantaged neighbors, in particular, becomes problematic for those who have less because individuals evaluate their own standing relative to that of others. For example, if children and families view themselves to be at a disadvantage, they may not be motivated to achieve and they may labeled negatively by their peers (Furstenberg and Hughes 1997). These models may also emphasize 'competition' or 'cultural conflict,' in that neighbors or peers are viewed as competing for limited resources.

Finally, 'institutional' models suggest that it is not the kind of neighbors that matters most, but instead the resources, services, and organizations available in a neighborhood (e.g., McLaughlin et al. 1994). Proponents of these models generally assume that higher SES neighborhoods have greater and higher quality organizations, services, and resources than lower SES neighborhoods, and that these are ultimately very beneficial for human development.

There are also models that downplay and even dismiss the likelihood of neighborhood effects altogether. These models, common in economics and anchored in theories of rational choice, assume that people do not base their decisions on what their neighbors value or how their neighbors behave. Instead, people base their decisions on their personal interests and circumstances. As long as individuals are able to find others who hold positive standards, they will be able to form positive lives. From this standpoint, the relative concentration of positive values and behaviors in a neighborhood is not as important as their mere presence.

3. *Conceptualizing Neighborhoods*

How neighborhoods are approached in research, policy, or practice will depend on how they are conceptualized. Neighborhoods may be conceived as sites, networks, cultures, or perceptions (Burton et al. 1997, Gephart 1997). The most common approach is to consider neighborhoods as sites and focus on their social and physical makeup. In these cases, the neighborhood unit is most often a set of block groups, zip codes, or census tracts; and data are drawn from a variety of largely official sources (e.g., census, crime reports, vital health statistics, social service agencies, school records). Recently, Sampson and Raudenbush (1999) also found innovative techniques for gathering systematic first-hand data via 'windshield' surveys.

When neighborhoods are approached as *networks*, the focus is on intimate ties, social relationships, and patterns of interaction. When neighborhoods are approached as *cultures*, the focus is on the behaviors, outlooks, rituals, and languages that characterize a neighborhood. Of course, neighborhoods can also be understood as *perceptions* and through the subjective experiences of neighborhoods as sites, networks, and cultures. Regardless of how neighborhoods are conceptualized, individual-level data are most often gathered through survey and interview methods and then aggregated. Indicators may be objective or subjective, quantitative or qualitative, static or temporal. There are strengths and limitations associated with each conceptualization of neighborhood, and the relevant unit to study and data to gather will depend on the outcomes or processes of interest.

4. *Measures of Neighborhood Composition, Structure, and Process*

Most studies of neighborhood effects have relied on single indicators or built composite measures of neighborhood composition. These have largely been based on secondary public sources, especially the census. Most research has focused on factors related to SES (e.g., percentage of families in poverty, median family income, percentage of male joblessness, median educational attainment of adults, percentage of adults in professional or managerial positions). It has also been common to emphasize racial composition (especially percentage of Black) and intersections between racial composition and indicators of SES. A few characteristics have also been used to index neighborhood commitment and stability (e.g., percentage of families in same dwelling for 5+ years, percentage of residents who own their homes).

Other researchers have built more informative measures of neighborhood structure and even process. These are most often based on individual-level survey data that are then aggregated. These studies have turned attention to dynamics around social cohesion, social control, social support, and disorganization, and to the availability of various types of resources, adult participation in local organizations, and stability in and satisfaction with the neighborhood. The most advanced research has attempted to link neighborhood composition, structure and process. For example, Wilson (1996) emphasizes the larger economic conditions that result in joblessness; the flight of middle-class families to the suburbs that reduces resources and investments from local and state governments; and the declining marital stability of families that results in a shortage of positive male role models.

These and other factors produce a local ethos that undermines social consensus and social control and produces social isolation. Similarly, Brooks-Gunn et al. (1993) suggest that the percentage of local adults who work in managerial and professional occupations affects the development of youngsters in positive ways because these adults are more likely to promote conventional attitudes and behaviors.

5. *Neighborhood Effects on Children and Adolescents*

Most scholarship has focused on the impact of neighborhood environments on the social outcomes of children and adolescents. Early childhood and late adolescence seem especially important periods in that they involve significant shifts in both family and school settings and individuals in these two periods differ in the degree to which they are susceptible to influences outside of the family (Brooks-Gunn et al. 1993). Neighborhood influences on development early in life also seem particularly important in that positive or negative effects may cumulate over time, generating further advantage or disadvantage over the life course.

Despite significant interest in neighborhood effects on the development of young people, research has not shown consistent or sizable effects (for overviews, see Gephart 1997, Mayer and Jencks 1989). (This may also relate to measurement problems, some of which will be highlighted in the conclusion of this article.) In childhood, neighborhoods have been shown to affect cognitive functioning, behavioral problems, child maltreatment rates, low birth weight rates, and infant death rates. In adolescence and early adulthood, neighborhoods have been shown to affect delinquency, drug use, school achievement and dropout, teen childbearing, health status, joblessness, poverty, self-efficacy, personal commitment to conventional attitudes and behaviors, and having conventional or delinquent friends.

Most literature is focused on negative outcomes (and on how to prevent negative outcomes) rather than on positive outcomes (and on how to foster positive outcomes). It is also focused on adolescence. Little attention has been paid to neighborhood effects on the development of adults outside of their early 20s. The exception to this trend relates to the other end of the life course: old age.

6. *Neighborhood Effects on Adults and the Elderly*

In the field of gerontology, there has been longstanding interest in the fit between older people and their environments. Where the subject matter of this entry is concerned, this fit relates to how residential environments might be designed or redesigned to better accommodate the physical, psychological, social, and economic changes that often accompany aging. It also relates to the challenges of relocation in later life, as significant numbers of older individuals move to nursing homes, government-subsidized apartments, or other residential settings. Indeed, many institutions for older people are explicitly aimed at achieving a sense of 'community' by design. These and other factors may lead to age segregation (rather than age integration) in residential and other environments, a phenomenon that has garnered attention in recent literature on aging (see Uhlenberg and Riley 2000).

Even those who age 'in place' have seen their neighborhoods through significant change, and often for the worse. Along these lines, there is also a tradition of research on crime and the elderly, both in terms of fear of crime (e.g., Ferraro and LaGrange 1992) and actual victimization (e.g., McCabe and Gregory 1998). The personal circumstances of many older people also constrain their residential options and their ability to move (especially health, marital and financial statuses). Recent scholarship on aging has also emphasized the differing needs and circumstances of rural and urban elders (e.g., Coward and Krout 1998) and the important roles that neighbors and satisfaction with neighborhood play in determining well being (Phillipson et al. 1999). Indeed, older people may be more likely to feel the absence of community and more likely to benefit from community life (Lucksinger 1994).

Relative to scholarship on childhood and adolescence, scholarship on neighborhood effects on older adults is significantly less advanced. While important research has been conducted on both ends of life, there is little research on the many decades between. Scholarship on adult development and aging should evaluate whether the models developed on childhood and adolescence can be extended to early adulthood, midlife, and later life, and how existing models must be reconsidered or examined anew to better understand development during these three periods.

7. *New Directions for Research, Policy, and Practice*

A significant problem in demonstrating neighborhood effects relates to important unobserved biases in empirical estimates, particularly those due to selection. Residents in specific neighborhoods differ in important ways before they arrive in those neighborhoods, yet those who choose to live in (or leave) particular neighborhoods are also likely similar in important ways. Depending on these dynamics, neighborhood effects may be overestimated or underestimated. Effects might also be small because the degree

of variation naturally found in a neighborhood is small. But because most research has been conducted at the level of census tracts, the degree of variation may instead be too great, making it equally difficult to ascertain neighborhood effects.

This point also draws attention to the fact that neighborhood phenomena must be understood in dynamic rather than static ways. Changes in neighborhoods, changes in individuals, and the connections between them must be mapped simultaneously. However, temporal multilevel models pose significant challenges for theory and research (for a discussion of these challenges, see Settersten 1999). Where neighborhoods are concerned, longitudinal frameworks are particularly challenging because the geographic mobility of individuals and families dramatically increases the number of neighborhoods that must be managed as a sample is tracked over time.

Research on neighborhood effects must also begin to consider both short- and long-term effects and the forms of these effects. It is important to explore the possibility of non-linear effects, including threshold models (i.e., points after which neighborhoods may become especially relevant or irrelevant to individual development). It is also important to examine connections between neighborhoods and individuals that are not only top-down, but also bottom-up or reciprocal in nature.

Neighborhoods must also be considered in combination with other contexts, and family and school settings are especially likely to moderate neighborhood effects. Because social contexts are complexly interwoven, scholarship must specify how various contexts are interrelated and how these interrelationships impact development during different periods of life. Scholarship on this topic must, therefore, produce a more differentiated view of how neighborhoods affect human development. It must address how specific aspects (or specific combinations) of neighborhood composition, structure and process might be tied to specific kinds of developmental outcomes (whether physical, psychological, or social). And it must address whether specific subsets of residents (e.g., by age, race, gender, social class, or their intersections) are more or less affected by neighborhood factors, whether they are affected in different ways, and why.

Local policies and initiatives aimed at building the social capital of communities ultimately improve the welfare and development of neighborhood residents. These should include efforts to foster or better economic security, housing, safety, schools and other institutions, services, social interaction and support, community service and leadership, or a sense of belonging and pride (Brown and Richman 1997). However, intensive community organizing is required for most community-based initiatives and must reflect a broad array of strategies to mobilize residents and to teach them how to plan and take action. State and federal policies also affect the welfare and development of neighborhood residents, especially via policies related to employment and its correlates (e.g., training, health insurance, transportation, childcare), public safety, drugs, and education (Lehman and Smeeding 1997). Together, policies and initiatives at each of these levels shape the experience of everyday life and the options, flexibility, and control that individuals have as they grow up and older.

See also: Community Environmental Psychology; Community, Social Contexts of; Community Sociology; Environmental Psychology: Overview; Life Course: Sociological Aspects; Lifespan Development, Theory of; Neighborhood: General; Residential Environmental Psychology

Bibliography

Brooks-Gunn J, Duncan G J, Klebanov P K, Sealand N 1993 Do neighborhoods influence child and adolescent development? *American Journal of Sociology* **99**: 353–95

Brown P, Richman H A 1997 Neighborhood effects and state and local policy. In: Brooks-Gunn J, Duncan G J, Aber J L (eds.) *Neighborhood Poverty*. Sage, New York, Vol. 2

Burton L M, Price-Spratlen T, Spencer M B 1997 On ways of thinking about measuring neighborhoods: Implications for studying context and developmental outcomes of children. In: Brooks-Gunn J, Duncan G J, Aber J L (eds.) *Neighborhood Poverty*. Sage, New York, Vol. 2

Coward R T, Krout J A (eds.) 1998 *Aging in Rural Settings: Life Circumstances and Distinctive Features*. Springer, New York

Crane J 1991 The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing. *American Journal of Sociology* **96**: 1226–59

Ferraro K F, LaGrange R L 1992 Are older people most afraid of crime? Reconsidering age differences in fear of victimization. *Journal of Gerontology* **47**: S233–44

Furstenberg Jr F F., Hughes M E 1997 The influence of neighborhoods on children's development: A theoretical perspective and research agenda. In: Brooks-Gunn J, Duncan G J, Aber J L (eds.) *Neighborhood Poverty*. Sage, New York, Vol. 2

Gephart M A 1997 Neighborhoods and communities as contexts for development. In: Brooks-Gunn J, Duncan G J, Aber J L (eds.) *Neighborhood Poverty*. Sage, New York, Vol. 1

Lehman J S, Smeeding T M 1997 Neighborhood effects and federal policy. In: Brooks-Gunn J, Duncan G J, Aber J L (eds.) *Neighborhood Poverty*. Sage, New York, Vol. 1

Lucksinger M 1994 Community and the elderly. *Journal of Housing and the Elderly* **11**: 11–28

Mayer S E, Jencks C 1989 Growing up in poor neighborhoods: How much does it matter? *Science* **243**: 1441–5

McCabe K A, Gregory S S 1998 Elderly victimization: An examination beyond the FBI's Index Crimes. *Research on Aging* **20**: 363–72

McLaughlin M W, Irby M A, Langman J 1994 *Urban Sanctuaries: Neighborhood Organizations in the Lives and Futures of Inner-City Youth*. Jossey-Bass, San Francisco

Phillipson C, Bernard M, Phillips J, Ogg J 1999 Older people's experiences of community life: Patterns of neighboring in three urban areas. *Sociological Review* **47**: 715–43

Sampson R J, Raudenbush S W 1999 Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American Journal of Sociology* **105**: 603–51
Sampson R J, Raudenbush S W, Earls F 1997 Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* **277**: 918–24
Settersten Jr R A. 1999 *Lives in Time and Place: The Problems and Promises of Developmental Science*. Baywood Publishing, Amityville, NY
Skogan W G 1990 *Disorder and Decline: Crime and the Spiral of Decay in American Neighborhoods*. Free Press, New York
Uhlenberg P, Riley M W (eds.) 2000 Essays on age integration. *Gerontologist* **40**: 261–307
Wilson W J 1996 *When Work Disappears: The World of the New Urban Poor*. Knopf, New York

R. A. Settersten Jr.

Community Power Structure

Two issues have been central to the study of urban politics and policymaking in the United States and, to a lesser extent, in Western Europe: Who holds power within the community? And how much autonomy do urban policy makers have to make meaningful decisions within national and global systems. Although the phrase 'community power structure' is commonly used to describe these discussions, the use of the term 'community' is somewhat misleading, since the jurisdictional level analyzed is usually that of the city.

1. *Origins of the Debate*

Consideration of the question of who has power within cities began with the debate between elitists and pluralists that developed after the publication of Robert Dahl's *Who Governs?* in 1961. Dahl examined the question of power in New Haven, Connecticut, by counting the participants in a series of important decisions. He focused on the issue areas of urban redevelopment, education, and political nominations and found that, while elites dominated in each area, only public officials subject to electoral sanctions had influence in all three. The system was therefore one of 'dispersed inequalities.' He further concluded that all social groupings had the potential to mobilize power resources. This finding was further elaborated by Nelson Polsby, who argued that urban society was 'fractured into congeries of hundreds of small special interest groups, with incompletely overlapping memberships, widely diffused power bases, and a multitude of techniques for exercising influence on decisions salient to them' (Polsby 1970, p. 118).

Polsby, in particular, contrasted the pluralist interpretation of urban power structure with elite theory, which contended that business leaders controlled public decision making. Hunter's (1953) study of Atlanta, Georgia, was seen to epitomize the elitist viewpoint. Hunter had asked informants who they thought made the important decisions in Atlanta. Each person interviewed was required to name those with power and to provide the names of additional informants. Hunter found that a small group of people was named in relation to a number of decisional areas and that the nominal participants in decision making were only the agents of those holding real power. According to Polsby , Hunter's use of this reputational method inevitably led to the result he found. Thus, the allegedly powerful had a reputation for power but did not necessarily really exercise power. In contrast, he claimed that the pluralists recorded actual behavior.

The pluralist contention of superior methodology, however, was challenged by Bachrach and Baratz (1970), who asserted that observing manifest behavior failed to capture 'the hidden face of power.' By exerting their authority behind the scenes or simply by shaping a set of expectations, truly powerful people structured the decisional agenda, insuring the exclusion of issues that could undermine their position. Community power, in this formulation, shifted from a set of overt actions involving the costly exercise of control to a more subtle manipulation of the context in which decision making took place.

2. *Shifting Terms of the Debate*

During the 1970s the analysis of community power evolved in two directions, both in response to events 'on the ground.' The initial community power debate took place against the backdrop of quiescent cities seemingly governed by a broad consensus over the contours of policy. Then suddenly cities in both the United States and Europe became arenas of social conflict. Within the United States divisions occurred primarily along racial and ethnic lines, with minority groups protesting their disadvantaged status as bureaucratic clients, as losers under urban renewal programs, and as the objects of discrimination (Fainstein and Fainstein 1974). In Europe urban social movements, based in both working class and student-radical strata, rebelled against sclerotic bureaucracies and rule by notables. They aimed at increasing collective consumption—that is, at improving the level and quality of state-sponsored benefits, especially housing. On both sides of the Atlantic demands for greater citizen participation challenged pluralist conclusions that, within democratic polities, previous low levels of mobilization indicated broad satisfaction with governance (see *Citizen Participation*).

Within the USA, heightened demands for state services, mobilizations around symbolic issues, and blocking of public programs by negatively affected groups led some observers to ask whether cities could be governed at all. Whereas earlier pluralists had regarded competition among groups as a benign exercise, this later group of 'hyperpluralists' regarded group conflict as counterproductive (Yates 1977, Wirt 1974, Judge 1995) (see *Locational Conflict (NIMBY)*).

In contrast, many other urbanists responded to urban unrest by sympathizing with the protestors. In Europe, the rise of urban social movements stimulated the application of Marxist theory to urban spaces, as part of a more general incorporation of spatial relations into Marxist structuralism (see Harvey 1973, Castells 1977). As a consequence, the question of power in cities shifted from the issue of who makes decisions to what were the forces that determined who benefited from urban change (Harloe 1977, Pickvance 1976). Pluralists and elitists had debated over who had power within cities. Now the question became whether local policy makers could actually control the destiny of a place. Instead cities were seen as the outcome of the impersonal operations of capital using space to maximize profits. Thus, cities experienced the benefits of capital investment or the depredations of capital withdrawal depending on factors mainly beyond the control of local decision makers (Harvey 1985). Moreover, capitalists benefited from the very volatility and competitiveness that jeopardized the wellbeing of urban residents.

Within the Marxist framework, the structural dependence of the local state on capital meant that politicians had little choice but to defer to the wishes of capitalists in making policy decisions (Foglesong 1986). In other words, the reliance by public officials on private investors for jobs for their constituents and, especially in America, for support of their tax base meant they had to conform to the wishes of capitalists. Only if they insured a 'good business climate' could they entice business leaders to invest in their locality.

The Marxist critique prompted the rise of the political economy approach to studying community power; within the US this mode of analysis branched into right and left forks. On the right, Paul Peterson (1981) argued that city governments needed to compete for business investment and were powerless to carry out redistributional policies, since the latter would create a intolerable burden on business. Unlike those on the left who considered cities as hostage to capital, however, Peterson regarded the pursuit of economic development as in the interest of all city residents. Thus, like the pluralists he maintained that public policy was made within a general agreement on the merits of economic development programs.

On the left, political economists ran a spectrum from neo-Marxist to neo-Weberian to eclectic. The neo-Marxists were particularly concerned with the contradictory demands on the local state to, on the one hand, provide the basis for capital accumulation through construction of infrastructure and subsidies to business, and on the other, to legitimate capitalist domination through social welfare expenditures. The outcome of this balancing act was a fiscal crisis of the state (Alcaly and Mermelstein 1977, Tabb and Sawers 1984). Capitalist domination of local decision making also manifested itself in land use planning, where urban redevelopment programs provided the basis for land grabs and gentrification (Dear and Scott 1981, Smith 1991).

Other critical approaches to urban political economy focused on local decision makers rather than the more abstract logic of capital. While partly Marxist-inspired, these works resembled the earlier pluralist and elite analyses in using case studies to locate the actors and processes that produced unequal results within particular cities. Influential works identified urban 'growth machines' or 'pro-growth coalitions' as the dominant agent in urban decision making (Logan and Molotch 1987, Mollenkopf 1983); property developers were central figures in these groupings. Comparative analysis of American cities indicated that the character of urban regimes changed historically and that the period of urban conflict in the 1960s and 1970s was marked in many cities by 'concessionary regimes' that responded to previously excluded population groups. With the decline of overt protest, however, 'conserving regimes,' while not totally eliminating redistributional programs from the earlier period, increasingly responded to business interests (Fainstein et al. 1986). Thus, the nature of community power did not remain stable and could not be captured in an ahistorical analysis (see *Urban Planning: Central City Revitalization*).

Attempts to apply the growth machine analysis to European cities found the concept less useful there (Harding 1994, 1995, Molotch and Vicari 1988). Because localities depended less on their own resources and government was considerably more centralized, the stakes of local politics were lower than in the United States. Nevertheless, some convergence toward the US model was found by Saunders (1979) in his study of development in Croydon (an outer borough of London).

As the US federal government withdrew from urban programs, city governments redefined their mission as promoting economic development rather than addressing social problems (Judd and Swanstrom 1998). The competition for footloose capital intensified, and public officials devised a variety of programmatic mechanisms to assist business development, primarily relying on off-budget tax subsidies. A number of analyses examined the 'public-private partnerships' supporting economic development efforts and concluded that the private side dominated the partnerships (Squires 1989, Beauregard 1989). Ultimately, however, the theoretical issue switched almost entirely

from the question of who governed to the extent of local agency (Harloe et al. 1990)—that is, what difference could local actors make, especially those who were not members of the business elite.

3. Urban Regimes and Local Autonomy

Political economic analysis seemed to make moot the issue of community power by attributing urban transformations to the effects of global economic restructuring. The implication was that human agency could not affect the most important factors determining people's lives (Goldsmith 1995). Critics, however, attacked the political economy approach as overly deterministic and pessimistic (Swanstrom 1993). They argued that local politics remained a meaningful endeavor and, like the early pluralists, that resources were available to be mobilized even by marginal groups. At the same time, however, analyses of the effects of African–American mayors on public policy indicated that changes in elected administrations did not change the basic contours of policy (Reed 1988). In particular, Clarence Stone's (1989) study of Atlanta demonstrated the persistence of a stable regime favoring business interests even as control of city hall passed from the white elite to rhetorically militant African Americans.

Stone (1993) laid out a theory of urban regimes that became the prevalent approach to urban power structure at the end of the century. According to regime theory, local politics mattered, and who dominated at the local level affected the welfare of urban citizens. Stone outlined a typology of possible regimes that included low-income opportunity regimes as well as business-dominated and progressive middle class ones. A regime consisted of both public and private actors, and it cut across a broad range of issues rather than, as the pluralists argued, being an issue-specific coalition. Although regime theory shared with pluralist theory a focus on the interaction of local political groupings, it differed in that Stone attributed a 'systemic bias' to urban power networks. Business interests, because of their control over investment allocations, usually dominated, and it took extraordinary effort on the part of other groups to overcome this natural tendency.

Regime theory thus fused political economy and pluralism. Like political economy it did not separate politics and economics, and it was concerned with who benefited from policy decisions, not simply who made them. Like pluralist theory it stressed the importance of building political coalitions and assumed that local decision-making could change people's lives. Stone (1993) also stressed that regime theory differed from pluralist theory in that it had a social production model of power—that is, power had to be created, was not simply a matter of control, and consisted of the ability to get things done. A careful reading of *Who Governs?* however, indicates that Dahl likewise emphasized the importance of coalition building; the mayor of New Haven was only able to achieve his urban renewal program through a process of social production similar to that described by Stone.

To date the application of regime theory has largely been restricted to American cities. Increasingly, however, the study of urban politics has taken on a comparative dimension, and analysis of local regimes offers a promising framework for comparison (Stoker 1995). Rational choice theory is its principal competitor as a theoretical basis for urban analysis. But while rational choice offers a method of analyzing the content and effects of public policy decisions, it does not lend itself to the scrutiny of power relationships.

The evolution of theory always reflects the development of an initial set of ideas and then a response. At the same time, it responds to actual events and real political conflicts. Perhaps in no area of political science has the social basis of different theoretical formulations been more apparent. The political quietude of the 1950s and the hegemony of anticommunism within the West underlay the pluralist rejoinder to the radical implications of elite theory. The social movements of the sixties and seventies stimulated both the neo-Marxist argument and the hyperpluralists' fears. The economic restructuring of the 1980s and the subsequent prevalence of neo-liberal ideology have framed the turn-of-the-century division of labor between political economists, who continue to focus on issues of class, race, and power in cities, and rational choice theorists, who focus on questions of strategies and their results.

See also: Community Sociology; Elites: Sociological Aspects; Planning, Politics of; Pluralism; Policy Process: Business Participation; Power in Society; Social Movements, Geography of; Urban Politics: United States; Urban Studies: Overview

Bibliography

Alcaly R E, Mermelstein D (eds.) 1977 *The Fiscal Crisis of American Cities*. Vintage, New York

Bachrach P, Baratz M S 1970 *Power and Poverty*. Oxford University Press, New York

Beauregard R A (ed.) 1989 *Atop the Urban Hierarchy*. Rowman and Littlefield, Totowa, NJ

Castells M 1977 *The Urban Question*. MIT Press, Cambridge, MA

Castells M 1983 *The City and the Grassroots*. University of California Press, Berkeley, CA

Dahl, R 1961 *Who Governs*? Yale University Press, New Haven, CT

Dear M, Scott A J (eds.) 1981 *Urbanization and Urban Planning in Capitalist Society*. Methuen, New York

Fainstein S S, Fainstein N I 1974 *Urban Political Movements*. Prentice Hall, Englewood Cliffs, NJ

Fainstein S S, Fainstein N, Hill R C, Judd D R, Smith M P 1986 *Restructuring the City*, (rev. edn.) Longman, New York
Foglesong R 1986 *Planning the Capitalist City*. Princeton University Press, Princeton, NJ
Goldsmith M 1995 Autonomy and city limits. In: Judge D, Stoker G, Wolman H (eds.) *Theories of Urban Politics*. Sage, London
Harding A 1994 Urban regimes and growth machines: Towards a cross-national research agenda. *Urban Affairs Quarterly* **29**: 356–82
Harding A 1995 Elite theory and growth machines. In: Judge D, Stoker G, Wolman H (eds.) *Theories of Urban Politics*. Sage, London
Harloe M (ed.) 1977 *Captive Cities*. Wiley, London
Harloe M, Pickvance C G, Urry J (eds.) 1990 *Place, Policy, and Politics: Do Localities Matter?* Unwin Hyman, Boston
Harvey D 1973 *Social Justice and the City*. Johns Hopkins University Press, Baltimore, MD
Harvey D 1985 *The Urbanization of Capital*. Johns Hopkins University Press, Baltimore, MD
Hunter F 1953 *Community Power Structure*. University of North Carolina Press, Chapel Hill, NC
Judd D, Swanstrom T 1998 *City Politics*, 2nd edn. Longman, New York
Judge D 1995 Pluralism. In: Judge D, Stoker G, Wolman H (eds.) *Theories of Urban Politics*. Sage, London
Logan J, Molotch H 1987 *Urban Fortunes*. University of California Press, Berkeley, CA
Mollenkopf J 1983 *The Contested City*. Princeton University Press, Princeton, NJ
Molotch H, Vicari S 1988 Three ways to build: The development process in Japan and Italy. *Urban Affairs Quarterly* **24**: 188–214
Peterson P 1981 *City Limits*. University of Chicago Press, Chicago
Pickvance C G 1976 *Urban Sociology*. St Martin's, New York
Polsby N W 1970 *Community Power and Political Theory*. Yale University Press, New Haven, CT
Reed A 1988 The black urban regime. In: Smith M P (ed.) *Power, Community and the City*. Transaction, New Brunswick, NJ
Saunders P 1979 *Urban Politics*. Hutchinson, London
Smith N 1991 *Uneven Development*, 2nd edn. Blackwell, Oxford, UK
Squires G 1989 *Unequal Partnerships*. Rutgers University Press, New Brunswick, NJ
Stoker G 1995 Regime theory and urban politics. In: Judge D, Stoker G, Wolman H (eds.) *Theories of Urban Politics*. Sage, London
Stone C N 1989 *Regime Politics*. University Press of Kansas, Lawrence KS
Stone C N 1993 Urban regimes and the capacity to govern: A political economy approach. *Journal of Urban Affairs* **15**: 1–28
Swanstrom T 1993 Beyond economism: Urban political economy and the postmodern challenge. *Journal of Urban Affairs* **15**: 55–78
Tabb W, Sawers L 1984 *Marxism and the Metropolis*. Oxford University Press, New York
Wirt F M 1974 *Power in the City*. University of California Press, Berkeley, CA
Yates D 1977 *The Ungovernable City*. MIT Press, Cambridge, MA

S. S. Fainstein

Community, Social Contexts of

1. *From Rural Decline to Globalization*

Community denotes first a local government (village, town, city). Second it conveys social integration—actual, or desired by the speaker—as in 'the global community.' Community sociology was shaped by Toennies, Weber, and Durkheim in the late nineteenth century over concerns for rural decline, industrialization, and urbanization. They held that small communities generated more social integration. These concerns continue today, but in the post-industrial word, globalization is reshaping the basic contours of social life, via global capitalism, worldwide migration, mass communication, and the Internet (Sassen 1991). For Weber and Durkheim the community was a strategic research site for refuting theories of pure individualism. Today as then, localities are strategic research sites for huge processes and forces—strategic first since one can observe global dynamics on a scale manageable for research. Second, because there are so many communities, they encourage comparison and thus scientific generalization.

The massive community research by sociologists in the twentieth century was shaped more by nineteenth century concerns than by new global concerns, which emerged only near 2000. This article considers three major themes from past community research still vital today:

(a) individual citizens and the nation;
(b) intermediary groups, like civic associations, and;
(c) new public agendas like environment issues.

All three illustrate transformations of community processes by global forces that shift power among participants and redefine the rules of the game.

Weber's and Durkheim's concerns continued for decades across much of the world, sparked by rapid urbanization and industrialization, which often brought explosive numbers of migrants to large cities, escalating crime, and political and social conflicts between new and old residents.

More abstractly, Weber and Durkheim used local communities to illustrate collective processes like shared values, migration, and the division of labor. They discussed these interrelated processes to articulate a focus for sociology, complementing the methodological individualism of Adam Smith's economics and the political contract theory of Hobbes, Locke, Bentham, and J. S. Mill.

Durkheim, Toennies, and many others contrasted a rising and seemingly unbridled individualism of the city against the socially-integrated (if perhaps idealized) small town, the Gemeinschaft. In this traditional community, they suggested, the individual was integrated via deep personal relations with family, neighbors, and strong institutions, especially churches. Such ties were broken when people migrated to cities

for new jobs. For George Simmel and Louis Wirth the city could thus offer 'Freie Luft': freedom and cosmopolitanism in the sophisticated diversity of urban life, liberating individuals from local provincialisms. Many contemporaries, from Baudelaire's poetry of alienation to Dreisen's novels of despair, projected individualism to radical extremes. Durkheim wrote of anomie and suicide, and Weber of harshly ascetic Protestant individualism.

Yet these big hypotheses, which seemed almost patently obvious when first stated, were surprizingly often not supported on careful investigation. The main surprize was discovering tight and powerful social relations among families and (selected) friends and new institutions in large cities. Close ties were not necessarily nearby; they might be miles or a phone call away; but they were often found when sociologists looked closely. This 'rediscovery of the primary group' emerged in ethnographic studies and later systematic surveys. Further, many rural residents reported considerable isolation and distrust, when studied closely. We are still unclear how much has changed, since more precise research emerged only at the end of the twentieth century (e.g. Fischer 1982). Thus it may be that some past small towns demonstrated greater interconnectedness and value consensus. Still as the range and depth of studies expand to include rural persons in Asia, Africa, and Latin America, reports are common of deep social conflicts and manipulations paralleling those in large urban areas. I have recently analyzed citizen reported trust in 34 countries, and found almost no differences by city size for towns from 2000 up to millions of residents. In just three countries do residents in the smallest towns report more trust. (In towns under 2000 inhabitants in the UK and China and under 10,000 in India, using the World Values Survey of 80,000 respondents. Item:

'Would you say that most people can be trusted or you can't be too careful in dealing with people.'

Most countries, like the US in Fisher's results, are more homogeneous. Still locations vary drastically in trust and social relations. Why, if not by community size?

Weak support for earlier hypotheses prompted new investigations. Most differ from the past by elaborating more precise intervening connections, like civic groups, but still use the community as a strategic research site to detail impacts of 'community context' like residents' educational level. Stressing context, as Abbott (1999) suggests, was a key element of Chicago urban sociology that continues today. We consider three illustrations of contextual variation in the next three sections. Transition from the first to the third illustrates the principle of hierarchy leveling.

2. *Citizenship and the Nation State*

T. H. Marshall identified 'citizenship' as a core concept linking citizens to the nation. He traced shifts in its meaning over two centuries. Every citizen of a nation has 'rights' which rose from minimal to extensive along with egalitarianism. The national welfare state is the agency largely responsible for implementing these rights, which extended employment guarantees, health care, retirement, and later, children and animal rights. As they grew, dependence on the locality paled.

Still national centralization varies substantially across countries. Napoleon declared it illegal for French citizens to meet publicly or to organize civic groups; Soviet Russia and Nazi Germany went further by making everyone a potential police informant, thus seeking to destroy social ties among all individuals. Some national leaders pursue such centralization to control dissent. But centralization springs from other sources: building on the social contract tradition, the philosopher John Rawls held that no thorough egalitarianism is possible unless all social ties are eliminated, including those in families.

This is an extreme form of the common criticism that community institutions preserve tradition, and should thus be weakened. In societies with authoritarian leaders and institutions, opposition groups often seek to level hierarchies through radical egalitarian programs. Many left, social democratic-type parties thus favored national control to level hierarchy and implement egalitarian citizenship in much of the twentieth century. Parallel arguments led most northern and eastern European countries to consolidate their small towns into 'metropolitan governments,' consciously eliminating traditional 'parish-based' loyalties. Following these and other centralizing policies, nation states generally rose in power, social responsibility, and budgets for the past two centuries, undermining local communities.

The strong nation legacy came under criticism in the last decades of the twentieth century, as people grew more affluent, educated, and sophisticated: they criticized national agencies for high costs, inefficiency, and non-responsiveness to citizens. Support for local autonomy thus rose. Globalization presses in the same direction: with more cross-national travel, communication, investment, and trade, the nation declines in its abilities to deliver 'egalitarian' welfare benefits, since ideal standards are less national and increasingly international 'Human rights' rises as a new standard. Yet the world is too large, as are regional entities like the European Union, to implement specifics. The net effect is thus to bolster regional and local entities. Still, why should they work?

3. *Searching for Democracy and Trust In Active Civic Groups*

An early critic of Napoleonic state centralization was Alexis de Tocqueville, whose *Democracy in America* in

the 1830s contrasted the French state-dependency of citizens with the rich array of American community associations. Local civic groups, linked to small and autonomous local governments, gave (ideally) all citizens the ability to participate meaningfully in decisions affecting them. Such experiences built networks of social relations and taught the values of participation, democracy, and trust in fellow citizens. Parallel ideas have been advanced in research on organizations, small groups, families, and socialization: more participation by all generates more social interaction among peers, and more social skills and trust. This is achieved better in small units, encouraging more focused exchange and meaningful social contacts. This has similarly led to analyzing specific types of intermediary groupings, from social classes to political parties, churches, and business groups.

Large literatures document these basic ideas in studies of different types of associations, causes and consequences of greater participation, citizen trust, and confidence in leaders. Over the twentieth century, many organizations shifted from the hierarchical and centralized to the smaller and more participatory. The community power literature from Floyd Hunter (1953) to Robert Dahl (1961) and beyond suggests a decline in the 'monolithic' city governance pattern which Hunter described in Atlanta. In contrast, Dahl documented a more participatory, 'pluralistic' decision-making process, where multiple participants combine and 'pyramid' their 'resources' which affects decisions in different 'issue areas' Dahl's quoted concepts highlighted freedom of choice and openness of leadership to popular initiatives. Some 160 studies were completed seeking to explore implications for Hunter vs. Dahl's concepts. Peter Rossi, Robert Crain and Terry Clark started national comparative research in 1967 with the Permanent Community Sample at the National Opinion Research Center, University of Chicago. The same US municipalities were resurveyed periodically to monitor change. These studies continue in the Fiscal Austerity and Urban Innovation (FAUI) Project, which by 2000 included surveys of over 10,000 communities in some 30 countries.

These comparative studies document support among citizens and leaders for more decentralized, egalitarian, participation by citizens the world over. *The New Political Culture* (Clark and Hoffmann-Martinot 1998) uses the FAUI data to chart the rise of new rules of the game for public decisions and building more active citizen-leader contacts via mechanisms like focused groups, block clubs, cable-TV coverage of local associations, and Internet groups. Putnam's (1993) related work documents clear differences across Italian communities, from hierarchical, state control in the South to Tocquevillian participation in the North. Fukukyama shows national differences in citizen trust to follow the pattern Tocqueville proposed: least in counties with national states that destroyed civic groups (France, Italy, former Soviet areas), more in countries with active civic groups (USA, Scandinavia).

Globalization forces have opened more hierarchically-organized societies to permit civic and political associations to emerge even if they oppose the national government. Ethnic separatist groups in Mexico, Yugoslavia, and former Soviet areas grew more active via intervention from international human rights associations, global media coverage of scandals, and international organizations like the UN and NATO. The global trend is toward the more decentralized, participatory patterns of culture, which rekindle local democracy and civic life.

4. *New Movements, Agendas, and Policies: Greening the World*

New Social Movements (NSMs) emerged in the 1970s to push new agendas: they championed ecology, feminism, peace, gay rights and other issues that 'old' groups and political parties ignored. The NSMs were initially less formally organized and often used confrontational tactics, sometimes even terrorism, generating the findings that those more organized and ready to use violence had more impact (Gamson 1975, 1990). More recent work stresses the 'political opportunity structure' accessible to social movements (McAdam D, McCarthy J D, Zald M N 1996) or the more general 'hierarchical leveling' propositions in Clark and Hoffmann-Martinot (1998). For instance in the 1970s in Italy even Communist and Socialist parties rejected the new issues; this encouraged radical tactics including marches and street violence. But as political parties and governments embraced many new social issues, the opportunity structure drastically shifted. Movement leaders then shifted from 'urban guerilla warfare' to participating in elections, lobbying, and advising governments. This expanded participation process encourages moderating demands.

In Europe, the national state and parties were the hierarchical 'establishment' opposed by NSMs. In the US, local business and political elites were more often targeted. The NSMs demanded more responsiveness to the environment and to more diverse lifestyles and ethnic concerns. From the 1970s to 2000, support grew for the new issues, but explaining why remains controversial. One interpretation in the US literature is that business elites redirect agendas toward growth (Logan and Molotch 1987). Similarly Stone (1989) held that leadership 'regimes' were slanted toward business. For Peterson (1981) competitive market capitalism forced all communities to pursue growth by lowering taxes and welfare-like benefits—or lose residents and jobs. Business dominance driving a 'pro-growth agenda' was commonly reported in the 1970s and 1980s, but by the 1990s the rise of new issues led to questioning past interpretations. Clark and Goetz

(1994) showed that 26 percent of US municipalities had anti-growth movements in the 1980s, which often had major impacts: on zoning, limits on size and number of new building, environmental impact hearings, historic preservation, and sensitivity to amenities like lakes. If green NSMs won, did business lose? Logan et al. (1997) reviewed 20 years of studies and found much less business impact than the growth machine hypothesis suggested. Similarly, Clark and Goetz found zero relations between business dominance and environmental policy, although business leaders did rank as powerful in many US cities. The resolution of this apparent paradox is that many business leaders do not promote growth, but support public agendas which vary across communities, reflecting local concerns.

Behind pressures from NSMs and organized groups lies the question of why such groups emerge? They are more common in localities with citizens more highly educated, reasonable affluent, who work in professional and technical occupations, where the media are powerful, and traditional parties, unions, and ethnic groups are weak (Clark and Hoffmann-Martinot 1998). Individuals with such characteristics similarly support new agendas, termed a new political culture or post-industrial politics, in more than 20 countries where detailed surveys are available (Clark and Rempel 1997). Inglehart (1997) has argued that such 'post-modern' values emerge more powerfully as income rises, since after material needs are met, then 'post-materialist' or post-modern concerns can be pursued, like democracy, a nicer environment, gay marriages, or abortion rights.

Note that most of these new agenda issues concern consumption and leisure more than work and jobs. Aesthetic and amenity concerns rise—like suburban sprawl, sports stadiums, and parks. These issues do not divide people into rich versus poor as did class politics and party politics for much of the twentieth century. Rather than national issues like Keynesian spending or income redistribution championed by past left versus right (national) political parties, the new agenda issues are led more by community civic groups. Indeed the largest US environmental organization, the Sierra Club, reported that its average member belonged to seven other environmental organizations, many local. Even in Japan, where NSMs and local communities are less salient, local ecology groups were nevertheless the driving force generating national air and water standards, reversing the normal centralization (Michio 1997). The agendas of community civic groups (including block clubs and neighborhood associations) are set more locally through more general participation, consciously contrasting with more hierarchical parties and unions—which alienate many younger activists. Voter turnout for elections organized by the classical national parties (which still control local candidate selection in most of the world) thus declined, while new issue-specific community associations mushroomed in the late twentieth century.

What about poor and minority groups? Some have shifted from the classic welfare/redistribution issues to embrace more civic and individual initiative, albeit more in the US than traditionally hierarchical societies (consistent with the leveling proposition). By 2000, Jesse Jackson, heir of Martin Luther King's movement, continued civil rights issues through classic marches and protests, but he added more speeches advocating self-help and hard work to rise economically and socially—even publishing a book in 2000 with his politically active son entitled *It's About the Money: How You Can Get Out of Debt, Build Wealth, and Achieve Your Financial Dreams*. Broadly analogous, some observers hold, are the dramatic spread of Protestant evangelism in Latin America and Asia, even China, based often on a strong self-help message.

Globalization heightens these trends, weakening traditional agendas and promoting the new issues in countries that might change slower without international contact. Local affiliates of NSMs like ecology groups drew tactics and policies from their international counterparts, via congresses, newsletters, and the Internet. Tourism and the media lead citizens to aspire to consumption standards based more on global than national standards. With the Cold War over, global markets increasingly replaced national states. Left political party leaders the world over often shifted from class and national issues to these new agendas. Presidents and Prime Ministers in the US, the UK, and Germany, illustrated this shift, as did many in Latin America and Asia, plus countless mayors. The new agenda thus spread globally, with slogans like 'think globally but act locally,' helping local communities re-emerge.

See also: Citizenship: Political; Civic Culture; Community, Expression of; Community Sociology; Durkheim, Emile (1858–1917); Globalization and World Culture; Neighborhood: General; Urban Activity Patterns; Urban Ethnography; Urban Sociology; Urban Sprawl; Weber, Max (1864–1920)

Bibliography

Abbott A 1999 *Department & Discipline*. University of Chicago Press, Chicago

Clark T N, Goetz E G 1994 The anti-growth machine. In: Clark T N (ed.) *Urban Innovation*. Sage, Thousand Oaks, CA

Clark T N, Rempel M (eds.) 1997 *Citizen Politics in Post-Industrial Societies*. Westview Press, Boulder, CO

Clark T N, Hoffmann-Martinot V (eds.) 1998 *The New Political Culture*. Westview Press, Boulder, CO

Dahl R A 1961 *Who Governs?* Yale University Press, New Haven, CT

Fischer C S 1982 *To Dwell Among Friends*. University of Chicago Press, Chicago

Gamson W A 1990 *The Strategy of Social Protest*, 2nd edn. Wadsworth, Belmont, CA

Hunter F 1953 *Community Power Structure*. University of North Carolina Press, Chapel Hill, NC
Inglehart R 1997 *Modernization and Post Modernization*. Princeton University Press, Princeton, NJ
Logan J R, Whaley B R, Crowder K 1997 The character and consequences of the growth regimes. *Urban Affairs Review*. **32**(5): 603–30
Logan J R, Molotch H L 1987 *Urban Fortunes*. University of California Press, Berkeley, CA
McAdam D, McCarthy J D, Zald M N (eds.) 1996 *Comparative Perspectives on Social Movements*. Cambridge University Press, Cambridge, UK
Michio M 1997 *Local Power in the Japanese State*. University of California Press, Berkeley, CA
Peterson P E 1981 *City Limits*. University of Chicago Press, Chicago
Putnam R D 1993 *Making Democracy Work*. Princeton University Press, Princeton, NJ
Sassen S 1991 *The Global City*. Princeton University Press, Princeton, NJ
Stone C N 1989 *Regime Politics*. University Press of Kansas, Lawrence, KS

T. N. Clark

Community/Society: History of the Concept

Community and society are key concepts in sociology and political and historical philosophy. In antique Greece, community and society were kept together in one concept *χοινωνι'α*. The antique Greek political philosophy did not distinguish between the institutions of community and society. *χοινωνι'α* was translated into Latin as sometimes *societas*, sometimes *communitas* without clear distinction. The synonymity—and the ambiguity—continued with the oldest Christianity and the scholastic philosophy. In the eighteenth century the complexity of the conceptual field around the equivalent terms of community and society had increased and included communion, participation, joining, taking part, sharing, liaison, union, corporation, co-operative, friendship, brotherhood, fellowship, etc. During the nineteenth century community and society split up and became counterconcepts expressing social and institutional alternatives. The roots of this separation were already in the eighteenth century. Together with another, closely connected, couple of counterconcepts, state and society, they played a crucial role in the reflection and the theory building in social sciences with clear ideological and political implications.

With Thomas Hobbes and Baroch Spinoza in the seventeenth century nature disappears as explanation of human collectivity. With the exception of the family all forms of socialization and communalization (collegium, corporation, craft, parish, municipality, partnership, trading company, etc.) were secondary formations within a state federation (*civitas*, *res publica*) constituted through a social convention. Hobbes talked about 'systems of people, numbers of men joined in one interest, one business.' Instead of the concept of society understood as synonymous with community a new concept of society emerged based on a conventional meaning identical to that of the state in the modern sense of the word.

From the middle of the eighteenth century and increasingly from the 1850s society and the social were seen as taking shape between individuals or households and the political order. Almost all social thought during this period was preoccupied with conceptualizing the collective entities that emerged within the 'social' in order to describe and analyze their nature, stability, and ties of cohesion. Besides community and society, concepts like nation, class, and culture were suggested. Societies were seen as organized by division of labor and social interests, by class struggle, or by cultural-linguistic identity. This was the more general framework of the debate on community and society.

John Locke emphasized the transformation from the traditional view on social relationships, based on natural right, to the modern thought based on rationalism and positive law, and replacing Aristotle's teleological view with the functional idea of collective security. The emphasis on collective security should be seen in the context of the growing role of private property. This development, where in political philosophy society became the preferred concept, did not mean that community disappeared. With the exception for political economy, where 'community' was narrowed down to trade company, the definition of society as community remained in everyday language until the end of the eighteenth century. Immanuel Kant viewed society, in its context of being based on reason and conventions among individuals, as a dynamic totality, the elements of which acted with and against one another. The 'unsocial sociality' (*ungesellige Geselligkeit*) of humans drove them to form societies. However, this drive towards convention-based collectivity provoked resistance, and the threat of the dissolution of the society was permanent. Kant's view was in this respect close to the school of English-Scottish moral philosophers (Adam Smith, David Hume, and Adam Ferguson).

Kant's question about a perfect society based on the free interaction among individuals was taken up by Johann Gottlieb Fichte at the beginning of the nineteenth century and put into a historical and linguistic-critical context. Although Fichte often used society and community as synonymous terms, he exposed the two concepts, and that of state, to linguistical-critical scrutiny observing the prevailing conceptual confusion.

As far as the German-speaking area is concerned, around 1800 Johann Gottfried Herder suggested a distinction between community and society. His point

of departure was the state-society connection, where 'state' emphasized the unity and closure of the collectivity, while 'society' had a focus on the multiple and open forms of collectivity including those based on the individual rights. In order to describe the relationship between state and church, Herder used society, based on Enlightenment reflection on the contextuality and historicity of religious forms. In order to emancipate the spirit and the mind, the *Geist*, from its entanglement with the state, and to rediscover the suprahistorical core of Christianity, Herder required the separation of society and community, *Gesellschaft* and *Gemeinschaft*, where the latter term was used to describe the community of spirits independent of the state. Herder's distinction, which again took up the old Greek and Biblical-Christian *χοινωνί'α* concept and infused it with new meaning, failed however to go beyond a specific religious vocabulary.

Kant's moral philosophy provoked criticism from G. W. F. Hegel, who rejected the idea of the autonomous individual. In Hegel's view the individual is certainly rational, although rationality is historical and trans-individual. The normative and legal order of a society could not have its point of departure in the idea of its individuals as *a priori* free, equal, and rational. Humans are free only within a strong moral community. Society is not based on convention but on moral. Such is Hegel's view. The human effort to be recognized as an individual results in the emergence of various communities, which are kept together by strong moral virtues, *Sittlichkeit*. The social community was linked to an ethical dimension, which, according to Hegel, appeared in various shapes through history, which was nothing but the development towards higher forms of ethical realization. The highest stage in this historical development was the state, which gave all its citizens true freedom through the confirmation and recognition of them in a meaningful totality. The individual is emancipated through subordination in the collective realization of *Sittlichkeit*.

From this condensed rational-romantic version of community, inspiration could a century later be picked up in Germany for extreme versions of the concept, where it was linked to the *Volk* ('people') concept and political action orientation. The difference between the individualistically oriented Kant and the collectivistic view of Hegel is not restricted to German history but is however more general and has been discernible in two alternative world views on community and society ever since. In the years around 2000 these alternatives are expressed in the debate between liberals and communitarians.

Emile Durkheim (1933 [1893]) developed Hegel's key question about the moral basis of modern societies in new directions. 'Modern society' is henceforth understood in terms of industrialization, division of labor, and democratization, processes that only had just begun when Hegel reflected on society, but were much more evident for Durkheim. His key question dealt with social solidarity, and how to bridge the distance between the state and the individual. The point of departure for Durkheim when he posed this question was the debate about the nature of the moral and political sciences. He was unsatisfied with the speculative character of political philosophy. He was convinced neither by the individualistic and utilitaristic approach developed by Kant or in the classical political economy, nor by Hegel's holistic enlightenment-romantic and mystical view on community, where the societies had goals, which were above and beyond those of the individuals. Durkheim looked in his positivist and empiricist approach for a position between Kant and Hegel for a society that was rather different than theirs. This debate was the very foundational moment of the social sciences in the contemporary sense.

Durkheim emphasizes the role of religion in premodern societies, in which, beyond the family, no other functionally cohesive institution other than religion existed. Religion provided 'mechanical solidarity,' which functioned as a social cement. This mechanical solidarity can be related to Hobbes's understanding of society (and nature) as clockwork. Post-Renaissance rationalism in the shape of Hobbes and others takes on new meaning when in the second half of the nineteenth century it was faced with a truly mechanical universe, i.e., the advances of physics, the division of labor, automized production, etc. In modern societies the division of labor changed the preconditions for social cohesion. A need for coordination of functional relationships emerged. The societies became ever more complex with growing divergence among individuals in their social orientation. Their interpretations of the world did not necessarily converge towards one collective will. The state was the only instrument for collective action when societies had developed a certain degree of complexity. Through the state organism reflexive thinking emerged. The 'stratification' of the ethnic nation was the modern beginning of this reflexivity, tearing the individual between instrumental-rational forms of belonging, and the organic, cultural, and social bonds. With this kind of reflection in the wake of the growing social division of labor organic solidarity arose alongside mechanical solidarity.

Durkheim problematizes the mediation between the individual and the society. The political philosophy of Enlightenment focused on concepts like people's sovereignty and political representation and on the issue of the institutional distribution of power. With intensified industrialization, capital concentration, the rise of the labor movement, and ideas of universal suffrage, the theoretical framework provided by the Enlightenment philosophers became insufficient. New forms of relationships between the individual and the state levels had to be developed. In this context

Durkheim evokes the notion of secondary institutions and develops ideas about what kind of organizational forms would fit an industrial society.

Taking his theoretical point of departure in his idea of division of labor as the basic principle of modern societies, he suggests interest-based representation through corporations rather than territory-based political representation. The corporations would reflect modern professions rather than traditional crafts and estates. In his functionalist approach Durkheim developed a moral theory in which he paid relatively little attention to questions of power and authority structures and social inequality. However his view on moral based on corporate representation and related to his idea of solidarity indicated the social question. The political function of the secondary institutions was both to keep the state sufficiently free from influences from the masses, so that the state could fulfil its role as the shaper of the society, and to protect the individuals from oppression by the state.

Hegel's problematization of the leading idea in moral philosophy and political economy, that society was based on the individuals and their agreement on conventions, constituted in the German-speaking area a step in the development from the traditional to the critical theory of society. The critical social theory, which emerged in particular in Germany around 1850, introduced new basic concepts and distinguished between society and community, between society and state, and between *bürgerliche* (the German term connotes in English translation both 'bourgeois' and 'civil') society and society. This critical social theory must be seen in the context of the social revolutionary movement, which emerged in the nineteenth century. This movement could be seen as a response to the crisis, which since the French Revolution, and with the industrial revolution, marked the everyday experiences in Europe and gradually also beyond Europe. Critical theory was an alternative reflection on society and community with respect to Durkheim's, based on ideas of functional differentiation and division of labor.

Politically the most long-term important element of critical theory was the conceptualization of a social history, or history of society, *Gesellschaftsgeschichte*, inscribed in the framework of the social sciences, by Karl Marx and Friedrich Engels. In their historical materialism, and in their emphasis on the political role of the proletariat, they depreciated the concept of society ('bourgeois'). They re-evaluated the community concept with its derivative communism, where socialism, with the same etymological origin as society, was seen as a transitional phase towards communism, the final stage, where history ceased. Marx and Engels criticized the view of society in earlier French and British social history, where society was seen as a set of mutual interactions. They criticized this view of society as being a society of exchange in which every member was a merchant. The society Marx and Engels discerned was societies in plural, where each of them was determined by the specific ways in which the production was organized. In each type of society specific power relationships prevailed. They followed one another in historical phases: the original society ('*Urgesellschaft*'), slave society, feudal society, bourgeois or capitalist society, socialist society, and communist community, where the merger into the old Greek *χοινωνι'α* was fulfilled in a new action- and future-oriented utopia.

A few years after the death of Karl Marx and after ever stronger experiences of industrial change and economic concentration, Ferdinand Tönnies (1887), analyzed quite differently, but still with a utopian embedding of the community concept, the connection between *Gemeinschaft* and *Gesellschaft*. In his view community took on a social romantic meaning connoting an archaic past based on the family and providing feelings of holism. His utopia was projected towards the past instead of the future. Society was the term used to describe the breaking-up of this past through industrialization, social disintegration, and experiences of atomism. Tönnies saw communism as a cultural system of community and socialism as the cultural system of society, i.e., communism was rather referred to as a historical or even archaic category, a kind of prototype, while socialism reflected the industrial society.

Tönnies' subordination of society to community had long-term political impact. Community was, in his view, in all sociological conceptualizations the superior and in all historical analyses the anterior concept. Community was the dialectical womb from which society arose in historical processes. The concept of society described the law-bound normal process of the decline of all community. The same process that Marx and Engels—in accordance with Hegel's idealistic philosophy and their critical-normative prognosis, where society became community—described as emancipation and a development towards higher stages of social organization, Tönnies analyzed as decay, as inexorable as the coming communism in the view of Marx and Engels. The only future hope that Tönnies could discern was weak: that the force of the community also within the era of society would remain although it was declining. The community could be promoted through the cooperative movement and through the union and solidarity idea in the labor movement.

Although Tönnies' concept of society includes the totality of the relationships among the individuals mediated through exchange, and in this respect can be historically situated, this historical location involves conceptual distinctions which cannot be justified either in genetic or in linguistic-critical terms. Both Max Weber and Georg Simmel observed this inconsistency and tried to solve it through a paradigmatic change of the sociological terminology. To consider social institutions like state, church, cooperative (*Genos-*

senschaft) as independent entities preceding the individual and attach them to a general society concept was in Weber's view a metaphysic residual, which needed to be overcome. He replaced society as a basic sociological concept with the term *Vergesellschaftung*, 'societalization,' an extremely general concept containing all forms of rationally motivated interest relationships under one umbrella concept. The generalization of the sociological conceptual framework around the term *Vergesellschaftung* certainly described processes and social change, but was basically a historical, where the analysis of the development was deprived of historical contextualization. Historical contexts were understood as 'cultural tragedy' (Simmel) or 'destiny of the rationalization of society through science' (Weber). The process of dehistorization of the concept of society that Tönnies began and Weber continued with his criticism of Tönnies had the effect of drawing growing attention to the community concept. *Gemeinschaft*, which can be seen as a kind of pseudo-historization, became topical. With the introduction of the concept of community in this way the sociological theory in Germany unintentionally opened up for the reactionary opposition against the industrial society. While in other European languages society and community remained synonymous concepts, *Gemeinschaft* became in Germany an ideological key concept connected to another highly ideologized crucial concept: the *Volk*. The ideological entanglement of the two concepts became a national-conservative instrument to mobilize the *völkische* movement after World War I against socialism, capitalism, and industrialism.

The cult and the dogmatization of the *Gemeinschaft* concept were also confirmed in German philosophy at this time. Max Scheler identified community as a person *sui generis*, 'a person of persons' as opposed to society, which was nothing but the sum of its individuals. He thus reinforced Tönnies' distinction. Community was based on blood, tradition, and historicity, while society was the waste product of internal disintegrative processes in the community. Society was synonymous with public, which continuously removed—Marx would have said alienated—the individual. Community was synonymous with personal solidarity, which reintegrated the individual.

Already before World War II, Talcott Parsons (1937) developed a view where community and society were seen as key concepts in structural-functional alternatives of value orientation, 'pattern variables': affectivity vs. affective neutrality, particularity vs. universalism, ascription vs. achievement, diffuseness vs. specificity. His approach was one, where, like in the German debate, the two concepts were separated, but for the rest his contribution was very different from the German debate.

In the 1950s René König (1955) was one of the leading figures in the reconstruction of the German sociology. He argued vehemently against Tönnies' dichotomy and use of the *Gemeinschaft* concept. He argued for an empirically oriented sociology with the American functionalism, where Talcott Parsons was a leading figure, as a source of inspiration, and warned against any attempt to conceptualize society as a totality, which only lead to speculation and dogmatism. Community in the German form of *Gemeinschaft* was in the 1950s rejected from the political and sociological vocabulary and society emerged instead in the 1960s as the key concept to describe and analyze the social organization.

At the same time as König in Germany warned against the *Gemeinschaft* concept community became the name of the European integration project in the 1950s. One of the key problems when this project was formulated was how to make Germany a strong ally in the Cold War and at the same time guarantee that the strength was controllable for its allies. The main architects of this design came from France, where the concern about the key problem for historical reasons was particularly great. There, *communauté*, for historical reasons, in the 1950s had a much more positive value charge than *Gemeinschaft* in Germany. In the form of the European Community in German translation, *Gemeinschaft* came back in as a political concept, which guaranteed the German incorporation in the West European Alliance in the Cold War, and which, therefore, was disconnected from its historical heritage in the German speech-area. When the European integration in the 1970s was exposed to strains, the idea of a European identity emerged connected to ideas of a supranational Europe. This intensification of the European Community concept was emphasized in the 1990s when community was changed for union. This European development can be understood as part of a modernization theory which is resurfacing with a communitarian twist, where culture is constructed around the idea of a community as a homogeneous discourse that somehow provides the glue to bind societies together. Such a theory of social integration, which takes for granted that cultural cohesion exists for national societies, and then simply transfers the idea of state-formation onto a supranational European level, built around the concept of European identity, echoes somehow Ferdinand Tönnies' idea of a sociopsychological community. After the Cold War the European Union is searching for new meaning and legitimization replacing the original rationale behind its foundation. This search is reflected on the conceptual level, where established concepts like community lose meaning and/or are charged with new content. Society seems in this respect to have become more stable and less ideologized.

The general theoretical framework of this discourse on European unity is the debate between libertarians and communitarians in social sciences since the 1970s, which debate, in turn, in a long historical view could be seen as continuity under transformation of the

bicentenary debate between Kantians and Hegelians. A point of departure for this theoretical debate was John Rawls's *A Theory of Justice* (1971). Erected in the tradition of contract theories and Kantian liberalism his theory bypassed the questions of political obligation and the welfare state through the idea of distributive rights in individualist terms. Utilitarian philosophers objected but in the long term the most substantive opposition to the neo-Kantian approach emerged from the 'communitarians.' The debate between the liberals and the communitarians became one of the most important issues in political philosophy in the 1980s. Methodologically the communitarians argued that the premise of individualism such as the idea of free and rational choice is wrong or biased and neglects the social, cultural, and historical contexts in which individuals act. In order to understand individuals one must look at their communities and communal relationships. In normative sense the communitarians argued that the premises of individualism give rise to morally unsatisfactory consequences. A. MacIntyre in *After Virtue* (1984) questioned the fundamental philosophical assumptions of liberalism. He and others questioned the claim of the priority of the right over the good and the image of the freely choosing individual. Political arrangements can not be justified without reference to common purposes. Freedom flourished within social interdependence not among free-floating individuals. However, the kind of community the communitarians drafted was one that often focused on the family, the tribe, or the neighborhood rather than on the nation, the state, or the class. The communitarian community is more than a mere association; it is a unity in which the individuals are members not in artificial or instrumental terms, but within its own intrinsic value. This is a view with similarities to Hegel's *Sittlichkeit* idea.

In the debate communitarianism has been dismissed both as little more than a corrective to the excesses of liberalism and as a form of old-fashioned collectivism. In the framework of the former view approaches have emerged, which can be described as a merger between liberalism and communitarianism. Michael Walzer sees communitarian thought as a permanently recurring corrective to liberalism (1983). Charles Taylor (1989) is another leading name in this approach. The trend at the end of the 1990s was that the divide that once seemed to separate communitarian and liberal theories began to close. Taylor and Walzer are often referred to as liberal communitarians. The theoretical debate has moved from the thick communitarian critique of thin liberalism to positions that combine elements of both. This trend has been nurtured not least from legal theory where a strong recent 'republican' turn aspires to hold both communitarian ideals and democratic participation together. The law is seen as sustaining community and promoting the common good to be negotiated through participation in a polity where liberal rights guarantee that the bargaining is uncoerced. Also the globalization discourse on the emergence of a globally rather than nationally organized economic order promoted the trend of closure between liberalism and communitarianism. The attempts to re-establish community, experienced as being lost to transnational globally operating forces, resulted in liberal politics of recognition of cultural alternatives in the name of multiculturalism.

Community has in this merger with liberal ideas been linked to the concept of citizenship, which borne in the French Revolution became a political key concept in the 1980s and 1990s. This linkage of community and citizenship is a dominant element not only of political and legal theory but of much current political debate as well, where both the Left and the Right try to establish positions of strength through various views, which emphasize both community and citizenship. This development is embedded in a broader theoretical and political reflection in the same direction around concepts like collective identity, network, trust, and social capital since the 1970s where the emphasis is on the historical emergence of reciprocal reliance among citizens. This trend around 2000 can be seen as attempts to establish a new merger between community and society from points of departure, which emphasize the role of the individual and plays down that of the state. Society is emphasized as civil society (in opposition to the state). The individual rather than the state has become the counter concept of society resulting in a migration of the society concept in the direction of the community concept, although this community concept is linked not only to the nation state but increasingly also to trans- and subnational levels of social and political organization.

See also: Communitarianism: Political Theory; Communitarianism, Sociology of; Community, Social Contexts of; Community Sociology; Community Studies: Anthropological; Durkheim, Emile (1858–1917); Hegel, Georg Wilhelm Friedrich (1770–1831); Individual/Society: History of the Concept; Individualism versus Collectivism: Philosophical Aspects; Kant, Immanuel (1724–1804); Law and Society: Sociolegal Studies; Libertarianism; Locke, John (1632–1704); Society/People: History of the Concept; State and Society

Bibliography

Avineri S, de Shalit A (eds.) 1992 *Communitarianism and Individualism*. Oxford University Press, Oxford, UK

Brunner O, Conze W, Koselleck R 1979 *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*. Band 2. E Klett-Cotta, Stuttgart, Germany

Christodoulidis E A (ed.) 1998 *Communitarianism and Citizenship*. Ashgate, Aldershot, UK
Durkheim E 1933 [1893] *The Division of Labour in Society*. MacMillan, London
Etzioni A (ed.) 1995 *New Communitarian Thinking. Persons, Virtues, Institutions, and Communities*. University Press of Virginia, Charlottesville, VA, and London
Gutman A, Taylor C (eds.) 1994 *Multiculturalism*. Princeton University Press, Princeton, NJ
Jaume L 2000 *La liberté et la loi. Les origines philosophiques du liberalisme*. Fayard, Paris
Koselleck R 1989 [1954] *Critique and Crisis*, 1st edn. MIT Press, Cambridge, MA
KönigR 1955 Die Begriffe Gemeinschaft und Gesellschaft bei Ferdinand Tönnies. Kölner Zeitschrift für Soziologie und Sozialpsychologie **7**
Lukes S 1985 *Emile Durkheim. His Life and Work*. Stanford University Press, Stanford, CA
MacIntyre A 1984 *After Virtue*. University of Notre Dame Press, Notre Dame, IL
Parsons T 1949 [1937] *The Structure of Social Action*. Free Press, Glencoe, IL
Poggi 2001 *Durkheim*. Oxford University Press, Oxford, UK
Powell W 1990 Neither market nor hierarchy: network forms of organization. In: Staw P B M, Cummings L (eds.) *Research in Organizational Behaviour*. JAI Press, Greenwich, CT, Vol. 12
Putnam R D (ed.) 2000 *A Decline of Social Capital? Political Culture as a Precondition for Democracy*. Princeton University Press, Princeton, NJ
Rawls J 1971 *A Theory of Justice*. Oxford University Press, Oxford, UK
Rorty R 1989 *Contingency, Irony and Solidarity*. Cambridge University Press, Cambridge, UK
Skinner Q 1998 *Liberty Before Liberalism*. Cambridge University Press, Cambridge, UK
Taylor C 1979 *Hegel and Modern Society*. Cambridge University Press, Cambridge, UK
Taylor C 1989 *Sources of the Self: The Making of the Modern Identity*. Harvard University Press, Cambridge, MA
Tönnies F 1991 [1887] *Gemeinschaft und Gesellschaft. Grundbegriffe der reinen Soziologie*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany
Wagner P 1994 *A Sociology of Modernity: Liberty and Discipline*. Routledge, London
Walzer M 1983 *Spheres of Justice: A Defence of Pluralism and Equality*. Robertson, New York

B. Stråth

Community Sociology

1. The Term Community

There are many meanings and uses of the term 'community.' Some have explored the implications of creating a 'global community' with governance based on human tolerance and understanding (The Commission on Global Governance 1995). Others have viewed economic 'communities' that promote national and corporate interests. A growing body of literature on rural communities, along with the International Rural Sociology Association, reflects the interests of researchers covering a range of educational, environmental, and social themes in developed and developing nations. However, this entry covers urban research concerning 'community' because such a selective approach enables a sharper thematic focus than would otherwise be possible with an all-encompassing survey of the wider research.

Today, contending theoretical perspectives complicate the task of arriving at a clear understanding of communities. Because the concept of 'community' is open to wide interpretation, it is frequently used as a term connected to urban citizenship (Lowndes 1995) or as a reference point in complex discourses about identity, place, and social meaning. For example, ethnic, racial, and gender affinities between groups may be regarded as aspects of 'community identity,' but these take on complex forms within the context of a wider range of cultural and social class affinities (Solomos 1993). Similarly, reference to geographic or 'place-based communities' implies the existence of groups with territorial interests or associations or communities based on formal local government boundaries, but place often plays only a small or even nonexistent part in locating 'community identity'. It is possible for individuals or groups to identify with others with whom they share particular attributes even though only vague geographical markers refer to a perceived 'community.'

Therefore, the diverse relationships between place and identity make theoretical conceptions of community highly contingent, especially when account is taken of global capital operating across national and local boundaries. Multifaceted economic and social change has challenged the basis of traditional social arrangements and attitudes (Soja 1996) and undermined old theoretical certainties and disciplinary boundaries in the discussion of place and identity (Allen et al. 1998) and in broader urban theory (Fainstein and Campbell 1996). The relationship between 'the local' and 'the global' is also more problematic as transnational corporations 'embed' their operations in localities and regions (Amin and Thrift 1994, Amin 1994) making it possible to define influential networks of corporations (Friedland 1982) and partnerships involving public and private sector groups. Elite groups are often seen as representing social class interests (Scott 1997) rather than more narrowly defined community interests. The question remains however, that despite profound social and economic changes is it still plausible to assume, as Etzioni (1993) does, that local communities can represent an alternative to radical individualism by providing individuals with a sense of purpose and a moral framework within which citizens can prosper?

2. Diverse Perspectives

This question is not new, and it relates to the basis of communities and social cooperation within them. Urban sociologists of the Chicago School in the 1920s and 1930s concentrated on the 'human ecology' of urban life and the social relationships and forms of organization in the metropolis. Robert Park, influenced by Georg Simmel and Émile Durkheim, presented a complex view of neighborhood organization where co-operation and competition underlay relationships between groups. The resulting tensions undermined social control, but communities achieved order through social specialization and interdependence. Later, Park (1952) emphasized the inherent 'biotic' competitiveness of urban life and the struggle to control scarce resources. Other Chicago School sociologists referred to human struggle within cities and the consequent adaptation of social organization to external environmental change. For example, Louis Wirth (1938) argued that cities represented the agglomeration of human activity, and that the social specialization of cities led to complex patterns of social segregation. Despite recognizing the coming together of groups with shared identities, Wirth stressed the superficiality of social relations and the resulting social psychology of urban life where class and ethnic affinities were less important than social heterogeneity. For Wirth, cities were large and impersonal, and divisions arose between people producing different ways of life that made true consensus and stable community life difficult to achieve (see Reiss 1964).

Wirth remained influential after World War II, but sociologists outside the human ecology tradition argued that the urban experience could be socially defined better by reference to the forces transforming traditional social class and community relationships. In the UK, Young and Willmot (1957) discovered strong working class community and family identities in the East End of London that contrasted with new housing projects where relationships were less intimate and not so cohesive. In the USA, Gans (1962) studied Boston's West End where a multiethnic 'urban village' had developed with a rich variety of social and political institutions.

Political scientist Robert Dahl (1961) provided the classic pluralist interpretation of local politics in a study of New Haven, CT, USA. There, 'polyarchy' represented the rule of the many where power was unequally distributed, but where groups could influence crucial local political decisions. However, elite theorists, disagreeing with pluralism, engaged in the 'community power debate' by arguing that local power-holders dominated the policy process by excluding many local community groups. The pluralist perspective was further questioned as social unrest affected big American cities and class conflicts and social exclusion defined the condition of large, disadvantaged cities. Urban riots in the 1960s, the expansion of militant community politics, and the rise of the black power movement highlighted an urban crisis while public policy shifted to favor public spending on aid programs for run-down communities. But, political scientist Edward Banfield (1970) provided a classic critique of government urban policies through controversial research that highlighted cultural trends in *The Unheavenly City* that undermined stable and cohesive urban communities.

Against the backdrop of the 'urban crisis,' research increasingly focused on race, social disadvantage, and the inaccessibility of local bureaucracies. Rex and Moore (1967) adopted a neo-Weberian approach viewing ethnic minority access to housing in the city of Nottingham, UK, where peoples' goals were influenced by the interaction of social and spatial conditions. Pahl (1975), also from a neo-Weberian perspective, regarded urban state bureaucracy as a mechanism restricting the access of local people to urban services.

There was also attention to broader urban processes. Gottdiener (1994a, 1994b) refers to the 'New Urban Sociology' in the 1970s and 1980s that integrated the analysis of social class, gender, cultural, economic and political change by building upon the work of neo-Marxist writers such as Castells (1977) and Harvey (1973). Castells studied the functions of cities and the role of social classes in the reproduction of capitalism. His early work was conducted at a time when the welfare state had expanded under capitalism. For Castells, local communities economically contributed to capitalist reproduction, so cities and regions were spaces supporting economic production as well as social organization.

Despite such contributions, the opposition to extensive state involvement in community development from the new right was echoed in research focused on the social problems of cities allegedly caused by government failure. Charles Murray's (1984) controversial notion of the 'underclass' identified the long-term plight of a growing 'underclass' of permanently welfare-dependent people in deprived urban communities. Community empowerment came high on the new right political agenda in the USA as an alternative to welfare dependency and government-funded solutions to urban problems. Conservatives argued that empowered communities could become more self-sufficient and better able to increase their capacities to make changes for themselves through free enterprise and local initiative.

3. Contested Boundaries

During the 1980s and 1990s, urban sociologists continued to draw on different paradigms and develop cross-disciplinary themes. The diversity of papers presented to panels organized for the Community Research Committee at the 1998 World Congress of Sociology illustrates the interdisciplinary interests of

urban and community researchers. The panels displayed the varied perspectives of sociologists, political scientists, policy analysts, geographers, and others. Contributions reflected the contested boundaries between disciplines and the broad horizons of researchers variously concerned with the relationships between social groups, local communities, post industrialism, post modernity, and globalization. Contemporary research has thus generated new insights into the spatial dimensions of urban change and the complexities of the issues confronting local communities.

Many sociologists, geographers, and political scientists argued that global social and economic change, multinational corporations, and financial institutions substantially influenced the urban agenda and the fate of local communities. Innovative research provided fresh and controversial perspectives on the changes affecting the global, regional, and local communities (e.g., Castells 1996, Soja 1996). Tensions arising from the expansion of local global capital and the defense by groups of local interests underscore the profound changes affecting communities. Local communities confront major changes as corporations pursue new strategies, regional economies restructure, and traditional social relations breakdown. In the new information economy, old certainties no longer hold. With economic reordering, there are social and political changes. Some see communities linking to governing corporate-dominated 'urban regimes' (Stone 1989). Others identify 'growth machines' (Logan and Molotch 1987, Ferman 1996). There have also been social dislocation, increased ethnic and racial competition, and discrimination against minorities. The notion of stable and harmonious neighborhoods is questioned as economic and social change fractures traditional communities and generates new urban problems and patterns of social intercourse. Studies of social disorder and urban riots, as in Los Angeles in 1992 (Baladassare 1994), illustrate the racial and ethnic cleavages that can result from rapid urban restructuring. Communal divisions often eclipse class solidarity, especially when groups compete for scarce resources and attempt to access urban services. Indeed, the 'community' is threatened as social change undermines consensus and stability. Political scientist, Robert Putnam (2000) argues that there is a decline in social capital in modern communities in the USA. Putnam argues that social capital, represented as the networks, norms, and trust in communities that help people to work for shared objectives, is in decline despite the efforts of governments and community leaders to increase the active participation of people in civic affairs.

4. *The New Political Culture*

An important theme in recent Community Research panels at the World Congress of Sociology has been the emergence of a new political culture (NPC). The NPC views political fragmentation against a background of fundamental social change affecting urban communities worldwide. According to sociologist Terry Nichols Clark and political scientist Vincent Hoffmann-Martinot (1998), the NPC increases the salience of issues such as the environment, citizen democracy, gay rights, abortion, and lifestyle politics. The politics of post industrialism places less emphasis on class and workplace politics and greater stress on single issues championed by diverse groups within a complex and changing sociopolitical environment. The NPC challenges traditional political parties and old state hierarchies and undermines the assumptions of the post war welfare state. People are more skeptical about large public bureaucracies and likely to support human-focused policies, smaller and more localized government units, and more responsive forms of local democracy. The NPC emerges 'more fully and forcefully in cities and countries with more highly educated citizens, higher incomes, and high-tech service occupations' (Clark and Hoffmann-Martinot 1998, p. 3). This marks a change in the ways local communities articulate social and political demands. Local groups are more likely to organize across social class lines as they adopt new methods of social and political action. Information technology opens new horizons and provides opportunities for more effective networking and communication.

5. *Networked Communities*

Technological change therefore produces new patterns of social life. In his later work, Castells combines an analysis of the global and local effects of information technology through the rise of a 'network society.' Castells, acknowledging Howard Rheingold (1993), argues that new information technologies integrate the world through global networks and a 'vast array of virtual communities' (Castells 1996, p. 22) that network groups and individuals by way of 'politics and religion to sex and research.' The Internet and 'electronics grass-roots culture' (Castells 1996, p. 354) enables individuals to interact globally and interactively, but also privately. For Castells, a new globalized and customized mass media permeates all levels of society directly affecting the local community and the individual.

Together with technological networks, social and inter-organizational policy networks influence local life and community organization. Social network analysis originated in social psychology and extended to sociology. White (1963) and Fischer (1982) developed social interaction models showing how networking influenced local social organization. Sustained sociological interest in the influence of complex networks in local communities (Scott 1991) accompanied interest in the role of networks in political

systems (Dowding 1995). Political scientists have argued that the collaboration of various well-organized corporate groups in local networks, together with public–private partnerships, help to bring about government policies favorable to those groups, sometimes with little local community involvement. Influential groups gain access to decision-makers and thereby manage to raise crucial issues up the political agendas of local governments.

6. *Communitarianism*

Studies of technological change, globalization, political networking, social capital, and the NPC all imply discontinuity and uncertainty in the development of local communities despite the different theoretical approaches adopted. Etzioni (1993), and a group of like-minded communiterians, have responded to the impersonal forces of globalization, social dislocation, and the undermining of community and family values by calling for a new 'moral, social, public order.' They seek a stronger sense of community to counter what they regard as a threat to the moral underpinnings of community life. Etzioni argues for a reassertion of community values by creating a new 'spirit of community' through rights and responsibilities that strengthen the moral foundations of society. Following the publication of the journal *The Responsive Community: Rights and Responsibilities*, the group advocated an alternative to market individualism by fostering values favoring mutual support and the family.

See also: Class: Social; Collective Identity and Expressive Forms; Communitarianism, Sociology of; Community Power Structure; Community, Social Contexts of; Community/Society: History of the Concept; Cooperation: Sociological Aspects; Ethnic Groups/Ethnicity: Historical Aspects; Ethnology; Ethnonationalism: Cultural Concerns; Identity and Identification: Philosophical Aspects; Identity Movements; Identity: Social; Mass Media and Cultural Identity; Multiculturalism and Identity Politics: Cultural Concerns; Neighborhood: General; Networks: Social; Power in Society; Race Identity; Regional Identity; Religion: Nationalism and Identity; Social Identity, Psychology of; Solidarity: History of the Concept; Underclass; Urban Sociology

Bibliography

Allen J, Massey D, Cochrane A et al. 1998 *Rethinking the Region*. Routledge, London and New York

Amin A (ed.) 1994 *Post-Fordism: A Reader*. Blackwells, Oxford, UK and Cambridge, MA

Amin A, Thrift N 1994 Living in the Global. In: Amin A, Thrift N (eds.) *Globalization, Institutions, and Regional Development in Europe*. Oxford University Press, Oxford, UK, pp. 1–22

Banfield E C 1970 *The Unheavenly City: The Nature and Future of our Urban Crisis*. Little, Brown, Boston, MA

Baldassare M (ed.) 1994 *The Los Angeles Riots: Lessons for the Urban Future*. Westview Press, Boulder, CO

Castells M 1977 *The Urban Question*. Edward Arnold, London

Castells M 1996 *The Rise of the Network Society*. Blackwells, Oxford, UK

Clark T N, Hoffmann-Martinot V (eds.) with Gromala M 1998 *The New Political Culture*. Westview Press, Boulder, CO

The Commission on Global Governance 1995 *Our Global Neighbourhood*. Oxford University Press, Oxford, UK

Dahl R A 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT

Dowding K 1995 Model or metaphor? A critical review of the Policy Network Approach. *Political Studies* **43**(1): 136–58

Etzioni A 1993 *The Spirit of Community: Rights, Responsibilities and the Communitarian Agenda*, 1st edn. Fontana Press, London

Fainstein S S, Campbell S (eds.) 1996 *Readings in Urban Theory*. Blackwells, Cambridge, MA

Ferman B 1996 *Challenging the Growth Machine: Neighborhood Politics in Chicago and Pittsburgh*. University of Kansas Press, Lawrence, KS

Friedland R 1982 *Power and Crisis in the City: Corporations, Unions and Urban Policy*. Macmillan, London and Basingstoke, UK

Fischer C S 1982 *To Dwell Among Friends: Personal Networks in Town and City*. Chicago University Press, Chicago

Gans H J 1962 *The Urban Villagers*. Free Press of Glencoe, New York

Gottdiener M 1994a *The New Urban Sociology*. McGraw-Hill, London

Gottdiener M (ed.) 1994b *The Social Production of Urban Space*, 2nd edn. University of Texas Press, Austin, TX

Harvey D 1973 *Social Justice and the City*. Edward Arnold, London

Logan J, Molotch H 1987 *Urban Fortunes: The Political Economy of Place*. University of California Press, Berkeley, CA

Lowndes V 1995 Citizenship and urban politics. In: Judge D, Stoker G, Wolman H (eds.) *Theories of Urban Politics*. Sage, London, pp. 160–80

Murray C 1984 *Losing Ground*. Basic Books, New York

Pahl R E 1975 *Whose City?* 2nd edn. Penguin, Harmondsworth, UK

Park R E 1952 *Human Communities*. Free Press, Glencoe, IL

Putnam R D 2000 *Bowling Alone: The Collapse and Revival of American Communities*. Simon and Schuster, New York

Reiss Jr. A J (ed.) 1964 *Louis Wirth On Cities and Social Life: Selected Papers*. University of Chicago Press, Chicago, IL

Rex J, Moore R 1967 *Race, Community and Conflict*. Oxford University Press for Institute of Race Relations, London

Rheingold H 1993 *The Virtual Community*. Addison-Wesley, Reading, MA

Scott J 1991 *Social Network Analysis: A Handbook*. Sage, London

Scott J 1997 *Corporate Business and Capitalist Classes*. Oxford University Press, Oxford, UK

Soja E W 1996 *Thirdspace: Journeys to Los Angeles and Other Real-and-Imagined Places*. Blackwells, Malden, MA

Solomos J 1993 *Race and Racism in Britain*, 2nd edn. Macmillan, Houndmills, UK

Stone C L 1989 *Regime Politics: Governing Atlanta 1946–1988*. University Press of Kansas, Lawrence, KS
White H C 1963 *An Anatomy of Kinship*. Prentice-Hall, Englewood Cliffs, NJ
Wirth L 1938 Urbanism as a way of life. *American Journal of Sociology* **44**: 3–24
Young M, Willmot P 1957 *Family and Kinship in East London*. Free Press, Glencoe, IL

B. D. Jacobs

Community Studies: Anthropological

The terms community and community studies in anthropology have two main denotations. One is generic, echoing the interest anthropology and anthropologists have always had in ordinary, stable, small-scale localized collectives. The other pertains to studies of rural populations (e.g., villages, parishes, counties) and urban enclaves (e.g., neighborhoods and quarters) within industrialized, developed Western countries.

1. *The Concept of Community in the Social Sciences*

Community, one of the most ubiquitous and imprecise terms in the social sciences, is usually associated with an array of positive connotations such as solidarity, familiarity, unity of purpose, interest, and identity. Earlier sociological debates of communities and their social significance stressed either a territorial base, i.e., groups whose members share a well-defined location (see Hawley 1950), or, later, community as state of mind—a collective sharing of an underlying feeling of sameness and belonging regardless of geographical dispersion (see Bell and Newby 1974). The latter option, while analytically promising and intellectually suggestive, produces a perplexing category. Since affiliation and identification with nonterritorial communities can be unclear even to the actors themselves—let alone to social scientists attempting to characterize and study them—the concept can often be used and interpreted as an omnibus for a variety of intended and unintended meanings.

Not surprisingly, some sociological attention was given to the distinctions between the informal nature of communities and the more rigid structure of organizations (Gottshalk 1975, Azarya 1984) as well as to the problem of indeterminacy of community boundaries (Hillery 1955). These notwithstanding, community studies emerged since the 1920s and well into the 1980s as a solid branch of sociological inquiry, not least through the consistent preoccupation of the Chicago school with modernization and its impacts on urban culture, subcultures, and urban society.

2. *Community (Re)Enters Anthropology*

Anthropology, with its early emphasis on intensive field study focusing on small-scale groups, whereby culture and society were assumed to be unproblematically linked to bounded space, had an inherent affinity with the view of communities as bounded territorial units. This may explain why the concept of community was not explored by anthropologists in earnest, receiving little explicit analytical and theoretical attention prior to the 1980s. Habitually working with bounded communities, anthropologists tended to develop context-specific tools for their analysis. Early emphasis on customs and manners, the subsequent emergence of the concept of culture, the paradigms of structural functionalism, transactionalism, symbolic interaction and structuralism *à la* Lévi-Strauss all offered, in their respective heydays, rich and convincing frameworks for studies of small-scale bounded groups.

Thus, while the term community was always present in the anthropological lexicon, its syntactic status tended to be relegated to noun or synonym. It remained a finite, noninterpretative idiom and was arrested from becoming a dynamic theoretical tenet. In short, engulfed in 'native' ways of explaining solidarity and social relations, and taken by the concept of an immutable culture occupying an ethnographic present, anthropologists had little use for the community. And while Lynd and Lynd's classic study of Muncie (eternalized as *Middletown* in Lynd and Lynd 1929), Lloyd Warner's Yankee City volumes (Warner and Lunt 1941, Warner and Low 1947) and Arensberg's work on the relationship between community and culture (see Arensberg and Kimball 1965) undoubtedly inspired ethnographically oriented studies such as Robert Redfield's *The Little Community* (1955), Marriott's edited volume on villages in India (Marriott 1955) or Frankenberg's *Communities in Britain* (1965), a strong and clear anthropological tradition of community studies was slow to materialize.

Two factors may have contributed to the emerging interest of anthropologists in the 1980s in the community as an analytical and theoretical category. One was the aforementioned emergence in the 1960s and 1970s within sociology of the view of the community as a more abstract entity, whereby meaning and a sense of solidarity became more salient than tangible physical contact. The other was a growing need amongst anthropologists, by then more willing than ever before to study societies and cultures in the core and on the margins of Europe, to develop a theoretical toolbox for this new challenge.

An edited volume and an ethnography by Anthony Cohen (1982, 1985) indicated the efficacy of the concept of community for a more sophisticated understanding of the relationship between local rural and urban communities and wider affiliations and identity,

not least the ones connected to the nation state. The edited volume (Cohen 1982) had six anthropologists writing two essays each about rural communities in Britain. One essay was incorporated into a section of the book entitled 'Belonging to the Part: Social Association within the Locality,' while the other was included in a section called 'Belonging to the Whole: The Community in Context.' The sections corresponded to the dual theme of the inquiry: how people develop a sense of belonging in a small community; and how they develop an identification and a set of practices that reflect their sense of place in the broader realm that lies beyond the local. The first theme naturally focuses on kinship, sect, regional histories, and affiliations. The second theme, pursued in Cohen's subsequent work, leads to a growing preoccupation with cultural boundaries and borders.

In terms of theory and analysis, the ethnographic attention paid by Cohen and his contributors to (mainly Celtic) rural Britain owes much to Clifford Geertz's emphasis on meaning and cultural interpretation. The community, according to Cohen's introduction to the volume, must be studied and interpreted not so much as a structured social organization but as an arena where experience is ordered, partly through a 'local' culture. This in turn produces a distinct local identity which is engaged in constant interplay with broader cultural formations, processes, and signals. Cohen's volume thus signposted a new field of inquiry, the form and theoretical approach of which were different from those of the adjoining anthropological arena of peasant studies, the latter focusing mainly on rural communities in semiperipheral areas, including the margins of Europe and Latin America.

The importance attributed by Cohen to meaning and imagination, and the Wittgensteinian supremacy he advocates of context of use over lexical occurrence, becomes even clearer in *The Symbolic Construction of Community* (Cohen 1985). Simple people too, he shows, are preoccupied with symbols, constantly using them to make comparisons (and sense) of appearance and reality, likeness and diversity, similiarity and difference. Once more, social structure collapses at the feet of culture, subverting the monopoly of sociology over the concept of community. Significantly, more emphasis than ever before was now directed at the borders. Writers came to recognize that it is near boundaries, and across them, that community culture and identity were being determined.

Community studies reduce methodological, analytical, and theoretical differences between anthropology and sociology. Conrad Arensberg's works of the 1950s and 1960s are by no means wholly sociological, and the same is true for Maurice Stein's *The Eclipse of Community* (Stein 1960). By the same token, the strongly ethnographic turn of Warner's *Yankee City* volumes, West's *Plainville* (West 1961) or the Lynds' *Middletown* (Lynd and Lynd 1929) probably warrant counting them as anthropological, while the preoccupation displayed by Cohen (1982, 1985), Frankenberg (1965) and Hannerz (1980) with broad sociological issues merit labeling their works at least partly sociological.

Some writers have alluded to an obvious shortcoming of anthropological (and, for that matter, sociological) studies of communities—their propensity to be a-political. Rich in analyses of inherent traits and characteristics, such studies lend themselves willingly to elaborate treatments of culture, cultural identity, and demarcation. This, however, can often serve as a cover for shying away from and even blindfolding themselves to issues of internal and external power disparities, political economy, and historicity (cf. Aronoff 1984, Cox 1987).

A strength that anthropological studies of communities have had, at least since the 1980s, has been their emphasis on abstract experiences, not least imagination, as idioms shaping human identity and agency. Major inspiration for this project undoubtedly came from Benedict Anderson's work, not least from his creative employment of this notion in his seminal work on nationalism, *Imagined Communities* (Anderson 1983). The focus on imagination resurfaced later in attempts to explore new ways in which anthropology could tackle the paradoxes and complexities of transnational experience and culture.

3. *Critical Anthropology and the Concept of Community*

Like 'ethnography,' 'the ethnographic present,' 'fieldwork,' 'informants,' and other key anthropological idioms, 'community' became, in the 1990s, a construct. Idealized as a composite, unproblematized pillar of old guard anthropology, it could now serve as a hammering block against which critical ideas could be articulated.

Malkki (1997) looks at the connection between the concept of community and anthropology's propensity to look for the ordinary, durable, 'everyday' routines and practices. This approach, which gained anthropology justified rewards and insights, gradually bred an expectation that all people studied by ethnographers are not just haphazard collections of humans somehow thrown together into the same location. Rather, they were expected to have always been part of a stable, permanent and—in particular—localized 'community.' But while such groups do of course exist, the emphasis on communities pushed anthropologists to overlook other possibilities, including those where life does create transient collections of humans randomly united. Such, Malkki (1997, p. 91) indicates, are what Barbara Myerhoff called *accidental communities*—Woodstock 1968 being a dramatic instance—and what Malkki herself found in refugee camps in Rwanda, where displaced people from a

variety of places of origin huddled together, reinventing a sense of communal rootedness they may or may not have had in the lives they had before their dislocation.

Arjun Appadurai (1996), in his seminal work on the relationship between bounded place and culture in a world that is rapidly becoming dislocated, transnational, and diasporic, likewise uses an idealized icon of the term community in anthropology to develop a more sophisticated concept of culture. Localities and communities, he asserts, must be historicized and contextualized, identified as processes of growth and shrinkage. Not least, anthropology should—and can—become more sensitive to what he calls the global production of locality.

Traditional concepts of community in the social sciences, as they had been applied to studies of rural Mexico, give Roger Rouse's (1991) study of migration into the USA a much needed structure against which to prop his concept of transnational migration. Earlier migration studies, he aptly indicates, had a bipolar picture, with culturally and demographically stable and homogeneous collectives on both ends. Immigrants, within this framework of analysis, simply moved—more or less successfully—between two 'communities.' His own ethnography, however, convinced him that what was taking place in fact was a more complex phenomenon, a translocal movement backward and forward. This productively subverts the wholesome view of the community—be it in rural Mexico or in the migrants' quarters of western USA—offering new tools for understanding of these complex and uneasy situations.

See also: Community, Expression of; Community, Social Contexts of; Community Sociology; Practice: Anthropological Aspects

Bibliography

Anderson B 1983 *Imagined Communities*. Verso, London

Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN

Arensberg C, Kimball S 1965 *Culture and Community*. Harcourt Brace, New York

Aronoff M 1984 Review of Cohen 1982 *Belonging: Identity and Social Organization in British Rural Cultures*. *Contemporary Sociology* **13**(1): 46–7

Azarya V 1984 *The Armenian Quarter of Jerusalem*. University of California Press, Berkeley, CA

Bell C, Newby H (eds.) 1974 *The Sociology of Community: A Selection of Reading*. Frank Cass, London

Cohen A 1982 *Belonging: Identity and Social Organization in British Rural Cultures*. Manchester University Press, Manchester, UK

Cohen A 1985 *The Symbolic Construction of Community*. Methuen (Tavistock), New York

Cox G 1987 Review of Cohen 1985 *The Symbolic Construction of Community*. *Sociologia Ruralis* **27**(1): 76–7

Frankenberg R 1965 *Communities in Britain: Social Life in Town and Country*. Penguin Books, London

Gottshalk S 1975 *Communities and Alternatives: An Exploration of the Limits of Planning*. Shenkman, Cambridge, MA

Hannerz U 1980 *Exploring the City: Inquiries Toward an Urban Anthropology*. Columbia University Press, New York

Hawley A 1950 *Human Ecology: A Theory of Community Structure*. Roland Press, New York

Hillery G A 1955 Definitions of community: Areas of agreement. *Rural Sociology* **20**(2): 111–23

Lynd R S, Lynd H M 1929 *Middletown: A Study in American Culture*. Harcourt Brace, New York

Malkki L 1997 News and culture: transitory phenomena and the fieldwork tradition. In: Gupta A, Ferguson J (eds.) *Anthropological Locations*. University of California Press, Berkeley, CA, pp. 86–101

Marriott M (ed.) 1955 *Village India: Studies in the Little Community*. University of Chicago Press, Chicago

Olwig K, Hastrup K (eds.) 1997 *Siting Culture: The Shifting Anthropological Object*. Routledge, London

Redfield R 1955 *The Little Community: Viewpoints for a Study of a Human Whole*. University of Chicago Press, Chicago

Rouse R 1991 Mexican migration and the social space of postmodernism. *Diaspora* **1**(1): 8–23

Stein M 1960 *The Eclipse of Community: An Interpretation of American Studies*. Princeton University Press, Princeton, NJ

Warner L, Low J O 1947 *Social System of the Modern Factory*. Yale University Press, New Haven, CN

Warner L, Lunt P 1941 *The Social Life of a Modern Community*. Yale University Press, New Haven, CN

West J 1961 *Plainville, USA*. Columbia University Press, New York

D. Rabinowitz

Comorbidity

The term 'comorbidity' was coined by Feinstein (1970) to refer to the situation in which an individual has two or more physical and/or mental illnesses. Comorbidity is clinically important for three reasons. First, some types of comorbidity complicate diagnosis because presenting complaints are a mix of symptoms of different comorbid conditions. Second, some types of comorbidity complicate treatment because standard therapies for one disorder in a comorbid set are contra-indicated for patients with another disorder in the set. Third, some types of comorbid disorders can magnify the functional impairment and adversely influence the course of other conditions.

1. The Causes of Comorbidity

Some comorbidities occur purely by chance. For example, given that approximately 15 percent of the

adult population in the United States suffers from seasonal allergies and that approximately 20 percent suffers from arthritis, one would expect to find 3 percent of adults who suffer from both disorders merely on the basis of chance association (i.e., 15 percent of 20 percent is 3 percent). There are also situations, though, in which certain comorbid disorder clusters occur at higher than chance levels.

Nonrandom comorbidity can be brought about by at least three broad classes of processes. The first is when one disorder directly effects the onset of a second disorder. For example, persistent alcohol abuse can lead to cirrhosis of the liver. A second class of processes involves indirect effects of one disorder on the onset of a second disorder. The stresses associated with the threat and lifestyle changes that follow from receiving a diagnosis of heart disease, for example, can predispose to the onset of an anxiety disorder. A third class of processes that can lead to comorbidity involves common causes. Traumatic life events, for example, can predispose to a wide range of comorbid anxiety and mood disorders. Biological factors can also be common causes. Twin studies, for example, show that there are common genetic factors that play an important part in bringing about the strong comorbidities found among many commonly occurring mental disorders (Kendler et al. 1995).

2. *Basic Patterns of Comorbidity*

Three types of comorbidity have been of particular interest to clinical psychologists: comorbidity between different mental disorders; comorbidity between mental disorders and substance use disorders; and comorbidity between mental disorders and physical disorders. All three types of comorbidity are quite common in the general population. Nearly half of all people with a mental disorder in the general population meet criteria for two or more such disorders, while more than half of people with a substance use disorder and as many as three out of four patients in treatment for substance abuse or dependence also meet criteria for a mental disorder (Kessler et al. 1996). Mental disorders, especially anxiety and depression, are also found in a substantial minority of people in treatment for physical illness and especially among high utilizers of primary care (Miranda et al. 1994).

2.1 *Mental–Mental Comorbidities*

Bivariate associations between pairs of hierarchy-free *DSM-III-R* disorders in general population epidemiological surveys reveal three general patterns (Kessler 1997). First, there is a generally positive pattern of associations among the vast majority of commonly occurring anxiety, mood, personality, and substance disorders. Second, associations are generally stronger within than between broad classes of disorders (e.g., anxiety disorders are generally more strongly related to each other than to mood disorders). Third, extremely strong comorbidities are found among disorders in clinically recognized syndrome clusters, such as panic and agoraphobia, double depression (major depression and dysthymia), and bipolar disorder (major depression and mania).

Some mental–mental comorbidities appear to be brought about by indirect causal processes, as in the case of panic predisposing to agoraphobia. Others appear to be due to common underlying causes, as in the case of the cycling episodes of depression and mania characteristic of bipolar disorder. Causes are less clear in other types of mental–mental comorbidity. There is currently a great deal of interest in using longitudinal twin and family research designs to investigate some of the more poorly understood of these comorbidities in an effort to refine nosological distinctions.

Many of the strongest comorbidities involve combinations of anxiety and depression—panic, generalized anxiety disorder and social phobia with depression. The anxiety disorder usually starts at an earlier age than the depression in all these cases. Whether this is true because anxiety is the primary disorder or because some types of comorbid anxiety depression are a single disorder in which anxiety is the earliest manifestation is not clear. The fact that certain antidepressant medications have been found to be effective in treating some anxiety disorders but not others, and the fact that the types of anxiety disorders that are affected vary from one antidepressant to another, strongly suggest that complex common neurobiological pathways are involved in these comorbidities. Although good progress is being made in tracing out these pathways, having enough knowledge to develop pharmacologic agents to treat specific types of comorbid anxiety depression is still a long way in the future (Boyer 2000).

2.2 *Mental–Substance Comorbidities*

Substance use disorders have strong comorbidities with a number of mental disorders (Kessler et al. 1996). These comorbidities are thought to come about largely because of a combination of indirect effects of mental disorders on substance use disorders (e.g., self-medication) and common causes (e.g., exposure to predisposing environmental experiences). There are also some cases in which substance use disorders bring about mental disorders either directly (e.g., brain kindling caused by heavy use of psychostimulants leading to panic) or indirectly (e.g., alcoholism leading to adverse environmental consequences that predispose to depression). However, these processes are

thought to be less common than those involving effects of mental disorders on subsequent substance use disorders.

The patterns and processes involved in mental–substance comorbidity differ by gender. Among women, anxiety and depression are the mental disorders most strongly related to substance use disorders (Kessler et al. 1996). These mental disorders typically occur prior to the onset of substance problems among women. Self-medication is thought to be the main process leading to the secondary substance problems. Among men, in comparison, acting out disorders such as oppositional-defiant disorder, conduct disorder, and pathological gambling are the mental disorders most strongly related to substance use disorders. Both sets of disorders are generally thought to be joint consequences of other influences involving such things as impulsivity, risk taking predisposition, and peer influences.

There is a good deal of evidence showing that patients with certain combinations of substance use disorders and comorbid mental disorders are more difficult to treat and have a more severe and chronic course than patients with pure disorders. This is true, in part, because continued use of substances can interfere with psychotherapy for mental disorders as well as because the existence of abuse potential limits the range of pharmacologic therapies available to the clinician to treat the mental disorder. In addition, treatment compliance is often lower and functional impairment is often higher among patients with mental–substance comorbidities than patients with pure disorders. These factors all predispose to a poor prognosis.

There is considerable interest among treatment researchers in using information about these processes to develop targeted interventions for patients with particular types of mental–substance comorbidity (Carroll et al. 1997). These processes are often quite complex. In the case of comorbid substance dependence and antisocial personality disorder, for example, there appear to be reciprocal effects in which antisocial behavior adversely affects the course of substance dependence as well as substance dependence adversely affecting the course of antisocial personality disorder. In addition, people with this type of comorbidity report earlier ages of first use of drugs, more rapid progression from initial drug use to dependence, and greater drug-related problems in daily living than other people with comparable drug problems. This type of comorbidity is also associated with high rates of relapse after treatment.

There has been some success at developing targeted treatment strategies linked to comorbidity. Relapse prevention programs, for example, have had success treating patients with primary anxiety disorders who use drugs to self-medicate their fears by helping these patients recognize that their relapses of substance problems are often triggered by stress reactions. Refinements of this approach have shown that drug abusers who use drugs as a form of self-medication of a mental disorder respond differently to treatment depending on the type of primary mental disorder they have. Large-scale mental–substance patient–program matching efforts that go beyond these cases, though, have had only mixed success so far (Project MATCH Research Group 1998).

Another treatment matching issue of considerable policy importance concerns the distinction between integrated treatment and parallel–sequential treatment. The evidence is clear that both mental and substance use disorders need to be treated in order to successfully reduce the persistence of mental–substance comorbidity. This being the case, integration of the treatment of the two types of disorders in a single setting makes sense both in terms of benefit to the patient and cost-effectiveness. However, integrated treatment remains the exception rather than the rule in the US today due to disciplinary controversies among the professionals who provide treatment for substance disorders and mental disorders. Differences in philosophy of treatment and concerns about control of resources have kept the two systems apart to the detriment of patients.

This is especially true in the public sector, where most states have separate departments of mental health and substance abuse services and where federal block grants to fund the two types of services are often given to separate organizations that refuse to collaborate. This separation has led in many states to a system characterized by parallel and sequential treatment in which the mental health treatment provider refuses to see the patient unless they are free of drugs and the substance abuse counselor ignores mental health treatment issues. Innovative strategies have been developed both to increase integration of these systems and to help patients negotiate these systems with the aid of a single case manager (Miller 1994). These strategies can be effective both in improving the quality of life of patients and in saving money by reducing inefficient duplication of effort across uncoordinated systems of care. However, entrenched systems move slowly and formidable barriers exist to system change.

2.3 Mental–Physical Comorbidities

It has long been known that high utilizers of primary medical care have high rates of anxiety and mood disorders (Katon et al. 1990). Although some of these patients suffer from physical disorders that indirectly predispose to their comorbid mental disorders, others have untreated primary mental disorders that lead to high medical care costs due to unnecessary medical tests and visits (Miranda et al. 1994). Primary care doctors who are aggressive in detecting and treating mental disorders among their patients generate signifi-

cantly lower overall per panel member health care expenditures than other physicians with comparable case mixes. Furthermore, treatment effectiveness trials designed to help primary care doctors improve their detection and treatment of mental disorders have documented significant reductions both in somatic complaints and in overall treatment costs for somatic complaints (Campbell et al. 2000).

Epidemiological studies have also documented significant comorbidities between particular mental–physical disorder pairs such as panic disorder with asthma and depression with diabetes (Ciechanowski et al. 2000). Prospective epidemiological studies of predisposing personality characteristics provide some evidence that early-onset mental disorders might be risk factors for some chronic physical disorders such as cancer and coronary heart disease. In other cases, such as comorbidity between asthma and panic disorder, the evidence suggests that the physical illness is causally primary.

The evidence is even stronger in showing that comorbid mental disorders, whether causally primary or secondary, adversely affect the course of chronic physical disorders. A variety of reasons have been proposed for this, including effects of mental illness on sensitivity to somatic symptoms, lack of adherence to medical regimens, reduction in immune competence, and increase in sympathetic nervous system reactivity (Rozanski et al. 1999). This evidence is strong enough in the case of some comorbidities to justify experimental trials aimed at evaluating the impact of treating comorbid mental disorders on the course of chronic conditions. For example, the government is currently sponsoring the Recovery in Coronary Heart Disease trial, in which heart patients with comorbid depression are being randomized either to usual care or to an enhanced program of care that includes cognitive-behavioral therapy for their depression.

3. *Implications of Comorbidity for Prevention of Secondary Disorders*

Many strong comorbidities begin with early-onset mental disorders that occur during the school years. One important implication of this fact that has not yet been exploited involves the possibility that temporally secondary disorders could be prevented by aggressive outreach and treatment of these prior mental disorders in school-based screening and intervention programs. This possibility would appear to be especially promising for comorbidities involving early-onset anxiety and mood disorders that predispose to secondary substance use disorders through self-medication. Epidemiological studies show that the time lag between the first onset of these primary mental disorders and subsequent substance abuse averages between six and 10 years (Kessler et al. 1996). This interval of time represents the window of opportunity during which detection and treatment of mental disorders can be used to help prevent or delay the onset of substance use disorders. Little is known about the likely success of these early treatment efforts due to the fact that most people who obtain treatment for anxiety or mood disorders do not receive early intervention. Efforts are currently underway to refine understanding of the transition from primary anxiety and mood disorders to secondary substance use disorders in epidemiological data and to carry out evaluations of early treatment programs in order to determine whether early outreach and treatment of these mental disorders can help prevent later substance abuse. Long-term follow-up of such programs might also provide important insights into the effects of such interventions on the later development of chronic physical disorders.

4. *Future Directions*

As knowledge about the neurobiology of mental disorders increases, some of the syndrome clusters presently considered mental–mental comorbidities will likely turn out to be single disorders. As this happens, a probable concurrent realization will be that some of the syndromes now considered single disorders are in fact heterogeneous. For example, several different types of depression might be identified, each characterized not only by a somewhat different set of depressive symptoms, but also by a characteristic profile of comorbid anxiety symptoms. The ultimate goal of this subtyping will be to refine treatment. As a result, advances in treatment, especially pharmacologic treatment, will spur on these refinements. Indeed, complex treatment trials evaluating the effects of patient–treatment matching will likely be the core research design used to validate subtypes. Initial trials of this sort already have shown the heterogeneity in newer pharmacologic treatments by their ability to treat various comorbid depression and anxiety disorders. Epidemiological twin and family aggregation studies will be useful to provide hypotheses about potentially important subtypes. Similar refinements will occur in studies of mental–substance comorbidities.

Research on mental–physical comorbidities is likely to proceed in a somewhat different way. Although research eventually may show that there are some common neurobiological pathways that predispose to certain closely related comorbidities of this sort, such as early-onset anxiety and later ulcers, most research on mental–physical comorbidities will probably continue to focus on mental disorders as complicating factors in the course of physical disorders. Chronic disease doctors increasingly are recognizing that anxiety and mood disorders are highly prevalent among their patients and strongly related to illness

course. The next logical stage is the development of programs for adjunctive treatment of mental disorders among the physically ill. A great increase in such programs in the future is anticipated. More widespread screening for mental disorders in primary care will likely occur in conjunction with these programs. Closer integration of mental health with physical medicine is an inevitable consequence, although the structure of this integration doubtlessly will be shaped by future developments in the organization and financing of services.

See also: Aging and Health in Old Age; Chronic Illness, Psychosocial Coping with; Health Outcomes, Assessment of; Health Surveys; Mental Illness, Epidemiology of; Mental Illness, Etiology of; Mental Illness, Genetics of; Stress and Health Research

Bibliography

Boyer P 2000 Do anxiety and depression have a common pathophysiological mechanism? *Acta Psychologica Scandinavica* **102**(suppl 406): 24–9

Campbell T L, Franks P, Fiscella K, McDaniel S H, Zwanziger J, Money C, Sobrero M 2000 Do physicians who diagnose more mental health disorders generate lower health care costs? *Journal of Family Practioners* **49**: 305–10

Carroll K M, Nich C, Rounsaville B J 1997 Variability in treatment-seeking cocaine abusers: Implications for clinical pharmacotherapy trials. *NIDA Research Monographs* **175**: 137–57

Ciechanowski P S, Katon W J, Russo J E 2000 Depression and diabetes. *Archives of Internal Medicine* **160**: 3278–85

Feinstein A R 1970 Pre-therapeutic classification of co-morbidity in chronic disease. *Journal of Chronic Disease* **23**: 455–68

Katon W, Von Korff M, Lin E, Lipscomb P, Russo J, Wagner E, Polk E 1990 Distressed high utilizers of medical care: DSM-III-R diagnoses and treatment needs. *General Hospital Psychiatry* **12**: 355–62

Kendler K S, Walters E E, Neale M C, Kessler R C, Heath A C, Eaves L J 1995 The structure of the genetic and environmental risk factors for six major psychiatric disorders in women: Phobia, generalized anxiety disorder, panic disorder, bulimia, major depression, and alcoholism. *Archives of Psychiatry* **52**: 374–83

Kessler R C 1997 The prevalence of psychiatric comorbidity. In: Wetzler S, Sanderson W C (eds.) *Treatment Strategies for Patients with Psychiatric Comorbidity*. Wiley, New York

Kessler R C, Nelson C B, McGonagle K A, Edlund M J, Frank R G, Leaf P J 1996 The epidemiology of co-occurring addictive and mental disorders: Implications for prevention and service utilization. *American Journal of Orthopsychiatry* **66**: 17–31

Miller N S (ed.) 1994 *Treating Coexisting Psychiatric and Addictive Disorders: A Practical Guide*. Hazelden Educational Materials, Center City, MN

Miranda J, Hohmann A A, Attkisson C C, Larson D B 1994 *Mental Disorders in Primary Care*. Jossey-Bass, San Francisco, CA

Project MATCH Research Group 1998 Matching alcoholism treatments to client heterogeneity: Treatment main effects and matching effects on drinking during treatment. *Journal of Studies on Alcohol* **59**: 631–9

Rozanski A, Blumenthal J A, Kaplan J 1999 Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. *Circulation* **99**: 192–217

R. C. Kessler

Comparable Worth in Gender Studies

Until recently, it was accepted practice to pay men more than women even for the same work. Men's wages reflected not only what their jobs were worth but also their status as breadwinners for their families. A woman's wage, when she worked for pay, was regarded as a supplement to those earned by the man she depended on and not the basis of her economic autonomy (Kessler-Harris 1990). In the 1940s, several US states enacted equal pay laws to ensure that men and women who performed the same jobs would receive the same pay. In 1963, the US Congress amended the Fair Labor Standards Act to include the right for equal pay for equal or essentially similar work. These laws proved severely limited in scope. Rarely did men and women do the same or essentially the same work. Occupational segregation is pervasive. And, it is a major factor accounting for the gender-based wage gap.

The National Research Council has concluded, 'not only do women do different work than men, but the work women do is paid less, and the more an occupation is dominated by women, the less it pays' (Treiman and Hartmann 1981). Using national US Census data, they found that each additional percentage point female in an occupation is associated with a $42 penalty in median annual earnings. Paula England's (1992) summary of many studies concludes that as much as 30 to 37 percent of the wage gap is a result of gender segregation at work combined with the tendency to pay women's jobs less, net of their requirements for education, skill, and difficult working conditions. According to a recent study done by the Institute for Women's Policy Research, a typical 29-year-old woman with a college degree stands to lose approximately $990,000 in wages over her working life as a result of the wage gap. Comparable worth, also known as pay equity, emerged in the US in the late 1970s and 1980s to correct for the gender-based gap in wages that is a by-product of occupational segregation.

Comparable worth broadens the scope of equal pay for equal work legislation. It requires that dissimilar

jobs of equivalent worth to the employer be paid the same wages. Comparable worth examines the extent to which work done primarily by women is systematically undervalued because that work has been and continues to be done primarily by women. Systematic undervaluation means that the wages paid to those who perform historically female work are lower than they would be if these jobs had been and were being performed by white males. Comparable worth involves correcting the widespread practice of paying women less than white men for work that requires equivalent skills, effort, responsibilities, and working conditions (Steinberg 1984). It seeks to remove 'femaleness' as a factor depressing the wage rate for historically female work. It attempts to base wages solely on the productive contribution of the work performed to the overall objectives of a work organization, removing the effect on wages of the sex of the typical incumbent.

1. Historical Background

Assumptions about gender that pervade the structure of compensation have their origins in nineteenth-century assumptions about the gender division of labor and the cult of domesticity. (For a fuller document see Steinberg and Figart 2000.) Specifically, by the end of the nineteenth century, occupational segregation had intensified, allowing for the development of a two-tiered wage structure by the gender of the typical job incumbent. Not only was the two-tiered wage structure taken for granted, it was also a legal workplace practice.

Several decades later, these breadwinner–homemaker ideologies were institutionalized into the wage structure through job evaluation procedures. Job evaluation systematically orders jobs as more or less complex and more or less valuable to the organization's mission for the purpose of paying wages. Jobs are described and assessed in terms of relevant job content characteristics. Wage rates are based on these assessments of job content. Individuals performing the jobs are assumed to have the required knowledge, skills, and character to perform the job competently.

Job evaluation systems can be traced back over 100 years, although their use in the private sector became widespread in the 1950s (Treiman 1979). Job evaluation systems in use today have their roots in systems first developed in the 1940s and 1950s (Steinberg 1992). Early evaluation systems and even most current systems use existing wage rates for jobs as the basis for determining what types of job content were complex. For example, if responsibility for supervising a work unit was found to be positively correlated to wages, then supervising a work unit was treated as complex work. At the other extreme, if a low-paying job involved working with the public, then working with the public was defined as simple work. Once a hierarchy of complexity had been established based on existing wages, all references to the derivation of the hierarchy were dropped (Schwab 1980). The fact that job content measures were a proxy for existing wages was lost.

Contemporary job evaluation systems appear to be scientific and objective in their assessment of job content, but actually have built into them gendered cultural assumptions from the 1940s and 1950s including the two-tiered wage structure. Modifications of these systems since the 1950s have been largely cosmetic. Kim (1989) found, for example, that 1986 salaries in California State public employment were affected by explicit assumptions about gender embedded in a compensation system established in 1931. These assumptions pervade the wage structure because over two-thirds of all employers use some form of job evaluation as the basis for their compensation policies.

Comparable worth first emerged as a policy objective during World War II, when the War Labor Board required that employers abide by equal pay for equal worth (Milkman 1987). This policy was honored in the breach. Several states did, however, enact equal pay for equal work laws, primarily in response to union pressure to protect male wages from being lowered when women replaced them during the War. Early drafts of the federal equal pay legislation, endorsed by the Kennedy administration, included language that required equal pay for 'work of comparable character on jobs the performance of which requires comparable skills. Final legislation substituted the word 'equal' for 'comparable' (Steinberg 1984, Hutner 1986).

The contemporary comparable worth movement in the US crystallized in the late 1970s. In 1974, the State of Washington conducted the first comparable pay study. In 1979, the newly constituted National Committee on Pay Equity (NCPE) convened a national conference. In 1981, the Supreme Court of Washington State in *County of Washington v. Gunther* found illegal differential pay between women prison guards and male prison guards, based on the content of their work. Also in 1981, the NAS report, *Women, Work and Wages: Equal Pay for Work of Equal Value*, was published (Treiman and Hartmann 1981). It found strong evidence for gender-based discrimination and legitimated the claims of comparable worth advocates.

The pace of activity on comparable worth accelerated in the US in the 1980s. By the end of that decade, all but five states had at least investigated gender differentials in their civil service pay scales. Over 50 municipalities, 25 counties, 60 school districts, and nearly 250 colleges and universities had investigated comparable worth inequities. Many enacted wage adjustments for historically female occupations.

In 1978, the federal government of Canada enacted the Canadian Human Rights Act, which established equal pay for worth of equal value. By 1990, 9 of the 12 Canadian provinces had adopted pay equity legis-

lation. The 1987 Ontario Pay Equity Act, the most far reaching pay equity legislation to date, requires that comparable worth be proactively established in all private firms with 10 or more incumbents, as well as in the public sector. It requires as well that job evaluation systems be gender-neutral in their assessment of work complexity (Cuneo 1990).

The pace of initial comparable worth efforts began to slow considerably in the late 1980s in the US and in the mid-1990s in Canada. In the US, this was due to a disappointing turnaround in legal decisions, to President Reagan's landslide victory in 1984, to President Clinton's reluctance to support labor policies that intervene directly in the labor contract, to the lessening fear among politicians at the state level of the gender gap at the polls, to the decline in membership and power of organized labor, and to the successful containment of reform implementation by employers (Steinberg 1991). Nonetheless, according to an IWPR study, in the US, $527 million in pay equity adjustments have been disbursed by 20 state governments.

2. *Job Evaluation and Gender Bias*

Job evaluation is the institutional mechanism that perpetuates wage discrimination, especially in medium-sized and large organizations. Approximately two-thirds of all employers use some form of job evaluation. A minority of work organizations set wages without using these formal processes. They rely instead on local market wages, employer biases, and institutional inertia in setting wages. However, if large employers modified job evaluation practices to pay women fairly, smaller employers who do not use such procedures would be affected by these outcomes. There are many aspects of gender bias that remain in virtually every traditional job evaluation system currently available to employers (Steinberg and Haignere 1987, Acker 1989, Steinberg 1990). Four of the most common forms of bias are discussed below.

First, the prerequisites, tasks, and work context of jobs historically performed by women have been ignored or taken-for-granted. Working with mentally ill or dying patients or reporting to multiple supervisors are not considered to be stressful working conditions. Working with noisy machinery, however, is considered stressful. University secretaries who protect the confidentiality of the records they work with do not receive compensation for that responsibility. When examining the skills associated with clerical work, most job evaluation system's treat as invisible requirements for knowledge of spelling and grammar, ability to compose straightforward correspondence, knowledge of the substantive work of the office, and knowledge of the organizational shortcuts within a bureaucracy. The ability of a public health nurse to break down complex and technical material to present to nontechnical audiences is also not acknowledged as a communication skill in traditional job evaluation. If job content remains unacknowledged in the process of describing and evaluating jobs, it is not taken into account when determining a wage rate.

Second, evaluation systems confuse the content of the job with stereotypic ideas about the characteristics of the typical jobholder. Since authority is associated with masculinity, managers are perceived as running offices and departments. The work of the secretary in the actual daily running of an office remains invisible, especially if she performs her work competently. The technical skills and life-and-death responsibilities required of a Registered Nurse are often subordinated to her perceived responsibility for the emotional comfort of her patients. Social psychological experiments confirm that the value of an activity is lowered simply by its association with women (Shepala and Viviano 1984).

Third, the content of work historically performed by women is recognized in the evaluation hierarchy of complexity, but, by definition, is assumed to be less complex than the content of work historically performed by men. For example, both women's and men's jobs require perceptual skills. Male jobs are more likely to require spatial perceptual skills, female jobs visual skills (England 1992). In traditional job evaluation systems spatial skills are treated as more complex than visual skills. Similarly, women's and men's work both require human relations skills, but the human relations skills associated with men's work often involve power and control, while the human relations skills in women's work often involve taking care of others. Job evaluation systems define power and control over others as involving more complex skills than emotional labor. Virtually every off-the-shelf system of job evaluation defines skill, effort, responsibility, and working conditions in way's that treat the job content of male work as more complex than that of female work.

Fourth, some job evaluation systems negatively value job content associated with female jobs in a way that actually lowers wages. One study of a major university found that staff whose jobs involved working with students on a regular basis actually 'lost' pay for that aspect of their job net of other aspects of their job content. Another study found that working with difficult clients and dying patients actually lowered pay independent of other job content. The ironic logic of this evaluation outcome is that the more an incumbent is required to perform the undervalued content, the less the incumbent earns.

3. *Gender Neutral Job Evaluation for Comparable Worth*

To achieve comparable worth, it is necessary to cleanse traditional compensation practices of those gender biases of which we are aware. Achieving this goal

requires recognition that systems of job evaluation have been socially constructed. Thus, they can be *socially reconstructed* to achieve gender neutrality. Early comparable worth initiatives attempted to modify traditional job evaluation systems. In Oregon, for example, a major compensation consulting firm, Hay Associates, agreed in principle to modify its own job evaluation system (Acker 1989). Yet, in practice, the consultants fought against all but minor changes proposed by the Task Force directing the project. When Hay modified its system, the changes had little bearing on the evaluations of the actual jobs being assessed. Some of the lower paid historically female jobs received wage adjustments, but the original hierarchy of job complexity was left intact. More recently, technical pay equity proponents have begun to design gender-neutral evaluation systems to measure job content. The new systems measure more accurately and positively value the invisible skills associated with historically female jobs. One way in which gender neutral job evaluation minimizes gender bias is by developing new dimensions of job complexity or new job factors to capture work differentially found in historically female work. An example is the construction of a factor for Emotional Effort. This factor measures the intensity of effort required to deal directly with the needs of patients, customers, clients, citizens, and coworkers in assisting, instructing, caring for, or comforting them. Hierarchies of complexity are built: dealing with clients who are unpredictably hostile or confused, or discussing death with the terminally ill and their families, is assessed as requiring more emotional effort than dealing with coworkers who are seeking information. The emotional effort of both female and male jobs is assessed in relation to a consistent metric.

A second way in which gender-neutral job evaluation reforms traditional job evaluation is by including and revaluing previously unacknowledged or undervalued characteristics of historically female work in measures of job content already included. For example, in a gender-neutral system, the measurement of Human Relations Skills would not only encompass the skills necessary to supervise subordinates and the skills necessary to speak with top-level executives of organizations. It would also include the skills required to deal effectively with or care for others or to shape, affect, or influence others' feelings, actions, or decisions. Similarly, a job factor on Communication Skills would not only recognize and positively value writing complexity, it would also recognize that presenting technical information to an audience where the level of understanding varies involves more complex communications skills than presenting technical information to a technical audience (Steinberg 1999).

Gender neutral job evaluation is almost never used in pay equity initiatives in the US and only rarely used in Canada. It remains a technical solution in search of a political base with sufficient power to implement it.

4. *Containment of Pay Equity*

Comparable worth advocates have not fared well in their efforts to reform the systems of job evaluation that are used in pay equity initiatives. Public sector employers that enacted pay equity early, such as Minnesota, Washington State, and San Jose, California, did so using unmodified traditional job evaluation. While political expediency in these jurisdictions did result in raising wages for those who work in historically female jobs, it did not remove most of the sources of gender bias from compensation systems. During the second phase of pay equity initiatives, advocates were well aware of the biases of traditional job evaluation, but political resistance to the modification of traditional job evaluation was formidable. Pay equity was not openly opposed. Instead, it was effectively contained, as illustrated by the Oregon example. Most strategies of containment involved manipulation of the process and results of traditional job evaluation. Employers gave the appearance of full cooperation with advocates, but structured decision making around the study in a way that would minimize advocate input and thus minimize both the economic costs of the reform and the changes in compensation practices (Steinberg 1991).

How did these containment tactics work? One strategy involved turning political decisions, such as what constitutes a female-dominated job, into technical decisions and making them nonnegotiable. How much should educational job requirements be valued relative to supervisory responsibilities? Determining the relative importance to overall job value of different dimensions of job content is one of the most important decisions involved in a job evaluation exercise. It is a political decision requiring debate and consensus. It is typically obscured, however, as a technical decision. A second strategy involves withholding information from political advocates until it is too late for them to do anything about it. In Oregon, by the time the pay equity advocates on the Task Force understood the implications of the actions of the consultants, their choice was to abandon the project entirely or accept the flawed results. They accepted the results.

A third strategy, perhaps the one most frequently used by employers, involves placing proponents in a minority position on a Task Force or dividing them among themselves. Most often, trade union advocates are initially placed in a privileged position and representatives of women's groups are marginalized. When they must unite in opposing employer efforts to undercut the reform, their distrust of each other makes common action difficult, if not impossible.

With the availability of Gender Neutral Job Evaluation, achieving pay equity becomes less a technical than a political challenge. But it is a formidable political challenge. Comparable worth is a political goal well worth fighting for. Cleansing compensation systems of gender bias would put an extra $1,500–

5,000 per year in the paychecks of those performing historically female work. Even flawed studies with gender biased evaluation systems have resulted in important wage increases for hundreds of thousands of women. For many women, it represents the difference between poverty and economic autonomy. It would empower women as decision makers in their families. It would make visible and positively reward the productive contribution of women's labor market work.

See also: Economic Development and Women; Economic Sociology; Education and Income Distribution; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Sociology; Gender, Economics of; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Sex Differences in Pay; Sex Segregation at Work; Sexual Harassment: Legal Perspectives

Bibliography

Acker J 1989 *Doing Comparable Worth: Gender, Class, and Pay Equity*. Temple University Press, Philadelphia, PA

Cuneo C J 1990 *Pay Equity: The Labour-Feminist Challenge*. Oxford University Press, Oxford, UK

England P 1992 *Comparable Worth: Theories and Evidence*. Aldine de Gruyter, New York

Hutner F C 1986 *Equal Pay for Comparable Worth*. Praeger, New York

Kessler-Harris A 1990 *A Woman's Wage Historical Meaning and Social Consequences*. University Press of Kentucky, Lexington, KY

Kim M 1989 Gender bias in compensation structures: A case study of the historical basis and persistence. *Journal of Social Issues* **45**: 39–50

Milkman R 1987 *Gender at Work: The Dynamics of Job Segregation by Sex During World War II*. University of Illinois Press, Urbana, IL

Schwab D P 1980 Job evaluation and pay setting: Concepts and practices. In: Livernash E R (ed.) *Comparable Worth: Issues and Alternatives*. Equal Employment Advisory Council, Washington, DC

Shepala and Viviano 1984 In: Remick H (ed.) *Comparable Worth and Wage Discrimination: Technical Possibilities and Political Realities*. Temple University Press, Philadelphia, PA

Steinberg R J 1984 'A want of harmony': perspectives on wage discrimination and comparable worth. In: Remick H (ed.) *Comparable Worth and Wage Discrimination: Technical Possibilities and Political Realities*. Temple University Press, Philadelphia, PA

Steinberg R 1990 The social construction of skill: Gender, power and comparable worth. *Work and Occupations* **17**: 449–82

Steinberg R 1991 Job evaluation and managerial control: the politics of technique and the techniques of politics. In: Fudge J, McDermott P (eds.) *Pay Equity: Theory and Practice*. University of Toronto Press, Toronto, Canada

Steinberg R 1992 Gendered instructions: Cultural lag and gender bias in the hay system of job evaluation. *Work and Occupations* **19**: 387–423

Steinberg R J 1999 Emotional labor in job evaluation: Redesigning compensation practices. *Annals American Academy of Political and Social Sciences* **561**: 143–57

Steinberg R, Figart D 2000 Pay Equity: The Policy and the Movement in North America. unpublished

Steinberg R J, Haignere L 1987 Equitable compensation: Operational criteria for comparable worth. In: Bose C, Spitze G (eds.) *Ingredients in Women's Employment Policies*. SUNY Press, Albany, NY

Treiman D J 1979 *Job Evaluation: An Analytic Review*. National Academy of Sciences, Washington, DC

Treiman D J, Hartmann H I (eds.) 1981 *Women, Work, and Wages: Equal Pay for Jobs of Equal Value*. National Academy Press, Washington, DC

R. J. Steinberg

Comparative History

Comparative history differs from other historical methods in that it takes an explicit line of questioning to compare two or more cases stemming from different contexts. The aim of this operation is either to bring out the similarities and differences of the different cases, or to determine the scope of social scientific theories or theoretical approaches. Seldom does comparison itself lead to the formulation of theories. These characteristics distinguish historical comparison monographic work concentrating on case studies, from historical syntheses with an international scope and from implicit comparison, which indeed belongs among the analytical instruments of all historical work.

1. The Story of 'Comparative History'

Although the comparative perspective has appeared repeatedly in the history of historiography, it has only presented itself as a methodical instrument of explicit theoretical comparison since the 1930s. Of course, the historiography of the Enlightenment made use of comparison in its endeavor to grasp the specific characteristics of the historical process by means of universal historical types; in this context, Montesquieu, Adam Smith, and the Cameralist J. H. G. Justi should be mentioned. In Scottish, English, French, and German Enlightenment literature as well, cultural areas and development stages were compared. Thus, just as the demarcation of the European model against non-European civilizations in the eighteenth century belonged to the arsenal of universal history, so did the belief that developed societies can trace their own history in that of less-

developed societies (e.g., Karl Marx). In this attempt, development thinking became closely intertwined with a Eurocentric perspective. This universal historical tradition persisted, carried on in the twentieth century by such diverse authors as Oswald Spengler and Arnold Toynbee, Karl Polanyi, and Schlomo Eisenstadt.

Modern comparative history, however, is not rooted in this tradition. Profoundly affected by the experience of World War I and the traumas suffered by parts of academia under nationalistic exclusionism, comparative history established its argument for comparison as a fruitful enterprise particularly based on the results produced in neighboring sciences. Henri Pirenne, a Belgian historian and one of the pioneers of internationally comparative historiography, did admittedly make reference to universal history in his opening speech at the fifth international conference of historians in Brussels in 1923, but with the intention of combating the narrowness of purely national perspectives and the subjection of historical research to functional ends. However, Marc Bloch and Otto Hinzte, two of the most influential forerunners in the field of comparative study, were even more active in taking up approaches from neighboring disciplines where comparison had long been a commonly practiced method. Historiography, as a historically comparative discipline, was a latecomer compared with other sciences. In fact, as early as the late eighteenth and early nineteenth centuries, various sciences had practiced comparative study successfully. Such works were produced not only in anatomy and physiognomy, but also in religious studies, jurisprudence, geography, and philosophy. Among them was that of F. Max Müller, the first *Introduction to the Science of Religion*, published in 1873. Thus it was no coincidence that Marc Bloch, in his 1928 essay advocating the practice of comparative history, does not cite historians, but rather Antoine Meillet and the broad range of scholars who had been carrying out internationally comparative linguistic research since as early as the nineteenth century.

In the development of modern comparative history, contact with the social sciences played the central defining role. Marc Bloch, whose works on the 'royal touch,' on the agrarian history of England and France, and on the different forms of feudal society in Europe up to 1940 provided important examples of comparative study, had received theoretical stimuli, above all from Emile Durkheim. Durkheim saw comparison as the sociological method *par excellence*, enabling the delineation of the fundamental characteristics of social types. Bloch used this method, for example, in his analysis of feudal society when he concentrated on its dominant dependency relationships. Otto Hintze, who was among the pioneers of historical comparison, also took the work done in the social sciences as his model. He adopted the concept of the ideal type from Max Weber and went on to construct an ideal-type concept of feudalism. He then applied this to the various manifestations of feudal domination, which he found to be particularly pronounced in Russia, Japan, and the Islamic countries.

In the period following World War II in particular, the number of theoretically based works with an internationally comparative perspective increased. The science of history bore a complex interrelationship to historical sociology. On the one hand, it profited from the broad scope of sociology's macro comparisons, which were limited neither to a specific epoch nor to a specific continent. These comparisons studied different national cases, using stringent argumentation to examine theoretical questions. Thus Barrington Moore inquired as to the role of agrarian development in the emergence of either democracy or dictatorial fascistic regimes in a broad comparison of six national cases, while Jack Goldstone compared revolutionary crises of the early modern age in Europe and Asia. On the other hand, however, it was only the historical research in monographic form that made available to historical sociologists the theses and research results that would enable them to carry out extensive comparisons. The stimuli obtained from historical sociology were also accompanied by critique of its generalizations and its handling of the historical state of research, which it did not always improve upon extensively. Nevertheless, historical sociology provided an important stimulus to comparative history. Its works elicited an increasingly favorable response in the 1960s to the extent that history was open to theoretical questions, and gradually after 1970, these works were also emulated.

The theory of modernization, which originated mainly in the United States in the 1950s, provided a powerful comparative impetus as well, and elaborated parameters of development against which it measured individual societies. The successful combination of market economy and political democracy was its criterion for modernity. In this sense, the vague but positively cast concept of 'modernity' was used widely in historiography on the nineteenth and twentieth centuries; it emphasized structures, functions, and interdependencies among the different parts of systems more strongly than their historical individuality. In individual historiographies—particularly in those countries catching up in industrial and political development—the comparison to the Anglo-Saxon 'leader' countries soon became one of the standard lines of argumentation. In the West German debate about the long-term and medium-term causes of National Socialism, structural deficits were identified in comparison to Anglo-Saxon development, which itself was admittedly highly typological and idealized: these included the feudalization of the German middle classes, a general deficit in the middle-class way of life, the exceptional proximity of middle-class life to the state, and the persistence of privilege-based social orders in the old and new middle class. Especially in

this debate about Germany's *Sonderweg* ('unique path')—but in Italian historiography as well—it was no longer just the linear and unceasing advance toward modernity that was at issue, but the ruptures, contradictions, and deficits in the development.

Within the science of history, comparative methods by no means penetrated all disciplines to the same degree or simultaneously. They emerged earlier in economic history, population history, and political history than in social history, gender history, or cultural history. As early as 1957, the Economic History Association dedicated one day of its annual conference to comparative economic history, and it was no coincidence that one year later, economic historian Sylvia Trupp founded the journal *Comparative Studies in Society and History* with the expressed goal of publishing studies of an interdisciplinary and comparative nature. In Europe, the English model of industrialization and its spread was the central topic of the comparative studies that appeared even before 1960. The most broad-based and influential comparative analysis was produced by Alexander Gerschenkron, who studied the basic similarities and differences within the European process of industrialization, inquiring into not just the different answers to similar challenges, but also the functional equivalents among European societies. He explains the differences in development on the basis of both the specific position of the individual economy in the system as a whole ('relative backwardness') and the influences generated by developed societies. Walt Rostow's 'stage theory' of economic growth also provided an impetus to comparative economic history after 1960, as did Simon Kuznets's quantitative studies. Rostow's thesis that every society has to go through similar phases of economic growth and that therefore each society's position can be identified against this background was almost a direct invitation for checking by historical comparison. Kuznets, in contrast, instead emphasized the uniformity of modern economic growth, although without ignoring its national diversity. His three-sector thesis also led to comparative studies.

In historical demography, in 1963 at Princeton, Ansley Coale began comparing 600 European regions to answer the question of whether and why the birth rate was declining in Europe; this was called the European Fertility Project. Although the project did produce convincing demographic results, the explanation of demographic behavior on the basis of historical macro processes came under criticism. The reasons for the early comparative studies of political movements and institutions can be found not only in the field of 'comparative politics,' which gained importance in the USA in the 1960s, but also—according to Klaus von Beyme—in a unique approach to research within political science. This approach applies to institutions that can be classified under the general type of democratic systems, but that nevertheless take on specific forms. Research on resistance movements and on social systems or constitutions paved the way for the advancement of comparison in social history .

Nevertheless, since the mid-1960s, sociohistorical comparison has taken on a leading role, showing a remarkable breadth of research and differentiation in approaches. In Europe, at least a handful of papers have been published every year since 1970 on comparative social history, none limited to a specialized area, but including the history of the family and social classes, and the development of the social state, the city and education. These studies have been focused mainly on three central issues. First, they deal with the explanation of national peculiarities, which often—because of political events—are seen as characteristic of the social development of the last few centuries. In the debate about the German *Sonderweg*, this orientation of comparative sociohistorical research has taken its most comprehensive and probably also most fruitful form as a scientific heuristic device. Second, they focus on the different pathways of individual European societies into the modern age: their distinctive features are identified above all in comparisons between different societies, and between European and non-European cultures. A third focal point has been the European model and its social characteristics. All these studies have demonstrated methodical diversity in both quantity and in quality; have integrated social, political, and economic factors; and have focused attention on both linear and cyclical developments. Furthermore, they have compared societies as well as parts of societies.

European historiographies have not all made use of the comparative method to the same degree. Surprisingly, Marc Bloch's appeal in France went unheard for many years, as French historiography was and continues to be strongly oriented toward monographic work and highly skeptical of theory. In Great Britain as well, historical comparison did not catch on because of the dominance of national history and the distance from the social sciences. In Italy, comparative history remained on the periphery—despite the proximity of many historians to debates in the social sciences—because of the strong regional orientation. In West Germany, Scandinavia, Switzerland, and Austria comparative work remained rather limited within total historical production, even though the number of comparative works was significantly higher than the number of other European historiographies. Although sociohistorical comparison was particularly strong in Germany, this was a result of both debate on the theory of the German *Sonderweg*, which was dealt with in many comparative historical essays, and the explicit discussions of social-scientific theories and proposed explanations that characterized modern German social history. In general, one may assume that internationally comparative history had a difficult time gaining a foothold wherever the relation to theory

was weak, and wherever attempts to explain national patterns of development in the philosophy of science remained only peripheral. Before 1989, this was the case in Iberian societies as well as in Eastern Europe.

The study of history in the United States, where, according to Raymond Grew, comparison is a weak point rather than a strong one, has not only brought forth a series of excellent comparative studies on European history, but also practices transepochal and transatlantic comparison more intensively than does European historiography. Among the comparative works on Europe, those of Charles Tilly, Peter Baldwin, and Susan Pederson are particularly worthy of mention. The study by Arno Mayer on the Holocaust, which he sees in the tradition of the Crusades and the Thirty-Years' War, takes a transepochal and internationally comparative approach and presents a comparative tableau of the situation of Jewish people in Eastern European societies in the 1930s. Transatlantic perspectives are opened up by those studies that place slavery in the South, the frontier problem, and the situation of migrants in a larger, multinational context.

In gender history, the comparison of male and female roles was for a long time the constitutive element of analysis, and only recently have theoretical, internationally comparative studies been produced. These have concentrated on institutions or sub-areas which, like social laws, have attained a special meaning for gender roles. Modern cultural history as well, which discusses the creation and relevance of stems of meaning, rituals, and symbols, has concentrated thus far on national configurations and ascribed little importance to international comparison.

In historiography on classical antiquity, the Middle Ages, and the early modern age, comparative studies were carried out, comparing rulers, regimes, and individual institutions. Because of the international character of political systems, the national aspect decreased in importance as a unit of comparison. However, theoretically oriented comparisons remained scarcer in this area than in the historiography on modernity.

2. *Functions of Historical Comparison*

The comparative method pursues two goals alternately. On the one hand, it seeks to accentuate the distinctive feature of each individual case, and on the other, attempts to derive evidence on general developments from case studies. While historians tend toward the first approach, the second is more prevalent among social scientists. Comparative historical studies do, however, also deal with the question of commonalities. Four functions are granted to comparison in these studies: a heuristic function, a contrastive function, an analytical function, and a distancing function.

Comparison plays a heuristic role when it alone can offer explanations and reveal phenomena that had been unknown or inadequately known up to that point. This function was already referred to by Marc Bloch, who, based on his knowledge of the English enclosure movement, sought an equivalent in French agrarian history. Founding his thesis on research in regional history, he dates the disappearance of collective rights in Provence as early as the fifteenth century, and concludes from this that similar movements took place in France not only earlier, but also under other conditions. Thanks to the comparative method, Bloch was able to discover a characteristic of French agrarian development through his studies of the agrarian history of other countries.

Historical comparison can be called contrastive when it serves to define more precisely the special features of a specific case. Above all in German–French comparison, this method was used to bring into bold relief the particularly xenophobic character of German nationalism in the early nineteenth century, the insurance character of Germany's social security system, and its unique characteristic of having a highly educated middle class. Comparison of Italian and Anglo-Saxon developments brought out the regressive state of Italian industrial development, its lack of a modern party system, and its unique feature of having a patrician, noble class. Depending on the country selected for comparison and the logic behind this comparison, specific characteristics of the mostly isolated national case but also of regional patterns of development are illuminated and brought into clear contrast against the totality. At least two problems emerge in this endeavor: the further the development stage in the reference country or region deviates from the individual case in question, the less suitable it is for defining the specifics of the case beyond mere identification of general deficits. If one measures the economic development in Italy in the nineteenth century using as a yardstick the conditions that promoted industrialization in England, one can only determine that these conditions were lacking in Italy, but cannot identify the specific conditions for the economic growth that took place there. The more that the comparative reality is used merely as a foil to highlight the contours of a specific case, the more typological and reductive the resulting picture. International case studies written in comparison to the German case with the aim of presenting a general view of the livelihood and benefits of German white-collar employees have only rarely provided convincing analyses of the individual national developments in the employee milieu. This inherent bias of comparison should be borne in mind in each case.

Comparison has an analytical character when it either tests a scientific hypothesis or identifies constellations of causes in a specific situation. The thesis that a causal relationship exists between capitalism and feudalism is relativized by the fact that strong fascist

movements did not emerge in all capitalist societies; they were able to develop only under specific conditions. Jürgen Kocka qualified this assumption in his comparison of American and German society between the two World Wars. In a study of the French, Russian, and Chinese revolutions, Theda Skocpol inquired as to the results of collapsed administrative apparatuses, broad peasant rebellions, and movements among political elites, thereby demonstrating the extreme diversity in the constellations of causes which explain the outbreak of modern revolutions.

In this analytical approach, comparison can also serve as an indirect experiment. When phenomenon *a* is ascribed to cause *b*, the historian can test their hypothetical connection by looking for constellations in other societies in which *a* appears without *b*, or where *b* exists without leading to *a*. This heuristically useful procedure is distinguished from experiments in the natural sciences in that here, the *ceteris paribus* conditions are seldom given.

Comparison has a distancing effect when it offers another perspective to observation and analysis. It can produce surprising discoveries as well as relativizing the tradition-based context of national historiographies. Especially for those historiographies that are deeply embedded in a national context, comparison can open up new and often broader vistas. When confronting comparisons with other reference cases—especially those from other cultures—not only does one gain experience with different types of question and method, but also fundamental assumptions of one's own historiography can be revealed and their problems expounded. The comparative view can also contribute new insights: Skocpol, for example, uses the comparative method to shed new light on the similarities between the French Revolution of the eighteenth century and the Chinese revolution of the twentieth century in her comparison of revolutions, and between the cities of France and Japan in the eighteenth century. Taking into consideration the variety of alternative pathways into the modern age prevents exclusive concentration on the European development path, and offers insight into those conditions of European development which cannot be generalized.

3. *Methodology of Comparison*

Although comparative history does not possess its own methodology, it does present particular methodological problems in bold relief.

Historians have used the proximity of their work to original sources as a special proof of its scientific nature. The goal of reconstructing history via a multitude of sources in as comprehensive a manner as possible is one of the fraternity of historians' standards. This fundamental methodological principle to a certain degree evades historical comparison. In so far as it compares more than two cases, it depends more heavily on secondary literature than on the evaluation of a consistent body of sources. Thus, it also faces the problem of having to assess adequately the historiographic context, where each individual work of secondary literature interprets the original sources in its own specific way.

Another of the principles of historical study is that individual aspects of reality cannot be understood separately from their place within a totality or within a historical development. The full meaning of details comes into focus only when looked at in synchronic and diachronic perspective. The method of isolating variables—a common one in empirical social research and political economy—can and should be used within the framework of research projects, but is inadequate for representing the results of research. Comparisons can trace diachronic development processes in their similarities or differences, but they can also examine trans-epochal cases along one specific line of questioning. For example, Raymond W. Goldsmith studies the premodern financial systems in ancient Mesopotamia, the Indian Mogul empire, and England under the rule of Elizabeth I. For him, what is of interest is not how an epoch can be characterized, but the contribution of different states and ages to a central line of questioning. Comparison can also evade both chronology and the total context in question: one separates out the individual cases, subjects them to examination from a specific perspective, reduces their complexity, and studies them as exemplary cases belonging to a 'universal,' the *tertium comparationis*.

Because of this isolation and reduction, the methodology of comparison enters into a relationship of unresolvable tension with fundamental principles of historical study. What has proven to be a rule of thumb is that the loss of context in comparative study is less severe the smaller the number of reference cases. By choosing medium-range theories and argumentation on a moderate level of abstraction, one limits as far as possible both the loss of concreteness and the distance from the object of study. To do justice to the diverse kinds of relationship affecting a particular case, comparative historical studies include processes rather than constellations, and link the analysis of individual situations with questions derived from the historical study of relationships. The integration of individual cases into the total social context can be intensified through close attention to the linguistic and conceptual specifics that define and influence these contexts.

Despite all of these methodical problems, historical comparison also has a positive impact on historical research by raising its consciousness about the extent and limits of different approaches, the implicit premises of scientific study, and the coherence of the total context. It opens up historical study to theoretical reflection.

The unit of comparison changes according to the line of questioning and the information being sought. For a long time, a national framework of study was dominant in the science of history, both because this framework influenced the problems being studied, and also because answers to questions of national history were sought in comparison. Research on nationalism and welfare states, democracy and religious denominations have been given a privileged place in the national framework. This national approach has been called into question by two developments: microhistory and intercultural history.

Microhistory has brought increased interest to local, often unique sources and problems which give a privileged role to the viewpoints of the people involved, and which—because of the emphasis on individuality—had been excluded previously from methodical comparison. In addition, this shows the importance of another area which calls into question the self-assuredness of past research on national historical developments. For the study of family constellations and famine unrest, local frameworks present themselves as the most appropriate, while for the analysis of linguistic dialects or industrialization, the regional level is preferable. As the study of history becomes increasingly global, the national level will decrease continuously in significance. This does not just apply to colonialism, the migration movements of the nineteenth and twentieth centuries, or the global economy that has been developing since 1880. Marc Bloch pointed out as early as 1928 that the development of manorial rule in the Middle Ages and early modern age in individual localities and regions cannot be understood independently of the decline of money rent throughout Europe.

In so far, then, as one seeks to identify the special features of Europe's development, it is necessary to include non-European cultures, which, however, should no longer serve merely as a foil as they did for Max Weber, but rather be examined with a differentiated view to their independence and specificity. In order to obtain a new perspective, Jürgen Osterhammel proposes a new approach: instead of analyzing non-European cultures from a European standpoint, one should take the opposite point of view and determine European specifics from the standpoint of non-European cultures. The methodical problems of such a perspective are as obvious as its advantages: these include on the one hand problems of languages, the knowledge of often very different historiographies, and aggregation of data that decrease in meaningfulness and are often statistically antiquated, and on the other, a new view of known developments, and the opening of a wider sphere of activity and relationships.

In particular, the history of transfer processes brought attention to Galton's problem of ethnologies when it questioned whether two or more cases can really exist independently alongside one another, or if they are enmeshed in a visible or invisible web of relationships that influence them decisively. The question of how specific constellations develop through direct influence, appropriation or imitation of foreign models does indeed increase the complexity of comparison, but in no way replaces it. As is well known, Marc Bloch extolled the study of societies in the same epoch which are neighboring, mutually influential, and subjected to the same macro changes as one of the ideal methods of historical comparison.

4. *Comparative History's Agenda and Unsolved Problems*

Comparison poses methodical and theoretical problems for the science of history, but at the same time challenges it, and can even work as a factor to increase method-consciousness and support theoretical reflection. Currently, an expansion of previous approaches and fields of comparison in the following directions could be considered:

Until now, comparative studies have limited themselves to defining the range and meaningfulness of theories. Such problems as whether Weber's class theory is appropriate for the history of the European middle class, whether Pierre Bourdieu's concept of cultural capital is useful for an analysis of intellectuals, or whether Marx's theory can be applied to modern revolutions—and whether all of these can explain more than just national cases—have been the focus of interest thus far. However, the question of what changes must be made in theoretical concepts if they are to explain more than one national development and encompass a multitude of empirical cases has seldom been asked. Therefore, comparative history has contributed little to the development of theory thus far, and has seldom taken the step to formulate theoretical alternatives by relativizing existing approaches.

In historical comparison, progressive types and structures—rather than the constitutive processes of historical facts and materials—were, for a long time, at the center of attention. With the increasing importance of constructivist approaches, individual sources such as reports and statistics are being interpreted to be more a representation of individual administrations or rulers than an account of social or political reality. Comparisons that strengthen this point of view would involve representations and strategies of representation, and prevent the naive use of sources and statistics. In this area, new fields of research are opening up.

In numerous comparative studies, the process of modernization is built in as the *tertium comparationis*. This enables that the trajectory of development of each particular case be identified and defined. The more that temporally staggered comparisons are used

in these studies, however, the less importance they ascribe to synchronic observation and the more they ascribe to the content of the line of questioning. In this way, comparison concentrates less on the characterization of phases of historical development, and more on different contexts—such as collectives' self-descriptions, social placement strategies, ascription of religious meaning—which it analyzes for several collectives.

Contemporary processes of globalization have left their mark above all on the capital and labor markets, but have also called into question the national contexts of thinking prevalent in the sciences. On the one hand, historiography in the European framework reacts to this relativization of the national: it can give a privileged role both to comparison between individual societies as well as to the homogeneity of powerful forces and processes. Transcultural comparison with non-European cultures also presents a promising means of isolating and identifying the factors of Europe's development. On the other hand, the individual societies themselves have become complex 'melting-pots,' as reception and diffusion of cultural models, social modernization, and social and cultural coding have merged in them, accompanied by national-historical and global-social influences. With the increasing demand for an adequate survey of contemporary society, the demands placed on a theoretically based, methodically differentiated comparative history will increase as well, expanding the spectrum of questions with which history will be interpreted.

See also: Civilization, Concept and History of; Civilizational Analysis, History of; Comparative Method, in Evolutionary Studies; Economic History; Global History: Universal and World; Historical Demography; History and the Social Sciences; Marx, Karl (1818–89); Modernization and Modernity in History; Modernization, Sociological Theories of; Polanyi, Karl (1886–1964); Social History; Weber, Max (1864–1920)

Bibliography

Atsma H, Burguière A (eds.) 1990 *Marc Bloch aujourd'hui. Histoire comparé et Sciences sociales*. Paris

Bloch M 1928 Pour une histoire comparée des sociétés européennes. In: Bloch M (ed.) *Mélanges historique*. Paris, I, pp. 16–46

von Beyme K 1988 *Der Vergleich in der Politikwissenschaft*. Munich, Germany

Espagne M 1999 Sur les limites du comparatisme en histoire culturelle. *Genèsis* **17**: 102–21

1988 German history before Hitler: The debate about the German *Sonderweg*. *Journal of Contemporary History* **23**: 3–16

Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Cambridge, MA

Goldstone J A 1991 *Revolution and Rebellion in the Early Modern World*. Berkeley, CA

Haupt H G, Kocka J 1996 *Geschichte und Vergleich. Ansätze und Ergebnisse international vergleichender Geschichtsschreibung*. Frankfurt, Germany

Hintze O 1964 Soziologische und geschichtliche Staatsauffassung. In: Oestreich V G (ed.) *Soziologie und Geschichte—Gesammelte Abhandlungen*. Göttingen, Vol. 2, pp. 239–305

Kaelble H 2000 *Der historische Vergleich*. Frankfurt, Germany

Kocka J 1977 *Angestellte zwischen Faschismus und Demokratie. Zur politischen Sozialgeschichte der Angestellten, USA 1890–1940 im international Vergleich*. Göttingen, Germany

Mayer A 1988 *Why Did the Heavens Not Darken? The 'Final Solution' in History*. New York

Moore B 1966 *Social Origins of Dictatorship and Democracy. Lord and Peasant in the Making of the Modern World*. Boston

Passato e presente, 1993, Vol. 28

Pederson S 1993 *Family, Dependence and the Origins of the Welfare State: Britain and France 1914–1945*. Cambridge, MA

Rostow W W 1960 *The Stages of Economic Growth*. Cambridge

Rossi P (ed.) 1990 *La storia comparata. Approcci e prospettive*. Milan

Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge

Tilly C 1984 *Big Structures, Large Processes, Huge Comparisons*. New York

van den Braembussche A A 1989 Historical explanation and comparative method: Towards a theory of the history of society. *History and Theory* **28**: 2–24

H-G. Haupt

Comparative Method, in Evolutionary Studies

Comparative methods seek evidence for adaptive evolution by investigating how the characteristics of organisms, such as their size, shape, life histories, and behavior, evolve together across species. They are one of evolutionary biology's most enduring approaches for testing hypotheses of adaptation. This article discusses the application and interpretation of comparative studies, reviews their historical development, then describes new methodologies for analysing comparative data.

One of the classic comparative relationships is that between the size of an organism's body and the size of its brain (Fig. 1): brain size increases steadily and predictably with body size in mammals, and this relationship also holds within most other animal groups. Much of the interest that comparative rela-

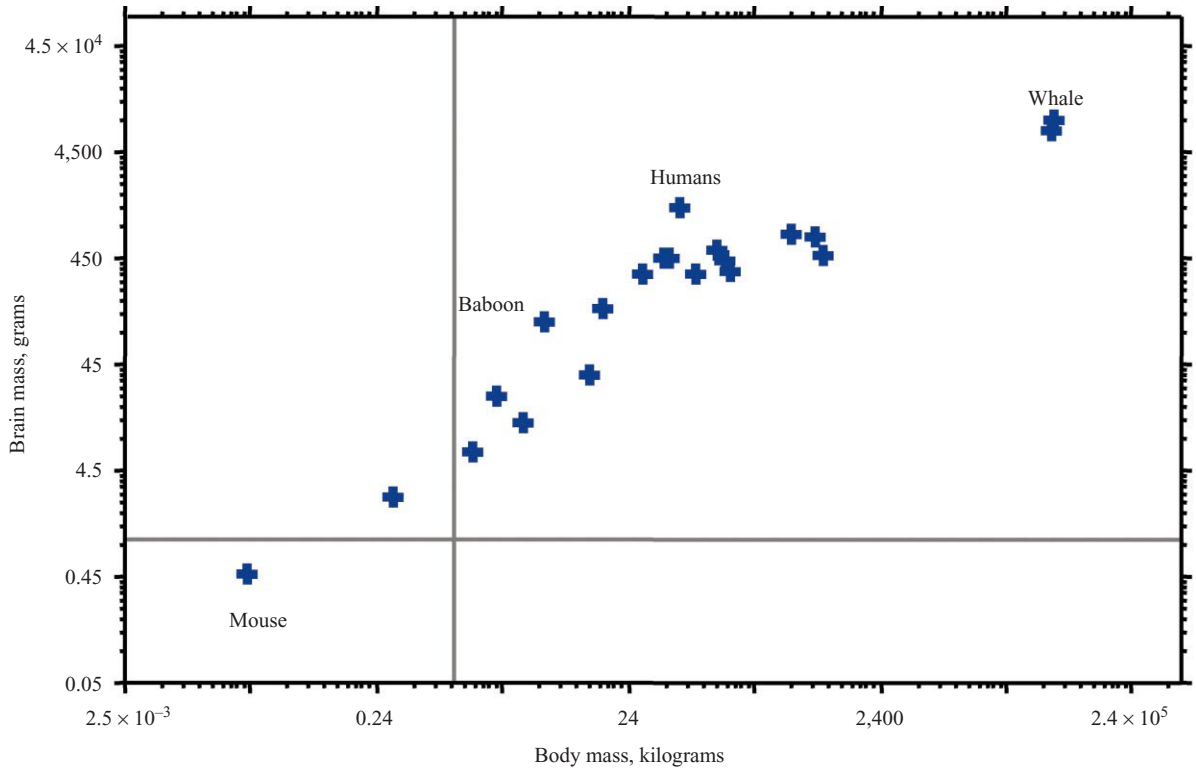

Figure 1
Plot of brain mass versus body mass in mammals. Selected points are labeled. Note the logarithmic scales. The correlation between the two variables is 0.96 and the slope of the line relating brain mass to body mass is 0.61 (see text and Table 1 for further discussion.

tionships such as this one generate derives from their ability to reveal broad or general evolutionary trends. Table 1 lists a small number of such trends and the species in which they have been found.

Comparative trends, in addition to describing the 'end-points' of evolution across contemporary species, can also be shown to have an historical interpretation: the relationship observed across a group of contemporary or extant species reflects the correlation that would have held in the ancestral species as they evolved over time up to the present (Pagel 1993). Thus, the relationships reported in Table 1 tell us that as organisms evolve and grow larger, their lifespan, ages at maturity, brain size, and basal metabolic rates increase, but their heart rates slow down.

The natural interpretation of comparative relationships is that they reveal adaptive trends. This is to say that the organisms that individually evolved to produce the trend were somehow more fit or suited for survival than those that did not. Thus, as animals evolve to become larger, a larger brain is required for 'housekeeping' functions as well as simply to monitor the increased number of neuronal signals coming from the organism's body and internal organs. Larger animals also tend to roam over wider areas and a larger brain may be advantageous for taking advantage of what the terrain has to offer. Small animals could have larger brains, but if the adaptive argument is correct, a relatively large brain in a small animal

Table 1
Selected comparative relationships

Trait	Slope or relationship to body size	Species in which measured
Brain mass	0.67–0.75	Mammals
Age at maturity	0.25	Birds, mammals
Lifespan	0.25	Birds, mammals
Heart rate	−0.25	Birds, mammals
Basal metabolic rate	0.75	Wide range

Note: the slope records how much the 'Trait' variable changes in response to a unit increase in body weight, when both variables are plotted on logarithmically transformed axes

might be a disadvantage. Brains are costly to produce and maintain (they must be supplied with blood and nutrients) and so animals are unlikely to evolve 'spare capacity.'

Some comparative trends reveal a relationship that is mediated by some other set of factors. It is not obvious why a longer lifespan is only advantageous to larger animals. Would not a small organism benefit from living a long time and thereby having more opportunity to reproduce? One possibility is that smaller organisms suffer higher rates of mortality, from natural variation in the environment to higher rates of predation. If this is true, it may not pay smaller organisms to evolve to have a long lifespan. Rather, natural selection may favour adaptations in small organisms that allow them to grow, mature, and reproduce rapidly: there is little point in a 'slow and steady' lifestyle if one is likely to be dead before reaching maturity. Conversely, a slow and steady lifestyle may be advantageous to longer-lived species. By putting resources into producing a large and healthy body, an individual may be able to produce a larger number of offspring over a long period of time.

When such additional factors are thought either to mediate a relationship or to offer an alternative explanation, comparative methods can still be applied. The relationship between two variables can be investigated, having statistically removed the effects of a third. Rates of mortality, for example, remain correlated with lifespan even after controlling for body size: thus, one can say that decreasing mortality rates, for a given body size, will be associated with a longer lifespan. These examples show that comparative studies can both test adaptive hypotheses and reveal trends whose precise causes can often be discovered by further investigation. Alternatively, a comparative trend might be regarded as evidence, not of adaptation, but rather of some form of constraint. Animals with larger bodies may have larger brains because the genes that make larger bodies also make larger brains. Similarly, larger animals may have a longer lifespan simply because it takes longer to grow a larger body than a small one. This tension between adaptive and nonadaptive explanations for observed comparative trends is what drives investigators to make ever more precise tests of their hypotheses. So-called nonadaptive hypotheses can often be tested by controlling for suitable 'third' variables.

1. *The Problem of Independence*

One of the strongest pieces of evidence that a comparative relationship represents an adaptive trend is to show that it has arisen independently on more than one occasion. Were this simply a matter of collecting information on a wide range of species and plotting a trend, the 'comparative method' would be little more than a commonsense set of rules for displaying information. However, the hierarchical nature of biological evolution means that individual species can seldom be considered independent instances of the adaptive relationship.

Over evolutionary time species evolve and give rise to daughter species which in turn give rise to more species. At each branching point in evolution, the ancestral species may live on independently of its daughter species, or it and its daughter species may evolve into something slightly different from the ancestral form. The collection of such ancestral-descendant relationships among a set of species is described by a phylogenetic tree. Phylogenetic trees are characterized by a series of branching points leading from the 'root' or common ancestor of the species up to the tips or contemporary organisms.

The phylogenetic tree diagram also makes clear why species cannot typically be considered independent instances of an adaptive trend. Consider the example of Fig. 2(a). Eight species, all descendant from a common ancestor take the values {A, B} on two traits. Eight other species, all descendant from a different common ancestor, take the values {a, b} on the same traits. The relationship counted across the sixteen species is perfect: 'A' always goes with 'B' and 'a' always goes with 'b.' But the phylogeny of these sixteen species reveals that the pairing {A, B} and {a, b} probably arose only once. For sake of argument, assume that the ancestor at the root of the tree was {A, B} and that the traits 'a' and 'b' evolved along the branch leading to the ancestor to the eight species on the right of the diagram. These traits were then retained in the eight descendants.

The phylogenetic perspective forces us to realize that what might have appeared to be a broad adaptive trend linking two traits could be little more than a chance pairing. For some reason 'a' and 'b' each evolved in the same branch, but quite possibly independently of each other. Their retention in the eight contemporary species is no evidence at all that they coevolve: each may be adaptive in its own right, but their joint presence is not evidence that they evolved together for some adaptive reason.

If the phylogenetic tree looked like that of Fig. 2(b), however, then there would be evidence that the relationship {a, b} had evolved repeatedly: that is, on a number of independent occasions, when 'a' evolved, so did 'b.' The summary of the number of species with traits {A, B} and {a, b} would be identical to that of Fig. 2(a). But, the result in Fig. 2(b) would compel us to seek evidence for a link between 'a' and 'b' in a way that was missing from Fig. 2(a).

These considerations lead to the following important conclusion: the strength of evidence for an adaptive relationship in a comparative study derives, not from knowing how many species come to inherit some set of traits, but in how many times the relationship between two traits has arisen independently. Nearly all of the effort among researchers who design methods

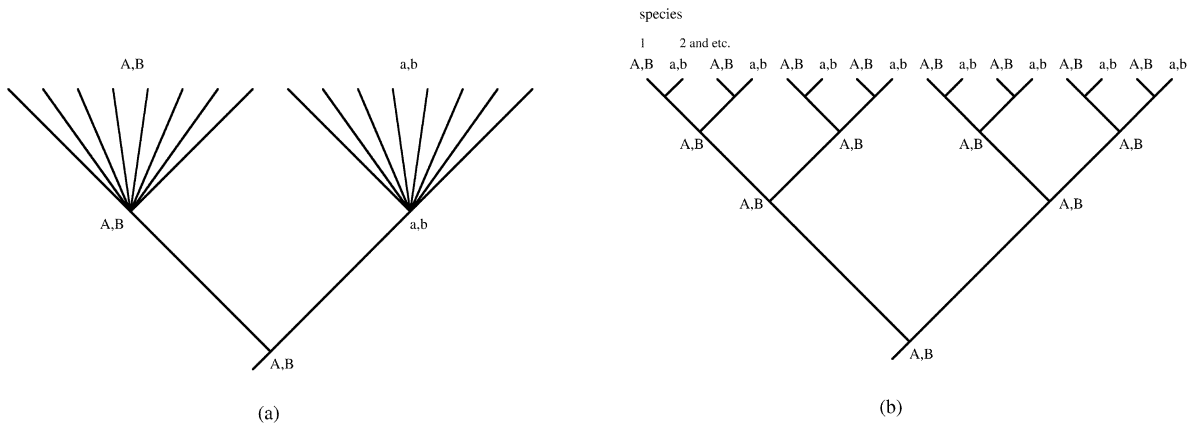

Figure 2
A phylogeny of sixteen species showing (a) two sets of eight species which have retained their ancestral values on each of two traits; (b) the evolution on eight separate occasions of the pair of traits {a, b}. Figure 2(b) provides the greater evidence for correlated evolution between the two traits, despite both figures having the same overall number of species with trait combinations {A, B} and {a, b}.

for comparative studies has been directed at this problem of independence and how to exploit phylogenies to reveal independent events of evolution.

2. *Phylogenies and Independent Evolution: A Brief History*

Three broad classes of method have been developed to identify independent events of evolution in phylogenies: so-called *higher-nodes* approaches, *phylogenetic-subtraction* techniques, and methods that partition the phylogeny into independent subsets of information. Of these three classes, those that partition the phylogeny came to be dominant. Very recently new statistical approaches have begun to replace these partitioning methods.

The *higher-nodes* approach seeks to identify a level of taxonomic classification that can be considered independent for statistical purposes and then restricts analysis to that level. For example, it is well established that in mammals, most of the diversity in brain and body size observed across species, is evident in differences among taxonomic 'orders' and 'families.' The Carnivora are an example of a Mammalian order and within the Carnivora the family *felidae* comprises the cats. The Artiodactyls and the *bovidae* (the cows) are another example of an order and family. Orders and families are commonly referred to as 'higher' taxonomic levels. In this instance, the higher nodes approach would proceed by analyzing differences among family means and counting families—not species—as the independent units of analysis. This would be analogous in Fig. 2(a) to comparing only the two nodes of the phylogeny. Here the procedure works by suggesting that there are only two independent data points in the phylogenetic tree, properly recognizing that the differences represented among the sixteen species arose independently perhaps only once.

Phylogenetic-subtraction methods proceed in a manner directly opposite to the higher-nodes in their attempts to identify independent evolution. These approaches seek to identify variation among species that is independent of their phylogenetic relationships to each other. Thus, taking again the example of brain and body size, the phylogenetic subtraction methods would subtract the appropriate taxonomic family-mean from each species data point. The resulting 'residual' value is taken to represent evolution that is independent of that species phylogenetic position. Applied to Fig. 2(a), the phylogenetic-subtraction methods would calculate the mean values at each of the two nodes, then subtract the appropriate mean from each species. Here the phylogenetic-subtraction methods would identify that none of the variability among species was independent of the immediately ancestral nodes.

Both the higher-nodes and phylogenetic-subtraction methods remove some portion of the variability among the species in an attempt to identify independent examples of evolutionary change among species. The successor to both of these classes of approach, *partitioning methods*, was based upon the recognition that the phylogeny itself provides a way of partitioning all of the variability amongst species into a set of, in principle, independent units of evolution. That is, none of the variability is discarded.

The best known partitioning method is that of *independent-contrasts*. A contrast is defined as a difference between two species, a species and a node of

the phylogeny, or between two nodes. Careful choice of the contrasts yields a set of difference scores such that each member of the set is independent of each other member and which collectively account for all of the observed variability amongst species.

Figure 2(b) shows how the independent-contrast methods work. The phylogeny shows that species 1 and 2 share an immediate common ancestor. Thus, the simple numerical difference between these two species on the trait of interest identifies some amount of evolution that occurred since they branched from their common ancestor. Similarly, differences among other pairs of species identify amounts of evolution that arose since they diverged from their common ancestor and which are independent of the other pairs of species in the phylogeny. The phylogeny identifies eight such differences among the tips of the tree. Moving back one level on the tree, four additional differences can be calculated, corresponding to pairs of nodes that immediately precede the tips of the tree. Each of these differences also measures a unique portion of the evolutionary change on the tree. Here, the comparisons reveal that no evolution has taken place on the traits between pairs of these more ancestral nodes. Moving back another level there are two more comparisons, and finally the last comparison, that between the nodes that represent the two descendants of the root or common ancestor of the whole tree.

In this manner, fifteen comparisons have been identified for the sixteen species. These fifteen comparisons, each one in principle independent of the others, exhaust the variability in the phylogeny, treating all of it as relevant to the hypothesis test. By comparison the phylogenetic-subtraction methods and the higher-nodes approaches always exclude some portion of the variability among the species' traits. In general, with a bifurcating tree (each node gives rise to two descendants) such as seen in Fig. 2(b), there will always be $n-1$ independent contrasts, given n species.

A set of contrasts can be calculated separately for each of two or more traits, then the correlation of the two sets of contrasts is used to test the hypothesis of coevolution. Importantly, each pair of contrasts provides an independent test of the hypothesis of coevolution: if, for example, brain size and body size covary, then each positive difference in body size between a pair of species (or nodes) should be accompanied by a positive difference in brain size. Further, larger differences in body size should be accompanied by larger differences in brain size.

3. *New Statistical Methods and the Analysis of Comparative Data*

The capability of independent-contrasts methods to use all of the variability in the phylogeny made them the favored technique of comparative analysis throughout the 1990s. A new set of statistical methods based upon the statistical approach known as *maximum likelihood* has since replaced these techniques. The new methods, like their predecessors, identify independent instances of evolution in the phylogeny. They do so without the mechanical procedure of forming contrasts or of calculating means at higher nodes and they have greater flexibility for manipulating the phylogeny and the species data. They also retain information that is lost in the independent-contrast approach. For example, contrast methods cannot estimate the ancestral state at the root of the phylogeny.

In words, the new methods use the phylogeny to specify the expected lack of independence among species. They then use that information to adjust the statistical analysis internally to reflect the lack of independence. The new methods have been developed for two broad classes of data: *discrete* or *categorical-variable* methods are applied to data that fall into a small number (typically two) of discrete classes or categories. Diet type, mating system, and the presence or absence of some feature or behavior are examples of this kind of data. The statistical Markov-model provides a natural framework for discrete traits (Pagel 1994). *Continuous-variable* methods are applied to traits that are measured on a continuous scale. Most traits relating to size, shape, timings, frequency, or concentrations are measured on continuous scales. For this kind of data a statistical technique known as *generalized least squares* or *GLS* is well suited to analyze trait evolution on a phylogeny (Garland 2000, Schluter et al. 1997, Martins and Hansen 1997, Pagel 1997, 1999)

3.1 *Correlated Evolution of Continuously Evolving Traits: The GLS Approach*

The most commonly used model for continuously evolving traits is the constant-variance random-walk process (sometimes called Brownian-motion). In the conventional random-walk model, traits evolve each short instant of time dt by some random amount that might be positive (for example, the trait gets larger) or negative. The successive random changes to a trait are independent of each other, they have a mean or average of zero, and some unknown and constant variance, denoted σ^2.

The random-walk is presumed to unfold independently at each instant of time and along each of the branches of the phylogeny. Given this, the phylogeny specifies how much a given species' trait value is expected to vary were evolution to be rerun as a random walk many times (this is the variance, σ^2), and how similar any pair of species are expected to be. These are the quantities that are needed to adjust the statistical analysis for nonindependence amongst species.

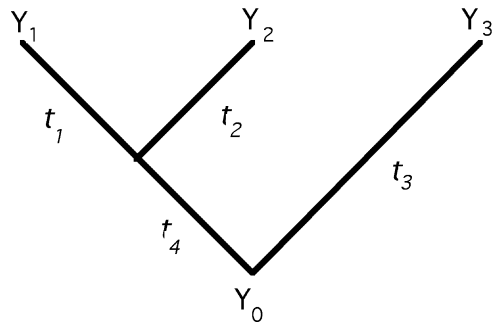

Variance-Covariance matrix

$$V = \begin{bmatrix} \sigma^2(t_1 + t_4) & \sigma^2 t_4 & 0 \\ \sigma^2 t_4 & \sigma^2(t_2 + t_4) & 0 \\ 0 & 0 & \sigma^2 t_3 \end{bmatrix}$$

Figure 3
The variances and covariances among species (matrix V) implied by the phylogeny (left), under the assumption of a constant-variance (Brownian motion) model of evolution. The symbols 'Y' represent some trait that is measured on the three species. The value Y_0 corresponds to the ancestral value at the root of the tree.

The variance of any species' trait value turns out to be given by $t\sigma^2$, where t is the total path length from the root of the phylogeny to the species (the summation of all the small dts); t might be measured in units of time or in units of 'operational time' such as genetic divergence. Closely-related species will tend to have similar trait values, even under a random-walk, owing to sharing most of their evolutionary history. The expected covariance (like a correlation) between two species in their trait values is given by $t_s\sigma^2$, where the subscript s denotes the shared path length from the root up to the point at which the two species separate.

The generalized least squares model represents the information about the expected variances and covariances of species in a matrix, usually denoted V. Figure 3 shows the matrix V that is implied by the simple phylogeny depicted there. GLS statistical methods use V to adjust internally for the expected lack of independence among species. This means that one can simply analyze the species data points as if they were independent, the nonindependence having been controlled for automatically by V. It also means that one can specify transformations of V that effectively change the underlying model of evolution (Pagel 1999).

3.2 Mammalian Brain-size Evolution

The evolution of mammalian brain size provides a good example of the use of the GLS approach to analyze comparative data. Theories of the evolution of mammalian brain size link brain volume to body mass either through body surface area or through basal metabolic rate. The surface area hypothesis requires that the slope of the line relating brain volume to body mass is 2/3; the metabolic rate hypothesis predicts that the slope of brain volume against body mass will be 3/4, owing to the relationships of basal metabolic rate to body and brain size. Empirically derived brain-body slopes generally lie somewhere in the 2/3 to 3/4 range.

Applied to the data of Fig. 1 and using a suitable phylogeny of the mammals, the GLS model returns a slope relating brain mass to body mass of 0.61. This is lower than the expected 2/3 to 3/4, despite a very strong correlation between the two traits (Table 2). Even in these comparatively small samples, the 95% confidence intervals exclude 3/4, and nearly exclude 2/3. The GLS model also estimates the ancestral mammalian brain and body sizes, here found to be a brain mass of 20 grams and a body mass of about 5 kilograms. By comparison to the brain and body size relationship, the slope of basal metabolic rate to body mass (Table 2) is near to the expected value of 0.75 (Table 1).

The GLS model returns unfavorable support for the surface-area and metabolic rate theories of mammalian brain size. Theories linking brain-adaptations to specific environmental demands may offer new ways of thinking about brain evolution.

3.3 Testing the Adequacy of the Model of Evolution

The GLS approach makes it possible to test the adequacy of the underlying constant-variance (Brownian motion) model and to test hypotheses about trait evolution (Pagel 1999; see also Table 3). This is done by transforming the elements of V. A parameter κ scales individual phylogenetic branch lengths and can be used to test for a punctuational mode of evolution; a parameter δ scales overall path lengths in the phylogeny—the distance from the root

Table 2
Slope of the regressions of brain-volume and basal metabolic rate against body size in mammals. Derived from directional GLS model. Ninety-five percent confidence intervals in brackets.

Variable	Number of species	correlation	GLS regression slope, 95% CI	Theoretical expectation
Brain volume	23	0.96	0.59 (0.52–0.67)	0.67–0.75
Basal metabolic rate	15	0.95	0.72 (0.60–0.84)	0.75

Table 3
Three scaling parameters available to the GLS model and their interpretation when applied to trait evolution on a phylogeny

Parameter	Action	0	< 1	1	> 1
λ (lambda)	Assess contribution of phylogeny	Star phylogeny (species independent)	Phylogenetic history has minimal effect	Default phylogeny	Not defined
κ (kappa)	Scale branch lengths in tree	Punctuational evolution	Stasis in longer branches	Default gradualism	Longer branches more change
δ (delta)	Scale total path (root to tip) in tree	Not defined	Temporally early change important (adaptive radiation)	Default gradualism	Temporally later change (species-specific adaptation)

to the species—and can detect adaptive radiations; and a parameter λ asks whether the phylogeny correctly predicts the patterns of variance and covariance. If species are evolving as if they are independent, this parameter will indicate that phylogenetic correction can be dispensed with, and will automatically do so.

For the brain and body size data, the 95 percent confidence intervals of all three parameters overlap 1.0, the default values under the constant-variance model: $\kappa = 1.06$ (0.42–1.63); $\hat{\delta} = 0.96$ (0.078–2.79); and $\hat{\lambda} = 1.0$ (0.94–1.0). These values justify using the phylogeny to correct for nonindependence, and indicate that trait evolution is proportional to the branch lengths.

3.4 Correlated Evolution of Discrete Traits: The Markov Model

The Markov-transition process is the statistical model widely used to describe the evolution of traits that adopt only a finite number of states. The Markov approach estimates the rates at which a trait or character makes transitions amongst its possible states as it evolves over time. A trait is assumed to begin some epoch of time in one of its possible states and then at each short epoch of time dt it can jump to one of its other states. The Markov transition rates are the rates at which these jumps are made. These rates are sufficient to calculate the most probable states at any node or tip of a phylogeny, given some starting point at the root. Similar to the situation with continuously evolving characters, closely related species will typically have similar trait values under a Markov model, whereas more distantly related species are not expected to be as similar.

To test for correlated evolution between two binary discrete characters the Markov model compares the fit of two models to the data. In one the two traits are allowed to evolve independently; in the other they evolve in a correlated fashion. Evidence for a correlation is found if the model of correlated evolution fits the data significantly better than the model of independent evolution (Pagel 1994).

For a trait that can take only two values (e.g., 0,1), two rates must be estimated, one for transition from '0' to '1,' and the other for transitions from '1' to '0.' Four parameters are required for two traits evolving independently. The model of correlated evolution considers the four possible states that two binary characters can jointly adopt (0,0; 0,1; 1,0; 1,1). It then allows one of the variables to change state in any branch of the tree, yielding eight possible transitions to be estimated (Fig. 4). These can be shown to be sufficient to calculate the probability of any kind of change in any branch of the tree, and they can be used to chart the most probable course of evolution from the ancestral state to the contemporary derived state.

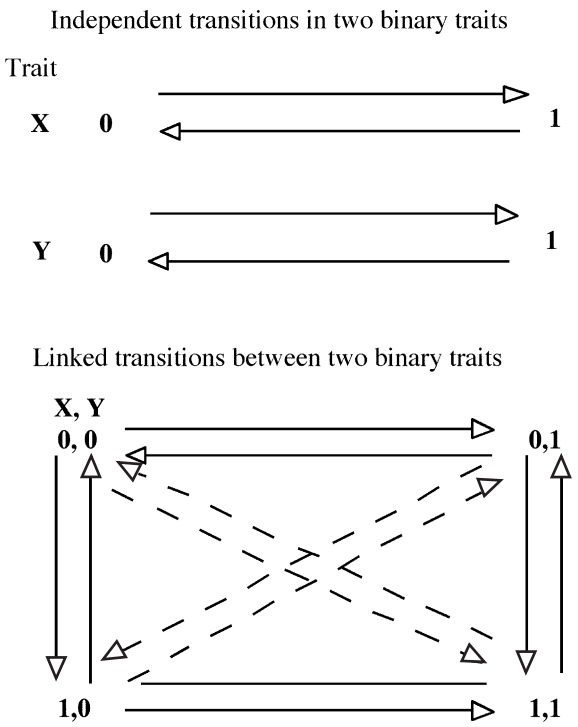

Figure 4
Independent transitions between two binary states in two traits (upper); Linked or correlated transitions in two binary traits (lower). Dashed lines are not calculated.

3.5 The Evolution of Lactose-tolerance

The evolution of lactose tolerance in human societies illustrates how to test for correlated evolution using two discrete traits. The lactase enzyme confers an ability to digest milk. Human infants can digest it, but most adults cannot (widespread tolerance to lactose amongst some European groups is exceptional). A dominant view is that adult lactose-tolerance in humans is an adaptation to reduced exposure to the sun: both the sun and the lactase enzyme promote calcium absorption. The hypothesis accounts for the lactose-tolerance of some northerly-dwelling cultural groups such as the reindeer-herding Lapps, and the prevalence in some European groups. An alternative to the latitude theory is that adult tolerance to lactose is advantageous in cultures that keep animals for milk. If milk forms a significant portion of the diet, selection pressures on adults to develop the ability to digest it could be strong. Thus, two binary traits are available (lactose tolerant/intolerant and dairy/nondairy subsistence) and the adaptive hypothesis is that they coevolve.

The Markov-process model applied to a phylogeny of human cultural groups shows that adult lactose-tolerance has arisen independently in humans up to three times in cultures that keep animals for milk, but never in nondairying cultures (Holden and Mace 1997). This association is sufficiently strong to reject the model that the two traits have evolved independently in human groups, in favor of the view that they have coevolved. The Markov approach also reveals the probable course of the evolution of lactose-tolerance: the ancestral condition of nondairying and no adult lactose-tolerance was first replaced by dairying, perhaps as early as 6,000–8,000 years ago, which then favored the evolution of tolerance to lactose. The data do not support the alternative scenario that human groups with adult lactose-tolerance were more likely to adopt dairying subsistence practices.

4. General Discussion

Comparative methods remain a popular set of techniques for investigating adaptation. New statistical approaches are easy to use and make use of more of the information in the data than previous methods. Coupled with increasingly accurate and informative phylogenies based upon gene-sequence information, comparative methods can be used to test new hypotheses and to re-examine old ones. The availability of statistical parameters for scaling branch lengths, and the phylogeny itself usher in a new range of hypothesis tests for comparative data.

See also: Body, Evolution of; Brain, Evolution of; Comparative Neuroscience; Comparative Psychology; Comparative Studies: Method and Design

Bibliography

Garland T Jr, Ives A R 2000 Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *American Naturalist* **155**: 346–64

Harvey P H, Pagel M 1991 *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, UK

Holden C, Mace R 1997 A phylogenetic analysis of the evolution of lactose digestion. *Human Biology* **69**: 605–628

Martins E P, Hansen T F 1997 Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist* **149**: 646–667

Pagel M 1993 Seeking the evolutionary regression coefficient: An analysis of what comparative methods measure. *Journal of Theoretical Biology* **164**: 191–205

Pagel M 1994 Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society Bio* **255**: 37–45

Pagel M 1997 Inferring evolutionary processes from phylogenies. *Zoologica Scripta* **26**: 331–48

Pagel M 1999 Inferring the historical patterns of biological evolution. *Nature* **401**: 877–84

Pagel M, Harvey P H 1989 Taxonomic differences in the scaling of brain on body weight among mammals. *Science* **244**: 1589–93

Schluter D T, Price A, Mooers Ø, Ludwig D 1997 Likelihood of ancestor states in adaptive radiation. *Evolution* **51**: 1699–1711

M. Pagel

Comparative Neuroscience

Ever since the brain was recognized as responsible in some way for behavior, awareness, and mental activity, the question has been alive: how do our brains compare with others, from apes to flat worms?

This question has frequently been investigated solely by comparing brain sizes. Specialists in comparative anatomy have also pointed out differences between major taxa, such as mammals and reptiles, in the layering of the cerebral cortex and the development of a corpus callosum, among other gross features. With the invention of a battery of new methods in experimental microanatomy that make possible the tracing of first order neuronal connections (those not interrupted by synapses), an extensive body of knowledge has accumulated on comparative hodology—the projections into and out of identified brain regions, nuclei, and laminae. There is, in addition, a more limited literature on the comparison of neurally active compounds, such as transmitters and modulators, among the major taxa, including phyla.

Still less advanced is knowledge of mesoscopic and microscopic organization, including physiological as well as anatomical features that differ among animals, especially those relevant to behavior. Examples are: (a) cell types, especially those defined by response properties; (b) local circuits intrinsic to a nucleus or laminated structure; (c) the degree of reciprocity of connections between structures at the same level—such as cortico–cortical—and at different levels—such as thalamo–cortical; (d) dynamical interactions (potentiation, induced rhythms, event-related potentials, brain waves); (e) kinds of information processing (command cells, head-position compensating cells, pattern-sensitive cells, face-selective cells); and (f) emergent properties (kindling, seizures, learning of each of several kinds). Virtually unknown and awaiting future study are the neural substrates of behaviorally defined traits such as (g) the number and resolution of discriminable stimuli like individual voices, smells, and places within the environments experienced—roughly knowledge; (h) the number and qualia of emotions, sets, moods, and status symbols that an individual can manifest and can discriminate in others; and (i) the variety, specificity, and longevity of memories of past events.

This list, obviously, is not exhaustive and serves simply as a framework upon which to review the state of comparative neuroscience today.

1. *Evolution of the Nervous System*

The nervous system and its sense organs and effectors have evolved through a greater span of complexity than any other system within living organisms, if we consider the animals with the simplest nerve nets and those with the most complex brains. It has not evolved linearly, steadily, or inevitably but with repeated instances of retrograde loss of complexity. But, overall and roughly, the more primitive taxa represented among surviving species have very simple nervous systems as well as behavior, relative to certain branches of the most derived taxa, which have extremely advanced (i.e., complex) brains and behavior.

This difference was already strong long before hominids or primates or mammals or vertebrates appeared. Even if we consider only gross anatomy and crude histology of the nervous tissue, a comparison of cnidarian medusae to polyclad flatworms, or of pulmonate snails to octopods, of agnathan hagfish to mormyrid teleosts, or of reptiles to mammals, dramatically underlines a diversity of grades of organization differing, above all, in complexity.

By complexity, in this context, we mean the sum of the kinds of elements and operations. This includes kinds of structural elements (distinguishable cell types on all possible criteria, plus groupings of cells, and structural parts with subdivisions at all levels), the kinds of functional connections and interactions, of physiological properties and influences, and of discriminable stimuli and behavioral actions. Each of these would be difficult to count but is concrete and not vague. On this tangible scale, one approachable by degrees, we can recognize higher and lower species, purely with respect to complexity.

We can thus estimate the degree of independence of evolution of traits and systems of traits, for example vision and movements, or vision and olfaction. We can approximate a distinction between major steps in grade (gastropods and cephalopods; reptiles and mammals), on the one hand, and on the other, adaptive specialization within roughly the same grade of overall complexity (e.g., life styles and adaptations among rodents or among bats). We can estimate that most evolution has been adaptation within grade, at the level of species, genera and families, whereas saltations in the major grades of complexity have been relatively few and clearest at the levels of (some) classes and (some) phyla. To see the big picture we should eschew arguable questions at the level of families and even higher taxa of roughly equivalent advancement, such as whether bats are a higher grade than rodents, or lobsters a higher grade than cock-

roaches. It is best to compare phyla and classes, sometimes orders, choosing the most advanced species in each and avoiding taxa that might be secondarily simplified.

With these caveats, made to dispel charges of anthropocentric or moralistic, value-laden reasoning or a simplistic, linear animal scale, we can look at the grand sweep. Most of the time we scrutinize evolution with respect to how it happens, or how it is adaptive or what phylogenetic branching is most likely. We tend to overlook the perspective of consequences in terms of major grades of complexity. A grand fact of animal evolution is the development of more and more complex brains and, probably, also behavior—slowly and erratically, without a clear and self-evident increase in survival value but spanning an overall series of levels of advancement almost impossible to overstate.

The main thesis of this essay is that modern biology has neglected to study the details of the differences between grades of complexity. Beyond the gross and macroscopic anatomical differences in classical textbooks, such as the mammalian hippocampus, corpus callosum, corticospinal tract, and six-layered isocortex, we lack an accumulated body of comparisons—at each of the relevant integrative levels from molecule to cognitive system. This is surely a rich vein of ore for future investigators to mine.

2. *Macroanatomy and Connectivity*

The state of knowledge of the gross and microscopic anatomy of the nervous systems of invertebrates (Bullock and Horridge 1965) and vertebrates (Butler and Hodos 1996, Nieuwenhuys et al. 1998) can be found in text and reference books, richly illustrated. Four levels of inquiry may be distinguished.

2.1 *Brain Size*

Brain size is the crudest, usually plotted against body size, since a regression of these two measures is general. Both measures have large uncertainties, including how much of the normal cerebrospinal fluid, blood, blood vessels and plexuses, meninges, and nerve roots should be included, how the brain was removed and preserved, how body weight is taken and how exoskeleton, horn, shell, gastrointestinal contents, fat, preservative, and other variables are treated. The regression is characteristically very low for plots of individuals of the same species, substantially higher for species of the same family, and still higher for higher taxa. This makes compensating for size, sex, age, nutritive state, and cutaneous appendages only an approximation. The general result is clear—that brains considered more advanced on other criteria tend to be larger in proportion to the body. To an unknown extent the larger brains are considered to reflect greater sensory input and a larger variety of discriminable stimuli plus a greater efference and larger variety of responses. These, however, are essentially untested hypotheses and it can be seriously doubted that these factors adequately account for the size differences.

2.2 *Brain Lobes and Nuclei*

Brain lobes, clumps of nerve cells (nuclei), and gross subdivisions of these are potentially more meaningful than sheer size to compare across taxa, life style, and behavioral adaptations. Many examples are known where prominence of optic lobes, auditory nuclei, areas of the cortex representing snout or whiskers, bird-song centers, or hippocampal place-memory enlargements correlate with behavioral adaptations. Yet many other features of comparisons between species are not obviously explicable in terms of the known habits of life—as for extreme differences in cerebellar size among families of elasmobranchs, size diversity in the habenula and corpus striatum, or diversity of layering in visual cortex among primates and in the dorsal cochlear nucleus among rodents. This level of analysis might be expected to be more sensitive than others to behavioral complexity, especially in the degree of differentiation of the relevant brain parts. It seems likely that enumeration by equivalent standards of the number of distinguishable subdivisions of the thalamus or hypothalamus, or of the cortex as between reptiles and mammals—taxa at quite distinct grades of complexity—would show a greater variety in the mammals. But we have, so far, very few attempts to make such comparisons. This would quantify the strong impression from qualitative histology that the brains in the most advanced teleosts are more complex than those in the most advanced amphibians, as well as elasmobranchs.

2.3 *Connections, Afferent and Efferent, into and out of Each Nucleus or Lamina*

Connections, afferent and efferent, into and out of each nucleus or lamina, indeed each cell type, are now extensively worked out in both invertebrates and vertebrates, thanks to an ever expanding battery of methods that reveal origins and destinations of axons, and even to some extent of dendrites, cell by cell. Hodology offers a multitude of comparisons, often complicated by a problem peculiarly serious at this level. That is the problem of what criteria are adequate for establishing which cell group in one family or order is equivalent—preferably homologous—to a cell group in another taxon. Besides the simple list of connections in and out of every structure, some significant general features of comparison are op-

portune. How do two taxa compare in the number of connections between higher and lower brain levels, ascending and descending, between regions at the same level (e.g., corticocortical), between cells of the same local region (often called interneurons or intrinsic neurons), and how localized are the terminations?

2.4 Identifiable Neurons

Widespread among invertebrate phyla and present to a limited extent in anamniote vertebrates is the phenomenon of identifiable neurons. These are neurons that can be recognized and named as the equivalent cell in each individual examined, hence one-of-a-kind cells, most often duplicated on left and right sides. Defined by a combination of anatomical and physiological features, especially their afferent and efferent connectivities, the cells with the same name in different individuals typically belong unequivocally to their unique class but are not identical. Identified neurons are conspicuous and common in leeches, probably in some polychaetes, insects, crustaceans, some gastropods, and a few other taxa. Some workers believe that most of the neurons in certain animals, such as the tectibranch gastropod, *Aplysia*, are potentially identifiable. A small number of identified neurons are known in lampreys and a pair in many teleosts. It is not clear what factors of phylogeny, or life style correlate with their occurrence. The question of intermediate classes of cells has been raised—those cells which belong, not to a class of 1 but 2 or 10 or 100. It is quite plausible that smallish equivalence classes are abundant even in mammals, taking account of all distinguishing features—immunological, dendrological, and especially functionally relevant connectivity. The significance of this notion is that the equivalence classes: (a) are in fact the biologically important kinds of cells in the nervous system; (b) may run into very large numbers—probably far beyond any other organ in the body; and (c) may represent, in principle, one of the best measures of evolutionary advancement, defined as complexity.

3. Cellular Processes and Messengers

The breakthrough of Hodgkin and Huxley gave us two specific ion channels with distinct roles. Simplifying history, Hagiwara added calcium channels in some tissues of some species. Today we have scores of distinct channels (see *Ion Channels and Molecular Events in Neuronal Activity*) and no clear picture of their distribution among cell types or animal species. It appears not to be the case that a wide array of different channels was evolved only in later, more derived or more complex animals. Studies on relatively primitive living groups show a high diversity of channels. It is still possible that general principles will emerge relating the relative abundance of different kinds of channels—their sensitivities, dynamics, and interplay—in different kinds of excitable cells and then among the taxa.

Much the same seems to be true for transmitters, modulators, and intracellular messengers. The half dozen or so transmitters, a rather stable list (see *Neurons and Dendrites: Integration of Information*; *Neurotransmitters*), are each quite widespread among phyla and differences in distribution are mainly between functional systems within a given brain. Modulators are increasing in number; currently there are scores. Several are known to be widespread but no simple picture is yet tenable as to evolution or distribution, particularly if we ask for evidence of what the compounds do neurally.

4. Mesoscopic Traits and Organization

If the vast differences between brains of simpler invertebrates and complex mammals are not obviously correlated with vast differences in the membrane and cellular machinery, could sheer numbers of cells and connections explain complexity? Multiple sources of evidence argue against this as adequate, even though mere numerosity doubtless contributes largely. Already mentioned is the domain of the increase, probably many orders of magnitude, in the number of kinds of neurons. Among the differences are permutations of neuronal personality traits that determine the output of a neuron as an integration of its weightings assigned to inputs and their spatiotemporal patterns and its own proclivities (see *Long-term Potentiation* (*Hippocampus*); *Long-term Depression* (*Cerebellum*); *Neurons and Dendrites: Integration of Information*; *Long-term Depression* (*Hippocampus*); *Long-term Potentiation and Depression* (*Cortex*)). Lists of integrative variables, each with a spectrum of possible states, run to well over 50 and new items continue to be discovered, such as the recently documented supralinear summation of multiple synaptic inputs upon dendritic gain mediated by an 'inward-rectifier' potassium current in leech pagoda cells (Wessel et al. 1999). Differentiation not only of kinds of cells but of assemblies, columns, and circuits appears to have resulted in emergent systems, and these in higher levels of systems. Between the microscopic, cellular, and subcellular levels of many components and interactions and the realm of human achievements, such as recognizing a national political opportunity, there are layers of mesoscopic organization that embody the main variety of evolutionary novelties in the grades of brains and behavior.

This mesoscopic level is where our knowledge of comparative neuroscience is weakest. We lack the

advantages of lower, cellular levels where the component elements, structures, and interactions are extensively known in well-studied species. The higher behavioral levels are more difficult to compare but at least provide many measurable variables, just because of their familiarity. The mesoscopic levels that hide the first-order explanations lack both of these advantages and lack even a recognized list of relevant measurable variables. As examples, to suggest the wealth of potential discoveries as well as their slipperiness, one thinks of electrical signs of assembly activity—brain waves, evoked and event-related potentials, sleep stages, seizure stages, patterns of cortical coherence during several kinds of thinking, localized increase in blood flow or oxygen consumption during subtle cognitive states, clinical observations of blindness to animate objects or denial of one's own left side after local lesions, and 'magic bullets' represented by neuroactive molecules that ameliorate clinical depression or obsessive-compulsive disorder. The comparative dimension will surely yield valuable perspectives to this level of analysis.

5. *Discriminable Stimuli, Responses, and States*

Any overall comparison or estimate of the degree of complexity of the nervous system as a whole, including special adaptations to a particular life style, must take the whole ethogram into account, particularly the range and variety of discriminable social situations, sign stimuli, food, enemy, habitat, and microclimatic signals, plus the range and variety of behavioral responses and internal states. The last named stand for the conative states of readiness, biases, changes of weightings, drives, and moods. The obvious difficulty of enumerating, even of discerning all of these—each representing a theoretically distinguishable spatio-temporal neural pattern—is no reason to neglect their recognition.

Even this list may allow us to forget a major measure of the capacity of the nervous system—one which is hardly touched in the bulk of the literature on comparative behavior and cognition. That is, the extent and depth and detail of the knowledge that an organism can acquire, retrieve, and use in its lifetime. This, like most of the foregoing list, may not be the task of the neuroscientist but comparative neuroscience should be based on the most complete ethology. In other words, it embraces neuroethology at every level up to the most subtle behavioral and cognitive abilities.

See also: Behavioral Neuroscience; Brain Development, Ontogenetic Neurobiology of; Brain, Evolution of; Classical Conditioning, Neural Basis of; Cognitive Neuroscience; Memory: Synaptic Mechanisms; Neural Development: Mechanisms and Models

Bibliography

Bullock T H, Horridge G A 1965 *Structure and Function in the Nervous Systems of Invertebrates*. W. H. Freeman, San Francisco

Bullock T H 1993 *How do Brains Work*? Birkhäuser, Boston

Butler A B, Hodos W 1996 *Comparative Vertebrate Neuroanatomy: Evolution and Adaptation*. Wiley-Liss, New York

Carew T J 2000 *Behavioral Neurobiology*. Sinauer, Sunderland, MA

Cohen M J, Strumwasser F 1985 *Comparative Neurobiology—Modes of Communication in the Nervous System*. Wiley Interscience, New York

Nieuwenhuys R, ten Donkelaar H J, Nicholson C 1997 *The Central Nervous System of Vertebrates*. Springer, New York

Roth G, Wullimann M F 2001 *Brain Evolution and Cognition*. Wiley, New York

Wessel R, Kristan W B Jr, Kleinfeld D 1999 Supralinear summation of synaptic inputs by an invertebrate neuron: Dendritic gain is mediated by an "inward rectifier" K+ current. *Journal of Neuroscience* **19**: 5875–88

T. H. Bullock

Comparative Psychology

Comparative psychology is a diverse field, one which is characterized by at least four different approaches to its subject matter. The most venerable of these approaches is to describe similarities and differences in the behavioral and psychological abilities or capacities of various species, irrespective of their evolutionary lineage (relationship) (see Sect. 1). A variant of this first approach is to study the behavior or learning abilities of various species with a view toward describing the advancement or progression of psychological functioning going from the more simple to the more complex, in lines that are not necessarily closely related in an evolutionary sense; this approach is called anagenesis (or the *levels* approach) and typically shows improvements in developmental plasticity, learning ability, etc. in more recently evolved forms. The point of this approach is to discern broad trends in evolution, especially psychological evolution, and it is based on a hierarchical conception of learning abilities, going from the simple to the complex.

A more recent variant of that approach is to describe similarities and differences in species that are closely related in an evolutionary sense, and to try to interpret these results with respect to the evolution of particular behavior (see Sect. 2). Another approach practiced by comparative psychologists is to study the species-typical behavior of animals in the field throughout their life cycle (see Sect. 3). That approach was made more popular by the zoologically-trained animal behaviorists in Europe, who called themselves etho-

logists to distinguish themselves from psychologically-trained animal behaviorists in the United States, many but not all of whom confined their studies to the laboratory. Finally, there is the animal model approach, which tries to find nonhuman species with behavioral or psychological repertoires that are similar to humans so that the results of experiments with the model may throw light on seemingly related behavior in human beings (see Sect. 4).

As stated by T. C. Schneirla (1962, p. 690R), 'Comparative psychology is studied because the behaviour of animals is in itself interesting, because of important applications in fields such as medicine and animal training, and also because of significant bearings on human psychology.' The remainder of this article gives examples of each of these approaches.

1. *Similarities and Differences in Distantly Related Species*

One of the most fundamental similarities in animals of all species is their ability to stop responding to a repeated stimulus in the absence of reward. This phenomenon is called habituation and it is, along with sensitization (described next), the lowest form of learning. It is of course adaptive for animals to cease responding to a stimulus event that poses no threat and offers no reward. In snails, for example, when the substrate upon which they are walking is given a mechanical shock, they withdraw their tentacles for a brief period. When the shock is repeated in exactly the same way at regular intervals, the extent and duration of tentacle withdrawal steadily decreases until finally the animal becomes completely indifferent to the stimulus. Sensitization is the obverse of habituation in that animals will orient to abruptly presented sights, sounds, smell, etc., and their orientation reaction will be quicker and longer if the stimulus changes somewhat from trial to trial. Both habituation and sensitization are nicely demonstrated in an experiment using the wiping reflex of frogs. When the frogs are touched with a hair on precisely the same zone of skin from trial to trial, their wiping reflex declines (fewer wipes per 100 stimuli). However, when the same hair is used to stimulate the frog's skin at somewhat different places each time, the wiping reflex is enhanced (more wipes per 100 stimuli). The next higher form of learning is called associative (stimulus–response or S–R) learning or conditioning, and was made popular by Ivan Pavlov's experiments with the salivary response of dogs. In these experiments, a bell or tone was sounded followed by the sight of meat powder. The hungry dog salivated at the sight and smell of the meat powder (unconditioned response). After numerous pairings of the to-be-conditioned stimulus (tone) with the unconditioned stimulus (meat powder), the dog began to salivate at the sound of the tone (conditioned response) in advance of the presence of the meat powder. Associative learning of this kind—called classical conditioning—has been demonstrated in all vertebrates from fish ('lowest' vertebrate) to humans ('highest' vertebrate). It has also been demonstrated in some invertebrate species, but not in all, so this must be considered a higher level of learning than habituation and sensitization. There is another form of conditioning—operant learning—that strengthens existing responses through reward (food for a hungry animal, e.g.) or reduces existing responses through punishment (electrical shock, e.g.). This form of conditioning is widely practiced among animal trainers and all vertebrates and some invertebrates are capable of it. The next psychological level as far as learning is concerned is trial-and-error or selective learning, in which there is a significant perceptual, perhaps cognitive, component, in the sense that the animal must discover a specific relationship between its activities and a goal object. As first studied by E. L. Thorndike in the late 1890s in connection with his doctoral dissertation, a variety of vertebrate species were placed in a puzzle box or a maze, from which they had to escape in order to get food or join companions. Thorndike's idea was that the animal had to learn to associate making certain movements (R) in a problem situation(s) (pulling a cord or unlatching a lock) with a subsequent 'satisfying' state of affairs (getting to food or joining companions). So the S–R association in this context is the situation and the successful movement leading to the satisfying state of affairs. While some invertebrate species can do trial and error learning, this class of learning reaches its peak in mammals. The highest category of learning ability, considered by some the true criterion of intelligent behavior, is reasoning or, in other terms, insightful problem-solving. This is seen in nonhuman animals when an organism solves a problem by using as a tool (means) some object(s) that never previously in that animal's experience had been used in that way. The most famous instance of this kind was Wolfgang Köhler's chimpanzee Sultan, who, in order to reach a banana high in his cage, stacked boxes one upon another to make a ladder enabling him to reach the banana. While no doubt occurring in humans and chimpanzees, comparative psychologists have been sparing in applying this appelation to other mammalian species.

There is a form of learning that takes place through mere exposure to physical and social objects; the exposure leads to subsequent attachment, preference, or acceptance of these objects (i.e., animals and humans not exposed to these objects do not show attachment, preferences, or acceptance of these objects). This familiarity mechanism is pervasive; it is found in many invertebrate species as well as vertebrates (Szentesi and Jermy 1989), and it is often referred to as imprinting, although, technically speaking, imprinting involves a sensitive period early in life

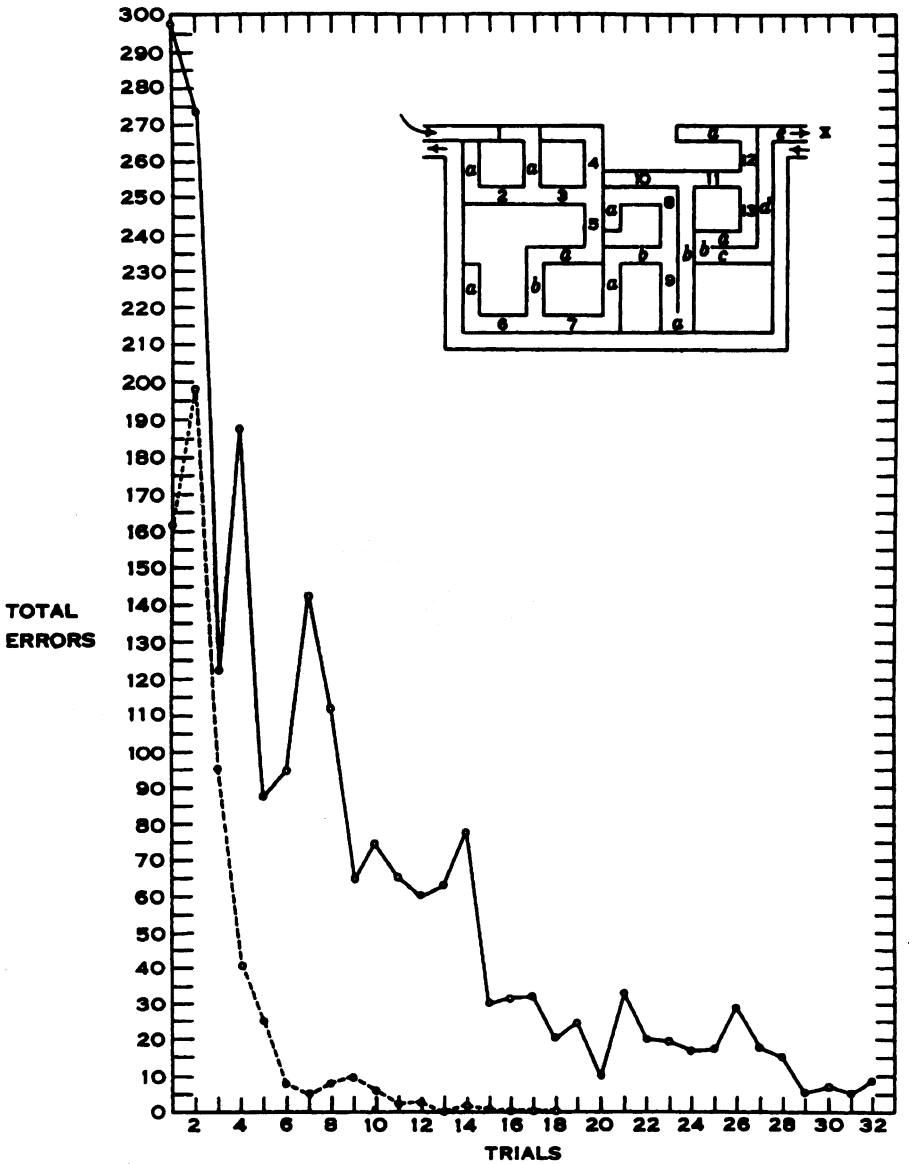

Figure 1
Six-blind maze problem used for testing both ants and rats. Solid line indicates total errors for eight ants; broken line indicates total errors of eight rats. Ants and rats learned maze by going in the direction 1 to 13*e*; rats to food at X, ants to their nest at X (from Schneirla 1953. Reproduced with permission of John Wiley & Sons, Inc.).

and is usually defined as inducing an enduring preference not merely acceptance (nonavoidance) of an object. While particularly striking in young animals (Sluckin 1965), exposure learning operates throughout life in some species such as our own (Hebb 1946, Zajonc 1971). Exposure learning is difficult to classify with certainty in the hierarchical scheme adopted here. Since it is so pervasive and is brought to bear by mere exposure, it probably fits in above the level of habituation and sensitization because of its requirement for a higher degree of perceptual differentiation and longer term memory. It would almost always be adaptive to form attachments, preferences, and acceptance of familiar physical and social objects.

There are two approaches to classification: lumping and splitting. The hierarchical classification of adap-

Table 1
Results of tests with displacement of food.

		Walks around the screen on the side					
		With food		Empty		Total	
Species	Total trials	Trials	Percentage	Trials	Percentage	Trials	Percentage
Pigeons	27	2	7.4	0	0	2	7.4
Chickens	129	35	27.1	32	24.8	67	51.9
Crow family	42	30	71.4	6	14.3	36	85.7
Rabbits	33	8	24.2	1	3.0	9	27.2
Cats	42	21	50.0	15	35.7	36	85.7
Dogs	18	14	77.8	2	11.1	16	88.9

Data from Krushinsky, 1965, p. 283.

tive or adjustive learning abilities into habituation, sensitization, exposure learning, conditioning, trial-and-error learning, and reasoning is an example of lumping. In 1971, Gregory Razran published a book, *Mind in Evolution: An East–West Synthesis of Learned Behavior and Cognition*, in which he described 11 forms or levels of learning. He included the ones described here (except exposure learning) and further broke down the various forms of conditioning and also added 'symboling' to include the role of language in what Razran called 'the learning of thinking, planning, and willing.' Razran's treatise is without equal in the literature concerning the comparative analysis of learning ability in a large number of species, as well as introducing Western scientists to the large literature that appears only in the Russian language.

A striking example of behavioral anagenesis is provided by an experiment in which the progression in learning a maze is observed in ants and rats (Schneirla 1962). As shown in Fig. 1, the ants took more than twice the number of trials to get their error rate down to near zero, as compared with the rats. Not only do the rats learn the maze much more quickly than the ants, they do it in a different fashion. Note the large reductions in errors from trial to trial in the rats' learning curve compared with the ants', in which there is slow and gradual progress.

Schneirla observed not only a quantitative difference in the behavior of the rats and ants but also a qualitative difference:

> Not only do the rats make fewer errors and require fewer runs than the ants to master this maze problem, but their learning process in it differs markedly in several respects from that of the ants. The three stages so clearly marked in the ants cannot be distinguished reliably in the rats. The rats, in fact, progress from the very first run in shortening and eliminating their entrances into the blind alleys; moreover they drop out some of the blinds quickly and almost as units and others 'by jumps' although less rapidly. The ant's pattern of eliminating blind-alley entrances gradually by small degrees is very unlike the rat's response to this problem (Schneirla 1962, p. 513).

Anagenic trends in learning (improvements over the course of evolution) have always been considered in light of evolutionary changes in brain structure and function. In an instructive experiment with three species of birds and three species of mammals, L. V. Krushinsky and colleagues (1965) examined what he considered the ability of these different species to solve an 'elementary logical problem' by what he called 'extrapolation' without any specific training. The task involved the animals facing a screen with a gap in its center. A bowl of food passed the gap moving either in the right or left direction, remaining out of sight behind the screen once having passed the gap. The animal was free to move behind either end of the screen during one minute after the bowl passed the gap. The question was whether the animal would walk around the screen on the side of the food bowl. The number of trials each animal responded and the side to which they responded was recorded. Each animal was tested two times within one week and were given 4–8 trials during each test, with the food bowl moving alternately in both directions from trial to trial.

As can be seen in Table 1, among the three bird species, the pigeons responded only twice in 27 trials, the chickens (hens) responded in slightly over one-half of the trials but their choice of sides was at a chance level, and the crows responded at a high rate (85.7 percent) and chose the correct side 71.4 percent of the time. Among the three mammalian species, the rabbits responded on only 27 percent of the trials (9) but they were correct on eight of the nine trials. The cats showed a high level of response but their choice was at a chance level (21 times to the side with food and 15 times to the empty side). The dogs responded on almost 90 percent of the trials and their correct performance (77.8 percent) was somewhat better than the crows (71.4 percent).

Psychologically speaking, Krushinsky's task involves at least three components: object permanence (the animal must be aware that an object still exists even though it has passed from view), logical inference

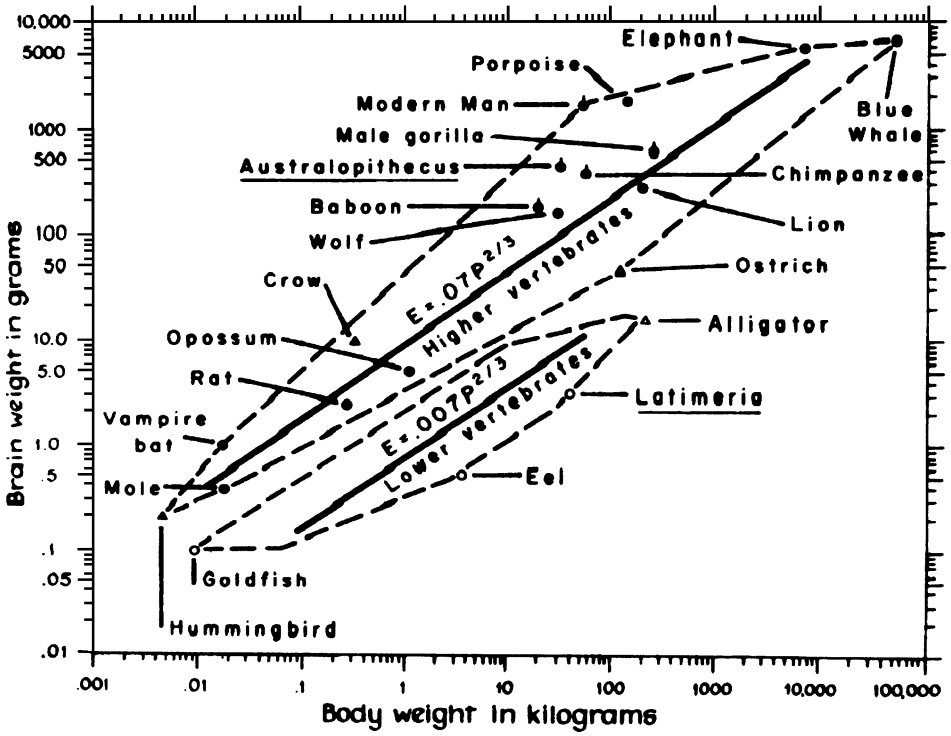

Figure 2
Brain–body ratios of birds and mammals ('higher vertebrates') vs. fish, amphibians, reptiles ('lower vertebrates') (from Jerison 1969. Reproduced with permission of University of Chicago Press).

(it must be on this side rather than that side, depending on the bowl's direction of movement), and detour learning (the animal must be able to move around an obstacle without being able to see the food). There is a great deal of variation among avian species in a structure of the brain called the wulst. The high level of performance of the birds in the crow family was attributed by Krushinsky to the high level of differentiation of their wulst in comparison to chickens and pigeons, the latter having the least developed wulst. The wulst is believed to be homologous (i.e., it is the 'same' organ in a different species) to the neocortex in mammals. Given the poor performance of the rabbits, if Krushinsky's reasoning is correct, one would predict a low level of neocortical development in rabbits compared with cats and dogs, and, to carry the comparison one step further, a difference between cats and dogs in favor of dogs. It is not known whether these predictions are correct or not.

The anagenesis of the brain that does not have precise parallels in psychological function and behavior is the increase in the ratio of brain size to body size that has occurred in the vertebrate phylum (fish, amphibians, reptiles, birds, and mammals). As shown in Fig. 2, the size of the brain of birds and mammals exceeds that of all other vertebrates at all body sizes. This rather gross statistic would suggest important psychological differences between birds and mammals and the other three classes of vertebrates, and has given some validity to the division of the phylum into 'higher' (birds, mammals) and 'lower' (reptiles, amphibians, fish) vertebrates. In the terms already developed in this article, one should expect to see, by and large, greater behavioral plasticity in birds and mammals than in lower vertebrates. As it happens, that prediction does accord rather well with learning ability, especially of a species' ability to show forms of learning above the level of conditioning: sensory preconditioning and learned stimulus configuring are possible only in birds and mammals, according to Razran's (1971) extensive review of the literature. In terms of exploratory behavior in the face of novel objects, in a study of over 300 animals in over 100 different species, there was a consistent finding of mammalian superiority over the other vertebrate species (Glickman and Sroges 1966). Finally, given the presumptive relationship between large brains and behavioral plasticity, species with large brains showing a greater degree of exploratory behavior, may be evidence of a faster evolutionary pace in these species than in species with smaller brains. That expectation accords rather well with the findings of Wyles et al. (1983) of an

Table 2
Brain size in relation to rate of anatomical evolution.

Taxonomic group	Relative brain size	Anatomical rate
Homo	114	> 10
Hominoids	26	2.5
Songbirds	23	1.6
Other mammals	12	0.7
Other birds	4.3	0.7
Lizards	1.2	0.25
Frogs	0.9	0.23
Salamanders	0.8	0.26

From Wyles et al. (1983). (Reproduced by permission of Professor J. G. Kunkel.)

almost perfect correlation between relative brain size and rate of anatomical evolution for a large number of vertebrate species.

As shown in Table 2, humans, the group with the largest relative brain size (Fig. 2), show the fastest rate of anatomical evolution, with the larger and older hominoid groups (human-like forms) ranking second in brain size and rate of evolution. What is perhaps most interesting is that the relatively recently evolved songbirds rank just below hominoids and well above other mammals and other birds on both relative brain size and rate of anatomical evolution. Finally, the classifications of 'other mammals' and 'other birds' show a relatively larger brain size than the lower vertebrates (lizards, frogs, salamanders) and a corresponding faster tempo of evolutionary change. Variations in early experience during the course of individual development also contribute in an essential way to exploratory behavior and learning ability in adulthood (Gottlieb 1992), so it is not solely the possession of a large brain that is responsible for these outcomes.

2. *Similarities and Differences in Closely Related Species: Importance of Early Experience and Context*

Humans share psychological and social characteristics with other primates. Language has always been considered the greatest difference between humans and other primates. Nonhuman primates do not share our vocal apparatus and thus cannot speak, so, in order to study linguistic competence in our relatives, manual sign language or tokens have been used. There are three components to language: syntax (grammar), semantics (meaning), and pragmatics (using language to attain goals). There is little doubt that chimpanzees in particular show evidence of the latter two and some primitive understanding of grammar (e.g., changing word order to alter the meaning of a phrase or sentence). What has hampered a genuine testing of the limits of chimpanzees' and other nonhuman primates' linguistic competence has been the age of the subjects when language instruction was started. Adults are inferior in language learning to juveniles, and it is only quite recently that a bonobo (pygmy chimpanzee) has been started in infancy and the as yet unpublished results are said to represent quite an improvement over previous attempts (Savage-Rumbaugh et al. 1998).

The importance of the presence and absence of early learning or experiential opportunities has not always been appreciated in comparative psychology, although the otherwise influential scientist T. C. Schneirla advocated the study of individual development as 'the backbone of comparative psychology.' An example of the influence of different developmental opportunities for learning is the use of pointing in chimpanzees as a communicative gesture to indicate an object of interest to a conspecific partner or human caretaker. Language-trained chimpanzees point with their index finger, as do humans aged 13–18 months. Language-naive chimpanzees point with their whole hand, as do humans aged 8–12 months (Leavens and Hopkins 1999). So developmental experience with linguistic training fosters pointing with the index finger in nonhuman primates and it parallels the beginning use of words in the older human infant (pointing frequently accompanies naming in human infants and their primary caretakers).

The importance of development and rearing context in comparing closely related species is shown in studies of domesticated animals compared with their wild (free-living) ancestors. Such studies uniformly show a reduction of aggressiveness, an increase in social tolerance, and a decrease in the domestic organism's physiological stress axes (Kunzl and Sachser 1999). These are inevitable accommodations to human-made housing and living conditions. The importance of environmental context is shown in a study by D. B. Miller (1977) in which he compared the courtship patterns of domestic and wild mallard ducks. In the barnyard situation, domestic mallards do not show courtship patterns and it was believed that these patterns were lost due to 'degenerate' genetic changes presumed to be inherent in the domestication process. Miller discovered that, when placed on a farm pond, domestic mallard not only showed all the courtship patterns of their wild progenitors but also showed them in the same order: the courtship and mating patterns were undistinguishable between the wild and domestic birds when the more natural species-typical context was provided.

3. *The Importance of Field Studies of Animals Living in their Usual Habitat*

While some comparative psychologists have done (and do) field studies, most comparative psychology was (and is) studied in the laboratory. Laboratory analysis

is more precisely controllable than fieldwork, and so it is a necessary adjunct to field observations. For a period of time in the mid-twentieth century, much of comparative psychology lost touch with natural ecological settings and, consequently, the relevance of the laboratory findings to a species' adaptation in its usual abode became dubious. (There was also an increased tendency to study a narrow range of readily available domesticated species on a narrow range of problems dictated by general theories of behavior that held sway for too many years in experimental psychology.) It is largely to the credit of the zoologically trained European ethologists that ecological relevance and the study of animals as they exist in nature became focal points. The observations in naturalistic settings provide the problems (questions) to be answered by experiments in the laboratory, so the results of the laboratory experiments are ecologically valid.

The study of imprinting in waterfowl and other precocial bird species is a case in point. Imprinting was believed to be the process whereby a young bird learned its species identification. It was thought that the young birds' perceptual system was wide open (like a blank slate), so all sorts of nonspecies-typical artifacts were used in the laboratory as imprinting objects, because it was convenient and the young incubator-hatched birds would indeed approach, follow, and become attached to these visual objects. Field observations (by a comparative psychologist!) indicated that, in a number of species, the mother uttered a special call to which the young were highly attracted. As subsequently shown in laboratory experiments, even incubator-hatched ducklings and chicks were attracted to the maternal call of their species, even though they had not previously heard it, having been hatched in an incubator (Gottlieb 1971). Those field observations, and the subsequent laboratory experiments, cast serious doubt on the ecological validity of the prior experiments on visual imprinting, especially when it was shown in the laboratory that the incubator-hatched birds were more attracted to the maternal call than to the visual aspect of a hen of their species. Because of their attraction to the maternal call, in nature young birds would always follow a hen of their own species, not just any object that happened to be passing by as implied by laboratory studies of visual imprinting.

4. *Animal Models*

The question of the utility of animal models for understanding human psychological and behavioral development does not seem tenable, if by animal model we mean we are modeling identical or even highly similar human psychological, social, and behavioral phenomena. On the other hand, the validity of animal models of human genetic, physiological, and anatomical outcomes seems much more likely because the analytic techniques are available to specify the identity of the genes, physiology, and anatomy in question. In that case, it is much easier to know whether one is right or wrong about the model. The certainty that animal models are faithfully mimicking their presumed human counterparts in the arena of psychological, social, and behavioral functions is always open to question. On the other hand, the outcome of programmatic developmental experiments with animals can sometimes provide interesting and significant food for thought—hypotheses not facts—about human development. Also, insofar as they do that they can also provide food for thought about development in other animal species. So, it is at the level of generalities, not specifics, that animal experiments are useful for thinking about human development.

Animal models may not do justice to human development when they emphasize (or seem to implicate) single developmental pathways to psychological or behavioral outcomes. Human development involves equifinality (more than one pathway to a given outcome), a principle of extreme importance, especially when dealing with problem populations (psychological disorders): there is likely not only one way to intervene (or develop the disorder in the first place).

An excellent example of how comparisons of animal models and human psychological functioning can mutually enrich each other is provided by experimental studies of prenatal and early postnatal experience in the development of intersensory perception in human and animal infants (Lickliter and Bahrick 2000). By taking a comparative approach, these authors were able to show how prenatal factors, brain–behavior relations, and context and experience contribute to infants' intersensory responsiveness. Another area in which human and animal comparisons has been mutually rewarding is the study of the role of early auditory experience in vocal development in young songbirds and human infants (Heaton and Brauth 1999).

See also: Animal Cognition; Autonomic Classical and Operant Conditioning; Behavior, Hierarchical Organization of; Body, Evolution of; Brain, Evolution of; Clinical Psychology: Animal Models; Cognitive Psychology: History; Cognitive Styles and Learning Styles; Comparative Method, in Evolutionary Studies; Conditioning and Habit Formation, Psychology of; Genes and Behavior: Animal Models; Human Cognition, Evolution of; Language Learning Impairment; Learning and Memory, Neural Basis of; Pavlov, Ivan Petrovich (1849–1936); Primates, Evolution of; Thorndike, Edward Lee (1874–1949)

Bibliography

Glickman S E, Sroges R W 1966 Curiosity in zoo animals. *Behaviour* **26**: 151–88

Gottlieb G 1971 *Development of Species Identification in Birds: An Inquiry into the Prenatal Determinants of Perception*. University of Chicago Press, Chicago
Gottlieb G 1992 *Individual Development and Evolution: The Genesis of Novel Behavior*. Oxford University Press, New York
Heaton J T, Brauth S E 1999 Effects of deafening on the development of nestling and juvenile vocalizations in budgerigars (*Melopsittacus undulatus*). *Journal of Comparative Psychology* **113**: 314–20
Hebb D O 1946 On the nature of fear. *Psychological Review* **53**: 250–75
Jerison H J 1969 Brain evolution and dinosaur brains. *American Naturalist* **103**: 575–88
Krushinsky L V 1965 Solution of elementary logical problems by animals on the basis of extrapolation. *Progress in Brain Research* **17**: 280–308
Kunzl C, Sachser N 1999 The behavioral endocrinology of domestication: A comparison between the domestic guinea pig (*Cavia aperea f. porcellus*) and its wild ancestor, the cavy (*Cavia aperea*). *Hormonal Behavior* **35**: 28–37
Leavens D A, Hopkins W D 1999 The whole-hand point: The structure and function of pointing from a comparative perspective. *Journal of Comparative Psychology* **113**: 417–25
Lickliter R, Bahrick L E 2000 The development of infant intersensory perception: Advantages of a comparative convergent-operations approach. *Psychological Bulletin* **126**: 260–80
Miller D B 1977 Social displays of mallard ducks (*Anas platyrhynchos*): Effects of domestication. *Journal of Comparative and Physiological Psychology* **91**: 221–32
Razran G 1971 *Mind in Evolution: An East–West Synthesis of Learned Behavior and Cognition*. Houghton Mifflin, Boston
Savage-Rumbaugh S E, Shanker S G, Taylor T J 1998 *Apes, Language, and the Human Mind*. Oxford University Press, New York
Schneirla T C 1953 Modifiability in insect behavior. In: Roeder K D (ed.) *Insect Physiology*. Wiley, New York
Schneirla T C 1962 Psychology, comparative. In: *Encyclopaedia Britannica*, 690Q–703
Sluckin W 1965 *Imprinting and Early Learning*. Aldine, Chicago
Szentisi A, Jermy T 1989 The role of experience in host plant choice by phytophagous insects. In: Bernays E A (ed.) *Insect–Plant Interactions*. CRC Press, Boca Raton FL, Vol. 2, pp. 39–74
Wyles J S, Kunkel J G, Wilson A C 1983 Birds, behavior, and anatomical evolution. *Proceedings of the US National Academy of Sciences* **80**: 4394–7
Zajonc R B 1971 Attraction, affiliation, and attachment. In: Eisenberg J F, Dillon W S (eds.) *Man and Beast: Comparative Social Behavior*. Smithsonian Institution Press, Washington, DC, pp. 143–79

G. Gottlieb

Comparative Research in Education: IEA Studies

Comparative research in education relates to the field of comparative education and the subfield of international education, but differs from them in so far as it is not concerned directly with teaching and training or the advancement of programs in developing countries. It has been strengthened by the movements that have occurred during the second half of the twentieth century for Westernization and modernization and more recently for globalization and it feeds back information to support these movements, but is independent both historically and structurally from them. This entry considers the formation of the field of comparative education, the place of comparative research in education with respect to that field, the theories and methods of comparative research in education, and the types of research studies that have been undertaken with a comparative and cross-national perspective. As an example of this body of research and scholarly inquiry, the studies conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) are considered from both theoretical and methodological perspectives and the key findings of these studies are listed and briefly discussed. The concluding section of this article addresses the issues arising in the future development of comparative research in education.

1. *The Field of Comparative Education*

The field of comparative education has identified as its founding father, Marc-Antoine Jullien, whose seminal study *l'Esquisse et vue préliminaire d'un ouvrage sur l'éducation comparée*, which was published in 1817, emphasized the collection of documents, objective observation, thoroughness in recording, and systematic analysis. This study established a strong research perspective in comparative education that has been maintained over time. Nevertheless, the ease of travel during the latter half of the twentieth century, together with greatly increased opportunities for communication, has led sometimes to superficiality rather than scholarly work in which comparisons are systematically examined, with sound theoretical foundations.

Leadership in comparative research in education has come traditionally from outside the USA, although the World Bank which is based in the USA has more recently supported the collection of information on the provision of education across developed and developing countries. Moreover, nongovernmental foundations as well as governmental agencies in the USA have supported financially the conduct of cross-national research studies. UNESCO, as a central organization, has sought directly to collect information on the provision of educational services across the world and its institutes, the International Institute for Educational Planning (IIEP) in Paris, the International Bureau of Education (IBE) in Geneva, and the UNESCO Institute for Education (UIE) in Hamburg have maintained for over half a century a research orientation in their studies of education across both the developed and developing world. Likewise, the

Centre for Educational Research and Innovation (CERI), which is established within the Organization for Economic Cooperation and Development (OECD) in Paris, has examined educational issues within and between the developed countries that are members of OECD. In the closing decade of the twentieth century, OECD sought to provide a stronger database through the calculation of educational indicators for use in its member countries and its publications 'Education at a Glance': OECD indicators have provided a valuable start in the provision of soundly based information.

Two publications from UNESCO, *Learning To Be: The World of Education Today and Tomorrow* (Faure et al. 1972) and *Learning: The Treasure Within* (Delors et al. 1996) have provided and are providing a basis for debate both within and between nations on the future of education from a world perspective. These publications emphasize the idea of lifelong learning in a world context that includes both developed and developing countries in which the seven tensions:

(a) between the global and the local,
(b) between the universal and the individual,
(c) between tradition and modernity,
(d) between long-term and short-term considerations,
(e) between the need for competition and concern for equality of opportunity,
(f) between the extraordinary expansion of knowledge and human beings' capacity to assimilate it,
(g) between the spiritual and the material

are very conspicuous in a world in which globalization is taking place rapidly (Delors et al. 1996). The monitoring of change in education across the world during the critical years of the first half of the twenty-first century is a major challenge for comparative research in education. Only with soundly based information on the magnitude and direction of change that could be provided through strong and well-recognized indicators can policy making and planning be carried out in all corners of the world to avoid the major catastrophies that could be caused by nuclear war, a population explosion, degradation of the biosphere, and widespread death from disease and famine.

2. *Comparative Methods in Research*

Research and scholarly inquiry in education uses two different strategies of investigation. One strategy is a holistic or systemic approach in which the characteristics or the parts are viewed with respect to the whole to which they belong. Within this strategy the pattern of relationships between the parts and the processes operating within the whole are examined. The issues involved in inquiry are whether an elementary part or characteristic can be considered to exist and whether a model of the whole built from the parts is an adequate representation of reality. The second strategy is a comparative approach in which the characteristics or the parts are compared across two or more research situations. If a difference is found between the two situations, with an adequate degree of likelihood or probability, then the origin or cause of the difference is sought. Moreover, if a difference is not found, then the reasons for the similarity may be of interest to the researcher. The comparative method is primarily concerned with the second strategy. However, the establishment of the existence of a characteristic, a part, or a relationship must precede the making of a comparison, although it might be argued that the detection of a difference through comparison can be considered to be evidence for the existence of an entity, a characteristic, or relationship.

Comparative research in education may be undertaken from either a scientific or humanistic perspective, where the former involves measurement and statistical analysis and the latter is largely descriptive or ethnographic. Furthermore, comparative research in education may draw on one or more of the major disciplinary fields, such as history, philosophy, sociology, and social psychology. There are, however, four specific ideas regarding the methods and nature of comparative education that must be considered to be fallacious.

First, there is the fallacy advanced by Holmes (1965) in his study, *Problems in Education: A Comparative Approach*, that there is a single scientific method for comparative research in education. He drew on the writings of John Dewey and advanced a sequence of eight major steps as a method of investigation that led to the production of new knowledge in education. While such systematic inquiry may be desirable in certain problem-solving studies, research in comparative education draws on more than one disciplinary field and their methods of investigation and cannot be constrained to a particular sequence of steps in inquiry.

Second, there is the fallacy that investigation in comparative research in education involves a choice between a quantitative or a qualitative approach or even a combination of the two approaches (Rust et al. 1999). The distinction between quantities and qualities is misconceived, when they are taken to be clear alternatives. Complex statistical procedures are now available for the rigorous examination of categorical or qualitative data, where information is recorded in terms of two or more categories that does not involve any attempt to measure, but merely employs classification. The advantage of measurement is that the improved precision increases the likelihood of detecting or rejecting a hypothesised relationship.

Third, there is the fallacy of making a clear-cut distinction between policy-oriented research and discipline-oriented research, or between decision-based research and conclusion-based research in comparative education. Holmes (1965) recognized that comparative education provided a strategy of inquiry that

led to a greater understanding of educational systems as well as providing information for the planned development of an educational system. However, the undue emphasis shown by politicians and administrators on the 'horse race' aspects of international studies of educational achievement, to the neglect of developing a greater understanding of the operation of the systems through comparative research, would seem to involve a serious misunderstanding of the nature of social action and the power of ideas. It assumes that research findings in the social and behavioral sciences, including education, are applied in the same way as the findings of research in the physical and biological sciences through technology and applied science. Consequently, there is a need for the UNESCO Institutes, The World Bank, and for the Centre for Educational Research and Innovation of the Organization for Economic Cooperation and Development to provide international forums for the discussion and debate of the findings of comparative research in education in terms that policy makers can understand and subsequently apply in both policy and practice.

Finally, there is the fallacy that comparative research in education is not concerned with the outcomes and products of education but only with the processes that are in operation. Nevertheless naive input–output models of education provide little of value for the making of policy or the improvement of practice, as does undue emphases in simplistic terms on achievement outcomes that are evident in the media presentations of the findings of some international studies of educational achievement.

3. *Paradigms and Theories in Comparative Research in Education*

Altbach (1991) has argued that in the late 1980s the theoretical frameworks employed in comparative research in education shifted away from a central role for sociology and from structural–functional and statistical paradigms towards greater heterogeneity. Paulston (1994) has more recently identified four world views in a taxonomy of knowledge perspectives in scholarly writings on comparative and international education. These four world views are (a) functionalist, (b) radical functionalist, (c) radical humanist, and (d) humanist perspectives. These four perspectives have been mapped by Paulston in a two-dimensional space with the many general theories involved in comparative and international education located within this two-dimensional map. The macromapping of paradigms and theories makes clear the multiplicity of general perspectives that, at the end of the 1990s, were present in the field of comparative and international education. Through the use of this map the broader movements that influence scholarly work in the field of comparative education are presented in a way that indicates the dynamic but adventitious nature of the field at a time when issues of globalization and the cross-national transfer of ideas and knowledge are being greatly facilitated by the new information technology. The challenge for comparative social theory is to make sense of the rapid changes that are occurring in the late 1990s and early 2000s.

4. *The Many Methods and Approaches of Comparative Research in Education*

There can be no argument about the fact that there are many different tactics and strategies being employed in comparative studies in education. Each of these different approaches has considerable utility in a field of inquiry that is actively pursued around the world, particularly along economic, social, and political dimensions. Of particular importance are the contributions that comparative research in education can make to the economic development, social development, and political development of a particular country, especially in those countries that are classified as developing countries.

In addition to those studies that might be classified as theory building and that address issues related to the structure of both knowledge and inquiry in the field and are sometimes described as theoretical or conceptual studies, there are studies that employ methods of inquiry derived from particular disciplines. These disciplinary studies involve historical, demographic, linguistic, economic, political, or other perspectives as they relate to education. There are, however, comparative research studies in education that extend across disciplines and that would seem to fall into one of three types: (a) descriptive studies, (b) developmental studies, and (c) analytical, relationship, and process studies.

4.1 *Descriptive Studies*

An important task in comparative education is to provide a descriptive account of how education is conducted in different countries, with generally only one country under survey in each study. However, the national systems of education of more than 130 countries are presented in the *International Encyclopedia of Education* (Husén and Postlethwaite 1994) and the derived *Encyclopedia of National Systems of Education* (Postlethwaite 1996) in a way in which education in each of the different countries is described within a common framework. Such descriptive studies serve an important purpose since they provide relatively concise accounts of educational policies and practices in each particular country and thus assist in

the building up of an understanding both within that country and elsewhere of the different systems of education that have been developed in different contexts across the world.

The comparative value of these descriptive statements is limited by the existence of many problems. The major problems involve: (a) the lack of a list of key educational indicators which have been agreed on by scholars and administrators in all countries, so that accurate and consistent data are collected; (b) the cost and practical difficulties of assembling data from many sources; (c) the lack of comparability of the data collected, in spite of efforts to maintain uniformity; (d) the tendency of ministry officials in each country to present the conditions of education within their country in the best possible light and as a consequence to withold data that are unfavorable; and (e) insufficient analysis and presentation of context without which the data recorded have only limited meaning.

4.2 Developmental Studies

The influence of the social and behavioral sciences is effected not through technology and direct application, but through the power of the ideas and relationships that arise from research, which are disseminated and debated and which give rise to collective social action. The contribution of education to this process of promulgation and discussion of ideas and relationships is greatly underestimated, and the nature of social action in democratic societies across the world is only gradually being understood. As a consequence, many of the societal developments that have occurred in different countries of the world and that have their origins in education still remain to be examined and reported.

Some attempts have been made to account for both changes in education and social life. Hans (1949), for instance, argued for a level of generality with the identification of factors influencing educational development: namely, (a) religious factors (Anglican, Catholic, and Puritan); (b) natural factors (racial, geographic, linguistic, and economic); and (c) secular factors (socialism, humanism, nationalism, and democratic ideals). Nevertheless, these factors do not provide an explanation in terms of the common or universal factors that have influenced the spread of educational provision across the world, and would seem to involve the unique factors which might explain why countries now differ with respect to educational provision in particular ways. Mallinson (1957) suggested that educational development within a country was influenced by a national mental constitution that had formed over time within the country. While this might help to account for particular features of German and Japanese education it would not seem to explain the marked growth in education that has taken place in so many countries since the latter half of the nineteenth century.

Meyer and Hannan (1979) have contended that the thrust for universal primary education must be seen in a cross-national context, and that the powerful ideas are to be found in the Darwinian theory of evolution, and the impact of Marx's ideas, particularly in China and the USSR. There is much more work to be done to explain the common developments in education that have occurred cross-nationally during the twentieth century.

4.3 Studies of Relationships and Processes

The systematic study of relationships and processes in comparative research in education has emerged only during the second half of the twentieth century. These studies seek to increase understanding of how variables relate to one another, to explain better the dynamics of teaching and learning and educational change, and to investigate how such relationships and processes vary across countries. Nevertheless, cross-national studies of educational achievement are difficult and costly to conduct because extensive cross-national consultation is required. The International Association for the Evaluation of Educational Achievement (IEA) has been the major pioneer of this type of investigation into the factors influencing achievement outcomes across the countries of the world.

5. The Research Program of IEA

In the late 1950s a group of educational research workers meeting at the UNESCO Institute for Education in Hamburg saw the need for comparative studies of the relationships and processes operating to influence the educational outcomes of achievement and attitudes. They recognized that there was generally insufficient variability within a particular country to detect the effects of specific factors on the outcomes of learning, using existing methods of statistical analysis. They argued that if research studies could be undertaken between countries, as well as within countries, then the greater variation that existed across countries and the different school systems within countries would permit the identification of common factors of importance which influenced learning outcomes.

The feasibility of conducting research studies across countries into factors influencing educational achievement was tested in a pilot study. The success of this study led to the establishment of the IEA. Over a period of nearly 40 years IEA has carried out a sustained program of research studies, which have in the main been survey and cross-national investiga-

tions, into factors that have an effect on educational achievement, participation rates, and attitudes towards school and school learning. The data collected have been examined across countries, between schools within countries, and between students within countries, and where possible relationships have been tested after controlling for other factors that might affect the educational outcomes under consideration. The specific relationships that have been tested fall into one of several types, namely: (a) relationships between predictor and criterion variables across countries; (b) similar common relationships between predictor and criterion variables at the levels of between students and schools within countries; (c) relationships between predictor and criterion variables at the levels of between students and between schools that are unique to a particular country or a small group of countries; and (d) the absence of an observable relationship where current opinion has hypothesized that such a relationship should exist.

Before the systematic examination of processes between students, between schools, and between countries can take place advances must be made along three fronts. First, it is necessary for theory to be developed and models formulated of specific processes of school learning. Second, it is necessary for data to be collected that are sufficiently comparable across countries, so that the testing of models in a comparative manner is worthwhile. Third, statistical procedures for the testing of models must be developed in order to provide strong estimates of effects and efficient estimates of error for the examination of statistical significance and effect size.

Several theoretical frameworks have guided the development of models in IEA research studies. From the Gränna Workshop conducted in Sweden by IEA in 1971, which examined in detail the seminal work *Handbook on Formative and Summative Evaluation of Student Learning* (Bloom et al. 1971) came the theoretical view of curriculum implementation that has been tested in several IEA studies. In this model the curriculum can be considered to exist at three levels: (a) the intended curriculum, (b) the implemented curriculum, and (c) the achieved curriculum, which are influenced by the antecedent and contextual factors as well as the learning conditions in the schools and teaching practices. A second highly influential model in IEA research has been Carroll's (1963) model of school learning, which was tested empirically with respect to the study of learning French as a foreign language, and has formed the basis for causal models in several IEA studies. A third model that has guided IEA research has been a cross-national model of educational achievement in a national economy advanced by Dahllöf (1967). Of particular interest for policy making are the frame or structural variables. Moreover, their dependence on (a) the environment and the economy, (b) demand for personnel, (c) curriculum content, and (d) the objectives of the school system, is argued on theoretical grounds by Dahllöf. However, only limited aspects of this theoretical model have been tested in either primary or secondary analyses.

IEA studies have consistently recorded the marked differences in levels of achievement between developed and developing countries, with much lesser variation between the countries within each group. Furthermore, it has been shown that a causal model, which explained well the influence of individual student factors on reading achievement at the 14-year-old level in developed countries, was less than adequate in two developing countries, India and Iran. These findings suggest that there are different processes in operation influencing educational outcomes in developed and developing countries.

As statistical procedures for the testing of models of causal processes have been developed, IEA has remained at the forefront of statistical analysis and has advanced and tested complex models to generate an understanding of the processes that operate to influence educational outcomes. It is noteworthy that such analytical procedures as blockwise regression analysis, partial least squares path analysis, and linear structural relations analysis were developed in collaboration with scholars at the Institute of International Education in Stockholm, who were working on the analysis of IEA data. Likewise, the advancement of procedures for the improvement of the measurement of educational outcomes in large-scale surveys, with equating across levels of test administration and over time owes much to the work of IEA.

6. *The Key Findings of the IEA Studies*

The IEA studies taken together have involved millions of students in testing programs, hundreds of thousands of schools, and thousands of research workers in the field of education from a very wide range of countries. In 1999, over 50 countries were members of IEA and its headquarters remained firmly established in Europe, although substantial financial support has been received from time to time but not exclusively from agencies and foundations in the USA. The school subjects involved in IEA testing programs have included mathematics, science, civic education, reading, French as a foreign language, English as a foreign language, computing, literature, and writing. In addition, there have been studies in preschool education.

There are marked differences in average levels of achievement between the students in school in the more developed countries (MDCs) and those in the less developed countries (LDCs). This occurs in spite of the fact that in the LDCs less than 100 per cent of the relevant age groups are enrolled in school.

The average level of achievement within a country at the terminal secondary school stage is inversely related to the proportion of the age group enrolled at

school or participating in the study of the subject under survey.

At the terminal secondary school stage, when equal proportions of the age groups are compared, there are only small differences in levels of achievement, irrespective of the proportions of the age groups retained at school. The best students do not suffer with increased retention rates.

Student achievement in mathematics, French as a foreign language, and science is positively related to the time given to the study of the subject at school, both in comparisons across countries and between students within countries.

The achievement of students is related to the time spent on homework after other factors influencing achievement have been taken into account.

The average level of student achievement across countries is positively related to the opportunity that the students have to learn the content of the items tested.

In LDCs the use of a textbook has an effect on student learning. However, the same effects have not been reported from studies in MDCs.

The level of reading resources of the home is positively related to student achievement, as are other indicators of language usage in the home, such as the use of a dictionary in the home, and whether or not the language of the home is the language of instruction.

Measures of the socioeconomic status of the home are positively related to student achievement in all countries, at all age levels, and for all subjects.

Although the effects of home background variables are similar across subject areas, the effects of the learning conditions in the schools differ between subject areas, and in some subject areas are equivalent to or greater in the size of their influence than the effects of the home.

For more detailed accounts of these 10 key findings from IEA studies and the more specific findings from particular studies, the report by Keeves (1995) should be consulted. Nevertheless, it is important that the vision of those who founded IEA for the advancement of research in a cross-national context is sustained at a time when the increased visibility of international educational achievement studies gives rise to serious misunderstandings, disputable conclusions for policy and practice, and undue emphasis on the rank ordering of national achievement (Beaton et al. 1999).

7. *Research in Comparative Education in the Future*

From an examination of the Delors report (Delors 1996), seven critical world issues emerge for which education would seem to provide the major contribution to their resolution. These problems involve:

(a) the feeding of the growing population of the world;

(b) the containment of population growth;

(c) the sharing of the world's wealth between the developed and the developing countries;

(d) the advancement of technological development in the poorer countries in order to raise standards of living;

(e) the building of schools and the training of teachers to achieve universal primary schooling;

(f) the establishment of greater equity between the sexes; and

(g) the establishment of greater equity between different social and racial groups.

Not only must comparative education help to monitor change with respect to these problems, but it must also play an active part in their resolution through the spread of knowledge. The rapid expansion of knowledge during the twentieth century demands the worldwide development of the learning society. The new information technology opens up new ways in which the learning society might operate. The field of comparative education provides the links across the countries of the world, but the exact manner in which comparative education through research might contribute is unclear. Nevertheless, it is necessary to recognize that the world's problems involve not just the wealthy and more highly developed nations, but all nations irrespective of size and wealth.

See also: Comparative Studies: Method and Design; Educational Systems: Asia; Educational Systems: Australia; Educational Systems: Europe; Educational Systems: North America

Bibliography

Altbach P G 1991 Trends in comparative education. *Comparative Education Review* **35**(3): 491–507

Beaton A E, Postlethwaite T N, Ross K N, Spearritt D, Wolf R M 1999 *The Benefits and Limitations of International Achievement Studies*. International Academy of Education and IIEP, Paris

Bloom B S, Hastings J T, Madaus G F 1971 *Handbook on Formative and Summative Evaluation of Student Learning*. McGraw-Hill, New York

Carroll J B 1963 A model of school learning. *Teachers College Record* **64**(8): 723–33

Dahllöf U 1967 Relevance and fitness analysis in comparative education. In: Super D E (ed.) 1967 *Toward a Cross-National Model of Educational Achievement in a National Economy*. Teachers College Press, New York

Delors J et al. 1996 *Learning: The Treasure Within*. Unesco, Paris

Faure E et al. 1972 *Learning To Be: The World of Education Today and Tomorrow*. Unesco, Paris

Hans N 1949 *Comparative Education: A Study of Educational Factors and Traditions*. Routledge and Kegan Paul, London

Holmes B 1965 *Problems in Education: A Comparative Approach*. Routledge and Kegan Paul, London

Husén T, Postlethwaite T N (eds.) 1994 *The International Encyclopedia of Education*. Pergamon, Oxford, UK

Keeves J P 1995 *The World of School Learning: Selected Key Findings from 35 years of IEA Research*. IEA, The Hague, the Netherlands
Mallinson V 1957 *An Introduction to the Study of Comparative Education*. Heinemann, Melbourne, Vic
Meyer J W, Hannan M T 1979 *National Development and the World System: Educational, Economic, and Political Change*. University of Chicago Press, Chicago
Paulston R G 1994 Comparative and international education: Paradigms and theories. In: Husén T, Postlethwaite T N (eds.) *International Encyclopedia of Education*, 2nd edn. Pergamon, Oxford, UK, pp. 923–33
Postlethwaite T N (ed.) 1996 *International Encyclopedia of National Systems of Education*, 2nd edn. Pergamon, Oxford, UK
Rust V D et al. 1999 Research strategies in comparative education. *Comparative Education Review* **43**(1): 86–109

J. P. Keeves

Comparative Studies: Method and Design

1. *Definition*

Comparisons are essential in any science to establish systematic similarities and differences between observed phenomena and, possibly, to develop and test hypotheses and theories about their causal relationships. Whereas this broad definition of comparative inquiry comprises several distinct methods, including case studies, statistical analysis, and experimental research, the comparative method in the narrower sense of the term characteristically involves a limited number of cases ('small *N*'), but usually very complex phenomena and relationships often at the 'macro'-level of entire societies or political systems. In this latter sense this method and its respective designs will be discussed here.

2. *Intellectual Context*

Comparative procedures in the historical and political sciences date back to antiquity, but their methodological backgrounds have only been developed systematically in the natural sciences in the eighteenth and nineteenth centuries based on studies, for example, by Linné in botany, Cuvier in anatomy, or Bernard in experimental medicine. The logical foundations for this method have been laid by Hume and, in particular, J. S. Mill's (1843) 'canons.' The first of these refers to the 'Method of Agreement' eliminating all similarities but one: 'If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon' (p. 390). By contrast, the 'Method of Difference' establishes the absence of a common cause or effect, even if all other circumstances are identical: 'If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon' (p. 391).

Both methods thus are concerned with the systematic matching and contrasting of cases in order to establish common causal relationships by eliminating all other possibilities. Both procedures are, however, somewhat extreme in the sense that they attempt to establish a single common cause, or its absence, by controlling all other possibilities and the entire environment. The method of difference can, to some extent, be approximated in 'pure' laboratory situations where such experiments may be conducted. The method of agreement is more applicable in 'field research' situations, but then the question of how to control the 'environment' becomes crucial (see *Field Observational Research in Anthropology and Sociology*).

Mill, therefore, devised a combination of the two which he called the 'Joint Method of Agreement and Difference' or the 'Indirect Method of Difference' which consists of a double employment of the Method of Agreement, once before and once after a certain event: 'If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance, the circumstance in which alone the two sets of instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon' (Mill 1974, p. 396). This 'quasi-experimental' design is, however, as Mill himself realized, less cogent than the pure Method of Difference.

A fourth method, which is a derivation of the Method of Difference, is the 'Method of Residues'. When cause and effect of other factors have been established by previous experiments, remaining variation in the outcome can be attributed to the 'residues' in the original experiments: 'Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents' (Mill 1974, p. 398).

Finally, for instances where a certain common external factor cannot be entirely eliminated by any of these procedures, its impact may at least be systematically varied from one situation to another (e.g., altitude above sea level, or heat or air pressure in certain experiments in physics) and then its varying influence assessed. This is the 'Method of Concomitant Variation': 'Whatever phenomenon varies in any manner whenever another phenomenon varies in some

particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation' (Mill 1974, p. 401).

3. Changes in Emphasis

Mill's 'canons' imply rather rigid 'positivist' assumptions about the relationships of cause and effect and the state of valid theory in any given area of research. These are: (a) contiguity between the presumed cause and effect; (b) temporal precedence, in that the cause had to precede the effect in time; and (c) constant conjunction, in that the cause had to be present whenever the effect was obtained (cf. Cook and Campbell 1979, p. 10). On the whole, such relatively mechanical and deterministic relationships can only rarely be established even in the 'hard' sciences.

By themselves, therefore, these methods do not produce any new *discoveries* unless some truly relevant factors have been included. Similarly, they may not *prove* any causal relationship unless a clear and complete (preconceived) model of such links has been tested and other factors were sufficiently 'controlled.' But even such a model may eventually be falsified and be integrated in a more comprehensive theory. So these methods may not produce any 'positive' results. They constitute, however, a valuable step to eliminate many irrelevant factors and to approximate the causal conditions in the 'real' world. In this sense they correspond to Popper's (1934) famous principle of 'falsification.' Or as it was expressed in another 'classic' of this period, Mill's methods are nevertheless '… of undoubted value in the process of attaining truth. For in eliminating false hypotheses they narrow the field in which true ones may be found. And even where these methods may fail to eliminate all irrelevant circumstances, they enable us with some degree of approximation to so establish the conditions for the occurrence of a phenomenon, that we can say one hypothesis is logically preferable to its rivals' (Cohen and Nagel 1934, p. 267).

If these qualifications have to be made from a general epistemological point of view, they apply even more to the social and behavioral sciences where 'objective' facts and their 'subjective' perceptions are closely interlinked and may change in the course of time. Mill himself was aware of the much higher complexity in the social sciences and the inapplicability of his methods in a very strict sense: in 'politics and history … Plurality of Causes exists in almost boundless excess, and effects are, for the most part, inextricably interwoven with one another. To add to the embarrassment, most of the inquiries in political science relate to the production of effects of a most comprehensive description, such as the public wealth, public security, public morality, and the like; results liable to be affected directly or indirectly either in *plus* or in *minus* by nearly every fact which exists, or event which occurs, in human society' (Mill 1974, p. 452, emphases in the original). He, therefore, found the Method of Difference inapplicable and the Methods of Agreement and Concomitant Variation inconclusive in the social sciences (pp. 881 ff.). Similarly, the Method of Residues requires a much more developed state of theory than is usually available in the social sciences (p. 884 ff.).

However, in the realm of the human and social sciences, Mill was willing to adopt a less deterministic and more probabilistic perspective:

> Inasmuch as many of those effects which it is of most importance to render amenable to human foresight and control are determined, like the tides, in an incomparably greater degree by general causes, than by all partial causes taken together; depending in the main on those circumstances and qualities which are common to all mankind, or at least to large bodies of them, and only in a small degree on the idiosyncracies of organisation or the peculiar History of individuals; it is evidently possible with regard to all such effects, to make predictions which will *almost* always be verified, and general propositions which are almost true. And whenever it is sufficient to know how the great majority of the human race, or of some nation or class of persons, will think, feel, and act, these propositions are equivalent to universal ones. For the purpose of political and social science this *is* sufficient (Mill 1974, p. 847, emphases in the original).

4. Emphases in Current Theory and Research

Two basic problems remain in any research of this kind, namely the questions as to what have been called the 'internal' and 'external' validity of any scientific findings (see also *Internal Validity; Causal Counterfactuals in Social Science Research*). The first refers to the 'approximate validity with which we infer a relationship between two variables is causal,' the second to the 'approximate validity with which we can infer that the presumed causal relationship can be generalised to and across alternate measures of the cause and effect as well as across different types of persons, settings and times' (Cook and Campbell 1979, p. 37). Even though both kinds of validity can only be approximated and remain subject to the possibility of further falsification, they can be established with some kind of certainty with the help of procedures of randomization.

As far as internal validity is concerned, Mill's original procedures can be enhanced by first randomly selecting two groups of subjects and then administering a certain stimulus or treatment to one of them. If a significant effect can be observed in this group, then the stimulus may be considered to have been the 'cause.' This corresponds to Mill's method of difference, but where all kinds of external influences can now be controlled, within certain limits, by the 'equalizing' procedure of random selection. This is the

realm of 'experimental research' proper (see *Laboratory Experiment: Methodology*). Similarly, if a random sample has been drawn from a larger universe of persons, as for example in survey research, the relationships which can be observed in the sample may be generalized to the larger universe from which it was drawn. This is the realm of 'statistical' analysis (see *External Validity*).

Unfortunately, however, both procedures of randomization are not feasible when we deal with the characteristic 'many variables–small *N*' situation of many kinds of current social research apart from the fact that many 'real' experiments with groups of human beings are not possible for practical, political, or ethical reasons either. We are thus left with Mill's original predicaments and have to attempt to cope with them in some other ways. This also depends, of course, on the kinds of cases and problems we are dealing with and the state of theory in any given area of investigation. In social research, 'cases' often refer to larger distinct groups of people, e.g., certain ethnic groups or entire societies and states, which have a proper name and in which we are interested because of their intrinsic relevance and not just as an anonymously selected unit out of a multitude of cases.

The kinds of problems we are dealing with vary according to the respective emphasis of the different social and behavioral sciences. What they have in common is the 'multidimensionality' (i.e., comprising 'objective,' 'subjective,' and often also normative dimensions) and the 'malleability' (i.e., possible changes over time and probabilistic internal relationships) of their subject matter (Almond and Genco 1977). The ways such relationships are perceived depend on the 'social,' i.e., also theoretical constructions of this reality. These, by necessity, can only be certain approximations as to their 'real' (and changing!) nature, but they have concrete referents and are not, in our view, entirely arbitrary whims of imagination as some of the more extreme 'postmodern' authors pretend (see *Postmodernism: Methodology*). *What* to compare thus differs from one discipline to the other and depends on the respective state of theory.

By contrast, here we are concerned with *how* to compare under particularly difficult conditions concerning problems of internal and external validity in the social sciences. From this perspective, a number of ways can be conceived to cope with the prevailing 'small *N*–complex cases' dilemma. First of all, the number of cases may be increased, including for example historical cases or investigating them at some lower level of analysis, e.g. certain districts rather than entire nation-states. But even then neither randomly selected control groups nor random sampling for statistical purposes are possible. In any case, we usually are interested in concrete societies and their problems and not just in any 'anonymous' units of investigation. A second strategy concerns the purposeful selection of cases and a third, the reduction of complexity by means of combining several variables or employing better theory.

At the outset of any investigation an area of homogeneity must be defined which establishes boundaries for the selection of cases. Cases must 'parallel each other sufficiently' and be comparable along certain specified dimensions. This is meant by the common saying that 'apples and oranges' should not be compared. In this regard, the subject matter and the problem we are interested in must first be specified in order to make any sense. Thus, these fruits may well be compared concerning their sugar or water contents, their nutritional value, etc. when these dimensions as the *tertium comparationis* have been made explicit. The specification of relevant cases at the start of an investigation thus amounts to an explicit or implicit hypothesis that the cases selected are in fact alike enough to permit comparisons. The primary consideration in delimiting cases for a small *N* comparative study is the dependent variable. For example, the breakdown or survival of democratic regimes in interwar Europe presupposes the prior existence of some form of democracy in the selected cases. In addition, some limitations in time and space can also enhance the homogeneity and thus the comparability of the cases examined. For example, certain kinds of colonial or other forms of external domination or religious-cultural influences may be useful criteria for selecting a specific group of cases.

A second consideration concerns the extent of diversity within the selected universe. In this regard, a maximum of heterogeneity for a minimum number of cases should be achieved. Taking the above example again, both survivors and breakdowns of democracy can be considered, and among the latter perhaps some more specific variants such as fascist vs. more generally authoritarian outcomes.

Within this universe two opposite strategies now become possible. One is the 'most similar,' the other the 'most different' systems design. These have been explicitly formulated and discussed by Przeworski and Teune: 'The most similar systems design is based on a belief that a number of theoretically significant differences will be found among similar systems and that these differences can be used in explanation' (1970, p. 39). By matching these similar cases as much as possible, most of the variables can be 'controlled.' Mill's 'indirect method of difference' where different outcomes may be attributed to the remaining factors which differentiate these cases now becomes applicable. Even though only rarely just a single factor will remain to which the effect can be attributed, at least many others can be excluded, and the remaining ones can be examined more closely in a theoretically guided qualitative manner. The 'internal validity' of the observed relationships can thus be greatly enhanced.

The opposite strategy is the 'most different' systems design: this design, '… which seeks maximal het-

erogeneity in the sample of systems, is based on a belief that in spite of intersystemic differentiation, the populations will differ with regard to only a limited number of variables or relationships' (Przeworski and Teune 1970, p. 39). This 'contrasting' of cases thus eliminates all factors across the observed range which are not linked to an identical outcome. In this way, more 'universal' explanations are sought as far as the selected area of homogeneity is concerned. To some extent, thus, the 'external validity' of some hypothesized causal relationship can be extended and the range of its applicability including certain limitations in time and space can be established.

Until recently, these designs had not, however, been fully operationalized. Several problems must be addressed which follow from the necessity of measuring the proximity or remoteness of pairs of cases in the heterogeneous, multidimensional space defined by the independent variables. These distance measures provide the basis for determining the 'most different' and 'most similar' pairs or groups of cases with regard to the respective outcome. The two main issues are: (a) choosing from among a variety of different ways of measuring the distance between pairs of cases in a multidimensional space; and (b) assigning relative weights to the variables that define this space. In this way the complexity of the data can be retained as much as possible in the complexity of the proximity measure. These procedures then make it possible to identify 'most different cases with the same outcome' (MDSO) and the 'most similar cases with different outcomes' (MSDO). The causes of different outcomes may now be assumed to lie in the commonalities that remain among MDSO cases and the differences that distinguish MSDO cases (for such operationalizations see, e.g., De Meur and Berg-Schlosser 1996).

Another strategy for dealing with the 'many variables—few cases' dilemma concerns the variables side. Here, for reasons concerning the 'internal validity' of Mill's methods, one should take a rather comprehensive approach in order not to leave out any possibly relevant variables. This, of course, depends on the state of theory in any given area of investigation, but usually in the social sciences there are several contending or supplementary hypotheses which should all be considered. Furthermore, not only bivariate relationships with an independent variable and the respective outcome, but also the possibility of more complex causal problems, e.g., 'multiple' (several variables combined have a certain effect) or 'conjunctural' ones (alternative combinations of different variables may lead to the same outcome) must be taken into account.

One technique is to look for 'constants' across the observed cases. If one variable turns out to be consistently linked to a particular outcome, it becomes a strong candidate for inclusion in any explanation as a necessary but perhaps not sufficient condition. A related approach focuses on correlations between the independent variables and the dependent variable. Like the search for constants, the examination of correlations may provide strong hints. More sophisticated techniques include discriminant analysis and logistic regression (applied to dichotomous outcome variables). A relatively large number of variables can be simultaneously treated in this way. These techniques are typically used to assess the net, additive contribution of each independent variable to some outcome. Variables that are uncorrelated with other variables but strongly correlated with the outcome are favored by these methods. Thus, these techniques are biased toward causal factors that act independently, not conjuncturally.

These statistical approaches are useful for identifying broad patterns, but they are of limited utility in situations of causal complexity. Most complex causal combinations are hidden from analyses based on correlations. For example, if a variable must be present in some conjunctures for an outcome but absent in others, the correlation between this cause and the outcome may be 0. 'Qualitative Comparative Analysis' (QCA), a recent comparative technique based on Boolean algebra (cf. Ragin 1987), can be used in situations of causal complexity. In contrast to the linear techniques just described, QCA focuses on configurations of variables. In QCA, an independent variable can be eliminated from an analysis if it does not uniquely distinguish any case that manifests an outcome from at least one case lacking the outcome.

Other techniques for reducing the number of variables focus on patterns among the variables. In the light of relevant theoretical and empirical knowledge, investigators may select a more limited number of major variables or reconstruct them in different ways. Such reductions may proceed statistically. For example, researchers often use factor analysis to combine correlated variables into single indices.

Qualitative techniques for constructing more encompassing causal variables are also possible. For example, in the interwar study the existence of a rural proletariat and large-scale landlords has been combined into the more encompassing concept of 'feudalism.' Similarly, a Boolean 'addition' of factors, constituting alternative constellations in combination with another variable, may be possible. For example, the presence of ethnic, religious, *or* regional social cleavages in the absence of any overarching ('*verzuiling*') structures was combined to the variable of 'social heterogeneity' in the study of democratic breakdowns and survivals (see also Berg-Schlosser and De Meur 1997). Use of ideal typic constructs can also be conceived of as reductions of complexity emphasizing some characteristics and de-emphasizing others. In this way, the overall number of variables may be reduced considerably while still retaining much of the original information. With the help of such strategies, both from the cases and the variables sides, the relationship between their respective numbers may

be reduced to manageable proportions, thus avoiding the problem of 'overdetermination' when the latter exceed the former.

QCA can also be employed for other purposes. It is based on dichotomized variables (which, however, can be constructed as 'dummies' from multichotomous or interval ones) and reduces, by employing certain algorithms, the observed complexity of possible independent factors with regard to a particular outcome to a minimal logical formula (or several alternative ones) taking account of the possible interactions of these factors. In this way it can: (a) examine a universe of cases with regard to a given output in an exhaustive manner; this description will be the most concise, including all possible alternate minimal expressions which allow for the most parsimonious interpretations. (b) test hypotheses by showing whether they are consistent with the combinations described; if many contradictions occur, the hypothesis fails. (c) deduce the shortest possible formulas, describing the simplest subset of actual, potential, or counterfactual cases not contradicting the respective outcome; in this way, more general expressions can be deduced by using 'simplifying assumptions.' QCA thus serves some of the most important aims of any science, namely systematic description and the falsification and construction of theories.

QCA shares some of the problems and limitations concerning the loss of information and a certain amount of arbitrariness in setting cutting-points for dichotomized variables. In contrast with common statistical procedures, however, it retains the full complexity of all cases and all variables considered. Nonunique formulas may emerge, but this must be considered a strength and not a weakness. It deals explicitly with 'outliers' and covers them fully. It can arrive at several 'conjunctural' constellations which explain the same outcome. When several different formulas cover a single case, the technique forces the researcher (as the MSDO technique) to take a closer look at the results and to *interpret* them in light of the historical knowledge of the case. In comparative inquiry, this approach is preferred to a purely mechanical procedure which, in many statistical analyses, entirely obscures the fate of particular cases. Here, in fact, begins the real *qualitative* work, depending very much on the training and quality of researchers, their in-depth knowledge of cases, but also their sensitivity and understanding.

5. *Methodological Issues and Problems*

Even though the 'small *N*' comparative and the statistical methods have their own realm of possible applications, each with its own advantages and problems, different 'schools' have formed around these approaches fraught with controversies. In part, these reflect certain biases; to some extent they are based, however, on real problems which cannot easily be resolved. Some of these issues have been stated as quantitative vs. qualitative, and 'case-oriented' vs. 'variable-oriented,' holistic vs. analytic, or deterministic vs. probabilistic research.

These distinctions pose some false alternatives and represent differences in emphasis rather than in kind. Thus, statistical procedures are, by their very nature, based on larger quantities, but 'small *N*' analyses similarly may deal with numbers and certain quantitative data. 'Qualitative' aspects here refer to the presence or the absence of a factor, as in QCA, but should not be confounded with 'qualitative' methods as, for example, hermeneutics or participant observation (see *Interpretive Methods: Micromethods*; *Participant Observation*) in a different sense of the word.

In the same way, 'case-oriented' means that emphasis is laid on a number of known and theoretically or historically particularly interesting and relevant cases, as opposed to, for example, the anonymous ones in random sampling, but all scientific inquiry is, of course, based on distinct units (cases) *and* variables and a number of relevant observations (King et al. 1994). 'Holistic' is also a misleading term referring to the usual higher complexity of case-oriented research but it does not mean in this context that cases can only meaningfully be understood in their 'entirety' (as for example in Confucian philosophy) as opposed to 'critical-rational' analysis.

Finally, 'deterministic' is a misnomer for procedures based on Millian methods or logical distinctions as in Boolean algebra. Here, only the presence or absence of a factor is referred to and not its 'probabilistic' nature with respect to its effect on the internal or external validity of any particular finding. 'Logically minimized' relationships found by Boolean procedures may be as 'spurious' as correlational analysis. The revealed patterns indicate different more complex 'qualitative' constellations, but they are not based on their relative frequency or 'probabilistic' inferences.

As to the 'real' problems, however, these persist and only certain approximations or tradeoffs between certain emphases are possible. Millian methods and similar ones can only serve to eliminate false hypotheses and to delineate the 'conditions for the occurrence' of a phenomenon (Cohen and Nagel 1934). Within this realm the truly 'qualitative' and interpretative skills of knowledgeable experts etc. are called for to provide meaningful assessments of the factors involved. These may be more tedious and less universalistic, but they are often more meaningful and less superficial than many of the 'macroquantitative' studies.

Most 'small-*N*' comparative studies, but also the majority of statistical analyses, have been 'cross-sectional' and 'static.' There have been some advances in the statistical field in recent years as far as 'pooled time series' and 'event history' modeling are concerned

(see also *Event Sequence Analysis: Methods*), but these share some of the limitations and the relative 'superficiality' of macroquantitative procedures. There are, as yet, no truly 'dynamic' procedures comparable to the MSDO or QCA techniques combining both cross-sectional and longitudinal aspects in a 'macro-qualitative' way. These still rely more on well-informed 'thick narratives' (see *Thick Description: Methodology*), but some specific 'path-dependent' techniques have been developed more recently (see *Structural Equation Modeling*).

A further limitation refers to the 'structure' ('macro'-) and 'actor' ('micro'-) dimensions of any social and, by implication, comparative analysis. Most comparative studies are concerned with the broader and presumably more enduring 'structures' of social processes both in time and space. The final outcome, nevertheless, may very well depend on any particular (group or individual) actor in the analyzed chain of events. Again, only certain approximations can be arrived at in this regard and certain tradeoffs between more 'universalizing' and more 'individualizing' approaches (Tilly 1984) have to be made. No 'macro'–'macro' shortcuts should be made in sociological explanations because always concrete subjects with their respective perceptions and preferences are involved, which only in a later step may be aggregated at a 'meso-' or 'macro'-level. Whether these actors should always be conceived of as 'rational' or as 'Resourceful, Restricted, Expecting, Evaluating, Maximizing Men' (RREEMM) or women (Esser 1993, Chaps. 6 and 14) need not concern us here. But at least the 'expecting' and 'evaluating' aspects point to the possibility that different things may be perceived, valued, and maximized in different cultures. In any case, the broader structural (and also larger contingent) aspects can be conceived as the 'opportunity set' within the boundaries of which actual actors operate and respective choices are made. In this regard, the 'scale' and detail of any particular analysis may vary as with maps for different purposes. Thus, a hiker's map may serve for a close orientation on the ground, whereas a pilot's map may only include the more important contours and 'structural' features, and even a global and 'universal' one may be devised, but it is important that all scales and features remain consistent with each other!

A final point must again be mentioned, the problem of inference. Another important and relatively recent work has stressed, once more, the 'unified logic of inference' of both quantitative and qualitative research (King et al. 1994). In spite of many important issues which are raised and clarified in this book, the experiences and recommendations of the authors lean somewhat too strongly on the 'quantitative' side of research. This relates both to the 'internal' and 'external' validity of the inferences to be drawn. In terms of the 'internal' causal chains which should be established, King et al. look at them, as has also been noted by Scharpf (1997, 22 ff.), from the side of the independent variables in a 'forward' sense towards their eventual effect on the outcome.

Much of small-*N* comparative research, as for example in policy studies, is, however, concerned with disentangling the different influences on an outcome from a 'backward' perspective in order to identify the different factors in their complex interactions over time involving the major 'structures' and 'actors,' but also many contingent and unique aspects. This resembles more a detective's investigation of a crime or an accident than any more 'nomothetic' generalizing scientific efforts and remains 'idiographic.'

The more complex pattern of causality which has been established in this way in terms of its internal validity, can then, however, serve to devise new more complex hypotheses which can be tested across more similar (comparable!) cases and, possibly, establish a certain range of their external validity through 'process tracing' and similar procedures. A much stronger theory may thus emerge. We concur, therefore, with Goldstone's conclusion that the comparative case study method does '… not stand on flimsy methodological foundations, but on a stronger combination of skill and insight in framing and testing hypotheses than many large-*N* statistical studies' (Goldstone 1997, p. 119).

6. *Probable Future Directions of Theory and Research*

The strength of the comparative method lies, therefore, in its critical applications to theory testing and development. The state of affairs of theory building in the social sciences remains relatively weak and is subject to specific problems and restrictions and to the general 'malleability' of its substance. Even if only certain approximations can be reached, the relevance of many macrosocial and political problems, for example in policy research, institution-building, 'constitutional engineering', etc., for potentially millions of people make it worth every effort. Better theory, in the longer run, then may also contribute to a better practice.

It is difficult, if not impossible, to predict any more distinct developments in this regard, but at least some of the *desiderata* which reflect some of the existing deficiencies can be outlined. First of all, many more existing propositions across the various fields of social inquiry based on other methods of data collection and theory building should be subjected to systematic testing by comparative methods as far as they refer to 'small-*N*' environments. Counterfactual examples and 'most different' cases should be sought to establish their respective range of external validity. Second, from this body of research, more coherent propositions may emerge delineating their respective range of application, both in space and time.

It seems likely that such 'medium range theories' will increasingly be based on regional and historical commonalities of certain countries and cultures. Some 'deep structures,' such as the social cleavages in Western Europe, and 'deep cultures,' such as certain characteristic submilieus in Germany or Italy, have persisted over centuries and can be linked to present developments (see also Flora 1999). What is needed are still better qualitative and quantitative methods to assess their continuities and changes over time and dynamic concepts and tools to establish their 'self-referential' nature, not just in abstract theory, but also in empirical reality. Such a combination of theoretical and historical concerns may also employ better 'analytic narratives' and lead to a new focus on the interactions between structures and actors.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Comparative History; Comparative Studies: Method and Design; Linguistics: Comparative Method; Mill, John Stuart (1806–73); Punishment, Comparative Politics of; Slavery: Comparative Aspects

Bibliography

Almond G A, Genco S J 1977 Clocks, clouds and the study of politics. *World Politics* **29**: 489–522

Berg-Schlosser D, De Meur G 1997 Reduction of Complexity for a Small-N Analysis. *Comparative Social Research* **16**: 133–62

Cohen M, Nagel E 1934 *An Introduction to Logic and Scientific Method*. Harcourt, Brace, New York

Cook Th D, Campbell D T 1979 *Quasi-Experimentation—Design and Analysis Issues for Field Settings*. Houghton Mifflin, Boston

De Meur G, Berg-Schlosser D 1996 Conditions of authoritarianism, fascism and democracy in inter-war Europe: systematic matching and contrasting of cases for 'small N' analysis. *Comparative Political Studies* **29**(4): 423–68

Durkheim E 1988 *Les Règles de la Méthode Sociologique*. Flammarion, Paris [originally published in 1894]

Esser H 1993 *Soziologie*. Campus, Frankfurt

Flora P (ed.) 1999 *State Formation, Nation-Building and Mass Politics in Europe—The Theory of Stein Rokkan*. Oxford University Press, Oxford, UK

Goldstone J A 1997 Methodological issues in comparative macrosociology. *Comparative Social Research* **16**: 107–20

King G, Keohane R O, Verba S 1994 *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press, Princeton, NJ

Mill J S 1974/75 A system of logic. *Collected Works of J. S. Mill*. Routledge and Kegan Paul, London, Vols. vii and viii [originally published in 1843]

Popper K R 1934/1968 *The Logic of Scientific Discovery*. American edn. Harper and Row, New York

Przeworski A, Teune H 1970 *The Logic of Comparative Social Inquiry*. Wiley, New York

Ragin C C 1987 *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. University of California Press, Berkeley, CA

Scharpf F W 1997 *Games Real Actors Play—Actor-Centered Institutionalism in Policy Research*. Westview Press, Boulder, CO

Tilly Ch 1984 *Big Structures, Large Processes, Huge Comparisons*. Russell Sage Foundation, New York

Weber M 1949 *The Methodology of the Social Sciences* [trans./eds. Shils EA, Finch HA]. Free Press, Glencoe, IL

D. Berg-Schlosser

Competencies and Key Competencies: Educational Perspective

In accordance with mainstream thought, we understand 'competence' as referring to combinations of those cognitive, motivational, moral, and social skills available to (or potentially learnable by) a person or a social group that underlie the successful mastery through appropriate understanding and actions of a range of demands, tasks, problems, and goals.

Because everyday schooling, work, and social life present countless tasks and goals, it is possible to postulate any number of competencies as necessary personal and/or social preconditions for successfully meeting task demands. Competence generally implies complex action systems with cognitive, motivational, social, and behavioral components, rather than basic cognitive abilities or simple skills. Because the psychological structure of such systems is frequently unclear, and the possibilities for substituting, compensating, and/or combining single components of competencies are unexplored, the psychological relations between external demands and personal competencies are often regarded as tautological. It seems as if the concept of competence reveals a negative correlation between its popularity and its precision.

What are the consequences of this state of affairs? One clear solution might be to avoid using the term competence and, of course, key competencies in scientific discourse, and to seek more appropriate concepts or terms for the intended construct. Would this constitute progress? Alternative concepts such as skill, ability, aptitude, or proficiency are not used in ways that are any more coherent or consistent. Moreover, the disadvantages of using the concept of competence are compensated somewhat by the semantic advantages. There have been many attempts to constrain and sharpen the concept of competence on phenomenological, definitional, or theoretical grounds. However, these attempts have failed to gain broad conceptual acceptance in the social sciences. This entry starts with a brief description of the concepts of 'competence' and 'key competence' in terms of their rather vague scientific meaning and their variable practical use. It then goes on to discuss both the psychological and the educational problems so often associated with the concept of key competencies

(key qualifications, core skills). The final section draws a few conclusions for educational practice, education processes, and instruction processes.

1. Definitions of the Concept of 'Competence'

In general, we know what the terms 'competence,' 'competencies,' 'competent behavior,' or 'competent person' mean, without being able to define them precisely or differentiate them clearly. The same can be said for terms such as 'ability,' 'qualification,' 'skill,' or 'effectiveness.' The use of these terms as synonyms is reflected in dictionary entries as well. For example, Webster's dictionary defines 'competence' as 'fitness or ability.' 'Capability,' 'capacity,' 'efficiency,' 'proficiency,' and 'skill' are given as synonyms or related terms.

An inspection of the Latin roots and historical variations in meanings reveals that competence is also understood to mean 'cognizance' or 'responsibility.' This concept of competence has been, and continues to be, used in very specific and arbitrary manners in biology, immunology, jurisprudence, and in several other scientific disciplines. However, this will not be elaborated further in this entry. Restricting our focus to the use of the term 'competence' in psychology and education still leaves a wide variety of definitions. Nonetheless, in many disciplines, competence is interpreted as a roughly specialized system of abilities, proficiencies, or skills that are necessary or sufficient to reach a specific goal. This can be applied to individual dispositions or to the distribution of such dispositions within a social group or an institution (such as a business company).

In recent decades, 'competence' has become a fashionable (and sometimes even modernistic) term with a vague meaning not only in lay usage, but also in many social sciences. One could even talk about a conceptual 'inflation' with the lack of a precise definition being accompanied by considerable surplus meanings. A typical illustration of this can be found in a report published in 1998 by an advisory committee for technology and innovation appointed by the German Chancellor. The report states:

> Competence can generally be understood as knowledge times experience times power of judgment. Knowledge is the necessary foundation of competence, and experience is the habitual way one deals with acquired and continuously changing knowledge. Power of judgment is a criterion for the independence of knowledge and its use. Thus, competence is always more than just knowledge or just experience (BMBF 1998, p.10).

Specific competencies mentioned in this report include: economic, technological, technical, and methodological competencies; social competencies; creativity and innovation skills; mobility and flexibility, combined with persistence, reliability, and precision.

In some respects, this report is typical of many current publications: although the term 'competence' is used to refer to both high-achieving individuals and successful social groups that master specific tasks and reach important goals, it is, at the same time, always concerned with the necessary learning dispositions that are available to individuals and/or the members of social groups and have to be used to solve demanding problems. The variety of meanings given to the concept of competence is seen not only in its many uses but also in the construction of compound words to express competence, such as: media competence, business competence, traffic competence, age competence, and also cognitive, social, motivational, personal, and many other competencies. It is not possible to discern or infer any coherent theory from these many uses. There is no basis for a theoretically grounded definition or classification from the seemingly endless inventory of ways in which the term 'competence' is used.

2. Key Competencies

In recent times, similar concepts such as key competencies, core competencies, key qualifications, and core skills have become popular in the social sciences and in educational policy. The concept of key competence is, however, no less vague or ambiguous than that of competence. Clear and well-reasoned distinctions between the two concepts are either arbitrary or nonexistent. Towards the end of the twentieth century, at least 654 different key competencies have been suggested in German literature on occupational training alone. These range from such constructs as creativity, logical thinking, problem-solving skills, achievement readiness, independence, and concentration abilities to foreign language skills, communication skills, and media competencies.

The search for key competencies is motivated by two notions: (a) the well-founded assumption that competencies acquired in school and vocational settings are learned and used in context-specific ways (e.g., within a discipline, within a vocation, within a company), and (b) that most activities over the life course take place in a variety of social and vocational contexts. This has led to the search for context-independent key competencies that are equivalent in their use and effectiveness across different institutions, different tasks, and under varying demand conditions.

Why is the concept of key competence so attractive? The term generally refers to multifunctional and transdisciplinary competencies that are useful for achieving many important goals, mastering different tasks, and acting in unfamiliar situations. For many teachers and politicians, there is also the hope that a curriculum bursting at the seams with the many

competencies necessary for modern life could be reduced if it were only necessary to transmit a limited number of key competencies.

Regardless of how much scientists and practitioners agree on the formal criteria for defining key competencies, there are major disagreements about which competencies should be classified as key competencies. If one considers the tasks and demands that most people have to master in order to achieve personal success in modern societies, the following key competencies can be identified as being mentioned frequently in the available literature: oral and written mastery of the mother tongue; mathematical knowledge; reading competence for rapid acquisition and correct processing of written information; mastery of at least one foreign language; media competence; independent learning strategies; social competencies; and divergent thinking, critical judgments, and self-criticism.

When key competencies are identified and defined within a normative-philosophical and socially critical frame of reference, one derives, for instance, those key competencies characterized by Canto-Sperber and Dupuy (in press), who sketch the conditions for the possibility of a good life in modern, complex democratic societies. The authors derive some personal competencies and five key competencies from an analysis of the features and characteristics of the general idea of good living. They are not concerned with particular basic skills, but with the dimensions that define an abstract, multidimensional space in which specific skills, abilities, and capacities can be localized.

The primary personal prerequisite for a good life is conceptual competence—defined either as general conceptual competency (conceived as a context- and content-free ability) or as special cognitive competency (conceived of as a special cluster of cognitive prerequisites and context-dependent capabilities). Conceptual competency is supplemented by procedural competency, motivational competency, and action competency. These competencies are not understood as cognitive goals; they are not curricular bounded knowledge or skills, but cross-curricular competencies.

In addition to these person-oriented competencies for a good life, Canto-Sperber and Dupuy (in press) identify five key competencies that are defined and characterized on a very high level of abstraction. These are competencies for coping with complexity, perceptive competencies, normative competencies, cooperative competencies, and narrative competencies.

3. *Psychological and Educational Problems With the Scientific Concept and the Practical Application of Key Competencies*

Regardless of whether one chooses a reality-based or a philosophical-normative perspective to identify important key competencies, there are always some questions and problems requiring principal or pragmatic answers. Competencies and key competencies follow no strict formal definitional constraints but are defined at very different levels of universality, generality, and abstraction. It is often noted that very abstractly formulated key competencies need to be further divided into subcompetencies. Such scientific plans have often failed in psychology, however. The underlying multilevel models can be reconstructed logically, but not validated psychologically. The different degrees of abstraction therefore mean a fundamental asymmetry in competence research: high abstraction: intellectually brilliant, pragmatically hopeless; low abstraction: pragmatically useful, intellectually unsatisfactory.

A related but independent issue is the frame of reference within which key competencies are defined. Competencies and key competencies may be identified from philosophical ideas about the nature of humankind, ideas about the good life and a desirable society, or even expectancies about present human life and social demands. There is a strong danger that the necessary skills for successful everyday life, for social and personal effectiveness, or for professional success will be trivialized when compared with normatively anchored universal competencies. Nonetheless, if one wants to go beyond an individual's adaptation level to the world of today with its limited possibilities of further development, and change the world by equipping people with the appropriate competencies, it is necessary to choose a normative starting point when defining key competencies rather than an empirical one.

Key competencies are always complex systems of knowledge, beliefs, and action tendencies, that are constructed from well-organized, domain-specific expertise; basic skills; generalized attitudes; and converging cognitive styles. Without a minimal hypothetical structure and process analyses, key competencies can easily deteriorate into catchwords that refer only to desired states. Frequently, cognitive abilities, cognitive styles, and emotional qualities are viewed as constituent components of key competencies. There are large individual differences in these psychological features, and it is doubtful whether and to what degree they can be modified through learning; whether deficits can be compensated; and whether people can indeed change in the required direction. If this question remains unasked or unanswered, there is a danger that an academic discussion of key competencies may trivialize the already enormous interindividual differences and thus lead to an upsurge in individual discrimination.

Frequently dangerous illusions regarding the possibilities and limitations of socialization and education are tied to the concept of key competencies. Many scientists suggest to an amazed public that it will no longer be necessary to acquire a large amount of world knowledge, expertise, and competencies or to work

hard to learn a lot. In the future it will be sufficient, so the story goes, to possess some key competencies, to have learned how to learn, and to have acquired enough media competence to obtain the necessary information at any time in an electronic form. This will create an attitude of personal self-confidence that one can always react appropriately and creatively in response to any difficult situations in life. Modern cognitive psychology would tell us that such an educational model is not only utopian but mostly just nonsense.

The more general a competence or strategy (i.e., the range of different types of situations to which it applies), the smaller its contribution to the solution of demanding problems. Over the last decades, the cognitive sciences have demonstrated convincingly that content-specific skills and knowledge play a crucial role in solving difficult tasks. Generally, key competencies cannot compensate adequately for a lack of content-specific competencies (Weinert 1998).

The lively discussions on the meaning of systematic vs. situated cognitions have demonstrated that general competencies have virtually no practical utility by themselves. Instead, specific knowledge, embedded in experience, is required to implement available competencies successfully when solving specific practical problems. For many key competencies, the question is whether and how they can be acquired through planned instructional programs. A typical example is critical thinking. Although there are special training programs for this key competence (Halpern 1998), their construction and efficiency remain controversial in scientific terms.

4. Concluding Remarks

Putting aside the philosophically and normatively derived key qualifications and concentrating on the methodologically functional key qualifications that can be used in the greatest possible range of different and important situations, one can name four essential criteria for their acquisition from an educational perspective. First, key qualifications should be capable of being learned by as many people as possible. This holds just as much for the oral and written forms of one's mother tongue and foreign languages as for basic mathematical and statistical skills or for media competence. However, not everything that can be learned can be taught; even though key qualifications have proved to be, in principle, both teachable and learnable. Things become problematic with key qualifications such as critical or logical thinking, because these abilities depend very strongly on the individual endowment with intelligence along with interindividual differences in intelligence. As a result, it cannot be assumed that the majority of people can acquire this key competency in a corresponding way. Second, key competencies have to be highly automatized if they are to be applied in very different situations. In other words, they have to be available in the behavioral repertoire and relatively easy to access. Third, key qualifications need to possess reserve capacities for acquiring the context knowledge that each case demands, building up new experiences, and acquiring additional skills. Fourth, key competencies not only have to be acquired in a way that makes them a domain-specific expertise, but they also have to be trained through forms of situational learning to make them adaptable to fit different occasions. Viewed from the standpoint of the psychology of learning, this can be attained only when methodological key competencies are acquired systematically, but their application is subject to permanent training in variable contexts.

See also: Co-constructivism in Educational Theory and Practice; Cognitive Styles and Learning Styles; Educational Learning Theory; Human Development, Successful: Psychological Conceptions; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values; Social Competence: Childhood and Adolescence; Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Bundesministerium für Bildung und Forschung (BMBF) 1998 *Kompetenz im globalen Wettbewerb* (*Competence in a Global Competition*). BMBF, Bonn Germany

Canto-Sperber M, Dupuy J-P (in press) Competencies for the good life and the good society. In: Rychen D, Salganik L (eds.) *Defining and Selecting Key Competencies*. Hogrefe & Huber, Göttingen, Germany

Halpern D 1998 Teaching critical thinking for transfer across domains. *American Psychologist* **53**: 449–55

Weinert F E 1998 Vermittlung von Schlüsselqualifikationen. In: Matalik S, Schade D (eds.) *Entwicklungen in Aus- und Weiterbildung—Anforderungen, Ziele, Konzepte* (*Beiträge zum Projekt "Humanressourcen"*) Nomos, Baden Baden, Germany, pp. 23–43

F. E. Weinert†

Competitive Strategies: Organizational

A firm's competitive strategy concerns how to compete in the business areas the firm operates. In other words, competitive strategy means to define how the firm intends to create and maintain a competitive advantage with respect to competitors. A business area identifies: (a) a distinct set of products or services sold

to a uniform set of customers; and (b) a well defined set of competitors to be faced. Holding a competitive advantage over competitors means to be more profitable than competitors over the long term. Superior profitability over competitors can be achieved in different ways: (a) a company sustains lower costs than competitors and this can lead to either selling the same volume as competitors (if the selling price is the same) with larger margins (because of lower costs) or selling a larger volume of products, obtained as a consequence of the reduction of the selling price (the selling price can be reduced as far as the difference between price and cost is that of competitors and therefore the company has the same margins as competitors); or (b) a company is able to commercialize a differentiated product from that of competitors: this advantage can be exploited either by increasing the selling price and therefore exploiting higher margins at the same volume or attracting a larger number of customers (which means larger volume) keeping the same price as competitors.

The definition of competitive strategy itself addresses the underlying dynamic character of the issue. Competitive advantages are created as the result of direct actions of the firm (proactive strategies) or as the result of the reaction of the firm to external changes (reactive strategies). As a consequence of the creation of a competitive advantage, competitors react trying to erode the advantage. This can be done either by imitating the holder of the advantage or doing actions that create a new advantage. Therefore, the protection of the advantage is also part of a competitive strategy. This leads to the key concept of the sustainability of the competitive advantage. The stronger the protection of the advantage, the longer the window of time in which the advantage can be exploited by the holder and the higher profitability for the holder. Moreover, an advantage can be maintained exploiting the time factor. Once an advantage has been created, the holder of the advantage can exploit the time competitors spend to close the gap to generate another advantage. The strategy in this case is staying ahead of competitors.

Therefore, a competitive strategy within a given business area is examined looking at two factors: the creation of the competitive advantage and the protection of the competitive advantage. As it can be argued from the discussion below, whatever type of competitive strategy, the ultimate source of competitive advantage and the potential of protection of such advantage reside in the resources of the firm. Resources may be tangible (such as physical or financial assets), intangible (such as technology, reputation, culture), human (specialized knowledge and competencies, communication and inter-relational capabilities, motivation), or organizational (organizational routines and business processes, organizational architecture, interrelations with suppliers and customers).

1. *Creating the Competitive Advantage*

The creation of competitive advantage within a given business area is examined looking at the two categories of competitive strategies as mentioned above: proactive and reactive competitive strategies.

1.1 *Proactive Competitive Strategies*

In proactive competitive strategies, the creation of competitive advantages is the result of the direct actions of the firm. Therefore, in this case, the sources of such advantages are internal to the business: the creation of competitive advantage is the result of the actions such as the introduction of a new product, or the use of a new production process. Such change is successful if it generates higher profitability for the firm than competitors (under one of the forms seen above): this means that the firm has created a competitive advantage over competitors (which is different from saying that 'the firm holds a competitive advantage' which means to maintain the advantage and have higher profitability than competitors in the long term). The competitive advantage is therefore the result of a competitive strategy deliberately meant to introduce changes in the business area in order to make the firm more profitable than competitors. It should be noted that such a form of competitive advantage creation not only generates advantages to the holder of the advantage, but may also destroy existing advantages, as the basis of competition may radically change.

The types of competitive advantages and the underlying competitive strategies can be of two types.

1.1.1 Improvement of performance (same game competitive strategy). If the business area is considered at a certain time, the competition among firms is played around some key performance factors such as cost, time, quality, variety, etc. (see below in this section). These are the performance factors by which customers measure their satisfaction.

A first type of competitive advantage creation concerns the improvement of product/service along one or more of the existing performance factors. The basis of competition is still the same as the performance factors are still the same. This form of competitive advantage creation is the most frequent in mature and established business areas, where the basis of competition tends to be more stable.

In more detail, major performance factors and therefore sources of advantage are the following:

Superior functional performance. Each product/service is related to specific functional performance. Products with superior functional performance with respect to competitors are sources of advantage.

Complementary services. A performance factor can concern the services added to products. Such services can be embodied into the product or separated from

the product. The ability to add services can be a source of advantage.

Quality (compliance). Another performance factor is the compliance of all products produced to the level of quality claimed by the firm. Firms able to commercialize products all of the same quality may create an advantage (it is also referred to as zero defect production).

Variety (mix). Another performance factor concerns the range of products the firm is able to offer within the business area considered. A wider mix of products can be a source of advantage.

Delivery time. A further performance factor is the ability to deliver products when customers want. The most desired time for customers is not always the earliest: if a firm produces industrial goods the required performance (and source of advantage) is to bring products to customers exactly when the customers use them in their production process.

Time-to-market. Time-to-market is the time spent to have a new product idea brought onto the market. Reducing time-to-market enables the firm to introduce new products earlier than competitors.

Environmental compatibility. In certain businesses the environmental compatibility can represent a source of advantage. Helping customers to comply with norms and legislation can be a form to differentiate products from those of competitors.

Cost. Cost is often a key performance factor in established businesses and reducing costs, a major competitive weapon. Major sources of cost advantage are here reported.

(a) Scale economies. Scale economies occur when an increase in the production scale determines a reduction of the cost per product unit. It means that an increase of the resources used determines a more than proportional increase of the output. The production scale at which the economies of scale are completely exploited is the minimum optimal scale (which is the minimum scale to obtain the minimum cost). This means that firms operating the minimum optimal scale have cost advantages with respect to firms operating lower scales.

(b) Learning curve. The learning curve shows the relation existing between cumulated production and cost per product unit. There is a reduction of 20–30 percent of the cost when the cumulated production doubles. Learning curve advantages mainly refer to savings in production, because of fewer scraps, shorter idle times, and incremental innovations. This means that firms which start producing a product before competitors can benefit from cost advantages because of learning effects.

(c) Process technologies. A certain production process using a defined process technology may be technically superior and more efficient than other process technologies.

(d) Input costs. There may be cost advantages related to the cost of input factors (materials or labor). Firms that directly own sources of raw materials or have access to input factors at lower cost have cost advantages.

(e) Localization. Cost advantages can also be based on the geographical location. The proximity to raw material sources or the localization of labor-intensive activities in low labor cost countries is a typical example.

(f) Production capacity utilization. If market demand fluctuates a major problem is the adaptation of production capacity. Firms with more flexible production systems benefit from cost advantages.

Finally, within a certain business area, a firm's competitive strategy can be based on a niche strategy. This means that competitive advantages are based on the ability to offer products/services tailored to the specific needs of a limited number of customers within the business area.

1.1.2 Change of the rules of the game (new game competitive strategy). Competitive advantages can be reated introducing a new performance factor in the business area. In such cases, the 'rules of the game' in the business area change as the basis of competition changes. It should be noted that there is a difference between cost and the other performance factor. The introduction of cost as a performance factor gives an advantage to the firm in absolute terms, whereas the introduction of another performance factor gives an advantage if that factor is perceived by the customers as critical.

1.2 Reactive Competitive Strategies

The creation of competitive advantages in proactive competitive strategies is the result of the actions of the firm. Competitive advantages can also be the result of the reaction to external changes.

External factors, which may determine profitability differences among firms, may be: changes in the consumer demand, price changes (for example, changes of the price of a raw material, price changes due to changes of the exchange rates), and technology changes. All these changes affect firms' profitability. The impact of such changes on firms operating the same business may strongly differ according to the responses of the firm to such changes, which in turn depend on the resource profile of the firms themselves. The quickness and the effectiveness of response to changes affect the profitability. The capability to respond quickly and effectively is based on a key resource (information) and a key competence (flexibility). Firms able to grasp information quickly are able to analyze weak signals through early warning systems. The ability to respond effectively depends on a firm's flexibility, which is in turn strongly linked to

characteristics such as organization, decision systems, culture, values. Such competitive advantages are purely the result of reactive strategies, i.e., strategies taking place as reaction to external changes.

2. Protecting the Competitive Advantage

The company that holds the advantage tends to maintain this position (in certain cases it can exploit a monopolistic position), whereas competitors tend to erode (destroy) the advantage. The erosion of the advantage can take place either through imitation or creation of a new advantage.

Therefore, whereas the holder of the advantage has to defend its competitive advantage, the competitors have to erode (destroy) this advantage. The time window in which the competitive advantage can be held over competitors brings to the concept of the sustainability of the advantage: the larger time window, the more sustainable advantage. The sustainability is related to the capability/opportunity to protect the firm's source of advantage. Factors of protection are the following.

Secrecy. It is a form of protection especially if the source of competitive advantage is the production process. Although information can be obtained from products and there is an inevitable leakage of know-how through a variety of channels (technical community, suppliers, etc.), firms can keep secrecy over that. Firms that do not open their plants to anyone external to the firm may succeed in keeping their own process innovations secret.

Accumulated tacit knowledge. If the knowledge at the basis of the advantage is strongly embodied into people and technical systems, and relies upon people's experience, it is difficult to make the knowledge explicit, and the related advantage can be highly protected. The well-known case of Italian small firms, often world leaders in niche markets and traditional manufacturing industries, is largely based on the accumulation of tacit knowledge. In very concentrated geographical areas (districts), there are networks of small firms that are strongly specialized in highly specific manufactures. Knowledge is, to a large extent, embedded in people living in that area.

Lead times. The ability to generate and quickly put new products on the market is a major source of advantage and is itself a major source of protection against imitation. It helps establish brand loyalty and credibility, accelerate feedback from customer use, accelerate learning effects, and consequently increase cost of entry for imitators. The market leadership is based on the ability to provide new generations of products frequently and stay ahead of competitors.

Complementary assets. The ability to make profits from the commercialization of a new product or service often depends on the availability of assets or competencies, such as production capabilities, marketing, distribution, after-sale service, which are complementary to the product and its technology. These complementary assets determine who exploits the new product introduction and the appropriation of the related benefits and to what extent. As a matter of fact, they are often the reason for failures of firms, that, having once generated the new product do not have the required complementary assets to exploit it or have to invest largely to acquire them. Both these conditions may cause the firm not to create a competitive advantage from the new product introduction. On the other hand, the availability of complementary assets for the firm may represent not only the means which ensure that profits are made from new product introduction but also a barrier against imitation by competitors especially when such assets may be difficult to acquire, access, or imitate.

Product complexity. In some sectors, product complexity is a major barrier against imitators. The large electromechanical equipment industry or the aircraft industry is a typical example.

Standards. In certain industries, making the firm's own product accepted as a standard opens and develops the market and raises barriers against competitors. This is especially true in network markets, where the compatibility of a product with related products (for example a software package with the related hardware) is a prerequisite for the product's success. Once the product is diffused throughout the market as a standard, it is difficult for competitors to erode the advantage. In fact, new products marketed by competitors should replace the existing product but may also involve the replacement of the other related products. This represents a major barrier. Computing, telecommunication, consumer electronics industries show examples in which standard setting is a key factor to appropriate the benefits of new product introduction and create a sustainable competitive advantage.

Pioneering radical new products. If a product is radically new, i.e., represents a strong discontinuity with the existing products/processes/services and lies upon a base of knowledge that is completely different from that behind traditional products, the firm is more protected against imitation. Potential imitators cannot rely on the existing knowledge base to replicate the product.

Strength of patent protection. Patenting is also a major means of protection. This is especially true in some sectors, chemical and pharmaceutical for example, where products can be clearly and rigorously described (often with a structural formula) and imitation by small differences is not easy as in most industries.

3. Other Competitive Strategies

Finally, it should observed that so far competitive strategies have been examined considering a given business area and that the actions or the reactions

involve only the firms operating in such business areas. In reality, higher profitability can be pursued implementing other strategies that are here mentioned but not treated in depth.

3.1 Creation of a Completely New Business Area (New Business Competitive Strategy)

A competitive strategy may be meant to create a completely new business area. If the business area is profitable for the firm, the firm creates a competitive advantage. The firm holds a monopolistic position as far as new firms enter the business area. This case especially occurs in highly innovative and emerging industries.

3.2 Enlarging the Geographical Scope or the Business Scope (Cross-market or Cross-business Competitive Strategy)

There are other two types of competitive strategies that can generate competitive advantage. They do not concern the specific business area considered but, in one case, the fact the firm operates that business in different geographical markets and, in the other, the fact that the firm operates different business areas.

In the first case, it is the geographical scope of the strategy that may represent the source of advantage. A firm can pursue a global strategy being present in different geographical markets. The global market is seen as one. The sources of advantage over competitors who compete on a local basis are mainly twofold: the homogeneity of market demand in different countries and the exploitation of economies of scale in different activities (R&D, production).

In the second case, the source of advantage resides in the range of business areas operated. The advantages are based on the interrelationships across distinct business areas. Interrelationships may concern the following: infrastructure (accounting, legal department, government relations, human resource management, finance); technology; purchased inputs; manufacturing (location of raw materials, production process, assembly, quality control, factory support activities); marketing (buyer, distribution channel, geographic market).

See also: Decision Making, Psychology of; Firm Behavior; Organizational Behavior, Psychology of; Organizational Decision Making

Bibliography

Abell D F 1980 *Defining the Business: The starting point of strategic planning*. Prentice-Hall, Englewood Cliffs, CA

Ansoff H I, McDonnell E T 1990 *Implanting Strategic Management*, 2nd edn. Prentice-Hall, Cambridge, UK

Grant R M 1991 *Contemporary Strategy Analysis. Concepts, Techniques, Applications*. Blackwell, Cambridge, MA, and Oxford, UK

Hall R 1992 The strategic analysis of intangible resources. *Strategic Management Journal* **13**(2): 135–44

Hamel G, Heene A 1994 *Competence Based Competition*. Wiley, Chichester, UK, and New York

Hamel G, Prahalad C K 1993 Strategy as stretch and leverage. *Harvard Business Review* March–April **71**(2): 75–84

Hax A C, Majluf N S 1991 *The Strategy Concepts and Process—A Pragmatic Approach*. Prentice-Hall, Englewood Cliffs, CA

Heene A, Sanchez R (eds.) 1997 *Competence-based Strategic Management*. J. Wiley, Chichester, UK, and New York

Leonard-Barton D 1992 Core capabilities and core rigidities: A paradox in managing new product development. *Strategic Management Journal* **13**: 111–25 Sp. I

Nonaka I 1991 The knowledge-creating company. *Harvard Business Review* November–December **69**(6): 96–104

Porter M E 1980 *Competitive Strategy: Techniques for Analysing Industries and Competitors*. The Free Press, New York

Porter M E 1985 *Competitive Advantage: Creating and Sustaining Superior Performance*. The Free Press, New York Collier Macmillan, London, UK

Prahalad C K, Hamel G 1990 The core competence of the corporation. *Harvard Business Review* May–June **68**(3): 79–91

Prahalad C K, Hamel G 1994 *Competing for the Future*. Harvard Business School Press, Boston, MA

Sanchez R, Heene A, Thomas H 1998 *Dynamics of Competence Based Competition: Theory and Practice in the New Management*. Elsevier Press, London

Stalk G, Evans P, Shulman L E 1992 Competing on capabilities: The new rules of corporate strategy. *Harvard Business Review*. March–April **70**(2): 57–69

Teece D, Pisano G 1994 The dynamic capabilities of firms: An introduction. *Industrial and Corporate Change* **3**(3)

Wernerfelt B A A 1984 Resource-based view of the firm. *Strategic Management Journal* **5**(2): 171–80

U. Bertelè and V. Chiesa

Compliance and Obedience: Legal

The law specifies what is acceptable behavior within a given society. A society enacts laws, and creates legal authorities, to regulate the conduct of its members by inhibiting the occurrence of those behaviors that are damaging to society, to the individuals within society, or to both. For the law to be effective in fulfilling its role of minimizing the occurrence of socially damaging behaviors, people must comply with the dictates of the law—i.e. the law must be obeyed.

If the law does not shape the behavior of most citizens most of the time, bringing those people's actions into conformity with legal rules, then the legal system is not effectively fulfilling its social regulatory function. In fact, the ability to secure widespread

compliance with the law is often viewed as a central characteristic of effective legal authorities. 'The lawgiver must be able to anticipate that the citizenry as a whole will … generally observe the body of rules he has promulgated' (Fuller 1971, p. 201), since the effective exercise of legal and political authority requires compliance from 'the bulk of (citizens) most of the time' (Easton 1965, p. 185). Hence, 'a judge's ruling means little if the parties to the dispute feel they can ignore it. Similarly, passing a law prohibiting some behavior is not useful it if does not affect how often the behavior occurs' (Tyler 1990, p. 19).

Recognizing the importance of gaining widespread compliance with the law, most discussions of compliance focus on the issue of why people obey or disobey laws. One straightforward and widely noted perspective on this issue is the deterrence, sanctioning, or social control perspective. That perspective argues that law gains its influence over people's behaviors because people are concerned about being caught and punished for wrongdoing. It is based upon the rational choice model of personal motivation, which views people as motivated by the desire to maximize their gains and to minimize their costs in social interactions.

From the deterrence perspective, two issues are relevant to decisions about whether or not to comply with the law. These issues are the expectation of gain, and the expectation of loss. The expectation of gain involves the person's estimate of the likely gains from engaging in law-breaking behavior. That estimate is a multiplicative combination of the estimated likelihood of gain and the expected amount to be gained. For example, a person contemplating robbing a bank has an estimated gain that reflects his sense of the likelihood that he will gain a particular sum of money from robbing a bank.

The expectation of loss involves the possibility of being caught and punished for engaging in rule-breaking behavior. This expectation can be broken down into two components: estimates of the likelihood of being caught and punished for law-breaking (the probability of punishment), and estimates of the expected magnitude of such punishment (the severity of punishment). These two estimates are typically multiplied to produce the subjective estimated risk associated with law-breaking behavior.

One way to reduce the amount of law-breaking behavior is to minimize the possibility of gain that is associated with such behavior. So, to minimize the risk of robbery, taxi drivers put their profits in a floor safe and people carry little pocket money when they go out at night. However, since the amount to be gained by law-breaking is typically beyond the control of legal authorities, studies of deterrence usually focus on the ways in which changing either the estimated probability of punishment or the estimated severity of punishment affect the occurrence of criminal behavior.

The likelihood of being caught and punished for breaking the law can be changed in many ways, such as by increasing the number of police officers seeking to solve crimes, by raising the likelihood that charges will not be dismissed, and by altering legal procedures to change the likelihood of conviction. The move toward a nonunanimous jury decision rule or the use of smaller jury, for example, increase the likelihood of being found guilty for a crime.

The severity of punishment can be altered by increasing the penalties for various crimes or by limiting or eliminating parole. The recent increase in the number of crimes for which criminals in the USA might receive the death penalty is one highly salient example of heightening the potential severity of punishment.

When legal authorities heighten either the actual likelihood of being caught and punished for wrongdoing or the severity of punishment, they are increasing the objective risks of rule-breaking. The authorities assume that these changes alter people's subjective estimates of the likelihood and severity of punishment for wrongdoing, and, as a consequence, lead to lower levels or rule-breaking behavior. Studies on the deterrent effect of estimated risk suggest that such deterrence effects do occur (see Nagin, 1998). While studies often find evidence of deterrence effects of rule-breaking behavior, the magnitude of those effects is typically small. For example, MacCoun (1993) examines deterrence studies in the area of drug use and estimates that approximately five percent of the variance in drug use behavior can be explained by variations in the expected likelihood and/or severity of punishment. Other studies using panel designs fail to find evidence of any significant independent deterrence effects (Paternoster et al. 1983, Tyler 1990). Hence, when deterrence effects occur, their magnitude is low.

The deterrence effects identified are more strongly associated with people's estimates of the probability of punishment than they are with the anticipated severity of punishment (Nagin and Paternoster 1991, Paternoster 1987, 1989, Paternoster and Iovanni 1986). This suggests that efforts to increase compliance with the law need to focus on increasing the presence of the police, on encouraging apprehension, and/or on raising the likelihood of conviction in the courts. Efforts to lower the crime rate by intensifying penalties—such as the proliferation of death penalty offenses—will be relatively less successful.

Why are risk judgments not more strongly related to law-breaking behavior? One reason is that judgments of risk have to be reasonably high to influence people's behavior—i.e., they have to exceed some threshold level to be psychologically meaningful (Ross 1982). However, in most actual situations, the objective risks of being caught and punished are quite low. For example, the objective risk of being caught, convicted and imprisoned for rape is 12 percent, for robbery 4 percent, for assault, burglary, larceny and motor vehicle theft, around one percent (Robinson and

Darley 1995). Of course, the psychological or subjective estimate of risk is the key to people's behavior—not the objective risk. Research suggests, however, that subjective risks are, if anything, lower than objective risks. As a consequence, subjective estimates of risk are quite low.

Ross (1982) applies the psychology of deterrence to the problem of drunk driving. He suggests that raising risk estimates enough to lower the rate of behavior, while not impossible, involves prohibitively high costs in terms of police manpower and citizens' willingness to accept state intrusions into their personal lives. He argues that it is difficult to implement deterrence approaches within democratic societies. Interestingly, Ross argues that even the intensive efforts of Scandinavian authorities, including random road blocks and other similar expensive and intrusive law enforcement measures, are insufficient to create and maintain subjective risk estimates that are high enough to deter drunk driving. Many of the problems associated with deterrence-based strategies are structural; they concern our general inability to monitor people's opportunities to break laws.

This argument suggests that there are situations in which deterrence strategies will be more or less effective in deterring law-breaking behavior. The two key variables that determine how conducive a situation is to a deterrence approach are the ease of behavioral surveillance and the level of resources that society is willing to devote to the task of surveillance. The effects of behavioral surveillance on the rate of law-breaking behavior can be illustrated by tax payments. If people are wage earners, their income is recorded, the possibility of hiding cheating is low, and the opportunities for effective deterrence of law-breaking behavior are high. The societal resource variable is illustrated by the case of murder. The objective risk of being caught and punished for murder is high (around 45 percent; Robinson and Darley 1997) because society has committed considerable resources to resolving this type of crime. Because society is willing to devote sufficient resources to investigating murders, the likelihood of being caught for committing a murder is high enough for deterrence to be effective in lowering the murder rate.

However, the case of deterring murder illustrates an important limit to deterrence strategies. They are more effective in shaping the crime rate for crimes that are committed for instrumental reasons. For example, car theft, burglary, and crimes of this type are typically motivated by calculations about the costs and benefits expected from law-breaking. However, other crimes are expressive; they are more strongly shaped by one's emotional state and events of the moment. Crimes of this type, such as rape and many murders, occur on the 'spur of the moment' and in the 'heat of passion.' Such crimes are not influenced by deterrence considerations, irrespective of the possibility of being caught and punished for wrongdoing. Deterrence approaches work most effectively in shaping the occurrence of instrumentally motivated crimes.

The problem faced by those responsible for the everyday enforcement of law in democratic societies is that, for most crimes, the resources devoted to law enforcement are low and the opportunities for cheating are high. As a consequence, deterrence strategies are unlikely to be a sufficient basis for an effective system of social regulation. Deterrence can form the foundation of efforts to maintain the legal order, but they cannot be a complete strategy for gaining compliance. To ensure sufficient public compliance, citizens must have other reasons for obeying the law.

Social psychologists typically divide the motivation for social behavior into two basic components. The first are people's assessments of the potential gains and risks associated with engaging in various types of behavior. The influence of these judgments has already been examined. The second are the internal values that shape people's feelings about what behavior is appropriate. Behavior based upon internal values is especially important to the legal system, since the desire to be consistent with internal values leads people to obey the law voluntarily. As a consequence, to the extent that people are following laws because of their internal values, society need not put its resources into supporting agents of social control. Further, such voluntary compliance is more reliable because it does not vary as a function of the person's circumstances. When one drives up to a stop sign on a deserted road at night, one's internal values lead one to stop, even when the possibility of punishment for ignoring the law is essentially nonexistent. It has been widely suggested that the ability of legal systems to function in democratic societies depends upon the system's ability to gain such voluntary compliance with the law (Easton 1975, Engstrom and Giles 1972, Parsons 1967, Sarat 1977, Scheingold 1974).

What type of internal values might lead people to obey the law? Two types of values have been recognized as central. The first value concerns how people view the morality of engaging in illegal behaviors. Morality is concerned with personal feelings about what is right or wrong. For example, murder would still be viewed by most people as immoral, even if it were not illegal; drug use, in contrast, is regarded by many people as moral, even though it is illegal. The second value concerns how people view the legitimacy of law and legal authorities. Legitimacy denotes people's sense of the degree to which they have a personal obligation to obey social rules.

1. *Morality*

Studies generally suggest that people's views about the morality of engaging in rule-breaking behaviors shape their willingness to engage in those behaviors. Further,

the influence of morality is typically found to be greater than that of the probability of punishment for rule-breaking. A classic study illustrating this point is that of Schwartz and Orleans (1967), which evaluated the effects of a persuasive appeal to pay taxes on taxpayers' behavior. Citizens received either no appeal, an appeal based on the risk of being punished, or an appeal based on the immorality of cheating. The influence of the appeal was established by examining the level of taxes paid before and after the appeal. In the control group which received no appeal, the average change in taxes paid was -$13 dollars. In the sanction-based punishment group, the average amount paid increased $181. In the appeal to moral conscience group, the average payment increased $804.

Other studies support the argument that moral values influence law-related behavior, and that they have a greater influence than do variations in the threat of punishment. Tyler (1990) compared the influence of sanction threats and moral judgments about law-breaking on people's everyday obedience to the law. He found that moral values significantly influenced law-related behavior and were more influential than were variations in the perceived risk of being caught and punished for wrongdoing. Paternoster (1989) also examined the influence of sanction threats and moral values on everyday law-related behavior and found that moral values were the primary factor shaping people's behavior. And Grasmick and Green (1980) and Grasmick and Bursik (1990) found that moral values were more central to people's law-related behavior than were their concerns about sanctions.

Paternoster and Simpson (1996) also found that moral values had an important influence on whether or not people committed corporate crimes. Their study involved interviews with business students and executives who were given hypothetical situations in which they might break laws. They found that moral considerations exerted a 'powerful and independent source of social control' (p. 579). They also found that moral judgments constrained the operation of risk judgments. If people felt that law-breaking was immoral, considerations of the gains and losses associated with law-breaking became 'virtually superfluous' (p. 549) to their decisions about whether or not to follow the law. On the other hand, if people did not feel that law-breaking was immoral, their law-breaking behavior was linked to their risk judgments.

The degree to which moral values support the law depends upon the degree to which the behaviors that the law defines as illegal correspond to public views about what is right and wrong. When these two elements coincide, people's motivation to obey the law should be especially strong, since law-breaking is then not only illegal, but also immoral. Robinson and Darley (1995) systematically examine the relationship between formal law and public views about what is right or wrong and demonstrate that there are a number of discrepancies. Such discrepancies, to the degree that people are aware of them, should lessen people's motivation to comply with the law.

Studies of public opinion about the law suggest that the most persistent sources of public discontent with law involve public perceptions that: (a) too many criminals go unpunished due to legal 'technicalities' such as the insanity defense and the exclusionary rule, and (b) the sentences given for crimes are too lenient. In both cases, studies suggest that public views exaggerate the discrepancy between public conceptions of what is morally appropriate and what actually occurs.

One compliance-enhancing strategy available to the legal system is public education about what the law actually says and what the courts actually do. Such a strategy may be beneficial in increasing public compliance with the law by showing people that the law is more consistent with their moral values than they think. However, the work of Robinson and Darley (1995) makes it clear that if the public accurately understood the totality of the criminal and civil law, they would recognize a number of discrepancies between public morality and the formal law. Hence, any long-term effort to bring law and public morality together must involve some efforts to reconcile true discrepancies between the law and public morality.

2. *Legitimacy*

The second internal value that shapes behavior toward the law is the legitimacy of the law and of legal authorities. A law has legitimacy when others feel obligated to obey it. A legitimate authority is one which is entitled to have its decisions and rules accepted and followed by others. In the case of law, people feel a personal responsibility to comply voluntarily with those laws that are created and enforced by legitimate legal authorities. In contrast to morality, legitimacy is a general acceptance of the right of the law to dictate appropriate public behavior. When authorities possess legitimacy, they are better able to regulate effectively the behavior of citizens.

Tyler (1990) examined the influence of views about the legitimacy of the law on citizens' everyday compliance with the law. He found that legitimacy had a significant influence on the degree to which people obeyed the law. Further, that influence was greater than were the effects of variations in the estimated probability of punishment. He has already asserted that legitmacy is a key value.

Studies suggest that the key to developing and to maintaining the legitimacy of law and of legal authorities lies in citizens' evaluations of the procedures through which legal rules are created and implemented. Studies of local legal authorities, such as the police and local courts, make it clear that the

central element that shapes public evaluations of their legitimacy are evaluations of the fairness of their procedures (Tyler 1990; Tyler et al. 1989). This is also true of citizens' evaluations of national level authorities, including the Supreme Court (Murphy and Tanenhaus 1969, Tyler 1993, Tyler and Mitchell 1994) and Congress.

The key to effective social regulation is recognizing that people's experiences with legal authorities not only shape their immediate compliance behavior, but also change their views about law and legal authorities in ways that lead to lasting compliance with the law. Studies indicate that experiencing fair procedures both leads to greater adherence to third-party decisions over time (Pruitt et al. 1993) and encourages long-term obedience to the law (Paternoster et al. 1997, Sparks et al. 1996). When legal authorities act in fair ways, they encourage the immediate acceptance of their decisions and enhance the values that lead to longterm compliance with the law.

3. *The Law-abiding Society*

The distinction between risk/gain estimates and internal values highlights the possibility of two types of legal culture. The first is a culture that builds public compliance on the basis of people's concerns about the possibility of being caught and punished for wrongdoing. Such a deterrence-based society depends upon the ability of legal authorities to create and maintain a credible threat of punishment for wrongdoing. The studies outlined demonstrate that, while deterrence influences law-related behavior, the social context of democratic societies makes it difficult for authorities to engage in the levels of surveillance needed to sustain a viable legal system.

To be viable, a legal system requires that citizens' behavior be influenced by internal values that promote compliance most of the time. Democratic societies depend heavily on people's self-regulatory motivations. Both public views about the morality of engaging in illegal behaviors and widespread feelings of obligation to obey legitimate legal authorities are central to such voluntary compliance with the law. Internal values, then, are critical in creating a society based upon the willing consent and cooperation of citizens. That cooperation develops from people's own feelings about appropriate social behavior, and is not linked to the risks of apprehension and punishment that people estimate to exist in their social environment.

A law-abiding society cannot be created overnight through changes in the allocation of resources within government agencies, because such changes only alter the expected gains and/or risks associated with compliance. It depends upon the socialization of appropriate social and moral values among children and the maintenance of those values among adults. Evidence suggests that the core elements in the creation and maintenance of such social values are: (a) the belief that law is consistent with public views about morality, and (b) the judgment that legal authorities follow fair procedures in exercising their authority.

See also: Crime: Knowledge about and Prevalence; Criminal Defenses; Criminal Law and Crime Policy; Deterrence; Law: Defense of Insanity; Prisons and Imprisonment; Punishment: Social and Legal Aspects

Bibliography

Easton D 1965 *A Systems Analysis of Political Life*. Wiley, New York

Engstrom R L, Giles M W 1972 Expectations and images: A note on diffuse support for legal institutions. *Law and Society Review* **6**: 631–6

Fuller L 1971 Human interaction and the law. In: Wolff R P (ed.) *The Rule of Law*. Simon and Schuster, New York

Grasmick H G, Bursik R J 1990 Conscience, significant others, and rational choice. *Law and Society Review* **24**: 837–61

Grasmick H G, Green D E 1980 Legal punishment, social disapproval, and internalization as inhibitors of illegal behavior. *Journal of Criminal Law and Criminology* **71**: 325–35

MacCoun R J 1993 Drugs and the law: A psychological analysis of drug prohibition. *Psychological Bulletin* **113**: 497–512

Murphy W F, Tanenhaus J 1969 Public opinion and the United States Supreme Court. In: Grossman J B, Tanenhaus J (eds.) *Frontiers in Judicial Research*. John Wiley, New York

Nagin D S 1998 Criminal deterrence research at the outset of the twenty-first century. *Crime and Justice: A review of research*. **23**: 1–42

Nagin D S, Paternoster R 1991 The preventive effects of the perceived risk of arrest: Testing an expanded conception of deterrence. *Criminology* **29**: 561–87

Parsons T 1967 Some reflections on the place of force in social process. In: Parsons T (ed.) *Sociological Theory and Modern Society*. Free Press, New York

Paternoster R 1987 The deterrent effect of the perceived certainty and severity of punishment. *Justice Quarterly* **4**: 173–217

Paternoster R 1989 Decisions to participate in and desist from four types of common delinquency. *Law and Society Review* **23**: 7–40

Paternoster R, Brame R, Bachman R, Sherman L W 1997 Do fair procedures matter? *Law and Society Review* **31**: 163–204

Paternoster R, Iovanni L 1986 The deterrent effect of perceived severity. *Social Forces* **64**: 751–77

Paternoster R, Saltzman L E, Waldo G P, Chiricos T G 1983 Perceived risk and social control: Do sanctions really deter? *Law and Society Review* **17**: 457–79

Paternoster R, Simpson S 1996 Sanction threats and appeals to morality: Testing a rational choice model of corporate crime. *Law and Society Review* **30**: 549–83

Pruitt D G, Peirce R S, McGillicuddy N B, Welton G L, Castrianno L M 1993 Long-term success in mediation. *Law and Human Behavior* **17**: 313–30

Robinson P H, Darley J M 1995 *Justice, Liability, and Blame*. Westview, San Francisco

Robinson P H, Darley J M 1997 The utility of desert. *Northwestern University Law Review* **91**: 453–99

Ross H L 1982 Deterring the drinking driver: Legal policy and social control. Lexington Books, Lexington, MA

Sarat A 1977 Studying American legal culture. *Law and Society Review* **11**: 427–88
Scheingold S A 1974 *The Politics of Rights*. Yale University Press, New Haven, CT
Schwartz R D, Orleans S 1967 On legal sanctions. *University of Chicago Law Review* **34**: 27–300
Sparks R, Bottom A, Hay W 1996 *Prisons and the Problem of Order*. Clarendon Press, Oxford
Tyler T R 1990 *Why People Obey the Law*. Yale University Press, New Haven, CT
Tyler T R 1993 Legitimizing unpopular public policies: Does procedure matter? *Zeitschrift für Rechtssoziologie* **14**: 47–54
Tyler T R 1994 Governing amid diversity. *Law and Society Review* **28**: 701–22
Tyler T R, Casper J D, Fisher B 1989 Maintaining allegiance toward political authorities. *American Journal of Political Science* **33**: 629–52
Tyler T R, Mitchell G 1994 Legitimacy and the empowerment of discretionary legal authority. *Duke Law Journal* **43**: 703–815

T. R. Tyler

Comprehension, Cognitive Psychology of

1. Definitions and Forms of Comprehension

'Comprehension,' often used interchangeably with 'understanding,' can generally be defined as the solution of the problem of 'the "how" and "why" of the connections observed and applied in action' (Piaget 1978), but like many other psychological terms, it is used differently by different authors. The term means both the process (act of comprehending) and the product (an achieved state of comprehension) with more emphasis on one or the other. More importantly, comprehension is characterized as either high-cost or low-cost cognition. In other words, the core process of comprehension is regarded either as controlled-effortful or automatic-effortless, and is contrasted to other terms in different ways accordingly. For instance, according to Collins et al. (1980), understanding is a process of progressive refinement of models, and involves a variety of problem-solving strategies (e.g., rebinding variables, questioning a default assumption), as in solving a crossword puzzle. In contrast, Kintsch (1998) assumes comprehension to be a process in which 'the majority of elements are meaningfully related to one another' through automatic and effortless spreading activation. Thus, to comprehend means to adopt a coherent and plausible interpretation of pieces of information obtained from both outside and inside the mind (i.e., to satisfy multiple constraints), but it may or may not be cognitively demanding.

It is sometimes useful or even necessary to distinguish a few forms or levels of comprehension. The depth of understanding may be expressed only in relative terms ('deeper' vs. 'shallower'), but a frequent qualitative distinction is whether comprehension is explicit or implicit (Greeno 1983). For instance, finding a coherent and plausible interpretation for a target object or phenomenon produces differentiated and appropriate predictions about the target but may not always lead to explanations for it. Such forms are called implicit understanding. Another dimension differentiating levels of comprehension is whether the interpretation involves a causal device or only an empirical generalization. For example, a patient's recovery from beriberi by consuming barley instead of rice can be explained either by referring to the well-known effectiveness of the altered diet or by referring to the causal device, that is, that barley contains considerable Vitamin B1, the lack of which causes beriberi. Moreover, even among causal explanations, some are considered to reveal a deeper understanding than others. As indicated by Miyake (1986), an explanation at one level is what needs to be explained at the next deeper level. We can ask further, for instance, why the lack of Vitamin B1 causes beriberi.

2. Processes of Comprehension

Comprehension is more complex and interactive than schema application, in which a relevant schema is triggered by one of its key words or headers and imposed on the given pieces of information. However, much of comprehension can be induced within a short period of time and often with little effort. Kintsch (1988, 1998) proposes the two-stage Construction–Integration (CI) model of comprehension to capture these features. The first stage of construction is for the generation of a variety of inferences (activation of related ideas through association). The second stage of integration is multiple constraint–satisfaction through spreading activation. As we will discuss below, this model fits especially nicely with the comprehension of ordinary (not experimentally deformed) discourse, because readers or listeners spontaneously and readily generate inferences as the discourse proceeds, and integrate inferred as well as presented pieces of information very rapidly.

However, both inference-generating and coherence-seeking processes may be time-consuming and effort-intensive for other tasks of comprehension, such as scientific inquiry. Those tasks induce what Hatano (1998) calls comprehension activity. Comprehension activity often offers multiple interpretations at one time, and compares their plausibility carefully by deriving and testing predictions from each of the interpretations being considered. It is not a process in which one holds onto the first idea that comes to mind or one waits for a single interpretation to emerge

through spreading activation. A comprehender may feel uncertainty or puzzlement during this enterprise.

Plausible interpretations that are to be examined can be induced in a variety of ways both in low-cost and high-cost comprehension. Sometimes they are derived directly from prior knowledge including schemas. Other times a comprehender relies on interpretation-generation heuristics, such as modifying a parameter and observing what occurs. Still other times, candidate interpretations are brought about by analogies.

When people encounter a novel object or phenomenon, they may try to understand it by likening it to a highly familiar object or phenomenon. For example, at the time of Iraq's invasion into Kuwait, many Americans (partly through the campaign by the government) likened Saddam Hussein to Hitler, and thus justified their bombing of Iraq. Interestingly, some objected to the military action by analogizing the Persian Gulf War to the Vietnam War. This illustrates how people try to understand a novel phenomenon by referring to the instances in their past experience that they think are most similar, and how their reactions to the phenomenon are mediated by their analogy (Holyoak and Thagard 1995).

The analogies people make may involve structurally inaccurate mapping, and induce biased reasoning. Even a highly productive scientist like Kepler was misled by analogies at the beginning of his career (Gentner et al. 1997). However, analogy-based interpretations are often highly useful in comprehension activity, both in science (Dunbar 1995) and everyday cognition (Hatano 1998).

3. *Experimental Studies on Comprehension*

In this section we review three categories of phenomena of comprehension that have been studied extensively. They are grouped in terms of inputs: a set of observations, a set of connected sentences, and patterns of actions and their consequences.

3.1 *Understanding as Identification Based on Observations*

Recognizing something that is ambiguous as an instance of a concept or schema constitutes understanding, because it is adopting a coherent and plausible interpretation of a set of observed features of the target based on relevant concepts or schemas. Once we understand the target object or phenomenon, we can make sense of the observations and can remember them much more easily (e.g., Bransford and Johnson 1972). Moreover, we can recognize the target's essence or underlying process, and decide reasonably how to treat it.

The diagnosis of disease by reading X-ray films is a good example of such tasks of categorization or identification. The same disease does not always produce the same appearance on the film, and there can be several diseases that have similar appearances. Therefore reading X-ray films involves many inferences and requires a constraint–satisfaction process to reach a diagnosis. This is undoubtedly a process heavily dependent on prior knowledge, in which experts and novices would be expected to differ.

Lesgold et al. (1988) compared the abilities of expert and novice radiologists to make accurate diagnoses by reading x-ray films. The participants were asked to report a diagnosis after being shown chest X-ray films just briefly, and their protocols were analyzed. It was found that compared with the novices, the experts reported more varied findings (i.e., attributed specific properties to the film or the patient), their reasoning chains were longer, and a greater number of their findings were connected to at least another finding.

However, experts and novices do not make the same observations and then interpret them differently based on their different amounts of knowledge in natural settings. In a further study, the participants were required to examine the film as long as desired, while thinking aloud, and to dictate their formal diagnostic reports. Qualitative analysis of the protocols revealed that the experts exhibited more flexibility and were able to better tune their schemas to the perception of features in the film that were relevant to the specific case with which they were working. They could discount most of the obvious film features that the novices were likely to be misled to judge as evidence of disease symptoms, by using a mixture of technical knowledge about how the films were made and a better developed perception of the features.

Diagnosing films requires apparently simple outputs, that is, whether there are pathological changes, and what kind of disease the changes are due to. However, the process of understanding is more dynamic and interactive than the top-down process of applying the relevant schema stored in memory or the bottom-up process of summarizing and abstracting collected pieces of information. Perception, inference, and memory retrieval are inextricably connected. This is probably true for the understanding of other targets.

3.2 *Discourse Comprehension*

Comprehension of discourse (including both narrative and expository text) is one of the most popular topics in the study of understanding. Here, the input information is in the form of a set of spoken or written sentences, and the comprehender's job is to form a coherent whole from individual sentences. Versions of schema theory (e.g., Schank and Abelson 1977) were predominant in earlier studies. However, as pointed

out by Kintsch (1998), such top-down views have limitations: Discourse comprehension is interactive if not bottom-up, and it is much more flexible and context-sensitive than predicted from the schema theory.

Most contemporary investigators assume that to understand a story (i.e., a narrative text) is to build a microworld in which the described events are likely to occur. Kintsch (1998) calls this microworld the 'situation model.' In addition to the textbase that consists of those elements and relations that are explicitly described in the text itself, the situation model involves pieces of information that make the textbase coherent, complete, and interpretable in relation to prior knowledge. These pieces of information are inferred from both the text and prior knowledge. Multiple situation models may be constructed from the same text (e.g., differing in the coverage or the mode of representation).

When people understand a story, they make inferences that are not given in the text. They sometimes infer that an event occurred, even though it is not written explicitly (e.g., from the sentence, 'He took a subway downtown,' the reader infers that the subject bought a ticket); other times, the reader infers how an event occurred, although the text does not give any concrete details. (From the sentence, 'He bought a ticket with coins,' one infers, 'He used a vending machine.') In other situations, readers connect two adjacent propositions. (From the two sentences, 'Taro wanted to buy some tuna for sashimi at a discounted price. He went to the fish-market,' one infers, 'Taro went to the fish-market to buy tuna there.') How many inferences are generated spontaneously may depend on the readers, texts, and modes of reading. Graesser et al. (1994) claim that the inferences that are needed to explain why given events occur and to establish the coherence of the text tend to be induced spontaneously as the text is being processed. How many inferences are based on conscious and deliberate attempts to build consistent and detailed situation models also depends on several factors. Kintsch (1998) proposes that a coherent situation model is built based on these presented and inferred propositions through spreading activation. However, studies using experimentally designed defective texts (e.g., in Collins et al. 1980) have revealed that text comprehension may require a number of effortful attempts to instantiate, coordinate, and even insert pieces of information.

Almost everyone is very experienced in discourse comprehension, because it is a major medium of human communication. However, knowledge about the topic of the discourse still makes a difference, especially in the generation of inferences. As a result, more knowledgeable people can build a richer situation model than less knowledgeable ones. They learn more, especially when the text is less coherent and thus requires comprehension activity on the part of the comprehenders. Schneider et al. (1989) demonstrated that students who knew a lot about soccer not only remembered the details of a given story about soccer better, but also made more inferences and recognized contradictions in the text more often than their agemates who knew little about soccer.

3.3 Comprehension of Procedures and their Target Entities

Just knowing the procedures required to achieve goals differs from knowing how and why they work. By possessing knowledge to build mental models of the procedure-target combination (what the target is like and how it is changed by various procedures), we can grasp how and why these procedures work, modify the procedures flexibly, and even invent new procedures. Building such models is usually accomplished through a process of comprehension.

Kieras and Bovair (1986) showed that having a 'device model' (a mechanical version of the procedure-target model) that describes the internal mechanism of a piece of equipment enhances not only the participants' learning of how to operate the equipment but also their inventing of more efficient procedures than the ones they have been taught. Their experiments compared two groups, that is, a model group and a rote group. The participants in the model group were presented with the device model in a fantasy context before being given procedural instruction, whereas those of the rote group learned the identical operating procedures of the equipment by rote. It was observed that the model group learned the procedures faster, retained them more accurately, and executed them more efficiently than the rote group. More interestingly, the model group could simplify inefficient procedures and infer new procedures much more often than the rote group, indicating the benefit of their deeper understanding. An additional experiment showed that what was critical was the knowledge about the internal structures of the components of the equipment, not the context of fantasy.

In the experiments by Kieras and Bovair (1986) described above, an externalized form of the standard device model was presented and explained. Active human minds often ask the questions 'how' and 'why,' even when they know that a given procedure works, and may construct a procedure-target model through comprehension activity. For example, the recipe for bonito sashimi sometimes induces comprehension activity, because it is not just the slicing of raw fish, but involves the steps of roasting the skin-covered surface quickly at a high temperature and putting the roasted side into ice-water (Hatano 1998). People may wonder why such a recipe is required, and if they comprehend it, they can modify it flexibly when they have to meet a different set of constraints, for example, when there is no ice or no source of searing heat. To understand a given procedure is, as in this example, to find a

plausible interpretation for how and why it works. In other words, it permits the comprehender to infer what kind of effect each step of the procedure has on the target entity of the procedure. Thus, trying to generate to oneself an explanation for how a procedure is applied tends to enhance its understanding (Chi et al. 1989).

We sometimes engage in an even more ambitious attempt to reconstruct a procedure from its product in our daily lives, which is similar to the scientific inquiry that tries to model a process from its observed outcomes. For example, we may want to reproduce an unfamiliar but tasty dish (Hatano 1998). In an experiment, college students who were given a slice of perch mousse and asked to find how it had been made by tasting it (more specifically, to indicate its recipe) seemed to engage in comprehension activity. They often proposed ideas that they eventually discarded themselves. For example, they offered a number of possible ingredients, but rejected some of them both from their taste and knowledge about cooking. They expressed their feelings of uncertainty or puzzlement, as well as insight.

Although there have been few studies, prior knowledge seems critical for the understanding of procedures. First, prior knowledge enables comprehenders to think of hypotheses. Second, it serves as constraints in the pursuit of coherence and plausibility.

4. *Comprehension Monitoring and Repair*

To understand something deeply, it is necessary to assess one's extent of understanding, and to take appropriate action based on that assessment. However, whether one understands is not obvious (whereas whether one has solved a problem can be recognized more clearly). To put it differently, every task of comprehension is inevitably ill defined. Therefore how well people can monitor their own understanding and how this monitoring ability develops are interesting questions in the study of comprehension.

Research on comprehension monitoring, following the pioneering work by Markman (1977), has shown that younger children tend to miss the insufficiency or inconsistency of a given message more often than older children or adults, but another line of research on metacomprehension has revealed that even college students tend to have this 'illusion of comprehension' (e.g., Glenberg and Epstein 1985). College students often believe that they understand a given text, though in fact they do not, at least as assessed by a multiple-choice test. This suggests that people have a general tendency to overestimate the adequacy of their own comprehension. People must be selective in directing prolonged comprehension activity, because it requires considerable time and effort. The illusion of comprehension guards them from engaging in the activity too often or in too diverse domains.

This tendency to overestimate the extent of comprehension may be due to the human ability to readily repair comprehension by supplementing pieces of missing information. Ackerman et al. (1991) observed that recovering coherence by adding a piece of new information by inference is frequent among both children and adults. People seldom suffer from the lack of comprehension for an extended period of time. This seems adaptive in the 'wild environment' in which quick decision-making and prompt action are required (Toda 2000).

As suggested by Schneider et al. (1989), it is likely that the more and less knowledgeable people differ in their ability to estimate accurately the degree of coherence among pieces of information and the adequacy of their comprehension. Humans seem to have a general tendency to make sense of their observations and actions, in other words, to derive coherent and plausible interpretations. However, judging whether the achieved understanding is adequate is a totally different issue. The criterion of adequacy may vary from context to context. Therefore only those who have performed many similar ill-defined tasks of understanding in the target domain can assess it accurately.

5. *Comprehension as a Paradigm of Cognition*

As Kintsch (1998) aptly points out, many forms of cognition can be regarded as a process of constraint–satisfaction, and thus a process of understanding. Comprehension as a process thus represents cognition in diverse areas, as does problem-solving through searching a problem space. However, as we have seen, the process of constraint–satisfaction is not necessarily executed through spreading activation, nor is it automatic–effortless. The process of understanding is in a sense to solve an ill-defined problem, that is, to find a 'reasonable' solution to a problem for which only an incomplete set of information is given, using a variety of constraints. A set of constraints is provided by prior knowledge, but, in addition, interpretations of many related observations can be used as mutual constraints, as in solving a crossword puzzle. It is a process of maximizing coherence among the pieces of information, whether they are presented, stored, or inferred.

In terms of the function, understanding is primarily to make a best guess about what the external world is like, based on observations, and constitutes a core function of the mind, more specifically, of central systems (Fodor 1983). Understanding enables us to build enriched, stable, coherent, and usable representations of the world, which often serve as the basis for solving problems, that is, for changing the world as we desire. In this sense, understanding is an investment in an unknown future, and the human tendency to seek deep understanding, spending much time and effort,

can be considered adaptive, even though it seems to be just intellectual luxury.

See also: Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Literary Texts: Comprehension and Memory; Logic and Cognition, Psychology of; Macrostructure in Discourse Comprehension, Psychology of; Narrative Comprehension, Psychology of; Piaget's Theory of Human Development and Education; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning; Problem Solving (Everyday), Psychology of; Problem Solving (Joint), Psychology of; Sentence Comprehension, Psychology of; Text Comprehension: Models in Psychology

Bibliography

Ackerman B P, Jackson M, Sherrill L 1991 Inference modification by children and adults. *Journal of Experimental Child Psychology* **52**: 166–96

Bransford J D, Johnson M K 1972 Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior* **11**: 717–26

Chi M T H, Bassok M, Lewis M, Reitmann P, Glaser R 1989 Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science* **13**: 145–82

Collins A, Brown J S, Larkin K M 1980 Inference in text understanding. In: Spiro R J, Bruce B C, Brewer W F (eds.) *Theoretical Issues in Reading Comprehension*. Erlbaum, Hillsdale, NJ

Dunbar K 1995 How scientists really reason: Scientific reasoning in real-world laboratories. In: Sternberg R J, Davidson J E (eds.) *The Nature of Insight*. MIT Press, Cambridge, MA

Fodor J 1983 *The Modularity of Mind*. MIT Press, Cambridge, MA

Gentner D, Brem S, Ferguson R, Wolff P, Markman A B, Forbus K 1997 Analogy and creativity in the works of Johannes Kepler. In: Ward T B, Smith S M, Vaid J (eds.) *Creative Thought: An Investigation of Conceptual Structures and Processes*. American Psychological Association, Washington, DC

Glenberg A M, Epstein W 1985 Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition* **11**: 702–18

Graesser A C, Singer M, Trabasso T 1994 Constructing inferences during narrative text comprehension. *Psychological Review* **101**: 371–95

Greeno J G 1983 Forms of understanding in mathematical problem solving. In: Paris SG, Olson G M, Stevenson H W (eds.) *Learning and Motivation in the Classroom*. Erlbaum, Hillsdale, NJ

Hatano G 1998 Comprehension activity in individuals and groups. In: Sabourin M, Craik F, Robert M (eds.) *Advances in Psychological Sciences. Vol. 2: Biological and Cogntive Aspects*. Psychology Press, Hove, UK

Holyoak K J, Thagard P 1995 *Mental Leaps: Analogy in Creative Thought*. MIT Press, Cambridge, MA

Kieras D E, Bovair S 1984 The role of a mental model in learning to operate a device. *Cognitive Science* **8**: 255–74

Kintsch W 1988 The use of knowledge in discourse processing: A construction-integration model. *Psychological Review* **95**: 163–82

Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK

Lesgold A, Glaser R, Rubinson H, Klopfer D, Feltovich P, Wang V 1988 Expertise in a complex skill: Diagnosing X-ray pictures. In: Chi M T H, Glaser R, Farr M J (eds.) *The Nature of Expertise*. Erlbaum, Hillsdale, NJ

Markman E M 1977 Realizing that you don't understand: A preliminary investigation. *Child Development* **48**: 986–92

Miyake N 1986 Constructive interaction and the iterative process of understanding. *Cognitive Science* **10**: 151–77

Piaget J 1978 *Success and Understanding*. Routledge and Kegan Paul, London

Schank R C, Abelson R P 1977 *Scripts, Plans, Goals and Understanding: An Inquiry to Human Knowledge Structures*. Erlbaum, Hillsdale, NJ

Schneider W, Korkel J, Weinert F E 1989 Domain-specific knowledge and memory performance: A comparison of high- and low-aptitude children. *Journal of Educational Psychology* **81**: 306–12

Toda M 2000 Emotion and social interaction: A theoretical overview. In: Hatano G, Okada N, Tanabe T (eds.) *Affective Minds*. Elsevier, London, UK

G. Hatano

Compression of Morbidity

The compression of morbidity paradigm (Fries 1980) envisions reduction in cumulative lifetime morbidity principally through primary prevention, by postponing the age of onset of morbidity to a greater amount than life expectancy is increased, largely by reducing the health risks which cause subsequent morbidity and disability. Recent data document slowly improving age-specific health status for seniors, indicate that postponement of the onset of disability by at least 10 years is feasible, and prove the effectiveness of some specific lifestyle modification interventions by randomized controlled trials.

The health of senior persons is the largest of all health problems in the developed nations. Over 80 percent of all illness, morbidity, mortality, and medical costs are concentrated in the years after age 65. The illness burden is mostly made up of chronic illness, complicated by the frailty of increasing age. In the twentieth century we saw a transfer of the illness burden from the acute infectious diseases prevalent in 1900 to ascendancy of chronic disease by mid-century, to a decline in major chronic illnesses, led by car-

diovascular disease, and now to the slow transition from chronic disease *per se* to the associated problems of aging (Fries 1980).

1. An Overview of the Aging Process

1.1 Mortality Changes over Time

Changes in survival curves in developed nations over the twentieth century are instructive, since they lead us beyond a simplistic, if popular, notion of ever-increasing life expectancy. Life expectancy from birth is affected most strongly by changes in infant mortality rates and in deaths in early and middle life. Successive survival curves (Fig. 1) have become more rectangular, as marked reduction in deaths at early ages flatten the initial part of the curve. There are now few deaths prior to age 50 or 60. At the same time, the downslope of the curves has become steeper, and the insertion point has changed little (Fries and Crapo 1981). A natural barrier to biologic longevity may be visualized through these successive curves. In the USA, life expectancy from birth increased from 47 to 76 years in the twentieth century. However, life expectancy from advanced ages has changed relatively little. From 1980 to 1998, life expectancy for females from age 65 increased only by 0.7 years in contrast to more rapid rises in prior decades. Life expectancy for both sexes from age 85 remained stable at 6.1 to 6.3 years in the USA in the 18 years from 1981 (Kranczer 1999). Projections for future increases vary widely from high (Manton et al. 1991) to low (Olshansky et al. 1990). Nevertheless, with increasing size of successive birth cohorts and with increased survival to age 65 or age 85 the absolute numbers of senior citizens will increase very substantially in coming years.

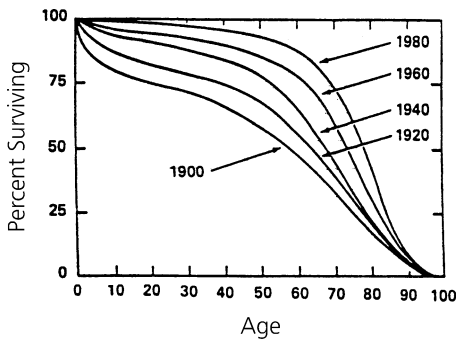

Figure 1
The rectangularization of survival curves. Successive curves, at 20 year intervals, illustrate (a) the flattening of the early part of the curves from decreases in premature deaths, (b) the steepening of the downslope as natural life limits are approached, and (c) the nearly constant insertion site of the curves

1.2 Declines in Organ Reserve Function with Age

Data from longitudinal studies of aging show a consistent decline in the maximum function of the various organs with age, the decline being essentially linear at a rate of 1.5 percent per year after age 30. Data on maximal performance, such as world record marathon times, similarly show a nearly linear decline with age at the same rate (Fries and Crapo 1981).

Physiologic normal values, however, remain constant with increasing age. Normal pH, blood chemistry, white count, and other homeostatic values do not vary over the lifespan, representing the internal physiologic environment essential for cellular function. However, with the linear decrease in organ reserve in multiple organs, the ability to respond physiologically to a perturbation decreases exponentially. As a result, mortality rates increase exponentially, with a doubling of mortality rates each seven or eight years after age 30 (Fries and Crapo 1981).

With these realities of human aging comes a paradox: the decline in organ reserve is inevitable, yet organ reserve can be increased quite readily, at almost any age (Rowe 1999). For example, an increase in exercise levels can increase cardiopulmonary reserve very substantially, even at advanced ages (Stewart et al. 1993, Fries et al. 1996).

1.3 Enhanced Personal Autonomy and Modifiable Determinism

In philosophic writings, a classical conflict has been between the advocates of free will and those of determinism, and this conflict now exists in modern dress. Determinism is represented by molecular genetics, with the sometime notion that your health over the lifespan is ultimately determined only 'by your genes.' Free will is represented by the advocates of health promotion, seeking voluntary changes in behavioral risk factors, such as lack of exercise, cigarette smoking, obesity, and excessive dietary fat; changes which can enhance organ reserve, preserve function, and extend life. In this view, health requires that you 'take care of yourself.' The tension between these two paradigms is complex and sometimes uneasy.

2. The Compression of Morbidity

2.1 The Hypothesis and the Paradigm

The compression of morbidity hypothesis was introduced in 1980 (Fries 1980) and has become a major paradigm underlying health improvement programs and health improvement policies directed at more successful aging (Fries et al. 1998). Most morbidity results from chronic processes of disease and frailty and is concentrated in the years prior to death. The

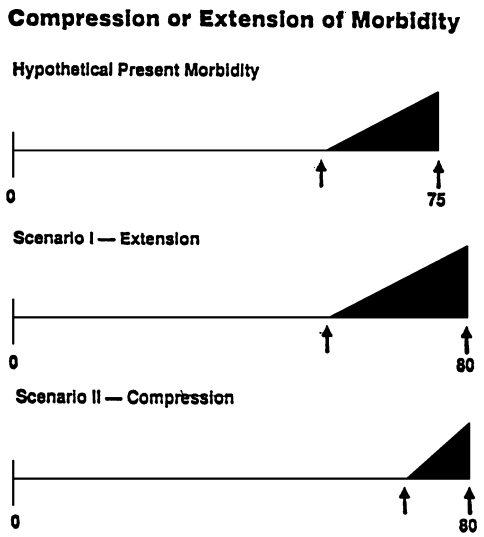

Figure 2
Compression of morbidity. The top line provides a graphic estimate of current lifetime morbidity, disability, or cost, expressed as a shaded area between onset of disability at an average age of 55 years and an average age of death of 75 years. Possible future scenarios are shown below. The second line represents the extension of morbidity if life expectancy is increased but the onset age of morbidity is not. The third shows a 'shift to the right'

goal is compression of morbidity between its time of onset, currently averaging 55 years, and the age of death, currently averaging 76 years, as diagrammed in Fig. 2. A crescendo of increasing morbidity occurs, on average, over this period.

With social and scientific progress it is likely that both the age of onset of infirmity and the age at death will increase over time. At issue is the relative movement of the two points. If mortality decreases predominate there may be more morbidity in the typical life, the so-called 'failure of success' (Gruenberg 1977, Verbrugge 1984, Scheider and Brody 1983), as shown in the second line of the figure. If postponement of the age of onset of morbidity predominates, then morbidity is compressed and the average illness burden is less, the period of adult vigor is prolonged, life quality is improved, and the need for medical care and associated costs may be reduced.

For many, the ideal life is one vigorous and vital over the lifespan, with a terminal collapse of physical and mental function limited to a few months or years immediately preceding death (Fries and Crapo 1981). A central issue, when considering the eventual impact of new scientific discoveries or behavioral changes, is whether they may add years of health or add years of illness to the average life. If morbidity is compressed by a new advance then the advantage is clear; if the process of dying is prolonged then major ethical issues arise (Verbrugge 1984).

2.2 *Trends in Morbidity Compression*

Since the early 1980s there has been some compression of morbidity in the elderly, as documented by a number of recent studies (Manton et al. 1997, Nusselder and Mackenbach 1996, Allaire et al. 1999). The major chronic illnesses, including heart disease, cancer, and stroke, are about 70 percent preventable. The true causes of death are not the chronic diseases themselves but the underlying causes of these diseases, which are led by cigarette smoking, sedentary living, obesity, and diets high in saturated fat and low in complex carbohydrates (McGinnis and Foege 1993). The potential for postponement of morbidity is now much better understood.

Freedman and Martin (1998), among others, have shown significant age-specific functional improvement in seniors over a recent seven-year period. Cross-sectional studies have suggested compression of morbidity in favored groups, such as those with higher levels of education (Leigh and Fries 1994), higher socioeconomic status (House et al. 1990), and those with fewer lifestyle health risks (Vita et al. 1998). Daviglus and colleagues (1998) showed substantial decreases in Medicare costs for those with few health risk factors in midlife. Reed and colleagues (1998) related healthy aging to prospectively determined health risks. Allaire et al. (1999) showed a reduction of disability of 25 percent across successive generations in the Framingham study. Clearly, social and behavioral health risks make a very major contribution to both morbidity and mortality. Selected medical advances also evidently contribute to compression of morbidity, from established therapies such as total joint replacement and cataract extraction to any other treatment which makes a major contribution to life quality. The central measure of success at compression of morbidity is reduction in average cumulative lifetime morbidity (Vita et al. 1998, Wang et al. 1999, Hazzard 1997). Hitt et al. (1999) analyzed morbidity in centenarians and found it to be mild, with a rapid terminal decline.

3. *How Long May the Onset of Morbidity be Postponed?*

Substantial numbers of aging seniors in two cohorts have been followed longitudinally to identify the factors which postpone onset of morbidity, the magnitude of the postponement, and the effects of lifestyle health risks, using disability as a definable and measurable surrogate for morbidity and cumulative

disability as a surrogate for cumulative lifetime disability.

The University of Pennsylvania Study followed 1,741 university attendees studied in 1939 and 1940, surveyed again in 1962, and followed annually since 1986. Health risk strata were developed for persons at high, moderate, or low risk, based upon cigarette smoking, body mass index, and lack of exercise, and assigned by risk status in 1962 (average age 43 years). Persons with high health risks in 1962 or in 1986 had twice the cumulative disability of those in the low risk strata. This 100 percent decrease in disability rates was offset only partially by a 50 percent decrease in mortality rates in the high risk strata, documenting compression of morbidity. Onset of disability was postponed by about seven years in the low risk stratum as compared with the high stratum (Vita et al. 1998).

In the Precursors of Arthritis Study 537 senior runners and vigorous exercisers and 423 age-matched community controls were studied over 13 years with careful controls for self-selection bias. The exercising group has had less than one-half the cumulative disability of the sedentary controls and differences between groups has actually increased over time. The proportion of those disabled was also reduced by more than one-half in the exercise groups. The postponement of disability for the exercising group was 8.7 years for minimum disability and approximately 12 years for higher levels of disability (Wang et al. 1999).

While these two studies are the only ones to directly assess the magnitude of morbidity progression they are entirely consistent with the larger literature (Stewart et al. 1993, Daviglus et al. 1998, Freedman and Martin 1998, Campion 1998, Allaire et al. 1999, Reed et al. 1998).

4. *Randomized Controlled Trials of Behavioral Interventions*

Epidemiologic studies of associations, no matter how strong and consistent, ultimately require randomized study of interventions for proof of causality, and a number of these are now available (Fries et al. 1998). Health promotion programs in senior populations directed at risk prevention, health improvement, and medical cost reduction have been studied in a number of large and well-performed randomized trials, with positive results. The Bank of America randomized study of 4,500 retired subjects reduced risks by 12 percent in the intervention groups in the first year compared with controls and reduced medical care requirements by even more, confirmed by study of medical claims (Fries et al. 1993). The California Public Employment Retirement System study (Fries et al. 1994), involved 57,000 subjects in a one-year randomized trial with similarly dramatic results confirmed by claims data endpoints. Over 42 carefully designed high quality health improvement programs have now been proven effective and cost-effective at preventing illness and at reducing medical need (Fries et al. 1998).

5. *Issues for Contemplation*

The human aging process, when not prematurely stopped by trauma or disease, moves toward multiple organ system frailty (Fries et al. 1998). The immediate cause of death shifts from external toward intrinsic factors. The formally assigned 'cause of death' becomes increasingly irrelevant compared with the underlying frailty, the inability of the aging organism to withstand even a minor perturbation. With frailty, an analogy is an old curtain, rotted by the sun, where an attempt to repair a tear in one place is followed by a tear in another.

William Osler considered pneumonia to be the 'old mans' friend,' terminating the infirmities of frailty. Influenza epidemics result in an increase in death rates, but they are followed by a six-month period of death rates which are actually below expected baselines. Pneumovax and influenza vaccines can prevent a number of specific perturbations and associated attributable deaths, but only in the context of multiorgan frailty. The ultimate health benefits of successful interventions in the terminally frail may prove in substantial part illusory when so many competing risks remain.

Health policies directed at healthy aging may now be visualized quite clearly. An underlying paradigm is the compression of morbidity. Prospective longitudinal studies now link health outcomes firmly with antecedent health risk factors, providing proof of concept. Randomized controlled trials of interventions in seniors document the ability to reduce health risks, to improve outcomes, and to reduce costs, providing proof of causality. National and international initiatives are now needed to provide these benefits to populations.

See also: Aging and Health in Old Age; Caregiving in Old Age; Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Human Development and Health; Old Age and Centenarians; Social Integration, Social Networks, and Health

Bibliography

Allaire S H, LaValley M P, Evans S R, O'Connor G T, Kelly-Hayes M, Meenan R F, Levy D, Felson D T 1999 Evidence for decline in disability and improved health among persons aged 55 to 70 years: The Framingham Heart Study. *American Journal of Public Health* **89**: 1678–83

Campion E W 1998 Aging better. *New England Journal of Medicine* **338**: 1064–6

Daviglus M L, Liu K A, Greenland P et al. 1998 Benefit of a favorable cardiovascular risk-factor profile in middle age with

respect to medicare costs. *New England Journal of Medicine* **339**(16): 1122–29
Freedman V A, Martin L G 1998 Understanding trends in functional limitations among older Americans. *American Journal of Public Health* **88**(10): 1457–62
Fries J F 1980 Aging, natural death, and the compression of morbidity. *New England Journal of Medicine* **303**: 130–5
Fries J F 1993 Compression of morbidity 1993: Life span, disability, and health care costs. *Facts and Research in Gerontology* **17**: 183–90
Fries J F, Bloch D A, Harrington H, Richardson N, Beck R 1993 Two-year results of a randomized controlled trial of a health promotion program in a retiree population: The Bank of America Study. *American Journal of Medicine* **94**: 455–62
Fries J F, Crapo L M 1981 *Vitality and Aging*. W. H. Freeman, San Francisco
Fries J F, Harrington H, Edward R, Kent L A, Richardson N 1994 Randomized controlled trial of cost reductions from a health education program: The California Public Employees' Retirement System (PERS) Study. *American Journal of Health Promotion* **8**: 216–23
Fries J F, Singh G, Morfeld D, O'Driscoll P, Hubert H 1996 Relationship of running to musculoskeletal pain with age: A six-year longitudinal study. *Arthritis and Rheumatism* **39**: 64–72
Fries J F, Koop C E, Sokolov J, Beadle C E, Wright D 1998 Beyond health promotion: Reducing need and demand for medical care. *Health Affairs* **17**(2): 70–84
Gruenberg E M 1977 The failures of success. *Milbank Memorial Fund Quarterly* **55**: 3–34
Hazzard W R 1997 Ways to make 'usual' and 'successful' aging synonymous: Preventive gerontology. *Western Journal of Medicine* **167**: 206–15
Hitt R, Young-Xu, Silver M, Peris T 1999 Centenarians: The older you get, the healthier you have been. *Lancet* **354**(9179): 452
House J S, Kessler R C, Herzog A R et al. 1990 Age, socioeconomic status, and health. *Milbank Memorial Fund Quarterly* **68**: 383–411
Kranczer S 1999 US life expectancy. *Statistical Bulletin* **Oct–Dec**: 20–27
Leigh J P, Fries J F 1994 Education, gender and the compression of morbidity. *International Journal of Aging & Human Development* **39**: 233–46
Manton K G, Stallard E, Tolley H D 1991 Limits to human life expectancy: Evidence, prospects, and implications. *Population and Development Review* **17**: 603–37
Manton K G, Corder L, Stallard E 1997 Chronic disability trends in elderly United States populations: 1982–1994. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 2593–98
Mcginnis J F, Foege W H 1993 Actual causes of death in the United States. *Journal of the American Medical Association* **270**(19): 2207–12
Nusselder W J, Mackenbach J P 1996 Rectangularization of the survival curve in the Netherlands, 1950–1992. *Gerontologist* **36**: 773–82
Olshansky S J, Carnes B A, Cassel C 1990 In search of Methuselah: Estimating the upper limits to human longevity. *Science* **250**: 634–40
Reed D M, Foley D J, White L R et al. 1998 Predictors of healthy aging in men with high life expectancies. *American Journal of Public Health* **88**(10): 1463–8
Rowe J W 1999 Geriatrics, prevention, and the remodeling of G medicare. *New England Journal of Medicine* **340**: 720–21
Scheider E L, Brody J A 1983 Aging, natural death, and the compression of morbidity: Another view. *New England Journal of Medicine* **309**: 854–6
Stewart A L, King A C, Haskell W L 1993 Endurance exercise and health-related quality of life in 50–65 year-old adults. *Gerontologist* **33**: 782–9
Verbrugge L M 1984 Longer life but worsening health? Trends in health and mortality of middle-aged and older persons. *Milbank Memorial Fund Quarterly* **62**: 475–519
Vita A J, Terry R B, Hubert H B et al. 1998 Aging, health risks, and cumulative disability. *New England Journal of Medicine* **338**: 1035–41
Wang B W E, Ramey D R, Fries J F 1999 Postponed development of disability in senior runners: A controlled 13-year longitudinal study. *Arthritis and Rheumatism* **42**(9): S74

J. F. Fries

Computational Approaches to Model Evaluation

A measure of scientific advancement in a discipline is the discovery of general laws and principles that govern the phenomenon of interest. As the underlying principles are typically not directly observable, they are formulated in terms of hypotheses. Models are formal expressions of such hypotheses, whose viability must be evaluated against data generated by the phenomenon. In this article we review computational approaches to model evaluation, with a special emphasis on the inductive inference perspective. A more in-depth treatment of the topic can be found in Cherkassky and Mulier (1998) and Haykin (1999). For a discussion of some of the issues presented here from the statistical perspective (see *Model Testing and Selection*, *Theory of*, which complements the present article).

1. *Modeling as Inductive Inference*

The goal of modeling is straightforward: given observed data, identify the underlying system that generated the data. Realistically, modeling is viewed as a learning problem in which we estimate an unknown function from observed data and then use the estimated function to predict the statistics of future data. When modeling, we reconstruct an unknown function based on several sample points in the data space (local subspace) and then apply it globally, throughout the data space. This process of generalizing from a particular instance to *all* instances is referred to as inductive inference, or predictive inference given the predictive nature of the process and the goals of the modeler. It is at the heart of the scientific enterprise in modern science as well as in engineering.

1.1 Ill-posedness of the Modeling Problem

Identifying an unknown function from data samples is an ill-posed problem, meaning that obtaining a unique solution is not in general possible. The reason stems from the reality that there is rarely enough data available (or information in the data) to discriminate between models. Multiple functions may provide equally good descriptions of the data in hand. This problem is known as the *curse of dimensionality*, which states that for a function defined in high-dimensional space, it becomes increasingly difficult to estimate accurately the function as dimensionality increases, with required sample sizes increasing exponentially with dimensionality (Cherkassky and Mulier 1998). This problem of ill-posedness can be overcome only by introducing external constraints on, and/or prior information about, the data-generating function.

1.2 Model Evaluation: Finding Best Approximation to Truth

An implication of the ill-posedness problem is that the 'true' model, the one that generated the data, can never be found in practice. Accordingly, the previously defined goal of finding the true function underlying observations must be modified. A more realistic goal is to find the best possible approximation to the truth, in which the best approximation is defined as the one that generalizes (i.e., infers) best in some defined sense.

2. Bias-variance Dilemma

Formally, a model is defined as a parametric family of probability distributions defined on random variables Y, representing data. An observed sample of size n denoted by a lower case vector $y = (y_1, \ldots, y_n)$ represent realizations (measurements) of Y, which are presumed to be generated from some true distribution $g(y)$ that the scientist seeks to unravel. The scientist, using a family of distributions $f(y|\theta)$'s where θ is a parameter vector on the parameter space Θ, seeks to construct the best approximation to g within the family f.

Let $t = E(y)$ denote the population mean vector of the true distribution $g(y)$, and let $z(\theta) = (z_1, \ldots, z_n)$ denote the vector predicted by the model atθ. Also, let the quadratic function

$$L(x, z(\theta_x)) = \min_{\theta} \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - z_i(\theta))^2 \right) \quad (1)$$

denote a measure of the discrepancy or error between a vector x and the prediction vector $z(\theta)$. Note that θ_x denotes the value of the parameter that minimizes the discrepancy measure given a vector x. This function is used to measure the effectiveness or risk of the model family f in approximating the truth g. For example, $L(y, z(\theta_y))$ represents the minimized mean squared error between observed and predicted data, that is, the best possible fit the model provides given data y. Because this quantity changes from sample to sample, a better measure of model effectiveness (or risk) is the expected value of $L(y, z(\theta_y))$, which can be expressed as the sum of two terms:

$$E[L(y, z(\theta_y))] = L(t, z(\theta_t)) + E\left[\frac{1}{n} \sum_{i=1}^{n} (z(\theta_y) - z(\theta_t))^2 \right] \quad (2)$$

In the above equation E denotes the expectation (i.e., average) taken over all data samples of size n which could be realized under the true distribution g, and t is the mean vector of the true distribution $g(y)$ that has generated the data. Note that the value of the prediction vector $z(\theta_y)$ varies across samples whereas $z(\theta_t)$ is a fixed constant vector, which can be replaced by $E[z(\theta_y)]$.

The left-hand side of the above equation is the expected risk in predicting all data samples given the model family f, and thus it measures the model's generalizability. The first term of the right-hand side of the equation is the discrepancy between the true mean vector t and the model's best approximation to t, $z(\theta_t)$. Accordingly, this term captures the model's inability to accurately approximate the truth g, and is called the *bias* or approximation error. If the model family f includes g as a special case, the bias term becomes zero, and takes on a positive value otherwise. The second term is the expected variability of the prediction vector $z(\theta_y)$ around its mean $z(\theta_t)$, and is called the *variance* or estimation error. It captures the lack of information in the sample of finite size about the truth g. As such, this term decreases to zero as sample size n goes to infinity.

To achieve good generalization performance, both the bias and variance of the model family f would have to be minimized at the same time. It turns out, however, that this is not possible because the bias can be minimized by increasing model complexity, whereas increasing complexity will have the detrimental effect of increasing the variance by capturing random error in the data, thereby reducing generalization performance. Complexity refers to the flexibility inherent in a model that enables it to fit diverse patterns of data. In general, a simple model with few parameters has a large bias but a small variance, whereas a complex model (one with more parameters) has the opposite. We thus have a *bias-variance dilemma* (Geman et al. 1992), with model complexity as the intervening variable that mediates between the two.

The bias-variance dilemma is a useful conceptual framework for understanding the trade-off between approximation error and estimation error, both of which contribute to a model's generalizability. How-

ever, the generalizability measure represented by the sum of bias and variance is not directly observable, because the truth g is unknown, and therefore it must be inferred from observed data.

3. Computational Approaches to Model Evaluation

In order to make the ill-posed problem of modeling into a well-posed one, additional constraints and prior information must be incorporated into the model evaluation process. Computational approaches to the problem differ from one another in how this is accomplished. *Structural risk minimization* (Vapnik 1992), *regularization theory* (Tikhonov 1963) and *minimum description length* (Rissanen 1989) are three major inductive methods of model evaluation.

Briefly, the structural risk minimization method imposes a constraint requiring that, among a set of candidate models, the one that minimizes an expected risk functional should be chosen. Regularization theory chooses among models based on the prior assumption that the unknown function underlying the data is smooth, in the sense that similar inputs map onto similar outputs. The minimum description length method prescribes the constraint that the description length of data should be minimized. Formalizations of these ideas are described in the following sections.

3.1 Structural Risk Minimization

The basic idea of structural risk minimization is to estimate an upper bound of the expected risk, $E[L(y, z(\theta_y)]$, in terms of the observable empirical risk, $L_{\text{emp}}(y, z(\theta_y))$, given data sample y and a confidence interval, the value of which depends on a measure of model complexity and sample size, n. For binary classification problems for which $y_i = 0$ or 1, the following inequality holds

$$E[L(y, z(\theta_y))] \leqslant L_{\text{emp}}(y, z(\theta_y)) + I_\alpha(h, n) \qquad (3)$$

where I is the confidence interval, α is the alpha level of inferential statistics, and h is a quantity known as the Vapnik–Chervonenkis (VC) dimension (Vapnik and Chervonenkis 1971). The model that provides the lowest value of the upper bound for the expected risk should be chosen.

The VC dimension for a given model family of classification functions is defined as the maximum number of distinct binary samples that the model can fit perfectly without error for all possible labelings of the samples. Thus the VC dimension may be viewed as a measure of the complexity or flexibility of a model. Interestingly, the VC dimension depends not only on the number of parameters of a model but also, importantly, on the functional form of the model equation. A model with only one parameter can have infinite VC dimension while a model with a billion parameters can have a low VC dimension. For linear models, the VC dimension is equal to the number of parameters.

Given and n, the empirical risk $L_{\text{emp}}(y, z(\theta_y))$ is a monotonically decreasing function of the VC dimension whereas the confidence interval $I(h, n)$ is a monotonically increasing function of the VC dimension. These relationships along with the above equation reveal how structural risk minimization works. Among a set of competing models, the method chooses the model that describes the data sufficiently well (i.e., smallest empirical risk) with the least complex structure (i.e., narrowest confidence interval), thereby formalizing the principle of Occam's razor. In practice, structural risk minimization is implemented by first creating a set of k nested models with successively increasing VC dimensions, $h_1 < h_2 < \ldots, < h_k$, and then calculating the empirical risk for each model given sample data. The model with the smallest expected risk is then selected.

3.2 Regularization Theory

Regularization theory's solution for solving the ill-posed problem is to secure the uniqueness of the solution by requiring it to satisfy some secondary condition that embeds prior information or assumptions about the solution. The form of prior information most often used is the condition that the unknown function being sought is sufficiently smooth. Formally, the regularization risk for a model is defined as

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - z_i(\theta))^2 + \lambda S(z(\theta)) \qquad (4)$$

where the vector $z(\theta)$ is the model's prediction at θ, is known as the regularization parameter, and $S(z)$ is a functional that measures smoothness of the solution. The quantity often used for $S(z)$ is the integral of the squared second derivatives with respect to an independent variable x on which smoothness of the solution is assessed, i.e.,

$$S(z) = \int \left\| \frac{\partial^2 z(\theta, x)}{\partial x^2} \right\|^2 \mathrm{d}x \qquad (5)$$

The regularization theory prescribes that the model with the smallest $R(\theta)$ should be chosen.

The first term of the above equation is a measure of the discrepancy between the data and the model's prediction, that is, the empirical risk of the model. The second term, $S(z)$, is a measure of complexity of the model. The larger the value of $S(z)$, the less smooth the prediction, and thus the more complex the model. The

regularization parameter λ represents the relative importance of the complexity of the model with respect to its empirical risk. With λ_0, the solution is primarily determined by observed data and thus tends to be highly complex in structure. On the other hand, with $\lambda \to \infty$, the smoothness condition entirely dictates the form of the solution, which tends to be simple in structure. In practice, a value between these two extremes is chosen based on some principled approach (Haykin 1999). In short, the regularization method, like the structural minimization method, uses complexity to steer the solution toward improved generalization. As such, it is another implementation of the principle of Occam's razor.

3.3 Minimum Description Length

Minimum description length (MDL), which originated from algorithmic coding theory in computer science, regards both model and data as codes. The idea is that any data set can be appropriately encoded with the help of a model, and that the more we compress the data by extracting redundancy from it, the more we uncover underlying regularities in the data. Thus, code length is directly related to the generalization capability of the model, where the model that provides the shortest description of the data should be chosen. Formally, the total description length of a model is given by

$$MDL = L(\text{data}|\text{model}) + L(\text{model}) \tag{6}$$

where $L(\text{data}|\text{model})$ is the description length of the data given the model, measured in bits of information, and $L(\text{model})$ is the description length of the model itself.

The above definition of minimum description length is broad enough to be applied to any well-defined models, even verbally defined, qualitative models. For a statistical model defined as a parametrized family of probability distributions, the minimum description length for such a model takes the following form (Rissanen 1996):

$$MDL = -\ln f(y|\theta_y) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln\int d\theta\sqrt{\det(I(\theta))} \tag{7}$$

In the equation θ_y is the parameter value that maximizes the likelihood function $f(y|\theta)$, k is the number of parameters, $I(\theta)$ is the Fisher information matrix and $\det(I)$ is the determinant of the matrix. The first term of the MDL can be interpreted as a measure of risk or lack of fit, and the second and third terms together are a measure of model complexity. Model evaluation is therefore carried out by trading off lack of fit for simplicity.

In summary, although the three approaches to model evaluation evolved from different conceptual frameworks and philosophies, they all share the common goal of achieving optimum generalizability by striking a balance between risk (or lack of fit) and complexity. Another commonality between these approaches is the aspect of a model that is captured by the complexity measure; they not only include the number of parameters in the model, but also, importantly, its functional form, which refers to the way in which the parameters are combined in the model equation.

4. Relations to Statistical Approaches

So far we have considered computational approaches that view model evaluation as an inductive inference problem, which is the predominant view in computer science and engineering. Model evaluation is also a topic of central interest for statisticians (statisticians prefer the term *model selection* to *model evaluation*). Approaches which have been developed within the statistical framework include the generalized likelihood ratio test (GLRT), cross-validation, the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and Bayesian model selection (see *Model Testing and Selection*, *Theory of*). There are many important differences between these methods, but in essence, they all implement a means of finding, explicitly or implicitly, the best compromise between lack of fit and complexity by trading off one for the other. Statistical model selection methods are entirely consistent with the computational modeling approaches discussed above (see Sect. 3).

Here are a few notable differences and similarities between some of the statistical methods and computational methods. First, GLRT, AIC, and BIC differ from the computational methods such as structural risk minimization and minimum description length in that the two statistical selection criteria consider only the number of parameters as a complexity measure, and thus are insensitive to functional form, which can significantly influence generalizability. Second, application of the statistical methods requires that each model under investigation be a quantitative model defined as a parametric family of probability distributions. On the other hand, the computational methods can be applied to qualitative models as well as quantitative ones. For instance, the minimum description length method defined in Eqn. (4) is applicable to evaluating the effectiveness of decision tree models or even verbal models (Li and Vitanyi 1997). Finally, there exists a close connection between Bayesian model selection and the minimum description length criterion defined in Eqn. (5). The minimum description length criterion can be derived as an asymptotic approximation to the posterior probability in Bayesian model selection for a special form of the parameter prior density. An implication of this connection is that choosing the model that gives the shortest explanation for the observed data is essen-

tially equivalent to choosing the model that is most likely to be true in the sense of probability theory (Li and Vitanyi 1997), see *Algorithmic Complexity*.

5. Conclusion

The question of how one should decide among competing explanations (i.e., models) of data is at the core of the scientific endeavor. A number of criteria have been proposed for model evaluation (Jacobs and Grainger 1994). They include (a) *falsifiability*: are there potential outcomes inconsistent with the model? (b) *explanatory adequacy*: is the explanation compatible with established findings? (c) *interpretability*: does the model make sense? Is it understandable? (d) *descriptive adequacy*: does the model provide a good description of the observed data? (e) *simplicity*: does the model describe the phenomenon in the simplest possible manner? and (f) *generalizability*: does the model predict well the characteristics of new, as yet unseen, data?

Among these criteria, computational modeling approaches described in this article, as well as statistical approaches, consider the last three as they are easier to quantify than the other three. In particular, generalizability has been emphasized as the principal criterion by which the effectiveness of a model be judged. Obviously, this 'bias' toward the predictive inference aspect of modeling reflects the dominant thinking of the current scientific community. Clearly one could develop an alternative approach that stresses other criteria, and even incorporates some of the hard-to-quantify criteria. There are plenty of opportunities in this area for the emerging field of model evaluation to grow and mature in the decades to come.

See also: Algorithmic Complexity; Information Processing Architectures: Fundamental Issues; Information Theory; Mathematical Psychology; Model Testing and Selection, Theory of

Bibliography

Cherkassky V, Mulier F 1998 *Learning from Data*. Wiley, New York, Chaps. 2–4

Geman S, Binenstock E, Doursat R 1992 Neural networks and the bias-variance dilemma. *Neural Computation* **4**: 1–58

Jacobs A M, Grainger J 1994 Models of visual word recognition-sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance* **20**: 1311–34

Haykin S 1999 *Neural Networks: A Comprehensive Foundation*. Prentice Hall, London, Chaps. 2, 4 and 5

Li M, Vitanyi P 1997 *An Introduction to Kolmogorov Complexity and its Applications*. Springer, New York

Rissanen J 1989 *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore

Rissanen J 1996 Fisher information and stochastic complexity. *IEEE Transaction on Information Theory* **42**: 40–7

Tikhonov A N 1963 On solving incorrectly posed problems and method of regularization. *Doklady Akademii Nauk USSR* **151**: 501–4

Vapnik V N 1992 Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems* **4**: 831–8

Vapnik V N, Chervonenkis A Y 1971 On the uniform convergence of relative frequencies of events to their probabilities. *Theoretical Probability and Its Applications* **17**: 264–80

I. J. Myung

Computational Imagery

Reasoning with diagrams (that is, image, pictorial, depictive, or analogical representations) has long been a topic of interest among psychologists and philosophers. Diagrammatic representations are important for both the storage and communication of information, whether they are stored externally, such as a pictorial representation on paper, or internally in the form of a mental image. Recently, researchers in Artificial Intelligence (AI) have also been considering how diagrams can be used in computational approaches to reasoning and problem solving.

This article considers diagrammatic or imagery-based reasoning primarily from the computational perspective. In particular, we consider the issues of how images or diagrams can be represented and used for automated reasoning. The applications of computational imagery in the context of problem solving is also considered. As this is a relatively new area of research in AI, there is not yet an agreed upon underlying theory or standard approach for problem solving with images. Therefore, this entry presents an overview of approaches to this form of reasoning along with a description of various applications. A more detailed discussion of some of the issues presented here can be found in the book *Diagrammatic Reasoning* (Glasgow et al. 1995).

1. Reasoning with Images

A primary characteristic of a good formalism for knowledge representation is that it makes relevant properties explicit. Traditional approaches to knowledge representation and reasoning were developed primarily for linguistic or sentential reasoning, whereas computational imagery is concerned with emulating the problem solving abilities associated with mental imagery. Larkin and Simon (1987) have suggested that diagrammatic representations allow for the explicit representation and direct retrieval of depictive information that is only implicit in the more traditional descriptive representations.

In the context of artificial intelligence, imagery can be considered as the ability to manipulate representations for the purpose of retrieving visual and spatial information that was implicitly, rather than explicitly, encoded in long-term memory. Marr (1982) has proposed three levels of analysis that are necessary to understand vision: computational theory, algorithmic and hardware (see *Marr, David (1945–80)*). Although all three levels of analysis can also contribute to the understanding of diagrammatic reasoning, this entry focuses on the algorithmic level, which is concerned with the underlying representations and computational processes involved in problem solving with images or diagrams. A thesis of much work in computational imagery is that some of the processes involved are facilitated by depictive knowledge rep resentation schemes that incorporate inferencing techniques based on direct inspection and transformation of the representation.

Computational imagery generally involves tools and techniques for visual and spatial reasoning, where depictive image representations are generated or recalled from longterm memory and then transformed, scanned, associated with similar forms (constructing spatial analogies), pattern matched, composed, and so on. Whereas visual reasoning is concerned with what an image looks like, spatial reasoning depends more on where an object is located relative to other objects in a diagram or a scene. This computational distinction is analogous to the distinction between visual and spatial models proposed in cognitive studies of imagery. Note that the representations generated to retrieve image information may correspond to real physical scenes or abstract concepts that are manipulated in ways similar to visual forms.

Funt (1980) proposed a computational approach to visual imagery that represents the state of the world as a diagram and actions in the world as corresponding to actions in the diagram. Similar to Kosslyn's computational theory (1980), Funt uses two-dimensional arrays to depict visual images. In a more recent study of imagery, Chandrasekaran and Narayanan (1990) describe how visual information can be represented within the computational framework of discrete symbolic representations in such a way that both mental images and symbolic thought processes can be explained. Visual knowledge in this model is structured as multilevel hierarchies consisting of imaginal (two-dimensional pixel arrays) and symbolic (shape primitives) descriptions of the image configuration.

Lindsay (1988) suggests that depictive representations possess properties that help avoid 'the combinatorial explosion of correct but trivial inferences that must be explicitly represented in a prepositional system.' Sloman's (1985) definition of analogical representations states that a representation is depictive if there is a correspondence between the structure of the representation and the structure of the image being denoted. More precisely, if D is a depictive representation of an image I, then (a) there must be a mapping of the parts in D (e.g., symbols or pixels) to parts in I, and (b) it must be possible to define a correspondence between properties or relations of parts of D and properties or relations of parts of I (e.g., shape, relative location, volume). In a *vivid* form of knowledge, there is also a one-to-one correspondence between symbols and objects in the world, and for every simple relationship of interest in the world there exists corresponding connections among symbols in the knowledge base. Levesque (1986) asserts that the main advantage of vivid knowledge bases is that they provide for efficient worst case reasoning behavior, since calculating what is logically implicit generally reduces to retrieving what is explicit.

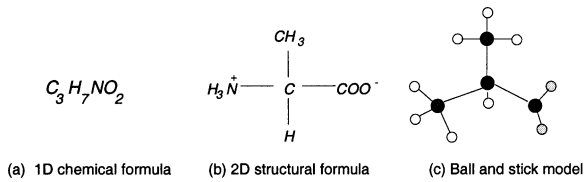

Figure 1
Alternative representations for molecule

As a criteria for assessing alternative representations for shape description in vision, Marr (1982) introduced the concept of accessibility, which is defined as the ability to derive a representation from an image at reasonable computing costs. A similar criteria can also be applied to potential representations for imagery. Although formulating an image representation is often straightforward, constructing one that is efficient in terms of the amount of information stored, and also effective for problem solving, may be a complex process. For example, in the molecular domain it is possible to represent a structure in several different ways; our choice of representation depends on the questions we wish to ask. Take, for example, the multiple representations of a molecular structure illustrated in Fig. 1. If we wish to determine how many atoms of carbon are contained in a molecule, then the formula in Fig.1(a) is sufficient. However, if we need to derive connectivity, angle, distance or shape information, then a more complex image representation, such as one corresponding to the illustration in Fig. 1(b) or (c), may be more appropriate.

Hayes (1974) discusses the notion of direct representations in which there exist similarities between what is being represented and the medium of the representation. Sloman (1985) has also argued the pros and cons of analogical representations, and has concluded that a variety of representation formalisms—including those specialized for spatial reasoning—are important to AI problem solving. Other hybrid approaches have been suggested for visual-spatial and model-based reasoning. Habel et al. (1993) have developed a system consisting of a propositional

and depictorial partonomy (organization of parts) for reasoning, where the depictorial partonomy reflects the hierarchy proposed in representations for visual processes. They suggest that the advantage of the depictorial representation in the system is that it facilitates an efficient attention-driven method for reasoning. Myers and Konolige (1992) treat model-based manipulations as a form of inference within a classical logic system. More specifically, they store partially interpreted sensor data using an analogical representation that interacts with a general-purpose sentential language. A technique for qualitative spatial reasoning, based on the directional orientation information made available through perceptual processes, has been presented by Freksa and Zimmermann (1993). In this work, orientations in two-dimensional space are defined by the relation between a vector and a point.

Motivated by cognitive studies of mental imagery, Glasgow and Papadias (1992) have proposed a hybrid knowledge representation scheme for computational imagery consisting of three representations, each appropriate for a different kind of image processing: an image is stored in long-term memory as a hierarchically organized, descriptive representation; the spatial representation denotes image components symbolically and preserves relevant spatial and topological properties; and the visual representation depicts information such as shape, relative distance, and size. Imagery-based reasoning is carried out using primitive and high-level processes that construct, transform, and inspect visual or spatial representations.

2. *Applications to AI Problem Solving*

Visual-spatial reasoning techniques have also been considered in the context of specific application domains. Funt (1980) represents and manipulates a visual analog of a world in order to predict potential instabilities and collisions in a physical domain. Others have also used diagrammatic reasoning to simulate potential solutions for problem solving.

Epstein and Gelfand (1998) use spatial representations to generate potential moves in game playing. Barwise and Etchemendy (1993) have proposed a system called Hyperproof which integrates diagrammatic reasoning with sentence-based logics. Hyperproof uses both diagrams and logic notation to teach students how to reason logically. They have also presented a formal semantics for reasoning with Hyperproof diagrams.

Research in geographical information systems and spatial databases has long been concerned with the issue of representing spatial knowledge (see *Geographic Information Systems: Critical Approaches*; *Spatial Data*). Samet (1989) has proposed a method for storing geographic knowledge based on the recursive decomposition of space. In this work, the term *quadtree* is used to describe binary array data structures that iteratively subdivide regions into segments until blocks are obtained that consist entirely of 1s or entirely of 0s. These structures (and their three-dimensional counterpart, termed octrees) are efficiently stored and implemented as trees, where each node of the tree corresponds to a region in the decomposition hierarchy. Another approach to reasoning in geographic systems has been described by Papadias and Sellis (1993). In this work, a symbolic two-dimensional array structure is used to preserve a set of spatial relations among geographic entities.

Spatial representations have also been considered in a machine vision research. According to Biederman (1987), the representation of objects can be constructed as a spatial organization of simple primitive volumes, called geons. The process of image analysis, as defined by Marr et al. (1982), depends on a series of representations culminating in a three-dimensional model of the spatial relations among entities which makes explicit what is where. As in Marr's approach to computational vision, research in molecular scene analysis (1993) is concerned with discovering what is present in the world and where it is spatially located. The act of determining the structure of a molecule is an interactive process consisting of a state space search of partially interpreted scenes, which can be represented and evaluated using processes from computational imagery.

Other problem domains where diagrammatic reasoning has been applied include qualitative physics, route planning, biology, human/computer interaction, architecture, geometry, and theorem proving.

3. *Future Research*

Research in the area of diagrammatic and imagery based reasoning is still in its infancy. Most of the work to date has focused on individual applications where visual or spatial reasoning can provide a guide to problem solving. There are still many unanswered questions and areas to pursue that would lead us to a better understanding of how reasoning with images could improve problem solving in artificial intelligence. Some of the questions we might consider are:

(a) Are there general representation and reasoning techniques that cross application boundaries? Forbus (1993) has stated the Poverty Conjecture that claims: 'there is no problem-independent, purely qualitative representation of space or shape.' Whether this claim is true or not is clearly a topic for future research.

(b) How does diagrammatic reasoning interact with other reasoning paradigms in AI? Already, researchers have been investigating how image representations and reasoning techniques may interact with approaches such as rule-based reasoning or case-based reasoning. More work needs to be carried out to

understand how such hybrid systems can best be constructed and applied.

(c) What can we learn from human cognitive models? The topic of mental imagery has long been studied in areas such as psychology, philosophy, and linguistics. Researchers in AI have much to learn from the understanding of how the human information processing system reasons with image representations.

(d) What are other potential domains of application for diagrammatic reasoning? Although several application areas have been considered, there are certainly other areas that would benefit from the ability to incorporate diagrams or images for the purpose of problem solving.

The area of diagrammatic reasoning is relatively new in AI. Prior to 1990 there were only a handful of papers related to this topic. Although interest and activity has greatly increased during the 1990s, there are still many questions to be posed and answered in this emerging area.

See also: Concept Learning and Representation: Models; Feature Representations in Cognitive Psychology; Imagery versus Propositional Reasoning; Knowledge Representation; Logics for Knowledge Representation; Mental Imagery, Psychology of; Mental Models, Psychology of; Mental Representations, Psychology of; Practical Reasoning: Philosophical Aspects; Propositional Representations in Psychology; Reasoning with Mental Models; Visual Imagery, Neural Basis of

Bibliography

Biederman I 1987 Recognition-by-components: A theory of human image understanding. *Psychological Review* **94**: 115–47

Chandrasekaran B, Narayanan N 1990 Integrating imagery and visual representations. In: *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 670–7

Epstein S, Gelfand J, Lock E 1998 Learning game-specific spatially oriented heuristics. *Constraints: An International Journal* **2**: 239–51

Forbus K D 1993 Image and substance. *Computational Intelligence* **9**(4): 309–33

Fortier S, Castleden I, Glasgow J, Conklin D, Walmsley C, Leherte L, Allen F 1993 Molecular scene analysis: The integration of direct methods and artificial-intelligence strategies for solving protein crystal structures. *Acta Crystallographica* **D49**: 168–78, Part 1

Freksa C, Zimmermann K 1993 On the utilization of spatial structures for cognitively plausible and efficient reasoning. In: *Proceedings of the IJCAI '93 Workshop on Spatial and Temporal Reasoning*, pp. 61–66

Funt B V Problem-solving with diagrammatic representations. *Artificial Intelligence* **13**: 201–30

Glasgow J, Papadias D 1992 Computational imagery. *Cognitive Science* **16**(3): 355–94

Glasgow W J, Narayanan N H, Chandrasekaran B (eds.) 1995 *Diagrammatic Reasoning: Cognitive and Computational Perspectives*. AAAI Press, Menlo Park, CA

Habel C, Pribbenow S, Simmons G 1993 Propositional and depictorial partonomies: A hybrid approach. In: *Proceedings of IJCAI '93 Workshop on Principles of Hybrid Representation and Reasoning*, pp. 62–71

Hayes P 1974 Some problems and non-problems in representation theory. In: *Proceedings of AISB Summer Conference*, pp. 63–79

Kosslyn S M 1980 *Image and Mind*. Harvard University Press, Cambridge, MA

Larkin J H, Simon H A 1987 Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science* **11**: 65–99

Levesque H 1986 Making believers out of computers. *Artificial Intelligence* **30**: 81–108

Lindsay R K 1988 Images and inference. *Cognition* **29**: 229–50

Marr D 1982 *Vision*. W. H. Freeman, San Francisco

Myers K, Konolige K 1992 Reasoning with analogical representations. In: *Proceedings of the Conference on Principles of Knowledge Representations and Reasoning*. Morgan Kaufmann, Los Altos, CA

Papadias D, Sellis T 1993 The semantics of relations in 2d space using representative points: Spatial indices. In: Frank A U, Campari I (eds.) *Spatial Information Theory: A Theoretical Basis for GIS: European Conference (COSIT '93)*, Marciana Marina, Elba Island, Italy, September 19–22 proceedings. Springer-Verlag, Berlin, Vol 716, pp. 234–47

Samet H 1989 *The Design and Analysis of Spatial Data Structures*. Addison–Wesley, Reading, MA

Sloman A 1985 Why we need many knowledge representation formalisms. In: *Research and Development in Expert Systems: Proceedings of the Expert Systems… the… Annual Technical Conference of the British Computer Society Specialist Group on Expert Systems*. Cambridge University Press, for the British Computer Society, Cambridge, UK

J. I. Glasgow

Computational Neuroscience

The term 'computation' in computational neuroscience refers to the way that brains processes information. Many different types of physical systems can solve computational problems, including slide rules and optical analog analyzers as well as digital computers, which are analog at the level of transistors and must settle into a stable state on each clock cycle. What these have in common is an underlying correspondence between an abstract computational description of a problem, an algorithm that can solve it, and the states of the physical system that implement it (Fig. 1). This is a broader approach to computation than one that is based purely on symbol processing (Churchland and Sejnowski 1992, Arbib 1995).

There is an important distinction between general-purpose computers, which can be programmed to solve many different types of algorithms, and special-purpose computers, which are designed to solve only a

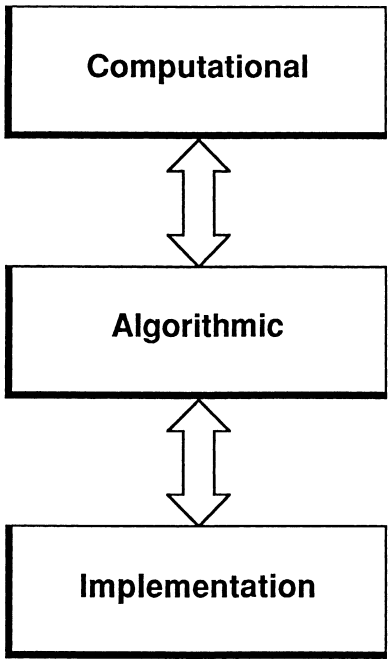

Figure 1
Levels of analysis. Marr (1982) advocated a top-down approach, starting with the computational level and proceeding down to the implementation level. The two-way arrows indicate that constraints between levels can be used to gain insights in both directions. Research in computational neuroscience often starts at the implementation level and proceeds up to the computational level

limited range of problems. Most neural systems are specialized for particular tasks; an example is the retina, which is dedicated to visual transduction and image processing. As a consequence of the close coupling between structure and function, the anatomy and physiology of a brain area can provide important clues to the algorithms that it implements and its computational function (Fig. 1); in contrast, the hardware of a general-purpose computer may not reveal its function, which also depends on software.

Another major difference between brains and general-purpose digital computers is that computers are hardwired, but the connectivity between neurons and their properties is shaped by the environment during development and remains plastic even in adulthood. Thus, as the brain processes information, it changes its own structure in response to the information that is being processed. Adaptation and learning are important mechanisms that allow brains to respond flexibly as the world changes—on a wide range of time scales, from milliseconds to years. The flexibility of the brain has survival advantage when the environment is nonstationary, and the evolution of some cognitive skills may depend deeply on genetic processes that have extended the time scales for brain plasticity.

Brains are complex, nonlinear dynamical systems with feedback loops, and brain models provide intuition about the possible behaviors of such systems. The predictions of a model make explicit the consequences of the underlying assumptions, and comparison with experimental results can lead to new insights and discoveries. Emergent properties of neural systems, such as oscillatory behaviors, depend on both the intrinsic properties of the neurons and the pattern of connectivity between them. For example, the large-scale coherent brain rhythms that accompany different states of alertness and sleep arise from intrinsic properties of thalamic and cortical neurons that are reciprocally connected (Destexhe and Sejnowski 2001).

1. *Models at Different Levels of Detail*

1.1 *Realistic Models*

Perhaps the most successful model at the level of the neuron has been the classic Hodgkin–Huxley model of the action potential in the giant axon of the squid (Koch and Segev 1998). Data were first collected under a variety of conditions, and a model was later constructed to integrate the data into a unified framework. This type of model requires that most of the variables in the model have been measured experimentally, and only a few unknown parameters need to be fitted to the data. Detailed models can be used to distinguish between different explanations of the data. In the classic model of a neuron, information flows from the dendrites, where synaptic signals are integrated, to the soma of the neuron, where action potentials are initiated and carried to other neurons through long axons. In these models, the dendrites are passive cables, but recently, voltage-dependent sodium, calcium, and potassium channels have been observed in the dendrites of cortical neurons, which greatly increase the complexity of synaptic integration (Stuart et al. 1999). Experiments and models have shown that these active currents can carry information in a retrograde direction from the cell body back to the distal synapses tree. Thus, it is possible for spikes in the soma to affect synaptic plasticity through the mechanisms discussed in Sect. 2.

Realistic models with several thousand cortical neurons can be explored on the current generation of workstation. The first model for the orientation specificity of neurons in the visual cortex was the 'feedforward' model proposed by Hubel and Wiesel, which assumed that the orientation preference of cortical cells was determined primarily by converging inputs from thalamic relay neurons. Although there is solid experimental evidence for this model, local

cortical circuits have been shown to be important in amplifying weak signals and suppressing noise, as well as performing gain control to extend the dynamic range. These models are governed by the type of attractor dynamics that was analyzed by John Hopfield (1982), who provided a conceptual framework for the dynamics of feedback networks.

Although the spike train of cortical neurons is highly irregular, and is typically treated statistically, information may be contained in the timing of the spikes in addition to the average firing rate. This has already been established for a variety of sensory systems in invertebrates and peripheral sensory systems in mammals (Rieke et al. 1996), but whether spike timing carries information in cortical neurons is an open research issue. In addition to representing information, spike timing could also be used to control synaptic plasticity through Hebbian mechanisms for synaptic plasticity, as discussed in Sect. 2.

1.2 Signal Processing Models

Other types of models have been used to analyze experimental data in order to determine whether they are consistent with a particular computational assumption. For example, a 'vector averaging' technique has been used to compute the direction of arm motion from the responses of cortical neurons (Georgopoulos et al. 1986), and signal detection theory was used to analyze the information from cortical neurons responding to visual motion stimuli (Movshon and Newsome 1996). Bayesian methods can effectively decode the position of a rat in a maze from place cell recorded in the hippocampus (Zhang et al. 1998). In these examples, the computational model was used to explore the information in the data but was not meant to be a model for the actual cortical mechanisms. Nonetheless, these models have been highly influential and have provided new ideas for how populations of neurons may represent sensory information and motor commands.

Neural network or 'connectionist' models that simplify the intrinsic properties of neurons can be helpful in understanding the computational consequences of information contained in large populations of neurons. An example of this approach is a recent model of parietal cortex based on the response properties of cortical neurons (Pouget and Sejnowski 2001). The parietal cortex is involved in representing the spatial location of objects in the environment and computing transformations from sensory to motor coordinates. The model examined the issue of which reference frames are used in the cortex for performing these transformations. The model predicted the outcomes of experiments performed on patients with spatial neglect following lesions of the parietal cortex.

Cognitive functions such as attention have also been modeled. Francis Crick (1994) proposed that the relay cells in the thalamus may be involved in attention, and has provided an explanation for how this could be accomplished based on the anatomy of the thalamus. Models of competition between neurons in cortical circuits can explain many properties of single cortical neurons in awake, behaving monkeys during attention tasks (Reynolds et al. 1999).

Finally, small neural systems have been analyzed with dynamical systems theory (Harris-Warrick et al. 1995). This approach is feasible when the numbers of parameters and variables are small. Most models of neural systems involve a large number of variables, such as membrane potentials, firing rates, and concentrations of ions, with an even greater number of unknown parameters such as synaptic strengths, rate constants, and ionic conductances. Where the number of neurons and parameters is very large, techniques from statistical physics become applicable in predicting the average behavior of large systems (Van Vreeswijk and Sompolinsky 1998). There is a midrange of system sizes where neither type of limiting analysis is possible, but where simulations can be performed (Bower and Beeman 1998). One danger of relying solely on computer simulations is that they may be as complex and difficult to interpret as the biological system itself.

2. Learning and Memory

One of the goals of computational neuroscience is to understand how long-term memories are formed through experience and learning. There is increasing experimental evidence that the strengths of interactions between neurons can be altered by activity, called synaptic plasticity. For example, high-frequency trains of stimuli at synapses in the hippocampus induce a form of long-term potentiation (LTP) that can last for days (Bliss and Lomo 1973). Moreover, these synapses require simultaneous presynaptic activity and postsynaptic depolarization (Kelso et al. 1986), as suggested by Hebb (1949). Models of associative memory that incorporate Hebbian synaptic plasticity have been widely explored (Anderson and Hinton 1981).

Hebbian synaptic plasticity can also be used to form maps and has been used to model the early development of the projection from the thalamus to the visual cortex. In particular, these models can explain why inputs from the right and left eyes form alternating stripes in primary visual cortex of cats and monkeys, called ocular dominance columns (Miller et al. 1989). Specific mappings arise in the cortex because temporal contiguity in axonal firing is translated into spatial contiguity of synaptic contacts.

The change in the strength of synapses in the hippocampus and neocortex depends on the relative timing of spikes in the presynaptic neuron and the postsynaptic neuron. Reliable LTP occurs when the presynaptic stimulus precedes the postsynaptic spike, but there is long-term depression (LTD) when the

presynaptic stimulus immediately follows the postsynaptic spike (Markram et al. 1997, Bi and Poo 1998). This temporal asymmetry in synaptic plasticity solves the problem of balancing LTD and LTP, since chance coincidences should occur about equally with positive and negative relative time delays. When sequences of inputs are repeated in a network of neurons with recurrent excitatory connections, temporally asymmetric synaptic plasticity will learn the sequence and the pattern of activity in the network will tend to predict future inputs. There is evidence for this in the hippocampus where place cells representing nearby locations in a maze may be linked together (Blum and Abbott 1996), and in visual cortex where simulations of cortical neurons can become directionally selective when exposed to moving visual stimuli (Rao and Sejnowski 2000).

The temporally asymmetric Hebbian learning rule can be used to implement the temporal difference learning algorithm in reinforcement learning and classical conditioning (Montague and Sejnowski 1994, Sutton and Barto 1998, Rao and Sejnowski 2000). The unconditioned stimulus in a classical conditioning experiment must occur before the reward for the stimulus–reward association to occur. This is reflected in the temporal difference learning algorithm by a postsynaptic term that depends on the time derivative of the postsynaptic activity level. The goal is for the synaptic input to predict future reward: if the reward is greater than predicted the postsynaptic neuron is depolarized and the synapse strengthens, but if the reward is less than predicted, the postsynaptic neuron is hyperpolarized and the synapse decreases in strength. There is evidence that in primates the transient output from dopamine neurons in the ventral tegmental area carries information about the reward predicted from a sensory stimulus (Schultz et al. 1997), and in bees, an octopaminergic neuron has a similar role (Hammer and Menzel 1995).

The temporal window for classical conditioning is several seconds—much longer than the window for LTP/LTD observed at cortical and hippocampal synapses. A circuit of neurons in the basal ganglia and frontal cortex may be needed to extend the computation of temporal differences to these long time intervals (Berns and Sejnowski 1998). It is surprising to find the same learning algorithm in different types of learning systems in different parts of the brain. This suggests that the temporal order of input stimuli is a useful source of information about causal dependence in many different learning contexts and over a range of time scales.

3. *Future Directions*

3.1 *Molecular/Synaptic*

Almost all of the proteins that make up synapses are now known, and the sequencing of the human genome will make it possible to complete that list within the next few years. Coupled with high-voltage electron microscopy and methods for labeling these proteins, it should be possible to develop a reasonably complete model for the neuromuscular junction and central synapses in a few more years. This will allow us to understand the synapses as a molecular machine and to understand the mechanisms that are involved in synaptic plasticity (building on a large existing body of research, for which the Nobel Prize in Medicine was given to Arvid Carlsson, Paul Greengard and Eric Kandel in 2000). It is highly likely that advances in modeling and computational theory at this level will have direct impact on the pharmaceutical industry through the design of new drugs and new approaches to mental disorders.

3.2 *Neuron/Network*

Progress in this area has been slower because experimental technique for recording from many neurons simultaneously still lags behind. However, the advent of large electrode arrays and optical recording techniques will make rapid progress possible so before long we should have a reasonable understanding of neural circuits in simple creatures and some parts of the vertebrate brain. Computer models are being used to understand how the intrinsic properties of neurons and the synapses that join them produce complex spatio-temporal patterns of activity and how these networks are used to encode, store and retrieve information in large neural systems such as cortical columns. Advances at this level could lead to a new generation of pattern recognition systems and sensorimotor control systems for robotic devices.

3.3 *Map/System*

The major breakthrough that has occurred recently in human brain imaging, primarily with the development of functional magnetic resonance imaging, is allowing rapid progress to be made in localizing cognitive functions within the human brain. Magnetic resonance technology is still at an early stage of development and new approaches, such as magnetic resonance spectroscopy, should make it possible to observe biochemical reactions occurring within the brain. However, until such time as resolution of these new techniques is improved, or they can be integrated with older techniques such as electroencephalography (EEG) and magnetoencephalography (MEG), it will not be possible to uncover the mechanisms that underlie the activity that is observed with functional imaging. It may take decades before techniques mature to the point where we can begin to understand the large-scale organization of primate brains. However, even before this, significant progress will be made using invasive techniques in nonhuman primates and other mammals. Understanding of brains at this level

will provide the ultimate insights into major philosophical questions about consciousness and autonomy (Crick 1994).

3.4 Technology for Brain Modeling

Do new properties emerge as the number of neurons in a neural system becomes large? This question can be explored with large simulations of millions of neurons. Parallel computers have become available that permit massively parallel simulations, but the difficulty of programming these computers has limited their usefulness. An approach to massively parallel models introduced by Carver Mead is based on subthreshold cMOS VLSI (Very Large Scale Integrated) circuits with components that directly mimic the analog computational operations in neurons. Several large silicon chips have been built that mimic the visual processing found in retinas. Analog VLSI cochleas have also been built that can analyze sound in real time. Analog VLSI chips have been built that mimic the detailed biophysical properties of neurons, including dendritic processing and synaptic conductances (Douglas et al. 1995). These chips use analog voltages and currents to represent the signals, and are extremely efficient in their use of power compared to digital VLSI chips. A new branch of engineering called *neuromorphic engineering* has arisen to exploit this technology.

4. Conclusions

Although brain models are now routinely used as tools for interpreting data and generating hypotheses, we are still a long way from coming up with explanatory theories of brain function. For example, despite the relatively stereotyped anatomical structure of the cerebellum, we still do not understand its computational functions. Recent evidence from functional imaging of the cerebellum suggests that the cerebellum is involved in higher cognitive functions and is not just a motor controller. Modeling studies may help in exploring these competing hypotheses. This has already occurred in the oculomotor system, which has a long tradition of using control theory models to guide experimental studies.

Digital computers have been increasing in power exponentially from around 1950 when the first machines, built from vacuum tubes and mercury delay lines, were introduced. The number of operations per second roughly doubles every three years. This vast increase in computer power, or, conversely, decrease in the cost of computing and data storage, has had a major impact on the study of the brain and on the development of brain models and theories. For example, functional magnetic resonance imaging, which has made it possible to study human brain activity non-invasively, would not have been feasible without fast digital computers that direct the pulse sequences, data acquisition and data analysis. The increasing power of computers is transforming our ability to analyze complex neural systems at many levels of investigation and will have far-reaching consequences on society.

It is very important to note that, at this stage in our understanding of the brain, a model should only be considered a provisional framework for organizing thinking. Many partial models need to be explored at many different levels of investigation, each model focusing on a different scientific question. As computers become faster, and as software tools become more flexible, computational models should proliferate. Close collaborations between modelers and experimentalists, facilitated by the internet, should lead to an increasingly better understanding of the brain as a computational system.

See also: Artificial Neural Networks: Neurocomputation; Behavioral Neuroscience; Cognitive Neuroscience; High Performance Computing; Information Processing Architectures: Fundamental Issues; Neural Networks and Related Statistical Latent Variable Models

Bibliography

Anderson J A, Hinton G E (eds.) 1981 *Parallel Models of Associative Memory*. Lawrence Erlbaum, Hillsdale, NJ

Arbib A M 1995 *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA

Berns G S, Sejnowski T S 1998 A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience* **10**(1): 108–21

Bi G-Q, Poo M-m 1998 Activity-induced synaptic modifications in hippocampal culture: dependence on spike timing, synaptic strength and cell type. *Journal of Neuroscience* **18**: 10464–72

Bliss T V P, Lomo T 1973 Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* **232**: 331–56

Blum K I, Abbott L F 1996 A model of spatial map formation in the hippocampus of the rat. *Neural Computation* **8**: 85–93

Bower J M, Beeman D 1998 *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System*, 2nd edn. TELOS, Santa Clara, CA

Churchland P S, Sejnowski T J 1992 *The Computational Brain*. MIT Press, Cambridge, MA

Crick F H C 1994 *The Astonishing Hypothesis: The Scientific Search for the Soul*. Charles Scribner's Sons, New York

Destexhe A, Sejnowski T J 2001 *Thalamocortical Assemblies: How Ion Channels, Single Neurons and Large-Scale Networks Organize Sleep Oscillations*. Oxford University Press, Oxford, UK

Douglas R, Mahowald M, Mead C 1995 Neuromorphic analogue VLSI. *Annual Review of Neuroscience* **18**: 255–81

Georgopoulos A P, Schwartz A B, Kettner R E 1986 Neuronal population coding of movement direction. *Science* **243**: 1416–19

Hammer M, Menzel R 1995 Learning and memory in the honeybee. *Journal of Neuroscience* **15**: 1617–19
Harris-Warrick R M, Coniglio L M, Barazangi N, Guckenheimer J, Gueron S 1995 Dopamine modulation of transient potassium current evokes phase shifts in a central pattern generator network. *Journal of Neuroscience* **15**(1): 342–58
Hebb D O 1949 *Organization of behavior: A neuropsychological theory*. John Wiley and Sons, New York
Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of the Sciences USA* **79**: 2554–8
Kelso S R, Ganong A H, Brown T H 1986 Hebbian synapses in hippocampus. *Proceedings of the National Academy of the Sciences USA* **83**: 5326–30
Koch C, Segev I 1998 *Methods in Neuronal Modeling: From Synapses to Networks*, 2nd edn. MIT Press, Cambridge, MA
Markram H, Lubke J, Frotscher M, Sakmann B 1997 Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**: 213–15
Marr D 1982 *Vision*. Freeman, New York
Miller K D, Keller J B, Stryker M P 1989 Ocular dominance column development: analysis and simulation. *Science* **245**: 605–15
Montague P R, Sejnowski T J 1994 The predictive brain: temporal coincidence and temporal order in synaptic learning mechanisms. *Learning and Memory* **1**: 1–33
Movshon J A, Newsome W T 1996 Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *Journal of Neuroscience* **16**: 7733–41
Pouget A, Sejnowski T J 2001 Lesioning a basis function model of spatial representations in the parietal cortex: Comparison with hemineglect. *Psychological Review* (in press)
Rao R P N, Sejnowski T J 2000 Predictive sequence learning in recurrent neocortical circuits. In: Solla S A, Lee T K, Muller K R (eds.) *Advances in Neural Information Processing Systems 12*. MIT Press, Cambridge, MA, pp. 164–70
Reynolds J H, Chelazzi L, Desimone R 1999 Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience* **19**: 1736–53
Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W 1996 *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA
Schultz W, Dayan P, Montague P R 1997 A neural substrate of prediction and reward. *Science* **275**: 1593–9
Stuart G, Spruston N, Hausser M (eds.) 1999 *Dendrites*. Oxford University Press, Oxford, UK
Sutton R S, Barto A G 1998 *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA
Van Vreeswijk C, Sompolinsky H 1998 Chaotic balanced state in a model of cortical circuits. *Neural Computation* **10**(6): 1321–71
Zhang K, Ginzburg I, McNaughton B L, Sejnowski T J 1998 Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *Journal of Neurophysiology* **79**: 1017–44

T. J. Sejnowski

Computational Psycholinguistics

Psycholinguistics is the scientific discipline that studies how people acquire a language and how they comprehend and produce this language. The increasing complexity of the models of human language processing which have been evolved in this discipline makes the development and evaluation of computer implemented versions of these models more and more important to understand these models and to derive predictions from them. Computational psycholinguistics is the branch of psycholinguistics that develops and uses computational models of language processing to evaluate existing models with respect to consistency and adequacy as well as to generate new hypotheses.

1. Cognitive Modeling and Computational Psycholinguistics

Not every computer model mimicking linguistic behavior can be declared as being psychologically relevant. For example, the ELIZA question-answering system (Weizenbaum 1966) is able to imitate the role of a therapist very well, although the system's behavior is based solely on simple pattern matching mechanisms that do not say anything about the linguistic and nonlinguistic processes underlying a dialogue. In order to be a psycholinguistically relevant computer model, the model must simulate a specific human cognitive function. As a prerequisite for the construction of such a model, formal analyses of the linguistic domain and empirical studies are required to determine the complexity of that domain and of the parameter values of the respective cognitive function to be modeled.

This research methodology corresponds to the typical methodology in cognitive modeling (see *Cognitive Modeling: Research Logic in Cognitive Science*). In fact, there is only one difference between cognitive modeling and computational psycholinguistics: While cognitive modeling deals with all aspects of human cognition, computational psycholinguistics is confined to architectures and mechanisms for human language processing.

1.1 Computer Models in Psycholinguistics

In essence, psycholinguistics addresses the questions of how speakers put pre-linguistic concepts into words, and then combine these words to larger units in order to produce the oral or written output, and how a listener parses the auditory or visual input into meaningful units and arrives at an understanding of the input.

Although psycholinguistics as part of the interdisciplinary field of the cognitive sciences shares the view that cognition should essentially be regarded as computation, the research methodology in psycholinguistics is primarily oriented at experimental studies with the computer as a tool for the selection and presentation of appropriate stimulus material, and the exact measurements of reaction times, etc. Computer modeling itself does not play a major role in model development. Surveys of psycholinguistic research

(e.g., Gernsbacher 1994) dedicate only a minor part to computer modeling in psycholinguistic research, and the use of computer models to evaluate models of (sub)tasks of language comprehension or production. This means that, compared with approximately 100 years of psycholinguistic research and 25 years of institutionalized cognitive science, computational psycholinguistics can be glossed as a brand new area of psycholinguistic research. Computer modeling turns out to be advantageous in model development for at least two reasons.

First, specifying a model only verbally makes it more difficult to check its completeness and overall consistency, if the model is appropriately complex. If such a model has been realized as a computer program and produces the desired input-output effect, it is not just a specification of relations in the model domain, but also a fully specified and consistent theory of the investigated cognitive function.

Second, the implemented model may be used to generate hypotheses: the compulsion of being explicit in every detail while developing the computer model may result in predictions under conditions which have empirically not been investigated before. These predictions can then be tested by new experiments that will either support or not support the predictions made by the program.

1.2 Techniques for Modeling Human Language Processing

In general, a reconstruction of an empirical fact by a model can be done in each of two ways: a narrow one or a process-oriented one. The former means that the computer model realizes an input–output relation that is identical to that of human beings. However, whether the processes underlying this mapping from informational inputs to the desired outputs correspond to the processes of humans or not is irrelevant. The latter means that the computer system does not only realize a correct input-output relation, but it also accounts for the internal representations and operations which will be used and performed accordingly, within the human mind. Artificial Intelligence typically develops models of the first class (see *Artificial Intelligence: Connectionist and Symbolic Approaches*), while computational psycholinguistics strives for models of the second class. Evaluation criteria for the resulting computer models are the same as for every model: their descriptive and explanatory adequacy, simplicity, generality, and falsifiability. An important additional criterion is the possibility to match empirical results to the system's behavior. Computer models of the same task can also be evaluated in a model-to-model comparison.

Techniques used in computational psycholinguistics range from all kinds of symbolic processing mechanisms like graph unification, planning, or deductive reasoning to connectionist approaches. The advantage of symbolic mechanisms is the achievement of a high level of abstraction that makes it easier to check the adequacy of the system's behavior. One disadvantage of symbolic systems is the possible rigidity so that exceptions to the rule require additional treatments. This robustness problem is problematic for several aspects of human language processing, because often decisions for the production or recognition of a specific linguistic form or structure turn out to be highly context-dependent (e.g., the modeling of typical speech errors).

Connectionist approaches, especially activation spreading, are also quite popular in computational psycholinguistics. The connectionist paradigm bypasses the robustness problem and it accounts for learnability. Its disadvantage is the lack of modeling of rule-based structural relationships that is assumed to be an essential characteristic of many tasks in human language processing.

2. Stages in Human Language Processing

Although the various proposed architectures of the human language processing systems and the respective mechanisms differ considerably in detail, there is not much disagreement on the basic stages involved in transforming concepts into a linguistic utterance and *vice versa*.

2.1 Stages in Language Production

Most psychological models of language production agree on the division of the production process into three major stages that are called, following (Levelt 1989), conceptualization, formulation, and articulation. Differentiation with respect to the subcomponents, interrelations, and functions of these components takes very different directions.

Conceptualization for language production comprises the selection, preparation, and linearization of pre-linguistic information. These processes lead to a conceptual representation which functions as input to a formulator.

Formulation involves the conversion of the output from the conceptual level into linguistic structures. It includes the selection of appropriate lexical items, and the retrieval of syntactic structures and phonetic information. It is widely accepted that the selection of lexical items consists of a two-step process. The first step involves the selection of lemmas, i.e., semantically appropriate items together with syntactic features. The second selection process concerns lexeme retrieval where phonological information becomes available. The relationship between both retrieval processes is still controversial, which finds its expression in the different architectural bases and the directionality of activation flow in the corresponding computer models of this task. The result of the formulation process

is an articulatory plan. Finally, articulation can be described as the realization of that phonetic plan and its execution at the motoric level.

One fundamental difference between the various production models relates to how the flow of information between the various processing stages is designed. Sequential models assume a unidirectional flow from the conceptualizer to the formulator, and from the formulator to the articulator. The enormous rate of human speech production is explained in these models by incremental processing, i.e., the processing of information in a piecemeal way: All modules work in parallel, but not on the same piece of information. Increments will be handed down to subsequent modules although the whole informational unit has not been completed yet.

In contrast to sequential models, interactive models allow direct feedback from subordinate to superordinate modules. Sequential models have the advantage of being easier to check empirically and simulate via computer programs. However, they have a completeness problem concerning the relationship between language and conceptualization. Sequential models must assume intermediate representations that reflect the previous decisions and the momentary knowledge. For example, all relevant conceptual decisions must have been concluded at the point where a conceptual fragment is handed down to the formulator. Nothing can be left unspecified which is relevant for selection processes in the formulator since the formulator itself does not provide criteria for decisions relating to content.

Both types of information flow are often directly related to the symbolic and connectionist processing paradigm, although there is no cogent link between a specific type of information flow and propositional or connectionist models.

2.2 *Stages in Language Comprehension*

From a logical point of view, the basic stages in comprehending a linguistic utterance must be the same as in language production, although there is overwhelming evidence that production is not simply the reverse of comprehension. The visual (in the case of reading) or auditory (in the case of listening) input must be perceived, the single words must be recognized, and the syntactic structure of the respective utterance units must be determined. Based on this structure the meaning of the utterances will be determined. During the discourse comprehension the meaning will be embedded into the overall discourse situation. The same issues as in production models arise in models of language comprehension with respect to information flow and feedback. Disagreements exist about the architecture of the human language comprehension system, as well as about the underlying specific processes.

However, one characteristic feature that models of language comprehension should support is incrementality. There are many studies showing that as soon as a word has been recognized, it is immediately attached to the preceding information (Crocker et al. 2000). Perhaps the most convincing argument for incremental processing in sentence comprehension comes from the so-called garden-path effect. People starting to read a sentence like 'The horse raced past the barn fell' (Bever 1970) interpret the initial sequence (The horse) (raced) (past the barn) as 'subject,' 'predicate,' and 'direction.' Only after the last word has been recognized do they have a feeling of being stranded (in other words they are 'led up the garden path'), because the syntactic analysis made so far has turned out to be false. In actuality, the sentence must be interpreted as 'the horse that was raced past the barn fell.'

3. *Computer Models of Human Language Processing*

Although most computer models in production and comprehension simulate processes around the level of lexical items (cf. Norris 1999), there are also several models at the sentence level. Computer models of the mechanisms in sentence comprehension simulate how people obtain a particular syntactic and semantic analysis for a sentence (e.g., Crocker 1996, Crocker et al. 2000). Computer models of sentence production simulate how people construct, depending on retrieved lexical information, the syntactic structure of a sentence while simulating typical speech errors that occur on the sentence level (cf. de Smedt and Kempen 1987). Computer models that go beyond the sentence level to the production or comprehension of discourse turn out to be much harder to develop, since it is very difficult to determine all discourse-related parameters in experimental studies. From a broader perspective, the phenomena to be considered on the discourse level are all related to the notion of inference; they range from the interpretation of referring expressions to listener modeling.

The computer models that are presented in this section are models of various stages in language processing. The list is far from being exhaustive: It does not include all computer-implemented models that have been developed for one processing task, nor do the models completely cover all tasks in human language processing. The models presented should give an impression of the advantages of computer modeling in psycholinguistic research.

Differences in the architectural basis and the techniques used depend on the theoretical background and the system's task. Each of the computer models presented below will be characterized along the following dimensions: first, the task in human language processing, which the model simulates, will

be outlined. The architectural basis for the processing mechanisms will be characterized and the technique used for the simulation will be presented. Finally, relations between the system's behavior and empirical data will be outlined.

3.1 Computer Models of Language Production

Since models of human conceptualization are in the early stages of development, elaborated computer models of conceptualization are currently not on the market. Issues related to the topic of how people organize their pre-linguistic knowledge in order to put it into language have been addressed primarily in Artificial Intelligence research, especially in Natural Language Generation, but the models developed in these disciplines are not intended to simulate human processing. Conceptualization is still very under-developed in computational psycholinguistics.

Computer modeling starts at the interface between conceptualization and formulation, viz. the retrieval of lemmas from the lexicon, given pre-linguistic information that needs to be expressed. Models of lemma retrieval come in two forms, symbolic and connectionist approaches, but only two connectionist models shall be characterized that are based on spreading activation networks, because they show the merits of computational modeling in psycholinguistics very well.

The decompositional model of (Dell and O'Seaghdha 1992) aims at the selection of appropriate lemmas so that the relevant conceptual features can be linguistically expressed. The system's architecture is non-modular and interactive. Hence, direct feedback is possible. The system uses a network consisting *prima facie* of nodes for conceptual features and nodes for the lemmas. Bidirectionality of the links between these two layers allows this model also to be used for language comprehension. After activating the conceptual features, activation spreads towards the lemma nodes and back. Depending on the chosen parameter values for the spreading formula, the model makes several empirical predictions with respect to the activation of semantically related lemmas and the speed of activation. However, the model also incorrectly retrieves hyponyms because all features that activate a word will also activate the word's hyponyms. Hyponyms are the semantically subordinated lemmas of a given lemma. For example, if the conceptual features activate the word 'pet,' they will also activate the words 'dog,' 'cat,' etc. These hyponyms can receive a higher activation value than the target word 'pet.'

The nondecompositional model of (Roelofs 1992) also aims at the selection of lemmas, but in contrast to the previous model, concepts are represented as one node, not as conceptual features, i.e., the model assumes that concepts cannot be decomposed into conceptual primitives. In this model, activation spreads towards the lemma nodes. The lemma will be selected which receives an activation level exceeding that of other nodes by some critical amount. The model correctly describes the activation of lemmas given some activation of concepts, and it does not suffer from the hyponym problem. Furthermore, the model makes some predictions on activation speed, which have been tested empirically and turn out to be correct.

Since an evaluation of both models along all dimensions is almost impossible, only the hyponym problem is discussed further. The missing hyponym problem of the non-decompositional activation model, together with the resulting correct empirical predictions, are strong arguments for the assumption that concepts do not consist of conceptual primitives. Whether mental representations of conceptualized objects, events, times, etc. are built up from conceptual primitives or not, is a topic that has been intensively discussed for a long time in linguistics and the philosophy of mind. Since the successful implementation of a cognitive function is the proof of the theory's consistency, the results argue against conceptual decomposition.

Two systems simulating the next step in language production, namely the construction of syntactic structures of the sentences to be expressed, are now presented.

The incremental parallel formulator (IPF) (de Smedt 1990) simulates how syntactic structures are constructed in a piecemeal way, if conceptual increments are given as input at specific time points. These increments are mapped onto single words or thematic roles. The system is based on a modular architecture and uses graph unification as operation. Graph unification is the standard operation in computational grammar formalisms. It allows the combination of two informational units to a third one if the information is not contradictory. The IPF formulator simulates parallelism in formulation. By means of the time-point and accessibility of incoming conceptual fragments the system also explains specific word order variations.

The flexible incremental generator (FIG) (Ward 1992) is a connectionist model with an interactive architecture. The task of incremental sentence construction is realized on the basis of an associative network. The input for the formulation process is conceptual information that is provided with an activation value. Contrary to the previous model, this input will not be offered incrementally but as a whole. However, the sentence production process is incremental because sentences are uttered word by word. Activation spreading through the network results in the activation of words; those words with the highest activation will be uttered. The model accounts for speech errors, because non-intended concepts might receive an activation.

Although the architectural basis and processing strategies of the models are completely different, they show that incremental processing is psychologically plausible. In the first model the time span between incoming fragments corresponds directly to incrementality effects. The second model describes the incremental selection of words. However, no predictions can be derived directly from either model. They differ in their assumption about where incremental processing is involved, namely on the level of conceptualization and formulation or during formulation only. With respect to the processing mechanisms that are used, both models are equally simple, because each system uses only one operation.

It has been claimed that connectionist approaches are more robust than symbolic ones. However, for incrementality effects during sentence construction, this statement is too general. The IPF model shows that a symbolic processing strategy does not necessarily impose constraints on the data to be explained.

3.2 Computer Models of Language Comprehension

Since comprehension begins with the recognition of words, this section deals first with models of spoken word recognition. In spoken word recognition, the models proposed must answer two main questions. First, they must describe how the sensory input is mapped onto the lexicon, and second, what the processing units are in this process.

The TRACE model (McClelland and Elman 1986) is an interactive model that simulates how a word will be identified from a continuous speech signal. The problem with this task is that continuity of the signal does not provide clear hints about where a word begins and where it ends. By means of activation spreading through a network that consists of three layers (features, phonemes, and words), the system generates competitor forms, converging to the ultimately selected word. Competition is realized by inhibitory links between candidates.

The SHORTLIST model (Norris 1994) is based on a modular architecture with two distinct processing stages. It uses spreading activation as well, but in a strictly bottom-up way. Contrary to TRACE, it generates a restricted set of lexical candidates during the first stage. During the second stage, the best fitting words are linked via an activation network.

Both models account for experimental data on the time course of word recognition. Assumptions on the direction of activation flow and its nature lead to several differing predictions, but the main different prediction concerns lexical activation. While TRACE assumes that a very large number of items are activated, SHORTLIST assumes that a much smaller set of candidates is available so that the recognition of words beginning at different time points is explained differently.

The last model is a model of sentence processing. In sentence processing, one of the fundamental questions is why certain sentences receive a preferred syntactic structure and semantic interpretation.

The SAUSAGE MACHINE (Fodor and Frazier 1980) is a parsing model that assumes two stages in parsing with both stages having a limited working capacity. The original idea behind the model is to explain preferences in syntactic processing solely on the architectural basis by means of the limitations in the working memories.

The SAUSAGE MACHINE is a quasi-deterministic model, because one syntactic structure for each sentence is generated. Only if the analysis turns out to be wrong, is a reanalysis performed. Since reanalyzing a sentence is a time consuming process, the system tries to avoid this whenever possible. The model accounts for garden path effects and the difficulty in understanding multiple center embedded sentences (like 'the house the man sold burnt down'). Furthermore, the model explains interpretation preferences by means of the limitation in working memories. However, it is now understood that the architecture of a system cannot be the only factor that is responsible for processing preferences, but additional parsing principles must be assumed (Wanner 1980). Newer computational models of sentence processing show that an explanation of several phenomena in sentence processing requires the early check of partial syntactic structures with lexical and semantic knowledge (Hemforth 1993).

4. Conclusion

The outline of different computational models of human language processing shows that psycholinguistics benefits in several respects from the use of simulation models. The influence of several factors in language processing such as the directionality of activation or information flow, the link between prelinguistic and linguistic knowledge, layers of information in specific processing tasks, limitations in working memory, and many others, have been tested successfully by simulation models.

The dominant use of activation spreading as a processing mechanism for stages in language processing below the sentence level is probably due to the derivation of computational models from psycholinguistic models, which have been developed on the basis of statistical methods. Using symbolic operations is primarily inspired by linguistic investigations. Models of human language processing that originate from linguistic research do not explain the time course of the underlying mechanisms or statistically significant effects, but do explain the linguistic competence underlying the skill of speaking and comprehending. Hence, computer models aiming at various effects in sentence processing predominantly use symbolic operations to explain this competence.

This situation suggests that one future direction in computational psycholinguistics could be the development of models combining symbolic with connectionist methods to describe the interplay of various tasks in language production or comprehension.

An excellent introduction to various computer models of subtasks in human language processing is given by the contributions in Dijkstra and de Smedt (1996).

See also: Artificial Intelligence in Cognitive Science; Cognitive Modeling: Research Logic in Cognitive Science; Connectionist Models of Language Processing; Macrostructure in Discourse Comprehension, Psychology of; Narrative Comprehension, Psychology of; Psycholinguistics: Overview; Sentence Comprehension, Psychology of; Speech Production, Psychology of; Speech Recognition and Production by Machines; Syntactic Aspects of Language, Neural Basis of; Text Comprehension: Models in Psychology

Bibliography

Bever T G 1970 The cognitive basis for linguistic structures. In: Hayes J R (ed.) *Cognition and the development of language*. Wiley, New York

Crocker M W 1996 *Computational psycholinguistics. An interdisciplinary approach to the study of language*. Kluwer, Dordrecht, The Netherlands

Crocker M W, Pickering M, Clifton C 2000 (eds.) *Architectures and Mechanisms for Language Processing*. Cambridge University Press, Cambridge, UK

Dell G S, O'Seaghdha P G 1992 Stages of lexical access in language production. *Cognition* **42**: 287–314

de Smedt K 1990 IPF: An incremental parallel formulator. In: Dale R, Mellish C, Zock M (eds.) *Current Research in Natural Language Generation*. Academic Press, London, pp. 167–92

de Smedt K, Kempen G 1987 Incremental sentence production, self-correction and co-ordination. In: Kempen G (ed.) *Natural Language Generation. New Results in Artificial Intelligence, Psychology and Linguistics*. Nijhoff (Kluwer Academic Publishers), Dordrecht, The Netherlands, pp. 365–76

Dijkstra T, de Smedt K (eds.) 1996 *Computational psycholinguistics*. Taylor & Francis, London

Fodor J D, Frazier L 1980 Is the human sentence parsing mechanism an ATN? *Cognition* **8**: 417–59

Gernsbacher M A 1994 (ed.) *Handbook of psycholinguistics*. Academic Press, San Diego, CA

Hemforth B 1993 *Kognitives Parsing. Repräsentation und Verarbeitung sprachlichen Wissens*. [Cognitive parsing. The representation and processing of linguistic knowledge.] Infix, St Augustin

Levelt W J M 1989 *Speaking*. MIT Press, Cambridge, MA

McClelland J L, Elman J L 1986 The TRACE model of speech perception. *Cognitive Psychology* **18**: 1–86

Norris D 1994 SHORTLIST: A connectionist model of continuous speech recognition. *Cognition* **52**: 189–234

Norris D 1999 Computational psycholinguistics. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 168–70

Roelofs A 1992 A spreading-activation theory of lemma retrieval in speaking. *Cognition* **42**: 107–42

Wanner E 1980 The ATN and the sausage machine: Which one is baloney? *Cognition* **8**: 209–25

Ward N 1992 A parallel approach to syntax for generation. *Artificial Intelligence* **57**: 183–225

Weizenbaum J 1966 ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM* **9**(1): 36–45

R. Klabunde

Computer Networking for Education

Computer networks such as the Internet and the World Wide Web (WWW) have an increasingly high impact in schools as well as in higher and further education. Teachers and researchers try to analyze both the current potential and future developments (e.g., Roschelle and Pea 1999) in computer networking in education. In this article we consider aspects of computer networking for education at the individual level, the group level, the level of the organization, and the societal level in turn. We report on important basic scenarios and developments as well as on approaches to improving learning and education.

1. Computer Networking and the Individual

1.1 Basic Scenarios

With the advent of the WWW, the learning individual has moved back into the focus of attention, especially in the context of teaching remote learners and using the WWW as a teaching resource.

1.1.1 Teaching remote learners. The internet is seen as an ideal medium for delivery and communication between teacher and learner. In this context teleteaching becomes increasingly more important. In teleteaching, a lecture is typically held in a university lecture-room and is synchronously transmitted through two-way interactive video to university students at another university. Another approach to teaching distant learners is that of online course delivery. In the tradition of distance learning, learners receive course material as computer files or on web pages. By e-mail or by videoconferencing they can ask their teachers for assistance.

1.1.2 Using the Internet as resource. The second major application of computer networks for individual learning is the use of existing Internet-based

material as resources to prepare and to elaborate on instruction. Students search the Internet in the sense of 'knowledge mining' (Roschelle and Pea 1999), trying to explore the knowledge of a certain expert community by analyzing web sites or newsgroup discussions. Teachers are using the Internet as resource as well. A large number of publications in education-related journals describe web-based teaching resources for use during classroom lectures. In many cases, annotated listings of useful Internet and web addresses are included.

1.2 Approaches to Further Improve Individual Learning

1.2.1 Reduce complexity by indexing. Free exploration of Internet resources proves to be overwhelming for many learners with little prior knowledge. Using the indexing strategy of web-based instruction, available topic-related web sites are selected, indexed, and evaluated by teachers or experts both for lecture preparation and follow-up, as well as for enriching classroom lectures.

1.2.2 Keep complexity, provide support. This approach highlights the importance of authentic and complex learning scenarios for the acquisition and transfer of knowledge. Instead of reducing the complexity of Internet-based resources, this approach emphasizes supporting learners to work effectively with these authentic materials. For instance, expert modeling of strategies, mapping techniques or strategy guidance have proved to be effective instructional tools in complex learning environments (see Mandl et al. 1999). Analysis and training of cognitive strategies for searching the web, and knowledge management strategies, are other variants of this approach (e.g., Hill and Hannafin 1997). In teletutoring or telecoaching approaches, the promotion of knowledge transfer to the workplace is the primary goal. Learners are at their workplaces and receive assistance via network technologies. This assistance is typically provided by a coach, who supports the learner in applying knowledge to new problems.

2. Computer Networking and the Group

2.1 Basic Scenarios

2.1.1 Technology-based learning within the classroom. A series of approaches focus on the new and enriching possibilities of network technologies for learning within traditional educational settings. Learners in a class are engaged in processes of collaborative knowledge construction and are supported by network technologies which are more or less specifically designed for group learning purposes. One central assumption is that in text-based communication in particular, social inequalities known from traditional settings can be partially resolved (e.g., Riel 1995). Variants of these approaches are 'classroom-meets-expert-community' projects. Students contact experts in the field via videoconferencing, by e-mail, or via newsgroups. It is an important goal of these projects to provide learners with acces to an appropriate expert community (e.g., Gomez et al. 1998).

2.1.2 Virtual learning environments. A further important use of computer networks is in connecting distant learners to form a virtual classroom, a virtual seminar, or a virtual tutorial (e.g., Harasim et al. 1995). In a large number of projects, the improved flexibility of learning with respect to both location and time is seen as the particular added value of computer networking in education. People can learn together who would hardly be able to do so in traditional instructional settings because of distance, disabilities, or incompatible time schedules.

2.1.3 Internet-based cultural encounters. In Internet-based encounters, socially or geographically distant groups or cultures are brought together on a classroom or seminar level by means of Internet facilities. In e-mail exchanges, learners should become aware of cultural diversity, are encouraged to reduce stereotypes, and to practice languages (Fabos and Young 1999). Moreover, electronic 'field trips' are arranged to explore a distant part of the world in order to enhance classroom lectures in different subject matter. Typically, an expedition team publishes its latest photos and reports on a journey on web pages and communicates by e-mail with several dozen classes all over the world.

2.2 Approaches to Further Improve Cooperative Learning

2.2.1 The moderation approach. In virtual cooperative learning environments, the role of the teacher or tutor will become more and more that of a facilitator or moderator, structuring group discussion according to certain theoretical principles.

2.2.2 The cooperative learning tools approach. This approach argues that today's networking tools are not genuinely designed for learning purposes (Roschelle and Pea 1999). There are efforts to develop, adapt, and apply networked collaborative learning tools to facilitate learning in groups. Shared active representation tools are designed to support learners and teachers in discussing complex subject matter by advancing domain-specific external representations, including text or graphic representations or mathe-

matical notations. Community-building tools are those which define social places and virtual worlds for new forms of cooperation (Roschelle and Pea 1999). Examples are MUDs (multiuser domains) as well as more complex environments using and combining technologies such as videoconferencing (Fischer and Mandl, in press), chat, or computer conferencing (Schwan 1997). Sociocognitive structuring tools organize learning activities according to successful patterns of tutorial dialogs or peer discussions. These tools prestructure or limit the kinds of possible contributions to a discussion and help to connect the messages in a meaningful way (Linn et al. 1999).

3. Computer Networking and Organizations

3.1 Learning Organizations

Schools, universities, and other places of learning are becoming 'learning organizations' themselves (e.g., Cavaleri and Fearon 1996). Learning organizations develop mechanisms helping them to adapt effectively to new situations and to use available resources more effectively to advance their members' development. It has been argued that such organizations will have crucial advantages in the competition among educational organizations. Computer networks (e.g., Intranets) may play an important role in these changes by supporting the sharing of resources, as well as by facilitating access to them. They enable new forms of communication and cooperation among learners and among educators as well as between learners and educators. Teachers and learners will have access to scientific resources and may contact other teachers as well as experts in the field. However, while implementing computer networks is an important step in establishing a culture of learning in an organization, it is not the decisive step. Students and teachers must be both enabled and motivated to contribute to shared resources of an organization.

3.2 Approaches to Further Improve Computer Networking in Organizations

An important way to advance a new culture of learning in educational organizations is represented by the 'learning communities' approach. In a learning community the major focus is on fostering collective knowledge and in that way supporting the growth of individual knowledge (Bielaczyc and Collins 1999). A learning community seeks out knowledge from multiple sources and shares its findings widely among the members. An excellent example of implementing a learning communities approach based on computer networks within a school context is the 'knowledge building communities' approach within the Computer Supported Intentional Learning Environments (Scardamalia and Bereiter 1994).

4. Computer Networking and Society

4.1 Globalization

Globalization has its impact on education as well. The possibilities of offering educational services and courses worldwide via computer networks are used more, and more particularly by large institutions. Increasing competition in the field of education might ensure that theoretical and practical innovations result in a positive modification of contents and methods of education. In recent work, however, both the clear dominance of Western culture and the issue of cultural specificity of education receive critical attention (Fabos and Young 1999).

An increasing reallocation of educational resources into network-based learning environments is frequently regarded as causing the emergence of a new 'two-class' society. People who do not have the basic knowledge necessary to use the new technologies have increasing difficulty in accessing relevant educational content. Fundamental prerequisites for the prevention of this development are missing at present; for instance, schools lack sufficient computers and network technologies.

4.2 Approaches to Improve Education

Providing educational institutions with an adequate supply of computer and networking technologies is often regarded as the basic condition for equal access to education. However, *media literacy* does not result from simply supplying the technologies. One can expect that all who take part in the processes of education—learners, teachers, researchers, parents—increasingly need media literacy to negotiate computer networks: knowledge and skills, which enable them to communicate, to learn and to become informed via such networks. Moreover, important capabilities of teachers and parents include being able to assess and reduce possible threats posed, for instance, by pornography and violence on the Internet.

Increasingly, technologies are seen as more or less flexibly developing tools which may advance societal visions and goals. For example, a number of American projects in the context of computer networking for education have aimed at nationwide educational reform goals such as equity or diversity (e.g. Garner and Gillingham 1996).

5. Outlook

A predominant trend in research on computer networking for education is to develop complex learning environments employing different tasks, as well as media mixes with multimedia, synchronous and asynchronous text-based communication or videoconferencing. With testbed designs or design experiments

(Gomez et al. 1998), researchers try to design and evaluate those prototypical scenarios based on instructional theories and tools as well as on constraints of the content to be learned, learners' needs, and organizational limitations. We believe that this important evaluative and design-related research should be accompanied by more controlled experimental studies analyzing specific interactive and cognitive effects resulting from specific types of cooperative learning tasks supported by specific technologies.

See also: Computer-assisted Instruction; Computers and Society; Information Technology; Internet: Psychological Perspectives; Learning: Organizational; Technological Innovation; Technology-supported Learning Environments

Bibliography

Bielaczyc K, Collins A 1999 Learning communities: A reconceptualization of educational practice. In: Reigeluth C M (ed.) *Instructional Design Theories and Models*, Vol. 2. Lawrence Erlbaum, Mahwah, NJ

Cavaleri S, Fearon D (eds.) 1996 *Managing in Organizations that Learn*. Blackwell Publishers, Oxford, UK

Fabos B, Young M D 1999 Telecommunication in the classroom: Rhetoric versus reality. *Review of Educational Research* **69**(3): 217–59

Fischer F, Mandl H (in press) Being there or being where? Videoconferencing and cooperative learning. In: Oostendorp H von (ed.) *Cognition in a Digital World*. Lawrence Erlbaum, Mahwah, NJ

Garner R, Gillingham M G 1996 *Internet Communication in Six Classrooms: Conversations Across Time, Space, and Culture*. Lawrence Erlbaum, Mahwah, NJ

Gomez L M, Fishman B J, Pea R D 1998 The CoVisProject: building a large-scale science education testbed. *Interactive Learning Environments*. **6**: 59–92

Harasim L, Hiltz S R, Teles L, Turoff M 1995 *Learning Networks: A Field Guide to Teaching and Learning Online*. MIT Press, Cambridge, MA

Hill J R, Hannafin M J 1997 Cognitive strategies and learning from the World Wide Web. *Educational Technology Research and Development* **45**(4): 37–64

Linn M, Bell P, Hsi S 1999 Using the Internet to enhance student understanding of science: The Knowledge Integration Environment. *Interactive Learning Environments* **1**(6): 4–38

Mandl H, Gräsel C, Fischer F 2000 Facilitating problem-oriented learning: Facilitating the use of domain-specific and control strategies through modeling by an expert. In: Flammer A, Perrig W J (eds.) *Control of Human Behavior, Mental Processes and Consciousness*. Lawrence Erlbaum, Mahwah, NJ

Riel M 1995 Cross-classroom collaboration in global learning circles. In: Star S L (ed.) *The Cultures of Computing*. Blackwell Publishers, Oxford, UK, pp. 219–42

Roschelle J, Pea R 1999 Trajectories from today's WWW to a powerful educational infrastructure. *Educational Researcher* **28**(5): 22–543

Scardamalia M, Bereiter C 1994 Computer support for knowledge-building communities. *The Journal of the Learning Sciences* **3**(3): 265–83

Schwan S 1997 Media characteristics and knowledge acquisition in computer conferencing. *European Psychologist* **2**(3): 277–86

H. Mandl and F. Fischer

Computer-assisted Instruction

1. Towards a Broader Conceptualization of Computer Use in Education

Since the introduction of computer use in education in the 1960s, its terminology has continuously evolved. Some contributors to the new field adopted the term Computer Assisted Instruction (CAI) for a specific approach, either focusing on a type of program or a type of instructional process. Steinberg (1991), for example, emphasized CAI as computer-presented instruction that is individualized, interactive, and guided. Related terms are also used, such as computer-based instruction, computer-based education, and computer-assisted learning. Often, these terms refer to broader conceptualization, of different kinds of computer use in education. Watson (1994), for instance, used the term computer-assisted learning for the whole variety of ways in which the computer is used for learning.

The rather confusing terminology is partly due to rapid technological changes. At the time when computers were introduced in education, big mainframe computers were still in use. Computers were conceived as being able to realize programmed instruction and to replace teachers. By the twenty-first century, computers have become personal and networked. The potential for education has particularly increased due to the integration of computers with communication systems including audio and video technology. The term computer technology is now replaced by information technology (IT) in North America, or by information and communication technology (ICT) in Europe. The latter term—ICT—will be used throughout this contribution.

2. Views on the Use of ICT in Education

It is generally accepted that the increasing impact of ICT on our society is also influencing teaching and learning. Some expect that ICT can reduce several problems in present day education, such as the need for individualized instruction and the poor transfer of school knowledge to real life (Sandholtz et al. 1997). Also, new demands from society require education to focus on information management and communication skills. ICT can be a valuable essential tool in learning these skills. Further, in constructivist

approaches to teaching and learning, students are held responsible for their own learning. Information technology is considered to offer promising environments and tools to support this process.

The high aspiration level of ICT in education is well reflected in Dede's (2000) list of ICT capabilities for enhancing learning: centering the curriculum around authentic problems parallel to those in real world settings; involving students in virtual communities of practice, using advanced tools similar to those in today's high-tech workplaces; facilitating guided, reflective inquiry through extended projects that inculcate sophisticated concepts and skills and generate complex products; utilizing modeling and visualization as powerful means of bridging between experience and abstraction; enhancing students' collaborative construction of meaning via different perspectives on shared experiences; including pupils as partners in developing learning experiences and generating knowledge; and fostering success for all students through special measures to aid the disabled and the disenfranchised.

Itzkan (1994) made an effort to interpret developments in ICT in terms of educational innovations. He postulated that the potential of technology directs the way education will develop. In his view, three stages can be distinguished. The 'substitution' stage marks the ICT applications that are able to automate parts of the instructional process. In this stage, ICT is solely being used as a replacement for tasks of the teacher, with the specific aim to make education more effective, or to provide more possibilities for individualizing education. In the stage of 'transition,' ICT is not only used for replacing existing tasks, but it also requires that instructional practices and content are changed. In this stage, ICT applications not only structure the learning process, but students themselves need to increasingly structure their own learning process. In the 'transformation' stage, not only must instructional practices and content be altered, but also the underlying rationale of education should change. An example of the latter is the development in network technology that may result in virtual learning communities where students, teachers, and scientists are involved. Such a situation affords greater opportunities for learning to take place in interaction with practicing professionals than might otherwise be possible with a few field trips per year.

3. *Applications of ICT in Education*

Early software packages for education were mainly text based and used only simple graphics. In the twenty-first century software is much more sophisticated. Different combinations of modalities—pictures, moving images, sound and text—are integrated into electronic environments. An important development is the increasing possibility to distribute ICT applications not only via stand-alone computers, but also in a networked environment or via the World Wide Web (WWW). The teachers involved are able to guide their students from remote locations, creating new possibilities for distance education.

The following types of applications are designed for education.

3.1 *Drill and Practice, Tutorials*

Software programs in this category serve as an assistant for teachers by taking over some of their tasks. The software often has facilities to provide feedback to the learners and to keep track of their performance. With these applications it is possible to individualize instruction. Drill and practice software is used to automate practicing basic knowledge and skills. Tutorials can be used for learning new educational content. The programs differ greatly in their complexity. However, drill and practice programs are usually quite simple, while tutorials can be very complex (Watson 1994).

3.2 *Simulations and Modeling Systems*

Simulations are representations of a part of reality. By manipulating variables in these representations and studying the effects, students can gain an understanding of the reality. Through simulations it may be possible to study parts of reality which cannot otherwise be studied for reasons of safety, cost, required apparatus, or scale. Simulations can be an aid in visualizing abstract concepts. They serve as a bridge between reality and the student's mental model of reality. Simulations in the twenty-first century are not developed for education only, but also many have been made for entertainment purposes, used at homes outside the school context. With modeling systems students can build their own (mathematical) model of a part of reality. In doing so they may gain understanding of complex relations. Modeling systems generally do not have built-in teacher control, while simulations offer a certain kind of teacher control through the model, the display, and the framework of interactions (Cox 1994).

3.3 *Multimedia Cases*

Multimedia cases offer the possibility to study parts of reality by presenting illustrations of real world practice that can be discussed and studied from several perspectives. Multimedia cases usually combine video segments, audio, and textual commentary together with random access facilities and hyperlinks, and so create an open-ended student-centered learning environment. Though used in many fields of study, they are

increasingly used in teacher education for studying the complexity of the classroom from various points of view.

3.4 Microcomputer-based Laboratories

These provide an environment for science education. The computer is used as a support tool for students' laboratory work. Real-time data collection is possible. It becomes much easier to repeat experiments, to measure different variables at the same time, to use a very short or a very long time range, to analyze data, and to represent data graphically. Instead of spending instructional time on data collection, it can be used for analyzing and interpreting data (Voogt 1996).

In addition to applications that are specifically designed for education, many generic software tools are available that can be used to support the learning process. Lacking a particular instructional content renders such tools multi-functional. They are increasingly applied as a substantial part of the students' learning environment, going beyond the function of paper and pencil or a calculator.

3.5 Word Processor, Spreadsheets, Presentation Software, Graphics Software

The software programs in this category are part of the basic application software of nearly every PC. Their use in education is growing fast (Becker et al. 1999). They support students in their learning process, because of their capability to structure information, and to easily present information in different modes. Research related to the use of word processors in education in particular shows the potential of this tool to stimulate students to write and to support the writing process.

3.6 E-mail, Videoconferencing, Discussion Boards

This software represents the increasing role of ICT in communication. They have shown to be particularly useful in the design of computer-supported collaborative learning environments (Koschmann 1996).

3.7 Digital Encyclopedias and Databases

This category of multimedia is used as a local information resource, to be compared with the function of the traditional school library. Closely related to these applications is the WWW as a source of information. The multimedia in this category anticipates the major role information has in our present day society. The potential scope of available information requires that students possess excellent information management skills.

4. Current Patterns in ICT Use

A worldwide survey on computers in education (Pelgrum and Anderson 1999) shows a rapid decline between 1995 and 1998 in student–computer ratios in all levels of education. Despite this fact, the actual use of computers in schools is still rather limited. Recent figures from the USA (Becker et al. 1999) show that only about one third of the teachers use computers on a regular basis, although the majority has a computer in their classroom. Remarkable is that software programs used most by teachers are not the programs that are specifically designed for education. From a worldwide perspective, it is known that in all levels of education, word processing is the most popular ICT application used in schools. Only in elementary education is drill and practice software still frequently used. The use of more sophisticated software, such as simulations or microcomputer-based laboratories, are only used in very small numbers of schools in many countries. Becker (2000) found that these kinds of applications are more likely to be used by those few teachers that hold a constructivist teaching approach. However, also among this group of teachers, generic software tools are still used considerably more than applications specifically designed for education.

5. Implementation of ICT: The Central Role of the Teacher

There are several obstacles that hinder the integration of ICT in education. An important limitation of many educational software programs is that they are poorly attuned to the curriculum or they only cover minor parts of the curriculum (Van den Akker et al. 1992). This makes it less attractive for teachers to invest their efforts (and budgets) in applying new media in their instruction. The international study of Pelgrum and Anderson (1999) shows the difficulties of integrating computers into classroom practices. Not only are many educational software programs quite isolated, but it also appears that teachers lack the competence of integrating computers in their instructional processes. Compared to the early 1990s, at the start of the twenty-first century many teachers have received training in basic ICT skills and knowledge. However, very few have been fortunate to receive continuous education that centered on the pedagogical use of ICT in instruction. Olson (2000) argues that the use of ICT in schools often does not fit into the existing teaching culture and that it may even undermine the teacher's sense of efficacy. The use of ICT often assumes a change in pedagogical approaches and classroom management strategies, implying that teachers have to change their routines (Sandholtz et al. 1997). One possible way to support such teacher learning is with carefully designed and validated curriculum materials that contain procedural specifications in order to guide teachers on essential, but vulnerable, aspects of ICT

integration (Voogt 1993). The impact of this kind of curriculum materials on teacher learning benefits from use and sharing in teacher networks that focus on the integration of ICT in education (Moonen and Voogt 1998).

6. Attainments for Students

It is widely assumed that ICT has high potential for improving education. However, research has not yet provided convincing evidence regarding the impact of ICT on student performance. Until now, most research has been restricted to simple drill and practice kinds of programs. In a meta-analysis Niemiec and Walberg (1987) found a moderate effect of CAI on enhancing student outcomes. Other findings focus on other possible student gains, such as an increase in the rate of learning, better retention, and a more positive attitude towards learning.

An important problem in establishing the cognitive attainments for students is that the use of ICT often aims to contribute to the mastery of complex cognitive skills. These types of skills cannot be determined by means of simple, standardized tests. The complexity of the problem is illustrated in the *Computer as a Learning Partner* project (Linn and Hsi 2000). Students involved in that project did not score better on multiple choice items in standardized tests that required recall, but they outperformed students on items that required interpretation. Yet, the outcomes of the project had much more impact on student learning than could be determined in standardized tests. The project could demonstrate that comparing subsequent versions of an ICT-rich curriculum for science education resulted in a 400 percent increase—over eight versions of the curriculum—in student understanding of the complex science concepts that were dealt with in the curriculum.

7. Research on ICT in Education

Much research has guided the development of new ICT tools for education. These studies are often long lasting, but relatively limited in scale. *The Computer as a Learning Partner* project (Linn and Hsi 2000) is an outstanding example of such a study that shows the potential of ICT for curriculum and instruction. Over a period of 15 years, an ICT-rich science curriculum focusing on understanding of complex science concepts for middle school students has been developed, extended and refined. Another promising example was the Jasper project in the mathematics domain (The Coginition and Technology Group at Vanderbilt 1997). However, the implementation of such curricula and related ICT tools in typical classrooms appeared very difficult to achieve. Also large-scale survey studies have been carried out that investigated the implementation of ICT mainly with the purpose to inform policy makers. Although these studies revealed general problems with ICT integration, they could not provide concrete and specific suggestions for improvement of the situation. More recent studies carried out by the Organization of Economic Cooperation and Development and the International Organization for the Evaluation of Educational Achievement focus on existing 'best practices' of ICT use in education with the purpose to get a deep understanding of the teaching and learning practices that contribute to effective and innovative use of ICT and the conditions that contribute or hinder its implementation. All these studies underline that the integration of ICT use in the instructional practice will remain a fascinating but difficult endeavor.

See also: Computer Networking for Education; Computers and Society; Human–Computer Interaction; Human–Computer Interface; Information Technology; Libraries: Digital, Electronic, and Hybrid; Technological Innovation; Technology-supported Learning Environments

Bibliography

Becker H J, Ravitz J L, Wong Y T 1999 *Teacher and Teacher-directed Student Use of Computers and Software*. University of California, Irvine, CA

Becker H J 2000 *Findings from the Teaching, Learning, and Computing Survey: Is Larry Cuban Right?* Revision of a paper written for the January 2000 School Leadership Conference of the Council of Chief Officers, Washington, DC

Cox M J 1994 The computer simulations and modeling. In: Husén T, Postlewthwaite T N (eds.) *International Encyclopedia of Education*, 2nd edn. Pergamon Press, Oxford, UK, pp. 985–8

Dede C 2000 Emerging influences of information technology on school curriculum. *Journal of Curriculum Studies* **32**(2): 281–303

Itzkan S J 1994 Assessing the future of telecomputing environments: Implications for instruction and administration. *The Computing Teacher* **22**(4): 60–4

Koschmann T 1996 *CSCL: Theory and Practice of an Emerging Paradigm*. Lawrence Erlbaum Associates, Mahwah, NJ

Linn M C, Hsi S 2000 *Computers, Teachers, Peers: Science Learning Partners*. Lawrence Erlbaum Associates, Mahwah, NJ

Moonen B, Voogt J 1998 Using networks to support the professional development of teachers. *Journal of In-service Education* **24**(1): 99–110

Niemiec R, Walberg H J 1987 Comparative effects of computer-assisted instruction: A synthesis of reviews. *Journal of Educational Computer Research* **3**: 19–37

Olson J 2000 Trojan horse or teacher's pet? Computer and the culture of the school. *Journal of Curriculum Studies* **32**(1): 1–8

Pelgrum W J, Anderson R A (eds.) 1999 *ICT and the Emerging Paradigm for Life Long Learning: A Worldwide Educational Assessment of Infrastructure, Goals and Practices*. IE, Amsterdam

Sandholtz J, Ringstaff C, Dwyer D C 1997 *Teaching with Technology: Creating Student-centered Classrooms*. Teachers College Press, New York
Steinberg E R 1991 *Computer-assisted Instruction: A Synthesis of Theory, Practice and Technology*. Lawrence Erlbaum Associates, Hillsdale, NJ
The Cognition and Technology Group at Vanderbilt 1997 *The Jasper Project. Lessons in Curriculum, Instruction, Assessment and Professional Development*. Lawrence Erlbaum Associates, Mahwah, NJ
Van den Akker J, Keursten P, Plomp T 1992 The integration of computer use in education. *International Journal of Educational Research* **17**(1): 65–76
Voogt J M 1993 *Courseware for an Inquiry-based Science Curriculum*. Doctoral dissertation. University of Twente, Enschede, The Netherlands
Voogt J M 1996 Microcomputer based laboratories and science education. In: Plomp T J, Ely D P (eds.) *International Encyclopedia of Educational Technology*. Pergamon Press, Oxford, UK, pp. 448–50
Watson D M 1994 Computer-assisted learning. In: Husén T, Postlethwaite T N (eds.) *International Encyclopedia of Education*, 2nd edn. Pergamon Press, Oxford, UK, pp. 988–92

J. Voogt and J. van den Akker

Computerized Test Construction

Like many other areas of psychology, the availability of cheap, plentiful computational power has revolutionized the technology of educational and psychological testing. It is no longer necessary to restrict testing to the use of items with a paper-and-pencil format in a group-based session. We now have the possibility to build multimedia testing environments to which test takers respond by manipulating objects on a screen, working with application programs, or manipulating devices with built-in sensors. Moreover, such tests can be assembled from banks with items stored in computer memory and immediately delivered to examinees who walk in when they are ready to take the test.

Computerized assembly of tests from an item bank is treated as an optimization problem with a solution that has to satisfy a potentially long list of statistical and non-statistical specifications for the test. The general nature of this optimization problem is outlined, and applications to the problems of assembling tests with a linear, sequential, and adaptive format are reviewed.

1. Test Assembly as an Optimization Problem

The formal structure of a test assembly problem is known as a constrained combinatorial optimization problem. It is an optimization problem because the test should be assembled to be best in some sense. The problem is combinatorial because the test is a combination of items from the bank and optimization is over the space of admissible combinations. Finally, the problem is constrained because only those combinations of items that satisfy the list of test specifications are admissible.

The quintessential combinatorial optimization problem is the knapsack problem (Nemhauser and Wolsey 1988). Suppose a knapsack has to be filled from a set of items indexed by $i = 1,\ldots, I$. Each item has utility u_i and weight w_i. The optimal combination of items is required to have maximum utility but should not exceed weight limit W. The combination is found defining decision variables

$$x_i \equiv \begin{cases} 1 & \textit{if item } i \textit{ is selected} \\ 0 & \textit{otherwise} \end{cases} \tag{1}$$

and solving the problem

$$\max \sum_{i=1}^{I} u_i x_i, \quad (\textit{maximum utility})$$

subject to

$$\sum_{i=1}^{I} w_i x_i \leqslant W, \quad (\textit{weight limit})$$

$$x_i = 0, 1, \quad (\textit{range of variables})$$

for an optimal set of values for variables x_i.

Problems with this structure are known as 0–1 linear programming (LP) problems (e.g., see *Linear and Nonlinear Programming*). Several test assembly problems can be formulated as a 0–1 LP problem; others need integer variables or a combination of integer and real variables. In a typical test assembly model, the objective function is used to maximize a statistical attribute of the test whereas the constraints serve to guarantee its content validity.

1.1 Objective Functions in Test Assembly Problems

Suppose the items in the bank are calibrated using an item response theory (IRT) model, for example, the two-parameter logistic (2PL) model:

$$p_i(\theta) \equiv \Pr(U_i = 1 \mid \theta) \equiv \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]} \tag{2}$$

where $\theta \in (-\infty, \infty)$ is the ability of the examinee, and $b_i \in (-\infty, \infty)$ and $a_i \in [0, \infty]$ are parameters for the difficulty and discriminating power of item i, respectively (e.g., see *Factor Analysis and Latent Structure: IRT and Rasch Models*).

A common objective function in IRT-based test assembly is based on the test information function, which is Fisher's measure of information on the unknown ability θ in the response vector $U_1,\ldots, U_n$, where n is the number of items in the test. For the 2-PL model the test information function is given by

$$I(\theta) \equiv I_{U_1 \ldots U_n}(\theta) = \sum_{i=1}^{n} \frac{[p_i'(\theta)]^2}{p_i(\theta)[1-p_i(\theta)]}$$

with $p_i'(\theta) \equiv (\partial/\partial\theta)p_i(\theta)$. Test information functions are additive in the contributions by the individual items, which are denoted by $I_i(\theta)$.

The first step in IRT-based test assembly is to formulate a target for the test information function. The next step is to assemble the test to have its actual information function as closely as possible to the target. Examples of popular targets are a uniform function over an ability interval in diagnostic testing and a peaked function at a cutoff score θ_c in admission testing. An objective function to realize the former is presented in Eqns. (4) and (5) (Sect. 2). The latter can be realized using objective function

$$\max \sum_{i=1}^{n} I_i(\theta_c)x_i \tag{3}$$

under the condition of an appropriate set of constraints on the test. Other possible objective functions are maximization of classical test reliability and minimization of the length of the test. For a review of these and other examples, see van der Linden (1998).

1.2 Constraints in Test Assembly Problems

Formally, test specifications can be viewed as a series of upper and/or lower bounds on numbers of item attributes in the tests or on functions thereof. An important distinction is between constraints on (a) categorical item attributes, (b) quantitative item attributes, and (c) logical relations between the items in the test. Categorical attributes are attributes such as item content, cognitive level, format, and use of graphics. Examples of quantitative attributes are statistical item parameters, expected response times, word counts, and readability indices. Logical (or Boolean) constraints deal with such issues in test assembly as items that cannot figure in the same test because they have clues to each other's solution or items that are organized as sets around common stimuli.

Let V_j be a set of items in the bank with a common value for an attribute. The general shape of a constraint on a categorical item attribute in a test assembly model with 0–1 decision variables is

$$\sum_{i \in V_j} x_i \gtreqless n_j$$

where n_j is a bound on the number of items from V_j. This type of constraint can also be formulated on intersections or unions of sets of items.

Constraints on quantitative attributes are typically on a function of their values for a set of items. For example, if a typical test taker has response time t_i on item i and the total testing time available is T (both in seconds), a useful constraint on the test is:

$$\sum_{i=1}^{I} t_i x_i \leqslant T$$

As an example of a logical constraint, suppose that W_s represents a set of items in the bank with a common stimulus s, and that n_s items have to be selected if and only if stimulus s has. This requirement leads to the following constraint

$$\sum_{i \in V_s} x_i = n_s z_s$$

with z_s being an auxiliary 0–1 decision variable for the selection of stimulus s.

In a full-fledged test assembly problem, constraints may also be needed to deal, for instance, with stimulus attributes or with relations between different test forms if a set of forms is to be assembled simultaneously. For a review of these and other types of constraints, see van der Linden (1998, 2000a).

2. Linear-test Assembly

Linear tests have a fixed number of items presented in a fixed order. For measurement over a larger ability interval, it is customary to choose a discrete set of target values for the information function, $T(\theta_k)$, $k = 1,\ldots, K$. In practice, because information functions are well-behaved continuous functions, target values at three to five equally-spaced θ values suffice. The need to match more than one target value simultaneously creates a multiobjective decision problem.

An effective way to deal with multiple target values is to apply a maximin criterion. This criterion leads to the following core of a test assembly model

$$\max y \quad (\textit{common factor}) \tag{4}$$

subject to

$$\sum_{i=1}^{I} I_i(\theta_k)x_i \geqslant T(\theta_k)y \quad k = 1,\ldots, K, \qquad (\textit{minimum information at } \theta_k) \tag{5}$$

where y is a common factor in the right-hand side bounds in (5) that is maximized and coefficients $T(\theta_k)$ control the shape of the information function (van der Linden and Boekkooi-Timminga 1989). Constraints to deal with the remaining content specifications

should be added to this model. For a large-scale testing program in education, it is not unusual to have hundreds of those constraints.

Methods to solve test assembly models can be distinguished into algorithms that have been proven to lead to optimality and intuitively plausible heuristics which typically select one item at a time. Well-known heuristics are those that pick the items with the largest impact on the test information function (Luecht 1998) or with the smallest weighted average deviation from all bounds in the model (Swanson and Stocking 1993). Optimal solutions can be found using a branch-and-bound algorithm (Nemhauser and Wolsey 1988), or, if the structure of the models boils down to a network-flow problem, a simplified version of the simplex algorithm (Armstrong et al. 1995). Several algorithms and heuristics are implemented in the test assembly package ConTest (Timminga et al. 1996).

Sometimes it is necessary to build a set of linear test forms, for instance, parallel forms to support different testing sessions or forms of different difficulties for use in an evaluation study with a pretest-posttest design. Sequential application of a model of linear-test assembly is bound to show a decreasing quality of the solutions. A simultaneous approach balancing the quality of the individual test forms can be obtained by replacing the decision variables in Eqn. (1) by

$$x_{if} \equiv \begin{cases} 1 & \textit{if item } i \textit{ is selected for } f \\ 0 & \textit{otherwise} \end{cases} \tag{6}$$

using these variables to model the test specifications for all forms $f = 1,\ldots, F$, and solving the model for all variables simultaneously. If no overlap between forms is allowed, logical constraints must be added to the model to prevent the variables from taking the value of one more than once. Efficient combinations of sequential and simultaneous approaches are presented in van der Linden and Adema (1998).

3. Sequential Test Assembly

Sequential test assembly is used in testing for selection or mastery with a cutoff score on the test that represents the level beyond which the test taker is accepted or considered to master the domain of knowledge tested, respectively. An obvious linear approach is to assemble a test using the objective function in Eqn. (3), but a more efficient procedure is to assemble the test sequentially, sampling one item from the bank at a time and stopping when the test taker is classified with enough precision.

If the items are dichotomous, the number-correct score of a test taker follows a binomial distribution

$$\Pr(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$$

with π being the success parameter and n the (random) number of items sampled. In a sequential probability ratio test (SPRT) for the decision to reject test takers with $\pi < \pi_0$ and accept those with $\pi \geqslant \pi_1$, with (π_0, π_1) being a small interval around cutoff score π_c, the decision rule is based on the likelihood ratio

$$\lambda_n \equiv \prod_{i=1}^{n} f(x_i|\pi_1)/f(x_i|\pi_0)$$

If $\lambda_n < A$ or $\lambda_n \geqslant B$ the decision is to reject or accept, respectively, whereas sampling of items is continued otherwise. The constants A and B are known to satisfy

$$\begin{aligned} A &\geqslant \beta(\pi_1)/\{1-\alpha(\pi_0)\} \\ B &\leqslant \{1-\beta(\pi_0)\}/\alpha(\pi_0) \end{aligned} \tag{7}$$

with $\alpha(\pi_0)$ and $\beta(\pi_1)$ being the probabilities of a false positive and false negative decision for test takers with $\pi = \pi_0$ and $\pi = \pi_1$, respectively. For more on sequential methods, see *Sequential Statistical Methods*.

Alternative sequential approaches to test assembly follow an IRT-based SPRT (Reckase 1983) or a Bayesian framework (Kingsbury and Weiss 1983).

4. Adaptive Test Assembly

If the items in the pool are calibrated using a model as in (2), adaptive test assembly becomes possible. In adaptive test assembly, the items are selected to be optimal at the ability estimate of the test taker, updated by the computer after each new response. Adaptive testing leads to much shorter tests; savings are typically over 50 percent relative to a linear test with the same precision.

To show the principle of adaptive testing, let $i = 1,\ldots, I$ denote the items in the pool and $k = 1,\ldots, K$ the items the test. It follows that i_k is the index of the item in the pool administered as the kth item in the test. The set $S_k \equiv \{i_1,\ldots, i_{k-1}\}$ contains the first $k-1$ items in the test. These items involve responses variables $U_{k-1} \equiv (U_{i_1},\ldots, U_{i_{k-1}})$. The update of the ability estimate after $k-1$ responses is denoted as $\hat{\theta}_{u_{k-1}}$. Item k in the test is selected to be optimal at $\hat{\theta}_{u_{k-1}}$ among the items in the set $R_k \equiv \{1,\ldots, I\}\backslash S_{k-1}$.

A popular criterion of optimality in adaptive testing is maximization of information at the current ability estimate, that is, selection of item i_k according to objective function

$$\max_{i \in R_k} I_i(\hat{\theta}_{u_{k-1}}) \tag{8}$$

Alternative objective functions are based on Kullback-Leibler information or on Bayesian criteria that use the posterior distribution of θ after $k-1$ items. These functions are reviewed in van der Linden and Pashley (2000).

Several procedures have been suggested to realize content constraints on adaptive tests. The four major approaches are: (a) partitioning the bank according to the main item attributes and spiraling item selection

among the classes in the partition to realize a desired content distribution; (b) building deviations from content constraints into the objective function (Swanson and Stocking 1993); (c) testing from a pool with small sets of items build according to content specifications; and (d) using a shadow test approach in which prior to each item a full linear tests is assembled that contains all previous items, meets all content constrains, and is optimal at the ability estimate, and from which the most informative item is selected for administration (van der Linden 2000b.)

Adaptive testing is currently one of the dominant modes of computerized testing. Several aspects of computerized adaptive testing not addressed in this entry are reviewed in Wainer (1990).

See also: Classical (Psychometric) Test Theory; Content Validity; Dimensionality of Tests: Methodology; External Validity; Internal Validity; Psychometrics; Reliability: Measurement; Test Administration: Methodology; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Armstrong R D, Jones D H, Wang Z 1995 Network optimization in constrained standardized test construction. In: Lawrence K D (ed.) *Applications of Management Science: Network Optimization Applications*. JAI Press, Greenwich, CT, Vol. 8, pp. 189–212

Kingsbury G G, Weiss D J 1983 A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In: Weiss D J (ed.) *New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing*. Academic Press, New York, pp. 257–83

Luecht R M 1998 Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement* **22**: 224–36

Nemhauser G, Wolsey L 1988 *Integer and Combinatorial Optimization*. Wiley, New York

Reckase M D 1993 A procedure for decision making using tailored testing. In: Weiss D J (ed.) *New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing*. Academic Press, New York, pp. 237–55

Swanson L, Stocking M L 1993 A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement* **17**: 151–66

Timminga E, van der Linden W J, Schweizer D A 1996 *ConTEST 2.0: A Decision Support System for Item Banking and Optimal Test Assembly* [Computer Program and Manual]. iec ProGAMMA, Groningen, The Netherlands

van der Linden W J 1998 Optimal assembly of educational and psychological tests, with a bibliography. *Applied Psychological Measurement* **22**: 195–211

van der Linden W J 2000a Optimal assembling of tests with item sets. *Applied Psychological Measurement* **24**: 225–40

van der Linden W J 2000b Constrained adaptive testing with shadow tests. In: van der Linden W J, Glas C A W (eds.) *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers, Norwell, MA, pp. 27–52

van der Linden W J, Adema J J 1998 Simultaneous assembly of multiple test forms. *Journal of Educational Measurement* **35**: 185–98

van der Linden W J, Boekkooi-Timminga E 1989 A maximin model for test design with practical constraints. *Psychometrika* **54**: 237–47

van der Linden W J, Pashley P J 2000 Item selection and ability estimation in adaptive testing. In: van der Linden W J, Glas C A W (eds.) *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers, Norwell, MA, pp. 1–25

Wainer H (ed.) 1990 *Computerized Adaptive Testing: A Primer*. Erlbaum, Hillsdale, NJ

W. J. van der Linden

Computers and Society

Rapid and continuing advances in computer technology, and the related convergence of computing and telecommunications, has enabled the development of an array of information and communication technologies (ICTs), from the hand-held computer to the World Wide Web. The significance of these ICT innovations and their application across all sectors of society made the computer one of the defining technological transformations of the late twentieth century. Widespread fascination with personal computers (PCs), video games, digital TV, cell phones, wireless Internet, and a multitude of other ICTs often focuses on the technical ingenuity of their designs and their growing capabilities. But a narrow focus on technical advances overlooks the social impacts of ICTs as well as the ways in which social factors shape technical change as well as its societal implications.

When the social implications of computers are discussed they are most often viewed in utopian vs. dystopian perspectives. For example, many influential protagonists have argued that the Internet and the Web promote global community, electronic democracy (see *Electronic Democracy*) and an information economy, bringing fulfillment, empowerment, and jobs for the twenty-first century. Others fear the new networks and services on the Internet and the Web are isolating individuals, undermining democratic processes, and destroying jobs. Still others dismiss the importance of ICTs.

Such views of technology either as an 'unalloyed blessing,' an 'unmitigated curse,' or 'not worthy of special notice' oversimplify its role and fail to provide an understanding of the 'actual mechanism by which technology leads to social change' (Mesthene 2000). While a continuing debate between utopian and dystopian perspectives has illuminated the social and economic issues at stake, it has failed to inform policy and practice. It is, therefore, important to develop perspectives on the relationship between computers and society as a means to address broad social, economic, and political issues, and also guide social and technical choices about ICTs in a variety of contexts.

1. The Social and Technical Shaping of Access

All technologies are inherently social in that they entail physical artifacts or equipment, which are designed, produced, and used by people. Moreover, the design and use of technologies involves know-how, that is, expertise, which is in itself a social attribute (MacKenzie and Wajcman 1985, p. 3). Technologies are also social in that they define, but do not determine, how people do things, making some paths more economically, culturally, or socially rational than others. For example, computers and related ICTs define ways in which people can get information and communicate with one another. A student can get the news online via the World Wide Web or through a newspaper. ICTs not only reshape how people do things, they also shape and reshape what, when, and where things are done and what people know, who they know, and what they consume.

1.1 Dimensions of Tele-access

ICTs involve much more than just access to information or the technology of the computer, implied by conventional discussion of the information 'haves' and 'have-nots.' ICTs shape an individual's, household's, firm's, or nation's access to information, people, services, and technology. The concept of tele-access highlights how ICTs shape access—both electronically mediated and unmediated—to a wide array of social and economic resources.

Social and technical choices about ICTs can reconfigure electronic and physical access to four interrelated resources: information, people, services, and technology (Dutton 1999). The most commonly recognized is access to information. ICTs not only change the way people get information, but also alter the whole corpus of what a person knows and the information available to an individual at any given time and place. ICTs play a role in making some people information rich and others comparatively information poor. But access to information is only one set of relationships shaped by ICTs, and not necessarily the most socially significant.

ICTs also shape access to people. Choices about the design and use of ICTs not only change the ways individuals communicate with one another, but also influence whom individuals meet, talk to, stay in touch with, work with, and get to know. ICTs can connect or isolate people. For example, throughout the 1990s, the most common use of the Internet was for electronic mail (e-mail), that is, for gaining access to people, not for access to information *per se*.

Third, ICTs shape access to services. ICTs do more than simply change the way people consume information, products, and services. They also influence what products and services a person consumes and whom an individual purchases them from. ICTs can render obsolete a local business or an entire industry, but also create a new business or industry.

Finally, access to particular technologies—equipment, know-how, and techniques—shapes access to other technologies as ICTs interconnect and depend on one another in many ways. For instance, the Internet can provide access to vast numbers of computers around the world, yet a person needs a computer and other ICTs (such as a telephone line, a cable connection, or wireless device) to access the Internet.

There are many other ways in which ICTs can reduce, screen, reinforce, or alter tele-access, such as the content and flow of information, by accident or design. ICTs do not just provide access to more information or more people, many of whom a person would not be in touch with otherwise: they change patterns of interaction between people, information, communities, and organizations. As a substitute for face-to-face communication, for example, ICTs can provide benefits such as reducing travel, saving time, and extending the geography of human community. They may replace valuable human contact with a much less rewarding form of communication, fostering social isolation, or permit communication among people who might never have an opportunity to meet face to face. Tele-access encompasses all these substitutions, enhancements, and much more, by highlighting how people make social and technical choices about ICTs in ways that will reshuffle society, influencing who's in and who's left out.

1.2 Social Factors Shaping Choices and Their Implications for Access

ICT innovations and their social implications are not random or unstructured. For example, patterns of tele-access, such as the distribution of information haves and have-nots, are enabled and constrained by the social and economic contexts within which relevant actors at all levels make choices. Major sets of factors can be identified that facilitate or constrain the design and use of ICTs in ways that shape tele-access (Dutton 1999).

One set includes a variety of economic resources and constraints, such as the size, wealth, and vitality of nations, companies, and other actors. These place major constraints on the development and use of ICTs in all arenas of activity. A second set consists of ICT paradigms and practices. Ideas can become the foundation of powerful belief systems or 'paradigms' which create a way of interpreting reality that is very different to that perceived by people whose thinking is embedded in another paradigm. Concepts such as the 'information society' and the emergence of 'virtual universities,' for instance, have been important factors in shaping social and political choices in the design and implementation of ICTs, irrespective of their

descriptive validity. At the same time, experience and knowledge about ICTs can influence or even create a paradigm shift, which will have important implications for how people design and use ICTs within particular settings.

Another set of factors entails conceptions and responses of users of ICTs. Technologies are designed with a more or less well founded conception of the users—whether workers, consumers, managers, citizens, viewers, or audiences—who play an active role in shaping the implications of ICTs in ways that would not be expected by simply extrapolating from the perceived potential of the technology. Misconceptions of the user can undermine the diffusion of ICTs.

A set of factors related to the geography of space and place can also facilitate or constrain tele-access; for example, a person's location shapes access to ICTs. The use of computers and telecommunication technologies can depend on where ICTs are placed within a household or around the world. ICTs do not erase geography, as is often argued, but often make geography more important (Goddard and Richardson 1996). For instance, since ICTs can make location less significant for gaining electronic access, they enable individuals or firms to be more strategic in shaping physical access; that is, as ICTs enable more flexibility in location decisions, a firm can place a higher priority on other criteria, such as facilitating face-to-face communication with customers, or gaining access to skilled or low-cost labor, in choosing where to locate particular jobs and functions.

Finally, a wide variety of institutional arrangements and public policies, such as first amendment traditions within the USA, constrain choices about access. Technical, social, policy, and organizational innovation are interdependent (Freeman 1996). For example, the design of an organization influences the use of ICTs, but ICTs also create a variety of new options for radically redesigning organizations and interactions between organizations.

All these and other social factors shape and are shaped by tele-access in ways that can have both immediate and cumulative long-range consequences on most areas of modern life. However, a focus on the social and technical shaping of access is but one way to understand and study the social implications of ICTs. It is also useful to compare this perspective with a number of more conventional perspectives on the 'information revolution' and what it will mean for society, defined by the kinds of problems, assumptions, and frameworks guiding those who work within these traditions.

Much of the work on computers and society can be usefully classified under one of the following broad and overlapping labels, which capture the primary focus of each perspective. These include perspectives focused on impacts, influence, technology, strategy, and information (Dutton 1999). These theoretical foci have influenced policy and practice around the world by telling people what to think about and what to ignore, and can all be contrasted with a perspective focused on the shaping of tele-access.

2. *The Social 'Impacts' of ICTs*

The classic dystopian novel of the twentieth century, George Orwell's *1984*, pivoted around the emergence of two-way television—the telescreen—as an instrument of propaganda and electronic surveillance. Half a century later, the growing centrality of the Internet has generated updated Orwellian visions of multimedia surveillance.

These scenarios appear highly plausible because they are often based on the actual capabilities of ICTs. However, the actual impacts of ICTs are more difficult to forecast because they are shaped by social and political choices and not simply by features of the technology. The same can be said of more immediate concerns over the implications of ICTs on privacy, freedom of speech, and employment. Some of the earliest research on the social impact of ICTs was led by specialists involved directly in the development of computing and telecommunications who sensed that the public were failing to understand the full potential of the computer and the risks it posed to society. They formed interest groups within their respective professional associations and attempted to explain how the technology works and how it could affect individuals and society in terms of major issues, such as privacy and surveillance, which still remain central to policy debates.

2.1 *The Limits of Technological Determinism*

The social impact approach usually was based on rational forecasts of the social opportunities and risks created by particular features of the technology. For instance, the enormous storage capacity of the computer raised the potential for organizations to create huge databanks of information about individuals. However, empirical research usually found that rational expectations based on such capabilities of the technology were seldom realized (Kling 2000). For example, early study of the impact of computers on privacy found that computerized organizations did not change their record-keeping practices in the ways anticipated by those who were alarmed about databanks. Computers did indeed enable organizations to create huge databanks, but managers and professionals used the computer within prevailing data processing paradigms, leading them to employ this equipment to simply increase the efficiency of their existing practices. It has taken decades for new ICT paradigms to emerge, which would fundamentally alter the ways in which organizations obtain and keep

records in ways that changed the kinds of information they collect and created major new threats to privacy and surveillance.

Likewise, forecasts of the paperless office were based on technical capabilities and a rational expectation that people would want to save the time and costs associated with paper. However, these forecasts ignored the habits and values of readers and writers and the countervailing technical advantages of paper, such as the relative difficulty of reading an electronic screen. No one knows how long a transition from paper to electronic screen may take, but it will not be determined simply by advances in the technical features, such as screen resolution. It is clear that individuals and organizations have become dependent increasingly on electronic media, such as by replacing mail with e-mail. However, at the turn of the twentieth century, the production of paper has continued to grow with advances in ICTs, and is often supported by these advances, such as with online book stores spurring book sales.

This poor track record in forecasting the implications of technical change has led to fundamental criticisms of theories based on deterministic assumptions that particular real or potential technical features will have predictable 'impacts.' The public is often told that technology is on a predetermined trajectory, but different scientists forecast different trajectories. By contrast, much social research and analysis on the design and impacts of technologies has emphasized the decisive role played by social choices in determining ultimate paths of technological development as well as their social implications (Kling 2000).

Studies of the social bases of choices shaping technological design have, therefore, led social scientists to no longer look only at the social implications of technological change, but also consider the psychological, social, political, geographical, and economic factors that influence the design, production, and use of technology. This research on the 'social shaping' of technological change has been highly critical of the simplified, linear models of cause and effect used by technological determinists. This has led to a greater emphasis on exploring the underlying 'processes' of technical and social change, such as captured by the shaping of tele-access, rather than predicting their long-term impacts.

2.2 *Technology Matters in Shaping Access*

The study of technological change suggests, therefore, that tele-access—which is entailed in the key social impacts attributed to ICTs—should be viewed as the outcome of an indeterminate social and political process rather than being set on a predetermined technological trajectory. For example, the much promoted views that technological advances are democratizing access to ICTs or isolating individuals are based on deterministic assumptions that are empirically questionable. They are also dangerous because deterministic perspectives on the future can undermine the political will to make hard choices to ensure positive social and political outcomes.

The design and development of ICTs is not as deterministic of social outcomes as is often assumed by early theorists. Neither are they neutral. Advances in ICTs reshape access, but they do not determine access. Technology is like policy, because it tells us how we are supposed to do things, and makes some ways of doing things more rational and practical than another. Also, the biases designed into technological artifacts and systems can be even more enduring than legislation (Winner 1986, pp. 19–39), and create a momentum that is difficult to reverse (Hughes 1994). For instance, the rise of e-mail is biased toward the speed up of interpersonal communication. Individuals can choose to slow their communication down, but this often takes a conscious effort and strong political will within organizations and networks accustomed to Internet speed.

3. *The Study of Influence*

Another early perspective on computers and society, closely related to the study of impacts, evolved out of research on propaganda and the political implications of the mass media of newspapers, radio, television, and film in the aftermath of World War II. This media research tradition created a number of useful models of media effects (see *Media Effects*), including notions of the two-step flow (see *Communication, Two-step Flow of*) and agenda-setting (see *Agenda-setting*), which highlights the critical role of the media as a 'gatekeeper,' such as in shaping access to the news.

This media studies 'influence' tradition has extended into work on computers and society. The tradition focuses on the content of messages conveyed through the media and the messages' influence on those exposed directly or indirectly to them, such as in studies of the violent content of video games targeted at children. Those looking at ICTs within this influence tradition often ask whether the interactive character of emerging media, such as envisioned for digital and Web-TV, will make them more engaging and, therefore, more powerful in shaping attitudes, beliefs, and values. Others have focused on the way the profusion of media and channels can segment audiences in ways that might erode the quality and integrative effect of the mass media, and no longer provide the common experiences or shared text of a community.

Concern over audience segmentation provides an example of how technological change has challenged research in the influence tradition, which is built on a presumption of access to a mass audience. McLuhan anticipated this problem in the early 1960s when he argued that there had been too much emphasis on the content of the message rather than the more significant effect of the medium itself. In claiming 'it is the

medium that shapes and controls the scale and form of human association and action', McLuhan (1964, p. 9) indicated his belief that television's ability to reconfigure access to messages was more significant than whatever message was conveyed.

In this respect, McLuhan's perspective resonates more with a focus on tele-access than with those working within the influence tradition that shaped communication research over the decades. But technological change has challenged assumptions of communication research, which tended to take access to a mass audience for granted. Gaining access to audiences has become a central preoccupation of the communication industries, from popular music to broadcasting, cable and telephone industries, and a focus for research on the social implications of ICTs, including the media. The very idea of 'mass media' has been challenged as communication researchers saw technical change reconfiguring access to audiences. The influence perspective becomes even less applicable as one moves outside media studies, such as to assess the implications of ICTs for the economy.

4. *Studies of Technology and Society*

Outside the communications field, social scientists generally have looked at the computer and ICTs more generally as just special cases in a more general concern with the study of technology and society. Treatments of technology and society have flourished since the 1960s.

4.1 *The Role of Technology in Society*

Philosophers and social scientists have focused on how society is affected by the growing centrality of technology—and how technology, as a means, has become an end in itself. Researchers in this field have also analyzed the impacts of specific ICTs, ranging from the videophone to the Internet. Such technology assessments bring to the surface a wide range of multidisciplinary issues that go far beyond concerns with the mass media; they show, for example, how ICTs have become an integral component of a 'high-tech society' that pervades all aspects of our lives.

Another major contribution to this literature has focused on control as one impact of technological change. Many dystopian novels and films, such as *2001: A Space Odyssey* and *Jurassic Park*, capture the fear that technology is out of control and has created a new, unelected, high-tech élite which increasingly dominates decision making.

4.2 *Technology, Expertise, and Social Control*

Early discussion of computers and telecommunications reflected this concern that modern technology is becoming an unstoppable force which would support increasingly centralized control in organizations and society. Many observers of the social impacts of more recent advances in ICTs argue the opposite—that the convergence of computing and networking has decentralized control over the technology and placed more control in the hands of users (de Sola Pool 1983). Nevertheless, many others continue to argue that ICT innovations are empowering a new elite of experts, who have the skills and knowledge to control networks in organizations and society.

Technology concerns more than just equipment. It also encompasses the knowledge that is essential to its use (MacKenzie and Wajcman 1985, p. 3). The control of technology is, therefore, bound up with issues of who has access to the skills, equipment, and know-how essential to design, implement, and employ technology. Changes in technology can, therefore, restrict access to all these resources. For instance, early mainframe computers depended on programmers with knowledge of specialized programming languages. But technological change can equally expand access, for example, by the way simple graphical user interfaces with the Web have made access to information stored on millions of computers available to users round the world at the click of a mouse button. Likewise, social choices, such as the decision to learn a new human or computer language, affect access to technology, jobs, and people. The ways in which technical and social choices shape access to ICTs has been central to the technological perspective, but was often lost in a body of work covering myriad concerns at all levels of analysis and across all technology.

5. *ICT Strategies*

A focus on the goals and strategies of actors, instead of the capabilities of the technology, has been most fully developed within the management field, where some of the earliest empirical studies into the design and use of ICTs were conducted. Management researchers began to study computers in the 1950s. From early on, leading management science thinkers highlighted the significance of the convergence into 'Information Technology' (IT) of computing, telecommunications, and related management science techniques, such as modeling and computer simulation, but focused on how managers could employ IT (what later came to be called ICTs) to accomplish their objectives.

5.1 *Management Use of ICTs*

Through the early decades of its use, the computer was seen by management theorists as a strategic tool for enabling executives to boost centralized control by giving them access to information about the organization's resources and everyday operations. Centralization was not expected as an inevitable consequence

of computerization, but as an extension of and means for realizing the prevailing management paradigm of the time. It was a technical fix for problems that seemed immune to structural reforms, such as efforts to centralize functions or redraw organizational charts (Danziger et al. 1982).

From this strategic perspective, the major implications of ICTs extend from the strategies of management rather than the characteristics of the technology. This view remains central to contemporary management research, but with an ICT management paradigm that places less emphasis on the value of top management control and more on the virtues of innovation and networking (Castells 2000).

Instead of finding that computers inevitably increased centralization, early empirical research on computers discovered that those who control decision-making tend to adopt and use ICTs in ways that follow and reinforce existing patterns of control within the organization, whether they be highly centralized or decentralized (Danziger et al. 1982). This research highlighted the malleability of ICTs and the degree to which top managers are only one set of actors within a more complex and interdependent set of actors making decisions about the design and use of IT in organizations. The organizational and social implications of ICTs are the uncertain outcome of a struggle among actors over the design and use of these technologies. It is not taking place on a level playing field, since there are inequalities in existing institutions, cultures, and social and technical systems that favor some actors and choices over others, but neither does any single actor have a monopoly of control.

5.2 *The Interaction of Strategies Within a Broader Ecology of Games*

This struggle for control over ICT design and use generally takes place on a variety of different fields at the same time. All actors are not involved in the same struggle. Instead, individuals and groups are pursuing different goals within their own domains in an ecology of games (Dutton 1999, pp. 14–16). For instance, a technical expert might be pursuing a technically elegant solution to a network's design, while a top manager is primarily seeking cost reductions. This places major constraints on the predictability of outcomes based on an assessment of strategic goals, unless the varied goals of different actors involved in the broader ecology of games is well understood and orchestrated.

Global advances in ICTs, the momentum of technologies already in place, and the widespread application of ICTs throughout society place limits on the ability of any individual, household, organization, or nation to control the design and implementation of ICTs in predetermined ways. It is more realistic, for example, to view tele-access, and its consequences, as the outcome of a process of social and technical choices by many different actors within a variety of separate but inter-related technical, organizational, social, and policy arenas. It is the outcome of an ecology of games.

6. *The Information Society*

A very different and enduring perspective was introduced in the 1970s by discussion of the information society as a new stage in economic development. The information society theorists cite many of the social threats identified by others, such as to personal privacy and information inequality, but they focused more attention on the role of technology, and ICTs in particular, in the economy and society.

6.1 *The Information Age Thesis and its Critics*

The American sociologist Daniel Bell wrote a seminal work on the information society, which he first called the 'post-industrial society' (Bell 1999). He said information would be the key economic resource in this new era—not raw materials, or financial capital as in earlier agricultural and industrial societies. Citing the US economy as an exemplary case, Bell identified major trends in the development of an information society (see *Information Society*).

The most significant trend towards an information society was the shift in the majority of the labor force from agriculture (the primary sector) and manufacturing (the secondary sector) to services (the tertiary sector), largely through the growth of 'information work,' defined to include a broad array of jobs related to the creation, transmission, and processing of information, ranging from programmers and software engineers, to teachers and researchers. Another trend has been the increasing importance of knowledge and its codification to the management of social and economic institutions. A third set of information society trends involved power shifts, particularly the growing prominence of a knowledge elite, which forms a professional and managerial class who understand and know how to work with data, knowledge, information systems, simulation, and related analytical techniques.

Changes in work, technology, and power posited by this thesis have all been challenged (Freeman 1996). Yet conceptions of the information society have increasingly defined the public's understanding of social and economic change tied to the computer and related ICTs. Public and corporate policies and academic institutions and journals have been built on the concept of information as a new strategic resource. However, this focus can mislead policy and practice, particularly by the way it has substantially shifted

attention to the information sector and away from other parts of the economy.

More generally, the identification of information as the key central resource of the economy has also been widely questioned (Dutton 1999). For example, information has a highly variable currency: it is not always wanted or valued. As a resource, information can lose its value, such as in the case of a day-old newspaper, or retain its value after repeated use, as with a literary classic. If defined narrowly in the sense of being about 'facts,' the term 'information' becomes far too limiting as a depiction of the social role of ICTs. But if defined very broadly, for instance, as anything that 'reduces uncertainty,' then information seems to be so all encompassing that it becomes virtually meaningless.

Moreover, information can create, rather than reduce uncertainty. For example, those more informed about a topic are often less certain about its properties than many who are less informed. There is no linear relationship between information and certainty, but rather what MacKenzie (1999) has called a 'certainty trough.' And information is not a new resource. It has been important in every sector of the economy throughout history. Popular conceptions of the 'information society'—like the virtual society, or the cyberculture—capture the social significance of the ICT revolution (see *Information Society*). While these concepts emphasize the increasing centrality of ICTs to society, they fail to provide insights about the role of computers and related ICTs in social change. Moreover, if taken literally, many prevailing conceptions of the information society are misleading. They suggest that information is a new economic and social resource, when there is nothing new about the importance of information (Castells 2000). What is new is how you get access to information, but also to people, to services, and to technology itself (Dutton 1999).

6.2 *The Politics of Information: Communicative Power*

As people understand the limitations of seeing information as the pivotal concept in the study of ICTs, the central role of processes that shape tele-access has become more evident. The cliché that 'information is power' is seldom examined critically. It suggests that experts will gain power over politicians, managers, and the lay public. Yet this thesis ignores the degree to which economic resources enables access to knowledge—information, expertise, and the means for creating knowledge. It is not information *per se*, but the ability to control access to information that seems key. Social factors, such as institutional arrangements and public policy, or geographical proximity, as in gaining tacit knowledge from direct observation, can shape access, as well as technologies.

The role of ICTs in shaping access to information might be called 'information politics' (Danziger et al. 1982, pp. 133–5). This is also the concern of Garnham (1999), who has argued for a research agenda focused on study of the social processes involving producers, distributors, and users in negotiations over how information is created, disseminated, and consumed.

However, as explained above, information is only one element in what can be viewed as a far more general politics of access. ICTs also shape access to people, services, and other ICTs. So it is critical to define information politics broadly to include the choices shaping tele-access and the factors that constrain and facilitate them.

6.3 *Perspectives on the Information Age*

The idea of an information society entails a staged progression of an economy from an agricultural, to an industrial, to an information society (or 'information age') and to whatever comes next, such as a 'knowledge society.' An alternative forecast of social and economic evolution is posed by the notion of long economic development cycles caused by successive waves of technological revolutions (Freeman 1996). This approach argues that the invention of a new technology, like the steam engine or the computer, can have applications across the economy that affect many facets of our lives, for instance ICTs having effects well beyond what might be labeled as information work or information processing.

The concepts of staged and cyclical theories of economic and technological change are both attractive because they promise some level of predictability. However, there is no consensus on the identification and validation of these stages or cycles of economic development. Also, there is a long lag of from one to two decades between the invention of a radically new technology and its impact on the way things are done. It takes time for people to change their habits and beliefs and to accept a new paradigm for how they do things. This led to a new focus of attention on the role of ICT paradigms in shaping the impacts of technology in organizations and society (Freeman 1996).

Ideas like the information society, economic cycles and tele-access are important in part because they shape views about the way the world works and, thereby, influence the decisions of individuals, firms, and governments. That is one reason why alternative perspectives on the role of computers in society, such as the shaping of tele-access, are more than competing theories. They are also ideas that can shape decisions in everyday life as well as in once-in-a-lifetime choices.

See also: Computer Networking for Education; Computer-assisted Instruction; Electronic Democracy; Information Society; Information Technology; Media Effects; Political Communication; Techno-

logical Determinism; Technology and Social Control; Telecommunications and Information Policy; Telecommunications: Policy

Bibliography

Bell D 1973/1999 *The Coming of Post-industrial Society: A Venture in Social Forecasting*. Basic Books, New York

Castells M 2000 *The Rise of the Network Society: The Information Age: Economy, Society and Culture*, 2nd edn. Blackwell, Oxford, UK, Vol. 1

Danziger J N, Dutton W H, Kling R, Kraemer K L 1982 *Computers and Politics*. Columbia University Press, New York

Dutton W H 1999 *Society on the Line: Information Politics in the Digital Age*. Oxford University Press, Oxford, UK

Freeman C 1996 The two-edged nature of technical change: Employment and unemployment. In: Dutton W H (ed.) *Information and Communication Technologies—Visions and Realities*. Oxford University Press, Oxford, UK and New York, pp. 19–36

Garnham N 1999 Information politics: The study of communicative power. In: Dutton W H (ed.) *Society on the Line: Information Politics in the Digital Age*. Oxford University Press, Oxford, UK, pp. 77–8

Goddard J, Richardson R 1996 Why geography will still matter: What jobs go where? In: Dutton W H (ed.) *Information and Communication Technologies—Visions and Realities*. Oxford University Press, Oxford and New York, pp. 197–214

Hughes T P 1994 Technological momentum. In: Smith M R, Marx L (eds.) *Does Technology Drive History?* MIT Press, Cambridge, MA, pp. 101–13

Kling R 2000 Social informatics: A new perspective on social research about information and communication technologies. *Prometheus* **18**(3): 245–64

MacKenzie D 1999 The certainty trough. In: Dutton W H (ed.) *Society on the Line: Information Politics in the Digital Age*. Oxford University Press, Oxford, UK, pp. 43–6

MacKenzie D, Wajcman J (eds.) 1985 *The Social Shaping of Technology: How a Refrigerator Got Its Hum*. Open University Press, Milton Keynes, UK and Philadelphia, PA

McLuhan M 1964 *Understanding Media: The Extensions of Man*. Routledge, London

Mesthene E G 2000 The role of technology in society. In: Teich A H (ed.) *Technology and the Future*, 8th edn. Bedford St. Martin's, Boston pp. 61–70

de Sola Pool I 1983 *Technologies of Freedom*. The Belknap Press of Harvard University Press, Cambridge, MA

Winner L 1986 *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press, Chicago and London

W. H. Dutton

Comte, Auguste (1798–1857)

Auguste Comte is the founder of the 'philosophy of positivism,' the disciplines of 'sociology' and the 'history of science,' and the 'Religion of Humanity.' These innovations were inspired by his desire to reorganize society during the period of upheaval following the French Revolution of 1789. Living through several revolutions, monarchical regimes, republics, and empires, Comte rejected purely political and institutional reforms as ineffective. Instead, he argued that social harmony could only occur if individuals were brought together by a consensus that was both intellectual and emotional in nature. Ideas as well as feelings had to be reordered. Comte's search for a consensus derived not only from the anarchy of the post-revolutionary period but from the instability in his own life, which led him to view inner harmony as a sign of health. He created positivism and its keystone, sociology, to construct a new cohesive society that would encompass the conservatives' call for order and the leftists' preference for progress. This article will investigate Comte's personal and intellectual development as well as his contributions to knowledge and politics.

1. Life and Main Works

The son of a minor bureaucrat, Comte was born on January 19, 1798 in Montpellier, France. Growing up in this center of counter-revolutionary activity, he was deeply affected by the chaos stemming from the French Revolution. By age 14, he renounced both the monarchy and God, alienating his royalist, Catholic parents. Comte's rebelliousness was shared by many members of his generation, who longed for a sense of community and new principles of legitimacy that would stabilize moral and political life. His insubordinate behavior led to his expulsion from the *École Polytechnique*, the prestigious Parisian engineering school, which had admitted him in 1814. After his dismissal in 1816, he committed himself to fulfilling the radical revolutionaries' dream of erecting a secular republic.

Living in Paris, in 1817 he accepted the position of secretary to Henri de Saint-Simon. They became close collaborators in many journalistic enterprises. However, in 1824 they parted ways after Comte accused him of trying to take credit for his 'fundamental opuscule,' the *Plan des Travaux Scientifiques Nécessaires pour Réorganiser la Société*, which outlined his social ideas. Although Comte later denied Saint-Simon's influence, the older reformer showed him that they were in an era of transition that had to be completed by constructing a new society founded on industry and a new system of scientific knowledge—the 'positive philosophy.' This philosophy centered on the scientific study of society. Using his talent for synthesis and system building, Comte brought to fruition the seminal ideas strewn haphazardly throughout Saint-Simon's often-incoherent works.

In 1826 Comte launched a course in positive philosophy in his apartment. Yet he soon went mad and spent eight months in an asylum. He struggled

against manic depression and paranoia the rest of his life. After working as a mathematics tutor and journalist, he secured a lowly teaching and administrative position at the *École Polytechnique*, where he worked from 1832 to 1851.

Between 1830 and 1842 he wrote his most important work, the *Cours de Philosophie Positive*. It covered the history of the sciences, demonstrating that it was time for the study of society to become a science as well. The Cours attracted the attention of several famous writers, including John Stuart Mill, but they eventually became estranged from him.

Another person who experienced a bitter rupture with Comte was his wife, Caroline Massin. A penniless director of a reading room, she married him in 1824. They had a tumultuous marriage and separated in 1842. To punish her, he later called her a prostitute.

Comte fell in love in 1845 with Clotilde de Vaux, who was 17 years younger than he. Having been abandoned by her husband, this aspiring writer grew close to Comte but resisted his sexual advances. After she died of tuberculosis in 1846, Comte made her into his muse.

The Revolution of 1848 spurred Comte to activism. He established his own club, the 'Positivist Society,' to launch the positivist movement, and he wrote a manifesto, the *Discours sur l'ensemble du Positivisme*, which was a more extensive introduction to this work than his *Discours sur l'esprit Positif* of 1844. Disappointed that scientists and other intellectuals neglected him, he targeted two new groups, workers and women, whose social roles were highlighted in the *Système de Politique Positive*. Published in four volumes between 1851 and 1854, this work expressed his political and social ideas. It promoted the importance of the emotions, established the Religion of Humanity, and made morality the seventh science. Many scholars blame his relationship with de Vaux for Comte's alleged decline into religious, sentimental thinking. Yet there was no sudden break in his career, for from the start he distrusted the purely scientific spirit, committed himself to moral reform, and sought to create a new belief system to replace traditional religion.

The last years of Comte's life were marked by increasing seclusion and conservatism. Afraid of anarchy, Comte supported Louis Napoleon in 1851. Disappointed in workers' allegiance to socialism, he tried to gain the support of women, for whom he wrote the *Catéchisme Positiviste* (1852). His *Appel aux Conservateurs* (1855) aimed to persuade conservatives to join the positivist movement. Finally, in 1856 Comte published the *Synthèse subjective*, which explored the importance of images and the emotions in the context of mathematics and logic. Dependent financially on a small group of followers, he died in Paris on September 5, 1857.

Auguste Comte was a brilliant visionary. However, his paranoia, egoism, and dogmatism impaired on his personal relations and on his ability to launch a political and intellectual movement. In effect, he sought to create in society the harmony and stability that he could not find in his private life.

2. *Positivism and the History of Science*

Comte's original version of positivism, which is far from the scientism with which it is often associated, had both a theoretical component and a practical, or political, aspect. It did not celebrate the sciences for their own sakes or seek to be value-free as scientific systems claim to be. Instead, positivism promoted a new republic imbued with 'altruism'—a word he coined—in order to counteract the sense of disconnectedness coming from the French Revolution.

An idealist, Comte believed that intellectual confusion was at the root of the postrevolutionary anarchy because ideas always determined the social structure. Rejecting the conventional abstractions and practices of religion and politics as useless, he sought to create an intellectual revolution that would first lead to a moral revolution and then to a political and social revolution. An intellectual revolution would occur if one could place knowledge of society on a sure foundation to attain more control over it. To this end, he recommended the application of the scientific, or 'positive,' method to the study of politics and society.

The 'positive' method designated the certain, precise, useful, relative (as opposed to the absolute), and constructive (as opposed to the negative) method. Positive laws must avoid references to first and final causes or substances, which were incomprehensible. These laws had to show relationships between phenomena that would allow scientists to make predictions affecting the world of action. Indeed, each science should eschew pure empiricism and replicate as much as possible the deductive, rational approach of mathematics. Criticizing isolated facts as useless, Comte sought to free people from the tedious task of direct observation to enable them to construct abstract laws and theories. Individuals could not even make observations and retain facts without first formulating a theory to guide their perceptions. Provisional hypotheses were especially important to advance scientific understanding when concrete, observed facts were missing. Furthermore, scientists should use their imagination to create 'scientific fictions,' or hypothetical cases, to fill in poorly understood areas. Comte also recognized the role of esthetics and the subjective in formulating scientific hypotheses. To avoid giving reason excessive importance in scientific research, he refused to offer universal logical rules of scientific procedure and scientific proof. His only injunction was that every positive theory had to ultimately refer on some level to real, concrete phenomena that could be observed directly or indirectly.

Reflecting the impact of nineteenth-century historicism, Comte asserted, moreover, that one could not truly understand a phenomenon without grasping its history. To comprehend the positive method, one had to see its development in the major sciences, each of which had to be considered in relation to its own history, the history of the other sciences, and the history of society in general. The Cours highlighted the history of science, which Comte believed was the most neglected and important part of the development of humanity.

The Cours outlined Comte's famous 'classification of the sciences' and his equally renowned 'law of three stages.' This law demonstrated that as the mind went from one mode of thinking to another, it generated a different theoretical system, which in turn shaped the political and social system. There were three stages of development. In the theological stage, people used supernatural ideas, such as God, to link their observations. (There were three phases of this stage: fetishism, where gods were located in phenomena, polytheism, and monotheism.) In a theological society, priests and military men ruled. The theory of divine right was the reigning political doctrine, while conquest was the principal material activity. In the transitional, metaphysical stage, God was replaced by a personified essence or abstraction, like Nature or Reason. Lawyers and metaphysicians dominated society. The doctrines of popular sovereignty and natural rights were important in politics. More material activities were geared toward production. Finally, in the positive stage, supernatural and abstract forces would be replaced by scientific laws describing how, not why, phenomena functioned. Industrialists and positive philosophers would run the republic. Production would replace conquest as the goal of society.

According to the 'classification of the sciences,' each science went through the theological and metaphysical stages and finally became positive according to the simplicity of its phenomena and the distance of these phenomena from humans. Mathematics, which was not a 'direct' natural science, demonstrated the characteristics of the positive method. The positive method was then established in astronomy, physics, chemistry, and biology in that order. Each science depended on the simpler sciences that came before it in the hierarchy of sciences and could not advance to the positive stage until the preceding ones did. Comte argued against reducing one science to another, for each had its own distinctive subject matter and methodology. Moreover, Comte warned against the widespread use of mathematics in scientific investigation, especially in biology and sociology.

The point of Comte's laws of three stages and classification of the sciences was to demonstrate the inevitable triumph of the positive method in every area of study because all aspects of knowledge were interrelated and the mind sought homogeneity. Now that mathematics, astronomy, physics, chemistry, and biology were positive, it was time for the positive method to be extended to the study of politics and society, the last stronghold of theology and metaphysics.

3. *Sociology*

Comte named his science of society 'sociology' in 1839, giving it a distinctiveness and importance in his system that it never had in Saint-Simon's work. Once it was established, the positive philosophy, which encompassed all of the sciences, would be completed and unified. All our ideas would be scientific and homogenous. In addition, sociology would help unite knowledge by making the sciences focus on improving humanity. Because everyone would agree on the most significant intellectual and moral principles, the establishment of sociology would lead to the intellectual consensus required to build a stable industrial society. Sociology would also provide the rational principles necessary to reform society.

Comte's sociology had two parts, social statics and social dynamics. They stressed the interconnectedness of the members of the human species in space and time in order to counter the egoism and selfishness of the modern age. While social statics gave people a sense of solidarity with other members of contemporary society, social dynamics provided people with a feeling of connection with past and future generations of the human species. Social statics was the study of the social order. It explored forces, such as the family, that held society together. Social dynamics focused on historical progress. Its main principle was the 'law of three stages.'

4. *Positive Politics and the Religion of Humanity*

According to the 'law of three stages,' the unique characteristics of the human species—intelligence and sociability—became more dominant through exercise within both the individual and society. This theory reflects Comte's multi-dimensional view of human nature, which was reinforced by romanticism. To him, the emotions were closely connected with the intellect, which they simulated. Because social needs were both intellectual and emotional, social reorganization had to involve the heart as much as the mind. In the first phase of his development, Comte emphasized the importance of creating a system of ideas that could be shared by all people. In this positivist system, sociology led everyone to focus their thoughts on the needs of society, or humanity. In his second phase, Comte stressed the necessity of reinforcing the intellectual consensus by encouraging emotional ties among individuals through the Religion of Humanity. Everyone should not simply think about humanity but love it as well. In short, an increase in intellectual well roundedness would lead to a growth of altruism.

Comte's promotion of universal love was inspired by his horror of the abuses of modern industrial and capitalist society. He believed that the division of labor, which was necessary for progress, increased specialization and made people egotistic, narrow-minded, and selfish. The lower classes suffered the most from the indifference of the rest of society. Existing political parties, tied to the old theological and metaphysical principles, could not solve social problems.

Comte created a secular religious culture to bring individuals together. He designed new sacraments and a calendar of secular saints, called on the arts to aid in the process of continual socialization, and insisted that public schools offer boys and girls of all classes a broad education that cultivated feelings of sociability. The cult of Humanity, including a cult of 'Woman,' would honor individuals who had contributed to civilization. The 'Religion of Humanity' also assumed features of fetishism, the most emotionally intense and concrete form of religion. Everyone had to respect the earth, the 'Great Fetish.' Positivism, the most advanced stage of history, would lead full circle to the beginning of civilization, reflecting Comte's doubts about the benefits of science and progress.

Although Comte's doctrine was scientific in nature, he insisted on the limits and dangers of the sciences and recognized the importance of practical activity to which most people were drawn. To prevent the emergence of a dictatorship of scientists, Comte recommended that there be two separate powers that would hold each other in check. The temporal power would be composed of industrialists, who would control the practical, material activities of society. The spiritual power would consist of positive philosophers, who would govern the world of ideas, opinions, and feelings. The philosophers' chief assistants would be workers, who would be finally incorporated into society, and women, who would make nurturing the goal of public life.

5. *Assessment and Influence*

Comte's system was a grand synthesis designed to appeal to diverse philosophical and political positions. It incorporated ideas from many European thinkers, including Bacon, Descartes, Hume, Montesquieu, Condorcet, the Ideologues, the Scottish Enlightenment philosophers, Maistre, Gall, Kant, and Herder. In terms of politics, Comte supported the moral concerns associated with the counter-revolutionaries and remained committed to the leftist tradition that political authority must aid the people. In searching for a synthesis—a synthesis that was both objective and subjective in nature—Comte pursued the middle way between extremes in all realms. He sought to balance the needs of the heart and those of the mind, rationalism and empiricism, materialism and idealism, religion and science, feminism and patriarchalism, democracy and authoritarianism, and individuality and sociability.

Comte's philosophy is authoritarian, dogmatic, and grandiose. Many of his so-called scientific assertions, including the law of three stages, are replete with assumptions. Paradoxically, the very diffuseness of his endeavor caused him to lose support in his own day. The left found him illiberal and religious, while the right considered him blasphemous and anarchical.

Nevertheless, Comte's thought was extremely influential. Because it was vast and complex, people took from it what they wanted. Mill and Emile Littré made positivism into a manifesto for the scientific age. Politicians of the Third Republic used positivism in their campaign against the Church. Charles Maurras employed positivist principles to enhance the proto-fascist doctrine of the *Action Française*. Elsewhere, positivism attracted support from both liberals and conservatives in the UK, the USA, and Latin America. The Brazilian flag displays Comte's motto 'Order and Progress.'

Comte also influenced the humanities and social sciences. Philosophers of science have been interested in his approach to hypotheses. Logical positivists shared Comte's concern with problems of verification. Positivism long dominated European historiography. Lucien Lévy-Bruhl introduced many of Comte's ideas into anthropology. Through Emile Durkheim, positivism influenced sociology. Other social scientists affected by Comte include Alfred Espinas, Herbert Spencer, Hippolyte Taine, Ernest Renan, Charles Booth, and Patrick Geddes. Comte's thought remains relevant today, when people still grapple with many problems that he highlighted, such as the role of affectivity in education and the sciences, the roots of moral crises, and the significance of collective memory. Particularly at a time when global interests are paramount, Comte's call for a commitment to humanity as a whole and the environment bears renewed examination.

See also: Empiricism, History of; History of Science; Positivism, History of; Positivism: Sociological; Scientific Disciplines, History of; Science, Sociology of; Sociology, History of; Spencer, Herbert (1820–1903)

Bibliography

Comte A 1998 *Cours de Philosophie Positive*, 2 Vols. Hermann, Paris

Comte A 1929 *Système de Politique Positive*, 4 Vols. Société positiviste, Paris

Gouhier H 1997 *La Vie d'Auguste Comte*. 2nd rev. edn. Vrin, Paris

Grange J 1996 *La Philosophie d'Auguste Comte: Science, Politique, Religion*. Presses Universitaires de France, Paris

Kremer-Marietti A 1982 *Entre le Signe et l'histoire: L'Anthropologie d'Auguste Comte*. Klincksieck, Paris

Petit A (forthcoming) *Le Système positiviste: La Philosophie des sciences d'Auguste Comte et de ses Premiers Disciples (1820–1900)*. Vrin, Paris
Pickering M 1993 *Auguste Comte: An Intellectual Biography*. Vol. 1. Cambridge University Press, Cambridge, UK
Scharff R C 1995 *Comte after Positivism*. Cambridge University Press, Cambridge, UK
Wernick A 2001 *Auguste Comte and the Religion of Humanity*. Cambridge University Press, Cambridge, UK

M. Pickering

Concept Learning and Representation: Models

Concept learning describes the process by which experience allows us to partition objects in the world into classes for the purpose of generalization, discrimination, and inference. When children learn that certain objects in the world are apples and others oranges, their newly acquired knowledge allows them to make informed guesses about new objects they encounter ('this is an apple but that isn't … '), to tell members of these categories apart, and to make decisions which support various actions ('this is an apple so I can bite it').

Models of concept learning have adopted one of three contrasting views concerning category representation (see also *Knowledge Representation*). In prototype theories, the concept learning process is assumed to yield an abstract representation corresponding to the central tendency of the category exemplars on each of the dimensions of variation. In exemplar models, the concept is simply the set of mental representations of all of the category exemplars that have been previously observed, with each instance assumed to be stored as a separate trace. And in decision rule models, the learner is assumed to construct a boundary or rule in psychological space which partitions it into different category regions. These different perspectives, and some of the evidence supporting each of them, are considered in turn. Next, the role of selective attention in categorization, and the way in which the different models deal with selective attention, is discussed. Evidence that categorization may be controlled by multiple mechanisms is evaluated, and finally the fine-scale dynamics (i.e., time course) and some aspects of the neuropsychology of categorization are reviewed.

1. Prototype Theories

Prototype theories make a simple assumption about concept representation, namely that the concept is represented as the 'ideal' or 'average' category exemplar. Thus, representations of specific category exemplars do not influence classification. Numerous forms have been suggested for the (probabilistic) response rule; here we describe one common example (for other examples, see *Categorization and Similarity Models*). Assume that the probability that stimulus i is classified in category J, $P(\mathrm{R}_J/\mathrm{S}_i)$, is given by the equation:

$$P(\mathrm{R}_J/\mathrm{S}_i) = \frac{b_J s_{iPJ}}{\sum_K b_K s_{iPK}} \tag{1}$$

where b_J $(0 \leqslant b_J \leqslant 1,\ \Sigma b_J = 1)$ represents the bias towards making category response J and s_{iPJ} is the similarity between exemplar i and the prototype of category J. The idea is that classification depends on the similarity between exemplar i and the category J prototype relative to i's similarity to the prototypes of all other categories. Similarity is usually measured with multidimensional scaling methods (see also *Multidimensional Scaling in Psychology*; *Signal Detection Theory: Multidimensional*).

One piece of evidence favoring prototype models is that a previously unseen prototype may in some circumstances be classified with higher accuracy in the test phase than the actual training stimuli. Moreover, compared to the training stimuli, the prototype may be particularly resistant to forgetting (Homa et al. 1981). Despite this, and notwithstanding their attractive simplicity, there is an abundance of evidence against prototype models of concept learning, at least within the domain of perceptual classification. Prototype models have often been shown to provide poorer fits than other models to large sets of classification data, they make a number of predictions that have been falsified, and they fail to account for a number of well-established phenomena (see Nosofsky 1992 for a thorough review).

2. Exemplar Theories

Although it may seem implausible that classification is mediated by a 'chorus' of category exemplars, there is an enormous body of evidence consistent with the idea that a trace is created in memory for each category member that is observed and that the effects of individual exemplars on later classification are long lasting. For example, judging whether a letter string is a word or a nonword is influenced by a single prior encounter with that string, and this influence can persist for many days and can be specific to the exact surface form (e.g., font) of the string (Tenpenny 1995). The best-developed formal instance-based model of concept learning is Nosofsky's Generalized Context Model (GCM). In the GCM, $P(\mathrm{R}_J/\mathrm{S}_i)$, is given by the equation:

$$P(\mathrm{R}_J/\mathrm{S}_i) = \frac{b_J \sum_{j \in J} s_{ij}}{\sum_K (b_K \sum_{k \in K} s_{ik})} \tag{2}$$

where s_{ij} is the similarity between exemplars i and j and, as before, b_J represents the bias towards making category response J. The idea is that classification depends on the summed similarity between i and each of the members of category J, relative to i's similarity to the exemplars of all other categories. As in prototype models, similarity is usually measured using multidimensional scaling methods.

A sizable amount of evidence supports exemplar models such as the GCM (Nosofsky 1992). These models have been shown not only to provide excellent fits to parametric data sets but also to predict a number of qualitative phenomena. For example, they predict that a previously unseen prototype may in some circumstances be classified with higher accuracy in the test phase than the actual training stimuli, and that classification of test items may sometimes be almost completely unrelated to recognition of the same items.

There is also more direct evidence of preservation of the training exemplars in memory. Thomas (1998) first trained participants in a two-dimensional classification task in which values on the dimensions were highly correlated across exemplars. Next, participants were shown stimuli that possessed only one of the dimensions, together with the appropriate category label, and asked to predict the value of the stimulus on the other dimension. Consistent with exemplar theories, Thomas found that many participants were able to perform this task accurately, predicting values for the missing dimension consistent with the correlation observed across the exemplars during training. This implies that they retained quite detailed information about the conjunctions or correlations of feature values across exemplars. Many theories in which information about specific training exemplars is discarded (e.g., prototype theory) would be unable to account for this finding.

The major difficulties faced by exemplar models are that they do not fully reproduce individual differences in the distributions of responses across test stimuli (Nosofsky et al. 1994) and they do not account for certain base-rate effects in categorization (Nosofsky et al. 1992). We will also see below that some further data obtained by Thomas (1998) are very problematic for exemplar theories.

3. Decision-bound Models

In decision-bound or rule-based models, concepts are represented as regions in psychological space. A rule or boundary demarcates the dividing line between one category and another. In their most general form, decision-bound models can assume any mathematical form for these boundaries, but in practice most researchers have assumed that people can only learn 'simple' boundaries, and have therefore attempted to study the plausibility of special cases in which they are constrained to be linear or quadratic. Another reason for studying these particular forms is that the optimal boundary for discriminating between two bivariate normal distributions is either linear or quadratic, and it is reasonable to speculate that many natural categories are multivariate normal. The theory then assumes that, in a two-category situation, if a stimulus (or, more accurately, the percept of the stimulus) falls on one side of the boundary the participant assigns it to the corresponding category and if it falls on the other side it is assigned to the other category. Importantly, the distance of the stimulus from the boundary is assumed to be irrelevant as far as $P(\mathrm{R}_J/\mathrm{S}_i)$ is concerned. In that sense, categorization is assumed to be deterministic.

Decision-bound models have proved difficult to discriminate empirically from exemplar-based models such as the GCM—indeed, under some circumstances they are formally equivalent (Ashby and Maddox 1993). Decision-bound models are favored by data from non-normally distributed categories (Maddox and Ashby 1993) and by evidence that participants in some circumstances do not retain information about specific training exemplars. One compelling example was obtained by Thomas (1998) in the study described previously. Although many participants in Thomas' study were able to predict one dimension value on the basis of another in stimuli drawn from categories with correlated dimensional values, others were quite unable to do this and instead responded by essentially giving the mean value of the missing dimension, despite the fact that they had learned the initial classification. This strongly suggests that during the course of category learning they had extracted a decision bound and discarded information about the specific training exemplars.

Alternatively, there is rather little evidence, even in one-dimensional perceptual categorization tasks, that participants use all-or-none cutoffs as decision-bound theory predicts (Kalish and Kruschke 1997).

4. Selective Attention in Concept Learning

It has been recognized for many years that concept learning increases the amount of attention paid to stimulus dimensions that are relevant for solving a classification problem and decreases the amount paid to irrelevant ones. Goldstone (1994), for example, found that perceptual discrimination of highly confusable stimuli in a same–different judgment task increased as a result of category learning if the stimuli were assigned to different categories whereas, at least in some cases, it decreased if the stimuli were assigned to the same category.

In the GCM, this process of selective attention is accommodated by assuming that the similarity between a test item and category exemplars stored in memory is a function of their separation in psycho-

logical space, and that this space can be stretched or shrunk (thus altering the similarities) along orthogonal dimensions as a consequence of selective attention. A similar process can be incorporated into prototype theories. Nosofsky and his colleagues (e.g., McKinley and Nosofsky 1996) have now provided abundant evidence on the one hand that this process of selectively weighting the dimensions of stimulus variation in the generation of category decisions is essential in fitting categorization data within the exemplar-based framework.

On the other hand, decision-bound models have been paradoxically successful at fitting categorization data *without* needing to assume an explicit selective attention process. On this theory, the structure of the classification problem affects the placement of the decision boundary, but no additional notion of selective attention to stimulus dimensions is necessary. It is perhaps unclear how decision-bound theories would explain results such as those of Goldstone, which provide direct evidence for changes in the perceived similarities of stimuli as a consequence of category learning, but selective attention does not seem to be required for predicting categorization responses *per se*.

One possible resolution is based on the distinction between selective attention as a perceptual or as a decisional process. The former sees attention as directly altering the perceptual representation of a stimulus, while the latter assumes that attention operates later in the processing stream, perhaps weighting certain features of the stimulus more heavily than others in making a classification decision. Whereas Goldstone's data provide strong support for selective attention affecting the perceptual representation of stimuli, direct evidence for an influence on decisional processes—as assumed in the GCM—is currently lacking. Therefore there is scope for decision-bound models to dispense with selective attention as a decisional process so long as some dispensation is made for it to operate perceptually, for example via changes in the variances of stimulus percepts on the different dimensions. For further discussion, see Maddox and Ashby (1998).

5. *Hybrid Models*

An obvious possibility suggested by the evaluation of exemplar and decision-bound models is that people can use different categorization strategies in different circumstances, and indeed there is now a wealth of evidence in support of this conjecture. For example, Johnstone and Shanks (2001) asked participants to classify letter strings as either 'grammatical' or 'ungrammatical' after exposing them to strings, all of which were consistent with a complex rule. With incidental training conditions, classification was controlled purely by featural similarity of studied exemplars, but with explicit 'code-breaking' training conditions, classification of exactly the same items was controlled by the complex rule and was unrelated to similarity of studied exemplars. Erickson and Kruschke (1998) review the growing body of evidence for this 'multiple-mechanism' view of category learning and representation.

A natural question which arises is how one might computationally model a multiple-mechanism categorization system, and several attempts have recently been made. Erickson and Kruschke (1998), for example, had some success with a 'mixture-of-experts' model which incorporates both an exemplar and a decision-bound module. In this model, an adaptive learning rule (*Connectionist Models of Concept Learning*) determines the extent to which the final category decision is based on one module or the other, and the weight assigned to each module is an inverse function of the prediction error it generates for that particular classification problem. This sort of modular view of categorization is also supported by recent brain imaging data (Smith et al. 1998).

6. *Time Course of Classification*

An important recent development in the study of category learning and representation is the application of exemplar and decision-bound models to the fine-scale dynamics—specifically, the time course—of categorization. Until quite recently, research concentrated almost exclusively on categorization probabilities, but new methodological techniques provide researchers with powerful tools for studying more microscopic aspects of the categorization process.

Two rather different proposals have been advanced to extend exemplar models to deal with this issue (Lamberts, 2000). One approach is to assume that when an item is presented for classification, a race begins to retrieve exemplars from memory. The time to complete the race depends on how similar the stored exemplars are to the test item. When one race is completed, and an exemplar retrieved, a new race commences, and a classification response is only emitted when a random-walk process is driven to one or another decision barrier. The alternative approach is to assume that there is a gradual accumulation of perceptual information about the test stimulus and that similarity to stored exemplars varies during the construction of its representation. On this account, classification at any given moment is governed by Eqn. (2) but the summed similarity term in that equation is dependent on how much information about the test stimulus has been accumulated.

Proponents of decision-bound theory have advanced a rather straightforward hypothesis about the time course of categorization, namely that classification response times depend on distance to the decision bound, with greater distances leading to faster responses.

Attempts to distinguish between these accounts are in their infancy, but each appears to have strong merits. The articles cited in this section provide detailed reviews. Further study of the temporal dynamics of categorization is sure to be a major part of future research.

7. *Neuropsychology of Category Learning and Representation*

Neuropsychology raises two important issues concerning categorization (see also *Categorization and Similarity Models: Neuroscience Applications*). First, there is evidence that the mental representations controlling categorization are functionally and neuraly distinct from those accessed in standard 'explicit' tests of memory such as recall and recognition. If this is true, it suggests that models of memory are unlikely to provide any useful input to our understanding of categorization. The main piece of evidence is that amnesic patients suffering from anterograde amnesia seem to behave normally in category learning tasks while being profoundly impaired on tests of recall and recognition (e.g., Squire and Knowlton 1995). For instance, one densely amnesic patient (EP) performed normally in learning to classify dot patterns generated from a prototype, while performing at chance in a recognition test for those patterns. It has been concluded from this that the hippocampus, which controls explicit memory and which is damaged in amnesia, plays no role in category learning.

Recently, however, the conclusion that categorization and memory are functionally and neuraly unconnected has been questioned on the basis of demonstrations that exemplar models such as the GCM can, just like EP, show dissociations between classification and recognition. In Nosofsky and Zaki's (1998) modeling approach, only a single parameter, the sensitivity parameter, was varied in order to account for a general learning difference between amnesics and controls. This parameter represents the ability to discriminate among distinct exemplars stored in memory, an ability which is plausibly reduced in amnesics as a result of their deteriorated memory capacity. Nosofsky and Zaki showed that decreasing the value of this parameter reduced classification performance only slightly, but reduced recognition performance considerably. Thus, the GCM successfully accounts for the dissociation between classification and recognition observed by Squire and Knowlton (1995).

Second, there is now a wealth of evidence concerning the brain localization of category knowledge. Brain imaging studies have suggested that knowledge is organized structurally at the neural level. Martin et al. (1996) studied neural organization of conceptual knowledge by presenting normal participants with pictures of animals and tools to name while brain activity was recorded by positron emission tomography (PET). As a control, participants saw nonsense objects. Compared to the control condition, brain activity for the meaningful objects was increased in the ventral temporal lobes bilaterally and in Broca's area in the left hemisphere. The latter is probably related to word generation. Relative to pictures of tools, pictures of animals produced increased activation in the left medial occipital lobe. Relative to pictures of animals, pictures of tools produced increased activation in the left dorsal temporal lobe and the left premotor region. This premotor area is also activated by imagined hand movements, suggesting that part of the conceptual representation of tools is a specification of the hand movements they afford. Overall, Martin *et al.*'s results confirm that gross category distinctions are present in the way the brain organizes conceptual knowledge.

For a fuller review of recent psychological and neuroscience research on models of concept learning and representation, see Lamberts and Shanks (1997).

See also: Categorization and Similarity Models; Connectionist Models of Concept Learning; Early Concept Learning in Children; Knowledge Representation; Representation: History of the Problem

Bibliography

Ashby F G, Maddox W T 1993 Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology* **37**: 372–400

Erickson M A, Kruschke J K 1998 Rules and exemplars in category learning. *Journal of Experimental Psychology: General* **127**: 107–40

Goldstone R 1994 Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General* **123**: 178–200

Homa D, Sterling S, Trepel L 1981 Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory* **7**: 418–39

Johnstone T, Shanks D R 2001 Abstractionist and processing accounts of implicit learning. *Cognitive Psychology* **42**

Kalish M L, Kruschke J K 1997 Decision boundaries in one-dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **23**: 1362–77

Lamberts K 2000 Information-accumulation theory of speeded categorization. *Psychological Review* **107**: 227–60

Lamberts K, Shanks D (eds.) 1997 *Knowledge, Concepts and Categories*. Psychology Press, Hove, UK

Maddox W T, Ashby F G 1993 Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics* **53**: 49–70

Maddox W T, Ashby F G 1998 Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky 1996 *Journal of Experimental Psychology: Human Perception and Performance* **24**: 301–21

Martin A, Wiggs C L, Ungerleider L G, Haxby J V 1996 Neural correlates of category-specific knowledge. *Nature* **379**: 649–52

McKinley S C, Nosofsky R M 1996 Selective attention and the formation of linear decision boundaries. *Journal of Exper-*

imental Psychology: Human Perception and Performance **22**: 294–317
Nosofsky R M 1992 Exemplar-based approach to relating categorization, identification, and recognition. In: Ashby F G (ed.) *Multidimensional Models of Perception and Cognition*. Erlbaum, Hillsdale, NJ, pp. 363–93
Nosofsky R M, Kruschke J K, McKinley S C 1992 Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**: 211–33
Nosofsky R M, Palmeri T J, McKinley S C 1994 Rule-plus-exception model of classification learning. *Psychological Review* **101**: 53–79
Nosofsky R M, Zaki S R 1998 Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science* **9**: 247–55
Smith E E, Patalano A L, Jonides J 1998 Alternative strategies of categorization. *Cognition* **65**: 167–96
Squire L R, Knowlton B J 1995 Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences* **92**: 12470–4
Tenpenny P L 1995 Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin and Review* **2**: 339–63
Thomas R D 1998 Learning correlations in categorization tasks using large, ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**: 119–43

D. R. Shanks

Conceptual Blending

Conceptual blending is a basic mental operation that leads to new meaning, global insight, and conceptual compressions useful for memory and manipulation of otherwise diffuse ranges of meaning. It plays a fundamental role in the construction of meaning in everyday life, in the arts and sciences, and especially in the social and behavioral sciences. The essence of the operation is to construct a partial match between two inputs, to project selectively from those inputs into a novel 'blended' mental space, which then dynamically develops emergent structure. It has been suggested that the capacity for complex conceptual blending ('double-scope' integration) is the crucial capacity needed for thought and language.

1. Some Simple Examples

1.1 Ritual of the Newborn Baby

In a European ritual, the newborn baby is carried up the stairs of the parents' house as part of a public event. The ritual is meant, symbolically, to promote the child's chances of rising in life. One input is the ordinary action of carrying a baby up the stairs. The other input is the schematic space of life, already structured so that living a life is metaphorically moving along a path, such that good fortune is up and misfortune is down. In a partial match between these inputs, the path up the stairs corresponds to the course of life, the baby is the person who will live this life, the manner of motion up the stairs corresponds to how the person 'goes through' life, and so on. In the symbolic ritual, the two inputs are blended, so that the ascent of the stairs is the course of life; an easy ascent is an easy rise in life for the person that the baby will become; stumbling or falling might take on extraordinary significance.

1.2 Boat Race

A famous example of blending is 'the boat race' or 'regatta.' A modern catamaran is sailing from San Francisco to Boston in 1993, trying to go faster than a clipper that sailed the same course in 1853. A sailing magazine reports:

> As we went to press, Rich Wilson and Bill Biewenga were barely maintaining a 4.5 day lead over the ghost of the clipper *Northern Light*, whose record run from San Francisco to Boston they're trying to beat. In 1853, the clipper made the passage in 76 days, 8 hours (Great America II, 1993 *Latitude 38* **190**: 100).

Informally, there are two distinct events in this story, the run by the clipper in 1853 and the run by the catamaran in 1993 on (approximately) the same course. In the magazine quote, the two runs are merged into a single event, a race between the catamaran and the clipper's 'ghost.' The two distinct events correspond to two input mental spaces, which reflect salient aspects of each event: the voyage, the departure and arrival points, the period and time of travel, the boat, its positions at various times. The two events share a more schematic frame of sailing from San Francisco to Boston; this is a 'generic' space, which connects them. Blending consists of partially matching the two inputs and projecting selectively from these two input spaces into a fourth mental space, the blended space (see Figure 1).

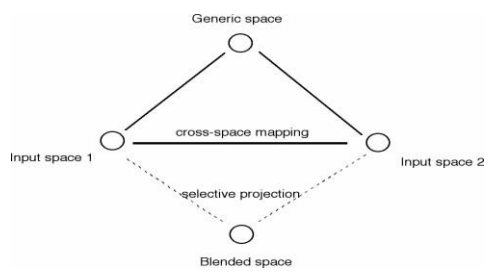

Figure 1
The blended space

In the blended space, we have two boats on the same course, that left the starting point, San Francisco, on the same day. Pattern completion allows us to construe this situation as a race (by importing the familiar background frame of racing and the emotions that go with it). This construal is emergent in the blend. The motion of the boats is structurally constrained by the mappings. Language signals the blend explicitly in this case by using the expression 'ghost-ship.' By 'running the blend' imaginatively and dynamically—by unfolding the race through time—we have the relative positions of the boats and their dynamics.

Crucially, the blended space remains connected to the inputs by the mappings, so that real inferences can be computed in the inputs from the imaginary situation in the blended space. For example, we can deduce that the catamaran is going faster overall in 1993 than the clipper did in 1853, and more precisely, we have some idea ('four and a half days') of their relative performances. We can also interpret the emotions of the catamaran crew in terms of the familiar emotions linked to the frame of racing.

The 'boat race' example is a simple case of blending. Two inputs share structure. They get linked by a cross-space mapping and projected selectively to a blended space. The projection allows an emergent structure to develop on the basis of composition, pattern completion (based on background models), and elaboration ('running the blend').

1.3 Clinton and Roosevelt

The type of conceptual blend just discussed (technically called a 'mirror network' because the same frame organizes both inputs) is very general. For example, a political comment on Clinton's presidency after he had been in office a few months might have been: *by this point*, *Roosevelt was far ahead of Clinton*. The two inputs are Roosevelt's and Clinton's presidencies. They are mapped onto each other in a natural way: starting points, mid-points, and so on are matched. In the blend, Roosevelt and Clinton are brought together within the same timeframe, so that they are competing against each other. Blends of this sort are routinely elaborated for reasoning purposes in political analysis.

1.4 Fictive Motion

Languages have means of describing static scenes in terms of fictive motion: *The fence runs all the way down to the river*. This works by having an 'imaginary' trajector move along the relevant dimension of an object, in this case the fence, or along some 'imaginary' path linking two objects. This is a remarkable mode of expression: it conveys motion and immobility at the same time. Objective immobility is expressed along with perceptual or conceptual motion. This apparent contradiction is a consequence of conceptual blending, which allows several connected but heterogeneous mental spaces to be maintained simultaneously within a single mental construction. An input space containing a static scene of a fence and a river is blended with an input space that contributes a moving trajector on a path with a reference point.

1.5 Counterfactuals

Human thought depends heavily on the capacity for counterfactual thought, and counterfactuals are complex blends. Most of us can effortlessly understand statements like 'In France, Watergate would not have hurt Nixon.' This counterfactual is intended to highlight some differences between the American and French cultural and political systems. It is a blend that brings in aspects of the French system from one input and the Watergate scandal and President Nixon from the other. In the blend, we have a Watergate-like situation in France. Running this blend delivers attitudes quite different from those in the American input, and so in the blend, the president is not harmed. Counterfactuals can blend frames and identities in powerful ways (*If I were you*, …). Such blends have been shown to play a major role for reasoning in the natural and social sciences.

1.6 Computer Interfaces

A nice example of conceptual blending in action and design is the 'desktop' interface, in which the computer user moves icons around on a simulated desktop, gives alphanumeric commands, and makes selections by pointing at options on menus. Users recruit from their knowledge of office work, interpersonal commands, pointing, and choosing from lists. All of these are 'inputs' to the imaginative invention of a blended scenario that serves as the basis for integrated performance. Once this blend is achieved, it delivers an amazing number of multiple bindings across quite different elements, bindings that seem, in retrospect, entirely obvious. A configuration of continuous pixels on the screen is bound to the concept 'folder,' no matter where that configuration occurs on the screen. Folders have identities, which are preserved. The label at the bottom of the folder in one view of the desktop corresponds to a set of words in a menu in another view. Pushing a button twice corresponds to opening. Pushing a button once when an arrow on the screen is superimposed on a folder corresponds to 'lifting into view.' Of course, in the technological device that makes the blend possible, namely the computer interface, there is no ordinary lifting, moving, or opening happening at all, only variations in the illumination of a finite and arranged number of pixels on the screen. The blend is not the screen: the blend is an imaginative

mental creation that lets us use the computer hardware and software effectively. In the blend, there is lifting, moving, opening, and so on happening, but imported not from the technological device at hand, which is only a medium, but from another input, namely our mental conception of work we do on a real desktop.

2. The Network Model

Conceptual blending is described and studied scientifically in terms of integration networks. In its most basic form, a conceptual integration network consists of four connected mental spaces: two partially matched input spaces, a generic space constituted by structure common to the inputs, and the blended space. The blended space is constructed through selective projection from the inputs, pattern completion, and dynamic elaboration. The blend has emergent dynamics. It can be 'run,' while its connections to the other spaces remain in place. Neurobiologically, it has been suggested that elements in mental spaces correspond to activated neural assemblies and that linking between elements corresponds to neurobiological binding (e.g. co-activation). On this view, mental spaces are built up, interconnected, and blended in working memory by activating structures available from long-term memory. Mental spaces can be modified dynamically as thought and discourse unfold.

Four main types of integration networks have been distinguished: Simplex, Mirror, Single-Scope, Double-Scope. In Simplexes, one input consists of a frame and the other consists of specific elements. A frame is a conventional and schematic organization of knowledge such as 'buying gasoline.' In Mirrors, a common organizing frame is shared by all spaces in the network. In Single-Scopes, the organizing frames of the inputs are different, and the blend inherits only one of those frames. In Double-Scopes, essential frame and identity properties are brought in from both inputs. Double-Scope Blending can resolve clashes between inputs that differ fundamentally in content and topology. This is a powerful source of human creativity. The main types of networks just mentioned are actually prototypes along a continuum that anchors our intuitive everyday notions about meaning to a unified understanding of the unconscious processes at work. Varieties of meaning traditionally considered unequal or even incommensurable—categorizations, analogies, counterfactuals, metaphors, rituals, logical framing, grammatical constructions—can all be situated on this continuum. Conceptual blending has been shown to operate in the same way at the highest levels of scientific, artistic, and literary thought, and at the supposedly lower levels of elementary understanding and sentence meaning. Elaborate blending is at work in superficially very simple expressions like 'safe gun' vs. 'safe child,' 'guilty pleasures,' 'caffeine headache,' or 'money problem.' Emergent structure arises routinely, as in 'This surgeon is a lumberjack', which suggests that the surgeon is incompetent, although *incompetence* is neither a feature of *surgeon* nor of *lumberjack*.

There are opposing pressures within an integration network to maximize topology matching, integration, unpacking of the blend, web connections, compression, and intentionality. More complex integration networks ('multiple blends') allow multiple input spaces and successive blending in which blends at one level can be inputs at another.

3. Compression

Blending is a remarkable tool of compression over vital relations like time, space, cause-effect, identity, and change. In the Newborn Ritual, time is compressed: an entire lifetime becomes, in the blend, the short time it takes to climb the stairs; in the Nixon counterfactual, space is compressed (Watergate can be in France); in the desktop interface, the complex sequence of events which move the mouse horizontally and cause an apparent vertical motion of the arrow (and other 'objects') on the screen is compressed and integrated into a single action, moving the arrow. This is a compression of space, cause-effect, and change.

4. Range of the Cognitive Operation

A systematic study of conceptual blending was initiated in 1993 by Gilles Fauconnier and Mark Turner (1998), who discovered the structural uniformity and wide application of the notion. Since then, important work has been done on the theory of conceptual blending (Coulson 2001), and its empirical manifestations in mathematics (Robert 1988), social science (Sweetser in press, Turner 1996a, 1996b), literature (Freeman 1997, Oakley 1998, Turner 2000), linguistics (Lidell 1998, Mandelblit 1997), and music (Zbikowski 1999). There have been proposals for the mathematical and computational modeling of the operation (Goguen 1999, Veale 1999), and experimental research within neuroscience on the corresponding neural and cognitive processes (Coulson 2001, Grush and Mandelblit 1997). Additional information and an extensive bibliography can be found on website (http://www.wam.umd.edu/~mturn/WWW/blending.html).

See also: Computational Imagery; Conceptual Change, Cognitive Psychology of; Counterfactual Reasoning, Qualitative: Philosophical Aspects; Imagery versus Propositional Reasoning; Mental Imagery, Psychology of; Metaphor and its Role in Social Thought: History of the Concept; Visual Imagery, Neural Basis of

Bibliography

Coulson S 2001 *Semantic Leaps: Frame-shifting and Conceptual Blending in Meaning Construction*. Cambridge University Press, New York and Cambridge

Fauconnier G 1997 *Mappings in Thought and Language*. Cambridge University Press, New York and Cambridge

Fauconnier G, Turner M 1998 Conceptual integration networks. *Cognitive Science* **22**(2): 133–87

Freeman M 1997 Grounded spaces: Deictic—self anaphors in the poetry of Emily Dickinson. *Language and Literature* **6**(1): 7–28

Goguen J 1999 An introduction to algebraic semiotics, with application to user interface design. In: Nehaniv C (ed.) *Computation for Metaphor, Analogy, and Agents*, series in Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, pp. 242–91

Grush R, Mandelblit N 1997 Blending in language, conceptual structure, and the cerebral cortex. The Roman Jakobson Centennial Symposium. *International Journal of Linguistics Acta Linguistica Hafniensia Volume* **29**: 221–37

Lidell S K 1998 Grounded blends, gestures, and conceptual shifts. *Cognitive Linguistics* **9.3**: 283–314

Mandelblit N 1997 Ph.D. dissertation, University of California, San Diego

Oakley T 1998 Conceptual blending, narrative discourse, and rhetoric. *Cognitive Linguistics* **9**: 321–60

Robert A 1998 Blending in the interpretation of mathematical proofs. In: Koenig J-P (ed.) *Discourse and Cognition*. Center for the Study of Language and Information (CSLI), Stanford, CA [distributed by Cambridge University Press]

Sweetser E in press Blended spaces and performativity. *Cognitive Linguistics*

Turner M 1996a Conceptual blending and counterfactual argument in the social and behavioral sciences. In: Tetlock T, Belkin A (eds.) *Counterfactual Thought Experiments in World Politics*. Princeton University Press, Princeton, NJ

Turner M 1996b *The Literary Mind*. Oxford University Press, Oxford, UK

Turner M 2000 Backstage cognition in reason and choice. In: Lupia A, McCubbins M, Popkin P (eds.) *Elements of Reason: The Science of the Mind and the Limits of Political Rationality*. Cambridge University Press, Cambridge, UK

Veale T 1999 Pragmatic forces in metaphor use: The mechanics of blend recruitment in visual metaphors. In: Nehaniv C (ed.) *Computation for Metaphor, Analogy, and Agents*, in series Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, pp. 37–51

Website for Conceptual Blending: http://www.wam.umd.edu/~mturn/WWW/blending.html

Zbikowski L 1999 The blossoms of 'Trockne Blumen': Music and text in the early nineteenth century. *Music Analysis* **18** (3): 307–45.

G. Fauconnier

Conceptual Change, Cognitive Psychology of

Conceptual change was one of the most popular notions in the areas of cognitive development and science education in the 1980s and 1990s. This notion, inspired by ideas proposed by philosophers of science such as T. Kuhn (1970), was originally introduced to denote a fundamental change in a concept, conception, or conceptualization with an increase in age, expertise, or instruction.

1. *Inevitability of Conceptual Change*

The key idea of conceptual change presupposes the domain-specificity of cognitive growth, instead of focusing on general logical structures across domains, as Piaget's developmental view assumed. It also presumes, unlike Piaget's view, that even young children have coherent bodies of knowledge about important aspects of the world. Proponents of conceptual change claim that such bodies of knowledge often constitute naive 'theories' containing ontological distinctions, coherent organizations among pieces of knowledge, and causal devices or explanations as essential components (Wellman 1990). They assume that humans have a basic tendency to construct a model or theory to make sense of an observed set of data. Some innate principles constrain this process of construction in a few selected domains that are critical for survival. However, because construction of the initial theory is based on a limited database, it has to be revised as more and more facts are incorporated into the body of knowledge, unless the initial set of observed facts constitutes a representative sample of all relevant facts. Some of the innate principles may be weakened or given up, as accumulated pieces of prior knowledge come to serve as constraints, which also makes conceptual change during childhood inevitable.

2. *Conceptual Change as Fundamental Restructuring*

The notion of conceptual change in cognitive development has been proposed as an alternative to 'enrichment views' (Carey 1985); it denotes that conceptual development involves not just enrichment or elaboration of the existing knowledge systems but their considerable reorganization or restructuring. Conceptual change involves change in core concepts, conceptions, or conceptualizations. Knowledge systems before and after the conceptual change may sometimes be locally incommensurable; that is, some pieces of knowledge in one system cannot properly be translated into the other, as exemplified by the shift from children's undifferentiated concept of heat/temperature to adults' separate concepts of heat and temperature, or the shift from the simple disjunction between animals and plants to a single, integrated concept of living things.

Conceptual change often takes the form of theory change, because concepts are considered to be embedded in theories; changing one core concept in a

theory generates changes in related concepts and eventually leads to a change in the whole set of concepts. Theory change involves changes in causal devices or explanations and/or changes in the phenomena or entities that are included. Four types of theory change can be distinguished with regard to the relationship between an old theory (as the prechange system) and a new theory (as the postchange system). First, a new theory emerges from an old theory with the latter being subsumed in, or replaced by, the former. For example, between ages 2 and 4–5, the early theory of mind, which is based solely on desire and perception, is transformed into the 'representational theory of mind', which includes beliefs as well (Gopnik and Wellman 1994).

Second, a new theory emerges and develops from an old one, and the latter continues to exist, although it becomes less salient. Sometimes the old theory is even extended by the new theory. For example, Perner (1991) claims that, although at about 4 years of age children's understanding of the mind changes from a 'situation theory' (where mental states are construed in relation to situations) to a 'representational theory' (where mental states are understood as serving representational functions), the latter does not replace the former but merely extends it; even adults may be situation theorists when possible, but they, unlike young children, can take a representational view when necessary. Another example is the conceptual change in naive biology that takes place approximately between ages 6 and 10 or older. Young children tend to understand biological phenomena by relying on vitalistic causality, whereas older children and adults use mechanistic causality. However, vitalistic causality continues to function as a basis of understanding and to be used in situations where people do not think they are required to reason using so-called scientific biology (Hatano and Inagaki 1996).

Third, a new theory emerges from an old one through differentiation, and new and old theories develop separately afterwards. For example, when considering biological phenomena, very young children often apply a theory in which psychology and biology are undifferentiated, whereas older children rely on a purely biological theory (Carey 1985). Another example is the emergence of a theory of matter from a theory of physics in which objects and the materials from which they are made are not fully distinguished (Smith et al. 1985).

Fourth, a new theory emerges through the integration of old subtheories. For example, young Israeli children consider plants neither living things nor nonliving things, but 'growers' (Stavy and Wax 1989). In contrast, they easily recognize animals as living by attending to their self-initiated movement. In other words, young children seem to possess different theories for animals and plants. As they grow older, they acquire a theory of living things by integrating these different subtheories of animals and plants.

It should be noted that conceptual change seldom occurs suddenly, just as it has taken years for concepts, conceptions, or conceptualizations to change in the history of science. The process of conceptual change tends to be slow and gradual, even if its end result is drastically different from its initial state.

3. *Spontaneous vs. Instructionally-based Conceptual Change*

Forms of conceptual change can also be distinguished in terms of whether the change occurs spontaneously or is induced by instruction (Hatano and Inagaki 1996, Vosniadou and Ioannides 1998). Spontaneous conceptual change is the change that results from children's enriched experience in their physical and sociocultural environment. In other words, it occurs without systematic instruction, though schooling certainly has some general facilitative effects on it. This form of change seems to occur readily, because it is usually found among most children growing up in highly technological societies. An example is the change that occurs in young children's naive biology, i.e., the change from a biology that is personifying and vitalistic in nature to one relying on inference based on complex, hierarchically organized biological categories and mechanistic causality (Hatano and Inagaki 1997). Another example, drawn from physics (Vosniadou and Ioannides 1998), is the change from the internal force model (that there is an internal force within stationary and moving heavy objects) to the acquired force model (that there is an acquired force within moving objects only).

Instructionally based conceptual change occurs as a product of systematic science education, more specifically, by incorporating scientific concepts and thereby correcting 'misconceptions' (alternative conceptions based on limited databases). The acquisition of synthetic mental models of the earth (Vosniadou and Brewer 1992) or of scientific biology having conceptual devices such as 'photosynthesis' or 'evolution' (Hatano and Inagaki 1996) is a good example of instructionally based forms of conceptual change. This form of conceptual change requires systematic teaching, and even with good teaching, only a limited portion of older children and adults may achieve it. Many studies in science education have reported that instructionally based conceptual change is difficult to induce (e.g., Clement 1982).

Cognitive developmentalists have been primarily concerned with spontaneous conceptual change, and have focused on specifying the initial state (initial theory) and just sketching the consequences of changes, rather than specifying the exact mechanisms of its change. Science educators, in contrast, have primarily been interested in instructionally based conceptual change, attempting to achieve the goal state (an understanding of scientific theory) as an

instructional aim. Both research groups have been working in parallel until recently. However, developmentalists have now become interested in instructionally based change as an aid for specifying the possible mechanisms of spontaneous conceptual change. Meanwhile, science educators are trying to incorporate findings on spontaneous conceptual change into the design of their instruction.

4. Mechanisms of Conceptual Change

We have three sets of data on how conceptual change occurs: from the history of science, conceptual development, and science education. Although these different sets of data differ markedly in terms of time span, the complexity of target entities and phenomena, the degree of intellectual sophistication of the individuals involved, and many other variables, they all illuminate some mechanisms of conceptual change.

It is clear that in every case increased knowledge is required for conceptual change, and that the prechange system serves as a cognitive constraint in conceptual change. A general mechanism for change is the spreading of the truth-value alteration. When new inputs change the truth-value of some pieces of knowledge, the changes bring about changes in the truth-value of other connected pieces, which may induce further changes in their neighbors. Thus, there can be a drastic change in almost all pieces through continued spreading and recurring effects. Those individuals who monitor the truth-value of pieces of knowledge constituting a theory (e.g., scientists) may be sensitive to and respond to a small amount of disconfirmation or discoordination. A good example is the 'thought experiment' including limiting case analyses, as indicated by Galileo's famous thought experiment showing that heavier objects do not fall faster than lighter ones (Nersessian 1992).

However, the spreading of the truth-value alteration works as a mechanism of conceptual change only when constituent pieces of knowledge of both old and new theories are available in the knowledge structure. Individuals may stick to the old theory even when predictions from it are not supported, if they do not think of any alternative theory, model, or interpretation. Thus, conceptual change may be triggered by analogy that suggests an alternative view. Using analogies, people map their knowledge about the source to the present new case (target) so that they can make a coherent interpretation of the set of observations for the target, or even build a tentative theory. Dunbar (1995) reports that scientists in real-world laboratories often rely on analogy to generate new ideas when faced with inconsistent experimental findings.

Social interactions, including group discussion, can be sources for generating inconsistencies or discoordinations in individuals' existing knowledge systems, because different perspectives are presented in the interactions, and they can also be sources for providing possible solutions. In this sense, discussion or question answering can contribute to the occurrence of conceptual change. Thus, conceptual change may be induced through discursive interactions even among less sophisticated thinkers (Hatano and Inagaki 1997). It can also be a powerful mechanism of inducing conceptual change among scientists, as reported by Dunbar (1995).

5. Issues Yet to Be Studied

Some psychologists emphasize the roles of motivational factors such as goals, interests, values, and so on as mediators of the process of conceptual change. A number of studies on classroom learning, for example, strongly suggest that learners' different goal orientations lead to different processes of conceptual change (Pintrich et al. 1993). However, we know very little about these variables.

Another unsolved issue is how the conscious monitoring of multiple knowledge systems and the flexible switches among the systems are acquired. One of the differences between young children's naive theory and adults' intuitive theory is that the latter has metaconceptual components, probably as a result of conceptual change. We do not know much about this issue because it has been neglected in conceptual change research.

See also: Cognitive Development: Child Education; Cognitive Development in Childhood and Adolescence; Cognitive Development: Learning and Instruction; Concept Learning and Representation: Models; Connectionist Models of Concept Learning; Discovery Learning, Cognitive Psychology of; Early Concept Learning in Children; Explanation-based Learning, Cognitive Psychology of; Intelligence, Prior Knowledge, and Learning; Piaget's Theory of Child Development; Piaget's Theory of Human Development and Education

Bibliography

Carey S 1985 *Conceptual Change in Childhood*. MIT Press, Cambridge, MA

Clement J 1982 Students' preconceptions in introductory mechanics. *American Journal of Physics* **50**: 66–71

Dunbar K 1995 How scientists really reason: Scientific reasoning in real-world laboratories. In: Sternberg R J, Davidson J E (eds.) *The Nature of Insight*. MIT Press, Cambridge, MA, pp. 365–95

Gopnick A, Wellman H M 1994 The theory theory. In: Hirschfeld L A, Gelman S A (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, New York, pp. 257–93

Hatano G, Inagaki K 1996 Cognitive and cultural factors in the acquisition of intuitive biology. In: Olson D R, Torrance N (eds.) *Handbook of Education and Human Development: New Models of Learning, Teaching and Schooling*. Blackwell, Oxford, UK, pp. 683–708
Hatano G, Inagaki K 1997 Qualitative changes in intuitive biology. *European Journal of Psychology of Education* **12**: 111–30
Kuhn T 1970 *The Structure of Scientific Revolutions*, 2nd edn. Chicago University Press, Chicago
Nersessian N 1992 How do scientists think? Capturing the dynamics of conceptual change in science. In: Giere R N (ed.) *Minnesota Studies in Philosophy of Science: Vol. 15. Cognitive Models of Science*. University of Minnesota Press, Minneapolis, MN
Perner J 1991 *Understanding the Representational Mind*. MIT Press, Cambridge, MA
Pintrich P R, Marx R W, Boyle R A 1993 Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research* **63**: 167–99
Smith C, Carey S, Wiser M 1985 On differentiation: A case study of the development of the concepts of size, weight, and density. *Cognition* **21**: 177–237
Stavy R, Wax N 1989 Children's conceptions of plants as living things. *Human Development* **32**: 88–94
Vosniadou S, Brewer W F 1992 Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology* **24**: 535–85
Vosniadou S, Ioannides C 1998 From conceptual development to science education: A psychological point of view. *International Journal of Science Education* **20**: 1213–30
Wellman H M 1990 *The Child's Theory of Mind*. MIT Press, Cambridge, MA

K. Inagaki

Conditioning and Habit Formation, Psychology of

Research on *conditioning and habit* broadly is concerned with how organisms learn the *what*, *where*, *and when* of their niche: what prey and threats they have to deal with, where they may be found, and when they are most likely to be encountered. These are complex problems but their solutions can be surprisingly simple, so long as they are clever.

Conditioning and habit involve learning about *contingencies*—if-then relationships—between biologically important stimuli and behavior. Conditioning involves learning signals for *stimuli*, *stimulus–stimulus* contingencies, such as between the odor of a lemon and acidity; habits involve learning about the consequences of behavior, *response–stimulus* contingencies, such as learning how to crack open an oyster to extract its flesh. The *mechanisms* of learning serve the organism by *selecting* the best signals for stimuli and the effective actions leading to reward. In terms of adaptation, conditioning helps the organism to prepare for forthcoming stimuli and habit learning allows the organism to learn from experience. Despite its apparent plasticity, learning is highly biologically constrained, every species is differently affected by particular stimuli and consequences as defined by its ecological niche, while the processes of learning themselves remain identical across species.

Learning is studied by arranging environmental *operations*, such as response–stimulus contingencies, and analyzing the behavioral *process* they generate. Behavior is sensitive to many interacting variables. For this reason, much research in learning has been carried out with animal subjects for whom stricter control of operations is possible, but most findings have been replicated with human subjects. Meanwhile, however, people with pragmatic interests in managing human behavior in educational, clinical, industrial, and correctional facilities, have benefited greatly from our understanding of learning.

1. History

Psychological investigation has been responsible for unraveling the intricacies of learning, but the central questions were first recognized by philosophers and biologists. The philosopher Descartes proposed two sources of behavior, a reflexive–mechanical body that connected the body to the physical world and a rational–spiritual mind that explained language, memory, and intellect. Intuitively, the Cartesian dualism captures the striking contrast between reflexes we all share, and the many idiosyncratic differences that make each unique. The first half of the Cartesian dichotomy was supported in the eighteenth century when Whytt and others isolated the spinal reflexes, but the assumption of native mental knowledge was harder to endorse. *Empiricist* philosophers such as Locke supplied a viable alternative position, that complex knowledge could be explained by the accumulated experience of *associated* ideas. By the mid nineteenth century Alexander Bain offered an associationist theory of *action*. Prompted by speculation of the physiologist, Johannes Müller, Bain proposed that in addition to reflexes, organisms randomly emitted spontaneous nonpurposive movements. When one of these movements produced a pleasurable outcome, it would be repeated because the 'muscle sense' produced by the action would be associated with pleasure. This early conception of habit acquisition gained acceptance amidst the rising tide of Darwinism by providing a mechanistic alternative to the creationist spirit implicit in Descarte's mentalism.

Approximately 40 years later, Edward L. Thorndike published the first experimental demonstrations showing how behavior could be affected by its consequences. One of his cat 'puzzle boxes' is shown in Fig.

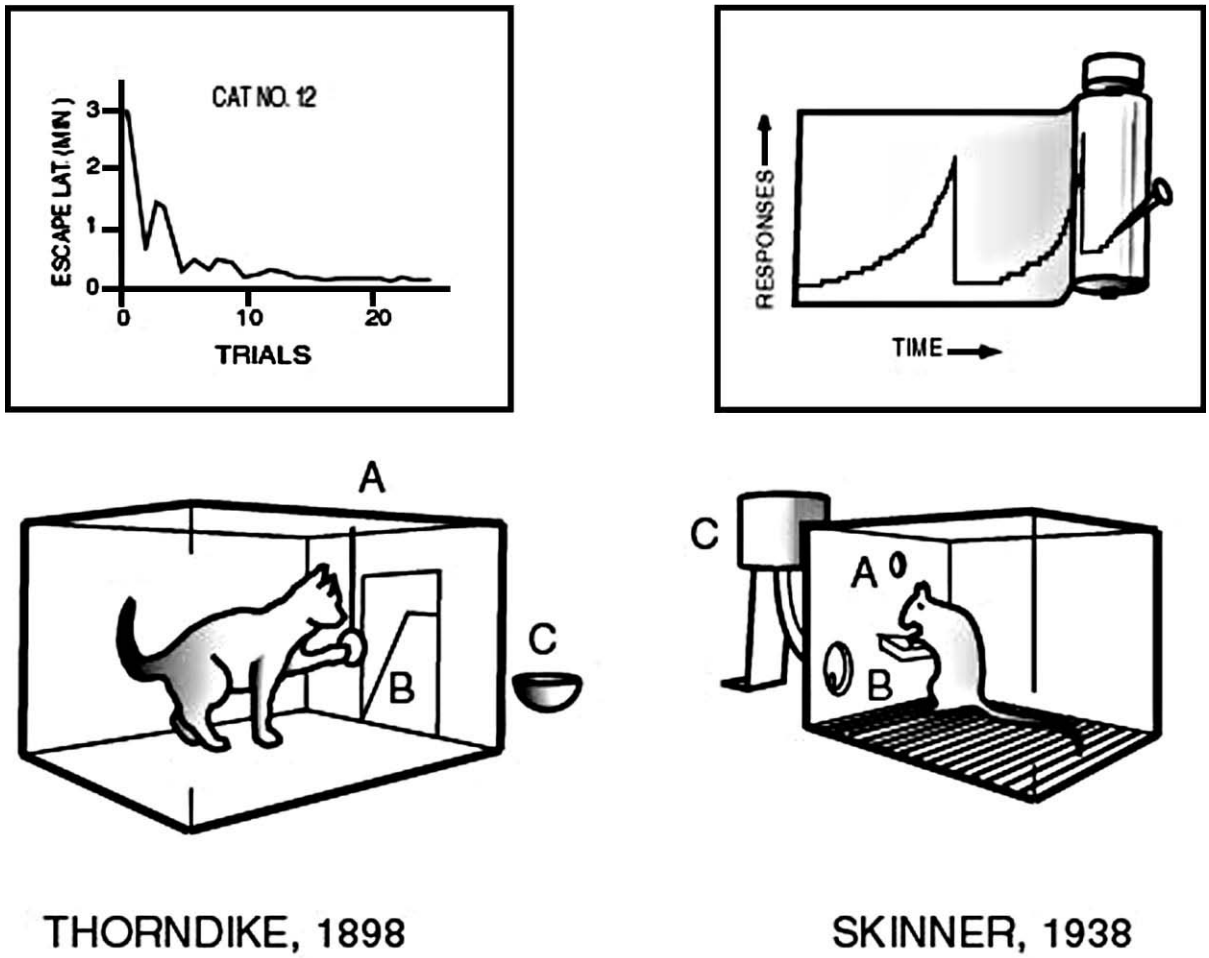

Figure 1
Thorndike's 1898 cat puzzle box and Skinner's 1938 rat lever box. Pulling on (A) the ring opens (B) the door, giving the cat access to (C) the food. The inset above the puzzle box is a learning curve, showing how latency to pull the ring decreased over trials. In the rat lever box, (A) the white light signals when pressing (B) the lever will produce (C) a food pellet. The inset above the lever box shows a cumulative record, recording lever presses and reinforcements in real time

1. Latency to escape from the box was long initially but gradually shortened. Thorndike called the behavior he observed in the puzzle-box, 'trial-and-error learning' and formulated the *Law of Effect*, stating that pleasurable consequences 'stamped in' responses and annoying consequences 'stamped out' responses.

Coincident with Thorndike's research, the Russian physiologist Ivan P. Pavlov accidentally discovered the *conditioned salivation reflex* in the dog. In the course of studying digestive reflexes, Pavlov noted that his dogs began to salivate while their food was being prepared, before the food made contact with the dog's mouth. He realized that more than a simple reflex was involved—the dog had learned a signal for food—initiating an extensive research program to study *conditioned reflexes*. Fig. 2 shows the setup he used to study stimulus–stimulus contingencies. Pavlov demonstrated conditioning with many other reflexes, leading him to claim that conditioned reflexes could explain all complex behavior including language.

Initially Pavlov's work eclipsed Thorndike's because of arguments that the Law of Effect might be explained by conditioned reflexes. But in the 1930s, B. F. Skinner endorsed the present view that Thorndike and Pavlov had discovered different processes. In time, conditioned reflexes became identified with *Pavlovian conditioning* and habit with *operant conditioning* (emphasizing that responses *operate* on

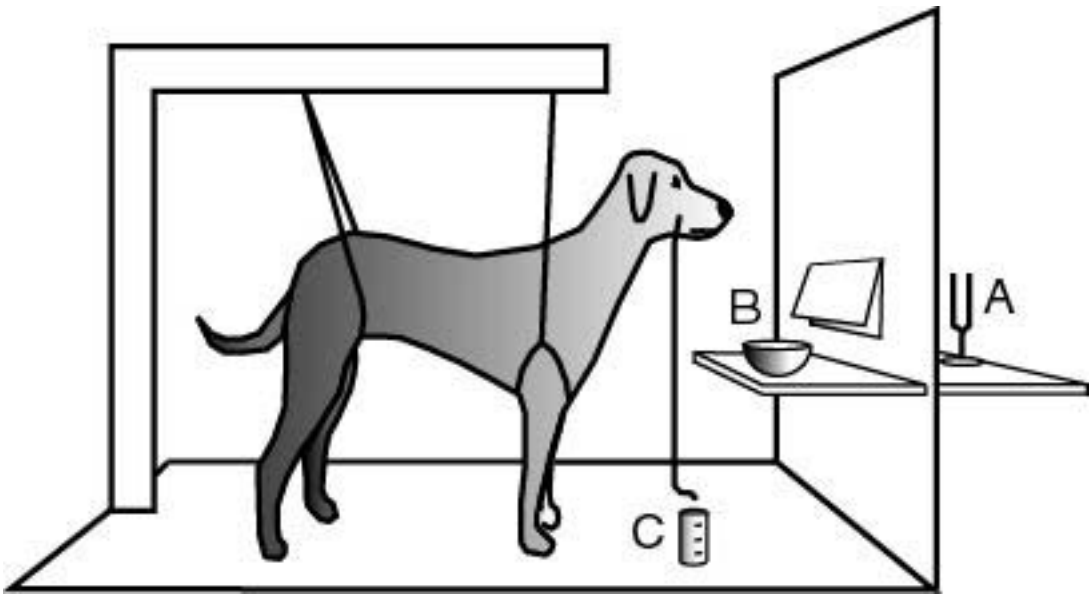

Figure 2
Pavlov's 1904 apparatus to study conditioned reflexes in the dog. In a typical conditioning trial, the dog was first presented with (A) a tone from a tuning fork, and closely followed by (B) food presented through a trap door. Saliva from the mouth was measured by (C) a recording device before and during each trial

the environment). Skinner made a number of important contributions to the study of operant behavior, including an automated experimental chamber to study reinforcement with rats and a *cumulative recorder* to record behavior in time (Fig. 1). His simplest box contained a lever, a food dispenser, and a light to signal the availability of food. Unlike puzzle boxes and mazes that studied learning in a trial-by-trial format, Skinner's automated apparatus permitted analysis of 'free-operant behavior,' the real-time interaction between behavior and the environment. He used his apparatus to conduct numerous critical experiments on reinforcement, including studies of discrimination learning and reinforcement schedules.

The genesis of experimental psychology in the early twentieth century was accompanied by disputes over the proper methods of investigation and the role of inference from behavior to mechanism. John B. Watson in 1913 proposed a 'Behaviorist' program, arguing that psychology should reject the introspective methods popular in his day and instead be founded only on publicly observable behavior. Watson's program is now called *methodological behaviorism*, and to the extent that psychological concepts are founded on observable behavior, his ideas have been widely accepted. Thorndike, Pavlov, and Watson, among many others, were committed to physiological *reductionism*, explaining behavior by inferring its underlying physiology. Skinner pushed for a truly independent science of behavior based entirely on the study of behavior–environment interactions, a program he called *radical behaviorism*. Behaviorism in turn was countered by *cognitivism* and *neo-behaviorism*. Both argue that behavior can be used to make reductionist inferences about nonphysiological mechanisms that mediate between the environment and observable behavior, such as *memory* and *associative* processes. The later half of the twentieth century has witnessed the rise of mathematical, dynamic models of behavior, called *quantitative behaviorism*, that describe mediators of behavior in terms of mathematical functions. This alternative to cognitivism seeks to model behavior–environment interactions mathematically, but the models are seen only as analytical tools to determine the minimum set of mechanisms required to explain behavior, and not to make mediational inferences. Despite philosophical differences, all of these views continue to have proponents and make contributions to the understanding of behavior.

Researchers and theorists have since broadened the relevance of Pavlovian and operant conditioning to economics, language, problem solving, and other varieties of complex behavior. At the same time, the science of learning has come increasingly to recognize that behavior is constrained by biological endowment. Despite repeated attempts to reduce Pavlovian and operant conditioning to a single process, they differ in essential ways. Although strong parallels occur in the effects of contingencies and contiguity, and in stimulus functions, response functions differ markedly.

2. Contingency and Contiguity

In a *reflex*, a stimulus *elicits* a response; in an *operant*, an *emitted* behavioral response produces a stimulus and the behavior becomes more or less likely to occur in the future, depending on whether the consequence is a *reinforcer* or a *punisher*. Fig. 3 illustrates these basic relationships.

Organisms are very sensitive to stimulus situations, or signals, present when stimuli occur. Signals impart enormous complexity in behavior by giving rise to the *conditioned reflex* and *discriminated operant* (Fig. 3). In Pavlov's conditioned reflex experiments with dogs, food was paired with a tone, and the dog began to salivate upon hearing the tone. In Skinner's discriminated operant experiments with rats, lever presses could produce food pellets when a stimulus light was on but not when the light was off, and the rats came to press the lever only when the light was on. For a conditioned reflex, the *conditioned stimulus* (CS) *elicits* a *conditioned response* (CR); for a discriminated operant, the discriminative stimulus (S^D) *occasions* an *emitted response*.

2.1 Contingency

Effects of operant and Pavlovian operations depend on *contingencies*, the conditional probabilities between stimuli and responses shown in Fig. 4. Operant behavior and Pavlovian conditioning require a *positive* contingency in which a consequence or a US is most likely after a response or a CS, respectively. A *zero* contingency, in which the occurrence of a stimulus is independent of other events, will not support operant behavior or Pavlovian conditioning. The inadequacy of a zero contingency shows that *contiguity*, the concurrence between an operant response and a

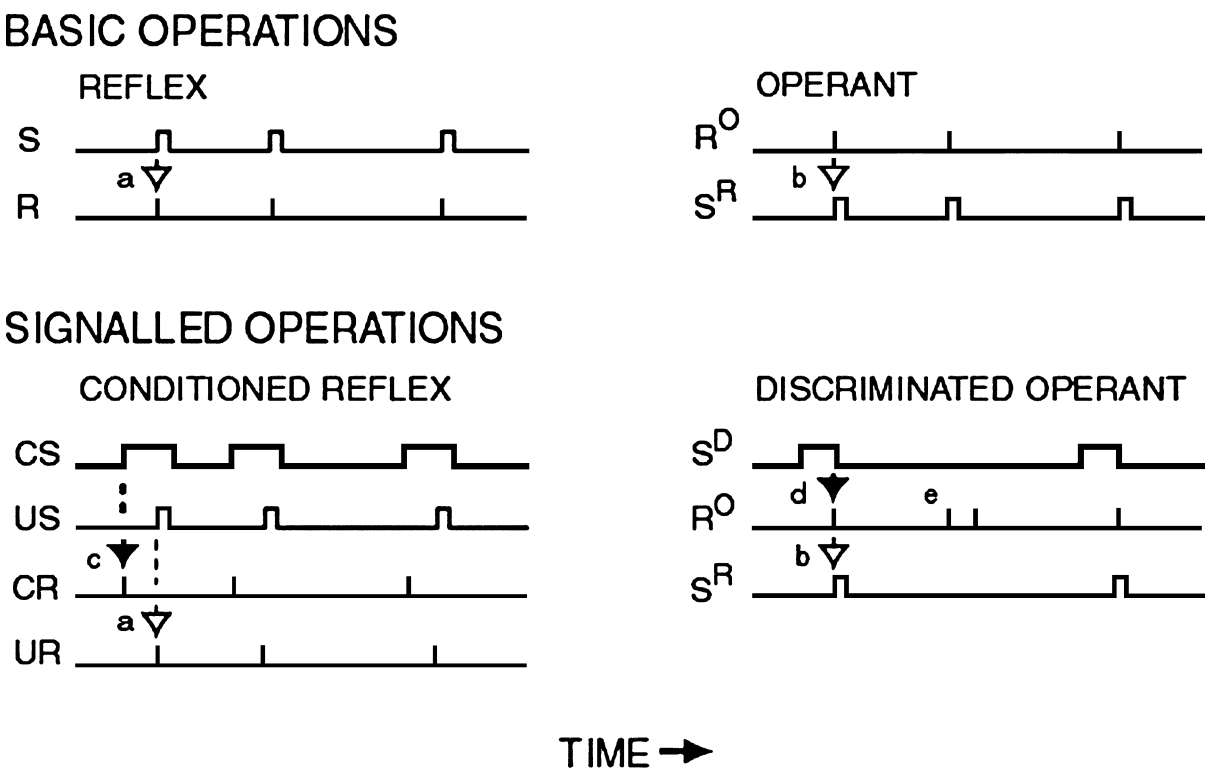

Figure 3
Reflex and operant relationships between stimuli and responses. Unfilled arrows in the upper diagrams show causal directions in reflex and operant behavior: in the reflex, a stimulus produces a response (R, *a*); in the operant, an operant response produces a reinforcing stimulus (S^R, *b*). Lower diagrams illustrate the addition of signals (indicated by bold lines). In the *conditioned reflex*, the conditioned stimulus (CS) signals the presentation of an unconditioned stimulus such as food (US); the US elicits an unconditioned response (UR; *a*) and the CS elicits a conditioned response (CR; *c*). In the *discriminated operant*, a discriminative stimulus (S^D) signals when responses (*d*) are scheduled to have consequences (S^R; *b*), responses occurring in the absence of stimuli (*e*) have no consequences. The S^D can also become a CS because of the way it is paired with the S^R

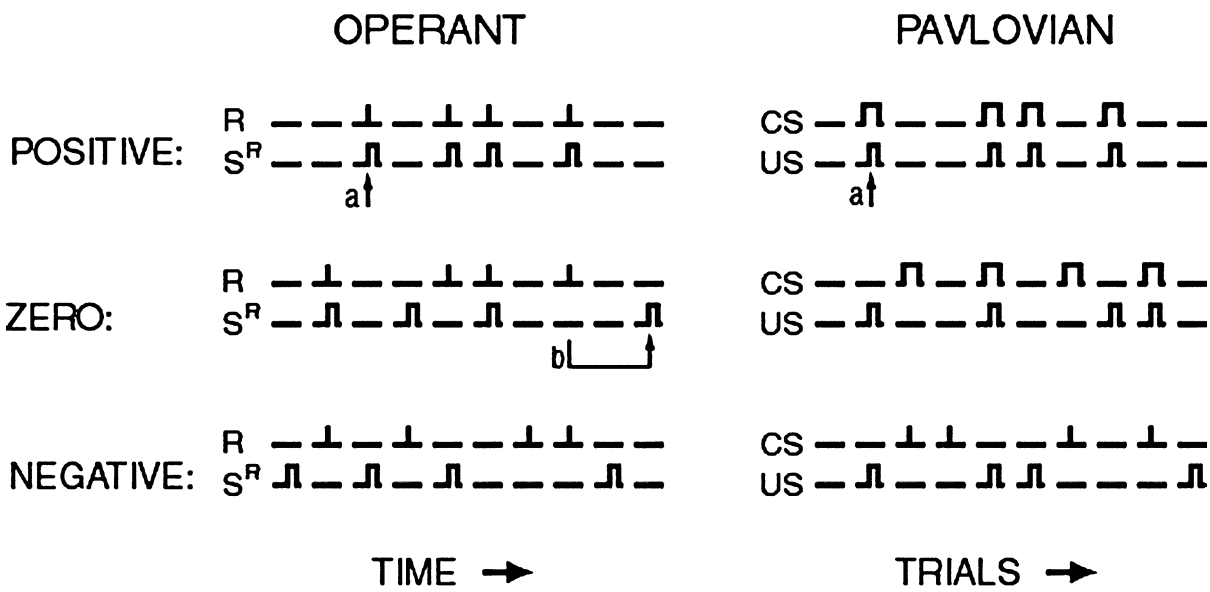

Figure 4
Operant and Pavlovian contingencies. Diagrams on the left depict the occurrence and non-occurrence of operant responses (R) and stimuli (S^R) over 10 intervals of time. Diagrams on the right show the occurrence and non-occurrence of a Pavlovian CS and US over 10 conditioning trials. Contiguous occurrences of events are indicated by arrows labeled *a*. In the *positive* operant contingency (upper left), there are four responses and each produces a *contiguous* reinforcer and there are no reinforcers without prior responses: the probability of a stimulus given a response (i.e., a consequence) is higher than the probability of a stimulus given no response. In the operant zero contingency (middle left), the two conditional probabilities are equal. And in the operant *negative* contingency, (bottom left) the probability of a stimulus given a response is lower than the probability of a stimulus given no response. Operant contingencies are complicated by the fact that a stimulus may follow a response after a delay (indicated by arrow *b*). In Pavlovian conditioning, positive, zero, and negative contingencies are defined in the same manner except that probabilities are calculated for the US given a CS and given no CS

consequence or between a CS and US, is not sufficient for learning; there must also be a positive contingency. While zero contingencies do not generate reinforcement effects, prior exposure to noncontingent, unpredictable stimuli can significantly retard subsequent learning of reinforcement contingencies.

2.2 Contiguity

Operant consequences are most effective if they are contiguous with responses and less effective if they are delayed—contingency is necessary but not sufficient. In animals, delays of just a few seconds can severely weaken a reinforcer; and reinforcers that are delivered independently of responses, as in a zero contingency, diminish the effect of contiguous reinforcers. Negative contingencies, also called *differential reinforcement of other behavior*, are an effective way to reduce problem behavior in applied settings.

Pavlovian conditioning is also sensitive to the temporal relation between the US and CS—contingency is not sufficient. Of the five possible temporal relations shown in Fig. 5, the most effective is *short-delay* conditioning in which the onset of the CS precedes the onset of the US by no more than a second. *Simultaneous* conditioning is least likely to produce conditioning; *backward* conditioning tends to make the CS an *inhibitory* stimulus because it is correlated with the absence of the CS; and *long-delay* and *trace* produce intermediate levels of conditioning. Taken together, these findings indicate that Pavlovian conditioning is strongest when the CS is a good predictor of the US in time. The one exception is *taste aversion learning*, an example of ecological *preparedness*—illness can produce aversion to a food even if it follows consumption by several hours.

2.3 Extinction

Conditioned reflexes and operant responses can be weakened by *extinction*. Operant behavior maintained by reinforcement can be reduced in frequency by terminating the reinforcement contingency or by delivering reinforcers independently of responses.

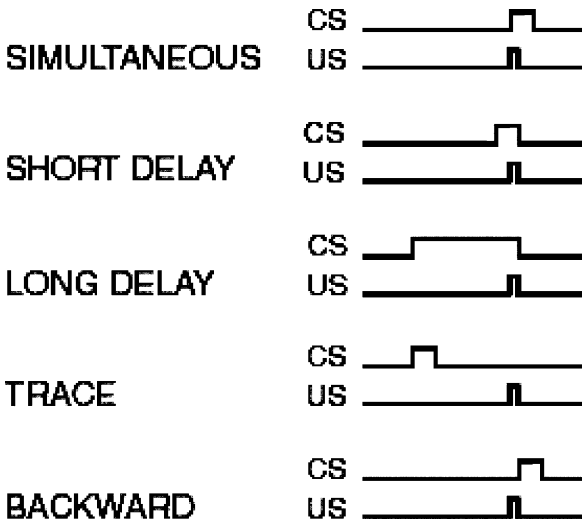

Figure 5
Temporal relations between CS and US in Pavlovian conditioning

Discontinuing a punishment contingency almost always results in the recovery of the punished behavior. A CS can be *extinguished* by presenting it without the US or by removing the correlation with the US. Extinction procedures do not erase the effects of contingencies but instead produce additional learning. In the case of operant behavior, organisms stop responding because they learn that the operant contingency is broken—extinguished responses reappear without additional training when a reinforcement contingency is reinstated. In the case of Pavlovian extinction, the organism learns that the CS now signals the absence of the US.

2.4 Operant–Pavlovian Interactions

Conditioned reinforcers and *conditioned punishers* can be generated by Pavlovian procedures. For example, if an odor is paired with food, the odor will elicit salivation as a CS but it can also serve as a conditioned reinforcer for an operant response. Conditioned punishers are created in the same way. Conditioned stimuli greatly enhance delayed consequences, for example, a delayed reinforcer will still be very effective as long as a stimulus bridges the gap between the response and the reinforcer (as in long-delay conditioning in Fig. 5). However, the effects are transitory unless the association between the conditioned stimulus and the reinforcer is maintained.

An operant S^D can become a Pavlovian CS for operant consequences such as food. The S^D then serves two functions that may produce competition between elicited and emitted responses. An example from the animal-training literature is shown in Fig. 6. A pig was reinforced with food for depositing coins in a bank. Shortly after learning the operant, the pig began 'rooting' the coins—pushing them along the ground, the pig's species-typical food handling behavior. As a result of the operant contingency the coin enters into a positive Pavlovian contingency with food and elicits rooting.

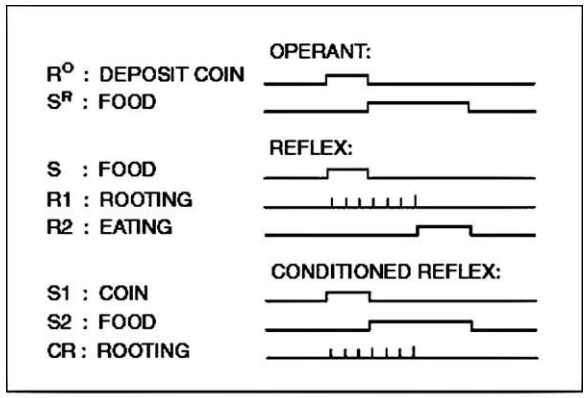

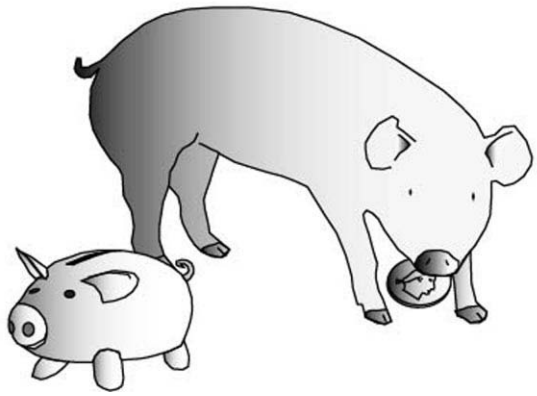

Figure 6
'Misbehavior' in the pig—conflicting operant and Pavlovian responses. The pig was trained to deposit coins in the bank for food reinforcers, but extended periods of rooting come to disrupt the operant contingency. The inset shows that, in the pig's natural behavior, food elicits reflexive rooting and eating. This creates a Pavlovian contingency between the coin CS and the food US, whereby the coin comes to elicit conditioned rooting

Punishment can reduce the frequency of operant behavior for a similar reason—a painful, punishing consequence reduces behavior because it elicits species-typical defensive behavior such as escape and avoidance. The result is that 'punisher' reduces the target behavior but not just by being a consequence of that behavior; it would still reduce the operant behavior even if arranged in a zero contingency. A

conditioned aversive stimulus, such as a tone paired with shock, will similarly interfere with operant behavior. As a rule, reinforcers are most effective when the responses they elicit are identical to or compatible with a reinforced response and punishers are most effective when the responses they elicit are incompatible with the punished response. For example, it is easy to reinforce pecking in the pigeon with food because food elicits pecking.

3. *Response Functions*

Consequences in operant behavior and US's in Pavlovian conditioning can be *appetitive*, like food, or *aversive*, like a very loud noise. The same stimulus can generate different kinds of adaptive behavior depending on the contingency, the kind of organism, and the organism's previous experience.

3.1 *Operant Responses*

In operant behavior, responses can accrue or remove stimuli, and the stimuli can be appetitive or aversive. In positive reinforcement and negative reinforcement response rate increases by addition of appetitive stimuli or removal of aversive stimuli; positive punishment and negative punishment decreases response rate by addition of aversive stimuli or removal of appetitive stimuli. Negative reinforcement includes escape, removal of an already present aversive stimulus, and avoidance, prevention or postponement of an aversive stimulus.

Operant responses can include everything from a rat's lever press maintained by food, an infant's crying maintained by maternal attention, and small talk maintained by social companionship. The form of an operant response depends on (a) environmental constraints, such as the height of a reinforcement lever and force required to operate it, (b) the behavior emitted by the organism, and (c) behavior elicited by the consequence (illustrated by Fig. 6). Once a response is reinforced, its form tends to be repeated in a stereotyped manner. Reinforcement will 'capture' whatever behavior the organism emits that is contiguous with the reinforcer; even if some features are unessential to the reinforcement contingency.

Reinforcement narrows the range of variability in emitted behavior by *selecting* its successful forms (not necessarily the most efficient behavior), making them more frequent and thereby displacing less effective forms. Predatory species, for example, show adjustments in hunting behavior as a result of their effectiveness. The *differential reinforcement* of some responses to the exclusion of others is called *shaping*. Shaping is analogous to natural selection in that different responses function like the phenotypic variations required for evolutionary change. Without variation, changes in reinforcement contingencies could not select new forms of behavior. Despite the tendency for reinforcement to produce stereotyped behavior, it never reduces variations in behavior entirely. Variability returns to emitted behavior in extinction, a potentially adaptive reaction to dwindling resources. Creativity can also serve as the basis for reinforcement (e.g., if reinforcers are contingent on the emission of novel behavior.)

Restriction operations, such as food deprivation, are necessary for reinforcement—the response elicited by the reinforcer, such as feeding, must have some probability of occurring for the reinforcer to be effective. In rats, restricting an activity such as running in a wheel will make the activity a reinforcer for drinking, even thought the rat is not water restricted. Findings like this suggest that reinforcers may best be understood in terms of the activities they produce and that the reinforcement process involves the organism's behaving in order to maintain set levels of activities.

Once an organism learns a response-reinforcer contingency, the likelihood of the response will reflect the value of the reinforcer. For example, if a rat drinks sugar water and becomes ill, it will no longer emit behavior previously reinforced with sugar water.

A great deal of behavior is maintained by negative reinforcement, including *avoidance* in which responses prevent the occurrence of aversive stimuli, and *escape* in which responses terminate an existing aversive stimulus. Much research has been directed to avoidance because it is maintained in the absence of contiguity—*responses prevent reinforcement*. In some cases, avoidance is maintained by deletion of an immediately present *conditioned* aversive (warning) stimulus, an avoidance condition that preserves contiguous reinforcement. But organisms will work to prevent or delay aversive events without the benefit of warning stimuli, showing that avoidance does not require a Pavlovian aversive contingency. Avoidance is closely linked to punishment because organisms frequently learn to avoid places associated with punishment, a pragmatic concern for advocates of punishment.

The timing, rate, and persistence of an operant response depends on its *reinforcement schedule*, a constraint on earning reinforcers that requires the organism to wait for the next available reinforcer, emit a number of responses, or some combination of both. Fig. 7 shows typical patterns of behavior on four simple schedules. In an *interval* schedule, a reinforcer is produced by the first response after the passage of a specified interval of time. In a *ratio* schedule, a reinforcer is produced after the emission of a specified number of responses. Time or response requirements can either be a fixed or variable. Response rate on both ratio and interval schedules increases with reinforcement rate up to a point, and then decreases again at high rates of reinforcement. At low rates of reinforcement or during extinction, behavior on ratio schedules alternates between bouts of rapid responding and

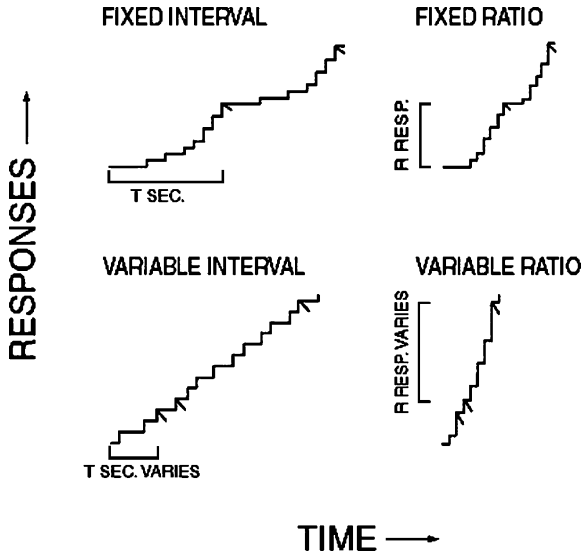

Figure 7
Typical response patterns on four simple schedules of reinforcement: fixed interval, fixed ratio, variable interval, and variable ratio. Ratio schedules produce higher rates of responding than interval schedules at a giver rate of reinforcement, and fixed schedules show longer pausing after reinforcement than variable schedules

pauses, while responses on interval schedules simply occur with lower frequency. Extinction is prolonged after exposure to reinforcement schedules—especially after exposure to long interval- or large ratio-schedules.

When rate of reinforcement in interval and ratio schedules is equated, organisms respond at higher rates on ratio schedules (although the difference is small at very high rates of reinforcement). Organisms are sensitive to the fact that the rate of reinforcement in ratio schedules increases directly as a function of response rate, but not in interval schedules.

Characteristic pausing or *waiting* occurs after earning a reinforcer. Waiting is determined by the currently expected schedule and not because of the reinforcer. Fixed-interval and fixed-ratio schedules produce long wait times, and variable-interval and variable-ratio schedules produce short wait times. Waiting on fixed-interval schedules is reliably a fixed *proportion* of the interval value. In VI and VR schedules, wait times are heavily influenced by the shortest inter-reinforcement intervals or ratios in a schedule. Waiting is obligatory—organisms will wait even when doing so reduces the immediacy and overall rate of reinforcement. In schedules that require a single response, with the reinforcer occurring a fixed time after the response, the optimal strategy is to respond immediately in order to minimize delay, but organisms still wait a time proportional to the fixed time before making the response that leads to reinforcement.

Schedules can be combined in almost limitless ways to study questions dealing with choice, conditioned reinforcement, and other complex behavioral processes. In experiments on choice between two *concurrent* variable-interval schedules, the relative preference of many organisms, including humans, will closely match the percentage of reinforcement provided by each schedule. With choices between different fixed-interval schedules, the preference for the shorter interval is far greater than predicted by the relative intervals, indicating that the value of a delayed reinforcer decreases according to a decelerating function over time. Conditioned reinforcers have been studied with *behavior chains* in which one signaled schedule is made a consequence of another, for example, a one-minute fixed interval schedule signaled by a red stimulus leads to another one-minute fixed interval signaled by a green stimulus, until a terminal reinforcer is obtained. Behavior extinguishes in the early schedules of chains with three or more schedules, to the extent that reinforcement is severely reduced. Possibly conditioned reinforcers must be contiguous with primary reinforcers to effectively maintain behavior; alternatively obligatory waiting in the early schedules extends the time to the terminal reinforcer, lengthening the time to reinforcement, and generating even longer waiting.

3.2 Pavlovian Responses

Pavlovian conditioning, based on reflexes that serve prey capture, feeding, defense, escape, reproductive behavior, and care of offspring, has been demonstrated with a vast number of conditioned stimuli. Conditioned responses include fear (e.g., acceleration in heart rate and perspiration), sexual arousal, and immunological responses. Pavlovian conditioning serves the organism by permitting it to anticipate the stimuli that elicit these reflexes, so it should follow that the conditioned response should resemble the unconditioned response. (A notable exception occurs with drug US's, for example, the UR to morphine is euphoria, but the CR to a CS for administration (e.g., a syringe) is illness rather than euphoria.)

The form of the conditioned response can be influenced by both the US and CS. For example, in the rat, a localized-light CS paired with a food US produces rearing on the hind legs, but the same localized-light paired with shock will produce muscular tension; while a tone paired food leads to an increase in activity. A Pavlovian positive contingency produces excitatory conditioning in that the CR becomes more likely; while a Pavlovian negative contingency produces inhibitory conditioning in that the CR becomes less likely (Fig. 4). For example, fear will be elicited by a tone CS that reliably precedes a shock US; but a light that reliably precedes periods of

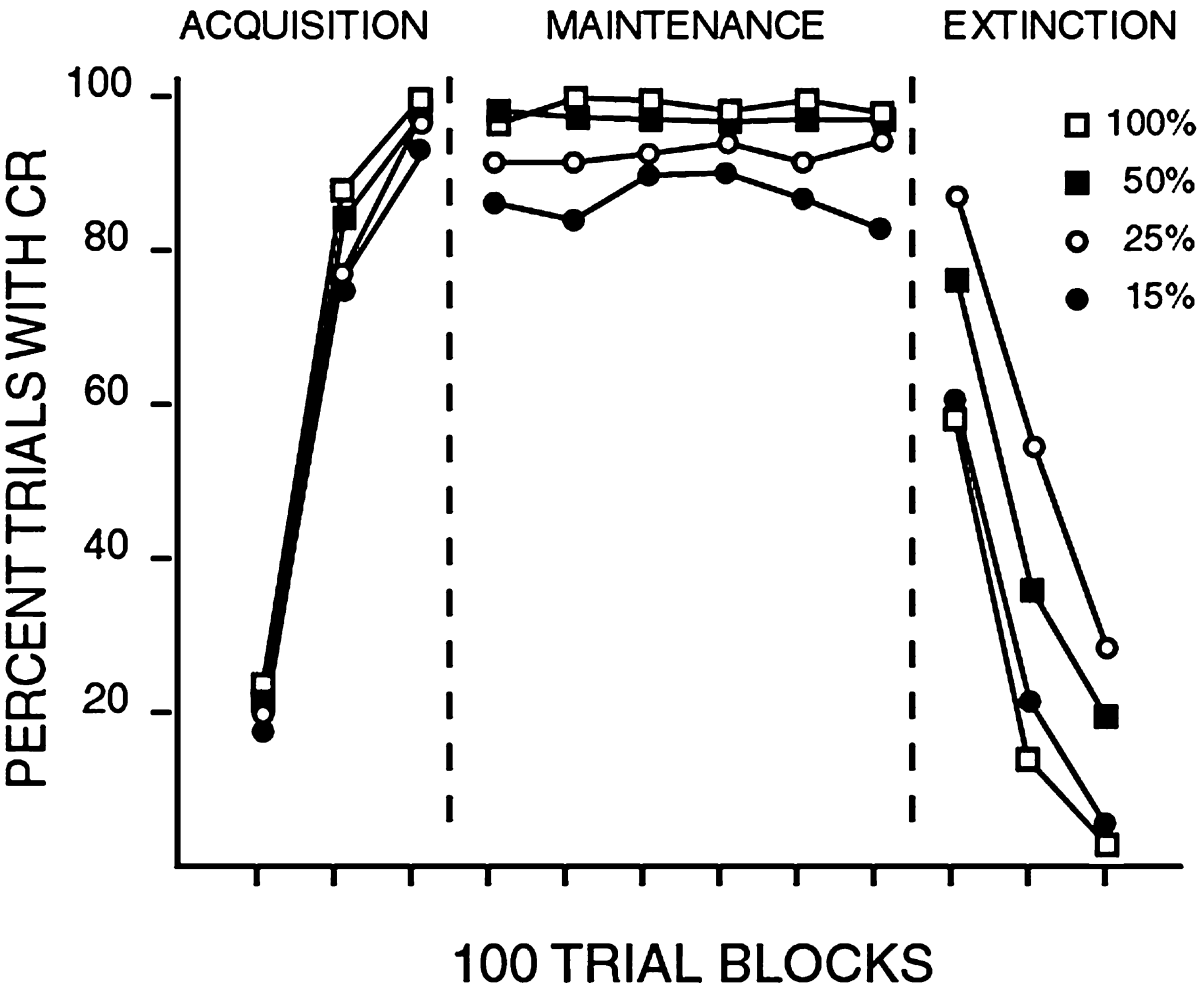

Figure 8
Acquisition, maintenance, and extinction of a Pavlovian CR, eye-blink in the rabbit, with different percent pairing between the CS and US (from Gibbs et al. 1978)

no shock will inhibit fear when paired with CS for shock. A conditioned inhibitor can also be produced by occasionally pairing a stimulus with an excitatory CS (a compound CS) while omitting the US or when a CS is extinguished.

Acquisition of a CR proceeds in a negatively accelerated pattern, illustrated in Fig. 8. Also shown in Fig. 8 is the effect of reducing the probability of a US given a CS (keeping the probability of a CS given no US at zero). Reducing the number of times a CS presentation is followed by a US has the effect of slowing acquisition, reducing the asymptotic number of trials with a CR, and slowing extinction. The acquisition and probability of the CR is also greatly affected by the ratio of CS duration divided by the inter-US interval, with smaller ratios producing higher CR probabilities.

A number of experiments have addressed the question of whether Pavlovian conditioning involves learning associations between the CS and the US, or between the CS and the UR. In experiments in which the CS is paired with a US while the UR is pharmacologically blocked, subjects show a CR to the CS after the block is removed, suggesting CS–US learning. But the matter has not been resolved because the US might still be activating efferent centers in the nervous system.

4. *Signal Functions*

Organisms can only respond adaptively to contingencies if they provide reliable signals in the form of CS's and discriminative stimuli. In order to maximize

efficiency, an organism must learn a signal after minimal trials, recognize variations, and reject similar but invalid stimuli. Moreover, the organism can't afford to make too many mistakes or spend too much time learning each variant. The ecological constraints are similar in Pavlovian and operant conditioning, so stimulus functions in both share many features.

4.1 Attention, Blocking, and Preparedness

Organisms do not attend to all stimuli that they can sense when a US or a reinforcer is presented; instead, they are selective. When discriminative stimuli contain multiple dimensions, such as shape and color, organisms may show attention to one dimension and ignore the other. A previously learned CS or discriminative stimulus can also block attention to a stimulus. For example, after pairing a light CS with a food US, further training with a light + tone CS will not produce learning of the tone CS. This blocking effect shows that contingency and contiguity are not sufficient for learning about signals—learning will only take place if there are no other reliable signals present.

Attention has a strong genetic component because contingencies associated with a particular niche tend to have predictable signals. Rhesus monkeys will fear a snake CS if it is presented with a fear response US from another monkey, but not a flower CS paired with the same US. In operant behavior, several food-storing birds selectively attend to the spatial locations of food sites and not their non-spatial aspects such as colors. In each case, the animal attends to the most ecologically relevant cue.

4.2 Discrimination and Generalization

An organism must generalize responding to stimuli that are different to the training stimulus, but not too different. With many stimuli, such as tones and colors, generalization shows a peak around the CS or discriminative stimulus that is called a *generalization gradient*—the probability of responding decreases as a function of difference from the training stimulus. Fig. 9 shows a typical generalization gradient for operant behavior. Discriminative stimulus control is tighter (peaked gradient) when reinforcement in the presence of the discriminative stimulus was alternated with periods without the stimulus (presence–absence training), in comparison to training without extinction periods (nondifferential training). This finding illustrates that discrimination requires differential training; without differential training, organisms generalize broadly.

Signals in the natural environment are likely to be plants, insects, odors, and so forth. Animals readily learn these *natural concepts*. Experiments investigating the basis of concept learning in animals find that some concepts such as cars require training with several examples but others, such as an oak leaf, can be learned after a single exemplar.

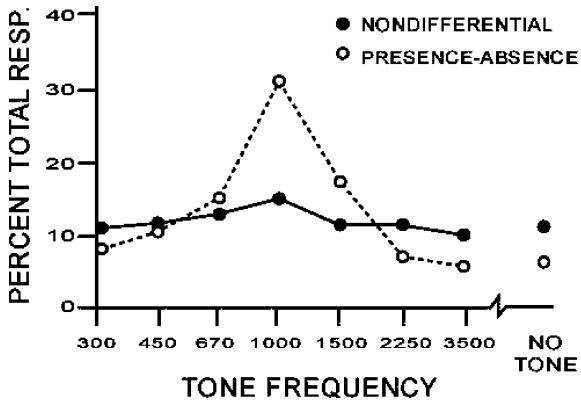

Figure 9
Generalization gradients from pigeons after nondifferential training with a 1000 Hz tone, and presence-absence training with and without the tone (from Jenkins and Harrison 1960)

4.3 Conditional Discrimination and Occasion Setting

Reinforcement can be made contingent on conditional relationships between stimuli. In an identity matching task, a pigeon is reinforced for responding on the basis of similarity: selecting the 'triangle' comparison is conditional on the 'triangle' sample. Conditional discrimination is the basis of symbolic behavior. For example, when prompted with 'What color?' in the presence of a blue square, we say 'Blue'; and with 'What shape?' we say, 'square.'

Animals and humans can learn conditional discriminations such as 'same' and 'different,' as well as symbolic matching between arbitrarily related stimuli such as words and their referents. However only humans can demonstrate certain kinds of emergent conditional discriminations. Once humans learn the symbolic stimulus relation, 'if stimulus A then B,' they can respond correctly to a symmetry test, 'if B then A,' without further training. The variables that explain symmetry are not yet understood; the ability may be a uniquely human accomplishment possibly emerging with language. Interestingly, however, both humans and other species show transitivity: 'if A then B' and 'if B then C'; therefore, 'if A then C.'

In Pavlovian conditioning, a stimulus that signals when a CS will be followed by a US will modulate the CR to the CS. For example, if tone + light CS is followed by a food US, but light by itself signals no food, then light will only show a food CR if tone is also present. Tone in this example is called an *occasion*

setter, and its effects can be shown to be generalized to modulate a CR to other CS's trained with other occasion setting stimuli and US's.

5. *Applications*

The principles of Pavlovian conditioning and operant behavior have been applied to solve numerous practical problems. Pavlovian extinction procedures have proven effective in the treatment of phobias; and clinics have found that conditioned nausea to hospitals after chemotherapy can be prevented by giving patients flavored candy to eat during administration, producing nausea to the flavor and not the hospital. Reinforcement procedures have similarly found many applications. The concept of conditioned reinforcement is behind token economies used in schools and correctional institutions; shaping procedures are being used in programmed instruction with computers; the matching principle has been applied to the reduction of maladaptive behavior in persons with mental retardation without the need for punishment; various principles have been applied to problems of treatment adherence in drug dependency programs; and principles of discrimination learning have been used to develop highly effective language training programs for children with autism. Our understanding of Pavlovian and operant conditioning mechanisms continues to provide simple yet sophisticated approaches to complex problems.

See also: Autonomic Classical and Operant Conditioning; Cardiovascular Conditioning: Neural Substrates; Classical Conditioning and Clinical Psychology; Classical Conditioning, Neural Basis of; Deinstitutionalization, Psychiatry of; Habituation; Operant Conditioning and Clinical Psychology; Pavlov, Ivan Petrovich (1849–1936); Skinner, Burrhus Frederick (1904–90).

Bibliography

Donahoe J W, Palmer D C 1994 *Learning and Complex Behavior*. Allyn & Bacon, Needham Heights, MA

Dragoi V, Staddon J E R 1999 The dynamics of operant conditioning. *Psychological Review* **106**: 20–61

Hineline P N 1981 The several roles of stimuli in negative reinforcement. In: Harzem P, Zeiler M D (eds.) *Predictability, Correlation, and Contiguity*. Wiley, New York, pp. 203–46

Rescorla R A 1988 Pavlovian conditioning: It's not what you think it is. *American Psychologist* **43**: 151–60

Staddon J E R 1983 *Adaptive Behavior and Learning*. Cambridge University Press, Cambridge, UK

Williams B A 1989 Reinforcement, choice, and response strength. In: Atkinson R C, Herrnstein R J, Lindzey G, Luce R D (eds.) *Stevens' Handbook of Experimental Psychology: Vol. 2. Learning and Cognition*. Wiley, New York, pp. 167–244

D. T. Cerutti

Conduct Disorder

Conduct disorder (CD) refers to antisocial behaviors in children and adolescents. These behaviors encompass a variety of acts that reflect social rule violations and actions against others such as fighting, lying, truancy, and stealing (Robins 1991, Kazdin 1995). Many of these behaviors are seen in most children over the course of development. CD is reserved for instances in which antisocial behaviors lead to impairment in everyday functioning, as reflected in unmanageability at home and at school or dangerous acts that affect others. This article discusses characteristics, onset, and course of CD and intervention strategies.

1. *Key Characteristics*

1.1 *Diagnosis and Prevalence*

CD is the psychiatric diagnostic category for antisocial behavior among children and adolescents (American Psychiatric Association 1994). Table 1 lists the main symptoms. A diagnosis of CD is provided if (a) the individual shows at least three symptoms; (b) within the past 12 months; (c) with at least one of the symptoms evident in the last 6 months.

The prevalence of a disorder refers to the percentage of cases in the population at a given point in time. Among community samples of school-age samples, the prevalence of CD is ~ 2–6 percent. The estimate is conservative in representing the scope of the problem. Children who approach but do not quite meet the

Table 1
Symptoms included in the diagnosis of conduct disorder

1	Bullying or threatening others
2	Fighting
3	Using a weapon that can cause serious physical harm to others
4	Being physically cruel to people
5	Being physically cruel to animals
6	Stealing and confronting a victim (e.g., mugging, purse snatching, extortion, armed robbery)
7	Forcing someone into sexual activity
8	Firesetting
9	Destroying property of others
10	Breaking into someone else's house, building, or car
11	Frequent lying or 'conning' others
12	Stealing without confronting a victim
13	Staying out late at night despite parental prohibitions
14	Running away from home
15	Being truant from school

Table 2
Long-term prognosis of youths identified as conduct disorder: overview of major characteristics likely to be evident in adulthood

Characteristic in adulthood[a]	Description
Psychiatric status	Greater psychiatric impairment, including antisocial personality, alcohol and drug abuse, and isolated symptoms (e.g., anxiety, somatic complaints); also, greater history of psychiatric hospitalization
Criminal behavior	Higher rates of driving while intoxicated, criminal behavior, arrest records, and conviction, and period of time spent in jail
Occupational adjustment	Less likely to be employed; shorter history of employment, lower status jobs, more frequent change of jobs, lower wages, and depend more frequently on financial assistance (welfare). Served less frequently and performed less well in the armed services
Educational attainment	Higher rates of dropping out of school, lower attainment among those who remain in school
Marital status	Higher rates of divorce, remarriage and separation
Social participation	Less contact with relatives, friends, and neighbors; little participation in organizations such as church
Physical health	Higher mortality rate; higher rate of hospitalization for physical (as well as psychiatric) problems

a These characteristics are based on comparisons of clinically referred children identified for CD relative to control clinical referrals or normal controls or from comparisons of delinquent and nondelinquent youths (Pepler and Rubin 1991, Peters et al. 1992)

criteria are likely to have significant impairment in everyday lives. Boys show ~ 3–4 times higher rates of CD than girls. The sex differences may be explained by differences in predispositions toward responding in aggressive ways. Age variations reveal interesting patterns in prevalence rates. Rates of CD tend to be higher for adolescents (~ 7 percent for youths aged 12–16 years) than for children (~ 4 percent for children aged 4–11 years).

1.2 Correlates and Associated Features

Children who meet diagnostic criteria for CD are likely to show a number of other problem behaviors (Robins and Rutter 1990, Pepler and Rubin 1991). They are likely to argue with adults, lose their temper, actively defy and refuse to comply with requests, deliberately annoy others, and are angry and resentful. A large percentage of children (e.g., 40–70 opening) diagnosed with CD also evince hyperactive behavior and meet criteria for attention-deficit/hyperactivity disorder (ADHD). The core symptoms of ADHD include inattention, impulsiveness, and hyperactivity.

Children with CD are also likely to show academic deficiencies. They are more likely to repeat a grade, to show lower achievement levels, and to end their schooling sooner than their peers. Such children often are socially ineffective in their interactions with adults (e.g., parents, teachers, community members) and engage in behaviors that promote deleterious interpersonal consequences such as peer rejection. In general, CD is associated with a wide variety of other child characteristics that reflect maladjustment in everyday life.

Several parent and family characteristics are associated with CD, including parent psychopathology and maladjustment, criminal behavior, and alcoholism. Parents of children with CD are likely to show especially harsh, lax, erratic, and inconsistent discipline practices. Within the families, there are often less supportive and more defensive communications among family members, unhappy marital relations, and interpersonal conflict. Untoward living conditions are also associated with CD including overcrowding, poor housing, and high-crime neighborhoods.

1.3 Onset and Course of Conduct Disorder

Research has identified subtypes in an effort to find meaningful ways of grouping various sets of symptoms and accounting for prevalence differences over the course of development. Two patterns are currently receiving major attention. *Child-onset CD* begins in early childhood and with such behaviors as stubbornness, noncompliance, and hyperactivity that can progress to conduct problems. Youths with child onset are more likely than those with adolescent onset to engage in aggressive and criminal behavior and are more likely to continue their dysfunction into adulthood. *Adolescent-onset CD* is more common than child onset. During adolescence, many youths engage in criminal behavior. For many of these youths, the acts are isolated and short-lived (Ketterlinus and Lamb 1994). Both child- and adolescent-onset CD youths engage in illegal behavior during adolescence. However, those with child onset are more likely also to engage in aggressive acts and to be represented primarily by boys.

Longitudinal studies show that CD in childhood predicts aggressive and antisocial behavior up to 30 years later. Among CD youths identified in childhood, slightly less than 50 percent continue their CD into adulthood. If all diagnoses are considered, rather than continuation of CD alone, slightly over 80 percent are likely to show a psychiatric disorder as adults. A broad set of characteristics CD youths are likely to show when they become adults are highlighted in Table 2. As the table indicates, individuals with a history of CD evince a broad range of untoward outcomes.

2. *Factors that Influence Onset of Conduct Disorder*

Risk factors refer to characteristics, events, or processes that increase the likelihood (risk) for the onset of a problem or dysfunction. Risk factors, as antecedents to the dysfunction, may provide clues as to development and progression of CD, possible mechanisms and processes through which the dysfunction comes about, and foci for possible intervention. Several factors that predispose children and adolescents to CD are highlighted in Table 3.

Even under very adverse conditions with multiple risk factors present, many individuals will not experience adverse outcomes. *Protective factors* refer to characteristics, events, or processes that decrease the impact of a risk factor and likelihood of an adverse outcome. Protective factors are identified by studying individuals known to be at risk show several risk factors and by delineating subgroups of those who do, versus those who do not, later show CD. Among a high risk sample, children are less likely to develop CD if they are first born, are perceived by their mothers as affectionate, show high self-esteem and locus of control, and have alternative caretakers in the family (than the parents) and a supportive same-sex model who played an important role in their development. Other factors that reduce or attenuate risk include above-average intelligence, competence in various skill areas, getting along with peers, and having friends.

It is not clear how risk and protective factors operate. As an exception, parenting practices (e.g., harsh punishment, attending to aggressive child behavior) have been well studied. These practices directly contribute to CD and altering these practices reduces aggressive and antisocial child behavior (Pepler and Rubin 1991, Patterson et al. 1992).

3. *Interventions*

3.1 *Treatment*

Conduct problem symptoms constitute the most frequent basis for which children and adolescents are referred for treatment. Consequently, there is a need for effective interventions. Many different treatments have been applied, including psychotherapy, pharmacotherapy (medications), home, school, and community-based programs, residential and hospital treatment, and assorted social services. Few treatments have been carefully evaluated in controlled studies and shown to reduce CD problems and to improve functioning of the child in everyday life (Kazdin 2000). Three of the more well studied treatments are highlighted here.

Parent management training refers to a treatment in which parents are trained to interact with the child in ways that promote prosocial behavior. The treatment focuses on the use of reinforcement (e.g., praise, tokens), mild punishment (e.g., brief time out, response cost), and related techniques (e.g., prompting, shaping) to develop child behavior. Extensive research has shown that many parent–child interaction patterns in the home unwittingly foster and escalate child aggressive behavior. Parent management training teaches skills to the parent, develops interactions between parents and the child that promote positive parent and child behavior, and in the process decreases aggressive and antisocial behavior.

Cognitive problem-solving skills training is based on research showing that conduct problem youths often show distortions in various cognitive processes (e.g., how individuals perceive, code, and experience the world, as reflected in beliefs, attributions, and expectations). A variety of cognitive processes pertain to interactions with others, including the ability to generate solutions to interpersonal problems and consequences of action (e.g., what would happen after a particular behavior). Problem-solving skills therapy develops skills in approaching interpersonal situations and teaches ways to identify prosocial or adaptive solutions and alternative consequences of actions. Children practice using the approach in treatment sessions and at home with their parents.

Multisystemic therapy focuses on the child behavior within the context of various systems (e.g., the family, peer group, schools) that may contribute to the child's problem behavior or could be used to help alter that behavior. A focus on the family as a system is designed to build better communication, to reduce negative interactions, and to improve the ability of the parents to function. Factors that can affect these interactions and the child's problems, such as stress that the parent experiences, marital conflict, association of the child with a deviant peer group, are focused on also. Many different techniques are applied to address these areas. Parent management training and problem-solving skills training, mentioned previously, are often incorporated into treatment.

3.2 *Prevention*

Prevention programs come in many different forms (McCord and Tremblay 1992, Mrazek and Haggerty

Table 3
Factors that place youths at risk for the onset of conduct disorder

Factor[a]	Description
Child factors:	
Child temperament	A more difficult child temperament (on a dimension of 'easy -to-difficult'), as characterized by more negative mood, lower levels of approach toward new stimuli, and less adaptability to change
Neuropsychological deficits and difficulties	Deficits in diverse functions related to language (e.g., verbal learning, verbal fluency, verbal IQ), memory, motor coordination, integration of auditory and visual cues, and 'executive' functions of the brain (e.g., abstract reasoning, concept formation, planning, control of attention)
Subclinical levels of CD	Early signs (e.g., elementary school) of mild ('subclinical') levels of unmanageability and aggression, especially with early age of onset, multiple types of antisocial behaviors, and multiple situations in which they are evident (e.g., at home, school, the community)
Academic and intellectual performance	Academic deficiencies and lower levels of intellectual functioning
Parent and family factors:	
Prenatal and perinatal complications	Pregnancy- and birth-related complications including maternal infection, prematurity and low birth weight, impaired respiration at birth, and minor birth injury
Psychopathology and criminal behavior in the family	Criminal behavior, antisocial personality disorder, and alcoholism of the parent
Parent–child punishment	Harsh (e.g., severe corporal punishment) and inconsistent punishment increase risk
Monitoring of the child	Poor supervision, lack of monitoring of whereabouts, and few rules about where youth can go and when they can return
Quality of the family relationships	Less parental acceptance of their children, less warmth, affection, and emotional support, and less attachment
Marital discord	Unhappy marital relationships, interpersonal conflict, and aggression of the parents
Family size	Larger family size, i.e., more children in the family
Sibling with antisocial behavior	Presence of a sibling, especially an older brother, with antisocial behavior
Socioeconomic disadvantage	Poverty, overcrowding, unemployment, receipt of social assistance ('welfare'), and poor housing
School-related factors:	
Characteristics of the setting	Attending schools where there is little emphasis on academic work, little teacher time spent on lessons, infrequent teacher use of praise and appreciation for school work, little emphasis on individual responsibility of the students, poor working conditions for pupils (e.g., furniture in poor repair), unavailability of the teacher to deal with children's problems, and low teacher expectancies

a The list of risk factors highlights major influences. The number of factors and the relations of specific factors to risk are more complex than the summary statements noted here (Robins and Rutter 1990, Peters et al. 1992, Kazdin 1995)

1994). Early intervention programs with the family have been effective in reducing conduct problems. High-risk families are identified, usually by such factors as low socioeconomic status, low educational attainment, and high-stress living conditions. Intervention programs sometimes begin before the child is

born to provide counseling regarding maternal care, to provide support in the home to reduce stress, and to prepare the parents for child-rearing demands. After the infant is born, the program may continue for a few years to help support parents, to develop cognitive skills of the child, and to enroll the child in a preschool program. Youths who have received such programs, as they become adolescents, show lower arrest rates, higher levels of educational attainment, and less substance use and abuse than other youths who did not receive early intervention.

Prevention programs are often conducted in the schools because there are opportunities to provide programs to youths in larger numbers, in the context of peers, and on a regular basis for extended periods. Programs often focus on developing positive skills and success experiences at home and at school. The reason for this focus is that bonding to deviant peers and poor connections with the family are risk factors for such behaviors as delinquency and substance abuse. Developing such success experiences in the schools among elementary school children has increased bonding to families and decreased rates of antisocial behavior and substance abuse. To date, the evidence shows that preventive interventions can have a significant impact on CD.

3.3 Social Policy

Social policy refers to governmental and legislative efforts to implement changes to benefit society or a particular segment of society and, in this sense, is a social intervention. There are many influences on the levels of aggression and violence that are society wide, depending on the country and culture. For example, the use of corporal punishment in child-rearing and school discipline, violence in the media (television and films), and social practices that permit or facilitate violence (e.g., availability of hand guns) influence aggressive and antisocial child behavior.

There have been many promising large-scale programs to curb influences on child conduct problems. For example, the use of corporal punishment (i.e., physical aggression against children) contributes to child aggression. Corporal punishment in child discipline at home and at school has been banned in a number of countries (e.g., Austria, Denmark, Finland, Norway, and Sweden) (Greven 1992). As another example, a large-scale effort was implemented in Norway to curb bullying among children. Parent, teachers, and peers were involved in a national effort that was effective in changing the ethos of interactions among youths. In the USA, preschool educational programs (Head Start) are designed to improve child development in disadvantaged families and to have a broad impact on child functioning. In general, social policy interventions are more difficult to implement and to evaluate than treatment and prevention. Policy practices are often subject to changes in laws, are often expensive, and require special monitoring procedures to ensure effective implementation (American Psychological Association 1993).

4. Conclusions

Conduct disorder in children and adolescents is a very active area of research. Efforts focus on understanding the emergence and course of disorder and the many child, parent, family, contextual, and societal influences that may operate. Research also focuses on interventions. Although treatment, prevention, and social policy changes were distinguished, the general view among professionals is that interventions at each of these levels are required in order to have a marked impact on the problem. For many individuals, severe antisocial behavior and associated dysfunction in multiple spheres represent a life-long pattern. Because the disorder is often passed from parents to children, efforts to understand and intervene are critically important. Many questions remain about CD and its course. Even so, much of the available knowledge can have an impact on the problem.

See also: Adolescence, Psychiatry of; Adolescent Development, Theories of; Adolescent Vulnerability and Psychological Interventions; Antisocial Behavior in Childhood and Adolescence; Behavior Therapy with Children; Social Competence: Childhood and Adolescence

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*, 4th edn. American Psychiatric Association, Washington, DC

American Psychological Association, Commission on Violence and Youth 1993 *Violence and Youth: Psychology's Response*. American Psychological Association, Washington, DC, Vol. 1

Greven P 1992 Exploring the effects of corporal punishment. *Child, Youth, and Family Services Quarterly* **15**(4): 4–5

Kazdin A E 1995 *Conduct Disorder in Childhood and Adolescence*, 2nd edn. Sage, Thousand Oaks, CA

Kazdin A E 2000 *Psychotherapy for Children and Adolescents: Directions for Research and Practice*. Oxford University Press, New York

Ketterlinus R D, Lamb M E (eds.) 1994 *Adolescent Problem Behaviors: Issues and Research*. Erlbaum, Hillsdale, NJ

McCord J, Tremblay R E (eds.) 1992 *Preventing Antisocial Behavior*. Guilford Press, New York

Mrazek P J, Haggerty R J (eds.) 1994 *Reducing Risks for Mental Disorders: Frontiers of Preventive Intervention Research*. National Academy Press, Washington, DC

Patterson G R, Reid J B, Dishion T J 1992 *Antisocial Boys*. Castalia, Eugene, OR

Pepler D J, Rubin K H (eds.) 1991 *The Development and Treatment of Childhood Aggression*. Erlbaum, Hillsdale, NJ

Peters R D, McMahon R J, Quinsey V L (eds.) 1992 *Aggression and Violence Throughout the Life Span*. Sage, Newbury Park, CA
Robins L N 1991 Conduct disorder. *Journal of Child Psychology and Psychiatry* **32**: 193–212
Robins L N, Rutter M (eds.) 1990 *Straight and Devious Pathways from Childhood to Adulthood*. Cambridge University Press, Cambridge, UK

A. E. Kazdin

Confessions: Psychological and Forensic Aspects

A confession is any written or oral statement in which a person admits to having committed some transgression, often indicating acknowledgement of guilt for a crime. In some settings, confessions are considered necessary for absolution, social acceptance, freedom, or physical and mental health, making it easy to understand why people often exhibit an 'urge to confess.' However, in other settings confessions predictably result in personally damaging consequences to the confessor—such as a loss of money, liberty, or life itself—making it difficult to understand this aspect of human behavior.

Confessions have played a multifaceted and powerful role throughout history. There are three venues of human social interaction in which one person's confession to another has proved important: religion, psychotherapy, and criminal justice. In religion, the scene of the penitent with the Catholic priest, occurring inside a small, private, and hallowed stall known as a confessional, serves to remind us that all of the world's major religions advise or oblige adherents to confess to their transgressions as a means of moral cleansing. In psychotherapy, the image of the emotionally distressed patient laying on a couch, in tears, while self-disclosing deep, dark, personal secrets to his or her analyst, illustrates the widely held belief in the healing power of 'opening up' the past—including memories of one's actual or imagined misdeeds. In criminal justice, of course, the classic image of the beleaguered suspect being grilled by police behind the locked door and under the bright light of the interrogation room, serves as a stark reminder that, in the law, the self-incriminating statement is considered the most potent evidence of guilt.

1. *Confession in Religion*

All major religions of the world—Buddhism, Christianity, Hinduism, Islam, and Judaism—provide a mechanism and encouragement for followers to acknowledge and disclose their transgressions. The purposes served by such confessions are twofold: to cleanse the individual's soul and to police the community, thus serving as a deterrent to wrongdoing.

Religions vary in terms of how, when, where, and to whom confessions are given, and even whether they are made in private or in public. Profound differences exist even within Christianity. For example, the Quakers, Unitarians, and others are encouraged to confess their sins to themselves, through private prayer. Others, such as Catholics and the Greek and Russian Orthodox have more formal rituals whereby followers confess, often at a designated time or place, to ministers or priests. The adoption of this model was particularly explicit in the year 1215 when the Roman Catholic Church, in the Fourth Lateran Council, made the rite of an annual confession obligatory for all adherents. In other religions, the confession to be given depends on the nature of the misdeed. Among American Southern Baptists, for example, people are required to disclose their sins to those they have specifically harmed—such as a spouse, an employer, or the entire congregation.

2. *Confession in Psychotherapy*

People in many parts of the world have long believed that confession is good not only for the soul but also for the body and the mind. Several years ago, anthropologist Weston La Barre (1964) found that many Native Americans of North and South America believed that physical and mental health required purity, which in turn required exposure—often through elaborate confession ceremonies involving shamans and witch doctors. Similar notions have permeated Western medicine ever since Josef Breuer and Sigmund Freud (1895) observed that patients often felt better after purging their minds of material buried in the unconscious. This discovery gave birth to Freud's *Psychoanalysis*, the first systematic form of 'talking cure,' and now forms the basis for many modern psychotherapies and social support groups.

Recent research confirms the healing power of opening up to someone else about one's problems, transgressions, victimizations, and other unpleasant truths. In a series of controlled experiments, James Pennebaker (1997) and others brought research subjects into a laboratory and asked them to talk into a tape recorder or write for 20 minutes either about past traumas or about trivial daily events. While speaking or writing, the subjects were upset and physiologically aroused. Many tearfully recounted accidents, failures, personal wrongdoings, instances of physical or sexual abuse, and other issues. Soon, however, the subjects felt better than ever. Thus, while systolic blood pressure levels rose during the disclosures, they later dipped below pre-experiment levels. The subjects even exhibited a decline in their number of doctor visits over the next six months.

Other studies, too, have shown that keeping confessional secrets can be stressful and that 'letting it out' and 'getting it off your chest' can have therapeutic effects on health—effects that are especially strong when the events being described are highly traumatic (Smyth 1998). For example, in a study of women who had an abortion, those who talked about it to an experimenter—compared to those who did not talk about it—were later less haunted by intrusive thoughts about the experience (Major and Gramzow 1999). In another study, researchers identified 80 gay men newly infected with the HIV virus but who had no symptoms, administered various psychological tests, and monitored their progress every six months for nine years. They found that in men who were partly 'in the closet'—compared with those who had openly confessed rather than concealed their homosexuality—the infection spread more rapidly, causing them to die sooner (Cole et al. 1996). This provocative correlation does not prove that 'coming out' is healthier than 'staying in.' In a controlled laboratory experiment, however, subjects who were told to suppress rather than express turbulent emotional thoughts exhibited a temporary decrease in the activity of certain immune cells (Petrie et al. 1998).

3. *Confessions in Law*

In criminal law, confession evidence is a prosecutor's most potent weapon—so potent, as one legal scholar put it, that 'the introduction of a confession makes the other aspects of a trial in court superfluous' (McCormick 1972). On the one hand, confessions play a vital role in law enforcement and crime control. On the other hand, they serve as a source of recurring controversy, with questions often arising about whether a statement is authentic, voluntarily given, the product of a competent waiver of rights, and in accord with the law. For all of these reasons, confessions have been described as 'troubling' (Brooks 2000).

3.1 *Rules of Confession Evidence and Procedure*

To guard the integrity of the criminal justice system, to protect citizens against violations of due process, and to minimize the risk that innocent people would confess to crimes they did not commit, American courts have erected guidelines for the admission of confession evidence. According to the historical overview proposed by Wigmore (1970), the modern law's treatment of confession evidence has evolved through a series of stages. In England, during the sixteenth and seventeenth centuries, there was no restriction on the use of confessions. All avowals of guilt were accepted at face value. Thus, at least through the middle of the seventeenth century, physical torture was used to extract self-incriminating statements. In contrast, by the nineteenth century, the judiciary was highly cynical of confessions and quick to repudiate them for a lack of reliability. Currently, as in the twentieth century, confessions are not accepted or rejected outright but, rather, each is judged according to a 'totality of the circumstances' and the requirement that it be considered voluntary. Thus, confessions are typically excluded if elicited by brute force, prolonged isolation, a deprivation of food or sleep, threats of punishment or harm, promises of immunity or leniency, or without notifying a suspect of his or her rights (Grano 1993 Kamisar et al. 1999).

In cases involving a disputed confession, a preliminary hearing is typically held in which a judge determines whether a confession was voluntary and hence admissible as evidence. In American courts, the trial judge will then either admit confessions deemed voluntary without special instruction, or he or she may specifically direct the jury to make an independent judgment of voluntariness and disregard statements found to be coerced.

3.2 *Methods of Police Interrogation*

Many years ago, interrogators used bright lights, brute force, the rubber hose, and physical intimidation to get confessions. Today, the police are required to warn suspects of their 'Miranda' rights to silence and to an attorney (Leo and Thomas 1998)—and the use of physical force has given way to psychologically oriented techniques such as feigned sympathy and friendship, appeals to conscience and religion, the presentation of false evidence, the alternation of good cop and bad cop, and other forms of trickery and deception. Today, a number of manuals are available for training the police in how to elicit confessions through interrogation. The most popular of the manuals is Inbau et al. (2001) *Criminal Interrogation and Confessions*, first published in 1962 and now in its fourth edition.

Inbau et al. (2001) advise the police to put suspects into a small, barely furnished, soundproof room—a physical environment designed to promote feelings of social isolation, helplessness, and discomfort. They then present a nine-step procedure for overcoming the suspect's resistance. According to this method, the interrogator begins by confronting the suspect with his or her guilt (Step 1), then develops psychological 'themes' that justify or excuse the crime (Step 2), interrupts all statements of denial (Step 3), overcomes the suspect's factual, moral, and emotional objections to the charges (Step 4), ensures that the increasingly passive suspect does not mentally withdraw (Step 5), shows sympathy and understanding, and urges the suspect to tell the truth (Step 6), offers the suspect a face-saving alternative explanation for his or her guilty action (Step 7), gets the suspect to recount the details of the crime (Step 8), and converts that statement into a full written confession (Step 9).

Kassin and McNall (1991) note that there are two general approaches contained within this method of interrogation. One is to pressure the suspect into submission by expressing certainty in his or her guilt—a goal that, at times, leads interrogators to falsely state or imply that they have damaging evidence such as fingerprints or an eyewitness. In this way, the accused is led to believe that it is futile to mount a defense. The second approach is to befriend and sympathize with the suspect and 'minimize' the offense by offering face-saving excuses—such as suggesting that the crime was provoked by the victim, or was accidental. This technique may lead a suspect to anticipate leniency, which provides an inducement to confess. Research shows that these tactics are often used (Gudjonsson 1992, Kassin 1997). In an observational study of 182 interrogations, for example, Leo (1996) found that detectives used an average of five to six tactics per suspect.

3.3 The Psychology of False Confessions

As prescribed, police interrogations should lead guilty suspects to incriminate themselves by reducing the perceived negative consequences of confessing while, at the same time, increasing the anxiety associated with continued deception. It could be argued that interrogation is not overly coercive in that innocent people do not confess to crimes they did not commit. This assumption is incorrect. Nobody knows the incidence of false confessions or has devised an adequate method of calculating their prevalence. The pages of history and current events make it clear, however, that there are many documented false confessions on the record (Borchard 1932, Gudjonsson 1992, Leo and Ofshe 1998, Radelet et al. 1992, Wrightsman and Kassin 1993). In fact, among the first 62 prisoners to be exonerated by DNA evidence, 15 had given full or partial false confessions (Scheck et al. 2000).

Regardless of the precise numbers of such cases, Kassin and Wrightsman (1985) identified three different types of false confession: voluntary, coerced-compliant, and coerced-internalized. This categorization scheme has provided a useful framework for the study of false confessions and has been adopted by other researchers in the field.

A 'voluntary false confession' is a self-incriminating statement that is offered without external pressure from the police. When Charles Lindbergh's baby was kidnapped in 1932, 200 people confessed. Why would anyone volunteer a false confession? There are several possible motives. Sometimes the goal is to protect a friend or relative—a problem often revealed in interviews with juvenile offenders. Other possible motives include a pathological need for fame, acceptance, recognition, or self-punishment.

In contrast to voluntary false confessions are those in which suspects confess as a result of intense interrogation pressures. Within the category of coerced false confessions, it is necessary to further distinguish between two forms of social influence: compliance and internalization. In the common language of social psychology, compliance refers to a change in one's public behavior for instrumental purposes. This form of influence was observed in early studies of conformity by Asch (1956) and research on obedience to authority by Milgram (1974). In contrast, internalization refers to a private acceptance of the beliefs espoused by others. This deeper form of influence was exhibited in early autokinetic studies on the formation of group norms (Sherif 1936) and later research on minority influence (Moscovici 1985).

'Coerced-compliant false confessions' occur when a suspect confesses to escape or avoid an aversive interrogation, or to gain a promised or implied reward. In these cases, the confession is an act of decision-making compliance by a suspect who privately knows that he or she is truly innocent. Psychologically, coerced-compliant false confessions are easy to understand, as they arise whenever a suspect comes to believe that the short-term benefits of confessing—such as being left alone, fed, or released—outweigh the long-term costs associated with prosecution, the loss of reputation, and incarceration (Hilgendorf and Irving 1981). Individuals who are characteristically predisposed to exhibit compliance in social influence situations are particularly vulnerable in this regard (Gudjonsson 1989). The pages of legal history are filled with documented false confessions of this nature. In one instance, four young men confessed independently to the massacre of six Buddhist monks in a Phoenix Arizona temple. In that case, a psychiatric inpatient called the police, volunteered a confession, and implicated three other men. His 'accomplices' were then seized and interrogated for several hours through the night. All eventually confessed and spent 70 days in jail—until the real murderers were discovered via ballistics evidence (Parloff 1993).

'Coerced-internalized false confessions' are those in which an innocent person—anxious, sleep-deprived, confused, and subjected to a highly suggestive interrogation that often includes the presentation of false evidence—actually comes to believe that he or she committed the crime. This type of false confession is particularly frightening because the suspect's memory may be altered, rendering its original contents potentially irretrievable. This phenomenon is closely related to recent laboratory and field studies involving the creation of false memories for words in a list, visual images, and past experiences (Garry et al. 1996, Hyman et al. 1995, Loftus 1993, Roediger and McDermott 1995).

Several cases of this sort have appeared in recent years. Consider, for example, the story of 18-year-old Peter Reilly, who returned home one night to find that his mother had been murdered. Immediately he called the police. Despite Reilly's distraught mental and

emotional state, the police interrogated him for 24 hours, during which time they administered a polygraph, which they described as infallible—and which, they claimed, he failed, thus indicating that he murdered his mother despite having no memory of having done so. Transcripts of the interrogation session revealed that Reilly underwent a stark transformation from denial through confusion, self-doubt, conversion ('well, it really looks like I did it'), and the signing of a full written confession. Two years later, independent evidence revealed that Reilly could not have committed the murder, that the confession he came to believe was false (Barthel 1976). Over the years, a number of other such cases have surfaced (Gudjonsson 1992, Kassin 1997, Leo and Ofshe 1998).

Is it really possible to convince people that they are guilty of an act they did not commit? Based on an analysis of past cases, Kassin and Kiechel (1996) theorized that two factors increase the risk: (a) a suspect who lacks a clear memory of the event in question, or whose memory is vulnerable to manipulation, and (b) the presentation of false evidence. To test this hypothesis, they recruited pairs of college students to work on a fast- or slow-paced computer task. At one point, the computer crashed, and students were accused of having caused the damage by pressing a key they had been specifically instructed to avoid. All students were truly innocent and denied the charge. In half the sessions, however, the second student (a confederate) said that she had seen the student hit the forbidden key. Demonstrating the process of compliance, many of the subjects confronted with this false witness signed a confession handwritten by the experimenter. Next, indicating internalization, some subjects later 'admitted' their guilt to a stranger (also a confederate) after the experiment was supposedly over and the two were alone. In short, innocent people made vulnerable to suggestion can be induced to confess and to internalize guilt by the presentation of false evidence—an interrogation tactic commonly used by the police.

3.4 Confessions and the Jury

Legal scholars have long speculated that confessions are the most persuasive form of evidence against a criminal defendant. Research shows that this assumption is correct. In a series of mock jury studies, Kassin and Neumann (1997) inserted a confession, eyewitness identification, or character testimony into various cases otherwise containing only weak circumstantial evidence. Consistently, they found that confessions raised the conviction rate more than did the other common and potent forms of evidence. Whenever a suspect confesses during a coercive interrogation but then withdraws the statement, claiming it was false, the jury is confronted with an attributional dilemma, needing to determine whether the defendant had confessed because he or she was truly guilty (a personal attribution) or because he or she merely sought to avoid the aversive consequences of silence (a situational attribution). According to the logic of attribution theory, jurors should lack certainty about confessions made in response to external pressure. But over the years, studies have shown that people, at least in Western cultures, frequently fall prey to the 'fundamental attribution error'—that is, they tend to make dispositional attributions for another person's actions, while neglecting the role of situational factors (Gilbert and Malone 1995, Jones 1990).

Is it similarly possible that jurors view suspects who confess as guilty, even if they were highly pressured by an interrogator? To examine this question, Kassin and Sukel (1997) had mock jurors read one of three versions of a murder trial. In a control version that did not contain a confession, only 19 percent voted guilty. In a low-pressure version in which the defendant was said to have confessed immediately upon questioning, the conviction rate rose considerably, to 62 percent. But there was a third, high-pressure condition in which subjects were told that the defendant had confessed out of fear and with his hands cuffed painfully behind his back. How did subjects in this situation react? Reasonably, they judged the confession to be coerced, and they said it did not influence their verdicts. Yet the conviction rate in this situation significantly increased, this time to 50 percent. In short, people were powerfully influenced by evidence of a confession—even while conceding that this confession was coerced.

The jury's reaction to confession evidence also depends on how that evidence is presented. Today, it is mandatory in the UK and common in some American police departments to videotape confessions for presentation in court. But how are these events staged for the camera? Attribution research has revealed that observers who watch two people engage in a conversation overemphasize the influence on that interaction of the person who is visually salient (Taylor and Fiske 1978). The same is true of interrogator–suspect interactions. Lassiter et al. (2001) reviewed a series of studies in which researchers taped mock confessions from three different camera angles so that either the suspect or the interrogator or both were visible. All subjects heard the same exchanges of words, but those who focused on the suspect saw the situation as less coercive—and the suspect as more guilty—than did those focused on the interrogator. When the camera directs all eyes on the accused, jurors are thus likely to underestimate the amount of pressure exerted by the 'hidden' interrogator.

3.5 Psychologists as Confession Experts

The problems raised by confession evidence are fundamental to criminal justice, substantial, and

important. Thus, there is a great deal about this aspect of the legal system that warrants the intervention of psychologists for educational and research purposes. It is important to note that intervention in this regard may take several forms, including but not limited to in-court testimony from 'confession experts.' In recent years, a growing number of psychologists have testified, often on behalf of criminal defendants who confessed but then retracted their statements and went to trial. Typically, these experts testify in general terms about the social influence processes of compliance, persuasion, and obedience to authority; suggestibility effects in memory; personality and psychopathology; and other relevant phenomena; in other cases, they offer opinions concerning a specific defendant based on interviews and clinical test results. As both the law and empirical foundations for such expert testimony evolve, it remains to be seen what impact psychological research on confessions will have on police, judges, juries, defendants, and others in criminal justice.

See also: Criminal Justice, Ethics of: Aspects of Human Dignity; Criminal Justice, Sociology of; Deception, Psychology of; Expert Testimony; Forensic Psychiatry; Forensic Psychology; Guilt; Juries; Law: Defense of Insanity

Bibliography

Asch S E 1956 Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs* **70**: 416

Barthel J 1976 *A Death in Canaan*. Dutton, NY

Borchard E M 1932 *Convicting the Innocent: Errors of Criminal Justice*. Yale University Press, New Haven, CT

Breuer J, Freud S 1895 Studies on hysteria. In: Strachey J (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London (Reprinted 1955)

Brooks P 2000 *Troubling Confessions*. University of Chicago Press, Chicago

Cole S W, Kemeny M E, Taylor S E, Visscher B R, Fahey J L 1996 Accelerated course of human immunodeficiency virus infection in gay men who conceal their homosexual identity. *Psychosomatic Medicine* **58**: 219–31

Garry M, Manning C G, Loftus E F, Sherman S J 1996 Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin and Review* **3**: 208–14

Gilbert D T, Malone P S 1995 The correspondence bias. *Psychological Bulletin* **117**: 21–38

Grano J D 1993 *Confessions, Truth, and the Law*. University of Michigan Press, Ann Arbor, MI

Gudjonsson G H 1989 Compliance in an interrogation situation: A new scale. *Personality and Individual Differences* **10**: 535–40

Gudjonsson G H 1992 *The Psychology of Interrogations, Confessions, and Testimony*. Wiley, London

Hilgendorf E L, Irving M 1981 A decision-making model of confessions. In: Lloyd-Bostock M (ed.) *Psychology in Legal Contexts: Applications and Limitations*. MacMillan, London, pp. 67–84

Hyman I E, Husband T H, Billings J F 1995 False memories of childhood experiences. *Applied Cognitive Psychology* **9**: 181–97

Inbau F E, Reid J E, Buckley J P, Jayne B C 2001 *Criminal Interrogation and Confessions*, 4th edn. Williams and Wilkins, Baltimore, MD

Jones E E 1990 *Interpersonal Perception*. Freeman, NY

Kamisar Y, LaFave W, Israel J 1999 *Modern Criminal Procedure: Cases, Comments and Questions*, 9th edn. West, St. Paul, MN

Kassin S M 1997 The psychology of confession evidence. *American Psychologist* **52**: 221–33

Kassin S M, Kiechel K L 1996 The social psychology of false confessions: Compliance, internalization, and confabulation. *Psychological Science* **7**: 125–8

Kassin S M, McNall K 1991 Police interrogations and confessions: Communicating promises and threats by pragmatic implication. *Law and Human Behavior* **15**: 233–51

Kassin S M, Neumann K 1997 On the power of confession evidence: An experimental test of the 'fundamental difference' hypothesis. *Law and Human Behavior* **21**: 469–84

Kassin S M, Sukel H 1997 Coerced confessions and the jury: An experimental test of the 'harmless error' rule. *Law and Human Behavior* **21**: 27–46

Kassin S M, Wrightsman L S 1985 Confession evidence. In: Kassin S, Wrightsman L (eds.) *The Psychology of Evidence and Trial Procedure*. Sage, Beverly Hills, CA, pp. 67–94

La Barre W 1964 Confession as cathartic therapy in American Indian tribes. In: Kiev A (ed.) *Magic, Faith, and Healing: Studies in Primitive Psychiatry Today*. Free Press, New York

Lassiter G D, Geers A L, Munhall P J, Handley I M, Beers M J 2001 Videotaped confessions: Is guilt in the eye of the camera? *Advances in Experimental Social Psychology* **33**

Leo R A 1996 Inside the interrogation room. *The Journal of Criminal Law and Criminology*, **86**: 266–303

Leo R A, Ofshe R J 1998 The consequences of false confessions: Deprivations of liberty and miscarriages of justice in the age of psychological interrogation. *The Journal of Criminal Law and Criminology*, **88**: 429–96

Leo R A, Thomas G C III (eds.) 1998 *The Miranda Debate: Law, Justice, and Policing*. Northeastern University Press, Boston

Loftus E F 1993 The reality of repressed memories. *American Psychologist* **48**: 518–37

Major B, Gramzow R H 1999 Abortion as stigma: Cognitive and emotional implications of concealment. *Journal of Personality and Social Psychology*, **77**: 735–45

McCormick C T 1972 *Handbook of the Law of Evidence*, 2nd edn. West, St. Paul, MN

Milgram S 1974 *Obedience to Authority: An Experimental View*. Harper and Row, NY

Moscovici S 1985 Social influence and conformity. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*, 3rd edn. Random House, New York, pp. 347–412

Parloff R 1993 False confessions: Standard interrogations by Arizona law enforcement officials led to four matching confessions to the murders of nine people at a Buddhist temple. *American Lawyer*, **May**: 58–62

Pennebaker J W 1997 *Opening up: The Healing Power of Expressing Emotions*. Guilford Press, New York

Petrie K J, Booth R J, Pennebaker J W 1998 The immunological effects of thought suppression. *Journal of Personality and Social Psychology* **75**: 1264–72

Radelet M L, Bedau H A, Putnam C E 1992 *In Spite of Innocence: Erroneous Convictions in Capital Cases*. Northeastern University Press, Boston
Roediger H L, McDermott K B 1995 Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 803–14
Scheck B, Neufeld P, Dwyer J 2000 *Actual Innocence: Five Days to Execution and Other Dispatches from the Wrongly Convicted*. Doubleday, New York
Sherif M 1936 *The Psychology of Social Norms*. Harper, New York
Smyth J M 1998 Written emotional expression: Effect sizes, outcome types, and moderating variables. *Journal of Consulting and Clinical Psychology* **66**: 174–84
Taylor S E, Fiske S T 1978 Salience, attention, and attribution: Top of the head phenomena. *Advances in Experimental Social Psychology* **11**: 249–88
Wigmore J H 1970 *Evidence* (revised by J. H. Chadbourn). Little, Brown, Boston, Vol. 3
Wrightsman L S, Kassin S M 1993 *Confessions in the Courtroom*. Sage Publications, Newbury Park, CA

S. Kassin

Confidentiality and Statistical Disclosure Limitation

1. *Data Quality Rests on Usefulness and Confidentiality Protection*

We capture enormous numbers of rich personal and proprietary records, store them in data warehouses measured in terabytes, analyze them using sophisticated computers and algorithms, and disseminate them worldwide via telecommunications channels that are cheap and fast. This explosive progress in information technology—look at the World Wide Web—heightens the tension between confidentiality and data access. Accepted principles of information ethics (Duncan et al. 1993) require that promises of confidentiality be preserved and that citizens have substantial access to information. Policies and procedures are needed to reconcile the data provider's interest in confidentiality and the data user's demand for data (Dalenius 1988). Government statistical agencies, such as the US Census Bureau, play a particularly significant and delicate brokering role in this reconciliation of competing demands. They often have legal authority to compel individuals and firms to provide data, and the data are both sensitive—like personal income—and so comprehensive and accurate to be in substantial demand by a broad range of data users. Academic researchers, corporate planners, journalists, and government analysts all seek such data. Other organizations, such as health care providers, necessarily collect data in treating patients that is highly personal. These data are of substantial research value for issues spanning medical quality assurance, cost containment, and access to medical treatment by the poor (Duncan and Kaufman 1996).

A compromise of the confidentiality pledge could harm the agency, the respondent, the public, or some other group. A statistical disclosure occurs when the data dissemination allows data snoopers to gain excessive new information about respondents whereby the snooper can isolate individual respondents and corresponding sensitive attribute values. Miller (1971, p. 176) illustrates the problem.

> Some deficiencies inevitably crop up even in the Census Bureau. In 1963, for example, it reportedly provided the American Medical Association with a statistical list of one hundred and eighty-eight doctors living in Illinois. The list was broken down into more than two dozen categories, and each category was further subdivided by medical specialty and area residence; as a result, identification of individual doctors was possible…

As shown in Fig. 1, the process of assuring confidentiality through statistical disclosure limitation has the following components:

(a) a data quality audit that, beginning with the original collected data, evaluates data utility and assesses disclosure risk;

(b) a determination of adequacy of confidentiality protection;

(c) if confidentiality protection is inadequate, the implementation of a disclosure limitation procedure; and

(d) a return to the data quality audit.

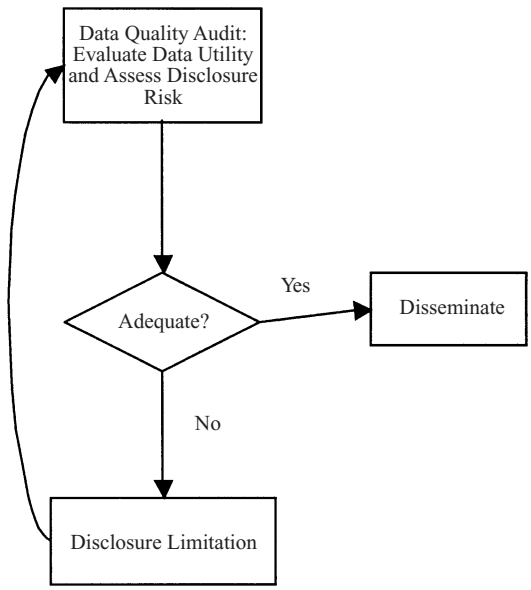

Figure 1
Statistical disclosure limitation process

2. Data Quality Audit: Data Utility and Disclosure Risk

A statistical agency audits the collected data, both to evaluate the utility of the data and to assess disclosure risk. Typically, with good survey design and implementation, the data utility is high. But, also, the risk of disclosure through the release of the original, collected data is too high, even when the data collected have been deidentified, i.e., apparent identifiers (name, e-mail address, phone number, etc.) have been removed. Reidentification techniques have become too sophisticated to assure confidentiality protection (Winkler 1998). A confidentiality audit will include identification of (a) sensitive objects, and (b) characteristics of the data that make it susceptible to attack.

Based on its confidentiality pledges, a statistical agency seeks to protect certain sensitive objects. An agency seeks to protect the instantiation of sensitive objects in the data from a data snooper. Sensitive objects can be the instantiations of a variety of variables associated with a subject entity (person, household, enterprise, etc.). Examples include the values of numerical variables, such as household income, a realization of an x-ray of a patient's lung, and a specific patent application. A second class of sensitive objects is instances of relationships, whether social or mathematical. An example of the former is 'Who is friends with whom?' in a social network; an example of the latter is an Internal Revenue Service formula to trigger a tax audit.

Collected data having particular characteristics pose substantial risk of disclosure. Characteristics that suggest vulnerability include:

(a) geographical detail (Greenberg and Zayatz 1992);

(b) longitudinal or panel structure;

(c) outliers;

(d) many attribute variables;

(e) population data, as in a census, rather than a survey with small sampling fraction; and

(f) existence of databases that are publicly available, identified, and share individual respondents and attribute variables with the subject data.

Data with geographical detail, such as census tract data, may be easily linked to known characteristics of respondents. Concern for this leads statistical agencies to place minimum population levels for geographical identifiers that they will release. Longitudinal data, which tracks entities over time, also poses substantial disclosure risk. Many electrical engineers lived in the Chicago area in 1998 and many electrical engineers lived in the Phoenix area in 1999, but few did both. Outliers, especially on variables like net worth, can easily lead to identifiable respondents. Data with many attribute variables allows easier linkage with known attributes of identified entities, and entities, which are unique in the sample, are more likely to be unique in the population. Further, population data, i.e., a census or near census, pose more disclosure risk than data that arise from a survey having a small sampling fraction. Finally, special concern must be shown when other databases are available to the data snooper and these databases are both identified and share with the subject data both individual respondents and certain attribute variables. Record linkage may then be possible between the subject data and the external database. The shared attribute variables provide the key.

3. Types of Statistical Disclosure

The legitimate objects of inquiry for statistical research are aggregates over individual records, for example, the median annual income of Hispanic social scientists in the United States. The statistical agency seeks to provide users with data that will allow accurate inference about such population characteristics. At the same time, because of confidentiality the statistical agency seeks to thwart the data snooper who might seek to use the disseminated data to draw accurate inferences about, say, the income of a particular Hispanic sociologist who now works in New York City. This capability by a data snooper represents a statistical disclosure.

There are two major types of disclosure—identity disclosure and attribute disclosure. *Identity disclosure* occurs with the association of a respondent's identity and a disseminated data record (Spruill 1983). *Attribute disclosure* occurs with the association of either an attribute value in the disseminated data or an estimated attribute value based on the disseminated data with the respondent (Duncan and Lambert 1989, Lambert 1993). In the case of identity disclosure, the association is assumed exact. In the case of attribute disclosure, the association can be approximate. Most statistical agencies place emphasis on limiting the risk of identity disclosure, perhaps because of its substantial equivalence to the inadvertent release of an identified record—a clear administrative slip up. On the other hand, an attribute disclosure, even though it invades the privacy of a respondent, may not be so easily traceable to actions of the agency.

4. Measures of Disclosure Risk

In the context of identity disclosure, disclosure risk can arise because a data snooper may be able to use the disseminated data product to re-identify some de-identified records. Spruill (1983) proposed a measure of disclosure risk for microdata: For each 'test' record in the masked file, compute the Euclidean distance between the test record and each record in the source file. Determine the percentage of test records that are closer to their parent source record than to any other source record. She defines the risk of disclosure to be the percentage of test records that match the correct

parent record multiplied by the sampling fraction (fraction of source records released).

More generally, and consistent with Duncan and Lambert (1986, 1989), the agency will have succeeded in protecting the confidentiality of a released data product if the data snooper remains sufficiently uncertain about a protected target value after data release. From this perspective, a measure of disclosure risk is built on measures of uncertainty. Further, the agency may model the decision making of the data snooper as a basis for using disclosure limitation to deter inferences about a target. Data snoopers are deterred from publicly making inferences about a target when their uncertainty is sufficiently high. Mathematically, uncertainty functions provide a workable framework for this analysis. Examples include Shannon entropy, which has found use in categorizing continuous microdata and coarsening of categorical data (Willenborg and de Waal 1996, p. 138).

Generally, a data snooper has *a priori* knowledge about a target, often in the form of a database with identified records (Adam and Wortmann 1989). Certain variables may be in common with the subject database. These variables are called *key* or *identifying*. When a single record matches on the key variables, the data snooper has a candidate record for identification. This candidacy is promoted to an actual identification if the data snooper is convinced that the individual is in the target database. This would be the case either if the data snooper has auxiliary information to that effect or if the data snooper is convinced that the individual is unique in the population. The data snooper may find that according to certain key variables that a sample record is unique. The question then is whether the individual is also unique on these key variables in the population. Bethlehem et al. (1990) have examined detection of records agreeing on simple combinations of keys based on discrete variables in the files. Record linkage methodologies have been examined by Fuller (1993) and Winkler (1998).

5. *Restricted Data Through Statistical Disclosure Limitation*

Direct transformations of data for confidentiality purposes are called *disclosure limiting masks* (Jabine 1993). With masked data sets, there is a specific functional relationship, possibly as a function of multiple records and possibly as a stochastic function, between masked values and the original data. Because of this relationship, the possibilities of both identity and attribute disclosures continue to exist, even though the risk of disclosure may be substantially reduced. The idea is to provide a response that, while useful for statistical analysis purposes, has sufficiently low disclosure risk. As a general classification, disclosure-limiting masks can be categorized as suppressions, recodings, or samplings.

5.1 *Suppression*

A suppression is a refusal to provide a data instance. For microdata, this can involve the deletion of all values of some particularly sensitive variable. In principle, certain record values could also be suppressed, but this is usually handled through recoding. For tabular data, the values of table cells that pose confidentiality problems are suppressed. These are the primary suppressions. Often, a cell is considered unsafe for publication according to the (n, p) dominance rule, i.e., if a few (n), say 3, contributing entities represent a percentage p, say 70 percent, or more of the total. Additionally, enough other cells are suppressed so that the values of the primary suppressions cannot be inferred from released table margins. These additional cells are called secondary suppressions. Tables of realistic dimensionality even with only a few primary suppressions present a multitude of possible configurations for the secondary cell suppressions. This raises computational difficulties that can be formulated as combinatorial optimization problems. Typical techniques that are employed include mathematical programming (especially integer programming), and graph theory (Chowdhury et al. 1999).

5.2 *Recoding*

A disclosure-limiting mask for recoding creates a set of data for which some or all of the attribute values have been altered. Recoding can be applied to microdata or to tabular data.

Examples of recoding as applied to microdata include data swapping, adding noise, and global recoding and local suppression. In data swapping (Dalenius and Reiss 1982, Spruill 1983), some fields of a record are swapped with the corresponding fields in another record. Concerns have been raised that while data swapping lowers disclosure risk it may excessively distort the statistical structure of the original data (Adam and Wortmann 1989). A combination of data swapping with additive noise has been suggested by Fuller (1993). Masking through the introduction of additive or multiplicative noise has been investigated (e.g., Fuller 1993). A disclosure limitation method for microdata that is used in the μ-Argus software is a combination of global recoding and local suppression. Global recoding combines several categories of a variable to form less-specific categories. Topcoding is a specific example of global recoding. Local suppression suppresses certain values of individual variables (Willenborg and de Waal 1996). The aim is to reduce the set of records where only a few agree on particular combinations of key values. Both methods

make the data less specific and so result in some information loss to the legitimate researcher.

Some common methods of recoding for tabular data are global recoding and rounding. Under global recoding, categories are combined. This represents a coarsening of the data through combining rows or combining columns of the table. Under rounding, every cell entry is rounded to some base b. The controlled rounding problem is to find some perturbation of the original entries that will satisfy (marginal, typically) constraints and that is 'close' to the original entries (Cox 1987). Multidimensional tables present special difficulties. Methods for dealing with them are given by Kelley et al. (1990).

Markov perturbation (Duncan and Fienberg 1999) makes use of stochastic perturbation through entity moves according to a Markov chain. Because of the cross-classified constraints imposed by the fixing of marginal totals, moves must be *coupled*. This coupling is consistent with a Gröbner basis structure (Fienberg et al. 1998). In a graphical representation, it is consistent with data flows corresponding to an *alternating cycle*, as discussed by Cox (1987).

5.3 Sampling

Sampling as a disclosure-limiting mask creates an appropriate statistical sample of the original data. Alternatively, if the original data is itself a sample, the data may be considered self-masked. Just the fact that the data are a sample may not result in disclosure risk sufficiently low to permit data dissemination. In that case, subsampling may be required to obtain a data product with adequately low disclosure risk.

Whether for microdata or tabular data, many of these transformations can be represented as matrix masks (Duncan and Pearson 1991), $\mathsf{M} = \mathsf{AXB} + \mathsf{C}$, where X is a data matrix, say $n \times p$. In general, the defining matrices A, B and C can both depend on the values of X and be stochastic. The matrix A (since it operates on the rows of X is a record-transforming mask, the matrix B (since it operates on the columns of X is a variable-transforming mask, and the matrix C is a displacing mask (noise addition).

6. Synthetic or Model-based Data

The methods described so far have involved perturbations or masking of the original data. These are called *data-conditioned* methods by Duncan and Fienberg (1999). Another approach, while, less studied, should be conceptually familiar to statisticians. Consider the original data to be a realization according to some statistical model. Replace the original data with samples (the synthetic data) according to the model. Synthetic data sets consist of records of individual synthetic units rather than records the agency holds for actual units. Rubin (1993) suggested synthetic data construction through a multiple imputation method. It remains an open research question as to what the impact of imputation of an entire microdata set is on data utility.

Rubin (1993) asserts that the risk of identity disclosure can be eliminated through the dissemination of synthetic data and proposes the release of synthetic microdata sets for public use. His reasoning is that the synthetic data carries no direct functional link between the original data and the disseminated data. So while there can be substantial identity disclosure risk with (inadequately) masked data, identity disclosure is, in a strict sense, impossible with the release of synthetic data. However, the release of synthetic data may still involve risk of attribute disclosure (Fienberg et al. 1998).

Rubin (1993) cogently argues that the release of synthetic data has advantages over other data dissemination strategies, because

(a) masked data can require special software for its proper analysis for each combination of *analysis* × *masking method* × *database type* (Fuller 1993);

(b) release of aggregates, e.g., summary statistics or tables, is inadequate due of the difficulty in contemplating at the data-release stage what analysts might like to do with the data; and

(c) mechanisms for the release of microdata under restricted access conditions, e.g., user-specific administrative controls, can never fully satisfy the demands for publicly available microdata.

The methodology for the release of synthetic data is simple in concept, but complex in implementation. Conceptually, the agency would use the original data to determine a model to generate the synthetic data. But the purpose of this model is not the usual prediction, control, or scientific understanding that argues for parsimony through Occam's Razor. Instead, its purpose is to generate synthetic data useful to a wide range of users. The agency must recognize uncertainty in both model form and the values of model parameters. This argues for the relevance of hierarchical and mixture models to generate the synthetic data.

7. Summary

A society based on democratic and free market principles cannot function without broad access to data, nor can it sustain those principles without affirming the individual autonomy assured by privacy and confidentiality. Statistical disclosure limitation methods work to ensure that the tension between data access and confidentiality can be resolved in ways that are favourable to both. This methodology draws on disparate fields of mathematics and is challenged to develop in response to the growing capability of information technology to capture, store and dis-

seminate data. To learn more about this topic, see *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics* (Duncan et al. 1993), two special issues of the *Journal of Official Statistics* (Jabine 1993 and Fienberg et al. 1998), the Federal Committee on Statistical Methodology 1994 and Duncan and Fienberg (1999).

Bibliography

Adam N R, Wortmann J C 1989 Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys* **21**: 515–56

Bethlehem J G, Keller W J, Pannekoek J 1990 Disclosure control of microdata. *Journal of the American Statistical Association* **85**: 38–45

Chowdhury S D, Duncan G T, Krishnan R, Roehrig S F, Mukherjee S 1999 Disclosure detection in multivariate categorical databases: Auditing confidentiality protection through two new matrix operators. *Management Science* **45**: 1710–23

Cox L H 1980 Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association* **75**: 377–85

Cox L H 1987 A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association* **82**: 38–45

Dalenius T 1988 *Controlling Invasion of Privacy in Surveys*. Department of Development and Research Statistics, Sweden

Dalenius T, Reiss S P 1982 Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6**: 73–85

Duncan G T, Jabine T B, de Wolf V A 1993 *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Panel on Confidentiality and Data Access, Committee on National Statistics, National Academy Press, Washington, DC

Duncan G T, Fienberg S E 1999 Obtaining information while preserving privacy: a Markov perturbation method for tabular data. Eurostat. *Statistical Data Protection '98, Lisbon 1999*. Office for Official Publications of the European Communities, Luxembourg, pp. 351–62

Duncan G T, Kaufman S 1996 Who should manage information and privacy conflicts?: Institutional design for third-party mechanisms. *The International Journal of Conflict Management* **7**: 21–44

Duncan G T, Lambert D 1986 Disclosure-limited data dissemination (with discussion). *Journal of the American Statistical Association* **81**: 10–28

Duncan G T, Lambert D 1989 The risk of disclosure of microdata. *Journal of Business and Economic Statistics* **7**: 207–17

Duncan G T, Pearson R 1991 Enhancing access to microdata while protecting confidentiality: Prospects for the future (with discussion). *Statistical Science* **6**: 219–39

Federal Committee on Statistical Methodology 1994 *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*. US Office of Management and Budget, Washington, DC

Fienberg S E, Makov U E, Steele R J 1998 Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* **14**: 347–60

Fuller W A 1993 Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**: 383–406

Greenberg B, Zayatz L 1992 Strategies for measuring risk in public use microdata files. *Statistica Neerlandica* **46**: 33–48

Jabine T 1993 Statistical Disclosure Limitation Practices of United States Statistical Agencies *Journal of Official Statistics* **9**: 427–54

Kelley J, Golden B, Assad A 1990 Controlled rounding of tabular data. *Operations Research* **38**: 760–72

Lambert D 1993 Measures of disclosure risk and harm. *Journal of Official Statistics* **9**: 313–31

Miller A R 1971 *The Assault on Privacy: Computers, Data Banks and Dossiers*. University of Michigan, Ann Arbor, MI

Rubin D B 1993 Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. *Journal of Official Statistics* **9**: 461–8

Spruill N L 1983 The confidentiality and analytic usefulness of masked business microdata. *Proceedings of the Section on Survey Research Methods, American Statistical Association* 602–7

Willenborg L, de Waal T 1996 *Statistical Disclosure Control in Practice. Lecture Notes in Statistics 111*. Springer, New York

Winkler W E 1998 Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Research in Official Statistics* **1**: 87–104

G. T. Duncan

Configurational Analysis

Configurational analysis (CA) operates at the interface of (a) Stern's (1911) general approach to analyzing variables on the one hand and (b) the person-oriented approach (Magnusson 1998, Bergman and Magnusson 1991) and Stern's differential approach to analyzing individuals on the other. Just as the standard, aggregate approach, CA uses variables to describe individuals. However, the inter-relationships among the variables are not assumed to be necessarily the same for each individual or group of individuals. In this article, CA is described, a theoretical embedding is provided, and methods of CA are introduced, specifically configural frequency analysis (CFA). Sample results from CA are taken from clinical research.

1. Profiles and Configurations

William Stern introduced in 1911 the distinction between the notions of variability and psychography. Variability is the focus when many individuals are observed in one characteristic with the goal to describe the distribution of this characteristic in the population. Psychographic methods aim at describing individuals in many characteristics. Stern also states that these two methods can be combined.

When describing an individual, results are often presented in form of profiles. For example, test results of the (Minnesota Multiphasic Personality Inventory)

personality test typically are presented in form of individual profiles, and individuals are compared to reference profiles. A profile describes the position of an individual on standardized, continuous scales. Thus, one can also compare the individual's relative standing across several variables. Longitudinally, one can ask, whether an individual's relative standing and/or the correlation with some reference change. For example, a profile may resemble the pattern typical of schizophrenics. Individuals can be grouped based on profile similarity.

In contrast to profiles, configurations are not based on continuous but on categorical variables. Specifically, the ensemble of categories that describes a cell of a cross-classification is called configuration (Lienert 1969). Consider, for example, the variables Gender (female–male), Marital Status (married–not married), and Happiness (yes–no). The cross-classification of these three variables has eight cells. The combinations of categories that describe these cells are the configurations. For example, the three categories female and married, and happy constitute the first configuration. The combination male and not married, and unhappy constitutes the last configuration.

CA investigates such configurations from several perspectives. First, CA identifies configurations. This involves creating cross-classifications, or, when variables are originally continuous, categorization and then creating cross-classifications. Second, CA asks, whether the number of times, a configuration was observed, could have been expected from some *a priori* specified model. Significant deviations will then be studied in more detail. Third, CA asks whether the cases found displaying different configurations also differ in their mean and covariance structures in variables not used for the cross-classification. This question concerns the external validity of configurational statements. Other questions are conceivable.

In the following section, a theoretical background for CA is provided. Specifically, CA will be embedded in differential psychology and the person-oriented approach. In Sect. 3, configural frequency analysis, a chief method for CA will be described. In Sect. 4, other methods such as cluster analysis and log-linear modeling will be described briefly.

2. *Configurational Analysis: Theoretical Embedding*

This section covers two roots of configurational analysis (CA), differential psychology, and the Person-Oriented Approach. The fundamental tenet of differential psychology is that 'individual differences are worthy of study in their own right' (Anastasi 1994, p. ix). This is often seen in contrast to General Psychology where it is the main goal to create statements that are valid for an entire population. General Psychology is thus chiefly interested in variables. The data carriers themselves, e.g., humans, play the role of replaceable random events. They are not of interest *per se*. In contrast, Differential Psychology considers the data carriers units of analysis. The smallest unit would be the individual at a given point in time. However, larger units are often considered, e.g., all individuals that meet the criteria of geniuses, alcoholics, and basketball players.

Differential psychology as both a scientific method and an applied concept presupposes that the data carriers' characteristics are measurable. In addition, it must be assumed that the scales used for measurement have the same meaning for every data carrier. Third, it must be assumed that the differences between individuals are measurable. In other words, it must be assumed that data carriers are indeed different when they differ in location on some scale. It is well known that this is often not the case with questionnaires and tests in the social sciences.

The Person-Oriented Approach (see Magnusson 1998, Bergman and Magnusson 1991, 1997, von Eye et al. 2001) is a relative of Differential Psychology. It is based on five propositions (Bergman and Magnusson 1997, von Eye et al. 2000; see also *Person-centered Research*) that stress the individual nature and complexity of functioning, process, and development (FPD); the lawfulness and structure of FPD; and the organization of FPD in patterns. Most important from the present perspective is Proposition 5, which states that some patterns and configurations will be observed more frequently than others are, or more frequently than expected based on prior knowledge or assumptions. These patterns can be called common types. CFA defines types as configurations that occur more often than expected from some chance model. Accordingly, there will be patterns or configurations that are less frequently observed than expected from some chance model. CFA terms these the antitypes.

Thus, the person-oriented approach and CA meet where (a) patterns of scores or categories are investigated, and (b) where the tenet of differential psychology is employed that it is worth the effort to investigate individuals and groups of individuals. The methodology employed for studies within the framework of the person-oriented approach is typically that of CA. The following section describes configural frequency analysis, one of the central statistical methods of CA.

3. *Configural Frequency Analysis (CFA)*

Configural frequency analysis is introduced by using a data example. Consider the three variables, Gender (G) (1 = Male, 2 = Female); Motive for Suicide Attempt (M) (1 = Illness, 2 = Psychiatric Disorder, 3 = Alcoholism); and Outcome of Suicide Attempt (1 = Survived, 2 = Dead) (see Lienert 1978, von Eye et al. 1996). Crossed, these three variables form a $2 \times 3 \times 2$

Table 1
CFA suicide data

Variable GMO[a]	Frequencies		Statistics		Type/ Antitype?
	Observed	Expected	z	$p(z)$	
111	64	59.42	0.635	0.263	
112	18	24.57	1.360	0.087	
121	76	67.49	1.117	0.132	
122	47	27.91	3.724	0.0001	T[b]
131	7	28.74	4.182	0.00001	A[b]
132	8	11.88	1.141	0.127	
211	86	70.76	1.962	0.025	
212	16	29.26	2.529	0.006	
221	61	80.37	2.367	0.009	
222	25	33.23	1.480	0.069	
231	47	34.23	2.265	0.012	
232	27	14.15	3.467	0.0003	T[b]

a G = Gender; M = Motive for Suicide Attempt, O = Outcome of Suicide Attempt. b T = CFA Type, A = CFA Antitype.

contingency table. This table appears in Table 1, along with results from CFA.

The first column of Table 1 lists the configurations. The first configuration is the combination of the categories, Male and Illness and Survived Suicide Attempt. The second configuration is the combination Male and Illness and Died, and so forth. The second column displays the frequency with which each of the configurations occurred. CA allows researchers to investigate cross-classifications of the kind presented in Table 1 from the perspectives listed above. The following paragraphs illustrate the first two perspectives using CFA as a sample method of analysis.

3.1 Identification of Configurations

The first perspective involved identification of configurations. The 12 configurations in Table 1 are identified as the cells of a $2 \times 3 \times 2$ cross-classification of three categorical variables. Each individual in the sample displays one of the 12 possible patterns and can therefore be unambiguously assigned to one of the cells.

3.2 Comparison of Observed with Expected Cell Frequencies: CFA Types and Antitypes

When inspecting the observed frequencies in Table 1, it becomes immediately obvious that the 12 configurations were observed at different frequencies. Whereas seven alcoholic males were counted to have survived a suicide attempt, 86 ill females were counted to have survived an attempt. While it does make sense to compare raw frequencies, more often researchers try to explain the observed frequency distribution based on an underlying model. Such a model is said to explain the observed data if no significant discrepancies between the observed distribution and the distribution predicted by the model exists.

CFA goes the opposite route. It specifies a so-called base model that is not supposed to explain the data (examples of base models follow in Sect. 3.3). Discrepancies from this model indicate what is going on, that is, the effects that created the observed distribution. These effects manifest in the above mentioned common types and common antitypes. More specifically, let f_i be the observed frequency for Cell i, and e_i the expected frequency for Cell i. The base model is used to estimate the expected cell frequencies. Then, if statistically $f_i > e_i$, the configuration that describes Cell i constitutes a CFA type. If $f_i < e_i$, the configuration constitutes a CFA antitype. If $f_i = e_i$, the configuration constitutes neither a type nor an antitype.

3.3 CFA Base Models

Admissible CFA base models meet the following three criteria (von Eye and Schuster 1998). The first criterion is that there be only one way to deviate from the base model for clearcut interpretation of types and antitypes. For example, first order CFA uses the log-linear main effect model for a base model. This model implies no interactions among variables. Deviations from this model are only possible if there are interactions.

The second criterion, introduced by von Eye and Schuster (1998), is that the sampling scheme be considered. Specifically, the multinomial and the product multinomial sampling schemes have major implications for the selection of base models and interpretation of results (von Eye et al. 1999).

Parsimony is a third criterion for the selection of base models. More specifically, a base model must take into account interactions of the lowest possible order.

Base models can be subdivided in two groups. The first group is that of global CFA models. Models in this group assign the same status to all variables in the analysis. There is a hierarchy of global models. It begins at the level where no effects are assumed at all. Thus, the expected cell frequency for Cell i is estimated as $e_i = t/N$, where t is the total number of cells in the cross-classification and N is the sample size. This model is termed zero-order CFA. At the next higher level, first-order CFA proposes that main effects can exist for each variable, but no interactions. First-order CFA is the original approach to CFA (Lienert 1969). The expected cell frequencies for Table 1 were estimated using the first-order CFA base model (this model will be explained in more detail later in this section). Accordingly, second order CFA proposes that first-order interactions exist between all pairs of variables, and so forth. In each of these models types and antitypes can emerge only if effects of orders higher than the ones assumed in the base model exist.

Regional CFA models distinguishes between two or more groups of variables. Examples of regional CFA models include Prediction CFA (P-CFA) which assumes that the variables are grouped into predictors and criteria. P-CFA types (antitypes) allow one to predict the occurrence (lack of occurrence) of criterion configurations from particular predictor configurations.

Most CFA base models can be expressed in form of simple log-linear models. Thus, expected cell frequencies can be estimated from the marginal frequencies of the variables involved in the cross-classification. A general expression for log-linear models and, thus, log-linear CFA base models, is $\log F = X\lambda$ where F is the array of expected frequencies under study (the third column in Table 1), X is the design matrix (comparable to a design matrix in analysis of variance), and λ is a parameter vector. Estimation is typically performed using maximum likelihood methods. Many CFA base models lead to very simple computational equations.

Once the expected cell frequencies are estimated, they can be compared with the observed frequencies. A large number of statistical tests have been proposed for these comparisons. One of the best of these tests is the binomial approximation of the z-test. The test statistic is $z_i = (f_i - e_i)/\sqrt{(e_i(1 - e_i/N))}$. Inserting the values from the data example in Table 1 yields for the first configuration the test statistic $z_{111} = (64 - 59.416)/\sqrt{59.416(1 - 59.416/482)} = 0.635$. This value is the same as the one for the test statistic for Configuration 111 in Table 1. Other tests include the exact binomial test, the χ^2-component test, Lehmacher's (1981) hypergeometric tests, and the F-test (cf. von Eye and Rovine 1988).

3.4 Statistical Decisions Concerning Types and Antitypes

For each of the configuration specific test statistics a tail probability is then calculated. Recent results suggest that standard testing as described above tends to underestimate the probability of antitypes (Indurkhya and von Eye 2000). Future research will have to devise methods that eliminate this asymmetry. One very important characteristic of the tests performed in CFA is that they are not independent of each other. Rather, they can be dependent because (a) the available information has already been exhausted or (b) capitalizing on chance. Therefore, measures must be taken to protect the significance level α. A number of adjustments have been proposed for protection of α. One of the most conservative yet most frequently employed adjustments is known as Bonferroni adjustment. This procedure creates an adjusted significance level, α^*, that is $\alpha^* = \alpha/t$, where α is the original, nominal significance level, and t is the total number of tests performed. In the example in Table 1 we test each cell. Therefore, $t = 12$. For $\alpha = 0.05$ we obtain $\alpha^* = 0.05/12 = 0.0041667$. Obviously, the adjustment makes the decision as to whether a type or antitype exists more conservative. Therefore, the Configurations, 211, 212, 221, and 231 for which the tail probabilities are below the usual $\alpha = 0.05$ fail to constitute types and antitypes. Still, even under this conservative decision rule, CFA manages to identify two types and one antitype in the data in Table 1. The first type is constituted by Configuration 122. It suggests that more men suffering from psychiatric disorders than expected from the base model of variable independence do not survive a suicide attempt. The second type is constituted by Configuration 232. It suggests that more alcoholic women than expected from the base model die in suicide attempts. The antitype, Configuration 131, indicates that fewer alcoholic men than expected from the base model survive a suicide attempt.

3.5 Interpretation of Types and Antitypes

When interpreting results from CFA, researchers use two sources of information. The first source includes the identified types and antitypes. Most typically (see, for example, Table 1), only a subset of configurations deviates from the CFA base model to the extent that types or antitypes are constituted. If there exist only few significant deviations, the base model is contradicted locally, and in the other sectors of the data space spanned by the cross-classification under study the base model holds.

The second important source of information is the CFA base model. In the analyses for Table 1 the base model proposed that the three variables, Gender, Motive, and Outcome of Suicide Attempt be unrelated throughout. The two emerging types and the one

antitype, however, suggest that there exist local associations such that particular categories of these variables, that is, particular configurations surface more often (and less often) than could be expected under the base model. Thus, the base model experiences local violations in three sectors, yet it holds in the remaining nine sectors. The base model for the CFA of the data in Table 1 can be violated by any relationship among the three variables.

It is important to realize that the application of CFA within the context of Configurational Analysis implies a perspective of variables that is different than the perspective taken when analyzing variables. Within CA, variable interactions and main effects are of interest only in so far as they are part of a CFA base model. The main focus of interest, however, lies in the possible local deviations from the base models. Such deviations are always interpreted at the level of configurations, that is, locally. There have been only a small number of attempts to interpret groups of types and groups of antitypes (von Eye and Brandtstädter 1997, Gutiérrez-Peña and von Eye 2000).

3.6 Variants of CFA

By far most CFA base models are derivatives of log-frequency models. However, there exist circumstances under which the estimation of expected cell frequencies from the observed frequencies might be problematic. An example where the probabilities of configurations are determined by information other than the observed distribution can be found in longitudinal CFA (von Eye 1990). When change scores are calculated for ordinal variables, the top and the bottom ranks have fewer possibilities for change. Whereas the middle ranks can stay the same, move up, or move down, the top rank can only stay the same or move down. Accordingly, the bottom rank can only stay the same or move up. These differences must be considered when estimating expected cell frequencies (for details see von Eye 1990, Chap. 6).

A second domain of development of methods for CFA is that of Bayesian CFA. Bayesian methods differ from standard, frequentist methods in that conclusions about a parameter or about unobserved data are made in terms of probability statements that are conditional on the observed value of the dependent measure. Implicitly, statements are conditioned on known values of the independent variable. In contrast, standard, frequentist statistical inference is based on an evaluation of the procedure used to estimate the parameter for the distribution of possible values of the dependent measure, conditional on the true yet unknown value of the parameter (Gelman et al. 1995). One interesting and important characteristic of Bayesian statistics is that prior information can easily be taken into account.

Thus far, there have been only a small number of attempts to formulate a Bayesian CFA (Wood et al. 1994, Gutiérrez-Peña and von Eye 2000, von Eye et al. 2000). Each of these approaches is (a) unrelated to log-linear modeling methods for estimation of expected cell frequencies, and (b) uses the so-called noninformative priors. The rationale for using this kind of priors is to let the data speak for themselves, that is, without consideration of information external to the data under study (Gelman et al. 1995). However, in many instances such information exists. Thus, future developments will discuss methods of incorporating prior information in the form of prior distributions.

A second route of development that has just begun to be traveled is that of defining types of types (and antitypes). Traditionally, CFA types and antitypes are defined in a way similar to residuals in log-linear models, that is, as significant discrepancies from some base model. In 1991, Goodman made explicit that there are many ways to deviate from expectancy. Three of these ways were used by von Eye, Spiel, and Rovine (1995) to define different concepts of types and antitypes. The first concept is a marginal-dependent relative of the correlation, ρ. The second is a marginal-free variant of the odds ratio, and the third concept is a marginal-dependent variant of the odds ratio. Results of simulation studies suggest that the patterns of types and antitypes identified under the three concepts can differ dramatically. In addition, these patterns can be different than the patterns identified using the original concept of CFA types and antitypes. Gonzáles Dében (1998) proposed a fourth approach to defining types and antitypes that is based on a residual definition by Rudas et al. (1994).

3.7 Alternative Methods of Configurational Analysis

Section 3 suggests that configuration frequency analysis is the chief method of data analysis in CA. However, other methods have been proposed and used. Two of these methods will be briefly reviewed here, log-linear modeling and cluster analysis.

Cluster analysis (see Hartigan 1975, von Eye et al. 1999) is a method that allows researchers to identify groups that were unknown to exist before analysis. Cluster analysis creates groups (clusters) that contain cases that are more similar to each other than to cases from other groups. The cluster formation process typically follows one of the following three strategies: (a) an existing sample is subdivided until some optimality criterion is reached; (b) the individuals in a sample are agglomerated until some optimality criterion is reached; and (c) density centers are formed and individuals gravitate to the nearest center. For more detail on the decision making involved when searching for clusters see von Eye et al. (1999). Cluster analysis and CA meet where researchers hold the assumption that data analysis at the aggregate level

can obscure interpretable differences between individuals or groups of individuals.

The typical application of log-linear modeling focuses on variables. Thus, this method finds more applications in standard, aggregate level analysis (Agresti 1990, Bishop et al. 1975). However, several variants of log-linear modeling can be fruitfully employed in CA. Two sample approaches will be briefly reviewed in this context. The first approach concerns models that specify effects that are specific to subtables. For instance, Clogg and Shockey (1988) proposed models that allow one to ask whether, in two-way cross-classifications, an association between the two variables that constitute the table is the same throughout. In an example with the variables Riding a Car for Fun and Drug Use, the authors showed that there is an association such that those respondents who ride a car often for fun also tend to consume stronger drugs. However, those respondents who never ride a car for fun tend to consume stronger drugs than those who ride a car a few times a year for fun do. This pattern suggests that the association between the two variables switch as one inspects different regions of a cross-classification.

The second approach is linked to nonstandard log-linear models (Rindskopf 1990). Standard log-linear models specify main effects and interactions in terms of coding schemes that are parallel to the ones used for analysis of variance (see, e.g., Christensen 1997). Nonstandard models include contrasts that cannot be expressed using these terms. Rather, these effects, comparable to planned mean comparisons, are added to the model to test specific hypotheses. These hypotheses typically concern effects that occur only in specific regions of a cross-classification. Von Eye and Brandtstädter (1998) proposed using nonstandard log-linear models to test hypotheses about causal effects. While such models may suffer from problems of parameter intercorrelations that can be viewed parallel to multicollinearity in analysis of covariance or multiple regression, they allow researchers to ask questions that go beyond standard questions of variable associations.

These and other approaches to log-linear modeling meet CA where hypotheses are formulated and tested that do not apply to everybody in a population but only to distinct groups. It is assumed that the variable relationships in these groups differ from the ones found for the rest of the populations.

To conclude, configurational analysis (CA) represents a methodology that allows researchers to counterbalance the far more frequently used methodology that focuses on variables rather than on individuals. Its chief representatives, cluster analysis and configural frequency analysis are well-developed statistical techniques that allow researchers to specify statements about individuals and groups of individuals. Variables play in these statements the role of carriers of information.

See also: Classification: Conceptions in the Social Sciences; Event Sequence Analysis: Methods; Multivariate Analysis: Discrete Variables (Loglinear Models); Person-centered Research; Statistical Clustering; Statistical Pattern Recognition

Bibliography

Agresti A 1990. *Categorical Data Analysis*. Wiley, New York

Anastasi A 1994 Geleitwort: Differential psychology: origin and sources. In: Pawlik K (ed.) *Die differentielle Psychologie in ihren methodischen Grundlagen*. Verlag Hans Huber, Bern, Germany

Bergman L R, Magnusson D 1991 Stability and change in patterns of extrinsic adjustment problems. In: Magnusson D, Bergman L R, Rudinger G, Toerestad B (eds.) *Problems and Methods in Longitudinal Research*. Cambridge University Press, Cambridge, UK, pp. 323–46

Bergman L R, Magnusson D 1997 A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology* **9**: 291–319

Bishop Y M M, Fienberg S E, Holland P W 1975 *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA

Cattell R B 1988 The meaning and strategic use of factor analysis. In: Nesselroade J R, Cattell R B (eds.) *Handbook of Multivariate Experimental Psychology*, 2nd edn. Plenum, New York, pp. 131–203

Christensen R 1997 *Log-linear Models and Logistic Regression*, 2nd edn. Springer, New York

Clogg C C, Shockey J W 1988 Multivariate analysis of discrete data. In: Nesselroade J R, Cattell R B (eds.) *Multivariate Experimental Psychology*. 2nd edn. Plenum, New York, pp. 337–65

Gelman A, Carlin J B, Stern H S, Rubin D B 1995 *Bayesian Data Analysis*. Chapman and Hall, London

Gonzáles Debén A 1998 Experiencias con un nuevo indice de falta de adjuste en el analysis de tables de contigencia. Unpublished Master's Thesis, Universidad de la Habana, La Habana

Goodman L A 1991 Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical Association* **86**: 1085–111

Gutiérrez-Peña E, von Eye A 2000 A Bayesian approach to Configural Frequency Analysis. *Journal of Mathematical Sociology* **24**: 151–74

Hartigan J A 1975 *Clustering Algorithms*. Wiley, New York

Indurkhya A, von Eye A 2000 The power of tests in configural frequency analysis. *Psychologische Beitraege* **42**: 301–8

Lehmacher W 1981 A more powerful simultaneous test procedure in configural frequency analysis. *Biometrical Journal* **23**: 429–36

Lienert G A 1969 Die 'Konfigurationsfrequenzanalyse' als Klassifikationsmethode in der klinischen Psychologie. Paper presented at the 26. Kongreß der Deutschen Gesellschaft für Psychologie in Tübingen 1968

Lienert G A 1978 *Verteilungsfreie Methoden in der Biostatistik*. Meisenheim am Glan, Hain, Vol. 2

Magnusson D 1998 The logic and implications of a person-oriented approach. In: Cairns R B, Bergman L R, Kagan J (eds.) *Methods and Models for Studying the Individual*. Sage, Thousand Oaks, CA, pp. 33–63

Rindskopf D 1990 Nonstandard log-linear models. *Psychological Bulletin* **108**: 150–62

Rudas T, Clogg C C, Lindsay B G 1994 A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society B Met* **56**: 623–39

Stern W 1911 *Die differentielle Psychologie in ihren methodischen Grundlagen*, 3rd edn. J A Barth, Leipzig, Germany

von Eye A 1990 *Introduction to Configural Frequency Analysis. The Search for Types and Antitypes in Cross-Classifications*. Cambridge University Press, Cambridge, UK

von Eye A, Brandtstädter J 1997 Configural frequency analysis as a searching device for possible causal relationships. *Methods of Psychological Research—Online* **2**: 1–23

von Eye A, Brandtstädter J 1998 The wedge, the fork, and the chain: modeling dependency concepts using manifest categorical variables. *Psychological Methods* **3**: 169–85

von Eye A, Indurkhya A, Kellam S 2001 Typifying developmental trajectories. (in preparation)

von Eye A, Indurkhya A, Kreppner K 2000 CFA as a tool for person-oriented research—Unidimensional and within-individual analyses of nominal level and ordinal data. *Psychologische Beiträge* **42**: 383–401

von Eye A, Rovine M J 1988 A comparison of significance tests for Configural Frequency Analysis. *EDP in Medicine and Biology* **19**: 6–13

von Eye A, Schuster C 1998 On the specification of models for configural frequency analysis—sampling schemes in prediction CFA. *Methods of Psychological Research–Online* **3**: 55–73

von Eye A, Schuster C, Gutiérrez-Peña E 2000 Configural frequency analysis under retrospective and prospective sampling schemes—frequentist and Bayesian approaches. *Psychologische Beitraege* **42**: 428–47

von Eye A, Spiel C, Wood P K 1996 Configural frequency analysis in applied psychological research. *Applied Psychology: An International Review* **45**: 301–27

von Eye A, Spiel C, Rovine M J 1995 Concepts of independence in configural frequency analysis. *Journal of Mathematical Sociology* **20**: 41–54

Wood P K, Sher K J, von Eye A 1994 Conjugate methods in configural frequency analysis. *Biometrical Journal* **36**: 387–410

A. von Eye

Conflict and Conflict Resolution, Social Psychology of

The term 'conflict' has two broad meanings: *overt conflict*, which refers to clashing actions by two or more parties, as in a fist fight or war, and *subjective conflict*, which refers to perceived divergence of interest or annoyance attributed to another party. Subjective conflict is often a source of overt conflict, as when annoyance about a neighbor's loud music leads to a shouting match. But subjective conflict can produce other outcomes as well, including use of the following four basic strategies: contending, problem solving, yielding, and inaction. Only one of these, contending, leads to overt conflict.

Conflict occurs at all levels of society, from the interpersonal to the international, and hence is of interest to most branches of the social and behavioral sciences. Conflict has attracted the attention of researchers because of its large costs and benefits. Costs are incurred when conflict escalates to the point of harming relationships, destroying property, or injuring people. Benefits accrue because conflict presides over most significant social change; contributes to solidarity within conflicting groups; and in its milder forms, often leads to the reconciliation of legitimate interests, thus strengthening relationships and safeguarding the peace.

1. Sources of Conflict

The modern social psychology of conflict began with a demonstration field experiment by Sherif (1966). Sherif ran several summer camps in which he produced overt conflict between two cabins of preadolescent boys and then resolved this conflict. Conflict was produced by means of the two sources of subjective conflict mentioned above: divergence of interest (for example, competitive games) and annoyance attributed to the other group (for example, the counselors vandalized one cabins' campgrounds and blamed it on the other cabin).

What conditions produce divergence of interest and annoyance from others? One answer is scarce resources. A second is role differentiation that produces disparate values, as between buyer and seller, parent and child, sales and production. A third is any condition that causes aspirations to rise rapidly (such as a sudden improvement in outcomes) or to become inconsistent with those of another party (such as ambiguity about which party is more powerful). A fourth is any source of distrust, because distrust tends to block cooperation and produce defensive behavior, which often annoys or frightens the other party.

When groups rather than individuals are involved, additional mechanisms encourage conflict. Groups are less trusted than individuals. Also, a perception that one's group is deprived relative to a reasonable standard (*fraternalistic deprivation*) has been shown to be a major source of political conflict. Arguing that Sherif's 'realistic conflict theory' is too narrow, Tajfel (1978) proposes that intergroup conflict often arises because group members compete with other groups in an effort to feel good about their social identity and hence about themselves. Another unique aspect of groups is the role of leadership. Leaders and would-be leaders often dramatize frustration at the hands of other groups, exacerbating conflicts while strengthening their position with their constituents.

2. Strategic Choice in Conflict

Rubin et al. (1994) have put forward a *dual concern model* of the psychological states that affect choice

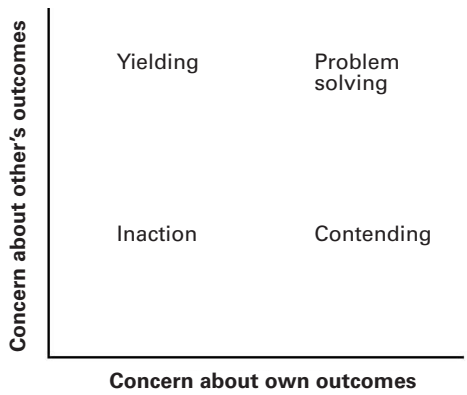

Figure 1
The dual concern model (after Pruitt and Carnevale 1993, p. 105)

among the four basic strategies mentioned earlier. Shown in Fig. 1, this model postulates two kinds of concern: concern about own outcomes ('self-concern') and concern about the other party's outcomes ('other-concern'), each ranging from weak to strong. People with strong self-concern have high, inflexible aspirations and are resistant to concession making. People with strong other-concern place importance on the other party's interests. Self-concern and other-concern are independent dimensions rather than opposite ends of a continuum. Beginning in the upper left-hand quadrant of Fig. 1, low self-concern coupled with high other-concern is assumed to produce a strong altruistic orientation, which encourages yielding. High self-concern combined with high other-concern is assumed to produce a strong cooperative orientation, which encourages problem solving. High self-concern combined with low other-concern is assumed to produce a strong individualistic orientation, which leads to the use of contentious tactics. Low self-concern coupled with low other-concern is assumed to produce inaction, an absence of efforts to achieve anything.

The dual concern model, though not a comprehensive theory of strategic choice, fits a lot of data from psychometric and experimental research. The psychometric tradition, which employs factor analysis and multidimensional scaling, views strategic preferences as individual differences in 'conflict style.' The experimental tradition, which involves studies of negotiation behavior, examines conditions that affect the likelihood of choosing each strategy. These studies have identified several sources of self-concern, including having high aspirations or fallback positions, being an accountable representative, and negatively framing the issues so that concessions are interpreted as loss rather than failure to gain. Sources of other-concern include having a close relationship with or dependence on the other party, and being in a positive mood. Other research suggests that contentious tactics are more likely to be adopted the greater the annoyance or perceived divergence of interest, the more dehumanized the target, and the more deindividuated the actor or target.

When increasingly harsh contentious tactics are used in a conflict, *escalation* is said to occur. Escalation usually results from a conflict spiral in which each party is reacting to the other party's most recent actions. However, some escalation consists of one party's reactions to persistent annoyance from another party (Pruitt 1998). Escalation is often accompanied by other changes: issues proliferate, negative attitudes and partisan perceptions develop, distrust sets in, goals change from doing well for oneself to hurting the other party, and broader communities polarize into hostile camps (Rubin et al. 1994). If groups are involved, ingroup solidarity usually increases and militant leaders often become more prominent. These changes tend to outlive the conflict in which they are generated, hurting the broader relationship between the parties and making the next conflict episode more likely to escalate (Coleman 1957).

3. *Conditions that Encourage Conflict Resolution*

Conflict resolution refers to any reduction in the severity of conflict or mitigation of its underlying causes. It may entail de-escalation of an overt struggle or reconciliation (in whole or in part) of divergent interests.

Social psychologists have long been interested in the role of *social contact* and *communication* in the resolution of conflict. Research has shown that these procedures help to alleviate mild conflict but may be worse than useless in severe conflict, allowing arguments and fights to develop (Deutsch 1973). Sherif (1966) found that a better way to reverse escalation in his camps was to get the boys to cooperate on *superordinate* (i.e., shared) *goals*. The value of this technique has been confirmed in other settings.

Zartman (1997), an international relations scholar who takes a psychological approach, has argued that severe conflict is most likely to be resolved when both sides become motivated to escape the conflict, a condition he calls *ripeness*. He finds that superordinate goals ('mutually enticing opportunities') seldom underlie ripeness in international conflict and civil war. Rather, ripeness is usually produced by a *mutually hurting stalemate*, in which the parties find that they are enduring unacceptable costs in a struggle they cannot win, sometimes augmented by the perception of an *impending catastrophe* if the conflict continues.

It can be argued that severe conflicts of the kind discussed by Zartman will only move toward resolution if ripeness is supplemented by *optimism* about the success of conciliatory actions. Optimism derives from two sources: *working trust*, a belief that the other

side is also motivated to resolve the conflict, and *perceived common ground*, a perception that a mutually acceptable agreement is actually possible.

Optimism can develop in a number of ways. Sometimes one of the parties takes *unilateral conciliatory initiatives* that encourage the other party to become more trusting. Such initiatives are especially effective when they are noticeable and unexpected, are fully explained, and cannot be construed as a trick or sign of weakness (Rubin et al. 1994). In severe conflicts, optimism more commonly arises out of intervention by *intermediaries*. Sometimes the psychological and social distance between adversaries is so great that a chain of two or more intermediaries is necessary to bridge the gap.

4. *Negotiation*

Conflict resolution can begin in a number of ways, but to be successful, it usually must culminate in *negotiation*. In other words, the parties must eventually try to work out their differences.

Research on negotiation has been the centerpiece of the social psychological approach to conflict for many years. Most of it is done in the laboratory with experimental games, but case studies and questionnaires are also sometimes used. At one time, research on negotiation was mostly performed in departments of psychology, but today, the main locus of this research is in management schools, where investigators who are basically social psychologists conduct dozens of studies each year. Early experimental research dealt with negotiation about a single issue (e.g., a price or a wage rate), where one party's gain is the other's loss. But current research usually looks at more complex settings that have *integrative potential*, in the sense that all parties can do well if they engage in problem solving. An important generalization, which applies to both kinds of settings, is that higher demands and slower concessions make it harder to reach agreement but, if agreement is reached, produce more favorable outcomes. Summaries of the psychological research on negotiation can be found in Bazerman et al. (2000) and Pruitt and Carnevale (1993).

5. *Third Party Intervention*

When conflict threatens to escalate and negotiation is ineffective, third parties often intervene. Five intervention procedures will be reviewed here: mediation, arbitration, relationship therapy, peacekeeping, and the design of conflict management systems.

In *mediation*, a third party assists the disputants with their negotiation. (Mediation includes the activities of intermediaries, a topic discussed earlier.) Among the most important mediator tactics are building rapport with the disputants, facilitating communication, questioning unrealistic aspirations, reframing the issues, encouraging disputant creativity, and thinking up new solutions if the disputants fail to do so. Several studies support a broad generalization, that vigorous mediator intervention tends to be effective when the parties are hostile toward each other or lack the motivation to escape their conflict but is counterproductive when the opposite is true (Rubin et al. 1994). Mediators are especially likely to succeed when the parties are highly motivated to resolve the conflict, resource shortages are not severe, and the issues do not involve general principles. Mediator neutrality usually contributes to the success of mediation but is not as essential, as formerly believed. Indeed, success is sometimes achieved because the mediator is closer to the side that must make the larger concessions (Kressel et al. 1989).

In *arbitration*, the third party renders a judgment about the proper solution to the conflict. Several studies have examined the impact of *med-arb*, a procedure in which binding arbitration occurs if agreement is not reached in mediation. This procedure usually motivates the disputants to try harder to solve their dispute during the mediation phase, because they fear loss of control over their outcomes if arbitration occurs (Kressel et al. 1989).

When disputants have a distressed relationship, third parties sometimes engage in *relationship therapy* (Pruitt 1998). Relationship therapy, which has been pioneered by marital therapists, usually involves problem solving training, in which the parties are taught skills of listening, clear communication, avoiding blame, analyzing needs, and brainstorming for solutions. The therapist may also help the disputants identify repetitive patterns of interaction that are contributing to the conflict. Kelman (1992) has developed *problem solving workshops*, a form of relationship therapy for intergroup and international conflict. In these workshops, selected members of both groups meet for several days to analyze the conflict.

When conflicts become violent or threaten to do so, third party intervention may take the form of *peacekeeping*. Examples include separating fighting children, arresting abusive spouses, and United Nations operations within ethnically divided countries. Fisher and Keashly (Fisher 1990) have proposed a normative model which recommends that third parties: (a) use peace keeping in violent conflicts, (b) provide incentives to encourage ripeness in heavily escalated but nonviolent conflicts, (c) offer relationship therapy (they call it 'consultation') in moderately escalated conflicts, and (d) mediate mildly escalated conflicts. This model assumes that each of these strategies produces the conditions necessary for the success of the next strategy in the list.

A final third-party role is the design of *conflict management systems* for large social entities such as

schools, factories, and communities. Guidelines put forward by Ury et al. (1988) include identifying potentially antagonistic groups, appointing and training prospective negotiators on each side who will meet if conflict develops, appointing and training prospective mediators to help these negotiators, and agreeing to observe an immediate cooling-off period whenever conflict develops. Such systems work better if they involve early warning systems that allow negotiators and mediators to mobilize in advance of hostilities.

6. *Disputant Preferences among Conflict Resolution Procedures*

A number of studies have examined disputant preferences among the procedures available for conflict resolution. A persistent finding is that arbitration is preferred to autocratic decision making. According to Lind and Tyler (1988), this shows that having a voice in the proceedings contributes to a sense of procedural justice. Preferences among negotiation, mediation, and arbitration depend on the circumstances. For example, high time pressure enhances attraction to arbitration, presumably because it can be finished up quickly. On the other hand, people in close relationships tend to reject arbitration, presumably because of its coercive features. Sequences of preferences are often found, a common one being to start with negotiation and move to third-party intervention if negotiation is unsuccessful (Pruitt and Carnevale 1993).

Some recent research has concerned cultural differences in preferences among conflict resolution procedures. For example, when dealing with ingroup members, people from collectivist societies prefer nonconfrontational procedures, hence, they are less attracted to direct competition and problem solving than people from individualist societies (Bazerman et al. 2000). But when dealing with outgroup members, collectivists tend to be more competitive than individualists (Pruitt and Carnevale 1993).

7. *Conclusions*

Conflict and conflict resolution are vigorous research fields. The interdisciplinary nature of these fields can be seen in the annual conferences of the International Association for Conflict Management and the pages of the *Journal of Conflict Resolution*, the *International Journal of Conflict Management*, and the *Negotiation Journal*. Social psychologists in psychology departments were more prominent in these fields at an earlier time than they are today. But the contribution of social psychology is readily apparent in the work of investigators from most other disciplines.

Three criticisms of these fields seem appropriate. The first is that they lack an overarching theory and mainly depend on ideas from other disciplines, a problem shared with many other branches of the social sciences. The other two criticisms concern the study of negotiation. One is that there is too much emphasis on the phase of actual negotiation, in which joint decisions are made, as opposed to the pre- and postnegotiation phases. Indeed, the latter two phases are often longer and more complex than the phase of actual negotiation and have more bearing on whether conflict is resolved (Druckman 1986). The other criticism is that there is too much research on negotiation altogether. Negotiation is, after all, only one star in a much larger universe.

See also: Conflict and Socioemotional Development; Conflict: Anthropological Aspects; Conflict/Consensus; Conflict Sociology; Dispute Resolution in Economics; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Negotiation and Bargaining: Role of Lawyers

Bibliography

Bazerman M H, Curhan J R, Moore D A, Valley K L 2000 *Negotiation. Annual Review of Psychology*. Annual Reviews, Palo Alto, CA

Coleman J S 1957 *Community Conflict*. Free Press, Glencoe, IL

Deutsch M 1973 *The Resolution of Conflict: Constructive and Destructive Processes*. Yale University Press, New Haven, CT

Druckman D 1986 Stages, turning points, and crisis: Negotiating military base rights, Spain and the United States. *Journal of Conflict Resolution* **30**: 327–60

Fisher R J 1990 *The Social Psychology of Intergroup and International Conflict Resolution*. Springer-Verlag, New York

Kelman H C 1992 Informal mediation by the scholar/practitioner. In: Bercovitch J, Rubin J Z et al. (eds.) *Mediation in International Relations: Multiple Approaches to Conflict Management*. St. Martin's Press, New York

Kressel K et al. 1989 *Mediation Research: The Process and Effectiveness of Third-Party Intervention*. Jossey-Bass, San Francisco, CA

Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum, New York

Pruitt D G 1998 Social conflict. In: Gilbert D T, Fiske S T, Lindzey G (eds.) The *Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, MA, Vol. 2

Pruitt D G, Carnevale P J 1993 *Negotiation in Social Conflict*. Brooks/Cole, Pacific Grove, CA

Rubin J Z, Pruitt D G, Kim S H 1994 *Social Conflict: Escalation, Stalemate, and Settlement*, 2nd edn. McGraw-Hill, New York

Sherif M 1966 *Group Conflict and Cooperation*. Routledge and Kegan Paul, London

Tajfel H 1978 *Differentiation between Social Groups: Studies in the Social Psychology of Intergroup Relations*. Academic Press, New York

Ury W L, Brett J M, Goldberg S 1988 *Getting Disputes Resolved: Designing Systems to Cut the Costs of Conflict*. Jossey-Bass, San Francisco, CA

Zartman I W 1997 Explaining Oslo. *International Negotiation* **2**: 195–215

D. G. Pruitt

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Conflict and Socioemotional Development

1. Conflict Defined

Conflict is a broad term, widely invoked in the common vernacular and in the scientific literature to suggest a state of disagreement or opposition. The term carries different connotations depending upon whether the experiencer is singular or plural. Intrapersonal conflict describes a state of emotional turmoil or intellectual dissent that occurs within an individual. Interpersonal conflict describes a social interchange that is marked by opposing goals and that involves two or more persons. The earliest English usage of each term dates to the fifteenth century; both have common roots in the Latin *conflictus*, meaning throw or strike together (Shantz and Hartup 1992). This shared etiology tends to muddy distinctions between these two distinct forms of conflict.

1.1 Intrapersonal Conflict

Intrapersonal conflict encompasses two broad domains. At the emotional level, conflict reflects competing impulses or desires. Freud (1965) attributed these tendencies to contradictory aggressive and sexual proclivities with origins in the id (the unregulated physical self) and to irreconcilable differences between these selfish urges and constraints imposed by the superego (the social mores conveyed by family and culture). At the cognitive level, conflict reflects dissonance produced by incongruent facts, opinions, or modes of thought. Piaget (1985) attributed these tendencies to equilibration, the innate tension between a preference for stability and a drive to master the environment. Assimilation encourages stability by incorporating new information into existing mental schema. Should this prove insufficient for mastering the environment, accommodation is required to alter mental schema in a manner consistent with the information available.

Intrapersonal conflict is by definition a private experience. It cannot be observed and it has no reliable metric. Early scientific psychologists relied on introspection to glean insight into thoughts, a practice modified and popularized by psychodynamic psychologists to probe unconscious emotional states. Subjective experience as a topic of study fell out of favor with the behavioral revolution. Reaction times, physiological cues, and projective responses were employed as proxies for inner conflict, but poor reliability and the absence of face validity undermined these efforts. With the cognitive (counter)revolution, new introspective methods gained legitimacy as advances in measurement and instrumentation were applied to mental processes. Scholars began to focus on the diverse experiences of different cultural and ethnic groups, and in so doing it became apparent that scientific inquiry into intrapersonal conflict cannot proceed in the absence of valid and reliable self-report data.

1.2 Interpersonal Conflict

Interpersonal conflict may take place at the level of the dyad or the group. Dyadic conflict is a social episode marked by overt behavioral opposition. The function of such conflicts and the manner in which they unfold vary according to characteristics of the participants and their relationship, the issue and context of the disagreement, the tactics employed, and their consequences (Deutsch 1973). Group conflict refers to negative or incompatible attitudes and behaviors directed by members or representatives of one group toward those of another group. The nature and significance of intergroup conflict differs according to leadership styles and prevailing norms, member cohesion, resource availability, and the individual and collective talents and experiences of each group (Sherif et al. 1961).

Interpersonal conflict is a quantifiable, ofttimes public event. It consists of a sequence of easily recognized and well-defined components including the initial opposition, and the tactics, resolution, and outcome. That said, empirical studies of dyadic and group conflict suffer from inconsistent operational definitions. The conflict measures employed differ in terms of whether they assess unilateral or mutual oppositions, and in terms of whether the frequency of an event is distinguished from its affective intensity (Laursen and Collins 1994). Distinct patterns of interpersonal conflict emerge when considering reciprocated as opposed to unreciprocated disagreements and when angry disputes are examined apart from the mundane. Early work on interpersonal conflict focused on groups and dyads in experimenter-contrived conditions designed to elicit disagreements. Concerns about the ethical treatment of participants and a lack of generalizability prompted improvements in ecological validity; unobtrusive observations in natural settings have been integrated with running assessments of psychophysiological states and moment-by-moment accounts of participant cognitions and affect (Gottman 1994). The utility of self-reports increased once scholars realized that participant reports of conflict differ not only from one another but also from observer reports, suggesting that what was once considered measurement error may in fact represent important individual differences in how conflict is perceived and experienced.

2. The Role of Conflict in Development

Conflict is inherent in growth, and human development cannot proceed without conflict. Wisdom,

emotional maturity, and social skills may be tied to a series of developmentally defined tasks and the conflicts specific to each. Scholars agree that intrapersonal and interpersonal conflict each shape developmental trajectories, but there is little consensus as to the exact mode of transmission. The developmental effects of conflict appear to be limited to specific domains of influence circumscribed by individual timetables and characteristics of the disagreement such as the issue and participants.

2.1 Intrapersonal Conflict and Development

Intrapersonal conflict is instrumental to the attainment of several important milestones in socioemotional development. First, conflict may prompt changes in internal schema and mental structures. Accommodation (Piaget 1985) exemplifies this process: A child confronted with a difficult moral dilemma may be forced to disregard views previously held dear in order to objectively consider different perspectives and reach a fair solution. Alternatively, conflict may give rise to changes in perceptions and priorities. Cognitive dissonance (Festinger 1957) typifies this mechanism: A child forced to select a single partner for an outing may find that, in so choosing, the esteem of one friend has been enhanced at the expense of others. Finally, conflict may alter affect and impulse control. A child who opts to delay gratification in the hopes of improved rewards will learn to succeed by cultivating strategies that minimize emotional arousal and divert attention from proximate stimuli.

Theory concerning the ontogenetic significance of intrapersonal conflict has outpaced research on the topic. There is less research on emotional development than on cognitive and social–cognitive development, but the available evidence supports the assertion that the influence of intrapersonal conflict is defined by the topic of disagreement and the cognitive or affective domain invoked. Furthermore, effects are qualified by individual timetables: Development mediates the influence of conflict. The likelihood that some developmental periods are more susceptible to alterations than others leads to the provocative suggestion that intrapersonal conflict is neither a necessary nor sufficient condition for development but rather a factor that may, at certain times, facilitate advances in the specific emotional or cognitive arena in which it arises (Shantz and Hartup 1992).

2.2 Interpersonal Conflict and Development

Social skills and social development are entwined with interpersonal conflict. Disagreements, especially those in close relationships, are one of the primary means whereby individuals define themselves and delimit their interactions with others. Three influence pathways merit mention. First, conflict may prompt changes in self-understanding and self-evaluation. Critical disagreements force participants to re-evaluate goals and tactics, offering opportunities to hone social perspective taking and interpersonal negotiation skills (Selman 1980). Second, conflict helps to determine patterns of social interaction in close relationships that may impact developmental outcomes. Coercive exchanges tend to be self-perpetuating and, once started, they interfere with the dividends that normally accrue from positive affiliations (Patterson 1982). Third, conflict may define and alter expectations about individuals and relationships. In some cases, roles and responsibilities are negotiated directly, but in other cases, behavior is evaluated for cues as to whether current expectations are consistent with the demands of the situation and the maturity of the participants (Hinde 1997).

There is considerable support for the premise that the developmental significance of conflict depends on the relationship in which it arises (Laursen and Collins 1994). Three relationship properties are germane: closeness, permanence, and power. Disagreements in close relationships have the potential to alter developmental trajectories profoundly; few disagreements in other relationships have such effects. Disagreements in voluntary relationships are more apt to bring about individual change than those in obligatory relationships; participants in the former must behave in a manner that affords mutually satisfactory outcomes if the affiliation is to continue, whereas participants in the latter worry little about the dangers of relationship dissolution. Disagreements in horizontal or mutual relationships provide especially fertile ground for improved social skills because when power is shared, negotiation is necessary for an amicable resolution; in vertical or unilateral relationships, resolutions are more often the product of power differentials than of negotiation abilities. These variations notwithstanding, developmental shifts in conflict man agement may be identified such that the practice of coercion is gradually replaced across childhood and adolescence by negotiation and withdrawal.

3. Conclusion: the Paradoxical Influence of Conflict

Conflict is often unpleasant and aversive. Given the option, most people would choose to avoid it. For this reason, it is commonly assumed that conflict hinders development. This assumption has been challenged on the grounds that linear models overlook the beneficent role that conflict plays in maturation. A curvilinear model of influence captures this alternative perspective: moderate conflict provides experience necessary for optimal socioemotional development, whereas too little conflict limits the potential for growth and too

much conflict overwhelms mechanisms for coping with it.

3.1 Future Directions

Most research on the significance of interpersonal conflict has focused on its potential for adverse consequences. Few studies have addressed the possibility that conflict promotes socioemotional development. Strong evidence supports the assertion that children who experience high rates of angry conflict suffer from low self-esteem, poor social skills, interpersonal rejection, and adjustment problems. Yet the nature of this research is such that it is impossible to determine whether maladaptation is the cause or the consequence of interpersonal conflict. Neither can it be determined whether there are sensitive periods during which conflict exerts a particularly strong influence on development, although a few studies suggest that conflict mediates difficult transitions through precocious puberty, marital reconfiguration, and school entry. Of particular importance to this field of inquiry is research that makes a full and nuanced account of the dynamic interplay between conflict and developmental outcomes. Studies of change over time are needed to specify the paradoxical influence of conflict on concurrent individual adjustment and on long-term developmental trajectories.

See also: Adulthood: Emotional Development; Conflict/Consensus; Infancy and Childhood: Emotional Development; Social Learning, Cognition, and Personality Development

Bibliography

Deutsch M 1973 *The Resolution of Conflict: Constructive and Destructive Processes*. Yale University Press, New Haven, CT

Festinger L 1957 *A Theory of Cognitive Dissonance*. Stanford University Press, Stanford, CA

Freud S 1965 *The Problem of Anxiety*. Norton, New York (Original work published 1936)

Gottman J M 1994 *What Predicts Divorce? The Relationship Between Marital Processes and Marital Outcomes*. Erlbaum, Hillsdale, NJ

Hinde R A 1997 *Relationships: a Dialectical Perspective*. Psychology Press, Hove, UK

Laursen B, Collins W A 1994 Interpersonal conflict during adolescence. *Psychological Bulletin* **115**: 197–209

Patterson G R 1982 *Coercive Family Process*. Castalia, Eugene, OR

Piaget J 1985 *The Equilibration of Cognitive Structures*. University of Chicago Press, Chicago (Original work published 1975)

Selman R 1980 *The Growth of Interpersonal Understanding: Developmental and Clinical Analyses*. Academic Press, New York

Shantz C U, Hartup W W 1992 *Conflict in Child and Adolescent Development*. Cambridge University Press, Cambridge, UK

Sherif M, Harvey O J, White F J, Hood W R, Sherif C W 1961 *Intergroup Conflict and Cooperation: the Robbers' Cave Experiment*. University of Oklahoma, Norman, OK

B. Laursen

Conflict and War, Archaeology of

Archaeology is the study of patterns of effects; repetitions of human behavior that leave an enduring mark on the physical world. War, the armed conflict between social units, is such a pattern and leaves very enduring effects. Such effects include the remains of victims of homicide and warfare, fortifications, specialized weapons, destroyed settlements, and depictions of combat. Ancient violent conflicts also had important consequences, correlates, and causes involving human ecology, the distribution of settlements, production and exchange, social organization, ideology, and symbolism.

The definitions of 'war' used by social scientists and historians can be grouped into two extreme categories: broad and narrow. Broad definitions, like that used in the introductory paragraph, classify violent and lethal conflicts between all social or political units as war. Broad definitions allow the survey and analysis of the widest range of deadly conflicts between human groups. They permit anthropologists studying prehistoric and recent small-scale societies to contribute to scholarly and popular debates about warfare. Narrow definitions restrict the term 'war' just to societies organized as states. States being political organizations that are hierarchically organized, centrally directed, class-stratified with a high degree of occupational specialization (especially, full-time military specialists) and which maintain a monopoly of deadly force. Narrow definitions restrict analysis of warfare only to societies roughly similar in scale and complexity to modern nations. They excuse ignorance about violent conflicts between nonstate societies. Broad definitions imply that warfare is a human universal while narrow definitions claim that collective violence is a scourge peculiar to states.

Until the development of scientific dating methods around 1950, the primary concern of archaeologists was establishing sequences of cultural changes for their respective regions. Their interpretations of social phenomenon, such as warfare, when they made any at all, tended to be off-hand with little basis in analysis or argument. Their attitudes toward prehistoric warfare were also affected by the Social Darwinist attitudes that characterized the heyday of Western Imperialism. Many kinds of implements and constructions were

blithely assumed to be weapons of war or fortifications. Many cultural changes were declared to be the result of prehistoric conquests. The shocks of World War II's savagery, postwar decolonization, and the 'atomic fear' of the Cold War fostered a severe aversion to war and conquest in the Western intellectual and popular cultures. Q. Wright (*A Study of War*, 1942, University of Chicago Press) and H. Turney-High (*Primitive War*, 1949, University of South Carolina Press) persuaded most postwar social scientists and historians that the fighting of prestate societies was ritualized, undangerous, and ineffective, a game or sport rather than a serious business. Both argued that fighting only became dangerous, terrible, and effective when societies became politically centralized and socially complex, especially as civilized states. 'Ethnographic analogy' based on these generalizations implied that prehistoric 'combat,' like games or sports, was politically, economically, and ecologically ineffective, unimportant, and frivolous. While archaeologists continued to record evidence of prehistoric conflicts, analysis or discussion of these traces became less and less common in the post-War era.

The development of radiometric dating after the war freed archaeologists to focus upon the actual functioning of the societies they studied as well as how and why they changed. For any number of reasons, since around 1950, the many theoretical currents among archaeologists have eschewed war and conquests as interpretations or explanations in favor of ecological, economic, and indigenous social evolutionary processes. Increasingly, fortifications were reinterpreted as ritual enclosures, and many obvious weapons were seen as merely symbols of wealth and status. Post-War archaeologists, through disregard or reinterpretation, pacified the past.

However, since around 1985 archaeologists around the world have uncovered some obvious, even shocking, remains left by ancient armed conflicts (see Sect. 1). These have led to a revival of interest by many archaeologists in ancient warfare. The relatively small-scale but horrific conflicts between ethnic groups and small nations that have appeared in many places after the end of the Cold War may also have encouraged this change. Discussions of methodologies for documenting prehistoric warfare as well as war's possible significance in prehistory have begun appearing in major archaeological journals and books.

1. *Human Remains*

The most unequivocal evidence of ancient homicides are human remains bearing traumas caused by weapons, for example, sword cuts, dents made by axes, and depressed fractures made by maces, etc. However, the most commonly noted weapons traumas, worldwide, involve projectile points of bone or stone embedded in the skeletons of victims. The absence of any healing or inflammation around such traumas indicates that the victim did not long survive them. The bones of early hominids do show healed and unhealed traumas (e.g., Neanderthals seemed to have been particularly accident-prone). But before the widespread use of stone and bone projectile tips (around 40 000 BP), it is impossible to determine whether these traumas were effected by human violence or by other more prosaic causes. Evidence of homicide (not necessarily warfare) appears as soon as modern humans appear on the scene. At Grimaldi, Italy, (around 32 000 BP), the skeleton of a child was found with a bone projectile point embedded in its spine. In the Nile Valley, Egypt, a male skeleton (around 20 000 BP) had several projectile points in his abdomen and a partially healed one embedded in his arm. The earliest clear evidence for warfare also comes from the Nile Valley from the Late Paleolithic (around 13 000 BP) cemetery at Gebel Sahaba. Over 40 percent of the 59 people buried there had clearly died from arrow wounds, often multiple. The victims included men, women, and children and were sometimes buried in small groups indicating that several were killed simultaneously. This gruesome evidence as well as many other examples from around the world (for example, Kennewick Man, around 9600 BP, from Washington State and the frequent homicide victims found in Late Mesolithic cemeteries in Europe) indicate that hunter-gatherers, even early ones, were anything but pacific. Homicide, then, is at least as old as modern humanity (i.e., *Homo sapiens sapiens*) while warfare appears in the archaeological record as soon as there are sufficiently large burial populations, concentrated in cemeteries, which allow its detection.

Such weapons traumas are sometimes accompanied by other traumas from perimortem (i.e., at the time of death) mutilations and/or trophy taking. One of the most common mutilations was 'overkill' with the victim struck several times with blows or projectiles, any one of which would have been sufficient to have caused death. Some of the men buried at Gebel Sahaba had been struck by over a dozen arrows. Decapitated corpses and disembodied heads, the result of heads being taken as war trophies, have been found in sites as diverse in time and space as the late prehistoric (around AD 1300) Norris Farm cemetery in Illinois and Mesolithic (5500 BC) skull caches found in Ofnet Cave, Germany. In North America, the practice of scalping was evidenced by characteristic cut-marks left on the skull by the scalping knife. Such cut-marks have been noted on many skulls of several thousand years old. The most complete mutilations suffered by war victims results from cannibalism. However, the thoroughgoing destruction of human remains associated with cannibalism usually removes or obscures any traces of the causes of death, including violence.

Very often corpses bearing weapons and mutilation traumas were interred 'unceremoniously,' that is, without grave goods or in a manner uncharacteristic of their respective archaeological cultures. For example, the people of the Linear Pottery Culture (around 5000 BC) of central and northwestern Europe usually buried their dead individually in a flexed position in oval plan graves, with the adults often accompanied by grave offerings that included pots, stone adze/axes and awls. At Talhiem, Germany, a large pit from this period contained the sprawled skeletons of 34 men, women, and children unaccompanied by any grave goods. They had all been struck on the head, the men always several times (i.e., over-kill), with axes and adzes of characteristically Linear Pottery shape. From the size and shape of the traumas, it was inferred that the wounds were inflicted by at least seven attackers. At Crow Creek in South Dakota the remains of over 500 men, women, and children dating to AD 1325 were found heaped in the fortification ditch of a large village of that period. Over 90 percent of these unceremoniously buried individuals bore scalping marks on their skulls and many showed traumas from axes and ball-headed clubs. The houses (50 +) and the palisade of this village had been burned into the ground and the site was never reoccupied. Occasionally, the remains of war victims were accidentally buried by the collapse of burned housed or fortification walls (see Sect. 2). The large number of simultaneous deaths caused by warfare overwhelmed their compatriots capacity to give the victims a 'proper' burial, while enemies seldom give their dead foes even the most cursory burial, if any at all.

It is not always easy from weapon traumas alone, to distinguish between the victims of simple murder and the collective homicide of war. The co-occurrence of weapon traumas with evidence of mutilation, trophy taking, and unceremonious burial (especially in groups), however, distinguish war from simple murder. When such gruesome human remains are recovered from areas and time periods that yield other indications of warfare, their interpretation as war casualties is more certain. For example, the Early Neolithic (around 5000 BC) village of Herxheim, Germany, was surrounded by two concentric ditches into which had been placed the disembodied skullcaps of at least 300 individuals, some children's. Given their regular placement in the already excavated portion (about 30 percent) of the ditches, the excavators estimate that the complete ditches contain about 1,000 skulls. The only one of these skullcaps completely studied so far bore two traumas from blows by blunt instruments and one trauma from the blow of an axe. Cut marks on this cranium also indicated that the scalp had been removed. One complete body had been simply thrown into one of the ditches and abandoned. Inside the village seven graves of the usual kind for this culture (Linear Pottery), some with grave offerings, were uncovered, as were the traces of five long-houses. Like Crow Creek, Herxheim produced evidence of weapons traumas, postmortem mutilation, trophy taking, and uncharacteristic or unceremonious treatment of the dead.

Interestingly there were time periods in some regions where, despite large burial populations, we find little or no evidence of weapon traumas, mutilation, or any other evidence of warfare. For example, In the Near East just as the people were becoming farmers during the final Epi-Paleolithic and Early Neolithic (around 11 000 to 7000 BC), life seemed to have been unusually nonviolent. This case and a parallel one involving the Hopewell-Adena period (around 300 BC to AD 500) in the eastern US directly contradict the very popular idea that the origin of warfare is tied to that of agriculture. The corollary that hunter-gatherers were inherently peaceful is belied by the many very homicidal, war ridden foraging societies known to archaeology, ethnography, and history. Indeed, over the ages, the bodies of homicide and war victims show no evidence that foragers, farmers, peasants, or city dwellers, the subjects of bands, chiefdoms or states, Europeans, Africans, Asians, Amerindians, or Oceanians were always, or even generally, more peaceful or warlike than anyone else.

2. *Fortifications*

The remains of fortifications are obvious traces of warfare. Since they must be collectively constructed by groups and defend a group's persons and property, fortifications are a symptom of war rather than murder. Such defenses often feature ditches, palisades, ramparts, and walls surrounding a settlement or other desirable locations. These barriers inhibit entry, shield the defenders from the weapons of the attackers, and screen the formers' numbers and movements from the latter's sight. Fortifications may also show projecting bastions at intervals which eliminate safe 'dead zones' by allowing defenders to direct flanking fire at attackers reaching the wall. Since the gates are the most vulnerable parts of fortifications, they often show special features that make them more dangerous to attackers. The oldest and most common of such entry defenses are 'baffle gates.' The simplest form involves having the curtain wall/palisades overlap at entry points in such a fashion that attackers can only enter the interior by passing one-by-one down a narrow passage. Often these baffles required attackers to expose their unshielded (i.e., right) side to defenders on the walls. Other gate defenses include towers, deadfall pits with removable bridges, narrow chutes projecting into the interior, etc. Even very ancient fortifications remain visible on the landscape or can be detected in aerial photographs. Nevertheless, excavation is required to assess their date, the details of their construction, and the plan of their gates.

However, fortifications are stationary fixtures and only protect a very small point on the landscape. Even a simple log palisade around a small settlement requires a large amount of labor and time to construct. Thus, peoples with mobile lifestyles, portable possessions, or small social groupings seldom construct fortifications. Because of their high labor demands, even sedentary peoples may only build them when the attacks are very frequent and/or costly. Probably for all these reasons, no examples of fortifications have been found dating to the Paleolithic when even the most sedentary hunter-gatherers were mobile during part of the year. Only settled farmers and, later, very sedentary hunter-gatherers have constructed fortifications. The earliest possible fortifications (around 8000 BC), consisting of a mud brick wall and tower, were found at Jericho in the Near East. However, some archaeologists interpret the wall as for flood control and the tower as a temple. During the Early Neolithic period (around 5000 BC) in northwest Europe many villages of early farmers were defended by ditches backed by palisades (which apparently were wattle-and-daubed) and had complicated (often 'baffled') gates. Somewhat later (around 4000 BC) exactly similar features in England, UK, were in at least two instances attacked by archers, stormed, and burned. In one instance, the body of a young man was found sprawled in the ditch under the debris of the collapsed burned palisade. He had been struck in the back by an arrow and had crushed the infant under him when he fell. Similar fortifications have been found surrounding late prehistoric settlements in the Eastern US and were observed by the first European explorers and invaders (e.g. De Soto's men in the Southeast). During the Bronze and Iron Ages in Asia, Europe, and Africa, walls of stone or brick, often with very elaborate gates, defended towns and chiefly residences.

3. *Weaponry*

Weapons of war have often been indistinguishable from those used for hunting or, in the case of axes, from more prosaic implements. Indeed, many implements were used for both military and everyday functions. Weapons specialized for warfare do appear in the archaeological record. Such implements have features that make them less suited for nonhomicidal activities and/or are only found in contexts that indicate that their functions were usually, or only, military. Because of their use in the most terrible and frightening of human activities, weapons of war usually acquire symbolic significance and roles in ritual. Projectile points used primarily for war often have stems and barbs that make them difficult to extract, or increase the likelihood that they will cause an infection. For example, many war points had easily slipped hafts so that when the arrow or spear shaft is pulled out, the head stayed in the body whereas hunting points were tightly hafted. For example, one type of Late Archaic (around 3000 BC) spear point from the US Southeast was found almost exclusively associated with homicide victims' skeletons. Unlike the more common other Late Archaic point types, this one had a short stem that would have easily slipped its haft. Judging from their food remains, the Early Neolithic (around 5000 BC) farmers of northwest Europe almost never hunted, yet a common find on their sites is a distinctive triangular arrowhead that would have easily slipped from its haft. A few human remains from this period have been recovered bearing traumas from points of this type. There is no profit in killing a prey animal days or weeks later by infection but, alas, the opposite is true regarding an enemy. There is archaeological evidence that some prehistoric people designed their weaponry with this point in mind.

Axes and adzes are common archaeological finds in many regions and time periods over the past 10 000 years. They are always interpreted as woodworking tools even though there is rarely any independent or direct evidence for this assumption except their form. But some prehistoric axes and adzes obviously also had more violent functions. The stone axes and adzes of the early Neolithic people of northwestern Europe and those of the Late Prehistoric Midwestern US were certainly used to kill people as the holes knocked into the heads of the victims at Talheim and Crow Creek (see Sect. 1) attest. More problematic are instances, such as those of the Late Prehistoric (after AD 1000) Great Plains and desert Southwest of North America, where trees were rare and wood-working a necessarily unusual activity yet labor-costly ground-stone axes of 'expensive' imported raw materials were nonetheless common. There are many ethnographic instances of such axes and adzes simultaneously serving as woodworking tools, media of exchange, items of wealth, symbols of (adult male) status, and weapons of war. Just because an implement was used as a tool does not preclude it being used as a weapon of war.

Some ancient implements of copper, bronze, and iron, such as swords, halberds, daggers, and broad axes, had no other possible or plausible direct use except as weapons of war. Increasingly since 1950 it has become fashionable to interpret these artifacts as nothing more than symbols of wealth (i.e., primitive monies) or (adult male) socio-political status. The geographic rarity of metal ores and the high labor costs of their reduction to implements would enhance the prehistoric economic and social value of any metal artifacts, even prosaic tools. Metal weapons of war could only be more valuable than the latter because war and warriors were known and valued by the general populace or the ruling elites. Similarly, their symbolic values only arise from their use or possible use in the most fearsome of human activities.

4. *Representational Art*

Artistic representations of homicide and warfare have been found on rock art, painted pottery, and carved into stone, bone, or ivory. The scenes may represent dreams, myths, or hoped for events rather than real fights. The frequency of depictions of combat in the corpus of a culture's artwork may bear little relationship with the contemporary frequency of conflict. For example, military scenes became less frequent on Classic Greek vases as combat became more common during the Peloponnesian Wars. However, representations of combat do indicate that the artists were at least familiar with warfare.

Human representations of any kind are extremely rare in the famous Upper Paleolithic (around 15000 BP) cave paintings of Western Europe. A couple of these paintings depict a man apparently skewered by several spears but whether these represent victims of warfare or murder is unclear. By the Neolithic (around 4500 BC), some cave paintings from Spain clearly show military scenes. The most celebrated is a painting of combat between two small groups, numbering four and three, respectively, fighting with bows. It also depicts 'tactics' by showing a simultaneous center advance and flank attack by the larger group. In the prehistoric (so-called 'Bushman') rock art of Southern Africa (date unclear) depictions of combat and cattle raiding are not uncommon. Some of these show smaller unshielded, bow-wielding San battling larger Bantu armed with shields, spears, and throwing clubs. In Peru, decorations on Moche (around AD 500) pots show warriors armed with maces carrying trophy heads. The Classic Mayan (around AD 800) wall paintings at Bonampak, Mexico, show a battle between Mayan warriors in one panel, the torture and beheading of prisoners in another, and a victory dance in a third. While rare, many different prehistoric peoples from many different regions of the world have left depictions of warfare and its aftermath. Some of these representations were created by hunter-gatherers, some by farmers, some by tribesmen, and some by subjects of chiefs and kings. Whatever their social and economic circumstances, these ancient artists were familiar enough with warfare that they depicted it so accurately that, even today, we can see that war was hell even when fought with wooden spears or bows.

5. *Contexts for Warfare*

It has often been argued that only farmers, with their fields and sedentary villages, had territories and resources worth fighting over. However, water-holes, fishing stations, plant gathering patches, hunting territories, and the like were equally essential to hunter-gatherers, as were pasture lands and their herds to nomadic pastoralists. The arguments propounding a close, if not causal, relationship between states and warfare are more complex. The general idea seems to be that political organizations that can concentrate such physical force and economic resources as states, devoted to death and taxes, should be more bellicose than societies in which power and resources are more diffused. Both archaeological and ethnographic data completely contradict both these claims—neither the frequency nor the viciousness of warfare can be correlated with any general type of economy or form of political organization.

Another popular claim is that warfare is somehow part of our genetic code. This idea implies that warfare is a constant feature of all human life. Archaeology, ethnography, and history indicate that warfare has long been a variable condition. Some places and periods show little or no warfare while deadly fighting was frequent and intense in other cases among people of the same genetic stocks. For example, in the eastern part of the central Mississippi drainage basin, warfare was rife just before 1000 BC, almost disappeared by 300 BC, began increasing after AD 500, and peaked between AD 1100 and 1400. The socio-biological argument also implies that the bravest, most regular, and ruthless warriors should have enjoyed a higher status and left behind more descendents (thus, 'war' genes) than less bellicose men. Also, it implies that the fiercest warriors should attract or rape more women rather than wealthier, more politically adept or handsome but less militant men. However, worldwide, prehistoric societies' most elaborately buried, higher status dead almost never died violently nor bore healed weapon traumas. Conversely, most skeletons with both healed and unhealed weapons traumas have been those of young men. This evidence implies that, for many thousands of years, war's usual casualties were warriors who had had no or little chance to pass on their genes. Instead, the persons accorded the highest status and having had the longest reproductive careers usually gave little or no evidence of having been warriors. Indeed, ethnography and history indicate that the men and women who have left the most descendents were not warriors but those versed in the arts of peace—craftsman(woman)ship, trade, child-rearing, diplomacy, oratory, generosity, and the like.

With the archaeology of warfare in the early stages of a revival, cases for comparative analysis remain rare. Some archaeologists have observed that periods of intense warfare were correlated with climatic and ecological deteriorations. Such 'hard times' include a period of frequent droughts in the American Southwest, declining precipitation on the US Great Plains, and a marine die-off along the coast of Southern California. Human remains from these regions and periods often show pathologies caused by inadequate nutrition. For example, the bones of victims of the Crow Creek massacre (see Sect. 1 showed signs of chronic undernourishment, while in the Santa Barbara region signs of anemia were common during the period

of frequent warfare. Moving frontiers between different cultures may have been especially war torn. For example, most fortified Linear Pottery settlements were concentrated on the frontiers of the colonist farmers' settlement zone. Other archaeologists have noted that warfare increases during periods of political consolidation, for example, when tribes were coalescing or being incorporated into chiefdoms, chiefdoms into states, or small states into empires. Conversely, warfare seems to have increased during the dissolution of larger political units, especially states and empires. Which, if any, of these interpretations apply to particular prehistoric situations demands much more excavation and analysis. The question of why some periods in some places were so peaceful has hardly been addressed. However, with the revival of archaeological interest in warfare, the necessary work has begun.

See also: Ceramics in Archaeology; Conflict: Anthropological Aspects; History of Technology; Military Geography; Military History; War: Anthropological Aspects; War: Causes and Patterns; War, Sociology of; Warfare in History

Bibliography

Keeley L 1996 *War Before Civilization*. Oxford University Press, New York

Martin D, Frayer D 1997 *Troubled Times*. Gordon and Breach, Amsterdam

Milner G et al. 1991 Warfare in late prehistoric Illinois. *American Antiquity* **56**: 583–603

Vencl S 1984 War and warfare in archaeology. *Journal of Anthropological Archaeology* **3**: 116–32

Vencl S 1991 Interpretation des Blessures Causee par les Armes au Mesolithique. *L'Anthropologie* **95**: 219–28

L. Keeley

Conflict: Anthropological Aspects

Conflict is an action based upon the perception of partially incompatible interests or intentions between two or more persons. This perception need not be shared by both sides. What should be considered as an interest or intention requires interpretation in respect of the context. For analytical purposes these concepts often remain projections; they should, however, be grounded in observations of statements or of chains of action which are open to operationalization and falsification. They are necessarily culture-bound and may even be a matter of debate within a given culture. Furthermore, the perception of conflict does not need real intentions or real persons in order to generate action. The assumption of witchcraft, e.g., may generate virtual feuds.

1. Goals of an Anthropological Theory of Conflict

An anthropological theory of conflict can

(a) Contribute to the explanation of violence. Conflict is, however, not the only element which will contribute to an explanation of violence. Violence may, for example, be needed as a strong symbolic marker for ritual purposes. And it has to be underlined against a popular perception of 'conflict' that not every conflict is violent.

(b) It can contribute to a theory of the attraction or displacement of emotions. The hatred against foes, for example, is not something natural but is an outcome of specific constructions of conflict.

(c) The theory of conflict competes with culturalist approaches as an explanation of social cohesion. Forms of conflict regulation can reduce the relative conflict intensity between persons and, thus, create their allegiance to the system or organization which provides for this conflict regulation.

(d) It contributes essentially to a theory of social evolution. The ways by which a society regulates conflict operate also as means for the selection of alternatives for its future.

2. The History of Anthropological Research on Conflict

Anthropological research on conflict is a history of many hopes lost and few results kept. It was a classic conviction (going back to Thomas Hobbes) that a 'primitive' or 'natural' state of society would be characterized by omnipresent violence. Konrad Lorenz (1966) proposed a biological paradigm insisting on a genetic programme for violence which Ashley Montagu (1968) convincingly opposed. We find a broad variance of behavior toward potentially conflictuous situations in all different forms of human social organization. If something is encoded, it is both a potential for aggression on the one hand, and a potential for flight and conflict avoidance on the other.

Sahlins' (1961) study of the Nigerian Tiv, of 'lineages of predatory expansion,' suggested to some that genealogical closeness generally made violent conflict less probable. Barth's (1959) study on Swat Pathans has shown, however, that this pattern cannot be generalized, not even to lineage societies. Confrontation may follow from alliances against one's closest competitors, even if they are close relatives.

Types of violent conflict had been associated with different 'cultural levels' even by Malinowski (1941). It was argued that on the first levels there would be less violence. Other authors (e.g., Chagnon 1968) assumed

for these societies a specific type of resource competition as a strong motive for conflict (cf. against these Hallpike 1973). Helbling (1999), however, has shown that the occurrence and intensity of conflict is not directly correlated with different types of social organization, but rather to population density within a given type of social organization. Thus swidden agriculturalists or hunter-gatherers with a high population density may also show a high incidence of conflicts.

Marcel Mauss (1923/4) built his theory of exchange ('the gift') in order to provide a theory of peace. Relations of reciprocity which serve as a basis for exchange including market exchange are, however, compatible with relations of warring. Polanyi (1968) showed this for the seasonal warfare of the kingdom of Dahomey in West Africa.

For a long time one of the most prominent anthropological theories of conflict was Max Gluckman's (1956) theory of cross-cutting ties which would hinder the outbreak of conflict because the potential foes were always also indirectly linked by alliances. Hallpike (1977, cf. Schlee 2000) has shown, however, that cross-cutting ties may as well work the other way round, as an organization of competition, thus fostering conflict.

A stable achievement of anthropology is Sumner's (1911) ethnocentrism theory. The organization of violence requires the definition of boundaries. Leach (1965) added dryly that killing is the most powerful tool of classification—for the survivors.

Another result, not so readily identified with a single name, is the perception that factors of social organization have a higher impact upon successful strategies of violent conflict than pure technology. Although early historical work insisted strongly upon the effect of the arrow, the warrior mounted on horseback and firing arms, or even more so bomber aircraft, later research, however, stressed rather the organization of access to technology, the transmission of knowledge and training of skills and, last but not least, the social organization of risk perception. Violent conflict requires that risks are played down against potential benefits.

3. *Methodological Problems of Conflict Research*

That a human action is categorized as conflictive requires that it fulfills in its form some preestablished patterns. That pattern perception also means that contingent action without conflictive intention is likely to be interpreted as conflictive if it is harmful and if a meaning may be interpreted into it.

Societies tend to present conflict as unforeseeable or even chaotic. Sometimes conflict is even excluded from a society's self-representation. Conflict may be characterized as evil, or as a psychological and not as a social phenomenon. The less conflict is 'normalized' in institutional forms of low intensity, the more difficulties will fieldworkers face when trying to learn something about existing conflicts or the pattern of their organization and resolution. The existence of conflicts may just be denied, or existent conflicts within a group may be hidden behind other conflicts with outsiders.

4. *Conflict as Embedded Social Action*

Anthropology has challenged the popular perspective of conflict as chaotic and disorderly. The image of chaos is, however, based upon experience: Successful conflict management in the interests of the actors requires surprise, which may create the impression of unforeseeability. Conflicts follow, however, socially ordered paths. Rather like in the case of avalanches, the exact moment and the intensity are difficult to foresee, but one knows in a given environment where they may happen and which path they will follow. Anthropological research has shown that even violent conflict follows culturally encoded patterns, has institutionalized forms, and is controlled and directed in its appearance. This has been called social embedding (Elwert et al. 1999).

Embedding is the ensemble of moral values, proper norms, and institutional arrangements which set limits to a specific type of action and make simultaneously the outcome of these actions calculable. The fact that conflict has both controlled and foreseeable aspects and an element of surprise creates a hybrid structure. Therefore one may speak of partial embedding.

Conflict situations are characterized by a relatively high autonomy of time use for one of the actors. The outcome in conflicts is unforeseeable in respect of the relative gains and losses. Society, on the other hand, creates foreseeability through routines and norms which channel the realization of conflict. Norms as a means of channeling conflict have a stronger effect than just routines, but are also much less flexible than routines. It is, therefore, inherent to the character of social norms that they will be 'underlived': One should therefore not expect norms regulating social conflict to be followed one hundred percent.

The concept of norms allows us to differentiate two types of conflict: normative conflict vs. actor conflict. In the former a person or a group of persons clashes with a norm when violating other persons' or groups' rights. In the latter individuals or groups clash and try to harm each other in a field of action which has normatively defined boundaries but where neither side can claim norms protecting its goals.

5. *Forms of Actor Conflict*

The multitude of concepts used in order to describe conflicts (war, genocide, competition, feud, legal battle

	More violence v	Less violence v
Stronger embedding >	warring	procedure
Weaker embedding >	destruction	avoidance

Figure 1
Differentiations of conflict

etc.) can be ordered in a field with four poles (see Fig. 1). To speak of poles means that the empirical cases and emic categories may be positioned between these abstract ordering categories. The criteria are more or less violence, and stronger or weaker embedding. The category 'warring' includes both feud and war. A war is a violent conflict which includes physical risks but is limited by some rules. A feud is a violent conflict with stronger embedding and in general also less violence than wars. It is defined as a violent conflict where the potential victims, the place of conflict, the time of conflict, or the weapons allowed may be the matter of a binding agreement. A necessary condition for the definition of feud is that the means to end the conflict is part of the agreement. Feuds and competitions cannot be clearly separated, as all these categories have a fluid transition. Competitions, however, should rather be seen as being part of procedures. Procedurally organized competition, measured in abstract parameters (e.g., money or speed), may allow for an almost continuous conflict without violence.

Whereas in legal studies procedures are defined by a specific—ritual—form, anthropology as a comparative science cannot only rely upon conceptualizations defined by the actors. An anthropological concept of procedure, therefore, has to take up suggestions formulated by Niklas Luhmann (1969). A procedure is distinct from daily interaction by its form and ordered sequence of action. Elections, court cases, and auctions have also been seen as conflict resolution procedures. During a procedure some power differentials are suspended. It may, for example, be excluded that a stronger person physically harms a weaker person during the procedure. A procedure has an outcome with consequences in action. Who will profit from the outcome is, in principle, open. It should be called procedure only when it ends in a conclusion which has meaningful consequences for action.

The conclusive procedure has to be distinguished from the pending procedure: something which has the ritual form of the procedure but which never comes to conclusions which are consequential. Thus, the pending procedure is not a way of conflict resolution but a way of conflict perpetuation. Pending procedures can have a high annoyance potential and are rather forms of warring with reduced violence.

Destruction (e.g., genocide) is distinguished from warring as the goal is the total annihilation of the other. There is no agreement which binds both partners together. Nevertheless, the ways the other side is destroyed follow a culturally encoded pattern. There is no human society which allows killing that does not simultaneously have some rules for 'proper' killing.

Avoidance is a concept rarely included in theories of conflict (Radcliffe-Brown 1952), but to have a complete conceptual structure we need to consider also this 'zero-case.' In some societies avoidance is the dominant mode of dealing with conflict (see Elwert 2001 on the Byalebe of West Africa). In a situation where avoidance is dominant it is more probable that people switch to destruction than that they switch to procedure.

6. *The Normative Conflict*

A normative conflict is a conflict in which a person's behavior toward another person or to a thing, which is meaningful to a person, conflicts with a norm. In this context, norm is meant in the strict sense; it requires both monitoring and sanction. Prescriptive or prohibitive formulas without a sanction probability should be considered moral values, not norms. This definition is different from those given within societies. There are, in general, more prescriptive or prohibitive formulas which are considered 'norms' than those whose infraction is really prosecuted.

The monitoring is necessary in order to create a sanction potential. Monitoring is a sensitive element of this process. In some societies monitoring is restricted to specialists or to persons with a recognized claim to competence (police, priests, soothsayers).

Reactions to norm infraction range over a wide spectrum. In state societies the dominant mode of reaction is typically the sanction by a central specialized apparatus, the coercion instance. Another solution is the 'sanction-service.' In state societies this may appear parallel to the centralized apparatus; it is then often considered to be illegitimate, for example as a Mafia (Krauthausen 1997). In stateless societies there may be specialists for sanction services, organized as corporations, age grades, or secret societies. A third option is self-help, which may take a form similar to feuding (Spittler 1980). Self-help requires some kind of broader agreement that a norm has been transgressed. This can be a court decision which then legitimates the persons considered as victims of the norm infraction to take action or to hire specialists for

this purpose (cf. Würth 2001 on shari'a courts in Yemen). A fourth option, which again is a zero-option, is ignoring the conflict. If there are no strong central instances or sanction services available and if the dividing line between the actor conflict and normative conflict is blurred, this is a common solution.

Sanctions are not only physical sanctions (as corporeal harm or mutilation, restriction of movement, or material and economic fines) but also loss of status or attribution of negative reputation. The reputation sanction is very common in those contexts where a person can hardly escape contact with those who share the negative evaluation of the culprit. Since societies are never homogeneous, subgroups may not share specific norms (Pospisil 1958). The reputation sanction is then limited in its efficiency. The redundancy of overlapping reference groups, each with a sanction capacity, can cope with that (Elwert 1980).

The normative conflict is a two-tier process. The first element is the infraction, the second is the sanction process. That the infraction is an infraction is, however, only definitely clear after the sanction process. The working of the sanction process may be expeditive (e.g., lynching) or procedural. The expeditive reaction is based upon a single decision or a turning point from when on a plurality of single decisions is valid enough for action. All human societies—not only the 'modern' ones—know, however, the advantages of procedural forms of sanctioning over the expeditive ones. Truth may be a matter of disagreement, and the loss or investment of time characteristic of procedures may allow information to be validated and emotions to cool down. Procedures may, however, also be used for the alternative purpose of steering the multiplicity of perceptions into just one established truth, and to stir up the emotions in such a way that sanctions become more probable or more violent.

7. *The Attraction and Displacement of Emotions*

To live with emotions and act upon them requires some societal agreement. The popular notion that conflict is motivated by emotions does not hold against anthropological evidence—at least insofar as recurring conflicts are studied. People hardly react to emotions, but to assumed or perceived goals of action. The reaction in a conflict situation has higher strategic qualities if it is not based upon emotion but upon cool calculation. To 'show' emotions may, however, be very important in order to legitimate one's action in a conflict situation, or in order to impress the adversary.

Actions in conflicts, however, can attract emotions. This is correlated with changes in the structure of relevance; it may help to concentrate resources to the conflict. Actor conflicts are strong attractors of emotion, even more so if violence is implied. In any procedural treatment of conflict, and especially in normative conflicts, emotions can be less intense. They may appear comparatively detached or displaced. A society which gives preference to normative conflicts and which privileges the procedural treatment allows for the displacement of emotions. The literary representation of conflict or sport competitions—a form of conflict with lesser implications to one's personal life-goals—may then attract stronger emotions than those linked to lived conflict.

8. *Conflict Motives*

Motives to enter conflicts can be grouped into three broad categories: honor, power and material gain. Each of these three general motives has its own definition of who won in a conflict situation. What, for example, an outside observer may see as an economic loss, may appear to the insiders as winning prestige. In most cases, however, there are several motives overlapping, and this can make it difficult to define who won in a given situation. To distinguish between these three categories of motives is, however, helpful in order to perceive different consequences of their priority for the respective organization of a conflict. If honor is the major goal—and even more, if honor may be even gained by a dead person declared to be a hero—then the propensity to risk increases. This may create patterns of recurrent warring, since new generations need new prestige (cf. Evans-Pritchard 1957 on Azande warfare). In the case that power, the control over persons, is the goal, then the end of the conflict has to be such that the winning side retains a coercion potential over the losing side. If material gain is the goal, then it may be reinvested, in a self-perpetuating process, to gain even more.

Systems of blood feud or vengeance are often dominated by a system of attribution of honor and shame. To kill a person may create honor. If this killing is not retaliated, this creates shame. Vengeance 'washes away the shame' and reestablishes honor (Schwandner 1999).

Behind a smokescreen of ideological motives, warlords exhibit a very clear priority for motives of economic gain, whether the means are peaceful or violent. This entrepreneurial activity creates a market without the typical embedding for peaceful markets, a market of violence (Elwert 1999). Whether warlords subjectively give priority to noneconomic goals or not, they need weapons and soldiers (even if they call them volunteers) in order to reproduce their activity. The gains from robbery, blackmail, hostage-taking, customs and trade (especially in valuable items such as precious metals or stones, drugs, and weapons) have to cover expenses for the violence and generate a profit.

Group conflicts may originate in divergences of interest between the groups or they may be produced by the construction of such contradictions. In the

latter case the purpose is a strategy which aims primarily upon the control of power within that group, an endostrategic mobilization. In such a process the outside conflict is heated up in order to homogenize the group, to push forward new norms or structures of leadership. Boundaries to 'the others' are created or become accentuated. Endostrategic mobilization is typically the instrument of would-be elites who want to gain power and therefore portray the actual elite as persons who do not sufficiently defend collective interests, or as agents of outside forces.

9. *Conflict Regulation and Cohesion*

It seems at first sight paradoxical to associate the social organization of conflict positively with social cohesion. Since anthropologists no longer believe that there is a natural harmony among human beings, however, we cannot escape the notion that the normalization and institutionalization of conflict in specific forms is an essential condition of social cohesion. Divergence of interest is omnipresent among human beings. If there is no channeled way to express it, then random conflict and—even more so—avoidance reduce chances for cooperation. Cooperation requires not that conflict is repressed but that it finds predictive ways of expression. Less destructive forms of conflict and especially forms which maintain both sides' interests in form of compromise ease the readiness to enter conflicts. They do so because the lower risk threatens less the conditions of cooperation during or after the conflict. Thus, an institutionalization of conflict may increase conflict incidence if conflict becomes less violent within the respective entity.

A social group with a common reference is not necessarily a 'conflict group,' a group capable of acting collectively against one other. A conflict group requires an internal sanction capacity. Without it, collective action would be undermined by freeriders who want to enjoy the benefits of a victory but avoid the risk of the conflict itself. The creation of such a conflict capacity through an internal sanction system is an important indicator of a mobilization potential for civil strife and civil war.

We can observe that we-groups, organizations, or societies tend to reduce the violence of conflict within their respective boundaries. But this tendency does not create a general pattern. As cooperation creates resources, material or immaterial (such as honor), these resources may motivate further conflict. Thus, the violence of conflict within groups may be stronger than the violence between them. As long as these conflicts do not turn into a systematic destruction of a given subgroup, the entity's cohesion may be maintained nevertheless.

The economist Albert Hirschman (1994) has observed something about modern societies which probably can be generalized to other societies as well: Group cohesion is strongly correlated with the access to accepted modes of conflict regulation. Shared cultural values, even if they are as emotionally binding as language or religion, do less to explain social cohesion than the presence of institutions which appear to give a fair chance to realize one's interests with reduced destructive risk. From an anthropological point of view one has to add that procedures create a higher cohesion than other channeled ways of living a conflict. Hirschman has further shown that those procedures which favor more-or-less solutions (incremental conflict resolution in our terminology) have a better chance to create cohesion than those with either–or structure (alternative conflict resolution). The compromise maintains elements of interest of both sides. This eases the identification with the outcome of the conflict regulation. This perspective has been applied in an anthropological perspective to modern nations. It could be shown in one case that irrespective of cultural parameters such as language or religion, it is the deprivation of access to normatively regulated institutions of conflict resolution which further processes of fission and separatism (Elwert 1997). For African states it has been shown that the classical ignition spark for civil war is the arbitrary use of violence by state agents (Wirz 1982).

10. *Institutions of Conflict Regulation as Central Determinants for Processes of Social Transformation*

Ralf Dahrendorf (1954) has argued that the forms of conflict regulation are central for the difference between paths of social transformation. The normative structure most often (but neither always nor exclusively) ensures a society's reproduction and its sameness over time ('identity'). Social change requires breaks in the reproduction of sameness. It can, however, not proceed by 'anything goes,' since the end of reproduction would undermine further change. Stimuli for change, whether they are endogenous or exogenous, have to be treated as alternatives to something, and not as unrelated events. Alternatives are never abstract, they imply persons' intentions. This gives the organization of social conflict the key role in social change to conceive and select alternatives.

Societies which give preference to the normative conflict increase the foreseeability of one's own actions. As long as one follows the norms, the risk of conflict is reduced. But this produces overall a higher conflict-processing capacity. There is less relatively randomized conflict and thus more possibility to engage into ordered conflict. A higher conflict-processing capacity means more variation. If the society uses these institutions in order to select not only winners and losers, but also in order to select elements of social order, forms of consumption, and new

products of human activity, then innovation can gain acceptance. It is the higher selectivity which reduces the fear of new elements of social order.

It cannot be stated, however, that such a society would be 'more highly developed' than others. Depending upon the type of social environment, different types of conflict regulation appear as optimal adaptations (Hallpike 1973). Avoidance-biased societies are among the most expansive types of social organizations in an environment characterized by raiding, especially slave raiding. Avoidance produces a high dispersion and thus hinders the population concentrations necessary to make slave raiding a profitable enterprise, e.g., in some regions of the Sahel and the Sudan. At the margins of aggressive imperialisms, feud-biased societies appear to be one of the most stable arrangements. These societies keep internally a high level of warring competence which act as deterrent or repellent to these imperialist forces—as can be shown in the case of the Caucasus or some regions of the Balkans. If societies interpenetrate each other by means of a reciprocal competition on their markets, a high innovation capacity gives an edge to this competition. Procedure-biased societies appear to be especially strong in this respect, e.g., some medieval oriental societies and some modern Western ones.

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict and War, Archaeology of; Conflict/Consensus; Conflict Sociology; Malinowski, Bronislaw (1884–1942); Mauss, Marcel (1872–1950); Simmel, Georg (1858–1918); Violence in Anthropology; War, Sociology of

Bibliography

Barth F 1959 Segmentary opposition and the theory of games. A study of Pathan organization. *JRAI* **89**(1): 5–21

Chagnon N 1968 *Yanomamö: The Fierce People*. Holt, Rinehart & Winston, New York

Dahrendorf R 1954 Out of utopia. *AJS* **64**: 115–27

Elwert G 1980 Die Elemente der traditionellen Solidarität. *Kölner Zeitschrift für Soziologie* **4**: 681–704

Elwert G 1997 Deutsche Nation. In: Schäfers B, Zapf W (eds.) *Zur Gesellschaft Deutschlands*. Leske & Budrich, Opladen, Germany, pp. 123–34

Elwert G 1999 Markets of violence. In: Elwert G, Feuchtwang S, Neubert D (eds.) *Dynamics of Violence. Processes of Escalation and De-escalation in Violent Group Conflicts*. Duncker & Humblot, Berlin, pp. 85–102

Elwert G 2001 Herausforderung durch das Fremde. In: Fikentscher W (ed.) *Begegnungen und Konflikt*. Beck, Munich, Germany

Elwert G, Feuchtwang S, Neubert D 1999 The dynamics of collective violence. In: Elwert G, Feuchtwang S, Neubert D (eds.) *Dynamics of Violence. Processes of Escalation and De-escalation in Violent Group Conflicts*. Duncker & Humblot, Berlin, pp. 7–31

Evans-Pritchard E 1957 Azande warfare. *Anthropos* **52**: 239–62

Gluckman M 1956 *Custom and Conflict in Africa*. Blackwell, Oxford, UK

Hallpike C 1973 Functionalist interpretation of primitive warfare. *Man* **8**: 451–70

Hallpike C 1977 *Bloodshed and Vengeance in the Papuan Mountains*. Oxford University Press, London

Helbling 1999 The dynamics of war and alliance among the Yanomami. In: Elwert G, Feuchtwang S, Neubert D (eds.) *Dynamics of Violence. Processes of Escalation and De-escalation in Violent Group Conflicts*. Duncker & Humblot, Berlin, pp. 103–16

Hirschman A 1994 Social conflicts as pillars of democratic market society. *Political Theory* **22**(2): 203–18

Krauthausen C 1997 *Moderne Gewalten*. Campus, Frankfurt Main

Leach E 1965 The nature of war. *Disarmament and Arms Control* **3**: 165–83

Lorenz K 1966 *On Aggression*. Harcourt, New York

Luhmann N 1969 *Legitimation durch Verfahren*. Luchterhand, Neuwied am Rhein, Germany

Malinowski B 1941 War—past, present and future. In: Clarksen J, Cochrane T (eds.) *War as a Social Institution*. Columbia University Press, New York

Mauss M 1923/24 Essai sur le don. *Année Sociologique* **1**: 30–186

Montagu A (ed.) 1968 *Man and Aggression*. Oxford University Press, Oxford

Polanyi K 1968 *Dahomey and the Slave Trade*. Montagu, Seattle, WA

Pospisil L 1958 *Kapauku Papuans and Their Law*. Yale University Press, New Haven, CT

Radcliffe-Brown A 1952 *Structure and Function in Primitive Society*. Cohen & West, London

Sahlins M 1961 The segmentary lineage: An organization of predatory expansion. *American Anthropologist* **63**: 322–45

Schlee G 2000 Identitätskonstruktionen und Parteinahme: Überlegungen zur Konflikttheorie. *Sociologus 2000* **1**: 64–89

Schwandner S 1999 Humiliation and reconciliation in Northern Albania. In: Elwert G, Feuchtwang S, Neubert D (eds.) *Dynamics of Violence. Processes of Escalation and De-escalation in Violent Group Conflicts*. Duncker & Humblot, Berlin

Spittler G 1980 Konfliktaustragung in akephalen Gesellschaften. *Jahrbuch für Rechtssoziologie und Rechtstheorie* **6**: 142–64

Sumner W 1911 *War and Other Essays*. Yale University Press, New Haven, CT

Wirz A 1982 *Krieg in Afrika—Die nachkolonialen Konflikte in Nigeria, Sudan, Tschad und Kongo*. Steiner, Wiesbaden, Germany

Würth A 2001 *As-Sari 'a fi Bab al-Yaman. Recht, Richter und Rechtspraxis*. Duncker & Humblot, Berlin

G. Elwert

Conflict/Consensus

Conflict refers to a situation in which there is disagreement over how to divide scarce resources. Conflict resolution thus decides who gets what (Lasswell 1936), and politics is the process of raising and settling this question. An individual experiences inner conflict

when he or she faces a choice between incompatible values. This psychic conflict may have political consequences, as when a voter is torn between lower taxes and more spending, but political scientists generally study conflict situations where two or more individuals, groups, political parties, or nation-states seek the same object or maintain incompatible goals. The stakes involved in a specific conflict thus vary widely.

Although conflict refers to disagreement and consensus to agreement among political actors, they are not mutually exclusive. Without agreement about what is desirable, there could be no conflict over its possession. Contact and interaction also are necessary for conflict; there can be no conflict between an earthling and the man in the moon. It also is important to distinguish between a *conflict situation* and the *behavioral responses* of the parties involved. The potential responses range from peaceful negotiation to unlimited aggression and violence, and outcomes vary from a stable compromise to total victory and destruction. For example, two states claiming the same territory may settle the conflict by partition or by war leading to occupation of the disputed land by the victor.

Since political conflict involves a struggle for advantage, generally arising from a challenge to the status quo, strategic behavior is a natural focus for analysis. The rational choice paradigm imported from economics (Elster 1986) and codified in game theory (Von Neumann and Morgenstern 1944, Axelrod 1984) represents an effort to explain how contending parties 'should' behave in zero-sum and nonzero-sum games (conflicts), two and n-person games, games with or without full information, and so forth. More psychologically oriented scholars (Simon 1957, Kahneman et al. 1982, Janis 1982) have emphasized the cognitive and motivational impediments to 'pure' rationality.

1. *Normative Aspects of Conflict*

Scarcity and selfishness assure that conflict is a perennial feature of society. Weber (1949) noted that 'conflict cannot be excluded from social life,' regarding peace as a momentary pause in the flow. Because conflict is disruptive and inevitably results in someone's disappointment, many theorists regard it negatively and stress the need for harmony and common values (Parsons 1952, 1960). Even Marx, for whom class conflict was the key explanatory variable in social development, rhapsodized about how the proletarian revolution would replace the government of people by the administration of things.

On the other hand, Simmel ([1908] 1955), Coser (1956), and Dahrendorf (1959) emphasize the positive functions of conflict. Conflict binds members of one contending group together, forming bonds of solidarity and fostering cooperation, even as it fuels antagonism toward the opponent. In fact, social identity theory in psychology (Tajfel and Turner 1986) posits that mere membership in a social category is sufficient to foster ethnocentric thinking. Furthermore, the appearance of harmony often masks entrenched social injustice, so conflict may be a necessary spur for the moral improvement of society. Finally, conflict may result in innovation, as when debate and negotiation stimulate the development of reforms that all contending parties accept.

When interactions among antagonists or potential antagonists persist over time, prudence tends to dictate the establishment and maintenance of reciprocal norms and understandings concerning conflict behavior. For example, there are international treaties concerning diplomatic immunity and the treatment of prisoners of war and refugees. Other limiting rules, such as a norm against assassinating leaders of other countries are more informal. Both the formal and informal arrangements confining conflict tend to evolve as the intentions and relative capabilities of the contending actors change.

Agreement on the rules of the game domesticates conflict by legitimating outcomes reached through accepted procedures. This insight leads scholars to argue that consensus on constitutional principles is a prerequisite for democratic stability (Lipset 1960). This consensus can develop in a number of ways. It can be the outcome of negotiations, of a recognition that the alternative to institutional arrangements limiting conflict may be violence and the risk of permanent loss, or, in a stable society, of socialization into a stable political culture.

2. *Foundations of Conflict*

Politics is the attempt to use state power to achieve a favorable distribution of scarce goods or values. Conflicts over such allocation frequently are termed realistic or objective and tend to group people according to shared positions with respect to the desired value. Perhaps the most common basis for conflict is economic, pitting town against country, worker against manager, and so forth. However, conflicts among religions, races, and cultures also are pervasive and often are more intense than conflicts over economic issues because they seem to touch more directly on people's fundamental sense of identity. Lines of political cleavage therefore reflect the structure of society, and as demographic, economic, and technological changes occur alignments shift and new layers of conflict are added to the old (Lipset and Rokkan 1967). Political party systems articulate the objective bases of conflict by aggregating coalitions of interests.

Since there are multiple sources of division in any complex society, the degree to which cleavages are cross-cutting rather than overlapping is important. Where religious and economic fault-lines coincide as,

for example, in Northern Ireland where Catholics tend to be lower on the income and occupational ladder than Protestants, conflict tends to be more intense and peaceful compromise harder to reach. By contrast, where the connections between religion or language and economic position are loose, individuals tend to have multiple affiliations and groups who are antagonists in one conflict may be allies in another, a circumstance that may moderate their behavior (Horowitz 1985).

Political scientists have contrasted instrumental with expressive conflicts. The former type refer to the pursuit of specific, tangible goals such as political office or economic gain. By contrast, expressive conflicts seek the redistribution of group recognition and esteem, and their outcome may be to supply psychic satisfaction for the many without changing their objective social circumstances (Edelman 1964). In the United States, ethnic group competition for social and political status hampered the development of class-based parties (Lipset and Marks 2000). Conflicts over collective ideals or a group's way of life are more likely to be infused with great emotion than struggles for personal advantage, and participants in such battles between 'good' and 'evil' are more likely to resort to aggressive or ruthless tactics (Coser 1956).

3. *The Impact of Consensus on Conflict*

Political consensus exists when a large proportion of the members of a society agree about how decisions regarding the distribution of values should be made. Such agreement may center on outcomes or on the procedures to reach them. The idea of political consensus has been used primarily to study the stability and effectiveness of democratic political processes. In that sense, a common hypothesis is that the existence of consensus mitigates the intensity of conflicts regarding the allocation of scarce resources. The function of consensus in this view is to maintain civil peace by reducing the probability of disagreements and by fostering a sense of mutual restraint among contending parties.

The crucial features of consensus refer to general principles and values rather than to preferences regarding specific policies. The degree to which there is common acceptance of existing institutions as legitimate, agreement that the society's political laws and rules are fair, and a shared sense of identity is the focus of empirical studies of political consensus. Political theorists diverge about the elements of social structure and culture that foster agreement on the allocation of values such as power, wealth, and status. Economic inequality and rigid barriers to upward mobility generally are viewed as impediments to consensus (Lipset 1960) whereas cultural unity often is regarded as facilitating agreement on political values (Schlesinger 1991). In traditional societies, religious unity tended to promote social harmony, but modern states usually have a plurality of religious and ethnic backgrounds, with nationalism emerging as an integrating ideology (Haas 1964). More generally, advocates of consensual patterns of belief and identity view these as a lubricant that enables political liberty and social complexity to coexist.

4. *The Relevant Contours of Consensus*

Social change poses a challenge to the prevailing political consensus by introducing new conflicts over the distribution of values. Rising groups seek power and recognition, while declining groups fight against displacement. As these conflicts are settled, the content of the political consensus shifts as resources are redistributed. So, for example, the general acceptance of welfare state policies mitigated class-based political conflict in industrialized democracies. A new consensus emerged over what benefits were untouchable and *hors de combat*. Similarly, the successes of the civil rights movement simultaneously produced consensus on the illegitimacy of official racial discrimination and initiated dissensus over the proper role of government in promoting equality of results rather than opportunity.

Since there is no natural harmony of interests among groups in a differentiated society, political consensus will always be incomplete. A certain number of people will always feel the prevailing norms are unjust. In addition, the belief systems of most people are only loosely integrated. There may be a consensus of values, such as liberty and equality, that collide in concrete situations without awareness of the ultimate need to make trade-offs. Furthermore, those sharing a political outlook differ in the intensity of their beliefs and in their level of political engagement. Hence, the significance of consensus for the government's effectiveness and stability depends not just on how widespread is popular agreement but on which groups in society have shared commitments and values.

In a study of US political culture, McClosky and Zaller (1984) found a broad consensus (defined as 75 percent agreement) within the general public on abstract principles of democracy and free enterprise. Agreement on the principle of minority rights tended to erode, however, when questions centered on what should be allowed to unpopular groups. There also was less consensus on questions of economic and social equality. Previously, Dahl (1961) and McClosky (1964) argued that consensus among the politically active was more significant than disagreement with democratic norms among the general public who often are apathetic and whose choices are structured by elites. The acceptability of existing institutions and important government policies to most significant social groups thus matters more than the universality of agreement within the entire population. In par-

ticular, since consensus within groups can enhance conflict across groups, it is important for agreement on fundamentals to encompass leaders of all the major political parties in order to encourage collaboration in the peaceful adjudication of conflicts.

There is, then, a reciprocal interaction between conflict and consensus. When disagreement over the allocation of values is relatively circumscribed, it is easier for feelings of unity and a common fate to develop. At the same time, agreement about what is fair and about accepted ways of pursuing one's interests facilitates settling conflicts of interest. This suggests that the significance of political consensus as a unifying or stabilizing force varies with the severity of the challenges confronting society. The benefits of consensus for democratic governments are felt mainly during times of external threat or internal convulsion.

5. *Creating Consensus*

Since political conflict never disappears, the problems of creating and sustaining consensus are ongoing. One source of consensus is the development of mutual trust among political actors through incremental steps of cooperation (Axelrod 1984). Another is the incorporation of previously disadvantaged groups into the polity through a combination of political and economic changes. Finally, consensus develops through the process of socialization that transmits values from one generation to the next. Families and schools are important agents in this process. The degree to which political regimes self-consciously strive to inculcate a supportive consensus varies, but the increasing difficulty of controlling access to divergent values makes the success of such efforts fragile.

See also: Cleavages: Political; Conflict and Conflict Resolution, Social Psychology of; Conflict: Anthropological Aspects; Conflict: Organizational; Conflict Sociology; Cooperation: Sociological Aspects; Integration: Social; Internal Warfare: Civil War, Insurgency, and Regional Conflict; Marx, Karl (1818–89); Political Culture; Simmel, Georg (1858–1918); Solidarity, Sociology of

Bibliography

Axelrod R M 1984 *The Evolution of Cooperation*. Basic Books, New York

Coser L A 1956 *The Functions of Social Conflict*. Free Press, Glencoe, IL

Dahl R 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT

Dahrendorf R 1959 *Class and Class Conflict in Industrial Society*. Stanford University Press, Stanford, CA

Edelman M J 1964 *The Symbolic Uses of Politics*. University of Illinois Press, Urbana, IL

Elster J (ed.) 1986 *Rational Choice*. Blackwell, Oxford, UK

Haas E B 1964 *Beyond the Nation-State: Functionalism and International Organization*. Stanford University Press, Stanford, CA

Horowitz D L 1985 *Ethnic Groups In Conflict*. University of California Press, Berkeley, CA

Janis I L 1982 *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin, Boston

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK

Lasswell H D 1936 *Politics; Who Gets What, When, How*. Whittlesey House, McGraw-Hill, New York

Lipset S M 1960 *Political Man: The Social Bases of Politics*. Doubleday, Garden City, NY

Lipset S M, Marks G 2000 *It Didn't Happen Here: Why Socialism Failed in the United States*. Norton, New York

Lipset S M, Rokkan S (eds.) 1967 *Party Systems and Voter Alignments: Cross-National Perspectives*. Free Press, New York

McClosky H 1964 Consensus and ideology in American politics. *The American Political Science Review* **58**(2): 361–82

McClosky H, Zaller J 1984 *The American Ethos: Public Attitudes Toward Capitalism and Democracy*. Harvard University Press, Cambridge, MA

Parsons T 1952 *The Social System*. Tavistock Publications, London

Parsons T 1960 *Structure and Process in Modern Societies*. Free Press, Glencoe, IL

Schlesinger A M Jr 1991 *The Disuniting of America*. Whittle Direct Books, Knoxville, TN

Simmel G [1908] 1955 *Conflict; The Web of Group Affiliations*. Free Press, Glencoe, IL

Simon H A 1957 *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in Society Setting*. Wiley, New York

Tajfel H, Turner J 1986 The social identity theory of intergroup behavior. In: Wrochel S, Austin W (eds.) *Psychology of Intergroup Relations*. Nelson Hall, Chicago

Von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

Weber M 1949 *Max Weber on the Methodology of the Social Sciences*, [Shils E A, Finch H A eds. and trans.]. Free Press, New York

J. Citrin

Conflict: Organizational

We make two explicit assumptions in this article. First, a life conducted in, or mediated through, organizations is one of the defining aspects of being a modern citizen. Second, and despite the best intentions of some theorists and practitioners, relations in organizations will always entail a degree of conflict and political maneuvering because '[r]easonable people can be expected to disagree on major, uncertain and ambiguously defined issues, and to argue strongly for their

personal convictions' (Buchanan and Badham 1999, p. 21). As a corollary of these assumptions it is likely that we have all been a participant in, or observer of, organizational conflict at some time and, in this respect, we may have an intuitive feel for what creates and diffuses conflict in organizations. In this article, however, we adumbrate the key characteristics of the following perspective on organizational conflict: (a) the orthodox social-psychology approach; (b) ethnographies of conflict that incorporate the social, cultural, and historical dimensions of organization; (c) neo-Darwinian studies that see conflict as an expression of our essential human nature; and (d) a neo-Machiavellian approach where conflict reflects the politicized character of organizational relations. A general theme of this article is that we should be wary of explaining conflict on the basis of universal or abstract laws of behavior at the expense of a subtle consideration of organizational history and context.

1. *Organizational Conflict and Social Psychology*

Orthodox social psychological approaches explore conflict through a hierarchy of dyadic interactions that start with the interpersonal and stretch to relations between discrete social entities, such as interest groups, unions, and even nation-states. The ways in which these interactions are played out can be mapped using the dimensions of *Process* and *Structure*, thus delineating the two main social psychological models of organizational conflict (Thomas 1976, 1994). The *Process* model is primarily empirical in approach, as it requires us simply to observe and record conflict as it unfolds. This involves categorizing various generic stages before relating these dynamic circumstances to particular outcomes or 'settlements.' These outcomes tend to be classified in economic terms so that resources—frequently the focus of the conflict in the first place under this approach—are distributed between parties, usually in a zero-sum arrangement. In contrast, the *Structural* model involves identifying the antecedents that shape the dynamics of a particular dyadic conflict. These include cognitive factors (for example, the actors' awareness of issues and possible settlements, their intentions, etc.) and affective states such as emotions and empathy. Underlying the operation of these structural features of conflict is the assumption of bounded rationality: on the basis of the information available to them, actors in a conflict will move toward satisfying outcomes. Thus, as Thomas (1994) suggests, conflicts are decision-making processes that evaluate possible settlements using a combination of rational-instrumental considerations (what in this for *me*?) with normative considerations (is this a 'good' outcome?).

In practical terms, the *Structural* and *Process* models are inextricably bound as the former provides the theoretical basis for the latter. Regardless of the aggregation of the unit (e.g., individual vs. individual, individual vs. group, group vs. group, etc.), this theoretical/empirical dependency remains intact to the extent that any dyadic conflict can be considered as:

> … a process which includes the perceptions, emotions, behaviors, and outcomes of the two parties. Specifically, in order to differentiate conflict processes from other processes … conflict is the process which begins when one party perceives the other has frustrated, or is about to frustrate, some concern of his (Thomas 1976, p. 891).

Thomas' use of the third person singular pronoun, in generic terms, unwittingly reveals two major assumptions underpinning the psychological approach. First, that we can empirically identify an entity that, much like a self-determining individual, possesses a coherent set of singular, stable, and universal interests, intentions, or affective states. Second, we can also empirically and unambiguously determine these specific psychological characteristics. The effect of these assumptions is to anthropomorphise *all* conflict relations to the extent that, at any level of aggregation, they will tend to follow certain psychological principles premised on the simplest person-to-person dyad. Beyond this obvious conceptual reductionism there are also strong implicitly functionalist assumptions, notably that the normal state for any collective entity is a healthy coherence that will be encouraged by good husbandry of those conflicts that do occur. Such assumptions give rise to the view that, so long as conflict is managed effectively using rational principles, we can keep the lid on organizational conflict and even use it to the organization's advantage.

Thomas (1976) offers three reasons why conflict can be functional for organizations. First, there may be a lack of tension in an organization that can create boredom and complacency while, in contrast, a certain level of conflict may stimulate creativity. Second, the existence of conflicting priorities and, hence, ideas may make for better solutions to organizational problems. Third, aggressive behavior need not necessarily be destructive, especially when it initiates conflict resolution procedures. Thomas also notes that, as beneficial side effects of conflict, organizations can identify systemic equity problems, address internal power imbalances, and, in situations where the whole organization is united in conflict with another, improve its internal cohesiveness.

The social psychological approach has spawned a whole industry of expert theory and practice that seeks better methods of containing conflict within the limits of acceptability. Generally speaking, however, two main methods of conflict containment are open to us: the resolution of conflict through negotiation or the regulation and minimization of conflict through

governance structures. Psychological studies have emphasized the former option to such an extent that the conceptual terrain of organizational conflict is now dominated by ideas of how to negotiate our way out of it! An indication of the importance of this shift of emphasis can be obtained by comparing the entry on 'Conflict' in the first edition of the *Handbook of Industrial and Organizational Psychology* (Thomas 1976) and the revised entry in the second edition (Thomas 1994). In this latter contribution the validity of the *Process* and *Structural* models is largely taken 'as read' and most of the space is devoted to outlining subsequent developments in negotiation theory and practice, especially the role played by third-party technocrats in bringing about conflict resolution. Social psychology perspectives typically provide little critical reflection on the adequacy of their models or on the validity of their behavioral assumptions. These shortcomings are particularly troubling: representing *all* organizational conflict as if it took place through the operation of a dyad fails adequately to capture the complexities of the social relations found in organizations. In the light of this, we shall now turn our attention to alternative treatments of conflict.

2. *Ethnographic Approaches to Organizational Conflict*

According to Barley (1991), one common misapprehension about the nature of organizational conflict is indicated by our tendency to 'bracket' it. By this Barley means that we see conflict as somehow being outside the usual day-to-day events of organizational life. It is to be seen as a special case to be treated in special ways. Barley considers this bracketing to be, indirectly, an artifact of the psychological discipline of conflict and negotiation studies: seeing the world through eyes used to studying dyadic negotiations in laboratory settings predisposes us to look for self-contained conflicts with well-defined antecedents and a limited life span. Barley argues that we need to bring the study of conflict back from the margins and place it at the heart of the quotidian experience of organizations. This means rejecting ideas of utility maximization and individual cognition in favor of an anthropology of specific conflict processes; an approach that involves developing an 'ethnography of disputing.' Barley regards the advantages of this approach to be the historicizing and contextualizing of the eternal conflict between organizational actors trying to make sense of their position in a negotiated order. Barley's position relies heavily on the work of Anselm Strauss and his earlier studies of the unstable divisions of labor found in many organizations. Strauss and his collaborators found that, rather than following externally imposed standardized rules and procedures, individual tasks and duties were more the product of bargaining and negotiation between the staff themselves (see *Organizations: Negotiated Orders*; *Individual Identities in Organizations*).

In general, an anthropological approach to organizational conflict places a good deal of emphasis on the social/cultural context in which the disputants find themselves. These include: socially constructed identities (especially 'insider' vs. 'outsider' status); the size and complexity of networks of dependency relations; the historical nature of social norms (especially those that define what a 'dispute' is and the rules by which it is handled); as well as the resources that actors who are seeking to influence the trajectory of a dispute have at their disposal (Barley 1991). Overarching these dimensions is an appreciation of the semantic and material implications of the language of conflict; in this way, the anthropological approach shares a number of methodological similarities with Actor Network Theory (Law 1994) and Domain Theory (Gray 1989).

The anthropological approach, as depicted by Barley, serves the important purpose of showing us that we do not have to think of conflict exclusively in psychological terms. Breaking the tyranny of the anthropomorphized dyad allows us to think about conflict as a complex nexus of social relations and avoids the functionalist view that it must be doing some 'good' somewhere. Nevertheless, the anthropological study of conflict still remains under-theorized. Indeed, it remains largely an empirical exercise and little theory is at work, an observation that is particularly striking when considering the dimension of power in shaping the antagonistic relations of everyday social settings. For example, Barley's (1991) conception of power extends little beyond the distribution of 'resources' available to parties to influence outcomes in competitive relations, a view that is also evident in some explicitly 'sociological' treatments of conflict (e.g., Tolbert 1991). Thus, it is little different from social psychological approaches in dealing with the complexities of the power relations mediating organizational conflict. Recently, several authors have addressed this shortcoming in anthropological studies of conflict: for example, Sarat and Felstiner's (1995) application of Michel Foucault to their studies of divorce negotiations. Nevertheless, the general lack of nuance in the treatment of power by theorists concerned with organizational conflict points to an area where studies can be greatly enriched.

The conceptual arena of power underwent profound changes in the late twentieth century, shifting from arid behaviorism to arcane abstractedness, before finally settling into fine-grained analysis (see Flyvbjerg 1998; also see *Organizations: Authority and Power*; *Organizational Control*). In part this shift was symptomatic of a return to a less moralistic and more Machiavellian view of power. We shall return to these themes in the last section of this article.

3. Organizational Conflict and Neo-Darwinism

The basic premise of neo-Darwinian approaches to human behavior, such as the Evolutionary Psychology movement (e.g. Buss 1995), is that the human mind is stuck in the prehistoric African Savannah of our hominid ancestors. Behavioral traits—for example, human aggression and a predisposition toward intersubjective conflict—survived because they conferred adaptive advantage on us as a species. Thus, we can established *ex post facto* explanations for almost any empirically observable form of behavior—on the basis of natural selection then, if it was not adaptive, it would have been bred out. Nicholson (1997) suggests that common characteristics of organizations, such as a predilection for gossip, the love of ritual, and even the existence of hierarchy, can be traced to biologically determined human nature.

The increasing influence of the neo-Darwinian approach (incidentally, a viewpoint that is opposed by many evolutionary biologists who see it as a misappropriation of Darwin's ideas) is already being felt in studies of organizational conflict. For example, Ehin (1995) considers the tension between conflict and cooperation to be part of our biological destiny. Moreover, he contends that, by studying the mutual dependence systems of our hunter-gatherer ancestors, we are able to define a set of desirable social principles for current organizations that are '…more congruent with human nature' (Ehin 1995, p. 667). Adopting a dualistic model of innate characteristics, Ehin argues that humans display either *agonic* (aggressive, individually territorial, and hierarchical) or *hedonic* (mutually supportive, group-centered, and egalitarian) tendencies. The extent to which we display one or the other is determined by context and, citing studies of chimpanzee behavior as evidence, Ehin contends that, if we can return to our 'natural' settings, the hedonic predisposition will prevail. It is only when we are removed from this setting—in this case, placed in a hierarchy of work organization—that we tend to display agonic characteristics. This work stands as the intellectual equivalent of Hobbes' primitive ruminations and the organizational implications are clear. First, conflict is an evolutionarily advantageous response to many 'unnatural' forms of organization and, second, if only we can return to a more natural form of organization, then conflict will be, to all intents and purposes, banished as we revert to our hedonic state. This reflects an increasingly popular view that presence of cooperation and the absence of conflict are not just virtuous aspirations within the modern society, they are biological necessities that ensure the continuation of our species (Fukuyama 1999). Not only is this view naive with respect to issues of power, it is also a contestable interpretation of Darwin's idea of natural selection. In particular, to two specific objections can be laid against it.

The first is fundamental: according to Darwinian adaptation, evolution occurs through random mutation rather than strategic intent and anything that involves intent cannot be explained by Darwinian ideas. Of course, axiomatically this rules out all social situations, including those where parties are in conflict over the relative moral status of competing norms or the distribution of resources (see *Industrial Relations and Collective Bargaining*). But, as if this misinterpretation of Darwin was not sufficient, such studies can never satisfy the requirements of evolutionary science. This point has been particularly well made by de Jong and van der Steen (1998) who persuasively demonstrate the inability of any psychological study to establish that anything so abstract and socially constructed as a behavioral trait (for example, a predisposition toward conflict) could be advantageous from an evolutionary perspective. Few would deny that heredity plays *no* part in behavior but it is stretching it beyond the conceptual and methodological bounds of Darwinism to suggest that the dynamics of organizational conflict will be *determined* by human genetics.

4. A Neo-Machiavellian View of Organizational Conflict

By control to the previous approaches to organizational conflict, we wish to offer a fourth framework for thinking about the problem. Our starting point is that we must acknowledge that conflicts matter in any organization and we are all players—even in the face of denial or ignorance (being devious or stupid does not disengage one from conflictual processes). There are four basic ways in which the political maneuvering of organizational conflict proceeds. The first three of these invariably lead to protagonists being ousted in conflict. The organizational agent who is not politically skilled will fail. Returning to the Machiavellian theme, the key implication is that skill in the arts of political engagement, strategy, and discourse should be central to organizational competencies.

The first typical mode is to be *in denial* of conflict. We can characterize this as unrealistic, for few rational actors could be insensitive enough to their organizational surroundings to survive in this state for long. Second, there is *passive acceptance without resistance* of what one is in conflict with. This is naive: it accepts what it does not seek to understand or counter. Third, there is *active recognition without action*, where one realizes the nature of the conflict but proposes no strategic counterpoint for dealing with it. We may characterize this as incompetent. Finally, there is *strategic cognition and treatment*, where one recognizes the phenomenon as conflict and treats it appropriately through devising strategies to manage it, with and through the use of power. We may characterize this,

after Buchanan and Badham (1999), as developing 'power-assisted steering.'

Essential to these competencies is the ability to be able to creatively improvise around organizational conflicts, through developing a diverse repertoire. This includes: image building; selective information release and dissemination; scapegoating of conflictual others; building formal alliances with potential allies; networking with friends and enemies of enemies; striking up appropriate compromises where necessary; being able to manipulate rules, and being ruthlessly pragmatic. Such tactics require deployment in specific organizational contexts, characterized by diverse stakeholders and variable political entrepreneurship. The exercise of power, in ongoing action, will be accomplished through the production of plausible accounts, the creation of viable reputations for political operators, and the achievement of outcomes regarded as 'wins' (Buchanan and Badham 1999). Essential to these capabilities is a capacity to analyze strategies of power discursively. Those who would be players need to be able to ask questions about the extent to which their power games—actual or projected—will be characterized as ethically acceptable, contextually reasonable, and plausibly accountable? What will the consequences be for those whose reputations are involved in the stakes: who will be weakened, strengthened, or unchanged?

Organizational life involves developing and deploying *political competence*. (The theoretical derivation of this view of organizations are elaborated in Clegg [1989] and Buchanan and Badham [1999].) Sometimes the skill will be in containing conflict or nipping it in the bud, other times ensuring where and when it flares, in what arenas, and with what participants. Conflicts represent opportunities for honing the skills of political competence in whatever normative order the organization constitutes. Clearly, these are highly variable—but even the most sacred may be as readily characterized in these terms as may be the most profane—as any reader of Trollope's 'Barchester' novels could attest. Political strategy does not depend on the normative environment in which it is lodged so much as the strategic capabilities deployed. On this reading, Machiavellian skills are appropriate in any organizational context.

See also: Administration in Organizations; Conflict and Conflict Resolution, Social Psychology of; Conflict: Anthropological Aspects; Conflict Sociology; Darwinism: Social; Organization: Overview; Organizational Behavior, Psychology of; Organizations: Authority and Power

Bibliography

Barley S R 1991 Contextualizing conflict: Notes on the anthropology of disputes and negotiations. *Research on Negotiation in Organizations* **3**: 165–99

Buchanan D A, Badham R 1999 *Power, Politics and Organizational Change: Winning the Turf Game*. Sage, London

Buss D M 1995 Evolutionary psychology: A new paradigm for social science. *Psychological Inquiry* **6**: 1–30

Clegg S R 1989 *Frameworks of Power*. Sage, London

de Jong H L, van der Steen W J 1998 Biological thinking in evolutionary psychology: Rockbottom or quicksand. *Philosophical Psychology* **11**: 183–205

Ehin C 1995 The quest for empowering organizations: Some lessons from our foraging past. *Organization Science* **6**: 666–71

Flyvbjerg B 1998 *Rationality & Power: Democracy in Practice*. University of Chicago Press, Chicago

Fukuyama F 1999 *The Great Disruption: Human Nature and the Reconstitution of the Social Order*. Free Press, New York

Gray B 1989 *Collaborating: Finding Common Ground for Multiparty Problems*. Jossey Bass, San Francisco

Law J 1994 *Organizing Modernity*. Blackwell, Oxford, UK

Nicholson N 1997 Evolutionary psychology: Toward a new view of human nature and organizational society. *Human Relations* **50**: 1053–78

Sarat A, Felstiner W L F 1995 *Divorce Lawyers and their Clients: Power and Meaning in the Legal Process*. Oxford University Press, Oxford, UK

Thomas K W 1976 Conflict and conflict management. In: Dunnette M (ed.) *Handbook of Industrial and Organizational Psychology*. Consulting Psychologists Press, Palo Alto, CA pp. 889–936

Thomas K W 1994 Conflict and negotiation processes in organizations. In: Dunnette M, Hough L M (eds.) *Handbook of Industrial and Organizational Psychology*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA, Vol. 3, pp. 651–717

Tolbert P S 1991 Negotiations in organizations: A sociological perspective. *Research on Negotiation in Organizations* **3**: 99–117

S. Clegg and G. Sewell

Conflict Sociology

When two or more social actors pursue incompatible interests they may be said to be in a relationship of conflict; such conflicts may remain potential, or they may result in various kinds of overt behavior. The sociological study of conflict is concerned with all these possibilities, though the study of the most extreme form of conflict—actual war—is usually the province of political science and international relations. The subject has been treated in diverse ways in sociological theory. These will be analyzed in terms of two sets of variables: conflict as exceptional or endemic; and as momentous or mundane.

1. Major Axes in the Analysis of Conflict

There are two central choices for sociological theory in the treatment of conflict. First, should it be treated as exceptional or as endemic? In the former case, the

theory assumes that normally social life will proceed without conflict, or at least without its overt expression in hostile actions. If conflict becomes evident, this is a sign that some or other institution is not functioning properly. Conflict is therefore seen as pathological. Alternatively, a theory may assume that social relations are likely to exhibit conflict at many points, and its expression will be expected to be frequent. In some cases the absence of conflict might be seen as pathological.

A second choice is whether, when conflict does occur, it should be regarded as of momentous importance, likely to result in major upheaval and possibly radical change; or whether it should be treated as mundane, that is, merely part of the events of everyday life, without particular moment.

These are by no means the only issues which have to be addressed by theories of conflict, but the two-by-two matrix which they form provides a useful heuristic for studying the position of different schools of thought (see Fig. 1).

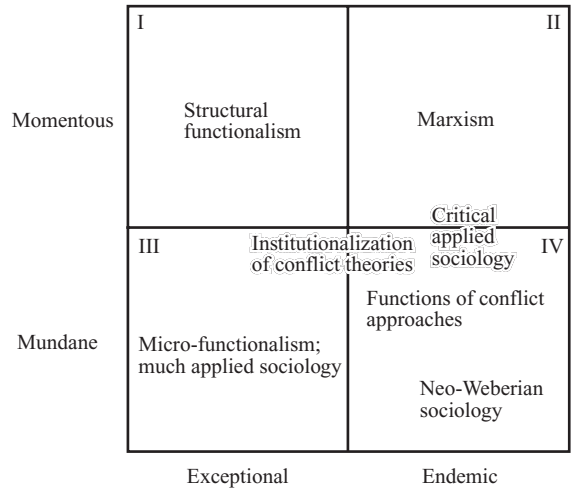

Figure 1
The main axes of sociological theories of conflict

2. *Conflict as Momentous—Whether Exceptional or Endemic*

The following discussion is organized in terms of the momentous versus mundane dimension, the second dimension being examined within this one. This is purely for convenience of presentation and does not imply any priority of importance.

2.1 *Conflict as Momentous and Exceptional: Functionalist Approaches*

Functionalist theories, particularly in the structural functionalist form which dominated US and much other sociology in the first two postwar decades, are clear and strong examples of the treatment of conflict as momentous and exceptional (see *Functionalism in Sociology*). These theories, particularly those of Talcott Parsons (in particular, Parsons 1964), have often been accused of an incapacity to analyze conflict. This is not true. As Parsons once remarked, he was aware of the enormous capacity for conflict in social relations and was surprised that so much cooperation in fact took place. The purpose of his theories was therefore to demonstrate how it was that, in practice, conflict did not overwhelm social relations. As a result, the emphasis of the theoretical tradition he launched was on how conflict was rendered exceptional.

Functionalism views social institutions as overwhelmingly interconnected and equipped with integration mechanisms. While for most of the time conflict does not occur, social relations are highly vulnerable should it do so. It is a sign that something is malfunctioning, that something pathological is occurring, and major disorder, possibly leading to change, can be expected (Cohen 1966). Structural functionalism fits box I of Fig. 1.

2.2 *Conflict as Momentous and Endemic: Marxist Approaches*

During the 1960s and 1970s there was a major polemical confrontation between structural functionalist and Marxist theorists, in both western Europe and the USA (see *Marxism in Contemporary Sociology*). For Marxists, conflict is endemic to all significant social relations, because these embody in various ways the class relations that are fundamental to social life and that are based on opposed interests. Marxist writers have therefore been unhappy with behaviorist approaches to social conflict, which regard a conflict as existing only when it is openly manifest. According to Marxists and other critical sociologists (see *Critical Theory: Contemporary*), this conceals from view conflicts of interest which the weaker parties to a relationship are powerless or fearful to express. Convinced that the conflict will in principle be there, such sociologists search for subtle and implicit expressions of it (Lukes 1974).

To take a simple example, a functionalist sociologist carrying out research on a factory or office would regard evidence of conflict between management and workers as evidence that something was wrong, that institutions governing social relations in the workplace were not operating correctly. A Marxist sociologist, on the other hand, would regard such instances of conflict as evidence that the reality of the relationship between management and workers was revealing itself; it would be the absence of conflict in such a

relationship that would require explanation as something denying that fundamental reality.

However, opposed as functionalism and Marxism might be, they are in fact agreed on the momentous–mundane dimension. For Marxist sociology as much as for functionalist, conflict is likely to be momentous in its consequences, bringing widespread and uncontainable disorder before it ushers in social change. This is because for Marxists conflict is the way in which the underlying contradictions of the class relations on which the social order rests are finally and catastrophically resolved. Marxism occupies box II in the matrix.

Therefore, contrary to initial appearances, structural functionalism and Marxism are not diametrically opposed in their treatment of conflict. One ironic consequence of this has been the way in which Marxist accounts of an ongoing social order—one in which the fundamental flaws have not yet been revealed—sometimes resemble functionalist ones. For both schools all social institutions are interconnected and, to the extent that (for Marxists) class domination is operating effectively, all institutions tend towards maintenance of the stability of the existing order. This inherent functionalism of much Marxist theory became particularly evident in the structural Marxism which developed in France during the 1970s, exemplified by such writers as Althusser (1968) and Poulantzas (1968). These depicted the totality of capitalist domination in such a way that social conflict seemed almost impossible.

3. *Conflict as Mundane—Whether Exceptional or Endemic*

An important development in the study of social conflict emerged during the 1950s with the idea of the institutionalization of conflict (e.g., Harbison 1954). This involved the hypothesis that, if mechanisms existed which separated different institutional areas from each other, conflicts appearing in one would be prevented from spreading into others, and conflict would therefore be limited. The idea can best be understood through an analogy with insulation or isolation in electricity. The cable wrapped around an electrified wire prevents the current from jumping from the wire to objects alongside it, which it might damage. Interest then focuses on what kinds of social mechanisms might serve as isolators separating one social area from another. If such isolators exist, then individual outbreaks of conflict are unlikely to have momentous implications for the wider society.

Interest in institutionalization developed from the turmoil of the first half of the twentieth century, which had seen considerable disruption caused by two political forces—communism and nazism–fascism—which owed their success, in part, to a capacity to amalgamate disparate issues into totalities that pitched whole groups of populations against each other. There was particular interest in mechanisms which might separate political from industrial conflict, this combination being seen as crucial to the success of communism.

In a closely related development, political sociologists were interested in ways in which different conflictual identities might either run together or cross-cut one another (Lipset 1964). In the former case conflicts might be expected to spread across whole institutional areas and to be intense. In the latter, conflicts between groups defined according to a particular characteristic would be offset and thereby contained and diminished by cross-cutting identities. For example, a society in which most blacks were Roman Catholics and also manual workers, while most whites were Protestants and bourgeois, would be expected to generate deeper and more intransigent conflicts than one in which ethnicity, religion and class divide the population in different ways.

These approaches made it possible to conceive of conflict of no major macro-social importance as mundane.

3.1 *Conflict as Mundane and Exceptional: Micro-functionalism and Applied Sociology*

Theories of conflict institutionalization often overlapped with the functionalist approach to conflict as pathological, and many writers in that school were concerned, both intellectually and in practice, to combat Marxism. The idea of institutional isolators was also quite compatible with the functionalist approach, such mechanisms being among the devices by which social integration was ensured. On the other hand, a stress on institutional separateness could lead theorists away from the emphasis on the overall interconnectedness of social institutions which was distinctive of structural functionalism.

As sociology expanded from the 1970s onwards, and spread into a large number of specialized subsectors, many practitioners of the discipline lost a sense of connecting their work to an overall perspective on society, not necessarily because they did not believe in such a perspective, but because the need to develop knowledge within specialisms required its neglect. Institutional separateness was pragmatically reinforced, even among those who maintained an essentially functionalist standpoint. Their theoretical stance becomes one of micro-functionalism. Functionalism's sense of conflict as exceptional and pathological is therefore not necessarily lost. These approaches therefore occupy box III in Fig. 1. The occurrence of strikes, divorces, crime, outbursts of violence and terrorism, can still be seen as empirical indicators of malfunctioning; but it is malfunctioning of a mundane kind. This is particularly likely to be the case with applied sociology, designed to be used for

the study of perceived social problems, in such areas as the sociologies of working life, ethnic relations, the family and marriage, poverty, social movements, crime and deviance.

3.2 Conflict as Mundane and Endemic: From Critical Sociology to the Normalization of Conflict

By no means all students of specialized fields in which conflicts may occur necessarily share a view of it as pathological. Some may be close to a Marxist or other radical position, likely to see expressions of conflict as demonstrating the realization of certain fundamental contradictions within institutions. This is, for example, likely to be the perspective of feminist sociologists on family conflict, or of observers from a number of perspectives on ethnic conflict. Work of this kind therefore marks a movement towards box IV (Fig. 1): conflict perceived as normal and mundane. However, to the extent that these approaches retain something of a Marxist or more generally 'critical' idea of conflict revealing inherent contradictions, they are unlikely to remain fully at the level of the mundane; conflicts will be expected to involve major disruption and possibly change. For example, if growing family conflict (both divorce and difficult intergenerational relationships) are seen as evidence of challenges to patriarchy, they are unlikely to be mundane conflicts that just rumble along without coming to some kind of crisis.

These differences in the extent to which conflict could be regarded as fully mundane and containable can be seen from contributions made to the study of industrial conflict by British sociologists. These have contrasted a unitary approach to industrial relations, characterized by implicit functionalist assumptions, with a pluralist one (Clegg 1975, Fox 1973). Under the former it was assumed that management and employees shared interests; as a result conflict was treated as pathological, something that should always be prevented. Under the latter, it was taken for granted that employers and employees had partially opposed interests; as a result conflict between them was endemic, and its occasional manifestation should be regarded as normal. There is then however disagreement among those who regarded such conflict as thoroughly containable and who belong fully to box IV in Fig. 1 (Clegg 1975), and those who took a more Marxist position and saw potentially momentous possibilities (Fox 1973, Hyman 1975).

More abstract theories and approaches to conflict belong fully to box IV (Fig. 1). As the Conclusion below will suggest, pursuing this path leads eventually to the disappearance of conflict sociology as such. However, the journey towards that position has involved some important contributions to the study of conflict.

Already in the late nineteenth century, Georg Simmel had pointed to a paradoxical way in which conflict might actually strengthen social ties: parties to a conflict had a stronger relationship than total strangers (see *Simmel, Georg (1858–1918)*). Many conflicts, even full-scale wars, were governed by minimal sets of rules and mutual understandings. From these insights developed the idea that conflict might be functional to social relationships. In the 1950s, Lewis Coser (1956, 1965) developed from Simmel a general theory of the functions of social conflict. At the macro-theoretical level, this was a direct challenge to the dominant view of conflict as pathological, while remaining within a broadly functionalist approach. More generally, it contributed to the perception of conflict as both endemic and mundane.

Also starting in the 1950s was a series of major contributions by Ralf Dahrendorf (1957, 1972), which took theory a significant step further. Not only could conflict be functional, but perhaps social order was actually sustained by conflict. Such an approach first required the concept discussed above of the institutionalization of conflict. Provided individual conflict zones could be mutually segregated in the way this concept anticipated, the open expression and working out of differences, difficulties and contradictions was the way in which people innovated, developed institutions and actually sustained social relations. If conflict was bottled up, there were no possibilities for doing this, relations would fail to be open and honest, and innovation would not take place. In radical contradiction with structural functionalism, this approach contended that it was the endemic and mundane character of conflict which ensured social order.

Dahrendorf was writing with the tragic history of his native Germany in the first half of the twentieth century very much in mind (Dahrendorf 1966). To that extent, his work can be seen as part of a more general tendency in postwar German sociology to stress open dialogue and communication, which necessarily implies the working out of conflicts (Habermas 1981). Conflict became normalized further by developments in the 1960s and 1970s which drew on the legacy of Max Weber (see *Weber, Max (1864–1920)*). There was for a long time a struggle between functionalists and others over who were the true heirs of Weber's approach, particularly in the English-speaking world where Parsons had been influential in the translation and interpretation of the German scholar's work Parsons 1937, Weber 1947).

As more thinkers came to read the original, as German sociology re-emerged during the post-war era, and as more neutral translations appeared, so Weber's legacy was prised away from functionalism. Weber had neither ignored conflict nor had a special theory of it. Conflict was there as an integral part of the collection of social phenomena that resulted from different actors pursuing their own interests as they defined them. Weber had not taken the functionalist step of trying to find overall integration processes that

would somehow bind these conflicting interests into a wider whole. Conflicts could rumble on indefinitely. If order was achieved, it was most likely to result from one group being able to impose its interests on others through the exercise of *Herrschaft*. This word translates easily as 'domination,' though Parsons had preferred the circumlocutory 'imperative co-ordination.'

Weber had also differed from Marxists in that he did not reduce social relations—including their conflictual aspects—to those of material class interests, nor indeed of any other essentialist forms. Conflict might be about class interest, but might also be about idealistic beliefs and symbolic orders. None were treated as ontologically privileged.

This interpretation of Weber became increasingly useful to those trying to avoid the hegemonic tendencies of grand theory. A good example is the work of John Rex (1961, 1981). Of political leanings which led him towards Marxism, Rex had been born in Zimbabwe (then Southern Rhodesia), and was preoccupied with racial and ethnic conflict. He spent most of his career in the British midlands, also a region of considerable ethnic complexity and, sometimes, tension. Marxist insistence on reducing all conflicts to class did not appear to him to cope with the independent reality of these issues. He found Weber's ideas of variously rooted status groups, and their use of a diversity of means to maximize their interests, far more useful. Ethnic relations, and by derivation many others, could therefore be seen as incorporating conflict as an intrinsic and normal aspect of attempts at interest definition and maximization.

4. *Conclusion: The Disappearance of a Distinctive Conflict Sociology*

Now that the great edifices of structural functionalism and Marxism seem finally to have collapsed, the agnostic, non-essentialist approach implied by the idea of conflict as endemic but mundane has probably become the dominant one in Western sociology. Yes, conflict is always likely to be a part of human relations; it can be 'about' almost anything; it does not necessarily have to be absorbed and tamed for some kind of social order to survive; it is not a special aspect of social life, but just a part of it. Research and theory based on assumptions of this kind cluster around the bottom right-hand corner of box IV (Fig. 1). Weber, perhaps along with Simmel, is probably the founding father whose approach to conflict appears most comfortable to contemporary sociologists. As heirs to the Marxist tradition seek to salvage something while dropping the dogmatism and historical certainty, they reach similar conclusions.

The more normal that conflict is seen, the more a specific conflict sociology will cease to exist. This becomes particularly true as sociological theory turns increasingly to theories based on models of rational action by calculating actors—approaches which are also found in Weber and Simmel (see *Rational Choice Theory in Sociology*). In these models actors are always depicted as having interests; and where there are interests, conflicts among them are normal and taken for granted (for example, Knight 1992).

For conflict sociology to continue to exist, the various specialisms of the discipline would need to draw on similar ideas and hypotheses when they come to instances of conflict within their particular area. This does happen to a limited extent. For example, studies of how protest groups engaged in difficult conflicts sustain their morale and endurance can be extended across from ethnic to industrial or environmental movements. Studies of the role of violence and the effect of violence on perpetrators, victims and outsiders may also be able to share knowledge across conflicts concerning very different issues. Beyond this however not much survives.

It has been noted at several points above how sociological approaches to conflict have often reflected developments in the world at large: obsession with society's capacity for upheaval after the first half of the twentieth century; Dahrendorf's concern to treat conflict as normal in the light of the German past; Rex's African concern to study racial conflict. Today's postmodern, post-Cold War world presents a scene in which conflict seems at once endemic but directionless, and sociological theory is reflecting that. Were the heartland of theoretical development to lie, not in the so-called 'advanced' societies, but in parts of the world where religious and ethnic struggles are fundamental to social life, the situation would probably be rather different (Ratcliffe 1994, Venkateswarlu 1992). And since it must not be expected that social change in the advanced societies has come to an end, it may well be that after a number of years the emphasis of the above account will appear very dated.

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict and War, Archaeology of; Conflict: Anthropological Aspects; Conflict/Consensus; Conflict: Organizational; Cooperation and Competition, Psychology of; Ethnic Conflicts; Geopolitics; Race and Gender Intersections; Violence, History of; War: Causes and Patterns; War, Sociology of

Bibliography

Althusser L 1968 *Lire le Capital*. Maspero, Paris

Clegg H A 1975 Pluralism in industrial relations. *British Journal of Industrial Relations* **XIII**: 3

Cohen P S 1966 *Modern Social Theory*. Heinemann, London

Coser L 1956 *The Functions of Social Conflict*. Free Press, Glencoe, IL

Coser L 1965 *Georg Simmel*. Prentice Hall, Englewood Cliffs, NJ
Dahrendorf R 1957 *Soziale Klassen and Klassenkonflikte in der industriellen Gesellschaft*. Enke, Stuttgart, Germany
Dahrendorf R 1966 *Gesellschaft und Demokratie in Deutschland*. Piper, Munich, Germany
Dahrendorf R 1972 *Konflikt und Freiheit auf dem Weg zur Dienstklassengesellschaft*. Piper, Munich, Germany
Fox A 1973 Industrial relations: a critique of pluralist ideology. In: Child J (ed.) *Man and Organization*. Allen and Unwin, London
Habermas J 1981 *Theorie des kommunikativen Handels*. Suhrkamp, Frankfurt am Main, Germany
Harbison F H 1954 Collective bargaining and American capitalism. In: Kornhauser A, Dubin R, Ross A M (eds.) *Industrial Conflict*. McGraw Hill, New York
Hyman R 1975 *Industrial Relations: A Marxist Introduction*. Macmillan, London
Knight J 1992 *Institutions and Social Conflict*. Cambridge University Press, Cambridge, UK and New York
Lipset S M 1964 *Political Man*
Lukes S 1974 *Power: A Radical View*. Macmillan, London
Parsons T 1937 *The Structure of Social Action*. McGraw-Hill, New York, Vol. 2
Parsons T 1964 *The Social System*. Routledge and Kegan Paul, London
Poulantzas N 1968 *Pouvoir Politique et Classes Sociales*. Maspero, Paris
Ratcliffe P (ed.) 1994 *'Race,' Ethnicity and Nation: International Perspectives on Social Conflict*. University College London Press, London
Rex J 1961 *Key Problems in Sociological Theory*. Routledge and Kegan Paul, London
Rex J 1981 *Social Conflict: a Conceptual and Theoretical Analysis*. Longman, London
Venkateswarlu D 1992 Conflict sociology in India—present state and future potentialities. *Indian Journal of Social Sciences* **2**: 241–248
Weber M 1947 *The Theory of Social and Economic Organization* [trans.] Henderson A M, Parsons T. Oxford University Press, New York

C. J. Crouch

Conformity: Sociological Aspects

Conformity is a central pattern of social life since it appears that individuals often behave in a similar way in similar circumstances. Why do they do so? Is it because there is some social force acting behind them, or is it because they can be seen as rational, this rationality being linked to some common tendency to act in a definite way in given situations?

1. *Conformity, Similarity, and Dissimilarity*

Conformity depends directly on the degree of similarity between an individual action and another, those actions being more or less frequent in a group. Tarde (1962) tries to stress the importance of imitation to explain such similar behaviors in social life. Imitation appears to him as a constant attitude which explains all social phenomena: all fields of action are involved, and innovation, or dissent, depends always on previous imitations. But imitation has such general contents that it is not very useful to explain the peculiarities of different situations. Durkheim (1982) does not accept this rather general idea of imitation. But for him, there is a similar strong social tendency toward the building of common patterns of behavior: moral, religious, linguistic, or esthetic rules express in a common way the fact that society is built up on the necessity of integration of individuals throughout shared values. Despite any subsequent difference of attitudes, conformity is an essential feature of society.

More or less supported by explicit values, or norms, conformity is strengthened by unequally strong rewards and sanctions. From death penalty to intellectual dissatisfaction, from imprisonment to snobbery, conformity is most often reinforced by social actions in different ways. But it is also observable that conformity can occur when no sanction at all is observable, for instance, people's breakfast habits, which differ from one society to another, this difference being related to a group attitude.

Moreover, when there is a claim to conformity, and sanctions (or rewards) towards potential conformists or deviants, the size of the reference group can have a variable extension (Mead 1962): for a Catholic, or a believer in human rights, humanity as such should comply with some specific norms, whereas the follower of an elitist fashion will stop liking it when the number of followers has reached a certain size (Coleman 1990).

Conformity is related to beliefs and to justifications: people are conformist or deviant because they believe they should be so, and others are asked to behave in the same manner when they are thought to belong to the relevant group to which the norm that introduces conformity should apply. Therefore, a study of the conformity problem must include the motives for which a justification is accepted or rejected. People can obey a norm because they believe the norm is valid, or because they are forced to comply with it. But they can refuse to obey a norm not because they do not acknowledge the validity of the norm, but because they have a stronger interest against conformity. On the other hand, people can refuse to obey a norm because they believe it is unfair, or useless, or unnecessary, or inadequate.

Compliance with similar patterns of behavior can be related to personal or group circumstances: people may think that a norm is illegitimate from the point of view of their position in society. For instance, those who cannot through their personal effort achieve a social promotion, will tend to believe that, since they do not have such opportunity, they cannot have any faith in a social order that builds its legitimacy on this

very self-promotion through effort (Merton 1957). Those conditions can be related to the position of an individual within a group, or to the members of a group within a larger group. Finally, there is a problem of comparison among completely different groups when it appears that they have completely different systems of values.

Because of this wide variety of situations, it seems difficult to accept only one general trend towards conformity or deviance, whether it should be located at an individual level (imitation) or at a social level (social integration). It is more promising to distinguish several fields where the conformity issue can be described in significantly different ways.

2. *Positive Beliefs*

A first domain for conformist attitudes is that of positive beliefs. In this field there is a strong claim to conformity when there are epistemic norms that are recognized from a universal point of view. If there is a norm of validity for the description of a fact, or a set of facts, it is not surprising that rational people will acknowledge the evidence regarding this norm. Conformity has no negative connotation in this case. For sociologists, who believe in science and the validity of their academic enterprise, it would be difficult, and arguably inconsistent, not to believe in such a norm. Subsequently, they will be interested, regarding positive descriptions and beliefs commonly shared by people, in the frequency of typical errors. The sociologist has many devices to explain such widely spread errors.

Tocqueville (1966), for instance, explains that people need general ideas to achieve their ordinary actions, but they do not have the means to verify the validity of all the beliefs they will refer to. In a society based on the dismissal of all authority, the majority of people will think that the most acceptable belief is the one that is commonly shared by others. There will be a tendency towards conformity since the adoption of majority's voice is less costly, both from an intellectual point of view and from a social point of view. Boudon (1990) has studied various ways social errors are built up on a rational basis.

3. *Social Rules in Moral, Political, and Religious Fields*

As previously mentioned, conformity implies a problem of legitimacy. There are strong social norms regarding social practices in moral, political, and religious fields. These norms tend to vary in a significant way from one society to another, and there is a historical evolution that must be taken into account. Sociologists can consider that this variety must lead to a complete relativity of values. But they can try also to pay attention to the reasons that lead to the acceptance or to the refusal of norms.

First of all, it seems that there are four relatively constant norms throughout all cultures (Cusson 1983): incest between mother and son, father and daughter, or brother and sister; rape of married women; murder, meaning the voluntary act of killing someone who belongs to one's own group; theft. Of course, the last case is the less accurate since the definition of legitimate belongings can vary in such a way that property itself can be seen as a theft.

If we depart from those universal features, labeling theorists (Becker 1963) interpret conformity as the result of a power relationship: the powerful impose on others conventions that are arbitrary. Arbitrary conventions have two effects. First, some people are labeled as deviants whereas others are praised as conformists. This expresses the structure of power in society, certainly not an intrinsic validity. Second, once people have been labeled as deviant, this label reinforces by stigmatization their deviant situation, since others can no more accept them as nondeviants. This narrows the range of actions that are available to them; eventually the stigmatized can themselves believe that they are deviants.

It is obvious that two situations can frequently appear. Some people tend to impose on others a norm that cannot be valid to them, and depend only on a power relationship. For instance, it is clear that a racist norm can hardly be seen as legitimate by its victims, since it has no other ground than the relation of power. But on the other hand, people tend to impose norms that they try to legitimize. The claim of the stigmatized must necessarily be a change of norm, that is, the replacement of a nonvalid norm by a valid one. For instance, critics of racism will support the idea that there is no legitimate ground for racism, and they will try to set up a norm against racism: if they succeed, racists will become deviants. If labeling theorists maintain the idea that there is never any intrinsic value to certain rules, they will have no basis for the replacement of a racist rule by a nonracist rule, and they will treat in the same manner the deviant who does not accept a prevailing racist norm and the deviant who does not accept a prevailing nonracist norm.

Therefore, historical evolution must be taken into account as it can be related to the relative strength of different arguments regarding norms. If we consider, for instance, norms regulating sexual behavior, Christian religious belief tends to condemn, on a metaphysical basis, premarital sex, divorce, homosexuality, and pedophilia altogether. In a society where people are submitted to such norms, if they do not share the religious beliefs, they might consider themselves as victims of arbitrary norms. In such an instance they might believe that they have a right to do whatever they wish to do, in spite of those forbidding norms. It can be noticed, therefore, that when there is

no consensus on religious norms for sexual behavior, the only path to maintain a legitimate norm that will be valid for everyone will be to accept the principle of personal choice. Religious people will act according to their choice, and the nonreligious to their other choice. Modern societies increase the principle of personal choice, and diminish the authority of traditional or religious norms. The evolution is significant: divorce tends to be widely practiced, and homosexuality is increasingly legalized. But one significant aspect of this evolution is that norms against pedophilia tend on the opposite to be strengthened in a liberal society. This is easy to explain, on one side, divorce or homosexuality are related to a freedom of choice among consenting partners. If a marriage is a contract between two adult persons, the notion of agreement can imply that the contract should be dismissed if the two partners do not agree any more, even if this is considered to be a failure from another point of view. On the contrary, pedophilia is definitely rejected because it involves children, legitimately seen as victims because of their lack of free choice.

Therefore it cannot be said that a labeling theory is the last word on the problem of conformity or deviance. Both depend on beliefs and possibilities of legitimacy. In other words, norms depend on reasons (Boudon 1990). In this case, if strong norms depend on religious beliefs, when they are imposed on others who are nonbelievers, there is a problem of legitimacy if those beliefs cannot be proved to be right or wrong. Therefore, rationalization will tend to lead to a situation where such norms will apply only to believers. In this case, the principle of choice tends to legalize some practices, and to reinforce the fact that others are forbidden.

According to Boudon's latest works, we can try to find the reasons for the acceptance or the rejection of a norm. Dissatisfaction can occur when there is a rejection of the elements on which the norm is based, or when the norm can be seen as valid in general, but not in the situation where the actors find themselves. Once again, according to Merton's description, someone will tend to reject a social order based on the idea of self-promotion through effort when it cannot be really achieved. Here a problem of consistency is faced: the norm is valid under certain conditions, but ceases to be so when those conditions are not met.

Hence, conformity should not be seen as static. Many norms depend on beliefs that are submitted to rationalization. One of the aspects of this rationalization is that some norms cannot be valid to certain people: for instance, if there is a norm supporting slavery, it will tend to be withdrawn in the long run since it is inconsistent with the source of any legitimacy for individuals, their ability to accept a norm, which implies a certain amount of freedom of choice. Some norms are seen as oppressive which were not before, because the belief on which they were based is no longer shared.

From a rationalistic point of view, there can be a discussion that shows that conformity or deviance depend on the perception of the validity of norms. It can be said that norms tend to be rejected in four major cases: when they just express a power relationship that is disguised in some other value; when they are grounded on false beliefs; when they are grounded on beliefs that cannot be proved and which have no self-evidence—in this case, they cannot be imposed on someone who does not share such beliefs; finally, a norm is not valid when it cannot be applied to someone's situation.

4. *Norms of Culture*

In the two previous sections, I have tried to isolate norms that can be discussed among people from the standpoint of their validity. But there are other norms that are connected to strong conformity that do not seem to be negotiated in terms of validity. For instance, Japanese women tend to walk in a way significantly different from European women. On one hand, there is no obvious reason to consider one way as better than the other; on the other hand, such behaviors express conformity, since they are group behaviors. How should we give an account of such attitudes applied to food, language, dress, architectural habits, etc.?

Three major elements can be stressed regarding this type of conformity. First, it can be related to practical constraints. A language must be shared among people to be efficient for communication. Here, conformity is just the result of the nature of the practice. There cannot be a private language (Wittgenstein 1967).

Second, similarity can express integration to a group. For instance, it is obvious that religious dress is linked to the sense of belonging to a community which can be turned into strategic devices. People tend to adopt some features to show that they belong to certain groups (Simmel 1923). This can be done because belonging to a group is seen as important, or because belonging to a group is a consequence of a common faith or a common goal, or because there is some sort of distinction that is looked for (Veblen 1953).

But finally, it must be said that in these fields conformity tends to be linked to innovation. It is obvious in the esthetic field: on the one hand, there is a huge amount of conformity, that is observable in the constituency of local schools, such as the general use of Romanesque architecture during certain periods in Europe. This conformity, however, does not exclude innovation; Gothic style sooner or later replaces the Romanesque. It would be too difficult in this article to assess all the elements that play a role in such an evolution. It is clear at least that the previous ones are important: interactive constraints, and the sense of

belonging to a community. But these allow the evolution of behaviors and the place for innovation.

See also: Collective Behavior, Sociology of; Collective Beliefs: Sociological Explanation; Compliance and Obedience: Legal; Control: Social; Influence: Social; Law: Anthropological Aspects; Law, Sociology of; Legitimacy, Sociology of; Norms; Status and Role, Social Psychology of; Status and Role: Structural Aspects; Values, Sociology of

Bibliography

Becker H 1963 *Outsiders*. The Free Press, New York
Boudon R 1990 *L'art de se persuader*. Fayard, Paris
Boudon 1995 *Le juste et le vrai*. Fayard, Paris
Coleman J 1990 *Foundations of Social Theory*. Harvard University Press, Cambridge, MA
Cusson M 1983 *Le contrôle social du crime*. Presses Universitaires de France, Paris
Durkheim E [1893] 1982 *The Rules of Sociological Method*. Macmillan, London
Mead G H [1934] 1962 *Mind, Self and Society*. University of Chicago Press, Chicago
Merton R K 1957 *Social Theory and Social Structure*, rev. edn. Free Press, Glencoe, IL
Simmel G 1923 Die Mode. In: *Philosophische Kultur. Gesammelte Essais*, 3rd edn. Potsdam, Kiepenheuer
Tarde G [1903] 1962 *The Laws of Imitation*. Peter Smith, Gloucester, UK
Tocqueville A de [1835–40] 1966 *Democracy in America*. Harper, New York
Veblen T [1899] 1953 *The Theory of the Leisure Class: An Economic Study of Institutions*, rev. edn. New American Library, New York
Wittgenstein L [1953] 1967 *Philosophical Investigations*. Blackwell, Oxford, UK

P. Demeulenaere

Congress: United States

Among the world's national legislatures, the United States Congress stands out as far and away the most powerful. It is distinctive in its independence from the executive, its complex internal organization and procedure, and a capacity to develop its own legislative proposals.

Congress is composed of two chambers: a Senate, in which each state has two senators; and a House of Representatives, in which the number of representatives from each state is determined by population. There are 100 senators and 435 representatives. To become law, bills must be passed by both chambers and be signed by the president. Bills vetoed by the president can become law, provided two-thirds of both the House and Senate vote to override; but that is rare. Senators are elected for 6-year terms, with a third of senate seats being filled every 2 years. Representatives are elected to 2-year terms, and all come up for re-election simultaneously.

1. Congressional Elections

Senators are elected at large from their states, while representatives are elected from single member districts (see *First Past the Post, in Politics*, and, by contrast, *Proportional Representation*. Also see *Apportionment: Political*). The 2-year terms of Representatives are the shortest term of any national legislature. The authors of the Constitution intended the House to be the 'people's' chamber, and meant to keep them on a short leash. The 2-year term does this most effectively. Members of the House are for practical purposes always running for reelection, and they must constantly pay attention to constituency concerns. Indeed, it is widely accepted that the desire of members to be reelected has a vast influence on congressional structure and policy making (Mayhew 1974, King 1997).

1.1 Primary Elections

In nearly all places in the United States, party nominations of candidates for Congress are done by means of primary elections, not by party committees or conventions. In other words, the party organizations have almost nothing to do with who runs for Congress under their banner. This system provides members of Congress with a degree of freedom from their party not enjoyed by members of the European parliaments. Members of Congress who fail to follow their leaders are not subject to the ultimate sanction of being denied renomination by their party. Instead of being constrained to satisfy their party, members of Congress are most concerned about pleasing their constituents (see *Primary Elections*).

1.2 Incumbency Advantage

Members of Congress enjoy a substantial 'incumbency advantage' that enables them to be reelected at high rates. Typically, 90 percent or more of House incumbents who seek re-election win. Senate incumbents are somewhat less successful, but still manage to be reelected with great regularity. Incumbency advantage stems from the greater name recognition that incumbents enjoy compared to challengers, from their ability to do favors for constituents, and, perhaps most importantly, from their superior ability to raise money to finance their campaign. Congressional campaigns have become extremely expensive in recent decades,

with the victorious House candidates spending on average $664,000 in 1998, and victorious Senate candidates spending on average $4.8 million (see *Political Money and Party Finance*). According to Jacobson, there is a self-fulfilling quality to the high reelection rate of congressional incumbents. The seeming invulnerability of incumbents, and the daunting challenge of raising the money necessary to mount an effective challenge, discourage many would-be challengers, leaving incumbents to face incompetent, underfunded challengers who have little or no chance of winning, thus further inflating re-election rates (Jacobson 1997).

2. *Internal Organization*

The United States Congress has a complex internal organization that has a great influence on how it translates the preferences of its members into legislation. Congressional organization is dominated by two partially competing institutions—committees and parties. Congress emerged more or less in its current, modern form at the end of the nineteenth century by means of a process of institutionalization (Polsby 1968). Boundaries were established as once rapid turnover declined; internal complexity developed as standing committees became more firmly established; and universalistic decision-making criteria developed as seniority took hold (see *Institutionalization*).

2.1 *Committees*

Both the House and the Senate have a set of standing committees with fixed legislative jurisdictions. Normally, all legislation, once introduced by a representative or senator, is referred to a committee with jurisdiction over its subject matter. All members of Congress are assigned to committees. Senators have three or four committee assignments, while representatives have one or two. The party organizations in Congress make assignments to committees, but generally members are placed on committees where they have some interest. Some committees have more important jurisdictions than others, and members naturally seek to be assigned to those committees. Seats on committees with jurisdiction over spending and taxation are among the most sought after, and the hardest to get.

Committees have great influence in shaping legislation and in determining what legislation will reach the floor—a bill is seldom considered by the entire House or Senate chamber unless it has been recommended favorably by the committee with jurisdiction. Committees do not have the power to force Congress to adopt legislation that its majority does not prefer, but they can block legislation a majority would favor, and thus they serve as an important veto point in the legislative process.

In both the House and the Senate, the chairperson of each committee is normally chosen according to a norm of seniority, whereby the member of the majority party with the longest continuous service on that committee is selected as its chairperson. The seniority norm is violated occasionally, especially when the majority party has just experienced a large influx of new members who have little commitment to seniority. This was the case in both 1975 and 1995, occasions on which the seniority norm was violated in three instances.

There is substantial debate among scholars about the consequences of delegating power to committees. One prevalent view holds that, because the members of a committee may not reflect the views of the entire chamber, they write legislation that advances their own interests, and not those of Congress as a whole. Another view holds that the chamber as a whole would not allow itself to be cheated in this way, and has tools to keep its committees in line and working for the benefit of the whole institution (see *Delegation of Power: Agency Theory*).

The committee system of Congress is commonly believed to strengthen Congress vis-à-vis the executive branch by encouraging its members to specialize. As issue specialists, members of Congress can better resist the executive branch, develop their own legislative proposals, and transform proposals submitted by the executive (Polsby 1975).

2.2 *Party Organization*

Nearly all members of Congress belong to the Democratic or Republican parties. Party is an extremely important aspect of congressional action, but political parties do not command the same degree of loyalty in Congress that they do in most other national legislatures. Party leaders in the House and Senate have a great deal of influence over the agenda, helping to determine what legislation will come to the floor, and when.

Recent party leaders have been motivated by varying conceptions of the leader's role. Newt Gingrich, Speaker of the House from 1995–8, saw it as his responsibility to articulate a vision for the Republican party, and he expected his party to follow him. This was an unusually expansive notion of congressional party leadership, and one which his party was not willing to support indefinitely. His replacement, Dennis Hastert, also a Republican, has returned to a more traditional leadership posture of seeking to identify a set of positions that unite his party, and endeavoring to pass bills that accomplish those purposes. Parties and their leaders sometimes try to enforce party discipline by threatening wayward members with loss of committee assignments, chairmanships, or loss of other perquisites, but this is unusual, and sometimes drives dissidents into the

waiting arms of the other party. The leaders' job might be better described as discovering rather than creating an agenda for the party.

The power of party leaders is limited by the willingness of party members to grant them power. With the members of Congress always eager to appeal to their own unique states and districts, they are reluctant to grant party leaders much power, for that might limit their own ability to pursue initiatives beneficial to their constituents.

To a substantial degree, the power of committees and party leaders is contradictory. To the extent that party leaders are able to increase their power, they do so at the expense of committees and committee chairs. When party leaders seek, as Gingrich did, to determine themselves what the agenda of the House will be, and what legislation will look like, they arrogate to themselves powers that committees and their chairpersons cherish.

2.3 Legislative Procedure

Because of their very different sizes, the House and Senate have fundamentally different procedures. In both, however, the rules of procedure have a profound impact on the legislative product. The House of Representatives, with its large size, cannot tolerate as much individual expression of its members as the Senate can, and its procedure is designed to ensure that the majority prevails. Consequently, the minority party and individual members often feel that they have been crushed by a legislative steamroller. The smaller Senate can more readily accommodate the peculiarities of its members, and accordingly it allows far more freedom for debate and amendment.

2.3.1 House procedure. After important bills are reported favorably from committee, they normally go to the Rules Committee. The Rules Committee drafts a resolution, called a 'special rule,' that, on adoption by the House, determines the conditions under which the bill will be debated, amended, and voted upon. The power and flexibility of special rules to alter, amend, and waive the rules of the House is staggering. Special rules can either make no amendments in order (a closed rule), or place no restrictions on amendments (an open rule). Alternatively, they can make a specified list of amendments in order, they can declare amendments adopted without a vote, or they can amend amendments without the permission of the amendment's author. Special rules can alter or waive the rules of the House for the purposes of considering a particular bill. In short, special rules are infinitely flexible, limited only by the creativity of the Rules Committee, allowing the majority to do nearly anything it wants. Rules written by the Rules Committee must be adopted by a majority vote of the House. In general, the majority party members work closely with their party leaders to ensure that the position preferred by the party is able to prevail, and that potentially troublesome amendments are either forbidden, or considered in such a way that they will do the least possible harm to majority party control.

2.4 Senate Procedure

While action in the House is structured closely by its extensive rules, the Senate, by contrast, has relatively few rules and frequently ignores what rules it has. The Senate gives individual senators an almost unbelievably large amount of discretion and power. This has the predictable consequences sometimes of paralyzing the Senate. The House works by assembling majorities and voting, but the Senate must work by a process of accommodating the individual senators, not by forming majorities, but by working to overcome the objections of individual senators.

The central feature of Senate procedure, around which all else is organized, is the possibility of the filibuster. Filibusters are efforts to defeat bills, not by voting them down, but by means of delay. The goal of filibusterers is to convince supporters of legislation that it will take so much time to pass the bill that they would be better off dropping it and moving on to something else.

Filibusters are possible because of three glaring omissions from Senate rules. The Senate lacks any means of limiting debating time; it lacks a 'germaneness' rule for interventions and amendments; and it lacks a previous question motion. Because the House rules have all of these key features, dilatory tactics are not possible there. The usual filibustering tactics include so-called 'extended debate,' whereby senators talk endlessly, and not necessarily about the bill in question, offering numerous amendments, which need not be germane to the bill, and offering numerous motions, such as suggesting the absence of a quorum, intended only to delay action. Filibusters can be ended by means of the 'cloture' process. If 60 senators are willing to vote for cloture, debate can be ended, but only after all amendments have been disposed of, and all senators have had an additional opportunity to speak for an hour.

Because even a single determined senator can cause a lengthy delay, losing the Senate the opportunity to consider other important legislation, the focus of Senate activity is not on breaking filibusters through cloture, but rather on avoiding them altogether. In advance of bringing legislation to the floor, bill sponsors often engage in extensive discussions with their colleagues in an effort to anticipate and satisfy objections. Most legislation passes the Senate with only minimal objection, in part because of efforts to build consensus early.

3. Congress and the President

By dividing the legislative and executive powers into separate branches, the authors of the US Constitution sought to encourage friction and disagreement in government. This plan has worked well for over two centuries, as Congress remains determinedly independent. Congressional independence is nowhere better demonstrated than in its handling of the federal budget. By statute, the president is required to submit to Congress each year a detailed budget proposal. Nothing, however, requires Congress to pay much attention to presidential proposals. Commonly in recent years the president's budget has been declared 'dead on arrival' at Capitol Hill, and Congress has gone on to draft its own, different budget, relying on its budget process, its committees, and its own budget agency. The president has influence over members of Congress, especially those from the president's own party; however, even members of the president's party are influenced strongly by their constituencies.

4. Congress and Foreign Policy

Congress defers to the executive branch to a great extent in foreign policy, but with important exceptions. Even though the Constitution gives Congress the power to declare war, power over warmaking drifted away from Congress to the president over the twentieth century, in part because of a need for speed and secrecy, and in part because of an unwillingness in Congress to challenge the president over the war power. In trade policy, Congress has since the 1930s mostly ceded trade negotiation power to the executive branch, but periodically becomes involved in trade issues such as automobiles or steel, especially when domestic jobs are threatened by imports.

5. Democratic Accountability

Ideally, individual members of Congress should be held accountable for the collective actions of Congress, but this does not necessarily occur (Arnold 1990). While members of Congress sometimes campaign for reelection by proclaiming the major legislative accomplishments of Congress, often they detach themselves from the institution as a whole and seek reelection for particular services rendered to the district or state. Adopting pork barrel projects such as dams, roads, and parks can help to make individual members popular, but the appearance of waste and venality can make Congress less popular. Fenno argued that representatives and senators distance themselves from Congress as an institution and seek to promote themselves separately, often at the expense of Congress as a whole (Fenno 1978). The accountability of Congress to voters is inherently difficult because Congress acts as an institution, but its members are judged individually.

See also: Accountability: Political; Democratic Party; Electoral Geography; Electoral Systems; Legislatures: United States; Majoritarianism and Majority Rule; Parliamentary Government; Parliaments, History of; Presidency: United States; Republican Party; Third Parties: United States

Bibliography

Arnold R D 1990 *The Logic of Congressional Action*. Yale University Press, New Haven, CT

Binder S 1997 *Minority Rights, Majority Rule: Partisanship and the Development of Congress*. Cambridge University Press, Cambridge, UK

Deering C J, Smith S S 1997 *Congressional Committees*, 3rd edn. Congressional Quarterly, Washington, DC

Edwards G C 1989 *At the Margins: Presidential Leadership of Congress*. Yale University Press, New Haven, CT

Fenno R 1978 *Homestyle: House Members in Their Districts*. Little, Brown, Boston, MA

Jacobson G 1997 *The Politics of Congressional Elections*, 4th edn. Longman, New York

King A 1997 *Running Scared: Why America's Politicians Campaign Too Much and Govern Too Little*. Free Press, New York

Mayhew D 1974 *Congress: The Electoral Connection*. Yale University Press, New Haven, CT

Polsby N W 1968 The institutionalization of the US House of Representatives. *American Political Science Review* **62**: 144–68

Polsby N W 1975 Legislatures. In: Greenstein F I, Polsby N W (eds.) *The Handbook of Political Science, Vol. 5, Governmental Institutions and Processes*. Addison-Wesley, Reading, MA

Tiefer C 1989 *Congressional Practice and Procedure: A Reference, Research and Legislative Guide*. Greenwood, New York

J. B. Gilmour

Conjoint Analysis Applications

1. Introduction

This article provides a general discussion of CA applications and identifies and reviews principal contributions, streams of research, and applications within the overall field. CA is a generic term for methods developed to model and measure preferences and trade-offs that share several common features:

(a) independent variables or 'attributes' that explain preferences are identified and defined (e.g., bus travel times and fares; MHz and prices of PCs);

(b) attribute 'levels' relevant to the research are assigned (e.g., 30, 60 min travel time or \$1, \$3 fare for buses; 600, 733 MHz and \$2,500, \$2,850 prices for PCs);

(c) statistical design techniques are used to combine the attribute levels and create descriptions of options (e.g., 15 min travel time for \$2 or 733 MHz for \$2,850);

(d) a sample of individuals evaluate all or a subset of the descriptions;

(e) individual- or group-level models are estimated from the evaluation data (e.g., transport modes or PC options); and

(f) the estimated models are used to measure subjective quantities (e.g., the subjective value of 600 vs. 733 MHz or implied willingness to pay to save 30 min travel time).

Despite commonalities, CA methods also differ in experimental designs, response tasks, model form(s), and measurement of subjective quantities. For example, initial CA applications were based on an axiomatic theory of the behavior of a ranking of all possible combinations of a discrete set of attribute levels (e.g., Krantz and Tversky 1971). All possible combinations of attribute levels are given by complete factorial enumeration, such that if $A_1(I)$, $A_2(J)$, ..., $A_N(N)$ represent the number of levels of attributes 1, 2, ..., N, then all combinations are given by the expression $I \times J \times \cdots \times N$. Thus, in a full factorial experiment the total combinations equal the product of the levels of each attribute. For example, if the attributes (levels) are bus fare (3), travel time (3), and service frequency (3), then there are $3 \times 3 \times 3$ (or 3^3) = 27 possible bus descriptions.

If a subject ranks all designed descriptions in order of preference (or another subjective dimension), and certain axiomatic conditions are satisfied, the ranking can be represented by an algebraic model defined by marginal and joint attribute effects. For example, for two or more factors, a ranking of a factorial set of options is additive in the attribute effects, if, and only if, the effects of all attributes on the ranking are independent of one another and there are no joint attribute effects (i.e., only marginal effects apply). That is, an additive process underlying a ranking of all combinations of, say, the ith, jth, and kth levels of attributes A_{1i}, A_{2j}, and A_{3k}, respectively, can be expressed as:

$$V_{ijk} = c + v(A_{1i}) + v(A_{2j}) + v(A_{3k}) \tag{1}$$

where V_{ijk} is the overall value of attribute level combination ijk, c is a constant that sets the response scale origin and $v(A_{1i})$, $v(A_{2j})$, and $v(A_{3k})$ are values of the ith, jth, and kth levels of each attribute.

Statistically, if the underlying process is additive, all attribute interactions will be nonsignificant. If the underlying process is multiplicative, the model is as follows:

$$V_{ijk} = c^* + v(A_{1i})v(A_{2j})v(A_{3k}) \tag{2}$$

where all terms were previously defined, except c^*, a constant to set the scale origin. Equation (2) requires all attribute main effects and interactions to be significant. Early applications lacked error theories, hence used goodness-of-fit tests to evaluate and compare models.

Full factorial designs proved too large, so ways to reduce numbers of combinations were proposed, such as fractional or other partial factorial designs, ranking all pairs of attribute levels or using incomplete block designs to select attributes and levels (e.g., Green 1974). Most applications rely on non replicated fractional factorials that cannot be used to test model forms and produce few observations relative to parameters for individuals (Louviere 1988). However one can estimate individual models if response, design, and model assumptions are satisfied (e.g., equal interval response scales, insignificant interactions, independent and identically distributed normal errors).

A new paradigm emerged when McFadden (1974) developed computationally tractable extensions of RUT to the multiple-choice case. This provided ways to estimate and test choice models with a formal theory of errors based on RUT. RUT assumes an unobservable value, U_i, for each of $i = 1, 2, \ldots, I$ options that is a sum of observable (explainable) and random (unexplainable) components:

$$U_i = V_i + \varepsilon_i \tag{3}$$

where V_i and ε_i are, respectively, observable and random components. RUT assumes that individuals try to choose their preferred options, but the random component implies that choices are inherently stochastic from a researcher's perspective: that is, the probability that a subject will choose an option from a set of options is predictable, not the exact choice. If $P(i|C)$ is the probability of choosing option i in set C, and Max = maximum operator, then

$$P(i|C) = P[(Vi + \varepsilon i) > \mathrm{Max}(Vj + \varepsilon j), \text{ for all } j \text{ options in } C] \tag{4}$$

Different discrete choice models are derived by making assumptions about the distributions of the ε's. For example, McFadden (1974) derived the multinomial logit model (MNL) by assuming that the ε's were independently and identically distributed Gumbel random variates. MNL choice probabilities have a simple closed-form, but if the ε's are non-independently and identically distributed normals, the nonclosed form multinomial probit (MNP) model results.

Louviere and Woodworth (1983) integrated RUT-based choice models with CA, and proposed ways to design choice experiments consistent with RUT-based choice models. This choice-based (CB) approach designs attribute level combinations (to estimate evaluation/utility functions) and selects choice sets in which they appear (to satisfy properties of choice models), and the error theory is used to select and test model forms. The CB approach typically estimates models from samples of subjects because individuals provide too few observations to satisfy asymptotic theory and obtain reliable model estimates.

Choice experiments mimic real choice situations. For example, to model choice of travel mode to work, one designs experiments to offer commuters sets of mode choices; to model choice of health insurance plan, one designs experiments that offer subjects sets of health insurance options. Thus, the CB approach designs attribute level combinations to describe sets of options (say M of them). If $M = 1$, subjects evaluate each option separately, making M binary choices (e.g., will/will not choose); if $M > 1$, subjects evaluate sets of multiple options (say S of them), making S multiple choices (Louviere and Woodworth 1983, Louviere et al. 2000).

Significant progress has been made in linking behavior in choice experiments and real environments. Specifically, in RUT models the magnitude (or 'scale') of the parameters are inversely proportional to choice variability. For example, if we denote the scale as λ, it can be defined as follows for MNL models (Ben-Akiva and Lerman 1985 pp. 104–5): $\lambda^2 = \pi^2/6\sigma_\varepsilon^2$ (π is the natural constant; σ_ε^2 = error variance). In empirical applications, analysts estimate the quantities $\lambda\beta_k$, not β_k (β_k is a k-element parameter vector associated with design matrix, X_{ki}, and the ith choice option).

Thus, λ is not identified in any one data source, and typically is set equal to one for convenience. Ben-Akiva and Morikawa (1990) showed that ratios of λ's can be identified and estimated from two or more choice data sources; hence, if experimental and actual choices differ only in response variability, parameters estimated from each will be proportional. Empirical tests in several fields suggest that parameter proportionality often holds to a close first approximation (see Louviere et al. 1999).

This property of RUT-based choice models allows many sources of preference and choice data to be compared, including traditional CA data. For example, one can transform CA ranks into implied choice sets by 'exploding' subjects' complete or partial orders into one or more choice sets (Louviere et al. 2000). That is, for a ranking of M options, the first is preferred to the other M–1, the second to the remaining M–2, etc; hence M rankings yields M–1 choice sets (if A, B, C, D are ranked 1, 2, 3, 4 the first set contains A, B, C, D with A chosen; the second set contains B, C, D with B chosen; etc). Such explosions transform preference responses into discrete choices that can be compared using RUT-based statistical models (e.g., Louviere et al. 2000). Recently Bayesian methods have been used to estimate individual-level parameters (Wedel et al. 1999), but the pros and cons of these methods in CA and CB applications have not been fully explored.

Statistical properties of choice experiments also remain elusive. In particular, humans interact with CA and choice experiments and particular designs can impact response variability. Statistically efficient designs may require many choices, which may increase choice variability; increases in choice variability may offset higher statistical design efficiency. Hence, research into behavioral impacts of designs and effects of non design issues like task layouts, formats, response tasks, etc, on choice variability also are needed.

See also: Decision Theory: Classical; Measurement Theory: Conjoint; Measurement Theory: History and Philosophy; Multidimensional Scaling in Psychology; Preference Models with Latent Variables; Psychometrics; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Anderson N H 1970 Functional measurement and psychophysical judgment. *Psychological Review* **77**: 153–77

Ben-Akiva M E, Lerman S 1985 *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA

Ben-Akiva M, Morikawa T 1990 Estimation of switching models from revealed preferences and stated intentions. *Transportation Research A* **24A**: 485–95

Green P E 1974 On the design of choice experiments involving multifactor alternatives. *Journal of Consumer Research* **1**: 61–8

Hensher D A, Louviere J J, Swait J 1999 Combining sources of preference data. *Journal of Econometrics* **89**: 197–221

Krantz D H, Tversky A 1971 Conjoint measurement analysis of composition rules in psychology. *Psychological Review* **78**: 151–69

Louviere J J 1988 Analyzing decision making: Metric conjoint analysis. Sage University Papers Series in Quantitative Applications in the Social Sciences, Newbury Park, CA, No. 67

Louviere J J, Woodworth G G 1983 Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research* **20**: 350–67

Louviere J J, Hensher D A, Swait J 2000 *Stated Choice Methods: Analysis and Applications in Marketing, Transportation and Environmental Valuation*. Cambridge University Press, Cambridge, UK

Louviere J, Meyer R, Bunch D, Carson R, Delleart B, Hanemann M, Hensher D, Irwin J 1999 Combining sources of preference data for modelling complex decision processes. *Marketing Letters* **10**(3): 205–18

Luce R D 1959 *Individual Choice Behavior*. Wiley, New York

McFadden D 1974 Conditional logit analysis of qualitative choice behaviour. In: Zarembka P (ed.) *Frontiers in Econometrics*. Academic Press, New York, pp. 105–42

Wedel W, Kamakura W, Arora N, Bemmaor A, Chiang J, Elrod T, Johnson R, Lenk P, Neslin S, Poulsen C S 1999 Discrete and continuous representation of unobserved heterogeneity in choice modelling. *Marketing Letters* **10**(3): 219–32

J. J. Louviere

Connectionist Approaches

1. Definitions

To facilitate the following discussion, it will be helpful to first define some terms. A typical connectionist network comprises a (potentially large) number of simple processing *units*. The units are often called (artificial) neurons, but that terminology begs the question of their relation to biological neurons, so it will be avoided in this article. In the most common case, the units form a weighted sum of their (quantitative) inputs and pass the result through a simple, nonlinear activation function, which limits the range of possible outputs. The resulting value is considered the activity of the unit, which may be transmitted to other units (through outgoing connections). In some cases the activity of a unit is a combination of its inputs and previous activity, which provides a kind of 'short-term memory' residing in the collective activities of the units.

The weighted sum results from the fact that each connection in the network has an associated weight (analogous to synaptic efficacy in biological neural networks), which multiplies the quantity transmitted by that connection. Positive weights correspond to excitatory connections and negative weights to inhibitory; zero-valued weights correspond to the absence of a connection. Mathematically, connection weights are often treated as a weight matrix **W**, with element W_{ij} being the weight of the connection to unit i from unit j. Learning and adaptation take place by modification of the weights according to some learning algorithm (Sect. 3); thus the connections constitute the network's 'long-term memory.' 'Connectionism' derives its name from the fact that knowledge resides in the patterns and weights of the connections.

Many connectionist networks are organized into layers, analogous to functional areas in the brain; information usually moves in lockstep from layer to layer. Although many networks are feed-forward, that is, the information moves through successive layers from input to output, other networks are recurrent, which means that there may be feedback connections from a layer to itself or to earlier layers. Recurrent networks are able to recognize and process temporally-extended patterns, that is, sequences of related inputs.

It must be stressed that there are exceptions to all of the preceding general statements about connectionist networks, and 'connectionist approaches' are best viewed as forming a Wittgensteinian 'family resemblance.'

2. History

A short history of connectionist approaches will be presented: first in the narrower context of cognitive science and artificial intelligence; then in the broader context of epistemology, linguistics and the philosophy of mind. Although the size of this article does not permit detailed citations of the literature, many of the seminal articles are collected in Anderson and Rosenfeld (1988) and Anderson et al. (1990). (See *Cognitive Psychology: History*; *Cognitive Science: History*.)

2.1 Narrower History

According to the second edition of the *Oxford English Dictionary*, the term 'connectionism' was first used by E. L. Thorndike, in his *Fundamentals of Learning* (1932), to refer to the reduction of mental processes to the connections between stimuli and responses (that is, to a form of associationism), and so connectionist theories have been set in opposition to cognitive theories. The term is used somewhat differently now (so that 'connectionist cognitive science' is not an oxymoron), but retains some similarities to associationism. However, to understand the relation it is better to look at connectionism from the perspective of neural network models of cognition.

In the early 1940s W. S. McCulloch and W. Pitts investigated the computation of logical functions by simple neuron-like elements; in effect they showed that these elements could compute logical 'and,' 'or,' 'not,' and so forth. In his *Organization of Behavior* (1949) D. O. Hebb suggested that learning takes place by the formation of cell assemblies and that this occurs through the strengthening of connections between simultaneously active neurons in neural networks, which are initially randomly connected. His description of this process inspired one of the simplest connectionist learning rules (Sect. 3.1).

In the late 1950s F. Rosenblatt began to investigate the application of simple neuron models called perceptrons to perceptual problems such as classifying printed letters (see *Perceptrons*). He developed a learning algorithm for simple (single-layer) perceptron networks, which iteratively adjusted the connection weights whenever the network made a mistake. He proved that, if the network were capable of solving the problem at all, then the algorithm would eventually find the connection weights to solve it. However, there

are many problems that a single-layer network cannot solve, and Rosenblatt never succeeded in finding a multilayer learning algorithm.

A key event in the history of connectionism was the publication of M. Minsky and S. Papert's *Perceptrons* (1969), which demonstrated limitations of simple perceptron networks. Specifically, they proved that single-layer perceptron nets could discriminate only those categories that are linearly separable (see *Linear Algebra for Neural Networks*; *Perceptrons*). Their proof did not apply to multilayer nets (for which there was still no learning algorithm), but they suggested that similar limitations would be found for these too. Nevertheless, their book was widely interpreted as showing the impotence of neural networks in general, and is commonly blamed for discouraging research in the field for a decade. (The extent to which it actually did so is, perhaps, a topic for historians of science.)

It will be worthwhile to comment on the role of holography in the development of connectionist approaches. As early as 1929 K. S. Lashley had conducted experiments suggesting that individual memory traces were not localized in any one place, and that degradation of memory was proportional to the amount of cortical mass destroyed, thus implying that individual traces were distributed over large areas of cortex. In his well-known 1950 paper, he despaired of ever understanding how such a nonlocal memory could operate. The principles of holography had been described by D. Gabor in 1948, but it was not until the advent of optical holography in the early 1960s that it began to be seen as a solution to Lashley's dilemma. Although an analogy between holography and memory had been suggested as early as 1963 by P. J. van Heerden, the 'holographic hypothesis' has been developed most extensively by K. H. Pribram and his colleagues since 1966 (see, e.g., Anderson et al. (1990), ch. 7). In the late 1960s and early 1970s holographic and holography-inspired models of associative memory were also investigated by H. C. Longuet-Higgins, D. J. Willshaw and others (Hinton and Anderson, 1989). Some connectionists were influenced by the critiques of traditional, rule-based AI by the phenomenologist philosopher H. L. Dreyfus (*What Computers Can't Do* 1972). Although he stressed the limitations of rule-based systems, he also suggested that some of these limitations would not apply to analog systems operating on principles similar to holography (see *Cognition*, *Distributed*).

Although a number of investigators (including J. A. Anderson, S. Grossberg, and T. Kohonen) continued connectionist research through the 1970s, the field was rejuvenated by the work of D. E. Rumelhart, J. L. McClelland, and other members of the 'PDP (Parallel Distributed Processing) Working Group,' many of whose publications were collected in a widely-read two-volume set (Rumelhart et al., 1986). The credibility of connectionist approaches was also enhanced by J. Hopfield's publication in 1982 of a simple recurrent net capable of associative memory and pattern completion.

2.2 Broader History

Although connectionism can be viewed as an approach to knowledge representation and inference that is of relevance only to cognitive science, in fact it has much broader implications, for it challenges assumptions about knowledge that have been largely unquestioned since ancient Greek philosophy. Already in the philosophies of Socrates, Plato, and Aristotle there is a preference for knowledge expressed as logical relations among discrete, language-like structures, and for a view of cognition as mechanized deduction. These ideas influenced many later philosophers, including Hobbes (who equated thinking with computation), Leibniz (who experimented with formalized systems of knowledge representation and mechanical deduction), and Boole (who invented mathematical logic).

These ideas were also influential in the development of logical positivism, which dominated the philosophy of science in the first half of the twentieth century. The idea persisted in the assumption that there must be a 'language of thought,' because no alternative is imaginable, and it is 'the only game in town.' Similarly, most research in artificial intelligence (AI) took for granted that intelligence resides in the structures of a *knowledge representation language*, and in deduction-like formal rules for their manipulation. Throughout the 1970s (the 'connectionist dark ages'), AI researchers concentrated their attention on expert systems, which depended on expertise represented symbolically (see *Expert Systems in Cognitive Science*). Disappointment with the performance of these systems was one of the motivations for the connectionist renaissance.

In summary, the Western tradition (with some exceptions) has displayed a kind of 'linguistic chauvinism,' which presumes that all knowledge and cognition can be expressed in language-like structures. Knowledge is expressed at a symbolic level, that is, in terms of atomic (indivisible), word-level categories related by sentence-like logical structures. On the other hand, most connectionist approaches represent knowledge at a *subsymbolic* level, that is, in terms of minute, quantitative features related by low-level, often statistical, connections. In other words, knowledge is more akin to an *image* than to a *sentence* (see *Imagery versus Propositional Reasoning*; *Mental Imagery, Psychology of*). Therefore, some of connectionism's advocates see it as a fundamentally new view of knowledge and cognition, which is leading to a paradigm shift in cognitive science and philosophy and is engendering a 'new AI.' Specific innovations of the connectionist approach are discussed in Sect. 5.1. See *Artificial Intelligence: Connectionist and Symbolic Approaches*; *Knowledge Representation*; *Production Systems in*

Cognitive Psychology; *Schemas, Frames, and Scripts in Cognitive Psychology*.

3. Mechanisms of Adaptation and Learning

Virtually all connectionist approaches incorporate adaptive mechanisms or learning algorithms, which allow the network to improve its performance; here a few will be discussed briefly.

3.1 Correlational Learning

Correlational learning ('Hebb's rule'), the simplest connectionist learning algorithm, takes its inspiration from Hebb's hypothesis that the simultaneous activity of two neurons strengthens the connection between them. It makes a change in connection weight proportional to the product of activities of the units it connects. Thus, the change in the weight W_{ij} of the connection to unit i from unit j is proportional to $y_i x_j$, where y_i is the activity of unit i and x_j is the activity of j. This learning rule can be viewed as a highly simplified model of long-term potentiation. See *Long-term Potentiation and Depression (Cortex)*; *Learning and Memory, Neural Basis of*; *Synaptic Efficacy, Regulation of*.

The effect of this rule is that the weight becomes a correlation coefficient between the activities of the units it connects. That is, the connection will become stronger (more positive) to the extent that the units are simultaneously positive or simultaneously negative. The connection will become more inhibitory (more negative) to the extent that one unit is positive while the other is negative. If there is, on average, no systematic relation between the activities of the two units, then the weight will tend toward zero, effectively disconnecting the units.

The correctional learning rule is the basis of a simple associative memory known as a *linear associator*. In this connectionist network, there are two layers of linear units, an input layer and an output layer; each input unit is connected to every output unit, so that the output is a linear function of the input, $\mathbf{y} = \mathbf{W}\mathbf{x}$. A series of pairs of pattern vectors $(\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{y}_p, \mathbf{x}_p)$ may be presented to the input and output layers of such a network and the weights adjusted according to the learning rule. The goal of the linear associator is that the network associate each $\mathbf{x}_k$ with the corresponding $\mathbf{y}_k$. In fact it can be shown that if the set of input patterns is orthogonal, then the output $\mathbf{W}\mathbf{x}_k$ will be proportional to the desired $\mathbf{y}_k$ (see *Linear Algebra for Neural Networks*).

3.2 Delta Rule

An improvement called the delta rule can be made in the linear associator; it illustrates a fundamental approach to connectionist learning. The idea is to define the error, as a function of the weight matrix, as the sum of the differences between the desired and actual outputs of the network, $E(\mathbf{W}) = \sum_{k=1}^{p} D(\mathbf{y}_k, \mathbf{W}\mathbf{x}_k)$. The weight matrix is then changed by gradient descent which means that the elements of $\mathbf{W}$ are changed in the relative proportion that causes a maximal incremental decrease of $E(\mathbf{W})$.

If the difference between patterns is measured by Euclidean distance squared, $D(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|^2$ (that is, the sum-of-squares error), then the delta rule is essentially equivalent to linear regression. If the input patterns are linearly independent, then the delta rule will converge to a weight matrix that associates perfectly, $\mathbf{y}_k = \mathbf{W}\mathbf{x}_k$. If they are not, then the weight matrix will be that which minimizes the total error; that is, it will be the best linear prediction of the output patterns from the input patterns (see *Linear Algebra for Neural Networks*).

The delta rule also illustrates an important characteristic of most connectionist networks: their ability to *generalize* to inputs other than those upon which they have been trained. The delta rule provides only linear generalization, but other algorithms, such as backpropagation (Sect. 3.3) can make nonlinear generalizations. Typically, connectionist categories are represented by concrete prototypes rather than by definitions in terms of necessary and sufficient conditions or other abstract symbolic structures. Network behavior then depends on similarity to the prototypes rather than on formal manipulation of symbolic structures.

3.3 Backpropagation

Gradient descent may be applied also to multilayer networks of nonlinear units, so long as the activation function is differentiable. The backpropagation algorithm (also called the *generalized delta rule*) efficiently computes the weight changes by starting with the last layer and working backward layer by layer. It has been rediscovered a number of times, perhaps first by P. Werbos in 1974, but its importance in connectionism began with its rediscovery in the early 1980s. There are also special adaptations of backpropagation for recurrent networks. Backpropagation has a number of limitations, including: (a) it may be quite slow, (b) it does not necessarily take the shortest path to an error minimum, and (c) it may get trapped in local minima. Nevertheless, it remains a fundamental learning algorithm and has been subject to many practical improvements.

3.4 Other Learning Algorithms

The preceding are examples of supervised learning procedures, which means that the 'correct answer' $\mathbf{y}_k$ is available for each training input $\mathbf{x}_k$. Although this is appropriate for modeling some cognitive processes and for many practical problems, for others unsupervised learning is preferable. In these procedures, the

network is not trained to produce any specific outputs, but it allowed to group or categorize inputs according to standards inherent in its design. Thus unsupervised learning is often equivalent to some kind of statistical clustering. Between these two extremes is reinforcement learning, in which the algorithm is told whether or not the output is correct, but not what the correct output is. In unsupervised and reinforcement learning, as in supervised learning, the network is normally expected to generalize reasonably to novel inputs. There are now hundreds of connectionist learning algorithms, of greater and lesser relevance to cognitive science and neuroscience, but this must suffice for an introduction.

4. *Example: Learning Past Tenses*

Since the early 1980s, connectionist networks have been used for an enormous number of practical applications and for modeling many aspects of cognition. One notable example must suffice here; see Rumelhart et al. (1986) for additional examples.

Rumelhart et al. (1986, Vol. 2, Chap. 18) trained a connectionist network to produce the past tenses of English verbs. The inputs were vectors representing phonological features of the present tenses and the outputs were vectors representing the phonological features of the corresponding past tenses. Learning was observed to pass through three stages. In the first stage, the most common verbs were learned, essentially by rote as individual special cases. In the second stage, the network learned how to form regular past tenses (for it was able to generalize to novel regular verbs), but over-generalized by treating the (previously correctly processed) irregular verbs as though they were regular. In the third stage, the network (re)learned the correct formation of the irregular past tenses without losing its ability to form regular past tenses. It is interesting and highly suggestive that children pass through these same three stages; the model also exhibited errors of the same kinds made by children.

Although this experiment can be (and has been) criticized on a number of grounds as a model of language learning, it is perhaps more valuable as a demonstration of how a connectionist network can exhibit apparently rule-like behavior, but not be following any explicit rules. In particular, exceptions to the rules are handled automatically without explicit accommodation. Thus it is paradigmatic of connectionist cognitive models.

5. *Issues*

5.1 *Connectionist Representations*

Connectionist approaches represent and process information in a way that is fundamentally different from *symbolic* approaches, in which knowledge is represented in discrete structures relating atomic lexical-level features (i.e., categories of the sort for which natural languages have words). In symbolic approaches, information is processed by formal logic-like rules, which rearrange these atomic units of meaning. In connectionist approaches, on the other hand, information is represented by patterns of activity distributed over large numbers of units, which individually have no lexical-level meaning. The latter are often termed microfeatures, but they are fundamentally different from features, which are supposed to be complete, context-free units of meaning. Microfeatures are just components of distributed representations and are usually individually uninterpretable.

Instead of rules, connectionist information processing is defined by quantitative connections between microfeatures and takes place at a subsymbolic level (Smolensky 1988). Cognition then is an emergent effect of large numbers of these interactions (which therefore constitute the microstructure of cognition). Indeed, according to this account, apparent symbolic processing is just such an emergent effect of these subsymbolic interactions.

Traditional (symbolic) models of knowledge have been criticized for their *brittleness*. For example, when a set of rules is formulated in an attempt to model the behavior of some expert, it is generally found that the rules do worse than the expert, since the expert applies rules flexibly and makes *ad hoc* exceptions to them as required by the particulars of the situation. Of course, additional rules can be formulated to cover the exceptions, but then these are likewise found to have exceptions, and so forth. The attempt to reduce flexible behavior to inflexible rules leads to a combinatorial explosion of possibilities which exceeds the capacities of brains and computers.

Connectionist approaches seem better able to account for the flexibility and context-sensitivity of natural intelligence. This is because the connection weights function as a large number of soft constraints, none of which is individually necessary or sufficient to produce a result. Therefore connectionist networks can accommodate inputs that are exceptional in various ways, either by ignoring the anomalous aspects, or by corresponding adjustment of their output. As a consequence, their performance is also robust in the face of noise in the input or damage to the network. Further, to the extent that microfeatures of the environment are represented in the input, the network can process information in a way that is sensitive to the context. (See *Artificial Intelligence: Connectionist and Symbolic Approaches*; *Knowledge Representation*; *Production Systems in Cognitive Psychology*; *Schemas, Frames, and Scripts in Cognitive Psychology*.)

5.2 *Criticisms*

Connectionist networks have been criticized for their *opacity* or *uninterpretability*. That is, when a connec-

tionist network has been trained to perform some task, it is difficult to extract human-interpretable rules from the network; although the network performs correctly, one cannot understand the 'rules' it is apparently following. The reason of course is that it is *not* following rules, and the individual units and weights represent microfeatures and constraints that are lexically meaningless (i.e., have no lexical-level meaning). There are mathematical procedures for extracting rule-like information from networks, but they give only approximations to the network's behavior. This is analogous to what has been observed during 'knowledge acquisition' for expert systems: after the fact, human experts can account in terms of rules for their decision making, but the rules do not account adequately for the expert's future decisions.

Quite naturally, some of the severest criticisms of connectionism have come from linguists and other cognitive scientists committed to a 'language of thought,' that is, to the hypothesis that cognition must be understood in terms of the manipulation of propositional or sentential symbolic structures. These criticisms have focused on the alleged inability of 'flat' connectionist representations to capture the rich hierarchical symbolic structure of human language and propositional attitudes, and the related sensitivity of cognition to their constituent structure. However, experiments in 'connectionist symbol processing' have shown that connectionist networks can be sensitive to the constituent structure of representations without explicit representation of that structure and without the use of explicit symbolic rules (e.g., Sect. 4). There is much more to the issue than this, however, and the early collection by Pinker and Mehler (1988) is still a good introduction to the anti-connectionist position. (see *Connectionist Models of Language Processing*).

5.3 Computability

If connectionism is viewed as a fundamentally new approach to information representation and processing, then the question arises of its power relative to conventional digital computation, as modeled by the Turing machine. At a basic level this question is easy to answer. On one hand, since connectionist networks are routinely simulated on digital computers, it is apparent that they have no greater power than a Turing machine. On the other hand, researchers have shown that various sorts of connectionist networks can simulate Turing machines, which therefore have no greater power than the networks. The conclusion would seem to be that connectionist networks are equivalent to Turing machines in computing power.

However, at a deeper level the question is problematic, for the Turing machine model is based on certain idealizing assumptions about what is significant and insignificant in models of computation, assumptions that are questionable when applied to connectionist models. In particular, the Turing machine model is based on the assumption that computation proceeds by the recognition of atomic tokens of definite type according to the discrete application of finite, definite rules; these processes are assumed to operate with complete reliability. These assumptions are a poor match to connectionist approaches, in which information is represented in distributed patterns of continuous activity, and in which recognition is a matter of degree. Nevertheless, questions of computability are also important in connectionism, but relevant answers may require the development of a new theory of computation that makes idealizing assumptions more relevant to connectionist approaches.

5.4 Relation to Biological Neural Networks

Connectionist networks are often called 'neural networks' and described in terms of (artificial) neurons connected by (artificial) synapses, but is this more than a metaphor? Generally, connectionist models have reflected the contemporary understanding of neurons. For example, McCulloch and Pitts focused on the 'all or nothing' character of neuron firing, and modeled neurons as digital logic gates. Newer connectionist models have had a more analog focus, and so the activity level of a unit is often identified with the instantaneous firing rate of a neuron. However, these models still ignore many important properties of real neurons, which may be relevant to neural information processing (Rumelhart et al., 1986′, vol. 2, Chap. 20). As a consequence neuroscientists have stressed the differences between biological neurons and the simple units in connectionist networks; the relation between the two remains an open problem. Nevertheless, it is much easier to envision neural implementations of connectionist networks than of symbol-processing architectures.

See Churchland (1986) and Quinlan (1991) for an introduction to connectionist approaches in philosophy and psychology. See *Connectionist Models of Concept Learning*; *Connectionist Models of Development*.

See also: Artificial and Natural Computation; Artificial Neural Networks: Neurocomputation; Connectionist Models of Concept Learning; Connectionist Models of Development; Mathematical Learning Theory; Mathematical Learning Theory, History of; Mathematical Psychology; Mathematical Psychology, History of; Neural Networks and Related Statistical Latent Variable Models; Neural Networks: Biological Models and Applications

Bibliography

Anderson J A, Rosenfeld E (eds.) 1988 *Neurocomputing: Foundations of Research*. MIT Press, Cambridge, MA

Anderson J A, Pellionisz A, Rosenfeld E (eds.) 1990 *Neurocomputing 2: Directions for Research*. MIT Press, Cambridge, MA

Churchland P S 1986 *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. MIT Press, Cambridge, MA

Dreyfus H L 1972 *What Computers Can't Do*. Harper & Row, New York

Gabor D 1948 A new microscopic principle. *Nature* **161**: 777–8

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Hinton G E, Anderson J A (eds.) 1989 *Parallel Models of Associative Memory*, rev. edn. Lawrence Erlbaum Associates, Hillsdale, NJ

Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA* **79**: 2554–8

Lashley K S 1950 In search of the engram. *Symposia of the Society for Experimental Biology* **4**: 454–82

Minsky M, Papert S 1969 *Preceptrons*. MIT Press, Cambridge, MA

Pinker S, Mehler J (eds.) 1988 *Connections and Symbols*. MIT Press, Cambridge, MA

Quinlan P 1991 *Connectionism and Psychology: A Psychological Perspective on New Connectionist Research*. University of Chicago Press, Chicago

Rumelhart D E, McClelland J L, PDP Research Group 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA

Smolensky P 1988 On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1–74

Thorndike E L 1932 *Fundamentals of Learning*. Teachers College, New York

B. J. MacLennan

Connectionist Models of Concept Learning

A category is a collection of objects or events. Concepts can be seen as mental representations of categories, and concept learning can be defined as learning to assign objects or events to categories. The study of concept learning has a long history in psychology. Experimental research on concept learning has revealed several regularities, which seem to apply across a wide range of stimuli and learning conditions. These regularities must be explained by any valid model of concept learning. A detailed overview of basic findings on concept learning is provided in *Concept Learning and Representation: Models*. Recently, connectionist models of concept learning have become particularly popular (e.g., Ashby et al. 1998, Gluck and Bower 1988a, 1988b, Kruschke 1992). In connectionist models, it is assumed that concept learning involves the development of associative links between objects and categories. These links form the basis of categorization, which is the process of assigning an object to a category. In this article, the basic principles of connectionist models of concept learning are presented, and their empirical and theoretical merits are discussed. There are so many different connectionist models of concept learning that it is impossible to cover them all here. Instead, the article will focus on a few models, which are well established and have been particularly influential. The discussion starts with a simple model of concept learning (the delta-rule model), which has formed the basis of much research on connectionist models of concept learning. Next, the fundamental limitations of the delta-rule model are highlighted. Finally, a recent alternative to the delta-rule model is presented.

1. The Delta-rule Model

One of the simplest and best understood learning rules in connectionist models is the so-called delta rule, which was proposed by Widrow and Hoff (1960) and which is formally equivalent to the learning rule in Rescorla and Wagner's (1972) model of associative learning (see Sutton and Barto 1981). The delta rule is an error-correcting rule. It assumes that feedback is available on the classification performance of a network, and that this feedback is available for each stimulus that must be classified. A learning scheme that fulfills this condition is called a supervised learning scheme.

To illustrate the delta rule, assume that we have the following simple model of concept learning (Fig. 1). The model takes the form of a connectionist network, with two connected layers of processing units. The input layer is used to represent or encode the stimuli

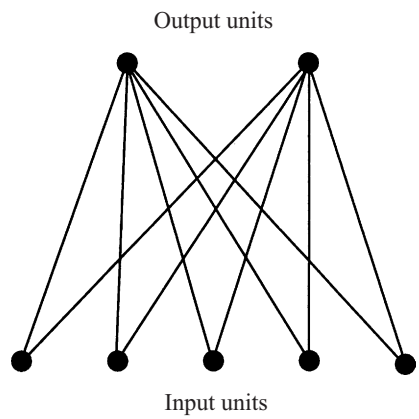

Figure 1
Delta-rule network architecture

that can be presented, whereas the output layer represents the response alternatives (in a categorization context, each response alternative typically corresponds to a single category). Input coding is achieved by a number of units, each of which corresponds to a single feature or dimension of the stimulus. Furthermore, each input unit is connected to each output unit. The connections in the network have associated strength values. The strength of a connection determines the effect that activation of its source or input unit has on its output unit. In a network of this form, category learning is assumed to involve the modification of the strengths of the connections between input and output units, in such a way that presentation of a stimulus to the input units will give rise to the activation of the output unit that corresponds to the category to which the stimulus belongs.

The input to the network on a given trial is represented by the pattern of activation across the input units. Suppose that each input unit represents a stimulus feature that can either be present or absent (such as 'red,' or 'large'). The set of features that characterizes a stimulus can then be represented by setting the activation of each input unit that corresponds to a feature that is present to 1, whereas the activation of the other units can be set to 0. Of course, other representation schemes are possible, depending on the nature of the stimuli and the requirements of the learning task.

Upon presentation of a stimulus, the activation of the output units is computed, as follows:

$$a_o = \sum_i w_{io} a_i \tag{1}$$

in which a_o is the activation of output unit o, a_i is the activation of input unit i, and w_{io} is the strength of the connection between input unit i and output unit o. Next, feedback about the network's output activation pattern is provided, and the discrepancy between the current output pattern and the desired output pattern is computed. In a concept learning task, the desired output would typically be that one output unit (corresponding to the correct category) is activated, whereas the other units remain inactive. For each output unit, an error value e_o is computed:

$$e_o = d_o - a_o \tag{2}$$

in which d_o is the desired output value for unit o. Finally, the connection strengths in the network are modified, as follows:

$$\delta w_{io} = \lambda a_i e_o \tag{3}$$

in which δw_{io} is the change in the connection strength and in which λ is a learning rate parameter.

After a number of presentations of the entire set of stimuli in a category-learning task, the delta rule is guaranteed to produce a set of connection strengths that is optimal for the category-learning task at hand, within the constraints imposed by the architecture of the network (Kohonen 1977). For some category structures, the simple network in Fig. 1 is sufficient to achieve perfect categorization of all stimuli. If perfect categorization is impossible (for reasons that will be explained below), the delta rule still converges upon a set of connection strengths that produces the best performance that can be achieved with the network to which it is applied.

2. *Evaluation of the Delta-rule Model*

Delta-rule networks have been evaluated in a large number of empirical studies on concept learning (e.g., Estes et al. 1989, Friedman et al. 1995, Gluck and Bower 1988a, 1988b, Shanks 1990, 1991), with considerable success. However, although the delta-rule model can explain important aspects of human concept learning, it has a major weakness: It fails to account for people's ability to learn categories that are not linearly separable. In a network of the kind described above, the activation of any output unit is always a weighted sum of the activation of the input units. This implies that the network can only learn categories that can be separated by a linear function of the input values. In human concept learning, linear separability does not appear to be an important constraint. It has been shown repeatedly that people can learn nonlinearly separable category structures without difficulty (e.g., Medin and Schwanenflugel 1981, Nosofsky 1987).

There are several ways in which delta-rule networks can be modified to handle nonlinearly separable categories. One way is to augment the input layer of the network with units that only become active when particular combinations of two or more features are present in the input (e.g., Gluck and Bower 1988b, Gluck 1991). If enough configural units are present (and if they encode the correct combinations of features), such networks can be made to learn any category structure using the delta rule. However, it is not clear that learning in such networks corresponds well to human learning, or that configural cue networks explain categorization after learning (Choi et al. 1993, Macho 1997, Nosofsky et al. 1994, Palmeri 1999). Moreover, the number of possible configural units grows exponentially as the number of stimulus dimensions becomes larger. Configural cue models are therefore not particularly attractive as models of human concept learning.

A second way of modifying delta-rule networks so that they can learn nonlinearly separable categories

involves the use of a layer of 'hidden' units, between the input units and the output units. Rumelhart et al. (1986) proposed a generalization of the delta rule for such networks. This back-propagation-of-error rule is used to determine how much the connection strengths between input and hidden units, and between hidden and output units should be changed on a given learning trial, in order to achieve the desired mapping between input and output. If the units in such multilayer networks have nonlinear activation functions, back-propagation networks can learn nonlinearly separable categories; in fact, they can learn arbitrary mappings between inputs and outputs, provided that they contain a sufficient number of hidden units (Hornik et al. 1989).

Despite their intuitive appeal and obvious computational power, backpropagation networks are not adequate as models of human concept learning (e.g., Kruschke 1993). Their most important shortcoming is that they suffer from 'catastrophic forgetting' of previously learned concepts when new concepts are learned (McCloskey and Cohen 1989). To overcome this difficulty, Kruschke (1992) has proposed a hidden-unit network that retains some of the characteristics of backpropagation networks, but that does not inherit their problems.

3. *ALCOVE*

ALCOVE (Kruschke 1992, 1993) is a connectionist network with three layers of units: an input layer, a hidden layer, and an output layer (see Fig. 2). The model is based on an exemplar theory of concept learning and categorization, Nosofsky's (1986) Generalized Context Model (GCM; see *Concept Learning and Representation: Models*). Each input unit i of ALCOVE encodes a single stimulus dimension and is gated by a dimensional attention weight α_i, which reflects the relevance of the dimension for the learning task at hand.

Each hidden unit has a 'position value' on each stimulus dimension, which means that each hidden unit corresponds to a particular stimulus or exemplar. Whenever an input pattern is presented to the network, each hidden node is activated according to the similarity between the stimulus it represents and the input pattern. In ALCOVE, similarity is defined as in Nosofsky's GCM:

$$a_j^{hid} = \exp\left[-c\left(\sum_i \alpha_i \,|\, h_{ji} - a_i^{in} \,|^r\right)^{q/r}\right] \tag{4}$$

in which a_j^{hid} is the activation of hidden unit j, h_{ji} is the position of hidden unit j on stimulus dimension i, c is a positive constant called the specificity of the hidden unit, a_i^{in} is the activation of input unit i, and where r and q determine the similarity metric and similarity gradient, respectively.

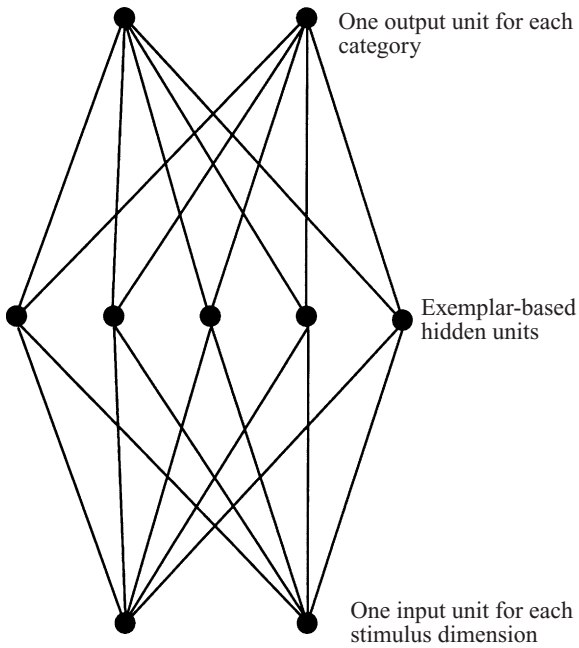

Figure 2
Architecture of the ALCOVE model of concept learning

Each hidden unit is connected to output units that correspond to response categories. The category units are activated according to the following rule:

$$a_k^{out} = \sum_j w_{kj} a_j^{hid} \tag{5}$$

in which w_{kj} is the association weight between hidden unit j and category unit k. This is the same activation rule used in the simple delta-rule network discussed above (see Eqn. (1)). Finally, category unit activations are translated into response probabilities by the rule

$$P(K) = \exp(\varphi a_K^{out}) \Big/ \sum_m \exp(\varphi a_m^{out}) \tag{6}$$

where φ is a scaling constant. ALCOVE employs a variation of the backpropagation learning rule to adjust dimensional attention weights α_i and association weights w_{kj} in the course of learning (see Kruschke 1992, for details of the learning rule). As a result of these adjustments, the network will eventually learn to classify each stimulus into the correct category.

ALCOVE has great advantages over the simple delta-rule network for concept learning. The model is not affected by the linear separability constraint. Like standard backpropagation networks, ALCOVE can

learn arbitrary mappings between stimuli and categories. Moreover, ALCOVE does not suffer from catastrophic forgetting (Kruschke 1993). Most important, however, is the close correspondence between ALCOVE's predictions about concept learning and human performance. In a large number of experiments, the model has been tested successfully (e.g., Choi et al. 1993, Kruschke 1992, 1993, Nosofsky et al. 1992, Palmeri 1999). ALCOVE ultimately derives its strength from its combination of the principles of exemplar-based processing with those of associative learning.

4. Conclusions and Future Directions

Whereas connectionist models such as ALCOVE can explain many important aspects of human concept learning, it is becoming increasingly clear that they also have fundamental limitations. The models that were reviewed here all assume that concept learning is an associative process, in which links between stimulus and category representations are modified. There is little doubt that many concepts are learned in this way. However, the associative model does not apply to the learning of all concepts. Some concepts are learned by a process of rule discovery, which has characteristics that are very different from those of connectionist models of learning. An important challenge for the future will be to determine when associative models and rule-based models of concept learning apply. There have been some recent attempts to develop hybrid models, which combine associative and rule-based learning principles (e.g., Erickson and Kruschke 1998), and it is likely that such models will become increasingly prominent.

See also: Categorization and Similarity Models; Categorization and Similarity Models: Neuroscience Applications; Concept Learning and Representation: Models; Connectionist Approaches; Early Concept Learning in Children; Implicit Learning and Memory: Psychological and Neural Aspects

Bibliography

Ashby F G, Alfonso-Reese L A, Turken A U, Waldron E M 1998 A neuropsychological theory of multiple systems in category learning. *Psychological Review* **105**: 442–81

Choi S, McDaniel M A, Busemeyer J R 1993 Incorporating prior biases in network models of conceptual rule learning. *Memory and Cognition* **21**: 413–23

Erickson M A, Kruschke J K 1998 Rules and exemplars in category learning. *Journal of Experimental Psychology: General* **127**: 107–40

Estes W K, Campbell J A, Hatsopoulos N, Hurwitz J B 1989 Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15**: 556–71

Friedman D, Massaro D W, Kitzis S N, Cohen M M 1995 A comparison of learning models. *Journal of Mathematical Psychology* **39**: 164–78

Gluck M A 1991 Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science* **2**: 50–55

Gluck M A, Bower G H 1988a Evaluating an adaptive network model of human learning. *Journal of Memory and Language* **27**: 166–95

Gluck M A, Bower G H 1988b From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General* **117**: 227–47

Hornik K, Stinchcombe M, White H 1989 Multilayer feed-forward networks are universal approximators. *Neural Networks* **2**: 359–66

Kohonen T 1977 *Associative Memory: A System Theoretical Approach*. Springer, New York

Kruschke J K 1992 ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* **99**: 22–44

Kruschke J K 1993 Human category learning: Implications for backpropagation models. *Connection Science* **5**: 3–36

Macho S 1997 Effect of relevance shifts in category acquisition: A test of neural networks. *Journal of Experimental Psychology: Learning, Memory and Cognition* **23**: 30–53

McCloskey M, Cohen N J 1989 Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation* **24**: 109–65

Medin D L, Schwanenflugel P J 1981 Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory* **7**: 355–68

Nosofsky R M 1986 Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* **115**: 39–57

Nosofsky R M 1987 Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **13**: 87–108

Nosofsky R M, Gluck M A, Palmeri T J, McKinley S C, Glauthier P 1994 Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins 1961 *Memory and Cognition* **22**: 352–69

Nosofsky R M, Kruschke J K, McKinley S C 1992 Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory and Cognition* **18**: 211–33

Palmeri T J 1999 Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin and Review* **6**: 495–503

Rescorla R A, Wagner A R 1972 A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black A H, Prokasy W F (eds.) *Classical Conditioning II: Current Research and Theory*. Appleton-Century-Crofts, New York, pp. 64–99

Rumelhart D E, Hinton G E, Williams R J 1986 Learning representations by back-propagating errors. *Nature* **323**: 533–36

Shanks D R 1990 Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology* **42A**: 209–37

Shanks D R 1991 Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **17**: 433–43

Sutton R S, Barto A G 1981 Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review* **88**: 135–70

Widrow C, Hoff M E 1960 Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record* **4**: 96–104

K. Lamberts

Connectionist Models of Development

Connectionism is a style of computational modeling inspired by principles of brain functioning and by the mathematics of statistical mechanics. It is implemented as networks of interconnected units, where units correspond roughly to neurons and connections correspond roughly to synapses. Momentary patterns of activity in units can be interpreted as cognitions or percepts; weighted connections between units can be viewed as long-term knowledge, modifiable by learning. Since the late 1980s, cognitive modelers have become interested in applying such neural networks to data on psychological development in children. Among the attractive characteristics of neural networks are graded knowledge representations, capacity for change and self-organization, feasibility of studying environment–heredity interactions, and neurological plausibility. This article reviews developmental connectionist models, showing how various modeling techniques have contributed to the understanding of a variety of developmental issues in cognitive and perceptual development.

1. *Techniques*

There are several neural network techniques that have been applied to developmental issues. Virtually all of them employ layered, feed-forward networks in which activation is passed forward from input units to hidden units and on to output units. Sometimes there are multiple layers of hidden units, other times no hidden units at all. These networks learn by processing examples of input–output pairs. Presentation of an input vector of activation results in a vector of output activation. Hidden and output unit activations are typically determined by summing the products of sending-unit activations and incoming weights and passing this weighted sum through a nonlinear activation function, such as a sigmoid or hyperbolic-tangent function. A linear activation function typically is used for continuous outputs.

In supervised learning, error at the output units is computed as the sum of squared differences between outputs and targets. Weights are adjusted to reduce this discrepancy, using the delta rule in which weight change is proportional to error. In multilayer networks, error is typically propagated backward to earlier layers of weights, a technique known as back-propagation. Unsupervised learning can also occur, without output targets, as networks learn to group together similar input patterns.

Networks can be designed by hand and remain static, or generated by a learning algorithm. One such generative algorithm, called cascade-correlation, grows by recruiting new hidden units that have learned to be active when the network is experiencing a high degree of error. Back-propagation of error is unnecessary here because incoming weights to a recruited hidden unit are frozen when training of the single-layer of output-unit weights resumes. Unit recruitment corresponds roughly to the formation of new neurons and new synapses.

A few developmental models employ auto-associator networks, in which units are fully interconnected (no layers) and learning occurs with either the delta rule or Hebb rule (in which weights are strengthened in proportion to the product of the activations of the sending and receiving units).

2. *Cognitive Stages and Perceptual Effects*

One of the first developmental connectionist models concerned a task that has become a benchmark for modeling developmental stages—the balance beam. Here a child is presented with a rigid beam on which a number of equal-sized weights are placed at certain positions to the left and right of a central fulcrum. The child is asked to predict which way the beam will tip when supporting blocks are removed, or whether the beam will remain balanced. A back-propagation model trained on examples of balance beam problems captured the first three stages observed with children (the use of only weight information, distance information when both sides have equal weight, both weight and distance information without resolving conflicts), but failed to settle into the final stage of mostly correct solutions (McClelland 1989).

A cascade-correlation model progressed through all four stages of balance beam performance and also captured the effect that problems with small weight × distance differences from one side to the other are relatively difficult at each stage (Shultz et al. 1994). Such perceptual effects are pervasive in cognitive tasks, but have not been comprehensively explained. Connectionist models produce these effects naturally in their method of passing activations.

There are several of these perceptual effects in the vast literature on conservation acquisition: the problem size effect (small number problems are easier than large number problems), length bias (nonconservers

choose the longer row as having more), and the screening effect (even nonconservers assume conservation when the transformation is hidden from view). A cascade-correlation model of conservation captured these phenomena, along with natural conservation acquisition and sudden spurts in performance, while showing how perceptual and cognitive information, although initially in conflict, are eventually integrated (Shultz 1998).

A modular cascade-correlation model of seriation (sorting) captured all four stages that children go through (random performance, partial sorts, complete trial-and-error sorts, and systematic sorts) as well as the fact that smaller differences between items make the task more difficult (Mareschal and Shultz 1999).

3. *Transition Mechanisms*

Perhaps the most difficult issue in developmental psychology concerns how children progress from one stage to the next. The longstanding debate on nativism vs. empiricism has been enlightened by connectionist analyses of different ways in which innate influences can be implemented—as neural representations, architectures, or timing of development (Elman et al. 1996). Interactions between learning and inheritance have been modeled by having successful neural networks reproduce over generations, showing that evolutionary progress is enhanced by learning ability (Nolfi et al. 1994). This belies the common assumption that connectionist approaches to development are strictly empiricist.

A constructivist approach to development has been implemented in several simulations using the cascade-correlation algorithm. These simulations show that it is possible to escape from Fodor's paradox of cognitive development—that experiential accounts of development are incoherent because it is computationally impossible to learn anything genuinely new. Generative networks escape from this paradox by increasing their computational power when needed (Mareschal and Shultz 1996).

4. *Non-normative Stages*

A curious feature of development is the presence of early nonnormative stages. Why would children regularly construct incorrect solutions to a problem? In one example, length bias in nonconservation resulted from a correlation between number and length for transformations that preserve density (Shultz 1998).

A cascade-correlation model of the integration of velocity, time, and distance cues for moving objects progressed through the non-normative stages observed with children on their way to the normative integration based on products and ratios of cues (Buckingham and Shultz 2000). For example, children first view velocity as being proportional to distance traveled, then as the difference between distance and time, and finally as the ratio of distance to time. Such stages are a natural consequence of the growth in computational power in cascade-correlation networks. Initially, these networks lack the power to represent the nonlinear relations among velocity, time, and distance and the ability to simultaneously represent the various directions of inferences based on these equations. Static feed-forward networks that adjust their weights, but do not grow, are unable to capture these stages.

5. *Developmental Lags*

Different methods of assessing stages in children sometimes produce different results. One example concerns the age at which infants develop the notion of permanent objects, objects that continue to exist even when they are no longer being experienced. Piaget had found that infants below about 9 months would not retrieve an object from under a screen, but Baillargeon has found that infants as young as 3.5 months look longer at events in which a moving object violates permanency, such as failing to emerge between occluding screens. A modular network model of infants' visual tracking of moving objects learned to predict the emergence of an object from behind an occluding screen (Mareschal et al. 1999). An unsupervised module learned to recognize objects, a recurrent back-propagation module learned to predict object trajectories, and another back-propagation module learned to retrieve objects by integrating information from the other two modules. The developmental lag was explained by noting that reaching requires integration of information about both the location and the nature of the object, whereas prediction requires only information about its location.

6. *Modularity*

With verbal theorizing, it is difficult to know when to claim that the operation of distinct modules is required for cognitive processing. Computational modeling suggests modularity whenever a homogenous processor fails to learn the task or capture the psychological data. Modular systems are known to simplify learning and processing of particularly difficult tasks.

Another example of modularity is a model of word learning using an unsupervised network to learn categories and an auto-associator network to learn category names (Schyns 1991). The output of each module served as input to the other. This model captured several phenomena in children's concept learning: strong recognition of a previously unseen

prototype, semantic overextension, interword competition, faster learning of a label for a distinctive category, and faster learning of a category with a distinctive label.

The seriation model presented earlier provides a third example of modularity, with one network module selecting a stick to move and another selecting a destination. Homogeneous networks did not progress through the proper psychological stages and did not learn to sort successfully.

7. *Self-organization*

Self-organization refers to spontaneous ordering tendencies in complex systems. Both brains and artificial neural networks, because of their complex dynamics, parallelism, and responsiveness to feedback are likely candidates for self-organization. Indeed, neural network models can be used to investigate self-organizing properties of the brain.

An example is a model with only input and output layers, a Hebb learning rule, and competition among output units that simulated the formation of ocular dominance columns in visual cortex (Miller et al. 1989). Groups of neurons in visual cortex eventually organize themselves into columns responding to information coming from one eye or the other.

8. *Integration of Diverse Findings*

Some connectionist models have been able to integrate large areas of psychological literature that have resisted the explanatory efforts of verbal theories. Examples of conservation and seriation were noted earlier. Another example is a cascade-correlation model of shift learning (Sirois and Shultz 1998). In shift learning, subjects learn a discrimination problem and are then shifted to another related problem to determine what was learned on the first problem. The network model integrated diverse developmental findings on a variety of shifts that could not be accounted for by any one of several verbal theories. The model was based on the idea that older children and adults spontaneously overlearn these problems.

9. *Explanation of Mystery Effects*

In smaller, contemporary literatures, there are often results that are difficult to explain. One curious finding concerns asymmetric exclusivity effects in infant habituation to categories of pictures, namely that dogs are not included in the category of cats, but cats are included in the category of dogs. This was captured by a back-propagation encoder network, and subsequently explained by determining how these effects arose within the model (Mareschal and French 1997). Encoder networks learn to reproduce their inputs on their output units using a smaller number of hidden units. A stimulus is recognized as familiar if it can be encoded onto the hidden units and then decoded accurately onto the output units. Habituation to repeated stimuli results in a decrease in network error; recovery of error to a new stimulus is interpreted as an indicator of novelty. The asymmetric exclusivity effects were explained in terms of the distribution of stimulus features. Because some cat-feature values are inside the range of dog-feature values but not the reverse, a system learning these statistics would naturally view cats as a subset of dogs.

10. *Resolution of Theoretical Disputes*

Theoretical disputes are often difficult to resolve at the level of verbal theories. One such recurrent debate concerns whether human cognition is based on symbolic rules and variables, or on sub-symbolic connections. Infant habituation and recovery to sentences in a simple artificial grammar have been covered by variety of standard neural models, without explicit rules and variables that others had argued were required (e.g., Shultz 1999, Sirois et al. 2000). Similar theoretical disputes have been addressed by models of conservation, seriation, and the balance beam presented earlier.

11. *Conclusions*

Connectionist models have simulated large varieties and amounts of developmental data while addressing important and longstanding developmental issues. Connectionist approaches provide a novel view of how knowledge is represented in children and a compelling picture of how and why developmental transitions occur. Like other modeling techniques, connectionism has increased the precision of theorizing and thus clarified theoretical debates.

As these models become more widely known, it is likely that many more of their predictions will be tested with children. It is also likely that connectionist models will be extended to a wider range of developmental phenomena. Although it is not yet clear whether these models will be able to cover phenomena in social development, there is a promising connectionist model of imprinting (O'Reilly and Johnson 1994). Much of the connectionist developmental literature concerns language acquisition, which is covered in another article. Some features lacking in current models will continue to receive attention: explicit rule use, genotypes, multitask learning, impact of knowledge on learning, embodiment, and neurological realism.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Connectionist Approaches; Connectionist Models of Concept Learning; Connectionist Models of Language Processing; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning

Bibliography

Buckingham D, Shultz T R 2000 The developmental course of distance, time, and velocity concepts: A generative connectionist model. *Journal of Cognition and Development* **1**: 305–45

Elman J L, Bates E, Johnson M H, Karmiloff-Smith A, Parisi D, Plunkett K 1996 *Rethinking Innateness: A Connectionist Perspective on Development*. MIT, Cambridge, MA

Mareschal D, French R M 1997 A connectionist account of interference effects in early infant memory and categorization. *Proceedings of the 19th Annual Conference of the Cognitive Science Society*. Erlbaum, Mahwah, NJ, pp. 484–9

Mareschal D, Plunkett K, Harris P 1999 A computational and neuropsychologial account of object-oriented behaviours in infancy. *Developmental Science* **2**: 306–17

Mareschal D, Shultz T R 1996 Generative connectionist networks and constructivist cognitive development. *Cognitive Development* **11**: 571–603

Mareschal D, Shultz T R 1999 Development of children's seriation: A connectionist approach. *Connection Science* **11**: 149–86

McClelland J L 1989 Parallel distributed processing: Implications for cognition and development. In: Morris R G M (ed.) *Parallel Distributed Processing: Implications for Psychology and Neurobiology*. Oxford University Press, Oxford, UK

Miller K D, Keller J B, Stryker M P 1989 Ocular dominance and column development: Analysis and simulation. *Science* **245**: 605–15

Nolfi S, Parisi D, Elman J L 1994 Learning and evolution in neural networks. *Adaptive Behavior* **3**: 5–28

O'Reilly R C, Johnson M H 1994 Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation* **6**: 357–89

Schyns P 1991 A modular neural network model of concept acquisition. *Cognitive Science* **15**: 461–508

Shultz T R 1998 A computational analysis of conservation. *Developmental Science* **1**: 103–26

Shultz T R 1999 Rule learning by habituation can be simulated in neural networks. *Proceedings of the 21st Annual Conference of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ, pp. 665–70

Shultz T R, Mareschal D, Schmidt W C 1994 Modeling cognitive development on balance scale phenomena. *Machine Learning* **16**: 57–86

Sirois S, Buckingham D, Shultz T R 2000 Artificial grammar learning by infants: An auto-associator perspective. *Developmental Science* **4**: 442–56

Sirois S, Shultz T R 1998 Neural network modeling of developmental effects in discrimination shifts. *Journal of Experimental Child Psychology* **71**: 235–74

T. R. Shultz

Connectionist Models of Language Processing

All connectionist models have in common that they consist of subunits that can, in one way or another, be likened to brain structures. The neuroanatomical structures the network elements stand for are either large brain regions and the pathways between them, or local clusters of neurons and their mutual connections, or individual neurons and the web of fibers in which they are embedded. There are at least three major meanings of the term 'Connectionist model': (a) *Classical connectionist models* proposed by nineteenth-century neurologists specify centers and pathways that are analogous to cortical areas and fiber bundles most relevant to language, and related modern *neurobiological approaches to language* propose intensely connected cell assemblies with different cortical distributions as the brain basis of language; (b) *Symbolic connectionist models* suggest single artificial neurons corresponding to linguistic units (language sounds, words, etc.); and (c) *Distributed connectionist models* represent such linguistic entities by activity vectors involving numerous neuronal elements. This article will explain these three different notions by introducing the respective research areas (Sects. 1–3). In the final Sect. (4), recent trends in the research on connectionist models of language will be highlighted.

1. Neurological and Neurobiological Models of Language

In the second half of the nineteenth century, connectionist theories were proposed by neurologists to summarize and model the effect of brain lesions on cognitive functions (Caplan 1987). The underlying idea was that local processing systems, so-called centers, specialize in particular cognitive processes. These were thought to be autonomous processors. The centers were linked through so-called pathways which allowed for information exchange between them. The pathways had their analogue in fiber bundles in the white matter of the brain. The most famous classical neurological model of language goes back to Wernicke (1874) and Lichtheim (1885). This model explains basic features of organic language disturbances, *aphasias*. The main proposal of the Wernicke–Lichtheim model was that there are two centers for language processing, one mainly involved in speech production and the other primarily contributing to language comprehension. This hypothesis could be confirmed by a wealth of clinical studies. It is now well-established that, in most right-handed subjects, the main or core language areas are located in the left inferior frontal lobe (Broca area for speech production) and in the left superior temporal lobe (Wernicke area for speech comprehension).

In the light of modern neuropsychological and neuroimaging research, early connectionist models turned out to be too crude to explain many patterns of brain activation induced by language processing, and the more fine-grained aspects of language disorders caused by brain lesions. It now appears more likely that the 'language centers' of Broca and Wernicke are not functionally independent but that they are mutually dependent when functioning properly. Furthermore, while these core language areas are certainly important for language, they are certainly not the only cortical areas contributing to and necessary for language processing. This is demonstrated by neuroimaging research (EEG, fMRI, MEG, PET) showing that, in addition to the core language areas, other areas light up when specific language stimuli are being processed, and by converging neuropsychological reports about patients with lesions outside their core language areas who showed category-specific linguistic deficits. These studies show that, in addition to the core language areas, there are additional or complementary language areas that are activated during language processing and whose lesion causes deterioration of aspects of language processing.

Mutual functional dependence of the core language areas and the category-specific role of complementary language areas are explained by a neurobiological model postulating that words and other language elements are cortically organized as strongly connected assemblies of neurons whose cortical distributions vary with word type (Pulvermüller 1999). Accordingly, concrete words referring to objects and actions are organized as widely distributed cell assemblies comprising neurons in sensory and motor areas involved in processing the words' meanings. In contrast, highly abstract grammatical function words and grammatical affixes are proposed to be more focally represented in the left-hemispheric core language areas of Broca and Wernicke (Pulvermüller 1999). In summary, neuron sets in core language areas appear to be relevant for all types of language-related processes, and complementary neurons in areas related to actions and perceptions regularly involved in language use may contribute to category-specific language processes.

2. *Symbolic Connectionist Models of Language*

According to modular approaches to cognitive psychology, the mental language processor consists of quasi-autonomous subprocessors or modules, and language processing therefore is being considered the result of quasi-autonomous subprocesses. The subprocesses envisaged to be involved in language comprehension are, for example, input-feature analysis, letter or phoneme analysis, word form processing, and semantic analysis (Morton 1967). These processes are assumed to occur sequentially or in a cascaded manner. A similar but reverse cascade has been assumed for the putative subprocesses of language production which finally results in movements of the articulators or the (writing) hand (Garrett 1984).

In symbolic connectionist models, the subprocessors of modular models have been replaced by layers of neuron-like elements, the assumption being that individual artificial neurons represent acoustic or visual features, phonemes or graphemes, word forms, and word meanings (see, for example, McClellend and Rumelhart 1981). Symbolic connectionist networks have been applied with some success to the modeling of word production (Levelt et al. 1999) and word recognition (Norris et al. 2000), as well as other language phenomena, for example speech errors observed in normal speakers and language-impaired neurological patients (see, for example, Dell 1986).

A still open issue is whether information should only be allowed to flow in one direction between the subcomponents of the postulated networks, from input to semantics in comprehension and from semantics to output in production. As an alternative, reciprocal connections and thus bidirectional information flow has been proposed to take place between the layers of the networks.

3. *Distributed Connectionist Models of Language*

The distributed network type most commonly used to model language resembles symbolic networks, because both network types are made up of layers of neurons and connections between layers. An important difference is the following: Symbolic networks include local representations—usually single artificial neurons—that represent elementary features of the input and output as well as more complex entities such as, letters, phonemes, and words. In contrast, distributed networks use activity vectors over large sets of neurons in a given layer to represent words and other more complex linguistic structures. The distributed networks most commonly used in language simulations do not include direct connections between the neurons of one layer. The active neurons defined by an activity vector are therefore not connected to each other and do not form a functionally coherent system. This distinguishes them from cell assemblies with strong internal links. However, as a consequence of associative learning, artificial neurons in different layers may strengthen their connections to neurons in other layers indirectly linking the active neurons defined by individual activity vectors.

The classical type of distributed network, the perceptron, consists of two layers of neurons. Each neuron in layer one is connected to each neuron in the second layer, though connection strengths (or weights) vary as a function of associative learning of pairs of input–output patterns. In the 1960s, Minski and Papert (1969) proved that perceptrons can learn to solve only a certain type of classification problem

called linearly separable. In the 1980s, Rumelhart and McClelland (1986a) showed that the addition of one 'hidden' layer of neural units between the input and output layers, and extension of the learning rule, allows for overcoming this limitation. The networks were now able to learn to solve more complex classification problems and this led to a strong interest in three-layer neural architectures. Three-layer perceptrons were used with some success to model language: There are models that classify speech signals (e.g., Waibel et al. 1995); others that mimic important aspects of the infant's learning of language specific information as described by elementary rules (e.g., Hare et al. 1995); simulations of the effects of focal brain lesions on language functions (e.g. Plaut and Shallice 1993) and of the recovery of language functions after stroke.

For solving more complex problems posed by syntax, the network architecture had to be modified once again. Three-layer architectures do not allow for keeping information in memory for a longer time span, but this is necessary for assessing syntactic dependencies between temporally distant language units as, for example, between the first and last words of a long sentence. To solve this problem, a 'memory layer' allowing for reverberation of activity and information storage has been added to the three-layer-architecture (Elman 1990). Such networks including an additional memory layer can be shown to be more powerful than three-layer perceptrons in storing serial order relationships. They are capable of learning subsets of syntactically complex sentence structures, for example aspects of so-called center-embedded constructions.

4. *Current Trends*

After introducing three types of connectionist models, the article will now highlight selected topics in connectionist research, where the three approaches offer somewhat different views and where the divergence in views has actually led to productive research.

Recent trends in connectionist research on language include the more detailed modeling of syntactic mechanisms and attempts at mimicking more and more properties of the actual neuronal substrate in the artificial models (Elman et al. 1996). Multidisciplinary research across the computational and neurosciences is necessary here. The strategy to copy the brain's mechanisms into the artificial neural network may be particularly fruitful for implementing those higher cognitive functions that, if implemented in the biological world, only arise from specific brain types. The brain's structure is information that may be of relevance for neuronal modeling.

The modeling of rule-like verbal behavior is an illustrative example for successful multidisciplinary interaction in connectionist research on language. It is sometimes assumed that symbolic algorithms are necessary for explaining the behavior described by linguistic rules. For producing a past tense form of English, one would, accordingly, use an abstract rule such as the following addition rule scheme:

$$\text{Present stem} + \text{Past suffix} = \text{Past tense form}$$

In particular, an algorithm of this kind could model the concatenation of the verb stem 'link' and the past suffix 'ed' to yield the past tense form 'linked,' and, in general, it could be used to derive any other regular past form of English. However, it is difficult to see how an irregular verb such as 'think' or 'shrink' could yield a past form based on a similar rule. In the extreme, one would need to assume rules for individual words to provide algorithms that generate, for example, 'went' out of 'go.' This would require stretching the rule concept, and linguists have therefore proposed that there are two distinct cognitive systems contributing to language processing, a symbolic system storing and applying rules and a second system storing relationships between irregular stems and past forms in an associative manner (Pinker 1997).

From the perspective of neural networks, however, one may ask whether two separate systems, for rules and exceptions, are actually necessary to handle regular and irregular inflection. Rumelhart and McClelland (1986b) showed that an elementary two-layer perceptron can store and retrieve important aspects of both past tense rules and exceptions. It can even produce errors typical for children who learn past tense formation, such as so-called overgeneralizations (e.g., 'go*ed*' instead of 'went.')

From a linguistic perspective, the two-layer model of past tense proposed by Rumelhart and McClelland has been criticized, for example because it does not appropriately model the fact that rule-conforming behavior is by far most likely to be generalized to novel forms. The past form of a newly introduced verb, such as 'dif,' will thus almost certainly receive an 'ed' ending if one intends to use it in the past tense ('diff*ed*.') This is even so in languages where most verbs have irregular past forms and only a minority of the verbs conform to the rule. The rule is nevertheless used as the default and generalized to novel forms and even rare irregular items. This is a problem for a subset of connectionist models, because the strongest driving forces in associative networks are the most common patterns in the input.

However, there are distributed three-layer networks that solved the problem of default generalization surprisingly well (Hare et al. 1995). An important determinant is that rule-conforming input patterns are maximally *dissimilar*, while the members of an irregular class resemble each other. Consider the different regular forms *to watch*, *talk*, and *jump* in contrast to the similar members of an irregular class

to sing, *ring*, and *sting*. Because the regulars are so heterogeneous, they occupy a wide area in input space. The representation in input space of a novel word is thus most likely to be closest to those of *one of the many different* regular forms, and this is one important reason why so many new items are treated as regular by the network. On the other hand, if a newly introduced item happens to strongly resemble many members of a regular class, for example the pseudoword *pling*, it is, in many cases, treated as regular. These observations may lead one to redefine one's concept of regularity: A rule is not necessarily the pattern most frequently applied to existing forms, but it is always the pattern applied to the *most heterogeneous set* of linguistic entities. The heterogeneity of the regular classes may explain default generalization along with the great productivity of rules.

The simulation studies of the acquisition of past tense and other inflection types by young infants suggest that neural networks consisting of one single system of layers of artificial neurons provide a reasonable model of the underlying cognitive and brain processes. In this realm, the single system perspective appears equally powerful as an approach favoring two systems, one specializing in rule storage and the other in elementary associative patterns.

Neuroscientific data and theories have recently shed new light on the issue of a single-system versus a double-system account of rule-like behavior. Important was the discovery of patients with brain lesions who were differentially impaired in processing regular and irregular past tense forms. Patients suffering from Parkinson's disease or Broca's aphasia were found to have more difficulty processing regulars, whereas patients with global deterioration of cortical functions as seen, for example, in Alzheimer's Disease or Semantic Dementia showed impaired processing of irregulars (Ullman et al. 1997; Marslen-Wilson & Tyler, 1997). This double dissociation is difficult to model using a single system of connected layers, but is easy to handle if different neural systems are used to model regular and irregular inflection.

Another argument in favor of a double system account comes from neurobiological approaches proposing that words and inflectional affixes are represented in the cortex as distributed cell assemblies. In this case, past tense formation can involve two types of connections, local within-area connections in the core language areas and long-distance links between the language areas and outside. It is known from neuroanatomy that two adjacent neurons are more likely to be linked through a local connection than are two distant neurons to be linked by way of a long-distance connection. This situation can be modeled by two pathways connecting the neuronal counterparts of present stems and past forms, for example a three-layer architecture with two pathways connecting input and output layers, one with higher and the other with lower connection probabilities between neurons in adjacent layers. Parameters are chosen appropriately, the two pathways or systems will differentially specialize in the storage of rules and irregular patterns. Similar to a two-layer perceptron, the low-probability system is best at storing the simple mapping between irregular present forms that resemble each other and their past forms. In contrast, the complex mapping between the heterogeneous regular stems and their past forms is best accomplished by the three-layer component with high connection probabilities. When the two components are differentially lesioned, the network produces the double dissociation between regular and irregular inflection seen in neuropsychological patients. This approach explains the neuropsychological double dissociation along with aspects of the acquisition of past tense formation by young infants (Pulvermüller 1998). This explanation is based on principles of cortical connectivity.

Together, the neuropsychological double dissociation and the neurobiological consideration argue in favor of a two-system model of regular and irregular inflection. In contrast to the modular proposal that each of two systems are exclusively concerned with regular and irregular processes, respectively, the neuroscientific variant would suggest a gradual specialization caused by differential connection probabilities. The ongoing debate between cognitive neuroscientists favoring single- or double-system accounts of rule-like knowledge clearly proves the importance of multidisciplinary interaction between the linguistic, cognitive, computational, and neurosciences.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Computational Psycholinguistics; Connectionist Approaches; Language Development, Neural Basis of; Neural Networks: Biological Models and Applications; Speech Perception; Speech Production, Neural Basis of; Syntactic Aspects of Language, Neural Basis of

Bibliography

Caplan D 1987 *Neurolinguistics and Linguistic Aphasiology. An Introduction*. Cambridge University Press, Cambridge, MA

Dell G S 1986 A spreading-activation theory of retrieval in sentence production. *Psychological Review* **93**: 283–321

Elman J L 1990 Finding structure in time. *Cognitive Science* **14**: 179–211

Elman J L, Bates L, Johnson M, Karmiloff-Smith A, Parisi D, Plunkett K 1996 *Rethinking Innateness. A Connectionist Perspective on Development*. MIT Press, Cambridge, MA

Garrett M 1984 The organization of processing structures for language production. In: Caplan D, Lecours R, Smith A (eds.) *Biological Perspectives On Language*. MIT Press, Cambridge, MA, pp. 172–93

Hare M, Elman J L, Daugherty K G 1995 Default generalisation in connectionist networks. *Language and Cognitive Processes* **10**: 601–30

Levelt W J M, Roelofs A, Meyer A 1999 A theory of lexical access in speech production. *Behavioral and Brain Sciences* **22**: 1–75

Lichtheim L 1885 Über Aphasie. *Deutsches Archiv für Klinische Medicin* **36**: 204–68
Marslen-Wilson W, Tyler L 1997 Dissociating types of mental computation. *Nature* **387**: 592–4
McClelland J L, Rumelhart D E 1981 An interactive activation model of context effects in letter perception: part 1. *Psychological Review* **88**: 375–407
Minsky M, Papert S 1969 *Perceptrons*. MIT Press, Cambridge, MA
Morton J 1969 The interaction of information in word recognition. *Psychological Review* **76**: 165–78
Norris D J M, McQueen J M, Cutler A 2000 Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences* **23**: 299–370
Pinker S 1997 Words and rules in the human brain. *Nature* **387**: 547–8
Plaut D C, Shallice T 1993 Deep dyslexia: a case study of connectionist neuropsychology. *Cognitive Neuropsychology* **10**: 377–500
Pulvermüller F 1998 On the matter of rules. *Network: Computation in Neural Systems* **9**: R 1–51
Pulvermüller F 1999 Words in the brain's language. *Behavavioral and Brain Sciences* **22**: 253–336
Rumelhart D E, McClelland J L (eds.) 1986a *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA
Rumelhart D E, McClelland J L 1986b On learning the past tense of English verbs. In: McClelland J L, Rumelhart D E (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA
Ullman M, Corkin S, Coppola M, Hickok G, Growdon J, Koroshetz W, Pinker S 1997 A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience* **9**: 266–76
Waibel A, Hanazawa T, Hinton G, Shikano K, Lang K J 1995 Phoneme recognition using time-delay neural networks. In: Chauvin Y, Rumelhart D E (eds.) *Backpropagation: Theory, Architectures, and Applications. Developments in Connectionist Theory*. Lawrence Erlbaum, Hillsdale NJ
Wernicke C 1874 *Der aphasische Symptomencomplex. Eine psychologische Studie auf anatomischer Basis*. Kohn und Weigert, Breslau, Poland

F. Pulvermüller

Conscious and Unconscious Processes in Cognition

Everyday experience suggests that we often seem to know more than we can tell. Riding a bicycle, playing tennis, or driving a car, for instance, all involve mastering complex sets of motor skills, yet we are at a loss when it comes to explaining exactly how we perform such physical feats. These dissociations between our ability to report on cognitive processes and the behaviors that involve these processes are not limited to action but extend to higher-level cognition as well. Most native speakers of a language are unable to articulate the grammatical rules they nevertheless follow when uttering expressions of the language. Likewise, expertise in domains such as medical diagnosis or chess, as well as social or aesthetic judgments, all involve intuitive knowledge to which one seems to have little introspective access.

We also often seem to tell more than we can know. In a classic article, social psychologists Nisbett and Wilson (1977) reported on many experimental demonstrations of the fact that accounts of our own behavior frequently reflect reconstructive and interpretative processes rather than genuine introspection. In one such study, Nisbett and Wilson asked patrons of a department store to choose the best-quality item among four identical pairs of nylon stockings in the context of what subjects believed was a consumer survey. There was a strong position effect: items located on the right, which were inspected last, were chosen much more often than items in other positions. Yet, when asked what motivated their choice, not a single subject mentioned position as a relevant factor, and many denied that it might have had an effect when asked directly about it. More generally, while it is usually agreed that cognitive *processes* are not in and of themselves open to any sort of introspection, Nisbett and Wilson further claimed that we can sometimes be '(a) unaware of the existence of a stimulus that importantly influenced a response, (b) unaware of the existence of the response, and (c) unaware that the stimulus has affected the response' (Nisbett and Wilson 1977, p. 231).

Just as for knowledge of our own cognitions, our knowledge of the world turns out to be much less complete that it might appear, thus further deconstructing Cartesian conceptions of cognition in which conscious awareness reigns supreme. In change blindness and inattentional blindness experiments, for instance, people are surprisingly often unable to detect large changes in visual stimuli that they are allowed to observe at their leisure. One such experiment (e.g., Simons and Chabris 1999) involved watching a movie depicting two teams of three basketball players and counting the number of times that a ball was passed between the players of each team. Under these conditions, that is with attention fully engaged in tracking one of the two balls, people often failed to notice that, halfway through the movie, an actor dressed as a gorilla appeared on the right of the screen and walked slowly towards the middle of the scene, where he repeatedly banged his chest, subsequently resuming his stroll to disappear on the left!

Demonstrations of dissociations between conscious awareness and stimulation and/or behavior have now been reported in many domains of psychology. Frith et al. (1999) have suggested organizing paradigms through which to study the 'neural correlates of consciousness' in nine groups resulting from crossing two dimensions: (a) three classes of psychological

processes involving, respectively, knowledge of the past, present, and future—memory, perception, and action, and (b) three types of cases where subjective experience is incongruent with the objective situation—cases where subjective experience fails to reflect changes in either (i) the stimulation or in (ii) behavior, and (iii) cases where subjective experience changes while stimulation and behavior remain constant.

Among such paradigms, the study of dissociations between conscious and unconscious perceptual processes has by far the longest history, extending back to the beginnings of scientific psychology in the nineteenth century. Subliminal perception (see Dixon 1971) is said to occur when stimuli presented below the threshold of awareness influence information processing, emotion, or action. In a classic but highly controversial series of studies, Anthony Marcel suggested that semantic priming, that is, the facilitation of decisions about semantic aspects of a target stimulus when it follows a related stimulus (the prime), can occur in conditions when the prime is presented so rapidly that subjects fail to be able to decide whether it was present or not (e.g., Marcel 1983). While many subsequent studies have reported similar findings with different stimuli such as spoken words or faces, many have also failed to replicate such findings, and the domain remains highly controversial (Holender 1986). Nevertheless, it appears that considerable information can be perceived and processed in the absence of conscious awareness when it is assessed through subjective measures: people can believe that they are guessing, yet perform better than chance would predict in a variety of tasks involving semantic and affective decisions.

Some forms of learning and memory can also be unconscious. Memory for previous events can be expressed explicitly, as a conscious recollection, or implicitly, as automatic, unconscious influences on behavior. Numerous studies have demonstrated dissociations between implicit and explicit memory. Amnesic patients, for instance, who exhibit severe or total loss in their ability explicitly to recall previous experiences, nevertheless retain the ability to learn novel procedural skills (Milner 1962) or to exhibit sensitivity to past experiences of which they are not conscious (e.g., Warrington and Weiskrantz 1968). Some studies have explored the extent to which patients undergoing general anesthesia are able to learn novel associations, and one study has even demonstrated learning in fetuses (Van Heteren et al. 2000).

Similar dissociations can be observed with normal subjects (see Cleeremans et al. 1998). Arthur Reber, in a classic series of studies conducted in 1967, first suggested that learning might be 'implicit,' to the extent that people appear to be able to learn new information without intending to do so, and in such a way that the resulting knowledge is difficult to express. In Reber's seminal study of artificial grammar learning, subjects were asked to memorize a set of meaningless letter strings generated by a simple set of rules embodied in a finite-state grammar. After this memorization phase, they were told that the strings followed the rules of a grammar, and were asked to classify novel strings as grammatical or not. In this experiment and in many subsequent replications, subjects were able to perform this classification task better than chance would predict, despite remaining unable to describe the rules of the grammar in verbal reports. This dissociation between classification performance and verbal report is the finding that prompted Reber to describe learning as implicit, for subjects appeared sensitive to and could apply knowledge (the rules of the grammar) that they remained unable to describe and had no intention of learning.

In a series of experiments that attracted renewed interest in implicit learning, Berry and Broadbent (1984) showed that success in learning how to control a simulated system (e.g., in one set of experiments, a system described to subjects as a 'sugar factory') so as to make it reach certain goal states was independent from the ability to answer questions about the principles governing subjects' inputs and the system's output: practice selectively influenced ability to control the system, whereas verbal explanations about how the system works selectively influenced ability to answer questions.

At the turn of the twenty-first century a third paradigm—sequence learning—has become dominant in the investigation of implicit learning. Nissen and Bullemer (1987) first demonstrated that subjects asked to respond as fast and as accurately as possible to a series of visual events learned progressively about the sequential structure of the stimulus sequence in spite of showing little evidence of being aware that the material contained structure. Numerous subsequent studies of this effect have indicated that subjects can learn about complex sequential relationships despite remaining unable fully to elicit this knowledge in corresponding direct, explicit tasks (e.g., Cleeremans and McClelland 1991).

One of the most convincing demonstrations of dissociations between conscious knowledge and behavior was obtained in an eye-blink conditioning situation (Perruchet 1985). In this experiment, people were exposed to a series of identical tones, 50 percent of which were followed after a short interval by an air-puff directed at the left cornea. Immediately after each tone was presented (and before the puff occurred in reinforced trials), people were asked to indicate the extent to which they expected the tone to be followed by an air-puff on a 0–7 points scale. A trial-by-trial analysis of the results indicated that eye-blink responses were increasingly more likely to occur after presentation of a tone, if the corresponding trial had been preceded by a series of reinforced trials (i.e., trials during which the tone had indeed been followed by an air-puff). In stark contrast, however, people's

subjective expectancy of the occurrence of an air-puff tended to *decrease* with the number of reinforced trials that preceded the trial under consideration. In other words, people's eye-blink responses were completely dissociated from their conscious expectations about when each tone would be followed by an air-puff.

Finally, the neuropsychological literature also contains many striking cases of dissociation between subjective experience and behavior. In addition to the dissociations exhibited by patients with amnesia, other syndromes have also attracted attention over the last decades. Blindsight patients, for instance, suffer from damage to the primary visual cortex, and report being unable to see in some portions of their visual field. Larry Weiskrantz (1986) showed that such patients are nevertheless often able to act appropriately when stimulated in their blind field, for instance by initiating correct reaching or grasping movements toward a target. Likewise, the study of patients with form agnosia or neglect has offered valuable insight into the neural correlates of consciousness. For instance, Milner and Goodale (e.g., 1995) have described a patient, D.F., who, while unable to offer verbal or manual reports on the orientation of a slot, nevertheless retains the ability actually to insert a card in it, accurately preparing her movement well before the slot is reached. These observations led Milner and Goodale to suggest that distinct 'what' and 'where' processing pathways exist in the human visual system, and that only one of these pathways (the ventral 'what' system) is associated with conscious visual experience.

1. *Consciousness and Theories of the Psychological Unconscious*

These different findings all suggest that unconscious influences on behavior are pervasive. This raises the question of how best to characterize the relationships between conscious and unconscious processes, and in particular whether one should consider that mental representations can be unconscious. The idea that mental life includes both conscious and unconscious events has been expressed most clearly by Freud (e.g., 1949), whose psychoanalytical theory has influenced profoundly both scientific thought and public conceptions of the mind. Even though one can find earlier characterizations of cognition in which computations that are not accompanied by awareness contribute to conscious decisions (e.g., Leibniz 1981, De Biran 1929), most thinkers up until Freud, including René Descartes and, arguably, William James, considered that mental life consists exclusively of conscious events. This position is still endorsed by some authors today, even though there is wide consensus about the idea that cognitive processing is largely unconscious.

Freud's characterization of the 'dynamic unconscious,' however, makes very specific assumptions—specifically that there exist unconscious mental representations, and that these representations can reflect semantic and affective dimensions of processing. Further, the unconscious, as Freud depicted it, is as dynamic and causally efficacious as the conscious—it is a mental bubbling cauldron of sorts, replete with repressed instincts, thoughts, and feelings vying to get access to awareness. Few thinkers would endorse Freud's characterization of the unconscious in full today—often choosing instead to deny its existence altogether, or considering that only shallow aspects of processing can take place unconsciously. In particular, while popular belief has often tended to ascribe powerful abilities to the unconscious, empirical exploration of the level of analysis at which processing occurs in the absence of consciousness has failed to offer convincing demonstrations that unconscious processing can be as flexible or as deep as conscious processing. Rather, the evidence suggests that unconscious processing can bias conscious processing in a way that reflects strong or habitual responses.

An important issue in this context is to determine the extent to which conscious and unconscious processing depend on separable neural systems, and what their relationship should be. Contemporary theories of consciousness make widely different assumptions about its underlying mechanisms. Atkinson et al. (2000) proposed to organize such theories along two dimensions defined by (a) whether the theory assumes that consciousness depends on the involvement of certain *processes* or rather on the properties of certain mental *representations*, and by (b) whether the theory assumes that consciousness depends on the involvement of *specific modules* or whether it *can occur anywhere in the brain*. Computational models of information processing play a significant role in fostering the development of novel theories of consciousness. In neural-network—or connectionist—models, for instance, task-relevant knowledge is embedded in the same structures that support processing itself. In contrast to traditional information-processing frameworks, knowledge is thus implicit in such models to the extent that it cannot be separated from the mechanisms that subserve processing.

If consciousness does not depend on neural systems specifically dedicated to subserve subjective experience, but should instead be viewed as an emerging property of the processing conducted by many different regions of the brain, one might expect to find indications that consciousness fails to be a unified, all-or-none phenomenon. Marcel (1993) described a relevant series of experiments probing the unity of consciousness in normal subjects. In these experiments people were asked to detect changes in the luminance of very faint stimuli in three different ways: by blinking, by pressing on a button, or by a verbal response ('yes'). When these responses had to be produced simultaneously, they often tended to be dissociated, with subjects answering positively, for

instance, through a blinking response, but negatively through the other modalities. Another experiment indicated that the probability of correct guessing varied significantly across modalities, thus suggesting that they each afforded varying degrees of availability to conscious awareness.

2. *Methodological Challenges*

Because there is no accepted operational definition of what it means for an agent to be conscious of something, complex measurement challenges arise in the study of the relationships between conscious and unconscious cognition.

First, consciousness is not a single process or phenomenon, but rather encompasses many dimensions of experience. A first important challenge thus arises in delineating which aspects of consciousness count when assessing whether a subject is aware or not of a particular piece of information: awareness of the presence or absence of a stimulus, conscious memory of a specific previous processing episode, awareness of one's intention to use some information, or awareness that one's behavior is influenced by some previous processing episode. Different aspects of conscious processing are engaged by different paradigms. In subliminal perception studies, for instance, one is concerned with determining whether stimuli that have not been consciously *encoded* can influence subsequent responses. In contrast, implicit memory research has been more focused on *retrieval* processes, that is, on the unintentional, automatic effects that previously consciously perceived stimuli can exert on subsequent decisions. In studies of implicit learning, it is the *relationships* between ensembles of consciously processed stimuli that remain purportedly unconscious. These subtle differences in which specific aspects of the situation are available to awareness illustrate the need to distinguish carefully between awareness during encoding and awareness during retrieval of information. Further, both encoding and retrieval can concern either individual stimuli or relationships between sets of stimuli, and both can either be intentional or not.

A second important challenge is to devise an appropriate measure of awareness. Most experimental paradigms dedicated to exploring the relationships between conscious and unconscious processing have relied on a simple dissociation logic aimed at comparing the sensitivity of two different measures to some relevant information: a measure *C* of subjects' awareness of the information, and a measure *P* of behavioral sensitivity to the same information in the context of some task. Unconscious processing, according to the simple dissociation logic, is then demonstrated whenever *P* exhibits sensitivity to some information in the absence of correlated sensitivity in *C*. There are several potential pitfalls with the simple dissociation logic, however. First, the measures *C* and *P* cannot typically be obtained concurrently. This 'retrospective assessment' problem entails that finding that *C* fails to be sensitive to the relevant information need not necessarily imply that information was processed unconsciously during encoding, but that, for instance, it might have been forgotten before retrieval. A second issue is to ensure that the information revealed through *C* is indeed relevant to performing the task. As Shanks and St. John (1994) have suggested, many studies of implicit learning have failed to respect this 'information' criterion. For instance, successful classification in an artificial grammar learning task need not necessarily be based on knowledge of the rules of the grammar, but can instead involve knowledge of the similarity relationships between training and test items. Subjects asked about the rules of the grammar would then understandably fail to offer relevant explicit knowledge.

A third issue is to ensure that *C* and *P* are both equally sensitive to the relevant information. At first sight, verbal reports and other *subjective measures* such as confidence ratings would appear to offer the most direct way of assessing the contents of subjective experience, but such measures are often difficult to operationalize in a sufficiently controlled manner. For instance, people might simply refrain from reporting on knowledge held with low confidence, or might offer reports that are essentially reconstructive in nature, as Nisbett and Wilson's experiments indicate. For this reason, many authors have advocated using so-called objective measures of awareness. *Objective measures* of awareness include forced-choice tests such as recognition, presence–absence decisions, or identification.

Even if these different conditions are fulfilled, however, it might be elusive to hope to obtain measures of awareness that are simultaneously *exclusive* and *exhaustive* with respect to knowledge held consciously. In other words, finding null sensitivity in *C*, as required by the dissociation paradigms for unconscious processing to be demonstrated, might simply be impossible because no such absolute measure exists. A significant implication of this conclusion is that, at least with normal participants, it makes little sense to assume that conditions exist in which awareness can simply be 'turned off.' Much of the ongoing debate about the existence of subliminal perception can be attributed to a failure to recognize the limitations of the dissociation logic.

It might therefore instead be more plausible to assume that any task is always sensitive to both conscious and unconscious influences. In other words, no task is *process-pure*. Two methodological approaches that attempt specifically to overcome the conceptual limitations of the dissociation logic have been developed. The first was introduced by Reingold and Merikle (1988), who suggested that the search for absolute measures of awareness should simply be abandoned in favor of approaches that seek to

compare the sensitivity of *direct* measures and *indirect* measures of some discrimination. Direct measures involve tasks in which the instructions make explicit reference to the relevant discrimination, and include objective measures such as recognition or recall. In contrast, indirect measures, such as stem completion in memory tasks, make no reference to the relevant discrimination. By assumption, direct measures should exhibit greater sensitivity than, or equal sensitivity to, indirect measures to consciously held task-relevant information, for subjects would be expected to be more successful in using conscious information when instructed to do so than when not. Hence, demonstrating that an indirect task is more sensitive to some information than a comparable direct task can only be interpreted as indicating unconscious influences on performance.

The second approach—Larry Jacoby's 'Process Dissociation Procedure' (Jacoby 1991)—constitutes one of the most significant advances in the study of differences between implicit and explicit memory. It is based on the argument that, just as direct measures can be contaminated by unconscious influences, indirect measures can likewise be contaminated by conscious influences: particular tasks can simply not be identified with particular underlying processes. The process dissociation procedure thus aims to tease apart the relative contributions of conscious and unconscious influences on performance. To do so, two conditions are compared in which conscious and unconscious influences either both contribute to performance improvement, or act against each other. For instance, subjects might be asked to memorize a list of words and then, after some delay, to perform a stem completion task in which word stems are to be completed either so as to form one of the words memorized earlier (the *inclusion* condition) or so as to form a different word (the *exclusion* condition). If the stems nevertheless tend to be completed by memorized words under exclusion instructions, then one can only conclude that memory for these words was implicit, since if subjects had been able consciously to recollect them, they would have avoided using them to complete the stems. Numerous experiments have now been designed using the process dissociation procedure. Collectively, they offer convincing evidence that performance can be influenced by unconscious information in the absence of conscious, subjective awareness.

3. *Treating Consciousness as a Variable: Qualitative Differences*

Rather than pursuing the elusive goal of demonstrating unconscious cognition, a more fruitful approach to exploring the relationships between conscious and unconscious cognition thus consists of treating consciousness as a variable, that is, of exploring the functional, neural, and computational differences that exist between tasks performed with awareness and without. This leads to the design of experiments in which qualitative differences between cognition with and without awareness can be established in carefully controlled conditions. In this respect, the 'contrastive approach' advocated by Baars (1988) and Frith et al. (1999) offers great promise in helping to solve the methodological challenges described above. Numerous qualitative differences between conscious and unconscious processes have been reported. For instance, based on behavioral methods, Arthur Reber and others have suggested that implicit cognition involves processes that are more resistant to insult or injury, that tend to be less sensitive to individual differences or affective state, and less sensitive to attentional demands. Recent advances in brain-imaging methods make it possible to go beyond purely behavioral methods and to integrate different *online* measures of awareness and of brain activity when attempting to contrast conscious and unconscious cognition. One of the most interesting paradigms through which to combine so-called first-person (subjective) and third-person (objective) methodologies is binocular rivalry—a phenomenon that takes place when different images are shown to each eye. Under these conditions, one's perception of the stimulus alternates spontaneously between the complete perception of one stimulus and the complete perception of the other. Using functional brain imaging, one can then correlate people's subjective reports of what it is they actually perceive at each point in the experiment with the cerebral regions that are most activated, and in so doing establish which of these regions correlate most with the unchanging stimulus and which correlate most with perception as assessed by subjective reports.

Numerous studies combining various online measures of brain activity have likewise attempted to identify which regions of the brain are most involved in processes such as conscious recollection, intentional retrieval, or nonconscious influences in different memory tasks (see Shacter et al. 1998 for a review). While it is too early to draw definite conclusions from such studies, it appears, for instance, that anterior prefrontal regions of the brain appear to be most clearly associated with processes of effortful, conscious retrieval. This and similar findings should help us further to develop theories of consciousness that have a firm rooting in our knowledge of the brain.

4. *Conclusions*

The study of differences between conscious and unconscious processing is a major endeavor for the cognitive neurosciences. Conscious and unconscious processing differ on several dimensions, including depth and specificity of processing. Action can be initiated without consciousness, and some aspects of

perception or memory can also take place without consciousness. Such unconscious processing always tends to reflect habitual or strong responses. From this perspective, unconscious processing is best characterized as the indirect effects of conscious processing. The extent to which conscious and unconscious processes involve distinct brain regions remains an open issue, but it is clear that some regions of the brain tend to be more associated with conscious processing than others. The distinction between conscious and unconscious processing thus appears to involve both a graded and a dynamic continuum (in terms of the underlying mechanisms) and a dichotomy (in terms of subjective experience). Functional brain-imaging methods, when used together with sufficiently sensitive behavioral methods, offer the promise of combining first-person and third-person perspectives to elucidate the relationships between conscious and unconscious processes in cognition.

See also: Cognitive Neuroscience; Cognitive Psychology: History; Cognitive Psychology: Overview; Consciousness and Sensation: Philosophical Aspects; Consciousness, Cognitive Psychology of; Consciousness, Neural Basis of; Evolution of Cognition: An Adaptationist Perspective; Implicit Learning and Memory: Psychological and Neural Aspects; Implicit Memory, Cognitive Psychology of; Mind–Body Dualism; Psychoanalysis: Overview; Theory of Mind

Bibliography

Atkinson A P, Thomas M S C, Cleeremans A 2000 Consciousness: Mapping the theoretical landscape. *Trends in Cognitive Sciences* **4**: 372–82

Baars B J 1988 *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge, UK

Berry D C, Broadbent D E 1984 On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology* **36**: 209–31

Cleeremans A, Destrebecqz A, Boyer M 1998 Implicit learning: News from the front. *Trends in Cognitive Sciences* **2**: 406–16

Cleeremans A, McClelland J L 1991 Learning the structure of event sequences. *Journal of Experimental Psychology: General* **120**: 235–53

de Biran M 1929 *The Influence of Habit on the Faculty of Thinking* [trans. Boehm M D]. Williams and Wilkins, Baltimore

Dixon N F 1971 *Subliminal Perception: The Nature of a Controversy*. McGraw-Hill, London

Freud S 1949 *An Outline of Psychoanalysis* [trans. Strachey J]. Hogarth Press, London

Frith C, Perry R, Lumer E 1999 The neural correlates of conscious experience: An experimental framework. *Trends in Cognitive Sciences* **3**: 105–14

Holender D 1986 Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences* **9**: 1–66

Jacoby L L 1991 A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language* **30**: 513–41

Leibniz G W 1981 *New Essays on Human Understanding* [ed. and trans. Remnant P, Bennet J]. Cambridge University Press, Cambridge, UK

Marcel A J 1983 Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology* **15**: 197–37

Marcel A J 1993 Slippage in the unity of consciousness. In: Bock G R, Marsh J (eds.) *Experimental and Theoretical Studies in Consciousness (Ciba Foundation Symposium No. 174)*. Wiley, Chichester, UK, pp. 168–86

Milner A D, Goodale M A 1998 *The Visual Brain in Action*. Oxford University Press, Oxford, UK

Milner B 1962 Les troubles de la mémoire accompagnant des lésions hippocampiques biltatérales. In: *Physiologie de l'hippocampe*. Centre National de la Recherche Scientifique, Paris, pp. 257–72

Nisbett R E, Wilson T D 1977 Telling more than we can know: Verbal reports on mental processes. *Psychological Review* **84**: 231–59

Nissen M J, Bullemer P 1987 Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology* **19**: 1–32

Perruchet P 1985 A pitfall for the expectancy theory of human eyelid conditioning. *Pavlovian Journal of Biological Science* **20**: 163–70

Reber A S 1967 Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* **6**: 855–63

Reingold E M, Merikle P M 1988 Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics* **44**: 563–75

Shacter D L, Buckner R L, Koutsaal W 1998 Memory, consciousness and neuroimaging. *Philosophical Transactions of the Royal Society of London, Series B* **353**: 1861–78

Shanks D R, St. John M F 1994 Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences* **17**: 367–447

Simons D J, Chabris C F 1999 Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception* **28**: 1059–94

Van Heteren C F, Boekkooi P F, Jongsma H W, Nijhuis J G 2000 Fetal learning and memory. *Lancet* **356**: 1169–70

Warrington E K, Weiskrantz L 1968 New method of testing long term retention with special reference to amnesic patients. *Nature* **217**: 972–4

Weiskrantz L 1986 *Blindsight: A Case Study and Implications*. Oxford University Press, Oxford, UK

A. Cleeremans

Consciousness and Sensation: Philosophical Aspects

1. Types of Consciousness

The term 'consciousness' picks out several different aspects of mental functioning. Though these aspects are related in various ways, it is crucial for any theoretical understanding of consciousness to distinguish the different things to which we apply the term.

The oldest and most widespread use of the term is to describe people, and other sentient creatures, as being awake and receptive to sensory input. This is also the least problematic mental phenomenon we refer to as consciousness. The term is used in this case simply to mark the distinction between a person or other creature being in an ordinary waking state and its being asleep, knocked out, anesthetized, or failing for some other reason to interact mentally with its environment in ordinary ways. The distinction between these conditions is doubtless to be explained in large measure in biological terms. Because this first notion of consciousness applies to people and other sentient creatures, we may refer to it as creature consciousness.

A second notion of consciousness has to do with a person or other animal's being conscious or aware of something, or its being conscious or aware that something is the case. A person is conscious of something when that person sees or hears it or perceives it in some other way. Following on that perceptual model, we also count a person as being conscious of something when that person has a thought about that thing as being present. Finally, a person is conscious that something is the case when that person has a thought that it is the case. This second notion of consciousness again applies to people and other creatures, but we describe this kind of consciousness in terms of some object of which a person or other creature is conscious. For that reason it is convenient to refer to this second kind of consciousness as transitive consciousness.

A third notion of consciousness applies not to people or other creatures at all, but to their mental states, states such as perceiving, sensing, thinking, desiring, and feeling. A mental state is conscious just in case one is conscious of it in a way that seems spontaneous and unmediated. An important theoretical question, taken up in Sect. 4, is how to explain the way one is conscious of one's mental states when those states are conscious. But it seems clear that no mental state counts as being conscious if the subject is in no way conscious of that state. Because this third use of the term 'consciousness' applies to mental states, rather than people or other creatures, it is useful to refer to it as state consciousness.

Some theorists have insisted that a person is never in any of these states without being conscious of them. This claim is perhaps best represented by philosophers such as René Descartes (1596–1650) and John Locke (1643–1704), although many thinkers today also hold this. One reason many find this claim tempting may be that in ordinary situations we care relatively little about any mental states that the subject is wholly unaware of. It seldom matters to anybody if a person thinks or feels something without being in any way aware of doing so; still less in the case of other creatures. But the temptation to insist that we are conscious of all our mental states also seems to reflect a certain confusion. When a person perceives something or has a thought about it, that person is in a state of being conscious of that thing. It may seem an easy step to pass from somebody's being in a state of being conscious of something to that state's being a conscious state.

But it is crucial to preserve the distinction between the things one's mental states are about and 'the mental states that are about those things. It is one thing to have a thought or perception of something and another altogether to be conscious of that thought or perception. Even if the thought or perception is not conscious, because one is in no way conscious of it, one's having that thought or perception will still make one conscious of whatever the thought or perception is about. One will then be conscious of something, but not consciously so. We mark this distinction by a third use of the term 'consciousness,' on which a mental state is conscious if one is conscious of that state in a spontaneous, seemingly unmediated way.

The condition about seeming to be spontaneous and unmediated is important. A thought or perception is not conscious if one is conscious of it by applying a theory to oneself, or because one takes somebody else's word for it. One must be conscious of oneself as being in the state in question without seeming to infer that from anything else.

In ordinary waking life, many mental states are conscious, but normally they are not conscious in any reflective or attentive way. When one focuses attentively on one's conscious states, so that one's interest is in the mental state one is in rather than on whatever that mental state makes one conscious of, one is said to introspect the mental state under scrutiny. In these cases one is not only conscious of the mental state one is in, but conscious also that one is conscious of that state. Ordinary state consciousness is not at all reflective in this way.

Both creature consciousness and transitive consciousness can be understood only if we first understand what it is to perceive and think about things. A person or other creature is conscious only if it is mentally responsive to its environment, and one is transitively conscious of something only if one is in a suitable kind of mental state. To understand state consciousness, by contrast, requires something in addition. One must understand what it is to think about or perceive something, and also understand the difference between thinking and perceiving consciously and doing so in a way that is not conscious. Because of disagreement about how to understand that contrast, state consciousness continues to provoke considerable controversy in psychological and philosophical theory, and seems to some even to resist scientific explanation altogether.

2. *Sensations and Consciousness*

The most difficult theoretical issues surrounding state consciousness arise in connection with the conscious-

ness of sensations. Sensations fall into two groups: bodily sensations and perceptual sensations. Bodily sensations include such mental states as pains, aches, tickles, itches, and feelings of nausea, dizziness, and the like. Perceptual sensations, by contrast, include those which pertain to the five exteroceptive senses of sight, hearing, taste, touch, and smell.

There are two additional kinds of sensation, which provide instructive borderline cases. Proprioceptive sensations pertain to the motions and position of our limbs, and enteroceptive sensations pertain to the feelings of our viscera and other internal organs. These two groups resemble bodily sensations in having to do with bodily conditions, but they also resemble perceptual sensations in providing information about the conditions of things, in this case one's own body. The existence of these borderline cases suggests the hypothesis, defended by some theorists (e.g., Armstrong 1962), that bodily sensations, along with proprioceptive and enteroceptive sensations, are simply perceptual sensations that are involved in perceiving bodily conditions. Pains, for example, reveal bodily damage, aches reveal bodily strains and stresses, and so forth for other types of bodily sensation.

Traditional theorists have usually distinguished between sensations and perceptions, though this distinction has been drawn in a variety of different ways. Some theorists have isolated sensations as states that occur in earlier stages of perceptual processing. But this is unhelpful unless one can specify exactly where in the early processing stream sensations are supposed to occur. Some theorists have sought to meet this difficulty by holding that sensations occur early enough in perceptual processing to represent things accurately, whereas perceptions may reflect distortions due to subsequent mental processing. But this is still unsatisfactory, since some distortion occurs even in the earliest stages of sensing. In any case, work on visual processing suggests that the early stages involve nothing that would ordinarily be called a sensation (Treisman 1986). An altogether different approach to the distinction holds that perceiving involves some conceptual content, whereas sensing has only qualitative character and no conceptual content. Even if sensing, so understood, often does not occur as a separate stage of visual processing, one can still abstract the purely qualitative aspect of perceiving and refer to that as sensing.

Conscious sensations exhibit characteristic qualitative character; for every type of conscious sensation, there is something it's like to have that sensation consciously. Many theorists hold that no informative account is possible of this conscious qualitative character, that what it's like to have a conscious sensation resists theoretical treatment of whatever sort (e.g., Chalmers 1996, Nagel 1974, Levine 1993, McGinn 1991). No theoretical treatment could capture or convey the knowledge of what it's like to have any particular type of conscious sensation; the only way one can come to have that knowledge is by having the conscious sensation itself (Jackson 1982).

Other theorists, however, have argued that one can give an informative, theoretical account of what it's like to have particular types of conscious sensation. Sensations play distinctive a role in perceiving and the various types of perceptions and perceptual beliefs are keyed to specific types of perceptible property. So it may be possible to give an informative theoretical treatment of the various types of sensation in terms of characteristic causal connections sensations have with perceptions and perceptible physical objects (e.g., Shoemaker 1975, 1984 and Sellars 1963, Chaps. 2, 3, and 5). Such an account fits well especially with the hypothesis that bodily sensations are a special case of perceptual sensations.

There are two ways theorists seek to distinguish among types of sensation: in terms of the characteristic role they play in perceiving and in terms of what it's like for one to have them. Some have held that these two ways of taxonomizing sensations might give different results. Consider sensations that play the same role in perceiving, say, sensations that result from visual stimulation by a ripe tomato in good lighting. Perhaps what it's like for one person to have such a sensation is different from what it's like for another person to have a sensation picked out by that perceptual role. Perhaps, more generally, what it's like for somebody to have a sensation bears no predictable relation to the characteristic perceptual cause of that type of sensation. The possibility that what it's like for different people might be independent of characteristic perceptual cause is known as the possibility of qualia inversion. (Qualia, singular 'quale,' are the conscious qualitative properties of sensations.) If such inversion occurs, what it's like for one person to have a sensation characteristic of seeing a ripe tomato might be the same as what it's like for another to have a sensation characteristic of seeing grass.

It may seem that such qualia inversion occurs all the time, since people do see things slightly differently. But these ordinary differences are readily detectable, relying only on the way people describe the things they see in perceptual terms. So these differences can be treated as a function of the perceptible properties people take physical objects to have. Qualia inversion would occur, by contrast, if what it's like to have various sensations were wholly independent of the connections those sensations have to perceptible properties. One would therefore be unable to detect such qualia inversion by reference to the way people describe their experiences in perceptual terms. If what it's like for people to have the relevant sensations were independent of the properties they perceive objects to have, people would use the same words to describe the objects they perceive, but what it would be like for them to perceive those subjects would be different.

It is unclear, however, that such undetectable inversion could occur. It is arguable that we are

conscious of our sensations in respect of the role they play in perceiving objects of characteristic sorts. But what it's like for one to have a sensation is a function of how one is conscious of that sensation. One can, for example, be conscious of sensations in more or less fine-grained ways; the visual sensation one gets from seeing a particular red object might be conscious just as some indiscriminate shade of red or as a highly specific shade, depending on factors such as how carefully one is attending, how alert one is, and so forth. If what it's like for one to have a sensation is a function of how one is conscious of it and one is conscious of one's sensations in respect of their role in perceiving, what it's like for one to have a sensation will not vary independently of that role in perceiving (Rosenthal 1999).

Mental states such as thoughts and beliefs do not exhibit qualitative properties, and some theorists insist that sensations and perceptions also exhibit no qualitative properties. This view, known as representationalism, holds that the only qualitative properties involved in sensing and perceiving are the qualities we take physical objects and processes to have (Dretske 1995, Harman 1990, Lycan 1996).

It is clear that no mental states have the qualitative properties of color, sound, and so forth that we perceive physical objects to have. But that does not show that sensations do not have a distinct, mental version of such qualitative properties. We do describe sensations in qualitative terms, visual sensations as being red and green, for example, and pains as being sharp or burning. These qualities of being red, green, sharp, or burning are not the same type of property that physical objects have, but sensations do fall under different types in respect of having one or another type of qualitative property.

It is likely that our grasp of the distinguishing properties of sensations trades on connections those properties have to perceptible properties of physical objects; red sensations occur in response to red objects, and so forth. But sensations sometimes occur in the absence of any objects of the kind that characteristically occasion them. So the sensations themselves must have distinguishing properties that parallel those of the objects that typically occasion them. These distinguishing properties are the qualitative properties of sensations.

3. *Do Mental States Ever Fail to be Conscious?*

It is generally recognized that some types of mental state can occur without being conscious. Thoughts, desires, wishes, expectations, and other such states all sometimes occur without the subject's having any conscious access to them. Indeed, novelists and playwrights have long described their characters as thinking and wanting things that they are sometimes wholly unaware of thinking and wanting.

Freud's theoretical and clinical arguments for positing unconscious thoughts and desires can be seen as continuous with this commonsense recognition of such unconscious states. Unconscious thoughts and desires, according to Freud, often make possible explanations of behavior that seems inexplicable by reference to conscious thoughts and desires (Freud 1915/1966–74). The unconscious thoughts and desires he posited interact to cause behavior in just the way that corresponding conscious thoughts and desires would to cause that same behavior.

Much in Freud's clinical and theoretical thinking remains controversial, but the general point seems sound. There are theoretically neutral examples of the operation of such unconscious thoughts and desires; verbal slips, for example, which Freud called parapraxis, sometimes provide evidence of thoughts or desires of which the subject is wholly unaware (Freud 1915/1966–74, pp. 166, 168–9).

More dramatically, work in social psychology has shown that people often report having thoughts and desires that it can be established experimentally they do not have (Nisbett and Wilson 1977, White 1988). Subjects in these cases confabulate thoughts and desires that seem to make *ex post facto* sense of their behavior, by rationalizing some situation or conforming to expectations or preconceived ideas that they or others have. When subjects confabulate thoughts and desires in this way, they typically remain wholly unaware of the thoughts and desires that actually cause that behavior. Experimental work on metacognition, which is how much subjects know of their own cognitive states and processes, also reveals nonconscious cognitive states (Metcalfe and Shimamura 1994 and Nelson 1992).

Even if thoughts and desires occur without being conscious, some theorists continue to doubt that this is possible with qualitative states, such as sensations. Even Freud held that, because of their qualitative nature, feelings and emotions cannot occur without being conscious (Freud 1915/1966–74, p. 177, 1923/1966–74, pp. 22–3).

There is something it's like for one to be in a qualitative state, such as an emotion or sensation, only when that qualitative state is conscious. When one is in no way conscious of a sensation, there is nothing it's like to have it, that is, nothing it's like for the subject. So, if we taxonomize sensations only in respect of what it's like for one to have them, sensations would have to be conscious to have the distinguishing characteristics of sensations at all. No state would count as a sensation or other qualitative state unless it was conscious.

But if we taxonomize sensations at least in part by reference to their role in perceiving, there is no reason to deny the occurrence of nonconscious sensations, since it is experimentally demonstrable that perceiving occurs without being conscious. It is well-established, for example, that when subjects are presented very

briefly with visual or auditory stimulus material followed by a second stimulus, the second stimulus may mask the first, blocking conscious access to it. Nonetheless, it can often be shown in these masked-priming experiments that the first stimulus has a distinctive perceptual effect (e.g., Marcel 1983a, 1983b).

Research on some subjects with a part of the primary visual cortex destroyed yields similar results. When such subjects are presented with visual stimuli in the area of the visual field served by the damaged part of the cortex, they report seeing nothing at all. Some subjects, however, can guess the visible characteristics of the stimulus with startlingly high accuracy (Weiskrantz 1986, 1997). This phenomenon is known as blind sight. In these cases and in masked-priming experiments, subjects' guesses, often in forced-choice experiments, are highly accurate in the absence of any relevant conscious mental states. Such experimental work leaves little doubt that nonconscious perceiving occurs and that whatever sensations occur in such perceiving also fail to be conscious.

4. *Theories of Consciousness*

A mental state is conscious only if one is conscious of that state in a way that seems to one spontaneous and unmediated. We do not regard as conscious any state the subject is in no way conscious of being in, or conscious of being in only as a result of inference or observation.

Theories of state consciousness differ chiefly in how they seek to explain the way people are conscious of their conscious states. As noted earlier, there are two ways one is transitively conscious of things: by sensing them and lay thinking about them as being present. Corresponding to those two ways of being transitively conscious of things are two hypotheses about the way we are conscious of our conscious states.

On one hypothesis, we are conscious of our conscious states by sensing them. This theory has a long history in philosophical writing. Aristotle's claim that we always perceive that we perceive (Aristotle 1968, 425b12) was developed by John Locke (1700/1975) and Immanuel Kant (1787/1998) into a theory of consciousness as a kind of 'inner sense.' On this view, represented by such contemporary writers as Armstrong (1978) and Lycan (1996), a mental state is conscious if one perceives that state. Because perceiving something seems to make us directly aware of that thing, if we perceive our conscious states, that would explain why it seems that we are directly aware of those states.

But we are also conscious of things when we think of them as being present. So it may be that a mental state is conscious if one has a thought to the effect that one is in that state (Rosenthal 1986, 1997). We would seldom be aware of such higher-order thoughts about our mental states, since they would not themselves be conscious thoughts unless we had third-order thoughts about them, which would be unusual. Very likely that would occur only when we are introspectively conscious of our conscious states. On this hypothesis, higher-order thoughts arise independently of any conscious inference, that is, any inference of which we are conscious. That would explain, without appeal to the apparent immediacy of perceiving, why it seems we are directly aware of our conscious states.

A variant of this model holds that actual higher-order thoughts are not required for a state to be conscious; all that is needed is that one be disposed to have a higher-order thought about that state (Carruthers 1996). But this is less satisfactory, since being disposed to have a thought about something does not result in one's being conscious of that thing, but only in one's being disposed to be conscious of it.

The higher-order thought model has certain advantages over the inner-sense hypothesis. Perceiving always seems to involve mental qualities, but we are never aware of any higher-order mental qualities. In addition, mental qualities all fall into families corresponding to the various sensory modalities, such as sight, hearing, bodily sensations, and so forth. But it is difficult to imagine what sensory modality could be involved in higher-order sensing or perceiving.

Another issue that divides theorists is whether our being conscious of our conscious mental states is intrinsic to the mental states we are conscious of. Looked at from the point of view of the person whose mental states are conscious, the property mental states have of being conscious seems to be intrinsic to those states, and it seems that they could not have occurred without having that intrinsic property. Accordingly, some theorists have adopted a model on which the higher-order thought or sensing in virtue of which we are conscious of our conscious states is intrinsic to those states (Brentano 1874/1973).

Most contemporary theorists, however, argue that a more satisfactory theoretical model results from assuming that the relevant higher-order states are distinct from the states they make us conscious of. It is well known, moreover, that commonsense impressions about which properties are intrinsic and which are relational are often wrong. As with physical properties such as weight, properties that initially seem to be intrinsic often turn out to be relational. So being conscious may well not be an intrinsic property of conscious states, despite its seeming, from the point of view of the person whose mental states are conscious, that it is.

Not all theorists appeal to inner sense or higher-order thoughts to explain how we are conscious of our conscious states. Some suppose that we are conscious of those states in some third way that is unique to being conscious of mental states. Such theorists often

appeal to metaphor to capture this third way of being conscious of things, these as a search light that illuminates mental states, thereby making them conscious. Such metaphorical description makes it difficult to evaluate these proposals. In the absence of some clear account of this third way of being conscious of things, it is likely that positing such a third way is simply a label for the problem of explaining how we are conscious of our conscious states, rather than a substantive explanation.

Some have gone farther, and reject even the commonsense idea that one is conscious of one's conscious mental states (Dretske 1995, Searle 1992). On this view, a mental state's being conscious consists not of one's being conscious of that state in some suitable way, but rather of its being a state in virtue of which one is conscious of something. But all sensing and perceiving and much if not all thinking makes one conscious of things. So this account of what it is for a mental state to be conscious obliterates the distinction between mental states that are conscious and those which are not.

Some theorists distinguish various ways in which mental states can be conscious. On one of the most influential accounts, a state's being access conscious consists of it being poised to play an active role in rationally guiding action, speech, and reasoning, whereas a state has phenomenal consciousness if it has qualitative character (Block 1995).

It is unclear, however, that either notion does justice to our pretheoretic, commonsense idea of a mental state's being conscious. Many states that are poised to guide action, speech, and reasoning are conscious states, but many are not. Nonconscious states often play an important role in the rational guidance of action, speech, and reasoning. So a state's being poised for these tasks cannot subserve that state's being conscious; intuitively, it is necessary also that one be in some way conscious of the state.

As for phenomenal consciousness, a state's simply having qualitative properties cannot ensure that the state is conscious; as masked-priming experiments and work on blindsight show, qualitative states occur without being in any intuitive way conscious. In addition, if phenomenal consciousness were just having qualitative properties, sensations could not occur without being conscious, since they cannot occur without qualitative properties. This sense of consciousness would again obliterate the distinction between a state's being conscious and not being conscious.

5. *Consciousness and Interpretation*

We often determine what thoughts and desires other people have by inferring from what they do to the thoughts and desires that would best explain or make sense of that behavior. Sometimes people also say what they think and desire, but often all we have to go on is that seeing others as having particular thoughts and desires would make sense of their behavior.

Daniel Dennett has argued attributing beliefs and desires to others in this requires seeing others as being rational in doing what will get them what they want, given what they believe. He argues that this explanatory strategy is effective with much behavior that cannot readily be explained by the physical stance, which appeals to how something operates as a physical object, or the design stance, which regards objects as well-designed machines. This explanatory framework, which Dennett labels the intentional stance, seeks to rationalize the behavior of people and other creatures by seeing that behavior as due to their thoughts and desires (Dennett 1987).

Dennett has developed a theory of consciousness, on which consciousness in effect results from people's adopting of the intentional stance toward themselves. A person's conscious thoughts and desires are those one is disposed to report having or in some other way takes oneself to have, thereby making sense of one's behavior. Consciousness is a matter of our spontaneously interpreting ourselves as being in various mental states. So there are, in this view, no determinate facts about consciousness beyond such self-interpretive behavior (Dennett 1991).

Spontaneously interpreting oneself as being in some mental state can be seen as resulting from one's having a spontaneous higher-order thought to the effect that one is in that state. So Dennett's view has much in common with the higher-order-thought model. But there are important differences. Dennett takes self-interpretative behavior to be the last word about what conscious states we are in. But because such behavior cannot by itself explain how one is conscious of one's conscious states, the higher-order-thought model takes that behavior to be a mere symptom of the higher-order thoughts in virtue of which one is conscious of those states. So on that model there will be determinate facts about consciousness independent of self-interpretative behavior.

Spontaneous self-interpretation is on both views the last word about how consciousness represents our mental lives, since a state's being conscious consists, on these views, of one's spontaneous interpretation of oneself as being in that state. But what states consciousness represents one as being in is not the last word on what states one is actually in. Thoughts, desires, perceptions, and feelings all have characteristic causal connections with stimulus conditions and behavior and with one another. These causal connections can determine what mental states a person is in independently of what states consciousness represents that person as being in.

If a mental state's being conscious involves spontaneous self-interpretation, it is natural to speculate that such self-interpretation may be influenced by the interpretations others make of us as being in particular

mental states. The interpretations others make of us may even play a role in our coming, as very young children, to learn to apply mental concepts to ourselves.

The hypothesis that a state's being conscious is a matter of one's spontaneously interpreting oneself as being in that state fits well with the way people sometimes confabulate having thoughts and desires that they do not actually have (Nisbett and Wilson 1977, White 1988). From the point of view of the person confabulating, these cases are subjectively indistinguishable from cases of actually having the relevant conscious thoughts and desires. Consciousness represents one as having thoughts and desires that one does not actually have. Some research suggests that the same may well occur with qualitative states, such as bodily or perceptual sensations (e.g., Staats et al. 1998, Holmes and Frost 1976).

Much work in social sciences such as sociology, cultural anthropology, social psychology, and history relies on the interpretation of people as being in particular mental states. This raises a problem if the mental states one is in can diverge from the mental states one is conscious of oneself as being in. One could interpret a person as actually being in a particular mental state or simply as being conscious of being in that state. If the mental states one is in can differ from those one is conscious of oneself as being in, one interpretation might be correct and the other not. So it is methodologically important to be clear about which interpretation is in question.

See also: Knowledge (Explicit and Implicit): Philosophical Aspects; Self-knowledge: Philosophical Aspects

Bibliography

Aristotle 1968 *De Anima* [trans. Hamlyn D W]. Clarendon Press, Oxford, UK

Armstrong D M 1962 *Bodily Sensations*. Routledge & Kegan Paul, London

Armstrong D M 1978 What is consciousness? *Proceedings of the Russellian Society* **3**: 65–76 [reprinted in expanded form in Armstrong D M 1980 *The Nature of Mind*. University of Queensland Press, St. Lucia, Australia, pp. 55–67]

Block N 1995 On a confusion about a function of consciousness. The *Behavioral and Brain Sciences* **18**(2): 227–47, with Open Peer Commentary, 247–72, and Author's Response, How many concepts of consciousness? 272–87

Brentano F 1874/1973 *Psychologie vom empirischen Staadpunkt* [McAlister L L (ed.) *Psychology from an Empirical Standpoint*, trans. Rancurello A C, Terrell D B, McAlister L L]. Routledge & Kegan Paul, London

Carruthers P 1996 *Language, Thought, and Consciousness: An Essay in Philosophical Psychology*. Cambridge University Press, Cambridge, UK

Chalmers D J 1996 *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York

Dennett D C 1987 *The Intentional Stance*. MIT Press/Bradford Books, Cambridge, MA

Dennett D 1991 *Consciousness Explained*. Little, Brown and Company, Boston

Dretske F I 1995 *Naturalizing the Mind*. MIT Press/Bradford Books, Cambridge, MA

Freud S 1915/1966–74 The unconscious. In: Strachey J (ed.) *The Complete Psychological Works of Sigmund Freud* [trans. Strachey J]. The Hogarth Press, London, pp. 166–215

Freud S 1923/1966–74 The ego and the id. In: Strachey J (ed.) *The Complete Psychological Works of Sigmund Freud* [trans. Strachey J]. The Hogarth Press, London, pp. 3–68

Harman G 1990 The intrinsic quality of experience. *Philosophical Perspectives* **IV**: 31–52; reprinted in Harman, *Reasoning, Meaning and Mind*. Clarendon Press, Oxford, UK, 1999, pp. 244–61

Holmes D S, Frost R O 1976 Effect of false autonomic feedback on self-reported anxiety, pain perception, and pulse rate. *Behavior Therapy* **7**(3): 330–4

Jackson F 1982 Epiphenomenal qualia. *Philosophical Quarterly* XXXII **127**(April): 127–36

Kant I 1787/1998 *Critique of Pure Reason* [trans. Guyer P, Wood A W]. Cambridge University Press, Cambridge, UK

Levine J 1993 On leaving our what it's like. In: Davies M, Humphreys G W (eds.) *Consciousness: Psychological and Philosophical Essays*. Basil Blackwell, Oxford, UK, pp. 121–36

Locke J 1700/1975 Nidditch P H (ed.) *An Essay Concerning Human Understanding* Oxford University Press, Oxford, UK, 4th edn.

Lycan W 1996 *Consciousness and Experience*. MIT Press/Bradford Books, Cambridge, MA

Marcel A J 1983a Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology* **15**: 197–237

Marcel A J 1983b Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology* **15**: 238–300

McGinn C 1991 *The Problem of Consciousness*. Basil Blackwell, Oxford, UK

Metcalfe J, Shimamura A P (eds.) 1994 *Metacognition: Knowing about Knowing*. MIT Press/Bradford Books, Cambridge, MA

Nagel T 1974 What is it like to be a bat? *The Philosophical Review* LXXXIII, **4**(October): 435–50

Nelson T O (ed.) 1992 *Metacognition: Core Readings*. Allyn and Bacon, Boston

Nisbett R E, DeCamp Wilson T 1977 Telling more than we can know: Verbal reports on mental processes. *Psychological Review* LXXXIV **3**(May): 231–59

Rosenthal D M 1986 Two concepts of consciousness. *Philosophical Studies* **49**(3): 329–59

Rosenthal D M 1997 A theory of consciousness. In: Block N, Flanagan O (eds.) *The Nature of Consciousness: Philosophical Debates*. MIT Press, Cambridge, MA, pp. 729–53

Rosenthal D M 1999 Sensory quality and the relocation story. *Philosophical Topics* **26**(1/2): 321–50

Searle J R 1992 *The Rediscovery of the Mind*. MIT Press, Cambridge, MA

Sellars W 1963 *Science, Perception and Reality*. Routledge & Kegan Paul, London

Shoemaker S 1975 Functionalism and Qualia. *Philosophical Studies* XXVII **5**(May): 292–315

Shoemaker S 1984 *Identity, Cause, and Mind: Philosophical Essays*. Cambridge University Press, Cambridge and New York

Staats P, Hekmat H, Staats A W 1998 Suggestion/placebo effects on pain: Negative as well as positive. *Journal of Pain and Symptom Management* **15**(4): 235–43
Treisman A 1986 Features and objects in visual processing. *Scientific American* **255**(5): B114–25
Weiskrantz L 1986 *Blindsight: A Case Study and Implications*. Oxford University Press, Oxford, UK
Weiskrantz L 1997 *Consciousness Lost and Found: A Neuropsychological Exploration*. Oxford University Press, Oxford, UK
White PA 1988 Knowing more about what we can tell: Introspective access and causal report accuracy 10 years later. *British Journal of Psychology* **79**(1): 13–45

D. M. Rosenthal

Consciousness, Cognitive Psychology of

The 'Hard Problem' of consciousness is how to explain a state of consciousness in terms of its neurological basis. Chalmers (1996) distinguishes between the Hard Problem and 'easy' problems that concern the function of consciousness. The Hard Problem (though not under that name) was identified by Nagel (1974) and further analyzed in Levine (1983).

There are two *prima facie* reasons for thinking that the Hard Problem has no solution.

(a) Actual failure. In fact, no one has been able to think of even a highly speculative answer.

(b) Principled failure. The materials we have available seem ill-suited to providing an answer. As Nagel (1974) says, an answer to this question would seem to require an objective account that necessarily leaves out the subjectivity of what it is trying to explain. We don't even know what would *count* as such an explanation.

1. *Perspectives on the Hard Problem*

There are many perspectives on the Hard Problem but I will mention only the four that comport with a naturalistic framework.

(a) *Phenomenal realism*, the view that consciousness is a substantial property that cannot be philosophically reduced in nonphenomenal terms. According to most contemporary phenomenal realists, consciousness plays a causal role and its nature may be found empirically as the sciences of consciousness advance. Phenomenal realism is compatible with the empirical reduction of consciousness to neurological or computational properties of the brain. (This is the perspective of this article.) This view accepts the Hard Problem but aims for an empirical solution to it. (Block 1995, Flanagan 1992, Nagel 1974, Searle 1992; McGinn 1991 argues that an empirical reduction is possible but that we can't find or understand it.)

(b) *Dualistic naturalism*. In this catch-all category, I include Chalmers' (1996) view that standard materialism is false but that there are naturalistic options to dualism such as panpsychism. Nagel (forthcoming) proposes that there is a deeper level of reality that is the naturalistic basis both of consciousness and of neuroscience.

(c) *Eliminativism*, the view that consciousness as understood above simply does not exist (Dennett 1979, Rey 1997). So there is nothing for the Hard Problem to be about.

(d) *Philosophical reductionism*. Philosophical reductionists (e.g., Dennett 1991) move closer than eliminativists to common sense by allowing that consciousness exists, but they 'deflate' this commitment—again on philosophical grounds—taking it to amount to less than meets the eye (as Dennett might put it). One prominent form of philosophical reductionism in this sense makes a conceptual reductionist claim: that consciousness can be conceptually analyzed in nonphenomenal terms. The main varieties of analyses are functional, representational, and cognitive. Many philosophical reductionists disparage *a priori* conceptual analysis, but nonetheless hold that we know on purely philosophical grounds (independently of the sciences of consciousness) that consciousness is constituted by the functional, representational, or cognitive.

Here are some examples of philosophical reductionism. Pitcher (1971) and Armstrong (1968) can be interpreted as analyzing consciousness in terms of beliefs. One type of prototypical conscious experience, as of the color blue, is a matter of an inclination (perhaps suppressed) to believe that there is a blue object in plain view. (See Jackson 1977 for a convincing refutation.) A different analysis appeals to higher order thought or higher order perception. These theorists take the concept of a conscious pain to be the concept of a pain that is accompanied by another state that is about that pain. A pain that is not so accompanied is not a conscious state (Armstrong 1968, Carruthers 1992, Lycan 1990, Rosenthal 1997 advocates a higher order thought view as an empirical identity rather than as a conceptual analysis). Another philosophical reductionist view that is compatible with the analyses in terms of beliefs concerns not the states themselves but their contents. Representationism holds that it can be established philosophically that the phenomenal character of experience is its representational content. Many representationists reject conceptual analysis, but still their accounts do not depend on details of the science of mind; if any science is involved, it is evolutionary theory (see Harman 1990, Dretske 1995, Lycan 1996, McDowell 1994, Tye 1995; Shoemaker 1994 mixes phenomenal realism with representationism in an interesting way). Conceptual functionalists say that the concept of consciousness is analyzable functionally (Lewis 1994).

According to these views, there is such a thing as consciousness, but there is no Hard Problem, that is, there are no mysteries concerning the physical basis of consciousness that differ in kind from run of the mill unsolved scientific problems about the physical/functional basis of liquidity, inheritance, or computation.

2. *How to Begin Solving the Hard Problem*

The main element in the approach I shall be suggesting to the Hard Problem is a distinction that is widely appealed to in discussions of Jackson's (1982) famous 'Mary' example. Jackson imagined a neuroscientist of the distant future who is raised in a black and white room and who knows everything scientific there is to know about color and the experience of it. But when she steps outside the room for the first time, she learns what it is like to see red. Jackson argued that since the scientific facts don't encompass the new fact that Mary learns, dualism is true.

The line of response to Jackson that I think wins the day (and which derives from Loar 1990/1997) involves appeal to the distinction between a property and a concept of that property. (See Churchland 1989, Lycan 1990, van Gulick 1993, Sturgeon 1994, Tye 1999, Perry forthcoming.)

A concept is a thought element in a way of thinking, a kind of representation. If you like, you can take concepts in my sense to be interpreted symbols in a 'language of thought.' This usage contrasts with a common philosophical usage in which a concept is something like a meaning. Concepts in my sense are individuated in part by meanings; x and y are instances of the same concept if and only if they are instances of the same representation and have the same meaning.

Someone could believe that *this color* is useful for painting pots but that red is not, even if *this color* = red. Our experiential concept of red differs from our linguistic concept of red. An experiential concept involves a phenomenal element, a phenomenal way of thinking, for example a mental image that is in some sense of red or an ability to produce such a mental image or at least an ability to recognize red—which arguably, could not be done without some phenomenal mental element.

Importantly, we can have an experiential concept *of an experience* (which we can call a phenomenal concept, phenomenal concepts being a subclass of experiential concepts) as well as an experiential concept of a color. And the very same mental image may be involved in both concepts. The difference between the phenomenal concept of the experience and the experiential concept of the color lies in the *rest* of the concept—in particular, the way the phenomenal element functions in the concept. This can be a matter of further concepts explicitly invoked in the concept—the concept of a color in one case and the concept of an experience in the other. One type of experiential concept (of a color or of an experience) involves a demonstrative plus a mental image plus a language-like representation, e.g. 'that [attention to a mental image] color' or 'that [attention to a mental image] experience' where the bracket notation is supposed to indicate the use of attention to a nondescriptive element, a mental image, in fixing the demonstrative reference. Loar (1990/1997) gives an example involving two concepts in which something like a mental image of a cramp feeling is used to pick out a muscular knot in one concept and in the other concept, the cramp experience itself. The two concepts in my notation would be 'that [attention to a mental image] cramp' and 'that [attention to a mental image] cramp experience.'

In these terms, then, Mary acquired a new concept of a property she was already acquainted with via a different concept. In the room, Mary knew about the subjective experience of red via physical concepts. After she left the room, she acquired a phenomenal concept of the same property. So Mary did not learn a new fact. She learned a new concept of an old fact. She already had a third person understanding of the fact of what it is like to see red. What she gained was a first person understanding of the very same facts. She knew already that corticothalamic oscillation of a certain frequency is what it is like to see red. What she learned is that this [attention to a mental image] is what it is like to see red. So the case provides no argument that there are facts that go beyond the physical facts.

Recall that there is a principled reason why mind–body identity seemed impossible: that a first person subjective property could not be identical to a third person objective property. But the concept–property distinction allows us to see that the subjective–objective distinction in this use of it and the first person–third person distinction are distinctions between kinds of concepts, not kinds of properties. There is no reason why a subjective concept and an objective concept cannot pick out the same property. Thus we can substitute a dualism of concepts for a dualism of properties.

3. *The Paradox of Recent Empirical Findings About Consciousness*

The most exciting line of experimental investigation of consciousness in recent years uses the phenomenon of binocular rivalry. If two different stimuli—e.g., horizontal and vertical stripes—are presented to each of one's eyes, one does not see a blend, but rather alternating horizontal and vertical stripes. Logothetis and his colleagues (Logothetis 1998) trained monkeys to pull different levers for different patterns. They then presented different patterns to the monkeys' two eyes, and observed that with monkeys, as with people, the monkeys switched back and forth between the two levers even though the sensory input remained the

same. Logothetis recorded the firings of various neurons in the monkeys' visual systems. In the lower visual areas (e.g., V1), 80 percent of the neurons did not shift with the percept. But further along the occipital-temporal pathway, 90 percent shifted with the percept. So it seems that later areas in the occipital-temporal pathway—let's call it the 'ventral stream'—are more dominantly part of the neural basis of (visual) consciousness than early areas.

However, what has also become commonplace is activation of the very same ventral stream pathways without awareness. Damage to the inferior parietal and frontal lobes have long been known to cause visual extinction in which subjects appear to lose subjective experience of certain stimuli on one side, viz., when there are stimuli on both sides. Extinction is associated with visual neglect in which subjects do not notice stimuli on one side. For example, neglect patients often do not eat the food on the left side of the plate.

Driver and Vuilleumier (2001) point out that the ventral stream is activated for extinguished stimuli (i.e., which the subject claims not to see). Rees et al. (in press) report studies of a left-sided neglect and extinction patient on face and house stimuli. Stimuli presented just on the left side are clearly seen by the patient, but when there are stimuli on both sides, the subject says he sees only the stimulus on the right. However, the 'unseen' stimuli show activation of the ventral pathway that is the same in location and temporal course (though lower in activation) as the seen stimuli. The paradox then is that our amazing success in identifying the neural correlate of visual experience in normal vision has led to the peculiar result that, in masking and neglect, that very neural correlate occurs without, apparently, subjective experience.

What is the missing ingredient, X, which, added to ventral activation, constitutes conscious experience? Kanwisher (forthcoming) and Driver and Vuilleumier (2001) offer pretty much the same proposal as to the nature of X, that the missing ingredient is binding perceptual attributes with a time and a place, a token event.

Whether or not any of these proposals are right, the search for X seems to me the most exciting current direction for consciousness research.

There is a very different approach to the nature of consciousness. Dennett (1994) postulates that consciousness is 'cerebral celebrity.' What it is for a representation to be conscious is for it to be widely available in the brain. Dehaene and Naccache (forthcoming) say consciousness is being broadcast in a global neuronal workspace.

The theory that consciousness is ventral stream activation plus e.g., neural synchrony, and the theory that consciousness is broadcasting in the global neuronal workspace, are instances of the two major approaches to consciousness in the philosophical literature: *physicalism* (though 'biologism' would be more appropriate in this context) and *functionalism*. The difference is that the functionalist says that consciousness is a role, whereas the physicalist says that consciousness is a physical or biological state that implements that role. Although functionalists are free to add restrictions, functionalism in its pure form is independent of implementation. Consciousness is defined as global accessibility, and although its human implementation depends on our biochemistry, silicon creatures without our biochemistry could implement the same computational relations. Thus functionalism and physicalism are incompatible doctrines since a nonbiological implementation of the functional organization of consciousness would be regarded as conscious by the functionalist but not by the physicalist. The big question for functionalists is: 'How do you know that it is broadcasting in the global workspace that makes a representation conscious as opposed to something about the *human biological realization* of that broadcasting?'

Many of us have had the experience of suddenly noticing a sound (say a jackhammer during an intense conversation), and at the same time realizing that the sound has been going on for some time even though one was not attending to it. If the subject did have a phenomenal state before the sound was noticed, that state was not broadcast in the global neuronal workspace until it was noticed. If this is right, there was a period of phenomenality without broadcasting. Of course, this is anecdotal evidence. But the starting point for work on consciousness is introspection and we would be foolish to ignore it.

In this section, we have seen a distinction between two concepts of consciousness, phenomenality and global accessibility. In the following section, we add a third, reflexivity.

4. *Phenomenality and Reflexivity*

Rosenthal (1997) defines reflexive consciousness as follows: S is a reflexively conscious state $\Leftrightarrow$ S is accompanied by a thought—arrived at noninferentially and nonobservationally— to the effect that one is in S. He offers this 'higher order thought' (HOT) theory as a theory of phenomenal consciousness. It is obvious that phenomenal consciousness without HOT and HOT without phenomenal consciousness are both conceptually possible. For examples, perhaps dogs and infants have phenomenally conscious pains without higher order thoughts about them. For the converse case, imagine that by biofeedback and imaging techniques of the distant future, I learn to detect the state in myself of having the Freudian unconscious thought that it would be nice to kill my father and marry my mother. I could come to know—noninferentially and nonobservationally—

that I have this Freudian thought even though the thought is not phenomenally conscious. Since there are conceptually possible counterexamples in both directions, the issue is the one discussed above of whether reflexivity and phenomenality come to the same thing in the brain.

If there are no actual counterexamples, the question arises of *why*. Is it supposed to be a basic law of nature that phenomenality and reflexivity co-occur? That would be a very adventurous claim. But if it is only a fact about us, then there must be a mechanism that explains the correlation, as the fact that both heat and electricity are carried by free electrons explains the correlation of electrical and thermal conductivity. But any mechanism breaks down under extreme conditions, as does the correlation of electrical and thermal conductivity at extremely high temperatures. So the correlation between phenomenality and reflexivity would break down too, showing that reflexivity does not yield the basic scientific nature of phenomenality.

The reader may wonder why Rosenthal includes 'noninferentially and nonobservationally' in the definition. The answer is: because one could find out that one is depressed by observing one's own behavior. This would not be conscious depression, but such a thought would be arrived at observationally and inferentially. However, this restriction is *ad hoc*, motivating the search for definition that is not *ad hoc*. Here is one: S is a reflexively conscious state $\Leftrightarrow$ S is phenomenally presented in a thought about S. This definition uses the notion of phenomenality, but this is no disadvantage unless one holds that there is no such thing apart from reflexivity itself.

We have seen three concepts of consciousness: phenomenality, reflexivity, and global accessibility. The Hard Problem arises only for phenomenality. The imaging work on consciousness engages phenomenality and accessibility. But many psychological experimental paradigms mainly engage reflexivity.

See also: Conscious and Unconscious Processes in Cognition; Consciousness and Sensation: Philosophical Aspects; Consciousness, Neural Basis of; Neglect, Neural Basis of; Reflexivity: Method and Evidence; Theory of Mind; Visual Systems: Dorsal and Ventral

Bibliography

Armstrong D M 1968 *A Materialist Theory of the Mind*. Routledge and Kegan Paul, London

Block N 1995 On a confusion about a function of consciousness. *Behavioral and Brain Sciences* **18**: 227–47

Block N, Flanagan O, Güzeldere G 1997 *The Nature of Consciousness: Philosophical Debates*. MIT Press, Cambridge, MA

Carruthers P 1992 Consciousness and concepts. *Proceedings of the Aristotelian Society*, Supplement **66**: 41–59

Chalmers D 1996 *The Conscious Mind*. Oxford University Press, New York

Churchland P 1989 Knowing qualia: A reply to Jackson. In: *A Neurocomputational Perspective*. MIT Press, Cambridge, MA, pp. 67–76

Dehaene S, Naccache L forthcoming Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*

Dennett D 1979 On the absence of phenomenology. In: Gustafson D, Tapscott B (eds.) *Body, Mind and Method: Essays in Honor of Virgil Aldrich*. Reidel, Dordrecht, The Netherlands

Dennett D 1991 *Consciousness Explained*. Little Brown, Boston

Dennett D 1994 Instead of qualia. In: Revonsuo A, Kamppinen M (eds.) *Consciousness in Philosophy and Cognitive Neuroscience*. Erlbaum, Hillsdale, NJ

Dretske F 1995 *Naturalizing the Mind*. MIT Press, Cambridge, MA

Flanagan O 1992 *Consciousness Reconsidered*. MIT Press, Cambridge, MA

Harman G 1990 The intrinsic quality of experience. In: Tomberlin J (ed.) *Philosophical Perspectives 4, Action Theory and Philosophy of Mind*. Ridgeview, Atascadero, CA, pp. 31–52

Jackson F 1977 *Perception*. Cambridge University Press, Cambridge, UK

Jackson F 1982 Epiphenomenal qualia. *Philosophical Studies* **32**: 127–36

Kanwisher N forthcoming Neural events and perceptual awareness. *Cognition*

Levine J 1983 Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–61

Lewis D 1994 Lewis, David: Reduction of mind. In: Guttenplan S (ed.) *Blackwell Companion to Philosophy of Mind*. Blackwell, Oxford, UK, pp. 412–31

Logothetis N K 1998 Single units and conscious vision. *Proceedings of the Royal Society of London Series B* **353**: 1801–18

Lycan W G 1990 Consciousness as internal monitoring. In: Tomberlin J (ed.) *Philosophical Perspectives 9*. Ridgeview, Atascadero, CA, pp. 1–14

Lycan W G 1996 *Consciousness and Experience*. MIT Press, Cambridge, MA

McDowell J 1994 The content of perceptual experience. *Philosophical Quarterly* April

McGinn C 1991 *The Problem of Consciousness*. Blackwell, Oxford, UK

Nagel T 1974 What is it like to be a bat? *Philosophical Review*

Nagel T forthcoming

Perry J forthcoming *Knowledge, Possibility and Consciousness*. Nicod Lectures, Paris, summer 1999

Pitcher G 1971 *A Theory of Perception*. Princeton University Press, Princeton, NJ

Rees G, Wojciulik E, Clarke K, Husain M, Frith C, Driver J in press Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. *Brain*

Rey G 1997 *Contemporary Philosophy of Mind*. Blackwell, Oxford, UK

Rosenthal D 1997 A theory of consciousness. In: Block N, Flanagan O, Güzeldere G (eds.) *The Nature of Consciousness: Philosophical Debates*. MIT Press, Cambridge, MA, pp. 729–54

Searle J 1992 *The Rediscovery of the Mind*. MIT Press, Cambridge, MA

Shoemaker S 1994 Self-knowledge and 'inner sense'; Lecture III: The phenomenal character of experience. *Philosophy and Phenomenological Research* **54**(2): 291–314

Sturgeon S 1994 The epistemic view of subjectivity. *Journal of Philosophy* **91**(5)
Tye M 1995 *Ten Problems of Consciousness*. MIT Press, Cambridge, MA
Tye M 1999 Phenomenal consciousness: The explanatory gap as a cognitive illusion. *Mind* **108**(432): 706–25
Van Gulick R 1993 Understanding the phenomenal mind: Are we all just armadillos? In: Davies M, Humphreys G (eds.) *Consciousness: Psychological and Philosophical Essays*. Blackwell, Oxford, UK

N. Block

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Consciousness, Neural Basis of

Consciousness is a puzzling state-dependent property of certain types of complex, adaptive systems. The best example of one type of such system is a healthy and attentive human brain. If the brain is anaesthetized, consciousness ceases. Small lesions in the brain stem and thalamus of patients can lead to a complete loss of consciousness, while destruction of circumscribed parts of the cerebral cortex of patients can eliminate very specific aspects of consciousness, such as the ability to be aware of motion or to recognize objects as faces, without a concomitant loss of vision in general. Given the similarity in brain structure and behavior, biologists commonly assume that at least some animals, in particular nonhuman primates, share certain aspects of consciousness with humans. Brain scientists, in conjunction with cognitive neuroscientists, are exploiting a number of empirical approaches that shed light on the neural basis of consciousness. Since it is not known to what extent artificial systems, such as computers and robots, can become conscious, this entry will exclude these from consideration.

1. *Some Common Neurobiological Assumptions*

By and large, neuroscientists have made a number of working assumptions that, in the fullness of time, need to be justified more fully.

(a) There is something to be explained; that is, the subjective content associated with a conscious sensation—what philosophers refer to as 'qualia'—does exist and has its physical basis in the brain. To what extent qualia and all other subjective aspects of consciousness can or cannot be explained within a reductionist framework remains highly controversial (Chalmers 1996).

(b) Consciousness is a vague term with many usages. It will, in the fullness of time, be replaced by a vocabulary that more accurately reflects the contribution of different brain processes (for a similar evolution, consider the usage of 'memory,' which has been replaced by an entire hierarchy of more specific concepts). Common to all forms of consciousness is that it feels like something (e.g., to 'see blue,' to 'experience a headache,' or to 'reflect upon a memory'); that is, it is usually about something. *Self-consciousness* is but one form of consciousness.

It is possible that all the different aspects of consciousness (smelling, pain, visual awareness, affect, self-consciousness, and so on) employ a basic common mechanism or perhaps a few such mechanisms. If so, then if one can understand the mechanism for one aspect, then one will have gone most of the way towards understanding them all.

(c) Consciousness is a property of the human brain, a highly evolved system. It therefore must have a useful *function* to perform. Crick and Koch (1998) assume that the function of the neuronal correlate of consciousness is to produce the best current interpretation of the environment—in the light of past experiences—and to make it available, for a sufficient time, to the parts of the brain which contemplate, plan, and execute voluntary motor outputs (including language). This needs to be contrasted with *on-line systems* that bypass consciousness but that can generate stereotyped behaviors (see below).

Note that in normally developed individuals, motor output is not necessary for consciousness to occur. This is demonstrated by *lock-in syndrome*, in which patients have lost (nearly) all ability to move yet are clearly conscious.

(d) At least some animal species posses some aspects of consciousness. In particular, this is assumed to be true for nonhuman primates such as the macaque monkey. Consciousness associated with sensory events in humans is likely to be related to sensory consciousness in monkeys for several reasons. First, trained monkeys show similar behavior to that of humans for many low-level perceptual tasks (e.g., detection and discrimination of visual motion or depth: Wandell 1995). Second, the gross neuroanatomies of humans and nonhuman primates are rather similar once the difference in size has been accounted for. Finally, functional magnetic resonance imaging of human cerebral cortex is confirming the existence of a functional organization in sensory cortical areas similar to that discovered by the use of single-cell electrophysiology in the monkey (Tootell et al. 1998). As a corollary, it follows that language is not necessary for consciousness to occur (although it greatly enriches human consciousness).

2. *Enabling, Modulating, and Specific Factors*

It is important to distinguish the general, *enabling* factors in the brain that are needed for any form of consciousness to occur from *modulating* ones that can up- or downregulate the level of arousal, attention, and awareness, and from the *specific* factors responsible for a particular content of consciousness.

An easy example of an enabling factor would be a proper blood supply. Inactivate the heart, and consciousness ceases within a fraction of a minute. This does not imply that the neural correlate of consciousness is in the heart (as Aristotle thought). A neuronal enabling factor for consciousness is the *intralaminar nuclei of the thalamus*. Acute bilateral loss of function in these small structures that are widely and reciprocally connected to the basal ganglia and cerebral cortex leads to immediate coma or profound disruption in arousal and consciousness (Bogen 1995).

Among the neuronal modulating factors are the various activities in nuclei in the brain stem and the midbrain, often collectively referred to as the *reticular activating system*, which control in a widespread and quite specific manner the level of noradrenaline, serotonin, and acetylcholine in the forebrain. Appropriate levels of these neurotransmitters are needed for sleep, arousal, attention, memory, and other functions critical to behavior and consciousness (Baars 1997).

Yet any particular content of consciousness is unlikely to arise from these structures, since they appear to lack the specificity necessary to mediate a sharp pain in the right molar, the percept of the deep, blue California sky, the bouquet associated with a rich Bordeaux, or a haunting musical melody. These must be caused by specific neural activity in cortex, thalamus, basal ganglia, and associated neuronal structures. The question motivating much of the current research into the neuronal basis of consciousness is the notion of the *minimal* neural activity that is *sufficient* to cause a specific conscious percept or memory (see below).

For instance, when a subject consciously perceives a face, the retinal ganglion cells whose axons make up the optic nerve that carries the visual information to the brain proper are firing in response to the visual stimulus. Yet it is unlikely that this retinal activity directly correlates with visual perception. While such activity is evidently necessary for seeing a physical stimulus in the world, retinal neurons by themselves do not give rise to consciousness.

Given the comparative ease with which the brains of animals can be probed and manipulated, it seems opportune at this point in time to concentrate on the neural basis of sensory consciousness. Because primates are highly visual animals and much is known about the neuroanatomy, psychology, and computational principles underlying visual perception, vision has proven to be the most popular model system in the brain sciences.

3. *Information Processing in the Brain that Bypasses Consciousness*

Cognitive and clinical research demonstrates that much complex information processing can occur without involving consciousness. This includes visual, auditory, and linguistic priming, implicit memory, the implicit recognition of complex sequences, automatic behaviors such as driving a car or riding a bicycle, and so on (Velmans 1991), and the dissociations found in patients with lesions in cerebral cortex (e.g., such as residual visual functions in the professed absence of any visual awareness known as *clinical blindsight* in patients with lesions in primary visual cortex: Weiskrantz 1997).

The cognitive scientist Jackendoff (1987) argues at length against the notion that consciousness and thoughts are inseparable and that introspection can reveal the contents of the mind. According to Jackendoff, what is conscious about thoughts are sensory aspects, such as visual images, sounds, or silent speech. Both the process of thought and its content are not directly accessible to consciousness. Indeed, one tradition in psychology and psychoanalysis—going back to Sigmund Freud—hypothesizes that higher-level decision making and creativity are not accessible at a conscious level, although they influence behavior.

Within the visual modality, Milner and Goodale (1995) have made a masterful case for the existence of so-called *on-line systems* that bypass consciousness. Their function is to mediate relative stereotype visuomotor behaviors, such as eye and arm movements, reaching, grasping, posture adjustments, and so on, in a very rapid, reflex-like manner. On-line systems work in egocentric coordinate systems, and lack certain types of perceptual illusions (e.g., size illusion) as well as possessing no direct access to working memory. This contrasts well with the function of consciousness alluded to above, namely to synthesize information from many different sources and use it to plan behavioral patterns over time. Milner and Goodale argue that on-line systems are associated with the *dorsal stream* of visual information in the cerebral cortex, originating in the primary visual cortex (V1) and terminating in the posterior parietal cortex.

4. *The Neuronal Correlate of Consciousness (NCC)*

The problem of consciousness can be broken down into several separate questions. Most, if not all of these, can then be subjected to scientific inquiry.

The major question that neuroscience must ultimately answer can be bluntly stated as follows: it is probable that at any moment some active neuronal processes in our head correlate with consciousness, while others do not; *what is the difference between them?* The specific processes that correlate with the current content of consciousness are referred to as the *neuronal correlate of consciousness*, or as the NCC. Whenever some information is represented in the

NCC it is represented in consciousness. The NCC is the minimal (minimal, since it is known that the entire brain is sufficient to give rise to consciousness) set of neurons, most likely distributed throughout certain cortical and subcortical areas, whose firing directly correlates with the perception of the subject at the time. Conversely, stimulating these neurons in the right manner with some yet unheard of technology should give rise to the same perception as before. Discovering the NCC and its properties will mark a major milestone in any scientific theory of consciousness.

What is the character of the NCC? Most popular has been the belief that consciousness arises as an emergent property of a very large collection of interacting neurons (for instance, Libet 1993) possibly in connection with exceeding some level of complexity (Edelman and Tononi 2000). In this view, it would be foolish to locate consciousness at the level of individual neurons. An alternative hypothesis is that there are special sets of 'consciousness' neurons distributed throughout cortex and associated systems. Such neurons represent the ultimate neuronal correlate of consciousness, in the sense that the relevant activity of an appropriate subset of them is both necessary and sufficient to give rise to an appropriate conscious experience or percept (Crick and Koch 1998). Generating the appropriate activity in these neurons, for instance by suitable electrical stimulation, would give rise to the specific percept.

Any one subtype of NCC neurons would, most likely, be characterized by a unique combination of molecular, biophysical, pharmacological, and anatomical traits. It is possible, of course, that all cortical neurons may be capable of participating in the representation of one percept or another, though not necessarily doing so for all percepts. The secret of consciousness would then be the type of activity of a temporary subset of them, consisting of all those cortical neurons which represent that particular percept at that moment. How activity of neurons across a multitude of brain areas that encode all of the different aspects associated with an object (e.g., the color of the face, its facial expression, its gender and identity, the sound issuing from its mouth) is combined into a single percept remains puzzling and is known as the *binding problem*.

What, if anything, can we infer about the location of neurons whose activity correlates with consciousness? In the case of visual consciousness, it was surmised that these neurons must have access to visual information, and project to the planning stages of the brain; that is, to premotor and frontal areas (Fuster 1997). Since no neurons in the primary visual cortex of the macaque monkey project to any area forward of the central sulcus, Crick and Koch (1998) propose that neurons in V1 do not give rise to consciousness (although V1 is necessary for most forms of vision, just as the retina is). Ongoing electrophysiological, psychophysical, and imaging research in monkeys and humans is evaluating this prediction.

While the set of neurons that can express any one particular conscious percept might constitute a relative small fraction of all neurons in any one area, many more neurons might be necessary to support the firing activity leading up to the NCC. This might resolve the apparent paradox between clinical lesioning data suggesting that small and discrete lesions in cortex can lead to very specific deficits (such as the inability to see colors or to recognize faces in the absence of other visual losses) and the functional imaging data that any one visual stimulus can activate large swaths of cortex.

Conceptually, several other questions need to be answered about the NCC. What type of activity corresponds to the NCC (it has been proposed as long ago as the early part of the twentieth century that spiking activity synchronized across a population of neurons is a necessary condition for consciousness to occur)? What causes the NCC to occur? And, finally, what effect does the NCC have on postsynaptic structures, including motor output.

5. *Experimental Approaches*

A promising experimental approach to locate the NCC is the use of *bistable percepts* in which a constant retinal stimulus gives rise to two percepts alternating in time, as in a Necker cube (Logothetis 1998). One version of this is *binocular rivalry*, in which a small image, say of a horizontal grating, is presented to the left eye and another image, say a vertical grating, is shown to the corresponding location in the right eye. In spite of the constant visual stimulus, observers 'see' the horizontal grating alternate every few seconds with the vertical one (Blake 1989). The brain does not allow for the simultaneous perception of both images.

It is possible, although difficult, to train a macaque monkey to report whether it is currently seeing the left or the right image. The distribution of the switching times and the way in which changing the contrast in one eye affects this leave little doubt that monkeys and humans experience the same basic phenomenon. In a series of elegant experiments, Logothetis and colleagues (Logothetis 1998) recorded from a variety of visual cortical areas in the awake macaque monkey while the animal performed a binocular rivalry task. In primary visual cortex, only a small fraction of cells modulate their response as a function of the percept of the monkey, while 20–40 percent of neurons in higher visual areas in cortex do so. The majority of cells increased their firing rate in response to one or the other retinal stimulus with little regard to what the animal perceived at the time. In contrast, in a high-level cortical area such as the inferior temporal (IT) cortex, almost all neurons responded only to the

perceptual dominant stimulus (in other words, a 'face' cell only fired when the animal indicated by its performance that it saw the face and not the pattern presented to the other eye). This makes it likely that the NCC involves activity in neurons in the inferior temporal lobe. Lesions in the homologue area in the human brain are known to cause very specific deficits in conscious face or object recognition. However, it is possible that specific interactions between IT cells and neurons in parts of the prefrontal cortex are necessary in order for the NCC to be generated.

Functional brain imaging in humans undergoing binocular rivalry has revealed that areas in the right prefrontal cortex are active during the perceptual switch from one percept to the other (Lumer et al. 1998).

A number of alternate experimental paradigms are being investigated using electrophysiological recordings of individual neurons in behaving animals and human patients, combined with functional brain imaging. Common to these is the manipulation of the complex and changing relationship between physical stimulus and the conscious percept. For instance, when an image is briefly flashed onto the screen and is immediately followed by presentation of a second image, the first image often remains invisible (it is said to be *masked*). Yet neurons can still respond in a selective manner to the first stimulus that is not consciously registered. Under other conditions, subjects will respond to a target although it is physically not present. The NCC in the appropriate sensory area should mirror the perceptual report under these dissociated conditions. Perceptual report can be influenced by delivering current pulses to the cerebral cortex in the absence of any physical stimulus or in the presence of an ambiguous stimulus, as explored in the context of elective surgery for epileptic patients (Penfield and Perot 1963) or in animal studies (Parker and Newsome 1998). Visual illusions constitute another rich source of experiments that can provide information concerning the neurons underlying these illusory percepts. A classical example is the *motion after effect*, in which a subject stares at a constantly moving stimulus (such as a waterfall) for a fraction of a minute or longer. Immediately after this conditioning period, a stationary stimulus will appear to move in the opposite direction. Because of the conscious experience of motion, one would expect the subject's cortical motion areas to be activated in the absence of any moving stimulus. Another approach, suitable for the establishment of a rodent model to study the NCC, relies on the differential involvement of consciousness that appears necessary to establish associative trace conditioning, but not for associative delay conditioning (Clark and Squire 1998). That is, subjects need to be aware of the temporal relationship between conditioning stimulus (CS) and unconditioned stimulus (US) if there is a delay between the end of the CS and the onset of the US, while this is not the case if the time course of the two overlap. Finally, understanding the specific actions of the different classes of anesthetic agents on cortical, thalamic, and basal ganglia neural networks will aid both the development of systems-level theories of anesthesia and the search for the NCC. Future techniques, most likely based on the molecular identification and manipulation of discrete and identifiable subpopulations of cortical and thalamic cells in appropriate animals, will greatly help in this endeavor.

Identifying the type of activity and the type of neurons that give rise to specific conscious percept in animals and humans would only be the first, albeit critical, step in understanding consciousness. One also needs to know where these cells project to, their postsynaptic action, how they develop in early childhood, what happens to them in mental diseases known to affect consciousness in patients, such as schizophrenia or autism, and so on. And, of course, a final theory of consciousness would have to explain the central mystery: why a physical system with a particular architecture gives rise to feelings and qualia (Chalmers 1995).

See also: Consciousness and Sensation: Philosophical Aspects; Consciousness, Cognitive Psychology of; Free Will and Action; Mind–Body Dualism

Bibliography

Baars B J 1997 *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, Oxford, UK

Blake R 1989 A neural theory of binocular rivalry. *Psychological Review* **96**: 145–67

Bogen J E 1995 On the neurophysiology of consciousness: I. An overview. *Consciousness and Cognition* **4**: 52–62

Chalmers D J 1996 *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York

Clark R E, Squire L R 1998 Classical conditioning and brain systems: the role of awareness. *Science* **280**: 77–81

Crick F, Koch C 1998 Consciousness and neuroscience. *Cerebral Cortex* **8**: 97–107

Edelman G M, Tomoni G 2000 *A Universe of Consciousness*. Basic Books, New York

Fuster J M 1997 *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe*, 3rd edn. Lippincott-Raven, Philadelphia, PA

Jackendoff R 1987 *Consciousness and the Computational Mind*. MIT Press, Cambridge, MA

Libet B 1993 *Neurophysiology of Consciousness: Selected Papers and New Essays*. Birkhauser, Boston

Logothetis N K 1998 Single units and conscious vision. *Philosophical Transactions of the Royal Society of London, Series B* **353**: 1801–18

Lumer E, Friston K J, Rees G 1998 Neural basis of perceptual rivalry in the human brain. *European Journal of Neuroscience* **10**(suppl): 331

Milner A D, Goodale M A 1995 *The Visual Brain in Action*. Oxford University Press, Oxford, UK

Parker A J, Newsome W T 1998 Sense and the single neuron: Probing the physiology of perception. *Annual Review of Neuroscience* **21**: 227–77
Penfield W, Perot P 1963 The brain's record of auditory and visual experience: A final summary and discussion. *Brain* **86**: 595–696
Tootell R B H, Hadjikhani N K, Mendola J D, Marrett S, Dale A M 1998 From retinotopy to recognition: fMRI in human visual cortex. *Trends in Cognitive Sciences* **2**: 174–83
Velmans M 1991 Is human information processing conscious? *Behavioral Brain Science* **14**: 651–68
Wandell B A 1995 *Foundations of Vision*. Sinauer Associates, Sunderland, MA
Weiskrantz L 1997 *Consciousness Lost and Found*. Oxford University Press, Oxford, UK

C. Koch and F. Crick

Conscription, Policy of

Conscription is the compulsory enlistment of people into military service. In modern democratic states, conscription is governed by a set of rules regulating eligibility, length of service, pay, and benefits. Conscription may involve universal military service, but it need not. The terms 'to conscript' and 'to draft' have become virtually synonymous.

1. Scholarship on Conscription

Conscription is one of the most important policies available to states. The very essence of the state, at least according to Max Weber, is its power to monopolize coercive power within a given territory, and this power depends on the state's control of an army, which in turn depends on the ability of government to recruit soldiers. Nonetheless, scholars concerned with developing a more adequate theory of the state have given relatively scant attention to military obligations and the institutions that enforce these obligations. Military service is demonstrably as important an aspect of the state–citizen relationship as any that exists. A government's decision to conscript and its policy implementing conscription have significant political and social implications for state building and for citizenship.

There has not, of course, been total neglect. Anthony Giddens, Michael Mann, Pierre Birnbaum, and other political sociologists have, on occasion, addressed the significance of military service and, more specifically, conscription for state building. Specialists in international relations and international political economy have certainly considered the importance of national defense and military power, but the state theorists among them have, on the whole, focused on civil vs. military domestic power, international cooperation, and trade. Military theorists and historians, including such luminaries as Thucydides, Niccolo Machiavelli, Carl von Clausewitz, and John Mahon and a host of more contemporary scholars, have discussed the political and social importance of military format as well as the role of nationalism and voluntary compliance in increasing acquiescence with conscription. However, to the extent there has been any systematic attention to the link between conscription and state building or citizenship, it has generally come from historically oriented political scientists, sociologists, and economists.

In addition to the investigations of conscription in state theory, there are two additional traditions of conscription scholarship. One focuses on contemporary policy issues: what role conscription should play in the recruitment of modern armies. The other deals with conscription data as a source of information about the population over time.

2. Conscription and State Theory

There are at least three strands to this literature. The first emphasizes the relationship between war, nation building, and state making. Here the major issues are the development of state capacity and the form of governance. The second emphasizes the relationship between conscription and citizenship. It, too, is a literature focused on the changing nature of the state, but the focus is on transformation of subjects into citizens. The third stresses the bases for and history of objection to conscription.

2.1 Nation and State Building

Otto Hintze (1994 [1906]), Samuel Finer (1975), Charles Tilly (1990), and Brian Downing (1991) capture the important role that military activity and institutions generally play in forging the infrastructure of governments. To wage war requires economic resources, including the ability to tax subjects for money and men. The way in which war is waged also has consequences for the form and power of the state. Joseph Schumpeter's (1954 [1918]) initial insight, while obvious, remains important: war permits increased government expenditure, which is sustained in peacetime, and increased government expenditure implies increased taxation. Moreover, the emergence of constitutional democracies or authoritarian regimes can be affected by the processes of extracting resources during wartime.

Conscription played a crucial role historically in the development of state capacity, particularly since the French and American revolutions. Throughout the

eighteenth century (and into the nineteenth), a single shilling pressed into the hand of a drunken man in a public house by a recruiting sergeant enlisted a soldier into the British army. In France, the dreaded *Milice royale* relied on a lottery to choose which peasants would be forced into the King's service. The army was not a popular institution, at least not for those who had to serve in its rank and file. Some joined because they liked the life, but most did so because they were coerced or needed the work or were fleeing from something worse. Those who refused did so because being in the army was inconvenient or actively repugnant. National patriotism is a relatively modern basis for service. It requires a sense of one's country as an entity that can legitimately demand contributions in the form of conscripts and taxes.

The use of coercion to compel young men to fight on behalf of a chief, a lord, or a king has long been a feature of human society. Modern conscription, however, emerged in the late eighteenth century. According to the OECD, the origin of the term conscription was in 1798 during the French Republic. At that time and place, conscription referred to a very particular process of military enrollment in which lots were drawn among the young men deemed eligible for service and which permitted those whose names were drawn to provide substitutes.

In the nineteenth century the countries of Europe and North America developed modern national states, and those in the Antipodes began the process. State building involved the creation of a military adequate to the requirements of relatively centralized government and the exigencies of international politics. The mass mobilization of the French Revolution and the Napoleonic Wars marked the beginning of the period in Europe; the American Revolution and the Civil War were critical moments for the United States. The governments of this period confronted the problems of turning feudal armies and local militias into national armies. A series of new institutions emerged to make orderly conscription possible: a census of eligible young men, rules of selection and eligibility, health examiners, and a bureaucratic apparatus capable of assigning, paying, and housing the recruits.

Historians have documented the creation of these new institutions and the problems they created for both legislators and the new bureaucrats (e.g., Chambers 1987, Schnapper 1968, Simkins 1988). Resistance to the census and to conscription was strong among those trying to avoid the reach of a centralizing state. Young men hid from the registrars, and they cut off fingers and toes to make themselves ineligible (e.g., Aron et al. 1972, Le Roy Ladurie 1979 [1973], Levi 1997). The long-term effect of the forms of conscription introduced in the nineteenth and early twentieth centuries, however, was the construction of more centralized states, whose extractive reach goes deep into the population.

Conscription can contribute to nation building as well as state building. The French army, for example, played a key role in transforming a population drawn from diverse regional and linguistic backgrounds into a single people. The British Army, on the other hand, seems to have reinforced ethnic divides in both the British Isles and its colonies. Today the armies of the United States, South Africa, and other countries with historically disenfranchized minorities are important institutions for the upward mobility of precisely those minorities once prevented from being soldiers. Investigations of other polities, including more recently developed countries, suggest reasons for variation in ethnic integration that range from the demand for troops to the differences between a professional and politicized military (e.g., Enloe 1980, Peled 1998).

2.2 Conscription and Citizenship

The institutions of conscription extend the state's social control of its people, but they can also lay the basis for the extension of citizenship. The exploration of the citizen-soldier has been the subject of considerable discussion, most notably in the work of Morris Janowitz, Samuel Stouffer, Charles Moskos, Elliot Cohen, and David Segal, but that literature has largely investigated civil–military relations and the motivations of soldiers to fight. Of more interest here is the ways in which the development of conscription interacts with the construction of the terms of citizenship.

Within the context of modern nationalism, someone who is willing to die for one's country has a claim on recognition as a full citizen of that country. Historically, conscription is often linked with demands for the franchise, social services, and other rights of citizenship. There is considerable evidence that ethnic minorities who have been exempted from conscription will become eager volunteers in order to demonstrate their loyalty and thus earn the privilege of eligibility for conscription. This may explain why so many blacks and Nisei Japanese enthusiastically signed up for United States Army during World War II, why some Druse are so eager to participate in the Israeli army, and why other notable ethnic minorities have also enlisted in their country's armies (Petersen 1989).

In France, Switzerland, several other continental European countries, and for many Israelis, citizenship, at least male citizenship, is closely tied to one's service as a conscript. There is an expectation of some form of conscription, even during peacetime. It is a rite of passage to have served in the military, and it is often a form of bonding for the cohort that drills together. The Anglo-Saxon democracies of Britain, the United States, Canada, Australia, and New Zealand have a quite different tradition. Until World War I, conscription could be imposed only during war and, moreover, only during a war of defense—although the

definition of 'war of defense' varies among these countries and among groups within these countries. Nearly all have had some extended period of conscription in the absence of major military conflagrations, but the draft remains controversial.

Conscription by democratic governments seems to succeed to the extent the policymakers design military service formats that, first, ensure relative equality of sacrifice and, second, provide a means for popular discussion and approval of the war and conscription policy. More often than not these more democratic moves are accomplished not through foresight but as revision of failed policies. With few exceptions, policymakers first introduced conscription that provided discriminatory exemptions to powerful constituents and laid nearly the whole burden of warfare on those without the franchise or other form of political clout. Such policies changed only in response to legislative conflict, political protests, and actual draft riots, such as those in New York in 1863 or in Toulouse in 1868. These acts compelled policymakers to revise their beliefs about the acceptability of permitting some part of the population to literally buy their way out of military service.

The high degree of citizen support necessary for conscription hinges on the perception of an acceptable policy bargain whose terms government actors are likely to uphold. The minimal terms of the democratic conscription bargain are that government will conscript according to some legislated and relatively equitable formula.

The history of conscription in France throughout the nineteenth and early twentieth century is one of ongoing negotiations between citizens and state officials over what makes for equitable rules governing conscription. Conscription was by lottery; a set number was chosen from the pool of eligible young men and constituted the cohort of that year. The registrars only went down the list if some of those chosen did not pass the medical examination. Throughout most of the nineteenth century, hiring a substitute was an acceptable practice. Indeed, many of those who received a 'bad' number in the lottery found substitution Pareto optimal. The unlucky recruit was generally covered by an insurance scheme his parents had bought and which provided the resources to pay the bonus or premium substitutes demanded. No one became newly eligible as a result, and young men accepted the substitution contract only if they found it sufficiently attractive—they were not compelled to serve. However, as France increasingly democratized and as the concept of equity altered in consequence, substitution was increasingly considered undemocratic and unfair.

In contrast, voluntary citizen soldiers had been the mark of the military format of the Anglo-Saxon democracies. The institution of conscription in Britain, the United States, Canada, and New Zealand during World War I represented a redesign of the democratic state and a redefinition of the obligations of citizenship. The Great War demonstrated the problems with voluntarism. It failed to produce enough enlistees or permit efficient mobilization of their labor. It led to unattractive and often inappropriate forms of moral suasion. It created bureaucratic mechanisms, such as the national registers, that may have discouraged as many volunteers as they encouraged. By publicizing the shortages in manpower, these censuses made evident the numbers who might be considered 'shirkers.' Such knowledge could make those who had enlisted as well as those who had not think twice about volunteering. If there were not to be enough men anyway, why make the sacrifice?

As states became more democratic, the relationship between conscription and citizenship changed. Conscription remained a basis for the demand for the vote and for social benefits, but it also came under more scrutiny as an obligation governments could impose, even in wartime. With increasing democratization came increasing questions of citizen approval of the wars being fought and the nature of the conscription system being imposed.

2.3 Objections to Conscription

Conscription solves manpower allocation problems during mass mobilization wars; it is a government tool for determining who fights and who works in essential industries. The institution of conscription creates as well as solves problems for democratic governments, however. There is one school of authors who object to conscription altogether as a form of unacceptable government compulsion, which abridges civil liberties and citizen choices and makes the state a Leviathan power. Others argue that conscription may be acceptable, but only during wars of defense recognized as such by the population. Finally, there are those who argue for exemptions from conscription based on conscientious objection; they make the case that individual rights here exceed the demands of universalism.

A major argument for conscription was—and is—its equity. Within democracies there develops a new standard by which to evaluate the implementation of conscription: equality of sacrifice. To the extent governments permit exemptions on nonuniversalistic grounds, they will experience an increase in refusal to consent. The pressure to offer exemptions is particularly strong during wars, such as Vietnam, that do not require total mobilization. Moreover, the fact that decisions about exemptions were made by decentralized draft boards in the United States, Britain, Australia, and elsewhere led to differential treatment of otherwise similar cases, thus confirming the sense of a discriminatory and inequitable process (e.g., Davis and Dolbeare 1968).

A related problem for democracies is the imposition of conscription without the approval of some of those it affects or despite the opposition of certain groups. There are numerous studies of resistance by particular ethnic, religious, racial, and class-based groups (see, e.g., Granatstein and Hitsman 1977, Karsten 1983, McKernan 1980). This, in the long run, is a problematic strategy for democracies, as the cases of the Irish in Britain, the francophones in Canada, the aboriginals in Australia, Canada, the United States, and South Africa, the blacks in the United States, among others, have demonstrated. Individuals who have lost or never possessed confidence that the playing field is level may be momentarily compliant due to coercion, but they are not in any sense consenting.

Principled refusal to comply with conscription dates at least from the early American militias. There are two primary grounds for refusal: religious and political. Religious objection to conscription, or what is generally called conscientious objection, requires membership of either a pacifist religion (Quakers, Mennonites) or one that denies the authority of the state over its members (Jehovah's Witnesses, certain sects of Judaism). Conscientious objection is thus most prevalent where dissenting religions thrive, as in the American colonies. Politically-based objection depends on a belief that there are reciprocal obligations between citizens and their governments and that consent can be withdrawn when the government has failed to meet its obligations. This kind of objection is most likely to arise in response to popularly contested wars, such as Algeria or Vietnam. Such factors make conscription and various forms of objection to conscription popular topics of discussion throughout the liberal democracies among politicians and the general public. There is, in addition, a significant philosophical, historical, and public policy scholarship (see, e.g. Simmons 1979, Klosko 1992, Sibley and Jacob 1962, Moskos and Chambers 1993, Levi 1997).

3. *Using Conscription Data*

For economic historians and demographers, there is an additional use for conscription. It is an important source of data. The registration and examination of conscripts provides information about the population over time as well as about the relationships between recruits and the state. It is particularly useful information for learning about height and health over history among different regions and classes (Aron et al. 1972, Floud et al. 1990, Mokyr and Grada 1996). However, it is also useful for inferring the relationship of different ethnic or religious groups to the state (Hanham 1973, Spiers 1980, Levi 1997).

What conscription data is not particularly useful for, however, is comparing the behavior of recruits across wars and countries. Most of the interesting studies are based on snapshots of a particular period in a specific place or rest very lightly on quantitative data. Despite the significant amount of data collected by armies, seldom is it collected the same way across time and or across countries or aggregated in a way useful to social scientists. Moreover, much information has been lost or remains restricted.

4. *Conscription and the Modern Army*

France epitomizes a country in which conscription and citizenship are nearly synonymous. Conscription has been part of the French political landscape since the French Revolution, except for one brief interlude. There have been significant policy shifts. Substitution and commutation were eliminated, and lengths of service and exemptions changed. In 1962, the French finally permitted conscientious objection, subsequently offering a variety of means of completing one's national service, including teaching in Africa, building *bidonville* in the South of France, and even working in a consul or bank outside of France. In the late 1990s, the greatest change of all occurred: the elimination of conscription altogether. This is largely for military reasons; the army is finding it costly to work with short-term recruits and prefers a highly trained, technologically sophisticated, and fully professional army. Belgium and The Netherlands abolished conscription prior to France, in the early 1990s, and for similar reasons. South Korea is another example from a very different continent and culture of the recent shift from a mass and conscripted army to a small and professional army.

The elimination of conscription in many contemporary states has raised a series of questions about the appropriate form of military service in democracies. One set of issues, played out largely in the pages of *Armed Forces and Society*, are the political and social implications for the civil–military relationship. Some scholars fear that increased reliance on a military elite and reduced dependence on the citizen soldier may dangerously enhance the power of the military. Others believe that the institutions of civilian government, at least in the advanced liberal democracies, will be sufficient as a countervailing power.

A related development is reflected in the all-volunteer armies of the United States and Britain. There are two issues here. The first, discussed primarily by economists and military sociologists, is the choice between conscription and volunteerism as the most efficient means for supplying manpower. Sophisticated econometric models (e.g., Ross 1994) are often developed to disentangle the costs and benefits of the different policy options. A second set of issues reflects the concern that the democracies are returning to armies unrepresentative of the population, armies composed disproportionately of the poor, the uneducated, and the minorities. There are advantages, of

course. By relying fundamentally on volunteers, more armies than the past are beginning the process of integrating women without facing issues of conscripting women, and they are enhancing the citizenship status of homosexuals as well as ethnic, religious, and racial minorities. The cost of this strategy may be high, however. Even when the army provides a means of upward advancement to such groups, participation in its ranks no longer constitutes an obligation of general citizenship and the distance between citizen and army increases.

There is no question that conscription, closely linked to citizenship in the liberal democracies during most of the nineteenth and twentieth centuries, is no longer part of the national consciousness in these countries. It is, however, a fact of life—and death—for the peoples of countries embroiled in some of the worst conflicts of the late twentieth century. The conscription of children into the armies fighting the racial and tribal conflicts in sub-Saharan Africa is a nightmare reversion to the impressment of the European past.

Conscription appears to have come nearly full circle in approximately 200 years. Modern, legitimate, and rule-bound conscription commenced with the French Republic, and it was part of the process of transforming subjects into democratic citizens. By the late twentieth century, the liberal democracies are returning to mercenary and professional armies, although severely altered from their eighteenth century forms. Countries without a tradition of democracy and without strong government institutions are using conscription solely as a tool of social control and particularistic power.

See also: Military and Politics; Military History; Military Sociology; Nation: Sociological Aspects; State Formation; State, History of; State, Sociology of the; War, Sociology of

Bibliography

Aron IJ-P, Dumont P, Le Roy Ladurie E 1972 *Anthropologie du conscrit français: D'après les comptes numériques et sommaires du recrutement de l'armée (1819–1826)*. Mouton, Paris

Chambers J W II 1987 *To Raise an Army: The Draft Comes to Modern America*. The Free Press, New York

Davis Jr J W, Dolbeare K M 1968 *Little Groups of Neighbors: The Selective Service System*. Markham Pub. Co., Chicago

Downing B M 1991 *The Military Revolution and Political Change: Origins of Democracy and Autocracy in Early Modern Europe*. Princeton University Press, Princeton, NJ

Enloe C H 1980 *Police, Military and Ethnicity: Foundations of State Power*. Transaction Books, New Brunswick, NJ

Finer S 1975 State and nation-building in Europe: The role of the military. In: Tilly C (ed.) *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ

Floud R, Gregory A, Wachter K 1990 *Height, Health, and History: Nutritional Status in the United Kingdom, 1750–1980*. Cambridge University Press, New York

Granatstein J L, Hitsman J M 1977 *Broken Promises: A History of Conscription in Canada*. Oxford University Press, Toronto, ON

Hanham H J 1973 Religion and nationality in mid-Victorian England. In: Foot M R D (ed.) *War and Society: Historical Essays in Honour and Memory of J. R Western 1928/1971*. Paul Elek, London, pp. 159–82

Hintze O 1994 [1906] Military organization and the organization of the state. In: Hall J A (ed.) *The State: Critical Concepts*. Routledge, London, Vol. 1, pp. 181–202

Karsten P 1983 Irish soldiers and the British Army, 1792–1922: Suborned or subordinate. *Journal of Social History* **17**(1): 31–64

Kiernan V G 1973 Conscription and society in Europe before the war of 1914–18. In: Foot M R D (ed.) *War and Society: Historical Essays in Honour and Memory of J. R. Western 1925/1971*. Paul Elek, London, pp. 141–58

Klosko G 1992 *The Principle of Fairness and Political Obligation*. Rowman & Littlefield Publishers, Lanham, MD

Le Roy Ladurie E 1979 [1973] *The Territory of the Historian*, trans. Reynolds B, Reynolds S. University of Chicago Press, Chicago

Levi M 1997 *Consent, Dissent and Patriotism*. Cambridge University Press, New York

McKernan M 1980 *The Australian People and the Great War*. Nelson, Melbourne, Vic, Australia

Mokyr J, Grada C O 1996 Height and health in the United Kingdom 1815–1860: Evidence from the East India Company Army. *Explorations in Economic History* **33**(2): 141–68

Moskos C C 1985 Citizen soldier and National Service. In: Suttles G D, Zald M N (eds.) *The Challenge of Social Control: Citizenship and Institution Building in Modern Society: Essays in Honor of Morris Janowitz*. Ablex Publishing Company, Norwood, NJ, pp. 149–60

Moskos C C, Chambers W J II (eds.) 1993 *The New Conscientious Objection: From Sacred to Secular Resistance*. Oxford University Press, New York

Peled A 1998 *A Question of Loyalty: Military Manpower Policy in Multiethnic States*. Cornell University Press, Ithaca, NY

Petersen R 1989 Rationality, ethnicity and military enlistment. *Social Science Information* **28**(3): 563–98

Ross T W 1994 Raising an army: A positive theory of military recruitment. *Journal of Law and Economics* **38**(April): 109–31

Schnapper B 1968 *Le remplacement militaire en France: Quelques aspects politiques, économiques et sociaux du recrutement au XIX siècle*. S.E.V.P.E.N, Paris

Schumpeter J 1991 [1918] The crisis of the tax state. In: Swezburg R (ed.) *The Economics and Sociology of Capitalism*. Princeton University Press, Princeton, NJ pp. 99–140

Sibley M Q, Jacob P E 1962 *Conscription of Conscience: The American Conscientious Objector, 1940–1947*. Cornell University Press, Ithaca, NY

Simkins P 1988 *Kitchener's Army: The Raising of the New Armies, 1914–1916*. Manchester University Press, Manchester, UK

Simmons A J 1979 *Moral Principles and Political Obligations*. Princeton University Press, Princeton, NJ

Spiers E M 1980 *The Army and Society, 1815–1914*. Longman, London

Tilly C 1990 *Coercion, Capital, and European States, AD 990–1990*. Basil Blackwell, Cambridge, MA

M. Levi

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Consensus Panels: Methodology

Should you avoid electromagnetic fields produced by home appliances and overhead transmission wires, or secondhand smoke from cigars and cigarettes because they may increase your risk of cancer? These are just two recent examples of the recurrent 'menace[s] of daily life' that have been resolved by consensus panels. A consensus is a shared belief that results from a group judgment. A consensus panel is a group created for the purpose of forming or reaching a consensus. The demand for formal consensus methods has greatly increased in recent decades as the complexity of the decisions, their importance, and the amount of conflicting information involved have also increased. Four prominent, widely used consensus panel methods are described.

1. *Important Consensus Panel Characteristics*

The two most critical aspects of consensus are: (a) how widely shared is the belief, and (b) how is the panel selected? With respect to the first question, while consensus is often thought of as a belief shared by all panel members (i.e., belief unanimity), it also can connote a majority opinion or even, in practice, an average of the group's belief. The answer depends on the procedure used to reach a consensus or the source of the panel's decision.

1.1 *Source of Panel Decision*

There are two sources or ways that a consensus decision can be achieved that result in either a majority or unanimous consensus. The first, and most commonly associated with the consensus process, is that the decision is internal to the panel arising from discussion and sharing of information. In such cases an *ad hoc* panel of experts could be convened to make a unanimous recommendation. The decision to administer the swine flu vaccine to the US public resulted from such a qualitative method (Neustadt and Fineberg 1978).

Since experts with divergent beliefs often disagree, an internal consensus is often impossible. In that case, a second source of decision-making uses a number of quantitative methods to derive a majority or average opinion external to the panel or entirely independent of it. The most common of these is the Delphi method (Dalkey 1969) developed at the RAND Corporation that was actually employed in parallel with the swine flu decision (Schoenbaum et al. 1976). A more recent quantitative method known as meta-analysis (Smith and Glass 1977) or research synthesis (Cooper and Hedges 1994) is widely used to develop an average consensus of the empirical, scientific literature produced by experts.

1.2 *Panel Composition*

Consensus panels are composed of experts or nonexperts. Expert or 'blue-ribbon' panels are most commonly employed since such groups presumably not only have the requisite knowledge to reach a correct consensus, but also the credibility required to convince others of the validity of their opinion.

Since selection of experts can easily result in bias (see discussion of *ad hoc* panels, nonexperts who are neutral on the issue have commonly been convened as a jury to evaluate the existing evidence and reach a decision or recommendation. This method has been used extensively by the US National Institutes of Health (NIH) to assess the effectiveness of medical innovations (Perry and Kalberer 1980). While selection bias is not a problem for the panel's composition, it is still possible that it could affect the information being considered.

2. *Consensus Panel Methods*

The four methods in Table 1 illustrate the contemporary methods for reaching group consensus that embody the characteristics described in the previous section. Experts and nonexperts are both directly involved in internal decisions while, by using quantitative techniques, they can indirectly produce a consensus externally. Both approaches have proven useful in different circumstances, and both have strengths and weaknesses.

2.1 Ad Hoc *Panel Method*

The use of *ad hoc* panels or groups is a widely-employed organizational method for consensus decision-making. Such panels are composed of experts convened in a face-to-face meeting to make a specific decision. The method can be either unstructured or structured. In an unstructured procedure, a nonrepresentative (or nonprobability) convenience sample of experts is gathered for one session to address a single, global question such as, 'Should the swine flu vaccine be produced and distributed?' Such convenience panels are often neither representative of the range of beliefs nor independent of the agency or organization requesting the consensus decision (Neustadt and Fineberg 1978).

In a structured procedure, an effort is made to create a panel containing members representing all the views on the topic. This involves an extended search and

Table 1
Consensus panel methods

Panel composition	Source of panel decision	
	Internal	External
Expert	*Ad hoc* panel	Delphi
Non-expert	Jury	Meta-analysis

nomination process, and often open public comment sessions that include the press as well. Rather than one global issue, the decision is also broken down into a set of subsidiary questions or issues with a procedure for addressing them. Typically, subgroups are formed to address these issues. They can either be members of the panel or another group of experts who are not part of the group and thus not subject to its potential biases (e.g., groupthink, the dominance of prestigious panel members, etc.).

The swine flu decision, as reviewed at government request (Neustadt and Fineberg 1978), illustrates both the strengths and weaknesses of this method. In early 1976 four cases of swine flu were reported among US Army recruits at Fort Dix. The Center for Disease Control (CDC) convened an *ad hoc* panel to determine if a vaccine should be produced and distributed to prevent a pandemic that fall. The panel was composed of members of four Federal and State agencies plus CDC staff. A month later this same group reconvened in a joint meeting with a standing CDC advisory committee appointed and staffed by the Director. Only one of the acknowledged elder statesmen in the field of virology participated in this meeting, and he strongly believed that the country was approaching the end of an 11-year cycle when a new pandemic would occur. At the end of a day's discussion a consensus emerged that 'the possibility of a pandemic existed' and that everyone was at risk. Therefore, the panel recommended that enough vaccine be produced to inoculate the entire US population.

A similar biased consensus unanimously endorsing a swine flu vaccine program also resulted from a third panel appointed by President Ford one month later. He wanted the advice of a 'representative group' of experts, but also had to rely largely on the CDC to select them. Although some famous scientists were added, there were still notable exceptions. Most of the panelists felt that the decision had already been 'programmed.' President Ford publicly announced the Federal program later that day. A series of disasters then followed including the deaths of three high-risk elderly people in Pittsburgh early that fall shortly after being vaccinated, and soon thereafter numerous reports of vaccine-related paralysis. In December, upon the recommendation of CDC, President Ford suspended the program.

A month later an outbreak of Victoria flu erupted in a Florida nursing home. The CDC advisory panel recommended 'limited resumption of the swine flu program' so that the Victoria vaccine that was included in the bivalent doses would be available. The new US Secretary of Health, Education, and Welfare decided to convene a new, fourth 'advisory group' that was more representative and independent than the first three—two CDC and President Ford's—panels. The *ad hoc* group was to be chaired by two of the 'nation's most distinguished scientists' who were not

part of the 'flu establishment' and the meeting was open to the public and the press. The 'improvised' consensus procedure led to a recommendation to lift the ban on the bivalent vaccine for 'high-risk' groups such as the elderly. It was considered 'a great success.'

2.2 Delphi Method

One of the major problems facing *ad hoc* panels such as those considering the swine flu vaccine program was the lack of information, particularly concerning a key issue—the probability of a swine flu pandemic. Other related problems concerned the influence of dominant group members, conformity, and hidden agendas. The Delphi method (Dalkey 1969) was created to solve all of these problems.

Group process artifacts are eliminated in the Delphi Method by avoiding a face-to-face meeting. According to Dalkey (1969), one of the developers of this group judgment method, the Delphi procedure has three important 'features': (a) anonymity of the panelist, (b) multiple sessions or 'interations,' and, as noted above, (c) an external, quantitative group decision. Panelists remain anonymous since their contact is by phone or mail through the Delphi coordinator. Multiple sessions are conducted in which feedback from the previous session or 'round' is provided to the individual panelists. Typically, the information is a quantitative summary of all the panelists ratings on important questions needed to make a decision. For example, to obtain an estimate of the probability of a swine flu pandemic five experts were surveyed using a mailed questionnaire with their responses obtained by telephone (Schoenbaum et al. 1976). The process was repeated until the change in the average (in this case, the median) was less than a specified criterion (here 10 percent).

The Delphi method has a number of other benefits. It is 'rapid,' inexpensive, and can produce a relatively 'objective' estimate of 'opinion' where there is a minimal amount of information available. The swine flue decision which was investigated using this method took only four weeks, comparable to the time taken by the *ad hoc* panels. It also produced a median probability estimate for the pandemic of 0.10—a figure too low to justify a Federal program to vaccinate the entire US population.

More recently researchers at RAND (Brook et al. 1986) have developed a newer consensus method using a variant of the Delphi technique. Nine-member expert physician panels were asked to rate the 'appropriateness' of an exhaustive set of 'indications' for performing (or not performing) various procedures (e.g., carotid endarterectomy, CE) using a nine-point interval scale. Panels were selected to balance speciality, practice, and location. There were two rounds of ratings. The first, typical of the Delphi method, involved anonymous ratings using the mail to provide the panelists relevant information including literature reviews. The second, however, involved a one- or two-day, face-to-face meeting at RAND to rerate the procedures. Each physician was provided a summary distribution of all the panelists' initial ratings with their own clearly indicated. After discussion, the indications were rerated. Four quantitative methods for determining unanimity were calculated (i.e., absolute with all ratings in a three-point region indicating 'inappropriate' (1–3), 'equivocal' (4–6), or 'appropriate' (7–9); scale indicating all ratings in a consecutive three-point region; and both absolute and scale with the highest and lowest ratings removed).

2.3 Jury Method

Juries are arguably the single most common method for consensus decision making given their pervasive use in legal proceedings. While such procedures may seem beyond the scope of a discussion of formal consensus methods, that is simply not the case. Not only have juries been used to make decisions on the same issues that have been dealt with by other consensus methods (Angell 1996), but a variant of the jury method, the 'Consensus Development Program' developed by the US NIH (Perry and Kalberer 1980), has been adopted by other western nations (Lomas et al. 1988, Vang 1986).

In both cases, the juries or panels are composed of nonexpert, neutral persons who listen to an adversarial discussion or presentation by experts over a number of days before reaching a judgment or decision. For example, juries have awarded billions of dollars in damages to women who have successfully sued companies making silicon breast implants claiming that the implants have caused a variety of major diseases (e.g., cancer, arthritis, lupus, etc.). Nevertheless, a 16-member expert scientific panel of the US National Academy of Sciences concluded in a unanimous consensus that 'women with implants are no more likely than other women [without implants] to develop these systemic illnesses' (Institute of Medicine 1999).

The major difference between legal juries and those more formal methods such as the NIH Consensus Development Program is the role of experts (Angell 1996). The law has a liberal definition of expert that allows those with strong beliefs, but not equally strong credentials or evidence, to present their views to juries. Since it is often difficult for both lawyers and juries to assess the credibility of the evidence for these views, persuasion and argumentation skills generally prevail. This is not solely the case in formal consensus procedures such as the one used by NIH since 1977. The neutral panelists have both the range of disciplinary background and skills to assess the evidence, and the expert presenters or witnesses focus on the evidence, usually their own research, rather than their interpretation or opinion.

Donald Fredrickson, the former director of NIH, developed the NIH Consensus Development Program based on his experience with the successful swine flu *ad hoc* panel described (Perry and Kalberer 1980). In particular, he incorporated into this jury procedure the use of a distinguished panel chair, a series of questions, and open meetings. The panel or jury meets in open session over a two-day period at NIH, listening to presentations on the various questions, and interrogating the speakers or presenters. On the third day the panel presents its verdict—a written draft consensus statement—to the public and the press for comment, and then revises it prior to publication.

2.4 The Meta-analysis Method

Meta-analysis is the most recent consensus method. Like the Delphi method, it uses quantitative methods to aggregate or summarize expert knowledge—in this case, the scientific research literature (cf., Cooper and Hedges 1994). An average effect of some human intervention or activity (e.g., the effectiveness of CE surgery to prevent stroke or the adverse health effects of secondhand smoke) is calculated by the meta-analyst, whose role is similar to the Delphi coordinator, from all available studies—both published and unpublished—that represent the views of all experts. Consequently, it is of most value in areas where a substantial amount of conflicting scientific research has been done on a controversial issue.

A welldone meta-analysis avoids the subjectivity inherent in the other consensus methods since it relies entirely on the actual evidence developed by experts rather than their opinions or interpretations, and can be subject to replication. Moreover, it overcomes well-known human cognitive limitations in aggregating large amounts of information (Smith and Glass 1977). For example, a comparison of meta-analysis to both *ad hoc* groups and the RAND modified Delphi method has shown meta-analysis to provide a more accurate estimate than those consensus panels (Antmann et al. 1992, Wortman et al. 1988).

Such summaries or research syntheses can either stand alone, be commissioned to answer a policy question similar to the breast implant issue, or be incorporated into a group decision process. A meta-analysis to assess the impact of US public school desegregation on the academic achievement of Afro-Americans illustrates all three of these uses (Cooper 1986). As the meta-analysis was being completed, the US government convened an expert panel to examine this issue.

The six experts selected to represent the range of opinions about the benefits of school desegregation on black achievement—with two in favor, two against, and two neutral—decided to replicate the meta-analysis using only the strongest evidence. Consequently, an initial unanimous consensus was reached to include only 18 of the original 31 studies. An additional study by two members of the panel that had originally been rejected as low in quality was added. All panelists then conducted their own meta-analyses with the assistance of the original meta-analyst as well as an outside methodological consultant. However, the panelists' prior beliefs affected their choice of evidence (Cooper 1986). Despite this bias, 'the panelists achieved a high degree of correspondence … [and] agreed the effect [of desegregation] was positive' (Cooper 1986 p. 348). Nevertheless, the panel moderator/methodologist was unable to persuade the panel to agree on a written consensus statement to that effect.

3. Conclusions

The descriptions of the four consensus methods indicate that they have differing strengths and weaknesses. One important weakness that potentially affects all of them, however, is selection bias, either in the composition of the panels, the questions asked, or the evidence retrieved. While selection bias in the first three methods involving human panels has been noted above, it is also the main problem in meta-analysis. Selection or publication bias may result in the omission of unpublished studies thus overestimating the average effectiveness of an intervention. The use of open meetings is one way to reduce this bias as is careful initial planning to seek out all relevant experts and evidence. In cases where evidence is available, a combination of methods such as the jury and meta-analysis may be warranted (Wortman et al. 1988). With the rapid increase in both information and access to large computer databases, both quantitative, external consensus combined with qualitative, internal consensus may be the next step in the development of consensus methods.

Bibliography

Angell M 1996 *Science on Trial*. W. W. Norton, New York

Antmann E M, Lau J, Kupelnick B, Mosteller F, Chalmers T C 1992 A comparison of the results from meta-analysis of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *Journal of the American Medical Association* **268**: 240–8

Brook R H, Chassin M R, Fink A, Solomon D H, Kosecoff J, Park R E 1986 A method for the detailed assessment of the appropriateness of medical technologies. *International Journal of Technology Assessment in Health Care* **2**: 53–63

Cooper H M 1986 On the social psychology of using research reviews: the case of desegregation and black achievement. In: Feldman R S (ed.) *The Social Psychology of Education*. Cambridge University Press, New York

Cooper H, Hedges L V 1994 *The Handbook of Research Synthesis*. Russell Sage Foundation, New York
Dalkey N 1969 An experimental study of group opinion: The Delphi method. *Futures* **1**: 408–26
Institute of Medicine 1999 *Safety of Silicon Breast Implants*. National Academy Press, Washington, DC
Lomas J, Anderson G, Enkin M, Vayda E, Roberts R, MacKinnon B 1988 The role of evidence in the consensus process: Results from a Canadian consensus exercise. *Journal of the American Medical Association* **259**: 3001–5
Neustadt R E, Fineberg H V 1978 *The Swine Flu Affair: Decision-making on a Slippery Disease*. US Govt. Printing office (No. 017–000-00210-4), Washington, DC
Perry S, Kalberer J T 1980 The NIH consensus development program and the assessment of health-care technologies: The first two years. *New England Journal of Medicine* **303**: 169–72
Schoenbaum S C, McNeil B J, Kavet J 1976 The Swine-influenza decision. *New England Journal of Medicine* **295**: 759–65
Smith M L, Glass G V 1977 Meta-analysis of psychotherapy outcome studies. *American Psychologist* **32**: 752–60
Vang J 1986 The consensus development conference and the European experience. *International Journal of Technical Assessment in Health Care* **2**: 65–76
Wortman P M, Smyth J M, Langenbrunner J C, Yeaton W H 1998 Consensus among experts and research synthesis: A comparison of methods. *International Journal of Technical Assessment in Health Care* **14**: 109–22
Wortman P M, Sechrest L 1988 Do consensus conferences work? A process evaluation of the NIH Consensus Development Program. *Journal of Health Politics, Policy, and Law* **13**: 469–98

P. M. Wortman

Consequentialism Including Utilitarianism

Roughly speaking, consequentialism is the theory that is the way to tell whether a particular choice is the right choice for an agent to make or have made is to look at the relevant consequences of the decision: to look at the relevant effects of the decision on the world. In directing us to consequences the theory is teleological in focus; the term comes from the Greek word *telos*, meaning end or goal. It is opposed to any theory that does not make the consequences into the ultimate touchstone of rightness and wrongness in choices. The classical antonym is the deontological approach—this term comes from the Greek work *deon*, meaning obligation or duty—but there are other antonyms too, as we shall see. Where consequentialism would assess a choice by looking at its consequences, a deontological approach would assess it by looking at how the choice measured up to the obligations incumbent on the agent.

1. The Definition of Consequentialism

Nonconsequentialism comes in many forms. Deontologists hold that agents ought to discharge certain duties: they ought to tell the truth, keep their promises, be nonviolent, etc. Kantians say that agents ought to act on the categorical imperative—act only on a maxim that they could accept as a general law of behavior—or ought to treat other people always as ends. Virtue ethicists say that they ought to manifest certain virtues in their behavior. Rights theorists maintain that they ought to respect certain rights that others have against them. Contractualists assert that they ought to conform to principles that no one could reasonably object to as the bases of social life. Theorists of special obligation say that they ought to deal in a certain way with those who are bound to them, such as their children, spouses, and friends. And so on.

What is the common thread in these positions? All nonconsequentialists speak, at whatever level of abstraction, about what any or every agent ought to do or be; in that sense they are universalists. All nonconsequentialists prescribe neutral patterns of behavior or psychology or relationships: that people act on the categorical imperative, manifest certain virtues, respect the rights of their dependants, nurture their friendships, and so on. The patterns are neutral in the sense that they can be understood in the same way by everyone. And all nonconsequentialists say that the right thing for an agent to do is to instantiate the prescribed patterns—so far as they are coinstantiable—in their own behavior or psychology or relationships. In particular, they say that that is the thing to do even if instantiating a pattern in their own life means, because of the perversity of the agent's circumstances, that the pattern will be less fully realized than otherwise in the world as a whole.

Consequentialism makes a dual assertion in response to such positions (Baron et al. 1997). First, it asserts that there are certain potentially shared values by means of which possible states of affairs can be ranked, although perhaps not completely. These may be neutral patterns in human behavior or psychology or relationships, of the kind that the nonconsequentialist prescribes, or they may be neutral outcome patterns that are more detached from agency: patterns such as the maximization of human happiness or the saving of uninhabited wildernesses. And second, consequentialism maintains that the right choice for an agent to take in any decision is one of those choices, assuming there is at least one—assuming incomplete ranking is not a problem—that promote the overall realization of such values or patterns (more below on what it is to promote values).

The basic difference, then, between consequentialists and their opponents is that while each side privileges certain general patterns—treats them as values, as it is natural to say—the consequentialist side

says that the important thing is to promote the realization of those patterns in the world at large, while the opposing side says that the important thing is rather to instantiate the relevant patterns in one's own life: in one's behavior or relationships or psychology, or whatever. Consequentialists say that privileged patterns should be treated as consequences to be promoted, nonconsequentialists that they should be treated as constraints to be satisfied.

The difference between the two approaches is particularly salient for simple cases where only one pattern is taken to be relevant, and where there is no issue about the relative importance of different patterns. Consider those cases, then, where nonviolence is assumed to be the only privileged pattern—the only value—that is relevant. And consider the different ways in which that pattern may be espoused among pacifists.

The nonconsequentialist pacifist says that nonviolence is to be instantiated by agents, even if this means that there is less nonviolence overall. The consequentialist counterpart says that nonviolence is to be promoted overall, even if this means instantiating violence: say, waging the war to end all wars. In practice the two sorts of pacifist will often converge in their recommendations, as nonconsequentialist Quakers converged with Bertrand Russell in their opposition to World War I. But there is always a possibility of their coming apart, as indeed those Quakers and Russell came apart in their attitudes to World War II.

Making the distinction between consequentialism nonconsequentialism by reference to whether the agent is required to promote certain privileged patterns or rather to instantiate them—to honor them—is only one among a number of ways of making it. Another common approach is to distinguish between agent-neutral values and agent-relative ones, and to say that consequentialists advocate the promotion only of agent-neutral values, nonconsequentialists the promotion of a mixed bunch, or of agent-relative values only (Nagel 1986).

Agent-neutral values are patterns that are identified without reference to who is doing the identifying: patterns like those involved in peace being achieved, happiness maximized, a wilderness preserved, and so on. Agent-relative values are patterns that are identified by back reference to the identifying person: patterns, to take the case where I do the identifying, like *my* doing my duty, *my* conforming to the categorical imperative, *my* manifesting certain virtues, *my* looking after my children, and so on. Where consequentialists advocate the promotion of agent-neutral values, and in particular agent-neutral values of a universal cast—they do not involve benefits just to a particular person or place or whatever—nonconsequentialists can be represented as advocating the promotion of agent-relative values: the promotion by every agent of *their* doing their duty, *their* conforming to the categorical imperative, *their* manifesting certain virtues, *their* looking after their children, and so on.

2. *A Theory of the Right, Not a Theory of the Good*

To say that something is good is to hold that it has a certain value, in particular a certain positive value. To say that something is right is to hold that among certain relevant alternatives—say, among the options in a choice—it is the one to have. And so rightness and goodness may come apart. An alternative that is wrong may yet be an object of great value; it may be wrong simply because another option is an object of even greater value. An alternative that is right, on the other hand, may not be something of great value; it may be right, simply because it is the least bad option among a very poor set of alternatives. A theory of the good, a theory of value, would enable us to determine the values of different entities. A theory of the right would enable us to determine, for any set of alternatives, which alternative or subset of alternatives is the right one.

When I speak of what is good, and of what has value, I have in mind an impersonal or agen t-neutral conception of goodness and value. If I value a prospect for the increase of happiness it promises, or even for a particular effect it will have, say on planet Earth, then I value it agent-neutrally. If I value it for the benefits it will have for me or mine, or for the fact that it will keep my hands clean, or for any reason of that self-referential kind, then I value it agent-relatively. The theory of the good, or the theory of the valuable, refers to the theory of what is agent-neutrally valuable.

Does consequentialism present itself as a theory of the good or as a theory of the right, or as something involving commitments in both areas? By the account given, the theory is concerned only with the right, not with the good. It says that the right alternative in any choice is that which promotes the good, without making any commitment on the nature of the good. While it makes the right a function of the good, it says nothing in itself on the matter of what is good.

This aspect of consequentialism is often found surprising, as many people tend to identify consequentialism with utilitarianism. But utilitarianism involves a double commitment. It says that the good should be equated with happiness, whether this is understood in the sense of pleasure or preference satisfaction, and then it adds that the right alternative in any set is that which best promotes such happiness (Sen 1979). Thus when G. E. Moore (1903) rejected utilitarianism, arguing that the good was not reducible to happiness—it was a non-natural property that allowed of no such reduction—he remained a consequentialist about the right. Many contemporary consequentialists have nonutilitarian theories of the good or the valuable.

Indeed they often look like nonconsequentialists to the extent that they may think that the good encompasses reflexive patterns, such as those involved in each person keeping their promises, respecting those with whom they deal, or giving their children special attention.

Robert Nozick (1974), himself an outstanding nonconsequentialist, supports this view when he distinguishes between a nonconsequentialist theory of rights—a theory of rights, proper—and a consequentialism or utilitarianism of rights. The consequentialist theory of rights would say that it is good that certain rights are enjoyed generally, and it would imply that the right option in any choice is that which promotes the enjoyment of such rights. This right option will itself instantiate the rights-observing pattern in most imaginable cases but it will not do so in those exceptional circumstances where the best way to promote the pattern is not to instantiate it oneself: where the best way to promote rights-observance is to breach that observance in one's own case.

3. *Consequentialism, Global, and Local*

Consistently with being a theory of the right, consequentialism can be thought of as having a smaller or a larger range of application to human agents (Pettit and Smith 2000). There are sets of alternative acts that we might perform in different choices, sets of alternative decision procedures that we might follow in choosing acts, sets of alternative motives out of which we might be led to act, sets of alternative rules that we might comply with in acting, and so on. Is consequentialism a local doctrine that just tells us what the best act is in any choice, for example—or the best decision procedure, or motive, or rule, or whatever—and that then says that the best alternatives in other sets are to be defined in a derived way? If acts are privileged, the best decision procedure might be that which produces the best acts, the best rule that which is consistent with the choice of the best acts, and so on through other categories. Or should consequentialism be taken to apply globally, telling us that as the best act is that which produces the best consequences among the alternatives on offer, so the same lesson holds for items in the other sets of alternatives?

Most consequentialist positions in the literature tend to be local in character. Act-consequentialism, as it is called, tends to say that the best act is the act with the best consequences, and that the best factor in any other set of alternatives involving an agent is to be defined in some way—a number of ways are possible—by reference to how it serves the choice of best acts. Rule consequentialism gives the priority to rules in a parallel way, and motive consequentialism gives the priority to motives.

The most natural position, however, would seem to be global consequentialism (Parfit 1984). It would say that no matter what the set of alternatives under consideration, the one to be preferred is that which does best by the consequences. The same lesson will apply, of course, to that set of alternatives whose members are different sets of act–procedure–motive–rule combinations; the right combination will be that which does best by the consequences overall. The act or procedure or motive or rule selected for any context under the best combination may not be the best possible one relative to the apparent alternatives in that context. But it will be the best possible one that is available consistently with the best possible combination overall.

4. *The Role of Consequentialism*

There is a long tradition among the opponents of consequentialist doctrines, in particular the opponents of utilitarianism, of suggesting that consequentialists are committed to holding that every choice should be made in a highly calculative, actuarial mode. F. H. Bradley made the point in the last century, writing about the utilitarian approach: 'So far as my lights go, this is to make possible, to justify, and even to encourage, an incessant practical casuistry; and that, it need scarcely be added, is the death of morality' (Bradley 1876, p. 107). Consequentialists have almost always resisted this charge. Thus, Henry Sidgwick in the last century and J. J. C. Smart in this have argued forcibly that utilitarianism does not require that agents all make their decisions by explicit reference to how the options will do by the promotion of happiness (Smart and Williams 1973). The point they have wanted to make was nicely summed up early in the nineteenth century by the jurisprude, John Austin, in defending the utilitarian thinker:

> Though he approves of love because it accords with his principle, he is far from maintaining that the general good ought to be the motive of the lover. It was never contended or conceived by a sound, orthodox utilitarian, that the lover should kiss his mistress with an eye to a common weal (Austin 1954, p. 108).

As recent consequentialists hve almost all taken the view that consequentialism is a theory of the right, not a theory of the good, so they have nearly unanimously argued that it is a theory for assessing the right alternative for an agent or an agency to have adopted, not necessarily a useful theory to be applied in choosing from among alternatives. The point is usually made with respect to acts but it can also be extended to alternatives in other categories. The claim is that making one's decision by reference to which option has the best consequences may be a way of

making one's decision that does not itself have the best consequences. Suppose one is concerned with one's own pleasure, for example. It is notorious that agents who make their decisions by reference to which option will promote the most pleasure may actually enjoy less pleasure than those who make their decisions in the more spontaneous fashion, say by reference to certain rules of thumb (Railton 1984, Pettit and Brennan 1986).

This aspect of consequentialism has been described as 'the paradox of teleology' (Scanlon 1998, p. 383, n. 15). It has important implications and these need to be appreciated, if the appeal of consequentialism is to be understood. It means that the consequentialist can look with approval on agents whose mode of decision making in everyday life involves a consideration of consequences only in periodic reviews of their performance or in times of crisis. These agents may be routinely available to their family and friends; they may be reliable truth tellers and promise keepers; and they may be dependably respectful of the claims of others upon them. The consequentialist will approve of such people to the extent that they are more likely to promote the good—whatever view of the good is adopted—than actuarial, calculative types.

5. *The Positive Case for Consequentialism*

Why be a consequentialist? Surprisingly, there has been very little discussion of this question in the literature. Consequentialists tend to assume that if one is to be serious about the enterprise of moral assessment, then there is no alternative to looking at the consequences—the consequences, neutrally, or impersonally characterized—of the choices assessed. How could it be rational, they ask, to neglect any such consequences? They assume that if nonconsequentialists think otherwise, that is because of an unquestioning commitment to a theological or a commonsense viewpoint. A theological viewpoint might suggest that human agents ought to fulfill prescribed obligations and let God look after the consequences. And a commonsense viewpoint might support the attitude that the important thing in moral decision making is to look after one's own moral standing, to keep one's own hands clean, and not to worry, except perhaps in exceptional circumstances, about any bad consequences that may follow on this.

There is also a second reason why consequentialists may not have concerned themselves much with providing a defense of their position. Consequentialists are a mixed bunch, as we saw, insofar as they differ widely among themselves on the question of which substantive sorts of consequences—which valuable patterns—are the ones by reference to which choices ought to be assessed. For example, they divide into utilitarians and nonutilitarians, depending on whether or not they accept the view that it is only consequences that bear on the utility of sentient beings which matter: only consequences that affect the happiness or preference satisfaction of such beings. Being a mixed bunch, consequentialists tend each to be concerned more with arguing for their particular view of what consequences count than with arguing for the shared consequentialist credo that it is indeed consequences that matter and not the sort of thing on which deontologists and other nonconsequentialists focus.

My own belief is that those who are attracted to consequentialism are moved by a different view of justification from those who are attracted to the different forms of nonconsequentialism (Baron et al. 1997). According to consequentialism, people will be able to think of their choices as justifiable to one another only insofar as they think that there are agent-neutral values that others ought to accept and that serve to justify those choices. According to nonconsequentialist positions, people will be able to think of their choices as justifiable under a weaker condition: if they think that there are agent-relative values such that others can be made to understand why they should be moved by those values—others are moved by corresponding values in their own case—and the values justify their choices. Thus consequentialists think that justification must ultimately be based on potentially shared or identical concerns: say, with overall happiness or peace or good relations between people. But nonconsequentialists think that it can be based, even at the most ultimate level, on concerns that are isomorphic rather than identical across people: say, on the individuals' concern that they should not break their promises, whatever the impact on promise keeping more generally, or that they should further the welfare of their children, whatever the repercussions for overall child care.

6. *The Negative Case for Consequentialism*

But if consequentialists have not done much in the way of arguing for their position, what have been the issues debated between them and opponents (Scheffler 1988)? These have mainly had to do with how far consequentialism can support firm commonsense intuitions about what is morally right. Opponents typically object to consequentialism on the grounds that in one way or another it would undermine otherwise compelling moral attitudes. Rather than providing positive reason for adopting their position, consequentialists have often been content to try to deal with those objections or to try to argue that while the points made are valid, they are not as important as deontologists make them seem.

There are three sorts of objection, in particular, that consequentialists have had to confront in this ex-

change. First, that their doctrine would make many intuitively wrong choices right, many intuitively right choices wrong; second, that even where it represents many intuitively right choices as right it would provide intuitively the wrong reasons for thinking them right; and third, that it would represent as obligatory various right choices that are supererogatory, that is intuitively beyond the call of duty (Scheffler 1982, Kagan 1989). Consequentialists have often argued in this exchange that our intuitions on relevant matters need to be revised but the more general line has been to maintain that on reflection consequentialism proves less counterintuitive than its opponents maintain.

7. *The Collective Face of Consequentialism*

Suppose we had a society of utilitarians, each of whom acted in a way that utilitarians would approve. Is it possible that the collective result of their so acting might be worse than the collective result of actions that did not individually earn utilitarian approval? Is it possible, in particular, that the collective result might be worse in utilitarian terms: that it might actually lead to less human happiness or preference satisfaction than certain alternatives? Is it possible that, however satisfactory as a theory of rightness for individual agents, it is not in this sense a collectively satisfactory theory (Hodgson 1967)?

Does it matter that a theory designed for individuals is collectively self-defeating Does it matter that agents who try to do what is best by the theory, or succeed in doing what is best by the theory, generate a collective result that is worse in their own terms than it might have been (Parfit 1984)? Some are happy to say that a theory can be self-defeating in this way and yet remain a compelling theory for individuals. Others will think that this is a matter of the first importance for the assessment of a theory.

Some may be willing to concede that consequentialism can be collectively self-defeating but most deny this, on variety of grounds (Lewis 1972, Singer 1972). One well-known line is that where collective self-defeat appears to threaten, it usually transpires that one of the options before the individuals is not being taken into account: say, the option of exploring possibilities of cooperation with others, rather than acting unilaterally, or the option of seeking further information on the prospects facing the agent (Regan 1980).

The prisoner's dilemma is often taken to show that what is individually rational may not be collectively rational: each of two prisoners confesses to a crime because, whatever the other does, confessing promises a better result than refusing to confess; yet each would be better off if both refused to confess than they are by both confessing. It is worth noting that where agents have distinct ends, as they will have under nonconsequentialism, then a prisoner's dilemma is a constant threat. The distinct ends that agents have may consist in individually wanting to do their own duty, individually wanting to instantiate the categorical imperative in their own behavior, individually wanting to do the best by their own children, and so on. Such ends are potentially competitive and so it is a permanent possibility that by each pursuing their own ends as best they can, they will secure a result under which neither does as well as they might have done in their own terms. By each trying to do the best they can by their own children, for example, they may each do less well than if they pooled their resources and did their best for children as such. Whatever problems are raised for consequentialism on the collective front, it is better protected in this respect against prisoner's dilemmas, for it assigns to each agent the same end of promoting the best consequences (Parfit 1984).

8. *Some In-house Issues*

There are a number of more or less in-house questions that are debated among consequentialists. One bears on how we should identify the alternatives to be evaluated in any decision. Should an agent do A or B, we ask. A, we answer, because its consequences do better on the relevant theory of the good. But what if there is some other option, C, which the agent might have chosen instead and which would have itself been better than A? This sort of example alerts us to the fact that in the consequentialist assessment of any choice, we have to be clear about what are the relevant alternatives to bring into consideration.

The question about alternatives has not received as much attention as it deserves in the literature, but it is undoubtedly of great importance (Bergstrom 1966, Jackson and Pargetter 1986). According to one view, we should assess any action against the alternative (or disjunction of alternatives) that the agent would have taken, had the action not been performed. According to another we should assess it against the alternatives that the agent could have performed instead: against the alternatives that were, in a relevant sense, within the power of the agent. From a third-person point of view the first approach would seem to be the most pertinent, although from a first-person point of view things are not so clear.

A second in-house question has been more widely debated in consequentialist circles. This is the issue as to whether in looking at the decision an agent or agency ought to have made, we consider the actual consequences of the option chosen and the consequences that would have followed on alternatives; or whether, rather, we look at the expected consequences at the time of choice (Lewis 1969, Jackson 1991). If the answer is that we should look at expected rather than actual consequences, then the question is whether

we should look at the subjectively expected consequences—in the manner of decision theory—or whether we should look at the consequences that were in some sense objectively expected (Menzies and Oddie 1992). Suppose that a doctor prescribes a drug for a nonfatal skin condition, which has the following features: there is a 10 percent chance that it will kill the patient, an 80 percent chance that it will make no difference, and a 10 percent chance that it will cure the complaint. Imagine that the drug works and the complaint is cured. Does consequentialism take its cue from actual consequences and say that the doctor made the right choice? Or does it look to expected consequences, whether subjectively or objectively expected, and say that the doctor made the wrong choice? That is the issue here.

The third more or less in-house issue is whether consequentialism should hold that the right option is that which does best by relevant consequences—that which is optimific—or whether it is sufficient for an option to be right that it does well enough, as we might say, by the relevant consequences. Some theories of rationality are maximizing theories, requiring rational agents to maximize some function, whereas others are satisficing theories: these require rational agents just to perform satisfactorily by some relevant criterion. The question here is whether consequentialism should not take the softer, satisficing line, rather than the stern, maximizing one (Slote and Pettit 1984).

See also: Decision Making, Psychology of; Utilitarian Social Thought, History of; Utilitarianism: Contemporary Applications

Bibliography

Austin J 1954 *The Province of Jurisprudence*. Weidenfeld, London

Baron M W, Pettit P, Slote M 1997 *Three Methods of Ethics: A Debate*. Blackwell, Oxford, UK

Bergstrom L 1966 *The Alternatives and Consequences of Actions*. Almqvist and Wiksell, Stockholm

Bradley F H 1876 *Ethical Studies*, 2nd edn., rev. Oxford University Press, Oxford, UK

Hodgson D H 1967 *Consequences of Utilitarianism*. Clarendon Press, Oxford, UK

Jackson F 1991 Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics* **101**: 461–82

Jackson F, Pargetter R 1986 Oughts, options and actualism. *Philosophical Review* **95**: 233–55

Kagan S 1989 *The Limits of Morality*. Oxford University Press, Oxford, UK

Lewis C I 1969 *Values and Imperatives*. Stanford University Press, Stanford, CA

Lewis D 1972 Utilitarianism and truthfulness. *Australasian Journal of Philosophy* **50**

Menzies P, Oddie G 1992 An objectivist's guide to subjective value. *Ethics* **102**: S12–33

Moore G E 1903 *Principia Ethica*. Cambridge University Press, Cambridge, UK

Nagel T 1986 *The View from Nowhere*. Oxford University Press, Oxford, UK

Nozick R 1974 *Anarchy, State, and Utopia*. Blackwell, Oxford, UK

Parfit D 1984 *Reasons and Persons*. Clarendon Press, Oxford, UK

Pettit P, Brennan G 1986 Restrictive consequentialism. *Australasian Journal of Philosophy* **64**: 438–55

Pettit P, Smith M 2000 Global consequentialism. In: Hooker B, Mason E, Miller D E (eds.) *Morality, Rules and Consequences*. Edinburgh University Press, Edinburgh, UK, pp. 121–33

Railton P 1984 Alienation, consequentialism and the demands of morality. *Philosophy and Public Affairs* **13**: 134–71

Regan D H 1980 *Utilitarianism and Cooperation*. Oxford University Press, Oxford, UK

Scanlon T M 1998 *What We Owe to Each Other*. Belkanap Press of Harvard University Press, Cambridge, MA

Scheffler S 1982 *The Rejection of Consequentialism*. Oxford University Press, Oxford, UK

Scheffler S (ed.) 1988 *Consequentialism and its Critics*. Oxford University Press, Oxford, UK

Sen A 1979 Utilitarianism and welfarism. *Journal of Philosophy* **76**: 463–89

Singer P 1972 Is act-utilitarianism self-defeating? *Philosophical Review* **81**: 94–104

Slote M, Pettit P 1984 Satisficing consequentialism. *Proceedings of the Aristotelian Society* **8**(Suppl.): 139–76

Smart J J C, Williams B 1973 *Utilitarianism, For and Against*. Cambridge University Press, Cambridge, UK

P. Pettit

Conservation: Waterways

Waterways include natural rivers and artificial waterways which may be rivers modified for navigation. Legally the word 'river' usually includes all natural streams however small, which have a definite and permanent course. Within each drainage basin, rivers and river channels are the surface expression of the way in which water from precipitation proceeds through the hydrological cycle to oceans, seas, or lakes. In the global hydrological cycle at any one time there is just 0.03 percent of the world's freshwater, or 0.0001 percent of the world's total water, stored in rivers and waterways. The largest river in the world, the Amazon, contributes 13 percent of the total river runoff reaching the oceans annually; the 10 largest world rivers according to their discharge account for more than 26 percent of the freshwater reaching the oceans each year; and the drainage areas or watersheds of the 10 largest rivers of the world occupy 20 percent of the earth's land surface (Leopold 1994).

Rivers are dynamic components of environment because the quantity and the quality of the water alter, and the river channel itself can change position. It is partly because of this dynamic character, but also because of their resource value, that rivers have been managed deliberately by a range of direct and indirect influences so that few world rivers are now completely natural and unmodified, with more than 60 percent of the world's total river flow being regulated. Recorded river modification began as early as 5,000 years ago with the advent of irrigation in the hydraulic civilizations of the Nile and the Indus basins. Subsequently there were many small-scale modifications of rivers for drainage schemes, fish weirs, water mills, navigation, and flotability, but the major control of rivers and river systems came with the Industrial Revolution when rivers were harnessed for industrial mills, power generation, cooling water, extensive irrigation, and water supply schemes. Technological advances in the twentieth century enabled even greater modification of rivers. For example, the impact of dam construction was relatively small until 1900, moderate to 1945, but substantial after the great increase of dam building between 1945 and 1971. Whereas in 1900 there were 427 dams higher than 15 m around the world, this increased to 5,268 in 1950 and to 39,000 in 1986 (ICOLD 1988). Building continues, for example, on the Qingjiang River, a tributary of the Yangtse, a large dam is planned for completion in 2009 at the same time as the Three Gorges dam in China. For the most part, however, the era of dam building came to an end in the developed world by the late 1980s and has been succeeded by the retirement of some dams.

Dam construction and engineering modification of waterways dominating the past 2 centuries followed from a 'technological fix' vision of the environment and its resources. More recently, conservation of rivers and waterways has become dominant, as has protection of the landscape more broadly to meet the needs of the present without compromising the opportunity for future generations (Newson 1992). This shift in management approach is illustrated by five themes pertinent to conservation of waterways.

1. *Waterway Conservation*

1.1 The Need to Control the Quantity of Water in River Channels

The growth of human populations and changes in land use adjacent to rivers and waterways increases floods and flood damage from storm events. Increasingly, therefore, river management necessitates not only the modification of river channels to convey a design flow or calculated discharge of water, but consideration of a range of complementary strategies including flood control schemes, land-use zoning, and choice between big or small dams to retain flood waters according to the nature of the physical environment. Despite increased expenditure on flood control and greater understanding of the methods for flood assessment and mitigation, an increase in losses often follows flood events because of the assumption that the implemented scheme guarantees no flood damage (Smith and Ward 1998). A flood prevention scheme is designed for events of a certain magnitude and recurrence interval. Events of a larger magnitude or more frequent occurrence than the design are not uncommon. In addition, design usually considers maintenance of low flows to minimize the effects of drought upon water supply and river uses, such as fishing and recreation.

1.2 The Need to Control the Quality of River Water

Waterways initially were seen as arteries to be used for disposal of waste products of society. The magnitude of waste has grown so large, however, that successive legislation has demanded more controls, to minimize the effects of serious levels of water pollution. Particular problems arise from point sources of pollution released into a river at an identifiable location, especially from wastewater treatment systems, industrial plants, agricultural pollution, and mining operations (National Research Council 1999). The culmination of such controls, for example, has witnessed the return of salmon to the Thames and Tyne rivers in the UK. Pollution is not always an easy problem to resolve because of the way in which pollutants relate to the sediment in the river channel. In northern New Mexico, for example, plutonium released from Los Alamos Laboratory on to nearby canyon floors was attached to sedimentary particles and in subsequent decades some of the sediments and the plutonium removed to the Rio Grande river raised concerns about environment and health hazards (Graf 1994).

1.3 The Need to Manage River Channels

River channelization involves modification, usually by engineering methods, for the purposes of flood control, drainage, navigation, and prevention of erosion. Direct channelization by engineering methods including straightening and reinforcing the bed and banks, led to the modification of the majority of lowland rivers in Europe over the past 2,000 years (Brookes and Shields 1996). Extension of channelization in the United States produced unwelcome consequences so that alternative schemes of river channel management were sought, including river restoration conceived as 'working with the river rather than against it.' Many river channels are now being restored to appear as 'natural' as possible and it is not unknown for rivers

which were straightened by engineering to be 'restored' to their original meandering form. The difficulty in such restoration is to prescribe such dimensions of the restored channel as will enable the channel to contain flows likely to be very different in character from those existing before engineering channelization (Brookes and Shields 1996).

1.4 The Need to Acknowledge that River Restoration Involves More than Returning a River Channel to its Original Condition

The dynamic nature of the river channel must be considered, aided by identifying and conserving areas with special features as exemplars of typical waterways and also by educating the public about natural riverscapes. Restoration requires that river channels are managed with an awareness of the way in which they change and develop: the dynamic river must be seen against a timescale when there may have been considerable changes in palaeohydrology (Gregory et al. 1995). Consideration of what is natural is required because public perception does not always correspond with reality, especially in forested areas where woody debris is normal in river channels but public perception tends to prefer debris-free channels. Conserved areas provide an important reference designated as SSIs or sites of special scientific interest in the UK (Gregory 1997). In the US, the National Park Service (NPS) maintains high quality recreational areas and facilities, which include rivers and river access, through the 'Rivers, Trails, and Conservation Program' of the Department of the Interior. As societies move into post industrial stages, social values for rivers are changing from rivers as water resources to more complex multipurpose ecosystems or landscapes that potentially serve many objectives (Graf 1996).

1.5 The Need to Manage Rivers and Waterways Holistically Within the Drainage Basin

This need arose with the appreciation that river management in one area may have effects in other parts of the basin so that schemes of Integrated Basin Management (IBM) were devised. The modern development of water resource systems proceeded from single purpose projects through multipurpose schemes towards more integrated river basin planning. Important in progress towards more integrated planning was the development of the Tennessee Valley Authority (TVA), created as an independent federal agency from 1933 (National Research Council 1999). In order to achieve the two original primary purposes of flood control and navigation development, the Tennessee River and its tributaries were developed into one of the most controlled river systems in the world. Electric power is generated at the center of a suite of projects including navigational improvements and flood controls on the Tennessee River proper and soil erosion control, reforestation, improvement in agricultural land use, and increased and diversified industrial development elsewhere in the watershed. The TVA blazed the trail of integrated basin management. Downs et al. (1991), however, advocate distinguishing comprehensive river basin development where several components are involved, from integrated basin management where schemes interact, and from holistic management which dynamically incorporates energetics, change, and human interactions. The need to integrate consideration of ecological, as well as morphological and hydrological aspects, within a dynamic approach to river management and conservation, can be accomplished within an overall integrated approach. A new strategy for watershed management has been advocated in the US with drainage basins or watersheds as the basis for planning and management (NRC 1999).

Waterways which were once seen as elements of the environment to be conquered and tamed, are now seen as integral parts of dynamic evolving systems to be treated in a way that is sustainable and to be enjoyed and appreciated by future generations. A holistic view of rivers and waterways within the basin is the best basis for conservation, but sustainable management requires water users to make choices that cannot be made solely on the basis of scientific evidence, in part because the results of science may be ambiguous at times. In addition to scientific evidence, sustainable watershed management requires community understanding and support. Therefore, IBM depends on partnerships between governments and the public, across disciplines and international borders, and among water users with different interests and values (Heathcote 1998). Paradoxically, as greater emphasis is placed on holistic approaches, the systems managed have been increasingly fragmented by human activity, for example into reaches controlled by dams with each reach exhibiting behaviour that can be markedly different from that of nearby reaches (Graf 1996). A further consideration necessary in relation to conservation is appreciation that rivers and waterways are valued in different ways by different cultures.

See also: Desertification; Environmental and Resource Management; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; Environmental Security; Environmentalism: Preservation and Conservation; Sustainable Development; Water Pollution: Terrestrial and Groundwater; Water Resources

Bibliography

Brookes A, Shields F D (eds.) 1996 *River Channel Restoration: Guiding Principles for Sustainable Projects*. Wiley, Chichester, UK, p. 433

Downs P W, Gregory K J, Brookes A 1991 How integrated is river basin management? *Environmental Management* **15**: 299–309
Graf W L 1994 *Plutonium and the Rio Grande. Environmental Change and Contamination in the Nuclear Age*. Oxford University Press, Oxford, UK, p. 329
Graf W L 1996 Geomorphology and policy for restoration of impounded American rivers: What is 'natural'? In: Rhoads B L, Thorn C E (eds.) *The Scientific Nature of Geomorphology*. Wiley, Chichester, UK, pp. 443–473
Gregory K J (ed.) 1997 *Fluvial Geomorphology of Great Britain*. Geological Conservation Review Series, Joint Nature Conservation Committee, Chapman and Hall, London, p. 347
Gregory K J, Starkel L, Baker V R (eds.) 1995 *Global Continental Palaeohydrology*. Wiley, Chichester, UK, p. 334
Heathcote I W 1998 *Integrated Watershed Management. Principles and Practice*. Wiley, Chichester, UK, p. 414
ICOLD 1988 *World Register of Dams*, Update. International Commission on Large Dams, Paris
Leopold L B 1994 *A View of the River*. Harvard University Press, Cambridge, MA, p. 298
National Research Council: Committee on Watershed Management 1999 *New Strategies for America's Watersheds*. National Academy Press, Washington, DC, p. 311
Newson M 1992 *Land, Water and Development. River Basin Systems and their Sustainable Management*. Routledge, London, p. 351
Smith K, Ward R C 1998 *Floods—Physical Processes and Human Impacts*. Wiley, Chichester, UK, p. 382

K. J. Gregory

Conservation: Wetlands

Wetlands is a generic name given to all types of terrain where land and water mix, and which share some of the characteristics of both. They include bogs, fens, levels, mangroves, marshes, mires, peatlands, potholes, pocosins, sloughs, swamps, river bottom lands, tidal flats, wet meadows, and wet prairies.

Wetlands cover about 7.8×10^6 km^2, or 6 percent of the world's surface, and are found in every climatic zone from the tundra mires to the tropical mangroves. They are highly fragmented and diverse, and with only a few exceptions (e.g., the Everglades, the Fenland, the Dutch polders, and major river deltas and lower valley bottoms) they rarely cover large stretches of land. Their diversity has been simplified and commonalities emphasized into the marine and estuarine wetlands of the coast, and the riverine, lacustrine, and palustrine of the interior (Cowardin et al. 1985, National Research Council 1995).

The need to conserve wetlands arises from their continuous and increasing conversion and loss. In the past, wetlands were considered wastelands, neither sound land nor good water, and therefore drained for agriculture, dredged-out for harbors or marinas, filled-in for urban or industrial purposes, or used as dumps for domestic and industrial waste. Conversion reached high proportions from c.1830 to 1960, when modern earthmoving machinery made excavation easy, steam (and later diesel and electrical) pumps lifted vast quantities of water over embankments, and the mass manufacture of clay tile drains and pipes made boggy ground firm. It is difficult to give precise figures as the process of the manipulation of hydrology is at least pre-Medieval in Europe (Darby 1983, Lambert 1985) and Asia (Ruddle 1987), and equally as ancient in the Americas (Denevan 1992). It is also difficult to define 'drained'—is it totally dry, flooded for a month, seasonally, or what? But of the estimated 1606×10^6 ha of land drained globally, 42 percent is in North America, 24 percent in Europe, 8 percent in Russia, and 20 percent in the South East Asian deltas, which leaves an infinitesimal amount of 6 percent in Africa and South America. Large as it is, however, it is still a mere fifth of the global total, the great bulk of which lies in the tundra wetlands, and to a lesser degree in the temperate coastal marshes and the tropical lowlands of the Amazon and Congo basins.

But since the early 1960s the beneficial, positive values of these productive ecosystems have been recognized increasingly. For example, their hydrological/physical functions include flood mitigation, coastal protection, recharging aquifers, and sediment trapping. Chemical functions include pollution trapping, removal of toxic residues, and waste recycling, while biological functions include biomass production and wildlife habitats. The socioeconomic consumptive benefits include food production, fuel from organic soils, fiber from hardwood forests, mangroves and reed swamps, and everywhere fish, fowl and fauna. But, perhaps above all else, it is the hard-to-quantify nonconsumptive benefits that have dominated the urge to conserve, and these include scenic, recreational, educational, aesthetic, archaeological, scientific, heritage, and historical benefits that abound in an environment where air, land, and water, and their fauna and flora, meet in an attractive and delicate way that has caught scientific and popular imagination.

Regret at the loss of game after reclamation was more than outweighed by pride in the creation of productive agricultural landscapes. Yet, paradoxically, the first concerted effort towards conservation came in the United States from the hunting lobby, aided by recreational and wildlife enthusiasts, with the active support of the Fish and Wildlife Service (FWS). Under the Migratory Birds Hunting and Conservation Act of 1934 the FWS had the right to sell 'duck stamps' which were purchased by all wildfowl hunters. Between 1934 and 1984 over $250 million was collected, the bulk being used to purchase 970,000 ha of waterfowl habitats, mostly wetland. In addition, the private duck-hunting lobby—'Ducks Unlimited' (with 600,000 members)—has raised funds ($400 million since 1937), and has purchased, created or manages over 100,000 ha of wetlands, and carried out a sustained campaign of lobbying and advertising, which

has been echoed by the Audubon Society and the Nature Conservancy Council.

After about 1960, the loss of wildfowl habitats alone was calculated to be about 121,000 ha/yr. Rising public awareness of their scenic/recreational attractions, and positive natural functions, led to public support for their conservation. Intellectual curiosity about wetlands had also been stimulated by integrated ecological studies, and research centers were established, such as the Center for Wetland Resources at the State University of Louisiana, Baton Rouge, and the Center for Wetlands at Florida University.

In addition, it was realized that the US Army Corps of Engineers (which regulated navigable rivers and coastal waters used in interstate commerce under the River and Harbor Act of 1899) was dredging annually about $300 \times 10^6 m^3$ of sediment out of coastal channels and navigable waterways and then backfilling the material in other wetlands to make land suitable for urban/residential development and marinas.

The federal Coastal Zone Management Act of 1972 allowed all states bordering the ocean and the Great Lakes to apply for federal grants for wetland preservation, provided they adopted strict land use and purchasing regulations. Coastal zone management plans were devised, which either prohibited coastal developments or allowed changes only by special permit (e.g., San Francisco Bay).

In time concern extended to the interior freshwater wetlands as toxic wastes were dumped in swamps, open space near urban areas was lost, and agriculturalists drained and infilled wildlife habitats, which incensed the powerful hunting, fishing and birdwatching lobby. But there was no legislation to control loss, although the regulation of water quality was possible.

Meanwhile in 1972, the Corps of Engineers was made responsible for regulating the discharge of dredged material under Section 404 of the Federal Water Pollution Control Act. Environmentalists argued that this implied total wetland protection, an interpretation resisted by the Corps, which argued that its responsibility did not extend to the headwaters of rivers where many wetlands were situated. Nevertheless, after legal action by environmental groups, the responsibility had to be admitted, and the Act was strengthened in 1977 (now called the Clean Water Act). From now on 'all the waters of the United States' were included, which was a legal and not a scientific concept. In addition, schemes for coastal and river alteration by the Corps could only occur after review and approved by the Environmental Protection Agency (EPA), the FWS and the National Marine Fisheries Service (NMFS), the EPA having an absolute veto over the use of any sites to be used for the dumping of dredged material.

Executive Order 11990 of 1977 required every federal agency 'to minimize the destruction, loss, or degradation of wetlands' by avoiding support of any new construction in wetlands. But coordination between agencies is poor and responsibility divided. The Corps, for example, is responsible only for those waters affected by interstate commerce, which excludes much of the coastal and nearly all inland waters. In addition, while it cannot dump dredged material in those wetlands it does control, it cannot regulate their reclamation. This explains the current massive conversion of the Mississippi River bottomlands, stimulated by the impetus of high crop prices.

It is clear too, that some federal policies, e.g., legislation dealing with agricultural tax credits, subsidies, and soil conservation has actually encouraged wetland destruction. Since 1985, however, the Food Security Act (under its 'Swampbuster' provisions), linked eligibility of farm subsidy benefits directly with what farmers did with their wetlands, preservation being rewarded and conversion being penalized. In addition, farm debts could be exchanged for conversion of land back to wetland. Increasingly, the link between the financial encouragement of agriculture and the destruction of wetlands is being realized.

Thus, although greater awareness of wetland protection is coming about, losses are still occurring. Annually, about 20,200 ha of coastal wetlands are still being converted, but over 100,000 ha in the unregulated inland wetlands, primarily for agriculture. The battle is far from won. Success will not come until the traditional approach to protection via either land use or water quality legislation is abandoned. Neither of these recognizes that it is the combination of land and water that is the unique characteristic of wetlands.

Compared to the United States, where wetlands have 'achieved divine status' (Kwong 1998), the same degree of awareness and concern about wetland loss, and hence protection, has not characterized the countries of Europe. Fragmentation of political units, the lack of a politically powerful or numerous hunting lobby, broadly 'rural' as opposed to environmental concerns, and the overwhelming dominance of food production and self-sufficiency as national goals, supported particularly by the Common Agricultural Policy of the European Union, have been the overwhelming characteristics of rural policy. Furthermore, a common awareness that centuries, if not millennia, of human manipulation have altered the European countryside and its soils, hydrology, and flora to its present state have been a corrective to any excessively sentimental notions of a 'pristine' nature.

Agricultural draining accounts for the bulk of the possible 300,000 ha lost annually, and there is now considerable awareness of the need to conserve the last remaining pieces of wetland. Most large-scale conversions, as in the Fens and Holland, are over, and it is tile draining that is nibbling away at the edges of the remaining freshwater wetlands, the filling-in of farm ponds that are perceived as redundant, and the draining of small areas of coastal salt marshes. But some coastal areas are under threat, e.g., around

Brittany and adjacent coasts that total over 500,000 ha, the Camargue and Etangs de Languedoc-Rousillon (170,000 ha) on the Mediterranean, the latter already radically altered in places by tourist development and marinas. Another unknown is the fate of the Shannon estuary in The Republic of Ireland, where large-scale industrial developments are proposed. In Ireland over 100,000 ha of peat bogs have been drained and water tables lowered prior to peat extraction, and the lowland raised bogs of Ireland and western Scotland have been severely reduced by afforestation schemes.

Actual losses are difficult to quantify because of fugitive and fragmentary sources. In total, perhaps as much as 615,135 ha were drained in France (5.2 per cent of the country) between 1970 and 1982 alone, and probably as much as 1,418,000 ha of Ireland (24.3 per cent of the country) between 1945 and 1980, of which nearly four-fifths was field drainage. In total, perhaps as much as half of the lowland fens and mires in France and the United Kingdom have been destroyed or damaged. Another measure of change is the number of species endangered or threatened. For example, 47 percent of all European bird species, all 111 of European Union breeding species of reptiles and amphibians, and 2,000 plant species are in decline.

Preservationist measures must be directed at the main culprit, agriculture and its support. In the past, CAP payments have provided generous subsidies for 'improvement' to boost production, particularly in the Less Favoured Areas which include, for example, the wetlands of western Ireland, the Islands and Highlands of Scotland, and parts of Mediterranean France and Italy. In addition support for farm improvement has been accompanied by support for prices. But since the later 1980s gross over-production, high storage costs of surpluses, adverse public opinion, and the drain that the CAP is having on the Community's finances has caused a reevaluation. Milk and food production quotas, set-aside schemes, reflooding to recreate wetlands, have all been undertaken. More significantly, the recognition of the unique fauna and flora in wetlands, and their place in the intercontinental migratory flight paths has led to their declaration as Sites of Special Scientific Interest (SSSI) in Britain, or Environmentally Sensitive Areas (ESAs) there and elsewhere throughout the EU. Birds, newts, and unusual plants are becoming the key to their preservation in a newly environmentally conscious society.

The complex infrastructure of legislation, legal precedent, scientific knowledge and active conservation groups has little counterpart in the less developed world. The need to expand food production in line with rising population have always made wetlands attractive potential farmland. In the past, at least 45,000 km^2 of the Sundarbans of the Ganges–Brahmaputra delta, and the Irrawaddy, Chao Phraya, and Mekong deltas have been transformed by millions of peasant settlers, and are only the most visible examples of a much greater process continuously at work along the entire coastline of South and Southeast Asia (Richards 1991). In more recent decades, foreign technological aid and capital inputs (especially large multi-purpose dams, barrages, canals, etc.) have often misjudged the physical capabilities and properties of wetlands. They have also ignored the traditional, indigenous, polycultural systems of agriculture–livestock–aquaculture which are sustainable, low energy-using, and frequently offer better returns than agricultural conversion. Large-scale schemes like the Kafue Gorge dam in Zambia, and the Jonglei canal across a bend in the Upper Nile, Sudan, have been highly controversial, and it is doubtful if their benefits have outweighed those of the systems they destroyed. The diseconomies of rapidly initiated and capital-intensive schemes are not confined to Western intervention. The government-sponsored transmigration schemes of Indonesia to move millions of farmers from Java to the outer islands will affect c. 200,000 km^2 of coastal wetlands. But soils are frequently nutrient-deficient, dry out too rapidly and shrink, and will not sustain the hoped-for intensive double cropping of rice. Rather, they need polycultural horticulture, planted with great sensitivity to local conditions.

Many interior wetlands are threatened in tropical Africa and South America, as are coastal tidal peats and mangroves in South and Southeast Asia. For example, in Africa there are some 224,000 km^2 of lightly used riverine flats (mostly papyrus marsh that seasonal flooding almost doubles in size) along the Nile, Niger, Zaire, and Zambesi rivers, as well as the Okavango basin. Similarly vulnerable are the floodable Gran Panatal on the upper reaches of the Rio Paraquai (c.100,000 km^2) and the flood plains of the Apure–Aranca tributaries of the Orinoco in Venezuela (c.70,000 km^2). The wetlands of the Amazon are incalculable.

Finally, the impetus to international cooperation over the conservation of wetlands has come once more through the preservation of wildfowl habitats, the one positive function that all parties can perceive, appreciate and agree upon. The Ramsar convention (1971) aims to protect Wetlands of International Importance, specifically for birds, and in 1980 the Cagliari conference widened the definition to include any endangered plant or animal species. An inventory is being compiled and maintained by the IUCN at the International Monitoring Center in Switzerland. These agreements have had some striking successes in preventing conversions in Italy, the Wadden Sea, and Pakistan, but equally they have limitations and are only as good as the compliance of the countries that sign-up to them, which currently is about only half the countries of the world.

See also: Desertification; Ecotourism; Environmental and Resource Management; Environmental Planning; Environmentalism: Preservation and Conservation; Precautionary Principle; Sustainable Development

Bibliography

Cowardin L M, Carter V, Golet F C, La Roe E T 1985 *Classification of Wetlands and Deepwater Habitats in the United States*, 2nd edn. Department of the Interior, Washington, DC

Darby H C 1983 *The Changing Fenland*. Cambridge University Press, Cambridge, UK

Denevan W M 1992 The pristine myth: The landscape of the Americas in 1492. *Annals of the Association of American Geographers* **82**: 369–85

Kwong J 1998 Back to a simple and pure nature. In: Anderson D, Mullen P (eds.) *Faking It: The Sentimentalisation of Modern Society*. Penguin Books, London

Lambert A M 1985 *The Making of the Dutch Landscape: An Historical Geography of the Netherlands*. Academic Press, London

Maltby E 1986 *Waterlogged Wealth: Why Waste the World's Wet Places?* International Institute for Environment and Development, London

Mitsch W J, Gosselink J G 1986 *Wetlands*. Van Nostrand Reinhold, New York

National Research Council (US) 1995 *Wetlands: Characteristics and Boundaries*. Committee on Characterization of Wetlands. Water Science and Technology Board, Board on Environmental Studies and Toxicology, Commission on Geosciences, Environment and Resources, National Research Council. National Academy Press, Washington, DC

OTA (Office of Technology Assessment) 1984 *Wetlands: Their Use and Regulation*. OTA-0-206, US Government Printing Office, Washington, DC

Richards J F 1991 Agricultural impacts in tropical wetlands: Rice paddies for mangroves in South and Southeast Asia. In: Williams M (ed.) *Wetlands: A Threatened Landscape*. Basil Blackwell, Oxford, UK

Ruddel K 1987 The impact of wetland reclamation. In: Wolman M G, Fournier F G A (eds.) *Land Transformation in Agriculture. SCOPE (Scientific Committee on Problems of the Environment) 32, SCOPE/ICSU (International Council of Scientific Unions)*. Wiley, Chichester, UK

Tiner R W Jr et al. 1984 *Wetlands of the United States: Current Status and Recent Trends*. National Wetlands Inventory Fish and Wildlife Service, U.S. Dept of the Interior, Washington DC

Williams M 1991 *Wetlands: A Threatened Landscape*. Basil Blackwell, Oxford, UK

M. Williams

Conservatism: Historical Aspects

1. Approaches to the Problem

An intuitive procedure for defining conservatism is to begin by listing the institutions which conservatives have sought to conserve. That will not get us very far, for conservatives have, at one time and place or another, defended a wide range of social, political, and economic institutions.

The study of conservatism has proved difficult for modern social science. Compared with the other great ideological phenomena of the last two centuries, such as liberalism, socialism, communism or fascism, the subject of conservatism is underilluminated by scholarship. One reason is because of the difficulty of arriving at meaningful generalizations about conservatism, which displays less obvious uniformity across national borders and tends to be more nationally particular than liberalism or socialism, which aspire to be universal in their reach. Moreover, since conservatism emphasizes the need for institutional and symbolic continuity with the particular past, its symbols and institutional ideals tend to be more tied to specific, usually national, contexts. For all of these reasons, studies of conservatism which go beyond one particular national context are scarce; as a result, conceptions of conservatism have tended to be parochial.

One of the earliest social scientific approaches to the issue was formulated by the Hungarian–German sociologist, Karl Mannheim, in his essay *Conservative Thought* (1927). He introduced the important distinction between traditionalism and conservatism, the former a near-universal psychological propensity to do things as they have traditionally been done, the latter an articulated set of inter-related political ideas, an ideology which arose, Mannheim thought, in response to the new and dynamic historical processes associated with the Enlightenment (see *Enlightenment*), capitalist modernization (see *Capitalism* and the bourgeois class (see *Bourgeoisie/Middle Classes, History of*. A pioneer of the sociology of knowledge, Mannheim claimed that conservative thought was linked to a worldview which could be correlated with class position, roughly that of the feudal nobility. The basic categories of each ideology, he maintained, were unconsciously determined by the experience of the world and the self-interest of the class which developed it. In the case of conservatism, it led to a focus on experience and the concrete, as opposed to universalistic, rationalistic theories of liberalism and the Enlightenment. Conservatives offered a critique of the rationalistic conceptions for reorganizing society proposed by advocates of radical reform, and focused on explaining the historical particularities of real, existing societies and the interconnectedness of their institutions.

Subsequent social scientific analysts of conservatism, even when they have not followed Mannheim's focus on class, have stressed the reactive nature of conservatism, the fact that it arises in response to a challenge (intellectual, political, or cultural) to existing institutions, on behalf of which conservative arguments are then developed. In one of the most perceptive scholarly analyses of the subject, Samuel Huntington (1957) argued that conservatism is best understood not as an inherent theory in defense of particular classes and institutions, but as a positional ideology. 'When the foundations of society are threatened, the conservative ideology reminds men of

the necessity of some institutions and desirability of the existing ones', Huntington suggested. He claimed that because 'the articulation of conservatism is a response to a specific social situation … The manifestation of conservatism at any one time and place has little connection with its manifestation at any other time and place.'

The reactive nature of conservatism is a premise of one of the finest historical studies of the subject, Klaus Epstein's *The Genesis of German Conservatism* (1966), which pays close attention to political and intellectual context and demonstrates how conservatism arose in German-speaking Europe as a response to the Enlightenment, commercial capitalism, and bourgeois liberalism. Taking the reactive approach a step further, Martin Greiffenhagen (1971), suggested that because self-conscious conservatism only arises once the institutions it values have lost their hold, conservative thought seeks not a preservation of the status quo, but uses an imaginatively transfigured conception of the past with which to criticize the present. According to Greiffenhagen, the romanticization of the past is thus an intrinsic and recurrent element of conservatism.

Two thinkers presented nonreactive conceptions of conservatism during the immediate post-war decade. In *The Conservative Mind* (1953), Russell Kirk, a US conservative man of letters, offered a conception of conservatism as based upon religion and 'belief in a transcendent order, or body of natural law, which rules society as well as conscience'—a conception which assimilated conservatism to religious orthodoxy, and was of scant use as an analytic tool. Kirk's conception was rejected by Michael Oakeshott, the most important British conservative philosopher of the era, in *On Being Conservative* (written 1956 and published in *Rationalism in Politics and Other Essays*, 1962), which presented a brilliant phenomenological description of conservatism as a disposition centered 'upon a propensity … to delight in what is present rather than what was or what may be.' Oakeshott sought to conserve a liberal and pluralist conception of society, and doubted that 'a disposition to be conservative in politics' depended on the religious and metaphysical assumptions set forth by Kirk.

2. *Definition and Characterization*

To understand conservatism as a distinctive mode of social and political thought, it is useful to distinguish it, ideal typically at least, from reaction and from orthodox. The orthodox theoretician defends existing institutions and practices because they are metaphysically true: the truth proclaimed may be based on particular revelation or on natural laws purportedly accessible to all rational men, it may be religious or secular in origin. The conservative defends existing institutions because their very existence creates a presumption that they have served some useful function, because eliminating them may lead to harmful, unintended consequences, or because the veneration which attaches to institutions that have existed over time makes them potentially usable for new purposes.

Conservatism can similarly be distinguished from reaction. The conservative seeks to conserve existing institutions, usually recognizing that the process of conservation may include the need for evolutionary reform. The reactionary, by contrast, is at odds with existing institutions, and seeks to return to some institutional status quo ante, often in a form transfigured by memory and ideology.

As misleading as the confusion between conservatism and orthodoxy is the false dichotomy of conservatism and Enlightenment. Contrary to the frequent characterization of conservatism as the enemy of the Enlightenment, it is more historically accurate to say that there were many currents within the Enlightenment, and some of them were conservative. Indeed, conservatism as a distinct mode of thought is a product of the Enlightenment. What makes social and political arguments conservative as opposed to orthodox is that the critique of liberal or progressive arguments takes place on the enlightened grounds of the search for human happiness, based on the use of reason.

3. *Characteristics of Conservative Thought*

Conservative theorists repeatedly decry the application to society and politics of a mode of thought which they characterize as overly abstract, rationalistic, and cut off from experience. Whether termed 'the abuse of reason' (by Burke 1790), 'rationalism in politics' (by Oakeshott 1962) or 'constructivism' (by Hayek), the recurrent conservative accusation against liberal and radical thought is that they are said to depend upon a systematic, deductivist, universalistic form of reasoning which fails to account for the complexity and peculiarity of the actual institutions they seek to transform.

For conservatives, the historical survival of an institution or practice—be it marriage, monarchy, or the market—creates a *prima facie* case that it has served some human need. The conservative emphasis on 'experience' is linked to the assumption that the historical survival of an institution or practice is evidence of its fitness in serving human needs. Since custom and habit are important features of human conduct, some of the usefulness of a practice comes from the fact that those engaged in it are already 'used' to it, and are apt to be discomfited by change.

Conservative thought has typically emphasized the imperfection of the individual, an imperfection at once biological, emotional and cognitive. Conservatives typically contend that human moral imperfection leads men to act badly when they act upon their uncontrolled impulses, and that they require the

restraints and constraints imposed by institutions as a limit upon subjective impulse. Conservatives, thus, are skeptical of attempts at 'liberation'. Conservatives have also stressed the cognitive element of human imperfection, insisting upon the limits of human knowledge, especially of the social and political world. They warn that society is too complex to lend itself to theoretical simplification, and that this fact must temper all plans for institutional innovation.

These assumptions explain the emphasis of conservative social and political thought upon institutions, that is, patterned social formations with their own rules, norms, rewards and sanctions. While liberals typically view with suspicion the restraints and penalties imposed upon the individual by institutions, conservatives are disposed to protect the authority and legitimacy of existing institutions because they believe human society cannot flourish without them. Unlike liberals, who favor voluntary, contractual social relations, conservatives emphasize the ongoing importance of nonvoluntary duties, obligations, and allegiances.

While there is no necessary link between conservatism and religious belief, conservatives have tended to affirm the social utility of religion. They make several arguments for the utility of religion: that it legitimates the state; that the hope of future reward offers men solace for the trials of their earthly existence and thus helps to diffuse current discontent which might disrupt the social order; and that belief in ultimate reward and punishment leads men to act morally by giving them an incentive to do so.

Conservatives typically emphasize the unanticipated negative consequences of deliberate social action. Such negative consequences, they argue, occur because reformers are unaware of the latent functions of existing practices and institutions. Reformers are insufficiently cognizant, it is said, of the contribution of the practice to the preservation or adaptation of the larger social system in which it is implicated. That contribution may be unintended by those engaged in the practice, so that its function may be unrecognized, or recognized only retrospectively, once the reform of the practice has brought about negative unintended consequences.

Together, the concepts of unanticipated negative consequences, latent functions, and functional interdependence serve as recurrent arguments against radical or wholesale reform, though not against reform as such.

Recurrent substantive themes of conservative social and political thought include:

(a) a skepticism regarding the efficacy of written constitutions, as opposed to the informal, subpolitical, and inherited norms and mores of society. For conservatives, the real 'constitution' of society lies in its historical institutions and practices, which are inculcated primarily through custom and habit;

(b) the central role of cultural manners and mores in shaping character and restraining the passions, and hence the political importance of the social institutions in which such manners and mores are conveyed;

(c) an emphasis on the family as the most important institution of socialization, and despite considerable divergence among conservatives over the proper roles of men and women within the family, the assertion that some degree of sexual division of labor is both inevitable and desirable;

(d) the legitimacy of inequality, and the need for elites, cultural, political, and economic; security of possession of property as a prime function of the political order;

(e) and the importance of the state as the ultimate guarantor of property and the rule of law, and hence the need to maintain political authority.

4. *Historical Overview*

Terms related to 'conservative' first found their way into political discourse in the title of a French weekly journal, *Le Conservateur* founded in 1818 by François-René de Chateaubriand with the aid of Louis de Bonald. 'The *Conservateur* upholds religion, the King, liberty, the Charter and respectable people (*les honnêtes gens*)', it proclaimed in its first issue. In England, the label 'Conservative Party' was first applied to the Tories by the publicist John Wilson Crocker in 1830. In Germany, the term 'conservative' came into use later in the same decade: it was first employed as an epithet against those designated as conservative, who were accused of seeking to preserve existing institutions at any price, but the term was eventually accepted by those it had been used to stigmatize. But as with many other historical developments, the phenomenon of conservatism long preceded the use of the term in political life.

The precursors of conservatism may be found in the Anglican critique of the Puritan contention that the elect or the inspired congregation, guided by their individual interpretations of the Bible, were entitled to exercise political authority. The thought of David Hume marks a watershed in the development of conservative social and political thought into a coherent, secular doctrine. Hume began by borrowing and expanding upon this critique of the politics of religious 'enthusiasm'. And he went on to criticize what he saw as its secular counterparts in the philosophically implausible and politically subversive doctrines of natural rights and of voluntary contract as the sole legitimate basis of political obligation. In 1757, Edmund Burke used similar terms to attack what he called the 'abuse of reason' among some enlightened thinkers. On the Continent, conservative social and political thought arose as a critique of the policies of enlightened absolutism. Among the foremost practitioners of this criticism was Justus Möser, himself a part of the German Enlightenment, and a

reformer of a conservative and corporatist tinge. In Holland, another Enlightenment figure, Jean de Luzac, articulated politically conservative arguments opposing the claims of 'patriots' for a more democratic constitution. The contention that conservatism arose in opposition to the Enlightenment and to the French Revolution reflects Edmund Burke's polemical characterization of the Enlightenment in his critique of the French Revolution. Though frequently reiterated, the contention is historically untenable. Conservatism arose not against the Enlightenment but within it. Much of post-1789 intellectual conservatism is continuous with the analytic strategies, if not with the tone, of pre-revolutionary conservative analysis.

It was the democratic radicalism of the French Revolution that evoked Burke's *Reflections on the Revolution in France* (1790), which became the most influential work in the history of conservative thought. Joseph de Maistre (1753–1821) and Louis Gabriel de Bonald (1754–1840) were the two leading figures in the tradition of French counter-revolutionary thought. The task with which they saw themselves confronted was rather different than the one that had motivated Burke's *Reflections*, and the nature of their conservatism varied accordingly. Burke's primary task was to conserve a British order which was substantially intact, though subject to intellectual attack. He could thus stress the historical utilitarian arguments that the existence of an institution creates a *prima facie* case that it has served some human need, and that historical continuity is intrinsically valuable because it increases the emotional hold of an institution upon its members. Similarly, Burke could underscore the importance of habit and custom. By the time de Bonald and de Maistre came to write their major works, the forces of the French Revolution had already overturned or transformed many of the key institutions of the Old Regime. The Counter-Revolutionary conservatives, therefore, were forced by circumstances to offer a rationale other than that of historical continuity, for the institutions of the old regime had been disrupted by the Revolution. De Maistre responded to the revolution by carrying to new lengths previous conservative arguments of epistemological modesty, historical particularism, and the unanticipated consequences of deliberate action.

Burke's defense of monarchy, aristocracy, and the established church stamped European conservative thought at the end of the eighteenth century and in the early decades of the nineteenth. Yet a similar critique of democratic radicalism appears, with a more republican but nevertheless recognizably conservative thrust, in the USA, first in the writings of the US Federalists, and later in the work of conservative Whig critics of Jacksonianism, such as Rufus Choate. Hume's combination of political conservatism with the championing of commerce was to remain characteristic of subsequent British and US conservatism. By contrast, the French, Catholic strains of conservatism of de Bonald and de Maistre, as well as the romantic German strains of conservatism remained less accepting of capitalist economic development than the British and US strains.

The development of conservative social and political thought after the mid-nineteenth century shows a movement from substance to function: from the defense of particular institutions to the defense of institutions in general; from the defense of the landed aristocracy to a defense of elites in general; from an emphasis on the role of the established Church to the function of culture in linking the individual to communal purposes; from monarchical authority to the authority of the state in general. In the last decades of the nineteenth century, conservatism underwent an important shift, reflecting a change in opponents and in institutional substance. The increasing dominance of industrial capitalism, the expansion of the suffrage, the rise of socialist movements and of a new brand of liberalism which was more economically redistributionist and more culturally permissive, all led to a substantial transformation of conservatism. Many of the economic institutions and policies once associated with European liberalism (see *Liberalism*) now became that which conservatives sought to conserve. This new brand of conservatism was pioneered in late nineteenth century Britain, and is reflected in the works of James Fitzjames Stephen and W. H. Mallock and by the US theorist, William Graham Sumner.

The rhetoric by means of which institutions were defended also changed. It placed less emphasis on the veneration of tradition as such, while seeking to increase the emotional hold of existing institutions by showing that they were in accord with the needs of historical development. In England and the USA, the language of common law and of inherited tradition was increasingly abandoned in favor of the language of science. The justification of the social utility of landed elites was replaced by a more general defense of the need for elites in the realms of economics and politics, while the defense of the established Church was transformed into a defense of cultural elites. In Germany, too, the substance of conservatism was transformed, from a defense of aristocracy, a social order based on fixed estates, and paternalist monarchism, toward an emphasis on the functions of the state as such, a transformation exemplified in the work of the twentieth century German theorist, Carl Schmitt.

One recurrent strand of thought which is related to conservatism, yet politically and analytically distinct from it, is radical conservatism, a strand particularly powerful in inter-war Europe, where its adherents often came to favor fascist regimes (see *National Socialism and Fascism*). Radical conservatives share some of the concerns of more conventional conservatism, such as the need for institutional authority and continuity with the past. But they maintain that the processes characteristic of modernity have

destroyed the valuable legacy of the past for the present, so that a restoration of the purported virtues of the past demands radical or revolutionary action. Hence the credo of one radical conservative, Moeller van den Bruck, that 'Conservative means creating things that are worth conserving.' Radical conservatism shares with conservatism an emphasis on the role of institutions in providing restraint and direction to the individual, but seeks to create new institutions which will exert a far stronger hold on the individual than do existing ones, which because of their relative tolerance are perceived by radical conservatives as 'decayed.'

After the defeat and discrediting of fascism, conservatism in the latter half of the twentieth century was defined not only by its opposition to communism, but by antipathy to the spread of the welfare state and attempts to bring about economic redistribution, and by skepticism regarding the solution of social problems through massive governmental action.

The study of conservative movements and conservative ideology provide an ongoing challenge to social scientists. For they must provide analyses of a phenomenon which is dynamic and changes over time, which varies between nations and often between regions within nations, and which often appears in combination with a variety of other trends (religious, nationalist) and under a variety of nomenclatures. These difficulties—together with the political predilections of the majority of social scientists, who tend to be antipathetic to conservatism—have led to a gap between the historical significance of conservatism and the extent of its exploration by scholars. This is lamentable, for not only have more-or-less conservative parties dominated western European and US politics for substantial portions of the last two centuries, but conservative thinkers have played a leading role in the development of modern thought.

See also: Liberalism; Liberalism: Historical Aspects; Mannheim, Karl (1893–1947); Political Culture; Traditions: Social

Bibliography

Burke E 1790 *Reflections on the Revolution in France*. J. Dodsley, London

Epstein K 1966 *The Genesis of German Conservatism*. University of Princeton Press, Princeton, NJ

Greiffenhagen M 1971 *Das Dilemma des Konservatismus in Deutschland*. R. Piper, Munich

Huntington S 1957 Conservatism as an ideology. *American Political Science Review* **51**: 454–73

Kirk R 1953 *The Conservative Mind*. Regnery, Chicago

Mannheim K 1927 *Das Konservative Denken* reprinted in K H Wolff (ed.) 1964 Mannheim, *Wissenssoziologie* Neuwied, Luchterhard; translated as Conservative Thought, In: Wolff K H (ed.) 1971 *From Karl Mannheim*. Oxford University Press, New York, pp. 132–222

Muller J Z 1987 *The Other God that Failed: Hans Freyer and the Deradicalization of German Conservatism*. University of Princeton Press, Princeton, NJ

Muller J Z 1997 *Conservatism: An Anthology of Social and Political Thought from David Hume to the Present*. University of Princeton Press, Princeton, NJ

Nisbet R 1986 *Conservatism: Dream and Reality*. University of Minnesota Press, Minneapolis, MN

Oakeshott M 1962 *Rationalism in Politics and Other Essays*. Methuen, London

Quinton A 1978 *The Politics of Imperfection: The Religious and Secular Traditions of Conservative Thought in England from Hooker to Oakeshott*. Faber and Faber, London

Quinton A 1993 Conservatism. In: Goodin R E, Pettit P (eds.) *A Companion to Contemporary Political Philosophy*. Blackwell, Oxford, UK

Scruton R 1991 Introduction: What is Conservatism. In: Scruton R (ed.) *Conservative Texts: An Anthology*. Macmillan, Houndmills, Basingstoke, UK

Sirinelli J-F 1992 *Histoire des Droites en France*. Gallimard, Paris, 3 Vols.

Stern F R. 1961 *The Politics of Cultural Despair*. University of California Press, Berkeley, CA

Vierhaus R 1968 Conservatism. In: Wiener D P (ed.) *Dictionary of the History of Ideas*. Scriber, New York Vol 1 pp. 477–85

Vierhaus R 1982 Konservativ, konservatismus. In: Brunner O, Conze W, Koselleck R (eds.) *Geschichtliche Grundbegriffe. Historisches Lexikon zur Politisch-sozialen Sprache in Deutschland*. Klett-Cotta, Stuttgart, Germany Vol. 3 pp. 531–65

Wolff K H (ed.) 1964 *Wissenssoziologie*. Luchterhand, Berlin

J. Z. Muller

Conservatism: Theory and Contemporary Political Ideology

1. Definition

An initial problem that arises in attempting to define conservatism is whether it is a theory or simply an attitude towards change. Most dictionary definitions incline toward the attitudinal definition, describing conservatism as a preference for existing institutions, and opposition to change. With this definition the content of conservatism will vary depending on whatever is the current institutional and ideological status quo. Yet the writers, parties, and movements that have described themselves as conservative share a number of common values and doctrines, which typically include theories of human nature, history, society, and politics.

The term 'conservatism' was first used in England after the 1833 Reform Act when the Tory Party, which had roots going back to the late seventeenth century, renamed itself the Conservative Party. The central doctrines of conservatism were expressed before the word was invented, most notably in Edmund Burke's

Reflections on the Revolution in France (1790), where we can find most of the central doctrines of conservatism, developed in response to the liberal and radical thinking of the French revolutionaries. In opposition to their demand for liberty, Burke stressed the values of order, tradition, and gradual change. Against the revolutionaries' call for equality, Burke stressed society's need for hierarchy, deference, class differentials, and a special position for the church and the nobility. In opposition to the radical individualism of the French, he insisted on the importance of community, tradition, and history. He criticized the antiinstitutional rationalism of the revolutionaries by stressing the importance of social institutions that have stood the test of time, and the weakness and fallibility of individual reason. Against their centralizing statism, he lauded the 'little platoons' of localism as defenders of liberty, and contrary to their anticlericalism, he praised the elevating effect of inherited religious beliefs.

In developing these ideas Burke was able to draw on the Western tradition, classical ideas of natural law, Christian views of hierarchy and man's inherited inclination to evil, as well as medieval practices of localism and organic corporate groups. Although himself an adherent of the Whig Party, his ideas were influential not only on the thinking of the British Conservative Party, but on conservative parties and movements elsewhere in Europe, and later in the United States. As originally expressed, their emphasis on order over freedom, hierarchy over equality, religion over anticlericalism, localism over centralism, and community over the individual stood in marked contrast to the liberal doctrines that they criticized.

2. *European Conservatism*

In the political struggles of nineteenth and twentieth centuries, European conservative parties drew on the conservative tradition to defend the interests of property, the upper classes, the military, and the church, and appealed to nationalist and patriotic feelings in support of their programs. Their efforts were not restricted to the defense of class and privilege. In 1867, hoping to secure lower class support for 'tory democracy,' Benjamin Disraeli, the Conservative Prime Minister in England, extended the suffrage. Otto von Bismarck initiated a program of social insurance in Germany to counteract socialist appeals; and Pope Leo XIII in his 1891 encyclical, *Rerum Novarum*, defended the living wage for the worker, and promoted the organization of trade unions (preferably under church auspices) while denouncing the extremes of liberalism and socialism. Less laudably, conservative parties in Germany, Austria, and elsewhere, in the twentieth century, sometimes allied themselves with or supported extreme rightists and even fascists and Nazis, seeing them as a bulwark against Communism. Hitler, for example, was able to come to power legally because of the votes of German conservatives.

The conservative impulse continues to be important in European politics. Since the 1833 Reform Act the British Conservative Party has been in power much longer than any other party. On the continent Christian Democratic centrist parties often work together with conservative groupings, linked by their commitment to religion, and to a 'social market economy.' Twentieth-century conservatives, despite earlier hostility to market forces, have increasingly embraced free enterprise economics. They have been critical of European economic and political integration on nationalist grounds. Nationalist resistance to foreign immigration also accounts for some of the support for conservative parties and movements.

3. *Conservatism in America*

In *The Liberal Tradition in America*, published in 1955, Louis Hartz argued that with the exception of a few Southerners, notably John C. Calhoun, Americans were all Lockean liberals, and that European-style conservatism was absent from the American political scene (Hartz 1955). Ironically at the very time that his book was published, a genuine and self-conscious conservative movement had begun to emerge on the American scene. Russell Kirk had published *The Conservative Mind* (Kirk 1953) arguing that Burke rather than Locke should be the model for America. At the University of Chicago, the German émigré political theorist, Leo Strauss, published *Natural Right and History* (Strauss 1953), which attacked Locke as a proto-Hobbesian and denigrated the natural-rights tradition in favor of classical models of politics, especially that of Plato. Another émigré, Erich Voegelin, appeared on the cover of *Time* magazine after he published *The New Science of Politics*—in which he associated communism, liberalism, and revolutionism with a Christian heresy, gnosticism, that attempted to establish heaven on earth (Voegelin 1952). Whittaker Chambers, who had accused Alger Hiss of spying for the Soviet Union, also linked Communism and liberal utopianism in his book, *Witness* (Chambers 1952). In 1955 William Buckley, the *enfant terrible* who had earlier published *God and Man at Yale*, an attack on the dominance of secular and leftist views at his *alma mater*, founded *The National Review* as a journal to promote conservative political and economic thought. Combining militant anti-Communism, the defense of traditional values, and the promotion of free enterprise economics, it became the principal organ of the conservative movement in America.

Together these publications formed the intellectual basis of the conservative movement in mid-twentieth-century America. As in the case of William Buckley, their conservatism was also linked to free-enterprise

economics, as defended in the writings of two University of Chicago economists, Friedrich Hayek and Milton Friedman. In *The Road to Serfdom* (Hayek 1944) and later works, Hayek argued that the growth of government threatened liberty and stifled the economy, while Friedman's 1955 lectures, published as *Capitalism and Freedom* (Friedman 1962) developed similar themes, as well as making innovative proposals to increase competition and privatize government functions, many of which were adopted in later decades. Their arguments were based on classical economic liberalism, but in the American context the two economists were viewed as conservative since they defended the existing capitalist system. In addition, their common hostility to Communism linked them to the social and political conservatives in the newly burgeoning movement.

The conservative movement became a major factor in national politics with the presidential candidacy of Barry Goldwater in 1964. After Goldwater's defeat, Ronald Reagan became the movement's leader. Promising to revitalize national defense against the Soviet Union, restore moral values, and downsize the welfare state and reduce the scope of the federal government, he led the conservative wing of the Republican party to victory in the presidential election of 1980. In the late 1970s, the conservative movement received support from a new group, the Neo-Conservatives, former liberal intellectuals who criticized what they perceived to be liberal weakness in the face of the Communist threat, and rejected the excesses of the New Left counterculture that had emerged in the 1960s (Kristol 1983). They provided key policy makers to the Reagan administration in the areas of defense and education, while the free market economists directed economic policy, cutting taxes and welfare programs.

4. *The Religious Right*

Conservatism in America had often been an elite phenomenon, but the Religious Right gave it a broader populist base. Building on the religious opposition to the Supreme Court decisions outlawing school prayer and recognizing a constitutional right to abortion, conservative churchmen, mostly but not exclusively evangelical Protestants, organized the Moral Majority, and later the Christian Coalition, to reverse what they saw as a liberal tendency to secularism and toleration of immorality. They also shared with other conservatives opposition to the Soviet Union, and a belief in free market economics. In the 1990s, they opposed the gay rights movements, and the increasing toleration of homosexuality in America. They became an important pressure group within the Republican Party, effectively vetoing candidates for national office who favored abortion, and pressing for the appointment of Supreme Court justices who were sympathetic to their views.

5. *Tensions within Conservatism*

American conservatism contains within it many different currents. Ronald Reagan and the Soviet threat could hold its various strands together, but with the end of the Cold War there were increasing tensions among them, in particular as to whether social and moral issues, or downsizing government and encouraging free markets, should be given priority in public policy. Concern about moral decline, a frequent conservative theme (Bennett 1993) and issues such as abortion, gay rights, and public prayer became the subject of *Culture Wars* (Hunter 1991) from local school districts to the Supreme Court.

The members of the Supreme Court, most of them appointed by Republican presidents, became deeply divided along conservative–liberal lines, both as to theories of constitutional interpretation—'original position' vs. 'the living constitution,' federal-state relations—and as to the rights of the individual against those of society. Decisions by a 5–4 majority were frequent, and the question of the ideological sympathies of future Supreme Court appointees was an issue in the 2000 presidential campaign.

6. *The Future of Conservatism*

A line in a Gilbert and Sullivan operetta asserts that 'every little boy and girl alive, is either a liberal or a conservative.' It seems that conservatism, in the sense of a reluctance to make rapid and wholesale change, is related to a continuing need in the human psyche. It also expresses a general set of preferences about politics and society that undergirds the ideology and program of many political parties and movements. While there may be a broad ideological consensus on the values of democracy and market economics, there will continue to be debate about the relation of the individual to society, the respective demands of freedom and order, the extent to which equality is necessary or desirable, the proper allocation of roles to state and non-state actors, and the need for mediating institutions and local self-government. As long as this is the case, a future role for conservativism as a theory and ideology seems assured.

See also: Conservatism: Historical Aspects; Liberalism: Historical Aspects; Orthodoxy; Republican Party; Tory; Tradition, Anthropology of; Traditions in Sociology

Bibliography

Bennett W J 1993 *The Book of Virtues*. Simon and Schuster, New York

Burke E 1984 (1790) *Reflections on the Revolution in France*. Viking Penguin, New York

Chambers W 1952 *Witness*. Random House, New York

Friedman M 1962 *Capitalism and Freedom*. University of Chicago Press, Chicago
Hartz L 1955 *The Liberal Tradition in America*. Harcourt Brace, New York
Hayek F A 1944 *The Road to Serfdom*. Routledge, London
Hunter J D 1991 *Culture Wars*. Basic, New York
Kirk R 1953 *The Conservative Mind*. Regnery, Chicago
Kristol I 1983 *Reflections of a Neo-conservative*. Basic, New York
Nash G 1976 *The Conservative Intellectual Movement in America*. Basic, New York
O'Gorman F 1986 *British Conservatism*. Longman, New York
Strauss L 1953 *Natural Right and History*. University of Chicago Press, Chicago
Voegelin E 1952 *The New Science of Politics*. University of Chicago Press, Chicago

P. E. Sigmund

		Structure of Society	
		Homogeneous	Plural
Elite Behavior	Coalescent	Depoliticized Democracy	Consociational Democracy
	Adversarial	Centripetal Democracy	Centrifugal Democracy

Figure 1
Lijphart's typology of democratic regimes (reproduced with permission of Yale University Press from Lijphart A 1977 *Democracy in Plural Societies: A Comparative Explanation*, p. 106, Fig. 2)

Consociationalism

The term 'consociational' dates back at least to 1603, when the German scholar and statesman Althusius used it to denote a polity consisting of an association of several societies. In modern political science, 'Consociational democracy means government by elite cartel to turn a democracy with a fragmented political culture into a stable democracy' (Lijphart 1969, p. 216). Although this situation has been labeled in a variety of ways ('Politics of Accommodation,' 'Power-sharing,' '*Pacificatie-democratie*' (in Dutch), '*Proporzdemokratie*' or '*Konkordanzdemokratie*' (in German)), consociational democracy (or consociationalism) is the term that is most widely used. Consociationalism is used empirically to explain stable democracy in deeply divided societies, and normatively to prescribe a solution to societies where democracy is at risk because of social segmentation. In this article these two uses are examined in turn, before discussing some of the criticisms passed on the theory.

1. *The Consociational Explanation*

Deep social cleavages, whether they are based on ideology, religion, class, language, or ethnicity, generally are regarded as unfavorable conditions for stable democracy (see *Cleavages: Political*). The risk is limited when cleavages are crosscutting (e.g., when members of a particular ethnic group belong to different social classes), because this is supposed to generate cross pressures which have a moderating effect on social conflicts (see *Cross-pressures: Political*). The risk is greatest when social cleavages coincide and reinforce each other (e.g., when all members of a particular ethnic group belong to one social class, and all members of another ethnic group to another). However, there are democracies that maintain political stability despite coinciding cleavages. In the late 1960s and early 1970s case studies of several such countries revealed a common pattern: social conflict was not moderated by crosscutting cleavages at the mass level, but by cooperation between the social segments at the elite level (e.g., Lehmbruch 1974, Lijphart 1975, Steiner 1974).

By looking at both the mass level and the elite level, a fourfold typology of democratic regimes was developed, as shown in Fig. 1. The situation of a homogeneous society and competing elites is most conducive to stable democracy. The United States and the United Kingdom serve as examples of such 'centripetal democracy.' One might argue that the combination of social homogeneity and elite cooperation guarantees stability even more, but the lack of democratic choice in such a 'depoliticized democracy' is likely to lead to antisystem opposition. The situation that is most dangerous for stable democracy is where society is divided deeply and the leaders of the social segments compete. Classic examples of this 'centrifugal democracy' are Weimar Germany and the French Fourth Republic. Consociational democracy represents the situation in which deep social divisions are prevented from destabilizing democracy by elite cooperation.

The list of countries that are claimed to be or to have been consociational democracies gradually has expanded and now includes Austria (1945–66), Belgium (from 1918), Canada (1840–67), Colombia (1958–74), Cyprus (1960 and 1963), Czechoslovakia (1989–1993), India (1947 to the late 1960s), Lebanon (1943–75 and after 1989), Luxembourg, Malaysia (1955–1969 and since 1971), The Netherlands (1917–67), and Switzerland (since 1943), with contemporary Canada and Israel being classified as semiconsociational. The

European Union has also been described as consociational. As can be seen from the dates, consociationalism has disappeared in many cases, either because it failed (often because of international interference: Lebanon, Cyprus) or because it was so successful in accommodating the social cleavages that it rendered itself superfluous (e.g., in Austria and the Netherlands). Because of the latter cases, consociationalism is sometimes seen as a transitional phase.

Generalizing from such case studies, the theory of consociational democracy has been developed (e.g., Nordlinger 1972, Lijphart 1977). Lijphart's contribution has been so influential that consociationalism often is equated with his work. He distinguishes four characteristics, each of which should be present for a country to be classified as consociational: (a) government by grand coalition (i.e., including representatives of all social segments); (b) segmental autonomy (federal arrangements in case of geographical concentration of subcultures (see *Federalism*), or subcultural autonomy with regard to schools, etc; (c) proportionality in the electoral system (see *Proportional Representation*), but also with regard to civil service appointments, distribution of public funds for schools, etc; and (d) minority veto (sometimes in the form of qualified majority requirements for certain types of legislation).

In a later offshoot of consociationalism, Lijphart elaborated a distinction between majoritarian democracy, which is characterized by 10 institutional features that concentrate power in the hands of the majority, and consensus democracy, which is characterized by an identical number of features that make decision-making as inclusive of all relevant parties or groups as possible (Lijphart 1999). Although the characteristics of consociational democracy are primarily behavioral whereas the characteristics of consensus democracy are largely institutional, and although consociationalism is also defined by the existence of deep social cleavages, both concepts clearly overlap and many consociational democracies are also classified as consensus democracies.

2. *The Consociational Prescription*

Consociational engineering is recommended as the only viable option for deeply divided countries that seek to achieve or maintain stable democracy. The two cases in which this advice appears to have been influential are Northern Ireland and South Africa (for the recommendations in South Africa, see Lijphart 1985). The 1998 Good Friday Agreement in Northern Ireland, which was endorsed by referendums in Northern Ireland and in the Irish Republic, contains several clearly consociational characteristics (dual premiership, a proportional electoral system (STV), requirements for 'parallel consent' in the legislature, etc.). In South Africa, the transition from apartheid (see *Apartheid*) was facilitated by a consociational interim constitution (including a government of national unity and an electoral system of proportional representation), but in the constitution of 1996 few consociational elements remain.

The prescriptive use of consociationalism has sparked a debate on the conditions that are favorable and unfavorable to the successful adoption of a consociational solution. First, it is more likely that elites refrain from intersegmental competition if they perceive the benefits of competition to be outweighed by its costs. This is the case when no social segment has a majority or can entertain any hope of achieving a majority. However, even if one segment has a (near) majority, its leaders may still prefer to share power with the minority if that minority has inflicted casualties or economic damage on the majority or can credibly threaten to do so, or if the minority's support is indispensable to face an external threat (cf. Nordlinger 1972, pp. 42–53). Second, it is more likely that a consociational solution succeeds when followers agree with their leaders' choice of intersegmental cooperation, or when the elites are relatively immune from intrasegmental competition. The latter is the case when elites control the mass media, or when dense networks of subcultural organizations or patron-client networks make it difficult for competitors to challenge the incumbent elites (cf. Nordlinger 1972, pp. 78–87).

Some other factors that are mentioned as favorable to the emergence of consociationalism are less self-evident (social segments of equal size, a small country, geographical concentration of segments, etc.; cf. Lijphart 1985, pp. 119–28). An important remaining question is whether some social cleavages are more amenable to elite accommodation than others. In view of consociationalism's potential for existing deeply divided countries, the question whether ethnic cleavages constitute a favorable or an unfavorable condition is of particular relevance (e.g., Barry 1975, pp. 502–3).

3. *Critiques of Consociationalism*

The central tenet of consociationalism (i.e., that political stability need not be safeguarded by cross-cutting cleavages at the mass level, but can also be maintained by intersegmental elite cooperation) has gone relatively unchallenged. In addition to controversy over favorable and unfavorable factors, most of the criticisms that have been passed on consociationalism are directed at the classification of some countries as consociational democracies, and at the way in which key concepts have been defined and operationalized, in particular by Lijphart (e.g., Barry 1975, Halpern 1986, Lustick 1997. For Lijphart's

reaction to most criticisms raised until then, see Lijphart 1985, pp. 83–117).

The country whose consociational credentials have been called into question most is Switzerland (e.g., Barry 1975). The debate centers on the question whether Swiss society is deeply divided, as religious and class cleavages cross-cut linguistic divisions, and as the linguistic segments had very few ambitions that brought them into conflict with one another.

Such controversies, it is argued, arise from the vagueness and elasticity with which many of the key concepts of consociationalism such as segmentation are defined (e.g., Halpern 1986). For example, one of consociationalism's four characteristics, grand coalition, includes not only all-party governments (as in Switzerland), but also oversized coalitions, all-party advisory commissions, joining forces in a catch-all party (e.g., India's Congress Party), or 'diachronic coalitions' in which power rotates, as in Colombia between 1958 and 1974. These different mechanisms can still be seen as functional equivalents that all facilitate elite cooperation, however. A more important problem is that elites are unlikely to cooperate all the time, on all issues, and in all arenas. Even in consociational democracies, elites tend to compete in elections, for example. Any classification of elite behavior as adversarial or coalescent therefore must rest on an often-impressionistic aggregation of elite behavior across issues and arenas.

Finally, the democratic quality of consociationalism is sometimes impugned because of the predominance of elites and the absence of opposition. From the perspective of democracy as a device to keep elites accountable and responsive to the masses, this is a pertinent criticism. However, in the deeply divided societies for which consociationalism offers a solution, the most salient distinction is not between elites and masses, but between social segments of a more or less ascriptive nature. In that situation democracy may be better served with a relative emphasis on inclusiveness, and in that respect consociational democracy (and consensus democracy) outperforms adversarial or majoritarian democracy.

See also: Cleavages: Political; Conflict/Consensus; Democracy; Democratic Theory; Majoritarianism and Majority Rule; Pluralism; Religion: Nationalism and Identity

Bibliography

Barry B 1975 Review article: Political accommodation and consociational democracy. *British Journal of Political Science* **5**: 477–505

Halpern S M 1986 The disorderly universe of consociational democracy. *West European Politics* **9**: 181–97

Lehmbruch G 1974 A non-competitive pattern of conflict management in liberal democracies: The case of Switzerland, Austria and Lebanon. In: McRae K D (ed.) *Consociational Democracy: Political Accommodation in Segmented Societies*. McClelland and Stewart, Toronto

Lijphart A 1969 Consociational democracy. *World Politics* **21**: 207–25

Lijphart A 1975 *The Politics of Accommodation: Pluralism and Democracy in The Netherlands*, 2nd edn. University of California Press, Berkeley, CA

Lijphart A 1977 *Democracy in Plural Societies: A Comparative Explanation*. Yale University Press, New Haven, CT

Lijphart A 1985 *Power-sharing in South Africa*. Institute of International Affairs. University of California Press, Berkeley, CA

Lijphart A 1999 *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*. Yale University Press, New Haven, CT

Lustick I S 1997 Lijphart, Lakatos, and consociationalism. *World Politics* **50**: 88–117

Nordlinger E A 1972 *Conflict Regulation in Divided Societies*. Center for International Affairs. Harvard University, Cambridge, MA

Steiner J 1974 *Amicable Agreement versus Majority Rule: Conflict Resolution in Switzerland*. University of North Carolina Press, Chapel Hill, NC

R. B. Andeweg

Constituent Structure: Syntactic

Although we perceive sentences in natural languages such as English, Chinese, Hopi, etc., as mere sequences of words, as in the division of Example (1) into six words, much if not all of language that is of interest involves structure that goes beyond the segmentation of the speech-stream into words. The grouping of words into phrases is called constituent structure.

(1) Old men and women visited Martha.

Constituent structure can be shown to determine the meanings of sentences, for example. Every theory of meaning requires that the account of meaning obeys Frege's Principle of Compositionality (Frege 1952), which states that the meaning of a complex expression is determined by its parts. In this connection, it is noteworthy that the sentence in (1) above is ambiguous, the ambiguity turning on the question of whether the women are old or not. In other words, the sequence *old men and women* has two interpretations, which can be represented with two different structures for the sequence. The two structures can be represented as follows:

(2) Old (men and women) visited Martha.

(3) (Old men) and women visited Martha.

In short, the ambiguity in (1) hinges on whether (a) the conjunction *and* is joining two nouns, the unity of which is modified by the adjective *old*, or (b) the conjunction *and* is joining a noun with an adjective-

plus-noun sequence. The example shows that one cannot specify the meaning of a sentence such as (1) by simply stringing the meanings of the words together. In this sense, natural languages such as English have a great deal in common with most systems of formal logic, which interpret sentences of logic by means of parentheses, which group the 'words' of the logical expression (Allwood et al. 1977). An example shows this for prepositional calculus, the simplest logical language. Beginning logic students are asked to determine the truth of the following expression, simplified here, reflecting different groupings of the logical words:

(4) Not (p or q) is equivalent to (Not p) or q.

Where p and q reflect simple propositions, which can be either true or false.

If two expressions are logically equivalent, they always have the same truth value. To determine whether or not (4) is true, therefore, one must assign p and q all possible combinations of truth values (both true, both false, p true and q false, or p false and q true), and given the semantic rules for *or* and *not*, one sees whether in all four cases, *Not (p or q)* has the same truth value as *(Not p) or q*.

The two sentences do not have the same truth value consistently, indicating that the parentheses, which reflect the grouping of elements, affect the meaning. Natural languages are similar to logical languages, in the sense that the meanings of sentences in both types of language depend on groupings of words—in short, constituent structure. There is a question, however, of whether or not constituent structure in natural language is represented in the same way that it is represented in logic. In logic, when complex sentences are constructed, as in (4), they are constructed in a series of steps, so that *Not (p or q)* is built by first disjoining p and q, and then negating the result. The rules of formation are called the syntactic rules, and hence syntax, in both logic and in natural language, is just the principles of sentence formation, as opposed to semantics, which interprets the results of the syntax. In logic, each syntactic rule is followed by a semantic rule. Hence, when the expression *p or q* is formed, *p or q* is interpreted. When *p or q* as a unit is combined with *Not*, the meaning of *Not* combines with the meaning of *p or q* to determine the meaning of the whole expression *Not (p or q)*. Given that each syntactic rule is followed by a semantic rule, it is unnecessary to actually represent the grouping by means of parentheses in a structure. The process of formation, known as a derivation, is what actually determines the meaning. The parenthesis is simply a way of describing the derivation of this sequence.

An open question, therefore, is whether constituent structure in natural language needs to be represented in an actual structure, or whether the process of forming sentences is enough to determine meanings of sentences. Most linguists assume that constituent structure must be represented by positing structures

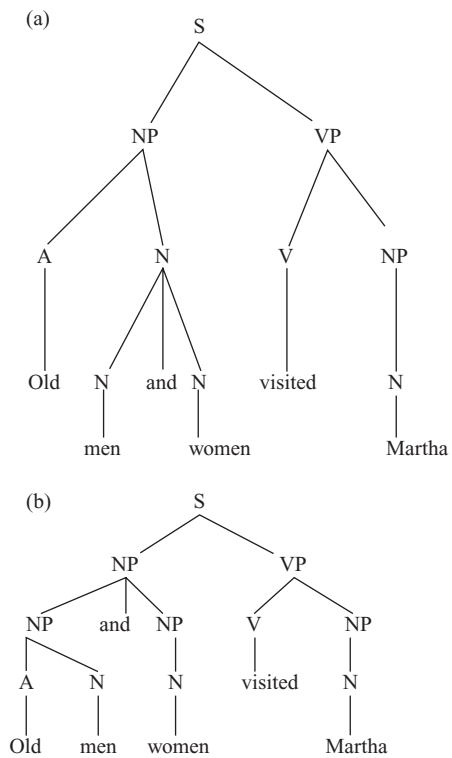

Figure 1
Tree-diagrams of (1)

(a) [S[NP[Adj old][N[N men] and[N women]]][VP[V visited][N Martha]]]

(b) [S[NP[NP[Adj old][N men]] and[N women]][VP[V visited][N Martha]]]

Figure 2
Labeled bracketings of (1)

for sentences known as phrase-markers, which are usually given as tree diagrams or labeled bracketings. The two tree-diagrams for (1), corresponding to both structures, are given Fig. 1, and they are equivalent to the labeled bracketings in Fig. 2.

A word of explanation is in order. NP stands for n(oun) p(hrase), VP for v(erb) p(hrase), and S for sentence. The diagrams in (7) and (8) are called phrase-markers, which are representations of the structures of sentences.

One piece of evidence for the words of the sentence being grouped has been based on meaning. However, there is evidence from a variety of sources for constituent structure. The approach to linguistics developed in the United States since roughly the 1930s (Bloomfield 1933), known as structuralism, has concentrated on the development of rigorous procedures for isolating and classifying the units of language, such

as grammatical categories. The method for classifying words into grammatical categories is based not on meaning but on distribution. A grammatical category (known from traditional grammar as a part of speech) is defined thus:

(5) A grammatical category is a class of elements whose members are mutually substitutable (i.e., interchangeable without loss of acceptability of the result) in a sufficiently wide range of environments.

So the class of, for instance, nouns is the class of elements that can appear in the environments in which nouns can appear, so that, for instance, *books*, *tables*, *chairs*, *sincerity*, *truth*, etc., can appear in enough of the same slots in a sentence to the exclusion of other words such as *from*, *at*, *and*, and *laugh* to justify putting them into a group, and calling this group nouns. (See *Word Classes and Parts of Speech*.)

The definition of grammatical category given in (5), however, which partitions the set of single words into classes, also defines sequences of words into members of the same grammatical class as single words. To illustrate, consider the underlined spaces in (6):

(6) (a) ____might interest me.
(b) I talked about____.
(c) He likes____.

Just as one can substitute the class of nouns that were exemplified above for one another, so one can substitute, for example *big books*, *the books*, *the big books*, *the big books about Nixon* for the word *books* in all of the environments in (8), as well as many others, leading us to conclude, from the definition of grammatical category given in (7), that the italicized sequences in this paragraph are all members of the same grammatical class as the simple noun *books*. We nevertheless must distinguish between simple members of a grammatical category, such as nouns, and the more developed phrasal instances of these grammatical categories. Hence, we use the term 'noun phrase' for the latter, and the simple term 'noun' for the former.

The same considerations that were illustrated above for noun phrases lead us to establish the other phrasal categories, such as verb phrases, adjective phrases, and prepositional phrases. There are a variety of processes that operate in grammar, which can move or delete linguistic elements. An example of movement is given in (7), and an example of deletion is given in (8):

(7) (a) Though he may laugh, it won't matter.
(b) Laugh though he may, it won't matter.

(8) (a) John laughed, and Bill laughed as well.
(b) John laughed, and Bill did____as well.

The (b) versions of (9) and (10) exemplify the movement and deletion of *laugh*. However, sequences of words can undergo these processes:

(9) (a) Though she may finish her homework, it won't matter.
(b) Finish her homework though she may, it won't matter.

(10) (a) John finished his homework, and Bill finished his homework as well.
(b) John finished his homework, and Bill did____as well.

The sequences of elements that undergo such grammatical processes are those sequences we would want to posit as constituents, so the movement process that is exemplified in (11b) is said to move verb phrases, and the deletion process at work in (12b) deletes verb phrases. For a variety of reasons, one would say that a verb phrase is being moved and deleted in (9b) and (10b), respectively, but the verb phrase is one that consists of a simple verb.

There is also evidence from the study of sentence processing in psycholinguistics for constituent structure. In one experimental paradigm (Fodor and Bever 1965), complex sentences (i.e., sentences with various types of subordinate clauses) are recorded, and clicks are superimposed on the recorded sentences by splicing a tape containing the click into a complex sentence. The location of the click was varied with respect to its location in the sentence, occurring from one syllable away from a clause boundary to a number of syllables away. For example, in sentence (11), the click would occur at various points away from the bolded clause boundary:

(11) John told Susan **that** Sally was hungry.

Experimental subjects always reported the click at the clause boundary, regardless of the click's objective location. These results indicate a salience of a structural point within the sentence as a whole.

While there is thus a great deal of evidence for constituent structure for languages such as English, there is less evidence from other languages, such as Warlpiri, which exhibit free word order (Hale 1983). Such languages are usually informally termed 'free word order' languages, but the term does not denote a totally unified class of languages. Japanese (Saito 1985) and Hindi (Mahajan 1990), for example, allow relatively free word order phrases within simple sentences, but require that phrases occur together. Hence, one finds subjects, objects, and indirect objects occurring in any order relative to one another, but subjects must stay together, as must objects, indirect objects, and indeed any phrasal units. Warlpiri, on the other hand, shows much less cohesion, so that the various parts of what are considered for instance, noun phrases, can be separated from one another by material from outside the noun phrase. The cohesion of noun phrases in Japanese and Hindi provides evidence for constituent structure in these languages, and evidence exists that Japanese has a verb phrase, regardless of the free word order among major constituents (Saito and Hoji 1983). The word-order variation in Japanese and Hindi has been analyzed by Saito and Mahajan, respectively, as arising from movement of the noun phrases from a fixed basic position. It is much more difficult to make this type of analysis for the truly free word order languages such

as Warlpiri. Baker (1995, 2000) has some suggestions. Languages with free word order of this sort have extensive case-marking, which would express the grammatical relations of the various noun phrases, so that the case-markers on what would be in English the parts of the noun phrase would signal the semantic units of Warlpiri sentences without making them constituents in the actual syntax.

Putting aside the question of the best treatment of truly free word order languages, we have seen that although a language is usually described as a set of strings (i.e., sequences of words) that is generated by a grammar, the strings are structured into constituents. Chomsky (1957) introduced the distinction between the weak generative capacity of a grammar, which is the set of strings that the grammar generates, and the strong generative capacity of the grammar, which is the set of structures that the grammar generates.

It is paradoxical that we only directly perceive the strings of a language, because most, if not all, of what is interesting about a language is defined in terms of strong generative capacity. We have seen that meaning is determined by structure, a hierarchical notion given by the constituent structure, rather than by linear order of words. Most grammatical operations have been shown crucially to implicate constituent structure, such as movement and deletion operations, as we have seen. Given that constituent structure is not directly perceptible, unlike linear order, it poses severe learnability problems, as pointed out by Pinker (1984), and is usually taken to be innate. This leads to a predicted universality of constituent structure, and this assumption has been used in recent years in work on comparative grammar, such as Pollock's work on French vs. English verb movement (Pollock 1989). Here Pollock studies the placement of adverbs with respect to verbs in both languages, assumes that the adverb is generated in the same structural position in both languages, and attributes the different positions of the verbs with respect to the adverbs to the scope of a verb-movement rule that applies much more generally in French than in English. Pollock shows that the differences between French and English verbs with respect to their linear order vis-à-vis adverbs correlates with other differences between French and English verbs, suggesting that verbs in the two languages differ in their structural positions.

It is also worthwhile to consider an alternative description of English which does not posit constituent structure, a model described in Chomsky (1957) as a finite-state grammar, based on the work of Shannon and Weaver (1963) on information theory. This model views a grammar, in generating a sentence, as a machine which enters successive states when constructing a sentence, and each word that is added to the sentence as it is created puts the grammar into a different state. Each word can be described in terms of its possible continuations, and the machine starts in state S0. A finite state description of, for example, sentence (1) would be as shown in Fig. 3.

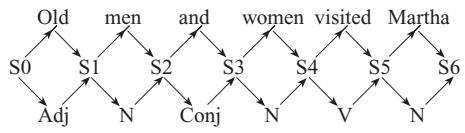

Figure 3
Finite-state grammar representation of (1)

The basic idea of a finite-state grammar is that there are a finite number of states because there is a limit on the number of words in a possible sentence. There does not seem to be, in fact, a limit on the length of a sentence in English or any other language that is due to grammar. For example, it is possible for a sequence of adjectives to occur before a noun in English, as in (12):

(12) Tall, beautiful, intelligent, rich, affable people.

Because there is no bound on the number of adjectives that can occur before a noun, an English sentence can be of any length whatever. Finite-state grammars can account for this obstacle to finiteness by allowing certain word-classes to loop back upon themselves. However, while finite-state grammars can account for iteration of words, they cannot account for iteration of phrases, as in (13), as Chomsky (1957) points out:

(13) John thinks that Fred knows that Mary believes that Max pretended that Joe was hungry.

In this case, the potentially infinite length is due to the containment of a sentence, a phrasal unit, within another sentence.

Chomsky's critique of finite-state grammars appeared shortly after an analogous critique occurred in psychology, in which Karl Lashley (1951) criticized associationism, which viewed behavior as being describable in terms of contiguous responses, so that, for example, going to the refrigerator to get a sandwich would be described, in associationist terms, as putting one step toward the refrigerator door, putting out one's hand, opening it, grasping the refrigerator door handle, pulling it toward one, etc. Lashley noted that such a view did not take into account the notion of an over-arching plan. Lashley was arguing that thought required constituent structure, and the finite-state grammar could be viewed as the linguistic implementation of associationism.

See also: Linguistics: Incorporation; Syntactic Aspects of Language, Neural Basis of; Syntax; Syntax–Semantics Interface

Bibliography

Allwood J S, Andersson L G, Dahl O 1977 *Logic in Linguistics*. Cambridge University Press, New York

Baker M C 1995 *The Polysynthesis Parameter*. Oxford University Press, Oxford, UK

Baker M 2000 The natures of non-configurationality. In: Baltin M, Collins C (eds.) *The Handbook of Contemporary Syntactic Theory*. Blackwell, Oxford, UK
Bloomfield L 1933 *Language*. H. Holt, New York
Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands
Fodor J A, Bever T G 1965 The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior* **4**: 414–20
Frege G 1952 On sense and reference. In: Black M, Geach P (eds.) *Translations from the Philosophical Writings of Gottlob Frege*. Blackwell, Oxford, UK
Hale K 1983 Warlpiri and the grammar of nonconfigurational languages. *Natural Language and Linguistic Theory* **1**: 5–49
Lashley K 1951 The problem of serial order in behavior. In: Jeffress L A (ed.) *Cerebral Mechanisms in Behavior: The Hixon Symposium*. Wiley, New York
Mahajan A 1990 *The A – A-Bar Distinction in Syntax*. Ph.D. dissertation, Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA
Pinker S 1984 *Language Learnability and Language Development*. Harvard University Press, Cambridge, MA
Pollock J-Y 1989 Verb movement, universal grammar, and the structure of IP. *Linguistic Inquiry* **20**: 365–424
Saito M 1985 Some asymmetries in Japanese syntax and their theoretical implications. Ph.D. dissertation, Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA
Saito M, Hoji H 1983 Weak crossover and move alpha in Japanese. *Natural Language and Linguistic Theory* **1**: 245–59
Shannon C E, Weaver W 1963 *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL

M. R. Baltin

Constitutional Courts

Constitutional Courts (CCs) are specialized courts with jurisdiction over constitutional matters.

In constitutionalist systems (constitutionalism) the constitution is considered to be the highest law with normatively binding force for political actors including the lawmaker. While one may detect antecedents in antiquity and in English common law the modern history of the 'higher law' nature of the constitution starts with the constitutions of the American colonies and the American and French revolutions. On a theoretical level, it is the distinction between *pouvoir constituant* and *pouvoir constitué* (Siéyès) that establishes the constitution's binding force over the lawmaker. In more recent constitutional theory the idea of the constitution as guaranty of the basic consensus and the protection of the democratic process and human rights against parliamentary majorities have become more important arguments. Today, most constitutionalist democracies use courts to guarantee the binding force of the constitution. While many countries follow the example invented by Marshall in *Marbury v. Madison* and give this function to the ordinary courts, others (and their number is increasing) use specialized CCs.

1. Constitutional Courts in Historical and Comparative Perspective

1.1 Democracy without Judicial Review

Judicial control of the political system is not a necessary feature of a democratic society. In systems with 'flexible constitutions' like the UK and NZ the constitution is at the disposal of the lawmaker. But even in countries with 'rigid' constitutions that are not at the disposal of simple majorities, the binding force can be conceptualized as a sanctionless imperative for the lawmaker. This was, for example, the case in France until the constitution of the fifth Republic introduced a (moderate) form of constitutional control. In other constitutionalist democratic systems judicial review is expressly excluded (NL, Switzerland on the federal level until 1999), still others know only an organized self-control of the lawmaker during the lawmaking process (Finland). However, the prominent examples of constitutionalist democracies without any form of judicial control are disappearing rapidly: with the constitutional reform of 1999 Switzerland has given its Supreme Court constitutional functions, Finland is considering a similar step and even the bulwark of parliamentarianism, Great Britain, has introduced a (very modest) form of judicial control of Parliament with the 1998 Human Rights Act. Increasingly, some form of judicial control of the political system has become a defining element of modern democracies.

1.2 Constitutional Control as Function of Ordinary Courts: The 'American' or 'Diffuse' Model

The introduction of judicial control by the US Supreme Court in *Marbury v. Madison* (1803) is not based on a constitutional authorization but rather on a logical conclusion drawn from the higher law status of the constitution. This power is, therefore, shared by all courts; there are no special procedures or remedies in constitutional matters; and decisions have no *erga-omnes*-effect. This system (called 'diffuse' in comparative constitutional law) has been adopted widely in legal systems whose constitutions are silent on judicial review, for example, common law countries and Scandinavia. Even though this system is regularly called the 'American' system, the US Supreme Court is in many ways an atypical case: it concentrates (through the *certiorari* system) on cases with constitutional implications, its judges are selected with their constitutional role in mind, and the *stare decisis* rule gives its judgments a *quasi-erga-omnes*-effect. In a socio-

logical sense it acts very much like specialized CCs and is, therefore, included in many comparative works on 'Constitutional Courts' without being one in a technical sense. This is atypical for the 'diffuse' system because for ordinary courts (including supreme courts) in such systems (e.g., in Scandinavia) constitutional questions are only a minor part of their function and judicial review is rarely exercized. Outside the US only a few Supreme Courts of this type (perhaps most notably the Indian Supreme Court) have assumed an active constitutional role.

1.3 Constitutional Courts: The 'Austrian' or 'Concentrated' Model

In contrast to the American model, CCs are the result of conscious constitutional design. Their history starts in 1920 with the Austrian Constitution based on a draft by Kelsen. The power to find laws unconstitutional was reserved to a special constitutional organ (therefore, this system is called 'concentrated'). In its original version only federal and state governments could appeal to the Court. It is no coincidence that CCs were invented by Kelsen. They are, on the one hand, a logical correlate to a hierarchical view of the legal system. On the other hand, Kelsen's pure theory of law tries to shield law against politics and, therefore, Kelsen was an opponent of the American model. The monopolization of constitutional control by a specialized court which could only be mobilized by a limited number of state organs for a constitution that did not contain a bill of rights was as much a barrier to the development of generalized judicial review as the institutionalization of constitutional control. Here again the name-giving case of the 'Austrian Model' is not representative: the vast majority of CCs created later were given much broader powers in constitutional systems with extensive substantive provisions. From a beginning with a much more restricted role than in the American system, today most CCs are more active political players than most courts in diffuse systems. The rise of CCs began after WW II: CCs were important elements in the democratic reconstruction of Germany and Italy. Their model was influential in the third wave of democratization, beginning with Spain and Portugal followed by the transformation in East Europe in which CCs played an important role. Constitutional reforms in other parts of the world similarly introduced CCs (South Korea, Thailand). South Africa gave its CC the unique function of certifying compliance of the new constitution with the principles agreed to in the peaceful abolition of Apartheid. In addition, other countries without a tradition of judicial review introduced CCs in constitutional reforms (*Cour de Arbitrage* in Belgium 1980). France created a restrictive form of preventive judicial control through a *Conseil Constitutionnel* in 1958. It is no coincidence that the rise of the CC-model is closely connected with situations of constitutional transition. In this situation there is a conflict between values and interests protected by old law and the legitimacy of the new constitutional order. In this situation a CC has obvious advantages over the American model because the ordinary judiciary will often be compromized by its involvement in the old order. While it is extremely difficult to transform legal cultures completely just by introducing a new constitution, a specialized court composed specifically for this task and with an institutional commitment to the new constitution can play a decisive role in bringing the inherited legal culture in conformity with the new model.

1.4 Mixed Systems

Some countries have left constitutional cases to the ordinary courts but have recognized the special nature of constitutional litigation by concentrating it in a Supreme Court or special chambers, by creating special constitutional remedies and by decreeing the *erga-omnes*-effects of judicial review. Examples can be found especially in Latin America (because of the competing influences of civil law inheritance and north-American constitutionalism) but also in Europe (Ireland, Estonia, Switzerland after the reform of 1999).

1.5 International Constitutional Jurisprudence

Since the end of WW II judicial control of the political system has gained an international dimension. With a decisive innovative approach the European Convention for Human Rights allowed individuals to sue their governments for violation of human rights in an international forum (a model later followed by the American and recently the African human rights systems). For countries without a bill of rights the ECHR-Organs served as a kind of substitute CC (and the desire to create a local forum to precede international proceedings has influenced the move towards judicial review, e.g., in GB).

With the European Communities, for the first time, an international organization created mandatory and binding court control over community organs and member states. The main function of the European Court of Justice is to safeguard the uniform application of European Law, but it also can be considered a CC of the European Union.

2. Organization and Powers of Constitutional Courts: An Overview

There is a broad range of different solutions for the organization of constitutional control through CCs.

2.1 Composition

In general, the rules for the selection of judges tend to address more openly their political function than those of ordinary courts. They try to square the circle between providing for judicial independence and democratic legitimacy. A majoritarian solution (election by the majority of the day) creates the least theoretical problems for democratic legitimacy but endangers the control function. Attempts to reconcile the two aims include the election with super-majorities (Germany) leading *de facto* to equal appointment rights between government and opposition; the distribution of nominating rights among different constitutional organs (Austria, France); and the inclusion of professional bodies in the nominating process (Italy). As the political element is stronger, appointment to CCs tends not to simply consist of advancement in the judicial hierarchy. Law professors and former politicians are especially highly represented. A relatively strong political influence on the appointment process is counterbalanced regularly by strong guarantees for judicial independence. Appointment or election is either for life or long-term.

2.2 Jurisdiction

In contrast to the 'American' system in which the constitutional question can only be raised incidentally during a case, in a CC system the constitution and constitutional court acts circumscribe the jurisdiction of the Court. There is a broad range of models from very restrictive to very broad opening of constitutional control.

The power of abstract judicial review as a means of objective control outside concrete cases on application from privileged applicants (state organs, entities of federal systems, quorums of members of parliament) is perhaps the one most typical for CCs (and most distinctive compared to the diffuse system). In its most restrictive, purely preventive form, the French *Conseil Constitutionnel* must be engaged to in the short interval between parliamentary vote and promulgation, thus reconciling constitutional control with the maxim *la loi est la loi*. Much more common is repressive abstract review after the law has come into force.

Concrete judicial review allows other courts to refer constitutional questions raised in their proceedings to the CC which holds the monopoly in finding laws unconstitutional.

Access to the CC is most open where it can be mobilized by individuals (constitutional complaint, *amparo*) in case of violations of human rights (usually only after exhaustion of other remedies).

In addition, CCs adjudicate disputes between state organs, the components of federal systems, have quasi-penal functions (impeachment, prohibition of parties) and supervise the election process.

3. Constitutional Courts and Political System

In political theory the role of CCs remains controversial. Their power to overrule the decisions of elected representatives of the people is difficult to reconcile with the theoretical premises of majoritarian democracy. In countries with a CC the debate tends to be less heated than in systems where judicial review has been invented by judges because the constitution makes express provision for jurisdiction. With the extension of judicial control in the last decades, rigorous opposition against it in the name of democracy has declined. Despite the theoretical merits of the majoritarian argument against judicial control over parliament, practical experience has shown it to be a workable reconciliation of majority rule with safeguards against the dangers of majority rule. The international human rights discourse has provided additional legitimacy to control over elected governments exercised in the name of a higher legal order. There remains, however, a permanent tension between the power of CCs to interfere in the political process and the tenets of democracy. But the debate has shifted from the merits of judicial control in the abstract to its proper boundaries. In jurisprudence this discussion is led in the framework of concepts like 'judicial restraint' and 'political questions' while in social science perspective the political role of CCs and their impact are studied. For this debate one of the oldest contributions, Justice Stone's footnote 4 in *Carolene Products* (304 US 144 [1938]), is still a valuable guide: the protection of the democratic process itself and that of 'insular' minorities not able to protect themselves in this process is easier to justify than the enforcement of competing value judgments.

4. The Sociology of Constitutional Courts

CCs (and constitutionally activist Supreme Courts) hold a special place in sociology of law. Their distinguishing feature is almost banal: they are very small. Different from other sociolegal court studies, with CCs one is not dealing with a social group but individually known actors. CCs apply a body of law that typically is more open textured and politically more controversial than normal law. This contrasts with other courts both facilitates and impedes their sociological analysis. On the one hand, CCs offer case-studies of the relationship between law and politics that make the relationship between judicial values and decisions impressively obvious. Therefore, sociology of law treatises often use constitutional jurisprudence to illustrate this relationship. On the other hand, they are so distinct from other courts that the application of sociolegal findings about judicial behavior to them is difficult, and generalizations from constitutional cases to other courts are dangerous.

4.1 Actors

Because of the political role of CCs the politics of elections and appointments is an interesting subject of research that is in this case more directed to concrete selection processes than to generalized study of recruitment patterns. While there are correlations between known judicial background (including party politics, religion, status) and decision there is no simple one-to-one relationship. The guaranteed independence (cf. 3a *supra*) of CC judges has often resulted in disappointment on part of selecting politicians. This can be explained with the self-image of judges and the institutional interests of CCs. For their impact in a political system CCs rely heavily on their perception as neutral arbiters and they would lose legitimacy if they appeared to be mere instruments of their selectors. The high prestige CCs enjoy in many systems (especially compared with politicians) very much depends on this aura of independence.

4.2 Mobilization

One of the major differences between courts and elected policy makers is the fact that courts are reactive, not active participants in the public arena. While this fact is often used as an argument by defenders of judicial review against majoritarians it appears to be of limited significance as a barrier for judicial involvement in politics. In systems with very restricted access to the court like in Kelsen's original model or the French *Conseil Constitutionnel* where only few institutional actors can mobilize the court, it holds true. But access to most CCs tends to be much broader (3b *supra*) and the different avenues combine to bring most politically controversial matters before the courts. When the German *Bundesverfassungsgericht* accepts less than 2 percent of the ca. 5000 complaints received annually one may assume that the court does not so much react but rather selects actively those issues it wants to address. In addition, CCs are inventive in finding ways to involve themselves in politically sensitive subjects without regard to restrictions under procedural rules. In a related development the more active courts do not restrict themselves to decide the case before the court but use *obiter dicta* to make long exhortations on future political action, thus departing from the judge's role of reactively deciding on the past for the politician's role of actively influencing the future.

4.3 Decision-making

Similarly, the binding force of law cited (especially by lawyers) as a restriction of judicial power is less important for CCs than for other courts. Judicial decision-making always has a political dimension but it is subject to constraints intended to promote a specific juridical rationality and to hinder idiosyncratic decision making by judges. These constraints appear to be less effective with CCs. Their decisions are less strictly programmed by law. While some constitutional law provisions (organizational norms, e.g., on elections) can be very specific, those texts that are relevant today in most constitutional proceedings—fundamental rights or constitutional principles—are very open textured (and traditionally in constitutional jurispurdence have been expanded further, rather than narrowed). CCs have to balance broadly worded rights against similarly broadly worded infringement-clauses and find the solution regularly by balancing private rights and public interest against each other under extremely open concepts like the principle of proportionality which give widest possible room of interpretation.

CCs have transcended the classical boundaries of judicial politics in another way. In principle, courts are different from political actors because they are restricted to yes/no decisions (lawful/unlawful) over the past which is a powerful incentive for judicial restraint (e.g., if a CC had only the alternative of either declaring a tax law void and risking public bankruptcy or accepting the law as constitutional). CCs have found ways around this problem. For example, they refrain from declaring laws void but order the lawmaker to change them in a given time. Or, they use the interpretation of laws or *obiter dicta* to enforce political compromises between proponents and opponents of a law under scrutiny. Such flexible forms of judicial decision free the court from the stark yes/no alternative and broaden its potential to actively involve itself in policy-making by influencing future policies, building coalitions, and instigating political compromises. A promising strategy for measuring judicial activism appears, therefore, to be the study to what extent CCs have given up the ideal-typical judicial style of decision making for a political one.

See also: Bill of Rights; Civil Liberties and Human Rights; Civil Rights; Constitutionalism; Constitutionalism, Comparative; Courts and Adjudication; Democratic Transitions; Fundamental Rights and Constitutional Guarantees; Supreme Courts

Bibliography

Beatty D (ed.) 1994 *Human Rights and Judicial Review: A Comparative Prospective*. Nijhoff, Dordrecht, Germany

Brewer-Carias A 1989 *Judicial Review in Comparative Law*. Cambridge University Press, Cambridge, UK

Bryde B-O 1999 *Constitutional Courts in Constitutional Transition*. Liber Amicorum van Houtte, Acco, Leuven, Germany

Cappeletti M 1971 *Judicial Review in the Contemporary World*. Bobbs-Merrill, Indianapolis, MI

Ely J H 1980 *Democracy and Distrust*. Harvard University Press, Cambridge, MA

Favoreu L 1986 *Les Cours constitutionnelles*. Presses Universitaires de France, Paris

Grewe C, Ruiz Fabri H 1995 *Droits constitutionnels Européens*. Presses Universitaires, Paris
Kelsen H 1928 La Garantie juridictionnelle de la Constitution. *Revue du Droit Public et de la Science Politique* **45**: 197–257
Klug H 2000 *Constituting Democracy. Law, Globalism and South Africa's Political Reconstruction*. Cambridge University Press, Cambridge, UK
Kommers D P 1989 *The Constitutional Jurisprudence of the Federal Republic of Germany*. Duke University Press, Durham, NC
Landfried C (ed.) 1988 *Constitutional Review of Legislation*. Nomos, Baden-Baden, Germany
McWhinney E 1986 *Supreme Courts and Judicial Law-making*. Nijhoff, Dordrecht, Germany
Sajó A 1995 Reading the invisible constitution: Judicial review in Hungary. *Oxford Journal of Legal Studies* **15**: 253–67
Smith E (ed.) 1988 *Constitutional Justice under Old Constitutions*. Kluwer, The Hague
Yang K 1993 Judicial review and social change in the Korean democratizing process. *American Journal of Comparative Law* **41**: 1–8

B.-O. Bryde

Constitutionalism

Constitutionalism must be distinguished from three somewhat related terms, a constitution, a constitutional text, and constitutionism. Every country, James T. Shotwell was fond of saying, has a constitution, most countries have a constitutional text, but very few have constitutionalism. And, we might add, almost all public officials claim to practice constitutionism.

Narrowly defined, a nation's constitution refers to its arrangement of public offices and powers as well as individual and group rights. More broadly, as Aristotle came to use the word, it signifies a way of political life, what modern theorists are apt to call a 'constitutional order,' which can vary from that of Mao's China and Stalin's Soviet Union, on the one hand, to the more benign systems of Australia, Canada, the Federal Republic of Germany, and the USA.

A constitutional text identifies the document (or set of documents) that supposedly spell out the nation's basic political principles, institutional arrangements, modes of selecting public officials, and the rights and duties of private citizens. The contents of these documents vary widely, as does their authority. At one extreme would be Stalin's constitutional text, a pious fig leaf; at the other extreme are the constitutional scripts of countries such as France, Ireland, Norway, and the USA, whose terms public officials usually try to follow.

The awkward neologism constitutionism refers to adherence to the terms of the constitutional text or those of the broader constitutional order. Because the basic values of both a constitutional order and of a constitutional text can range from those of totalitarianism to a loose libertarian political system, officials of most governments, even Chile's, under Agosto Pinochet, can often pretend to be practicing what they consider constitutionalism, while in fact they are only following constitutionism, that is, being true to the constitutional order.

The concept constitutionalism, however, differs from constitutionism in demanding more than adherence to any given constitutional text or order. Constitutionalism is a normative political theory that endorses a special kind of constitutional order, one that counts among its practical tenets that every exercise of governmental power, whether representing the will of a single leader, an elite, or an overwhelming majority of citizens, is subject to important substantive limitations. In short, constitutionalism contends that the central value of any society must be human dignity. It follows that there are some things that government cannot do, no matter how faithfully it follows procedures specified by a constitutional text and a larger constitutional order, or even if its actions perfectly mirror the deliberate judgment of a charismatic fuehrer, benevolent junta, or the majority of voters. Obviously, this normative theory is incompatible with most forms of authoritarian government. Moreover, it co-exists uneasily with its usual partner, representative democracy.

Constitutionalism is closely related to the concept of limited government. Indeed, constitutionalism can be viewed as one among a variety of species of that much older genus (Gordon 1999). There are many possible ways of restricting political power, countervailing force being among the more obvious. Because the people of Israel had a deserved reputation for waging war, the Roman Empire tried to avoid open conflict by according Jews certain special civil rights, exempting them from requirements to worship pagan gods and to pay taxes to maintain local shrines, and also according them the privilege of sending money to the Temple at Jerusalem. Theology may provide another check. Anyone reading Homer must be struck by the irascible Olympians' jealous responses to human hubris. Later in Western history, the prophets tried to impose restraints on the kings of Israel, as did the Catholic Church on European monarchs through the middle ages. Custom may provide an additional limitation, recognized in a ruler's promise to do justice to his subjects even as his father did justice to their fathers in return for his subjects' pledges of fealty.

The possibility of armed conflict may give a practical impetus to the acceptability of certain ideas, and custom and theology are typically major ingredients of a political culture. But again, their collective relationship to the norms of constitutionalism depends on the substance of the customs, theology, and political culture. Certainly, neither the frequently malevolent behavior ascribed to Homer's gods nor the murder

and rapine celebrated in 'heroes' like Achilles and Odysseus would support the ideals of constitutionalism. For them, human beings were typically mere things to be exploited and discarded. On the other hand, the foundations of modern constitutionalism rest on the theological claim of Judaism and Christianity that all men and women are made to the image and likeness of an all-loving God—a basis that does not prevent people of other (or no) religions from availing themselves of constitutionalism's benefit of controlling government.

What we know as constitutionalism also grew out of the efforts of Western Europeans to curb absolute power by creating political institutions that could express customs and moral codes. Thus, one can look on constitutionalism as a political theory made operative through the institutionalization of certain norms into the political system. And, although its medieval and early modern roots are visible, efforts to restrict government began much earlier in Western history. Without doubt, the citizens of ancient Athens did not have a concept of universal human dignity as rich as that of people living in modern constitutional democracies; still, their constitutional order did not accept government as unlimited in authority. Athens had in place institutions that could check even the Assembly, a legislative body in which every citizen had a right to speak and vote on any proposed public policy. Among these institutions was a version of judicial review: a citizen could challenge the constitutionality of any decree the Assembly had approved by filing a claim that would be heard by a jury, ranging in size from 500 to 1500, randomly selected from those citizens who were willing to sit. In that forum, the issue of the decree's congruence with the polity's constitution could be further debated, decided, and voided if deemed to violate the constitutional order (Hansen 1991).

In the modern world, the open processes of representative democracy are often hailed as the most effective protector of individual rights against government. 'The mass of the people,' Thomas Jefferson wrote John Taylor in 1816, 'are the safest depositary of their own rights.' The reasoning is that, if allowed access to a free press, to speak freely among themselves, to organize for political purposes, and to vote in honest elections, citizens will zealously guard that deposit. Democratic theories thus stress that a political system should choose its principal decision makers through open political processes and, in turn, those decision makers should follow similarly open processes in making public policies and then standing for periodic re-election. Democratic theorists preach that, because the reach of governmental power and individual rights are contestable, the legitimacy of governmental action is better resolved through elected representatives responsible to the people than by an elite's judging a policy's congruence with its understanding of the polity's underlying norms. The voters can affirm or reverse constitutional interpretation at the next election.

Constitutionalism shares concerns about process as well as many of the substantive values democratic theorists endorse. Constitutionalists, however, are more pessimistic about human nature; they are constantly concerned, even obsessed, with the human penchant to act selfishly and abuse power. They fear that even the 'mass of the people' are sufficiently clever to oppress others without hurting their own rights. As James Madison wrote to Jefferson:

> In our Governments the real power lies in the majority of the Community, and the invasion of private rights is chiefly to be apprehended, not from acts of Government contrary to the sense of its constituents, but from acts in which Government is the mere instrument of the major number of its constituents (Meyers 1973, p. 206).

The fundamental dispute between constitutionalist and democratic theorists centers on how best to protect rights against government. For democrats, the right to political participation is the fundamental right because is it 'protective of all other rights.' Constitutionalists respect the right to political participation but are more apt to view 'the right to be let alone' as more fundamental. They tend to view democratic political processes as 'at best regulated rivalry … Political power rapidly accumulates and becomes unequal; making use of the coercive apparatus of the state and its law, those who gain the advantage can often assure themselves of a favored position.' Private funding of political campaigns aggravates these problems (Rawls 1971, p. 226).

Constitutionalism, therefore, demands 'a system of regularized restraints [beyond free elections] upon those who exercise political power' (Friedrich 1964, p. 17). A constitutional text with a bill of rights and judicial review are the most visible but not the sole constitutionalist restraints. Many constitutionalists, like most democratic theorists, also emphasize the importance of political culture. As James Madison told the First Congress when he introduced the Bill of Rights.

> It may be thought that all paper barriers against the power of the community are too weak to be worthy of attention … [Y]et, as they have a tendency to impress some degree of respect for them, to establish the public opinion in their favor, and rouse the attention of the whole community, it may be one means to control the majority from those acts to which they might otherwise be inclined (*Annals of Congress*, I, 440f, 1789).

Separate institutions competing for power (Jones 1990) and federalism are also typical constitutionalist weapons. Democrats are likely to speak of the 'sovereignty of the people,' while constitutionalists tend to regard the very concept of sovereignty as a dangerously fuzzy concept. They prefer a political

system in which power is so dispersed that the supreme law making power, if it exists at all, will be only found in the midst (and mist) of jealous interactions among popular and elitist institutions.

No nation has fully adopted constitutionalism any more than any nation has fully embraced democracy. Indeed, with its distrust of human nature and insistence on the primacy of individual rights, pure constitutionalism might make governance in the modern world impossible. So, too, unmodified democracy might soon commit suicide by allowing today's majority to restrict the minority's political rights and tomorrow's new majority, in turn, to restrict tomorrow's minority (though it had been part of yesterday's majority) until finally only a small junta or single dictator were governing. The most common arrangement in the modern world is becoming constitutional democracy, a marriage between representative democracy with its notion that the people should rule through their freely chosen representatives, and constitutionalism, with its norm that if the people govern, they should not govern too much. Like most marriages this one constantly endures tension and often experiences conflict, even turmoil.

See also: Democracy; Democracy: Normative Theory; Democratic Theory; Judicial Review in Law; Legal Issues: Public Opinion; Political Culture; Rule of Law

Bibliography

Casper G 1986 Constitutionalism In: Levy L et al. (ed.) *The Encyclopedia of the American Constitution*. Macmillan, New York, Vol. II, pp. 473–84

Dahl R A 1989 *Democracy & Its Critics*. Yale University Press, New Haven, CT

Elster J, Slagstad R (eds.) 1988 *Constitutionalism and Democracy*. Cambridge University Press, New York

Friedrich C J 1964 *Transcendent Justice*. Duke University Press, Durham, NC

Gordon S 1999 *Controlling the State: Constitutionalism from Ancient Athens to Today*. Harvard University Press, Cambridge, MA

Hansen M H 1991 *The Athenian Democracy in the Age of Demosthenes: Structure, Principles and Ideology*. B. Blackwell, Oxford

Jones C 1990 The separated presidency. In: King A (ed.) *The New American Political System*, 2nd edn., American Enterprise Institute, Washington, DC, p. 3

McIlwain C H 1947 *Constitutionalism, Ancient and Modern*. Rev. edn, Cornell University Press, Ithaca, NY

Meyers M (ed.) 1973 *The Mind of the Framer*. Bobbs-Merrill, Indianapolis, IN

Murphy W 1993 Constitutions, constitutionalism and democracy. In: Greenberg D et al. (eds.) *Constitutionalism and Democracy: Transitions in the Contemporary World*. Oxford University Press, New York, pp. 1–25

Pennock J, Chapman J (eds.) 1979 *Constitutionalism*. New York University Press, New York

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Sartori G 1987 *Democratic Theory*. Praeger, New York

W. F. Murphy

Constitutionalism, Comparative

1. Introduction

The idea and practice of comparative constitutionalism is a new field propelled by the political and technological changes of the late twentieth century (Jackson and Tushnet 1999), yet it is rooted in antiquity, in the methodology of Aristotle and his students at the Lyceum. While not a well established area of academic study (ACLS, 1990) comparative constitutionalism exists both as a normative and legal practice as well as a form of comparative political and legal analysis.

Comparative constitutionalism is now widely practiced by both constitution-makers and constitutional interpreters. Comparative constitutionalism became prominent in the late twentieth century, expanding exponentially in practice with the wave of constitution-making that accompanied the end of the Cold War. While comparative constitutionalist practices are initially applied in the process of constitution making, in which participants seek to apply models and experiences from other contexts, it is perpetuated in the debates over the meaning of each new constitutional provision from the moment of its adoption. This practice is institutionalized in the jurisprudence of many constitutional and supreme courts who through their decisions effectively engage in a transnational comparative constitutional discourse on constitutional meanings and functions.

Academically, a single definition of comparative constitutionalism is even more difficult to formulate. In philosophy, political science, and law, comparative constitutionalism is used both normatively to evaluate constitutional provisions, and empirically to describe, contrast, and reflect upon the consequences of the different rights and institutional provisions of constitutions around the globe. As a normative exercise, the survey and comparison of constitutions remains close to the original Aristotelean method of comparing different constitutional provisions and institutions in order to define a better form of government. This approach has been supplemented with historical accounts which explore the development of constitutionalism through time (Pennington 1993) and across national borders (van Caenegem 1995) A normative emphasis is also prevalent in a range of

studies that focus on some of the foundational questions underlying the adoption and implementation of constitutions: the relationship between constitutionalism and democracy (Elster and Slagstad 1988; Habermas 1996); the origin of the notion of fundamental rights and the role of civil society (Mamdani 1991); and even the idea of a diversity of forms of constitutionalism that might be achieved through demands for cultural recognition in a post-colonial age (Tully 1995). While increasingly of an interdisciplinary nature and central to the idea of constitutionalism, these studies tend either to focus on the conceptual and theoretical issues or to adopt a country by country approach rather than a comparative perspective (Gordon 1999, Pennock and Chapman 1979).

However, with the emergence of globalization as both an empirical phenomena and as an object of study (see *Globalization: Legal Aspects*), the debate on constitutionalism has taken a comparative turn (Singh 1989). Comparative constitutionalism in this latter sense has two distinct foci: on the one hand there is a focus on comparing the different institutions that are employed to secure democratic forms of government (Sartori 1994), or their role in shaping particular social issues such as identity and diversity (Rosenfeld 1994, Habermas 1998). On the other hand there is an emerging focus on constitutional borrowing, adaptation, and hybridization (Badi 2000, Klug 2000).

The dispersal of sovereignty, which is inherent in the process of globalization, provides further difficulties for the definition of constitutionalism, as may be witnessed in the debate over citizenship and democracy in the European Union, or in the ideas of shared sovereignty which have been introduced into the debate over the future of Jerusalem. The post-twentieth century emphasis on the negotiated settlement of disputes over governance and power, whether at an international or local level, provides fertile ground for debating constitutional alternatives. Here, comparative constitutionalism becomes a pragmatic alternative to continuing conflict.

2. *The Global Context*

Alarm over the Americanization of the world became a central theme for many critics of globalism in the 1990s. At the same time, many lawyers participated in and celebrated the new emphasis on the 'rule of law,' presuming the global adoption of law and legal processes as a precondition for successful transitions from authoritarian regimes and command economies into a new world order. The adoption of bills of rights and the empowerment of courts, which characterized many of these political transitions, led many to believe these developments indeed reflected the embrace of American values in the post-Cold War era. Bruce Ackerman heralded the emergence of a 'world constitutionalism,' arguing that the 'enlightenment faith in written constitutions is sweeping the world' (Ackerman 1997). Another prominent United States constitutional scholar, Cass Sunstein, argued that the 'right kind of constitution could play an important role in fueling economic development and democratic reform; indeed, under current conditions, it may be indispensable to them' (Sunstein 1993).

With the widespread adoption of constitutional forms genetically tied to the innovations of the American revolution, it is not surprising that both advocates and detractors of the American experience assume the United States to be the hegemonic model. However, American lawyers and scholars who participated in a billion-dollar program in the 1990s to promote the rule of law around the world found themselves in competition with alternative constitutional models. Although the bulk of activity in this 'rule of law' movement focused on judicial and legislative reform designed to promote market economies, it is important to recognize that a central product of this dual process of political transformation and rule of law advocacy was fundamental constitutional reform in half the world—in approximately 104 of the 188 members of the United Nations. Of these, a staggering 78 percent, or approximately 81 countries, adopted completely new constitutions.

Given a common belief in the United States Constitution as a 'source of inspiration and ideas … studied and appraised by scholars, invoked by publicists, adopted or adapted by revolutionaries, nation builders and statesmen' (Henkin 1989), many of those who joined the multitude of legal expeditions to assist in constitution making in Eastern Europe, Africa, and elsewhere were surprised to find themselves confronted with alternative constitutional models. As exporters, advocates of the American model found that instead of enjoying an unassailable status, they faced open competition from advocates of the German, Canadian, Indian, or other constitutional experiences. Given that these alternative constitutional experiences are to a large extent themselves based on the American model of constitutionalism, it seems counterintuitive to regard the dominant model, the American experience, simply as one of a number of alternative models. Yet this is the case, and may be explained by the emergence of comparative constitutionalism. Constitution-makers, interpreters, and litigants actively compared their own conditions and experiences with the current and historical experiences of other countries, in order to advocate one or another constitutional solution, whether a particular constitutional institution, the specific form of a constitutional clause, or the meaning of particular constitutional language. These practices became increasingly common through the 1990s as states around the globe engaged in one of the most profound processes of constitutional and political

reconstruction ever. It is this context that has framed the new intellectual interest in comparative constitutionalism.

3. Constitutions, Constitutionalism, and the Politics of Comparison

Comparative constitutionalism has shifted its emphasis from being a philosophical debate over the nature and form of good government to a more legal focus on the application and meaning of particular constitutional structures and rights, as well as on the role of the judiciary in the project of limited government (Sajo 1999). While the question of judicial review, and particularly the debate over the relationship between constitutionalism and democracy, dominated much of the academic discourse in the twentieth century, the changing social and political context has brought greater attention to the interaction of constitutional ideas and institutions across different legal systems and national boundaries (Schedler et al. 1999, Hanafin and Williams 1999, Teubner 1997). This new trend is beginning to define a new field of comparative constitutionalism in theory and practice.

Comparative constitutionalism originated in western thought with Aristotle. In the *Nicomachean Ethics* (Aristotle 1996) he made a case for the comparative study of constitutions, arguing that by studying the 'constitutions we have collected … we shall perhaps be more likely to see which constitution is best, and how each must be ordered, and what laws and customs it must use' (Aristotle 1181b, 13–22). Although only the *Constitution of Athens* (Aristotle 1996) remains from this early venture, the product of this analytical exercise is presented by Aristotle in *The Politics* (Aristotle 1996) and remains a principle source for the study of comparative government, which still largely aims to construct a normative definition of institutional and political practices that are necessary in order to secure the good life.

The Aristotlean notion of constitutionalism, combining both a distinctive political order and the idea of a political regime in general is, according to Carl Friedrich, the source of the 'modern idea of a constitution as the organization of a government, its offices, and the relation of the offices' (Friedrich 1968a). For Friedrich, however, this understanding, which recognizes all forms of government as exhibiting one or other form of constitutionalism, was inadequate, for it incorporated dictatorial forms of governance within the idea of constitutionalism. Instead, Friedrich (1968b), along with McIlwain (1947) and Wormuth (1949), emphasized the relationship between constitutionalism and individual liberty. For them, the essence of constitutionalism is not the organization of power but rather the idea of limited power. From the idea of restraints on power to the separation and division of power, the focus among Anglo-American theorists through the middle decades of the twentieth century was on the guarantee of individual liberty which they took to be the product of various means of limiting state power (Pound 1957).

Writing first in the shadow of Nazi Germany, then in the aftermath of World War II, and during the Cold War, Friedrich shifted the focus away from the comparative perspective inherent in the Aristotlean approach by posing the question: what is the political function of the constitution? From this starting point, Friedrich identified the protection of the individual member of the political community as the primary object of constitutional government, thus redefining constitutionalism as a 'modern phenomena' whose central concern is the practice and theory of limited government (Friedrich 1968b). This focus empowered western theorists to reject constitutional forms, particularly the 'symbolic' state socialist constitutions that failed to limit state power, as aberrations of the constitutionalist tradition. Thus constitutionalism was once again given a normative form, but at the cost of limiting the concept only to those constitutional arrangements that successfully limited and controlled state power.

The failure of post-colonial constitutions in Africa and other parts of the Third World, despite the inclusion of clauses formally limiting, separating, and controlling state power, produced an alternative, 'socio-legal' approach to constitutionalism. Legal academics and political scientists argued that it was important to understand the economic and social context that determined the effective constitution of power in these newly independent states (Ghai 1991). Taken further, this analysis seeks to demonstrate how the underlying colonial and neocolonial conditions of social organization, particularly the dual legal system and lack of land tenure security, restrict the capacity of civil society in Africa (Mamdani 1996, Shivji 1998). Without changes in these conditions, it is argued, the notion of legal rights against the state and restraints on state power are ephemeral. Despite this disillusionment with formal constitutions and constitutionalism, struggles for democracy and interethnic strife in many developing countries has fueled a continual search for understanding the place constitutionalism may play in the reconstruction of post-colonial states. Yash Ghai has sought constitutional solutions to problems of ethnic strife, both in his practice as a constitutional advisor and in his academic work (Ghai 2000). He has also applied this socio-legal perspective to analyzing the role of constitutional formulations produced to define the relationship between the central government in China and a reintegrated Hong Kong (Ghai 1997).

With the protection of the individual as its central goal, and the restraint of political power as its essential feature, the study of constitutionalism lost steam in the Anglo-American world. It split in two distinct directions, one legal, the other political. On

the one hand, an interest in judicial review became the central concern of legal constitutionalists. Here the idea of comparative constitutionalism became, and remains so to some extent, equated with comparative constitutional law (Capelletti and Cohen 1979). In this context the comparative approach either focuses on the specific problem of judicial review, engaging in a comparison of the process of judicial review, exercised by courts around the world (Cappeletti 1989), or compares the case law of different jurisdictions in order to explore different means of the constitutional resolution of particular issues from the protection of property rights (Allen 2000) to abortion and hate speech (Jackson and Tushnet 1999). On the other hand, the fiscal crisis of the state and political conservatism in the late 1970s and 1980s led to a direct political attack on the role of the state in society. Despite challenges of cultural imperialism by those advocating an 'Asian model' of autocratic paternalism, or by those who argued that 'constitutionalism is largely a cultural phenomena and not simply the product of properly designed institutions and structures of government' (Franklin and Baun 1995), it was the emphasis upon the need to restrict state power that created an intellectual and political environment in which the constitutional protection of individual rights and the institution of judicial review became the dominant features of a post-Cold War international political culture.

The emphasis on the legal form of constitutionalism, and on judicial review in particular, shifted the focus onto the role of the judiciary, encouraging both a practical emphasis on judicial reform as well as in interest in the global expansion of judicial power (Tate and Vallinder 1995). However, at the same time, the failure and even collapse of governments in Africa and the Asian economic crisis in the 1990s brought a renewed focus on the question of power and the need for effective governance. This focus raised the question of how the constitutional allocation of power should in fact empower the state to achieve a level of social coordination necessary for individual security and freedom.

4. *The Problem of Governance and the Role of Comparative Constitutionalism*

If constitutionalism is, historically, the struggle to regulate political competition, then it is unsurprising to discover that different constitutional options in fact reflect continuing ideological and political alternatives. The politics of constitution making (Elster 1995) is thus festooned with the claims of competing political advocates who project their ideological perspectives and political goals into constitutionalist forms (Klug 2000). While the post-World War II response to the violation of human rights produced a movement that could insist on the recognition and promotion of human rights through both international and domestic fora, the re-emergence of a nineteenth century liberalism, in the guise of a 1980s neoliberalism, had profound implications for the politics of constitution-making in the post-Cold War era. In this context, Friedrich von Hayek's conception of politics as threatening a liberal order that is simultaneously complex and delicate (Hayek 1960) provided a central argument for those who sought to shrink the state. Although democratic theorists and constitutional scholars had long highlighted the anti-democratic implications of constitutional review, the neoliberal project emphasized the 'rule of law'—requiring an independent judiciary with powers of judicial review—and demanded a constitutionalization of many other institutions and areas of governance as a means to limit the destabilizing impact of politics.

Despite the gains made by neoliberalism at the end of the Cold War, the politics of constitution making remained eclectic. Although the vast majority of states undergoing reconstruction through processes of constitutional change in this period seemed to accept the liberal paradigm of individual human rights and multiparty democracy, this did not preclude the simultaneous inclusion of a range of alternative constitutional elements, including socioeconomic and cultural rights, which reflect alternative traditions (see *Civil Liberties and Human Rights*). Significantly, the dominant liberal democratic tradition itself contains different trends, sometimes complementary, sometimes contradictory, advanced by different social forces and reflecting a plurality of interests. Of these, the two trends that had the most direct impact on processes of post-Cold War state reconstruction emphasized liberty and equality. While there is a significant overlap in liberalism's ideological commitment to liberty and equality, emphasis on one or the other provides a range of alternatives within the liberal democratic tradition. These extend from classic nineteenth-century liberalism, with its emphasis on individual freedom and property rights, to the claims of the democratic collectivity inherent in the social democratic liberalism of the post-World War II era. While both aspects of this tradition emphasize electoral democracy and the protection of individual rights, they also contain contradictions with important consequences for the shape and role of the state.

While the struggle for equality, whether aimed at economic or racial inequalities, elicited the power of the state to address entrenched private power, the struggle for individual freedom has eschewed the state, often characterizing governmental power as the very source of oppression. These different responses to the state have obvious consequences for the structure of government, the most dramatic being the emphasis on the need to downsize or limit the role and capacity of the state. Combined with a reaction against the commanding role held by the state in socialist societies

and strengthened by the fiscal crisis of the state in western democracies, this tendency significantly influenced the international political culture that framed the post-Cold War process of state reconstruction. By contrast, the classic elements of a social democratic state, including the state's social welfare orientation and regulatory role in the relationship between capital and labor, including the creation of corporatist institutions which gave organized labor an important voice in social organization, fell into political and ideological disrepute and began to be dismantled.

In this context, comparative constitutionalism provides an opportunity to rethink many of the longstanding assumptions about traditional constitutional forms so as to understand their role in the construction and maintenance of different constitutional orders. There are three sets of values that frame traditional notions of constitutionalism: (a) federalism, or the spatial division of power, (b) the separation of powers between different branches of government, (c) the notion of constitutional rights. Each of these address different aspects of the 'problem' of power—its allocation, application and restriction—within the nation state. While these constitutional values, with their diverse historical origins, have become increasingly 'universalized,' their application in increasingly varied historical, cultural, and political contexts has produced a diversity of approaches and understandings which provide a diverse 'global text' for political actors, constitution makers, constitutional interpreters, and litigants.

5. *The Practice and Theory of Comparative Constitutionalism*

As a field of empirical academic research and legal practice, comparative constitutionalism suffers many of the methodological challenges that have encumbered the fields of comparative law and politics. Apart from the more obvious difficulties of language and the different legal traditions which plague the comparative law project, each country's historical, social, political, and economic contexts require attention in studying the origins and meanings of particular constitutional forms. More recently, the increasing comparative jurisprudence of constitutional and supreme courts in a number of countries has broadened the scope of comparative constitutional law, and has raised questions about the theory and practice of constitutionalism in the context of a globalizing world. On the one hand, there is a growing attempt to formulate constitutional theories to throw light on the 'increasingly cosmopolitan character of constitutional interpretation' (Choudhry 1999). On the other hand, there is a call for Anglo-American lawyers to recognize the need for a broader legal theory of the state or civil-law type *Staatslehre* (Dyzenhaus 1998) as an essential step towards the formulation of an adequate theory of the relationship between law and political power. It is this relationship that is at the heart of the idea of constitutionalism and the object of research and practice in the growing field of comparative constitutionalism.

See also: Aristotelian Social Thought; Aristotle (384–322 BC); Constitutional Courts; Constitutionalism; Fundamental Rights and Constitutional Guarantees; Globalization: Legal Aspects; Governments; Judicial Review in Law; Sovereignty: Political

Bibliography

Ackerman B 1997 The Rise of World Constitutionalism. *Virginia Law Review* **83**: 771

Allen T 2000 *The Right to Property in Commonwealth Constitutions*. Cambridge University Press, Cambridge, UK

American Council of Learned Societies 1990 *The ACLS Comparative Constitutionalism Project: Final Report* (ACLS Occasional Paper, No. 13)

Aristotle 1996 *The Politics and the Constitution of Athens*. Eversons (ed.) Cambridge University Press, Cambridge, UK

Badi 2000 *The Imported State: The Westernization of the Political Order*. Stanford University Press, Stanford, CA

Banting K G, Simeon R (eds.) 1985 *Redesigning the State: The Politics of Constitutional Change in Industrial Nations*. University of Toronto Press, Toronto, Canada

Cappelletti M 1989 *The Judicial Process in Comparative Perspective*. Clarendon, Oxford

Cappelletti M, Cohen W 1979 *Comparative Constitutional Law*. Michie, Charlottesville, VA

Choudhry S 1999 Globalization in search of justification: Toward a theory of comparative constitutional interpretation. *Indiana Law Journal* **74**: 819–92

Dyzenhaus D (ed.) 1998 *Law as Politics*. Duke University Press, Durham, NC

Elster J 1995 Forces and mechanisms in the constitution-making process. *Duke Law Journal* **45**: 364

Elster J, Slagstad R (eds.) 1988 *Constitutionalism and Democracy*. Cambridge University Press, Cambridge, UK

Franklin D P, Baun M J (eds.) 1995 *Political Culture and Constitutionalism*. Sharpe, Armonk, NY

Friedrich C J 1968a *International Encyclopedia of the Social Sciences*. Macmillan, New York

Friedrich C J 1968b *Constitutional Government and Democracy* (4th edn.) Blaisdell, Waltham, MA

Ghai Y 1991 The theory of the State in the Third World and the problem of constitutionalism. *Connecticut Journal of International Law* **6**: 411

Ghai Y 1997 *Hong Kong's New Constitutional Order*. Hong Kong University Press, PRC

Ghai Y (ed.) 2000 *Autonomy and Ethnicity*. Cambridge University Press, Cambridge, UK

Gordon S 1999 *Controlling the State*. Harvard University Press, Cambridge MA

Greenberg D, Katz S N, Oliviero M B, Wheatley S C (eds.) 1993 *Constitutionalism and Democracy: Transitions in the Contemporary World*. Oxford University Press, New York

Griffin S M 1996 *American Constitutionalism*. Princeton University Press, Princeton, NJ
Habermas J 1996 *Between Facts and Norms* (Rehg W, trans.), MIT, Cambridge, MA
Habermas J 1998 *The Inclusion of the Other*. Cronin C, de Greiff P (eds.), MIT, Cambridge, MA
Hanafin P J, Williams M S (eds.) 1999 *Identity, Rights, and Constitutional Transformation*. Ashgate, Aldershot, UK
Hayek F A von 1960 *The Constitution of Liberty*. University of Chicago Press, Chicago
Henkin L 1989 *Constitutionalism and Rights*. Henkin L, Rosenthal A N (eds.), Columbia University Press, New York, Introduction
Jackson V C, Tushnet M 1999 *Comparative Constitutional Law*. Foundation, New York
Klug H 2000 *Constituting Democracy*. Cambridge University Press, Cambridge, UK
McIlwain C H 1947 *Constitutionalism: Ancient and Modern*. Cornell University Press, Ithaca, NY
Mamdani M 1991 Social movements and constitutionalism: The African context. In: Shivji I G (ed.) *State and Constitutionalism: An African Debate on Democracy*. Sapes, Harare, Zimbabwe
Mamdani M 1996 *Citizen and Subject*. Princeton University Press, Princeton, NJ
Nwabueze B O 1973 *Constitutionalism in the Emergent States*. Hurst, London
Okoth-Ogendo H W O 1993 Constitutions without constitutionalism: Reflections on an African paradox. In: Greenberg D, Katz S N, Oliviero M B, Wheatley S C (eds.) *Constitutionalism and Democracy*. Oxford University Press, New York
Pennington K 1993 *The Prince and the Law 1200–1600*. University of California Press, Berkeley, CA
Pennock R J, Chapman J W 1979 *Nomos XX: Constitutionalism*. New York University Press, New York
Poggi G 1978 *The Development of the Modern State*. Stanford University Press, Stanford, CA
Pound R 1957 *The Development of Constitutional Guarantees of Liberty*. Yale University Press, New Haven, CT
Rosenfeld M 1994 *Constitutionalism, Identity, Difference, and Legitimacy*. Duke University Press, Durham, NC
Sajo A 1999 *Limited Government*. Central European University Press, Budapest, Hungary
Sartori G 1994 *Comparative Constitutional Engineering*. New York University Press, New York
Schedler A, Diamond L, Plattner M F (eds.) 1999 *Self-restraining State: Power and Accountability in New Democracies*. Lynne Rienner Publishers, Boulder, CO
Shivji I G (ed.) 1991 *State and Constitutionalism: An African Debate on Democracy*. Sapes, Harare, Zimbabwe
Shivji I G 1998 *Not Yet Democracy: Reforming Land Tenure in Tanzania*. International Institute for Environment and Development and HAKIARDHI, University of Dar es Salaam, Tanzania
Singh M P 1989 *Comparative Constitutional Law*. Eastern, Lucknow, India
Sunstein C R 1993 On property and constitutionalism. *Cardozo Law Review* **14**: 907–35
Tate C N, Vallinder T (eds.) 1995 *The Global Expansion of Judicial Power*. New York University Press, New York
Tuebner G (ed.) 1997 *Global Law Without a State*. Dartmouth Publishing, Brookfield, NT
Tully J 1995 *Strange Multiplicity: Constitutionalism in an Age of Diversity*. Cambridge University Press, Cambridge, UK
Van Caenegem R C 1995 *An Historical Introduction to Western Constitutional Law*. Cambridge University Press, Cambridge, UK
Wormuth F D 1949 *The Origins of Modern Constitutionalism*. Harper, New York

H. Klug

Constraint Satisfaction Models

A designer creates something by a particular time, under a specified cost, making use of only certain materials. Problems like this can often be solved by progressive satisfaction of specified constraints. Virtually any problem can be stated in terms of constraints, and solutions are essentially problem states that satisfy constraints. Additional complications may arise when some constraints are more important than others or when there is no perfect solution in which all constraints can be completely satisfied. This article stresses that solution of such problems can be modeled in a natural way with constraint-satisfaction neural networks.

1. *A Neural Constraint-satisfaction Algorithm*

Each unit in a constraint-satisfaction neural network can represent a hypothesis that is interpreted as true when the unit is active and false when the unit is inactive. Connection weights between the units can represent implicational constraints between the hypotheses. These connections are typically exhaustive, bidirectional, and symmetrical. There can also be external input to a unit, reflecting external evidence, and a bias input, reflecting *a priori* tendencies. The relative importance of these various constraints is implemented by the sizes of weight, input, and bias values. Goodness is the degree to which the network satisfies all the constraints.

In a generic constraint-satisfaction algorithm, the goodness contributed by a particular unit i is defined as

$$goodness_i = \sum_j w_{ij} a_i a_j + bias_i a_i + input_i a_i \qquad (1)$$

where a_i is the activation of receiving-unit i, a_j is the activation of sending-unit j, w_{ij} is the weight between them, $bias_i$ is the bias of unit i, and $input_i$ is the external input to unit i. Goodness increases when units connected by positive weights achieve similar activations and units connected by negative weights achieve dissimilar activations, and as unit activations come to resemble their biases and their external inputs. Goodness over the entire network is defined as the sum of

individual goodness values across all receiving-units i in the network

$$goodness_o = \sum_{ij} w_{ij} a_i a_j + \sum_i bias_i a_i + \sum_i input_i a_i \quad (2)$$

Such functions go by different names in different fields: Lyapunov functions in dynamic systems, Hamiltonian functions in statistical mechanics, cost functions in optimization theory, and fitness functions in evolutionary biology (Hertz et al. 1991).

Unit activations are updated so that overall goodness increases or eventually stays the same. The update rules employ net input to a unit i, defined as

$$net_i = int\left(\sum_j w_{ij} a_j + bias_i\right) + ext(input_i) \quad (3)$$

where *int* and *ext* modulate the relative influence of internal and external factors, respectively. With a positive net input, the activation of a unit i is computed as

$$a_i(t+1) = a_i(t) + net_i(ceiling - a_i(t)) \quad (4)$$

where $t+1$ indexes the updated activation, t indexes the previous activation, and *ceiling* is the highest activation that units can achieve. Equation (4) specifies that, before being added to the previous activation, net input is scaled by distance to the ceiling, thus preventing unit activations from rising above the ceiling. With a negative net input, unit activation is updated by

$$a_i(t+1) = a_i(t) + net_i(a_i(t) - floor) \quad (5)$$

where *floor* is the lowest activation value. Here, activations are kept from falling through the floor by scaling net input by distance from the floor before adding net input to the previous activation. Thus, goodness increases because positive net inputs drive unit activations toward the ceiling while negative net inputs drive them toward the floor. Overall goodness in the network continues to increase until units reach their extreme values or net inputs fall to 0. Such states are called stable (or fixed-point) attractors, and a network reaching such a state is said to have settled into a solution, although the solution may only be a local maximum of goodness.

The dynamics of network behavior can be conceptualized within a multidimensional space in which the activation of each unit is represented as a particular direction. Each point in this space, an activation vector, represents a possible state of the network. Besides converging on fixed-point attractors, constraint-satisfaction networks, depending on their weights and inputs, can also settle into limit cycles (characterized by periodic motion of the activation vector) or strange attractors (characterized by chaotic movement of the activation vector) (Hertz et al. 1991).

2. Algorithmic Variations

There are many possible variations on this basic algorithm. Activation values can be continuous or binary. Activation updates can be synchronous or asynchronous, deterministic or stochastic, and can minimize energy or maximize goodness. Energy is basically the negative of goodness. A fixed-point attractor is a low energy point within a basin of attraction, and any input pattern within a particular basin is transformed into the attractor state for that basin. Constraint-satisfaction networks are sometimes called *Hopfield* networks after the person who developed the first energy functions for them (Hopfield 1984).

A variety of different nonlinear activation functions can implement updates, e.g., sigmoid or hyperbolic-tangent functions. Initial activations can start at 0 or be preset to other values. Unit biases, inputs, decay, self-connections, and internal and external modulators are optional. Connections can be symmetric or asymmetric.

The search for a global goodness maximum can be facilitated by simulated annealing, a process of gradually lowering the temperature (or gain) of the activation update function in order to move networks with stochastic, binary units out of local maximums (Hinton and Sejnowski 1986). Such networks are called Boltzmann machines because the probabilities of the states are characterized by the Boltzmann distribution in statistical mechanics.

Connections can be determined by conventional learning algorithms, such as the Hebb rule or the delta rule (Hertz et al. 1991), or be set by a programmer, perhaps on the basis of psychological principles. Finally, activation updating can be incorporated into multi-layered feed-forward networks that learn with back-propagation of error, in a technique called recurrent back-propagation (Hertz et al. 1991). This allows for the inclusion of hidden units, enabling the learning of nonlinear patterns. Boltzmann machines can also use hidden units, to the same advantage.

3. Symbolic Constraint Satisfaction

In Artificial Intelligence (AI), problem solving typically is viewed as search through a problem space, where a particular state of the problem can be transformed into another particular state by applying some operator. Such searches may generate solutions in the form of paths through the search space. When a search reaches a dead end without a solution, the problem solver backtracks to a more promising state to continue the search. Constraints can limit the size and complexity of search by eliminating parts of the search space, in some cases eliminating the need for backtracking (Kumar 1992). Symbolic constraint-directed search has been applied to a wide range of

AI problems including planning, scheduling, design, reasoning, belief maintenance, map coloring, and machine vision.

Although many interesting algorithms, heuristics, and issues have arisen from this research on symbolic constraint satisfaction, it has not generated any notable models of human performance, perhaps because symbolic methods for search and constraint propagation have psychologically unrealistic working-memory requirements. Also, symbolic methods do not cope naturally with constraints that conflict with each other or vary in importance.

4. *Neural Constraint-satisfaction Models*

A large number and variety of psychological phenomena have been simulated with constraint-satisfaction neural networks. Content-addressable memories are stored as fixed-point attractors, retrievable from cues containing part of the memory content (Hopfield 1984). Values for particular slots in schemas are completed in a way that satisfies constraints imposed by external and internal information (Rumelhart et al. 1986). Analogies are mapped on to new problems by satisfying constraints imposed by similarity, structure, and purpose (Holyoak and Thagard 1989). Favored explanations are those that maximize coherence among hypotheses and evidence (Thagard 1989). Decisions and plans maximize coherence among goals, subgoals, and actions (Thagard and Millgram 1995). Forming impressions of people is a matter of reconciling information about their behavior and characteristics and the stereotypes of the perceiver (Kunda and Thagard 1996). Attitudes change in a way that increases consistency across attitudes and beliefs (Spellman et al. 1993). Cognitive dissonance is reduced as constraints imposed by attitudes, beliefs, and their implications are employed to increase consonance (Shultz and Lepper 1996). Objects are recognized by matching visual scene information with model features in a consistent way (Chen and Lin 1994). An input string of letters is transformed within a recurrent feed-forward network into an attractor representing a pronunciation of the corresponding word (Plaut et al. 1996). Damaging the units or connections of recurrent feed-forward networks alters the basins of attraction, simulating a variety of neuropsychological deficits such as dyslexia and prosopagnosia (Plaut 1995).

5. *Future Trends in Constraint-satisfaction Research*

Current activities can be extrapolated to predict some near-future trends in research with constraint-satisfaction models. First, it is likely that constraint-satisfaction models will be extended to other psychological phenomena. These might include cognitive balance theory, belief maintenance and perseverance, problem solving, and additional neuropsychological deficits.

For both new and current domains, it can be expected that constraint-satisfaction models will be run, not only to cover existing psychological phenomena, but increasingly to predict new phenomena that could guide empirical research.

Efforts might continue to expand the limited memory capacity of current constraint-satisfaction networks. The number of patterns that *Hopfield* networks can store and accurately retrieve is only 13.8 percent of their number of units (Hertz et al. 1991).

Neural network modelers might use some of the heuristics discovered in symbolic constraint-satisfaction research to enhance the power of neural models to cover the guessing and backtracking required on more difficult problems.

Finally, the neurophysiology of constraint satisfaction could be more fully explored to the mutual benefit of both modeling and brain research. Evidence has been found for attractors in the olfactory system of rabbits and cats (Freeman 1991) and in monkey visual cortex (Miyashita 1988). Brain imaging of humans suggests that the anterior cingulate cortex is involved in the monitoring of cognitive conflict and interference, and this can be modeled by neural network energy (Botvinick et al. 1998).

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence: Search; Linear Algebra for Neural Networks; Neural Networks, Statistical Physics of; Problem Solving and Reasoning, Psychology of.

Bibliography

Botvinick M M, Braver T S, Carter C S, Barch D M, Cohen J D 1998 Evaluating the demand for control: Anterior cingulate cortex and crosstalk monitoring. Technical Report PDP.CNS.98.1. Carnegie Mellon University, Pittsburgh, PA

Chen T-W, Lin W-C 1994 A neural network approach to CSG-based 3-D object recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **16**: 719–26

Freeman W J 1991 The physiology of perception. *Scientific American* **264**(2): 78–85

Hertz J, Krogh A, Palmer R G 1991 *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA

Hinton G E, Sejnowski T J 1986 Learning and relearning in Boltzmann machines. In: Rumelhart D E, McClelland J L, PDP Research Group (eds.) 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA, Vol. 1

Holyoak K J, Thagard P 1989 Analogical mapping by constraint satisfaction. *Cognitive Science* **13**: 295–355

Hopfield J J 1984 Neurons with graded responses have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences* **81**: 3008–92

Kumar V 1992 Algorithms for constraint satisfaction problems: A survey. *AI Magazine* **13**(1): 32–44
Kunda Z, Thagard P 1996 Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint satisfaction theory. *Psychological Review* **103**: 284–308
Miyashita Y 1988 Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* **335**: 817–20
Plaut D C 1995 Lesioned attractor networks as models of neuropsychological deficits. In: Arbib M A (ed.) *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA
Plaut D C, McClelland J L, Seidenberg M S, Patterson K 1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115
Rumelhart D E, Smolensky P, McClelland J L, Hinton G 1986 Schemata and sequential thought processes. In: Rumelhart D E, McClelland J L, PDP Research Group (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA, Vol. 2
Shultz T R, Lepper M R 1996 Cognitive dissonance reduction as constraint satisfaction. *Psychological Review* **103**: 219–40
Spellman B A, Ullman J B, Holyoak K J 1993 A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf War. *Journal of Social Issues* **49**: 147–65
Thagard P 1989 Explanatory coherence. *Behavioral and Brain Sciences* **12**: 435–466
Thagard P, Millgram E 1995 Inference to the best plan: A coherence theory of decision. In: Ram A, Leake D B (eds.) *Goal-driven Learning*. MIT Press, Cambridge, MA

T. R. Shultz

Constructivism/Constructionism: Methodology

What makes a particular research method 'constructivist' or 'social constructionist' in orientation? In a sense, the answer lies less in the method or technique itself than in the philosophy with which the technique is used. For example, an interview can be conducted to count the number of positive self-statements in the interviewee's verbal output, or to classify him or her according to some extant diagnostic system. However different these aims might be from one another, both could be considered to express traditional research goals associated with applying or creating universally 'true' or valid categories for human experience. In contrast, other interviews might be conducted to tease out a network of personal meanings unique to a particular respondent, to reveal his or her shifting reliance upon different narratives over the course of therapy, or to raise questions about the apparent coherence of a particular line of political discourse. Such interview methods might well be constructivist, as they focus on distinctive patterns or processes of constructing (or deconstructing) meanings in a given personal or social context, without the presumption of universality that attends traditional knowledge claims in psychological science.

More specifically, a research method might be classified as constructivist or social constructionist to the extent that it (a) elucidates 'local' as opposed to 'universal' meanings and practices, (b) focuses upon provisional rather than 'essential' patterns of meaning construction, (c) considers knowledge to be the production of social and personal processes of meaning-making, and (d) is more concerned with the viability or pragmatic utility of its application than with its validity *per se*. This emphasis on local, provisional, and pragmatic assessment of (inter)personal meanings can be illustrated by a closer consideration of three core techniques associated with a constructivist or social constructionist approach, each of which encompasses many different variations. These include repertory grid technique, analysis of narrative processes in spoken or written 'text,' and discourse analysis.

1. Repertory Grid Technique

Developed within personal construct theory, the repertory grid technique represents a widely used method for studying personal and interpersonal systems of meaning. Because of their flexibility, repertory grids (or *repgrids*) have been used in literally thousands of studies of a broad variety of topics, ranging from children's understandings of physical science principles and consumer preferences to formal structures of self-reflection within cognitive science and the mutual validation of belief systems between friends. However, their most consistent area of application has probably been in the clinical domain, where grids have been used to assess the meaning systems of clients in psychotherapy, and how these change over the course of treatment. The 'reptest' was initially designed by George Kelly, the author of personal construct psychology, as a means of assessing the content and structure of an individual's repertory of role constructs—the unique system of interconnected meanings that define his or her perceived relationships to others.

In its simplest form, the reptest requires the respondent to compare and contrast successive sets of three significant people (e.g., my mother, my father, and myself), and formulate some important way in which two of the figures are alike, and different from the third. For example, if prompted with the above triad, a person might respond, 'Well, my father and I are very self-reliant, whereas my mother is very dependent on others.' This basic dimension, *self-reliant vs. dependent*, would then be considered one of the significant themes or constructs that the person uses to organize, interpret, and approach the social world, and to define his or her role in it. By presenting

the respondent with a large number of triads of varying 'elements' (e.g., a previous romantic partner, best friend, a disliked person, one's ideal self), the reptest elicits a broad sampling of the personal constructs that constitute the person's outlook on life and perceived alternatives. These constructs can then be interpreted impressionistically, used as the basis for further interviewing of the respondent, or coded using any of a number of reliable systems of content analysis, conducted either manually or using available computer programs.

Although the analysis of construct content is often revealing, most contemporary users prefer to extend the method beyond the simple elicitation of constructs by prompting the respondent subsequently to rate or rank each of the elements (e.g., people) on the resulting construct dimensions. For example, a respondent might generate a set of 15 constructs, on which 10 important elements (e.g., my mother, father, self, partner) could be rated, yielding a matrix of 150 specific ratings that would then be amenable to a wide range of analyses. Although the repgrid was originally devised as an interview-based or paper-and-pencil measure, most contemporary users rely on any of a number of computer programs for their elicitation and analysis, such as the popular *WebGrid III* program available via the Internet.

Although analysis of the grid can focus on simple element ratings (e.g., observing that the respondent views herself as *self-reliant*, *hard-working*, *emotionally controlled*, etc. on a sample of her personal constructs), it is typically more helpful to conduct a comprehensive analysis of the grid to discern larger patterns. This might involve correlating and factor analyzing the matrix of ratings to see at a glance which constructs 'go together' for the respondent (e.g., *self-reliance* is associated with being *psychologically healthy*, whereas *dependence* implies *weakness*), or to learn what people are most and least alike in the respondent's view. These linkages among constructs often suggest why people remain 'stuck' in symptomatic patterns, as when a client resists reconstruing himself as *happy* instead of *depressed*, because the former is associated with being *superficial* as opposed to *deep*. Similarly, patterns of identification among elements in a grid can be clinically informative, with some of these (e.g., degree of correlation between actual self and ideal self) providing useful indices of progress in psychotherapy.

At the most general level, grids can be useful for identifying how 'tightly' or 'loosely' a given person's meaning system is organized, as revealed, for example, by the average degree of correlation among the constructs that comprise it. Like most constructivist methods, repgrids focus on personal or local (e.g., family) systems of meaning, with no assumption that these generalize to other people. Moreover, repgrids measure construct systems at a particular moment in time, demonstrating meaningful sensitivity to variation across education and psychotherapy, rather than presuming to reflect invariant psychological essences. Finally, most users of repertory grids consider the ultimate hallmark of their value to be the degree to which they prompt relevant insights for the respondents who complete them, although some investigators have also demonstrated their validity and reliability in more conventional terms. Assessment methods having some affinity to the grid technique include the *laddering technique*, which elicits the higher-order value implications of the respondent's constructs of behaviors and social roles, and the *self-confrontation method*, which evaluates an individual's sense of agency and communion with others through an analysis of important life events. A common feature of these methods is their focus on *personal structure*, viewing meaning in systemic terms that can be depicted visually to enhance its comprehension and discussion by research participants.

2. *Narrative Analysis*

Narrative psychology considers thinking processes to be guided by the structures and elements of which stories are composed. Instead of viewing logic as the guiding force behind information processing and meaning making, narrative psychologists view people as organizing life events into plot structures and creating themes that give them significance. This model of organization can be classified as constructivist due to the agentic, creative processes involved in the formation of narratives.

Individuals create stories to help them comprehend and communicate their experiences and interactions in the world. Narratives are generated through the use of language, common scripts, and symbols derived from different cultures and traditions, and therefore are necessarily context dependent. Stories produced are seen as reflecting truths unique to their individual authors and contexts and as rooted within distinctive lived histories.

The narrative perspective has become popular only over the last 20 years, but has grown rapidly into a movement in psychology, partly due to its broad relevance across cognitive, developmental, and clinical areas of research. In the cognitive domain, investigators have demonstrated that people inherently organize ambiguous experiences using story schemas, and have studied the way that narrative formation structures memory as well as perception of events. Developmental researchers have traced the emergence of children's story-telling competencies, and their progressive shift toward accounting for the perspective of listener vs. narrator. In the domain of psychotherapy, researchers have developed methods for mapping the topical and thematic shifts that characterize client accounts of their experience, relating these to good and poor outcome. The *Narrative Process*

Coding and System illustrates the use of constructivist methods to identify and characterize narratives from transcribed psychotherapy sessions.

First, an investigator using the system segments the session dialogue into topic units that are identified through shifts in protagonists and themes. For instance, a client may begin a session discussing a conflict she is having with a colleague and then shift to examining a marital issue. The identification of topic segments can allow researchers to consider how themes are maintained or changed through discussion and how the interpersonal psychotherapeutic process can act to facilitate narrative development. Second, the researcher codes these topic segments into one of three narrative processes. External narrative sequences are dominated by event description. The recounting of the steps entailed in registering for university would be an example of this narrative sequence type. Internal narrative sequences focus upon emotional and experiential states. For instance, a portrayal of the different feelings present on the first day of classes would exemplify this narrative process. Finally, reflexive narrative sequences entail analysis and interpretation of events and internal reactions in order to understand their significance. An exploration of the meaning of attaining a higher education would be classified as reflexive. Within different psychotherapy orientations, research has suggested specific patterns of narrative process associated with positive outcome. For instance, a shift between internal and reflexive exploration has been associated with successful humanistic psychotherapy.

Like other methods, whether or not the narrative process coding system is 'constructivist' depends on the epistemic intent that guides its use. For example, the system can be used within an objective, experimental model of science to compare the distribution of narrative sequences within different transcripts. Within such a model, results can be tabulated and statistical arguments for generalization of findings can be developed that aspire toward more universal implications. Although useful in some respects, such an application departs from the emphasis on individuality that characterizes most constructivist research. Alternately, this narrative tool can be used in a discovery-oriented framework, and themes identified that represent local, provisional truths. For instance, dominant narrative themes could be identified within the discourse of women with eating disorders on the topics of food and body image, and these themes could be explored in relation to the lives of these individual women and their social contexts.

Other narrative coding methods include the identification of *core conflictual relationship themes* implicit in client accounts of problems with other people, and the coding of interviews to establish the degree to which the narrator experiences him or herself as an *origin* or *pawn* in connection with important events. In each case, the epistemic framework within which findings are presented and interpreted will dictate whether or not a method is being used in a constructivist spirit.

3. *Discourse Analysis*

Whereas constructivist methods tend to focus on the meaning-making efforts of individuals, families, and small groups, social constructionist methods shift attention to the broader systems of 'languaging' that characterize public speech, disciplinary discourse, or cultural contexts. Even when applied to individual language users, such methods tend to have a critical thrust, such as when careful 'deconstructive listening' is used to bring to light the hidden assumptions that constrain a speaker's view of what positions are permissible for self or others. Social constructionist methods are in this sense more radical than constructivist methods, attempting to uproot oppressive and marginalizing forms of thinking and speaking that have dehumanizing consequences.

An example of social constructionist methods is provided by *discourse analysis*, the critical evaluation of social 'texts' to reveal their concealed constraints and contradictions. For example, analyses of published debates in the South African parliament in the early 1980s have elucidated the shared assumptions about the superiority of whites held by both liberal and conservative MPs. Likewise, feminist discourse of the 1970s has been 'deconstructed' to expose the ways in which 'butch/femme' lesbians were marginalized to support a politic that advocated lesbian androgyny. Interestingly, discourse analysts have been especially active in undermining traditional assumptions in psychology, such as the largely unexamined convictions that people ideally function as rational 'selves' and that psychotherapy involves a move toward greater 'reality contact.' As a method, discourse analysis appears to be more an art than a science, borrowing from deconstructive and critical literary theory in order to understand and resist 'dominant narratives' that limit the terms of debate or the permissible range of human variation that is considered 'normal.' It is therefore sometimes used by psychotherapists concerned with helping clients free themselves from the oppressive narratives of their lives, and recover a sense of themselves as the authors of their experience. Related social constructionist methods include the *analysis of rhetoric*, the tropes or moves used to advance a persuasive argument, and *textual analysis*, which reveals the way in which authors skillfully position themselves and others within the moral discourse of an account, so as to present themselves to best advantage.

In summary, constructivist and social constructionist methods in psychology are multifaceted in form and diverse in application, but tend to express a shared commitment to revealing the structure of personal and

collective meanings and the processes by which they are constructed. Given their convergence with postmodern trends in the human sciences, it is likely that their influence will continue to grow in the future.

See also: Constructivist Psychotherapies; History of Science: Constructivist Perspectives; Linguistic Turn and Discourse Analysis in History; Narrative, Sociology of; Narratives and Accounts, in the Social and Behavioral Sciences

Bibliography

Angus L, Levitt H, Hardtke K 1999 Narrative processes coding system. Research applications and implications for psychotherapy practice. *Journal of Clinical Psychology* **55**: 1255–70

Burr V 1995 *An Introduction to Social Constructionism*. Routledge, London

Fransella F, Bannister D 1979 *A Manual For Repertory Grid Technique*. Halstead Press, New York

Gottschalk L A, Lolas F, Viney L L 1986 *Content Analysis of Verbal Behavior*. Springer-Verlag, New York

Luborsky L, Crits-Christoph P 1998 *Understanding Transference: The Core Conflictual Relationship Theme Method*. American Psychological Association, Washington, DC

Neimeyer R A 1993 Constructivist approaches to the measurement of meaning. In: Neimeyer G J (ed.) *Constructivist Assessment: A Casebook*. Sage, Newbury Park, CA, pp. 58–103

Neimeyer R A, Raskin J D (eds.) 2000 *Constructions of Disorder*, 1st edn. American Psychological Association, Washington, DC

R. A. Neimeyer and H. Levitt

Constructivism in Cognitive Psychology

Constructivist thinking and reasoning have played a major role in psychology for several years; for example, in William James' 'pragmatic method' as well as in Barlett's 'construction of remembering.' In modern cognitive psychology constructivist principles and concepts are the basic elements of research, especially in the fields of learning, language and memory. Rumelhart (1980) analyses cognition as an interpretation process controlled by cognitive schemes. Johnson et al. (1999) examines the construction of mental models for 'extensional reasoning' of 'naive individuals.' Constructivist principles and concepts are also relevant for scientific knowledge because 'knowledge of all kinds, including scientific knowledge, is a construction of the human mind' (Scarr 1985, p. 499). The concept that all knowledge and thinking is a result of construction is certainly the core assumption of all versions of scientific constructivism.

1. What does Constructivism Mean?

Constructivist approaches refer to an epistemological position in which knowledge is regarded as constructed. These approaches concentrate on the analysis of single processes or functions. Constructivist approaches present themselves through a remarkable spectrum in psychology. Therefore, our understanding of constructivism as a unifying principle is questionable. Instead we prefer to distinguish varieties of constructivist reasoning in psychology (and philosophy), which share at least the first two of the following characteristics: they analyse the ability of individuals to act and to demonstrate self-control, partly with the consequence of taking distal variables like genetic variation into special consideration (Scarr 1985, p. 510). They are based on the assumption that knowledge is the result of constructive processes; and the consequence of this for the analysis of person-situation-interaction is that these are constructed on the basis of experiences, i.e., the same context is seen and applied differently.

Finally, interventions and instructions will be viewed here in relation to individually-constructed experiences and their applicability and fit. These three characteristics are core features of constructivist reasoning. The following representation of constructivist perspectives in psychology refers to the consideration of these features and not only to the fact that the author calls himself a constructivist. This manner of thinking is legitimate because on the one hand there are numerous authors who do not call themselves constructivists but are closely associated with this perspective. On the other hand, there are considerable differences between various constructivist approaches, especially between radical constructivism and social constructivism as well as between constructivist approaches to learning and knowledge acquisition.

Radical constructivism—which views realism in science as inadequate and assumes an informational closeness of cognitive systems (Maturana & Varela 1987)—plays a subsidiary role in psychology, even though it is postulated as the prototype of constructivist perspectives. Radical constructivists like Maturana or Varela always question the possibility of objectivity and truth because only information which is viable and serves to support the system is, according to them, processed. Truth, objectivity and knowledge are thus constructed. Radical constructivism focuses on epistemology and studies the relationship between knowledge and the world, rejecting realism and any form of ontology.

A second version of constructivism is named 'constructionism' or 'social constructionism'. This version

follows the tradition of the social psychology of Cooley and Mead. It analyses the processes 'in which the individuals describe and explain the world in which they live or how they see it' (Gergen 1985, p. 3f). Social constructionism, in contrast to radical constructivism, focuses on the acting and thinking individual and describes the construction of knowledge as the result of social interaction in contexts which form the foundation of shared knowledge. This view has proven itself remarkably fruitful for numerous psychological fields, for example, in the analysis of cognitive development or the development of prejudice and self-impression.

A third version of constructivism is engaged in exploring the correlation of learning and instruction, especially in the area of mathematic knowledge (Cobb et al. 1997, Resnick 1994). Knowledge acquisition is discussed here in close connection with situated learning and related instructional approaches. In this review, only the two last-presented versions of constructivism will be discussed, due to the fact that these have been the most influential in cognitive psychology. First of all, approaches which define themselves as constructivist will be presented. Second, studies will be discussed which contain the core features of constructivist reasoning. Constructivism emphasizes the active experience-based knowledge construction and its embedding in social contexts. From this point of view intervention and instruction are chosen with regard to their individual fit. In this article cognitive tools which are typical for constructivistic approaches will be discussed first: conceptual metaphors, cognitive schemes, and subjective theories. In the second part they will be described in different psychological application areas. Therein the potential of constructivistic approaches for the research on text comprehension and the development of domain-specific knowledge can be shown. Moreover a better understanding of the relationship between construction and instruction can be gained.

2. *Principles for the Construction of Knowledge*

'The business of cognitive science is to give empirically grounded explanations of the phenomena of mind, in the broadest sense of that term' (Fernandez-Duque and Johnson 1999). Thus the goal of cognitive psychological studies is the pursuit of principles and concepts which describe and explain the construction of knowledge. Piaget compares a child's development of intelligence to the development of scientific theories. Bruner (1990) suggests the use of concepts which are useful to the analysis of narrative structures for the analysis of mind. Personal constructs (Kelly 1955), cognitive schemes (Rumelhart 1980), metaphors (Ortony 1980), folk theories of mind and behavior (Bruner 1990, Lillard 1998) and theory-theories (Gopnik & Wellman 1994) have proved important for the construction of everyday knowledge and scientific knowledge. Through these concepts, active experience-based processes of knowledge construction, rooted in different contexts, are analyzed.

Kelly points out in his personal construct theory in 1955 that individuals basically behave like scientists: they too develop theories about their environment and use personal constructs of prediction and interpretation to control their environment. Kelly postulates these predictions by forming and using personal constructs as active psychological processes which lead to different choices. Individuals cannot always change their environment, but they can always construct it differently. Kelly (1970) described this as constructive alternativism. Psychological research is especially useful to facilitate effective psychotherapy. Personal constructs represent the experiences of individuals and form the precondition for active, sometimes alternative construction of knowledge. Rumelhart uses the concept of cognitive schemes in a constructivist manner to analyse knowledge representation and how that representation facilitates the use of knowledge in particular ways (Rumelhart 1980, p. 34). Cognitive schemes are understood as data structures. They are active computational devices which represent meaning and procedural knowledge. Following Barlett with his 'gist story,' Rumelhart describes the effects as interpretation and as a reconstruction of the original interpretation. Rumelhart later (1999) uses the concept of metaphor in a more complex way, as a theoretical model by employing the computer of the computer as a model of mind. Such metaphors of mind (Ortony 1980) are currently used for describing different research approaches, i.e., attention and metaphors for learning.

Fernandez-Duque and Johnson (1999, p. 84) understand the conceptual metaphor 'as conceptual mapping of entities, properties, relations and structures from a domain of one kind (the source domain) onto a domain of a different kind (the target domain).' They understand it as a tool for the construction of scientific theories. Here you can also find typical arguments for constructivist approaches, even though the researchers do not regard themselves as constructivists. Fernandez-Duque and Johnson show 'that a highly constrained set of conceptual metaphors is constitutive and definitive of the theoretical perspectives taken toward attention and toward the search programs based on these theories' (Fernandez-Duque and Johnson 1999, p. 84). Conceptual metaphors like the ones described above not only structure folk models but also 'our most sophisticated scientific theories.' Fernandez-Duque and Johnson show this by using metaphors such as the filter metaphor, the spotlight metaphor, the spotlight-in-the-brain metaphor and the attention-as-vision metaphor. These metaphors guide research questions and influence the search for data and its interpretation.

Metaphorical reasoning is clearly central to the construction of ordinary and scientific knowledge, for mathematic as well as for cognitive psychology (Moreno & Mayer 1999). Ortony (1980) as well as Sternberg (1990) emphasize the function of metaphors in the creation of new knowledge and promote viewing incidents in a new light, because 'without an understanding of the metaphors that generate or at least guide various theories' (Sternberg 1990 18), psychological theories of intelligence can not adequately be related to each other. Mayer (1992) describes the theoretical development of learning as a development of three metaphors; the metaphor of response acquisition, knowledge acquisition and knowledge construction.

The construction of new knowledge by cognitive schemes and metaphors can as well be found in approaches in which knowledge construction in everyday life is basically viewed as equivalent to the development of scientific knowledge. This equivalence of folk psychology and scientific psychology (Bruner 1990) means that new knowledge is constructed as meaningful knowledge and that narrative formats are regarded as key tools for the construction of meaning. In his representation of the theory of mind, Bruner emphasizes the assumption that ordinary knowledge is equivalent to theoretical knowledge and is generated domain-specific and in the range of cognitive development. Children develop concrete theories of everyday psychology or about mathematics, which they constantly change.

Those naive theories, as well as metaphors and cognitive schemes, encompass different but related coherent concepts which are used as tools for the construction of knowledge. Thus these concepts are 'building blocks of cognition' (Rumelhart 1980), which describe and explain the phenomena of mind (Fernandez-Duque & Johnson 1999). The views related to these concepts describe remarkably closely the core features of constructivism, specified at the beginning of this article.

3. *Fields of Application for Constructivist Approaches in Psychology*

In the following part, some of the most important fields of application for constructivist approaches in psychology will be analyzed. These include remembering and knowledge representation, language and text comprehension, learning and instruction, and cognitive development as well as social cognition, social biases and social stigmata. Cognitive processes are still the most important aspects of research using constructivist approaches. In developmental and social psychology a considerable amount of social constructivist research can be found which expand on the theories of Piaget, Vygotsky, and Mead (Berger and Luckmann 1966).

In 1932 Bartlett stated that remembering 'is an imaginative reconstruction, or construction, built out of the relation of our attitude towards a whole active mass of organised past reactions or experience' (Bartlett 1932, p. 213). Bartlett used the concept of cognitive schemes but did not think it as important as Rumelhart did (1980). Remembering is, according to Bartlett, less the recall of knowledge memorized in cognitive schemes and more of a transactional process which zigzags between the perception of past and new experiences and which leads to revised construction of knowing and acting. Psychologists arguing in a constructivist way often refer to Bartlett, however, they come to very different assumptions about the functional meaning of cognitive schemes and their evaluation of 'stored schema view' (Clancey 1997).

Clancey (1997) regards the 'stored schema view' as inadequate and undynamic. Thomas (1999, p. 224) argues in a similar way, claiming 'it is not the schema but the activity it supports that carries intentionality and embodies our perceptual and imaginal awareness'. In contrast to activity-oriented approaches, the mental model theory states that 'what we perceive depends on both what is in the world and what is in our heads—on what evolution has wired in our nervous system and what we know as a result of experience. The limits of our models are the limits of our world' (Johnson-Laird 1989, p. 471). Johnson-Laird et al. (1999) understand mental models as cognitive representations of relevant situations, which have a structure and content. At present these types of moderate constructivist approaches represent the mainstream of cognitive psychology. This is also demonstrated by the fact that the situated approaches have problems with certain phenomenon, especially with the phenomenon which Landauer and Dumais (1997) call 'Plato's problem' or 'the poverty of the stimulus,' i.e., 'how people acquire as much knowledge as they do on the basis of as little information as they get.'

This problem is the basis of Chomsky's nativist-constructivist theory of language (Chomsky 1991). Chomsky emphasized the creative aspect of language, because a person always hears for the first time the combination of sentences which another speaker verbalizes. Thus Chomsky and his school propose 'a language acquisition device, an innate mechanism which prepares linguistic competences like a language instinct.' The theories for text comprehension emphasize both language production and the active construction of knowledge. Kintsch and his associates (1988) developed a theoretical model of text comprehension which discriminates between a text basis, i.e., the relations deduced from semantics, and a situational model. They have also evaluated the model in various studies.

Deep understanding is not possible until a situational is constructed, which integrates a textual basis and the reader's knowledge. To this end Kintsch (1988) presented a theoretical approach which he

calls 'construction-integration model.' In the above-described sense, this model of text comprehension is a constructivist model which describes in detail knowledge-construction and the activation of knowledge. In addition, there are studies which consider the third core feature of constructivist reasoning, which involves instruction in the ideal proportion of individually-constructed experiences and their individual usability and fit. McNamara et al. (1996) showed that 'high-knowledge readers could be stimulated by difficult, less coherent texts to active processing.' Studies like the one described above show that language reception and text comprehension for the purpose of generating situational models can be adequately supported by constructivist instruction. It is at this point that constructivist-oriented theories of learning and instruction begin. These theories analyse which learning environment supports the construction of knowledge without affecting active processing. These approaches, which view themselves as constructivist, are subsumed under the notation 'situated learning.' The particular characteristics of situated learning are:

(a) learning is an active and constructive process (the situational perspective shares this assumption with the cognitive perspective);

(b) the process focuses on the sharing of distributed knowledge in social groups;

(c) learning is analyzed in the framework of learning environments and described as an adaptation to the constraints and opportunities of the context;

(d) such adjustments are based on the participation of the members of the learning group. Their participation is at first characterized as peripheral and then as central as experience is gained.

Analyses of situated learning prefer to focus on the analysis of effective learning environments and their characteristics. Some of these characteristics have been realized in the following approaches: 'cognitive apprenticeship' (Collins et al. 1989), 'collaborative learning culture' (Brown 1997) and 'authentic learning' (Bruner 1990).

The theory of situated learning is first of all constructed as a theory of instruction and of the mode of operation of learning environments. This theory shares with the cognitive view the assumption that learning is a constructive and knowledge-based process. A further basic assumption is that the use of new media for the development of constructivist learning environments is taken into consideration (Moreno & Mayer 1999).

The development of learning and knowledge acquisition is another subject of constructivist-oriented approaches in long-term studies. This is particularly important for the constructivist theory of mind (Schwanenflugel et al. 1998), which is generated by children for different knowledge domains: about nature (physics, biology) or about memory, comprehension and the intentionality of interpersonal relationships (psychology).The theory of mind is at its core a constructivist approach, which is based on the following assumptions:

(a) knowledge is organized in a similar manner to theory and refers to a core set of beliefs;

(b) the theory of mind is relatively coherent, uses constructs and serves as a basis for predictions and explanations for the behavior of other persons;

(c) the theory of mind is at least partly innate or innately predisposed;

(d) Parts of the theory of mind clearly underlie cultural variations, for example the fundamental ideas of what causes action (Lillard 1998).

Theories of mind are often equated by researchers to scientific theories and conceptual change as a change of paradigm in the sense of Thomas Kuhn. Such analogies can be found in Piaget, Kelly, Bruner and representatives of the theory-theory. Equating the theory of naïve actors and scientific theories is highly problematic if questions of validity, falsification or operationalization are regarded. This position also underestimates the efficacy of social biases (for details see Nickerson 1999) and how they are analyzed by the research of social judgement.

The effect of social cognition in the construction of individual reality has a long tradition, especially in social psychology. For Gergen, a basic assumption for social construction is that 'the terms in which the world is understood are social artifacts, products of historically situated interchanges among people' (Gergen (1985, p. 4). These artifacts are of various types. They range from quality attribution to implicit personality theories to social processes of stigmatizing. One of the most robust findings of individual perception research is the 'lay dispositionism' (Choi et al. 1999). It is upon this finding that typical judgement failures like the correspondence bias, i.e., the belief that behavior of other persons can be explained by personality traits or other internal attributes, are based. Modern social judgment research shows clearly the manner in which individuals construct their knowledge about other persons. This research also demonstrates how individuals impute their own knowledge to others (Nickerson 1999) and retain conflicting attitudes which coexist in the different evaluations of the same attitude object. Wilson et al. (2000) convincingly describe in their model of dual attitudes how old, stored attitudes interact in different ways with new constructed attitudes and thus remain stable while affecting behavior.

Stereotype correspondence biases, dual attitudes and implicit personality theories are, like the above-mentioned metaphors and cognitive schemes, important tools for the construction of the individual reality. This construction is faulty and these errors are subject to revision in the framework of 'bargaining processes' in cooperative groups, especially 'if anchor and adjustment heuristics are used' (Nickerson 1999). These constructions of reality become problematic if they lead to social stigma. Stigmatizing is commonly

understood as a process by which the members of a group mark a stigmatized individual as a consequence of the internalized judgments (Crocker et al. 1998). Research on the construction of coherent knowledge about one's own personality by judgments of others, stereotypes and lay dispositionism (Choi et al. 1999) show remarkably clearly how moderate constructivist psychological approaches give rise to empirically-based explanations of the phenomena of mind (Fernandez-Duque and Johnson 1999).

4. Conclusion

There are considerable differences within this research domain, especially with regard to the meaning of cognitive representation and the question of the extent to which plans control actions. The question regarding the degree to which individual action is influenced by universal principles is also contentious. For Chomsky and his school, language acquisition is controlled by a powerful innate mechanism. Resnick (1994) speaks in this regard of 'biological constraints and prepared schemes and skeletal and hardwired structures which influence learning by protomathematic dispositions.' In contrast, Scarr emphasizes the importance of 'genetic variation and genotype-environment effects for the development of intelligence' (Scarr 1985). In addition, ideas can be found regarding the effect of the social stereotype and the blindness of individuals due to correspondence bias and lay dispositionism (Choi et al. 1999). Bargh's pessimism becomes apparent when he speaks about a cognitive monster—an enormous amount of once-constructed stereotypes, which are resistant to change and once activated, produce automatic stereotypes (Bargh 1999).

These two poles characterize the wide range of constructivist thinking in psychology and show once again the importance of constructivist-oriented instructional support. By the promotion of instruction—which offers room for the construction of knowledge, self-control, self-organization and shared knowledge in collaborative groups—the gulf between nativism and the cognitive monster can be bridged. This is the position of a moderate constructivism, which is particularly accepted by modern psychology in connection with multimedia. This position bears little relation to the radical constructivism and its refusal to recognize reality and objectivity.

On the other hand, the moderate constructivism, which we represent, opens new doors to understanding the phenomenon of mind. The moderate constructivism emphasizes processes like knowledge acquisition, the individual activity potential and the importance of reality construction without claiming the equivalence of naive and scientific knowledge.

See also: History of Science: Constructivist Perspectives; Constructivism/Constructionism: Methodology; Constructivist Psychotherapies; Constructivism in Cognitive Psychology; Piaget, Jean (1896–1980); Piaget's Theory of Human Development and Education; Piaget's Theory of Child Development; Vygotskij, Lev Semenovic (1896–1934); Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Bargh J A 1999 The cognitive monster: the case against controllability of automatic stereotype effects. In: Chaiken S, Trope Y (eds.) *Dual Process Theories in Social Psychology*. Guilford Press, New York, pp. 361–382

Bartlett F C 1932 [1967] *Remembering: A Study in Experimental and Social Psychology*. (reprint edn.) Cambridge University Press, Cambridge, UK

Berger P L, Luckmann T 1966 *The Social Construction of Reality*. Doubleday, New York

Brown A L 1997 Transforming schools into communities of thinking and learning about serious matters. *American Psychologist* **52**(4): 399–413

Bruner J 1990 *Acts of Meaning*. Harvard University Press, Cambridge, MA

Choi I, Nisbett R E, Norenzayan A 1999 Causal attribution across cultures: variation and universality. *Psychol. bull.* **125**(1): 47–63

Chomsky N 1991 Linguistics and cognitive science: problems and mysteries. In: Kasher A (ed.) *The Chomskyan Turn*. Blackwell, Cambridge, MA

Clancey W 1997 *Situated Cognition. On Human Knowledge and Computer Representations*. Cambridge University Press, Cambridge, UK

Cobb P, Gravemeijer K, Yackel E, McClain K, Whitenack J 1997 Mathematizing and symbolizing: the emergence of chains in signification in one first-grade classroom. In: Kirschner D, Whitson J A (eds.) *Situated Cognition*. Erlbaum, Mahwah, NJ, pp. 151–233

Collins A, Brown J S, Newman S 1989 Cognitive apprenticeship: teaching the crafts of reading, writing, and mathematics. In: Resnick L B (ed.) *Knowing, Learning, and Instruction*. Hillsdale, NJ, Erlbaum, pp. 453–494

Crocker J, Major B, Steele C 1998 Social stigma. In: Gilbert D, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, vol. 2, pp. 504–553

Fernandez-Duque D, Johnson M L 1999 Attention metaphors: how metaphors guide the cognitive psychology of attention. *Cognitive Science* **23**(1): 83–116

Gergen K J 1985 Social constructionist inquiry: context and implications. In: Gergen K J, Davis K E (eds.) *The Social Construction of the Person*. Springer, New York, pp. 3–18

Gopnik A, Wellman H 1994 The theory-theory. In: Hirschfeld L A, Gelman S (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, Cambridge, UK, pp. 257–293

Johnson-Laird P N 1989 Mental models. In: Posner M I (ed.) *Foundations of Cognitive Science*. MIT Press Cambridge, MA, pp. 469–499

Johnson-Laird , Legrenzi P, Girotto V, Legrenzi M S, Caverni J-P 1999 Naive probability: a mental model theory of extensional reasoning. *Psychological Review*. **106**(1): 62–88

Kelly G A 1955 *The Psychology of Personal Constructs*, 2 vols. Norton, New York

Kelly G A 1970 A brief introduction to personal construct theory. In: Bannister D (ed.) *Perspectives in Personal Construct Theory*. Academic Press, New York, pp. 1–29
Kintsch W 1988 The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. rev.* **95**(2): 163–182
Landauer T K, Dumais S T 1997 A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. rev.* **104**(2): 211–240
Lillard A 1998 Ethnopsychologies: cultural variations in theories of mind. *Psychol bull.* **123**(1): 3–32
Maturana H, Varela F J 1987 *The Tree of Knowledge*. Shambhala, Boston, MA
Mayer R E 1992 Cognition and instruction: their historic meeting within educational psychology. *Journal of Educational Psychology* **84**(4): 405–412
McNamara, D S, Kinstch E, Songer N B, Kintsch W 1996 *Cognition and Instruction* **14**(1): 1–43
Moreno R, Mayer R E 1999 Multimedia-supported metaphors for meaning making in mathematics. *Cognition and Instruction* **17**(3): 215–248
Nickerson R S 1999 How we know—and sometimes misjudge—what others know: imputing one's own knowledge to others. *Psychol. rev.* **125**(6): 737–759
Ortony A 1980 Metaphor. In: Spiro R J, Bruce B C, Brewer W F (eds.) *Theoretical Issues in Reading Comprehension*. Erlbaum, Hillsdale, NJ, pp. 349–365
Resnick L B 1994 Situated rationalism. Biological and social preparedness for learning. In: Hirschfeld L A, Gelman S A (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, Cambridge, UK, pp. 474–493
Rumelhart D 1980 Schemata: the building blocks of cognition. In: Spiro R J, Bruce B C, Brewer W F (eds.) *Theoretical Issues in Reading Comprehension*. Erlbaum, Hillsdale, NJ, pp. 33–58
Rumelhart D 1989 The architecture of mind: a connectionist approach. In: Posner M I (ed.) *Foundations of Cognitive Science* MIT Press, Cambridge, MA, pp. 133–159
Scarr S 1985 Constructing psychology. *American Psychologist* **40**(5): 499–512
Schwanenflugel P J, Henderson R L, Fabricius W V 1998 Developing organization of mental verbs and theory of mind in middle childhood: evidence from extensions. *Developmental Psychology*. **34**(3): 512–524
Sternberg R 1990 *Metaphors of Mind*. Cambridge University Press, Cambridge, UK
Thomas N J T 1999 Are theories of imagery theories of imagination? An active perception approach to conscious mental content. *Cognitive Science* **23**(2): 207–245
Wilson T D, Lindsey S, Schooler T Y 2000 A model of dual attitudes. *Psychol. bull.* **107**(1): 101–126

J. Gerstenmaier and H. Mandl

Constructivist Psychotherapies

As a philosophical position that emphasizes both personal and social processes of meaning-making, constructivism has influenced several contemporary traditions of psychotherapy. For this reason it is more accurate to consider constructivism as a general approach to understanding people, conceptualizing psychological distress, and fostering human change than to view it as a distinctive 'school' of psychotherapy associated with a particular theorist, preferred method, or specific set of problems requiring treatment. Thus, constructivism is best viewed as a 'metatheory' that encompasses many late twentieth century developments in clinical theories as diverse as psychoanalysis, existential-humanistic psychotherapy, cognitive-behavioral therapy, and family systems approaches. In addition, a number of novel psychotherapies have been devised along avowedly constructivist lines, from personal construct theory, which was formally set forth in the 1950s, to narrative therapy approaches that have become prominent only in the last 10 to 15 years. Appreciating the contributions of these diverse models therefore requires a consideration of their core philosophic similarities and the range of concepts and strategies that shape their expression at the level of clinical practice.

1. *Constructivist Philosophy*

If there is a unifying theme that links constructivist forms of psychotherapy, it is at the level of their epistemology, or theory of knowledge. Although most constructivists acknowledge that a 'real world' exists outside of human consciousness or language, they are much more interested in the nuances in people's construction of the world than they are in evaluating the extent to which such constructions are 'true' in representing a presumably external reality. This emphasis on the active, form-giving nature of the mind dates back at least to the Italian rhetorician and historian Giambattista Vico (1688–1744), who traced the development of thought to the attempt to understand the world by projecting upon it human motives, myths, fables, and eventually linguistic abstractions. The German philosopher Immanuel Kant (1724–1804) likewise emphasized the transformative character of the mind, which necessarily imposes spatial, temporal, and causal order on the phenomena of experience. From these philosophers, constructivists borrowed a model of knowledge as an active structuring of experience, rather than a passive or receptive assimilation of a 'noumenal' reality of 'things in themselves,' uncontaminated by human knowing.

At the threshold to the twentieth century, these themes were elaborated by the German analytic philosopher, Hans Vaihinger (1852–1933), whose Philosophy of 'As If' asserted that people develop 'workable fictions' (e.g., of mathematical infinity or God) to order and transcend the hard data of experience, and establish distinctively human goals. A similar emphasis on the distinction between our linguistic 'map' of experience and the 'territory' of the

world was made by the Polish intellectual Alfred Korzybski (1879–1950), whose system of general semantics focused on the role of the speaker in assigning meanings to events. From these thinkers, constructivists drew the implication that human beings operate on the basis of symbolic or linguistic constructs that help them navigate in the world without contacting it in any simple, direct way. Moreover, they suggested that such constructions are viable to the extent that they help us live our lives meaningfully and find validation in the shared understandings of others in our families, communities, and societies. 'Postmodern' thinkers who follow in this constructivist vein further stress the extent to which we live in a world constituted by multiple social 'realities,' no one of which can claim to be 'objectively' true across persons, cultures, or historical epochs. Instead, the constructions on the basis of which we live are at best provisional ways of organizing our 'selves' and our activities, which could under other circumstances be constituted quite differently.

2. *Constructivist Psychotherapies*

The first person to develop a thoroughgoing theory of psychotherapy along constructivist lines was the American clinical psychologist, George Kelly. Working in the relative isolation of rural Kansas in the 1930s and 1940s, Kelly confronted the overwhelming psychological needs of farming communities that had been devastated by the twin crises of the Dust Bowl and the Great Depression. This prompted Kelly to design efficient psychotherapeutic procedures in which clients were coached to enact fictional identities in their daily lives for a fixed period of time (usually only 2 or 3 weeks). Working from a 'self-characterization' written by the client, the therapist and a consulting team first drafted an 'enactment sketch' of an imaginary person whose way of construing life differed from, but did not necessarily contradict, the client's own. The client and therapist then practiced this novel role in a series of therapeutic enactments with the therapist playing various acquaintances, authority figures, and family members, while the client simultaneously generalized the role to the actual relationships in his or her life. At the end of this period, the role was consciously set aside, and the therapist and client discussed the extent to which the client's sense of self and social world were permeable and provisional constructions that might be lived out quite differently if viewed in alternative ways. Kelly's 'fixed role therapy' was therefore the first form of brief therapy, and it foreshadowed the use of dramatic and narrative strategies of change incorporated in many contemporary constructivist therapies. Eventually, Kelly drafted a comprehensive psychology of personal constructs that placed these procedures in a rigorous theoretical context, and suggested diagnostic, therapeutic, and research methods targeting the unique meaning systems that individuals devised to structure and anticipate the themes of their lives. Interest in Kelly's work remains strong throughout North America, Europe, and Australasia, with contemporary personal construct theorists challenging, broadening, and applying his initial theory to a range of current problems and disorders.

A second tradition of constructivist therapy began to gain momentum in the late 1970s as an outgrowth of cognitive and cognitive-behavioral therapies emphasizing the role of the person's interpretations of events in a range of disorders. Michael Mahoney, an American pioneer in the cognitive trend, began to critique its reliance on rationalistic epistemologies that presumed that emotional adjustment was a straightforward matter of making one's cognitions realistic and in line with an observable world. In particular, Mahoney, along with colleagues such as the Italian theorist Vittorio Guidano, began to turn attention to the 'core ordering processes' by which persons construct and maintain a sense of self in a social field. This gave rise to more developmental forms of constructivist psychotherapy, which focused on the periodic perturbation of an individual's conscious identity through confrontation with difficult or traumatic live events. At a strategic level, these therapies involve a variety of procedures for enhancing clients' self awareness and helping them reconstruct an emotionally coherent sense of their identities across time. Significantly, this emphasis on identity construction redirected attention to the often tacit, intuititive, and non-conscious processes by which people organize their experience of self and others, extending the more common focus of cognitive-behavioraltherapies on people's conscious 'inner dialogue' or 'automatic thoughts.' By the 1990s, a focus on cognitive-emotional schemes of self-and-relationships had begun to augment more traditional cognitive therapies, and even initially rationalistic approaches such as the American psychologist DonaldMeichenbaum's cognitive-behavioral modification began to be revised along constructivist lines (see *Cognitive Therapy*).

As the more relativistic philosophy of constructivism became widespread, it began to permeate even long-standing traditions of psychotherapy, such as psychoanalysis. One harbinger of this trend was the American analyst Donald Spence who argued that analysis could not unearth the historical truth of the patient's life, but only constitute a procedure for disclosing its narrative truth in the eyes of its author—the client. In keeping with the psychoanalytic tradition, however, constructivist analysts still tend to emphasize the usefulness of accessing and reviewing emotionally significant memories of one's early life, but view these less as veridical 'insights' than as inventions subject to the demand for narrative

'smoothing.' One interesting recent expression of this approach is the depth-oriented brief therapy developed by the American psychologists Bruce Ecker and Laurel Hulley, which uses a procedure of radical questioning to reveal the hidden implications of the client's pro-symptom and anti-symptom positions. Like other contemporary constructivist approaches, this integrative model of therapy relies more heavily on experiential exploration of the client's meanings (e.g., using visualization exercises) than on the interpretive stance preferred by classical psychoanalysts.

Another long-standing tradition to be influenced by a constructivist position is humanistic-existential psychotherapy, whose emphases on human choice, agency, and phenomenology or the study of experience converged with core themes in personal construct theory and related forms of constructivist theory. One outgrowth of this convergence is the self-confrontation method devised by the Dutch personality theorist Hubert Hermans which uses open-ended questions to elicit those 'valuations,' or important units of meaning that a person uses to structure a sense of her or his past, present, and future. The client then rates critical life experiences on a standard list of affect terms, which yields measures expressing the degree of self-enhancement, union with others, negative emotions, and positivity associated with each, providing a basis for self-reflection or a method for mapping therapeutic changes in meanings over time. A second example of the infusion of constructivist themes into humanistic therapy is the dialectical approach of the Canadian psychologist, Leslie Greenberg, which emphasizes the role of personal experience and internal conflict in the creation of subjective meaning (see *Experiential Psychotherapy*). For example, therapy might attempt to synthesize two strongly-felt and mutually contradictory internal experiences, such as self-contempt and the self-prizing. This could involve helping the client vividly experience both stances through the use of a 'two chair' technique, in which the client successively speaks from each position and fosters their integration into a new structure (e.g., self-acceptance). In keeping with constructivist themes, the 'selfhood processes' that are the focus of theseapproaches are viewed as continually evolving andprovisional, rather than a static 'essence' to be simply discovered or actualized.

Finally, family systems therapies have been revolutionized by a social constructionist perspective, adopting a non-authoritarian view of therapy as conversation whose goal is to alter what the American psychotherapists Harlene Anderson and Harry Goolishian termed the 'problem determined system.' In such approaches, the therapist functions less as an expert dispensing answers, than as a conversation manager who promotes the sort of exchanges among family members that 'dissolve' old problems by 'languaging' about them in a new way. Other family therapists have adopted the constructivist metaphor of lives as stories, and have devised novel means of helping clients free themselves from the 'dominant narratives' that originate in particular families and cultures, and that keep them from feeling like the authors of their own lives. For therapists such as Michael White in Australia and David Epston in New Zealand, externalizing the problem by reconstruing it as something separate from the client's 'self' provides a useful first step toward recognizing its destructive impact on clients and the relationships in which they are engaged. Therapy can then turn toward recognizing and validating those exceptional moments when a person begins to resist the dominant narrative, and rewrite his or her life story along more hopeful lines. A common theme in these systemically informed variants of constructivism is their emphasis on the resourcefulness of individuals, families, and communities in resolving problems in living. Accordingly, clinical work is cast as a search for solutions, rather than a diagnosis of dysfunction.

3. *Critical Perspectives*

Despite the creative contributions of constructivism to contemporary clinical practice, challenges to this perspective have been raised by realists who question what they view at its 'anything goes' relativism. From the vantage point of these critics, the constructivist emphasis on multiple realities risks undermining conventional understandings of 'truth' as well as 'objective' procedures of research into therapy process and outcome. However, constructivists have responded that a respect for the ways in which various people and cultures construe life differently accords well with the diversity and multiculturalism of contemporary life, and actually provides a psychological basis for twenty-first century ethics. Likewise, constructivists have noted that, their criticism of naïve realism notwithstanding, psychologists working within this tradition have carried out literally thousands of empirical studies, on topics as diverse as the structure of personal meaning systems and the intricacies of change in the process of psychotherapy. Significantly, these efforts have often required the development of new tools for revealing individual and conversational meaning-making, such as the role construct repertory grids devised by personal construct theorists, and the narrative process coding schemes validated by constructivist psychotherapy researchers. Thus, it seems unlikely that constructivism will erode the disciplined study of psychological phenomena in an uncritical attempt to embrace all knowledge claims as equally valid.

Tensions also arise within the constructivist camp, particularly between theorists who champion the individual's role in meaning-making, and those who emphasize the extent to which meanings reside in systems of culture and language that precede any given individual. Indeed, the debates engendered by the

latter social constructionist position are likely to continue, as they tend to undermine the straightforward depiction of the self-as-agent endorsed by constructivists with more humanistic leanings. This same emphasis on the social and discursive construction of reality, however, has sensitized constructivism to issues of familial, ethnic, and cultural meanings, thereby enhancing its relevance and inclusiveness. Ultimately, the challenges faced by constructivist theories and therapies can be viewed as the predictable 'growing pains' encountered by any significant new perspective that promises to help shape the future, as well as the present, of psychotherapy.

See also: Behavior Psychotherapy: Rational and Emotive; Constructivism/Constructionism: Methodology; Constructivism in Cognitive Psychology; Eclectic Psychotherapy: Psychological Perspectives; History of Science: Constructivist Perspectives; Interpersonal Psychotherapy; Person-centered Psychotherapy

Bibliography

Angus L, Levitt H, Hardke L 1999 The narrative processes coding system: research applications and implications for psychotherapeutic practice. *Journal of Clinical Psychology* **55**: 1255–70

Ecker B, Hulley L 1996 *Depth-oriented brief therapy*. Jossey Bass, San Francisco, CA

Eron J B, Lund T W 1996 *Narrative solutions in brief therapy*. Guilford, New York

Franklin C, Nurius P S (eds.) 1998 *Constructivism in practice*. Families International Press, Milwaukee, WI

Guidano V F 1991 *The self in process*. Guilford, New York

Hermans H J H, Hermons-Jansen E 1995 *Self-narratives*. Guilford, New York

Hoyt M 1998 *Constructive psychotherapies*. Jossey Bass, San Francisco, CA

Journal of Constructivist Psychology. Taylor & Francis. Philadelphia, PA

Kelly G A 1955 *The Psychology of Personal Constructs*. Norton, New York. Reprinted By Routledge, London, 1991

Mahoney M J 1991 *Human Change Processes*. Basic Books, New York

McNamee S, Gergen K J (eds.) 1992 *Therapy as Social Construction*. Sage, Newbury Park, CA

Neimeyer R A, Mahoney M J (eds.) 1995 *Constructivism in Psychotherapy*. American Psychological Association, Washington, DC

Neimeyer R A, Neimeyer G J 1990–2000 *Advances in Personal Construct Psychology* (Vols. 1–5). JAI/Elsevier, Greenwich, CT

Neimeyer R A, Raskin J (eds.) 2000 *Constructions of Disorder: Meaning-making Frameworks for Psychotherapy*, 1st edn. American Psychological Association, Washington, DC

White M, Epston D 1990 *Narrative Means to Therapeutic Ends*, 1st edn. Norton, New York

R. A. Neimeyer

Consumer Culture

The term consumer culture refers to the culture of the consumer society. It suggests that the representations (signs and images) and values of contemporary societies revolve around consumption: the purchase and enjoyment of goods for the construction of lifestyles. Advertising is important here in supplying a whole repertoire of images and representations of the good life that have become associated with particular commodities and brands. Consumer culture is driven by a modernist dynamic to expanding the range of goods and meanings on offer through the search for novelty, technological innovation, and the exotic. Many of the activities of everyday life are dominated by the consumption of goods and experiences in a variety of sites such as malls, city centers, department stores, theme parks, tourist resorts, heritage centers which offer a range of simulations, spectacular spaces, and sign play. These 'dream worlds' lead to the aestheticization of everyday life with the promise of new experiences, the loosening of emotional controls, and the exploration of desire. At the same time, consumer culture also provides resources for lifestyle construction and the promise of self-improvement and renewal: for reworking the body, identity, and relationships. Resources for repositioning the self within a consumer world in which the constant supply of new material goods, signs, and images make the task of classification, of 'reading the signs,' and locating people and things within stable status hierarchies, all the more difficult and compelling.

1. History of the Term

Production and consumption are often seen as fundamentals of social life. All societies produce some forms of material goods that are exchanged and consumed. Yet the identification, theorization, and naming of specific activities as production, exchange, or consumption is of relatively recent origin and takes its impetus from the emergence of the discipline of economics since the eighteenth century. Historically, the weight of interest has been upon production and exchange, revolving around questions of delineating the general forms and 'laws' seen as governing economic activities. From this perspective, consumption remains relatively unproblematic: people are seen as having 'natural' or infinitely expanding needs and desires for new and better goods. The object of all production is assumed to be consumption, yet why and how people consume, what they do with the consumer goods, given the premise of rational economic action, has only been of peripheral interest, as for example in allegedly marginal situations where people refuse to consume, or purchase and consume 'irrationally.' The sociological critique of economics which

developed in the late nineteenth and early twentieth centuries, focused on the limits of rational action as the sole mode for explaining social life. Durkheim, Weber, Marx, Simmel, Pareto, Parsons, and Veblen analyzed the social and cultural parameters within which economic action was embedded and in various ways began to raise questions of consumption.

For many intellectuals and academics the topic of consumption remained intertwined with discussions of mass culture. The logic of capitalist mass production was seen as leading to a proliferation of standardized consumer goods, in which the aim of marketing was to generate uniformity of taste. Hence culture became produced by industrialized techniques and became a commodity like any other. The culture industries sought to manipulate and control consumption, which became associated with passivity and massification leading to the elimination of individuality and creativity. With the post-World War II society development of a fully-fledged consumer society, especially in the United States, Japan, and parts of Europe, the analyses and critiques of consumption gathered momentum. In the late 1950s and 1960s a number of works came into prominence which sought to diagnose and challenge the logic of the 'affluent society' with its perceived manipulative advertising, status seeking soulless conformism, and suppression of alternative definitions of the good life (Packard 1958, Galbraith 1961, Marcuse 1964).

The term consumer culture itself seems to have originated in the 1970s and gained impetus from the work of Ewen (1976) and Bell (1976) who added historical and analytical depth to these more polemical critiques. Yet the term had little impact on the majority of scholars at the time. In the field of sociology, the sub-fields of the economy, industry, class, and power had long dominated the center of the field. The general tendency in sociology was to operate with a reductive and epiphenomenal view of culture. Yet from the late 1970s onwards the study of culture, everyday life, gender, emotions, leisure, and consumption that had long been marginalized and confined to the periphery, now became seen as legitimate areas of study and began to move towards the center of the field. A key influence here in the 1980s was postmodernism which problematised the role of culture in social life with its focus on the instability and mixing of codes, the overproduction of sign and images, along with the effacement of the boundary between art and everyday life. For some critics postmodernism became seen as the 'cultural logic' of the consumer society, reinforcing the link between image saturation, consumption, and postmodern fragmentation (Jameson 1984, Baudrillard 1993).

Yet for others, the negative evaluation of consumer culture became more problematic. This was especially the case for those writing from within cultural studies, anthropology, and women's studies, who sought to contest taste hierarchies which denigrated popular culture, and allegedly wasteful and mundane activities such as watching television soap operas or shopping. There was also an interest in examining the roots of modern consumption by historians and sociologists, who refused to see the consumer society as an unproblematic outgrowth of industrialization, rather for some a consumer revolution was seen to precede the revolution in production. The origins of consumer culture were traced back to eighteenth-century England (McKendrick et al. 1982), sixteenth-century Elizabethan England (McCracken 1988), and even linked to the rise of a romantic ethic (Campbell 1987). Questions about the particularity of Western consumer culture and the rise of consumer cultures in the big cities of Song and Ming Dynasty China and Tokagawa Japan have also begun to be raised (Myoshi and Harutoonian 1989), along with a greater sense of the longstanding historical centrality of East Asia in an interconnected global economy (Gunder Frank 1998).

In the 1990s the investigation of consumer culture became a part of the legitimate study of culture, as the importance of consumption in providing resources for identity construction and forms of sociability, both mainstream and oppositional, became acknowledged. In response to this more nuanced sense of the role of consumption, a spate of publications, textbooks, edited collections, monographs, and journal articles occurred, along with the founding of specialist journals on consumption. The term began to circulate more widely, gaining legitimacy and currency in the media through the efforts of cultural intermediaries who used the term in exhibitions, museums, and popular culture. Consumer culture has increasingly become a part of everyday language.

2. *Material Goods, Cultural Classification, and the Body*

A central feature of consumer culture is the proliferation of material goods. While people in all societies use material goods, their production, exchange, and circulation are usually limited and surrounded by strong rules and rituals. There have been many instances of sumptuary laws that restrict the consumption of goods, such as the wearing of certain types of clothes (e.g., in the European Middle Ages silks, velvet, lace, bright colors were forbidden to peasants and other groups at the lower end of the social spectrum). The development of a culture around consumption is premised upon the removal of restrictions to access for goods, and the assumption that the capacity to pay the price in the market is the sole arbiter of use. In addition, it is based upon the assumption that the volume and significance of goods in circulation steadily increase, that mass produced goods become readily available and used in everyday

practices to play a central role in the construction of narratives and identities. A key feature here is novelty: new goods are constantly being produced and sought after; something that increases the amount of information in circulation that is necessary to make sense of new goods and integrate them into lifestyles. Hence advertising and background discussions aimed at helping to educate consumers into new tastes, take place within mass media such as newspapers, magazines, television, and radio programs.

The consumer society and its culture are often criticized for materialism: that people thirst after new goods to the neglect of the spiritual side of life. Yet it can be argued that the proliferation of goods has not been driven by a relentless empty materialism or greed, but has as much to do with cultural meaning systems. It is the relational matrix of the set of goods that is important, the ways in which goods are used not only to mark status divisions, but also to delineate the boundaries of social relationships and cement in-groups against out-groups. In effect goods can act as both doors and bridges to exclude or build relations with others. This suggests that consumer goods can be seen as resources for classifying the social world in which the logic of signs and cultural categories are used to create or reinforce social distinctions between groups. It also pointed to the process of investment in acquiring the skill to manipulate, read, and display the taxonomies of goods, dress, and demeanor: an approach which has its basis in anthropology (Douglas and Isherwood 1980, Bourdieu 1984).

Yet it can be argued that the classification system of consumer culture is inherently unstable, it has all sorts of gaps and ambiguities when compared to the relative closure of the primitive classifications described by Durkheim and Mauss, or the classical systems of science as described by Foucault. Here we can point to both the quest for innovation and novelty driven by the dynamics of the fashion system and the impact of new technologies that constantly disturb taxonomies. If goods have been seen from the point of view of economics as possessing both a use value (in terms of the purpose or need they fulfill) and an exchange value (monetary value), the progressive development of consumer culture complicates the picture. Advertising, it can be argued, attaches a further meaning to the commodity, a secondary 'ersatz-use-value' or 'sign-value,' which can occlude the original use (Adorno, Baudrillard). The replacement of use value by sign value implies that the meaning of goods becomes unstable, given the vast repertoire of themes and imagery (romance, exotica, beauty, adventure, the good life, excitement, fulfillment, solidarity, etc) which advertisers can draw upon.

There has been a general shift within the history of advertising away from direct description of the qualities of the product, to the use of more indirect associational imagery, in which meaning becomes elided and goods become associated with vague, but potent notions of lifestyle. The instabilities of the imagery associated with goods and the expansion of the media through which imagery can be created and recreated (especially with television, and the newer electronic forms such as the Internet) are seen by some as leading to an overproduction of signs and images. Such 'culturally saturated societies,' are often characterized as typically postmodern, with their fluid, depthless culture in which meaning refuses to stabilize or cohere (Baudrillard 1993, Jameson 1984). The oft-quoted archetypal example here of postmodern sign play is MTV (music television), with its endless series of pop videos in which images constantly change and flow into one another before the viewer can recover them into a coherent narrative.

Yet postmodern sign play constitutes just one end of the continuum of consumer culture activities. At the other end are highly formalized and ritualized consumption practices, in which every act and body movement, such as in eating, drinking, dancing, is heavily coded and carries weighty significance for the participants. The elaborately coded consumption of court societies would be a good example here. The latter example also helps us see that while consumption can entail watching the media in private, a good deal of consumption is more directly social; it involves others, people who come together in social practices in which there are varying demands for reciprocity. Consumer culture offers a large repertoire of resources for the construction and renewal of identities and presentation of the self as people move between public and private settings. Hence consumer culture promotes a world in which spatial and social mobility feature strongly: individuals are invited to imagine themselves operating in different settings, with different people and trying out different experiences. In an anonymous urban world, appearances become more important, and the possibility of 'passing,' of masking, reworking appearance, becomes both possible and acceptable. The values of personality, the development of charm, fascination, attractiveness become preferred to the more traditional virtues of character: integrity, duty, hard work, and thrift. Consumer culture through magazines, newspapers, and self-help books offers a mass of literature on self-improvement: how to reform the self, the body, sex life, and relationships. Here the body becomes more important as its 'seen and seeing' quality becomes central to the construction of what is heralded as an essentially embodied self. In the consumer culture view of the world the connection between body image and self image is presented as a crucial one, with the positive benefits of 'look good, feel good' body maintenance health and fitness routines extolled. The body then is a key resource for identity building (Featherstone 1982).

At the same time we should not underestimate the ways in which individuality and conformism are both presented as necessary, something that requires a complex integrative structure to achieve the right

balance for each situation. Consumer goods can be used as resources to cement collective relationships and renew the social order: wearing the right clothes and demeanor for each occasion, executing the correct performance, having appropriate accessories and material objects to hand which stamp the person as part of a status group. In Japan, arguably one of the most highly developed consumer societies, these aspects are very important, with consumer culture goods used to reinforce stable and lasting relationships (Clammer 1997). Department stores in Japan are big promoters of gift giving and provide elaborate services to help choose, package, and deliver the appropriate gift to cement relationships at work, in the family, between friends, and for a whole range of intermediaries and service providers. Consumer culture then, needs to be understood in its different social contexts, and we cannot assume that everything that occurs within Western modernity is immediately exportable around the world. Self-cultivation, reflexivity, and life projects, which are associated with consumer culture in the West, cannot be assumed to develop readily everywhere, and the instabilities and transgressive sign play associated with the 'dream worlds' of consumption cannot be assumed to be always as threatening to identities and the social order.

3. *Spaces of Consumption, Shopping, and Gender*

A key aspect of consumer culture has been the development of new spaces for consumption, shopping, and entertainment: the arcades, city centers, and department stores which originated in the nineteenth century and the wider range of sites such as cinemas, cafés, resorts, leisure, and theme parks and malls that developed in the twentieth century. These were sites for purchasing goods and experiences, for looking at goods, for seeing and being seen, for enjoying the swirl of the crowd and the reading and misreading of the rich fabric of urban life with its vivid color advertisements, street decorations, and neon signs. They were relatively safe spaces that enabled women to venture out into public in greater numbers. The influential writings of Benjamin (1999) and Simmel (1997) provide a view of the experience of modernity as an essentially urban one, in which the crowds, movement, shocks, provided a bombardment of the senses with new impressions. The sensitivity to the heightened visual quality of modern life is captured in the *flâneur*, the stroller, who first appeared in the new consumer spaces of the nineteenth century to take in the fleeting fragmentary impressions of the crowds, the street, the buildings, and minutiae of the urban landscape.

If advertising was important in educating consumers into new tastes and desires, the department stores which appeared in Europe, the United States, and Japan in the second half of the nineteenth century and the early decades of the twentieth century, can be seen as playing an equally important role in the formation of consumer culture. Unlike traditional shops, the department store was a space in which it was possible to walk around and examine goods on display. A key feature here was the opulence of the setting; department stores were designed spaces that sought to house the expanding range of cheap mass-produced consumer goods in novel ways. They were built on a massive scale with impressive architecture, layout, and decor as new public-private spaces that invited the consumer to stroll around, to stop and look, to take in the latest goods framed in exotic and luxurious display settings. They helped to expand the time frame of shopping so that it could become a pleasurable activity in itself, whether it be the actual purchase of goods or the activity of 'window shopping,' 'just looking,' and experiencing the 'Hollywood set' quality of the displays. In addition, stores saw themselves as artistic and entertainment places where people could witness performances and spectacles, such as ballet, concerts, art exhibitions, and fashion shows. They also advertised themselves as communal centers, places where women could meet, engage in sociability, and even spend the whole day. It is no accident that department stores were referred to as 'dream worlds,' 'fantasy palaces,' and 'cathedrals of consumption' through their capacity to dominate the urban landscape and generate and legitimate a new consumer gaze and set of pleasures (Nava 1997, Williams 1982).

With the emergence of city centers as sites of consumption in the late nineteenth and early decades of the twentieth centuries, women were able to move more freely in the new public and public-private spaces. Department stores and other shopping, entertainment, and expanding service sector sites such as cinemas, restaurants, hotels, cafés, also provided new opportunities for women's employment. These were all places where emotional labor and presentational skills were important, and where women had to learn how to look and be looked at. Places of encounters with customers and members of the public, and of the viewing of the latest fashions and spectacular displays, provided new forms of knowledge and embodied competencies which could foster confidence and independence and generate a sense of an adventurous more exciting life. This formed the basis for a new definition of gender incorporating both a greater sense of self-control and expressivity. Something, which reached a peak in the 1920s with the flapper, the hedonistic young woman who had bobbed hair, wore a short skirt, and hyperactively danced to the jazz beat, which started to circulate around the world through the efforts of the nascent Hollywood film industry with movies such as Clara Bow's *It!* Consumer culture started to globalize: in Japan for example the department store became a central institution and along with the cinema, helped to provide stimulus for the emergence of a new form of femininity, the *moga* girl,

the modern girl who incorporated some of the elements of the flapper (Silverberg 1991)

The popularity of cinema going in the early decades of the twentieth century was important, as the majority of movie audiences were women. Like the department store the cinema was a legitimate place where women could come together in public relatively free from the surveillance of men. Like the department store it was a site that nurtured dreams and desires. It was also a powerful vehicle for fostering identification with others across gender, class, age, and ethnic lines along with the simulation of emotions. Something which has given rise to the view that the cinema helped educate women into new forms of public life which can be seen as having positive social implications. In effect the cinema led to the development of a cinematic public sphere, a key element in the formation of a women's public sphere (Hansen 1991). Here a sense of public responsibility for the other could be seen as not merely developing through the alleged rational argumentation of the (male) public sphere, but through emotional identifications and aesthetic experiences.

This positive view of the ways in which consumer culture spaces offering greater freedom and self-expression for women squares uneasily with accounts of consumption and mass culture as forms of deception, offering seductive and intoxicating pleasures. In accounts, such as those of the Frankfurt School, the fear was that mass culture and consumer culture amounted to forms of regression, with the dangers of loss of emotional control and the dissolution of the ego and social boundaries, and even the loss of civilized codes and sexual abandon (Huyssen 1986). This perspective misses the ways in which the decontrol of the emotions can take place in controlled ways, in which the swings between rigid rational control, and dangerous unbounded emotions are managed in subtler, playful, and less threatening ways. At the same time it can be argued that this capacity for emotional decontrol developed in men too, with men increasingly asked to cultivate their appearance and engage in bodily maintenance. In the second half of the twentieth century with the expansion of consumer culture, we have the emergence of new masculine ideals, the celebrities of leisure, sport, and the media who cultivate 'personality' (Featherstone 1995).

Shopping offers an interesting impure hybrid form: it mixes purposiveness, pleasure, and play. It is also often seen as offering the short-term pleasures of the purchase and conspicuous consumption, the vulgar tasteless pleasures of the new rich, in which high price is taken as the only guide to the real value of goods. Something which can be contrasted to the elaborately coded taste hierarchies and finely graded distinctions of the connoisseur or collector, who works hard to place the object in a context and assess its immanent value. Yet this perspective misses the ways in which particular forms of shopping can involve discrimination and careful judgment of taste. When people shop for clothes, for example, it can often entail the use of elaborately coded knowledge of the fashion system and the latest styles. Here the pleasures of knowledge, of manipulating the classification system come together with the pleasures of the body, with knowledge of one's own bodily form and what suits it. A different form of pleasure from flicking through a mail order catalogue or surfing through the Internet's 'click and purchase' shopping sites. Going out to shop puts the emphasis upon performance and enactment. Crucial here is the capacity for movement in spaces, for the sensual (especially visual) pleasures of physical movement, of handling and touching the goods. In the postwar era, malls and shopping centers have become important consumer sites through their capacity to offer secure, comfortable spaces for shopping, entertainment and sociability (see *Entertainment*). While recently there has been some decline in the popularity of malls in the United States and a shift back to city centers, with the expansion of gated communities and globalization of divided, more dangerous cities, malls have expanded in many parts of the world. They have become important secure public-private spaces largely for the new middle classes, and particularly for women, to engage and experiment with consumer culture lifestyles and new forms of sociability.

4. Aestheticization Processes and the Cultural Intermediaries

Consumer culture can be seen as offering and legitimating a wide range of aesthetic experiences and bodily pleasures, something that has become designed into goods and consumer spaces by the growing ranks of cultural intermediaries. Yet this involves a highly differentiated set of practices with the rhythms of the day, the week, the year pulling different spaces of consumption into view. Buying a birthday present or gifts on holiday takes different types of judgment about price, taste, and aesthetics than the weekly supermarket visit or the purchase of a utility such as a washing machine. Consumer spaces include the warehouse form of hypermarket such as Wal-Mart, a de-aestheticized environment with minimal decoration and display setting, where cheapness and 'value-for money,' are everything. On the other hand, they also include sites in which every effort is made to give goods or a brand a quasi-sacred significance, as for example we find in the Nike Museum in Chicago, where trainers are displayed like art objects.

It is possible to envisage a continuum here, ranging from sites in which standardization of infrastructure and product are to the fore, where people are processed in and out rapidly, to sites where the consumer is invited to linger and have an experience, where consumption involves consuming the setting, the excitement, and aesthetic pleasures of discovery of

novelty. At the standardized end we have McDonald's and Wal-Mart, the phenomenon of the chain or brand that offers the simplification of complex consumer choices through its alleged reliability and value for money (see *Fast Food Enterprises*). Near this end are the 'no place' spaces of airports, chain hotels, and resorts, which also seek to offer the insecure customer, who has no time or inclination to browse or experiment, predictability and reliability.

At the other end we have the experiential sites that offer the ambience of 'amazing spaces;' sites where every effort is made to disguise standardization and uniformity and cultivate the exotic and the new. Here the pleasure can be both in the participation in the experience, or in anticipation through learning processes (browsing guidebooks, planning, imaginative constructions, and dreaming). Opening up exotic worlds through foreign travel and exploring art, counter-cultural, and bohemian experimental lifestyles can involve not just excitement, but also the risk of identities and relationships. Yet, consumer culture seeks to market the risk and adventure too, both directly, through the 'ultimate experience' of high risk sports, such as paying large sums of money to be guided to the top of Mt. Everest by companies such as Mountain Madness or Adventure Consultants, and indirectly in all sorts of more controlled simulations such as 'white knuckle rides.'

The provision of various blends of excitement and novelty along with standardization and the familiar is found in many consumer culture sites: ranging from theme parks, resorts, hotels, bars, restaurants, shops, malls, to museums and heritage sites. Consumer culture's capacity to offer 'dream worlds,' reaches a new high point with theme parks such as Disneyland, where 'imagineers' have been able to produce hyper-real full-scale simulations of fairy tales, novels, and historical settings in which visitors are invited to 'step inside the dream,' and enjoy the movie set quality and full sensory immersion of another world. The retro bar, restaurant, shop, and even mall can work off this logic of seeking to construct spaces which provide a unique simulated experience. They work parasitically off the theme park, heritage park, and movie set design skills in constructing or reconstructing another world. They draw on content from a wide range of sources: from sacred tourist and historical sites, to the vulgarity of Las Vegas strip's popular culture, to the vitality and creativity of bohemias and urban subcultural spaces. The logic here is more one of seeking to provide mass customization than mass consumption. The capacity built into the infrastructural building technologies and the media image technologies is increasingly one that permits not only standardization and reproduction, but also reformatting and re-creation. Today's digital technologies can easily and rapidly spin out variations of codes, not just to provide the multimedia aspect of the experience; digitalization also becomes an integral element of the overall design process as architectural design itself increasingly works through virtual and simulated forms.

Consumer culture can be seen to work in a dual way through representations and sites. It draws on a wide range of representations to construct a culture around material goods, weaving a range of imagery and sign play to make commodities more enticing and exciting (advertising in its various forms being the key here) and provides a whole range of publicity material to educate consumers into enjoying new tastes. It also requires specific spaces to sell, display, and house goods and experiences (hypermarkets, department stores, malls, theme parks, resorts, etc.). Yet many of the sites become enriched with representations, with attention given to the décor and the setting to the extent that the act of purchase becomes an experience or that what one purchases or seeks out is an experience. If consumer culture has helped to swell the ranks of service sector occupations, then within this group design and associated cultural intermediaries have become increasingly salient. These groups, which include those members of the new middle classes working in the media, fashion, design, advertising, marketing, and the culture industries, are those who design the consumer culture representations, the informational aspect of consumer goods, and the experiential dimension of consumer sites.

These new cultural intermediaries have been referred to as 'para-intellectuals' given their admiration for artists and intellectuals and their constant search for knowledge of new and different lifestyles, sensations, and cultures (Bourdieu 1984, Featherstone 1991b). Their predisposition to explore the full range of global cultures, to seek out new styles, or recycle traditions, and search for the exotic through travel makes them have dispositions that could be labeled postmodern, through their aesthetic interest in playing with signs and cultures. Like artists and intellectuals they have an interest in travel and experimental bohemian lifestyles, to provide the stimulus for new ideas and themes to work into consumer culture imagery. Here, the city as the home of both the bohemian and the *flâneur* is seen as an important repository or database, in providing some of the dangerous and exciting places and personae with which to generate new cultures and lifestyles. The new cultural intermediaries have been in part responsible for the 'democratization of bohemia,' for feeding transgressive bohemian themes and images of the sexualized or fantasy city back into everyday consumer culture (Pels and Crebas 1987, Wilson 1998).

5. *Globalization and the Limits of Consumer Culture*

It can be argued, that since the 1960s, the pace of the globalization of consumer culture has increased (Featherstone 1995). This has been particularly evi-

dent since the 1980s with the lowering of tariffs through GATT and the WTO, state deregulation, and the extension of the market. The collapse of the Soviet bloc furthered the extension of market capitalism. Now the consumer culture, however unevenly it has occurred within particular societies, has become the main provider of images of the good life. Globalization has also meant the rise of the information order, with spatially dispersed more elaborate production and consumption networks coordinated through the use of electronic communication. Consumer brands and imagery have become globalized, yet the vast majority of humanity, who cannot gain access to the loop, can only experience consumer culture on this level of images and dreams.

Yet what if the excluded were to become included? Would this mean that we run into the limits of the consumer society? Some argue that we have already reached the limit, that it is not merely resource depletion, but global warming and the disturbance of the planetary ecological balance that are occurring. If global 'bads' are seen to accompany the spread of global goods, then it is the unintended consequences of unregulated consumption, the accumulation of risks through the drive for increased technological innovation and scientific progress, which have become evident. The consumer society, then, is also a risk society and increasingly we are becoming conscious of the global scope of risk chains and networks. At the same time there would also seem to be the beginnings of the emergence of a global public sphere, where concerns about the pace of change, the exploitation of natural resources, the dangers of the biotechnological revolution, are starting to trouble some groups. Something akin to global public opinion is in the process of formation, as we saw in 1999 in Europe and the United States with the reaction against genetically modified food. The Internet with its speed, cheapness, and scope is becoming important here in providing consumers with the means for developing associations, networks, and campaigns that can raise questions about the need for more responsible forms of consumption.

Yet consumer culture is a long way from becoming a conserver culture. The dynamic of consumer culture is still one in which the search for new sensations and change dominate. Increased global flows of culture and the mobility of images through the media and information technology are accompanied by increased mobility of people, whether in pursuit of work or tourism. The demand for greater speed through information and transportation technologies has serious resource implications, along with new unintended risks and integral accidents. Serial and multiple lives in which people seek to experience intensified more exciting sensations of short duration, fit well with consumer culture. This seems to have been one of the unintended legacies of artists, intellectuals, and bohemians, however much the content of their work seeks to provide an urgent critique of consumer culture. Cultural modernism and postmodernism betray their origins in Western modernity and have yet to work through the dynamic of global post-humanism and a planetary humanism, in which the responsibility is to think the limits of 'the new,' the limits of mobility, movement, and the consumer. If there is something beyond the modern, then it implies thinking through new forms of consumption and developing a different relation to the object world of nature and technology.

See also: Advertising and Advertisements; Advertising: Effects; Advertising: General; Consumer Economics; Consumer Psychology; Consumption, Sociology of; Culture, Sociology of; Entertainment; Fast Food Enterprises; Leisure and Cultural Consumption; Leisure, Psychology of

Bibliography

Baudrillard J 1993 *Symbolic Exchange and Death*. Sage, London
Bell D 1976 *Cultural Contradictions of Capitalism*. Heinemann, London
Benjamin W 1999 *The Arcades Project*. Belknap Press of Harvard University Press, Cambridge, MA
Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste*. Routledge, London
Campbell D 1987 *The Romantic Ethic and the Spirit of Modern Consumerism*. Blackwell, Oxford, UK
Clammer J 1997 *Contemporary Urban Japan: A Sociology of Consumption*. Blackwell, Oxford, UK
Douglas M, Isherwood B 1982 *The World of Goods*. Norton, New York
Ewen S 1976 *Captains of Consciousness. Advertising and the Origins of the Consumer Culture*. McGraw-Hill, New York
Featherstone M 1982 The body in consumer culture. *Theory, Culture & Society* **1**(2)
Featherstone M 1991a The body in consumer culture. In: Featherstone M, Hepworth M, Turner B S (eds.) *The Body: Social Process and Cultural Theory*. Sage, London
Featherstone M 1991b *Consumer Culture and Postmodernism*. Sage, London
Featherstone M 1995 *Undoing Culture: Globalization, Postmodernism and Identity*. Sage, London
Galbraith J K 1961 *The Affluent Society*. Penguin, Harmondsworth, UK
Frank A Gunder 1998 *Re-ORIENT: Global Economy in the Asian Age*. University of California Press, Berkeley, CA
Hansen M 1991 *Babel and Babylon: Spectatorship in American Silent Film*. Harvard University Press, Cambridge, MA
Horkheimer M, Adorno T 1971 *Dialectic of Enlightenment*. Herder and Herder, New York
Huyssen A 1986 Mass culture as women: modernism's other. In: (eds.) *After the Great Divide*. Indiana University Press, Bloomington, IN
Jameson F 1984 Postmodernism, or the cultural logic of late capitalism. *New Left Review* **146**: 53–92
Jameson F 1990 Postmodernism, or the cultural logic of late capitalism. In: Jameson F (ed.) *Postmodernism*. Verso, London
McCracken G 1988 *Culture and Consumption*. Indiana University Press, Bloomington, IN

McKendrick N, Brewer J, Plumb J H 1982 *The Birth of a Consumer Society: The Commercialization of Eighteenth-century England*. Hutchinson, London
Marcuse H 1964 *One-dimensional Man*. Routledge, London
Myoshi M, Harutoonian H (eds.) 1989 *Postmodernism and Japan*. Duke University Press, Durham, NC
Nava M 1997 Women, the city and the department store. In: Falk P, Campbell C (eds.) *The Shopping Experience*. Sage, London
Packard V 1957 *The Hidden Persuaders*. D. McKay, New York
Pels D Crebas A 1987 Carmen: Or the invention of a new feminine myth. *Theory, Culture & Society* **4**
Pels D, Crebas A 1991 Carmen: Or the invention of a new feminine myth. In: Featherstone M, Hepworth M, Turner B S (eds.) *The Body*. Sage, London
Silverberg M 1991 The modern girl as militant. In: Bernstein G L (ed.) *Recreating Japanese Women, 1600–1945*. University of California Press, Berkeley, CA
Simmel G 1997 The metropolis and mental life. In: Frisby D, Featherstone M (eds.) *Simmel on Culture*. Sage, London
Williams R 1982 *Dream Worlds: Mass Consumption in Late Nineteenth Century France*. California University Press, Berkeley, CA
Wilson E 1998 Bohemian love. In: Featherstone M (ed.) *Love and Eroticism*. Sage, London

M. Featherstone

Consumer Economics

1. Introduction

Consumer economics is concerned with the allocation of the scarce means of the consumer over a variety of different commodities and services. These goods and services may be material as well as immaterial. What matters is that they are not available to the consumer for free. They carry a price. What also matters is that they are desirable, that is, that the consumer is willing to pay a price for them. As a rational being, the consumer will select that set of goods and services that pleases him or her most among the ones he or she can afford.

Two basic ingredients are involved in this selection: limited means and consumer preferences. The first one, limitation of the budget, expresses the fundamental scarcity that is the *raison d'être* of modern microeconomics. The second one, the ordering of the available alternatives according to preferences, reflects the role played by personal attitudes, biological needs, social aspirations, cultural trends and so on in determining how the scarce resources are being allocated. In the marketing approach to consumer decision making the emphasis is mostly on what shapes preferences. In the theory of consumption and saving, the focus is on the lifetime wealth constraint and to a lesser degree on the explanation of (time) preference. On the level of generality used here both ingredients will be discussed in an even-handed way.

Neither the budget nor the preferences are fixed or necessarily exogenous to the economic system. In the present context, however, the budget or total available means will be taken as given for the consumer, while also the preferences are not an object of choice for her or him.

Consumer theory is about individual behavior. Empirical data frequently refer to an aggregate of consumers, for example, a nation. Without some additional assumptions the theoretical results do not carry over to such an aggregate and the empirical relevance of the theory is minimal. The additional assumptions imply the existence of the *representative consumer*, who can be treated as if he or she were the individual consumer of the theory.

Empirical applications are abundant. Davenant (1699) is the earliest known example. It explains the price of corn as a function of the amount harvested. For a long time empirical work and theory were virtually unrelated. It was only just before World War II that the applications were first based on theory. Still, one usually focused attention on the demand for a single good. In more recent times attention has also been directed at the explanation of demand for all goods simultaneously, for which there is an explicit theory that offers the appropriate framework.

2. The Budget

The total means available, the *budget*, are used to pay for the quantities of the various goods and services, goods for short, that the consumer wants to acquire. It is assumed that the number of different goods is finite and fixed. The set of quantities is called a *bundle*. The budget is then the total amount to be paid for a bundle. Theory assumes the budget to be given for the consumer. Not quite correctly, 'income' is frequently used as a synonym for total means.

The *price* is the amount of money to be paid for a unit of a particular good. The prices are assumed to be positive, with negative and zero prices not admitted. Quantities are most of the time taken to be positive as well. The framework can be extended to also cover negative demands: 'bads,' or the supply of home-produced goods, for example. Here the focus is on positive demands. With positive prices and positive demands the budget must be positive.

Theory assumes that the goods are fully divisible. This is a realistic assumption in the case of fluids like gasoline, but not empirically valid for many types of household durables. Theoretically, one can replace the durables themselves by the services these render, which are continuously divisible, but empirically the stream of services is not directly observable. It has been

attempted to reformulate consumer theory in terms of integer quantities. This development has not entered the mainstream of consumer analysis.

Most of the time the quantities are taken to be the objects of choice and the prices given for the consumer, at least in the short run. There are cases where the quantities are fixed and the prices are endogenous. The Davenant study just mentioned is an example where the prices are being explained and the harvests of corn being fixed. This reversal of causality is specifically of relevance for nonstorable agricultural produce. If for all goods the quantities are the variables explained one speaks of *regular* demands. The case of endogenous prices is one of *inverse* demands. When for some goods the quantities are endogenous and for other goods the prices are endogenous, one speaks of *mixed* demands. Although inverse and mixed demands are easily handled by the theory, the mainstream is concerned with regular demands.

3. *Preferences*

According to the basic postulate of consumer theory, the consumer selects that bundle among the available alternatives that gives the most satisfaction. Assumptions are made that guarantee the existence of a complete, reflexive, and transitive *preference order* defined on the set of bundles. The requirement of completeness, that is, the ability to order all possible alternatives, is somewhat stronger than needed. It is enough if the consumer is able to order the bundles in the neighborhood of the set of bundles that fit in the budget. The assumptions of reflexivity and transitivity are of a logical nature.

The assumptions about the preference order do not exclude the possibility that bundles occupy the same rank in that order. All bundles with the same order form an *indifference class*. In the case of the choice between only two goods the graphical representation of the indifference class is the *indifference curve*. The set of all indifference classes is named the indifference map. Theory specifies a continuity postulate for this map. This amounts to the property that a drop in the quantity of one good of a bundle can always be compensated for by a change in the quantities of one or more other goods to stay within the same indifference class. This assumption rules out *lexicographic preferences*, where the preference for one good absolutely dominates the desire for other goods.

The preference order is equipped with two further properties. One is that of *desirability*. More of any good is preferred, keeping the quantities of the other goods the same. There is no satiation. Moreover, it excludes the existence of intersecting indifference classes. It also implies that the budget is always used fully. The other property is that of *convexity of preferences*: a linear convex combination of two bundles is preferred to the least preferred one of the two bundles themselves. This property is of special importance to guarantee a unique solution to the consumer's choice problem.

4. *Utility*

Under the assumptions made about the preference order there exists what is known as the *utility function*. It assigns a (real) value, utility, to bundles in such a way that for any pair of bundles the one that is preferred to the other obtains a higher value. In the case that the consumer is indifferent between the two bundles they obtain the same value.

Given the desirability property of preferences the utility function is *monotonously increasing* in the quantities of the goods. Convexity of preferences implies that the utility of a convex combination of two bundles obtains more utility than the one with the least utility. It is known as the property of *quasiconcavity*.

The name 'utility function' is slightly misleading. It is a preference indicator. Utility is not 'produced' by a bundle. The utility function associates with a bundle a certain value that reflects the rank of its indifference class. There is also no uniqueness of the index. Adding to all rank values the same value yields another utility function that corresponds with the same preference order as the original one. In more general terms: given a particular utility function for a preference order any order preserving or monotone increasing transformation of it will yield an equally valid utility function. As far as consumer demand is concerned only *ordinal* properties of a utility function matter. *Cardinal* properties that are not invariant under order preserving transformations have no effect on choice behavior.

In fact, consumer theory does not need a utility function at all. However, a utility function appears to be a convenient tool for describing the consumer's choice. No damage is done using a particular utility function, as long as it correctly reflects the preference order.

A related concept is that of *marginal utility*. Mathematically, it is simply the first derivative of the utility function with respect to the quantity of a certain good. It represents the increment in utility for an infinitesimal increase in that quantity. In view of the desirability of preferences it is always positive.

5. *Demands*

Taking the case of exogenous prices, the consumer selects that bundle among the bundles that can be acquired with the given budget that he or she prefers most. Under the assumption made for the preference order the selected bundle is unique. It can be written as a set of functions, one for each good, of the budget and of the set of prices. As the selected bundle fits in the budget also the demand functions will fit in it. This is

the *adding-up* property of demand. The composition of the bundles that satisfy the budget will remain the same if all prices and the budget change by the same positive factor. This then also holds for the set of demand functions. These are *homogeneous* of degree zero in the budget and the prices.

One of the implications of the assumptions made is that for two bundles in the same indifference class the amount demanded of a certain good is less if its price is higher. This is the *Law of Demand*. Note that it concerns bundles belonging to the same indifference class. It is the most important qualitative result of consumer theory. The more general version stating that demand for a good is a negative function of its price is not true with the same degree of theoretical validity. This law was put forward for the first time by Davenant (1699).

Observe that the adding-up, homogeneity, and negativity conditions have been formulated without the use of the concept of utility. For the derivation of further properties of demand functions the utility function proves to be convenient.

6. Consumer Equilibrium

With the utility function as a tool the consumer's choice problem can be described as maximizing utility with respect to the quantities of the various goods under the condition that the bundle chosen fits in the budget. The first-order condition for this constrained optimum is the proportionality of the respective marginal utilities with their prices. This result is known as the *Second Law of Gossen*. It characterizes the consumer's equilibrium. It implies that for all goods the marginal utilities divided by the prices are the same.

One interpretation of this outcome is that an extra unit of currency is allocated over the various goods in such a way that it yields the same increase in utility per unit of currency for all goods. The marginal utility of a unit of money spent on each good is the same.

The marginal utilities are cardinal concepts, but their ratio is ordinal. There is a tendency to work with this ratio that is known as the *marginal rate of substitution*. In equilibrium the marginal rate of substitution between two goods is then equal to the price ratio of these goods.

Proportionality of marginal utilities and prices is not only the equilibrium condition in the case that the quantities are endogenous and the prices exogenous, but also when the prices and quantities reverse roles in this respect. In fact, this condition also holds when for some goods the prices are exogenous and the quantities endogenous, while for other goods the opposite is true.

Under the usual assumptions made, the equilibrium conditions can be solved for a unique set of values for the endogenous variables given the values for the exogenous variables and the shape of the utility function. In the case of endogenous goods the solution can be expressed as a set of demand functions in terms of total means available and the prices. These demand functions have a number of interesting properties, which play a role in empirical analysis.

7. Properties of Demand Functions

Under rather general conditions the demand functions are differentiable with respect to income or total means and prices. The derivatives satisfy certain conditions. Frequently one expresses those conditions in terms of *elasticities*, defined as the proportional change in demand as a consequence of a (small) proportional change in means (*income elasticities*) or in prices (*price elasticities*), respectively. The elasticities can be easily derived from the derivatives of the demand functions. Note that neither derivatives nor elasticities are necessarily constants according to the theory. The choice of constant coefficients is a matter of empirical expedience. Income elasticities can be negative. In this case, an increase in income reduces demand for the good. This usually happens if there exists a similar product that is much more attractive. If the income elasticities are negative one speaks of *inferior* goods. Goods with positive elasticities are *superior* or *normal* goods. Income elasticities can be larger than one. Then as income increases, the budget share, being the share of expenditure on the good in question in total expenditure, will increase. A good with such a property is named a *luxury*. A *necessity* is a good with an income elasticity less than unity.

In Sect. 5 some properties were mentioned. The adding-up property reflects that the demands exhaust the budget. The *Engel* adding-up condition states that the sum of the income elasticities multiplied by the respective budget shares adds up to one. The *Cournot* adding-up condition refers to the price elasticities. The sum of the price elasticities for the prices of the various goods with respect to the price of one good, say good i, multiplied by the budget shares, equals minus the budget share of the ith good. These properties imply that if one knows all but one of, say, the income elasticities, one can reconstruct the remaining one as a residual.

The homogeneity condition entails that the sum of the income and price elasticities equals zero. The same proportional increase in income and all prices will have no effect on the quantity demanded. Thus, knowing all but one of those elasticities enables one to find the remaining one.

The price effects can be split up in a part that represents the move from one indifference class to the other and in a part that reflects substitutions within an indifference class. The first effect can be neutralized by an adequate change in means, such that one stays in the original indifference class. It is named the *income effect of price changes*. The other part of the effect of a price change is known as the *substitution effect of price changes* or as the *compensated effect of price changes*.

The income effect of price changes can be combined with the effect of an income change to form the effect of a *real income* change. This represents the move from one indifference class to the other.

The substitution effects can be collected to form a matrix, the *Slutsky* matrix. On its diagonal one finds the substitution effect on the demand for a good by its own price. On the basis of the assumptions made it should be negative, see Sect. 5. For superior goods the income effect of the own price change is also negative. Then the two effects work in the same, negative, direction. This cannot be generalized. If the good is an inferior commodity, the income effect will be positive and may, in absolute value, be larger than the substitution effect. The total own price effect is then positive. Such a good is known as a *Giffen good*. No empirical cases are on record, but they cannot be ruled out on theoretical grounds. For practical purposes the total own price effect may be taken to be negative.

A further property of the Slutsky matrix is its symmetry, first stated by Slutsky (1915). It is a mathematical property of great empirical consequence, but with hardly any intuitive appeal. The empirical importance can be illustrated for the case that there are 20 goods in the budget. The number of Slutsky cross-price effects is then 20*20–20 = 380. The property that these are pairwise equal reduces the empirical problem by 190 fewer coefficients to estimate separately.

The elements of the Slutsky matrix contain a considerable amount of information about the preferences of the consumer. Two instances of this will be taken up in the following sections.

8. *Complementarity, Substitution, and Independence*

Preferences are specifically of interest when they refer to interactions between goods. If two goods reinforce their desirability, one speaks of *complementary* goods. Examples are wine and cheese. Two goods are *substitutes* if one can replace the other to a certain degree; beef and pork, for instance. In the case of *independence* there is no specific relation between the two goods. How are these fundamental relations between goods reflected in empirical demand behavior?

The early marginalists started off from the marginal utility of a good. If it goes up if one has more of another good, the two are complements. If it goes down, the pair are substitutes. Unfortunately, this rather straightforward definition is not an ordinal one. Depending on the type of utility indicator corresponding to the same preference order, two goods may be complements rather than substitutes. The original meaning of the terms 'complements' and 'substitutes' clearly refers to the preference order and not to a particular representation of it.

It has been attempted to associate complementarity and substitution with the sign of the marginal rate of substitution, an ordinal concept. This definition is remote from intuition and the consequence for the signs of the compensated price elasticities are not easy to trace. Hicks (1939) cut that Gordian knot by using the sign of the elements of the Slutsky matrix. A negative sign means complementarity and a positive one substitution. However, there are problems with this definition. One of these is that positive off-diagonal elements of the Slutsky matrix tend to be larger or more numerous than negative elements, reflecting the tightness of the budget, which, as such, has nothing to do with preferences. The approach of Allais (1943) appears to meet most of the difficulties, but is not that simple to apply.

In the early days of empirical application of full-sized demand systems the assumption of complete independence was used frequently. Under this assumption there exists a utility function that is additive in (functions of) the various goods. The best known example is the Samuelson–Klein–Rubin utility function. It is a weighted sum of the logarithms of the quantities of the goods as far as they exceed some minimum value. The implied demand functions are correspondingly also more simple than in the case of real preference interaction between goods. The impact of price changes other than that of the own price takes the form of a change in a price index where the identity of the other price change does not matter. A consequence of complete independence is the limited number of coefficients to estimate. Given the paucity of data, typical for the early years of empirical demand analysis, a small number of unknown parameters was an attractive feature, even if the underlying assumption was not quite realistic.

9. *Separability of Preferences*

Terms like 'wine' and 'cheese' actually refer to conglomerates of different types of wine and different types of cheese. If the preferences for the various types of wine are independent of the amounts of the diverse cheeses consumed, one speaks of *separability* of preferences. This does not mean that the total amount spent on wines as a group does not depend on the total amount of cheese consumed. For the preference interaction between wines and cheeses the wine nature of the wines and the cheese nature of the cheeses dominate the relation and the specific brands do not matter. As a consequence the consumer's behavior can be modeled as a stepwise procedure. First, the consumer decides on the total amounts of wine and cheese he or she wants to consume. Next, given those amounts and the prices, he or she determines how much of each type of wine and of each type of cheese he or she will purchase. For the first level the consumer needs to know total available (real) income and all the prices. However, those prices can be represented by a price

index for each group. The consequence is that the impact of a change in the price of, say, Gouda cheese on the demand for, say, a high-class Burgundy wine is only by way of the effect of a change in the price index of cheese. The origin of the price change does not need to be identified.

It is a simple matter to extend this framework to more than two levels. For example, one can think of a hierarchy of the type: total consumption, food, meat, beef, with beef taken as the elementary good.

For the full set of demand relations separability means a rather simple constraint on the Slutsky coefficients for interactions between goods belonging to different groups.

Separability is very convenient for empirical work. One can estimate the first level model separately from the lower level models and in this way control the dimensionality of the estimation problem. In fact, the data usually are only available in already aggregated form. The assumption of separability justifies working with those aggregates as if they were elementary goods.

A special case of separability is *strong separability*. There the aggregates have no special type of interaction like that of substitution or complementarity discussed in the previous section. They are all independent vis-à-vis each other. The Slutsky matrix then becomes even more constrained than for the case of ordinary or weak separability.

Under separability the utility function can be written as a function of a set of subutility functions, one for each group. Under strong separability the preference order can be represented by a utility function that is additive in those subutility functions.

10. Indirect Utility and Expenditure Functions

The utility function is a function of the quantities of the various goods. One can replace these quantities by the *optimal* quantities, the demand functions. One obtains the *indirect utility function*, an expression in terms of total means and all prices. It indicates the maximum attainable utility for the given total means and these prices. Its derivatives are the marginal utility of total means and the marginal utilities of the various prices. The former shows how much utility can be increased if total means are slightly increased. The marginal utility of a price is the reduction in utility when the price goes up. These properties of the indirect utility function are based on those of the demand functions. Another property is its homogeneity of degree zero in means and prices. Otherwise said: an increase of means and prices by the same factor leaves utility unchanged.

The indirect utility function is a cardinal concept like the original utility function. The properties just mentioned, together with a curvature assumption, are ordinal, however. The *Identity of Roy* can be used to derive the demand functions from a well-specified indirect utility function. To obtain the demand function for good i one simply takes the negative of the ratio of the marginal utility of the price of good i and that of total means. It is a convenient way to arrive at a proper functionally specified set of demand functions. The alternative to deriving the demand functions as the optimal values for the direct utility function under the budget turns out to be practical only for very simple functional specifications of that function. The Roy identity can be seen as a convenient short cut.

A concept similar to the indirect utility function is the *expenditure function*, also known as *cost function*. It yields the minimum outlay or means to obtain for a given set of prices a given level of utility. It can be obtained from the indirect utility function by fixing the level of utility and solving for the means. It has, *mutatis mutandis*, similar properties to the indirect utility function. It reacts positively to both an increase in the level of utility to be reached and to price increases. It also has to satisfy a certain curvature condition. Corresponding to the identity of Roy there is *Shephard's lemma*. By differentiating the expenditure function with respect to the price of good i, keeping utility constant, one obtains the optimal demands as a function of the prices and the level of utility.

To distinguish these demand functions from the ones that are expressed in prices and *means*, they are named *Hicksian* demand functions. The former ones are then the *Marshallian* demand functions. One can shift easily from the Hicksian to the Marshallian demand functions by replacing the level of utility by the expenditure function. Likewise, one may change from the Marshallian demand functions to the Hicksian ones by substituting the total means variable by the expenditure function. For empirical purposes the Hicksian demand functions are not so useful because one would need a value for the utility level, which would be rather arbitrary. However, starting from a well-specified expenditure function one can derive the Hicksian demand functions in which one replaces the utility level by the indirect utility obtained from the expenditure function.

From a theoretical point of view, the Hicksian demand functions are of considerable interest. Their price derivatives represent the effect of price changes with utility kept constant. They are precisely the Slutsky coefficients introduced in Sect. 7. The Slutsky matrix turns out to be the matrix of second-order price derivatives of the expenditure function.

Another interesting feature of the expenditure function itself is its role in the derivation of the *constant utility price index*. Consider the case of a consumer who is in equilibrium. The expenditure function gives the minimal amount of expenditure to reach a given level of utility with the current prices. Take another set of prices but keep utility constant. The expenditure function will now show the minimum outlay required to reach the original utility level with the new prices.

The ratio of the latter and former amounts can be seen as a price index. It indicates the factor by which one has to multiply the total means to keep the same level of welfare. The factor could be more or less than one. In the case of a general price increase it will be larger than one. The practical use of the constant utility price index meets with the difficulty of specification of the expenditure function. However, its theoretical appeal cannot be questioned.

11. Empirical Validity

Do data on consumer behavior reflect the properties that theory postulates? In the case of *demand functions for a single good* the properties of interest are the negativity in the response to a change in the own price for the good in question and the homogeneity condition, implying no response to an equiproportional change in income (budget) and the prices. Usually negativity does not create serious problems and the demand equations in estimated form readily display this property.

Statistical tests tend to be less lenient for the homogeneity condition. This is somewhat puzzling because of the plausibility of this property, which is sometimes interpreted as the absence of a monetary veil. Homogeneity justifies the proposition that allocation of means is based on relative prices rather than on absolute ones, which is a central issue in microeconomics. Many reasons have been advanced to explain away this dilemma such as inappropriate data, specification errors (absence of explicit dynamics, for instance), and the use of incorrect test statistics. Correction of some of these shortcomings appears to have achieved some success.

In the case of *systems of demand equations* explaining demand for a set of goods simultaneously, negativity does also not create a serious validity problem, but homogeneity is less easily accepted. Additional properties of empirical interest here are Slutsky symmetry and separability. Symmetry appears to agree with the data. This may be more a matter of the wide confidence intervals of the estimates of the estimated Slutsky coefficients than of the truth of the symmetry hypothesis itself. Separability of preferences has such attractive properties for applied work that it is frequently used in spite of its demonstrated lack of empirical validity.

Also in consumer analysis, theory and application are being handled in a balanced way, reflecting doubts about both theory and facts and about the procedures to reconcile them.

Bibliography

Allais M 1943 Les données général de l'économie pure. Traité d'économie pure. Imprimerie Nationale, Paris

Barten A P, Böhm V 1982 Consumer theory. In: Arrow K J, Intriligator M D (eds.) *Handbook of Mathematical Economics*, II. North Holland, Amsterdam

Davenant C 1699 *An Essay Upon the Probable Methods of Making a People Gainers in the Ballance* (*sic*) *of Trade*. J. Klapton, London

Deaton A, Muellbauer J 1980 *Economics and Consumer Behavior*. Cambridge University Press, Cambridge, UK

Hicks J R 1939 *Value and Capital. An Enquiry into Some Fundamental Principles of Economic Theory*. Clarendon Press, Oxford, UK

Samuelson P A 1947 *Foundation of Economic Analysis*. Harvard University Press, Cambridge, MA

Slutsky E 1915 Sulla teoria del bilancio del consomatore. *Giornale degli economisti* **51**: 1–26

Theil H 1975 and 1976 *Theory and Measurement of Consumer Demand*. North-Holland, Amsterdam

A. P. Barten

Consumer Psychology

Consumer behavior refers to the acquisition, consumption, and disposal of time and offerings (goods, services, and ideas) by decision-making units (individuals, dyads, and larger groups). Consumer behavior is pervasive and universal involving, as it does, choices made during the waking moments of virtually all human beings in all societies and cultures. In addition to overt observable behaviors (e.g., making a purchase, wearing a shirt, engaging in word-of-mouth communications regarding products), consumer behavior also includes the covert psychological states and processes (e.g., motives, learning, attitudes, decision making, past experience) that underlie and influence the overt behaviors. Last, while consumer behavior involves trying, rebuying, and increasing some behaviors, it also includes the cessation or lessening of some of those same behaviors (e.g., smoking, unsafe sex, eating fatty foods).

Though also referring to the covert psychological processes, when used to describe a scholarly endeavor, consumer psychology refers to using distinctively psychological concepts and methods to understand consumer behavior. These days, scholars who employ clearly psychological concepts and methods to study consumer behavior are more likely to do so from venues other than departments of psychology.

1. The Complexity of Consumer Behavior

Consumer behavior includes often-overlooked layers of complexity. Offerings may be acquired in ways other than via purchasing, including via renting, borrowing, trading, finding, stealing, as a gift, and so forth. Similarly, consumption—which occurs over time periods of varying duration (from the almost instantaneous munching and swallowing of a potato

chip, through the wearing of a shirt for several years, to buying and living in a home for decades)—generally involves distinct facets, including using, maintaining (e.g., cleaning a suit), storing, and retrieving, all or some of which may be repeated over time. Once consumption ends, one may be faced with disposing of the item (or its container). One can decide to throw it away, convert it to serve some other purpose, trade or exchange it for something else, sell it, give it away, and so forth. Consumer behavior also reflects cyclical and contingent qualities. Having purchased, consumed, and disposed of a tube of toothpaste, we acquire a replacement and repeat the cycle. Car insurance and gasoline are generally acquired contingent upon having acquired a car.

Time is finite; how we choose to spend it has tremendous implications for virtually all other aspects of our lives. In some respects, one can actually buy and trade time, e.g., paying someone to mow our lawn and free our time for something else.

Offerings come in many forms and are not limited to products and services provided by commercial or for-profit entities. Offerings also include political candidates, television programs, newspaper articles, leisure time activities, different articles within the same newspaper, and so forth.

Engaging in consumer behavior involves several roles, including behaving as the decision maker, purchaser, and/or user. Often, these roles are confined to a single individual, as when one person chooses, purchaser, and then consumes a soft drink. Alternatively, they may involve two or more individuals (e.g., wife decides, husband purchases, child uses).

Last, consumer behaviors occur not in a vacuum, but in the context of everyday life where they tend to be influenced by numerous external factors, including cultural, social, political, environmental (e.g., not shopping during a snow storm) and physical (e.g., not wanting to stoop for a product on a low shelf) considerations.

2. *Why Study Consumer Behavior?*

As understanding consumer behavior is often the key to business success, many who study consumer behavior come from disciplines (e.g., economics, marketing, advertising, retailing) that support business activities. But why do psychologists find it worthy of study? Discussed in greater detail elsewhere (Jacoby 1975, pp. 981–5), a few of the reasons are as follows.

Consumer behavior is a domain providing innumerable opportunities for important, socially meaningful research. Learning how to influence the underprivileged in the USA and elsewhere to adopt and increase consumption of products and services good for them and to cease consumer behaviors that are less appropriate or even harmful is one example. Studying why there is so high a rate of noncompliance with physician prescription dosage instructions and studying how to reduce this rate is another.

Attempting to understand consumer behavior forces one to consider a broader array of factors than is typical in much psychological research. This occasionally leads to seeing relationships that might not have been noted had one not been compelled to approach the subject using a broader (or at least different) perspective. Consumer theory and research thus have the potential to contribute to the development and extension of psychology itself. As an example of a conceptual contribution, consider a richer, more dynamic way of conceptualizing stimulus–organism–response (Jacoby 2002). As an example of an empirical contribution, consider the development of process methods that enable psychologists to better test dynamic theories regarding how information acquisition leads to the development of, and molecular changes in, higher order cognitive processes (Jacoby et al. 2002).

Consumer research provides numerous opportunities for working with dependent variables that, from the perspective of our test subjects, are meaningful and consequential. Further, because of the universal and pervasive nature of consumer behavior, this arena provides excellent opportunities for examining the validity and limits of psychological theory.

3. *Evolution of the Scholarly Approach to Studying Consumer Behavior*

The origins of scholarly work on consumer behavior may be attributed to the economists of the eighteenth and nineteenth centuries (e.g., Adam Smith, Jeremy Bentham, Alfred Marshall). These economists held that a person's desire for goods and services generally exceeded the person's ability to pay. Therefore, buying decisions were theorized to result from a process during which a value was assigned to each desired offering based upon the person's 'rational' assessment of that offering's ability to satisfy his or her needs and desires. This want satisfying ability was termed 'utility' and it was held that consumers would seek to maximize utility. According to these economists, the price of the offering was the key element in the maximizing utility equation. Raising an offering's price was predicted to lower its sales; lowering its price would increase sales. Though seemingly commonsensical, this proposition suggests that, given changes in price, consumers respond like billiard balls, buying when prices are lowered and refraining when prices rise. Little (perhaps nothing) else matters.

In terms of the stimulus–organism–response (S–O–R) model familiar to psychologists, it can be seen that the 'economic man' conception is severely limited, suggesting as it does a relatively empty organism whose behavior is determined by economic stimuli and little else. According to the classic S–O–R

formulation, external stimuli (e.g., price changes) first need to be perceived, understood, evaluated, and integrated with other psychological phenomena (e.g., past experience, current needs, attitudes, and expectations, etc.) before having impact, if any, on the individual's behavior. Thus, while one can see various reasons why a price decrease would lead to increased purchase (as predicted by economic theory), one can also see various reasons why it would have either no effect (e.g., the price change does not represent a 'just noticeable difference'; the consumer would purchase the same amount of the items regardless), or the reverse effect (e.g., sales decrease because the consumer interprets the price decrease as reflecting a decrease in quality, or because the consumer views the item as no longer having the desired snob appeal, etc.). To best understand and predict consumer behavior, we would have to know something about the consumer's motives, *a priori* expectations, how the incoming information is perceived and interpreted, and so forth—namely, the factors traditionally studied by psychologists and then some. (Parenthetically, when this is done, it is often found that other factors are more important than price. Examination of contemporary consumer behavior texts reveals that most devote fewer than a half dozen or so pages to price considerations.)

Responding to limitations in the 'economic man' model, both theorists and practitioners began adding variables to the stimulus and especially organism portions of the S–O–R model. During the 1930s through the 1960s, a number of European émigrés trained in Freudian psychoanalysis, collectively known as 'motivational researchers' (e.g., Dichter 1960), became influential among the business community. Other psychological constructs (including attitude and personality) also were introduced during this period. By the mid-1960s, the emerging models (developed not by psychologists, but by professors of marketing) reflected primarily psychological and/or sociological factors (e.g., Andreasen 1965, Engel et al. 1968, Howard and Sheth 1969). Engel et al.'s, book now in its seventh printing, was the most influential introductory text for perhaps 15 years and remains popular. The 1960s also represented the time when an increasing number of psychologists with a scholarly bent began studying consumer behavior. Reflecting evolution, a number of today's leading texts (Hoyer and Maclnnis 1997, Peter and Olson 1999, Solomon 1999) are authored or co-authored by PhD psychologists (Hoyer, Olson, Solomon).

Around 1970, two developments did much to shape the emerging field. First came the formation of the Association for Consumer Research (ACR), today considered to be the principal scholarly organization in the field. Starting with a few hundred members in 1969–70, ACR's membership consists of researchers from a wide variety of disciplines and now numbers nearly 2,000, including several hundred psychologists. Though its members focus on different issues (e.g., marketing researchers tend to be more interested in understanding purchase decisions, communication researchers more interested in interpersonal or mass media communication, home and family economists more interested in postpurchase issues, etc.), their shared frame of reference is that, influenced by elements in a complex environment, consumers are information-processing, problem solving, decision-making, choice-making beings. Its annual series, *Advances in Consumer Research*, best illustrates the diversity of the field.

The second major development was the introduction of the *Journal of Consumer Research*, the first journal devoted specifically to consumer research. Prior to the early 1970s, scholarly work on consumer behavior appeared sporadically in journals spanning many disciplines. Work having a behavioral science (including psychological) perspective tended to appear in the *Journal of Applied Psychology* and, to a lesser extent, in the *Journal of Marketing* and *Journal of Marketing Research*. While these journals continue to publish some of the better consumer research, their principal focuses lie elsewhere. A collaborative venture by a number of national societies, the first issue of the *Journal of Consumer Research* appeared in 1974 and quickly became the field's flagship journal. Using the first 15 years of JCR publications, Hoffman and Holbrook (1993) provide an interesting perspective on the intellectual structure of the field. A 1991 *JCR* article by Cote et al. (1991) employ a citation analysis to assess the influence of the journal across the behavioral sciences. As a field of inquiry, consumer research is now interdisciplinary and intersector, being conducted by researchers operating from within academia, government, industry, and public not-for-profit groups.

4. *The Introduction of Psychology to Studying Consumer Behavior*

Interest by psychologists in consumer behavior and the factors that influence such behavior extends one hundred years. In 1900, H. Gale published work on the position of an advertisement on the printed page in 'The University of Minnesota Studies on Psychology.' Walter Dill Scott (whose dissertation on attention and persuasion was with Wundt), introduced the concepts of needs and motives in his 1903 book, *The Theory of Advertising*. Scott's 1908 book, *The Psychology of Advertising in Theory and Practice*, emphasized psychology by placing the word in its title. However, except for occasional articles in the *Journal of Applied Psychology*, psychologists devoted little scholarly attention to consumer behavior during the first half of the twentieth century. When studied at all, consumer behavior was often viewed as convenient for testing more fundamental psychological propositions. For

example, when the USA was experiencing shortages in traditional cuts of beef during World War II, Kurt Lewin's classic studies on public versus private commitment assisted both government and industry in making internal meats (e.g., brain, liver, kidney) more palatable to the US consumer. With the exception of Katona (1951, 1960), who was essentially a voice in the wilderness, few psychologists studied consumer behavior for its intrinsic interest. Situated in industry, most members of APA's Division of Consumer Psychology had an applied orientation and were interested in research that assisted business.

This emphasis began changing in the mid-1960s. The first chapter on consumer psychology appeared in *Annual Review of Psychology* (Twedt 1965). The second appeared (Perloff 1968) three years later, as did a chapter in the *International Encyclopedia of the Social Sciences* (Krugman 1968). In addition to the traditional focus on the consumer *qua* purchaser (i.e., as the object of business interests), these contributors focused greater attention on the consumer *qua* consumer. This emphasis became entrenched during the decade that followed (Jacoby 1975, 1976), gaining even greater strength in the intervening quarter century (Jacoby et al. 1998). By the early 1980s, the dominant focus of APA's Division of Consumer Psychology (which had been renamed The Society for Consumer Psychology) had shifted to studying the consumer *qua* consumer. In 1992, it began publishing the *Journal of Consumer Psychology*.

Concurrent with increasing numbers of psychologists finding that consumer behavior merited their scholarly attention, scholars in other disciplines eagerly embraced psychological theory and methods for their own consumer behavior research, with many advising their doctoral students to minor in psychology (particularly social, cognitive, or measurement psychology). Cross-fertilization was enhanced as increasing numbers of newly minted PhDs in social, cognitive, and experimental psychology assumed faculty positions in departments of marketing and other disciplines where they studied consumer behavior. A number subsequently became highly regarded scholars and leaders in the field. Adding to this trend, prominent psychologists began participating in Society for Consumer Psychology and Association for Consumer Research conferences, occasionally publishing articles in consumer behavior journals; examples include William McGuire (1976), Robert Zajonc (Zajonc and Markus 1982), and Daniel Kahneman (Ratner et al. 1999).

5. *Current Content and Focii of Consumer Psychological Research*

A psychologist reading today's introductory texts would feel right at home. Even those not (co-) authored by psychologists most often devote chapters to familiar subjects, including motives, attention, perception, categorization and comprehending, learning, memory, attitudes, persuasion, and decision-making. Large portions of other chapters are devoted to personality, values, cognitive dissonance, the self, affect, and symbolic considerations. Much of the remaining content tends to be sociocultural in nature, with chapters typically covering cultural and cross-cultural influences, subcultures, social class, reference and other group (e.g., age, gender, race, religion) influences, and family and household considerations.

Though consumer researchers study a wide variety of psychological phenomena and topics, some have dominated. Attitude theory has enjoyed sustained popularity for more than three decades. During the early 1970s, hardly an issue of the *Journal of Consumer Research*, *Journal of Marketing*, or *Journal of Marketing Research* appeared without at least one (and often more than one) article making heavy use of Fishbein's seminal work (cf. Fishbein and Ajzen 1975). Since the late 1980s, attention has shifted to more contemporary attitude theories (Chaiken 1980).

In the early 1970s, research on various aspects of cognitive processes (including judgment and decision-making) surfaced, quickly providing what currently remains the dominant intellectual focus in the field. Summaries of this work, as well as work on other topics, may be found in *The Handbook of Consumer Behavior* (Robertson and Kassarjian 1991). More recently, mood, affect, and the nature of the consumption experience also have captured substantial researcher attention. It is expected that all these emphases will continue to remain strong.

Another thrust has been increasing attention devoted to how consumer research may be used to develop, evaluate, and influence public policy affecting consumer well-being. While research of this sort sometimes appears in the journals noted above, much is now published in the *Journal of Public Policy and Marketing*, a journal established for this purpose in around 1980.

The field has experienced shifts in research methodology. Early nonpsychological research placed great reliance on surveys. As psychological factors assumed greater prominence in the field's models, experimentation became and remains the dominant methodological approach. Since the 1990s there has been a gradual acceptance of qualitative research methods of various sorts—ethnographic, interpretivistic, introspective, etc.). Approximately 15 to 20 percent of the currently published work utilizes such methods, and this trend is expected to continue.

6. *Emerging Focii in Consumer Research*

In addition to the emphases mentioned above, other anticipated trends include the following. First, though we do not expect many to forsake reliance on static

verbal report methods, it is expected that consumer researchers will gradually display greater use of qualitative and quantitative methods that capture the dynamic nature of the behavior and underlying psychological processes. Second, while some work has always focused on socially important issues and improving social welfare, we expect to see an increase in such research. Third, we expect consumer researchers to be drawn into applying their expertise in a number of other realms. Such a trend has been emerging in the legal arena, where consumer psychologists are called on to study phenomena such as deceptive advertising, trademark confusion, and product misuse/liability. The Internet and e-business represent another realm where considerable consumer research can safely be predicted.

Last, as noted, the Association for Consumer Research is widely acknowledged to be the central organization for consumer researchers of any stripe. Unlike the Society for Consumer Psychology, which subscribes to the American Psychological Association's Code of Ethics, the governance of ACR has resisted efforts to develop and implement any such code. This is regrettable, especially given the tremendous opportunities that consumer researchers, as researchers and consultants, have for affecting all consumers. Accordingly, one would hope and expect ACR's neglect to change, preferably sooner rather than later.

See also: Consumer Culture; Consumption, History of; Consumption, Sociology of; Decision Making, Psychology of; Decision Research: Behavioral; Family and Consumer Sciences: United States; Heuristics for Decision and Choice; Leisure and Cultural Consumption; Psychology and Economics; Psychology and Marketing; Social Psychology

Bibliography

Andreasen A R 1965 Attitudes and consumer behavior: A decision model. In: Preston L (ed.) *New Research in Marketing*. Institute of Business and Economic Research, University of California, Berkley, CA, pp. 1–16

Chaiken S 1980 Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology* **39**: 752–66

Cote A, Meng Leong S, Cote J 1991 Assessing the influence of *Journal of Consumer Research*: A citation analysis. *Journal of Consumer Research* **18**(3): 402–10

Dichter E 1960 *The Strategy of Desire*. Doubleday, New York

Engel J F, Kollat D T, Blackwell R D 1968 *Consumer Behavior*. Holt, Rinehart, Winston, New York

Fishbein M, Ajzen I 1975 *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA

Hoffman D L, Holbrook M B 1993 The intellectual structure of consumer research. *Journal of Consumer Research* **19**: 505–17

Howard J A, Sheth J N 1969 *The Theory of Buyer Behavior*. Wiley, New York

Hoyer W D, Maclnnis D J 1997 *Consumer Behavior*. Houghton-Mifflin, Boston, MA

Jacoby J 1975 Consumer psychology as a psychological sphere of action. *American Psychologist* **30**(10): 977–87

Jacoby J 1976 Consumer psychology: An octennium. *Annual Review of Psychology* **27**: 331–58

Jacoby J (2002) Stimulus–organism–response reconsidered: An evolutionary step in modeling (consumer) behavior. *Journal of Consumer Psychology* **12**

Jacoby J, Johar G V, Morrin M 1998 Consumer psychology: A quadrennium. *Annual Review of Psychology* **49**: 319–44

Jacoby J, Morrin M, Jaccard J, Gurhan Z, Maheswaran D (2002) Attitude formation as a function of incremental information input: A procedure for studying on-line processing models of attitude formation. *Journal of Consumer Psychology* **12**(1)

Katona G 1951 *Psychological Analysis of Economic Behavior*. McGraw-Hill, New York

Katona G 1960 *The Powerful Consumer*. McGraw-Hill, New York

Krugman H E 1968 Consumer behavior. In: *International Encyclopedia of the Social Sciences*. Macmillan, New York, pp. 349–54

McGuire W J 1976 Some internal psychological factors influencing consumer choice. *Journal of Consumer Research* **2**(4): 320–39

Perloff R 1968 Consumer analysis. *Annual Review of Psychology* **19**: 437–66

Peter J P, Olson J C 1999 *Consumer Behavior and Marketing Strategy*, 5th edn. Richard D. Irwin, Chicago

Ratner R K, Kahn B E, Kahneman D 1999 Choosing less preferred experiences for the sake of variety. *Journal of Consumer Research* **26**(1): 1–15

Robertson T S, Kassarjian H H 1991 *Handbook of Consumer Behavior*. Prentice-Hall, Englewood Cliffs, NJ

Solomon, M R 1999 *Consumer Behavior*, 4th edn. Prentice-Hall, Upper Saddle River, NJ

Twedt D W 1965 Consumer psychology. *Annual Review of Psychology* **16**: 265–94

Zajonc R B, Markus H 1982 Affective and cognitive factors in preferences. *Journal of Consumer Research* **9**: 123–31

J. Jacoby

Consumption, Environmentally Significant

1. Introduction

The nineteenth and twentieth centuries have seen an unprecedented surge in the number of people on the planet and a corresponding boom in the construction of buildings, roads, machines, and other artifacts. Both outcomes relied on technologies that mastered nature by controlling disease and facilitating access to material resources. This intensification of human activities has left its mark on the air, soil, water, and all

living systems. Some effects are minor or at least reversible, but others result in dramatic, long-lasting changes that may be demonstrably undesirable. There are widespread concerns today that, on balance, human activities are degrading the environment substantially.

Two familiar categories of activities are at the root of the problem: production and consumption. In the modern world, more and more of people's time and energy is spent producing goods and services to earn a living and then using this income to purchase an unprecedented quantity and variety of items for consumption. These are the principal activities involving human use of fuels, materials, water, and soil and therefore the basis for environmental disruption. For this reason, production and consumption are the obvious targets for changes in behavior.

Production and consumption cannot be called to a halt, but they can be carried out differently. Which changes, and how much change, are desirable are not questions that can be answered in a definitive way. Bold ideas are needed to break the cycle of more people, more production, more damage. The social and natural sciences can provide guidance for trying out these ideas in what is essentially a social experiment on a global scale. Contemporary information technologies make possible an unprecedented quality of public access to information and participation in debate.

Many efforts to reduce systematically harm to the environment focus on creating technological options for production activities (see *Industrial Ecology* and *Industrial Metabolism*). Another active area of investigation produces policy advice for public and private-sector executives. The focus of attention on consumption is relatively recent. It holds the promise of technological and policy opportunities, such as more fuel-efficient cars or tax breaks for better insulated homes. However, in addition, it also opens up the new prospect, which is far less developed, of systematically exploring lifestyle options for the direct benefit of household decision-makers. Thus the following questions can be posed. Which environmental changes are most problematic? How is consumption related to environmental change? Can people change their lifestyles and consumption patterns? The final section describes how social scientists can enhance the prospects for changes in consumption.

2. *Which Environmental Changes are Most Problematic?*

The environment consists of air, water, soil, minerals, microorganisms, animals, and plants—all bound by the great cycles and overlapping systems of which they are part. Changes to the environment, whether due to human activities or not, exhibit an unlimited variety. For example, 'water pollution' may involve contamination by a variety of wastes, each with different properties (see *Water Pollution: Marine*; *Water Pollution: Terrestrial and Groundwater*). The concentration of a particular waste material (as measured in g m^{-3} of water) depends on the amount discharged and the condition of the water—for example, whether it is still or moving. Another variable is whether other foreign materials are also present in the water and in what quantities. Polluted water could be confined to a remote location, or it could be heavily utilized for industrial purposes, irrigation, or even household use. Furthermore, different individuals will react differently even upon similar exposure to the same pathogen or toxic material.

Thus a specific act (like discharging a volume of chlorine from a paper mill into a river) will under different circumstances be associated with a variety of changes to the environment, and these changes may have either benign or ill consequences. The interactions of several different activities are common and compound the unpredictability. We would like to rank-order environmental problems according to a single, unambiguous criterion and resolve them one after the next starting with the most serious, but this is not a realistic expectation (see *Human–Environment Relationship: Indicators*). Consider the set of problems that are candidates for the most threatening to life on earth, such as climate change, the contamination of fresh water, or the ubiquity of toxic materials. Like the broom of the sorcerer's apprentice, each one subdivides into a large number of component problems, most associated with the confluence of multiple activities.

The only approach to preventing environmental damage that can hope to match the enormity of the challenge is the wholesale change of habits so as to eliminate many problems in one fell swoop. This would require a limited number of powerful solution concepts, each of which is capable of governing innumerable individual changes. Recent thinking about how to reduce the environmental impact of production activities provides examples of such concepts. Industrial Ecology promotes the transformation of industrial systems by copying a few basic principles that are characteristic of ecological systems. One is to close production loops, reclaiming wastes of one process for productive reuse in others, instead of discarding them as useless. Design for the Environment includes environmental criteria in the design process for industrial products, alongside the conventional criteria of performance, cost, and esthetics. Finally, Life-Cycle Analysis describes the environmental impact of a product from the extraction of materials for its fabrication to its disposal. Based on such studies, many analysts have concluded that producers should take responsibility for products over their full life cycle. Such solution concepts are working their way into the language and customs of corporate

cultures. Some firms make changes only when they are cost saving in the short-term or required by law. With the passage of time, however, and for a variety of reasons, others are seizing the initiative to reduce the environmental impact of their production activities. In this article a parallel path toward greater citizen initiative is described.

3. *How is Consumption Related to Environmental Changes?*

There are systematic differences in the environmental pressures associated with lifestyles of poverty and of affluence. In this article the concern is mainly with the consumption patterns typical of the industrialized countries because they serve as models for aspirations and emulation around the globe.

Purchased goods and services satisfy not only basic needs but also the desire for mobility, novelty, physical comfort, and luxury. The use of consumer items entails environmental degradation beyond that associated with production. The automobile provides the most important illustration. It is responsible for noise, traffic congestion, urban sprawl, and disposal problems in addition to the emission of a variety of pollutants that enter the atmosphere and get swept into soils and water supplies. Food consumption likewise presents major challenges quite distinct from the formidable ones associated with agriculture. What people choose to eat has major implications for land use, and packaging for foods (as well as for other household items) ends up after a short service life in landfills. The disposal of half-used containers of household chemicals or pharmaceuticals is also a serious problem. In recent years there have been systematic efforts to quantify the different ways households use energy, materials, and water and the various kinds of garbage that people discard (see, for example, Dillman et al. 1983, Schipper 1996, Wernick 1996, Rathje et al. 1992).

One of the most striking differences between production and consumption activities is the enormous amount of time, effort, and money that have gone into developing technological alternatives in the production environment and studying the implications of choosing different options. One reason for this amount of scrutiny is that a new plant may cost hundreds of millions of dollars and serve for many decades. Households, by contrast, are offered mainly commonsense advice in popular books or by interested parties (for example, through mailers that accompany utility bills). There are important opportunities, which should be exploited, such as better insulation and more energy-efficient appliances. Perhaps the best example is the real prospect for substantially more fuel-efficient cars. However, increasing numbers of researchers are convinced on the basis of their analyses (e.g., Duchin and Lange 1994) that these technological changes cannot offset (and through lowering costs may even stimulate) the growth in demand.

More fundamental changes in consumer behavior, especially regarding mobility (linked to cars and, increasingly, air travel) and material goods, are possible, but they are difficult to achieve for two reasons. First, consumer behavior reflects underlying lifestyle decisions that are intimately tied up with the sense of personal identity. Individual consumer purchases are part of a pattern, and large changes cannot be made one element at a time. For example, if a household decided to give up its cars, it might be necessary to change jobs or to move from a house to an apartment near public transportation or nearer to work. Second is that lifestyle changes are strongly constrained externally, notably by policy decisions like zoning and tax legislation or investment in highways rather than public transportation infrastructure. Some countries have experimented with nationwide incentives for eliminating unnecessary packaging and reusing or recycling the rest (Germany; see Fishbein 1994) or with extensive and convenient public transportation systems intended to displace car travel (The Netherlands). These are efforts of historic significance, but their success is inconclusive. To realize the potential of such initiatives, people need to examine their own households' lifestyle decisions, not just their consumption choices, in a context of well-articulated alternatives, in the same way that firms have begun to actively consider organizational and technological alternatives aimed at reducing their environmental impact. If households come to favor new alternatives, their choices can be expected to influence legislation and public investment priorities.

4. *Can People Change their Lifestyles and Consumption Patterns?*

Starting in childhood, people experience relationships with family, friends, neighbors, bullies, shopkeepers, thieves, police, librarians, and teachers. They develop values and opinions, acquire habits, hobbies, and skills. Later they may engage in paid work, all sorts of unpaid work, and political and community activity, choose where they will live and with whom, establish a household, and possibly have children of their own. They may frequent fast food restaurants, forego meat, prefer bowling or opera, drive tens of thousands of miles a year, etc. All of their attitudes and activities, constrained as they are by such things as legislation and public infrastructure, jointly comprise their lifestyle. The pattern of consumption, with its direct implications for the environment, is one expression of lifestyle choices. If lifestyle does not change, changes in consumption, in particular as they affect the environment, are likely to be slow and incremental. Many people today suspect that changes in lifestyle could bring about decisively better outcomes.

All people have the same basic material needs for food, clothing, shelter, and fuel and are daily exposed to common cultural messages. Even in a modern society that strives for individual self-realization and is relatively tolerant of unconventional living arrangements, only a limited number of distinctive lifestyle patterns emerge. We can define a household's lifestyle in terms of how it carries out the activities that satisfy its members' needs and desires. A taxonomy for describing these activities is proposed by Duchin (1998):

(a) Food
(b) Care of old and young
(c) Healthcare
(d) Personal care
(e) Education
(f) Household administration
(g) Clothing (garments and cleaning)
(h) Recreation and entertainment
(i) Vacations
(j) Housing (systems and maintenance)
(k) Household appliances and furnishings
(l) Household cleaning
(m) Transportation
(n) Home work for pay
(o) Home work for community
(p) Work outside home
(q) Idleness.

Market researchers in the US have provided a uniquely holistic view of American household lifestyles, laying the groundwork for a systematic approach to the questions that interest us. Based on the statistical analysis of a large database compiled from diverse sources, they concluded that the lifestyles of the approximately 100 million American households could be described in terms of only a few dozen categories. They gave these categories colorful names such as Furs and Station Wagons (described as 'new money in metropolitan bedroom suburbs'), Gray Power ('upper middle-class retirement communities'), Single City Blues ('downscale, urban singles districts'), and Public Assistance ('America's inner-city ghettos') (Weiss 2000). Today, consumer-marketing strategies use these categories routinely to target client households that are likely to purchase particular products. Households with similar consumption patterns tend to have similar impact on the environment.

Households with distinct lifestyles have different prospects for changing their lifestyle scripts. Individuals move from one category of household to another at turning points in the life cycle, for example upon graduation from high-school or university or at retirement. The challenge is for people to consider a new set of lifestyle scripts, especially at these critical junctures when change is more easily achieved.

It is an open question as to whether people will be engaged by this challenge rather than mainly responding to initiatives taken by corporate executives and politicians. There is a wide range of views about the extent to which individuals act in their personal self-interest or take a wider set of concerns into account (for an overview see Friedman 1996). People behave differently with their loved ones and with strangers and are generally less selfish in family settings than in business decisions. They often trade off immediate gratification for longer-term considerations and are willing to make sacrifices for the benefit of others—not only intimates or acquaintances but anonymous others or society at large. In fact, some people will go out of their way to save a river or risk their lives to prevent an accident. Can a critical mass of people be expected to travel less or eat less meat voluntarily in pursuit of a better quality of life, or for the good of the planet?

The standard economic proposition is that self-interested behavior is 'rational' behavior and that people act 'altruistically' only in exceptional circumstances. According to this view, the only way to influence firms or individuals to act differently—for example, to use less energy or generate less waste—is by the coercion of regulations or, better still, by affecting the money bottom line through subsidies and other positive incentives, or through charges and taxes. But is it true that this is the most effective way to prod households, like profit-making firms, to act in the public good?

Psychologists, anthropologists, and sociologists view the personal decision process to be a complex balance among considerations that frequently overrule individual self-interest. They provide reason to believe that appeals to the full set of motivations and behaviors—accompanied by an analysis of bold options—can encourage lifestyle decisions that reduce pressures on the environment. While different cultural settings call forth different patterns of motivation and behavior, Fiske argues persuasively that they can all be covered by a four-category taxonomy that includes communal sharing, authority ranking, and equality matching, as well as the kinds of behavior associated with what he calls market pricing (Fiske 1991).

When there are not powerful motivations to act otherwise, people often try to do the best they can for themselves, and 'rational choice' is a compelling explanation for many everyday decisions. But large numbers of people are passionate in their concern about the environmental dangers they perceive and ready to take initiative to transform the situation. This is what it means to say that there is an environmental movement (Stern et al. 1999). Nonetheless, consumption behavior has changed relatively little compared to attitudes about the environment. Arguably the most important obstacle is the difficulty of imagining new scripts and removing the obstacles to actually living them. What is needed are a detailed and sustained dialogue about lifestyle options, demonstration projects designed and implemented with the collaboration of a variety of experts, and the celebration of pioneers in all walks of life. How can such an unprecedented dialogue and experiment be initiated?

5. What Can Social Scientists Contribute?

This article has argued that environmental problems are too numerous and varied to yield to policies designed to deal with them one by one. Taxes and subsidies act only indirectly and slowly. A direct focus on a few powerful solution concepts can, in principle, resolve many problems at once. A great deal of research by engineers and applied physical scientists is already aimed at radical technological innovations in production (and consumption) and demonstration projects to illustrate their feasibility. Now it is time to turn to lifestyles and consumption.

Social scientists are interested in how different kinds of households carry out the full range of everyday activities. They can provide a coherent and systematic description of types of households and household lifestyles as a basis for understanding how we now live. The development of this structure makes it possible to use quantitative methods to analyze the implications of adopting new lifestyle scripts, ones that capture the public imagination. The objective of the analysis is to see how much difference it would make if particular new scripts were adopted. Feedback from analysis can sustain a more realistic public dialogue than one based only on anecdote and advocacy.

This is the age not only of the environment but also of the Internet, with its capability for sharing information and networking of both physically proximate communities and virtual ones with members around the globe. This medium might provide a powerful way to launch and maintain relevant public dialogue about individual and community decisions. If facts, figures, demonstration projects, and analysis inform this dialog, the combined effort of experts and citizens could actually make a difference. In fact, universities—and social science scholars in particular—might profitably reinvent their disciplines in the process.

See also: Consumer Culture; Consumer Economics; Consumer Psychology; Consumption, Sociology of; Desertification; Environmental and Resource Management; Environmental Policy; Environmental Psychology: Overview; Human–Environment Relationship: Carrying Capacity; Leisure and Cultural Consumption; Leisure, Psychology of; Sustainable Development

Bibliography

Consumption and Environment 1999 *Journal of Ecological Economics* **28**(3): 363–466

Dietz T, Stern P C, Guagnano G A 1998 Social structural and social psychological bases of environmental concern. *Environment and Behavior* **30**(4): 450–71

Dillman D, Rosa E, Dillman J J 1983 Lifestyle and home energy conservation in the U.S. *Journal of Economic Psychology* **3**: 299–315

Duchin F 1998 *Structural Economics: Measuring Change in Technology, Lifestyles, and the Environment*. Island Press, Washington, DC

Duchin F, Lange G 1994 *The Future of the Environment*. Oxford University Press, New York

Fishbein B K 1994 *Germany, Garbage and the Green Dot: Challenging the Throwaway Society*. Inform, New York

Fiske A P 1991 *Structures of Social Life: The Four Elementary Forms of Human Relations*. Free Press, New York

Friedman J (ed.) 1996 *The Rational Choice Controversy: Economic Models of Politics Reconsidered*. Yale University Press, New Haven, CT

Goodwin N R, Ackerman F, Kiron D (eds.) 1997 *The Consumer Society*. Island Press, Washington, DC

Rathje W L, Hughes W W, Wilson D C, Tani M K, Archer G H, Hunt R G, Jones T W 1992 The archeology of contemporary landfills. *American Antiquity* **57**(3): 437–47

Schipper L 1996 Life-styles and the environment: the case of energy. *Daedalus* **125**(3): 113–38

Stern P C 1992 Psychological dimensions of global environmental change. *Annual Review of Psychology* **43**: 269–302

Stern P C, Dietz T, Abel T, Guagnano G A, Kalof L 1999 A value-belief-norm theory of support for social movements: the case of environmentalism. *Human Ecology Review* **6**: 81–97

Weiss M J 2000 *The Clustered World: How We Live, What We Say, and What It All Means About Who We Are*. Little, Brown & Co., New York

Wernick I 1996 Consuming materials: the American way. *Technological Forecasting and Social Change* **53**(1): 111–22

Wilk R 1998 Emulation, imitation, and global consumerism. *Organization and Environment* **11**(3): 314–33

F. Duchin

Consumption, History of

The history of consumption, though a distinct field of historical research and writing only since the 1980s, has recently emerged as an especially vigorous area of inquiry. Stimulated by the visibility and power of the consumer economy in contemporary societies, scholars have redoubled their interest in investigating its past—to locate its beginnings, trace its growth, and assess its historical effects. In principle, then, to write a history of consumption is to shed light on all of those practices, institutions, identities, and meanings attached to the merchandising, acts of purchase, and final use of goods. Earlier traditions of social history addressed some of these concerns. The Annales School stands out for its abiding interest in material culture, especially in the early modern world (Braudel 1977, Roche 2000).

British social historians too, in the course of debating working class standards of living and moral economy during the industrial revolution, were clearly concerned with consumer habits, as were historians of France when writing about the subsistence crises

and bread riots of the old regime and Revolution. Increasingly, subjects such as leisure-time use, once considered separately, are now grouped along with the study of tourism, cinemagoing, television viewing, and other types of spectatorship and reception, reading included, as properly belonging to the history of consumer cultures, especially insofar as they are produced under capitalist exchange relations. The coherence of the field derives, thus, not from any single topic, but rather from the large and complex role that consumer practices are now interpreted as playing in forming collective identities, social movements, and individual subjectivity.

1. Approaches and Concepts

Because there is no general theory for why people consume, and the determinants of consumer practice are recognized as ranging from the economic calculations involved in obtaining commodities to the psychological satisfactions derived from their possession, the methodologies deployed to study the history of consumption are decidedly eclectic. Social history is still the dominant approach. But because of the complex nature of the subject matter and the fact that the timing of the study of consumer cultures coincided with the rising fashion for cross-disciplinary and multicultural studies, histories of consumption draw on an unusually wide range of disciplinary influences. These range from economics, marketing research, anthropology, and sociology to cultural and cinema studies and art and design history.

Evidence of past consumer behaviors is provided by sources as diverse as industrial archeology to recover the shards of past material life; probate records to discover the legacies and bequests of the dead, thereby to estimate the value they put on goods; business records to reveal merchandising strategies; and advertising copy, both to disclose the imagination of taste brokers and to reveal the embedding of 'social tableaux' in the popular imagination. Topics such as leisure-time use, which were once treated as specialized subject areas, are now grouped along with tourism, cinema-going , television-viewing, and other types of spectatorship and reception, reading included, as properly belonging to the history of consumer cultures, especially insofar as they are produced under capitalist exchange relations. The coherence of the field derives, thus, not from any single topic or debate, but rather from the belated discovery of the large and complex role that consumer practices play in forming collective identities, demands for social entitlement, and individual subjectivity.

The subfield's rapid establishment evidences a convergence of pressures. One stimulus came from social historians critical of neoclassical economists who, by treating consumer choices as ahistorical, reducible to demand curves based on the price and availability of goods, ignored the symbolic and social value of commodities. Social historians were also critical of the previous generation of historical scholars who, by focusing on social relations around production, failed to capture the significance of consumer practices to class, politics, ideology, and values.

Another stimulus arises from the so-called postmodern or postscarcity sensibility of the late twentieth-century West. The generally high living standards, the rapid shift from manufacturing to service-oriented industries, and the densely symbolic consumer cultures of late twentieth century capitalism seem to have profoundly altered conceptions of individual and social identity such as to pose a double question. One is whether the postmodern epoch has not engendered an entirely new relationship with material life, so much are modern cultures in the thrall of material abundance to the exclusion of nonmarket values; the other is whether the ambivalence about goods so deeply rooted in Western culture did not previously inhibit study of the power of things to shape social life (Featherstone 1991).

At first, the failure of socialist-type planned economies, which construed themselves as the only fundamental alternative to free-market, individualist consumer capitalism, reinforced a post-Cold War Consensus that consumer freedom was an unalloyed social good, engendering a celebratory attitude toward consumer abundance. As time passes, however, the globalizing trends that go under the names 'Americanization,' 'Coca-colonization,' or 'McDonaldization' has generated worries that the rapid-fire spread of Western-style consumer habits will not just destroy indigenous ways of life suited to local resources and customs, widening the gap between rich and poor nations, but also ruin the global environment.

Finally, feminist historiography has opened up whole new vistas on consumer practices by highlighting the gendered division of labor around domestic goods, provisioning, and shopping, the bridges between market-driven consumption models and household decision making, and the timeworn tropes associating the female figure with the fickleness of changing styles of culture and consumption (de Grazia and Furlough 1996).

The self-made, often empirical, character of the new field belies the influence long exercised by political economy and historical sociology on ways of thinking about consumption. Adam Smith's famous dictum from the *Wealth of Nations* (1776) that it is innate for people to 'treaty, barter, and purchase' (Bk. 1, Chap. 2) is often cited to legitimate Western models of acquisitive individualism, as are his remarks that 'Consumption is the sole end and purpose of all production … The maxim is so perfectly self-evident that it would be absurd to attempt to prove it' (Bk. 4, Chap. 8). Karl Marx's distinction in *Capital* Vol. 1

(1867) between use-value and exchange value is still influential among critics of capitalist consumerism, used to differentiate between so-called real and false needs, the latter fostered by bourgeois stratagems to fuel 'false consciousness.'

The 'fetishism of commodities,' meaning the process by which exploited human labor is embodied in goods to give the appearance that they circulate the market under their power, their value thereby derived from relations of exchange rather than the alienated labor that originally went into their creation, is still an important reference point for historians critical of the commodification of culture and other goods. The Marxist tradition is also referred too for having emphasized production as determining of class identities to the neglect of relations around consumption.

Non-Marxist German historical sociology also stands as an important reference point. Max Weber is often cited for his paradoxical argument that capitalist accumulation was spurred by an Puritan asceticism. Seventeenth century Dutch indulgence in their 'embarrassment of riches' belies Weber's argument (Schama 1987). Consumerism, a recent critic of Weber argued, is central to the modern Western world outlook, but its origins lie not in capitalist relations of exchange or the Protestant religious ethic so much as in the eighteenth century romantic structure of feeling, combined of alternations of 'anticipatory pleasure and consumatory disappointment,' (Campbell 1987) Weber's more lasting contribution is to recognize the signal importance of socially differentiated consumer practices, based on *de facto* elite monopolies over certain goods, to reinforcing social status and power.

This intuition is most fully developed in the French sociologist Pierre Bourdieu's notion of 'cultural capital' and the ways in which late twentieth century elites manage it to establish class 'distinctions.' Maurice Halbwachs too, writing early in the century (1912, 1933), made the important point on the basis of on cross-national comparisons of lower-class standards of living that different notions of lifestyle, based on class identities, might be more important to shaping consumption practices than differences of income. The influence of his contemporary, the anthropologist Marcel Mauss, is still notable for his distinction between commodity exchanges under capitalist systems and the 'gift economies' based on reciprocity of non capitalist or precapitalist exchange systems. No historian of consumer culture can forego familiarity with the US sociologist Thorstein Veblen's *Theory of the Leisure Class* ([1899] 1953).

A mordant critic of the turn-of-the century US elites, he excoriated them for renouncing the craft spirit to revel in 'conspicuous consumption.' Rivals for place and power in an ever more fluid social structure, they engaged in 'pecuniary emulation,' generating the phenomenon, reproduced down the social ladder of what later would colloquially be called 'keeping up with the Joneses.' Veblen's contemporary, the German historical sociologist Werner Sombart, is equally interesting for hypothesizing that different types of exchange relationship (moving from feudalism to the 'high capitalism' of his epoch) yielded very different overall outlooks toward wants and needs (Sombart [1913, 1967]). He was the first to argue, on the basis of his study of the US working class in *Why is there no Socialism in America?* (1904) at the beginning of the twentieth century, that an abundance of goods, consumed within a more or less democratic framework, could significantly condition political ideologies.

2. *Periodization*

Like historians of industrialization, historians of consumption struggle over the timing of the moment of birth of a mass consumption society. Some argue the breakthrough was as early as the sixteenth century, fruit of the circulation of images related to the printing press and the cultural exchange around the Mediterranean world and with Northern Europe (Muckerji 1983). Others argue that the eighteenth century was when the collective mentality of the medieval world based on scarcity, austerity, and self denial, evolved into what the philosopher Charles A. Taylor calls the 'culture of ordinary life.'

Historians of the early nineteenth century forcefully place the first consumer revolution within the context of the vast shifts in terms of trade that saw Europe's emergence as the center of the first global world economy. That even the habits of urban lower classes and perhaps even rural cottagers changed in this period is debated. Optimists argue that the demand side was as important as the supply side in the industrial revolution, and modern manufactures such as Wedgewood porcelain ware benefited not just from aristocratic and bourgeois patronage, but also from the custom of a broad band of people capable of purchasing 'populuxe goods' (McKendrick and Fairchild 1993). That workers in the British Empire early acquired new tastes for sugar, tea, and rum, and rural cottage workers were willing to exploit themselves harder to enter the market to purchase exotic wares, is now well-demonstrated. Still, both the British and American colonial experience show that emergent consumer culture was not an unalloyed progress.

Sugar, as it sweetened tastes, also toughened the bonds of empire. The boom in demand was an incentive to expanding slave systems in the colonies; sugar and tea were also stimulants that eased the pangs of early industrial manufacturing's appalling work rhythms (Mintz 1993). American Colonial records show that householders traded off declining health standards for an improved sense of psychological well-being; tea drinking and smart tea sets

spread while nutrition declined (Shammas 1993). What seems sure is that as material life changed dramatically in the course of the first and second industrial revolution, the gap between the conditions of living of a goods-rich bourgeoisie and a goods-poor working class and peasantry was hugely enhanced. This divergence of conditions would only begin to diminish from the first third of the twentieth century.

To speak of the first mass consumer society, one has to turn to the USA around the beginning of the twentieth century. There, the 'consumer revolution' was favored by a peculiar combination of vast economic resources, quasidemocratic outlooks, and the rapidly expanding needs of a huge, mobile, urbanized emigrant population. Workers were decently enough paid, class stratifications around consumption enough diminished, and markets highly protected to encourage large units of manufacturing and retailing to orient themselves to standardized low cost per unit production for a mass market. Cheap, abundant food and ample housing reduced the family budget share devoted to nutrition and shelter (which summed together in Western Europe still consumed as much as 75 percent of the working class family budget through the 1950s). This freed monies for other goods. By the late 1920s, as many as three-fifths of the US people were 'consumers,' with all due attention to huge differences of income. The Ford Model T (1908) was the paradigmatic good, and US production and domestic consumption of automobiles outstripped the whole world's combined down to the 1950s.

Not surprisingly, the USA pioneered the invention of practices and institutions related to mass consumption (though not the department store, which was the fruit of the nineteenth century bourgeoisie and highly developed in Western European large cities as well as Japan). Chain-store retailing started as early as the 1840s. The cash register began to be widely marketed from 1884. Retail credit had become widespread by the 1920s. Scientific advertising campaigns drawn up by complete advertising agencies dated from before World War I. Mass marketing using survey research and public opinion polling dated from the 1920s. The first government-sponsored national census on distribution occurred in 1929, compared to the British (1951) and the French (1965). Supermarkets flourished from the early 1930s.

Government aggressively promoted exports not just of consumer goods but also images and institutions of the consumer-centered US way of life, the belief that there was no real freedom without consumer abundance becoming a key tenet of US anticommunist foreign policy by the early 1940s. More and more, entitlement as citizens was couched as a right to a decent standard of living, to be achieved not through economic leveling or by government provision, but through free markets and higher incomes. Cold War 'consensus history' attributed the USA's purported superiority as a 'people of plenty' to the relatively harmonious 'communities of consumption' by contrast with Europe's class divisiveness. Recent historical studies are invaluable to weigh the formation of and cracks in this consensus. Whereas studies of the 'brokers' of the new circuits of consumer culture, including advertising and department store managers, etc, put the emphasis on the blunting of class consciousness (Leach 1993), studies of immigrant and working-class culture show that the language of commodities could be effectively used as civic dialog to refashion demands for social justice as demands for entitlement to a high standard of living (Cohen 1990).

Western Europeans would not achieve similar changes in standards of living until the 1960s, in the wake of the destruction of the most rigid aspects of the bourgeois class order of prewar Europe, reconstruction under the hegemony of US consumer models, the growing confidence of entrepreneurs in the face of widening and deepening markets, the pressure of labor movements on behalf of higher wages, not least government belief in Keynesian growth politics that spurred demand by supplementing low wages with generous welfare state protections.

If there is still no one historical account of the great transformation of the twentieth century which turned postwar Western Europe into a relatively affluent society, recent studies nonetheless pick up on several singular features. One is the sharp class stratifications carried over from the aristocratic into bourgeois regimes of consumption (Roche 2000, Auslander 1996). Another is the centrality of the world of small shopkeepers well into the twentieth century. Before the so-called 'Auto-Fridge Revolution' of the 1950s, there was the commodified leisure and spectatorship of the interwar period. Entry into the world of mass consumption was an arduous struggle even in the early 1950s. Evidence that affluence dimmed working class consciousness is at best ambiguous, as Goldthorpe showed in his famous study of early 1960s British auto worker (*The Affluent Worker*) although it certainly undercut earlier subcultures.

The study of the history of consumer practices in Eastern Europe picked up only in the wake of the collapse of Soviet-type planned economies. The story of the appalling scarcities of the Stalinist years in the former USSR is now well documented. However, to measure postwar, especially post-1950s, developments strictly by Western European and especially US standards misreads the extent of social provision, the pressure from ideologically competing models, the very different point of departure of development, and the paradoxes of finally delivering significant quantities of consumer durables in the 1960s and 1970s just as Western Europe passed to a postscarcity economy. More generally, new research promises to weigh the importance not just of the political regime but of deeply culturally determined ideas of market exchange in shaping wider notions of consumer culture (Merkel 1999).

3. Consumption Regimes Outside the West

The history of consumption in non-Western countries moves in two directions. One is to show the influence of Western consumer culture, the other to show the relative autonomy of developments by delving deeper into national and regional legacies.

The first instinct of writers on non-Western consumer culture was to emphasize the imperialistic nature of Western consumption habits. Aggressive corporate capitalism, most notably US-based media and communications corporations, not only sought uncontested economic power, using force to sustain it, but also fostered a hypertrophic consumer culture so out of proportion to local economic developments that it was capable of snuffing out indigenous cultures. Western consumer culture thereby perpetuated if not aggravated, semicolonial dependency. Closer study, critical of the Eurocentric presuppositions of 'cultural imperialism,' highlights the complexity of the interaction between international commodity flows and indigenous material life.

Subaltern studies posit the essential autonomy of postcolonial subjects; ethnographic studies dissect the multiplicity of interpretative responses to the imported consumer mores; anthropological studies probe the creative uses of material culture as a 'dialog around development.' None would dispute that the major direction of cultural flows is currently from the West, notably from the USA in the form of US 'infotainment' industries. However, studies focusing on practically all of the world's regions make a strong case for seeing significant variation in local vernaculars, notably in the cinema and popular music. Major differences turn on how much freedom local cultures have to refashion, 'creolize,' or hybridize material and cultural commodities (Smart 1999).

As historical research on consumption widens geographically, more and more emphasis will be put on the relative autonomy of developments, the evolution of sensibilities about material goods different from in the West, the factors that explain the disposition to or resistance against using Western goods (Bayly 1986, Hamilton and Chi-Kong Lai 1989). If the measure is not the quantity of goods and durability of structures but the quality of life that provided for physical well-being but used resources economically, the Japanese compared favorably with the British in the nineteenth century (Hanley 1997). Japan clearly had a distinctive trajectory of development. Urban markets and print culture flourished as early as the Tokugawa period in spite of sumptuary laws (ended only in 1868) that prescripted behaviors on everything from the size and location of residence to style of headgear and footwear. The emergence of mass consumption before World War II was disjoined from democratization. The males of the modern middle-salaried class were notably new goods oriented, out of eagerness to show off their capacity to master Western-type technologies and thereby legitimate their place in the social hierarchy.

4. New Agendas

The history of consumer movements stands as an important chapter of the history of consumption. Consumer cooperatives have a venerable history, starting from 1844, when the impoverished spinners of Rochdale in the north of England formed associations to secure their food supply against speculators and the exploitative practices of company stores. From the late nineteenth century, often in connection with the socialist movement, cooperativism spread globally, sometimes to fall victim to political repression, sometimes to succeed so well as to become practically indistinguishable from capitalist enterprises except in its democratic membership statutes and payout of dividends. Movements on behalf of consumer protections were pioneered in the USA as early as the Progressive era.

The goals were to regulate against corporate malfeasance, regulate against consumer fraud, promote truth in labeling laws, and to protect domestic manufacturing and labor unions with Buy National campaigns. Public interest campaigners, most famously the American Ralph Nader, overturned the age-old rule of *caveat emptor* through campaigns and law suits to hold manufacturers liable for consumer safety. Aside from studying actual movements of protest or reform, historians of consumer cultures constantly engage with the vexed question as to whether mass consumption, like the 'bread and circuses' in Roman times or the provisioning boards of absolutist monarchies, has blinded subaltern classes to their exploitation or whether goods, services, and occasions of consumption cannot equally be used, like other social activities, to create communities of opposition, facilitate communication (and thereby to provide alternative perspectives on prevailing power arrangements), and propose new standards of living that, implicitly at least, challenge the existing distribution and styles of wealth and power.

If one is to find fault with this vital field, it is that its rapid growth had caused it to reflect too little on its own genealogy and on the issues that spurred its development. If, initially, the field referred largely to the contemporary period, it now embraces Greco–Roman antiquity as well as postcolonial Central America, the superabundant material culture of the West, and the economies of scarcity of Africa and Southeast Asia. What then, if any, are the shared concerns? One certainly is methodological, common to social history generally, namely to overcome the bias toward the object-laden as opposed to the objectless in the reconstruction of the past. Another concern is to consider the symbolic charge of goods in tandem with the inequalities and asymmetries of power that to a greater or lesser degree are common to

all societies. Historically, people have always used goods to think about their society. Ultimately, the historian's charge is to discover, classify, and compare the extraordinarily different range of ways this has occurred.

See also: Capitalism; Consumer Culture; Consumer Economics; Consumer Psychology; Consumption, Sociology of; Income Distribution; Leisure and Cultural Consumption; Leisure, Psychology of; Mass Society: History of the Concept; Poverty in History; Social Inequality in History (Stratification and Classes); Social Stratification; Wealth Distribution

Bibliography

Appadurai A (ed.) 1986 *The Social Life of Things: Commodities in Cultural Perspective*. New York

Auslander L 1996 *Taste and Power: Furnishing Modern France*. Berkeley, CA

Bourdieu P 1979 *Distinction: A Social Critique of the Judgement of Taste*. Paris

Braudel F 1977 *Afterthoughts on Material Civilization and Capitalism*. Trans. Ranum P M. Baltimore

Braudel F 1982 *Civilization and Capitalism, 15th–18th Centuries*, 2 Vols. Trans. Reynolds S. New York

Brewer J, Porter R (eds.) 1993 *Consumption and the World of Goods*. London

Campbell C 1987 *The Romantic Ethic and the Spirit of Modern Consumerism*. Oxford

Cohen L 1990 *Making a New Deal: Industrial Workers in Chicago, 1919–1939*. New York

de Grazia V, Furlough E (eds.) 1996 *The Sex of Things: Gender and Consumption in Historical Perspective*. Berkeley, CA

Featherstone M 1991 *Consumer Culture and Postmodernism*. London

Halbwachs M 1912 *La classe ouvrière et les niveaux de vie*

Halbwachs M 1933 *L'Evolution des besoins dans les classes ouvrières*. Paris

Hamilton G, Chi-kong Lai 1989 Consumerism without capitalism: Consumption and brand names in late Imperial China. In: Rutz H R, Orlove B S (eds.) *The Social Economy of Consumption*

Hanley S B 1997 *Everyday Things in Premodern Japan*. Berkeley, CA

Leach W 1993 *Land of Desire: Merchants, Power and the Rise of a New American Culture*. New York

Mauss M [1924] 1990 *The Gift: The Form and Reason for Exchange in Archaic Societies*. Trans. Halls W D. New York

Merkel I 1999 *Utopie und Bedürfnis: Die Geschichte der Konsumkultur in der DDR*. Cologne, Germany

Muckerji, C 1983 *From Graven Images: Patterns of Modern Materialism*. New York

Roche D 2000 *A History of Everyday Things: The Birth of Consumption in France, 1600–1800*. Cambridge and New York

Schama S 1987 *The Embarrassment of Riches: An Interpretation of Dutch Culture in the Golden Age*. New York

Siegrist H, Kaelble H, Kocka J (eds.) 1997 *Europäische Konsumgeschichte*. Frankfurt, Germany

Sombart W [1913] 1967 *Luxury and Capitalism*. Trans. Dittmar W R. Ann Arbor, MI

Smart B (ed.) 1999 *Resisting McDonaldization*. London

Veblen T [1899] 1953 *Theory of the Leisure Class: An Economic Study in the Evolution of Institutions*. New York

Weber M [1904–5] 1958 *The Protestant Ethic and the Spirit of Capitalism*. New York

Weber M 1978 Classes, status groups and parties. In: Runciman W G (ed.) *Max Weber: Selections in Translation*. Cambridge, pp. 43–61

V. de Grazia

Consumption, Sociology of

The sociological study of consumption is as old as sociology itself. Its empirical roots lie in the early development of statistical methods. In the middle of the nineteenth century, information about household budgets in different parts of Europe began to be collected with the purpose of determining the subsistence minimum. These studies concentrated on the consumption of food. This is how the statistical regularity named after its inventor Engel was established. It showed that a rise in income correlates with a decline in the share of food consumption of all household expenses. This seems to hold true even today (Ilmonen and Pantzar 1986).

Although household budget research had theoretically interesting results, it had a pretheoretical character. Theory was introduced into sociological research by Marx (1976). According to him, a commodity had both a use value and an exchange value. He further argued that in a market economy the latter began to dominate at the same time as the production relations between people turned into commodity relations. Marx called this phenomenon commodity fetishism. His views had an important effect on the sociology of consumption in the 1900s.

For Marx, consumption was subordinate to production. This idea was common in his time. A human being was understood primarily as a producer—as a *homo faber*—and the concept of consumer had not yet been invented. Its invention has been attributed to Condillac. However, his concept of *consommateur* is ambiguous, as it also refers to eating. Perhaps Condillac, like Engel, saw eating as a basic form of consumption. More importantly, however, he conceived the consumer as an abstract entity that satisfied its needs with objects produced for the market.

1. The Classic Stage of Sociology: The First Wave of the Sociology of Consumption

It was only in the classical period of sociology that consumption began to be analyzed as a distinct sphere of action. This period witnessed the first wave of the sociology of consumption. Several of the major sociologists of the time (e.g., Sombart, Weber, Tarde) analyzed consumption as part of the birth of capital-

ism and the modern way of life, paying special attention to fashion and the consumption of luxury goods. From the viewpoint of the sociology of consumption, the most important of them was Simmel (1930), whose *magnum opus* dealt with money as the mediator between consumption and various qualitative relations. He also analyzed the significance of consumption as a manifestation of city life and noted how central a mechanism fashion was in the modern world.

Simmel approached fashion from the perspective of social class. The upper strata rejected the prevailing fashion when it began to trickle downwards in the social hierarchy. It is Veblen (1926), however, who is regarded as the main formulator of this 'trickle down' theory. He called the way the upper class strived to distinguish itself from other classes 'conspicuous consumption.' Its central characteristic is the acquisition of expensive luxury goods with the aim of improving one's social standing. As the consumption of these goods becomes increasingly common first in the middle class and later in the working class, the upper class has constantly to struggle to show its standing by means of consumption. This is the driving force of consumption. Mills (1976) later emphasized its importance when he wrote about the 'status panic' of the middle class.

2. *The Second Wave of the Sociology of Consumption*

The second wave of the sociology of consumption spanned the interwar period and the period immediately following the World War II. Mauss (1925), better known as an anthropologist, was developing a general theory of exchange and gift giving. According to him, the latter is not as selfless an activity as might appear at first glance. The person giving a gift is, in practice, expecting a return present, even though they may deny this. Mauss' observations were to have a far-reaching influence on the sociology of consumption.

Another line of European sociology that touches upon consumption is connected with the critical theory of the Frankfurt school. On the one hand, Benjamin (1972) continued Simmel's work in describing city life. He focused on shopping alleys and popular culture as manifestations of modern life. He also studied fashion, but as an indicator of time rather than class. Blumer (1969) continued to develop this theme in the US. On the other hand, Adorno and Horkheimer (1979) criticized the spreading of commodity fetishism through popular culture. They did not restrict their criticism to consumption, though. They used it to show how capitalist society comes to an impasse when it subordinates use value to exchange value. This theme was developed further in both the German sociological tradition, including Haug's (1971) critique of commodity aesthetics, and in the American sociological tradition that dealt with the artificial aging of products and the manipulation of consumers. This approach has long roots. It reaches back to Adam Smith's critique of 'vanity consumption' and extends to the twenty-first century in the persistent criticism of marketing and consumerism.

In another American research tradition consumption is conceptualized more broadly. Consumers are not considered mere victims of the 'captains of markets' but are seen as actively expressing their values and goals through consumption. Lazarsfeld (1982), for example, pointed out that market communication is not etched in consumers' minds as on an empty board; instead, its message progresses in two stages through opinion leaders to people at large. Even then its reception depends on existing tastes and preferences. This realization led to another: that consumption provides excellent material for the analysis of social change. Riesman (1969), among others, used all sorts of consumption data to describe the moral change in American society. He argued that there remained only one and at the same time the most modern personality type: the 'other-directed' personality, for which consumption had become a central dimension in life.

3. *The Differentiation of the Sociology of Consumption into a Sphere of Its Own: The Third Stage of the Sociology of Consumption*

The third stage of sociological consumption research and, at the same time, the formation of the subdiscipline of the sociology of consumption occurred at the end of the 1970s. While previous sociological analyses were characterized by a great deal of moralizing about consumption, consumption now came to be seen as a meaningful social activity that offers a valuable perspective on society.

The new phase was ushered in by Bourdieu, who had also come to be known as an anthropologist. He ended up describing, by means of consumption, the maintenance of the class structure of French society in a way that resembles Simmel's and Veblen's view of status commodities trickling down in the social hierarchy. However, whereas Veblen stresses that consumption becomes an important means of making social distinctions in a situation in which differences are disappearing, Bourdieu (1979) argues that distinctions are embodied in specific habituses and internalized in consumption choices. He is, of course, not the only one who sees consumption as a means of maintaining class barriers and social distinctions. Concurrently with his research, similar studies were being carried out in the Anglo-Saxon world. These studies showed that consumption reflects important social divisions and temporal categories. This research tradition (e.g., Sahlins 1976) departs from Bourdieu's approach in that it leans primarily on the Durkheimian

legacy. Consumption is seen within it as mirroring not only all sorts of social distinctions (class, age, gender, etc.), but also social time and qualities of its social context (feast or everyday life, dining in or out, etc.).

In the 1980s and early 1990s we saw the sociology of consumption characterized by at least four research orientations. First there was the approach, connected with the postmodern turn and influenced by Simmel and Benjamin, which focused on the meanings of consumption. In line with Baudrillard's theorizing (1988), it argued in the poststructuralist spirit that consumption is not a language. Clothing, food, etc., do not constitute a language-like sign system open to a uniform reading by all members of the same culture. Rather, consumption forms mere systems of classification that become increasingly important as the significance of consumption increases. Attention is thus focused on the role of consumption in the formation of social identities and on the embodiment of consumption. When seen as a crucial building block of identity, consumption has been regarded as an important expression of *Zeitgeist*. This perspective has been easily adaptable when defending the idea of a postmodern turn.

Then there is the empirically oriented approach centered on food. It analyzes a theme made topical by globalization and by the changing basis of social order: the stability and change of consumption patterns. The concrete objects of research include the symbolic dimension of foodstuffs, the national and cultural aspects of meals, the gender-, age-, and class-related forms of eating, eating habits, the 'problem' of omnivorousness and eating disorders, eating out, and preparing meals at home, and the fate of the family meal. Sociologists of food have found that food choices are surprisingly stable despite the improved availability of all kinds of foodstuffs. They have also proved that the direction of change is not only from centers to peripheries ('Macdonaldization') and that the flow of luxury goods is not only from top to bottom in the social hierarchy: in many cases a dish originally favored by the common people has become popular among upper classes too (like pizza and traditional Chinese meals).

A third research orientation is rooted in the Durkheimian and Maussian tradition dealing with the circulation of commodities. It elaborates on the central features of the institution of gift-giving, focusing on the ways in which an object, *qua* present, loses the status of commodity, as it is bad taste in Western societies to give presents that are merely commodities with an exchange value. Wrapping the present up in an ornamental package signifies the loss of the commodity status. It is also thought that a gift must be 'special.' For it to be experienced as valuable, it must reflect the personality of its giver.

This research tradition has also examined hobbies and exceptional consumption patterns, especially collection. It has been found that any product or object can become a collectible, thereby losing its commodity status and gaining a new meaning in relation to the already existing or imaginary parts of a collection. When a collection is seen as 'complete,' it can be commodified and sold on the market.

The fourth line of research in the sociology of consumption deals with the questions: What are the aesthetic and other characteristics that make commodities appealing to us? These reflections have resulted in the rejection of various theories of basic needs and pre-established need hierarchies. Needs are, by contrast, seen as mediated by the cultural filter and as historically changing forces that are as much the outcome as the starting point of our activity. In addition, this research orientation addresses the way in which unsatisfied needs give rise to ultimately unsatisfiable desires and how this process contributes to the creation of the consumer society. In this context Campbell (1987) has pointed out how modernization has refined the quantity-oriented hedonism based on the Protestant ethic into quality-oriented hedonism. Contrary to Weber's view, the romantic 'spirit' propitious to self-realization and consumption too has played an important part in this process of refinement.

4. *Trends in the Sociology of Consumption*

The meaning–centricity of the sociology of consumption began to be criticized as too cognitivist at the end of the 1990s. This cognitive paradigm is committed to the ideas of agency and reflexivity, neglecting the fact that only part of consumption can be reduced to them. Furthermore, it reduces the entire sphere of consumption to meanings. The focus has been on the external meanings of commodities, usually provided by researchers themselves. From these models researchers have deduced the meanings consumers must have intended to convey to others. Although the subjective dimension of consumption (its personal significance) has been studied, the emphasis has still been on its cognitive aspects at the expense of its experiential side. Finally, only 'public,' or 'visible,' consumption has managed to kindle the interest of cognitively oriented sociologists.

Attempts have therefore been made to broaden the perspective of the sociology of consumption. Of course, the meanings attached to consumption and their connections with the spatiotemporal and social structures of everyday life and identity formation will continue to be analyzed. It is, after all, impossible to imagine consumption without meanings. The central mechanisms of consumption and their connection with social order are also likely to continue to receive attention. Nevertheless, there is evidence to suggest that the sociology of consumption is fragmenting into several approaches.

First, it is evident that researchers are more and more interested in the experiential dimension of

consumption and thereby in the emotional bond in the commodity–person relationship. This opens up a new perspective on the consumption-driven process in which individual and collective identities are constructed historically, more or less on the basis of personal emotional sensation.

Second, the nonreflexive aspects of consumption, or consumption routines, are a new focus of research, which addresses the spatial and temporal arrangements of everyday consumption: the gendering of consumption routines (cleaning, cooking, etc.) and their variation with age, ethnic background, and religion. Routines, however, are constantly shaken by change. The accelerating social change, the increasing health awareness, and risks produced by gene technology and product forgery make consumers reflect on their routines and sometimes change them. This has thrown a challenge to sociologists, forcing them to pay attention to people's ways of dealing with consumption risks. The latter theme has brought the concept of trust to the center of the sociology of consumption too, paving the way for the analysis of trust as a means of routinizing consumption. This research orientation is formulating a theoretically interesting tension between reflexive and nonreflexive action in general.

A third research focus attracting increasing interest is the change in the spatiotemporal experience triggered by the new supergoods—products of information technology that have become part of everyday life. This change not only destroys old forms of sociality and practices connected with them (such as shopping); it is also creating new kind of virtual sociality independent of space and time. How this is constituted and reflected back on information technology is a new challenge for sociology.

The fourth rising line of research in the sociology of consumption is connected with the end results of consumption: wastes, waste management, and the recycling of used products. Both are now billion dollar industries, and research on consumer society is being complemented with research on 'rubbish society.' It opens up an interesting perspective on not only waste treatment routines but also on consumers' opportunities of commodifying their wastes and selling them on waste markets.

The sociologists of consumption do not lack objects of research. Moreover, it should be pointed out that consumption not only reflects social practices but has a dynamic of its own as well. It seems to be a self-sustaining phenomenon. For instance, the idea of a petrol-powered car implies petrol stations. These, again, imply filling up the tank, and so on. This autonomy of consumption, which Simmel noted at the beginning of the twentieth century, is drawing more and more attention, founded on the realization that present consumption choices not only guide future production decisions but also shape our common future in general.

See also: Class: Social; Consumer Culture; Consumer Economics; Consumer Psychology; Critical Theory: Frankfurt School; Economic Sociology; Fashion, Sociology of; Influence: Social; Lazarsfeld, Paul Felix (1901–76); Leisure and Cultural Consumption; Leisure, Psychology of; Marx, Karl (1818–89); Simmel, Georg (1858–1918)

Bibliography

Adorno T, Horkheimer M 1979 *The Dialectic of Enlightenment*. Verso, Thetford, UK

Baudrillard J 1988 Jean Baudrillard: *Selected Writings*. Polity Press, Cambridge, UK

Benjamin W 1972 *Gesammelte Schriften*. Suhrkamp, Frankfurt, Germany

Blumer H 1969 Fashion: From class differentation to collective selection. *The Sociological Quarterly* **10**(3): 275–91

Bourdieu P 1979 *A Social Critique of the Judgment and Taste*. Harward University Press, Cambridge, MA

Campbell C 1987 *The Romantic Ethic and the Spirit of Modern Consumerism*. Blackwell, Oxford, UK

Ilmonen K, Pantzar M 1986 Determinants in the modern choice of food. *Journal of Consumer Studies and Home Economics* **10**: 97–110

Lazarsfeld P F 1982 *The Varied Sociology of Paul F. Lazarsfeld* (Writings Collected and Edited by Kendall P L). Columbia University Press, New York

Marx K 1976 *Capital*. Penguin, Harmondsworth, UK

Mauss M 1925 *The Gift: Form and Functions in Archaic Societies*. Cotren and West, London

Mills C W 1976 Status panic. In: Feldman S D, Thielbar G W (eds.) *Life Styles. Diversity in American Society*. 2nd edn. Little & Brown, Boston

Riesman D 1969 *The Lonely Crowd*. Yale University Press, New Haven, CT

Sahlins M 1976 *Culture and Practical Reason*. The Chicago University Press, Chicago and London

Simmel G 1930 *Philosophie des Geldes, Fünfte, unveränderte Auflage*. Verlag von Duncker & Humblot, Munich and Leipzig

Veblen T 1926 *The Theory of Leisure Class: An Economic Study of Institutions*. Allen and Unwin, London

K. Ilmonen

Contemporary History

1. Appearance of the Term and the Activity

Contemporary history (*Zeitgeschichte*, *Histoire contemporaine* or *Histoire du Temps Présent*, *sovremennaia istoriia*, *storia contemporanea*, etc.) has recently become a designation for a branch of historical research and writing that emerged in the period after 1945. This period signifies the history of events leading up to, and including, World War II and the period

since. The term, however, is far older. The 'idealist' philosopher of history, Benedetto Croce (Croce 1917) claimed that all history must be contemporary history. By this he meant that since the historian's mind always mediated any reflection on the past, history must reflect the mentality of the historian: his (there was no her considered) political concerns, social position, etc. Contemporary history thus seems to occupy an unstable niche between the safely past and the historian's present. During the early days of 'professional' historians, i.e., university-trained, the term appeared largely in titles by amateurs; the professionals tending to dismiss it as an impossible and 'unscientific' enterprise.

Those critical of the field reproach it with its lack of perspective and its paucity of sources, above all since most nationally preserved archives are closed for a customary 30 years or more. Critics also complain that there is no dividing line between contemporary history and mere political partisanship or 'journalism'—a term usually meant to suggest that an account must be ephemeral and partial. These critiques will be considered at the end of this article. Recall, however, that one of the very greatest of ancient historians, Thucydides, insisted that explaining and narrating the great civil war still engulfing the Hellenic world as he wrote, was a valid historical exercise.

Contemporary history has, in fact been a historiographical enterprise as venerable as any other. The great French medievalist Marc Bloch also denied a radical break between past and present. Contemporary history is contrasted with the history of a relatively remote past. (Bloch, 1949, p. 89, Noiriel, 1999, pp. 7–28). Those who distrust contemporary history suggest that 'the past is another country,' by which it is meant that the ways of thinking and acting, the principles for collective association, the bonds of loyalty and governance, are different enough from our own, such that the historian must work his way into them through hermeneutic dealing with sources. But this remoteness is no less true for many societies in the world of the twenty-first century. Nor can the remote past be totally strange; otherwise there would be no way of arriving at statements of motivation or descriptions of behavior. No history can presuppose radical alterity.

The time period included in 'contemporary history' has never been rigidly defined in any national usage. Contemporary history usually signifies a contrast with 'modern history,' a term that most Western historians now apply to the entire period since the French Revolution, although before 'contemporary history,' came into usage, 'modern history' was sometimes extended to the entire epoch since the close of the Middle Ages. In 1775, the German historian J. G. Büsch distinguished modern from medieval and ancient history, and within modern history, an epoch even more recent. After the French Revolution, some historians, including Ranke, labeled post-revolutionary history as the 'newest' or more recent history within the newer history dating from 1492. Not only Leopold von Ranke, but also many of the major German historians of the nineteenth century devoted lectures to the most recent events, whether the revolution of 1848 or the founding of the German Empire in 1871. Defending the study of contemporary history in 1886, Emile Bourgeois contrasted its application by scholars and teachers to events after 1789 with the popular notion of the history 'of our own generation.' The latter, he proposed, faced pitfalls, but what he preferred, following the Germans, to call 'modern history' since 1789, could be studied as a science. (Bourgeois 1886 pp. 12–4) In today's contexts the term has become a synonym for twentieth-century history or Western history since 1914 and the outbreak of World War I. The *Journal of Contemporary History*, founded by Walter Laqueur and George Mosse in 1966, has taken the twentieth century as its purview and its first issue focused on the rise of fascism. In other contexts, it remains the history of the ideological conflicts that led to the rise of fascism and Nazism, the Second World War and afterward. In the Soviet Union the Bolshevik or 'October' Revolution formed an ideologically defined boundary. Still other historians, above all in the North American context, apply it to post-World War II events, i.e., to the period since 1945. Outside Europe and the Americas it can refer to the period of national independence since the end of colonial rule.

Sometimes contemporary history is construed as comprising the era from which historical participants are still alive. Modern, as well as ancient historians have considered the availability of live witnesses to be essential: Thucydides cited the questioning of witnesses as one of the scientific sources of his researches. *The Institut de l'Histoire du Temps Présent*, reorganized in 1980, defined the presence of witnesses as a crucial criterion for its investigative mission. But if the starting point is not fixed, the end point is. Common to all definitions of contemporary history is the idea that it must extend, whether as events or as the account of those events, until the present day. It is designed to explain current social, political, or cultural development. While 'post-modern' has become a general term in cultural studies, post-contemporary remains oxymoronic, although in fact some German advocates of '*Zeitgeschichte*' argued that the post-1945 period constituted a 'newer' rather than an older contemporary history.

2. *What Qualities Define Contemporary History?*

Beyond the criterion of recentness or actuality, is there any distinguishing characteristic to contemporary history? Does it have any real specificity on non-temporal grounds? Contemporary history is history that is usually deemed to provide an explanatory or

interpretive background for events still being contested or whose ramifications are still affecting current political or social development. But non-contemporary history also performs this function. The history of African-American slavery would not be described as contemporary history even though that institution still casts a long and powerful shadow over current American politics. All major developments permanently structure today's alternatives and all history presupposes such an enduring influence. Most historians seek to defend the 'relevance' of their inquiry—sometimes by claiming analogies with the past, sometimes by claiming a real, even if remote, impact of past events on the present. Such a claim alone cannot distinguish contemporary history. Nonetheless, contemporary history often has an impact—and is intended to have an impact—on current political outcomes. Contemporary history comprises narratives and explanations that lead to policy decisions, electoral contests, and sometimes to political trials. It is a history with consequences. This does not mean that contemporary history is less capable of escaping partisan judgements than history of the remote past. Historians of all eras bring to bear their values and political temperaments. They cannot shed these although they can and should achieve self-insight into the way their psychological make-up conditions their interpretation. In this respect the contemporary historian can muster an equal degree of self-insight. Certainly his or her history can prove as encompassing and detached as the history of more remote eras. Still, contemporary history bears an intimate relationship to the choices facing the society in which the historian is rooted. Such choices often seem particularly urgent because they are the legacy, not only of long trends and development, but also of episodes of recent conflict, and sometimes violence.

Thus, while the particular beginning point of contemporary history can be subject to debate, it is usually and logically taken to be a significant 'rupture' or caesura in the recent past. (This suggests that contemporary history, now so often defined as history since 1945, may within a few decades be interpreted as history since 1989. The first issue of the Italian journal *Contemporanea* cited the events of the late 1980s and 1990s as constituting 'a turning point in world history, imposed with the force of a periodizing caesura.' (p. 3)) Such ruptures, often accompanied by radical change and intense political or international conflict, require explanations and narratives designed to account for the sudden transition of institutions or regimes. Even where a new regime decides to forswear any judgmental effort about the past, the transition itself, the process by which a seemingly stable set of institutions, gave way to another, requires intellectual mastery.

Not surprisingly, it was in France, where the Third Republic remained a contested regime until World War I, that the political stakes of contemporary history most starkly emerged. For conservatives, the French defeat at the hands of Prussia in 1870 demonstrated the Revolution of 1789 had undermined national cohesion. Hippolyte Taine's *Origines de la France Contemporaine* (1875–1893) explained the debility of French society as a legacy of the Revolution; the *Société d' Histoire Contemporaine* (1890) remained under Catholic conservative influence. But the Left did not simply cede contemporary history to the Right; Emile Bourgeois (1886) excoriated Taine's indictment, anticipated all the later criticisms of post-1989 history, but claimed the enterprise was no less valid than 'doing a writing' of the history of earlier epochs. The Republicans in power created a Chair in Modern and Contemporary History in 1884 and ensconced in the University, the historians of the Left founded the *Revue d'Histoire Moderne et Contemporaine* (1890). Pro-Republic historians fought for control of the official history texts taught in schools, among them Ernest Lavisse's 10-volume *Histoire de la France contemporaine de la Révolution à 1919* (Noiriel 1998, p. 14).

Contemporary history has thus often originated as an effort to determine responsibilities, often for disastrous outcomes (either on the part of those acting for the old order or for the new holders of power). Institutions established to research contemporary history have had the assignment of examining the origins and responsibility for recent human-rights abuses. If such painful episodes as dictatorship or state-sponsored terror, revolutionary upheaval or war and sometimes the collaboration of citizens with an occupying power had not taken place, contemporary history might be a less compelling enterprise. Sponsorship of contemporary history often involves a society's self-examination—despite the reticence on the part of those wishing to avoid confronting their own complicity—after political catastrophe, whether humiliating defeat in war or domestic recourse to terror or authoritarianism.

Such a self-reckoning encounters resistance among historians as well as from former power holders. Nominally designating the field as a special branch of history allows some of this reticence to be overcome, whether political or academic. In Germany and France, the respective terms '*Zeitgeschichte*' and '*Histoire du Temps Présent*,' were adopted in part to differentiate their fields of study from history that had been academically credentialled. In the German case, early postwar researchers into the National Socialist period confronted a widespread resistance among professional colleagues to study the Third Reich. The alleged reasons given were the closeness of the period and the passions it aroused—in fact, many scholars, of course, had been highly supportive of the regime. Still, *émigré* scholars, new political leaders, and an emerging generation of historians understood that the Nazi period could not be bracketed off on the grounds of being too close to study.

In fact, the proximity of the Third Reich led some historians to argue, not that one must be more distant, but that rather only those who had experienced the temptations of National Socialism might understand how complicated the moral as well as historical issues were in order to really understand its history. Outsiders and those who came later would allegedly lack the capacity for '*Verstehen*' or empathetic immersion into the events that German Historismus had always prescribed. As the author of a major discussion of postwar German and Japanese historiography writes (Conrad 1999, p. 217): 'objectification without the mediating role of empathy (*Einfühlung*), according to this view, had to produce a distorted image: the Germans would appear merely as a distant object and no longer the legitimate subject of history.'

3. *Institutes, Progress, and Controversies*

In 1950, the German Federal and state governments financed the Munich-based Institute for the History of the National Socialist Era, renamed simply the *Institut für Zeitgeschichte* in 1952. It became the leading research institute into National Socialism, engaged some of the most promising researchers, and since 1953 has issued the indispensable '*Vierteljahrshefte für Zeitgeschichte*,' devoted to publishing documents as well as narrative and interpretative studies, and a series of book-length monographs. The Institute and its quarterly represented recognition that the Nazi era formed part of German history but presented particular historiographical challenges. The concept of '*Zeitgeschichte*' or 'contemporary history' legitimized opening up scholarly research on this painful, indeed often shameful, past, but simultaneously recognized some theoretical distance from more distant and settled subjects. From its reappearance in 1949 to 1960, the traditional *Historische Zeitschrift* only published one substantive article on the NS period. Even the leading Third Reich researcher Hans Buchheim, argued that the NS period was not appropriate for doctoral research although courses were offered—often by visiting or returning *émigrés* such as Hans Rosenberg—at the university level. The topics usually involved the factors undermining Weimar, the foreign and military policy of the Third Reich, and the unsuccessful Twentieth of July plot. The publication of Karl Dieter Bracher's magisterial *Dissolution of the Weimar Republic* (1955) finally presented the success of the Third Reich within the long-term structures of German history; nonetheless, the author was trained as a political scientist; he understood his task less burdened by the inhibitions of the historical profession. Bracher's work duly arrayed structural, ideological, and contingent factors. Still, to envisage Nazism as a consequence of the long-term continuities of modern German history—that is, its institutional and social structures and not just supposedly general leanings toward authoritarianism—remained the agenda of subsequent historians. For the researchers of the 1960s, the buffer of 'contemporary history' became less essential: Nazism represented only the latest and most extreme recourse of old and new elites who resorted throughout the modern period to national expansion and authoritarian measures. And only by the decades of the 1970s and 1980s did historians seek to probe the 'mentalities' of ordinary Germans under the regime, to measure the acquiescence or enthusiasm for the regime and for its anti-Semitic program. The *Institut für Zeitgeschichte* carried out some of its most original work within the so-called Bavarian project organized by director Martin Broszat, which was an effort to explore the categories of enthusiasm, acceptance, and '*Resistenz*' (or small-scale noncompliance) in the Third Reich.

This research program, for all the insights it generated, revealed that the conceptual difficulties raised by the Nazi experience still lingered as special challenges for *Zeitgeschichte*. Whether in scholarly works such as those of Ian Kershaw or popular depiction of the Nazi era in the German television series *Heimat*—and contemporary history is increasingly presented by the visual media—Nazi rule still appeared to impinge as a force from outside society. It did not really penetrate the local community. Widespread acts of '*Resistenz*' suggested a society that endeavored to maintain its autonomy from politics. Contemporary history had begun as an effort to account for the highly disturbing victory of the National Socialists; three decades later, it focused on the limits of totalitarianism. Nonetheless, the most recent studies of denunciation and police activity (Fitzpatrick and Gellately 1997) have again placed the accent on the self-policing of society rather than its resistance.

In France the *Comite d'histoire de la Second Guerre Mondiale* was transformed by the initiative of Rene Rémond and other historians into the *Institut pour l'Histoire du Temps Présent*, which in the 1980s and 1990s became the leading foyer for research on contemporary France. Its debut marked a new willingness to focus on the Vichy period, which had been subject to euphemism and obscurity. Perhaps because it was early taken in hand by a younger generation of researchers—a development that occurred only later in the trajectory of the *Institut für Zeitgeschichte*—its conceptual categories for dealing with traumatic history were less evasive. It has had a remarkable series of *colloques* in which actors and historians have met together to consider such subjects as labor under Vichy or the early efforts at postwar economic and social transformation. Such countervailing and critical interaction would probably not have been possible in Germany; where in effect younger historians gradually had to press for deeper investigations. Still, French research was also subject to equivalent resistance to full disclosure and confrontation.

Postcommunist societies have devoted major academic resources to assimilating the historical legacy of the long decades of communist rule. In East Germany popular revelations about Stasi collaboration, the debates about trying functionaries of the former GDR, and the insistence of the civic movements of 1989, have all encouraged serious historical examination of the former regime. The collapse of the GDR brought a rare opportunity to do contemporary history since the documentary sources of the regime were put at the disposal of researchers within a few years of its collapse. Aside from numerous memoirs and popular accounts, a major center for carrying on this research has been the *Zentrum für Zeitgeschichtliche Forschung* in Potsdam (originally *Schwerpunkt Zeitgeschichtliche Forschung*), an institute that incorporated researchers from the former Academy of Sciences of the GDR and was entrusted with the mission of studying the GDR as a regime. While its personnel and approach were severely contested at first by dissidents, who felt that it was quasi-apologetic, the ZZF has now brought out significant edited monographs on culture and society in the former GDR as well as more traditional accounts of international relations and politics.

War origins and responsibilities have traditionally been a major source for contemporary historical research ever since Thucydides, although the investigations are not always thought of as 'contemporary history' as such. World War I led to rival publications of diplomatic documents and then narrative accounts of the international relations that led to the conflict. These were often partial, but based upon recently opened—if often very incomplete—archival sources, but the history that emerged was not conceived of as a particular or tentative branch of history. US historiographical revision, somewhat analogous to the difficult reassessments in Germany and France, also marked the late 1950s and 1960s. 'Revisionist' historians (the term 'revisionist' has had successive applications over time) aroused harsh denunciations when they suggested that Americans, not Soviets were largely responsible for the Cold War, whether because of the personality of Truman or the 'needs' of American capitalism for world-wide markets. A particularly bitter controversy arose over the use of the atomic bomb, which Gar Alperowitz suggested was designed to intimidate the Soviets, not end a war with Japan that had shortly to yield victory in any case (Maier 1970).

The emergence of the 'new' historians in Israel prior to 1990 should also be understood as part of this worldwide wave of reconsidering taboo subjects, pre-eminently the fate of the Arab population living in the area of Palestine that would become the Jewish state. The traditional accounts of the 1947–48 founding of the country overwhelmingly alleged that the Palestinians had fled from their homes in the Jewish regions on instructions from Arab leaders who promised that they would be resettled after Israel was defeated. The 'New Historians' have instead blamed Israeli leadership for aggressive action against Arab civilians in an effort to uproot as many as possible.

In Japan, debate over the origins of the Pacific War was often lively in university milieus, but more inhibited in schools and public forums. An initial postwar willingness to question Japanese policies was not followed to the same degree as in Germany and Japan has often been depicted as a victim of the war, a view reinforced by the atom bombing of Hiroshima and Nagasaki. Nonetheless, among professional historians and increasingly in the wider public, more self-critical accounts have been emerging. The Nanjing massacre has brought sustained historical debate, in which the Japanese profession revealed itself as unready to reconsider the dissident view of the war.

Thus while 'contemporary history' is a term that denotes no more than particular chronological criteria, recent output suggests that its implicit subject matter has been the opening up of the earlier 'shameful' experiences of a society, at least those within the memory of living participants. Thus the whole question and to what extent citizens cooperated remains morally and analytically troubling. Until the end of the 1960s, historians and political analysts tended to emphasize the power of the totalitarian regime—its capacity for violence, atomization, conquest, and subjugation of an autonomous society. Thereafter, analytical histories argued for the limits of totalitarian rule and the persistence of 'niches' where individuals could conduct a normal private life whether under National Socialism or Communism. By the 1990s contemporary historians were tending to examine not the immunity of society from dictatorship but rather the self-policing of society. The legacies of 1945 and 1989 in Europe have made Nazism, Fascist, collaborationist, and Communist regimes—the problem of 'totalitarianism'—perhaps the leading theme for contemporary historians, as the dismantling of South African Apartheid and military dictatorships in Latin America has done elsewhere. Contemporary history has wrestled with the issue of how to describe the nature of acquiescence, enthusiasm, and dissent under totalitarianism. Research has come to center on the relationship between civil society (that category rediscovered for scholarship by 1989) with political authority (Fitzpatrick and Gellately 1997).

4. *Contemporary History as 'Collective Memory'*

Given its highly charged political content, research into contemporary history has surged into new forums, which produce not a historical narrative as traditionally understood, but important efforts to reconstruct and critically evaluate the national past. Distinctive since around 1980 has been the wide recourse to official bodies to investigate—and to

acknowledge publicly—the human-rights abuses of authoritarian regimes that have been dismantled. The work of the South African Truth and Reconciliation Commission can be construed as an investigation into contemporary history (the case is less persuasive for commissions that took their evidence in secret). Its objective was not primarily to establish a historical narrative but to hear testimony from victims and acknowledge the violence done to those testifying. By being established with the right to confer amnesty, the TRC encouraged perpetrators to describe their role in enforcing the Apartheid regime.

Truth commissions or other committees of inquiry can, but need not, produce traditional historical accounts. The TRC, for example, elicited many partial narratives of personal experience and explanations of why and how both perpetrators and victims behaved. The findings presented are not usually cross-examined or subject to critique of sources and methods. A contemporary history of Apartheid, for example, would require an effort at synthesis and contextualization within an overarching temporal account that fully revealed. For instance, such structural and background issues as sources of investment and labor, migration of whites and blacks, long-term ideological formation, and the Anglo-Boer clash. In short, the whole relevant history of the society before and during Apartheid. It would sacrifice immediacy for more encompassing perspective (Truth vs. Justice 2000; Rotberg and Thompson 2000). Nonetheless, this sort of public hearing has increasingly served, in effect, as the first approach to a historical account of the previous regime.

The motivation behind such efforts remains more admittedly political than most academic historians would feel comfortable acknowledging. Former victims and dissenters urge exposure of the past as a moral and political as well as historical acquisition. The East German Citizens Movement (*Bürgerbewegung*) pressed for a large-scale German parliamentary inquiry, the Enquete-Commission, into the functioning of the East German regime that produced volumes of testimony and analytical essays. In this respect much of the impetus for contemporary history, as the history of a military experience or dictatorship, emerges not just from professional historians but from advocacy groups.

Such efforts represent an effort not just to advance historical research, but to establish versions of what the French sociologist Maurice Halbwachs called 'collective memory,' the construction of a group's version of the past (and not just a retrieval of events gone by). As an activity, contemporary history thus involves, along with traditional research and official boards of inquiry, many volunteer efforts to memorialize historical sites and record interviews for local history (the 'history-workshop' movement; see especially *History Workshop Journal*). The themes can be work and family experience, the impact of major events, such as war or political programs on local life. History is also increasingly presented in a museological context. Museum curators or those charged with exhibitions have pressed agendas of historical reassessment, often to encounter opposition and charges of distorting the past. The 1995 'Enola Gay' exhibit at the Aerospace Museum that was to focus on the American bombing of Hiroshima seemed critical enough of American policy to incite veterans groups and their Congressional supporters to demand closing the exhibit, which, in fact, was substantially modified. The Wehrmacht exhibit produced by the Hamburg Institute, which featured a travelling set of photographs of German army atrocities on the Eastern Front, likewise provoked criticism and reaction (Thiele 1997). Whether or not contested, such exhibits and museums play a powerful role in historical education—perhaps preeminent now is the US Holocaust Museum, which has become one of the most frequently visited Washington sites.

Finally, it is important to note that the prolific debates over museums, likewise over the concepts for memorials, have themselves become one of the major topics for contemporary history. The so-called history of memory or more precisely the history of 'collective memory' has developed apace along with contemporary history. (See the journal, *History and Memory*.) It is its most self-reflective branch, and thus appeals to the intensely self-reflective trend of postmodernist culture, historical and otherwise. Obviously each of these themes opens up not just a political but a social history of power.

5. *Methodology and Objections*

Advocates of contemporary history have sometimes suggested that it has a distinctive methodology. But in fact it must rely on the evaluation of sources that historians learn for all periods. As with the history of earlier eras, the historian must learn to compensate for the absence of some sources and to treat those available with caution. The contemporary historian can use economic data, interviews, public-opinion surveys, the media, and available sources. Other recently developed historical approaches designed for other periods—'everyday history (*Alltagsgeschichte*),' cliometric history, 'microhistory' and anthropology, can be exploited for the contemporary period as well as earlier periods.

Still, contemporary history does present particular problems: the contradiction between closeness and objectivity. How often do historians who lecture on the recent past not encounter the questioner who states, 'I was there, I know … !' The witness will usually complain that the contemporary historian does not really have the story right. But precisely because he or she was there, the witness has experienced only one level of the story. He or she cannot usually rise above

personal experience, often one that has left deep psychological scars; to attain the perspective needed for a general history. This is not a question of intent or even of mere distortions left by imperfect memory. The problem of one-sided history does not lie with the reliability of the witness or lack thereof, but with the failure to balance a story. The historian can decide that every witness or victim has, in fact, left an accurate memoir, and still have concern about the marshalling of these accounts as the major source. Those historians of Stalinist purges and the Gulag who stress their enormous human toll have given a higher value to memoirs and autobiography of 'victims' than so-called revisionists who have contested the high number of victims and feel that victims' memoirs have colored the objectivity needed about the period (Getty and Manning 1993). The issue is not the reliability of the sources or the accuracy of memory, but of contextualizing the memoirs. The German historians who documented the expulsion of Germans from East Central Europe in 1945 took special care to test the reports of Soviet brutality and rape for accuracy. But that was hardly the point: what made these reports problematic was the fact that the authors showed no equivalent concern for the victims on the other side and thereby could hardly furnish a satisfactory account of causality (Aly 1997, Conrad 1999). Contemporary history can be afflicted with a misplaced and misleading stress on the reliability of witnesses at the expense of contexualization.

Nonetheless, most critics reproach the field its paucity of traditional sources, above all since most nationally preserved archives remain closed for a customary 30 years or more. This is often true, but it is also true to an even greater extent for periods in the remoter past. The mass of printed documentation, of statistical inquiries, of public opinion and media reports, of possible interviews, and now telephone and other records is in fact overwhelming. True, written minutes and archives will reveal debates within governmental agencies that may not be fully disclosed to the contemporary history. The quality of history that can be written once these become available will be more informed. But this cannot disqualify contemporary history from counting as respectable scholarship.

Critics also complain that there is no dividing line between contemporary history and mere political partisanship or journalism—a term usually meant to suggest that an account must be ephemeral and partial. But of course the issue is what sort of journalism. Historians journalistic reporting and analysis use as source material in its own right. There are historians whose narration of the past remains as superficial, as keyed to headlines without any probing of deeper explanations, as the most superficial newspaper report. And there are journalists who endeavor to uncover and integrate in a written text models of explanation as deep and persuasive as what historians propose. The valid distinction is not between history and journalism, but a history or journalism that is content with day-to-day reporting—a disclosure of events—and a history or journalism that seeks to probe and place immediate events into a framework of long-term analysis.

Contemporary history, it is charged, cannot transcend the ideological passions of the day and is thus less 'objective' than history of earlier times. Any reflection, however, on the history of major events will show that while consensus does finally build around some attributions of responsibility, many deep interpretative divisions persist or are renewed. There is no transcendent view of whether the Gracchi were social reformers or mere demagogues; whether revolutionary violence was a response to real enemies or a pure paranoia; whether the problem of American slavery might have been resolved without civil war; whether Germany was particularly responsible for the outbreak of World War I; and whether the Americans do or do not bear some of the responsibility for the Cold War. Political divisions are fundamental to human societies and there can be no more value-free a resolution of the issues involved for historians than there was for contemporary actors. But this does not mean that historians are condemned to partisanship. Whether dealing with events of the moment or those long since over, researchers can follow certain safeguards. They can report the evidence that might contradict their chosen narrative interpretation or current political preference, and they can seek to include the reasoning of adversaries within an overall historical account such that it becomes multi-voiced.

It is true that contemporary history is likely to be superseded. But this is true of all historical treatments including those of antiquity—each will be rewritten as the interests and preoccupations of a current generation of scholars change. All history is subject to revision: revision in questions, revision in methods, and revision in perspectives.

Contemporary history enjoys no less sanctity. Granted, the contemporary historian must remain particularly aware that the questions that motivate him or her may be dictated by current controversies and thus soon superseded. The contemporary historian should self-consciously seek to anticipate, say, frameworks that successive generations might pose and not just the agenda of the moment. Some anticipation of the future, at least as a spectrum of possible issues and inquiry, as well as contextualization in terms of the remote past, remains necessary for the staying power of contemporary history. Thucydides understood that. Thucydides abides.

See also: Bloch, Mark Léopold Benjamin (1886–1944); Collective Memory, Psychology of; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); National Socialism and Fascism

Bibliography

Aly G 1997 *Macht—Geist—Wahn. Kontinuitäten deutschen Denkens*. Argon, Berlin

Bloch M 1949 *Apologie de l'histoire*. Armand Colin, Paris

Bourgeois E 1886 *L'histoire contemporaine et la science de l'histoire*. Armand Colin, Paris

Brozszat M, Fröhlich E, Wiesemann F (Vol. 1), Grossman A (Vols. 3–4), 1977–81 *Bayern in der NS Zeit*. Oldenbourg, Munich and Vienna

Conrad S 1999 *Auf der Suche nach der verlorenen Nation: Geschichtsschreibung in Westdeutschland und Japan, 1945–1960*. Vandenhoeck and Ruprecht, Göttingen, Germany

Croce B 1917 *Teoria e Storia della Storiografia*. Laterza, Bari, Italy

Enquete-Kommmisssion vom Deutschen Bundestag 1995 *Aufarbeitung von Geschichte und Folgen der SED-Diktatur in Deutschland, neun Bände in 18 Teilbänden*. Nomos Verlagsgesellschaft, Baden-Baden, and Suhrkamp Verlag, Frankfurt am Main, Germany

Fitzpatrick S, Gellately R 1997 *Accusatory Practices: Denunciation in Modern European History, 1789–1989*. University of Chicago Press, Chicago

Fogel J (ed.) 2000 *The Nanjing Massacre in History and Historiography*. University of California Press, Berkeley, CA

Getty J A, Manning R T (eds.) 1993 *Stalinist Terror: New Perspectives*. Cambridge University Press, Cambridge, UK

Jarausch K, Sabrow M (eds.) 1999 *Weg in den Untergang: Der innere Zerfall der DDR*. Vandenhoeck and Ruprecht, Göttingen, Germany

Kershaw I 1983 *Popular Opinion and Political Dissent in the Third Reich: Bavaria 1933–1945*. Clarendon Press, Oxford, UK

Maier C 1970 Revisionism and the interpretation of Cold War origins. *Perspectives in American History* **IV**: 313–47

Morris B 1987 *The Birth of the Palestinian Refugee Problem, 1947–1949*. Cambridge University Press, Cambridge, UK

Noiriel G 1998 *Qu'est-ce que l'histoire contemporaine?* Hachette, Paris

Rotberg R, Thompson D (eds.) 2000 *Truth vs. Justice: The Morality of Truth Commissions*. Princeton University Press, Princeton, NJ

Thiele H-G 1997 *Die Wehrmachtsausstellung: Dokumentation einer Kontroverse*. Zentrale für politische Bildung, Bonn, Germany

C. S. Maier

Content Analysis

Content analysis is a class of techniques for mapping symbolic data into a data matrix suitable for statistical analysis. When the term is used, one refers to the content analysis of cultural artifacts (e.g., books, architectural styles, discourse on prime-time television, etc.). That is, one refers to a mapping of non-numeric artifacts into a matrix of statistically manipulable symbols. Thus, content analysis involves measurement, not 'analysis' in the usual sense of the word. Whereas one might perform a content analysis of campaign slogans, one uses the data generated in such a content analysis to analyze (i.e., not 'content' analyze) the slogans' effects on public opinion. Texts and transcripts have been by far the most common objects of content analysis. For this reason, the focus of this article will be on the content analysis of texts (i.e., on *quantitative text analysis*). (For discussions of qualitative text analysis techniques, see *Linguistic Turn and Discourse Analysis in History*; *Event History Analysis: Applications*; *Ethnography*; *Qualitative Methods, History of*).

1. Classical Content Analysis

The first large-scale applications of content analysis came with the Second World War and US government commissioned propaganda analyses performed by Lasswell, Berelson, George, and other content analysis pioneers. During the first decade of the postwar era, a classical approach to content analysis emerged that involved the generation of data matrices of a particular type. These matrices typically exhibit a row for each sampled text block and a column for each theme (or meaning category) that may or may not have occurred in each block. Cells in the matrix are counts of the number of times the theme associated with its column occurred in the text block associated with its row. Thus, classical content analysis typically produces matrices of word- or phrase-counts, albeit sometimes ones embellished with value and salience weights (cf. Pool 1955, Stone et al. 1966, Holsti 1969, Smith 1992).

During a conference held in 1955 at the University of Illinois' Monticellos Allerton House many of these text-analysis pioneers gathered to develop solutions to the methodological problems of the day. Pool (1959) is the scholarly legacy of this conference. The most influential, if not the largest faction among the participants, was a group of Harvard researchers who made extensive use of what they called 'contingency analysis.' After generating a matrix of theme occurrences, these researchers would compute a matrix of associations between pairs of themes. Contingency analysis then proceeded as the researcher developed (usually *post hoc*) explanations of why some themes co-occurred (i.e., were positively associated) and why others were disassociated (i.e., were negatively associated).

In a casual but extremely insightful remark at the conference, Osgood (1959) noted, 'As a matter of fact, we may define a method of content analysis as allowing for "instrumental" analysis if it taps message evidence that is beyond the voluntary control of the source and hence yields valid inferences despite the strategies of the source.' And later, regarding contingency analysis, 'The final stage, in which the analyst interprets the contingency structure is entirely subjective, of course.'

In other words, classical content analysts treat words as symptomatic instruments from which the researcher can diagnose the source's possibly unconscious or unacknowledged characteristics.

2. Representational vs. Instrumental Interpretation

Shapiro (1997) has extended Osgood's meaning of 'instrumental analysis' by differentiating it from 'representational analysis' in which the researcher attempts 'to classify, tag, or retrieve the intended meanings of the authors.' At issue in this distinction is whether it is the source's or the researcher's perspective that is used to interpret the texts under analysis. When a researcher understands texts representationally, they are used to identify their sources' intended meanings. When a researcher understands texts instrumentally, they are interpreted in terms of the researcher's theory. Thus, Namenwirth exemplifies the latter approach in his argument that the sources of his texts 'are unfamiliar with many fundamental properties of their own culture and prove unable to specify its structural rules … To recover culture's properties and rules, we cannot ask culture's participants to answer these questions. Instead, we must rely on outsiders as investigators and use their methods, however unreliable these may prove to be' (Namenwirth and Weber 1987). Accordingly, instrumental text analysis methods are used to identify individual and societal characteristics about which society's members may be unaware; representational methods are used to characterize texts in ways that their sources intended them to be understood.

Osgood and Shapiro's representational/instrumental distinction helps direct attention to the import that researchers' analytic frameworks have for their findings. For example, consider a small subpopulation of male novelists who only write books with elderly female heroines. In seeking to understand their choices of leading characters, the researcher with a representational orientation might scour the novelists' prefaces, possibly discovering references there to a need for elderly female heroines as a corrective to an overuse of young male heros in contemporary literature. In contrast, a researcher with an instrumental, Freudian orientation, would likely find the choice of matronly heroines to be symptomatic of Oedipus complexes among the novelists. Note that in the former case the novelists' own meanings are assigned to their writings, whereas in the latter case the researcher's (Freudian) perspective is applied.

Methodological challenges related to representational text analysis are those associated with the researcher's ability to sympathetically understand the sources of his texts (see *Ethnography*; *Qualitative Methods, History of*). In both representational and instrumental text analyses the researcher is responsible for developing an explicit method for systematically applying the sources' or researcher's perspectives in interpreting the texts under analysis.

3. Data Matrices Produced in Content Analyses of Texts

Independent of whether it is representational or instrumental, a content analysis always involves the production of a data matrix. That is, it requires that words be mapped into a two-dimensional matrix representation suitable for statistical analysis. This (i.e., content analysis's 'data matrix requirement') greatly simplifies the task of delineating the domain of possible questions that quantitative text analyses are able to address. In particular, it allows this domain to be defined according to the various ways in which the columns (or variables) and rows (or units of analysis) of this data matrix can be defined. The balance of this article is devoted to further developing this observation, and consequently to helping the text analyst identify the most appropriate text analysis technique(s) for the research question at hand.

3.1 Variables

During the 1970s to the 1990s semantic and network text analysis methods have added to classical content analysis' word/phrase counts, other types of variables for the statistical analysis of linguistic data. Whereas in a thematic text analysis (of which classical content analysis is an instance) one examines occurrences of themes, in a semantic text analysis the examination is of sentences (or clauses) in which themes are interrelated. Moreover, in a network text analysis the examination is of themes' and/or sentences' locations within networks of interrelated themes. The three are not mutually exclusive, but may be combined in the same analysis.

The type of data matrix generated in a thematic text analysis has already been described in the first paragraph of Sect. 1. In brief, it is a matrix having one row for each randomly sampled block of text, and one column for each theme (or concept) that may occur in these text blocks. Cells in the data matrix indicate the number of occurrences of a particular theme within a specific block of text. As computer power has grown, and key-word-in-context searches have become easier, researchers have developed text analysis software with which *ad hoc* dictionaries (i.e., sets of theme categories) can be constructed interactively (Popping 1997). When the themes in these dictionaries are constructed to reflect the meanings intended by the texts' sources, consequent analyses are 'representational thematic text analyses.' When dictionaries' themes are constructed to reflect the researcher's perspective for interpreting the texts, consequent analyses are 'in-

Table 1
A data matrix from a semantic text analysis

ID-number	Subject	Action	Object
1	10	34	53
2	8	21	78
3	13	34	61
4	10	35	61
5	16	29	44

strumental thematic text analyses' (also known as classical content analyses). In all cases it is essentially a matrix of word-counts that forms the basis for a thematic analysis of texts.

Analyses of word-counts yield inferences about the predominance of themes in texts. For example, Namenwirth and Weber's (1987) cultural indicators research reports shifts in the prevalence of various political and economic themes over time. Yet note that if, e.g., certain types of 'political protest' are found to occur in texts in which one also finds mentions of 'economic inflation' one is unable (on the basis of a data matrix of word-counts) to determine if the protests are *mentioned in the texts as* being the cause or the effect of inflation. This is because information on semantic relations among themes such as 'political protest' and 'economic inflation' is not afforded by aggregated word-count data.

Relations among themes are encoded in a semantic text analysis, however. In a semantic text analysis the researcher begins by constructing a template (a semantic grammar). Themes from sampled texts are then mapped as syntactic components within this template. For example, Markoff et al. (1974) have developed a two-place semantic grammar for 'grievances' that contains one syntactic component for the object of the grievance (i.e., what is being grieved about) and another for the action that should be taken toward this grievance. The reader is referred to Franzosi (1990) and Roberts (1997) for more detailed descriptions of the application of semantic grammars in text analysis.

Table 1 illustrates a data matrix that might have been generated in a semantic text analysis. Note that the cells in the data matrix do not contain indicators of theme occurrences, but contain discrete codes for the themes themselves. The column in which a specific theme's code appears indicates the theme's syntactic role within the researcher's semantic grammar. In generating Table 1, blocks of text were encoded as sequences of subject–action–object triplets. Inferences from such a data matrix might be made within randomly sampled newspaper accounts of labor disputes, by comparing the odds that representatives of management vs. labor initiate collective bargaining. Within transcribed speech from a sample of minutes of prime-time television content, inferences might be drawn regarding the odds that blacks vs. whites refer to themselves as targets of aggression. More generally,

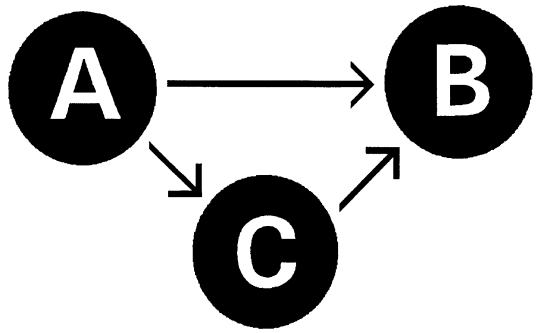

Figure 1
A network of causal relations among themes.

and in contrast to thematic text analysis, semantic text analyses yield information on how themes are related according to an *a priori* specified semantic grammar.

Writings on semantic text analysis methods date back to Gottschalk's instrumental analyses in the 1950s on psychological states and traits (e.g., Gottschalk and Kaplan 1958). Gottschalk has since developed highly automated, parser-based text-encoding software for measuring (according to his perspective on states that are reflected in how people relate words) such psychological states as hostility, depression, and hope (Gottschalk and Bechtel 1989). The Kansas Events Data System (KEDS) is another special-purpose, parser-based program that can be used either instrumentally (in conjunction with the standard World Events Interaction Survey (WEIS) coding scheme) or representationally (based on user input content categories gleaned from one's data) in analyzing event data from news reports (Gerner et al. 1994). Most other uses of semantic text analysis are representational, however. For examples, see Franzosi's (1997) research on labor disputes, Roberts's (1989) on ideology shifts, and Shapiro and Markoff's (1998) work on public opinion in eighteenth century France.

Network text analysis originated with the observation that once one has a series of encoded statements, one can proceed to combine these statements into a network. Moreover, once text blocks are rendered as networks of interrelated themes, variables can be generated to measure the 'positions' of themes and theme-relations within the networks. For example, let us imagine that we construct a network of themes in which all linkages indicate causal relations. Assigning the names theme-A and theme-B to any pair of themes in the network, one could develop a measure of 'the causal-salience of theme-A on theme-B' as the proportion of all sequences of causal linkages that are ones in which theme-A is the cause and theme-B is the effect.

For example, the simple three-theme network in Fig. 1 contains four sequences of causal linkages, namely $A \rightarrow B$, $A \rightarrow C$, $C \rightarrow B$, and $A \rightarrow C \rightarrow B$.

Table 2
A data matrix from a network text analysis

ID-number	Causal salience measures					
	A on B	A on C	C on B	B on A	C on A	B on C
1	.50	.25	.25	.00	.00	.00
2	.25	.00	.50	.00	.25	.00
3	.00	.00	.25	.25	.50	.00
4	.00	.00	.00	.50	.25	.25
5	.00	.25	.00	.50	.00	.25
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Accordingly, the 'causal salience' of theme-A on theme-B equals 0.5, which is the proportion of the four sequences in which A is the cause and B is the effect. Note how calculation of this measure draws on more than isolated semantically-linked themes in a corresponding block of text. It incorporates information on all themes and links within the text-block's network representation.

Thus, a data matrix such as that in Table 2 might be generated from a sample of networks that contained variables measuring the causal-salience of each pair of the texts' themes. Network text analysts have developed many other measures of network characteristics. A theme's 'conductivity' is one example, referring to the number of linkages that the theme provides between pairs of other themes (Carley 1997). Another is of theme linkages that are logically implied, but not explicitly stated, in each block of text (Kleinnijenhuis et al. 1997). In addition to pioneering work in this area by Osgood et al. (1956), nearly all quantitative network text analysis research has been conducted either by Carley or by a group of Dutch researchers in Amsterdam (Carley 1986, van Cuilenburg 1986, 1988, Carley and Palmquist 1992). All of these researchers do representational network text analysis. The dawn of instrumental network text analysis awaits a social scientist who both develops a network representation of language (analogous to a semantic grammar) and uses an interpretive perspective in developing rules for 'filling in' the network's arcs and nodes given various combinations of sentences in text blocks.

3.2 Units of Analysis

The discussion thus far has yielded a 2×3 taxonomy of quantitative text analysis methods. These methods are distinguished on the first dimension according to whether the source's or the researcher's perspective is the basis for interpreting texts, and on the second dimension according to whether they use variables that reflect occurrences of themes, themes in semantic roles, or network-positions of themes or theme-relations. Let us now proceed to specify the types of units of analysis that are possible in a content analysis of texts.

In the earliest stages of every quantitative text analysis, the researcher is confronted with a 'mountain of words' (a text-population) about which statistical inferences are to be drawn. On the one hand, this text population may be an initially undifferentiated mass (e.g., a sample of minutes of speech on US prime-time television could be drawn at random from the undifferentiated 'mountain of words' that were uttered during a year's time). On the other hand, the text population may consist of clusters of sentences such as newspaper editorials, transcripts of interviews, diary entries, and so on. In either case, a representative sample can only be drawn once the text population has been divided into distinct text-blocks, each of which is then assigned a unique number and sampled at random.

Yet the text population should not be mindlessly divided, even if it appears to the researcher as already a collection of discrete text-blocks such as editorials, interviews, or diary entries. This is because the statistical inferences that the researcher may legitimately draw from texts depend fundamentally on the units into which the text population is divided initially.

Imagine a researcher, who wishes to analyze prime-time television data according to performers' mental models (or conceptual frameworks). In such a case, the researcher would begin by dividing the text population into blocks associated with each *performer*. On the other hand, if the researcher wished to analyze the narratives (or story lines) depicted on prime-time television programs, the researcher would begin by dividing the text population into blocks associated with each *program*. Yet performers may appear in more than one program, and programs will involve many performers. As a result, the researcher's inferences will differ, depending on this initial division of the text population.

In short, the researcher who wishes to ask a substantive question of a population of texts must not only consider the thematic, semantic, and network variables required to address the question, but also the units of analysis yielded when this population is divided into text-blocks.

Thus far two types of units of analysis have been mentioned that might be yielded when a text population is divided; namely the conceptual framework of the text's source (e.g., the television performer's mental model), and the message that the text conveys (e.g., the program's story line). But of course, there are others.

Consider Lasswell's (1948) oft-cited depiction of communications research as the study of 'Who says what, in which channel, to whom, and with what effect?' In his article-long answer to this question, Lasswell argued that communications' effects can be largely understood as functions of their source, message, channel, and audience. Moreover, these four

aspects of communication are the most common *contextual variables* used in analyses of texts and transcripts:

(a) characteristics of source (gender, affiliation, biases, etc.),

(b) characteristics of message (local vs. domestic news, descriptive vs. evaluative orientation, etc.),

(c) characteristics of channel (radio vs. TV news, public vs. commercial network, written vs. spoken medium, etc.), and

(d) characteristics of audience (sociocultural and/or historical setting within which text appeared).

Thus, other than source- and message-identifications, the researcher may also identify text-blocks according to their channel and/or intended audience.

However, comparisons among texts' sources, messages, channels, and audiences are only possible if each text-block under analysis can be clearly identified according to its type of source, message, channel, and/or audience. The key in selecting a unit of analysis is not to assume that one's population of text is comprised *a priori* of clearly distinguishable text-blocks. On the contrary, it is the researcher's responsibility to divide this population into blocks—blocks that can be uniquely identified according to the contextual variables required for addressing the research question at hand.

4. *What Questions Can Content Analyses Answer About Text Populations?*

Content analyses of texts yield data matrices with text-related variables and, almost surely, contextual variables. Text-related variables in the matrix may measure occurrences of themes, theme-relations within a semantic grammar, and/or network-positions of themes and theme-relations. Contextual variables may indicate the source, message, channel, and/or audience uniquely associated with each text-block under analysis. Accordingly, content analyses of texts afford answers to questions about 'what themes occur,' 'what semantic relations exist among the occurring themes,' and 'what network positions are occupied by such themes or theme relations' among texts with particular types of source, message, channel, or audience. Within these limits, the decision of which question to ask is, of course, where the researcher's own imagination takes hold.

See also: Interpretive Methods: Macromethods; Interpretive Methods: Micromethods

Bibliography

Carley K 1986 An approach for relating social structure to cognitive structure. *Journal of Mathematical Sociology* **12**: 137–89

Carley K M 1997 Network text analysis: The network position of concepts. In: Roberts C W (ed.) *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Lawrence Erlbaum, Mahwah, NJ, pp. 79–100

Carley K, Palmquist M 1992 Extracting, representing, and analyzing mental models. *Social Forces* **70**: 601–36

van Cuilenburg J J, Kleinnijenhuis J, de Ridder J A 1986 A theory of evaluative discourse: Towards a graph theory of journalistic texts. *European Journal of Communication* **1**: 65–96

van Cuilenburg J J, Kleinnijenhuis J, de Ridder J A 1988 Artificial intelligence and content analysis: problems of and strategies for computer text analysis. *Quality and Quantity* **22**: 65–97

Franzosi R 1990 Computer-assisted coding of textual data: An application to semantic grammars. *Sociological Methods and Research* **19**: 225–57

Franzosi R 1997 Labor unrest in the Italian service sector: An application of semantic grammars. In: Roberts C W (ed.) *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Lawrence Erlbaum, Mahwah, NJ, pp. 131–45

Gerner D J, Schrodt P A, Francisco R, Franciso R A, Weddle J L 1994 Machine coding of events using regional and international sources. *International Studies Quarterly* **38**: 91–119

Gottschalk L A, Bechtel R 1989 Artificial intelligence and the computerization of the content analysis of natural language. *Artificial Intelligence in Medicine* **1**: 131–37

Gottschalk L A, Kaplan S M 1958 A quantitative method of estimating variations in intensity of a psychologic conflict or state. *Archives of Neurology and Psychiatry* **79**: 688–96

Holsti O R 1969 *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley, Reading, MA

Kleinnijenhuis J, de Ridder J A, Rietberg E M 1997 Reasoning in economic discourse: An application of the network approach to the Dutch press. In: Roberts C W (ed.) *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Erlbaum, Mahwah, NJ, pp. 191–207

Lasswell H D 1948 The structure and function of communication in society. In: Bryson L (ed.) *The Communication of Ideas*. Institute for Religious and Social Studies, New York, pp. 37–51

Markoff J, Shapiro G, Weitman S 1974 Toward the integration of content analysis and general methodology. In: Heise D R (ed.) *Sociological Methodology, 1975*. Jossey-Bass, San Francisco, pp. 1–58

Namenwirth J Z, Weber R P 1987 *Dynamics of Culture*. Allen & Unwin, Boston

Osgood C E 1959 The representational model and relevant research methods. In: Pool I de S (ed.) *Trends in Content Analysis*. University of Illinois Press, Urbana, IL, pp. 33–88

Osgood C E, Saporta S, Nunnally J C 1956 Evaluative assertion analysis. *Litera* **3**: 47–102

Pool I de S 1955 *The Prestige Press: A Comparative Study of Political Symbols*. MIT Press, Cambridge, MA

Pool I de S (ed.) 1959 *Trends in Content Analysis*. University of Illinois Press, Urbana, IL

Popping R 1997 Computer programs for the analysis of texts and transcripts. In: Roberts C W (ed.) *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Erlbaum, Mahwah, NJ, pp. 209–21

Roberts C W 1989 Other than counting words: A linguistic approach to content analysis. *Social Forces* **68**: 147–77

Roberts C W 1997 A generic semantic grammar for quantitative text analysis: Applications to East and West Berlin radio news content from 1979. *Sociological Methodology* **27**: 89–129

Shapiro G 1997 The future of coders: Human judgments in a world of sophisticated software. In: Roberts C W (ed.) *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Erlbaum, Mahwah, NJ, pp. 225–38

Shapiro G, Markoff J 1998 *Revolutionary Demands: A Content Analysis of the Cahiers de Doléances of 1789*. Stanford University Press, Stanford, CA

Smith C P (ed.) 1992 *Motivation and Personality: Handbook of Thematic Content Analysis*. Cambridge University Press, Cambridge, UK

Stone P J, Dunphy D C, Smith M S, Ogilvie D M 1966 *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA

C. W. Roberts

Content Validity

Validation is a process of disciplined inquiry. In the social sciences, it entails the systematic assembly of logical and empirical evidence to support the intended use of the test scores. Content validation is the process of obtaining such evidence in the form of expert judges' evaluation of test content in relation to a defined domain of knowledge or universe of behavior. Suppose an examinee takes a test of foreign language proficiency, and must translate sentences and identify the meaning of vocabulary words. The test user wants to draw inference from performance on the test items at hand to a broader universe of behavior represented by these tasks. Practical constraints of testing time, examinee fatigue, examinee motivation, and cost of item development limit the number of items on the test. Thus these items and response formats must represent a hypothetical pool of tasks that could have been presented covering the same content. Validity arguments for this inference hinge largely on whether a clear domain of language proficiency was defined, and how adequately this domain was sampled by the exercises.

Using expert judges to review the domain description and evaluate specific features of the items in relation to this domain, as well as collecting and summarizing their judgments, is known as content validation. Content validation evidence is useful in (a) assessment of educational achievement; (b) assessment of job knowledge or skills by employers; and (c) professional certification and licensure testing. It may also play a role in establishing construct validity of instruments for assessing affective traits.

1. *History and Terminology*

Attempts to define content validity were prominent in the measurement literature in the mid-twentieth century (Ebel 1956, Gulliksen 1950, Lennon 1956). From 1950s–1980s measurement experts (e.g., Cronbach 1971) identified three distinct approaches to validation: (a) content validation; (b) criterion-related validation, both predictive and concurrent; and (c) construct validation. Throughout this period, however, a few theorists continued to challenge whether content judgment studies fit within the definition of validation (e.g., Fitzpatrick 1983, Messick 1975), but this view was not dominant.

Originally, content validation was applied primarily in educational achievement testing. Clearly, the usefulness of test scores as indicators of student learning depended upon the linkage between items presented on the test and the body of knowledge which the items sampled. The 1970s witnessed a growing trend for personnel psychologists to employ content validation to justify use of test results in personnel selection and advancement decisions, despite some notable reservations (Ebel 1977, Fitzpatrick 1983, Guion 1977). This application grew in inverse proportion to waning confidence in the adequacy and legal defensibility of other validation procedures for many jobs and employment settings. Licensing and certification testing also relied on content validation throughout this period (Shimburg 1982).

Approaching the twenty-first century, a more unitary view of validity prevailed (Messick 1989). One theme of this perspective is that proficiency in an academic subject or job performance domain should be viewed as a construct. Although not new, the idea gained mainstream acceptance. From this viewpoint, all validation efforts are designed to provide evidence of construct validity, but there are multiple aspects to the process of construct validation (and multiple types of evidence). Content is consistently the first aspect and involves determination of the construct relevance, representativeness, and technical quality of the test using expert judgment (Messick 1995).

Other terms are associated with content validity, but not synonymous. 'Face validity' is the extent to which items appear relevant to a trait in the eyes of laymen or examinees who are not necessarily experts in the subject tested (Mosier 1947). 'Instructional validity' and 'curricular validity' were spawned in the mid-1970s during landmark litigation over the use of high-school competency examinations in Florida (see Yalow and Popham 1983). Curricular validity refers to congruence between item content on an achievement and curricular objectives and textual materials. Instructional validity refers to congruence between item content and instruction from teachers.

2. *Design of a Content Judgment Study*

Content judgment studies typically are undertaken by test developers as part of their initial validation effort for a new test of knowledge or proficiency. Test users may also conduct local content validation prior to the

decision to adopt an existing, published test. These studies may require replication over time to demonstrate that changes in curricula or job requirements have not undermined the meaning of the test scores. A content judgment study involves systematic collection of observations. Steps in the process may include:

(a) Specifying the questions to be addressed.
(b) Defining the domain of interest.
(c) Identifying the features of the exercises to be examined in the judgment process.
(d) Developing a structured framework for the review process.
(e) Selecting qualified experts for that domain.
(f) Instructing the experts in the review process.
(g) Collecting and summarizing the data.

2.1 Domain Definition

With achievement tests, domain definition typically is extracted from instructional objectives commonly used by those who teach the subjects; review of curricular materials and published textbooks also may play a central role in domain definition. Interviews or surveys of instructors have also been used in this process (see Yalow and Popham 1983). Test specifications (blueprints), used by test developers to maintain the desired proportional coverage of various content areas and cognitive performance levels, may also be used in domain definition. In personnel or licensure testing, domains, or required or prerequisite knowledge may be defined through a detailed job description provided by experts, or a job analysis based on detailed observations of individuals performing the job (Kane 1997); however, in licensure testing the domain is restricted typically to skills and knowledge that are essential to public safety (Smith and Hambleton 1990). A 'critical incidents' technique may also be employed to capture important knowledge or skills that spell the difference between successful and failing job performance. Thorough domain definition also includes the format(s) in which individuals may be expected to display their knowledge: i.e., it covers both a stimulus and response domain (Guion 1977).

2.2 The Judgment Process

The judgment process focuses upon at least three features of the test items: (a) relevance of item content to the domain; (b) balance of coverage or the extent to which the collection of items adequately sample the content domain; and (c) technical quality of the items, response formats and scoring procedures (Messick 1989, 1995).

Relevance of item content is assessed by evaluating the importance of the knowledge or skills required to answer the test item in relation to successful performance in the domain. Judges may be asked to rate frequency and criticality of the skill or knowledge tapped by the exercise. When the domain is complex, the study may involve having a sample of experts (a) evaluate the appropriateness of the objectives or list of behaviors that comprise the domain; and (b) match exercises to those objectives.

Representativeness (balance of coverage) is ascertained by having the judges review the entire test to identify areas of the domain that may be over- or under-represented. It is possible for all items on a test to be relevant when reviewed individually, but yet for the coverage of the domain to be disproportionately limited to only a few objectives or skills. A mathematics test, for example, might stress computational skills, but place little emphasis on problem-solving, even though both skills are stressed in the curriculum. The proportions of items matched to each objective or job behavior are examined to assess appropriateness of coverage. The format for examinee response to exercises may also be considered in evaluating representativeness. In judging technical quality of exercises, experts attend to features of the question, response format, scoring rubrics or keys, and administration and scoring instructions that might interfere with examinees' ability to display their knowledge of the correct response.

2.3 Selecting and Training Judges

Judges' qualifications and the overall representative nature of the panel of judges in terms of demographic characteristics and educational/professional backgrounds are important issues in the design of a credible content validation study. The instructions given to the judges and training they receive before undertaking their review should also be planned and documented. Finally, in large-scale content validations, level of agreement across the panel is useful information.

3. Analyzing Data from Content Judgment Studies

Thorough content validation studies yield substantial amounts of data. Each item should have a record containing multiple judges' ratings for each feature of the item that was reviewed. Marshalling these data for analyses to summarize results at the individual item level and for the total test requires forethought. At the simplest level, indices such as percentage of items classified to objectives may be computed and reported. Judges' mean ratings for each item and over the test as a whole on frequency or criticality may be computed. More sophisticated treatments of data from content validation studies also have been suggested. Crocker et al. (1987) described various quantitative indices used in content judgment studies, including measures of (a) overall fit between test and curriculum; (b)

degree of fit for individual items to content domain; (c) fit between test specifications and examinee performance; and (d) error estimation in judges' ratings. In addition, new technical approaches to analyzing job analysis data are emerging (e.g., Kane 1997, Šireci and Geisinger 1995).

3.1 Evolving Standards for Content Validation

Standards for evaluating test development and use have been articulated since the 1950s by a series of joint committees of the American Association of Educational Research (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). These standards have received global recognition. In the first published committee report, each recommendation was labeled as ESSENTIAL, VERY DESIRABLE, or DESIRABLE (APA 1954, p. 20). The primary recommendation for content validation was:

> If a test performance is to be interpreted as a sample of performance in some university of situations, the manual should indicate clearly what universe is represented and how adequate the sampling is. ESSENTIAL.

Additional recommendations included that the descriptions be provided for sources of items, criteria for inclusion or exclusion of items, and method of sampling items from the domain. A decade later, the *Standards* (APA 1966) specified that content validity information should include:

1. The universe that the test items represented (ESSENTIAL).
2. The adequacy of the procedure for selecting a sample of items from the universe (ESSENTIAL).
3. The dates of publication of any textbooks from which items topics were selected (ESSENTIAL).
4. The professional qualifications of the expert judges (VERY DESIRABLE).
5. The directions that they received (VERY DESIRABLE).
6. The extent of agreement among judges (DESIRABLE).
7. Any classification or taxonomy used in item selection (DESIRABLE).

The next edition of the joint standards (APA 1974) generally upheld the previous standards for content validity but recognized increasing applications of content validation in personnel selection:

> When a test is represented as having content validity for a job or class of jobs, the evidence of validity should include a complete description of job duties including relative frequency, importance, and skill level of such duties. ESSENTIAL. (p. 46)

In 1985, the joint *Standards* downplayed the notion of 'content validity' as a separate test quality; but recognized the uniqueness of its form of evidence (AERA et al. 1985). In addition, these standards covered domain definition and content validation for licensure testing. The final edition published in the twentieth century (AERA et al. 1999) includes two standards for validity evidence from content judgment studies. First,

> When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should be clearly explained and justified. (p. 18)

Second,

> When a validation rests in part on the opinions or decision of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of the procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth. (p. 19)

Shifts in wording and focus of criteria for evaluating content judgment studies reflect trends in the measurement community's view of validity *per se*. Yet, given advances in the theory, quantitative models, and technology of modern psychometrics, it is remarkable that the underlying logic, methods, and criteria for content judgment studies have survived with so little change. Evidently the concept possesses enduring utility for a broad range of test developers and test users.

See also: Curriculum Evaluation; Educational Evaluation: Overview

Bibliography

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) 1985 *Standards for Educational and Psychological Testing*. American Psychological Association, Washington, DC

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) 1999 *Standards for Educational and Psychological Testing*. American Psychological Association, Washington, DC

American Psychological Association (APA) 1954 Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin* **5**(2), supplement

American Psychological Association (APA) 1966 *Standards for Educational and Psychological Tests and Manuals*. American Psychological Association, Washington, DC

American Psychological Association (APA), American Educational Research Association (AERA) and National Council on Measurement in Education (NCME) 1974 *Standards for Educational and Psychological Tests*. American Psychological Association, Washington, DC
Crocker L M, Miller D, Franks E A 1989 Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education* **2**: 179–94
Cronbach L J 1971 Test validation. In: Thorndike R L (ed.) *Educational Measurement*, 2nd edn. American Council on Education, Washington, DC, pp. 443–507
Ebel R L 1956 Obtaining and reporting evidence for content validity. *Educational and Psychological Measurement* **16**: 269–82
Ebel R L 1977 Comments on some problems of employment testing. *Personnel Psychology* **30**: 55–63
Fitzpatrick A R 1983 The meaning of content validity. *Applied Psychological Measurement* **7**: 3–13
Guion R M 1977 Content validity: The source of my discontent. *Applied Psychological Measurement* **1**: 1–10
Gulliksen H 1950 Intrinsic validity. *American Psychologist* **5**: 511–17
Kane M 1997 Model-based practice analysis and test specifications. *Applied Measurement in Education* **10**: 5–18
Lennon R T 1956 Assumptions underlying the use of content validity. *Educational Psychological Measurement* **16**: 294–304
Messick S 1975 The standard problem: Meaning and values in measurement and evaluation. *American Psychologist* **30**: 955–66
Messick S 1989 Validity. In: Linn R L (ed.) *Educational Measurement*, 3rd edn. American Council on Education, Washington, DC, pp. 13–103
Messick S 1994 The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* **23**(2): 13–23
Messick S 1995 Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice* **14**(5): 5–8
Mosier C I 1947 A critical examination of the concept of face validity. *Educational and Psychological Measurement* **7**: 191–206
Shimburg B 1982 Licensing and certification. *Encyclopedia of Educational Research*, 5th edn. American Educational Research Association, Washington, DC, pp. 1084–92
Sireci S G, Geisinger K F 1995 Using subject matter experts to assess content representation: A MDS analysis. *Applied Psychological Measurement* **19**: 241–55
Smith I L, Hambleton R K 1990 Content validity studies of licensing examinations. *Educational Measurement: Issues & Practice* **9**(4): 11–14
Yalow E S, Popham W J 1983 Content validity at the crossroads. *Educational Researcher* **12**: 10–14

L. Crocker

Contextual Studies: Methodology

Within the social sciences, 'context' refers to the physical, social, cultural, and temporal environment in which behavior occurs. As a term, context is used rather loosely, and can refer to an interpersonal relationship (e.g., children's emotional expression in the *context* of the parent–child relationship), a social group (e.g., interpersonal conflict in the *context* of a peer group), a physical setting (e.g., sex role behavior in the *context* of the classroom), a locale (e.g., crime in the *context* of neighborhood poverty), or a broad expression of time or culture (e.g., the role of religion in the *context* of twentieth century America). The study of context in the social sciences, once the province of disciplines such as sociology and anthropology, has now become an accepted focus of inquiry within perspectives that formerly had been concerned exclusively with the study of individuals in controlled, laboratory settings, such as psychology. This article examines the ways in which context is taken into account in contemporary social science research.

1. Historical Underpinnings

The contemporary study of context has its origins in a number of different disciplines, not only sociology and anthropology, but also *ethology* (the study of animal behavior in natural habitats), *ecology* (the study of biological organisms and their environment), and urban geography (the study of the development and functioning of neighborhoods in cities). All of these disciplines emphasize the fact that any human or animal behavior cannot be understood outside of the setting in which it occurs. In this sense, the organism and context are seen as inextricably linked.

Psychology, as a discipline, was slow to embrace a contextual point of view. Although numerous writers had attempted to move the field in this direction—most notably, Kurt Lewin (1951)—the study of psychology was dominated for much of the twentieth century by experimental scientists, who studied human behavior in the laboratory. Context was seen as a nuisance variable—something that got in the way of pure science—and the goal of experimental research was to study behavior under conditions that permitted the investigator to control any variable that was seen as extraneous to the research focus. If one wanted to study, for example, how individuals make decisions, the researcher would need to control the setting in which the decision-making took place, so that each subject in the study could be observed under identical circumstances. The simplest way to do this was to perform research in a laboratory, where each and every element of the context could be controlled.

The benefit of controlled laboratory research is that it permits the scientist to isolate pure behavioral phenomena, uncontaminated by the outside environment. But this benefit must be weighed against the fact that phenomena observed under controlled experimental conditions may not be the same as phenomena observed in their natural context. The way in which individuals make decisions in a psychology laboratory, for instance, may not resemble the way they

make decisions at home, at work, or in a store. Just as ethologists would not want to generalize observations of caged animals to the behavior of that same species in the wild, social scientists must be careful about generalizing findings derived from controlled experiments to human behavior outside the laboratory. This is not to question the importance of laboratory-based research but, rather, to note that the laboratory is itself a context, and that features of this context, like any other context, necessarily influence the behavior that occurs within it.

During the 1970s, concerns about ecological validity—the extent to which an observed phenomenon in a research investigation is a valid reflection of that phenomenon as it occurs in natural contexts—were voiced by numerous investigators working in a variety of areas. In the study of intelligence, for example, scholars raised questions about whether standardized intelligence tests adequately captured intelligent behavior as it occurs in the real world. Several researchers found that individuals who performed poorly on standardized tests of mathematics frequently demonstrated the ability to use very sophisticated mathematical reasoning when observed during day-to-day situations, such as shopping. Similarly, scholars of social behavior discovered that individuals' behavior was considerably influenced by the context in which it took place, and that conclusions drawn from studies of behavior in one context (e.g., observations of young children at home) could not always be extended to other settings (e.g., observations of these same youngsters in the preschool classroom). Ecological psychologists drew attention to the fact that individual behavior was profoundly shaped by the structure of the physical environment in which people lived. Several classic studies of schools indicated, for example, that students' behavior in small schools was markedly different from students' behavior in large schools (Barker and Gump 1964).

In the late 1970s, psychologist Urie Bronfenbrenner drew widespread attention to the role of context in human development and behavior by noting that the scientific literature on child development was based almost entirely on the study of children interacting with strangers in strange situations for the briefest amounts of time. How, Bronfenbrenner wondered, could social scientists hope to contribute to the solution of social problems if they insisted on studying human behavior in a contextual vacuum? Although his comments were aimed at students of child development, they applied equally well to other disciplines. In his book, *The Ecology of Human Development*, Bronfenbrenner (1979) called on the field to think in more sophisticated fashion about the settings in which development took place, and he outlined a model of context that researchers could use to guide their thinking and empirical investigation. This work, and the thinking it would subsequently inspire, led to the development of the ecological perspective on human development, a perspective that grew more and more influential over the next several decades (Bronfenbrenner and Crouter 1983).

2. *The Ecological Perspective*

Within the ecological perspective, the context is construed as a series of nested environments, with each level of contextual influence embedded within a larger level, not unlike a set of Russian dolls. At the innermost, or most proximal level of contextual influence, are the immediate settings in which individuals spend time—the family, the friendship group, the classroom, the workplace, the laboratory, and so forth. Most studies of the role of context in shaping human behavior and development are studies of single contexts examined in isolation from one another (e.g., the impact of the family context on child development, the impact of teaching practices on student achievement).

The ecological perspective provides a conceptual framework within which to investigate more complex interactions between persons and environments, however. Three notions are especially important in this regard. First, proximal settings are seen as connected to each other, because events that take place in one setting often have ramifications for individual behavior and development in another. Socialization in the family context influences how adolescents behave in the peer group, peer influence processes affect how adolescents behave in the classroom, experiences in the classroom affect students' performance in extracurricular settings, and so on. Within the ecological perspective, therefore, an important focus of inquiry is on the ways in which, and the extent to which, different settings are linked, and the impact of these links on individual behavior.

Second, within the ecological perspective, these proximal settings are seen as contained within broader institutional and community contexts that shape the structure of settings and influence what takes place within them (Duncan and Raudenbush 1998, Jencks and Mayer 1990). Families and friendship networks reside within neighborhoods, for example, and qualities of neighborhoods influence how families and friendship networks function. Parents interact with their children differently in poor neighborhoods than they do in more affluent ones, for example, and, as a consequence, child development may vary across neighborhood contexts. Similarly, schools exist within communities, and characteristics of communities—urban vs. rural, tightly knit vs. impersonal—necessarily influence the ways in which schools are organized and operate. Within the ecological perspective, therefore, one needs to examine relations among nested contexts as well as between individuals and settings.

Finally, proximal settings and the broader environments that contain them are located within particular historical, social, economic, political, geographical, and cultural contexts—'macro-systems'—that influence the nature, structure, function, organization, and influence of all levels of the environment. Schools in Japan function differently than those in Canada, and differences between Japanese and Canadian schools have implications for the ways that teachers and students interact with each other, which in turn have implications for understanding student achievement in each context. A city's neighborhoods today may bear little resemblance to these very same neighborhoods 50 years ago, and the ways in which neighborhoods have changed over historical time may have important implications for understanding transformations in family life and parent-child relationships. Important social and economic events—wars, economic depressions, natural disasters—shape proximal and distal settings in important ways that have ramifications for individuals' behavior and development. Within the ecological perspective, researchers are interested in the influence of macro-systems on the settings located within them, and, consequently, in the rippling effects of these influences on individual behavior and development.

3. *Methodological Challenges in Contextual Research*

As more and more social scientists began to incorporate an increasingly sophisticated model of context into their research designs, it became clear that the scientific study of context presented a number of new and formidable methodological challenges. The discussion shifted from *whether* one needed to study context (the consensus was that one did) to how best to do so. Among the most important of these concerns involved deciding upon the most appropriate way to assess context, developing models of the transactional nature of contextual influence, and disentangling the effects of hierarchically organized, or nested, environments.

3.1 *Objective vs. Subject Assessment of Context*

One of the longstanding debates in the study of context concerned the relative merits of objective vs. subjective assessments of the environment. Objective assessments attempt to separate the measurement of context from the assessment of the individuals whose behavior is the target of investigation. A researcher interested in the influence of the parent–child relationship on child development might conduct impartial observations of family interaction in the home and draw his or her own conclusions about the relationship on the basis of what was observed. Similarly, a student of neighborhood stress and its impact on mental health might draw upon official statistics on crime, poverty, unemployment, and housing density and look at these in relation to residents' mental health problems.

Subjective assessments of context, in contrast, rely on the research participants' perceptions of their environment. These perceptions can be assessed by asking individuals to complete questionnaires or by interviewing them about the settings in which they live. Thus, a researcher interested in parent–child relationships might interview parents about their socialization practices or ask children to complete questionnaires about their parents' behavior. Similarly, someone interested in neighborhood stress and mental health might collect information from neighborhood residents on their perceptions of the community's dangerousness, economic health, or housing quality and examine the links between these perceptions and one or more indices of mental health.

These different approaches to the assessment of context have particular strengths and weaknesses. The advantage of using objective assessments inheres in the fact that the assessment of context is not contaminated by the characteristics of the individuals whose behavior is being studied. Suppose, for instance, a researcher interested in neighborhood stress and mental health gathered information about each by interviewing a sample of residents and found that depression was more prevalent in dangerous areas. Because depressed individuals may describe their neighborhoods in more negative terms than nondepressed individuals, however, it would be difficult to know whether the observed link between dangerousness and depression was genuine or merely an artifact of biased reporting. If, however, the relationship between dangerousness and depression were confirmed in a study using official crime statistics, one would have more confidence in the validity of the finding.

The advantage of using subjective assessments of context is that such assessments capture the context as it is experienced by the participant. Some social scientists argue that it is the subjectively experienced context, and not the objective context, that has the most profound impact on behavior and development. Individuals' perceptions of their neighborhood as safe vs. dangerous, for example, may well influence how they behave far more than the actual level of crime as measured by official statistics. As the sociologist W. I. Thomas wrote early in the twentieth century, if individuals 'perceive situations as real, they are real in their consequences' (Thomas and Thomas 1928). Of course, one problem that arises when researchers rely on subjective assessments of context concerns the reconciliation of different subjects' perceptions of the same context. In family research, for example, it is common to find that different family members characterize their family context in different ways.

Ultimately, the most thorough assessment of context is one that employs multiple sources of infor-

mation and uses multiple methods of assessment. By combining information from objective and subjective sources, from multiple vantage points, and gained through observational, self-report, and archival data, researchers are often able to provide richer and more accurate characterizations of a setting than are achieved through one approach alone.

3.2 The Transactional Nature of Contextual Influence

A second methodological challenge involves capturing the transactional nature of contextual influence. Although researchers often claim to be studying the influence of context on the individual, in reality the situation is far more complicated than this, in at least three important ways. First, individuals are not the passive recipients of contextual influence, but are active agents who shape the contexts in which they live. At the same time, contexts are not static entities, but settings that change and develop over time. Thus, while children are no doubt influenced by their family environment, family environments are themselves influenced by the children who inhabit them. Similarly, friendship groups both affect, and are affected by, their members. Finding ways to model the reciprocal and dynamic relations between individuals and their environments has been an especially difficult challenge for researchers.

Second, it is often difficult to separate the effects of contextual selection from those of contextual influence (Jencks and Mayer 1990). In laboratory experiments individuals can be randomly assigned by the researcher to different groups and then compared to see whether group membership affects individual behavior. Outside the laboratory, however, people often pick the contexts in which they spend time. Consider, for example, the often reported finding that delinquent teenagers often are members of peer groups composed of other delinquent adolescents. Is this because delinquent peers influence their friends to engage in antisocial behavior, or is it due, instead, to the propensity of delinquent individuals to choose to affiliate with others who share similar attitudes and values? Obviously, it is not possible to answer this question without randomly assigning individuals to different peer groups (not a feasible option in most research) or observing individuals and contexts over time. Similar problems exist in attempts to document the influence of neighborhoods on their residents. Because individuals have some control over where they live, it is difficult to disentangle the effects that neighborhoods have on individuals from the factors that lead some people to select particular neighborhoods in which to settle.

Third, different individuals are often affected by the same contextual factors in very different ways, making it extremely difficult to generalize about the effects of various contextual factors on individual behavior and development. A highly active and impulsive child may fare poorly in a relatively unstructured classroom, whereas this same context may be perfectly suitable for a more passive, more placid youngster. Similarly, an employee who is self-reliant and self-directed may thrive in a work environment that permits a great deal of autonomy, but this same work environment may prove problematic for an employee who needs a great deal of guidance and supervision. The recognition that contextual influence varies as a function of differences between individuals has stimulated a good deal of interest in the study of 'person-environment fit'—the extent to which a given context meshes well with the individual's personal characteristics.

3.3 Hierarchical Nature of Contextual Influence

The third methodological challenge concerns the hierarchical and interactive nature of contextual influence. The recognition that individuals are affected by multiple contexts simultaneously has necessitated the development of more sophisticated conceptual and statistical models in order to best describe the complicated ways through which individuals are influenced by the contexts in which they live (Duncan and Raudenbush 1998). Some contexts, for example, affect individuals directly, as when the behavior of a parent directly affects the development of a child. Other contextual effects occur through their impact on some other setting, however; stress in the parent's workplace may affect the child indirectly, through its impact on the family. In this instance, the second setting (in this case, the family) is said to 'mediate' the impact of the first setting (i.e., the workplace). And still other contextual effects occur by transforming, or 'moderating,' the impact of other contexts. Neighborhoods, for example, often exert their strongest impact on individuals not directly, but by attenuating or enhancing the impact of other contexts, such as the family or school. Within the context of an affluent and safe neighborhood, close parental monitoring may have deleterious effects on the child (because such monitoring may be experienced by the child as overprotective or intrusive), but the very same level of parental monitoring may have beneficial effects within the context of a dangerous, economically stressed community. In order to analyze direct, mediating, and moderating effects, researchers have had to develop new classes of multivariate statistical procedures, such as path analysis, structural equation modeling, and hierarchical linear modeling.

4. Concluding Note

The growing appreciation of the importance of context in the study of human behavior and development stimulated a wave of new research in the 1980s and 1990s that sought to understand how individuals are

affected by different levels of the environment, both independently and in interaction with each other. Multidisciplinary and interdisciplinary research that brought together social scientists from different disciplines became increasingly common, as scholars realized that understanding the complex interrelations among individuals and their environments necessitated drawing on the methods and theories of multiple disciplines. Over time, the study of individuals in natural settings has come to dominate research in fields in which controlled laboratory experiments had once been the dominant paradigm. Once considered a nuisance that interfered with the pursuit of 'pure knowledge,' the context of human behavior is now recognized as a necessary and enormously important focus of contemporary social science inquiry.

See also: Ecology, Cultural; External Validity; Human Ecology; Lewin, Kurt (1890–1947); Situated Knowledge: Feminist and Science and Technology Studies Perspectives; Situation Model: Psychological

Bibliography

Barker R G, Gump P V 1964 *Big School, Small School: High School Size and Student Behavior*. Stanford University Press, Stanford, CA

Bronfenbrenner U 1979 *The Ecology of Human Development: Experiments By Nature and Design*. Harvard University Press, Cambridge, MA

Bronfenbrenner U, Crouter A 1983 The evolution of environmental models in developmental research. In: Kessen W, Mussen P H (eds.) *Handbook of Child Psychology; History, Theory, and Methods*, 4th edn., Wiley, New York, Vol. 1, pp. 357–414

Duncan G, Raudenbush S 1998 *Neighborhoods and Adolescent Development: How can we Determine the Links?* Working Paper Series, Institute for Policy Research, Northwestern University, Evanston, IL

Jencks C, Mayer S 1990 The social consequences of growing up in a poor neighborhood. In: Lynn L E, McGeary M G (eds.) *Inner-City Poverty in the United States*, National Academy Press, Washington, DC pp. 111–86

Lewin K 1951 *Field Theory and Social Science*. Harper, New York

Thomas W I, Thomas D S 1928 *The Child in America*. Knopf, New York

L. Steinberg

Contractarianism

Contractarianism offers an intuitively appealing account of political and moral rules and obligations, based on the idea of consent or agreement. Social institutions and moral codes are to be explained and justified by reference to their *eligibility*—the property that they would be chosen or agreed by individuals of a particular type in particular circumstances. Contractarianism was developed by Hobbes and Locke, and by Rousseau and Kant, and a range of contemporary contractarian positions have emerged in the second half of the twentieth century to challenge utilitarian and other idealist approaches to political and moral philosophy. Contractarianism has clear connections with ideas such as individualism, democracy, and pluralism.

1. Introduction

The voluntary act of entering into a contract is the act of choosing to be involved in a specified joint scheme or commitment. Any argument that places an idea of contract, agreement, or consent in a central role in explaining or justifying a particular joint scheme or commitment might then be seen as 'contractarian' in nature. Most prominent among the joint schemes and commitments that might be explained or justified are political, social, and legal institutions such as government, and moral values. Contractarian arguments are distinctive in terms of their emphasis on considerations of eligibility, so that demonstrating that a particular arrangement or commitment was, or might have been, chosen, at least by certain types of individuals in certain circumstances, carries weight in explicating and justifying that arrangement or commitment. But contemporary contractarian arguments differ from each other along several dimensions, most obviously according to the domain of the argument, the role of the idea of contract or agreement in the argument, and the specific nature of the contract in question. These dimensions map out a space that includes many potential contractarian positions. Not all points in this space are tenable, but a number of them are occupied by contemporary contractarians.

In the domain dimension, a very modest contractarianism would seek to explain or justify some particular activity, policy, or other aspect of social life against a background that already includes a rich variety of social institutions and norms. By contrast, a moderately ambitious or *political* contractarianism seeks to explain and justify basic social and political institutions themselves, rather than merely the activities that occur within them. Finally, a highly ambitious or *moral* contractarianism seeks to provide a theory of morality itself.

In the dimension that identifies the role of agreement in any contractarian argument, we note two major possibilities—agreement might play either a shallow, indicative role or a deep, constitutive role. An indicative role might be relevant in cases where the agreement serves as a useful but inessential signal of

some further and more basic property (the signal may be sufficient but not necessary for the realization of the underlying property). By contrast, contract would play a deeper, more constitutive role where the idea of agreement is essential to the argument (so that the voluntary contract is both necessary and sufficient).

In the dimension that identifies the specific nature of the relevant contract, the most obvious range extends from the case of the fully explicit, actual, or historical agreement, to various specifications of tacit, implicit, or hypothetical agreements. But there is also scope for qualifying the nature of agreement in other ways. Thus, some contractarian arguments will emphasize the informational context of agreement, or the extent to which the agreement is grounded in fully rational deliberation.

The basic idea or vision that unites all contractarianism is the conception of society (in the case of political contractarianism) or morality (in the more ambitious case) as human constructions, so that, for example, society is 'a co-operative venture for mutual advantage' (Rawls 1971, p. 4). This places contractarianism firmly in an individualist setting in which society/morality is to be explained and justified by appeal to the extent to which it responds to human needs, supports cooperation, and suppresses various forms of conflict. From this common starting point, contemporary contractarianism may be divided into two major branches—one originates with Hobbes and may be described as interest-based contractarianism, while the other derives from Kant and may be described as impartial contractarianism.

In the Hobbesian tradition the interests of individuals are taken as basic so that agreement is seen as identifying mutuality of interest. The most basic version of such a theory would begin in presocial state of nature, allow individual interests to be whatever the individuals concerned perceive them to be, and would require real, explicit, and unanimous agreement between individuals. Such a theory is open to a range of obvious criticisms: that agreements based on mistaken or misperceived interest cannot carry the normative weight required by the argument; that since explicit, historical agreements of the required type are not a common feature of the real world, the contractarian argument cannot explain or justify the political or moral features of the real world; and so on. The point to be emphasized here is just that, for contractarians in the Hobbesian interest-based tradition, the idea of contract or agreement is important as a means by which to investigate the prospects for mutual advantage for rational individuals.

In the Kantian contractarian tradition, attention shifts from mutual advantage to impartiality or justice. The basic idea here is that the rational contemplation of the world from an impartial standpoint will reveal what political arrangements are just or what is moral. There is no real attempt to link the contractarian idea to real, explicit agreements. The essential nature of the Kantian contractarian method is to construct a hypothetical, impartial position from which to consider what is just.

2. *Contract and Compliance*

Although the possibility of cooperation is a hallmark of the contractarian vision of society in both its Hobbesian and Kantian forms, it is the nature of the conflict that threatens cooperation that sets the more detailed agenda for contractarianism. This agenda may be illustrated by a sequence of simple games which have become staples of the literature—the pure coordination game, the mixed-motive game, and the prisoners' dilemma. In developing these illustrations, I will use the language of a simple Hobbesian, interest-based contractarian argument, where the contract is explicitly entered into.

In the pure coordination game each player has an interest in coordinating with others, and no other relevant interest. This is illustrated in the simple two-person, simultaneous-play game of Fig. 1, where the pay-offs from each combination of actions are shown in the form (payoff to row, payoff to column).

All that matters in this game is that the two players should take the same action—both should X, or both should Y—since the interests of the two players are perfectly aligned. In this setting is easy to see how preplay communication between the players might allow an agreement which identifies a *convention* in the sense of Lewis (1969). So there is little difficulty in identifying the potential grounds for agreement (although there might be a minor difficulty in identifying which of the actions is to be coordinated on). It is equally easy to see that such an agreed convention will be self-enforcing in that neither player has any reason to depart from the convention once it is established. There is no compliance problem here. This, then, is a model in which a simple contractarian argument promises to work well—agreement seems likely to pick out social or political conventions that are self-enforcing in the relevant population. However, the obvious criticism is that a model of pure coordination is an inappropriate starting point for even moderately ambitious political contractarians since, although it may capture the essential features of some social and political issues, it can hardly claim to capture the

		Column	
		X	**Y**
Row	**X**	**(1,1)**	**(0,0)**
	Y	**(0,0)**	**(1,1)**

Figure 1
Pure coordination

		Column	
		X	Y
Row	X	(2,1)	(0,0)
	Y	(0,0)	(1,2)

Figure 2
Mixed motives

		Column	
		X	Y
Row	X	(2,2)	(0,3)
	Y	(3,0)	(1,1)

Figure 3
Prisoners' dilemma

fundamental nature of society where competition and conflict must be recognized.

A move towards incorporating conflict is made in Fig. 2, which represents the mixed-motive game. The benefits of cooperation are still clear—each player still has an interest in doing what the other does—but now the two players have divergent interests concerning the action to coordinate on. Row has an interest in coordinating on X, while Column has an interest in coordinating on Y. This introduces a difficulty into the contractarian argument, the difficulty that any preplay agreement must now overcome an essential conflict of interests. If we are willing to assume that some agreement is reached, we might still argue that such an agreement forms a convention and is self-enforcing. Consider the case in which Row and Column agree that both should X, Column now has good reason to suppose that Row will X and, given this belief, Column will also have good reason to X. There may be no compliance problem here despite the *ex ante* conflict between players.

A further step in the direction of conflict is illustrated in the standard prisoners' dilemma game of Fig. 3. Here each individual has a private reason to Y, since this is the dominant strategy. The potential and mutual gains from cooperation in this case come from shifting the behavior of both players from Y to X. In a sense, the problem here is the opposite of the problem in the mixed-motive game. The nature of any agreement seems clear enough, precisely because the coordinated shift from Y to X does offer mutual benefits, while no other route to mutual benefits exists. But now the existence of a severe compliance problem may undermine any agreement to X. If both agree to X in preplay negotiations, it is not clear that either will comply with the agreement, since it will be in their individual interests to renege on the agreement regardless of whether they believe the other will keep to the agreement or not (this is just what it means to say that Y is the dominant strategy). The preplay agreement may be dismissed as 'cheap talk' that does not actually impact on individual decision making. If the compliance problem is so severe as to render prior agreement ineffective, there will be no reason for individuals to waste their time in reaching ineffective contracts.

These simple illustrative games serve to identify two of the basic issues facing contractarians, and the relationship between these issues. On the one hand, the issue of the nature and content of any agreement given the likely presence of *ex ante* conflict between individuals and, on the other, the issue of compliance with any agreement given the *ex post* conflicting private incentives facing individuals.

3. *Contemporary Contractarian Positions*

The last two sections have identified two broad traditions within contemporary contractarianism—the Hobbesian interest-based tradition and the Kantian impartiality-based tradition—and two broad issues facing contractarians—the *ex ante* agreement issue and the *ex-post* compliance issue. The remainder of this article will briefly consider contemporary contractarian positions—representing both traditions and offering different approaches to the basic issues. The intention is both to sketch some salient features of contemporary contractarianism, and to reveal some of the strengths and limitations of contractarian argument.

3.1 Hobbesian Contractarianism—Buchanan and Gauthier

Both Buchanan's and Gauthier's contractarianism are clearly and explicitly Hobbesian in nature, but an initial difference between these two contemporary contractarians lies in the scale of their ambition. Buchanan ties his contractarian argument to political constitutionalism, so we might view Buchanan as an essentially *political* contractarian. Buchanan takes contractarianism as providing a justification for political and constitutional structures, rather than merely an explanation, so that he is committed to contractarianism as a means of moral enquiry; but he does not set out to derive any specific or substantive moral principles from the idea of agreement. Buchanan is certainly a moral individualist, but he may be interpreted as allowing individuals to hold whatever moral views they may happen to have (including none at all), insisting only that agreement or contract is the means by which individual moral values, as well as individual desires or preferences, must be brought together in any exercise in normative social or political theory. By contrast, Gauthier's ambition is that of a full moral

contractarian. *Morals by Agreement* sets out an argument (with some of the details revised and extended in later work) for the contractarian construction of morality. More specifically, 'Moral principles are introduced as the objects of fully voluntary *ex ante* agreement among rational persons. Such agreement is hypothetical ..... But the parties to the agreement are real, determinate individuals, distinguished by their capacities, situations and concerns ..... As rational persons understanding the structure of their interaction, they recognize a place for mutual constraint, and so for a moral dimension in their affairs' (Gauthier 1986, p. 9). These differences in ambition between Buchanan and Gauthier relate to further differences in the treatment of both *ex ante* agreement and *ex post* compliance.

A basic distinction in Buchanan's argument is that between 'constitutional' and 'in-period' choices. Constitutional choices are described as choices of rule-systems, while in-period choices are described as choices within rules. The *ex ante* agreement issue, for Buchanan, is then the question of why individuals who may differ markedly in their interests (and values), and in their in-period behavior, may nevertheless agree on rules. Buchanan's approach to this issue focuses on considerations of the nature of the objects of choice rather than the choosing agents. As already noted, Buchanan does not impose on his individuals any restriction in the set of interests or values that they might have. On the contrary, Buchanan frequently insists that 'we start from here' and that individuals must be taken 'as they are', recognizing their diversity. Rather, Buchanan argues that the nature of constitutional choice itself will present alternatives in such a way as to reduce conflict between individuals, and so enhance the prospects for agreement. One key element of this argument is that constitutional provisions are 'general', in that they are expected to apply both across many particular in-period choice situations and across time, so that any individual contemplating the choice between rule-systems will have good reason to judge the alternative rule-systems in terms of their general properties, rather than in terms of the implications for any particular in-period decision. Thus, for example, in choosing a voting rule to be applied to a broad class of in-period political decisions, any individual at the constitutional stage will have good reason to balance the implications of each potential voting rule across a wide range of issues, and this suggests that different individuals may judge voting rules similarly, even if their specific preferences are rather different. This argument seems secure—surely the move from the level of the particular in-period decision to the more abstract constitutional level will tend to suppress at least some of the potential conflict between individuals and so allow potential agreement. However, this apparent success in at least diluting the *ex ante* agreement problem may simply transfer the problem to an arena of compliance. The *ex ante* agreement problem is still more severe for Gauthier. To the extent that Gauthier simply argues for the 'recognition of mutual constraint' at the *ex ante* level, then there is no substantial difference between Gauthier and Buchanan—'mutual constraints' may be identified with constitutional rules without much loss. However, Gauthier needs to go a step further to argue that individuals will come to agree on moral principles rather than just particular, institutionalized constraints or rules. To take this extra step Gauthier relies on a model of bargaining (although the specific theory of bargaining is an area in which Gauthier has revised his views) and a version of the Lockean proviso. The proviso is of particular interest since it seems to be a moral principle, but is not argued to arise out of agreement. It is, rather, a precondition for agreement. Essentially, the proviso states that initial bargaining between individuals should be constrained so that no individual should be worse off in the initial bargaining situation than they would be in a nonsocial context of no interaction. This effectively introduces a minimal set of rights into the original bargaining situation. Gauthier argues that such a proviso has to be accepted by each individual if agreement is to be possible. In one sense this proviso might be thought of as a technical device to ensure that the bargaining outcome actually realizes mutual advantages, and therefore stays true to the underlying Hobbesian idea; but in another sense it reveals a potential weakness of any bargaining model in that the outcome will reflect bargaining power relative to some baseline rights. If such power is to be viewed as illicit in some cases—and ruled out by the proviso—why should the outcome of power be granted moral status in other cases?

The basic compliance question for any Hobbesian contractarian is simply why any individual should feel him- or herself to be bound or obligated by a prior agreement (whether on political rules or moral principles), even if that agreement was explicit and real. This question may be put in terms of the *ex post* motivational impact of the *ex ante* agreement, and whether the motivational impact is sufficient for the agreement to be self-enforcing. If interests are the currency both of *ex post* decision making and *ex ante* agreement, it would seem to place these two activities on all-fours with each other so that the claimed priority of the *ex ante* contract over the *ex post* decision is threatened.

The prospects for wholly self-enforcing contracts seem to be hopeless if we are to take individual interests either as being entirely unspecified, so that any individual may have any interests at all, or if we insist on interests being modeled in terms of narrow self-interest along the lines of the standard depiction of *homo economicus*. In either of these cases, the fact of the contract (if it is a fact) and the content of the contract (whatever it might be) would seem to have no independent or reliable role in determining in-period behavior in situations such as the prisoners' dilemma

where in-period incentives are clear. In order to generate reliable compliance in such cases, we seem to need to assume at least some motivational relevance or status for the contract itself. Without such a move, the range of political and moral principles that might be explained or justified by a Hobbesian contractarian would seem to be limited to those which can be analyzed as self-enforcing conventions. Gauthier attempts such a move by arguing for what he terms 'constrained maximization' rather than straightforward maximization as the appropriate interpretation of rational decision making, where 'A constrained maximizer has a conditional disposition to base her actions on a joint strategy, without considering whether some individual strategy would yield her greater expected utility ....in other words, a constrained maximizer is ready to co-operate in ways that, if followed by all, would yield outcomes that she would find beneficial and not unfair, and she does co-operate should she expect an actual practice or activity to be beneficial' (Gauthier 1986, p. 167). The basic point here is that Gauthier is attempting to ground his moral contractarianism in a specific interpretation of what rationality requires. The detailed claim about rationality is clearly contentious, but without some similar claim, there would be little prospect for contractarianism that goes much beyond the securing of conventions.

A second aspect of the question of compliance relates to the possibility of enforcement. Clearly this is of much more relevance to a political contractarian than a moral contractarian since a moral contractarian must by necessity be concerned with the acceptance and internalization of moral principles but, even for a moral contractarian, it is clear that enforcement may play some role. So, can enforcement resolve the compliance problem? As usual, the answer is both yes and no. Of course, the possibility of more complex contracts which include enforcement mechanisms will extend the range of political and social situations that can be treated. Such complex contracts need raise no new issues at the *ex ante* stage. For example, if a group of individuals contract to engage in a joint venture that has the characteristics of the provision of a public good (the draining of a swamp, perhaps), it will be appropriate for them to recognize that a simple contract in which each person agrees to contribute to the in-period effort and cost may founder since each individual will face strong in-period incentives to free-ride. In such a case, a form of enforcement, perhaps raising taxes against the credible threat of some further punishment if taxes are not paid, seems both to solve the *ex post* compliance problem and be something that offers appropriate mutual advantage so that it will be eligible at the *ex ante* stage. However, this is too fast. Enforcement of the type required cannot simply be assumed to be available. Typically, enforcement requires enforcers and their motivations and decisions to be explicitly considered. Assuming that judges will virtuously apply the law, or that policemen will act regardless of their own individual interests, is not to solve the compliance problem but simply to relocate it. This is not to deny that enforcement is sometimes possible and useful, but merely to suggest that while a political or social arrangement may involve enforcement within its structure, the structure as a whole must be seen to be self-enforcing in the sense that the behavior of all individuals—including those who act as enforcers—are seen to be consistent with their motivational characters and the incentives facing them.

3.2 Kantian Contractarianism—Rawls and Scanlon

Some of the discussion of the Hobbesian positions of Buchanan and Gauthier applies with only minor modification to the Kantian positions of Rawls and Scanlon. In this section I will concentrate on rather different concerns to avoid repetition.

The Kantian nature of Rawls's contractarianism is most obvious in the formulation and discussion of the original position. This position, with its relatively thick veil of ignorance, is the basic means by which Rawls attempts to overcome the potentially conflictual nature of the *ex ante* agreement and generate the impartial conditions under which to explore what justice requires. To (over)simplify, the original position attempts to strip away all of the distinctive and partial characteristics of specific individuals, their particular circumstances, interests, and so on, so reducing the extent of conflict between individuals. In the limit, this process reduces the idea of rational agreement among individuals in the original position to the idea of the rational choice of a single hypothetical and representative individual. This then places almost all of the burden of identifying the institutions and principles that should govern society on the idea of rationality. The danger here is that the original position must satisfy two rather conflicting conditions. On the one hand it must be sufficiently removed from the real world to satisfy the requirement of impartiality, that is, it must strip away almost all the recognizably human aspects of the individual, while on the other hand it must retain sufficient structure to allow the idea of rationality and rational choice to be meaningful and recognizable. This is no trivial task. The standard thin theory of rationality favored by economists and endorsed by Rawls is one in which rational choice is defined in terms of the desires and beliefs of the particular individual; absent of specific desires and beliefs, it is not clear what the thin theory of rationality entails.

Scanlon's contractarianism differs from that of Rawls in several key respects—first in terms of its ambition where Scanlon is explicitly more ambitious than Rawls in arguing for a moral rather than a merely political contractarianism, and second in his formu-

lation of the idea of *ex ante* agreement which makes no major use of the construction of an original position. Scanlon offers the following definition of his account of contractarianism: 'An act is wrong if its performance under the circumstances would be disallowed by any system of rules for the general regulation of behaviour which no one could reasonably reject as a basis for informed, unforced general agreement. This is intended as a characterisation of the kind of property which moral wrongness is' (Scanlon 1982, p. 110). Contractarianism, in this view, is concerned with providing a distinctive basis for moral reasoning by individuals. Thus, for example, 'A satisfactory moral philosophy will not leave concern with morality as a simple special preference, like a fetish or a special taste, which some people just happen to have. It must make it understandable why moral reasons are ones that people can take seriously' (Scanlon 1982, p. 106) and again, 'According to contractualism, the source of motivation that is directly triggered by the belief that an action is wrong is the desire to be able to justify one's actions to others on grounds they could not reasonably reject' (Scanlon 1982, p. 116). This formulation ties together the *ex ante* agreement and *ex post* compliance issues in a particular way. By formulating the idea of agreement in terms of principles 'which no one could reasonably reject', Scanlon allows a degree of flexibility—there is no argument to suggest that there exists a unique set of principles that could not be reasonably rejected—and so weakens the requirement of agreement in Rawls' formulation. But, at the same time, the specification that the hypothetical 'agreement' (or nonrejection) is something that occurs in the minds of real individuals (rather than individuals in an original position) who are also motivated to justify their actions to others provides a route to compliance. It is the desire to justify action to others that seems to drive the motivational side of the problem, and the requirement of 'reasonableness' in the nature of justification that bears the strain in terms of identifying particular justifying arguments. But the dangers here are clear: that the desire to justify my action to others is simply smuggled in as a 'simple special preference like a fetish or a special taste', which is not itself grounded in any contractarian logic; and that the idea of 'reasonableness' is insufficiently detailed to bear the strain placed on it.

There is a final concern that is shared by moral contractarians, whether Hobbesian or Kantian. Even if we grant that it may be possible to parlay rational agreement, in whatever interpretation of that phrase, into moral principles, will the moral principles that emerge be recognizable as our moral principles. In particular, will the resources of contractarianism be sufficient to ground the range of moral views that we actually have and, if not, what would be the implications of that fact? Would it point to the possibility that some of our 'moral' views are not really moral after all, but mere tastes or preferences? Or would it point to the possibility that there may be other means by which moral principles may be grounded, so that contractarianism might not be the only means of approaching moral principles? Of course, this issue only arises once substantive moral principles have been derived from the contractarian method, and there is certainly no consensus, even among contractarians, as to precisely which substantive moral principles can be so derived.

4. *Finale*

Contemporary contractarianism offers approaches to political and moral theory that share as their basic commitments a form of individualism and the recognition of the central role of rationality in human affairs. In a simple sense, contractarianism is dedicated to the question of how rational people can live together. This simple and appealing structure goes a long way to explaining the continuing appeal of contractarianism. Of course, not all accept this vision. Communitarians and others who take society as prior to the individual, for example, must object to contractarianism in all of its forms by reason of its individualism. Equally, not everyone accepts that the basic idea of eligibility is compelling or even interesting. A question to all such critics of contractarianism is how they might expect to persuade individuals (even beyond the range of philosophers and political theorists) that their preferred alternative is valid without falling into a form of contractarianism themselves. For if a particular substantive moral theory or political constitution can be justified to others, then it might be presumed that it could be derived by a contractarian argument that emphasizes that mode of justification. This is simply a way of pointing to the fact that debate between contractarians and others is sometimes guilty of a mismatch between the contractarian method and particular substantive positions that might, at least in principle, be derived from that method.

Contemporary contractarianism is engaged in the study of the relationships between the idea of rationality, the nature of individual interests and dispositions, self-interest and morality, political and social rules, and moral psychology. While there are many areas in which these relationships are still obscure, and while there are surely noncontractarian approaches that may prove rewarding, it seems difficult to deny the importance of these relationships, or the potential of the contractarian enterprise to illuminate them.

See also: Contracts: Legal Perspectives; Hobbes, Thomas (1588–1679); Kantian Ethics and Politics; Kant, Immanuel (1724–1804); Law and Morality: A Continental-European Perspective; Law and Morality: An Analytic Perspective; Trust: Philosophical Aspects

Bibliography

Barry B 1989 *Theories of Justice*. University of California Press, Berkeley, CA
Barry B 1995 *Justice as Impartiality*. Clarendon Press, Oxford, UK
Buchanan J.M 1975 *The Limits of Liberty*. Chicago University Press, Chicago
Gauthier D 1986 *Morals by Agreement*. Clarendon Press, Oxford, UK
Hampton J 1986 *Hobbes and the Social Contract Tradition*. Cambridge University Press, Cambridge, UK
Lewis D 1969 *Convention*. Harvard University Press, Cambridge, MA
Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA
Scanlon T 1982 Contractualism and utilitarianism. In: Sen A, Williams B (eds.) *Utilitarianism and Beyond*. Cambridge University Press, Cambridge, UK

A. Hamlin

Contracts: Legal Perspectives

A contract is a legally recognized agreement made between two or more parties which may be enforced in law, the default of which may give rise to legal liability. In limited circumstances, a contract may be legally unenforceable because it is illegal, contrary to the Statute of Frauds, or procured by fraud. In practical terms, contracts are legally binding because legislatures and courts so determine. Agreements that are not legally binding still may have moral, social, or economic significance.

Contracts must satisfy specified legal requirements in order for there to be a binding agreement. These requirements are contained in legal principles, rules of law, and standards of application. The key legal principle is the liberty of parties to contract (Atiyah 1979). Applicable rules relate, among others, to the exchange of offers between the parties and their performance under the contract. Legal principles and rules of application, in turn, are subject to standards of reasonableness. The liberty to contract must be exercised reasonably. Offers and acceptance must be reasonably certain. Performance must comply reasonably with the promise. What is reasonable, however, is a matter of construction that may vary according to the context.

A contract can be described analytically according to the correlative rights and duties that arise from it. A contract arises when one, but usually two, parties assume correlative rights and duties, usually towards each other (Trakman and Gatien 1999). Contract rights and duties arise in individual or discrete transactions between parties. They also arise in ongoing relationships between parties. Whether contracts are discrete or relational (Macneil 1980), they entail correlative rights and duties between the parties to them. The seller has a duty to provide the goods or services promised; and the right to be paid for them. The buyer has the right to the goods or services promised by the seller; and the duty to pay for them.

Contracts also entail the rights and duties of third parties. For example, parties to marriage contracts owe duties, not only towards each other, but also to the state and other third parties to their contract. They have a duty to register their marriage contract and not to commit bigamy. The state, in turn, owes them a duty to enforce it, but not if one party is a bigamist. Others, such as employees, have a duty not to discriminate on the basis of marital status. The rights and duties of third parties arise as a consequence of the contract, not under the contract itself. Only the parties to a contract ordinarily have rights and duties in terms of it.

Contracts ideally are freely and voluntarily concluded. Such freedom entails the right of parties to choose the terms and conditions of their contracts. For example, parties to marriage contracts cannot be forced to marry, or to agree to particular terms of marriage. However, they can be required to comply with legal rules governing marriage, such as not to enter into incestuous marriages.

Some argue that contracts should be efficient (Kronman 1980). However, efficiency is not a legal requirement of a contract. The terms of a contract must be reasonable, even if they are inefficient (Dworkin 1980). For example, a court may declare that a contract is unenforceable in law for violating public policy even though it is efficient. What is reasonable engages the subjective determination of the party reaching it. Courts, however, are expected to comply with principles of contract law, and rules of application, in deciding what is reasonable.

1. *Legal Requirements of a Contract*

Contracts are concluded between real and artificial persons. Individuals, companies, and governments all enter into contracts. There are many different kinds of contracts. Some relate to family, religious, and cultural life. Some relate to business dealings, others to relations with the state. Parents and children enter into domestic contracts when they agree to share car expenses. Consumers and supermarkets conclude consumer contracts to purchase groceries and other household goods. Merchants enter into business contracts when they purchase goods and services for resale, when they insure them, ship them, and secure credit. Governments enter into contracts when they grant individuals and companies licenses to trade, when they conclude agreements with multinational corporations or other governments, and when they sign multilateral conventions.

Some contracts are subject to formal requirements. Marriage contracts must be solemnized publicly through registration. Sales of property must also be registered. Most contracts, however, are not subject to registration. Some must be in writing. Most are concluded informally. A nod of one's head may constitute a legal agreement. Even complex business contracts may be concluded informally, such as on the golf course or at a party.

Contracts may be subject to substantive requirements. One substantive requirement is that the parties to contracts must intend to create binding relations, or such an intention may be imputed to them. A related requirement is that they must exchange promises, or one party must perform an act or acts in terms of a promise by the other. A further requirement is a bargained-for-exchange or consideration in common-law jurisdictions (Sutton 1974). Civil law jurisdictions require a *causa* or reason for contracting (von Mehren 1959). A contract is distinguished from a gift in both common and civil law. The parties to a gift do not ordinarily enter into a bargain or an exchange. Nor do they have a reason for contracting. What is a contract or a gift, however, is a question of construction upon which jurists may differ.

Most contract requirements are satisfied in stages. The formation stage occurs when a party makes an offer or promise which the other party accepts. For example, a party offers to sell goods to another, and the other party accepts the offer. The performance stage occurs when the parties perform the contract, such as when the seller delivers the goods and the buyer pays for them. A nonperformance stage *may* occur if the seller fails to deliver the goods or the buyer fails to pay. This failure to perform ordinarily gives rise to a duty, such as to compensate the other party for losses incurred. Absent nonperformance, the final stage, is the termination of the contract, as when all goods are delivered and paid for under the contract.

Non- or defective performance can give rise to litigation for default or breach of contract. For example, a seller may deliver goods which the buyer alleges are defective. The buyer may sue for the contracted goods, substitute goods and/or damages for losses. In Anglo-American jurisdictions, the most frequent remedy is for the defaulting party to pay damages to compensate the other party for breach (Fuller and Perdue 1936). In civil law jurisdictions, the most common remedy for default is specific performance. The seller in default is required to perform specifically in terms of the contract, for example, by delivering the goods as promised. Both civil and common law jurisdiction, however, recognize a range of remedies, varying from compensatory to punitive damages, and from specific to modified performance.

Law and economics scholars sometimes conceive of contract default as *amoral*, not immoral (Coleman 1980). They maintain that a seller should be entitled to default, without moral judgment, so long as the seller compensates the buyer for damages arising from that default. This attempt to legitimize efficient breach is suspect, however, when it produces inequities and losses that cannot readily be quantified, such as sentimental or emotional loss. Breach may also be inefficient when an 'innocent' party accepts inadequate compensation in response to litigation threats by the other party.

2. *Construing Contracts*

In the nineteenth and early twentieth centuries, the interpretation of contracts was grounded in a subjective theory of contracting. A contract was formed as a result of ascertaining the subjective or intention of the parties to it. Significantly, this subjective theory was replaced over the twentieth century by an objective theory. Here, the intention of the parties is determined, less by their subjective intention, than according to the objective circumstances surrounding their contract. Judges resorted to this objective intention because the subjective intention of the parties was allegedly ambiguous, or to prevent a party from manipulating its subjective intention at the time of dispute. In some measure, the objective theory incorporates a tort standard of due care into contract law. Contracting parties are required to communicate with due care and not to rely on a doubtful subjective intention (Sharp 1939). Ultimately, the objective intention of the parties derives from the courts' construction of what it believes is reasonably intended in the circumstances. Their objective intention is expressed subjectively by the court. At its most objective, the court examines the parties' course of dealings, past practices, and usages of trade (Devlin 1951). At its most subjective, the court applies its own conception of reasonableness in determining the objective intention of the parties.

This does not infer that, while parties may manipulate their subjective intention, courts manipulate the objective theory. Courts ordinarily choose to apply the subjective or objective theory of contracting along a continuum. The presence of clear and unambiguous evidence of a subjective intention by both parties supports use of the subjective theory. The absence of such evidence supports use of the objective theory. The application of the objective theory used, in turn, depends on the contractual circumstances. These include, among others, the complexity of the transaction, the bargaining power of the parties, and changed circumstances since the date of contracting (Trakman 1985). In applying the objective theory of contracting, courts are likely to be influenced significantly by an implied duty of care owed by the parties. For example, courts are likely to construe the subjective intention of a producer in a take-it-or-leave-it consumer contract restrictively, in accordance with its implied duty of care towards its customers. They

are more likely to construe the subjective intention of the parties expansively, in the absence of such adhesive elements (Kessler 1943).

Courts use different methods of interpretation to identify the intention of the parties. At one extreme, they use literal methods to establish what the parties actually intended, based on the plain word meaning of their contract. On the other extreme, they use open-textured methods, varying from party practices and trade usages, to the courts' own construction of what is reasonable (Farnsworth 1968). These open-textured methods of interpretation ordinarily give rise to judicial 'gap-filling.' Judges fill gaps in contracts on grounds that contracts are incomplete and reasonably warrant filling (Costigan 1920). They make different assumptions in doing so. One assumption is that, had the parties considered the gap, they would have filled it, as did the court. Had the seller and buyer anticipated a breakdown in supply, they would have provided for it by contract. Another assumption is that the buyer and seller knew of the gaps, but did not fill them in order to clinch the deal. Business people sometimes fail to provide for default in the hopeful expectation that default will not arise (Macaulay 1963). Courts fill contract gaps on behalf of such parties. A further assumption is that judges fill gaps in 'interpreting' contracts to avoid injustice. They imply warranties of merchantability into contracts of sale to prevent sellers from excluding their liability to customers.

The extent to which courts ought to fill gaps in contracts is controversial. Excessive gap filling is considered incompetent contracting and a threat to the liberty of contract. Judges who avoid gap filling are viewed as opening the door to between the parties bad faith dealings. Courts respond to this controversy according to the perceived nature of the gap, the parties' anticipation of it, the injustice arising from not filling it, and the courts' discretion in filling it (Farnsworth 1968).

3. *From Form to Substance*

The nineteenth and early twentieth centuries were marked by formalism in contract law. Contracts consisted of distinct forms associated with categories of contracts, such as marriage contracts and contracts for the purchase and sale of property. Each form of contract had its own characteristics, and each gave rise to distinct legal consequences (Atiyah 1979).

The pervasive use of formal contracts declined in the early years of the twentieth century. Jurists emphasized increasingly the substance of contracts. This shift from form to substance developed, in part, in response to the evolution of legal pragmatism. Pragmatists argued for a functional approach towards law (James and Castell 1948, Dewey 1924). Some stressed the substantively position of contracting parties and the functional role of courts in relation to them (Kessler 1943). Others scrutinized how law-makers and judges ought to address business risks arising in contract (Patterson 1924). Yet others challenged the classical theory of liberty of contract. Their concern lay in substantive abuse in contract practice. They maintained that, for there to be true liberty of contract, contracting parties ought to be free *both* to conclude contracts, *and* to influence their terms (Pound 1908, Williston 1921). The result was a functional analysis of fine-print disclaimer and warranty clauses used by dominant parties to limit or exclude their liability by contract. The purpose was for consts to construe such adhesive contract clauses against the drafter and in favor of the weaker party.

This functional approach towards contract law was supported by the belief that the law should be developed and applied in light of contract behavior. Contract law was regarded as less an aloof juridical science, than a response to social behavior. Contract practices were to be regulated, less through principled ends inhering in contract law, than through social ends imputed by law (von Jhering 1913, Pound 1943). This engineering of law through social science was supported by two schools of thought: sociological jurisprudence; and in regard to the judiciary, legal realism. Sociological jurisprudence identified the relationship between social behavior and the law. Legal realism concentrated on predicting how judges made law, including contract law. Both emphasized the interdependence between the law and social or business practice (Powell 1956, Devlin 1951). This interdependence also became central to an evolving Law and Society movement (Stone 1966, Selznick 1959) and contract law scholars within it. Some established the distinction between long-term relationships and discrete transactions (Macneil 1980). Others scrutinized the manner in which business communities devised noncontractual as distinct from contractual relations (Macaulay 1963).

The relationship between law and social sciences gave rise to critical attacks on the liberal roots of modern contract law in the late twentieth century. Critical Legal Theorists charged that *laissez-faire* principles and standards of modern contract law favored the privileged elite, disempowering workers and consumers (Unger 1986). Irrationalists demolished such principles as internally contradictory (Singer 1984). Critical scholars deconstructed contract law rules and standards as incoherent (Dalton 1985). Philosopher kings reconstructed contract law values away from individual autonomy and towards communal solidarity (Unger 1986).

The end of the twentieth century saw an uneasy balance between ideological conflicts over liberty of contract and subset individual and community values. Critical scholarship facilitated the partial reconstruction of contract law principles along communitarian lines. Strident communitarians who confronted *laissez-faire* contracting (Sandel 1982) were offset by

moderates who balanced individual free choice against family and religious values (Glendon 1991).

The early twenty-first century is likely to rekindle neoclassical models of free choice theory. Innovations in contracting, such as Internet transacting, are likely to be heralded as exemplifications of *laissez-faire* practices at work. New technologies are likely to enhance modes of communicating, including contracting. Courts are likely to resort to more refined plain-word methods of interpreting contracts and fewer open-textured ones. Judicial gap-filling in contracts is likely to be constrained. Disparities in contract bargaining power are likely to be construed more narrowly on grounds that consumers enjoy improved levels of literacy, foresight of risk, and bargaining power (Bernstein 1996).

These neoclassical developments likely will produce healthy tensions in contract law. Those who champion a neoclassical freedom *to* contract will accentuate the right of parties to conclude contracts of their own free will. Those who accentuate the abuse of free choice will argue for freedom *from* contract in an imperfect market. The debate over 'the rise and fall of freedom of contract' likely will be revitalized. Practices, such as the exchange of marriage vows over the Internet, will evolve; as will restrictions imposed on them in the interests of public policy and morality. Nation-states are likely to sanction the Internet as an instrument of contract practice. But they will try to constrain excesses in its usage.

The early twenty-first century is likely to include greater resort to structuralism in contract to redress internal contradictions in contract law principles and undue variance in substantive rationality (Shamir 1993). It is also likely to work towards greater analytical coherence in relating contract law to contract practice.

4. Conclusion

Almost a century ago, Sir Henry Maine (1917) wrote that the shift from ancient to modern society involved a shift from status to contract. The twentieth century represented a partial shift back to status. A model of perfect competition, applied in the early Industrial Revolution, was replaced by a model of imperfect competition in the post-Industrial Revolution period. The assumption that contracting parties had perfect knowledge and equal ability was replaced by the assumption that they enjoyed unequal education, legal know-how, and economic and political power. Against this background, jurists stressed the substance of contract law over its form. They evaluated contracts functionally, according to the nature of contract relations and discrete transactions. They construed them in light of the imperfect nature and value of the bargain, and the disparate positions of parties.

The early twenty-first century has revitalized the formal rationality of a century earlier. This renewal comes at a time of a new technological revolution. New technologies have opened the door to new contract practices. Contract bargains may be reached through inventive methods of communication that transcend differences in legal systems and national boundaries. Contract formalism is an understandable response to such developments. The new technological era has augured the appearance of a panacea in which sophisticated, informed, and educated parties contract freely across the globe in the absence of regulation. This perception has rejuvenated the image of *laissez-faire* contracting. Parties choose the terms of their contracts freely. Courts construe them literally according to their plain-word meaning; and nation-states enforce them.

This *laissez-faire* revitalized formalism in contract relations is likely to diminish over the course of the twenty-first century. Debate over the freedom *to* contract and the freedom *from* contract in unequal bargaining relations is likely to renew. Lawmakers are likely to challenge the assumption that the masses who have access to sophisticated technology are equally able to use and enjoy it in concluding contracts. Despite the access of billions of consumers to new technologies, including ways of contracting, their disparate understanding and use of those technologies will expand over time, not decline. Nation-states and their courts will return, perhaps with greater vengeance than before, to examining the substance of contract relationships. Contract forms will be scrutinized for their palpable social impact, beyond the form itself. The reasonable contracting party will be subject to increasing standards of due care. Motivated by both public and self-interest, nation-states will extend their regulation of contracts. Debate will rage over the extent to which contract law should inform, or be informed, by contract practice. Contract jurists will take sides; and resolution will remain elusive.

See also: Contractarianism; Industrial Relations and Collective Bargaining; Markets and the Law; Negotiation and Bargaining: Organizational Aspects

Bibliography

Atiyah P S 1979 *The Rise and Fall of Freedom of Contract*. Oxford University Press, Oxford, UK

Bernstein 1996 Merchant law in a merchant court: Rethinking the code's search for immanent business norms. *University of Pennsylvania Law Review* **144**: 1765–821

Coleman J L 1980 *Markets, Morals and the Law*. Cambridge University Press, Cambridge, UK

Costigan G P 1920 Implied in fact contracts and mutual assent. *Harvard Law Review* **33**: 376–400

Dalton C 1985 An essay on the deconstruction of contract doctrine. *Yale Law Journal* **94**: 997–1114

Dewey J 1924 Logical method and law. *Cornell Law Quarterly* **10**: 17–27

Devlin P 1951 The relation between commercial law and commercial practice. *Modern Law Review* **14**: 249–66
Dworkin R 1980 Wealth as value. *Journal of Legal Studies* **9**: 191–226
Farnsworth E A 1968 Disputes over omissions in contracts. *Columbia Law Review* **68**: 860–91
Fuller L, Perdue W R 1936 The reliance interest in contract damages. *Yale Law Journal* **46**: 52–96
Glendon M A 1991 *Rights Talk*: *The Impoverishment of Political Discourse*. Free Press, New York
James W, Castell A 1948 *Essays in Pragmatism*. Hafner, New York
von Jhering R 1913 *Law as Means to an End* [trans. Husik I]. The Boston Book Co., Boston, MA
Kennedy D 1982 Distributive and paternalistic motives in contract and tort. *Maryland Law Review* **41**: 563–658
Kessler F 1943 Contracts of adhesion—Some thoughts about freedom of contract. *Columbia Law Review* **43**: 629–42
Kronman A 1980 Wealth maximization as a normative principle. *Journal of Legal Studies* **9**: 227–42
Llewellyn K N 1930 What price contract? An essay in perspective. *Yale Law Review* **11**: 704
Macaulay 1963 Non-contractual relations in business: A preliminary study. *American Sociological Review*. **28**: 55
Macneil I 1980 *The New Social Contract*. Yale University Press, New Haven, CT
Macneil I 1982 The efficient breach of contract: Circles in the sky. *Virginia Law Review*. **68**: 947–69
Maine H 1917 *Ancient Law*. Everyman, London
von Mehren A T 1959 Civil law analogues to consideration: An exercise in comparative analysis. *Harvard Law Review* **72**: 1009–78
Patterson E W 1924 The apportionment of business risks through legal devices. *Columbia Law Review* **24**: 335–59
Pound R 1908 Liberty of contract. *Yale Law Journal* **18**: 454–87
Pound R 1943 Sociology of law and sociology jurisprudence. *University of Toronto Law Journal*. **1**: 1–20
Powell R 1956 Good faith in contracts. *Current Legal Problems* **9**: 16–38
Sandel M J 1982 *Liberalism and the Limits of Justice*. Cambridge University Press, Cambridge, UK
Selznick P 1959 The sociology of law. In: Merton R K, Broom L, Cottrell L S (eds.) *Sociology Today: Problems and Prospects*. Basic Books, New York
Shamir R 1993 Formal and substantive rationality in American law: A Weberian perspective. *Social & Legal Studies* **2**: 45–72
Sharp M P 1939 Promissory liability I. *University of Chicago Law Review* **7**: 1-23, 250–80
Singer J W 1984 The player and the cards: Nihilism and legal theory. *Yale Law Journal* **94**: 1–70
Sutton K C T 1974 *Consideration Reconsidered*. University of Queensland Press, St. Lucia, Qld
Stone J 1966 *Law and the Social Sciences, The Second Half-Century*. University of Michigan Press, Ann Arbor, MI
Trakman L E 1985 Winner take some: Loss sharing and commercial impracticability. *Minnesota Law Review* **69**: 471–19
Trakman L E, Gatien S 1999 *Rights and Responsibilities*. University of Toronto Press, Toronto, ON
Unger R 1986 *The Critical Legal Studies Movement*. Harvard University Press, Cambridge, MA
Williston S 1921 Freedom of contract. *Cornell Law Quarterly* **6**: 365–80

L. E. Trakman

Control Behavior: Psychological Perspectives

The scope of control-related phenomena is extensive. It encompasses numerous concepts such as perceived control, personal control, cognitive control, vicarious control, illusory control, proxy control, outcome control, primary and secondary control, decisional control, and predictive control (Skinner 1996). In her review of a decade of control-related research, Skinner (1996) identified some 100 different control-related constructs, and cautioned that 'the large number of terms has produced some theoretical confusion about the boundaries of the topic of control, about the interrelationships among constructs, and even about which constructs can be appropriately included in the study of control' (Skinner 1996, p. 550).

1. Conceptual Distinction Between Control Behavior and Perceived Control

The field of control-related research has suffered from conceptual confusion between constructs addressing control behavior on the one hand and constructs about perceptions or beliefs about control on the other (see also review in Schulz and Heckhausen 1999). Although a functionalistic analysis of control-related behavior would view control behavior as more critical for determining outcomes, research on control since the 1970s has been dominated by constructs and assessments of perceived control and related concepts (e.g., self-efficacy, Bandura 1977). This almost exclusive emphasis on perceived control rather than control-directed behavior or realized control has lead to widespread conceptual confusion, in that all kinds of control are subsumed and not differentiated from perceived control (see, e.g., Skinner 1996).

Before reviewing the key findings from the different bodies of research, specific definitions for the major control constructs are given. *Control-related behavior* is directed at producing contingencies between own behavior and events. When the events targeted are located in the outside world, the control-related behavior is *primary control behavior*. In contrast, when internal events are targeted it is *secondary control behavior*. *Control* is realized contingency between behavior and event. *Perceived control*, by contrast, refers to the extent to which an individual believes to control behavior–event contingencies. It thus becomes clear that perceived control does not reflect behavior, but internal representations, and may thus be involved in secondary control.

2. Research About Perceived Control

Perceived control and related constructs have a long-standing history in the psychology of motivation and self-regulation. Five major conceptions have guided

research in this area: locus of control (Lefcourt 1981, Rotter 1990) (see *Control Beliefs*: *Health Perspectives*), causal attributions (Weiner 1985) (see *Motivation and Actions*, *Psychology of*), learned helplessness (Seligman 1975, Abramson et al. 1978) (see *Control Beliefs*: *Health Perspectives*), self-efficacy (Bandura 1977) (see *Control Beliefs*: *Health Perspectives*), and perceived own intelligence as fixed vs. malleable (Dweck and Elliott 1983). Locus of control refers to the issue whether an event is controlled or produced by factors inside or outside the individual. Causal attributions refer to the retrospective explanation of outcomes of behavior by internal factors such as effort or ability and external factors such as task difficulty and luck. Learned helplessness focuses on the issue of whether events are contingent on the organism's behavior, and further distinguishes between stable and unstable and specific and unspecific causes of events. Finally, the distinction between fixed vs. malleable intelligence addresses inter-individual differences in the tendency to view intelligence as either an unchangeable entity or as a modifiable resource for performance. All these characteristics addressed by the five theories have been found to hold important implications for behavior and the evaluation of outcomes (Weiner 1986, Heckhausen 1991, Skinner 1995, Dweck 1999). Depending on the outcome, internal, stable, and general control beliefs are conducive, in the case of success, or detrimental, in the case of failure, to effort investment, ambitious task choice, persistence, and improvement in performance. Moreover, the temporal direction of control perceptions is critical, with prospective expectations of control influencing goal choice, and persistence, and retrospective attributions of control guiding the way in which success and failure are interpreted as indicative of future control potential.

Recently, researchers in this field have called for conceptual and domain-specific distinctions to be discussed in more detail here (for more detailed information about research traditions, see *Control Beliefs: Health Perspectives*; *Motivation and Actions*, *Psychology of*). Perceptions about personal control involve two components, 'agents of control, and means of control' (Skinner et al. 1988, Skinner 1996, p. 552). These two components are often not sufficiently differentiated, but confounded in a summary construct of perceived control. First, control beliefs imply conceptions about the controllability and the relevant means to attain an outcome. This component of control beliefs is captured by various closely related concepts (see review in Skinner 1996): *contingency judgment*; *means–ends beliefs*, and *strategy beliefs*. Second, control beliefs involve conceptions about the agent's resources for control in terms of having access to means for bringing about an outcome. This component of control beliefs is referred to as *competence judgment*, *agency beliefs*, and *capacity beliefs*.

Research about perceived control falls into two broad areas: (a) normative age changes in perceptions of control across the lifespan, and (2) the association of perceived control and various functional outcomes, such as task-performance, school grades, subjective well-being, and health (see also review in Schulz and Heckhausen 1999).

Research on childhood development of perceived control has shown that control beliefs become increasingly realistic, thus moving away from illusory gross overestimations of control in early childhood (Weisz 1983). However, the extent to which perceptions of personal control become realistic in mid-childhood, adolescence, and adulthood is a function of cultural variation (Little et al. 1995). School children in the United States, for instance, hold beliefs about their competence to attain good grades (i.e., agency beliefs), which have little relation to their actual school grades, whereas children in the former German Democratic Republic reported agency beliefs which closely resembled their actual school achievements.

In addition, developmental research has revealed that the agency and the means–ends components of control beliefs, as well as beliefs about different causal factors (e.g., ability, effort, luck, teacher) were increasingly differentiated during mid-childhood and exhibit unique developmental trajectories and relations to outcomes. Skinner et al. (1998), for instance, showed that during mid-childhood, cycles of influences between perceived personal control and school achievement were focused on different components at different ages. Beliefs about effort dominated the cycles in younger school children, whereas beliefs about ability were the focus of the regulatory cycles in older children (see *Self-regulation in Childhood*).

As to normative age changes during adulthood the findings to date provide a mixed picture (see review by Schulz and Heckhausen 1999). While earlier reviews of the literature reported little evidence for age-related changes in generalized measures of control (Lachman 1986a), more recent research has identified decreases in perceived control across adulthood (e.g., Mirowsky 1995). However, Lachman was the first to emphasize and provide evidence for domain specific perceptions of control and their age-related trajectories (Lachman 1986b). Lachman's research shows that older adults report decreasing perceptions of control for those domains for which their more constraint social roles and declining physical fitness yield deteriorated opportunities and resources for control.

The association of perceived control and health has been studied extensively using field experimental paradigms involving interventions improving objective control opportunities (e.g., Langer and Rodin 1976, Schulz and Hanusa 1978, Baltes and Baltes 1986), quasi-experimental methods exploiting naturally occurring environmental changes of control potential (e.g., Timko and Moos 1989), and large sample surveys (e.g., Krause 1987, Menec and Chipperfield 1997). Findings from all these studies, research paradigms, and assessments of perceived control and health

outcomes converge to show that having high perceptions of control benefits one's health, and having low perceptions of control can be damaging (see also review in Schulz and Heckhausen 1999) (see *Control Beliefs*: *Health Perspectives*; *Health*: *Self-regulation*; *Self-regulation in Adulthood*).

3. *Evolutionary Origin and Ontogenetic Potential of Human Control Behavior*

A common characteristic of all motivated behavior is that the organism attempts to achieve outcomes in the environment by its own activity. In all activities relevant for survival and procreation, such as foraging, attracting a mate, or competing with a rival, the organism strives for control in terms of bringing about desired outcomes and preventing undesired ones. Thus, the most fundamental and universal motivational tendencies relate to this basic strive to control the environment referred to as *primary control striving* (Heckhausen and Schulz 1995, Heckhausen 2000).

The striving for control is shared with a broad range of species and goes back far into the phylogenetic past, probably as far back as to those species which first acquired a notable flexibility in their behavior programs (see review by Heckhausen 2000). This assumption converges with White's (1959) classical article on the motivation for competence, effectance, and mastery as a universal striving of humans and mammals in general.

The fundamental status of control striving in behavior regulation is supported by findings about neonatal behavioral preferences (see review by Heckhausen and Schulz 1995). Neonatal humans are able to detect behavior–event contingencies, learn head movements to control milk reinforcement, and, most importantly, continue to show the control behavior even when after complete satiation the milk reinforcement has lost its consummatory value and merely provides a contingency event to the infants' head-turning behavior.

Humans hold a unique potential for control in terms of pursuing goals, which can involve multiple subgoals and cover extensive time spans. This capacity is based on the mental representation of goals, which renders the organism independent of immediate external response elicitation and feedback, and sets it free for self-regulation.

Another important characteristic of humans' control-related behavior is their sense of self, and their quest for learning about and feeling good about the characteristics of their own self (e.g., about competence). The sense of self is acquired in early childhood, and in itself becomes a motivational incentive for control striving (see review by Heckhausen and Schulz 1995). Control goals are pursued not only for the sake of their outcomes, and the behavior–event contingency they represent, but also in view of the anticipated enhancement or confirmation of self-ascribed capacities to control outcomes. Mental representations of one's own control capacity and of the control potential inherent in a situational challenge are referred to as *perceived control* (Skinner 1996).

Particularly for challenging control pursuits, the anticipation of positive information about one's own competence implied in a successful completion of the control striving becomes a major resource for motivation. Anticipatory positive self-evaluation motivates the individual persistently to invest effort into acquisitional processes. Only this way can time-extended acquisitional processes of expertise development be sustained motivationally. However, this ability to develop conceptions about one's own competence and control also renders the human motivational system vulnerable to experiences of failure (see *Self-regulation in Childhood*). From about 4 years of age, humans perceive failure as negative feedback about their own competence and react with embarrassment and shame (see review by Heckhausen and Schulz 1995). For humans older than 4 years of age, failure may thus undermine self-ascribed competence and perceived control, both important motivational resources needed for persistent primary control striving.

The system of control behavior in humans therefore needs to involve processes that protect and restore conceptions of own competence and mastery. Heckhausen and Schulz in their lifespan theory of control (Heckhausen and Schulz 1995, Schulz and Heckhausen 1996) have proposed to subsume diverse, mostly cognitive processes, which serve to protect motivational resources and focus them on a given goal (i.e., volitional commitment) of primary control striving, under the construct of *secondary control striving*. Among such secondary control processes are those that help the individual to disengage from a futile goal (e.g., devalue goal, increase value of alternative goals) and protect the self against unfavorable implications of failure and loss (e.g., downward social comparison, egotistic causal attributions). In addition, secondary control strategies are also needed for generating and maintaining motivational commitment to a chosen goal (e.g., enhance value of chosen goal, devalue alternatives, overestimate perceived control for goal attainment, Gollwitzer and Kinney 1989, Heckhausen 1999, Wrosch and Heckhausen 1999).

4. *Research About Control of Development*

The study of actual control behavior and control striving, as distinct from perceived control, is an emergent field of research (Heckhausen and Schulz 1995, Brandtstädter 1998, Schulz and Heckhausen 1999). Two major models of developmental regulation, which address processes of individual control

in developmental change, are considered here: the dual-process model of coping by Brandtstädter and colleagues (overview by Brandtstädter 1998), and the life-span theory of control by Heckhausen and Schulz (Heckhausen and Schulz 1995, Schulz and Heckhausen 1996).

4.1 The Dual-process Model: Assimilation and Accommodation

Brandtstädter and colleagues (Brandtstädter 1998, Brandtstädter et al. 1999) view personal control of development as motivated by the fundamental goal of preserving personal continuity and identity. In the service of self-continuity and consistency, the individual can profit from two processes, assimilation and accommodation, which both help to bring desired and actual situations or states into congruence. Assimilation encompasses all intentional action directed at achieving goals. In contrast, accommodative processes are subintentional and help the individual to protect the self by adjusting mental representations (e.g., lowering goal aspiration) to external constraints. These two processes are conceptualized as antagonistic but complementary modes of coping.

In the context of psychological aging, Brandtstädter proposes that accommodative processes play a major role in bringing about the seeming paradox of aging-related decline without costs to psychological well being. He argues that it is through processes of downscaling goals, and adjusting aspirations to attainable outcomes, that individuals maintain their sense of consistent self. Accordingly, the two modes of coping follow inverse developmental trajectories across the lifespan: assimilation decreases, and accommodation increases towards old age.

A particularly important contribution of Brandtstädter's model is its inclusion of subintentional processes (Brandtstädter et al. 1999), by which evaluations of goals and aspiration levels can be adjusted, when goal attainment has become uncontrollable. Without intent and even awareness, the attractiveness of a goal, which has become unattainable in old age, might fade away. Also, cognitions that immunize the self against failure experiences (e.g., deflect self-blame) unwittingly emerge in the individual's mental account of a loss-relevant episode. Such subintentional and subconscious accommodations have been demonstrated in experimental studies, and were shown to be closely associated with superior psychological well being.

4.2 The Lifespan Theory of Control

Heckhausen and Schulz (Heckhausen and Schulz 1995, Schulz and Heckhausen 1996) have developed the lifespan theory of control which proposes that the desire to exert control over ones environment and thus realize primary control rules the system of control behavior. The lifespan theory of control elaborates the distinction between primary and secondary control, originally proposed by Rothbaum et al. (1982) and applies it to the human lifespan. Primary control *striving* is expected to remain the dominant motivator of behavior across the lifespan, while the potential to realize primary control undergoes radical changes across age (see also Heckhausen 1999). The potential for primary control, in contrast, increases during childhood, adolescence, and young adulthood, then reaches a maximum and plateau during mid-life, and declines with the loss of social roles and physical fitness associated with old age. The increasing discrepancy between primary control striving and primary control potential at older ages provides a challenge for the individual, in that it can only be managed by disengagement from age-inappropriate goals and engaging with new more age-adapted goals (see *Coping across the Lifespan*). Such cycles of goal engagement and disengagement can only be mastered by activating specific processes addressing the internal world of the individual in terms of goal hierarchies and motivational commitment. Such strategies are conceptualized as *secondary control strategies*. Brandtstädter's conception of accommodative coping processes emphasizes the nonintentional nature of some goal disengagement and self-protective processes, which he contrasts with assimilative coping that comprises intentional attempts to control developmental losses (Brandtstädter 1998) (see *Coping across the Lifespan*).

Heckhausen and Schulz (Heckhausen and Schulz 1995, Heckhausen 1999) distinguish between four types of control strategies. *Selective primary control* refers to the investment of internal behavioral resources such as effort, time, and skills in the pursuit of a goal. *Compensatory primary control* is required when internally available behavioral resources of the individual are insufficient to attain the goal, and external resources such as help or advice from others or technical aids (e.g., wheelchair) have to be recruited. *Selective secondary control* serves to enhance and maintain motivational commitment to a chosen action goal, particularly when it is challenged by unexpected obstacles or attractive alternatives. Selective secondary control strategies include enhanced valuation of the chosen goal, and devaluation of nonchosen alternatives, as well as positive illusions about one's control, potential for the chosen goal. *Compensatory secondary control* is required when the individual experiences a loss of control and when the goal becomes unattainable or excessively costly. Compensatory secondary control can be attained by disengaging from the obsolete goal. In addition, compensatory secondary control involves specific self-protective strategies, such as self protective causal attribution (avoid self blame; see also *Attributional Processes: Psychological*) and

downward social comparisons, which deflect the potential negative effects of failure experiences on important motivational resources of affective balance and self-esteem.

The theory of control behavior can be integrated with modern motivational action theory, which views motivation behavior as organized in action cycles of goal engagement and goal disengagement Heckhausen 1991). Specifically, goal engagement involves the strategies of selective primary and selective secondary control as well as, in cases of obstacles, compensatory primary control. Goal disengagement relies of the compensatory secondary control strategies of goal distancing and self-protection.

The use of primary and secondary control strategies was investigated in various developmental ecologies during the lifespan (Chipperfield et al. 1999, Heckhausen 1999, Wrosch and Heckhausen 1999). Heckhausen (1999) found in a socially heterogeneous sample of East and West German adults that striving for primary control remained constant across age groups, whereas the flexibility of adjusting one's goals, a key component of compensatory secondary control, increased with higher ages. Chipperfield et al. (1999) investigated the effectiveness of primary and secondary control strategies in a sample of older adults experiencing either acute temporary or chronic severe health stress. They found that primary control strategies had positive health implications for the young–old, but that the same strategies appeared to be detrimental to health in advanced old age.

One way to study the responsiveness of control behavior to different control potential is to study developmental regulation across distinct life-course transitions in control potential (Heckhausen 1999) (see *Self-regulation in Adulthood*). Developmental regulation across the lifespan is organized along a timetable of waxing and waning opportunities to attain important developmental goals, such as graduating from school, building a career, finding a permanent partner, and having a child. The age-graded opportunity structures for various developmental goals involve final time constraints, conceptualized as developmental deadlines. An example is the so-called 'biological clock' for child-bearing.

Developmental deadlines were shown to be the watersheds between urgent predeadline goal engagement and postdeadline goal disengagement and self-protection (Wrosch and Heckhausen 1999). Individuals approaching a developmental deadline, such as the early 40s in the case of child-bearing, activated goal engagement strategies of control (i.e., selective primary, selective secondary, compensatory primary control) fervently striving to attain the goal before the time runs out. In contrast, once a deadline has been passed, (e.g., women in their early 40s or 50s with regard to child-bearing) control strategies of goal disengagement, such as devaluing the goal and self-protective downward comparison with others were preferred. Moreover, it was shown that the employment of phase-congruent control strategies (i.e., goal engagement in the predeadline phase and goal disengagement in the postdeadline phase) was associated with greater psychological well-being and mental health (e.g., less depressive symptoms, more positive affect) than phase-incongruent control strategies.

5. *Future Directions of Research*

Future research about control behavior and control perceptions should take a functionalistic perspective and analyze the consequences that objective changes in opportunities for control have for domain-specific components of perceived control and their behavioral correlates. Specifically, future research can exploit the model of action cycles. In both developmental and non-developmental contexts, control processes associated with goal engagement and goal disengagement can be investigated in terms of their congruence with control opportunities. This approach also allows to study interindividual differences in the ability to recognize changes in opportunity structures and in the flexibility of switching between goal engagement and goal disengagement. This may provide an avenue to investigate interindividual differences in vulnerability versus resilience to developmental transitions.

See also: Control Beliefs: Health Perspectives; Coping across the Lifespan; Self-efficacy; Self-efficacy and Health; Self-regulation in Adulthood; Self-regulation in Childhood; Stress and Coping Theories

Bibliography

Abramson L Y, Seligman M E P, Teasdale J D 1978 Learned helplessness in humans: critique and reformulation. *Journal of Abnormal Psychology* **87**: 49–74

Baltes M M, Baltes P B (eds.) 1986 *The Psychology of Control and Aging*. L. Erlbaum Associates, Hillsdale, NJ

Bandura A 1977 Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* **84**: 191–215

Brandtstädter J 1998 Action perspectives on human development. In: Lerner R M (ed.) *Handbook of Child Psychology: Vol. 1. Theoretical Models of Human Development*, 5th edn. Wiley, New York, pp. 807–63

Brandtstädter J, Wentura D, Rothermund K 1999 Intentional self-development through adulthood and later life: Tenacious pursuit and flexible adjustment of goals. In: Brandtstädter J, Lerner R M (eds.) *Action and Self-development*. Sage, Thousand Oaks, CA, pp. 373–400

Chipperfield J G, Perry R P, Menec V H 1999 Primary and secondary control enhancing strategies: Implications for health in later life. *Journal of Aging and Health* **11**: 517–39

Dweck C S 1999 *Self-theories: Their Role in Motivation Personality and Development*. Psychology Press/Taylor and Francis, Philadelphia, PA

Dweck C S, Elliott E S 1983 Achievement motivation. In: Hetherington E M (ed.) *Handbook of Child Psychology: Vol.*

4. *Socialization Personality and Social Development*. Wiley, New York, pp. 643–91
Gollwitzer P M, Kinney R F 1989 Effects of deliberative and implemental mind-sets on illusion of control. *Journal of Personality and Social Psychology* **56**: 531–42
Heckhausen H 1991 *Motivation and Action*. Springer-Verlag, New York
Heckhausen J 1999 *Developmental Regulation in Adulthood: Age-normative and Sociostructural Constraints as Adaptive Challenges*. Cambridge University Press, New York
Heckhausen J 2000 Evolutionary perspectives on human motivation. *American Behavioral Scientist* **43**: 1015–29
Heckhausen J, Schulz R. 1995 A life-span theory of control. *Psychological Review* **102**: 284–304
Krause N 1987 Chronic strain, locus of control, and distress in older adults. *Psychology and Aging* **2**: 375–82
Lachman M E 1986a Personal control in later life: Stability, change and cognitive correlates. In: Baltes M M, Baltes P B (eds.) *The Psychology of Control and Aging*. Erlbaum, Hillsdale, NJ, pp. 207–36
Lachman M E 1986b Locus of control in aging research: A case for multidimensional and domain-specific assessment. *Psychology and Aging* **1**: 34–40
Langer E J, Rodin J 1976 The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting. *Journal of Personality and Social Psychology* **34**: 191–8
Lefcourt H M (ed.) 1981 *Research with the Locus of Control Construct: Vol. 1. Assessment Methods*. Academic Press, New York
Little T D, Oettingen G, Stetsenko A, Baltes P B 1995 Children's action control beliefs about school performance: How do American children compare with German and Russian children? *Journal of Personality and Social Psychology* **69**: 686–700
Menec V H, Chipperfield J G 1997 The interactive effect of perceived control and functional status on health and mortality among young-old and old-old adults. *Journal of Gerontology: B—Psychology* **52**: P. 118–26
Mirowsky J 1995 Age and the sense of control. *Social Psychology Quarterly* **58**: 31–43
Rothbaum F, Weisz J R, Snyder S S 1982 Changing the world and changing the self: A two-process model of perceived control. *Journal of Personality and Social Psychology* **42**: 5–37
Rotter J B 1990 Internal versus external control of reinforcement: a case history of a variable. *American Psychologist* **45**: 489–93
Schulz R, Hanusa B H 1978 Long-term effects of control and predictability-enhancing interventions: Findings and ethical issues. *Journal of Personality and Social Psychology* **36**: 1194–201
Schulz R, Heckhausen J 1996 A life-span model of successful aging. *American Psychologist* **51**: 702–14
Schulz R, Heckhausen J 1999 Aging, culture and control: Setting a new research agenda. *Journal of Gerontology: Psychological Sciences* **54B**: P139–45
Seligman M E P 1975 *Helplessness: On Depression Development and Death*. Freeman, San Francisco
Skinner E A 1995 *Perceived Control Motivation Coping*. Sage, Thousand Oaks, CA
Skinner E A 1996 A guide to constructs of control. *Journal of Personality and Social Psychology* **71**: 549–70
Skinner E A, Chapman M, Baltes P B 1988 Control, means–ends, and agency beliefs: A new conceptualization and its measurement during childhood. *Journal of Personality and Social Psychology* **54**: 117–33
Skinner E A, Zimmer-Gembeck M J, Connell J P 1998 Individual differences and the development of perceived control: Introduction and overview. *Monographs of the Society for Research in Child Development* **63**: 1–220
Timko C, Moos R H 1989 Choice, control, and adaptation among elderly residents of sheltered care settings. *Journal of Applied Social Psychology* **19**: 636–55
Weiner B 1985 An attributional theory of achievement motivation and emotion. *Psychological Review* **92**: 548–73
Weiner B 1986 *An Attributional Theory of Motivation and Emotion*. Springer, New York
Weisz J R 1983 Can I control it? The pursuit of veridical answers across the lifespan. *Life-span Development and Behavior* **5**: 233–300
White R W 1959 Motivation reconsidered: The concept of competence. *Psychological Review* **66**: 297–333
Wrosch C, Heckhausen J 1999 Control processes before and after passing a developmental deadline: Activation and deactivation of intimate relationship goals. *Journal of Personality and Social Psychology* **77**: 415–27

J. Heckhausen

Control Beliefs: Health Perspectives

Control refers to the ability to influence what is happening or what will happen. *Beliefs* about control in the context of health refer to the thoughts (or cognitions) an individual has regarding the ability to influence health behavior, health status (or other health outcomes), or health care. Control beliefs are one of a number of determinants of health behavior and health outcomes. A person's health status can be influenced by and can influence his or her control beliefs. Individuals who believe they have control over their health behavior are more likely to engage in healthy behavior; thus, they are more likely to have better health outcomes (see *Health Behaviors*). Better health outcomes are associated with more positive control beliefs. Also relevant are beliefs about control over health-care delivery. Believing that one can influence one's own health care is a primary determinant of satisfaction with health care. Perceived control over one's health behavior, health outcomes, and/or health care is different from *desire* for control over these aspects. It is assumed that most individuals desire control over their health behavior and health outcomes; no such assumption is made regarding control over how health care is delivered.

1. Locus of Control

'Locus' refers to the *place* where control (over outcomes) resides—either 'internal' to the individual (i.e., based on his or her own behavior or enduring characteristics) or 'external' to the individual (e.g., due

to other people's behavior, fate, luck, or chance). Although it was once thought that *internal* and *external* beliefs were at opposite ends of a continuum (Rotter 1966), it is now understood that these two belief orientations are independent of one another. That is, a person might simultaneously hold internal and external beliefs about the locus of control of a given phenomenon (e.g., his or her health status; Wallston 1989, 2001). For example, a person newly diagnosed with lung cancer might attribute the diagnosis to such internal factors as a lifetime of smoking cigarettes in addition to poor dietary and exercise habits. At the same time, the person might attribute the lung cancer to such external causes as 'bad genes' (inherited from one's parents), punishment from God, pollution in the environment, and/or poor luck. Furthermore, one's locus of control orientation about the cause of a health problem (its etiology) is unrelated to one's beliefs about responsibility for the course of the problem (i.e., whether it will get better or worse). It is possible to attribute the etiology of a disease to external causes (e.g., a virus) and to adopt an internal orientation to dealing with the condition once diagnosed (e.g., learn everything one can about the condition and its treatment). Many patients with a medical condition believe that health-care professionals either fully or partially control what happens to their health status.

Better health outcomes are typically associated with more internal beliefs about control of one's health. However, an internal health locus of control orientation does not necessarily signify that one attributes only good health outcomes to one's own behavior. For some individuals, having 'internal' beliefs may mean blaming oneself for one's poor health outcomes. Being 'responsible for' one's outcomes does not always translate into being in control of those outcomes. Furthermore, internal beliefs about one's health are only weakly predictive of engaging in healthy behavior or avoiding health risks. Internal beliefs are somewhat more predictive of healthy behaviors for people who highly value having good health compared with those for whom good health is less important, but locus of control only explains a relatively small amount of variance in health behavior even for people who highly value good health (Wallston 1992). Valuing good health and believing that one's health is contingent on one's own actions are necessary, but not sufficient, determinants of health behavior.

2. *Self-efficacy, Mastery, and Competence*

Locus of control is only one type of control belief. Other psychological constructs that are similar to locus of control are *self-efficacy* (Bandura 1977), *mastery* (Pearlin and Schooler 1978), and *competence* (White 1959). Self-efficacy, or the belief that one *can do* a specific behavior in a specific situation, is much more predictive of actually engaging in that behavior in that situation than is an internal locus of control orientation (see *Self-efficacy and Health*). Both mastery and competence are more generalized constructs than self-efficacy, and they encompass control over the situation and outcomes as well as control over behavior. Individuals with a sense of self-efficacy, mastery, or personal competence (and also those with an internal locus of control orientation) generally feel very good about themselves (i.e., have high self-esteem and psychological well-being) and are receptive to learning about and engaging in new behaviors. An internal locus of control orientation coupled with a high degree of self-efficacy, mastery, or personal competence is a potent resource for helping individuals cope with health-related stressors.

3. *Perceived Behavioral Control*

Ajzen (1988) introduced the construct 'perceived behavioral control' into his theory of planned behavior as a determinant of both behavioral intention and of the behavior itself. On a conceptual basis, perceived behavioral control is similar to self-efficacy—both constructs refer to the person's belief that the behavior in question is under his or her control—but, operationally, perceived behavioral control is often assessed by the ease or difficulty of the behavior (e.g., 'I find it difficult to exercise three times a week'), while self-efficacy is operationalized by the individual's confidence in being able to carry out the behavior in the face of extenuating circumstances (e.g., 'I am confident that I can exercise three times a week even when I am away on vacation').

Like Ajzen's theory of planned behavior, most psychosocial theories of health behavior (e.g., the health belief model (Rosenstock 1990), the protection motivation theory (Prentice-Dunn and Rogers 1986), and the health-action process approach (Schwarzer 1999) have incorporated perceived behavioral control as a major determinant of intention to engage in a health behavior and/or as a determinant of actually engaging in the behavior, but these last three theories explicitly label this construct as 'self-efficacy.' Regardless of the label attached to the construct, believing that one has control over a behavior is highly associated with actually doing that behavior.

4. *Perceived Situational Control*

Beliefs about control encompass control over one's situation, as well as control over one's behavior and outcomes. One example of a situation is being a patient in a clinical facility. Does the patient know why he or she is there? Does the patient know what will happen to him or her, and why? What it will feel like? Does the patient have any choice over what will

happen? The answers to these questions partially determine the patient's degree of perceived situational control. In turn, the more control the patient believes he or she has in that situation, the less distressed and more satisfied the patient will be, and the more likely the patient will be able to participate in his or her own care, including adhering to situational demands. This is true regardless of how much control the patient wants in that situation (cf. Wallston 1989).

5. *Helplessness*

Helplessness is the belief that there is nothing that anyone can do to improve a bad situation (such as being diagnosed with an illness). In many ways, then, helplessness is a belief that control over the situation or its outcomes is impossible. Like all beliefs, helplessness is learned (Seligman 1975). Learned helplessness is analogous to a chance locus of control orientation (i.e., the belief that one's outcomes are due to fate, luck, or chance). Helplessness beliefs can be either *universal* (i.e., there is nothing that *anyone* can do) or *personal* (i.e., there is nothing that *I* can do). Either type of helplessness belief is associated with motivational, behavioral, and/or affective deficits. People with learned helplessness are not inclined to learn or engage in new, potentially effective behaviors, and they exhibit higher than normal levels of anxiety and depression.

6. *Assessing Control Beliefs*

A number of well-developed, standardized instruments exist for the purpose of assessing individual differences in control beliefs applicable to health and health-care settings (Wallston 1989, in press). Some of these measures (e.g., Pearlin and Schooler's mastery scale) are general, personality-like instruments, consisting of a single dimension. Others, such as the health locus of control scales developed by the author and his colleagues, are multidimensional and can be adapted to assess control over one's health status in general or control over specified health conditions (e.g., cancer or diabetes). Researchers have also had some success in asking one or two pointed questions about subjects' perceptions of control (see Wallston 2000 for examples), but the psychometric soundness of this approach is questionable. Measures of desire for control are only weakly correlated with measures of perceived control

7. *Altering Control Beliefs*

Because control beliefs are learned, they can easily change as a function of life events (e.g., receiving a new medical diagnosis), or they can be systematically modified through the application of some form of cognitive-behavioral therapy. Shapiro and Astin (1998) have developed an integrated approach to psychotherapy, health, and healing that they term 'control therapy.' Their therapeutic approach is based on a unifying theory of human control 'which is organized around three broad postulates: (a) gaining and maintaining a sense of control is a major motivational force across the human life cycle; (b) there are both higher and lower levels of control-related goals, desires, and strategies by which people seek to gain a sense of control; and (c) there are individual differences with respect to how and why control is sought' (Shapiro and Astin 1998, p. 20).

See also: Depression, Hopelessness, Optimism, and Health; Explanatory Style and Health; Health Behavior: Psychosocial Theories; Health: Self-regulation; Learned Helplessness; Self-efficacy; Self-efficacy and Health

Bibliography

Ajzen I 1988 *Attitudes, Personality, and Behavior*. Dorsey Press, Chicago

Bandura A 1977 Self-efficacy: toward a unifying theory of behavior change. *Psychological Bulletin* **84**: 191–215

Pearlin L I, Schooler C 1978 The structure of coping. *Journal of Health and Social Behavior* **19**: 2–21

Prentice-Dunn S, Rogers R W 1986 Protection motivation theory and preventive health: Beyond the health belief model. *Health Education Research, Theory and Practice* **1**, 153–61

Rosenstock I M 1990 The health belief model: explaining health behavior through expectancies. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research, and Practice*. Jossey-Bass, San Francisco, pp. 39–62

Rotter J B 1966 Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs* **80**: 1–28

Schwarzer R 1999 Optimism, goals, and threats: How to conceptualize. Self-regulatory processes in the adoption and maintenance of health behaviors. *Psychol Health*, **3**: 759–66

Seligman M E P 1975 *Helplessness*. W. H. Freeman, San Francisco

Shapiro D H Jr, Astin J A 1998 *Control Therapy: an Integrated Approach to Psychotherapy, Health, and Healing*. Wiley, New York

Wallston K A 1989 Assessment of control in health care settings. In: Steptoe A, Appels A (eds.) *Stress, Personal Control and Health*. Wiley, Chichester, UK, pp. 85–105

Wallston K A 1992 Hocus-pocus, the focus isn't strictly on locus: Rotter's social learning theory modified for health. *Cognitive Therapy and Research* **16**: 183–99

Wallston K A 2001 Conceptualization and operationalization of perceived control. In: Baum A, Revenson T A, Singer J E (eds.) *The Handbook of Health Psychology*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 49–58

White R W 1959 Motivation reconsidered: the concept of competence. *Psychological Review* **66**: 297–333

K. Wallston

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Control Variable in Research

An important goal of empirical research is to support causal assertions about the effects of explanatory variables on outcome variables (see *Causation (Theories and Models): Conceptions in the Social Sciences*; *Explanation: Conceptions in the Social Sciences*). Extraneous variables that influence the outcome represent potential confounds whose effects must be eliminated before causal assertions regarding the explanatory variable can be made. In its broad sense, the term 'control' refers to a researcher's systematic attempts to eliminate the confounding effects of extraneous variables associated with sampling, measurement, experimental manipulation, randomization, and statistical adjustment. The term 'control variable' is used most commonly in the narrow context of statistical adjustment of potential confounds.

The current presentation of 'control variables' is guided by three broad tenets. (a) Strategies of statistical control and its interpretation depend on an understanding of the nature of the relationship between extraneous variables and the explanatory and outcome variables. The term control variable is best understood within this broader explanatory framework. (b) The nature of potential confounds and therefore the need for their statistical control depends on the specific research design. The need for statistical control increases as the inherent controls imposed by the research design diminish. (c) The potential confounding variable may be observed or unobserved. Although most methodological and empirical literature on statistical control focuses on the control of observed variables, this article is organized around prototypical methods of controlling for both observed and unobserved sources of extraneous variability. This approach highlights the importance of a causal framework for understanding statistical control and also pinpoints the most common cause of the failure of observed variables methods, i.e., the presence of unobserved sources of confound.

1. *Controlling for Observed Variables*

Several methods of control are available when it can be assumed that the potential confounding variables are measured including regression, matching, and propensity score. Regression (Pedhazur and Pedhazur 1991) or analysis of covariance (ANCOVA; Hutiema 1980) based approaches are the most commonly used methods of statistical control. The purpose and interpretation of regression adjustment depends on the specific research design employed.

1.1 *Regression and ANCOVA Based Approaches*

Regression of a dependent variable Y on an independent variable X provides an estimate of the causal effect of X on Y under the critical assumption that all omitted variables C that influence Y are uncorrelated with X. If an omitted extraneous variable C that influences Y is also correlated with the independent variable X, then the regression parameter is biased as a measure of direct effect of X on Y. In this case, the effect of explanatory variable X is confounded with the extraneous variable C. The true direct effect of X on Y (a in Fig. 1) may be obtained by measuring and including C in the model. The key assumption made in regression based approaches is that all extraneous variables are measured without error and included in the analysis.

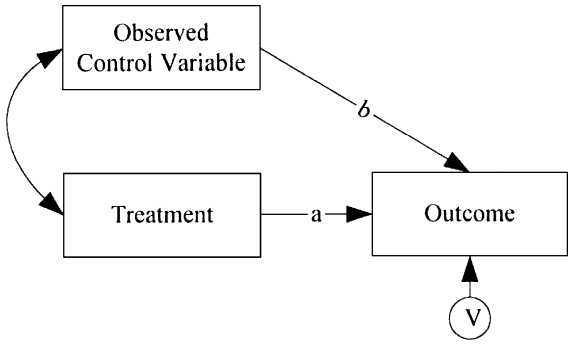

Figure 1
Path '*a*' shows the direct effect of the treatment on outcome after controlling for the effect '*b*' of the observed covariate

1.1.1 Regression adjustment in randomized experiments. Randomized experiments (Cook and Campbell 1979; see *Laboratory Experiment: Methodology*; *Random Assignment: Implementation in Complex Field Settings*) offer the most robust method for making causal inferences and imply two powerful means of control. These are: (a) the independent or explanatory variable is controlled by direct manipulation. (b) All other sources of extraneous variability are controlled by randomization or chance. Random assignment, *if successful*, makes experimental and control groups probabilistically equal with respect to all potential confounding variables and hence minimizes bias due to initial differences among treatment conditions. As a result, extraneous variables are expected to be uncorrelated with the manipulated variable or the treatment assignment and their exclusion does not bias the estimates of treatment effects. On the other hand, including such variables in analysis reduces the unexplained within-group variability in the outcome, thus increasing the power for detecting treatment effects. Pretest measures of the dependent variable are frequently used for this purpose. It is important to note that randomization may fail to make the experimental conditions equal with

respect to various confounders. In addition, noncompliance and attrition may render various treatment conditions noncomparable.

1.1.2 Regression adjustment in quasi-experimental and nonexperimental designs. In quasi-experimental designs, the independent variable is manipulated in the absence of random assignment (see *Nonequivalent Group Designs*). In these designs, subjects may be nonrandomly assigned to or may self-select themselves into different conditions. For example, in a study investigating the effects of remedial education on achievement, students who are expected to benefit the most may be assigned by the researcher into the treatment condition (see *Selection Bias, Statistics of*). As a result, the groups may not necessarily be equal at baseline with respect to various observed and unobserved characteristics that may influence the outcome of the treatment. In nonexperimental settings, the researcher may wish to infer causation from the observed association between a hypothesized independent variable and an outcome, e.g., the effect of smoking on lung cancer (see *Selection Bias, Statistics of*).

The purpose of regression adjustment in nonexperimental designs is to make the experimental conditions statistically equal with respect to various observed characteristics that may also influence the outcome. The success of regression based approaches in quasi-experimental and observational research depends on the plausibility of the assumption that all potential confounding variables are measured without error and included in the analysis. This assumption may not be tenable in most quasi-experimental settings where self-selection or assignment to treatment may be based on unmeasured expectations about outcome under treatment and control conditions (see *Selection Bias, Statistics of*).

The observed characteristics that are typically controlled to include the pretest measure of the outcome variable and background characteristics such as gender, age, and socioeconomic status. The control variable frequently serves as a proxy for some other unmeasured theoretical variable of interest. This method of controlling for theoretical variables is a common practice in applied settings and leads to a number of additional problems. Proxy variables, unlike the measured indicators of a construct, contain *both* random error as well as unspecified sources of systematic errors. The presence of random measurement error in regressors leads to biased estimates of the regression parameters. On the other hand, the presence of systematic variability other than the hypothesized construct introduces new and unspecified sources of bias. As a result, the use of background characteristics as proxies in evaluation research may result in contradictory findings regarding the effects of the explanatory variable in question depending on the specific set of covariates included in the analysis.

1.1.3 Control variable: validity and differential treatment effectiveness. The use of control variables for statistical adjustment is motivated primarily by a desire to increase the internal validity of the study (see *Internal Validity*). An alternative way of eliminating confounding due to extraneous variables is to include only those individuals at a specific level of the confounding variable. For example, if ethnicity and gender are related to the treatment assignment and to outcome, the researcher may choose to include only white males in the study. Such control by exclusion limits the generalizability of the findings to the population actually included in the study. In contrast, the regression based approach allows generalization across all levels of the controlled covariate present in the sample if the treatment effects are the same within each level of the covariate (see *External Validity*). In other words, the possibility of differential treatment effects across levels of the covariate must be ruled out before the results can be generalized. For example, a specific intervention for treating childhood aggression may be more effective at higher levels of initial aggression. In this case, the initial level of aggression could be included in the analysis along with an interaction term with the manipulated variable to test if the magnitude of the treatment effect depends on the level of the observed covariate. When the interaction between the independent variable and a covariate is significant, the covariate is said to moderate the effect of the independent variable on the outcome (see *Moderator Variable: Methodology*; *External Validity*). In this situation, the arbitrary distinction between an explanatory and a control variable begins to blur. The researcher must now explore the mechanism of differential treatment effectiveness and ascribe proper causal status to both the covariate and the explanatory variable.

1.2 Propensity Score

In contrast to the regression approaches, the propensity score approach explicitly models the treatment selection/assignment process using the *observed* background characteristics (see *Selection Bias, Statistics of*). The *propensity score* $P(\mathrm{T}|\mathrm{C})$ is the probability that an individual with a given set of background characteristics C will be assigned to a particular treatment group (Rosebaum 1995). If C also has an effect on the outcome variable, then conditional on the propensity score, the treatment assignment will be random with respect to the outcome. The propensity score captures the collective influence of all observed covariates on

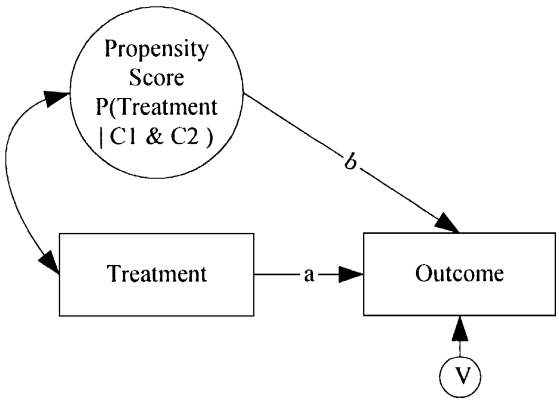

Figure 2
Path '*a*' shows the direct effect of the treatment on outcome after controlling for the effect '*b*' of the Propensity Score estimated using two observed variables C1 and C2 (not shown)

treatment assignment. In other words, by controlling for the propensity score, it is possible to control for all observed characteristics that determine treatment assignment. This can be achieved by including the propensity score in the regression equation (Fig. 2) or by matching on propensity score. The propensity score may be estimated using logit or probit regression by regressing observed treatment assignment on the set of covariates and their interactions.

2. *Controlling for Unobserved Variables*

Both the regression and propensity score methods make a strong assumption that all potential sources of confound are measured and included in the analysis. It is likely that selection into or exposure to a treatment depends on a subject's expectations regarding the outcome. For example, an individual's implicit beliefs regarding expected outcomes in the treatment and control groups may influence their decision to participate in a particular treatment as well as the eventual outcome. Even in randomized experiments, treatment compliance and attrition may be related to such unobserved variables. In this situation, observed background characteristics may not contain relevant information regarding a subject's decision process. As a result, methods of control based on observed variables are likely to produce biased estimates of treatment effects. Rosebaum (1995) provides methods of assessing the sensitivity of the results to potential hidden biases in the context of observational studies. The problem of isolating true causal effects from observed association in the presence of latent confounds can be thought of as an identification problem. If one is willing to make certain assumptions about the

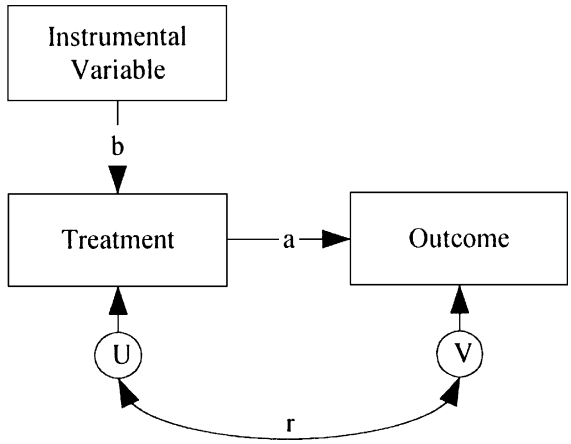

Figure 3
Path '*a*' shows the direct effect of the treatment on the outcome independent of the correlation (*r*) between the residuals for the treatment (*u*) and the outcome (*v*). This effect is identified by the presence of an instrumental variable that is presumed to influence the outcome exclusively via its effect on the treatment assignment (path *b*)

nature of treatment effects under various conditions, it is possible to express treatment effects using inequalities or bounds (Manski 1995).

Alternatively, the identification issue can be resolved by measuring additional variables that are causally related to the explanatory and outcome variables but are *assumed* to be independent of the unobserved confounds. In other words, the presence of such variables allows the true causal effects of an independent variable to be isolated from the unobserved sources of confounds. Such variables can help identify the true causal effects under two circumstances: (a) if the effects of these variables on the outcome are entirely mediated by the independent variable or (b) if the variable mediates the effect of the independent variable on the outcome.

2.1 *Instrumental Variable Method*

The instrumental variable approach for controlling unobserved sources of variability is the *mirror opposite* of the propensity score method for controlling observed variables (Angrist et al. 1996, Winship and Morgan 1999). Unlike an observed control variable, an instrumental variable is *assumed not* to have any *direct effect* on the outcome. Instead, the instrumental variable is thought to influence only the selection into the treatment condition. In other words, the effect of the instrumental variable on the dependent measure is entirely *mediated* via its effect on treatment assignment (*b* in Fig. 3). This condition is also known as *exclusion*

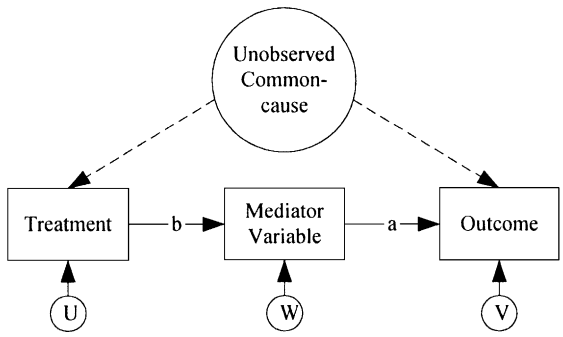

Figure 4
The direct effect of the mediator on outcome is represented by '*a*' and that of the treatment assignment on the mediator is represented by '*b*'. The unobserved common-cause of treatment assignment and outcome is assumed not to influence the mediator. As a result, the mediated effect of the treatment on the outcome ($a \times b$) is independent of the confounding due to plausible common cause

restriction (see *Endogeneity: Methodology*). If the independent variable were regressed on the instrumental variable, the residual would contain all unobserved sources of variability that determine treatment assignment and also influence the outcome variable (represented by the correlation *r* in Fig. 3). As a result, the existence of an instrumental variable *identifies* or *isolates* the average direct effect (*a* in Fig. 3) of the treatment on the outcome independent of the unobserved sources of variability. The success of this strategy rests on the reasonableness of the assumption of exclusion restriction.

2.2 Mediator Variable Method

The effects of unobserved confounds may be isolated in a radically different fashion by modeling the causal mechanism by which the treatment influences the outcome (Pearl 2000). In this approach, the goal is to find an observed variable that *mediates* the effect of the independent variable on the outcome *but is unrelated to the unobserved common cause* (see *Mediator variable*; Fig. 4). For example, in the past cigarette companies have claimed that the relationship between smoking and lung cancer is spurious: a result of a gene (common-cause) that leads to craving for tobacco and also causes lung cancer. If such a common cause in fact existed, it is still possible to isolate the independent effect of smoking on lung cancer by identifying a variable that mediates the effects of smoking on cancer but is *not influenced by the hypothesized common-cause*. For example, tar in lungs is one such variable known to cause cancer which accumulates in lungs as a result of smoking, passive smoke inhalation, and environmental pollution. It is reasonable to argue that the level of tar in lungs is not caused by the common-gene, especially among the passive smokers. The existence of such a mediator variable can statistically identify the effect of the independent variable on the dependent variable ($a \times b$ in Fig. 4) even in the presence of unobserved sources of confounding. It is to be noted that the instrumental and mediator variables are not control variables themselves. Instead, the use of these variables identifies true treatment effects by isolating the confounding effects of other unobserved sources of variability.

Bibliography

Angrist J D, Imbens G W, Rubin D D 1996 Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**: 444–72

Cook T D, Campbell D T 1979 *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Rand McNally College Pub. Co, Chicago

Hutiema B E 1980 *The Analysis of Covariance and Alternatives*. Wiley, New York

Manski C F 1995 *Identification Problem in the Social Sciences*. Harvard University Press, Cambridge, MA

Pearl J 2000 *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK

Pedhazur E J, Pedhazur S L 1991 *Measurement, Design and Analysis: An Integrated Approach*. Lawrence Erlbaum Associates, NJ

Rosebaum P R 1995 *Observational Studies*. Springer-Verlag, New York

Winship C, Morgan S L 1999 The estimation of causal effects from observational data. *Annual Review of Sociology* **25**: 659–707

P. D. Mehta

Control: Social

An important concept of US sociology at the beginning of the twentieth century, social control referred to those aspects of social life contributing to the social order. It was used to answer a perennial question in sociology: How is it that social order is possible? In the 1970s, a shift in meaning became noticeable. Sociologists and historians studying state institutions designed to deal with deviants, like prisons and mental health facilities, gave it a pejorative connotation: Social control came to mean coercion, manipulation, labeling, and stigmatization. At the beginning of the twenty-first century, the term is more often used in criminology than in sociology. Two different meanings can be found. First, in control theories, it means informal social controls imposed on juveniles within the family and the school. Second, it is sometimes used

as a synthetic term to include all the actions and institutions intended to curb crime.

1. The Original Meaning in Sociology

In the eyes of the early sociologists, the question of how society curbs individual passions and interests so as to perpetuate itself in peace, was of paramount importance. Social control 'referred to the capacity of a society to regulate itself according to desired principles and values' (Janowitz 1975, p. 82). In 1901, E. Ross published a book called *Social Control. A Survey of the Foundation of Order* which was influential in putting the concept at the center of US sociology. The question he undertook to answer was: 'How is it that people come to live together and cooperate with that degree of harmony we see about us?' A kind of natural order emerges first from people living together with their capacities for sympathy, sociability, sense of justice, and resentment. Second, society develops a certain measure of control over the individual. The means of control surveyed by Ross (1901) cover a whole range of social practices and institutions: public opinion, law, belief, social suggestion, education, religion, personal ideals, ceremony, art, social valuation, etc. According to Ross, social control includes all those aspects of social life inducing conformity amongst individuals and contributing naturally to the social order.

Ross shared the preoccupations of Emile Durkheim, who was also trying to discover the roots of social harmony. The French sociologist used the term 'social constraint' to refer to the influence of the 'collective conscience' on individuals. The pressures of society on the individual does not take only the form of external constraints, it also creates a sense of moral obligation to accept duties and to obey the rules (Durkheim [1893] 1960, 1895, 1923).

Following Ross, sociologists of the Chicago School, like Park, Burgess, and Thomas, continued to place social control at the center of sociology. 'All social problems turn out to be problem of social control' (Park and Burgess 1921, p. 785).

While developing the concept of social control, the early sociologists had another agenda besides explaining social order. They wanted to demonstrate the capacity of sociology to explain social facts more completely than economics or plain common sense. They first wanted to show the limitations of Adam Smith's analysis of the natural harmony of interests fostered by the famous invisible hand of the market. Economic self-interest cannot fully account for the existence of the social order (Pitts 1968, Janowitz 1975). They also wanted to debunk the common sense assumption that order resides in the hands of the police, judges, prison guards, and the like. Cooperation, harmony, and organization do not rest predominantly on coercion nor on utilitarian calculations but on subtle pressures, expectations, and informal sanctions in the family, in the peer group, the workplace, the local community, etc. However, if social control is neither market regulation, nor state control, nor coercion, we are left with a notion referring essentially to informal social control. But if we stick to the ambition of taking into account most of the social order, how is it possible to ignore the utilitarian and coercive ways of achieving it?

Another problem emerges: by its very nature, social control is a static concept. When it occupies a central place, sociology becomes the science of social order, unable to analyse social disorder or social change. During the 1920s and 1930s, 'sociologists became so caught up with the source of cohesion that they neglected to examine the source of change' (Rothman 1983, p. 108).

Those problems cooled the enthusiasm of sociologists for the concept. Some restricted it to the control of deviance. In many contemporary textbooks and handbooks of sociology, one finds a chapter on deviance and social control (e.g., Scull 1988). After the 1970s, the few sociologists attempting to write on general theory do not make more than a passing mention of the term. The search for the foundation of social order is no longer conceived in terms of the control of society on individuals. A major sociologist like Coleman (1990) explains the social order with the concepts of interdependence; transaction (including exchange, bribes, threats, promises, and resource investment); authority relations; rights; and norms.

2. The Skeptic Reaction

In the 1960s and 1970s, sociologists critical of the rhetoric and reality of the institutions dealing with deviants (mainly prisons and mental hospitals) gave a whole different tone to the term. Social control came to mean first those stigmatizing actions putting in motion an amplification of the deviance process, and second those bureaucracies pretending to reform and treat deviants but, in fact, indulging in manipulation and coercion.

2.1 Labeling

In labeling theory, social control refers mainly to those reactions of police officers, judges, psychiatrists, etc. which consist of arresting, condemning, classifying, and incarcerating. Somebody comes to be classified as criminal, mentally disturbed, dangerous, or deviant in one way or another, and is set apart in a prison, an asylum, or any other place of exclusion. Those reactions mark such individuals, they stigmatize them, they attach a label to their personality. Doing so, they not only modify the social identity of the individuals but also their self-perception, their way of life, and their behavior. 'The experience of being caught and publicly labelled as a deviant … has important conse-

quences for one's further social participation and self-image' (Becker 1963, p. 31). 'Treating a person as though he were generally rather than specifically deviant produce a self-fulfilling prophecy. It sets in motion several mechanisms which conspire to shape the person in the image people have of him' (Becker 1963, p. 34). Instead of producing order, the attempt at social control amplifies the deviant tendencies of the labeled person (Wilkins, 1964). The arrest, the sentence, the psychiatric diagnostic, the imprisonment, or the institutionalization in a mental hospital puts individuals subjected to one of those treatments at risk of internalizing the infamous label while putting them at the margin of society. They might feel obliged to become part of a deviant group and act as one of the gang. Accordingly they might become more deviant after being labeled than before.

The labeling theory has been criticized mainly on two counts. The labeling process is not totally arbitrary nor completely unrelated to the pre-existing deviant behavior and characteristics of the labeled individual. More often than not, the court condemns accused people who really committed a crime, and generally, the more serious the crime, the more severe the sentence. The number of crimes committed in the past by an individual is also related to the severity of the punishment. If such is the case, the question to be asked is: Does the label create a new social identity or is it just a disclosure of a deviant identity already there?

Empirical tests of the labeling hypothesis provide an inconsistent picture, with some researchers finding a weak effect of social reaction on subsequent offending, others failing to find one. The overall impact is small. 'The preponderance of research finds no or very weak evidence of labeling effects. The more carefully the research keeps other factors constant, the less likely it is to find evidence that labeling has significant effect on criminal or deviant behavior' (Akers 1994, p. 134, see also Braithwaite 1989).

2.2 The History and Sociology of State Dealing with Deviants

A related trend of research focuses not so much on the effects of social control but on the characteristics and evolution of state institutions dealing with offenders, mental patients, and other deviants. In this context, social control is defined as 'the organized ways in which society responds to behavior and people it regards as deviant, problematic … or undesirable' (Cohen 1985, p. 1). As with labeling theory, the concept carries strong pejorative overtones. Gone are the benevolent informal controls to be replaced by coercion, the repressive state, and the manipulation used by the powerful against the lower classes. The negative overtones come also from a disenchantment with the rehabilitative ideal and from skepticism concerning the so-called reform movement with its self-serving ideologies (Rothman 1983). The historians and sociologists working in this perspective embark upon a work of demystification, confronting the official discourses with the real functioning of the social control apparatus. Psychiatrists and correctional officials pretend to treat and rehabilitate; in fact, they cannot help but keep on putting mental patients in total institutions and incarcerating offenders. Reformers speak of decarceration, decriminalization, and diversion, but in fact the old prison system remains, more crimes are created by the law, and new deviants are caught in the social control net. 'Overall, the system enlarges itself and becomes more intrusive, subjecting more and newer groups of deviants to the power of the state and increasing the intensity of control directed at former deviants' (Cohen 1985, p. 38).

This discrepancy between discourse and reality, between what officials say they do and what they do in fact, is an enlightening exercise provided that it goes beyond plain conspiracy theory.

The main contribution of authors working in this perspective is in the history of changes in punishment in Europe and the USA beginning at the end of the eighteenth century. Foucault (1975) and Ignatieff (1978, 1983) documented the following trends: (a) a sharp decline in physical punishments inflicted in public in order to inspire terror in the spectators; (b) the emergence of the prison as the dominant punishment for serious crimes; (c) the creation of the penitentiary, an institution pretending to reform, deter, and isolate prisoners not only from the outside world but also from the criminogenic influence of their fellow prisoners.

In a related vein are sociological studies of the internal functioning of mental hospitals and prisons, like *Asylum* by Goffman (1961). The thesis of the book is that institutions like prisons, mental hospitals, concentration camps, and even monasteries share some basic characteristics. They are all total institutions in which guards, nurses, doctors, etc. are given the mission and the power to control the inmates and isolate them from the outside world. In those total institutions, the dominating group organizes everyday life in minute details for maximizing their control over inmates. They impose detailed regulations so as to make the inmate's behavior most predictable. Swift punishment and degrading measures are visited upon the rebellious.

3. Social Control in Criminology

Nowadays, it is mostly in criminology that the term social control is found. It is often used in the same skeptical spirit in which the term was used by sociologists criticizing the state's apparatus of control. Some criminologists, though, remain faithful to the orientation of the early sociologists preoccupied with the foundation of social order. One aspect of this

social order which, for obvious reasons, is the focus of criminologists, is a society having a tolerable level of crime. Accordingly, social control is conceived as the means to achieve it.

3.1 Control Theory

Instead of asking the usual question, 'Why do people commit crimes?,' criminologists who developed the control theory turned the problem around and asked 'Why do people obey the rules of society?' 'Deviance is taken for granted; conformity must be explained' (Hirschi 1969, p. 10). Basically, this is the kind of questioning which was at the center of the works of Hobbes, Durkheim, and Ross.

Control theory was mostly used to explain juvenile delinquency or, more precisely, its absence. Such was the intention of Reiss (1951). He was able to show that delinquency is a result of 'a failure of personal and social control.' Of interest is his definition of social control: 'the ability of social groups to make norms or rules effective.'

It is Hirschi's (1969) formulation of control theory that is the most influential in criminology. His basic proposition is that 'delinquent acts result when an individual's bond to society is weak or broken' (p. 16). For adolescents, the bond materializes in the family, at school, and with their peers. This bond is made of four elements: attachment, commitment, involvement, and belief. Attachment means the close emotional ties to parents, school teachers, and peers, and caring about their expectations. Commitment is the investment in conformity, in education, and in a future career that would be jeopardized by violating the law. Involvement is the time spent in legitimate activities, like studying, taking part in sports, or leisure activities in the family. Belief refers to the adhesion to the rules of society, to the conventional values, to the idea that law has a moral value and should be obeyed. Hirschi provided the measures for those elements of the bond and a persuasive demonstration that the stronger young people's ties to society (in fact, the family and the school) are, the fewer delinquent acts they report.

Later research found many confirmations of the theory, especially with the attachment and involvement constructs. It is beyond contest that juveniles attached to their family and school do not tend to indulge in delinquency. Also, strong involvement, measured by educational and occupational aspirations, is related to a low level of delinquency (Akers 1994, pp. 116–20, Braithwaite 1989, pp. 28–9).

3.2 Promising Trends

Control theory refers to informal social control, leaving aside the law and other formal controls. This is not to say that the latter are neglected in contemporary criminology. On the contrary the focus is increasingly on official controls and the evaluation of their effectiveness. The revival of deterrence theory is a case in point. In its classic formulation, the deterrence hypothesis predicts that the frequency of crime will vary inversely with the certainty, celerity, and severity of punishment handed by the state. Recent researchers cannot find much deterrent effect of severity on crime, but crime rates tend to be low when the certainty of punishment (often measured by the police clearance rate) is high. It is also the case that there is much more evidence of a general effect of certainty on crime rate (general deterrence) than on the level of recidivism of punished offenders (specific deterrence) (Andenaes 1974, Gibbs 1975, Blumstein et al. 1978, Cook 1980, Nagin 1998).

However, there are other ways to limit crime than deterrence and informal controls. The state uses nonpunitive ways to keep crime at bay, like therapies to rehabilitate or reform offenders and crime prevention measures such as street lighting, surveillance of public areas, gun control, and regulations on car antitheft systems. In civil society, crime control activities are very common. Everybody devotes time, money, and effort to protect their goods against theft and to safeguard their person against aggression. People put their money in banks, lock the doors of their houses, avoid dangerous places; in some cases, they even buy guns for self-protection. Businesses and other organizations spend considerable amounts of money to protect their assets from loss or crime: They hire private guards, control access to their premises, use safes, install CCTV, etc.

All those actions and precautions—public and private, formal and informal, repressive and preventive—are clearly aimed at reducing the probability of a crime occurring. As such, their common goal is crime control. This leads us to a last definition: the social control of crime refers to all the means specifically aimed at reducing the probability or severity of crime.

The reader will note that social control is now defined by its intention or aim, not by its results. By doing this, we follow Gibbs (1989, pp. 23–4) who criticizes the sociological concept for belittling the intentional quality of social control. In its common use, the term 'control' conveys intention: one tries deliberately to control, to direct, to influence another. Actions having the unintended effect of preventing crime no doubt exist. For example, Felson (1998) explains the sharp decline of crime rates starting in 1994 in the USA by the coming of a cashless society. People use more credit cards and the like, so they have less cash in their pockets and their houses. Having less cash to steal, the offenders become less active. In this instance, we should not speak of social control but of an unintended preventive effect of an economic evolution.

Results (more or less crimes) are important matters but should not be included in the definition of social

control. Attempts at social control, including failures, are social control. The latter's impact is not a matter of definition but of evaluation. Because of their exclusive focus on scientific evaluation, Sherman et al. (1998, p. 2) chose another path. They define crime prevention not by its intention but by its consequences. It is 'any practice shown to result in less crime than would occur without the practice.'

If social control is made of intentional actions and choices, is it possible to conceive its impact on offenders in terms of actions and choices? The potential offender—that is, the individual having the intention to commit an offence—might choose to do it despite social controls or not to do it because of them. Such an individaul is a decision maker acting under the constraints of social control.

The impact we try to have on offenders when we attempt to control crime is essentially to: (a) increase the effort of commiting crime (e.g., by target hardening or gun control); (b) increase the risks (by surveillance, punishment, burglar alarms, and the like), (c) reduce the anticipated rewards of crime (by target removal, identifying property, etc.), and (d) removing excuses used by offenders to minimize the moral opprobrium cast on crime (e.g., by rule setting and public condemnation of crime) (Clarke 1997). If potential offenders live in a well-ordered society where those impacts are achieved, they will find themselves in a radically different choice situation than in a disorganized society where social controls are erratic. Most of the time they will find commiting crime difficult, risky, unrewarding, and reprehensible. If they are minimally rational (Cornish and Clarke 1986), they will tend to look for noncriminal alternatives. This means that where and when the social controls operate reasonably well, they shape the alternatives of choice for social actors. They close most of the criminal options for of us. They attach quite negative utilities (in the economic sense) to the criminal options. To commit crime in such a situation, one needs a fair amount of greed, temerity, disregard for long-term consequences, or plain foolishness.

Logically, the more serious a type of crime is, the more it should be worthwhile controlling. In fact, we find that guardianship is lax when minor valuables are to be protected and increased when major valuables or life must be protected. Police detectives work harder investigating murders than burglaries. The positive relation between the gravity of crime and the severity of punishment, as well as between the gravity of crime and the certainty of punishment are basic facts of research on penal decision making (Gottfredson and Gottfredson 1980). This heightened pressure of social controls on the most serious crimes will give incentive to offenders to choose the least criminal option, the least severe offence in the eventuality that they persist in crime. That should explain the inverse relation one finds between the frequency of a type of crime and its severity (there are fewer murders than robberies, and fewer robberies than burglaries). Those pressures on criminal choices can be called the structuring effects of social control (Cusson 1993).

However, the lesson learned by sociologists and historians, showing that social controls often operate in an erratic manner, should not be forgotten. Formal and informal controls are not in place where they should be for a number of reasons: groups are too disorganized, resources are lacking, the acts do not follow the rhetoric. This is to mean that the quality and intensity of social control have every reason to be highly variable in space and time. In turn, this uneven quality and intensity of social control should not be unrelated to the uneven distribution of crime rates in space and time.

See also: Crime: Sociological Aspects; Crime, Sociology of; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Delinquency, Sociology of; Justice and Law; Law and Everyday Life; Law: Economics of its Public Enforcement; Law, Sociology of; Norms; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Sanctions in Political Science; Stigma, Social Psychology of; Totalitarianism

Bibliography

Akers R L 1994 *Criminological Theories: Introduction and Evaluation*. Roxbury, Los Angeles, CA

Andenaes J 1974 *Punishment and Deterrence*. University of Michigan Press, Ann Arbor, MI

Becker H 1963 *Outsiders*. The Free Press, New York

Blumstein A, Cohen J, Nagin D (eds.) 1978 *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. National Academy of Science, Washington, DC

Braithwaite J 1989 *Crime, Shame and Reintegration*. Cambridge University Press, Cambridge, UK

Clarke R V 1997 Introduction. In: Clarke R V (ed.) *Situational Crime Prevention. Successful Case Studies*. Guilderland, New York, pp. 1–44

Cohen S 1985 *Visions of Social Control*. Polity Press, Cambridge, UK

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Cook P 1980 Research in criminal deterrence: Laying the groundwork for the second decade. In: Morris N, Tonry M (eds.) *Crime and Justice*. University of Chicago Press, Chicago, Vol. 2, pp. 211–68

Cornish D B, Clarke R V (eds.) 1986 *The Reasoning Criminal*. Springer-Verlag, New York

Cusson M 1993 The structuring effects of social control. *International Annals of Criminology* **31**(1–2): 45–57

Durkheim E [1893] 1960 *De la division du travail social*, 7th edn. Presses Universitaires de France, Paris

Durkheim E 1895 *Les règles de la méthode sociologique*. Presses Universitaires de France, Paris

Durkheim E 1923 *L'éducation morale*. Presses Universitaires de France, Paris

Felson M 1998 *Crime and Everyday Life*, 2nd edn. Pine Forge Press, Thousand Oaks, CA

Foucault M 1975 *Surveiller et punir*. Gallimard, Paris

Gibbs J P 1975 *Crime, Punishment, and Deterrence*. Elsevier, New York
Gibbs J P 1989 *Control: Sociology's Central Notion*. University of Illinois Press, Urbana, IL
Goffman E 1961 *Asylum*. Doubleday/Anchor, Garden City, NY
Gottfredson M, Gottfredson D M 1980 *Decision-making in Criminal Justice: Toward the Rational Exercise of Discretion*. Ballinger, Cambridge, MA
Hirschi T 1969 *Causes of Delinquency*. University of California Press, Berkeley, CA
Ignatieff M 1978 *A Just Measure of Pain: The Penitentiary in the Industrial Revolution, 1750–1850*. Pantheon, New York
Ignatieff M 1983 State civil society and total institutions: A critique of recent social histories of punishment. In: Cohen S, Scull A (eds.) *Social Control and the State*. Basil Blackwell and Martin Robertson, Oxford, pp. 75–105
Janowitz M 1975 Sociological theory and social control. *American Journal of Sociology* **81**(1): 82–108
Nagin D S 1998 Criminal deterrence research at the outset of the twenty-first century. In: Tonry M (ed.) *Crime and Justice. A Review of Research*. University of Chicago Press, Chicago, pp. 1–42
Park R E, Burgess E W 1921 *Introduction to the Science of Sociology*. University of Chicago Press, Chicago
Pitts J R 1968 Social control: The concept. In: *International Encyclopedia of the Social Science*. Macmillan and The Free Press, New York
Reiss A Jr. 1951 Delinquency as the failure of personal and social control. *American Sociological Review* **16**(2): 196–206
Ross E A 1901 *Social Control: A Survey of the Foundation of Order*. Macmillan, New York
Rothman D 1983 Social control: The uses and abuses of the concept in the history of incarceration. In: Cohen S, Scull A (eds.) *Social Control and the State*. Basil Blackwell and Martin Robertson, Oxford, UK, pp. 106–17
Scull A 1988 Deviance and social control. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA, pp. 667–95
Sherman L W, Gottfredson D C, Mackenzie D L, Eck J, Reuter P, Bushway S D 1998 Preventing crime: What works, what doesn't, what's promising. *National Institute of Justice. Research in Brief*. US Department of Justice, Office of Justice Programs, Washington, DC
Wilkins L 1964 *Social Deviance*. Prentice-Hall, Englewood Cliffs, NJ

M. Cusson

Conventions and Norms: Philosophical Aspects

The predominant view of 'convention' is that it describes a kind of equilibrium behavior among individuals in a strategic situation, a basic feature of which is the existence of at least one potential alternative equilibrium. A convention is therefore an arbitrary but stable social regularity. 'Norm' has several meanings. It can refer to a mere statistical regularity or, in moral theory, a prescriptive rule for human behavior. In the sense pursued here, however, a norm describes a regularity of behavior among a population of individuals, the central feature of which is that most or all of the individuals approve conformity to the regularity and/or disapprove nonconformity. Thus, conventions and norms are distinct but overlapping social regularities.

1. Introduction

A few (putative) conventions will identify the topic and illustrate its importance. A common example is the regularity of driving on only one side of the road (left or right). Another is monetary currency (see Hume [1740] 1978, p. 490). The convention here is to treat a particular form of metal or paper as having value sufficient to exchange it for goods and services whose value is otherwise (without the convention) far greater. In each case, an individual wishes to follow the regularity once it arises—to avoid the danger of driving into oncoming traffic or the inefficiency of barter. Yet these conventions are arbitrary: everyone driving on the left works like everyone driving on the right; everyone accepting one paper form as currency works much like everyone accepting another.

Language is similarly thought to be conventional because different sounds could be used to refer a given thing, but to communicate successfully, an individual is better off using the sound that others use. One might object to the idea that *all* of language is a convention if a convention is thought to require a *convening*, i.e., an actual agreement. An agreement to use certain sounds as language must be expressed in language, so some language must exist prior to any agreement based convention (see Quine [1936] 1967). Lewis (1969) and others have since argued that a convention can occur without any agreement, suggesting that language may indeed be conventional.

Language provides one example of the significance of conventions. A clear understanding of conventions may illuminate a variety of other important social phenomena. Hume ([1740] 1978, pp. 477–513) contended that justice, particularly the institution of property, was based on convention (see Vanderschraaf 1998). Some theorists give conventions a central role in the theory of morality or law (see *Law and Morality: An Analytic Perspective* and *Law and Morality: A Continental-European Perspective*; *Ethics and Values*). Conventions and norms (discussed below) also explain 'spontaneous order' that arises without a centralized force such as the state. If so, conventions (or norms) might sustain some forms of cooperation in the 'state of nature' and thereby complicate or refute social contract theories that justify the state by the disorder that exists in its absence (see Skyrms 1996, Taylor 1987). Once the state exists, spontaneous order may limit the domains in which state power is necessary

(see Ellickson 1991), or contrarily, state action may be necessary to restrict or unsettle undesirable conventions (or norms) (see McAdams 1997, pp. 409–32).

2. Defining 'Convention'

2.1 Convention as Agreement or Jointly Accepted Principle of Action

In one view, a convention is a 'jointly accepted principle of action' (Gilbert 1989, p. 377). A principle of action is a rule prescribing (at least presumptively) how to behave in certain situations. An individual can adopt a principle of action unilaterally or jointly. Gilbert (1990, p. 16) gives an example of the latter: *A* chastises *B* for failing to adhere to a previous regularity they had observed (about who phones back when they are accidentally disconnected) without prior agreement. *B* acquiesces in the criticism and apologizes. By this communicative exchange, Gilbert says the parties have created a convention by 'jointly accepting' a principle of action requiring behavior they previously engaged in without convention.

Joint acceptance does not, however, explain the common intuition that conventions are in some sense arbitrary. Suppose two individuals face the prospect of a recurrent situation in which neither can survive unless each takes some specific action. On this view, if the individuals jointly accept the only principle of action that will save their lives, the principle is a convention although it is in no sense arbitrary. In addition, Gilbert's example reveals how close this understanding of convention is to the idea that conventions are agreements. Arguably *A* and *B* have implicitly agreed because each believes that *A*'s criticism impliedly asserted the existence of a mutual obligation and that *B*'s acquiescence revealed *B*'s assent to this assertion. The joint acceptance idea clarifies how agreements produce conventions, but not how conventions arise without agreement.

2.2 Conventions as Solutions to Recurrent Coordination Games

Lewis' *Convention* (1969) is the origin of an alternative view of convention, one that explicates the way in which they are arbitrary and contends that no explicit or implicit agreement is necessary. Lewis (1969, p. 42) initially modeled conventions as solutions to recurrent coordination games. A *game* is a situation in which the outcome of any one's action depends on the decisions of other individuals as well as one's own. In game theory, *players* (usually) rationally choose among alternative *strategies*. Each combination of the players' strategies produces an *outcome*, which consists of a utility index for each player (see *Rational Choice Explanation: Philosophical Aspects* and *Game Theory: Noncooperative Games*).

	C1	C2	C3
R1	1 1	0 0	0 0
R2	0 0	1 1	0 0
R3	0 0	0 0	1 1

Figure 1
A pure coordination game

Pure coordination games involve purely common interests among the players. Consider people who want to meet each other but who have failed to communicate in advance where to meet. Lewis (1969, p. 9) uses the matrix in Fig. 1 to describe such a coordination game involving two players and three possible meeting places. One player—'R' for Row-chooser—can play strategies R1, R2, or R3, which consist of going to places P1, P2, or P3, respectively. The other player—'C' for Column-chooser—can play strategies C1, C2, and C3, which consists of going to the same three places, respectively. The players choose their strategies simultaneously, not being able to observe what the other has done. Each receives more utility from meeting (1) than from not meeting (0). (The number in the lower left corner of a cell represents the utility of Row-chooser; the upper right number states the utility of Column-chooser.) Each player therefore has an interest (equal in this example) in coordinating his or her action with that of the other player.

Specifically, R wants to coordinate by playing R1 if C plays C1, R2 if C plays C2, and R3 if C plays C3. C wants to play C1 if R plays R1, C2 if R plays R2, and C3 if R plays R3. Taken together, these observations identify three pure-strategy *Nash equilibria*, outcomes in which no one player can gain by changing his or her strategy: R1/C1, R2/C2, and R3/C3. For example, R1 is the best response R can make to C1 and that same strategy—C1—is the best response C can make to R1. Thus, given R1/C1, neither player would want to switch strategies unilaterally. The same is true for R2/C2 and R3/C3. (There are also *mixed strategy*

Nash equilibria, as where R and C each go to each of the three places with a probability of one-third.)

This analysis reveals the need for *coordinated expectations*. Except by accident, players 'solve' a coordination game only if they have accurate expectations about what the other(s) will do. In the example, each player needs to *expect* the other to go to the same place, for example, P1. But if each player knows the other wishes to meet, then, except by accident, it seems that R will expect to meet C at P1 only if R believes that C expects to meet R there as well. But C will expect to meet R at P1 only if C believes that R expects C to go there. Lewis (1969, pp. 52–60) claims the need for *common knowledge*. Common knowledge of a fact x means that everyone in the relevant population believes x, believes that everyone believes x, believes that everyone believes that everyone believes x, and so on, *ad infinitum*. Except by accident, coordination seems to require that each player have common knowledge of what the other(s) will do.

Finally, consider how a coordination game can recur. Such a situation would exist if, on repeated occasions, R and C find themselves involuntarily separated in a crowd and in need of a meeting place. More generally, the question of which side of the road to drive on, what paper to use as currency, and what sounds to use to communicate, all arise repeatedly.

A convention may then be provisionally understood as a pure-strategy Nash equilibrium to a recurrent coordination problem. A convention consists of the coordinated expectations that produce the equilibrium and the regularity of behavior the equilibrium represents. Because the situation is a coordination game, the equilibrium is arbitrary in that there is at least one alternative equilibrium.

2.3 Convention as an Arbitrary Equilibrium

In several ways, Lewis' approach is narrow. First, Lewis (1969, p. 78) requires common knowledge even of the fact that there is an alternative equilibrium. This condition is not necessary to make the regularity stable; knowing the alternative might even make it less stable. The condition thus excludes an arbitrary regularity from being a convention merely because most individuals never discover, or later forget, what the alternative is. On this view, a population's language would not be conventional if the population were unaware of the possibility of other languages (see Burge 1975).

Moreover, certain aspects of a pure coordination problem seem unnecessary to the definition of convention. Lewis' final definition (1969, p. 78) generalizes beyond coordination games, but only to overcome a problem in describing the recurrent situation when the behavior is continuous rather than discrete. Lewis retains two significant features of pure coordination games, requiring that (a) 'almost everyone has approximately the same preferences regarding all possible combinations of actions,' and (b) 'almost everyone prefers that any one more conform to [the regularity], on condition that almost everyone conform to [it].' The latter point means that most individuals conditionally prefer both their own conformity *and* the conformity of others. Each requirement excludes regularities that are plausibly conventions.

	C1	C2
R1	1 5	0 0
R2	0 0	5 1

Figure 2
A simple mixed motive game

First, there is an element of coordination in many games that also involve conflict. As long as there is some common interest, one may wonder why the equilibrium outcome, when others are possible, is not a convention. For example, in Fig. 2 (sometimes called 'the battle of the sexes' based on an early illustration), each player prefers coordinating at Nash equilibria R1/C1 or R2/C2 to either other outcome. But each player also strongly prefers one equilibrium to the other and these preferences conflict. Thus, the players do not have 'approximately' the same preferences for outcomes (which they would if we replaced each 5 with, say, 1.05). Lewis would not count a solution to this game as a convention even though it were one of two equilibria and the players each prefer coordinating at an equilibrium than reaching a nonequilibrium outcome.

Another weakness in Lewis' account is his requirement that the players conditionally prefer *other* players' conformity even though the stability of a regularity depends only on players preferring their own conformity. With a mix of common and conflicting interests, players may prefer that other players deviate from any of the multiple equilibria. In Fig. 3, R1/C1 and R3/C3 are the only Nash equilibria. As in Fig. 1, the players are indifferent between these equilibria and prefer to coordinate on either than to reach other outcomes that are worse for each (R1/C3, R2/C2, and R3/C1). But unlike Fig. 1, in each of these equilibria, the players do *not* prefer that the other conform to the equilibrium because each would gain by the other playing a different strategy. At R1/C1, R

	C1	C2	C3
R1	4 5	2 7	3 1
R2	5 3	0 0	5 2
R3	2 4	3 7	4 5

Figure 3
Adapted from Kornhauser 1996, p. 209

prefers that C plays C2 and C prefers that R plays R2. The same is true at R3/C3.

Lewis (1969, pp. 45–6) acknowledges the point with this example: suppose those planning to attend a party wish to wear the same style of clothing that others wear. But suppose many 'nasty' people prefer that many others fail to conform to the regularity and thereby suffer embarrassment. Because these nasty people do not conditionally prefer that most others conform, Lewis would not consider it a convention even if, based on coordinated expectations, everyone arrives at the party dressed in the same style. The more common intuition is that, regardless of nasty preferences, a stable and arbitrary clothing regularity is a convention (see Miller 1986, p. 123).

These examples suggest dispensing with the requirements that players have approximately the same preferences for all outcomes and conditionally prefer that others conform. A convention may then be broadly defined as *the coordinated expectations that sustain a pure-strategy Nash equilibrium, in circumstances where multiple pure-strategy equilibria are possible, and the behavioral regularity that the equilibrium represents*. Some would add stability conditions that narrow the definition. Others would expand the equilibrium concept (see Vanderschraaf 1995) and possibly include mixed strategy equilibria where individuals play different pure strategies, each with some probability.

On such a broad view, many more regularities are conventions. For example, in an iterated prisoner dilemma game, there can be multiple equilibria, some of which involve strategies of contingent cooperation. Contingent cooperation can then be conventional even though the players do not have approximately the same preferences for all outcomes. Iterated 'hawk/dove' games may produce conventions even though players do not always prefer others to conform (see Bicchieri 1993, Sudgen 1986). The definition still excludes behavioral regularities (a) that do not arise from situations of strategic interaction (e.g., breathing and sleeping) and (b) that arise in games with only one equilibrium (e.g., a one-shot prisoner's dilemma).

3. *How Conventions Emerge*

The simplest way to produce the coordinated expectations necessary for a convention is by *agreement*. When the parties have sufficiently (not necessarily purely) common interests, there is no need to enforce the agreement. If C and R desire to meet, the agreement to meet at P1 may give each a reason to expect the other to go to P1. Neither has a reason to defect from the agreement. But the game theory account of convention purports to explain their emergence without agreement. Some conventions—language and other forms of spontaneous order—seem to require such an origin. How could people form coordinated expectations without agreement?

Much of game theory addresses how players reach equilibrium (see *Rational Choice Explanation: Philosophical Aspects* and *Game Theory: Noncooperative Games*). The issue is complex and unsettled. It is often quite difficult to demonstrate that players will reach any equilibrium using the common assumption that players are perfectly rational. In a single instance of a coordination game, for example, there is no purely rational basis for expecting a player to select one particular equilibrium outcome. Game theorists believe, however, that players might coordinate on a 'salient' equilibrium or, if the game is recurrent, that players' strategies might evolve to an equilibrium.

Empirical data suggests that people can solve coordination problems without agreement. Schelling (1960, pp. 54–8) experimented with such problems and found that the psychological 'prominence' of certain solutions provided the players 'focal points' for coordinating. For example, when the game was selecting a single time of day to meet another, most players selected noon; when asked to select the same positive number, two-fifths selected 1; and when asked to pick a meeting spot on a map, most players selected the only unique feature on the map (for similar results, see Mehta et al. 1994). Apparently, people sometimes solve coordination problems by selecting what each believes the other is most likely to perceive as unique. Lewis (1969, pp. 35–6) calls this psychological feature *salience*. As he observes, what is salient need not be the most desirable, but may even be the least desirable equilibrium (but is still more desirable for each than noncoordination).

The empirical effectiveness of salience seems sufficient to demonstrate that conventions can arise

without explicit or implicit agreement. But salience itself is difficult to analyze. Salience does not seem to work by rationality alone. Knowing that one equilibrium strategy 'sticks out' from the rest does not justify expecting other players to select it unless one knows that those players favor salient outcomes, which is not required by rationality. Also, rationality alone does not explain what *is* salient, that is, it does not explain how to select among various dimensions by which to evaluate uniqueness. Similarly, in recurrent games, *precedent* may supply a means of coordinating, but the selection of precedent requires something beyond rationality. Before expectations are coordinated, there is always uncertainty about whether or which past behavior is a precedent for the present behavior. There are many dimensions from which to evaluate whether or which past incident 'sticks out' as being like the present one (see Sugden 1998, Gilbert 1989).

For recurrent games, theorists increasingly offer evolutionary accounts to explain the emergence of an equilibrium. The evolution in these models typically does *not* occur through greater sexual reproduction of players who enjoy greater success in the recurrent game. Rather, evolution occurs by players, often imperfectly rational, *learning* from experience in the game and switching to strategies that provide greater expected utility. (see). Evaluating the success of these models is beyond the scope of this article. Some argue, however, that the same problem of induction that affects salience also infects evolutionary models, to the extent they require players to detect the patterns that constitute a game and other players' strategies in that game (see Sudgen 1998).

4. *Defining 'Norm'*

As noted above, a norm can be a mere behavioral regularity among individuals, as the fact that most people in a population drink coffee or tea nearly every morning. In moral theory, a norm can be a prescriptive rule or reason for action. Thus, one could sensibly assert the existence of a moral norm obligating certain behavior even though no one actually adheres to the norm. In a third sense, a norm is a behavioral regularity but not 'merely' one. It is a regularity sustained in part by the fact that individuals generally approve (and otherwise reward) conformity and/or disapprove (and otherwise punish) nonconformity. Norms are, therefore, perceived by individuals in the relevant population as obligatory, a regularity one 'ought' to follow, whether or not they are justified in moral theory. This rest of this article discusses this third sense of the term.

The definition of norm (in the sense indicated) entails at least three elements (see Pettit 1990). First, it is a behavioral regularity generally followed by the individuals in a population. This requirement distinguishes a society's mere aspirations for behavior from the norms that its members actually follow. One might want to expand the definition, however, to include cases where most do not follow the regularity but most *either* follow the regularity *or* bear significant costs to avoid having their nonconformity discovered. Second, a norm is a regularity in which conformity is approved and/or nonconformity is disapproved. On a broad view, the approval and disapproval need not be morally justified or even based on principle but may include the simple attitudes of being positively or negatively inclined toward—liking or disliking—the behavior. Pettit (1990, p. 731) suggests a third requirement: that the pattern of approval and disapproval 'helps to ensure' the existence of the regularity. The causal contribution may be unconscious and contingent, in that other considerations may usually suffice to sustain the regularity but the (dis)approval pattern fills in when those considerations fail.

One may also follow Pettit (1990, p. 751) in requiring some kind of common belief or knowledge of the existence of the above three elements. Even given these elements, if most people were unaware of the regularity, thinking that contrary behavior is equally common, it would seem odd to label the covert regularity a norm. The same might be true if people were aware of the regularity but not of the pattern of (dis)approval, or perhaps even if people were aware of both but attributed the regularity solely to other forces, thinking the pattern to be epiphenomenal. Finally, people might generally believe that all three elements exist, but might make mistakes about what others believed of these facts, or what others believe other believe, and so on. One may require some higher order beliefs, or at least the absence of contrary beliefs, although not necessarily full-scale common knowledge.

Norms may then be distinguished from conventions, which do not require approval of conformity and/or disapproval of nonconformity, but may be sustained entirely by other means. There is substantial room for overlap, but individuals may approve of conformity to behavioral regularities that are not conventions and they may follow conventions without approving others' conformity or disapproving their nonconformity.

5. *The Origin of Informal Norms*

Legal norms are formal; they are promulgated and enforced by the centralized authority of the state. Nongovernmental organizations also formally promulgate and enforce norms. But some norms arise and are enforced informally and without centralized authority. Informal negative sanctions include censure, gossip, ostracism, and violence. Positive sanctions include praise and the provision of social and economic opportunities. Beyond external sanctions, one may

approve or disapprove one's own conformity or nonconformity, respectively, and experience pride or guilt. Informal norms are important because they contribute to spontaneous order, which has implications for, among other things, political theory.

A key question is how informal norms first arise. What causes the behavioral regularity and what causes people to embrace it as obligatory? Some theorists emphasize nonrational forces and, indeed, view norm-based explanations of behavior as an alternative to rational choice explanations. As implied above, norms can be 'internalized' so that one feels guilt or anxiety from violating a norm, and pride or self-satisfaction from complying, even if no one else is aware of one's behavior. Elster (1989, pp. 97–151) emphasizes this emotional feature of norms. He claims that, contrary to rational behavior, a defining characteristic of norm-based action is that it is not 'outcome-oriented.' Internalization, however, seems a poor candidate for explaining norm origin. However it occurs, it seems more likely to follow than to precede the approval of conformity and/or disapproval of nonconformity.

By contrast, rational choice theorists have offered two nonexclusive ways of explaining norm origin (see Pettit 1990). They differ by whether one explains the behavioral regularity as arising before or after the pattern of approval and disapproval.

5.1 *Deriving Norms from Regularities*

The common way to derive a norm from interest is to explain first how a regularity arises and second how people come to approve conformity to the regularity and/or disapprove nonconformity. Game theory addresses the first issue—regularities—by its study of the mechanisms by which equlibrium is achieved. The (dis)approval patterns may subsequently arise for various reasons. Lewis (1969, pp. 97–100), for example, notes that conditions of a convention may easily give rise to approval of conformity. Even if conventions need not include his requirement that most individuals conditionally prefer that *others* conform (discussed above), it will often be the case that the condition holds. For example, one driving on the conventional side of the road desires that others follow the convention. In this circumstance, it is easy to imagine that self-interested individuals begin to *approve* of others conforming to the convention, given the danger they each face from nonconformity. Similarly, if contingent cooperation is a convention in an iterated prisoner's dilemma game, one will not only expect but also prefer that another engages in the conventional cooperation (see Bicchieri 1993).

A consequence of this analysis is that a regularity might first be a convention, become both a convention and a norm, and end up as a norm only. After the regularity became a norm, it might cease to be arbitrary in the sense described above. The incentives created by the norm—particularly if it is ultimately internalized—might change the payoffs so that the only equilibrium outcome was conformity with the norm regularity.

5.2 *Deriving Norms from Patterns of Approval and Disapproval*

A less common means of deriving norms from interest is to explain first how a pattern of approval and disapproval arises and second how this produces a behavioral regularity. Various mechanisms might produce the attitudinal pattern, the most obvious being the egoistic effect just discussed. Despite there being no regularity at this stage, it is easy to imagine that one person disapproves (approves) of another for conduct that harms (benefits) the former. Thus, prior to a regularity, an individual may negatively regard a neighbor who makes loud and disturbing noises and positively regard those who clean up common areas. The same result could occur through moralistic rather than egoistic reasoning, if individuals each tend to disapprove (approve) of behavior they believe generally harms (benefits) others.

These patterns of (dis)approval can create behavioral patterns if individuals sufficiently value *esteem*, i.e., the approval and good opinion of others, as an end. In some circumstances, individuals may (a) acquire information about the behavior of others without bearing any cost for that purpose, and (b) form an esteem judgment of approval or disapproval, which, being essentially reflexive and involuntary, is also costless. If so, the allocation of esteem creates a costless means of punishing or rewarding others. The desire for approval and the fear of disapproval may cause individuals to conform to a known pattern of esteem judgments (see Pettit 1990, McAdams 1997).

Elster (1989, p. 133) observes that *expressing* disapproval is always costly, as it takes time and risks the retaliation of the target. One might add that selectively expressing approval risks retaliation by those not approved. But these costs need not prevent esteem-judgments from operating as a sanction and producing a norm. First, judgments of disapproval need not be expressed to be perceived. Introspection may suffice: if *A* hates having her sleep disturbed by a loud noise, she may simply guess that others will disapprove of her for waking them with a loud noise (see Pettit 1990, p. 740). Second, there may be a cost to concealing one's esteem judgment that offsets the costs of expression. Suppose *B* prefers to socialize only with people he esteems. For *B*, inviting to a social gathering people he does not esteem (or forgoing the gathering altogether) is more costly than revealing his negative esteem judgments. *B* therefore rationally invites only those whom he esteems and, despite the costs, *indirectly* expresses his negative judgments about others. As to both points, if individuals value esteem, they will seek evidence of

how to get it, so even a trivial clue may cause people to form beliefs about the judgments being made. Once individuals perceive a pattern of approval and disapproval, they may gain (net) approval by directly expressing views consistent with the pattern.

Because individuals will tend to approve commonly beneficial behavior more than commonly harmful behavior, the discussion implies that norms derived from (dis)approval patterns will tend to be socially beneficial. A more detailed analysis would reveal several offsetting effects. However norms are derived, they can cause harm in a variety of ways, particularly by allowing one group to benefit itself at the expense another (see Hardin 1995, Ullman-Margalit 1977, pp. 134–97).

See also: Economics and Ethics; Evolution, Natural and Social: Philosophical Aspects

Bibliography

Bicchieri C 1993 *Rationality and Coordination*. Cambridge University Press, Cambridge, UK

Burge T 1975 On knowledge and convention. *Philosophical Review* **84**: 249–55

Ellickson R C 1991 *Order Without Law: How Neighbors Settle Disputes*. Harvard University Press, Cambridge, MA

Elster J 1989 *The Cement of Society: A Study of Social Order*. Cambridge University Press, Cambridge, UK

Gilbert M 1989 *On Social Facts*. Routledge, London

Gilbert M 1990 Rationality, coordination, and convention. *Synthese* **84**: 1–21

Hardin R 1995 *One for All: The Logic of Group Conflict*. Princeton University Press, Princeton, NJ

Hume D [1740] 1978 *A Treatise of Human Nature*. Oxford University Press, Oxford, UK

Kornhauser L A 1996 Conceptions of social rule. In: Draybrooke D (ed.) *Social Rules: Origin; Character; Logic; Change*. Westview Press, Boulder, CO

Lewis D K 1969 *Convention: A Philosophical Study*. Harvard, University Press, Cambridge, MA

McAdams R H 1997 The origin, development, and regulations of norms. *Michigan Law Review* **96**: 338–433

Mehta J, Starmer C, Sugden R 1994 The nature of salience: An experimental investigation of pure coordination game. *American Economics Review* **84**: 658–73

Miller S R 1986 Conventions, interdependence of action, and collective ends. *Noûs* **20**: 177–40

Pettit P 1990 *Virtus Normativa*: rational choice perspectives. *Ethics* **100**: 725–55

Quine W V 1967 [1936] Truth by convention. In: Lee O H (ed.) *Philosophical Essays for Alfred North Whitehead*. Russel & Russel, New York

Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA

Skryms B 1996 *Evolution of the Social Contract*. Cambridge University Press, Cambridge, UK

Sugden R 1986 *The Economics of Rights, Co-operation and Welfare*. B. Blackwell, Oxford, UK

Sugden R 1998 The role of inductive reasoning in the evolution of conventions. *Law and Philosophy* **17**: 377–410

Taylor M 1987 *The Possibility of Cooperation*. Cambridge University Press, Cambridge, UK

Ullmann-Margalit E 1977 *The Emergence of Norms*. Clarendon Press, Oxford, UK

Vanderschraaf P 1995 Convention as correlated equilibrium. *Erkenntnis* **42**(1): 65–87

Vanderschraaf P 1998 The informal game theory in Hume's account of convention. *Economy and Philosophy* **14**: 215–47

R. H. McAdams

Conversation Analysis: Sociological

Conversation analysis (CA) is a field of sociology concerned with the norms, practices and competences underlying the organization of social interaction. Notwithstanding the name, CA is concerned with all forms of spoken interaction including not only everyday conversations between friends and acquaintances, but also interactions in task-focused social contexts such as medicine, the mass media, education and socio-legal contexts. Originating in the 1960s in the privately circulated lectures of Harvey Sacks (1992), CA has grown into a field of research that is practiced worldwide.

1. Background

CA emerged from the confluence of two theoretical initiatives in sociology. The first derives from Erving Goffman who, in a long series of theoretical writings, argued that social interaction forms a distinct institutional order comprised of normative rights and obligations that regulate conduct in interaction, and that functions as the medium for the operation of other societal institutions. From Goffman, CA adopted the essentially Durkheimian perspective that these normative conventions are autonomous and independent of the social and psychological characteristics of persons and their particular motivations and projects, and indeed are the vehicles through which the particular characteristics of interactants are made manifest in conduct. The second influence derives from Harold Garfinkel's ethnomethodology, which stresses the contingent and socially constructed nature of both action and the understanding of action in the social world. From Garfinkel, CA adopted the perspective that a common body of normative conventions and practices are the basic resources for the methodical production and recognition of action, and for the achievement of common understandings of joint activities in a dynamic social context that is maintained or altered with each successive contribution.

These two perspectives were fused into a methodology that focuses on the sequential production of interaction. Analysis of the normative practices through which interaction and its outcomes are built is possible because participants display their understandings and analyses of one another's conduct in each successive contribution to interaction. The systematic 'choices' involved in each successive move are resources for grasping the practices—conceived as a domain of massive order and empirical regularity—through which these choices are implemented.

2. *Basic Assumptions*

Underlying this sequential methodology are a number of fundamental ideas. First, in performing some current action, participants normally project (empirically) and require (normatively) the production of a 'next' or range of possible 'next' actions to be done by another participant. Second, in constructing a turn at talk, they normally address themselves to preceding talk and, most commonly, the immediately preceding talk. Participants design their talk in ways that exploit this basic positioning, thereby exposing the fundamental role of this sequential contextuality in their utterances.

Third, by the production of next actions, participants show an understanding of a prior action and do so at a multiplicity of levels—for example, by an 'acceptance,' an actor can show an understanding that the prior turn was possibly complete, that it was addressed to them, that it was an action of a particular type (e.g., an invitation) and so on. These understandings are (tacitly) confirmed or can become the objects of repair at any third turn in an on-going sequence. Within this framework, the grasp of a 'next' action that a current projects, the production of that next action, and its interpretation by the previous speaker—are the products of a common set of socially shared practices. CA analyses are thus simultaneously analyses of action, context management and intersubjectivity because all three of these features are simultaneously, if tacitly, the objects of the actors' actions.

Embedded in this perspective is the notion that social action is both context-shaped and context-renewing, and that social context is not a simple 'container' of social interaction, but rather something that is dynamically created, sustained and altered across an interaction's course. Similar conclusions hold for the relevant social identities of the participants, which are also activated, sustained or adjusted on a temporally contingent basis. Finally, the procedures that inform these activities are normative in that actors can be held morally accountable both for departures from their use and for the inferences which their use, or departures from their use, may engender. The sequential methodology arising from these assumptions has proved highly productive in generating findings about the basic organization of interaction, and its modifications in a wide range of more specific institutional settings.

For the first two decades of its existence, CA research focused on ordinary or casual interaction between friends and family members, rather than on the intersection between interaction and specific social institutions. This focus on the institution of conversation reflected the stance that ordinary conversation is in various ways a fundamental domain for the analysis of interaction, and that its analysis is a basic resource for the extension of CA into more specialized interactional domains (Sacks et al. 1974). There are various reasons to treat ordinary conversation as fundamental, and indeed as a 'primordial form of human sociality' (Schegloff 1992). It is the predominant form of human interaction in the social world and the primary medium of communication to which the child is exposed and through which socialization proceeds. It thus antedates the development of other, more specialized, forms of 'institutional' interaction both phylogenetically in the life of society and ontogenetically in the life of the individual.

Moreover the practices of ordinary conversation appear to have a 'bedrock' or default status. They are less historically mutable and subject to discursive justification (by reference, for example, to equity or efficiency) than practices of communication in legal, medical, pedagogical and other institutions manifestly are. Research is increasingly showing that communicative conduct in more specialized social institutions embodies task- or role-oriented specializations and particularizations that generally involve a narrowing of the range of conduct that is generically found in ordinary conversation (see below). The latter thus embodies a diversity and range of combinations of interactional practices that is unmatched elsewhere in the social world. Communicative conduct in institutional environments, by contrast, embodies socially imposed and often irksome departures from that range.

3. *Domains of Research on the Study of Ordinary Conversation*

CA research into ordinary conversation addresses practices of interaction that are organized as systems. These practices are clustered in a range of domains. These include the following:

(a) Turn-taking comprises the organization of coordinated access to the interactional floor and its social regulation, together with the management of failures in coordination and regulation (manifested in overlapping talk and various forms of competitive interruption).

(b) Sequence organization focuses on the practices through which courses of action are built, and how

alternative pathways within them are activated and structured. Beginning with simple paired actions—such as question-answer, or invitation response—the study of sequence organization focuses on the various ways in which the interior structure of these sequences are managed and expanded. Many sequence organizational practices, including expansions, are oriented to issues of affiliation and social solidarity between the speakers, handled under the related topic of 'preference.' Sequence expansions, for example, are, in large part, designed to avoid, minimize or otherwise mitigate disaffiliative actions which are potentially ubiquitous, and which would otherwise undermine the social solidarity of interactants. For example, the simple pre-expansion inquiry 'Are you doing anything tonight?' permits a recipient, by responding with an alternative activity, to forestall an upcoming suggestion and thus to avoid the face-threatening action of rejection. Correspondingly, a 'No' response permits a would-be inviter to proceed with more confidence since a less face-threatening opportunity to reject the invitation has already been passed.

(c) Repair concerns the resources with which participants deal with problems in speaking, hearing and understanding talk, including the interactional mechanics of self- and other-initiated repair, and the ongoing management of problems in sustaining intersubjective understanding.

Other domains of research include story-telling and narrative, the internal structuring of turns at talk, word selection, prosody, and the intersection between talk and body behavior. Taken together, this research has generated a very large range of findings that link the sociological study of social interaction to convergent perspectives in linguistics and communication. These findings have increasingly been enriched by parallel analyses from other languages, including most European languages and a number of prominent Asian languages, including Japanese, Korean, and Mandarin.

4. *Interaction and Social Institutions*

During the past two decades, the findings emerging from the study of ordinary conversation have found increasing application in the study of interaction in various kinds of institutional settings, including education and social work, socio-legal contexts, police work, the mass media, medicine, and the human-machine interface. These are all areas in which social interaction is a prominent feature of the working activities of participants. Research in this field addresses the ways in which the work of institutions is realized in interaction.

Initial interest in the analysis of social institutions focused on how participants' actions embodied their understandings of relevant aspects of the settings which, through these same actions, they constitute as real and binding on one another. As noted earlier, institutional interactions tend to involve reduction in the kinds of practices and actions that are deployed in ordinary conversation and a specialization of those that remain. In this context, research initially focused on the ways in which turn-taking arrangements and basic facets of sequence organization undergo modifications in such contexts as classrooms, law courts, news interviews and business meetings, and the ways in which these modified arrangements structure opportunities for action among the participants that impact on other dimensions of interaction in these settings.

Subsequent research has focused on the ways in which the particular demands of institutional roles, tasks and identities are managed in the sequential structure of interaction, in the specific design of turns at talk, and in the ways that the different knowledge bases possessed by the interactants play out as bases for inference and action. More recent work increasingly ties specific practices of interaction to participants' perceptions and expectations, and to the outcomes of interactions, particularly in medicine and related fields where outcomes are measurable. Finally, with the passage of time and the growth of an historical record of recorded data, it is becoming possible to chart changes in interactional practices over time and to examine them as indices and vectors of social change.

5. *Methodologies*

CA embodies a significant methodological departure from most extant sociological perspectives. First, it is premised on the assumption that all aspects of social interaction are organized and orderly and that social participants attend to interaction in fine detail, albeit with varying degrees of reflexive awareness. Thus gaze, body posture, prosody, intonation, pacing, self-interruption, word selection, grammar, etc., have been shown to be the objects of systematic orientation in both the production and interpretation of interaction. This conception of order contrasts markedly with other positions in the social sciences—from Parsons to Chomsky—which start from the presumption that social interaction is 'noisy' and needs to be cleansed of accidental detail if analysis is to be successful. It effectively mandates that CA investigations be based on audio- and video-recordings of interaction, rather than data arrived at through invention, reconstruction, recollection or precoding, because all the latter foreclose on the levels of detail and organization that conversation analysts have repeatedly found to be relevant to participants in recorded specimens of interaction.

Beyond this point, there are some divergences. The study of ordinary interaction focuses on practices which are putatively of general provenance, and usable

by anyone who knows how to talk. While large numbers of specimens of such practices are valuable in determining their nature and function, statistical analysis has played little role in the field, largely because in the matter of interactional practices, as in the case of biological species, large numbers are not essential to establishing their existence. In the field of institutional interaction, however, and especially in its applied aspects, where practices of interaction are being linked to tasks, identities, perceptions, expectations and outcomes, a more statistically-focused methodology is appropriate and is increasingly used.

6. Conclusion

CA is a large and diversifying field of study. Its basic outlook and methodology have achieved 'paradigmatic' status in Thomas Kuhn's sense, and is widely accepted across the range of social science disciplines to which it has contributed. It seems likely that its methodology will continue to evolve and that it will contribute, not only to sociology, communication science and linguistics, but also to medicine, neuroscience, artificial intelligence and the life sciences with which its practitioners are in increasing contact.

See also: Action, Theories of Social; Goffman, Erving (1921–82); Interactionism: Symbolic; Labor, Division of; Sociology, Epistemology of; Theory: Sociological

Bibliography

Atkinson J M, Heritage J (eds.) 1984 *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, Cambridge, UK

Couper-Kuhlen E, Selting M (eds.) 1996 *Prosody in Conversation*. Cambridge University Press, Cambridge, UK

Drew P, Heritage J (eds.) 1992 *Talk at Work*. Cambridge University Press, Cambridge, UK

Goffman E 1983 The interaction order. *American Sociological Review* **48**: 1–17

Goodwin C 1981 *Conversational Organization: Interaction Between Speakers and Hearers*. Academic Press, New York

Goodwin C (ed.) 2002 *Conversation Analysis and Communication Disorder*. Oxford University Press, Oxford, UK

Garfinkel H 1967 *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ

Heath C, Luff P 2000 *Technology in Action*. Cambridge University Press, Cambridge, UK

Heritage J, Maynard D (eds.) 2002 *Practicing Medicine: Structure and Process in Primary Care Encounters*. Cambridge University Press, Cambridge, UK

Ochs E, Schegloff E A, Thompson S A (eds.) 1996 *Interaction and Grammar*. Cambridge University Press, Cambridge, UK

Sacks H 1992 *Lectures on Conversation*. Blackwell, Oxford, UK

Sacks H, Schegloff E A, Jefferson G 1974 A simplest systematics for the organization of turn-taking for conversation. *Language* **50**: 696–735

Schegloff E A 1992 Repair after next turn: The last structurally provided for place for the defense of intersubjectivity in conversation. *American Journal of Sociology* **97**(5): 1295–345

ten Have P 1999 *Doing Conversation Analysis*. Sage, London

J. Heritage

Conversation: Linguistic Aspects

Conversations are social creations. They are produced one step at a time as people carry out certain joint activities. A joint activity is one in which two or more people have to coordinate with each other to succeed (Clark 1996). These include not only waltzing, playing a piano duet, playing tennis, but gossiping, planning a party, and negotiating a contract. In waltzing, duets, and tennis, people coordinate moment by moment by means of gesture, touch, and other actions; but in gossip, planning, and negotiation, they use speech as well—they converse. What people do and say is not determined beforehand. It emerges as they negotiate their way through these activities.

Conversations reflect the joint activities they coordinate. Every joint activity has participants—the people actually taking part, who are distinct from nonparticipants (bystanders, onlookers, overhearers)—and so do the conversations that emerge from them. The participants take particular roles, such as doctor and patient, teacher and student, or friend calling and friend called, and the roles constrain what the participants do and say. Every joint activity has public goals—mutually agreed-upon purposes for carrying them out. The overall goal may be to exchange gossip, plan an outing, or negotiate a contract, and these have subgoals. Although some of these goals are set from the start, most get established as the participants go along. The participants also have private goals—to be polite, not to lose face, or to finish quickly, for example—and these, too, constrain what the participants do and say. Finally, people often engage in two or more joint activities at a time—such as gossiping and eating dinner together—so their conversation switches back and forth between them.

Working together in a joint activity takes commitments and actions by all the participants. Joint activities have boundaries—distinct beginnings and ends, and transitions from one part to the next—but these boundaries don't exist until the participants agree to them. To enter a planning session, for example, two people must agree on (a) what the joint activity is to be, (b) who is to take part, (c) in what roles, (d) at what time and (e) at what place, and (f) whether or not they are each committed to taking part. They must reach these agreements at each transition point as well. What is remarkable is that people

Table 1
An adjacency pair

Ann	whereabouts are you going,
Betty	I've got a job at Columbia University, in New York, -
Ann	have you,

Table 2
A pre-question

Duncan	Edgar, I I . may I ask you a question,
Edgar	surely,
	[Duncan and Edgar exchange fifteen turns as Duncan describes an author's claim about Danish usage.]
Duncan	it is it it really so, that that is the common usage in Denmark, I don't know if you've heard, of that,
Edgar	- - umm . I'm . I would be surprised, if it was accepted usage,

accomplish all this locally, turn by turn (Sacks et al. 1974).

One basic unit of conversation is the *adjacency pair* (Schegloff and Sacks 1973), as in the spontaneous example shown in Table 1 (all examples are from Svartvik and Quirk 1980).

Adjacency pairs each have two parts, by different speakers, where part 2 is conditionally relevant given part 1. Part 1 is a *proposal*, and part 2 is expected to be the *uptake* of that proposal. In Table 1, Ann proposes that Betty tell her whereabouts she is going, and Betty takes up the proposal by saying that she's got a job at Columbia University. In just two turns, Ann and Betty manage to coordinate on the content, participants, roles, time, place, and commitments of their joint action. They would have failed if Betty had replied 'What do you mean?' or 'You mean me?' or 'I don't know' or 'I won't tell you.'

Adjacency pairs are available for a wide range of joint actions. These include not only requests for information (as in Ann and Betty's exchange), but greetings ('Hi,' 'Hi'), farewells ('Bye,' 'Bye'), offers ('Have some cake,' 'Thanks'), orders ('Sit down,' 'Yes, sir'), and apologies ('Sorry,' 'Oh, that's okay') (Stenström 1994). They are used for even the simplest exchanges of information ('I've got a job … ' 'Have you?').

Adjacency pairs take only two turns, but they can be used to project larger sections, as in Table 2 (the disfluencies are in the original).

Duncan's first turn is a *pre-question* (Schegloff 1980). With it he proposes to ask Edgar a question, and Edgar agrees. Duncan now has the freedom to take up preliminaries to his question, and it takes the two of them fifteen turns to do that. Only then does he ask his question, and Edgar answers it. Pre-questions project not only the eventual question, but preliminaries to that question.

Table 3
Types of pre-sequences

Pre-request	Customer	Do you have hot chocolate?
	Waitress	Yes, we do
Pre-invitation	Man	What are you doin'
	Woman	Nothin.' What's up?
Pre-narrative	June	Did I tell you I was going to Scotland
	Kenneth	No
Pre-conversation	Caller	[rings telephone]
	Recipient	Miss Pink's office

Table 4
Opening of telephone conversation

Karen:	(rings Charlie's telephone)
Charlie:	Wintermere speaking? -
Karen:	hello?
Charlie:	h*el*lo
Karen:	Charlie
Charlie:	yes
Karen:	actually it's
Charlie:	hello Karen
Karen:	it's me
Charlie:	m
Karen:	I (- laughs) I couldn't get back last night, [continues]

Pre-questions and their responses are one of a large family of so-called *pre-sequences*. Table 3 gives a few more examples.

Each pre-sequence prepares the way for another joint action. The pre-request sets up a request ('I'll have one'); the pre-invitation sets up an invitation ('Would you like … '); the pre-narrative sets up a narrative; and the pre-conversation sets up an entire telephone conversation. So pre-sequences are useful in organizing longer sections of conversation.

Opening a conversation takes special coordination as two or more people move from not being in a conversation to being in one. Table 4 gives the opening of a telephone conversation between acquaintances.

First, Karen and Charlie coordinate contact through a proposal to have a conversation (the telephone ring) and its uptake ('Wintermere speaking?'). Next, they mutually establish their identities. Karen tells Charlie that she recognizes him in line 5, but Karen has to say 'hello?' 'Charlie,' and 'actually it's' before he identifies her in line 8. Only then does Karen introduce the first topic. It took ten turns for them to coordinate on the participants, roles, time, place, and commitment to the conversation.

Closing conversations is no easier (Schegloff and

Table 5
Closing of telephone conversation

June	and I'll . I'll ring again, as soon as I can on the tenth, uhh to definite confirm it,
Kathryn	right,
Kathryn	okay,
June	right,
June	thanks a lot,
Kathryn	r . right,
June	bye bye,
Kathryn	bye

Table 6
A pre-narrative exchange

Nancy:	I acquired an absolutely magnificent sewing-machine, by foul means, did I tell you about that
Kate:	No
Nancy:	well when I was . doing freelance advertising - [proceeds to give a five minute narrative]

Table 7
A pre-narrative exchange

Connie:	did I tell you, when we were in this African village, and [- they were all out in the fields, - the]
Irene:	[yes you did, yes, - yes]
Connie:	babies left alone, -
Irene:	yes .

Sacks 1973), as illustrated in Table 5 (an ending to a telephone conversation).

Although June and Kathryn finish a topic in lines 1 and 2, they cannot hang up without agreeing to hang up. So in line 3, Kathryn proposes to close the conversation ('Okay'), and although June could now introduce a new topic in line 4, she agrees to Kathryn's proposal ('Right'). That opens up the closing in which the two exchange thanks ('Thanks a lot' 'Right') and then goodbyes. The two must *agree* to close the conversation before they can actually close it.

Conversations are opportunistic: The paths people take depend on the opportunities that become available with each agreement. Consider the pre-narrative and its response in Table 6.

Nancy proposes to tell Kate a story ('Did I tell you about that'), and Kate accepts ('No'). But partners can also decline, as in Table 7.

Irene interrupts Connie (the speech in brackets is overlapping) to say that she *has* heard the story, and the two of them then go down a different path. Nancy and Connie used their pre-narratives to discover how best to proceed and, receiving different replies, took different routes.

Table 8
A side sequence

Duncan	well wha- uhh what shall we do about uh this boy then, - -
Ed	Duveen?
Duncan	m,
Ed	well I propose to write, [continues]

People help signal which opportunities they are taking by using *discourse markers* (Schiffrin 1987). For example, Nancy used 'well' to signal that she was introducing a change in perspective as she began her story. Other discourse markers indicate such boundaries as the start of a new topic (e.g., 'so,' 'then,' 'speaking of that'), the start of a digression ('incidentally,' 'by the way'), or the return from a digression ('anyway,' 'so'). All of these help in coordinating on what is happening next.

People carry out joint activities against their common ground—their mutual knowledge, mutual beliefs, and mutual assumptions. They infer their common ground from past conversation, joint perceptual experiences, and joint membership in cultural communities. When Ann asked Betty 'Whereabouts are you going?' she presupposed certain common ground—for example, that Betty has been looking for a university job in America, and that Ann doesn't know where. Ann also adds to their common ground that she wants to know. Conversations proceed by orderly increments to the participants' current common ground.

Conversations cannot succeed, therefore, unless the participants *ground* what they say. To ground what is said is to establish the mutual belief that the addressees have understood the speakers well enough for current purposes (Clark and Schaefer 1989). One technique for grounding is the adjacency pair itself. When Betty said 'I've got a job at Columbia University,' she displayed to Ann how she had interpreted her question. If Ann hadn't been satisfied with that interpretation, she could have corrected it, e.g., by replying 'No, I meant … ' (Schegloff et al. 1977). In fact, she followed up Betty's reply with 'have you,' displaying to Betty that she accepted her interpretation. Another technique is the *back-channel response*, *acknowledgment*, or *continuer* (Schegloff 1982). In two-party conversations, addressees are expected to add 'uh huh' or 'mhm' or 'yeah' at or near the ends of certain phrases. With these, they assert that they have understood the speaker so far and imply that the speaker should continue.

Grounding is sometimes achieved through *side sequences* (Jefferson 1972), as illustrated in Table 8.

When Ed didn't understand to whom Duncan was referring as 'this boy,' he initiated an embedded adjacency pair, a side sequence, to clear up the problem. Only when he had cleared it up did he answer

Duncan's original question. Side sequences are initiated to clear up not only mishearings and misunderstandings, but also other pre-conditions to taking up the first part (e.g., 'Why do you want to know?'). Grounding is also accomplished sometimes by overlapping speech. When Irene interrupted Connie's offer 'Did I tell you … ' to say, 'yes you did, yes, yes,' she was signaling to Connie that she already understood and Connie didn't need to go on.

The structure of conversations, in summary, is not determined from the outset. It emerges step by step as people coordinate on each new move in joint activities. People have to coordinate on the content, participants, roles, time, place, and commitment to each joint action, and they do that in a sequence of local, opportunistic agreements. It is these techniques that lead conversations to have the structure they do.

See also: Conversation Analysis: Sociological; Expression and Interaction: Microfoundations of Expressive Forms; Inferences in Discourse, Psychology of; Verbal Communication, Psychology of

Bibliography

Clark H H 1996 *Using Language*. Cambridge University Press, Cambridge, UK

Clark H H, Schaefer E R 1989 Contributing to discourse. *Cognitive Science* **13**: 259–94

Jeffeson G 1972 Side sequences. In: Sadrow D (ed.) *Studies in Social Interaction*. Free Press, New York, pp. 294–338

Sacks H, Schegloff E A, Jefferson G 1974 A simplest systematics for the organization of turn-taking in conversation. *Language* **50**: 696–735

Schegloff E A 1968 Sequencing in conversational openings. *American Anthropologist* **70**(4): 1075–95

Schegloff E A 1979 Identification and recognition in telephone conversational openings. In: Psathas G (ed.) *Everyday Language: Studies in Ethnomethodology*. Irvington, New York, pp. 23–78

Schegloff E A 1980 Preliminaries to preliminaries: 'Can I ask you a question?' *Sociological Inquiry* **50**: 104–52

Schegloff E A 1982 Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In: Tannen D (ed.) *Analyzing Discourse: Text and Talk. Georgetown University Roundtable on Languages and Linguistics 1981*. Georgetown University Press, Washington, DC, pp. 71–93

Schegloff E A, Jefferson G, Sacks H 1977 The preference for self-correction in the organization of repair in conversation. *Language* **53**: 361–82

Schegloff E A, Sacks H 1973 Opening up closings. *Semiotica* **8**: 289–327

Schiffrin D 1987 *Discourse Markers*. Cambridge University Press, Cambridge, UK

Stenström A-B 1994 *An Introduction to Spoken Interaction*. Longman, London

Svartvik J, Quirk R (eds.) 1980 *A Corpus of English Conversation*. Gleerup, Lund, Sweden

H. H. Clark

Cooperation and Competition, Psychology of

1. *Introduction*

During the twentieth century, there have been a series of social forces that have highlighted the importance of cooperation and competition and often created a creative tension between them. In the early 1900s, the advent of the industrial revolution and the ending of colonialism focused attention on (a) the importance of well functioning, cooperative groups, and (b) competition as the secret to increased economic success. During the 1920s and 1930s, with the worldwide depression and the rise of dictatorship in Europe, the importance of cooperation was again highlighted. Business and industry groups, such as the Liberty League in the USA, increased their emphasis on competition in business and industry, education, and other areas of life. In the aftermath of the Second World War there was great concern about the future of democracy and the need to better understand how democratic organizations could be made to function more effectively. The health of a democratic society, it was thought, depended on the effectiveness of its component groups. Strengthening of the quality of cooperation in the family, the community, and the multitude of groups within society was viewed as the primary means of ensuring the vitality of democracy and the successful solving of the existing social problems. At the same time, the rise of social Darwinism promoted competition as the natural order of life. In the early 1960s, anthropology and ethnology proposed that competition over weapon systems sparked human evolution. The civil rights movement and concerns about the possibility of nuclear war, on the other hand, created great interest in cooperation among diverse individuals. The tension between, and interest in cooperation and competition sparked discussions in three major theoretical orientations: cognitive-developmental, behavioral, and social interdependence.

2. *Cognitive-developmental Theory*

The 'cognitive developmental perspective' is largely based on the theories of Piaget (1950), Vygotsky (1978), cognitive science, and academic controversy (Johnson and Johnson 1995). To Jean Piaget (1950), cooperation is the striving to attain common goals while coordinating one's own feelings and perspective with a consciousness of others' feelings and perspective. From Piaget and related theories comes the premise that when individuals co-operate on the environment, socio-cognitive conflict occurs that creates cognitive disequilibrium, which in turn stimulates perspective-taking ability and cognitive devel-

opment. Cooperative learning in the Piagetian tradition is aimed at accelerating a student's intellectual development by forcing him or her to reach consensus with other students who hold opposing points of view about the answer to the school task.

Lev Semenovich Vygotsky (1978) and related social constructionist theorists claim that our distinctively human mental functions and accomplishments have their origins in our social relationships. Mental functioning is the internalized and transformed version of the accomplishments of a group. Knowledge is social, constructed from cooperative efforts to learn, understand, and solve problems. A central concept is the 'zone of proximal development,' which is the zone between what a student can do on his or her own and what the student can achieve while working under the guidance of instructors or in collaboration with more capable peers. Unless students work cooperatively, they will not grow intellectually and the time students work alone on school tasks should therefore be minimized.

From the cognitive science viewpoint, cooperative learning involves modeling, coaching, and scaffolding (conceptual frameworks provided for understanding what is being learned). The learner must cognitively rehearse and restructure information for it to be retained in memory and incorporated into existing cognitive structures (Wittrock 1990). An effective way of doing so is explaining the material being learned to a collaborator. Tutoring, when it is viewed from the perspective of the benefits that accrue to the tutor, is also a form of cooperative learning.

'Controversy theory' (Johnson and Johnson 1995) posits that being confronted with opposing points of view creates uncertainty or conceptual conflict, which creates a reconceptualization and an information search, which results in a more refined and thoughtful conclusion. The key steps are organizing what is known into a position, advocating that position to someone who is advocating the opposing position, attempting to refute the opposing position while rebutting the attacks on one's own position, reversing perspectives so that the issue may be seen from both points of view simultaneously, and creating a synthesis to which all sides can agree.

3. *Behavioral-learning Theory*

The 'behavioral learning perspective' assumes that individuals will work hard on those tasks for which they secure a reward of some sort and will fail to work on tasks that yield no reward or yield punishment (Bandura 1977, Skinner 1968). Cooperative efforts are designed to provide incentives for the members of the group to participate in a group effort since it is assumed that individuals will not intrinsically help their classmates or work toward a common goal. Skinner focused on group contingencies, Bandura focused on imitation, and Homans as well as Thibaut and Kelley focused on the balance of rewards and costs in social exchange among interdependent individuals. Recently, Bandura (2000) has conceptualized cooperation as collective agency. In 'collective agency' individuals have to work together to secure what they cannot accomplish on their own.

The most developed behavioral theory of cooperation and competition was developed by Kelley and Thibaut (1978). They assumed that individuals act to maximize their self-interests by behaving in ways to maximize their rewards and minimize their punishments or costs. 'Rewards' are defined as positive interaction experiences that yield outcomes that promote physical or emotional well being, self-esteem, or personal growth such as enjoyment, fulfillment, or contentment. 'Costs' are defined as negative interaction experiences that yield physical or emotional pain or failure. People attempt to recreate experienced interactions that yield pleasure and avoid interactions that yield pain. 'Cooperation' is thus defined as acting in ways perceived to maximize joint rewards and minimize joint costs, 'competition' as acting in ways to maximize ones own rewards and minimize ones own costs relative to others, and 'individualistic efforts' as acting in ways to maximize ones own rewards and minimize ones own costs with little or no regard of the outcome for others. In evaluating outcomes, individuals have a 'comparison level' about what constitutes an acceptable outcome and a 'comparison level for alternatives' about how those outcomes compare to other possible outcomes. Cooperation tends to occur when outcomes are correspondent and competition tends to occur when outcomes are noncorrespondent.

4. *Social Interdependence Theory*

The roots of social interdependence theory lie in the early 1900s when Kurt Koffka (one of the founders of the Gestalt School of Psychology) proposed that groups were dynamic wholes in which the interdependence among members could vary. In the 1920s and 1930s, Kurt Lewin refined Koffka's notion, proposing that the essence of a group is the interdependence among members (created by common goals), and that interdependence results in the group being a 'dynamic whole' so that a change in the state of any member or subgroup changes the state of any other member or subgroup. Continuing Lewin's work, in 1949 Morton Deutsch (1949, 1962) formulated social interdependence theory.

'Social interdependence' exists when individuals share common goals and each individual's outcomes are affected by the actions of the others (Deutsch 1962, Johnson and Johnson 1989). It may be differentiated from social dependence (i.e., the outcomes of one person are affected by the actions of a second person but not vice versa) and 'social independence' (i.e., individuals' outcomes are unaffected by each other's

actions). There are two types of social interdependence: cooperative and competitive. The absence of social interdependence and dependence results in individualistic efforts.

'Cooperation' exists when individuals work together to accomplish shared goals. When a situation is cooperatively structured, individuals' goal achievements are positively correlated; individuals perceive that they can reach their goals if, and only if, the others in the group also reach their goals. Thus, individuals seek outcomes that are beneficial to all those with whom they are cooperatively linked. 'Competition' exists when individuals work against each other to achieve a goal that only one or a few can attain. When a situation is structured competitively, individuals' goal achievements are negatively correlated; each individual perceives that when one person achieves his or her goal, all others with whom he or she is competitively linked fail to achieve their goals. Thus, individuals seek an outcome that is personally beneficial but detrimental to all others in the situation. 'Individualistic efforts' exist when individuals work by themselves to accomplish goals unrelated to the goals of others. When a situation is structured individualistically, there is no correlation among participants' goal attainments. Each individual perceives that he or she can reach his or her goal regardless of whether other individuals attain or do not attain their goals. Thus, individuals seek an outcome that is personally beneficial without concern for the outcomes of others.

5. *Interaction Patterns*

The basic premise of social interdependence theory is that the type of interdependence structured in a situation determines how individuals interact with each other which, in turn, determines outcomes (Deutsch 1949 1962, Johnson and Johnson 1989). Positive interdependence tends to result in promotive interaction; negative interdependence tends to result in opposition or contrient interaction, and no interdependence results in an absence of interaction. 'Promotive interaction' occurs when members help and assist each other, exchange resources, give and receive feedback, challenge each other's reasoning, and encourage increased effort. Group members focus on maximizing the success of their groupmates as well as their own. 'Oppositional interaction' occurs as individuals discourage and obstruct each other's efforts to achieve. Individuals focus both on increasing their own success and on preventing anyone else from being more successful than they are. 'No interaction' exists when individuals work independently without any interaction or interchange with each other. Individuals only focus on increasing their own success and ignore as irrelevant the efforts of others. Each of these interaction patterns creates different outcomes.

Deutsch (1949, 1962) noted that depending on whether individuals promote or obstruct each other's goal accomplishments, there is 'substitutability' (i.e., the actions of one person substitute for the actions of another), 'cathexis' (i.e., the investment of psychological energy in objects and events outside of oneself), and 'inducibility' (i.e., openness to influence). Essentially, in cooperative situations the actions of participants substitute for each other, participants positively cathect to each other's effective actions, and there is high inducibility among participants. In competitive situations the actions of participants do not substitute for each other, participants negatively cathect to each other's effective actions, and inducibility is low. When there is no interaction, there is no substitutability, cathexis, or inducibility. The relationship between the type of social interdependence and the interaction pattern it elicits is assumed to be bidirectional. Each may cause the other.

6. *Outcomes*

The investigation of cooperation and competition is the longest standing research tradition within American social psychology. Between 1898 and 1989, researchers conducted over 520 experimental studies on social interdependence and over 100 correlational studies (Johnson and Johnson 1989). Since research participants have varied widely as to sex, economic class, age, and cultural background, since a wide variety of research tasks and measures of the dependent variables have been used, and since the research has been conducted by many different researchers with markedly different orientations working in different settings and in different decades, the overall body of research on social interdependence has considerable generalizability.

The numerous outcomes studied may be subsumed within three broad categories (Johnson and Johnson 1989): (a) effort to achieve; (b) positive relationships; and (c) psychological health. In terms of productivity and achievement, a meta-analysis of all available studies (Johnson and Johnson 1989) found that the average person in a cooperative situation performed at about 2/3 a standard deviation above the average person learning within a competitive (effect size = 0.67) or individualistic situation (effect size = 0.64). These results held constant when group measures of productivity were included as well as individual measures, for short-term as well as long-term studies, and when symbolic as well as tangible rewards were used. In terms of relationships, over 180 studies indicate that cooperative efforts, compared with competitive and individualistic experiences, promoted considerable more liking among individuals (effect sizes = 0.66 and 0.62, respectively) even when individuals were from different ethnic and historical backgrounds, or were handicapped. In addition, over 106 studies indicate that cooperative experience promoted greater task-oriented and personal social support than did competitive (effect size = 0.62) or

individualistic (effect size = 0.70) efforts. Finally, the research indicates that (a) working cooperatively with peers and valuing cooperation result in greater psychological health than does competing with peers or working independently and (b) cooperative attitudes are highly correlated with a wide variety of indices of psychological health. Competitiveness was in some cases positively and in some cases negatively related to psychological health, and individualistic attitudes were negatively related to a wide variety of indices of psychological health. In addition, over 80 studies indicated that cooperative experiences promoted higher self-esteem than do competitive (effect size = 0.58) or individualistic (effect-size = 0.44) experiences. Cooperative experiences also tended to increase perspective-taking ability while competitive and individualistic experiences tended to promote egocentrism (being unaware of other perspectives other than your own) (effect sizes of 0.61 and 0.44, respectively). It is through cooperative efforts that many of the attitudes and values essential to psychological health (such as self-efficacy) are learned and adopted. Social skills and competencies tend to increase more within cooperative than in competitive or individualistic situations (Johnson and Johnson 1989). These outcomes of cooperative efforts (effort to achieve, quality of relationships, and psychological health) form a gestalt where each influences the others and they are therefore likely to be found together.

7. *Basic Elements*

In order for cooperative efforts to be effective, five elements must be structured in the situation: positive interdependence, individual accountability, promotive interaction, social skills, and group processing. 'Positive interdependence' exists when one perceives that one is linked with others in a way so that one cannot succeed unless they do (and vice versa) and that one must coordinate one's efforts with the efforts of others to complete a task (Johnson and Johnson 1989). Cooperative efforts may be built on outcome interdependence (i.e., goals and/or reward interdependence or each group member's success in achieving a goal or receiving a reward depends on the joint efforts of all group members) and means interdependence (i.e., resource, role, and task interdependence or group members being tied together by complementary resources, interlocking roles, or a division of labor).

'Individual accountability' exists when the performance of each individual student is assessed and the results given back to the group and the individual. It is important that the group knows who needs more assistance, support, and encouragement in completing the task and group members know that they are responsible for doing their share of the work. 'Promotive interaction' may be defined as individuals encouraging and facilitating each other's efforts to complete tasks and achieve in order to reach the group's goals. Promotive interaction is characterized by group members helping and assisting each other, exchanging information and resources, and encouraging each other to learn.

'Social skills' are the actions needed to effectively interact with other people. Leadership, decision-making, trust building, communication, and conflict-management skills are examples. Social skills form the basic nexus among individuals, and if individuals are to work together productively, they must have a modicum of these skills. 'Group processing' may be defined as reflecting on group efforts to (a) describe what member actions were helpful and unhelpful in achieving the group's goals and maintaining effective working relationships among members, and (b) make decisions about what actions to continue or change. The purpose of group processing is to clarify and improve the effectiveness of the members in contributing to the joint efforts to achieve the group's goals.

8. *Enhancing Factors*

Deutsch (1962, 1973, 1985) has posited that trust, conflict, and distributive justice are mediators of the effectiveness of cooperation. The greater the trust, the more effective cooperative efforts tend to be, the more constructive conflicts tend to be, and the more positive relationships tend to be. Effective cooperative requires that conflicts of interest are resolved constructively and when decisions involve diverse points of view individuals strive to reach agreement through a process of argument and synthesis. In cooperative efforts the distribution of benefits is based on equality and need while in competitive situations benefits are distributed according to equity (high performers receive more than low performers).

9. *Competitive and Individualistic Efforts*

Competition tends to result in constructive consequences when it occurs within a broader cooperative context, clear and fair rules, and criteria for winning are present, the task is appropriate, the task may be completed individually, competitors have an equal chance of winning, and winning is of low importance (Johnson and Johnson 1975/1999; Stanne et al. 1999). The conditions under which individualistic efforts have an advantage over cooperative efforts are unclear.

10. *Future Directions*

The tension between cooperation and competition has not ended. Considerably more research is needed on the nature of cooperation and the variables that mediate its effectiveness. There is new interest in

collective agency (Bandura 2000). The nature of competitive and individualistic efforts the conditions under which they are effective need to be further investigated. There is a current interest in the values inherently taught by participating in cooperative and competitive situations (Johnson and Johnson 1999). Cooperation, for example, may promote a concern for the well-being of others and a desire to contribute to the common good, while competition may promote a desire to be better than others and a concern for self-gain at the expense of others. Finally, one of the least understanding aspects of cooperation and competition is inducibility, the willingness to be influenced by others. Inducibility is the center of constructive power, which occurs, in cooperative but not competitive situations. World events, furthermore, such as increasing economic interdependence among countries but fractionalization within countries, will keep interest in the study of cooperation and competition high in the future.

See also: Competitive Strategies: Organizational; Cooperation: Sociological Aspects; Cooperative Learning in Schools

Bibliography

Bandura A 1977 *Principles of Behavioral Modification*. Holt, Rinehart, and Winston, New York

Bandura A 2000 Exercise of human agency through collective efficacy. *Current Directions in Psychological Science* **9**(3): 75–8

Deutsch M 1949 A theory of cooperation and competition. *Human Relations* **2**: 129–52

Deutsch M 1962 Cooperation and trust: Some theoretical notes. In: Jones M (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, pp. 275–319

Deutsch M 1973 *The Resolution of Conflict*. Yale University Press, New Haven, CT

Deutsch M 1985 *Distributive Justice: A Social Psychological Perspective*. Yale University Press, New Haven, CT

Homans G 1974 *Social behavior: Its Elementary Forms*, 2nd edn. Harcourt Brace Jovanovich, New York

Johnson D W, Johnson R 1989 *Cooperation and Competition: Theory and Research*. Interaction Book Company, Edina, MN

Johnson D W, Johnson R 1995 *Creative Controversy: Intellectual Challenge in the Classroom*. Interaction Book Company, Edina, MN

Johnson D W, Johnson R 1999 *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning*. Allyn and Bacon, Boston

Kelley H, Thibaut J 1978 *Interpersonal Relations: A Theory of Interdependence*. Wiley, New York

Piaget J 1950 *The Psychology of Intelligence*. Harcourt, New York

Stanne M, Johnson D W, Johnson R 1999 Social interdependence and motor performance: A meta-analysis. *Psychological Bulletin* **125**(1): 133–54

Vygotsky L 1978 *Mind and Society*. Harvard University Press, Cambridge, MA

Wittrock M 1990 Generative processes of comprehension. *Educational Psychologist* **24**: 345–76

D. W. Johnson and R. T. Johnson

Cooperation: Sociological Aspects

Individuals cooperate if each willingly acts in a manner that contributes to the others' welfare. Cooperation is one of the oldest and most revered topics in sociology even though it has appeared under a variety of headings, such as 'social order' and 'solidarity.' The classical sociological answers to the question of how cooperation among individuals comes about differ in detail but share a common core: Cooperation is the result of shared values and norms and norm-conforming behavior.

A problem with this 'obligation-centered' solution is that there is no explanation of where norms come from and why the degree of norm conformity varies even for the same individual over time. Theories that might solve these problems were not developed until the 1970s when a sea change in the basic assumptions on human action took place within sociology. At that time, analyses of cooperation among rational egoists (mostly rooted in game theory) began to shed new light on the way one might analyze cooperation, norms, and norm-conforming behavior. For some time, game-theoretic analyses of cooperation had become the most serious attempt to explain cooperation. Big advances were made but many problems remained, especially the problem that individuals do not seem to be as self-centered as they are assumed to be in game theory. Still, the game-theoretic analyses were powerful enough to suggest that there was no way back to explanations of cooperation that ignored human rationality all together. More recently, analyses based on some form of social rationality have arisen in order to address some of the unsolved problems. The following sections first present the most important aspects of sociological research on cooperation among rational egoists, and then discusses some contributions on cooperation with social rationality.

1. Cooperation Among Rational Egoists

It may not be quite realistic to assume that individuals are always looking out for themselves and that they do so by weighing the costs and benefits of their behavior. However, this assumption of 'rational egoism' seems to have four strong points in its favor. First, it allows one to pinpoint why cooperation is an interesting problem. Second, it allows rigorous theory formation with tractable structures of argumentation. Third, everyday experience and research give this assumption

enough support to make it worthwhile to use it in the light of its analytic advantages, especially when analyzing interaction among interdependent individuals. Fourth, at later stages of theory formation, this assumption can be relaxed in favor of more realistic assumptions.

Game theory has furnished the most important instruments for such analyses, and since the late 1970s—inspired by developments in political science and economics—sociologists have begun to use it to study questions concerning cooperation, typically bringing sociological aspects into the analyses, such as network embedding. The crucial first step in the 'new' analysis was to identify typical situations in which cooperation is problematic: social dilemmas. A social dilemma is—roughly speaking—a situation in which the rational decisions of individual actors are likely to lead to a collectively suboptimal outcome. An outcome is suboptimal (i.e., not 'Pareto-optimal') if one or more actors could improve their position without worsening the position of any other actor(s). In a social dilemma, individual and collective rationalities are in conflict. Cooperation can now be succinctly defined: An actor cooperates if and only if he or she chooses a course of action that will lead to a collectively rational outcome when other actors behave cooperatively as well. Social dilemmas occur when Pareto-superior behavior cannot be achieved by enforceable contracts. This may have many different reasons. For example, it may be impossible to communicate with others in order to come to an agreement; or, if communication is possible, transaction costs may be too high to actually reach an agreement. If agreements are within reach, it may be practically or epistemically impossible to specify all relevant future contingencies in the contract. Even if a contract can be drawn up, there may be no court with sanctioning power; or if there is such a court, it may not be possible to prove uncooperative behavior in court.

There are structurally different variants of social dilemmas, and game theory helps to distinguish them. For example, there is the *common's dilemma* (Hardin 1968, Ostrom 1990) in which various actors have free access to a resource (such as fishers from different nations in international waters); there are *step-level public goods* (Taylor 1987), for which only a limited number of actors have to cooperate in order to produce the public good for a much larger group. A special case of this is the *volunteer's dilemma* (Diekmann 1986), for which only one volunteer is necessary to produce the public good for all, such as rescuing a drowning child with a crowd of onlookers; there is the *trust game* (Dasgupta 1988) in which the trustee can honor or opportunistically exploit the cooperation of the other (as when you lend money to an acquaintance who promises to pay it back with interest).

The *Prisoner's Dilemma* (PD) is by far the best known variant of a social dilemma. It involves an arbitrarily large number of people. The special case of a two-person PD has been explored the best so far. In a two-person PD, both players have the option to cooperate or to defect (see Fig. 1). If both cooperate, each gets a payoff of value R. If one defects while the other cooperates, the defector gets T and the cooperator gets S. If both defect, both will get P, with $T > R > P > S$. There is a dominant strategy for each player (i.e., an individually rational choice, no matter what the other does) leading to a 'Nash equilibrium' (in which no player can unilaterally improve his or her position by deviating from the equilibrium strategy) and to a Pareto-inferior result. The assumption of rational egoism leads to two motives in a PD: the defensive motive to shield against exploitation (if the other defects you are better off defecting yourself), and the aggressive motive to exploit a cooperative player (if the other cooperates, you get even more by defecting). Together, these motive result in the conflict between individual and collective rationality. Mutual cooperation would be the Pareto-optimal result but the two motives lead to mutual defection.

	C	D
C	R, R	S, T
D	T, S	P, P

Figure 1
Prisoner's Dilemma Game. In the Axelrod simulation is T (temptation) = 5, R (reward) = 3, P (punishment) = 1, and S (sucker's payoff) = 0

1.1 The Embedding of Cooperation

Raub and Weesie (2000) have suggested that game-theoretic work on solutions to social dilemmas can be divided into three kinds of embeddings of potential cooperation: embedding in time, in institutions, and in social networks. We follow their suggestion for this overview.

1.1.1 Time. When actors find themselves in a social dilemma that is iterated indefinitely over time, new sources for cooperation arise. First of all, defection can be sanctioned by the other player(s) in later moves, and the resumption of cooperation by the defector can be rewarded by a similar response of the

other player(s). Friedman (1977) and Taylor (1987) have shown that if the value of future outcomes is high enough, this opportunity to make one's cooperation conditional on that of the other leads to cooperative strategies with Nash equilibrium and Pareto optimality. Axelrod (1984) built on this result to ask the question which strategy would be most conducive to lead to cooperation in an iterated PD. In order to answer this question, he conducted computer simulations in which various strategies played against each other for about 200 rounds. Axelrod had invited suggestions for the strategies that would create the most cooperation. The winner was the conditional strategy 'tit for tat' (TFT) suggested by the mathematician and psychologist Anatol Rapoport.

TFT is very simple: (a) always begin with cooperation; (b) for the present round, copy the move of your partner in the previous round. TFT has a number of distinct features. It is 'nice,' because it starts with cooperation; it cannot be provoked, because it sanctions exploitation; and it is forgiving, because it returns to cooperation after the other has done so. Contrast this for example with 'permanent retaliation' in which defection is answered by unforgiving defection, a strategy that did much worse than TFT. TFT won against other strategies but it is not ideal under all circumstances. It does not do well against all sorts of strategies and it has not the property of 'evolutionary stability.' The heart of Axelrod's results, however, is thoroughly in line with a classical sociological insight: that the norm of reciprocity is probably the most essential norm generating cooperation in time-embedded interactions. As sociologists have argued for quite some time, homogeneity of a group in terms of important characteristics (including the use of 'nice' strategies) will increase the likelihood of reciprocity, as does smallness of the group because it helps identification of defection. Similarly, a low rate of fluctuation of membership helps reciprocity because it increases the shadow of the future. A related argument has been made in terms of 'social capital' (Coleman 1990).

1.1.2 Institutions. When reciprocity norms work, they are self-enforcing because no player can do better by unilaterally deviating from the cooperative strategy (Nash equilibrium). However, under many conditions, norms of reciprocity will not work properly. For example, when future contingencies are so complex that it is difficult to determine *a priori* what constitutes defection and/or difficult to detect defection, a reciprocity norm is not enough to secure cooperation. Similarly, such self-enforcing norms do not work well for the production of collective goods in larger groups, the less so the more unhomogeneous and fluctuating the group membership. In such situations, expectations concerning the relative importance of one's own and others' contribution are likely to be low and individuals tend to free ride rather than to contribute to the production of the collective good.

When self-enforcing norms do not work, cooperation may be achieved by embedding interactions in formal and/or informal rules which create 'selective incentives,' that is, added rewards for cooperating and/or punishments for not cooperating. This argument was first forcefully made by Olson (1965) and has been refined by sociologists and other social scientists (e.g., Hechter 1987, Marwell and Oliver 1993, Ostrom 1990). For example, the state punishes tax evasion and provides some standard instruments for contracting; in many countries labor unions offer extra benefits (such as cheaper insurance) for joining the labor union. The formation of institutional solutions has been explained by Hechter (1987) as a two-stage process. First, groups for the joint production of private goods are established (such as insurance). They are small enough that noncooperators can be detected and excluded (hence no free rider problem, see Lindenberg 1998). The profits made from the joint production can then be used to establish selective incentives for the joint production of collective goods. In this way, labor unions may be born out of mutual insurance groups. Abell (1996) adds to this kind of analysis that a proper institutional explanation of cooperation should include the explanation of how cooperation can be combined with rivalry and competition. More work in this direction is needed.

Williamson (1985) has pointed to the importance of 'private orderings' in which, in the shadow of the law, contract partners may voluntarily restrict their own options, thereby creating selective incentives for their own cooperative action. For the partner, such measures work if there are 'credible commitments,' for example, by posted hostages and granted warranties. Sociologists have used this idea successfully to study contractual behavior (e.g., Weesie et al. 1998).

Often, credible commitments are difficult to establish (due to the size of the group) and formal institutional arrangements (with enforced sanctions) may not be available. In such groups, informal rules (social norms) that govern selective incentives for cooperation are needed. Like all rules, informal rules only work if they are enforced, and this creates the second-order free rider problem. Sanctioning noncooperators is itself a (step-level) collective good. There seems to be an infinite regress. How do groups get out of this problem? There is no handy solution to this problem within the rational egoist approach. However, as Heckathorn (1996), following Marwell and Oliver, has shown, public good production is not always subject to the first and second order free rider problems. In fact, the production function (i.e., the function that relates the proportion of cooperators to the level of the public good production) is likely to be S-shaped and only for certain values of the function will the situation have the structure of a PD. The reason for this is that an S-shaped production function

implies a changing marginal utility from cooperation as the proportion of cooperators increases. Interestingly, Heckathorn also found that in certain phases 'hypocritical cooperation,' in which people criticize others for not cooperating even though they do not cooperate themselves, is quite effective in eliciting contributions to the collective good (see also Yamagishi 1986). The main lesson to be learned from Heckathorn's analysis is that collective action should be analyzed dynamically, a conclusion one can also draw from Hechter's analysis of institutional controls mentioned above.

1.1.3 Social networks. It has long been known in sociology that it often matters to people what other people say about them. Cooperation is helped by the effects of reputation within networks. For example, DiMaggio and Louch (1998) found in a national sample that consumers, when they are uncertain about products and performance, prefer to buy goods and services from suppliers with whom they have social ties outside the transaction itself. If the suppliers in such a network are clearly unreliable, the consumers are likely to have heard about it. If these suppliers begin to be unreliable, the network relations can be turned against them. Thus, networks have an information function (people learn about the cooperativeness of certain others) and a control function (people can use the network to sanction noncooperators).

Both aspects have been worked out game-theoretically by Buskens (1999). For dyadic transactions, characteristics of the network (such as density, 'indegree,' and 'outdegree,' i.e., the number of others with whom people are connected) play an important role but it turns out that reputation effects are quite complex and there are indications that they might be less important than the temporal embeddedness. For economic transactions, Uzzi (1996) has found interesting additional effects of networks on tacit knowledge and joint problem solving. Network embeddedness for collective action has apparently not been studied yet in any rigorous way.

Yet another use of networks for the explanation of cooperation has been suggested by Coleman (1990). He points to the importance of intermediaries for the creation of trust and thus for cooperation. Intermediaries can play a number of roles: as advisors who introduce interested parties and facilitate the interaction necessary to create trust; as guarantors who absorb the risk if trust turns out to be misplaced; and as entrepreneurs who combine the resources of several actors to place them into the hands of others who are expected to realize gains. For future research, a sociologically interesting question is under what conditions structures with certain kinds of intermediaries will emerge.

2. *Cooperation Among Socially Rational Individuals*

There is a growing sociological literature on cooperation that holds on to the assumption of rational action but adds specific 'bridge theories' on preferences, learning, and/or framing, the effect of which is to render regard for others quite prominent in the analyses. Hence the term 'social rationality.' While the traditional literature on the effects of socialization was unrelated to the analysis of scarcity and choice, the social rationality literature connects effects of socialization with purposeful action in the face of scarcity. In this literature the trade-off between empirical richness and analytical tractability is different than in most game-theoretic analyses. In the latter, added analytical power and tractability is often achieved by the 'rational egoist' assumption at the expense of some loss in closeness to reality, whereas in the former, some analytical power and tractability is sacrificed in favor of added closeness to reality. Clearly, each kind of approach profits from the existence of the other. In this review, there remains room for the discussion of only a few developments.

2.1 *Goals (Preferences)*

By far the largest amount of literature on cooperation in the 'social rationality' category is concerned with 'social preferences.' Although the term 'preference' is often used in this context, a better term would be 'social goals.' One variant of this approach, found in social-psychological and sociological research, is the 'social orientation' research, in which individuals are differentiated according to their orientation towards others. This orientation is often assumed to be more or less stable and the joint result of nature and nurture. Different types of orientation are distinguished, most frequently the following three: 'cooperative' (the goal is to maximize joint payoffs); 'individualistic' (the goal is to maximize individual payoff); 'competitive' (the goal is to maximize the positive difference between own and other's payoff). Such 'types' can be combined with game-theoretic analyses and/or with analyses of evolutionary learning at the population level.

Sociologists and economists focus especially on questions of signaling and the spread of strategies over time, some of them using genetic algorithms taken from evolutionary biology . This makes *trust* a central concept in the analysis of cooperation. For example: 'Are the tendencies to trust and to honor trust correlated?' (Snijders 1996); 'How can cooperative types convincingly signal their type to others so that they will be trusted and cooperation comes about?' Or 'Under what conditions will effective rules for trusting (i.e., conventions for detecting cooperators) evolve in neighborhood interactions and spread across the population through contact with strangers?' (see Macy

and Skvoretz 1998). Another question concerns the explanation of preferences themselves. For example, Frey (1997) has linked a long tradition in sociology to study the effects of voluntary sacrifice (the 'gift') on others' willingness to cooperate with psychological research on autonomy. He argues forcefully that a cooperative orientation (he speaks of 'intrinsic motivation') is driven out by monetary rewards and by forced cooperation. Conversely, Raub and Voss (1990) investigated circumstances that would create or increase a pro-social orientation.

One problem with the concept of social orientation is that it does not allow for situational influences. Couldn't there be situations in which one is pro-social and others in which one would be quite egoistic? A different approach to social goals incorporates answers to this problem. It consists of assumptions on general human goals that imply different social orientations, most notably the achievement of status (leading to an emphasis on outcome differences), behavioral confirmation (leading to an emphasis on own outcomes) and affection (leading to an emphasis on joint outcomes) (see Lindenberg and Frey 1993). Even if a person has a particular disposition, it now depends on the situation which goal and thus which orientation becomes particularly salient. Note that all three goals are inherently directed towards others.

2.2 Framing Effects

The willingness to cooperate may also depend on how people define a situation. Of particular importance is the difference between gains and losses. Kahneman and Tversky have shown that individuals are keener on avoiding losses than on achieving gains, *ceteris paribus*. This leads to the hypothesis that it is easier to bring about cooperation concerning the defense of a common pool resource than the achievement of a (new) public good. For example, it should be easier to bring about a campaign to preserve a rainforest than a campaign to create a wildlife preserve. The evidence for this is somewhat mixed (see van Assen 1999), but when the disturbing effects are carefully controlled, the framing prediction finds support, especially for initial stages of cooperation (see Sell and Son 1997). There is also a claim that people often cognitively transform a PD into a game with conditional cooperation ('assurance game'). The interesting question then is how the beliefs that others will cooperate are influenced by social circumstances or by cultural beliefs (see, for example, Hayashi et al. 1999), and what influence certain thresholds of cooperation have for triggering conditional cooperation in a larger group (see Yamaguchi 2000).

A different framing effect comes from 'framing through goals' (see Lindenberg 1998). Goals are assumed to have a strong impact on the way individuals define a situation and thereby create a difference between 'foreground' and 'background' information and preferences. What is in the foreground will have a much stronger impact on behavior than what is in the background, *ceteris paribus*. For example, if the goal (the action frame) 'to act appropriately' is salient, then information on normative expectations and ways to meet them are in the foreground, whereas the goal to minimize costs (of meeting these expectations) is in the background. As a result, when the action frame is normative (i.e., driven by the goal to act appropriately), costs of cooperation must be much larger in order to reduce people's willingness to cooperate than when the action frame is, say, gain. In addition, people are likely to signal their action frame to others through their behavior (through 'relational signals') which puts a high premium on including symbolic behavior in the analyses of cooperation, even in game-theoretic analyses.

All in all, in sociology, studies on cooperation are both based on the assumption of rational egoists and on some kind of social rationality. The latter are clearly on the rise in number and sophistication.

See also: Altruism and Self-interest; Cooperation and Competition, Psychology of; Dispute Resolution in Economics; Emotions, Sociology of; Game Theory; Game Theory: Noncooperative Games; Groups, Sociology of; Institutions; Network Analysis; Networks and Linkages: Cultural Aspects; Networks: Social; Norms; Organizational Decision Making; Organizations, Sociology of; Prisoner's Dilemma, One-shot and Iterated; Psychoanalysis in Sociology; Rational Choice Theory in Sociology; Rationality in Society; Social Relationships in Adulthood; Solidarity, Sociology of; Trust, Sociology of

Bibliography

Abell P 1996 A model of the informal structure (culture) of organizations: Help, trust, rivalry and team spirit. *Rationality and Society* **8**: 433–52

Axelrod R 1984 *The Evolution of Cooperation*. Basic Books, New York

Buskens V 1999 *Social Networks and Trust*. Thela Thesis, Amsterdam

Coleman J S 1990 *Foundations of Social Theory*. Harvard University Press, Cambridge, MA

Dasgupta P 1988 Trust as a commodity. In: Gambetta D (ed.) *Trust: Making and Breaking Cooperative Relations*. Blackwell, New York, pp. 49–72

Diekmann A 1986 Volunteer's dilemma. *Journal of Conflict Resolution* **29**: 605–10

DiMaggio P, Louch H 1998 Socially embedded consumer transactions. *American Sociological Review* **63**: 619–37

Doreian P, Fararo T (eds.) 1998 *The Problem of Solidarity: Theories and Models*. Gordon and Breach, Amsterdam

Frey B S 1997 *Not Just For the Money*. Elgar, Brookfield, VT

Friedman J W 1977 *Oligopoly and the Theory of Games*. North-Holland, Amsterdam

Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8

Hayashi N, Ostrom E, Walter J, Yamagishi T 1999 Reciprocity,

trust, and the sense of control. *Rationality and Society* **11**: 27–46
Hechter M 1987 *Principles of Group Solidarity*. University of California Press, Berkeley, CA
Heckathorn D 1996 The dynamics and dilemmas of collective action. *American Sociological Review* **61**: 250–77
Lindenberg S 1998 Solidarity: Its microfoundations and macro-dependence. A framing approach. In: Doreian P, Fararo T (eds.) *The Problem of Solidarity: Theories and Models*. Gordon and Breach, Amsterdam, pp. 61–113
Lindenberg S, Frey B 1993 Alternatives, frames, and relative prices: A broader view of rational choice. *Acta Sociologica* **36**: 191–205
Macy M, Skvoretz J 1998 The evolution of trust and cooperation between strangers. *American Sociological Review* **63**: 638–60
Marwell G, Oliver P 1993 *The Critical Mass in Collective Action*. Cambridge University Press, Cambridge, MA
Olson M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK
Raub W, Voss T 1990 Individual interests and moral institutions: An endogeneous approach to the modification of preferences. In: Hechter M, Opp K-D, Wippler R (eds.) *Social Institutions*. Aldine de Gruyter, New York, pp. 81–117
Raub W, Weesie J 2000 The management of matches: A research program on solidarity in durable relations. *Netherlands' Journal of Social Sciences* **36**: 71–88
Sell J, Son Y 1997 Comparing public goods with common pool resources: Three experiments. *Social Psychology Quarterly* **60**: 118–37
Snijders C 1996 *Trust and Commitment*. Thela Thesis, Amsterdam
Taylor M 1987 *The Possibility of Cooperation*. Cambridge University Press, Cambridge, UK
Uzzi B 1996 The source and consequences of embeddedness for the economic performance of organizations. The network effect. *American Sociological Review* **61**: 674–98
Van Assen M F 1999 Effects of individual decision theory assumptions on predictions of cooperation in social dilemmas. *Journal of Mathematical Sociology* **24**: 1–11
Weesie J, Buskens V, Raub W 1998 The management of trust relations via institutional and structural embeddedness. In: Doreian P, Fararo T (eds.) *The Problem of Solidarity: Theories and Models*. Gordon and Breach, Amsterdam, pp. 113–38
Williamson O E 1985 *The Economic Institutions of Capitalism*. Free Press, New York
Yamagishi T 1986 The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* **51**: 110–16
Yamaguchi K 2000 Subjective rationality of initiators and of threshold-theoretical behavior of followers in collective action. *Rationality and Society* **12**: 185–225

A. Diekmann and S. Lindenberg

Cooperative Learning in Schools

Cooperative learning, instructional methods in which students work in small groups to learn academic material, is one of the most extensively studied and widely used classroom innovations. Hundreds of studies have compared cooperative learning to other instructional methods on various social and academic measures. By far the most frequent objective of cooperative learning research has been to harness its effects on student achievement. Field and laboratory studies of the achievement effects of cooperative learning have been conducted in every major subject, at every grade level, and in all types of schools the world over. Cooperative learning is not only the subject of research and theory, the knowledge generated by that labor is used at some level by hundreds of thousands of teachers. A national survey of US teachers found that 79 percent of elementary teachers and 62 percent of middle school teachers reported making some sustained use of cooperative learning (Puma et al. 1993).

While there is little doubt about the facilitating effect of certain forms of cooperative learning on student achievement, there is still great disagreement over how and why cooperative learning affects achievement. Perhaps more importantly, there is intense debate over the conditions under which cooperative learning practices are effective. Some researchers describe theoretical mechanisms underlying the achievement effects of cooperative learning that are completely different and at odds with those assumed or described by others. In particular, some scholars emphasize the role of incentive structure while others attribute achievement effects to students' interactions in the types of task structures used in cooperative learning. Though often working in parallel, there has been remarkably little cross-fertilization in the work of cooperative learning researchers. Applications and investigations of cooperative learning emerging from different perspectives typically differ in their operationalization of aspects of both the incentive and task structures. As a result, it is difficult to determine which is responsible for which outcomes.

This article summarizes and updates earlier discussions of the major theoretical perspectives that seek to explain the facilitating effects of cooperative learning (Slavin 1995). It describes motivational, social cohesion, and cognitive elaboration perspectives on cooperative learning and explores the conditions under which each may operate. Finally, it suggests future research that will be needed to sustain the study and practice of cooperative learning in the twenty-first century.

1. Four Major Theoretical Perspectives on Cooperative Learning and Achievement

1.1 Motivational Perspectives

Motivational perspectives on cooperative learning emphasize the reward and/or goal structure of the

learning environment (Slavin 1995). In order to achieve personal goals, each student must attend not only to his or her own efforts but to those of other group members as well. Cooperative incentive structures (e.g., group rewards based on the learning of all group members) encourage students to engage in behaviors that help the group achieve its goals. These behaviors include maintaining self-motivation, giving or withholding praise, sanctioning, assisting, and encouraging others. According to the motivational perspective, the beneficial effects of cooperative learning for achievement are primarily due to the incentive provided for students to encourage and help one another to learn.

The motivationalist critique of traditional classroom organization holds that competitive grading and the informal reward system common in classrooms creates and maintains negative goal interdependence among students. Since one student's success lessens the opportunities others will have to succeed, students may hope for others' failure. They are discouraged from supporting each other's efforts. Such conflicts may ultimately lead to peer norms opposing academic achievement.

Motivational theorists build criterion-based group rewards into their cooperative learning methods. In strategies developed at Johns Hopkins University (Slavin 1995), students work to earn certificates and other recognition for their efforts if the average score of each team member exceeds a pre-established criterion (see also Kagan 1995). Methods developed by Johnson and Johnson (1994) and their colleagues often give grades based on group performance. The theoretical rationale for these group rewards is that if students are encouraged to value the success of the group, they will encourage and help one another to achieve. This mode of operation stands in contrast to the negative goal interdependence common in traditional competitive classrooms.

1.1.1 Empirical support for the motivationalist perspective. There is a great deal of empirical support for the motivationalist assertion that group rewards are essential to the success of cooperative learning in classrooms, with one critical qualification. A number of studies have concluded that the use of group goals/rewards enhances achievement only in cases where the group reward is based on the individual learning of each student in the group (Slavin 1995). Most often this means that team scores are derived from quizzes that the team members take individually, without help. When group rewards are given based on a single group product (i.e., without individual accountability), there is little incentive for students to cooperate and one or two group members may do all of the work. In student team-achievement divisions (STAD) (Slavin 1995), students work in mixed ability teams to master material initially presented by the teacher. After working in their teams, students take individual quizzes on the material. Teams can only succeed if they make sure that all team members have learned, so the work sessions focus on team members explaining concepts to one another.

A review of 99 cooperative learning studies that involved durations of at least four weeks compared achievement gains in cooperative learning and control groups. Of 64 studies of cooperative learning methods that provided group rewards based on the sum of group members' individual learning, 50 (78 percent) found significant positive effects on achievement and none found negative effects (Slavin 1995). The median effect size for those studies for which effect sizes could be computed was $+0.32$. Studies which gave rewards based on a single group product, or which gave no rewards, found few positive effects, with a median effect size of only $+0.07$. Studies which compared different modes of cooperative goal structure within the same study found similar patterns. Group goals with rewards based on individual learning were important to the instructional effectiveness of cooperative learning models.

1.2 Social Cohesion Perspective

According to Cohen (1994 pp. 69–70), 'if the task is challenging and interesting, and if students are sufficiently prepared for skills in group process, students will experience the process of group work itself as highly rewarding… never grade or evaluate students on their individual contributions to the group or product.' Cohen's work may be described as social cohesion theories. They emphasize that the effects of cooperative learning on achievement are largely mediated by the cohesiveness of the group. As such, they downplay the role of incentives and accountability (Cohen 1994, Johnson and Johnson 1996). As such, their work often emphasizes team-building activities in preparation for cooperative learning and processing or group self-evaluation during and after group activities.

Cohen et al. (1992) all use forms of cooperative learning in which students take on individual roles within the group. A major purpose of these techniques is to create interdependence within the group and in the classroom. Slavin (1983) terms this 'task specialization.' In Aronson's Jigsaw method (1978), students study material on one of several topics distributed among group members. They meet in expert groups to share information on their topics with members of other teams who have been assigned the same topic, and then take turns presenting their topic to their respective teams. In Sharan's group investigation method, groups take on topics from a unit presented to the entire class and further subdivide them among members of the group. They then

investigate their assigned topics as a group and ultimately present their findings to the class.

The social cohesion and motivational perspectives differ in that social cohesion theorists assume that students will work together because, and to the extent that, they care about each other, the group, and achievement. Motivational theorists, on the other hand, assume that students help their classmates at least in part because it serves their own interests to do so. Both perspectives focus primarily on motivational rather than cognitive explanations for the instructional effectiveness of cooperative learning. Both models treat motivation as the critical variable to be manipulated toward fostering achievement and both regard cooperation as a way to affect student motivation. The relationship (and tension) between the social cohesion and motivationalist perspectives is apparent in the work of Johnson and Johnson (1994). While their models do make use of group goals and group incentives such as recognition and grades, their own and related theoretical writings emphasize the development of group cohesiveness and social group processing strategies.

1.2.1 Empirical support for the social cohesion perspective. The achievement outcomes of cooperative learning methods that emphasize social and task variables are inconsistent. Concerning task specialization, there is evidence that, when well implemented, group investigation can significantly increase student achievement (e.g., Sharan and Sharan 1992). However, research on the original form of Jigsaw has not typically yielded positive effects on achievement, except in cases where group rewards have been added to the original model (Mattingly and Van Sickle 1991). In studies of at least four weeks' duration, the Johnson and Johnson (1994) methods have not been found to increase achievement more than individualistic methods unless they incorporate group rewards (grades) based on the average of group members' individual quiz scores (see Slavin 1995).

Research on the practical applications of methods based on social cohesion theories provide inconsistent support for the proposition that building cohesiveness among students through team-building without group reward incentives will enhance student achievement. Yager et al. (1986) found that group processing activities did enhance the achievement effects of cooperative learning. However, other investigations of the Jigsaw model have found that team building activities had no effect on achievement outcomes.

With the exception of the group investigation model, there is inconsistent evidence that social cohesion and group process forms of cooperative learning that do not include individual accountability are superior to traditional instruction. Further, group investigation methods include an evaluation based on group efforts that is composed of individual team member contributions. This may amount to a type of group reward with individual accountability, and may thus reconcile the findings with the assertions of motivationalist theory.

1.3 Cognitive Perspectives

The cognitive perspectives offer explanations for the beneficial effects of cooperative learning which are quite different from those presented thus far. Under the general heading of cognitive perspectives, these theories can be said to place the value of cooperative learning in the interactions among students during group work. These interactions are said to affect students' mental processing of information to be learned rather than their motivation to learn the material. Within that general heading some work is focused on cognitive development, while other work concerns cognitive processes more generally.

Developmentalist theorists work under the fundamental assumption that children's interactions around appropriate tasks increase their mastery of critical concepts. Vygotsky's 'zone of proximal development' is the distance between the child's actual developmental level as determined by independent problem solving, and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers (Vygotsky 1978, p. 86). Collaborative activity among children promotes growth because children of similar ages are likely to function in each other's zones of proximal development, modeling in the collaborative behaviors more advanced than each could perform as individuals.

Similarly, Piaget (1926) held that social-arbitrary knowledge—language, values, rules, morality, and symbol systems—can only be learned in interactions with others. Interaction is also essential to logical and mathematical thought development in disequilibrating a child's conceptions and in providing feedback about the validity of logical constructions. For example, if a child has an incorrect conception, interaction with peers increases the chances that the child will move to a more sophisticated understanding. There is considerable support for the notion that peer interaction can help nonconservers become conservers, and for the importance of peers working in each other's zones of proximal development.

On the basis of these and other findings, Piagetian and Vygotskian, theorists have prescribed increased use of cooperative activities in schools. Damon (1984) draws on these ideas in proposing a conceptual foundation for peer-based learning which explicitly rejects the use of extrinsic incentives. Nonetheless, despite wide theoretical and laboratory support for the cognitive models of cooperative learning, there is as yet little evidence from classroom experiments done for meaningful intervals that 'pure' cooperative meth-

ods, without group rewards and individual accountability, produce higher achievement. Still, it seems likely that the cognitive processes described by developmental theorists underlie the effectiveness of cooperative learning in any form; motivation to help or encourage others is likely to increase the quantity or quality of peer interactions, which in turn lead to cognitive growth.

1.4 Cognitive Elaboration Perspective

Research in cognitive psychology has long held that that if information is to be retained and integrated in memory, the learner must engage in some sort of cognitive restructuring or elaboration of the material. It is also widely accepted that one of the more effective means of elaboration is explaining the material to another. Noreen Webb (1989) found that the students who gained the most from cooperative activities were those who gave elaborated explanations to others. This contention also is supported in empirical work concerning peer tutoring, and cooperative scripts. Classroom investigations have yielded similar findings in writing instruction and in reciprocal teaching for reading comprehension.

2. Reconciling the Perspectives

The perspectives discussed above are typically debated in the theoretical and empirical literature as if they were contradictory. They arise out of deeply differing metatheoretical orientations, each with well-established theoretical bases and sufficient supporting evidence to assure that they apply in some circumstances. By the same token, their variety assures that none of them is both necessary and sufficient in every setting. The debate is complicated because research in each tradition tends to establish settings and conditions that favor that perspective. For example, most cooperative learning research from the motivationalist and social cohesion perspectives is conducted in classrooms where students may be accustomed to working for extrinsic rewards, and over extended time periods since social cohesion may be expected to take time to develop.

In contrast, studies from developmental and cognitive elaboration views tend to be brief, making issues of motivation moot. These latter paradigms also tend to use pairs, rather than groups of three or more; pairs may involve more simple interaction dynamics and require less time to develop modes of working together. Developmental research almost exclusively uses young children and traditional Piagetian tasks which bear little resemblance to the social-arbitrary learning that characterizes most classroom learning. Cognitive elaboration research mostly involves college student samples.

The various perspectives should be seen as complementary, not contradictory. While cognitively oriented strategies may not explicitly focus on motivation, motivational strategies invariably include peer interaction and cognitive engagement among groupmates. Figure 1 depicts a path through which group goals might directly affect students' behavior to in turn enhance cooperative learning outcomes (from Slavin 1995).

Motivational theorists would not argue that the processes prioritized in cognitive models are unimportant, they simply argue that motivation drives such cognitive processes, which in turn produces learning

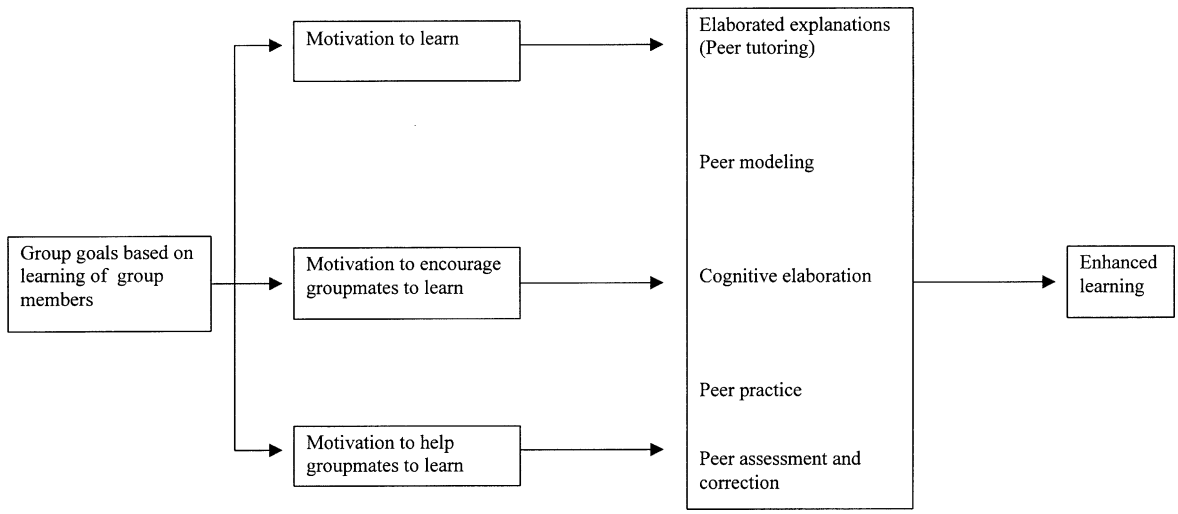

Figure 1
Effect of group goals on students' behavior with respect to cooperative learning outcomes

effects. They would argue that over an extended period, students may not engage in the types of elaborated explanations that Webb (1989) credits for cooperative learning effects, without strategies directed at enhancing motivation. By extension, motivational theorists would hold that extrinsic incentives might affect achievement indirectly, by motivating students to engage in peer modeling, cognitive elaboration, and/or practice with one another. Extending this chain of influence farther, the provision of group goals may lead to group cohesiveness, increase caring, and concern among group members, and make groupmates feel responsible for one another's achievement, thus motivating students to engage in cognitive processes that enhance learning. In addition, group goals may motivate students to take responsibility for one another independently of the teacher, thereby solving important classroom organization problems and providing increased opportunities for cognitively appropriate learning activities.

Referring to the model in Fig. 1, researchers from the cognitive perspective attempt to intervene directly on mechanisms considered mediating variables in the motivational view. For example, social cohesion theorists attempt to directly increase group cohesiveness by engaging in elaborate teambuilding and group processing training. One group investigation study suggests that this can be done successfully, but that it takes a great deal of time and effort. In this study, teachers were trained over the course of a full year, and then teachers and students used cooperative learning for three months before the study began. Earlier research on group investigation failed to provide a comparable level of preparation of teachers and students, and the achievement results of these studies were less consistently positive.

Cognitive theorists would hold that the cognitive processes that are essential to any theory relating cooperative learning to achievement can be created directly, without the motivational or affective changes discussed by the motivationalist and social cohesion theorists. This may turn out to be accurate, but at present demonstrations of learning effects from direct manipulation of peer cognitive interactions have mostly been limited to studies of very brief durations and to tasks that lend themselves to the cognitive processes involved. Most of the Piagetian conservation tasks studied by developmentalists have few practical analogs in the school curriculum. At the same time, there is research on reciprocal teaching in reading comprehension which shows promise as a means of intervening directly in peer cognitive processes. Long-term applications of Dansereau's (1988) cooperative scripts for building comprehension of technical material and procedural instructions also seem likely to be successful.

The previous discussion has summarized evidence that generally supports the motivationalist view that group goals and individual accountability can be important to realizing the potential benefits of cooperative learning. This position is by no means absolute and there are conditions under which achievement gains have been found for cooperative learning treatments that lacked group goals, individual accountability, or both. Structured dyadic tasks (e.g., Dansereau 1988) and voluntary study groups, as are often seen among college and graduate school students, are two other instances where factors inherent to the task or scenario may affect students sufficiently to render instructor-imposed group goals and individual accountability unnecessary. By the same token, there is little evidence that the presence of group goals or individual accountability would reduce the outcomes of these methods, and they are likely to enhance their effects. There is a clear need to explore such cases further, if only because the deeply rooted mistrust among educators for extrinsic incentives makes reward-free cooperative learning a practically valuable goal.

The practice of cooperative learning is entering a new phase. On one hand, it is so widely used and so infused in preservice and inservice education that it is no longer seen as an innovation. On the other, much of the application of cooperative learning is informal, unstructured, and inconsistent with the principles of effectiveness derived from any of the schools of thought described here. Cooperative learning often means students sitting together with no particular goal or structure. Where cooperative learning is most appropriately used is in more structured, comprehensive approaches that provide extensive professional development and materials designed specifically for use in cooperative learning. Examples include reciprocal teaching (Webb and Palincsar 1996), writing process, Cooperative Integrated Reading and Composition, and Success for All (Slavin and Madden 2001). These and related programs may represent the future of cooperative learning in the twenty-first century, where cooperative learning is becoming a basic element of comprehensive reform rather than only a generic process to be applied to every subject and grade level to the extent teachers wish to do so.

Research on cooperative learning is needed both to further understand cognitive and motivational mechanisms underlying its effects on achievement and to use this research to develop and evaluate new and increasingly effective forms of cooperative learning. After more than 30 years of research, the potential power of cooperative learning has been demonstrated and causal mechanisms have been extensively explored, but there is still much to be done to achieve the full potential of students working with students to achieve.

The evidence is strong indicating that cooperative learning has the potential to serve as a foundation upon which education in the twenty-first century will be built. Research must strive to provide practical, theoretical, and intellectual guidance to the efforts of

educators the world over in the service of both traditional and innovative goals.

See also: Classroom Climate; Cooperation and Competition, Psychology of; Environments for Learning; Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Processes in the Classroom; Learning Theories and Educational Paradigms; Learning to Learn; Motivation: History of the Concept; Motivation, Learning, and Instruction; Open Classroom: Learning and Teaching; School Achievement: Cognitive and Motivational Determinants; School Learning for Transfer; Situated Learning: Out of School and in the Classroom

Bibliography

Aronson E, Blaney N, Stephen C, Sikes J, Snapp M 1978 *The Jigsaw Classroom*. Sage Publications, Beverley Hills, CA

Cohen E G 1994 *Designing Groupwork: Strategies for the Heterogeneous Classroom*, 2nd edn. Teachers College Press, New York

Damon W 1984 Peer education: The untapped potential. *Journal of Applied Developmental Psychology* **5**: 331–43

Dansereau D F 1988 Cooperative learning strategies. In: Weinstein C E, Goetz E T, Alexander P A (eds.) *Learning and Study Strategies: Issues in Assessment, Instruction, and Evaluation*. Academic Press, San Diego, CA, pp. 103–20

Johnson D W, Johnson R T 1994 *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning*, 4th edn. Allyn & Bacon, Boston

Johnson D W, Johnson R T 1996 Cooperative learning and traditional American values: An appreciation. *NASP Bulletin* **59**: 63–5

Kagan S 1995 *Cooperative Learning Structures for Classbuilding*. Kagan Cooperative Learning, San Juan Capistraumo, CA

Mattingly R M, Van Sickle R L 1991 Cooperative learning and achievement in social studies; Jigsaw II. *Social Behavior* **55**(6): 392–5

Piaget J 1926 *The Language and Thought of the Child*. Harcourt Brace, New York

Puma M J, Jones C C, Rock D, Fernandez R 1993 *Prospects: The Congressionally Mandated Study of Educational Growth and Opportunity: The Interim Report*. Abt Associates, Washington DC

Sharan Y, Sharan S 1992 *Group Investigation: Expanding Cooperative Learning*. Teacher's College Press, New York

Slavin R E 1983 *Cooperative Learning*. Longman, New York

Slavin R E 1995 *Cooperative Learning: Theory, Research, and Practice*, 2nd edn. Allyn & Bacon, Boston

Slavin R E, Madden N A 2001 *One Million Children: Success for All*. Corwin, Thousand Oaks, CA

Vygotsky L S 1978 *Mind in Society*. Harvard University Press, Cambridge, MA

Webb N M 1989 Peer interaction and learning in small groups. *International Journal of Educational Research* **13**: 21–39

Webb N M, Palincsar A S 1996 Group processes in the classroom. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Simon & Schuster Macmillan, New York, pp. 840–73

Yager S, Johnson R T, Johnson D W, Snider B 1986 The impact of group processing on achievement in cooperative learning groups. *Journal of Social Psychology* **126**: 389–97

R. E. Slavin, E. A. Hurley, and A. M Chamberlain

Coping across the Lifespan

'… how a man rallies to life's challenges and weathers life's storms tells everything of who he is and all that he is likely to become' (Saint Augustine).

Coping did not appear as a term in *Psychological Abstracts* until 1967. Since that time, research has been voluminous. A search of *PsycINFO* on 'coping' from 1967 to 1998 yielded about 23,000 entries, making it one of the fastest growing areas in psychology (Sommerfield and McCrae 2000). Studies span infancy to oldest age, and examine coping in almost every imaginable life domain, including health, work, finances, family, parenting, school, peers, bereavement, unemployment, and aging, as well as nonnormative life events, like accidents, crime victimization, cultural upheaval, economic downturns, war, and relocation. (See *Chronic Illness, Psychosocial Coping with*; *Coronary Heart Disease (CHD), Coping with*; *Disasters, Coping with*; *Job Stress, Coping with*; *Illness: Dyadic and Collective Coping*; *Well-being and Health: Proactive Coping*.)

1. Definitions

Coping is a process integral to adaptation and development. As an organizational construct, coping includes 'all those efforts to deal with environmental pressures that [can] not be handled by reflexes or organized skills, but [involve] struggles, trials, persistent focused energy directed toward a goal. So responses to threats and dangers, frustrations and defeats, obstacles, loss, strangeness and the new or unknown, demands from adults and others in the environment, all required study' (Murphy 1974, p. 71). No consensus exists about a more precise definition. Given the current prominence of transactional views, coping is often defined as 'constantly changing cognitive and behavioral efforts to manage specific external and/or internal demands that are appraised as taxing or exceeding the resources of the person' (Lazarus and Folkman 1984, p. 141). Closely related terms include stress, stressor, stressful life events, defense, challenge, threat, adversity, risk, resilience, mastery, vulnerability, thriving, hardiness, and social support. (See *Stress and Coping Theories*.)

Coping incorporates processes from many levels. It

reflects evolution; humans come prepared to recognize and react to environmental demands in ways that promote survival. Coping has a tightly integrated physiological, psychological, and social basis. Many features of psychological functioning, such as emotions, motivation, attention, volition, cognition, and communication are organized in ways that allow them to contribute to coping. For example, emotions signal and organize interactions of adaptive significance. Many aspects of social relationships and cultural systems also contribute to coping. For example, attachment relationships provide a safe haven of protection during times of distress.

Coping has both normative and individual difference features. Normatively, humans are prepared to cope adaptively. In fact, stress is likely necessary for the kinds of sustained constructive interactions that facilitate development. However, human responsiveness to stress is also a vulnerability; too much stress can overwhelm and damage people. In addition, there are enormous differences in how physiological, psychological, social, and cultural resources for coping are distributed across individuals and social groups.

Finally, coping is a process that unfolds over time. Coping takes real time; any specific stressful transaction involves sequences of moment-to-moment interactions. Coping also takes place over episodic time; dealing with demands encompasses a protracted series of specific real-time transactions. As noted by White, 'Described not inappropriately in military metaphors, adaptation often calls for delay, strategic retreat, regrouping of forces, abandoning of untenable positions, seeking fresh intelligence, and deploying new weapons' (White 1974, p. 50).

Given its complexity and centrality to adaptation, it is not surprising that coping has been approached from a wide range of theoretical perspectives. It has been defined as a specific person–context transaction, personality in action under stress, a repertoire of strategies, a hierarchically organized set of ego processes, an indicator of competence, a function of emotion, an outcome of temperament, an expression of stress physiology, and a quality of action regulation. Although it has sometimes been suggested that coping is primarily a manifestation of other phenomena, most researchers vigorously argue that coping processes have an independent causal status; more specifically, they maintain that coping makes a material difference to how stressful interactions are resolved and is a key factor influencing long-term mental and physical heath, well-being, and development in the face of adversity. (See *Stress and Coping Theories*.)

2. *Historical Contexts*

Historically, work on coping started well before the 1960s. It has early forerunners in psychoanalytic work on defenses and has been influenced heavily by the study of stress. Work with children and the elderly reflects a focus on development. Theories have been informed by a host of interventions designed to help people of all ages cope more adaptively with trauma and stressful life events. (For historical overviews, see Lazarus 1993, Lazarus and Folkman 1984, Murphy 1974, Parker and Endler 1996, Snyder 1999.)

2.1 *Coping and Defense*

The concept of defense, as it first appeared in psychology in the early writings of Freud ([1894] 1962), referred to psychological mechanisms (such as repression or projection) used to protect the psyche from disturbing thoughts and emotions, mainly by distorting reality or removing the distressing material from conscious awareness. These ideas were expanded by Anna Freud (1936/1946), who argued that individuals have particular styles of defense which are linked to specific forms of psychopathology.

This formulation influenced several generations of ego-psychologists (e.g., Haan 1977, Valliant 1986, see Cramer 1998). These theorists argued explicitly for a hierarchical model of ego processes, in which some defenses are more mature than others. For example, Haan (1977) posited a three-tiered taxonomy of ego-processes: coping, considered more forward looking, flexible, largely conscious, and attentive to reality; defense, considered to be inherently organized around issues from the past, rigid in operation, unconscious, and distorting of reality; and ego-failure or 'fragmentation.'

Many notions from this approach appear in current theorizing. They include the idea that coping occurs in response not only to environmental demands, but also to intrapsychic pressures; that some ways of coping are inherently superior; and that people consistently show characteristic styles of coping with threat that reflect personality or ego development. Moreover, this work introduced several themes which are resurfacing in current discussions of coping, including the idea that some modes of adaptation are unconscious and unintentional, that distortions of reality can be adaptive, and that the self (or ego) and its regulatory functions are central to processes of coping.

2.2 *Stress and Coping*

A second strand of work on coping emerged from research on stress, a concept prominent in the social and health sciences since the 1930s. Systematic attempts to link stress exposure to specific psychological or somatic outcomes revealed marked individual differences in stress effects. Explicit recognition that living organisms 'fight back,' as captured in terms like 'host resistance,' galvanized interest in the social and individual factors that determine vulnerability or

resistance to the effects of stress. An important branch of this work contributed to the study of coping (Coelho et al. 1974, Lazarus and Folkman 1984, Moos and Billings 1982, Pearlin and Schooler 1978). (See *Stress and Coping Theories*.)

The transactional perspective, currently the most prominent view, was articulated by Lazarus and his colleagues, and held that the experience of stress is shaped by interpretations or appraisals of the meaning of environmental demands. Coping itself is seen as a transactional process, iterating with (re)appraisals and environmental changes, and as having two primary functions: to deal directly with the environmental stressor (referred to as problem-focused coping) and to deal with the emotional consequences of the stressful transaction (emotion-focused coping).

The basic tenets of a transactional perspective include the notion that coping does not directly reflect personality but varies in accordance with the situations in which it occurs; that no ways of coping are inherently better; that any stressful event (such as illness or job loss) gives rise to a complex host of specific and changing demands; and that, in order to cope with these, any individual can and does use a wide variety of strategies. Perhaps most critical is the emphasis on coping as a process that changes over time.

Several methodological shifts in research accompanied this new approach. The measurement of coping shifted from clinical assessment to self-report, and efforts were made to unconfound measures of coping from assessment of outcomes. In addition to a focus on coping with traumatic events (e.g., life-threatening illness or victimization), research also came to include the study of chronic stressful conditions, and relatively minor everyday stressors or 'daily hassles.' Moreover, research became more specialized by type of stressor (e.g., health problems, work stress, or bereavement; see *Disasters, Coping with*; *Job Stress*, *Coping with*). Even within domains, research has become more finely segregated; for example, different medical conditions are often studied separately. (See *Chronic Illness, Psychosocial Coping with*; *Coronary Heart Disease (CHD), Coping with*; *Illness: Dyadic and Collective Coping*; *Well-being and Health: Proactive Coping*.)

2.3 Coping and Development

Child psychologists have had a long-standing interest in the impact of adversity on children and, since the early 1900s, have attempted to document the effects, for example, of maternal deprivation, serious illness, hospitalization, and wartime conditions, as well as more recent attention to the effects of parental mental or physical illness, unemployment, poverty, and divorce. The first formal study of children's coping is often traced to Murphy and her colleagues at the Menninger Clinic, who used intensive observations and interviews to conduct an 18-year longitudinal study of ways in which normal children cope (Murphy and Moriarity 1976).

The 1980s witnessed an explosion of quantitative work, sparked by the publication of seminal works on stress and coping in children by Garmezy and Rutter (1983) and by Compas (1987). Although the current wave of research on children closely resembles work on adults, it nevertheless highlights several important themes. It makes evident that coping is inherently shaped by social partners, close relationships, families, and communities (Hetherington and Blechman 1996) and that cognitive coping strategies, so prominent in adults, are not commonly used by children until late childhood. Research on children also encourages a renewed focus on physiological and temperamental antecedents of coping, as well as attention to coping with normative developmental tasks (Wolchik and Sandler 1997).

Finally, a shift in focus to development across the lifespan spurred interest in how adults successively deal with life transitions, social change, and the challenges of aging (Aldwin 1994, Baltes and Baltes 1990, Brandtstadter et al. 1999). Research on successful aging, in documenting that most people maintain a high level of psychological functioning despite objective biological declines and social losses, focuses attention on positive coping strategies, like selection, compensation, planning, and proactive coping. It also highlights the importance of 'accommodative' modes of coping which allow people to relinquish goals and withdraw from goal pursuit.

2.4 Interventions to Improve Coping

A relatively distinct strand of work involves psychological and behavioral medicine interventions which have succeeded in helping people deal adaptively with traumatic events and their associated ongoing stressors. In contrast to traditional therapy, these interventions focus on providing information and social support, and on helping people build coping resources and skills for dealing more effectively with particular stressors, such as illness, chronic pain, divorce, or bereavement (Sandler et al. 1997).

3. Research on Coping

Since the advent of the transactional approach, mainstream research has focused on the correlates of coping (Compas et al. 2001). Typical studies examine a single population at a single time, assess self-reported ways of coping (e.g., problem solving, help seeking, avoidance) with either a variety of unidentified stre-

ssors or a prespecified narrow class of stressors, and then examine their associations with self-reports of personal and social predictors (e.g., perceived control or social support) and with various outcomes (usually indicators of mental or somatic functioning, such as emotional adjustment or physical symptoms).

This work has identified several clusters of adaptive coping. One is organized around problem solving (information seeking, strategizing, planning, prevention) and seems to be supported by opportunities for control and perceptions of efficacy. (See *Hope and Hopelessness*.) A second is organized around relying on trusted others for comfort, discussion, help, advice, and aid; these ways of coping are supported by the presence of warm and caring social partners and the perception that they are available. A possible third cluster is organized around processes of accommodation (negotiation, compromise, prioritizing, and goal setting) and may be supported by a sense of self-determination and autonomy support from others. In addition, several maladaptive ways of coping have been identified, including catastrophizing and panic, social isolation, venting, rumination, blame of self or others, and escape. These ways of coping are exacerbated by certain situational features and interpretational biases. Such reactions to stress are maladaptive because they tend to exhaust personal and social resources and to interfere with recruiting and implementing effective action.

As a body, however, this work has come under attack on both conceptual and methodological grounds. Studies have been criticized for being light on theory, for inconsistent use of coping categories, for relying heavily on self-reports, for using designs that cannot examine change, and for including only a narrow range of short-term outcomes. (See *Coping Assessment*.) As noted by Sommerfield and McCrae (2000, p. 620), 'The seemingly boundless enthusiasm for coping research seen in the 1980s has been replaced by widespread disaffection, intense scrutiny, and corresponding calls for change.' Cutting-edge research has already moved away from this empirical format and reflects the use of several interesting new approaches.

3.1 Categories of Coping

One factor impeding research has been widespread disagreement about taxonomies of coping. No consensus exists about core categories. Years of exploratory factor analyses have failed to converge on a set of higher-order categories, nor do theorists agree on systems for rational classification. Given the seemingly infinite variety of specific coping responses, even the possibility of identifying a core subset of ways of coping has been questioned. (See *Coping Assessment*.)

One promising approach uses theory and confirmatory factor analysis to identify multiple core categories. For example, researchers attempt to identify a finite set of basic adaptive processes and then to classify multiple heterogeneous ways of coping as belonging to a 'family' of ways of coping that serve each function. Examples of theories used in these attempts include functionalist theories of emotion, theories of motivation, regulation, primary and secondary control, and ego processes. Such discussions expand on the distinction between problem-focused vs. emotion-focused coping and on coping styles (e.g., repression vs. sensitization) as well as reopening discussion of the functions of coping in adaptation.

3.2 Individual Ways of Coping

One productive empirical strategy has been to make particular categories of coping the focus of detailed process-oriented programs of laboratory and naturalistic research. Such studies have produced relatively well-accepted conclusions about positive ways of coping, such as mastery-oriented thinking, disclosure and discussion, constructive thinking, accommodation, and optimism, and about the generally maladaptive consequences of rumination, helplessness, blaming others, catastrophizing, and social isolation. Programs of research on other ways of coping, such as help- or support-seeking, positive illusions, avoidance, denial, and relinquishing control, have produced less conclusive results, including evidence that their effects depend on characteristics of the situations in which they are employed.

It should be noted that almost every activity that is considered a way of coping also has its own well-developed body of research. These include problem solving, information seeking, planning, negotiation, compliance, self- and other-blame, help seeking, social support, opposition, venting, selective optimization, compensation, rumination, disclosure and discussion, primary and secondary control, optimism, and positive illusions. Most of these areas of research are largely independent of work on coping, although some examine the functioning of these processes under stress.

3.3 Coping as a Process

Despite consensus that coping unfolds over time, research methodologies able to capture process are just beginning to (re)appear. These include case studies, ethnographic narratives, interviews, and observations, which were popular decades ago (Folkman 1997, Murphy and Moriarity 1976, Valliant 1986) as well as more recent quantitative methods, such as intra-individual time series analysis (Tennen et al. 2000). Designs are longitudinal, repeating measures over short times (e.g., daily) or over several years, often including markers of the progress of stressful

events (e.g., diagnosis and surgery, or caretaking, deterioration, and death of a partner).

For microgenetic approaches, patterns of intra-individual relations among variables (e.g., problem- and emotion-focused coping) across time are determined. Then interindividual difference variables (e.g., depression or optimism) are used to form groups which are compared on their patterns of intraindividual relationships. Observations of coping interactions with social partners (e.g., mothers and their children) can also be used to examine the sequential effects of individual coping efforts and partner reactions over time. Narratives and interviews capture a wider range of experiences and can track their progress over longer intervals. The yield from this labor-intensive research seems promising. For example, such studies already suggest that both generation of and focus on positive experiences are critical in dealing constructively with chronic stress; that depression influences day-to-day reactions to chronic pain; and that maternal interpretations of negative events shape children's subsequent appraisals and coping.

Additional strategies for capturing coping processes may be suggested by related research, in which fine-grained studies focus on how children and adults deal with certain classes of environmental demands, such as noncontingency, failure, unpredictability, separation, loss, interpersonal conflict, competition, and coercion. Relevant studies come from areas as diverse as learned helplessness, self-efficacy, perceived control, reactance, self-determination, compliance, attachment, interpersonal problem solving, attributional biases, Type-A personality, and self-regulated learning. Many consider the study of coping as a process to be the most important goal of the next generation of coping researchers (Lazarus 1993).

3.4 Coping as Regulation

An explicit theme resurfacing in discussions and research is the notion of coping as an inherently regulatory process (Aspinwall and Taylor 1997, Barrett and Campos 1991). A focus on adaptation suggests that the cognitions, appraisals, motivation, volition, behaviors, and emotions, which are clearly parts of coping, must be assembled into actions which are sequentially coordinated with internal and external conditions. Labels under which such processes are studied include ego functioning (e.g., ego-control and ego-resilience), conscience, functionalist theories of emotion, and theories of volition and regulation of behavior, attention, emotion, and action. Conceptions of coping as a dynamic regulatory process may contribute to the next phase of coping work. Features of the self involved in recognizing and creating 'relational meaning' (Lazarus, 1993) may be studied as integral organizers of regulation. (See *Self-regulation in Childhood*; *Self-regulation in Adulthood*.)

3.5 Future Directions

'Stress' is not inherently damaging to people, and the primary function of coping is not simply to shield people from stressful experiences. Coping is a process which is intrinsic to psychological and relational growth and one which individuals and groups are predisposed to do well. Current interest in positive coping (Snyder 1999) and the study of growth from adversity or transformation in the face of trauma, can be viewed as attempts to conceptualize and study the growth facilitating functions of coping. Analysis of the role of coping in resilience in children (Haggerty et al. 1994) and the relationship of coping and development across the lifespan (Aldwin 1994, Skinner and Edge 1998) may help put these issues in a larger perspective. Future research will likely feature process-oriented studies in both naturalistic and lab settings, using a variety of assessments (interviews, narratives, daily reports, observations) and examining intra-individual relations and change over time. The promise of the area of coping, and its challenge to researchers, is to examine coping as an engine of development or a catalyst for change, to determine if conceptions of coping can provide a process mechanism which explains how people (individually and collectively), in the face of adversity, are able to find and create opportunities for development.

See also: Adulthood: Developmental Tasks and Critical Life Events; Chronic Illness, Psychosocial Coping with; Coping Assessment; Social Support and Stress; Stress and Coping Theories; Stress and Health Research

Bibliography

Aldwin C M 1994 *Stress, Coping, and Development: An Integrative Perspective*. Guilford, New York

Aspinwall L G, Taylor S E 1997 A stitch in time: Self-regulation and proactive coping. *Psychological Bulletin* **121**: 417–36

Baltes P B, Baltes M M (eds.) 1990 *Successful Aging: Perspectives from the Behavioral Sciences*. Cambridge University Press, Cambridge, UK

Barrett K C, Campos J J 1991 A diacritical function approach to emotions and coping. In: Cummings E M, Greene A L, Karraker K H (eds.) *Life-span Developmental Psychology: Perspectives on Stress and Coping*. Erlbaum, Hillsdale, NJ, pp. 21–41

Brandtstadter J, Wentura D, Rothermund K 1999 Intentional self-development through adulthood and later life: Tenacious pursuit and flexible adjustment of goals. In: Brandtstadter J, Lerner R M (eds.) *Action and Self-development: Theory and Research Through the Life-span*. Sage, Thousand Oaks, CA, pp. 373–400

Coelho G V, Hamburg D A, Adams J E (eds.) 1974 *Coping and Adaptation*. Basic Books, New York

Compas B E 1987 Coping with stress during childhood and adolescence. *Psychological Bulletin* **101**: 393–403

Compas B E, Connor-Smith J K, Saltzman H, Thomsen A H, Wadsworth M E 2001 Coping with stress during childhood

and adolescence: Problems, progress and potential in theory and research. *Psychological Bulletin* **127**: 87–127
Cramer P 1998 Coping and defense mechanisms: What's the difference? *Journal of Personality* **66**: 895–918
Folkman S 1997 Introduction to the special section. Use of bereavement narratives to predict well-being in gay men whose partners died of AIDS: Four theoretical perspectives. *Journal of Personality and Social Psychology* **72**: 851–4
Freud A [1936] 1946 *The Ego and the Mechanisms of Defense*. International Universities Press, New York
Freud S [1894] 1962 The neuro-psychoses of defense. In: Strachey J (ed. and trans.) *The Standard Edition of the Complete Works of Sigmund Freud*. Hogarth Press, London, Vol. 3, pp.161–85
Garmezy N, Rutter M (eds.) 1983 *Stress, Coping and Development in Children*. McGraw-Hill, New York
Haan N 1977 *Coping and Defending: Processes of Self-environment Organization*. Academic Press, New York
Haggerty R, Sherrod L, Garmezy N, Rutter M 1994 *Stress, Risk, and Resilience in Children and Adolescents*. Cambridge University Press, New York
Hetherington E M, Blechman E A (eds.) 1996 *Stress, Coping, and Resiliency in Children and Families*. Erlbaum, Mahwah, NJ
Lazarus R S 1993 Coping theory and research: Past, present, and future. *Psychosomatic Medicine* **55**: 234–47
Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York
McCubbin H I, Thompson A I, McCubbin M A 1995 *Family Assessment: Resiliency, Coping, and Adaptation*. University of Wisconsin Publishers, Madison, WI
Moos R H, Billings A G 1982 Conceptualizing and measuring coping resources and coping processes. In: Goldberger L, Breznitz S (eds.) *Handbook of Stress: Theoretical and Clinical Aspects*. Free Press, New York, pp. 212–30
Murphy L B 1974 Coping, vulnerability, and resilience in childhood. In: Coelho G V, Hamburg D A, Adams J E (eds.) *Coping and Adaptation*. Basic Books, New York, pp. 47–68
Murphy L, Moriarity A 1976 *Vulnerability, Coping, and Growth: From Infancy to Adolescence*. Yale University Press, New Haven, CT
Parker J A, Endler N S 1996 Coping and defense: A historical overview. In: Zeidner M, Endler N S (eds.) *Handbook of Coping*. Wiley, New York, pp. 3–23
Pearlin L I, Schooler C 1978 The structure of coping. *Journal of Health and Social Behavior* **19**: 2–21
Sandler I N, Wolchik S A, Mackinnon D, Ayers T S, Roosa M W 1997 Developing lineages between theory and intervention in stress and coping processes. In: Wolchick S, Sandler I (eds.) *Handbook of Children's Coping: Linking Theory, Research and Intervention*. Plenum Press, New York
Skinner E A, Edge K (eds.) 1998 Special section on Coping and development across the life-span. *International Journal of Behavioral Development* **22**: 225–356
Snyder C R 1999 *Coping: The Psychology of What Works*. Oxford University Press, New York
Sommerfield M R, McCrae R R 2000 Stress and coping research. Methodological challenges, theoretical advances, and clinical applications. *American Psychologist* **55**: 620–25
Tennen H, Affleck G, Armeli S, Carney M A 2000 A daily process approach to coping: Linking theory, research, and practice. *American Psychologist* **55**: 626–36
Valliant G E 1986 *Empirical Studies of Ego Mechanisms of Defense*. American Psychiatric Association, Washington, DC
White R W 1974 Strategies of adaptation: An attempt at systematic description. In: Coelho G V, Hamburg D A, Adams J E (eds.) *Coping and Adaptation*. Basic Books, New York, pp. 47–68
Wolchik S A, Sandler I N (eds.) 1997 *Handbook of Children's Coping: Linking Theory, Research, and Intervention*. Plenum Press, New York

E. A. Skinner

Coping Assessment

Coping with stress has become a major subject of health and clinical psychology over the last decades. Coping assessment is not only an active area of scientific research, but has also found practical application in the fields of health and clinical psychology as well as in organizational and educational counseling. Its primary object is to discover the relative competence or deficit of patients and clients in mastering stressful situations. The omnipresence of stressful events in modern life gives particular importance to coping competence and its assessment.

1. What Is Assessed by Coping Assessment?

Coping with stress as an adaptive response refers to demands appraised as exceeding or taxing the subject's resources (Monat and Lazarus 1991). These demands can consist of microevents such as daily hassles, critical life events, i.e., divorce or loss of a loved person, and chronic strain such as work stress. The responses, termed 'coping with stress,' include different psychological processes and activities: attention processes such as vigilance or cognitive avoidance with respect to threatening stimuli; appraisal of the stress agent (stressor) concerning its significance for one's well-being; cognitive activities altering the meaning of stressful encounters and actions influencing the stress-loaded person–environment relationship.

Several coping responses classifications circulate in research and applications. The most influential distinction emphasizes two fundamental categories: the problem-focused and the emotion-focused coping responses (Lazarus and Launier 1978). Some researchers add a third category: the 'appraisal-focused coping.'

The types of questions which coping assessment should answer concern two major goals: (a) reliable and valid methods for observing, recording, and describing relevant features of coping processes; and (b) methods allowing the estimation of the probability of particular appraisal or coping modalities characterizing individual subjects (interindividual differences).

These probabilities may or may not depend on situation specific conditions.

Coping research, developed over the last decades, uses different strategies for accessing psychological characteristics of coping tendencies and processes. The methods differ according to (a) the information sources, (b) the types of measures, (c) the developmental stage of the subjects, and (d) the diagnostic techniques.

2. Data Sources

The first data source that can be distinguished is *self-report*. Typical, and especially retrospective self-report data, mirror the subjects' cognitive representation of their stress experience and coping behavior, and not how they really behaved. The assessment takes place outside the stressful situation, and memory plays a primordial role.

A second source is direct *systematic self-observation* inside the current stressful situation. This observation focuses on cognitive, emotional, and behavioral processes involved in concrete stress and coping episodes from daily life or experimentally arranged situations. The third data source is assessing coping behavior by systematic *external recording* of psychological features by trained observers. Such methods have been commonly used in the laboratory, for coping at the industrial workplace, and also applied to the family and school context. A fourth potential data source is the direct recording of *psychophysiological and biochemical parameters*.

3. Measurement Type in Relation to the Stressful Situation

A central aspect that differentiates the gamut of assessment procedures concerns the question: how does the method assess coping with respect to the stressful situation?

Some methods measure coping tendencies using *hypothetical situations* that describe potentially stressful scenarios. Participants are asked to imagine how they would probably act in well-described hypothetical situations. These procedures presuppose that the person is able to report reliably and validly his or her probable responses to hypothetical stressful situations. The 'Mainz Coping Inventory' (ABI) of Krohne (1993) measures separately 'vigilance' and 'cognitive avoidance' with respect to ego-threatening and physical-threatening hypothetical situations. ABI contains, for both situation types, four hypothetical scenarios that are followed by ten binary scale items describing possible coping actions. The reliability scores are satisfactory. Miller's Behavioral Style Scale (MBSS; Miller 1987) also measures, on the base of two hypothetical physical and two ego-threat situations, the coping modalities 'monitoring' and 'blunting' that are close to the vigilance–avoidance concept.

Other questionnaires focus on coping with more *complex stressful domains* (including chronic strain), for example the family (Buehler 1990), workplace (Semmer et al. 1996), or critical life events. Such assessments try to answer how people cope with a rather complex and temporally enduring situation, for example an illness or a difficult transition in family development (e.g., the birth of the first child). Most instruments use self-report measures and some self-observation or external observation (Semmer et al. 1996). For example, the Family Crisis Oriented Personal Evaluation Scales (F-COPES) from McCubbin et al. (1987) were created to assess problem-solving and behavioral strategies of families in problematic situations. Thirty coping items focus on features of intra and extrafamilial coping activities. Parent and adolescent norms have been elaborated on the basis of large samples for the five subscales and the total scale.

Various assessment approaches refer the assessed coping to real *experienced stressful situations* in the recent past or to actual ongoing stress encounters. The Ways of Coping Questionnaire (WCQ), the successor of the Ways of Coping Checklist (WCC) from Folkman and Lazarus (1988) can be considered as a variant of such measures. WCQ contains eight subscales measuring coping strategies such as 'Confronting Coping,' 'Positive Reappraisal,' or 'Seeking Social Support.' Participants have to respond to 50 items on a four-point Likert scale on how they coped with a recently experienced stressful situation.

Approaches to assessing ongoing daily coping, record coping activities by daily checklists, or by the experience sampling method (ESM). Here participants record, using a time sampling schedule, at least six times a day for several days, their emotional states, coping activities, and other relevant information.

A fourth group of instruments does not specify the type of stressful situation. These *situation-unspecific* methods ask with what coping modalities participants respond if they encounter a stressful or upsetting situation. For measuring situation-unspecific coping tendencies, Endler and Parker (1990) devised the Coping Inventory for Stressful Situations (CISS), which focus on 'task-oriented,' 'emotion-oriented,' and 'avoidance-oriented' coping in stressful situations. Coping tendencies, for stressful situations in general, are also measured by the questionnaire Stressverarbeitungsfragebogen (SVF) from Janke et al. (1997), which assess 20 different types of coping modalities. Both procedures inform on stable interindividual differences and have good psychometric qualities.

The last group of situations, to which coping assessment has been related, are *laboratory situations* with real (or video-recorded) stressors. The laboratory setting is destined for research purposes. The validations of some questionnaires has profited from such laboratory assessments (e.g., Miller1987).

Table 1
Data resources and situation types characterizing assessment approaches

Data source	Situation type				
	Hypothetical situations	Complex stressful domains and macro events	Concrete life situations in the natural setting	Situation unspecified	Laboratory situation
Self-report	Krohne (1993), Miller (1987), Reicherts and Perrez (1993)	McCubbin et al. (1987)	Folkman and Lazarus (1988)	Endler and Parker (1990), Janke et al. (1997), Brandtstädter and Renner (1990)	Lazarus et al. (1962)
Self-observation (actual)		Cartwright et al. (1995)	Perrez et al. (2000)		
External observation of psychological features		Semmer et al. (1996)	Fabes and Eisenberg (1992)		Miller (1987), Krohne (1993)
External observation of psychophysiological features			Fahrenberg (1996)		Lazarus et al. (1962)

Table 1 summarizes the different data sources and the different types of situations that are used in coping assessment, illustrated with selected examples of instruments.

4. *Coping Assessment Across Life Span*

The type of technical procedure, item format, and content in questionnaires depends not only on the theoretical background and the type of stressors studied, but also on the age of the population which is the focus of the instrument. If it is evident that the same technical procedures can not be applied for illiterate preschool children as for adults, it is less clear, and indeed controversial in the literature, if the quality of coping modes changes across life span, or if at some life phases different frequencies predominate as a function of development-specific needs.

For preschool children, the only means available of assessing coping behavior are systematic observation of the overt behavior in the natural family setting or care institutions, and the use of rating scales by parents who evaluate the coping behavior of their children. Fabes and Eisenberg (1992) describe the modalities to observe systematically the coping behavior of 3.5- to 6-year-old children in anger-eliciting situations at full-day care institutions with categories such as 'active resistance attempts,' 'adult-seeking attempts,' 'revenge-attempts,' and others. The Coping Inventory of Zeitlin (1985) assesses measures for self- or environment-oriented coping tendencies by 3 to 15 year old children and adolescents with respect to the features 'productivity,' 'activity,' and 'flexibility.' 'Coping with Self' concerns the behaviors a child uses to meet personal needs, and 'Coping with Environment' the behaviors to adapt to the demands of the world. Zeitlin considers these properties as relevant for adaptive behavior independent from situation and age. Parents and teachers serve as data sources. An adapted version can be used as self-rating by adolescents and adults.

Most instruments were devised for adults. For the elderly, only few instruments have been developed and evaluated. One of them is the questionnaire of Brandtstädter and Renner (1990) to assess preferences for assimilative and accommodative modes of coping. Assimilation and accommodation processes are supposed to be of first importance to reduce the discrepancy between aspirations and achievements, which are progressively imbalanced in later adulthood. The two statistically independent 15-item scales assess the disposition for tenacious goal pursuit (TGP-scale) and for flexible goal adjustment (FGA-scale). The psychometric qualities of the scales are satisfactory. The above-described instruments illustrate the dependency of the measures (with respect to their content and their technical format) on theoretical assumptions and the developmental stage of the subjects chosen.

5. Technical Procedures

Most assessment procedures work on the base of self-report questionnaires. For an overview see Schwarzer and Schwarzer (1996). They mainly assess coping features and tendencies using multi-item scales. By measuring a coping modality using a set of items loading on the same coping factor, the reliability of the measure is improved. For constructing such scales, classical psychometric methodology offers a rich range of procedures and parameters, such as reliability, validity, and measurement quality evaluation.

Only a few methods observe appraisal and coping behavior within the situation by systematic self-observation. One example is the Family Self-Monitoring-System (FASEM-C), which runs on a palmtop computer. FASEM-C records information on current emotional and somatic states, on antecedents of (stress-) emotions such as the setting, events and appraisal, individual and social adaptive, or maladaptive coping responses (what subjects actually did), and on their short-term effects. All family members (or couples) have to record simultaneously, prompted by an acoustic signal based on a time sampling schedule, six times a day for a weekly period, their own experience and behavior, and the coping behavior of other family members present. Coefficients for reliability, reactivity, and validity are satisfactory (Perrez et al. 2000).

6. What Questions Can be Answered by Coping Assessment?

The types of questions coping assessment can answer depend on the accuracy with which it can observe the underlying psychological phenomena. Most present assessment methods have focused on the cognitive representation that participants have of their appraisal and coping tendencies with respect to specific situations or to stressful situations in general.

Questionnaires with normative data from reference populations allows reliable information on *inter-individual differences*, e.g., CISS or SVF; others permit information on *within-subject descriptions* of appraisal and coping modalities (intraindividual measures), sometimes with respect to one single episode (e.g., WCQ). Questionnaires specially for intra and inter-individual measures have been developed for different age groups (Aldwin 1994).

A third type of information concerns *process-related measures*: process measures assess features of the appraisal-coping processes that can be observed in concrete processes.

The development of such instruments is still in its infancy. The Stress and Coping Process Questionnaire (SCPQ) models the perception/appraisal-stress emotion and coping sequence, starting with a hypothetical stressful situation that the participant has to imagine (Reicherts and Perrez 1993). The subject gives relevant scaled information about his of her probable cognitive, emotional, and behavioral responses. Then the situation is presented with an evolved second and third stage asking similar information. The questionnaire, with satisfactory psychometric qualities, allows the computing of scores on individual differences for process-relevant features. It has the disadvantage of being somewhat time-consuming for the participants. Alternative approaches to access process-relevant features are computer aided self-monitoring procedures.

7. Perspectives on Future Research

As far as the theoretical frameworks deal more with cognitive and behavioral processes and activities than with their subjective representation, it can be expected that the future development will give more emphasis to information on within-situation recorded cognitive activities, emotions, and behaviors rather than on information depending on post-situation reports and the participant's own cognitive representation. The actual method-theory discrepancy is due to parsimonious reasons that have been in favor of representation-oriented questionnaires. Modern miniaturized computer technology now opens rich possibilities, including the concurrent assessment of psychological and physiological measures under real-life conditions (Fahrenberg 1996), but also the simultaneous recording of appraisal and coping behavior of dyads and small groups such as the family. One interesting development, which goes beyond the pure assessment function, is the development of ambulatory expert systems. These systems can aid the user in improving his or her coping behavior.

See also: Behavioral Assessment; Chronic Illness, Psychosocial Coping with; Coping across the Life-span; Coronary Heart Disease (CHD), Coping with; Disasters, Coping with; Job Stress, Coping with; Psychometrics; Psychophysiology; Stress and Coping Theories; Stressful Medical Procedures, Coping with

Bibliography

Aldwin C M 1994 *Stress, Coping, and Development*. Guilford, New York

Brandtstädter J, Renner G 1990 Tenacious goal pursuit and flexible goal adjustment: Explication and age-related analysis of assimilative and accommodative strategies of coping. *Psychology and Aging* **5**: 58–67

Buehler C 1990 Adjustment. In: Touliatos J, Perlmutter B F, Straus M A (eds.) *Handbook of Family Measurement Techniques*. Sage, Newbury Park, CA, pp. 393–574

Cartwright S, Cooper C L, Murphy L R 1995 Diagnosis a healthy organization: A proactive approach to stress in the workplace. In: Murphy L R, Hurrell J (eds.) *Job Stress Interventions: Current Practice and Future Directions*. American Psychological Association, Washington, DC, pp. 217–33

Endler N S, Parker J D A 1990 *Coping Inventory for Stressful Situations (CISS): Manual*. Multi Health Systems, Toronto, Canada

Fabes R A, Eisenberg N 1992 Young children's coping with interpersonal anger. *Child Development* **63**: 116–28

Fahrenberg J 1996 Concurrent assessment of blood pressure, physical activity, and emotional state in natural settings. In: Fahrenberg J, Myrtek M (eds.) *Ambulatory Assessment*. Hogrefe and Huber, Seattle, WA, pp. 165–87

Folkman S, Lazarus R S 1988 *Manual for the Ways of Coping Questionnaire*. Consulting Psychologist Press, Palo Alto, CA

Janke W, Erdmann G, Ising M (eds.) 1997 *Stressverarbeitungsfragebogen (SVF 120)* [Coping with stress questionnaire]. Hogrefe, Göttingen, Germany

Krohne H W 1993 Vigilance and cognitive avoidance as concepts in coping research. In: Krohne H W (ed.) *Attention and Avoidance: Strategies in Coping with Aversiveness*. Hogrefe and Huber, Seattle, WA, pp. 19–50

Lazarus R S, Speisman J C, Mordkoff A M, Davison L A 1962 A laboratory study of psychological stress produced by a motion picture film. *Psychological Monographs: General and Applied* **76** (34, Whole No. 553)

Lazarus R S, Launier R 1978 Stress-related transactions between person and environment. In: Pervin L A, Lewis M (eds.) *Perspective in Interactional Psychology*. Plenum, New York, pp. 287–327

McCubbin H I, Olson D H, Larsen A S 1987 F-COPS. Family Crises Oriented Personal Evaluation Scales. In: McCubbin H I, Thompson A I (eds.) *Family Assessment Inventories for Research and Practice*. The University of Wisconsin-Madison, Madison, WI, pp. 194–207

Miller S M 1987 Monitoring and blunting: Validation of a questionnaire to assess styles of information seeking under threat. *Journal of Personality and Social Psychology* **52**(2): 345–53

Monat A, Lazarus R S 1991 Stress and coping—Some current issues and controversies. In: Monat A, Lazarus R S (eds.) *Stress and Coping. An Anthology*. Columbia University Press, New York, 3rd edn, pp. 1–15

Perrez M, Schöbi D, Wilhelm P 2000 How to assess social regulation of stress and emotions in daily family life? A computer assisted family self-monitoring-system (FASEM-C). *Journal of Clinical Psychology and Psychotherapy* **4**: 326–39

Reicherts M, Perrez M 1993 *Fragebogen zum Umgang mit Belastungen im Verlauf (UBV)*. Hans Huber, Bern (English translation in Perrez M, Reicherts M 1992 *Stress, coping, and Health*. Hogrefe and Huber, Seattle, WA)

Schwarzer R, Schwarzer C 1996 A critical survey of coping instruments. Coping and defense: A historical overview. In: Zeidner M, Endler N S (eds.) *Handbook of Coping*. Wiley, New York, pp. 107–32

Semmer N, Zapf D, Greif S 1996 'Shared job strain.' A new approach for assessing the validity of job stress measurements. *Journal of Occupational and Organizational Psychology* **69**: 293–310

Zeitlin S 1985 *Coping Inventory: A Measure of Adaptive Behavior*. Scholastic Testing Service, Bensenville, IL

M. Perrez

Coronary Heart Disease (CHD), Coping with

1. Introduction

Coronary heart disease (CHD) is a chronic illness with acute manifestations. Two clinical aspects are angina pectoris (pain) and myocardial infarction (heart attack). CHD refers to illnesses caused by the narrowing of the coronary arteries. Coronary arteries supply the heart with oxygen and when narrowing occurs, the flow of oxygen is reduced. A mild shortage of oxygen causes pain in the chest and arm, while a severe shortage may result in a heart attack. CHD accounts for about 40 percent of all deaths in North America. Both men and women are prone to CHD, although the age of onset for women is about 15 years later than it is for men. Many premature deaths occur as a result of CHD.

2. Antecedents of CHD

2.1 Genetic and Biochemical Factors

CHD is more common among individuals who have a parental history of the disease. This is partially due to genetic predisposition and to family lifestyles that may include risk factors for CHD. Additionally, personality factors related to CHD (e.g., hostility) may be heritable. As one ages, environmental factors (i.e., lifestyle choices) play a more important role than genetics (Rice 1997). Furthermore, family history is an important unmodifiable factor in determining risk for CHD, as are age and gender (Rice 1997). Other risk factors, like lifestyle choices, can be modified.

2.2 Poor Lifestyles

Many behavioral health risk factors have been linked to the development of heart disease, including smoking, alcohol consumption, high serum cholesterol level, hypertension (high blood pressure), obesity, and a sedentary lifestyle. Review studies have shown that risk for CHD is doubled for inactive compared to active persons (Leon 1997). The incidence of heart attacks is directly related to the average number of cigarettes smoked, and smoking affects other risk factors: it results in increased cholesterol levels, increased heart rate, and increased blood pressure (see *Health Behaviors*).

2.3 Stress

Selye (1976; as cited in Endler 1997) proposed his 'General Adaptation Syndrome' in which stress was

defined as the body's nonspecific response to any demands placed on it. Furthermore, Selye distinguished between good stress (eustress) and bad stress (distress).

There are both physical stressors, which include environmental factors, and psychological stressors. Situations that are stressful induce a physiological arousal that mobilizes a fight or flight reaction. This can be adaptive; however, under continued stress, this arousal can be harmful (see *Stress and Health Research*).

Heart rate accelerates when a person is under stress. Exposure to chronic stress has been linked to the development of CHD, and acute stressors have been found to precipitate heart attacks. The amount of stress one experiences at work (occupational stress) has been found to be related to the incidence of CHD. Job factors such as job strain, low job security, and high work pressure are related to increased risk for CHD (Taylor 1999).

2.4 Personality Factors

In the 1970s and 1980s, researchers thought that a particular behavioral style, the Type A behavior pattern, was related to the development of CHD. The Type A behavior pattern consists of a hostility component, a competitiveness component, and a need for everything to be done in a hurry (Friedman 1996). More recently, other researchers have concluded that it is the hostility component (a cynical form) of Type A behavior that is most strongly implicated in the development and progression of CHD (Taylor 1999). This cynical hostility consists of anger, suspiciousness, distrust towards others, and resentment. People who are hostile and have a defensive personality style have been shown to be at a particularly high risk for CHD. Other research has shown that it is the expression of hostile emotions that is related to CHD, not merely being in a hostile state. This research has been predominantly carried out in men, and the relationship between hostility and CHD in women has not yet been clearly established (see *Coronary-prone Behavior, Type A*).

Hostile people manifest more physiological reactivity in response to stress. Stress causes both an increased heart rate and blood vessel shrinkage in some individuals, causing the heart to transfer more blood through narrower blood vessels. Other mechanisms of action include sympathetic nervous system activation resulting in the transfer of lipids into the bloodstream and hormonal based reactions. Alternatively, hostility may not be related to CHD through cardiovascular reactivity, but rather through the relationship between hostility and poor lifestyle choices (i.e., smoking, consumption of alcohol, higher weight, higher lipid level, and hypertension).

Strong relations have also been found between CHD and negative emotional states other than hostility (i.e., anxiety, depression, and aggression). Additionally, optimism has been found to play a positive role. Individuals who are optimistic recover faster following surgery, have better adherence to the treatment regimen, and have better health outcomes (Irvine and Ritvo 1998).

3. Consequences of CHD

3.1 Physical and Psychological Changes

CHD may result in mortality or disability leading to a restriction of activities. Individuals with CHD who have experienced a heart attack may be restricted in social and leisure activities and may be forced to quit working. These physical, social, and occupational changes may affect an individual's self-concept, self-worth, and self-efficacy. There also may be an affect on interpersonal relationships.

4. Coping

4.1 Coping: General Styles vs. Specific Strategies

Coping mediates between antecedent stressful events and distressing consequences. A coping style is a typical manner of confronting a stressful situation and dealing with it. There are three basic coping styles: task-oriented, emotion-oriented, and avoidance-oriented (Endler 1997). Task-oriented coping consists of efforts aimed at solving the problem. Emotion-oriented coping involves emotional reactions; and avoidance-oriented coping involves activities and cognitions aimed at avoiding the stressful situation and can be of a distraction or social diversion nature. One can also conceptualize and investigate coping in terms of how individuals respond to specific stressful situations, such as coping with an illness (see *Stressful Medical Procedures, Coping with*).

4.2 Coping with Illness

There are four strategies that individuals use for coping with illness: (a) distraction coping involves attempts to think about anything other than the health problem or to engage in unrelated activities; (b) palliative coping involves strategies aimed at alleviating the unpleasantness of the illness; (c) instrumental coping involves a task-oriented approach such as seeking medical advice; and (d) emotional preoccupation coping consists of affective strategies such as ruminating about the health problem (Endler and Parker 2000).

5. *Factors Influencing Coping*

5.1 *Personal and Situational Factors*

The individual's personality in a stressful situation affects how one copes with the event. Personality characteristics such as optimism, perceived control, and self-esteem, in part, determine how one copes with stressful situations. There are particular types of situations for which specific coping behaviors are more likely to be effective than others. External factors such as socioeconomic status, time, and social support also affect coping methods. Thus, persons, situations, and their interactions each affect coping styles.

5.2 *Social Support Systems as Resources*

Social support, a resource for coping, is important for adjustment. People seek advice from others, emotional support from others, and others provide opportunities for distracting activities. People with high levels of social support from family, friends, and/or the community may experience less stress and cope more effectively than people without strong support networks. Recovery from cardiac surgery is facilitated by social support (Schröder et al. 1997). In addition to social support, individuals with CHD should seek medical and possibly psychological support (see *Social Support and Recovery from Disease and Medical Procedures*).

6. *Rehabilitation and Prevention*

6.1 *Effective Coping Styles*

Avoidance or distraction coping has been shown to be effective in the early stages of CHD, especially during the first few days following a heart attack. At later stages of CHD, avoidance-oriented coping strategies are not as effective. At this point, an instrumental or task-oriented coping approach is more adaptive (Zeidner and Endler 1996). Furthermore, individuals coping with a chronic illness who use a variety of coping strategies may cope better than those who use one predominant coping style (Taylor 1999).

6.2 *Changes in CHD Risk Behavior*

There are a number of lifestyle changes that result in decreased risk of CHD. These include: (a) changing to a low-fat diet, (b) quitting smoking and decreasing excessive drinking, (c) reducing stress, (d) exercising, and (e) psychological counseling.

The three primary CHD risk factors are hypertension, elevated cholesterol levels, and smoking (Rice 1997). Furthermore, the presence of these factors is of a synergistic nature. Fortunately, all three of these risk factors can be modified. Changes in diet and exercise have been found to reduce blood pressure and decrease cholesterol levels. Exercise tends to decrease heart rate, reduce blood pressure, and improve the overall efficiency of the heart (Leon 1997). Exercise also typically results in weight reduction, which consequently results in the reduction of cholesterol. To reduce blood pressure, the reduction of both salt intake and alcohol consumption are also indicated. The cessation of smoking, which may not be easy, also reduces the risk of CHD substantially (see *Smoking and Health*).

Cognitive-behavioral stress reduction programs have been found to modify personality factors which contribute to CHD (Johnston 1997). Stress and pain management techniques have been shown to be effective in reducing angina and may be part of a rehabilitation program. Rehabilitation programs are directed at both altering risk factors for and reducing symptoms of CHD. These programs include medical evaluation, prescribed exercise, education regarding risk factors, and counseling (Leon 1997). Furthermore, medication such as beta-blocking agents, ACE inhibitors, calcium channel blockers, or diuretics may be prescribed to control blood pressure and for chest pain.

7. *A Systems Model for Reducing Prevalence of CHD*

7.1 *Stress, Anxiety, and Coping Interaction Model*

The multidimensional interaction model of stress, anxiety, and coping is process-oriented and consists of person variables (e.g., heredity, cognitive style) which may interact with one another and with situation variables (e.g., illness, hassles). The interaction of person and situation variables leads to a perception of threat and consequent changes in state anxiety. Changes in state anxiety lead to a modification of physiological reactions and coping responses (Endler 1997).

Stress, anxiety, and coping with illness all involve complex processes that interact with one another. Therefore, it is important to consider all aspects of this process when facing a stressor, especially one as significant as CHD. Person variables interact with situation variables in determining the coping strategies used to face the stressor and affect the physiological arousal and illness experienced.

8. *Conclusions*

There are several factors that one cannot control in terms of risk for CHD (e.g., heredity) but many factors can be controlled, either in terms of prevention or for rehabilitation. The cessation of smoking, engaging in physical activity, and following an appropriate diet all

decrease the risk for CHD. Furthermore, avoidance-oriented coping strategies appear to be more adaptive early on in the illness, while task-oriented coping strategies appear to be most adaptive later on. Psychological counseling may be effective in reducing personality factors, such as hostility, which are related to increased CHD risk.

See also: Cardiovascular Conditioning: Neural Substrates; Coping across the Lifespan; Coronary Heart Disease (CHD): Psychosocial Aspects; Coronary-prone Behavior, Type A; Gender and Cardiovascular Health; Rehabilitation (Cardiac): Psychosocial Aspects; Stress and Cardiac Response; Stress and Coping Theories; Stressful Medical Procedures, Coping with

Bibliography

Endler N S 1997 Stress, anxiety and coping: The multidimensional interaction model. *Canadian Psychology* **38**: 136–53

Endler N S, Parker J D A 2000 *Coping with Health Injuries and Problems (CHIP): Manual*. Multi-Health Systems, Toronto, ON, Canada

Friedman M 1996 *Type A Behavior: Its Diagnosis and Treatment*. Plenum, New York

Irvine J, Ritvo P 1998 Health risk behaviour change and adaptation in cardiac patients. *Clinical Psychology & Psychotherapy* **5**: 86–101

Johnston D W 1997 Cardiovascular disease. In: Clark D M, Fairburn C G (eds.) *Science and Practice of Cognitive Behaviour Therapy*. Oxford University Press, NY, pp. 341–58

Leon A S (ed.) 1997 *Physical Activity and Cardiovascular Health*. Human Kinetics, Champaign, IL

Rice P L 1997 *Health Psychology*. Brooks/Cole Publishing, Pacific Grove, CA

Schröder K E E, Schwarzer R, Endler N S 1997 Predicting cardiac patients' quality of life from the characteristics of their spouses. *Journal of Health Psychology* **2**: 231–44

Taylor S E 1999 *Health Psychology*, 4th edn. McGraw-Hill, Boston

Zeidner M, Endler N S (eds.) 1996 *Handbook of Coping*. Wiley, New York

N. S. Endler

Coronary Heart Disease (CHD): Psychosocial Aspects

1. Introduction

A comprehensive review of the epidemiology and prevention of cardiovascular disease is given in Labarthe (1998). In industrialised countries, the commonest cause of death is coronary heart disease (CHD). This article will thus focus on CHD and its associated risk factors. Consideration of all of these risk factors is relevant, as virtually all are strongly influenced by psychosocial processes. Large differences in disease rates between countries, evidence that risk of disease changes on migration, large differences within countries according to socioeconomic position and area of residence, and relatively rapid changes over time, suggest that the disease is preventable.

CHD has been defined as 'myocardial impairment due to an imbalance between coronary blood flow and myocardial requirements caused by changes in the coronary circulation.' Atherosclerosis is an important underlying pathological process. The process of atheroma deposition in coronary arteries, however, cannot be observed directly in life without the use of invasive clinical procedures. Even so, the study of symptomatic disease may actually be more appropriate. The clinical disease is, after all, what is experienced and may represent the culmination of a number of pathological processes.

Large differences in disease rates between countries, evidence that risk of disease changes on migration, large differences within countries according to socioeconomic position and area of residence and relatively rapid changes over time, suggest that the disease is preventable. To illustrate this international variation the rates of death attributed to CHD in men and women from the populations participating in the World Health Organisation MONICA (monitoring trends and determinants in cardiovascular disease) study are shown in Fig. 1.

CHD was until recently viewed largely as a twentieth century epidemic. In retrospect it seems clear that CHD occurred before this. William Heberden described the typical symptoms of angina pectoris in 1768. Later in the eighteenth century Dr Samuel Black of Newry, County Down, described many cases of angina pectoris and produced a list of factors associated with increased and decreased susceptibility to angina, which are similar to those currently considered plausible (Evans 1995). While angina pectoris occasionally was recorded as a cause of death, less than 2 percent of all deaths attributed to heart disease in Britain fell into this category even by the early part of the twentieth century.

Atherosclerosis has been demonstrated in Egyptian mummies, but the first contemporary descriptions occurred in the sixteenth century. As CHD is associated with advanced coronary atherosclerosis, the increase in CHD mortality in the first half of the twentieth century would be expected to be accompanied by increases in severe atheroma. However, studies of postmortem records of the London Hospital suggested that there was no such increase. Later data comparing the coronary arteries at postmortem of US soldiers who died in the Korean and Vietnam War reported more atheroma in the early 1950s than the

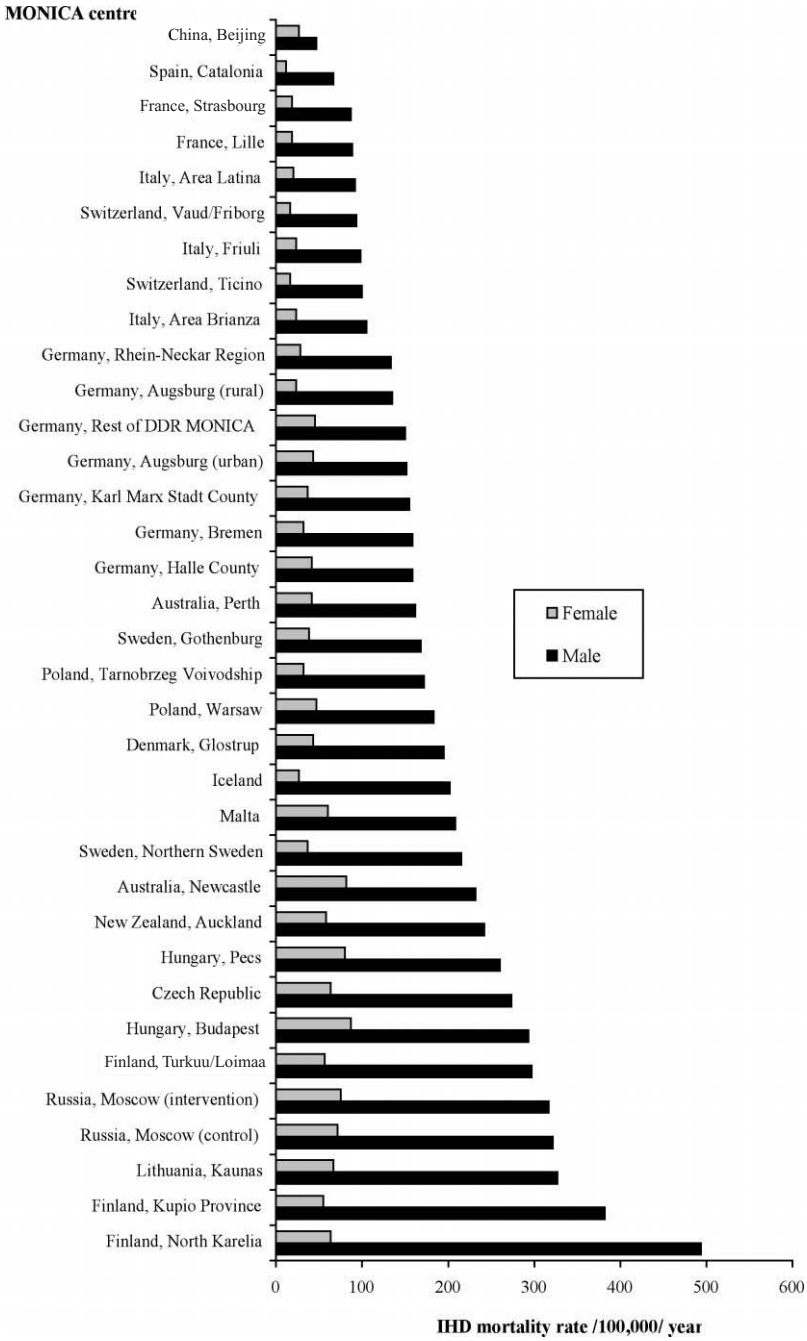

Figure 1
Male and female IHD (ischemic heart disease) mortality rates in men and women aged 35–64 in 35 MONICA populations in the 1980s Source: The World Health Organisation MONICA project. Ecological analysis of the association between mortality and major risk factors of cardiovascular disease. *International Journal of Epidemiology* 1994; **23**: 505–516

1960s, while CHD mortality in the USA was high and stable over this period. The relative contribution of atherosclerosis and other pathological processes, such as thrombotic tendency, to trends over time and differences between countries in CHD mortality is unclear.

2. International Trends in CHD

Nonsocialist developed countries have generally shown a similar pattern of rise and then fall in CHD mortality. Some, like the United States, experienced an early and rapid fall, whereas others, such as the UK, experienced a relatively late onset decline. Rates in Japan and the Mediterranean countries of Europe are low. These international time trends for men are illustrated in Fig. 2. The MONICA (Monitoring Trends and Determinants in Cardiovascular Disease) study has shown that the recent trends in mortality are attributable to both changes in event rates and changes in case fatality. Changes in case fatality suggest that medical interventions are reducing mortality in groups with existing CHD.

3. Migrant Studies

Migrant studies provide valuable insights into the heritability of disease and the stage of life at which risk is determined. Studies of Japanese migrants to Hawaii and mainland America suggest that as they adopt Western lifestyles their CHD risk increases. More puzzling is the fact that migrants to Australia from Greek islands, despite adopting many of the detrimental habits of their host country, are at lower risk of CHD. It may be that Greek Australians adopt healthier lifestyles in other respects that counterbalance their increased fat consumption and smoking prevalence (Powles 1994).

4. Overview of Risk Factors for CHD

There are a number of personal attributes, often described as risk factors, associated with the development of CHD. Some of these—like age—are fixed. Others, like smoking, are modifiable. And some, like serum cholesterol, though modifiable, are a product of the interplay between an individual's genetic make-up and environmental exposures. A list of some of the

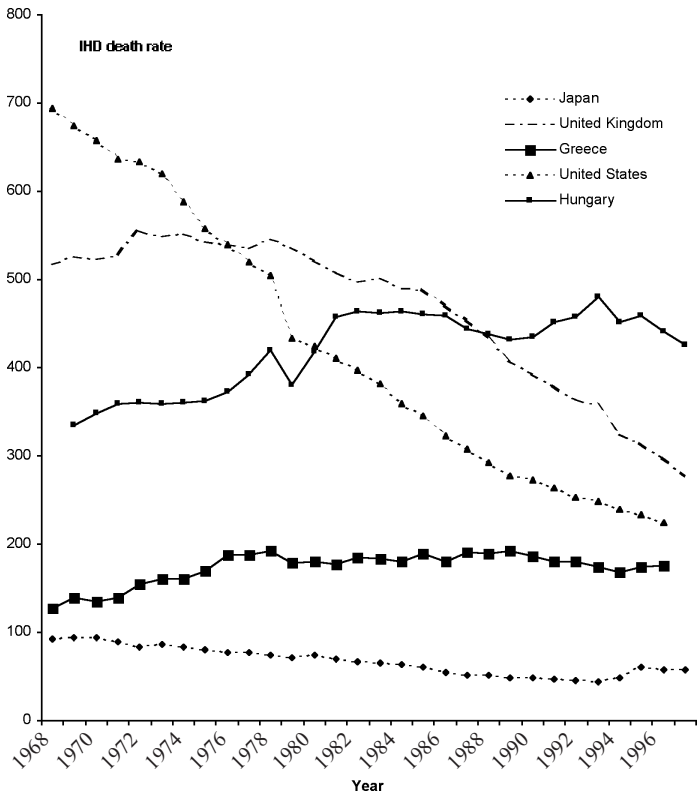

Figure 2
Age standardized death rates attributed to ischemic heart disease from 1968–1997 in men aged 35–74 in selected countries. Source: Peterson S, Mockford C, Rayner M 1999

attributes associated with CHD is set out in Table 1. This list is not dissimilar to the list proposed by Samuel Black 200 years earlier, although Black considered that more affluent people were at greater risk of CHD.

Stamler suggests that the major modifiable risk factors are smoking, high blood pressure, high serum cholesterol, and diet. These major risk factors are powerful predictors of CHD. In a large study of middle-aged US men, smokers in the highest quintiles for serum total cholesterol and systolic blood pressure were around 20 times more likely to die from CHD over the next 12 years as nonsmokers in the lowest quintiles of serum cholesterol and blood pressure (Neaton and Wentworth 1992). Though they are powerful individual predictors of risk these factors only accounted for 23 percent of the variance in CHD incidence in men and 14 percent of the variance in women observed between populations in the WHO MONICA study.

4.1 Age

Age is a strong risk indicator for CHD incidence and mortality. Compared to men aged 40, 50-year old men have five times the risk, 60-year old men have 15 times the risk, and 70-year old men have over 40 times the risk of dying from CHD (Labarthe 1998). A similar steep gradient with age is seen for women. Risk continues to increase into older age groups. Recent evidence suggests that the elderly benefit from risk factor control. This combined with their higher absolute level of risk means that sizeable absolute reductions in CHD risk can be achieved in older people and where therapies may be particularly cost effective.

4.2 Gender

Rates of death attributed to CHD in men are consistently three to four times higher than those in women. There are, however, marked variations in the rates of death between countries, such that women in Northern Ireland or Scotland have CHD death rates that are four times higher than those experienced by Japanese men. The rates of death attributed to CHD in women correlate closely with those in men, suggesting that common environmental factors explain these international differences.

Despite the fact that women have been under represented in epidemiological studies of CHD, it nevertheless appears that risk factors such as serum total cholesterol, blood pressure, and smoking perform similarly in women. But as the disease is commoner in men, the differences in absolute risk are much greater for men.

Men are more likely to smoke than women are, but differences in smoking and other coronary risk factors do not fully explain the observed excess of CHD seen in men. The lower risk of coronary heart disease among women has understandably led to studies of sex hormones. There is at present, however, little evidence to support a decisive role for sex hormones. Many observational studies (of women who had elected to take hormone replacement therapy (HRT)) appeared to support this contention, but the findings of randomized controlled trials (where women were allocated at random to HRT or placebo), while still preliminary, suggest no such protection is given by HRT (Barrett-Connor 2001).

4.3 Family History and Genetic Factors

A family history of CHD in first-degree relatives is associated with an increased risk of CHD (Labarthe 1998). Currently identified major single-gene conditions, for example, familial hypercholesterolemia, are uncommon, so while they are associated with a large increase in risk for an individual, they contribute little to population CHD rates. Other polymorphisms with lower relative risks but greater prevalence are under investigation. Individuals with a family history of premature CHD are at higher risk and thus may have more to gain from risk factor control.

4.4 Early-life Influences

The majority of epidemiological research on CHD has focussed on risk factors acting in adulthood. More recently, it has been appreciated that environmental factors acting *in utero* or in childhood may be important. A series of studies has demonstrated that birth weight is inversely related to risk of CHD, suggesting that suboptimal intrauterine environments result in offspring who, many years later, are more likely to succumb to CHD (Barker 1998). Furthermore, several CHD risk factors have been shown to be related to birth weight. However, taking these risk factors into consideration does not appear to fully account for the influence of fetal development on CHD risk, which seems to involve other, as yet unspecified, pathways. It is also possible that genetic factors could underlie both low birth weight and increased disease risk in adulthood. Whether modifying fetal development will modify later CHD risk is a matter which requires further investigation.

Exposures acting in infancy, childhood, and adolescence may also be of importance. A frequently replicated and long-standing observation is that taller individuals are at reduced CHD risk. Recently, it has been shown that stature in childhood is inversely related to later CHD risk, indicating that the association between height and CHD mortality is not due to differential shrinkage occurring in adulthood amongst individuals most prone to CHD (Gunnell et al. 1998). Growth occurs during infancy and childhood, being influenced by nutrition and infections, as well as genetic factors. Childhood growth may also be

Table 1
Personal attributes associated with ischemic heart disease

Fixed factors	Exposure	Composite markers of risk
Age	Smoking	Serum lipids
Sex	Physical inactivity	Blood pressure
Family history	Diet	Lung function
Socioeconomic-position	Psychosocial stress	Hemostatic factors
Genotype	*In utero* experience	Obesity
	Infections	Insulin resistance
	Climatic factors	Hormonal status
		Homocysteine

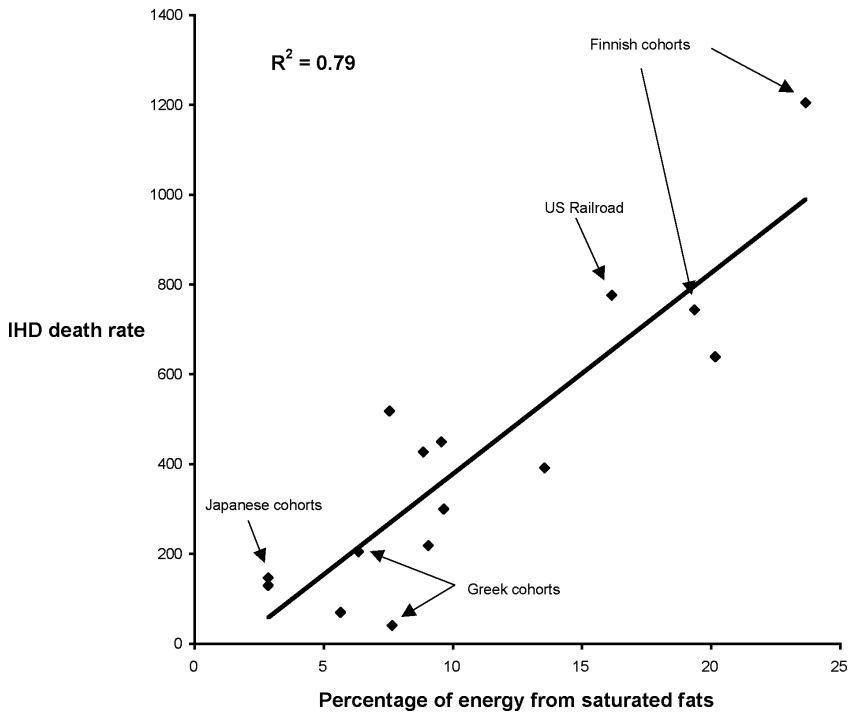

Figure 3
Ecological comparison between percentage of calories from saturated fat consumed at baseline and IHD mortality over fifteen years in 15 cohorts from seven countries recruited between 1958 and 1962 (Source: Keys A, Menotti A, Karvonen M J, et al. The diet and 15-year death rate in the seven countries study. *American J Epidemiol* 1986; **124**: 903–915)

influenced by stress. Of the infections acquired in childhood, *Helicobacter pylori* infection has been most widely investigated, but the overall evidence does not strongly point in the direction of a direct causal association between this infection and CHD risk.

4.5 Diet

Various dietary characteristics have been associated with risk of CHD. These include alcohol, dietary fat, and dietary antioxidants such as ß-carotene.

Moderate consumption of alcoholic drinks appears to protect against CHD, with higher rates of CHD in teetotallers and heavy drinkers. Ecological comparisons suggest that wine may exert additional protection but cohort studies appear to show that the type of beverage is unimportant. People who indulge in sporadic—or binge—drinking appear to be at increased (rather than reduced) risk of CHD.

The Seven Countries Study compared the diets of individuals from populations with different rates of CHD. It showed a strong ecological association between CHD mortality and saturated fat consump-

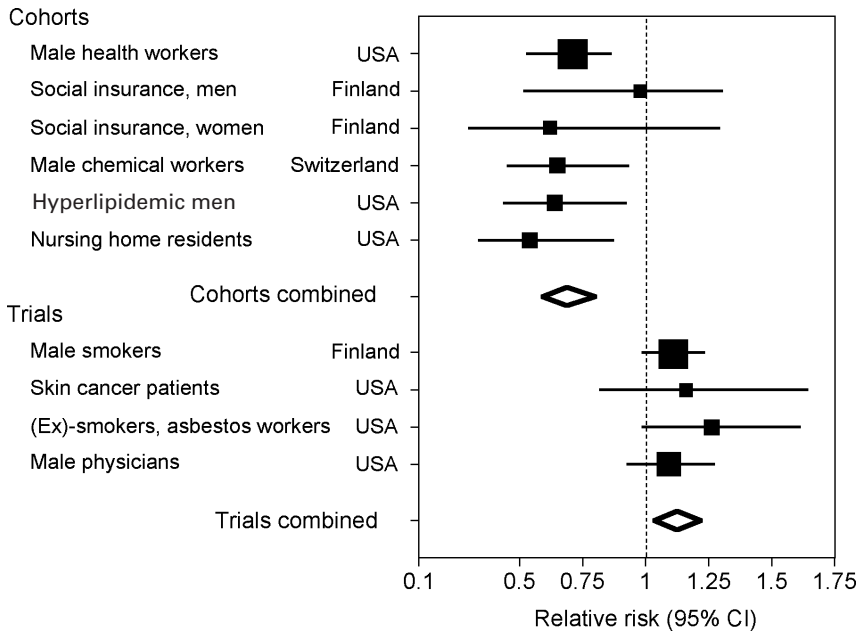

Figure 4
Meta-analysis of results of observational cohort studies of beta-carotene intake and cardiovascular mortality and of randomised controlled trials of the same issue

tion. The results of this study are shown in Fig. 3. The relationship between fats and CHD may be complex with different fatty acids increasing, decreasing or having no effect on CHD risk.

Prospective studies of β-carotene intake showed a significantly lower risk of cardiovascular death among those consuming more β-carotene. The results from the trials, however, indicated a moderate adverse effect of β-carotene supplementation. The results are summarized in Fig. 4

It is likely that the apparent protective association in observational studies represents confounding evidence, as people who eat diets rich in β-carotene or who take supplements are more socioeconomically advantaged and adopt a number of protective health-related behaviors. This finding is of importance both in itself, and as a demonstration of the kinds of spurious findings which can be observed in observational epidemiological studies. The same cautious approach to the interpretation of evidence is certainly required with respect to studies of psychosocial factors.

4.6 Smoking

Smoking was identified as a cause of lung cancer by German researchers in the 1930s and 1940s. Since the results of prospective epidemiological studies in the 1950s it has been clear that heavy cigarette smokers have approximately twice the risk of CHD (Labarthe 1998). This risk is reduced by giving up cigarettes, and returns to that of non-smokers after about 10 years.

4.7 Obesity

Very obese individuals have an increased risk of CHD. A U-shaped association between adiposity and CHD mortality has been seen in many studies. It seems likely that the high mortality among thin people occurs because some thin people are ill and some thin people smoke or drink heavily. Obesity is at least partially related to higher CHD risk because it is associated with higher blood pressure and an unfavorable blood lipid profile (Labarthe 1998). Hereditability models suggest that obesity is, to a remarkable degree, a genetic characteristic of individuals, with only a small proportion of the variance being accounted for by environmental factors. However, the substantial changes in the prevalence of obesity which have occurred within populations such as the UK over a relatively short period of time indicate that environmental factors (notably low levels of physical activity) are of great importance.

4.8 Physical Inactivity

Initially, it was thought that exercise predisposed to coronary heart disease, through cardiac strain. How-

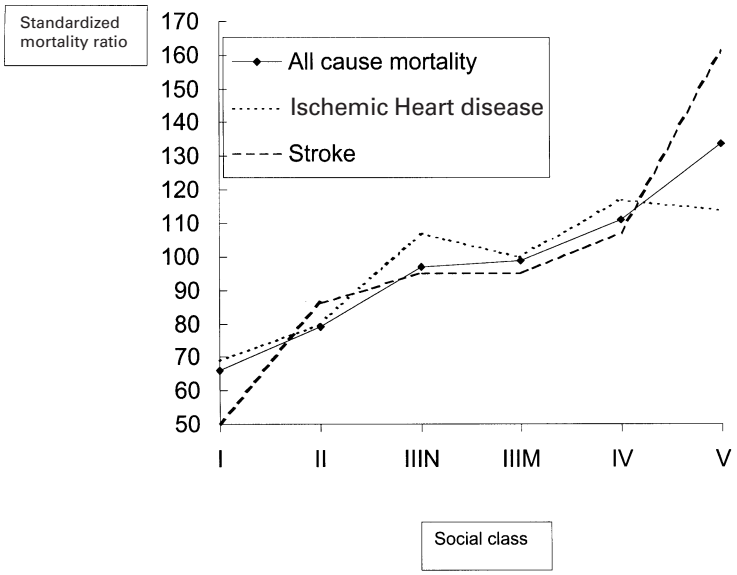

Figure 5
Social class differences in mortality of men 15–64 between 1976 and 1989 from various causes of death
Source: Harding S. Social class differences in mortality of men: recent evidence from the OPCS Longitudinal Study. *Population Trends* 1995; **80**: 31–37

ever, research starting in the mid-twentieth century suggested that occupational physical activity was associated with reduced risk of CHD. Clearly, health-related selection is a serious problem in studying the relationship between physical activity and coronary heart disease, since those with early signs of the disease will be less active. Engagement in leisure time physical activity is also associated with other important risk factors for CHD, such as socioeconomic position. However, the evidence suggests that increased physical activity does protect against CHD, although there is uncertainty about the type, intensity, and frequency of exercise required. A formulation consistent with the current data is that physical activity, which is at a sufficient level to produce cardio-respiratory training, reduces the risk of coronary heart disease, while the value of lower intensity physical activity is unclear. The level of activity which is protective will therefore depend upon the fitness level of an individual. For the unfit individual, even very mild exercise would produce a cardio-respiratory training effect. As the individual gets fitter, a higher degree of exercise will be required.

4.9 Cholesterol

Circulating blood cholesterol is strongly and positively associated with CHD risk. A series of large therapeutic trials of cholesterol reduction demonstrated that this process is reversible and risk reductions of the magnitude predicted from the observational data can be produced. A 20 percent reduction in cholesterol is associated with around 25 percent reduction in coronary heart disease mortality risk (Ebrahim et al. 1998). Early epidemiological studies only measured total circulating cholesterol, but it is evident that subfractions of cholesterol have differential effects on CHD risk. The adverse effects are restricted to the low-density lipoprotein cholesterol fraction, with high-density lipoprotein cholesterol (HDL) levels being inversely associated with coronary heart disease risk.

4.10 Blood Pressure

There is a strong dose–response relationship between increased blood pressure and increased risk of CHD. Large randomized trials of pharmacological blood pressure reduction in middle age have confirmed that blood pressure reduction reduces subsequent risk of CHD (Collins and MacMahon 1994). Several observational studies have also shown that increased blood pressure in early adult life is associated with increased CHD mortality in later life, suggesting that risk trajectory may be set early (McCarron et al. 2000).

4.11 Hemostatic Factors

In the middle of the nineteenth century, Karl Freiherr von Rokitansky of Vienna proposed that CHD had a thrombogenic basis. Since the 1980s, a number of cohort studies have examined the association between

hemostatic factors and coronary disease (Labarthe 1998). There is a consistent independent association between increased fibrinogen levels and increased risk of coronary disease. Some prospective studies have reported increased coronary risk with factors VII and VIII but the data are less extensive and less consistent. Smokers have higher fibrinogen levels than non-smokers and this may in part explain the excess coronary risk observed in smokers. However, the causal nature of these associations is still unresolved, as both risk factors and pre-existing disease may alter hemostatic factor levels, generating associations between hemostatic factors and disease which are noncausal.

4.12 Socioeconomic Position

In most industrialized countries CHD risk is higher amongst people living in worse social circumstances. Whether the apparent gradients in the opposite direction seen during the early stages of the CHD epidemic in industrialized countries, and in newly industrialized countries currently, are real or reflect differential disease coding and reporting, remains uncertain (Davey Smith 1997). Social disadvantage can influence CHD risk in a wide variety of ways, and act across the entire lifecourse. People born into worse social circumstances are likely to have lower birthweight. Socioeconomic deprivation in childhood is associated with poor nutrition and growth, and possibly higher incidence of childhood infections. Behavioral patterns—in particular with relation to diet, physical activity, alcohol consumption, and smoking—are developed in childhood in a socially patterned manner and track into adulthood. In most industrialized countries smoking and dietary patterns associated with increased CHD risk are more common among poorer adults.

Given this wide array of exposures associated with socioeconomic deprivation which could increase the risk of CHD, it is not surprising that studies which have taken into account a limited number of these find they fail to account fully for the socioeconomic distribution of CHD. The large (and in many countries increasing) socioeconomic differentials in CHD risk provide both a model for testing etiological hypotheses and also evidence of an important potential locus for public health interventions. Figure 5 shows the threefold difference in CHD mortality between men in unskilled manual occupations in Britain compared to those in managerial and professional occupations, and the graded difference in risk between these occupational extremes. Clearly, reducing CHD risk among the less socioeconomically advantaged to the same level as that of the most advantaged would have a dramatic influence on overall diseases rates in the population. A further implication is that since socioeconomically disadvantaged people have a considerably increased risk of CHD, which is not purely dependent upon conventional risk factors, they have the most to gain from intervention (see *Socioeconomic Status and Health*).

4.13 Stress

Stress is considered by the general public to be an important cause of CHD. Stress is, however, difficult to define or measure. Epidemiological studies have examined a wide range of potential exposures that fall under the general heading of 'stress.' These include measures which are essentially of personality traits—such as the Type A behavior pattern—through to records of stressful life events or global measures of perceived stressfulness of daily activities. A broader conceptual category of 'psychosocial factors' is now widely used in the epidemiological literature. This includes aspects of social life such as the strength of social support networks or the level of control over their work which people have.

The type A behavior pattern was investigated from the 1950s as a potential cause of CHD. Type A individuals—those who are involved in an incessant struggle to achieve more and more in less and less time—were found to have higher risk of CHD in several early prospective studies (Labarthe 1998). More recent studies have failed to find any association between type A behavior and CHD. It may be that this association no longer exists because CHD is not now considered a disease of the wealthy, stressed, businessman, who in the past was more likely to be diagnosed with the condition. Research on behavioral traits now focuses on those elements of the original type A classification, which are related to adverse social background (such as hostility). These may predict CHD risk because of this, rather than because of any causal link with the disease. Ongoing studies should help to resolve this issue in particular trials of stress management instituted to establish whether interventions in this area can produce beneficial changes (see *Coronary-prone Behavior, Type A*).

Measures of self-reported global stress and other self-reported indicators such as low control at work or poor social networks tend to be related to self-reported CHD symptoms. They less consistently relate to objective measures of CHD. This may reflect an underlying reporting tendency such that people who report higher levels of adversity also report higher levels of symptoms. People with low control over their work (e.g., shift workers in a factory) are, in general, in less favorable social locations than those with higher control over their work. It is not surprising that the former have higher rates of CHD, given the strong social patterning of the disease. Whether these adverse psychosocial factors are one of the causes of the high CHD mortality in the less advantaged or is merely

another consequence of material inequalities is currently not clear.

Further, more methodologically robust, studies employing objective measures of both stress and CHD outcomes are required. It is remarkable that the supposed biological mediators between stress and disease–disturbed functioning of the hypothalamo-pituitary-adrenal axis and possible immune system outcomes of this—have not themselves been related to CHD risk in prospective observational studies. The current controversy—with enthusiastic proponents of stress as a cause of CHD on one side lined up against the majority of academic and research cardiologists and epidemiologists who dismiss the association as spurious—will only be resolved when studies employing better methods report their findings.

5. *Lifecourse Approach*

The long incubation period for CHD has been recognized for many years (Rose 1982), and a lifecourse approach to its etiology is a natural extension of this view. In this discussion we have focused on studies for which the main focus is the social patterning of CHD risk.

However, most of the important risk factors identified for CHD are socially patterned. Table 2 summarizes factors which are putative or established CHD risk factors and/or are of particular interest from a lifecourse perspective, according to their period of influence. It is clear that CHD can be considered the archetype of diseases whose determinants should be sought across the entire lifecourse—from conditions existing at the time of conception and during intrauterine development, through nutrition, growth, and health in childhood, to social conditions, occupation, diet, physical activity, and smoking throughout adult life. In Table 3, we combine a marker of early life circumstances—father's social class—and behavioral and socioeconomic factors acting in adulthood and demonstrate their joint influence on CHD mortality. If factors acting at different stages of life—such as childhood social circumstances, education, deprivation level of area of residence in adulthood, adulthood occupational social class, smoking, and drinking in adulthood—are combined, then very large differences in cardiovascular disease risk can be observed (Davey Smith and Hart 2001). Socioeconomic inequalities in CHD can only be understood through consideration of how a variety of exposures which increase CHD risk are influenced by social circumstances, how this social dependence leads to them clustering across time, such that some individuals are adversely influenced by a wide array of risk factors, and thus how the accumulation of—and interaction between—influences acting at different stages of life determines the pattern of CHD within and between populations.

Table 2
Lifecourse risk factors for coronary heart disease

Maternal health, development and diet before and during pregnancy
Parental history of CHD
Low birth weight
Socioeconomic deprivation from childhood onward
Stress from childhood onward
Poor growth in childhood
Short leg length in childhood
Obesity in childhood
Certain infections acquired in childhood
Diet from childhood onwards
Blood pressure in late adolescence
Serum cholesterol in late adolescence
Smoking from late adolescence onwards
Little physical activity from late adolescence onwards
Blood pressure in adulthood
Serum cholesterol in adulthood
Obesity in adulthood
Job insecurity and unemployment in adulthood
Short stature in adulthood
Binge alcohol drinking in adulthood
Diabetes and components of syndrome X in adulthood
Elevated fibrinogen and other acute phase reactants in adulthood
Certain infections acquired in adulthood

Table 3
Age adjusted relative rates of CHD mortality by father's social class and later life risk factors in the West of Scotland Collaborative study

Father's social class	Smoking	
	Other	Current cigarette
Nonmanual	1	2.01 (1.46–2.77)
Manual	1.56 (1.18–2.07)	2.78 (2.12–3.63)
	Alcohol	
	< 15 units/week	⩾ 15 units/week
Nonmanual	1	1.17 (0.80–1.70)
Manual	1.44 (1.18–1.75)	1.86 (1.50–2.31)
	Screening social class	
	Nonmanual	Manual
Nonmanual	1	1.53 (1.06–2.21)
Manual	1.46 (1.18–1.82)	1.78 (1.45–2.17)

6. *The Prevention of Coronary Heart Disease*

There are two differing but complementary approaches to prevention: the high-risk approach and

the population approach. The high-risk approach seeks to identify those at highest risk of developing disease and intervene to prevent disease. With this approach it may be easier to encourage people at high risk to alter their lifestyles as these people have a considerable amount to gain personally. The recent results from the WHO MONICA study seem to confirm that the widespread use of effective treatments can reduce the number of symptomatic episodes and deaths attributed to CHD. Such an approach, however, requires that those at high risk be identified and ignores the people who develop symptomatic disease who are at moderate risk. For example, in the Whitehall study only around 40 percent of the coronary deaths occurred in the 20 percent at highest risk.

The alternative approach is the population approach, which seeks to alter behavior across the whole population. The advantage is that it is likely to produce large improvements in population health, for fairly modest changes in behavior. The disadvantage is that the absolute benefit for each individual who makes a lifestyle change is small. This mismatch between the likely size (Rose 1985) of the benefits to the individual and population has been called the 'Prevention paradox.'

It has been shown that brief advice from physicians leads to small, but worthwhile, reductions in smoking rates. Nicotine replacement gum can also aid smoking cessation. School-based antismoking programs have, however, shown disappointing results. The detection and treatment of high blood pressure reduces the risk of CHD and has the added benefit of substantially reducing the risk of stroke. Similarly, lowering cholesterol levels with the statin drugs reduces CHD risk in primary prevention.

Health-related behaviors are influenced strongly by societal, legislative, and fiscal forces, perhaps more so than by individual levels of knowledge and degree of willpower. This is the case with smoking, where cigarette use in a country is responsive to the competing incentives offered by tobacco companies and government policies. The prevention of smoking can only be ultimately successful if the world community accepts the need to restrict the growing, marketing, and use of tobacco (see *Smoking and Health*).

The differences in both behaviors and CHD disease rates clearly illustrate the capacity for populations to change. While simple invocations have modest effects on behavior, social change can result in profound improvements in health. Conversely, where cultures fail to adapt or social structures collapse (as happened in Russia with the demise of the Soviet Union) CHD can increase. Effecting societal change is complex but the potential public health benefits are profound. The concentration of CHD among those in society who are least advantaged suggest that any such policies will be more effective if we seek to improve their lot by improving their income, opportunities, and self-esteem. More inclusive and extensive programs that seek to reduce inequalities in health and improve health for all offer real opportunities to reduce CHD.

See also: Cardiovascular Conditioning: Neural Substrates; Coronary Heart Disease (CHD), Coping with; Coronary-prone Behavior, Type A; Gender and Cardiovascular Health; Health Behaviors; Stress and Cardiac Response; Stress and Health Research; Stress: Psychological Perspectives

Bibliography

Barker D J P 1998 *Mothers, Babies and Health in Later Life*. Churchill Livingstone, Edinburgh, UK

Barrett-Connor E 2001 Hormone replacement therapy: Risks and benefits. *International Journal of Epidemiology*. **30**: 423–6

Collins R, MacMahon S 1994 Blood pressure, anti-hypertensive drug treatment and the risks of stroke and coronary heart disease. *British Medical Bulletin* **50**: 272–98

Davey Smith G 1997 Socioeconomic differentials. In: Kuh D, Ben-Shlomo Y (eds.) *A Life Course Approach to Chronic Disease Epidemiology*. Oxford University Press, Oxford, UK, pp. 242–73

Davey Smith G, Hart C 2001 Lifecourse socioeconomic and behavioral influences on cardiovascular disease mortality: The collaborative study. *American Journal of Public Health*

Ebrahim S, Davey Smith G, McCabe C, Payne N, Pickin M, Sheldon T A, Lampe F, Sampson F, Ward S, Wannamthee G 1998 Cholesterol and coronary heart disease: Screening and treatment. *Quality in Health Care* **7**: 232–9

Evans A 1995 Dr. Black's favourite disease. *British Heart Journal* **74**: 676–7

Gunnell D J, Davey Smith G, Frankel S, Nanchahal K, Braddon F E M, Pemberton J, Peters T J 1998 Childhood leg length and adult mortality: Follow up of the Carnegie (Boyd Orr) survey of diet and health in pre-war Britain. *Journal of Epidemiology and Community Health* **52**: 142–52

Labarthe D R 1998 *Epidemiology and Prevention of Cardiovascular Disease: A Global Perspective*. Aspen, Gaithersburg, MD

McCarron P, Davey Smith G, Okasha M, McEwen J 2000 Blood pressure in young adulthood and mortality from cardiovascular disease. *Lancet* **355**: 1430–1

Neaton J D, Wentworth D 1992 Serum cholesterol, blood pressure, cigarette smoking and death from coronary disease: Overall findings and differences by age for 316,099 white men. *Archives of International Medicine* **152**: 56–64

Peterson S, Mockford C, Rayner M 1999 *Coronary Heart Disease Statistics*. British Heart Foundation, London

Powles J W 1994 Greek migrants in Australia: Surviving well and helping their hosts. In: Marks L, Worboys M (eds.) *Migrants, Minorities and Medicine: Historical and Contemporary Perspectives*. Routledge, London

Rose G 1982 Incubation period of coronary heart disease. *British Medical Journal* **284**: 1600–1

Rose G 1985 Sick individuals and sick populations. *International Journal of Epidemiology* **14**: 32–8

G. Davey Smith and A. R. Ness

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Coronary-prone Behavior, Type A

In the late 1950s, two cardiologists working in San Francisco described what they referred to as an 'action–emotion complex' that seemed to characterize patients who developed coronary heart disease (CHD) at an early age. Meyer Friedman and Ray Rosenman suggested that this set of behavioral characteristics, which they labeled the Type A Behavior Pattern, represented a risk factor for CHD. In their view, individuals characterized by intense achievement striving and competitiveness, a persistent desire for recognition and advancement, constant involvement in multiple tasks and activities, a rapid pace of speaking and behaving, time urgency and impatience with others, and easily aroused hostility and aggressiveness were at greater risk of developing CHD than were more relaxed, easy-going, and agreeable persons, who they labeled 'Type B.' In the following 20 years, this concept had a major impact on the emerging fields of health psychology and behavioral medicine, as well as the science and practice of cardiology. The concept even penetrated popular culture, where the labels Type A and B have become widely recognized.

1. *Rebirth of a Mind–Body Hypothesis*

The notion that characteristic patterns of emotion and behavior could confer risk for the development of CHD was not novel at the time of Friedman and Rosenman's initial description of the Type A pattern. Over 100 years ago, Sir William Osler, one of the pioneers in the clinical description of CHD, observed that this illness befell, 'not the neurotic, delicate person … but the robust, the vigorous in mind and body, the keen and ambitious man, the indicator of whose engine is always at full speed ahead.' Prior to the 1950s, psychoanalytically oriented clinicians had also suggested that personality traits such as excessive ambition and aggressiveness seemed to characterize coronary patients and hence might indicate that personality traits somehow contributed to the development of the disease. However, these observations were not subjected to thorough scientific investigation, and as a result had limited impact on behavioral or medical science and even less on the clinical care of coronary patients.

Early in their program of research, Friedman and Rosenman reported controlled studies in which patients with and without CHD differed in the behavioral characteristics included in their description of the Type A pattern. Further, they found that Type A and B individuals differed in physiological characteristics that could be involved in the development of the disease. For example, Type A persons displayed greater daily levels of excretion of catecholamines and other stress hormones compared to Type B individuals. In this early work, Friedman and Rosenman intentionally avoided framing their research as an investigation of personality characteristics because of the limited impact and skepticism surrounding the previous, psychoanalytic speculations about the role of personality in CHD. Most importantly, in the late 1960s they launched a large prospective, longitudinal study of the association between the Type A pattern and subsequent CHD. The prospective design provided a more rigorous test of the hypothesis that Type A behavior somehow influenced the development of CHD when compared to the cross-sectional research they had conducted up to that point. The Western Collaborative Group Study (WCGS: Rosenman et al. 1975) demonstrated that among initially healthy men, Type A's were about twice as likely to develop CHD over an eight-year follow-up period than were Type B's. This relationship remained significant even when established CHD risk factors (e.g., smoking) were statistically controlled. This study provided the first prospective evidence that the Type A pattern was an independent risk factor for cardiovascular disease, at least in a sample of predominately white, North American, middle-class males.

The results of the WCGS prompted a variety of other investigations, and evidence regarding the health implications of the Type A pattern accumulated rapidly. In 1978, the National Heart Lung and Blood Institute sponsored a panel of neutral experts and charged them with the task of reviewing this evidence. The panel's report concluded that the Type A pattern was indeed a reliable, independent risk factor, and that it was about as important in the development of CHD as were the established risk factors such as smoking and high blood levels of cholesterol (Review Panel 1981). This represented a unique development in the study of CHD, as it was the first time a psychological construct had been endorsed as a risk factor by a panel of biomedical experts. The Type A pattern quickly became a central focus of a rapidly emerging emphasis on psychological aspects of cardiovascular disease (Steptoe 1981). However, as discussed below, subsequent years witnessed the accumulation of a more mixed body of evidence concerning the association between Type A behavior and CHD, and a reformulation of what are the key aspects of the Type A syndrome related to CHD and other health outcomes.

2. *Assessment and Underlying Mechanisms*

A variety of methods have been developed to assess Type A, coronary prone behavior. The 'gold standard' of these assessment procedures is the Type A Structured Interview. This approximately 10-minute interview consists of a series of questions concerning the individual's work habits, daily routines, and aspects of the Type A pattern. A semistructured set of prompts and follow-up questions are also included, and many questions are presented in such a way as to evoke Type

A responses from predisposed persons. For example, the interviewer is occasionally somewhat challenging and even provocative, or hesitates and delays the completion of a rather obvious question. The content of the individual's answers is not the primary source of information in making a judgement about the presence and degree of Type A characteristics. Rather, the assessment is based primarily on the style of response. A loud, rapid, and vigorous vocal style, interruptions and 'talking over' the interviewer, and subtle or nonverbal expressions of impatience (e.g., an exasperated sigh), hostility or irritation toward the interviewer (e.g., sarcasm or condescension) are weighted heavily in the classification of an individual as Type A or B, although self-reports of competitiveness, impatience, overinvolvement in work, and easily provoked anger are considered as well. This emphasis on displayed rather than self-reported characteristics is consistent with Friedman and Rosenman's prior clinical observations that Type A individuals were often unable or unwilling to give accurate descriptions of their emotions and actions, perhaps reflecting a lack of insight or a defensive, uncooperative, or mistrusting nature. Although the Structured Interview procedure can produce quite reliable ratings, it has obvious limitations stemming from the need for expert interviewers and raters. Hence, several self-report questionnaires have been developed as alternatives to the Structured Interview. The most widely used of these in the history of Type A research is the Jenkins Activity Survey (JAS: Jenkins et al. 1974). This instrument consists of several multiple-choice questions about the various components of the Type A pattern. Factor analyses have indicated that the primary dimensions of the JAS are hard-driving competitiveness, impatience and speed of activity, and job involvement. Importantly, the JAS does not thoroughly sample the aggressive verbal style, anger, and easily evoked hostility that are weighted heavily in the ratings derived from the Type A Structured Interview. Like many other self-report measures of the Type A pattern, the JAS scores are not closely correlated with the Structured Interview ratings, and are not consistently related to CHD in prospective studies. Further, the various self-report scales are typically only modestly correlated with one another, and in many instances have unique correlates with other psychological characteristics. This suggests that although often grouped together as Type A research, studies using different assessment procedures are likely to be examining what are actually only weakly related personality and behavioral characteristics. These important differences across measures of the construct have contributed to inconsistency in the body of research on the Type A pattern, in part because they identify possibly quite distinct persons as Type A. For example, the self-described competitive, achievement striving, overextended, hurried, and impatient person who scores high on the JAS may or may not have the emphatic, rapid, dominant, and unfriendly vocal and interactional style that would warrant a classification of Type A in The Structured Interview.

A variety of models of the underpinnings of the Type A pattern have been proposed. Among the more influential of these views is that of Glass (1977). He suggested that Type A characteristics reflected a strong desire to exercise control over the people and events of one's life, and a low threshold for perceiving threats to control. From this perspective, the overt Type A characteristics themselves represent a well-developed aggressive style of asserting control in response to perceived threats of loss of control. In another influential model, Price (1982) suggested that a set of beliefs comprise the psychological core of the Type A pattern. She argued that Type A's believe that one must continuously demonstrate self-worth through achievement, that the resources and opportunities for such achievement are limited and often scarce, and that there is no universal moral principle to ensure that other people will be fair in their pursuit of these valued achievements. As a result, in contrast to the more relaxed and trusting Type B, Type A individuals are engaged in a continuous struggle to demonstrate their self-worth, through competitive and adversarial interactions with unprincipled others. Hence, the overt Type A characteristics are the manifestations of this continuous struggle to sustain externally-referenced self-esteem or value, in real or imagined competition with adversaries. These conceptual models have stimulated a great deal of research on the psychological basis of Type A persons, and have served as guides to the development of interventions for the reduction of Type A behavior (Rhodewalt and Smith 1991).

In most models of Type A behavior, the impact of this psychological variable on the development of CHD is seen as occurring through the mechanism of physiological responses to psychological stress. Briefly, Type A's are believed to respond to challenges, threats, and demands with larger, sympathetically-mediated increases in heart rate, blood pressure, and various stress hormones. Over long periods of time, the repetition of this physiological reactivity can initiate and hasten the development of coronary artery disease (CAD) or coronary atherosclerosis, the underlying condition common to the various clinical manifestations of CHD. CAD begins with microscopic injuries to the inner lining, or endothelium of the coronary arteries, and continues through the progressive buildup of fatty deposits at these injury sites. In later stages, a hard, fibrous cover forms, and these advanced plaques can rupture or bleed. At the location of these coronary artery lesions, the passage through which blood would otherwise pass becomes narrowed and oxygen supply to the heart muscle is consequently impeded. Once CAD has advanced to a sufficient degree, exaggerated psychophysiological reactivity can increase the heart's need for oxygen as it beats

more rapidly and forcefully during sympathetic activation. Further, this reactivity can contribute decreased blood and oxygen supply to the heart through the transient narrowing of the coronary artery at plaque sites. Through either or both increased demand or decreased supply of oxygen, psychological stress can evoke an insufficient supply of oxygen to the heart muscle. The resulting myocardial ischemia can provide the clinical manifestations of CHD, such as angina pectoris (chest pain), myocardial infarction (death of heart muscle), or sudden coronary death. A large body of evidence suggests that compared to Type B's, Type A persons do respond to potential stressors with more pronounced psychophysiological responses (Houston 1988). Further, other evidence indicates that such responses can contribute to the initial development of CAD and to myocardial ischemia in individuals with advanced CAD (Rozanski et al. 1999).

In an extension of these models of the connection between Type A behavior and CHD, Smith and Anderson (1986) suggested that Type A individuals not only respond to everyday stressors with heightened physiological reactivity; relative to Type Bs, they are also exposed to more frequent, severe, and prolonged stressors. This greater exposure to challenges, demands, and hassles in everyday life reflects a stress-engendering process, in which the ambitious, competitive, controlling, and hostile style of Type A's leads to more tasks, commitments, and obligations, and to competition, opposition, and antagonism in interactions with others. Hence, the exaggerated psychophysiological reactivity that links Type A behavior to disease comes from two sources—the stressors all individuals face and the additional challenges, threats, and demands Type A's create for themselves.

Another important view of the association between Type A behavior and CHD was suggested by Krantz (Krantz and Durel 1983). He argued that the overt manifestations of Type A behavior are the consequence rather than the cause of the psychophysiological reactivity associated with this pattern. In this constitutional vulnerability model, Type A behaviors are produced by an underlying biologic predisposition to heightened sympathetic nervous system (SNS) responsivity. This underlying responsivity, in turn, also fosters the development of CAD and CHD. Consistent with this view, twin studies have indicated that the Type A pattern is at least partially heritable, and pharmacological studies suggest that the pattern is related to an underlying SNS responsivity (Smith and Gallo 2001).

3. *Evolution of Type A Research*

Shortly after the initial conclusion that the Type A pattern was a reliable, independent risk factor for CHD, several large prospective studies failed to replicate this effect. This led to skepticism in the scientific community about the actual importance of the construct. However, a careful analysis of the evidence revealed possible explanations for the inconsistencies in this literature. In a quantitative review of the available evidence, Miller et al. (1991) found that the Structured Interview was a much more reliable predictor of subsequent CHD than the JAS and other self-report measures of Type A, and that the association between Type A and CHD was larger in initially healthy samples than it was among high-risk samples and among samples with established CHD. The apparent temporal trend of increasing failures to replicate the association between Type A behavior and CHD apparently reflected a growing use of high-risk samples and self-report measures of the behavior pattern. When defined by the Structured Interview, the combined evidence still indicated that Type A behavior was a reliable risk factor for the development of CHD among initially healthy individuals, and that the weaker association in high-risk individuals and clinical samples of CHD patients could be attributed to methodological problems inherent in such study designs. Hence, although the emergence of inconsistent evidence led many researchers and clinicians to lose interest in the Type A concept, this waning importance seems to have been unwarranted.

Inconsistent findings and growing skepticism also had a very positive consequence for the study of coronary-prone behavior. Following earlier seminal work by Matthews et al. (1977), many investigators began to examine the predictive utility not of the global Type A pattern but of its individual components. In these more specific and refined analyses, hostility quickly emerged at the most 'toxic' component among the array of individual Type A characteristics, confirming a centuries-old hypothesis about the negative health effects of anger (Siegman 1994). As an umbrella term, hostility refers to individual differences in the tendency to experience anger, negative beliefs and attributions concerning the motives and intentions of others (e.g., cynicism and mistrust), and overtly aggressive verbal and physical behavior. Achievement striving, competitiveness, and impatience, in contrast, did not seem to be reliable risk factors for cardiovascular disease.

Because measures of hostility existed in several large studies in which individuals had completed personality tests decades earlier and were followed over many years, independent tests of the health consequences of hostility could be conducted without the need to begin additional long-term, prospective studies. For example, hostility scales could be derived from the true–false item pool comprising the Minnesota Multiphasic Personality Inventory (MMPI), and this psychological test had been in wide use for many several decades, nearly throughout the world. Hence, a literature on this element of the Type A pattern grew rapidly. As with the global Type A pattern, these

prospective studies of hostility and subsequent health included mostly positive or confirmatory findings regarding the negative health consequences of hostility, but some failures to replicate the effect as well. However, in a recent quantitative review of this literature, behavioral ratings of hostility and self-report measures of the construct were found to be reliable predictors of the later development of CHD and premature mortality (Miller et al. 1996).

As in the case of the global Type A pattern, research on the mechanisms linking hostility and CHD has focused on psychophysiologic responses to potential stressors (Williams et al. 1985). In response to potential social stressors administered in the laboratory such as disagreements, conflicts, debates, and harassment, hostile persons respond with larger increases in heart rate, blood pressure, and stress hormones (e.g., circulating catecholamines) than their less hostile counterparts. Further, ambulatory studies indicate that this exaggerated psychophysiological reactivity is also characteristic of the responses of hostile persons to their natural, daily environments (Smith and Gallo 2001). Other research has indicated that among persons with CAD, episodes of anger can precipitate myocardial ischemia and even infarction (Rozanski et al. 1999), and that hostile persons with CAD may be particularly prone to such responses. Hence, the psychophysiological responsiveness of hostile persons may be the key mechanism through which they are at increased risk of disease.

Hostility is also associated with greater exposure to interpersonal stressors such as heightened levels of conflict with and reduced levels of social support from friends, family members, and co-workers (Smith and Gallo 2001). Further, much of this 'psychosocial vulnerability' associated with hostility may stem from the impact of the hostile person's thoughts and actions on his or her social environment. Given their mistrusting, cold, and antagonistic interpersonal style, hostile persons may undermine sources of social support and create conflicts, thereby creating more frequent, prolonged, and extreme social stressors to which they react with heightened reactivity. This discordant, unsupportive social environment would also tend to maintain and perhaps even exacerbate the hostile personality style.

Hostility may not be the only 'toxic' element within the global Type A pattern. In a reanalysis of the original Structured Interviews from the WCGS, Houston et al. (1997) found that behavioral ratings of hostility and a vigorous, dominant interactional style were separate predictors of later CHD and premature death. These authors suggested that hostility and social dominance are separate coronary-prone characteristics. The possible health consequences of social dominance have a clear parallel with one of the best-developed animal models of stress and individual differences in social behavior as risk factors for cardiovascular disease. Kaplan et al. (1994) demonstrated that socially dominant, male cynomolgus monkeys develop more severe coronary artery disease in response to chronic social stress (i.e., recurrent exposure to novel cage mates) than socially subordinate males. Further, this differential impact of social stress among dominant males on vulnerability to coronary disease is apparently mediated by sympathetic nervous system activity. When dominant male animals receive pharmacologic blockade of beta-adrenergic input to the heart, they no longer display vulnerability to coronary atherosclerosis even when under chronic social stress. Hence, chronic sympathetic nervous system activation of the heart and vasculature, as dominant males recurrently assert their status and respond to related challenges and threats, may underlie this psychosocial effect on atherosclerosis. In humans, efforts to assert dominance or control over others is also associated with increases in blood pressure and heart rate (Smith et al. 2000), again suggesting that psychophysiological reactivity could connect a second 'toxic' component of Type A behavior—social dominance—to the later development of CHD.

4. *Application of Type A Research*

The most obvious benefit of research identifying risk factors for the development and progression of CHD is the design of related, risk reducing interventions. In the case of Type A, coronary-prone behavior, this actually raises two questions. First, can Type A behavior be changed? Second, if so, do such changes reduce the risk of CHD? These questions were the focus of a large, controlled clinical trial conducted by Friedman et al. (1984)—the Recurrent Coronary Prevention Project (RCPP). Compared to a control group receiving typical information and counseling to make appropriate changes in health habits to reduce cardiac risk (e.g., smoking cessation, diet, exercise), postmyocardial infarction patients who had undergone group therapy designed to identify and reduce Type A behavior were indeed able to change these characteristics, as measured by Structured Interview ratings. Further, patients receiving this psychological intervention were much less likely to experience a recurrence of their coronary disease such as a second myocardial infarction. Among patients with mild initial coronary disease, the intervention also reduced the incidence of subsequent coronary deaths (Powell and Thoresen 1988). Based on the conceptual model of Price described above, this intervention consisted of cognitive restructuring directed toward altering the beliefs underlying the Type A pattern, reducing exposure to environmental cues that activate Type A tendencies, rehearsing alternative, Type B responses to such situations, and instruction and practice of relaxation exercises and other stress management skills.

A variety of other programs have been developed to modify the Type A pattern, but few have been evaluated in large controlled trials with definitive medical outcomes as in the RCPP. However, the general approach of stress management and related psychological interventions for coronary patients has been found to be successful in reducing recurrent events in this population (Linden et al. 1996). For example, Blumenthal et al. (1997) found that a stress management intervention reduced the frequency of episodes of myocardial ischemia occurring during the daily activities of patients with advanced CAD. Although not specifically targeted at Type A behavior, such interventions are designed to reduce complications from heart disease by altering emotional and physiological stress responses.

5. *The Future of Research on Coronary Prone Behavior*

Since the height of scientific and popular attention in the 1970s and 1980s, the number of research studies published each year on Type A behavior has steadily decreased. However, this reflects an evolution rather than growing neglect. The focus on hostility as the toxic component within the global Type A pattern continues in epidemiological, psychophysiological, and intervention research. Further, other individual differences have emerged as reliable risk factors as the search for coronary prone behavior continues and broadens. For example, symptoms of anxiety and depression appear to be reliable risk factors for the initial development of CHD, and are associated with poor prognosis among patients with established disease (Januzzi et al. 2000). These other coronary-prone behaviors will also be the focus of additional studies of their predictive utility, the psychobiological mechanisms through which they impact the development and course of CHD, and the effectiveness of related interventions. The addition to the specific research discussed above, the seminal work of Friedman and Rosenman represents a cornerstone of the study of behavioral aspects of cardiovascular disease. In fact, despite an inconsistent reputation in the scientific community and the popular press, the Type A concept played a central role in the rapid development of the broader fields of health psychology and behavioral medicine. It also has been influential in the expanded attention to psychological factors in the clinical care of patients with CHD.

See also: Coping across the Lifespan; Coronary Heart Disease (CHD), Coping with; Coronary Heart Disease (CHD): Psychosocial Aspects; Gender and Cardiovascular Health; Personality and Health; Stress and Cardiac Response; Stress and Coping Theories; Stress and Health Research; Well-being and Health: Proactive Coping

Bibliography

Blumenthal J A, Jiang W, Babyak M et al. 1997 Stress management and exercise training in cardiac patients with myocardial ischemia: Effects on prognosis and evaluation of mechanisms. *Archives of Internal Medicine* **157**: 2213–23

Friedman M, Thoresen C E, Gill J J, Powell L H, Ulmer D, Thompson L, Price V A, Rabin D, Breall W, Dixon T, Levy R, Bourg E 1984 Alteration of Type A behavior and reduction in cardiac recurrences in post-myocardial infarction patients. *American Heart Journal* **108**: 237–48

Glass D C 1977 *Behavior Patterns, Stress, and Coronary Disease*. Erlbaum, Hillsdale, NJ

Houston B K 1988 Cardiovascular and neuroendocrine reactivity, global Type A, and components of Type A behavior. In: Houston B K, Snyder C R (eds.) *Type A Behavior Pattern: Research, Theory & Intervention*. Wiley, New York, pp. 212–53

Houston B K, Babyak M A, Chesney M A, Black G, Ragland D R 1997 Social dominance and 22-year all-cause mortality in men. *Psychosomatic Medicine* **59**: 5–12

Januzzi J L, Stern T A, Pasternak R C, DeSanctis R W 2000 The influence of anxiety and depression on outcomes of patients with coronary artery disease. *Archives of Internal Medicine* **160**: 1913–21

Jenkins C D, Rosenman R H, Zyzanski S J 1974 Prediction of clinical coronary heart disease by a test for the coronary-prone behavior pattern. *New England Journal of Medicine* **290**: 1271–5

Kaplan J R, Botchin M B, Manuck S B 1994 Animal models of aggression and cardiovascular disease. In: Siegman A W, Smith T W (eds.) *Anger, Hostility, and the Heart*. Lawrence Erlbaum, Hillsdale, NJ, pp. 127–48

Krantz D S, Durel L A 1983 Psychobiological substrates of the Type A behavior pattern. *Health Psychology* **2**: 393–411

Linden W, Stossel C, Maurice J 1996 Psychological interventions for patients with coronary artery disease: A meta-analysis. *Archives of Internal Medicine* **156**: 745–52

Matthews K A, Glass D C, Rosenman R H, Bortner R W 1977 Competitive drive, Pattern A, and coronary heart disease: A further analysis of some data from the Western Collaborative Group Study. *Journal of Chronic Disease* **30**: 489–98

Miller T Q, Smith T W, Turner C W, Guijarro M L, Hallet A J 1996 A meta-analytic review of research on hostility and physical health. *Psychological Bulletin* **119**: 322–48

Miller T Q, Turner C W, Tindale R S, Posavac E J, Dugoni B L 1991 Reasons for the trend toward null findings in research on Type A behavior. *Psychological Bulletin* **110**: 469–85

Powell L H, Thoresen C E 1988 Effects of Type A behavioral counseling and severity of prior acute myocardial infarction on survival. *American Journal of Cardiology* **62**: 1159–63

Price V A 1982 *Type A Behavior Pattern: A Model for Research and Practice*. Academic Press, New York

Rhodewalt F, Smith T W 1991 Current issues in Type A behavior, coronary proneness, and coronary heart disease. In: Snyder C R, Forsyth D R (eds.) *Handbook of Social and Clinical Psychology*. Pergamon, New York, pp. 197–220

Rosenman R H, Brand R J, Jenkins C D, Friedman M, Straus R, Wurm M 1975 Coronary heart disease in the Western Collaborative Group Study: Final follow-up of 82 years. *Journal of the American Medical Association* **233**: 872–7

Rozanski A, Blumenthal J A, Kaplan J 1999 Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. *Circulation* **99**: 2192–217

Siegman A W 1994 From Type A to hostility to anger: Reflections on the history of coronary-prone behavior. In: Siegman A W, Smith T W (eds.) *Anger, Hostility and the Heart*. Lawrence Erlbaum, Hillsdale, NJ, pp. 1–22

Smith T W, Anderson N B 1986 Models of personality and disease: An interactional approach to Type A behavior and cardiovascular risk. *Journal of Personality and Social Psychology* **50**: 1166–73

Smith T W, Gallo L C 2001 Personality traits as risk factors for physical illness. In: Baum A, Revenson T, Singer J (eds.) *Handbook of Health Psychology*. Lawrence Erlbaum, Mahwah, NJ

Smith T W, Ruiz J M, Uchino B N 2000 Vigilance, active coping, and cardiovascular reactivity during social interaction in young men. *Health Psychology* **19**: 382–92

Steptoe A 1981 *Psychological Factors in Cardiovascular Disorders*. Academic Press, London

Weiss B M 1981 Coronary-prone behavior and coronary heart disease: A critical review. *Circulation* **63**: 1199–215

Williams R B Jr, Barefoot J C, Shekelle R B 1985 The health consequences of hostility. In: Chesney M A, Rosenman R H (eds.) *Anger and Hostility in Cardiovascular and Behavioral Disorders*. Hemisphere, Washington, DC, pp. 173–85

T. W. Smith

Corporate Culture

1. *How to Think About and Define 'Culture'*

To understand corporate culture one must first understand the concept of culture. A chronic issue in conceptualizing culture is whether to think of culture as a static *property* of a given organization—its shared customs, beliefs, norms, values, and tacit assumptions, or whether to think of culture as a *dynamic human process* of constructing shared meaning (Frost et al. 1985). Culture *creation* is one of the unique characteristics of humans, being based on our capacity to be self-conscious and able to see ourselves and others from each others' points of view. It is this reflexive capacity of humans that makes culture possible. At the same time, it is the human need for finding meaning that creates the motivation for culture stabilization. Without some predictability, social intercourse becomes too anxiety provoking. Developing shared meanings of how to perceive, categorize, and think about what goes on around us is necessary to avoid the catastrophic anxiety that would result from reacting to everything as if it were a new phenomenon (Weick 1995).

Given these human characteristics, it then becomes clear that culture is both a process and a state. In new situations, shared meanings must be constructed through a social learning process. As these meanings help the participants to make sense of their world they become stabilized and can be viewed as 'states.' At the same time, as the members of a group interact, they not only recreate and ratify prior meanings but also construct new meanings as new situations arise.

A useful way to think about this issue is to take a cue from the anthropologist Marshall Sahlins (1985), who argues that one cannot really understand certain social phenomena without understanding both the historical events and the cultural meanings attributed by the actors to those events. While it is undeniably true that we produce and reinforce culture through perpetual enactment and sense making, it is equally true that the actors in those same social events bring to them some prior meanings, stereotypes, and expectations that can only be understood in a historical context. Culture production in the enactment sense, then, is either the perpetuation of or a change of some prior state, but that prior state can be thought of as 'the culture' up to that point. And one can describe that culture as if it were a 'state' of the existing system, even though one knows that the system is dynamic and perpetually evolving. The direction of that evolution will be a product of several forces: (i) technological and physical changes in the external environment; (ii) changes in the internal dynamics of the social system; and (iii) historical circumstances that are fortuitous or serendipitous.

For example, let me offer my oversimplified summary of Sahlins's very sophisticated analysis of the death of Captain Cook at the hand of the Hawaiians. Because Captain Cook was viewed as a God (as predicted in the Hawaiian mythology), the sexual favors offered to his sailors by the Hawaiian women were viewed as gifts and as opportunities to relate to the divine. The sailors' cultural background defined this as a version of prostitution, however, for which they felt they should pay. When they offered the women something in exchange for the sex, the women asked for something that was scarce in the society, namely metal. Once the loose metal on board the ships had been used up, the sailors began to pull nails from the ship itself, weakening it structurally. Hence, when Captain Cook set sail he discovered that the ships needed repair and ordered a return to harbor. In Hawaiian mythology a God returning under these circumstances had to be ritually killed. At the same time, Hawaiian social structure was undergoing change and became permanently altered because the subordinate role of women in the society was altered by their ability to acquire metal, a scarce resource that gave them social power.

When one contemplates this wonderful analysis it appear pointless to argue whether culture should be viewed as a state of the system or as a process of enactment. Clearly there was a culture in Hawaii and a different culture on board the British ship, and clearly the interaction of these two cultures produced events that had a profound impact on both of these cultures.

2. Implications for 'Organizational' Culture Analysis

The major lesson of Sahlins's analysis is that *when we have access to historical data we should use it*. Organizations have defined histories. Therefore, when we analyze organizational cultures we should reconstruct their histories, find out about their founders and early leaders, look for the critical defining events in their evolution as organizations, and be confident that when we have done this we can indeed describe sets of shared assumptions that derive from common experiences of success and/or shared traumas. And we can legitimately think of these sets of assumptions as 'the culture' at a given time. That description will include subcultures that may be in conflict with each other, and there may be subunits that have not yet had enough shared experience to have formed shared common assumptions. In other words, culture as a state does not have to imply unanimity or absence of conflict. There can be some very strongly shared assumptions, and large areas of conflict and/or ambiguity (Martin 1992) within a given cultural 'state.'

At the same time, we can study the day-to-day interactions of the members of an organization with each other and with members of other organizations to determine how given cultural assumptions are reinforced and confirmed, or challenged and disconfirmed. We can analyze the impact of these perceptual interactive events in order to understand how cultures evolve and change. This process could be especially productive in mergers, acquisitions, and joint ventures of various sorts. Whether one chooses to focus one's cultural research on building typologies of cultural 'states,' categories that freeze a given organization at a given point in time, or on analyzing the moment to moment interactions in which members of a given social system attempt to make sense of their experience and, in that process, reinforce and evolve cultural elements, becomes a matter of choice. Both are valid methodologies and in practice they should probably be combined.

3. A Formal Definition of Culture

Culture as a property or process of any group can now be defined as: a pattern of shared basic assumptions that the group has learned as it solved its problems of external adaptation and internal integration, that has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems (Schein 1992).

Once these patterns have been learned they function to reduce anxiety and provide moment-to-moment meaning and predictability to daily events. Or, as Trice and Beyer (1993) put it: 'Human cultures emerge from people's struggles to manage uncertainties and create some degree of order in social life' (p. 1). Note that patterns of overt *behavior* are not part of the formal definition because such patterns can arise through other causes such as common instinctive patterns (e.g., ducking when we hear a loud noise) or by common reactions to a common stimulus (e.g., everyone running in the same direction to avoid some threat). If one understands the shared basic assumptions, one can determine which regularities of behavior are 'cultural' and which ones are not. One cannot, however, infer the assumptions just by observing behavior.

4. A Conceptual Model for Analyzing Culture

Whether one chooses to analyze culture as a state or as a process, it is helpful to differentiate the levels at which culture as a shared phenomenon manifests itself. Any group, organization, or larger social system can be analyzed in terms of (i) its visible and feelable 'artifacts,' (ii) its espoused beliefs and values; and/or (iii) its less visible, taken for granted shared basic assumptions (see Fig. 1).

This model is useful in two ways: (i) it helps a novice observer of organizational culture to differentiate the superficial manifestations and espoused values that most organizations display from the cultural substrate or essence, the tacit, shared assumptions that drive the day-to-day behavior that one observes; and (ii) the differentiation of levels is a necessary conceptual tool in helping an organization to decipher its own culture.

To further understand these distinctions, especially the distinction between espoused values and shared tacit assumptions, it is necessary to take a historical evolutionary point of view toward culture formation. If we take a typical business organization, the process starts with one or more entrepreneurs who found a company based on some personal beliefs and values. As they hire others to work with them they will either choose those others on the basis of their compatibility or will socialize newcomers to the beliefs and values that they regard as core to running the new business. The founders' beliefs and values will cause the organization to make decisions in its environment, and if those decisions are successful, the newcomers will begin to entertain collectively the idea that the founder's beliefs and values must be 'correct.' As the environment continues to reinforce the behavior of the organization, what was originally the founder's *personal* beliefs and values gradually come to be *shared* and imposed on new members as the organization grows. If this success cycle continues, these shared beliefs and values gradually come to be taken for granted, drop out of awareness, and can therefore be thought of as 'taken for granted assumptions' that become increasingly non-negotiable.

This process of culture formation will occur not only in reference to the organization's primary task in

ARTIFACTS

THE VISIBLE, HEARABLE, FEELABLE MANIFESTATIONS OF THE UNDERLYING ASSUMPTIONS, e.g., BEHAVIOR PATTERNS, RITUALS, PHYSICAL ENVIRONMENT, DRESS CODES, STORIES, MYTHS, PRODUCTS, etc.

SHARED VALUES

THE ESPOUSED REASONS FOR WHY THINGS SHOULD BE AS THEY ARE, e.g., CHARTERS, GOAL STATEMENTS, NORMS, CODES OF ETHICS, COMPANY VALUE STATEMENTS

SHARED BASIC ASSUMPTIONS

THE INVISIBLE BUT SURFACEABLE REASONS WHY GROUP MEMBERS PERCEIVE, THINK, AND FEEL THE WAY THEY DO ABOUT EXTERNAL SURVIVAL AND INTERNAL INTEGRATION ISSUES, e.g., ASSUMPTIONS ABOUT MISSION, MEANS, RELATIONSHIPS, REALITY, TIME, SPACE, etc.

Figure 1
The levels of culture

its various external environments but also with respect to its internal organization as well. Those founder beliefs and values, which make life livable and reduce interpersonal anxiety, will gradually come to be taken for granted and come to be tacit assumptions about the 'correct' way to organize. But in both domains the driving force is what works. Culture is the *result* of successful action. If things do not work out the group will disappear.

A basic need in human groups is to justify what they do. And, paradoxically, the justifications often are not the same as the tacit assumptions that actually determine the behavioral regularities. Thus groups create ideologies, aspirations, visions, and various other kinds of 'espoused values' which may or may not correspond isomorphically with the tacit assumptions. The most difficult aspect of deciphering organizational cultures, then, is to determine to what extent the claimed espoused values actually correspond to the behavioral regularities observed and, if not, to determine what the shared tacit assumptions are.

5. *How to Describe Cultures*

There are three approaches to describing cultures: (i) profiling them on various preselected dimensions; (ii) creating conceptual typologies into which to fit given cultures; and/or (iii) clinical or ethnographic descriptions that highlight unique aspects of a given culture.

Among the profiling approaches, one of the most widely used is Hofstede's four dimensions based on factor analyzing questionnaire responses—Individualism (vs. Groupism), Masculinity (the perceived gap between male and female roles), Tolerance for Ambiguity, and Power Distance (the perceived distance between the most and least powerful in the society). These dimensions used originally to describe national cultures but have also been extended to descriptions of organizational cultures (Hofstede 1980, 1991).

A different approach has been to use sociological dimensions developed by Parsons (1951) and elaborated by Kluckhohn and Strodtbeck (1961) as a way of creating profiles of a given organization and identifying critical value issues—universalism vs. particularism, individualism vs. collectivism, affectively charged vs. neutral, specific vs. diffuse, time orientation, and degree of control over nature (Hampden-Turner and Trompenars 1993, Trompenars 1993). A third version of this approach is to take one or two dimensions, consider them to be conceptually central, and create a typology based on a two-by-two table with those dimensions. Cameron and Quinn (1999) use (a) Flexibility vs. Stability and Control and (b) Internal Integration vs. External Differentiation to create four types of cultures—Clan, Adhocracy, Hierarchy, and Market. Goffee and Jones (1998) use the dimensions of Degree of Solidarity and Degree of Sociability to create four types—Networked, Communal, Fragmented, and Mercenary. Among the typologies we have Likert's Systems 1 to 4, Harrison

and later Handy's 'Gods of Management,' McGregor's Theory X and Theory Y, Ouchi's Theory Z, and the more popular distinctions between 'command and control' vs. 'committed and empowered' (Likert 1967, Handy 1978, McGregor 1960, Ouchi 1981).

The clinical or ethnographic approach attempts to deal with any given culture as a unique pattern of shared assumptions around issues that are not known before the clinician/researcher is on the scene as an observer and active inquirer (Schein 1987, 1992, 1999). The prime objection to questionnaires as research tools for the study of *culture* is that they force us to cast the theoretical net too narrowly. The advantage of the ethnographic or clinical research method is that we can train ourselves consciously to minimize the impact of our own models and to maximize staying open to new experiences and concepts we may encounter. In the end we may well sort those experiences into the existing categories we already hold. But at least we will have given ourselves the opportunity to discover new dimensions and, more importantly, will have a better sense of the relative salience and importance of certain dimensions within the culture we are studying. The issue of salience is very important because not all the elements of a culture are equally potent in the degree to which they determine behavior. The more open group oriented inquiry not only reveals how the group views the elements of the culture, but, more importantly, tells us immediately which things are more salient and, therefore, more important as determinants.

As to the categories themselves I have found it empirically useful to start with a broad list of 'survival functions'—what any group must do to survive in its various environments and fulfill its primary task, and 'internal integration functions'—what any group must do to maintain itself as a functioning system. This distinction is entirely consistent with a long tradition of empirical research in group dynamics that always turns up two critical factors in what groups do—(i) task functions; and (ii) group building and maintenance functions. Ancona and others have pointed out that we must add a third set to these two—boundary maintenance functions (Ancona 1988). Task and boundary maintenance functions are external survival issues, and group building and maintenance functions are internal integration issues. We may then construct different lists of what specific dimensions of behavior, attitude, and belief we will look for in each domain, but at least we have a model that forces us to cast the net widely, and a reminder that culture is, for the group, the learned solution to all of its external and internal problems.

If we then look a little deeper, drawing again on anthropology and sociology, we find broad cultural variations around deeper, more abstract issues such as those developed by Kluckhohn and Strodtbeck. Also useful is the work of England (1975) on managerial values which deals specifically with how in a given culture one arrives at 'truth.' If one combines these dimensions with some of the dimensions identified by Edward Hall (1966, 1976, 1983) on concepts of space and spatial relationships, and with more recent concepts about the nature of the 'self' in different cultures, one has a pretty good template of what culture covers at this deeper level.

6. *Does Corporate Culture Matter?*

Several claims based on various different kinds of research have tried to show a connection between the strength and/or type of culture and economic performance (Deal and Kennedy, 1982, Denison 1990, Kotter and Heskett 1992, Collins and Porras, 1994). The problem is that different kinds of cultural dimensions relate to different kinds of environments in ways that are not entirely predictable. The best way to summarize, then, is to say that culture certainly influences economic performance, but the manner in which this occurs remains highly variable. A culture that can be very functional in one environment or at one stage in a company's evolution can become dysfunctional and cause that same company to fail. Culture needs to be analyzed and understood, but until much more research has been done, one cannot make generalizations about its impact.

See also: Bureaucracy and Bureaucratization; Bureaucracy, Sociology of; Corporate Governance; Entrepreneurship; Ideology, Sociology of; Leadership in Organizations, Psychology of; Management: General; Myths and Symbols: Organizational; Organizational Climate; Organizational Culture; Organizational Culture, Anthropology of; Organizations, Sociology of; Population Ecology

Bibliography

Ancona D G 1988 Groups in organizations: Extending laboratory models. In: Hendrick C (ed.) *Annual Review of Personality and Social Psychology: Group and Intergroup Processes*. Sage, Beverly Hills, CA

Cameron K S, Quinn R E 1999 *Diagnosing and Changing Organizational Culture*. Addison-Wesley, Reading, MA

Dennison D R 1990 *Corporate Culture and Organizational Effectiveness*. Wiley, New York

England G 1975 *The Manager and His Values*. Ballinger, New York

Frost P J, Moore L F, Louis M R, Lundberg C C, Martin J (eds.) 1985 *Organizational Culture*. Sage, Thousand Oaks, CA

Goffee R, Jones G 1998 *The Character of a Corporation*. Harper, New York

Hall E T 1966 *The Hidden Dimension*. Doubleday, New York

Hall E T 1976 *Beyond Culture*. Doubleday, New York

Hall E T 1983 *The Dance of Life*. Doubleday, New York

Hampden-Turner C, Trompenars A 1993 *The Seven Cultures of Capitalism*. Doubleday-Currency, New York

Handy C 1978 *Gods of Management*. Pan Books, London

Hofstede G 1980 *Culture's Consequences*. Sage, Beverly Hills, CA

Hofstede G 1991 *Cultures and Organizations*. McGraw-Hill, London
Kluckhohn F R, Strodtbeck F L 1961 *Variations in Value Orientations*. Harper, New York
Kotter J P, Heskett J L 1992 *Corporate Culture and Performance*. Free Press, New York
Likert R 1967 *The Human Organization*. McGraw-Hill, New York
Martin J 1992 *Cultures in Organizations*. Oxford University Press, New York
McGregor D 1960 *The Human Side of Enterprise*. McGraw-Hill, New York
Ouchi W G 1981 *Theory Z*. Addison-Wesley, Reading, MA
Parson T 1951 *The Social System*. Free Press, Glencoe, IL
Sahlins M 1985 *Islands of History*. University of Chicago Press, Chicago, IL
Schein E H 1987 *The Clinical Perspective in Fieldwork*. Sage, Newbury Park, CA
Schein E H 1992 *Organizational Culture and Leadership* (2nd edn.). Jossey-Bass, San Francisco, CA
Schein E H 1999 *The Corporate Culture Survival Guide*. Jossey-Bass, San Francisco, CA
Trice H M, Beyer J M 1993 *The Cultures of Work Organizations*. Prentice-Hall, Englewood Cliffs, NJ
Trompenars A 1993 *Riding the Waves of Change*. Economist Books, London
Weick K E 1995 *Sensemaking in Organizations*. Sage, Beverly Hills, CA

E. H. Schein

Corporate Finance: Financial Control

Corporate finance is concerned with financial decisions made by corporations. These decisions can be grouped easily into two major categories: investment decisions and financing decisions.

Regarding the first major decisions a firm must take, each company needs real assets (tangible and intangible) to carry on its business and keep its operations going. How much a firm should invest and what specific assets a firm should invest in are the typical questions related to investment decisions. Investment decisions, then, determine the total amount of assets held by a company and the composition of these assets. The economic evaluation of investment projects aims at orienting corporations' choices connected to their asset allocation process. This process that leads a company to allocate its capital to investment proposals whose benefits are to be realized in the future is generally referred to as *capital budgeting*.

All the assets we mentioned above are to be paid for. To raise the money necessary to their investments policy, corporations issue securities on financial markets. These financial assets have value since they represent claims on companies' real assets. Securities issued by corporations include not only shares of stock but also bonds, bank loans, lease obligations, and any other form of financing. How the cash required for investments should be obtained is the answer that financing decisions must give. Financing decision, in fact, are concerned with determining the best mix of financial sources in the form of equity and debts. In other words, financing decisions define the total amount and the composition of a company's capital structure. The definition of capital structure referred to is that which takes into account the explicitly costly forms of financing raised from the company's investors. For sake of simplicity, they are divided into financial debts and equity without dwelling on the different form of financing.

According to what has been highlighted about corporate finance decisions, financial management attempts to find specific answers to needs arising from both investment and financing areas.

After defining what corporate finance decisions are concerned with, the next step in this analysis will be to reread investment policies and capital structure decisions in the light of the theory of shareholders' value. Since the principle of maximizing shareholders' wealth largely is accepted as the rational guide for running a business and for the efficient allocation in an economy, it is assumed to be the objective in considering how financial decisions should be made.

Investment decisions and financing decisions must contribute together to create value for the company's shareholders. From one hand, a financial manager that acts in its shareholders' interest should invest in those projects that increase the overall firm's value and, then, its shares' value. From the other hand, the mix of financial resources that maximizes the company's value should be adopted. From this point of view, it becomes crucial to understand how corporate borrowing affects its shareholders' wealth according to Modigliani and Miller proposition II. In their article, Modigliani and Miller (1963) state that, in a world with taxes, the value of a levered firm increases in proportion to the debt–equity ratio (Debt/Equity). As we will show in Sect. 6 this is due to the fact that financial charges are tax deductible whereas dividends and retained earnings are not; an increase in debts will reduce the company's cost of capital and raise the firms' value.

In addition to that, it must be kept in mind that investment decisions and financing decisions interact, sometimes producing different consequences on the company's value. In fact, any change in the firm's capital structure influences the cost of capital that a company applies to discount investments cash flows. Consequently, any rise in the cost of capital could make any investment less profitable, reducing its contribution to the overall company's value.

From what has been said, it can be concluded that both investment and financial decisions have to be analyzed in relation to their impact on the firm's major objective: to make the firm's shareholders as well off as

possible. As a result, an optimal combination of the two areas of decision will maximize the value of the firm for its shareholders.

Consistent with that, the financial manager must act as an intermediary between the firm's operations and capital markets where the firm's securities are traded. The purpose of capital markets is allocating savings efficiently in an economy, from ultimate savers to ultimate users of funds who invest in real assets (companies). In fact, companies issue securities and raise cash by selling those financial assets to investors. Then the cash is invested in the firm's operations and used to purchase real assets. Later, if the firm does well, the real assets generate cash inflows that more than repay the initial investment. Finally, once the firm's operations generate new cash, this is either reinvested in the company or returned to investors who hold the securities originally issued. The last decision is not completely free for the financial manager, but it depends on several factors such as the characteristics of the specific financing, the dividend policy, the financial needs of the company.

1. Corporate Finance Decisions and Financial Control

In order to make optimal financial decisions in keeping with the major objective of the firm (i.e., maximizing its shareholders' wealth), the financial manager must implement inside the firm an effective process of analysis, planning, and control and make use of some analytical tools. In fact, financial analyses are a necessary condition for making appropriate financial decisions in the meaning as stated before. From a very simplistic point of view, in order to create value for all the company's investors (debtholders and stakeholders), the asset allocation process must yield more than the company's cost of capital.

Since the financial manager has to grant that the cash flow ability of the firm will properly remunerate its financial sources over the long run, they must evaluate the size and timing of cash flows and monitor the company's capital structure, its major sources and uses of funds, and its profitability over time. As a measure of the company's profitability ROI (return on investments = operating income/capital invested) is taken into consideration. According to that, the financial manager is concerned on one hand with the profitability of the company's investments and with the efficiency of the implemented asset management decisions. It means that they have to make use of rational tools to evaluate capital budgeting decisions. On the other hand, the financial manager is interested in keeping constantly under control the changes in the company's capital structure that can affect both the firm's cost of capital and the yield required by capital suppliers. If the company's cost of capital is defined as the weighted average cost of capital (WACC) and represents the net cost of debts and equity weighted by their percentage on the total capital structure, an increase in the percentage of debts, for instance, can reduce the WACC since they are less costly than equity, due to their fiscal deductibility; at the same time, an excessive rise in debts can make the company be perceived as a riskier investment and thus induce investors to require higher yields on debt and equity. Both these effects have to be taken into account once the financial manager is approaching any financing decision.

In the light of what financial managers must do to maximize shareholders' value, *financial control* can be defined as the whole of the activities and tools that enable financial managers to carry out financial analyses for the purposes of internal and external control. By internal control, we mean evaluating the investments profitability and the efficiency/effectiveness of the capital budgeting process. In addition to that, a form of internal control is considered as any analysis turned to keep under observation the composition of the company's capital structure. External control is any aspect of financial analysis that outside suppliers of capital take into consideration in evaluating the firm and that the financial manager should be interested in the endeavor of bargaining more effectively for outside funds.

So far it has been stated that the main corporate goal is to increase the shareholders' wealth. According to that, a company creates value if its shareholders' wealth rises. It is worth mentioning that, recently, financial markets and investors have taken into account new parameters in the evaluation of financial investments. These parameters add a social and ethical appraisal of financial investments to a pure financial valuation. Such parameters aim at reminding companies of their social responsibilities and spur them to give constructive contributions to society and bring positive energy to the solutions of its problems.

As a consequence of the new approach to interpreting the corporate role within the society, investors choose to invest their money only in those companies that behave ethically. On the other hand, financial products suppliers are tailoring new financial products which are labeled as ethical, and that respond to the new evaluation requirements.

The Dow Jones Sustainability Index (DJSGI) can be seen as an example of this new trend inside financial markets. That index, in fact, consists of more than 200 companies that represent the top 10 percent of leading sustainability companies in 64 industry groups in the 33 countries covered by the DJSGI.

The companies included in the DJSGI have shown over time a strong commitment on the achievement of the five corporate sustainability performance principles: innovation, high standard corporate governance, shareholders' requirements, best practices, and social behaviors.

From what has just been highlighted, managers must be aware that new principles beyond those suggested by financial theories have to be pursued in running their business.

2. An Integrated System of Financial Analyses

An effective financial analysis, planning, and control process is critical to enhance enterprise value and, then, shareholders' wealth. To make rational decisions in the shareholders' interest, financial managers need to be supported by an integrated system of financial analyses and reports. These tools enable financial managers constantly to control both the profitability of the company (in the meaning of the firm's ability to generate cash flows from its operating activities) and the cost of resources applied to undertake operating investments that determine the company's cost of capital.

Since they need to keep under control both the profitability of capital invested and the cost of resources raised from investors, financial managers make use of an integrated system of analyses that includes financial statements, ratios, financial planning, investment valuation, and capital structure analysis. Each of these tools plays a key role with reference to the financial control activity, providing the financial manager with a wide range of information that helps in orienting the company's investment and financing decisions. Even if financial analysis, planning, and control activities are often considered as different tools, they have to be viewed as a whole since they together aim at accomplishing the maximization of shareholders' value.

In large companies it often happens that different managers carry out the above-mentioned analyses in different departments. It becomes important, then, to construct an integrated system of data exchange that allows the different tools to work jointly. To assess an investment's profitability, for instance, it is necessary to forecast the company's market share in order to determine its potential incremental revenues and cash flows. In addition to that, the composition of the expected capital structure will be also needed to calculate the new cost of capital and discount future cash flows at that rate.

3. Financial Accounting

Financial accounting (financial statements, ratios, and cash flow statements) gives the financial manager a picture of the firm's past performance and its current standing. These analyses help the financial manager to understand the company's current economic and financial situation and make forecasts for the future. It also allows the manager to represent the firm's position according to its competitors and its evolution over time.

The reorganization of financial statements (assets and liabilities statement and income statement) according to some predefined schemes is the first step of a process that, starting from accounting information, ends up with the appraisal of a company's performance from both income and cash flow perspective.

One of the criteria applied to reorganize the company's financial statements is, for instance, the activity-related method. With relation to this approach, we group the balance sheet entries in current and noncurrent assets and liabilities according to their pertinence or otherwise to the operating cycle of the company. The same criterion applied to the items contained in the income statement allows us to separate revenues and costs in relation with the four activity areas that can be identified in an ongoing business: current, financial, subsidiary, and extraordinary activities.

Once financial statements have been reorganized, it is possible for the financial manager to combine income statement and balance sheet entries in order to obtain significative ratios. Second, from reorganized financial statements, a cash flow statement can be drawn up as well.

In particular, a complete system of ratios describes the current state of health of the company according to different points of view: liquidity, profitability, leverage. This analysis offers a static diagnosis of the firm that can be helpful for managerial control and for a better understanding of what outside suppliers of capital may expect in terms of a company's financial condition and performance. In fact, both the financial manager and the company's investors are interested in measuring the return on investments of the company, its capability of repaying debts in the short run and in the long run, and the relation between debt and equity. Consistent to that aim, ratios are mostly employed for *ex-post* analysis, to verify the effect of investment and financing decisions on the company's profitability, and for the composition of its assets and liabilities.

The final step in the above-mentioned process of financial analysis is represented by the construction of the cash flow statement. Cash flow analysis can be seen as a dynamic analysis that highlights the company's ability of generating cash flow. Using cash flow, the financial manager identifies the sources and uses of monetary resources during a given period. The cash flow statement groups all the inflows and outflows according to their pertinence to the current activity, the financing and reimbursement activity, the investment activity (operating and nonoperating), and to the remuneration of debt holders and stakeholders. In other words, this tool leads to the quantification of the cash inflows that the company is currently able to generate from its operating activity, and the comparison of them to the flows absorbed by the in-

vestment policy, the debt repayment, and the dividend policy. In fact, it is one thing if a company can provide the monetary resources autonomously to undertake investments and adequately remunerate the capital invested, while it is quite another if the same company regularly needs to find outside sources of capital in order to avoid financial distress caused by an insufficient production of cash inflow.

Consistent with that, the cash flow analysis can be interpreted as an *ex-post* analysis that helps the financial manager to appraise the impact of investment and financing decisions in terms of financial flows.

4. Financial Planning

Financial planning enables financial managers to assess the effects of financing and investment decisions *ex-ante*. Financial planning is a necessary analysis as investment and financing decisions interact and they should not be made independently. Financial planning is concerned with analyzing financial flows of the firm as a whole, forecasting the consequences of investment and financing decisions, and weighing the effects of different alternatives. In fact, any financial decision that affects the investment policy of the firm or its mix of financial sources raised from investors influences the company's future profitability as well as its capital structure (and its cost of capital). The financial manager has to define both a target financing mix and a target return on investments and undertake only the investments/financings consistent with those targets. In order to keep under control any unexpected change that can modify the company's return on investments or its cost of capital, the financial manager has to analyze any financial strategy very carefully and to take into consideration any economic and financial impact that it may have. Moreover, financial planning makes possible the comparison of different investment and financing strategies in order to evaluate which one maximizes shareholders' wealth.

The tangible tool applied in the planning process is the financial plan describing the firm's financial strategy and projecting its future consequences in a proforma balance sheet, income statement, and cash flow statement. In particular, the forecasted cash flow statement provides an effective method for the financial manager to assess the growth of the company and its resulting financial needs as well as to determine the best alternative to finance those needs.

In fact, the plan sets financial goals and it represents a benchmark itself for evaluating subsequent performance. Helping the company to achieve these goals is one of the chief responsibilities of financial planning.

Finally, the planning horizon depends on the company and its business. High-tech companies, for instance, produce much shorter financial plans than most companies do as a consequence of technologies life cycle, whereas public utilities and energy companies make longer financial plans due to the long lead time of their capital projects.

5. Investments Valuation

Capital budgeting involves the entire process of planning capital expenditures whose returns are normally expected to extend beyond one year. Since we stated that investment decisions must be made so that they maximize shareholders' value, capital budgeting decisions forcedly must be related to the firm's overall strategic planning. Capital budgeting must be integrated with strategic planning as excessive investments or inadequate investments could cause serious consequences for the future of the firm. If the company invested too much in fixed assets overestimating its potential growth, it would incur unnecessarily heavy expenses that would reduce its return on investments. On the other hand, if a company has not invested enough in new production capacity it may lose a portion of its customers to rival firms. A related problem is how to properly phase the availability of capital assets in order to have them 'working' at the correct time. An effective process of capital budgeting aims at improving the timing of asset acquisitions and the quality of assets purchased. Another reason that highlights the relevance of the capital budgeting process is that asset expansion typically involves substantial expenditures. Before a company undertakes an expensive investment such as the purchase of a new plant, the launch of a new product, or the signature of a commercial agreement, it needs to plan how to raise funds required for that specific project. A number of factors make capital budgeting one of the major financial management decisions. In fact, not only does any investment decision contribute to determine the company's profitability (and its ROI) and, then, the shareholders' value, but also the capital budgeting process vitally affects all departments of a firm (such as production, marketing, finance, to mention a few). Since financial managers need to keep under control the asset allocation process so as to grant the maximum value to their shareholders, it can be said confidently that one of the most important corporate finance functions is the economic valuation of investment projects. In fact, the first step in undertaking any allocation project is the analysis of its economic and financial profile. The latter aspect is taken into account in drawing up the financial plan, where the financial manager verifies the compatibility of the cash flows related to the project under consideration with those that happen at the overall company's level.

The financial manager must carry out those projects that maximize the incremental value for the company's shareholders. In order to increase shareholders' wealth, a project must yield more than the cost of

funds used to undertake the capital expenditure. It is, then, possible to conclude that the financial manager must judge a certain investment project according to its expected return. If this is higher than the company's cost of capital, the specific investment will create value for the company's shareholders.

As the asset allocation process is led on a continual basis, the financial manager must measure constantly the profitability of the firm's investment activity. Such a form of control makes use of appropriate methods of analysis like the NPV and the IRR. The NPV (net present value) is the present value of the cash lows stemming from the investment project; the IRR (internal rate of return) is that rate that makes the NPV equal to zero and it can be seen as the gross return of an investment project.

The implementation of the above-mentioned tools of financial control on investments prevents the funds being misallocated. In fact, when many alternatives of investment are evaluated, the most profitable will be chosen first, then the others one by one according to their expected rate of return. This *modus operandi* becomes a necessary approach if the company has a limited amount of funds available to finance investment projects (*capital rationing*).

6. *Capital Structure Analysis*

Capital structure refers to how the firm's assets are financed. If the monetary resources that the current activity is able to produce are not enough to offset the cash-outs due to the company's investment activity, the firm needs to raise new funds from investors. Accordingly, capital structure is concerned with explicitly costly forms of financing just as financial debts (bonds, loans, overdrafts) and equity are. In fact, both debt holders and shareholders want to be remunerated for the money they invested into the company. It is then possible to determine the cost of capital as the average of the net cost of debt and equity (WACC). To calculate the average, the impact of debts and equity on the total capital structure has to be taken into account.

The composition of capital structure influences the company's cost of capital from two different points of view. First, any change in the amount of debts and equity will make their weights vary so that we come up with a new WACC, other things being equal. Second, the cost of capital will differently change according to which form of financing is chosen. Since financial charges are tax deductible whereas dividends and retained earnings are not, an increase in debts will reduce the WACC. However, it must not be concluded that a company should be 100 percent debt-financed. In fact, the more debts rise the higher the risk of financial distress will be. This higher risk perceived by investors leads debt holders and stakeholder to ask for being granted a higher rate of return.

The composition of capital structure impacts on the shareholders' value since it influences the cost of capital invested in the company. If the return on investments remains equal, a rise in the cost of capital will cause a decrease in the company's profitability and in shareholders' wealth. The financial manager has to take into consideration the effect on the capital structure when any financing decision is evaluated. Once a financial need arises from the planning activity, the financial manager should simulate what impact a debt or equity issue may have on the overall company. This activity enables the financial manager to quantify the effects of financing alternatives in terms of the firm's cost of capital and financial riskiness.

We are assuming that there are no capital constraints that affect the company so that the firm can count on an unlimited access to both the equity market and the debt market. According to this hypothesis, the financial manager can choose the financing alternative that maximizes the shareholders' value and minimizes the cost of capital. Sometimes, however, the firm has to face with a limited availability of financing alternatives to raise money from investors. In fact, many factors such as capital markets features or aspects related to companies' ownership and control could make some financing decisions unfeasible. At the same time, other aspects such as corporate taxation can make some sources of financing preferable in comparison to others.

See also: Commercial Banking; Corporate Governance; Finance, Geography of; Financial Accounting and Auditing; Financial Institutions in Economic Development; Organizational Decision Making

Bibliography

Bernstein L A 1978 *The Analysis of Financial Statements*. Dow Jones-Irwin, Homewood, IL

Brealey R, Myers S *Principles of Corporate Finance*, 5th edn

Broke M Z 1990 *Handbook of International Financial Management*. Macmillan

Cheng F L 1985 *Financial Analysis and Planning*. Addison-Wesley, Reading, MA

Dallocchio M 1996 *Corporate Finance. Managing Corporate Decisions*. Egea, Milan, Italy

Modigliani F, Miller M H 1958 The cost of capital, corporate finance and the theory of investment. *American Economic Review* **48**(3): 261–97

Modigliani F, Miller M H 1963 Corporate income taxes and the cost of capital: A correction. *American Economic Review* **53**(3): 433–43

Ross S A, Westerfield R W, Jaffe J 1996 *Corporate Finance*. Irwin

Solomon E 1963 *The Theory of Financial Management*. Columbia University Press, New York

Van Horne J C 1989 *Financial Management and Policy*. Prentice Hall, Englewood Cliffs, NJ

Weston J F, Thomas T E *Managerial Finance*, 9th edn

Weston J F, Thomas T E 1992 *Financial Theory and Corporate Policy*. Addison-Wesley, Reading, MA

M. Renzetti

Corporate Governance

The term 'corporate governance' refers to the legal rules, institutional arrangements, and practices that determine who controls business corporations, and who gets the benefits that flow from them. Corporate governance issues include how major policy decisions are made in business corporations, how various stakeholders can influence the process, who is held accountable for performance, and what performance standards are applied.

1. *The Governance Problem*

Governance problems arise because corporations are fictional entities, with a legal status that is separate from any of the individuals involved in them. They are not, themselves, actual persons, yet numerous different groups of individuals—investors, employees, managers, suppliers, customers—may have interests at stake in them. Often, they are managed by hired executives. Hence, decision-making authority in corporations generally is separated from personal responsibility and liability. Neither managers (the active decision makers), nor other employees, nor investors (who enjoy the protection of 'limited liability') necessarily bear the full costs of corporate actions.

Hence, the governance problem to be solved in setting up any corporation is to create a mechanism for selecting and overseeing the firm's managers that fosters cooperative behavior among multiple participants, protects the investments made by them, discourages abuses by decision makers, and still provides sufficient freedom of action to encourage innovation and risk taking. Such a mechanism should allocate decision and control rights to parties who have the appropriate incentives and the necessary information to make wealth-enhancing decisions for the firm, while at the same time ensuring that the controlling parties are accountable to all of the other participants who are exposed to risk by the actions of the corporation. Most governance arrangements, of course, do not achieve this ideal, and in practice, governance arrangements are responsive to a variety of political, as well as economic, pressures.

Legal rules that influence corporate governance vary from country to country, but generally they include corporate and securities law, as well as rules from contract, bankruptcy, and antitrust areas of law. Other corporate governance practices may be driven by market pressures, or tradition, or by the structure of equity ownership. Observed governance arrangements and practices at a given corporation may be prescribed by specific terms of that corporation's charter, or they may be merely customary.

2. *Goals of Corporate Governance*

The phrase corporate governance came into prominent use in the 1980s, and is often used narrowly to refer to the mechanisms and rules that govern relationships among direct corporate participants in publicly-traded firms, especially shareholders, directors, managers, and sometimes employees. But, historically, questions about social control over corporate behavior have been quite important. Since the corporate form first emerged as the dominant way to organize big business enterprises in the second half of the nineteenth century, policy concerns about corporations have, at various times, focused on antitrust, consumer protection, pollution control, worker and/or investor protection, corporate involvement in the political process, and corporate contributions of resources to charitable causes.

Broadly speaking, there are two schools of thought about the appropriate goals of corporate governance institutions and arrangements. The 'shareholder primacy' view, which has been dominant in the US and other English-speaking countries since the 1980s, has focused on the set of governance problems that arise in publicly-traded corporations in which equity shares are held and traded by numerous individuals who have little or no management connection to the firm. This problem has been referred to in the literature as the 'agency cost problem' (Jensen and Meckling 1976) resulting from the 'separation of ownership from control' (Berle and Means 1932).

By contrast, in Europe and Asia (as well as in the US in the middle of the twentieth century), corporations are more likely to be viewed as institutions with a quasipublic character and role. Hence, the proper goal of corporate governance is more likely to be seen as a balancing of interests among all of the corporate stakeholders. This so-called 'stakeholder' view of the firm was widely dismissed by leading legal, economics, and finance scholars in the US in the 1980s and 1990s as not being well grounded in a convincing theoretical model of the firm. However, recent scholarship in the US is providing a stronger theoretical basis for this view.

3. *Recent Developments*

The explosion of interest in corporate governance since 1980 has been driven by a number of developments. The first was the dramatic emergence in the US of unsolicited and unwanted tender offers, which became a common threat to corporate independence in the 1980s. In these so-called 'hostile takeover' transactions, outsiders attempted to buy up enough of the outstanding, publicly-traded shares to gain voting

control over a company's board. Closely related to the rise of hostile takeovers was the development of numerous innovative financial instruments used to finance these transactions.

The threat of hostile takeover became a tool by which investors were able to compel managers to cut staff, close plants, redeploy assets, or take other unpopular actions in order to increase corporate profitability. Initially, corporate directors and managers resisted the takeovers, and invented a variety of new mechanisms to prevent the outsiders (whom they pejoratively dubbed 'raiders') from gaining control. This led to numerous court challenges and decisions, and an intense legal and policy debate about the appropriate response of corporate directors to takeover threats.

The second development was the rapid increase in the power and influence of financial investors in general, and institutional investors in particular. By the 1980s, institutional investors such as pension funds and mutual funds had grown to become the largest shareholders in many US corporations, and they began to use their new clout to influence corporate agendas. Among the changes institutional investors sought were corporate governance rules and executive compensation plans designed to make corporate executives more sensitive to shareholders' interests, especially their interest in higher share prices.

The third development was the shifting patterns of corporate performance worldwide. In the 1980s, many US firms lost substantial market share to international competitors, especially in Europe and Japan. In response, scholars and policy analysts asked whether Japanese and European corporate governance systems were somehow better than the US system. The situation reversed in the 1990s. Japanese and European economies slumped, while the US economy, led by a very dynamic high-tech sector, put on one of its best postwar performances. Scholars and policymakers then began asking what aspect of the US system makes it so much better at fostering and financing business innovation. Toward the end of the 1990s, financial crises in numerous countries, especially the formerly high-growth countries of southeast Asia, again forced policymakers and academics to rethink their beliefs about how corporate governance systems work, and what makes one system work better than another.

The fourth development was the collapse of the former Soviet Empire, and the effort by nearly all communist and socialist countries to convert to capitalism. In the early 1990s, Western advisors rushed into Eastern Europe and Asia to help the governments of the transition countries develop corporate laws and securities markets so that they could convert state-owned enterprises to Western-style capitalist corporations. The effort to create new corporate governance structures and practices in transition economies encouraged these policy analysts and scholars to think more deeply about how the corporate governance systems in advanced economies actually work, and what other legal and institutional infrastructures are needed to support them.

The fifth development has been the continued globalization of capital markets and, closely related, the surge in cross-border mergers and acquisitions, both of which create pressures for harmonization of corporate governance standards across countries.

4. *Mechanisms of Governance*

The most important mechanisms of corporate governance in all capitalist economies are boards of directors with fiduciary duties, independent auditors, financial disclosure requirements of securities laws, and the rules determining voting rights. Securities exchanges impose additional disclosure and other governance requirements on firms that list their securities for trading.

4.1 *Boards of Directors*

Although statutory requirements and practices differ from country to country, in every country that has incorporation laws, some provision is made for the establishment of a board of directors at the time the firm comes into being. The board then becomes the locus of legal accountability for the corporation. In the US, state-level incorporation statutes generally are 'enabling' laws, giving founders of corporations wide latitude to establish almost any governance arrangements they want, as long as they create a board, and set up rules by which directors are to be chosen. It is almost always a requirement under state law in the US that directors are elected by shareholders, and a common practice is for directors to be elected for one- to three-year terms by shareholders, on a one-share, one-vote basis.

Directors have the legal responsibility for making virtually all major substantive decisions about the future direction of the company, and for the distribution of corporate income and assets. They decide whether, when, and how much to pay out in dividends; they are responsible for hiring, firing, and compensation of the chief executive officer; and they are responsible for approving the broad strategy, major initiatives, investments, and transactions undertaken by the firm. Other important duties of boards include responsibility for ensuring that adequate books and records are kept, and, in the case of publicly-traded companies, providing for adequate public disclosure and compliance with other laws.

In the US, the authority of duly elected directors over corporate decisions is nearly absolute. They must satisfy the requirements of their corporate charter and bylaws (which are normally quite general and un-

restrictive) and of any external laws and explicit contracts with outside parties. And they must fulfill their fiduciary duties of 'care' and 'loyalty.' The duty of loyalty is taken seriously by the courts in the sense that directors must not engage in self-dealing transactions with the firm (unless such transactions are both substantively and procedurally fair), and must not take advantage of business opportunities that properly belong to the corporation. Beyond that, the courts require little other than that directors must exercise a reasonable amount of diligence in carrying out their duties. If they satisfy these requirements, directors are almost completely protected by the so-called 'business judgment rule,' a common law rule that says that the courts will defer to the 'business judgment' of directors in strategic, policy, or investment decisions made on behalf of the firm.

In the US, anyone can serve on a corporate board, although in recent years, it has become the accepted wisdom that, in publicly-traded firms, a majority of directors should be outsiders, who are not part of the company's management structure. In Europe and Asia, it is common for representatives of large investors, such as wealthy families or large financial institutions such as banks or other companies, to sit on corporate boards. European corporate law also often specifies certain aspects of the size, structure, and composition of corporate boards. In Germany, for example, publicly-traded companies (AGs) and large limited liability companies (GmbHs) are required to have two-tier boards, with up to half of the positions on the upper-tier, or supervisory board reserved for employee representatives. (Most German corporations are also required to have 'works councils,' which provide another formal mechanism for employee voice in some corporate decisions.)

Nominally, corporate law in Japan is similar to that in the US, and in particular, boards of directors in Japanese firms are chosen in a similar fashion and have a similar degree of control and discretion. But as a matter of custom and practice, Japanese directors typically are chosen from among the ranks of senior employees.

4.2 Independent Auditors

One of the most important mechanisms that boards of directors use to ensure that corporations are accountable to their various investors and stakeholders is arranging for audits of records and corporate performance indicators by external, professional auditors. Since the board bears the ultimate responsibility for ensuring the firm's compliance with the law and for the integrity of its public financial disclosures, the board should arrange for and cooperate with auditors to improve the reporting systems (OECD 1999a). In recent years, investors and professional organizations generally have come to agree that the audit function should be performed by individuals who are unaffiliated with the company, and should be under the control of an audit committee with outside directors in the majority. It is a listing requirement at several important stock exchanges (for example, including the New York Stock Exchange) that listing firms have audit committees dominated by outside directors.

4.3 Financial Disclosure

In the US and other common law countries, publicly-traded corporations are required to make regular and frequent public filings of reports providing detailed financial information about the firm. Key parts of this information must be certified by independent auditors, and must conform to standard accounting practices, as determined by independent standard-setters within the accounting profession such as the Financial Accounting Standards Board in the US. Sanctions for officers and directors of companies that report fraudulent information can be severe, and both securities regulators and private investors participate in policing the disclosures. Moreover, insiders are prohibited from trading on information before it is made public. In general, the disclosure requirements are not as strict in other countries and the restrictions against insider trading not as well enforced. An intense debate is underway in the international community about whether these requirements ought to be brought up to the US standards.

Outside the US, the International Accounting Standards Committee and the International Organization of Securities Commissions are working to develop a core set of accounting standards for use in cross-border financings and for listings in global markets.

One of the controversial questions at stake in the policy discussions about accounting standards is the problem of providing better information about 'intangible' assets. Intangibles are assets such as intellectual property rights (patents, copyrights, brand names), reputational capital, organizational capital (including, for example, proprietary software or other management systems), and organizational and intellectual assets, such as the unpatented results of research and development projects or the tacit knowledge and skills of employees. Current accounting rules do not provide a mechanism for such assets to be recorded on the books of firms. Rather, any costs involved in developing these assets are treated for accounting purposes as expenses that, implicitly, do not provide any future return. By some estimates, intangible assets now account for a substantial and growing share of the wealth being generated by the corporate sector, particularly in developed countries. Investors are being provided with very little information about these assets, however, even in countries with relatively transparent accounting systems and strict disclosure requirements.

A related, and equally controversial question at stake has to do with whether corporations should be required to provide public reports on their social and environmental performance. This question links the debate about accounting for intangibles to the debate about shareholder primacy and the social role of corporations.

4.4 Voting Rights

One of the few legal constraints on boards of directors is that the board itself must be elected, and must subject a few important decisions to a vote by affected participants. Generally shareholders must elect directors on a one-share, one-vote basis, for example. In special circumstances, however, other stakeholders may participate in choosing directors. For example, holders of preferred shares or other investors may sometimes have the right to vote for directors if a company fails to pay preferred dividends. Employee representatives to the supervisory boards in Germany are chosen by a process that involves voting both by employees and by their unions. And in a handful of US companies, labor unions have negotiated for the right to choose at least one director.

Shareholders and sometimes other participants often have the right to vote on major corporate transactions, such as a merger or a dissolution. Whatever the formal rules are, however, the voting rules and nomination procedures are easily manipulated so that, in practice, US directors have found they can often get around voting requirements by structuring transactions in different ways. Additionally, control over the nomination process gives existing directors substantial influence over the process of choosing new directors. It is possible for an outsider to wage a campaign to elect an alternate slate of directors, or to amend the charter, or to recommend some other corporate action. Such challenges to boards are called 'proxy fights' because the process requires challengers to solicit other shareholders to gain permission from them to vote their shares by 'proxy.' Proxy fights happen occasionally in the US, but they are costly, and hence, rare.

Nonetheless, shareholders have other informal mechanisms of influence that in most instances are more useful and effective than proxy fights. Financial investors who own large blocks of shares, for example, have the capacity to transfer a controlling block to a raider by selling out, or to dump shares on the market and thereby drive down the price. This inevitably gives large investors considerable informal influence and market control.

4.5 Listing Requirements

Other important mechanisms of corporate governance are the listing requirements for stock exchanges. If a company wants access to public equity markets, it must not only meet disclosure requirements for issuance of securities to the public, it must also make sure that its securities can be traded in organized securities markets. Stock exchanges typically impose governance requirements on firms beyond the requirements of corporate law. As already mentioned, the New York Stock Exchange requires that listed firms have audit committees composed of outside directors.

5. Patterns of Shareholdings

Apart from the formal governance mechanisms, the factor that may have the greatest influence on how decisions get made in publicly-traded corporations, and on the weight given to the interests of various corporate constituencies in those decisions, is the pattern of ownership of equity shares. There is a significant difference between the pattern that prevails in the US and other English-speaking countries, and the pattern that prevails in continental Europe and Asia. In the US, there are thousands of publicly-traded corporations with shares that tend to be very widely held. Typically, no single institutional or individual shareholder controls more than 50 percent of the voting shares, and even 20 percent blocks are relatively uncommon. When no large investor holds a substantial block, it may be hard for shareholders to monitor boards and managers. Moreover, as already noted, it is costly for shareholders to organize themselves in order to wage a proxy fight. These factors tend to make corporate officers and directors quite autonomous and independent. On the other hand, the fact that shares are widely held usually means that there is a liquid market in the shares, and dissatisfied shareholders can easily sell. The market price of the shares provides a constant signal about how well investors think the company is doing.

By contrast, in continental Europe and Japan, there are many fewer publicly-traded companies (even relative to the size of their economies), and companies that are publicly-traded often have very concentrated shareholdings. In German firms, for example, more than half of all publicly-traded companies have a large shareholder with 50 percent or more of the shares. In Japan, groups of industrial companies hold substantial amounts of each others' shares, and in Italy, pyramidal groups headed by families, coalitions, or the state often hold controlling interests in firms. The large shareholders in these firms often have seats on the boards of directors, and inevitably exercise considerable influence on the boards.

Corporate governance scholars have debated the reasons for this pattern. One theory holds that financial institutions in the US were prohibited from becoming large shareholders in industrial companies for political reasons (Roe 1994). The result has been that shares are widely held, and shareholders have

tended to be passive in the US. Managers and directors, in turn, generally operate with much greater impunity than they could if there were more large-block shareholders actively involved in governance. But the fact that shares are widely held and actively traded on deep markets also creates the possibility that outsiders can get control of companies, even against the wishes of boards and managers, by engaging in tender offers. Hence, the 'market for corporate control' may be an effective substitute for direct monitoring.

An alternative theory holds that shareholdings are unconcentrated in the US and other Anglo-Saxon countries because, all else equal, investors would prefer to diversify their risks and make their investments more liquid by holding only small amounts of many different companies in their portfolios. Since their holdings in any one company are typically small (both relative to the size of their portfolios, and relative to the size of the company), investors would prefer to be passive (Black 1990). However, investors could not afford the luxury of being passive if they were not confident that they would be protected from exploitation both by managers and by controlling shareholders. Advocates of this theory note that common law countries have the deepest, most transparent financial markets and the best protections for minority shareholders in the form of strict securities laws. Consequently, these same countries tend to have widely-dispersed shareholders. By this theory, shareholding is more concentrated in European and Asian countries because those countries do not provide sufficient protection for minority shareholders to encourage small investors to hold equities (Coffee 1999).

6. *Supporting Institutions*

These arguments suggest an overarching issue, which is that corporate governance arrangements must be seen as imbedded in a larger set of social and political institutions. In the early 1990s, many of the Western scholars and policy analysts who advised transition economy governments about transforming their economies into capitalist-style free market economies recommended 'shock therapy' approaches. These advisers believed that if productive enterprises could be taken out of the control of governments and placed under the control of private investors, and if subsidies and cross-subsidies could be ended, market pressures would come into play and quickly drive private actors to reorganize themselves into much more efficient configurations. In fact, this had rather conspicuously failed to happen in Russia by the end of the 1990s, and the results were spotty, though generally less disastrous, in other Eastern European countries. As a result, scholars are now turning their attention to the important role played by supporting institutions such as the development of accounting standards; independent and professional accounting firms; court systems that can adjudicate complex contracts and enforce fiduciary duties; professional norms and standards; well-regulated financial institutions; and securities laws with effective systems of enforcement. Empirical studies have shown that in countries without these institutional supports, the corporate sector tends to remain small, and the corporations that do exist tend to be closely held and controlled by families or other networks of economic elites.

7. *Which Way Works Best?*

While the corporate governance systems that have emerged in modern developed economies are obviously superior to the systems in developing and transition countries, scholars are still debating the relative merits of the various systems in developed countries. Is the common law system in the US and other English-speaking countries, with its scattered shareholding, deep and transparent financial markets, and relatively autonomous boards and managers, better in any sense than the large-shareholder dominated, civil code systems of continental Europe and Asia? Each may be an internally consistent system with its own advantages and disadvantages. Where supporting institutions, such as an independent accounting profession and courts that can adjudicate complex contracts and enforce securities disclosure rules, are weak, it may be impossible to provide adequate protection for minority investors. In such circumstances, outside equity investors may only be willing to put capital into firms in which they can get seats on the board and exercise some control. On the other hand, if the legal and social system can effectively limit or discourage self-interested transactions by insiders, then the advantages of widespread shareholding and fluid capital markets can be realized.

But there are tradeoffs. Civil code systems may be more stable in the short run because it is harder in these systems for financial investors to move capital out of companies in search of higher returns. For the same reason, it may be easier to impose 'social' regulations and requirements (to protect jobs, for instance, or to clean up the environment) on companies in civil code countries.

In contrast, capital is extremely mobile in common law countries. In the US in the 1980s and 1990s, the mobility of capital made it easier for financial interests to force corporations to restructure and downsize, a development whose costs and benefits to society as a whole are still being debated (although it seems pretty clear that the restructuring of the corporate sector enhanced wealth for the shareholding class). In other words, capital may, counter intuitively, be more powerful in the long run in countries where shareholdings are relatively small, dispersed, and mobile,

than in countries where large block shareholders have much more direct influence over managers.

The mobility of capital in common law countries has probably made it easier for start-up firms to get new capital, thereby contributing to the dynamism of the high-tech sector in the US in the 1990s. However, increasing the rate of growth and technological change quite possibly reduces social and income stability, and widens the disparities in income distribution—effects that may not be politically acceptable in all societies. Hence, there may be an unavoidable trade-off of social goals, and which path a society chooses is partly a matter of politics and ideology.

On a theoretical level, numerous scholars and commentators, especially in the US, approach corporate governance questions from the perspective of a 'principal-agent' view of the firm, in which shareholders are understood to be the 'owners' of firms, and directors and officers their 'agents' (e.g., Easterbrook and Fischel 1991, Shleifer and Vishny 1997). Analysts who have adopted this point of view tend to be more libertarian and more comfortable with the social ordering that emphasizes private property and limited role for government. The normative implications of this perspective are that corporations should be run solely for the purpose of maximizing value for shareholders, and that corporate governance systems should support this goal alone, and not other social agendas. Hence, the relative impunity with which directors can control the nomination and voting process and resist takeovers is often viewed as a flaw in the US system because shareholders are not given sufficient control to enforce shareholder primacy (e.g., Jensen 1993). However, the freedom with which capital moves is viewed as a strength offsets this weakness. Meanwhile, the legal rules and practices in Japanese and German firms that have made them much more protective of employee interests relative to shareholders' interests are viewed as flaws by scholars and policy analysts who adopt this point of view.

On the other side, a small but growing body of scholarship has argued that the principal-agent perspective is misleading both as a matter of law and of economic theory. Under US law, the role of corporate directors is unique, but is much closer to that of 'trustees' than that of 'agents.' Outside the US, shareholders rarely are referred to as 'owners' of corporations (as if corporations were merely pieces of property). An emerging theoretical perspective emphasizes that the productive activity undertaken by corporations is a form of 'team production,' in which numerous parties in addition to shareholders are involved in the productive activity of corporations and have assets (whether tangible or intangible) at risk in the enterprise. This alternative view has somewhat different normative implications. For example, it suggests that corporate governance should not just be about maximizing value for shareholders. Rather, it should be about creating structures that foster and encourage wealth creation for all of the 'team' members. One mechanism by which this may be achieved in the US is the very mechanism so maligned by agency cost theorists: board independence. Mechanisms that insulate directors from pressures from all stakeholders, including shareholders, may serve to protect minority shareholders (seen as important even by shareholder primacy advocates) and to protect the interests of stakeholders other than shareholders (Blair and Stout 1999, Coffee 1999, Hansmann and Kraakman 2001). Even though shareholders do not exercise tight control over directors or managers of publicly-traded corporations, they enter into the relationship freely, thereby yielding legal control rights to the board. This allows them to make a credible commitment not to use their control over capital to unfairly squeeze out other stakeholders who may make irreversible investments in the enterprise. Hence, shareholders in firms where directors are relatively autonomous are more likely to win the cooperation of other stakeholders in the first place (Blair and Stout 1999).

The question is more than a theoretical exercise. As financial investors have grown in influence, the voices of shareholder primacy advocates have been quite prominent among those who want to see corporate governance standards harmonized worldwide. These voices have clout because capital has become so mobile internationally. Thus, the political question of whether corporate governance rules and standards should be harmonized across countries, and if so, whether international standards should be made to look more like the US or another model, is caught up in powerful economic forces, making it hard for countries to resist financial market pressures to move to a US standard.

See also: Business Law; Capitalism: Global; Corporate Finance: Financial Control; Corporate Law; International Law and Treaties; International Organization; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; International Trade: Geographic Aspects; Law and Economics; Law and Economics: Empirical Dimensions; Markets and the Law; Monetary Policy; Regulation, Economic Theory of; Regulation: Working Conditions; Stockholders' Ownership and Control; Venture Capital

Bibliography

Becht M 1997 Strong blockholders, weak owners and the need for European mandatory disclosure. European Corporate Governance Network Executive Report. European Commission, Brussels, Belgium

Berle A A Jr, Means G C 1932 *The Modern Corporation and Private Property*. Commerce Clearing House, New York

Black B S 1990 Shareholder passivity reexamined. *Michigan Law Review* **89**: 520–608

Blair M M 1995 *Ownership and Control: Rethinking Corporate Governance for the Twenty-first Century*. Brookings Institution Press, Washington, DC
Blair M M, Stout L A 1999 A team production theory of corporate law. *Virginia Law Review* **85**: 247–328
Coffee J C Jr 1999 The future as history: The prospects for global convergence in corporate governance and its implications. *Northwestern University Law Review* **93**: 641–707
Easterbrook F H, Fischel D R 1991 *The Economic Structure of Corporate Law*. Harvard University Press, Cambridge, MA
Fukao M 1995 *Financial Integration, Corporate Governance, and the Performance of Multinational Companies*. Brookings Institution Press, Washington, DC
Hansmann H 1996 The Ownership of Enterprise. Belknap Press of Harvard University Press, Cambridge, MA
Hansmann H, Kraakman R 2001 The end of history for corporate law. *Georgetown Law Journal* **89**: 439–68
Jensen M C 1993 The modern industrial revolution, exit, and the failure of internal control systems. *Journal of Finance* **48**: 831–80
Jensen M C, Meckling M H 1976 Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* **3**: 305–59
Milgrom P, Roberts J 1992 *Economics, Organization, and Management*. Prentice Hall, Englewood Cliffs, NJ
Millstein I M 1995 The professional board. *Business Lawyer* **50**: 1427
Organization for Economic Co-Operation and Development 1999a *Principles of Corporate Governance*
Organization for Economic Co-Operation and Development 1999b *Corporate Governance: Improving Competitiveness and Access to Capital in Global Markets*. A Report to the OECD by the Business Sector Advisory Group on Corporate Governance. OECD, Paris
Roe M J 1994 *Strong Managers, Weak Owners: The Political Roots of American Corporate Finance*. Princeton University Press, Princeton, NJ
Shleifer A, Vishny R W 1997 A survey of corporate governance. *Journal of Finance* **52**: 737–83
Williamson O F 1985 *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. Free Press, New York

M. M. Blair

Corporate Law

The term 'corporate law' refers to the body of legal rules that govern the creation, life, and extinction of business corporations. Generally speaking, a corporation is a particular type of business firm that is designed to address certain fundamental and universal economic exigencies, primarily the need for entrepreneurs to obtain capital. Thus, it is possible to consider generally the framework of matters regulated by corporate law, even though the nuances of specific countries' corporate law will depend on the distinctive features of their legal systems, histories, and traditions.

An understanding of corporate law must begin by examining the general problems of the theory of the firm, which defines the financing, managing, and organizing needs of business enterprises generally. With this background, it is possible in the second part to examine some basic features of corporate law—including the limited liability of shareholders and the contractual nature of corporate enterprises—and to consider the regulatory competition among corporate law policy makers. In the third section, internal organizations problems will be discussed, in particular the separation of ownership and control and the consequent issues of corporate governance. Finally, the fourth and final section will address the regulation of the financial structure of the corporation, specifically the problems raised by bonds issuance and the role of financial markets such as the market for corporate control.

1. *The Theory of the Firm and Corporation*

The reason why business ventures in modern economies are carried on by organized firms, instead of by individuals, is best understood through the theory of transaction costs (Coase 1937). The basic insight of this theory is that through the firm, it is possible to transform part of the transaction costs of running a business into what might be called control costs or agency costs. A useful example is the employer–employee relationship. Imagine that an entrepreneur has two possible choices to accomplish a particular task. Either the entrepreneur can buy the required service from an independent provider (contracting on the market), or he can hire an employee (creating an organization) to perform the same service. Under the first option, the entrepreneur must draft a detailed contract that specifies the relationship between the parties. Creating a separate contract involves significant expenditures of time and money to draft, monitor and possibly enforce the agreement. In economic parlance, these expenses are called 'transaction costs.' These transaction costs will be further increased every time the entrepreneur must negotiate a new agreement for a different service from a different provider.

Under the second option, if the entrepreneur hires an employee to perform the necessary services, his transaction costs will go down. An employment contract is an 'open' contract, in which the details of the performance do not need to be delineated with great particularity. Instead, an employer can simply direct the work of the employee according to specific and perhaps changing needs. Although this approach reduces transaction costs, it generates a new form of costs. These costs arise because, in contrast to an external provider who likely has an incentive to perform efficiently, an employee has less incentive to be efficient. As a consequence, the employer will need to monitor the employee's performance, thereby expending both the cost of sanctions for deviations and the cost of incentives for proper performance. The increased costs that a principal incurs in monitoring to

ensure that an agent adequately performs are called 'control costs' or 'agency costs.'

In addition to monitoring expenditures by the principal, agency costs also include what might be termed 'bonding expenditures' by the agent. (Jensen and Meckling 1976). Bonding expenditures are costs incurred by the agent to expand resources that guarantee the agent will not take certain actions, like quitting. The classic example of bonding expenditures is training or professional education in the field, which creates a high level of specialization and can make it more costly for the agent to change jobs.

Generally, a firm will prefer to internalize production whenever the agency costs will be lower than the transaction costs of an external relationship. Given the incentives and means for reducing transaction costs, the expected result is an increase in agency relationships for business enterprises, meaning an increasing preference for the firm structure over an integrated network of independent entrepreneurs. The firm can thus be understood as a method of organizing production (Posner 1998).

In addition to organizing production, another fundamental need that firms have is to raise capital with which to operate or expand. Again examining the economics of available choices, the nature and purpose of the corporate structure will be revealed. If an entrepreneur lacks financial resources or does not want to risk his own money in the business, he can borrow capital. Economic analysis demonstrates, however, that if the entrepreneur finances his business only through borrowed capital, the interest rate will rise to an uneconomic level because of the high risks of an unbalanced financial structure (Posner 1998).

There are two potential solutions to this problem. Either the entrepreneur can draft a very detailed loan agreement, which will lower the level of risk faced by the lender, or he can admit partners to the enterprise, who will provide (and risk) the necessary capital, but will also share the profits of the business. This second option can be economically preferable in several respects. First, it lowers the entrepreneur's contracting costs because he no longer has to negotiate the details of a loan agreement, and because it is not necessary to bargain around the interest rate since the partners' return on their investment will be determined by the business's success. In addition, the presence of creditworthy partners can improve the credit rating of the firm, thus lowering the cost of satisfying future capital needs.

If the need for capital increases significantly, however, this option becomes less desirable and the partnership structure shows its limits. The first problem is that partners in a firm have unlimited liability, which means significant agency or control costs. Because they are personally and unlimitedly liable for the conduct of the business, they will seek to intrude in the management of the firm to protect themselves and their investment. Moreover, because the financial strength of the partnership depends on the personal assets of the different partners, before investing everyone of them will want to investigate the other partners' wealth. Additional limitations are inherent in the nature of partnership, which is strictly dependant on the identity of the individual partners and can, for example, be dissolved on the death of one of the partners: the impermanence of the partnership agreement may thus affect the ability of the firm to raise capital from financial institutions.

For all these reasons, it becomes clear that the partnership structure is not suitable for larger, capital-intensive enterprises. Instead, the corporation is the form designed to facilitate further capital-raising activities. The two defining features of the corporate structure that allow it to do this are the limited liability of its owners and the division of ownership into 'shares,' meaning securities that represent a stake in the corporation but can easily be bought and sold. Corporate law is the set of mandatory and default rules that creates these features and regulates the consequences of them.

2. *Basic Features of the Corporation and Corporate Law*

2.1 *Limited Liability*

Notwithstanding attempts to trace back the modern corporation to ancient forms of organization of business enterprises, 'limited liability,' as the term is understood today, was an invention of the seventeenth century, granted to legal entities established for the purpose of raising funds to finance colonial expansions. The increasing need for new capital, which could neither be found in the treasuries of western monarchies nor raised through taxes, made resort to private saving an imperative resource. In order to collect such a considerable amount of money it became necessary to involve also small and medium bourgeoisie classes. Because these groups were risk-averse with their money, it was necessary to ensure legally that if they invested, only the amount of the investment would have been at risk. Limited liability was considered a royal privilege, an exemption from the general principle of unlimited liability, and as a consequence, it was subject to the discretion of the sovereign. It was only later that the 'power' to grant limited liability to corporations became a nondiscretionary power and was entrusted, in continental Europe, to the judiciary.

Borne out of historic necessity, limited liability has remained critical to the development of the modern corporation, and consequently to the development of western capitalism. The predominant economic explanation for the modern importance of the corporate structure is that it reduces the costs generated by the

separation and specialization of ownership and management functions (Easterbrook and Fischel 1985, Woodward 1985). A more detailed account of this economic effect involves several distinct but complementary considerations.

The first major consideration is that limited liability reduces the investor's need to incur monitoring costs. True to its name, 'limited liability' defines and restricts to the original investment the total amount of money an investor can lose. With the fear of unlimited liability gone, the investor no longer has a great incentive to interfere with the management of the corporation or to monitor the financial wellbeing of fellow shareholders.

Another important consequence of limited liability is that it makes freely transferable shares possible. In a setting with unlimited liability, shares would not be readily transferable because it would be highly relevant who is selling the shares and who composes the shareholders' group (they are not, in other words, 'commodities'). Meanwhile, when shares are readily transferable, the investor's incentives to monitor management are further reduced because his shares can be sold if he is dissatisfied with the corporation's performance, which can be easy if there is an active market for the shares. Another consequence of the reduction in the need to monitor and the availability of ownership through shares is that they permit the investor to diversify his investments into several companies, which further attenuates the consequences of poor management or adverse external events in one particular company.

The free transferability of shares, in turn, makes it possible to create a market for corporate control, which can be a powerful mean of external governance of the corporation. In fact, if shareholders are dissatisfied with the management of the corporation, they can sell their shares, thus causing a decline in the price of the shares. The corporation can thus be exposed to hostile takeovers by newcomers that believe to be able to make a profit buying the shares at a low price, manage it efficiently and thus enjoy a capital gain in the value of the shares. Thus, the presence of a market for corporate control represents an incentive for managers and controlling shareholders to act efficiently and not oppress minorities because, as long as shares have a voting right attached, poorly managed companies are exposed to the risk that shareholders will sell their shares and cause a change in control.

Corporate law not only creates limited liability to promote these aims, it also regulates some potentially harmful side-effects of limited liability. One consequence of limited liability and ownership through shares is that ownership can become more and more separate from management: the typical shareholder is not knowledgeable about the firm and, because there are reduced incentives to monitor management, does not expect to participate in its management. This separation between management and control—as will be discussed below—brings with it some potential for problems.

While the basic economic feature of firms and the need to raise capital justify the legal concept of limited liability as a general matter, there are also scenarios experienced by all systems that counsel in favor of exceptions to limited liability. For example, if the corporate structure is used to mislead creditors or perpetrate fraud, or provides inappropriate incentives to corporate actors, the protection of limited liability is no longer justified and it is instead better policy to permit the creditors of the corporation to reach shareholders' personal assets. In light of these potential consequences, all European and North American legal systems provide a mechanism 'to pierce the corporate veil' or to eschew a corporation's limited liability.

In civil law systems, the circumstances in which this mechanism is available are usually strictly defined by law. For instance, under Italian law, article 2362 of the Civil Code permits the corporate veil to be pierced in the case of corporate insolvency if there is a single shareholder. In common law systems, by contrast, there is greater flexibility in determining when the corporate scheme can be disregarded. A general overview of US and British cases on this issue suggests that the most important situation in which courts pierce the veil is when separate incorporation misled creditors. There are also a range of other exceptions, such as when the debtor corporation acted as a mere *alter ego* of the shareholders, when the corporation is not adequately capitalized or when shareholders have deceived creditors on the financial solidity of the corporation. In addition, US courts have demonstrated a willingness to disregard corporate limited liability in tort cases reasoning that without the possibility of such liability, corporations have less incentive to avoid disastrous harm to victims who they will never be required to compensate beyond the limited value of the corporation. In that situation, instead of simply attracting capital, limited liability creates the potential for a 'moral hazard' (Hansmann and Kraakman 1991).

2.2 The Corporation as a Standard Contract

Corporate law provides for limited liability, regulates its consequences and provides for exceptions to it through a set of rules that provide for what can be understood as a standardized contract. Corporate statutes enacted by the different legislators can be seen as a set of default rules that apply once the company is incorporated in a particular state. These comprehensive rules reduce the costs of contracting that parties would face if they had to individually negotiate all of the terms that are provided in the corporate law. The fundamental problem in this respect concerns mandatory versus enabling rules. Should corporate

law provide for some mandatory fundamental rules concerning shareholders' rights, minorities protection, corporate governance, and so on that parties can not opt out? Or, should corporate law only propose some enabling rules, that apply only if parties do not agree to the contrary? Finding the appropriate balance between mandatory and enabling rules is one of the fundamental problems of corporate law.

On one side, providing for mandatory rules ensures a higher level of protection of minorities in systems in which there is no active market for corporate control. When shareholder minorities are not organized and are thus unable to play an active role in controlling managers and shareholder majorities, mandatory legislative protection becomes a necessary incentive for investment. On the other hand, the lack of flexibility that characterizes mandatory rules can be inefficient if those rules differ from the ones that the parties would prefer. For instance, the presence of a mandatory rule concerning corporate disclosure can interfere with the willingness of a risk-averse investor to finance an enterprise that would prefer to opt for a lower level of disclosure. In such situations, default rules that are modifiable by the parties might produce a more efficient result. Enabling rules promote efficient regulation, however, only when investors are able to make their decisions based on accurate assessments of the regulation chosen by managers and controlling shareholders. This prerequisite suggests that default rules will only be efficient when there are organized minorities (usually institutional investors), who do not face information asymmetries and problems of collective actions in 'contracting' around default rules.

Against this basic economic background, it is clear why in common law countries, and the US in particular, the presence of more developed financial markets and of stronger 'organized' minorities (through institutional investors such as mutual funds) made possible corporate statutes that contain primarily default rules, in contrast to the more rigid mandatory systems that prevail in continental Europe. Things are changing at a rapid pace in Europe, and with the development of financial markets and the increasing intervention of foreign investors, even more traditional civil law countries are moving toward a more flexible corporate law that relies more on default rules (Easterbrook and Fischel 1989, Coffee 1989).

2.3 Regulatory Competition in Corporate Law

An important consequence of the fact that different legislators provide for different corporate law statutes, which strike different balances between mandatory and enabling rules, is that a 'regulatory competition' has developed to attract corporate charters (Cary 1974, Winter 1977, Romano 1987). Regulators may desire to have business enterprises incorporating within their boundaries for different reasons, such as incorporation fees, taxes, or indirect effects on the economic system. To achieve their goal, regulators have incentives to enact a body of corporate law that fits the needs of those who will choose the state of incorporation. One risk in this situation, however, is that those who have the power to choose (the controlling shareholder, the directors or the managers) have different incentives than the minority shareholders. Consequently, they may be tempted to choose a jurisdiction that maximizes their utility, but not that of the minority investors. The result is that regulatory competition can lead either to a race to the top (where rules maximize the wealth of all the subjects involved) or a race to the bottom (where rules attract corporations at the expense of minority investors). Between these two possibilities, the final result will depend on the ability of investors to 'penalize' corporations that choose a regime that is bad in terms of minority protection.

In the US, regulatory competition has been possible because corporations are free to incorporate and reincorporate in the different states. This freedom of regulatory competition has produced the so-called 'Delaware effect,' meaning that a majority of corporations, regardless of their physical location, have chosen to incorporate in Delaware (Macey and Miller 1987). The effects of this competition on shareholders' wealth are still discussed by scholars, but it is clear that corporations demonstrate a preference for statutes that provide more default rules.

In Europe regulatory competition has not been possible so far because of a more 'protectionist' approach, based on the so-called *siege reel* principle, which requires that corporation be governed by the law of the state where they have their principal place of business. Some recent decisions of the European Court of Justice seem to suggest a more opened approach that would promote regulatory competition. Most particularly, the so-called Centros Case, decided by the European Court of Justice on March 9th 1999, affirmed the possibility for a business to incorporate in England even though it operated only in Denmark and did not have contacts with England. This case suggests that changing world economic pressures will encourage wider acceptance of enabling rules and regulatory competition in corporate law in Europe.

3. Separation of Ownership and Control and the Reduction of Agency Costs

The separation of ownership and control (Berle and Menas 1932) raises problems of agency costs, analogous to, perhaps even more extreme than, the agency costs discussed above when an employer monitors an employee. Investors are principals who entrust their savings to majority shareholders, directors or managers (depending on the level of the analysis). Internal governance structures, such as corporate

governance rules, and external governance structures, such as the market for corporate control, can be seen as devices designed by corporate law to reduce these costs, allowing principals to oversee, directly or indirectly, the activity of their agents.

3.1 Internal Governance Structures: Corporate Governance and Internal Organization

Corporate governance, strictly speaking, is the set of rules that governs the relationships between and the balances the interests among shareholders, directors, managers and employees. In a more general sense, corporate governance also includes social controls on firms regarding such issues as antitrust, consumer protection, and pollution control. For the purposes of this discussion of corporate law, analysis will be limited to the three major areas that affect the internal organization of corporations: shareholder voting, the board of directors, and control and auditing.

3.2 Shareholders Voting

Shareholder voting regulations address issues such as who should vote, what issues should they vote on, and which procedures they should follow. From an economic standpoint, scholars contend that voting power is granted to the group that has the appropriate incentives (collectivity problems to one side) to make a discretionary decision that maximizes the corporate wealth (Easterbrook and Fischel 1983). For this reason, in most but not all systems, the shareholders' meeting usually has the power to deliberate major changes in the financial and legal structure of the corporation. In common law countries, which are characterized by efficient and strong capital markets, directors are granted expansive discretionary powers in relation to shareholders. By contrast, in the civil law systems of continental Europe, the shareholders' meeting is still seen as the 'central' organ in corporate law. The shareholders' meeting is entrusted exclusive powers concerning the juridical and financial structure of the corporation (such as to amend the corporate charter, and to deliberate over the issuing of new shares).

One of the fundamental problems concerning shareholders' voting is the issue of collective action and individual opportunism. Either when there is a controlling shareholder or widely dispersed ownership, the marginal shareholder will have no incentives to collect costly information in order to vote intelligently. Instead, the marginal shareholder is likely to rely on the possibility that other shareholders will incur those costs and follow their lead in voting. In economic terms, this situation is referred to as the 'free-rider problem.' While rational at the individual level, the choice at the aggregate level can preclude minority shareholders from ever using their *voice* to oversee the management of the corporation. Instead, they will sell the shares or 'exit' if they are dissatisfied with the way the corporation is run.

Meanwhile, for several reasons, institutional investors who hold a minority but non-trivial share can exercise active control on management through voting and it is often rational for them to do so. First, institutional investors encounter significantly lower information costs than individual investors and the availability of a non-trivial number of votes may make it possible to actually influence the management of the corporation. The other traditional alternative for minority shareholders, selling shares, can be difficult without having a bearish effect on the market. The presence of institutional investors, a development of the 1960s and in the 1970s in the US and more recently in Europe, is an important development that affects the effectiveness of corporate governance tools. In fact, without qualified minorities capable of activating them, instruments permitting direct control by minorities, such as derivative suits and proxies, would be absolutely useless (Black 1992).

Proxies and proxy fights are considered one of the most important tools through which minorities can participate in the life of a corporation, even if available data suggests that they are used less frequently than tender offers to transfer corporate control (Pound 1988). A proxy fight occurs when management and dissidents seek to obtain authority to act as designated voting representatives of voting shareholders at shareholder meetings. Virtually all the legal systems provide rules to regulate proxy fights. The regulation of proxy fights aims both to permit the solicitor of proxies an opportunity to reach all shareholders, which can be difficult with public companies if shares are recorded in 'street' names or the name of the nominee that administers the transfer of shares for clients (i.e., brokerage or bank). Corporate law also seeks to ensure that shareholders are provided with fair information and granted the possibility of an unbiased vote.

Another important feature of corporate voting is the potential for categories of shares with different rights attached. Almost all the legal systems statutorily prohibit cumulative voting and provide that every share should have no more than one vote. Nonetheless it is permitted, and indeed common, that corporations issue more than one class of stock, granting different rights for each class (for example, one class of voting shares and one class of non-voting shares with stronger economic rights). Typically, a class of shares granted with limited voting rights is issued for public shareholders, and another class with full voting rights is issued to majority shareholders interested in the actual control of the company (Gordon 1988). Notably, however, not all stock exchanges are willing to list limited voting shares. The New York Stock Exchange, for instance, until recently had very strict rules concerning limited voting shares.

In comparing regulation of proxies and the presence of non-voting or limited voting shares, there are two possible approaches to protecting minorities. On one hand, it is possible to acknowledge minority shareholders' lack of interest and ability in managing or controlling the corporation, and to protect them by reducing their administrative powers but enhancing their economic privileges, such as dividends and residual claims in case of dissolution. On the other hand, it is possible to allow minorities to play an active role in controlling majority shareholders' and managers by providing for specific tools such as proxy fights. The latter option generally requires the presence of institutional investors and organized minorities capable of actually using these tools in an informed and competent way. These different approaches are not mutually exclusive and in fact are often both present in corporate law statutes.

3.3 Boards of Directors

Directors are elected by the shareholders' meeting on a one-share-one-vote basis. They are entrusted with the power to manage the corporation and to represent it to third parties. Usually, directors are organized in a board and, as discussed above, different legal systems allocate competence between the shareholders' meeting and the board differently. Continental European systems usually provide for more extensive competence for the shareholders, while the US and other common law systems permit the board of directors to make the key decisions for a corporation without deferring to the shareholders' meeting.

The internal organization of a board of directors also differs among national systems. In the US, for instance, the flexibility of corporate regulation allows for the possibility of establishing one or more committees within the board of directors, which are designated specific tasks such as compensation or auditing. Another peculiar model is the board of the German corporation, which is characterized by a so-called dualistic system. The *Vorstand* has managing and representation powers, but its members are elected and controlled by the *Aufsichtsrat*, a controlling board directly appointed by the shareholders that can remove them for cause (*wenn ein wichtiger Grund vorliegt*, according to the § 84.3 AktG). German workers are also granted representation by law in the *Aufsichtsrat*, demonstrating a commitment to co-management between the providers of capital and labor (*Mitbestimmung*). In France the dualistic German-like model can be chosen, but is rarely used in practice. Instead, the board of directors is often dominated by the *President Directeur General*, who is both the chairman of the board and the chief executive officer of the corporation. This role was introduced during the German occupation for ideological reasons, but it has nonetheless proven to be efficient (for comparative references Weigmann 1997).

Normally, directors have two major duties, the duty to act with reasonable care in the managing of the corporation's businesses, and the duty to act loyally, which includes to avoid in conflict of interests. Breach of either of these duties can result in civil liabilities either to the company or its shareholders. While the duty of loyalty is strictly enforced, the duty of care is often blurred. Most systems provide for a 'business judgment rule,' which acts as a safe-harbor that prevents judges from second-guessing the merits of the directors' choices.

3.4 Auditors

Finally, most legal systems (but not all) provide for a separate board of auditors entrusted with the task of controlling the activity of the directors, especially with reference to accounting practices. The board of auditors is elected by the same majority that appoints directors, which can lead to problems since the elected auditors by the majority do not necessarily have any incentive to act in the interest of minority shareholders. For this reason, some legislation provides a mandatory rule, at least for public companies, under which the board of auditors must include a member appointed by minority shareholders. An example of this approach is the new Italian law on financial markets and listed corporations (the d. lgs. n. 58 of 1998, art. 148).

4. External Governance Structures: The Market for Corporate Control

Until this point, discussion has been limited to those internal governance structures that reduce the cost of control faced by corporate investors. These costs are also reduced by external forces, most significantly the market for corporate control. The inability of managers or controlling shareholders to create wealth for all shareholders can be reflected in the market price of the shares. As noted above, if shareholders are dissatisfied with the way their agents act, they can exit the investment. The sale of their investment will lead to a drop in the price of the shares, which might allow a shareholder who believes to be better able to manage the corporation to buy control and change management. The threat to managers and present controlling shareholders of loosing control should provide an incentive to align their preferences with those of minority shareholders (Manne 1965, Jensen 1988, Jarrel et al. 1988).

Given this possibility, making transfers of control easy could be seen as the best way to reduce agency costs, although this solution has its own problems. When a new shareholder takes control, he usually pays a premium for control. If the market is not very efficient, liquid, or quick in adjusting shares prices to the new control, it is possible that control can be acquired directly from the controlling shareholder. In

that situation, the premium for control will be paid directly to the controlling shareholder and minority shareholders will not enjoy any of the premium. For this reason, some legal systems have a mandatory tender offer rule, which seeks to distribute the control premium among all the shareholders. Typically in Europe, where there are less efficient markets, corporate legislation requires a public offer to buy shares to acquire control. In the US, by contrast, where the market is considered more efficient and able to minimize exploitation of minority shareholders, there are no—or less stringent—rules requiring specific compulsory offers in case of acquisition of control.

5. *Financing the Corporation: The Issuing of Bonds*

To complete a discussion of corporate law, it is necessary to note briefly an alternative financing option for corporations that is permitted by the corporate law of all legal systems: the issuance of bonds, which is essential to provide a mixed financial structure, thus leveraging financial risk.

Bond-holders can be seen as economic actors who are willing to lend money to the corporation but, different from the shareholders who accept a variable rate of return depending on the profits of the corporation, prefer to receive a fixed interest rate. In economic terms, the reason for this preference could be that they do not want to undergo the controlling expenditures faced by shareholders and prefer instead to contract in advance for a fixed rate of return. Bond contracts are thus written with the explicit purpose of avoiding the bondholder-stockholder conflict (Smith and Warner 1979). The US system provides an interesting illustration of the different interests of shareholders and bond–holders: while Delaware corporate law is usually chosen because its default rules allow greater flexibility, New York law is usually chosen to govern bond issuing because it is considered particularly protective of creditors and its mandatory rules preclude the need for very detailed contracts.

See also: Business Law; Contracts: Legal Perspectives; Corporate Finance: Financial Control; Corporate Governance; Delegation of Power: Agency Theory; Firm Behavior; Law and Economics; Liability: Legal; Organizations and the Law; Organizations, Sociology of; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulation Theory in Geography; Regulatory Agencies

Bibliography

Bebchuk L A 1990 *Corporate Law and Economic Analysis*. Cambridge University Press, Cambridge, UK

Berle A A Jr, Menas G C 1932 *The Modern Corporation and Private Property*. Commerce Clearing House, Inc, New York

Black B S 1990 Shareholder passivity reexamined. *Michigan Law Review* **89**: 520–608

Black B S 1992 Agents watching agents: the promise of institutional investor voice. *UCLA Law Review* **39**: 811

Cary W 1974 Federalism and corporate law: reflections upon Delaware. *Yale Law Journal* **83**: 663

Clark 1986 *Corporate Law*. Toronto, ON

Coase R H 1937 The nature of the firm. *Economica* **4**: 386

Coffee J C Jr 1989 The mandatory/enabling balance in corporate law: an essay on the judicial role. *Columbia Law Review* **89**: 1618

Coffee J C Jr 1999 The future as history: the prospectus for global convergence in corporate governance and its implications. *Northwestern University Law Review* **93**: 641–707

Eastbrook F H, Fischel D R 1983 Voting in corporate law. *Journal of Law and Economics* **26**: 395

Eastbrook F H, Fischel D R 1985 Limited liability and the corporation. *University of Chicago Law Review* **89**: 52

Eastbrook F H, Fischel D R 1989 The corporate contract. *Columbia Law Review* **89**: 1416

Eastbrook F H, Fischel D R 1991 *The Economic Structure of Corporate Law*. Harvard University Press, Cambridge, MA

Gordon J N 1988 Ties that bond: dual class common stock and the problem of shareholder choice. *California Law Review* **76**: 2

Gordon J N 1989 The mandatory structure of corporate law. *Columbia Law Review* **89**: 1549

Gower 1992 *Gower's Principles of Modern Company Law*. London

Guyon 1994 *Droit des affaires. Droit commercial général et sociétés*. Paris

Hansmann H 1988 Ownership of the firm. *Journal of Law, Economics, and Organization* **4**: 267

Hansmann H, Kraakman R 1991 Toward unlimited shareholder liability for corporate torts. *Yale Law Journal* **100**: 1879–934

Henn G 1994 *Handbuch des Aktienrechts*. Heidelberg

Jarrel G A, Brickley J A, Netter J M 1988 The market for corporate control: the empirical evidence since 1980. *Journal of Economic Perspectives* **2**: 49

Jensen M C 1988 Takeovers: their causes and consequences. *Journal of Economic Perspectives* **2**: 21

Jensen M C, Meckling M H 1976 Theory of the firm: managerial behavior, agency cost and ownership structure. *Journal of Financial Economics* **3**: 305–59

Macey J R, Miller G P 1987 Toward an interest group theory of Delaware corporate law. *Texas Law Review* **65**: 469–523

Manne H G 1965 Mergers and the market for corporate control. *Journal of Political Economy* **73**: 110

Posner R A 1998 *Economic Analysis of Law*. Aspen Law and Business, New York

Pound J 1988 Proxy contents and the efficency of shareholder oversight. *Journal of Financial Economics* **20**: 237

Roe M J 1994 *Strong Managers, Weak Owners: The Political Roots of American Corporate Finance*. Princeton University Press, Princeton, NJ

Romano R 1987 The state competition debate in corporate law. *Cardozo Law Review* **8**: 709

Smith C W, Warner J B 1979 On financing contracting: an analysis of bond covenants. *Journal of Financial Economics* **7**: 117

Uría 1993 *Derecho Mercantil*. Madrid

Weigmann R 1997 Società per azioni. In: *Digesto delle discipline privatistiche*. Sezione commerciale, Utet, Torino

Williamson O E 1984 Corporate governance. *Yale Law Journal* **93**: 1197–230

Williamson O F 1985 *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. Free Press, New York
Winter R 1977 State law, shareholder protection, and the theory of the corporation. *Journal of Legal Studies* **6**: 251
Woodward S E 1985 Limited liability in the theory of the firm. *Journal of Institutional and Theoretical Economics* **141**: 601

P. Marchetti and M. Ventoruzzo

Corporate Taxation

Corporate taxation is an important source of government revenue around the world, and a major consideration in planning business activities. This article identifies the economic incentives created by the taxation of corporate income, and reviews available evidence of their behavioral impact.

1. *Scope and History*

Corporate tax obligations consist chiefly of fractions of corporate income. The corporate income-tax rate structure is usually progressive, meaning that average tax rates rise with income, typically reaching a maximum rate rapidly enough that almost all of the income of large corporations is subject to tax at the highest rate. In the United States in 1999, corporations earning net income up to $50,000 paid 15 percent tax; this tax rate rose to 34 percent on income over $100,000, and was 35 percent on incomes above $10 million. Other countries tax corporate income at similar rates that can vary according to the size, location, and industry of the corporate taxpayer. Thus, desires to rectify geographic income disparities led Germany, Italy, and other countries to offer tax concessions for investment in depressed regions, while mineral-rich countries such as South Africa and Papua New Guinea subject mineral extraction to taxation at supranormal rates.

Much popular and public policy attention is devoted to the tax planning activities and corporate tax obligations of big businesses—and with good reason, since corporate income tends to be concentrated in a relatively small number of large companies. For example, in 1997, there were 4.7 million active corporations in the United States, of which approximately 9,000 had assets exceeding $250 million. These 9,000 large corporations accounted for 86 percent of total US corporate assets, 80 percent of net corporate income, and 78 percent of total corporate tax payments.

The American federal corporate income tax was introduced in 1894 but was found to be unconstitutional the following year; it reappeared as a gross receipts tax in 1909, and was modified to become a genuine income tax following ratification of the 16th Amendment to the US Constitution in 1913. The US corporate tax rate in 1913 was 1 percent. The corporate tax rate rose over time, though it remained below 15 percent until World War II, at which point it rose sharply. Corporate tax was a major revenue source for the US federal government between World War II and the late 1960s, when corporate tax collections consistently represented more than 20 percent of federal government revenue. Corporate tax collections have fallen as a fraction of total government revenue since then, though they exceeded $188 billion in 1998, representing approximately 11 percent of US federal government revenues, or 2.2 percent of US gross domestic product (GDP). This pattern is repeated in other high-income countries, many of whose corporate income tax rates and tax provisions are remarkably uniform. In a typical recent year (1994), top marginal corporate tax rates among the 12 member countries of the European Union ranged from a low of 33 percent to a high of 45 percent, with just a single country (Germany) taxing corporate income at a rate in excess of 40 percent.

2. *Impact on Economic Behavior*

The taxation of corporate income encourages entrepreneurs and managers to structure and conduct their business operations in ways designed to avoid taxes. Corporations generally reduce their tax obligations, and those of their shareholders, by using debt rather than equity finance, investing in assets that can be depreciated rapidly for tax purposes, and those for which generous tax credits are available, and avoiding dividend payments or other tax-disadvantaged distributions to investors.

The major tax consideration in corporate finance is that interest payments to bondholders are deductible from taxable income, while dividend payments to corporate shareholders are not deductible. As a result, corporations generally have tax incentives to issue more debt than they would otherwise, and there is ample evidence (Auerbach in press) that debt/equity ratios are higher as a result. Since dividend payments generate tax obligations for recipients, corporations also have incentives to buy back shares, or to use other non-dividend methods of distributing income to shareholders.

Tax incentives to issue debt are mitigated in situations in which firms anticipate nonrefundable tax losses, or in which there is strong investor preference for returns in the form of capital gains and dividends rather than interest income. The strength of investor preference for stock ownership depends on circumstances. Individual investors most strongly prefer owning a firm's stock to owning its bonds if they reside in countries (such as Australia, France, or the United Kingdom) that permit dividend recipients to claim credits for corporate taxes paid on income generating dividends. In countries without such personal and

corporate tax integration, the favorable tax treatment of accruing capital gains means that some investors, typically those in the highest tax brackets, prefer stock ownership (despite the associated corporate tax obligations) to holding interest bearing debt. With sufficient numbers of such investors, it follows that asset markets will price stocks and bonds so that the economy's debt/equity ratio increases with the corporate tax rate, while any individual corporation is indifferent between debt and equity finance (Miller 1977).

Taxation influences the timing, magnitude, and composition of corporate investment in plant and equipment, inventories, research and development, and other business assets. Higher tax rates generally reduce investment, though this depends on tax treatment of investment expenditures. Investors generally do not reduce taxable income immediately by the full amount of spending on business investment, instead amortizing the cost of investment spending over a period of years in which 'depreciation allowances' are permitted. These depreciation allowances are most valuable if claimed soon after an investment is undertaken, since they are seldom adjusted for inflation and the time value of money. Hence governments that seek to encourage business capital formation often do so by permitting investors to depreciate their investments for tax purposes over short periods of time. Governments also often provide special tax credits for investments in selected asset categories, such as research and development (R&D) (available in Canada, Japan, Spain, the United States, and elsewhere). Depreciation allowances and investment credits together have the potential to stimulate investment by reducing required rates of return (Hall and Jorgenson 1967). These tax considerations are captured succinctly by 'effective tax rates' (Auerbach and Jorgenson 1980) on business investment, which usefully can be compared over time and between countries (King and Fullerton 1984). The available evidence indicates that rates of business investment are inversely related to effective tax rates (Hassett and Hubbard in press).

Corporate taxation also affects business organization by discouraging the incorporation of profitable businesses that can be organized in noncorporate form. In the United States, business organizations whose income is not subject to the corporate income tax include small ('S') corporations, partnerships, sole proprietorships, and limited liability companies.

3. *The Taxation of Multinational Corporations*

Multinational corporations—those with active business operations in more than one country—pose special problems for tax systems, since it is necessary to determine the location and character of taxable income, as well as the means by which double tax relief is to be provided. Double tax relief is essential, because a corporation's home country claims the right to tax all of its income, including the income earned by its foreign affiliates, while host countries in which foreign affiliates are located insist on their rights to tax affiliates' incomes. In the absence of special tax relief, income earned by foreign affiliates would be subject to taxation both by host countries and by home countries at cumulative rates that might approach or exceed 100 percent.

There are two practical systems of international double tax relief, both of which permit countries to tax fully any income earned by economic activity undertaken within their borders. Double taxation is avoided whenever home countries forgo taxing multinational corporations on income earned by their foreign affiliates, or else, when home countries tax the incomes of foreign affiliates but permit taxpayers to claim credits for foreign taxes paid. Use of the credit method effectively subjects income earned by foreign affiliates to home-country taxation at a rate equal to the difference between home and foreign tax rates. Actual methods of double tax relief often resemble a combination of these two methods. A majority of the world's countries exempt from tax most of the income earned by foreign affiliates of domestic multinational corporations, though several major capital exporting countries, including the United States, Japan, the United Kingdom, Italy, and others, subject such income to taxation but permit credits to be claimed for foreign taxes.

There is ample evidence that international tax rate differences influence the timing and location of investment by multinational corporations (Hines 1999). Countries and subnational jurisdictions that tax corporate income at low rates receive unusually large volumes of foreign direct investment, particularly from countries that do not tax foreign income. Furthermore, multinational firms arrange their business transactions to avoid international taxes. Thus, foreign affiliates located in high-tax countries tend to be financed with higher debt/equity ratios than are investments in low-tax countries, and are more likely to remit profits in the form of (deductible) royalty payments than (nondeductible) dividends. Only a small fraction (16 percent in 1984) of the foreign affiliates of American companies pay dividends to their parent companies each year, and those that do tend to have particularly tax advantaged situations (Hines and Hubbard 1990). And the pretax profitability of foreign affiliates is correlated negatively with host country tax rates (Hines and Rice 1994), which is suggestive of tax motivated transfer pricing and unlikely to be the outcome of ordinary investment responses to tax rate differences.

4. *Incidence of the Corporate Income Tax*

The requirement that corporations pay taxes does not mean that owners of corporations necessarily bear the

burden of the corporate tax, since this burden might be shifted, partially or entirely, to consumers in the form of higher prices, or to workers in the form of lower wages. Indeed, there are reasonable situations in which the existence of the corporate tax actually enhances returns to capital owners, including owners of corporations (Harberger 1962).

Corporate tax obligations contribute to the cost of investment and thereby encourage substitution of other productive factors (such as labor) for capital used by corporations. Labor expenses are deductible against taxable income, so the corporate tax does not affect the marginal condition characterizing a firm's decision of whether or not to employ additional labor. While the substitution of labor for capital depresses the demand for capital and thereby reduces its after-tax market return, there exists a separate effect of corporate taxation that stems from its induced intersectoral reallocation of resources.

Corporate taxation increases the cost of producing corporate output, thereby raising output prices, depressing demand, and shifting output from the corporate sector of the economy to the noncorporate sector. This reallocation affects factor demands to the extent that factor input ratios differ between the corporate and noncorporate sectors of the economy. If the corporate sector of the economy has a lower capital/labor ratio than the noncorporate sector, then the introduction of a corporate tax shifts resources into the noncorporate sector and thereby raises the demand for capital. If this effect is large enough, it has the potential to exceed in magnitude the countervailing impact of factor substitution, thereby implying that higher rates of corporate tax are associated with greater after-tax returns to capital—including capital invested in corporations. It would then follow that labor bears the burden of the corporate tax in the form of lower wages. The distribution of the corporate tax burden is an important area for future empirical research. The limited available statistical evidence is inconclusive (Krzyzaniak and Musgrave 1963), thereby leaving open the possibility that corporate taxation ultimately improves the after-tax profitability of owning corporations.

See also: Corporate Finance: Financial Control; Corporate Law; Income Distribution; Income Taxes; Multinational Corporations; Policy Process: Business Participation; State and Local Taxation; Wealth Distribution

Bibliography

Auerbach A J in press Taxation and corporate financial policy. In: Auerbach A J, Feldstein M (eds.) *Handbook of Public Economics*. North-Holland, Amsterdam, Vol. 3

Auerbach A J, Jorgenson D W 1980 Inflation-proof depreciation of assets. *Harvard Business Review* **58**: 113–18

Hall R E, Jorgenson D W 1967 Tax policy and investment behavior. *American Economic Review* **57**: 391–414

Harberger A C 1962 The incidence of the corporation income tax. *Journal of Political Economy* **70**: 215–40

Hassett K A, Hubbard R G in press Tax policy and business investment. In: Auerbach A J, Feldstein M (eds.) *Handbook of Public Economics*. North-Holland, Amsterdam, Vol. 3

Hines J R Jr. 1999 Lessons from behavioral responses to international taxation. *National Tax Journal* **52**: 305–22

Hines J R Jr., Hubbard R G 1990 Coming home to America: Dividend repatriations by U.S. multinationals. In: Razin A, Slemrod J (eds.) *Taxation in the Global Economy*. University of Chicago Press, Chicago

Hines J R Jr., Rice E M 1994 Fiscal paradise: Foreign tax havens and American business. *Quarterly Journal of Economics* **109**: 149–82

King M A, Fullerton D (eds.) 1984 *The Taxation of Income from Capital: a Comparative Study of the United States and the United Kingdom, Sweden, and West Germany*. University of Chicago Press, Chicago

Krzyzaniak M, Musgrave R A 1963 *The Shifting of the Corporation Income Tax*. Johns Hopkins University Press, Baltimore, MD

Miller M H 1977 Debt and taxes. *Journal of Finance* **32**: 261–75

J. R. Hines, Jr

Corporatism

Corporatism can be defined as a specific relationship between organized interests and government where interest associations coordinate some of their key activities with government policy. Such coordination most often presupposes a specific structure of organizations permitting their leaders to represent the particular interests of their members while restraining their demands and obliging them for the pursuit of overarching collective purposes. The concept achieved prominence in the study of industrial relations, of comparative political economy, and of public policy and has in particular been employed to investigate the impact of industrial relations on macroeconomic policy. But it covers a broader variety of sectors and has also been given the wider meaning of self-regulation through 'private interest government.'

1. Prehistory of the Concept

The concept of corporatism (or 'corporativism') initially designated institutional blueprints for 'functional representation' of organized interests as they throve since the late nineteenth century. Responding to the growing intensity of social conflicts between

capital and organized labor, conservative as well as socialist authors propagated the representation of interest groups by organizations with compulsory membership and their insertion into an institutional framework for peaceful conflict regulation. Among the influential proponents of a 'corporative order' were Catholic social theorists whose ideas found their expression in the papal encyclicals *Rerum Novarum* (1891) and *Quadragesimo Anno* (1931). The concept was also endorsed by conservative and fascist ideologues who proposed the replacement of representative parliamentary democracy by a 'corporative state.' It served as a basic institutional formula in the authoritarian regimes of Mussolini, Franco, Salazar, and Pétain (though not in Nazi Germany) and thus had for a long time a distinctly antiliberal and antidemocratic connotation. Similar corporatist features were also introduced in some authoritarian regimes of Latin America (Brazil, Argentina). But in the 1930s, corporatist ideas were advocated even in some liberal parties (e.g., in Switzerland).

The concept acquired a new—and this time explanatory rather than prescriptive—significance since the economic crisis of the 1930s, and notably in and after World War II (in the context of the management of shortages in the war and postwar economy). At that time, institutionalized forms of cooperation between governments and organized interests were observed in a number of democratic countries, especially in Western Europe. Often associations choose voluntary cooperation (notably in wage policy) in order to avoid dirigiste interventions of government. Subsequently such 'social pacts' acquired increasing importance as an instrument of Keynesian macroeconomic policy. Among the social scientists who pointed to the resemblance between these practices and some aspects of the 'corporativist' designs of the interwar years one of the first was Shonfield (1965), later followed by political scientists (also Lehmbruch 1974 and in particular Schmitter 1974) who developed the typological distinction between 'state corporatism' and 'societal corporatism' (or 'authoritarian' vs. 'liberal' corporatism). What distinguished these 'neocorporatist' arrangements from their forerunners was both their more or less informal character and the complete absence of an underlying articulate doctrine. Nowadays the terminological convention of social scientists most of the time ignores the prehistory and employs the term 'corporatism' for these new organizational patterns.

2. *Typological Approaches to the Study of Corporatist Interest Intermediation*

Political scientists employed this more recent version of the concept of corporatism to explain empirical observations that did not square with the established interpretation of interest politics often called 'pluralist.' In the tradition of the 'group approach,' political interest groups were defined as 'shared attitude groups that make claims upon other groups in the society,' eventually 'through or upon any of the institutions of government' (Truman 1951, p. 37), and hence the 'access' of interest groups to government (or, with a somewhat pejorative connotation, 'pressure politics') was a key analytical question. However, such an approach was not particularly suited to explain the emergence, in different countries, of 'social contracts' between government and organized interests for voluntary incomes policy as an instrument of Keynesian economic policy. Since in these instances government was making claims upon interest groups the vectors of influence were inverted.

Generalizing from such observations, the concept of corporatism describes a type of policy formation where organized interests are encouraged to (explicitly or preemptively) coordinate their activities with government and to foster its policy objectives. Thus the analytical problem is not *access* but a process of *exchange* in which interest associations pledge organizational compliance and are often rewarded with specific organizational privileges. As a complementary terminological consequence of this theoretical reorientation of interest group research, the established notion of 'interest representation' was substituted by 'interest intermediation' (Schmitter 1979, p. 93, endnote 1).

A corporatist exchange relationship between government and organized interests tends to benefit from specific structural properties of organizations. In Schmitter's well-known definition of corporatism, a limited number of associations enjoy a monopoly in the representation of specific interests, membership may be compulsory, and leadership is clearly centralized. This enables a corporatist association to obligate the organization in its relationship with government and other associations and permits the government to hold it responsible for obtaining the compliance of the membership. Pluralism, on the other hand, is characterized by a multiplicity of associations that may compete for turf in representing specific interests, in which membership is voluntary, and leadership positions may be subject to internal contestation. A crucial element of Schmitter's influential contribution was thus the conceptualization of 'corporatism' as one building block of a comparative typology of organized interests.

From the original fourfold typology with its strongly deductive flavor, subsequent research has only retained the distinction of 'corporatism' and 'pluralism,' while the constructs of 'monism' and 'syndicalism' turned out to be of limited relevance for empirical research. In this typology, 'pluralism' carries a specific connotation derived from earlier US research on interest associations. In particular the emphasis of the 'group approach' (as it was developed by David Truman 1951) on society as a 'mosaic of overlap-

ping groups of various specialized sorts,' with an inescapable 'proliferation of associations' that may eventually split and compete among each others, formed a striking contrast not only to the organizational monopolies of older 'state corporatism' but also to the all-encompassing 'peak associations' of some European democracies.

The concept of 'corporatism' should not be construed as an alternative to the theory of 'pluralist democracy' or 'polyarchy' with which it is, on the contrary, largely compatible. Rather, the typological distinction from 'pluralism' aimed specifically at modifying the 'pluralist' paradigm of interest group politics: 'pluralism' was no longer treated as a general paradigm but rather as a special type of interest group activity. The swift adoption of the typology in political science was not least due to its innovative contribution to comparative empirical research, where it facilitated the development of measurement concepts that so far had largely been missing.

2.1 *The Measurement of Corporatism*

In empirical research, the constructs of corporatism and pluralism should be considered as extreme types permitting the location of individual cases on (nominal or ordinal) scales. The typology has served not only to describe individual associations but also the structure of 'interest systems' (Schmitter 1979) of whole countries. Indeed, within particular countries the organizational properties of the most important associations and of their relationship with government and among each other are often characterized by a considerable degree of 'organizational isomorphism' (as defined by DiMaggio and Powell 1991). The location of countries on a corporatism scale has, in particular, become a preferred instrument of comparative research on public policies for the investigation of the impact of structures of interest intermediation on policy outcomes.

Operationalization of these typological constructs is subject to the availability of standardized indicators and data. The degree of organizational centralization of labor unions or of the collective bargaining system are most easy to observe and have in particular been employed by economists focusing on the impact of collective bargaining on macroeconomic performance (Calmfors and Driffill 1988). To describe more complex patterns of interest intermediation, some political scientists have preferred to elaborate multidimensional indexes for measuring the strength of corporatism in specific countries, focusing either on associational properties (Schmitter 1981) or on associational participation in policy making (Lehmbruch 1984).

3. *Macrocorporatism and Mesocorporatism*

Corporatist arrangements have been observed mostly in industrial relations, be it on the national level (*macrocorporatism*) or on the level of industries (*mesocorporatism*). But they are also found in other policy domains, such as the social security system or agricultural policy (sectoral corporatism).

3.1 *Macrolevel Concertation and 'Social Pacts'*

As a pattern of policy formation, corporatism was originally largely tantamount to *macrocorporatism*, defined as an interorganizational arrangement in which the national peak associations of business and labor coordinate their autonomous policies through central-level bargaining, with government being often (at least informally) involved in tripartite 'social pacts.' Although the degree of formalization can vary, their central objective is always the consensual coordination of wage setting across the economy, as a decisive contribution to the pursuit of macroeconomic goals such as monetary stability, employment, and economic growth. Such 'concertation' (a term derived from the French *économie concertée*) might also include connected areas such as labor market or tax policy. Up to the 1970s, corporatist concertation was mostly found in the European welfare states with a strong social-democratic tradition, and for a considerable period they succeeded, under the conditions of Keynesian macroeconomic policy, to neutralize the trade-off between economic stability and high employment.

When the objective of full employment was increasingly put into question and Keynesian economic policy was abandoned it was often assumed that, as a consequence of the decline of macroeconomic policy, peak-level concertation would become redundant. In particular it was believed that the monetary instruments available to central banks after the end of fixed exchange rates might be powerful enough to induce spontaneous wage restraint. And when in the 1980s highly centralized bargaining systems were abandoned in several countries this was interpreted as a crisis of macrocorporatism. But subsequent research has maintained that under specific conditions corporatist wage bargaining has remained successful in macroeconomic stabilization.

In Europe, moreover, speculative predictions about the demise of corporatism were challenged by the emergence of a new generation of 'social pacts.' This new type of concertation aims at a negotiated reorganization of welfare state policies, notably of the old-age pensions systems and health policy. In this context, coordinated wage policies have remained important because of the interdependence of wages, nonwage labor costs, and employment. But whereas central tripartite negotiations remain the prevailing forum for the conclusion of social pacts about consensual welfare reform, the macrocoordination of wage policy is not necessarily identical with centralized peak-level wage-setting because there are also decentralized forms of coordination that may even yield superior effects (Soskice 1990).

3.2 Mesocorporatism

The participation of associations in the governance of particular industries has been termed *mesocorporatism*. This covers quite different phenomena: It includes the participation of organized labor and business in the management of industrial crises where government support may be vital (e.g., steel or mining), but also self-regulation by associations that have often been granted binding powers ('private interest government'). Associational participation in the governance of markets is particularly found in agriculture. In manufacturing industries, standard setting is another well-researched field of corporatist 'private interest government.' Moreover, in a number of European countries *sectoral corporatism* is also found in policy domains such as social security or health policy.

4. Cross-national Distribution and Longitudinal Trends

The strength of corporatism varies longitudinally as well as cross-nationally. Variations across time have been the object of many speculations: in the 1970s some authors predicted a secular trend toward corporatism in industrial democracies, whereas in the following decades observers frequently detected a general decline of corporatism. In cross-national empirical research we are on much firmer ground, and the well-established result is considerable variability of the strength of corporatism across industrialized countries. The acknowledged examples of 'interest systems' characterized by strong corporatism are those of Austria, the Netherlands, and the Scandinavian countries; whereas Germany has weaker corporatism, and the USA is clearly characterized by a dominantly 'pluralist' system of interest intermediation. These differences have, in the majority of cases, been relatively stable since the end of World War II and can quite often be traced back to the emergence of modern interest organizations in the late nineteenth and early twentieth century.

However, as far as the countries with strong corporatism are concerned certain trends can indeed be observed. In specific phases, notably in situations of crisis, an older pluralism of interest associations was increasingly superseded by organizational monopolies for the most important associations. The war economy of the two world wars, the world economic crisis of the 1930s, and the economic reconstruction after World War II have contributed most to these developments. While theories of the transformation of the political economy of highly developed industrial societies (e.g., 'organized capitalism') have often overlooked these institutional variations across nations, research on corporatism suggests the existence of a contingent plurality of development paths due to specific historical circumstances. Thus, corporatism was more likely to emerge where premodern forms of interest organization (such as corporations and guilds) have been able to survive and transform themselves instead of being destroyed in the transition to a capitalist market economy. Also, different 'state traditions' are characterized by different approaches to the intermediation of interests (Crouch 1993). Among others, corporatism appears to be strongly related to consociational or 'consensus democracy.'

5. Beyond Corporatism: Comparative Political Economy and Public Policy

5.1 Effects of Corporatism and Composite Models of Coordination in Market Economies

Comparative research on the effects of corporatist interest intermediation on macroeconomic performance and income distribution seems to indicate that macrocorporatism contributes to low unemployment and to lower levels of income inequality. However, the relationship with inflation rates seems to vary over time. Recent research in the policy performance of macrocorporatism has increasingly come to adopt more complex models in which the structure of interest intermediation is assumed to interact with other variables. Such conceptualizations distinguish different production regimes (Hollingsworth et al. 1994) or the coordinated vs. uncoordinated (or liberal) market economies (Soskice 1990). In particular, effects of corporatism on macroeconomic performance are better explained if the interaction with central bank independence is taken into account (Iversen and Pontusson 2000).

5.2 Corporatism and the Borders of State and Society

The emphasis of corporatist theory on the contribution of interest associations to the pursuit of collective goals has raised the question whether the traditional dichotomy of 'state' and 'society' could still be used as an adequate conceptual framework for the analysis of public policies in highly developed industrial democracies. This has led to the replacement of the traditional image of a hierarchical sovereignty of the 'state' by notions such as *governance*, understood as 'all those activities of social, political and administrative actors that can be seen as purposeful efforts to guide, steer, control or manage (sectors or facets of) societies' and where public and private actors work in 'co-arrangements' (Kooiman 1993, p. 2). Corresponding notions employed in European social science are *Steuerung* in the German, or *régulation politique* in the French literature. Research on policy networks has

developed more rigorous analytical instruments to compare corporatist to other patterns (Knoke et al. 1996).

See also: Authoritarianism; Interest Groups; Interest Groups, History of; Pluralism; Policy History: State and Economy; Policy Process: Business Participation; State and Society

Bibliography

Calmfors L, Driffill J 1988 Bargaining structure, corporatism and macroeconomic performance. *Economic Policy* **6**: 14–61

Crouch C 1993 *Industrial Relations and European State Traditions*. Clarendon Press, Oxford, UK

DiMaggio P J, Powell W W 1991 The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. In: Powell W W, DiMaggio P J (eds.) *The New Institutionalism in Organizational Analysis*. University of Chicago Press, Chicago, pp. 63–82

Hollingsworth J R, Schmitter P C, Streeck W (eds.) 1994 *Governing Capitalist Economies: Performance and Control of Economic Sectors*. Oxford University Press, New York

Iversen T, Pontusson J 2000 Comparative political economy: A Northern European perspective. In: Iversen T, Pontusson J, Soskice D W (eds.) *Unions, Employers, and Central Banks: Macroeconomic Coordination and Institutional Change in Social Market Economies*. Cambridge University Press, Cambridge, UK, pp. 1–37

Knoke D, Pappi F U, Broadbent J, Tsujinaka Y 1996 *Comparing Policy Networks: Labor Politics in the US, Germany, and Japan*. Cambridge University Press, Cambridge, UK

Kooiman J 1993 Social-political governance: Introduction. In: Kooiman J (ed.) *Modern Governance: New Government–Society Interactions*. Sage, London, pp. 1–9

Lehmbruch G 1974 Consociational democracy, class conflict, and the new corporatism. In: IPS Association (ed.) *Round Table on 'Political Integration'*. Jerusalem

Lehmbruch G 1984 Concertation and the structure of corporatist networks. In: Goldthorpe J (ed.) *Order and Conflict in Contemporary Capitalism*. Oxford University Press, Oxford, UK, pp. 60–80

Schmitter P C 1974 Still the century of corporatism? *Review of Politics* **36**: 85–131

Schmitter P C 1979 Modes of interest intermediation and models of societal change in Western Europe. In: Schmitter P C, Lehmbruch G (eds.) *Trends Toward Corporatist Intermediation*. Sage, Beverly Hills, CA, pp. 63–94

Schmitter P C 1981 Interest intermediation and regime governability in contemporary Western Europe and North America. In: Berger S (ed.) *Organising Interests in Western Europe: Pluralism, Corporatism and the Transformation of Politics*. Cambridge University Press, Cambridge, UK, pp. 287–330

Shonfield A 1965 *Modern Capitalism: The Changing Balance of Public and Private Power*. Oxford University Press, London

Soskice D 1990 Wage determination: The changing role of institutions in advanced industrialized countries. *Oxford Review of Economic Policy* **6**: 36–61

Truman D B 1951 *The Governmental Process: Political Interests and Public Opinion*. Knopf, New York

G. Lehmbruch

Corpus Linguistics

Corpus linguistics encompasses the compilation and analysis of collections of spoken and written texts as the source of evidence for describing the nature, structure, and use of languages. Developing out of a long tradition of using texts rather than introspection, intuition, or elicitation as the empirical basis for linguistic description and argumentation, corpus linguistics is a methodology or approach for undertaking linguistic description. It is not a theory of language or a separate branch of linguistics such as morphology or sociolinguistics. Indeed, by providing sources of authentic, not made-up, examples of linguistic items, structures, or processes, the use of corpus-based evidence can complement other methodologies used by linguists. This evidence can be used for testing hypotheses, validating the accuracy of descriptions, making generalizations, and answering research questions which arise from theories of language or practical language-related issues in society.

With texts in electronic form as the source of linguistic data, and the speed and reliability of computers to assist analysis, the particular contribution of modern corpus linguistics has been to bring an additional quantitative dimension to linguistic description. On the basis of their occurrence in corpora, linguistic features can be shown to have different probabilities of occurrence in a language according to genres or domains of use. Thus, for example, Francis and Kučera (1982) showed that in US written English almost 15 percent of the verb predications in informative prose occur in the passive voice whereas in fiction, only about 4 percent of the predications are passive. Further, even within informative prose genres there can be considerable variation, with press reportage averaging about 13 percent passives and academic writing across many subject fields averaging 22 percent. Passive voice sentences can occur with an agent (e.g., '*These results were anticipated by the market*'). However, in texts from academic sources over 85 percent of passive voice predications typically occur without an agent. Corpus-based linguistic descriptions typically show not only what items, structures, or processes are possible, but which ones are probable in particular domains of use, thus inviting explanations of why this should be so. It can be argued that by including a quantitative dimension as part of descriptive adequacy, corpus linguistics has helped bring to the study of language a concern for distribution which is characteristic of other sciences.

1. Pre-electronic Corpora

Although corpus-based linguistics since the 1960s has been associated with texts stored and analyzed on computers, the use of texts as the basis for linguistic

description goes back well before electronic computers became available (see Kennedy 1998). In particular we may note the use of texts in lexicography, the study of the meaning and use of words. For example, the first edition of the *Oxford English Dictionary*, which was completed in 1928 after 70 years' work, was based mainly on the analysis of a collection of citations, totaling perhaps 50 million words, from texts from the canon of English literature, representing the use of written English over a period of 800 years.

Educationalists and linguists in the USA also used corpora of up to 18 million words in the first half of the twentieth century to discover the most frequently used words in English so as to develop better curricula for improving literacy education. This work of Thorndike, Sapir, and others was also very influential in the development of language teaching methodology. Their work was carried further in the UK by West and others and had a major impact on the methodology used for the teaching of English to speakers of other languages in substantial parts of Africa and Asia.

The first corpus compiled on a principled basis to be 'representative' of a language was the *Survey of English Usage Corpus* (SEU). Begun in1959 the SEU consists of 200 samples, each of 5,000 words, from some 24 genres of spoken and written British English. Although the corpus was not initially in machine-readable form, the spoken half of the SEU was eventually made available in electronic form as the *London–Lund Corpus*, until recently the largest available corpus of transcribed spoken language. The SEU marked the beginning of modern corpus linguistics and led to a large number of studies of English, culminating in the most extensive modern description of English grammar—*A Comprehensive Grammar of the English Language* (Quirk et al. 1985).

2. Computer-based Corpus Linguistics

Computers have given new impetus to corpus linguistics. The first computerized corpus designed for linguistic research was the *Brown Corpus* of contemporary written American English, compiled by Francis and Kučera. The one-million-word *Brown Corpus* consists of 500 text samples, each of 2,000 words representative of material printed in the USA in the year 1961. It became available in 1964. The careful design of the *Brown Corpus* with texts sampled from some 42 genres of informative and imaginative prose, the speed with which it was compiled and the way it was made freely available for linguistic research set new standards. It is noteworthy that although this first electronic corpus was compiled in the USA and there was an academic tradition there of text-based linguistic analysis and educational applications, most of the major developments in corpus linguistics for the next three and half decades came from Europe. These included the compilation of corpora for linguistic research, software development, and published research. This was partly because the direction of linguistic theory and research in the USA became dominated from the early 1960s by the influence of Noam Chomsky and attention shifted away from the observation of linguistic behavior ('performance') to a more epistemological focus ('competence'), with the study of performance ceasing to be normative except within subdisciplines such as sociolinguistics.

Following the compilation of the SEU and *Brown* Corpora, one million words became for a time the typical corpus size. This was brought about partly because one-million-word corpora were found to provide a reliable basis for many types of linguistic description, and because initially there were limitations in computer processing capacity and difficulty in getting machine-readable versions of texts before electronic text storage and optical scanning became more readily available. Until the mid-1980s most corpora had to be compiled through laborious keyboarding. A number of one-million-word corpora with the same basic design as the *Brown Corpus* were compiled in the 1970s and 1980s. These included, for example, the *Lancaster–Oslo–Bergen Corpus* (LOB) of written British English, the *Kolhapur Corpus* of written Indian English, and the *Wellington Corpus of Written New Zealand English*. All of these corpora and others of similar size compiled for synchronic or diachronic linguistic research purpose are available on CD-ROM from the International Computer Archive of Modern and Medieval English (ICAME) which also distributes the most complete bibliographic resources on corpus-based descriptions of English.

Although a high proportion of corpus-based research has been on English, there are now hundreds of corpora in existence, not only in English but in many other languages. Some have been compiled for commercial purposes such as dictionary making. Many others are available for research purposes and can be analyzed on personal computers (see Biber et al. 1998). Further information about corpora can be found in the journals *International Journal of Corpus Linguistics* and *Computational Linguistics* and through organizations such as the Linguistic Data Consortium and the European Corpus Initiative (see Bibliography).

In addition to the central focus on linguistic description, the study of the phonology, morphology, syntax, and discourse structure of languages, contemporary work in corpus linguistics has focused on four other main areas of activity:

(a) corpus design and compilation;

(b) the development of automatic grammatical annotation of corpora by means of word-class tagging and parsing to assign constituent structures;

(c) methodology for linguistic analysis using corpora; and

(d) applications of corpus-based linguistic descrip-

Table 1
Structure of the *International Corpus of English* (500 samples, each of 2,000 words)

Nature of texts	No. of texts	Nature of texts	No. of texts
Spoken		Written	
Dialogue		Nonprinted	
Private		Nonprofessional writing	
Direct conversations	90	student untimed essays	10
Distanced conversations	10	student examination essays	10
Public		Correspondence	
Class lessons	20	social letters	15
Broadcast discussions	20	business letters	15
Broadcast interviews	10	Printed	
Parliamentary debates	10	Informational (learned)	
Legal cross-examination	10	humanities	10
Business transactions	10	social sciences	10
Monologue		natural sciences	10
Unscripted		technology	10
Spontaneous commentaries	20	Informational (popular)	
Unscripted speeches	30	humanities	10
Demonstrations	10	social sciences	10
Legal presentations	10	natural sciences	10
Scripted		technology	10
Broadcast news	20	Informational (reportage)	
Broadcast talks	20	press news reports	20
Speeches (not broadcast)	10	Instructional	
Total	300	administrative/regulatory	10
		skills/hobbies	10
		Persuasive	
		press editorials	10
		Creative	
		novels/stories	20
		Total	200

tions to assist the development of natural language processing and language teaching.

(a) Corpora which are designed for linguistic analysis differ according to the purpose for which they are compiled. Some are highly structured and attempt, by judicious sampling in different domains of use, to be representative of a language or language variety. Such general or 'balanced' corpora are used for making generalizations about a wide variety of linguistic phenomena.

Corpora are also compiled for particular specialized research, such as how speakers bid for turns in conversation, how language develops in children, how teachers use directives in classrooms, the vocabulary and grammatical structures most frequent in academic writing in the social sciences, telephone conversations among familiars, or how a language such as English has developed across the centuries. Corpora are compiled in different ways, with some consisting, for example, of samples from many texts, others consisting of whole texts, the complete works of an author, or all the issues of a newspaper over decades (see Kennedy 1998). By way of illustration, Table 1 shows the structure of one of the most recent and ambitious structured corpus projects, the *International Corpus of English* (*ICE*). This project, when completed, will include 20 one-million-word corpora of English (60 percent from spoken sources) compiled in different parts of the world from texts produced in the 1990s. In addition to the whole corpus forming a representative sample of English as an international language, the 20 individual subcorpora can be the basis of comparative studies of regional varieties of English. The British section of ICE, the first section publicly available, comes equipped with sophisticated search and retrieval software for automatic lexical and grammatical analysis.

For lexicographical purposes and the study of low-frequency words or processes, however, corpora need to be much larger than one million words, in order to provide enough instances of individual items to enable valid and worthwhile generalizations to be made. One of the most notable large structured corpora is the *British National Corpus* (BNC) This corpus of 100 million words (10 percent spoken sources) is carefully designed to represent British English in the 1990s. it has been sponsored by a consortium of government, leading universities, and publishers and has brought

significant changes to our understanding of the lexical, grammatical, and discoursal resources of British English.

However, size or structure does not necessarily make a corpus representative. A major issue in corpus linguistics has been whether corpora should be structured at all. Corpora such as the *Bank of English* are huge, heterogeneous and rapidly expanding collections of texts of all kinds added to the corpus as opportunity arises. Such 'dynamic' or 'monitor' corpora do not have a particular structure or finite size. In some ways, they resemble the even larger text databases of parliamentary proceedings, literary works or newspaper archives now available which can also be used for linguistic research. The rapid expansion of the Internet also provides potential sources of text for certain types of linguistic description. For example, where it is difficult to find enough tokens of low frequency words in a corpus to provide reliable lexicographical description, Internet search engines can often provide more than enough. In the final analysis, because any corpus is useful only so far as it makes possible answers to particular questions, corpora may be best categorized by the uses to which they can be put rather than by their design features.

(b) The automatic grammatical annotation of texts has become increasingly important. Specifically, the identification and counting of words, collocations and grammatical structures in a corpus can be greatly facilitated when variant forms of the same word (e.g., *book/books*, *go/went*) are represented by a single form ('lemmatization') and when all words in a corpus are tagged to show their appropriate word class, and parsed to show their grammatical function as part of the constituent structure of sentences. Such text annotation not only serves linguistic description but can be a prerequisite for other applications of corpus analysis associated with natural language processing.

(c) Particular techniques and procedures have come to be associated with corpus-based analysis. To some extent, available computer software has influenced these methodologies. In addition to producing word lists showing the frequency of occurrence of each word type, corpus-based analysis has depended heavily on concordancing. A concordance is a list of all the occurrences ('tokens') of a particular word or phrase in a corpus in the context of a given number of preceding and following words. Concordancing software has led to a focus on words and on 'collocations,' the company a word typically keeps. For example, linguists have noted that English intensifiers such as *very*, *quite*, *pretty*, *terribly* have a tendency to precede particular adjectives (e.g., *terribly nice*, *pretty horrible*). Corpus-based research on collocations has led increasingly to the recognition that a high proportion of what people say or write consists of prefabricated 'formulaic' groups of words (e.g., *pouring rain*, *strong tendency*, *at the moment*), thereby blurring somewhat the distinction between lexis and grammar.

(d) Important applications of corpus linguistics have occurred in the fields of natural-language processing (NLP) and second-language teaching. Research has been undertaken by computational linguists and others to make progress on the automatic lemmatization, tagging, and parsing of text. Rule-based approaches to annotation have sometimes been used, but increasingly some of the most promising NLP research has employed corpus-based statistical information on the probability of occurrence of linguistic items or processes as a basis for advances in automatic word-class tagging, parsing and disambiguation. Such information is prerequisite for developments in the retrieval of information from databases, speech recognition, speech synthesis, and automatic language translation.

Within second language teaching for the last 80 years there has been a continuing but by no means universal concern among curriculum designers and teachers as to which linguistic items or processes are most likely to be needed by language learners. In English for example, fewer than 100 words typically account for half of all the words used, and about 3,000 words account for 90 percent. According to Biber et al. 1999, just 12 English verbs account for nearly 45 percent of all lexical verbs in conversation. Many language teachers have known since the 1960s that although English verb phrase morphology appears complex and fraught with possibilities, the simple present and simple past tenses account for over 80 percent of all verb phrase structures in most corpora. Similarly, more recent work has shown that there is a much greater tendency for *prepositional phrases rather than relative clauses to be used in English postnominal modification* (e.g., *The person in the car was alone* is likely to be much more frequent than *the person who was in the car was alone*).

3. *Current Issues and Directions*

Corpus research has resulted in descriptive studies of all levels of language. Some of the most innovative work has been on linguistic variation. Biber (1988), for example, pioneered the corpus-based study of how the distribution of linguistic items in texts and the patterns of co-occurrence of these items could be used to characterize the ways in which languages vary diachronically and synchronically in different regional and sociolinguistic contexts. However, frequency of occurrence is not an end in itself. With easier availability of corpora, more powerful computers and more sophisticated software, corpus linguists have new opportunities to move beyond describing the structure and quantitative distribution of linguistic features. Some of the best work in corpus linguistics continues to involve the judicious use of manual analyses supported by the computer. There are opportunities for corpus linguists to illuminate the signifi-

cance of distribution, and to use corpora to test hypotheses and the robustness of linguistic theories. In focusing on the ways in which language users habitually express themselves, corpus linguistics has already thrown new light on how languages vary systematically in different regional and sociolinguistic contexts, genres, and registers. There is potential to provide further insights on the speed and direction of language change and on the factors which contribute to change. Corpus linguistics may also come to give greater support to work in cognitive linguistics through further exploration of the semantic roles and functions of linguistic features.

See also: Database, Core: Linguistics and CHILDES; Databases, Core: Linguistics, Online Grammars, and Dictionaries

Bibliography

Aijmer K, Altenberg B (eds.) 1991 *English Corpus Linguistics*. Longman, London

Biber D 1988 *Variation across Speech and Writing*. Cambridge University Press, Cambridge, UK

Biber D, Conrad S, Reppen R 1998 *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, Cambridge, UK

Biber D, Johansson S, Leech G, Conrad S, Finegan E 1999 *Longman Grammar of Spoken and Written English*. Longman, London

European Corpus Initiative. http://www.cogsci.ed.ac.uk/elsnet/eci.html

Francis W N, Kučera H 1982 *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston

Garside R, Leech G, Sampson G (eds.) 1987 *The Computational Analysis of English: A Corpus-based Approach*, Longman, London

Greenbaum S (ed.) 1996 *Comparing English Worldwide: The International Corpus of English*. Clarendon Press, Oxford, UK

International Computer Archive of Modern and Medieval English (ICAME). http://nora.hd.uib.no./whatis.html

Johansson S, Hofland K 1989 *Frequency Analysis of English Vocabulary and Grammar*. Clarendon Press, Oxford, UK

Johansson S, Stenström A-B 1991 *English Computer Corpora: Selected Papers and Research Guide*. Mouton de Gruyter, Berlin

Karlsson F, Voutilainen A, Heikkilö J, Antilla A (eds.) 1995 *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin

Kennedy G 1998 *An Introduction to Corpus Linguistics*. Longman, London, UK

Linguistic Data Consortium. http://www.ldc.upenn.edu

Quirk R, Greenbaum S, Leech G, Svartvik J 1985 *A Comprehensive Grammar of the English Language*. Longman, London, UK

Scott M 1996 *WordSmith*. Oxford University Press, Oxford, UK

Sinclair J 1991 *Corpus, Concordance, Collocation*. Oxford University Press, Oxford, UK

Svartvik J (ed.) 1992 Directions in corpus linguistics. *Proceedings of Nobel Symposium* 82, Stockholm, 4–8 August 1991. Mouton de Gruyter, Berlin

Thomas J, Short M (eds.) 1996 *Using Corpora for Language Research*. Longman, London

Tottie G 1991 *Negation in English Speech and Writing*. Academic Press, London

G. Kennedy

Correspondence Analysis

Correspondence analysis (CA) handles research data that have the form of rectangular tables containing indications of association strength between row entries and column entries (correspondence tables). The association measure is assumed to be some non-negative quantity. Cells of a correspondence table may contain transition or confusion frequencies (forming a contingency table), binary choices among dyads of persons, for example in sociometric studies, or preference judgments for options by experts on a rating scale from weak to strong endorsement. In other words, response scales have to be unipolar, ranging from zero, indicating absence of association, to some maximal positive value, indicating strongest possible association. Although the scope of application of CA is very broad, it has been most typically used and developed as a method for analyzing contingency tables. Hence some core concepts refer to the frequency domain, even though other justifications of the method exist.

1. Row and Column Profiles

As an illustration consider the eight discrete distributions in Table 1(a) taken from van Ijzendoorn and Kroonenberg (1988). Several studies using the Strange Situation paradigm (Ainsworth et al. 1978) in various countries seemed to show marked cross-cultural differences in distributions of attachment classifications. Most of the studies used a classic three-fold attachment classification into Avoidant (A), Secure (B), and Resistant (C). These three types characterize the enduring socio-emotional bond between child and regular care-giver. To examine the cross-cultural heterogeneity more closely, the authors selected 32 samples from eight countries, representing 1,990 Strange Situation classifications, which have been aggregated in Table 1 to the level of countries. To study similarities and differences of distributions regardless of sample size, it is necessary to divide frequencies by their row total (see Table 1(b)). A row of proportions is called a row profile; analogously, a column profile is formed by dividing the cell frequencies in a column by the column total. The natural comparison standard to assess a row profile against is the marginal profile of the columns, that is, each of the

Table 1
Attachment classifications in eight countries

	(a) Frequencies				(b) Row-proportions		
Country	A	B	C	Total	A	B	C
Germany	48	77	11	136	0.35	0.57	0.08
Great Britain	16	54	2	72	0.22	0.75	0.03
Netherlands	66	169	16	251	0.26	0.67	0.06
Sweden	11	38	2	51	0.22	0.74	0.04
Israel	8	76	34	118	0.07	0.64	0.29
Japan	5	65	26	96	0.05	0.68	0.27
China	9	18	9	36	0.25	0.50	0.25
United States	260	797	173	1230	0.21	0.65	0.14
Total	423	1294	273	1990	0.21	0.65	0.14

Legend: A = avoidant; B = secure; C = resistant.

column totals divided by n, the grand total. Similarly, column profiles are first compared with the marginal profile of the rows, that is, each of the row totals divided by n. From the row profiles it is immediately clear that Germany has a relatively large proportion of Avoidant, for example, Great Britain and Sweden have an overrepresentation of Secure and an underrepresentation of Resistant, while Israel, Japan, and China have an overrepresentation of Resistant.

The aim of CA is to assess similarities and differences of row profiles (and/or column profiles), not only with respect to their marginal profile, but also amongst each other, by constructing a parsimonious approximation of the cell frequencies. The approximation consists of decomposing the correspondence table into one or more components, which through graphical display allows a deeper insight into the original categories by suggesting specific groupings, orderings, or more complicated arrangements.

2. *Deviations from Independence*

The decomposition of CA has two major parts: one consisting of a multiplicative combination of row and column proportions, the usual form used in studying statistical independence, the other consisting of row–column interaction effects. To be more explicit, the following notation is introduced.

The frequency in cell (i, j) of the contingency table is denoted by n_{ij}, where the categories of the row variable A have range $i = 1, \ldots, I$ and the categories of the column variable B have range $j = 1, \ldots, I$, and the categories of the column variable B have range $j = 1, \ldots, J$. Row totals are defined as $n_{i+} = \Sigma_j n_{ij}$ and column totals as $n_{+j} = \Sigma_i n_{ij}$. We also have the grand total $n_{++} = \Sigma_i n_{i+} = \Sigma_j n_{+j}$, frequently denoted simply by n. To keep the results invariant under changes in sample size, CA works with the quantities $p_{ij} = n_{ij}/n$, each of which is the (estimated) probability mass in the cell (i, j) of the bivariate distribution of A and B, $r_i = n_{i+}/n$, the relative mass of row profile i, and $c_j = n_{+j}/n$, the relative mass of column profile j. The CA decomposition is based on the tautologous identity

$$p_{ij} = r_i c_j (1 + q_{ij}) \tag{1}$$

where the quantities q_{ij} are defined as $q_{ij} = (p_{ij} - r_i c_j)/(r_i c_i)$. The q_{ij} are standardized deviations from independence: if the two variables A and B are independent, we have $p_{ij} = r_i c_j$ and every q_{ij} would vanish. Dependences between A and B show up specifically in the q_{ij}: categories i and j are more strongly associated than expected under independence when $q_{ij} > 0$, and less strongly when $q_{ij} < 0$. Spatial and graphical representations that result from CA are representations of the deviations $[q_{ij}]$, not of the data $[p_{ij}]$ themselves.

3. *Basic Geometric Model*

In the CA model each row profile initially is a point $\mathbf{a}_i$ in a subspace of dimensionality $J-1$, with coordinates $a_{ij} = n_{ij}/n_{i+}$. Analogously, each column profile is a point $\mathbf{b}_j$ with coordinates $b_{ij} = n_{ij}/n_{+j}$ in another subspace, of dimensionality $I-1$. The total spread of the points in space is measured by their inertia $\mathscr{I}$, a concept from physics defined as (for the row points)

$$\mathscr{I} = \sum_i r_i \sum_j \frac{1}{c_j} (a_{ij} - c_j)^2 = \sum_i r_i \|\mathbf{a}_i - \mathbf{c}\|^2_{C^{-1}} \tag{2}$$

where $\mathbf{c}$ is the point with coordinates c_j that represents the marginal profile of the columns. The function $\|\cdot\|_{C^{-1}}$ defined implicitly in Eqn. (2) is a Euclidean norm with weights $1/c_j$. It downweights column differences associated with large column mass, because large frequencies tend to be less reliable. There is a parallel expression of inertia for column points, which can be

shown to yield the same numerical value of $\mathscr{I}$. In both cases, large overall differences among the distributions will lead to a spread-out configuration of points with large $\mathscr{I}$, while small overall differences will correspond to a tight concentration of points around their center of gravity, with small $\mathscr{I}$.

Dividing the usual chi-squared statistic by n, so that lack of homogeneity (or independence) is measured regardless of sample size, we have the equality

$$\chi^2/n = \sum_i \sum_j \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}} = \mathscr{I} \tag{3}$$

In statistics, χ^2/n is known as Pearson's mean square contingency. The relation in Eqn. (3) between the chi-squared statistic χ^2 and the weighted sum of squared distances towards the center of gravity $\mathscr{I}$ defined in Eqn. (2) is the reason that one speaks of heterogeneity being measured in the chi-squared metric $\|\cdot\|_{C^{-1}}$.

Mutual distances between row profiles are determined in the same chi-squared metric, and hence called chi-squared distances. For two profiles i and k we have

$$d_{\chi^2}(\mathbf{a}_i, \mathbf{a}_k) = \|\mathbf{a}_i - \mathbf{a}_k\|_{C^{-1}} = \sqrt{\sum_j \frac{n}{n_{+j}} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{kj}}{n_{k+}} \right)^2} \tag{4}$$

with again an analogous expression for the column points. In Eqn. (4) columns with small marginal frequency are seen to contribute more to overall lack of similarity than columns with large marginal frequency. For an exhaustive treatment of the geometry of CA, see Benzécri (1992, pt. I).

4. Weighted Least Squares Approximation

Low-dimensional CA approximations of the full-dimensional model are obtained by least squares, where each cell is weighted by the product of the row and column masses (see Andersen 1990). The least squares solution consists of two sets of standardized coordinates, $z_{it}(A)$ and $z_{jt}(B)$, where t indexes the components, with t having range $t = 1, \ldots, \min(I-1, J-1)$. The components are ordered by the size of a third set of quantities σ_t, called singular values. Standardized coordinates $z_{it}(A)$ have the property $\Sigma_i r_i z^2_{it}(A) = 1$; analogously, for $z_{jt}(B)$ we have $\Sigma_j c_j z^2_{jt}(B) = 1$. Also, the coordinates are in deviation from their weighted mean and uncorrelated across different values of t. The singular values σ_t indicate the relative importance of component t in the approximation of $[q_{ij}]$. Practical CA solutions are obtained by dropping components associated with small singular values.

Graphical displays are based upon a mapping from $\mathbf{a}_i$ to $\mathbf{y}_i(A)$, a model point with principal coordinates defined as $y_{it}(A) = \sigma_t z_{it}(A)$, with $t = 1, \ldots, T$ (some chosen dimensionality). Principal coordinates for column points are $y_{jt}(B) = \sigma_t z_{jt}(B)$. For maximal dimensionality $T^* = \min(I-1, J-1)$ it can be shown that

$$\|\mathbf{a}_i - \mathbf{a}_k\|_{C^{-1}} = \|\mathbf{y}_i(A) - \mathbf{y}_k(A)\| \tag{5}$$

i.e., Euclidean distances among model points are equal to chi-squared distances among profile points. In approximations with $T < T^*$ the former are always smaller than latter (see Meulman 1982).

5. Some Important Additional Concepts

The CA solution for the attachment data is shown in Fig. 1, in which the row points are plotted as open circles. Since in this example we have $T^* = 2$, the model exactly reproduces the data ($\chi^2 = 102.42$ with $df = 14$, $\mathscr{I} = 0.051$). Approximation in $T = 1$ implies projecting all points on the horizontal axis, which here accounts for 86 percent of the inertia (see Fig. 2). With the US and China forming the center, the largest contrast in both solutions is between the Western European countries vs. Israel and Japan, while Fig. 1 shows a second contrast between China and Germany versus Great Britain and Sweden related to deviations in their Secure classifications.

5.1 Transition Formulas

Row points and column points reside in two different but intimately related spaces. Their coordinates are connected by the equations

$$y_{it}(A) = \frac{1}{r_i} \sum_j p_{ij} z_{jt}(B) \tag{6}$$

$$y_{jt}(B) = \frac{1}{c_j} \sum_i p_{ij} z_{it}(A) \tag{7}$$

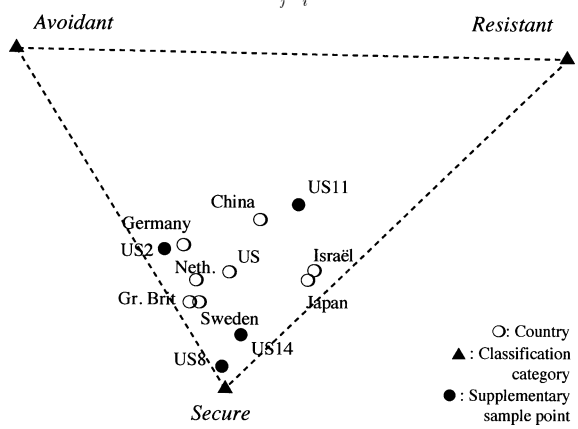

Figure 1
Full-dimensional CA model of countries-by-attachment-classification data, including four supplementary sample points (countries in principal coordinates, categories in standardized coordinates)

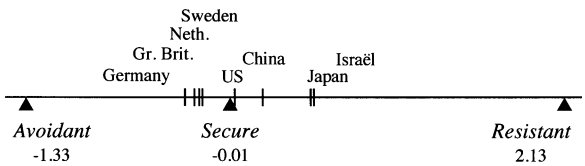

Figure 2
One-dimensional approximation of countries-by-attachment-classification data, accounting for 86 percent of the inertia (countries in principal coordinates, categories in standardized coordinates)

These equations are called transition formulas, since they show how the principal coordinates, e.g., $y_{it}(A)$, of one set of points can be obtained from the standardized coordinates, e.g., $z_{jt}(B)$, of the other set of points: by weighted averaging with the data as coefficients. If the correspondence table is binary, principal coordinates are plain averages of standardized coordinates, and CA becomes equivalent to the method of reciprocal averages, which has firm roots in psychometrics (Horst 1935, Guttman 1941), with continued research under the name dual scaling (Nishisato 1994).

5.2 Joint Plot and Coherent Normalization

In Fig. 1 the country points are plotted in principal coordinates, so that their Euclidean distances are equal to the chi-squared distance of Eqn. (5). The distances between the country markings along the line in Fig. 2 approximate these chi-squared distances rather closely. The plots also show the three attachment classes in standardized coordinates. Superposition of row and column points forms a joint plot, also called a 'biplot,' which is easily interpretable due to the transition formulas. For example, Germany is located relatively close to Avoidant, because its location is a weighted average of the three attachment points, weighted with [0.35, 0.57, 0.08] rather than with the average profile [0.21, 0.65, 0.14], which forms the center of the plot. It is also possible to (re)scale the coordinates so as to make the two clouds of points more evenly mixed, but this has to be done with some care (Heiser and Meulman 1983). Rescaling is coherent if we choose points $\mathbf{x}_i(A)$ and $\mathbf{x}_j(B)$ with coordinates $x_{it}(A) = \sigma_t^{\alpha} z_{it}(A)$ and $x_{jt}(B) = \sigma_t^{1-\alpha} z_{it}(B)$ for any α, because it remains true that $\mathbf{x}_i(A)^{\mathrm{T}}\mathbf{x}_j(B) = q_{ij}$. But it should be noted that we no longer approximate chi-squared distance unless $\alpha = 0$ or $\alpha = 1$.

5.3 Supplementary Points

It can be useful to add new profiles to an existing joint plot. These supplementary points are easily obtained if scaling is chosen as $\alpha = 0$ or $\alpha = 1$, since then Eqns. (6) or (7) can be used to find coordinates. In Fig. 1, points were added for four divergent individual US studies, showing that there is at least as much variation within the US as between countries. While US2 closely resembles the total German group, US11 is much more like China, for example.

6. Extensions and Applications

The case of more than two variables, called multiple CA, was pioneered by Guttman (1941). Under the name 'homogeneity analysis,' Gifi (1990) generalized Guttman's technique to ordinal data, and introduced extensions based on partitioning of variables. Greenacre (1984) and Lebart et al. (1984) bridged the gap between the early French work and the inter national literature. The edited volume of Greenacre and Blasius (1994) contains social science applications, versions of CA with various constraints, and versions that achieve clustering of categories.

See also: Multidimensional Scaling in Psychology; Multivariate Analysis: Discrete Variables (Correspondence Models); Multivariate Analysis: Discrete Variables (Overview); Scaling: Correspondence Analysis; Scaling: Multidimensional

Bibliography

Ainsworth M D S, Blehar M C, Waters E, Wall S 1978 *Patterns of Attachment, a Psychological Study of the Strange Situation*. Erlbaum, Hillsdale, NJ

Andersen E B 1990 *The Statistical Analysis of Categorical Data*. Springer, Berlin

Benzécri J-P 1992 *Correspondence Analysis Handbook*. Dekker, New York

Gifi A 1990 *Nonlinear Multivariate Analysis*. Wiley, New York

Greenacre M J 1984 *Theory and Applications of Correspondence Analysis*. Academic, London

Greenacre M, Blasius J (eds.) 1994 *Correspondence Analysis in the Social Sciences*. Academic, London

Guttman L 1941 The quantification of a class of attributes: A theory and method of scale contribution. In: Horst P et al. (eds.) *The Prediction of Personal Adjustment*. Social Science Research Council, New York, pp. 319–48

Heiser W J, Meulman J J 1983 Analyzing rectangular tables with joint and constrained multidimensional scaling. *Journal of Econometrics* **22**: 139–67

Horst P 1935 Measuring complex attitudes. *Journal of Social Psychology* **6**: 369–74

Lebart L, Morineau A, Warwick K M 1984 *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. Wiley, New York

Meulman J J 1982 *Homogeneity Analysis of Incomplete Data*. DSWO, Leiden, The Netherlands

Nishisato S 1994 *Elements of Dual Scaling: An Introduction to Practical Data Analysis*. Erlbaum, Hillsdale, NJ

van Ijzendoorn M H, Kroonenberg P M 1988 Cross-cultural patterns of attachment: a meta-analysis of the strange situation. *Child Development* **59**: 147–56

W. J. Heiser

Corruption: Political and Public Aspects

Although there is fairly general agreement that corruption is a trans-systemic phenomenon endemic to all societies, regimes, and countries, and salient in different periods of the history of mankind, the study of corruption reveals a weaker consensus as to its definition, measurement, and explanation. Empirical studies constantly are faced with three major issues that may hinder further comparative understanding of the phenomenon. First, corruption lacks a precise definition or agreement upon its meaning. Like many other concepts in social sciences, corruption is volatile and subject to historical, social, and cultural connotations. Second, corruption is susceptible to variations across time and space. Its measurement, both in terms of volume as well as the standards qualifying that form of behavior, is prior to any attempt at explaining how the phenomenon expands, grows and mutates contextually. Third, without clearing the first two methodological steps, the empirical testing of a derived hypothesis is bound to show rather limited or disputed results.

1. *Definition*

Idiomatic definitions differ substantially from country to country, a problem first raised by anthropologists, who denoted variations resulting from different social, religious, and cultural backgrounds. Corruption as a political phenomenon adds to that definitional complexity. The different usages of corruption by political scientists, with a broader meaning than those present in most contemporary penal codes, have raised potential ambiguities.

1.1 *Contemporary Social Sciences Definitions*

One could group this rich diversity in three basic definitional types: public-office-centered, market-centered, and public-interest-centered.

1.1.1 Public-office-centered definitions. Traditionally, the largest group of political scientists has opted for a definition focusing on those practices/behaviors that violate the legal/formal standards regulating public office. The most prominent definition is that Nye: 'Corruption is behavior which deviates from the normal duties of a public role because of private-regarding …, pecuniary or status gains; or violates rules against the exercise of certain types of private-regarding influence. This includes such behavior as bribery … nepotism … and misappropriation …' (1967).

Public-office definitions may help to disentangle the complexity of corruption by restricting the variety of potential denominations to the (most) visible or condemnable types. In practice, the standards covered by this set of definitions are not too far from those stated in most Penal Codes. More often than not, penal reforms are a *post facto* attempt to deal with corruption. Rules governing public office have been revised or stiffened, when certain manifestations of corruption have become rampant, provoking widespread convulsion and contempt in political society.

Public-office definitions are stable and objective and, for those reasons, have attracted considerable popularity amongst political scientists and criminologists. Normative aspects are simplified by legal reductionism. Legal norms, it is argued, translate a tangible value system, that can be considered, *a priori*, as an expression of the mores and ethico-juridical structure of a society at a given time. But even then one cannot escape the interpretative differences and imperfections of legal norms and how these evolve through time. The question arises: 'Which statement of the rules and norms governing public office is to be employed?' For most authors, the normative discourse of legislators, judges, and other competing authorities should be used as a definitional criterion. Legal norms are most likely to reflect the ethical standards of a particular group—the ruling elite—hence socially disputable and susceptible to revision.

The usage of legalistic definitions highlights the weaknesses of black-and-white modes of conceptualization. If compliance with public-office rules relies on the degree to which these are embedded in the ethical environment of the society in question, then one is bound to end with a rigid definition of corruption unrepresentative of the ensemble of manifestations occurring in that legal/social context. As Sorokin remarked, '… There always is some discrepancy between the situation as it is depicted in the "official law" and in the psychosocial mentality of members of the society. And the discrepancy is the greater, the quicker the socioethical life of the society changes' Sorokin (1991). Public-office definitions may have an analytical advantage from other definitions as they ease comparison across countries, societies, or regimes, yet, they bear the cost of leaving too many things undefined and, consequently, unexplained.

1.1.2 Market-centered definitions. A more restricted group of academics, deriving their premises from economic theory, conceptualized corruption in terms of an exchange of money for political decisions (asset), which private actors seek to acquire

(demand), and public agents are just too willing to sell (supply) by avoiding being caught (liability). The definition embarks on a public choice assertion that, *ceteris paribus*, all agents will seek rationally to take an advantage from their privileged position in office (rent maximizing) whenever there is a window of opportunity. As van Klareven argues, 'A corrupt civil servant, regards his public office as a business … a "maximizing unit." The size of his income depends … upon the market situation and his talents for finding the point of maximal gain on the public's demand curve' Van Klareven (1989).

The definition is articulated convincingly when dealing with specific types of transaction corruption (bribery, kickback, *pots-de-vin*) in which the legal criterion of infringement–condemnation may not be so clear-cut or nonexistent. Rose-Ackerman, an exponent of this type of definition, stresses that market corruption is essentially about 'all payments (legal or illegal) to agents that are not passed on to superiors,' but includes also a series of 'legal activities with similar public policy consequences,' such as, 'campaign contributions and lobbying' Rose-Akerman (1978).

Definitions of this sort appear to bypass normative considerations, but in reality they do not. Corruption is seen essentially as those exchanges, often between public officials and private entrepreneurs, likely of being susceptible to some disciplinary or criminal punishment. However, if corruption were defined as a rational behavior of maximization of personal advantages, then what would be the stimuli behind the acceptance of a pecuniary benefit that can be sanctioned by law? The definition also neglects 'moral costs' affecting the rationale of actors to engage in corruption such as public condemnation, disapproval by superiors and colleagues in the working place, and individual ethics. Some acknowledge the existence o'moral costs of breaking the law' (Rose-Ackerman 1978), but fail to provide a clear reference of what those 'moral costs' are.

The reliance on market concepts undermines that rules and principles binding public and elected officials are considerably distinct from those applicable to market actors. Market(s) of corruption—a disputable concept (Cartier-Bresson 1995)—can be defined as parallel exchanges where scarce 'goods' are allocated at prices determined illegally, contrary to governmental regulations and affecting its public policies (i.e., 'black market'). Such definition may not even correspond to the mechanisms and concepts operating in a free market situation. The very designation of 'black,' i.e., obscure and sanctionable, runs counter to the 'freedom of choice' qualifying market economies.

1.1.3 Public-interest-centered definitions. In recent years, political scientists increasingly have felt that both these types of definition were narrowly conceptualized. On the one hand, market-centered definitions were useful in so far as one dealt with transaction corruption, but insufficient once manifestations referred to nonpecuniary social exchanges and their moral implications. On the other, there was an upsurge against the tendency to reduce corruption to public-office rules applicable to a limited number of practices/behaviors, most likely criminal. According to these authors, corruption is a deviant behavior or practice characterized essentially by the subjugation of public to private interests (Rogow and Lasswell 1963, Mény 1992). The classical concept of public interest is revived and argued vividly as a criterion not only useful, but also indispensable to illustrate the essence of corruption. The public interest invites men to overcome their selfishness, to confront their preferences with those of others, and to evaluate options with equal disinterest and impartiality.

Moral issues are taken into account. Friedrich wrote, 'Corruption, being in fact a decomposition of the body politic through moral decay, is a general category to include all kinds of practice which are believed … morally corrupt' Friedrich (1989). The immediate questions are 'whose morals?' determine 'what corruptions?' In order to avoid associating corruption as value system degeneration with the classical idea of a transcendental and pre-established set of mores, the concept is made tangible through the predominant public perceptions in society. Morality is regarded as a social construction and what is or is not corruption will depend on the observer's opinion. However, replacing 'public interest' by a more fluid and vague concept of 'public opinion' does not prove elucidating either. The usage of public opinion pertains to the conceptualization of corruption, but only in terms of its operationalization, in other words, as a soft criterion used to measure the nature and scope of corruption.

Critics insist that the public interest is a weak reference, since in most contexts there is no definition of what the concept stands for: the notion of public is difficult to ascertain, least to say, that interest is another jargon of pluralist theories of the State. The concept is a battled one and open to broad interpretations and for that reason tends to be discarded from comparative measurements. The notion of public interest—its values, perceptions, and laws—vary across societies and that has implications in terms of what manifestations of corruption are identified across cases. However, this should not constitute an obstacle to the explanation of why societies have different standards defining corrupt practices/behavior or why these standards change through time.

1.2 Operationalization

The plurality of definitions available, sometimes controversial, may become an obstacle to empirical research. An alternative is to operationalize the con-

cept with the utmost goal of enhancing broad-based comparisons. This is not an easy task as most measures do not 'travel' easily in space and time. Scott wrote, 'Corruption, we would all agree, involves a deviation from certain standards of behavior. The first question which arises is, What criteria shall we use to establish those standards?' Scott (1972). Empirical studies on corruption have too often relied on legal/formal criteria to qualify standards of corruption (Nye 1967, Scott 1972). These tend to be regarded as stable and accurate indicators of what are the standing standards of corruption at a given time slot and society. Yet, even legal standards are not monoliths. The concept of corruption as defined in most penal codes is no longer particularly distinct from other types of crimes, such as fraud, influence trafficking, peculate, extortion. There has been equally a tendency towards enlarging its definitional parameters in order to embrace a wider range of pervasive manifestations (e.g., the inclusion of nonpecuniary inducements, the distinction of passive and active initiatives, and licit or illicit outcomes).

Legal criteria are not a precise volume indicator either: most corrupt transactions are obscure and clandestine; and shifts in volume may be a product of greater or more efficient prosecution. By excluding public reaction from the conceptual framework of corruption one risks to fall into legal dichotomising (i.e., legal 'noncorrupt vs. illegal' corrupt). But not everything that is legal is approved morally by citizens (Gardiner 1992). A purely legal conceptualization would omit all corrupt behavior which does not necessarily imply a breach of law or of codes of conduct, but still involve 'a serious violation of the standards and expectations associated with a public role' (Friedrich 1966, Jos 1993).

Corruption is also about perceptions. It can be many things to many people and has been used incessantly as an offensive political weapon. Another aspect of public opinion condemnation is that of labeling a cluster of distinct practices/behaviors as corruption, independently of the standing legal standards. Greater public perception depends largely on 'how?' and 'what kind?' of manifestations has been made visible to the public. Public opinion is fluid and particularistic, susceptible of being politically mobilized. Hence, the understanding of the 'corruption of politics' passes through an attentive analysis of the 'politics of corruption' (Nelken and Levi 1996). For most times, public opinion is an insufficient criterion based upon unstable, vague, if not contradictory, standards which oscillate between diversified perceptions in society, but criticisms fall short of a flexible alternative to measure qualitative variations. The variety of social representations on corruption is a sign that there is no harmonious definition, but competing standards. There is no static definition, but a qualitative gradation in terms of public condemnation and there is no necessary concordance between social/cultural standards and the standing legal rules condemning corruption in society.

However, a purely public opinion-based conceptualization is not the solution either. In as much as public opinion is important to understand changes in ethico-juridical structure, the condemnation of corruption is not just a matter of perception. Public accusation goes beyond the allegations raised by prominent newspapers, the rhetoric of electoral campaigns, or that of fervent parliamentary debates, and condemnation can be translated into criminal prosecution.

Corruption cannot be defined exclusively by legal or public opinion criteria. Corruption as a deviant behavior/practice concerns both the breach of rules governing public office and the infringement of noncodified, widely accepted ethical norms (Gardiner 1992, Jos 1993). The concept has to be flexible enough to include both set of norms 'constituting a society's system of public order' and the way these two ethical components are being challenged continuously and revisited (Johnston 1996).

Heidenheimer's (1989) definition of black, gray, and white types of corruption in relation to a variation of an elite- and public-based standards attempts to bridge both criteria. His typology is innovative in so far it denotes a certain qualitative gradation of the phenomenon product of value change. Some practices/behaviors, (often) collusive and subject to prosecution, are regarded as severe violations of the legal and moral norms set in society (black corruption). In this case, standing legal standards are likely to be reinstated by competent authorities when infringement is exposed. When toleration mounts, coding certain corrupt practices/behaviors 'a way of life,' legal condemnation gradually loses its meaning (white corruption). Standards degenerate. Neither of these two situations will necessarily trigger a revision of standards. It is rather the 'grayness' of corruption that promotes the demand for reform in the system, because existing standards are no longer seen as operational by political elites and sectors of public opinion. The waves of reform and control that swept across most Western European countries by the late 1980s and early 1990s can only be understood in the light of a significant value discrepancy between ruling elites and masses (gray corruption).

2. *The Measurement of Corruption*

The measurement of corruption goes hand in hand with the observation of qualitative and quantitative variations in space and time. Is country 'A' more corrupt than country 'B'? Is there more corruption at the end of the twentieth century than in the 1980s? Was illegal party financing equally condemnable as political corruption across different pluralist democracies? Political scientists increasingly are concerned

to come about with a measurement of corruption as a pre-analytical step towards a more comprehensive, empirically based, comparative explanation of corruption.

Priority has been given to the measurement of levels of corruption, but this was not without its problems. Attempts at measuring corruption across countries with significantly different legal and moral norms have been criticized as an academic tool of stereotyping, which political actors and others, at the national or international levels, are just too willing to have at their disposal as an offensive (or defensive) labeling weapon. The problem of stereotyping emerges at the level of operationalization, but it also has to do with the parameters of comparison (small n/large N). How to include the greatest number of cases without losing consistency and comparability? Different countries are governed by different rules and norms, so the immediate problem is: whose standards should be taken as reference to comparison?

This brings back the problematic that surfaced earlier comparative works: how to avoid 'cultural relativism' without falling into 'cultural imperialism?' Taking a lower denominator—each country's rules governing office or national perceptions of the phenomenon—makes it difficult to envisage meaningful generalisations about corruption. Cultural specificities are to be taken seriously; yet, this has often permitted a degree of toleration towards some forms of corruption regarded as beneficial to the societies in question. Taking Western values as an higher denominator, which other countries may try to emulate, raises doubts about the validity of results attained, either because those standards are not yet embedded in developing societies or because such *alta vista* neglects different social/cultural realities.

Transparency International (TI), a non-governmental organisation (NGO) devoted exclusively to the study and combat of corruption worldwide, has developed a corruption perception index based on several surveys conducted in as many as 85 countries (cf. http://www.transparency.de). The TI's index is the first attempt to measure levels of corruption, but its reliance on public opinion poses serious limitations: (a) there is a problem of operationalization (the index is concerned only with transaction corruption, namely, bribery of foreign officials); (b) there is a problem of sample (it relies on an homogeneous sector of public opinion, i.e., international businessmen, financial journalists); (c) it reinforces (or creates) stereotyping. Despite undeniable methodological weaknesses, political scientists have not precluded its use as a starting point for further qualification and understanding of corruption.

Unfortunately, there have been few attempts at measuring qualitatively standards of corruption and establishing comparative frameworks of control measures implemented and their effectiveness. After all, corruption as a policy issue is part of a cumulative learning process of past control experiences, their successes, and failures. Timescale measurements of qualitative variations equally have been lagging. The understanding of why corruption emerges in a particular conjuncture leading to widespread condemnation and revision of standards, or inversely, why after periods of enforced political morality, ethical standards shade into oblivion, has been mainly a task of some political historians and public policy analysts (Woodiwiss 1988, Mény 1993).

3. *The Evolution of Theoretical Approaches*

3.1 *The Corruption of the 'Good Polity'*

For the founders of Western political thought, such as Plato, Aristotle, and Thucydides, corruption expressed the degeneration of right regimes, those who rule for the 'good of the whole,' into perverted ones, those who rule for 'the good of themselves' (i.e., tyranny, oligarchy, demagogy). Whereas the conceptual overlap between the broader institutional decay and corruption, *stricto sensu*, was explained in terms of behaviors, practices, and laws that debased from the 'common good,' little was said about the moral components of that intangible axiom. This tradition persisted throughout the development of Western political thought, from Aquinas to Rousseau, from Montesquieu to Burke, from the twentieth century philosophers of *décadence* (Aron, Malraux) to the contemporary theorists of democracy (Dahl, Sartori). Their conceptual weakness was that of not establishing empirically what standards defined political corruption, because that was secondary to their theoretical concern with the broader moral corruption.

This contributed largely to its unpopularity in earlier comparative works, because it 'sacrificed clarity to brevity, in so far as it leaves too implicit how or why behavior is deviant from which norms' (Heidenheimer et al. 1989). The decolonization processes in the aftermath of World War II brought a vast number of new entities to the world's political map, whose values of democracy emulated were not necessarily those upon which Western democracies had been historically founded. The moralist (or institutionalist) approach became an easy target by 'those whose conceptions are not based on Aristotelian ideal types' (Heidenheimer et al. 1989). With the autonomy of political science in the 1950s, the classical approach was gradually replaced by new political science approaches (functionalist, market, and systemic).

3.2 *Contemporary Political-science Approaches*

The attempt to locate studies about corruption according to schools of thought does not seek over-

simplification of the variety of methods and perspectives through which the complexity of corruption is made tangible. Typologizing aims at helping those who are strange to the subject or have come across its problematic in the course of their research, to be aware of their adhesion to each alternative available as well as the insufficiencies and partial judgements recurring from their own choice.

The specificity of each theoretical approach depends largely upon the definition chosen to interpret the phenomenon. Their evolution denotes a value change in the way the phenomenon has been identified and explained.

3.2.1 Functionalist approach. Viewing corruption as 'another form of political influence,' functionalism, during the 1960s, tolerated certain forms of corruption in so far these were held to enhance economic development and political integration. Cultural relativism added to this complacency by considering corruption to be inherent in specific cultures/societies, and ignoring it elsewhere.

The normative argument surrounding the degree of legitimacy of the authoritative allocation of values and goods to society was a question which functionalists avoided at all costs. Corruption thus concerned the same 'legitimate apparatus' in two different facets of political activity: 'Whereas party manifestos, general legislation, and policy declarations are the formal façade of the political structure, corruption stands in sharp contrast to these features as an informal political system in its own right' (Scott 1972). Here, one could identify a Machiavellian vision of politics. But whereas the Florentine thinker observed that the struggle, exercise, and maintenance of power opened the door to a number of strategies, common to both democratic and authoritarian regimes, which were not necessarily confined to the legal and ethical norms set in society, functionalists went further to actually claim that corruption had its positive effects to the political system, despite being an 'immoral thing.' Corruption, functionalists concluded, 'greased the wheels of the system' in countries where the administrative process was overloaded and overstretched and it helped to hold shaky systems together, by mobilizing voters into the democratic process through party patronage and clientelism (Leff 1964, Bayley 1966, Huntington 1968).

Today, it would be difficult to advocate the benefits of conventional forms of corruption, in countries where toleration of those practices furthered the expansion and institutionalization of corruption into other areas of public life and paved the collapse of the political system. Moreover, the belief that corruption in developing regions, such as Latin America, Asia, and Africa, promoted economic development, political participation, and administrative efficacy proved historically wrong. Functionalists had never accounted for the long-lasting negative effects of corruption.

3.2.2 Market approach. The market approach focuses essentially on the principal-agent-client (PAC) corrupt transactions taking place at the bureaucratic level, coupling positive economic theory with neo-liberal theories of the state (Rose-Ackerman 1978). Although PAC theorists criticized vividly the functionalist view on benefits of corruption for development, their approach also sought to evade moral issues by centering their utilitarian arguments on the rational choice of individuals to undertake corrupt transactions. Accordingly, corruption depended on two major determinants: the existing structures of opportunity and the costs/incentives involved. From a theoretical standpoint, the rational choice approach has underestimated personal ethical standards constraining the public official's 'choice' or willingness to engage in corruption. Certainly, individual standards of probity, what the Anglo-Saxons termed 'civil service spirit,' go beyond cost-benefit calculations and constitute an important element in any micro-interpretation of corruption. Its theoretical weakness, however, had to do more with a drawback of neoliberal theories of the State.

The Welfare State crisis and the collapse of socialist economies offered propitious ground for a conceptualization of corruption as an illegal tradeoff between public and private spheres. Corruption, it was argued, expanded where the State extensively had permeated the market sphere, directly, through interventionist policies and ownership, and indirectly, via 'obstructive' market regulation. The remedies advocated were too simplistic: if the reason for corruption was excessive State intervention and regulation over the market economy, the solution was, *ipse dixit*, to privatize and deregulate. By the late 1980s, however, PAC theorists were less prone to support the neoliberal assumption of 'less State, less corruption': privatization and deregulation processes taking place worldwide had created new corrupt incentives and opportunity structures. The contribution of political economists to the study of corruption did not preclude with the neoliberal theories of the State, nor public choice interpretations of the motivation for corrupt behavior. Current economic theory is re-examining the very concept of 'market of corruption' through an analysis of the networks of corruption unrepresentative of a market and by introducing nonrational elements to motivation (Cartier-Bresson 1995).

3.2.3 Systemic (or neoclassical) approach. Both these approaches have avoided an important conceptual aspect, namely, moral condemnation. They had fallen into a reductionist and mechanic vision of the

democratic process and ignored that democracy, other than being a set of procedures and institutions, is also a normative system with ultimate goals to be realized in society (Mény 1992). Current literature on democracy theory looks attentively at normative aspects and qualities, including its pathologies and the measures to prevent its degeneration. This is not because corruption is more frequent in democracies than in nondemocracies, but because its consequences seem to be more damaging to the basic principles of democracy: equality before the state apparatus, accountability and transparency of public decisions, and the legitimacy of political institutions.

The revival of the moralist approach goes together with the growing belief that the ethical foundations of democracy have degenerated during the past decades. Academic concern about the growth of corruption in the post-Cold War environment is associated closely with the processes of state collapse and democratic consolidation in Central and Eastern Europe (Holmes 1993) and a period of crisis faced by the model of liberal democracy—weakened political participation and solidarity, party and party system crisis, crisis of the State of Law, crisis of the Welfare State (Della Porta and Mény 1995).

Everywhere, corruption was no longer product of an occasional 'rotten apple,' but a systemic phenomenon often with greater success than the legitimate political system from which it draws its resources (Della Porta and Mény 1995). It no longer appeared to be an exception, but a norm: threatening democratic foundations, causing institutional crisis, destabilizing the global economy, helping 'organized crime' to breed from clandestinity to illicit government, fomenting public deception about political life, and undermining the idea of public good. Corruption celebrated the communion between power and economy: money increased the possibility to privileged access to power, and power allowed the possibility to enrich (Della Porta 1992).

Political scientists were invited to revise their definitions and to include normative concerns in their empirical account of corruption (Jos 1993). Politico-institutional crisis and reform were the key conjunctural features urging the need for a normative analysis of corruption and this was not detached from the world economic downturn of the early 1990s. 'What changed in the early 1990s was the public's readiness to tolerate political corruption' (Heywood 1997): if the economic boom of the 1980s had been propitious for complacent views about corruption, the decreasing 'feel-good' factor of the 1990s hastened perceptions and condemnation, if not only as a reaction to widening social disparities.

By the early 1990s, corruption was condemned widely at all levels of decision-making: local, national, regional, and international. The discrepancy between the legal/formal standards and those upheld by the society at large had grown wide (Johnston 1991, Mény 1996). Most democracies became engaged in a spiral-web of media exposure, scandal, political inquiries, and judicial prosecution. Governments, even those that until then considered corruption 'a problem affecting others,' reacted through the implementation of a series of control measures and devoted their efforts to the moralization of political life.

The growing concern about corruption by international governmental (the Council of Europe, the European Union (EU), the Organisation for Economic Co-operation and Development (OECD), the World Bank, and the International Monetary Fund (IMF)) and nongovernmental (the International Chamber of Commerce (ICC), some NGOs, and multinational corporations (MNCs)) entities was contemporary to other transnational issues, such as environmental degradation, human rights, fraud, money laundering, fiscal evasion, and organized crime. Although analytically distinct from these cross-border phenomena, the mechanisms and ways in which corruption takes place are often intertwined. Not surprisingly, bribery by MNCs linked to the establishment of petrol and chemical industries or infrastructural projects in developing countries have resulted in vast environmental degradation. Similarly, countries condemned internationally for human rights violations are also known for their predatory practices towards national and foreign investors. At the EU level, official reports have indicated that the growth of fraud against Community financial interests is related partly to the diversion of funds by criminal organizations, which have been used, *a posteriori*, for illegal political financing.

3.2.4 Corruption and its public policy dimension. Current political research is particularly concerned with the way corruption affects governmental ability and efficacy to deliver policy outputs, varying accordingly to the singular organisational and attitudinal characteristics of each country. In a few words, corruption runs counter to good governance.

Other than a moral issue, corruption is a public policy problem. Corruption grows where public ethics have degenerated, where there are no clear rules through which public business is carried out, and where public or private activities lack proper modes of regulation to guarantee due process and fairness. The changing nature of the State and the blurring of the public/private divide have been at the core of the problematic of corruption in recent years, alerting decision makers to the need to strengthen controls and revise standards in public life (Mény 1992, Della Porta 1992).

The complexity of modern decision-making process—different layers, actors, and mechanisms of political influence—added to the lack of transparency in which public decisions and procedures have been

carried out, permitted a privileged access to State resources and undermined public confidence in democratic government.

The costs of corruption to the public purse are unquantifiable. Besides being a burden to taxpayers, corruption consolidates inefficiency and foils value-for-money considerations. It inflates market prices for public services and goods, supplied directly or contracted out to market actors; it fosters clientelism at the cost of fair competition. It allows the uncompetitive to thrive in a context of political protection.

Policy-makers have tried, in many countries, to regulate and control the situations conducive to corruption: procurement policies, regulation of public utilities delivery, control of markets. Other forms of soft corruption (various types of conflict of interest, *pantouflage*, *cumul des mandats*, lobbying) have also been subject to tighter regulation.

At the international level, the United Nations (UN) incessantly has reported and condemned vividly corruption as a product of mismanagement and bad policy-making at all levels. International economic/financial organizations (the OECD, the IMF, the World Bank) equally have realized that their rehabilitation programs could not be accomplished solely through the recruitment of expert economists and that specific social, cultural, political, and institutional factors conditioning development had to be brought into account. Financial aid to underdeveloped economies, rather than contributing to development, was helping to sustain corrupt elites in power. The European Commission has insisted on good governance as a criterion to its preferential financial treatment and trading benefits with developing countries, and a guiding principle to enlargement/accession strategies.

The European Parliament has also taken a recriminatory position in regard to corruptive practices within EU institutions, more precisely, the Commission.

See also: Ethics and Values; Journalism; Media, Uses of; Political Culture; Political Money and Party Finance; Political Trials; Public Opinion: Political Aspects

Bibliography

Bayley D H 1966 The effects of corruption in a developing nation. *Western Political Quarterly* **19**(4): 719–32

Cartier-Bresson J 1995 A economia da corrupção. In: Della Porta D, Mény Y (eds.) *Democracia e Corrupção na Europa*. Inquérito, Mem Martins, Portugal

Della Porta D 1992 *Lo scambio occulto*. Il Mulino, Bologna, Italy

Della Porta D, Mény Y (eds.) 1995 *Democracia e Corrupção na Europa*. Inquérito, Mem Martins, Portugal

Friedrich C J 1966 Political pathology. *Political Quarterly* **37**: 70–85

Friedrich C J 1989 Corruption concepts in historical perspective. In: Heidenheimer A J et al. (eds.) *Political Corruption*. Transaction Publishers, New Brunswick, NJ

Gardiner J A 1992 Defining corruption. *Corruption and Reform* **7**(2): 111–24

Heidenheimer A J 1989 Perspectives on the Perception of Corruption. In: Heidenheimer A et al. (ed.) *Political Corruption*. Transaction Publishers, New Brunswick, NJ

Heywood P (ed.) 1997 *Political Corruption*. Blackwell, London

Huntington S P 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, CT

Holmes L 1993 *The End of Communist Power*. Polity Press, Cambridge, UK

Johnston M 1991 Political corruption: historical conflict and the rise of standards. *Journal of Democracy* **2**(4): 48–60

Johnston M 1996 The search for definitions: the vitality of politics and the issue of corruption. *International Social Science Journal* **48**(3): 321–35

Jos P H 1993 Empirical corruption research: beside the (moral) point? *Journal of Public Administration Research and Theory* **3**(3): 359–75

Leff N 1964 Economic development through bureaucratic corruption. *American Behavioral Scientist* **8**(3): 8–14

Mény Y 1992 *La Corruption de la République*. Fayard, Paris

Mény Y 1993 La décennie de la corruption. *Le Débat* **77**: 15–25

Mény Y 1996 'Fin de siècle' corruption: change, crisis and shifting values. *International Social Science Journal* **48**(3): 309–20

Nelken D, Levi M 1996 The corruption of politics and the politics of corruption: an overview. *Journal of Law and Society* **23**(1): 1–17

Nye J S 1967 Corruption and political development: a cost-benefit analysis. *American Political Science Review* **61**(2): 417–27

Rogow A A, Lasswell H-D 1963 *Power, Corruption and Rectitude*. Prentice-Hall, Englewood Cliffs, NJ

Rose-Ackerman S 1978 *Corruption: A Study in Political Economy*. Academic Press, New York

Scott J C 1972 *Comparative Political Corruption*. Prentice-Hall, Englewood Cliffs, NJ

Sorokin P 1991 *Social & Cultural Dynamics: A Study of Change in Major Systems of Arts, Truth, Ethics, Law, and Social Relationships*. Transaction Publishers, New Brunswick, NJ

Van Klareven A J 1989 The Concept of Corruption. In: Heidenheimer A et al. (eds.) *Political Corruption*. Transaction Publishers, New Brunswick, NJ

Woodiwiss M 1988 *Crime Crusades and Corruption: Prohibitions in the United States 1900–1987*. Pinter, London

Y. Mény and L. de Sousa

Cortical Activity: Differential Optical Imaging

Introduction

Primates are extraordinarily visual animals. Hence it is not surprising that we (as humans) prefer visual

representations of everything we try to understand. Whether in the form of text, images, or graphs, visual representations are an integral component of our efforts to comprehend. As a consequence, it is easy to understand the appeal of techniques for mapping activity in the brain. It is even easier to understand the appeal of optical imaging, which maps activity patterns directly by monitoring changes in reflected light.

The roots of optical imaging can be traced back at least 60 years, to the observations of Wilder Pennfield (1933) who noted minor changes in the appearance of exposed cortex during epileptic seizures. On account of associated vasospasms, the seizing regions looked alternately light and dark. Though such changes must be large indeed to be visible to the naked eye, they resemble smaller ones that accompany normal events. This wasn't recognized at the time—indeed these changes were deemed pathological. But even if it had been, the technologies needed for using them to map normal activity would not become available for many years.

A few years after Pennfield's initial observations, Kelin and Millikan found activity dependent changes in the optical properties of cytochromes. Hill and Keynes (1949) subsequently recorded the first activity dependent optical signals, in specially prepared nervous tissues. However, these and subsequent efforts soon were diverted by an emerging focus on single neurons. Hence, the possibility of using light to monitor slow changes in the activation of cortical tissue was neglected until Blasdel and Salama (1986) showed that slow, activity-driven changes in light reflectance can be used to visualize organizations that had intrigued neuroscientists for decades: the arrangement of ocular dominance and orientation slabs in the primary visual cortex. Though their initial observations have been replicated many times, using a variety of slightly modified procedures, the fundamental strategy stays the same. All rely on the small, activity driven reflectance changes resembling those noted by Pennfield, all try to drive these alternately between two extremes, and all try to map associated reflectance changes with some differential imaging technique.

1. *Primary Visual Cortex*

It is not an accident that the first comprehensive optical maps were obtained from the primary visual cortex of macaques—an area ripe for this kind of analysis, after decades of brilliant discoveries showing the importance of ocular dominance and orientation selectivity, and the likely organization of these properties in slabs (see Fig. 1). In addition, the primary visual cortex of macaque monkeys is particularly well suited to optical investigation since most of it is remarkably large and flat and, due to its location on top of the occipital lobe, conveniently accessible. It also was not an accident that this was first achieved by visual neuroscientists, since many of the problems with differential imaging resemble those in vision—problems whose solutions are most familiar to those thinking about them full-time.

In order to appreciate the nature of the problem at hand, it is important to understand a few aspects of primate area V1. Light from the outside world is refracted onto the retina, where it directly forms the first of several representations including one in a layer of ganglion cells, whose axons form the optic nerve and project centrally to the lateral geniculate nucleus (LGN). Neurons in six separate but retinotopically aligned layers of the LGN receive this information and project directly to V1. The thing to keep in mind, therefore, is that V1 is not the first two-dimensional representation of visual space, nor does it reflect retinotopic arrangements in any simple way. Instead, it reflects the cumulative consequences of successive transforms, each one of which modifies the representation of space in simple but important ways.

These modifications lie beyond the scope of this review. However, they can be summarized as a tendency to reduce some form of redundancy with each and every step. The consequence of a center-surround receptive field in the retina or LGN, for example, is to minimize responses to diffuse or gradually changing light. Light that doesn't change with distance is redundant and can be ignored. This makes it possible to concentrate on more important events, like those where light changes suddenly, associated with an edge.

The reduction in redundancy continues in the LGN and primary visual cortex (also known as the striate cortex, area 17, and V1), where redundant views from the two eyes are merged. Because the convergence is usually incomplete, there is a residual tendency for one eye to dominate in alternating 0.5 mm wide slabs. These are known as 'ocular dominance slabs' or 'ocular dominance columns' and are illustrated in Fig. 1a.

At the time information converges from both eyes to form a single binocular map, the cortex sees the emergence of a new response property—orientation selectivity—that is characterized by selective responses to particular orientations of edges. This property too can be characterized as a reduction in redundancy, since strings of discontinuous events that occur along a single axis in space can now be summarized by a single neuron. Like ocular dominance, this property is characterized by vertically aligned neurons, with those lying above one another responding only to similar orientations. What is particularly important about the vertical alignment of ocular dominance and orientation columns, in regard to optical imaging, is that their depth invariance makes it possible to ignore the depth of light penetration, which is difficult to de-

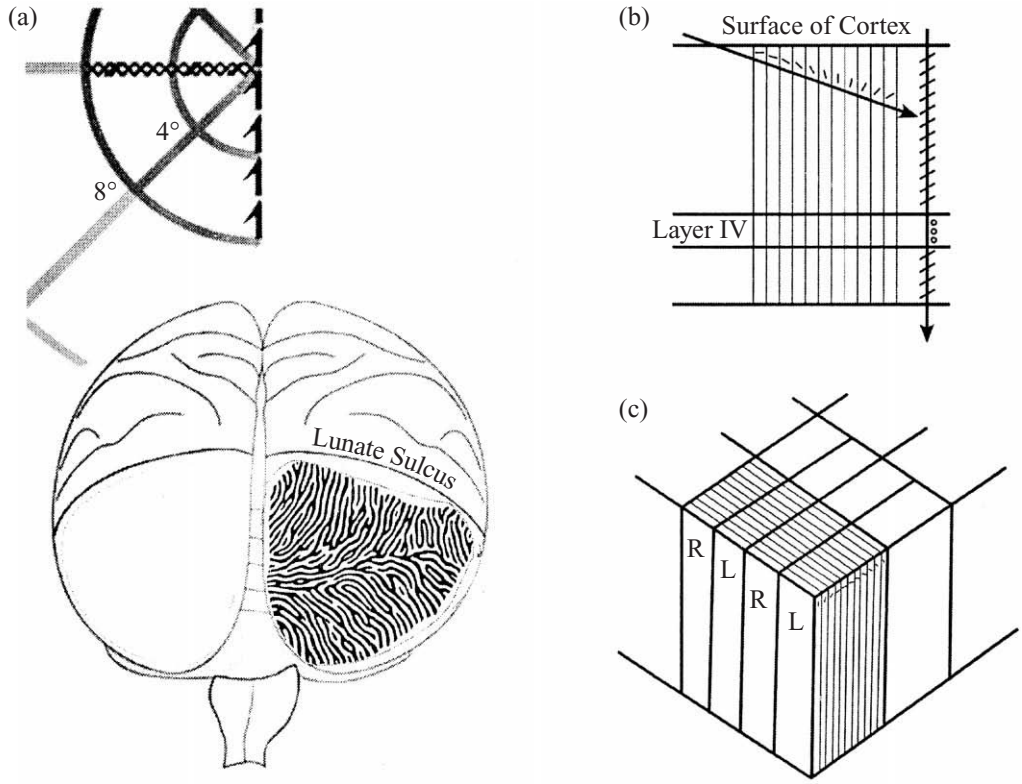

Figure 1
Organizational schemes for the representation of visual space, ocular dominance and orientation preference columns. (a) Monkey striate cortex occupies a region half the size of a credit card at the back of the animal's head; each hemisphere represents the contralateral hemifield. The representation is topographic (neighboring points in space are represented by neighboring points in cortex. However, it is also characterized by a regular arrangement of half millimeter wide ocular dominance slabs—shown here for the right hemisphere representation of left visual space—that interleave striped representations of the right and left eyes. A characteristic feature of ocular dominance columns in monkey striate cortex is that they intersect the boundary of this area at right angles (modified from LeVay et al. 1975 and from Hubel and Freeman 1977). (b) Hubel and Wiesel's results for electrode penetrations conducted perpendicular and parallel to the cortical surface. For perpendicular penetrations, the preferred orientation remains constant (except in layers 4a and 4c, where the selectivity for orientation is weak or absent). For lateral penetrations, the preferred orientation rotates continuously for distances of 0.5–1.0 mm. (c) Orthogonal stripe-like representation of orientation and ocular dominance in primary visual cortex, proposed by Hubel and Wiesel (1974). This model has four important components: 1) slab-shaped regions preferring one eye or one orientation project vertically between pia and white matter, 2) slabs representing the two eyes lie parallel to one another, 3) slabs representing different orientations lie parallel to one another, and 4) ocular dominance and orientation preference slabs intersect at right angles. These illustrations are re-drawn from those appearing in Hubel and Wiesel (1974, 1977)

termine. In spite of their clear importance, ocular dominance and orientation selectivity patterns had proven difficult to map.

2. *Differential Video Imaging*

A solution to this problem is provided by the small changes in reflectance—similar to those described by Pennfield (1933)—that accompany normal activity as well. Due to recent advances in the development of practical imaging technologies (e.g., cameras, image processors, computers, etc.) reflectance differences as small as 0.01 percent—approximately 500 times smaller than ones that can be seen by the unaided eye—can easily be seen (Blasdel 1992a, b). Accordingly, any area of cortex that can be exposed and activated physiologically, can be mapped with this technique.

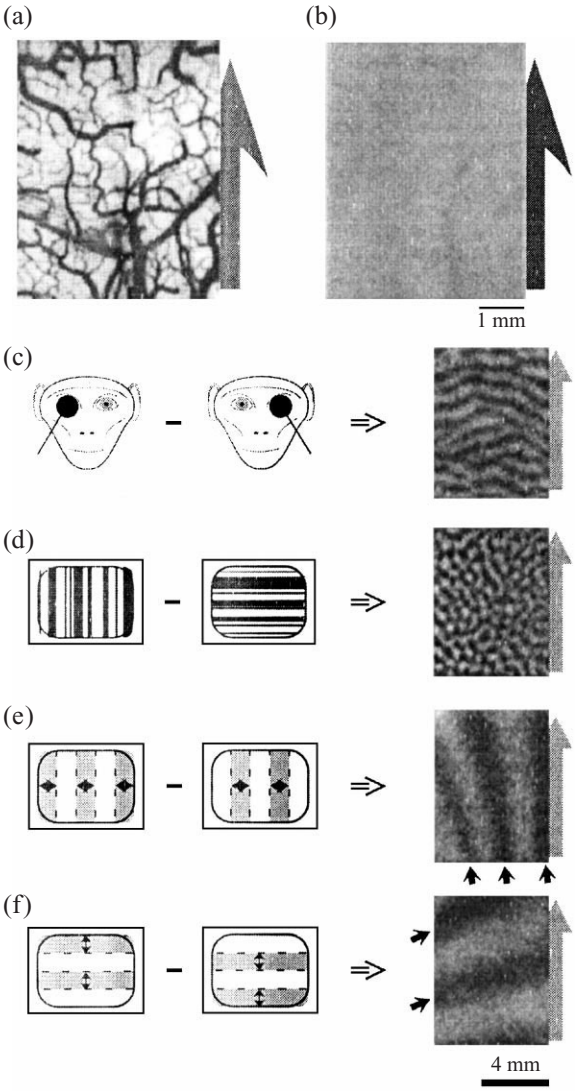

Figure 2
Patterns mapped from the surface of primary visual cortex. (a) The appearance of 3 mm × 4 mm region of cortex under 540 nm light, where the blood vessels appear dark due to the strong absorption of this wavelength by hemoglobin. (b) The same region of cortex under near-infra red (720 nm) light—the wavelength at which optical maps are obtained—where blood vessels appear fainter due to weak absorption by hemoglobin. (c) Visualization of ocular dominance bands by stimulating right and left eyes on alternate trials and differentially imaging the small changes in reflectance that occur. Note the characteristic width of these bands and their tendency to intersect the border of cortex (represented by gray arrow) at right angles. (d) Differential images of orientation visualized by stimulating both eyes alternately with orthogonal orientations (in this case, vertical and horizontal) and differentially imaging the changes that emerge.
(e) Differential images of retinotopic space, obtained by stimulating complementary stripe like compartments on alternate trials and differentially imaging the changes. Due to the retinotopic organization of cortex, the stimulation of vertical stripes in space induces vertical bands (that run parallel to border, represented by arrow), while (f) The stimulation of horizontal stripes in spaces produces bands that intersect the V1 border at right angles

Though a number of procedures have been described for doing this, all entail minor variations of the same technique. All exploit small, activity-dependent changes that cause active cortex to reflect less light; all seek to modulate them between two extremes, and all seek to average responses to increase ratios of signal to noise. The specific mechanisms that couple reflectance to activity are controversial and likely to reflect different contributions at different wavelengths. Since shorter wavelengths (600 nm) are strongly absorbed by hemoglobin, signals obtained with them are dominated by vascular events: changes in blood volume and possibly even the oxygenation of hemoglobin. Since longer wavelengths (720 nm) are absorbed much less strongly by hemoglobin, signals monitored with them are freer from vascular events and more likely to reflect differences in scattering due to changes in ion concentration, water movement, and the volumes of active cells; they are also less pronounced (for a review, see McLoughlin and Blasdel 1998).

The first apparatus used for differentially imaging these changes relies on a low power microscope (Blasdel 1992a) that uses co-axial illumination to irradiate the cortex with monochromatic light (720 nm ± 20 nm). Reflected light is collected by an objective that relays it to an imaging device—usually a TV or CCD camera—that transduces it into a video signal that is digitized and recorded by an image processor that averages thousands of frames for each response to improve ratios of signal to noise.

The point of differential imaging is to see changes that are too small to be seen by themselves. This is especially true in the case of cortical responses, where the optical changes associated with orientation preference or ocular dominance may be a hundred times smaller than those associated with cortical activation alone. By stimulating the cortex twice, however, with stimuli that activate it to approximately the same degree, it is possible to modulate a single variable and visualize the small reflectance changes that develop in phase. In the case of ocular dominance patterns, for example, the responses to either eye alone may be huge. Yet because common mode changes to each eye are similar, a differential subtraction (i.e., images achieved during stimulation of one eye are subtracted

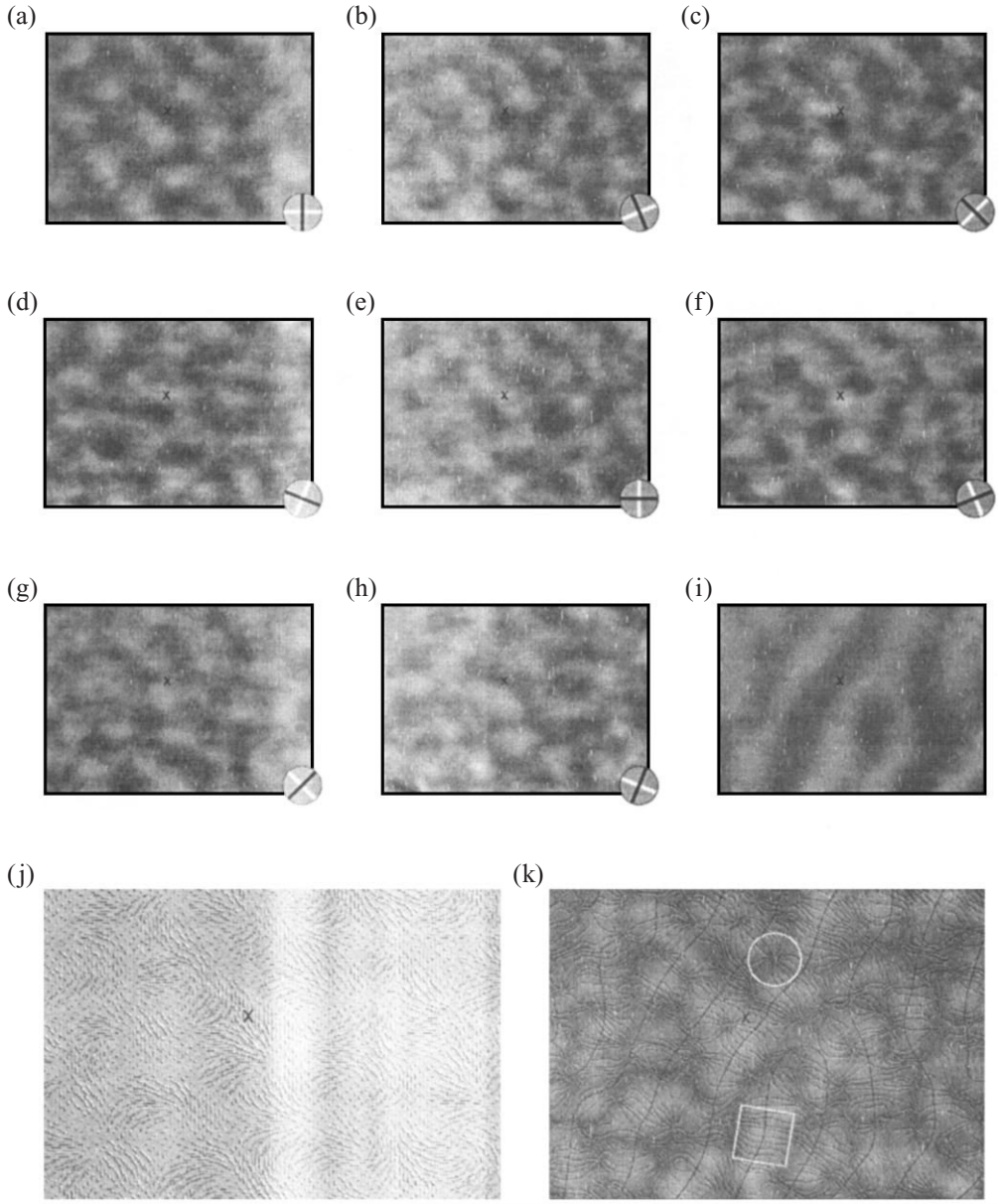

Figure 3
Differential images of orientation and ocular dominance from the same region of macaque V1. The horizontal axis in each frame corresponds to 4.4 mm. Each of the 8 images in (a–h) was obtained by differentially imaging responses to orthogonal edges, moving bilaterally at 1.5°/sec. Dark and light values reflect preferences for each of the two orientations, which are indicated by dark and light bars in the lower right hand corner, and which are rotated by 22.5° in successive frames—i.e., 0°–90°, 22.5°–112.5°, 45°–135°, 67.5°–157.5°, 90°–0°, 112.5°–22.5°, 135°–45°, and 157.5°–67.5° (from Blasdel 1992b). Note how similar these patterns look, even though actual positions of dark and light regions change. This is made clear by the relative location of an **'x'** appearing at the center of a 'Vertical' region in all frames. This coincides with a dark patch in (a), a light patch in (e), and intermediate locations in between. (i) Differential image of ocular dominance from same region. (j) By turning each of the images in (a–h) into vectors, to keep track of angle phase, it is possible to add them to achieve a field of orientation vectors that

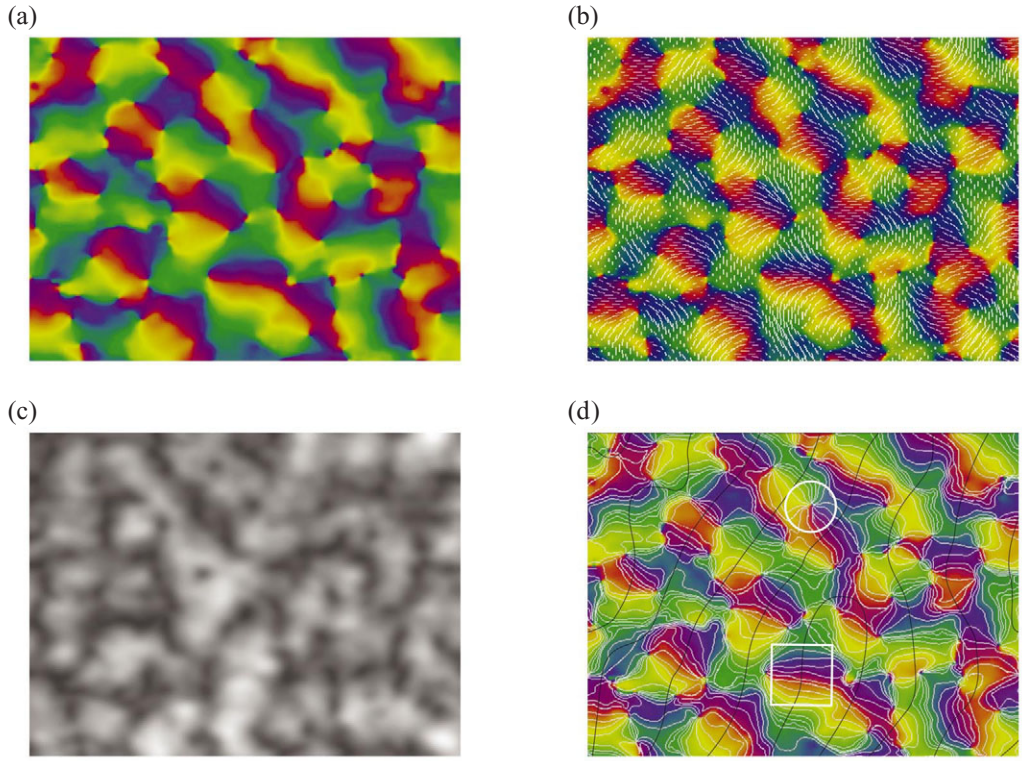

Figure 4
Because orientation preferences change cyclically, they cannot be illustrated by gray levels. Hence, the orientation preferences (calculated from images in Figs. 3a–3h), are illustrated more easily by using opponent colors to represent orientations that are perpendicular. (a) This color-coded map contains the same information as Fig. 3j, but represents preferences for vertical and horizontal with green and red, preferences for left and right oblique with blue and yellow, and preferences for intermediate orientations with colors in between. (b) This shows the degree of selectivity in gray-scale: lighter regions correspond to longer lines in Fig. 3j, indicating greater selectivity for the indicated orientation in Fig. 3j, and (a), above. (c) Superposition of the short lines from Fig. 3j, which appear in white, on the color-coded representation from (a). (d) Iso-orientation contours (Fig. 3k), re-drawn over color-coded map

from images averaged during stimulation of the other) can remove these common mode events and reveal differences specific to the two eyes.

Once the cortex has been exposed (by removing the scalp, cranium, and overlying meninges) and immobilized by a thin plastic or glass plate, it is possible to see the surface of the cortex (Fig. 2a) and differentially image activity. While the dense covering of superficial blood vessels might seem to present a problem, they become virtually transparent at 720 nm.

indicate the degree of selectivity and the preferred orientation at each location. Though these usually are represented by color-coded maps, they also can be represented by short lines, whose lengths and orientations reflect selectivity and preferred orientations at corresponding locations. Note that the **'x'** from the images above appears in a field of vertical lines. (k) Iso-orientation contour map (derived from Obermayer and Blasdel 1993) by calculating contours along which the preferred angle (e.g., vertical) does not change and representing it by a white line superimposed over maps of orientation selectivity (grayscale) from the same region. The thin black lines represent ocular dominance borders for the same region, derived from (i). Important differences from expectations based on Hubel and Wiesel's (1974, 1977) work—see Fig. 1c—include: 1) Iso-orientation slabs (white lines) are limited to 0.5–1.0 mm in length, 2) Iso-orientation slabs converge at regular orientation *singularities* (indicated by white circle) in the centers of ocular dominance bands, 3) Iso-orientation slabs do intersect ocular dominance borders at right angles in *linear ones* (indicated by white square), which lie equidistant between singularities and where they appear particularly well aligned

Due to the weak absorption of 720 nm light by hemoglobin, the same blood vessels that dominate Fig. 2a become virtually transparent in Fig. 2b. This is the image seen by the camera imaging all responses that are compared.

Differential imaging works by comparing two images of the cortex (see Figs. 2c–2f) in slightly different states that under ideal circumstances differ with respect to only one variable. If done carefully, the resulting difference image reveals a pattern specific to the variable. By subtracting images of the cortex responding to the left eye from images of it responding to the right eye, for example, one obtains an image of ocular dominance columns, as illustrated in Fig. 2c. By subtracting images of the cortex responding to horizontal edges from images of it responding to vertical edges, one obtains a pattern associated with orientation selectivity, where dark and light regions represent selectivity for vertical and horizontal (Fig. 2d). Similarly, by differentially imaging responses to complementary parts of space, it is possible to visualize their cortical representations. The vertical stripes in Fig. 2e, for example, elicited bands running parallel to the V1/V2 border (which represents vertical), while horizontal stripes (Fig. 2f) induced bands that intersect the V1/V2 border at right angles.

Fig. 3 shows 8 differential images (a–h) of orientation, obtained by comparing responses to each of two perpendicular gratings, with the orientation of each pair rotated by 22.5° in successive frames. As one can see, the small 'x' placed at the center of a vertical iso-orientation band lies at the center of a dark patch in Fig. 3a, but moves gradually to the side in successive frames until it lies over a white patch in Fig. 3e. From the relative displacements of light and dark regions relative to this x, it is possible to see that these patterns depend *only* on the orientations of stimuli used to induce responses that are compared.

3. *Orientation Preference Patterns*

As one might note from the images in Fig. 3, the gray values at various locations reveal the degree to which each of the two opposed orientations (indicated in the lower right hand corner of each frame) was effective at driving responses. They do not indicate which orientation was optimal, however, since this information can only be inferred by comparing responses to many orientation pairs. Because this information appears distributed across *all* the images in Fig. 3a–3h, it is easiest to extract from all of them at once, by converting the pixel values in each image into vectors (which use phase to keep track of the angle used), that can be added to reveal the distribution of preferences and selectivities at all locations (see Blasdel 1992a for more details). The outputs of this operation are illustrated in Fig. 3j, where each short line indicates the degree of selectivity and preferred orientation, through its length and orientation. As one can see, some of these are quite short, indicating no selectivity at all; while many are quite long, indicating sequences of orientation preference that rotate gradually in swirls, from side to side. In order to estimate the shapes of orientation 'columns,' it is possible to calculate contours along which orientation preferences do not change. The resulting iso-orientation contours (Obermayer and Blasdel 1993), which resemble Hubel and Wiesel's (1974) iso-orientation slabs, are illustrated for the same region of cortex in Fig. 3k.

As one can see, this pattern differs from that expected (see Fig. 2c). Rather than appearing in long arrays of infinitely long columns, iso-orientation slabs appear limited to 0.5–1.0 mm in length and to sequences that are no more than 0.5–1.0 mm across. This wasn't known before, and has become apparent only on account of optically imaged patterns, with results from all laboratories showing approximately the same thing (see also T'so et al. 1990, Bartfeldt and Grinvald 1992). A particularly surprising phenomenon concerns orientation singularities that occur at regular intervals in the centers of ocular dominance columns where iso-orientation contours converge (Blasdel 1992b; Bonhoeffer and Grinvald 1993). One example appears highlighted by a white circle in Fig. 3k. These are the main events that disrupt iso-orientation contours, and prevent them from getting longer. Nevertheless, iso-orientation contours seem to congregate in equally regular linear zones that subtend 0.5–1.0 mm and predominate near ocular dominance borders. One example is outlined by a white rectangle in Fig. 3k. Within these zones, iso-orientation contours appear particularly well organized in linear arrays of slabs, more or less as Hubel and Wiesel (1974) predicted (see Fig. 1c). Note, that within these zones, iso-orientation slabs also have a striking tendency to cross ocular dominance borders at angles of approximately 90°—more or less as suggested by Hubel and Wiesel's model, in Fig. 1c.

From this relatively simple analysis of optical imaging maps, it is possible to conclude that orientation preferences organize according to one of *at least* two competing schemes—a linear one, confined to 0.5–1.0 mm patches, where orientation preferences change linearly along one axis while remaining constant along the other; and a discontinuous, circularly symmetric one, that entails continuous changes in preferred orientation around a point singularity. The linear scheme predominates in two-dimensional patches, 0.5–1.0 mm across, while the pinwheel singularities that disrupt them lie somewhere in between. Also, these results make it clear that within the linear zones iso-orientation contours and ocular dominance borders cross at right angles.

The true strength of optical imaging is apparent from the comparisons that it enables. Since it only takes a few minutes to obtain each pattern with a technique that appears to be relatively benign, it is

possible to collect many different types of response profile from the same cortical regions. This is what makes it possible to calculate maps of orientation preference, for example, where responses to many different orientations need to be compared. It also is what enables comparisons between ocular dominance and orientation maps from the same regions, allowing them to be related precisely for the first time. By comparing these with maps of retinotopic representation from the same region (see Figs. 2e, 2f) it is possible to obtain unprecedented insights into cortical organizations.

See also: Electrocephalography: Basic Principles and Applications; Functional Brain Imaging; Occipital Lobe (Visual Cortex): Functional Aspects; Visual Perception, Neural Basis of

Bibliography

Bartfeld E, Grinvald A 1992 Relationships between orientation-preference pinwheels, cytochrome oxidase blobs, and ocular-dominance columns in primate striate cortex. *Proceedings of the National Academy of Science USA* **89**: 11905–9

Blasdel G G 1992a Differential imaging of ocular dominance and orientation selectivity in monkey striate cortex. *Journal of Neuroscience* **12**(8): 3115–38

Blasdel G G 1992b Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience* **12**(8): 3139–61

Blasdel G G Campbell D in press Functional retinotopy of monkey visual cortex. *Journal of Neuroscience*

Blasdel G G, Salama G 1986 Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature* **321**: 579–85

Bonhoeffer T, Grinvald A 1993 The layout of iso-orientation domains in area 18 of cat visual cortex: Optical imaging reveals a pinwheel-like organization. *Journal of Neuroscience* **13**(10): 4157–76

Grinvald A, Lieke E, Frostig R P, Gilbert C, Wiesel R M 1986 Functional architecture of cortex revealed by optical imaging of intrinsic signals. *Nature* **324**: 351–64

Hill D K, Keynes R D 1949 Opacity changes in stimulated nerve. *Journal of Neuroscience* **7**: 3378–415

Hubel D H, Freeman D C 1977 Short communications: projection into the visual field of ocular dominance columns in macaque monkey. *Brain Research* **122**: 336–43

Hubel D H, Wiesel T N 1972 Laminar and columnar distribution of geniculo-cortical fibers in the macaque monkey. *Journal of Comparative Neurology* **146**: 421–450

Hubel D H, Wiesel T N 1974 Sequence regularity and geometry of orientation columns in the monkey striate cortex. *Journal of Comparative Neurology* **158**: 267–93

Hubel D H, Wiesel T N 1977 Ferrier lecture. Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London Series B Biological Science* **198**(1130): 1–59

Hubel D H, Wiesel T N, Stryker M P 1978 Anatomical demonstration of orientation columns in macaque monkey. *Journal of Comparative Neurology* **177**: 361–80

LeVay S, Hubel D H, Wiesel T N 1975 The pattern of ocular dominance columns in macaque usual cortex revealed by a reduced silver stain. *Journal of Comparative Neurology* **159**(4): 559–76

McLoughlin N P, Blasdel G G 1998 Wavelength-dependent differences between optically determined functional maps from macaque striate cortex. *Neuroimage* **7**: 326–36

Obermayer K, Blasdel G G 1993 Geometry of orientation and ocular dominance columns in monkey striate cortex. *Journal of Neuroscience* **13**: 4114–29

Penfield W 1933 The evidence for a cerebral vascular mechanism in epilepsy. *Annals of Internal Medicine* **7**: 303–10

Schiller P H, Finlay B L, Volman S F 1976 Quantitative studies of single-cell properties in monkey striate cortex. III. Spatial frequency. *Journal of Neurophysiology* **39**: 1334–51

Ts'O D Y, Frostig R D, Lieke E E 1990 Functional organization of primate visual cortex revealed by high resolution optical imaging. *Science* **249**: 417–20

Wilson H R, Levi D, Maffei L, Rovamo J, DeValois R L 1989 The perception of form: retina to striate cortex. In: Spillman L, Werner J S (eds.) *Visual perception: The Neurophysiological Foundations*, Academic Press, New York, pp. 231–72

G. Blasdel

Cortical Development: Transplantation and Rewiring Studies

The mammalian neocortex, a folded sheet that in humans contains over 10 billion neurons, is the seat of the highest sensory, motor, and cognitive abilities. Cortical development involves the formation of many discrete areas that process different kinds of information uniquely. Individual areas are characterized by specific sets of inputs, processing networks, and outputs. The development of these features involves factors that are intrinsic to the tissue or are regulated by extrinsic or environmental influences. This article examines the relative roles of such influences, employing in particular evidence from experiments that have probed developmental mechanisms by transplanting cortical regions or by rewiring inputs to a part of the cortex.

1. Development of the Cortex

The principal method for studying mechanisms of development is to manipulate developmental processes. In general, it is not possible to separate intrinsic cleanly from extrinsic aspects of developmental programs of an organism: no gene acts independently of an environment, and the environment always acts on a scaffold. A classic paradigm in developmental biology for clarifying the role of the local environment in determining the features of a tissue is to translocate the tissue to a novel environment.

Transplanting one part of the mammalian cortex to another at different stages of development has been a valuable tool for determining whether or not unique features of a cortical area derive from intrinsic properties of the cortex, or are induced by new inputs in the host environment. A powerful recent method that complements transplantation studies is to route sensory inputs that normally drive one part of the cortex to another part. Thus, instead of altering the location of a target and preserving host inputs, the inputs to existing targets are altered. Together, these studies demonstrate that the features that make a cortical area unique are expressed progressively as development proceeds. Importantly, they show that electrical activity can have an instructive role during the development of cortical networks, so that specific spatial and temporal patterns of activity regulate specific patterns of connections between cortical neurons.

2. *Formation of Cortical Layers*

The adult mammalian neocortex consists of six layers. These are generated by a heterogeneous population of precursor cells that lie along the walls of the lateral ventricles very early in development. The time of neurogenesis in the ventricular zone regulates the final disposition of cells in individual layers. Cells which are born earlier form the deeper cortical layers, whereas later-born cells form the progressively more superficial layers. Transplantation experiments have shown that there is a progressive restriction of cell fate in the precursor population: Early progenitors are able to give rise to neurons which come to reside in any cortical layer, whereas later progenitors give rise only to neurons of more superficial layers.

One key set of experiments (McConnell and Kaznowski 1991) demonstrates that transplanting neurons from the ventricular zone of a donor animal at an age when layer 6 is being born, into a host in which layer 2/3 is being born, causes the donor cells to settle into either host-specific layer 2/3 or donor-specific layer 6, depending on the stage of the cell cycle at which donor cells are extracted. If donor cells are extracted prior to or during the phase of DNA synthesis, they are able to switch their fate and migrate to the layer that is being generated in the host (layer 2/3), due presumably to environmental cues in the host. If donor cells are extracted later in the cell cycle, their fate is sealed prior to transplantation and they settle into the layer determined in the donor (layer 6). The precise nature of extrinsic signals that determines or alters the laminar fate of newborn cells during a certain part of the cell cycle is unknown.

Events occurring at the ventricular zone can contribute in other ways to regional differences in cortical architecture. The dynamics of the cell cycle show regional variation. Most notably, the germinal zone of the primary visual cortex (V1) of primates, a region with twice the neuronal density of other areas, has a notably higher rate of cell production which is associated with changes in the parameters of the cell cycle. Thalamic afferents can exert a mitogenic effect on cortical progenitors, resulting in an increase in the number of cell divisions. Thus, thalamic inputs are one important source of extrinsic influences on cortical development, starting with the cell cycle and lamina formation.

3. *Formation of Cortical Areas*

There is considerable evidence suggesting that at least the initial broad parcellation of the cortex into discrete areas is regulated by molecular determinants that are intrinsic to the proliferative zone of the developing cortex. For example, evidence from transplantation and co-culture studies suggests that the initial expression of region-specific markers or phenotypes is not dependent on specific thalamic innervation. A number of genes are expressed differentially between cortical areas before thalamic afferents are believed to have invaded the cortical plate. Similarly, the expression of several genes can occur normally even in the absence of thalamocortical innervation. Finally, gradients of gene expression in the early cortex can regulate initial cortical parcellation, and such genes can also serve to attract appropriate sets of inputs from the thalamus, or direct appropriate outputs from the cortex (Rubenstein and Rakic 1999).

While the expression of genes and molecules may delineate broad areas of cortex at very early stages of development, subsequently thalamic afferents can influence the size and even the identity of specific areas (O'Leary et al. 1994). Transplants of extremely immature cortex—at embryonic day (E)12 in rats—from the limbic to the somatosensory cortex may take on inputs and express molecules characteristic of the host region rather than the region of origin. This ability is lost if the transplants are derived at a later developmental stage. Transplanting visual cortex into somatosensory cortex in rats causes the graft to accept inputs from the somatosensory thalamus and form 'barrels,' a characteristic of the rodent somatosensory cortex. Other work shows that the transplant needs to occur early, before E16, for the graft to form substantial connections with the somatosensory thalamus, and later grafts are already committed to form connections with the visual thalamus. With respect to output connections, occipital to frontal transplants at E12 can form projections to the spinal cord, a characteristic of the host cortex rather than the donor; however, by E14 that capacity has been lost and projections to the tectum, a characteristic of the region of origin, are formed instead. Interestingly, a similar age-dependent switch between E12 and E14 influences the corticocortical connections made by transplanted

perirhinal cortex into the parietal region. These findings indicate that there is an early time window during which the input and output connections of a cortical area may be sculpted by thalamic and local or regional signals to produce an area-specific phenotype. Thus, cortical areas arise by progressive specification of their region-specific phenotype from a multipotent phenotype (Levitt et al. 1997).

4. *Formation of Cortical Networks: Ocular Dominance and Orientation Columns in Visual Cortex*

The cortex differentiates progressively into layers, and as layer 4 appears in the cortical plate, thalamic innervation specifies the principal sensory areas. Concurrently, descending projections from the cortex innervate specific thalamic nuclei, primarily by the targeting of axons of layer 5 cells to principal and/or association nuclei, followed by the development of layer 6 projections to principal relay nuclei. Thus, a thalamocortical loop is set up very early in development, and the initial specification of cortical areas is fundamentally a specification of unique feedforward and feedback connections between thalamus and cortex. The development of intracortical circuitry follows thalamic innervation; considerable recent evidence, primarily from the visual cortex, demonstrates that a combination of intrinsic and extrinsic factors (such as electrical activity in thalamic inputs) is responsible for the formation and maintenance of specific thalamocortical and intracortical connections.

Ocular dominance columns—regions within layer 4 of V1 in higher mammals that receive input exclusively from one eye or the other via the lateral geniculate nucleus (LGN) of the visual thalamus—are paradigmatic of thalamocortical patterns that specify individual cortical areas. In primates they are present by birth, suggesting that visual experience is not required for their formation. Recent work suggests that they may also be set up before eye opening in ferrets, soon after geniculocortical axons innervate layer 4 (Crowley and Katz 2000), and similarly in cats. Surprisingly, they may not even require the presence of the eyes for their initial establishment. Monocular enucleation does not degrade these early columns immediately, and binocular lid suture does not reduce their development in cats for the first three weeks. One possibility is that ocular dominance columns are set up initially by the molecular matching of inputs from different layers of the LGN (each layer representing a given eye) with appropriate target regions in V1.

An alternative view is that ocular dominance columns are set up by activity-based rules of development. Thus, if inputs from one eye are correlated with each other and uncorrelated with inputs from the other eye, application of a Hebbian developmental rule for strengthening connections (and of a complementary rule for weakening connections), together with local excitation and long-range inhibition in the cortex, can lead to cortical stripes that resemble ocular dominance columns. Correlated inputs in the form of spontaneous waves of activity exist in the retinae prior to eye opening. In the LGN, the firing of neurons within a given eye-specific layer is well correlated, and uncorrelated with the firing of neurons in an adjacent layer which represents the other eye. Interestingly, ablation of the cortex abolishes these correlations, whereas ablation of inputs from the eyes has little effect. The pattern of electrical activity transferred to the cortex, therefore, appears to be regulated strongly by interactions between the thalamus and the cortex, and activity in the thalamocortical loop may be sufficient for eye-specific patterning. Significant support for the hypothesis that ocular dominance columns can arise, at least in principle, by activity-dependent sorting of inputs comes from the finding that eye-specific stripes form in the optic tectum of 'three-eyed' frogs forced to receive input from two eyes (Law and Constantine-Paton 1981). Whether or not ocular dominance columns in V1 require electrical activity to instruct their formation, there is convincing evidence that electrical activity is required to maintain them (Sherman and Spear 1982).

Orientation selectivity is a major emergent property in visual cortex that is created by aligned thalamic inputs whose activity is amplified by local intracortical connections. Orientation selectivity is present in V1 of primates at birth, and to a degree in cats and ferrets at or before the time of natural eye opening, although selectivity increases with visual experience (Chapman et al. 1999). Orientation-selective cells in visual cortex are linked by long-range horizontal intracortical connections in the superficial layers to form an orientation map. The adult pattern of clustered horizontal connections is present at birth in primates. In cats and ferrets, crude clusters appear just prior to eye opening, and are refined after the onset of vision; binocular deprivation prevents this refinement. The orientation map, revealed by optical imaging of intrinsic signals, develops in parallel with the development of orientation selectivity in single cells.

Monocular lid suture after orientation maps have already formed disrupts the map from the closed eye; reverse lid suture restores the map precisely. In addition, matching orientation maps for the two eyes develop even in cats raised under an alternating-suture paradigm, so that the two eyes never have common visual experience. Together, these studies indicate that the emergence of orientation-selective responses in single cells and the overall layout of the orientation map does not require patterned vision, although visual experience is necessary for orientation selectivity and maps to fully mature. However, blocking electrical activity in V1 reduces both the orientation tuning of single cells and the clustering of horizontal connections. Thus, spontaneous electrical activity in cortex or

in the thalamocortical loop is required for the initial establishment of local and long-range intracortical connections.

5. Patterned Electrical Activity and the Development of Cortical Networks

The vast majority of experiments on the influence of activity on visual cortical networks have involved manipulations such as lid suture or activity blockade that alter the *amount* of activity in the visual pathway. Relatively few experiments have examined the influence of the *pattern* of activity on cortical network development and maintenance; such manipulations include artificial strabismus, specific rearing paradigms, and rewiring of visual projections to the auditory pathway.

Strabismus refers to misalignment of the two eyes' optical axes, so that activity from corresponding retinal loci in the two eyes are no longer temporally correlated. Artificially induced strabismus causes neurons in V1 to become almost exclusively monocular, and ocular dominance columns to be more sharply delineated and have altered spacing compared to normal animals. Long-range horizontal connections within the superficial layers of V1 are also affected. Whereas in normal cats these connections cluster to link regions with similar orientation preference but do not align with ocular dominance columns, in strabismic cats the connections come to link columns with similar eye dominance and orientation preference (Schmidt et al. 1999). Thus, correlations in input activity are used to sharpen and organize ocular dominance columns, and to shape intracortical horizontal connections.

Rearing kittens in a visual environment consisting of alternating black and white stripes at specific orientations restricts pattern vision and appears to shift the orientation preference of cortical cells towards the experienced orientation. While a shift could be caused by a passive loss of responses to nonexperienced orientations, there is an expansion of cortical columns devoted to the experienced orientation (Sengpiel et al. 1999), indicating an instructive effect of patterned visual experience on orientation selectivity and the orientation map.

5.1 Rewiring the Cortex

A unique paradigm for investigating the role of patterned afferent activity in the development of cortical circuitry and function is experimentally to redirect afferents carrying information about one sensory modality to central targets and pathways that normally process a different sensory modality (Schneider 1973). Retinal axons can be induced to innervate the auditory thalamus in the ferret (Sur et al. 1988), a carnivore with a highly organized visual pathway and which is born at a very early stage in development. The routing of retinal projections to the auditory pathway can be induced at birth, allowing the probing of the role of patterned activity in the *establishment* (and not simply the *maintenance*) of cortical circuitry. Importantly, visual input is relayed from the auditory thalamus to the primary auditory cortex (A1) via thalamocortical projections whose physical identity is unchanged but which provide spatiotemporal patterns of electrical activity to the auditory cortex that are very different from normal (Fig. 1).

Visual driving leads to emergent responses in 'rewired' A1 that demonstrate the role of patterned activity in shaping cortical networks. An orderly two-dimensional retinotopic map develops in A1 of rewired ferrets, in cortex that normally maps a one-dimensional surface, the cochlea. The retinotopic map in rewired A1 arises despite widely dispersed and overlapping thalamocortical projections, suggesting that well-localized receptive fields and their orderly progression are created by correlation-based mechanisms that operate intracortically to select and strengthen a specific subset of synapses from a broader set available anatomically.

Visual cells in rewired A1 have orientation-tuning, direction-tuning, and velocity-tuning properties that are quantitatively indistinguishable from V1 cells, suggesting that similar mechanisms operate in the generation of receptive field properties in the two cortices. Orientation-tuned neurons in A1 are organized into an orientation map that resembles in important respects the map in V1 (Sharma et al. 2000). The differences between orientation maps in V1 and rewired A1 reflect differences in the underlying organization of superficial-layer long-range horizontal connections in these two cortices. In V1, horizontal connections are patchy, spatially periodic, and extend mediolaterally. In contrast, horizontal connections in normal A1 are more band-like in organization, show little spatial periodicity, and are anisotropic anteroposteriorly. Horizontal connections in rewired A1 show features that are intermediate between V1 and normal A1: The connections in rewired A1 form much smaller and more regular patches than in normal A1, though the patches are less tightly clustered and are larger in size than in V1. Thus, horizontal connections within the rewired cortex are significantly altered by visual input, but in a manner that appears to be constrained by intrinsic features of the auditory cortex.

Rewired ferrets also provide an opportunity for examining whether the behavioral role of a cortical area is set by intrinsic determinants or by the pattern of afferent activity during development. Behavioral experiments, supported by several kinds of control experiments, indicate that rewired ferrets interpret visual stimuli which activate the rewired projection as visual rather than as auditory (von Melchner et al.

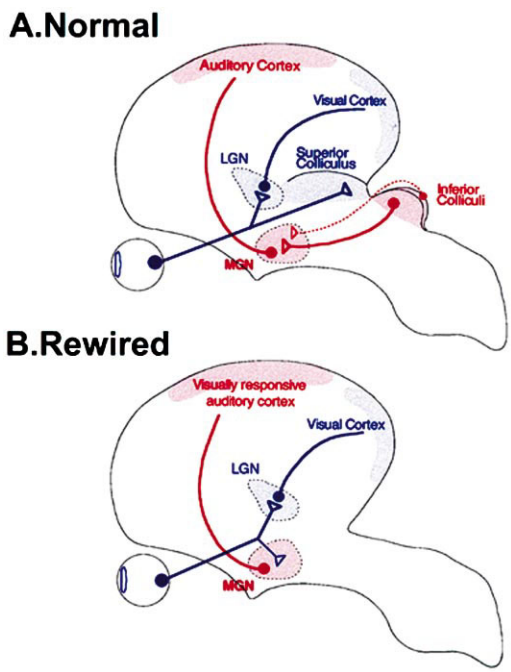

Figure 1
Retinal inputs routed to the auditory thalamus in rewired ferrets drive the auditory cortex with visual activity without altering thalamocortical projections. **A:** Visual and auditory pathways in normal ferrets. The retina projects predominantly to the lateral geniculate nucleus (LGN) and superior colliculus. The LGN projects to the primary visual cortex. The medial geniculate nucleus (MGN) receives most of its subcortical afferents from the ipsilateral inferior colliculus, although afferent projections also arise from the contralateral inferior colliculus. The MGN projects to the primary auditory cortex. **B:** Visual pathways in rewired ferrets. Subcortical inputs to the MGN in one hemisphere are removed in early postnatal ferrets. This induces retinal axons to innervate the deafferented MGN. The MGN still projects to the primary auditory cortex, as in normal ferrets, but in rewired ferrets it relays visual rather than auditory inputs. This change in spatiotemporal patterns of input activity early in development has a profound effect on networks in auditory cortex, and on its function. Adapted from Angelucci et al. (2000)

2000). Thus, the function of a cortical area is dependent fundamentally on the spatiotemporal pattern of activity it receives during development. It is possible that all 'auditory' pathways central to the thalamus in the rewired hemisphere are turned 'visual,' including the cortex and downstream structures, with a concomitant respecification of their perceptual identity.

Cortical development involves a progressive shaping of the fate of the cortical epithelium into discrete cortical areas with specific inputs, outputs, and local networks. Developmental manipulations, which include, importantly, transplantation and rewiring studies, demonstrate a continual interplay of intrinsic and extrinsic factors at all stages of development—at the ventricular zone during the cell cycle, in the cortical plate during the parcellation of cortical areas, and within the cortex during the formation and maintenance of cortical networks. The nature of extrinsic signals varies with developmental time, and likely includes intercellular signals in the ventricular zone that influence the formation of layers, trophic or permissive electrical signals in early area formation, and instructive electrical signals in late network development, which persist into adulthood as substrates for learning and memory.

See also: Brain Implants and Transplants; Cerebral Cortex: Organization and Function; Cognitive Development in Infancy: Neural Mechanisms; Neural Development: Mechanisms and Models; Prefrontal Cortex Development and Development of Cognitive Function; Topographic Maps in the Brain; Visual Development: Infant

Bibliography

Angelucci A, Sharma J, Sur M 2000 The modifiability of neocortical connections and function during development. In: Kaas J H (ed.) *The Mutable Brain*. Harwood, Amsterdam

Chapman B, Godëcke I, Bonhoeffer T 1999 Development of orientation preference in the mammalian visual cortex. *Journal of Neurobiology* **41**: 18–24

Crowley J C, Katz L C 2000 Early development of ocular dominance columns. *Science* **290**: 1321–24

Law M I, Constantine-Paton M 1981 Anatomy and physiology of experimentally produced striped tecta. *Journal of Neuroscience* **1**: 741–59

Levitt P, Barbe M, Eagleson K 1997 Patterning and specification of the cerebral cortex. *Annual Review of Neuroscience* **20**: 1–24

McConnell S K, Kaznowski C E 1991 Cell cycle dependence of laminar determination in developing neocortex. *Science* **254**: 282–85

O'Leary D, Schlagger B, Tuttle R 1994 Specification of neocortical areas and thalamacortical connections. *Annual Review of Neuroscience* **17**: 419–40

Rubenstein J, Rakic P 1999 Genetic control of cortical development. *Cerebral Cortex* **9**: 521–23

Schmidt K, Galuske R, Singer W 1999 Matching the modules: Cortical maps and long-range intrinsic connections in visual cortex during development. *Journal of Neurobiology* **41**: 10–17

Schneider G E 1973 Early lesions of superior colliculus: Factors affecting the formation of abnormal retinal projections. *Brain, Behavior and Evolution*. **8**: 73–109

Sengpiel F, Stawinski P, Bonhoeffer T 1999 Influence of experience on orientation maps in cat visual cortex. *Nature Neuroscience* **2**: 727–32

Sharma J, Angelucci A, Sur M 2000 Induction of visual orientation modules in auditory cortex. *Nature* **404**: 841–47

Sherman S M, Spear P D 1982 Organization of visual pathways in normal and visually deprived cats. *Physiology Review* **62**: 738–55

Sur M, Garraghty P E, Roe A W 1988 Experimentally induced visual projections into auditory thalamus and cortex. *Science* **242**: 1437–41

von Melchner L, Pallas S L, Sur M 2000 Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature* **404**: 871–76

M. Sur

Cosmopolitanism

Cosmopolitanism is a western notion that epitomizes the need social agents have to conceive of a political and cultural entity, larger than their own homeland, that would encompass all human beings on a global scale. Cosmopolite in Ancient Greece meant citizen of the world. Cosmopolitanism presupposes a positive attitude towards difference, a desire to construct broad allegiances and equal and peaceful global communities of citizens who should be able to communicate across cultural and social boundaries forming a universalist solidarity. Its inclusive drive is most evident in moments of crisis of other modes of representing and ascribing membership to existing sociopolitical and cultural units. Much of the malaise and misunderstanding cosmopolitanism may provoke are related to its ambiguity, that is, its unique way of uniting difference and equality, an apparent paradox of wishing to reconcile universal values with a diversity of culturally and historically constructed subject positions.

The composition of the Greek term, *cosmopolis*, already indicates this unsolved tension: *cosmos*, a natural universal order, is related to *polis*, society's variable order. As a consequence, from the Greek democratic city-state to the global village, the idea of a cosmopolite has been haunted by questions such as *whose* world this is. Can homogeneous external expansionist forces coexist with heterogeneous localized ones? A truly cosmopolitan answer would imply a permanent interest in difference and the recognition that globals and locals depend on each other to exist.

Since its inception, cosmopolitanism has been a category marked by a need to negotiate with 'others' and has reflected tensions between local and supralocal realities, ethnocentric and relativist perspectives, and particularism and universalism. Historically, cosmopolitanism has mirrored the ideologies of different periods and modes of integration to larger, imperial or global, political entities. As a category mostly held by elites, it often means the sophistication that results from familiarity with what is different. It has become a metaphor for mobility, migrancy, sensitivity, and tolerance to otherness, independence from specific authorities, and transcultural and transnational realities and claims. Its opposing concepts have often indicated xenophobia, fixity, parochialism, restricted sovereignty, and allegiance to a motherland, or a nation-state.

1. *A Brief History*

The history of the relationships between local and supralocal conceptions is probably as old as humankind. A strong trend towards local reality, particularism, diversity, and context may oscillate, such as at the end of the Renaissance or during the Enlightenment, towards an emphasis on general formal timeless statements that pretend to be universal. Having its roots in Ancient Greece, cosmopolitanism has been variably present in western philosophical or political discussions.

The military conquests of Alexander the Great (356–323 BC) opened up the conditions for the existence of a 'world empire' that allegedly aimed at uniting East and West into an enlightened commonwealth. Greek became the *lingua franca* of the Hellenistic age (fourth to first century BC), a period that lasted until the establishment of Roman hegemony. Although cosmopolitanism was an issue for Greek philosophers before Stoicism, this school of philosophy established in Athens by 300 BC systematized cosmopolitan theories advancing visions such as that of a world city, an ideal state where everyone would be a citizen. Stoics were instrumental in criticizing Greek ethnocentrism towards barbarians and fostered a sense of brotherhood, a vision of humankind that was transmitted to Romans and predated Christianity's claims to universalism. From the Roman Empire through Medieval Europe cosmopolitanism passed on to different political and intellectual elites. The Christian church played a major role in the reproduction of cosmopolitan ideals and apparatuses by forming transcultural sacred imagined communities and diffusing Latin as the language of a transnational European power.

Some of the fundamental historical processes of long duration that strengthened cosmopolitanism were related to the establishment of modernity, itself a civilizatory cosmopolitan notion and force. The printed book, a new technology of communication developed in mid-fifteenth century, put into question parochialisms in time and space, performing a typical cosmopolitan operation: the enhancement of the consciousness of diversity and the construction of larger imagined communities. Sixteenth-century European expansion gave impulse to the world capitalist system through the incorporation of new territories and populations, and established global colonialism,

multiplying the number of images of and contact with exotic others. Science, technology, and reason began their path to hegemony in the construction of universalizing discourses. Market places and urban centers emerged with citizens that experienced new forms of individuality, etiquette, and public space aspiring to new secular ideologies and modes of Republican, democratic governments. The American (1776) and French (1789) Revolutions formally marked this cleavage in time. The Enlightenment and its Encyclopedia represented a climax of universalist ideals.

Impersonal, anonymous relations, mobility, and estrangement were to become the rule for proletarians and the bourgeois in metropolitan centers that were being transformed by an emergent industrial revolution. The time came for capital to intensify its global reach through the networks of new colonial and imperial cosmopolitan elites linked to the development of the nation-state as the dominant form of organizing the equation territory-culture-ideology-politics-state. Expansion of the nation-state had started with the Peace of Westphalia in 1648 and consolidated in the nineteenth century. This movement was to be completed only in the twentieth century with the last great wave of decolonization in the 1960s and 1970s. The notable growth of nationalism invigorated its counterpart, internationalism, especially in moments of crisis of the interstate system. After the two World Wars, organisms were inaugurated to regulate the world system and establish new forms of global governance and hegemony: the League of Nations (1920–1946); the United Nations (1945); the International Monetary Fund (IMF) and the World Bank (1944). These agencies are greatly responsible for the dissemination of transnational ideologies and initiatives.

In all these processes, many thinkers have played important roles. They held lay, religious, idealist, positivist, socialist, liberal, conservative, or revolutionary positions, clearly demonstrating the plasticity and efficacity of cosmopolitanism across political and theoretical persuasions. They thought of an international court and army (Abbe de Saint Pierre), a federation of nations and eternal peace (Kant), technological solutions for rationally organizing the planet (Saint-Simon, Goethe), global alliances against oppression (Marx), world government and state (H. G. Wells, Lippmann), world economic regulations (Keynes).

The need for cosmopolites to communicate with a myriad of others has been a crucial leitmotif from Leibniz (a universal system of characters) to Sapir (an international auxiliary language). This ancient quest for a 'debabelized' world, a peaceful community of communication constructed out of heteroglossia, has proved to be unviable given the highly complex ways language communities are reproduced. Moreover, the uneven distribution of power within the world system structures the role certain languages—Greek, Latin, and English, for instance—have played in the construction of cosmopolitan/transnational creoles, communities, and imaginaries, leaving little room for initiatives such as Esperanto.

2. *Anthropological and Sociological Considerations*

The sentiments cosmopolitanism evokes are not restricted to the western world. Human beings are always interested in knowing where people come from. The ability to locate oneself or others in geographical, cultural, and political terms depends on an array of classificatory categories that are culturally and historically constructed. These modes of representing membership to sociocultural and political units, modes that bind people to collectivies and territories, can be conceived as a continuum of concentric circles ranging from local, phenomenological immediacies, to more distant regional, national, international, and transnational levels of integration whose influence is variably present in the lives of social agents. Given the present nature of world system integration, all these levels are simultaneously present allowing for feelings of multiple belonging, generally conceptualized in terms of 'hybridity.' Participation on one level triggers certain forms of social representations and allegiances that are relational and circumstancial, and that define rigid or flexible identity boundaries and subject positions that in turn inform social cooperation or competition. A person can simultaneously hold allegiances to a neighborhood, a city, a region, a country, or a continent, or be a transmigrant in a world city or, yet, a global nomad, an employee of a transnational corporation. Much of the discussion on whether cosmopolitanism is possible or not rests on a disconsideration of the simultaneous presence of these forces of identity construction, which resonates with essentialist conceptions of identity. The existence of broad-based encompassing forms of integrating people and territory under the same political and symbolic umbrella does not mean the end of narrower ones. Otherwise, how could one explain the persistence of regional and ethnic separatism within nation-states? It should be clear, though, that while everyone is local, not everybody is global.

Many forces have made contemporary cosmopolitanism possible: individualism with its relative detachment from immediate, narrow solidarities; global expansion of economic and political systems by military, commercial, and religious means; development of transportation and communication technologies that have exacerbated time-space compression and, consequently, the circulation of people, information, and commodities on a planetary scale; growth of global cities and the heightening of cultural and ethnic heterogeneity they have brought; the empire of the mass media especially of global television

and the emergence of the information era with its worldwide virtual web; new political actors such as nongovernmental organizations that are fueled by transnational organizations and ideologies.

Two criticisms are commonly made against cosmopolitanism: one, that it is an elitist social representation; the other, that it is an impossible project. Arguments can be offered to counterbalance both these points. The intense global migratory movements of the last two centuries generated large numbers of uprooted people, complex urban and national ethnic segmentations, transnational networks, and diasporic cultures that, intermixed with the works of the mass media, created a popular cosmopolitanism, fueling processes and visions of globalization from below. It is thus necessary to explore the existence of several cosmopolitanisms. Popular cosmopolitanisms differ from corporate ones, which in turn differ from those of bourgeois tourists, business magnates, or international scholars. There is no doubt that exposure to difference and cultural diversity is increasing rapidly as is the number of transmigrants and differentiated groups, most often occupational ones, among whom allegiance to a nation-state is secondary. The efficacy of historical and sociological facts (such as the Atomic Bomb, satellite planetary integration, the global interconnection of stock markets, global consumer culture, the European Union) and of new universalist ideologies (such as environmentalism and the defense of human rights), together with the emergence of new political subjects and social movements, has further stimulated articulations and actions of a new transnational activism. For some, all this will result in the organization of a global civil society. In spite of the fact that these processes also generate relativist critiques in face of the unequal global distribution of power (the South/North divide in activism also draws a line of asymmetrical relationships and access to visibility and infrastructure), they engender a more concrete framework for cosmopolitanism than in any previous period. Extraterritoriality is another real challenge to legal systems firmly set on national legislations and jurisdictions in a world system where imperial powers or multinational military alliances operate.

3. Cosmopolitanism, Globalization, and Transnationalism

With the completion of the nation-state system in the 1970s, cosmopolitanism clearly meant transcending this entity. Nation-states were then viewed as a homogenizing force by an arising postmodern critique tired of the universalist metanarratives of the Enlightenment based on discursive matrices, such as progress, embodied by contemporary nation-states. French-inspired postmodernism hit mainstream academia, especially in the 1980s and 1990s, in the United States, favoring visions of global heterogeneity and multiculturalism. More than ever, the discussion on cosmopolitanism became intimately intertwined with debates on transnationalism. In the late 1980s, the end of 'really existing socialism' helped to propagate the image of a unified world, monopolized by a triumphant capitalism under the hegemony of powerful transnational corporations and financial capital. In the kingdom of flexible, postfordist capitalism, globalization turned into a mantra and much of the typical tensions inherent to cosmopolitanism as a notion were dramatized within the localist/globalist framework of analysis, sometimes regarded as antithetic polarities, othertimes as complimentary and articulated terms. Notions of transnational classes, cultures, and identities entered the horizon of the social sciences.

There are two major currents of interpretation and promotion of transnational cosmopolitanism. The first is dominated by transnational capitalists and associated elites that praise a borderless neoliberal world, that is, unrestricted access to domestic markets and natural and social resources, and the empowerment of global actors and agencies such as the World Bank, the IMF, and the World Trade Organization. The second is composed of intellectuals, some of them within academia (working mostly in the fields of anthropology, cultural studies, economics, geography, international relations, political science, and sociology), others in nongovernmental organizations, who are advancing visions of heterogeneity, heteroglossia, migrancy, cultural diversity, or empowerment of local actors, and claiming a need for global governance and a global civil society to regulate the power of transnational deterritorialized elites. In different ways, both sides feed on universalist ideologies such as development (with its promises of unlimited fortune and technological transcendence), republicanism, liberalism, socialism, environmentalism, and the defense of human rights. They are also strengthened by new transnational frameworks of cultural and political activity engendered by the new technologies of communication of late twentieth century. The latter have become a focus of explorations on global culture, on the emergence of new fragmented and flexible identities, electronic public spaces and interactions, cultural hybridty and cosmopolitan political communities. The Internet has brought about the possibility of a transnational imagined virtual community: a decentered cosmopolitan crowd, synchronized by cyberspace, interacting on real-time, and engaging in global commercial, cultural, and political exchanges.

Cosmopolitanism frequently is conflated with the imperial inclinations of a historical period; a fact that stresses its contradictory relationships with power. In the early twenty-first century, it will be criticized increasingly as a disguised form of Americanism or of serving transnational capital. Will the future see a

homogenizing global state or a heterogeneous planetary civil society? Be that as it may, one thing is certain, cosmopolitanism—the need to transcend received loyalties and attachments in favor of imagining distant unknown others, of constructing broader solidarities and global notions of citizenship—will continue to furnish, with its pacifist, egalitarian, utopian reverberations, a powerful ideological window into the future of a shrinking world.

See also: Cross-cultural Psychology; Cultural Assimilation; Cultural Diversity, Human Development, and Education; Global History: Universal and World; Globalization and World Culture; Metropolitan Growth and Change: International Perspectives; Transnational Citizenship: Cultural Concerns

Bibliography

Bohem M H 1942 Cosmopolitanism. *Encyclopedia of the Social Sciences* **4**: 457–61

Brennan T 1997 *At Home in the World. Cosmopolitanism Now*. Harvard University Press, Cambridge, MA

Cheah P, Robbins B (eds.) 1998 *Cosmopolitics. Thinking and Feeling Beyond the Nation*. University of Minnesota Press, Minneapolis, MN

Clifford J 1992 Traveling cultures. In: Grossberg L, Nelson C, Treichler P (eds.) *Cultural Studies*. Routledge, New York

Featherstone M (ed.) 1990 *Global Culture*. Sage, London

García Canclini N 1990 *Culturas Híbridas*. Grijalbo, México

Hannerz U 1996 *Transnational Connections*. Routledge, London

Harvey D 1989 *The Condition of Post-modernity*. Blackwell, Oxford, UK

Rosenau J N 1990 *Turbulence in World Politics*. Princeton University Press, Princeton, NJ

Toulmin S 1990 *Cosmopolis. The Hidden Agenda of Modernity*. Free Press, New York

Wolf E R 1982 *Europe and the People Without History*. University of California Press, Berkeley, CA

G. L. Ribeiro

Cost–Benefit Analysis

Cost–benefit analysis (CBA) is a method for assessing the economic efficiency of public policies through the systematic measurement of social costs and social benefits. When economic efficiency is the only relevant social goal, CBA provides an appropriate decision rule: choose the policy, or set of policies, that maximizes net social benefits. Although the conceptual foundations of CBA continue to be refined, most of the basic concepts were in place when it was explicitly introduced in the 1930s to evaluate water resource projects in the USA. Advances in empirical research methods, especially with respect to the measurement of the benefits of goods not traded in markets, have steadily expanded the plausible range of application of CBA.

1. Foundations of Cost–Benefit Analysis

CBA is concerned primarily with efficiency in the allocation of resources. As such, it follows directly from the body of neoclassical economic theory (or resource allocation theory). Yet whereas the main body of neoclassical theory deals with the nature of *private* decisions by households and firms, CBA focuses on *public* decisions. It sets forth the requirements for any public decision, or policy, to be efficient; that is, for the policy to contribute to an 'improvement' in the allocation of economic resources. CBA seeks to answer the question: will an alteration in the current pattern of resource allocation improve efficiency in the use of national resources? Insofar as CBA is directed at allocative efficiency, it can be viewed as an attempt to replicate for the public sector the decisions that would be made if private markets worked satisfactorily. In private markets, no explicit attention is given to social goals such as equity. Indeed, private markets tend to respond to actual demands and resource constraints, where demand is partly a reflection of the distribution of income, and economists do not generally regard private markets as operating unsatisfactorily simply because they fail to produce goods for people who cannot afford to pay for them. Similarly, the focus of cost–benefit analysis is on allocative efficiency, and implies that public sector decisions should be guided by these same demand (not desire) and resource cost considerations. It is presumed that if there are concerns about the acceptability of the income distribution that gives rise to the pattern of demands, or to the impact of the policy's benefits and costs on the income distribution, these concerns will be taken into account through other policies with explicitly redistributive goals.

As with the neoclassical theory that guides market analysis, CBA rests on the analytical underpinnings of modern welfare economics. Modern welfare economics shares the utilitarian presumption that aggregating the wellbeing (utilities) of the individuals who make up society is a valid measure of social wellbeing. CBA provides a basis for inferring changes in utilities, which are unobservable, from economic choices and outcomes, which are observable. These outcomes, especially if they are revealed in markets, express the value of changes in utilities in terms of a money metric. At one level, then, CBA can be thought of as the accounting protocol of welfare economics for measuring in terms of a money metric the changes in utilities caused by changes in public policy.

The principle of *Pareto efficiency* lies at the heart of welfare economics (see *Pareto, Vilfredo (1848–1923)*). An allocation of goods is Pareto-efficient if there exist no alternative allocations that make at least someone

better off without making anyone else worse off. The intuitive appeal of Pareto efficiency is obvious: one would have to be malevolent to oppose a reallocation that betters the situation of someone without harming anyone else.

Although the Pareto criterion seems unobjectionable as a basis for judging the relative desirability of public policies, it would be largely useless as the basis for choice because policies without some losers are extremely rare in the real world. This limitation of the Pareto criterion prompted efforts to develop a normative welfare criterion with more general applicability to actual decisions. A more serviceable criterion for CBA arose out of proposals by Nicholas Kaldor and J. R. Hicks: a change in the allocation of resources should be regarded as increasing welfare if either the Pareto criterion is met or the persons who have gained by the reallocation could compensate those who have been harmed by it so as to leave everyone at least as well off as they would have been without the reallocation. Any resource allocation that meets this criterion is said to be a 'potential Pareto improvement.' This criterion expands the set of situations and policies about which statements may be made as to whether economic welfare has been increased, decreased, or left unchanged.

In resting on this criterion, CBA becomes an accounting protocol that enables the analyst to determine whether a reallocation has the potential to be Pareto improving. As a decision rule, this criterion seeks the greatest potential Pareto improvement by choosing projects that maximize the excess of social benefits over social costs.

CBA gained a distinct identity in the 1930s as a formal technique for systematically measuring benefits and costs of flood control projects sponsored by the US federal government (for an overview, see Steiner 1974). As the scope of governments grew, and economists gained influence as policy analysts, the use of CBA expanded well beyond the evaluation of public works projects. In the USA, for example, Executive Order 12291 (February 17, 1981) requires that federal regulatory agencies subject proposed regulations with estimated costs of over $100 million to CBA. Continuing advances in empirical research methods further expand its range of application. Natural resource economists, for instance, use survey methods to estimate benefits necessary for applying CBA to environmental policies.

2. *Social Benefits and Costs: Concepts and Measurement*

2.1 *The With/Without Principle of Evaluation*

The evaluation of public projects in terms of the potential Pareto improvement principle requires that analysts determine if the value of the output in the economy *with* the proposed project would be greater than the value of the economy's output *without* the project. Note that this comparison must be made *ex ante* (based on the prediction of results). It therefore requires the prediction of the future under two scenarios: one with the project and one without it. This with/without principle is important for defining and measuring both social benefits and social costs.

Although CBA is forward looking, its application often requires analysts to draw on past experience as a basis for prediction. In such cases, the outcome of a similar policy previously implemented is observed but the counterfactual—what would have occurred without it—is not. Analysts must infer the counterfactual in order to make the with/without comparison. For example, in order to assess the social costs and benefits of an existing irrigation project, analysts must infer the value of crops and environmental quality that would have been produced without the project for comparison with those observed to occur with it. This *ex post* CBA (based on actual results) could then be used as an empirical basis for predicting the social costs and benefits for an *ex ante* CBA of a proposed irrigation project with similar features.

2.2 *A Comprehensive Framework: Standing*

Costs and benefits in CBA derive from the preferences of the individuals who make up society. But what defines society? And are all preferences legitimate? These questions have come to be known as issues of 'standing' (Whittington and MacRae 1986). Recognizing that the citizens of a country share a common constitution, which sets out values and rules for making collective choices, and participate in the same national economy, most analysts define society at the national level. The evaluation of social benefits andu social costs requires that the protocol for the analysis be a comprehensive one in which all of the adverse and beneficial effects of the policy are tabulated, regardless of who in the society is harmed or helped by them. It follows that all effects accruing to national citizens (or perhaps residents) be counted.

Comprehensively taking account of effects accruing to all those with standing means that CBA differs from financial analysis. Considering only the revenue flows to government agencies is generally inadequate because it does not take account of all the impacts of policies on citizens. It also means that analysts working for sub-national governments should resist pressures to ignore national costs and benefits that accrue outside of their jurisdictions. For example, a city may view a grant from the central government as only a benefit, while from the social perspective it involves an offsetting national cost.

An accounting that accepted all preferences as legitimate would be too broad. For example, most observers would not view the pleasure that psychopaths get from physically assaulting people as an

appropriate part of the utilitarian calculus. Where to draw the line between legitimate and illegitimate preferences, however, is sometimes difficult. One approach is to rely on the existing legal framework for guidance (Zerbe 1998), yet doing so assumes that existing laws are moral.

2.3 The Concept of Social Benefits: Willingness to Pay

The fundamental concept for guiding the measurement of social benefits is the aggregate *willingness to pay* of those with standing for the impacts of a policy. The analyst wishes to know the maximum amounts individuals would be willing to pay for policy impacts that they view as beneficial, and the minimum amounts individuals would be willing to accept as compensation for policy impacts that they view as harmful. In other words, the analyst would like to know how much each individual's wealth would have to be adjusted in conjunction with the policy change so that the individual's utility remained unchanged. The algebraic sum of these wealth adjustments for all individuals measures the social value of the impacts of the policy in terms of dollars. Although some theoretical caveats apply (Blackorby and Donaldson 1985), policies with larger positive sums have larger social benefits.

Note that the amounts that people are willing to pay to obtain a policy's impacts depends on their levels of wealth. For example, provision of a public good, such as improved environmental quality, may convey the same physical effect on all people. However, people with more wealth would be willing to pay more for this benefit than those who are poorer. An obvious implication of the dependence of individuals' willingness to pay on their levels of wealth is that the aggregate willingness of a society to pay for a policy impact depends on the distribution of wealth among its members. Thus, the social benefits used in CBA are contingent upon the particular distribution of wealth that exists in the society.

2.4 Categories of Social Benefits: Active and Passive Use

The impacts of public policies can be divided into two broad categories. In the first category are impacts that change observable behaviors of individuals. These behavioral changes are labeled 'active use.' For example, preserving a wilderness area may facilitate birdwatching, an observable behavior. Any such behavior provides at least the possibility of making an inference about willingness to pay through its observation.

Many policies also produce impacts that are valued by people without changing observable behaviors. The realization of these values is labeled 'passive use.' Analysts have found it useful to identify three distinct categories of passive use.

First, a policy may have an 'option value' (Weisbrod 1964). People may value a good because they anticipate the possibility of using it in the future. Only when they actually use it, if ever, will their behavior be observable. For example, a person may be willing to pay something now to preserve a national park in Alaska with the hope of visiting it someday. The person's circumstances, however, may never allow the visit to be made.

Second, a policy may have 'existence value' (Krutilla 1967). People may value a good even if they never anticipate actively using it. For example, a person may be willing to pay to preserve a wildlife habitat that would be completely closed to the public. One motivation for the willingness to pay may be a belief in the intrinsic value of the natural order. Another motivation may be altruism: a desire to make the good available either to the current generation, or to bequeath it to future generations.

Third, a policy may have a 'donor value' (Hochman and Rodgers 1969). People may be willing to pay something for a policy that distributes goods in a way they view as desirable. The motivation may be altruistic, say in providing more to those who otherwise would be consuming less. It may also be motivated by values of equality or fairness. The inclusion of donor values integrates individual preferences over the distribution of wealth into the efficiency framework, broadening it substantially from a normative perspective. While some studies have attempted to attach explicit weights to benefits and costs depending on the characteristics of who bears them, most analysts suggest an alternative approach—namely that the distributional consequences of a proposed resource reallocation be displayed as a supplement to the cost–benefit calculation based purely on efficiency. Such a display enables decision-makers to perceive fully the distributive implications of the proposals being discussed.

2.5 The Measurement of Social Benefits

How can willingness to pay be measured? It would be ideal if analysts could elicit willingness-to-pay amounts directly from individuals through structured conversations. If securing this information from all citizens were too costly or impossible, surveys of population samples could be used as the basis for estimating aggregate willingness to pay for the population. This approach, called 'contingent valuation,' is conceptually attractive because it potentially allows the analyst to elicit a willingness-to-pay amount for impacts that involve both active and passive use.

The use of contingent valuation to measure willingness to pay has become increasingly common, especially in the evaluation of environmental policies

(Bateman and Willis 2000). A blue ribbon panel convened by the US National Oceanic and Atmospheric Administration concluded that contingent valuation could be a reasonable basis for estimating passive use values in natural resource damage assessment (Arrow et al. 1993).

Nevertheless, the use of contingent valuation remains controversial. One source of concern is the cognitive demands that it places on respondents in terms of understanding the policy being valued. Unless respondents understand what is being valued, and believe their answers to be consequential in influencing the policy choice, they cannot be expected to give meaningful answers to serve as the basis for an inference about the willingness to pay of the entire population. Considerable research, mainly by environmental economists, is contributing to the craft of designing effective questionnaires for contingent valuation. Another source of concern is the fear that respondents will answer strategically in the sense of giving false answers that they believe will lead to a better outcome for themselves than would result from truthful answers. Both theory and evidence suggest that giving respondents referendum-type questions about public goods minimizes the dangers of strategic responses. That is, rather than directly asking a respondent to state her willingness to pay, she is given a random dollar amount and asked if she would vote for the policy if it would cost her that amount in higher taxes or increased costs for other goods. An estimate of mean or median willingness to pay for the sampled population can be inferred statistically from sample responses.

The conceptual and practical problems in conducting contingent valuations commonly lead analysts to estimate willingness-to-pay amounts through inferences from observation of the behavior of individuals, especially in markets. Demand analysis plays an especially important role in such estimation. A person's demand schedule for a good tells how much of a good the person wishes to purchase as a function of the price of the good, holding all other prices and the person's utility constant. The height of the demand schedule at each level of consumption gives the person's willingness to pay for an additional unit of consumption. The difference between the willingness to pay for this unit and the amount that the consumer actually pays is its 'consumer surplus.' Adding up the surpluses for each of the units consumed gives the total consumer surplus that accrues to the person from participation in the market or experiencing services produced by the public sector.

If a policy measure either satisfies a demand that has not been met, or changes the price of a good or service that a person is already consuming, the consumer surplus of the person will be changed. This change is the person's willingness to pay for the impact of the policy measure. (Graphically, the consumer surplus is the area under the demand schedule but above the market price from zero units to the number of units actually purchased.) If the good or service produced by the public sector is a private good, in that it does not generate external, or spillover, costs or benefits, then the market demand schedule for the good (that is, the horizontal sum of all the individuals' demand schedules) provides the basis for estimating changes in consumer surplus, and hence changes in the aggregate willingness to pay of individuals for the effects of the policy.

The market demand schedule must be estimated from observed price and quantity data. The standard econometric procedures result in the direct estimation not of the demand schedule described above, which holds the utilities of individuals constant (the Hicksian or compensated demand schedule), but of one that holds their incomes constant (the Marshallian or market demand schedule). Quite often these demand schedules are sufficiently similar so that measuring consumer surplus in terms of the market demand schedules provides good approximations of the aggregate of individuals' willingness-to-pay amounts (Willig 1976).

Estimation of willingness to pay is more difficult when policy impacts do not correspond to changes in markets for traded goods. For example, though most people would be willing to pay positive amounts for improvements in environmental quality, there is no market for this public good. People may convey information about their willingness to pay for changes in environmental quality, however, by their action in other markets. So, for instance, housing values may change to reflect changes in levels of noise pollution near airports, thereby providing a so-called 'shadow' price for quiet. In general, shadow prices refer to inferences of marginal social value when it is not revealed through market prices.

This approach is based on what is called the 'hedonic price model,' which provides a theoretical basis for statistically isolating the independent effects of the various characteristics of a product on price. It has been used extensively to estimate the benefits of air quality improvements and health risk reductions from wage and housing value variation. An important use of the hedonic price model in CBA has been to estimate the 'value of life,' defined as the amount that people are willing to pay for reductions in the risks of death that they face (see Viscusi 1993). Analysts can take advantage of prior research to find estimates of relevant shadow prices (Boardman et al. 1997). For example, a large number of studies conducted across a number of countries suggest that commuters value reductions in their travel times at between 40 and 50 percent of their after-tax wage rate (Waters 1996).

2.6 The Concept of Social Costs: Opportunity Costs

The implementation of policies usually requires the use of economic inputs that could be used to produce

other things. CBA uses the concept of 'opportunity cost' to value these particular policy impacts. The opportunity cost of an input is the value of the goods and services that the input would have produced in its best alternative use. This value is just the willingness to pay for these goods and services on the part of those who would have consumed them if they had been produced.

Opportunity cost may or may not correspond to the accounting cost of inputs used to implement policies. If purchasing the input in a competitive market does not alter the price in the market, then opportunity cost and accounting cost are identical. For example, buying books for a local reading project is unlikely to affect prices so that the purchase cost of the books represents their opportunity cost. If purchasing the input in a competitive market increases the market price, then opportunity cost will be less than the accounting cost based on the post-purchase price. For example, buying concrete for a dam in a local market may drive up the price of concrete, as the marginal costs of producing concrete increase. However, some of the payments at the higher price will be to more efficient suppliers with lower marginal costs. These payments in excess of marginal costs, called 'rents,' are simply transfers from the project budget to the efficient suppliers and therefore do not contribute to opportunity cost. The opportunity cost of inputs purchased in noncompetitive markets may be larger, smaller, or equal to their accounting cost. For example, the opportunity cost of the time of people compelled to serve on juries may be larger or smaller than the legally set payments they receive from the court.

2.7 *The Measurement of Social Costs*

The problems encountered in measuring social opportunity costs are very similar to those confronted in measuring social benefits. A commonly encountered problem in CBA is the measurement of the opportunity cost of unemployed labor. The conceptually correct measure is the value of the forgone leisure to each worker. The accounting cost of the project is certainly an overestimation of the opportunity cost of otherwise unemployed labor. Alternatively, zero is probably an underestimation of the opportunity cost. Rarely do analysts have enough information about the value of the forgone leisure to estimate the conceptually correct opportunity cost for unemployed labor so they usually use some fraction of accounting cost as an approximation (Haveman and Krutilla 1968).

Another commonly encountered problem in CBA is the appropriate measurement of the opportunity cost of government revenues. Standard practice is to treat a dollar of accounting cost raised through taxation as the equivalent of a dollar of opportunity cost. Raising revenue through taxes, however, usually involves inefficiencies, or 'deadweight losses,' that reduce social welfare. The amount by which the social cost of a dollar of tax revenue exceeds its nominal value is called its excess marginal burden. Expenditures funded by tax increases, and revenues allowing tax reductions, should be adjusted by the marginal excess burden of the relevant tax. A number of estimates of marginal excess burden have been made for various components of the US tax system (Jorgenson and Yun 1990).

An appropriate accounting of opportunity costs helps clarify when the inclusion of so-called secondary benefits is correct (Haveman and Weisbrod 1975). The primary benefits of a project derive from its direct effects. For example, the primary benefit of stocking a lake with game fish is to increase the supply of fishing days. The improved fishing might very well have secondary effects, such as the construction of a nearby hotel. It is usually incorrect to attribute any benefits to the secondary effects because they involve offsetting opportunity costs—the hotel requires resources that could have been used to build it elsewhere. Such secondary effects may involve benefits, however, if they use otherwise unemployed resources such as labor. In such cases, the benefits produced may not be fully offset by opportunity costs. A final point on opportunity cost: resources already expended have zero opportunity cost if they are no longer available for other uses. These previously expended resources are known as 'sunk costs.' Because it is forward looking, CBA ignores sunk costs.

3. *Timing of Costs and Benefits: Discounting*

The social benefits and costs generated by a public program often extend over long periods. Because waiting is costly, the amount that people value a dollar of benefits or costs in the future is less than if that value were obtained in the present. In an ideal economy, markets for lending and borrowing funds (also known as capital markets) balance the preferences of lenders who require a payment for giving up use of their funds (hence, delaying consumption) and the preferences of borrowers who invest in projects with future payoffs. In this situation, the market interest rate that is the outcome of this balancing serves as the basis for adjusting costs and benefits occurring in the future to make them commensurate with costs and benefits occurring in the present. This adjustment for time is called discounting. In particular, if $\mathbf{r}$ is the annual market interest rate, then a dollar of cost or benefit that will be realized after N years in the future would have a present value of $1/(1+\mathbf{r})^N$. Maximizing the present value of net benefits would maximize potential Pareto efficiency.

This simple solution results if capital markets are competitive and not restricted or distorted, for example because of taxes on savings (consumption) or on the returns from investment. When such distortions

exist, the solution becomes more complicated. Such taxes insert wedges between the marginal valuation of funds by lenders and borrowers. With such wedges, it is not clear which of the values is appropriate for discounting: the before-tax rate of return of borrowers who invest (the marginal rate of return on private investment), or the after-tax return faced by lenders (the marginal rate of pure time preference). Another distortion is general price inflation. Market interest rates reflect expectations about inflation. They correspond to projections of costs and benefits in inflated (nominal) dollars. When costs and benefits are projected in terms of current (real) dollars, typically the most convenient method, nominal discount rates must be converted to real discount rates.

Although some controversy remains, most analysts do seem to agree that the social discount rate should reflect the fact that both consumption and investment are foregone when the public sector uses resources. Thinking of the discounting problem in terms of shadow price of capital provides a way of taking account of these effects (Bradford 1975, Boardman and Greenberg 1998). It involves using the marginal rate of return on private investment to convert any foregone investment to the stream of future consumption that it would have produced. The consumption equivalents of foregone investment are then added to any foregone consumption to express all resource costs in terms of consumption, which can then be appropriately discounted using the marginal rate of social time preference.

4. *Accounting for Uncertainty*

As CBA involves prediction, it can rarely be done with certainty. The standard approach to dealing with uncertainty is to estimate expected net benefits. The first step is to convert the problem from one of uncertainty to one of risk by assuming a plausible distribution of benefits. The second step involves finding the expected value, basically the weighted average, over this distribution. Third, society is assumed to be risk-neutral, so that the expected values can be treated as commensurate with certain values. For example, a project with certain opportunity costs of \$15 million and a 20 percent chance of yielding \$5 million in benefits and an 80 percent chance of yielding \$20 million in benefits would have expected benefits of \$17 million and expected net benefits of \$2 million.

The calculation of expected net benefits, however, does not fully take account of individual preferences for risk. Individuals may not be willing to trade certain amounts for expected values of the same magnitude. In such cases, the appropriate measure of benefits is 'option price,' the individual's willingness to pay for the particular distribution of outcomes associated with a policy (Graham 1981). As individuals are generally risk averse, policies that reduce risk generally have option prices larger than expected benefits. The difference, 'option value,' is a premium individuals are willing to pay for the reduced risk. As option price can only be elicited through contingent valuation surveys, it is rarely estimated. Unfortunately, relatively little progress has been made in the theoretical determination of the magnitude or even sign of option value. For policies involving substantial risks, especially if the risks are social in the sense that everyone realizes the same outcome, the use of expected benefits may introduce substantial error into CBA. Even when the policy being evaluated is not inherently risky, analysts may be uncertain about their predictions of its costs and benefits. It is common, therefore, for analysts to test how sensitive their estimates of costs and benefits are to particular assumptions. This process, called 'sensitivity analysis,' typically involves changing the assumed values of a few key parameters to see how net benefits change.

5. *The Maximum Net Benefits Rule*

Assuming that all of the costs and benefits of a policy have been correctly measured, discounted, adjusted for uncertainty, and aggregated, positive net benefits indicate that the policy is potentially Pareto improving: the excess of benefits over costs would permit the policy to be implemented in conjunction with a set of compensatory transfers such that at least someone would be better off without anyone being worse off. Thus selecting policies on the basis of positive net benefits, the Kaldor-Hicks criterion discussed above, guarantees that adopted policies are potentially, but not necessarily actually, Pareto-improving.

The rule can be stated as follows: choose from among all possible policies those that offer the largest sum of net benefits. In the special case of mutually exclusive policies, the one with the largest net benefits should be selected. In the special case of completely independent policies, all those with positive net benefits should be selected. Note that this rule says nothing about a benefit–cost ratio, which is a common calculation made by CBA analysts. The use of a benefit–cost ratio criterion is appropriate only if the policies are independent of each other, in which case the set of projects with benefit–cost ratios greater than one corresponds to the rule of adopting policies with positive net benefits.

6. *Further Reading*

A number of textbook-length treatments of CBA are available for those wishing to explore it in much more depth (Boardman et al. 2001; Dinwiddy and Teal 1996, Zerbe and Divily 1994, Gramlich 1990, Sugden and Williams 1978).

See also: Optimization in Evolution, Limitations of; Pareto, Vilfredo (1848–1923); Policy History: State and Economy; Rationality in Society; Welfare: Philosophical Aspects; Welfare Programs, Economics of

Bibliography

Arrow K, Solow R, Portney P, Leamer E, Radner R, Schuman H 1993 Report of the NOAA Panel on contingent valuation. *Federal Register* **58**: 4601–14

Bateman I J, Willis K G (eds.) 2000 *Valuing Environmental Preferences: Theory and Practice of the Contingent Valuation Method in the US, EC, and Developing Countries*. Oxford University Press, Oxford, UK

Blackorby C, Donaldson D 1985 Consumers' surpluses and consistent cost–benefit tests. *Social Choice and Welfare* **1**: 251–62

Boardman A E, Greenberg D H 1998 Discounting and the social discount rate. In: Thompson F, Green M T (eds.) *Handbook of Public Finance*. Marcel Dekker, New York

Boardman A E, Greenberg D H, Vining A R, Weimer D L 2001 *Cost–Benefit Analysis: Concepts and Practice*. Prentice-Hall, Upper Saddle River, NJ

Boardman A E, Greenberg D H, Vining A R, Weimer D L 1997 Plug-in shadow price estimates for policy analysis. *Annals of Regional Science* **31**: 299–324

Bradford D F 1975 Constraints on government investment opportunities and the choice of discount rate. *American Economic Review* **65**: 887–99

Dinwiddy C, Teal F 1996 *Principles of Cost–Benefit Analysis for Developing Countries*. Cambridge University Press, New York

Graham D A 1981 Cost–benefit analysis under uncertainty. *American Economic Review* **71**: 715–25

Gramlich E M 1990 *A Guide to Benefit–Cost Analysis*. Prentice Hall, Upper Saddle River, NJ

Haveman R H, Krutilla J V 1968 *Unemployment, Idle Capacity, and the Evaluation of Public Expenditures*. Johns Hopkins University Press, Baltimore, MD

Haveman R H, Weisbrod B A 1975 Defining benefits of public programs: some guidance for policy analysts. *Policy Analysis* **1**: 169–96

Hochman H M, Rodgers J D 1969 Pareto optimal redistribution. *American Economic Review* **59**: 542–57

Jorgenson D W, Yun K Y 1990 Tax reform and US economic growth. *Journal of Political Economy* **98**: S151–93

Krutilla J V 1967 Conservation reconsidered. *American Economic Review* **57**: 777–86

Steiner P O 1974 Public expenditure budgeting. In: Blinder A S, Solow R M, Break G F, Steiner P O, Netzer D (eds.) *The Economics of Public Finance*. Brookings Institute, Washington, DC

Sugden R, Williams A H 1978 *The Principles of Practical Cost–Benefit Analysis*. Oxford University Press, New York

Viscusi W K 1993 The value of risks to life and health. *Journal of Economic Literature* **31**: 1912–46

Waters W G 1996 Values of travel time savings in road transport project evaluation. In: Hensher D et al. (ed.) *World Transport Research*. Elsevier, New York, Vol. 3

Weisbrod B 1964 Collective consumption services of individual consumption goods. *Quarterly Journal of Economics* **78**: 71–7

Whittington D, MacRae D 1986 The issue of standing in cost–benefit analysis. *Journal of Policy Analysis and Management* **5**: 665–82

Willig R D 1976 Consumer's surplus without apology. *American Economic Review* **66**: 589–97

Zerbe R O 1998 Is cost–benefit analysis legal? Three rules. *Journal of Policy Analysis and Management* **17**(419): 56

Zerbe R O, Divily D D 1994 *Benefit–Cost Analysis in Theory and Practice*. HarperCollins, New York

R. H. Haveman and D. L. Weimer

Counsel, In-house

1. Introduction

In-house counsel are legally trained individuals who have responsibilities for some of the law jobs of the organizations employing them. Organizations having needs for legal services may resolve them either internally or with legally trained individuals whose services are purchased in the market. Organizations may chose to 'make' rather than 'buy' legal services, as they do in hiring in-house counsel. Hiring in-house counsel is an instance of organizations 'professionalizing' their workforce. Sociological research on in-house counsel has focused on in-house counsel being both organizational and professional actors.

2. Delimiting the Role of In-house Counsel

Many individuals with legal training work for organizations in other than a distinctly legal capacity. In most of the world, university legal study usually does not lead graduates into a career as a legal professional. These graduates take a facility in the law into whatever positions they assume. Others may leave a career in law to work for organizations in positions other than that of in-house counsel. In some countries, including the USA, individuals may combine being lawyers with careers as organizational actors.

In-house counsel do not differ from these other law-trained employees because they are primarily employed for their law training and practice experience. Corporate finance departments, for example, may employ many law graduates, some with significant careers as lawyers, to perform law work. These individuals, however, have not been considered to be in-house counsel. Although they perform legal work and are law graduates, they are responsible to the corporate finance department, not to the organization's general counsel. An organization's law department is the home of in-house counsel.

In-house counsel are responsible for only a limited part of the law jobs of an organization. Legal work is performed in organizations by other than in-house counsel. Executives, staff members, regulatory specialists, managers, and salespeople all may have degrees

and experience in law and be responsible for a number of law jobs. Other law jobs may be performed by individuals who do not have any formal legal training. They may negotiate, write, and sign contracts, or deal with banks and enter into loan agreements. Organizations may have such continuing relations with regulators that they develop specialized subunits, such as employment or environmental relations, with responsibilities for legal compliance and dispute resolution.

With in-house counsel, organizations delimit a distinctly legal capability, but in-house counsel's responsibilities are neither co-extensive with the organization's law jobs nor comprehensive of organizational law compliance. In-house counsel are employed lawyers who provide the organization with a professionally managed legal capability (Mackie 1989). Organizations vary greatly in how they manage in-house counsel. Depending on the legal regime and its professionalization project among other factors, in-house counsel differently attest to the organization's legal function.

3. *In-House Counsel as Organizational Actors*

In-house counsel vary in their work responsibilities. Some organizations make in-house counsel rather than buying their legal services across the market, in order to increase the economic and efficient meeting of recurrent predictable legal needs. Others may hire in-house counsel because of the firm-specific knowledge needed in providing legal services. Others may hire in-house counsel in imitation of other organizations. Yet others, risking reducing organizational flexibility, hire in-house counsel to increase the organization's coordination and control of legal services (Chayes and Chayes 1985).

The extent to which in-house counsel assume organizational responsibilities varies greatly. In-house counsel perform a staff function, but, either individually or organizationally, they also may command executive functions. Hiring in-house counsel attests to the jurisdictional claims of law graduates. In-house counsel may extend beyond fulfilling a staff function by making professional jurisdictional claims, for example, gain a voice in the choice of remedy for legal problems.

In the USA, in-house counsel make professional claims to organizational power. The rules of the legal profession require in-house counsel and legal departments to have access to the Board of Directors. Some in-house counsel secure these claims by emphasizing their skills in handling legal risk. Others do so through their unique and nontransferable skills. In gaining organizational responsibility, in-house counsel become managers of aspects of the organization's legal services.

Today, in the USA business sector, many in-house counsel are involved in compliance issues and strategic decisions that gain them access to the ultimate organizational decision makers. US corporate counsel have won the right to monitor and review many legal services produced internally or purchased across the market by the organization. Often, they are sought for their judgment about decisions where legal risks do not predominate (Rosen 1989).

In many economies, corporate in-house counsel are less frequently ceded executive responsibility (see e.g., Van Houtte 1999, Daly 1997). In most of the business world today, in-house corporate counsel resemble US corporate counsel of an earlier era. Although situations varied, depending on the trust and confidence particular in-house counsel attained, in-house counsel performed scrivener and secretarial tasks, carrying out decisions reached without their significant input (Rosen 1989). But even in economies where in-house counsel have few responsibilities, global corporations located there use in-house counsel as do modern US ones (Daly 1997).

Outside the business sector, in-house counsel for organizations—especially public interest organizations—face the possibility—and are cognizant of the attendant responsibilities—of performing executive tasks in the organizations or causes they take as clients (Sarat and Scheingold 1998).

Even where in-house counsel have executive power, they share control with others. As a matter of professional ethics, lawyers cede some forms of control to their client, the organization. Decisional authority over the scope and content of legal work, however, may be allocated within an organization by dynamics driven by other than the norms of the legal profession. Students of organizations may study inside counsel to gain information about organizational decision making, especially in regards to certain aspects of regulatory compliance.

4. *In-house Counsel as Professional Actors*

Professions whose professionalization projects stress autonomous employment as a hallmark of the profession—the 'liberal' professions—may denigrate employed lawyers for lacking professional independence. During the progressive era, in the USA, in-house counsel were called 'kept lawyers.' In the European Union, unlike self-employed lawyers, in-house counsel are not legally privileged to protect the confidentiality of their work product. In many civil law countries, a lawyer must resign from the bar before taking salaried employment. A 'heteronomously employed' legally trained individual who performs legal work may not be professionally respected as a 'lawyer.'

The law jobs performed by in-house counsel may be bought by contracting with external providers. In the marketplace, there are sellers of legal services in professional law firms, multi-disciplinary practices, and competing professions. In-house counsel compete with autonomously employed lawyers who compete with accounting-consulting firms, regulatory engin-

eering specialists, and business advisors of many sorts. In the USA, this competition has altered the distribution of deference within the legal profession, with independently employed lawyers no longer denigrating in-house counsel.

The substantive claim in the professional denigration of in-house counsel is their supposed lack of professional independence. Organizational control over the work of professionals, from a legal perspective, creates a potential conflict of interest. Although the control may better align legal services with the interest of the organization, it may cause conflicts with legal interests, sometimes crystallized in the personal and professional interests of in-house counsel. In-house counsel may experience conflicts of interest that will shape their professional judgments.

The professional independence of in-house counsel is relative to both that which can be bought in the market and the organizational role in-house counsel assume.

The independence of external providers of legal services is relative to the organization's abilities to control them. By increasing the organization's ability to monitor the legal services purchased, in-house counsel may decrease external providers' abilities to effect their independence. Sometimes, in-house counsel may demand that external providers be independent. Other times, their influence on whom and when to hire in the external market can decrease external providers' willingness to exert their independence. In-house counsel also decrease the abilities of external providers to be independent by inhibiting personal relations with organizational members or by limiting either the time frame in which the external providers respond or the problem posed. By retaining authority to select external providers, define the goals of their work and define the problem that is the subject of legal work, organizations control what legal services they purchase in the market, decreasing the professional independence of external providers (Kagan and Rosen 1985).

The independence of in-house counsel is also relative to the organizational roles they assume. First, in-house counsel may be delegated or may assume organizational decisional authority. In-house counsel vary in the extent to which they are assigned thick responsibility in their organizations. Second, organizations differ not only in their enacted cultures, but also in the extent to which they seek to gain the advantages of in-house counsel commitment by supporting job-enrichment schemes mirroring professional privileges. Third, in-house counsel differ not only in their professional orientations, but also in their abilities to draw on professional resources of independence. Fourth, in-house counsel differ in the human capital they bring to their jobs. In-house counsel may be trusted organizational actors, using that trust to exercise their professional independence. Fifth, organizations differ in their internal politics, so that in-house counsel may have opportunities to assert their independence by bargaining. And sixth, in-house counsel may be induced to assert their independence as the addressees of laws or as a result of threats of their own legal liability.

Legal professions differ in where and how they establish the lack of plasticity in professional work that is deemed to warrant the assertion of professional independence. During much of this century, there have been calls that in-house counsel serve as consciences of their organizations. In the US business sector, where in-house counsel have attained organizational power, this is no longer on the agenda. Nor on the agenda are limits on what information about the law in action in-house counsel may professionally provide. Students of professions may study inside counsel to gain information about the plasticity of the professional legal role.

In their work, in-house counsel attest to the law's legitimacy for the organizations they serve. In their work, in-house counsel apply professional knowledge. Inside counsel and their work create subjects for research for students of both organizations and professions.

See also: Lawyers; Lawyers, Regulation of; Legal Professionalism; Organizational Decision Making; Professions in Organizations; Professions, Sociology of

Bibliography

Chayes A, Chayes A H 1985 Corporate counsel and the elite law firm. *Stanford Law Review* **37**: 277–300

Daly M C 1997 The cultural, ethical and legal challenges in lawyering for a global organization: The role of the general counsel. *Emory Law Journal* **46**: 1057–111

Donnell J 1971 *Corporate Counsel: A Role Study*. Bureau of Business Research, Graduate School of Business, Indiana University, Bloomington, IN

Kagan R, Rosen R E 1985 On the social significance of large law firm practice. *Stanford Law Review* **37**: 399–443

Kolvenbach W 1979 *The Company Legal Department: Its Role, Function and Organization*. Kluwer, Deventer, The Netherlands

Mackie K J 1989 *Lawyers and Business and the Law Business*. Macmillan, London

Rosen R E 1989 The inside counsel movement, professional judgment and organizational representation. *Indiana Law Journal* **64**: 479–553

Sarat A, Scheingold S (eds.) 1998 *Cause Lawyering: Political Commitments and Professional Responsibilities*. Oxford University Press, New York

Spangler E 1986 *Lawyers for Hire: Salaried Professionals at Work*. Yale University Press, New Haven, CT

Van Houtte J 1999 Law in the world of business: Lawyers in large industrial enterprises. *International Journal of the Legal Profession* **6**: 7–25

R. E. Rosen

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Counseling and Psychotherapy: Ethnic and Cultural Differences

1. Introduction

Scholars and mental health professionals have argued that racial-cultural issues have been obstacles for those seeking help (Sue and Sue 1999). When counseling has been available, the focus has often been placed upon failures to create realistic understanding of members of racial-cultural groups in the United States, practitioners' unfamiliarity of historical, social, and political discrimination and oppression (Guthrie 1998), narrow professional cultures, and biased theoretical models (Carter 1995).

In many regards, scholars, educators, and practitioners have been focusing upon changing population demographics as a driving force necessitating multicultural inclusion. For example, the US Bureau of the Census data (1990) has shown significant increases in the number of Blacks, American Indians, Asian/Pacific Islanders, and Latinos. Between 1970 and 1990, the Latino population increased 145 percent, the Asian/Pacific Islander population by 375 percent with 22.3 million Latinos and 7.2 million Asian/Pacific Islanders now living in the USA. Currently one in three US Americans has been either born outside of the US or are members of visible racial-cultural groups. Given the present rate of immigration as well as fertility rates, researchers at the Bureau of Immigration have predicted by the year 2040 Whites will only comprise 62 percent of the total US population.

Recognizing the importance of meeting the needs for all seeking mental health services, scholars and educators have developed competency guidelines and ethical mandates for integrating multicultural perspectives into psychology research, practice, and training (American Psychological Association 1991, Sue et al. 1982, 1992, 1998), educational programs (Carter 1995, Ridley et al. 1994), and evaluation instruments (Pope-Davis and Dings 1995).

2. Culture

What is culture? Culture at its core understands human behavior through causes and effects *and/or* content and processes. While cultural content is focused on knowledge about learned and acquired beliefs, art, morals, customs, and habits, cultural processes attend to the process and perpetuation of pattern transmission. Thus, psychosocial cultural perspectives are socially transmitted attitudes and behaviors emerging from interpersonal interactions transmitted generationally, not through biological inheritance but by formal and informal methods of teaching and demonstration. Thus, culture can include historical dimensions emphasizing traditions, normative dimensions focusing upon unique and shared experiences, psychological dimensions including adjustment, learning and group behaviors, structural dimensions attending to the composition of and interrelated patterns in organizations and, genetic dimensions focusing on the origins of culture as interactive adaptive processes between groups and their environment. Broadly speaking, culture has bound groups of people by common explicit (e.g., tangible artifacts, clothing, food, festivals, and behaviors readily observed by outsiders) and/or implicit (e.g., intangible conscious and unconscious values), assumptions, attitudes, beliefs, and feelings and has endured over time and space. Additionally, while people may be identified with one or more groups, the intensity and resilience of belongingness can vary with the interaction of self-investment and degree of group acceptance.

How people's racial-cultural assumptions shape the way in which culture is embraced, rejected, and ignored has been described in Carter's (1995) classification system. In the *Universal* paradigm, shared similarities are affirmed as universal human experiences and differences as secondary intragroup variations. In the *Ubiquitous* paradigm, socially derived circumstances constitute cultural groups. From this perspective, people embrace one or more socially derived cultural group affiliations. Unlike the Universal paradigm, the Ubiquitous paradigm acknowledges and accentuates differences by promoting cultural sensitivity to increase people's awareness of variation. Influenced by anthropology, the *Traditional* paradigm equates culture with geography (e.g., country) and adaptive socialization (e.g., birth, upbringing). Since cultural membership circumscribes personality development by experiencing 'cultures' through exposure and cultural informants, culture is understood as personal circumstances superseded by the dominant culture. In the *Race-based* paradigm, culture is a function of transcending racialized experiences. Becoming aware of the effects of racism, oppression, and promoting racial identity development, culture is understood by how race and racism influence power and authority in the USA. In the fifth *Pan-National* paradigm, culture is constructed as how racial group affiliation and the distribution of power influence global issues.

2.1 Cultural Pluralism and the Tyranny of the Shoulds

Historically, group relations between White Anglo and all other racial-cultural groups in the USA have influenced psychological, social, and cultural adjustment for members of dominant (e.g., mainly White Anglo-Saxon) and subordinate cultural (e.g., visible racial-cultural) groups. For example, from the mid 1700s to the early 1880s, the melting pot concept

(Zangwill 1910) assumed all entering US inhabitants would willingly relinquish indigenous heritages and adopt existing White Anglo-Saxon cultural norms. From the late 1880s to the mid-1950s, this biologically blended melting pot became '*Americanized*' (Fairchild 1926). Through Americanization, the ideology of *cultural pluralism* or *cultural unity* advocated including visible racial-cultural groups into 'mainstream' American culture. Symbolic appreciation for diversity, especially from other civilizations beside White-European societies through congressional legislation and Supreme Court decisions in the 1960s and 1970s furthered the cultural pluralist movement. However, while White European immigrants were expected to assimilate dominant White Anglo-Saxon cultural norms, visible racial–cultural groups were still excluded and institutionally segregated.

Thus, cultural pluralism ideology was contradictory to the realities of the American 'melting pot.' While preserving cultural and social heritages as a means to creating a new society, economic, political, and educational separatism for visible racial-cultural groups was creating a 'tyranny of the shoulds'—idealized images and standards imposed upon visible racial-cultural people by the dominant society resulting in the suppression of the 'true self' (Horney 1950). Despite conformist trends throughout the nineteenth century, voices of dissent and the 'tyranny of the shoulds' was heard in the words of W.E.B. Dubois as:

> … a peculiar sensation, this double consciousness, this sense of always looking at one's self through the eyes of others, of measuring one's soul by the tape of the world that looks on in amused contempt and pity. One ever feels his two-ness – an American, a Negro; two souls, two thoughts, two unreconciled strivings; two warring ideals in one dark body, whose dogged strength alone keeps it from being torn asunder (1989, p. 3).

At some level, cultural pluralism may have enabled visible racial-cultural people to exist in threatening environments. However, the consequences of vigilance against repressive environments also created unnecessary personal and group self-doubt, alienation, and isolation from resources and support, thus limiting a person's ability to achieve their full potential.

2.2 Cultural Pluralism and Cross-cultural Psychology

While cross-cultural, diversity, multicultural, and cultural pluralism have been used interchangeably, these terms have generally referred to people who are members of visible racial-cultural groups (e.g., American Indians, Blacks, Asians, Latinos). Recognizing the health disparities of these visible racial-cultural groups, the cross-cultural psychology movement has been focused upon improving mental health services for members of these groups. Thus, cross-cultural psychology presumed that 'cross-cultural counseling problems were most likely to occur when there was a low degree of client–counselor assumed similarity… [where] the practitioner is…a member of the majority group and the client is a minority group member…[and] include[s]…counselor/therapist and client…minority individuals, but represent[ing] different minority groups ([B]lack-Hispanic, Asian-American-American Indian, (Sue et al. 1982, p. 47). By appreciating group differences, cross-cultural psychologists believed that understanding racial-cultural group specific knowledge (e.g., values, customs, communication patterns) would eliminate racial-cultural mental health disparities. For example, misunderstandings arising from cultural variations in values and language were suggested as antecedents to early psychotherapy terminations at rates greater than 50 percent compared to 30 percent for White clients (Sue and Sue 1999).

Recognizing the majority of mental health educators and practitioners are White, professional organizations developed guidelines endorsing cross-cultural culture-specific heritage education about specific visible racial-cultural groups (e.g., Blacks, American Indians, Asian and Pacific Islanders, Latinos) in hopes that sensitizing White practitioners to issues of '*culturally different*' people, practitioners would become more effective counselors (American Psychological Association 1991, Sue et al. 1982, 1992). For example, a plethora of information exists about group specific racial-cultural knowledge about self-expression, norms, family structures, and values for 'special populations' (e.g., restraint of emotions and hierarchical relationships for Asian American clients, extended family and religion for Latino and Black clients and harmony and cooperation among American Indians). (It should be noted that these categories have been recently expanded to include Whites, socioeconomic status, ethnicity, religion, gender, age, etc. and effects of institutional racism, classism, and sexism (Sue and Sue 1999).)

Thus, cross-cultural educational strategies have virtually used 'othered' focused (Sue and Sue 1999) or universal (Fischer et al. 1998, p. 561) perspectives when teaching about cultural issues. If cultural pluralism and cross-cultural psychology have been effective in equity and inclusion of underserved visible racial-cultural groups, why does racial-cultural separatism continue to exist?

Several reasons can account for this ongoing separatist phenomenon. Overall, cultural pluralism has not provided meaningful principles about psychological well-being. Beyond obvious axioms such as all humans have language, social systems, emotions, and values and cultural norms, identification of group understanding perceived specific racial-cultural groups by group comparisons has neglected existing within group variation. The results have been simplistic

caricatures about service delivery, cultural norms, and group homogeneity. Consequently, using a social constructivist approach, multicultural psychologists have begun developing more inclusive, affective, and intrapsychic approaches to applied mental health.

2.3 Moving From Cross-cultural to Multicultural Psychology

Understanding people from a multicultural perspective has meant simultaneously recognizing individual and group identities. From this perspective, culture accounts for internal and external personal and social realities by becoming distinct from biological and geographical heritage. In this regard, the importance of racial-cultural self-exploration of understanding explicit content knowledge and implicit affective processes has been most illustrative in the clinical bias literature where scholars have shown mental health professionals susceptibility to clinical bias (Oakes et al. 1994). Given issues of professional and personal bias, applied psychology and other mental health training programs have been remiss when addressing how personal assumptions influence professional development. Lacking has been a counselor-focused framework incorporating a multiculturally inclusive perspective. By challenging people's beliefs and assumptions about who they are as cultural people, mental health professionals can more realistically contextualize how social, historical, and political processes have influenced their personal and professional development

While the cross-cultural competencies (Sue et al. 1982) were expanded from 11 to 31 in 1992 (Sue et al. 1992), the 1998 multicultural competency guidelines (Sue et al. 1998) have taken a fundamental shift towards being more culturally inclusive by focusing upon issues of within group variation (e.g., counselor *and* client racial–cultural identity development into counseling relationships, interventions, and processes) and proposing greater latitude in defining professional roles and responsibilities. Here, multicultural unlike cross-cultural psychology has suggested mere intellectual understanding of visible racial-cultural groups norms does not meaningfully translate into adequate or sufficient counseling skills.

Thus, the movement from cross-cultural to multicultural psychology has involved a fundamental shift of cognitive learning about (e.g., between group differences) visible racial-cultural groups to an integrated cognitive-affective racial-cultural identity (within group/individual variation) orientation. Thus, clarifying ones' own racial and cultural identities, developing a sense of comfort and self-acceptance within a social, historical, and political framework, and respectfully relating and functioning with all racial-cultural groups have become essential prerequisites for developing multicultural competence (Carter 1995, Sue et al. 1998).

3. Multicultural Psychology and Racial-cultural Identity Development

People throughout time have been challenged to understand their sense of self and life experiences in unifying ways. As critical tools for cultural understanding, racial-cultural identity development models have been considered 'one of the most promising approaches to the field of multicultural counseling/therapy' (Sue and Sue 1999, p. 123). By incorporating sociopolitical contexts, racial identity theories have provided a structure of how to understand life experiences, perceptions, attitudes, and behaviors of one's own and other racial-cultural groups, identity formation and intra-group variation.

Diagnostically, racial-cultural identity models also have clinical utility by framing how individuals' experiences of counseling are influenced by their racial-cultural identity and not simply linked to presumed racial-cultural group affiliations (Carter 1995).

Racial-cultural identity attitude theories are dynamic identity models representing developmental ego differentiation and have existed for Blacks, other visible racial-cultural people (e.g., Asian, Latino, women), and Whites since the early 1970s (Sue and Sue 1999). Racial identity attitudes are the psychological consequences of racial socialization.

Whereas people are socialized in racialized environments, racial-cultural ego identity development is the psychological reaction to the benefits and consequences of racial-cultural social systems. Involving simultaneous perspectives of oneself in relation to members of dominant and subordinate groups, people with less developed racial-cultural ego identity statuses hold more simplistic and unexamined notions about race and race relations while greater differentiated racial-cultural ego identity statuses are based on more accurate and examined personal and group information and experiences. Thus, a person's presumed racial-cultural phenotype does not preclude racial-cultural orientation. Rather, a person's overall psychological self that has been validated, denied, or ignored influences the manner in which racial identity has been integrated into one's personality and psychological well-being. By exploring the psychological aspects of how feelings, thoughts, and behaviors people have about their own and other's race, a more realistic understanding of *who* a person rather than *what* a person is becomes possible.

3.1 Racial Identity Theory

Based upon internal racial identification levels, racial identity models have provided a framework for understanding intra-group socio-racial psychological perspectives. According to Black racial identity theory (and more broadly, People of Color Identity Theory), Black people develop from (a) overvaluing and idea-

lizing White cultural norms and internalizing negative stereotypes of Blacks (Preencounter); (b) feeling confusion as they try to understand their previously held beliefs of White superiority and Black inferiority (Encounter); (c) affirming 'Blackness' and denigrating 'Whiteness' by idealizing African cultural heritage and withdrawing from White society (Immersion); (d) accepting strengths and weaknesses of Black people and incorporating a more affirming and realistic Black identity (Emersion); (e) intellectually assessing and responding to Whites (internalization); and (f) integrating a positive Black identity while maintaining a balanced perspective of 'Whiteness' motivated by personal preferences rather than racial group self-denial (Integrative Awareness) (Carter 1995).

In White racial identity theory, identity development involves the abandonment of racism and the development of a non-racist White identity by moving from (a) being unaware of their own racial group and ignoring the race of others (Contact); (b) becoming aware of the saliency of Whiteness, feeling guilty over personal internal standards and societal norms about race and over-identifying with Blacks (Disintegration); (c) rejecting Blacks and over idealizing Whiteness (Reintegration); (d) intellectualizing rather than emotionally understanding race (Pseudo-Independence); (e) exploring biases and redefining Whiteness by trying to change Whites instead of Blacks (Immersion/Emersion); and (f) emotionally and intellectually internalizing a positive White identity through appreciation for and respect of racial differences and similarities (Autonomy) (Carter 1995).

Overall, for visible racial-cultural people, as members of less powerful sociocultural groups, the primary socioracial identity issue in developing racial-cultural identity means overcoming internalized negative stereotypes associated with their racial-cultural group affiliation. For Whites, as members of the dominant sociocultural group, racial-cultural identity development means overcoming entitled stereotyping, learning self-value and appreciation as a White person. Therefore, in efforts to serve as racial-cultural change agents, all people must seek and promote racial-cultural affirmation—the experience of having various facets of their racial-cultural identities validated for themselves and by others. Thus, the strength of racial-cultural identity models has relied upon their dynamic nature of understanding within group racial-cultural variation within a social, historical, and political context as well as their diagnostic value for informing counseling processes and outcomes.

Whether its socio-political forces vs. poor psychological functioning, racial-cultural issues are silent or powerful environmental and psychological realities in a person's development. By examining how racial-cultural issues influence psychological phenomenon, therapists can help clients discover how their racial–cultural ego identity forms a filter which thoughts, feelings, and behaviors organize perceptions and interpretations of thoughts, emotions, and psychological life. To this end, racial-cultural identity models provide a framework understanding personal racial-cultural intrapsychic phenomena with in and between people.

4. *Conclusions*

Since the 1970s, scholars, educators, and applied mental health professionals have become increasingly aware that for the needs of all clients to be met, multicultural psychology is a needed and essential change for the profession (American Psychological Association [APA] 1991). However, since the original cross-cultural counseling competencies position paper (Sue et al. 1982) to the latest multicultural competences (Sue et. al. 1998), educators' ambivalence towards programmatic multicultural education has continued through decreasing multicultural-related courses in counselor preparation programs (Hollis and Wantz 1994), lack of comprehensive training models (Ridley et al. 1994) and professionals who continue to perceive themselves as having limited multicultural counseling competence providing mental health services (Allison et al. 1994).

Since intellectual knowing does not necessarily translate into sufficient multicultural counseling competence, multicultural psychology in the twenty-first century must mean looking beyond the cross-cultural 'tyranny of shoulds' of cultural groups differences toward racial–cultural self-reflection. Since mental health professionals' worldviews are neither neutral or culture free, questions posed, studies designed, and results interpreted will be inherently influenced by their racial-cultural perspectives.

As Gordon (1999) so aptly states:

> I like to remind people who cannot recognize it that I am an African-American male, born in the United States in the 20th century, and that these things influence me and what I do. They influence the way I think about things. When I identify a problem, to investigate, I bring my African-American, 20th century, male person to bear on it. That person is likely to be somewhat different from some of my colleagues who do not share my particular background and identity.... each of us bring to our work may be different. Most important is that those perspectives may influence our respective findings as well as our interpretations – even though we may like to think of findings as being objective (p. 150).

Thus, cognitively and affectively understanding ones' own racial and cultural identities and developing a sense of comfort and self-acceptance are essential prerequisites for developing multicultural competence. Accordingly, cognitive-affective reflective learning of personal and social identities has been shown to be a powerful and effective method of developing multicultural counseling competence (Carter 1995).

For example, an ecological person-process training framework has been successful in allowing trainees to

gain multicultural counseling competence (Tomlinson-Clarke and Ota Wang 1999). Allowing examination of *who* a person is, this model has emphasized how intra-/intergroup processes influence personal and social racial-cultural identity development. Here, people have been conceptualized within interlocking dynamic systems of: personal individual interactions (microsystem); institutions directly influencing individuals (mesosystem); institutional policies (exosystem); and the overall zeitgeist (Tomlinson-Clarke and Ota Wang 1999).

Unfortunately, while everyone has potential to develop multicultural competence, sociocultural environments have often imposed barriers of *conformity*, *prejudice*, and *oppression*, limiting essential cognitive and racial-cultural flexibility. Moreover, to varying degrees, people have been overtly or covertly subjected to these obstructions. For example, in *conformity*—the prevalence of expected societal, institutional or self-imposed mythical ideals (e.g., White, blonde, blue-eyed people are smarter than those who have darker complexions) have often resulted in feelings of inferiority for those who do not fit these ideals and of superiority for those who do. Beyond limiting, these mythical ideals have also prevented others from recognizing the value of diversity. In *prejudice*—those who hold the power in society and in institutions have denied equity to people who are different from them by conveying messages of 'you can't be successful unless your phenotype, attitudes, and behaviors are similar to ours.' Because prejudice promotes group and cultural superiority, its destructive forces have artificially separated racial-cultural groups. And with *oppression*, this barrier has been socially and personally destructive due to mythical prescriptions for visible racial-cultural people to be *what they are not* while exploiting them for *being who they are* as means for keeping them from true full participation in society.

So how can social scientists nurture the development of multicultural psychology? A discipline hoping to understand, and value diversity must identify and eliminate overt and covert conformity, prejudice, and oppression in its institutions and members. Additionally, programmatic multicultural programs with goals of recognizing, respecting and learning from individual and cultural similarities and differences will combat the present day mythical social ideals and stereotypes resulting in the exploitation, coercion, and oppression of all people. While difficulties exist in eradicating xenophobia and notions of individual, group, and cultural superiority/inferiority, these obstacles will be outweighed by the ultimate rewards of all people having the opportunity to develop into their fullest personal, social, and racial-cultural potential. Thus, educators will need to continue developing training programs to incorporate cognitive/affective racial-cultural learning so counselors will be prepared to realistically work with inherent human variation. By clarifying ones' own racial and cultural identities and developing a self-acceptance, counselors will be able to respectfully and effectively work with all racial-cultural groups. Thus, the growth of multicultural psychology will depend on those who teach, research, and practice to continually develop themselves as racial-cultural people. Institutional support will provide the fertile ground necessary to do this work of moving beyond ignorance, tolerance, to meaningful understanding. Ultimately, the results will be a world where multicultural psychology is a lifestyle lending itself to an openness of understanding *who people are* rather than *what they are imagined to be*.

See also: Conflict Sociology; Ethnicity, Race, and Health; Gender, Class, Race, and Ethnicity, Social Construction of; Racial Relations

Bibliography

Allison K W, Crawford I, Echemendia R L, Knepp D 1994 Human diversity and professional competence. *American Psychologist* **49**: 792–6

American Psychological Association 1991 *Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations*. American Psychological Association, Washington, DC

Carter R T 1995 *The Influence of Race and Racial Identity in Psychotherapy: Toward a Racially Inclusive Model*. Wiley, New York

Dubois W E B 1989 *The Souls of Black Folks*. Bantam Classic, New York

Fairchild H P 1926 *The Melting Pot Mistake*. Little, Brown and Co., Boston, MA

Fischer A R,, Jome L M, Atkinson D R 1998 *The Counseling Psychologist* **26**: 525–88

Gordon E W 1999 *Education and Justice: A View from the Back of the Bus*. Teachers College Press, New York

Guthrie R V 1998 *Even the Rat was White: A Historical View of Psychology*. Allyn and Bacon, Boston, MA

Hollis J W, Wantz R A 1994 *Counselor Preparation 1993–1995*. Accelerated Development, Muncie, IN, Vol. II

Horney K 1950 *Neurosis and Human Growth; the Struggle Toward Self Realization*. Norton, New York

Oakes P J, Haslam S A, Turner J C 1994 *Stereotyping and Social Reality*. Blackwell, Oxford, UK

Pope-Davis D B, Dings J G 1995 Assessment of multicultural counseling competencies. In: Ponterotto J G, Casas J M, Suzuki L A, Alexander C M (eds.) *Handbook of Multicultural Counseling*. Sage Publications, Thousand Oaks, CA, pp. 287–311

Ridley C R, Mendoza D W, Kanitz B E 1994 Multicultural training: Reexamination, operationalization, and integration. *The Counseling Psychologist* **22**: 227–89

Sue D W, Sue D 1999 *Counseling the Culturally Different: Theory and Practice*. 3rd edn. J. Wiley and Sons, New York

Sue D W, Arredondo P, McDavis R J 1992 Multicultural counseling competencies and standards: A call to the profession. *Journal of Counseling and Development* **70**: 477–86

Sue D W, Bernier J E, Durran A, Feinberg, L, Pedersen P, Smith E J, Vasquez-Nuttall E 1982 Position paper: Cross-cultural counseling competencies. *The Counseling Psychology* **10**: 45–52

Sue D W, Carter R T, Casas J M, Fouad N A, Ivey A E, Jensen M, LaFromboise T, Manese J E, Ponterotto J G, Vasquez-Nuttal E 1998 *Multicultural Counseling Competencies*. Sage, Thousand Oaks, CA

Tomlinson-Clarke S, Ota Wang V 1999 A paradigm for racial-cultural training in the development of counselor cultural competencies. In: Kiselica M S (eds.) *Confronting Prejudice and Racism During Multicultural Training*. American Counseling Association, Alexandria, VA, pp. 155–67

US Bureau of the Census 1990 *Census of the population: Supplemental report. Race of the population by state*. US Government Printing Office, Washington, DC

Zangwill I 1910 *The Melting Pot*. Macmillan, New York

V. Ota Wang

Counseling Psychology

This article is about the science and profession of counseling psychology. Key distinguishing features of counseling psychology are described, as are the main roles and work settings of counseling psychologists. In addition, basic areas of scientific inquiry are discussed.

Counseling psychology is a specialty field within the broader discipline of psychology. This specialty is related to, but also different from, several other psychology specialties that are often characterized by the term 'applied.' The psychology specialties to which counseling psychology is most closely related are clinical, community, school, and industrial/organizational psychology. Although each of these fields seeks to study and understand human behavior empirically, the fields also share the mission of applying scientific understandings to the solution of human problems and the betterment of peoples' lives. It is in this sense that they may be termed applied specialties.

As is the case with any field or specialty, counseling psychology originated from a broad matrix of social, economic, and psychological forces. Its beginnings may be traced to the early part of the twentieth century. However, the designation of counseling psychology as a specialty within the American Psychological Association (APA) first occurred in 1952. For historical analyses, the reader is referred to Gelso and Fretz (2000).

1. Counseling Psychology: Distinctiveness and Diversity

In this section, the features that make counseling psychology a unique specialty within psychology, as well as the breadth and diversity of this field, are described.

1.1 The Unifying Themes of Counseling Psychology

As a way of clarifying the features that make for its distinctiveness, Gelso and Fretz (2000) proposed five unifying themes of counseling psychology. These themes, which have received substantial endorsement in the field, are briefly described below. The themes are best viewed as central tendencies, for there is also great variability around each theme.

The first unifying theme is a focus on intact, as opposed to profoundly disturbed, personalities. In earlier days, counseling psychology was thought of as a field that focused only on 'normal' problems. Today, psychologists trained in this fieldwork with clients at all levels of disturbance. However, in comparison to other health service fields such as clinical psychology and psychiatry, counseling psychologists continue to work more with clients closer to the normal range, and to devote more time to issues of psychological health and wellness.

The second unifying theme is the focus on assets and strengths, regardless of clients' level of disturbance. Indeed, attention to strengths has been a hallmark of this specialty, and is most evident when counseling psychologists engage in preventive and developmental interventions, as opposed to remedial ones. Preventive and developmental interventions, rather than emphasizing emotional problems that have already emerged, seek to prevent problems and enhance growth. A few of the numerous examples of such interventions are enhancement treatments for couples who do well but want to improve further, career counseling for clients who want to understand fields that fit them best, and consultation with business organizations to help create healthy work environments.

The third unifying theme is emphasis on relatively brief interventions. Counseling psychologists engage in a wide range of interventions, for example, counseling, consultation, guidance. Also, many practitioners in this specialty conduct long-term psychotherapeutic treatments. However, a hallmark of the field has always been an emphasis on brevity. The use of the term 'counseling' is a reflection of this tendency, and currently popular brief therapies fit closely with such a counseling orientation. Generally, therapy lasting less than six months is seen as brief, although most counseling psychology interventions are of much briefer duration.

Emphasis on person–environment interactions is a fourth unifying theme. From its inception, counseling psychology has sought to balance the intrapsychic and the environmental, both in understanding development and in formulating interventions. Theories that have focused largely on intrapsychic explanations, such as psychoanalysis, have not had wide appeal. Current attention to the role of culture and multicultural issues is an example of counseling psychology's commitment to incorporating the role of situation and environment in understanding personality and development, and in providing effective interventions.

The final unifying theme, perhaps more than the others, reflects counseling psychology's distinctive-

ness. The field's emphasis on educational and career development and environments has often been seen as its most distinctive feature. Throughout its history, counseling psychologists have theorized about and empirically investigated issues around career choice, problems, interests, development, and interventions.

1.2 Fundamental Features of the Field

There are certain features of this field that, while not necessarily distinctive, are so fundamental as to require mention. First, counseling psychology is a doctoral level field in the USA in that the large majority of US psychologists who use the title counseling psychologist, and who join the Division of Counseling Psychology of APA, hold a doctoral degree. Second, since the specialty's beginnings, it has strongly endorsed the scientist-practitioner model of training, which dictates that doctoral programs seek to train students as scientific researchers as well as practitioners of counseling psychology.

A third fundamental feature of the field is its attention to issues of culture, multiculturalism, and diversity. Although all fields within psychology have begun to pay attention to diversity in terms of race/ethnicity, sexual orientation, gender, religious beliefs, and physical disability, counseling psychology has been at the forefront in these areas for many years. Attention to issues of diversity pertains to, for example, recruitment of doctoral students and faculty, empirical research and theory, and counseling practice. The fundamental emphasis on culture and diversity in the specialty is evidenced in chapters in the most recent edition of the *Handbook of Counseling Psychology* (Brown and Lent 2000) on the role of race and social class in development; gender, sexuality, and sexual orientation; gender issues in counseling; and models of multicultural counseling.

1.3 The Roles and Functions of Counseling Psychologists

Information on where counseling psychologists work and what they do is basic to understanding this specialty. Gelso and Fretz (2000) present data on job settings indicating that over half the counseling psychologists who were members of the Division of Counseling Psychology (DCP) of APA were employed in college and university settings, either in academic departments (35 percent) or counseling centers (17 percent). A major change in recent years has been the increased appeal of independent practice, with approximately 20 percent of DCP members claiming this as their primary job. Furthermore, since many counseling psychologists who are in independent practice do not join the DCP, the 20 percent figure is likely a significant underestimate.

The remainder of counseling psychologists are distributed across a wide range of job settings, including hospitals (e.g., general, psychiatric, Veterans Administration), clinics, government agencies, business and industry, and rehabilitation facilities. Surveys over several decades indicate that these job patterns have been generally quite consistent, except for the increase in independent practice since about the mid-1980s, and a decline in the numbers of counseling psychologists in Veterans Administration centers.

Regarding job activities, the most basic finding of all surveys is that counseling psychologists are involved in an extensive range of activities. A significant number devote at least some of their work time to counseling and psychotherapy (individual, group, couples, family); consultation; psychological testing, assessment, and evaluation; research and writing; teaching, training, and supervision; and administration (Gelso and Fretz 2000). At the same time, the activities performed by the greatest numbers for the greatest amount of time are counseling and psychotherapy.

2. Scientific Inquiry

The major areas of theory and research in counseling psychology reflect the above discussion. Below are described key scientific issues and topics in this field.

2.1 Methodological Considerations

A fundamental quality of science in counseling psychology since the late 1970s has been its emphasis on diversity of research methods and approaches (Heppner et al. 2000). As pointed to by most reviewers, beginning in the late 1970s, there has been an increasing press to incorporate research methods that go beyond the 'received view,' for example, quantitative-experimental research.

Within the context of this methodological pluralism, two notable patterns may be seen. First, stemming from postmodern thought that is evident in all scholarly disciplines, there has been a growing interest in the use of qualitative research approaches (Morrow and Smith 2000). Such approaches use verbal or even visual rather than quantitative data, gather data by observing and listening to what people do and say, and present results in the form of verbal descriptions and theories rather than statistics.

A second pattern actually moves in the opposite direction from qualitative research. That is, there has been a major increase in the use of highly sophisticated quantitative methods. Multivariate statistical methods now appear to be as common as univariate approaches. Especially prominent are techniques that seek to permit causal inferences from correlational data, for example, structural equation modeling.

2.2 Counseling and Therapy: Theory and Research

Regarding psychological interventions, counseling psychologists study essentially the same therapeutic process and outcome questions as do other applied counseling fields, for example, clinical psychology. After many years of outcome research, it is clear that, on the whole, therapy is effective; researchers now seek to discover what treatments are most effective when offered by what therapists for which clients experiencing what problems (Gelso and Fretz 2000).

Beyond this general statement, there are some therapy theories and topics that have been most influential in counseling psychology. Broadly speaking, cognitive-behavioral and experiential-humanistic theories of counseling have had the greatest impact (Heesacker 2000). As part of the strong cognitive-behavioral influence, theories such as Albert Ellis' rational-emotive behavior therapy (see *Behavior Psychotherapy: Rational and Emotive*), Aaron Beck's cognitive therapy (see *Cognitive Therapy*), and Albert Bandura's social cognitive theory (see *Social Cognitive Theory and Clinical Psychology*) have been particularly appealing to researchers and practitioners. Psychoanalytic theories have historically been less attractive to counseling psychologists, partly because these theories have conflicted with some of the major emphases of the field (e.g., strengths and assets, brevity, environmental influences). Fundamental paradigmatic changes in psychoanalysis in recent years, however, have made some of its theories (e.g., psychoanalytic self-psychology) more appealing (Gelso and Fassinger 1992). Social psychological conceptions have also had great appeal in recent decades. Theories such as social influence theory have been creatively extrapolated to counseling and therapy, and their propositions extensively investigated (Heesacker 2000).

2.3 Multicultural Counseling and Research

As noted, counseling psychologists have closely attended to diversity for many years. Since the early 1980s, however, there has been a mushrooming of theory and research on counseling and development as related to culture and diversity. The topic that has gathered the most scientific momentum is that of race and ethnicity in psychological development and treatment. This momentum is evidenced in the form of the high rate of articles on race, ethnicity, and culture in leading journals; major handbooks on the topic (e.g., Ponterotto et al. 1995); the creation of new journals; and the publication of texts and significant chapters. Although race and ethnicity are the most frequently studied topics under the umbrella of diversity, there are others. In recent years there has been a significant surge of research on gay, lesbian, and bisexual issues in counseling and development. The study of gender, which may be placed within the diversity category, has been central in counseling psychology for nearly three decades (Heppner et al. 2000).

2.4 Career Development and Intervention

Whereas the development of theory and research on diversity topics has increased at a dramatic pace in recent decades, the topic of career development and intervention has been a mainstay of counseling psychology since its inception. Many consider this the strongest and most empirically mature area within the specialty (Heppner et al. 2000). As part of this strength and maturity, robust theories of career development and choice guided theory and practice over many years (e.g., theories by Holland, Super, and Krumboltz). There exist a number of journals, handbooks, and texts aimed at studying and elaborating career development and intervention processes. In keeping with the diversity movement, recent decades have witnessed much attention to career development in the lives of marginalized people. The study of career development and intervention in womens' lives has been particularly prominent over many years.

2.5 Training in Practice and Science

The tandem topics of counselor training and researcher training are key areas of investigation. Of the two, counselor training and supervision has had the longer life. In particular, theory and research on the process and outcome of counselor supervision have been prominent since the late 1970s. Especially notable have been developmental models of counseling supervision, wherein therapists are theorized as developing in certain stages, and supervision approaches need to be matched to those stages (Stoltenberg et al. 1997). Regarding research training, theory and research have examined both person and research training environment factors that contribute to graduate students' and psychologists' interest in, sense of efficacy for, and production of research and science (Gelso and Lent 2000).

2.6 Healthy Functioning and Human Effectiveness

Research and theory in this general area are natural extensions of two key unifying themes described earlier: focus on intact personalities and attention to people's strengths. Although attention to positive qualities is a deep and widely shared part of the specialty, theory and research on the topic within counseling psychology had waned for a number of years, to the point that Gelso and Fassinger (1992)

viewed this area as counseling psychology's 'unfulfilled promise.' However, there has been an upsurge of scientific attention to well functioning. The study of positive constructs such as subjective well-being, career self-efficacy, and problem solving all fit this thrust. So, too, does the emerging examination by counseling psychologists of biopsychosocial factors that promote positive physical health (Hoffman and Driscoll 2000).

See also: Clinical Psychology in North America, History of; Counseling and Psychotherapy: Ethnic and Cultural Differences

Bibliography

Brown S D, Lent R L (eds.) 2000 *Handbook of Counseling Psychology*, 3rd edn. J. Wiley, New York

Gelso C J, Fassinger R E 1992 Personality, development, and counseling psychology: Depth, ambivalence, and actualization. *Journal of Counseling Psychology* **39**: 275–98

Gelso C J, Fretz B 2000 *Counseling Psychology*, 2nd edn. Harcourt College Publishers, Fort Worth, TX (1st edn. published 1992)

Gelso C J, Lent R W 2000 Scientific training and scholarly productivity: The person, the training environment, and their interaction. In: Brown S D, Lent R L (eds.) *Handbook of Counseling Psychology*, 3rd edn. Wiley, New York, pp. 109–39

Heesacker M 2000 *Counseling Psychology: Theories*. In: Kazdin A E (ed.) *Encyclopedia of Psychology*. American Psychological Association, Washington, DC

Heppner P, Casas M, Carter J, Stone G 2000 The maturation of counseling psychology: Multifaceted perspectives, 1978–1998. In: Brown S, Lent R (eds.) *Handbook of Counseling Psychology*, 3rd edn., Wiley, New York, pp. 3–49

Hoffman M A, Driscoll J 2000 Health promotion and disease prevention: A concentric biopsychosocial model of health status. In: Brown S D, Lent R L (eds.) *Handbook of Counseling Psychology*, 3rd edn., Wiley, New York, pp. 532–67

Morrow S, Smith M 2000 Qualitative research in counseling psychology. In: Brown S D, Lent R L (eds.) *Handbook of Counseling Psychology*, 3rd edn., Wiley, New York, pp. 199–232

Ponterotto J, Casas J, Suzuki L, Alexander C (eds.) 1995 *Handbook of Multicultural Counseling*. Sage, Thousand Oaks, CA

Stoltenberg C, McNeil B, Delworth U 1997 *IDM Supervision: An Integrated Development Model for Supervising Counselors and Therapists*. 1st edn. Jossey-Bass, San Francisco

C. J. Gelso

Counterculture

First, the word: 'counterculture' came into common usage largely as a result of the success of historian Theodore Roszak's bestseller, *The Making of a Counter Culture* (1969), a book about the movements of the 1960s. Earlier, J. Milton Yinger had used the term 'contraculture' in an article in *The American Sociological Review* (1960) concerned with distinctions among subcultures. But in a later comprehensive treatise, *Countercultures* (1982), Yinger yielded to the broad adoption of Roszak's usage.

Whether counter- or contra-, the prefix suggests opposition, and in the 1960s there were plenty of oppositional movements in Europe and the Americas against what was taken to be the dominant culture. That culture was conceived as a set of values—and practices which either (a) manifested those values (in which case the values were opposed as oppressive or repressive), or (b) contradicted them (in which case the practices were condemned as hypocrisy).

In either case, the oppositional movements, which ranged across the institutional landscape, emphasized 'liberation' from the constraints of dominant norms which were, more or less, enforced by authorities in each of the institutional domains. What had previously been regarded as 'deviance' from those conventional norms becomes countercultural when the deviants themselves begin to defend and celebrate their values and practices as wholesome, appropriate, desirable, or healthy, and seek to attract adherents.

Any ambiguity between counterculture as the word specifically designating the oppositional spirit of the 1960s, and its more general oppositional meaning, can be clarified by observing that the movements of the 1960s had predecessors to which they are historically linked. Similarly, one of the accomplishments of culture theory in recent decades has been to highlight the diversity of cultures (i.e., subcultures) contained within the notion of a national (or 'dominant') culture. Indeed, part of the dominance of a dominant culture is that *its* culture is taken to be *the* culture when, in fact, almost any group (e.g., ethnic, occupational, or age groups)—being sufficiently insulated from the dominant culture and/or interacting primarily or exclusively with its own members—may be expected to generate (or imperfectly sustain if already established) a set of values and practices—sometimes even an argot or lingo of its own—that defines its distinctive mode of life.

Seen in this context, countercultures are a subcategory of subculture, having a distinctively oppositional posture toward the dominant culture, unlike most other subcultures, which may differ from the dominant culture in several respects but are seldom essentially or overtly oppositional. The counterculture of the 1960s affirmed the Dionysian character of rock music and shared (among other things—an exhaustive listing is not practical) a taste for psychedelic drugs, consensual nonpossessive sex, emotional candor, colorful dress, ecstatic dance, racial equality, avant-garde esthetics, religious mysticism, and communal living. Along with their more politically militant comrades in the New Left, these counterculturalists

shared a distaste for the Vietnam war, bureaucracy, authority, impersonality, consumerism, careerism, corporate capitalism, some (though not all) technology, and 'suburban' blandness. Any such description has the status of what Max Weber called an 'ideal type': no specific counterculture group is likely to carry every one of these features—they characterize an 'ideal' counterculture.

The values and practices of this counterculture were of course regarded by many as outrageous. That the behavior was usually not actually illegal made it difficult to suppress, especially in societies that guaranteed civil liberties. Rather than violating statutes, some ostensibly common set of implicit rules and sense of propriety were being violated, propriety being one of the important dimensions of social life that culture defines. Countercultures are part of continuing effort to redefine that dimension.

Conceived as a kind of cultural radicalism, then, countercultures have a long tradition. Like all traditions, contemporary expressions of counterculture incorporate some (though not all) of its ancestral features and adds some new wrinkles of its own. Its ecstatic Dionysian character is continuous though any specific expression of it is inevitably affected by current conditions, for example the way in which certain forms of rock music, by appealing to its counterculture audience, is designed to offend others. The milieu of counterculture is evoked in Warren Beatty's film *Reds*, about the Russian revolution. In art it is evident in futurism, surrealism, dadaism, and abstract expressionism; it was there at the birth of cultural modernism in Vienna in the 1890s, and at the turn of the nineteenth century in Munich's then Bohemian district, Schwabing—whose denizens at that time regarded Paris as provincial. Some of it can be seen in the revolutions of 1848 and back beyond that to the Parisian Bohemia of the 1830s whose roots can be found in the cafes of Paris two generations earlier, before the French Revolution: that peculiar mix of populist and democratic, anarchic and communal sympathies combined with 'advanced' tastes in art and 'liberated' ideas about sex, drugs, and general lifestyles. Malcolm Cowley had evoked the tradition in his book *Exile's Return* (1934) about the experience of his 'lost generation' of friends in Europe in the 1920s. Yinger finds evidence of counterculture as early as the seventeenth century in the Ranters of that time.

Since the late nineteenth century, eruptions of counterculture have usually been attributed to 'generations'—usually to a 'younger' generation rebelling against some features of a status quo. While it is true that young people, frequently students, have been prominent partisans of counterculture movements, they have never been anywhere close to a majority of their age-cohort, but rather mostly well-born youth with more education than most of their contemporories. Still, like other group-elites they have often been successful in 'speaking for' their age group; majorities don't matter much in generational rhetoric. Moreover, in order to achieve any success in altering some of the values and practices of the dominant culture, counterculture youth require wider support from groups other than their age-mates. By themselves, 'generations' don't pass legislation, stop unpopular wars, decriminalize marijuana, destigmatize premarital sex, or even abolish the *in loco parentis* functions of college and university administrators.

To achieve such essentially political objectives, counterculture radicals must make common cause with political radicals, but the alliance is an uneasy one and often short-lived. The eventual disdain in the early nineteenth century between the Saint Simonians and Paris's literary Bohemians foreshadowed the mutual disdain between the 1960s counterculture and the New Left ('Heads' and 'Fists', respectively) after they broke apart around 1968. In the 1920s the cultural avant-garde of the Weimar Republic could find no political home; they were an embarrassment to the left-wing parties who feared alienating their culturally conservative working-class constituencies.

There are good structural reasons for the fragilities of the alliance between counterculturalists and political radicals. Counterculture objectives usually focus on matters like art, sex, manners, mores, and other lifestyle issues that involve mostly claims for private rights and liberties not usually requiring basic structural changes in the society. Political radicals, on the other hand, usually aim at achieving basic change in the structures of political power and the distribution of wealth. The different agendas (and therefore the constituencies) of these two versions of radical oppositional movements create tensions between them which are difficult to resolve. The natural constituencies of political radicals are the oppressed or disadvantaged classes with deep grievances and resentments against the privileged classes.

The constituencies of counterculture radicals are harder to designate: not the disadvantaged classes, not 'the people,' but small sectors of the high bourgeoisie, educated professionals, some students at elite universities, some sectors of contemporary culture industries (e.g., film, theater, advertising, fashion) seeking marketable innovation, and other hard-to-classify groups who patronize their efforts, partially adopt their style, ambivalently admire their 'wildness,' and otherwise aid and abet such swings in the fashions of culture.

Because most forms of counterculture are not major threats to the political and economic interests of ruling groups they are frequently (though far from always) tolerated by established powers. Bismarckian and Wilhelmine Germany tolerated a lot of youthful cultural radicalism. The same was true of Franz-Joseph's Austro-Hungarian Empire, and even, for a short time, of the Soviets. In the US, attacks on The National Endowment for the Arts and efforts to suppress the exhibition of 'indecent' or 'pornographic'

art in some museums at the end of the twentieth century succeeded only in dividing powerful elites. Countercultures, then, do not typically need political and economic transformations to achieve their moral and esthetic aims. What they do need are special audiences and publics, free spaces, tolerant ambiances and milieux which are either insulated against the larger political community, or functionally important to it—for example, in the way that colorful 'Latin Quarters' of cosmopolitan cities attract tourist business to them, are defended by active merchants' associations, and are therefore at least tolerated, if not wholly respectable, parts of their economies.

See also: Conflict: Anthropological Aspects; Conflict Sociology; Conformity: Sociological Aspects; Culture, Sociology of; Generations: Political; Generations, Relations Between; Generations, Sociology of; Groups, Sociology of; Subculture, Sociology of; Values, Sociology of

B. M. Berger

Counterfactual Reasoning: Public Policy Aspects

There is nothing new about counterfactual inference. Historians have been doing it for at least 2500 years. Thucydides invoked counterfactuals when he argued that had it not been for the far-sighted policies of Themistocles, the Persians would have defeated the Athenians at the battle of Salamis—an outcome that many, although not all, twentieth century observers believe might have nipped Western civilization in the bud (cf. Hanson 2002, Strauss 2002). Contemporary social scientists—from Max Weber ([1905] 1949) to Robert Fogel (1964)—have also long been aware of the pivotal role that counterfactuals play in scholarship on such diverse topics as the causes of economic growth and the diffusion of religious and philosophical ideas. Interestingly, though, some scholars have still sternly warned against 'what-might-have-been' questions. They claim that making sense of history is tough enough as it is—as it actually is—without worrying about how things might have worked out differently in this or that fictitious scenario (Carr 1961).

1. The Unavoidability of Counterfactual Reasoning

The ferocity of the critics is a bit unnerving. Moreover, they are certainly right that counterfactual inference is often dauntingly difficult. But they are wrong to say that one can avoid counterfactual reasoning at acceptable cost. And they are wrong that all counterfactuals are equally absurd because they are equally hypothetical (Fisher 1970). Analysts can avoid counterfactuals only if they eschew all causal inference from historical data and limit themselves to strictly noncausal narratives of what actually happened (no smuggling in causal claims under the guise of verbs such as 'influenced,' 'responded,' 'triggered,' 'precipitated,' and the like). Putting to the side the question of whether any coherent and compelling narrative can be noncausal, this prohibition would prevent observers from drawing exactly the sorts of lessons from history that scholars and policy makers regularly draw concerning such topical areas as the best ways to encourage economic growth, to preserve peace, and to cultivate democracy. Without counterfactual reasoning, how could observers know whether state intervention accelerated or retarded growth in a set of developing countries, whether deterrence policies prevented or provoked aggression, or whether the loan policies of international institutions such as the World Bank reduced the risk of financial contagion or increased the risk of moral hazard? In brief, counterfactual reasoning is a prerequisite to any form of learning from history. To paraphrase Robert Fogel's (1964) reply to the critics of counterfactual history in the 1960s, everyone does it and the alternative to an open counterfactual model is a concealed one.

This article surveys the diverse roles that counterfactual arguments play in policy analysis and in political debate. A useful place to begin is by clarifying what is meant by counterfactual reasoning. A reasonably precise philosophical definition is that counterfactuals are subjunctive conditionals in which the antecedent is known or presumed for purposes of argument to be false (Goodman 1983). As such, an enormous array of politically consequential arguments qualify as counterfactual. Consider the following rather representative sample of counterfactuals that have loomed large in recent policy controversies:

If Boris Yeltsin had followed the fiscal and monetary advice of the advocates of shock therapy policies in early 1992, Russian inflation in 1993 would have been a small fraction of what it was;

If the USA had not dropped atomic bombs on two Japanese cities in August 1945, the Japanese would still have surrendered roughly when they did;

If it were not for the affirmative action programs that have existed in the USA since the 1970s, the black middle class would have grown much more slowly than it did;

If Reagan had not adopted such a hard-line posture toward the USSR in the early 1980s, it is very unlikely that the USSR would have been as conciliatory as it was in the late 1980s;

If the liberal social welfare policies of the 1960s had not implicitly encouraged single parenthood, the percentage of out-of-wedlock births in the USA in the 1980s would have been much lower than it was.

2. Distinguishing Sound from Unsound Counterfactuals

What standards of evidence and proof might be used for judging the quality of the counterfactual arguments that underlie so many causal claims in policy debates? In a comprehensive review of the methodological and philosophical literature bearing on counterfactual inference, Tetlock and Belkin (1996) identify six classes of criteria for making such judgments. These are outlined below.

2.1 Clarity

Good counterfactual arguments should specify and circumscribe the independent and dependent variables (the hypothesized antecedent and consequent). This initial recommendation might strike readers as obvious. Like actual experiments, thought experiments should manipulate one cause at a time, thereby isolating pathways of influence. Although excellent advice in principle, implementing it is often deeply problematic. As Jervis (1997) documents, there is no way to hold 'all other things equal' when we perform thought experiments on social systems that are densely interconnected. To invoke the terminology of experimental design, the 'independent' variable in interconnected systems cannot be manipulated without creating ripple effects that alter the values taken on by other potential causes in the historical matrix, thereby creating 'confounding' variables that render the interpretation of the original thought experiment problematic. For example, if Boris Yeltsin had lived in a Russia in which he felt it was feasible to pursue shock therapy policies in 1992, there is a good case to be made that this hypothetical Russia differs from the actual Russia in a host of other ways (less corruption and reduced influence of special interests) that might also have been more conducive to better economic performance.

2.2 Logical Consistency or Cotenability

Good counterfactual arguments should specify the connecting principles that link the antecedent with the consequent with sufficient precision that it is possible to gauge whether those connecting principles are cotenable with the original antecedent. Every counterfactual claim represents a condensed or incomplete argument that requires connecting principles that sustain, but do not imply, the conditional assertion. When explicitly articulated, these connecting principles are often complex, even in the case of such deceptively simple counterfactuals as 'if the match had been struck, it would have lighted.' The connecting principles specify, within reasonable limits, everything else that would have to be true to sustain the counterfactual, including the necessary amount of friction generated by the contact of the match against the surface, the chemical composition of the match, the absence of water, and the presence of oxygen.

These usually invisible connecting principles are sometimes in direct tension with the historical antecedent that has been altered in the counterfactual. One example is Jon Elster's (1978) critique of Robert Fogel's (1964) counterfactual assertion that if the railroads had not existed, the American economy in the nineteenth century would have grown only slightly more slowly than it did. Elster did not show that Fogel was wrong but he did show that it is nonsensical to postulate as a supportive connecting principle the proposition that the internal combustion engine would have been invented earlier in a USA without railroads because that proposition presupposes a theory of technological innovation that undercuts the original antecedent. If we have a theory of innovation that permits the invention of cars 50 years earlier, why does it also permit blinding people to the potential of railroads? In a similar vein, some critics have taken to task John Mueller (1988)for his counterfactual claim that, even in the absence of nuclear weapons, the Cuban missile crisis would not have escalated into war. This is an even more straightforward violation of cotenability of antecedent and connecting principles. Why would the USSR go to all the trouble of placing conventionally armed intermediate-range missiles in Cuba? Why take so large a risk for so small an advantage?

Taken together, the previous two standards—the logical clarity and cotenability standards—are helpful for screening out ambiguous and oxymoronic counterfactuals. But they define a bare minimum for quality-control purposes. Additional, historical and theoretical standards are needed for winnowing out other forms of 'what-if' foolishness.

2.3 The Minimal Rewrite Rule

Good counterfactual arguments should specify antecedents that could possibly have been implemented given the political and economic constraints at work in the situation. As shall soon be evident, there are occasions when scholars and policy analysts want to override this rule. However, historians do frequently apply it and many policy makers sympathize. What is the point of wondering about the effects of a policy that never had the remotest chance of being implemented? There is indeed little point in doing so if one is interested solely in judging whether policy makers selected the best of the politically feasible options at that historical moment. But there is good reason for doing so if the goal is to delineate the conditions under which particular policy interventions are likely to have particular effects. For instance, political leaders in industrialized countries during the Great Depression

may have correctly surmised that they had little choice but to resort to protectionist policies. To suppose that they could have acted otherwise may be historically naive. But from the standpoint of drawing useful policy lessons for the future, it is important to assess to what degree the economic downturn of the 1930s was exacerbated by the proliferation of beggar-thy-neighbor tariffs that thwarted application of the logic of comparative advantage.

2.4 Theoretical Consistency

Compelling counterfactuals should rest on connecting principles that are consistent with well-established theoretical generalizations relevant to the hypothesized antecedent–consequent linkage. Otherwise, analysts cannot rule out counterfactuals that start from reasonable antecedents but end in far-fetched consequences by invoking preposterous principles of causality such as: 'If Oswald had not shot Kennedy, then someone else would have done so, because Kennedy was astrologically fated to die by assassination' or 'If North Korea had conquered South Korea in 1950, the economy of the South would have grown even more rapidly than it actually did because of the wisdom of the policy of self-sufficiency of the Great Leader Kim Il Sung.'

Ideally, all counterfactual inferences could be grounded in extensively validated scientific laws of the sort drawn upon in the match-lighting conditional. Occasionally, this is possible: Witness physicist Richard Feynmann's compelling thought experiment (with the assistance of a glass of ice water) that demonstrated the key role that cold-sensitive O-rings played in the explosion of the space shuttle Challenger in 1986 (Vaughan 1996). More often, however, policy analysts must rely on more fragile—and controversial—connecting principles drawn from the behavioral and social sciences.

2.5 Statistical Consistency

Compelling counterfactuals should rest on connecting principles that are consistent with well-established statistical generalizations bearing on the antecedent–consequent linkage. In many contexts, we rely not on theoretical laws but on statistical generalizations to fill in 'what would have happened if this rather than that event had occurred.' One obvious form such reasoning takes is reliance on base rates and patterns of co-variation. For instance, an observer might justify the counterfactual, 'If Bill Clinton had lost the presidential election of 1992, he would have been disappointed,' by observing in a two-by-two contingency table that when people fail to achieve a goal for which they have worked long and hard, the overwhelming majority experience disappointment, but when people do achieve their goals, there is markedly less disappointment. This discovery is hardly startling; statistical counterfactual arguments can, however, be made more controversial. Econometric modeling techniques can be used to reconstruct how the US economy might have performed if the invention of railroads had been delayed or whether slavery might have disappeared as a result of the operation of market forces (no Civil War necessary?). Or regression and time-series techniques might be used to infer how extensive out-of-wedlock births would have been in late twentieth century USA if welfare benefits had been lower or higher across states since the 1970s (inferences that rest on controversial but at least well-defined analytical assumptions).

2.6 Projectability

Good counterfactual arguments should lead to testable implications that help to make sense of antecedent–consequent links in other contexts. Theory evaluation and counterfactual evaluation are inextricably entangled. Sound counterfactuals require sound theories that provide the law-like generalizations that fill in the missing data points in our thought experiments. How can social scientists judge, however, whether these law-like generalizations are robust enough to support counterfactual inferences? Here Nelson Goodman's (1983) concept of projectability is helpful. Goodman draws a sharp distinction between coincidental generalizations that just happen to be true at a particular time and place (and are therefore unprojectable) and robustly law-like generalizations that hold up over a range of circumstances and permit projection into the past and future. An example of a merely coincidental generalization is 'All the coins in my pocket yesterday were silver.' Nothing follows from this observation—certainly not 'If this penny were in my pocket yesterday, it would be silver.' The counterfactual fails because 'if this penny were in my pocket yesterday,' the reader would simply assume that the original generalization—'all the coins in my pocket yesterday were silver' was false. By contrast, a robustly law-like generalization—such as that oxygen is a necessary but not sufficient condition for fire—inspires confidence when the reader moves either backward in time (if there had been no oxygen, the Great Fire of London would not have occurred) or forward in time (if we cut off any future fire's source of oxygen, the fire will expire).

Many social scientists view projectability as the acid test of scientific legitimacy. The same causal principles that allow observers to retrodict the past should make it possible to predict the future. Indeed, the strong Popperian form of this argument asserts that observers should take counterfactual claims seriously if and only if the law-like generalizations supporting the claims

yield falsifiable forecasts. This classic philosophical argument has resurfaced in the recent methodological advice of King et al. (1994), who urge scholars to search aggressively for the observable implications of their causal constructs by regularly asking themselves, 'If my argument is correct, what else should be true?' Counterfactuals that are devoid of testable implications in the actual world leave historical observers marooned in hypothetical worlds of their own subjective making. Projectability, from this vantage point, stands as the pre-eminent criterion for judging the value of counterfactual speculation.

3. *'Consciousness Raising'*

There are special classes of counterfactual arguments that thoughtful people may not want to subject to the conventional historical and social science tests—counterfactuals that are designed primarily to open otherwise closed minds. Tetlock and Belkin (1996) dub such arguments 'gadfly counterfactuals.' Here it is, however, worth noting the intense skepticism of many psychologists. They doubt the capacity of the human mind to escape its preconceptions when it undertakes mental simulations of possible worlds. It is just too easy for political partisans to conjure up answers to 'what-if' questions that reinforce their prejudices: what could be more natural than for liberals to fill in the missing counterfactual control conditions of history with ideologically consonant scenarios ('of course, subtract out the Reagan presidency and replace it with a two-term Carter presidency and a Mondale follow-up, and the USSR disintegrates pretty much in the same way it did') and for conservatives to reach the opposite counterfactual conclusion. Indeed, such beliefs are often held with such conviction that they feel factual, not counterfactual.

But there are also psychological reasons for suspecting that people are not always fated to be prisoners of their preconceptions (Tetlock 1998, 1999). Here it is worth dwelling on some intriguing parallels between the mental process of simulating what would have happened in possible worlds and that of constructing computer simulations of natural or social phenomena (Cederman 1997). Both the computer software and the mental wetware exercises require making explicit previously implicit and ill-thought-through assumptions. In the course of running these computer or mental simulations, one can in principle learn a great deal about hitherto concealed gaps and contradictions in one's knowledge. As Kahneman (1996) points out, mental simulations derive their persuasive force and power to surprise people by revealing unnoticed tensions between explicit, conscious beliefs and implicit, unconscious ones. In this sense, people discover aspects of their political belief systems in mental simulations that would otherwise have gone undiscovered. It is useful to distinguish three specific ways in which mental simulations can yield insights into one's own thought processes: by revealing double standards in moral judgment, contradictory causal beliefs, and the influence of unwanted biases such as certainty of hindsight.

3.1 *Counterfactual Morality Tales*

Mental simulations can compel people to acknowledge embarrassing inconsistencies in their application of moral rules. The paradigmatic example is the identity-substitution thought experiment that manipulates either the perpetrator or victim of a policy act and asks the audience to contemplate whether they had the same emotional reaction to what actually happened as they would have to various hypothetical events. For example, 'if Bosnians were bottle-nosed dolphins (or if Rwandans were white or if Chechnyans were Lithuanians …), the public never would have tolerated the slaughter of innocence so long.' Insofar as the audience detects a discrepancy in their reactions to the scenarios, and insofar as the audience firmly believes that the mentally manipulated cause should be irrelevant, the audience will deem the discovery of a differential emotional reaction to be a disturbing fact about themselves, that is, to be cognitively dissonant. Of course, not all political leaders fold when they are accused of moral inconsistency: The famously anti-Bolshevik Winston Churchill responded to those who probed how uncomfortable he felt about being aligned with Stalin in World War II by invoking a counterfactual argument of his own to the effect that if Hitler were to invade Hell, Churchill would find a way of making a favorable reference to the Devil in the House of Commons.

3.2 *Counterfactual Consistency Probes*

These mental simulations reveal contradictions between causal beliefs that may have previously coexisted within a belief system. The paradigmatic example is the syllogism that traces the logical implications of one set of beliefs to the point where the contradiction becomes undeniable. For example, 'if you really believe there is that much indeterminacy at the microlevel (factors that determine the outcome of battles, the success of firms, the policies of bureaucratic units of government), even invocations of the 'miracle of aggregation' do not allow you to argue for so strongly deterministic a position at the macrolevel (the outcomes of wars, the performance of economies, the political stands of governments).' Or, 'if you adopt an extreme structural-realist position that denies any significant role to domestic politics in shaping properties of the international system, then you would have to belief that even if the USSR had been a democracy in 1945, the Cold War would still have occurred.' In these cases, the thinker has the dissonance-reduction

options of changing one or both sets of beliefs, introducing new cognitions that neutralize the contradiction, or disengaging from the simulation exercise by simply ignoring the contradiction.

3.3 Counterfactual Exercises as Means of Stimulating the Imagination

Retrospective scenario generation is a form of mental stimulation in which the goal is to prevent premature cognitive closure induced by certainty of hindsight. Some scholars (e.g., Weber [1905] 1949) have long suspected, and cognitive psychologists have now demonstrated (e.g., Fischhoff 1975), that 'outcome knowledge' contaminates our understanding of the past. Once people learn the outcome of an event, they not only perceive that outcome as more likely *ex post* than they did *ex ante* (which might be defended as a rational Bayesian updating of subjective probabilities), they systematically misrecall their *ex ante* assessments as more consistent with that actual outcome than those assessments really were (the 'I-knew-it-all-along' effect).

Counterfactual thought exercises can check the 'creeping determinism' of certainty of hindsight. Asking people to think of how things could have worked out differently becomes a means of preventing the world that did occur from blocking our views of the worlds that might well have occurred if some antecedent condition had taken on a different value. Indeed, there is a 'debiasing' literature in experimental psychology that assesses the usefulness of encouraging people to imagine that the opposite outcome occurred. There is also a literature in literary studies on the concepts of sideshadowing, foreshadowing, and backshadowing in narratives that makes a strikingly similar normative point (for reviews, see Tetlock and Belkin 1996). Sideshadowing calls attention to what could have happened, thereby locating what did happen in the context of a range of possibilities that might, with equal or even greater likelihood, have taken place instead. Sideshadowing serves as a valuable check on foreshadowing (the tendency, in extreme form, to reduce all past events to harbingers of the future) and backshadowing (the even more insidious tendency to judge historical actors as though they too should have known what was to come).

4. Closing Remarks

The topic of counterfactual thought experiments in public policy leaves many social scientists feeling a bit uneasy, even defensive. To be blunt, it feels like epistemological slumming. There is, in the behavioral and social sciences, a prestige hierarchy of methods for drawing causal inferences. At the top of the pecking order is experimentation in which investigators can manipulate hypothesized causes and then either hold everything else constant (*ceteris paribus*) or randomize extraneous influences across treatment conditions. Experimental control of this sort is occasionally possible in the domain of public policy: There have been some well-known experimental attempts to assess the impact of various drug treatment programs and social welfare programs. By and large, however, experimentation is hopelessly impractical. We cannot rerun the tape of history to assess the causal impact of Gorbachev's policies of glasnost and perestroika or to assess the impact of Reagan's defense build-up or tax cuts.

Social scientists often resort to statistical control when experimentation is ethically or practically problematic. But statistical arguments invariably rest on counterfactual assumptions and are extraordinarily difficult to justify in many policy contexts. For example, what kind of regression or time series analysis will allow us to estimate the causal contribution of nuclear weapons to the 'long peace' between the USA and USSR between 1945 and 1991? There are too many confounding variables—a problem that can be alleviated but not eliminated through judicious selection of comparison cases and meticulous examination of decision-making protocols (Fearon 1991).

So where does this leave the field? Probably still feeling uneasy: Social scientists and policy analysts seem to be stuck with quite literally a third-rate method, i.e., counterfactual thought experimentation. The control groups exist—if 'exist' is the right word—only in the imaginations of the policy analysts who are left with the daunting task of trying to reconstruct how history would have unfolded if causal variables from the past had taken on different values from the ones that they actually did. The whole exercise starts to look hopelessly subjective, circular, and nonfalsifiable. What is to stop ideologues from simply inventing counterfactual outcomes that serve to justify their political biases and predilections? There appear to be large classes of policy questions for which experimental control is out of the question and statistical control is of limited usefulness (assuming that one can find suitable comparison cases and can reliably operationalize the key conceptual constructs). These questions are too important to ignore but apparently too difficult to answer in a fashion that transcends traditional ideological divisions.

Too often, the response to the dilemma is to embrace extreme solutions (Strassfeld 1992): either to reject all counterfactual arguments as fanciful suppositions, mere conjecture, and frivolous figments of policy analysts' imaginations or to assume confidently that we know exactly what would have happened if we'd gone down an alternative path, sometimes going so far as to project several steps deep into hypothetical causal sequences. The former response leads to futile efforts to exorcize all counterfactuals from historical inquiry; the latter response leads at best to error (we

ignore the compounding of subjective probabilities at our peril) and at worst to the full-scale politicization of counterfactual argumentation (as advocates claim *carte blanche* to write hypothetical histories that advance their ideological agenda).

This article has articulated a principled compromise between these extremes. On the one hand, it is necessary to acknowledge that counterfactual thought experiments inevitably play key roles in the causal arguments that animate many policy controversies. On the other hand, it is necessary to acknowledge that counterfactual thought experiments are often suffused with error and bias. But, that said, it is not appropriate to conclude that things are hopeless—that it is impossible to draw reasonable causal inferences from historical data when experimental and statistical control are not available. Rather, the preferred conclusion of this article is that disciplined use of counterfactuals—grounded in explicit standards of evidence and proof—can be enlightening in specific policy settings.

See also: Causal Counterfactuals in Social Science Research; Causal Inference and Statistical Fallacies; Causation (Theories and Models): Conceptions in the Social Sciences; Explanation, Cognitive Psychology of; Explanation: Conceptions in the Social Sciences

Bibliography

Carr E H 1961 *What is History?* Knopf, New York

Cederman L E 1997 *Emergent Actors in World Politics: How States and Nations Develop and Dissolve*. Princeton University Press, Princeton, NJ

Elster J 1978 *Logic and Society: Contradictions and Possible Worlds*. Wiley, New York

Fearon J D 1991 Counterfactuals and hypothesis testing in political science. *World Politics* **43**: 169–95

Fischhoff B 1975 Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance* **1**: 288–99

Fisher D H 1970 *Historians' Fallacies: Toward a Logic of Historical Thought*. Harper and Row, New York

Fogel R 1964 *Railroads and American Economic Growth: Essays in Econometric History*. Johns Hopkins University Press, Baltimore, MD

Goodman N 1983 *Fact, Fiction, and Forecast*. Harvard University Press, Cambridge, MA

Hanson V 2002 One man saves the West. In: Tetlock P E, Lebow R N, Parker G (eds.) *Unmaking the West: Exploring Alternative Histories and Counterfactual Worlds*. Columbia University Press, New York

Jervis R 1997 *Systems: Dynamics and Effects*. Princeton University Press, Princeton, NJ

Kahneman D 1996 Varieties of counterfactual thinking. In: Roese N J, Olson J M (eds.) *What Might Have Been: The Social Psychology of Counterfactual Thinking*. Erlbaum, Mahwah, NJ, pp. 408–24

King G, Keohane R O, Verba S 1994 *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press, Princeton, NJ

Mueller J 1988 The essential irrelevance of nuclear weapons: Stability in post-war world. *International Security* **13**: 55–79

Strassfeld R N 1992 If … Counterfactuals in the law. *George Washington Law Review* **60**: 339–416

Strauss B 2002 The resilient West: Salamis without Themistocles. Classical Greece without Salamis, and the West without Classical Greece. In: Tetlock P E, Lebow R N, Parker G (eds.) *Unmaking the West: Exploring Alternative Histories and Counterfactual Worlds*. Columbia University Press, New York

Tetlock P E 1998 Close-call counterfactuals and belief system defenses: I was not almost wrong but I was almost right. *Journal of Personality and Social Psychology* **75**: 639–52

Tetlock P E 1999 Theory-driven reasoning about possible pasts and probable futures: Are we prisoners of our preconceptions? *American Journal of Political Science* **43**: 335–66

Tetlock P E, Belkin A 1996 *Counterfactual Thought Experiments in World Politics*. Princeton University Press, Princeton, NJ

Vaughan D 1996 *The Challenger Launch Decision*. University of Chicago Press, Chicago

Weber M [1905] 1949 Objective possibility and adequate causation in historical explanation. In: *The Methodology of the Social Sciences*. Free Press, Glencoe, IL, pp. 163–88

P. E. Tetlock

Counterfactual Reasoning, Qualitative: Philosophical Aspects

1. Supposition

To reason about an hypothesis counterfactually is to reason about what *would be true*, on the supposition that the hypothesis *were true*. Such reasoning is naturally expressed in the form of conditional sentences couched in the subjunctive mood, for example:

If this glass had been struck, then it would have shattered.

The implication is that, in fact, the glass was not struck and did not shatter; hence the term *counterfactual*, or contrary-to-fact, conditional. However the same kind of reasoning applies in situations where the truth–value of the antecedent is an open question, as for example in the open subjunctive conditional:

If this glass were struck, then it would shatter.

This suggests that the terms *subjunctive conditional* and *suppositional reasoning* might more appropriately delineate the topic. But since the term 'counterfactual' is well entrenched, it will suffice to stipulate that one may reason 'counterfactually' about hypotheses that turn out to be true.

The philosopher Frank Ramsey once remarked: 'If two people are arguing "If p will q" and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q' (Ramsey 1990). Ramsey seems to have had open conditionals in mind, but the comment has been taken

by some to suggest a general pragmatic account of conditionals, that has become known as the Ramsey test: a subjunctive conditional is accepted in a belief state if and only if the consequent of the conditional is accepted when the belief state is adjusted so as to accommodate the hypothesis that the antecedent be true.

'Accommodating' the hypothesis must be more than just a matter of adding it to one's stock of beliefs, since to extend Ramsey's idea beyond the open case is to admit that the hypothesis may be inconsistent with one's prior beliefs. The 'adjustment' should, at the very least, leave the belief state consistent. But the adjustment might also reasonably allow for what Isaac Levi (1996) calls *inductive expansion*, the addition of further beliefs that are not entailed by the hypothesis.

This might suggest that the problem of counterfactual reasoning is one and the same as the problem of rational revision of belief. But there is a catch. Rational revision of belief is generally held to be subject to the following condition:

> Preservation: If the evidence to be accommodated is consistent with the prior belief state, then any proposition accepted in the prior belief state will still be accepted in the revised belief state.

Example: if you come to believe that Phar Lap won the 1930 Melbourne Cup, and this piece of information is consistent with everything you previously believed, then you should not respond by rejecting any of your old beliefs. Information is valuable and should not be discarded unnecessarily.

The catch is this: the Preservation condition and the Ramsey test are inconsistent with one another on pain of triviality. This result is known as Gärdenfors' theorem (Gärdenfors 1988, Sect. 7.4). It is closely related to the Lewis trivialization theorem described in the companion article.

The moral of Gärdenfors' theorem is that counterfactual reasoning has a logic of its own, distinct from the theory of the rational revision of belief. Supposition is not the same thing as belief revision. Entertaining the hypothesis A counterfactually is a matter of determining what one believes would be true, were A true. This must be distinguished from the different process of determining what one would believe to be true if one came to believe that A is true. Consider, for example, the difference between reasoning about what would have happened if Oswald *had not* killed Kennedy, and reasoning about what did actually happen if (surprise!) Oswald *did not* kill Kennedy.

2. *Imaging*

Revision is a matter of moving from a prior belief state to a revised belief state so as to accommodate input in the form of propositional evidence. Widely accepted criteria for the rational revision of belief have been developed by Alchourrón, Gärdenfors, and Makinson (Gärdenfors 1988). These criteria are usually referred to in the literature as the *AGM axioms*. Among them is the Preservation rule, along with the following natural conditions:

(a) The proposition to be accommodated must be accepted in the adjusted belief state.

(b) The adjusted belief state should be consistent if both the prior belief state and the proposition to be accommodated are consistent.

Gärdenfors' theorem tells us that supposition cannot satisfy all of the AGM axioms, since it is a method of belief adjustment that fails to satisfy Preservation. So what can be said about the logic of supposition?

Agents typically agree on some matters of belief and disagree on others. In general, given two agents, we may talk about the beliefs that they *share*. A shared belief is simply a proposition accepted by both agents. Those shared beliefs may themselves be thought of as forming a potential belief state; the strongest belief state compatible with what both agents believe. Now if you and I both adjust our beliefs in order to accommodate the same proposition there will be an adjustment to the beliefs that we share. How might the adjustment to our shared beliefs be determined? One simple answer to that question would be provided by the following principle:

> Composition: the adjustment to what is shared between belief states is what is shared by the adjustments to those belief states.

Call a method of belief adjustment that satisfies (a), (b), and the Composition principle an *imaging* method. The central result of the logic of supposition may now be stated: in the presence of conditions (a) and (b), the principle of Composition is equivalent to the Ramsey test condition. So, by the Gärdenfors theorem, AGM revision cannot be compositional. But supposition will be. Supposition—that is to say, counterfactual reasoning—is always a matter of imaging.

Example: it is a sunny day and Alice and Bob are driving to Versailles. They have been there twice before. They share the belief that it was raining on exactly one of those occasions. Alice believes it was raining on their first visit, but Bob believes it was raining the second time. Alice says: 'If it *had* rained the second time we were there, then since it rained the first time too, it would have been raining on both occasions.' Bob does not have to adjust his beliefs to entertain the hypothesis that Alice considers counterfactual; he thinks it is true. Cher is sitting in the back seat. She does not know whether Alice or Bob is correct, but she is tired of their argument. Cher thinks to herself, 'If Bob is right then if it had rained the first time they were there, it would have rained both times, but if Alice is right, then it would have still have been fine the second time.' Cher concludes, 'So, if it *had* rained the first time they were there it might have

rained the second time and it might have been fine the second time.' But Cher also thinks, 'If it *did* rain the first time they were there, then Alice is right and it was fine on their second visit.' Cher's belief state consists of the beliefs shared by Alice and Bob. When Cher supposes that it rained the first time, the suppositional state is obtained by counterfactual adjustment of the shared beliefs, which by Composition is the state that consists of those beliefs common to the states obtained after Alice and Bob make that supposition separately. This suppositional state is distinct from the revised belief state that Cher would be in if she were to learn that it is Alice who is remembering events correctly.

This section concludes with a demonstration of the close connection between counterfactual supposition as imaging and the Stalnaker–Lewis semantics for the subjunctive conditional described in the companion article.

Consider the subjunctive conditional: 'If this glass were struck, then it would shatter.' By the Ramsey test, I accept this conditional when my supposing that the glass be struck leads to my concluding that the glass would shatter. Since supposition goes by imaging, this means that the image of my belief state on the hypothesis that the glass be struck entails that the glass shatter. But imaging is compositional, so taking any more detailed account of the way things might be that is consistent with what I believe, and forming the image of that on the hypothesis that the glass be struck, must also result in a state that entails that the glass shatters. This remains true for the most detailed accounts of all, which might as well be called 'possible worlds.' To accept the conditional is to believe that the actual world is such that its image on the hypothesis that the glass be struck entails that the glass shatters. Call those worlds compatible with the image of the actual world on the antecedent the *closest* worlds at which the antecedent is true. Then to accept the conditional is to believe that all the closest worlds at which its antecedent is true are worlds at which its consequent is true. The conditional whose acceptance conditions are given by the imaging Ramsey test is revealed as none other than the Stalnaker–Lewis conditional.

3. *Subjunctivity*

The role of subjunctivity in nonconditional contexts is illuminated from a rather different angle by attempts to correct the expressive limitations of conventional formal languages. Crossley and Humberstone (1977) observe that one sense of the three-way ambiguous sentence:

It is possible for every red thing to be shiny

cannot be rendered in the standard language of modal logic. The possibility operator may be positioned inside or outside the scope of the universal quantifier, but neither formulation captures that third sense whose expression seems to require recourse to the logic of *actually*:

It is possible for everything *that is actually red* to be shiny.

If the modal operator 'possibly' functions like an existential quantifier over possible worlds, allowing us to roam freely over the domain in search of a world at which the embedded sentence is true, then 'actually' functions to *switch off* that operator temporarily. In the example just given, it forces us to return to our own world when it comes to determining what the relevant red things are. Humberstone (1982) remarks that 'actually' may be viewed as an *inhibitor*: a sentential operator that shields whatever is within its scope from the influence of other external operators.

Thus 'actually' has a function that is the exact opposite of the kind of semantic role Castañeda (1967) ascribes to subjunctivity. As Humberstone puts it, expressive inadequacies like the above are overcome in languages in which the subjunctive mood functions to *switch on* or to *activate* an otherwise dormant external operator. So, to adapt an example of Quine (1956), the ambiguity of the English sentence:

I am looking for a dog that talks

is resolved in Spanish via the indicative/subjunctive distinction. 'Busco un perro que habla' ('*habla*' indicative) means that there is a particular dog I am looking for that just happens to talk. 'Busco un perro que hable' ('*hable*' subjunctive) suggests that I do not have any particular dog in mind, any talking dog will do. The subjunctive element activates the 'looking for' operator at that point in the sentence, signaling that the dog's talking is to be understood as part of the content of my seeking. In Spanish (at least) desire contexts are among those naturally activated by the subjunctive.

4. *Desire*

DAB: To desire something is to believe it desirable.

An argument similar to the proof of the Gärdenfors theorem establishes that DAB also conflicts with Preservation (Collins 1988). But DAB is the simplest form that a cognitivist theory of motivation might take. So if cognitivism is not to be ruled out on purely logical grounds, 'desirability' must be interpreted suppositionally. Thus: a possibility is believed desirable just in case the supposition that the possibility be true leads to the conclusion that ones desires would be satisfied. Similarly, the expected desirability of an option is a belief-weighted average of the desirabilities of more specific ways the option might be realized. A *suppositional decision theory* obtains the expected desirability of an option by imaging on the supposition that the option be chosen. But this is none other than the *causal decision theory* (described in 'Newcomb's problem'). There is an argument for causal decision

theory that bypasses the usual discussions of Newcomb's problem.

See also: Causal Counterfactuals in Social Science Research; Counterfactual Reasoning: Public Policy Aspects; Counterfactual Reasoning, Quantitative: Philosophical Aspects; Nonstandard Reasoning; Piaget, Jean (1896–1980); Tversky, Amos (1937–96)

Bibliography

Castañeda H-N 1967 Actions, imperatives, and obligations. *Philosophical Studies* **19**: 25–48

Collins J 1988 Belief, desire, and revision. *Mind* **97**: 333–42

Crossley J N, Humberstone I L 1977 The logic of 'actually.' *Reports on Mathematical Logic* **8**: 11–29

Gärdenfors P 1988 *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, MA

Humberstone I L 1982 Scope and subjunctivity. *Philosophia* **12**: 99–126

Levi I 1996 *For the Sake of the Argument: Ramsey Test Conditionals, Inductive Inference, and Nonmonotonic Reasoning*. Cambridge University Press, Cambridge, UK

Quine W V O 1956 Quantifiers and propositional attitudes. *Journal of Philosophy* **53**: 177–87

Ramsey F P 1990 General propositions and causality. In: Mellor D H (ed.) *Philosophical Papers*. Cambridge University Press, Cambridge, UK, pp. 145–63

J. Collins

Counterfactual Reasoning, Quantitative: Philosophical Aspects

This article should be read in conjunction with its companion (see *Counterfactual Reasoning, Qualitative: Philosophical Aspects*). Here, after a general introduction and historical overview, we emphasize the role of counterfactual reasoning within the quantitative frameworks of probability theory, decision theory, and game theory.

1. Counterfactuals

Counterfactuals are a species of conditionals. They are propositions or sentences, expressed by or equivalent to subjunctive conditionals of the form 'if it were the case that A, then it would be the case that B,' or 'if it had been the case that A, then it would have been the case that B'; A is called the *antecedent*, and B the *consequent*. Counterfactual reasoning typically involves the entertaining of hypothetical states of affairs: the antecedent is believed or presumed to be false, or contrary-to-the-fact, but its truth is imagined or supposed. Counterfactual reasoning is thus a form of *modal* reasoning, kindred to reasoning about necessity or possibility, and in constrast to reasoning about the way things actually are.

The philosophical study of conditionals goes back at least as far as the Stoics of ancient Greece, although their systems of logic apparently did not accord the counterfactual any emphasis. The rise in interest in counterfactuals has been a rather recent phenomenon, as it started to become clear to philosophers that counterfactuals are implicated in a host of other important concepts—laws of nature, confirmation, causation, scientific explanation, knowledge, perception, dispositions, free action, etc. The significance of counterfactuals has also become increasingly appreciated in the social sciences—by psychologists (e.g., in understanding emotions such as regret), by historians (e.g., in studying the incremental contribution by some commodity to economic growth), in the law (e.g., in apportioning responsibility), etc. However, a widely agreed upon analysis of counterfactuals has proved elusive. The so-called *problem of counterfactuals*, raised notably by Chisholm and by Goodman in the 1940s, is that of providing adequate (in particular, noncircular) truth conditions for the counterfactual.

Let '$A \Box\!\!\rightarrow B$' denote a counterfactual with antecedent A and consequent B. Fixing the truth value of A and B clearly does not fix the truth value of $A \Box\!\!\rightarrow B$. For example, 'if I were to drop this fragile glass, it would break' is true, 'if I were to drop this fragile glass, it would turn into a hamster' is false, yet the antecedents ('I drop this fragile glass') and consequents ('the glass breaks,' 'the glass turns into a hamster') are all false. This rules out as an adequate analysis the material conditional $A \supset B$ (defined to be equivalent to 'not-A or B'), which is automatically true when A is false. According to *metalinguistic* approaches, $A \Box\!\!\rightarrow B$ is true if and only if A, together with suitable further premises (on one version, stating appropriate laws of nature and initial conditions), implies B. Mackie takes counterfactuals to be themselves elliptical presentations of such arguments, and thus lacking truth values altogether. The advent of possible worlds semantics for modal logics in the 1960s gave new direction to the analysis of counterfactuals. Truth conditions for them are proposed in classic works by Stalnaker (1968) and Lewis (1973).

Stalnaker motivates his account by Ramsey's recipe for counterfactual reasoning ('Ramsey's test'): First, add the antecedent (hypothetically) to your stock of beliefs; second, make the minimal adjustments required in the rest of your beliefs to maintain consistency; finally consider whether or not you then believe the consequent. Stalnaker's truth conditions are intended to apply generally to conditionals of the form 'if A, then B,' which Stalnaker denotes '$A > B$,' pragmatics distinguishing between indicative conditionals and counterfactuals. He assumes that for each world w there is an ordering of worlds according

to their similarity, or 'closeness,' to w. Call a world at which A is true an *A-world*; the proposition A is regarded as the set of A-worlds. His guiding idea is that $A > B$ is true at a world w if and only if B is true at the closest A-world to w.

Lewis rejects Stalnaker's assumption that, for any given w and A, there is exactly one closest A-world to w. His truth conditions take as primitive the notion of comparative closeness of worlds: $A \Box\!\!\rightarrow B$ is true at w if and only if some AB-world is closer to w than any $A\bar{B}$-world. For both Stalnaker and Lewis, the counterfactual is vacuously true if the antecedent is impossible.

2. *Probabilities of Counterfactuals*

One important line of research considers counterfactuals within the quantitative framework of probability theory. In particular, there has been much interest in the connection between *probabilities of conditionals* (of which counterfactuals are an important class) and *conditional probabilities*. Let P be your subjective probability function, so that $\mathrm{P}(A \Box\!\!\rightarrow B)$ is your degree of belief in $A \Box\!\!\rightarrow B$. There are various reasons (linguistic intuitions, considerations stemming from Ramsey's test, etc.) for thinking that counterfactual reasoning is, or should be, subject to the following constraint: (*) $\mathrm{P}(A \Box\!\!\rightarrow B) = \mathrm{P}(B|A)$ for all A, B in the domain of P such that $\mathrm{P}(A) > 0$.

Stalnaker proposed a version of this identity, which can be shown to vindicate his truth conditions above, and to vitiate Lewis's. In response, Lewis (most famously in 1976) proved a number of 'triviality results': under certain assumptions, the only classes of probability functions that can sustain (*) consist solely of members that take at most four values, and that are thus trivial. Since then, a number of authors have strengthened these results (see Eells and Skyrms 1994 for a collection of relevant articles).

Nonetheless, an identity somewhat in the spirit of (*) does hold for the Stalnaker conditional. Start with probability function P. Among other things, P assigns probabilities to individual worlds. Define a new probability function P_A as follows: for each world w, P_A shifts the probability that P assigns to w to the closest A-world to w. P_A is said to be derived from P by *imaging* on A. Some authors model counterfactual reasoning by imaging (eschewing conditional probabilities as in (*)). Lewis (1976) shows that there is an intimate connection between the Stalnaker conditional and imaging: $\mathrm{P}(A > B) = \mathrm{P}_A(B)$ (for all P, A and B). The exploration of tenable variants of (*) is likely to continue to be a lively area of research.

3. *Counterfactual Reasoning in Decision Theory*

The distinction between probabilities of conditionals and conditional probabilities is important for decision theory, a normative theory of how one's beliefs and desires in tandem determine what one should do. Such a theory typically combines an agent's utility function u and probability function P to give a figure of merit for each possible action, called the *expectation* or *desirability* of that action. Let $S_1, S_2, \ldots, S_n$ be a partition of possible states of the world. *Evidential decision theory*, as presented by Jeffrey (1966), measures the choiceworthiness of action A by a conditional probability-weighted average of the utilities of its possible outcomes (the conjunctions A and S_i):

$$\mathbf{V}(A) = \sum_i u(A \& S_i)\mathrm{P}(S_i|A) \qquad (1)$$

It thus prescribes that an agent acts so as to maximize $\mathbf{V}$. Various authors believe that evidential decision theory gives incorrect verdicts in Newcomb problem cases, in which an action is thought to be evidence for the obtaining of a given state, without in any way causing that state (see *Newcomb's problem*). This has led to the development of rival versions of decision theory, sharing the name *causal decision theory*, that replace the 'evidential' conditional probability weights with weights that are supposed to capture the agent's degrees of belief about the causal efficacy of the action in bringing about each possible state of affairs. Gibbard and Harper (1981), following Stalnaker, use probabilities of counterfactuals as the weights

$$\mathbf{U}(A) = \sum_i u(A \& S_i)\mathrm{P}(A \Box\!\!\rightarrow S_i) \qquad (2)$$

They then advocate the maximization of $\mathbf{U}$. Proponents of causal decision theory believe that it agrees with evidential decision theory in unproblematic cases, and that it delivers intuitively correct answers in Newcomb problem cases.

4. *Counterfactual Reasoning in Game Theory*

Counterfactual reasoning can play a crucial role in the rational deliberation of two or more agents strategically playing against each other, their pay-offs depending on what they all do—the province of *game theory*. Paralleling the distinction we found in decision theory, Stalnaker (1996) distinguishes two sorts of counterfactual reasoning, *causal* and *epistemic* (the latter could also be called *evidential*). In the former, agents reason about what would happen if they or others were to act in ways that they know or believe to be nonactual. In the latter, they consider how their own or others' beliefs would change if they were to learn things that they expect not to learn, for example, contrary to what is standardly taken to be common knowledge (and thus regarded as certain) by all the players, an irrational choice is hypothetically made. Indeed, such (epistemic) counterfactual reasoning leads naturally to some important refinements of Nash equilibrium, such as Selten's subgame perfect and ('trembling hand') perfect equilibrium (see *Game Theory*).

Various models have been proposed of how an agent's beliefs are disposed to change in the face of unexpected actions by others, and the consequences of such dispositions for rational action—see, for example, Bicchieri (1988) (for extensive form games) and Stalnaker (1996) (for normal form games). Closely related is Shin's (1989) theory of equilibrium in normal-form games. He gives a counterfactual-based account of rationality—a rational agent never acts in such a way that he would have done better had he acted differently—and proves that his account is equivalent, under certain assumptions, to those that employ game theory's standard solution concepts of Nash equilibrium and correlated equilibrium. Skyrms (1998) integrates the themes that we have discussed in this and the previous two sections. He gives a probabilistic theory of subjunctive conditionals that finds its inspiration in (*), with 'assertability value' replacing the 'P' on the left-hand side, and expectation of conditional chance replacing the right-hand side. He then uses this theory to give a unified treatment of counterfactual reasoning in individual decision theory, and in normal form and extensive form games.

Finally, in what might be taken as an indicator of a future trend in this field, more and more sophisticated probabilistic models are being developed for the representation of conditional beliefs in games. The economists Battigalli and Siniscalchi (1998), for example, construct a space of infinite hierarchies of conditional probability systems in which one can represent any statement about the players' dispositions to hold (arbitrarily high-order) beliefs regarding the other players.

See also: Causal Counterfactuals in Social Science Research; Counterfactual Reasoning: Public Policy Aspects; Decision Theory: Bayesian; Decision Theory: Classical; Federalism; Nonstandard Reasoning; Piaget, Jean (1896–1980); Probability and Chance: Philosophical Aspects; Probability: Interpretations; Rational Choice Explanation: Philosophical Aspects

Bibliography

Battigalli P, Siniscalchi M 1998 Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games. EUI Working Paper ECO No. 98/29, European University Institute, France, Italy

Bicchieri C 1988 Strategic behavior and counterfactuals. *Synthese* **76**: 135–69

Eells E, Skyrms B (eds.) 1994) *Probability and Conditionals*. Cambridge University Press, Cambridge, UK

Gibbard A, Harper W L 1981 Counterfactuals and two kinds of expected utility. In: Harper W L, Stalnaker R, Pearce G (eds.) *Ifs*. D. Reidel, Boston

Jeffrey R C 1966 *The Logic of Decision*. Chicago University Press, Chicago

Lewis D 1976 Probabilities of conditionals and conditional probabilities. *Philosophical Review* **85**: 297–315

Lewis D K 1973 *Counterfactuals*. Basil Blackwell, Oxford, UK

Shin H S 1989 Two notions of ratifiability and equilibrium in games. In: Bacharach M, Hurley S (eds.) *Foundations of Decision Theory*. Blackwell, Oxford, UK

Skyrms B 1998 Subjunctive conditionals and revealed preference. *Philosophy of Science* **65**: 545–74

Stalnaker R 1968 A theory of conditionals. In: Cornman J (ed.) *Studies in Logical Theory*. Basil Blackwell, Oxford, UK

Stalnaker R 1996 Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy* **12**: 133–63

A. Hajek

Courtroom Consultation: Psychology

This article is written from a US perspective. A domain now about three decades old, 'psychology and courtroom consultation' involves applying psychological theory and techniques in order to assist litigants present maximally effective trial presentations to juries that are as sympathetic as possible to litigants' claims. Also called 'trial consultation,' 'litigation consultation,' 'jury (de)selection,' and 'jury science,' the field grew dramatically toward the close of the twentieth century along several dimensions, including number of practitioners, scope of activities, level of public recognition, and debate over its ethical and legal legitimacy.

1. Historical Context

The first well-publicized instance of social scientists serving as jury consultants was the trial of Philip Berrigan and seven other defendants charged with criminal conspiracy involving numerous anti-Vietnam War activities. Faced with defending their clients, known as the 'Harrisburg Seven,' in the conservative community the federal government had chosen for the trial, the defense attorneys asked a group of social scientists, headed by Jay Schulman, a Columbia University sociologist, to help select a favorable jury from a largely unfavorable community. Using special public opinion surveys, the team interviewed hundreds of citizens to determine what background characteristics were correlated with pro-prosecution and pro-defense attitudes. They then used these empirical relationships as the basis for recommending which prospective jurors the defense attorneys should 'strike' from the panel. The trial ended in a hung jury (the vote was 10–2 for acquittal). Based on this outcome and several other highly visible trials throughout the 1970s in which politically liberal defendants relied on social scientists for jury selection advice and were found not guilty, trial consulting blossomed into a cottage industry.

In the first decade of psychology and courtroom consultation, consultants typically worked for poor defendants or socially unpopular individuals charged with political crimes. Their work was often done *pro bono*, and after the death penalty was approved by the Supreme Court in 1976, a particular focus of what was then called 'jury selection' was on defendants charged with capital crimes. Capital litigation proved to be a major training ground for jury consultants in part because the jury selection process (known as *voir dire*) is expanded in the typical death-penalty trial and in part because juror attitudes tend to correlate more strongly with beliefs about penalties than with assessment of evidence. As jury consultation became better known, its application spread to civil trials and to litigants with greater financial means. A landmark was MCI's antitrust suit against ATT in 1980, in which the MCI plaintiff team relied heavily upon jury consultant recommendations and won a nearly two billion dollar judgment.

By the 1990s, what had begun as a largely volunteer activity born from social consciousness and intellectual curiosity had become a large commercial enterprise associated with trials carrying enormous financial stakes. Jury consultants were credited with major roles in several trials involving the wealthy and the privileged, including O. J. Simpson, William Kennedy Smith, Imelda Marcos, and a host of large corporations. Consultants such as Jo-Ellan Dimitrius and Donald Vinson received a level of media attention normally reserved only for celebrities. Some consulting firms now employ scores of doctoral-level researchers and consultants and generate millions of dollars in fees annually. Formed in 1982, the American Society of Trial Consultants is the field's main professional organization; it has published a *Code of Professional Standards* that is a statement of aspirational practices rather than an enforceable set of regulations. On-line trial consulting was introduced in the 1990s; one service impanels 'cyberjuries' to listen to evidence and render mock decisions that can be analyzed by consultants and trial attorneys (Strier 1999).

2. *Techniques of Psychology and Courtroom Consultation*

Courtroom consultation is not limited to weeding out unfavorable prospects from juries although that goal has received inordinate attention in popularized accounts. Recognizing that jurors' demographic and personality characteristics usually do not correlate strongly with group verdicts in general, litigation consultants have turned their focus to studying intensively the dynamics and issues of each specific case (Vinson and Anthony 1985). Therefore, courtroom consultation applies many techniques to the detailed analysis of the individual case and considers how the main issues in that case will be evaluated by different types of jurors. Among the techniques employed by trial consultants, the following five are the most important:

2.1 *Public Opinion Surveys*

Consultants conduct public opinion surveys, usually by telephone, to measure the attitudes and beliefs of residents in communities from which jurors subsequently will be drawn. These surveys yield empirical profiles of favorable and unfavorable jurors, which are then used to guide the exercise of *peremptory challenges*, the mechanism by which attorneys arbitrarily deselect a certain number of individuals from serving on a jury. Although demographic and personality characteristics tend to show modest-at-best correlations with case-relevant beliefs, life experiences (e.g., personal victimizations, individual litigation history), attitudes about social issues, and general values can be fairly robust predictors of verdict tendencies and therefore are the focus of many surveys.

2.2 *Focus Groups and Mock Juries*

A favored tool is the pretrial focus group or *mock jury* in which the consultant arranges for a jury-sized group to evaluate an abbreviated,'dress rehearsal' version of the upcoming trial, including opening statements, evidence, and closing arguments. The focus group then discusses the evidence and makes individual and group decisions as the consultant observes the process. These simulations allow the consultant to pre-test several dimensions of the upcoming trial—impact of attorney styles, coherence of the theory of the case, clarity of exhibits, and persuasiveness of the overall presentation—before a group of people selected to resemble the real jury. Focus groups are sometimes repeated more than once before a trial, giving the litigation team the opportunity to gauge the effects of variations that were suggested from initial iterations of this methodology. In a few complex, lengthy trials, consultants have impaneled a mock jury to listen to the case simultaneously as it unfolds before the real jury, providing day-by-day data on how the trial is being perceived and enabling suggestions for mid-course corrections and adjustments.

2.3 *Pretrial Investigations; In-court Observations; Post-trial Interviews*

Consultants use several techniques to gather information about individual jurors. Pretrial investigations of jurors' backgrounds or interviews with key informants who know specific jurors are employed, but both methods have two drawbacks—resentment over what jurors perceive as an invasion of privacy and

limitations in the reliability and validity of informant data.

Consultants emphasize the in-court observation of prospective jurors as a basis for their advice on exercising peremptory challenges. A key element in this strategy is to evaluate prospects' answers to questions that have been crafted carefully to reveal their beliefs, biases, and preferences. In addition to the content of the answer, the nonverbal cues associated with jurors' responses reveal attitudes as well. For example, aloofness toward one type of person coupled with deference toward another may suggest authoritarianism, a juror quality generally not favorable for criminal defendants. Another technique is to conduct post-trial interviews with the jury to assess the validity of the consultant's decisions and to learn how the jury evaluated the trial. This method is not always available because some judges prohibit such interviews.

2.4 *Modification of Attorney Style; Enhancement of Voir Dire; Preparation of Witnesses*

In addition to studying and predicting juror behavior, litigation consultants often attempt to modify aspects of litigant behavior or influence the litigation process itself. In both instances, the purpose is to use psychological knowledge or techniques to gain an edge over the opposing litigant. For example, consultants can teach attorneys how to ask questions of prospective jurors so they are comfortable revealing their true feelings about sensitive topics of crime, race, justice, and punishment. Open-ended questions, often prefaced by a self-disclosure by the attorney, are effective in this regard as are those active-listening techniques used by contemporary psychotherapists. Similarly, consultants often recommend how *voir dire* can be conducted so that the maximum amount of candid information can be obtained from prospective jurors without biasing other members of the panel. Creating special questionnaires to be completed by prospective jurors prior to trial is a prime example. Another strategy is to seek permission to question jurors one at a time on key issues rather than to interrogate them as a group. Whether any of these innovations is used in a given trial is a matter over which the trial judge has almost total discretion so the consultant's role is simply to suggest that the attorney request the judge to allow them.

Most controversial among these techniques is witness preparation. Some consultants help prepare witnesses for their testimony with the intention of assisting them make as effective a presentation of the facts as possible. Other consultants are reluctant to prepare witnesses because of ethical concerns about the possibility of unintentionally encouraging or suggesting changes in the 'facts' to which a witness will testify. In general, most consultants believe that interventions aimed at 'how' the facts are presented or at the 'style' of testimony are acceptable.

2.5 *Development of a Case Theme*

Essential to the success of litigation consulting is the development of a case theme to which prospective jurors will be sympathetic. The modern trial has been likened to a morality play in which an audience of jurors writes the final scene (Nietzel et al. 1998). The first task is to deselect from this audience people whose personal characteristics and experiences might make them unfavorable 'playwrights.' Second, armed with the knowledge gleaned from public-opinion surveys and mock juries, the consultant will attempt to develop a case theme that promotes the plot lines and final resolution most favorable to the consultant's 'side.' This theme will then be reinforced in every way possible. Through the selection of *voir dire* topics, the framing of an opening statement, the structure of evidence presentations, and the building of a closing argument, the consultant will help the attorney repeatedly stress the most salient psychological factors of the case.

Jurors tend to form a narrative or private story that summarizes trial evidence into a coherent, emotionally and cognitively compelling account (Pennington and Hastie 1988). To the extent that a consultant helps an advocate present evidence that advances a captivating narrative, that advocate should improve the chances of a favorable verdict. For example, in the sentencing phase of capital trials, defendants strive for narratives that emphasize personal mitigators of culpability—e.g., youthfulness, psychological disturbance, a history of abuse, and deeply-felt remorse. Conversely, prosecutors fashion stories that emphasize the brutality of the crime, the loss suffered by the victim's family, and the deterrent value of capital punishment. The side that promotes the most convincing narrative enhances its chances of jurors finishing the script with the moral resolution it desires.

3. *Effectiveness of Psychological Courtroom Consultation*

The effectiveness of psychological courtroom consultation is disputed. Its initial successes in the political trials of the 1970s lent it an aura of potency that was exaggerated by the fact that political conspiracy charges are difficult to prove before juries of any kind. An additional confound is that attorneys who make the extra effort to employ trial consultants might also be more diligent in preparing other facets of their cases than attorneys who forego litigation consulting services. In fact, little empirical research evaluating the effectiveness of litigation consulting has been published, and what is available is plagued by questionable methodology (Cutler 1990; Strier 1999), including an over-reliance on mock jury studies. Consequently, it is still not certain whether and to what extent 'scientific jury selection' surpasses the effectiveness of traditional selection strategies. However, Horowitz (1980) did

compare scientific jury selection to traditional attorney methods in four simulated criminal trials. Neither approach was superior across all trials. Psychological consultation was more effective when there were clear-cut relationships between personality and demographic variables and jurors' votes, but when these relationships were weak, consultants' accuracy did not surpass the conventional wisdom of practicing attorneys.

One common claim is that jury consultants are unlikely to exert significant effects because the weak relationships between individual juror characteristics and verdicts would vitiate almost any empirical selection strategy (Saks 1976). In cases where the evidence is about equally persuasive for both sides, or where the trial issues provoke strong attitudes, individual juror characteristics have been shown to exert small to moderate influences (Hastie et al. 1983, Moran and Comfort 1982, Penrod 1990) that provides an opportunity for consultants to have an impact. Further, even when a collection of individual difference variables account for only a small amount of verdict variance, this relationship can still translate into trial outcomes that are of substantial practical importance.

One study of defense-team psychological consultants examined 35 jury verdicts in capital trials (Nietzel and Dillehay 1986). Juries recommended the death sentence in 61 percent of verdicts in trials where no consultants were employed compared to 33 percent of verdicts in trials with jury consultants, producing an effect size of 0.26. Because the cases differed on several variables besides the use of consultants, it is not possible to conclude that the effect was due to consultants, but the results are consistent with claims that consultation might be effective in cases in which juror attitudes are important, as they are likely to be when a jury must choose between life and death.

4. *Ethical and Legal Criticisms*

As the practice of psychology and courtroom consultation expanded, so too have legal and ethical objections to it. Critics have condemned these techniques as 'jury rigging' or 'jury tampering' and claim that they undermine the fairness of trials and the public's confidence in jury verdicts (Adler 1994). Another objection is that courtroom consultants have become affordable only by wealthy individuals or large corporations, producing a fundamental disadvantage for poor or even average-income litigants.

Fueled by these criticisms as well as by outrage over what have appeared to be miscarriages of justice in a few trials where consultants played a major role, public concern about litigation consultation has led to several suggested reforms. The following proposals are the most common (Strier 1999): (a) a complete ban on courtroom consultation; (b) a restriction of *voir dire* such that only limited questioning by the judge rather than the attorneys is permitted; (c) a curtailment of peremptory challenges so that there is less opportunity to remove jurors perceived as unfavorable; (d) a requirement that survey data or other information developed by a trial consultant be shared with the opposing litigant; (e) a prohibition against pretrial investigations of prospective jurors; (f) a requirement that litigants disclose publically their use of consultants; and (g) a suggestion that courtroom consultation be regulated through professional licensure and practice standards.

Few, if any, of these 'reforms' have been implemented, at least in the US, nor is it clear that they would be beneficial or yield fairer justice. The American system of justice is fundamentally adversarial in nature; each litigant is expected to present its case as zealously as possible, within the confines of the law, with the goal of persuading the jury to return a verdict favorable to that litigant. Attorneys on both sides of a case are expected to use every legal means available to woo the jury, present evidence, offer arguments, and counter the opponent in order to win the dispute. The objection that courtroom consultation gives one side an advantage ignores the fact that one side almost always has advantages—a better attorney, more persuasive expert witnesses, more skilled investigators. To single out courtroom consultation, *voir dire*, or peremptory challenges for special curtailments is a slippery slope premised on the faulty notion that each side in a trial can be 'required' to be equal. It is also important to remember that consultants cannot 'stack' juries in their favor; they can only prune juries of individuals they believe are unfavorable. The ability to eliminate jurors perceived as biased through the skillful use of peremptory challenges might be worth preserving, particularly if it enhances the general perception of trial fairness.

Bibliography

Adler S J 1994 *The Jury: Trial and Error in the American Courtroom*, 1st edn. Times Books, New York

Cutler B L 1990 Introduction: The status of scientific jury selection in psychology and law. *Forensic Reports* **3**: 227–32

Hastie R, Penrod S D, Pennington N 1983 *Inside the Jury*. Harvard University Press, Cambridge, MA

Horowitz I A 1980 Juror selection: A comparison of two methods in several criminal cases. *Journal of Applied Social Psychology* **10**: 86–99

Moran G, Comfort J C 1982 Scientific jury selection: Sex as a moderator of demographic and personality predictors of impaneled felony juror behavior. *Journal of Personality and Social Psychology* **43**: 1052–63

Nietzel M T, Dillehay R C 1986 *Psychological Consultation in the Courtroom*. Pergamon Press, New York

Nietzel M T, Hasemann D, McCarthy D M 1998 Psychology and capital litigation: Research contributions to courtroom consultation. *Applied and Preventive Psychology: Current Scientific Perspectives* **7**: 121–34

Pennington N, Hastie R 1988 Explanation-based decision making: Effects of memory structure on judgment. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition **14**: 521–33
Penrod S D 1990 Predictors of jury decision making in criminal and civil cases: A field experiment. *Forensic Reports* **3**: 261–78
Saks M J 1976 The limits of scientific jury selection. *Jurimetrics Journal* **17**: 3–22
Strier F 1999 Whither trial consulting? Issues and projections. *Law and Human Behavior* **23**: 93–115
Vinson D E, Anthony P K 1985 *Social Science Research Methods for Litigation*. Michie, Charlottesville, VA

M. T. Nietzel

Courts and Adjudication

1. Introduction

Adjudication is a type of dispute resolution that relies on an independent, third-party facilitator (Eckoff 1965, Shapiro 1981). Such triadic relationships range from a mutual acquaintance who may serve as a go-between to help conflicting friends, to a powerful state which imposes a resolution on warring neighbors. Within the legal context there are three distinct types of dispute resolution triads: the mediator, the arbiter, and the judge. While all are engaged as third-party resolvers of disputes, they represent a continuum along which the freedom of the parties to the dispute is progressively limited. A mediator will facilitate communication between conflicting parties and perhaps propose a solution, but the parties themselves must agree to this solution. In contrast, an arbiter is a third-party decision-maker, selected and granted authority by the parties to impose a binding resolution.

Judges presiding over proceedings in courts share much in common with mediators and arbiters. All three are independent and strive to bring about a resolution of the dispute that will be perceived to be fair. But adjudication is distinguished from both mediation and arbitration in several ways. Judges are externally appointed by the state and parties have no or very little control over who is selected to hear their dispute. Judges impose authoritative outcomes, but unlike arbiters, judges are expected to render their decisions according to pre-established norms. Furthermore, unlike mediation, adjudication does not require the consent of the parties either to participate or to resolve the disputes. Indeed, typically adjudication is triggered by the action of one party against the other. In some cases, for instance in criminal proceedings, a public prosecutor acting on behalf of the state initiates a proceeding not only against the will of the accused but without the consent of the victim.

Because their proceedings do not require the consent of all parties, courts face a crisis of legitimacy. As a consequence, courts must find substitutes for consent. These substitutes can be understood as a set of principles (Cappalletti 1989). The more important of these are the principle of party participation, the principle of the rule of law, the principle of impartiality, and the principle of independence. Indeed these principles have come to be understood as the defining characteristics of courts. The principle of party participation assures that those likely to be affected by a court's decision have an opportunity to make arguments and offer proofs in their behalf, but it also requires that the court respond with reasoned arguments to these concerns (Fuller 1981). This principle may help account for the propensity of courts almost everywhere to encourage settlements in lieu of full-fledged trials and to value confessions of guilt or acknowledgements of liability. Although such practices have the obvious benefit of reducing the costs and risks of trial, they also provide–even if only symbolically–the active consent of the defendant to the process. The principle of the rule of law prohibits judges from making *ad hoc* decisions; it requires them to render decisions according to pre-existing norms, the rules of the game which both parties have used (or are free to use) to organize their expectations and to arrange their affairs. Both party participation and the rule of law can be understood as approximations for consent. Participation provides parties with a voice and requires that their concerns be addressed; the rule of law presumes a shared set of norms or at least a set of rules under which all the parties operate. As we shall see, the rule of law is not a necessary condition for judicial independence, though it is closely associated with the notion of judicial independence.

The principle of impartiality requires that judges have no interest in the outcome and no reason to favor one party over the other. This is usually achieved through the nature of the office; full-time judges, whose compensation is not dependent upon the outcome of the controversy, do not have any personal interest in the outcome of the controversy or ties to the parties of the dispute.

The related principle of judicial independence holds that the judge is not only independent of one or the other of the parties, but is also independent of other influences as well; the judge's full fidelity is to the rule of law. This poses an obvious and important challenge and may be the most difficult of the principles to implement. At times those not party to a dispute, such as an industry or trade association, may have an interest in pressing for a particular outcome in a case or for a particular interpretation of a rule. More challenging still is that, as officials of the state, judges may be under pressure to adhere to political concerns rather than to the rule of law. Indeed, these two concerns are not always clear and separate. Almost everyone would agree that judicial independence is likely to be jeopardized if political authorities dismiss or force the resignation of judges with whom they disagree or if courts routinely uphold edicts of the

regime that have not been duly enacted. A more subtle problem of judicial independence arises when the state enacts new laws, resists interpretative approaches of the judges, removes issues from the jurisdiction of courts, and seeks to influence judicial selection.

Similarly the issue of independence arises when judges make policy. Typically judges make policy when they confront new situations not covered by existing doctrine or when existing doctrine contradicts deeply held personal or widely supported social norms. In such situations, courts may continue to apply outmoded norms. However, courts might alternatively create new doctrine through any of several institutional devices, such as the use of fictions (Fuller 1967), reliance on metaphors, and reasoning by analogy (Levi 1961), or treat vague and open-ended provisions in constitutions or statutes as grants of authority authorizing them to make new law (Feeley and Rubin 1998). In any of these situations, judges can be criticized for abandoning the rule of law and hence for compromising their independence. Yet judges who insist on following outmoded or restrictive laws also run the risk of being criticized for failing to act independently and for being subservient to an unpopular regime or a new regime. Independence surely means more than freedom to apply a narrowly-prescribed process of interpretation.

No one seriously argues that the modern state lacks authority to enact a wide variety of laws (though constitutional safeguards, conceptions of individual rights, and the like may severely constrain what the state may do and how it must enact laws). Still, the rule of law requires that laws be duly enacted, and publicized. At a minimum an independent judiciary is one which honors and upholds these procedural limits on state power. Still, governments, particularly new regimes following major political realignments or revolutions, may exert strong pressures on the judiciary to fall in line with the 'spirit' of the new regime. The new regime may be impatient with the slow pace of change through the rule of law or may react to what it perceives as a conservative judiciary wedded to the old order; it might exert pressure upon the judiciary to look at things in a new way. Such pressures can take any number of forms, including requesting judicial resignations, altering the procedures for judicial selection and advancement, selecting new judges whose views are aligned with those of the new regime, demoting judges deemed to be recalcitrant, or simply ignoring judicial actions that are deemed unacceptable. This is more difficult in democratic regimes since they are all more or less committed to respecting judicial independence.

Thus it would be a mistake to treat the issue of judicial independence as a problem faced only by authoritarian or totalitarian regimes. In modern states everywhere, judges are governmental officials. In democracies, both the law and the constitution are understood to be products of popular will. Courts everywhere are likely to be affected by, shaped by, and at times in conflict with larger governmental processes in ways that raise issues about the independence of the judiciary. Indeed, since much of the legitimacy of modern government is closely tied to a belief in majoritarian principles and to the rule of law, conflicts between government and courts are all but guaranteed in all sorts of political regimes. This complex relation between law and government, adjudication and politics, and courts and governmental regimes is addressed in the sections below.

2. *Judicial Independence*

The quest for judicial independence in common law and civil law systems has taken quite different forms, though the differences may be due less to the particular legal forms than to the political systems in which they are situated. With the rise of absolutist states in early modern Europe and continuing well into the twentieth century, the judiciary was strongly integrated into the machinery of government and courts were well-known as arms of the ruling regimes. With the rise of support for constitutional democracy in Europe, reformers sought to institutionalize the principles of the rule of law and judicial independence. One of the main objectives of constitutionalism is 'to limit the arbitrary exercise of power and make it legally accountable' (Sartori 1987). Although the rule of law is embraced in all constitutional systems, it has taken quite different forms. In countries with traditions of strong states (which historically have drawn their judiciaries into the service of the state), judicial independence has been pursued by integrating the judiciary into the structure of the civil service, by developing comprehensive codes of law, so that laws are understood as bureaucratic-like rules, and by characterizing judges as technical experts. Judges are free to pursue their duties, but within these severe constraints. This conception of the separation of powers and judicial independence assumes the primacy of the legislature; the legislature represents popular will, and courts play a subordinate although independent role in finding facts and applying rules. Thus in strong states, judicial independence is purchased at the price of a narrow conception of the judicial role.

One can see this in the judiciaries of countries with a civil law tradition, where law is understood as a set of technical rules and judges are considered to be technicians (Merryman 1985). And one can see it in some authoritarian states, where judges have sought to protect themselves by defining their roles narrowly and by insisting on strict interpretations of the law. But where the law involves politically salient issues—such as labor relations or economic regulations—governments everywhere are likely to remove jurisdiction from ordinary courts and establish specialized tribunals more directly responsive to political

controls to deal with disputes. Thus in all countries, a large group of disputes between citizens and the state do not fall under the jurisdiction of independent courts but are assigned to specially created administrative tribunals that are more responsive to governmental direction. Similarly, issues of constitutional interpretation and judicial review, i.e., the power of courts to declare actions of other branches of government unconstitutional, have generally been removed from ordinary courts and assigned to more politically responsive bodies. As a result, many disputes between citizens and the state do not fall under the jurisdiction of ordinary courts, but are instead assigned to special administrative tribunals responsive to governmental direction. Thus, in disputes between individuals which are not likely to raise political controversy, judicial independence is protected, but where issues which may be politically charged, a more politically responsive judicial institution has been established. In a variation on this theme, Damaska has argued that, in regimes with a tradition of a strong or activist state, courts are designed and expected to implement government policy (1977).

In common law countries with traditions of weak states, judicial independence has been protected in a quite different way. Ironically, such common law political systems tend to foster 'strong' courts. Under the common law, the judiciary—and indeed, the entire legal system—has been understood as something distinct and separate from the state. The idea of the rule of law incorporates more than the belief in constitutionalism (i.e., that government officials are constrained by and must act within the law). Rather, it envisions a legal system in which law itself is separate from and independent of government. In these countries the independence of law is protected not so much through a careful civil service-style judiciary and minutely crafted legal codes, but through a robust legal profession and with close associations with the judiciary. Although here too judges are officials of the state, they are also regarded as an integral part of civil society and are closely associated with an autonomous legal profession (Halliday and Karpik 1998). Constitutionalism and the rule of law are understood not only as procedural constraints on the government; they are external limits on the state as well. Under such a system, courts are also likely to have judicial review. Ironically, this conception of the limited state produces a strong judiciary and an expansive notion of judicial independence (Damaska 1986).

It is important to underscore some of the salient differences produced by these two conceptions of courts and of judicial independence. On the European continent, independence is secured by subordinating the role of judges to the democratic principle—the will of the people expressed through parliament—and by insisting upon detailed codes so that judges have no need to make law but only to apply it. Typically this is reinforced by a career judiciary whose initial selection and subsequent promotions are based on seniority or merit. In contrast, common law systems entrust substantial law-making powers to the courts, since the law itself is conceived as something independent of the state and politics. Although there is an admitted political feature to judicial selection, judges are recruited from senior members of the bar, continue to maintain close ties to the bar, and are not dependent upon political authorities for support. Countries elsewhere have often adopted a hybrid of American-like and Continental institutions, though in all but former British colonies, the result has been much closer to civil law than common law style structures.

The differences between civil law and common law systems are familiar to anyone with some exposure to comparative law. Yet these differences only identify central tendencies; legal systems are too varied and too complex to be divided into any such neat categories. For instance in continental systems, even the most comprehensive code is not likely to anticipate all problems that may come within its scope; as a result a judge may have to make law to fill in the gaps. Similarly, law-making in modern common law countries is no longer primarily judge-made; modern law everywhere is increasingly statutory and administrative. Moreover, since the late twentieth century, the independence of civil law judges has been strengthened. In many countries, the government no longer controls judicial careers and judges have begun to play a more active role in law-making.

Furthermore, judicial independence—of either variety outlined above—is much more likely to occur in regimes with competitive political parties. Almost everywhere, one-party or predominant-party states have constrained the independence of the judiciary and sought to use it for their own political purposes. Conversely, countries with competitive parties which alternate in controlling the government or where political parties are weak (e.g., Italy and Israel in the late twentieth century) are more likely to have independent judiciaries. Although all political victors may be tempted to convert the judiciary for their own use, those knowing that they may hold office only temporarily may resist the temptation in order not to jeopardize their long-term interests. Recognizing the harm that each party could impose on the others, parties in competitive political systems all have an incentive to maintain an independent judiciary (Landes and Posner 1975).

3. *The Business and Organization of Courts*

The popular image of trial courts is that they try cases and resolve conflicts. But a careful examination of the process of adjudication reveals a more complex picture. The vast majority of all cases disposed of by courts—perhaps as many as 98 percent— are resolved by the parties themselves, without a full-dress trial.

Indeed, in many cases, there is not even any real controversy. In civil cases, court filings are often treated as signals to the opposing party that the other person is serious about pursuing his or her grievance. This is often enough to precipitate serious negotiations and resolution of the legal conflict. Similarly, in criminal cases in both civil law and common law countries, most matters are disposed of without sustained judicial involvement, through unilateral prosecutorial decisions to drop charges, plea negotiations, or simply pleas of guilty. Indeed, in many countries, vast numbers of lesser criminal cases are disposed of through the mail without even a single court appearance by the accused (Feeley 1979, Felstiner 1979). In many of these situations, some combination of the prospects of the cost, time, and embarrassment of a trial is enough to encourage the parties to negotiate a settlement in the shadow of the courthouse (Mnookin and Kornhauser 1979). In such situations, the court may have been actively involved in the preliminary stages of the conflict and may even have handed down important pretrial rulings dealing with the scope of discovery, the admissibility of evidence, and the like. Such rulings provide cues to the participants as to the likely outcome of their case and thus precipitate negotiations and settlement rather than a prolonged trial process. In such cases, although the court did not formally resolve the conflict, it did nevertheless play an important role in its resolution.

The reasons for negotiations in the shadow of the law are fairly obvious. Filing a case often provides each party with the first opportunity to determine the strength of the opposition. This experience leads weak plaintiffs to drop their cases or to try to settle, and weak defendants to admit liability quickly or to settle. Cases that continue to trial are likely to be close cases in which both parties believe that they stand a good chance of winning. Economists have formalized this model, hypothesizing that, all things being equal, defendants would win (or lose) just as often as plaintiffs in full-fledged trials (Priest and Klein 1984). Although uncertainty, changes in legal rules, and the degree to which an initiated case is controlled by the court can affect this dynamic, in the long run a stable equilibrium should prevail.

At the outset of this article, we noted that one of the problems of adjudication, in contrast to mediation and arbitration, is that the parties neither resolve their own dispute nor select the third-party decision-maker and that this can generate an aura of illegitimacy around the judicial process. If true, this may account for the ubiquitous tendency for parties to settle in the shadow of the courthouse and for courts to foster settlement (Shapiro 1981). Despite this, the tiny fraction of cases that go to full-dress trial provides the public with its perception of the nature of courts and adjudication.

There are still other types of adjudication which for all practical purposes do not involve any real disputes at all. In these situations, rather than adjudicate a controversy, the court simply ratifies or oversees a legal transaction that has been made elsewhere, and often hears no contested issues. Thus for instance, in many countries, trial courts are required to certify divorces, even if they are no-fault. Similarly, trial courts are often required to certify changes of name, uncontested child custody arrangements, conveyances of property, probate and bankruptcy proceedings, and to oversee a host of administrative matters. Many of these actions should be understood more as notary-like activities than as adjudication.

In both civil and common law systems, trial courts are divided into two levels: lower courts which handle minor issues and upper-level courts which are reserved for a handful of more important matters. Although a substantially higher proportion of cases are resolved without formal or full-fledged trial in the lower courts, most issues reaching upper level courts are resolved through negotiations or uncontested proceedings as well. Indeed, when all negotiated settlements and uncontested proceedings at both levels of courts are considered together, one finds that only a tiny portion of cases in trial courts of any sort are resolved through full-fledged adjudication (Trubek et al. 1983). One should however bear in mind that adjudication is a process that involves several steps. A party to a case may vigorously contest a pretrial ruling over, for example, the admissibility of evidence. Depending on the outcome of this ruling, it may then decide to drop the case or to admit liability. Although disposition is formally recorded as a dropped case or a negotiated settlement, this outcome was in fact the result of an adjudicated decision. This is not at all uncommon. For instance, in the United States, many so-called plea bargains in criminal cases take place only after a vigorously contested probable cause hearing or a hearing on a motion to suppress evidence. As trial procedures have become more complex over the past two centuries, much of the most important adjudication has come to take place in the pretrial process.

All modern judicial systems provide for some system of appeals. The scope of reasons for appealing a case varies widely in different legal systems, but it is safe to say that a primary purpose of appellate courts is to listen to those who lost at trial and those who claim that the trial court made a mistake. In some legal systems, an appellate court is restricted to consider only claims that the trial court erred in interpreting the law; if it finds an error, it must overturn the trial court's verdict or remand the case to the trial court for further consideration. In other systems, the appellate court can act like a trial court, reconsidering facts not properly considered at the original trial or examining additional facts brought to light since the trial. Whatever their precise duties, appellate courts provide losing parties with an opportunity to ask another court to review the decisions of the trial court and to correct its errors. But in addition to correcting errors,

appellate courts provide the judicial system with a way of overseeing and supervising the activities of lower courts. Appellate courts provide the political regime with a way of checking the independent powers of a highly decentralized system of trial courts. That is, by securing appellate court judges sympathetic to its aims, the ruling regime can gain some assurances that trial courts will act within bounds of toleration. This is an additional reason why civil law systems—that is active, policy-implementing states—tend to have more appeals (Damaska 1977).

4. Conclusion

Adjudication is a triadic structure, involving two disputants and a judge who is charged with imposing a solution on the parties. The process is inherently unstable; the losing party is likely to feel that the judge sided with the other party. In modern societies, where judges are government officials and the state is often a party in a case, the problem of judicial legitimacy can be considerable. Courts have responded to this inherent instability in two ways. They have fostered judicial independence through the formalism of the rule of law ('Don't blame me, blame the law!') and they encourage negotiations in which the contesting parties reach settlement on their own accord. As we have seen, judicial independence can take quite different forms; in civil law countries, it has led to judges being perceived and perceiving themselves as highly circumscribed civil servants 'applying the rules' made elsewhere. In other countries, perhaps most notably some common law countries, judicial independence means primarily judicial autonomy from the state.

In all countries, however, courts tend to encourage parties to arrive at consensual or negotiated outcomes so that they do not have to impose a solution upon the parties. A cursory look at judicial statistics around the world reveals that courts are fairly successful at this. The overwhelming bulk of all cases, of all types, are resolved short of full-fledged adjudication. However, such figures should not be interpreted to mean that the courts are irrelevant or play only a small role in this process. Legal claims are often resolved in the shadow of the courthouse and against looming alternative of a full-fledged trial. Furthermore, they are often settled only after a court has issued a preliminary ruling that has signaled its view of issues and hence a possible outcome.

See also: Common Law; Constitutional Courts; Courtroom Consultation: Psychology; Dispute Resolution in Economics; Justice and Law; Legal Process and Social Science: United States; Plea Bargaining; Procedure: Legal Aspects; Prosecution; Rule of Law; Supreme Courts

Bibliography

Cappalletti M 1989 *The Judicial Process in Comparative Perspective*. Clarendon Press, Oxford

Damaska M 1986 *The Faces of Justice and State Authority*. Yale University Press, New Haven, CT

Eckoff T 1965 Impartiality, separation of powers, and judicial independence. *Scandanavian Studies in Law* **9**: 11–48

Eckoff T 1966 The mediator, the judge, and the administrator in conflict resolution. *Acta Socologica* **10**: 158–66

Felstiner W L F 1979 Plea contracts in West Germany. *Law and Society Review* **13**: 309–26

Feeley M M 1979 *The Process is the Punishment: Handling Cases in a Lower Court*. Russell Sage Foundation, New York

Feeley M M, Rubin R L 1998 *Judicial Policy Making and the Modern State*. Cambridge University Press, New York

Fuller L 1967 *Legal Fictions*. Stanford University Press, Stanford, CA

Fuller L 1981 Forms and limits of adjudication. In: Winston K (ed.) *The Principles of Order: Selected Essays of Lon Fuller*. Duke University Press, Durham, NC, pp. 91–134

Galanter M 1983 Reading the landscape of disputes: What we know and don't know (and think we know) about our allegedly contentious and litigious society. *UCLA Law Review* **31**: 4–71

Gluckman M 1967 *The Judicial Process among the Barotse of Northern Rhodesia*. Manchester University Press, Manchester, UK

Guarnieri C 2001 Judicial independence in Latin countries of Western Europe. In: O'Brien D, Russel P H (eds.) *Judicial Independence: Critical Perspectives from Around the World*. University Press of Virginia, Charlottesville, VA, pp. 147–79

Halliday T, Karpik L 1998 *Lawyers and the Rise of Western Political Liberalism*. Oxford University Press, New York

Jacob H, Blanckenburg E, Krtizer H, Provine D M, Sanders J 1996 *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT

Landes W, Posner R 1975 The independent judiciary in an interest-group perspective. *Journal of Law and Economics* **18**: 875–901

Merryman J H 1985 *The Civil Law Tradition*, 2nd edn. Stanford University Press, Stanford, CA

Mnookin R, Kornhauser L 1979 Bargaining in the shadow of the law. *Yale Law Journal* **87**: 950–87

Nader L 1969 Styles of court procedure. In: Nader L (ed.) *Law in Culture an Society*. Aldine Publishing Co., Chicago, pp. 69–92

Priest G, Klein B 1984 The selection of disputes for litigation. *Journal of Legal Studies* **13**: 1–23

Sartori G 1987 *The Theory of Democracy Revisited*. Chatam House, Chatam, NJ

Shapiro M 1981 *Courts: A Comparative and Political Analysis*. University of Chicago Press, Chicago

Trubek D, Sarat A, Felstiner W, Kritzer H, Grossman J 1983 The costs of ordinary litigation. *UCLA Law Review* **31**: 73–129

M. M. Feeley and C. Guarnieri

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Cr

Craft Production, Anthropology of

The familiar definition of 'craft' as a skill in making something, usually by hand, is essentially that used in anthropology. Political economists have distinguished between craftsmen who make things for use or sale to the consumer, and manufacture where things are made for merchants/capitalists to sell. Manufacture often involves several stages of production, either in the processing of raw materials (e.g., thread into cloth; weaving; dyeing) or of components (parts of a watch each made by a specialist craftsman; parts of a car). Before the industrial revolution the stages of manufacturing were coordinated and controlled by merchants or by the state, who had the necessary capital. While today manufacture refers to making things by machines, the literal meaning is 'hand-made' (from Latin *manus*, 'hand' and *facere*, *factum*, to make). Pre-industrial manufacture involved coordinating the stages of handcraft production, and was widespread in some societies with complex economies (India, China, medieval Europe). Production of goods by hand today is often coordinated by middlemen (see below). Thus it is useful to include in craft production both the making of goods by skilled craftsmen, and the coordinated manufacture of handcrafted goods.

1. Craft Specialization, the Division of Labor, and Commodization

The division of labor by gender and age has been basic in human societies from the simplest hunter-gatherers through pastoralism, simple and more complex forms of agriculture, to urban societies up until the industrial revolution, and remains important today. In small-scale societies where all adult men and women practice the same craft skills, these tend to be carried out as part of the domestic economy. Kin roles are also economic roles. In this 'domestic mode of production' (Sahlins 1972), craft products are made for both use and exchange. Craft skills are learned from older same-sex kin in the course of assisting with craft production. The developmental cycle of the domestic group reproduces the unit of craft production.

As societies become larger and more complex, emerging urban centers can support craftsmen who specialize in a single product (iron knives, silk cloth, pottery for cooking). Repeatedly we find that specialized production is associated with refinements in technology. With a specialist division of labor, craft products become commodities as they are exchanged for other commodities or enter the market. Hart writes that the central feature of increasing economic complexity is the integration of the division of labor in a nexus of commodity production and exchange; and that there is a tendency towards the progressive incorporation of ever more aspects of social life as commodities. 'The commodity is human labor embodied in a good or service offered to society rather than be consumed by its producer. As such, commodisation may be defined as the progressive abstraction of social labor' (Hart 1982).

As the nexus of production and exchange becomes more elaborate, new patterns of the division of labor appear. This pattern can be seen in the independent towns of medieval Europe and the Mediterranean. In Europe in the Middle Ages, textile production for an expanding market became organized as manufacture, first in specialist guild workshops, and then in rural cottage industry which took two forms. Cottage-craftsmen and women raised or bought the raw materials which they spun and wove into cloth, and sold locally to merchant buyers. The cottage-craftsmen owned raw materials, tools, and the finished cloth, planned their own work, and dealt directly with the market, although vulnerable to its fluctuations. In the other form of cottage industry, merchant manufacturers supplied the thread, often the loom, and owned the finished cloth. Weavers were paid for their work, in effect selling their labors as cottage-laborers. Rural cottage craft production organized by merchant capital is an example of what has been termed 'proto-industrialization.' With the harnessing of water- and then steam-power the invention of machines to replace handcrafts skills became possible. Such machines brought workers into factories, but the organization of components of production in manufacturing remained in many ways the same (machines now did the work of spinning, warping, weaving, finishing, dyeing) (Goody 1982).

Clearly, the industrial revolution has radically changed the conditions of craft production. It is no longer central to the production and exchange of commodities in industrial societies. Further, Western manufactured goods and industrial production have become integral to the economies of developing countries. Yet a number of niches of craft production remain. Social anthropology, grounded in participant observation, has addressed the topic of craft production during the twentieth century in several ways.

2. Theoretical Frameworks of the Anthropology of Craft Production

2.1 Material Culture and Folklore

The nineteenth century excitement with evolution was shared by anthropology, and carried over into the twentieth century with the Folklore school in which the study of material culture was central, especially in the United States and the Scandinavian countries. The focus was on detailed description and mapping of material culture, with a view of preserving a record of rapidly disappearing 'early' forms of society. Where craft production was treated it was within this framework.

2.2 Ethnographies in Social Anthropology

The ethnographies of this tradition took as problematic the social structure and functional integration of societies studied. Ethnographic accounts treat religion, kinship, economics, and political dynamics as all important and related. However social anthropologists differed in the weighting given to each domain, as did the societies they studied. Significantly, those ethnographers who treat craft production as a significant feature tend to have worked in relatively complex societies. Among them, Nadel's account of the kingdom of Nupe (Nigeria) is pre-eminent for its detailed accounts of a number of craft industries, and the way these relate to the Nupe court. Each craft is discussed in terms of the access to raw materials, technical processes, the organization of the division of labor within the workshop, authority and transmission of skills among craftsmen, design and fashions as these relate to consumer demand, and patronage by nobles and the court where this is relevant. Some crafts are 'open' in the sense that any man can learn by assisting a skilled worker (weaving, bead making); others are 'closed,' restricted to members of certain kin groups (blacksmiths, brass smiths). Craft skills are learned in informal apprenticeship. Closed crafts restrict apprenticeship to kin; this also serves to keep ritual and secrete knowledge within the craft kindered (Nadel 1942). Miner's account of craft production in Timbuctoo in the same period is less detailed, but closely parallels that for Nupe (Miner 1965 [1953]). Slave production of food stuffs and cotton was important, particularly in northern Nigeria, but does not appear to have been organized for craft production. However rural industry of the cottage craftsman kind in weaving and dyeing is documented around the city of Kano where certain villages were famous as centers of dyeing. The town of Daboya played a similar role as a center of dyeing in the precolonial textile industry of northern Ghana. (See also chapters by Pokrant and by Goody, in Goody 1982.)

2.3 Marxist Studies of Craft Production

With its focus on modes of production, Marxist theory offers social anthropology a framework for the study of craft production. A mode of production consists of the productive forces (natural, material, and technical) and the social relations of production. Most work in this area is by French anthropologists, exemplified by an English collection of papers by French neo-Marxists (Seddon 1978). The goal of this volume is to lay the groundwork for 'a truly scientific study of society … which will lead to the discovery of laws that underlie the development of … social and economic formations … (and) identify the relations of production characteristics of pre-capitalist formations' (Seddon 1978, p. ix). There are chapters about self-sustaining agricultural societies, the peasant mode of production, the Asian mode of production, and an African mode of production. Based on detailed ethnography, most are rich and insightful discussions of the relations between economic and political factors, several foregrounding kinship in new ways; this is especially true of those in the section based on fieldwork in Africa. However, perhaps because their goal is the identification of general laws, craft production is not a focus of description or analysis.

Following the Seddon volume a number of anthropologists have written ethnographic studies of craft production from a Marxist perspective. Three can be noted briefly: Kahn's *Minangkabau Social Formations* (1980) is a study of village blacksmiths in Sumatra. He describes in detail technology, organization, and marketing—treating relations of production within the workshop, the pattern of learning craft skills, relations between blacksmiths, external influences such as imported materials, and tools. The analysis is set in the context of Minangkabau matriliny, a subsistence agriculture based on rice, and economic policies of colonial overlordship. Based on individual or small workshop production for a local market or to traders, this is petty commodity production.

The next example is a study of Scottish Hebridean Island weavers of Harris tweed. Ennew (1982) traces the development of this industry from the first sales of Harris tweed in 1881. From the beginning it was based on cottage-laborer production, with merchants supplying the wool, dyes, and often the loom, and deducting the cost from payment for the finished cloth. Demands from the external market (London, overseas markets) led first to the introduction of a wider, foot-powered loom, then to removal of warping to the spinning mills in order to control quality and patterns. 'From the moment of wool-purchase the production process of Harris Tweed is entirely in the control of mill-owning capital, except for the slight loss of control during the outworking [weaving] phase' (Ennew 1982, pp.189–90).

Very different relations of production characterize the manufacture of carved hardwood furniture in

Hong Kong. In the factory which Cooper joined as an apprentice wood carver, identical sets of tables were carved for tourists and export. He speaks of the proletarianization of the industry. Workers were paid daily at an hourly rate; the factory products were standardized, carvers producing endless legs and tops for tables, with no involvement with design or contact with customers. However he does feel that seven months as a factory carver gave him insight into the Maoist distinction between manual labor and mental labor, into proletarian culture as well as ideology (Cooper, in Coy 1989).

It seems that craft production is incorporated in several modes of production: petty commodity, cottage laborer, and even as proletarian labor.

2.4 The Rebirth of Material Culture Studies: The Cultural Meanings of Commodities

Material culture has again become a central concern for anthropology, but now the focus is on the cultural and personal meanings of things as commodities. Objects serve to objectify cultural meanings in individual experience; individual identities are constructed through the consumption of commodities. Indeed it has been argued that the study of consumption and commodities represent a major transformation in the discipline of anthropology and may come to replace kinship as the core of anthropology. Appadurai, in his Introduction to *The Social Life of Things: Commodities in Cultural Perspective* (1986) reviews economic theories which focus on the production of commodities and argues, following Douglas and Isherwood, that the interesting questions are rather about why and how people consume goods. The factors that drive consumption (the 'demand' of classical economics) are to be found in socially and historically shaped cultural meanings—in the political economy of the social life of things. The Introduction concludes: 'Ever since Marx and the early political economists, there has not been much mystery about the relationship between politics and production. We are now in a better position to demystify the demand side of economic life' (Appadurai 1986, p. 58). Given this perspective it is not surprising that among these anthropologists the few who are interested in production deplore the scarcity of relevant studies (but see chapters by Spooner and Bayly in Appadurai 1986).

2.5 Learning and Apprenticeship in Craft Production

The learning of craft skills has consistently been central to craft production itself—with the domestic mode of craft production in simple societies, in apprenticeships in craft workshops, and again in the domestic mode of cottage industry production. Craft production is labor intensive, and the work of unskilled children and youths frees craftsmen to concentrate on more complex skilled tasks, assembly of components, and organization of both production and distribution to consumers. A Vygotskian view sees learning as based on working alongside an adept: observing, assisting with very simple task elements, gradually taking over responsibility for these elements while continuing to observe, and then assisting with more complex tasks. This model corresponds closely to the learning of weaving by fostered apprentices in the Daboya textile industry (Goody, in Goody 1982). It also appears to fit accounts of domestic craft production by blacksmiths (Nadel 1942, Kahn 1980, Coy 1989); it is consistent with formal apprenticeship as in industrial wood carvers (Cooper, in Coy 1989) and Japanese folk potters (Singleton, in Coy 1989). The contributors to Coy's edited volume on *Apprenticeship* (1989) are all anthropologists who did fieldwork while themselves apprenticed, and several learned handcrafts. Their work supports the view that in craft production the learner is central to the production process—by providing unskilled labor in labor-intensive production; by systematically enforced subordinate status which legitimizes labor demands and supports the expert's control of synchronizing elements of production; and by embodying the social reproduction of craft skills and knowledge.

3. Trajectories of Craft Production in a Global Economy

Before industrialization, virtually all artifacts—for daily use, exchange, luxuries—and the tools for making these, were the products of craft skills. With the industrial revolution, craft manufacture ('hand making') of most commodities has been replaced by machine production in factories. Yet today craft production still occurs in the many niches left by industrialization, sometimes in surprising places.

There are many craft production niches in both industrial and developing societies; this is an open set. The conditions under which handcraft production has advantages (of cost, of detail work, of skilled knowledge) are varied, and change in response to both local and global factors. Social anthropologists have studied craft production in some of these niches; other awaits research. Contemporary anthropology is responding to the complexity of globalization and postmodern skepticism of both theory and observation with a proliferation of particular interests. While none of the major current themes focuses on production, there are good individual studies in several of these niches. It is only possible here to sketch these briefly.

3.1 Niches of Contemporary Craft Production in 'Developing' Societies

3.1.1 Artisan/craft production for local use in developing societies. In partially industrialized third world

societies many commodities for local use are made by individual craftsmen and workshops, and in small factories, using handcraft tools and techniques. In Bolivia, tinsmiths make pails and basins, blacksmiths make farm implements and locks, others make kerosene lanterns, dancing masks, clothing, shoes, etc. (Buechler, in Coy 1989). The largest single example must be the Indian Khadi cottage industry making handspun, hand woven cloth. Beginning in the ideology of self-sufficiency with which Gandhi led the movement for independence, this continues to be centrally organized. As with other attempts to make household craft production economically viable alongside industrial production, partially mechanized processes have been introduced to support handcraft skills (e.g., multiple-spindle spinning jennies are wired to electricity by villagers, Bayly, in Appadurai 1986).

3.1.2 Artisan/craft production in developing countries for the global market. There are two links to this market, tourists buying locally, and international buyers for stores in industrialized countries. The two are related, since tourists want to buy similar craft products on returning home, and consumers of craft producers in industrialized countries look for them as tourists. Social anthropologists have recently become interested in tourism. Among the studies which treat craft production are several in Graburn's *Ethnic and tourist arts* (1976); see especially the chapters by Graburn (on Inuit soapstone carving), Kent (Puebio and Navajo weaving in the American Southwest), and Salvador (on Mola applique blouses made by Cuna Indians).

Handwoven and hand-printed textiles have for centuries been among India's most successful export industries. In the twentieth century some firms concentrating on garment manufacture attempted to secure advantages of efficiency and control through factory production. However, there is a pattern of reverting to commissioning both weaving and making up of garments from smaller units—cottage weavers and small workshops. Swallow's study of one such company finds that in a volatile export market the costs of factory production (maintaining plant and machinery, labor troubles) remain higher than those of cottage craftsmen and cottage laborers managed by intermediate capitalist merchants, under conditions of population pressure and lack of alternative unemployment (Swallow, in Goody 1982).

An increasingly common pattern is merchant or factory production of crafts sold locally to Western buyers for retail in industrial countries. Best studied is the production and marketing of luxury goods such as oriental rugs/carpets (see Spooner in Appadurai 1986) and carved hardwood furniture (Cooper 1980). Less elaborate, cheaper craft products (simple jewelry, pottery, glass, textiles) are made by cottage craftsmen, cottage laborers, or in workshops for sale to wholesale buyers who market them in bulk to retailers at trade fairs in industrialized countries. These are finally sold to consumers in Western gift shops, department stores, and clothing stores (buyer, personal communication). Although this mode of craft production for final sale in the commodity-rich markets of the West probably far exceeds in volume that of luxury goods, it seems as yet not to have been studied by anthropologists.

3.1.3 Development aid and craft production. International development agencies, nongovernmental organizations (NGOs), and national government agencies have seen craft production as a potential focus for income-generating projects, usually for women. This may mean building on existing craft skills; often crafts judged as marketable are unfamiliar locally, and have to be taught. Project effort is directed towards teaching skills, securing access to materials and tools, and marketing. The development of some form of local cooperative structure is often integral to the project. Although anthropologists often participate, little published material is available on such projects, except in the form of reports and evaluations.

3.2 Niches of Craft Production in Industrialized Societies

3.2.1 Outwork and workshop handcrafts in industrial societies. In industrialized societies craftsmen—here usually craftswomen—continue to work at home at piecework rates on a number of crafts: sewing on industrial machines (owned or rented), knitting (often by machine), and the making of small cheap goods like Christmas crackers. This modern cottage industry is cheaper for manufacturers for the familiar reasons: low rates of pay; a dependent labor force lacking employment alternatives; no need to maintain plant or machines; flexibility of labor force which is only given work when demand is high. Some manufacturers, particularly in textiles, bring craftspeople into workshops where sewing and knitting machines are provided and supervision is possible, making changes in design easier.

3.2.2 Production of folk crafts for the national market. The 'Arts and Crafts movement' associated with William Morris in England in the second half of the nineteenth century had counterparts in the United States, and later Japan. This movement reacted to industrial production by romanticizing craft production, seeing it as preserving old techniques and a classical vision of beauty expressed in individual hand skills. In order to manufacture craft products

so as to make them accessible to an enlightened market, Morris taught himself to weave, embroider, print textiles, make tied carpets, make glazed tiles, and block print designs used for wallpaper. Apprenticeship was initially rejected, since they were seeking a pure re-creation of the essence of each craft. Although it was Morris' dream to make craft products cheaply so all could afford them, the cost of production with skilled labor and costly materials meant they were always luxury goods.

This vision survived. It subsequently led to handcraft manufacturers using traditional patterns and materials but of relatively modest cost: of textiles (Harris tweed), pottery (exemplified by the craft of Bernard Leach), and in the United States of Quaker furniture. In Japan it was reflected in the deliberate creation of a word of 'folkcraft' (*mingei*) after discussions among artists and philosophers (including Bernard Leach), and the establishing of a national museum of folkcraft which worked especially with traditional potters to develop aesthetic and practical adaptations of old forms. This folk pottery was made in traditional rural workshops, and became commercially popular. In his study of several rural Japanese pottery-making communities, Singleton found that high demand led in some to an opening of apprenticeship to outsiders, with a decline in the authority of the master (Singleton, in Coy 1989).

3.2.3 Persistence of customized craft production. Factory production for mass consumption has become the major source of commodities in industrialized societies. Customised handcraft production continues alongside this, especially of men's clothing, women's fashions, shoes, and jewelry. Women's fashions have become international, with a few famous designers in Paris, Rome, New York, and Tokyo holding annual shows which set the latest fashion. At the luxury end of this mode of production the craftsmen designs a prototype made in his workshop; this may be commissioned by a customer, offered for sale through a show, or in exclusive shops. Successful designs may be copied in other workshops for sale in select stores; such as garment with a 'designer label' is still considered exclusive. A few designs are more widely copied and sold through chains of stores to the public. This 'trickle-down' of fashions through dressmakers' workshops has always linked dress to social mobility. Lace, from the sixteenth to the nineteenth centuries, moved from courts to town gentry to village bonnets. Maher describes this for Turin in the early twentieth century, where seamstresses copied the dresses made for customers to wear themselves—thus passing socially as ladies (Maher 1987). Customized craft production also continues to be important, for instance in fine furniture, jewelry, pottery, and textiles. These are sold in specialist shops, and increasingly in shops alongside replicas of museum holdings. Their makers are coming to be known as designer craftsmen,' and also as 'artist craftsmen.' There is a debate, as with some of the ethnic crafts, as to whether this is 'art' or 'craft.'

3.2.4 Individual craft production for pleasure in using craft skill. The anthropological preoccupation with conceptualization of identity in terms of commodities has displaced earlier anthropological interest in occupational and personal identity based on skilled craftsmanship. Yet there is a resurgence of popular interest in learning craft skills like sewing, knitting, embroidery, lace-making, weaving, pottery, jewelry-making, etc. A number of indices reflect this enthusiasm: where they used to be learned from kin or in apprenticeship, such crafts are now taught in community centers and in summer workshops; practitioners form associations concerned with style and pattern, access to tools and materials, teaching and examining craft skills; there are magazines and books devoted to patterns and techniques; practitioners hold exhibitions of their work. The main motivation seems to be development and exercise of individual handcraft skills. Products are used by the maker or given as gifts. While occasionally a seamstress, weaver, or potter becomes skilled enough to sell skills or products on the market, for most commodization is not a factor. Indeed, individual craft production may be seen as the separation of craft skills from the market.

See also: Apprenticeship: Anthropological Aspects; Economic Anthropology; Economic Sociology; Labor History; Labor, Division of; Mode of Production; Tourism; Work and Labor: History of the Concept; Work: Anthropological Aspects; Work, History of

Bibliography

Appadurai A (ed.) 1986 *The Social Life of Things: Commodities in Cultural Perspective*. Cambridge University Press, Cambridge, UK

Cooper E 1980 *The Wood-carvers of Hong Kong: Craft Production in the World Capitalist Periphery*. Cambridge University Press, Cambridge, UK

Coy M 1989 *Apprenticeship: From Theory to Method and Back Again*. State University of New York Press, Albany, NY

Ennew J 1982 Harris Tweed: Construction, retention and representation of a cottage industry. In: Goody E N (ed.) *From Craft to Industry*. Cambridge University Press, Cambridge, UK

Goody E N (ed.) 1982 *From Craft to Industry: The Ethnography of Proto-Industrial Cloth Production*. Cambridge University Press, Cambridge, UK

Graburn N H H (ed.) 1976 *Ethnic and Tourist Arts: Cultural Expressions from the Fourth World*. University of California Press, Berkeley, CA

Hart K 1982 On commodisation. In: Goody E N (ed.) *From Craft to Industry*. Cambridge University Press, Cambridge, UK
Kahn J 1980 *Minangkabau Social Formations*. Cambridge University Press, Cambridge, UK
Maher V 1987 Sewing the seams of society: Dressmakers and seamstresses in Turin between the wars. In: Collier J, Yanagisako S, Block M (eds.) *Gender and Kinship: Essays Toward a Unified Analysis*. Stanford University Press, Stanford, CA
Miner H 1965 [1953] *The Primitive City of Timbuctoo*. Anchor Books, New York
Nadel S F 1942 *A Black Byzantium*. Oxford University Press, London
Sahlins M 1972 *Stone Age Economics*. Akline, Chicago
Seddon D (ed.) 1978 *Relations of Production: Marxist Approaches to Economic Anthropology*. Frank Cass, London

E. N. Goody

Creationism, Evolutionism, and Antievolutionism

1. Creationism, Evolution, and Antievolutionism

1.1 Creationism

Broadly defined, *creationism* refers to the idea that a supernatural agent or agents created the universe, living things, and/or human beings. To Leibnitz's question, 'Why is there something rather than nothing?' most human societies answer with creation stories, wherein gods, culture heroes, or other powerful forces bring about the Universe from nothing, or modify some existing matter to produce the Earth and living things. More narrowly defined, in North America, creationism has come to connote the theological doctrine of special creationism, a particular kind of creationism in which God created the Universe—galaxies, stars, the solar system, Earth, living things, humans—all in a brief time in its present form. 'Young Earth creationism' is the most frequently encountered form of special creationism; its proponents believe that the creation event took place during six 24-hour days, a relatively short time ago (within 10,000 years). This reflects a literal interpretation of Genesis, the first book of the Bible, and although most common in Christianity, special creationism is part of the theology (but thought not to be a central part) of the other Middle Eastern monotheisms, Judaism and Islam.

1.2 Evolution

Evolution is a scientific view of the Universe as having reached its present form gradually over time. Evolution, then, is ultimately about history: galaxies, stars, the planet Earth, and living things are different today from what they were in the past; there has been change through time. Biological evolution, a subset of this larger view of change through time, is the theory that living things have descended with modification from common ancestors. As with all biological entities, human beings have a shared common ancestry with other living things; our closest nonhuman relatives are the living apes. The ape/human common ancestor shared common ancestors with monkeys and other primates, other mammals, and then (more distantly) with reptiles, amphibians, fish, and invertebrates. Biological evolution affirms a genealogical relationship of species through time. It thus contrasts sharply with special creationism, which requires that all living things be created in their present forms, rather than having descended with modification from common ancestors.

Religious views other than special creationism may be, and often are, compatible with evolution. Theistic evolution, the doctrine that God worked purposefully through natural law to bring about the Universe via evolution, is part of the accepted theology of the Catholic Church, mainline Protestant Christian churches, and Reformed and Orthodox Judaism. It is also acceptable in some interpretations of the Koran. In these theologies, Genesis is not interpreted literally, and the separate creation of 'kinds' of living things is not important. The focus is on God as Creator, rather than *how* God created.

1.3 Antievolutionism

Although evolution is widely accepted within the scientific community, evolution has come under attack in North America, and in a growing number of other countries, by holders of certain religious views. In North America, the controversy plays primarily in the arena of precollege public-school education, although antievolutionism is not uncommonly expressed in informal science education centers such as National Parks, museums, zoos, and science centers.

Young-Earth Protestant Christians following a biblical literalist religious tradition form the core of the antievolutionism movement, but in the USA at least, antievolutionism extends beyond biblical literalists to include some mainline Christians. These nonliteralist Christian antievolutionists are uneasy about evolution because it appears to them to have serious existential consequences (Pennock 1999). They reject theistic evolution in favor of special creationism because they believe that if humans evolved through the same process that brought about oak trees and salamanders, then we would be less special in the eyes of God. Being less special to God implies to them that there is no special purpose for human beings, and thus

that human life has no more meaning than that of any other creature on the planet.

For life to be a purposeless and meaningless —, and God to be at best impersonal, rather than directly concerned about and accessible to individuals, — is unacceptable. Forced to choose between creationism and existential fulfillment on the one hand, and existential emptiness and evolution on the other, they reject evolution. Other Christian traditions resolve these issues in a variety of ways, finding no theological problem with God's working through natural law to produce humans, but still sustaining the Universe and interacting with humans in a meaningful way.

It can fairly be said that antievolutionism does not flow from scientific weaknesses in evolutionary data and theory, but from religious or philosophical objections. This is illustrated by recent secular 'postmodernist' critiques of evolution, which focus on alleged negative implications for society of Darwinian evolution (evolution through Darwin's mechanism of natural selection). Much as did members of the progressivist movement at the beginning of the twentieth century, postmodernist antievolutionists criticize Darwinism as the source of racism, sexism, exploitation of workers, and other social evils. Ironically, conservative Christians criticize evolution for being the source of Marxism, while leftist postmodernists criticize evolution for being the source of 'nature red in tooth and claw' capitalism!

2. *History*

Antievolutionism has been most successful in the United States, for historical as well as cultural reasons. The United States was settled in part by religious dissidents, whose churches were strongly congregational rather than hierarchical in structure. The period since the early 1800s has seen the origin of a variety of unique American Christian theological traditions, and although many (Shakers, Millerites) have passed into history, some have continued into the twenty-first century (Mormons, Christian Scientists, Seventh Day Adventists). Decentralization also characterizes the American educational tradition, because isolated settlements were forced to create their own schools, hire their own teachers, and develop their own curricula. The United States has never had a national curriculum in science, or in any other field, and curricula may vary considerably from district to district even within a single state. Evolution may or may not be taught in a given school district, even if it is required in the district's curriculum. Teachers often omit evolution from instruction because it is controversial in their communities.

A major historical influence on American antievolutionism was the development during the second decade of the twentieth century of fundamentalism, a uniquely American Protestant religious tradition. Written between 1910 and 1915, the 12 pamphlets known as 'The Fundamentals' laid out a back-to-basics Protestant theology that relied heavily on the inerrancy of the Bible. Emerging during the unsettled social and political conditions preceding World War I, fundamentalism's comforting tenets were accepted rapidly by large numbers of Americans. Biblical literalism, however, had not been part of British Anglican, or continental Lutheran or Catholic traditions, and fundamentalism remained an American phenomenon.

By the beginning of the twentieth century, the theory of evolution had become widely accepted by scientists both in Great Britain and in the USA (Moore, 1979). In Great Britain, the Anglican church adjusted relatively quickly to evolution, accepting within a few decades of the publication of Darwin's *On the Origin of Species* that God created through the process of evolution. Textbooks in the USA at that time routinely discussed evolution as the organizing principle of biology. But the rise of fundamentalism was to change that, because evolution contradicted a literal interpretation of Genesis. Because Darwinism was associated in the public mind with *social* Darwinism, Christian social progressives such as William Jennings Bryan opposed it as they sought to ban child labor, improve the lot of workers, provide for women's rights, and promote other progressive causes. By the early 1920s, efforts were being made in many states to ban outright the teaching of evolution.

2.1 *Banning Evolution*

Although between 1921 and 1929 antievolution legislation was introduced into 37 states, only five states passed such laws: Oklahoma, Tennessee, Florida, Mississippi, and Arkansas. Tennessee's law was challenged by a high-school teacher named John Scopes, which culminated in the famous Scopes trial of 1925. Scopes lost, and the Tennessee law remained on the books for over 40 years.

Even though during the Scopes Trial the popular press pilloried the creationist view as old-fashioned and ignorant, antievolutionism did not go out of favor. In fact, the Mississippi and Arkansas antievolution laws were passed *after* the Scopes trial. The coverage of evolution in textbooks declined sharply after the Scopes trial, until by the mid-1930s, hardly any textbooks included the topic. Antievolutionism remained triumphant until the late 1950s, when the Soviet Union's launch of the first orbiting satellite, Sputnik, caused the federal government to reexamine its science policy and to begin a major science education reform movement. This movement included the production of textbooks written by scientists and master teachers rather than in-house writers in publishing companies.

The federally-funded Biological Sciences Curriculum Study (BSCS) cited evolution as an organizing principle in its textbooks, and soon commercial publishers followed suit. By the late 1960s, evolution was once again prominent in high-school textbooks. In 1968, antievolution laws such as the Tennessee law under which Scopes had been tried were overthrown by the Supreme Court in *Epperson vs. Arkansas*.

2.2 Creation Science

The resurgence of evolution in textbooks caused an antievolutionary backlash. A strategy developed: if creationism could be presented as an alternative scientific view, then it would deserve a place in the curriculum. 'Creation science' first appeared in developed form in 1961 in *The Genesis Flood* by John Whitcomb and Henry R. Morris, in which the authors argued that modern geological features could be explained by Noah's flood. Its mix of theology and science characterized future creation science efforts, and Morris became the foremost proponent of creation science in the latter part of the twentieth century.

Creation science is associated closely with 'young Earth creationism,' in which the six days of Genesis took place about 10,000 years ago. Most creation science literature focuses upon supposed flaws in evolution, the reasoning being that evidence against evolution is evidence for creationism.

In 1972, Henry Morris and others founded the Institute for Creation Research in San Diego, CA, as an antievolution ministry. It remains the flagship antievolution ministry, but there are other national organizations such as Answers in Genesis, the Bible-Science Association (recently renamed Creation Moments, Inc.), and several regional and local ministries that also promote the young Earth creationist views of Henry Morris. A sizable corpus of antievolutionary material consisting of books, videos, compact discs, filmstrips, tape recordings, posters, and curricula is publicly available from these organizations.

By the late 1970s, creationists instituted a movement to propose 'equal time' legislation in state legislatures, requiring that if evolution were taught, creation science would also have to be taught. At least 23 states had equal-time legislation introduced, and two states passed these bills: first Arkansas and then Louisiana. Although both laws were challenged in the courts, only the Louisiana decision reached the Supreme Court, which ruled in 1987 in *Edwards vs. Aguillard* that equal-time laws violate the First Amendment of the US Constitution, which requires that public institutions be religiously neutral. Creationism, being inherently a religious view, cannot be advocated in the public schools; thus laws promoting equal time for creationism are unconstitutional. Efforts to pass such legislation largely ceased, and school districts around the country that had required the teaching of creation science quietly changed their regulations.

2.3 Neocreationism

Although public-school teaching of creation science itself has been ruled unconstitutional, antievolutionism continues in the United States. More recent antievolutionary strategies avoid reference to any form of the root word 'creation' and go directly to trying to discourage the teaching of evolution without banning it directly. Whereas in the 1980s and early 1990s, school districts and teachers were exhorted to 'teach evolution and teach creation science to balance it out,' now antievolutionists encourage them to 'teach evolution and the "evidence against evolution" to balance it out.' In some districts, teachers are required to read a disclaimer that evolution is 'only a theory' to students before teaching the subject, and in fact, 'only a theory' disclaimer stickers have been placed in textbooks in Alabama and in several school districts around the country.

During the mid-1990s, a movement to 'teach evolution and "intelligent design theory" to balance it out' appeared and began to spread. 'Intelligent design theory' is a restatement of William Paley's 1802 argument from design, an apology for the existence of God. Paley argued that because complex, intricate structures such as the vertebrate eye could not have occurred naturally, they must have been created specially by God. Intelligent design creationists update Paley by claiming that the structure of DNA or functions of the cell are 'too complex' to have been brought about by the alleged 'random' process of natural selection.

3. International Antievolutionism

Antievolutionism has been strongest in the United States, but there are movements in several other countries as well. American creation science materials have been circulated in Canada, Australia, and New Zealand, and as translations into languages other than English have become available, into dozens of other countries including Korea, the Philippines, Turkey, the nations of Europe, and in the former Soviet Union. Antievolutionism has not taken root in Europe, probably because secular culture predominates there, but in the current disarray in which Russia and other former components of the Soviet Union find themselves, antievolutionism is increasing in proportion to the increase of adherents to conservative Christian sects.

International antievolutionism is largely the result of activities of American conservative Protestant Christian missionaries. The withdrawal of official

Catholic approval for 'liberation theology' has caused many in Central and South America to search for a religious alternative to Catholicism, and Protestant missionaries have stepped up their proselytizing efforts in this part of the world. Missionaries distribute antievolution materials as well as more traditional tracts. In Russia, Protestant missionaries have long been teachers of English, and with the breakdown of the central government's oversight (and the increase in interest in religion since the dissolution of the Soviet Union), antievolution materials have proliferated. The Institute for Creation Research has cultivated antievolutionists in Russia as well as many other nations, in order to spread creation science as well as biblical literalist Christianity.

Of interest is the growth of non-Christian antievolutionism, as seen in a small but vocal movement in Islamic Turkey, as well as in Israel from ultra-Orthodox Jews. In both cases, the literal interpretation of Genesis is foundational to the opposition to the teaching of evolution in schools. Of the two, the Turkish movement is more organized, and uses translations of American creation science materials. In fact, ICR employees have made several trips to Turkey to cultivate the antievolution movement, apparently without promoting Christian conversion.

One of the stranger chapters in modern antievolutionism is the development of a form of creationism based on Hinduism. A Krishna Consciousness organization in the United States, the Baktividanta Institute, contends that modern evolutionary science is not true. But instead of substituting creation science, Institute personnel argue for the literal truth of the Hindu Vedas. According to these religious scriptures, the world and living things have come and gone according to a series of cycles, and we are currently the midst of one. Humans were created 350 million years ago, and did not evolve only within the last few million years. Ironically, in a movie shown on American television during the 1990s ('Mysterious Origins of Man'), antievolution creation science arguments were used, but stripped of their biblical references, and also of their young-earth references. Like creation science, 'Krishna creationism' claims that modern scientists are in a conspiracy to hide the 'truth' of the lack of scientific support for evolution from the public.

4. *The Future*

Will antievolutionism increase or decrease over time? The future of social movements is always difficult to predict, but no predictions are possible without an understanding of the roots and motivations of such movements. Clearly, the foundation of antievolutionism is the concern (if not fear) that religious views will be proven untrue, or at least require major modification. This is most apparent with biblical literalists, who tie salvation to the literal truth of the book of Revelations, and tie Revelations to the literal truth of Genesis. But, as mentioned earlier, nonliteralist Christians may also be uneasy about evolution if evolution is seen to imply or require that only natural processes could be involved in the production of the Universe—that God is completely removed from Creation. Beliefs about such existential issues will be very influential in either the increase or decrease of antievolutionism.

Antievolutionism is correlated with conservative religious belief. Although sacred texts such as the Bible have been interpreted as making claims about the natural world, as science has grown and proven its worth, most varieties of Judaism and Christianity have gradually ceased to attempt to explain the natural world through revelation.

Even though there are religious traditions within which evolution is not acceptable, it is probable that theological change will continue. There are only a handful of Christians who believe that the Bible incorporates geocentrism, and perhaps on some future day, all Christians will look back on the era of religious opposition to evolution much as they do upon Galileo's conflict with the Church of his time: as an example of the gradual replacement of theological statements about the natural world with scientific data and theory.

Not surprisingly, polls show high acceptance of evolution in nations where religious belief is low. In a world of nonbelievers, one would find little antievolutionism. But faith does not seem to be on the decline. Secularism is more common in economically developed nations than in less developed countries. The United States, in fact, is atypical among developed countries for its high percentage of individuals who profess belief in God. But even with increased prosperity among developing nations, one would be unwise to predict a linear increase in secularism. Adherence to religion waxes and wanes through time, depending on a network of factors affecting the social and economic environment.

In Russia under the communists, for example, there was reportedly a very low level of religious adherence. Of course, there were reduced opportunities to practice religion in a state where religion was actively discouraged. But with the collapse of the communist system and government, more churches and synagogues are opening in Russia and more people are attending services at the start of the twenty-first century than for decades previously. Reportedly, the Chinese communist government has not been successful in fully stamping out religious adherence. It would be premature indeed to generalize the secularization seen in developed nations in Europe to other nations: traditional religious beliefs are very resilient.

So a decrease in support for antievolutionism is most likely to occur, if it occurs, because of changes in current religious views now hostile to evolution, rather than because of large numbers of people's abandoning

religion. Antievolutionism may in fact *increase* in the future, if there is increased adherence to traditions requiring the literal interpretation of sacred texts (the Bible, the Koran, the Torah, the Vedas, and perhaps others).

Antievolutionism is also related to the public understanding of science: people who understand evolution as a scientific idea that explains natural phenomena rather than as an inherently antireligious idea are less likely to reject it. Thus, if there is improvement in the public understanding of science, antievolutionism may decrease. An improvement in science education in countries in which antievolutionism is established or growing may reduce its effect.

See also: Darwin, Charles Robert (1809–82); Darwinism; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionary Epistemology; Evolutionism, Including Social Darwinism; Myth in Religion; Religion and Science; Religion, History of; Science and Religion; Secular Religions

Bibliography

Larson E J 1985 *Trial and Error*. Oxford University Press, New York

Montagu A (ed.) 1984 *Science and Creationism*. Oxford University Press, New York

Numbers R L 1992 *The Creationists*, 1st edn. Knopf, New York

Pennock R T 1999 *Tower of Babel: The Evidence Against the New Creationism*. MIT Press, Cambridge, MA

Strickberger M W 2000 *Evolution*, 3rd edn. Jones and Bartlett, Sudbury, MA

E. C. Scott

Creativity and Cognition

Cognitive theories of creativity focus on the intellectual structures and processes that lead to insights, solutions, and ideas that are original and useful. Creative work may involve more than cognition, but certain cognitive processes may be necessary for virtually all creativity. Moreover, some of the extracognitive contributions to creative behavior (e.g., the motivation to identify and develop a creative idea) depend on cognition. With this in mind it is not surprising that many theories of creativity focus on cognitive processes. These describe creative thinking and problem solving in terms of divergent thinking, analogical thinking, metacognition (especially strategy), lateral thinking, associative thinking, or simply as a special kind of information organization and reorganization. This article briefly describes the various cognitive underpinnings of creative thinking. It also describes the major issues and questions, including the one already mentioned about the interdependence of cognition and motivation.

1. Information Processing and Perception

In the most general terms cognition involves the storage and processing of information. Often information processing is described with three stores (or memory banks) and three transfer processes. The first store is the sensory buffer, which hold raw information which has not yet been recognized or interpreted. Next is the short-term memory (STM), and last the long-term memory (LTM). Attention is the processes by which information is transferred from the sensory buffer to short-term memory. Rehearsal (in various forms) transfers information from STM to LTM. Retrieval transfers information back to STM from the LTM.

Different theories of creative cognition focus on different components or stages of thought. Creative cognition may occur very early in the process. Creative persons are, for example, often somewhat sensitive; sensitivity is one of the 'core characteristics' that personality researchers have found in many creative samples. This sensitivity may lead individuals to certain experiences, or more accurately, make them notice subtle stimuli and associations which others will overlook. In this sense sensitivity can dramatically influence the perceptual process in such a way as to make creative thinking more likely.

In strict cognitive terms, perception occurs when sensory information (based on experience and the environment) is interpreted. Perception always involves an interpretation of information; and as the example above suggests, interpretations are often significantly influenced by the individual's background, expectations, and personality. Creative persons may very well have certain perceptual and interpretive tendencies which lead them to original interpretations of experience. Other components of the cognitive process are also tied to creative thinking.

Attention, for example, plays an important role. This is because information is not processed unless the individual attends to it. By attending to information, we focus our limited cognitive resources, selecting some information and ignoring other information. Creative cognition may benefit from such selections. It is not, however, a matter of being especially focused. In fact, there are reports of creative persons having 'wide attention deployment strategies.' This means that they attend to a wider range of information and are not as selective as less creative persons. They probably have more information available to them at any one time, and more diverse information as well. Another way of describing this tendency is as 'defocused attention.' Again the idea is that there is less

selectivity and a broader range of available information. A broad range of associations may also result; the information may be environmental and experiential, or it may be internal and drawn from memory.

2. *Memory, Information, and Expertise*

The various memory stores hold information. Information in turn is frequently required of creative cognition. Some creative achievements, for example, require expertise, which can be loosely defined in terms of information or knowledge. Referring back to the model of cognition used earlier in this article, knowledge is information stored in long-term memory. Expertise is usually domain-specific; the individual has expertise within one field or another. It takes time to develop, though some fields take less time than others. Physics represents a highly technical field and expertise in that field requires years of effort; musical creativity often appears much earlier precisely because there is less technical information required.

An efficient memory, containing a great deal of information, does not always facilitate creative cognition. Information, knowledge, and expertise can benefit the individual but sometimes actually inhibit creative thinking. Sometimes expertise leads the individual to make assumptions, and assumptions minimize the mindful thought that is typically required for original problem solving. Expertise also often provides the individual with routines; but here again these can preclude the flexible perspective and questioning attitude that more often support creative thinking. The trick is to be well-informed but at the same time to avoid assumption and fixed routine.

3. *Metacognition*

Routine and assumption are easy to avoid. They follow naturally from expertise, and they often allow the individual to behave in an efficient manner. They will not inhibit creative cognition if they are used mindfully. This means that individuals must monitor their thinking and know when they are indeed relying on assumptions and routines. When original thinking is desired, assumptions should be questioned and routines broken.

Such mindful use of assumption and routine is relatively easy. This is because individuals develop metacognitive skills early in adolescence. Metacognition is thinking about thinking, or cognition about cognition. It is used whenever individuals reflect on their own cognition. It is especially useful for the kind of monitoring mentioned just above, and even more so for the direction and guidance of one's own thinking. Metacognition allows the individual to be strategic and tactical, as well as mindful.

A number of tactics are particularly useful for creative thinking. Some of these are metaphorical, such as the tactic for 'lateral thinking,' which asks individuals to 'dig elsewhere, not deeper.' Other tactics are more direct and quite simple, including two given above: 'question assumptions' and 'avoid routine' (when original thinking is desired). If the individual is having difficulty with a particular problem, for instance, he or she might ask, 'am I making inappropriate assumptions?' Other tactics include working backwards on a problem (begin with the objective and then consider what state must immediate precede it, and what must precede that, and so on) or changing one's perspective. Original ideas very frequently result from a change of perspective, whether it results from discussion with other people, travel, a change in the way the problem is represented, or from time away (incubation).

Consider the Wright brothers. They were extremely tactical in their work. They broke the large problem of flight down into small problems, for example (power, control, weight). They used analogies and looked to nature (studying birds during flight). They collected data by writing to many other persons working on the problems of flight and by experimenting. They even build a wind-tunnel in their bike shop! These same tactics—breaking a large problem into smaller ones, using analogies, collecting more information—can be used with many of our own day-to-day problems to facilitate the finding of original solutions.

4. *Associative Theory and Divergent Thinking*

Something more should be said about less mindful (cognitive rather than metacognitive) contributions to creative cognition. There are, for example, theories describing creative thinking as associative. Clearly, individuals often follow chains of thought, associating one experience or idea to a related one. That second idea or experience may in turn lead to a third idea, which then leads to a fourth, and so on.

In one associative theory creative insights result from long ideational chains. In this view the first associations are the most obvious (black is associated with white; dog is associated with cat). Creative insights, on the other hand, tend to be 'remote associates.' They are further along the associative chain. Research suggests that individuals differ in associative tendencies. Creative persons, for example, move quickly to remote associations. In theory, each of us can tactically and strategically postpone quick judgments, incubate (take one's time), and try to allow long chains of ideas and associations. In simplest terms the tactics would be 'take your time; don't rush; don't accept the first solution that comes to mind.'

Associative theory was used in the development of the Remote Associates Test, known as the RAT

(Mednick 1962. In the RAT, three words are given (e.g., blood, river, money) and the examinee is asked to find a common associate. Validation studies have offered only marginal support for the RAT, perhaps because of a potential verbal bias, but this does not imply that the theory which led to the RAT is invalid.

Associative theory was also used when tests of divergent thinking were developed in the 1950s (e.g., Guilford 1950, 1968). These tests also allow individuals to follow associative paths but they are less biased by verbal ability and scored for original ideas rather than correct associations. Correct associations, like those on the RAT, may result from 'convergent thinking' rather than 'divergent thinking.'

Divergent thinking is probably used any time an individual is faced with an open-ended task. On tests of divergent thinking, open-ended tasks include requests like, 'name as many round things as you can think of,' or 'how are a potato and carrot alike?' Clearly these questions do not have one correct or conventional answer. The examinee can generate a large number of ideas, all of which are fitting to the task at hand. Some are likely to be original and even creative. Thus, from this perspective, creative cognition is operationalized as divergent thinking.

Divergent thinking is further operationalized in terms of the individual's 'ideational fluency,' 'ideational originality,' and 'ideational flexibility.' Ideational fluency is defined in terms of the individual's productivity. The fluent individual will produce a large number of ideas. Originality is defined in terms of unusualness or uniqueness. The original individual will produce ideas that are unusual in the sense that other people do not think of the same things. Flexibility is defined in terms of variety. The flexible person produces varied ideas, ideas which tap a range of categories or cross diverse themes.

The inflexible person will respond to the task 'name things which are round' with 'tennis ball, basketball, baseball, softball, volleyball, golf ball, soccer ball,' and so on, while a flexible person will respond to the same question with something like 'eyeball, pea, Venus, golf ball, atoms, peaches,' and so on. Fluency is the most commonly recognized feature of divergent thinking, in part because it is predictive of originality and flexibility; but originality and flexibility are very important as well. Originality is, after all, the most widely accepted dimension of creativity. Creative things are always original. Flexibility is important in part because varied ideation may make it easier to find original solutions and insights. Also, flexibility seems to preclude the 'functional fixity' which keeps some people in intellectual ruts.

Empirical investigations have provided stronger evidence of validity for tests of divergent thinking than they have for the RAT, though the actual validity coefficients are only moderate. Still, no one claims that all creative thinking requires divergent thinking. Divergent thinking can lead to original ideas, but divergent thinking is not synonymous with creative thinking.

5. *Issues Concerning Creative Cognition*

One of the issues in this area of study concerns the relationship of creative cognition with problem-solving skill. There are theorists who believe that all creativity is a kind of problem solving. There are others who argue that problem solving and creativity are frequently independent. Problem solving may be creative, but sometimes is not. Support for this second view was mentioned above, with the distinction of divergent and convergent thinking. Sometimes problem solving requires one correct answer; this is convergent rather than divergent thinking, with little or no room for originality. The debate concerning the relationship of creativity and problem solving has been reviewed by Runco (1994).

The arts are unambiguously creative and typically seem to rely more on self-expression than problem solving. Then again, it is easy to see that artists are often solving problems. They may be expressing the problem in the content of the art, or exploring alternatives, or they may be experimenting with technique and solving problems of communication.

Better support for the independence of problem solving and creativity is provided by examples of 'problem finding.' Problem finding is the general term for effort that precedes problem solving. Long ago Wallas (1926) suggested that the creative process requires 'preparation,' along with 'incubation,' 'illumination,' and 'verification.' The preparation stage of this process may involve identifying a problem to solve, or the definition of a problem that allows original solutions to be found. Apparently there are notable individual differences in problem-finding tendencies, and persons who devote more attention to carefully defining problems are often more creative than other persons. It may even be the case that creative solutions are only possible when there is a original problem. Problem finding does seem to be distinct from problem solving, and it seems to play a role in many creative efforts. It is best not to equate creative thinking with problem solving.

Another issue involves the relationship of cognition proper with extracognitive influences on creative thinking. Creative efforts are, for example, often tied to intrinsic motivation. The creative person is not reacting to incentives, rewards, or other extrinsic motives but is instead driven by a personal need to understand or solve a problem. The issue reflects the priority of cognition and motivation. It has been suggested that the former depends on the latter because we don't think about what we don't care about. It has also been suggested that the latter depends on the former because we don't care about things we can't understand.

Creative cognition probably involves both motivation and cognition *per se*. In the model of the creative process presented by Runco and Chand (1995), a primary tier of cognitive components includes problem finding, ideation, and evaluation skills, with a secondary tier of important influences including motivation (both intrinsic and extrinsic) and information (both factual and procedural, or tactical). Although some parts of creative cognition are in fact more cognitive than others, realistically, extracognitive influences, and their interaction with the cognitive components, need to be recognized.

6. *Conclusions*

Creative thinking includes some basic cognitive processes, including perception, attention, and memory. Creative cognition involves perceptual processes when a creative insight is a direct result of the individual's original interpretation of experiences or associations. Creative cognition may reflect attentional selections when the individual benefits from the recognition of experiences and associations that can be used for original ideation. Creative thinking can depend on memory and the storage of information as well. Some creative achievements actually depend on expertise and require years (to master the knowledge which is required by a field). Recall what was said above, however, about expertise occasionally working against creative insights by leading to routine and assumption. Other basic processes are involved in associative and divergent thinking. In addition to these basic processes, creative cognition may be metacognitive and tactical. In this sense we can exert a degree of control over our own thinking and direct our cognition to the generation of original and useful ideas, insights, and solutions. We can think creatively.

See also: Creativity Training; Problem Solving: Deduction, Induction, and Analogical Reasoning

Bibliography

Guilford J P 1950 Creativity. *American Psychologist* **5**: 444–54
Guilford J P 1968 *Creativity, Intelligence and their Educational Implications*. EDITS/Knapp, San Diego, CA
Martindale C 1999. In: Sternberg R J (ed.) *Handbook of Creativity*. Cambridge University Press, New York, pp. 138–52
Mednick S A 1962 The associative basis for the creative process. *Psychological Review* **69**: 200–32
Mumford M D in press Creative cognition. In: Runco M A (ed.) *Creativity Research Handbook*, Vol. 2. Hampton Press, Cresskill, NJ
Runco M A (ed.) 1994 *Problem Finding, Problem Solving, and Creativity*. Norwood, NJ
Runco M A (ed.) in press *Critical Creative Processes*. Hampton Press, Cresskill, NJ
Runco M A, Chand I 1995 Cognition and creativity. *Educational Psychology Review* **7**: 243–67
Wallas G 1926 *The Art of Thought*. Harcourt, New York

M. A. Runco

Creativity and Innovation in Organizations, Management of

At the root of the development of our species from our primitive beginnings to the recent stunning advances in technology, communication, and social complexity, have been creativity and innovation—the development and implementation of ideas for improved processes, products, or procedures (Drazin and Schoonhoven 1996). In this article the factors that influence individuals, groups, and whole organizations to be creative and innovative in their work, are described. The range of research in the social sciences on this topic is enormous partly because of the positive connotations and practical benefits associated with innovation.

However, the notion that creativity and innovation are easy, positive, opportunistic processes is inappropriate. Conflict is a common characteristic of innovation, observable principally in resistance to change. Innovation by definition represents a threat to the status quo, so for a group or organization to implement innovation successfully, its members must manage conflict with the attendant emotional pain and difficulty, overcome resistance to change, and persist in ensuring the successful implementation of their innovative proposal. And innovations do not always have beneficial outcomes. An innovation is often a mistake after the event. Moreover, even creativity appears to be a consequence of hard work or as Csikszentmihalyi and Sawyer (1995, p. 331) suggest '… 99 per cent perspiration and 1 percent inspiration.'

To make sense of this domain of social science research without recourse to complex academic definitions it is useful to distinguish between creativity and innovation implementation. Creativity can be seen as the development of new ideas, while innovation implementation is the application of those new ideas in practice. Using this distinction between creativity and innovation implementation it becomes clearer that creativity is more a characteristic of individuals, while innovation implementation tends to be accomplished by groups, organizations, or societies. Indeed, there is evidence that people are most creative when they work alone.

Creativity requires individuals with creative characteristics, who feel free from threat and pressure, and

work in a supportive environment. Innovation requires diversity, integration, external challenge or demand, and practical support for innovation.

1. Individual Innovation

1.1 Individual Characteristics

The innovation process begins with the creativity of individuals, so the generation of a new idea is a cognitive process, located within individuals, albeit fostered by interaction processes, for example, in teams (Mumford and Gustafson 1988). Thus, first and foremost, innovative individuals are both creative and innovative (i.e., they don't just have creative ideas, they also try to implement them). These are people who have a preference for thinking in novel ways; who think globally instead of locally (distinguishing the wood from the trees); and who have intellectual abilities that include synthetic abilities (to see problems in new ways and escape the bounds of conventional thinking), analytic abilities to recognize which ideas are worth pursuing, and the practical contextual abilities to persuade others of the value of their ideas (Sternberg and Lubart 1996).

To be innovative and creative also requires sufficient knowledge of the field to be able to move it forward, while not being so conceptually trapped in it as to be unable to conceive of alternative courses (Mumford and Gustafson 1988). A legislative thinking style is required, or a preference for thinking in novel ways of one's own choosing. It also helps to be able to think globally as well as locally, and to be able to recognize which questions are important and which are not (Sternberg and Lubart 1996).

People who are confident of their abilities are more likely to innovate in the workplace. Researchers have found that confidence and motivation to develop knowledge and skills predicted innovation following job change. Innovative people also tend to be self-disciplined in matters concerning work, with a high degree of drive and motivation, and a concern with achieving excellence (Mumford and Gustafson 1988). This perseverance against social pressures presumably reduces the dangers of premature abandonment. Minority influence theory in social psychology suggests that this perseverance acts to bring about change in the views of majorities, and is a necessary behavioral style among innovators (see Nemeth 1986).

Tolerance of ambiguity, widely associated with creativity, enables individuals to avoid the problems of following only mental ruts, and increases the chances of unusual responses and the discovery of novelty (Barron and Harrington 1981).

Innovative people tend to be self-directed, enjoying and requiring freedom in their work (Mumford and Gustafson 1988). They have a high need for freedom, control, and discretion in the workplace and appear to find bureaucratic limitations or the exercise of control by managers frustrating. Such people need clear work objectives along with high autonomy to perform well in work.

1.2 Work Environments

Creative cognitions occur when individuals feel free from pressure, safe, and positive (Claxton 1997). Experimental manipulations of stress levels in experiments have shown that higher levels of stress lead to greater reliance on habitual solutions. Moreover, psychological threats to face or identity are also associated with more rigid thinking, and time pressure increases rigidity of thinking on work-related tasks such as selection decisions. Speculation (a critical creative process) makes people in work settings feel vulnerable because they often tend to experience their workplaces as unsafe. Questioning the person who comes up with an idea too closely, joking about the proposal (even in a light way), or simply ignoring the proposal, can lead to the person feeling defensive.

The characteristics of work tasks also impact upon people's creativity. Oldham and Cummings (1996) found that the following five job characteristics predicted levels of individual innovation at work: skill variety and challenge, task identity, task significance, autonomy task feedback.

Skill variety refers to the degree to which a job requires different activities in order for the work to be carried out and the degree to which the range of skills and talents of the person working within the role is used. Thus a nurse working with the elderly in their homes may need to use the professional skills of dressing wounds, listening, counseling, being empathic, and appraising the supports and dangers in the person's home.

Task identity is the degree to which the job represents a whole piece of work. It is not simply adding a rubber band to the packaging of a product, but being involved in the manufacture of the product throughout the process, or at least in a meaningful part of the process.

Significance of the task in terms of its impact upon other people within the organization, or in the world at large, has an influence on creativity. Monitoring the effectiveness of an organization's debt collection is less significant than addressing the well-being of elderly people in rural settings, and may therefore evoke less creativity.

Autonomy refers to the freedom, independence, and discretion for the people performing tasks, in determining how to do their work and when to do it. Their level of autonomy directly determines the extent to which people are creative and innovative in their work.

When people receive feedback on their performance they are more likely to become aware of the 'performance gaps.' Consequently they are more attuned

to the need to initiate new ways of working in order to fill the gaps. Of course this also implies that they have clear job objectives.

While some theories of creativity and flow suggest that creative work is primarily sustained by intrinsic motivation (Amabile 1988), emerging research evidence suggests that extrinsic rewards can complement intrinsic motivation. Rewards appear to be counterproductive only if they serve to displace attention from the task toward the reward. Despite much popular belief to the contrary, there is considerable evidence that extrinsic rewards encourage both creativity and innovation implementation.

Human beings minimize effort in their activities and therefore some external stimulus is necessary to prompt the extra effort required to innovate. Andrews (1979) found that moderate time pressure was generally associated with higher creativity amongst R and D scientists. Many studies have shown that work demands explained more of the variance in innovation than any other work role factor. Necessity may well be the mother of invention.

Job security is a positive predictor of innovation. People on short-term contracts are much less likely to innovate than those with longer tenure. At the same time, however, job tenure is often a negative predictor. In other words, the longer people are in their jobs, the less likely they are to innovate. Paradoxically, organizational tenure seems to be a positive predictor. Research suggests that job insecurity reduces the motivation to innovate. However, the longer people stay in the same job within the organization the less they innovate.

1.3 Work Group or Team Innovation

Group or team innovation occurs when a diverse group (in terms of task relevant knowledge, skills and abilities) experiences both high external demands and high levels of internal integration and psychosocial safety (and this may largely be determined by the knowledge skills and abilities members have for working in teams). Put succinctly, creativity requires safety, while innovation implementation requires high demands (whether in the form of threat or large reward). Groups will be creative primarily when their task is sufficiently interesting, motivating, and challenging and when the group's internal environment or processes are experienced as safe. Diversity is necessary to ensure there is sufficient difference and richness of input to encourage creative and innovative outputs (Mumford and Gustafson 1988).

Research on climates for innovation in teams has revealed the importance of shared objectives in teams, high levels of participation (information-sharing, shared influence over decision making and interaction), an orientation to excellent task performance, and support for innovation.

West and Anderson (1996) also found that these team climate factors predicted levels of innovativeness over a six-month period, in top management teams in 27 hospitals—innovation was monitored via tape recordings of team meetings, interviews with team members, and study of the minutes of team meetings (see also the seminal UNESCO studies reported in Andrews 1979). Moreover the proportion of innovators in the team (members scoring high on a measure of innovativeness) predicted the radicalness (changes to status quo) of innovations they introduced into the hospitals.

Of course, one of the implications is that organizations need to encourage teams by focusing not just on productivity, but also on creativity and innovation as important performance outcomes. Research examining the climate of UK university departments over an eight-year period showed that external ratings of performance, based on quantitative measures of scientific productivity and excellence, led members of those departments to report an increase in the perceived effectiveness of those departments. But they also reported that their departments became more formalized (more rules and regulations) while levels of participation in decision making and support for innovation declined. These are precisely the conditions under which innovation and creativity will also decline.

A key indicator of innovation in work groups is reflexivity. Team reflexivity is the extent to which team members collectively reflect upon the team's objectives, strategies, and processes as well as their wider organizations, and adapt them accordingly. There are three central elements to reflexivity: reflection, planning, and action or adaptation:

(a) Reflection consists of attention, awareness, monitoring, and evaluation of the object of reflection.

(b) High reflexivity exists when team planning is characterized by greater detail, inclusiveness of potential problems, hierarchical ordering of plans, and long-range as well as short-range planning. The greater the details of implementation intentions or plans, the greater the likelihood that they will manifest in innovation.

(c) Action refers to goal-directed behaviors relevant to achieving the desired changes in team objectives, strategies, processes, organizations, or environments.

As a consequence of reflexivity, the team's reality is continually renegotiated during team interaction. Understandings negotiated in one exchange between team members may be drawn upon in a variety of ways in order to inform subsequent discussions and offer the possibility of helpful and creative transformations in meanings as well as the development of shared mental models.

Leadership is also important to team performance and creativity. For example, research reveals that chief executive officers' ages, flexibility, and perseverance are all positively related to the adoption of tech-

nological innovation in their organizations. How the leader manages team meetings can influence group processes—directive chairs inhibit teams from achieving shared mental models and inhibit innovation by restricting the multiplicity of interactions in free-flowing teams. Recent theories of leadership depict two dominant styles: transformational and transactional. Transactional leaders focus on transactions, exchanges, contingent rewards and punishments to change team members' behavior. Transformational leaders influence group members by encouraging them to transform their views of themselves and their work. They rely on charisma and the ability to conjure inspiring visions of the future. Evidence suggests that it is transformational leaders who have most influence on consensus decision-making processes, and thereby innovation.

Dissent is a potential stimulus for group innovation when it occurs in a cooperative context. Minority influence theory provides an understanding of how dissent and minority influence intragroup processes lead to creativity and innovation (Nemeth and Owens 1996). It implies that individuals in small groups can bring about change in the views of the majority through being consistent and persistent. It suggests that minorities and dissenting views in organizations should be tolerated, and also encouraged, in order to develop seedbeds for innovation and creativity. By drawing upon this area of theory and research, powerful conceptual insights into innovation implementation strategies can be developed.

2. Organizational Innovation

2.1 Organizational Climate

Organizations create an ethos or atmosphere within which creativity is either nurtured and blooms in innovation, or is starved of support. Psychological research has revealed that supportive and challenging environments are likely to sustain high levels of creativity (Mumford and Gustafson 1988), especially those that encourage risk taking and idea generation (Kanter 1983). Employees frequently have ideas for improving their workplaces, work functioning, processes, products, or services. Where climates are characterized by distrust, lack of communication, personal antipathies, limited individual autonomy, and unclear goals, the implementation of these ideas is inhibited (Amabile 1988).

2.2 Organizational Structure

The greater the complexity and more differentiated the organization (in terms of departments, groupings, etc.) the easier it is to cross boundaries and the greater the number of sources from which innovation can spring. Yet, as organizations become more complex, senior managers tend to increase formalization and control systems and this constrains innovation (Kanter 1983, Van de Ven et al. 1989).

Companies operating in uncertain environments will require flexible, decentralized, and informal work practices (organic rather than mechanistic forms of organizing) in order to respond effectively through innovation. Collaborative idea development across an organization is often cited as a precondition for organizational innovation (Kanter 1983, Zaltman et al. 1973). There is support for the notion that high centralization is a negative predictor of innovation (Burns and Stalker 1961), and Lawrence and Lorsch's (1967) case studies showed that tightly coupled interdepartmental relationships fostered new product development in organizations. However, research also suggests that centralization may also be necessary to ensure innovation implementation. Zaltman et al. (1973) call this the innovation dilemma. Decentralization at local level is necessary for creative ideas to be developed but centralization may be required for the effective implementation of those ideas in the wider organization. The failure of many organizations to innovate may be a consequence of a failure to recognize this inherent tension.

The resolution of the dilemma may be team-based organizations. Teams provide the sources for ideas (especially cross-functional teams) while the team-based organization also offers simultaneously centralized and distributed decision-making structures that enable successful innovation. Indeed, the extent of team-based working in organizations appears to be a good predictor of innovation.

2.3 Organizational Size and Age

Large organizations have difficulty changing their forms to fit changing environments. Yet organizational size has been a positive predictor of both technological and administrative innovations. Innovative agility is more a characteristic of smaller organizations (Rogers 1983). Size may be a surrogate measure of several dimensions associated with innovation such as resources and economies of scale. However, in large organizations, decentralization and specialization are not sufficient to ensure innovation. Integration across groups, departments, specialisms, is also necessary for communication and sharing of disseminated knowledge, and this requires some centralization or else the sophisticated development of team-based structures. More recent research, examining all 35 US firms that produced microprocessors between 1971 and 1989, showed that smaller organizations were more likely to be the sources of innovation.

The longer human social organizations endure, the more embedded become their norms and the more resilient to change become their traditions. Consequently, mature organizations will have more difficulty innovating and adapting. Evidence from both UK and

US studies suggests that younger organizations (years since start up) are likely to innovate in products, production technology, production processes, work organization, and people management.

2.4 *Market Environment and Uncertainty*

Low market share predicts higher levels of product innovation while environmental uncertainly predicts administrative innovation (innovation in work organization and people management systems). These findings point to the importance of external threat or demand in stimulating innovation in organizations.

2.5 *Management of Quality*

The philosophy of continuous improvement embodies a central commitment to innovation. A wide variety of methods and techniques are deployed by total quality management (TQM) adherents, including improvements in product and manufacturing design, upgraded technology, responsibility for quality at the point of production, statistical process control, and the development of a production culture based on continuous improvement in both products and processes. The more sophisticated and pervasive the management of quality, the more innovative organizations tend to be.

2.6 *Support for Innovation*

Organizations must provide appropriate resources for creative efforts and encourage independent action in order to facilitate the implementation of the creative ideas of those who work within them. There are generally positive and strong correlations between R and D expenditure in organizations and innovation.

Companies use a wide range of schemes to support innovation, and the need for such practical schemes to stimulate, support, and ensure the implementation of innovation is clear. Many organizations include a commitment to innovation in their philosophies, but the practical realization of their commitments is poor. The tendencies to inertia in organizations and resistance to change require clear, practical, and effective strategies to promote innovation.

A central contextual factor for encouraging innovation is reward for innovation implementation. Outcomes like money, fringe benefits, public recognition, and preferred work assignments are likely to lead to innovation implementation.

2.7 *Networks and Alliances*

Increasingly researchers are examining the relationships between organizations (e.g., between suppliers and customers and in joint ventures) and this research suggests that such alliances are an important source of innovation. However, mergers, acquisitions, and divestitures by organizations are associated with lower levels of innovation, probably because managers divert attention and resources away from innovation in order to manage structural changes.

3. *Conclusions*

Findings across a range of levels and sectors make it clear that innovation only occurs where there is strong practical and cultural support for efforts to introduce new and improved products and procedures. At the same time, opportunities to develop and implement skills in the workplace and to innovate are central to the satisfaction of people at work (Nicholson and West 1988), while innovation is vital to the effectiveness of organizations in highly demanding and competitive environments.

See also: Creativity and Cognition; Creativity Training; Innovation and Technological Change, Economics of; Innovation: Organizational; Innovation, Theory of; Organizational Climate; Teamwork and Team Training; Technological Innovation

Bibliography

Amabile T M 1988 A model of creativity and innovation in organizations. *Research in Organizational Behaviour* **10**: 123–67

Andrews F M (ed.) 1979 *Scientific Productivity*. Cambridge University Press, Cambridge, UK

Barron F B, Harrington D M 1981 Creativity, intelligence and personality. *Annual Review of Psychology* **32**: 439–76

Burns T, Stalker G M 1961 *The Management of Innovation*. Tavistock, London

Claxton G 1997 *Hare Brain, Tortoise Mind: Why Intelligence Increases When You Think Less*. Fourth Estate, London

Csikszentmihalyi M, Sawyer K 1995 Creative insight: The social dimension of a solitary moment. In: Sternberg R J, Davidson J E (eds.) *The Nature of Insight*. MIT Press, Cambridge, MA, pp. 329–63

Drazin R, Schoonhoven C B 1996 Community, population, and organization effects on innovation: A multilevel perspective. *Academy of Management Journal* **39**: 1065–83

Kanter R M 1983 *The Change Masters: Innovations for Productivity in the American Corporation*. Simon and Schuster, New York

Lawrence P R, Lorsch J W 1967 *Organization and Environment: Managing Differentiation and Integration*. Harvard University, Boston

Mumford M D, Gustafson S B 1988 Creativity syndrome: Integration, application and innovation. *Psychological Bulletin* **103**: 27–43

Nemeth C, Owens P 1996 Making work groups more effective: the value of dissent. In: West M A (ed.) *Handbook of Work Group Psychology*. Wiley, Chichester, UK, pp. 125–42

Nicholson N, West M A 1988 *Managerial Job Change: Men and Women in Transition*. Cambridge University Press, Cambridge, UK

Oldham G R, Cummings A 1996 Employee creativity: Personal and contextual factors at work. *Academy of Management Journal* **39**: 607–34
Rogers E M 1983 *Diffusion of Innovations*, 3rd edn. Free Press, New York
Sternberg R J, Lubart T I 1996 Investing in creativity. *American Psychologist* **51**: 677–88
Van de Ven A H, Angle H L, Poole M S (eds.) 1989 *Research on the Management of Innovation: The Minnesota Studies*. Harper & Row, New York
West M A 1997 *Developing Creativity in Organizations*. British Psychological Society, Leicester, UK
West M A, Anderson N R 1996 Innovation in top management teams. *Journal of Applied Psychology* **81**(6): 680–93
West M A, Farr J L (eds.) 1990 *Innovation and Creativity at Work: Psychological and Organizational Strategies*. Wiley, Chichester, UK
Zaltman G, Duncan R, Holbeck J 1973 *Innovations and Organizations*. Wiley, London

M. A. West

Creativity Training

Creativity training focuses on the enhancement of creative potentials and performances. The training may be short- or long-term. It may involve explicit instructions or be less direct (e.g., simply suggesting that individuals appreciate originality and look to maximize it in their daily lives). Creativity is multifaceted (Runco and Albert in press), and enhancement efforts tend to focus on one component of what has been described as the creativity syndrome, or complex. Creativity often results from particular cognitive operations and processes, for example, but these are not used unless the individual is motivated to use them. In that sense, creativity is an affective, as well as cognitive phenomenon (Runco and Richards 1997). Particular personality traits, such as openness to experience, may also be involved in creative work, and many individuals studying creativity now assume that there are social, interpersonal, and contextual influences as well. This article first describes the issue of enhancement, and then reviews the cognitive, affective, and contextual training targets and concerns.

1. *Training as the Enhancement of Potential*

In the present context, 'enhancement' is best viewed as the fulfillment of potential. Such fulfillment may be intentional and systematic, as is the case in most experimental investigations of the topic and in those educational programs that attend to creativity. Enhancement also occurs in relatively unsystematic ways, as is the case in families or organizations which appreciate creativity but do not explicitly target original behaviors (Basadur 1994).

Defining enhancement as the fulfillment of potential allows all increases in creativity—systematic and unintentional, short- and long-term—to be recognized. It also circumvents the controversy over the nature and nurture of creativity. This is because most human potentials are both genetic and experiential, with experience (in the present context, the experience that leads to enhancement) working within genetic boundaries. Biologists refer to the 'range of reaction,' with the range genetically determined and the reaction within that range determined by the individual's experience. In this light, the enhancement of creativity merely fulfils an individual's genetic potential. Some will have more of that potential than others; not everyone can be Mozart or Einstein. Everyone can, however, fulfil his or her potential for original and creative expression. Given the importance of creativity for personal growth and social and technological evolution, we should consider carefully how best to fulfil the relevant potentials.

2. *Enhancement of Cognitive Components of Creativity*

There are various cognitive theories of creativity. Some involve basic process, such as memory and information processing, and some focus on higher-order thought, including problem solving, metaphor, and divergent thinking. Metacognitive processes are frequently involved in cognitive theories of creativity. This is because they refer to strategic and intentional operations, and creative work is often both of those. Many creative persons focus on doing what has not been done before, for example, or on stretching and extending ideas, approaches, or work into entirely new directions. They intentionally, and often strategically, target creative results.

As a matter of fact, given the intentionality assumed by all strategic and metacognitive efforts, it is likely that a count of all systematic enhancement efforts would show that the majority attempt to improve thinking by providing tactics. Table 1 presents a list of some common tactics.

Note that many of the tactics in Table 1 are interrelated. Many of them, for instance, involve some sort of change in perspective. This is true of points (a)–(h), (l), (m), (r), (t), (w), (x), and (z)—and the benefit of several others may follow from a shift in perspective. New perspectives break mental sets and routines, and thereby minimize the chances of fixity. Original ideas often result.

Several of the tactics focus on originality. This is because originality is necessary, though not sufficient, for creativity. Consider points (i), (j), and (t)–(v). Keep in mind, however, that original insights are not always creative. Some original insights are unusual, novel, or unique precisely because they are worthless or inappropriate. Creativity involves more than orig-

Table 1
Example tactics to improve creative thinking

(a) Change your perspective.
(b) Work backwards, from the end of the problem (the solution) to the beginning.
(c) Turn the problem upside down.
(d) Question your assumptions.
(e) Get another opinion. Talk to someone from another field, or with a different perspective.
(f) Incubate. Put the problem aside for a while.
(g) Think like a kid. (What would your younger brother say?)
(h) Break the problem into small parts.
(i) Try new combinations.
(j) Explore new connections.
(k) Let your mind wander. Trust your intuition.
(l) Be playful, foolish, uninhibited.
(m) Use trial and error, especially the latter—and then learn from those mistakes.
(n) Gather more data.
(o) Think big! Think small! 'Honey I shrunk the kids!'
(p) Find an analogy (The Wright brothers did it; they looked to birds!).
(q) Look to nature.
(r) Visualize. Try 'the mind's eye.'
(s) Change the medium. (Draw a picture. Make up a song about that problem.)
(t) Explore the bizarre.
(u) Be a 'contrarian.' Walt Disney was one—that's how we got Mickey Mouse.
(v) Bend the rules! Dr. Seuss made up how many new words?
(w) Focus on quantity for a while. A high-quality idea may find a way out, eventually.
(x) Borrow an idea and then make it better.
(y) Ask questions. Ask: why? how? why not? what if?
(z) Change the problem.

inality. Truly creative insights are original and somehow fitting or useful (Runco and Charles 1993).

One final note about the list of tactics given in Table 1. Some of them involve a problem, and others focus on a solution. It has been said that creative ideas require creative problems. Sometimes redefining the problem guarantees an original solution. With this in mind, it is not surprising that some tactics lead the individual to change the problem or its representation.

3. *Problem Solving and Finding*

There is a larger issue: namely that of the relationship of creativity with problem solving (and problem finding). Many definitions of creativity presume that it is a kind of problem solving. The alternative is that problem solving is sometimes creative, but that creativity may be involved when an individual has the need or urge for self-expression. In this alternative view, creative thinking and problem solving are sometimes unrelated; creative skills may be used for more than problem solving.

Tactics such as those listed above (and there are many others) are often included in the various systematic programs which are designed to enhance creativity. Synectics, for example, encourages the use of analogical thinking. It is easy to see the power of analogical thinking in inventions such as the steam engine (a tea kettle), Velcro (clingy burrs), and the airplane (birds, to whom the Wright brothers looked many times). Here again there is an alternative, which in this case is that some creative thinking uses analogical reasoning, and some does not.

There are many programs for enhancing creativity, in addition to synectics (see Meador et al. 1999). There is the Odyssey of the Mind, for example, Future Problem Solving, and Creative Problem Solving (CPS). The last of these is probably the best known. It relies heavily on *brainstorming*. When brainstorming, the quantity of ideas is targeted by a group; quality is de-emphasized. Judgment and evaluation are postponed, and 'piggybacking' (extending ideas given by others, stretching a line of reasoning) is encouraged. Empirical research on brainstorming suggests that it is only modestly effective (Rickarts and deCock in press). In fact, if an organization really needs a large number of ideas, they may be better off having individuals work alone. Brainstorming may be good for team building, but it does not seem to work as well as individual problemsolving. Too often, individuals in brainstorming groups loaf ('social loafing') and assume that others will do enough work to get the job done.

4. *Information Processing*

Enhancement efforts occasionally focus on basic facets of information processing. These efforts usually sug-

gest changes in information transfer and say little about information storage (memory). Sometimes enhancement is expected to result when an individual stops treating information as factual in the static sense. This is another way of saying that the individual questions assumptions and changes his or her perspective. Information is mindfully examined and perhaps even altered, rather than merely transferred without question.

Other times, enhancement is expected when the individual is simply better informed, or at least has the most relevant data. Good information is very important for many creative solutions. It is, however, a matter of quality more than quantity. In fact, individuals can apparently have too much information! When this occurs, there is a kind of expertise that typically leads the expert to more assumptions and a reliance on routines (Rubenson and Runco 1995). These preclude mindful and spontaneous thinking. Creative thinking thus often requires an optimal amount of information; more is not always better. Perhaps it would be most appropriate to conclude that enhancement is most likely when the individual is informed but, in addition, avoids the assumptions and routines that may preclude original thinking.

5. *Motivation*

The best enhancement efforts communicate the value of creative solutions and ideas, in addition to simply showing how the individual or organization can be more original. They demonstrate why and how, as well.

This brings us to the motivational and attitudinal components of creativity. Many enhancement efforts attempt to change attitudes about creative work and assume that improved attitudes will motivate the person or organization to utilize the strategies that have also a part of the enhancement. Attitudinal changes may be necessary because people will need to expend effort to change their ways. They will need to break routines, to question assumptions, to take risks.

They may also assume that creative persons are eccentric, or even difficult to work with (because they do things their own way); and in this light, one of the critical attitudes targeted in enhancement efforts is tolerance (Richards in press). People in organizations wishing to increase innovation and creativity must tolerate independence, risks, and perhaps a lack of structure. Educators must tolerate divergence in the thinking of the students; and this is especially difficult, because so much of the educational curriculum (and assessment tradition) is convergent. Tests, for example, are most objective when there is one correct answer, rather than originality.

It can be very difficult to motivate an individual to do creative work. If that individual is already motivated, it may be a matter of supplying tactics, direction, or resources. But if the individual is not motivated, or motivated to work in other ways (e.g., 'satisficing,' or just getting the job done), the motivation to do creative work may be difficult to arouse. This is because creative efforts are typically intrinsically, rather than extrinsically, motivated. Extrinsic motivation may result from pay, reward, incentive, or the like—contingencies and promises made by others. Intrinsic motivation, on the other hand, comes from within. It is personal and independent of contingencies and promises. Moreover, extrinsic factors can undermine intrinsic motivation. For this reason, the creativity that results from intrinsic motivation to solve a problem or express oneself can be lost if it is not respected and if an effort is not made to reinforce or redirect it. This can make enhancement very difficult. It means that enhancement usually requires that demonstrations and the like deal with impersonal examples. They may need to leave the application of the tactic by the individual to his or her own life to that individual.

6. *Environmental and Interpersonal Factors*

The need to respect intrinsic motivation and the need for tolerance in educational settings or organizations implies that there are interpersonal and contextual influences on creativity. Indeed, enhancement may result from the manipulation of context and interpersonal processes as much as from direct and explicit training.

In general terms, creativity is supported by environments which allow autonomy and which provide the resources and respect necessary for the individual to maintain independence of thought (Rickards and deCock in press, Rubenson and Runco 1995). Tolerance for autonomous and diverse work styles and interests may be necessary, and individuals may need time. Deadlines and imposed schedules can constrain thinking. Admittedly, there are individual differences. Some famous creative persons have claimed that they work best under pressure (e.g., with some competition).

Importantly, creative results are often unexpected. They are, after all, original, and original results may be at best loosely associated with previous efforts. But the unexpected nature of creative insights means that there is a risk. Hence, a tolerance of risk, on the part of the individual and those working with the individual, is often necessary.

7. *Conclusions*

Enhancement may involve explicit instructions (e.g., 'think of things no one else will,' 'find an analogy in

nature'), or it may result from a particular environment, context, and interpersonal stance. Enhancement efforts may focus on the cognitive, metacognitive, motivational, or attitudinal facets of the creative process. The cognitive facets may be the easiest to target (e.g., by suggesting tactics) and the motivational the most difficult. The most effective enhancement will take most or all facets of the creative process into account.

Educators or managers may recognize the need for enhancement efforts. They may investigate the programs which are designed to maximize creative efforts, or they may take it upon themselves to create a stimulating environment and provide the respect, resources, and tolerance that will allow creative thinking. Enhancement is, however, probably most effective over the long run. Short-term programs may be effective, but they may fail to generalize and may not be maintained for very long. Creativity is the most likely when enhancement is supported over a long period of time. In this sense parents may be in the position to best enhance creative thinking. They can model creativity, provide opportunities for original problem solving and self-expression, and appropriately appreciate autonomy and risk taking—and they can do so for an extended period of time. If the benefits of enhanced creativity are recognized on a larger scale, by society as a whole, enhancement is virtually assured. If that occurs, parents, educators, managers, and supervisors will all tolerate, appreciate, and support enhanced creative thinking.

See also: Creativity and Cognition; Creativity and Innovation in Organizations, Management of; Discovery Learning, Cognitive Psychology of; Intrinsic Motivation, Psychology of; Problem Selection in the Social Sciences: Methodology

Bibliography

Adams J L 1986 *Conceptual Blockbusting: A Guide to Better Ideas*. Norton, New York

Basadur M 1994 Managing the creative process in organizations. In: Runco M A (ed.) *Problem Finding, Problem Solving, and Creativity*. Ablex, Norwood, NJ, pp. 237–68

Meador K S, Fishkin A S, Hoover M 1999 Research-based strategies and programs to facilitate creativity. In: Fishkin A S, Cramond B, Olszewski-Kubilius P (eds.) *Investigating Creativity in Youth: Research and Methods*. Hampton Press, Cresskill, NJ, pp. 389–416

Richards R in press Millennium as opportunity: Chaos, creativity, and Guilford's structure-of-intellect model. *Creativity Research Journal*

Rickards T, deCock C in press Understanding organizational creativity: Towards a multi-paradigmatic approach. In: Runco M A (ed.) *Creativity Research Handbook*. Hampton Press, Cresskill, NJ, Vol. 2

Rubenson D L, Runco M A 1995 The psychoeconomic view of creative work in groups and organizations. *Creativity and Innovation Management* **4**: 232–41

Runco M A (ed.) 1997 *Creativity Research Handbook*. Hampton Press, Cresskill, NJ

Runco M A, Albert R S (eds.) in press *Theories of Creativity*, rev. edn. Hampton Press, Cresskill, NJ

Runco M A, Charles R E 1993 Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences* **15**: 537–46

Runco M A, Richards R (eds.) 1997 *Eminent Creativity, Everyday Creativity, and Health*. Ablex, Norwood, NJ

M. A. Runco

Creolization: Sociocultural Aspects

The concept of creolization, so called by analogy with the creolization of languages, has been used in anthropology to refer to the process of social change that takes place in societies characterized by rapid social flux, and where cultural influences from various origins are integrated into a new system of meaning locally produced. Cultural creolization is found particularly in urban centers in association with sustained cultural contact, more typically produced within the hegemonical conditions of colonization, of World system influences, and of globalization.

1. The Origin of the Concept

The concept of 'creolization' is borrowed from linguistics, where it refers (a) to a process of transformation of a pidgin language into the main language (and often, into the mother tongue) of a new linguistic community, and (b) to the creation of a creole language out of different languages (at least two) without an intermediate pidgin state (see *Pidgin and Creole Languages*). It is this second meaning that has been borrowed by anthropology, but as this process is rather rarer than the preceding one, it would have been more accurate to use the term of pidginization (formation of a pidgin language) to refer to the cultural phenomena that are now described with the concept of creolization (Jourdan 1987, McKellin 1991). A pidgin or a creole develops as a response to communication pressures, in situations of intense cultural contact, and incorporates features from languages available to the pidgin/creole makers: the new language is distinct from the languages that have shaped it, yet has retained sufficient features of them, that their origin can sometimes be traced and identified. Theories of pidgin and creole genesis are highly debated in creole studies, yet scholars agree that the characteristics of pidgin and creole languages include the speed with

which they form and develop; the sociological conditions that have shaped them, most typically, colonization; and an almost instantaneous creativity.

2. Creolization and Anthropology

Despite the epistemological difficulties inherent to the borrowing of theories and analytical models across disciplines, the concept of creolization seems to have captured the imagination of some scholars as a way of talking about and analyzing rapid cultural change associated with contact between societies of different ideological and technological orientation. The fact that the label 'creolization' was retained, rather than the more technically appropriate 'pidginization' is linked to two factors: first, the word 'creole' already existed in anthropology to refer to the people and plantation societies of the Caribbeans and the Indian Ocean, typically characterized by a high degree of cultural diversity and intense contact, but nevertheless full-fledged cultural formations, most often sustained by creole languages. Second, for a long time pidginization has been thought of as a process of linguistic impoverishment, whereas creolization was always associated with linguistic enrichment and elaboration. Using the former label could have been seen as a pejorative and inadequate label to describe a process of cultural creativity where 'loss' had no place. In this case, the metaphorical power of creolization won over the more technically correct pidginization. The earlier examples of such usage appear in the works of Drummond (1980), and Hannerz (1987). In parallel, French linguists and anthropologists working in West Africa and in the Pacific have introduced the concept of *créolité* to talk about the creole-like features of the societies where they work (Mannessy 1987), of 'creolicity' to talk about the socioeconomic and ideological conditions fostering the development of creolized societies (Jourdan 1985) and of cultural *métissage* (Amselle 1990).

In anthropology, creolization has been used in two different ways. First, as a metaphor capturing elusive processes of integration of new cultural forms, a synonym of hybridization associated with globalization. Second, as an analytical model for the study of social change directly mapped onto the linguistic model of creolization–pidginization described above. In all cases, creolization is used to refer to social change. An example, yet not representative of all types and domains of creolization, would be a Solomon Island business man sending printed invitations to the Catholic wedding of his daughter in the biggest hotel of the city of Honiara, but who would make sure to collect the bridewealth from the family of the groom, and would subsequently redistribute it to his kin.

Used as a metaphor, the meaning of creolization remains very akin to that of hybridization and creates the same type of conceptual ambiguities: is it the process or is it the results of structural modification or reorganization? The image that there is something that gets mixed with something else to produce yet another thing (e.g., the new form of marriage in Honiara), is inherent to the metaphor. As such, it is an interesting shortcut into the description of phenomena not easily definable, but which seem to be clearly associated with the complexification of social relationships. Used only as a metaphor, creolization (and hybridity) fail to explain how what is happening warrants the usage of this new term rather than intersystem (Drummond 1980), syncretism (see *Syncretism*), or métissage, those other concepts central to a study of social change that focuses on the in-between. Creolization, in that sense, is used to refer to a recombination of structures, practices and meanings: a case of 'new meanings in old shells' or 'old meanings in new shells.' In as much as they focus on the dynamic intermixing and restructuring of meanings, both concepts are similar. Both capture the image of the in-between, of the neither/nor, of flux and vanishing boundaries, but focus on the result.

A stronger reading of the concept of creolization, closest to the way the concept is used in linguistics, proposes ways of identifying the cultural logic which provides answers to the following questions: What changes and what does not, and under which conditions? What is the nature and the pace of the change? If culture is practice, how does change affect the practice of identity? If culture is text, is every aspect of culture affected in the same manner and to the same extent? If culture is grammar, are there parameters and constraints for change? What makes the change acceptable? This version of creolization rests on a set of assumptions: (a) that moral–existential aspects of culture can be differentiated from instrumental aspects; (b) that the latter change at a faster pace than the former; (c) that, with regards to the pace of change, they can be compared with the syntax and lexicon of a language respectively. In situations of creolization, local instrumental aspects of culture are often replaced, along with their attending symbols and meanings, with those from another source. The moral–existential aspects of culture are now sustained by new symbols, and may be given slightly different meanings, but will still be perceived as essential to the group's definition of identity. The bridewealth payment mentioned above is essential to the father's standing in one of his networks (here his lineage) and justifies his traveling back to the home village to collect it. All participants in the transaction agree that the meaning of such bridewealth is now very different from what it used to be, and that the amount is inflated. Yet bridewealth has to be paid, even against the wishes of the parish priest, because this form of exchange is essential to the integration of lineages, and therefore to the father's identity as a member of the group. As with any form of social process, creolization rests on people's agency and ability to engage the

world, and is associated with a reconfiguration of self and of the practice of identity. Undoubtedly, there is a distinct advantage in identifying the processes that lead to a redefinition of people's practice of identity, even if, initially, this identification may start with an analysis of the features and elements that are being transferred across cultures, with or without attendant localization.

3. *Essentialism*

Contra critics have argued that creolization theory is essentialist because it presupposes the existence of pristine cultures that will subsequently join, mix, overlap, and be reshaped; proponents of creolization reject this notion, and point to culture as a locus of permanent change, made even more spectacular at particular times in the history of societies. They also point to the shortcomings of a conception of culture as a bounded pristine entity.

If creolization were an essentialist theory, the concept of cultural loss would be integral to its definition. Yet, it is not, for the simple reason that talking about loss in connection with social change implies a move away from the ideal, but unreal, pristine state of tradition. Loss belongs to a theoretical framework in which cultures are seen as essential wholes, neatly bounded self-reproducing structures, impervious to the passage of time and deriving legitimacy from a primal pristine state. Cultures are not less authentic because they are creolizing. The new cultural practices, and the new cultural forms, that have emerged out of the creolization process are not any more authentic or spurious than the ones they are derived from, or more spurious or authentic than societies that have changed in other ways. They simply exist. What is revealed by the creolization concept is the endless negotiation of meaning that takes place when groups and individuals engage in social relations within societies. The pace and scope of these cultural negotiations are more important in situations of rapid social change. The strength of creolization theory is that it points to (or foregrounds) elements that are not easily captured by other models: heterogeneity rather than homogeneity, process rather than stasis, fragmentation rather than integration, multiple rather than singular identities, and above all at the domain of culture as contested ground rather than public space.

4. *The World System and Globalization*

Typically, but not exclusively (see Hannerz 1996) the creolization concept has been applied to the analysis of social change in the so called Third and Fourth World societies, in colonial and postcolonial situations. Colonial, postcolonial and hegemonical contact relations are particular in the type and speed of the sociocultural change they provoke and are best suited to be studied with the paradigm of creolization that holds the speed of change as a defining element. Critics have argued that the quasi-exclusive application of creolization theory to these societies disempowers them and reinforces the *fait accompli* of colonization by strengthening stereotypes. The argument is usually framed in relation to the World System theory and to the concepts of core and periphery. There may be a reason for this affinity: the very sociocultural formations and ideologies that have fostered the development of creole languages have reappeared under a new guise, that of the World System, economic and cultural, which creates even more pressures for change in these societies than anywhere else at the same moment. These types of contact-induced ideological and experiential pressures resemble the superstrate and substrate influences at work in the formation of creole languages in that they juxtapose and confront practices and meanings that are produced by different social systems and ideologies. This raises the question of the relationship of creolization to globalization. Creolization shares with globalization an increase in available modes of social organization and cultural representation (Pieterse 1995), both offer more options. Both are associated with more culture and not less, even if, on the surface, things look different. But with globalization residing in a movement of ideas, as much as in people's engagement with new sets of ideologies and practices, one could identity, as Hannerz does in Stockholm (1996), creolizing formations in nontypical creolization sites.

5. *Urbanization*

It is in cities that creolization is the most obvious, if only because of the rate and scope of change. As cultural market places, cities are the loci of very fast sociocultural mutations that lead some observers to talk of deculturation, and not only of adaptation to an urban way of life. Creolization is neither. When full-fledged towns and cities develop over mere 40 years, as is the case in many parts of the Third World, in places where there is no local tradition of urbanization, for instance, it requires on the part of urbanites a reassessment of the parameters of their identity (as with the father's behavior at the time of his daughter's wedding) and subsequent engagement in new types of social relations of the type and scale that is not found in other forms of sociocultural change. What is typical of cities is the ebullition of cultural creativity. Only a close study of the superstrate and substrate influences could reveal why, despite the onslaught of a generic cultural and economic World System (the superstrate) onto local traditions (the substrate), the cultural formations of urban Nigeria (Hannerz 1987) Solomon Islands (Jourdan 1985), and Guyana (Drummond 1980) are similar in very many ways, and yet very different from

one another. This begs the question of the specificity of creolized societies: are they any different from other types of societies, given the fact that all societies are changing, and that many contemporary non-Third World societies have, at some point in their history, lived in situations of colonization, invasion or hegemony? May be they are, if only because of the speed and the depth of change that characterize them, and give the impression of a permanent flux. The lifespan of creolization is also a subject of debate: is creolization a transitional period in the life of societies, at particular moments of their history, or is it a permanent state of affairs? One has to ask whether anthropology will be better served by using creolization as a universal synonym for fast cultural change associated with globalization, or whether the concept is not more useful when applied to some social contexts of change and not to others.

See also: Cultural Assimilation; Culture Shock; Globalization and World Culture; Hegemony: Cultural; Hybridity; Pidgin and Creole Languages; Syncretism

Bibliography

Amselle J L 1990 *Anthropologie de l'identité en Afrique et ailleurs*. Payot, Paris

Drummond L 1980 The cultural continuum: a theory of intersystems. *Man* **15**: 352–74

Hannerz U 1987 The world in creolization. *Africa* **57**: 546–59

Hannerz U 1996 *Transnational Connections*. Routledge, London

Jourdan C 1985 Sapos iumi mitim iumi: creolization and urbanization in the Solomon Islands. Ph.D. thesis. The Australian National University, Canberra, ACT

Jourdan C 1987 Des plantations à la ville. *Journal de la Société des Océanistes* **85**: 243–53

Mannessy G 1987 Créolisation et créolité. *Études Créoles* **10**: 25–38

McKellin W 1991 Hegemony and the language of change: the pidginization of land tenure among the Managalase of Papua New Guinea. *Ethnology* **30**: 313–24

Pieterse J N 1995 Globalization as hybridization. In: Featherstone M, Lash S, Robertson R (eds.) *Global Modernities*. Sage, London, pp. 45–68

C. Jourdan

Crime and Class

Does social class affect criminal behavior? If so, why has the relationship been so difficult to document in self-report crime surveys? If not, why are prisoners drawn almost exclusively from the ranks of the poor? These questions are important for stratification researchers as well as those studying law, criminology, and deviance. Criminologists are interested in class because it is a fundamental indicator of social position. If crime is related to social class, then social structural factors such as inequality, poverty, and labor markets are likely to be important causes of crime. If not, criminologists must look to factors that may cut more evenly across class lines, such as individual pathology or dysfunctional family processes. Stratification researchers are becoming interested in crime for analogous reasons. Individual-level evidence suggests that crime and the societal reaction to crime can disrupt individual attainment (Hagan 1993), while macro level studies show that crime and punishment may alter unemployment rates and economic performance (Western and Beckett 1999). This article first reviews empirical generalizations about class and crime, and the conceptual and methodological tools used to interpret them. It then outlines recent theoretical and empirical developments and diverse research strategies to further elaborate the relationship, concluding that the relation between criminal behavior and social class is likely to be reciprocal or endogenous.

1. Measurement

1.1 Measurement of Crime

The relationship observed between crime and class is sensitive to the measurement of each concept. There are three primary sources of crime data: the offenders, the official enforcement agencies, and the victims. Self-report studies, typically based on samples of high-school students, generally report very weak relationships between social class and delinquency (Tittle et al. 1978). However, social class is a strong and significant predictor of official law violation, as measured by arrest and punishment, all over the world (Braithwaite 1981). Surveys of prison inmates show that they are educationally and economically disadvantaged relative to the general population. Prior to their most recent arrest, only two-thirds of inmates in the USA had completed high-school, one-third were not employed, and less than half reported annual income greater than $10,000 (US Department of Justice 1993). The relation between social class and criminal victimization depends on the type of crime under consideration. In 1998, in the USA, the violent victimization rate among households reporting annual income greater than $35,000 was only half the violent victimization rate of households earning less than $7,500. For property crimes such as theft, however, the highest victimization rate occurs among households earning $50,000 or more per year (US Department of Justice 1999).

It has proven difficult to reconcile the conflicting pictures of the class–crime relationship provided by self-report, official, and victimization data. Some argue that the discrepancy arises because self-reports generally capture a different domain of behavior than official data. When self-reports are compared with police records, the results show high rates of congru-

ence for white, in-school youths, but less agreement for black males with an official history of delinquency (Hindelang et al. 1981). Some attribute these discrepancies to under-reporting by black males, while others point to race biases in arrest records that overstate the arrests of blacks relative to whites. Most agree that greater police attention to lower-class areas exacerbates class (and race) discrepancies in self-report and official data.

The nonsignificance of class in self-report surveys may also be a product of the compressed frequency and severity range of offending typically recorded in self-report surveys (Elliott and Ageton 1980). In support of the argument that self-reports and official data tap different domains of criminal behavior, some investigations find that middle-class and upper-class delinquency is less serious than delinquency by less advantaged youth. Nevertheless, the strong negative relation between social class and official delinquency and the far smaller class differences in self-report delinquency are well-founded empirical generalizations.

Any discussion of class and criminal behavior would be incomplete without noting that social class is one of the defining characteristics of white-collar crime, or 'crime committed by a person of respectability and high social status in the course of his occupation' (Sutherland 1983, p. 7). By definition, this conceptualization assumes that some degree of power and influence is necessary to commit crimes such as bank embezzlement or insider trading. Because the structure of opportunity to commit such crimes is bound up with social position, however, it is difficult to identify meaningful comparison groups to determine whether white-collar, occupational, or corporate criminals are any more or less deviant than the general population.

Finally, most self-report studies have been conducted on youth, often in school settings. Unfortunately, we know comparatively little about the class distribution of self-reported adult offending in the general population. Class differences in the longitudinal sequence or criminal careers of individual offenders have rarely been explored, although lower-class offenders may have higher crime frequency and participation rates, earlier ages of onset, and longer career durations (Janson and Wikström 1995). With the aforementioned caveats regarding occupational and adult offending, however, the results of self-report studies are generally consistent, reporting weak negative relations between social class and crime. Much of the remaining variation among these studies can be attributed to differences in the conceptualization and measurement of social class.

1.2 Measurement of Class

The effects of social class on self-reported crime depend on whether researchers use gradational or relational class measures. Researchers using continuous gradational measures, such as socioeconomic status (SES), generally find little association between SES and crime. Investigations that operationalize class as a relational measure, based on Marxian ownership or Weberian authority generally report somewhat stronger associations (Brownfield 1986) (see *Class and Law*). One problem with many class indicators, however, is that they omit the most disadvantaged individuals and families—the underclass or surplus population. Theories of crime such as Merton's (1938) anomie model point to the importance of sustained poverty and limited opportunity within a cultural context of common success symbols. Along these lines, some research suggests that parental unemployment is a positive predictor of self-reported delinquency (Farnworth et al. 1994).

Most of the research considered so far actually considers social class origins rather than the social positions or interests of the offenders themselves (Stark 1979). In part, this is because the peak age of offending for many crimes occurs in the mid-to-late-teens, preceding the transition to full-time employment in many societies. To the extent that the social class positions of adolescents have been considered in the delinquency literature, schools are generally the primary (if not exclusive) setting. Nevertheless, many adolescents and young adults have established independent class positions outside the educational arena, particularly those who have left home and school. Research on homeless youth that assesses class in terms of unemployment, hunger, and lack of shelter, reports strong 'foreground' class effects on street crime (Hagan and McCarthy 1997).

In summary, the strongest relations between social class and crime have been observed in studies that use relational rather than gradational class measures. In fact, the most important class distinction is likely between the most disadvantaged and everybody else. Though some specifications of the 'underclass' concept blur the distinction between class and criminal behavior by including law violation (in the form of violence, drug use, or other crime) among its many manifestations, such tautological reasoning is unnecessary to establish the relationship. Sustained parental unemployment in an individual's background or more immediate and extreme material deprivation places persons at greater risk of offending.

2. Sampling, Units of Analysis, and Analytic Approaches

2.1 Sampling

As the preceding discussion suggests, class effects on criminal behavior are unlikely to be observed unless the research methodology ensures an adequate rep-

resentation of extremely disadvantaged respondents and a meaningful comparison group. 'Street criminology' differs from 'school criminology' (Hagan and McCarthy 1997) in both theoretical and analytic orientation. In particular, each approach implies a different domain of criminal behavior and a correspondingly different sampling strategy. Student samples can yield elegant tests of theory among nationally representative samples, although results may not generalize to more extreme law violation. Street samples are more difficult to obtain, in part because complete sampling frames are rarely available. Nevertheless, the latter approach is best suited to identifying respondents from truly disadvantaged class origins and more marginalized current class positions.

2.2 Units of Analysis

Studies conducted at the aggregate rather than the individual level of analysis have generally shown stronger relationships between social class and crime. Housing costs affect the population composition of a neighborhood, which, in turn, influences the distribution of social groups with differential risks for offending (Janson and Wikström 1995). At the community level, the studies of Shaw and McKay (1942) showing delinquency rates in Chicago by neighborhood economic status remain among the most powerful statements in this line of research. At least part of the individual-level discrepancy between self-reported and official crime and delinquency is thus explained by greater police attention to lower-class areas (Sampson 1986).

Emerging multilevel methods are beginning to shed light on the intervening mechanism connecting class with neighborhood crime rates. Research merging individual-level data with neighborhood characteristics shows that the effect of poverty and unemployment rates on violent crime is mediated in large part by 'collective efficacy,' or a community's capacity for informal social control and mutual trust (Sampson et al. 1997). Finally, national-level research on class structure and crime suggests that inequality (Braithwaite 1979), state welfare spending (Messner and Rosenfeld 1997), and unemployment (Archer and Gartner 1984) bear some relation to official crime rates.

2.3 Qualitative and Quantitative Analytic Approaches

Many qualitative researchers take the association between class and crime for granted, focusing their efforts on gangs or individuals in extremely marginalized circumstances. Increasingly, however, such work also includes a comparative dimension. Mercer Sullivan (1989) for example, shows how early criminal behavior can disrupt adolescent education and employment, which in turn increases adult criminal behavior. By examining cliques of boys from similar class positions but different neighborhoods and ethnicities, Sullivan identifies contingencies such as local job networks and illegal markets that mediate the relation between class origins and crime. Other ethnographic work illustrates how street crime is woven into the everyday lives of some individuals and neighborhoods (Venkatesh 1997), identifying the critical turning points in their criminal histories. Because such insights are difficult to glean from the quantitative analysis of survey data, several prominent quantitative criminologists have added a qualitative element to their investigations (Hagan and McCarthy 1997, Sampson and Laub 1993).

3. Promising Avenues for Research

3.1 Reciprocal Relations

The reciprocal relationship between attainment and deviance that Sullivan (1989) noted among disadvantaged adolescents has been explored in a number of important quantitative studies. Analyzing a cohort of London boys born in the 1950s, Hagan (1993) finds that adolescent criminal 'embeddedness,' measured by delinquent activities, friends, and parental criminal convictions, has a strong effect on unemployment in young adulthood. Similarly, in a sample of American delinquent and non delinquent boys born in the 1920s, Sampson and Laub (1993) report that early juvenile incarceration decreases later job stability, which, in turn, exacerbates adult criminal behavior. Thus, social class and crime are likely to be reciprocally determined.

The type and severity of offending however, remains important in determining crime effects on class. Substance use and minor delinquency have little effect on the life outcomes of a predominantly middle-class sample, partly because they are likely to receive numerous 'second chances' (Jessor et al. 1993). Similarly, Hagan (1991) finds that identifying with a subculture of delinquency reduced occupational attainment among working-class boys, but identifying with a 'party subculture' had a small positive effect on attainment among non-working-class boys. In such cases, the type and severity of offending determine the sign as well as the magnitude of the effects of crime on social position.

3.2 Elaborating the Relationship

Criminological researchers have elaborated the relationship between class and crime by specifying and

testing interactions between social class and other factors and by seeking the intervening factors that interpret the relation. With regard to interaction, the relationship between social class and crime may be conditional on background characteristics, such as race, sex, or other characteristics. Power-control theory, for example, posits that social class interacts with gender to affect common delinquency, receiving mixed empirical support (Hagan et al. 1987) (see *Crime and Gender*; *Crime and Ethnicity (Including Race)*).

Regarding intervening factors, Wright et al. report both positive and negative social-psychological paths connecting socioeconomic status and delinquency, characterizing the relationship as 'causation but not correlation' (1999, p. 179). High SES increases delinquency by increasing risk taking and social power and by decreasing conventional values, but reduces delinquency by diminishing alienation and economic strain, and increasing aspirations (Wright et al. 1999). Other research suggests that class effects on crime and deviance are largely mediated by family factors and school performance (Hagan and McCarthy 1997).

Even if the bivariate relationship between class and crime is close to zero, as some self-report research suggests, further research on interactions is necessary to explain how we get to zero—to specify the positive and negative paths linking social class and criminal behavior. Where the bivariate relationship between class and crime is stronger, as in studies using parental unemployment or official crime data, further research is needed to identify the catalysts that transfer class origins into criminal outcomes.

3.3 Cross-national and Comparative Work

A great deal of the research discussed above has been conducted on English-speaking samples, such as Hagan's work on British (1993) and Canadian (1991) data, the Wright et al. (1999) piece on New Zealanders, and the many American studies reviewed by Tittle et al. (1978). Nevertheless, Braithwaite's comprehensive international review suggests that patterns observed in these nations are replicated across the world with a 'degree of consistency which is unusual in social science' (1981, p. 38). To date, however, few truly cross-national or comparative efforts have been undertaken. Most studies have taken nation states as the level of analysis, identifying the structural-level correlates of homicide rates (e.g., Archer and Gartner 1984). One individual-level effort, comparing working-class males in London and Stockholm, found remarkable similarity in the criminal career patterns of the two groups (Farrington and Wikström 1994).

It is difficult to locate the comparable samples and indicators necessary to test the stability of individual-level class effects across national contexts. Nevertheless, such research will be necessary to generalize knowledge of class and crime and to begin to disentangle the relation between social structure and criminal behavior in varying cultural settings.

4. Taking Stock

The concepts of class and crime are fundamental to understanding social position and social order and disorder. Not surprisingly, their relationship is complex and multifaceted. Social class is a strong predictor of official law violation by all measures and a weaker but consistent predictor of self-reported law violation by gradational measures. Researchers are elaborating the relationship by specifying the interactions and the intervening mechanisms connecting class and crime. In particular, it has become increasingly clear that criminal behavior and social class are likely to be reciprocally related throughout the life course.

See also: Class and Law; Class: Social; Crime and Delinquency, Prevention of; Crime and Ethnicity (Including Race); Crime and Gender; Crime: Knowledge about and Prevalence; Crime: Sociological Aspects; Crime, Sociology of; Durkheim, Emile (1858–1917); Personality and Crime; Quetelet, Adolphe (1796–1874); Race and the Law; Statistical Systems: Crime

Bibliography

Archer D, Gartner R 1984 *Violence and Crime in Cross-National Perspective*. Yale University Press, New Haven, CT

Braithwaite J 1979 *Inequality, Crime, and Public Policy*. Routledge & Kegan Paul, London

Braithwaite J 1981 The myth of social class and criminality reconsidered. *American Sociological Review* **46**: 36–57

Brownfield D 1986 Social class and violent behavior. *Criminology* **24**: 421–38

Elliott D S, Ageton S S 1980 Reconciling race and class differences in self-reported and official estimates of delinquency. *American Sociological Revie* **45**: 95–110

Farnworth M, Thornberry T P, Krohn M D, Lizotte A J 1994 Measurement in the study of class and delinquency: Integrating theory and research. *Journal of Research in Crime and Delinquency* **31**: 32–61

Farrington D P, Wikström P-O H 1994 Criminal careers in London and Stockholm. In: Weitekamp E G M, Kerner H-J (eds.) *Cross-national Longitudinal Research on Human Development and Criminal Behavior*. Kluwer, Dordecht, The Netherlands, pp. 65–90

Hagan J 1991 Destiny and drift: Subcultural preferences, status attainments, and the risks and rewards of youth. *American Sociological Review* **56**: 567–82

Hagan J 1993 The social embeddedness of crime and unemployment. *Criminology* **31**: 465–91

Hagan J, McCarthy B 1997 *Mean Streets: Youth Crime and Homelessness*. Cambridge University Press, Cambridge, UK

Hagan J, Simpson J, Gillis A R 1987 Class in the household: A power-control theory of gender and delinquency. *American Journal of Sociology* **92**: 788–816
Hindelang M J, Hirschi T, Weis J G 1981 *Measuring Delinquency*. Sage, Beverly Hills, CA
Janson C-G, Wikström P-O H 1995 Growing up in a welfare state: The social class-offending relationship. In: Blau Z S, Hagan J (eds.) *Current Perspectives on Aging and the Life Cycle*. JAI Press, Greenwich, CT, Vol. 4, pp. 191–215
Jessor R, Donovan J, Costa F 1993 *Beyond Adolescence: Problem Behavior and Young Adult Development*. Cambridge University Press, New York
Merton R K 1938 Social structure and anomie. *American Sociological Review* **3**: 672–82
Messner S F, Rosenfeld R 1997 Political restraint of the market and levels of criminal homicide: A cross-national application of institutional-anomie theory. *Social Forces* **75**: 1393–416
Sampson R J 1986 Effects of socioeconomic context on official reaction to juvenile delinquency. *American Sociological Review* **51**: 876–85
Sampson R J, Laub J H 1993 *Crime in the Making: Pathways and Turning Points Through Life*. Harvard, University Press, Cambridge, MA
Sampson R J, Raudenbush S W, Earls F 1997 Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* **277**: 918–24
Shaw C, McKay H D 1942 *Juvenile Delinquency and Urban Areas*. University of Chicago Press, Chicago
Sullivan M L 1989 *Getting Paid: Youth Crime and Work in the Inner-City*. Cornell University Press, Ithaca, NY
Stark R 1979 Whose status counts? *American Sociological Review* **44**: 668–9
Sutherland E S 1983 *White Collar Crime: The Uncut Version*. Yale University Press, New Haven, CT
Tittle C R, Villemez W J, Smith D A 1978 The myth of social class and criminality: Empirical assessment of empirical evidence. *American Sociological Review* **43**: 643–56
US Department of Justice 1993 *Survey of State Prison Inmates, 1991*. Bureau of Justice Statistics, Washington, DC
US Department of Justice 1999 *Criminal Victimization* 1998. Bureau of Justice Statistics, Washington, DC
Venkatesh S 1997 The social organization of urban street gang activity. *American Journal of Sociology* **103**: 82–111
Western B, Beckett K 1999 How unregulated is the US labor market? The penal system as a labor market institution. *American Journal of Sociology* **104**: 1030–60
Wright B R E, Caspi A, Moffitt T E, Miech R A, Silva P A 1999 Reconsidering the relationship between SES and delinquency: Causation but not correlation. *Criminology* **37**: 175–94

C. Uggen

Crime and Delinquency, Prevention of

Juvenile delinquency and crime are legal definitions rather than specific behavioral or psychiatric syndromes. Since common law is based on theological law, society has historically responded to juvenile delinquency and crime based on moral and religious beliefs regarding the age at which juveniles are criminally responsible rather than from scientific knowledge regarding the developmental course of delinquency and crime. Historically, the age at which children are criminally responsible has been subject to much debate and has varied significantly both within and between nations (Rutter et al. 1998). Moreover, this debate and variability has led to ongoing tension within society and its institutions between a focus on punishment or treatment and rehabilitation. However, since the early 1970s scientists and interventionists in the fields of criminology, sociology, and psychology, who have focused on the prevention of juvenile delinquency and crime, have made significant methodological and theoretical progress. This progress has included identifying specific antecedents and mediators and their developmental sequencing, and the development of precise interventions (see Stoff et al. 1997). Society and its institutions will need to use this accumulated scientific knowledge as the foundation for developing policy and interventions that are intended to prevent delinquency and crime.

1. Historical Background in the Prevention of Delinquency and Crime

Since the rise of delinquency as a social problem during the industrial revolution, the prevention of delinquency and crime has had a history of swinging between two philosophical poles: punishment and rehabilitation (Bernard 1993). Advocates of punishment have relied on the intellectual power of classical theory, whereas advocates of rehabilitation have relied on learning and attachment theories. Classical theory and its criminological offspring, social control and strain theory (Gottfredson and Hirschi 1990), posits that individuals are responsible for their own behavior and are motivated by cognitive appraisals of the potential costs and benefits of delinquent or criminal acts. Attachment and learning theories and their most recent progeny, social interactional or social developmental theory (Patterson et al. 1992), posit that the probability of performing a delinquent or criminal act is highly related to a child's history of reinforcement for certain behaviors within a family and early associations with negative peers. That is, complex but observable reinforcement contingences are responsible for the initiation and then maintenance of chronic antisocial behavior.

Currently, social control theory appears to be most influential among the public and politicians both in the USA and in other Western countries (Howell 1997, Rutter et al. 1998). However, in the scientific literature social developmental theory appears to have more empirical support. Policy and intervention implications of a reliance on social control theory (i.e., harsher punishments such as training schools, juvenile jails, reprimanding juveniles into adult courts and jails,

mandatory sentencing, etc.) have been shown to be neither in the best financial interests of society nor to have reduced rates of juvenile delinquency and crime. Whereas, recent cost–benefit analyses of interventions based on social developmental theories suggest cost effectiveness, particularly in the long run (Howell 1997). Therefore, there has been a re-emergence of social developmental theoretical models of delinquency and crime. These models emphasize the importance of interactional contingencies operating within family, peer groups, schools, neighborhoods, and communities in the development and maintenance of antisocial behavior (e.g., Patterson et al. 1992) versus theoretical models of delinquency and crime based on economic opportunity (e.g., Gottfredson and Hirschi 1990).

2. *The Problem of Chronic Offenders*

Recent research has shown that a small percentage of children with antisocial behavior, including those who are arrested as juveniles, account for a large proportion (50–75 percent) of juvenile crime (Huizinga et al. 1995). Researchers have also found similar percentages in adult samples. Because this smaller proportion of criminals accounts for a large percentage of crime, it has been suggested that communities often waste their limited resources on noncareer juvenile delinquents who are unlikely to commit further offenses. By the time these noncareer criminals are arrested and/or targeted for intervention, they are often at the end of their short offending span. However, it is also argued that intervention may be easier and more likely to work with the non-chronic group. Some researchers have also suggested that the youth in the chronic group account for a disproportionate share of mental and physical health problems, divorce, domestic violence, and drug and alcohol problems in their age cohorts.

It has become clear that it is rare for a persistently delinquent youth to become a chronic offender without moving through an increasingly well-validated developmental sequence (i.e., early childhood noncompliance, antisocial or aggressive behavior in childhood including covert antisocial behavior, and early arrest, Stoff et al. 1997, Chap. 20). For example, Patterson and his colleagues at the Oregon Social Learning Center have developed two models, an early-onset and a late-onset model, that may account for differences between life-course persistent and transient offenders. In the Oregon Youth Study sample Patterson found that boys with the highest rates of antisocial acts in childhood are at the greatest risk of being arrested before the age of 14. In addition, Huizinga et al. (1995) reported that from 39–62 percent of chronic violent offenders self-report committing their first act of violence at or before age 9. Patterson has also found early arrest (before age 14) is associated with the increased probability of being a chronic offender and, in turn, being a chronic offender is positively correlated with both violent offending and being arrested as an adult (Howell 1997, Rutter et al. 1998).

3. *Prediction and Stability of Antisocial Behavior*

Since the early 1970s there has been increasing agreement among researchers that, to a great extent, externalizing conduct disorders in childhood precede chronic delinquency (Patterson et al. 1992), and chronic delinquency precedes serious and violent juvenile crime and adult criminality (Loeber and Farrington 1998). In addition, there is impressive international evidence that antisocial behavior, including criminal behavior, is relatively stable across stages of the life course (Rutter et al. 1998). The stability of delinquent behavior is particularly strong when antisocial behavior begins early in the life course. Researchers have also found that the development of childhood antisocial behavior and juvenile delinquency leading to adult criminality is multiply determined and involves the convergence of many risk factors (Howell 1997, Loeber and Farrington 1998, Patterson et al. 1992, Rutter et al. 1998). However, recent strides have been made in identifying specific risk factors for the initiation and maintenance of antisocial behavior and perhaps most important the developmental sequencing of their influences.

4. *Developmental Sequencing*

Several well-designed prospective longitudinal studies (e.g., Huizinga et al. 1995, Patterson et al. 1992) have shown that levels of common parenting practices starting early in a child's life and continuing through adolescence are associated with the reduced probability of delinquent behavior. These parenting practices include appropriate discipline (e.g., application of mildly aversive consequences for problem behaviors, structuring, reinforcement of prosocial behaviors, and lack of abusive tactics such as humiliation and hitting), supervision (e.g., monitoring children's behavior and peer affiliations in school and community environments), and developing a positive relationship with a child (e.g., having reasonable levels of positive reinforcement for prosocial behavior, involvement in positive leisure activities). These longitudinal studies along with other recent research (Dishion et al. 1999) have shown that delinquent peers become a significant factor in the development and maintenance of delinquency, and this effect is most significant during adolescence.

In addition, researchers have found an array of complex individual, psychosocial, and society-wide influences that are related to externalizing behavior

problems, offending rates, and a variety of antisocial and delinquent behaviors. These influences include genetic influences, obstetric complications, cognitive abilities, temperament, peer relationships, biased cognitive processing, autonomic reactivity, serotonin metabolism, teenage parenting, large family size, broken homes, abuse and neglect, coercion and hostility, ineffective parenting and supervision, peer groups, poverty and social disadvantage, mass media, school effects, and accessibility of guns and drugs (Rutter et al. 1998, Stoff et al. 1997). Although most of these influences have been shown to have robust associations with delinquent behavior, their individual influences on rates of delinquency are often small (i.e., often accounting for less than 10 percent of the variance in delinquency and crime).

What is significant about the individual influences is that most have been shown to increase in importance as they accumulate for any one child. Longitudinal research also suggests that individual influences are most important if they interact negatively with parent, peer, school, and community relationships. Contingent interactions (many of which have been shown to be bidirectional, Patterson 1982) that occur in these settings have been implicated in the emergence and subsequent maintenance of chronic childhood antisocial behavior, delinquency, and crime. For example, recent research strongly suggests that parent–child interaction patterns starting in early childhood that take place in the context of typical parenting tasks are associated with positive and negative child development outcomes (Patterson et al.1992). It has also been hypothesized that biological and genetic variables play a role in the development of problematic parent–child interaction patterns (Rutter et al. 1998). That is, infants with difficult temperaments and/or hyperactivity are harder to parent, and when matched with low-skilled parents, they have been shown to have more negative developmental outcomes such as higher levels of noncompliance and parental rejection. It is also interesting that, studies by Werner (1989) and others have shown that difficult infants raised by reasonably skilled parents are not at particular risk for negative developmental outcomes.

Many of the individual contextual influences that have been historically associated with rates of delinquency and crime (e.g., poverty, high crime neighborhoods, parental antisocial behavior and psychopathology), in large part, have been shown to be mediated to children through their disruptive effect on parenting practices and family functioning (e.g., parental depressed mood, poor parental discipline, marital conflict) (e.g., Patterson et al. 1992, Sampson and Laub 1994). Some theorists even postulate that economic downturns increase both the number of difficult-to-raise infants (because of increases in infant health problems) and the number of low-skilled parents (through the complex effects of transitions and mobility). These low-skilled parents then drift into disorganized neighborhoods (where rents and costs are lower) and where there are high crime rates and low levels of social support that create further isolation and stress (Sampson 1992). Patterson has suggested that further research on the relationship between economic downturns and density of difficult-to-raise infants and low-skilled parents may solve an embarrassing problem for most current theories of delinquency: that these theories have been unable to account for significant variance in societal rates of juvenile offending.

5. *Prevention and Intervention*

The outcome of three decades of intense research has been a general agreement that prevention of delinquency and crime needs to involve multiple components that target multiple risk factors, and that no single-component intervention will have enough strength to alter the course of seriously antisocial or delinquent behavior. Because many of the early risk factors directly or indirectly involve parent–child interactions, early prevention efforts are increasingly targeting parenting and other contextual variables that indirectly or directly affect parenting. As the children age, their interactions with adults and peers in schools, neighborhoods, and the larger community also become important in the further development and maintenance of antisocial and delinquent behavior (Stoff et al. 1997, Chap. 20). Therefore, all serious current prevention and intervention models for antisocial and delinquent behavior target multiple variables (e.g., family interactions, school success, peer group affiliations, social support, graduated sanctions) in multiple settings (e.g., home, school, neighborhoods, and communities) (Loeber and Farrington 1998).

5.1 *Gender*

International research has shown that young males commit a higher proportion (up to 80 percent) of officially reported juvenile crime. This ratio is somewhat lower when self-report data is used to document criminal acts. It is also known that these ratios are falling at the end of the twentieth century and the disparity between male and female delinquency was less than a third of what it was even 40 years before (Rutter et al. 1998). There is argument about why there is an increase in officially reported and self-reported female antisocial behavior (see Stoff et al. 1997). Regardless of why increases in female antisocial behavior are being reported, the long-term detrimental effects of such behaviors are well-documented (Sampson and Laub 1994). In particular, female antisocial behavior, especially delinquency, has been linked to the development of later psychiatric, educational, and

employment/occupational problems, substance abuse, early pregnancy, the selection of antisocial partners, early marriage, and early death.

However, because developmental research in the area of delinquency and crime has primarily involved boys (particularly prospective longitudinal studies), very little is currently known about the development of juvenile offending and criminality in female populations. The existing literature, however, does suggest that there are different effects for parenting, sibling, and peer relations on antisocial behavior in boys and girls. It also shows that the developmental histories of delinquent boys and girls differ in other respects such as levels of abuse, neglect, and trauma (Stoff et al. 1997, Chap. 46). Researchers also have proposed that the same predictors and mechanisms may not explain these gender variations. Further, even though females contribute proportionally less than males to delinquency and crime rates, some theorists believe it is critically important to increase the study of girls. Not only are antisocial or delinquent young women at grave risk for a plethora of negative developmental outcomes, they also may play a key role in the intergenerational transmission of criminality through pairing with antisocial boys, early pregnancy, and low parenting skills (Rutter et al. 1998).

6. *Future Directions*

Continued use of multiple measures and multiple methods within passive and experimental longitudinal studies will further identify malleable variables and validated intervention technologies at all levels of delinquency and crime prevention (i.e., primary, secondary, and tertiary) (Loeber and Farrington 1998). Studies will also focus on the further development and validation of gender-specific predictive and prevention models. These models will likely lead to the development of screening criteria that can identify the small percentage of youths who account for the most delinquency and crime and those youths most likely to be involved in the transgenerational transfer of delinquency and crime. These improved screening measures will reduce to acceptable levels the current high false-positive rates associated with previous attempts to develop screening tools for early identification of children at risk of persistent delinquency and crime. However, until this scientific task is accomplished, results from programmatic longitudinal research using causal modeling suggests that targeting children and families who have both biological and socioenvironmental risk factors will be most cost effective. Although the Gluecks and their colleagues argued for this approach many years ago (Glueck and Glueck 1972), accomplishing it appears possible at the beginning of the twenty-first century, given recent improvements in measurement theory, research design and analysis, and intervention technology.

In addition, future theoretical and methodological advances will include an increase in research on the interaction between biological/genetic variables and social environments (both naturally occurring and therapeutically designed) and the results of this interaction on antisocial developmental trajectories and rates of delinquency and crime (Stoff et al. 1997). Also, empirically validated prevention interventions targeting delinquency and crime will be developed for all ages and levels of seriousness. Furthermore, any intervention used by society will be seen as potentially a preventive intervention if it has a positive effect on the developmental trajectory of antisocial behavior once it has begun (Stoff et al. 1997, Chap. 32). Finally, long-term solutions to the problem of delinquency and crime will lie in the scientifically informed cost–benefit balance between prevention, rehabilitation, and graduated sanctions (Howell 1997, Loeber and Farrington 1998, Rutter et al. 1998).

See also: Antisocial Behavior in Childhood and Adolescence; Crime: Sociological Aspects; Crime, Sociology of; Delinquency, Sociology of; Juvenile Justice: International Law Perspectives; Personality and Crime

Bibliography

Bernard T J 1993 *The Cycle of Juvenile Justice*. Oxford University Press, New York

Dishion T J, French D C, Patterson G R 1995 The development and ecology of antisocial behavior. In: Cicchetti D, Cohen D J (eds.) *Developmental Psychopathology. Vol 2: Risk, Disorder, and Adaptation*. Wiley, New York, pp. 421–71

Dishion T J, McCord J, Poulin F 1999 When interventions harm: Peer groups and problem behavior. *American Psychologist* **54**: 755–64

Glueck S, Glueck E G (eds.) 1972 *Identification of Predelinquents: Validation Studies and Some Suggested Uses of Glueck Table*. Intercontinental Medical Book Corporation, New York

Gottfredson M, Hirschi T 1990 *A General Theory of Crime*. Stanford University Press, Stanford, CA

Howell J C 1997 *Juvenile Justice and Youth Violence*. Sage, London

Huizinga D, Loeber R, Thornberry T 1995 *Recent Findings from the Program of Research on Causes and Correlates of Delinquency* (Report to the Office of Juvenile Justice and Delinquency Prevention). Office of Juvenile Justice and Delinquency Prevention, Washington, DC

Loeber R, Farrington D P 1998 *Serious and Violent Juvenile Offenders: Risk Factors and Successful Interventions*. Sage, London

Patterson G R 1982 *A Social Learning Approach to Family Intervention: III. Coercive Family Process*. Castalia, Eugene, OR

Patterson G R, Reid J B, Dishion T J 1992 *A Social Learning Approach: IV. Antisocial Boys*. Castalia, Eugene, OR

Rutter M, Giller H, Hagell A 1998 *Antisocial Behavior by Young People*. Cambridge University Press, Cambridge, UK

Sampson R J 1992 Family management and child development: Insights form social disorganization theory. In: McCord J (ed.) *Advances in Criminological Theory. Vol 3: Facts, Frameworks, and Forecasts*. Transaction, New Brunswick, NJ, pp. 63–91
Sampson R J, Laub J H 1994 Urban poverty and the family context of delinquency: A new look at structure and process in a classic study. *Child Development* **65**: 523–40
Stoff D M, Breiling J, Maser J D (eds.) 1997 *Handbook of Antisocial Behavior*. Wiley, New York
Werner E E 1989 Children of the garden island. *Scientific American* **260**: 107–11

K. J. Moore

Crime and Ethnicity (Including Race)

This article explores research regarding race, ethnicity, and crime. It focuses on the disproportionately high incidence of racial and ethnic minorities caught up in the justice systems of Western countries and on the reasons behind these disparities.

1. Race, Ethnicity, and Crime and the Limitations on Research

The term 'race' once referred to the 'major biological divisions of mankind' (Caucasoid, Negroid, and Mongoloid), as characterized by skin color, hair texture, and other physical features (Walker et al. 1996). This notion has fallen into disrepute for lack of scientific validation. Race is now recognized as a social construct used by groups seeking to delineate themselves from others. Ethnicity is a similar concept that principally refers to the countries from which an individual's heritage can be traced (e.g., 'Hispanic' refers to those with Spanish ancestry), although it is also associated with language, religion, cultural practices, and self-perception. The meanings of these terms continue to be debated, and usage is not consistent internationally. In Germany, for instance, the term *Ausländer* (foreigner) includes ethnic and racial differences in its vague and elastic scope. In England and Wales, researchers consider Blacks to be ethnic minorities, but offenses against them are sometimes referred to as being 'racially' motivated.

Crime typically is defined as 'a social harm that the law makes punishable' (Garner 1999). Examples of crime range from white-collar crimes and organized crime, to murder and hate crimes. This article will focus on conduct often referred to as 'street crime' (also known in the US as Index Crime: murder and nonnegligent manslaughter, forcible rape, robbery, burglary, aggravated assault, larceny-theft, motor vehicle theft, and arson), because current records provide the best data for analysis of variation in offending and victimization by race and ethnicity.

Analysis is constrained by the difficulty in obtaining reliable crime statistics. Such data are collected regularly in Western countries; however, even where crime data is available, the race and ethnicity of victims or offenders are rarely recorded. Only the US officially records and publishes data regarding an individual's race beginning at arrest and continuing to incarceration, and even there data has reliability problems (see the subsection entitled *The Quality of Criminal Justice Data on Race and Ethnicity*, in Walker et al. 1996). Lawmakers in many countries have chosen not to record racial and ethnic identification data for ethical reasons. Though the reluctance of Europeans to compile data regarding the ethnicity of offenders and victims is hardly surprising, given the legacy of Nazi Germany, the decision to forgo collection is equally problematic. Comparing existing data cross-nationally can be difficult because each country has different laws and processes, research traditions, and immigration histories, but these concerns are not insurmountable.

1.1 Racial and Ethnic Disproportion in Justice Systems

Tonry (1997) notes that 'members of *some* disadvantaged minority groups in every Western country are disproportionately likely to be arrested, convicted, and imprisoned for violent, property, and drug crimes.' Whether or not this is a more universal phenomenon is as yet unknown, as research has largely been limited to English-speaking and European countries. Regardless, what is now known is disturbing enough. Racial disparities in arrest and incarceration have historically been of particular concern to researchers in the US, and it is with that data that this investigation begins.

Table 1
Proportion of arrestees for FBI index crimes by racial category (percent of general population in parenthesis)

	Whites	Blacks	American Indian or Alaska Native	Asian or Pacific Islander
1976	72.3% (88.3)	25.3% (11.5)	1.4% (n.a.)	0.09% (n.a.)(a)
1986	71.3% (84.7)	27% (12.2)	1.0% (b)	0.7%
1996	66.9% (82.8)	30.7% (12.6)	1.3% (c)	1.2%

Source: Bureau of Justice Statistics: Source book of Criminal Justice Statistics: 1978, 1987, and 1997
(a) Chinese and Japanese only (b) Other racial categories comprised 3.1 percent in 1986 (c) Other racial categories comprised 4.5 percent in 1996

The Federal Bureau of Investigation (FBI) has collected crime data since 1930 and annually produces the Uniform Crime Report (UCR). The UCR details arrest data for four racial groups (White, Black, Native American, and Asian), but it does *not* include ethnicity. These data are necessarily incomplete, as they only reflect crimes reported to the police, yet they reveal the stubborn stability of racial disproportion in US arrestees for Index crimes (Table 1).

Clearly, White Americans are the most numerous arrestees, but African Americans are arrested at rates that greatly exceed their representation in the population (about 12 percent). Though the data in Table 1 are aggregated, the racial disparity holds in each Index crime category. In 1996, African Americans were arrested at a rate disproportionate to their representation in the general population ranging from 1.9 times for arson to 4.6 times for robbery.

Research indicates racial and ethnic disparities in US jail and prison populations as well. In 1990, 14 percent of jail inmates were Hispanics, who composed only 9 percent of the general population. More notable is the fact that in 1992, 27 percent of federal prison inmates were Hispanic. US jail and prison data revealed that in 1990, 47 and 51 percent of each population was African American, respectively; the percentage of African American prison inmates has exhibited an upward trend since the 1920s. In 1997, 1,083 of every 100,000 American Indians were in jail—the highest rate for any race.

The US is not alone in exhibiting racial and ethnic disparity in arrests and imprisonment. Australian records, for example, indicate disparity in the arrest of Aborigines—considered a racial minority in that country. Aborigines were almost 8 times more likely to be arrested than non-Aborigines in 1990; the rate increased to 9.2 by 1994. Additional data show that Aboriginal juveniles particularly are likely to be arrested or detained. Data from 1993 reveal ethnic or racial disparities in the prisons of England, Wales, and Canada. Blacks (people of West Indian, Guyanese, and African origin) represented 11 percent of the male English and Welsh prison population, while comprising only 1.8 percent of the young adult males in the general population. Aggregating across all Canadian provinces, researchers found that in 1993–1994, when about 3.7 percent of the Canadian general population was Aboriginal, Aborigines composed 17 percent of provincial prison admittees. Blacks are also overrepresented. Data from Ontario correctional facilities reveal a prison admission rate of 705 per 100,000 for whites, 1,992 per 100,000 for Aboriginals, and 3,686 per 100,000 for Blacks.

Notably, while *some* minorities are overrepresented in various criminal justice systems, others are underrepresented, either in the aggregate or in arrest rates for specific offenses. In England, the rates of imprisonment among Indians, Pakistanis, and Bangladeshis are generally *lower* than that of whites, despite their generally poorer economic position. Likewise, a 1992 profile of US prison populations showed that Asians were underrepresented. Asians comprised 3 percent of the general population, yet they made up only about half of one percent of state inmates and just over 1 percent of federal inmates.

Thus, minorities are over and underrepresented in Western justice systems—usually the former. Why?

3. *Reasons for Racial and Ethnic Disproportion in Justice Systems*

The debate concerning minority overrepresentation in justice system processing has largely centered around two questions. Are racial and ethnic disparities the result of discrimination by criminal justice system actors? Or are they the result of disproportionate offending? Current research strongly suggests that racial and ethnic disproportionate representation in international justice systems results from both factors.

3.1 *Evidence of Discrimination*

The evidence does *not* suggest that bias is systematic: it does not afflict Western justice systems at all levels, in all cases, and at all times. Rather, it suggests that in certain cases, places, and times, minorities are more likely than majority members to face discrimination in interactions with justice system components. Four arenas evidencing bias are worth discussing briefly: 'Neutral' processes, police practices, drug law enforcement, and the death penalty.

3.1.1 'Neutral' processes. Research has confirmed that two practices, designed to be 'race neutral' in implementation, operate to the disadvantage of minorities on a regular basis: pretrial detention, and plea-bargaining. Though it seems logical to hold before trial those most likely to flee, international scholars have discovered that the characteristics associated with flight risk are more likely to be present in the minority population than in the majority, which results in racial disparities in pretrial detention. Similarly, minorities are less likely to plead guilty. Research indicates that this tendency is shaped by and contributes to the beliefs many minorities have about justice system bias.

3.1.2 Police practices. The first experience many minorities have with the criminal justice system comes through contact with the police (see *Policing*). The police function entails vast enforcement discretion, so the possibility of abuse of authority is ever-present. In the countries in which minorities are

overrepresented in the justice system, there is also minority dissatisfaction with police practices. Reports of brutality and unnecessary stops, searches, and arrests fuel these views. In a 1994 US study, 47 percent of African Americans surveyed reported having been harassed by police, while only 10 percent of Whites reported harassment. Researchers in Maryland found that African Americans were the subject of 77 percent of vehicle searches conducted by state troopers in 1995; no contraband was found in 67 percent of the cases. Similarly, the police forces of Glasgow, London, Ottawa, and Toronto are among those who have come under scrutiny recently.

It is possible that police practices impacting primarily minorities also affect the opinions of the police held by the majority population. In a survey conducted in 1981 and 1990 in 20 countries, only a minority of *all* citizens polled, ranging from a high of about 40 percent in Britain to a low nearing 5 percent in Argentina, expressed 'a great deal' of confidence in their nation's police; in all but four of these countries the percentage had *declined* by 1990.

3.1.3 Drug law enforcement. Since the 1980s, the problem of race, ethnicity, and crime in the US and elsewhere has been driven by drug laws and policies that have resulted in the rapid rise in the incarceration rate of minorities. The increase in the Black prison population in the US, England, and Wales is obviously tied to drug law violations. In these countries minorities tend to be convicted of relatively low-level drug offenses at rates much higher than those of the majority of individuals. Scholars have concluded that the stark racial disparities in arrests and incarceration resulting from drug law enforcement are difficult to justify given the small gains in crime control from these policies (Tonry 1995, Meares 1998).

3.1.4 The death penalty. This issue is of particular concern to scholars in the US, as it is the only Western democracy still imposing this sentence. Examination of sentencing practices indicates that the *victim*'s race, not the offender's, most often determines whether or not the defendant will receive the death penalty. Those who murder Whites are more likely to receive the death penalty than those who murder Blacks. African American defendants charged with murdering Whites are most likely to receive the ultimate sanction.

Despite this evidence, the US Supreme Court has refused to void the death penalty, finding in *McCleskey v. Kemp* (1987) that defendants must demonstrate, in addition to discriminatory effect, that the prosecutor's decision to pursue the death penalty in a particular case was the result of racial animus. Today, US death rows are still predominantly minority. Not one death row defendant has yet been able to demonstrate the evidentiary requirements stipulated by *McCleskey* (see *Death Penalty*).

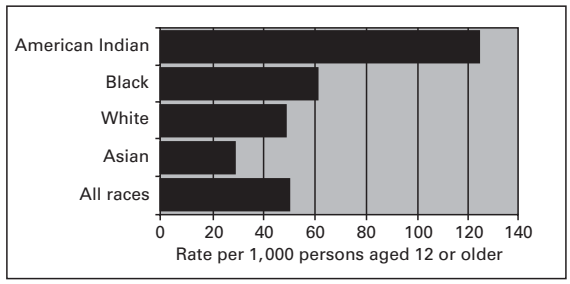

Figure 1
Annual average rate of violent victimization by race of victim, 1992–96. Source: Bureau of Justice Statistics, *American Indians and Crime*. Washington, DC: USGPO, 1999.

3.2 Evidence from Victimization Studies

To determine the extent of discrimination in international justice systems, some scholars have looked to victimization surveys, reasoning that if arrest data matches victimization reports, the racial disparities evidenced in arrests could be explained by disproportionate offending rather than by bias.

Among Western countries, only the US, England, Canada, and The Netherlands conduct long-term national victimization studies. US data shows that minorities fall victim to violent crime at a rate that exceeds their representation in the population (Fig. 1). US data also shows that violent crime is largely an intraracial event. From 1976 to 1996, 94 percent of cases involving Black murder victims involved a Black offender. The corresponding percentage was 85.6 for Whites.

The victimization data largely points away from invidiously or blatantly discriminatory targeting of offenders by the justice system. Instead, the data suggests that differential rates of offending by racial and ethnic groups contribute a great deal to the disparity in the US justice system. Hindelang's (1978) classic study shows significant correspondence between the distribution of the perceived race of offenders from victimization survey results and the distribution of arrestees by race from official statistics. While there are well-documented limitations to this approach, existing research has prompted Tonry (1997) to conclude that 'in countries in which research has been conducted on the causes of racial and ethnic disparities in imprisonment, group differences in offending, not invidious bias, appear to be the principle cause.'

4. Why do Minorities Disproportionately Offend?

Minority groups overrepresented in justice systems tend to be economically and socially disadvantaged, especially compared to the majority. In many cases this is the result of historical discrimination, or is tied to the group's recent immigration to their new national home. Often racial and ethnic minority groups are segregated geographically from other groups. *Social disorganization theory* postulates that the geographical concentration of poverty, along with factors such as joblessness and family disruption, negatively impacts the ability of community-level institutions to mediate crime (Shaw et al. 1929, Sampson 1989). This theory helps to explain why some disproportionately minority areas—especially urban ones—exhibit high crime rates. But it does not explain why particular individuals do or do not engage in crime. It is primarily for this reason that attempts to incorporate the notion of 'rotten social background' as a legal excuse for crime commission have been unsuccessful.

Other explanations for minority involvement in the criminal justice system emphasize power and social relationships over economic disadvantage. *Conflict theory* suggests that discriminatory justice practices are simply another expression of the disproportionate power distribution in society. The dominant group can and does use available means to maintain its grip on power (Liska 1992). *Cultural conflict theory* maintains that heterogeneous societies tend toward disparate values and are thus more vulnerable to crime (Sellin 1938). This theory suggests that some minority groups, for various reasons, do not accept majority values, increasing the likelihood of criminal behavior. Finally, *social strain theory* sees increased criminality as one of several responses to the tension inherent in a society in which many lack the means to obtain dominant goals in the generally agreed upon manner (Merton 1957). Rebellion and retreatism (e.g., through drug use) are among the other reactions to this type of societal stress. While each of these theories has certain merits, none does a particularly good job of explaining why some disadvantaged minorities are disproportionately *underrepresented* in criminal justice systems.

5. Conclusion

Racial and ethnic disparities apparent in Western justice systems will increase rather than decrease, at least in the short term, as a result of increasing rates of immigration to Western countries by racial and ethnic minorities and the worldwide trend toward urbanization. This problem is not confined to the Western countries emphasized here. One need only consider the social landscape of South Africa (and the ongoing conflict between Israel and the Palestinian Liberation Organization) to see this. The main impediment to exploration of these issues worldwide is the profound lack of data. Thus, the challenge of promulgating useful crime policy without unnecessarily impacting racial and ethnic minorities will continue to be considerable without good data to help determine the scope and substance of the problem.

See also: Crime: Knowledge about and Prevalence; Crime: Organized; Crime: Sociological Aspects; Crime, Sociology of; Crime: White-collar; Ethnic Cleansing, History of; Ethnic Conflict, Geography of; Ethnic Conflicts; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Ethnicity, Crime, and Immigration; Hate Crimes

Bibliography

Bureau of Justice Statistics 1999 *American Indians and Crime*. USGPO, Washington, DC

Garner B A 1999 (ed.) *Black's Law Dictionary*. West Gray, St. Paul, MN

Gottfredson M R 1979 (ed.) *Sourcebook of Criminal Justice Statistics—1978*. US Department of Justice, Bureau of Justice Statistics, Washington, DC

Gross S R, Mauro R 1989 *Death and Discrimination: Racial Disparities in Capital Sentencing*. Northeastern University Press, Boston

Hindelang M 1978 Race and involvement in common-law personal crimes. *American Sociological Review* **43**: 93–109

Inglehart R (ed.) 1994 *World Values Survey, 1981–1984 and 1990–1993*. Inter-University Consortium for Political and Social Research, Ann Arbor, MI

Jamieson K M, Flanagan T J (eds.) 1998 *Sourcebook of Criminal Justice Statistics—1987*. US Department of Justice, Bureau of Justice Statistics, Washington, DC

Liska A E (ed.) 1992 *Social Threat and Social Control*. State University of New York, Albany, NY

Maguire K , Pastore A L (eds.) 1998 *Sourcebook of Criminal Justice Statistics—1997*. US Department of Justice, Bureau of Justice Statistics, Washington, DC

Mann C R 1993 *Unequal Justice: A Question of Color*. Indiana University Press, Indianapolis, IN

Meares T L 1998 Social organization and drug law enforcement. *American Criminal Law Review* **35**: 191–227

Merton R K 1957 *Social Theory and Social Structure*. Free Press, New York

Russell K K 1998 *The Color of Crime: Racial Hoaxes, White Fear, Black Protectionism, Police Harassment, and other Macroaggressions*. New York University Press, New York

Sampson R J 1989 Community structure and crime: Testing social disorganization theory. *American Journal of Sociology* **94**: 774–802

Sellin T 1938 *Culture Conflict and Crime*, Bulletin 41. Social Science Research Council, New York

Shaw C R, Forbaugh F, McKay H D 1929 *Delinquency Areas*. University of Chicago Press, Chicago

Tonry M 1995 *Malign Neglect: Race, Crime and Punishment in America*. Oxford University Press, New York

Tonry M (ed.) 1997 *Ethnicity, Crime and Immigration: Comparative and Cross-national Perspectives*. The University of Chicago Press, Chicago

Uniform Crime Reports for the United States 1996 USGPO, Washington, DC
Walker S, Spohn C, DeLone M 1996 *The Color of Justice: Race, Ethnicity, and Crime in America*. Wadsworth, Washington, DC

T. L. Meares

Crime and Gender

One of the most consistent and strong findings in criminology is that females commit much less crime and juvenile delinquency than males. This gender gap in law-violation is found using data on arrests, convictions, self-reported crime, and victims' reports about offenders. It also appears to exist across nations and over time.

Yet, until the 1980s, gender differences in crime received scant attention. In the 1970s, feminist critiques of criminology called attention to the neglect of females (e.g., Smart 1976). Since that time, there has been a trend toward taking more seriously the issue of female offending and the gender gap in crime and delinquency.

This article summarizes the research on gender and crime that has been conducted since these critiques. It first identifies some of the major issues that have emerged in this research, and identifies three typical approaches to the study of gender and crime. It then discusses key findings and arguments in research taking each of these approaches. Note that space limitations restrict the scope of this essay to research on the causes of gender differences in offending; it does not address the literatures on gender and victimization, or gender and criminal justice processing. For a review of this literature, see Kruttschnitt (1996).

1. *Approaches to Studying Gender, Crime, and Delinquency*

A major issue in research on gender and offending centers on the level of analysis. One line of research focuses on the *macro-level* of analysis, emphasizing patterns of crime over time and space, and seeking to explain these as the outcomes of broad social patterns. Specifically, this line of research examines changes in the gender gap in crime over time, and reports some convergence between female and male crime rates since the 1950s and 1960s in various nations, including England and Wales, Canada, and the United States (Box and Hale 1983, Fox and Hartnagel 1979, Steffensmeier 1993). At some times during this period, the increase in female crime rates has exceeded the increase in male rates; at other times, male crime rates have declined more than have female rates. Both of these patterns have resulted in some narrowing of the gender gap in crime, although male offending exceeds female offending throughout the period. This pattern holds for most types of serious crime, except murder (Heimer 2000); however, it is more pronounced during some periods for certain property crimes, such as larceny and fraud (Steffensmeier 1993). Almost all existing research on changes in the gender gap uses annual arrest or conviction rates, aggregated to the national level. The typical explanations offered for the changes include women's liberation and economic circumstances.

Another line of research on gender differences in offending uses individual-level data and proposes *social–psychological* explanations. This research typically is based on quantitative data from self-report surveys administered to large samples. These surveys show that males are more likely than females to commit minor, as well as serious, law violations. However, the gender gap is most pronounced for the most serious offenses (see Smith and Visher 1980). Individual-level research has focused on how social–psychological factors—such as family processes, beliefs, and attitudes—contribute to the gender gap in law-violation.

A second major issue in the study of gender and crime is the type of methodology employed. Whereas the two lines of research discussed above rely mostly on *quantitative* methods, a third line of research uses *qualitative* methods, such as the analysis of open-ended interviews and ethnographic techniques. This research most often focuses on either male or female offenders, and thus addresses differences in crime among groups of females or groups of males rather than the comparison of offending across gender.

The remainder of this essay briefly reviews the three lines of research identified above. It then, discusses attempts to bridge the macro- and individual-levels of analysis. Finally, it suggests that fruitful avenues for future research are continuing to bridge levels of analysis, and moving toward greater synthesis of quantitative and qualitative research findings.

2. *Research on the Narrowing of the Gender Gap Over Time*

Research on the narrowing of the gender gap in crime rates over time proposes three explanations—women's liberation, economic marginalization, and the decline in chivalry. The *liberation hypothesis* argues that women's crime rates increased during the 1960s and 1970s because women gained increased economic and social independence from men. One variant proposes that a convergence in gender roles led to a convergence in the gender gap in all types of crime (Adler and Adler 1975). Another variant proposes that women's increased labor-force participation created new opportunities for crime in the workplace (Simon 1975). Both variants have been criticized for assuming that labor-force participation and improved economic

conditions increase crime by females, whereas these conditions reduce crime by males (Chesney-Lind 1997). Moreover, the narrowing of the gender gap is most pronounced for crimes that are consistent with women's preliberation roles—including larceny (theft), fraud, and forgery (see Steffensmeier 1993). Perhaps not surprisingly, therefore, empirical research provides little evidence that women's labor-force participation increases their crime (e.g., Box and Hale 1983; for review, see Steffensmeier and Allan 1996).

Consequently, contemporary studies often focus on the *economic marginalization hypothesis*, which proposes that the increased financial hardship of women relative to men has contributed to the narrowing gender gap in crime (Box 1987, Steffensmeier 1993). Very few studies have assessed this hypothesis empirically, and existing studies use restricted measures of economic marginalization. A review of these studies argues that future research needs to address the interplay between the specific elements of women's poverty that affect their crime rates, including increasing rates of single motherhood, persistent gender inequality in wages, and changes in social welfare policies (Heimer 2000).

A third explanation of the narrowing of the gender gap in crime is the *decline of chivalry hypothesis*, which proposes that the observed changes are due in part to changing attitudes among law enforcement personnel. Specifically, the gender gap in arrest rates reflects the chivalry of law enforcement personnel toward women (e.g., hesitation to arrest women), and the narrowing of the gap reflects a decline in this attitude over the years (Steffensmeier 1993). Empirical evidence suggests that the decline of chivalry likely has had some effect on changes in the gender gap in crime, although economic factors are at least as important (Box and Hale 1983, Steffensmeier and Allan 1996). However, there remains debate over the veracity of the chivalry hypothesis, with some researchers arguing that females are treated more harshly than males for some types of offenses (Chesney-Lind 1997).

3. *Qualitative Research on Women's Offending*

Consistent with the economic marginalization hypothesis, most ethnographic and qualitative studies of women's crime highlight structural disadvantages. These studies typically focus on women offenders only and do not address directly the gender gap in offending. Instead, they conduct an in-depth analysis of the factors that are associated with chronic offending among women. As a group, these studies suggest that poverty, unemployment, and single motherhood combine to create conditions of extreme economic hardship, which in turn are associated with events and conditions of women's lives that increase the chances of a life of crime (Miller 1986, Carlen 1988, Daly 1994, Maher 1997, Baskin and Sommers 1998). For instance, several studies emphasize the connections between abuse during childhood, economic hardships, and criminal careers among women (Daly 1994, Maher 1997). Other studies show how living in a marginalized community facilitates entry into criminal networks and peer groups, which play an important role in initiating women into lives of crime and sustaining these lifestyles in the future (Miller 1986, Baskin and Sommers 1998). Finally, several studies suggest that turning to crime is a rational attempt to support children and other family members when confronted with severe economic hardship (e.g., Miller 1986, Carlen 1988). These studies offer rich insight into the social world of female chronic offenders and, as we note in Sect. 5, open the door for the development of theories that can address macro-level as well as social–psychological processes.

4. *Self-Report Research on Gender and Juvenile Delinquency*

Most research based on self-report surveys focuses on juvenile delinquency, especially common and less serious forms of delinquency. The emphasis is on understanding the social–psychological mechanisms that contribute to the gender gap in delinquency. Recent research on these mechanisms suggests that the gender gap reflects both gender differences in *exposure* to factors that make law violation more likely, as well as gender differences in the *impact* of these factors.

The social–psychological mechanisms that have received the most attention in research on gender and delinquency occur within the family. Specifically, some research shows that part of the gender gap in delinquency is explained by the greater exposure of girls than boys both to direct parental controls, such as supervision and monitoring, and to indirect controls, such as emotional bonding (e.g., Jensen and Eve 1979, Hagan et al. 1985, 1987). Other research shows that the gender gap is not the simple outcome of girls being subject to higher levels of familial controls, but also reflects the differential impact of these controls on girls versus boys. For instance, even though girls are more closely supervised, boys' delinquency appears to be influenced more strongly by supervision and monitoring; by contrast, girls' delinquency is influenced more strongly by emotional bonds to their families (Heimer and De Coster 1999; see also Hagan 1989). In short, it may be that parents control girls' misbehavior through subtle, covert control mechanisms (e.g., emotional bonding), whereas controlling boys' misbehavior requires more direct, overt strategies (e.g., supervision).

Another individual-level factor that has received attention is gender roles. The argument, related to the liberation thesis discussed above, is that feminine roles restrain delinquency more than masculine roles do (e.g., Shover and Norland 1979). Many studies of

gender roles and delinquency appeared in the late 1970s. However, the empirical research was inconsistent, perhaps because gender roles were operationalized in diverse ways across studies, often including traits, attitudes, and familial controls as well as behavioral expectations or roles.

The more recent trend has been to focus on cultural definitions of gender or hegemonic gender definitions, which are widely-accepted beliefs and attitudes that support the subordination of females to males under patriarchy. Hegemonic gender definitions depict a feminine ideal that includes a high capacity for nurturance, passivity, connectedness to others, and physical weakness; by contrast, the masculine ideal emphasizes competitiveness, aggressiveness, independence, and strength. Definitions of femininity thus are more inconsistent than definitions of masculinity with victimizing others physically or taking their property. Some empirical research shows that acceptance of these gender definitions is an important contributor to the gender gap in law violation (Simpson and Elis 1995, Heimer and De Coster 1999). Other research suggests that boys who have internalized hegemonic definitions of masculinity are more likely to engage in crime and violence (Messerschmidt 2000).

A third individual-level factor that has received attention in the literature on gender and delinquency is attitudes toward risk-taking. The hypothesis here is that youths who develop a taste for risk will be more likely than others to engage in thrill-seeking behavior, including delinquency. The power–control theory of gender and delinquency, discussed below, proposes that one reason males are more likely than females to be delinquent is that males are socialized to prefer risk-taking (Hagan 1989).

5. *Bridging Levels of Explanation*

To date, most research on gender and criminal offending has focused on either the macro-level or social–psychological explanations discussed above, and rarely has pursued links between the two. One notable exception is Hagan's power–control theory, which argues that parents' positions in the workplace shape their parenting practices, including their supervision of children, emotional bonding with children, and the inculcation of taste for risk in children (Hagan et al. 1985, 1987). More specifically, when fathers have more power than mothers in their work or professional lives, mothers experience paternalistic control within the family. They then reproduce this arrangement with their children by subjecting their daughters to higher levels of control and by encouraging taste for risk more among their sons. When parents have more equal power in their work lives, female and male children are treated more equally. Thus the gender gap in delinquency will be smaller in the latter type of family. Research supports many of these arguments, although there is some debate over some of the specific predictions (e.g., Jensen and Thompson 1990, Morash and Chesney-Lind 1991).

Another emerging perspective on gender and crime with the potential to bridge levels of analysis is the work of Messerschmidt (e.g., 1993, 2000). He argues that males may attempt to display masculinity through crime and violence when legitimate channels for claiming their gender are blocked by structural barriers, such as unemployment and poverty (Messerschmidt 1993). Messerschmidt (2000) further explores links between structural barriers and social–psychological mechanisms in his analysis of boys' violence. He argues that much of this violence is a result of the interplay between patriarchal structures, social class, and social–psychological processes, such as alienation at school, victimization, body image, and hegemonic masculine ideals. Although this line of work does not address female offending, it moves toward bridging levels in analyzing masculinity and crime, and suggests a pathway for a similar analysis addressing femininity and crime.

Some of the qualitative research discussed in Sect. 3 also could serve as a springboard for the development of theories that bridge the macro- and social–psychological levels of analysis. Many of these studies examine how economic hardship and cultural contexts influence individuals' perceptions and decisions about crime (e.g., Miller 1986, Maher 1997, Baskin and Sommers 1998). The next step would be to account for these findings by specifying a theory of the links between the macro- and individual-level mechanisms leading to women's crime, and then broaden this explanation to address gender differences in crime.

6. *Directions for Future Research*

Up to the end of the twentieth century the strands of research discussed above have remained relatively isolated from one another. In addition, research has been aimed more at uncovering empirical relationships than at building comprehensive theories of gender, crime, and delinquency. Future research will need to develop rigorous theoretical arguments that link levels of analysis, drawing links between the mechanisms identified by macro-level research on changes in the gender gap and individual-level survey research. Moreover, research must bridge the divide between quantitative and qualitative research, combining findings from both to generate more complete explanations of gender, crime, and delinquency.

See also: Crime and Ethnicity (Including Race); Crime, Geography of; Crime: Sociological Aspects; Crime, Sociology of; Criminology: Psychopathological Aspects; Delinquency, Sociology of; Gender and the Law; Gender, Class, Race, and Ethnicity, Social Construction of; Gender Differences in Personality and Social Behavior; Gender-related Development; Personality and Crime

Bibliography

Adler F, Adler H M 1975 *Sisters in Crime*. McGraw-Hill, New York

Baskin D R, Sommers I B 1998 *Casualties of Community Disorder*. Westview Press, Boulder, CO

Box S 1987 *Recession, Crime and Punishment*. MacMillan Education, London

Box S, Hale C 1983 Liberation and female criminality in England and Wales. *British Journal of Criminology* **23**: 35–49

Carlen P 1988 *Women, Crime and Poverty*. Open University Press, Milton Keynes, UK

Chesney-Lind M 1997 *The Female Offender: Girls, Women and Crime*. Sage, Thousand Oaks, CA

Daly K 1994 *Gender, Crime, and Punishment*. Yale University Press, New Haven

Fox J, Hartnagel T F 1979 Changing social roles and female crime in Canada: a time series analysis. *Canadian Review of Sociology and Anthropology* **16**: 96–104

Hagan J 1989 *Structural Criminology*. Rutgers University Press, New Brunswick, NJ

Hagan J, Gillis A R, Simpson J 1985 The class structure of gender and delinquency: toward a power control theory of common delinquent behavior. *American Journal of Sociology* **90**: 1151–78

Hagan J, Simpson J, Gillis A R 1987 Class in the household: a power control theory of gender and delinquency. *American Journal of Sociology* **92**: 788–816

Heimer K 2000 Changes in the gender gap in crime and women's economic marginalization. In: *Criminal Justice 2000. Vol. 1. The Nature of Crime: Continuity and Change*. National Institute of Justice, Washington, DC, pp. 427–83

Heimer K, De Coster S 1999 The gendering of violent delinquency. *Criminology* **37**: 277–317

Jensen G J, Eve R 1979 Sex differences in delinquency: an examination of popular sociological explanations. *Criminology* **13**: 427–48

Jensen G F, Thompson K 1990 What's class got to do with it? A further examination of power-control theory. *American Journal of Sociology* **95**: 1009–23

Kruttschnitt C 1996 Contributions of quantitative methods to the study of gender and crime, or bootstrapping our way into the theoretical thicket. *Journal of Quantitative Criminology* **12**: 135–61

Maher L 1997 *Sexed Work: Gender, Race, and Resistance in a Brooklyn Drug Market*. Clarendon Press, New York

Messerschmidt J W 1993 *Masculinities and Crime: Critique and Reconceptualizan of Theory*. Rowman & Littlefield, Lanham, MD

Messerschmidt J W 2000 *Nine Lives: Adolescent Masculinities, the Body and Violence*. Westview Press, Boulder, CO

Miller E M 1986 *Street Woman*. Temple University Press, Philadelphia

Morash M, Chesney-Lind M 1991 A reformulation and partial test of the power-control theory of delinquency. *Justice Quarterly* **8**: 347–77

Shover N, Norland S 1979 Sex roles and criminality: science or conventional wisdom? *Sex Roles* **4**: 111–25

Simon R 1975 *Women and Crime*. Lexington Books, Lexington, MA

Simpson S S, Elis L 1995 Doing gender: sorting out the caste and crime conundrum. *Criminology* **33**: 47–81

Smart C 1976 *Women, Crime and Criminology: A Feminist Critique*. Routledge and Kegan Paul, London

Smith D A, Visher C A 1980 Sex and involvement in deviance/crime: a review of the empirical literature. *American Sociological Review* **45**: 691–701

Steffensmeier D J 1993 National trends in female arrests: 1960–1990: assessments and recommendations for research. *Journal of Quantitative Criminology* **9**: 411–41

Steffensmeier D, Allan E 1996 Gender and crime: toward a gendered theory of female offending. *Annual Review of Sociology* **22**: 459–87

K. Heimer and S. De Coster

Crime, Geography of

Crime is a form of deviance that involves the infringement of rules or laws that a specific society has created; both the act of a crime and the behavior of the criminal are therefore socially defined. The geographic study of crime draws a distinction between the geography of offences, where crimes are committed, and the geography of offenders, where criminals live. Different offences have different geographies. The places where shoplifting, robbery, and residential burglary occur are all different: the teenage mugger is likely to come from a different kind of neighborhood than the tax fraud. This diversity has stimulated study of the geography of crime at a variety of scales. Environmental criminology, with its focus on the places at which crimes occur, offers one organizing framework; the problematic concept of the 'crime area' another.

1. Origins

The origins of the geography of crime are found in early nineteenth century interest in cartographic criminology. A French statistician, Guerry (1833), produced maps showing the regional concentrations of crime. His observation that crimes of violence were disproportionately represented in the south of France led him to link criminal behavior and climate. In England, the social reformers such as Mayhew (1861-62) noted the strong clusters of crimes and criminals in specific areas of cities. Mayhew described the 'rookeries' of London where children were 'born and bred to the business of crime.' The Chicago studies of Shaw and McKay (1942) established a social science basis perspective on crime or delinquency areas, urban spatial patterns, and gradients and a set of 'theories' that underpinned the ecology of crime in the city. The basic premises were that crime had a spatial order; it was not distributed randomly but showed strong and persistent trends to cluster in poorer, disadvantaged environments. Subsequent studies have shown a remarkable persistence of these basic qualities of a geog-

raphy of crime and, further, have demonstrated that spatial order, clustering, and a link with 'place' typify related aspects of the criminal justice system such as sentencing patterns.

2. Information Bases

The geography of crime shares with criminology the problematic nature of the available data. Official crime statistics are suspect because the police record a relatively small minority of offences that actually occur and apprehend an even smaller number of criminals. Approximately 80 percent of offences that occur never enter police records. This 'dark area' varies by type of offence. It is very high for offences such as shoplifting or vandalism and low for car theft or homicide and arises from the unwillingness of the public, as victims or observers, to report crime to the police. Reasons for under-reporting range from embarrassment to fear of reprisal to apathy. Police themselves actually observe few criminal events and therefore rely on the public, broadly defined, to notify them when offences occur. Even the remaining 20 percent is further diminished by offences not proceeded with. This under-recording of crime may mean that official data contains biases and deficiencies. Does it over-record offences by disadvantaged youths? Does it under-record white-collar crime? Official crime statistics may reveal more about those who define criminality and administer the criminal justice system than about the actual incidence and seriousness of crime.

The official 'antidote' to these concerns about criminal statistics has been the introduction of national crime surveys. Dating in the main from the 1980s, these build upon an older *ad hoc* tradition of self-report studies and take the form of sample surveys of victimization experience and related attitudes towards crime, justice, law, and order. National crime surveys have confirmed under-recording and have qualified statements on rising crime rates, which may reflect differential recording practice rather than actual changes in crime. They have also augmented our knowledge of topics such as fear of crime. Whereas national crime surveys have not remedied the deficiencies of official criminal statistics, they have allowed their interpretation in more refined and constructive ways. For geographers, official statistics and national crime surveys are useful for studies of national and regional variations, but many geographies of crime focus on neighborhood variations and 'crime areas' within cities.

3. Theoretical Bases

The geography of crime draws upon the generic theories of criminology and social science and its own disciplinary concepts. Criminology has nurtured many theories and a significant lack of consensus. Most embrace a 'multifactor' explanation, which, despite diffuseness, is probably correct. There is no single causal explanation for criminality but a 'package' of inter-related conditions of varying applicability. Structural theories see criminal behavior as the inevitable outcome of societies that create inequalities. Yet these macrotheories fail to account for the fact that some individuals who suffer from inequality will react criminally, whereas others will not. Geographers have favored those criminological theories that lend themselves to analyses that allow a focus on place. Social disorganization is one such theory. Originating in the Chicago studies of the 1920s and 1930s, this theory linked crime with areas that lacked social order, established values, leaders, and social networks. Comparisons were made with residential areas that did possess these qualities, were 'socially organized' and experienced low rates of crime. Subcultural theory, proposing that crime may be prevalent because it is part of the normal behavior pattern within specific groups, can be found in embryo form in Mayhew's study of London's rookeries. Although now argued most consistently in relation to social groups, subcultural theory can be articulated in terms of areas or neighborhoods. Social control theory shares common strands with social disorganization and subculture, emphasizing the positive impacts of social control mechanisms in reducing the incidence of criminal behavior. The mechanism may be good parental supervision, an effective school or teacher, or the effects of voluntary organizations or community groups. It is again a theory that can be adapted to a focus on place and has a central concern with the offenders and where they live.

Environmental criminology offers the closest approximation to a geographical 'theory' of crime. It emphasises the criminal act and the location at which it occurs, the 'where' rather than the 'why' of crime (Bottoms 1994). Place is central to an approach that acknowledges that each offence leaves a new point on the map and is a clue to the activity space of the offender. The approach draws in considerations of the routine spatial activities of offenders and victims, the concept of journey to crime, and related ideas of territoriality and surveillance. At the present time environmental criminology is more a methodology than a theory. As such it draws heavily on basic geographical concepts such as space and place, and lends itself to detailed spatial analysis through geographical information systems. Conceptual issues include the problem of attaching any independent significance to space and understanding the full range of meanings attached to place.

4. Approaches to the Geography of Crime

Studies that merit the description of a geography of crime have changed along with broader paradigm shifts within the social sciences. Cartographic crimi-

nology, basically concerned with the mapping of criminal statistics and the representation of spatial patterns, has evolved to reflect wider use of statistical methods including the sophisticated technology of geographical information systems (GIS). Police forces recognize the value of GIS as a part of crime data analysis and the geocoding of criminal events for computer storage has become standard practice. There are clear advantages in the rapid portrayal of crime patterns and the related ability both to measure the features of the spatial pattern and to explore correlation with other environmental indicators. One police analysis of 79 crimes involving a single serial rapist was reported as involving 790,000 calculations using probabilities of journey to crime to link all points on a grid to the known crime sites (Grescoe 1996). Research into the statistical correlates of area crime rates has progressed through a wide range of single and multivariate analyses. The areal units used in such analyses include nations, police force areas, urban car beats, and census tracts; each raises methodological issues of data reliability and measurement. Scale effects, the modifiable areal unit problem, and 'ecological fallacy' continue to qualify exercises of this kind. Broadly, the outcomes of these analyses show a high level of consistency revealing strong correlations with indicators of a poor residential environment. Measures such as high unemployment rate, low income, substandard housing, and low educational attainment recur and, less frequently, indicators such as lone-parent families and broken homes. Correlates of this kind have been interpreted as indicative of some more general condition such as social disorganization or weak social control, and particular significance is often attached to single measures such as high unemployment. For the most part, however, the results of this type of analysis remain descriptive; they have value in identifying groups and areas from which offenders are likely to come but are in effect markers of some more general social condition as is the crime rate itself.

Statistical analyses have also been used as a first stage in the study of crime or delinquency areas. After identifying clusters of offenders and the conditions under which they live, the research task then is to probe more deeply with different research methodologies into the nature of the 'crime area.' One approach has identified particular neighborhood characteristics that might predispose residents to criminology. Herbert (1976) compared 'crime' and 'noncrime' areas in a city, controlling for sociodemographic profiles, and found a consistent difference in terms of sets of values and forms of behavior between the two types of area. In the 'crime' areas, parents adopted different approaches to control, had a clear tendency to accept some forms of misbehavior, and the children were lower achievers educationally. Other neighborhood studies have focussed on specific issues such as lack of parental supervision. More recently, what might be termed 'crime area' studies have taken a known problem area and sought to unravel both its origins and present form. In the United Kingdom, the so-called 'problem housing estates' lend themselves to this kind of analysis. Such estates often originated as local housing managers 'dumped' known problem families or relocated a slum clearance population to a specific estate. Early episodes of misbehavior reinforced labeling and reputations that have persisted. The Meadowell Study (Barke and Turnbull 1992) demonstrated a life-style that was permeated by welfare dependency, drug-trading, petty crime, and vandalism. Often self-selection of tenants who opt for this kind of lifestyle builds the area profile.

Understanding why criminal events occur where they do requires different approaches. Criminals respond to opportunities or targets in the environment and the task is to identify these 'cues.' A large majority of such crimes are opportunist and involve minimal planning; a minority is 'professional,' involving experienced criminals who may spend significant amounts of time preparing for a single criminal event. It is the former that lends itself to analyses based on opportunism and environmental cues. What makes some areas or residential units more vulnerable to crime? First, areas that offer more obvious or lucrative targets for a crime such as burglary are not likely to experience most offences. In part this is a function of security, wealthier neighborhoods are able to defend themselves with security devices and forms of control such as block or neighborhood watch. Research also shows that burglars tend not to travel far to commit an offence (Herbert and Hyde 1985). The offender–residence hypothesis suggests that vulnerable areas are either the areas or are close to areas where offenders live. Neighborhoods have 'cores' and 'borders' and the border-zone hypothesis suggests that the borders are more vulnerable. The local social control hypothesis suggests that ordered, well-organized neighborhoods with a strong sense of community identity experience less crime; the area variability hypothesis suggests that 'mixed' areas experience more. Hypotheses of these kinds, focusing on the places at which criminal events occur, blend into the general approach of environmental criminology. So too do studies of incivilities or signs of disorder in the urban environment (Herbert 1993) and the idea that these may indicate a deteriorating situation in neighborhoods. The concept of a 'community career' suggests that some localities move through a form of life cycle with highs and lows of criminal activity (Bottoms et al. 1992).

As criminals react to cues in the environment so do potential victims. Fear of crime arises from the perception that parts of the city, types of individuals or groups, and times of the day, are dangerous. People will modify their behavior to cope with fear and to achieve feelings of community safety. Early studies identified residents' perceived stress areas and showed that these were avoided in everyday routines. Fear of

crime is known to be high among old people and women with young children (Valentine 1997). Such groups experience less crime than youths but adopt avoidance strategies to reduce risk. Police have responded with initiatives such as 'do not speak to strangers' to protect children, but part of the problem may lie in the design of city environments and conventional attitudes towards time and place.

The step from environmental criminology to 'designing out crime' and crime prevention has proved a short one. In his thesis on 'defensible space,' Newman (1972) argued that places became more vulnerable if they had bad design features, lacked a sense of territoriality, had poor surveillance, and were stigmatized by their externalities. If design could incorporate the positive qualities, it could create defensible space and become less vulnerable to crime. The most comprehensive test of defensible space principles was provided by Coleman's (1985) study of housing projects in inner London. She developed a 'disadvantagement' score that was made up of measures on a range of indices from presence or absence of overload walkways to alignments of doors and windows. Coleman demonstrated close correlations of disadvantagement with crime and argued that less disadvantagement meant less crime. Although criticized for neglect of social dimensions, this work had a high impact on policy and on perceived community safety and quality of life. More restricted research designs, such as those to assess the impact of improved street lighting or CCTV camera surveillance, have produced similar positive results. The caveat with such measures is that instead of reducing crime they can simply displace it to other less well-protected areas. The ultimate in designing out crime is evident in rich suburban enclaves where high technology surveillance devices, walls, gates, and guards are used to protect residents. Davis (1990) wrote of fortress Los Angeles, divided between the protected enclaves of the very rich and the territories of the criminalized poor. Geographical research on social conditions also translates easily into crime prevention policies. Block or neighborhood watch is a direct transfer of ideas on community, surveillance, and social control.

The geography of crime has extended to consideration of policing, sentencing, and the criminal justice system. Early studies showed that sentencing was harsher in certain states and localities: where one committed a crime affected the sentence received. At international scale there are clear differences; Islamic law, for example, punishes an offence differently from European law. Within the United States, Florida tends to have more prisoners on death row and more executions (Harries and Brunn 1978). Studies show that jury composition affects verdicts (Herbert 1994). Magistrates drawn from one class judge offenders from another. Policing as an activity has strong spatial connotations. Police carry images of the city, derived from their experience and passed-on wisdom; these influence the ways in which they police the city. As police patrol areas, they need a detailed understanding of those places; the distributions in space of offenders and offences are key considerations for them. The police often focus their resources on specific areas and/or crimes for limited periods both in terms of community policing initiatives and crime prevention or target hardening. Geographers have studied policing initiatives such as police consultative committees and the ways in which they are constructed. Studies of the so-called 'victimless crimes' of prostitution (Shumsky and Springer 1981) suggest that policing practice and the variable workings of the criminal justice system shape their geographies. Beyond policing and the courts is the penal system, and as prison populations rise, their distribution and access are significant.

5. *Future Directions*

The geography of crime started from a few simple observations of spatial distributions and extended to a variety of applications of basic disciplinary concepts of space, place, and environment. Current thrusts involve developing pattern analysis through sophisticated information technologies, focusing on the meanings and relevance of place, and examining the interfaces between crime, criminals, and the criminal justice system.

See also: Crime and Class; Crime and Ethnicity (Including Race); Crime: Sociological Aspects; Crime, Sociology of; Delinquency, Sociology of; Police, Sociology of; Policing; Violence: Public

Bibliography

Barke M, Turnbull G 1992 *Meadowell: The Biography of an Estate with Problems*. Avebury, Aldershot, UK

Bottoms A E 1994 Environmental criminology. In: Maguire M, Morgan R, Reiner R (eds.) *The Oxford Handbook on Criminology*. Clarendon Press, Oxford, UK, pp. 585–656

Bottoms A E, Claytor A, Wiles P 1992 Housing markets and residential community crime careers: a case study from Sheffield. In: Evans D J, Fyfe N R, Herbert D T (eds.) *Crime, Policing and Place: Essays in Environmental Criminology*, Routledge, London, pp. 118–44

Coleman A 1985 *Utopia on Trial*. Shipman, London

Davis M 1990 *City of Quartz: Excavating the Future of Los Angeles*. Verso, London

Grescoe T 1996 Murder he mapped. *Canadian Geographic* **September/October**: 49–52

Guerry A M 1833 *Essai sur la Statistique Morale de la France avec Cartes*. Crochard, Paris

Harries K D, Brunn S D 1978 *The Geography of Laws and Justice*. Praeger, New York

Herbert D T 1976 The study of delinquency areas: a social geographical approach, *Transactions of the Institute of British Geographers* **NS 1**: 472–92

Herbert D T 1993 Neighborhood incivilities and the study of crime in place. *Area* **25**: 45–54
Herbert D T, Hyde S W 1985 Environmental criminology: testing some area hypotheses. *Transactions of the Institute of British Geographers* **10**: 259–74
Herbert S 1994 Venue v vicinage: the siting of trials and the pursuit of justice. *Urban Geography* **15**: 411–4
Mayhew H 1861–62 *London Labour and the London Poor*. Griffin, London, Bohn
Newman O 1972 *Defensible Space*. Macmillan, New York
Shaw C R, McKay H D 1942 *Juvenile Delinquency and Urban Areas*. University of Chicago Press, Chicago
Shumsky N L, Springer L M 1981 San Francisco's zone of prostitution. *Journal of Historical Geography* **7**: 71–89
Valentine G 1997 My son's a bit dizzy and my wife's a bit soft: gender, children and the culture of parenting. *Gender, Place and Culture* **4**: 37–62

D. Herbert

Crime: Knowledge about and Prevalence

Systematic knowledge about crime and its prevalence relies largely on legal definitions of crime, which—like commonsense definitions—are socially constructed. After describing sources of information on crime, this article summarizes what is known about its historical and social distribution, and how this distribution has been explained.

1. *Defining Crime*

In popular discourse, crime is behavior that violates fundamental moral precepts and deserves to be denounced and punished. This commonsense definition implies that certain behaviors are inherently criminal and necessarily invite sanctions. Because moral precepts, as well as responses to their violation, vary substantially over time and place, commonsense definitions of crime cannot denote a fixed and natural category of behaviors.

In legal discourse, crime is behavior which leads to particular legal procedures and for which the law permits imposition of criminal penalties. Here crime is constituted by legal reactions, not by characteristics inherent in behaviors. Legal definitions of crime appear to be both more precise and more objective than commonsense definitions: crime is what the law defines and treats as such. Because legal systems are products of particular historical epochs, and legal definitions are created and applied through social and political processes, crime in this definition is both historically and socially constructed.

Social and behavioral scientists who study crime typically rely on legal definitions, and restrict their attention to acts defined as violations of the criminal law. They differ, however, on the questions they ask about crime. Some are interested primarily in behaviors defined as crime and people who engage in them, whereas others are interested in the processes through which particular behaviors and people are defined as criminal.

Determining how much crime occurs and explaining why people commit it have traditionally been the principal concerns of those who study crime. Because there is no common behavioral attribute that ties all crimes together, the task of explanation becomes quite complex. How, for example, can one account for behaviors as diverse as possession of heroin, price fixing, vandalism, bribery of public officials, and spouse assault? Each of these violates the criminal law of many countries.

The difficulty of explaining crime as behavior suggests why other researchers direct their attention to the social and political processes whereby acts and people are criminalized. Here, crime is both a product of, and a discourse used in struggles over how social order is defined and achieved; its lack of any essential attribute highlights its ideological character (van Swaaningen 1997). This approach, however, 'implies a virtually unlimited legislative power to create crimes without regard to any restraining principles' (Hughes 1983, p. 301) and neglects the near-universal criminalization of certain behaviors, such as murder, assault, theft, and arson.

Notwithstanding the conceptual challenges of defining crime, the combination of two elements is ordinarily recognized as critical: human behavior, and a state-sanctioned response to it. Crimes of greatest concern in most societies also involve a third element: *intention* to harm or wrong another. The more serious of these crimes—what are called *mala in se* (acts wrong in themselves)—are the target of substantial measurement efforts designed to answer questions about their prevalence, distribution, and causes.

2. *Measuring Crime*

Efforts to measure crime are usually guided by legal definitions within the jurisdiction of interest and rely on crime reports of legal officials. Most nations compile annual reports on the number of crimes known to police and the number of persons proceeded against. There are two principal difficulties with this approach to measuring crime. First, while legal definitions and official statistics provide a reliable guide to what is designated as criminal behavior *within* a given society, they may not measure accurately differences *across* societies in similar types of behavior. This is because legal definitions of even relatively serious crimes vary across societies.

The second difficulty is that, for a number of reasons, official statistics do not measure accurately

levels of crime, even within a single society. Characteristics of official agencies that collect information on crime, such as the police, inevitably influence how these acts are interpreted and recorded. Some acts that fit legal definitions of crime may be subject only infrequently to criminal enforcement, such as many white-collar or corporate crimes. Since the public is the primary means by which officials learn about crime, popular views on crime and the agencies that deal with it also influence which acts are reported to legal officials. So, for example, well under a half of all physical and sexual assaults are reported to police in most nations (Mayhew and van Dijk 1997). As a result, official statistics on crime do not reflect its true prevalence.

Unofficial sources of crime data, such as surveys of the general population, avoid some of these problems and shed light on the 'dark figure of crime'—that is, crimes that do not come to the attention of legal officials. Victimization and self-report surveys, which ask people about their criminal victimization, or offending, over a specified period of time, yield higher estimates of the prevalence of crime than do official statistics. Victimization surveys also show that the likelihood that a crime will become known to officials depends in part on the characteristics of the victim, the offender, and their relationship; and that the more serious the crime, the more likely it is to be reported to officials. But because they rely on the honesty, memory, and perceptions of respondents, such surveys have their own limitations and biases.

All efforts to measure crime inevitably confront the problem of *defining* a crime: someone—a victim, a police officer, a witness, or a researcher—must apply a legal definition to a complex set of circumstances and human relationships. The difficulties in measuring crime, then, are not simply technical or statistical, but conceptual—and unavoidable. Consequently, it is impossible to determine the total amount of crime in any jurisdiction at any time. One can, however, count and describe with some accuracy events which are crimes according to the law, and which come to the attention of legal officials, as well as criminal events which people report to survey researchers. These estimates form the basis of knowledge about the prevalence and distribution of crime over time, place, and persons.

3. *The Distribution of Crime*

Is there more crime now than in the past? For reasons outlined above, this question has no simple answer. To the extent that contemporary criminal law defines a wider range of behaviors as criminal, there is likely to be more crime. The most reliable historical evidence pertains to serious violent crimes, such as homicide, robbery, and serious assault, in Western Europe and North America. Their trends over the last six or seven centuries have undoubtedly been downward, albeit punctuated by short-term waves of violent crime after wars and other major social dislocations (Gurr 1989).

This long-term decline reversed somewhat in the late twentieth century. In the early 1970s, rates of serious violent crime rose and then plateaued in most Western nations but in many Eastern European nations, violent crime rates appear to have increased dramatically since the late 1980s. Nevertheless, violent predation has become a much less customary part of daily life over the long term. Whether the same can be said of property crime is unclear, because reliable historical data do not exist. But the best evidence suggests that, as societies developed, property crime rates increased, in part because consumer goods were more available and more portable.

Are some societies more crime-prone than others? Cautious comparisons can be made of direct-contact, predatory crimes by individuals using data from international victimization surveys (Mayhew and van Dijk 1997), and public health statistics on homicides (World Health Organization 1997). According to these data, rates of serious violent crime vary greatly, even across relatively similar societies. At the end of the twentieth century, for example, homicide rates were four times higher in the USA than in Canada, three times higher in Finland than in Norway, and four times higher in Venezuela than in Chile. Among advanced capitalist nations, homicide rates are higher in the New World countries (the USA, Canada, and Australia). Countries in North Africa and the Middle East tend to have relatively low rates of serious violent crime, as do many Asian countries (Heiland et al. 1992).

Homicides and other serious violent crimes make up only a small part of the total crime rate, however. In all nations, most crimes are crimes against property; national differences are much less pronounced for property crimes than for serious, violent crimes, especially in advanced capitalist countries. While the property crime rate is not necessarily a good predictor of the violent crime rate, nations with high rates of property crime are unlikely to have low rates of violent crime. Although there is little research and almost no systematic data on national differences in white-collar, corporate, and political crimes—for example, theft by corporations, environmental destruction, or illicit violence by state agencies—these appear to be more common in developing countries. In part, this results from lax enforcement and the ability of offenders to avoid prosecution. For similar reasons, activities of criminal organizations, such as trafficking in drugs, people, and weapons, money laundering, and cross-border terrorism, may also be more prevalent in countries with emerging or transitional economies.

As this suggests, the economic and political organization of a nation produces particular patterns of crime. For example, although the pre-1990s socialist countries of Eastern Europe had low rates of violent

and predatory street crimes, other behaviors—worker sabotage, theft of nationalized property, bribery, and black market activity—were relatively widespread. Violent crime rates are higher in nations with greater inequalities in income and wealth, and in nations with fewer restrictions on the ownership and use of firearms. Nations that go to war more often also tend to have higher homicide rates during peacetime.

Many of the social processes, conditions, or policies commonly thought to affect crime rates often do not do so, or affect only some types of crime. Urbanization and unemployment, for example, typically are not accompanied by higher rates of violent crime, although they are associated with higher rates of property crime (see *Crime and Class*). Conversely, some social policies that are intended to decrease crime in fact do not. For example, there is little evidence that the death penalty deters homicides (see *Death Penalty*). During the twenty-first century, the social process likely to have the greatest impact on crime—from street crimes to crimes of transnational corporations—is globalization. 'Globalization,' according to Findlay (1999, pp. 1–2), 'creates new and favourable contexts for crime' and enables 'the transformation of crime beyond people, places and even identifiable victims.'

Are some people more likely to engage in crime than others? Once again, the heterogeneity of the acts comprising crime demands a cautious answer. Serious, direct-contact, predatory crimes are the domain of young, economically disadvantaged males in virtually all nations. White-collar, corporate, and political crimes are also male-dominated, though these are, for obvious reasons, rarely the preserve of the young and the poor. It is largely for victimless crimes, such as prostitution, gambling, or drug use, that women's involvement at times rises to the level of men's (see *Crime and Gender*).

4. *Explaining the Distribution of Crime*

Most explanations for the distribution of crime feature one of three mechanisms: motivations, controls, or opportunities.

Motivational explanations focus on social conditions or processes that generate criminal tendencies in societies, groups, or individuals. Criminal motivations have their roots in both social structure—institutional patterns and processes that organize interactions and relationships—and in culture—accumulated traditions and shared experiences that shape the values of a society or group. Economic inequality is a structural factor thought to increase motivations for crime by fostering alienation and hopelessness (Hagan and Peterson 1995). Norms that advocate criminal acts under certain conditions are a cultural factor that may motivate criminal behavior.

Control explanations focus on social conditions and processes that weaken internal and external, informal and formal controls on crime. Both structural and cultural factors can affect social control (Sampson and Raudenbush 1999). Crime tends to be lower in societies with stable families and integrated community networks because these strengthen controls on criminal behavior. Crime tends to be higher where group norms excuse crimes against certain victims (e.g., against spouses, organizations, sexual minorities, etc.) because these cultural features weaken controls.

While motivational and control perspectives are concerned primarily with potential offenders, opportunity explanations focus on the circumstances in which crimes occur. Situations in which a motivated offender encounters a suitable target in the absence of capable guardians create opportunities for crime (Cohen and Felson 1979). Patterns of routine daily activity that take people away from home foster opportunities for burglary, mugging, and vehicle theft. Criminal opportunities are also greater where small consumer items are mass-produced.

No general theory of the distribution of crime has been tested adequately enough to command widespread support. Indeed, there is still considerable debate over whether a general explanation of crime is possible or even desirable, given the heterogeneity of crime. Nevertheless, some generalizations about the sociocultural contexts that appear to discourage direct-contact, predatory crimes can be drawn. These are characterized by: (a) strong systems of informal social control reinforced by highly consensual normative systems, (b) strong networks of communal obligation and mutual interdependence, and (c) cultural orientations that discourage interpersonal aggression. Societies with these characteristics tend to be organized around collectivist rather than individualistic principles, which are reflected in their political, economic, and legal ideologies and practices (Pampel and Gartner 1995).

These generalizations are limited by their grounding in research on a restricted range of criminal behaviors. Too little is known to be able to develop similar generalizations about many types of crime, such as those perpetrated by transnational corporations and national governments. Crimes that cross national borders have also been neglected by researchers, although this is changing. New forms of crime emerging in some societies—cyber-crimes, for example—also need to be studied. Until documentation of these crimes is taken seriously, understanding of the prevalence and distribution of crime will be incomplete.

See also: Crime and Class; Crime and Delinquency, Prevention of; Crime and Ethnicity (Including Race); Crime and Gender; Crime, Geography of; Crime: Organized; Crime: Sociological Aspects; Crime, Sociology of; Crime: White-collar; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Statistical Systems: Crime

Bibliography

Cohen L, Felson M 1979 Social change and crime rate trends: A routine activity approach. *American Sociological Review* **44**: 588–608

Findlay M 1999 *The Globalisation of Crime*. Cambridge University Press, Cambridge, UK

Gurr T R 1989 Historical trends in violent crime: Europe and the United States. In: Gurr T (ed.) *Violence in America: The History of Crime*. Sage, Newbury Park, CA, pp. 21–54

Hagan J, Peterson R D (eds.) 1995 *Crime and Inequality*. Stanford University Press, Stanford, CA

Heiland H-G, Shelley L I, Katoh H (eds.) 1992 *Crime and Control in Comparative Perspective*. Walter de Gruyter, New York

Hughes G 1983 Concept of crime. In: *Encyclopedia of Crime and Justice*. The Free Press, New York, pp. 294–301

Mayhew P, van Dijk J J M 1997 *Criminal Victimization in Eleven Industrialized Countries: Findings of the 1996 International Crime Victims Survey*. Ministry of Justice, WODC, The Hague, The Netherlands

Pampel F N, Gartner R 1995 Age structure, socio-political institutions, and national homicide rates. *European Sociological Review* **11**: 243–60

Sampson R J, Raudenbush S W 1999 Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American Journal of Sociology*. **105**: 603–51

van Swaaningen R 1997 *Critical Criminology: Visions from Europe*. Sage, London

World Health Organization 1997 *World Health Statistics Annual: Vital Statistics and Causes of Death*. WHO, Geneva, Switzerland

Zvekic U, Alvazzi del Frate A 1995 *Criminal Victimization in the Developing World*. UNICRI, Rome

R. Gartner

Crime: Organized

Relative to other kinds of crime groups (e.g., a burglary crew), law enforcement practitioners and social scientists typically define an organized crime group as follows: (a) it has some manner of formalized structure whose primary objective is to obtain money through illegal activities; (b) it has continuity over time; (c) it maintains its position through use of violence or threat of violence, corruption of public officials, and extortion. Additional important elements may include: (d) its operations tend to be interjurisdictional; (e) key members are bufffered by forms of organizational structuring and by the use of 'fronts' and nonmember associates; and (f) long-term goals stress the maximization of profit and power through monopolization, or cartelization, of criminal enterprises and markets, as well as control over legitimate (noncriminal) businesses and sectors of the economy (Lupsha 1995).

Organized crime groups may perpetrate almost any kind of swindle, theft, or illicit activity imaginable. But the two main types of activities involve the provision of illicit goods and services like illicit drugs and gambling, and the infiltration of legitimate business (including the illegal manipulation of labor unions).

Most organized crime is never detected and so its costs to society are unknown. It generally is recognized, however, that the costs of organized crime in terms of illicit profits, unpaid taxes, disruption of legitimate business, and political and commercial corruption far exceed the costs incurred from conventional crime and street crimes.

1. Scope and Structure of Organized Crime Groups

The size, structure, and cohesion of criminal organizations has been the subject of much debate. At one end of the spectrum, organized crime is described in terms of large hierarchical organizations that are structured rather like traditional corporations. At the other end, organized crime is described in terms of small, fluid networks that are rather amorphous in structure and are characterized by considerable opportunism.

This dichotomy is unnecessary, however. Organized crime as a concept is a matter of degree, not a matter of kind. Criminal organizations fall along a continuum from large to small and from formal hierarchical (even bureaucratic) structure to fluid network organizations (Steffensmeier 1995). Criminal organizations falling at either end of the continuum may be equally adaptable to the exigencies of the illicit marketplace. Groups may combine elements of a formal structure at certain levels, with more amorphous and flexible networks at the lower levels. The Colombian cartels and the US Mafia, for example, have demonstrated remarkable adaptability in spite of their large size and organizational sophistication.

In addition to size and degree of formal structure, criminal organizations can be identified and compared relative to their scope of criminal activities. Some groups are highly specialized and focus on one kind of activity, such as prostitution or drugs. Other groups engage in a broader range of criminal activities such as gambling, extortion, drug trafficking, financial scams, and infiltration of legitimate businesses.

Criminal organizations or networks also vary in terms of their geographic scope—local, national, or transnational. A powerful transformation is occurring whereby organized crime increasingly is becoming transnational in character. Less and less are the criminal activities of criminal organizations constrained by national borders. Further, organized crime groups which are still predominantly domestic in scope can be expected in the future to increase their exploitation of global opportunities.

A final dimension concerns the level or scale of impact on the citizenry and on the political and economic structure of a nation. The US Mafia and the

major transnational crime organizations represent what has been termed large-scale organized crime, to refer to those powerful criminal groups 'able to engage in political and economic activity both through legitimate official structures and by illegal means, and which can adversely affect whole areas of productive, social and institutional life' (Arlacchi 1986, p. 213).

2. *Organized Crime Groups/Networks in the USA*

The bulk of research and theory about organized crime has centered on the USA, where four main forms of organized crime networks now exist: the US Mafia, emerging or previously ignored crime groups, local racketeering syndicates, and transnational criminal organizations (Steffensmeier 1995). These forms are analytically distinct but in practice may merge one into another.

2.1 *The US Mafia*

Italians are but one in a long succession of groups to become involved in organized crime in the USA. The Irish were the first group to become involved on a large scale, followed by Eastern European Jews, and then by Italians. In recent years Hispanic and Chinese immigrants have made a place for themselves in organized crime. But Italian–American organized crime, of which Cosa Nostra is the archetype, has achieved broader dominance over a wider range of criminal activities than that of any other ethnic group, and for a longer period of time.

At the beginning of the twenty-first century, about two dozen Cosa Nostra families are centered in urban areas with relatively large Italian–American populations, such as New York and Chicago. The networks of these families are expansive, moreover, so that localized Mafia groups and Mafia-connected criminal activities exist in a substantial number of US cities. Legislation used to counter organized crime, notably the Racketeer Influenced Corrupt Organizations (RICO) Act, was designed with Cosa Nostra in mind and has enabled law enforcement to develop cases against Mafia groups and other crime organizations as criminal enterprises (Jacobs 1994). Until about the 1970s, 'Cosa Nostra' and 'organized crime' were considered to be synonymous terms by many law enforcement authorities.

Newer evidence confirms and refines the picture of Cosa Nostra provided by earlier sources (see Cressey 1969). Cosa Nostra exists; possesses a hierarchical chain of command; is self-perpetuating and has an overarching structure which is independent of criminal activity; and becoming a 'made' member involves initiation rites and pledges of loyalty, secrecy, and violence. A Cosa Nostra 'family' operates autonomously in many US cities, but individual groups are joined together in a crime confederation in which a limited amount of authority has been ceded to a governing national commission. The commission regulates interfamily conflicts but does not attempt to run the day-to-day operations of the syndicates.

The bureaucratic analogy sometimes applied to Cosa Nostra, however, has been overstated. Its structure is looser and more decentralized than the organizational charts imply (Abadinsky 1999). Day-to-day operations of organized crime are often quite haphazard, both because of inadequate managerial skills on the part of many participants and because of the need for secrecy—which sets limits on managerial control. Relationships between family hierarchy and members, as well as associations with other family groups, are flexible and fluid. Individual members have sufficient autonomy to build their own power bases and connections. They make their own deals with other criminals, businesses, labor leaders, and government officials based on potential profits or services required.

It appears that the criminality and influence of a number of Cosa Nostra families is less at the beginning of the twenty-first century than it was in the early 1990s. Factors contributing to this decline include intense federal enforcement efforts directed at the Mafia; changing illicit markets (especially drugs); advancing age and attrition in membership; the decline of ethnic communities in urban areas; the shift of political power from cities to the federal government, including the decline in the local 'political machine'; and the decline of unions. These together have reduced three main historical assets of the Mafia—centralized corruption of police and local officials, a disciplined workforce, and reputation (Reuter 1994). However, some recent prosecutions and law enforcement reports indicate that the increased law enforcement activity has had a tremendous impact on the hierarchy of the families but little effect on Mafia-related activity (Kelly 1999, Pennsylvania Crime Commission 1991).

2.2 *Previously Ignored and Emerging Crime Groups*

Besides traditional Italian organized crime, it is widely recognized that there exists in many US metropolitan areas a collage of groups—Asians, bikers, Blacks, Colombians, Dominicans, Jamaicans, Mexicans, Syrians, Lebanese, and Russians—involved in organized crime, notably in the drug trade (Pennsylvania Crime Commission 1991, Steffensmeier 1995). Recent immigration to the USA, in particular, has spawned ethnic crime groups who often direct much of their criminal activity against members of their own ethnic group and maintain ties to criminal groups in their homelands. Some of the nontraditional groups are highly organized, while others are not. While many of

these groups are considered 'new' or 'emerging' in that they have become active in recent years, other groups are 'new' in the sense that they have only recently been recognized as crime groups. The outlaw motorcycle gangs, for example, have been operating for many years in organized crime activities and are now well established.

2.3 Local Racketeering 'Syndicates'

As far back as 1951, the US Senate's Kefauver Committee concluded that much organized crime is localized and indigenous to the community in which it exists. A largely unrecognized but important dimension of the organized crime phenomenon is the significant and longstanding presence of local racketeering groups which dominate extortion and vice activities in many localities and cities throughout the USA, doing so both through legitimate official structures and by illegal means (Pennsylvania Crime Commission 1991, Steffensmeier 1995). Core 'membership' in these racketeering groups includes not only mafioso and professional criminals but also business people (e.g., car dealers, contractors, restaurant owners), police and city officials, attorneys, and other individuals from 'respectable' society who are influential in local politics.

These local crime syndicates generally represent distinct entities, although they often overlap and may even be fragments of a large-scale crime organization such as Cosa Nostra. The syndicates tend to be dominated by a handful of entrepreneurs in which the illegal activities of individual racketeers are carried out through a network of individuals associated with them in that activity. With this type of operation, it frequently is the activity rather than the particular people that fosters the relationship. A number of substantive criminal relationships or 'partnerships' may develop, linking joint criminal operations within a network and across other crime networks dominated by other local criminal figures. Overlapping partnerships for specific illegal ventures are stitched together through opportunistic and fluid alliances. In some instances there may be a central racketeering figure, but more likely there is shared power and a shifting power structure depending, in particular, on who at the time is most influential with local law enforcement and city officials (Steffensmeier 1986).

3. Transnational Crime Organizations (TCOs)

Criminal organizations, particularly in the USA and other rich countries, are increasingly engaging in activities that cross national borders (Martens 1991, National Research Council 1999, Williams 1999). Factors contributing to this development include the globalization of the economy and financial markets; the rise in the numbers and heterogeneity of immigrants, who bring with them social and commercial networks that make law enforcement more difficult and that facilitate conspiracy; improved communications technology, which makes borders more permeable and criminal activities less conspicuous; the end of the Cold War and triumph of capitalism, which have increased the porosity of national boundaries and created new markets for both licit and illicit commodities.

The term 'transnational' refers to the movement of information, money, physical objects, people, or other tangible or intangible items across national boundaries (Williams 1999). 'Transnational organized crime,' in turn, refers to large-scale crime organizations or networks that operate across national borders or that have a home base in one nation but that operate in other nations or states where there are favorable market opportunities. Their principal activities at the turn of the twenty-first century include: smuggling—in commodities, drugs, and protected species; contraband such as stolen cars and tobacco products; and services—in immigrants, prostitution, indentured servitude, and money laundering (see Kelly 1986, National Research Council 1999).

Transnational crime organizations (TCOs) vary considerably in the structure, geographic scope, and scale of their operations. Some are organized hierarchically, while others are fluid in structure. The TCOs typically operate out of 'national sanctuaries'—in areas where there is uneven justice and limited law enforcement capacities, and where they are relatively immune to law enforcement efforts. By acting transnationally, the groups gain access to lucrative markets for illicit commodities or services where the profits are very high and then channel the proceeds from illicit activities through the global financial system.

The most widely recognized transnational groups are: Chinese triads, Colombian cartels, Japanese yakuza, Sicilian Mafia, and some Russian criminal organizations (which vary greatly in organizational structure and sophistication). The Asian gangs and the Sicilian Mafia have long traditions as compared to the Colombian cartels and Russian gangs, which are relatively young.

The Sicilian Mafia, the most global and perhaps the most complex of the transnational organizations, actually refers to four distinct, often overlapping, groups. The most important of these groups is the Sicilian Mafia or Cosa Nostra, followed in order by the Neapolitan Camorra (a group of predominantly local organizations), the Calabrian N'drangheta (a federation of families involved in illegal drugs, contraband tobacco, kidnaping, and extortion), and the 'Sacra Corona Unita' in Apulia (based on breakaway factions from the Camorra and N'drangheta). Like the US Mafia, these groups entail a relatively simple hierarchical structure and are based on close associations that are partly functional, partly personal and

familial, and partly based on fear (Arlacchi 1986, Paoli 1999). Recent trials both in the USA and Italy (e.g., the 'Pizza Connection' case in 1986) have revealed the scope of the Sicilian Mafia's involvement in international drug trafficking (especially heroin and cocaine) and the considerable extent to which its members associate and interact in criminal activity with the US Mafia. Although its power base remains in Sicily and southern Italy, the Sicilian Mafia has increasingly become a transnational enterprise and its Mafia groups can now be found in virtually every country where there resides a moderately sized population from Sicily (Savona 1993, Sterling 1990).

The Colombian cartels are foremost in the drug business, in contrast to the Sicilian Mafia and most other transnational criminal organizations which engage in a range of illegal activity. The cartels—and particularly the Cali cartel that has become the predominant group in the cocaine industry—have also amalgamated corporate and criminal cultures more than any other group (Williams 1994). They resemble modern legitimate firms in utilizing business management principles such as elaborate staffing and division of labor, administrative oversight, strategic planning, and meticulous accounting procedures. The cartels have developed a very specialized cell structure to carry out activities such as logistics, marketing, and money exchange; the cell stucture also provides protection by disbursing information on a need-to-know basis (see Pennsylvania Crime Commission 1991).

The recent rise and expansion of major international crime networks represent the most important and threatening development in organized crime at the turn of the twenty-first century. First, very large fortunes have been amassed by some TCOs—reflecting both the extraordinarily large profits obtainable in the illegal economy and the ease with which money can be laundered or 'cleaned.' Second, the expanding demand for criminal labor generated by the increased scale of activities has fostered integrative links between the international mafias on the one hand with common and juvenile crime on the other (Arlacchi 1986). The changes toward an older profile and adult involvement in youth gangs in many US cities reflect this development. Third, the most powerful criminal groups (e.g., in Colombia, Russia, Sicily) have shown that they can influence society and economic processes to a greater extent than they are determined by them (Arlacchi 1986).

Fourth, the large-scale criminal organizations—through their huge economic resources and by means of private violence—are able to impose their power and to manipulate large sectors of the political system and the state apparatus. These groups not only arouse general fear and anxiety in many countries but also threaten the governability of individual countries and the world order more generally. The threat to democratic institutions and the erosion of government's ability in many parts of the world 'to govern, to provide even the basics of civil peace, economic opportunity, or justice' (Olson 1994, p. 1) are probably the most detrimental repercussions of the transnational criminal organizations.

4. *Roots of Organized Crime*

At least four main factors, singly or in combination, contribute to the rise and persistence of organized crime. First is a certain level of social-cultural disorganization (Arlacchi 1986). Important here are the consequences of (a) the industrialization and urbanization that have taken place over the last two centuries in countries in the West; (b) the so-called 'modernization' that is presently taking place in certain large Asian, Latin American, and African nations; and (c) the spreading democratization and free-market capitalism in the Eastern European nations. Their effects—growing social injustice, poverty, and violence—are a seedbed in particular for the recent emergence of large-scale criminal organizations at the global level.

A second factor is 'weak' government. Criminal organizations are more likely to emerge and flourish where the government is lacking in legitimacy or is not fully in control of the territory under their jurisdiction. Organized crime in many countries, as in Sicily, has been generated by despotic governments which give rise to groups formed for self-protection and for providing quasi-governmental services. These conditions also help to foster an ethos of cultural underpinnings for organized crime (Cressey 1969, Pennsylvania Crime Commission 1991, Steffensmeier 1995). Main tenets include distrust of outsiders and government, clannishness, proclivity for sanctioned violence and private revenge, a sense of how to coopt the law and achieve immunity by corrupting influential persons, and savvy to cultivate the goodwill of local residents while simultaneously exploiting them for personal advantage.

Third, organized crime arises because it is 'functional'—either to the community or large subgroups in the society (Cressey 1969). These functions are varied, but may range from providing desired but illicit goods and services, social control and protection (e.g., against certain predatory crimes) in areas where government or official agencies have failed, opportunities for political and economic participation, and welfare-type assistance (e.g., in finding employment, short-term loans to prop up precarious small businesses).

Fourth, there is an easy affinity between organized crime and market-based economies. Organized crime is about the exploitation of business or market opportunities. The necessity to maintain and extend one's share of the market governs entrepreneurship in organized crime as it also does in the legitimate market (Cressey 1969, Smith 1980).

By their very nature, many organized crime activities, especially financial crimes, require the tacit or open cooperation of legitimate organizations and actors. In the abstract, organizations could be perceived as either legitimate groups that commit no crime or as criminal groups with purely criminal purposes. In the real world, however, the activities of many organizations, both legitimate and criminal, fall between these extremes (Steffensmeier 1986, National Research Council 1999).

These factors together suggest that the organized crime phenomenon is primarily economic and institutional, rather than criminal (see Arlacchi 1986, Cressey 1969). Control strategies most likely to be effective will target the real structures and seats of power of organized crime—governmental weakness and official corruption, financial markets and accumulated capital, and the institutional degeneration that allows the problems to continue and grow worse.

5. *Ethnicity and Organized Crime*

Ethnicity affects organized crime in three related ways (Pennsylvania Crime Commission 1991, Steffensmeier 1995). First, organized crime is a multiethnic phenomenon in the sense that a wide variety of ethnic groups are involved. Second, ethnicity typically plays an important role in the formation and function of organized crime groups. Ethnicity contributes to the organization of crime, a feeling of camaraderie built on blood ties, a common cultural base, insularity, mutual trust, a distrust of government, and an exclusivity that reinforces bonds and secrecy within the organization. Third, in both the USA and other rich countries, foreign and immigrant criminal groups have become principal figures in transnational organized crime.

5.1 *Alien Conspiracy vs. Ethnic Succession*

The role of ethnicity in shaping a country's organized crime problem has been the subject of some dispute. The alien conspiracy view holds that organized crime doesn't so much emerge out of a country's culture or economic realities and politics, but is thrust upon it by particular ethnic, immigrant groups who bring forms of organized crime with them. In the USA, for example, the theory holds that a secret organization originating in Sicily made its way into the USA during the period of massive Italian immigration around the start of the twentieth century. Popular in the 1950s and early 1960s, the Mafia conspiracy view has since been widely rejected. In 1969 organized crime expert Donald Cressey stated that while 'the similarities between the two organizations are direct and too great to be ignored, the American Mafia is not merely the Sicilian Mafia transplanted' (Cressey 1969, p. 8).

A contrasting view characterizes organized crime not as a foreign import but as an expedient avenue for social mobility that is an integral part of a country's social and economic system (Ianni 1972). Ethnic succession refers to the process by which different ethnic immigrant groups have used, and continue to use, organized criminality as a means of assimilation and advancement. Using their illegal gains and as more legitimate opportunities opened, legitimate advances with all their advantages were pursued as alternatives to criminal activities.

The process of ethnic succession adequately describes the Irish, the Germans, and the Jews (who preceded the Italian migration into the USA), as well as some of the recent ethnically based crime groups both in the USA and elsewhere. But the persistence of the Italians in organized crime is at odds with ethnic succession.

6. *Alternative Models of Organized Crime Groups*

Recent writings on organized crime distinguish among three main paradigms of organized crime (Albanese 1998).

In the *hierarchical structure model*, organized crime is a more or less government-like structure with graded ranks of power or authority; organized illegal activities are conducted with the blessing of higher-ups, illicit activities are 'protected' through the influence of the hierarchy, and profits or 'tribute' are funneled to superiors (Cressey 1969).

An alternative approach has been to describe organized crime as a *network of patron–client relationships* wherein people's connections provide them with access to a particular illicit activity (Albini 1971, Ianni 1972). Relationships develop around a central figure, or patron, and the recipients of his favors who are dependent upon him for contacts, protection, and resources. In return, these clients offer loyalty and allegiance.

The *enterprise model* constitutes a third paradigm of organized crime. Organized crime is a 'business' activity that represents a deviant variation on legitimate activity (Smith 1980). The focus is on the 'enterprise' and its functional needs, such as revenue and investment requirements, rather than the group undertaking it. Notably, the size and structure of an organized crime group is basically determined by the activity in which it is involved (i.e., function determines structure).

These models are analytically distinct and offer competing pictures of reality, but are better seen as complementary. All organized crime is entrepreneurial and is shaped by market forces; moreover, both patron–client and kinship relations are typically key features of hierarchical crime organizations. One key difference is that the enterprise model focuses more on how organized crime activities are organized (e.g., a

drug smuggling operation, a bookmaking business), whereas the hierarchical model focuses more on how crime groups are organized.

7. *Success and Survival of Organized Crime Groups*

Five factors help explain why some crime networks, like the Cosa Nostra or Japanese yakuza groups, succeed and survive for a lengthy period of time (see Reuter 1994, Steffensmeier 1995). The first is historical circumstance—e.g., Italian mobsters arrived on the US scene at a critical time (i.e., Prohibition) and were able to entrench themselves in the rackets of urban USA. Once in place, they warded off competing crime groups through advantages derived from their dominant position.

A second factor involves a unique cultural ethos—a set of traditions and organizational principles supportive of organized crime and very adaptable. A key element of this ethos, anchored in kinship and ethnicity, is a bounded solidarity and enforceable trust that group members can rely on for illicit activities. Other elements include distrust of government, clannishness, and proclivity for sanctioned violence.

The third factor is reputation—a set of beliefs or capacities that represent advantages the mafiosi as entrepreneurs enjoys over their competitors. The set includes the credible threat of violence, the ability to neutralize law enforcement, and the wherewithal to 'get things done.'

A fourth factor is the ability to recruit a disciplined workforce and prospective leaders. Mafia groups have been able to draft, almost effortlessly, reliable and able recruits from within kin networks and Italian-Sicilian neighborhoods where Mafia-based values and traditions have been in force, and where an abundance of young 'toughs' covet the fear and respect that mafiosi still engender.

A final element is durability itself. A crime network (e.g., the Mafia) that has been around a long time gains credibility by showing it can withstand either efforts to repress it or the loss of members from time to time.

These five factors also can be treated as analytic devices for broadly evaluating the strength of organized crime groups and their 'peculiar threat' relative to each other and to other forms of crime. The more these conditions can be said to characterize a group, the greater its strength, staying power, and threat relative to other organized crime groups. Furthermore, today's aggressive law enforcement effort to dismantle established crime groups will be successful only if it greatly reduces the competitive advantages described above. Likewise, strategies aimed at preventing the emergence or stabilization of comparably powerful crime groups will be effective only if they prevent the build up of such assets.

See also: Control: Social; Crime: Sociological Aspects; Crime, Sociology of; Crime: White-collar; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Delinquency, Sociology of; Drugs: Illicit Use and Prevention; Drugs: International Aspects; Ethnicity, Crime, and Immigration; Gangs, Sociology of; Law, Sociology of; Mafia

Bibliography

Abadinsky H 1999 *Organized Crime*. Nelson-Hall, Chicago

Albanese J 1998 *Organized Crime in America*. Anderson, Cincinnati, OH

Albini J 1971 *American Mafia: Genesis of a Legend*. Appleton, Century Cross, New York

Arlacchi P 1986 *Mafia Business: The Mafia Ethic and Spirit of Capitalism*. Biddles, Guildford, UK

Cressey D 1969 *Theft of the Nation*. Harper and Row, New York

Ianni F 1972 *A Family Business*. Russell Sage Foundation, New York

Jacobs J 1994 *Busting The Mob: U.S. v. Cosa Nostra*. New York University Press, New York

Kelly R 1986 *Organized Crime: A Global Perspective*. Rowman and Littlefield, Totowa, NJ

Kelly R 1999 Organized crime: Some current trends. In: Einstein S, Amir M (eds.) *Organized Crime: Uncertainties and Dilemmas*. Office of International Criminal Justice/University of Illinois, Chicago, pp. 203–30

Lupsha P 1995 Organized crime. In: Bailey W (ed.) *The Encyclopedia of Police Science*. Garland, New York, pp. 493–7

Martens F 1991 Transnational crime and the elimination of frontiers. *International Journal of Comparative and Applied Criminal Justice* **15**: 11–21

National Research Council 1999 *Transnational Organized Crime*. Commission on Behavioral and Social Sciences and Education/National Academy Press, Washington, DC

Olson W 1994 The crisis of governance: Present and future challenges. Testimony to Senate Foreign Relations Committee, Washington, DC, April

Paoli L 1999 The future of Sicilian and Calabrian organized crime. In: Einstein S, Amir M (eds.) *Organized Crime: Uncertainties and Dilemmas*. Office of International Criminal Justice/University of Illinois, Chicago, pp. 155–83

Pennsylvania Crime Commission 1991 *The 1990 Report—Organized Crime in Pennsylvania: A Decade of Change*. Pennsylvania Crime Commission. Commonwealth of Pennsylvania, Conshohocken, PA

Reuter P 1994 Research on American organized crime. In: Kelly R, Schatzberg R, Chin K (eds.) *Handbook of Organized Crime in the United States*. Greenwood Press, Westport, CT, pp. 91–120

Savona E 1993 *Mafia Issues*. International Scientific and Professional Advisory Council of the United States Crime Prevention and Criminal Justice Program, Milan, Italy

Smith D 1980 Paragons, pariahs, and pirates: A spectrum-based theory of enterprise. *Crime and Delinquency* **26**(4): 358–86

Steffensmeier D 1986 *The Fence: In the Shadow of Two Worlds*. Rowman and Littlefield, Totowa, NJ

Steffensmeier D 1995 Public policy agenda for combating organized crime. In: Barlow H (ed.) *Crime and Public Policy*. Westview Press, Boulder, CO, pp. 269–95

Sterling C 1990 *Octopus: The Long Reach of the Sicilian Mafia*. Simon and Schuster, New York

Williams P 1994 Transnational organized crime and international society. *Survival* **36**(1): 96–113
Williams P 1999 Getting rich and getting even: Transnational threats in the twenty-first century. In: Einstein S, Amir M (eds.) *Organized Crime: Uncertainties and Dilemmas*. Office of International Criminal Justice/University of Illinois, Chicago, pp. 19–63

D. Steffensmeier and F. Martens

Crime: Sociological Aspects

The study of sociological aspects of crime is marked by controversy, beginning with definitional issues. There is general agreement that crimes are 'events and actions that are proscribed by the criminal law of a particular country' (Wilkins 1968, p. 476). The latter proviso, however, and the continually evolving nature of criminal law, render crime difficult to measure, describe, and explain. Laws vary greatly, between and within nation-states. Not all crimes are defined in criminal law. Entire classes of crimes, e.g., white collar and corporate crimes, encompass many behaviors that are defined by administrative and regulatory law. Correspondingly, 'criminals' in such cases are often corporations and other organizations. Nevertheless, the focus on law—law making and breaking, law breakers and their victims, and the control of law breaking and law breakers—is a unifying theme for an otherwise extremely diverse discipline and profession.

Some who study crime argue that crime should be defined as harm, whether or not specified in the criminal law. This ideological position charges that criminologists too often support 'existing structures of power and oppression,' and urges instead active promotion of change in these structures (Henry and Einstadter 1997, p. 418).

The vast majority of sociologists and other scholars who study crime and criminals reject this view, holding—in spirit, if not always in practice—to the cannons and the methods of science. The scholarly study of crime has become more truly interdisciplinary and, despite variations in substantive foci, methodological and theoretical approaches, attempts to integrate theories from a variety of disciplines are now common. Sociologists are at the forefront of such efforts (see, e.g., Tittle 1995).

1. *Sociological Issues in the Study of Crime and Criminals*

Sociological aspects of crime can be divided roughly into the following broad categories: (a) sociology of law; (b) conceptual issues; (c) measurement issues; (d) the social distribution of crime; (e) explanations of crime; and (f) issues related to crime control. Because they are so closely related to each other, conceptual and measurement issues are discussed together, as are the social distribution and explanations of crime.

1.1 *Sociology of Law*

The sociology of law focuses on the foundations of legal orders, patterns of legal change, and the role of law in social life. Definitions of behavior as criminal vary greatly in different times and places. Because laws define what is criminal or otherwise subject to sanction by governmental action, study of the law making process and of the effect of law on behavior is of particular interest to sociologists (Selznick 1968).

Law changes continuously, influencing and influenced by societal contexts. It is both enabling and constraining, conferring both rights and obligations. Legal ferment in many societies is related to changes in such matters as the status of females, the young, racial and ethnic minorities, homosexuals, and the indigent (McIntyre 1994). Understanding how and why such changes occur are major challenges to sociologists of law. Conflict and Marxist theories of crime draw much of their support from evidence that status inequalities bias law making and enforcement against the less powerful. The effects of legal sanctions on offenders and on those who might offend is also of interest to theorists of crime and policymakers alike.

1.2 *Conceptual and Measurement Issues*

Because crime consists of a very large number of heterogeneous criminal acts, no single index can capture its nature. Scholars and crime control agencies therefore find it necessary to reduce crime into more homogeneous categories. How these categories relate to one another is important both for understanding and controlling crime.

Official crime counting systems are designed primarily to serve institutional purposes such as documenting police or court activity, justifying budget requests, and evaluating institutional performance. They reflect the laws and practices of the jurisdictions with which they are associated. Data on crime and criminals are correspondingly affected. A great deal of effort has gone into standardizing definitions and reporting systems in order to enhance their comparability and research usefulness.

Having participated in developing measures of crime and crime reporting systems in many countries, sociologists continue to seek their refinement. Most countries now collect data on crimes known to the police and persons arrested, persons appearing before juvenile and adult courts, and persons convicted and sentenced to various forms of correctional sanctions. Interpretation of such data is guided by two principles.

(a) The further removed from its commission is a measure of crime, the less reliable and valid it is of the phenomenon. (b) Because the police must rely on victims and others to report most crimes, only data on serious crimes, which are most likely to be reported, are regarded as reliable. Therefore, measures based on serious crimes reported to the police or in surveys of victims of crime are favored when the object of study is the *distribution* of crimes. Similarly, for purposes of studying offenders, study of persons *arrested* for serious crimes are regarded as superior to data on persons who are incarcerated for those crimes.

Dissatisfaction with official data sources motivated the search for other measures. *Self-reports* of involvement in criminal and delinquent behaviors were developed in the 1950s and have been refined and solicited from better samples since that time. Reports of criminal *victimization* began to be developed during the latter part of the twentieth century and are now common in many countries.

Because crime is so varied, alternative measures and classification systems seek to reduce such heterogeneity (see Wilkins 1968). Thus, police classify separately more serious offenses against the person (homicide and non-negligent manslaughter) from other such offenses (aggravated assault; assault with intent to kill; and forcible rape). Offenses against property are similarly distinguished according to seriousness.

Typologies of criminal behavior systems yield important descriptive material on criminal techniques, careers, and lifestyles, but none has succeeded in capturing the full range of criminal behavior. Moreover, most crimes are committed not by specialists but by persons who vary the crimes they commit, contingent on their needs, abilities, opportunities, internal and external constraints.

1.3 The Social Distribution and Explanation of Crime

The basic data for sociological explanations of crime rest heavily on how crimes are distributed among such social categories as age, gender, social class, racial and ethnic status, over time, and in differing social and cultural circumstances. Most ordinary crime in all modern societies is committed primarily by young males, and is concentrated among those who are disadvantaged by economic and/or racial and ethnic status. The major exception is 'white-collar' crime, first identified as crimes committed by corporations or persons of high status, but later focusing primarily on violations of fiduciary responsibility or public trust.

Scholars associated with a variety of disciplines and professions study and theorize about crime and its control. The primary data employed by sociologists are *crime rates* and their distribution among groups, organizations, communities, and other political units. This *macrosocial* level of explanation seeks to determine what it is about such entities—their cultures, subcultures, and structural conditions—that explains varying crime rates. Sociologists who study the *individual* level of explanation focus on the influence of social and cultural factors, social interaction and individual learning experiences. Until late in the twentieth century, the lines between disciplines were clearly drawn and controversial. A third level of explanation, the *microsocial*, focuses on criminal behavior as a function of characteristics of situations, and of ongoing interaction among actors in situations (Short 1998). Theories that seek to integrate explanations of crime challenge disciplinary hegemony. Bridging theoretical levels of explanation and theoretical integration nevertheless remain difficult and contentious.

1.4 Issues Related to Crime Control

Crime control brings the sociological study of crime full circle, back to the law. The effects of law, law enforcement, and the imposition of legal sanctions on offenders and on others are major sociological concerns. Extensive research on the effects of legal sanctions on criminals, and on those who might offend, have resulted in few firm conclusions. Public concern with rising crime rates during the second half of the twentieth century resulted in higher rates of incarceration of convicted offenders, which accelerated during the last quarter of the century. A National Research Council report concluded that increments to crime control from incapacitation of offenders were modest (the size of the inmate population in the US doubled over a decade, resulting in an estimated reduction in crime of 10 percent), extremely expensive, and socially destructive (see Blumstein et al. 1986). Evaluations of other 'get tough' measure of crime control reach similar conclusions. As crime has increasingly become an issue of political contention, the relationship between knowledge about crime, criminals, and crime control has become controversial among scholars, criminal justice professionals, and the public. Efforts to bridge these sometimes conflicting interests nevertheless continue, theoretically (Braithewaite 1989) and in the form of experimental programs supported by governmental agencies in many countries.

2. The History of Explanations of Crime

Although eighteenth century philosopher/reformers and physicians/physiologists introduced modern thinking about crime, scientific study came later when European scholars such as Guerry, Quetelet, Durkheim, and Tarde began to search for regularities in the distribution of social phenomena. Durkheim and Tarde, especially, set the tone for a distinctively

sociological approach to the study of crime (Durkheim 1951/1897, Tarde 1886). Durkheim's insistence that social facts such as suicide and homicide rates were *sui generis* drew a boundary between sociology and psychology that profoundly affected the development of both disciplines.

Sociology was one of the earliest disciplines to include crime and criminals among its research and theoretical interests, and sociology continues to be the primary field of the majority of criminologists. Criminologists also study other basic social science disciplines, however, and in criminal justice programs. The professional field most closely associated with the study of crime, aside from criminal justice professionals, is law.

Several early strands of sociological research form the basis for much current theorizing. The first, building on the earlier work of the moral statisticians, sought to establish and interpret the social distribution of crime and delinquency. Following the framework of sociological and ecological theory developed by Thomas (Thomas and Znaniecki 1920) and Park and Burgess (1924), Shaw and his associates (Shaw et al. 1929, Shaw and McKay 1942) related crime and delinquency rates to social conditions in community areas in Chicago and elsewhere. Their chief findings—that rates of delinquency were highest in local communities characterized by poverty and high rates of population heterogeneity and turnover—and the theory advanced to account for them, social disorganization, remain of vital importance (Sampson et al. 1997).

Social disorganization theory hypothesized that socially disorganized areas lacked the means of both formal and informal social control because population turnover and heterogeneity were assumed to weaken social institutions, networks of primary relationships, and channels of communication, all vital to the supervisory capabilities of communities.

Case studies of delinquent boys led to a second theoretical conclusion, viz., that socially disorganized areas lacked sufficient socializing capabilities to protect children from involvement in delinquency and crime. Evidence of this sort led Sutherland (1939) to focus on the process by which individuals become criminal, a process he called 'differential association.' That crime and delinquency are learned behaviors, acquired in the course of interaction with others, remains an important principle. The further insight that definitions and labeling as criminal (in law, in interaction with others, and by authority figures) has behavioral consequences is the basis, as well, for theories stressing the importance of one's self-concept for behavior, and for labeling theory. In order to illustrate the applicability of differential association to crimes that were significantly different from ordinary crime, Sutherland also studied white-collar crime and professional crime, thus encouraging the strategy of studying different patterns of criminal behavior.

A third major sociological perspective on crime causation appeared with the publication of Merton's (1938) seminal theory of 'Social structure and anomie,' focusing on the consequences of disparities between culturally emphasized goals and structural limitations on the means for their achievement. Merton noted that the cultural emphasis on achievement of economic success posed special problems for those who were disadvantaged by virtue of their social class, racial, or ethnic status. Theoretical elaboration of anomie theory, as Merton's formulation has come to be known, have taken a variety of forms. Messner and Rosenfeld (1994) theorize that the domination of institutional life by economic institutions weakens the ability of other institutions to control crime. Cloward and Ohlin's (1960) 'opportunity theory' joins anomie theory with both social disorganization and differential association to explain variations in delinquent subcultures. Others shift the focus of Merton's formulation to individual strain, viewing crime as a means for reducing psychological strain caused by failure to achieve a variety of goals, such as autonomy, status, respect, and reciprocity in relationships with others.

3. *Continuity and Change in Criminological Theory and Research*

Much sociological theory about crime continues to be based on these theoretical traditions. A large-scale community study in Chicago, for example, found that a measure of 'collective efficacy' (willingness to intervene for such common goods as supervision of children and maintaining public order) had a strong effect on suppressing violent crime (Sampson et al. 1997), thus systematically documenting impressions gained from ethnographic research (Anderson 1990).

Rapid social change and new developments in research and theory continually influence sociological criminology. New and improved governmental statistical series inform criminal offending and victimization, and sophisticated self-report and observational studies provide insights into the lives of delinquents, criminals, and their victims, and the institutional and community contexts of crime. New techniques of statistical analysis yield new interpretations of empirical studies. Large-scale funding by governmental agencies and foundations permit more ambitious studies, including longitudinal data sets, experiments, and multiple methodologies and theoretical approaches. Theoretically integrative work combines insights from a variety of theoretical perspectives and levels of explanation.

Social disorganization, control, and opportunity theory continue to evolve as massive social changes alter the conditions of crime. Disciplinary boundaries become less distinct as sociologists find it necessary to take into account the impact of a rapidly changing global economy and new political alignments;

developmental and learning theorists must accommodate findings made possible by technological advancements in brain and neuropsychological research and by studying life-course patterns of individuals who were studied when they were very young.

A major impediment to theoretical advance is the lack of adequate data to measure new concepts and theoretical formulations that have been stimulated by the inquiries they have made possible. Theories such as 'control balance' (Tittle 1995) reach beyond criminal behavior and require measurement of the amount of control exercised by an individual, the amount of control experienced by that individual, ratios between these measures, and a large number of motivational, opportunity, and situational contingencies that are hypothesized to relate to a variety of forms of deviance, including crimes. 'Integrative shaming' (Braithewaite 1989), extending labeling theory and insights from control theory, requires measures of social integration of communities, their ability to shame norm violators, and their willingness to reintegrate them into the community, as well as of these same motivational, opportunity, and situational contingencies. 'Social capital' theory requires measures of the capacity of communities and institutions to provide supportive interpersonal and institutional relationships that bridge the generations while socializing the young.

Knowledge of substantive areas such as organized and professional crime and white-collar crime continue to suffer because systematic data are unavailable or inappropriate for extant theories. Rapid social change, driven by advances in science and technology, create new opportunities for crime even as they enhance crime control efforts. The same technological advances that greatly enhance computational abilities and complex data analyses pose new challenges for institutions of social control even as they improve their reach and their efficiency. Communication and data transfer technologies, for example, have transformed criminal justice agencies and created new opportunities for criminal behavior. Other technologies, such as nuclear and biotechnologies, lend themselves to widespread victimization, blackmail, and terrorist attack. Traditional distinctions between individual and corporate behavior that is defined as crime, and the behavior of nations that, although harmful to citizens of those nations previously was considered beyond the purview of other nations, are changing rapidly. New laws and new rules continually evolve as global markets alter commercial activity.

See also: Control: Social; Crime and Class; Delinquency, Sociology of; Law, Sociology of

Bibliography

Anderson E 1990 *Street Wise: Race, Class, and Change in an Urban Community*. University of Chicago Press, Chicago

Blumstein A, Cohen J, Roth J A, Visher C A 1986 *Criminal Careers and 'Career Criminals'*. National Academy Press, Washington D.C.

Braithewaite J 1989 *Crime, Shame, and Reintegration*. Cambridge University Press, Cambridge, UK

Cloward R A, Ohlin L E 1960 *Delinquency and Opportunity: A Theory of Delinquent Gangs*. Free Press, Glencoe, IL

Durkheim E 1951/1897 *Suicide: A Study in Sociology* [trans. Spaulding J A, Simpson G]. The Free Press, Glencoe, IL

Henry S, Einstadter W (eds.) 1997 *The Criminology Theory Reader*. New York University Press, New York

McIntyre L J 1994 *Law in the Sociological Enterprise: A Reconstruction*. Westview Press, Boulder, CO

Merton R K 1938 Social structure and anomie. *American Sociological Review* **3**: 672–82

Messner S F, Rosenfeld R 1994 *Crime and the American Dream*. Wadsworth, Belmont, CA

Park R E, Burgess E W 1924 *Introduction to the Science of Sociology*. University of Chicago Press, Chicago

Sampson R J, Raudenbush S, Earls F 1997 Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* **277**: 918–24

Selznick P 1968 The sociology of law. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 9, pp. 50–9

Shaw C R, McKay H D 1942 *Juvenile Delinquency and Urban Areas*. University of Chicago Press, Chicago

Shaw C R, Zorbaugh F M, McKay H D, Cottrell L S 1929 *Delinquency Areas*. University of Chicago Press, Chicago

Short Jr J F 1998 The level of explanation problem revisited. *Criminology* **36**: 3–36

Sutherland E H 1939 *Principles of Criminology*. Lippincott, Philadelphia

Tarde G 1886 *Le Criminalité Comparée*. F Alcan, Paris

Thomas W I, Znaniecki F W 1920 *The Polish Peasant in Europe and America*. Richard Badger, Boston

Tittle C R 1995 *Control Balance: Toward a General Theory of Deviance*. Westview Press, Boulder, CO

Wilkins L T 1968 Offense patterns. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 3, pp. 476–83

J. F. Short, Jr.

Crime, Sociology of

The sociology of crime is the study of crime causation in relation to social determinants, as opposed to biological or psychological factors. Causation is understood in a nondeterministic way as the necessary and sufficient antecedents or conditions of criminal conduct. Not being much older than a hundred years, the sociology of crime is a recent undertaking and its findings are best presented as a series of opposites, the first of these bearing on the definition of crime. For detailed treatment of the various elements discussed within these opposing perspectives and for further reference see *Anomie*; *Conflict Sociology*; *Control: Social*; *Crime: Sociological Aspects*; *Crime: Organized*;

Crime: White-collar; *Delinquency, Sociology of*; *Sociology: Overview*; *Criminal Justice, Sociology of*; *Gangs, Sociology of*.

1. *Sociology of Crime vs. Sociology of Criminalization*

Although crime is discussed as early as Plato's *Laws*, its empirical study really began at the turn of the last century with the Italian Positivist School. The most prominent members of this school articulated an objectivist perspective according to which crime was a universal fact that shared identical features across cultures. Hence, Garofalo developed his theory of the 'natural crime' and Lombroso his views on the 'born criminal.' In contrast, the French sociologist Émile Durkheim defined crime as a legal construct resulting from the social obloquy directed at certain forms of behavior (Durkheim 1958). Crucially, Durkheim not only asserted that social reaction varied according to context but that it was a force that asserted itself independently of the character of the behavior defined as deserving penal sanction: in a cloister of saints, the most benign misconduct would still attract harsh punishment. Both positions—crime as a natural fact and crime as a cultural construct—have shaped sociological thought on crime.

Sociologists tend to fall on either side of this divide largely on the basis of the type of crime that they study. The more predatory the crime the more realistic will be the perspective, either in respect to the crime itself or to its harmful consequences for victims (Young 1994). In contrast, the sociology of victimless crime and lifestyle delinquency stresses the constructive aspects of their legal definition. It must be emphasized that the sociology of crime has traditionally devoted a significant part of its inquiries to youth crime and juvenile gangs.

2. *Predetermined Behavior vs. Learned Behavior*

The concept of causation used in the sociological study of crime allows for many interpretations. On the one hand, the determinants of criminal behavior may be viewed as factual antecedents that do not depend on language communication to influence behavior. The various forms of deprivation associated with criminal behavior—such as lower-class status, unemployment, poverty and poor housing conditions, dysfunctional families and child sexual abuse, lack of education—are such relatively immediate determinants. On the other hand, criminal behavior can be seen as the result of a learning process which is progressively effected through language. In 1939, Edwin Sutherland famously objected to the causation through deprivation theory that it only applied to lower-class crime (if it did at all) and utterly failed to explain the crimes of the powerful (Sutherland 1983). In later work, he attempted to demonstrate that his own theory of differential association—according to which criminal behavior is learned in association with those who define such behavior favourably and in isolation from those who define it unfavourably—accounted as much for the crimes of the powerful as for the crimes of the powerless. The theory of differential association, which Sutherland and his collaborator, Donald Cressey, constantly refined, has exercised a powerful influence on the sociology of crime.

2.1 *Multifactorialism vs. Generalizing Theorists*

This opposition is closely related to the previous one. In his 1939 address on white-collar crime, Sutherland ridiculed sociological research that accrued so-called independent variables in order to produce a deterministic account of a selected part of criminal activities. As mentioned, he preferred to build an all-embracing general model that accounted in theory for all forms of crime. He was not the only sociologist of crime to favor theory building over the collection of heterogeneous sets of data. Most explanations of crime were developed by researchers who relied on broad single concept approaches, whether it was social disorganization (the Chicago School), anomie (Durkheim 1952, Merton 1967, and the subculture theorists, Cloward and Ohlin 1960), cultural and class conflict (Sellin 1938, Marxist criminology), the labeling process (Becker 1963), or routine activities (Cohen and Felson 1979). However, despite Sutherland's work, the crimes of the powerful have remained a largely underdeveloped field in the sociology of crime. The adherents of multifactorialism have pursued their unabated data collection in order to establish meaningful correlations between crimes of the powerless, which make up the overwhelming majority of reported crime, and various forms of deprivation.

3. *Consensus vs. Conflict*

There is one interpretation of social causation that stretches the concept of cause to its logical limit. It is the view that crime is the result of a conflict between different norms of conduct. Crime is then considered less as a thing *per se*—a given kind of behavior—than as a relationship between behavior and norms of conduct, where it is the latter that triggers the definition of certain kinds of behavior as being criminal (Sellin 1938). Within this framework, it is entirely conceivable that new rules of conduct may be established, which retrospectively define as being criminal certain kinds of behavior that were not previously seen as rule violating (e.g., private ownership in postrevolutionary Soviet Russia). In this case, it would appear that the causative

element follows in time the crime that it paradoxically generates.

This conflict of norms has been seen by sociology in opposite ways. Durkheim (1964) first proposed that conduct was criminalized because it strongly offended the collective sense of morality. His position implied that nearly all members of a given society shared a common ethos from which the lawbreakers were excluded. This view of the rules governing society as being the outcome of a consensus later came under increasing criticism. Rule definition was no more seen in the light of an opposition between the collective and the individual but in terms of power: those who were dominant within society, either through their numbers or, more importantly, through their status or their wealth, imposed their own definitions of proper behavior upon those whom they dominated. This latter perspective, known as 'conflict sociology,' reached its apex between 1960 and 1990 in Marxist studies, which upheld that criminalization was intended to perpetuate the domination of the ruling class over the working class (Greenberg 1981). Elements of the working class were actually defined as 'dangerous classes' in the sociological inquiries undertaken during the nineteenth century, and reappeared under the designation of the 'underclass' in the twentieth century.

3.1 Labeling Theory

Labeling theory carried the logic of conflict sociology to its logical term by not only asserting that social groups created deviance in making the rules that defined it but also by concluding that deviance was not a quality of the act a person commits but a consequence of the application by others of rules and sanctions to a labeled 'offender' (Becker 1963). Labeling theory was presaged in the work of Lemert (1951 and his concept of secondary deviance. According to Lemert, by being stigmatized as 'criminals,' offenders were excluded from mainstream society and left with no alternative but to sink deeper in crime. Thus the trapping of offenders in the ghetto of labeled criminals was a source of induced reoffending (secondary deviance).

4. Strain Theory vs. Control Theory

No dichotomy has attracted as much attention as the opposition between strain and control theories, which was articulated by Hirschi (1969). Strain theories rest on the concept of anomie, which was borrowed from Durkheim by Merton (1938) in an article that was revised several times in subsequent work (Merton 1967). Although Merton's conception of anomie differs much from Durkheim's, they are both framed in terms of a discrepancy between means and ends. For Durkheim, anomie is the hallmark of a troubled social predicament where people have unlimited cravings and limited means to fulfill them. This strain generates a normative vacuum which is for some conducive to self-destruction. For Merton, anomie results from a breakdown in the relationship between culture goals and the legitimate or institutionalized means to achieve them. This breakdown, particularly when it occurs between goals that are accepted and legitimate means to achieve them that are either unavailable or rejected, is a source of strain that may erupt in criminal behavior.

Anomie and strain theory basically tries to answer why and how people embrace criminal behavior. Therein lies the fundamental disagreement between strain theorists and control theorists such as Hirschi. For the latter, the question worth addressing is not why people break the law but why they actually conform to it. Then, the central concern of a sociology of crime should not be deviance but conformity. The sources of deviance can nevertheless be derived from breakdowns in the machinery that fosters conformity. Control theory originally answered its question about the roots of conformity with reference to civil society. Hirschi explained conformity by social bonding relationships that strengthened the ties between an individual and law-abiding society. These relationships consisted of attachment to family, school, and peers; commitment to conventional lines of action; involvement in conventional activities; and belief in the values that triggered these social bonding mechanisms and kept them functioning. Although that was not the path originally chosen by Hirschi, conformity could also be explained in political and legal terms, that is, by pointing to the fear of sanctions meted by the state upon lawbreakers. In this way control theory is connected with deterrence theory, and more generally with research into how people are coerced to conform by the state apparatus (see *Criminal Justice, Sociology of*; *Police, Sociology of*). In this process, the sociology of crime draws a full circle: it first begins as a sociology of deviance and then, striving to explain conformity, it is transformed into a sociology of social bonding and, ultimately, a sociology of the efficacy of the penal system. Furthermore, the notion of the positive delinquent whose behavior is impelled by multiple determinants is superseded by a perception of delinquents as individuals who exercise rational choice. Deterrence theory is predicated upon such a perception.

4.1 Delinquent Subcultures vs. the Subculture of Delinquency

As reinterpreted by Merton, anomie resulted from a breakdown between culturally valued goals and legitimate avenues of access to them. By combining the

acceptance or the rejection of culture goals with the acceptance or the rejection of institutionalized means, Merton generated a typology of four modes of social adaptation to strain: 'conformity' (goals and means are both accepted); 'innovation' (goals are accepted, but means rejected); 'ritualism' (goals are rejected, but means accepted); and 'retreatism' (goals and means are both rejected). To these, Merton added a fifth mode of adaptation, 'rebellion,' which is triggered when goals and means are not only rejected but also replaced. In combination with differential association theory, Merton's typology developed into the theory of differential opportunity and into the sociology of delinquent subcultures. Merton's mode of innovation would then beget a subculture of crime against property, rebellion could lead to conflict and to a subculture of violence, and retreatism to a subculture of drug taking (Cloward and Ohlin 1960).

It was quickly objected to subculture sociology that there may be a subculture of delinquency but that it was not a delinquent subculture (Matza 1964). The main point of this objection is that the notion of a delinquent subculture obscures the similarities between the conventional culture and the subculture of delinquency. In the subculture of delinquency, the acceptance of norms and, particularly, of legal norms is undermined by techniques of neutralization such as the negation of the offense, the sense of injustice, rudimentary conceptions of the sense of tort, and the primacy of custom. Yet the result of the application of these techniques is generally provisional. Delinquency is characterized by its deep ambivalence, the values promoted by conventional culture, and by the subculture of delinquency being convergent on key points.

4.2 *Microsociology vs. Macrosociology*

The critique of the sociology of delinquent subcultures was also indirectly critical of the general theories that underlaid this sociological trend, that is, the theory of anomie and of differential association. Such criticism caused a further split in the sociology of crime. In Matza's subsequent work, and in phenomenological and ethnomethodological research, the search for explanations was not only replaced by descriptive work but the descriptions eschewed quantitative generalizations for qualitative in-depth investigation of cases selected for their significance. Such investigations were influenced by the methods of ethnography and ethnolinguistics.

Concurrently, however, grand theorizing was being rediscovered with a premium by the Marxist sociology of crime. Sociologists influenced by Marxism attempted to develop a political economy of crime that ranged over the whole field of the social sciences; they also tried to articulate the notion of a structural or systemic mode of causation. Since it is impossible to divorce Marxist science from a Marxist project of class emancipation, this trend in the sociology of crime began to lose strength with the collapse of communist societies.

5. *Recent Trends*

The sociology of crime has evolved from a dichotomous field into a fragmented and an eclectic one. At least, the theories proposed on either side of the divides that we have been acknowledging were often highly integrated. At the beginning of the twenty-first century, the sociology of crime is in the grip of various centrifugal forces.

5.1 *A Wealth of Objects*

A factor that favored the theoretical integration of the sociology of crime during a long period of its development was that research focused on relatively few objects, youth crime providing an enduring reference. Social and political concern about forms of deviance as diverse as organized crime, terrorism, computer crime, and collective violence is now splintering the field into various specialties that defy integration. Some of these objects, such as organized crime and terrorism, strongly resist definition and they are nearly impervious to empirical research.

5.2 *The Shock of the Feminist Critique*

There are too many strands of feminism to present it as an integrated approach. Moreover, the sociology of women's crime is still in its infancy, despite important pioneering work (Carlen 1988). A significant part of feminist research has been directed at women considered either as victims of male predators or as subjects of oppression by criminal justice rather than as crime perpetrators. Nevertheless, the feminist critique has succeeded in casting a doubt on whether extant sociological theories of crime apply as much to women's crime as to men's and whether they account for the obvious gender difference in the perpetration of crimes. The answer to such questions will require a serious effort in research. It is dubious that in the short term this effort will foster the integration of the sociological theory of crime.

5.3 *Beyond the Sociology of Crime?*

One innovative theory that was developed during the late 1970s was the 'routine activities' theory (Cohen and Felson 1979). 'Routine activities' refer to the set patterns of behavior within the spatial environment of three kinds of social actors: (a) motivated offenders; (b) capable guardians of persons or property; and (c) suitable targets of criminal victimization (targets may

be persons or places). The interconnection of these patterns may account for criminal activities. For instance, the growth of the service industry since the end of World War II has multiplied evening work shifts that keeps people away from home as guardians of their property, while increasing their accessibility as targets when they come back home at night. Routine activities theory include 'guardians' as a variable explaining the occurrence of crimes. Guardians, who may be public police or private security agents, are however as much a factor in explaining crime as they are in accounting for crime prevention. This is why routine activities theory may act as a bridge between the sociology of crime and the sociology of crime control and of the production of security. This bridge is problematic as it may serve to integrate both disciplines into a consistent whole or to dissolve any claim to consistency that either of them has made.

See also: Law: Economics of its Public Enforcement; Law, Sociology of; Youth Culture, Sociology of; Youth Gangs

Bibliography

Becker H S 1963 *Outsiders: Studies in the Sociology of Deviance*. Free Press of Glencoe, London

Carlen P 1988 *Women, Crime and Poverty*. Open University Press, Philadelphia

Cloward R A, Ohlin L E 1960 *Delinquency and Opportunity: A Theory of Delinquent Gangs*. Free Press, Glencoe, IL

Cohen L E, Felson M 1979 Social change and crime rates trends: A routine activity approach. *American Sociological Review* **44**: 588–608

Daly K, Chesney-Lind M 1988 Feminism and criminology. *Justice Quarterly* **5**: 101–43

Durkheim E 1952 *Suicide: A Study in Sociology*. Routledge K Paul, London

Durkheim E 1958 *Rules of Sociological Method*. Free Press, New York

Greenberg D F (ed.) 1981 *Crime and Capitalism: Readings in Marxist Criminology,* 1st edn. Mayfield, Palo Alto, CA

Hirschi T 1969 *Causes of Delinquency*. University of California Press, Berkeley, CA

Lemert E M 1951 *Social Pathology*. McGraw-Hill, New York

Matza D 1964 *Delinquency and Drift*. Wiley, New York

Merton R K 1938 Social structure and anomie. *American Sociological Review* **3**: 672–82

Merton R K 1967 *Social Theory and Social Structure*. Free Press, New York

Sellin T 1938 A sociological approach to the study of crime causation. *Social Science Research Council Bulletin* **41**: 17–32

Sutherland E H 1983 *White Collar Crime: The Uncut Version*. Yale University Press, New Haven, CT

Young J 1994 Incessant chatter: Recent paradigms in criminology. In: Maguire M, Morgan R, Reiner R (eds.) *The Oxford Handbook of Criminology*. Oxford University Press, Oxford, pp. 69–124

J.-P. Brodeur

Crime: White-collar

The term 'white-collar crime' was coined by Edwin H. Sutherland, in a 1939 presidential address to the American Sociological Society, as 'crime committed by a person of respectability and high social status in the course of his occupation' (Sutherland 1940, 1949, p. 9). In the better part of the century that followed, scholars continued to disagree about this definition of the concept—from its focus, not on illicit acts, but on specific characteristics of individual lawbreakers, to its characterization of their misdeeds as 'crimes.' Though more eschewed than embraced his early definition, Sutherland's compelling label persists. But a vast collection of synonyms giving linguistic expression to conceptual, theoretical, and ideological differences proliferated—upperworld, elite, suite, of-the-powerful, corporate, organizational, occupational, employee, avocational, trust, business, economic, official, governmental, or political crimes, illegalities, offenses, violations, deviance, misconduct, harms, or violence. After decades of fruitless conceptual debate, most scholars have agreed to disagree.

This babel of competing verbiage and the disparate phenomena it embraces would be a mere annoyance were it not for the fact that the term was born, as Robert Merton once observed (Merton 1957, pp. 90–2), of an effort to 'liberate' scholarship from 'cognitive misbehavior' reflected in the well-accepted but spurious correlation between poverty and crime. Concepts tell us where to look, and where we look affects what we find. Sutherland asserted that criminologists, obstructed by conceptual blinders, had been looking in the wrong places:

> …[white-collar crimes] are not included in the criminal statistics nor are individual cases brought to the attention of the scholars who write the theories of criminal behavior. The sample of criminal behavior on which the theories are founded is biased as to socio-economic status, since it excludes these business and professional men. This bias is quite as certain as it would be if the scholars selected only red-haired criminals for study and reached the conclusion that redness of hair was the cause of crime (Sutherland 1949, pp. 8–9).

In other words, if scholars look only at crimes committed by the poor, it is not surprising that they conclude that poverty and its associated pathologies cause crime. For Sutherland, this selection bias is the *raison d'être* of inquiry into white-collar crime. These offenses evoke that nagging sense of all that is ignored by traditional criminology and that cause us to misunderstand the nature, causes, and consequences of crime and its control.

In the ensuing years, white-collar crime has become a subject of inquiry in its own right, of course. But selection bias continues to symbolize the essence of white-collar crime scholarship, simultaneously signaling its promise and foreshadowing its vulnerability, accounting for why the work is provocative and

groundbreaking, and yet so often misguided. This entry explores this dual tension, thereby featuring a handful of the major contributions of this scholarship and the challenges that social scientists face in doing it well.

1. Removing the Blinders

We begin by considering how some of the standard topics in criminology are transformed when one looks in the unconventional places captured by the varied conceptions of white-collar crime.

1.1 The Social Organization of Offending

Although Sutherland's early definition explicitly referred to individuals, his empirical work (Sutherland 1949) on the offenses of the seventy largest manufacturing, mining, and mercantile corporations in the United States took organizations as the unit of analysis, a legacy reflected in the generations of scholars he spawned. Though organizations surely commit misdeeds, they cannot do so without the complicity of individuals, often countless individuals plucked from disparate branches of the corporate tree and from tiers spanning the organizational hierarchy, indeed individuals often situated outside of the organization altogether—lawyers, accountants, investment bankers, brokers, agents, consultants, advertisers, competitors, even regulators. But coordinating the activities of such a diverse collection of collaborators situated across the globe, many of them ignorant of the illicit activities to which they contribute and often engaged in concerted action spanning years or decades, is not easy. Organizations need to maintain secrecy, lest the activities be exposed. They must insure that some conspirators remain unwitting of the consequences of their actions, lest they refuse to participate or blow the whistle to regulators inside or outside the organization. They must maintain loyalty and commitment and cool out those with little taste for offending, lest disgruntled or disaffected conspirators sell out their colleagues. They must buffer those in positions of power from those carrying out their directives to shield the former from accusations of culpability or to build a paper trail that deflects responsibility away from those most complicit. They must deceive and cover up their activities from victims to insure their continued victimization and to postpone discovery. They must destroy or doctor evidence to subvert regulatory discovery and to undermine the prosecution of their misdeeds, if discovered. These are massive organizational tasks perhaps better suited to the theoretical paradigms emanating from organizational sociology than from traditional criminology and that provoke unique challenges for systems of social control (Stone 1975).

1.2 Victimization

White-collar crimes exploit impersonal trust, asymmetric relationships in which principals entrust strangers to provide specialization, intermediation, and collectivization (Shapiro 1990). Victims of white-collar crime, therefore, typically lack expertise and access to the information necessary to evaluate the conduct of trustees. Because many participate in 'futures transactions' such as investment and insurance where payoff is delayed or protracted, victims have no tangible data in the interim to assess the likelihood that commitments will eventually be honored. Victims of white-collar crimes, therefore, tend to be unwitting. Because victims of common crimes are readily aware that they have been robbed, burglarized, assaulted, or victims of larceny, and members of their social networks know when associates have been killed, systems of social control (and researchers as well) tend to be reactive. They wait for victims to tell them that a crime has occurred. Reactive systems of social control yield little, stale, and biased data (about more blatant misdeeds) when most victims are unwitting. The policing (and study) of white-collar crime therefore requires different methods than those characteristic of common crimes.

White-collar offenses also differ from common crimes in the recruitment, pattern, scale, and impact of victimization, and in the relationships between victims and offenders. Because some relationships of trust collectivize the interests of multiple principals—in financial institutions, mutual and pension funds, insurance pools, and the like—victimization from the abuse of trust is often massive. Moreover, because white-collar crimes tend to be delivered by organizations whose activities touch countless consumers, clients, employees, investors, and neighbors (who drink the water and breathe the air poisoned by the corporation), victimization is often extensive and its cost enormous. Research finds that the organizational complexity of an offense is 'the single most powerful predictor of victimization' (Weisburd et al. 1991, p. 93). Hence, the organization itself becomes the 'weapon' in white-collar crime. It is 'for white-collar criminals what the gun or knife is for the common criminal—a tool to obtain money from victims' (Wheeler and Rothman 1982, p. 1406).

1.3 Enforcement

Systems of social control can be differentiated into those which seek 'compliance'—preventing the occurrence of law violations—or 'deterrence'—reacting to violations that have already occurred (Reiss 1984). Because white-collar offenses portend significant victimization with respect to cost, harm, consequences, irreversibility, and numbers of victims, enforcement systems frequently opt for compliance over deterrence. Regulators would rather prevent toxic waste or abort

a securities swindle than punish those responsible, especially when the unwittingness of victims often means that significant harm has been done by the time reactive deterrence systems learn of the illegality. And many find that cooperative enforcement techniques are better able to secure compliance than those structured to deliver punishments (e.g., Grabosky and Braithwaite 1986).

Compliance systems are more common in the enforcement of white-collar than common crimes as well because the former are more predictable. Because would-be organizational offenders are relatively stable and visible, as they seek to market their wares, they can be licensed, given preclearance, inspected, monitored, and subject to regulations and mandated procedures.

The policing of white-collar crime is also more diversified than that of common crime, which is undertaken largely by the state and private police (e.g., security guards). The costs and consequences of white-collar crimes, the penalties imposed, and the loss of reputation and the diminution of institutional trust (which are the lubricants of continued business) are so great, that those likely to bear these costs often undertake their own social control. Insurers, stock and commodity exchanges, professional or trade associations, and organizations themselves undertake or mandate self-regulation or buy the regulatory services of outside auditors, inspectors, bond-rating firms, and the like.

1.4 Prosecution and Punishment

Scholars object, with good reason, to Sutherland's classification of white-collar offenses as 'crimes.' Many—perhaps most—so-called white-collar 'crimes' are, in fact, not prosecuted as crimes at all or subject to criminal penalties. Compliance systems of social control have little use for the criminal justice system (aside from the incentive that potential criminal penalties give for undertaking self-regulation or abstinence). Criminal sanctions come too late, after the damage is done; they require standards of proof that are difficult to meet, especially with extensive obfuscation and cover-up that shield those most complicit; they do little to remedy the harm, lessen its impact, or impede future wrongdoing. Moreover, the choice between imprisonment and relatively small criminal fines offers limited options when the criminals are wealthy organizations. White-collar offenses are often dealt with by civil injunctions or administrative remedies that are better able to abort incipient or ongoing violations, contain the damage quickly, remedy the harm, provide restitution to victims, and fashion remedial solutions that will reduce opportunities for future wrongdoing without exacerbating the harm to innocent parties. Moreover, because wayward organizations or their insurers often have deep pockets, victims and the private plaintiff's bar have strong incentives to undertake civil litigation against white-collar malefactors. Private litigation often plays a significant role in rectifying the misdeeds, securing compensation, meting out punishment, and deterring future misconduct.

1.5 Criminological Theory

As even these few selected features of white-collar crime scholarship make clear, Sutherland's *raison d'être* for drawing attention to these offenses was prescient. By broadening the expanse of crime and its control from those patterns and institutions evoked by the stuff of traditional crime statistics to those insinuated in organizational offending and abuse of trust, the limitations of criminological theory are readily exposed. The challenge of incorporating into criminological theory the phenomena observed once the conceptual blinders were removed has been less successful.

Over the years, there have been sporadic attempts to develop a general theory of crime that accounts for both white-collar and common crimes. Sutherland's theory of differential association—that the techniques and motivations for criminal behavior are learned through contact with others—represents the earliest attempt (Sutherland 1949, p. 234). Some scholars have extended Merton's anomie theory of deviance or variants of strain theory, proposing that, like ordinary deviance, white-collar crimes arise when legitimate means to achieve material success or to attain organizational goals are blocked or unavailable. And other theories of white-collar lawbreaking have drawn on opportunity, neutralization, and conflict theories originally developed to account for traditional crimes. A more recent effort is seen in Gottfredson and Hirschi's (1990) general theory of crime, which asserts that white-collar crime, like all crime and deviance, is impulsive, arising from the self-interested pursuit of pleasure and the tendency to seek short-term, immediate gratification, coupled with low self-control. Ironically, their erroneous conclusion that there are no differences between those who offend in the streets and those in the suites (Hirschi and Gottfredson 1987, pp. 970–1), brings us full circle, as it results from the same selection bias—reliance on conventional Uniform Crime Report arrest statistics—that Sutherland exposed a half-century earlier.

These varied theoretical paradigms have been subject to extensive and justifiable criticism for imprecision, tautological reasoning, inability to account for compliance and law-abiding behavior, confusing correlation and cause, or lack of empirical support. Most white-collar crime scholars eschew such grand theories, developing explanations for specific misdeeds or those that account for variation within particular datasets.

2. Methodological Pitfalls

The empirical challenges to the study of white-collar crime far outweigh the theoretical ones. First, because of victim unwittingness, relationships of impersonal trust, the protection of crime in the suites by institutions of privacy, complex cover-up schemes, the commission of crime with organizational weapons rather than physical ones, and the ability to buffer misdeeds in long-latent futures transactions, the dark figure of undetected white-collar crime probably exceeds that of all other crimes combined. And even where law enforcement officials have discovered white-collar offenses, they are often unable to ascertain or pin the blame on those most responsible. The most common sources of data on traditional crimes—arrest records and crime statistics, victimization surveys, and observation studies—therefore yield little and very biased data on white-collar crime.

With limited access to data, researchers often rely on materials in the public domain—typically, sentencing data, reports of administrative agencies, journalistic accounts, hearings, and official inquiries. But selection bias is the price of easy access. Journalists produce what sells—the egregious, scandalous, shocking, clever, costly, and devastating offenses committed by notorious or colorful characters, where those in high stations fell from grace, where safety nets failed, where victims should have known better, where offenders received excessively lenient treatment—many of the same features of offenses that incite public hearings and legislative inquiry. Hardly the mundane stuff of white-collar crime day in and day out. In selecting research subjects from the news media and political tribunals, scholars replicate Sutherland's red-haired crime fallacy. We come to believe that these crimes are committed (with impunity) by the rich and powerful, when a less biased data source would reveal that they are typically carried out by members of the middle class, and that, contrary to expectation, the severity of punishment of white-collar criminals appears to increase with social standing (Weisburd et al. 1991).

Turning to public records of convicted criminals incurs other selection biases. Convictions represent a very small and unrepresentative tip of the iceberg. Research on the US Securities and Exchange Commission, for example, indicates that not even five of every 100 parties found after investigation to have violated the securities laws are convicted of a crime (Shapiro 1984, pp. 155–61). Criminal courts are less likely to process the misdeeds of organizations (than of individuals), that were uncovered by compliance (rather than deterrence) systems of social control, and that were committed by those atop corporate hierarchies who were able to conceal their responsibility or inculpate others in exchange for immunity from criminal prosecution. When offenses are ongoing, victimization is extensive, and the impact is significant, law enforcers opt for compliance—halting the misdeeds, forestalling, and remedying the harm, minimizing the likelihood of recurrence, and doing so with more immediacy, less litigation, and lower standards of proof—over criminal proceedings. Therefore, although serious significant offenses can certainly be found at the tip of the iceberg, so can some of the most trivial for which no remedy other than criminal prosecution is available. Researchers concentrate on the tip of the iceberg at their peril.

Because of these biases embedded in the social construction of white-collar offenses as 'crimes,' many scholars choose not to concentrate solely on criminal cases. Some study offenses that carry criminal penalties, even if they have been resolved instead by private settlements, warnings, injunctions, civil or administrative proceedings, lawsuits, or other remedies. But a slippery slope looms below the tip of the iceberg. As the labels shift from 'crimes,' to 'violations,' 'offenses,' 'misconduct,' 'harms,' 'wrongs,' 'deviance,' 'violence,' and so on, the literature becomes cluttered with studies of torts, liability, malpractice, accidents, mistakes, or inadvertence committed by corporations or those in high-status occupations. We dilute white-collar crimes with all manner of misfortune. To include these harms among crimes in the suites is like including fender benders in an ice storm among crimes in the streets.

Sadly, the ideological passions that animated Sutherland's work and that of many of his intellectual progeny, which cause us to be suspicious of corporations and individuals of wealth, status, and power, also cause us to get sloppy. They permit us to confuse mistake or accident with intentional noncompliance. They lead us to believe what comports with our assumptions, relying on the thin anecdotal data that are readily available rather than probing deeper, using published reports and secondary data rather than collecting primary materials and triangulating data sources. They blind us from questioning and making problematic what seems intuitively true. As a result, we are destined to learn that some of the classic and repeatedly cited 'landmark narratives' of corporate crime—for example, the Ford Pinto affair and the Challenger disaster—actually got the story wrong, confusing misconduct with mistake, and venality and amoral calculation with banal normalized organizational routines (Lee and Ermann 1999, Vaughan 1996).

We are also learning what it takes to get the story of long-running technologically, financially, and organizationally complex potential white-collar crimes right. In the case of the Challenger disaster, it took 10 years of persistent, painstaking, methodical research which probed the entire decade that preceded the shuttle explosion that cold January morning; systematic analysis of more than 122,000 pages of documents, 2,500 pages of hearing transcripts, and 9,000 pages of depositions; dozens of new interviews

and site visits; the commitment to learn technical science and engineering; and the willingness of a publisher to print an affordable 600-page scholarly book with almost 100 pages of technical charts, diagrams, memos, and thousands of footnotes to transform a seductive but erroneous account of organizational misconduct into a creditable one of the normalization of deviance and the sociology of mistake (Vaughan 1996). Vaughan's chronicle, with the conventional (erroneous) account in boldface type juxtaposed against the fuller complex documentary and ethnographic narrative, starkly demonstrates what is at stake in the social construction of white-collar crime scholarship and what is required of future scholars.

3. *White-collar Crime and the New Social Science*

> …societies have been undergoing an organizational revolution. Just as the forests and fields of the physical environment are being replaced by streets and skyscrapers, the primordial institutions around which societies have developed are being replaced by purposively constructed social organization (Coleman 1990, p. xv).

Perhaps nowhere has James S. Coleman's influential call for a new social science attuned to the transformation of social life occasioned by the displacement of natural persons by 'purposive corporate actors' been better realized than in the study of white-collar crime. White-collar crime scholarship does not hold a monopoly on the study of organizational offending, of course. Even traditional criminology (which concentrates on the natural persons who populate the primordial social organization which Coleman argues is 'withering' away) has produced the occasional study of gangs, organized crime, terrorism, the distribution of illicit goods (e.g., drugs, gambling) and services (e.g., prostitution, fencing), and the like. But white-collar crime scholarship more than any other area of criminological inquiry has made problematic the difficulties of regulating organizational life and insuring norm compliance within organizational contexts, areas of investigation which will become ever more important in the new century. White-collar crime offers a bridge to the new social science and a compelling window onto the changing structure of social life. As we gaze through the window, we realize the myopia our blinders have imposed.

See also: Control: Social; Crime and Class; Crime: Organized; Crime: Sociological Aspects; Crime, Sociology of; Criminal Justice, Sociology of; Delinquency, Sociology of; Electronic Economy: Governance Issues; Law: Economics of its Public Enforcement; Law, Sociology of; Technology and Social Control

Bibliography

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Gottfredson M R, Hirschi T 1990 *A General Theory of Crime*. Stanford University Press, Stanford, CA

Grabosky P, Braithwaite J 1986 *Of Manners Gentle: Enforcement Strategies of Australian Business Regulatory Agencies*. Oxford University Press, Melbourne, Australia

Hirschi T, Gottfredson M 1987 Causes of white-collar crime. *Criminology* **25**: 949–74

Lee M T, Ermann M D 1999 Pinto 'madness' as a flawed landmark narrative: An organizational and network analysis. *Social Problems* **46**: 30–47

Merton R K 1957 *Social Theory and Social Structure*. Free Press, Glencoe, IL

Reiss Jr A J 1984 Selecting strategies of social control over organizational life. In: Hawkins K, Thomas J M (eds.) *Enforcing Regulation*. Kluwer-Nijhoff, Boston, pp. 23–35

Shapiro S P 1984 *Wayward Capitalists: Target of the Securities and Exchange Commission*. Yale University Press, New Haven, CT

Shapiro S P 1990 Collaring the crime, not the criminal: Reconsidering the concept of white-collar crime. *American Sociological Review*. **55**: 346–65

Stone C D 1975 *Where the Law Ends: The Social Control of Corporate Behavior*. 1st edn. Harper and Row, New York

Sutherland E H 1940 White-collar criminality. *American Sociological Review* **5**: 1–12

Sutherland E H 1949 *White Collar Crime*. Holt, Rinehart and Winston, New York

Vaughan D 1996 *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press, Chicago

Weisburd D, Wheeler S, Waring E, Bode N 1991 *Crimes of the Middle Class: White-collar Offenders in the Federal Courts*. Yale University Press, New Haven, CT

Wheeler S, Rothman M L 1982 The organization as weapon in white-collar crime. *Michigan Law Review* **80**: 1403–26

S. P. Shapiro

Criminal Defenses

Modern legal systems worldwide accept, at least in principle, the notion that punishment should correspond to the extent of an individual's responsibility for an act. Unless the crime is an absolute or strict liability offense, what must be established during the criminal trial is the mental state of the individual at the time he or she committed the action in question. Whether there were any circumstances which would warrant holding a defendant less responsible, thereby resulting in less punishment, also has to be established. In every case the issue is what constitutes condign punishment.

In order for an individual to be found guilty, the prosecution must establish that the defendant had the requisite *mens rea* or intent and that the defendant committed the *actus reus* or act. The law in the Anglo-American tradition is only concerned with whether or not the accused intended to commit the act, which the law defines as a crime. The reason why a defendant intended to violate the law is not deemed relevant to the question of guilt (except where motive is part of the definition of the crime, i.e., for self-defense where one kills in order to save one's own life and for specific intent crimes such as child sexual abuse where the touching must be for the purpose of sexual gratification). The twentieth century view of *mens rea* as merely legal guilt, represents a shift in thinking from earlier times when a defendant had to be shown to have both legal guilt or intent as well as moral guilt or motive. The trend toward a narrower version of *mens rea* may reflect a practical judgment that it is difficult to prove motives.

The criminal defenses that have developed over the past century are often predicated on the idea that an individual's alleged criminal act was not entirely voluntary. The social science evidence presented in court, ordinarily through the testimony of expert witnesses, is designed to show that the defendant lacked the necessary intent to be held fully responsible for the offense. In some cases there may be doubt as to the legitimacy of a particular defense, and there may also be doubt as to whether the defendant properly invokes the defense. Nevertheless, whether or not it is possible to prove that an individual acted under an imperative, legal systems guarantee the right to invoke these defenses during the trial.

Through colonialism, countries with common law and civil law traditions imposed their legal systems, along with their defenses, on many parts of the world. Consequently, many, though not all, of the defenses discussed have been employed in courts across the globe. Yet, even where one finds the same defenses in different countries, the use of the defenses varies by jurisdictions.

Most of the defenses discussed are invoked in both common law and civil law types of legal systems.

1. *Justification: Self Defense*

Some defenses are classified as justifications, meaning that the person was correct in his or her action and should therefore not be considered culpable at all. The classic example is self-defense, under which it is justifiable for a person to use force to save his or her own life. Though self-preservation may justify the use of force, the law generally holds that a person should only use as much force as is necessary to save his or her life. Not only must there be proportionate use of force, but the situation must present an imminent danger of unlawful bodily harm. For example, some courts have held that a person has a duty to retreat where possible.

2. *Excuses*

When raising this second type of defense, defendants admit their action was wrong but maintain they should be excused because of the circumstances.

2.1 *The Insanity Defense*

One of the most important criminal defenses is the insanity defense, which is based on the proposition that a defendant with a mental defect should not be held criminally responsible (see *Law: Defense of Insanity*). Most formulations of the insanity defense are based on either cognitive or behavioral impairment (for analysis of the German 'biological test,' see Silving 1967). With respect to cognitive impairments, it is considered unjust to punish a person who suffers from a mental defect that prevents him or her from comprehending the nature of his or her action. Where the insanity defense relies on a cognitive dysfunction, the famous M'Naghten test is frequently used: did the defendant know the nature and quality of the act, and if he did, did he know that what he was doing was wrong? One ambiguity with the M'Naghten test has to do with the interpretation of 'wrong.' When a defendant argues that he knew the act violated the law but felt that he was morally right, courts accept this insanity defense only in cases where the defendant claims to have been following a divine commandment.

Where the defendant was aware that the action was wrong but could not make his or her behavior conform to the law, this is known as a volitional dysfunction. A common version of the volitional impairment is the irresistible impulse test, and a kleptomaniac exemplifies the sort of person who would invoke the volitional insanity defense.

The insanity test has remained controversial for several reasons: (a) the fear that malingerers will avoid punishment (i.e., 'beat the rap'); (b) the concern that poorer defendants will be less able to retain expert witnesses who are as persuasive as those provided by the state (see *Expert Witness and the Legal System: Psychological Aspects*); and (c) the erroneous belief that those committed to mental institutions will be released earlier than those who are incarcerated in prisons.

The meaning of the insanity defense has evolved as new conditions are documented. For example, the medical and psychiatric communities have recognized the existence of culture-bound syndromes such as 'running amok.'

2.2 *Diminished Capacity/Partial Responsibility*

The diminished capacity defense may be used where a defendant cannot be considered insane according to

the strict tests in effect in a given jurisdiction, but nevertheless suffers from some sort of mental abnormality. There are two variants of the diminished capacity doctrine: the '*mens rea*' variant and the 'partial responsibility' variant. The essential distinction between them, is that in the *mens rea* variant the defendant negates one of the elements of the crime (and so he is not guilty of the crime charged), while in the partial responsibility variant, the defendant argues that his responsibility is lessened by virtue of his mental abnormality at the time of the commission of the crime (and so he deserves lesser punishment).

The rise of new versions of diminished capacity is worth noting. The defenses are frequently versions of post-traumatic stress disorders including a *post partem* defense, an intoxication defense, a Vietnam Veteran's defense, and the Battered Woman defense. The automatism defense is sometimes regarded as a type of insanity defense but is different in character. Automatism refers to the behavior of an individual who acts either unconsciously or by reflex or spasm, including conditions such as somnambulism. The standard view of the automatism defense is that it does not relate to *mens rea* but rather to the *actus reus*. The notion is that a person who acts while unconscious should not be held responsible for illicit conduct because the act was not a voluntary one, not because of a failure to form the necessary intent.

2.3 Mistake of Law and Mistake of Fact

The traditional maxim 'ignorance of the law is no excuse' is followed for the most part. If a person was unaware that his or her conduct was proscribed by law, that is not an excuse. This strikes some as unfair as it is unrealistic to expect individuals to know all the laws in their justice systems, even assuming that was possible in earlier times.

By contrast, a mistake of fact can excuse otherwise criminal conduct, as long as the mistake is objectively reasonable. To determine whether mistake of fact excuses depends on the following: if the facts were as the defendant believed them to be, the act would not have been a crime; the mistake of fact excuses the behavior. Contrariwise, had the facts been as the defendant imagined still been a crime, then the mistake furnishes no excuse. In rape cases, men have invoked this defense to claim that they were mistaken as to the consent of the women.

2.4 Necessity and Duress

Some defenses are meant to show that individuals had no real choice when they committed the crimes. The defense of necessity was used, for example, in the famous nineteenth-century 'Case of the Mignonette,' when sailors in a dinghy without provisions devoured a cabin boy. When they were later prosecuted for murder, the argument was that either one had to die or they all would have died. Thus, the choice was one of lesser evils. Although the court rejected the defense of necessity in this case, necessity has been interpreted as both a justification and an excuse. Modern legal systems treat it as an excuse.

In contrast to necessity, which is a consequence of natural forces, duress requires that the pressure of an unlawful threat come from another human being. Defendants argue duress when they are ordered by someone else to kill and must comply in order not to be killed themselves.

2.5 Partial Excuse—Provocation

Provocation, also known as the heat of passion rule, is one of the most ancient defenses in the criminal law. Found both in common law and in most homicide statutes, the 'provocation/passion formula,' which can reduce a murder charge to one of manslaughter, is based on the idea that a person who is provoked to kill does so without the malice aforethought required for the crime of murder, and is therefore less culpable. Often raised by jealous husbands who kill their wives, the provocation defense is the classic example of a partial excuse. Because the infidelity of a wife is considered likely to elicit a violent reaction on the part of husbands, the provocation defense is said to be 'cross-culturally ubiquitous, perhaps universal' (Daly and Wilson 1997). This defense is found around the world (Isedonmwen 1988, Vasdev 1968, Feltoe 1983–1984, Stanley 1992).

If the jury is persuaded both that the defendant was provoked and that the so-called 'objective reasonable person' would have been provoked by the action in question, then a charge of murder is reduced to manslaughter. The defendant is not exonerated, but rather is convicted of a lesser charge. But it is sometimes difficult to prove that one's response is that of a reasonable person, as in the case of culturally specific insults that are offensive only to members of a particular ethnic community.

2.6 The Witchcraft Defense

Another controversial defense, which has been widely used in colonial and formerly colonial settings, is the so-called witchcraft defense. Much of the literature about the defense concerns its use in African contexts (see, e.g., Aremu 1980, Chukkol 1983, Mutungi 1977), but it has also been invoked in other parts of the world (see, e.g., Clark 1980, Hobgin 1935). When invoking the defense, a defendant claims to have killed a person because that person used supernatural powers against him or his family. The defense has frequently been formulated as a version of self-defense or of insanity.

Its use has often been problematic either because courts conclude that the 'reasonable person' would not believe in witchcraft or that a belief in witchcraft does not demonstrate insanity on the part of the defendant since sane members of the community strongly believe in witchcraft. Although the defense was largely unsuccessful in colonial settings, in the past it was considered acceptable for the state to kill witches in Europe and the US (Currie 1968).

3. *Controversial Defenses*

Attorneys sometimes suggest new defenses such as the brainwashing defense. When defendants invoke the brainwashing defense, they argue that they should not be held responsible because their criminal actions were a consequence of mind control. The 'rotten social background' defense, championed by Judge Bazelon and Professor Richard Delgado in the US, depends on the idea that individuals who grew up in the ghetto may be predisposed to use force as part of a survival strategy. As a consequence, defendants argue that they should not be held fully responsible for their illicit conduct. There is also the 'Black rage defense' which lawyers sometimes use when they wish to attribute their clients' illegal activities to racism.

The cultural defense, which has received little support, is employed primarily by immigrants and refugees (see *Cultural Rights and Culture Defense: Cultural Concerns*). The defense takes as its premise the idea that individuals were socialized to behave in particular ways, some of which may clash with the law of the new homeland. If an individual is predisposed to act in ways that violate the law, or is required by customary law to avenge a death, then the cultural defense would require that the legal system should at least take the cultural imperatives into account in determining the correct degree of culpability.

Religious defenses, which have been invoked when members of religious minorities are prosecuted for using drugs in religious ceremonies, for hunting endangered species like eagles and panthers, or for marrying more than one woman, fare no better because governments' interests in enforcing the law apparently override the right to religious liberty.

The apparent proliferation of new criminal defenses has led some critics to condemn the 'abuse excuse.' Critics find this trend deplorable as they believe criminals should accept responsibility for their actions and not seek to blame others for their misconduct. The counter-argument has been that legal systems have long sought to achieve individualized justice by tailoring punishment to the particular attributes of defendants (e.g., adults are distinguished from juveniles; the sane treated differently from the insane). The acceptance of the battered woman defense also shows an acknowledgment of gender differences in the perception of, and reaction to, violence.

Whether the defenses are formally incorporated in legal systems is important to criminal defendants. If officially recognized, this means in practical terms that defendants will be entitled to present specific kinds of evidence in court. Without a formal defense, judges are more likely to exclude the evidence as irrelevant.

4. *Expert Witnesses*

The success of defenses often hinges on the presentation of evidence by expert witnesses (see *Expert Witness and the Legal System: Psychological Aspects*). Social scientists play a crucial role in the presentation of information relevant to particular defenses, which serves to provide the context for the defendant's action. If the defense is a novel one, there may be some question as to whether the expert satisfies the requirements for serving as an 'expert.' This determination usually turns on the degree to which the evidence constitutes a body of knowledge widely accepted within the relevant professional or scholarly discipline.

5. *Role of Criminal Defenses*

Criminal defenses are important to guarantee that the accused only receives as much punishment as is deserved. Should the defense fail, however a judge can still mitigate the punishment during the sentencing phase of the trial. Thus, even if a defendant is convicted of a crime that is really too serious, given the nature of his or her act, the defendant may ultimately receive less punishment than one might expect. Indeed, judges sometimes hold criminals responsible, finding them technically guilty, even though they impose suspended sentences or probation. By so doing, they maintain the rule of law, the idea that justice is applied in an even-handed way to all defendants, even though in reality they mete out varying sentences.

6. *Future Uses of Criminal Defenses*

The question of what defenses are acceptable emerged anew in the 1990s during the drafting of the Statute for the International Criminal Court. As efforts were underway to harmonize competing legal conceptions from the common law and the civil law, a decision had to be made as to which defenses should be part of the operating procedure of the international tribunal. Human rights advocates found disturbing the acceptance of the defense of superior orders, even under limited circumstances, a defense according to which soldiers should not be held responsible for war crimes and crimes against humanity if they were simply following the orders of their superior officers.

The question of which criminal defenses are legitimate remains an open one in most legal systems. With

the advent of new medical, technological, and psychological breakthroughs, it is likely that new defenses will continue to emerge on the horizon.

See also: Courtroom Consultation: Psychology; Criminal Justice, Ethics of: Aspects of Human Dignity; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Cultural Rights and Culture Defense: Cultural Concerns; Expert Testimony; Expert Witness and the Legal System: Psychological Aspects; Guilt; Justice and its Many Faces: Cultural Concerns; Justice: Philosophical Aspects; Law: Defense of Insanity; Trials: Cultural

Bibliography

Aremu L O 1980 Criminal responsibility for homicide in Nigeria and supernatural beliefs. *International and Comparative Law Quarterly* **29**: 112–31

Berner S H 1971 The Defense of Drunkenness—A Reconsideration. *University of British Columbia Law Review* **6** (1): 309–51

Bolgar V 1967 The present function of the maxim *Ignorantia iuris Neminem Excusat*—a comparative study. *Iowa Law Review* **52**: 626–56

Brown B 1964 The 'ordinary man' in provocation: Anglo-Saxon attitudes and 'unreasonable non-Englishmen.' *International and Comparative Law Quarterly* **13**: 203–35

Clark D S 1980 Witchcraft and legal pluralism: The case of Celimo Miquirucama. *Tulsa Law Journal* **15**: 679–98

Currie E 1968 Crimes without criminals: witchcraft and its control in Renaissance Europe. *Law and Society Review* **3**: 7–32

Connell J G III, Valladares R (eds.) 2000 *Cultural Issues in Criminal Defense*. Yonkers, Juris, New York

Daly M, Wilson M 1997 Crime and conflict: Homicide in evolutionary psychological perspective. *Crime and Justice* **22**: 51–94

Delgado R 1979 Ascription of criminal states of mind: Toward a defense theory for the coercively persuaded defendant. *Minnesota Law Review* **63**: 1–40

Delgado R 1985 Rotten social background: Should the criminal law recognize a defense of severe environmental deprivation? *Law and Inequality: A Journal of Theory and Practice* **3**: 9–90

Farhad M 1994 *The Concept of Islamic International Criminal Law. A Comparative Study*. Graham and Trotman/Martinus Nijhoff, London, Dordrecht, Boston

Feltoe G 1983–1984 Criminal law policy in relation to the defense of provocation. *The Zimbabwe Law Review* **1, 2**: 140–57

Ferguson G 1990 The insanity defense in Canada, Malaysia and Singapore: A tale of two codes. *Journal of Malaysian and Comparative Law* **17**: 1–28

Fingarette H T 1973 *The Meaning of Criminal Insanity*. University of California Press, Berkeley, CA

Fletcher G P 1974 The individualization of excusing conditions. *Southern California Law Review* **47**: 1269–1309

Gaw A C (ed.) 1993 *Culture, Ethnicity, and Mental Illness*. American Psychiatric Press, Washington, DC and London, UK

Goldstein A S 1967 *The Insanity Defense*. Yale University Press, New Haven, CT and London

Harris P 1997 *Black Rage Confronts the Law*. New York University Press, New York and London

Hart H L A 1964 *The Morality of the Criminal Law* (the Lion Cohen Lectures). Oxford University Press, Oxford, UK

Herman D H J 1983 *The Insanity Defense: Philosophical, Historical and Legal Perspectives*. Charles C. Thomas, Springfield, IL

Howard C 1961 What color is the 'reasonable man?' *Criminal Law Review* 41–8

Husak D 1989 Motive and criminal liability. *Criminal Justice Ethics* **8**: 3–14

Isedonmwen E O 1988 A requiem for provocation? *Journal of African Law* **32**: 194–207

Jayaratnam Neelakanthi 1979 Irresistible impulse. *Colombo Law Review* **5**: 103–14

Knoops G-J G J 2001 *Defenses in Contemporary International Criminal Law*. Transnational Publishers, Ardesley, New York

LaFave W R 2000 *Criminal Law*, 3rd edn. West, St. Paul, MN

Obi N, Ignatius Ebbe 2000 *Comparative and International Criminal Justice Systems: Policing, Judiciary and Corrections*. 2nd edn. Butterworth-Heinemann, Boston

O'Regan R S 1972 Ordinary men and provocation in Papua and New Guinea. *International and Comparative Law Quarterly* **21**: 551–7

Perkins R M, Boyce R N 1982 *Criminal Law*, 3rd edn. Foundation Press, Mineola, NY

Ploscowe M 1930 An examination of some dispositions relating to motives and character in modern European penal codes. *Journal of the American Institute of Criminal Law and Criminology* **21**: 26–40

Robinson P H 1982 Criminal law defenses: A systematic analysis. *Columbia Law Review* **82**: 199–291

Robinson P H 1984 *Criminal Law Defenses*. West, St. Paul, MN 2 Vols

Silving H 1967 *Essays on Mental Incapacity and Criminal Conduct*. Charles C. Thomas, Springfield, IL

Stanley M H 1992 Yeo lessons on provocation from the Indian penal code. *International and Comparative Law Quarterly* **41**: 615–31

Tadiar A F 1984 Legal consequences of insanity. *The Philippine Law Journal* **59**: 365–76

Simpson A W B 1984 *Cannibalism and the Common Law*. Penguin Books, Harmondsworth, UK

Vasdev K 1968 Provocation as a defense in Sudan criminal law. *Sudan Law Journal and Reports* : 167–229

Walker L 1979 *The Battered Woman*. Harper and Row, New York

Williams G 1949 Homicide and the supernatural. *Law Quarterly Review* **65**: 491–503

Williams G L 1953 *Criminal Law: The General Part*. Stevens, London

A. D. Renteln

Criminal Justice, Ethics of: Aspects of Human Dignity

Important ethical issues can be identified at many levels of the theory and practice of crime control in the modern state. One important contrast is between

substance and procedure, with different procedural systems generating professional roles with distinctive ethical problems. The adversarial system of criminal procedure in Anglo-American regimes is associated with such problems for lawyers and government officials who must honor duties to adversarial interests as well as to the achievement of just overall outcomes. The accusatorial systems of Continental Europe generate different types of role conflict by confining in a single professional role responsibility for both the repression of crime and the assurance of fair criminal procedure. Across all varieties of criminal procedure, the enormous discretions vested in all modern punishment systems present ethical problems of great importance and indeterminate resolution. The only way to reduce discretion is with mandatory punishment rules that may be more problematic than the practical outcomes of discretionary systems.

This article will not survey the wide variety of ethical issues found in modern criminal justice. Instead we wish to emphasize one central concern: the strategic importance of imposing limits on the punishment of criminals to ensure the respect of government for citizenship and individual dignity. Serious criminal offenders present the least attractive case for claims to limit government power, but they are, for that reason, the most important frontier for defending the limits of that power. If the claim of human rights is only as strong as its weakest link, that will almost always make the rules and limits of criminal punishment of central importance to respect for individual dignity as a limiting principle throughout government.

This article will first sketch out the argument for considering criminal punishment as a critical arena for human rights and then discuss those areas within criminal justice where the claims for limiting government are most important. A second section addresses the politics of governmental limits on punishment. A third section focuses on conditions of imprisonment. A fourth section contrasts two styles of characterizing the role of government in punishing criminals.

1. *Appeals for Government Power*

To the extent that power held by government comes at the expense of restrictions on the liberties of individual citizens, there is a natural tendency for citizens to resent increases in governmental power and to desire limits on government. Where citizens have continuing power over governmental policy, new powers must be approved on the basis of the need for government to produce common benefits that citizens value more highly than the liberties they surrender. The need to justify governmental power in terms of collective gains that are more important than lost liberty is spread over the full range of governmental powers from taxation and spending to military force. Crime and punishment have a special strategic importance in this dynamic to establish the limits of governmental power, because criminals are the most frightening and most unpopular citizens in a democratic state. Fear of crime means that citizens will cede authority to government for crime control that they might not be willing to surrender for other reasons. The public hostility toward criminals will support harsh punishments administered by government because criminals are to receive them. Citizens will approve conduct in the punishment and control of crime that they would hesitate to support or tolerate in other domains.

Because criminal punishments are the most extreme deprivations that government will inflict on any citizen, most enforceable limits in governmental power will be limits on criminal punishments. Further, because of hostility to criminal offenders, public opinion and democratic institutions will support punishments of high magnitude. Expanding power to punish in democracy will be easier than in other arenas of government, and limiting power through democratic institutions will be more difficult.

That the punishment enterprise is the boundary territory for the maximum exercise of negative government power explains the historic importance of criminal justice in the basic architecture of limited government. The Bill of Rights to the US constitution has eight amendments dealing with the relationship of government and individual citizens. Five of these eight amendments (4, 5, 6, 7, and 8) are chiefly concerned with criminal justice.

2. *An External Constraint*

Concern about excessive governmental power is external to the usual topics of effectiveness and individual desert in the determination of appropriate punishment. To worry about whether it is proper for modern governments to extinguish life as a criminal sanction is not to address either topics like deterrence or incapacitation that concern the effects of punishment on crime rates or to address what particular punishment might be morally justified by the commission of a particular act. Restraints on government power of this sort are prior conditions to setting a scale of punishments based on individual desert or deterrence. Citizens are potential victims in this setting both of criminals and of government; and they need not imagine themselves as criminal offenders to feel threatened by a government that is without restraints in pursuing its policies.

The external concerns that motivate restraints on punishment can produce debates in which the contestants are talking past each other, and in which differences of opinion exist even about the topic being debated. In many disputes about capital punishment,

retentionists typically assume that the topic is crime control and therefore the issues are individual desert and deterrence, while abolitionists see the topic as the appropriate limits of power in the modern state, a concern that has little use for criminological data or expertise. Often the debate is decided by which side succeeds in establishing its view of the topic under discussion.

What we are calling a 'human rights' perspective tends to be most prominent in circumstances where citizens have special concerns about the organization and limits of government. Change of regime is one obvious circumstance where citizens have pressing concerns about limiting governmental power. The widespread abolition of capital punishment in Central and Eastern Europe after the end of Soviet domination is one recent example of this. A second is that the abolition of the death penalty in Europe after World War II was first evident in the nations that were defeated and changed regimes, that is in Germany, Italy, and Austria (Zimring and Hawkins 1986). Only later did abolition spread to the major powers that were victorious, to the UK and France.

3. *The Politics of Governmental Limit*

All the major democracies in the world have substantial constitutional limits on the government's power to punish criminals that include prohibitions of torture, and all the major democracies except the USA and Japan prohibit the death penalty. All the developed democracies impose substantial minimal procedural conditions that must be met before punishing. The current pattern is that those nations with substantial democratic influence on government power also have more limits on punishment than nations with less citizen control on government, but this is no paradox. Authoritarian regimes do not choose to limit themselves and their citizens lack the power to force limits.

More interesting are the motives and mechanisms which limit punishment power in democracies. If a government responds to popular will, why not give it total discretion in the choice of punishment for criminal offenders? One problem with that position is that changing political circumstances is a risk that cannot be discounted. Governments may cease to be accountable, so that whatever safety came from popular control might disappear. But there is also a possibility of majority will supporting excessive use of force. In almost all nations, discrete minority populations support restrictions on government power in part because of a fear of popular control. And while self-interested enthusiasm for limits on government power is easiest to imagine for those with conspicuous minority identities, there are few of us immune from being outvoted in a matter of important personal preference. All citizens are potential minorities on contested issues.

So the reason democratic regimes have more restrictions on government power than autocratic regimes is that democracies allow such limits to exist. We know this because the strongest support for limits on government will come from those who have recently experienced autocratic governmental excesses. Limits on government power are quickly imposed once iron curtains fall.

How might a government constrained by limits on power tempt its citizens to remove such limits? Fear of crime and of criminals is one obvious technique, and this creates a strong motive for governments that desire additional power to exaggerate the danger of crime and criminals. The political task is to convince citizens that criminality is a greater threat to them than government excess.

Not all areas of criminal justice are of equal concern from a human rights perspective. The maximum dangers cluster at the extreme edges of the system—the most serious offenders, the most fear-inspiring crime problems, and the most extreme punishments. The most serious and hated criminal offenders are the justification for new incursions of state power. The serial killer, the murderer of children, the predatory sex offender, and the drug lord are the principal arguments put forward for extensions of the severity of punishments available and the length of incapacitation. Terrible crimes are always urged as the justification for terrible punishments. But more than the heinousness of individual offenders, the justification for new intensities of punishment depends on the public's sense of insecurity in combating the particular problem that is urged as the necessity for the extension of state punitive power. It is citizens' sense of vulnerability more than their distaste for particular criminal offenders that justifies the extension of state power. A sense of public emergency thus becomes the leading enemy of moderation and the limitation of government power.

The importance of public feelings of vulnerability about crime to support for the expansion of punitive power explains not only the incentive of ambitious governments to scare their citizens but also the cyclicality of the public sense of emergency about particular crime problems. From the war on drugs, to gang violence, to the Mafia, each campaign to make citizens feel acutely threatened has a relatively short effective life after the threatened Armageddon does not happen. So governments intent on consistent expansion of punishment powers will need to rotate the particular problems that are pushed to justify penal inflation. When drug wars are followed by panics about juvenile violence, the shift in topic may be the only way supporters of expansion in governmental power can generate consistent levels of public anxiety.

While extreme forms of criminality are the focus of support for extensions of governmental power, extreme forms of punishment attract the attention of the enemies of unlimited governmental power, particularly when a punishment is distinct from more standard penal measures in kind as well as degrees. Torture, beatings, and execution have been the targets of choice for human rights reformers, practices which inspire citizen empathy with the subject of the punishment, and types of deprivation that are easy to distinguish from more common penal measures such as imprisonment and fines.

Penal practices which differ from those in common use in extremeness but not in kind are more difficult targets for the human rights reformer to identify and to attack. An illustration of this phenomenon of current importance concerns what has come to be called 'the maxi-maxi prison' that has proliferated throughout the USA since the 1980s. Designed to impose unprecedented levels of individual isolation and psychological deprivation, the maxi-maxi prison is nonetheless a more difficult target for the human rights reformer than canings or torture because the institution can be defended by government as merely another form of prison.

The general acceptability of imprisonment as a penal sanction becomes that major premise in a syllogism; the minor premise is that these maxi-maxi institutions are prisons; and the conclusion is that therefore they must be acceptable penal measures. To date, these extraordinary institutions have not received the attention they deserve from either reformers or the wider community.

4. *Impulse and Obligation in Conditions of Imprisonment*

Imprisonment is the most serious penal sanction that is frequently imposed in the modern state and a stunning example of the conflict and complexity involved in the protection of individual dignity during punishment. The imprisoning state administers an institution where the totality of the offender's life is under its control and restraint. For this reason, sociologists describe such facilities as 'total institutions.' For the modern prison, the same government that condemns and wishes to punish offenders controls every aspect of their daily life: eating and drinking, social contact, excreting, communication with the world outside the prison, exercise, light and darkness are all state-administered conditions in the modern prison.

The totality of state power over individual life means that the punitive impulse that inspires imprisonment could in theory influence every aspect of the conditions of confinement. Why should those who have murdered and raped get decent meals or the opportunity to communicate with family and friends? What is the argument for allowing those who are being punished by imprisonment to enjoy movies on cable television or build their muscles in exercise rooms that have been equipped at public expense? One version of objections to positive conditions in prison acquired the label 'less eligibility': an argument that conditions in confinement should not be better than those available to the poorest members of the law-abiding community. But the broader argument is why not make the conditions of confinement in penal institutions punitive in all their details?

Two points of importance in considering the limits of punitive agendas on conditions of imprisonment echo general concerns which we adverted to earlier in this article. The first is that the source of any limits on the punitive content of imprisonment will be external to the purposes of punishment. One cannot derive compelling arguments for cable television in prison through analysis of concepts like deterrence and retribution. The second point is that the popular will in democratic government will be an unlikely source for conferring amenities on prisoners. The general hostility to the subjects of imprisonment is not self-limiting.

The limiting principal necessary to minimum standards of confinement is the governmental obligation to respect the human dignity of the prisoner. The content of this obligation is not easy to determine and will vary with economic resources and other social conditions. Deprivations that are not necessary to maintaining the restriction on liberty that make prisons punitive are gratuitous, and the claims of the prisoner as human being against the state make most gratuitous punitive conditions of confinement seem problematic. But the most that one can conclude from a discussion of the purposes of punishment is the lack of necessity to push further, a sense of what might not be necessary in order to condemn the punished offender or to deter others.

Yet more than a general notion of parsimony informs prohibitions on torture and execution. And the use of torture is prohibited even when it might prove quite useful to interrogations. As with torture, we believe that gross conditions of isolation and sensory deprivation that are intentionally imposed in hypermaximum security prisons in the USA also violate free-standing conceptions of human dignity that are no less important because they are imprecise.

But enforcing restrictions on punitive conditions of confinement through the political process is difficult not only in autocratic political regimes, but also in democracies. The authoritarian government is unwilling to constrain the tactics it uses in conflict with domestic enemies. In democratic regimes, citizens will be unwilling to spend resources to confer benefits on hated criminal offenders, particularly when politicians portray spending money on prison conditions as a status competition in which the citizen must choose

between the interests of prisoners and the interests of crime victims (Zimring et al. 2001 Chap. 11). Under such circumstances books for prison libraries and educational programs in penal institutions might be regarded as a direct affront to the victims of crime.

Enforceable limits on the punitive content of imprisonment can be maintained in two ways even when they are unpopular. The first is by a commitment of government to constitutional principles that render some conditions of confinement beyond the reach of governmental discretion. This is a 'bill of rights' approach to the limitation of gratuitous punishment. It requires enforcement somewhere in government, but not necessarily by the executive branch that administers prisons.

A second strategy for protecting imprisonment from gratuitous punitive content is to make the detailed conditions of imprisonment the responsibility of expert governmental actors whose discretion is insulated from populist politics (Zimring et al. 2001 Chap. 10). This has the advantage over a 'bill of rights' approach of allowing detailed consideration of conditions of confinement by governmental officials who have ongoing responsibility for administering imprisonment. The disadvantage of delegating to the warden and the prison guard is the threat of conflict of interest, that prison administrators will themselves prefer unjustifiably punitive conditions for their least favorite prisoners for reasons of animus or administrative convenience. The recent history of maxi-maxi prisons in the USA suggests that administrative controls within the executive branches of government have been manifestly insufficient protection of the human dignity of prisoners.

5. *State and Society in the Punishment Enterprise*

When governments push for increased power to punish criminals, they assert an identity of interest between the state and society in the punishment process. Executions are necessary to protect the community from murderers and rapists, and preventive detention is required because individual citizens demand safety rather than for any distinctive self- interest of the government. The plea for power is premised on an identity of state interests, community interests and individual interests.

Those who oppose the expansion of state punishment power will typically emphasize the differences between the interests of government and of social groups. In this view, state power can suppress rather than facilitate the interests of social groups and individuals. State power is presented not as an extension of social power but as power over society exercised by outsiders. This contrast in the rhetoric of punishment limits suggests a relationship between state and society that should predict the appeal of restraints on punishment power. In settings where there is an ideology of either identity or strong overlap between state and community, limits on government power to punish will be a hard sell, even if the populace has some control over government policy. States which govern simple and homogenous societies in accord with traditional principles will be difficult places to argue for a bill of rights, just as the identity of state and society assumed by theocratic and totalitarian regimes makes it difficult to imagine the appeal of limits on government to those who govern.

By contrast, limitations on government power have a natural constituency for citizens and social groups when the interests of government are regarded as separate from those of significant social groups. The simple example here would be the popularity of limited government to a population that has been recently been occupied by an alien political regime. For the same reasons, states which govern complex and plural societies will have a much more difficult time convincing citizens that the government will instinctively serve the interests of constituent social groups. The larger the number of groups and the more diverse the values and interests of groups within the boundaries of a political state, the more likely will be a consensus that limits on governmental power to punish serve the common good.

This simple hypothesis can help to explain the relative popularity of limits on government power to punish both cross-sectionally and over time. Those nations that encourage complexity and diversity also embrace limits on the exercise of governmental power generally, and punishment power specifically. The theocratic regimes of Islam are notoriously inhospitable to constitutional limits on government because the apparatus of state power is regarded as the faithful servant of a single dominant social group. In totalitarian states, on the other hand, limitations on governmental power are unthinkable because social groups are regarded as the faithful servants of the state.

Trends over time in state and society provide some reason for hope, but not without cautionary countertrends. Those areas of the world where pluralism and complexity are encouraged have become the breeding ground for domestic and international limits on government. The European community and the British Commonwealth nations are examples of political development in the direction of constitutionally limited regimes. Globalization and multinational economic organizations favor the development of limited government. But the march of history has not been inevitably in that direction. Totalitarian regimes were a twentieth-century development that often succeeded simpler social frameworks, and there are also strong theocratic trends in some areas of the world that run counter to theories of inevitable liberalization.

6. Penal Restraint and the Human Rights Agenda

In one key respect, the punishment of criminals is not a frontier issue for consideration of the obligation of the state to its citizens. Criminal justice is not the area where governments can be expected to innovate in the extension of opportunities and entitlements to citizens. Such innovations will usually first be extended to dependent populations with much higher social reputations than criminals—the elderly, children, and to those handicapped through no fault of their own. The strategic role of the criminal offender is in defining the absolute minimum obligation of state to citizen, so that adequate provisions for prisoners is a defensive necessity. Assuring the human dignity of the murderer and the rapist is not a strategy for expanding the entitlements of schoolchildren and senior citizens; instead it is a strategy to prevent the erosion of citizen claims against government, to prevent regressions applied to the least popular of dependent populations which might thereafter be applied more broadly.

The defensive significance of offenders' interests in the human rights dynamic has two implications. The first is that the entirety of a human rights agenda can never be exclusively based on offenders' interests. The defence of prisoners' rights is always a necessary but never a sufficient condition for a human rights agenda.

A second corollary of the defensive strategic position of offenders' interests is that efforts to protect this particular flank will be much more important in bad times than in good times. In periods characterized by good will and the growth of government concern for citizens, the advocacy for offenders' rights will be somewhat easier in its own right and less necessary for the protection of other vulnerable populations.

When a social climate turns threatening, the defense of offenders' interests becomes at once tactically more difficult and strategically more important. It is during eras of bad feeling that the interests of vulnerable populations are most at risk. The path of least resistance to the erosion of individual rights will always be to target those domestic enemies assembled in the criminal justice process.

See also: Civil Liberties and Human Rights; Criminal Justice, Sociology of; Human Rights, Anthropology of; Human Rights, History of; Human Rights: Political Aspects; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects

Bibliography

Garland D 2001 *The Culture of Control: Crime and Social Order in Late Modernity*. University of Chicago Press, Chicago

Temple W 1934 The ethics of penal action. First Clarke Hall Lecture. London

Zimring F E, Hawkins G 1986 *Capital Punishment and the American Agenda*. Cambridge University Press, New York

Zimring F E, Hawkins G, Kamin S 2001 *Punishment and Democracy: Three Strikes and You're Out in California*. Oxford University Press, New York

F. Zimring and G. Hawkins

Criminal Justice, Sociology of

1. The Origins of the Concept

The term 'criminal justice system' (hence designated as CJS) is a recent acquisition of the scientific and professional vocabulary. Since the publication of the United States. President's Commission on Law Enforcement and Administration of Justice report 'The Challenge of Crime in a Free Society' (1967), the term has been used with a somewhat precise meaning.

However, in the process, the Commission's investigation led to the claim that the goal of the 'war on crime'—the fight against insecurity and the foundation of public safety of the persons, goods, and moral integrity of the citizenry—necessitated a comprehensive concept inclusive of all the institutions concerned in this objective. The breadth and complexities of this conceptualization led to the identification of incongruities between some of its parts, allowing critics to call it a 'nonsystem,' referring to the apparent incoherences and contradictions in the functioning of the overall system.

It became evident for the Commission that there existed a real and a virtual inter relationship, reflecting an interdependency between legislations, institutions, organizations, and scattered offices throughout the machinery of criminal justice, whose goal is the maintenance of public order, the safeguard of liberty, physical integrity, and concern for the victims of the failings of the protective shield, extended by the law and protected by the Constitution. Recognizing these historically and nonlogically ordered institutions, laws, services, and so forth, the Commission wanted to give a degree of coherence to all of them in order to secure a better, more efficient coordination, built around a common goal: assuring, under the supremacy of the law and the Constitution, equal justice and protection to everybody. Despite the criticism, the concept of the CJS was born. Thanks to this organizational principle (a 'system' necessitates some logic), overlapping jurisdiction, duplication of services, and irrelevant goals or functions were identified. The Commission's logo—with its input and output/entrance and exit design—illustrated graphically how, from arrest to release, from sentence to punishment, the whole process of criminal justice constituted an administrative and logical 'whole.' This holistic approach provided the reformer with a blueprint and a compelling agenda for reform.

However, these recent origins of the CJS should not hide the basic fact that the inspiring and commanding center of their systematizing approach can be understood as a development of the *code d'instruction criminelle of 1808*. This code was characterized by (a) rigorous penal procedural rules; (b) the availability of material and technical conditions to implement those rules; (c) a reasonable degree of coherence among the constitutive elements of the system, principally the offices of the prosecutors, magistrates, and other officers of the judiciary. Similar transformations took place in other Western countries that incorporated a systematic approach. The principle of the separation of powers, the presumption of innocence, and the independence of the judiciary were nowhere threatened by the emergence of the CJS, with the notable exception of course in the totalitarian countries.

2. *Practical Impact of the Concept of the Criminal Justice System*

A radical transformation in public administration began as early as the middle of the nineteenth century, accelerating after World War I and continuing beyond World War II. One result was the emergence of an activist and interventionist social policy. A similar transformation in criminal justice administration did not occur until later. In the process of its consolidation, it confronted the cultural revolutions of the 1960s and the 1970s, and even some basic principles of the penal procedural law underwent some changes.

Among the consequences of the development of legal structures and philosophies were a loosening up of the system, allowing an increased role for a defense attorney (human rights legislation had a major impact on the proactive role of the defense councils), and resulting in an increase in the burden carried by law enforcement agencies to accommodate the increasing demands of defense attorneys as well as rules edicted by commissions of human rights (the need for 'foolproof' proofs required by tribunal became more cumbersome and proceedings slowed down notably). The increase of the workload in the prosecutor's offices, as well as at the courts, was not in step with the necessary budgetary requirements allowing a decent discharge of the increased workload devoluted to the CJS in charge of the 'post-arrest' phases.

Additionally, there was a growing incoherence between segments inside the diverse components of the CJS. Undue competition has arisen between services, which should have been submitted to rules prescribed by their common goals. Moreover, a growing politicization has taken place. The power of unions is increasing everywhere: the judges, the prosecutors, and police are organized in their own unions. Some unions of magistrates, particularly in France and in Italy, are playing an active and sometimes radical role inside of the CJS. Thus the coherence of the CJS has suffered from the centrifugal forces in its own ranks. This worldwide phenomenon is affecting open societies, and while the European continental nations, in line with their centralizing traditions, are seeking new measures in tune with their traditions, all of them are also facing similar problems.

3. *Components and Goals of the Criminal Justice System*

The 1967 Presidential Commission, which is *de facto* the originator of modern research on crime, delinquency, victimology, and criminal justice agencies, introduced and somehow accredited a rather artificial distinction between research carried out on the CJS and criminology. This discipline was related, if not assimilated, to etiological research on crime causation. Its continental origins were stressed. Criminal justice research, born of the common law tradition, postulated the existence and interdependence of all components of the CJS, including law enforcement, criminology, criminal justice administration, prosecutor's offices management, sentencing guidelines for judges, and so forth. Of course, 'corrections,' a new designation for penology, was a major aspect of research on the CJS. These distinctions also reflect a growing attention to the evolution of ideologies, to new tendencies of penal philosophies, justifying research on alternatives to the actual CJS, which answers to the needs of victims and new sentencing philosophies. The analysis of the performance of the CJS, evaluation of alternatives for this system, and all kinds of innovations, were also an integral part of research on the CJS. Also, the treatment of offenders, their resocialization, programs of crime prevention, and experiences with 'restorative justice' became the spearheads of research in criminal justice.

Many of the major changes which the CJS underwent are related to the sudden surge of the world economy and commerce in the direction of globalization. Accordingly, the problems facing the CJS became global, embracing the whole world. One has to face the fact that crime and justice have to be addressed in an open world, at the level of the global economic and financial system working with the tools of the newest information and technology. The sociopolitical system and the political power structures are closely related to the new financial, commercial, and economic trends and powers. The magnitude of the challenge facing all the components of the CJS is tremendous. Consider, after 1980, that the volume of the sales of foreign stocks and bonds increased drastically on the market (Adler et al. 1994).

However, many countries are largely decentralized and have residues of common law traditions. There, the integration and internal coherence of the CJS is limited; rather the tendency is toward internal incoherence and diversity. Thus it is not surprising that

these countries, illustrated by the USA, are much slower to ratify or contribute to international criminal and penal jurisdiction. An example is the US government's refusal to adhere to the Naples' convention sponsored by the UN to create an international penal tribunal.

4. Criminal Justice System and the Courts

In this context of global integration, the thrust toward internationalization increasingly impacts local crime scenes (arms dealing, etc.) so that, directly or indirectly, they become integrated into international activities through the operation of off-shore banking, among other devices, intermingling, sometimes, with secret service operations (e.g., the Panama-Noriega scandal); and with illegal merchandising traffic, tax frauds, illegal financial contributions to political campaigns or parties, and so forth. Money laundering is the lifeline of all these criminal and illegal activities. Out of necessity a new international CJS is emerging with the support of the UN and the European Union (EU). Other tribunals in Africa and Asia are in process of being established.

One can see that open democratic societies are building new lines of defenses using legislations and sharpening legal tools available to the police and prosecuting attorneys. A new Europe-wide legal system is being created and is slowly and systematically building structures and defenses, which are both reactive and proactive. The building of such a coherent system was, in part, a response by the CJS to challenges initiated in continental countries under the auspices of the EU.

5. The Debate on the Definition of Deviance and Criminality

In the 1970s sociology questioned what distinguishes a 'criminal' act from a 'deviant' one, thereby adding a substantial burden to the functioning of the CJS. Such questioning created difficult dilemmas for the CJS, whose task is to enforce the existing laws. Problems arose as to where to draw the limits of the interventions of the jurisdictions of the CJS agencies in controlling or repressing the 'criminal' and unlawful activities of large, and certainly vocal, minorities. The frontiers between crime and deviant behavior became blurred. The law enforcement agencies, occupying the first line of defense in securing public safety, became hesitant in the application of the law. This environment of questioning and doubt encouraged the increasing activism of human rights defense groups, critical of the law. And when the police tried to enforce questionable laws, they were frequently confronted by activist groups, sometimes with violence. Many court actions followed violent confrontations in the streets. Also, accusations against the police for being prejudicial created a major impediment to efficient policing.

The manifold increase in the charges burdening the prosecutor's offices, mainly in major cities, reflected itself in the overflow of cases brought in front of the judges. Plea bargaining before the trial (essentially a kind of out of court settlement between the prosecutor and the defense council) became an accepted way of proceeding. For several decades it has been the only way to avoid the complete paralysis of the whole judicial process. These negotiations are questionable, morally as well as legally, and sometimes they are clearly illegal, although they are still currently practiced in the US courts system. In contrast, these practices are strongly rejected by the European continental legal system, but the future of its integrity is, unfortunately, an open question. The caseload of the courts is increasing constantly and political power negatively confronts the judiciary in several important Western European countries. The main issues are political corruption, bribery, and abuse of power. We have yet to see the outcome of this type of conflict inside of the continental system where political power negatively confronts the CJS.

6. Changes in the Philosophy of Punishment

Imprisonment was a notable advance over the all-too-common previous method of punishment: the death penalty. Virtually from its beginnings imprisonment developed in response to religious concerns (largely from nonconformist denominations—Methodist, Congregationalist, Presbyterians, Quakers, Baptists, etc.), and was informed by ideas of the continental Enlightment, as well as English social contract and utilitarian theories, and the views of Scottish moral philosophers.

In the north-east sector of the USA these religious and intellectual forces combined to encourage a massive social experiment in punishment, which was grounded in the emerging belief that humans are social beings possessing 'natural' or inalienable rights, capable of reason, and inherently good or virtuous: the infant cannot be evil. From this perspective the logic of imprisonment would not only punish the wrongdoer, but seek also to rehabilitate that wrongdoer, as well as to deter others. Elmira, Auburn, and Cherry Hill were famous attempts to express this evolving penal philosophy, and de Tocqueville's tour of these institutions was an example of European continental interest.

This complex and integrated humanistic view, issuing from the Enlightment, was challenged by developments in sciences such as medicine and psychiatry. These disciplines understood criminal acts to arise from personality disorders rooted essentially in social, biological, or psychological conditions. In a sense, the 'medical model' retained but transformed

elements of the 'humanistic model.' The goal of rehabilitation was retained, but its justification was to be found and legitimized in scientific methodology. One consequence of this scientific approach, as it gained influence, was to impact the CJS through legislative reforms and judicial practices, for example, the rapid spread of indeterminate sentences, and the transferal of the principle of rehabilitation from being a concern of the larger society (religious activists, legislators, educators, etc.) to become the concern of medical sciences.

7. *Emergence and Hegemony of the Medical Model: Conflicts of Goals with the Legal Tradition*

The nature of the scientific method is the permanent questioning of existing proofs and practices. For example, in medicine, it results in a continuing revision of therapeutic practices and rehabilitative techniques. Naturally, a continuing tension and conflict develops between the ever-changing medical-therapeutic approach and the desire of the CJS to retain humanistic principles of equity, rights, and equal treatment before the law. This situation threatens its coherence and creates an increasing vulnerability for the CJS, especially to its claims of legitimacy.

The challenge to legitimacy was not limited to the confines of the CJS, but permeated the entire open, pluralistic, humanistic society. Everywhere, a profound crisis of values occurred: in the family, and in institutions of authority, including the medical and academic communities. The identification of who or what was responsible for this turmoil emerged from a variety of diverse sources, making it impossible for a generally agreed upon understanding to be reached, and so disarray prevailed.

The impact on the CJS, which became closely associated with the medical model, was disastrous in that it allowed for both an immunization of the 'underclass' against claims of criminal behavior, and the emergence of new prophets with new social and cultural constructs. An attempt to correct this situation, and to devise a means of bridging the chasm separating the reality of crime and the CJS, was Von Hirsch's (1976) doctrine of 'just deserts.' This doctrine promised law and order while accommodating the new forms of behavior. The end result was a temporary compromise in which the CJS maintained and protected the core values concerning material and personal security, while adapting itself to the demands of the 'post-Woodstock' society.

8. *Some Future Perspectives*

The real and symbolic ruins that were the reality of the CJS under the conditions of this temporary compromise were, fortunately, addressed by H.L. Packer (1968). Packer identifies two models. The first is the 'due process model,' characterized by the jurisprudence of the Warren Court, in which most of the odds were placed against the defender of public safety and order. This model amplified the 'rights' of the individual, and encouraged a critical, reactive attitude on the part of the judiciary. This model has a rule-oriented rather than a substantive-oriented approach.

The second is the 'crime control model,' which emphasizes the existence and exercise of official power. Its validating authority is ultimately legislative (although approximately administrative). It can be viewed as consistent with, and a continuation of, humanistic-Enlightment principles. In light of the lessons of the eighteenth century, and because Packer's theories are deeply rooted in humanistic-Enlightment principles, one can surmise that his crime control model will prevail in the near future.

9. *Conclusion*

In conclusion, the contributions of the social sciences to the CJS are varied. In continental Europe, the social sciences, as an academic discipline, had limited influence, notwithstanding contributions of outstanding scholars. In North America the social sciences have established a strong academic position and exercised a direct influence on the CJS. This influence also included a contribution to the intellectual and moral crises which characterized the 1960s and the 1970s. This was largely due to an epistemological crisis, which shook the very foundations of social and behavioral sciences. Whether or not they will recover their empirical scientific tradition is still an open question. Hopefully, the positive renewal of the CJS, influenced in part by Packer's writings, may have a spillover affect into the social and behavioral sciences.

See also: Control: Social; Courts and Adjudication; Crime and Class; Crime and Ethnicity (Including Race); Crime, Geography of; Crime: Knowledge about and Prevalence; Crime: Sociological Aspects; Crime, Sociology of; Delinquency, Sociology of; Imprisonment: Sociological Aspects; Justice and Law; Juvenile Justice: International Law Perspectives; Law: Economics of its Public Enforcement; Law, Sociology of; Police and Politics; Police, Sociology of; Policing; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Rehabilitation of the Offender

Bibliography

Adler F, Mueller G O W, Laufer W S 1994 *Criminal Justice*. McGraw-Hill, New York

Bridges G S, Weis J G, Crutchfield R D (eds.) 1996 *Criminal Justice Readings*. Pine Forge Press, Thousand Oaks, CA

Cohen J 1983 Incapacitation as a strategy for crime control: Possibilities and pitfalls. In: Tonry M, Morris N (eds.) *Crime and Justice: An Annual Review of Research*. Chicago University Press, Chicago
Cullen F T, Gilbert K E 1982 *An American History of Correction*. Anderson Publications, Cincinnati, OH
Jacoby J 1979 The changing policies and prosecutors. In: McDonald W F (ed.) *The Prosecutor*. Sage, Beverly Hills, CA
Maguire M, Morgan R, Reiner R (eds.) 1994 *The Oxford Handbook of Criminology*. Clarendon Press, Oxford, UK
Menninger K 1968 *The Crime of Punishment*. Viking Press, New York
Mouledoux J 1972 E. Livingston (1764–1836). In: Mannheim H (ed.) *Pioneers in Criminology* 2nd edn. enl. Patterson Smith, Montclair, NJ
Packer H L 1968 *The Limits of the Criminal Sanction*. Stanford University Press, Stanford, CA
Schrag C 1961 Foundations for a theory of corrections. In: Cressey D R (ed.) *The Prison: Studies in Institutional Organization and Change*. Holt, Rinehart & Winston, New York
Sykes G M 1958 *The Society of Captives*. Princeton University Press, Princeton, NJ
United States. President's Commission on Law Enforcement and Administration of Justice 1967 *The Challenge of Crime in a Free Society*. U.S. Government Printing Office, Washington, DC
Von Hirsch A 1976 *Doing Justice: The Choice of Punishment* 1st edn., Hill and Wang, New York

D. Szabo

Criminal Law and Crime Policy

1. Introduction

While the forms and functions of criminal justice institutions vary significantly around the world, trends can be observed that reveal a shrinking planet where countries are heavily interdependent, capitalism is triumphant, and pressure is keen for universal norms of political, economic, and social life. Although data from developing nations are often either inadequate or nonexistent, and much crime policy seems to be in flux, the available information suggests similar patterns in both Western and non-Western nations (with exceptions for some Islamic states). Whether the society is democratic or authoritarian, industrialized or developing, its police and courts are expected to deal strictly (often harshly) with law-breakers—particularly those who have committed violent offenses and violations of drug laws—and to respond to demands for protection from the public, ordinary citizens as well as elites. Rapid developments in transportation, communications, and trade have provided new opportunities for transnational crime and local crime increasingly to reflect and shape global events. In response, the nations of the world are trying to coordinate strategies to address crimes like drug trafficking, arms smuggling, and fraud that cross national borders. Crime policy is affected by political change, too. The spread of constitutional democracies on several continents and the breakdown of traditional mechanisms of informal social control in the Third World are creating a new awareness of rights in nations where they have previously been unknown, a development with significant implications for peoples' expectations that criminal justice will operate efficiently and fairly.

As with the policy measures of the early twentieth century, get-tough morality—this time in the form of a worldwide war on drugs, fought with treaties and international cooperation as well as with conventional law enforcement—coexists with efforts to find alternatives. Today, however, the innovations are often taking place outside state-centered justice systems, in the spread of private policing and the use of community dispute resolution in both traditional societies and industrialized urban areas. New paradigms for dealing with incidents that disrupt community harmony—principally the restorative justice ideals behind programs like 'family conferencing' in New Zealand, Australia, and elsewhere—seem likely to have a communitarian influence on criminal justice. The full effect of globalization on official social control remains to be discovered.

2. The Punitive Trend

No clear relationship exists internationally between levels of crime and the prevalence of official punishment. A 1997 analysis of 34 countries conducted as part of the Fifth UN Crime and Justice Survey found no significant associations between sentencing rates (per 100,000 population) and crime rates (Newman 1999, p. 91). Similarly, harshness in penal law may accompany either increasing or decreasing crime rates. Crime control policies in South Africa and the United States, for instance, show quite similar directions, even though trends in the incidence of crime are very different. In the 1990s, violent crimes like robbery, assault, and car-jacking were on the rise in South Africa but falling in the United States. Both countries, however, adopted longer and mandatory sentences, asset forfeiture laws, and limits on bail eligibility. Penalties for property crimes continued to be relatively light throughout Western Europe during that period, despite significant increases in rates of offending in several countries, while in North America lower rates for such crimes did not reduce sentences.

These illustrations of statistical patterns suggest two important realities of crime policy around the world at the millennium. One is that it is generally quite punitive, at least if one defines 'punitive' to mean the criminalization of many behaviors deemed undesirable and heavy reliance on the deprivation of

liberty. The other is that the United States—in many respects the most punitive of the developed countries (and growing more so all the time)—has tremendous direct and indirect international influence.

In the 1960s Packer articulated two conceptions of the criminal justice process—the crime control model, which gives priority to the efficient capture and containment of potential and actual law-breakers, and the due process model, which looks beyond immediate community protection and seeks to guarantee fundamental fairness to all those who come within the reach of the criminal justice system (Packer 1968). While scholars and policymakers are still wrestling with the tension between these two concepts of the function of the criminal sanction, it seems clear that the crime control perspective dominates the operations of police and courts around the world.

According to data from the United Nations Surveys of Crime Trends and Operations of Criminal Justice Systems (UNCJS)—admittedly limited to those countries that respond and compromised by inevitable differences in reporting standards, methods, and capabilities—prison is a more popular sanction in most countries than noncustodial penalties like fines (a common alternative) or community supervision. (International numerical comparisons in this section come from the Fifth UN Crime and Justice Survey (Newman 1999, pp. 89–119) unless otherwise noted.) Whether it is Vietnam—where smuggling as little as 100 grams of illicit drugs can warrant the death penalty—or the United States—where federal guidelines instruct judges to sentence youth more harshly than adults for commission of the same crimes—harsh punishments have great appeal.

One standard measure of harshness is the rate of incarceration per 100,000 population. On this score Russia and the United States are the clear leaders, Russia with a rate of 411 in 1994 and the US with a rate of 468 in 1999 (as of June 30), the last years for which data are available (Bureau of Justice Statistics 2000). Many countries' incarceration rates rose substantially during the first half of the 1990s, particularly in Eastern Europe; in Italy the rate increased 100 percent between 1990 and 1994, perhaps due to a new and harsher drug law passed in 1990 (repealed 3 years later).

Despite opposition from the Catholic church, international political bodies, and human rights organizations, the death penalty continues to be used widely in developing countries and in the United States, where 3,652 people were on death row at the start of year 2000 (NAACP Legal Defense and Education Fund 2000, p. 1). Amnesty International reports that 1,813 people were known to have been executed in 31 countries in 1999, with China, Congo, Iran, Saudi Arabia, and the United States accounting for 85 percent of that number (Amnesty International 2000). Many countries show considerable ambivalence towards capital punishment. Russia abolished and reinstated it three times in the twentieth century, and the Philippines recently began executing people after a 20-year moratorium. More countries are likely to abolish the death penalty since it is a requirement for membership in the European Union.

Criminal justice policy can also be assessed by noting procedural protections provided to accused persons. Although the common law countries tend to emphasize constitutionally-mandated defendants' rights (like the presumption of innocence and rigorous standards of proof at trial), it is by no means clear that in practice they are more protective than civil law countries, which often incorporate similar protections into the inquisitorial process in less formal ways. It is in English-speaking countries, in fact—Britain, South Africa, India—that the right of the accused to remain silent has recently come under attack; in the United States, the imperatives of the war on drugs have caused significant erosions of the prohibitions against unreasonable search and seizure and self-incrimination. In 1999, the British Home Secretary called for an end to the ancient and automatic right to trial by jury. He also criticized the prohibition against trying a defendant more than once for the same crime (double jeopardy), a constitutional provision in the United States that has been weakened by a number of court-approved exceptions.

The most obvious influences of United States' criminal law and crime policy on other countries' punitive trends can be seen in innovations related to the war on drugs, among them the forfeiture of suspects' assets, money-laundering laws, and elaborate sting operations. The long arm of US policy has reached into other areas as well. Laws that require community notification when a sex offender is released from prison have gone into effect in the United Kingdom and Canada. Another American export is habitual offender legislation (the 'three-strikes-and-you're-out' policy that helped to swell California prisons at the end of the twentieth century), which is now common in Europe and has been adopted recently in Australia and New Zealand. American dominance meets with occasional resistance, but the trend is likely to continue as long as federal law enforcement personnel staff American embassies and provide training in crime control both at home and abroad.

The punitive trend is not universal. UNCS data from 1994 indicate that fines comprise more than 70 percent of sentences in a number of countries on three continents; the survey found that in Japan fines were imposed in 95 percent of cases. One study of increases in violent juvenile crime and policy responses in Europe and the US found that, although in Britain, the Netherlands, and the US sentences have been lengthened and procedural protections narrowed since the late 1980s, most of the countries studied have retained programs that provide options besides secure detention and are working to develop social and vocational services intended to prevent juvenile crime

(Pfeiffer 1998). Ideas about alternatives to incarceration that have become standard for minor offenses in Japan are taking hold in developing countries, though they have yet to materialize as policy. In India in 1996, the Law Commission recommended legislation providing for noncustodial punishments. In Kenya in 1999, a conference of criminal justice professionals, UN representatives, lawyers, and judges pledged to work toward the decongestion of prisons by expanding alternatives to incarceration for minor and first-time offenders. (It should be noted, however, that these sentences are not always less punitive than short prison stays and that they rarely reduce incarceration rates.)

3. *Policing and the Public*

A prominent comparative scholar of policing recently has warned that most of the world's police forces are 'regime police,' answerable to government and its needs and 'organized to be responsive upward' (Bayley 1999, p. 3). The undemocratic essence of such policing is its inability to respond to public demands and the resulting unwillingness of members of the public to turn to police in anything less than emergency situations. Two worldwide developments—community policing and government investigations of police behavior—may soon alter Bayley's conclusion. Through these innovations, policymakers, concerned citizens, and international organizations are seeking greater police efficiency, accountability, and transparency with unprecedented vigor, though it is difficult at this point to say to what effect.

'Community policing' is an international buzzword that may mean anything from citizen patrols in Kuala Lumpur and Singapore to the highly professional commitment of police working with residents and city agencies in Chicago neighborhoods (Skogan and Hartnett 1997). The vagueness of the term and the potential for abuse—empowering vigilante groups, dividing neighborhoods by building networks of citizen informers—gives warning of its antidemocratic possibilities, and flaws in concept and implementation are notable everywhere. But in theory, community policing promises visible patrols, decentralized command, and the chance for the public to communicate with police about community problems (Skogan and Hartnett 1997, p. 6). Its appeal is an indicator of the public's desire for participation in and oversight of the institution that is at least theoretically charged with the politically neutral protection of a community's security. In volatile and crime-ridden Bogota, Colombia, the police chief sees community policing as a remedy for public distrust; in Lusaka, Zambia, businessmen assist the officers assigned to patrol the commercial district; in March 2000, the Israeli Minister of Internal Security announced his intent to reallocate some resources from defense to community policing. Western countries that have already embraced a paradigm shift from the old aims of professionalism and paramilitary discipline in policing to the new community-oriented perspective are coming to the aid of less developed countries that want to get on board. The US has helped El Salvador initiate foot patrols and community relations efforts, the UK has held training sessions for Chinese police chiefs, and officials from the Czech Republic have been consulting with the Canadian Mounted Police.

Greater public accountability is also furthered by governmental concern over the police brutality, bias, and corruption that are endemic in many countries. In the US, the Crime Act of 1994 has given the federal government unprecedented authority to investigate abuses in local police departments; investigations of police practices in Los Angeles, New York, and Buffalo, among other cities, have resulted. In Britain, the race relations law has been expanded to include liability for police bias, following the charge of institutional racism made against the London Metropolitan Police after a botched investigation of the murder of a black teenager killed by whites. Anticorruption campaigns in Latin America and Africa include a focus on police, and local corruption in Japan led the National Police Agency to institute random inspections of prefecture management and conduct. While it is unlikely that major changes in police culture will occur soon as the result of these measures, they are reinforced by international pressures, some of which carry powerful incentives for change. Turkey's effort to comply with requirements for admission to the European Union includes new legislation that would encourage the prosecution of police who torture.

4. *Future Crime Policy Issues*

Even as the public police around the world are subjected to increasing scrutiny, reliance on them may be diminishing. Policing is no longer solely a state function but has been 'pluralized' (Bayley and Shearing 1996). In South Africa, Australia, the US, and many other industrialized countries, private guards outnumber public police several times over, providing security for companies, government buildings, and affluent neighborhoods. The policing function—combined with adjudication of minor crimes in some developing countries—is also carried out by volunteer groups, who are sometimes informal community associations, sometimes adjuncts of the criminal justice system. To some degree, these institutions are reflections of a community's view that the formal justice system is unresponsive or unfair, but they may also manifest a need for indigenous solutions to conflict that is fundamental to a particular culture (Ebbe 1996, Brogden 1996). Developing regulatory

schemes for these domains of policing seems likely to be a policy challenge of the future.

The globalization of trade, communications, and transportation has brought more than international consulting opportunities to criminal justice policymakers and practitioners. Controlling transnational crime has become a major issue. Celebrated incidents like the 1993 terrorist bombing of the World Trade Center and more routine (but often violent) crimes like drug trafficking across national borders have brought about calls for internationally coordinated crime control strategies (Muller et al. 1997, p. 137), but law enforcement in this context—the intersection of criminal law and international law—presents problems of sovereignty, proportionality, and privacy. Without adequate regulation, pre-existing agreements between countries may be unnecessarily abrogated and amorphous behaviors (like conspiracy) over-criminalized. Constraints on police authority and conduct, which are difficult to impose in a single country, will be even more difficult to enforce at the international level. International law enforcement efforts may be rejected or ineffective because of an insufficient understanding of the particular political and cultural traditions of individual countries.

Recognizing these potential pitfalls but acknowledging the need for curbing illegal traffic in drugs, arms, and immigrants, as well as bribery of public officials, the International Penal Law Association in 1999 adopted a resolution on transnational organized crime. It recommended a parsimonious definition of international crimes; the regulation of extraterritorial jurisdiction and police cooperation; the adoption of standards for judicial assistance and cooperation; and the protection of defendants' rights in international criminal proceedings (International Enforcement Law Reporter 1999). The issues raised by this resolution are likely to be debated for the foreseeable future.

Cooperation in fighting transnational crime must obviously also include information exchange, often electronic. The US Attorney General has proposed a Global Criminal Justice Information Network that would link criminal and intelligence databases from all over the world and would be run by the US Department of Justice. But the prospect of globally-shared investigative data and international data surveillance raises difficult issues of privacy, accuracy, and dissemination. It is unclear at this time what criteria would be used for the protection of data or how they would be enforced.

Two other issues highlight the increasing internationalism of efforts at crime control. Countries in transition from authoritarian to democratic regimes often experience rising crime. Borders that were formerly closed become permeable, encouraging the growth of multination organized crime syndicates. More generally, freedom brings license; Nelson Mandela, commenting on the increase in violent crime in South Africa's affluent white suburbs, has noted that liberation frees criminals, too. Corruption in democratizing countries is also common, though what appear to be an increase in incidence resulting from regime change may in reality be the effect of placing a new light on an old problem, as seems likely in some of the Eastern European countries. The determination of whose standards to apply in addressing these problems and how to arrive at international consensus is a challenge for international policymakers.

A related issue is the handling of war crimes, genocide, and crimes against humanity. The recent prosecutions of the former dictators of Chile (Augusto Pinochet) and Chad (Hissein Habre) by the courts of other countries may signal an end to the traditional deference countries show to one another's sovereignty where serious violations of human rights have occurred. The international law principle of universal jurisdiction—the idea that a former head of state can be tried anywhere for atrocities committed while in office—is becoming operational. Empowered by the United Nations Convention against Torture and national laws that confer jurisdiction over human rights crimes, a number of European courts have charged foreign officials with crimes committed in Rwanda and the former Yugoslavia. Furthermore, the establishment of a permanent International Criminal Court (in 2002) to try these crimes will further institutionalize a far-reaching concept of international criminal liability. Proponents of the court also believe that it will have a deterrent effect on heads of state, though its authority is likely to be weakened by the refusal of several countries, including the US and China, to support it.

5. *Conclusion*

At the dawn of the twenty-first century, crime policy in a number of countries reflects a departure from some of the assumptions of modern, liberal criminal justice systems. The aim of both civil and common law traditions is to elicit the truth of what happened in a particular situation that gave rise to a criminal charge and to determine the legal consequences of that truth for the individual offender (and perhaps for the victim, whose concerns are now getting increased attention). But legislators, judges, and scholars are beginning to depart from the exclusive goal of individualized justice and articulate a larger concern for what happens to community life as a result of the incident under consideration and others like it. This focus can take a number of directions, two of which are showing up as policy innovations with potentially far-reaching consequences.

Sometimes called 'actuarial justice,' the more common of these directions has entered the mainstream in American criminal law and criminal justice practice (Feeley and Simon 1994). Sentencing, probation, and parole have become risk assessment functions. Deter-

mining penalties and how offenders are supervised as they pay those penalties depends on the degree to which the defendants pose a risk of reoffending, given not only who they are but what their salient demographic characteristics are (age, employment status). The 'New Penology,' as two prominent critics have called it, no longer seeks to change or to punish lawbreakers, but 'to regulate groups as part of a strategy of managing danger' (Feeley and Simon 1994, p. 173). The other approach—a burgeoning movement in a number of countries for the resolution of minor disputes and deviance—is restorative justice, in which the victim and the offender join in a ceremony that acknowledges the wrong that has been done but seeks social reintegration of the offender and assistance to the victim, restoring harmony through a mutually beneficial process (Braithwaite 1989). One researcher reported in 1999 that at least 300 victim–offender reconciliation programs are operating in North America and over 500 in Europe (Umbreit 1999).

The managerial aspects of the actuarial model are obvious; the crucial decision is made by legally empowered figures in a centralized hierarchy of state institutions. They place primary emphasis on the outcome as the force that will assure community security and pay only cursory attention to individual autonomy. Restorative justice, on the other hand, relies on decisions made locally and jointly by the parties and others involved in the incident, and the community is expected to benefit from the process of arriving at the decision as well as from the resolution of the problem. If surveillance dominates actuarial justice, empathy and what Braithwaite calls 'reintegrative shaming' characterize restorative justice.

It is unclear which of these models of criminal justice will prevail in the next decades, or whether they will continue to co-exist. Taken together, their departure from the individualistic models of twentieth-century jurisprudence and criminology is stunning, overshadowing even the changing nature of policing, the privatization of prisons, globalizing influences on criminal justice, and get-tough sentencing policy. Their emergence, and the ways in which they will interact with the other developments mentioned, points to our century-spanning generation as a decisive moment in the history of criminal justice.

See also: Crime and Delinquency, Prevention of; Crime: Knowledge about and Prevalence; Criminal Justice, Ethics of: Aspects of Human Dignity; Criminal Justice, Sociology of; Death Penalty; Prisons and Imprisonment

Bibliography

Amnesty International. News release. April 18, 2000 (ACT 50/09/00)

Bayley D H 1999 The contemporary practices of policing: A comparative view. In: Burack J, Lewis W, Marks E (eds.) *Civilian Police and Multinational Peacekeeping—A Workshop Series: A Role for Democratic Policing*. Office of Justice Programs, US Department of Justice, Washington, DC

Bayley D H, Shearing C D 1996 The future of policing. *Law and Society Review* **30**(3): 585–606

Braithwaite J 1989 *Crime, Shame and Reintegration*. Cambridge University Press, Cambridge, UK

Brogden M 1996 The indigenisation of policing in South Africa. In: Marenin O (ed.) *Policing Change, Changing Police*. Garland, New York

Bureau of Justice Statistics, US Department of Justice 2000 http: www.ojp.usdoj.gov/bjs/prisons.htm

Ebbe O N I 1996 The judiciary and criminal procedure in Nigeria. In: Ebbe O N I (ed.) *Comparative and International Justice Systems: Policing, Judiciary and Corrections*. Butterworth-Heineman, Boston, MA

Feeley M, Simon J 1994 Actuarial justice: The emerging new criminal law. In: Nelkin D (ed.) *The Futures of Criminology*. Sage, London

International Enforcement Law Reporter 1999 International Penal Law Association adopts resolutions on transnational organized crime. *International Enforcement Law Reporter* **15**(2)

Mueller G O W, Friday P C, McCormack R, Newman G, Ward R H 1997 American crime problems from a global perspective. In: (ed.) *Critical Criminal Justice Issues*. US Department of Justice, Washington, DC

NAACP Legal Defense and Education Fund 2000 Death row U.S.A. (unpublished document)

Newman G (ed.) 1999 *Global Report on Crime and Justice* (United Nations Office for Drug Control and Crime Prevention). Oxford University Press, New York

Packer H 1968 *The Limits of the Criminal Sanction*. Stanford University Press, Stanford, CA

Pfeiffer C 1998 Juvenile crime and violence in Europe. In: Tonry M (ed.) *Crime and Justice: A Review of Research*, Vol 23, University of Chicago Press, Chicago, IL, pp. 255–325

Skogan W G, Hartnett S M 1997 *Community Policing, Chicago Style*. Oxford University Press, New York

Tonry M 1999 Rethinking unthinkable punishment policies in America. *UCLA Law Review* **46**: 1751–91

Umbreit M 1999 Restorative justice though juvenile victim–offender mediation. In: Bazemore G, Walgrave L (eds.) *Restorative Juvenile Justice: Repairing the Harm of Youth Crime*. Criminal Justice Press, Monsey, NY

D. Gordon

Criminology: Psychopathological Aspects

1. Introduction

A study on the relationship between criminology and psychopathology would have to start from questioning its existence, its terms, and whether it is one of causality or of association. A causal relationship could be based on nosographic parallelism because psychopathology and criminology share mutual grounds

so that symptoms are equal to crimes and vice versa, or on epidemiological decisions that some mental states cause criminality. However, a relationship of association would have to specify the factors through which the association holds. Aristotle (1941) already expressed an interest in the relationship between psychopathology and criminology, but the large number of mental patients in prisons and the publicity generated by major crimes in which a mental patient has been involved have increased the interest in later years. In addition, depictions of criminals with twisted minds in movies and television spread the opinion that crime and mental illness are related. Unfortunately, very infrequently does the media provide realistic portrayals of the abuse, homelessness, social disenfranchisement, or social neglect that are common reasons why some mental patients are arrested. In this article, therefore, the author has two purposes: first to disentangle the existence and parameters of a relationship between psychopathology and criminology and second, to review risk factors that could determine the relationship.

2. General Statement on Psychopathology

2.1 Definitions

Criminology is the scientific study of crime and criminals and their motivations for criminal behavior. Psychopathology is the study of personality factors that are somewhat out of regular conscious awareness and that lead to behavior outside the norm in a particular social group. It studies abnormal emotional or personality phenomena that underline symptoms or behavior. Psychopathology provides criminology with methods to analyze elements of a criminal act within the personality structure and aims to find out the motivations for the crime. The interest is centered on the individual, or criminal, as somebody who is different from the rest because of personal or social deficiencies that have led him or her to commit a criminal act. Act and actor are the focus of study and intervention. In this type of criminology of difference the crime is a fact, and the criminal is singled out as an object of study.

From a wider perspective, any criminological study would ideally include the context in which the criminal behavior has taken place. This interpretation involves the criminal as wrongdoer, sometimes even as victim, the identified victim, the determinant elements of the social circumstances in which the crime has been committed, and the deciding role of the law. In this criminology of process, the crime is the result of a conflict and the criminal is the subject, or actor, caught in a web of social circumstances and legal definitions that are required to identify the act as criminal. No act could be defined as a crime without the social and legal identification that places the actor within a conflict with the group that, eventually, qualifies the act as infraction deserving of sanctions (Debuyst 1997).

2.2 Psychopathological Methods

It would be tempting to study every crime and its actor from the point of view of a criminology of process, but this approach has countless ramifications about the meaning of the act; and teleological implications about criminal responsibility. Although this perspective has high heuristic value, it would contribute little if the aim of the study is to gain a deep knowledge of the criminal. From a practical standpoint, singling out a person for study and dissecting the motivations underlying the act makes a statement that the criminal is different and in need of being separated from the rest of the group. Separateness is an objectifier that demands knowledge of psychopathological methodologies to do the separating and the dissecting.

There are crimes where the motivations could be placed entirely outside the range of personal psychopathology, but these are rare. For example, highly politically organized crimes take the victim as the representative of an order considered oppressive by a majority. In these crimes the immediate actor may not have personal motivations, but acts as an instrument of the group. Although such pure, politically motivated crime on behalf of the group could be studied from a historical, political, and social perspective, it would still have to be examined against a possible backdrop of social pathology, and the motive of the immediate actor would have to be examined to determine personal pathology. Regicides and magnicides committed by a lone criminal without social backing usually relate to personal psychopathology. Other forms of political crime, such as a leader's need to eliminate his or her opponents, although couched on political terms, would have to be examined as to whether the leader's thrust for power at all costs, including the need to preserve power that allows for satisfaction of plain personal greed, is rooted on psychopathological demands. In Mafia executions, for example, when somebody is 'rubbed out?' strictly 'for business,' the leader's need to preserve personal power may collude with the executioner's own psychopathology and callous disregard for life. Similarly, a leader's psychopathology could be linked to social pathology as in cases of genocide, when atavistic fears or deep resentments and anger against a group are brought to the surface in a cauldron of hatred fed by the leader's own hatred and psychopathological features. Ordinarily the leader uses charismatic power to manipulate the masses which allows for the emergence of primitive impulses of revenge, violence, and destruction. In turn, this frenzy proves the leader right and consolidates power. Political gain and revenge are the usual motivations of terrorism that depend on the

ability of the terrorist to objectify the victim, in itself a psychopathological construct.

Any other type of crime involves usual motivations such as greed, envy, lust, power, anger, overpossessiveness, and sometimes, even love. In any crime, however, the criminal, the victim, and society are inextricably tied in a web of emotional and social interactions mediated by the criminal act. Social sciences such as history, criminology, sociology, law, anthropology, psychology, and psychiatry converge on an act to label it criminal, and on the criminal to study the motivations.

The sciences diverge from that point on. For history, sociology, anthropology, and criminology look for answers in wider systems of social, political, and legal domains whereas psychiatry, law, psychology, and applied criminology are pragmatic sciences that explicitly objectify the criminal in order to justify interventions to change criminal behavior. In the elaboration of a psychopathological historiography, psychopathology seeks to understand the criminal act as a chain of personal events that have preceded the act and that will repeat in the future, if no intervention is undertaken. Sometimes a chain of similar acts has already been committed. In this 'package concept' of the criminal and crime, the instant event is only one expression, or symptom, of subjacent maladaptive processes that permeate all aspects of the personality. Maladaptive processes such as impulsivity, or callousness, are 'transnosographic' constructions (Widlöcher 1997) that span the spectrum of the personality. Constructs of this nature are rooted in character structure and may gain expression through criminal behavior or through symptoms of more obvious pathology such as depression or schizophrenia.

There are two types of psychopathological populations. First, the patients whose psychopathology is easily diagnosable as a mental condition such as anxiety, affective disorders, schizophrenia, or dementia. Patients are usually cared for in psychiatric units or in mental hospitals by general psychiatry. Second, criminals whose psychopathology may not be so obvious. They constitute the subject matter of penitentiary psychiatry in charge of studying the behavior, no matter how aberrant. Criminals are usually affected by deep personality disturbances that lead them repeatedly to crime such as thieving, assaultiveness, extreme forms of violence, substance dependencies, or deviated sexual practices. These two psychopathological populations are not usually so neatly separated because many persons suffer from co-morbid or dual conditions, or because there is an overlap whereby mental patients, like anybody else, could commit crimes and criminals, like anybody else, could become mentally ill.

Psychopathological explanations of crime have to be placed together with the sociological explanations that give meaning to the circumstances, historical and social, that surround the actor, or to the immediate triggers of the criminal act. Not to take these factors into account would do no justice to the criminal and it would deny an essential foundation to diagnostic interpretations and to treatment recommendations that might follow. A crime does not exist without a context and no criminal can be excised from the environment in which the crime took place.

2.3 Applications

A psychopathological evaluation of the criminal and the meaning of crime is an essential component, if not the *raison d'être* of forensic psychiatry. Forensic psychiatrists evaluate the mental condition of an accused for purposes of determining competence as required for legal purposes. To the contrary, treatment is not a regular concern of forensic psychiatrists. In correctional and penitentiary psychiatry evaluations do not dwell extensively on diagnostic issues, but rather, on the motivations for the crime to determine best treatment alternatives. Correctional treatment has two prongs: clinical treatment for symptoms of obvious psychopathology such as anxiety or depression, and correctional treatment to prevent or decrease recidivism usually through psychological or criminological methods. Unfortunately, sometimes clinicians in prisons are driven by a belief on rehabilitation and provide treatment without properly evaluating outcomes post-release over time.

3. Parameters of the Relationship

Psychopathological expressions could be subtle and difficult to elucidate, or other persons could be enrolled to demonstrate the pathology as in *follie à deux*, or in families where one member becomes the identified patient expressing the pathology of the other members. Psychopathology among criminals range from serious personality disorders to outright symptoms of mental illness. In between, there is a gamut of symptoms and crimes where the boundaries between criminology and psychiatry become blurred. Five levels of interactions could be described.

3.1 Crime as Symptom

These are cases in which a one-to-one relationship exists in the sense that the expression of the symptom constitutes the crime. For example, diagnosing some sexual pathologies such as pedophilia or exhibitionism is to acknowledge that a crime has been committed. A frotteur, for example, will not get sexual satisfaction by masturbating or by rubbing himself against an inanimate object; the compulsion is to do it against another person, yet anonymously, hence the crime happens in crowded places such as public transports.

The symptom is a crime against public morals and an assault for undue touching. A person who presents with fantasies about children, or about compulsively exhibiting in public, has not yet expressed the symptom and has committed no crime. To enter a diagnosis of pedophilia or exhibitionism at this stage would not be correct and could cause the person innumerable social and legal complications if the diagnosis is found out. In many crimes of this group the motivation is expressed as a desire, even compulsion, to indulge on the symptom and clearly precedes and determines the crime; thus, when psychopathology is present, the symptom is the crime.

3.2 Crime as Possible Symptom

One-to-one interactions could be observed in some other crimes. For example, firesetting is indicative of psychopathology, usually of an obsessive nature, and is found with some frequency associated with mental retardation, sexual pathology, or psychotic conditions. A diagnosis of firesetting could not be entered without immediately thinking that a crime, arson, has been committed. However, arson could also be a political or economical crime. Some shoplifters suffer from depression and anxiety, but shoplifting could simply be the result of a survival need, or even a business.

3.3 Crime as Correlate to Symptom

The third level of interaction presupposes a more distant relationship and takes place when a crime could be committed in order to indulge the symptom. An addict, for example, cannot do without the drug, but for as long as the drug is obtained legally because funds are readily available or because goods such as sexual favors could be bartered, the addict is free of crime. An addict commits a crime, either against property or sometimes against the person, when the drug cannot be purchased due to lack of funds or if there is nothing to barter. Sometimes, a child is sold or bartered as sexual pawn, or the addict agrees to a major violent crime in order to secure the drug. In these situations, the relationship between psychopathology and criminality is established not by the symptom itself, craving the drug, but by the drug-obtaining behavior at any cost. A similar situation could be envisaged among persons with developmental disabilities who display impulsive violent or destructive behavior. Usually the intervening factor is low frustration tolerance to irritating or annoying situations, or a poor grasp of heterosocial interactions leading to a sexual crime when willing partners, usually with similar disabilities, cannot be found.

3.4 Crime as Result of Symptom

This level involves a relationship where the crime is not the symptom nor is the crime related to the symptom through a criminal need in order to indulge the symptom, but it is the end result of several paths found in the course of some mental conditions. These include the following.

3.4.1 Crimes that follow the dictates of a symptom. Symptoms that could lead to crimes are only part of a multitude of other symptoms that accompany a particular mental condition such as paranoid states, schizophrenia, or major affective disorders. The majority of these patients may never be involved in any offences so crimes are not part of the definition of the condition. Furthermore, if the symptom is present, the patient could learn to control it by recognizing it as part of the illness, or through other forms of treatment such as medications. However, patients in this group could commit a crime when they follow the dictates of a symptom as happens in cases in which psychotic symptoms lead a person to believe in a certain state of affairs and, eventually, dictate that the person take action so as to modify a perceived negative course of the events.

3.4.2 Unexpected, unmotivated crimes. Sometimes a crime results from impulsive acts that are committed unpredictably and without time to foresee the consequences. This could be observed in persons suffering from intermittent explosive disorders, in the unreasonable, and unpredictable crimes of some patients suffering from schizophrenia, or in homicides, sometimes followed by suicide, of seriously depressed persons.

3.4.3 Crimes of interaction with social conditions. Finally, crimes could be part of the general social expression of a mental illness when social factors such as homelessness, deprivation, or poverty lead chronic mental patients to commit minor violations, usually motivated by survival needs punishable by a few days in a prison, or when a patient reacts violently to abuse or victimization. In this latter case, the crime may be caused not by the mental condition, but by a completely normal motivation of personal protection, or the mental condition may have affected the modulation of affect and aggravated the extent of the response.

3.4.4 Crime as an expression of being. Although characterizing crime as fate may seem pessimistic, criminal tendencies are so intimately linked to personality structures in some individuals that nothing

else could be expected, but a life career of crime. Being-in-the-world for these persons constitutes an extravagant display of predatory behavior in order to satisfy needs and sometimes perversions that are egosyntonic and whose satisfaction, which in the extremes could involve the humiliation, mutilation, and death of another, produces no guilt or remorse, but an urge to further predations. Many of these individuals present a range of behaviors, from the mildly antisocial to the extremely psychopathic and sadistic. Serial and sadistic murderers are usually found in this category.

4. *Epidemiological Evidence*

While a relationship between mental illness and criminality could be ascertained through the use of epidemiological methodologies, a relationship between psychopathology and crime can only be ascertained through the painstaking study of the individual criminal. Epidemiological studies cannot answer questions about fine personality traits or about motivation, which are usually buried in the very abyss of the human soul.

Epidemiological studies, however, can give a picture of the limit of psychopathology, as measured by the prevalence of criminality among psychiatric patients, or of mental illness among criminals, whether in or out of institutions (Arboleda-Flórez 1998). Analytical epidemiological methods have been used to determine causal factors between mental illness and violence, but much more refinements have to be introduced in the methodology of these studies before a statement of causality could be made (Arboleda-Flórez et al. 1998).

5. *Biological Determinants of Criminality*

Criminal behavior can be explained on sociological or anthropological terms or as responding to biological determinants. Criminal behaviors with a psychopathological substratum require a biological layer of explanation through biological mechanisms that involve personality aspects, genetic and hereditary factors, physiological alterations, cerebral damage caused by perinatal insults or head trauma, or acute intoxication.

5.1 *Personality*

Some physiological correlates of personality are also related to criminality, especially by individuals with antisocial and psychopathic personalities. Higher thresholds for excitability and higher impulsivity accompanied by lower levels of skin conductance, pulse, and electroencephalographic markers for excitability than found among normal individuals are found among psychopaths which could explain their thrill-seeking behaviors (Hare 1982). Similarly, low pulse has been found to be related to regulatory patterns of inhibitory conduct among adolescents, juvenile delinquents, and those diagnosed with conduct disorders.

5.2 *Cerebral Damage*

Neuropsychological tests could demonstrate cerebral damage, but this is best done with imaging. MRI, PET scans, and SPECT examinations have demonstrated frontal lobe damage among extremely violent individuals, serial murderers, and non-psychotic personality disordered persons (Goyer and Semple 1996). Cerebral pathology is related to intermittent explosive disorders, and frontal lobe insults have been associated with levels of impulsivity (Barratt 1987) which in turn have been related to levels of serotonin (Brown et al. 1989).

5.3 *Genetics and Heredity*

A familiar component has been described relating antisocial behavior, criminality, and violence, which in turn are related to paternal violence, poverty, single parent families, and rough neighborhoods. These interfamily variation factors, as known in genetic epidemiology, change from family to family but remain constant as a load in one single family. It is not possible, yet, to differentiate within members of a family the quantities that could be attributed to the genetic load (genotype) from that attributed to the environment and that result on a particular form of behavior (phenotype). Link and association studies demonstrate that some genetic disorders such as alcoholism, Gilles de la Tourette syndrome, and the fragile-X syndrome could be related to antisocial behavior and violence (Carey 1994). Furthermore, adoption studies of twins indicate that there exists a genetic relation between antisocial personality and alcoholism (Cadoret et al. 1986).

5.4 *Intoxicating Substances*

Toxic substances are common correlates of criminal behavior either ingested voluntarily or through exposure in the environment. The impact of alcohol on infants *in utero* resulting in fetal alcohol syndrome or intellectual damage suffered by normally born babies exposed to high levels of lead in the environment are well documented. Developmental delays and injuries to the developing brain are risk factors for criminality. Alcohol has a well-known impact on the frontal lobes (Baron and Richardson 1994) which are in charge of inhibitory functions, and on the limbic structures in charge of vegetative and instinctive functions. The effects of alcohol and drugs are well known in forensic psychiatry because of violent crimes including sexual attacks committed while intoxicated and the emerg-

ence of dissociative and automatic states of mind (Arboleda-Flórez 1999). Figures bear this relationship, 42 percent of arrested individuals are inebriated and 29 percent are under the effects of drugs at the time of arrest. Furthermore, 57 percent had used drugs within one month prior to their arrest, and 83 percent of inmates in state prisons and 73 percent of those in federal prisons had utilized drugs at some time in their lifetimes (Mumola 1998).

5.5 Biological Reductionism

Some authors have expressed a concern that biological explanations could lead to a fruitless search for easy solutions to the complexities of crime. They point out that such way of thinking could distort the truth because of the weight paid to biological facts, could raise false hopes for easy solutions, and could engender fears, prejudice, and apprehensions in the population worse than the actual dangers represented by crime (Brain 1984). Myths surrounding biological investigations of crime have retarded the development of this profitable area of research (Stoff and Cairns 1996).

6. Social Determinants of Criminality

Social explanatory models of criminality demonstrate that there is a relationship between social situations and crime. Anthropological and social definitions of crime at a larger scale indicate that some forms of behavior are acceptable in some cultures or localities, but not in others. For example, corporal punishment of children and spousal abuse may be considered appropriate in some subcultures or countries, but frowned upon and punishable as serious crimes in others. Cultures of crime exist in rough neighborhoods and social definitions of acceptable behavior follow local definitions of morality. These broad determinants are independent of other generally acceptable correlates of criminality such as labeling, poverty, deprivation, exposure to violence at home or in society in general, breakdown of social lines, or the social fabric at times of political, economical, or wartime unrest, and availability of guns. Unfortunately, none of these models have wholesome parsimonious value although most can explain crime under some circumstances.

7. Conclusions

Criminality responds to multideterminant and multidimensional processes that demand a broad understanding and knowledgeable approach to its study, away from reductionistic beliefs, be they biological or social, and away from solipsistic solutions that would deny possibilities for study and conflict resolution. Learning about the psychopathological, biological, and social correlates of crime opens the door to instituting remediation strategies aimed at rehabilitation and successful reintegration of the criminal into society.

Bibliography

Arboleda-Flórez J 1998 Mental illness and violence: An epidemiological appraisal of the evidence. *Canadian Journal of Psychiatry* **43**: 989–96

Arboleda-Flórez J, Holley H, Crisanti A 1998 Understanding causal paths between mental illness and violence. Special Supplement of *International Journal of Social Psychiatry and Psychiatric Epidemiology* **33** Suppl. 1: 38–46

Arboleda-Flórez J 1999 *Forensic Psychiatric Evidence*. Butterworth, Toronto, ON

Aristotle 1941 The Nicomachean ethics. In: McKeon R (ed.) *The Basic Works of Aristotle*, Random House, New York

Baron R A, Richardson D R 1994 *Human Aggression*. Plenum Press, New York, Chap. 8, pp. 277–308

Barratt E 1987 Impulsiveness and anxiety: Information processing and electroencephalograph topography. *Journal of Research in Personality* **21**: 453–63

Brain P F 1984 Biological explanations of human aggression and the resulting therapies offered by such approaches: A critical evaluation. In: Blanchard R J, Blanchard D C (eds.) *Advances in the Study of Aggression*. Academic Press, Orlando, FL, Vol 1, pp. 63–102

Brown C, Kent T, Bryant S, Gevedon R, Campbell J, Felhous A, Barratt E, Rose R 1989 Blood platelet uptake of serotonin in episodic aggression. *Psychiatry Research* **27**: 5–12

Cadoret R J, Troughton E, Gorman T W, Heywood E 1986 An adoption study of genetic and environmental factors in drug abuse. *Archives of General Psychiatry* **43**: 1131–6

Carey G 1994 The genetic association study in psychiatry: Analytical evaluation and a recommendation. *American Journal of Medical Genetics* **54**: 311–17

Debuyst C 1997 Preface. In: Albernhe T (ed.) *Criminology et Psychiatrie* [English translation: Criminology and Psychiatry]. Ellipses, Paris, pp. 9–11

Goyer P F, Semple W E 1996 PET studies of aggression in personality disorders and other nonpsychotic patientes. In: Stoff D M, Cairns R B (eds.) *Aggression and Violence*. Erlbaum, Mahwah, NJ, Chap. 11

Hare R D 1982 Psychopathy and physiological activity during anticipation of an aversive stimulus in a distraction paradigm. *Psychophysiology* **19**(3): 266–71

Mumola C J 1998 *Substance Abuse and Treatment, State and Federal Prisoners, 1997*. Bureau of Justice Statistics, US Department of Justice, Washington, DC

Stoff D M, Cairns R B 1996 *Aggression and Violence*. Erlbaum, Mahwah, NJ, pp. 13–15

Widlöcher D 1997 Le fait criminel est-il un fait psychopathologique? [English translation: Is the criminal act a psychopathological act?]. In: Albernhe T (ed.) *Criminology et Psychiatrie* [English translation: *Criminology and Psychiatry*]. Ellipses, Paris, p. 29

J. E. Arboleda-Flórez

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Critical Legal Studies

'Critical legal studies' (CLS) is an umbrella term for a variety of movements of legal intellectuals, most of them in law faculties in the USA, Britain, and France. The movements originated in the late 1970s and achieved their peak of influence in the 1980s and 1990s. Their adherents are generally on the political left, critical of existing distributions of wealth, power and status rewards and of patterns of domination and subordination by class, race, gender, and other 'illegitimate hierarchies.' Unlike classical Marxists, they do not treat law as a merely 'superstructural' reflection of a material ' base' of determining economic relations, or as simply an instrument of ruling-class power (see *Marxism and Law*), ; they take the content and forms of law and legal reasoning seriously, as relatively autonomous active force fields in the construction and interpretation of the social world. Critics (or 'crits') target their critiques to the dominant modes of legal reasoning, scholarship, and teaching in their legal cultures, which usually pretend to be neutral and apolitical, but which the critics see as expressions of conflicting political ideologies. The ideologies hidden in law often serve to produce, reproduce, and justify inequalities and illegitimate hierarchies, but they also express utopian aspirations and point to plausible alternative social arrangements. The critics' general aims have been to bring into view the hidden ideological biases and effects of the existing legal order, to capture and re-orient enclaves of legal scholarship, teaching, and practice, and to use them to reimagine and to implement alternative legal ideas and policies.

Within these broad commonalities of outlook and concern, the CLS movements have tended to fragment into many different schools, informed by diverse intellectual antecedents and committed to divergent views of the role of law in social life and of the political ambitions of intellectual critique and reform.

The US movements receive most attention here, since they attracted the most members, produced the largest literature, and penetrated the furthest into centers of legal education; also, their types of scholarship and internal quarrels parallel those of CLS movements elsewhere. CLS-USA came into being in 1977 at a conference of law teachers rebelling against their own legal education and its seeming irrelevance to the contemporary ferment of social movements (anti-Vietnam War, civil rights, feminist, environmental, etc.). By the 1980s critical teachers made up a significant fraction (four or more faculty) at Buffalo, Harvard, Miami, Northeastern, Stanford, UCLA, and Wisconsin law schools and a token presence at many more, eventually producing students who went into law teaching themselves, a substantial bibliography (Kennedy and Klare 1984, Bauman 1996), and a series of national conferences.

1. *Antecedents of CLS*

To critique the orthodoxies of their legal culture, American critics in the first phase of their movement found resources in various European theories, especially Western neo-Marxist theories of law as 'ideology' and an instrument of 'hegemony' (Gramsci 1971), humanist social histories of subordinated groups (Hay 1975, Thompson 1975), structuralism, and phenomenology. But their main resource turned out to be their domestic forerunners, the American legal realists of the 1920s and 1930s who critiqued the classical–formalist legal orthodoxy of their own day (see *Legal Realism*; *Legal Formalism*). That orthodoxy had claimed that the operating rules of the legal system could be derived from a few basic principles induced from a scientific arrangement of precedents, such as the protection of property, the basing of liability to others on 'fault,' and the enforcement of the 'free will' of parties to contracts. This system added up to a natural and neutral mode of ordering private life, and legislative interferences with this system of rules would be struck down as unconstitutional unless they met strict tests of validity.

The realists' critique, in brief, was that (a) the principles were too vague, and the precedents too flexible, to determine outcomes, and in actual operation produced results varying over time and with contexts; (b) the actual determinants of results were latent and unarticulated policies, ideologies, and social vision; and (c) the 'private' law rules of property-tort-contract that undergirded the natural-seeming private sphere of free choice in free markets, protected against interference by constitutional law, were no less artificial, no less exercises of state policy, and no less coercive, than the 'public' rules of legislation and administration. A property right, for example, delegated to its owner the ultimate power to use state force (self-help, private guards, the police or the army) to exclude others (such as union members or black people) he did not want to hire or deal with, and to condition access to his property (tools, land, wages, the assets of a marriage) on those others' submitting to his orders. The coercive state was always involved in the market and business firms and family life; the issues for legal policy were thus not *whether* the state should be involved, but *how*, and *to what ends*—issues that courts, as well as legislatures and administrative agencies, could not avoid and had to resolve.

By the1950s and 1960s, however, the leading currents of American legal thought had settled into a new orthodoxy, one that absorbed elements of the realist critique but blunted its more radical implications (see *Legal Scholarship*). Some postwar schools held that law (including law made by judges), although not autonomous from clashing social interests, had neutral craft techniques for balancing those interests; others that law properly interpreted served social purposes on which there was presumed to be a general con-

sensus; and others that courts differed from legislatures in that judge-made law promoted general 'principles' rather than specific 'policies.' In the late 1960s, emerging schools of law-and-economics sought to rationalize legal doctrine, especially the private law of property, tort, and contract, as serving the aims of efficiency or wealth maximization. (For an overview see Duxbury 1995; see *Law and Economics*.)

2. *Early CLS*

CLS began as a set of critiques of these postwar modes of legal reasoning. Unlike other movements that carried on the legal–realist critiques, such as the law-and-society movement (which investigated empirically the social influences that helped to form, and the actual social effects of, legal rules), the critics deployed an internal critique: attacking the prevalent modes of doctrinal and policy argument practiced by law teachers, judges, and other officials on their own terms, aiming to develop counterarguments to teach to law students and practicing lawyers.

Following neo-Marxist writers and the critical social theorists of the Frankfurt School, the critics pictured legal doctrines and arguments ('legal discourses') as ideology, a partial vision of the world that presents itself as a universal vision. These discourses function so as to portray existing constellations of rights, powers, privileges, and immunities recognized by the legal system as if they were natural and inevitable ('false necessity'), and as close to being good as they could be ('false legitimation'). By such means the law 'reifies' and thus helps to freeze in place scripted social roles such as manager and worker, landlord and tenant, husband and wife, consumer debtor and bank creditor, property owner and picketing laborer, that blind the ordinary people who play those roles to alternative possibilities for action and organization (see e.g., Gabel 1980).

Like the legal realists, the critics' method was to unearth and make transparent the latent subtexts of legal doctrines and policies, the hidden framing conceptions, empirical assumptions, and social visions on which conventional legal reasoning, and the scripted social roles it gave people to play, relied. But the aim was not simply to unmask; it was to develop alternative doctrinal and policy arguments in particular fields of law, and by so doing to open a path for a revisionary politics based on different views of social reality and human potential.

In its first decade, CLS developed a distinctive model of such critiques, which, taught to and adapted by adherents and disciples, was replicated in about 100 articles covering a wide variety of legal fields. The model critique identified some deep-level tensions, polarities, or contradictions underlying a field of legal policy or argument. In a pioneering article, Kennedy argued that contract law doctrines tended to cluster around opposing conceptions of human nature, social obligation, and the functions of law: one favoring formal rules and 'individualist' views of economic transactions and social duty, the other favoring informal standards and 'altruist' ethics of sharing and cooperation (Kennedy 1976). The law in force at any given time oscillates between the opposing poles as it moves from one subject to another, but tends to privilege one pole and suppress the other, reserving it for exceptional or marginal situations.

Doctrines from the suppressed pole, however, are always available as a 'dangerous supplement' which at any moment may, and sometimes do take over the privileged position (Kelman 1987). One pioneer (Unger 1975) generalized the critique across legal fields to 'liberalism' as a whole; but most practiced it as a form of local critique, whose content varied with the field of law that was being criticized. The model also spawned its own genre of structuralist legal history, showing how legal fields come to be structured around contradictory organizing principles and social visions. For example, doctrines reflect people's need for other people and fear of them; the desire for a state powerful enough to protect rights that will not also trample them, and for contract and property law that will carry out parties' intentions and underwrite their freedom of action but also protect their (and others') security. The legal system tries to 'mediate' the contradictory doctrines by confining each to a limited field of operation, yet such systems become unstable as the mediating devices become incoherent and exceptions spill over limits to swamp the rules (see e.g., Alexander 1987, Kennedy 1979, Mensch 1982).

This strand of CLS was often misunderstood. Critics argued that the structuring of legal fields around such 'contradictions' meant that law was 'indeterminate' because it could (and historically often did) enable judges, advocates, or reformers to make valid arguments (according to internal criteria of the legal system) for opposed positions. Doctrinal or policy arguments could always be 'flipped' by appealing to principles at the opposite pole. This did not mean that law was a facade, sham, or arbitrary meaningless jumble, or that doctrine was just a mask for raw political choice; rather, its indeterminacy was patterned and systemic, an effect of law's (or liberalism's) contradictory structures. Nor did it mean that law was unpredictable, because arguments drawn from one pole or another tended to be stable in the medium term. Nor, finally, was it an indictment of a legal system to say that it was 'indeterminate.' It was merely a fact, one papered over by mediating devices leading to the appearance of natural, necessary, and almost-as-good-as-it-could-be law.

3. *Later CLS*

CLS also extended the method beyond private law doctrine to the powerful emerging new forms of policy

analysis, legal economics, attacking certain core premises of the new science including the Coase theorem, cost-benefit analysis of legal entitlements, public choice theory, and submission to the will of others as rational consent (see e.g., Kennedy and Michelman 1980). Others extended it to institutions: Simon identified opposed models of welfare administration, mechanical rule-application by low-skilled workers and the professional informal discretion of caseworkers. He showed how the welfare system had evolved from discretion to rules with damaging results for clients and argued for restoration of discretion (Simon 1983).

Critics also revived and extended the legal realists' critique of the distinction between 'public' (coercive regulatory) and 'private' (freely–willed contractual consensual market) relations. Following the realist R. L. Hale (see Fried 1998 for an intellectual biography), they argued that all 'free choice' is made by actors with property and power endowments and under baseline conditions structured by the legal system, even choice in the most intimate spheres of life such as marriage relations. Their aim was to supplement Foucault's (1977) account of the pervasive regulation of social interactions through 'disciplines' with an account of the pervasive structuration of transactions and interactions in business and personal life through law.

Much critical effort went into history (see Gordon 1984). Some critical historians worked in the tradition of the social history of subordinated groups, using historical examples to show how ruling groups had used neutral- and equal-sounding legal doctrines to extend and justify their rule. Horwitz (1977) showed how early-nineteenth-century law had reworked private law rules of property, tort, and contract to favor entrepreneurs in the early nineteenth century, then later froze the new rules against redistributive revision by imposing constitutional restraints on legislation. Klare (1978) showed how conservative judges had drained the National Labor Relations (Wagner) Act of the New Deal of its potential as a charter of workers' rights to participate in management. Critical historians often read the past as a repository of lost social experiments, long since crushed or forgotten, which might serve as models for reconstructing aspects of social life—as did Frug with the nineteenth-century model of powerful autonomous city governments (Frug 1980), and Simon did with 'social–republican' property models of landholding and producers' cooperatives (Simon 1991). Other histories explored the legal structuration of social relations, especially those of master and slave, employer and employee, husband and wife (see e.g., Steinfeld 1991, Olsen 1983).

By the mid-1980s CLS had entered a new phase. One group of critics, perhaps responding to accusations that CLS was 'nihilist' in its attachment to destructive criticism ('trashing'), turned to concrete programmatic work. Unger's (1998) was once again the most ambitious, proposing large-scale programs for the democratization of the economies of both developed and developing societies; most others were small-scale proposals for reforming specific legal policy fields, such as worker participation in ownership and management (Stone and van Wezel 1988), treatment of workers in plant closings (Singer 1988), regulation of injurious speech (Abel 1994), intellectual property (Boyle 1996), disability law (Kelman and Lester 1997), and property rights in ex-Communist societies (Alexander and Skapska 1994).

Another group focused on developing critical theories of law to a new level of sophistication, integrating them with postmodernist and poststructuralist theories, especially those of Foucault and Derrida (see e.g., Balkin 1987, Heller 1984, Peller 1985).

4. *Challenges and Erosion*

Around the same time, CLS experienced intensifying challenges from neighboring left legal–intellectual movements, especially feminism and critical race theory. The new movements asserted differences from CLS in focus and method—differences that provoked vigorous arguments throughout the 1990s. CLS had criticized liberal legalism for its claims to have solved the problem of domination by constructing 'rights.' Some critics said rights increased the alienation of individuals by constituting them as monads isolated from one another (Gabel 1984). Others argued that rights mystified the realities of power relations, while still others argued that social movements' crusades for rights (such as the black civil rights movement) often ended up giving them only symbolic victories and unenforceable claims still others contended that rights, like other legal doctrines and policies, were always in conflict—the right to equal treatment in accommodations, housing, and employment, for example, with the right to associate—and that the conflicts could not be resolved by appeals to rights (Tushnet 1984). Race theorists (e.g., Crenshaw 1988) counterargued that whatever might be the weaknesses of rights-rhetoric, claims using the language of rights had been essential to the gains of subordinated groups. CLS scholars had also stressed the indeterminacy, manipulability, and contingency of social arrangements and legal discourses, the fluidity of meanings and identities. Some feminist and critical-race theorists, by contrast, wished to insist at least provisionally on the grounded objectivity of experiences of subordination, the common shared elements that made up sexual or racial identity and the authenticity of personal narratives of the lives of the subordinated (see *Critical Race Theory*; *Feminist Legal Theory*).

CLS had helped to open up and to institutionalize space—in the form of places for faculty positions, room for articles in law reviews, slots on the programs of conferences—for a legal–academic left, a novel

phenomenon in the USA (and wholly missing in other professional faculties such as business schools). By the end of the 1990s, however, most of that space, like most of the energy going into novel work, was occupied by gender and race studies. CLS ideas continued diffusely to pervade both mainstream and critical fields of scholarship and policy analysis—notably its critiques of the public–private distinction and of the legal foundations of neo-liberal political economy. But as a distinct school CLS had faded away.

5. *European CLS*

European critical schools of law followed similar paths, with important differences due to national variations in the strength and programs of left-wing political parties and intelligentsia, and the objects of critique—traditional modes of legal training and reasoning—on the other. (A European Conference on Critical Legal Studies was organized in 1981.) In Britain, critical legal studies were first concentrated in social science faculties and grew out of the radical sociology of law, oriented in its early years to the critique of the 'State' and its legal controls of poor and working people through labor, policing, criminal law, and professionalized social services, later evolving into broader critiques of law as ideology (see e.g., Cain and Hunt 1979). In the 1980s CLS was taken up by law teachers, especially in the polytechnics. They organized the Critical Legal Conference in 1984, proclaiming 'a shared rejection of the dominant tradition of Anglo-American legal scholarship, the expository orthodoxy or … "black-letter law" tradition' (Fitzpatrick and Hunt 1987, p. 1). Believing that their American counterparts focused too narrowly on the internal critique of legal doctrine and reasoning, British critics maintained their connections to socio-legal studies and to radical law practices; they also developed new alliances with feminist, deconstructionist, psychoanalytic, and literary approaches. Critical legal studies in the UK, as in the USA, are a very fragmented enterprise, but unlike its American counterpart, it is flourishing (for exemplary collections and overviews see Fitzpatrick and Hunt 1987, Douzinas et al. 1994).

In France the combined influence of (a) Marxist traditions treating law as 'merely superstructural' and (b) separation of faculties of legal philosophy and theory from faculties of law, held back the development of left-wing legal studies until the late 1960s, when radical social movements and the surge of interest in the social theories of Althusser, Gramsci, and Foucault led to critiques of law as forms of ideology that structure and legitimate the functioning of capitalist order. After the 1980s the movement shifted emphasis from general theory to concrete case studies of the norms, reasoning modes, and institutions that went into the construction of everyday legal relations. French CLS has not deeply penetrated the centers of university training of the legal elite, but has gained a foothold in some faculties and in research centers at St. Etienne, Lyon, and Montpellier (see Arnaud 1993 for a concise summary).

See also: Critical Race Theory; Critical Theory: Contemporary; Critical Theory: Frankfurt School; Feminist Legal Theory; Hermeneutics, Including Critical Theory; Law and Economics; Law and Economics: Empirical Dimensions; Law and Society: Socio-legal Studies; Law as Constitutive; Legal Formalism; Legal Realism; Legal Scholarship; Marxism and Law

Bibliography

Abel R 1989 *American Lawyers*. Oxford University Press, New York

Abel R 1994 *Speech and Respect*. Sweet & Maxwell, London

Alexander G 1987 The transformation of trusts as a legal category. *Law & History Review* **5**: 303

Alexander G, Skapska G 1994 *A Fourth Way? Privatization, Property and the Emergence of the New Market Economies*. Routledge, New York

Arnaud A-J 1993 Critique du Droit. In: Arnaud A-J (ed.) *Dictionnaire Encyclopedique de Théorie et de Sociologie de Droit*, 2nd edn. Libre générale de droit et de jurisprudence, Paris

Balkin J 1987 Deconstructive practice and legal theory. *Yale Law Journal* **96**: 743

Bauman R W 1996 *Critical Legal Studies: A Guide to the Literature*. Westview Press, Boulder, CO

Boyle J 1996 *Shamans, Software and Spleens*. Harvard University Press, Cambridge, MA

Cain M, Hunt A (eds.) 1979 *Marx and Engels on Law*. Academic Press, London

Crenshaw K 1988 Race, reform and retrenchment. *Harvard Law Review* **101**: 1331

Douzinas C, Goodrich P, Hatchamovitch Y (eds.) 1994 *Politics, Postmodernity, and Critical Legal Studies*. Routledge, London

Duxbury N 1995 *Patterns of American Jurisprudence*. Oxford University Press, New York

Fitzpatrick P, Hunt A (eds.) 1987 *Critical Legal Studies*. Blackwell, Oxford, UK

Foucault M 1977 *Discipline and Punish: The Birth of the Prison*. Pantheon, New York

Fried B H 1998 *The Progressive Assault on Laissez-Faire: Robert Hale and the First Law and Economics Movement*. Harvard University Press, Cambridge, MA

Frug G 1980 The city as a legal concept. *Harvard Law Review* **93**: 1057

Gabel P 1980 Reification in legal reasoning. *Research in Law and Sociology* **3**: 25

Gabel P 1984 The phenomenology of rights-consciousness and the pact of the withdrawn selves. *Texas Law Review* **62**: 1563

Gordon R W 1984 Critical legal histories. *Stanford Law Review* **36**: 57

Gramsci A 1971 *Selections from the Prison Notebooks*. In: Hoare Q, Nowell-Smith G (ed. and trans.). International Publishers, New York

Hay D 1975 Property, authority and the criminal law. In: Hay D, Linebaugh P, Rule J G, Thompson E P, Winslow C (eds.) *Albion's Fatal Tree*. Pantheon, New York

Heller T C 1984 Structuralism and critique. *Stanford Law Review* **36**: 127
Horwitz M J 1977 *The Transformation of American Law, 1780–1860*. Harvard University Press, Cambridge, MA
Kelman M 1987 *A Guide to Critical Legal Studies*. Harvard University Press, Cambridge, MA
Kelman M, Lester G 1997 *Jumping the Queue*. Harvard University Press, Cambridge, MA
Kennedy D 1976 Form and substance in private law adjudication. *Harvard Law Review* **89**: 1685
Kennedy D 1979 The structure of Blackstone's commentaries. *Buffalo Law Review* **28**: 205
Kennedy D, Klare K 1984 A bibiliography of critical legal studies. *Yale Law Journal* **94**: 461
Kennedy D, Michelman F 1980 Are property and contract efficient? *Hofstra Law Review* **8**: 711
Klare K 1978 The deradicalization of the Wagner Act and the origins of modern legal consciousness. *Minnesota Law Review* **62**: 265
Mensch E 1982 The colonial origins of liberal property rights. *Buffalo Law Review* **31**: 635
Olsen F 1983 The family and the market: a study of ideology and legal reform. *Harvard Law Review* **96**: 1497
Peller G 1985 The metaphysics of American law. *California Law Review* **73**: 1151
Simon W H 1983 Legality, bureaucracy and class in welfare administration. *Yale Law Journal* **93**: 1198
Simon W H 1991 Social republican property. *UCLA Law Review* **38**: 1335
Simon W H 1992 The prospects of pension-fund socialism. *Berkeley Journal and Employment & Labor Law* **14**: 251
Singer J W 1988 The reliance interest in property. *Stanford Law Review* **40**: 611
Steinfeld R 1991 *The Invention of Free Labor*. University of North Carolina Press, Chapel Hill, NC
Stone K, van Wezel 1988 Labor and the corporate structure. *University of Chicago Law Review* **55**: 73–173
Thompson E P 1971 *Whigs and Hunters*. Pantheon, New York
Tushnet M 1984 An essay on rights. *Texas Law Review* **62**: 1363
Unger R M 1975 *Knowledge and Politics*. Free Press, New York
Unger R M 1998 *Democracy Realized*. Verso, New York

R. W. Gordon

Critical Psychology

The term 'critical psychology' has been applied to various approaches to psychology which, in the last decades of the twentieth century, typically share hardly anything more than a resistance to so-called mainstream or traditional psychology, an affinity to the moral ingredient of the Marxist critique of capitalism, and an aspiration to fight for the oppressed and the excluded.

Critical psychologies of whatever nuance are not connected to I. Kant's critical philosophy or to neo-Kantian approaches to psychology despite the fact that one of the authorities of critical psychology, S. L. Rubinshtein, started out as a student of neo-Kantianism in Marburg.

One division in the heterogenous stream of critical psychologies is named *Critical Psychology*. This is a school of psychology that chose to spell its name with capital letters; it originated in the 1970s in West Berlin, and is centered around the work of K. Holzkamp. It describes itself as being Marxist–Leninist, scientific (in a Marxist sense), and critical of what it calls bourgeois psychology. Critical Psychology will be the primary object of this article as it is so far the branch of critical psychology with the most noticeable degree of social coherence and of theoretical consistency.

1. Origins and Early History of Critical Psychology

The second half of the 1960s witnessed a rapid increase in the university student population in the Western world. The resulting tension and uneasiness combined with the televised atrocities of the Vietnam war to produce an international agitation for global revolutionary changes in a world engaged in Cold War.

An important factor in the student revolt, especially in Germany, and also in the formation of critical psychology and Critical Psychology, was the Critical Theory of the Frankfurt School, a strain of Marxism embodied by T. W. Adorno, M. Horkheimer, H. Marcuse, and J. Habermas. Especially pertinent at universities was their critique of positivism in sociology and in other social sciences, and generally of positivism in the sciences as prominently advocated by K. Popper. The positivist emphasis on the priority of observable events and empirical data over hypothetical constructs and theoretical terms was criticized as a deceptive disguise for sinister theory, especially for reactionary ideology in the service of capitalism. Therefore the key reproach against positivism was its presumed lack of scientific relevance and of relevance for social change.

The Frankfurt School was not interested in the academic discipline of psychology. But psychology students watched the positivism controversy and applied it to their own field. A student Action Committee Critical Psychology disturbed the customary serenity of the congress of the German Society for Psychology at Tübingen in 1968, and, in order to coordinate various local initiatives, students met for a singular and tumultuous Congress of Critical and Oppositional Psychology in Hanover in 1969.

The term 'critical psychology' appeared for the first time as a book title *kritische psychologie* in 1969. This is an anthology compiled by Bochum university psychology students comprising pirated texts by T. W. Adorno, W. Reich, K. Holzkamp and others, and various resolutions exchanged at the Hanover Congress and at earlier occasions. Sold in many psychological institutes, it established the term as a trademark for revolt, and stepped up rebellious behavior which few university teachers could stomach. The concept of

critical psychology covered a multitude of different convictions, from an engagement for the weak and the betterment of society, to the use of psychology in a Marxist class struggle, to the conviction that psychology is unredeemably reactionary and should be dumped on the dungheap of history. Views of the future of psychology varied accordingly.

Among the few university professors who engaged in extended dialog with critical psychology students was Klaus Holzkamp (1927–95), holder of a chair of psychology at West Berlin's Free University. He had, in 1964 and in 1968, published a critique of psychological experimentation wherein he denounced the lack of criteria for the evaluation of the relevance of psychological experiments for psychological theory. He tried to improve the diagnosed situation by following the constructivist philosophy of science of H. Dingler, the most prominent European proponent of operationalism.

In an invited speech at the Berlin section of the German association of professional psychologists in 1968, Holzkamp applied the concepts of technical and emancipatory relevance, which he had taken from J. Habermas, to psychology and declared emancipatory relevance superior to all other forms of relevance. Through his discussions with critical students Holzkamp was introduced to an analysis of society and its dominance relations according to Marxist lines. In his collection of essays, *Critical Psychology* (Holzkamp 1972), he called his change in outlook the *Critical-Emancipatory Turn of Psychology*. Activity at the West Berlin psychological institute was characterized by studies in the history of psychology which strove to uncover its quality as a tool of oppression in the ongoing class struggle (Geuter and Mattes 1984), and by practical engagement for the underprivileged. This approach to the practice of psychologists remained what Holzkamp later called 'voluntaristic,' that is, appealing to the good will of psychologists and a willful decision for partisanship for the weak and needy.

As the antiauthoritarian and nonconformist student movement slowly became organized by political parties and splinter groups in the early 1970s, so were the critical psychologists. At the psychological institute in West Berlin, one faction drifted into the sphere of influence of the Marxist–Leninist party (SED) ruling East Germany (DDR) and its West Berlin vassal party (SEW); another faction gathered around Beijing-oriented Maoist lines; others become Trotskyites, while a good many remained nonaligned. The psychologists around Holzkamp enlisted in the Marxist–Leninist camp. They withdrew from dialog with other factions, fashioned their own publication outlets and, to emphasize their distinctness, from the middle of the 1970s spelled their field as Critical Psychology. This group has also been known as the Holzkamp school, or the Berlin critical psychologists, despite other critical psychologists protesting that it was merely one offshoot of a much broader critical psychology unjustifiably trying to usurp the appellation and to eclipse the other approaches. The Marxist–Leninist turn has led Critical Psychology indubitably away from Adorno, Critical Theory, and the New Left.

2. *Creation of Critical Psychology*

Critical Psychology defines itself as a Marxist–Leninist psychology (Holzkamp 1983, p. 27). It looks toward the writings of K. Marx, F. Engels, and V. I. Lenin for orientation, and adopts as its scientific and epistemological foundation or superstructure dialectical materialism and historical materialism. It embraces as pioneering works in this kind of psychology the publications of the historicocultural approach devised during Stalin's reign by L. S. Vygotsky and his disciples, in particular S. L. Rubinshtein and A. N. Leont'ev. These authors were *de rigueur* for prefaces to be considered ideologically impeccable in the psychology of the DDR and the USSR, even though few serious scientists paid more than lip service to the ideological superstructure.

Holzkamp's turn away from critical-emancipatory psychology towards Critical Psychology became visible with his book on *Sensory Cognition* (1973) where K. Marx is the most often cited author, and *Das Kapital* provides the model of human consciousness as dependent on and determined by the material and historical foundation of economic and social practice. Traditional psychology, now named bourgeois psychology, however, is perceived as erroneously proclaiming the independence of consciousness from this material basis.

Critical Psychology had not only an ideological superstructure, but a rarely considered material basis as well, namely the chair of psychology at the Free University in West Berlin held by Holzkamp and the power and privileges that went with it. A German university professor is a civil servant, but at the same time enjoys a very high degree of academic freedom in research and in teaching, as guaranteed by the constitution. This freedom, however, may have one limitation: A professor of psychology is obliged to train and examine future psychologists according to the canon of the Examination Regulations.

Critical psychology's critique of traditional or bourgeois psychology had purportedly unveiled the latter's theoretical hollowness and its practical use as a weapon of the ruling classes against the proletariat. This brought Holzkamp into a 'dilemma' (1973, p. 13). He could either go on teaching and training in bourgeois psychology and thereby lose his credibility as a critical psychologist, or he could stop training psychologists altogether, limit himself to criticizing psychology and face the loss of his professorship and its benefits. He resolved this 'fatal alternative' (1973, p. 15) in a way that he thought allowed him to preserve

critical credibility and to keep the material basis as well. He devised Critical Psychology, a new brand of psychology professedly consistent with the results of the critique of traditional psychology and also suitable as a foundation for the training of future practical psychologists.

Holzkamp and his co-workers had to prove that Critical Psychology could deliver. They formed a secluded group and started out to present to the public the various areas of psychology mentioned in the Examination Regulations according to Critical Psychology.

3. *Organization of the School of Critical Psychology*

Critical Psychology organized itself as a school of psychology with defined teacher–pupil relations and clear boundaries between insiders and outsiders. Its center was the chair of psychology occupied by Holzkamp, who was surrounded by assistants, doctoral students, and students. The notable names are gathered in the reader *Critical Psychology* (Tolman and Maiers 1991). The West Berlin case remained unique. Critical Psychology could never gain a similar foothold in another university institute.

A series of textbooks, *Texte zur Kritischen Psychologie*, started to appear in 1973. It was supplanted in 1977 by another series, *Studien zur Kritischen Psychologie*, which appeared in the publishing house of Marxist–Leninist orthodoxy. The journal *Forum Kritische Psychologie* began to appear irregularly in 1975, and was turned into a biannual journal in 1977. A congress of Critical Psychology was held in the West German university town of Marburg in 1977; further congresses, now named 'International,' were held there in 1979 and 1984, and a fourth congress, not 'International,' in Berlin in 1997.

The 1980s and 1990s seem to have been a time of stagnation for Critical Psychology. But steps were undertaken to introduce an international public to Critical Psychology (Tolman and Maiers 1991, Tolman 1994). The latest developments may be followed at *www.critical-psychology.de*.

4. *Themes and Topics of Critical Psychology*

Work done in Critical Psychology is to an uncommon degree conceptual work. Basic categories of psychology are put under scrutiny in the so-called functional-historical *analysis of categories* with the aim of reconstructing them under more satisfactory conditions. Psychological categories (or rather what they contain), like perception, motivation, learning, emotion, cognition, are considered in their phylogenic, historical, and ontogenic evolution. At the end of this process, newly defined or rehabilitated categories are offered for the construction of an encompassing system of psychology.

Functional-historical analysis blends two things. One is C. Darwin's theory of evolution. It is utilized mainly for armchair paleozoology and paleoanthropology, and amplified and revised by the other constituent, Hegelian dialectics in its specific modification formulated by K. Marx, F. Engels, and later luminaries of Dialectical Materialism like G. V. Plekhanov. Functional-historical analysis therefore includes controversial Marxist tools of thought like negation in nonconceptual matters (antithesis), contradiction as causal agent (thesis–antithesis clash), or qualitative leaps or transitions (resulting in syntheses as unities of opposites, and forming the new theses for more dialectics). Such qualitative leaps are, among others, the transition from prepsychical to psychical organisms, the differentiation of special psychical functions, the evolution of learning, and the emergence of the societal nature of human beings.

These tools of thought are usually disregarded in science. It is therefore difficult for the reader not trained in their applications to follow the train of cogitations in which they are used. It is also not feasible to paraphrase them adequately in nondialectical prose, and therefore it may be fitting to let Critical Psychology speak for itself. This is a summary of the five steps of the functional-historical method according to Holzkamp (1983, p. 78):

> 1. Identify the real historical dimensions within the preceding developmental stage of the organisms at which the quantitative transformation in question took place; that is, determine the 'position' that is dialectically 'negated' by the qualitative transformation and thereby bring the specificity of the new developmental stage into relief.
> 2. Identify in the external conditions the objective alterations that constitute the 'environmental pole' of the inner developmental contradiction that causes the new quality to emerge. Such alternations involve a 'moderate discrepancy' that, on the one hand, demands 'compensations' from the organismic system, yet is supportable within its capacity. If it is not, the overall organism–environment system collapses, the inner contradiction is turned into an external opposition, and its antipodes will no longer be mediated as 'poles' in further development.
> 3. Identify the change in function of the relevant dimension demonstrated in step 1 as the 'organism pole' of the developmental contradiction and with it the origin of the first qualitative leap, that is, the development of the specificity of the new function under the altered conditions. This dialectical negation of a previously prevalent function does not determine the overall process but, in a sense, still serves to maintain the system at the earlier stage.
> 4. Identify the change of dominance between the earlier function characteristic of the system's maintenance and the new function, which then determines the specifics of the system. (This change presents itself as a discontinuous reversal of the relationship between two continuously changing dimensions.)
> 5. Identify the ways in which the overall development of the system is restructured and assumes new direction (that is,

> analyze the 'specific-secondary alterations,' and so on) after the qualitatively more specific function has become determining for the system's maintenance. With this identification of the qualitatively new dimensional structure, in which further qualitative transformations will occur, follows a return to the first step of the analysis but at a higher level. (Tolman and Maiers 1991, p. 39f.).

Categorial analysis leads to three characteristics of the human societal mode of existence: action possibility, action potency, and subjective situation. The latter is a personal assessment of one's own action possibilities and of where one stands in a particular objective situation. This derivation leads to a novel understanding of subjectivity and of the famous quote from Marx: 'The essence of man is conscious activity.'

In personality psychology and psychotherapy, Critical Psychology tries to utilize Freudian psychoanalysis. After a deft functional-historical analysis of its categories and of the fumbling efforts of Freudo-Marxists, it presents itself as the Marxist alternative to psychoanalysis.

The enthusiasm of early critical psychology to investigate the history and function of psychology is not followed by Critical Psychology. There is no critical historical analysis of the fate of psychology and psychologists in the USSR, especially during the times of Stalinism and anticosmopolitanism when Vygotsky initiated the historicocultural approach, and Rubinshtein and Leont'ev became its proponents. Nor is the practical application of psychology in the USSR or DDR, particularly in their armed forces and secret services, examined.

5. *Critique of Critical Psychology*

There have been numerous controversies around Critical Psychology, among them the following.

Holzkamp and the not yet capitalized critical psychology came under attack by H. Albert and others (Albert and Keuth 1973) who championed a Popperian position for psychology and the social sciences. The ensuing debate in psychology reiterated the main positions of the positivism controversy between Popper and the Frankfurt school.

Holzkamp's turn to Critical Psychology and the associated Marxist-Leninist position provoked the criticism of other tendencies in the spectrum of critical psychology. This resulted in a number of more or less ephemeral publications, the most accessible of which is edited by Busch et al. (1979). Points of criticism were the usurpation of the name 'critical psychology,' the political orientation on Marxist-Leninist lines, and a perceived retrocession to bourgeois conceptions of psychology.

Disapproval of Critical Psychology by psychologists working in the DDR could not be uttered before the collapse of Real Socialism and the Soviet empire. But then they could report that party officials reprimanded their local psychologists for staying aloof from party doctrine and advised them unpleasantly to follow the ideologically blameless example of Critical Psychologists.

6. *Other Critical Psychologies*

It is not appropriate to call Critical Psychology German critical psychology, as Teo (1998) does. There are and always have been other critical psychologists in Germany who have no affiliations with Critical Psychology. They gathered around the journal *Psychologie und Gesellschaft*, which was established in 1977 and renamed *Psychologie und Gesellschaftskritik* in 1978. Rexilius (1988) assembled some of the non-aligned voices.

An early voice in critical psychology was that of Peter Brückner (1922–82). From 1967, he occupied the chair of psychology at the Technical University in Hanover. There was no professional training of psychologists. Brückner did not publish systematic treatises on his view of psychology or on standard examination topics, but he authored a *Social Psychology of Capitalism* (1972), and papers on political psychology and on psychoanalysis (1983). Although he had some disciples, they did not organize a school.

A comparative portrayal of three early critical psychologies, Vygotsky's, Holzkamp's, and K. F. Riegel's dialectical psychology, was published by Ijzendoorn et al. (1984).

There is a recent tendency in English-speaking countries for various New Left tendencies to rally under the name of 'critical psychology.' Information on present trends may be found in Fox and Prilleltensky (1997), Sloan (2000), the newly launched *Annual Review of Critical Psychology* (from 1999), the *International Journal of Critical Psychology* (from 2000), or the Critical Psychology website at *www.sar.bolton.ac.uk/Psych/Main/WebRing.htm*.

It seems to be too early to extract consensus and divergence among the positions found here, although the Marxist moral critique of capitalism appears to be a pervasive constituent, not infrequently connected with the moral part of Lenin's critique of imperialism. The theoretical pillars of Marxism, however, like the labor theory of value, the theory of class struggle as the motor of social change, the theory of the approaching collapse of capitalism, and dialectical and historical materialism seem to be losing plausibility lately.

7. *Concluding Remark*

The various critical psychologies present worthwhile case study material for research in the predicament of the discipline of psychology or the dialectics between psychology as a science and as a profession. When in the twentieth century psychology was hastily metamorphosed into a practical profession, the university institutes had to be converted from pure research

institutions into training establishments—and they had to somehow cope with the egregious gap between the high-minded expectations of application-minded students and the comparatively limited amount of effective tools psychology had to offer. This left many students dissatisfied and a prey of neophilism.

See also: Critical Theory: Contemporary; Critical Theory: Frankfurt School; Dialectical Thinking, Psychology of; Hermeneutics, Including Critical Theory; Marx, Karl (1818–89); Marxism and Law; Marxism/Leninism; Marxist Social Thought, History of; Psychology: Historical and Cultural Perspectives; Psychology: Overview; Vygotskij, Lev Semenovic (1896–1934)

Bibliography

Albert H, Keuth H (eds.) 1973 *Kritik der kritischen Psychologie*. Hoffmann und Campe, Hamburg, Germany

Brückner P 1972 *Zur Sozialpsychologie des Kapitalismus: Sozialpsychologie der antiautoritären Bewegung*. Europäische Verlagsanstalt, Frankfurt am Main, Germany

Brückner P 1983 *Zerstörung des Gehorsams: Aufsätze zur politischen Psychologie*. Wagenbach, Berlin

Busch T, Engelhardt W, Geuter U, Mattes P, Schulte D 1979 *Zur Kritik der Kritischen Psychologie. Psychologie, Erkenntnistheorie und Marxismus*. Oberbaumverlag, Berlin

Fox D, Prilleltensky I 1997 *Critical Psychology: An Introduction*. Sage, London

Geuter U, Mattes P 1984 Historiography of psychology in West Germany: Approaches from social history. *Storia e Critica della Psicologia* **5**: 111–26

Holzkamp K 1972 *Kritische Psychologie: Vorbereitende Arbeiten*. Fischer-Taschenbuch-Verl, Frankfurt am Main, Germany

Holzkamp K 1973 *Sinnliche Erkenntnis—Historischer Ursprung und gesellschaftliche Funktion der Wahrnehmung*. Athenäum Fischer, Frankfurt am Main, Germany

Holzkamp K 1978 *Gesellschaftlichkeit des Individuums: Aufsätze 1974–1977*. Pahl-Rugenstein, Cologne, Germany

Holzkamp K 1983 *Grundlegung der Psychologie*. Campus Verlag, Frankfurt am Main, Germany

Ijzendoorn M H van, Veer R van der, Goossens F A 1984 *Main Currents of Critical Psychology: Vygotskij, Holzkamp, Riegel*. Irvington, New York

Kritische Psychologie 1969 Fachschaft Psychologie, Bochum University, Germany

Rexilius G (ed.) 1988 *Psychologie als Gesellschaftswissenschaft: Geschichte, Theorie und Praxis kritischer Psychologie*. Westdeutscher Verlag, Opladen, Germany

Sloan T (ed.) 2000 *Critical Psychology: Voices for Change*. St. Martin's Press, New York

Teo T 1998 Klaus Holzkamp and the rise and fall of German critical psychology. *History of Psychology* **3**: 235–53

Tolman C W 1994 *Psychology, Society and Subjectivity: An Introduction to German Critical Psychology*. Routledge, London

Tolman C W, Maiers W 1991 (eds.) *Critical Psychology: Contributions to an Historical Science of the Subject*. Cambridge University Press, Cambridge, UK

H. U. K. Gundlach

Critical Race Theory

In contemporary US social life, it is popularly accepted that racism is a form of injustice and that state law has the moral obligation, not only to refrain from perpetrating racist acts against citizens, but to provide remedies in certain circumstances to the victims of non-state racism as a matter of public policy. Most debates over antidiscrimination law and policy focus on whether the state's obligation to reject racism requires eschewing all race-conscious state and non-state public action, or whether race-conscious action should be permitted or even required in certain circumstances in order to remedy past and continuing racial discrimination. Both positions implicitly assume that racism is not integral to contemporary law or society and can be eliminated with determined efforts.

Writers who associate themselves with critical race theory take the position that racism is ordinary and normal in American society (Delgado 1994), indeed perhaps integral to it. Accordingly, critical race theorists attempt to show how contemporary law—including contemporary antidiscrimination law—accommodates and even facilitates racism. The critical race theory project inevitably calls into question fundamental assumptions about legal institutions and legal reasoning, about the trajectory of American history, and about the nature of injustice.

One tendency within critical race theory, deriving from the tradition of civil rights activism and jurisprudence, is strongly reformist in bent and seeks to galvanize moral outrage and direct it toward constructive legal change. A second influence on critical race theory is the poststructuralist tradition, within which the very possibility of overcoming racism through discursive structures such as law is considered a false hope. Critical race theory thus reflects a distinctive mix of reformist zeal and critical pessimism (Harris 1994). For example, Derrick Bell has urged civil rights activists to accept that, although there is a moral duty to challenge racism, racism is a permanent feature of American society (Bell 1992).

1. Origins of Critical Race Theory

Critical race theory emerged in the USA in the 1980s as a product of political, intellectual, and sociological developments in American legal academia. Politically, the 1980s were a time when many American civil rights activists and left-wing legal scholars felt themselves caught up in a conservative backlash against the gains of the 1960s. The scholars who would later call themselves critical race theorists sought explanations for why formal legal equality had produced only modest success in improving the lived experience of most African–Americans and other people of color,

and why hopes for social 'integration' with whites seemed largely to have faded (Crenshaw 1988).

Intellectually, the early 1980s saw the heyday of Critical Legal Studies (CLS) in the American legal academy. CLS introduced into legal scholarship post-structuralist ideas that had already permeated other disciplines focused on the interpretation of texts, including philosophy, literary criticism, and anthropology. Within legal studies, CLS scholars challenged the divide between 'law' and 'politics,' arguing that legal rules were radically indeterminate. CLS called for the deconstruction or 'trashing' of legal doctrine in order to expose the reification of legal concepts and the hegemonic function of traditional legal reasoning (Kelman 1987). CLS also provided an intellectual opening for broad-based critiques of law based on hidden norms of race and gender exclusion. For example, the first collection of essays in critical race theory appeared under the title 'Minority Critiques of the Critical Legal Studies Movement'; its authors were CLS sympathizers who nonetheless were disappointed with CLS's failure to extensively engage with racism (Delgado 1987, Williams 1987).

Sociologically, the 1980s saw the gradual entry into legal academia of African–Americans and other people of color, albeit in small numbers. Many of these new scholars were drawn to study race and the law, were frustrated by the disappointing outcomes of traditional avenues of legal reform in an era of backlash, and perceived traditional civil rights scholarship as a literature that ignored the voices of people of color (Delgado 1984). Several key figures in what would become critical race theory were radicalized as students by the departure of Derrick Bell—one of the pioneering African–American scholars critical of existing civil rights jurisprudence—from Harvard Law School, and by the administration's failure to hire an adequate replacement. These Harvard students created their own course on race, racism and the law, and staffed it with visiting scholars from around the country. The movement known as 'critical race theory,' and the term itself, emerged from this working group of students and teachers (Crenshaw et al. 1995).

2. *Themes of Critical Race Theory*

One of the earliest themes to appear within critical race theory was epistemological in nature, and was reflected in a distinctive method. Why were people of color much more likely to see American society as racist and to be pessimistic about the possibility of eliminating racism than whites? Did they know something whites did not? Was there a distinctive 'voice of color' that was missing in public policy, indeed in legal scholarship? Many critical race theorists argued that existing antidiscrimination law omitted the perspectives of the racially subordinated, and that subordinated groups in any social system have knowledge that the privileged lack. Thus, in a famous essay Mari Matsuda argued that civil rights scholarship and practice should 'look to the bottom' of society for leadership and insight (Matsuda 1987).

Critical race theorists also frequently abandoned the third-person objective voice in their legal writing in order to openly 'tell stories.' For example, Derrick Bell's 'Civil Rights Chronicles' featured an imaginary character who could travel through time, and who used her magical powers to criticize American civil rights doctrine from the founding of the Constitution to the present (Bell 1985, 1992). The telling of personal anecdotes or fables was intended not only to convey ideas that could not be expressed in traditional scholarly language, but also to criticize the assumption that the existing norms of legal scholarship, legal reasoning, and legal institutions operated objectively and produced neutral assessments of public policy. Critical race theorists hoped to show that the legal scholar's 'perspectivelessness' (Crenshaw 1989b) was in fact a perspective aligned with white privilege. Thus, for example, Patricia Williams, trained in post-structuralist theory and CLS, used personal stories, a dense web of allusions, and literary techniques to critique the social and legal construction of racial realities (Williams 1991).

The assertion of a distinctive voice of color proved controversial; critical race theory, like CLS before it, was bitterly attacked as racial 'special pleading' and as degrading the rigorous standards of argument and research established by traditional legal scholarship (see, for example, Kennedy 1989). Legal 'storytelling' was similarly attacked as sloppy scholarship and as an attempt to destroy the legacy of the Enlightenment. In this attack critical race theory was taken to be a postmodernist assault on truth itself (see, for example, Farber and Sherry 1997).

Most of the writing within critical race theory, however, has not taken on philosophical questions but rather has consisted of a sustained attack on the US Supreme Court's recent jurisprudence, particularly its constitutional jurisprudence. For example, throughout the 1980s and 1990s, the Court asserted its fidelity to four constitutional 'traditions' (Hayman Jr. 1995). First, the Court has increasingly favored a position of 'colorblindness,' under which all government racial classifications are presumptively unconstitutional, except perhaps when necessary to compensate a group of plaintiffs who can demonstrate the existence of conscious, deliberate racial bigotry that caused material disadvantage. Second, the Court has frequently deferred to legislatures, to state and local governments, and to the discretion historically exercised by public actors such as criminal prosecutors and juries when these actors are charged with racial discrimination. Third, the Court has deferred to private choices and free markets to produce the optimal allocation of resources. Fourth, the Court has

supported traditional indicators of 'merit' when these are challenged as racially biased.

Critical race theorists have painstakingly attempted to show how these jurisprudential developments block the possibility of political, economic, and social redistributions that would disrupt systemic racism. Some critiques have focused on unpacking the rhetorical structure of judicial opinions concerning race, or the deep structure of the Court's contemporary race jurisprudence (Ross 1989, Siegel 1997). Other scholars have mounted sustained attacks on recent doctrinal developments. For example, in an influential article Charles Lawrence III attacked the requirement that plaintiffs invoking the 14th amendment equal protection clause must prove that the defendants harbored the conscious desire to discriminate on the basis of race (Lawrence 1987). Lawrence, Mari Matsuda, Richard Delgado, and Kimberle Crenshaw have criticized the Court's apparent elevation of first amendment speech rights over 14th amendment norms of equality; the authors suggest that racist hate speech may constitutionally be sanctioned (Lawrence and Matsuda 1993).

In keeping with CLS's denial of any meaningful distinction between 'law' and 'politics,' critical race theorists have also sought to place jurisprudential developments in a larger political and historical context. For example, the Supreme Court trends just described coincided with political attacks on affirmative action by conservatives, a 'war on drugs' that locked up more and more African–American and Latino men in American prisons and jails, social panic about Asian and Mexican immigrants to the USA (reflected in nativist measures such as English-only laws and cutbacks of social welfare benefits for noncitizens), and the noisy public debate over whether curricula in American history and literature should be 'multicultural.' Critical race theorists sought to connect the Supreme Court's jurisprudence with larger political developments in a narrative of resurgent racism—a racism that constantly changes in form and expression (Calmore 1997, Siegel 1997).

A third theme within critical race theory has been an effort to understand how racism interlocks with other forms of oppression, including those focusing on gender, class, and sexuality. From the beginning, many critical race theorists regarded themselves as feminists, and several addressed the failure of mainstream feminism to respond to the interests and perspectives of women of color (Crenshaw 1989a, Harris 1990). Later writers in this tradition sometimes refer to themselves as 'critical race feminists' (Wing 1997), and have produced a body of work that criticizes legal institutions and legal doctrine from the perspective of women of color, particularly African–American women (Roberts 1997). Most of the work in this tradition has come to treat 'race' as only one axis of a multidimensional system of status oppression (Hutchinson 1997). A notable exception is Regina Austin, who treats racial identity as the primary focus of her work and examines state and nonstate constraints on 'the black public sphere' (Austin 1994).

Most recently, an offshoot of critical race theory—Latina Critical Theory or 'LatCrit' as it has been familiarly dubbed—has devoted itself explicitly to exploring the interplays of race, ethnicity, religion, national origin, class, and sexuality (Valdes 1998). The appearance of LatCrit also marks a move beyond the traditional 'black/white paradigm.' Much traditional US civil rights jurisprudence, and much early critical race theory, focused on caste racism against African–Americans. Scholars associated with LatCrit have explored the racialization of Latinos, Asian–Americans, and American Indians, and have treated immigration law as a branch of race law (Williams 1991, Perea 1997). These scholars identify nativist racism as a phenomenon distinct from, although related to, caste racism (Chang and Aoki 1998). Critical race theorists have also sought to recover and examine aspects of American race history, such as Asian exclusion and the colonization of Puerto Rico, that have been neglected because they do not fit the traditional emphasis on African–American civil rights (Chin 1998, Roman 1997).

Finally, a fourth theme within critical race theory has been the effort to work out the implications for legal theory of understanding 'race' as a social construction, rather than a natural fact. Some scholars have imported the work of sociologists Michael Omi and Howard Winant on 'racial formation' into the legal academy, studying the influence of state law on the racialization of 'whites' and 'Mexicans' (Haney Lopez 1996, 1998). Others have traced the differing meanings of 'race' within judicial discourse and within social and political discourse, to show how conflicts often involve an unacknowledged manipulation of the term (Gotanda 1991, Hernandez 1998).

3. *Future Directions for Critical Race Theory*

One recently emerging interest for critical race theorists is reestablishing the 'praxis' link between the now-burgeoning theoretical literature and the world of practicing civil rights attorneys, for whom much critical race theory is foreign and intimidatingly 'academic.' Scholar–activists such as Eric Yamamoto have called for a reconstruction of civil rights practice in light of critical race theory's skepticism about legal reasoning and legal institutions, its emphasis on coalition-building among groups suffering from discrimination, and its suggestion that law is only a weak and uncertain tool in the greater political struggle for empowerment (Yamamoto 1997).

Critical race theorists have also recently begun to incorporate in their analyses the facts that not all groups racialized as nonwhite experience racism in the same way, and that racial minorities are frequently in relations of competition and subordination rather

than solidarity (Yamamoto 1999). As the white population of the USA shrinks and the Asian–American and Latino populations increase, the binary relationship of white versus nonwhite will be increasingly inadequate to understand racialized relations of power. Similarly, critical race theorists have begun to address the issue of class, ethnic, and color distinctions within particular 'races.'

A third recent development within critical race theory has been a growing interest in 'white privilege.' One aspect of this development is an interest in exploring white identity, facilitated by the growing literature on 'critical white studies' outside the academy (Flagg 1998, Delgado and Stefancic 1997). Additionally, there is growing interest in taking racial 'privilege' rather than 'discrimination' as the object of study and critique (Wildman 1996). The traditional focus on racial victimization leaves unexamined the everyday ways in which persons who are not direct perpetrators of racism nevertheless benefit from racial hierarchies. In keeping with the structuralist ambitions of critical race theory, the turn toward privilege moves beyond the focus on immediate perpetrators and victims.

Moreover, critical race theory seems to be moving toward placing the US experience in the context of a world system of economic and status inequality. Scholars of international human rights and international relations have begun to incorporate critical race theory into their work (Hernandez-Truyol 1996, Gordon 1997). LatCrit scholars have followed their interest in immigrant communities within the USA to the transnational relationships of labor and capital that govern immigrant flow. Critical Race Theory has begun to converge with postcolonial studies in developing a historical perspective on the USA that treats it as a country still in recovery from its colonial history. Critical race theory has also begun to see the USA as part of an international system that divides the 'North' from the 'South' and that concentrates economic and political power in the North. In contrast with its beginnings as a movement largely rooted in and focused on the USA, critical race theory in the twenty-first century seems to be ready to 'go global.'

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on its Effectiveness; Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Conflict Sociology; Critical Legal Studies; Discrimination: Racial; Ethnic and Racial Social Movements; Feminist Legal Theory; Feminist Theory; Gender, Class, Race, and Ethnicity, Social Construction of; Hermeneutics, Including Critical Theory; Minorities; Race and the Law; Race: History of the Concept; Race Identity; Race Relations in the United States, Politics of; Racial Relations; Racism, History of; Racism, Sociology of

Bibliography

Austin R 1994 'An Honest Living': Street vendors, municipal regulation, and the black public sphere. *Yale Law Journal* **103**: 2119–31

Bell D A Jr. 1985 Foreword: The civil rights chronicles. *Harvard Law Review* **99**: 4–83

Bell D A Jr. 1992 *Faces at the Bottom of the Well*. Basic Books, New York

Calmore J O 1997 Exploring Michael Omi's 'Messy' real world of race: An essay for 'Naked people longing to swim free.' *Law and Inequality: A Journal of Theory and Practice* **15**: 25–82

Chang R S, Aoki K 1998 Centering the immigrant in the inter/national imagination. *La Raza Law Journal* **10**: 1395–447

Chin G J 1998 Segregation's last stronghold: Race discrimination and the constitutional law of immigration. *UCLA Law Review* **46**: 1–74

Crenshaw K W 1988 Race, reform, and retrenchment: Transformation and legitimation in antidiscrimination law. *Harvard Law Review* **101**: 1331–87

Crenshaw K W 1989a Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine. *University of Chicago Legal Forum* 1989: 139–67

Crenshaw K W 1989b Foreword: Toward a race-conscious pedagogy in legal education. *National Black Law Journal* **11**: 1–14

Crenshaw K, Gotanda N, Peller G, Thomas K 1995 *Critical Race Theory: The Key Writings That Formed the Movement*. The New Press, New York

Delgado R 1984 The imperial scholar: Reflections on a review of civil rights literature. *University of Pennsylvania Law Review* **132**: 561–78

Delgado R 1987 The ethereal scholar: Does critical legal studies have what minorities want? *Harvard Civil Rights—Civil Liberties Law Review* **22**: 301–22

Delgado R 1994 Bibliographic essay: Critical race theory. *SAGE Race Relations Abstracts* **19**(2): 3–28

Delgado R, Stefancic J 1997 *Critical White Studies: Looking Behind the Mirror*. Temple University Press, Philadelphia, PA

Farber D, Sherry S 1997 *Beyond All Reason: The Radical Assault on Truth in American Law*. Oxford Press, New York

Flagg B 1998 *Was Blind But Now I See: White Race Consciousness and the Law*. New York University Press, New York

Gordon R 1997 Saving failed states: Sometimes a neocolonialist notion. *American University Journal of International Law and Policy* **12**: 903–74

Gotanda N 1991 A critique of 'Our constitution is colorblind.' *Stanford Law Review* **44**: 1–68

Haney Lopez I F 1996 *White By Law: The Legal Construction of Race*. New York University Press, New York

Haney Lopez I F 1998 Race, ethnicity, erasure: The salience of 'Race' to LatCrit theory. *La Raza Law Journal* **10**: 57–125; *California Law Review* **85**: 1143–211

Harris A 1990 Race and essentialism in feminist legal theory. *Stanford Law Review* **40**: 581–616

Harris A 1994 Foreword: The jurisprudence of reconstruction. *California Law Review* **82**: 741–85
Hayman R L Jr. 1995 The color of tradition: Critical race theory and postmodern constitutional traditionalism. *Harvard Civil Rights—Civil Liberties Law Review* **30**: 57–108
Hernandez T K 1998 'Multiracial' discourse: Racial classifications in an era of color-blind jurisprudence. *Maryland Law Review* **57**: 97–173
Hernandez-Truyol B E 1996 Natives, newcomers, and nativism: A human rights model for the twenty-first century. *Fordham Urban Law Journal* **23**: 1075–137
Hutchinson D L 1997 Out yet unseen: A racial critique of gay and lesbian legal theory and political discourse. *Connecticut Law Review* **29**: 561–645
Kelman M 1987 *A Guide to Critical Legal Studies*. Harvard University Press, Cambridge, MA
Kennedy R L 1989 Racial critiques of legal academia. *Harvard Law Review* **102**: 1745–819
Lawrence C R III 1987 The Id, the Ego, and equal protection: Reckoning with unconscious racism. *Stanford Law Review* **39**: 317–88
Lawrence C R III, Matsuda M J 1993 *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*. Westview Press, Boulder, CO
Matsuda M J 1987 Looking to the bottom: Critical legal studies and reparations. *Harvard Civil Rights—Civil Liberties Law Review* **22**: 323–99
Perea J 1997 *Immigrants Out! The New Nativism and the Anti-Immigrant Impulse in the United States*. New York University Press, New York
Roberts D 1997 *Killing the Black Body: Race, Reproduction, and the Meaning of Liberty*. New York University Press, New York
Roman E 1997 Empire forgotten: The United States's colonization of Puerto Rico. *Villanova Law Review* **42**: 1119–211
Ross T 1989 The Richmond narratives. *Texas Law Review* **68**: 381–413
Siegel R 1997 Why equal protection no longer protects: The evolving forms of status-enforcing state action. *Stanford Law Review* **49**: 1111–48
Valdes F 1998 Under construction: LatCrit consciousness, community, and theory. *La Raza Law Journal* **10**: 1–56
Wildman S M 1996 *Privilege Revealed: How Invisible Preference Undermines America*. New York University Press, New York
Williams P J 1987 Alchemical notes: Reconstructing ideals from deconstructed rights. *Harvard Civil Rights—Civil Liberties Law Review* **22**: 401–33
Williams P J 1991 *The Alchemy of Race and Rights*. Harvard University Press, Cambridge, MA
Williams R A Jr. 1991 Columbus's legacy: Law as an instrument of racial discrimination against indigenous people's rights of self-determination. *Arizona Journal of International and Comparative Law* **8**: 51–75
Wing A K 1997 *Critical Race Feminism: A Reader*. New York University Press, New York
Yamamoto E K 1997 Critical race praxis: Race theory and political lawyering practice in post-civil rights America. *Michigan Law Review* **95**: 821–900
Yamamoto E K 1999 *Interracial Justice: Conflict and Reconciliation in Post-Civil Rights America*. New York University Press, New York

A. Harris

Critical Realism in Geography

Critical realism is a philosophy of and for the social sciences that has had an impact on methodology in geography since the 1980s, although there have always been realist elements in geographical thought. With the rise of critical realism, these and further elements were developed and promoted self-consciously, while non- or anti-realist elements were attacked.

1. *A Definition*

Critical realism argues that the natural and social worlds we study exist largely independently of the researcher, and that there is necessity in the world which enables and constrains what can happen. Causation is to be understood in terms of the powers that objects possess rather than in terms of empirical regularities among events; if and when these 'causal powers' are activated, the results depend on contexts, and are unlikely to be regular. Since social phenomena such as actions and ideas are intrinsically meaningful, unlike the subject matter of the physical sciences, their meaning also has to be interpreted. The world can only be understood in terms of available discourses. While there are no privileged routes for discovering the truth about the world, and empirical tests are themselves fallible, it is often possible to distinguish better from worse accounts in terms of their adequacy for informing material practice.

2. *Intellectual Context*

Critical realist philosophy was developed by Roy Bhaskar out of his earlier work on the philosophy of natural science ('transcendental realism') (Bhaskar 1978, 1989). The latter sought an alternative to both those positions which treat knowledge as straightforwardly based on observation and experience (empiricism) and for which truth is a simple matter of correspondence between statements and the objects to which they refer, and various relativist positions which treat the truth of theories as purely relative to discourses. Critical realism also steers a course between positions which regard social science as the same as natural science and those which treat it as entirely different.

3. *Realism and Fallibilism*

The most fundamental realist assumption is that much of the world can exist independently of our knowledge of it. The plausibility of this assumption rests upon the

fallibility of knowledge; the fact that knowledge claims sometimes prove to be mistaken and generate false expectations implies that the world is not a product of our thought but is whatever it is regardless of what we think about it. Realists reject the empiricist view that reality is whatever we can experience since this implies an improbable coincidence between the world and the limits of our sensory powers, and implies that either knowledge is already complete, or that new knowledge creates new objects. Critical realists accept that our observations and knowledge are always formed through available discourses or are 'theory laden', so that we can never escape from our conceptual frameworks to see how they compare with the world they seek to represent. However, that observation is theory laden does not necessarily mean that it is entirely determined by theories and so it can still be possible to compare the adequacy of different theories from within discourses and find observations which contradict expectations. This offers a way of avoiding the relativist view that truth is merely relative to theoretical systems. It also implies that science can make progress in explaining the world even though absolute truth is not an intelligible goal.

Critical realism therefore does not involve claims to have a privileged access to the truth about the world, a uniquely 'realistic' view, for the very independence of the world from our knowledge of it renders such claims problematic. It must necessarily view all knowledge as in principle fallible, although this does not mean, of course, that it is all false or all equally true or untrue.

4. *Key Arguments and Their Influence on Geography*

Critical realism became influential in geography partly for the above reasons, for geographers, like other scientists, were seeking a way of understanding science that avoided the unappealing alternatives of empiricism and relativism. However, a further important factor was probably its view of causation and systems. Some of the first discussions of critical realism in geography emphasized this issue and its implications for the nature of theorizing in geography (Sayer 1982, 1985).

The scientific revolution in geography, exemplified by *location theory* and *spatial analysis*, had largely assumed that causation was a matter of empirical regularities or 'constant conjunctions' among causes and effects. It was expected that understanding of the regularities in spatial organization of phenomena would grow to the point where it would be possible to formulate laws about them. This regularity-based view of causation had dominated the philosophy of science despite the fact that, particularly in social sciences like human geography, there appear to be few enduring precise empirical regularities, so that the search for laws of social and spatial behavior has been singularly unsuccessful. For realists this is unsurprising. For them, causation is a matter of what produces change, not a matter of empirical regularities in sequences of events. Objects have causal powers, that is, the power to act in certain ways. A seed has the power to germinate, a person the power to work. These objects have these powers in virtue of their structures, such as the biochemical constitution of the seed, or the anatomical and mental structures of a person. Whether these powers are ever activated depends on conditions whose presence is contingent, that is, neither necessary nor impossible, such as moisture and solar energy in the case of the seed or a need for food in the case of the person. Furthermore, the particular effects which follow when these causal powers are activated depends on conditions which again are contingent. The germination of the seed may be encouraged by fertilizer or halted by herbicides, the attempt of the person to work may be hampered by lack of tools or encouraged by their availability.

This alternative conception of causation has far-reaching implications. First, since it dissociates causation from empirical regularities, it enables understanding of how unique events, as much as repeated ones, can be caused. Second, it focuses on necessity rather than regularity or the search for empirical order. Third, in view of this, explanation does not require repeated observations: what makes something happen has nothing to do with the number of times it has or has not been observed to happen. It follows from this that the old debate in geography between idiographic and nomothetic approaches—the former a supposedly nonscientific study of the unique, the latter a supposedly scientific search for laws—was falsely based. Regardless of whether phenomena were unique or widely replicated, they could be explained in the same way.

The uniqueness of many geographical phenomena no longer posed a threat to geography's scientific status. Fourth, it became possible to see that the search for generalizations is distinct from explaining how causes work; generalizations, as the term suggests, attempt to tell us how common or extensive certain phenomena are, and this is different from explaining how they are produced. The greater part of standard literature on the philosophy and methodology of science had confused these two things, and many geographers had done so too. Fifth, the realist account offers a way of understanding how it is possible for the same cause to produce different effects and for different causes to produce the same effects, since the relationship between causal powers and empirical outcomes is not one of automatic regularity but is dependent on the conditions within which they are located.

An example of this can be given from a kind of economic geography which was popular at the same

time critical realism became influential. The geography of industrial restructuring looked at changes in economic geography in terms of how competitive pressures upon firms to restructure produced responses (e.g., Massey and Meegan 1982). In realist terms, competitive pressures on profit margins were considered as the main causal power of interest here, with their effects being mediated by a wide range of contextual circumstances, such as the particular cost structures of the firms under pressure, the scope for technological change, and the availability of various kinds of labor. Thus, in an industry such as clothing, which had little scope for further automation or speeding up work, and in which wages were a major element of costs, a common response was to seek out cheaper sources of labor, sometimes by relocating.

By contrast, in industries which were more technology intensive and had scope for further automation, a more common response was technological innovation. By such means, it was possible to explain how the same or similar causes could produce radically different results according to context, whereas approaches which rested upon the discovery of empirical regularities foundered. Similar research also illustrated how a single kind of outcome—job losses—could have completely different causes, e.g., a drop in demand for the firm's products, or the introduction of automation during steady or rising demand, or reorganization of work routines to eliminate idle time and intensify work (see also Henry 1992). For further discussion of these topics see *Economic Restructuring: Geographic Aspects*.

One of the most important and lasting influences of critical realism on geography has been on its empirical research designs (Sayer 1992). Hitherto, the dominance of conceptions of explanation as tied to generalizations about empirical regularities had been used to privilege 'extensive' research designs which sought to uncover such regularities or patterns of covariation. Although these are, as the term indicates, useful for establishing how extensive certain phenomena are, they do not necessarily identify causes, as opposed to spurious correlations. The latter are likely to be thrown up by extensive research because it preselects groups of objects taxonomically, that is, on the basis of similarities (e.g., industries of a certain kind), rather than functional connection. By contrast, intensive research designs, including ethnographic case studies, focus on causal rather than taxonomic groups and follow up causal connections wherever they lead—often between heterogeneous objects, such as firms of different kinds. Intensive research is typically more exploratory, since the purpose is to discover the relevant causal groups. It is often time consuming to do this for large numbers of cases, and therefore it is not always a good basis for generalization. However, as we have seen, this lack of generalizability does not undermine the status of causal explanations of the particular cases which are studied intensively. Consequently, having formerly been regarded as unscientific, intensive studies have now become more accepted in geographical research.

Critical realism acknowledges that the methodology of social science must diverge from that of the natural sciences because human, social phenomena are intrinsically meaningful. While the objects that, say, geologists study, are what they are regardless of any meaning people may give to them, objects such as socially segregated urban areas depend on the meanings actors give to them. To understand what these are, one has to know what their meaning is in society, whereas the nature of a rock does not depend on how its constituents understand one another, as they do not. Consequently, an important part of the methodology of human geography, no less than other social sciences, is interpretive understanding. For example, to understand the geography of the Middle East, we would have to understand what significance its land had for Jews and Arabs in terms of their identities.

Although this interpretive understanding has also been the focus of *humanistic geography*, critical realists make certain qualifications which the latter might not. The three most important are as follows. First, critical realists argue that although we have to interpret the understandings that actors have of their situations in order to make sense of how they act, their understanding is not necessarily a good one, indeed it may be systematically flawed. Moreover, these misunderstandings may have important effects which actors cannot explain. Thus a mysogynist may believe that the proper spatial location for women is in the home, on the false grounds that they are incapable of coping with life in the public sphere; the falsity of this understanding makes a difference to what happens and therefore has to be acknowledged in scientific accounts. Further, different actors may give contradictory accounts of the same situation: where this is the case the researcher cannot sit on the fence and say both are correct. Actors' accounts of situations therefore face social scientists' accounts both as objects of study and rivals. It is therefore likely that researchers' accounts will be at least implicitly critical of actors' accounts—a conclusion which humanistic geographers are generally unwilling to accept. It is primarily for this reason that Bhaskar's philosophy of social science is called *critical* realism.

Second, among the intrinsically meaningful aspects of social phenomena are actors' reasons. Critical realists argue that reasons can be causes, since they can be responsible for producing change, by prompting actions. This, of course, implies that realists do not limit causation to physical processes. To deny that reasons are causes is to imply that the giving of reasons by actors is ineffectual and redundant, which of course makes it unclear why they are given. Third, an insistence on the necessity of interpretive understanding should not allow us to forget that social life is also material, embodied, and spatially situated, and that

things happen *to* people, so that, for example, they can be dominated, sometimes regardless of their understanding of the situation.

The realist facility for explaining irregular processes was particularly attractive to a discipline such as geography, concerned as it is with spatial differentiation and uneven development, and hence unable to ignore—as, say, economics tries to do—the problems of applying equilibrium models or making generalizations which are supposed to hold across time and space. The systems which geography and other social sciences study are *open*, that is, their elements can themselves be undergoing internal qualitative change—actors aging and learning, for example—and the relations among the elements are not constant, with the result that any empirically regular behaviors are unlikely to be durable or widespread. Evolutionary change and disequilibrium rather than equilibrium are the norm.

Where previous attempts at scientific explanation in geography had tried to evade the prevalence of differentiation and irregularity, on the mistaken grounds that explanation required the discovery of regularity, human geography began to re-embrace differentiation, nowhere more strikingly than in Doreen Massey's influential *Spatial Divisions of Labour* (Massey 1984). This was closely followed by a burgeoning literature on locality studies, which typically focussed on how economic restructuring worked out in different localities, and how those differences impacted back on the restructuring process (e.g., Cooke 1986).

5. Critical Reception

Some geographers feared that such concerns heralded an abandonment of theory and generalization and a slide towards the empiricist documentation of facts (e.g., Harvey and Scott 1989), indicating a resistance to the critical realist de-emphasis of generalization. These fears were not merely methodological but partly about a weakening attachment in economic geography to a particular kind of theory—Marxism—and also a tendency to treat critical realism as a source of substantive theory rather than as a philosophy, when it had no such pretensions. At the same time, some of the attempts to use critical realism in localities studies invited justified criticism by arbitrarily translating it into the study of general, supralocal causal processes interacting with local contingent conditions. However, critical realism did not license such an extraordinary alignment of scale and necessity: the supralocal is no more or less the realm of necessity than the local, nor is the local any more the realm of contingency than the supralocal.

Towards the end of the 1980s a newer generation of geographers became influenced by postmodernism (see *Postmodernism in Geography*). In some ways this continued the emphasis on differentiation—only phrased as 'difference' which had connotations of social differences of identity. Societies were seen as inherently messy, and skepticism towards grand universalizing accounts of the world grew. One of the positive features of this literature was that it gave greater concern to the textual nature of accounts of the world (Sayer 1989, Stones 1996). At the same time, some postmodernist writers rekindled relativist, anti-realist tendencies (e.g., Barnes 1996, Graham 1990), although sometimes by mischaracterizing realism as claiming privileged access to the truth rather than as having a fallibilist view of knowledge. The debate continues (e.g., Peet 1998).

See also: Geographic Information Systems: Critical Approaches; Humanistic Geography; Phenomenology: Philosophical Aspects; Postmodern Urbanism; Postmodernism: Methodology

Bibliography

Barnes T J 1996 *Logics of Dislocation*. Guilford Press, New York

Bhaskar R 1978 *A Realist Theory of Science*. Harvester Press, Hassocks, Sussex

Bhaskar R 1989 *The Possibility of Naturalism*, 2nd edn. Harvester Wheatsheaf, Hemel Hempstead, UK

Cooke P 1986 *Localities: the Changing Face of Urban Britain*. Unwin Hyman, London

Graham J 1990 Theory and essentialism in Marxist geography. *Antipode* **22**: 53–66

Harvey D, Scott A J 1989 The practice of human geography: theory and empirical specificity in the transition from Fordism to flexible accumulation. In: Macmillan B (ed.) *Remodelling Geography*. Blackwell, Oxford, UK, pp. 217–29

Henry N 1992 The new industrial spaces: locational logic of a new production era. *International Journal of Urban and Regional Research* **16**: 375–96

Massey D 1984 *Spatial Divisions of Labour*. Macmillan, London

Massey D, Meegan R 1982 *Anatomy of Job Loss*. Methuen, London

Peet R 1998 *Modern Geographic Thought*. Blackwell, Oxford, UK

Pratt A C 1994 *Uneven Re-production: Industry, Space and Society*, 1st edn. Pergamon, Oxford, UK

Sayer A 1982 Explanation in economic geography. *Progress in Human Geography* **6**: 68–88

Sayer A 1985 Realism and geography. In: Johnston R J (ed.) *The Future of Geography*. Methuen, London, pp. 159–73

Sayer A 1989 The 'new' regional geography and problems of narrative. *Environment and Planning D: Society and Space* **7**: 253–76

Sayer A 1992 *Method in Social Science: a Realist Approach*, 2nd edn. Routledge, London

Stones R 1996 *Sociological Reasoning: Towards a Post-modern Sociology*. St Martin's Press, New York

A. Sayer

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Critical Theory: Contemporary

Critical theory was originally associated with the Frankfurt Institute in the 1930s, the key figures of which included Max Horkheimer,Theodor Adorno, Walter Benjamin, and Herbert Marcuse. Critical theory embodied the Hegelian and psychoanalytic reorientation of Western Marxism during the 1930s in which theory was transformative, rather than contemplative, engaging with historical processes to enhance human emancipation (Horkheimer 1972). Unlike more orthodox Marxism, it accorded central significance to culture, including mass media, authoritarianism, science and technology, and gender. Central to their thinking was the apparent eclipse of the revolutionary proletariat in the wake of Stalinism, Nazism, and mass consumption capitalism. Many of these themes, expressed for example in Horkheimer and Adorno (1973), were to become central to social theory some 25 years later. As this happened, the distinctive character of critical theory became less clear, although it remains an important focus of critical theoretical debate.

1. *Contemporary Critical Theory*

Much contemporary critical theory though has been shaped by Jürgen Habermas (b. 1929) who, despite often being described as the 'heir' to earlier critical theory, has criticized many of its claims. He regards the apocalyptic despair of Horkheimer and Adorno's work as having ignored the 'unfulfilled potential' of Western modernity to expand rationality and democracy (Habermas 1987, p. 113). They too readily accepted a Weberian nightmare of an iron cage of total domination, in which philosophy could only keep alive the possibility of a better society while lamenting the lost opportunity for emancipation (Benhabib 1981). For Habermas though, the very possibility of critique must presuppose the existence of a rational criteria of a better society against which the present is being judged (Habermas 1987, p. 113). Habermas aims to recover and explicate this potential, to which end he has undertaken an elaborate reconstruction of the grounds of everyday communication.

1.1 *Communicative Turn*

Habermas' early work focused on the constriction of the public sphere through technocratic politics and commercial mass media. Though less central to later work, this theme is manifest in Habermas' distinction between two social orders, of lifeworld and system. The former is the sphere of communication and culture, mediated by norms and language. The latter refers to forms of social 'steering' by money and power, which are not available to normative discourse. While Habermas insists that the market and state are essential in modern societies, there is a continual tension along the boundaries of lifeworld and system along which new social movements arise. An example would be environmental protests over nuclear reprocessing that open debates about the rationality and morality of projects previously driven by technological and financial criteria.

In recent work though Habermas has identified emancipatory practice not by appeal to agents of change (such as the proletariat) but through 'universal pragmatics'—the reconstruction of rationality implicit in speech acts. Habermas argues that the potential for rational agreement is present whether we acknowledge it or not, in sociolinguistic rules of communication. He uses Chomsky's linguistics and Austin and Searle's theory of speech acts to reconstruct communicative rules, arguing that the elementary units of communication (speech acts) involve validity claims that are 'naively accepted' in conversations. For conversations to occur we assume agreement about grammar and illocution (we recognize something *as* a promise, an assertion, an order, etc.). Similarly, judgments about the comprehensibility of actions involve evaluations as to whether good reasons have been provided. Although moral choices are often regarded as matters of personal conviction, Habermas agues that the requirement to provide good reasons, subject to public scrutiny, renders these choices 'capable of truth.' The evaluation of moral choices is comparable to the evaluation of empirical claims, in that in both our proffered reasons are subject to critical testing (Habermas 1992, p. 76). Thus Habermas seeks a practical and cognitive basis for critical theorizing.

These are not timeless linguistic structures, but unfold historically through the formation of post-traditional (modern) worldviews. Unlike Marxist theories, which view social development as the succession of modes of production, Habermas' theory refers to stages of cognitive and moral learning, which parallel childhood development. Premodern forms of moral authority involve conventional attachments to externally given value systems, underwritten by sacred beliefs. Echoing Kant, 'maturity' involves post-conventional distance from rules and the capacity to evaluate and argue whether adherence to a moral injunction is appropriate. Successful post-traditional sociality institutionalizes new levels of cognitive and moral learning, thus normative structures are the 'pacemaker of social evolution' (Habermas 1979, p. 120). Modern worldviews dissolve sacred authority into rational justifications of morality.

This process of the 'linguistification of the sacred' is potentially liberating because it releases the 'argumentative force of language' grounded in the capacity of assent or dissent from any proposition, whether empirical, moral, or aesthetic. In each of these spheres any attempt to reach an understanding of the truth of

a proposition, the sincerity of the speaker, or the appropriateness of the utterance can be disputed. Habermas claims that entering into dialog about the validity of utterances involves a specific type of discussion (*Diskurs*) oriented towards examining these conditions (Habermas 1984, p. 28). But here is the nub of the argument. Such practices will be successful only if we assume equality of access to speech acts and freedom to move from level to level in *Diskurs*, since it is only under such conditions that consensus could be motivated by the force of better argument. This enables Habermas to introduce the 'general symmetry requirement' (also called the 'ideal speech situation') which states that if a consensus is to be reached guided only by the force of better argument, then whether we recognize it or not we routinely assume certain conditions. That is:

everyone has an equal chance to deploy, initiate, and perpetuate speech acts;
utterances are comprehensible;
their propositional content is true;
what is said is legitimate and appropriate;
it is sincerely spoken.

Habermas realized, of course, that these conditions of communicative competence are counterfactual and rarely present in actual speech, but it is precisely this that justifies two further claims. The first is that the goal of normal communication is disturbed by power relations, which intrude to prevent questions being raised or validity claims being tested. Second, these constraints on communication are self-defeating since they contradict the underlying assumptions that make dialog possible at all. From this it follows that rationality can be measured by the degree of openness or closure in communication. The goals of truth, freedom, and justice are not mere utopian dreams, but are anticipated in ordinary communication; and therefore the goal of emancipation is presupposed in the constitution of the species as linguistic beings (Habermas 1991, p. 244). If the unspoken authority of all communications can potentially be challenged, then perhaps democratic social relations are already implicit in everyday communications. For this insight to have practical significance though, it would need to be linked to the formation of emancipatory social movements: feminism, ecology, civil rights, and democratization movements (Ray 1993, pp. 57–78).

1.2 How Critical is Critical Theory?

Contemporary critical theory has developed a wide range of debates and projects in the course of which Habermas' has been extensively criticized. For some (Hullot-Kentor 1989) he has abandoned the earlier critical thrust of Horkheimer and Adorno. For others, he privileges communicative functions of language over others, such as irony or aesthetic expression (Thompson 1982) and has misunderstood or misused speech act theory (Rasmussen 1990, p. 37). One of the most frequent criticisms is that if Habermas envisages domination-free dialog as an attainable state, this is an implausible utopian dream (Jay 1988, p. 31). If it is not an attainable state then the force of his critique is lost. Similarly, Habermas' emphasis on communication and language is idealist and perhaps ignores the importance of material structures (Dux 1991). Brian Fay suggests that emancipation through rational reflection encounters limitations imposed by 'embodiment,' since authority is inscribed into conditioned patterns of activity, which cannot readily be subject to rational refection (Fay 1987, p. 148). Two areas of controversy in particular serve to highlight the relevance of critical theory to contemporary debates.

2. Critical Theory and Feminism

Some of the most interesting critiques of Habermas have come from feminist critical theorists, who share his commitment to the explication of power, knowledge, and morality, but differ with him on other issues. A central issue is that Habermasian theory stands in the Enlightenment tradition which many feminists have rejected (Meehan 1995). Universal pragmatics deals with highly abstract and depersonalized speech rules. By contrast, feminist critics have identified discursive forms that represent experience of embodied female subjects. For example, Nancy Fraser (1985) argues that Habermas fails to examine the gendering of his core concepts such as citizenship, universalism, public sphere, and civil society. The role of citizen is premised on male occupation of the public sphere in which a civil contract amongst brothers is contrasted with the feminization of the private sphere.

Seyla Benhabib criticizes what she sees as the disembodied transcendental rational ego that is the subject of ideal speech. Drawing on communitarian critics of the 'unencumbered self' (Taylor, MacIntyre), she sees in Habermas's attempt to reconstruct universal morality an autonomous male ego, which universalism privileges as a point of reference. Since subjects are finite and social, contexts of gender and community are central and people are not only generalized others (moral persons with reason) but also concrete others. The negotiation between the 'generalized other' and collective concrete others takes place in a male public sphere from which women disappear (Benhabib 1992, p. 13). Nonetheless, Benhabib applauds Habermas' shift to communication, which departs from reason as a psychological attribute of consciousness to an intersubjective view of ethics and dialog.

A similar point is made by Kate Soper, who notes the 'all male monastical atmospherics of the Frankfurt School,' but sees nonadversarial dialog as similar to consciousness-raising, and his distinction between normatively secured and communicatively achieved

dialog as a means of opening up unquestioned conventions (e.g., family) to explicit discussion. Indeed, there is some anxiety among these critics that the theory risks losing sight of universalism in favor of Romantic particularism. Benhabib asks, 'would a moral theory restricted to the standpoint of the concrete other be sexist, racist, culturally relative, and discriminatory?' 'No,' she replies, 'rights must be an essential component of any such theory.' The problem here (cf. Soper 1992) is that 'rights' must involve abstraction from particularity otherwise the grounding for the demand for equal treatment is removed.

3. *The Debate over Modernity*

Critical theory seeks to salvage Enlightenment rationality. Habermasappears to be the only contemporary theorist willing to defend the tradition of modernity, and is frequently called to do so in debates with theorists like Lyotard, Gadamer, and Foucault. Yet critical theory shares with postmodernism an attention to knowledge and communication media, the fragmentation of culture and identity, and the end of historical narratives. Nonetheless, for critical theory, postmodernism denies the grounds of its own critique and permits celebration of spontaneous power of imagination. That, on the contrary, modernity is not exhausted but unfulfilled, was demonstrated by the anticommunist 'rectifying revolutions' (*die Nachholende Revolution*) of 1989. They posed anew the core issues of modernity—rights, civil society, market governance, cultural autonomy, and democratic association (Habermas 1990, 1994, p. 62).

The robust defense of modernity and enlightenment is, paradoxically, as much a challenge to earlier critical theory as to postmodernism. An evaluation of contemporary critical theory should assess its success in identifying emancipatory impulses neglected by earlier theorists. Habermas gestures towards social movements as representing new communicative structures, although he says very little about these in detail. The key issue is whether the theory of communicative action is a plausible reconstruction of the emancipatory potentials inherent in everyday speech. Even if we find these arguments persuasive, the question remains whether these are cultural conventions or deeper structures constitutive of discourse.

See also: Adorno, Theodor W (1903–69); Communication: Philosophical Aspects; Critical Theory: Frankfurt School; Cultural Critique: Anthropological; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Marxist Social Thought, History of; Modernity; Postmodernism: Philosophical Aspects; Rationality in Society; Theory: Sociological

Bibliography

Benhabib S 1981 Modernity and the aporias of critical theory. *Telos*. **49**: 38–61

Benhabib S 1992 *Autonomy, Feminism and Postmodernism*. Blackwell, Oxford, UK

Dux G 1991 Communicative reason and interest. In: Honneth A, Joas H (eds.) *Communicative Action*. Polity, Oxford, pp. 74–96

Fay B 1987 *Critical Social Science—Liberation and its Limits*. Cornell University Press, Ithaca, NY

Fraser N 1985 What's critical about critical theory? Habermas and gender. *New German Critique*. **35**: 97–131

Habermas J 1979 *Communication and Evolution of Society*. Heinemann, London

Habermas J 1984 *The Theory of Communicative Action*: *Reason and the Rationalization of Society*. Heinemann, London, Vol. 1

Habermas J 1987 *Philosophical Discourse of Modernity*. MIT Press, Cambridge, MA

Habermas J 1990 What does socialism mean today? The rectifying revolution and the need for new thinking on the left. *New Left Review*. **183**: 3–21

Habermas J 1991 A reply. In: Honneth A, Joas H (eds.) *Communicative Action*. Polity, Oxford, pp. 214–64

Habermas J 1992 *Moral Consciousness and Communicative Action*. Polity, Cambridge, UK

Habermas J 1994 *The Past As Future*. Polity, Cambridge, UK

Horkheimer M 1972 Traditional and critical theory. In: Horkheimer M (ed.) *Critical Theory*. Herder and Herder, New York

Horkheimer M, Adorno T W 1973 *Dialectic of Enlightenment*. Allen Love, London

Hullot-Kentor R 1989 Back to Adorno. *Telos*. **81**: 5–29

Jay M 1988 *Fin de Siècle Socialism*. Routledge, London

Lovibond S 1994 Feminism and the 'crisis of rationality'. *New Left Review*. **207**: 72–86

Meehan J 1995 Introduction. In: Meehan J (ed.) *Feminists Read Habermas*: *Gendering the Subject of Discourse*. Routledge, New York

Rasmussen D M 1990 *Rereading Habermas*. Blackwell, Oxford, UK

Ray L J 1993 *Rethinking Critical Theory*. Sage, London

Soper K 1989 Feminism as critique. *New Left Review*. **176**: 91–112

Thompson J 1982 Universal pragmatics. In: Thompson J B, Held D (eds.) *Habermas—Critical Debates*. Macmillan, London, pp. 116–33

L. Ray

Critical Theory: Frankfurt School

Critical theory has a narrow and a broad meaning in the history of the social sciences. 'Critical theory' in the narrow sense designates several generations of theorists in the Western European Marxist tradition known as the Frankfurt School. According to these theorists, a 'critical' theory may be distinguished from a 'traditional' theory according to a specific practical

purpose: a theory is critical to the extent that it seeks 'to liberate human beings from the circumstances that enslave them' (Horkheimer 1982, p. 244). Because such theories aim to explain and transform *all* the circumstances that enslave humans, many 'critical theories' in the broader sense have been developed. They emerge typically in connection with social movements that identify varied dimensions of the domination and oppression of humans in modern societies. In both the broad and narrow senses, however, a critical theory is a form of social inquiry with a particular practical and normative intent.

This critical and practical orientation gives rise to three questions about critical social inquiry. First, is critical theory a distinctive form of social inquiry? Second, what sort of knowledge does such inquiry provide in order to provide insight into social circumstances and justify social criticism? Finally, what sort of verification does critical inquiry require? The debates about theoretical uniqueness, the epistemic superiority of the critics, and the practical verification of criticism have produced different forms of critical inquiry with different political goals.

1. *What is a Critical Theory?*

Although myriad in its forms, 'critical theory' (or more generally critical social inquiry) has a distinctive purpose and overall structure. Critical social theorists generally aim at constructing social theories that link explanation and criticism and thus have both normative and explanatory features. Furthermore, such theories must also be 'practical.' 'Practical' here does not simply mean useful or that critical theories are connected to practice generally, but rather to a particular purpose of human emancipation (Horkheimer 1982, Habermas 1971). The philosophical problem of critical social inquiry is to identify precisely those features of its distinctive explanations that give them their distinctive normative character in underwriting social criticism. A closer examination of paradigmatic works across the whole tradition from Marx's *Capital* (1871) to the Frankfurt School's *Studies in Authority and the Family* (1939) and Habermas's *Theory of Communicative Action* (1984, 1987) reveals neither some distinctive form of explanation or a special methodology that provides the necessary and sufficient conditions for such inquiry. Rather, the best such works employ a variety of methods and styles of explanation and are often interdisciplinary in their mode of research (Horkheimer 1993). What then makes them works of critical social science?

There are two competing answers to this question. The first is that they employ a distinctive comprehensive theory that unifies such diverse approaches and explanations and underwrites the epistemic authority of the critic. The second is practical. According to this view, such theories are distinguished by the form of politics in which they are embedded and which also provides the method for practical verification. Even if critical social science is best unified practically and politically rather than theoretically or epistemically, it is not thereby reducible to democratic politics. It is rather the moment of social inquiry from within the normative perspective of a participant within such practices.

As evidenced by the turn to historical materialism in Marx and its constant reconstruction thereafter by critical theorists, the practical solution has not been the favored one in the history of critical social science. The theoretical solution provides the critic with an epistemically superior status over and above the limits of the participants' perspective; it underwrites the claims that such criticisms are 'scientifically justified.' It also provides the basis for going beyond a 'mere' pluralism and for adjudicating among the contradictory claims of theories and explanations in the social sciences; those explanations that cannot be organized into the unity of the comprehensive social theory are rejected by theorists such as Marx as inadequate to justify criticism. This theoretical conception is not devoid of politics. Rather, corresponding to this single comprehensive theoretical framework is the comprehensive political goal of human emancipation. If we reject such theoretical and political convergence, where does critical social science begin amidst such theoretical and political pluralism? Such pluralism requires that we abandon what Marx saw as the criterion justifying the superiority of historical materialism: a unique fit between critical explanation and the goals of a particular political practice.

The alternative practical account links critical inquiry to cooperative and reflexive practices. Indeed, any cooperative practice demands ongoing inquiry into the social conditions and normative basis of its activities. As practices of inquiry embedded in other practices, all the social sciences establish specific relationships between inquirers and other social actors. Habermas and Marcuse consider such social inquiry that only develops strategic and instrumental solutions to be 'technocratic' (Habermas 1970, Marcuse 1964). It models the social scientist on the engineer, who masterfully chooses the optimal solution to a problem of design according to criteria of success as set by independent standards of rationality. Only the practical stance of the reflective participant is fully dialogical, giving inquirer and agent equal standing as agents in a shared practice (whatever the goals of that practice may be). It is this political relationship of equal agency that critical inquirers seek to establish, as they address their criticism to actors who will engage in public reflection on the nature and goals of the practices in which inquirer and actor both participate. As in Kuhn's distinction between normal and revolutionary science, second-order critical reflection considers whether or not the framework for coop-

eration itself needs to be changed. Such criticism is directed at the framework of current institutions as well as towards formulating new terms of political cooperation under which problems are solved.

2. *The Status of the Critic: Theoretical and Methodological Pluralism*

An examination of the history of critical theory shows that the shift away from the distinctiveness of some comprehensive theory to a more practical orientation leads to a pluralistic account of social inquiry. This shift has three phases. The first phase is the Frankfurt School's 1930s program of an 'interdisciplinary historical materialism.' This phase is characterized by a pluralism of social scientific methods within a unified theory of historical materialism (Horkheimer 1993) and this synthesis proved unstable; the Frankfurt School gradually abandoned historical materialism (Horkheimer and Adorno 1972, Wiggershaus 1994). As the leading theorist of the second generation, Habermas constructed a theory of communicative action for explanatory and normative purposes (Habermas 1984, 1987). While comprehensive, this theory actually unifies many different theories: a theory of rationality, of modernization as social and cultural rationalization, and so on. The final phase is the recognition by Habermas of the full implication of both methodological and theoretical pluralism. This thoroughgoing critical pluralism shifts normative weight to the role of the critic in the pluralist practice of democracy, and the public sphere as a social location in which social criticism can take place and have emancipatory effects (Habermas 1989).

The justification of this two-sided (pluralistic, yet comprehensive) approach is worked out in more detail in Habermas's writings on the philosophy of social science and is the common thread that connects *On the Logic of the Social Sciences* (1988) to *The Theory of Communicative Action* (1984, 1987). The problem of pluralism is not new in the philosophy of social science. As Habermas sees it, Max Weber was among the first to recognize that the social sciences combine various cognitive operations: explanation as well as interpretation, historical comparison as well as transhistorical theoretical terms. This mixed status leads social scientists to combine seemingly contradictory and heterogeneous methods, aims, and theories into more or less coherent wholes. Just as in the analysis of forms of inquiry tied to distinct knowledge-constitutive interests (Habermas 1971), Habermas accepts that various theories and methods each have 'a relative legitimacy.' The most fruitful approach to social scientific knowledge is to bring all the various methods and theories into relation to each other: 'Whereas the natural and the cultural or hermeneutic sciences are capable of living in mutually indifferent, albeit more hostile than peaceful coexistence, the social sciences must bear the tension of divergent approaches under one roof …' (Habermas 1988, p. 3). In *The Theory of Communicative Action*, Habermas casts critical social theory in a similar pluralistic, yet unifying way. In discussing various accounts of societal modernization, critical theory takes on the task of critically unifying the various theories and their heterogeneous methods. 'Critical theory does not relate to established lines of research as a competitor; starting from its concept of the rise of modern societies, it attempts to explain the specific limitations and relative rights of those approaches' (Habermas 1987, p. 375).

Such a pluralist approach need not deny the possibility of fruitful large-scale theories, so long as their success is established by a wide range of empirical evidence and by the diversity of sub theories that they fruitfully integrate (Bohman 1991). Indeed, the mixed status of the social sciences as combining explanation and interpretation indicates that elaborating such combinations and connections is the proper goal of social theory. It suggests a theoretically modest, but practically robust pragmatism about critical inquiry. First of all, comprehensiveness does not ensure explanatory power. The second problem is that a close examination of standard critical explanations, such as the theory of ideology, shows that they typically appeal to many different social theories and methods at many different levels (Bohman 1991). Inquiry located in democratic practice suggests a different model: that criticism motivated by whatever theory or method must be verified by those participating in the practice and that this demand for practical verification is part of the process of inquiry itself.

Such a form of inquiry does not need to deny that there is an important role for general theories as 'general interpretive frameworks' in light of which it is possible to construct 'critical histories of the present' (McCarthy and Hoy 1994, pp. 229–30). Such interpretations are validated by the extent to which they open up new possibilities of action that are themselves to be verified in public debate and democratic inquiry. General theories are then best seen as practical proposals whose critical purchase is not moral and epistemic independence but practical and public testing according to criteria of interpretive adequacy.

3. *The Practical Verification of Social Criticism*

Merely to identify a number of different methods and a number of different theories connected with a variety of different purposes and interests leaves the pluralist critical theorist in a rather hopeless epistemological dilemma. Either the choice among theories, methods, and interests seems utterly arbitrary, or the critical theorist has some special epistemic claim to survey the

domain and make the proper choice for the right reason. The former, more skeptical horn of the dilemma is one endorsed by 'new pragmatists' like Rorty (1991) who sees all knowledge as purpose-relative, and by Weber in more decisionistic moments in his methodological writings (1949). The latter demands objectivist claims for social science generally, and for the epistemic superiority of the critical theorist in particular, that Habermas and other critical theorists have been at pains to reject (Habermas 1973, p. 38).

The problems of a critical pluralism cannot be avoided simply by shifting the debate to practical criteria. Not only are there many distinct practical interests or purposes, the social sciences may also be practical in many different senses. To the extent that we can identify the epistemological basis of the social sciences in forms of practical knowledge, the dilemma of pluralism becomes more tractable as a problem: identifying the type of knowledge required by the specific social context of inquiry. The analysis of interests takes us some of the way toward a solution (Habermas 1971). Technical knowledge does not only represent a particular interest, say the interest in controlling outcomes; in the case of the social sciences, it also presupposes a particular practical relation between the social scientists and the subjects of their inquiry. In order to gain democratic control over certain social processes, a particular context of inquiry must be created: emancipation as the regulative control of institutions over certain kinds of practices or processes. As second-order reflection, critical social inquiry addresses the subjects of inquiry as equal reflective participants, as knowledgeable social agents. In this way, the asymmetries of the context of technical control are suspended; this means that critical social inquiry must be judged by a different set of practical consequences, appealing to increasing the 'reflective knowledge' that agents already possess to a greater or lesser degree. As themselves agents in the social world, social scientists participate in the creation of the sort of contexts in which their theories are publicly verified or rejected as critical proposals.

The goal of critical inquiry is not to control social processes or even to influence the sorts of decisions that agents might make in any determinate sort of way. Instead, its goal is to initiate public processes of self-reflection (Habermas 1973). Thus, critical social science can measure its success against the standard of attaining such a practical goal. In most cases, the goal may be attained only through a functioning public sphere; hence, the success of critical social inquiry may be measured by its practical consequences for the quality of discussion and debate in the public sphere. Such a standard is the same measure of success for the contributions of any reflective participant in the public sphere. But it may not always be the case that a well-functioning public sphere exists; and even if one exists it may be difficult to initiate reflection on various social themes or self-reflection on aspects of the public sphere. Such cases are the subject matter of most standard critical theories, such as the theory of ideology, leaving the critic once again in a position of epistemic superiority. This problem, too, is less intractable from a pragmatic point of view: no such strong claims are required, since the critic's theory does not provide the practical warrant independent of a public process of verification and reflection.

On the practical view of critical theory, social scientific knowledge helps agents to see their circumstances differently, especially at that point when mounting problems indicate that some change is practically necessary. Habermas argues that participant-critics cannot employ reflexive theories in an 'objectivating fashion.' The fact that agents do not now accept some proffered interpretation of their practices does not refute such a criticism; the practical conception of verification allows that agents could be mistaken about the character of their practices. Just as technical knowledge depends on creating the social conditions of a community based on free and open critical examination in order to validate its predictions, so too critical inquiry aims at creating the appropriate social conditions under which agents themselves may verify or falsify the criticism offered. Agents in this new social context can retrospectively judge their past knowledge as reflectively inadequate; but apart from undertaking this process of practical verification that involves taking into account many different perspectives on problematic situations, there is no way of knowing the truth or falsity of a criticism as a 'fact of the matter.' Rather, a criticism is verified if it is acceptable to those to whom it is addressed under the appropriate reflective conditions in a public process of judgment and validation.

A further advantage of the practical account is that it makes it easier to see why there are many different critical theories in different historical contexts. When criticisms are justified solely on the basis of the superiority of theoretical knowledge, the critic has no foothold in the social world and no way to choose among the many competing approaches and methods. On a practical account, critical inquiry aims at creating the reflective conditions necessary for its own practical verification. In practical verification, agents may not in the end find these insights acceptable and thus may not change their second-order self-understandings. The publicity of such a process of practical verification entails its own particular standards of critical success or failure. This sort of critical success also helps explain the political pluralism of the many social movements, all of which employ critical social theories and methods as the moment of inquiry of their democratic politics. As new forms of critical theory emerge related to racism, sexism, and colonialism, reflective social agents have transformed these same democratic ideals and practices in the interest of emancipation.

See also: Adorno, Theodor W (1903–69); Authoritarianism; Authority, Social Theories of; Communication: Philosophical Aspects; Critical Psychology; Critical Race Theory; Critical Theory: Contemporary; Hermeneutics, Including Critical Theory; Marx, Karl (1818–89); Marxism and Law; Marxism in Contemporary Sociology; Marxist Social Thought, History of; Pragmatist Social Thought, History of

Bibliography

Bohman J 1991 *New Philosophy of Social Science: Problems of Indeterminacy*. MIT Press, Cambridge, MA

Bohman J 1996 *Public Deliberation: Pluralism, Complexity and Democracy*. MIT Press, Cambridge, MA

Habermas J 1970 *Towards a Rational Society*. Beacon Press, Boston

Habermas J 1971 *Knowledge and Human Interests*. Beacon Press, Boston

Habermas J 1973 *Theory and Practice*. Beacon Press, Boston

Habermas J 1984 *The Theory of Communicative Action*. Beacon Press, Boston, Vol. 1

Habermas J 1987 *The Theory of Communicative Action*. Beacon Press, Boston, Vol. 2

Habermas J 1988 *The Logic of the Social Sciences*. MIT Press, Cambridge, MA

Habermas J 1989 *Structural Transformation of the Public Sphere*. MIT Press, Cambridge, MA

Horkheimer M 1982 *Critical Theory*. Contininum, New York

Horkheimer M 1993 *Between Philosophy and Social Science*. MIT Press, Cambridge, MA

Horkheimer M, Adorno T W 1972 *Dialectic of Enlightenment*. Seabury, New York

Marcuse H 1964 *One Dimensional Man*. Beacon Press, Boston

McCarthy T, Hoy D 1994 *Critical Theory*. Blackwell, London

Rorty R 1991 Inquiry as recontextualization. In: Hiley D, Bohman J, Shusterman R (eds.) *The Interpretive Turn*. Cornell University Press, Ithaca, NY

Weber M 1949 *The Methodology of the Social Sciences*. Free Press, Glencoe , IL

Wiggershaus R 1994 *The Frankfurt School*. MIT Press, Cambridge, MA

J. Bohman

Critical Thinking, Cognitive Psychology of

Critical thinking is often used as a synonym for good or clear thinking. When viewed from the perspective of cognitive psychology, critical thinking is the use of those skills and strategies that increase the probability of a desirable outcome, where the definition of what is desirable depends on individual goals and values. Thus, it is generally acknowledged that the quality of one's thinking is evaluated within a value system. The content of critical thinking courses in cognitive psychology is embedded in the theory and empirical research of the discipline. It includes instruction in scientific thinking, formal and informal logic, thinking probabilistically, evaluating the quality of information, generating and selecting alternatives and goals, and analyzing arguments for the soundness of conclusions. The use of the word 'critical' represents the notion of a critique or evaluation of the processes used in thinking and the outcomes of the thinking process. It is not meant to be negative (as in a 'critical person'), but is used in the sense of a reflective review or informational feedback. The cognitive psychology of critical thinking has a long history. In 1909, John Dewey used the term 'reflective thinking' in a discussion of good thinking that was surprisingly similar to the definitions of critical thinking being used by cognitive psychologists almost a century later.

1. The Critical Need for Critical Thinking

Changes in necessary workplace skills and the increasing complexity of a technology-mediated society have created an increase in the demand for citizens with the abilities to think effectively. It is in this context that critical thinking has emerged as a standard course in many college curricula, often offered as a 'general education' requirement for every college student. Despite the popularity of critical thinking courses, the faculty and administrators who teach the courses and organize knowledge into academic disciplinary categories often disagree over the content and objectives of these courses and the transferability of critical thinking skills and knowledge within and across domains of knowledge.

1.1 The Politics of Critical Thinking

Unlike many other topics in cognitive psychology, critical thinking has gained widespread popularity as, arguably, the most important subject matter and cognitive skill needed for effective citizenship and economic growth in the third millennium. The development of critical thinking skills was identified as one of the primary goals of Education 2000, a national agenda for education in the United States, with support from both Republicans (e.g., President George Bush) and Democrats (President Bill Clinton) (National Education Goals Panel 1991). Unfortunately, and despite bipartisan political support, the goal to improve the critical thinking skills of educated adults was never funded. The failure to fund critical thinking as a national priority touched off a (somewhat cynical) debate over the question of whether politicians really wanted an electorate that was capable of

critical thinking. There are political implications to the development of an academic discipline that purports to improve adult-level thinking skills. Future interaction between academics concerned with the cognitive psychology of critical thinking and politicians concerned with the politics of critical thinking could have profound effects on the development of this academic discipline.

2. *Interdisciplinary Perspectives*

No single discipline can claim sole ownership of or be the sole authority on 'thinking.' For cognitive psychologists, the study of thinking originated in the German laboratory of Wilhelm Wundt, the generally acknowledged 'father of psychology.' Modern cognitive approaches to the study of thinking have generally followed two different approaches—descriptive and prescriptive. The descriptive approach has catalogued common fallacies, such as the reliance on small samples and confusion of correlation and cause, and biases, such as the preference for evidence that confirms a favored hypothesis. The descriptive approach to critical thinking has focused on the ways in which 'thinking goes wrong,' and instruction using this framework usually involves admonishments to prevent fallacious thinking. By contrast, the prescriptive approach focuses on principles for good thinking, such as the appropriate use of base rate information and scientific principles of evidence. When the prescriptive approach is used, there is less emphasis on the pragmatics of thinking, such as the way familiarity and abstractness influence how people think.

The philosophical study of reasoning predates the ancient Greeks, so not surprisingly, critical thinking courses are sometimes offered in departments of philosophy. Content in these courses tends be heavily weighted with the principles of formal and informal reasoning and their application. Education, English, and communication departments also offer courses under the critical thinking rubric, emphasizing skills of persuasion in writing and oral formats. Despite these differences in emphases and historical origins, the central concern in all courses on critical thinking is the enhancement of thinking skills, with the general goal that students will be better thinkers because of the course content.

3. *Evidence that Thinking Can Be Improved With Instruction*

When critical thinking is taught within the framework of cognitive psychology, the principles of scientific reasoning are a primary component of the course. These principles require an empirical basis for sound conclusions. Accordingly, critical thinking instruction needs to be evaluated by its own criteria. Despite considerable skepticism, there are numerous studies that have shown that thinking can be improved when students learn to recognize and apply appropriately the skills and strategies of critical thinking. A general list of skills that are useful across contexts would include understanding how cause is determined, recognizing and criticizing assumptions, analyzing means–goals relationships, and recognizing problems. The evidence for the beneficial effects of critical thinking instruction shows that specific instruction designed for transfer to novel contexts can lead to better thinking in a variety of settings. There are numerous empirical results that support this conclusion: evaluations of students' thinking and writing about novel topics conducted by evaluators who did not know which of the students had received instruction in critical thinking (i.e., 'blind' evaluators); changes in the organization of knowledge so that it comes to resemble the structure of expert knowledge; and gains in adult cognitive development (see Halpern 1996, 1998 for a review of the literature).

4. *A Network Model of Thought and Knowledge*

The cognitive approach to critical thinking uses basic concepts in cognitive psychology as a way of organizing content and designing instruction to improve thinking. It is built upon a model of thought and knowledge in which meaning is represented in memory as a network of interconnected semantic concepts called 'knowledge structures.' The internal representation of meaning in memory determines the way information is used and the circumstances in which some memories are retrieved and others are not, depending on context cues. Meaning is embedded in the interconnections among knowledge structures, which get their meaning(s) from the other knowledge structures to which they are connected. The network model of knowledge representation is an analogy for the patterns of neurochemical activity or biological bases of thought. It is an 'as if' statement for the biological underpinnings of cognition. The network model is a useful way of thinking about critical thinking: a faulty thinking pattern, such as confusing correlation with causation, is represented by connections between the knowledge representation for 'co-occurrence' of two events and the knowledge representation for 'causation.' In correcting this faulty thinking path, the connection needs to be weakened or pruned and replaced with connections between concepts like 'random assignment to conditions' and 'causation.' The faulty connection is particularly strong or salient when the co-occurrence and causation link is consistent with an individual's other beliefs about the cause of a particular outcome (e.g., marijuana use causes the use of 'harder' drugs).

As critical thinking skills are developed, new connections are added to the network and additional knowledge structures are formed and new retrieval

cues effectively access stored memories. The network model of thought and knowledge is dynamic in that individuals are continually modifying their knowledge structures and remaking connections among knowledge structures. It also posits the learner/thinker as an active seeker of information who applies thinking skills or strategies when 'thinking through' the maze of available information stored in memory. Individual differences in one's willingness to engage in the effortful cognitive 'work' of critical thinking and in the individual nature of stored knowledge are also an important part of these models.

5. *Four Components of Critical Thinking*

A four-part model of critical thinking based on the network model of thought and knowledge has been proposed. It includes: (a) the dispositions for the effortful thinking and learning needed for critical thought; (b) knowledge of critical thinking skills; (c) structure training to facilitate cross-contextual transfer; and (d) metacognitive monitoring.

5.1 *Dispositional or Attitudinal Component*

Critical thinking requires an attitude or disposition to engage in effortful, active, and prolonged mental effort because many of the skills of critical thinking require extended thought, which includes a concern for accuracy and attention to details. The dispositional or attitudinal component blurs the distinction between cognitive psychology and topics in emotion and motivation, for example, antidogmatism, the willingness to evaluate information fairly, and the insight to distinguish between rationalization and reasoning. These noncognitive aspects of a critical thinking disposition include the ability and willingness to see a problem from multiple perspectives and a tolerance for ambiguity and uncertainty. These components of critical thinking overlap with emotional and affective topics in psychology; cognitive and emotional processes operate in concert, exerting mutual influences on each other. Instruction in dispositional aspects of critical thinking includes an emphasis on planning, suppression of impulsivity, and the use of situational and social context variables in directing thought and action.

5.2 *Skills Component*

Critical thinking skills are sometimes referred to as 'higher order skills' to differentiate them from 'simpler' (i.e., lower order) skills, such as rote memorization or routinization. Critical thinking skills require judgment, reflection, analysis, synthesis, and attention to context. The complexity in thinking critically often comes from the multidimensional nature of the problem or the need to make a decision or solve a problem when the information that is available is incomplete, probabilistic, or not completely credible. A short taxonomy of critical thinking skills includes the verbal reasoning skills needed to defend against common persuasive techniques, argument analysis skills that include determining the strength of conclusions based on reasons and evidence, skills used in scientific thinking, the use of likelihood and uncertainty to make probabilistic decisions, and the ability to generate multiple alternatives and goals in problem solving and decision making situations. The focus is on component skills that can be specified, identified, practiced, and used spontaneously across domains of knowledge. Although there are several different lists of necessary skills for critical thinking, there are commonalities among them. Fisher and Scriven (1997), for example, list reasoning, problem solving, decision making, communication, evaluation, explanation, analysis and synthesis, metacognition, and formal and informal reasoning as examples of such skills. Most texts target similar skills, with differences among them occurring mostly in their emphases and in the perspectives that they take on these topics.

5.3 *Structure Training to Promote Transcontextual Transfer*

The skills of critical thinking and the dispositions to use them are not sufficient for successful critical thinking, unless these skills are used appropriately in novel contexts. The general question of transfer of skills within and across domains of knowledge has been a knotty problem for cognitive psychology and education. There is ample evidence that the transfer of critical thinking skills is not a necessary or automatic outcome of thinking skills instruction within a single domain of knowledge or traditional content-based courses (Nisbett 1993 provides ample evidence for this conclusion).

Critical thinking skills need to be deliberately taught in ways that enhance transfer. The probability of transfer can be increased with the use of examples from multiple contexts and by making the underlying structure of problems more salient so that problem structure can act as a retrieval cue that triggers memories for the appropriate skills. The structure sensitivity of thinking, especially as a problem solving strategy, was made clear by Hummel and Holyoak (1997) who believe that, 'First thinking is structure sensitive. Reasoning, problem solving, and learning … depend on a capacity to code and manipulate relational knowledge' (p. 427). Relational structures among meaningful concepts are primary in cognitive models of comprehension and thinking. Structure

training is based on models of human cognition in which meaning is conceptualized as being embedded in the relationships among concepts stored in memory. In order to trigger the 'reminder' that a particular skill is needed, the structure of the problem, not its topical information, needs to serve as a retrieval cue. The goal of structure training is to learn to recognize when particular skills may be useful (e.g., regression to the mean or the need to consider multiple alternatives) regardless of the content of the problem.

5.4 Metacognitive Monitoring

The executive or control function of critical thinking is 'metacognition,' which refers to knowledge about one's own cognitive abilities and the use of this knowledge as a guide for thinking and learning. It includes decisions about the allocation of time and resources to a particular problem and the type and extent of effort expended for the cognitive task. Good metacognitive knowledge will include an assessment of what is known and what needs to be known and a determination of goals and a plan for achieving them. Critical thinkers also know when they do not have enough information to reach a meaningful conclusion as well as the difference between a probable outcome, an unlikely outcome, and a certain outcome. Metacognitive monitoring is a conscious consideration of one's cognitive abilities and an accurate assessment of cognitive resources, hence it serves as the boss or 'executive function.'

Metacognition directs three activities that are critical to the execution of critical thinking: monitoring, controlling, and knowing. Monitoring is important during the acquisition of knowledge—it assesses how well new material is being learned and how completely a difficult concept is being understood. The controlling function is important because it provides direction for cognitive effort—if the learning or understanding is not going well, then additional attention and effort needs to be directed to these processes. The knowing function is used to make judgments about one's confidence in his or her confidence in whether a solution is satisfactory. Thus, metacognition underlies multiple, successive processes that are essential to critical thinking.

6. Creative and Critical Thought

Creative thinking is the production of unusual and good responses to problems. This definition of creative thinking is compatible with the definition of critical thinking being used in this article, differing only in that creative thinking requires the use of unusual strategies or skills. By comparison, the thinking skills used in critical thinking could be either common or unusual. Thus, critical thinking is a superordinate of creative thinking. Critical thinking requires creativity (unusual and good solutions) in many aspects—generating alternatives to problems, redefining goals, and recognizing which critical thinking skills are needed in novel situations. There is no clear demarcation between critical and creative thinking because it is impossible to determine where one type of thinking ends and the other begins; commonly used thinking skills and unusual ones lie along a single continuum.

In an analysis of creative thinking, which includes several case studies, Weisberg (1993) traced the origins of creative thought to experiences that generalized across domains of knowledge. He acknowledged that far or remote transfer of thinking skills is unusual, which is the defining attribute of creative thinking—an unusual solution or insight. It is for this reason that structure training designed to improve transferability of thinking skills across contexts is a critical component to instruction designed to improve critical thinking. When these skills are applied in a novel domain of knowledge, their use in that domain may be unusual, or, in other words, creative. Thus, heuristics that encourage novel ways of making associations and the use of examples from many different contexts are an important part of critical and creative thinking courses.

7. The Role of Context and Judgment

Critical thinking instruction recognizes the need for different sorts of skills, depending on the context, and the ability to recognize that a thinking skill that was used in one context (e.g., determining the effectiveness of a drug) is useful in a very different context (e.g., deciding if trade embargoes will improve political relations with a foreign country). In this way, instruction in critical thinking differs from instruction in formal logic, with its heavy use of syllogisms and linear orderings, which are fully rule dependent and do not vary with context. Critical thinking is more complex than rule-driven deductive logic. To be ecologically valid, critical thinking would require judgment, usually under conditions of uncertainty and ambiguity, two conditions that are common in most real-life situations. Critical thinking is sometimes referred at as 'authentic thinking' because it incorporates some of the 'messiness' of real-life problems. Practice examples often come from a variety of real-life scenarios, such as the arguments presented in newspaper editorials, decisions taken from real life such as which medical procedure to select from a set of imperfect alternatives, and solving problems like making one's money last until the next pay check. Some of the antagonism between critical thinking as cognitive psychology and critical thinking as philosophy is over the extent to which each discipline prefers rule-based vs. heuristic-based (probabilistic) approaches to the subject matter.

8. Critical Thinking as a Formal Discipline

The principles of cognitive psychology have provided researchers and educators with a useful model for the enhancement of critical thinking. There is a considerable body of evidence showing that better thinking can be an outcome of education when critical thinking is explicitly taught for transfer across contexts, when the disposition to think critically is enhanced, and when metacognitive monitoring is used as a guide to the thinking process. Critical thinking is cognitive psychology's finest offspring. It offers the best hope for the future because the ability to think well is the best preparation for a future that is rapidly changing and becoming increasingly complex. Perhaps the disasters that doomsday experts are predicting for the third millennium can be prevented with large-scale efforts to help more people become better thinkers. Critical thinking is emerging as a formal discipline with its own content area and adherents. Its current status as a subtopic under cognitive psychology is likely to change over the coming decades, when it will emerge as a major theoretical perspective on how people acquire and use information more effectively, so that they can become better thinkers and learners.

Bibliography

Dewey J 1909 *How We Think*. Heath, Boston, MA

Fisher A, Scriven M 1997 *Critical Thinking: Its Definition and Assessment*. Edgepress, Point Reyes, CA

Halpern D F 1996 *Thought and Knowledge: An Introduction to Critical Thinking*, 3rd edn. Erlbaum, Mahwah, NJ

Halpern D F 1998 Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist* **53**: 449–55

Hummel J E, Holyoak K J 1997 Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* **104**: 427–66

National Education Goals Panel 1991 *The National Education Goals Report*. National Education Goals Panel, Washington, DC

Nisbett R E (ed.) 1993 *Rules for Reasoning*. Erlbaum, Hillsdale, NJ

Weisberg R W 1993 *Creativity: Beyond the Myth of Genius*. Freeman, New York

D. F. Halpern

Cross-cultural Psychology

Cross-cultural psychology is the systematic study of behavior and experience as it occurs in different cultures, is influenced by culture, or results in changes in existing cultures. This subdiscipline of psychology is used by those interested in studying the relationship of culture with psychological processes.

1. Relationship with Cultural and Indigenous Psychology

Cross-cultural psychology (CCP) is different from cultural psychology (CP) (see *Cultural Psychology*) and indigenous psychology (IP) in a number of ways. CCP is closer to psychology and uses a universalistic theoretical framework that assumes the psychic unity of humankind. CP and IP are closer to anthropology and use relativist frameworks and examine how culture and psychology 'make each other up.' Some versions of CP and IP do not assume the psychic unity of humankind. CCP is widely used by social and organizational psychologists, CP is favored by developmental psychologists.

CCP searches for universal (etic) as well as culture specific (emic) findings. It tries to study them using both emic and etic measurements. It emphasizes equivalence of measurement of the constructs in each culture. In addition it aims at culture-sensitive measurement in each culture by making use of both emic and etic items (questions, stimuli) in each culture that correspond to etic constructs. Only etic items can be used to compare cultures. However, both emic and etic items can be used to describe cultures.

CCP uses methods that include little context (e.g., questionnaires, projective tests, experiments), while CP and IP use methods that emphasize context (e.g., participant observations in real life situations, content analyses).

CCP conceives of culture as outside the person, i.e., as a context for perception and action that influences universal psychological processes. For example, the probability of an act is a function of the frequency of the reinforcement of the act. But in some collectivist cultures (see *Collectivism: Cultural Concerns*) the reinforcement may be given to a valued member of the ingroup rather than to the individual whose action is being studied. In individualistic cultures (Triandis 1995), the reinforcement must be given to the person under study, otherwise it has little impact. Thus, in the universal process the frequency of reinforcement is linked to the probability of an act; in the culture-specific process that operates in some collectivist cultures, one need not reinforce the person who is acting, but can obtain the same results by reinforcing a significant other of that person.

Cultures have many elements. Some elements change slowly, perhaps after a few generations, and others change very quickly, such as particular ways of using words. CCP focuses on the more stable elements; CP and IP focus mainly on the changeable elements, especially on the meanings of key concepts. CCP is favored by social and industrial psychologists who study people who have been members of their cultures for a long time. CCP considers differences in meaning, in different cultural communities, as barriers to research. These barriers must be overcome by using specially developed methods of data collection, e.g.,

local (culture specific) standardization of items. Some people have argued that CCP is a method and does not have specific content.

CCP is interested in individual differences. CP and IP do not focus on individual differences, and assume that members of a culture are more or less alike. The cultural distance between the researcher and the culture under investigation is very different in the case of these three perspectives. Cultural distance refers to differences in language (e.g., are the two languages members of the same language family or belong to very different families), differences in social structure (e.g., monogamy vs. polygyny), differences in religion, politics, economic conditions, level of literacy, and the like. Psychologists who study IP are usually members of the culture they study. By contrast, psychologists who use the CP perspective study cultures that are very distant from their own culture. Psychologists who take the CCP perspective hold an intermediate position, studying cultures that are somewhat distant but not too different (e.g., managers in Japan and France). Since cross-cultural psychologists use psychological methods, they are limited to studying people with whom such methods can be used. Such methods can be used only with the people who are more or less familiar with them. Standardized tests, experiments, and the like are meaningful only in the cultures in which they were developed or in similar cultures. On the other hand, participant observations or interviews are less problematic in distant cultures. Thus, psychologists who use the IP or CP perspective use these proceedures more frequently, while CCP researchers use them less frequently.

2. *History*

Since antiquity people have been interested in the manners and morals of people different from their own, as well as in their own psychology. Herodotus in the fifth century BC had the insight that all humans are ethnocentric (use their own culture as the standard for the evaluation of other cultures, and evaluate those cultures that are like their own more positively than the cultures that are very different from their own). In the nineteenth century Darwin's theory of evolution stimulated great interest in mental measurements in different parts of the world. Expeditions by Rivers (1901, 1905) supplied the first data. Jahoda (1993) provided an excellent history of the field.

Wilhelm Wundt published a *Voelkerpsychologie* in 10 volumes (1900–14), which summarized findings of anthropologists, linguists, and historians that had relevance for psychology. After the First World War a movement developed that rejected ethnocentrism and stressed the understanding of cultures in their own terms. Notable events in the history of CCP are the publication of Klineberg's (1954) social psychology text, which did include a great deal of cultural information; the publication of the *Cross-Cultural Psychology Newsletter* (Triandis 1967), which became the *Cross-Cultural Psychology Bulletin*; the publication of the first *Directory* of individuals interested in cross-cultural research (Berry 1968); the publication of the *Journal of Cross-Cultural Psychology*(Lonner 1970); the establishment of the International Association for Cross Cultural Psychology (Dawson 1972); and the publication of a code of ethics for cross-cultural research (Tapp et al. 1974). The *Annual Review of Psychology* started chapters reviewing cross-cultural work in 1973, the establishment of the *Association pour la Recherche Interculturelle* (ARIC) occurred in 1984, the publication of the first *Handbook of Cross-Cultural Psychology* (Triandis 1980–1981) in six volumes, and the second edition (Berry et al. 1997) in three volumes, consolidated the field. The avoidance of ethnocentrism by searching for different perspectives on a phenomenon and developing appreciation of emic points of view; the local standardization of psychological instruments; the development of methods that use both emic and etic items; the use of converging multimethod measurements; the testing of rival hypotheses; and the acceptance of testing of hypotheses using data from the Human Relations Area Files as well as data obtained from research participants became important distinguishing features of cross-cultural psychology.

3. *Examples of Significant Work in CCP*

3.1 *Theory and Method*

The desirability of the CCP, IP, and CP perspectives discussed above is one of the topics being debated. In addition, issues such the extent to which anthropological, genetic, evolutionary, and other approaches must be integrated with CCP are subjects of debate. The methodology of the field was presented in Triandis and Berry (1980). Van de Vijver and Leung (1997) summarized newer methodological developments. Berry et al. (1997) reviewed the most recent issues about theory and method in CCP. Amir and Sharon replicated several social psychological studies published in the *Journal of Personality and Social Psychology* in Israel, and found that only about half the findings replicated. Adamopoulos and Lonner identified numerous cultural universals.

Irvine did much work on psychological testing across cultures, showing that elaborate pretests, parametric studies, and careful analyses are needed if data from different cultures are to be compared meaningfully. On the subject of methodology readers may wish to study the work of the following individuals in particular: John Berry, Michael Bond, Don Campbell, Kwok Leung, Roy Malpass, Raoul Naroll, Fred Strodtbeck, Ype Poortinga, Harry Triandis, and Fons

Van de Vijver. Each method has its own meaning in each culture, so that multimethod measurements and convergence of findings across methods are essential. Local standardization of the measurements (Triandis 1992) is most important. Checking if the cultural differences that have been identified are due to culture rather than to some other factors that happen to be associated with culture, requires the testing of many rival hypotheses.

Berry (1976) developed an ecological framework that linked aspects of the ecology, such as whether people make a living by hunting or by agriculture, to attributes of cultures and attributes of culture to psychological adaptation. People in stable, agricultural cultures conform more than people in hunting cultures.

3.2 Basic Processes

Berry et al. (1997) reviewed the basic and developmental processes using both the CP and the CCP perspective. Specifically, Segall, Campbell and Herskovits (1966) showed that the perception of visual illusions is influenced by culture. For instance, people in carpentered environments (with many right angles), such as those found in industrial cultures, are quite susceptible to the Mueller–Lyer illusion while people in environments where most buildings are round, as in Africa, are not so susceptible. Exposure to many right angles during socialization has the effect of incorrectly seeing the ends of the lines in that illusion as representing right angles.

The relationship of language and thought was investigated by Hunt and Agnoli, M. Cole, Gay, Glick, Scribner, Sharp, and others investigated the effects of literacy on cognition. Literacy was shown to be a major factor in shaping cognition because the unschooled have difficulty making abstract judgments. Extensive studies of different educational systems, such as those of East Asia, the Qur'anic schools, and comparisons with Western education have shed light on the advantages and disadvantages of each educational system. Studies by Stevenson and his associates have shown that the mathematical performance of East Asian students is far superior to the performance of American students, yet American parents are quite satisfied with their schools, while parents in East Asia are dissatisfied with their schools.

Ekman, Frijda, Kitayama and Markus, and Matsumoto, among others, linked culture and emotion. It appears that some aspects of emotions are universal and other aspects are culture specific. In most cultures individuals are emotionally sensitive to certain contingencies (e.g., danger of rejection by the group) and respond in similar ways (e.g., crying, or hostile behavior). But there are culture-specific norms specifying what emotions may be displayed. In other words, there is evidence that the antecedents of emotion have similarities across diverse cultures, but there is also evidence that no emotion term has an exact correspondence to any other emotion term, even when the languages are very similar, such as German and Swiss-German. Although basic emotional processes appear to have universal components, emotional expression is definitely controlled by culture-specific norms.

Osgood, May, and Miron investigated universals of affective meaning. The meaning of any concept can be conceived as the projection of the concept on the dimensions of evaluation, potency, and activity, emically operationalized in each culture. The study was carried out in 27 diverse cultures. An *Atlas* of affective meaning was developed that consists of 600 concepts judged on evaluation, potency, and activity in the 27 cultures. The information is archival, and has been used by many researchers for different purposes. It can be obtained from the Osgood Center of the Department of Psychology, University of Indiana/Purdue, in Indianapolis, Indiana, USA.

3.3 Developmental Psychology

Most developmental psychologists prefer the CP perspective (see *Cultural Psychology*), so here only a few examples that use the CCP perspective will be mentioned. John Whiting, Beatrice Whiting, Irving Child, William W. Lambert, and Leigh Minturn examined the settings within which child development takes place. For example, the presence or absence of grandparents constitutes a difference in setting. Observations and interviews converged in showing that nuclear vs. extended household type was associated with the contrast between authoritarian aggressive and social-intimate child behavior. Ron Rohner used three kinds of data: questionnaire responses within culture, ethnographies of different cultures, and the Human Relations Area Files (summaries of available ethnographies classified by topic). All three data sets indicated that parents who are warm, supportive, and cuddling have children who are optimistic and well adjusted; in contrast, parents who are cold and unavailable or rejecting have children who are pessimistic, emotionally unresponsive, hostile, aggressive, and poorly adjusted. Jerome Bruner, Patricia Greenfield, Joan Miller, Barbara Rogoff, and Charles Super showed that cultural differences increase throughout the life span of development.

Eckensberger, Miller, and Shweder have reported studies of moral development in different cultures. For example, Miller showed that in India helping another person is obligatory in more situations than it is in the USA. Kohlberg's stage theory of moral development was tested in many cultures, but the materials were translations with minor adjustments of the dilemmas

used to test it in the West. Critics have emphasized that it is both a Western and an androcentric theory. Eckensberger has developed a method of testing that is more culture sensitive. It uses the core of a dilemma and an interview which allows the participant to reveal his or her thinking about the dilemma. The answers are then scored according to a stage model. The data suggest that there are stages of moral development, but they do take culture-specific forms, especially at the higher levels of moral development.

3.4 Social Processes

Berry et al. (1997) reviewed this topic (see also *Collectivism: Cultural Concerns*). Emiko and Yoshi Kashima found relationships between culture and the use of language (e.g., in collectivist cultures, 'I' is not required in sentence construction), as did Semin and Zwier (e.g., collectivists frequently use action verbs and individualists use state verbs). Giles and Gudykunst investigated cultural differences in communication behaviors. Poyatos (1988) has edited a volume on nonverbal behaviors such as gestures in different cultures.

Bilingualism was investigated by W. E. Lambert (e.g., middle-class English speakers that went to a French-speaking school became fluent in both languages; the same person speaking in French or English is perceived quite differently). Triandis (1972) found differences in the meaning of roles and social behaviors in Greece and the USA; he also investigated the meaning of value terms in Japan, India, Greece, and the USA and found that words that could refer to either a collective or an individual attribute (e.g., 'progress') were perceived as attributes of a collective in collectivist cultures and as individual attributes in individualist cultures.

Doob as well as Levine examined the meaning of time in different cultures. For example, empirical work has shown that there are fast- (Switzerland, Ireland, Germany, Japan), moderate- (Hong Kong, Costa Rica, Taiwan, Singapore, USA), and slow-moving cultures (China, Syria, Indonesia, Brazil, Mexico). Within the USA, Boston, Buffalo, and New York are fast, Nashville, San Diego, and East Lansing moderate, and San Jose, Sacramento, and Los Angeles slow. In large cities, rich, individualist cultures, and cold climates, people tend to be faster than in rural settings, poor, collectivist cultures, and warm climates.

Berry provided a typology for the study of acculturation that has implications for social policy. Social policies might favor the melting pot concept of assimilation, i.e., those who come from culture (A) and come in contact with a dominant culture (B) are expected to drop culture A and only use culture B. A sounder policy for mental health, according to Berry's findings, is integration or biculturalism, i.e., A + B. Other undesirable policies are segregation, i.e., keeping A and rejecting B, and anomie or marginalization, i.e., rejecting both A and B.

Williams and Best showed that gender stereotypes are quite similar in different cultures. Men are stereotyped as strong, aggressive, cruel, coarse, and adventurous; women are stereotyped as weak, appreciative, softhearted, gentle, and meek. They argue that the pancultural similarities greatly outweigh cultural differences. The biological differences between men and women can be amplified or diminished by cultural influences.

Inkeles studied modernity and found that in all cultures there are segments of the population that are modern. The more educated segments of the society, factory workers, and those who read the newspapers respond to attitude questions in a more modern way than those who have not had these experiences.

Brewer and Campbell explored ethnocentrism, social perception, and the similarity–attraction link in a large number of African cultures. All humans are ethnocentric to some degree. In the case of extreme ethnocentrism, people consider the norms of their culture as the only ones that are 'natural' and 'correct'; they see their customs as universally valid, and believe that it is natural to cooperate with ingroup members and feel proud of their ingroup, whereas they are distrustful and even hostile of outgroups. Ethnocentrism can be reduced by exposure to different cultures, especially 'successful' cultures from the point of view of the people under study.

The more similar two African cultures are to each other, the higher is their evaluation of each other. The autostereotypes (how members of a target culture see their own culture) of the groups were correlated with the heterostereotypes (how members of other cultures see the target culture) received by these groups from the other groups. This correlation was higher for cultures that were in frequent contact than for groups that were not in much contact. This suggests that stereotypes become more accurate as contact increases.

Buss linked the desirable attributes of marital partners to culture and showed that the theory of evolution can explain the observed differences. Specifically, women in all cultures are looking for a mate who is reliable and well established; men in all cultures look for a mate who is physically attractive. Since women invest much time in rearing children they look for mates who will provide for them; men use physical attraction as a cue to good health and probable fertility.

Earley, A. Fiske, Hofstede, Kagitcibasi, Markus and Kitayama, U. Kim, and Triandis carried out extensive work on collectivism (see *Collectivism: Cultural Concerns*). It will not be reviewed here. Leung has shown that dispute resolution is different in different cultures. People in collectivist cultures tend to avoid confrontations and seek to resolve disputes by using mediation by respected third persons. Foa, Leung,

and Hui have studied cultural differences in the use of equity (to each according to contribution), equality, or need in the distribution of resources. For example, equity is used when individualists distribute resources or when collectivists distribute resources to outgroup members; equality is used when collectivists distribute resources to ingroup members.

Miller, Morris, and Peng showed that culture influences the way people make attributions. For example, people in individualist cultures are more likely to use internal factors (e.g., ability, attitudes, or personality), whereas people in collectivist cultures are more likely to use external factors (e.g., what other people did, the task, the norms of the group) in explaining social behavior.

Segall, Ember, and Ember examined aggression and reported that aggression is ubiquitous, all societies attempt to control it, but punishment is not effective in controlling it. Attending to fictional violence contributes to aggression. High frequency of warfare is predicted from resource unpredictability and from territorial and ethnic conflict. In societies where there is severe punishment for wrongdoing and family violence, there is more warfare.

4. *Applications*

Draguns and Marsella investigated culture and psychopathology relationships. Culturally invariant and variable components of psychopathology, especially depression and schizophrenia, have been identified. Some symptoms appear only in one or a few cultures. Psychopathology is manifested in different cultures somewhat differently and the dimension of collectivism-individualism seems promising in organizing and integrating the findings. Pedersen and Leong explored the relationship between culture and counseling.

Bhawuk, Brislin, Landis, as well as Triandis, published extensively about cross-cultural training. For example, culture assimilators have been developed to aid in such training. They consist of scenarios that describe the interaction of members of two cultures, the culture of the trainee, and the target culture. The trainee reads the scenarios and picks one of four or five attributions that may explain why members of the target culture acted the way they did in the scenario. After selecting an attribution, the trainee receives feedback that explains why the attribution is correct or incorrect from the perspective of the target culture. This way the trainee learns to make attributions that resemble the attributions that members of the target culture make when they think of the scenarios. Thus, trainees learn to give the same meaning to behaviors that occur in the target culture as members of the target culture do, and that increases the effectiveness of the trainee's work abroad, as well as reducing the culture shock experienced by the trainee. Culture-general assimilators describe events in many cultures. Culture-specific assimilators are limited to a particular culture, and sometimes to specific demographic categories, e.g., a European-American supervisor working with African-American first-time employees.

Adler, Earley, Erez, and Wilpert investigated the effects of culture on organizational behavior. The main topics included differences in motivation, the meaning of work in different cultures, organizational commitment, communication, the effectiveness of ethnically heterogeneous groups, leadership in different cultures, participative management and industrial democracy in different cultures, reward schemes and their effectiveness in different cultures, organizational development, and intercultural negotiations.

5. *Conclusions and Future Direction of Theory and Research*

A conclusion is that almost every aspect of psychological functioning has both universal and culture-specific components. This perspective suggests that CCP provides an important way to study the relationship of culture and psychological processes.

An important issue is the extent CCP, IP, and CP will develop as separate subdisciplines or become integrated into a coherent approach that uses all three types of investigation. Perhaps the future will bring the integration of these approaches.

The theory of convergence holds that cultures are becoming more similar over time. However, most cross-cultural psychologists believe that convergence is occurring in only a few elements of culture, and that divergence remains in the other elements.

See also: Cross-cultural Research Methods; Cross-cultural Study of Education; Cultural Diversity, Human Development, and Education; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Cultural Expression and Action; Cultural Psychology; Cultural Relativism, Anthropology of; Cultural Studies: Cultural Concerns; Culture and Emotion; Industrial and Organizational Psychology: Cross-cultural; Personality Psychology; Personality Structure; Personality Theories

Bibliography

Berry J W 1976 *Cross-Cultural research and Methodology Series: III. Human Ecology and Cognitive Style: Comparative Studies in Cultural and Psychological Adaptation*. Sage, New York

Berry J W, Poortinsa Y H, Pandey J, Dasen P R, Saraswathi T S, Segall M H, Kagitçibasi C 1997 *Handbook of Cross-cultural Psychology*, 2nd edn. Allyn and Bacon, Boston, 3 Vols.

Jahoda G 1993 *Crossroads Between Culture and Mind: Continuities and Change in Theories of Human Nature*. Harvard University Press, Cambridge, MA
Klineberg O 1954 *Social Psychology*, rev. edn. Holt, Rinehart & Winston, New York
Poyatos F (ed.) 1988 *Cross-Cultural Perspectives in Nonverbal Communication*. Hogrefe, Toronto, ON
Rivers W H R 1901 Primitive color vision. *Popular Science Monthly* **59**: 4458
Rivers W H R 1905 Observations on the senses of the Todas. *British Journal of Psychology* **1**: 321–96
Segall M H, Campbell D T, Herskovits M H 1966 *Influence of Culture on Visual Perception*. Bobbs-Merrill, Indianapolis, IN
Tapp J L, Kelman H C, Triandis H C, Wrightsman L, Coelho G 1974 Continuing concerns in cross-cultural ethics. *International Journal of Psychology* **9**: 231–49
Triandis H C 1972 *The Analysis of Subjective Culture*. Wiley Interscience, New York
Triandis H C 1980–81 *Handbook of Cross Cultural Psychology*. Allyn and Bacon, Boston, 6 Vols
Triandis H C 1992 Cross-cultural research in social psychology. In: Granberg D, Sarup G (eds.) *Social Judgment and Intergroup Relations: Essays in Honor of Muzafer Sherif*. Springer Verlag, New York, pp. 229–44
Triandis H C 1995 *Individualism & Collectivism*. Westview Press, Boulder, CO
Triandis H C, Berry J H 1980 *Handbook of Cross-Cultural Psychology: Methodology*. Allyn and Bacon, Boston
Van de Vijver F J R, Leung K 1997 Methods and data analyses of comparative research. In: Berry J W, Poortinga Y H, Pandey J (eds.) *Handbook of Cross-Cultural Psychology*, 2nd edn. Allyn and Bacon, Boston, pp. 257–300

H. C. Triandis

Cross-cultural Research Methods

Cross-cultural studies involve persons from different countries and/or ethnic groups; a defining characteristic is their comparative nature. Most studies employ quantitative methods of data collection and analysis. Studies of cultural topics that are noncomparative and apply a qualitative methodology can be found in sociology ('cultural studies,' e.g., Alasuutari 1995), cultural psychology (Greenfield 1997), and cultural anthropology (Naroll and Cohen 1970) see *Phenomenology in Human Science*; *Ethnomethodology: General*).

The range of instruments used in comparative studies is very broad, ranging from highly standardized psychological tests, to observation schedules, and free interviews. In many studies existing Western instruments (mental tests, survey questionnaires, personality inventories) are administered in a new cultural context, either or not adapted to enhance their cultural appropriateness.

If two persons from different cultural groups show different scores on a reliable and valid measure of subjective well-being, these score differences may refer to individual differences in subjective well-being. However, the score differences may also arise from differential social desirability or some other response style, inappropriate translation, or inadequacy of the item to measure well-being in both groups. The example illustrates a central problem in cross-cultural research: observed score differences are often susceptible to multiple explanations (Campbell and Stanley 1966, Cook and Campbell 1979, Poortinga and Malpass 1986). When the same instrument has been administered to persons from different ethnic groups, it cannot be taken for granted that the same scores obtained in different cultural groups have the same psychological meaning.

The ambiguity of interpretation is a consequence of the methodological nature of culture as an independent variable. In laboratory studies researchers randomly assign subjects to experimental treatments (see *Laboratory Experiment: Methodology*). The random assignment leads to a firm control of ambient variables; ideally an experimental and control group are matched on all outcome-relevant characteristics (e.g., personality characteristics and socioeconomic status), except for the treatment variable studied (see *Internal Validity*). However, like gender and other intrinsic subject characteristics, culture is not an experimental treatment that can be manipulated. Groups with a different cultural background tend to differ on a variety of outcome-relevant characteristics. These differences may constitute rival explanations of observed cross-cultural differences. Without precautions to rule out these rival explanations, observed cross-cultural differences are open to multiple interpretations. Findings in cross-cultural research are more convincing when rival explanations have been more adequately dealt with.

Bias is the generic name of an important family of rival explanations (see *Test Bias*). It refers to the common problem in the assessment of nonequivalent groups that scores obtained in different cultural groups are not an adequate reflection of the groups' standing on the construct underlying the instrument. If scores are biased, their psychological meaning is group dependent and group differences in assessment outcome are to be accounted for, at least to some extent, by auxiliary psychological constructs or measurement artifacts. A closely related concept is equivalence which refers to the absence of bias and hence, to similarity of meaning across groups. The two concepts have somewhat different historical roots and areas of application. Whereas bias usually refers to nuisance factors, equivalence has become the generic term for metrical implications of bias.

Bias and equivalence are not inherent properties of an instrument but arise in a group comparison with a particular instrument. Score comparisons of groups that differ in more test-relevant aspects will show a higher susceptibility to bias.

1. Sources of Bias

There are three bias sources in cross-cultural research. The first is called construct bias; it occurs when the construct measured is not identical across groups. Ho's (1996) work on filial piety (psychological characteristics associated with being a good son or daughter) provides a good example. The Western conceptualization is narrower than the Chinese, according to which children are supposed to assume the role of caretaker of their parents when these grow old and become needy. Construct bias precludes the cross-cultural measurement of a construct with the same measure. An inventory of filial piety based on the Chinese conceptualization will cover aspects unrelated to the concept among Western subjects, while a Western-based inventory will leave an important Chinese aspect uncovered.

An important source of bias, called method bias, can result from sample incomparability, instrument characteristics), tester and interviewer effects, and the method (mode) of administration (see *Test Administration: Methodology*). Examples are differential stimulus familiarity in mental testing and differential social desirability in personality and survey research. Some sources of method bias can be dealt with by careful preparation of the assessment instrument and its instruction manual (e.g., proper test instruction with a clear specification of what is asked from participants, standardization of administration, and adequate training of testers and interviewers). Yet, it may be impossible to eliminate all outcome-relevant sample characteristics, in particular when the cultural distance of the countries involved is large. There are indications that a country's Gross National Product (per capita) is positively related to its mean score on mental tests and negatively to its mean score on social desirability. Particularly in comparisons of culturally highly dissimilar groups it may be hard or even impossible to eliminate the impact of sources of method bias such as sources familiarity and social desirability.

Finally, bias can be due to anomalies at item level (e.g., poor translations); this is called item bias or differential item functioning. According to a definition that is used widely in psychology, an item is biased if persons with the same standing on the underlying construct (e.g., they are equally intelligent) but coming from different cultural groups, do not have the same

Table 1
Strategies for identifying and dealing with bias in cross-cultural research

Type of bias	Strategies
Construct bias	Decentering (i.e., simultaneously developing the same instrument in several cultures) Convergence approach (i.e., independent within-culture development of instruments and subsequent cross-cultural administration of all instruments)
Construct and/or method bias	Use of informants with expertise in local culture and language Use samples of bilingual subjects Use of local surveys (e.g., content analyses of free-response questions) Nonstandard instrument administration (e.g., 'thinking aloud') Cross-cultural comparison of nomological networks (e.g., convergent/discriminant validity studies, monotrait-multimethod studies, connotation of key phrases)
Method bias	Extensive training of adminstrators (e.g., increasing cultural sensitivity) Detailed manual/protocol for administration, scoring, and interpretation Detailed instructions (e.g., with sufficient number of examples and/or exercises) Use of subject and context variables (e.g., educational background) Use of collateral information (e.g., test-taking behavior or test attitudes) Assessment of response styles (e.g., social desirability, acquiesecence) Use of test–retest, training and/or intervention studies
Item bias	Judgmental methods of item bias detection (e.g., linguistic and psychological analysis) Psychometric methods of item bias detection (e.g., Differential Item Functioning analysis) Error or distracter analysis

Source: Van de Vijver and Tanzer 1997 *European Review of Psychological Assessment* (reproduced with the permission of Swets and Zeitlinger)

average score on the item. The score on the construct usually is derived from the total test score. If a geography test administered to pupils in Poland and Japan, contains the item 'What is the capital of Poland?' Polish pupils can be expected to show higher scores on the item than Japanese students, even when pupils with the same total test score would be compared. The item is biased as it favors one cultural group across all test score levels. Of all bias types, item bias has been the most extensively studied; various psychometric techniques are available to identify item bias (e.g., Camilli and Shepard 1994; Holland and Wainer 1993).

An overview of common ways of addressing bias is given in Table 1.

2. *Types of Equivalence*

Elaborating on categorizations in the literature, four different types of equivalence are proposed here (cf. Van de Vijver and Leung 1997). The first type is labeled construct inequivalence. It amounts to comparing apples and oranges (e.g., the comparison of Chinese and Western filial piety, discussed above). Comparisons lack an attribute for comparison, also called tertium comparationis (the third term in the comparison). The second is called structural (or functional) equivalence. An instrument administered in different cultural groups shows structural equivalence if it measures the same construct in both groups (e.g., Raven's Standard Progressive Matrices test has been found to measure intelligence in various cultural groups). Exploratory factor analyses followed by target rotations or confirmatory factor analysis of correlation matrices (structural equation modeling) may be applied to examine structural equivalence. Structural equivalence does not presuppose identity of measures across groups. The measures may use different stimuli across groups. If different operationalizations have been chosen, structural equivalence can be examined by comparing nomological networks across groups.

The third type of equivalence is called measurement unit equivalence. Instruments show this type of equivalence if their measurement scales have the same units of measurement and a different origin (such as the Celsius and Kelvin scales in temperature measurement). This type of equivalence assumes interval- or ratio-level scores (with the same measurement units in each culture). It applies when a bias factor with a fairly uniform influence on the items of an instrument affects test scores of different cultural groups in a differential way. Social desirability and stimulus familiarity may exert this influence. Observed group differences in scores are then a mixture of valid cross-cultural differences and measurement artifacts. When the relative contribution of both sources cannot be estimated, the interpretation of group comparisons of mean scores remains ambiguous. Multigroup comparisons of confirmatory factor analytic models have been utilized to examine measurement unit equivalence. This technique, based on comparisons of covariance matrices, can identify bias sources that affect the covariance of items or tests but it cannot differentiate between valid differences in mean scores and bias sources with a uniform influence on all parts of an instrument.

Only in the case of scalar (or full-score) equivalence can direct comparisons be made; it is the only type of equivalence that allows for the conclusion that average scores obtained in two cultures are different. This type of equivalence assumes the same interval or ratio scales across groups. Conclusions about which of the latter two types of equivalence applies are often difficult to draw and controversial. For example, racial differences in intelligence test scores have been interpreted as due to valid differences (scalar equivalence) and as reflecting measurement artifacts (measurement unit equivalence). Scalar equivalence assumes that the role of bias can be safely neglected. The demonstration of scalar equivalence draws on inductive argumentation. Therefore, it is easier to disprove than to prove scalar equivalence. This can be made plausible by measuring presumably relevant sources of bias (such as stimulus familiarity or social desirability) and showing that they cannot statistically explain observed cross-cultural differences in a multiple regression or covariance analysis.

The distinction between the latter two types of equivalence is immaterial when comparing experimental conditions or changes across cultures (e.g., developmental trajectories or training effects). Scales that show measurement unit equivalence, measure changes at the level of full score equivalence.

Structural, measurement unit, and scalar equivalence are hierarchically ordered. The third presupposes the second, which presupposes the first. Moreover, higher levels of equivalence are more difficult to establish. It is easier to demonstrate that an instrument measures the same construct in different cultural groups (structural equivalence) than to demonstrate numerical comparability across cultures (scalar equivalence). On the other hand, higher levels of equivalence allow for more detailed comparisons of scores across cultures. Whereas in the case of structural equivalence, only factor structures and nomological networks can be compared, scalar equivalence allows for more fine grained analyses of cross-cultural similarities and differences, such as comparisons of mean scores across cultures in t tests and analyses of (co)variance.

2.1 *Linguistic Equivalence*

The concept of linguistic equivalence is developed in multilingual studies. Versions of an instrument in

different languages show linguistic equivalence if these have the same characteristics that are relevant for the measurement outcome such as meaning, connotations of words and sentences, comprehensibility, and readability. Linguistic equivalence can be jeopardized by various sources, such as incorrect translations, words that are hard or impossible to translate (e.g., the English 'distress' does not have an equivalent in many languages), idiomatic expressions and metaphors (e.g., 'feeling blue'), and imprecise quantifiers ('rather often').

Because linguistic equivalence is not always guaranteed by a literal translation, it has become increasingly popular to utilize adaptations. In an adaptation, parts are changed (instead of literally translated) with the aim to improve an instrument's suitability for a target group.

Many studies employ a translation–back-translation procedure (Brislin 1986, Werner and Campbell 1970). This amounts to a translation, followed by an independent back-translation and a comparison of the original and back-translated version, possibly followed by some alterations of the translation. Such a procedure provides a powerful tool to enhance the correspondence of original and translated versions that is independent of the researcher's knowledge of the target language. Yet, it also has some disadvantages. It puts a premium on literal reproduction; this may give rise to a stilted language use in the target version that lacks the readability and natural flow of the original. The problem may be compounded by translators' awareness of their involvement in a translation–back-translation procedure.

A second problem involves translatability. The use of idiom (e.g., the English 'feeling blue') or references to cultural specifics (e.g., references to country-specific public holidays) or other features that cannot be represented adequately in the target language challenges translation–back-translations designs (and indeed all procedures in which existing instruments are translated). During the 1990s there was a growing awareness that translations and adaptations require the combined expertise of social and behavioral scientists (with a competence in the construct studied) and experts in the target language(s) and culture(s). In this so-called committee approach in which the expertise of all relevant disciplines is combined, there is usually no formal accuracy check of the translation. Usage of the committee approach is popular in large international bodies in which texts are translated in many languages like the United Nations and the European Union.

When versions in all languages are developed simultaneously, a procedure called 'decentering' (Werner and Campbell 1970) can be used. No single instrument or cultural group is then taken as starting point; individuals from different cultures jointly develop an instrument, thereby reducing the risk of introducing unwanted references to a specific culture.

The judgmental evidence of the designs to enhance linguistic equivalence can be easily combined with statistical approaches to establish equivalence. Reports of multilingual studies often provide a combination of judgmental–linguistic and empirical–statistical evidence to demonstrate the adequacy of an instrument and its translation. There is a tradeoff between the type of translation and the level of equivalence that can be obtained. When most or all questions of an instrument are adapted, structural equivalence is the highest level possible. When most or all items are literally translated, measurement unit and structural equivalence can be obtained. Recent statistical advancements in item response theory and structural equation modeling have made it possible to retain scalar equivalence, even when not all stimuli are literally translated (provided that all items measure the same underlying construct in each group).

3. *Sampling Cultures and Subjects*

Cross-cultural studies can apply three types of schemes to sample cultures. Three types of sampling can be envisaged. The first is probability (or random) sampling. Because of the large cost of a probability sample from all existing cultures, it often amounts to stratified (random) sampling of specific cultures (e.g., Western cultures). The second and most frequently observed type of culture sampling is convenience sampling. The choice of cultures is governed here by availability and cost efficiency: researchers decide to form a research network and all participants collect data in their own country. In the third type, called systematic sampling, the choice of cultures is more based on substantive considerations. A culture is deliberately chosen because of some characteristic, such as in Segall et al.'s (1966) study in which cultures were chosen on the basis of features of the ecological environment such as openness of the vista.

In survey research there is a well-developed theory of subject sampling (Kish 1965). In the area of cross-cultural research three types of sampling procedures of individuals are relevant as they represent different ways of dealing with confounding characteristics. The first is probability sampling. It consists of a random drawing from a list of eligible units such as persons or households. Confounding variables are not controlled for. The second type is stratified sampling. A population is stratified (e.g., in levels of schooling or socioeconomic status) and within each stratum a random sample is drawn. The purpose of stratification is the control of confounding variables (e.g., matching on number of years of schooling). The procedure cannot be taken to adequately correct for confounding variables when there is little or no overlap of the cultures (e.g., comparisons of literates and illiterates). The third procedure combines random or stratified

sampling with the measurement of control variables. The procedure enables a statistical control of ambient variables (e.g., using an analysis of covariance).

4. Individual- and Country-level Studies

The 1990s saw a growing interest in studies combining individual- and country-level data. Two kinds of studies have been reported. In multilevel hierarchical models (Bryk and Raudenbush 1992), a regression model is used to explain individual variation (e.g., explaining pupils' achievement scores on the basis of their intelligence) and class, district, or country variation (e.g., the explanation of the relationship between intelligence and achievement by means of school quality indicators). The second line of research involves the study of the same phenomena at different levels of aggregation (multilevel covariance structure analysis or multilevel factor analysis, Muthén 1994). From a conceptual point of view this line is more complicated, because it is well documented that (des)aggregatation can lead to methodological artefacts such as the ecological fallacy (see *Ecological Fallacy, Statistics of*), the incorrect application of culture-level characteristics to individuals. In each country a proportion of women is pregnant, but obviously, this proportion does not apply to any individual woman. In addition, equivalence issues have to be dealt with (Leung 1989). Hofstede's famous study of values is based on country characteristics; the four dimensions he reported (individualism, power distance, uncertainty reduction, and masculinity–femininity) are characteristics of countries and their applicability to individual behavior cannot be taken for granted and has to be established. Multilevel covariance structure and factor analysis are powerful tools to examine the equivalence of phenomena at different levels of aggregation, answering questions like the structural (in)equivalence of concepts like individualism–collectivism at individual and country level. For these and various other concepts the structural equivalence at different levels of aggregation is still unresolved.

See also: Psychometrics; Test Bias

Bibliography

Alasuutari P 1995 *Researching Culture. Qualitative Method and Cultural Studies*. Sage, Thousand Oaks, CA

Brislin R W 1986 The wording and translation of research instruments. In: Lonner W J, Berry J W (eds.) *Field Methods in Cross-cultural Research*. Sage, Beverly Hills, CA

Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park, CA

Camilli G, Shepard L A 1994 *Methods for Identifying Biased Test Items*. Sage, Thousand Oaks, CA

Campbell D T, Stanley J C 1966 *Experimental and Quasi-experimental Designs for Research*. R McNally, Chicago

Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. R McNally, Chicago

Greenfield P M 1997 Culture as process: Empirical methods for cultural psychology. In: Berry J W, Poortinga Y H, Pandey J (eds.) *Handbook of Cross-cultural Psychology*, 2nd edn. Allyn and Bacon, Boston, Vol. 1

Hambleton R K 1994 Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment (Bulletin of the International Test Commission)* **10**: 229–44

Ho D Y F 1996 Filial piety and its psychological consequences. In: Bond M H (ed.) *Handbook of Chinese Psychology*. Oxford University Press, Hong Kong

Hofstede G 1980 *Culture's Consequences: International Differences in Work-related Values*. Sage, Beverly Hills, CA

Holland P W, Wainer H (eds.) 1993 *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ

Kish L 1965 *Survey Sampling*. J Wiley, New York

Leung K 1989 Cross-cultural differences: Individual-level vs. culture-level analysis. *International Journal of Psychology* **24**: 703–19

Muthén B O 1994 Multilevel covariance structure analysis. *Sociological Methods and Research*. **22**: 376–98

Naroll R, Cohen R (eds.) 1970 *A Handbook of Cultural Anthropology*, 1st edn. American Museum of Natural History, Garden City, New York

Poortinga Y H, Malpass R S 1986 Making inferences from cross-cultural data. In: Lonner W J, Berry J W (eds.) *Field Methods in Cross-cultural Psychology*. Sage, Beverly Hills, CA

Segall M H, Campbell D T, Herskovits M J 1966 *The Influence of Culture on Visual Perception*. Bobbs-Merrill, Indianapolis, IN

Van de Vijver F J R, Leung K 1997 *Methods and Data Analysis for Cross-cultural Research*. Sage, Thousand Oaks, CA

Van de Vijver F J R, Tanzer N K 1997 Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology* **47**: 263–80

Werner O, Campbell D T 1970 Translating, working through interpreters, and the problem of decentering. In: Naroll R, Cohen R (eds.) *A Handbook of Cultural Anthropology*, 1st edn. American Museum of Natural History, Garden City, New York

F. van de Vijver

Cross-cultural Study of Education

Results of educational research from Western cultures may not be universally valid. Extending research to other cultures may reveal relevant phenomena and processes and allow the specification of universals and culture specificities as influenced by genetic and environmental factors relevant to development and learning.

Education is understood here as cultural transmission resulting from enculturation and socialization in informal or formal learning contexts. It includes intentional and unintentional influences on a person's thinking and behavior. Culture is not conceived of as a set of independent variables 'outside' of education influencing people in a unidirectional and unavoidable way. Culture is understood here as a certain commonality of meaning, customs and rules (not a homogeneous entity) shared by a certain group of people and setting a complex framework for learning and development.

1. Formal and Informal Education

Formal and informal educational contexts are of special interest for studies on education. Both contexts provide opportunities for and constraints on individual learning and development. These may function in different ways, depending on the cultural context.

1.1 Differentiation of Formal and Informal Education

Formal education, following the 'Western model,' consists of systematic instruction in 'universal' knowledge and skills, provided at prearranged times and places by specialists. Formal education also exists as indigenous education including initiation ceremonies, instruction provided in age-grade societies, and various forms of religious schooling, such as Quranic schools. There are cultural differences in the institutionalization of formal education, e.g., in the organization of the educational system, the relation between public and private schools, the length of compulsory schooling, the number of years of elementary and secondary schooling, the school curriculum, teaching, and instruction.

Informal education includes life-long socialization influences from interacting, e.g., with family members or peers in settings related to culture-specific values and norms useful for everyday experience. In all cultures, education takes place informally in everyday situations, in response to momentary needs, and sometimes simply involves observation and imitation (for a summary, see Segall et al. 1999, pp. 78–88). *Informal traditional* education including moral and spiritual aspects as well as manual labor, is in essence adapted to the local cultural system which it tends to perpetuate. According to Désalmand (1983), characteristics of traditional (African) education as compared with schooling are as follows: it is provided everywhere, all the time, and by everyone; it is closely tied to the environment, integrated with productive work, and addresses the needs of the society. It emphasizes cooperation rather than individual competition; everyone can be successful. Parents and elders play an important role, relations among participants are personalized, and occurs in the local language.

As Bruner (1996, p. ix) remarks: 'Schooling is only one small part of how a culture inducts the young into its canonical ways. Indeed, schooling may even be at odds with a culture's other ways of inducting the young into the requirements of communal living.'

1.2 Transmission of Know-how in Informal Education

1.2.1 Apprenticeship. An essential process in informal education is apprenticeship. Informal education can transmit specific, economically useful knowledge and often follows a very structured, albeit implicit, pedagogy. This was demonstrated in Greenfield and Lave's (1982) study of weaving apprenticeships among Zinacanteco girls in Mexico and tailoring in Liberia. Masters employ scaffolding to effect a progression within each step, allowing trial and error only if there is no material risk. Informal education thus involves more than just observation and imitation. Verbal instruction is an integral part of the apprenticeship process, but the pupils seldom speak and do not ask questions.

Based on her studies with Nahua Indians of Mexico, Chamoux (1986) provides a typology of know-how (or skill) and its transmission that distinguishes between 'incorporated know-how' and 'the mastery of algorithms.' The former results from personal, concrete experience; one may have a certain skill but not be able to describe it. It is therefore not easily transmitted by formal teaching and is best acquired through apprenticeship, i.e., by participating in the activity.

Incorporated know-how may be general or particular, i.e., transmitted to everyone or only to certain individuals who become specialists. Transmission occurs either with a master or through gradual immersion; the latter is possible in the case of cultural activities that can frequently be observed.

1.2.2 Ethnotheories. 'Indigenous theories of education' structure the pedagogical modalities in a coherent fashion (Chamoux 1986, Super and Harkness 1997). In some types of traditional informal education (e.g., for Nahuas), observation is preferred to instruction as a learning method, and may be combined with low pressure to learn and high preference for the will of the apprentice.

1.2.3 Guided participation in cultural activity. The way caretakers structure the activities of children is

another characteristic, studied by Rogoff et al. (1993), who observed caregivers and their toddlers in four settings: Mayan Indians in a small town in Guatemala, a tribal village in India, and two middle-class urban neighborhoods in Turkey and the USA. In the first two communities, children were nonverbally encouraged to observe adult activities; they were basically responsible for their own learning. In the two middle-class settings, mothers tended to structure explicitly their children's learning, used verbal interactions, and provided 'lessons' removed from the context of ongoing activities. They also tended to organize their schedule to separate adult activities from time devoted to interacting with children. In the two rural communities, adults shared their attention among a variety of activities, managing several tasks at the same time, including socializing with other adults and facilitating the children's involvement and learning. The behavior of the middle-class mothers can be seen as a preparation for schooling. Rogoff et al. (1993) suggest that many school reforms move in the direction of group work in which the children become more responsible for their own learning.

1.2.4 Schooling and apprenticeship. The transferability between practical and school-based knowledge is of special importance in education research. Delbos and Jorion (1984) observed how empirical knowledge is conveyed in traditional professions in France in contrast to scientific, school-based knowledge. In activities where competence is acquired through experience within the family, knowledge is transmitted not explicitly but through participation in daily life and as if it were an unquestionable truth. While this knowledge is specific, school knowledge is rather general and abstract but transferability does not always live up to the facts.

2. *Educational Effects*

One aim of education research is to specify conditions for development and learning. However, it is difficult to derive general conclusions on the basis of only Western-world theorizing and methodology.

2.1 *Effects of Different Educational Systems*

What are the effects when an educational system is implanted into another culture? The export of Western-type schooling into 'traditional' societies is often criticized because of its lack of cultural relevance and practical use. Julius Nyerere (1967) (a former President of Tanzania and also a teacher) was one of the first critics. Serpell and Hatano (1997) provide excellent coverage of the inportation of Western-type schooling during colonial times (in what they call 'hegemonic imposition') and summarize the present situation for Africa: 'The entire design of public schooling in Africa tends to bear very little relevance to the demands of economically productive adult life in rural settings, where the vast majority of school-leavers will spend their lives' (p. 363).

In contrast, Kagitçibasi (1996) points out the many positive effects of schooling, notably in terms of health promotion and women's rights, and is concerned that critiques of schooling may impede the implementation of policies to attain the UNESCO goal of 'education for all.'

Effects of formal and informal education are difficult to disentangle since both interact. Sometimes cumulative effects on learning can be observed when the educational goals of family and school fit together. This may be one reason for the higher achievement of Asian as compared with North American students (Stevenson 1988). For example, in Japan and Korea, teachers and mothers share the same educational values and coordinate informal and formal education.

Transfer of education between different contexts is easier when the same goals are functional. For example, the traditional Quranic education attempts to train memorizing by teaching fixed formulas, thus providing elements of social and cognitive skills required by modern schooling (Wagner 1988).

Negative effects on learning occur in the case of conflicting or nonmatching educational goals, e.g., in the case of involuntary acculturation when children from a minority group attend a majority-group school. The attainment of certain educational goals (e.g., acculturation, achievement) profits from the coordination of informal and formal education by cumulating their functional properties.

The effects of formal or informal education or their combination differ according to developmental age and depend on the person, the environmental conditions, and the adaptive function of the specific skill. According to studies by Wagner and Spratt (1987) language background and preschool experience have different effects on the reading acquisition of rural and urban Moroccan children in the first year of primary school. 'Modern' and 'traditional' preschooling need not differ with respect to higher level reading measures. The traditional Quranic preschool can be beneficial for literacy acquisition in Arabic. However, differences in literacy between urban and rural (or Arabic-speaking and Berber-speaking) children in Morocco caused by maternal language and preschooling could not be eliminated by mainstreaming modern primary schools. The effects of an urban environment were longer lasting than those of language background and preschool experience. These studies are relevant for analyzing effects of multicultural education.

Formal schooling can influence general and domain-specific skills (e.g., literacy, logical thinking). Studies on Quranic-schooled adults in Liberia which isolate the effects of schooling and literacy show that literacy increases specific memory skills but can be acquired without Western schooling (Scribner and Cole 1981). Formal schooling can profit from informal education in contexts which have practical significance for arithmetic skills. However, informally acquired arithmetic skills in comparison to school-based algorithms have their limits (for a review, see Segall et al. 1999).

2.2 International Comparison of School Achievement

Formal schooling usually follows the 'Western model,' giving preference to such cognitive skills that facilitate success in highly technological and scientifically structured environments.

The Program for International Student Assessment (PISA) aims to provide nationally representative data from OECD countries on 15-year-old students' knowledge and skills in mathematics, science, reading, and cross-curricular competencies. Three large-scale comparative studies (First, Second and Third Mathematics and Science Study; FIMS, SIMS, TIMSS) by the policy oriented International Association for the Evaluation of Educational Achievement (IEA) demonstrate significant differences in school achievement (for a review, see Baumert 1998). Even among societies with a common language (Switzerland, Germany, Austria), significant differences occurred. Comparisons of mathematical achievement demonstrate significantly higher achievement of Japanese and Taiwanese as compared with North American children (Stevenson and Stigler 1992). High-achieving students from Asia as compared with Germany and North America show better mathematical performance. The test scores of Chinese and Japanese students in mathematics and science are higher than those of their American peers at all age groups (Stevenson et al. 1993).

These differences are presumably related to several factors, including differences in educational systems, teaching and learning styles, time use, cultural values, and relations between informal and formal education (e.g., Stevenson 1988). East Asians place a higher value on education, focus on the role of effort in academic achievement, hold higher aspirations, devote more time in academic effort, and report more parental involvement in their children's academic work (Stevenson et al. 1993). Differences in cognitive skills can be mediated by motivational and socioemotional factors (Kornadt et al. 1980, Kornadt and Trommsdorff 1996). Whether these are also affected by education has to be studied in future research.

3. Outlook

An increasing number of studies include culture in educational research. However, systematic comparisons are rare. Also, results on effects of culture and education are not consistent. This is related to several problems.

One problem lies in the fact that the focus of research was on either informal or formal education as *separate settings* ignoring their specific and possibly cumulative effects. School quality, family, and peer groups interact in various ways depending on the cultural context, thus affecting development and learning (Chamoux 1986).

Another problem lies in the fact that possible effects of genetic dispositions have rarely been studied and the role of education often has been overestimated. Research in the tradition of the nature–nurture controversy focusing on either biological or environmental conditions ignore the role of the active, self-reflective person (the learner as the producer of his/her development). Future research should take into account the impact of genetic dispositions on the development of abilities and the readiness for learning certain skills, and on interactions between environment (education) and person (Weinert and Perner 1996).

A further theoretical and methodological problem is the validity of the data. This is partly due to experimental artifacts and the context dependence of the behavior studied. Also, much cross-cultural research is based on Western theories and methods without explicit reference to culture. Often the role of certain skills in a specific culture is neglected. The functional equivalence of variables and measurements needs to be ascertained for valid comparisons. Here problems of etic (comparative) and emic (culture-specific, indigenous) methods are to be solved. Both approaches are necessary to arrive at universal and culture-specific theorizing (Trommsdorff 1986, Segall et al. 1999).

Education research needs a solid theoretical frame. An eco-cultural approach seems promising since it relates cultural conditions to educational and psychological processes. One basic assumption is that in the case that educational goals and practices are relevant in a culture, one can expect qualified development of related skills. For example, in hunting as compared with agricultural societies, spatial skills are needed for survival; they are more valued and more efficiently learned. In some cultures, holistic rather than analytic problem solving and group discussion rather than individual reflection are valued; educational goals are defined accordingly (for a review, see Serpell and Hatano 1997). Comparative research on intelligence is another example of how the culture-specific meaning was clarified (for a review, see Berry 1984). Conceiving intelligence as the adaptation of mental activities to the environment requires culture-specific measures of

'intelligent' action and appropriate educational conditions for its development.

An eco-cultural approach starts from the insider's understanding and specifies what has to be accomplished (in various domains of behavior) in order to live in a certain cultural context. This allows one to link cultural context and cognitive skills and to study indigenous actions as part of the respective culture and as an indicator of 'practical intelligence.' A 'practice theory of culture and cognition' attempts to analyze educational processes affecting cognitive skills (Wagner and Spratt 1987). Related research on 'everyday cognition' follows the sociohistorical approach rooted in Vygotsky's writings (for a review, see Segall et al. 1999).

These approaches are linked to the theoretical framework of the developmental niche proposed by Super and Harkness (1997) who assume that three subsystems surround the child: the physical and social contexts in which the child lives, the cultural educational practices, and the psychological characteristics of the caretaker, including their ethnotheories.

The developmental niche adapts to social change and thus allows for adaptation on the individual level, including acculturation. This theorizing is related to Bronfenbrenner's work and the eco-cultural framework suggested by Segall et al. (1999).

These theoretical approaches assume that education is linked to the given context which provides constraints and resources for learning and development. This needs further empirical testing. Once a solid theory has been developed, problems of measurement and the application of education research will be more easily solved.

See also: American Studies: Education; Apprenticeship and School Learning; Comparative Research in Education: IEA Studies; Comparative Studies: Method and Design; Cross-cultural Psychology; Cross-cultural Research Methods; Cultural Diversity, Human Development, and Education; Education: Cultural and Religious Concepts; Latin American Studies: Education; Soviet Studies: Education

Bibliography

Baumert J 1998 Internationale Schulleistungsvergleiche. In: Rost D H (ed.) *Handwörterbuch Pädagogische Psychologie*, Beltz, Weinheim, Germany, pp. 219–25

Berry J W 1984 Toward a universal psychology of cognitive competence. *International Journal of Psychology* **19**: 335–61

Bruner J 1996 *The Culture of Education*. Harvard University Press, Cambridge, MA

Chamoux M N 1986 Apprendre autrement: Aspects des pédagogies dites informelles chez les Indiens du Mexique. In: Rossel P (ed.) *Demain l'Artisanat?* Presses Universitaires de France, Paris, pp. 211–335

Delbos G, Jorion P 1984 *La Transmission des Savoirs*. Maison des Sciences de l'Homme, Paris

Désalmand P 1983 *Histoire de l'Éducation en Côte d'Ivoire*. CEDA, Abidjan, Ivory Coast, Vol. 1

Greenfield P, Lave J 1982 Cognitive aspects of informal education. In: Wagner D A, Stevenson H W (eds.) *Cultural Perspectives on Child Development*. Freeman, San Francisco, pp. 181–207

Kagitçibasi C 1996 *Family and Human Development Across Cultures: A View from the Other Side*. Erlbaum, Hillsdale, NJ

Kornadt H-J, Eckensberger L H, Emminghaus W B 1980 Cross-cultural research on motivation and its contribution to a general theory of motivation. In: Triandis H C, Lonner W (eds.) *Handbook of Cross-cultural Psychology: Vol. 3. Basic Processes*. Allyn and Bacon, Boston, pp. 223–321

Kornadt H-J, Trommsdorff G 1996 Cross-cultural approaches to human development. In: DeCorte E, Weinert F E (eds.) *The Encyclopedia of Developmental and Instructional Psychology*. Pergamon, Oxford, pp. 51–5

Nyerere J M 1967 *Education for Self-reliance*. Government Printer, Dar es Salaam, Tanzania

Rogoff B, Mistry J, Göncü A, Mosier C 1993 Guided participation in cultural activity by toddlers and caregivers. *Monographs of the Society for Research in Child Development* No. 50 (Serial No. 236)

Scribner S, Cole M 1981 *The Psychology of Literacy*. Harvard University Press, Cambridge, MA

Segall M H, Dasen P R, Berry J W, Poortinga Y H 1999 *Human Behavior in Global Perspective: An Introduction to Cross-cultural Psychology*, 2nd edn. Allyn and Bacon, Boston

Serpell R, Hatano G 1997 Education, schooling, and literacy. In: Berry J W, Dasen P R, Saraswathi T S (eds.) *Handbook of Cross-cultural Psychology: Vol. 2. Basic Processes and Human Development*, 2nd edn. Allyn and Bacon, Boston, pp. 339–76

Stevenson H W 1988 Culture and schooling: Influences on cognitive development. In: Hetherington E M, Lerner R M, Perlmutter M (eds.) *Child Development in Life-span Perspective*. Erlbaum, Hillsdale, NJ, pp. 241–58

Stevenson H W, Chen C, Lee S Y 1993 Mathematics achievement of Chinese, Japanese, and American children: Ten years later. *Science* **259**: 53–8

Stevenson H W, Stigler J W 1992 *The Learning Gap: Why Our Schools are Failing and What We Can Learn from Japanese and Chinese Education*. Summit Books, New York

Super C M, Harkness S 1997 The cultural structuring of child development. In: Berry J W, Dasen P R, Saraswathi T S (eds.) *Handbook of Cross-cultural Psychology: Vol. 2. Basic Processes and Human Development*, 2nd edn. Allyn and Bacon, Boston, pp. 1–39

Trommsdorff G 1986 German cross-cultural psychology. *German Journal of Psychology* **10**: 240–66

Wagner D A 1988 'Appropriate education' and literacy in the Third World. In: Dasen P R, Berry J W, Sartorius N (eds.) *Health and Cross-cultural Psychology: Toward Applications*. Sage, Newbury Park, CA, pp. 93–111

Wagner D A, Spratt J E 1987 Cognitive consequences of contrasting pedagogies: The effects of Quranic preschooling in Morocco. *Child Development* **58**: 1207–19

Weinert F E, Perner J 1996 Cognitive development. In: Magnusson D (ed.) *The Lifespan Development of Individuals: Behavioral, Neurobiological, and Psychosocial Perspectives*. Cambridge University Press, New York, pp. 207–22

G. Trommsdorff and P. Dasen

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Cross-modal (Multi-sensory) Integration

1. Multiple Sensory Representations in the Brain

Large brains generally are found in large animals. This is primarily because the body evolves in concert with the neural architecture necessary to handle the information it provides and to control its movement. Although there are certainly exceptions to the rule, and the human brain is disproportionately large, the actual acumen of the host is not as relevant to overall evolution of brain size as one might think, and even big dim-witted animals can have big brains.

Because animals have multiple senses, and each requires a good deal of neural circuitry to make them work, much of even the largest brain ends up being devoted to representing the sensory complex. Thus, the cerebral cortex, generally considered the highest stage of information processing and the seat of perception, appears to have a large portion of its area organized into a succession of higher order sensory (and motor) processing stations. Thus far, a minimum of 32 visual, 15 auditory, and 8 somatosensory areas have been identified in the primate cortex. Somehow, though no one really knows how, the co-ordinated operation of multiple cortical areas, each responsible for dealing with a limited set of stimulus qualities, produces a unified perceptual experience.

That so much of the brain is devoted to sensory processing and its associated motor commands is due to the fact that survival depends on the speed and accuracy with which an organism can evaluate external events and properly react to them. Consequently, having many senses is generally better than having few and, despite the increased metabolic cost of building and maintaining the necessary underlying brain circuits, there are significant overall gains. For one thing, sensory modalities can substitute for one another in different circumstances (e.g., sight and touch substitute for vision in the dark). For another, multiple senses allow the organism to simultaneously sample very different forms of environmental energy, changes in any one of which may provide important information about an event. Multiple inputs from different senses also increase the likelihood of correctly identifying a given event, for it may be difficult to distinguish on the basis of a single sensory dimension (two events may sound alike, but look different).

2. Multi-sensory Integration and Perception

Yet, the brain is able to make more of its multiple senses than might at first be apparent, and by synthesizing information from two or more modality-specific inputs it creates something entirely new. At the level of sensation/perception, the new product of this 'multi-sensory integration' is a unique experience. For example, what we think of as 'taste' is not due solely to the signals the brain receives about the food-induced chemical changes on the tongue. Rather, taste represents a synthesis of signals derived from these chemical changes (gustatory) and those from the aroma (olfactory) and texture (somatosensory) of the food. Thus, multi-sensory integration provides a mechanism for producing much richer and more diverse sensory experiences than would otherwise be possible. At the neural level, multi-sensory integration appears to create a new, amplified signal that is greater than that produced in response to either of its constituent modality-specific cues. In fact, this signal is often greater than the sum of the constituent responses; an excellent biological example of the whole being greater than the sum of its parts. Such an enhanced sensory signal is strongly correlated with an increased probability of a behavioral response. Furthermore, as will be shown below, the multi-sensory amplification process is selective; it increases the brain's ability to detect and identify events by amplifying only those cross-modal sensory signals whose spatial and temporal properties indicate that they are likely to have a common origin. Indeed, those cross-modal signals that appear to be derived from different events can be minimized via these same synthetic mechanisms. When the stimulus conditions do favor an enhanced multi-sensory signal, not only is the detection of the event and the probability of an appropriate response enhanced, but the speed of the response is increased as well. These multi-sensory consequences have been demonstrated in both animals (Frens et al. 1998, Stein et al. 1989) and humans (Frens et al. 1995, Goldring et al. 1996, Hughes et al. 1994). Furthermore, multisensory integration has been associated with enhancing object recognition (Giard and Peronnet 1999, O'Hare 1991), increasing the apparent intensity of a signal (Stein et al. 1996) and aiding in speech perception (see Massaro and Stork 1998).

It has been known for some time that multiple sensory inputs must be bound together to produce what we experience as a 'unitary' perceptual experience, and that this 'binding' process can have potent perceptual consequences. Yet, we are largely unaware of its actual operation. Perhaps we don't notice the process because it is the normal mode of brain operation. For example, seeing a speaker's face enhances our ability to perceive what is being said (see Sumby and Pollack 1954). Yet, we are rarely conscious of this except, perhaps, in a noisy room where the benefit of combined visual-auditory inputs is particularly important and, as a result, we try very hard to focus on the speaker's face. The consequence is an enhancement of the brain activity that underlies speech perception (e.g., see Calvert et al. 1999, Sams et al. 1991) induced by the congruent visual (lip movement) and auditory (spoken words) stimuli we receive and integrate. Indeed, visual-auditory integration in speech perception is so seamless that we are quite

surprised when we experience the so-called 'McGurk effect.' Here, the coupling of the sound of one word with the lip movements of another produces a cross-modal product that differs from either. One example of this is hearing the word 'bows' while seeing lip movements that form the word 'goes' and perceiving 'doze' or 'those.' This cross-modal illusion underscores the potent multi-sensory interactions that normally occur between vision and audition.

Whereas the perceived meaning of the signal is altered in the McGurk effect, the perceived location of the signal is affected in another compelling visual-auditory illusion, the 'ventriloquism effect' (Howard and Templeton 1966). In this illusion a visual cue (e.g., movement of the lips) provides a signal that alters the apparent location of the source of the sound. We commonly experience this illusion when watching a movie. Voices always seem to be coming from the appropriate characters, regardless of their location or movement on the screen and despite the fact that all the sounds actually originate from the same location—the speakers at the sides of the screen. There are many other compelling illusions involving different senses, such as the 'parchment skin' illusion, in which the apparent texture of the skin is changed as a consequence of coupling the sensation of rubbing of the fingers with a grating sound that is played through earphones the subject is wearing (Jousmaki and Hari 1998), or the oculogravic illusion (Graybiel 1952), in which vestibular cues alter visual perceptions. Although the neural basis of these individual illusions is poorly understood at present, we have begun to understand some of the general principles by which individual neurons in the brain combine simple stimuli from different sensory inputs to produce a multisensory signal. It is believed that understanding these principles will help us better understand the behavioral and perceptual consequences of multi-sensory integration.

3. *Multi-sensory Neurons and the Superior Colliculus (SC)*

In order to effect multi-sensory integration at the level of the single neuron, there must be sites in the brain where neurons receive converging inputs from different senses. Such regions have been found at nearly every level in the central nervous system (see Stein and Meredith 1993), thereby paralleling areas that are devoted to processing information for each individual sensory modality. As a consequence of these seemingly parallel systems, higher animals have developed a curious duality: some areas of the brain are specialized for processing information on a sense-by-sense basis, and probably give rise to the peculiar subjective impressions that are so closely bound to one modality or another (e.g., the perceptions of color, pitch, tickle, or itch). Conversely, other areas are specialized for information processing regardless of the specific sense or senses from which that information is derived. These latter areas depend on a convergence of inputs from different senses onto a common set of (multisensory) neurons.

One of the best known structures in which multisensory convergence occurs is found in the midbrain: the superior colliculus (SC). In fact, researchers have leaned heavily on the responses of multi-sensory SC neurons to develop a model that is helpful in understanding the general neural properties of multisensory integration, and this has proved to be an excellent strategy for a host of reasons. First, the SC is richly endowed with converging visual, auditory, and somatosensory inputs, and thus has many multisensory neurons that can be studied electrophysiologically. Second, the SC is also known to play an important role in producing motor actions that are guided by sensory stimuli. Specifically, it is involved in generating orienting behaviors to stimuli in contralateral space and, thus, for example, mediates rapid shifts of gaze (i.e., eye and head movements) to objects of interest. These factors make it possible to use the same structure to explore the principles governing multi-sensory integration at the neuronal and behavioral levels.

A multi-sensory SC neuron has multiple receptive fields, one for each sensory modality to which it is responsive. A receptive field defines the area of sensory space (e.g., visual space, auditory space, somatosensory space) where a given stimulus will be effective in activating the neuron. For multi-sensory neurons, their different receptive fields are in spatial register with one another. For example, a visual-auditory neuron with a visual receptive field in central space will be responsive to auditory cues in an overlapping region of auditory space as shown in Fig. 1. The spatial register among the modality-specific receptive fields of a multi-sensory neuron plays a critical role in its synthesis of cross-modal cues. Such register is characteristic of multisensory neurons regardless of where they are found in the nervous system, and is maintained even if, for example, one set of peripheral sensory organs moves. Thus, for example, any eye movement will shift visual space relative to auditory (head-axis) or somatosensory (body-axis) space because the retina and its receptor cells have moved. Rather than disrupt receptive field alignment, movement-induced misalignment of sensory reference frames sets into motion compensatory neural mechanisms that attempt to maintain receptive field register. Thus, a shift in eye position to the right produces a rightward shift in the auditory receptive field of an SC neuron (Harrington and Peck 1998, Hartline et al. 1995, Jay and Sparks 1984). Similarly, there is even evidence that the somatosensory receptive fields of visual-somatosensory neurons show compensatory changes as a consequence of eye movements (Groh and Sparks 1996). Although the mechanisms

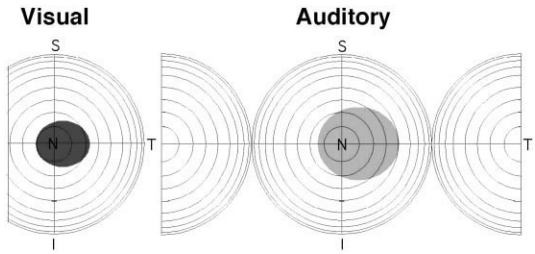

Figure 1
The individual receptive fields of a visual-auditory neuron are in good spatial register. The circles shown here are schematic diagrams of visual (left) and auditory (right) space. The intersection of the horizontal and vertical meridians depicts a point directly in front of the subject, and each concentric circle represents 10°. The frontal half of auditory space is shown by the central sphere, and the caudal half is represented by the two hemispheres that have been split and folded forward. The respective receptive fields of a typical multisensory neuron are shown by the shading. Note that superimposing the diagrams of visual and auditory space will show that most of the two receptive fields overlap each other. S = superior, I = inferior, N = nasal, T = temporal

underlying these dynamic shifts in receptive fields are not well understood, using the retina as a referent for shifting nonvisual receptive fields in the SC reflects its principal role in generating gaze shifts to targets of any modality, including those that are multi-sensory. This is not the case everywhere multi-sensory neurons are found. In regions of premotor cortex, where visual-somatosensory neurons are found and whose activity generates limb rather than eye movements, movements of the limb produce compensatory changes in the location of visual receptive fields (Gentilucci et al. 1983, Graziano et al. 1994). Once again, cross-modal receptive field register is a key element in the maintenance of a modality-independent coordinate frame to properly code movement.

4. *Principles of Multi-sensory Integration in Single SC Neurons*

4.1 *Spatial*

The enhanced multi-sensory response of a neuron depends on stimuli having access to its two (or more) receptive fields simultaneously or within a short window of time. Given the spatially overlapping nature of receptive fields in multi-sensory neurons, the activating stimuli should derive from the same location in space, such as when they originate from the same event. The consequent multi-sensory response typically is a train of discharges that contains more impulses than one evoked by either of the modality-specific stimuli. Indeed, if the modality-specific responses are themselves weak, the response to the combined stimulus often exceeds the arithmetic sum of the modality-specific responses (Fig. 2).

If, however, the stimuli are derived from different locations (e.g., originate from different events), so that one stimulus falls outside its receptive field and is spatially discordant with the other stimulus, no such response enhancement is produced. In fact, under such circumstances there is a good probability that the opposite effect, response depression, will occur. In such cases, even vigorous visual responses can be abolished (Fig. 2). These changes at the single neuron level are reflected at the behavioral level as enhanced attentive and orientation responses to spatially coincident cross-modal stimuli, and depressed attentive and orientation responses to spatially disparate cross-modal stimuli (Stein et. al. 1989).

These opposing effects of the same stimuli are believed to be due to a neuron receiving synergistic excitatory inputs when its two input channels are stimulated by spatially aligned stimuli (both stimuli fall within the excitatory receptive fields of that neuron), but, because some receptive fields have inhibitory regions at their borders, response depression can be explained by the antagonism between the inhibitory input from the stimulus outside the receptive field and the excitatory input produced by the stimulus within its receptive field (Kadunce et al. 1997).

4.2 *Temporal*

Integrated multi-sensory responses also depend on the relative timing of the stimuli. When the interval between the occurrence of the stimuli is too large, they are recognizable in the neuronal response as discrete events. In other words, no integration takes place. Nevertheless, for any given neuron this temporal 'window' may be several hundred milliseconds and generally is predictable from the duration of its response to the stimulus that causes the earliest excitation. So, for example, if a somatosensory stimulus elicits a response that lasts 200 milliseconds, a multi-sensory interaction will occur if the activation produced by a stimulus of a second modality occurs during that period (Meredith et al. 1987, Wallace and Stein 1997).

This is important because even though different senses may be activated by the same event, the speed with which the modality-specific components activate a multi-sensory neuron can differ substantially (auditory and somatosensory information reaches SC neurons in 10–25 ms, while visual information requires

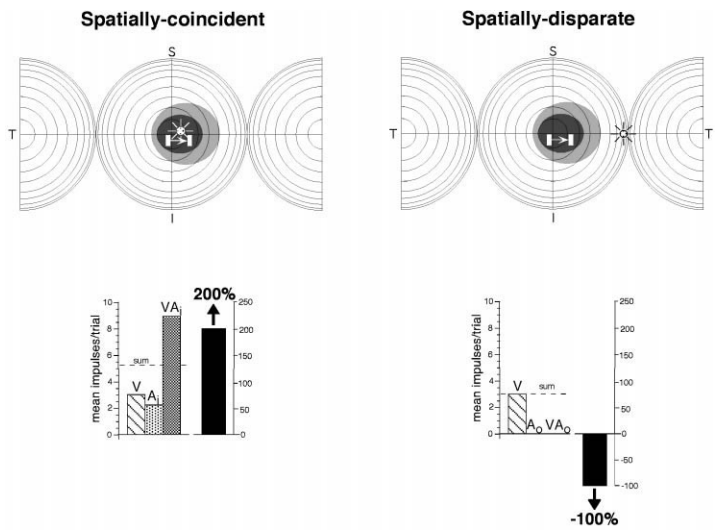

Figure 2
The presentation of two visual-auditory stimulus pairs, that differed only in their spatial coincidence, resulted in very different neural responses. Top left: within the visual (dark shading) and auditory (lighter shading) receptive fields of this multi-sensory neuron are icons depicting the close spatial coincidence of a moving bar of light (arrow) and a stationary auditory stimulus (speaker). At the bottom are shown the neuron's average responses to each stimulus alone (V = visual, Ai = auditory within the receptive field) as well as to their combination (VAi). The dark bar shows that the multi-sensory response exceeded the arithmetic sum of the individual visual and auditory responses (multi-sensory enhancement). Right: when the auditory stimulus was presented far from its visual counterpart and outside its receptive field (Ao = auditory outside the receptive field) it abolished the visual response (multi-sensory depression)

40–120 ms). If precise coincidence were required, very few multi-sensory interactions would take place. Nevertheless, the temporal interaction window is finite and different neurons respond best to different stimulus asynchronies. Generally speaking, the magnitude of the interaction reflects the degree to which the modality-specific activation patterns overlap in time. Thus, the magnitude of the interaction diminishes rapidly as this period of temporal overlap declines (Fig. 3). In some cases, temporal intervals that are transitional between those that produce enhanced responses and those that produce two independent responses give rise to response depression. In the case of visual-auditory neurons, the ability to respond differentially to stimuli with different input latencies suggests a tuning to events occurring at different distances from the animal. Whether there is some systematic organization of these neurons that may be helpful in localizing stimuli in 3-dimensional space remains to be determined.

4.3 Inverse Effectiveness

The degree to which the response to a multi-sensory stimulus exceeds that to either of the unimodal component stimuli is most often inversely related to the effectiveness of the modality-specific stimuli. This makes intuitive sense if we consider the possible functional consequences of multi-sensory enhancement. By definition, very effective stimuli evoke strong neural responses and are very likely to elicit a behavioral response. Thus, if the objective of multisensory enhancement is to increase the probability of detecting and/or responding to a stimulus, there would be little gained from augmenting an already vigorous modality-specific response. On the other hand, substantial benefit might be had by combining the influences of weak modality-specific stimuli. But what is to be gained from this? One could argue that increasing the sensitivity to relatively weak stimuli would make the system susceptible to environmental and/or neural 'noise.' However, one could counter this argument by noting that the probability that spatially and temporally coincident activation occurring on independent modality-specific channels by chance is low and that, more often than not, this circumstance would correspond to a real event in the external world. It is this sort of reasoning that has provided the impetus for many organizations (most prominently, industry and the military) to develop

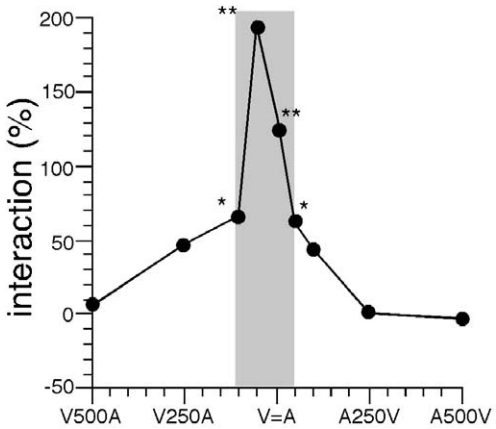

Figure 3
The temporal window for a multi-sensory interaction in one visual-auditory neuron. The magnitude of a multi-sensory interaction is plotted here as a function of the interval between the onset of the two stimuli. Positive values represent response enhancements, negative values represent response depressions. Note that in this neuron simultaneous onset of the visual and auditory stimuli (V = A) produced the greatest multi-sensory enhancement. The enhancement was lost in this particular neuron when the visual stimulus preceded the auditory stimulus by more than 100 milliseconds or when the auditory stimulus preceded the visual stimulus by more than 50 milliseconds. Asterisks depict significant interactions, and the shaded region represents the temporal window within which significant multi-sensory interactions were obtained. Different neurons showed their maximum multi-sensory response at different cross-modal intervals (generally within 50–100 milliseconds of simultaneity) and had different temporal windows (generally 100–500 milliseconds)

sensor-fusion devices for navigation or to aid in searching for objects embedded in 'noisy' environments such as a forest or sea floor.

5. *The Essential Circuit For Multi-sensory Integration in SC Neurons*

One might reasonably assume that any neuron receiving inputs from two different sensory modalities would have the ability to integrate these inputs. However, this is not the case. Multi-sensory neurons in the SC receive converging modality-specific inputs from many sources (see Stein et al. 1993). However, recent experiments in cats have shown that modality-specific inputs derived from a region of cerebral cortex play a critical role in mediating multi-sensory integration in SC neurons (see Jiang et al. 1999, Wallace and Stein 1994). As shown in Fig. 4, when the SC is

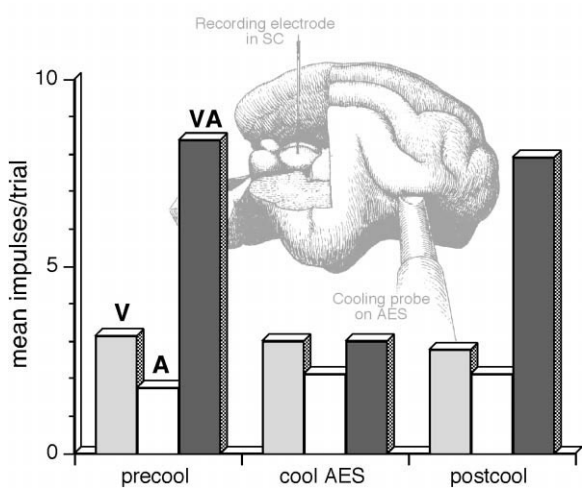

Figure 4
Inputs from association cortex are critical for multi-sensory integration in the SC. Shown here is a representative example of an SC neuron whose sensory responses were examined before, during, and after temporarily deactivating association cortex by cryogenic blockade (cooling). Under control conditions (precool), this neuron exhibited a substantial multi-sensory response enhancement to the visual-auditory stimulus combination. When a region of association cortex (the anterior ectosylvian sulcus, AES) was deactivated (cool AES), this enhancement was abolished, even though the neuron retained its normal responses to the visual and auditory stimuli when presented individually. Following rewarming of AES (postcool), the enhanced multi-sensory response was reinstated.

deprived of inputs from this region of the cortex (by reversible cryogenic blockade of cortical activity), SC neurons retain the ability to respond to inputs from different sensory modalities, but lose their ability to integrate them and thereby enhance their responses. Rather, they respond to combinations of cross-modal stimuli in much the same way as they would respond to the most effective of these modality-specific stimuli in isolation.

Presumably, these cortical deactivation-induced physiological changes in SC neurons would disrupt SC-mediated multi-sensory behaviors. A test of this was made by training animals to respond to visual cues in the presence or absence of auditory stimuli, and then testing them with or without this region of cortex deactivated (Fig. 5). When the cortex was temporarily deactivated (with either a local anesthetic injection or cryogenic blockade) there was no change in the animal's responses to the visual cue, but the ability of the auditory cue to enhance (spatially coincident stimuli) or degrade (spatially disparate stimuli) re-

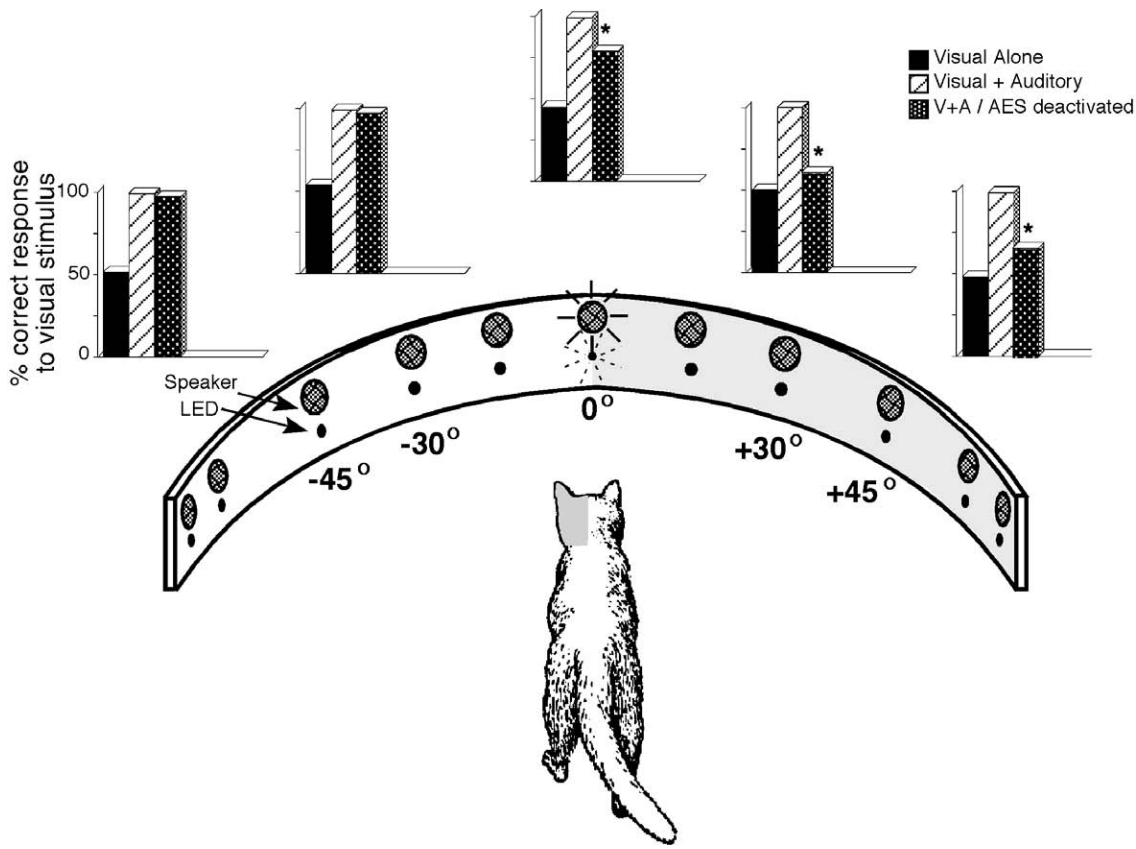

Figure 5
Multi-sensory orientation behaviors are affected by cortical deactivation as predicted from physiological studies. Three cats were trained in a perimetry device (bottom) to orient to and approach an illuminated light-emitting diode (LED) for a food reward. During testing, the intensity of each LED was manipulated until each animal responded correctly to that LED on approximately 50 percent of the trials (black bars, visual alone). The pairing of the auditory (speaker) with the visual stimulus at the same spatial location dramatically enhanced the probability of a correct response (visual + auditory). However, during deactivation of the left AES (shading on animal's head), these multi-sensory behavioral response enhancements were compromised in the contralateral hemifield (V + A/AES deactivated). Data is shown for 5 stimulus locations from left to right: 45° left (−45°), 30° left (−30°), directly ahead (0°), 30° right (+30°) and 45° right (+45°). Asterisks denote where statistically significant effects of cortical deactivation were obtained. In agreement with the results of physiological studies (see Fig. 4), responses to a modality-specific (visual) stimulus alone were unaffected by cortical deactivation (not shown) (modified from Wilkinson et al. 1996)

sponses was severely disrupted (Wilkinson et al. 1996). These behavioral observations, coupled with the electrophysiological results described above, demonstrate that the neural circuit in the SC essential for multisensory integration must include inputs from the cerebral cortex.

6. *Commonalities in Multi-sensory Integration Across Species*

Although many of the specifics that were presented above were drawn from experiments in cats, most of the sensory features that characterize the SC of this animal are also found in every other laboratory species that has been examined thus far. Particularly important in the current context is the spatial register among the different receptive fields of multi-sensory neurons. This has been found in rodents, carnivores, and primates. The widespread nature of this scheme for coupling inputs from different sensory modalities, and its presence in neurons of the nonmammalian homologue of the SC, the optic tectum (see Stein and Meredith 1993), suggest that it was preserved during

the transition from nonmammalian to mammalian forms and is adaptable to a wide variety of ecological circumstances. Furthermore, its utility is not restricted to neurons involved in SC-mediated behaviors, as it is also characteristic of multi-sensory neurons in many neocortical areas (e.g., Bruce et al. 1981, Duhamel et al. 1991, Fogassi et al. 1996, Graziano et al. 1997, Ramachandran et al. 1993, Rizzolatti et al. 1981a, 1981b, Stein et al. 1993, Wallace et al. 1992), where such neurons are likely to be involved in functional roles very different from those involving the SC. The few experiments that have examined multi-sensory processes in cortical neurons have yielded similar results. Indeed, despite the observation that spatially disparate stimuli produce a weaker inhibitory interaction in cortex than SC, many of the same principles of multi-sensory integration appear to be operative in both areas. This lends credence to the notion that the fundamental principles of multi-sensory integration are likely to supersede structure, function, and species.

This does not mean that there are not substantial species differences in the stimuli that are effective in activating multi-sensory neurons or in differences in the roles played by multi-sensory neurons in different regions of the brain. Nevertheless, it does appear likely that despite the various specialized characteristics of multi-sensory neurons in different animals and/or structures, there exists a common set of principles by which they pool information from their different sensory channels. The presence of a common set of integrative principles that operate at different levels of the neuraxis is likely to provide a comparatively simple means of simultaneously enhancing or degrading (i.e., coordinating) the activity of different brain areas that cooperate in the production of an integrated behavior. It now remains to be seen whether that postulate is confirmed.

See also: Auditory System; Neural Synchrony as a Binding Mechanism; Olfactory System; Perceptual Organization; Somatosensation; Taste and Smell, Psychology of; Taste, Neural Basis of; Vestibular System; Vision, High-level Theory of; Vision, Psychology of

Bibliography

Bruce C, Desimone R, Gross C G 1981 Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology* **46**: 369–84

Calvert G A, Brammer M J, Bullmore E T, Campbell R, Iversen S D, David A S 1999 Response amplification on sensory-specific cortices during crossmodal binding. *NeuroReport* **10**: 2619–23

Duhamel J-R, Colby C L, Goldberg M E 1991 Congruent representations of visual and somatosensory space in single neurons of monkey ventral intraparietal cortex (area VIP). In: Paillard J (ed.) *Brain and Space*. Oxford University Press, New York

Fogassi L, Gallese V, Fadiga L, Luppino G, Matelli M, Rizzolatti G 1996 Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology* **76**: 141–57

Frens M A, Van Opstal A J, Van der Willigen R F 1995 Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception and Psychophysics* **57**: 802–16

Gentilucci M, Scandolara C, Pigarev I N, Rizzolatti G 1983 Visual responses in the postarcuate cortex (area 6) of the monkey that are independent of eye position. *Experimental Brain Research* **50**: 464–8

Giard M H, Peronnet F 1999 Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience* **11**: 473–90

Goldring J E, Dorris M C, Corneil B D, Ballantyne P A, Munoz D P 1996 Combined eye-head gaze shifts to visual and auditory targets in humans. *Experimental Brain Research* **111**: 68–78

Graybiel A 1952 Oculogravic illusion. *American Medical Association Archives of Ophthalmology* **48**: 605–15

Graziano M S, Hu T, Gross C G 1997 Visuospatial properties of ventral premotor cortex. *Journal of Neurophysiology* **77**: 2268–92

Graziano M S, Yap G S, Gross C G 1994 Coding of visual space by premotor neurons. *Science* **266**: 1054–7

Groh J M, Sparks D L 1996 Saccades to somatosensory targets. III. Eye-position-dependent somatosensory activity in primate superior colliculus. *Journal of Neurophysiology* **75**: 439–53

Harrington L K, Peck C K 1998 Spatial disparity affects visual-auditory interactions in human sensorimotor processing. *Experimental Brain Research* **122**: 247–52

Hartline P H, Pandey Vimal R L, King A J, Kurylo D D, Northmore D P M 1995 Effects of eye position on auditory localization and neural representation of space in superior colliculus of cats. *Experimental Brain Research* **104**: 402–8

Howard I P, Templeton W B 1966 *Human Spatial Orientation*. Wiley, London

Hughes H C, Reuter-Lorenz P A, Nozawa G, Fendrich R 1994 Visual-auditory interactions in sensorimotor processing: saccades versus manual responses. *Journal of Experimental Psychology: Human Perception and Performance* **20**: 131–53

Jay M F, Sparks D L 1984 Auditory receptive fields in primate superior colliculus shift with changes in eye position. *Nature* **309**: 345–7

Jiang W, Wallace M T, Jiang H, Vaughan J W, Stein B E 1999 Two cortical areas contribute to the multisensory integration in superior colliculus neurons. *Society for Neuroscience Abstracts* **25**: 1413

Jousmaki V, Hari R 1998 Parchment-skin illusion: Sound-biased touch. *Current Biology* **8**: R190

Kadunce D C, Vaughan J W, Wallace M T, Benedek G, Stein B E 1997 Mechanisms of within- and cross-modality suppression in the superior colliculus. *Journal of Neurophysiology* **78**: 2834–47

Massaro D W, Stork D G 1998 Sensory integration and speech-reading by humans and machines. *American Scientist* **86**: 236–44

Meredith M A, Nemitz J W, Stein B E 1987 Determinants of multisensory integration in superior colliculus neurons: I. Temporal factors. *Journal of Neuroscience* **10**: 3215–29

O'Hare J J 1991 Perceptual integration. *Journal of the Washington Academy of Science* **81**: 44–59
Ramachandran R, Wallace M T, Stein B E 1993 Distribution and properties of multisensory neurons in rat cerebral cortex. *Society for Neuroscience Abstracts* **19**: 1447
Rizzolatti G, Scandolara C, Matelli M, Gentilucci M 1981a Afferent properties of periarcuate neurons in macaque monkeys. I. Somatosensory responses. *Behavioural Brain Research* **2**: 125–46
Rizzolatti G, Scandolara C, Matelli M, Gentilucci M 1981b Afferent properties of periarcuate neurons in macaque monkeys. II. Visual responses. *Behavioural Brain Research* **2**: b 47–63
Sams M, Aulanko R, Hamalainen M, Hari R, Lounasmaa O V, Lu S-T, Simola J 1991 Seeing speech: Visual information from lip movements modified activity in the human auditory cortex. *Neuroscience Letters* **127**: 141–5
Stein B E, Meredith M A 1993 *The Merging of the Senses*. MIT Press, Cambridge, MA
Stein B E, London N, Wilkinson L K, Price D D 1996 Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience* **8**: 497–506
Stein B E, Meredith M A, Honeycutt W S, McDade L 1989 Behavioral indices of multisensory integration: Orientation of visual cues is affected by auditory stimuli. *Journal of Cognitive Neuroscience* **1**: 12–24
Stein B E, Meredith M A, Wallace M T 1993 The visually-responsive neuron and beyond: Multisensory integration in cat and monkey. *Progress in Brain Research* **95**: 79–90
Sumby W H, Pollack I 1954 Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* **26**: 212–5
Wallace M T, Meredith M A, Stein B E 1992 Integration of multiple sensory modalities in cat cortex. *Experimental Brain Research* **91**: 484–8
Wallace M T, Stein B E 1994 Cross-modal synthesis depends on input from cortex. *Journal of Neurophysiology* **71**: 429–32
Wallace M T, Stein B E 1997 Development of multisensory neurons and multisensory integration in cat superior colliculus. *Journal of Neuroscience* **17**: 2429–44
Wilkinson L K, Meredith M A, Stein B E 1996 The role of anterior ectosylvian cortex in cross-modality orientation and approach behavior. *Experimental Brain Research* **112**: 1–10

B. E. Stein, M. T. Wallace, and T. R. Stanford

Cross-pressures: Political

1. The Fundamental Conceptual–Theoretical Structure

1.1 Basic Propositions

Cross-pressure theory has been employed primarily in the analysis of decision-making, asking how the decision-making process is affected by the presence of divergent and conflictual antecedent factors. The theory entered the study of politics in the 1940s and 1950s via the then new quantitative analysis of voting behavior.

As originally formulated, cross-pressure theory predicted three effects on voting behavior: (a) vacillation; (b) delay; and (c) withdrawal. In other words, the cross pressured voter (a) will shift back and forth in his/her voting choice; (b) will delay making a decision; and (c) will completely avoid making a decision and withdraw from participating in the election altogether.

The predictions of cross-pressure theory derive from three basic propositions:

(a) Cross-pressures (motivational inconsistencies), as all conflicts, give rise to psychic tension.

(b) High tension levels and tension increases are experienced as painful and low tension levels and decreases in tension are experienced as pleasurable.

(c) All living organisms (including humans) prefer pleasure to pain. They will act so as to keep tension low, by trying to prevent conflicts from arising. They will also act so as to reduce high levels of tension, by trying to resolve, remove, or disclaim conflicts (including cross-pressures).

These three prepositions, in one form or another, are part of practically all anticonflict theories, including cross-pressure theory. (For a discussion of the anticonflict nature of cross-pressure theory, see Sperlich 1971, pp. 179–221.) Unfortunately, propositions (b) and (c) are factually incorrect.

Proposition (b) embraces as axiomatic that high tension levels and tension increases will necessarily be painful. It ignores first, that adequate proof in support of this proposition has not been developed, and second, that there is considerable evidence to the contrary, to be found in various forms of deliberate stimulus and adventure seeking, voluntary risk-taking and exploratory behavior, as well as in the frequently unpleasant nature of low tension states (stimulus deprivation). For an extended review, see Sperlich 1971, pp. 52–9.

Proposition (c) takes unconstrained hedonism as axiomatic. It lacks any recognition that—however important pleasure may be—it is not the only goal that human beings can, or do pursue. It altogether ignores the vast compendium of philosophical and empirical inquiries into the complexity and multiplicity of human motivations. As examples can serve the various studies of gratification delay, of demonstrations of (costly) loyalty, of the keeping of (burdensome) promises, and of altruistic behavior—all of which involve choosing lower pleasure and/or greater pain (Sperlich 1971, pp. 85–9).

1.2 Externally and Internally Generated Cross-pressures

Several different types and constellations of factors are said to produce cross-pressures in persons, in-

cluding demographic, social, and cognitive (psychic) variables. The first two of these are experienced as external pressures brought to bear on the deciding person; the third factor is internal. Examples of demographic factors are occupation, education, religion, age, and ethnicity. Examples of social factors are memberships in family and friendship groups. (The crucial distinction here is comprehensive personal interaction, lacking for members of social groups.) The internal factors include preferences, attitudes, beliefs, and informational elements.

In respect to the external factors, traditional cross-pressure theory simply assumes what can neither be proven nor automatically granted: that the norms and preferences of the demographic and social aggregates to which the person belongs will generally be subjectively real. But to produce cross-pressures, these orientations first must be perceived, and second must become active elements in the person's decision-making processes. If a given voter belongs to religion A (the norm of which is to vote for party X) and also belongs to ethnic group B (the norm of which is to vote for party Y), he/she must first of all perceive the existence of the groups' majority preferences, and second must incorporate these positions into his/her own decision-making activities. Lack of information about group norms, as well as indifference to these norms, may invalidate the three cross-pressure predictions noted earlier. Traditional cross-pressure theory has neglected to take these matters into account, as it has failed to take into account the issue of causal ambiguity prominent in cross-pressure analyses (compare Lipset 1954, pp. 584–98, vs. Campbell et al. 1960, p. 83)—which is one of the explanations of its predictive failures.

Internal cross-pressures constitute a clearer causal case. It is not a vast conceptual leap to infer the presence of cross-pressures, if the voter states that he/she prefers the economic policies of candidate A, but the foreign policies of candidate B. Yet, other matters should be taken into account. For example, fluctuation, avoidance, and withdrawal are not likely to occur if, in the present example, the voter only cares about economic policies and is wholly uninterested in foreign policy matters. Again, traditional cross-pressure theory ignores the question of the relative strength of the elements in conflict—one more explanation of its predictive failures.

1.3 Cross-pressure Valences

Cross-pressures can have different valences, that is, different evaluative directions. The economic vs. foreign policy dissonance is an example of approach/approach (+/+) conflicts. Of course, it is also possible that the voter dislikes candidate A for one reason and candidate B for another—an example of avoidance/avoidance (−/−) conflicts. However, few voters will perceive only positive or only negative qualities in candidates or parties. The most realistic explication of cross-pressures is the double approach/avoidance (+ −/+ −) scenario, where the voter likes certain aspects and dislikes other aspects of each candidate or party. Or to state the matter in motivational language:

(a) Approach–approach conflict: the conflicting motivations simultaneously push the individual toward (make the individual desire) two incompatible choices, e.g., favoring two candidates, but being able to only vote for one of them.

(b) Avoidance–avoidance conflict: the conflicting motivations simultaneously push the individual away from (make the individual disdain) all available choices, one of which, however, must be chosen, e.g., disliking both candidates, but being obligated to cast one's vote for one of them.

(c) Double approach–avoidance conflict: the conflicting motivations simultaneous push toward, and push away from both candidates, e.g., when the voter is aware of negative as well as positive aspects of each candidate.

It is quite obvious that each of these three types of conflict is likely to be experienced differently and to have different consequences. The avoidance/avoidance conflict, for example, is likely to be more unpleasant—and more likely to produce withdrawal—than the approach/approach conflict. Traditional cross-pressure theory ignores the question of the qualitative nature of the conflict—yet another explanation for its predictive failures.

The potential costs of the errors of cross-pressure theory are not limited to the effects on political science and other academic disciplines. Cross-pressure theory is concerned only with reducing tensions, maintaining low tension states, 'and thereby maximizing pleasure—to be accomplished by eliminating conflicts, disagreements, and inconsistencies. The theory fails to acknowledge higher-level human functions and concerns. It is a partial and rather primitive conception of human nature. Necessarily, the theory is unable to understand and predict actual human behavior. In effect, the theory supports and encourages policies of cultural conformity. It tends to view differences between the individual and the community as problems of individual mental health, to be managed by personal adjustment, and even forcefully corrected by behavioral therapies or similar frights.

Politics and conflict are closely related. Politics is the conflicts and tensions between the various groups and their interests, as well as the efforts to find workable compromises. The concept of freedom makes little sense unless one has the ability and right to differ. Cross-pressure theory, however, regards diversity and its conflicts as societal dysfunctions. The implicit preference is for a conflict-free and harmonious system. What the theory fails to understand is that complete harmony and uniformity can only be achieved (if at all) by force and oppression.

2 *History of Cross-pressure Analysis*

Cross-pressure theory entered political science via the analysis of voting behavior at Columbia University (Lazarsfeld et al. 1948, Berelson et al. 1954) and the University of Michigan (Campbell et al. 1954, Campbell et al. 1960). The basic notion of cross-pressures, however, is considerably older. The idea that persons and other living entities, who are pushed (or pulled) in two (or more) directions at once, and hence are burdened by special cognitive and behavioral difficulties, is ancient. It can be found, for example, in Aesop's fable of the donkey starving to death between two equidistant piles of hay. In the more recent academic literature, the concept of cross-pressures can be traced to the writings of Georg Simmel. Simmel (e.g., 1921), however, did not employ the German equivalent of 'cross-pressures.' He used '*Kreuzung sozialer Kreise*,' which strictly translated means 'the crossing of social circles.' It was Bendix (in Simmel 1955), who translated Simmel's phrase into 'cross-pressures.'

The general idea of conflicting pressures received systematic attention almost simultaneously in political science and sociology, as well as in psychology and social psychology. In sociology the theory was stated mostly in terms of 'status equilibrium' and 'status crystallization' (Lenski 1961, 1964). The psychological fields generated the concept and theory of *cognitive dissonance* (e.g., Festinger et al. 1956, Festinger 1957. Cognitive dissonance theory has had a more illustrious career than its political-science and sociological equivalents. Indeed, the concept of cognitive dissonance has entered popular speech, even political campaigns. There has been serious and significant work in respect to cognitive dissonance, including sustained attention to systematic testing, improved measurement, and theoretical refinement of the propositions of cognitive dissonance (Davis 1993, Rabin 1991, Shultz and Lepper 1996). Such efforts are largely lacking in respect to cross-pressures. Only a single monograph exists exploring the nature, accuracy, and utility of cross-pressure theory (Sperlich 1971).

After an initial flurry of interest in studies of voting behavior—where it generally failed to produce confirmed predictions and useful explanations (see Sperlich 1971, pp. 9–26)—cross-pressure theory has fallen into almost universal disuse, even in voting behavior studies. Benney (1956), for example, provided a detailed analysis of social class voting, of multiple influences on voting choice, and of those who changed their voting intention during the campaign. He also used the term 'group pressure'—not, however, any real cross-pressure analysis. Similarly, Butler and Stokes (1974) spoke occasionally about 'cleavages [that] cut across the electorate,' but did not engage in actual cross-pressure analysis in their study of British voting behavior. Books dealing explicitly with conflict show a similar neglect (for example, Lipset 1985, Porter and Taplin 1987, Reilley and Sigall 1976), as do works in most other sub-fields. Kourvetaris (1997), for example, presents a conflict model of political sociology, but makes no use of cross-pressures concepts; psychologically oriented examples include; Freedman and Freedman 1975, Renshon 1974, Stone 1974, Moran 1985), rejects what he calls a 'civility model' and develops a 'conflict model' instead, dealing extensively with social cleavages, such as those associated with class, religion, race, and nationality.

Moran's book seemed well suited for the use of the idea of cross-pressures—but it did not make any such use. Similarly, Mayer (1977) included discussions of class conflict, cognitive dissonance, and cross-cutting vs. segmental cleavages, but not of cross-pressures (for other examples see: Almond and Verba 1963, Duverger 1972, and Meny and Knapp 1998). Rokkan (1970) developed a 'cleavage model' of politics without the use of cross-pressures. Numerous general discussions of politics, political systems, and political behavior entirely, or largely, omit discussions of cross-pressures (e.g., Alker 1973, Bellamy 1993, Delury 1987, Eulau 1996, Needler 1996, and Polsby et al. 1963).

3. *Disuse and Problems*

Upon closer inspection, the extensive non-use of cross-pressure theory is not surprising. It is theoretically inadequate. Two of its three basic propositions are in error. It ignores the problems of causal directionality and subjective reality when dealing with external cross-pressure factors. It has developed no measurement of the relative strength of a cross-pressure conditions, only recognizing a dichotomized presence or absence of cross-pressures. Further, it has paid no attention to the valence of the conflicts.

Proposition (b) is in error when it assumes that high tension and tension increases are invariably unpleasant. (Though, of course, some high-tension situations in respect to voting may indeed be experienced as unpleasant, particularly if the person is already subject to high tension levels in other aspects of his/her existence, and if the conflict is of the $-/-$ variety.) Proposition 2 needs to be revised as follows:

(a) A given tension level may be experienced as comfort or discomfort, depending on the tension requirement of the organism at the time; appropriate tension magnitudes will be experienced as comfortable, inappropriate ones (not enough, as well as too much tension) as discomfort.

(b) A change (increase or decrease) in the tension level may be experienced as pleasure or pain, depending on the direction of the change in respect to the tension requirements of the organism at the time; changes in the direction of the appropriate tension magnitude will be experienced as pleasurable, changes away from that magnitude as painful.

Proposition (c) is in error when it assumes that hedonism (maximizing pleasure and minimizing pain) is the only existing human motivation. It needs to be revised as follows:

(a) Human beings aim at a large variety of goal states and their actions are guided by a multiplicity of motives. Maximization of pleasure and minimization of pain are important, but are not the sole aims of human striving. There are acts to which the pleasure/pain calculus is irrelevant.

(b) The more dominant in any given situation the aims and motives associated with ego and conscience, the less will behavior be guided by organismic (id) demands for pleasure and gratification, such as expressed in attempts to resolve all conflict.

Research has shown that as long as the theory fails to take into account causal directionality, subjective reality, the strength of the cross-pressure conditions, and the valence of the conflicts, it cannot generate correct predictions of human behavior (Sperlich 1971, pp. 9–26) The introduction of better measures and the adoption of the necessary theoretical revisions, however, have been shown to yield highly successful predictions (Sperlich 1971, pp. 90–168).

See also: Cleavages: Political; Cognitive Dissonance; Participation: Political; Party Identification; Pluralism; Social Stratification; Voting, Sociology of; Voting: Turnout

Bibliography

Alker H R, Deutsch K W, Stoetzel A H 1973 *Mathematical Approaches to Politics*. Jossey–Bass, San Francisco

Almond G A, Verba S 1963 *The Civic Culture*. Princeton University Press, Princeton, NJ

Bellamy R (ed.) 1993 *Theories and Concepts of Politics*. Manchester University Press, UK

Benney M 1956 *How People Vote*. Routledge & Kegan Paul, London

Berelson B, Lazarsfeld P F, McPhee W N 1954 *Voting*. University of Chicago Press, Chicago

Butler D, Stokes D 1974 *Political Change in Britain. The Evolution of Electoral Choice*. 2nd edn., Macmillan, New York

Campbell A, Gurin, G, Milter W E 1954 *The Voter Decides*. Row, Peterson, Evanston IL

Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. John Wiley, New York

Davis W L 1993 Economists' uses for cognitive dissonance: An interdisciplinary note. *Psychological Reports* **73**

Delury G E (ed.) 1987 *World Encyclopedia of Political Systems and Parties*. 2nd edn. 2 Vols. Facts on File Publications, New York

Duverger M 1972 *The Study of Politics*. Thomas Y. Crowell, New York

Eulau H 1996 *Micro-macro Dilemmas in Political Science*. University of Oklahoma Press, Norman, OK

Festinger L 1957 *A Theory of Cognitive Dissonance*. Stanford University Press, Stanford, CA

Festinger L, Riecken H W, Schachter S 1956 *When Prophecy Fails*. Harper & Row, New York

Freedman A E, Freedman P E 1975 *The Psychology of Political Control*. St. Martin's Press, New York

Gurr T R (ed.) 1980 *Handbook of Political Conflict*. Free Press, New York

Kourvetaris G A 1997 *Political Sociology*. Allyn and Bacon, Boston

Lazarsfeld P F, Berelson B, Gaudet H 1948 *The People's Choice*. Columbia University Press, New York

Lenski G 1961 *The Religious Factor*. Doubleday, New York

Lenski G 1964 The four socio-religious groups. In: Raab E (ed.) *Religious Conflict in America*. Doubleday, New York

Lipset S M 1954 Opinion formation in a crisis situation. In: Katz D (ed.) *Public Opinion and Propaganda*. Dryden Press, New York

Lipset S M 1985 *Consensus and Conflict: Essays in Political Sociology*. Transaction Books, New Brunswick, ME

Mayer L C 1977 *Politics in Industrial Societies: A Comparative Perspective*. John Wiley, New York

Meny Y, Knapp A 1998 *Government and Politics in Western Europe*. Oxford University Press, New York

Moran M 1985 *Politics and Society in Britain*. Macmillan, London

Needler M C 1996 *Identity, Interest, and Ideology*. Praeger, Westport, CT

Polsby N W, Dentler R A, Smith P A 1963 *Politics and Social Life*. Houghton Mifflin, Boston

Porter J N, Taplin R 1987 *Conflict and Conflict Resolution*. University Press of America, Lanhom, MD

Rabin M 1991 *Cognitive Dissonance and Social Change*. Working Paper, University of California, Berkeley, CA

Reilly T A, Sigall M W 1976 *Political Bargaining*. Freeman, San Francisco

Renshon S A 1974 *Psychological Needs and Political Behavior*. The Free Press, New York

Rokkan S 1970 *Citizens, Elections, Parties*. David McKay

Shultz T R, Lepper M R 1996 Cognitive dissonance reduction as constraint satisfaction. *Psychological Review* **103**

Simmel G 1921 *Der Konflikt der modernen Kultur*. Duncker and Humblot, Berlin

Simmel G 1955 *Conflict and the Web of Group-affiliations*. Free Press, New York

Sperlich P W 1971 *Conflict and Harmony in Human Affairs: A Study of Cross pressures and Political Behavior*. Rand McNally and Co., Chicago

Stone W F 1974 *The Psychology of Politics*. The Free Press, New York

P. W. Sperlich

Crowding and Other Environmental Stressors

We typically describe environmental problems as endangerments to physical health amenable to technological solutions. Yet human behavior occupies center stage for virtually every major environmental

problem confronting humankind. On the one hand, both personal and societal decisions and actions are the root cause of most ecological problems, and on the other, environmental quality can affect psychological health and functioning. The focus of this entry is on the impacts of environmental quality, specifically crowding, noise, and pollution upon psychological functioning.

The construct of psychological stress (see *Stress: Psychological Perspectives* for further details) has proven useful for understanding how the physical environment can affect human well-being. Psychological stress refers to behavioral and physiological reactions to environmental demands that exceed or tax the adaptive capacities of the organism. These demands are termed stressors. Adaptive capacities of the organism encompass genetic predispositions, cognitive and behavioral coping strategies, and social characteristics (e.g., social support) that can alter the balance between environmental demands and individual capacity.

1. Psychological Consequences of Crowding, Noise, and Air Pollution

The primary index of crowding is residential density (people/room). Other measures of crowding, particularly those at a larger scale, such as people per acre, have proven inconsequential for health and behavioral outcomes. This is one reason why demographic and epidemiological approaches to crowding and health have largely proven futile—for example, density statistics based upon census tract data are typically uncorrelated with a host of health outcomes, including disease, mental illness, and crime (Evans in press). Use of aggregate level density indices also leads to misleading cross-national comparisons of crowding and pathology. As an illustration, comparisons of crime rates across countries reveal lower crime rates in Hong Kong in comparison to the USA even though the number of people per acre in the USA is markedly lower than people per acre in Hong Kong. On the other hand, people per acre in Manhattan is not lower. Most likely it is the immediate experience of unwanted or uncontrollable social interaction that lies at the root of the crowding experience. This can be illustrated by the following dilemma: Would you rather share a very large room with one other person or have a small, even cramped room for yourself? Exposure to traffic congestion may be a related form of crowding stress largely caused by interference with progress towards one's destination. In this case exposure to peak traffic levels rather than average levels of congestion is decisive.

Noise, which is defined as unwanted sound, is usually measured in terms of sound intensity metrics such as decibels (dBA). Both average levels of sound intensity (Leq or Ldn) as well as peak levels of sound intensity exposure appear to be consequential. Leq is an arithmetic average, and Ldn is a weighted average incorporating a night time noise exposure penalty. Ldn is the most typical index of noise exposure used around airports in the USA. It is important to note that decibels is a logarithmic scale. An increase of 10 dBA is perceived as approximately twice as loud. In terms of hearing loss, the duration of exposure is as important if not more so than the actual level of sound intensity, unless extreme. Unfortunately, the implications of exposure duration for stress effects of noise have received insufficient attention.

Air pollution exposure is measured by atmospheric concentrations of known pollutants (e.g., sulfur dioxide) or toxins that are a major constituent of pollution (e.g., ozone as a component of photochemical smog). Both average levels of exposure as well as peak exposure levels are of importance for human well-being.

Precise estimation of individual exposure to environmental stressors like crowding, noise, and air pollution has proven difficult because of spatial and temporal mobility of human samples. Ambient air pollution illustrates well the challenge of accurate exposure estimation. Not only do people move around during the day, they spend time commuting as well as transferring back and forth between interior and outdoor settings. Each of these factors alters personal exposure to ambient pollutants.

For all three environmental stressors, human psychological functioning has been assessed across a range of experimental studies in the laboratory, cross-sectional field studies and, in a few instances, longitudinal studies with assessments of changes in exposure associated with outcomes of interest. Therefore caution is warranted in interpreting the conclusions described here since many of them are based primarily on acute, short term laboratory studies, or on cross-sectional field comparisons. On the other hand, convergence between laboratory and field studies is generally high (Evans 2001).

1.1 Psychophysiological Stress

Both laboratory and field research indicate that high density elevates blood pressure and neuroendocrine indices (e.g., epinephrine) of stress. Trends suggest that these effects are heightened for males and can occur in children as young as six years of age (Baum and Paulus 1987, Evans 2001). Parallel psychophysiological data have been found in field studies of traffic congestion.

Acute noise elevates cardiovascular and neuroendocrine parameters, but these effects rapidly habituate unless experimental participants are placed under moderately intense physical or cognitive

demands. There is also some evidence that noise-sensitive individuals do not readily habituate to acute noise exposure. Cardiovascular data from occupational noise exposure are inconsistent and lie on a foundation of poorly designed studies rendering definitive conclusions impossible. Better designed community studies show modest elevations in blood pressure associated with airport noise exposure. Interestingly, these psychophysiological effects do not habituate with longer exposure duration. Most of the community noise studies have examined children attending elementary schools in airport noise-impact zones. Few community studies have examined adults or neuroendocrine parameters (Evans 2001).

Research on air pollution as a stressor has not examined psychophysiological markers of stress. Understandably most health research on pollutants has focused on respiratory functioning or, in the case of toxins, neurological parameters.

1.2 Cognitive Performance and Learning

Short-term exposure to crowding in the laboratory impairs performance under a limited set of conditions. Performance is degraded only when substantial cognitive load is placed upon the individual, for example by having to monitor multiple signals or to respond to rapid or ambiguous signals (Baum and Paulus 1987). Parallel results have been found for noise (Smith and Jones 1992). Experimental work on pollutants and task performance has primarily been restricted to neurotoxicological testing.

Residential crowding is associated with achievement deficits among school children (Evans et al. 1998, Saegert 1982), and numerous studies have shown associations between chronic noise exposure and reading deficits (Evans and Lepore 1993). The noise studies include some longitudinal evidence, dose-response data, and some work on underlying cognitive processes to account for the findings. Speech perception and language acquisition appear to be negatively impacted by chronic noise exposure. Most of the noise and reading studies ruled out hearing impairment as an alternative explanation.

1.3 Interpersonal Relationships

Crowding induces negative affect, heightens irritability and hostility, and may provoke aggression (Baum and Paulus 1987). Residential crowding is associated with more conflicted family relationships including more corporal punishment (Evans et al. 1998, Saegert 1982). Noise is an irritant that, when coupled with provocation or modeling, can lead to elevated aggression (Cohen and Spacapan 1984). Noise itself does not produce aggression but can amplify hostile reactions already present. Attempts to associate crowding or noise exposure with crime have generally proven unsuccessful. There is also evidence that crowding (Baum and Paulus 1987) and noise exposure (Cohen and Spacapan 1984) can suppress altruism.

Perceptible air pollution can increase negative affect and, if not at high levels, elevate aggression (Evans 1994). The nonlinear relation between air pollution levels and aggression may reflect underlying negative affect (Rotton 1983).

1.4 Psychological Health

Given both experimental and field evidence linking environmental stressors to negative affect and conflicted interpersonal relationships, many investigators have endeavored to demonstrate associations between chronic exposure to environmental stressors and mental health. For all three environmental stressors, the data are mixed, although careful scrutiny reveals some consistent patterns. Studies of residential crowding and mental health that have relied upon individual-level data reveal modest relations. Studies of more serious outcome measures of psychiatric impairment as well as those relying on aggregate-level indices of crowding (e.g., people/acre) typically find no assoication with mental health (Evans 2001).

The data on noise and mental health are more equivocal, with no discernable trends, although many studies have used poor estimates of noise exposure and relied upon catastrophic mental health outcome measures (e.g., psychiatric hospitalization). There is some evidence that those who are noise-sensitive, which is related to neuroticism, may be especially vulnerable to elevated psychological distress in relation to noise exposure (Stansfeld 1993). Air pollution exposure may also be related to psychological distress, although not much data have been collected (Evans 1994, Rotton 1983).

1.5 Summary

Acute and chronic exposure to crowding or noise can elevate cardiovascular and neuroendocrine indicators of stress. Insufficient research has examined similar reactions to pollution. Short-term exposure to crowding or noise can impair performance when cognitive load is high. Chronic crowding has been linked to achievement deficits among children in a few studies, and numerous studies have revealed noise-related reading deficits. Crowding, noise, and air pollution, both in the laboratory and under naturalistic conditions, increase annoyance, irritability, and hostility. Aggression may be more likely to occur when prior provocation or modeling of conflict has occurred. In the case of pollution, evidence suggests that the

relation is nonlinear and may be mediated by negative affect. Residential crowding and exposure to pollutants appear to be associated with elevated psychological distress. The evidence for mental health sequels of noise exposure is equivocal.

2. Underlying Mechanisms

The initial phase of behavioral research on environmental stressors was focused on documenting potential adverse effects on human functioning along with examining possible individual differences in responses. Several reliable gender differences have been uncovered for crowding, suggesting that men react more strongly physiologically, whereas women may suffer stronger negative affective responses. Noise-sensitive individuals tend to react more negatively to noise in a variety or ways in comparison to their less sensitive counterparts. Attempts to identify systematic personality or dispositional differences in reactions to environmental stressors have generally been unsuccessful. The strongest data to date suggest that persons seeking higher levels of arousal or sensation are less adversely affected by crowding and noise (Mahrabian and Russell 1974).

Current research has shifted emphasis, focusing primarily on underlying, explanatory psychosocial processes to help account for the apparently stressful, adverse effects of environmental stressors on human well-being. The two underlying psychosocial processes that have garnered the most attention to date are control and social support.

2.1 Control

Research on all three environmental stressors implicates lack of control as an important component of stress reactions. Two general approaches have led to this conclusion. In the first approach, studies of noise (Cohen et al. 1986), crowding (Baum and Paulus 1987), and air pollution (Rotton 1983) have shown that the negative impacts of each of these stressors are substantially attenuated by providing individuals with a sense of control over the stressor. For example, in a classic set of experiments, Glass and Singer (1972) demonstrated that the negative effects of noise were essentially eliminated if subjects believed they could shut off the noise with a switch. Just the perception of control was sufficient to ameliorate the harmful impacts of noise, regardless of whether subjects actually exercised the control option.

The second approach to examining control as an underlying explanatory process for the stress effects of environmental stressors has been to study motivational deficits associated with learned helplessness among those chronically exposed to environmental stressors. As explained in more detail in the entry *Learned Helplessness*, repeated exposure to uncontrollable stimuli can lead to decrements in motivation caused by the belief that the outcomes of one's actions are independent of those actions. Children chronically exposed to noise or to crowding are less likely to persist in their attempts to solve challenging puzzles (Evans 2001). Adults living under more crowded conditions adapt a withdrawal strategy in group game-playing scenarios. Their degree of withdrawal is closely matched by self-reports of feelings of loss of control over the immediate social environment in their residences (Baum et al. 1981). Physiological stress reactions to traffic congestion have also been shown to be partially attributable to diminished perceptions of control (Evans 2001).

2.2 Social Support

Several studies indicate that residential crowding is associated with diminished social support (Evans 2001). This is believed to occur as a consequence of a common coping strategy for dealing with crowding, namely social withdrawal. Long-term crowding appears to precipitate social withdrawal which, in turn, leads to a breakdown of socially supportive relationships in the home. There is also evidence that deteriorated social support mediates the linkage between residential crowding and psychological distress.

Analogous research has not been completed with noise or air pollution, but there are reasons to suspect that similar processes may be operating. Both of these stressors are clearly irritating and can affect one's propensity for negative social interaction. Furthermore, noise interferes with communication, a fundamental building block of socially supportive relationships.

3. Future Directions

From both a methodological and a policy perspective, there is critical need for prospective longitudinal field studies of environmental stressors and human psychological functioning. Ideally these studies will also generate dose-response functions in order to understand critical values of environmental stressors for human behavior. At the present we know environmental stressors are inimical to optimum psychological functioning, but we cannot say currently precisely how much exposure to a particular stressor is sufficient to produce harm.

Research is likely to continue on underlying processes to explain how and why environmental stressors can interfere with human behavior. Hopefully this kind of research can also incorporate better estimation of human exposure to environmental stressors. Ambulatory monitoring and use of time budget methodologies may prove helpful in this regard.

The contexts in which exposure to environmental stressors occur warrant greater attention. Exposure to environmental stressors is not random. Most of these

stressors are experienced jointly (e.g., noise and crowding) and frequently accompany poverty. More psychological analysis of multiple environmental and social stressors exposure is called for.

The social and behavioral sciences will continue to provide insights that expand our understanding of environment and health. Psychological stress provides one theoretical framework that has proven useful for doing so. Our attitudes towards the natural environment, along with models of ecologically responsible behavior, are two additional topics where the social and behavioral sciences are likely to continue making important contributions to the environmental sciences. Human behavior plays a pivotal role both in the etiology of environmental problems and as an index of their impacts.

See also: Community Environmental Psychology; Environmental Adaptation and Adjustments; Environmental Challenges in Organizations; Environmental Psychology: Overview; Environmental Stress and Health; Stress and Coping Theories; Stress and Health Research; Stress: Measurement by Self-report and Interview; Stress: Psychological Perspectives

Bibliography

Baum A, Gatchel R, Aiello J R, Thompson D 1981 Cognitive mediation of environmental stress. In: Harvey J (ed.) *Cognition, Social Behavior and the Environment*. Erlbaum, Hillsdale, NJ, pp. 513–31

Baum A, Paulus P B 1987 Crowding. In: Stokols D, Altman I (eds.) *Handbook of Environmental Psychology*. Wiley, New York, pp. 533–70

Cohen S, Evans G W, Stokols D, Krantz D S 1986 *Behavior, Health, and Environmental Stress*. Plenum, New York

Cohen S, Spacapan S A 1984 The social psychology of noise. In: Jones D M, Chapman A J (eds.) *Noise and Society*. Wiley, New York, pp. 221–46

Evans G W 1994 The psychological costs of chronic exposure to ambient air pollution. In: Isaacson R L, Jensen K F (eds.) *The Vulnerable Brain and Environmental Risks: Vol. 3: Toxins in Air and Water*. Plenum Press, New York, pp. 167–84

Evans G W 2001 Environmental stress and health. In: Baum A, Revenson T, Singer J E (eds.) *Handbook of Health Psychology*. Erlbaum, Hillsdale, NJ, pp. 365–85

Evans G W, Lepore S J 1993 Nonauditory effects of noise on children: A critical review. *Children's Environments* **10**: 31–51

Evans G W, Lepore S J, Shejwal B R, Palsane M N 1998 Chronic residential crowding and children's well being: An ecological perspective. *Child Development* **69**: 1514–23

Glass D C, Singer J E 1972 *Urban Stress*. Academic Press, New York

Mehrabian A, Russell J 1974 *An approach to Environmental Psychology*. MIT Press, Cambridge, MA

Rotton J 1983 Affective and cognitive consequences of malodorous pollution. *Basic and Applied Social Psychology* **4**: 171–91

Saegert S 1982 Environment and children's mental health: Residential density and low income children. In: Baum A, Singer J E (eds.) *Handbook of Psychology and Health*. Erlbaum, Hillsdale, NJ pp. 247–71

Smith A P, Jones D M 1992 Noise and performance. In: Smith A P, Jones D M (eds.) *Handbook of Human Performance, Vol. 1: The Physical Environment*. Academic Press, London, pp. 1–28

Stansfeld S A 1993 Noise, noise sensitivity, and psychiatric disorder: Epidemiological and psychophysiological studies. *Psychological Medicine: Monograph Supplement* **22**: 1–44

G.W. Evans

Crowds in History

Crowds do play an important role in history. Particularly during times of crisis and revolution, they can function as prime movers in historical–political conflicts. As a constituent feature of everyday public life, the omnipresent crowd is a highly complex phenomenon, especially in modern urbanized societies. In view of the innumerable forms that crowds take, in the most varied of contexts, for pragmatic reasons it is sensible to distinguish between: (a) more trivial everyday crowds as well as crowds of a recurring, institutionalized, and routine nature; and (b) exceptional, unforeseeable crowd events.

For research purposes, crowds can be defined as relatively short-term gatherings of large groups of people (on streets or squares or in the countryside—in any event, in the open air), whose actions are goal-oriented and, as a rule, conflictual. Crowds are assemblages that interrupt the regular course of everyday life and disturb the public order, or at least threaten to disturb it. Under certain circumstances, of course, an everyday or routine crowd can also transform itself into one that will be conflictual.

Crowd history was (and will likely remain) primarily indebted to the social history paradigm: theory-oriented, but seldom fixated on theory—in fact, with a fair bit of skepticism regarding theory; and utilizing a variety of empirical methods, singularly or in combination, from 'thick description' to extensive statistics on protest. Central concerns are the precise reconstruction of crowd events, the involved groups and their identities, and a micro-historical embedding in local contexts. Dominating questions concern exact times, size, duration, frequency, and topography or geography of crowds. Research topics are their physical composition (social status, ethnic identity, religion, age, sex) and the—considerably heterogeneous—behavior of individuals or smaller groupings in these events.

Other focuses are the forms and motives of action (violent tendencies, emotionality, resentment, mentalities, visions), in addition to occasions and possible structural reasons for common action. Furthermore, crowd history investigates the targets of crowd action, as well as the presence and function of peacekeepers in such confrontations. And last but not least, crowd

history discusses the results and long-term consequences of crowd events in consideration of high politics and society.

1. Phases and Schools: Changing Images of Crowds

Academic crowd research was first begun toward the end of the nineteenth century, and its beginnings are conspicuously multidisciplinary: philosophy, psychology, criminology, sociology, and anthropology occupied themselves with the phenomenon; only gradually did historical studies turn their attention to the subject. Against the background of opposing political–ideological positions, several academic schools disputed the 'true essence' or the 'actual character' of crowds. Depending on one's position, interpretations ranged from an extreme damning of crowds as purely negative and regressive destroyers of civilization, to their enjoying an idealized historical–philosophical role as genuine creative agents, as essential driving forces of historical progress. From the historian's perspective, there are at least four important phases.

The starting point was the so-called 'Latin School' of mass psychology around 1900, represented by such authors as the Italian criminologist Scipio Sighele (*La Folla Delinquente*, 1891), the French popular scientist Gustave LeBon (*La Psychologie des Foules*, 1895), and the French criminologist and sociologist Gabriel Tarde (*L'Opinion et la Foule*, 1901). Relying for their historical evidence primarily on one-sided studies of the French Revolution by the conservative historian and philosopher Hippolyte Taine, their speculative theories concerning crowds were a projection of bourgeois-conservative fears of social subversion at the *fin de siècle*: 'Their crowds loomed as violent, bestial, insane, capricious beings whose comportment resembled that of mentally ill, women, alcoholics, or savages' (Barrows 1981 p. 5).

The emerging social–historical school of French Revolution research opposed the conservative LeBon mass psychology by de-demonizing the crowd debate through empirical case studies addressing the character and role of crowds during the French Revolution. One of this school's path-breaking authors was Georges Lefebvre. In *La Grande Peur* (1932) and especially in *Foules Révolutionnaires* (1934), he showed how rebellious groupings were stimulated through various types of oral communication (e.g., spreading of rumors, suspicions, and mental images), and how the *foules pures* of everyday life transmogrified into the *foules révolutionnaires* when seized by a 'collective mentality.'

After the World War II, a modern classic school of crowd research was constituted that traced its ancestry partly to the social–historical school of French Revolution research. The classic school's first important propagator was an occasional student of Lefebvre's, George Rudé, whose 1950 doctoral dissertation treated the Paris workers and their uprisings of 1789–91. In his studies on crowds in the French Revolution and on crowds generally in France and England from 1730–1848 (Rudé 1964), he identified for the first time the 'faces in the crowd.' Together with Eric J. Hobsbawm's early studies of the pre- and early-industrial workers groups whose rebellious behavior was plausibly labeled 'collective bargaining by riot' (Hobsbawm 1952), as well as the related work of Edward P. Thompson, especially his paradigmatic study of *The Moral Economy of the English Crowd in the Eighteenth Century* (Thompson 1971), the classical canon of a new history from below emerged. These historians understood themselves to be an open and undogmatic variant of a Marxist historiography of the West. At first an oppositional subdiscipline, from the 1960s to the 1990s it quickly gained a wide reputation in academic circles, inspiring numerous other studies.

There was a parallel development in the American social sciences, using modernization theory and quantitative measurements in historical–sociological research on crowds. Crucial was a more systematic investigation of possible causal connections between large processes such as industrialization, urbanization, or social crises, and the frequency and geography of crowd occurrences. In a critical test of theory, then-current concepts of collective behavior and of social movement research (relative deprivation, among others) were checked against historical evidence. This led to more refined methods of crowd event analysis, accumulating more precise protest statistics over longer periods and/or testing its geographic diffusion for strategically chosen regions. Exemplary is the work of Tilly et al., in particular the comparative study of France, Italy, and Germany in *The Rebellious Century* (1975), or the strictly quantitative and simultaneously Thompsonian case study of English riots 1790–1810 (Bohstedt 1983).

2. The Unwritten Program: 'Subsistence' and 'Liberties'

For the longest time, European societies of the early-modern period, including the nineteenth century, figured as the classic time frame for the bulk of crowd research. The essential scope of crowd activities was above all defined by the secular processes of state centralization and national state-building, the expansion of market economies, social polarization, and proletarianization. Viewed from the lower-class angle, one is able to speak of a European age of persistent material shortages. Under conditions of subjection and/or political immaturity, the lower classes found themselves in permanent conflict with authorities and social elites for subsistence guarantees, communal rights and freedoms, and cultural autonomy.

Rudé's and Hobsbawm's 'new picture of the crowd' was and remains largely a counter-image to the previously dominant LeBon tradition. Crowds, it was now shown, were not—as a rule—criminal rabble but people of established residence, average working men and women, poor but orderly folk. Their motivations were a frequent mixture of Jacobin egalitarianism and conservative monarchism. Their collective behavior was one of direct action, and they pursued their thoroughly rational goals with fitting means, this sometimes translating into violence. However, it is said that their violence was predominantly symbolic and a question of efficacy. Moreover, their successes were of limited nature, particularly as their actions were so short-lived and lacking of far-reaching visions. Thus, conflict crowds were to be considered forerunners of later social and political movements with lasting organizational structures and long-term social–political programs.

These and related studies established a solid canon of conflicting crowds, a European repertoire of riots. These included: peasant uprisings against noble estates, anti-tax protests against mounting state prerogatives, religious riots between competing confessions, subsistence-based actions of the rural and urban lower classes, moral rebukes in the tradition of *charivari* (rough music), strikes and boycotts of artisans and workers, machine-wrecking, but also aggressive acts of popular monarchic loyalty against alleged enemies of king and church, and xenophobic outbursts against 'strangers.' As the latest research for the UK, France, and Germany has shown, in no other early-modern conflict of any duration were crowds so involved as in food riots.

More than 3,000 food riots have been documented for France between 1560–1860. At their center were the blockades of inter-regional grain trade to prevent scarce resources from leaving their place of origin. Participating in these blockades were mainly the rural and urban lower classes from the highly commercialized farming regions of the far-reaching agrarian hinterlands of Paris. For four decades, ever since 1760, the power struggle over access to limited food resources became permanent in France. The attempt of Controller General Turgot in the fall of 1774 to deregulate the national grain trade according to physiocratic liberal doctrines, ended in the fiasco of the Guerre des Farines (Flour War) of 1775, when over 300 food rebellions forced the government to back down its policy (Bouton 1993).

There have been about 1,000 food riots documented for England in the period 1530–1820, the majority of them taking place in the last 80 years of that time span. With regard to the years around 1800, Bohstedt speaks of a communally anchored 'protocol of riot,' of 'orderly disorders,' with widely accepted public rules for all of the participants. With crowd action as an essential element of community politics, poorer consumer groups, especially in small and medium-sized market towns, participated successfully in the creation of reasonably tolerable living conditions (Bohstedt 1983).

For German regions (1700–1870) over 400 food conflicts have been substantiated. They culminated in the 1840s (shortly before and during the Revolution of 1848), a decade that has been characterized as the 'angry forties.' Particularly in Prussia's East-Elbian provinces, the eighteenth-century myth of a paternalistic 'social monarchy' quickly faded. Market-driven resource scarcities, not least the result of agricultural exports to Western Europe, forced a sharp internal 'dissent about markets.' Even more so than the urban lower classes, the rural poor frequently saw themselves robbed of whole food entitlements (Gailus 1990).

All in all, these recurring pre- and early-industrial mass conflicts attested to not only a simple struggle for calories among the abject poor and hungry, but symbolized a central conflict of values and goals. On the one side were liberal principles and Utopias envisioning abundant markets and property for all, while opposed to it were nontransferable subsistence rights and the blows to the human dignity of rapidly expanding market-dependent classes in an increasingly urbanizing market society.

Probably no other model has so persuasively explained the social and cultural meaning of pre- and early-industrial crowd activities as Edward P. Thompson's classic notion of moral economy, which he developed through an analysis of the rebellious behavior of the working poor in eighteenth-century England. According to his study, the customs and rituals practiced by the crowds in their numerous food riots, their angry outbursts and anxious visions, their claims and demands, were embedded in 'a popular consensus as to what were legitimate and what were illegitimate practices in marketing, milling, baking, etc. This [consensus] in its turn was grounded upon a consistent traditional view of social norms and obligations, of the proper economic functions of several parties within the community' (Thompson 1971 p. 79). For Thompson the word 'riot' is too narrow a concept to adequately describe all the issues at stake in these confrontations. In order to assure their 'way of survival,' crowds (or people) were in permanent collective negotiations with aristocracy and gentry, with officials and the middle classes. The crowds would typically remind the authorities of their traditional paternalistic duties vis-à-vis the weaker members of the community. In the ritualized forms of conflict with ruling elites, in expressive and symbolic confrontations of political 'theatre' and 'counter-theatre,' they actively participated in the social organization of their own living conditions and thus functioned as an influential collective subject. Crowds were exercising direct influence on the epoch's balance of power, which was shaped by the political hegemony of the gentry (Thompson 1991). One can with good reason speak of 'participation by riot' or of popular

'participation before participation.' Despite some critical objections, the moral economy model has asserted itself as the most successful interpretive concept for historical studies of conflict crowds—and not solely for food riots and for European conflict patterns. More recent work attest to its yet fertile analytic potential (Randall and Charlesworth 1996).

Focusing the implications of these old European crowds on a few essential aspects would above all include 'subsistence' (entitlements to food, rights of social share) and 'liberties' (in the sense of additional rights and popular customs). This collective establishing of values was both asserted and defended against outside 'disturbances' and all actual and imagined threats to existence through dynamic penetration of the two great powers, the national state and the expanding modern market. This culture of resistance of conflicting crowds has contributed in manifold ways to the formation of basic political and social rights of modern constitutional states.

3. *Conceptual Constrictions—Widening of Perspectives*

The advance of the modern social–historical school (1960–90) had a lasting effect on the historical image of crowds: the 'bad crowds' of the older conservative orthodoxy turned into the 'good crowds,' possessing a predominantly positive, emancipating function. However, completely or partially obscured were other aspects of crowd history: aggressive, conservative, and nationalistic or chauvinistic mass events, as well as excessively violent and ethnically motivated outbursts. Also criticized has been crowd history's exclusive concentration on rebellious crowds. Too little attention has been given to other forms of nonconflicting crowds. In addition, the 'great watershed' of 1850, allegedly occasioning a transformation of crowds, has hardly been proven. Despite a sneaking institutionalization of crowds over the following decades, they have remained critically important in the industrial societies of the late nineteenth and twentieth centuries (Holton 1978, Harrison 1988).

Crowd history has to pay more attention to violent actions against minorities and outsiders. Excessive crowd violence was mostly reserved for pogrom-like persecutions of the Jewish minorities. These assumed an almost nation-wide character at least twice in nineteenth-century Germany: in the form of 'Hep-hep riots' (1819) against the political emancipation of the Jews, and the forcible expulsion of south German village Jews within the wider context of the rural mass movements of the 1848 revolutions (Rohrbacher 1993). As of the second half of the nineteenth century, pogroms established themselves in eastern Europe with much greater militancy—in Russia, especially the Ukraine, and in parts of Poland—its chief phases encompassing the periods 1881–84, 1903–06, and 1919–21. Until the end of the nineteenth century these actions were perpetrated by crowds of similar compositions to those in other riots (Klier and Lambroza 1992).

There has been an increasing shift of focus to nonconflicting crowds, a development promising to greatly expand the field of research. Dynastic governmental ceremonies, middle-class-inspired national festivals, traditional celebrations, religious feast days, and new-style leisure activities could easily assume the form of mass events, though in different guises and in radically transformed public functions: staged by the ruling class and social elites and hierarchically structured; composed of a mixture of social groups, neither exclusively nor primarily from the lower classes; symbolically strengthening the existing rule and its established social order, as opposed to rebellious behavior protesting destitution and political impotence; in their effects, socially integrative rather than polarizing. The lower classes, usually self-determining actors at the very nucleus of crowd activities, figured here primarily as public, as mere backdrop.

A study of Bristol 1790–1835 shows that next to peaceful 'election crowds,' 'royal and military crowds,' and 'recreational crowds,' unruly 'crowd riots' were among those events occurring with the highest frequency in the city. However, at the same time, 'crowd riots' formed only a minority (22 percent) of a sample that covered all types of crowd occurrences taking place within the city. The local elites of Bristol had ample skill and experience in successfully mobilizing large crowds onto the main streets and squares for purposes of their own (Harrison 1988).

The general European 'nationalization of the masses' in the nineteenth century included a wealth of identity-forming national celebrations and mass manifestations.

A comparative study of middle-class movements for national memorials in France and Germany has demonstrated the precarious and paradoxical strategy of including the lower classes in the mythical 'nation,' while simultaneously drawing sharp social distinctions over and against these very same lower classes. It was above all through the militarization of the bourgeois festival culture that the elites were able to overcome their deep-rooted fear of the masses (Tacke 1995).

Expanded political participation, as well as transformations within the public sphere in industrial and post-industrial societies of the twentieth century in no way contributed—as often asserted—to the disappearance of protesting crowds. Rather, one can speak of modifications, and of coexistence of direct action with newly learned and more organized and representative political forms. After 1850, aside from periodic recurrences of crowd activity in line with the older model (for example, food riots in France, and especially in Italy and Spain), one finds numerous

newly-invented forms of crowd action, above all strikes, boycotts, mass manifestations and the like. However, soon after 1900 the power potential of crowds began to appear less threatening to the established order (in the sense of general overthrow). Crowds were governable, crowds could be manipulated—this was the new slogan of ruling elites.

An empirical study of Berlin 1900–14 shows the high presence of crowd events in early twentieth-century streets. The events fell into three major categories—405 street disorders, 74 violent strikes, and 61 political mass manifestations. The ongoing struggle for the city streets is interpreted as conflict between a 'street politics from below' and a restrictive 'street politics from above' carried out by the state power and its police. It was on the political terrain of the street that exaggerated fear of subversion, strict anti-revolutionary measures, and socially discriminatory police practice clashed with collective demands for greater social justice and expanded political participation (freedom of assembly, the right to demonstrate, universal suffrage). In a much wider sense, these were demands for respect and self-determination fought out by the proletarian majority in the public forum of the street, the traditional political arena of the common people (Lindenberger 1995).

For the extremely crisis-ridden interwar period, one can even speak of a culmination of 'street politics,' increasingly transformed (militarized) through the advent of radical political–ideological movements from both the left and the right. In this civil war atmosphere of rabid party-political action, with the air filled with bellicose slogans such as 'self-help' and 'self-defense,' authentic collective crowd action threatened to be instrumentalized and finally suffocated in the battle of totalitarian movements bent on the conquest of state power (for a comparative study of Berlin and Paris 1919–33/39, see Wirsching 1999).

For some time now the broadening of perspectives has meant the inclusion of non-European regions of Asia, Africa, and Latin America, which have a history of crowds of their own. Important basic forms resemble the European riot patterns: in the center stood collective claims of peasant producers who rebelled against landlords or large land-holding farmers, urban money-lenders or state organs for reasons of land scarcity, imposed duties, debts, or tax pressure. In the East Asian context, in a series of studies, James C. Scott has shown how peasant behavior was culturally embedded in an all-encompassing subsistence ethic with precise norms regarding access to land, acreage, and a fair apportionment of their own produce. The threats posed to these cultural norms in the context of firmly established clientele relationships might have led to open collective resistance, even if—in consideration of the manifold covert and individual 'weapons of the weak'—this remained more the exception than the rule (Scott 1976).

In some regions (particularly of China since the eighteenth-century, and to some degree of Japan and of India in the nineteenth and early twentieth-centuries) the food riot tradition was of considerable historical importance, thus offering comparisons with early-modern Europe (Wong 1983). At the same time, subsistence conflicts of many Third World regions continue to this day, as confirmed by many contemporary eruptions since the 1970s in North and Central Africa (Egypt, Morocco, Sudan, and Tunisia) and Latin America (Dominican Republic, Venezuela, Brazil, and Argentina). In several cases it was food-riot crowds—whose agitation came in the wake of reduced state food subventions and resulting higher prices—that shook and even toppled ruling governments.

4. *Outlook*

The long and hard-fought interpretive battle concerning crowds has been marked by a double disappointment: neither have crowds been shown to be the great negative power, an absolute culture-destroying evil; nor have they been—teleologically speaking—a collective subject to fulfill any historical missions. Crowds have always shown themselves to be multi-layered, ambiguous, and variously interpretable phenomena. And naturally, a certain ambivalence remains. For although emancipatory aspects in the sense of subsistence gains, political participation, and self-determination may dominate the phenomenon of crowds, dangerous exclusionary efforts at expulsion or even destruction of religious, ethnic, and other minorities can also be part of the crowd repertoire. The demythification of the crowd, the gradual opening of political structures to greater popular participation, the modification of collective behavior through historical experience (learning processes that include a growing self-control and adoption of the principle of nonviolence), and other factors may have contributed to crowds having increasingly forfeited their attractiveness as a catch-all for fearful visions or as an instrument for political and ideological agendas imposed from without.

As historical actors in their own right, crowds lay claim to a considerable degree of legitimacy. At the center of this legitimacy stood the collective demands for 'subsistence' (later on: social sharing) and 'liberties' (communal rights, cultural autonomy and, later, political participation) in the context of given governmental and living conditions. Insofar, conflicting crowds were always—more unconsciously than intentionally—indirect collective reformers. Obviously, this seemingly paradoxical role is part of the peculiarities of rebellion and revolution. The political culture of the West—with its human rights norms embodied in the welfare state and democratic institutions since the end of the nineteenth century—would have never attained to the stature it occupies

today had it not been for persistent collective intervention from below. Crowds have helped to draw up the agenda for social and political human rights, and they continue to do so.

Where comprehensive systems of social sharing and political participation have been historically established and long in place, crowds tend toward a shift in their forms and motives. From a purely quantitative standpoint, protesting crowds in recent times have been anything but on the decline. To the contrary, they have been on the upswing, becoming a conventional part of political culture. Although violence has receded, it has not fully disappeared. Rating high are questions other than elementary ones of subsistence and freedom. Demands for social and political egalitarianism have for some time taken a backseat to new values and fresh demands based on 'new social movements.' Those groups affiliated with crowd events are no longer predominantly or exclusively from the lower classes, but are a great mix of people with considerable representation from the public sector, education, science, and culture. Fitting in well with these trends is also the upsurge of class non-specific leisure crowds in the context of new event cultures.

As an elementary form of the public sphere (*Versammlungsöffentlichkeit*), crowds will endure. Certainly modern mass media have continued to occupy strongly the public arena, becoming themselves more and more an embodiment of the entire public. But talkshows have hardly rendered crowds superfluous. Albeit today's media presence largely determines the public perception of crowd events, thereby heavily influencing their effectiveness. In their endless variety, crowds represent the unmediated, face-to-face public. Such gatherings of people for purposes of social and communicative self-awareness is apparently a deep-seated human need that—despite the relative dwindling of traditional conflicting crowds—will not soon disappear.

See also: Collective Behavior, Sociology of; Groups, Sociology of; Labor Movements, History of; Public Sphere: Nineteenth- and Twentieth-century History; Revolution; Revolutions, History of; Revolutions, Sociology of; Social Movements, Geography of; Social Movements, History of: General; Social Protests, History of

Bibliography

Barrows S 1981 *Distorting Mirrors: Visions of the Crowd in Late Nineteenth-Century France*. Yale University Press, New Haven, CT and London

Bohstedt J 1983 *Riots and Community Politics in England and Wales 1790–1810*. Harvard University Press, Cambridge MA

Bouton C A 1993 *The Flour War: Gender, Class, and Community in Late Ancien Regime French Society*. Pennsylvania State University Press, PA

Gailus M 1990 *Strasse und Brot: Sozialer Protest in den deutschen Staaten unter besonderer Berücksichtigung Preussens 1847–1849*. [English translation: *Street and Bread: Social Protest in the German States with Special Reference to Prussia 1847–1849*]. Vandenhoeck & Ruprecht, Göttingen, Germany

Harrison M 1988 *Crowds and History: Mass Phenomena in English Towns 1790–1835*. Cambridge University Press, Cambridge, UK

Hobsbawm E J 1952 The machine breakers. *Past and Present* **1**: 57–70

Holton R J 1978 The crowd in history: Some problems of theory and method. *Social History* **3**: 219–33

Klier J D, Lambroza S (eds.) 1992 *Pogroms: Anti-Jewish Violence in Modern Russian History*. Cambridge University Press, Cambridge, UK

Lefebvre G 1934 Foules Révolutionnaires [English translation: Revolutionary Crowds]. *Annales historiques de la Révolution française*. **11**: 1–26

Lindenberger T 1995 *Strassenpolitik: Zur Sozialgeschichte der öffentlichen Ordnung in Berlin 1900 bis 1914*. [English translation: *Street Politics: A Social History of Public Order in Berlin 1900 to 1914*]. Dietz Nachfolger, Bonn, Germany

Randall A, Charlesworth A (eds.) 1996 *Markets, Market Culture and Popular Protest in Eighteenth-Century Britain and Ireland*. Liverpool University Press, Liverpool, UK

Rohrbacher S 1993 *Gewalt im Biedermeier: Antijüdische Ausschreitungen in Vormärz und Revolution 1815–1848/49*. [English translation: *Violence in the Biedermeier Era: Anti-Jewish Riots during Vormrz and Revolution 1815–1848/49*] Campus, Frankfurt, Germany and New York

Rudé G 1964 *The Crowd in History: A Study of Popular Disturbances in France and England 1730–1848*. Wiley, New York & London & Sydney, Australia

Scott J C 1976 *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia*. Yale University Press, New Haven, CT and London

Tacke C 1995 *Denkmal im sozialen Raum: Nationale Symbole in Deutschland und Frankreich im 19 Jahrhundert*. [English translation: *Monument in the Social Space: National Symbols in Germany and France in the 19th Century*] Vandenhoeck & Ruprecht, Göttingen, Germany

Thompson E P 1971 The moral economy of the English crowd in the eighteenth century. *Past and Present* **50**: 76–136

Thompson E P 1991 The patricians and the plebs. In: Thompson E P (ed.) *Customs in Common*. Merlin Press, London, pp. 16–96

Tilly C, Tilly L, Tilly R 1975 *The Rebellious Century 1830–1930*. Harvard University Press, Cambridge, MA

Wirsching A 1999 *Vom Weltkrieg zum Brgerkrieg? Politischer Extremismus in Deutschland und Frankreich 1918–1933/39: Berlin und Paris im Vergleich*. [English translation: *From World War to Civil War? Political Extremism in Germany and France 1918–1933/39: Berlin and Paris in Comparative Perspective*]. Oldenbourg Verlag, Munich, Germany

Wong R B 1983 Les émeutes de subsistances en Chine et en Europe occidentale. [English translation: Food Riots in China and Western Europe]. *Ann Econ Soc Civil* **38**: 234–58

M. Gailus

Cuban Revolution, The

The Cuban Revolution, as Fidel Castro proclaimed in January 1959, was a true revolution. A new society was forged with the triumph of Castro's organization,

the Twenty-sixth of July Movement, over other competing political organizations, and with the institutionalization of a Marxist-Leninist government. Cuba was involved in some of the most serious international confrontations of the last half of the twentieth century. The massive exodus that started in 1960 created a diaspora of more than two million Cubans. This article explores the historical background of the revolution and the policies that made it both a success and a failure.

1. Key Events

1.1 From Independence to Revolution

From 1895 to 1897, the War of Independence between Cuba and Spain was a protracted conflict. The war ended when the USA intervened and occupied the island in 1898. This intervention occurred as the war was starting to turn in favor of the Cubans. Although the Cubans collaborated with the US troops and constituted the fundamental land force that permitted the swift US victory, they were ignored during the peace negotiations at the Treaty of Paris of 1898. The USA aided in the reconstruction of the war-torn country, improving health conditions and education. US investments and the return of Spanish capital contributed to economic recovery. When US occupation (1898–1902) ended, there was only one objectionable point: the Platt Amendment. Imposed by the USA as a requisite for turning over the country to the Cubans, the delegates to the 1901 Constitutional Assembly attached it to the Cuban Constitution. Annexationist interest and the notion held by US officials that Cubans were incapable of self-rule, led to the imposition of the amendment. The amendment, drafted by US Senator Orville Platt, gave the USA the power to intervene militarily in the island and restricted the Cuban government in its international relations.

Traditional and pseudo-Marxist historiography reduces US–Cuban relations to the events that occurred in and after 1898. In this view, 1898 marks the date in which Cuba stopped being a Spanish colony to become a US one. The truth is far more complex. Cuba's economic dependence on US trade was a fact by the mid-nineteenth century. Cuba was the world's foremost producer of sugar and the USA imported 90 percent of Cuba's production. The USA had also been a refuge for Cubans who aspired to independence. The first Cuban political party was created by exiles in the USA under the leadership of José Martí. Martí, although not anti-USA, feared and staunchly opposed North American intervention. At the end of the war, the economic ties between the countries continued to play a direct role in US–Cuban relations.

Liberalism, political participation, the absence of strong partisanship, electoral fraud, administrative graft, and civil unrest characterized the early republic. The postindependence concentration of US and Spanish capital in the sugar industry, together with the Spaniards' domination of banking and commerce, rendered participation in the government the main vehicle for economic success for many Cubans. The creole elite had been nearly destroyed. Spain did not compensate Cubans for the loss of property during the war. Armed struggle was often a means to ensure circulation of elites in public office. The USA intervened militarily to establish order on two occasions and occupied the island for a second time from 1906 to 1909.

Cuba's economy continued to depend on sugar exports. By 1902, the almost miraculous recovery of the sugar industry had turned Cuba into the world's second largest producer. The preferential tariff treatment accorded by the USA in the Reciprocity Treaty of 1903 reinforced the preeminence of sugar, hindered economic diversification, and heightened Cuba's dependence on the USA.

Sugar production and prices rose during and after World War I. The prosperity created a Cuban-born entrepreneurial class with a nationalist agenda. In the 1920s, the collapse of sugar prices and the Cuban banking crash plunged the nation into economic chaos. By 1923, students and workers had joined industrialists and intellectuals demanding abrogation of the Platt Amendment, curtailment of government concessions to foreigners, and the end to the country's dependence on sugar.

General Gerardo Machado's dictatorship (1925–33), coupled with the US government's interference and plummeting sugar exports were the catalysts for the revolution of 1933. A provisional revolutionary government was installed and president, Dr. Ramón Grau abrogated the Platt Amendment unilaterally. His administration, which enacted social, political, and economic reforms, encountered the unswerving opposition of the USA. Benjamin Sumner Welles, the US ambassador, and other opponents denounced the measures as too radical. Although Welles requested US military intervention, Washington opted to withhold diplomatic recognition of Grau's administration. It was axiomatic that no Cuban government could survive without US support. Strikes continued, and in the eastern provinces workers seized some sugar mills and established soviets. On January 1934, Fulgencio Batista, promoted to army chief of staff, deposed Dr. Grau. Although short-lived, the 1933 revolution challenged US hegemony, introduced changes in Cuba's socioeconomic and political structure, and rekindled nationalism. In 1934, the USA abrogated the Platt Amendment, signaling a curtailment of Washington's interference in Cuban affairs.

Batista ruled by proxy until 1940, when he was elected president. The Partido Revolucionario Cubano (Auténtico) came to power with the presidential election of Dr. Ramón Grau (1944–48). Cubans

expected a continuation of the reformist policies enacted during Grau's short administration in the 1930s. Those reforms, incorporated into the 1940 constitution, reflected the revolutionary spirit of the 1930s. Unfortunately, many of the significant political and socioeconomic advances of the Auténticos were combined with governmental graft, corruption, and political gangsterism, which was also the case with the auténtics administration of Carlos Prío Socarrás (1948–52).

In 1947, Eduardo Chibás, a disaffected Auténtico, founded the Partido del Pueblo Cubano (Ortodoxo). The Ortodoxos represented the hopes of a generation for whom the promises of the 1933 revolution remained unfulfilled. Political legitimacy and honesty in public office were at the core of his movement. Chibás's suicide in 1951 created a leadership vacuum and renewed the frustration that plagued Cuban society. A coup, on March 10, 1952, led by General Fulgencio Batista, ended the constitutional experience two months before the 1952 elections. On July 26, 1953, Fidel Castro, a young Ortodoxo, who had run for office in the 1952 elections pre-empted by General Batista's coup, organized the unsuccessful attack on the Moncada barracks in Santiago de Cuba. The daring act provided publicity for Castro and for his nascent anti-Batista organization, the Twenty-sixth of July Movement. When Batista refused the demands for free elections made by a coalition seeking a nonviolent solution to the crisis, the Federación Estudiantil Universitaria (FEU) called for a student strike, and violence escalated. In 1956, Castro and José A. Echeverría, FEU president and head of the Directorio Revolucionario Estudiantil (DRE), joined forces. By December, Castro and his followers had established guerrilla operations in the mountains of the Sierra Maestra, in Cuba's westernmost province of Oriente. On March of 1957, the Auténticos and the DRE stormed the presidential palace and Echeverría was killed in a separate incident during the failed assault. A growing number of the professionals and entrepreneurs joined the fight as the government's repression mounted. Other guerrilla fronts were then established and the underground initiated a terrorist campaign. Several conspiracies in the armed forces were discovered between 1956 and 1957. In 1958, the USA imposed an arms embargo on the Cuban government; a military offensive against the rebels failed and they in turn mounted a counteroffensive. Batista was deposed by a coup in December. On January 1, 1959, the army surrendered to the Rebel Army and the unanticipated communist revolution began its course. The Cuban communist party, Partido Socialista Popular (PSP), had been critical of Castro's Moncada attack and barely participated in the struggle against Batista until 1958. Raul Castro, Fidel's brother, however, was a communist and may have established contacts with the international communists at the Sierra Maestra.

On the eve of the revolution, Cuba was a developing nation ranking among the top five countries in Latin America. The movement to depose President Batista was not anti-Yankee, nor anticapitalist, antibourgeois, or nationalist. By the 1950s, radical nationalism was mostly confined to the intellectual elite. Although the anti-Batista movement had ultimately a heterogeneous class base, members of the middle classes seemed to have comprised the core of the active participants, particularly in the urban underground. Among the insurrectionists, there was also an obvious presence of individuals of middle and upper class backgrounds, such as the Castro brothers and some of their closest collaborators. The Twenty-sixth of July Movement and the other anti-Batista groups focused on political freedom, constitutionality, and fiscal integrity, which were the issues concerning the Cuban middle classes. Castro's social and economic planks were based on the implementation of the provisions of the constitution of 1940, which included agrarian reform.

1.2 Revolution (1959–1970)

Between 1959 and 1961 Cuban society was transformed. Foreign and domestic properties were gradually expropriated. Political parties, elections, freedom of the press, and labor's right to strike were eliminated. Universities, labor unions, and professional associations were purged. The 1940 constitution was proclaimed to be outdated, and the USA was identified as Cuba's enemy. A new currency destroyed the power of the upper and middle classes. All criticism was declared as counterrevolutionary, and political opponents were imprisoned or executed. The first massive flight into exile of Cuba's entrepreneurial and professional sectors was initiated.

Cuba's foreign relations changed just as rapidly. Soviet military advisers and weapons began arriving by September of 1960. In October, the USA imposed a trade embargo and diplomatic relations between the two countries were severed in January 1961. That same year, an invasion was launched by Cuban exiles that had received covert military training and support from the US government. The failure of the Bay of Pigs invasion of April of 1961 consolidated the revolution. The underground was crushed, as an estimated 100,000 persons were detained. On the eve of the invasion, Castro proclaimed the socialist nature of the revolution. One month later, diplomatic relations were reestablished with the USSR, and on December 2, Castro declared himself, for the first time, a Marxist-Leninist.

In 1962, after a US Air Force reconnaissance aircraft discovered the construction of missile-launching sites and offensive Soviet intermediate ballistic missiles in Cuba, an international crisis between the two superpowers ensued. The outcome of the missile

crisis had a dual impact on Cuba's foreign policy. When the USA promised the USSR not to invade Cuba in exchange for the Soviet removal of missiles, any viable external threat was eliminated; however, the Cubans felt betrayed, as the compromise was reached without consulting them. Cuba then sought new allies by promoting and supporting revolutionary movements throughout Latin America. In 1967, these policies and the arrests of several members of the Cuban communist party that followed the Soviet line strained Cuban–Soviet relations. The USSR retaliated by decreasing the amount of petroleum deliveries to the island.

Cuba attempted to chart an independent course during the 1960s. It was, however, facing trading problems similar to those experienced during colonial times. It had changed a trading partner 140 km from its shores for one more than 4,500 km away. By 1963, an ill-conceived industrialization program and the neglect of Cuba's main cash crop caused the progressive indebtedness of the country. The leadership turned again to sugar and a ten million-ton sugar harvest was programmed for 1970. The sugar industry, however, was in shambles by 1964. Professional administrators had gone into exile or were substituted with politically trustworthy personnel, while expert cane-cutters had been transferred to other activities or incorporated into the militia. Already facing the failure of its international policies and domestic economic strategies, coupled with the US embargo, the impact of the Soviet sanctions on the regime was catastrophic. In 1968, Castro supported the USSR invasion of Czechoslovakia, signaling Cuba's capitulation to Soviet hegemony and Castro's determination to remain in power at any cost.

By the mid-1960s the revolution was giving unequivocal signs of fatigue. Juvenile delinquency and crime were on the rise. Disaffected Cubans, together with priests, seminarians and homosexuals, drafted into Military Units to Aid Production (UMAP), were interned in concentration camps from 1965 to 1967. The 1968 'Revolutionary Offensive' nationalized the remaining small businesses. Low worker morale and absenteeism led to low productivity and poor product quality. The strategy of moral incentives failed to mobilize the majority of workers, who were no longer imbued by the revolutionary zeal. The revolutionary process had spent itself by 1970.

1.3 Institutionalization and Foreign Military Adventurism

After 1970 Soviet political and economic models were adopted. Some material incentives were implemented, and peasant markets opened. Institutionalization, however, continued under the charismatic guidance and legitimization of Castro's leadership. In 1974, Leonid Brezhnev's visit consolidated Cuban–Soviet relations. A year later Cuba's African campaigns began with Soviet support, and in 1979 Castro was elected president of the Nonaligned Movement. Cuba's support of the Soviet invasion of Afghanistan, however, diminished its prestige and its claims to nonaligned status.

Although increases in Soviet aid led to improvements in living standards, over 120,000 Cubans abandoned the island during the 1980 Mariel Boatlift. By 1986, the failure of the Soviet models was indisputable. The crisis was compounded by the US embargo being turned again into a powerful economic weapon by the collapse of the Soviet bloc and the breakup of the USSR. By the 1980s Soviet subsidies had reached $3 billion a year. The Soviet Union supplied: 90 percent of the island's oil needs and an inordinate proportion of basic foodstuffs. To confront the crisis, Castro launched the 'Rectification of Errors Campaign' at the Third Party Congress. Material incentives and market reforms were eliminated, and mass mobilization and volunteer work were reintroduced.

In 1989, the arrest, trial, and execution of Division General Arnaldo Ochoa and of the interior ministry's Colonel Antonio de la Guardia signaled an institutional crisis. They were accused of drug dealing and breaching national security. A massive purge followed and several government agencies were dismantled. Many Cubans suspected that General Ochoa had supported political liberalization and economic reforms.

1.4 Back to the Future: Authoritarianism and Capitalism

The 'rectification' measures proved insufficient to preclude an economic collapse. In 1990 a 'special period in peacetime' was declared and policies to develop an autarkical state were adopted. To alleviate food shortages and decrease energy consumption, thousands of urbanites were forced to move to the countryside and to work in food production. Austerity and mass mobilization were expected to extricate the country from the crisis, until foreign investments, tourism, and an incipient biotechnological industry could generate the necessary hard currency.

At the 1991 Party Congress, Castro revealed that he was a 'slave' of power, entrusted with preserving socialism. There would be no political or economic reforms. By 1993, economic reforms were inevitable. Although private initiatives continued to be restricted, free markets for agricultural and consumer products, duty-free zones and 'dollarization' were introduced. Access to dollars, mainly through remittances from exile relatives, tourism, and illicit activities such as black-market participation and prostitution, provided an economic respite to a growing segment of the population. As Castro had suspected, economic liberalization started eroding the state's coercive power. Religious participation increased, critics became more vocal, human rights groups proliferated,

and short-lived rebellions erupted. In 1994, another mass exodus occurred when more than 30,000 people abandoned the island in unique makeshift innertube rafts.

In the 1990s the regime encountered growing criticism from international leaders and organizations. The International Labor Organization, Amnesty International, and the United Nations repeatedly condemned the Cuban Government for its human rights violations. Pope John Paul II urged Cuba to open to the world and condemned the US embargo. In 1996, the embargo had been reinforced by the passage of the Helms-Burton bill, introduced after Cuban Migs had shot down two US light planes in international waters, killing three US citizens. The planes belonged to 'Brothers to the Rescue,' an exile organization that flies over the Strait of Florida in search of rafts carrying Cubans fleeing the island.

2. *Rebellions and Revolutions*

'Rebellions' and 'revolutions' are terms often used to identify a series of political struggles in Cuba's history. As with China, Cuba's official, Marxist historiography proclaims one single revolutionary movement starting with the Ten Years' War (1868–98). Thus, that first war, which was restricted to the western provinces and was a classical war for independence, organized by members of the white middle and upper middle classes, primarily made up of professionals and landowners, is interpreted as a revolutionary crusade. The War of Independence (1895–8) is depicted as a continuation of that struggle, despite its broad popular base comprising workers, the middle class, peasants, and the black and mulatto populations. This second anticolonial, nationalist war, which sought to establish an independent, democratic republic, is then portrayed as a socialist revolution, interrupted by colonial, imperialist, and bourgeois interests until Castro's victory in 1959.

Another taxonomy could divide Cuba's revolutionary history, beginning in 1868, into two anticolonial, nationalist, strictly political revolutions (Ten Years' War and War of Independence); a protorevolution (1933 revolution), which was revolutionary in intent but did not achieve radical social, economic,or political changes; and a true revolution after Castro's triumph in 1959.

3. *Origins of the Cuban Revolution*

The complexity of Cuban society and of revolutions in general becomes evident as one surveys the literature on the causes of the Cuban revolution. Leo Huberman and Paul Sweezy (1961) see it as the outcome of peasant mobilization. Theodore Draper (1962) claims that it was a middle-class phenomenon. Andrés Suárez (1967), among others, argues that it was the product of Castro's charismatic leadership and political maneuvering. To Luis Aguilar (1972), it was the frustrated 1933 revolution and the absence of reforms. Jorge I. Domínguez (1978) emphasizes that it was the fragmentation and individual strength of the labor unions, professional organizations, and other interest groups in prerevolutionary Cuban politics that allowed the revolutionary process to crystallize after 1959. Marifeli Pérez-Stable (1999) sides with contemporary Cuban historiography, and claims the existence of a linear relationship between the republican failure to achieve the aspirations of the wars of independence and the revolution. In essence, there is no agreement among scholars regarding the causes of the Cuban Revolution.

4. *Revolutionary Dynamics*

The Cuban revolutionary leaders implemented strategies commonly used by totalitarian regimes. Castro's policies are reminiscent of Nazi, fascist, and Falangist tactics. The revolution marked a new beginning and required a radical transformation in the collective consciousness. Ernesto Ché Guevara (1971) expounded on the need to create both a new economic base and a 'new man' to build communism. Destruction of the historical memory was fundamental to the creation of a new society. The process of deculturation strove to uproot prerevolutionary organizational and interpersonal social patterns. The institutions of work and education, radically transformed, became the focus of Cuban life. Institutionalized religion was crippled, while attempts were made to destroy the traditional structure and values of the Cuban family. Places and events embody symbolic meanings that provide continuity and solidarity. The names of streets, sugar mills, and other landmarks were changed. The government monopolized all means of communication. Slogans, in which the names of key figures were always present, were incessantly repeated through printed media, billboards, radio, and television. Patriotic and religious holidays were replaced by revolutionary events. In 1968, the October 10, which marks the beginning of the Ten Years' War, was reinstated and the leaders appropriated this patriotic date to proclaim themselves the true liberators of Cuba 'after 100 years of fighting.' Through symbolic manipulations Cuba became synonymous with revolution and the revolution synonymous with Fidel Castro. Thus, through this syllogism, Cuba, Castro and revolution, (Fidel–patria–revolution according to Pérez-Stable 1999) became one in the public imagery.

Mass organizations, which incorporated, voluntarily or coercively, all members of society into the process, were responsible for transmitting the new ideology, introducing new patterns of behavior, conducting constant vigilance, and mobilizing the population. The euphoria that the revolutionary triumph

inspired promoted an environment of fanatical solidarity in which there was no room for dissension. The 'terror' stage, common to revolutionary upheavals, was unleashed by the trials, long jail sentences, executions, and the feeling of being under constant surveillance. A process of self-repression was set in motion among the dissenting population. After the Bay of Pigs fiasco, the exodus intensified, with the massive flight of the urban middle class, including those who had fought against previous oppressive regimes. Their density of networks and their experience in underground activities made them a target of the regime. The government, reluctantly, allowed them to leave, although their flight represented a devastating loss of human capital. The impact of these policies, although not as desired by the government, has been, nonetheless, sufficiently powerful to preclude organized domestic challenges for several decades.

5. *Conclusion*

In the 1990s a Cuban joke that defined socialism as 'the longest road between capitalism and capitalism' summarized the tragedy of a revolution that had gone full circle. Cuba's economy is returning to capitalism and continues to depend on sugar and tourism. Cuban women and men, now from all walks of life, often prostitute themselves to make a living, or in the hope that contacts with foreigners may provide them with the means to leave the country. The revolution's greatest failure has been its historical incapacity to retain the nation's greatest asset: its human capital. Making political survival its main priority, however, has won the leadership its greatest success; it has stayed in power for four decades.

See also: Cold War, The; Communism, History of; Dictatorship in History; Nation-states, Nationalism, and Gender; Revolution; Revolutions, History of

Bibliography

Aguilar L E 1972 *Cuba 1933. Prologue to Revolution*. Cornell University Press, Ithaca, NY

Bonachea R L, San Martín M 1974 *The Cuban Insurrection, 1952–1959*. Transaction Books, New Brunswick, NJ

Bunck J M 1994 *Fidel Castro and the Quest for a Revolutionary Culture in Cuba*. Pennsylvania State University Press, University Park, PA

Domínguez J I 1978 *Cuba: Order and Revolution*. Harvard University Press, Cambridge, MA

Draper T 1962 *Castro's Revolution: Myths and Realities*. Praeger, New York

Fagen R 1969 *The Transformation of Political Culture in Cuba*. Stanford University Press, Stanford, CA

Guevara E 1971 Man and socialism in Cuba. In: Silverman B (ed.) *Man and Socialism in Cuba: The Great Debate*. Atheneum, New York

Huberman L, Sweezy P M 1961 *Cuba: Anatomy of a Revolution*. Monthly Review Press, New York

Medin T 1990 *Cuba: The Shaping of Revolutionary Consciousness*. Lynne Rienner, Boulder, CO

Oppenheimer A 1992 *Castro's Final Hour. The Secret Story Behind the Coming Downfall of Communist Cuba*. Simon and Schuster, New York

Pérez L A 1988 *Cuba: Between Reform and Revolution*. Oxford University Press, New York

Pérez-Stable M 1999 *The Cuban Revolution. Origins, Course and Legacy*. Oxford University Press, Oxford, UK

Pino Santos O 1973 *El asalto a Cuba por la oligarquía financiera yanqui*. Casa de las Américas, Havana, Cuba

Suárez A 1967 *Cuba: Castroism and Communism, 1959–1966*. MIT Press, Cambridge, MA

Thomas H 1971 *Cuba: The Pursuit of Freedom*. Harper and Row, New York

M. R. Moreno Fraginals and
T. P. Moreno Fraginals

Cultural Assimilation

The term is best understood in contrast with close sociological concepts, some more general (cultural diffusion) and others more specific (acculturation). Cultural diffusion denotes the passage of ideas, norms and practices from one group to another in general. Acculturation is the processes whereby one cultural system conquers the minds of an individual or group. 'Cultural assimilation' is used generally to denote the situation in which a group gradually acquires some traits of a larger society's repertoire of concepts and norms. Two main features are specific to assimilation situations. First, there is a clear asymmetry, often between a majority and a minority but, more importantly, between what are felt to be center and periphery, a dominant culture and a less central one. This is often visible in the way the central group's identity remains unnamed, or else carries the name of the overall society. In this way the leading group in the Austro-Hungarian Empire had no explicit ethnic label, unlike the Czechs, Jews, or Hungarians they ruled.

Similarly, although white Americans of European descent are clearly the majority background against which minorities are defined in the USA, there is no commonly-used ethnic category for them ('Caucasian' is not really used by anyone outside the census bureaucracy). Second, assimilation is the project whereby a group tries to gain a seat at the common table, as it were, but not to dissolve itself. But there may be considerable anxiety that achieving this political and social goal will inevitably dissolve the cultural traits that underpin a sense of common identity and destiny.

The most acute of these debates were conducted among the Jewish minorities of central Europe at the turn of the twentieth century. From this situation in which a general, often violent oppression of Jews as a group combined (in some European polities) with

unprecedented opportunities for Jews as individuals, emerged two different views of assimilation. One recommended a more intense participation in the social and political processes of the wider polities, trusting that common legal arrangements would shield the minority from the worst effects of personal prejudice. The end-goal was to make Jewishness, if not invisible, at least irrelevant to most contexts of social interaction, and therefore accepted in practice. The other, largely Zionist, view assumed that such a project would either fail altogether (inasmuch as the ever-more visible presence of Jews in the larger society would fan the fire of resentment and prejudice) or, on the contrary, succeed all too well, at the price of dissolving Jewish identity and annihilating Jewish culture.

This tension between participation in the wider social forum, often seen as a condition for survival, and the preservation of identity, construed as equally urgent prescriptions, informs American debates about the depth, meaning and consequences of the cultural assimilation of minorities. The phrase 'melting pot' was coined in the 1910s. It was thought to describe the essentially American and altogether laudable practice of combining norms and practices from different groups in order to create a distinct American political entity as well as a distinct culture. It was assumed that this cultural alchemy would preserve the best of each culture and create an even better one, in the same way as a free market weeds out inefficiency and fosters good practices.

The 'melting pot' ideology emerged more or less at the time when it ceased to be an American reality. Whereas the unification of American groups of different European origins (Scandinavian, English, Irish, etc.) in the course of the nineteenth century more or less corresponded to what the 'melting pot' was supposed to describe, the destinies of groups such as Asian-Americans, Native Americans, Hispanics and African-Americans in the twentieth century were to follow a distinctly separate trajectory. For these groups there was no melting pot, as many liberal critics of American politics observed. These critics often saw the sorry plight of minorities as a straightforward consequence of this failure to mix or assimilate.

However, many members of these minority groups had a diametrically opposed view of the assimilation process. For them, acquiring political leverage and legal protection requires, not their dissolution into larger social membership, but exactly the opposite; an attempt to make their membership of a particular group visible and audible, so that minorities could be counted and allocated political power commensurable to their demographic composition. This change in assimilation patterns (whereby modern assimilation no longer entails acculturation) does not reduce to the general contrast between old groups and newly arrived ones (called the 'old-new' model in sociology). Hence a distinctively American nonmelting pot, where *de facto* cultural similarities are ever-increasing, but social-ethnic groups also maintain strict boundaries and a heightened sense of distinctive identity, and seek an corresponding political influence.

This modern American situation is equidistant from two European political projects that informed the connections between nation and culture. One such project was romantic nationalism. Born in European nations in the eighteenth and nineteenth centuries, this ideology celebrated the ideal of a community united by land, language, and cultural traditions. It suggested that the most 'natural' or most viable political units would correspond to such communities. There was no need to provide specific political recipes for the management of such communities since each cultural tradition already incorporated its own political wisdom, time-tested and therefore appropriate to the nation concerned.

The opposite project was that of the Enlightenment and French Revolution. Here the nation was a *de facto* political unit, with a predefined political ideal (equality and democracy) but no connection to any particular cultural tradition or language. The nation was therefore construed as emerging from the participation of different individuals *qua* individuals. This is the sense of the famous statement of the French Constituent Assembly, requiring that the emancipation of the Jews, while granting them citizenship and making religiously-based oppression illegal, would 'give them everything as citizens, but nothing as a group.' In such a political project there is space for various associations of citizens based on common interests or ideology, but no place for ethnically-defined groups as political participants.

Both European projects seem to have reached the limits of their political relevance by the end of the twentieth century. Romantic nationalism, the foundation stone of many central European nations, flounders in situations characterized by ethnic patchwork, as in the Balkans, and leads to deluded ideologies of domination and the annihilation of the other. Enlightened, supposedly antiethnic nations are reinvested with ethnic and cultural qualities by many participants. This is why many citizens of the Federal Republic in Germany or the French Republic find certain of the cultural practices of Turks or Arabs repulsive or culturally threatening. In both countries political debate now centres on the way distinct ethnic groups could be granted some measure of influence or cultural autonomy without threatening the liberating possibility of participation as a free citizen, with no specific group affiliation. These debates are made more pressing by the expansion of the European Union, with the attendant decrease in the powers of the states and the increase of political influence from regions.

In Europe and other parts of the world such debates are also prompted by concerns over 'globalization.' However vaguely defined, this phenomenon of world-

wide cultural integration and political interdependence is seen as desirable (or inevitable) by many ruling elites and every bit as detrimental (or resistible) by others. From both positions, it seems that even established nations, where cultural and political identities coincide, find themselves confronted with an assimilation process. That is, some of their characteristic political or cultural arrangements are seen as an obstacle to entry into a global economic and political world. So most places become a periphery relative to a center defined as everywhere else.

Historians, sociologists, and political scientists have focused mainly on the social groups concerned, on the interaction between center and gradually assimilated entities. As a consequence, assimilation is often construed, albeit implicitly, as a one-dimensional process, whereby a given group gets closer to the cultural mainstream or maintains its distance from it (or is kept at the periphery). Such metaphors of proximity and distance may not do justice to the multidimensional nature of participation in a large polity. There may be many different dimensions along which a given individual places him- or herself at a particular point relative to a social mainstream. Many African-Americans feel more culturally assimilated to mainstream America than Asian-Americans, yet they are accepted much less in terms of political participation.

To make matters even more complex, it may be insufficient to describe cultural assimilation in terms of explicit concepts and norms. So far, social scientists have focused principally on those notions or preferences that individuals are aware of, and consider to be their distinctive character: their cultural traditions, their language, their common political past. But a cultural unity, a sense of being the same kind of person as others, may also be based on concepts and norms that are not consciously accessible; on ways of acting and even reasoning that are transmitted from one generation to the next without much conscious realization. Experimental social psychologists have documented such subtle cultural differences.

For instance, (white middle-class) American subjects do not react to simple computer animations in quite the same way as Chinese or Japanese subjects. When such animations show a school of fish tightly packed together and another fish at a distance from the group, Americans see the latter as a laudably independent, self-motivated agent. Many Asians, however, tend to see it as pitifully isolated and probably lost. Such results may seem largely explained by explicit ideologies of individualism and self-reliance (in the West) contrasted with group cohesion and solidarity (in the East). Other differences are less easy to interpret, such as a tendency in Asians to look at the overall pattern of a computer display, when Americans are more likely to focus on details and ignore its overall coherence.

Such experimentally-revealed differences do not necessarily mean that socialization in different groups result in fundamentally different cognitive processes. But they do show that assimilating a culture, either that of one's own group or that of a larger polity, cannot be reduced to adherence to explicit norms and cultural concepts. This connects with a concept of culture that is familiar to anthropologists but less so to other social scientists as encompassing, not just explicit norms and habits, but also what goes without saying, the tacit norms and unconscious cognitive habits that structure one's everyday world.

There is still little systematic study of the impact of migration and political assimilation on such elusive cognitive differences. Yet this is a crucial domain of investigation. Social and political interaction between central and noncentral groups may depend on agreement or discrepancies between these tacit norms every bit as much as they depend on explicit differences. Most social scientists are eager to show that assimilation processes require agents, that is, they do not happen outside the social and political projects of various people. It is not, for instance, just American culture, as an abstract entity, that somehow permeates cultural boundaries but actual Americans and members of other groups that join (or do not join) common social projects. Without neglecting this sense of active decision-making on the actors' part, it may be necessary to take into account aspects of cultural inertia that are not affected by people's decisions, because they are not really visible to the actors themselves, yet influence their integration into a wider cultural environment.

See also: Assimilation of Immigrants; Ethnic and Racial Social Movements; Ethnic Conflicts; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Immigrants: Economic Performance; Immigration; Immigration and Migration: Cultural Concerns; Immigration: Consequences for Fiscal Developments in the Receiving Population; Immigration: Public Policy; International Migration by Ethnic Chinese; International Migration by Ethnic Germans; International Migration by Ethnic Russians; International Migration by Ethnic South and Southeast Asians

Bibliography

Choi I, Nisbett R E, Smith E E 1997 Culture, category salience, and inductive reasoning. *Cognition* **65**(1): 15–32

Gellner E 1983 *Nations and Nationalism*. Blackwell, Oxford, UK

Glazer N, Moynihan D P 1963 *Beyond the Melting Pot*. Basic Books, New York

Lieberson S 1963 The old-new distinction and immigrants in Australia. *American Sociological Review* **28**(4): 550–65

Morris M, Peng K 1994 Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology* **67**: 949–71

P. Boyer

Cultural Contact: Archaeological Approaches

The focus of culture contact, as currently defined, is the study of culture change within the context of contact situations involving different cultures. Much of the recent effort in this field has been on early encounters of indigenous peoples with Europeans in colonial contexts. The current archaeological approach integrates lines of evidence from archaeology, ethnography, ethnohistory, linguistics, biological anthropology, and native oral traditions. Beginning with 'acculturation,' the article alludes to subsequent theoretical and methodological trends. While no single theoretical position captures the field, certain ubiquitous issues include: (a) world economy and the intensification of regional exchange and innovations in material culture; (b) demographic collapse, lethal epidemics, and changing natural landscapes; and (c) implications of demographic collapse for social disruption and reformulation including native responses of resistance, domination, and cooperation. A discussion of how issues relating to culture contact articulate with the greater theoretical agendas that dominate anthropology today leads to the recognition of methodological issues or problems as well as future directions involving historical anthropology and an increased interest in the development and negotiation of status hierarchies, ethnic identities, and gender relations in pluralistic colonial communities.

1. Introduction

The focus of culture contact, as currently defined, is the study of culture change or persistence within the context of contact situations involving different cultures and diverse types of interaction. This research has been and continues to be motivated by the overarching and longstanding desire within the field of anthropology to understand how and why culture changes. Much of the recent effort in this field has been on early encounters of indigenous peoples in colonial contexts. Numerous factors were involved in the early exploration and eventual settlement of landscapes all over the world including the discovery and exploitation of resources, the religious or spiritual conversion of Indians, trading, and gaining access to sea routes or ports. In some cases military objectives often necessitated the allegiance of the native population or the use of native peoples to buffer between European military territories. In other cases it was more expedient to use missionaries to control native groups than to establish forts and use military might (Milanich 1995, p. 165). Native encounters with European, African, and Asian peoples in early contact settings can be used to address issues regarding the founding and development of the resultant multiethnic colonial communities.

2. Historical Perspective

As late as 1938 the intensive study of contact between peoples was a 'relatively recent development in the anthropological repertory' (Herskovits 1937, p. 1). One term often used within these studies was 'acculturation.' Herskovits, however, found this word to be 'peculiarly American' and noted that the British had consistently preferred to employ the term 'culture-contact' (Herskovits 1937, p. 2). The subtle differences in these terms mark a trajectory that influences culture-contact perspectives to this day. While acculturation may be taken to refer to the ways in which some cultural aspect is taken into a culture and adjusted and fitted, it also implies some relative cultural equality between the giving and receiving cultures. In 1928 the word 'acculturation' was defined as 'the approximation of one human race or tribe to another in culture or arts by contact.' However, a later edition of *Webster's Unabridged Dictionary* (1934) added the phrase 'to cause or induce a people to adopt the culture of another' (Herskovits 1937, p. 2).

In 1940 acculturation was viewed as a process within a paradigm that saw cultures as 'infinitely perfectable' and in which rates of change will differ in 'so-called primitive cultures' (Linton 1940, p. 468). A sense of the developing colonial and unidirectional emphasis that characterized early culture contact studies might be inferred from a statement made by Malinowski in 1945. 'The study of culture change must take into account three orders of reality: the impact of the higher culture; the substance of native life on which it is directed; and the phenomenon of autonomous change resulting from the reaction between the two cultures' (1945, p. 26).

By 1955 American archaeologists had begun to recognize the complexity of contact contexts and identified eight types of culture contact situations: including intrusive culture, resident culture, site-unit intrusion, and trait-unit intrusion (Woodbury 1955). By 1961 the notions of 'differential change and both directed and nondirected situations' that have since been adopted were being acknowledged by many anthropologists and archaeologists (Spicer 1961).

3. Current Theory and Research

Today, culture contact studies are recognized as being well suited to evaluate and refine theoretical and methodological approaches to culture change and persistence in archaeology. The current archaeological approach subscribes to the potential of multiple lines of evidence by integrating data from archaeology, ethnography, ethnohistory, linguistics, biological anthropology, and native oral traditions. While no single theoretical position captures the field, certain ubiquitous issues include: (a) world economy; intensification of regional exchange and innovations in material culture, (b) demographic collapse; changing natural landscapes and lethal epidemics, and (c) implications of demographic collapse for social disruption and reformulation including native responses of resistance, domination, cooperation, and cultural transformation.

3.1 World Economy and the Intensification of Regional Exchange and Innovations in Material Culture

Many recent archaeological studies of contact contexts have been informed by a colonialist perspective of core–periphery relationships as exemplified in World Systems theory (Wallerstein 1974, Wolf 1982). Wallerstein's world systems theory remains the most influential work in archaeological studies of core–periphery relationships. While these studies have expanded the scope of research, some problems remain. Specific problems with colonialist models include (a) insular models of culture change that treat frontiers as passive recipients of core innovations, (b) the reliance on macroscales of analysis employed in frontier research, and (c) the expectation of sharp frontier boundaries visible in material culture (Lightfoot and Martinez 1995, p. 471). The study of material culture is the unique domain of archaeology and often the influence of material culture on social dynamics is under-rated. Material items are active symbols in broadcasting and even negotiating a person's identity in culture contact situations. However, the impact of introduced materials in these contexts is becoming well recognized. For example, the impact of the introduction of brass and iron kettles by French traders was so great that the cessation of ceramic production among aboriginal populations often occurred in less than a generation (Walthall and Emerson 1992, p. 6). Or, as Hantman suggests, copper could have been a key symbol, if not a source, of power and authority in matters secular and sacred—a critical variable in the dynamics of Jamestown (1990, p. 685). Not only items or categories of materials but architecture, roads, villages, and the spatial distribution of materials and villages on the landscape may reflect or influence similar changes.

3.2 Demographic Collapse, Lethal Epidemics, and Changing Natural Landscapes

Colonists brought with them pathogens that escaped from their original hosts to invade new territory. Because the natives lacked immunity to the introduced viruses and germs, they often succumbed in large numbers (Dobyns 1983, p. 8). Epidemics spread along established trade routes throughout the larger regions, and major outbreaks occurred at varying intervals, some striking at least once every generation. These events often left 'niches' to fill or made inoperative many conventional understandings evolved by large populations (Dobyns 1983). Any respectable theory that attempts to explain the European demographic advance has to provide explanations for the frequent demoralization and often annihilation of indigenous populations, as well as the 'stunning' success of European agriculture in colonialist territories (Crosby 1986).

3.3 Implications of Demographic Collapse for Social Disruption and Reformulation Including Native Responses of Resistance, Domination, and Cooperation

Archaeology has a critical role in addressing change in these contexts of social disruption and reformulation. It is only through analysis of the past that we can assess the precontact social structure, evidence of change, and the accuracy of the ethnohistoric or documentary record. As for native responses, the range of variability available to all people must be attributed to indigenous populations. These would include the possibilities of resistance, domination, and cooperation. For example, sedentary farming societies fit better into the colonial mission system than groups who were not sedentary agriculturalists, and attempts to establish missions among nonagriculturalists was not always successful (Milanich 1995). Revolts, resistance to tribute demands, conscripted labor, and threats to power and status were all potential responses. Brumfiel (1994) has summarized the importance of factional competition in the transformation of societies. These strategies include intermarriage, the creation of exchange partnerships, and the development of work cooperatives.

4. Some Problems and Future Directions

A discussion of how issues relating to archaeological approaches to culture contact articulate with the greater theoretical agendas that dominate anthropology today leads to the recognition of methodological issues or problems. While recognizing the great potential for an archaeology of pluralism, two issues were raised recently about the theoretical models and methodological practices employed in the study of

material culture in multiethnic contexts. The first issue concerned the development of more sophisticated 'contextual' approaches to study pluralism, which can compliment and broaden our current focus beyond artifact-based analyses (Lightfoot 1995, pp. 202–10). The second issue involved the promotion of multiscalar research that may enable archaeologists to address not only macroscale processes of world systems and colonial policies, but also microscale practices of individual intentionality and social action that are critical components of 'encounters' in pluralistic settings (Lightfoot and Martinez 1995, pp. 477–88).

For example, one approach builds on a crucial tenet of practice theory—that individuals will enact and construct their underlying organizational principles and worldviews in the ordering of daily life. By examining the organization of daily life in archaeological contexts—the spatial layout of residential space, the reordering of domestic tasks, the structure of trash disposal—we can critically evaluate the nature and magnitude of culture change and persistence in contact settings (Lightfoot et al. 1998, pp. 216–17).

For other scholars, changes in economic activities are changes in the structure of meaning. Debates over historical process often center on whether internal or external agents are precipitators of historical change. However scholars approach historical processes, they realize now that they were wrong in the past to assume that any culture has been isolated from the rest of the world or that change was unidirectional. Future directions involve issues relating to a redefined 'historical anthropology' and an increased interest in the development and negotiation of status hierarchies, ethnic identities, and gender relations in pluralistic colonial communities.

In summary, the current archaeological approaches to culture contact embrace a 'contextual' and holistic approach that integrates multiple lines of evidence and multiple perspectives.

See also: Cultural Assimilation; Ecology, Cultural; Hegemony: Anthropological Aspects; Hegemony: Cultural; World Systems Theory

Bibliography

Brumfiel E M 1994 Factional competition and political development in the New World: an introduction. In: Brumfiel E M, Fox J W (eds.) *Factional Competition and Political Development in the New World*. Cambridge University Press, Cambridge, UK

Crosby A W 1986 *Ecological Imperialism: The Biological Expansion of Europe, 900–1900*. Cambridge University Press, Cambridge, UK

Dobyns H F 1983 Widowing the coveted land. In: Dobyns H F (ed.) *Their Numbers Become Thinned: Native American Population Dynamics in Eastern North America*. The University of Tennessee Press, Knoxville, TN in cooperation with the Newberry Library Center for the History of the American Indian, Chicago, pp. 8–32

Hantman J L 1990 Between Powhatan and Quirank: Reconstructing Monacan culture and history in the context of Jamestown. *American Anthropologist* **92**: 676–90

Herskovits M J 1937 The problem. In: Herskovits M J (ed.) *Acculturation: The Study of Culture Contact*. J. J. Augustin Publisher, New York, pp. 1–32

Lightfoot K G 1995 Culture contact studies: Redefining the relationship between prehistoric and historical archaeology. *American Antiquity* **60**(20): 199–217

Lightfoot K G, Martinez A 1995 Frontiers and boundaries in archaeological perspective. *Annual Review of Anthropology* **24**: 471–92

Lightfoot K G, Martinez A, Schiff A M 1998 Daily practice and material culture in pluralistic social settings: an archaeological study of culture change and persistence from Fort Ross, California. *American Antiquity* **63**(2): 199–222

Linton R 1963 Acculturation and the process of culture change. In: Linton R (ed.) *Acculturation in Seven American Indian Tribes*. D. Appleton-Century Co., New York, pp. 463–82

Malinowski B 1945 *The Dynamics of Culture Change*. Yale University Press, New Haven, CT

Milanich J 1995 *Florida Indians and the Invasion from Europe*. University of Florida Press, Gainesville, FL

Spicer E H (ed.) 1961 *Perspectives in American Indian Culture Change*. (*A Summer Research Seminar*), (Albuquerque 1956) University of Chicago Press, Chicago

Wallerstein I 1974 *The Modern World-System*. Academic Press, New York, Vol. 1

Walthall J A, Emerson T E 1992 Indians and French in the Midcontinent. In: Walthall J A, Emerson T E (eds.) *Calumet and Fleur-de-lys: Archaeology of Indian and French Contact in the Midcontinent*. Smithsonian Institution Press, Washington, DC, pp. 1–13

Wolf E 1982 Iberians in America (Chapter 5)/The fur trade (Chapter 6). In: Wolf E (ed.) *Europe and the People without History*. University of California Press, Berkeley, CA, pp. 127–95

Woodbury R B (ed.) 1955 *Memoirs of the Society for American Archaeology*, no. 11, Seminars in Archaeology

A. Martinez

Cultural Critique: Anthropological

1. Cultural Critique and the Ethnographic tradition

Cultural critique, the use of anthropology to draw critical attention to institutions that readers take for granted, is as old as the origins of the modern discipline. It was a major purpose of Boas' students, especially Margaret Mead and Ruth Benedict. Its antecedents range from modernism to the critical theory of the Frankfurt school and the French surrealists between the wars; and the roots of critique lie in eighteenth-century philosophy.

What is 'cultural critique?' It is to examine the foundations of culture by having recourse to judgment. Judgment in turn is the ability to form an opinion on the basis of careful consideration; beyond

that to discern relations linking particulars to more general principles. Although not indifferent to fact and logic, judgment requires consideration of worth. A judge is respected for his or her wisdom and apparent objectivity, that is for the ability to transcend mere opinion, even by giving expression to universal truth. A cognate expression is critic (which is, after all, derived from the Greek word for judgment). But criticism most often connotes for us the formation of opinions about works of art and the term 'cultural critique' suggests a direct link between anthropology and literary criticism. Indeed, critical anthropology has involved of late a shift in emphasis from life to text, reflecting a trend driven by poststructuralist discourse from the 1970s onwards.

Ethnography has been the dominant form of anthropological enquiry in the twentieth century. It literally means writing about peoples considered to constitute natural units, on the basis of extended participation in and observation of their societies. It arose in opposition to the method that preceded it, evolutionary history. This last sought to explain how the world came to be unified by Western imperialism by postulating a racial hierarchy in which 'our' cultural superiority (having developed a machine technology based on scientific reason) was attributed to biological difference. World War II undermined the credibility of this claim of Western civilization to be founded on universal reason and opened up a space for anthropologists willing to assert the virtues of ways of life previously considered to be 'primitive' and inferior. Bronislaw Malinowski is rightly considered to be the pioneer of a new ethnographic approach stressing the integrity of contemporary exotic cultures. In *Argonauts of the Western Pacific* (Malinowski 1961) he made clear his aim to challenge the assumption that Western economic norms were either universal or inherently superior. Subsequently Edward Evans-Pritchard drew on liberal myths of Saxon resistance to Norman rule to dispute the identification of political order with the state, finding 'ordered anarchy' among the Nuer of the Sudan (Evans-Pritchard 1940).

The academic norm of scientific ethnography was made necessary by anthropologists' drive towards legitimacy as a social science discipline. They eschewed fiction of all kinds. The basis of their truth claims was participant observation, the fact that they had joined the people in some exotic location and shared their life as a means of studying them. Students were encouraged to read the classical monographs as scientific truth. The exercise of judgment was unnecessary, since professional vetting of the empirical object was sufficient guarantee of a monograph's validity. Within this framework of objectivity, ethnography was potentially critical of Western civilization's claim to universality. By stepping out of normal society, ethnographers pointed a sharp light on the assumptions of their readers, showing, for example, that witchcraft accusations made sense or that gifts may be instruments of domination. The general aim of anthropologists at this time seems to have been to discredit the racist assumptions underlying colonial empire. In the USA this trend was associated with a position known as 'cultural relativism': every way of life, however barbarous, has a right to exist and it is not for outsiders to judge it as inferior for any reason.

The founder of modern American anthropology, Franz Boas, was himself committed to educating public opinion (Boas 1929); but his students took this further by setting out to reach a general US audience with the surprising results of their own and others' ethnography. Americans think of their culture as natural, i.e., inevitable; by showing them cultural variation elsewhere they are forced to ask if their own culture is malleable. This perhaps accounts for why Ruth Benedict's *Patterns of Culture* (1934), an impressive collation of cultural variety worldwide, was the anthropology bestseller of all time. Its subtitle is particularly instructive: 'an analysis or our social structure as related to primitive civilizations.' In *Coming of Age in Samoa* (1928), her colleague Margaret Mead's topic was adolescence. Americans think of teenagers' turbulence as the natural result of sexual development; Samoan youth experienced adolescence more peacefully; so maybe they ought to think again about the reasons for teenage behavior. In a comparative work, *Male and Female: A Study of the Sexes in a Changing World* (Mead 1949), Mead piled example on example to demonstrate that the sexual division of labor thought to be natural in Western societies was a special case. It is not hard to imagine how such a book might enter the campaign for female emancipation. Mead has been subjected to savage criticism of late for the inadequacy of her fieldwork and the exaggeration of her deconstructive cultural contrasts. These attacks have come precisely when the cultural critique she pioneered has sought to undermine the premises of scientific ethnography itself, as she never did.

2. *Writing Culture*

Most recently, cultural critique in anthropology has been associated with an American school of ethnographers led by George Marcus. With James Clifford, he published *Writing Culture: The Poetics and Politics of Ethnography* (Clifford and Marcus 1986), a book which has come to be seen as marking the postmodern shift in anthropology. At much the same time, Marcus published, with Michael Fischer, *Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences* (Marcus and Fischer 1999; see also Marcus 1999). The growing prominence of cultural critique has been ascribed to a 'literary turn' associated with the work of Clifford Geertz since the early 1970s (Geertz 1988). Clifford has carved out his own contribution to resolving 'the crisis of representation' in a series of works, notably *The Predicament of Culture*

(1988). Paul Rabinow's *Reflections on Fieldwork in Morocco* (Rabinow 1977) prefigured the critical reassessment of ethnographic method. The journal *Critique of Anthropology* grew out of collaboration between radical anthropologists in Britain, Holland and the USA beginning in the 1970s.

The critique launched by *Writing Culture* had as its focus three aspects of the ethnographic enterprise: poetics, politics (both in the subtitle), and epistemology. By 'poetics,' attention is drawn to the fact that ethnographers make up what they write. Geertz had been more or less explicit about his ambitions as a creative writer for over a decade before. Language, in such a perspective, is never simply descriptive, but persuasive also. If ethnography is literature, the strident assertion by positivists of its factual (nonfictional) basis must be questioned. The 'politics' of ethnography refers to a triadic relationship between anthropologists, their subjects and their readers. The political economy of global inequality must be foregrounded in ethnographic descriptions, and author–reader relations are more problematic than the original formula allows for. Who are the monographs for, and what possibilities for collaboration exist beyond the stereotype of the ethnographer as a lone ranger in a pith helmet? Finally, the status of ethnography as a source of knowledge (epistemology), its methodology and configuration of knowledge as power, are called into question. Cultural critique in these terms examines the very intellectual foundations of modern anthropology.

What then have been the poetic and political purposes of the cultural critics and how effective was their epistemological critique? First, they were Americans who wished to engage more directly with American society. That is, they proposed to carry out ethnographic fieldwork in and write about mainstream activities and institutions. This required developing experimental methods, summed up as 'reflexivity,' the need for the ethnographer to be continuously self-monitoring in the light of feedback from society. Their aim was similar to Mead's in seeking to disturb conventional assumptions about the USA's place in the world. Techniques of authorship proved hard to subvert in practice. Although the boundary between fact and fiction became blurred, few anthropologists have openly embraced the novel as a model. The radical rhetoric of the cultural critics, their attempts to develop new constituencies and forms of collaboration ran up against the continuing dominance of the universities as anthropology's institutional home.

The ideal of the 'citizen ethnographer,' boldly synthesizing fieldwork-based research, critical theory and political commitment at home, remains a somewhat distant goal. If we can no longer read ethnographies as factual truth, how do we acquire the judgment to make sense of them, especially when cultural relativism has told us we should not judge others as the Victorian imperialists used to? Much postmodern writing adopts an ironic tone, and perhaps this is because it straddles unresolved contradictions of this kind. For critics are by definition judgmental, yet convinced relativists have nowhere to stand.

The main controversial aspect of this program is thus its confrontation with the realism, naturalism and objectification inherent in the positivist tradition of scientific ethnography. The traditional premise of ethnography, the study of other cultures for the sake of our own self-enlightenment, became more untenable under postcolonial circumstances. It fell to Marcus and his associates to make explicit the faults of a paradigm that had been unraveling for decades. The value of this American movement is much debated. Although the protagonists have sometimes tried to distance themselves from the 'postmodern' tag, preferring to speak of 'high modernism' or 'reflexive modernity,' cultural critique has been seen by its positivist detractors as evidence of a postmodern collapse into fragmentation, confusion and extreme relativism (Gellner 1992). Others seem to agree that we have entered a postmodern age and that cultural critique is an appropriate response to it.

3. *Other Services of Critical Anthropology*

Cultural critique has not been absent from the other main national traditions of anthropology. Indeed, in France the public's reception of intellectuals has always been readier than was ever the case among the Anglo–Saxons. Moreover, French anthropologists have never embraced a narrow cultural relativism, preferring to keep alive the Enlightenment's legacy of an appeal to universal reason. The recent publication of Marcel Mauss' political writings (Mauss 1997) shows how deep-seated this public engagement was for him. As his principal successor, Claude Lévi-Strauss wrote a popular classic, *Tristes Tropiques*, and exposed to critical view the central institution of French culture, table manners (Lévi-Strauss 1976, 1978). Among contemporary anthropologists, Marc Augé (1995) has successfully bridged the gap between the academy and popular culture, explaining the symbolic significance of Princess Diana's death and other contemporary events to readers of the serious newspapers. Britain produced the only systematic critique of contemporary civilization by an anthropologist so far in Edmund Leach's brilliant BBC lectures, *A Runaway World?* (Leach 1968). His attempt to address general fears concerning the explosion of machine production was premised on the need for his listeners to come to terms with their human potential and for anthropology's scope to match that potential. No one has come close to doing that since.

In the last decades of the twentieth century, feminist anthropology was in the forefront of cultural critique.

Indeed the absence of women from *Writing Culture* provoked a justified response (Behar and Gordon 1995). It was, after all, the women's movement that declared in the 1960s that 'the personal is political' and launched a devastating critique of Western institutions on grounds of the invisibility, exclusion and exploitation of women. These broader criticisms were readily applied to anthropology. An example of the sophistication reached by feminist and post-feminist discourse is Marilyn Strathern's *The Gender of the Gift* (Strathern 1988), where the confounding of Western gender stereotypes in Melanesian cultures is taken as a point of departure for a much wider critique of such core conceptual pairs as individual–society and nature–culture. The 1980s were a decade of deconstruction, what Hegel called 'negative dialectic,' in which the conventional categories of the modern synthesis became confused and discredited. Both inside and outside the academy, this task was performed to a disproportionate extent by women scholars.

Anthropology does not stand alone within an academic division of labor that has recently added 'cultural studies' to its ranks. An enhanced focus on the contemporary forms of cultural production has revived interest in the Frankfurt school of Adorno and Benjamin. Their concern with exposing the uses of culture in masking the contradictions of capitalism remains a powerful one. Perhaps even more relevant to critical anthropology is the work of the West Indian writer, C. L. R. James (Grimshaw 1992, Grimshaw and Hart 1993). In *American Civilization* (James 1993), he identified a growing conflict between vast anonymous bureaucracies and people's aspirations for an extension of democracy into all areas of their lives, an aspiration that could only be met indirectly through the movies and other forms of popular culture. For James, this was a crisis for modern society as a whole, most acutely revealed in the USA; and he addressed it through a critique ranging from Hollywood films and comic strips to psychoanalysis and the classical novels of Herman Melville.

4. *Eighteenth Century Roots of Critique*

As long as ethnographers reached the remotest corners of the planet, their work could claim to be anthropology in the sense of providing the widest possible context of cross-cultural comparison. It is not easy to see what is anthropological about the indigenous critique of American Society, when the term 'ethnography' is also used by geographers, sociologists, and literary critics. The limitations of cultural critique so far could be said to lie in two weaknesses. The epistemological critique of ethnographic method has been somewhat introverted, an in-house discourse maintained by a minor academic profession, when the issues raised are potentially of universal significance. And the aspiration to grasp the human condition in general seems to have been abandoned in favor of competing for public attention in the USA. Both of these deficiencies might be redressed if critical ethnographers examined their roots in the eighteenth-century philosophy that launched modern anthropology, as well as the notion of critique.

This is not the place to explore how Kant developed his original conception of critique, most notably in his great *Critique of Pure Reason*, but also, for our purposes, in the *Critique of Judgment* (see Cassirer 1981 for an inspiring introduction to each). In seeking a metaphysical ground between reason and understanding, Kant emulated Copernicus' revolution, showing how, instead of the object revolving around the spectator, the spectator could be seen as revolving, while the object remained at rest. There are so many echoes of Kantian subjectivity in the anthropological moment of cultural critique. What is missing is the aspiration to a universality of method. This is doubly inexcusable in that Kant invented the term 'anthropology' in its modern sense (Kant 1978). Surely a discipline with the remit of offering knowledgeable guidance on human teleology should not content itself with a jester's role, *épater les bourgeois*.

No one would dispute Jean-Jacques Rousseau's standing as the founder of modern cultural critique. He showed that a critical perspective on society, a refusal to take as inevitable things as they are, requires us to devise new methods of studying and writing about a transitional present. The boundary between fact and fiction has to be blurred if we are to talk about the evolution of possible worlds out of the actual one we live in. In developing a revolutionary critique of politics, education, sexuality and self-identity (*Le Contrat Social*, *Emile*, *La Nouvelle Héloïse*, *Les Confessions*), Rousseau came up with a different genre of writing for each topic.

Critical anthropologists today might emulate his example, were they not forced to justify themselves to the academic bureaucracy. The two discourses that launched Rousseau's career, however, are the principal sources for a renewal of critical anthropology, combining as they do a critique of corrupt civilization with an anthropology whose aim is to redress global inequality. In particular, his *Discourse on the Origins and Foundation of Inequality among Men* (Rousseau 1984), which has inspired anthropologists from Morgan and Engels to Lévi-Strauss, deserves to be seen as the first great work of the modern discipline.

Cultural critique has been an invigorating stimulus to new thinking in anthropology. It is uncertain whether it can flourish while its practitioners remain securely locked up in their academic enclaves. It is time in any case for a return to the universal outlook of the eighteenth-century pioneers.

See also: Anthropology; Benedict, Ruth (1887–1948); Critical Theory: Contemporary; Ethnography;

Gender and Feminist Studies in Anthropology; Geoarchaeology; Mauss, Marcel (1872–1950); Mead, Margaret (1901–78); Participant Observation

K. Hart

Bibliography

Augé M 1995 *Non Places: Introduction to an Anthropology of Supermodernity*. Verso, London

Behar R, Gordon D (eds.) 1995 *Women Writing Culture*. University of California Press, Berkeley, CA

Benedict R 1934 *Patterns of Culture: An Analysis of Our Social Structure as Related to Primitive Civilizations*. Penguin Books, New York

Boas F 1929 *Anthropology and Modern Life*. Allen and Unwin, London

Cassirer E 1981 *Kant's Life and Thought*. Yale University Press, New Haven, CT

Clifford J 1988 *The Predicament of Culture: Twentieth-century Ethnography, Literature and Art*. Harvard University Press, Cambridge, MA

Clifford J, Marcus G (eds.) 1986 *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA

Evans-Pritchard E 1940 *The Nuer*. Clarendon Press, Oxford, UK

Geertz C 1988 *Works and Lives: The Anthropologist as Author*. Stanford University Press, Stanford, CA

Gellner E 1992 *Postmodernism, Reason and Religion*. Routledge, London

Grimshaw A (ed.) 1992 *The C. L. R. James Reader*. Blackwell, Oxford, UK

Grimshaw A, Hart K 1993 *Anthropology and the Crisis of the Intellectuals*. Prickly Pear Press, Cambridge, UK

James C L R 1993 In: Grimshaw A, Hart K (eds.) *American Civilization*. Blackwell, Oxford, UK

Kant I 1978 *Anthropology from a Pragmatic Point of View*. University of Southern Illinois Press, Carbondale, IL

Leach E 1968 *A Runaway World?* BBC, London

Lévi-Strauss C 1976 *Tristes Tropiques*. Penguin, Harmondsworth, UK

Lévi-Strauss C 1978 *The Origin of Table Manners*. Jonathan Cape, London

Malinowski B 1961 *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagos of Melanesian New Guinea*. Dutton, New York

Marcus G (ed.) 1999 *Critical Anthropology Now: Unexpected Contexts, Shifting Constituencies, Changing Agendas*. School of American Research Press, Santa Fe, NM

Marcus G, Fischer M 1999 *Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences*. University of Chicago Press, Chicago

Mauss M 1997 In: Fournier M (ed.) *Ecrits Politiques*. Fayard, Paris

Mead M 1928 *Coming of Age in Samoa*. Mentor Books, New York

Mead M 1949 *Male and Female: A Study of the Sexes in a Changing World*. Dell, New York

Rabinow P 1977 *Reflections on Fieldwork in Morocco*. University of California Press, Berkeley, CA

Rousseau J-J 1984 *A Discourse on Inequality*. Penguin, Harmondsworth, UK

Strathern M 1988 *The Gender of the Gift: Problems with Women and Problems with Society in Melanesia*. University of California Press, Berkeley, CA

Cultural Diversity, Human Development, and Education

Humans are a unique species in that they can accumulate experiences in the form of culture, a special medium of life that consists of a set of interrelated and shared artifacts that their offspring can access and incorporate. Human development is thus achieved by incorporating these artifacts into the mind. The human species, more than any other, acquires behavioral characteristics through learning. Moreover, humans organize learning opportunities for their young to socialize or enculturate them in particular ways. Education has emerged from this species-specific tendency (Serpell and Hatano 1997). Hence there should be close, mutually supporting relationships between diverse cultures, differently organized ways of educating children, and correspondingly differentiated courses and products of individual development.

However, two standardizing factors have made the relationships among culture, human development, and education much more complex. First, contemporary cultures, and especially their educational aspects, share a large number of features that are needed to meet the demands of mass production–consumption economies and the values reflecting them. To survive in the present competitive world, most cultures are eager to promote in their children the learning of science and technology, and thus to enhance literacy, numeracy, and a beginning of an understanding of science. As a result, education in the form of institutionalized, basic public schooling has become almost universal across cultures. This form of schooling is, for example, conducted by teachers, who are regarded as specialists and are in charge of choosing their course contents and methods of teaching. The teachers usually rely on didactic methods that require students to manipulate symbols instead of handling real objects in meaningful contexts (Serpell and Hatano 1997).

Second, although the human mind develops by being socialized or enculturated, it has some universal features as a product of the species' evolutionary history. It is now accepted generally that human development is constrained by a set of innate tendencies and preferences, and as a result, there is one human mind, although there are many mentalities. This implies that even diverse cultures may produce a highly similar pattern of individual development. In

this article we will discuss these complicated relationships between culture, human development, and education, taking a few domains of growth as examples.

1. Socialization and Education in Infancy and Early Childhood

It has been universally documented by empirical research that affective relationships between young children and socialization agents enhance the children's susceptibility and tendency to comply with these agents which include parents, siblings, peers, and teachers. A nurturing person to whom a child has attached can induce the child's response to her or his signals, proposals, or demands and convey social influences. In fact, infants earnestly try to acquire information from an attached figure through social referencing when they interpret a situation as being unusual or urgent. Moreover, mothers of securely attached toddlers use less power-assertive and more affective strategies when they interact with their children than do mothers of insecure children. In turn, securely attached toddlers are more compliant with their mothers than are their insecure counterparts.

Although the affect-facilitation principle can be applied to the socialization of children in every culture, the content of socialization must vary between cultures. That is, there are cultural preferences in the goals of child development and also in the patterns of socializing interactions between children and caregivers.

For example, Weisner et al. (1994) investigated children's gender typing in conventional and nonconventional families, and found that the nonconventional parents encouraged their children to develop gender egalitarian values while displaying fewer gender-stereotyped behaviors than did their traditional counterparts. Miller et al. (1997) analyzed narratives directed to a target toddler by family members who were European Americans or Taiwanese. They found that the Taiwanese family members mentioned the past transgressions of the child, treating the transgressions as a didactic resource for conveying the moral standard, while American subjects referred to the past transgressions less frequently and with less seriousness, or they recalled children's past failures as entertainment for the family. Greenfield and her colleagues (e.g., Greenfield 1999) investigated cross-generational changes in the cultural transmission of weaving in a community in Mexico. Their observations indicated that the ecological transition of the society from an agricultural subsistence economy to commercial entrepreneurship over two decades had changed both the goals (the standard of high quality weaving) and the ways of teaching (imitation vs. creation) in this community.

Moreover, educational institutions for young children, nursery schools and kindergartens, have been constructed on every continent under the influence of Western philosophers of education, such as Pestalozzi, Montessori, and Piaget. Through participation in these educational institutions, children obtain knowledge and develop social skills. Some of the goals, tools, and processes of early childhood education are universal across cultures, but others vary owing to historical traditions, customs, and the everyday practices of the peoples of each culture. Through those experiences, children become appropriate successors for each society. For example, Western researchers who observe Japanese preschools are often impressed by the schools' emphasis on cooperation and group-living skills. These researchers speculate that participation in such cultural practices will help Japanese children to develop friendships and an interdependent sense of self, which make them able successors of the so-called interdependent culture (e.g., Tobin 1992).

2. Literacy, Schooling, and Human Development

Formal education in the contemporary world is often construed as a cumulative progression with two broad phases (Serpell and Hatano 1997): (a) the fundamental acquisition of basic literacy skills that are subsequently deployed by the individual, and (b) the acquisition of cultural knowledge and of a higher level of understanding. Thus, much of the initial phase of schooling focuses on the teaching and learning of basic reading and writing. This is because literacy is often construed as a passport to full participation in 'the modern world.'

However, whether the acquisition of literacy skills, both within the community and at the individual level, plays a critical role in human development is still debatable. Although a number of leading scholars in the field assert that literacy establishes a relationship between a word and its referent that is more general and more abstract than in oral communication, such views (called 'the great divide theories') have been criticized because they imply that nonliterate cultures or individuals are primitive or immature. Moreover, it has been demonstrated that the acquisition of literacy skills *per se*, separated from schooling and used in restricted situations, does not produce general cognitive effects. It seems reasonable to assume that orality has its own virtues (Olson and Torrance 1991).

It is reasonable, however, to claim that literacy induces some specific cognitive changes, depending upon what people do with it (Olson 1994). In technologically advanced societies many practices, including administration, religion, and science, require their participants to use the written language, so the acquisition of literacy often enables people to enter one of these practices. This will inevitably enhance the development of the body of knowledge and cognitive skills that are used in the target practice. As in-

formation technology develops further, new forms of literacy (e.g., what is called computer literacy) are now being sought.

Another interesting issue related to the acquisition of literacy in terms of the triangular relationships between culture, human development, and education is how different writing systems are learned differently, and how each culture supports such learning (Harris and Hatano 1999). Writing systems vary widely, even in the contemporary world. For example, as is well known, English is an orthographically irregular language, that is, there is an inconsistent (and, in some cases, arbitrary) relationship between letters and sounds. In contrast, in other languages such as Spanish and Italian, there is a highly regular relationship between sounds and letters. When learning an orthographically regular language, children become fluent readers once they master all the relationships between spelling and sound. It is possible for them to spell any word that they hear and to pronounce any word that they see written down.

There are, in addition, some writing systems in which letters or characters do not represent phonemes but a larger unit. That unit is either a morpheme or a syllable. The best known example of a writing system in which a single character represents a morpheme (basic unit of meaning) is Chinese. To read a logographic orthography such as Chinese, children do not have to segment a spoken word into phonemes to discover the script–sound correspondence. Thus, learning to read the first 100 or so words in these systems is easier than in any alphabetical writing system. However, becoming a competent reader in such a writing system is a very time-consuming process with a large memory load, since it contains a very large number of different characters that cannot easily be broken down into smaller subcomponents as in the case of alphabetically written words. The second type of nonalphabetic writing system is a syllabic system in which each syllable in the language is represented by one character. It is easy for young children to learn to read in the syllabic system of writing because syllables are more readily accessible to awareness than phonemes. Syllabaries in which each character represents a syllable or mora (subsyllabic rhythmic unit) are extensively used in Japan.

For different writing systems, different sets of preconditions are required on the part of the learner. For example, whereas a less sophisticated phonological segmentation strategy is needed for learning Chinese characters than that for alphabets, the ability to discriminate visually complex figures is more critical. Accordingly, different play and study activities for enhancing the development of the relevant preconditions are provided by the cultures. For example, the ability to segment a word into an onset and a rhyme is facilitated by nursery rhymes in English-speaking countries. Similarly, to make up for the difficulty of learning a large number of Chinese characters, the ability to read and write them earns the learner some special respect in the cultures in which these characters are used.

3. *Cross-national Differences in Mathematics Teaching and Learning*

Humans seem to possess some protoquantitative schemata, by which they can count, add and subtract. These schemata serve as the basis for buying and selling, as observed in street mathematics (Nunes et al. 1993). The educational practice of mathematics teaching and learning might be expected to be essentially the same all over the world, wherever schooling is instituted. However, a number of studies have yielded striking differences in mathematical achievements between American and East Asian children. For example, in the first study by Stevenson's group, Chinese and Japanese first-graders clearly outperformed American first-graders. At the fifth-grade level, the difference in performance between the American and the Asian children was even greater—even the highest average score for an American fifth-grade classroom was below the lowest average score of an Asian fifth-grade classroom (Stevenson and Stigler 1992).

There are also remarkable differences in a number of features, both quantitative and qualitative, of the mathematics teaching and learning practices in China, Japan, and the USA. Asian children spend a larger number of hours on mathematics, and are led by teachers more often than their American counterparts. Whereas American teachers have a tendency to ask as many questions as possible, Asian teachers often spend an entire class discussing one or two problems. Also, whereas American teachers often give immediate feedback to students' answers, Chinese and Japanese teachers are likely to ask students to explain their answers and then call upon other students to evaluate their correctness.

These observed differences in the teachers' behavior may be attributed to their different folk-pedagogies as regards the best way for students to learn mathematics (Hatano and Inagaki 1998). They may constitute part of more general cultural beliefs about learning and teaching. Stigler and Hiebert (1999) suggest that, based on a video study of eighth-grade mathematics teaching in Germany, Japan, and the USA, which was part of the Third International Mathematics and Science Study, American teachers believe that the more exercises given, the more smoothly learning will proceed, whereas Japanese teachers believe that students learn best by struggling to solve a problem and also hearing about the pros and cons of different methods.

To summarize, although apparently highly similar public schooling is institutionalized almost everywhere

in the world, the process of actual learning and teaching in mathematics differs even among technologically advanced countries, owing to their different cultural backgrounds. Such differences tend to produce widely different levels of mathematical achievements.

4. *Intervention for Citizenship*

For the creation and continuation of society, every society needs competent members. However, the kinds of knowledge that are required and the extent to which they have to be established depend on various cultural backgrounds, including history, politics, economics, and religion.

In technologically advanced countries, this need has been conceptualized as acquiring citizenship and has provoked relevant research and education. Social scientists and educators are examining whether people, including young children, can develop social awareness through understanding the history, politics, and economics of the society, and also urgent problems such as human rights and environmental issues. It is expected that, as members of society, citizens should recognize these problems as well as be familiar with their entire social system. In addition, they should understand how everyday practices in which they participate relate to other people and other social systems, and know that they are more or less responsible for social events that happen anywhere on earth.

However, it has been documented that it is not easy for ordinary people to acquire accurate societal knowledge. For instance, in his monumental book, Furth (1980) showed that even adolescents, who had supposedly acquired formal operations in cognitive development, could not spontaneously reach the stage of fully understanding social institutions or mechanisms. Moreover, studies on people's understanding of the banking business, as representative of the kind of economic knowledge needed to grasp the present capitalistic world, indicated that adolescents and young adults from middle-class, salaried families in Glasgow, Leiden, and Tokyo possessed only limited knowledge. They developed naive theories about the banking business by borrowing rules from everyday personal interactions that were basically governed by ethical and humanistic considerations (Takahashi and Hatano 1999).

However, it has also been documented that children who are involved in economic activities and live in a highly commercialized area have an enhanced economic understanding. For example, Jahoda (1981) found that children in Harare, Zimbabwe who were working as sellers understood profit-making mechanisms of shops more precociously than did their less-experienced counterparts in Scotland and New Zealand. In addition, it has been observed that people of different social classes develop their own social representations as to social events and systems. For instance, adolescents from middle-class families tend to offer more individualistic explanations for poverty than do their working-class counterparts. Thus, we can assume that formal and informal interventions will advance and elaborate some kinds of societal understandings. In fact, Youniss and Yates (1997) studied how a school-based social activity, i.e., community service experiences at a soup kitchen for the homeless, provoked reflection on poverty, racism, and various social problems among high-school students from a major city in North America. In this case, in a year-long project on social justice, students were both taught in a classroom and required to work four times at a soup kitchen. In Germany, there are many kinds of educational programs for students to learn human rights after Auschwitz. In one project, high-school students cleaned a Jewish cemetery to learn the fate of the Jewish people in Nazi Germany (Rathenow and Weber 1998).

In addition, awareness of inequality and unfairness among the weak in society, i.e., females, children, elderly persons, and handicapped people, has gradually transformed conventional views and beliefs regarding the nature of human beings and human development. For example, Gilligan (1982/1993) pointed out that social sciences have long neglected the fact that females have been socialized and educated to have different voices from males. Although there are no decisive biological differences between the sexes that can rationalize different socialization and education, many studies have indicated that from the birth on, children in almost all cultures are differently socialized and educated through both visible and invisible biased programs of gender typing. Some progressive researchers in Japan have shown by analyzing the content of school textbooks that there are male-dominant biases in both the texts and illustrations.

Furthermore, there are a variety of other opportunities to enhance our societal understanding. For instance, empirical data about the weaker members of society advance people's understanding of human rights. In fact, certain findings have shown that people over 70 years old are, despite the negative image of old age, cognitively and socially competent. This result surely promotes fresh perspectives of successful aging. In addition, narratives about the so-called handicapped may challenge the conventional idea of normality. A book written by a Japanese student who has congenital tetra-amelia, but enjoys an active and satisfying college life, has deeply moved people. Over 5 million copies have been sold. It is reasonable to expect that passing laws and insisting on international treaties that protect women and children will advance people's understanding of the meaning of human rights.

5. Conclusion

Socialization and education are ubiquitous to the extent that people convey knowledge, skills, beliefs, and attitudes. They seem to be based on human species-specific properties. Although we are often impressed by cultural diversity, that is, cultural differences in developmental goals and methods, patterns of human development are also surprisingly similar across cultures. In the era of globalization, these universal aspects of human development should be taken seriously in education as well as in politics and the economy.

Since the end of the Cold War, people have been allowed to identify themselves with ethnic and other local groups within the nation-state and/or with even larger groups extending beyond national borders. To participate and function fully in this complex, hierarchically organized global community, people have to respect other cultures while preserving their own culture. How we can cultivate a sense of global citizenship among children should be a universal task of education.

See also: Literacy Education; Mathematical Education; Writing Systems; Writing Systems, Psychology of

Bibliography

Furth H G 1980 *The World of Grown-ups: Children's Conceptions of Society*. Elsevier, New York

Gilligan C 1982/1993 *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA

Greenfield P M 1999 Cultural change and human development. In: Turiel E (ed.) *Development and Cultural Change: Reciprocal Processes. New Directions for Child and Adolescent Development* **83**: 37–59

Harris M, Hatano G (eds.) 1999 *A Cross-linguistic Perspective on Learning to Read and Write*. Cambridge University Press, Cambridge, UK

Hatano G, Inagaki K 1998 Cultural contexts of schooling revisited: A review of the learning gap from a cultural psychology perspective. In: Paris S G, Wellman H M (eds.) *Global Prospects for Education*. American Psychological Association, Washington, DC, pp. 79–104

Jahoda J 1981 The development of thinking about economic institutions: The bank. *Cahiers de Psychologie Cognitive* **1**: 55–73

Miller P H, Wiley A R, Fung H, Liang C 1997 Personal storytelling as a medium of socialization in Chinese and American families. *Child Development* **68**: 557–68

Nunes T, Schliemann A L, Carraher D 1993 *Street Mathematics and School Mathematics*. Cambridge University Press, New York

Olson D. R 1994 *The World on Paper*. Cambridge University Press, Cambridge, UK

Olson D R, Torrance N (eds.) 1991 *Literacy and Orality*. Cambridge University Press, New York

Rathenow H, Weber N H 1998 Education after Auschwitz: A task for human rights education. In: Holden C, Clough N (eds.) *Children as Citizens: Education for Participation*. Jessica Kingsley, London, pp. 95–112

Serpell R, Hatano G 1997 Education, schooling and literacy in a cross-cultural perspective. In: Berry J W, Dasen P R, Saraswathi T S (eds.) *Handbook of Cross-cultural Psychology, Vol. 2: Basic Processes and Human Development*. Allyn and Bacon, Boston, pp. 339–76

Stevenson H H, Stigler J W 1992 *The Learning Gap*. Summit Books, New York

Stigler J W, Hiebert J 1999 *The Teaching Gap*. Free Press, New York

Takahashi K, Hatano G 1999 Recent trends in civic engagement among Japanese youth. In: Yates M, Youniss J (eds.) *Roots of Civic Identity: International Perspectives on Community Service and Activism in Youth*. Cambridge University Press, Cambridge, UK, pp. 225–44

Tobin J 1992 Japanese preschools and the pedagogy of selfhood. In: Rosenberger N R (ed.) *Japanese Sense of Self*. Cambridge University Press, Cambridge, UK, pp. 21–39

Weisner T S, Garnier H, Loucky J 1994 Domestic tasks, gender egalitarian values and children's gender typing in conventional and nonconventional families. *Sex Roles* **30**: 23–54

Youniss J, Yates M 1997 *Community Service and Social Responsibility in Youth*. University of Chicago Press, Chicago

G. Hatano and K. Takahashi

Cultural Evolution: Ethnogenesis

The term 'ethnogenesis' is an obvious construction between two classical Greek roots, *ethnos* for tribe or nation and *genesis* for birth or beginning, and consequently the word has been invented and reinvented several times. In general, the term has been used over the last two centuries to describe the processes by which tribes, nations, and ethnicities have come into existence in history and prehistory. However, the various major schools of thought, French, Soviet, Central European, ethnohistorical, and sociocultural, have proceeded using different methodologies, and their subject matter has also been somewhat different. Intellectually, ethnogenetic theories have tended to serve as counterpoints to racial theories of human and national origins. More than most scholarly concepts, these theories have frequently been invoked by one side or another in hotly contested political disputes (see *Cultural Evolution: Phylogeny versus Reticulation*).

1. The French Tradition

Probably the first to construct the term ethnogenesis was André-Marie Ampère, the French physical scientist who published his *Essai sur La Philosophie des Sciences* in 1834. In the course of outlining 'all human knowledge,' he created a typology in which the science of ethnology was said to comprise ethnography, geography, and *ethnogénie*. His conception of ethnogenesis was actually quite modern, since he hypothe-

sized that a new ethnic group would usually arise from people of shared descent, or having common beliefs, or occupying a shared territory. But he did not allege, as later theorists did, that these three characteristics necessarily went together.

Several works of history explicitly labeled ethnogenetic were created later in nineteenth-century France, but it is not clear if they were directly inspired by Ampère. Whether or not, a standard method of ethnogenetic research was invented in this period, called the chronological method, which consisted of the following steps: (a) identification of possible ethnic antecedents for modern and recent nations by analyzing and criticizing ancient writings, particularly those in Latin; (b) evaluation of which antecedents were most influential in creating the culture of a descendant nation, and (c) assertion of the cultural primacy or hegemony of certain sets of cultural institutions over others, for example concluding that France is primarily a Celtic nation, or a Roman one, on the basis of language or legal principles.

A good illustration of this methodology is contained in *Les Origines des Aquitains* by Alphonse Castaing, published in Paris in 1885. Taking as his challenge the idea that the Aquitains of southwestern France were originally Basques from the Iberian Peninsula, Castaing mobilized passages from Caesar and Pliny. He identified 'cognates' among the ethnonyms used by classic authors (Galls, Belges, Bretons, Basques, Gascons, etc.) and put the cultural and demographic influences of different immigrations into chronological order. Then he concluded that 'the Aquitains are Gallish, as Caesar declared and as the ancient doctrine admits.'

Another work in the French school is *Ethnogénie Gauloise* by de Belloguet (1872) that emphasizes linguistic evidence. The author concluded that the *langue Gauloise* is 'not Germanic, but entirely Celtic.' Throughout the French school, we find the assumption that every society has a cultural core which can be reliably assigned to just one antecedent. This assumption is analogous to the premise in historical linguistics that every language has a syntactic and semantic core that can reliably be assigned to one and only one language family.

2. *Racial Histories*

It is no accident that French ethnogenetic histories appeared soon after the publication of racial histories by Gobineau and other German authors beginning in 1853. In these racial histories, the French and other European nationalities are cast as imperfect Germans, the results of interbreeding between the Nordic or Aryan 'race' and indigenous Celtic and Slavic peoples. The more general evolutionary theory is that there were originally in antiquity a small number of allegedly 'pure' human races which interacted historically, culturally, and biologically over the centuries to create the distribution of European ethnicities observed in the nineteenth century. In this tradition, which continued through Houston Chamberlain and Herman Wirth, and which finds its ominous climax in the writings of Nazi theorist Alfred Rosenberg, the defined races are said to be not only biologically distinct, but culturally distinct, and each is endowed with a specific level of intellect, and a unique morality and character (Shipman 1994). While Rosenberg emphasized the superiority of the Nordic 'Sun-People,' who he said originated in Atlantis, now sunk between Iceland and Greenland, the Italian Fascists emphasized Roman antecedents for modern Italians, and they alleged that Romans were culturally and biologically superior to Celtic, African, and other especially Mediterranean peoples (Cecil 1972)

These racial histories embodied theories of cultural degeneration that hark back to traditional Christian ideas about the descendants of Shem, Ham, and Japheth, and the wandering tribes of Israel embraced in previous centuries. The general idea is that once upon a time there were clearly defined races in a world where everyone was in their correct place, geographically and socially. But then migration and interbreeding gradually eroded this cultural balance, producing 'mixed races' which were inferior to the originals. These, then, are not theories of cultural evolution, but of cultural devolution, promising that the 'superior' races could reemerge under something like the Nazi political program of segregation and genocide. By contrast, the ethnogenetic theories, with some notable exceptions, have tended to be progressive in content. In a manner analogous to biological evolution, new 'species' of ethnic groups are said to arise continually, pushing aside preexisting ethnicities until the new ethnic formations are in turn pushed aside. Each new ethnic group is supposed to represent a cultural improvement over previous ones, so that overall this process is supposed to propel the human species toward something called 'civilization.'

A prominent but bizarre theory of evolution, explicitly labeled 'ethnogenetic,' emerged in Brazil in the mid-twentieth century. Writing in 1940, Domingos Magarinos argued that the human species evolved in the Central Plateau of Brazil during the late Tertiary. Further, 'languages, alphabets, beliefs, laws, everything originated in Brazil and from Brazil it was diffused to the rest of the world.' As strange as this sounds in light of modern archaeology and palaeontology, such extreme statements call attention to the essentially chauvinist and patriotic nature not only of this book, but of many other essays in the traditions of ethnogenesis and racial history.

3. *The Soviet Tradition*

The Soviet conception of ethnogenesis or *etnogenés* was apparently not derived directly from the French tradition, despite some similarities, but developed in-

dependently from Marxist discussions of 'the national question.' In its Stalinist form, a nation or 'ethnos' was defined as '… a historically constituted, stable community of people, formed on the basis of a common language, territory, economic life, and psychological make-up manifested in a common culture' (Stalin 1975). Under Soviet law, those nations and ethnic groups that qualified for non-Russian ethnic status were supposed to be given certain privileges of language and culture, and to enjoy limited autonomy and sovereignty within the boundaries of the Soviet Union. Over time, these minority groups were supposed to become increasingly Russified, voluntarily losing their own languages and cultures.

Like the French tradition, the Soviet tradition arose to some extent as a response to racial histories. In this regard it is interesting to note that Alfred Rosenberg was resident in Moscow as a student during the 1917 revolution. Contemptuous of Slavic peoples, Rosenberg wrote blueprints in the late Nazi period for a territorially expanded Third Reich in which Russians and Poles would serve as manual laborers for the German master race. Rosenberg alleged that the Slavic peoples, as well as the Lithuanians, Hungarians, Gypsies, Czechs, and other 'inferior races,' had no history of their own, except that which they derived from interactions with more 'advanced' races

After Soviet scholarship was reorganized on a 'revolutionary' basis in 1934, the task of writing official histories for the various component nations of the Soviet Union was split between the discipline of history, which included archaeology, and ethnography, which was increasingly focused on Soviet minorities in Siberia and Central Asia. History was especially entrusted with refuting 'German nationalist' conceptions of European history, and to this end the theory of ethnogenesis was invoked to describe and explain the actual origins of the Russian people (Mehnert 1952)

To refute the 'Nordic' view, Soviet historians denied that the original 'Rus' were from Scandinavia, asserting instead that the ethnogenesis of the Russian ethnicity had been entirely 'autochthonous.' In this view, the ancient Slavic peoples had always lived where they were located in modern times, and were the same as the Scythians of classical times ($\sim$ 500 BC). Most importantly, for political purposes, it was alleged that while the ancient Slavs had interacted with Varengian traders from the north and with literate societies such as Byzantium, they had never been under their domination but had always been a coequal partner in the development of Western civilization

To describe ancient Russian and Soviet history, which became more and more synonymous during the Stalin period, historians utilized the evolutionary periods defined by Marx and Engels: primitive, slave, feudal, capitalist, socialist, and communist. The stages of Russian history and the status of their historical allies and antagonists were also characterized in these terms. Scores if not hundreds of Soviet historians cooperated in this effort, producing hundreds of revisionist and Soviet-flattering books and articles between 1935 and 1990, predominantly in Russian.

While the word ethnogenesis soon faded in Soviet historical studies in favor of a more conventional vocabulary involving 'nation' and 'nationality,' the word and its embodied theory blossomed in Soviet anthropology. Although the terms 'ethnos' and 'ethnogenesis' were proposed for Soviet anthropology as early as 1921 by S. M. Shirokogoroff, the theory did not truly catch fire in the Soviet Union until after World War II. In the meantime, there was a complex jockeying for position among Soviet academic disciplines, and there was the ongoing problem of how to maintain a theoretical and philosophical position within an academic discipline that was compatible with the ever-changing party line of official Soviet Marxism-Leninism (Slezkine 1991).

Practically and philosophically, these problems were solved for anthropology in the late 1950s and early 1960s by Soviet theorist Julian Bromley, who proposed that ethnography (which included what we would call ethnology in English) should focus on the ethnos and especially should investigate the processes of ethnogenesis among small-scale Soviet societies. He defined the ethnos in an orthodox Marxist manner which reflected earlier discussions of the national question, as follows: 'Ethnos (in the narrow sense of the term) can be defined as a firm aggregate of people, historically established on a given territory, possessing in common relatively stable particularities of language and culture, and also recognizing their unity and difference from other similar formations (self awareness) and expressing this in a self-appointed name (ethnonym).' Contrasting the process of ethnogenesis with racial history, he says that ethnogenesis 'shows that all modern peoples trace their origins to a diversity of ethnic components and are mixed in composition, and consequently refutes racist, chauvinist inventions about the "racial purity," "indigenous ancestry," and "national exclusiveness" of certain peoples.' This perspective on the various ethnic groups of the Soviet Union gained official support in the Communist leadership, and resulted in increased approval and support for Soviet anthropology. Essentially, the discipline of anthropology was given the franchise for investigating the social and cultural status of Soviet minorities. Their findings and recommendations concerning the language and culture of over a hundred minority groups often found a sympathetic audience among Soviet bureaucrats entrusted with administering the 27 'republics' and 'autonomous regions' within the Soviet Union.

As leader of the Institute of Ethnography in Moscow, Bromley commissioned hundreds of anthropological field trips throughout the Soviet Union to collect data on language, demography, biology, economics, and culture, resulting in a large number of

publications. Between 1977 and 1982, for example, Soviet anthropologists produced more than 500 books and articles, many of them directly addressed to issues of ethnogenesis. It is no exaggeration to say that during its most active period, anthropologists connected with the Institute produced more pages of ethnogenetic study than the rest of the world before or since. However, because of the Cold War and a lack of literacy in Russian among Western scholars, these publications have not had the academic impact that they might have had under other circumstances.

Ironically, and as a post-Soviet footnote to the era of ethnogenetic studies, the fragmentation of the Soviet Union has seen the use of earlier ethnogenetic theories to deny Russian influences on former Soviet minorities and allies. Recent Romanian historical scholarship, for example, has returned to an emphasis on their Latin roots, minimizing the significance of their historical relationships with Slavic peoples. Peoples of the Caucasus have done the same. And so ethnogenesis is now being used to allege social and even biological 'purity,' purposes opposite to those intended by Bromley.

4. *The Central European Tradition*

Another independent ethnogenetic school of thought developed in the countries of Central Europe after World War II, largely Austrian and mostly published in German. This scholarly florescence was, once again, a reaction against the race-based culture histories offered by Fascist and Nazi historians and pre-historians before 1945. This postwar school of *ethnogenese* tends to consider the same range of problems as the earlier French ethnogenetic tradition, but without as much concern for general theory or for creating an explicit shared definition of ethnogenesis.

The results of these efforts, presented at periodic *Limeskongresses* in Central Europe since 1949 and as part of other activities sponsored by the Austrian Academy of Sciences, constitute an important reconsideration of the origins of all European nations as well as a sprinkling of other societies around the world (Wolfram and Pohl 1990). In their vocabulary and range of interests, this network of ethnogenetic scholars is much more historical and archaeological than it is anthropological. Their implicit evolutionary theories are largely progressive and somewhat Eurocentric. They see 'progress' in terms in increases in social scale, improvements in infrastructure, the development of capitalist democracy, and higher literacy.

5. *Ethnohistory*

In recent decades, ethnogenetic studies of tribal ethnohistory have had a renaissance in the United States, inspired largely by a classic 1971 article by William Sturtevant. Taking his cue from Soviet scholars, Sturtevant used the concept to describe the separation of the Seminoles from the Mvskoke Creeks in the period 1780–1820, and he defined ethnogenesis simply as 'the establishment of group distinctiveness.' In his writing, he has elaborated this simple definition to include distinctiveness in territory, economy, and political structure.

In the hands of American scholars such as Nancy Hickerson and Ernest Burch, ethnogenetic theory has lost much of its political baggage and become a tool for doing ethnohistory, a hybrid field between history and ethnology which analyzes tribal and ethnic histories during that shadowy period between prehistory and history, where there exist only cursory and distorted ethnographic descriptions. During the period of Spanish colonialism, for example, Hickerson uses Spanish documents to show how the so-called 'Jumanos' underwent two centuries of migrations from Mexico to Canada and a series of ecological revolutions and basic ethnonymic transformations before emerging on the Central Plains as the 'Kiowa Indians' in the early nineteenth century. On a more limited time scale, Ernest Burch and his colleagues have shown how a group of Athapaskans in Alaska were ethnogenetically transformed into 'Eskimos' between 1860 and 1880 (Hill 1996).

6. *Sociocultural Approaches*

In 1962, American sociologist Lester Singer once again reinvented the word 'ethnogenesis' to describe 'the process whereby a people, that is an ethnic group, comes into existence.' This theme, and the word, was picked up by other sociologists and some sociocultural anthropologists for the study of ethnic or national minorities within complex societies. Independently, Belgian anthropologist Eugeen Roosens has developed his own brand of ethnogenetic studies, addressing himself not only to tribal societies but to European immigrant minorities. His work, like that of sociologists, is more clearly inspired by theories of ethnicity than of ethnogenesis, since it considers only a shallow time depth and is not much concerned with evolutionary theory.

7. *Remaining Issues*

Currently, the global spectrum of ethnogenetic studies ranges from the strictly humanistic to the thoroughly scientific. On the humanistic side, historians and some sociologists are largely indifferent to the task of defining general principles of ethnogenesis. The more politicized scholars of ethnogenesis, in fact, often assert that the processes that produced their particular nation or ethnic group were absolutely unique. At the other end of the spectrum, prehistorians, ethnohistorians, and ethnologists of a scientific orientation are very much preoccupied with developing and refining theories of ethnogenesis that are consonant with and

in fact are constituent parts of general evolutionary theory. From the great volume of data already in hand and still being collected, specific and elaborate scientific theories of ethnogenesis, clearly articulated with evolutionary theory, are sure to emerge.

The most imminent current problem for ethnogenetic scientists, however, is determining how many processes of ethnogenesis there might be, and characterizing the differences among them. Perhaps there are some aspects of ethnogenesis that are universally typical of the process, no matter what its scale or geographical location, just as there are recurrent features of millenarian processes and political revolutions. Once these ubiquitous features are accounted for, the more significant problem for theories of cultural evolution is whether the remaining unique or variable aspects of ethnogenesis can be placed in some kind of chronological order, or otherwise seriated in some unilinear or multilinear and universal sequence. If they can, then ethnogenesis can be counted among the most important processes of cultural evolution.

See also: Cultural Evolution: Phylogeny versus Reticulation; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Ethnogenesis in Anthropology; Ethnology; Ethnomethodology: General; Gender, Class, Race, and Ethnicity, Social Construction of; Group Processes, Social Psychology of; Race: Genetic Aspects; Racism, Sociology of

Bibliography

Ampère A-M 1834 *Essai sur la Philosophie des Sciences*. Chez Bachelier, Paris

Arutiunov S 1994 Ethnogenesis: Its forms and rules. *Anthropology and Archeology of Eurasia* **33**(1): 70–93

Bromlei I 1982 *Present-day Ethnic Processes in the USSR*. Progress Publishers, Moscow

Castaing A 1885 *Les Origines des Aquitains*. Maisonneuve freres and ch, Paris

Cecil R 1972 *The Myth of the Master Race*. Dodd Mead, New York

Hill J D 1996 *History, Power and Identity: Ethnogenesis in the Americas, 1492–1992*. University of Iowa Press, Iowa City, IA

Mehnert K 1952 *Stalin Versus Marx*. George Allen and Unwin, London

Michael H N (ed.) 1962 *Studies in Siberian Ethnogenesis*. University of Toronto, Toronto, Canada

Roosens E E 1989 *Creating Ethnicity: The Process of Ethnogenesis*. Sage, London

Shipman P 1994 *The Evolution of Racism*. Simon and Schuster, New York

Singer L 1962 Ethnogenesis and Negro-Americans Today. *Social Research* **29**(4): 419–32

Slezkine Y 1991 The Fall of Soviet Ethnography, 1928–38. *Current Anthropology* **32**(4): 476–84

Stalin J 1975 *Marxism and the National-Colonial Question*. Proletarian Publishers, San Francisco

Tishkov V 1998 US and Russian anthropology: Unequal dialogue in a time of transition. *Current Anthropology* **39**(1): 1–17

Wolfram H, Pohl W 1990 *Typen der Ethnogenese unter besonderer Berücksichtigung der Bayern, Teil 1*. Verlag der Österreichischen Akademie der Wissenschaften, Wien

J. H. Moore

Cultural Evolution: Overview

Cultural evolution is one of the most hotly debated of all concepts in the social and behavioral sciences. There are many reasons for the debate, but some of the most important issues stem from the divergent answers of scholars to three basic questions.

1. What is Cultural Evolution?

First, there is debate over what the expression 'cultural evolution' refers to. What is it that evolves? This domain of debate is hardly surprising since 'culture' itself is notoriously difficult to define and delimit. Historically, however, the issues here cluster around two main schools of thought. One school follows a broad, holistic definition of cultural evolution that includes changes over space and time in the combined cultural and social systems of a human population. This definition covers big terrain, including both people's values, ideas, and beliefs—the ideational phenomena of cultures—as well as the organization and structure of societies (as reflected in the size and centralization of polities, the degree of stratification, level of social integration, and so on). This school looks for either universal tendencies across vast expanses of space, time, and human experience—so-called 'general evolution'—or for variable trajectories of change within particular societies or groups of societies— 'specific evolution.' Credible examples of this approach include Johnson and Earle (1987) and Sahlins and Service (1960). Today, the first school is commonly called 'social evolution' (see *Social Evolution: Overview*).

The other school takes a narrower definition of cultural evolution and focuses on the dynamics of change in the ideational systems of populations. This school attempts to trace the history of ideas, values, and beliefs that are held by groups of people, and to ask what processes have guided the historical trajectories of change in these phenomena? What processes have made certain ideas and beliefs prominent, in the sense of being widely held and acted on within a group, and others ephemeral or unimportant? This school is sometimes described by analogy: if one regards the cultural system of a human population as a collection of socially transmitted texts in the minds of culture bearers then one can ask, what is the social history of these texts? How were they written and rewritten, and by whom? What are the main processes and actors that

have historically edited the texts? And so on. Such questions today are at the heart of this school of cultural evolution, sometimes called 'evolutionary culture theory' (Durham 1990, 1992). This second school, focused on socially-transmitted information in human populations—though not exclusive of socially-conveyed information in nonhuman populations (see, for example, Bonner 1980, Heyes and Galef 1996, Wrangham et al. 1994)—is the topic of the remainder of this entry.

2. *In What Sense Do Cultural Systems Evolve?*

A second question in cultural evolution concerns the meaning of 'evolution' itself, another one of those words that can mean many different things (see *Evolution, History of*). Curiously, the word 'evolution' has had two quite different categories of meanings as it applies to the study of culture change. First, the term has referred to predictable trajectories of change in the social and cultural properties of human populations. In this category, evolution has been taken to mean stage-wise progression from earlier to later forms through regular transitional steps. This meaning of evolution is especially prevalent in the unilinear trajectories (e.g., from simple to complex, from sexual promiscuity to the conjugal family, from savagery to civilization, etc.) of the 'classical evolutionists' of the nineteenth century (e.g., John Lubbock, Henry S. Maine, John F. McClennan, Herbert Spencer, Edward E. Tylor—reviewed in Stocking 1987). A similar meaning of the term is also found among so-called 'neo-evolutionists' of the twentieth century (including Leslie White, Julian Steward, Elman Service, and Kent Flannery), who commonly propose parallel, multilinear trajectories of change (reviewed Sanderson 1990). Because these arguments, old and new, feature the changing social organization of populations, again they are more commonly called 'social evolution' today (see *Social Evolution: Overview*).

The second category of meanings for cultural evolution features a close analog of Darwin's original definition, 'decent with modification,' only as it applies to the cultural (ideational) features of distinct populations. Here the focus is on transformational change either as (a) one cultural system changes over time and space; (b) one cultural system differentiates into two or more distinct systems; or (c) two or more cultural systems merge into one new one. The first of these transformations (a) is also called cultural 'microevolution'—change within a given cultural system over time. The second and third transformations, (b) and (c) together, comprise what is sometimes called cultural 'macroevolution'—that is, pattern and process in the diversification (or homogenization) of new cultural systems either through the division (branching) of one system or the fusion (reticulation) of two or more cultural systems. Because new ideational systems often emerge through the formation (whether by fission or fusion) of new and distinct 'ethnic groups,' readers interested in cultural macroevolution see *Cultural Evolution: Ethnogenesis*; *Cultural Evolution: Phylogeny versus Reticulation*. The remainder of this article focuses on cultural microevolution.

3. *What Causes Cultural Microevolution?*

This third question, what causes change within a given cultural system over time, is an area of active research and discussion in the social sciences. The question entails at least two main areas of contemporary debate. The first debate concerns the nature of the elements of which cultural (ideational) systems are composed—the so-called 'units of culture.' If cultural systems are composed of socially transmitted information, so the argument goes, then what exactly are the dimensions and properties of the information under conveyance? Are cultural phenomena best modeled as fairly uniform little pieces or 'particles' during social conveyance? Or is it more accurate to view cultures as interconnected systems or hierarchies of information that do not readily 'chunk' into uniform small pieces during social transmission? This productive, ongoing debate (reviewed in Weingart et al. 1997) has spawned what might be called another school of cultural evolution theory—the 'meme' school, or 'mimetics.' Mimetics considers cultural systems to be little more than aggregations of 'cultural viruses' that human (or nonhuman) minds as their hosts (see *Memes and Cultural Viruses*). The debate has been useful for drawing attention to several interrelated questions, all without clear answers at the time of writing. If cultural histories are replete with 'viral outbreaks,' have culture carriers evolved a diversity of immune (protective) responses? If so what are these immune systems and how do they work? Are they effective enough to prevent people's thoughts and actions from being governed by 'selfish memes?' To the extent that cultural evolution favors selfishness on the part of cultural entities, genetic and cultural evolution are predicted to show divergent tendencies (Durham 1991). How often does this occur, in what contexts, and with what outcomes? The implications are clearly substantial for the social sciences: there is a pressing need for more good work in this area.

The second big area of debate concerns alternative models and processes for understanding the 'editing activity' of cultural microevolution. The challenge that has emerged from this debate is to identify that process or processes that are akin to Darwin's 'main motor' of biological evolution, natural selection. Looking across the vast expanses in space and time of cultural conveyance in human populations (or nonhuman), what process or processes have served as 'editor in chief?' Also, does the role of 'editor in chief' vary in predictable ways by circumstance, as defined by ecological, sociopolitical, and economic conditions? A wide range of different approaches have

appeared in recent years to answer these questions; useful views and reviews include Cronk (1999), Durham et al. (1997), see also *Cultural Evolution: Theory and Models; Genes and Culture, Coevolution of*.

One good way to summarize some of today's leading models, but by no means the only good way, is according to the models of other processes that inspired them. One approach, for example, called an 'epidemiology of representations' (Sperber 1996), draws its inspiration from models and theories governing the spread of contagious disease in host populations. It regards cultural systems as 'contagions of ideas' that spread through human population under the guiding influence of ecological and psychological conditions. In this epidemiological model:

> All the information that humans introduce into their common environment can be seen as competing for private and public space and time—that is, for attention, internal memory, transmission, and external storage. Many factors affect the chances of some information being successful and reaching a wide and lasting level of distribution, of being stabilized in a culture .... The most general psychological factor affecting the distribution of information is its compatibility and fit with human cognitive organization. (Sperber 1996, p. 140)

The approach has much to recommend it: certainly many cultural phenomena spread 'horizontally' (between peers) within societies, like fads and fashions; certainly the success of contagious cultural elements depends to some extent on their psychological appeal or attraction; and surely cognitive processes in individual minds actively transform the 'contagion' as it spreads. The challenge to this approach remains to illustrate by theory and example the relative force of psychological attraction compared to other forces, and to produce convincing empirical examples across a spectrum of cultural phenomena (for one example, see Boyer 1994).

A second approach uses the term 'coevolution' to develop a model of cultural evolution that is explicitly parallel to, though different from, Darwinian theory in biology (Durham 1991). The model views human populations as having two 'inheritance systems,' one genetic and one cultural. Each form of inheritance is conceptualized as having its own channel of information transmission, one consisting of genes passed along via reproduction and the other consisting of ideational phenomena (ideas, values, beliefs) passed along via social transmission. Each form of inheritance is subject to a variety of processes producing change over space and time. In the genetic channel, these processes include the familiar Darwinian foursome: mutation, migration, drift (chance fluctuations), and differential reproduction or 'natural selection.' In the cultural channel, coevolutionary theory proposes the roughly analogous processes of innovation, diffusion (with or without migration), cultural drift, natural selection (cultural change via differential reproduction), and cultural selection (change through value-driven decision making by culture carriers). Durham's formulation (1991, p. 198) emphasizes the role of human beings as decision makers:

> Provided they are given both a range of options and an opportunity to choose, people can participate in the transmission process as selecting agents. In other circumstances, where options and choice are largely pre-empted [by power relations], cultural evolution may still be guided by human decisions—that is, by the imposed decisions of powerful individuals or groups. But either way, culture changes under 'human' direction; people, not 'nature' do most of the selecting.

The strengths of this approach are several: it builds social structure and asymmetrical power relations into the evolutionary dynamics of cultural change; it emphasizes key similarities and differences with biological evolutionary theory without overly 'Darwinizing' cultural dynamics; and it provides a ready explanation for the culturally-guided directionality so characteristic of cultural evolution. In Durham's view, the most important values governing the decision-making processes are previously evolved 'cultural' values. Further the framework has particular value in analytical applications (see Durham 1992). While not as quantitative as other models, it has been useful in more and more thorough empirical analyses than the other formulations.

The third approach was also the first historically, and the most quantitative (see also *Cultural Evolution: Theory and Models*). Taking its main inspiration from Mendel, mixed with influence from Darwin and epidemiology, the approach focuses on the cultural transmission process and the ways in which the context of transmission (who teaches, who learns) produces evolutionary change in cultural systems (Cavalli-Sforza and Feldman 1981). Quantitative models have been proposed for three different contexts of social transmission: vertical (meaning from parents to offspring), oblique (from nonparents of one generation to offspring in the next), and horizontal (among peers of the same generation). The models predict, among a variety of other things, that vertical transmission is generally most conservative in its evolutionary effects (i.e., causing least change from generation to generation) while horizontal transmission promotes relatively rapid evolutionary change—a finding nicely corroborated in a later empirical studies (see, e.g., Hewlett and Cavalli-Sforza 1986). These baseline models have been modified and extended to include conformity effects, continuous traits, and group-level selection (see discussion in *Cultural Evolution: Theory and Models*). Similar transmission models developed by Boyd and Richerson (1985), also drawing inspiration from Mendel, emphasize various forms of 'bias' that can result from variability in the context of social transmission. In 'indirectly biased transmission,' for example, individuals are viewed as adopting for

themselves the cultural variants that are used or exhibited by 'successful' individuals whom they emulate. Boyd and Richerson show how this biasing mechanism can lead to the unstable 'runaway' cultural evolution of highly exaggerated features (such as growing huge yams or participating in elaborate tattooing, in the examples they offer). In 'frequency dependent biased transmission,' individuals adopt for themselves the most common or most popular cultural variant in a population, a process that readily leads to conformist effects. An interesting result of conformist transmission is that it may increase the magnitude of intergroup differences, potentially contributing to differential rates of group-level survival and propagation (i.e., group selection, see Boyd and Richerson 1985, Chap. 7).

In a further extension of so-called transmission models, Feldman and Cavalli-Sforza, and Aoki after them, proposed a combination model that allows genetic and cultural transmission to produce joint 'genophenotypes' which can then be used to model coevolutionary dynamics (see *Genes and Culture, Coevolution of*). Recently, Laland et al. (2000), building on Durham (1991), have added self-induced habitat change, or 'niche construction,' to the interactive processes of biological and cultural evolution. The schema formally recognizes the way that cultural processes produce long-lasting habitat modifications, thus providing a kind of 'environmental inheritance' that permits human beings to 'co-direct our own evolution.'

In conclusion, one may fairly state that theories of cultural evolution themselves continue to undergo rapid evolutionary change. Their potential value to social science is substantial, particularly if they eventually help to bridge the intellectual gap between biological and social sciences. Further theoretical developments will surely contribute to the bridge, but good empirical analyses and longitudinal case studies will probably be even more persuasive. At the time of writing, theories of cultural evolution remain surprisingly peripheral to mainstream social science. Hopefully that, too, will change.

See also: Adaptation, Fitness, and Evolution; Cultural Evolution: Theory and Models; Cultural History; Culture as Explanation: Cultural Concerns; Culture in Nonhuman Organisms; Evolution and Language: Overview; Evolution, History of; Evolutionary Social Psychology; Genes and Culture, Coevolution of; Memes and Cultural Viruses

Bibliography

Bonner J T 1980 *The Evolution of Culture in Animals*. Princeton University Press, Princeton, NJ

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Boyer P 1994 *The Naturalness of Religious Ideas: A Cognitive Theory of Religion*. University of California Press, Berkeley, CA

Cavalli-Sforza L L, Feldman M W 1981 *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press, Princeton, NJ

Cronk L 1999 *That Complex Whole: Culture and the Evolution of Human Behavior*. Westview Press, Boulder, CO

Durham W H 1990 Advances in evolutionary culture theory. *Annual Review of Anthropology* **19**: 187–210

Durham W H 1991 *Coevolution: Genes, Culture, and Human Diversity*. Stanford University Press, Stanford, CA

Durham W H 1992 Applications of evolutionary culture theory. *Annual Review of Anthropology* **21**: 331–55

Durham W H, Boyd R, Richerson P R 1997 Models and forces of cultural evolution. In: Weingart P et al. (eds.) 1997 *Human By Nature: Between Biology and the Social Sciences*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 327–53

Hewlett B S, Cavalli-Sforza L L 1986 Cultural transmission among Aka Pygmies. *American Anthropologist* **88**: 15–32

Heyes C M, Galef B G Jr 1996 *Social Learning in Animals: The Roots of Culture*. Academic Press, San Diego, CA

Johnson A W, Earle T 1987 *The Evolution of Human Societies: From Foraging Group to Agrarian State*. Stanford University Press, Stanford, CA

Laland K N, Odling-Smee J, Feldman M W 2000 Niche construction, biological evolution, and cultural change. *Behavioral and Brain Sciences* **23**: 131–75

Sahlins M D, Service E R (eds.) 1960 *Evolution and Culture*. University of Michigan Press, Ann Arbor, MI

Sanderson S K 1990 *Social Evolutionism: A Critical History*. Basil Blackwell, Cambridge, MA

Sperber D 1996 *Explaining Culture: A Naturalistic Approach*. Blackwell Publishers, Oxford, UK

Stocking G W Jr 1987 *Victorian Anthropology*. The Free Press, New York

Wrangham R W et al. (eds.) 1994 *Chimpanzee Cultures*. Harvard University Press, Cambridge, MA

Weingart P et al. 1997 Units of culture, types of transmission. In: Weingart P et al. (eds.) 1997 *Human By Nature: Between Biology and the Social Sciences*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 299–325

W. H. Durham

Cultural Evolution: Phylogeny versus Reticulation

Why do human cultures and languages form patterns of differentiation through time and across space, and how do such patterns form? Why do language families and archeological cultures exist? Application of biological concepts such as 'phylogeny' (evolutionary history of a genetically-related group of organisms) within the anthropological, archeological, and linguistic disciplines might seem a little unwarranted since culture is transmitted through both Mendelian (phylogenetic) and Lamarckian (reticulate) processes (Boyd and Richerson 1985, Cavalli-Sforza and Feldman

1981, Durham 1991). Thus, the transmission of cultural knowledge and language from generation to generation cannot be precisely equated with genetic inheritance in terms of mechanism. But the terms phylogeny and reticulation can have direct value in understanding the creation of many historical and ethnographic situations.

For instance, since AD 1600 , societies have developed in the UK, North America, and Australasia, which clearly share some degree of common history (phylogeny). This is evident in the almost universal use of the English language, in the direct biological ancestry of many inhabitants of these regions from a British source, and in many aspects of material culture which originated in ways of doing things widely current in Britain between approximately AD 1600 and 1850. These societies in North America and Australasia originated historically via a process of colonization and early political control emanating from the British Isles. Similar generalizations can be made about the spread of Iberian culture, language, and biology through much of Central and South America.

At the beginning of the twenty-first century, it is quite obvious that by no means all of the cultural and biological patterning found in these regions is of a circumscribed British (or Spanish, or Portuguese) origin. Indigenous peoples have contributed, as have migrants from other regions. Cross-cultural borrowings abound and occur continuously, forming a network of horizontal transmission, which in this article is termed reticulation.

Thus we have two conceptual directions of transmission for the elements of both cultural and biological patterning—vertical (phylogenetic or descent-based) and horizontal (reticulative or interaction-based) (Bellwood 1996a, Dewar 1995, Mace and Pagel 1994). Naturally, in a real-life situation such as the modern USA it is impossible to untangle the inherited from the borrowed in an overall sense and to present the results in the form of a statistical statement. To attempt to do so would in any case be futile. These are simply concepts, which allow us to model the formation of cultural patterns through time, essentially to understand the main trends of history.

It might seem obvious that cultures at any given time will reflect variation from both ancestral and borrowed sources. But there is one very significant implication of the concept of cultural phylogeny. In those situations where common ancestry is demonstrable for the member units within widely spread entities, such as language families, major religious complexes, or highly specific style repertoires within archeological complexes, we are obliged to infer dispersal of an ancestral configuration (such as a 'protolanguage') from a homeland region. This is especially the case where the entities concerned are of very broad extent. Explanations which involve the convergence of genetically unrelated units into a single entity such as a language family do not work in these large scale, sometimes transcontinental, situations which extend beyond the ranges of regional communication networks.

The linkages between phylogeny and dispersal can therefore become extremely important in explanations for particular long-term and large-scale patterns in human prehistory. Any evidence for fairly rapid and far-flung dispersal will obviously militate against an explanation for cultural patterning expressed in terms of gradual evolutionary change at a constant and uniform rate. By its very existence, phylogenetic patterning carries significant implications favoring a punctuated trajectory for cultural evolution, focused on periodic episodes of dispersal and change alternating with trajectories of relative equilibrium (Bellwood 1996b; Dixon 1997).

It is important in this regard, however, to differentiate between spreads of people, languages, and other aspects of culture. Ethnographic and historical records offer examples of situations where patterns of biological, linguistic, and sociocultural variation do not correlate at all, or correlate only slightly. On the other hand, they can sometimes correlate with almost 100 percent precision, as in the early stages of population expansion into new territories. Whole populations already *in situ* geographically can adopt new languages and cultural forms through contact-induced change (Ross 1997), whereas at the other extreme whole populations can colonise new territories with closely correlated physical attributes and cultural characteristics.

In terms of human prehistory, it is important to remember the significance of scale. On a small scale, the inhabitants of a village or small river basin could, and indeed have in many historical and ethnographic examples, adopt a new language and undergo genetic change due to intermarriage with members of an outsider, perhaps of higher status, group of people. Ethnic groups can also 'fuse' under situations of depopulation or outsider-imposed stress, as has happened very frequently in the period of tribal shrinkage which history offers us in the form of the 'ethnographic present' (Atkinson 1989, Moore 1994, Nicolaisen 1977–78).

But, on a much larger scale, a chain-reactive process of language and cultural shift of this kind will not explain why, for instance, related languages are spoken over vast regions of the earth in the form of major language families such as Indo-European, Austronesian or Uto-Aztecan (see Fig. 1 in *Archaeology and the History of Languages*). These language families had reached their outer geographical limits before the beginnings of written history, and no historical processes such as military conquest or mass media persuasion can hope to explain them. They are records of major human dispersals, spreading foundation languages and archeological cultures. Many undoubtedly owe their existences to population increase and dis-

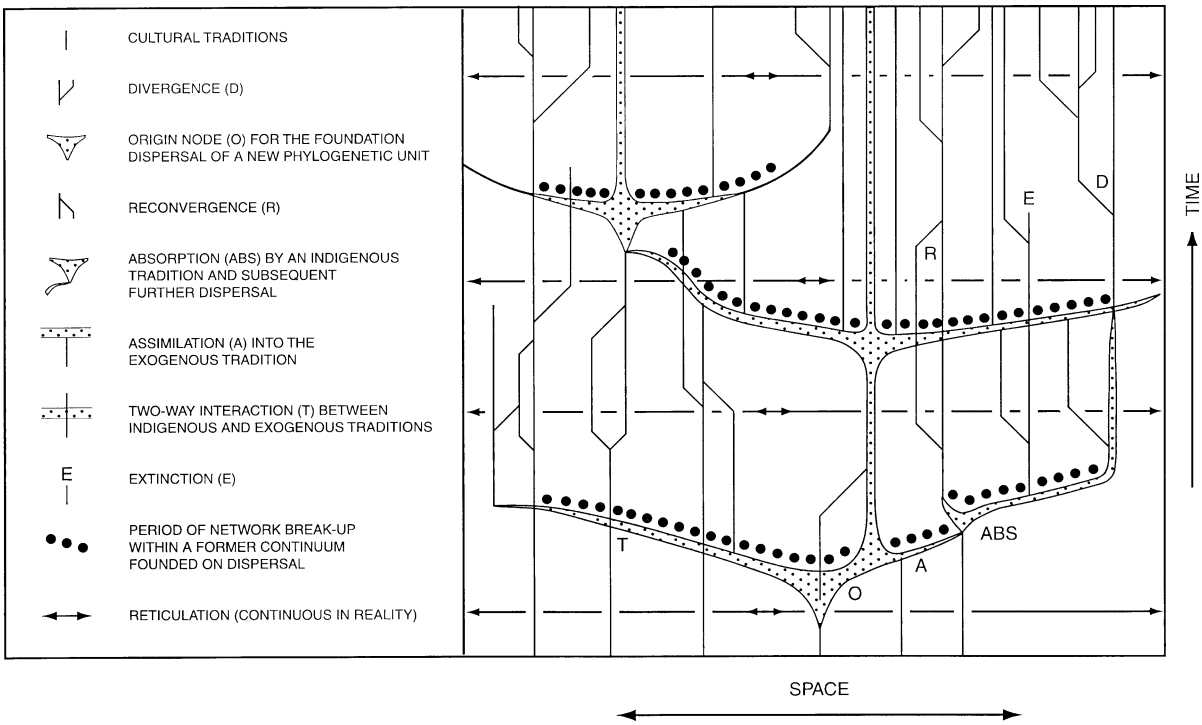

Figure 1
Phylogeny and reticulation in time and space perspective. (NB: A diagram such as this cannot hope to replicate the full complexity of cultural evolution in any real- life situation. It is intended as a schematic model only.)

persal consequent upon the regional beginnings of agropastoral food production (Renfrew 1997). But this is not true for all, since many languages such as English and Spanish have spread over vast distances in recent historical times and can also be regarded as potential foundation languages for future language families. Comparative linguistics and ethnohistory also reveal some excellent examples of large-scale hunter-gatherer dispersal in the form of the Eskimo-Aleut and Athabaskan language families in North America.

The point to be reinforced is that wherever phylogenetic relationships can be demonstrated to exist between the members of an array of languages or cultures (including archeological cultures), then dispersal of an ancestral complex must be inferred. Whether or not this ancestral dispersal directly related to a movement of actual people is something which must be decided on a case-by-case basis. In the case of extremely widespread phylogenetic entities, population movement to some degree can hardly be denied. In addition, most language families clearly spread relatively quickly in overall chronological terms. From internal evidence and archeological correlations, most reached their geographical limits in 4,000 years or less (for example, between 2,000 and 4,000 years from origin regions to outer limits in the cases of Indo-European and Austronesian). Speed of dispersal, in this regard, relates directly to the intensity of punctuation.

The comparative linguistic record can also provide clear structural evidence for rapid dispersal, especially when the component language subgroups within a family do not share any unique linguistic innovations which allow them to be subgrouped phylogenetically above the individual subgroup level (but still below that of the language family as a whole). 'Descent trees' created to show the structure within such language families will be extremely rake-like, rather than tree-like with many successive branches. In such cases, claimed for instance for several of the major Austronesian subgroups of Southeast Asia and the Pacific (Pawley 1999), there is a strong likelihood that the resulting pattern reflects an initial history of very rapid and far-flung dispersal, followed by a formation of independent subgroups, each sharing a common ancestry in the initial dispersal but all reflecting an independent history thereafter.

Direct demonstration of rapid dispersal can also be identified in the archeological record, as for example when a particular aspect of artefact style and decoration too complex to reflect functional selection alone

is distributed very widely in a short time. Examples include the Olmec stylistic complex which spread through Mesoamerica after 1250 BC (Sharer and Grove 1989), the Early Neolithic Linearbandkeramik cultural complex which spread through central Europe about 5000 BC (Bogucki 1996), and the Lapita cultural complex with its distinctive and elaborately decorated pottery which spread through the western Pacific about 1000 BC (Kirch 1997). In contextual terms it is apparent that these dispersals of material culture correlated with dispersals of people and perhaps language, even though the nature of the archeological record does not allow this to be stated as more than a logical possibility rather than a demonstrated fact. A parallel with the linguistic diversification of language families from a common source also comes in the archeological record when very widespread complexes such as those mentioned above are followed by a breakdown of widespread homogeneity into regional diversity, this diversity becoming more marked over time.

Finally, reasons for dispersal in any given historical situation are probably knowable, even if they are often heatedly debatable. Clearly it is important to distinguish between populations who dispersed into territories devoid of prior human inhabitants and those who moved into already-settled areas. In both situations the inception of a successful trajectory of dispersion/colonization would obviously have required the acquisition of some form of advantage allowing demographic growth and range expansion. In the case of many of the examples discussed above the advantage seems to have been the acquisition of systematic agriculture, but the whole story is of course far more complex. The initial human colonizations of empty continents like Australia (60,000 years ago) and the Americas (15,000 years ago) must also have led to the dispersal of phylogenetically-related languages and cultures, of which some faint traces might still survive in present cultural patterns.

1. Some Examples of Phylogenetic Studies of Archeological and Ethnographic Cultures

The ethnographic and archeological records are replete with examples of cross-cultural contact via trade, warfare, and intermarriage, indeed by 'diffusion' and gene flow in their many forms. Past humans have been just as interested in communicating with their neighbors, whether as friends or foes, as have recent ones. But evidence that ethnographic and prehistoric cultures form phylogenetic arrays indicative of dispersal is not always so obvious once one moves beyond the evidence from comparative linguistics (Mace and Pagel 1994). Normally, teasing out of the nonlinguistic aspects of the pattern requires attention to details from archaeology and biological anthropology, assisted of course by information from comparative ethnology and, where relevant, historical records and oral tradition (e.g., Flannery and Marcus 1983, Romney 1957).

As a case study, a debate has emerged in recent years amongst anthropological scholars working in the islands of Polynesia. Polynesians form a population of many island societies in the great triangle bounded by Hawaii, Easter Island, and New Zealand, including Tonga and Samoa, with a few additional 'Polynesian Outliers' existing to the west amongst the larger islands of Melanesia. That Polynesian languages share a common ancestry is not in doubt, but beyond language the explanations for the Polynesian cultural pattern are more contested. Two recent perspectives on the Polynesians—the 'phylogenetic' and the 'reticulative'—can be counterposed.

The archaeologists Kirch and Green (1987, 2001) have presented Polynesia as a 'phylogenetic unit' of closely related cultures, languages, and peoples all sharing a common origin derived from the settlement of the Tongan, Samoan, and other nearby Western Polynesian islands by Lapita populations at about 1000–800 BC. Working from the archeological and linguistic records, Kirch and Green suggest that the Polynesian societies of ethnographic times formed a closely related set in terms of material culture, social organization, biological phenotype, and most importantly, language. Societies so related can be conceived as something akin to the genus of biologists—a group of 'species' defined as a group through possession of certain uniquely shared features (synapomorphies in biological terminology). From this perspective, Polynesian societies all share a single origin, and since that period of origin have differentiated due to a wide range of historical processes reflecting both internal change and external contacts with other societies both inside and outside Polynesia. Because Polynesian societies are geographically so clearly bounded by sea, and because archaeology and linguistics lead us to believe that only Polynesians ever inhabited Polynesia prior to European contact, the region obviously lends itself ideally to such a phylogenetically oriented historical perspective.

But not all agree entirely with a phylogenetic perspective. The archaeologists Terrell et al. 1997 (see also Terrell and Welsch 1997) take a tangentially opposed and rather more reticulated viewpoint, pointing out that ancient Pacific Island (including Polynesian) societies were never isolated and would always have maintained some degree of contact between different regions. Virtually all archaeologists would agree with this statement in simple form, because of course the archeological, linguistic, and genetic records carry widespread evidence of trade, borrowing, inter-marriage, and so forth between different island populations. But the reticulative perspective is taken further by Terrell and his colleagues to draw the conclusion that interaction will erase any recognizable traces of cultural phylogeny (see also Moore 1994).

For these scholars, the phylogenetic ancestry of cultures is generally unknowable, and thus cannot contribute to an understanding of cultural history. In addition, as they rightly point out:

> We need to be wary about thinking that history and diversity in the Pacific can be reduced to a few grand moments of genesis and immigration. (Terrell et al. 1997, p. 175)

The perspective offered in this entry agrees that a masking of phylogenetic history can occur in cases where actual creolization has taken place, in the form of mixing between two cultural entities of entirely different origin and ancestry. Such intricate mixing on the level of material culture is believed to have occurred, for instance, in some of the Austronesian- and Papuan-speaking societies of lowland New Guinea (Welsch et al. 1992, but see also Roberts et al. 1995 for a contrary view). But when 'like meets like,' as would generally have been the case in Polynesia where all cultures shared a recent common origin, then erasure of all traces of phylogenetic relationship becomes much harder to imagine. Even in the case of lowland New Guinea, the fact that the Austronesian and Papuan languages have not in themselves generally become creolized tells us something about the strength of cultural, especially linguistic, transmission through time, and from generation to generation.

2. *Future Research Prospects*

At the beginning of the twenty-first century, there is widespread interest in the prospects of combining linguistic, archeological, and genetic perspectives on the human past. Elsewhere in the Old World a main supporter of a dialog has been the archeologist Colin Renfrew in his writings on Indo-European languages (Renfrew 1987) and broader macrofamily relationships (Renfrew 1992a, 1992b). Renfrew's position, like that of Bellwood for Austronesian (1996a, 1996b, 1997) and that of geneticist Luca Cavalli-Sforza (Cavalli-Sforza and Cavalli-Sforza 1995) for world populations in general, is that the prehistories of large scale ethnolinguistic populations can be approached from the independent, but overlapping perspectives of genetics, archaeology, and linguistics. The basic model here focuses on a homeland concept, followed by phylogenetic descent to the present through a dispersive sequence involving mainly fissions and divergences, albeit constantly tempered with the results of interpopulation contact. But as Bateman et al. (1990) point out, in discussing the problems of reconciling biological and linguistic phylogenies:

> Such phylogenies are likely to be 'noisy,' reflecting complicating factors such as hybridisation, the less constrained evolution and competitive displacement among cultural traits, and the fact that the contemporaneity of human gamodemes and languages cannot be assumed. (Bateman et al. 1990, p. 13)

An attempt is made finally to encapsulate some of the cultural trajectories and differing opinions discussed above in the form of Fig. 1. This diagram highlights both divergent and convergent evolutionary situations, which should be regarded as overlapping possibilities rather than as absolutely separate processes. Divergent situations include the following: (a) Cultural divergence following the break-up of a communication network; (b) Large-scale operation of the above as a result of the foundation by dispersal of a new phylogenetic unit (e.g., the dispersal of the foundation dialect chain for a major language family). Situations of convergence or amalgamation include the following: (c) Periodic cultural reconvergence of traditions already related phylogenetically; (d) Absorption of some elements of a cultural dispersal (e.g., language, food production system) by indigenous populations, usually in peripheral locations with respect to the source of the cultural dispersal (i.e., the indigenous dominates the exogenous). Such absorption can then be followed by renewed dispersal; (e) Assimilation with no continuing separate cultural identity of the assimilated tradition (i.e. the exogenous dominates the indigenous); (f) Two-way interaction (the middle road between absorption and assimilation). Also to be taken into account is; (g) Cultural extinction, normally as a result of environmental problems, naturally or humanly stimulated, sometimes also reflecting strong assimilation (concepts such as pidgin and creole formation in linguistics are not considered here since they are held to belong mainly to recent episodes of colonial population translocation).

Four other processes are shown schematically in Fig. 1: (a) Generation to generation transmission of cultural traditions through time; (b) Schematic horizontal representation of interaction (reticulation). Reticulation is continuous and universal; (c) The relative potential for population growth (rapid during population dispersal); (d) The period during which the breaking-up of a cultural network, resulting from the foundation dispersal of a future phylogenetic unit, will give rise to separate cultural traditions.

The aim of the diagram is, in part, to indicate that human prehistory cannot be purely regarded as a process of local evolution through all time, but that major episodes of dispersal of the foundation cultures for future phylogenetic arrays punctuate the mesh of gradual reticulate ethnogenesis. Without such dispersals there could be no major language families, and none of the rapid and large-scale spreads of early agricultural economies and cultural complexes which we read from the archeological record in many parts of the world.

See also: Archaeology and the History of Languages; Creolization: Sociocultural Aspects; Cultural Evolu-

tion: Overview; Evolution and Language: Overview; Genes and Languages; Languages: Genetic Classification

Bibliography

Atkinson R R 1989 The evolution of ethnicity among the Acholi of Uganda: The precolonial phase. *Ethnohistory* **36**: 19–43

Bateman R, Goddard I, O'Grady R, Funk V, Mooi R, Kress W, Cannell P 1990 Speaking of forked tongues: The feasibility of reconciling human phylogeny and the history of language. *Current Anthropology* **31**: 1–24

Bellwood P 1996a Phylogeny and reticulation in prehistory. *Antiquity* **70**: 881–90

Bellwood P 1996b The origins and spread of agriculture in the Indo-Pacific region. In: Harris D (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London, pp. 465–98

Bellwood P 1997 *Prehistory of the Indo-Malaysian Archipelago*, 2nd edn. University of Hawaii Press, Honolulu, HI

Bogucki P 1996 The spread of early farming in Europe. *American Scientist* **84**: 242–53

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Cavalli-Sforza L L, Cavalli-Sforza F 1995 *The Great Human Diasporas*. Addison-Wesley, Reading, MA

Cavalli-Sforza L L, Feldman M 1981 *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press, Princeton, CT

Dewar R 1995 Of nets and trees: Untangling the reticulate and dendritic in Madagascar's prehistory. *World Archaeology* **26**: 301–18

Dixon R M W 1997 *The Rise and Fall of Language*. Cambridge University Press, Cambridge, UK

Durham W H 1991 *Coevolution: Genes, Culture, and Human Diversity*. Stanford University Press, Stanford, CT

Flannery K V, Marcus J 1983 *The Cloud People*. Academic Press, New York

Kirch P V 1997 *The Lapita Peoples*. Blackwell, Oxford, UK

Kirch P V, Green R C 1987 History, phylogeny, and evolution in Polynesia. *Current Anthropology* **28**: 431–56

Kirch P V, Green R C 2001 *Hawaiki: Ancestral Polynesia*. Cambridge University Press

Mace R, Pagel M 1994 The comparative method in anthropology. *Current Anthropology* **35**: 549–64

Moore J H 1994 Putting anthropology back together again: The ethnogenetic critique of cladistic theory. *American Anthropologist* **96**: 925–48

Nicolaisen I 1977–78 The dynamics of ethnic classification: A case study of the Punan Bah of Sarawak. *Folk* **19–20**: 183–200

Pawley A K 1999 Chasing rainbows: Implications of the rapid dispersal of Austronesian languages for subgrouping and reconstruction. In: Zeitoun E, Li P J-K (eds.) *Selected Papers from the Eighth International Conference on Austronesian Linguistics*. Institute of Linguistics, Academia Sinica, Taipei, pp. 95–138

Renfrew C 1987 *Archaeology and Language*. Jonathan Cape, London

Renfrew C 1992a Archaeology, genetics, and linguistic diversity. *Man* (NS) **27**: 445–78

Renfrew C 1992b World languages and human dispersals: A minimalist view. In: Hall J A, Jarvie I C (eds.) *Transition to Modernity*. Cambridge University Press, Cambridge, UK, pp. 11–68

Renfrew C 1997 World linguistic diversity and farming dispersals. In: Blench R, Spriggs M (eds.) *Archaeology and Language I: Theoretical and Methodological Orientations*. Routledge, London, pp. 82–90

Roberts J M, Moore C C, Kimball Romney A 1995 Predicting similarity in material culture among New Guinea villages from propinquity and language: A log-linear approach. *Current Anthropology* **36**: 769–88

Romney A K 1957 The genetic model and Uto-Aztecan time perspective. *Davidson Journal of Anthropology* **3**: 35–41

Ross M 1997 Social networks and kinds of speech-community event. In: Blench R, Spriggs M (eds.) *Archaeology and Language I: Theoretical and Methodological Orientations*. Routledge, London, pp. 209–61

Sharer R, Grove D (eds.) 1989 *Regional Perspectives on the Olmec*. Cambridge University Press, Cambridge, UK

Terrell J, Hunt T, Gosden C 1997 The dimensions of social life in the Pacific. *Current Anthropology* **38**: 155–96

Terrell J, Welsch R 1997 Lapita and the temporal geography of prehistory. *Antiquity* **71**: 548–72

Welsch R L, Terrell J, Nadolski J A 1992 Language and culture on the North Coast of New Guinea. *American Anthropologist* **94**: 568–600

P. Bellwood

Cultural Evolution: Theory and Models

More than 150 definitions of 'culture' have been proposed in the literatures of anthropology, archeology, sociology, and other disciplines. Most of these have been made with a rationale that is neither predictive nor quantitative, but rather descriptive and/or comparative. In the following account, the article will take as culture those aspects of thought, speech, behavior, and artifacts that can be learned and transmitted. This process of transmission may include the action of genes but may also purely cultural and will therefore produce results that are, in general, different from those of models in evolutionary biology. The transmission may involve any or all of imprinting, conditioning, observation, imitation, or direct teaching.

Cavalli-Sforza and Feldman (1973, 1981) developed a quantitative theory for the cultural transmission and evolution of traits within populations. Such traits may be discrete valued or take values in a continuum. The idea here was to seek laws or principles that could explain observed levels of variation or estimated rates of change in a manner analogous to (but, of course, different from) those that have proven useful in evolutionary biology.

Durham (1991) has argued that the use of 'cultural traits' as the units of cultural transmission 'pits nature

against nurture.' But if one simply uses 'traits' as units, then there is no implication as to how the trait is determined. This determination is defined by the model of transmission, which may be purely cultural, purely genetic, or some combination of these (Feldman and Cavalli-Sforza 1976, 1979). In principle, Durham's alternative 'ideational units' are amenable to the same kind of quantitative analysis as that originated by Cavalli-Sforza and Feldman.

Boyd and Richerson (1985, p. 36) exclude 'behavior and the products of behavior' from the definition of culture because 'behavior is contingent upon both patterns of thought and feeling and environmental circumstances.' In following the example of Geertz (1973, pp. 44, 143–6) this exclusion would, if observed religiously, prevent most reasonable measurement and hence quantitative inference about cultural transmission and change. Many of the examples used by Boyd and Richerson, however, do actually involve behaviors as either the transmitted entity or the way in which the hypothesized entity is revealed and transmitted. Formally, at least, the structure of the theory in Boyd and Richerson (1985) is very similar to that of Cavalli-Sforza and Feldman (1973, 1981) and Feldman and Cavalli-Sforza (1976).

Units of cultural transmission have been assigned a variety of names beyond the 'trait' designation originally used by Cavalli-Sforza and Feldman (1973) and Feldman and Cavalli-Sforza (1976). Dawkins (1976) introduced the term 'meme,' which he later defined as a unit of information that is conveyed from one brain to another during cultural transmission. Lumsden and Wilson (1981) use the term 'culturgen' because they require that such traits are processed through 'genetically determined procedures that direct the assembly of the mind.' Each of these terms has disadvantages: 'meme' conveys a sense of imitation or involuntary allocation to memory, while 'culturgen' places too heavy emphasis on the 'gen,' and we know very little about which, or how many genes might be involved in human information processing. If a term is necessary, the recently introduced term 'seme' advocated by Hewlett et al. (2001) is attractive, with its connotations of semantics, signal, and communication.

1. Types of Cultural Transmission and their Consequences

Consider individuals in a large population at time t, where t is on a scale of generations, and generation t represents the parents of generation $t+1$. It is often useful to assume that these generations are non-overlapping with generation t physically replaced by generation $t+1$. Of course, for culturally transmitted entities, this may not preclude a contribution of generation t to generation $t+2$, for example. It is usually assumed that the population is censussed after cultural transmission has taken place. In the simplest models, the sexes are not distinguished, and each instantiation of the cultural trait may be regarded as taking a value from a set C, say.

Table 1
Transmission of a dichotomous trait

Marriage			Offspring probability**	
Mother	Father	Frequency*	0	1
0	0	$x_0^2+mx_0x_1$	$b_{00,0}$	$b_{00,1}$
1	0	$x_0x_1(1-m)$	$b_{10,0}$	$b_{10,1}$
0	1	$x_0x_1(1-m)$	$b_{01,0}$	$b_{01,1}$
1	1	$x_1^2+mx_0x_1$	$b_{11,0}$	$b_{11,1}$

* The subscript t for the parental generation is suppressed. $x_0+x_1=1$.
** $b_{ij,0}+b_{ij,1}=1(i,j\in\{0,1\})$.

1.1 Vertical Transmission

If an offspring's probability of acquiring trait l in C depends on the traits j and k of its parents, we may represent the set of such transmission probabilities or rates as $\Gamma=\{\beta_{jk,l}\}$. Given the frequencies of each representation of the trait in the parental generation at time t, $\{x_{i,t}\}$, Γ determines $\{x_{i,t+1}\}$ and the sequence of frequency vectors $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t \ldots$ fully describes the evolution of the trait. Marriage between a pair of parents may occur at random, or we may assume, as is likely for many culturally transmitted traits, that individuals of the same trait value prefer to marry one another. This is called 'assortative' mating and occurs with probability m, say. An example of a dichotomous trait, that is one for which C contains the elements 0 and 1, is given in Table 1.

Then, if we write x_i' for the frequencies in the offspring, that is, generation $t+1$, we have

$$x_i' = x_0^2B+x_0C+b_{11,0}+mBx_0(1-x_0) \qquad (1)$$

where $B=b_{00}-b_{10}-b_{01}+b_{11}$, $C=b_{10}+b_{01}-2b_{11}$.

Iteration of Eqn. (1) produces the complete evolutionary trajectory for the population frequencies of this trait. Note that $b_{00,0}=1, b_{01,0}=b_{10,0}=\frac{1}{2}, b_{11,0}=0$ would be the expected transmission rule if 0 and 1 were genetic alleles of a sexual haploid biological organism. The closer the bs are to these values, the more the dynamic resembles a purely genetic evolutionary trajectory. But the system shows interesting flexibility; for example, if the probability that an offspring is unlike its parents is high enough, a cyclic trajectory results (Feldman and Cavalli-Sforza 1976).

Within the framework of vertical cultural transmission, the parameters $\{\beta_{ij,k}\}$ may be disaggregated into separate additive contributions from each parent, although other partitions are conceivable. Estimation of these transmission parameters, for example under an assumption of equilibrium, is possible from familial correlations such as father–offspring or sib–sib correla-

tions, while m is usually measured as the correlation between spouses. However, this may be biased by marital convergence.

Empirical studies designed to assess the importance of vertical transmission of everyday preferences among Stanford University students revealed that only religious and political attitudes were strongly vertically transmitted. Spousal correlations among the designated parents of these students were reasonably high, suggesting a moderate value for m (Cavalli-Sforza et al. 1982). Among the Aka pygmies, an African group of hunter-gatherers, however, vertical transmission of many customs was estimated by Hewlett and Cavalli-Sforza (1986) to be strong.

Evolution under long-continued vertical transmission should be conservative relative to that expected under other modes of transmission in the sense that change over generations is restricted by the extent of parent–offspring concordance. In particular, the more closely $\{\beta_{ij,k}\}$ resembles a genetic transmission system, the slower will be the evolution. In fact, an exactly genetic transmission structure, in the absence of other forces, results in no temporal change at all. These points will be revisited in connection with continuous traits.

1.2 Horizontal and Oblique Transmission

Suppose that after vertical transmission the frequency of trait variant i in the population at time $t+1$ is $x^V_{i,t+1}$ and suppose that this particular variant, once acquired, cannot be lost. Then, in analogy with epidemic models of disease transmission, we suppose that individuals of type j (different from i) who did not acquire i by vertical transmission contact those who did and are converted to i at rate f_{ji}, the rate of horizontal transmission. It is usual to assume $0 < f_{ji} < 1$. Then, after this transmission the new frequency of i is $x^H_{i,t+1}$ with

$$x^H_{i,t+1} = x^V_{i,t+1} + \sum_{j \neq i} x_{i,t+1} x_{j,t+1} f_{ji} \tag{2}$$

If the trait has only two variants 0 and 1, for example, then

$$x^H_{0,t+1} = x^V_{0,t+1} + f_{10} x^V_{0,t+1}(1 - x^V_{0,t+1}) \tag{2a}$$

in which form the logistic nature of the dynamic becomes obvious.

Within the population, horizontal transmission will generally result in more rapid evolution, that is faster change in the frequencies of trait variants, than vertical.

Horizontal transmission modeled by Eqn. (2) occurs between members of the same demographic cohort. If the transmitters are in generation t (or earlier) and are not the parents of the acquirers of the trait, then we have oblique transmission. In most cases, oblique transmission will have similar consequences to same-cohort transmission, with a slight delay. However, an important special case of oblique transmission we have termed 'one-to-many' or 'teacher-type' transmission, which, in the strongest case, entails that a single individual at time t transmits to all individuals at time $t+1$. This results in exceptionally rapid evolution equivalent to the kind of cultural inbreeding often associated with modern mass media.

An important case of horizontal transmission is often called 'cultural diffusion.' The term generally applies to the trajectory of new cultural variant, an alternative to one that already exists and is dominant in the population. It can apply to the spread of the new variant within a single population or, if the population consists of several subpopulations between which there is some communication, it can refer to the spread across the collection of subpopulations. A logistic dynamic is often observed for the change in frequency of the new variant as it diffuses (e.g., Rogers 1995). An important question for many anthropologists is whether this diffusion may occur in the absence of migration of the people who carry the trait, and it is this case for which the term cultural diffusion is often reserved. On the other hand, if people move and carry their variant of the trait with them, and this variant diffuses with its carriers, the term 'demic diffusion' is commonly used. Distinction between these is usually difficult to make, but if genetic data are available that suggest a cline or gradient in allele frequencies from a putative source population of the human migrants (Cavalli-Sforza et al. 1993), then it might be reasonable to infer that demic diffusion has occurred.

The empirical basis of the conversion parameters $\{f_{ji}\}$ in Eqn. (2) has been the subject of extensive sociological analysis. Rogers (1995) introduces the notion of awareness-knowledge, which is information that the new variant of the trait exists. Then there is the adoption phase, in which the decision to adopt may depend critically on local history of the variant, on the education level or economic status of earlier adopters vis-à-vis the decision-maker. Clearly, for the dynamics to be interpretable, simplification of the large number of parameters involved in these details is necessary. Hence the choice of $\{f_{ji}\}$.

It is possible to incorporate the idea of reinforcement through repeated transmission in both vertical and horizontal cases simply by iterating recursions Eqn. (1) or Eqn. (2) a number of times within a generation and censussing after this number of repeated operations. The rate of change across the new census points will be faster than that due to a single iteration per generation.

1.3 Group Effects and Conformity

The pressure on individuals within a population to adopt a belief or perform a task in a certain way may be affected by how many others in that population

believe or act in that way. The result is that the transmission parameters, instead of being constant as in Eqns. (1) and (2), become functions of the frequency (or possibly frequencies). That is, for the dichotomous trait with horizontal transmission, f_{10} in Eqn. (2a) might be an increasing function of $x^V_{0,t+1}$ when $x^V_{0,t+1}$ is large enough. This would amount to a pressure to conform to the majority. Boyd and Richerson (1985, Chap. 7) point out that this can accelerate the rate of cultural evolution. If the frequency-dependent transmission operates in different directions in different places, because of chance effects of environmental differences, this kind of conformity may also augment geographical differences between isolated groups. Applications to economic theory in terms of norms and preferences that are regarded as cultural entities have been initiated by Bowles (1998).

1.4 Innovation

In biological evolution, the ultimate source of all variation is mutation, nondirected random changes in the DNA that may or, more often, may not result in functional change in the organism. The equivalent for a cultural trait is the appearance of a new cultural variant, an innovation. The fate of an innovation within a population depends on its rules of transmission, which in turn may depend on a host of structural features of the trait itself. The analogue to the genetic phenomenon epistasis, which measures interaction between variants of different genes, is an example of a structural feature, but others would bear little relationship to the structural features that affect the fate of biological mutations. For example, stylistic variation in tool manufacture or behaviors that are coerced economically or by direct force would fall into this category.

In terms of the transmission law of Table 1, if $b_{00,1} > 0$ or $b_{11,0} > 0$ an innovation has occurred in the offspring of the first or fourth matings, respectively. On a more general scale, an analogue to the infinite alleles models of population genetic theory (Ewens 1979, Chap. 3) is more appropriate to describe the stochastic dynamics of innovations, although this is an area that has had little formal study.

2. Continuous Traits and Linear Gaussian Models

If the trait under study varies continuously, then the transmission scheme is more difficult to represent. It is usual to represent the values of an offspring's trait as X_{t+1} and to specify that it is a weighted average of two randomly chosen parental values augmented by a random term:

$$X_{t+1} = b_m X_{m,t} + b_p X_{p,t} + \varepsilon \qquad (3)$$

where b_m and b_p are maternal and paternal weights, and ε is the error random variable with zero mean and variance σ^2. Recursion (Eqn. (3)) corresponds to vertical transmission, and it is easy to include correlation between the parents (that is, assortative mating) as well as a kind of oblique transmission. The latter may be represented as a group effect g:

$$X_{t+1} = (1-g)[b_m X_{m,t} + b_p X_{p,t}] + g\bar{X}_t + \varepsilon \qquad (4)$$

where $\bar{X}$ is the mean of the trait at generation t.

Several features of the dynamics resulting from models like Eqns. (3) and (4) are worth noting. The existence of a group-effect constrains the variance within the population (Cavalli-Sforza and Feldman 1981). If a number of groups are considered, however, within each of which transmission occurs as in Eqn. (4), then the variance, V_t, of the average across groups relative to the original group mean increases linearly with time: $V_t = kt\sigma^2$, where k depends on g, b_m, and b_f.

Linear models analogous to Eqn. (3) or (4) allow correlations between relatives to be computed rather simply, especially under Gaussian assumptions for the trait distribution. Indeed, the most widely used framework for studying familial aggregation of quantitative characters represents the genetic contribution to the character as a linear factor complementary to a linear environmental factor. With appropriate assumptions on the genetic term, variances and covariances between relatives can be computed that mimic those calculated under the simplest genetic hypotheses (Fisher 1918). Controversy remains, however, about which joint linear model of genetically transmitted, culturally transmitted, and nontransmitted contributions to traits that aggregate within families is appropriate (Feldman and Otto 1997), while few attempts to use nonlinear representations have been made.

3. Natural and Cultural Selection

Most intuition about the effects of natural selection on genotypes comes from an interpretation of fitness as a relative probability of survival from fertilized egg to adult. However, a considerable body of experimental evidence (reviewed by Lewontin 1974, Chap. 5) suggests that differential fertility, sexual selection, and frequency-dependent effects are more important than differential survival abilities, although the former are more difficult to measure experimentally and to analyze theoretically. Despite these difficulties, it is reasonable to view natural selection on biological variation in terms of differential reproduction and survival. More details on the components of selection may be found in Hartl and Clark (1997, Chap. 6).

For cultural variants, natural selection on a trait may be understood in the same Darwinian way. To the

extent that the variants of a trait affect the reproduction or survival of their carriers differentially, the frequencies of the variants will change over time accordingly. By contrast, cultural selection refers to differences in the rates of transmission of the cultural variants. As examples, such differences may be manifest in the rates at which individuals become aware of the alternatives or in the decision processes leading to the adoption of one or other of the alternatives. The results of cultural selection will be changes in the frequencies of the cultural variants in a population over time. One obvious difference from natural selection is that cultural may act within a generation rather like an epidemic (Rogers 1995). The parameters that summarize the process of cultural selection are the vertical, horizontal, or oblique transmission rates previously defined.

One area of theoretical debate concerns the possibility of conflicts between cultural and natural selection acting on the same variant of a trait. Thus, for the cultural trait of fertility, the variant whose carriers adopt fertility control (e.g., via contraception) is at a clear disadvantage in natural selection but has spread, especially in developed countries, by virtue of its advantage in cultural selection. Other examples are discussed by Cavalli-Sforza and Feldman (1981), Boyd and Richerson (1985), and Durham (1991). The debate, which cannot be settled with standard experimentation, centers on whether natural and cultural selection should act in the same direction most of the time. Alternatively, should the process of natural selection during human evolution have generally shaped cultural transmission in such a way that those variants favored by cultural selection are adaptive in the Darwinian sense?

A second contentious topic in debates about selection on a cultural trait concerns the role of group selection in the maintenance of variants which, within a group, are individually disadvantageous. Two kinds of models have been considered; in one there is extinction and recolonization of groups according to a stochastic process. If the probability of extinction of a group decreases with increasing frequency of the individually disadvantageous trait, then competition between the groups may preserve the disadvantageous trait. The usual example used for such analyses is altruism (e.g., Boyd and Richerson 1985, Chap. 7; Uyenoyama 1979). Maintenance of such a trait by extinction-type group selection is much less likely than by graded group selection, in which extinction does not occur but the expected relative contribution of a deme to the whole set of demes increases with the frequency of the altruist (Uyenoyama 1979). In the latter formulation, after reproduction there is redistribution of individuals to subgroups according to a stochastic process whose outcomes depend on the numbers of the disadvantageous types (Cohen and Eshel 1976). In other models, fluctuating environments in each of the subpopulations have been shown to allow group selection to occur under more general mixing conditions (Uyenoyama 1979). As pointed out by a number of authors, there are great similarities between these theories of group selection, and the Shifting Balance Theory of genetic evolution due to Sewall Wright (1931, 1935, 1948). On balance, there is a recent trend towards accepting the efficacy of an appropriately formulated model of graded group selection in maintaining traits that are individually disadvantageous, whether culturally or genetically transmitted. In particular, frequency-dependent cultural transmission, for example of the conformist type, may amplify differences between groups, thereby increasing the strength of group selection relative to that of natural selection within groups (Boyd and Richerson 1985).

4. *Some Metatheoretical Considerations*

Since around 1965, three main theoretical frameworks have formed the basis for the study of behavioral evolution and, in particular, human behavioral evolution. The first is kin selection theory, the second is evolutionary game theory, and the third is gene–culture coevolutionary theory.

Kin selection theory is due to W. D. Hamilton (1964) and is based on the idea that genetic alleles whose effect is disadvantageous on a carrier may be preserved if that carrier acts to increase the survival rate of other carriers of those alleles. These other carriers will most likely be relatives of the first carrier—hence the name 'kin selection.' By the mid 1970s, this theory had been extended beyond its original insect and animal boundaries to include human behaviors. In this process, the position that a wide array of behaviors would generally be genetically determined and adaptive, came to be called sociobiology (Wilson 1975).

During this period, the application of game theory, the second framework, to animal and human behaviors also became widespread. Here, in pairwise interactions for example, each individual was thought to optimize its behavior relative to the alternatives the other could adopt. The notion of optimization became prevalent in discussions of behavioral evolution (see, e.g., Parker and Maynard Smith 1990). Such treatments did not consider evolutionary dynamics as processes of frequency change but as processes of maximization of functions chosen by the investigator, usually to represent a measure of fitness.

Although such traits as altruism may well have a culturally transmitted basis, they have generally been studied in the theoretical literature using genetic models. Gene–culture coevolutionary theory was introduced as an alternative to the two former frameworks by Feldman and Cavalli-Sforza (1976). By allowing each individual a phenogenotype, the investigator is free to incorporate transmission mech-

anisms, selection functions, and dynamics that span the spectrum from purely genetic to purely cultural. Besides the obvious gain in flexibility to describe the evolution of behaviors, this framework also permits a natural statistical analysis of correlations between relatives *à la* Fisher (1918). The advantage here is that nongenetic effects are modeled more realistically than is possible using standard linear regression theory. In the coevolutionary dynamics with vertical cultural transmission, an association measure between the cultural and genetic variants is important and the order of magnitude of this association as a function of the comprehensive transmission rule can be derived. More details on the connections between these theories and the following examples can be found in Feldman and Laland (1996).

5. *Examples of Gene–Culture Evolutionary Theory*

The evolution of lactose absorption was shown by Aoki (1986) to be well understood as a case of gene–culture coevolution. Here the gene in question is an autosomal dominant with lactose absorption the dominant phenotype at this locus. The cultural dichotomy is the presence or absence of a history of dairy farming. The frequency of lactose absorbers in populations with dairy traditions often reaches over 90 percent, while in those populations without such a tradition it is typically below 20 percent. In modeling the joint evolution of the gene and the custom of dairying, both natural selection due to increased nutrition provided by milk and the cultural transmission of milk use, are affected by the genotype at the lactose absorbing locus. Analysis of this coevolution suggests that the fate of the absorption allele critically depends on the rate of vertical transmission of milk use. If this rate is very high, then a significant fitness advantage to absorbers can shift the frequency of the dominant allele to high frequencies in as few as 300 generations.

The spread of agriculture into geographical areas previously occupied by hunter-gatherers has been a fruitful subject for students of genetic and cultural history. Aoki and Shida (1996) used a reaction-diffusion process to model the spread of farming. Their analysis included a horizontal transmission process converting hunter-gatherers to farming as well as different intrinsic growth rates for initial farmers, converted farmers, and hunter-gatherers. Their analysis produced conditions under which farming would advance at a constant rate and, in addition, a gene frequency cline would simultaneously form. This elegant study represents a major advance over the earlier proposal of Ammerman and Cavalli-Sforza (1971), which was based on R. A. Fisher's famous model for the advance of an advantageous gene (Fisher 1937).

A third example of gene–culture coevolution concerns the well-documented cultural bias against daughters in South and East Asia (Das Gupta 1987, Tuljpurkar et al. 1995). Under the assumption that a preference for sons is vertically transmitted, Kumm et al. (1994) studied how this transmission might affect the evolution of sex-determining genes. They showed that if parents act to increase the proportion of sons, irrespective of the biologically set sex ratio among their children, and if they compensate for lost daughters by having further children, the sex ratio will evolve to be female-biased. On the other hand, if parents try to achieve a desired sex ratio among their children and in so doing have fewer children than impartial parents, the primary sex ratio will become male-biased.

See also: Cultural Evolution: Overview; Culture as Explanation: Cultural Concerns; Culture in Development; Culture in Nonhuman Organisms; Culture, Sociology of; Genes and Culture, Coevolution of; Memes and Cultural Viruses

Bibliography

Ammerman A J, Cavalli-Sforza L L 1971 Measuring the rate of spread of early farming in Europe. *Man* **6**: 674–88

Aoki K 1986 A stochastic model of gene–culture coevolution suggested by the 'culture historical hypothesis' for the evolution of adult lactose absorption in humans. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 2929–33

Aoki K, Shida M 1996 Travelling wave solutions for the spread of farmers into a region occupied by hunter-gatherers. *Theoretical Population Biology* **50**: 1–17

Bowles S 1998 Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature* **36**: 75–111

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Cavalli-Sforza L L, Feldman M W 1973 Cultural versus biological inheritance: Phenotypic transmission from parents to children (A theory of the effect of parental phenotypes on children's phenotypes). *The American Journal of Human Genetics* **25**: 618–37

Cavalli-Sforza L L, Feldman M W 1981 *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press, Princeton, NJ

Cavalli-Sforza L L, Feldman M W, Chen K H, Dornbusch S M 1982 Theory and observation in cultural transmission. *Science* **218**: 19–27

Cavalli-Sforza L L, Menozzi P, Piazza A 1993 Demic expansions and human evolution. *Science* **259**: 639–46

Cohen D, Eshel I 1976 On the founder effect and the evolution of altruistic traits. *Theoretical Population Biology* **10**: 27–302

Das Gupta M 1987 Selective discrimination against female children in rural Punjab, India. *Population and Development Review* **13**(1): 77–100

Dawkins R 1976 *The Selfish Gene*. Oxford University Press, New York

Durham W H 1991 *Coevolution: Genes, Culture and Human Diversity*. Stanford University Press, CA

Ewens W J 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin
Feldman M W, Cavalli-Sforza L L 1976 Cultural and biological evolutionary processes, selection for a trait under complex transmission. *Theoretical Population Biology* **9**: 239–59
Feldman M W, Cavalli-Sforza L L 1979 Aspects of variance and covariance analysis with cultural inheritance. *Theoretical Population Biology* **15**: 276–307
Feldman M W, Laland K N 1996 Gene–culture coevolutionary theory. *Trends in Ecology and Evolution* **11**: 453–7
Feldman M W, Otto S P 1997 Heritability from twins and other relatives. *Science* **278**: 1383–4
Fisher R A 1918 The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**: 399–433
Fisher R A 1937 The wave of advance of advantageous genes. *Annals of Eugenics*. **7**: 355–69
Geertz C 1973 *The Interpretation of Culture*. Basic Books, New York
Hamilton W D 1964 The genetical evolution of social behavior: Parts I and II. *Journal of Theoretical Biology* **7**: 1–16, 17–52
Hartl D L, Clark A G 1997 *Principles of Population Genetics*, 3rd edn. Sinauer, Sunderland, MA
Hewlett B S, Cavalli-Sforza L L 1986 Cultural transmission among the Aka pygmies. *American Anthropologist* **88**: 922–34
Hewlett B S, de Silverti A, Guglielmino C R 2001 Semes and genes in Africa. To appear
Kumm J, Laland K N, Feldman M W 1994 Culturally transmitted prejudices for offspring gender and the evolution of primary sex ratios. *Theoretical Population Biology* **46**: 249–78
Lewontin R C 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York
Lumsden C J, Wilson E O 1981 *Genes, Mind and Culture*. Harvard University Press, Cambridge, MA
Parker G A, Maynard Smith J 1990 Optimality theory in evolutionary biology. *Nature* **348**: 27–33
Rogers E M 1995 *Diffusion of Innovations*, 4th edn. The Free Press, New York
Tuljpurkar S, Nan L, Feldman M W 1995 High sex ratios in China's future. *Science* **267**: 874–6
Uyenoyama M K 1979 Evolution of altruism under group selection in large and small populations in fluctuating environments. *Theoretical Population Biology* **15**: 58–85
Wilson E O 1975 *Sociobiology*. Harvard University Press, Cambridge, MA
Wright S 1931 Evolution in Mendelian populations. *Genetics* **16**: 97–159
Wright S 1935 Evolution in populations in approximate equilibrium. *Journal of Genetics* **30**: 257–66
Wright S 1948 On the roles of directed and random change in gene frequency in the genetics of populations. *Evolution* **2**: 279–94

M. W. Feldman

Cultural Expression and Action

Attention to action in the sociological study of culture (and in the social sciences more generally) typically has meant taking the individual as a theoretical starting point. Max Weber (1968, p. 3) defined 'social action' as follows: 'We shall speak of "action" insofar as the acting individual attaches a subjective meaning to his behavior—be it overt or covert, omission or acquiescence. Action is "social" insofar as its subjective meaning takes account of the behavior of others and is thereby oriented in its course.' Cultural and structural accounts of action are often contrasted, but action theorists have been interested in precisely how structural constraints and cultural meanings interact. A focus on culture and action also points toward questions of how culture affects action, rather than toward the interpretation of culture for its own sake.

The focus on action raises a set of important questions around which this article is organized. (a) What kinds of meanings or cultural elements shape action? (b) How are meanings brought to bear on action? (c) What is the relation between consciously articulated symbols and implicit, unarticulated practices in guiding action? (d) What is the best way to conceptualize the relation between culture and social structure? (e) Are there cultural 'logics' and if so, how do they operate? (f) How is culture organized differently, such that it may have differing effects on action, at different levels of social organization?

1. *Kinds of Meanings*

What *kinds of meanings* or cultural elements shape action? There are several candidates for the key component of culture that influences action, each with different implications for the shape that influence will take. Max Weber attempted to understand the subjective meanings that governed action, focusing on the effects of particular, historically specific *ideas*. Thus, he argued, the beliefs of Calvinist Protestants, or Hindu mystics, or Confucian officials had consequences for the ways they acted. Ideas are not the only influence on action, but they provide its premises: 'not ideas, but material and ideal interests directly govern men's conduct. Yet the "world images" that have been created by "ideas" have, like switchmen, determined the tracks along which action has been pushed by the dynamic of interest' (Weber 1946 [1922–23], p. 280).

The difficulty with Weber's formulation, however, comes in specifying how ideas actually guide action. Weber had emphasized that ideas such as religious doctrines channeled actors' motives by defining the ends or goals of action and shaping beliefs about relevant means. Thus Western religious traditions, positing a powerful personal God who is creator and judge, conceive of individual salvation as something to be earned by active conduct in this world. Many Eastern religious traditions conceive divinity as immanent in reality, and their adherents seek to merge with the divine through contemplative or mystical practices (Weber 1993). In Weber's most famous example, *The Protestant Ethic and the Spirit of Capitalism* (Weber 1958), Protestant anxiety about salvation, channeled through the Calvinist doctrine of predestination, produced powerful motivations for

disciplined work in a calling. Later theorists, however, noted that belief systems could and did support diverse lines of action, that people in fact hold a variety of often inconsistent or poorly-integrated beliefs, and that even civilizations Weber thought to be otherworldly and mystical develop disciplined, rational economic activity.

Ideas are rarely coherent or consensual enough to account for the paths actors take (see Wuthnow 1987, Skocpol 1985), although Swidler (1986) has argued that such consistent, intense, effective ideologies are more common in 'unsettled' historical periods. Indeed, even John Calvin, who for Weber epitomized rigorous submission to the logic of ideas, adapted his doctrines continuously to his vision of how the Genevan community should be ruled (see Walzer 1973). Furthermore, the Protestant ethic of anxious, disciplined self-scrutiny long outlasted Calvinist doctrine, casting doubt on the causal centrality of Calvinist ideas in themselves.

Talcott Parsons, reinterpreting Weber's work along with that of Emile Durkheim, focused not on particular, historically contingent ideas, but on broad, general *values* which he felt characterized whole societies. Such values as individual achievement or equality provided the core goals that motivated both individuals and societies. Parsons (1937, 1951) developed a voluntarist theory of action which described actors making choices in a situation. While he understood that there were many noncultural constraints on action, ranging from the biological nature of the human organism to characteristics of individual psychology, and to constraints and opportunities provided by power or wealth, he saw the distinctive contribution of sociological analysis as study of the cultural values and norms that shaped what people sought and the means they thought legitimate.

Parsons' formulation raised numerous questions among them are: to what degree cultural values are shared; what values are and where they are located (how they can be observed); how they are internalized in individuals; and how they constrain individual thought and action. Some critics argued that 'values' had no independent standing as causal elements in human action, since the only evidence for the existence of values that purportedly guided action was the action itself (Cancian 1975). Others objected to an overly deterministic theory which, despite the metatheoretical commitment to 'voluntarism' saw individuals as constrained by socially imposed norms and values such that no real choice was possible (Wrong 1961).

Other theorists question the image of individuals as choosing their actions one at a time in order to attain their ultimate ends, in favor of broader but looser conceptions of how culture shapes human action. Geertz (1973) emphasizes the *ethos* of a people, especially the mood or tone of their collective life, the emotional attitudes engendered by their ritual practices, and the experience of the world encouraged by their artistic styles, their everyday practices, and their sacred symbols. He still sees a role for 'world view,' but he does not seek to draw direct links between cultural beliefs and action, nor to trace social choices to shared, overarching values. Swidler (1986) argues that not values but styles, skills, and habits (culturally cultivated capacities) are the most direct cultural influences on action. She argues that individual actors do not usually start with culturally defined goals or values (as Weber and Parsons had assumed) and then build lines of action around them. Rather, actors are constrained by their culturally learned styles, skills, and habits, and they tend to adopt goals or values which are compatible with those cultured capacities. By locating cultural influence in skills and habits rather than in ultimate values or ideas, Swidler allows for the possibility that cultures contain many diverse elements, some of which may be contradictory, so that culture should be thought of as a 'tool kit' upon which actors draw, rather than a coherent set of ends or values.

New ways of thinking about culture have become important since the 1980s as concepts of *discourse* and *practice* have replaced the older interest in ideas and values. The term discourse, although it sometimes designates symbolic communication as opposed to embodied practices, usually means a transpersonal ordering of symbols and meanings which sets the terms within which particular things may be said. In this usage the term points away from action theory and actors altogether. A discursive system does not exist in the mind of any particular actor and does not depend on any one user (see Wuthnow 1987). Nonetheless it presumably constrains what actors can think, say, and do, although the precise nature of these constraints is not made clear. The emphasis on practice (Bourdieu) or practices (Foucault), as Ortner (1984) suggests, envisions culture as both deeper (embodied in unconscious habits rather than in consciously held ideas and values) and more fraught with power and conflict than earlier approaches had recognized. Practice theorists have stressed the active, strategic agency of social actors (their 'resistance') while also emphasizing their subjection to dominant discourses and practices.

A more recent challenge to action theory comes from a revitalized social constructionist perspective (see Berger and Luckmann 1967) which reintroduces culture as a general set of *cognitive rules and recipes* in terms of which agents, institutions, and structures are constituted. John Meyer and his collaborators (Meyer 1983, 1986, 1987, Meyer et al. 1997) locate culture in the 'environment' of cognitive rules and recipes that define the world, rather than in the consciousness of individual actors. They argue that the major effects of culture are to define institutional structures (such as nation states, organizations, or such institutions as schooling) rather than to guide individual action

directly. (On the contrasts between Parsonian understandings and Meyer's 'new institutionalism' see DiMaggio and Powell 1991.) Thus, in the examples favored by Meyer and his collaborators, nation states are constituted according to global rules imbedded in the international system; organizations are created according to schemas constituted within nation states; and individuals themselves are constituted according to rules imbedded in a global culture of 'human rights,' citizenship, and the rationalized Western account of the sources and purposes of economic growth and progress. Rather than offering an alternative explanation of how culture affects action, such neo-institutionalist approaches dismiss the explanation of individual and collective action as uninteresting. Neo-institutionalist theory has offered a broad description of changes in the contemporary world system, while remaining vague about the mechanisms by which institutional rules and recipes emerge, change, or decline, or the concrete processes involved in their adoption, alteration, or rejection by particular groups or actors.

2. *How Culture Is Internalized*

The second major question for any theory of action is *how cultural elements come to govern action*. Do they become internalized, and if so how? Weber assumed that cultural ideas operate by channeling human motives (for health, wealth, salvation from suffering) along lines suggested by a world view. Weber assumed that world views were offered by religious and ideological entrepreneurs, but that once the ideas had become established they constrained people by shaping what they thought was real, what they thought was possible, and what they thought desirable. Parsons (along with collaborators such as Freed Bales) developed a complex Freudian theory of family dynamics to explain how values are internalized in individuals through socialization. The most innovative contribution along these lines has been that of Pierre Bourdieu (1977, 1984). Bourdieu sees human actors as active strategists, not following cultural rules, but manipulating rules to their own advantage. Nonetheless, the skills needed to manipulate a cultural code, according to Bourdieu, require a deep process of internalization during which people in essence internalize their own position in the social structure. Thus the conditions for agency—a fluid mastery of a system of social meanings—are the conditions for being dominated by its criteria of evaluation. Bourdieu sees culture not as values but as cultural skills and capacities, particularly the skills that allow innovative performances within established cultural patterns. Bourdieu incorporates a notion of social constraint into the heart of his model of the actor as active, strategic, and innovative, by arguing that culture becomes deeply internalized, not in the mind as values, but in the body (or rather 'bodily schemas') as *habitus*: as a system of 'transposable dispositions' which make 'possible the achievement of infinitely diversified tasks, thanks to analogical transfers of schemes permitting the solution of similarly shaped problems' (1977, p. 83). For Bourdieu the *habitus* marks one both because the whole set of internalized bodily dispositions, tastes, and habits are particular to one's social location, but also because learning to manipulate the social world means internalizing it and its standards.

Randall Collins (1981, 1988) has also sought to understand how relations of power and status inequality are reproduced at the micro level of personal interaction. His model, however, depends less on internalization and more on exchange. Collins argues that every interaction leaves actors either enhanced or diminished emotionally, so that those in powerful or high-status roles dominate interactions and find that each interaction increases their emotional reserves, while those in subordinate positions lose out and experience themselves as less vital or powerful. Bourdieu and Collins thus both show how elements of social structure (particularly power and inequality) are incorporated directly into the person at a visceral or bodily rather than an intellectual level.

Hans Joas' *The Creativity of Action* (1996) offers a rich description of action that does not require the presumption of an autonomous, goal oriented actor who chooses actions in a situation that is somehow external to his or her action. Joas reconstructs pragmatist social theory to show how sociality, physicality, and intentionality are inherent in human action: 'vague dispositions toward goals which are constantly at work even when we have not set ourselves any immediate goals, are located in the personal body of the human being…Situations do not trigger our actions, but nor do they merely provide the terrain on which we carry out our intentions. Our perception of the situation is predefined in our capacities for action and our current dispositions for action' (p. 161).

3. *Discourse and Practice*

What is the *relation between articulated discursive elements and implicit, unarticulated practices* in guiding action? Are the meanings that shape action primarily ideas and symbols, or are they embodied practices—habitual ways of acting not necessarily guided by conscious thought? Michel Foucault (1965, 1978, 1983) has argued that discourse and practice reinforce each other. New institutional practices, such as asylums that separate the sane from the mad, make plausible the associated discourses and their ways of categorizing human beings. Practices, such as that of confession in the church or the segregation of lepers in the Middle Ages, persist to form the basis of related practices (psychoanalysis; the insane asylum) and their associated discourses in later historical eras. At the same time, new kinds of knowledge—what Foucault (1980) calls 'power/knowledge'—such as the academic

disciplines that study, analyze, and categorize human behavior, the biological organism, or social formations—provide the bases for new practices, which in turn make the categories plausible. John Mohr and Vincent Duquenne (1997) have mapped the mutual constitution of categories and practices, showing how categories of the poor are constituted by welfare agencies' practices, and agencies' practices are defined by the categories of the poor they treat. Mohr and Duquenne also demonstrate how this structure of mutually constituted practices and categories changes over time.

Some scholars have tried to disentangle the causal role of discourses and practices. Richard Biernacki (1995) has argued that the differing ways labor was constituted as a commodity in Germany and England (in Germany employers purchased labor time, while in England they purchased the labor embedded in commodities) had profound effects on both thought and action. The cultural constitution of labor shaped the design of factories, the ways foremen behaved, and the ways workers went out on strike, as well as the contrasting intellectual perspectives of Marx vs. the British political economists. Yet, he argues, this profound difference in the nature of labor as a commodity was not a matter of differing ideas or discourses about labor, but of the 'silent practices' which reproduced the differing ways labor was constituted as a commodity, independently of how men and women tried to think about their experience. In particular, the ways factory owners and workers negotiated wages, with German workers and managers contending over compensation for workers' labor time, while British workers and owners fought over the price of cloth of different qualities, reproduced differing understandings of the nature of labor. These patterns emerged in turn out of the practices for mobilizing labor that were available to owners at the start of the industrial era. These different understandings of labor as a commodity ended when, during World War I, governments suspended wage negotiations and thus interrupted the arena of practice in which these differing forms of labor had been reproduced.

Dobbin (1994) also privileges practice over discourse in arguing that nations base their industrial policies on cognitive models of the sources of order and progress. These models in turn are based on understandings of national sources of political order. While neither study resolves the question of the relationship of practice and discourse, they suggest additional ways of thinking about how culture constitutes patterns that in turn constrain social action.

4. *Culture and Structure*

What is the relation between *culture and social structure*? The fourth problem raised by any theory of action is that of the relation between the cultural influences on action and the influence of what are sometimes called 'structural' or 'material' factors. This problem is an old one in the social sciences, from Marx's insistence that ideas were the product not the determinant of material realities to Max Weber's discussion, quoted above, of the ways 'material and ideal interests' govern men's conduct. For many decades, sociological debate revolved around how to weigh 'material' vs. 'ideal' influences on action, or somewhat later, the relationship of 'culture' and 'social structure.' Talcott Parsons developed the most sophisticated solution of this problem in its original Weberian form in his 'cybernetic' model of action (see Alexander 1983 for a fuller discussion). Parsons proposed paired hierarchies of factors that 'condition' vs. 'control' action. Conditioning factors are necessary but not sufficient conditions for action, such as being alive or having the material capacity to act—high in 'energy' but low in 'information.' Controlling factors, like ideas, culture, or the genetic code, are high in information but low in energy. High-energy factors determine the capacity for action, but the high information factors—a person's ideas, an architect's plans, or a chromosome's genetic structure—convey the information necessary actually to guide action.

Recent theories have sought to re-examine or eliminate altogether the distinction between culture and structure in theories of action. The most important effort along these lines is that of William Sewell Jr. (1992), drawing on Bourdieu (1971) and Anthony Giddens (1984). Bourdieu noted that structure itself has to be reproduced, and argued that culture plays a role in that reproduction. Giddens introduced the idea that structures are 'dual,' constituted simultaneously by 'schemas' which are 'virtual' and material resources which are actual. This formulation maintains the distinction between material and ideal, but instead of seeing them as separate factors in action, sees them as something like matter and form in the ancient Greek sense, in which entities do not acquire their properties unless matter is given form through cultural meanings. In this formulation, structures themselves become cultural artifacts whose creation, reproduction, and meanings are constituted by cultural schemas. At the same time, schemas are instantiated only when they structure some set of material resources, if only print on a page or sound waves in air.

Sewell (1992) advances Giddens' discussion in several ways. First, he follows Giddens in noting that schemas are transposable from one context or situation to another, but he makes this one of the crucial foundations for an argument about human agency. Each transposition of a schema to a new context (even the reproduction of a structure in a new time or place) is a creative adaptation in new circumstances. Second, cultural schemas are 'polysemic' or 'multivocal' so that each time a schema is abstracted from its current embodiment and reproduced, there is human agency involved in interpreting or reading the schema-bearing

structure. Indeed, Sewell's most powerful idea may be that structures themselves are 'read' like texts for the schemas they encode, and that such readings may vary. Third, any instantiation of a schema may produce unpredictable material outcomes, and those outcomes themselves may then be read as embodying an altered schema. Fourth, unlike material resources, human resources always involve some element of agency since human beings attribute meaning to their action and retain some measure of control over their capacities and commitments.

The notion that culture 'constitutes' social structure has become something of a truism, but Sewell has begun to explore the sense in which it is true. His conception of structure still covers an extraordinarily wide range of phenomena, from handshakes, to capitalism, to nation states, but he is right to point to the cultural schemas that organize and order each of these structures, and which at least potentially make possible their transformation. If existing structures are read for the cultural schemas they instantiate, this also suggests how material transformations, whatever their source, can constitute a new cultural text.

The most creative use of this framework is Sewell's (1996) paper on the transformation of structures during the French Revolution. By analyzing the 'event' of the fall of the Bastille as it unfolded over a period of weeks, Sewell shows how an intense collective crisis transformed social meanings. This transformation was made possible by the possibilities of existing patterns of symbolic meaning (the fact that the 'people' ['*le peuple*'] could mean both the crowd or rabble and the legitimating foundation of the nation), but the new meanings emerged only in an intense, public ritual where new symbolic meanings could be validated by public enactment. Sewell also shows the role of material resources in establishing a new schema. He argues that the popular violence at the Bastille was retrospectively redefined as legitimate revolutionary fervor only when the National Assembly, in the days following July 14, 1789, recognized that the people had saved the Assembly from the King's National Guard. Thus the revolutionary transformation of structure resulted from the agency of actors caught up in an intense, fraught, consequential public ritual, but the reproduction of those transformed meanings also depended on their material consequences.

5. *Coherence and Cultural 'Logics'*

There is wide disagreement about *how coherent* cultural systems are. Are they 'systems' at all, and if so, in what sense? One challenge for those who argue that cultural systems have their own logic is to specify what a cultural logic is, how it operates, and how it constrains thought and action. Alexander (Alexander and Smith 1993), along with anthropologists such as Schneider (1976) and LeVine (1984), have insisted on cultural coherence, while Swidler (2001), DiMaggio (1997), and others have seen culture as being more disorderly and unsystematic. Sewell (1999) has argued that culture is necessarily a system in the analytic sense of semiotic codes whose meanings are inter-related. Any concrete culture, however, is made up of many different codes which may be inconsistent, partial, or conflicting.

Theorists who insist on culture's coherence rely on the notion of a 'cultural logic' as the link between culture and action. But sociologists who use the idea of a cultural logic have done little to define, or even to examine explicitly, what cultural logics are, and how they constrain action. Weber (1946) argued that world views had greater influence on action as they became more rationalized and thus had stronger, more inescapable logical implications. Bellah (1964) developed powerfully this idea in his analysis of 'Religious Evolution,' arguing that as symbol systems became more differentiated and coherent, and as specialized institutions for generating and transmitting religious meanings emerged, those meanings had greater independent leverage over action. But logical-deductive coherence is not the only, nor the most plausible, candidate for describing how cultural meanings might actually be organized. Levi-Strauss and his followers have described binary oppositions as the fundamental organizing principle of all cultures, while Clifford Geertz has insisted that each culture's logic is distinctive. Such a notion of unique logics might seem to vitiate the notion of a 'logic' altogether, except that Geertz (1973, pp. 133, 399, 405–8) frequently describes a kind of existential logic in which some ways of experiencing the world are compatible with, or even require, others.

Other theorists have proposed narratives as organizing cultural logics (see Turner 1974, Ortner 1989); still others have seen cultural logics as arising from institutional patterns (Friedland and Alford 1991, Swidler 2001). The anthropologist Bradd Shore (1996) has argued that cultures contain multiple, intersecting logics (see also Friedland and Alford); Pierre Bourdieu (1990) has argued that when people seek to solve urgent problems, the logic of practice brings to bear all sorts of contradictory schemas. If there are many overlapping or intersecting cultural logics, do actors just pick and choose among them? Are some more compelling than others? Or do 'external factors' determine when and how cultural logics are constraining? Some notion of a cultural logic or logics seems essential for understanding how culture influences action. But existing approaches leave unresolved problems of how cultural logics operate and how or when they are brought to bear on action.

6. *Levels of Analysis*

For those who accent the multiplex, incoherent, and conflicting aspects of culture, one challenge is to specify the contexts that organize meanings and

sometimes bring them to bear in a consistent way on action. Swidler (2001) points to the incoherence of actual cultures, both the variety of cultural resources available (the opposing adages of conventional wisdom; platitudes vs. the cynical knowledge of insiders; competing voices in the political arena; contradictory messages in the media) and the confused, often contradictory aspirations and understandings of individuals. She argues that to explain when and how cultural meanings constrain action, one needs to specify the contexts or institutions that sometimes bring culture to bear in a consistent way on action, even when it remains incoherent or incomplete in the minds of individuals.

The key to analyzing problems of culture and action is to differentiate levels of analysis—to see that culture may be organized differently at different levels of social organization. On the one hand, individual actors draw on repertoires of cultural meanings to construct lines of action which allow them to navigate the wider social space. At the same time, contexts external to actors (such as public arenas, institutional demands, or even publicized social codes) may structure meanings and thus influence action, independently of whether culture is deeply incorporated into the actor. Swidler (1995) offers the example of polarized public meetings as contexts in which cultural meanings may become organized coherently, so that voicing an opinion one held yesterday may 'mean' something very different today; public debate may reorganize ideas around new 'logics'; and actors may find themselves constrained by the public ordering of ideas they personally do not hold.

Some social meanings may be anchored in established social practices, without those practices themselves being linked to larger cultural logics. Asking what anchors cultural meanings raises the question of why culture is sometimes much more fluid, negotiable, and transformable than at other times. Biernacki (1995) has shown that practices such as those of wage negotiations between owners and workers in England vs. Germany were enduring and were capable of organizing wide domains of thought and action, yet had no necessary logical tie to other social patterns. Sewell (1999) has suggested that the state is in the business of imposing order on cultural meanings, attempting to produce a more unified cultural logic in the public arena.

7. Collective Action Schemas

The basic paradigms that describe logics of collective action (that is, the mobilization of a community outside institutionalized channels) may have a distinctive logic that is not necessarily homologous, and does not necessarily change in rhythm with more institutionalized patterns of culture and action. In essence, the schemas that guide collective action are a kind of social 'default option'—what everyone knows that everyone else knows that everyone knows. Such schemas of collective action may endure through many institutional transformations because they are not created by routine social experience and they always depend on simultaneous, public knowledge to establish new patterns. Dramatic public ritual events, like those Sewell (1996) has analyzed, may redefine paradigms for collective action, while the culture that undergirds ordinary strategies of action may remain largely unaffected (see Swidler 1992).

The Weberian theory of action saw culture as shaping social life by shaping the ideas and thus channeling the motives of individual actors. Later theorists of action have attempted to find a tighter linkage between the individual and the social—sometimes, like Parsons, seeing the individual as a carrier of consensual social values; sometimes, like Sewell and Swidler, emphasizing again the autonomy of individual actors. Practice theory attempted to solve the relation of actors to social structure by seeing social structure itself as deeply, unconsciously internalized. But understanding how culture is organized in actors' minds, bodies, and experiences is not sufficient for a full theory of culture and action.

Recognizing that culture is organized, and organized *differently*, at multiple levels may help to specify the autonomous causal role of culture, taking account of the agency of actors while acknowledging separate processes through which external contexts bring culture to bear on action. Such an approach would take cultural meanings seriously while analyzing the contexts that bring culture to bear on action, the situations that organize or reorganize meanings, the schemas that encode logics of collective action, and the institutional demands that reproduce cultural supports for action.

See also: Collective Behavior, Sociology of; Cultural Expression and Action; Cultural Psychology; Culture and Emotion; Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Culture as Explanation: Cultural Concerns; Culture, Sociology of; Parsons, Talcott (1902–79); Values, Sociology of; Weber, Max (1864–1920)

Bibliography

Alexander J C 1983 *The Modern Reconstruction of Classical Thought: Talcott Parsons*. University of California Press, Berkeley, CA

Alexander J, Smith P 1993 The discourse of American civil society: A new proposal for cultural studies. *Theory and Society* **22**: 151–207

Bellah R N 1964 Religious evolution. *American Sociological Review* **29**: 358–74

Berger P L, Luckman T 1967 *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Doubleday, Garden City, NY

Biernacki R 1995 *The Fabrication of Labor: Germany and Britain, 1640–1914*. University of California Press, Berkeley, CA
Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK
Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste* [trans. Nice R]. Harvard University Press, Cambridge, MA
Bourdieu P 1990 *The Logic of Practice*. Stanford University Press, Stanford, CA
Cancian F M 1975 *What Are Norms? A Study of Beliefs and Action in a Maya Community*. Cambridge University Press, London
Collins R 1981 On the microfoundations of macrosociology. *American Journal of Sociology* **86**: 984–1014
Collins R 1988 The micro contribution to macro sociology. *Sociological Theory* **6**: 242–53
DiMaggio P 1997 Culture and cognition. *Annual Review of Sociology* **23**: 263–87
DiMaggio P J, Powell W W 1991 Introduction. In: Powell W W, DiMaggio P J (eds.) *The New Institutionalism in Organizational Analysis*. University of Chicago Press Chicago pp. 1–38
Dobbin F 1994 *Forging Industrial Policy: The United States, Britain and France in the Railway Age*. Cambridge University Press, Cambridge, UK
Foucault M 1965 *Madness and Civilization: A History of Insanity in the Age of Reason* [trans. Howard R]. Random House, New York
Foucault M 1978 *The History of Sexuality* [trans. Hurley R]. Pantheon, New York, Vol. 1
Foucault M 1980 *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977* [Gordon C (ed.)] Pantheon, New York
Foucault M 1983 Afterword: The subject and power. In: Dreyfus H, Rabinow P (eds.) *Michel Foucault: Beyond Structuralism and Hermeneutics*, 2nd edn. University of Chicago Press, Chicago, pp. 208–26
Friedland R, Alford R R 1991 Bringing society back in: Symbols, practices and institutional centralization. In: Powell W W, DiMaggio P J (eds.) *The New Institutionalism in Organizational Analysis*. University of Chicago Press, Chicago, pp. 223–62
Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York
Giddens A 1984 *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, Berkeley, CA
Joas H 1996 *The Creativity of Action* [trans. Gaines J, Keast P]. University of Chicago Press, Chicago
LeVine R A 1984 Properties of culture: An ethnographic view. In: Shweder R A, LeVine R A (eds.) *Culture Theory: Essays on Mind, Self, and Emotion*. Cambridge University Press, Cambridge, UK, pp. 67–87
Meyer J W 1983 Conclusion: institutionalization and the rationality of formal organizational structure. In: Meyer J W, Scott W R (eds.) *Organizational Environments: Ritual and Rationality*. Sage, Beverley Hills, CA, pp. 261–82
Meyer J W 1986 Myths of socialization and of personality. In: Heller T C, Sosna M, Wellbery D E (eds.) *Reconstructing Individualism: Autonomy, Individuality, and the Self in Western Thought*. Stanford University Press, Stanford, CA, pp. 208–21
Meyer J W 1987 The world polity and the authority of the nation state. In: Thomas G M, Meyer J W, Ramirez F O, Boli J (eds.) *Constituting State, Society, and the Individual*. Sage, Beverly Hills, CA, pp. 41–70
Meyer J W, Boli J, Thomas G M, Ramirez F O 1997 World society and the nation-state. *American Journal of Sociology* **103**: 144–81
Mohr J, Duquenne V 1997 The duality of culture and practice: Poverty relief in New York City, 1888–1917. *Theory and Society* **26**: 305–56
Ortner S 1984 Theory in anthropology since the sixties. *Comparative Studies in Society and History* **26**: 126–66
Ortner S B 1989 *High Religion: A Cultural and Political History of Sherpa Buddhism*. Princeton University Press, Princeton, NJ
Parsons T 1937 *The Structure of Social Action*. McGraw-Hill, New York
Parsons T 1951 *The Social System*. Free Press, Glencoe, IL
Schneider D M 1976 Notes toward a theory of culture. In: Basso K H, Selby H A (eds.) *Meaning in Anthropology*. University of New Mexico Press, Albuquerque, NM, pp. 197–220
Sewell W H Jr 1992 A theory of structure: Duality, agency, and transformation. *American Journal of Sociology* **98**: 1–29
Sewell W H Jr 1996 Historical events as transformations of structures: Inventing revolution at the Bastille. *Theory and Society* **25**: 841–81
Sewell W H Jr 1999 The concept(s) of culture. In: Bonnell V E, Hunt L (eds.) *Beyond the Cultural Turn: New Directions in the Study of Society and Culture*. University of California Press, Berkeley, CA, pp. 35–61
Shore B 1996 *Culture in Mind: Cognition, Culture, and the Problem of Meaning*. Oxford University Press, New York
Skocpol T 1985 Cultural idioms and political ideologies in the revolutionary reconstruction of state power: A rejoinder to Sewell. *Journal of Modern History* **57** (March): 86–96
Swidler A 1986 Culture in action: Symbols and strategies. *American Sociological Review* **51**: 273–86
Swidler A 1992 Inequality in American culture: The persistence of voluntarism. *American Behavioral Scientist* **35** (March/June): 606–29
Swidler A 1995 Cultural power and social movements. In: Johnston H, Klandermans B (eds.) *Social Movements and Culture*. University of Minnesota Press, Minneapolis, MN, pp. 25–40
Swidler A 2001 *Talk of Love: How Culture Matters*. University of Chicago Press, Chicago
Turner V 1974 *Dramas, Fields and Metaphors: Symbolic Action in Human Society*. Cornell University Press, Ithaca, NY
Walzer M 1973 *The Revolution of the Saints*. Harvard University Press, Cambridge, MA
Weber M 1946 [1922–3] The social psychology of the world religions. In: Gerth H H, Mills C W (eds.) *From Max Weber*. Oxford University Press, New York, pp. 267–301
Weber M 1958 [1904–5] *The Protestant Ethic and the Spirit of Capitalism* [trans. Parsons T]. Charles Scribner's Sons, New York
Weber M 1968 [1920–2] *Economy and Society: An Outline of Interpretive Sociology* [Roth G, Wittich C (eds.)] University of California Press, Berkeley, CA
Weber M 1993 [1922] *The Sociology of Religion*. Beacon Press, Boston
Wrong D 1961 The oversocialized conception of man in modern sociology. *American Sociological Review* **26**: 183–93
Wuthnow R 1987 *Meaning and Moral Order: Explorations in Cultural Analysis*. University of California Press, Berkeley, CA

A. Swidler

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Cultural Geography

Cultural geography is a subdisciplinary category that is credited as being 'perhaps the most ambiguous term in the discipline's lexicon' (Price and Lewis 1993). There are two reasons for this ambiguity. First, the term 'culture' is often not defined; there is simply an 'easy coupling with geography in which both refer to a general sensitivity to issues of context, difference, and the local' (Barnett 1998, p. 631). Second, there has been a proliferation of cultural geographical approaches since the 1980s: 'the revival of cultural geography in a radically new form was one of the most striking [disciplinary] developments of the 1980s' (Gregory 1994, p. 98).

1. *The Berkeley School*

The term 'cultural geography' can be traced to nineteenth century German scholarship, but is thought to have been introduced into American geography in the 1920s by Carl Sauer (Price and Lewis 1993). Sauer, his colleagues in the geography department at Berkeley, and their students dominated American cultural geography until the 1980s. (This is not to say that this was the only strand of 'traditional' cultural geography. For a contemporary overview of a broader range of traditional cultural geographies, see Foote et al. 1994). Sauer (1925) framed cultural geography in opposition to environmental determinism, which he criticized for subjecting the diversity of cultures to a monocausal, nomothetic theory. Sauer instead advocated historical synthetic accounts of human interaction with natural environments. For Sauer, landscape study was an exercise in historical reconstruction that sought to show how a particular cultural group, working on and through the natural landscape, enacted a cultural landscape. The relevant unit of observation was the cultural region, defined as an area over which a functionally coherent way of life dominated. Much of Sauer's fieldwork was located at the US/Mexican border, and he explored the border as a cultural, rather than a physical divide. In the prologue to his important book, *The Early Spanish Main*, Sauer wrote that: 'the same mountains and desert, pine forest, and woodlands, scrub, and grasslands extend north and south; the difference is the people and their ways' (Sauer 1966, p.V). Members of the Berkeley School explored a range of themes in mostly rural locations: the migration of human groups and the process of adapting a familiar culture to a new land; the domestication of plants and animals; the development, intensification, and dispersion of agriculture; ancient cultivation systems; and the genesis and diffusion of cultural traits and material culture (e.g., folk housing). (For fuller accounts of the Sauerian tradition, see *Ecology, Cultural*; *Cultural Landscape in Geography*; *Cultural Landscape in Environmental Studies*.)

1.1 *Recent Assessments of the Berkeley School*

In the early 1980s, as new approaches to cultural geography were developed, a (contested) critique of the Berkeley School emerged. Berkeley School cultural geography was charged with conceiving cultural groups and regions as unities, and deploying a superorganic concept of culture that invested agency in culture rather than individuals. Power relations were given short shrift (Duncan 1980). The accuracy of this critique has been disputed (Price and Lewis 1993) but, even among those who accept it, there is now a renewed appreciation of the cultural ecology tradition that grew out of the Berkeley School. Crang (1998) draws attention to the ways in which Sauer bound the material and social, viewing crop species and gene pools, for example, as 'material expression and embodiment of social processes and knowledge' (p. 16). Willems-Braun (1997, p. 705) acknowledges the significance of studies of pre-contact indigenous cultural transformation of nature; they offer important tools for those who wish to disrupt the equation of 'native' and 'nature.' Even the disdainful treatment given to the Berkeley School's preoccupation with establishing cultural hearths deserves reconsideration; Woods (1998), for example, deploys the notion of cultural hearth in relation to 'blues epistemology' as an indigenous knowledge. Woods insists that the US south, in particular the Mississippi Delta, is the center of African American culture and that its 'lands, rivers, streams, air, plants, and animals … must be restored to their sacred status' (p. 290). This is an empirical and strategic claim, which he poses in opposition to white writers who focus on the degradation of African American culture and thus the necessity for cultural assimilation. 'Not only is working-class African American culture not broken and deviant, it is a powerful world culture held together, and expanded, by repeated blues movements' (p. 289). The Mississippi Delta must be understood as a 'Mecca', and African American blues culture as a model for sustainable communities.

2. *New Cultural Geographies*

Recent tendencies toward accommodation mute a sense of subdisciplinary transformation that existed in the late 1970s and 1980s, when a new generation of cultural geographers began to engage with a diverse range of philosophical traditions, and social and cultural theories. In some ways paralleling Sauer's rejection of nomothetic theory, humanistic geographers positioned themselves against spatial science and drew upon phenomenology, existentialism, and symbolic interactionism to recover the role of individual agency in the construction of place and cultural meanings (see *Humanistic Geography*). Hist-

Figure 1
Scripting Travel on the Nile (Charles Dana Gibson) (This is historical material that is out of copyright).

orical materialism, and then an increasingly diverse array of cultural theories have also been drawn upon, such that the subdisciplinary—indeed disciplinary—boundaries of cultural geography are now exceedingly blurry. Some of this history and excitement can be organized around the themes of representation, power, and cultural identity.

2.1 Representation

The linguistic turn of the 1980s was felt throughout the social sciences and humanities; in cultural geography it focused attention on the landscape as text and the politics of the representations that geographers produce (i.e., on the culture of geography).

2.1.1 Landscape as text. There is a long tradition of 'reading' the landscape in cultural geography. Peirce Lewis (1979, p. 12), for example, described the landscape as 'our unwitting autobiography.' Since the 1980s, however, the metaphor of landscape as text has been pursued more rigorously through a fuller engagement with literary and cultural theory, for example, in Duncan's (1990) study of the Kandyian Kingdom in early nineteenth century Sri Lanka. Duncan describes how the king of Kandy concretized two intertwined discourses on kingship through a massive city-building program, in an attempt to secure his political power; this landscape transformation was then interpreted differently by the king, nobles, and peasants, through the lens of the two key texts. The story, then, is of intertextuality, of the interplay of discourses enacted in landscape and texts.

This textualization of landscape has itself been criticized, less in relation to Duncan's specific empirical application than as a general theory of landscape interpretation. It is criticized for erasing process (Gregory 1994), for over-emphasizing the coherence of texts and landscapes, and for suppressing traces of non-human others: 'it treats the landscape as a blank page that only human actors can read and write upon' (Demeritt 1994, p. 170). Nevertheless, the metaphor has been reworked, for example around the notion of theatre and the script, to draw out the open, performative possibilities of text—partly structured and partly improvised. Gregory (1999), for example, writes about the scripting of Egypt in the nineteenth century by a growing tourism industry. European tourists not only read voluminously when traveling down the Nile (Fig. 1), the guidebooks also provided stage directions for transforming *dahabeeah* (floating barges) into 'secure viewing platforms' (e.g., English travelers were advised to bring 60 yards of chintz as a means of effecting this), and the remains of ancient Egypt were literally and materially staged as an 'extended exhibition.' It is this attention to the materiality and spatiality of writing, in this case travel writing, that distinguishes the work of cultural geographers (Duncan and Gregory 1999).

Figure 2
Cartoon of the looting of the Stockholm Exhibition of 1897. (Material out of copyright)

2.1.2 Landscape and maps as ways of seeing. Humanistic geographers of the early 1980s were drawn to the textual metaphor; Cosgrove (1985) entreated them to pay closer attention to the visual, as a way of understanding one of their key, though untheorized, concepts, that of landscape (see *Cultural Landscape in Environmental Studies*). In Gregory's (1994, p. 98) assessment, this recognition of the 'conceptuality' of landscape is one of the 'cardinal achievements' of the revival of cultural geography in the 1980s. Cosgrove traced the emergence of landscape as a visual ideology, as a way of seeing, in fifteenth and sixteenth century Europe. The discovery of linear perspective enabled a realist illusion of space: 'Landscape is thus a way of seeing, a composition and structuring of the world so that it may be appropriated by a detached individual spectator to whom an illusion of order and control is offered through the composition of space according to the certainties of geometry' (Cosgrove 1985, p. 55). This way of seeing coincided with and supported a transition from feudalism to capitalism, and new social relations with nature, and land as property.

What landscape paintings achieve aesthetically 'maps, surveys, and ordnance charts achieve practically' (Cosgrove 1985, p. 55); this recognition has led to a critical reassessment of another of geographers' representational forms: the map. The assumed link between reality and representation has been broken, and maps are now read as 'mechanisms for defining social relationships, sustaining social rules, and strengthening social values' (Harley 1992, p. 237), in short, as technologies of power. Harley has used deconstructive methods to read a simple road map, specifically a map of the State of North Carolina. He begins his interpretation at the margins of the map—for example, with the pictures of points of interest (that include a Cherokee woman making beaded jewelry, and a sand dune), a ferry schedule, and a message from a state official—to argue that the map does more than help ordinary motorists find their way. Through the marginal, mundane notes, and images, the map constructs a mythic geography: a landscape full of points of interest; it affirms State sovereignty; and it sustains the American 'love affair' with the car. Maps are such effective technologies of power because they 'operate behind the mask of seemingly neutral science' (Harley 1992, p. 238) (see *Cartography*; *Cartography, History of*).

2.1.3 Other ways of seeing. Cultural geographers have sought to unmask this neutrality through their own representations; one of the most adventurous has been Pred (1995,) who experiments with font size, repetition, and montage to engage us in (re)cognizing European modernities. In his essay on the Stockholm Exhibition of 1897, Pred invites the reader to participate in the 'phantasmagorical dream' of the exhibition by leading us through a tour of endless lists of commodities and visual pleasures. This description is jostled by fragments of quotations from cultural theory, newspaper and government reports, and personal anecdote. Rather than being declarative and closed, Pred's analysis emerges through a series of unanswered questions. An unlabeled cartoon (Fig. 2) is

repeated incessantly, its interpretation shifting in juxtaposition to different portions of the text. We eventually learn that it is a cartoon of the extensive looting that took place at the close of the exhibition. We more fully understand Pred's repetition of the cartoon when he asks, in closing the essay, whether the looting—which was largely erased from official history—is not in fact an overdetermined metaphor for the exhibition. Pred's experiments in montage are an effort to 'force discordant fragments to 'whisper' and 'shout' at each other in polylogue' (1995, p. 26) in order to re-present contemporary commodity society.

2.2 *Material and Symbolic Geographies as Technologies of Power*

It should now be clear that power and politics have been central concerns for contemporary cultural geographers (Jackson 1989). Landscapes not only express social relations, they are also an important means of enacting them. And landscapes, like maps, are such effective technologies of power because they tend to naturalize these relations (Cresswell 1996, Mitchell 1996). Pred (1995), for example, understands Stockholm's exhibitions as training grounds for citizen-consumers. The exhibition of 1897 afforded visitors new perspectives on the city: 'Spy on others without their knowledge. Observe restaurant customers shamelessly flirting with waitresses … Be titillated in these and other ways while simultaneously (re)learning the lessons of modern power, of panoptical bureaucratic power … Be reminded that the power of the State lies not so much in its ability to rally police or locally stationed troops at points of actual or potential disturbance, as in the ability of its constituent agencies to keep an eye on, to maintain bureaucratic surveillance over, each and every subject' (p. 61). By the 1930 exhibition, the geography lessons are quite different. The dominant architecture does not enable a bird's-eye view; it is a 74-meter advertising mast that gathers the crowd around the commodity: 'A mast to command all, to be seen from below … a mast to be dwarfed by, to integrate one with the everyday masses, to instill a sense of solidarity' (p. 122). Cultural geographies organize people in space so as to communicate and effect social relations.

Considerable attention has been given to the various geographies that constitute colonial relations, in the past and in the 'colonial present' (see *Postcolonial Geography*). Sparke (1998), for example, traces the multiple material and symbolic geographies deployed by the Canadian State to extinguish two First Nations—the Wet'suwet'san and Gitksan—claims to sovereignty and land rights in an extended 374-day trial in 1987. The geographies include the relocation of the trial from Smithers (a small community in the heart of the claimed territory) to Vancouver, a distance of over 1200 kilometers, that imposed costs on many of those testifying for the First Nations, removed those testifying from their support networks, and placed the trial squarely within the metropolitan 'center.' This material geography, among others, was supplemented by symbolic ones, for example, instances of ridiculing First Nations' counter-maps, in which they translated House songs of territory into map format. The lawyer representing the Canadian federal government dismissed these maps: 'The place names, the names of creeks, rivers, and hills and all the other features, are none of them geographic names, they are Gitksan names' (Sparkae 1998, pp. 474–5) (see *Law and Geography*). Ironically, when these maps of traditional boundaries of Gitksan and Wet'suwet'-san territories were introduced in court, they were labeled, by the presiding judge, as 'the map that roared,' and Sparke argues that they did indeed roar in profoundly disruptive ways. It is this potential for resistant, transgressive, and transformative 'roaring,' and for the eruption of social conflict in and through conflicting claims to space, that captures considerable interest among cultural and social geographers (Cresswell 1996).

2.3 *Cultural Identity and Place*

A persistent theme for cultural geographers has been cultural identification with place (see *Place in Geography*); in the 1970s, humanistic geographers reactivated concerns about attachment to place, (sub)-cultural variations in use of and meanings attributed to places, and alienation in modern, bureaucratized landscapes such as suburbs (see *Humanistic Geography*). The latter concern also permeates a now vast literature on the 'disneyfication' of urban space, leading to the criticism that a romantic nostalgia for authentic landscapes runs through it (see *Postmodern Urbanism*).

What is newer is a profound unease about the bounding of cultural identity and place. This unease is tied to the ways that claims to territory function as mechanisms for social exclusion and control (see *Territoriality: Geographical*). In an early attempt by a cultural geographer to engage with Said's analysis of Orientalism, Anderson (1991) argued that the creation of Chinatown as a space apart in early twentieth century Vancouver was constitutive of Chinese people as a race. It is not only that Chinatown 'expressed' racist sentiment or racial ideology; the definition of Chinatown as unhygienic (marked, for example, by listing it as a separate entry in public health records—it was the only neighborhood to be listed in this way) and a place of vice (which altered police practices in ways that produced more crime; see *Crime, Geography of*) was part of the process of racializing Chinese people as 'other' than Canadian. Using the psychoanalytical concept of abjection, Sibley (1995) fur-

ther explores the compulsion on the part of dominant groups to 'purify' space. Designating a group, such as gypsies, to a marginal, polluted place is quite literally part of the process of marginalizing the group (see *Social Geography*; *Ethnic conflict, geography of*; *Locational Conflict (NIMBY)*).

Mitchell (1995) traces this concern back to the concept of culture itself, arguing that it is a modernist concept, integrally tied to notions of otherness: 'culture is a concept to stop flux in its tracks, creating stability and "ways of life" where before there had been changes and contest. The idea of culture demanded a mapping of boundaries and edges, the specification of a morphology: culture had to become a bounded object that ultimately differentiated the world' (p. 107). He urges geographers to resist conceiving culture as an ontological entity, and to focus instead on the work done by the idea of culture. In whose interests is the idea of culture deployed? Who has the capacity to define cultural boundaries?

These concerns have led to attempts to conceive of identities that exceed fixed boundaries; for cultural geographers this also entails imagining the spaces in which this might happen. A striking example of such an exercise is Law's (1998) discussion of one bar in Cebu City in the Philippines. Law criticizes the discourses currently used by Western feminists and non-governmental health organizations to frame sex workers in the Philippines (as either victim or economically empowered), in part because sex workers are unable to recognize themselves within them. She draws on her ethnography of one bar to represent both dancers and their male, mostly Australian, customers in more nuanced terms. This involves conceiving shifting relations and identities so as to render neither dancer nor customer as powerful or powerless. The space of the bar is more than a backdrop to this ambiguous dance of identities and power relations; it is the liminal space in which it can happen. By pursuing the spatialities of hybrid, mobile identities, geographers have a good deal to contribute to a more free-floating discussion in cultural studies. In Law's case, she offers a cultural representation that might lead to more effective communication between dancers and health organizations around HIV/AIDS education (see *AIDS*, *geography of*).

Attempts to rethink the relations between identity and place have been provoked also by the realizations not only that cultural identities are multiple and shifting, and that multiple identities inhabit the same place, but also that places are open and unbounded. The global flows through the local, in terms of financial capital, commodities, ideas, and migration of individuals and households (see *Globalization: Geographical Aspects*). The attachments that individuals may feel to multiple places (captured by the term polylocality) enrich and complicate increasingly multi-cultural locales. Tensions arising from multi-culturalism often become manifest as landscape conflict. Ley (1995) has written an exemplary case study of one such conflict over the transformation of the affluent neighborhood of Shaughnessy in Vancouver. Shaughnessy was built in the early twentieth century, by mainly foreign-born English immigrants, to the standards of conservative English landscape tastes. This English theming has little cultural salience for ethnic Chinese immigrants who have come to Vancouver in increasing number since 1980 from nations in the Pacific Rim. Ethnic Chinese elites have rebuilt parts of Shaughnessy along a different landscape aesthetic, and the construction of what have been called 'monster houses,' sometimes necessitating the removal of trees, has been hotly contested by the old, Anglo-identified elite. Ironically, this elite—themselves immigrants or descendents of immigrants—deployed the rhetoric of roots, nature, and preservation to naturalize their claims to place and to convey their deeply-held sentiment for the landscape. A strength of Ley's analysis is that he conveys the old elite's topophilia, love of place, and the integrity and intensity of their griefing for a lost landscape, while simultaneously problematizing their exclusive claims to an urban—indeed national—space, and drawing out the interwoven colonial histories of the Anglo and Chinese elites. Ultimately, Ley makes an incident involving the felling of two sequoias on a residential lot speak to the problem of constructing a homogeneous national citizen in what has been identified as the first 'postmodern' State (Canada). This movement between scales, and the demonstration of the interwoven construction of scale, are features of the best contemporary work in cultural geography (see *Scale in Geography*).

3. *Beyond Dualisms*

If Sauer defined the cultural against environmental determinism, and humanistic geographers defined the cultural against economism, recent theorizing disrupts both sets of implicit dualisms, leaving Matless (1996, p. 380) to ask: 'Is "cultural geography" then a nonsensical term?' Rather than nonsense, it is gaining a new sense in relation to, rather than in opposition to, nature and economy.

3.1 *Nature and Culture*

Geographers have drawn from Bruno Latour and Donna Haraway to interrogate the assumed duality between nature and culture (Demeritt 1994, Mitchell 1995, Matless 1996, see *Nature–Society in Geography*). There are two issues at stake. First, the division between nature and culture enforces the view that humans are the sole social agents. An artifactual understanding of nature pluralizes agency; cultural geographers are now willing to consider that we live in a livelier world in which nonhuman actors also have

agency. The notion of agency has been reworked, away from that of a conscious and controlling self, to one of having effects. Second, conceptualizing the boundary between nature and culture as a social construction has opened up a rich set of investigations around both the production of the boundary and slippages across it. Anderson (2000), for example, examines the ways in which racial ideology has been interwoven with discourses of animality and nature, and Emel (1995) argues that the eradication of wolves in the USA has been underwritten by norms of masculinity.

3.2 Economy and Culture

The lines between economy and culture are no less blurry. Economic development is increasingly about culture, whether it be in the form of tourism or the redevelopment of urban areas for the purposes of spectacle and consumption (see *Postmodern Urbanism*). Access to jobs and job performances are interpreted as cultural phenomena. Economic theory, models, and methods are being read for the ways in which they are structured by metaphor (see *Economic Geography*). Gibson-Graham (1996), for example, argues that recent discussions of the globalization of capitalism function through the metaphor of rape (see *Globalization: Geographical Aspects*). Drawing on feminist attempts to conceive a perspective on rape in terms other than victimization, Gibson-Graham entreats us to imagine possible economic futures that exceed the 'rape' of capitalism.

The spirit of the Gibson-Graham analysis, of conceiving of non-violent trespass, is perhaps a good note on which to end. Cultural geography is currently extremely lively, no longer operating as a bounded sub-discipline but as a critical perspective on the production of boundaries and processes of differentiation.

Bibliography

Anderson K 1991 *Vancouver's Chinatown*. McGill-Queens University Press, Montreal, PQ

Anderson K 2000 'The beast within': Race, humanity, and animality. *Environment and Planning D: Society and Space* **18**: 301–20

Barnett C 1998 Cultural twists and turns. *Environment and Planning D: Society and Space* **16**: 631–4

Cosgrove D 1985 Prospect, perspective and the evolution of the landscape idea. *Transactions of the Institute of British Geographers* **10**: 45–62

Crang M 1998 *Cultural Geography*. Routledge, London

Cresswell T 1996 *In Place/Out of Place: Geography, Ideology and Transgression*. University. of Minnesota Press, Minneapolis, MN

Demeritt D 1994 The nature of metaphors in cultural geography and environmental history. *Progress in Human Geography* **18**: 163–85

Duncan J S 1980 The superorganic in American cultural geography. *Annals of the AAG*. **70**: 181–98

Duncan J S 1990 *The City as Text: The Politics of Landscape Interpretation in the Kandyian Kingdom*. Cambridge University Press, Cambridge, UK

Duncan J S, Gregory D (eds.) 1999 *Writes of Passage: Reading Travel Writing*. Routledge, New York

Emel J 1995 Are you man enough, big and bad enough? Ecofeminisms and wolf eradication in the USA. *Environment and Planning D: Society and Space* **13**: 707–34

Foote K E, Hugill P J, Mathewson K, Smith J (eds.) 1994 *Re-Reading Cultural Geography*. University of Texas Press, Austin, TX

Gibson-Graham J K 1996 *The End of Capitalism (As We Knew It)*. Blackwell, Cambridge, MA

Gregory D 1994 *Geographical Imaginations*. Blackwell, Cambridge, MA,

Gregory D 1999 Scripting Egypt: Orientalism and the cultures of travel. In: Duncan J, Gregory D (eds.) *Writes of Passage: Reading Travel Writing*. Routledge, New York

Harley J B 1992 Deconstructing the map. In: Barnes T J, Duncan J S (eds.) *Writing Worlds: Discourse, Text and Metaphor in the Representation of Landscape*. Routledge, New York

Jackson P A 1989 *Maps of Meaning: An Introduction to Cultural Geography*. Unwin Hyman, London

Law L 1997 Dancing on the Bar: Sex, money and the uneasy politics of third space. In: Pile S, Keith M (eds.) *Geographies of Resistance*. Routledge, London

Lewis P 1979 Axioms for reading the landscape. In: Meinig D W (ed.) *The Interpretation of Ordinary Landscapes*. Oxford University Press, New York

Ley D 1995 Between Europe and Asia: The case of the missing sequoias. *Ecumene* **2**: 185–210

Matless D 1996 New material? Work in cultural and social geography, 1995. *Progress in Human Geography* **20**: 379–91

Mitchell D 1995 There's no such thing as culture: Towards a reconceptualization of the idea of culture in geography. *Transaction of the Institute of British Geographers* **20**: 102–16

Mitchell D 1996 *The Lie of the Land: Migrant workers and the California Landscape*. University of Minnesota Press, Minneapolis, MN

Pred A 1995 *Recognizing European Modernities: A Montage of the Present*. Routledge, New York

Price M, Lewis M 1993 The reinvention of cultural geography. *Annals of the AAG* **83**: 1–17

Sauer C O 1925 The morphology of landscape. *University of California Publications in Geography* **2**: 19–54

Sauer C O 1966 *The Early Spanish Main*. University of California Press, Berkeley, CA

Sibley D 1995 *Geographies of Exclusion*. Routledge, London

Sparke M 1998 A map that roared and an original atlas: Canada, cartography, and the narration of nation. *Annals of the AAG* **88**: 463–95

Willems-Braun B 1997 On cultural politics, Sauer, and the politics of citation—reply. *Annals of the AAG* **87**: 703–8

Woods C 1998 *Development Arrested: The Cotton Blues Empire of the Mississippi Delta*. Verso, New York

G. Pratt

Cultural History

It is not easy to define 'cultural history' in its specificity. Should it be done by designating objects and practices whose study would constitute the very nature of this

history? The risk is great, therefore, of failing to draw a definite and clear line between cultural history and other histories: for instance, the history of ideas, the history of literature, the history of art, the history of education, the history of media, or the history of sciences. Should we, consequently, change perspectives and consider that all history, whatever its nature—economical or social, demographic or political —is cultural, insofar as the most objectively measurable phenomena are always the result of the meanings that individuals attribute to things, to words and behavior? In this fundamentally anthropological perspective, the problem is not so much to define the particular sphere of cultural history, differentiated from that of its neighbors, but rather avoid an imperialist definition of the category.

Between these two stumbling blocks, the road is narrow. From there, the course followed in this article consists of marking the shifts that have characterized the historiographical practices designated, in their time or subsequently, as belonging to cultural history.

1. *The History of Mentalities*

The very lengthy genealogy of cultural history generally stops at a certain number of precursors: the historians of the nineteenth century (Michelet, Burckhardt), Voltaire's, *Siècle de Louis XIV*, the legal practitioners of the 'perfect history' during the sixteenth century, even Herodotus himself. The course is not without illusion, measuring the 'anticipations' of certain precursors according to a posterior state of historical science. So as to avoid this trap, it is preferable to limit the research to the twentieth century and to follow the different faces taken by the historiographical projects which intended to focus on the phenomena left aside by the classical forms of political, economical, or social history. The history of mentalities was the first amongst them.

1.1 The Founders of the Annales and the History of Mentalities

Unlike a given idea, the paternity of the category of the 'history of mentalities,' insofar as it indicates a particular area of history, is not to be attributed to the founders of the *Annales* Lucien Febvre and Marc Bloch. It is the invention of Robert Mandrou and Georges Duby at the end of the 1950s. The expression appears for the first time in the title of a university course in 1956 with the election of Robert Mandrou in the 6th Section of the Ecole Pratique des Hautes Études to the chair of 'Social History of Modern Mentalities', while at the same time, Georges Duby was opening a seminar on 'medieval mentalities' at the University of Aix-en-Provence. In 1961, when Robert Mandrou published his book *Introduction à France, Modern*, he presented it as a response to the 'requests formerly made by Lucien Febvre in favor of the history of collective mentalities' (while using another notion as the subtitle of the study, 'psychological history' which was also dear to Febvre). Thus, designated as the inspirer of this new form of history, Lucien Febvre, in fact, had rarely used the exact term 'mentality,' preferring the adjective 'mental' added to words like 'equipment,' 'material,' but above all 'tool,' or even, in the plural form, 'habits' and 'needs.' In 1942, in the *Problème de l'incroyance au XVIe siècle: Religion de Rabelais* he characterized the 'mental tools' by expressing two essential statements: that the ways to perceive and reason are neither invariable nor universal, and that there is no continuous and necessary progress in the succession of mental tools. The book makes an inventory of the instruments and conceptual categories which are the different supports for thinking: first, the state of language, with its vocabulary and syntactic particularities, the tools and the languages available in the operations of knowledge, and finally the value and the credit attributed to each sense. From there, the conclusion: 'So close to us in appearance, the contemporaries of Rabelais are already distant by all their intellectual properties. And even their structure was not ours.' In each time, the ways of thinking and feeling outline in a specific way the limits between nature and the supernatural, or between what is possible and impossible.

Paradoxically, Lucien Febvre, taken to be the father of the history of mentalities, uses the term less often than Marc Bloch who uses it as much in the *Société féodale* (where we encounter 'mental atmosphere' or 'religious mentality') as in *Apologie pour l'histoire ou le métier d' historien*. He prefers 'sensitivity' that defines the subject of 'historical psychology' which seems for him the only capable way of avoiding the culpable anachronism which equips men and women of the past, not only with knowledge and conceptions which were impossible for them, but also with feelings and emotions that were unknown to them. Febvre stigmatizes such an error in *Amour sacré, amour profane: autour de l'Heptaméron* (1944) by concluding: 'In fact, a man of the sixteenth century should be intelligible not in relation to us, but in relation to his contemporaries.'

In his book of 1961, Robert Mandrou does not dissociate the study of the mental tool from that of sensations, emotions, and passions which make up the mentality. Without separating these two dimensions, he distinguishes the common elements to all men (and women) in a shared time and place and those that are particular to each generation, to each profession, to each social group, or to each class. For him, 'all historical psychology, all history of mentalities is social history.' We should, therefore, once the common mental tool has been identified, describe 'the mental horizons which are characteristic of the dif-

ferent social groups.' The 'history of mentalities,' or of the 'visions of the world,' another term often used by Mandrou, was thus strongly anchored in the differences between social classes, defined more by the unity of life style and the feeling of belonging than by a strict economical determination.

1.2 The Golden Age of the History of Mentalities

From the 1960s onwards, the notion of mentality imposes itself to designate a history whose object is neither ideas, nor socioeconomical realities. This 'French' history of mentalities reposes on a certain number of ideas more or less shared by those who practice it (see Le Goff 1974). In the first instance, the object of the history of mentalities is defined by Le Goff as the opposite of that of classical intellectual history: 'The level of the history of mentalities is that of everyday life and the automatic, which is what escapes individual subjects as it shows the impersonal content of their thoughts.' These ideas, which result from the conscious elaboration of a singular spirit, are therefore opposed to mentality, always collective, which regulates, without them knowing, the immediate perceptions of social subjects. Such an expression is not very far off the definition of collective representations in the tradition of the sociology of Durkheim as the accent is placed on the contents or the methods of thinking which result from the unconscious incorporation of unknown determinations in each member of a community, which set up their common manner of classifying and judging. Now, the second characteristic underlined by Le Goff: the possibility that the history of mentalities or the historical psychology 'link themselves to another important trend of historical research today: quantitative history.' Having as subjects collective, automatic, and repetitive actions, the history of mentalities should and must be serial and statistical. It is part of the heritage of the history of economies, populations, and societies that, on the horizon of the major crisis of the 1930s, followed by the mutations after the war, constituted the most innovative field of historiography. When, in the 1960s, the history of mentalities and the historical psychology defined a new, promising, and original area of study, they did so often by recapitulating the methods which ensured the conquests of socioeconomical history: the techniques of regressive statistics and the mathematical analysis of series.

From the importance given to series, and therefore to the establishment and treatment of homogenous data, repeated and comparable at temporal regular intervals, two consequences follow. The first is the privilege given to massive sources, widely representative and available over a long period, for example, the inventories after death, wills, library catalogues, legal archives, etc. The second consists of the attempt to articulate, according the braudelian model of different times (long term, conjuncture, event), the long period of mentalities which often resists to change, with the short period of brutal ruptures or of rapid transfers of belief and sensitivity.

The withdrawal of witchcraft as a criminal act in France during the seventeenth century (see Mandrou 1968), the transfer of attitudes before childhood or death (see Ariès 1964, 1977) or the dechristianisation of France during the second half of the eighteenth century (see Vovelle, 1973) illustrate the articulation of the different periods of the history of mentalities. In each case, the problem lies in understanding how, in the long-term stability of mental structure, an essential transfer occurs: thus, the transformation of the representations of the world in the mileu of magistrates, the invention of childhood, and the concealment of death, or the indifference towards devotion practices, moral injunctions, and catholic beliefs.

A third characteristic of the history of mentalities in its golden age lies in its ambiguous way of considering connection to society. The notion seems, in fact, dedicated to erasing the differences in order to find the categories shared by all the members of a same era. 'The mentality of a historical individual, a great man, is exactly what he has in common with other men of his time' writes Le Goff (1974), adding, as examples, 'it is what Caesar and the last soldier of his legions, Saint Louis and the peasant of his domains, Christopher Columbus and the sailor all have in common.' Amongst all the practitioners of the history of mentalities, Philippe Ariès is no doubt the one who made the greatest attachment to such an identification of the notion as 'common feeling' or 'general feeling.' The recognition of the 'archetypes of civilisation,' shared by a whole society, does certainly not signify the cancellation of all differences between social groups or between clerics and laymen. However, these differences are always considered inside a long-term process which produces representations and behavior which become common. Postulating the fundamental unity (at least tendentious) of the 'collective unconscious,' Philippe Ariès reads the texts and images, not like demonstrations of individual peculiarities, but in order to 'decipher, beyond the will of authors or artists, the unconscious expression of collective sensitivity,' or to 'find, beneath the ecclesiastic language, the ordinary set of common representations which are obvious' (see Ariès, 1975). The sensitivity and the collective gestures which are disclosed should be understood at the cross-roads of biology and the mental, at the meeting point of demographic realities (birth, death, etc.) and psychological investments (the forms of self-consciousness, the representations of life after death, the feeling of childhood, etc.). With Philippe Ariès, mentality refers to 'currents of the deep' which govern, without them necessarily being aware of it, the most essential attitudes of men and women of a same period.

For other mentality historians, more directly located in the heritage of social history, the essential elements lie in the link between the differences between the ways of thinking and feeling and social differences. Such a perspective organizes the classifying of mentality facts into divisions established by the analysis of society and then the superposition postulated as necessary between social boundaries that separate groups or classes and those which differentiate mentalities. This social cutting out is no doubt the most precise trace of the dependence of the 'history of mentalities' in relation to social history in French tradition. It was possible to understand it at a global and macroscopic level—and so in research aiming to characterize a mentality, a religion or a 'popular' literature, opposed finally to that of the dominant or the elite—or in a more fragmented way, in reference to the hierarchy of conditions and professions. However, in both cases, the study of mental horizons reproduces the divisions proposed by the history of societies.

1.3 History of Mentalities or Historical Psychology?

Mentalities, sensitivities, visions of the world: the unstable plurality of vocabulary indicates, at the same time, both the difficulty in defining objects of a new historiographical approach and the will to link, in the same perspective, intellectual and psychological categories.

When Alphonse Dupront proposes, in 1960, at the International Congress of Historical Sciences at Stockholm, to constitute historical psychology as a whole discipline within human sciences, he gives it a maximal extension as it must be 'the history of values, mentalities, forms, symbolics, myths.' Such a definition reduced the distance established by the founders of *Annales* between mentality and ideas, as the latter participate fully in 'collective mentality' of men of a period. The ideas, perceived through the circulation of words which designate them, situated in their social rooting, considered in their affective and emotional load as much as in their intellectual content, therefore become, just like myths or values, one of those 'collective forces via which men live their time,' one of those elements which Dupront, in words borrowed from Jung, called 'collective psychic.' An expression exists there which, while claiming to be loyal to the project of the *Annales*, surpasses the old oppositions by giving a fundamental psychological definition of mentality and by reintroducing the ideas in the exploration of the collective mental.

Such a perspective (without the word 'mentality') appears in the work of Ignace Meyerson whose importance, perhaps underestimated today, was central for the renewal of historical studies of Antiquity—in particular his book *Les Fonctions psychologiques et les oeuvres* of 1948. A first relationship lies in the assertion of the fundamental historicity of mental categories and the psychological functions. It is this essential historicity of psychological objects which allows Meyerson to define it as a 'historical anthropology': 'The psychological functions have a history and they have had different forms throughout this history. Time and memory have a history. Space has a history. Perception has a history. The person has a history.' The work of historical psychology does not, therefore, consist of finding different modalities or expressions of functions considered as stable and universal. It attempts to understand, in their discontinuity and their singularity, the emergence and the economy of each of these functions. In this way, by applying the perspective of Meyerson to the question of the person in ancient Greece, Jean-Pierre Vernant writes: 'There is not, there cannot be a model person, exterior throughout human history, with differences, variations according to places, transformations due to time. Research should therefore not establish if the person, in Greece, is or is not, but should research what the ancient Greek person is, how he is different, in the multiplicity of his features, of today's person' (see Vernant 1965).

On the other hand, Ignace Meyerson radically modifies the location and comprehension of the mental and psychological categories. To their immediate, existentialist, phenomenological grasp he opposes their knowledge based on symbolical forms, works, and acts in which they are objectivized. Analyzing psychological functions via productions (institutional, religious, legal, aesthetic, linguistic) allows for the rupture with the idea of universal and abstract men, and with the universalization of a particular form of the personality.

1.4 Success and Criticisms of the History of Mentalities

How can one explain the infatuation, of historians and readers, in France and outside France, for the history of mentalities, whatever be the designation, in the 1970s and 1980s? No doubt because such an approach allowed for, in its very diversity, the introduction of a new balance between history and social sciences. Contested in its intellectual and institutional superiority by the development of psychology, sociology, and anthropology, history coped by annexing the topics of the discipline which questioned its domination. The focus then moved towards objects (systems of belief, collective attitudes, ritual forms, etc.) which, until then, belonged to the neighbors but which fully entered into the program of a history of collective mentalities.

Adapting to the approaches and analysis methods of socioeconomical history while transforming the historical questionnaire, the history of mentalities (in its widest definition) was able to occupy the front part of the historiographical scene and constitute an effective response to the challenge launched by social sciences.

However, there were many critics of its principles and methods. The first came from Italy. In 1970, Franco Venturi denounced the obliteration of the creative force of new ideas for the benefit of simple 'mental structures' 'lacking dynamism and originality' (see Venturi, 1970). Some years later, Carlo Ginzburg magnified the criticism (see Ginzburg 1976). He refused the notion of 'mentality' for three reasons: first, for its exclusive insistence on 'elements which are inert, obscure and unconscious of a determined vision of the world,' which lead to reducing the importance of rationally and consciously expressed ideas; second, for the 'interclass' character which unduly assumes the sharing by the whole society of the same 'mental equipment'; and lastly, for the alliance with the quantitative and serial approach, which, all together, reifies the contents of thought and attaches itself to the most repetitive expressions and ignores singularities. Historians were thus invited to privilege individual appropriations more than statistical distributions, to understand how an individual or a community interpreted, according to its own culture, ideas and beliefs, and texts and books circulating in their society.

In 1990, in the book with the provocative title, *Demystifying Mentalities*, Geoffrey Lloyd, historian of Greek philosophy and science, hardened the indictment once again. The criticism lies within the two essential principles of the 'history of mentalities': on the one hand, allocating to a whole society a stable and homogenous set of ideas and beliefs; on the other hand, considering that all thoughts and all conducts of an individual are governed by a unique mental structure. The two operations are the very condition allowing a mentality to be distinguished from another and permitting the identification within each individual of the mental tool shared with his contemporaries. However, such a way of thinking erases, in the repetitions of the collective, the originality of each singular expression and it encloses within an artificial coherence the plurality of belief systems and ways of reasoning that a same group or a same individual can successively mobilise.

Lloyd therefore proposed to substitute for the notion of 'mentality' that of 'styles of rationality' whose use depends directly on the contexts of discourse and the domains of experiences. Each of them lays down their own rules and conventions, defines a specific form of communication, supposes particular expectations. This is why it is quite impossible to bring back the plurality of methods of thinking, knowing, and arguing to a homogenous and unique mentality.

The case was well pleaded but is it really justified? On one hand, the history of mentalities did not only detain the single globalizing definition of the notion, as it inherited it from Lévy-Bruhl, author in 1922 of *La Mentalité primitive, ou des psychologues* (Charles Blondel, Jean Piaget, and Henri Wallon). If Lucien Febvre surely was tempted by the 'interclass' definition of mentality—in particular in *Le Problème de l'incroyance au XVIe siècle*—and Philippe Ariès after him, Marc Bloch and Robert Mandrou were very attentive to the social differences which command, in a same society, different ways of thinking and feeling, or diverse visions of the world.

On the other hand, French historians have not always ignored the possible presence, within the same individual, of several mentalities, distinct or even contradictory. Le Goff strongly expresses it: 'The coexistence of several mentalities at a same time and in a same spirit is one of the delicate but essential elements of the history of mentalities. Louis XI, who, in politics, revealed a modern 'Machiavellian' mentality, in matters of religion revealed a very traditional superstitious mentality.'

The critical examination of the contributions and limits of the 'history of mentalities' should neither reduce the diversity of it nor simplify the expressions of it.

2. *From the History of Mentalities to Cultural History*

The path of the history of mentalities is, therefore, made up of paradoxes. While it claimed to be clearly differentiated from other historical practices, it was never able to define clearly and unanimously its objects, methods, and concepts. It was during the years when it underwent the most criticism and when expression itself retreated for the benefit of other categories that the multiplication of works which explore the topics it designated were seen. It is no doubt due to the plasticity of its definition and the diversity of its uses that this history, omnipresent and inaccessible, was able to durably characterize a form of *historiographical* work. It is, therefore, no longer possible today, in a time when the notion of 'cultural history' has become dominant, even sometimes supreme. This is what must now be examined.

2.1 *An Impossible Definition*

The difficulty in defining cultural history lies fundamentally in the even larger difficulty in defining what the object 'culture' itself is. The innumerable meanings of the term can be diagrammatically distributed into two families of meaning: one which designates the works and gestures which, in a given society, avoid economical or symbolical urgencies of daily life and are submitted to an aesthetic or intellectual judgement, and one which aims at ordinary practices through which a community, whichever community it may be, lives and thinks out its relations to the world, to others, or to itself.

The first order of meanings leads to building the history of texts, works, and cultural practices like a history with two dimensions, as Schorske suggests:

'The historian seeks to locate and interpret the artefact temporally in a field where two lines intersect. One line is vertical, or diachronic, by which he establishes the relation of a text or a system of thought to previous expressions in the same branch of cultural activity (painting, politics, etc). The other is horizontal, or synchronic; by it he assesses the relation of the content of the intellectual object to what is appearing in other branches or aspects of a culture at the same time' (Schorske 1979). We must, therefore, consider each cultural production in the history of its genre, discipline, or field as well as in context of its relationships with the aesthetic, or intellectual productions and the cultural practices which are contemporary to it.

The latter leads to the second family of definitions of culture. It strongly relies on the meaning that symbolical anthropology gives to notion—and in particular Geertz: 'The culture concept to which I adhere [...] denotes an historically transmitted pattern of meanings embodied in symbols, a system of inherited conceptions expressed in symbolic forms by means of which men communicate, perpetuate, and develop their knowledge about and attitudes towards life' (Geertz 1973). It is, therefore, the entirety of languages and of the symbolical actions of a community which constitute its culture. From there, for historians, the attention is transferred to collective expressions where a cultural system is expressed in a paroxysmal way: rituals of violence, rituals of passages, carnivalesque festivals, etc. (Davis 1975, Darnton 1984).

What the different approaches try to consider today is the paradoxical articulation between a difference—the one by which all societies, in varying methods, separated a field characterized by particular experiences and delights—and subordination—those which make the aesthetic and intellectual invention possible and intelligible by noting it in the social world and in the symbolical system particular to a time and a place.

2.2 *Plurality of Practices, Common Questions*

According to historiographical traditions and heritage, cultural history favored different objects, fields, and methods. Making an inventory of it is an impossible task—and partially futile as it would cause repetition of the assessments presented in the numerous articles of this encyclopedia. More significant, perhaps, is the pinpointing of the common questions to these very diverse approaches.

A first stake is the necessary articulation between singular works and common representations. There are several ways to conceive this: by concentrating on the particularities of each social space where works develop and circulate (guilds, the court, the academies, the market), by situating them in relation to texts and ordinary practices with which they are in 'negotiation' (to use a term which is dear to 'New Historicism'), or, using Elias's method, by understanding how aesthetic conventions refer to psychological economy and structure of personality in a time and a place.

A second question, very widely shared, lies in the relationships between popular and learned culture. The ways of perceiving them can be dealt with by using two large models of description and interpretation. The first, trying to abolish all forms of cultural ethnocentrism, treats popular culture as a coherent and independent symbolical system, which is organised according to irreducible to logic, that of well-read culture. The second, concerned remembering the existence of relationships of domination and inequalities of the social world, understands popular culture by its subordination and its weaknesses compared to the culture of the dominant. Therefore, on the one hand, popular culture is considered as a symbolical independent system, enclosed within itself, and on the other, it is entirely defined by its distance opposed to cultural legitimacy. For a long time, historians fluctuated between these two perspectives. Then, the work carried out on religion or literature treated as specifically popular and the construction of an opposition, repeated through time, between the golden age of a free and vigorous popular culture and the times of censorship and constraints which condemn it and dismantle it. Distinctions that are so clear are no longer accepted without doubts today, which leads us to consider all the mechanisms which cause internalization by the dominated to be of their own illegitimacy and expressions via which a dominated culture manages to save something of its symbolical coherence. The lesson is valuable for the confrontation between the elite and people in the old Europe (Ginzburg 1966, 1976) and for the relationships between the dominated and the dominant in the colonial world (Gruzinski 1988).

A final challenge for cultural history, whatever be the approaches and objects, lies in the articulation between practices and discourse. The questioning of ancient certainties took the form of a 'linguistic turn' which reposed on two essential ideas: that the language is a system of signs whose relationships themselves produce multiple and unstable meanings, beyond all intention or all subjective control; that 'reality' is not beyond discourse but is always built by discursive practices (Baker 1990).

Opposed to such a position, numerous are the historians who, following the distinctions proposed by Foucault between 'discursive formations' and 'non-discursive systems' (see Foucalt 1969) or by Bourdieu between 'practical sense' and 'scholastic logic' (see Bourdieu 1997), marked the difference between the logic of practices and that which governs the discursive production and which underlined the irreducibility between the reality which was (or is) and the discourse which intend to organise it, censure it, or represent it.

The fundamental object of a history attempting to recognize the way in which social actors give meaning

to their practices and discourse is, therefore, found in the tension between the inventive capacities of individuals or communities and, on the other hand, the constraints and conventions which restrict—more or less tightly according to the position that they occupy in their domination relationships—what is possible for them to think, express, and do. The acknowledgment is valid for well-read works and aesthetic creations as well as for ordinary practices—which is another way of expressing the double definition of the objects of cultural history.

See also: Collective Beliefs: Sociological Explanation; Collective Memory, Anthropology of; Collective Memory, Psychology of; Cultural Psychology; Culture, Sociology of; History and Memory; Intellectual History; Psychohistory

Bibliography

Ariès P 1964 *L'enfant et la vie familiale dans l'Ancien Régime*. Plon, Paris

Ariès P 1975 *Essais sur la mort en Occident du Moyen Age à nos jours*. Paris

Ariès P 1977 *L'homme devant la mort*. Seuil, Paris

Baker K M 1990 *Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century*. Cambridge University Press, Cambridge, UK

Bourdieu 1997 *Meditations Pascaliennes*. Editions du Seuil, Paris

Darnton R 1984 *The Great Cat Massacre and Other Episodes in French Cultural History*. Basic Books, New York

Davis N 1975 *Society and Culture in Early Modern France*. Stanford University Press, CA

Foucalt 1969 *L'Archéologie du savoir*. Gallimard, Paris

Geertz C 1973 *The Interpretation of Culture*. Basic Books, New York

Ginzburg C 1966 *I Benandanti. Stregoneria e culti agrari tra Cinquecento e Seicento*. Giulio Einaudi editore, Torino

Ginzburg C 1976 *Il formaggio e i vermi. Il cosmo di un mugnaio del'500*. Giulio Einaudi editore, Torino

Gruzinski S 1988 *La colonisation de l'imaginaire. Sociétés indigènes et occidentalisation dans le Mexique espagnol, XVIe–XVIIIe siècle*. Gallimard, Paris

Le Goff J 1974 Les mentalités. Une histoire ambiguë. In: Le Goff J, Nora P (eds.) *Faire de l'histoire*, Vol. III. Gallimard, Paris, pp. 76–94

Mandrou R 1961 *Introduction à la France moderne, 1500–1640. Essai de psychologie historique*. Albin Michel, Paris (re-edited in 1998 with an Afterword by Cottret M, Joutard P, Lecuir J)

Mandrou R 1968 *Magistrats et sorciers dans la france du XVIIe Siécle*. Plon, Paris

Schorske C 1979 *Fin-de-siècle Vienna. Politics and Culture*. Random House, New York

Venturi F 1970 *Utopia e riforma nell'Illuminism*. Einaudi, Tonno, Italy

Vernant J-P 1965 *Mythe et pensée chez les Grecs. Étude de psychologie historique*. Maspero, Paris

Vovelle M 1973 *Piété baroque et déchristianisation en Provence au XVIIIe siècle*. Plan, Paris

R. Chartier

Cultural Landscape in Environmental Studies

The concept cultural landscape has at least two meanings. It refers both to an empirical object of analysis, and to an approach to studying that object. Cultural landscape as an object of analysis refers to the material–cultural expressions of human modifications of 'nature' as they appear on a particular surface of the earth. The cultural landscape is generally thought to include all of the elements of the built environment (e.g., buildings, roads), as well as land-use patterns. In this usage, cultural landscape is juxtaposed with natural landscape, although the distinction between the two is problematic (hence the quotes). By contrast, the concept of landscape used in ecology tends to be more focused on nature and natural processes instead of the emphasis on culture and cultural mechanisms that forms the basis of the concept used in geography and the social sciences more generally. The cultural landscape as an approach, generally refers to an interpretive and inductive strategy for understanding the meaning of those cultural expressions, as opposed to a more scientific and deductive approach. These two meanings have an intertwined history, although the former meaning (object of analysis) preceded, and led to, the second. This essay is not divided into two parts, but rather how and why the different meanings emerged and often converged are pointed out (when relevant).

Introduced into American geography in the early twentieth century, the importance of the concept cultural landscape has waxed and waned in relation to the shifting contours and contexts of the history of the social sciences, expressed primarily in the discipline of geography. It has taken on diverse meanings, been subject to rigorous criticisms, and emerged in the late twentieth century as a vital and politically-charged concept. This essay traces the changing interpretations and uses of the concept 'cultural landscape,' before analyzing contemporary and future trends.

1. Historical Development

American geographer Carl Sauer's influential essay 'The Morphology of Landscape,' originally published in 1925 (1963), provided the first formal introduction of the concept of the cultural landscape into American social science, and laid a foundation on which it would build for the next 50 years. Partly in reaction to the methodological and philosophical flaws of what had become the reigning paradigm in geography—environmental determinism—Sauer attempted to situate human-environment relationships more firmly as a science by putting forward the concept of the cultural landscape as its distinctive object of analysis. Sauer posited a natural landscape, comprised of land forms,

climate, waterways, and vegetation, that was acted upon by human cultures to create a cultural landscape, composed of material expressions of population density and mobility, housing plan and structure, production, and communication. In his now famous phrase: 'The cultural landscape is fashioned from a natural landscape by a culture group. Culture is the agent, the natural area is the medium, the cultural landscape is the result' (Sauer 1925, p. 343). To make clear his position vis-à-vis environmental determinism, Sauer explicitly stated that nature is only important in that it provides the raw material that people use to fashion the cultural landscape: 'The shaping force, however, lies in the culture itself' (Sauer 1925, p. 343). So by positioning landscape studies in this way, Sauer was putting forward an empirically verifiable object of analysis (the cultural landscape), and was shifting the explanatory variable away from nature to culture. Sauer's 'Morphology of Landscape' proved particularly influential in American geography, partly because he drew on German and French traditions which at the time were thought to be more sophisticated in conceptualizing human–environment relationships, and partly because he provided a core concept that gave geography its own object of analysis distinct from other disciplines. For the next 30 or so years, the concept of the cultural landscape as advanced by Sauer was central to the emerging discipline of geography in the USA, and to human–environment studies more broadly. Sauer and the students he trained at the University of California at Berkeley formed a particularly influential group of geographers, and the methods and approaches they used to study cultural landscapes, influenced by the anthropology of A. Kroeber and R. Lowie, became almost paradigmatic.

As former students and others attempted to 'operationalize' Sauer's ideas in the 1930s and 1940s, two distinct trajectories became evident: the first stressed the enumeration of elements of the cultural landscape in order to delineate homogenous regions, while the second stressed spatial and temporal processes that created the cultural landscape. The first trajectory, often referred to as chorology, or areal differentiation, offered a way of studying the cultural landscape systematically by enumerating its various elements, looking for similarities in forms and structures (morphology), and deducing from this analysis areas or regions. The second interpretation looked at a particular cultural landscape as the end result of a set of spatial and historical processes. The significant spatial process in Sauer's approach was that of diffusion—the movement of people, ideas and things across space—and the historical processes referred to the role of cultures in transforming a 'natural' landscape into a cultural landscape. For example, geographers in the 1930s to 1960s studied the origins and diffusion of agricultural patterns, house types, and fences (Zelinsky 1973), while others reconstructed sequences of past cultural landscapes leading up to the present-day, and outlined the transformations from a so-called 'primitive landscape' to a 'cultural landscape.'

As these studies matured, the two trajectories became more divergent. The delineation of regions on the basis of cultural landscape elements (the first trajectory) moved away from the focus on cultural landscape elements, and more to the systematic analysis of spatial patterns. The quantitative revolution that swept through the social sciences in the 1960s was taken up by the disciplines more focused on cultural landscape, leading to diminished attention. The term 'cultural landscape' retained its currency for those who emphasized historical processes, and those who continued to use the three-dimensional definition of cultural landscape (the second trajectory). In these studies, the concept of the cultural landscape as an object of study and as an approach to that study converged. In other words, researchers concerned with outlining and analyzing how, over time, culture groups shaped their landscapes, and how cultural landscape forms can be interpreted as indicators of cultural norms, looked to historical and interpretive methods to answer their questions. Their emphasis on the particularities of landscapes, and of its three-dimensional qualities, was not amenable to abstract, generalizable models. The subfield of cultural geography coalesced around these studies and approaches, with the concept of cultural landscape forming an integral component.

Sauer and the students he had trained at Berkeley dominated these studies of the cultural landscape. Fred Kniffen at Louisiana State University and Wilber Zelinsky at Penn State University were particularly prolific, publishing articles on a range of cultural patterns from the New England Barn to the Covered Bridge. Kniffen presented a conceptual framework for these studies in his publication 'Folk Housing: Key to Diffusion' (1965), while the statements put forward by Phillip Wagner and Marvin Mikesell (also Sauer's former students) in their introduction to the edited collection *Readings in Cultural Geography* (1962) provided a firm definition of the concept: 'This term comprehends both the distinctive features that serve to set a region apart and those not confined to a particular region but occurring there. The cultural landscape, then, connotes the geographic content of a determined area, or a geographic complex of a certain type, in which the choices made and changes worked by men as members of some cultural community are manifested' (10). By the 1960s, the term cultural landscape had become synonymous with idiographic studies of the 'content of a determined area,' and formed a central tenet of the subfield of cultural geography.

2. *Criticisms and Reconceptualizations*

Criticisms of cultural landscape as the object of analysis for cultural geography were first given voice

by an Australian geographer Harold Brookfield. In his 1964 article, he questioned why cultural geography was limited to a study of the visible material elements of culture, and argued that only by examining issues on the 'human frontiers,' that is, those that are concerned with social, cultural, political factors, could geographers ever hope to answer the how and why of landscape patterns (Brookfield 1964). A variant on this theme was taken up in 1980 by geographer James Duncan, in a penetrating philosophical critique of the use of the superorganic form of explanation in cultural landscape studies and cultural geography more broadly. According to Duncan, these studies were fundamentally flawed since their explanations relied on a reified notion of culture that assigned it 'ontological status and causative power' (Duncan 1980, p. 181). By so doing, culture writ large could be said to explain landscape patterns, overlooking any in-depth analysis of the workings of cultural processes, and foreclosing the possibility of individual agency. Duncan's critique was particularly aimed at studies of vernacular material culture, studies that had relied on a fairly naïve understanding of 'folk' cultures to explain landscape patterns. According to Duncan, these studies provided rather sweeping explanations of patterns, but not in-depth understanding of process. As a result of such critiques, and of the quantitative shift in geography, the discipline that most focused on the topic, the concept cultural landscape as object of analysis and as approach was marginalized throughout the 1970s and 1980s.

This is not to say that the concept was moribund. Some of the most compelling cultural landscape studies were published during this time—essays that implicitly asked and answered questions that went beyond Brookfield's 'human frontiers' and that included human agency and cultural processes in their explanations. J. B. Jackson had begun publishing such essays in the 1950s in his journal *Landscape*, and throughout the 1960s and 1970s scholars from a range of disciplines including architecture, history, landscape architecture, urban planning, and geography, contributed provocative writings to the journal. Jackson was very influenced by the French school of landscape history that looked at the cultural landscape as the embodiment of a *genre de vie*—that is, of an economic, social, and moral order. For example, in distinction to the studies of vernacular house types that were used to delineate cultural regions, Jackson (1952) saw the house as the most significant element in the cultural landscape because it was the most tangible symbol of a people's way of life. In this way, the cultural landscape was not simply the passive imprintment of a society, but was an active expression, both reflecting and expounding the economic, social, and moral order of a group of people. Scholars such as Grady Clay (1980) took up this idea and generally applied it to the urban built environment at large—that is, to the way everything from street patterns to commercial buildings could be 'read' as statements about the way of life of the people responsible for building them. Even in disciplines seemingly unrelated, such as literary criticism (Conron 1973) and photography (Corner and MacLean 1996), scholars explored the idea of the cultural landscape as a potent visible symbol of a particular way of life.

Reconceptualizing the concept cultural landscape as a symbol opened it to new arenas of analysis and new methods. Geographer Donald Meinig, in the influential group of essays he edited, *The Interpretation of Ordinary Landscapes* (1979), spoke of the cultural landscape as 'a panorama, a composition, a palimpsest, a microcosm' (6), and argued for historical methods of analysis. For geographer Marywn Samuels (1979), the cultural landscape could be thought of as the physical manifestation of innumerable, individual decisions, and its interpretation could take the form of a 'biography,' with scholars outlining the contexts in which those individual decisions had been made. Drawing on a large range of writings, from religious studies to American literary criticism, geographer Yi-Fu Tuan (1974) undertook a sustained examination of human meanings expressed in cultural landscapes, and human experiences and responses to those meanings. For Tuan, the cultural landscape was more than the sum of material artifacts; it was full of psychological, religious, aesthetic, and moral meanings that people had given to it, and it evoked those meanings to other people. For these scholars, cultural landscape studies offered a way to explore cultures and societies beyond what was offered by the quantitative work on spatial patterns that dominated the discipline at the time.

A different sort of reconceptualization of cultural landscape began in the UK in the 1980s. Heavily influenced by the cultural materialism of Raymond Williams, cultural and social geographers reconceptualized the cultural landscape as the expression of, and a constituent force within, the complex social, economic and political structures that characterize cultures. In *Social Formation and Symbolic Landscape* (1984), British geographer Denis Cosgrove took on the idea of landscape as a particular way of seeing, and positioned that way of seeing within its European historical context, making clear its connections to the 'real historical world of productive human relations' (2). Cultural landscapes were seen not only as the result of human labor, but also as cultural products that 'expressed and supported a range of political, social and moral assumptions' (1). In distinction to a strict Marxist view, the culture of cultural landscapes was not seen as divorced from the 'real' world of socio-economic formations and thereby dismissed as unimportant, but was instead regarded as an integral component of that world.

This infusion of cultural materialism into cultural geography brought new life to the concept of cultural landscape. In a statement that set the way for a reconstituted 'new' cultural geography (in distinction

to the 'old' or Sauerian school), Denis Cosgrove and Peter Jackson (1987) positioned the cultural landscape concept prominently, and offered a 'new' way of interpreting that 'allows us to disclose the meanings that human groups attach to areas and places and to relate those meanings to other aspects and conditions of human existence' (96). Their suggestions of two such approaches to interpretation—using the metaphor of text to conceptualize the cultural landscape so that it could be 'read' as a social document, similar to what Clifford Geertz had proposed in anthropology; and using the art historian's notion of iconography to interpret the cultural landscape as a set of visual images that can be interpreted at various levels of significance—was highly influential in the work on cultural landscapes that would follow in the late 1980s and 1990s. But perhaps more influential was their shifting of the agenda of cultural geography, and by implication of cultural landscape work, away from its emphasis on rural, historical studies of fixed and homogenous places, to the urban, contemporary scene, and to the contested nature of its production. In Cosgrove's and Jackson's words, the new cultural geography should be 'interested in the contingent nature of culture, in dominant ideologies and in forms of resistance to them (1987, p. 95). These statements removed the stigma that had been attached to cultural landscape studies—as benign, but irrelevant work done by antiquarians in remote and unimportant places—and highlighted the importance of the cultural landscape, as both object of analysis and approach, in the poststructural, politically charged atmosphere of the 1980s and 1990s social sciences. With the 'cultural turn' that had taken hold in the social sciences, the concept cultural landscape was suddenly relevant and trendy.

3. *Current Usage*

The two approaches suggested by Cosgrove and Jackson—landscape as text and the iconography of landscape—have generated a host of empirical studies. Both approaches were borrowed from the humanities (iconography from art history and 'text' from literary criticism), and drew on what could be called a postmodern stance on interpretation, i.e., that the meanings that can be drawn from analyses of landscapes are never straightforward, nor are they unchanging. 'Instead of providing a transparent window on a "real" world,' Cosgrove and Jackson argue, cultural landscapes, like 'images and languages are now regarded as a sort of sign that presents a deceptive appearance of naturalness and transparence concealing an opaque, distorting, arbitrary mechanism of representation, a process of ideological mystification' (1987, p. 98). Influenced by the writings of such scholars as Jacques Derrida and Michel Foucault, practitioners began to see the cultural landscape as a representation (not a mimetic expression) of time and place-specific, contested cultures. The 'cultural turn' in the social sciences brought an awareness that landscape was, in the words of W. J. T. Mitchell, 'a medium of cultural expression' (1996, p. 4). But because landscape, 'does not easily accommodate political notions of power and conflict, indeed it tends to dissolve or conceal them,' the role of the researcher is to explore and reveal its 'duplicity' (Daniels 1989b, p. 196). For example, Daniels' essay on the landscape of Georgian England points out how the 'typical' landscape representation of the time—the tradition of the picturesque—served both to engage some of the political and economic problems of early nineteenth-century England, at the same time that its imagery served to 'obscure social and economic issues' (1989a, p. 73).

Drawing on similar poststructuralist assumptions about interpretation, but with insights gained from psychoanalytic theory and film criticism, feminist geographer Gillian Rose (1993) criticized the 'iconographic' cultural landscape school for its implied masculinism. Rose argues that despite the progressive aim of these new cultural landscape studies, they embody an approach that by definition is implicated in power relations and oppression of women. That approach is one of an objective viewer, coded in Western thought as masculine, 'gazing' at and thereby in power over, a 'feminized' scene or cultural landscape. These implied power relations make problematic any appreciation of or pleasure derived from the actual viewing of landscapes as image. Catherine Nash (1996) countered this criticism by showing how understanding the particular contexts of viewing and creating cultural landscape images can subvert or 'unsettle any easy claim of oppression in any act of seeing or visual representation' (152). Her analysis of two landscape images created by women that utilize traditional landscape genres, but intentionally conjoin landscape with images of the male body, shows how 'to view body as land or land as body has no essential meaning, yet neither can it ever be innocent' (167).

In a different vein, but still informed by feminist theory, Heidi Nast (1996) 'reads' the cultural landscape of the Kano Palace in northern Nigeria circa 1500 to reveal its active role in constituting and replicating the gendered and sexual politics of its time. By reading the changing cultural landscape of the Kano Palace in relationship to changes in Islam, sexual practices, and gender roles, Nast shows how spatial evidence like cultural landscapes can be used to interpret the past worlds of people who were excluded from power, and therefore were unable to leave any written record. Similarly, Lynn Stewart (1995) reveals how the cultural landscapes of Plantations in antebellum American South reinforced the legal codes that gave all eminent rights to the white land owners. However, Stewart also shows that within this oppressive landscape slave men and women were able to

create their own social spaces that allowed for self-expression and at times forms of transgression.

Others have taken an iconographic and textual approach to studying cultural landscapes within a postcolonial framework. Rod Neumann (1995), for example, examines how the British importation of ways of seeing landscape was integral to their imperialist efforts at the turn of the century. The British appropriated lands for national parks in east Africa by constructing them as landscapes as consumption, free of human labor and visually satisfying, in distinction to the lands they developed for agricultural production. By remaking African land into one that satisfied an English landscape tradition where production and consumption were spatially separated, the British undermined indigenous land rights and land-use practices, a 'phenomenon inseparable from the recasting of African society' (1995, p. 153). In a related fashion, Jane Jacobs (1996) takes a postcolonial look at the changing meaning of cultural landscapes both at the 'heart' of British imperialism—the City of London—and at its peripheries—Aboriginal land claims in Australia. She argues that the power relations of nineteenth-century imperialism endure and are evident even in our postmodern cities, and that these relationships are articulated and negotiated through debates about the meanings of cultural landscapes.

4. Future Directions

If the existence of debate is any indication of intellectual currency, then the concept cultural landscape will certainly enter the third millennium as a vital part of the geographic discipline and the social sciences more broadly. Recent controversies have centered on the 'correct' way to interpret and think about cultural landscapes—whether analyses have moved too far beyond materialist explanations (Mitchell 1997), and the degree to which 'cultural' explanations incorporate the political, economic, and social (Schein 1997). These debates signal that the concept is politically charged, both in the sense of it serving as a focal point of political debate among academics, but more importantly, as it has now come to be seen as a key concept in our understanding of power relations and potential challenges to power. Whether it concerns the actual material production and transformations of cultural landscapes, or the ongoing re-interpretations of its meaning, the cultural landscape will be a key source for understanding dominant ideologies, and challenges to them. The cultural landscape embodies and expresses relationships of power along all of its axes—of gender, sexuality, 'race,' class, ethnicity, disability, age—and future research lies in explorations of these power relations as they are constituted by, and reflected in, cultural landscapes, and to the ways that a cultural landscape approach and subject matter can help us understand and articulate challenges to those power relations.

See also: Architecture; Art History; Cultural Geography; Ecology, Cultural; Globalization and World Culture; Human–Environment Relationships; Humanistic Geography; Location Theory; Social Geography

Bibliography

Brookfield H C 1964 Questions on the human frontiers of geography. *Economic Geography* **40**: 283–303

Clay G 1980 *Close-Up: How to Read the American City*. University of Chicago Press, Chicago

Conron J 1974 *The American Landscape: A Critical Anthology of Prose and Poetry*. Oxford University Press, New York

Corner J, MacLean A S 1996 *Taking Measure Across the American Landscape*. New Haven, Yale University Press, CT

Cosgrove D E 1984 *Social Formation and Symbolic Landscape*. Croom Helm, London

Cosgrove D E, Jackson P 1987 New directions in cultural geography. *Area* **19**: 95–101

Cosgrove D E, Daniels S (eds.) 1988 *The Iconography of Landscape: Essays on the Symbolic Representation, Design, and Use of Past Environments*. Cambridge University Press, Cambridge, UK

Daniels S 1989a The political iconography of woodland in later Georgian England. In: Cosgrove D, Daniels S (eds.) *The Iconography of Landscape: Essays on the Symbolic Representation, Design, and Use of Past Environments*. Cambridge University Press, Cambridge, UK

Daniels S 1989b Marxism, culture and duplicity of landscape. In: Peet R, Thrift N (eds.) *New Models in Geography*. Unwin Hyman, London

Duncan J 1980 The superorganic in American cultural geography. *Annals of the Association of American Geographers* **70**: 181–98

Jackson J B 1952 Human all too human geography. *Landscape* **2**: 2–7

Jacobs J M 1996 *Edge of Empire: Postcolonialism and the City*. Routledge, London

Kniffen F 1965 Folk housing: key to diffusion. *Annals of the Association of American Geographers* **55**: 549–77

Meinig D W (ed.) 1979 *The Interpretation of Ordinary Landscapes: Geographical Essays*. Oxford University Press, New York

Mitchell D 1996 *The Lie of the Land: Migrant Workers and the California Landscape*. University of Minnesota Press, Minneapolis, MN

Mitchell W J T 1997 Imperial landscape. In: Mitchell W J T (ed.) *Landscape and Power*. University of Chicago Press, Chicago

Nash C 1996 Reclaiming vision: looking at landscape and the body. *Gender, Place and Culture* **3**: 149–70

Nast H 1996 Islam, gender, and slavery in West Africa circa 1500: a spatial archeology of Kano Palace, northern Nigeria. *Annals of the Association of American Geographers* **86**: 44–77

Neumann R 1995 Ways of seeing Africa: colonial recasting of African society and landscape in Serengeti national park. *Ecumene: A Journal of Environment, Culture and Meaning* **2**: 149–70

Rose G 1993 *Feminism and Geography: The Limits of Geographical Inquiry*. University of Minnesota Press, Minneapolis MN

Samuels M 1979 The biography of landscape. In: Meinig D W (ed.) *The Interpretation of Ordinary Landscapes*. Oxford University Press, New York

Sauer C O 1963 (originally published 1925) The morphology of landscape. In: Leighly J (ed.) *Land and Life*. University of California Press, Berkeley, CA

Schein R 1997 The place of landscape: a conceptual framework for interpreting an American scene. *Annals of the Association of American Geographers* **87**: 660–80

Stewart L 1995 Louisiana subjects: power, space and the slave body. *Ecumene: A Journal of Environment, Culture, Meaning* **2**: 227–46

Tuan Y F 1974 *Topophilia: A Study of Environmental Perception, Attitudes, and Values*. Prentice-Hall, Englewood Cliffs, NJ

Wagner P L, Mikesell M W 1962 *Readings in Cultural Geography*. University of Chicago Press, Chicago

Zelinsky W 1973 *The Cultural Geography of the United States*. Prentice-Hall, Englewood Cliffs, NJ

M. Domosh

Cultural Landscape in Geography

1. *The Concept*

The term, 'cultural' landscape is a scientific concept that has acquired an array of distinct meanings over time as it developed in geography and spread to other scholarly disciplines. This multiplicity of meanings has vastly enriched contemplation of the relation of humans to their earthly environment, but not without ambiguity and periodic confusion. It has entered the scientific literature of many languages, most notably Dutch (*kulturlandschap*), French (*paysage humanisé*), German (*Kulturlandschaft*), Italian (*paesaggio culturale*), Japanese (*bunka-keikan*), Russian (*landshaftovedeniye*), Spanish (*paesaje cultural*), and Swedish (*kulturlandscapet*). A cultural landscape is the successive conversion over time of the material habitat of a sedentary human society responding with growing strength and variety to the challenges of nature, the society's own needs and desires, and the historical circumstances of different regions in different times.

The cultural landscape, as part of the noösphere or human environment (in contrast to the Earth's physiosphere, containing, among other things, the natural landscape), is an objectivization of socially organized and environmentally interactive humankind. It represents, therefore, a transformation of human thought into physical matter, a vast and complex assemblage of natural materials shaped by human struggle, con-

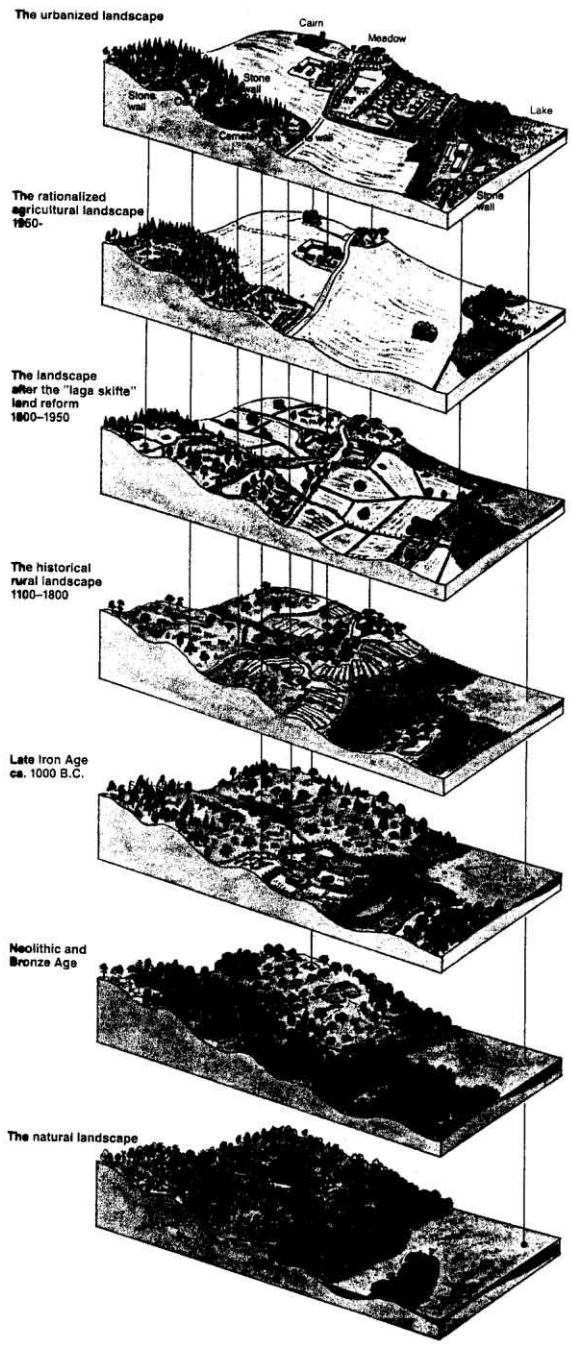

Figure 1
Cultural landscape development through time (Source: Helmfrid S (ed.) 1994 *Landscape and Settlements*, © National Atlas of Sweden, Stockholm. Nils Forshed, artist. Reproduced with permission)

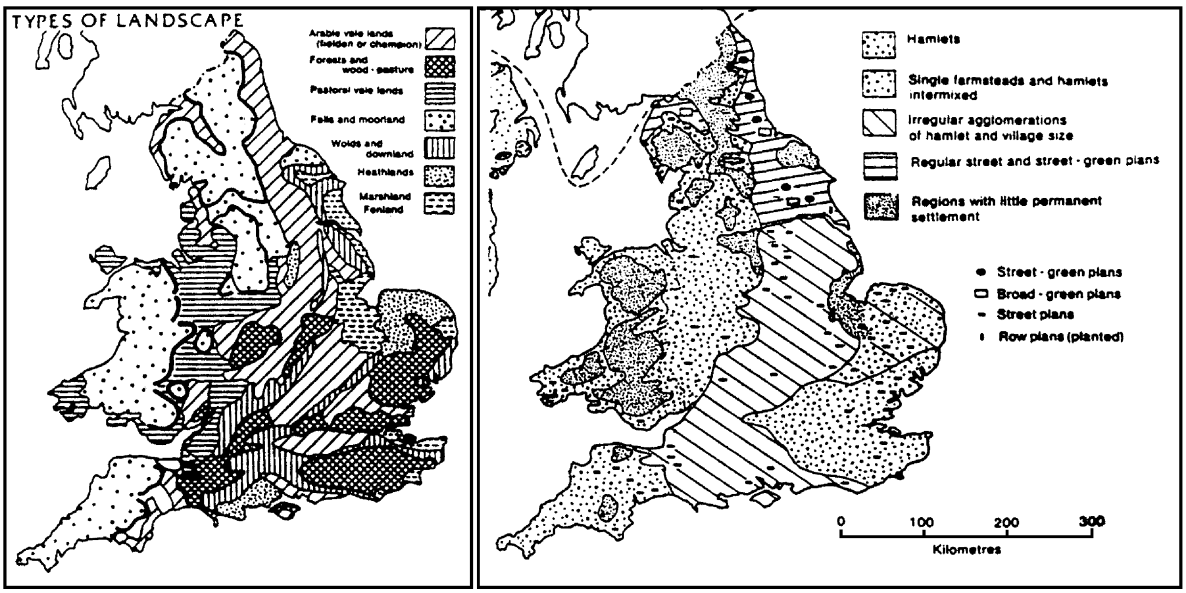

Figure 2
Geographical distribution of cultural landscape types in England and Wales (Source: based on Roberts B K 1996 *Landscapes of Settlement*, © Routledge, London. Reproduced with permission)

flict, cooperation, and ingenuity. It includes a heritage of valued material possessions passed selectively from generation to generation, representing unique societal assets.

The cultural landscape keeps its resident societies fairly rooted in place for varying periods of time, sometimes millennia, because of the continuous labor it assimilates in construction and reconstruction, as well as the collective psychological associations it inspires. Sometimes, this rootedness withstands major historical upheavals. Continuity or repetition of human occupation produces historical stratification (see Fig. 1), making the cultural landscape an invaluable multidimensioned source of cultural knowledge, aside from written documents, inviting interpretation from numerous perspectives. It consists of an extraordinary variety of human settlement forms providing for livelihood, shelter, and cultural sustenance superimposed on the land surface. These include buildings of all kinds, fields, fences, gardens, roads, bridges, factories, quarries, and a host of special land uses, plantings, and structures in rural and urban settings. Variations in age, design, social provenance, and symbolism are practically infinite, but in any given area generally fall into discernable patterns.

With the historical growth of societies, and their progressive integration in broad-scale territorial entities, specialization has created a diverse array of functional types of cultural landscapes, nested within a regional geographical matrix spanning the landmasses of the globe. Few parts of the contemporary world fail to exhibit cultural landscapes, many of them of great intensity in relation to the encompassing natural environment. Figure 2 presents the example of England and Wales, in which landscapes have been classified by two criteria: (a) general physio-economic characteristics highlighting their dominant vegetation and land use appearance, and (b) rural settlement forms, highlighting degrees of nucleation or dispersal and village layout.

The concept has affinities with other constructs: landscape is related to but not identical with nature; it is a scene, but not identical with scenery; it is around us, but not identical with environment; related to but not identical with place, and related to but not identical with region, area, or geography (Meinig 1979, pp. 2–4). For many, the interest lies in the study and explanation of the landscape itself; for others, in the way we look at it, given human ideas, attitudes, and aesthetics.

Numerous mental constructs of landscape in general have evolved over time, and the cultural landscape is implied in nearly all of them. The most important meanings can be grouped as follows: (a) the pictorial representation of an earthly scene by an artist (usually, that taken in at a glance from one point of view); (b) the sensory impression of the everyday environment; (c) the outward appearance of a district or region; (d) the natural condition of an area; (e) the cultural imprint of a region; (f) the total character of a region; (g) a

bounded territory; (h) a political-legal territorial entity (now rare, but historically significant); (i) an areal or distributional extent of a particular feature or characteristic; and (j) a 'geo-ecosystem' and an interaction system (Schmithüsen 1976, pp. 74–6, Leser et al., 1995, p. 346). In modern times the term landscape has gained popular metaphorical usage as (k) a stretch of abstract intellectual territory, as for example in the 'inner landscape' of the mind or the 'political landscape.'

2. *Historical Development*

The scientific term 'cultural landscape' came into use at the end of the nineteenth century in Germany, but its emergence rests upon a long history in which the primary term, landscape, evolved numerous meanings, many of which endured for long periods, some down to the present. The history of the specialized scientific term also mirrors shifts in understanding and popularity, and today the term enjoys wide currency in both scholarly discourse and general use.

2.1 *The Prehistory of the Scientific Concept*

The graphic depiction of landscapes for symbolic purposes is almost as old as human society, being known from the Assyrian, Egyptian, and Minoan cultures during the second millennium BC. From Greek vases in the sixth century BC to frescoes and pictorial mosaics of the Roman Empire, landscape motifs appeared with frequency. The word for landscape appeared in Chinese literature around 1000 BC by the fusion of the signs for mountain and river, and stylized landscape renderings entered Chinese art during the Han period (202 BC–220 AD).

In the western world, the word, landscape, emerged in various Germanic forms during the post-Roman period, including an ancient Frisian form 'lant-sceap,' and generally meant area, territory, tract of country, or historically constituted district. European art first depicted stylized landscape elements as scenic background for religious subjects during the thirteenth century, and in the following century for local historic figures, as with the Tuscan hill towns in Simone Martini's fresco of Guidoriccio da Fogliano painted in 1328. By the fifteenth century, Dutch, Italian, and German art of the Renaissance began to incorporate direct field observation, exemplified by the work of Jan van Eyck, Leonardo da Vinci, and Albrecht Dürer. Now came realistic portrayal of scenes from nature and social life, and even landscapes without people. In this era, just as printing made its appearance, usage of the word, landscape, shifted to the new scenic painting, which focused on the content of localities, there being plenty of synonyms to sustain the territorial meaning. The new graphic portrayals strengthened the use of the German word *Landschaft* in its meaning as picture, physiognomy, or character of an area. By 1480, European maps joined pictures as representations of regions, with da Vinci and Dürer equally skilled at both.

The age of exploration spurred rapid development in the scientific quality of maps as representations of geographical areas and their contents, particularly with the abstract addition of place names. The sailor-soldier Hans Staden in 1557 was seemingly the first to incorporate '*Landtschafft*' in a book title, referring to the contents of a region seen during his adventures in Brazil. By 1598 landscape became the 36,963rd word added to the English vocabulary (Finkenstaedt et al. 1970). During the eighteenth century geographers became seriously interested in scientific regional description, with an emphasis on physical features and the interrelationship between creative forces. However, through the works of Alexander von Humboldt (see *Hebb*, *Donald Olding* (1904 – 1985)) and Carl Ritter, there was no explicit scientific concept of landscape as such.

2.2 *The German Scientific Tradition: Systematization, Hegemony, and Near-collapse*

German geographical thought in the decades following the deaths of Humboldt and Ritter in 1859 centered on the scientific advance of physical geography grounded in systematic field observation. Ferdinand von Richthofen developed the concept of 'chorology,' understanding regions and places through their interconnectedness in the context of the earth's surface as a whole, an approach carried further by Alfred Hettner. Friedrich Ratzel developed an 'anthropogeography' based heavily on the relations of physical earth conditions and human movements. In sharp reaction to these perspectives, with their basis in method rather than subject matter and their potential over-reliance on environmental explanation, Otto Schlüter (1906) proposed a focus on *Landschaftskunde* (landscape science) as a more grounded study of the interrelations of things in particular areas. Attention to the visible features of the landscape, both natural and cultural, he felt, would give geography distinctive subject matter and logical structure. Nonobservable features would be relevant only to the extent that they helped explain the visible landscape. The approach would be evolutionary and permit study of both the *Naturlandschaft* (natural landscape) and the development of the *Kulturlandschaft* (cultural landscape) as the human modification of the original *Urlandschaft* (the wild landscape at the outset of human colonization).

Schlüter's concept was widely adopted, thanks in part to the spread of an 'objective-realistic' world view in central Europe, and by the 1930s it dominated the practice of German geography. But there was strenu-

ous debate over its limitation to evidence observable in the field, and the role of nontangible and moveable features, such as social customs and trade, and, indeed, the precise way in which the term *Landschaft*, with its multiple historical meanings, should be used. Awkwardly, the concept validated the idea that *Landschaft* is made up of manifest, visible phenomena while at the same time signifying a homogeneous area or region: thus, a region is more than a mere mental construct, it is a concrete reality. Nevertheless, Schlüter is considered to have founded modern cultural geography, and was the first to raise the landscape-forming activity of man to a methodological principle. The cultural landscape could now be examined from the standpoints of its morphology, physiology, and developmental history, just as the visible phenomena of nature in the build of the landscape as a whole. This gave human geography the same systematic order as physical geography (Waibel 1933). Siegfried Passarge wrote extensively on the topic, particularly at the world scale, which brought him international attention.

By World War II most German geographers subscribed to the scientific *Landschaft* concept whether they included or excluded nonmaterial features in their studies, exemplified by Schlüter's own master-work on the early settlement landscapes of central Europe published during the 1950s. *Kulturlandschaften* rode high in the saddle of German geography and the idea of cultural landscapes diffused elsewhere in European geography. This philosophical hegemony in Germany would persist well into the 1960s, when it fell sudden victim to a widespread reformulation of geography as the science of man's spatial organization (a paradigm shift initiated in Anglo-American geography), and suffered as a result greatly reduced popularity (Schmithüsen 1976).

The concept survived, however, particularly in the shrunken subfield of historical geography, and in recent times has made a dramatic comeback under new impulses to reunite physical and human geography through frameworks such as geoecology, dynamic human-environment systems, and landscape synthesis (Neef 1967, Hard 1973, pp. 15–181, Leser et al. 1995, pp. 34–355). In this light, the cultural landscape is seen as the highest level of integration of anthropogenic geofactors. It includes the study of cultural landscape elements, cultural areas, cultural landscape organization, cultural landscape cells, cultural landscape making and landscape change (Leser et al., 1995, pp. 332–3).

2.3 *Diffusion of the Concept*

The German scientific concept of the cultural landscape found receptive minds elsewhere in European geography during the course of the twentieth century, usually as an explicit theoretical importation grafted on to related indigenous ideas. As a result it has broadened and mutated to fit intellectual and historical conditions in different major language areas, as a few cases will illustrate.

French geography in the early twentieth century was concerned with the *genre de vie* of small regions *pays*, led by Vidal de la Blache, who saw landscape as the physical domain, set alongside 'civilisation.' But Schlüter's concept was quickly adopted by Vidal's student Jean Bruhnes, and a tradition of interpreting *paysage* emerged, through Pierre Deffontaines, Roger Dion, Marc Bloch, to Augustin Berque (1995). Berque discriminates between civilizations with and without landscapes (a distinctly French notion), the former necessitating the existence of specific vocabulary for landscape, a descriptive literature, landscape painting, and a culture of garden design. Italian geographical concern with landscape existed already in the late nineteenth century, but applied to physical scenery and aesthetic responses to it. During the 1930s the German landscape debates seeped into Italian geography and fluoresced ultimately in the writings on *paesaggio antropogeografico* of Aldo Sestini, Lucio Gambi (*paesaggio umano*), and later writers on *paesaggio culturale* such as Silvio Piccardi (1986).

German thought on the cultural landscape made a particular impression on American geography through Carl O. Sauer, who soaked it up as a student in Chicago before World War I. His (1925) essay, 'The morphology of landscape,' belatedly reflected this influence. Later Sauer would favor culture history over landscape as the central pursuit of human geography, but his many students were to extend American geographical contributions to cultural landscape research well into the 1970s. Richard Hartshorne (1939, pp. 145–74), siding with Hettner, sharply attacked this focus for the discipline, favoring instead relative location, as expressed in 'areal differentiation,' and thereby helped mute wider interest in the cultural landscape for some decades. J. B. Jackson (1997), however, a lay geographer who founded *Landscape* magazine in 1955, attracted a devoted following within and beyond geography for his unabashed passion for understanding cultural landscapes. Wilbur Zelinsky, David Lowenthal, Peirce F. Lewis, and Donald W. Meinig (1979) have since been prominent in the field.

British geographers were strikingly disinterested in the German concept of the cultural landscape during most of the twentieth century—perhaps monolingualism and two world wars took a toll—notwithstanding Robert E. Dickinson's (1939) introduction to the German literature. Vaughan Cornish, a British lay geographer, had broached the issue of landscape aesthetics earlier, but it was H. Clifford Darby's (1940) approach to historical geography which provided an intellectual home for British thinking about the cultural landscape, with his emphasis on major transformations—draining fens, clearing woodlands, reclaiming heaths. Cultural landscape research was well embedded in British geography before the historian

William G. Hoskins published his landmark *The Making of the English Landscape* (Hoskins 1955, Muir 1999).

In human geography elsewhere in Europe—Scandinavia, the Low Countries, Spain and pre-communist Eastern Europe—the cultural landscape as defined in Germany became a generally accepted, if not always core, concept by mid-century. Only in Russia, under Soviet conditions, did the concept find little favor, landscape science being defined as a physical domain to assist natural resource mobilization. In Latin America, Africa, and Asia, the cultural landscape concept also gained some currency, often under the rubric of historical geography. Australia and New Zealand followed British precedent.

Most effort in the history of the concept of the cultural landscape has been given to it as an objective, concrete reality, seeking means to define, measure, and interpret its morphology, functional structure, dynamics, and meaning as a physical reflection of variable culture across the face of the earth. That both this concept and its subject matter serve also as social constructions, with consequences for scientific objectivity as well as, more profoundly, for understanding the social meanings of landscape over time and place, has been examined intensively only since the 1970s or 1980s. This shift coincides with the rise of poststructuralism and postmodernism in the social sciences and humanities worldwide. Although the new perspectives have individual precedents, even among the German writing of the 1920s and 1930s, most of the recent scholarship has been introduced in the British and American literature. While extensive research continues unabated in many of the long-established spheres of study, suggesting its continued value, less attention tends to be paid now to the relations between human societies and the physical environment, and more to the subjective aspects of the concept, particularly the mind-states and collective visions of human actors in and interpreters of landscape making through the ages.

2.4 Landscape as Social Construction

Landscape may be regarded as 'a way of seeing,' and an ideological one, not just an image or an object; a sensibility dating to the Italian Renaissance (Cosgrove 1984, 13). Therefore, the art-forms used to represent and symbolize landscapes, as well as the artifice of the landscapes themselves, over time and space are laden with their own significance. From this vantage point landscape can be seen to reflect the unequal distribution of social and political power, where the forces of production are slyly obscured and the struggles and achievements of the less powerful hidden, patterns demonstrably starker in autocracies than democracies.

Scholarship in this vein has focused on the culturally-produced imagery of landscape and treated it, for the most part, as an essentially closed circuit of cultural invention, visualization, marketing, and social contest seen from a largely aspatial communications perspective. Whatever the ground truth of the specific earth landscapes being represented at the time are or may have been remains conspicuous by its absence and seeming irrelevance, and would be difficult anyway to include (see, for example, Daniels's 1993 study of paintings and national identity). The analytical substance here is not the concrete realm of the objective landscape, but rather the mental realm of world-view, perceptions and stereotypes, values and aesthetics, symbolism and semiotics, project and discourse (e.g., Duncan 1990).

It is in this subjective domain of landscape interpretation that British and American geographers, in particular, over the last two decades have explored interdisciplinary terrain with vigor, starting with art history. They have found common ground with scholars in other social sciences and the humanities, who were discovering simultaneously the role that 'space' plays in the problematics of culture (for further discussion of this conceptual arena, see *Cultural Landscape, in Geography*; *Cultural Landscape in Environmental Studies*).

3. Interdisciplinary Interest

Cultural landscapes are so intimately intertwined with the human societies inhabiting them as to have attracted increasingly interdisciplinary attention. But not all disciplines have put direct value on the cultural landscape concept, so the interest has been uneven, bypassing some social and behavioral sciences while including some in the physical and biological sciences and humanities.

Long before the scientific concept of cultural landscape, liberally-educated geologists and naturalists such as Archibald Geikie in Britain and George Perkins Marsh in the United States wrote about the aesthetic character and educative value of landscapes. This interest has returned in modern times with the environmental movement and finds expression in landscape ecology and historical ecology.

In the social sciences beyond geography, historians were first sensitized to landscape through Hoskins in the 1950s. Since then landscape history has become a popular pursuit in Britain and continental Europe (Hoskins 1955). In America, historical interest still emphasizes the human actors in relation to individual landscape elements (such as parks) rather than the artifact's complex organization. Landscape history has drawn in archaeologists interested in scales beyond that of the 'site,' and the British lead is now matched by similar interest in Europe and elsewhere. Anthro-

pologists have also recently discovered landscape as a 'place' context for relating community studies to broad patterns of cultural transfer. In the behavioral and medical sphere, landscape perception has been recognized as a component of environmental psychology (Porteous 1977) and mental health (Gesler 1992).

In the humanities, landscape representation has long been a fixture in art history, and more recently has been tied to landscape aesthetics and environmental values and behavior more generally (Tuan 1974, Appleton 1975, Penning-Rowsell and Lowenthal 1986). Cultural landscapes have also found a place in architectural history with the belated recognition of the importance of vernacular forms (Brunskill 1981). Landscape as a theme in literature is ancient, and affective landscape description in 'regional' literature of long standing. Recently, as metaphor, it has gained new attention in cultural studies. Linguistics and folklore have periodically contributed significantly to cultural landscape history.

4. *A Semiautonomous Field*

Over the last century cultural landscapes have spawned a burgeoning interdisciplinary scientific and scholarly literature. Their study, however, has not attained disciplinary autonomy, although it exhibits many features of an independent field: distinct scientific organizations devoted to its furtherance, dedicated publication series, public policy relevance, and in some countries an avid lay following. The less-than-autonomous status of the field reflects perhaps the extreme diversity of landscape's physical and social character, the diversity of analytical skills brought to its elucidation, and the fragmentation of managerial entities relevant to its management. Besides, disciplinary associations remain strong and vital. Geography is still the source of much, if not most, conceptual thought applied to cultural landscapes, and as a spatially-integrative discipline remains well placed to assimilate and interpret the diverse findings of landscape research on all fronts. Archaeology, history (particularly art, architectural, and landscape design history), and geoecology bring essential questions and methods to the subject.

The investigative research literature of cultural landscape studies is supported by at least a dozen international scientific journals, foremost among them being *Landscape, Landscape History, Landscape Research, Landscape Journal, Ecumene*, and a new entry *Landscapes*. Dedicated research groups exist, such as the Landscape Reseach Group, the Society for Landscape Studies, the Vernacular Architecture Forum, the Society for Commercial Archaeology, and many others. And some large-scale research programs have yielded remarkable publications on such subjects as the multinational, multivolume *Historic Towns Atlas*, the Swedish Ystad Project (*The Cultural Landscape During 6000 Years in Southern Sweden*) and an entire volume of the National Atlas of Sweden devoted exclusively to the nation's cultural landscapes (Vol. 13), and the *Atlas of the Irish Rural Landscape*.

5. *Other Uses of the Concept*

Since the 1950s studies using the cultural landscape concept have gained relevance beyond academia in the arena of public policy, education, and business. Detailed survey-driven documentation studies, consisting of cultural resource inventories (organized in GIS/geographic information systems), along with regional interpretation of the cultural landscape features past and present support two key social-environmental enterprises: the design professions such as architecture (building traditions) and landscape architecture (land planning and design), and activities that come under the rubric of *landscape management*. These include environmental conservation and protection (national parks, urban open space), historic preservation (cultural parks, open air museums, ornamental gardens, historic districts), and general regional planning (especially of transportation, land use, recreation and tourism). For purposes such as these techniques of landscape assessment (or appraisal, evaluation), including land and structure classification, surveys of landscape aesthetics (measuring visual complexity, 'quality,' and preferences), and landscape impact analysis—all ultimately grounded in theories of cultural landscape history, dynamics, and perception—provide wide-ranging applications of this fundamental, integrating concept in the social and behavioral sciences.

See also: Cities: Cultural Types; Local Economic Development; Planning Ethics; Prison Life, Sociology of; Social Protests, History of

Bibliography

Appleton J H 1975 *The Experience of Landscape*. Wiley, New York

Berque A 1995 *Les raisons du paysage: de la Chine antique aux environnements de synthèse* [The Meanings of Landscape: From Ancient China to Synthetic Environments]. Hazan, Paris

Brunskill R W 1981 *Traditional Buildings of Britain*. Gollancz, London

Cosgrove D 1984 *Social Formation and Symbolic Landscape*. Croom Helm, London

Daniels S 1993 *Fields of Vision: Landscape Imagery and National Identity in England and the United States*. Princeton University Press, Princeton, NJ

Darby H C 1940 *The Draining of the Fens*. Cambridge University Press, Cambridge, UK

Dickinson R E 1939 Landscape and society. *Scottish Geographical Magazine* **55**: 1–15

Duncan J S 1990 *The City as Text: The Politics of Landscape Interpretation in the Kandyan Kingdom*. Cambridge University Press, Cambridge, UK
Finkenstaedt T, Leiss E, Wolff D 1970 *A Chronological English Dictionary*. C. Winter, Heidelberg, Germany
Gesler W M 1992 Therapeutic landscapes: medical issues in light of the new cultural geography. *Social Science & Medicine* **34**: 735–746
Hard G 1973 *Die Geographie: eine wissenschaftstheoretische Einführung* [*Geography: A Conceptual Introduction*]. de Gruyter, Berlin
Hartshorne R 1939 *The Nature of Geography*. Association of American Geographers, Lancaster, PA
Hoskins W G 1955 *The Making of the English Landscape*. Hodder & Stoughton, London
Jackson J B 1997 *Landscape in Sight: Looking at America*. Yale University Press, New Haven, CT
Leser H, Haas H D, Mosimann T, Paesler R 1995 *Dierke-Wörterbuch der allgemeinen Geographie*, Band 1 A–M [*Dierke Dictionary of Systematic Geography*, Vol. 1: A–M]. Westermann, Braunschweig
Meinig D W (ed.) 1979 *The Interpretation of Ordinary Landscapes: Geographical Essays*. Oxford University Press, New York
Muir R 1999 *Approaches to Landscape*. Macmillan, Basingstoke
Neef E 1967 *Die theoretischen Grundlagen der Landschaftslehre* [*The Theoretical Basis of Regional Geography*]. Haack, Gotha/Leipzig
Penning-Rowsell E C, Lowenthal D (eds.) 1986 *Landscape Meanings and Values*. Allen and Unwin, London
Piccardi S 1986 *Il Paesaggio Culturale* [The Cultural Landscape]. Pàtron Editore, Bologna
Porteous J D 1977 *Environment and Behavior: Planning and Everyday Life*. Addison-Wesley, Reading, MA
Sauer C O 1925 The morphology of landscape. *University of California Publications in Geography* **2**: 19–53
Schlüter O 1906 *Die Ziele der Geographie des Menschen* [The Goals of the Geography of Man]. R. Oldenbourg, Berlin
Schmithüsen J 1976 *Allgemeine Geosynergetik: Grundlagen der Landschaftskunde* [General Geosynergetics: Principles of Regional Geography]. de Gruyter, Berlin
Tuan Y F 1974 *Topophilia: A Study of Environmental Perception, Attitudes and Values* Prentice-Hall, Englewood Cliffs, NJ
Waibel L 1933 Was verstehen wir unter Landschaftskunde? [What do we understand by regional geography?]. *Geographische Anzeiger* **34**: 197–207

M. P. Conzen

Cultural Policy

The study of cultural policy is concerned with the instruments (legal, administrative, economic) through which governments provide, regulate, and manage cultural resources and the uses to which they are put. The objectives that are pursued by these means can be divided into three broad, but overlapping, categories—the symbolic, the social, and the economic. In exploring these objectives and the political rationales on which they rest, this article examines the history of their development as concerns of government, and the varying interpretations they have received in different political regimes and philosophies.

1. Cultural Policy and State Symbolism

The cultural policies of contemporary governments are most evidently continuous with those of earlier political systems when supporting cultural activities that symbolize the virtues of a particular nation, people, or political system. The most obvious precursor for these concerns consists in the use of the arts as instruments of state policy associated with the absolutist monarchies of early modern Europe. These were characterized by two contrasting orientations. The first aimed to weaken the nobility by involving its members in a socially demanding and economically debilitating cultural life centered on the court. The second aimed to symbolize the principles underlying the political order through a public theater of power (statues, processions, palaces) which magnified the king's potency so as to incline his subjects to obedience. Culture, here, symbolizes sovereign power with a view to enhancing its effects.

If this connection between sovereign power, culture, and the arts of governing was most fully developed in France, it was given a new interpretation during the French Revolution. This consisted in the role that was developed for the arts, and culture more generally, of symbolizing the more abstract form of sovereignty that was now vested in a democratic citizenry (McClellan 1994). In this new conception—evident in the policies developed for the Louvre as much as in the debates over language policies, public monuments, and architecture—culture served as the means through which a democratic citizenry made its own power as a sovereign people manifest to itself. But it also allowed that citizenry to exercise a direction over itself through the ideals—of, for example, heroism and sacrifice—the citizen was exhorted to emulate.

These relations between the arts, culture, and sovereign power have undergone a number of mutations in the subsequent development of the rationales underlying state support for specific kinds of cultural provision. The logic underlying continuing state patronage of the art forms that had been most closely linked to court society—ballet, theatre, opera, and the art museum—remains one of forging particular forms of elite sociability, bonding the powerful through shared rituals and symbols. The evidence in all societies where relevant studies have been conducted shows that participation in these cultural institutions is mostly limited to elite social strata (Bourdieu and Darbel 1991). Their role in symbolizing a cultural distinction between such strata and the generality of the population plays an important part in the cultural dynamics of class societies. This is, therefore, an aspect of government involvement in the cultural sphere that

is often regarded as in conflict with the democratic principle of equal cultural entitlements for all citizens which, in the postwar period, has become a guiding principle for cultural policy development.

Government support for more spectacular cultural forms, as a means of symbolizing the power of a particular people or nation, has remained an enduring aspect of the cultural policies of modern polities. The festivals of the French Revolution constituted a symbolic use of public space which, in making a citizenry visible to itself, disputed the earlier symbolism of royal processions. The subsequent development of participatory forms of public spectacle—international exhibitions, national celebrations, and spectator sports—has yielded a range of contexts in which a people is made visible to itself and its virtues celebrated in ways which pit them in competition with other nations, races (the prime cases being Nazi Germany or fascist Italy), and political regimes (especially during the Cold War).

Important questions of publicness concerning the forms in which the identity and qualities of a people, citizenry, or nation are to be represented are at stake in these festival, celebratory, and commemorative forms. The relations between the policy structures and processes through which these matters are resolved and the perspectives of political constituencies organized in terms of relations of race, class, gender, or religion have played a crucial role here in redefining both who gets to be counted as citizens and the exemplary virtues that are held up as models for their emulation. A continuing history connects those who criticized the revolutionary festivals for a public iconography of womanhood which portrayed women as mere helpmates to a republican brotherhood of men to late nineteenth-century feminist criticisms of the male-centered norms of publicness associated with international exhibitions and museums. And there is a continuing history connecting both of these to contemporary debates centered on the adequacy of the forms of representativeness—whether in terms of gender, race, or religion—that are to be aimed for in major forms of national commemoration or celebration. The debates occasioned by Georges Pompidou's *grands projets*, by Aboriginal engagements with Australia's 1988 bicentennial celebrations, or by Britain's Millennium Dome are all cases in point. As such, they illustrate how ways of making power manifest inherited from an earlier political regime based on a singular source of authority are imbued with a more contested content as a consequence of their relations to a civil society that is plural and divided.

2. *Acting on the Social*

As well as investing in culture to symbolize power, modern governments have also invested in culture as itself a power that is capable of bringing about changes in conduct and ways of life. These are not entirely different functions. Debates concerning how best to symbolize a citizenry—how best to present to them their collective power and virtue—have been concerned centrally with reinforcing some kinds of conduct and changing others. The early modern period, however, also saw the development of an orientation toward culture that, although it had precedents in earlier forms of pastoral power, represented a new departure so far as secular forms of cultural administration were concerned. This consisted in the role that was accorded culture in the *Polizeiwissenschaft* (science of police) of the seventeenth and eighteenth centuries. This aimed to bring a new order to the social body in place of that which had characterized the relationships between different estates in the old social order with its clear rules regarding the forms of behavior, dress, duty, and demeanor appropriate to particular stations in life. It thus concerned itself with ways of life both negatively and positively: negatively in prohibiting particular kinds of behavior (for example, public drunkenness), and positively by providing the cultural means through which problematic forms of conduct might be transformed through programs of self-improvement. An early nineteenth-century Prussian regulation puts the point well:

> In their capacity as 'Polizei' authorities of the land the governments are charged with the care and well being of our loyal subjects, in positive as well as negative aspects … . Governments therefore have supervisory powers in matters of public improvement, education, and culture (Knemeyer 1980).

Within the emerging class dynamics of nineteenth-century industrial societies, such conceptions, when translated into the programs of liberal government, typically split the field of culture into two. On the one side, the culture of the working classes presented itself as ways of life that needed to be acted on, often with varied aims in view (combating political sedition, inducing habits of thrift and sobriety), as a means of re-forming the social. On the other, the rational and improving culture that had once formed a part of the bourgeois public sphere (Habermas 1989) appeared in a new historical guise as an instrument of government, providing the means whereby individuals might be led to transform their own conduct (Minihan 1977).

This was the general impetus behind the nineteenth-century development of public libraries, museums, art galleries, concert halls, and public schooling through which, in Europe, governments undertook to make cultural provision for the population as a whole. It was, however, the needs and aspirations of government, rather than any sense of democratic cultural rights of the kind that had been asserted during the French Revolution, that supplied the rationale for this provision. Government action in the cultural sphere was based on the expectation that the improving influence of a rational, enlightening, and elevating

culture would lead, through the accumulation of thousands of individual acts of self reform, to a transformation of working class ways of life and, consequently, a re-ordered social body. The guiding principle of state cultural provision in this period accordingly was focused mainly on leading 'the people' away from cultural activities and traditions that were popular with subordinate social strata and towards greater participation in elite cultural forms that had received the official sanction of state approval.

It is not until the 1930s that questions concerning the responsibility of governments for maintaining and promoting popular pastimes, traditions, and ways of life came centrally to the fore. The contexts in which this occurred were often nationally specific: the development, in the context of the New Deal in the United States, of community-based cultural programs (Harris 1995); the programs of leisure-time organization (*dopolavoro*) of fascist Italy; and the role of the Popular Front in France in opening up a new politics of popular leisure (Looseley 1995). These tendencies were continued in the Second World War which, if it was 'the people's war,' also brought the question of the people's culture—and its role in maintaining their morale—to the fore in new ways that made it increasingly imperative for democratic governments both to know about popular ways of life and to nourish them so that they might thrive. This trajectory toward more democratic understandings of the concerns of cultural policy has been continued throughout the postwar period, although by no means consistently or without reversals. Indeed, exceptions to this tendency have often been institutionalized as the administrative apparatuses established to develop government support for art, culture, and leisure have reflected the lingering influence of older elite values. The histories of the Arts Council in Britain and the Australia Council are cases in point. In both cases, the criteria of excellence that have guided their allocations of funds have been at odds with their obligation to meet the cultural needs of all sections of society (Hawkins 1993).

The predominating tendency of the period since the 1960s has been to call into question these socially exclusive conceptions in favor of more democratic conceptions of the range of cultural tastes and values governments should support (see *Cultural Policy: Outsider Art*). While these criticisms initially centered on questions of social class, feminist criticisms brought questions of gender to the fore in the 1970s while, since the 1980s, questions of race and ethnicity have tended to displace the importance that was earlier accorded social class. The responses to these pressures have varied. The cultural policies developed by Jack Lang as a member of Francois Mitterand's governments provided an influential model in the 1980s in view of the accent they placed on the values of pluralism and democracy (Looseley 1995). The cultural policies developed in Australia under the Labor governments of Bob Hawke and Paul Keating were also innovative in developing cultural policies that recognized the diverse needs of a society with a history of culturally varied migration and an indigenous population with highly distinctive cultural needs and traditions.

The recognition and promotion of cultural diversity—now the major international leitmotif of cultural policy discourse, and one actively promoted by UNESCO (World Commission on Culture and Development 1995)—is not, however, simply a matter of recognizing cultural rights. On the contrary, programs of cultural diversity are now central components in the governing strategies of modern polities. They supply a means for acting on the social—by celebrating cultural diversity, and marshalling a variety of legal and moral sanctions against racist forms of conduct—that is widely accepted as crucially important for managing the relationships between ways of life in contexts where the mix of peoples is increasingly diverse (Gunew and Rizvi 1994).

However, these tendencies are neither universal nor irreversible. The policies and politics of ethnic cleansing are a testimony to the continuing influence of monocultural approaches to managing the relationships between peoples with different ways of life that had characterized earlier forms of racism. The new right's criticisms of political correctness in the United States have also challenged cultural diversity policies. The links that had earlier connected cultural access policies to social justice principles are now also weaker in Europe where policies based on the logic of social inclusion often fail to acknowledge the relevance of social class to either social or cultural distributive issues.

3. *Culture and Economic Development*

Whatever the importance of the social and civic issues associated with cultural policies, these are now eclipsed by the economic significance of the cultural and media industries. This is not entirely new. Tourism has played a role in the calculations that have guided investments in large cultural infrastructures since the eighteenth-century spa movement and the museum and exhibition boom of the nineteenth century. The industrial issues at stake in the development of new communications technologies have also been clear from the outset: witness the variety of legal and technical barriers developed in Europe to combat the dominance of American cinema.

However, the economic aspects of cultural policy have assumed an altogether unprecedented significance in the closing decades of the twentieth century. There are a number of reasons for this. Some have to do with the demise of more traditional industrial sectors and the increasing importance of cultural industry development to the employment strategies of

once prosperous industrial regions. Some have to do with the changing relations of work and leisure, resulting in increased effective demand for cultural activities. This is reflected in the increasing role accorded cultural tourism as a means of attracting visitors to particular tourist destinations. But the key driving force over the 1980s and 1990s has consisted in the intersections of a number of technological developments that have revolutionized communications in ways that have affected profoundly all aspects of the cultural and media sectors of advanced societies (Collins and Murroni 1996). The role of computing and the Internet in the development of the 'information economy,' and the convergence of these with new developments in television and telecommunications, have been especially important.

The cumulative effect of these developments has been to forge a stronger connection between cultural policies and economic policies. This is especially evident in contexts where local or national cultural industries are threatened by the more global pattern of cultural flows associated with the new international media order. In France, Australia, and Canada the case for government investment in cultural industries— for example, film and popular music—has been urged as a means of retaining a viable base for their future expansion in the face of American and British imports. The same rationale lies behind policies that restrict the amount of foreign content that can be broadcast on radio or television, or regulations limiting foreign ownership of media. However, it is also clear that, in these cases, the considerations at stake are not only economic. In view of their role in circulating not merely economic value but also cultural values and meanings, viable cultural and media industries are seen as necessary for maintaining and developing nationally, regionally, or locally distinctive heritages and identities. In other contexts, however, economic and cultural considerations may be in tension with one another and in need of mediation. The requirement that tourism policies should be 'culturally sustainable'—that is, not be promoted to an extent that will detract from the quality of the specific culture or way of life of a particular locality—is a case in point.

Here too, then, questions of ways of life and the relations between them are at stake. The same is true of the increased interest in questions of intellectual property associated with the new media. A part of the concern here has been with the extent to which intellectual property regimes developed in an earlier media environment can provide for sufficient security of title to sustain the viability of investment in cultural forms whose reproducibility opens up the prospects of their free circulation via the Internet. A second area of distinctive concern, however, has been with the intellectual property protocols that are needed to mediate the relations between specific cultural producers and the marketplace in ways that do not prejudice the cultural traditions of the former. The debates that have accompanied the use of the Internet to market the art of the indigenous peoples of Australia, Canada, and New Zealand are a case in point, comprising a subset of a broader group of policy questions concerning the relations between new media and indigenous cultures (Michaels 1994).

Considerations of this kind, however, do not address what is perhaps the most important issue associated with the new media order: the increasing division between 'information rich' and 'information poor' societies. The fact that there are few, if any, international mechanisms that are capable of addressing this issue underlines the extent to which cultural policies remain, in the main, the instruments of national governments.

4. *Culture, State, and Civil Society*

The development, over the eighteenth and nineteenth centuries, of these different kinds of governmental involvement in the cultural sphere has occasioned debate concerning the extent to which such involvement is justified and the means through which it is most appropriately exercised. These questions are rehearsed, in the specialist literature, in assessments of the respective virtues of the different models of support for culture that have been developed in connection with different political regimes and philosophies. These range from the view, prevalent in the United States and Japan, that the role of government should be purely facilitative, regulating cultural and media markets by legal and fiscal means to provide the conditions in which the arts and culture might flourish, to the assumption of direct responsibility for funding and directing cultural activities associated with communist regimes.

The main point at issue in these debates concerns the relations between state and civil society that different kinds of state action in the cultural sphere entail. Within the formulations of the science of police, the allocation of a responsibility to the state for the cultural well being and development of the population formed part of a more general program of government in which the state was to become involved directly in the administration of all aspects of life. There is a significant continuity between these formulations and the approaches to the administration of culture associated with communist regimes of the mid-to-late twentieth century in their aspiration to replace market mechanisms by means of a cultural plan developed by state functionaries in the light of officially-sanctioned criteria of aesthetic and political excellence. This construction of the relations of culture and government entails a denial of the autonomy of civil society as a realm separate from the state. Instead, in the cultural policies of communist regimes, the institutions of civil society, including those of culture, are inte-

grated into the state in order that the citizen might be molded directly in accordance with cultural ideals endorsed and supported by the state. This integration of the institutions of civil society and of the citizen into the state was also a characteristic of fascist cultural policies: the *dopolavoro* was a clear assault on the autonomy of Italian working-class culture and an attempt to integrate it into the state where it could be subjected to more direct forms of regulation (De Grazia 1981).

In western democracies, the role of liberalism and its criticisms of the totalizing aspirations of the earlier science of police have given rise to a different set of relations between state, culture, and civil society. Within liberal conceptions of government, the autonomy of civil society is to be not so much respected as constructed, organized as a zone of rights and freedoms on which the activity of government must not trespass. This has meant that cultural policies that have aimed to bring about specific changes of conduct have done so in ways which have accepted the independence of civil society. This has entailed fashioning cultural resources into means of acting on the social that work indirectly, through the agency of the citizen's voluntary activity rather than via state imposition, and at a distance in the sense of accepting the autonomy of the social as a realm of behaviors and relationships that exist independently of the state.

These principles of liberal government have been a constant feature of the cultural policies of the United States, Britain, Canada, France, and Australia. However, their implementation has been, at different historical moments, either muted or modified in view of the expansion of the state's functions associated with its role in the provision of social welfare and, since the Second World War, corporatist strategies of economic management. They have, however, come into the ascendancy once again with the phase of neoliberalism inaugurated by the Thatcher and Reagan administrations and subsequently greatly assisted by those accounts of globalization which suggest that the global flows of both culture and capital now exceed the regulatory capacities of national governments. The process of rolling the state back out of the lives of its citizens has resulted in a tendency to reduce the levels of direct government funding and support for specific forms of culture in favor of the operation of private cultural markets and corporate or philanthropic forms of sponsorship. In these circumstances,'audience development' has become the policy buzzword of the 1990s. At the same time, where government funding remains a significant factor, new relations of competition have been fostered to make the institutions of public culture more responsive to the effects of market forces. This has been most notably true of the restructuring of the British broadcasting industry over the period since the 1990 Broadcasting Act (Goodwin 1998). Taking a broader perspective, however, perhaps the most significant changes in the relations of state, culture and civil society at the close of the twentieth century are those evident in China where new forms of cultural consumption associated with the development of new cultural markets now increasingly take the place of state-provided culture.

However, if the policy catch phrase of the 1980s and 1990s has been deregulation, this tendency has not gone unchallenged. Indeed, the consequences of deregulating the culture and media industries have often been sufficiently problematic to lead to more-or-less instant demands for their re-regulation. There is, however, no escaping the historical shift that is entailed in a policy discourse that now sees the state's cultural role primarily in terms of the exercise of its regulatory capacities rather than as itself directly a significant agent of cultural production.

See also: Art, Sociology of; Cultural Expression and Action; Cultural Landscape in Geography; Cultural Resource Management (CRM): Conservation of Cultural Heritage; Culture, Sociology of; Mass Media, Representations in; Public Opinion: Microsociological Aspects; Regulation: Sexual Behavior

Bibliography

Bourdieu P, Darbel A 1991 *The Love of Art: European Art Museums and their Public*. Polity Press, Cambridge, MA

Collins R, Murroni C 1996 *New Media, New Policies: Media and Communications Strategies for the Future*. Polity Press, Cambridge, MA

De Grazia V 1981 *The Culture of Consent: Mass Organization of Leisure in Fascist Italy*. Cambridge University Press, Cambridge, UK

Goodwin P 1998 *British Television under the Tories. Broadcasting Policy: 1979–1997*. British Film Institute, London

Gunew S, Rizvi F (eds.) 1994 *Culture, Difference and the Arts*. Allen & Unwin, St. Leonards, NSW, Australia

Habermas J 1989 *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. MIT Press, Cambridge, MA

Harris J 1995 *Federal Art and National Culture: The Politics of Identity in New Deal America*. Cambridge University Press, Cambridge, UK

Hawkins G 1993 *From Nimbin to Mardi Gras: Constructing Community Arts*. Allen & Unwin, St. Leonards, NSW, Australia

Knemeyer F L 1980 Polizei. *Economy and Society* **9** (2): 172–96

Looseley D L 1995 *The Politics of Fun: Cultural Policy and Debate in Contemporary France*. Berg, Oxford, UK

McClellan A 1994 *Inventing the Louvre: Art, Politics and the Origins of the Modern Museum in Eighteenth-century Paris*. Cambridge University Press, Cambridge, UK

Michaels E 1994 *Bad Aboriginal Art: Traditions, Media and Technological Horizons*. Allen & Unwin, Sydney, Australia

Minihan J 1977 *The Nationalization of Culture: The Development of State Subsidies to the Arts in Great Britain*. Hamilton, London
World Commission on Culture and Development 1995 *Our Creative Diversity*. UNESCO, Paris

T. Bennett

Cultural Policy: Outsider Art

Within the broad framework of public policy, government support of culture varies from one nation to another. In authoritarian or totalitarian regimes culture is usually a tightly controlled instrument of ideology, in liberal democracies expenditures of public funds are justified on other grounds, for example, the role of culture in strengthening civil society. Until recently, cultural policy has traditionally been concerned with art forms whose standing was uncontested. The art genre that has come to be known as 'outsider art' is of particular interest because of its divergences from either of these historic patterns. It serves as the focus of this article.

1. Introduction

1.1 Construction of the Outsider Art Genre

(a) The term 'outsider art' has been used to signify the art of the 'insane' or more generally the creations of certain kinds of untrained artists. Its main components, noted from the late nineteenth century on, consist of works by mental patients, children, and 'primitive' peoples from cultures outside of Western civilization. Recently, the genre has been enlarged by the cultural creations of what are usually thought of as self-taught creators considered marginal to or marginalized by their society: naïve artists, folk craftsmen or women, urban or rural isolates, hobbyists, the homeless, prison inmates, institutionalized elderly (Zolberg and Cherbo 1997).

(b) In the last quarter of the twentieth century, outsider works began to gain fame, to the monetary and symbolic benefit of their owners. Works are exhibited in museums, featured in art magazines, and analyzed in books that detail and assess their quality. In the process, a 'tradition' of outsider art has emerged centered upon certain canonic works that sell for extraordinary prices. In contrast to what are classified as professional artists, outsider artists are generally believed to be unaware of their artistry, not promoting their own career, but spontaneously following their creative impulse (Becker 1982). They become known largely through the discovery and *marketing* practices of art dealers, gallery owners, art critics, scholars, museums, and government agencies (Ardery 1997).

(c) At one extreme, their creators inhabit a world of isolated naïves, whose works take shape as the result of personal visions (Becker 1982). At the opposite, as inmates of total institutions (Goffman 1961), they gain access to the possibility of creation when it is offered by their keepers. It is the latter who select what merits public display. A third class of canonical outsider art is 'primitive' art, a form whose meanings derive from completely different societal and cultural traditions. It was through convenient misunderstandings of their origin and significance that primitive art was lumped together with the works of the 'insane,' adult naïves, and children (Zolberg 1997, 53–72).

1.2 State and Market in the Making of the Genre

(a) Outsider art flourishes or languishes depending upon the political regime and the art market in which it comes to light. The complexity both of art and of its context makes it necessary to clarify the nature of their interactions by tracing the genre's institutional and structural foundations. The institutions and policies of governments play a part, but government policies, vital as they are for fostering or inhibiting culture, can seldom determine cultural outcomes. Their importance lies in the fact that they set the conditions under which their creation and dissemination may take place. In liberal states where commercial processes are permitted and fostered, the power of the art market is at least as decisive as is governmental policy *per se* in providing gatekeepers, agencies of reward, recognition and legitimation.

(b) The principal question is whether, how, and with what consequences the public sector came to enter into the domain of outsider art. Equally important is the role of art market structures and processes. Although this combination of domains has not been directly drawn together into a unified analysis, it appears that together these gatekeeper agencies and groups are engaged in the discovery as well as construction of new forms of outsider art.

2. The Social Construction of Outsider Art

2.1 An Aesthetic Response to Modernization

(a) Interest in outsider art is an aspect of attempts by observers and scholars to grasp the meaning and impact of modernizing trends in society. Theorists who laid the groundwork for understanding the phenomenon came at it from various perspectives and disciplines, though, with the exceptions of Max Weber (Gerth and Wright 1946) and Georg Simmel (Wolff 1950), they generally did not refer explicitly to the arts.

One of the most influential formulations was that of Ferdinand Toennies, whose analysis of the growing dominance of modern society over what he perceived as a declining communitarian rural life world, summarizes a widespread unease prevalent among many nineteenth-century intellectuals (Toennies 1963). Cultural opinion leaders such as William Morris and his followers tried to overcome what they saw as the cold ugliness of industrialization and its products by reverting to medieval inspiration for design, and work organization based on handcraft. Forms and genres that they associated with folk culture appealed to their romantic sensibilities and, for some, their socialist ideas.

(b) An imagined peasantry and its way of life was not the only victim to the seemingly inexorable march of modernization. Integral to much nineteenth-century longing was a search for 'authenticity' in contrast to the 'artificiality' of urban civilization. Seemingly embodied in artistic creations of children and of asylum inmates, authenticity was sought by therapeutically oriented scholars of the 1920s. Hans Prinzhorn, for example, a leading proponent of the artistry of the mentally ill (Prinzhorn 1972) was not alone in his appreciation of their gifts. Walter Morgenthaler, a physician in a Swiss institution, publicized the achievements of one of the first major outsider artists to be 'discovered,' his patient, Adolph Wölfli (Morgenthaler 1992).

(c) The third principal form of outsider art, works by non-Western peoples, has a very different source and trajectory. It entered European consciousness through nineteenth- and twentieth-century colonialist imperialism, when much of it was rarely recognized as art, but rather as superstitious fetishes or indications of the innate childishness of primitive people. These interpretations provided additional justification for European domination, since they intersected with conventional stereotypes, and rationalized the claim that colonialism would elevate the subjects to a civilized state. Primitive artistry was 'discovered' largely in the early twentieth century when, spearheaded by *avant garde* artists (Cubists, Fauvists, Expressionists), these works were re-interpreted in appropriate aesthetic terms (Vogel 1988). It was not long before other artists, art historians, and critics turned their attention to this blossoming genre (Zolberg 1997, 53–72).

2.2 Intellectual and Scholarly Perspectives on Outsider Art

(a) English critic Roger Cardinal (1979), French *avant garde* artist Jean Dubuffet (1988), art historian and museum curator Robert Goldwater (1986), historians and critics John MacGregor (1989), Michel Thévoz (1976) and many others have provided a multifaceted scholarship on these works. Anthropologists, psychologists, and sociologists have contributed their understandings: Howard Becker (1982), Sander Gilman (1982), James Clifford (1988), Anne Bowler (1997), Vera Zolberg (1997), Julia Ardery (1997). Several of them perceive affinities between outsider art and genres and works of 'maverick' artists (Becker 1982) of the turn of the century *avant garde* movements (Bowler 1997, Tuchman and Eliel 1992, Hall and Metcalf 1994).

(b) While interest in outsider genres parallels concerns about modernization, technology, and anxieties about ecology, free-floating concerns are insufficient to account for the burgeoning of interest in these genres. Without the creation of aesthetic institutions dedicated to state building, and their transformation with the rise of art markets, it is difficult to imagine the emergence of outsider art as an admired genre.

2.3 Structural Context for Recognition

(a) The modern system of public institutions and agencies (the academic system), which had been enacted by nation states, became the authoritative gatekeeper in establishing and maintaining artistic status hierarchies. Exclusionary and narrow in their definition, academic establishments were eventually sufficiently shaken by challenges from successive waves of *avant garde* art movements (Poggioli 1971) to be obliged to open themselves to new styles. This came about largely as they faced competition from the private sector of an emergent, free-wheeling, commercial dealer system (see, for example, *Fine Arts*).

(b) Psychotherapy converged with the rise of *avant garde* art movements. Resulting imagery was interpreted as spontaneous expressions of outsiders who were seeking a visionary experience. To these unschooled, apparently spontaneous creations of institutionalized mental patients and children, were added tribal arts of Africa, the southern Pacific (Goldwater 1986), and rustic Europe.

(c) By the latter half of the twentieth century, as the 'insider' art worlds of academies and of dealer systems became increasingly open to stylistic and genre innovations, official and conventional art categories were transformed. This was true as well for unconventional *avant garde* artists, whose claim to standing as arbiters of fine art was challenged by even more unconventional expression that rejected pure aestheticism and linked their creation to other domains. Barriers between high and low art, art and politics, and art and religious rite, art and emotional expression, art and life itself are repeatedly breached. Art historians, aestheticians, social scientists, and policy makers face complex challenges when they try to delineate what is Art, what it includes or excludes, whether and how it is to be evaluated, and the relative importance to assign to different genres (DiMaggio 1987). These uncertainties are directly implicated in controversies

about how and whether governments should provide official support for more recent kinds of artistic creation, and their dissemination to larger audiences.

3. *Public Policy and the Arts*

3.1 *Pitfalls of Public Culture*

(a) In recognizing certain forms of culture as symbolically significant, the state confers official approbation and respect. France very early established a sphere of official cultural policy for the glorification of the state (Mesnard 1974). The French example influenced much of western Europe and left a durable heritage on its former colonies. Inevitably, officially recognized cultural forms were viewed as hierarchically superior to those receiving support from other sources, whether religious, provincial, or commercial (see, for example, *Fine Arts*). Integration into its ranks gains artists public or private patronage rather than leaving them dependent on commercial forces.

(b) While most liberal national governments recognize and support some art forms, rarely is the institutional structure of support as centralized as in authoritarian or despotic regimes. Nazi Germany suppressed the principal forms of outsider art and their creators. Under the dominance of their leaders' racist beliefs and eugenicist pseudoscience, the state's agencies used their apparent kinship with the stylistic and philosophical orientations of *avant garde* forms (Expressionism) to justify official suppression of both (Barron 1991). Labeling the art of vanguard artists and the art of the mentally ill, the naïve, and African works as 'degenerate,' the Nazi regime persecuted and frequently murdered artists, and sold their works abroad or simply destroyed it.

(c) The relationship of outsider art to contemporary political agendas in most other societies is considerably more benign. Thus, in relatively liberal regimes outsider art forms have come to be associated with heart-warming outcomes of social work or psychotherapy; some have come to be regarded as having aesthetic value in their own right. It is in this process that they have become incorporated into the gallery system.

3.2 *Cultural Policy US Style*

(a) European nations, with their feudal and monarchical traditions, have emulated the French structures of official academic art, though rarely in as centralized a form (Corvisier 1978). Even in a nation as divergent from European traditions and, in particular, from state centralization, as the USA, France's reputation as patron of the arts was admired by artists. This is understandable, since in contrast to France, nineteenth century US artists found themselves dependent upon a market for their livelihood, or at best on insufficiently appreciative patrons (Morgan 1978, p. 71). For many US politicians, however, far from being the touchstone, France was an anti-model. Not only was an official culture incompatible with the distrust of the state that pervaded US political thought (Kammen 2000), but the arts were thought of as a private pleasure in whose support the state had no right to disburse public monies (Meyer 1979).

(b) It is not surprising that cultural subsidies are a relatively new development in the USA. Prior to the Great Depression of the 1930s, it had been confined, as in other nations, to visual and aural forms of relatively routine expressions of national sovereignty: designs for currency, the official seal, military uniforms, postage stamps, architecture of government buildings and monuments, and musical compositions for military bands and a national anthem. For the most part, designs were made by civil servants in the course of their duties or, later, by commissions to artists or architects as needed. The spurt of activity generated by New Deal art programs was unprecedented. Not surprisingly, they were justified not so much on the ideal of artistic creation, but on the practical ground of generating jobs for unemployed artists.

(c) With a preference for fairly popular styles to appeal to an unsophisticated public, most New Deal artists, composers, and writers worked largely in the most recognizable fine art forms: painting (Marling 1982), and background music for documentary films. In addition, writers were commissioned to produce regional guide books, in part to enhance the appeal of distant or isolated US regions for prospective tourists. The New Deal supported scholars and serious amateurs, who collected and recorded music in rural areas among agricultural people, fishermen, small town churchgoers, and prisoners in chain gangs. Beside research in outlying areas, the New Deal employed musicians and artists to teach in schools and settlement houses, as well as adapting some art forms to therapeutic ends (McDonald 1969). The preservation of folk music, arts, and crafts echoed the search for authenticity by artists, writers, and musicians. Perhaps inadvertently, by preserving folk music, the federal government helped nurture one of the classic genres that later came to be incorporated into the domain of outsider art.

(d) Recognizing the connections between certain art forms and the groups which are their source, the Natiojnal Endowment for the Arts (NEA) seeks out originality in the relatively uncontroversial domain suitable to school curricula: 'folk arts, storytelling, and historic preservation' (NEA 1999). Since it is the principal federal agency charged with providing support for the arts, its artistic selections are likely to lend high visibility and legitimacy to the creative activities it embraces. Its budget, originally handled on a national level, is now partially allocated among state and local arts agencies in all 50 states and possessions, for grass roots groups, and especially for educational projects.

4. Artistic Genres Multiplied: Boundary Blurring

4.1 Cultural Contradictions or Expansion of Civil Society?

(a) As important as official government agencies may be in recognizing new art forms, public policy alone does not account for the rise of outsider art. It is only one ingredient among others, of which the most important are art market initiatives. Although therapists played a vital part in recognizing the creations of their patients, the discovery of outsider art hinged on collectors, dealers, and scholars in the context of an expanding art world. These developments were vital to broadening the domain of the arts more generally, including the entry of new collectors. This is indicated by the growth of interest among African–American collectors in African tribal art as well as in works by African–American artists (Halle 1993). The exclusionary definition of art proposed by some of the most influential philosophical estheticians and buttressed by the academic system has ceded to one more sensitive to nuance, process, and inclusivism: not high vs. low but high and low; not insider vs. outsider but insider and outsider; not art vs. politics but art and politics. These boundary blurring innovations are seen by some as a danger, indications that much of the Western world has entered a period of 'cultural declassification—an unraveling and weakening of ritual classifications.' The discomfort that this may create is expressed by the sociologist Daniel Bell, who regrets this symptom of the decline of the fine arts into no more than the impermanence of fashion. He takes it as a sign of an esthetic malaise because of a 'disjuncture of culture and social structure' (Bell 1996). Other sociologists, such as Paul DiMaggio, see declassification from a more dispassionate standpoint, as a predictable result of the intimate relation between them. DiMaggio attributes it to a combination of factors, among which the most salient are the transformation of local upper classes into a US elite, anchored in organizations and less in community; an increase in the influence of commercial principles of classification as popular culture industries expand; the emergence of relatively autonomous and competitive high culture art worlds; the growth of mass higher education; and changes in the US state (DiMaggio 1987, p. 452).

(b) While an air of elitism still clings to the arts, both artists and the public they seek have expanded to the point where earlier conceptions of exclusivity seem to many to be in need of revision. To some degree, this was driven by the new challenge of the 1960s when abstractionism was confronted by Postmodernism. In Andreas Huyssen's analysis, Postmodernism challenged the *avant garde* autonomous fine art sphere, which had preserved traditional strictures of uniqueness and originality from the illegitimate importations of technology. In challenging this position, certain artists launched a revolution that brought the techniques of mass media into the domain of the fine arts. No longer would the quasi-sacred realm of fine art be clearly distinguished from that of commerce. Instead, it came to include everyday consumer goods, on the same level as the aura-laden fine art to which Walter Benjamin had directed his attention (Benjamin 1969). In the process, they blurred the line—the Great Divide—between fine art and commercial art (Huyssen 1986, Cherbo 1997). Moreover, makers of prints, color lithographs, and photographs that reproduce unlimited numbers of copies were not content to be mere adjuncts for reproducing already acknowledged works of art. Instead, they claim these media of design and advertising are themselves Art. Upholders of the older, exclusive *avant garde* ideals became the chief detractors of Postmodernism. Their efforts were largely in vain: the day of clearcut boundaries between fine and other art forms had passed.

(c) Postmodernism has another side that is closely connected to the make up of societal membership. Although the situation varies considerably from one nation to another, in many the exclusion from the esthetic core of certain art forms on the basis of their creators' categorical memberships—based on gender, race, class, status, or socially defined handicaps—have gained attention of policy makers and scholars. Since art worlds are embedded in, and encompass a part of, the social fabric, it is not surprising that the tensions and demands of under-represented groups and their art works are revealed in them. These forms of the Postmodern turn have implications for all the arts, but outsider art may be the most salient because it reveals the multitude of domains in which different forms of artistic excellence are recognized.

4.2 The Triumph of the Transitory

(a) The dynamic of modernism was centered on the very dismantling of a guiding canon, the blurring of the boundary between a governing center; and recurring waves of outsiders struggling to become insiders. In the 'tradition of the new' art could be intended or unintended, made by either professionals or nonprofessionals (Rosenberg 1964). Unexpectedly, however, vanguard innovations beginning in the last decade of the nineteenth century, for better or worse, did so much violence to the renaissance and enlightenment heritage underlying the cultural structure of the arts, that it was altogether reasonable to characterize their effect as constituting 'the shock of the new' (Hughes 1981). Startling as it was at the time, that shock did not in itself immediately destroy the consensus surrounding the fine arts. A century later, however, the visual and aural revolution heralded at its outset has thrown into question the very idea of a high aesthetic realm. The range and density of artistic

change have reached their apogee in the form of a permanent revolution.

(b) The existence of outsider art implies an insider art, one in which a canon serves as a focus around which artistic products and their makers are evaluated. But it is difficult today to identify a canon that governs art as opposed to non-art. Without an autonomous domain of fine art based on a consensus of aesthetic standards and criteria, in a world in which anything can potentially be art, it is to be expected that outsider art is not more likely to be excluded than a potentially infinite variety of other genres. This does not imply that artistic recognition no longer exists, but that recognition and legitimation are no longer identifiably situated in a single institution such as an academy. Rather, they inhabit a domain composed of a plurality of gatekeepers—organizations, influential individuals, publications, media, popular and commercial or elite and scholarly, each of which may be local, national, or international in reach. Insider–outsider distinctions have become multidimensional; they are matters of degree rather than of kind. Recognition may be founded on the fame and glamour of stardom, commercial success based on sales, critical or scholarly appreciation, depending upon the trajectory of creation and reception.

(c) No longer hemmed in by a single canon governing fine art, competing groups promote forms or styles that they identify as their own. Shifting power centers support claims to validity to a wide range of outsider activities and mentalities (Huyssen 1986, p. 218). In the process, the European autonomous sphere of fine art has come to exist as one cultural structure among several. The artistic tradition in which both integrated professional artists and mavericks furthered their own creativity and success (Becker 1982) has also enabled the arts to become available for other purposes: for therapists using art, music, or theatrical performance for prisoners, the elderly, the ill, to improve their sense of self worth or to reinforce a sense of ethnic identity (Zolberg and Cherbo 1997). However, this does not exclude the probability from an administrative perspective that the arts may be a means of social control in custodial institutions. What is clear is that the dynamic of insiders and outsiders extends beyond the bounded art world of objects that may be bought and sold, that gain or lose value. The conjunction of government programs with agendas not specifically intended to construct this genre, and commercial forces that capitalize on the spending power of an enlarged clientele with empathy for authenticity and spontaneity, have played a role in creating this genre.

See also: Art and Culture, Economics of; Art: Anthropological Aspects; Art, Sociology of; Censorship and Transgressive Art; Cultural Expression and Action; Fine Arts

Bibliography

Ardery J S 1997 Loser wins: Outsider art and the salvaging of disinterestedness. *Poetics* **24**: 329–46

Barron S 1991 *Degenerate Art*. The Los Angeles County Museum of Art, Los Angeles, CA

Becker H S 1982 *Art Worlds*. University of California Press, Berkeley and Los Angeles, CA

Bell D 1996 *The Cultural Contradictions of Capitalism*. Basic Books, New York

Benjamin W 1969 The work of art in the age of mechanical reproduction. In: Arendt H (ed.) *Illuminations*. Schocken Books, New York

Boime A 1971 *The Academy and French Painting in the 19th Century*. Phaidon, New York

Bowler A 1997 Asylum art: the social construction of an aesthetic category. In: Zolberg V L, Cherbo J M (eds.) *Outsider Art: Contesting Boundaries in Contemporary Culture*. Cambridge University Press, New York, pp. 11–36

Cherbo J M 1997 Population art: Ugly duckling to swan. In: Zolberg V L, Cherbo J M (eds.) *Outsider Art: Contesting Boundaries in Contemporary Culture*. Cambridge University Press, New York, pp. 85–97

Clifford J 1988 *The Predicament of Culture*. Harvard University Press, Cambridge, MA

Corvisier A 1978 *Arts et Sociétés dans l'Europe du XVIII Siècle*. Presses Universitaires Françaises, Paris

Crane D 1987 *The Transformation of the Avant-Garde: The New York Art World 1940–1985*. University of Chicago Press, Chicago

Danto D C 1986 *The Philosophical Disenfranchisement of Art*. Columbica University Press, New York

DiMaggio P 1987 Classification in art. *The American Sociological Review* **52**: 440–55

Gerth H, Wright M C 1946 *From Max Weber*. Oxford University Press, New York

Gilman S 1982 *Seeing the Insane*. John Wiley & Sons in association with Brunner/Mazel, New York

Goffman, E 1961 *Asylums: Essays on the Social Situation of Mental Patients and other Inmates*. Doubleday, Garden City, NY

Goldwater R 1986 *Primitivism in Modern Art*. The Belknap Press of Harvard University Press, Cambridge, MA

Graña C 1964 *Bohemian and Bourgeois: French Society and the French Man of Letters in the 19th Century*. Basic Books, New York

Hall M D, Metcalf E W Jr (eds.) 1994 *Creativity and the Boundaries of Culture*. Smithsonian Institution Press, Washington, DC

Halle D 1993 *Inside Culture: Art and Class in the Modern American Home*. University of Chicago Press, Chicago

Hughes R 1981 *The Shock of the New*. Alfred A. Knopf, New York

Huyssen A 1986 *After the Great Divide: Modernism, Mass Culture, Postmodernism*. Indiana University Press, Bloomington, IN

Kammen M 2000 Culture and the state in America. In: Bradford G, Gary M, Wallach G (eds.) *The Politics of Culture*. The Center for Arts and Culture and New York: The New Press, Washington, DC

Larson G 1983 *The Reluctant Patron: The US Government and the Arts, 1943–1965*. University of Pennsylvania Press, Philadelphia

1997 *American Canvas*. The National Endowment for the Arts, Washington, DC

Marling K A 1982 *Wall-to-Wall America: A Cultural History of Post-Office Murals in the Great Depression*. University of Minnesota Press, Minneapolis, MN
McDonald W F 1969 *Federal Relief Administration and the Arts*. Ohio State University Press, Columbus, OH
Mesnard A H 1974 *La Politique Culturelle de l'État*. Presses Universitaires de France, Paris
Meyer K E 1979 *The Art Museum: Power, Money, Ethics*. William Morrow, New York
Morgan H W 1978 *New Museums: Art in American Culture*. University of Oklahoma Press, Norman, OK
Morgenthaler W 1992 *Madness and Art: The Life and Work of Adolf Wölfli* (trans. Esman A H, Spoerri E). University of Nebraska Press, Lincoln, NE
National Endowment for the Arts 1999 Home Page: NEA Strategic Plan-Accomplishments
Poggioli R 1971 *The Theory of the Avant-Garde*. Harvard University Press, Cambridge, MA
Prinzhorn H 1972 *Artistry of the Mentally Ill* (trans. Brockdorff E). Heidelberg, New York and Springer-Verlag, Berlin
Rosenberg H 1964 *The Tradition of the New*. McGraw-Hill, New York
Smithsonian Center for Folklife Programs and Cultural Studies 1999 Home Page on Folkways
Toennies F 1963 *Community and Society*. Harper and Row, New York
Tuchman M, Eliel C S (eds.) 1992 *Parallel Visions: The Modern Artist and the Outsider*. Los Angeles County Museum of Art and Princeton University Press, Los Angeles, CA
Vogel S 1988 *ART/artifact*. The Center for African Art, New York
Wolff K H 1950 *The Sociology of Georg Simmel*. The Free Press, Glencoe, IL
Zolberg V 1990 *Constructing a Sociology of the Arts*. Cambridge University Press, New York
1993 Censorship in the United States: politics, morality, and the Arts. In: Calhoun C, Ritzer G (eds.) *Social Problems*. McGraw-Hill, New York
Zolberg V L 1997 African legacies, American realities: art and artists on the edge. In: Zolberg V L, Cherbo J M (eds.) *Outsider Art: Contesting Boundaries in Contemporary Culture*. Cambridge University Press, New York
Zolberg V L, Cherbo J M (eds.) 1997 *Outsider Art: Contesting Boundaries in Contemporary Culture*. Cambridge University Press, New York

V. L. Zolberg

Cultural Property and Ancestral Remains, Repatriation of

Repatriation refers to the return of objects of cultural heritage to the living descendants of their creators. This seemingly simple concept is made enormously complex by the emotional, legal, and political environment in which claims for repatriation take place.

1. *Heritage of the Group or of Humanity?*

This environment is conditioned by the historical interplay of two basic human desires: (a) the urge of individuals to give meaning to their lives through their sense of belonging to a special group; and (b) the fascination that human beings have for the ancient and the exotic, the beautiful and the bizarre, and the unknown and the faraway.

People declare their group identity through physical objects that symbolize the special relationships they share. Ultimately, these objects take on an almost mythical significance with the possession and curation of them considered essential to the physical well being and the spiritual continuity of the group. Throughout history, conquerors in recognition of the importance of such objects have mutilated or destroyed them and confiscated them as lasting reminders of conquest and subjugation.

Thousands of objects have been removed from their homelands as souvenirs. Explorers and travelers have filled the museums of Europe and America. A global tourist industry has enabled people everywhere to participate in a culture of humanity-at-large, a shared world heritage. Western European countries, inspired by the art and literature of the classical world, helped both to create and to disseminate this composite heritage of humanity. Objects from the ancient Mediterranean region now grace the halls of museums worldwide.

These influences created an appreciation of other cultures and sometimes a better understanding of them. Unfortunately, what is seen by Europeans and Americans as appreciation and understanding is often interpreted as domination, exploitation, and appropriation by many indigenous people and former colonial subjects. However, these two versions of heritage are not antithetical. One may appreciate the cumulative global human achievement and at the same time participate fully in one's chosen group.

2. *Setting the Stage for Repatriation*

The exploration that introduced exotic people and their cultural property to Europeans spawned a period of economic expansion, imperial conquest, and colonial exploration. European explorers and conquerors brought home strange and beautiful objects from all over the world. Among these objects were often the very physical symbols that maintained the social, political, and spiritual identity of the subjugated. Objects were acquired as the booty of war, legitimate diplomatic and personal gifts, honorable and forced purchase, and outright thefts. Many were the motives for transferring objects of cultural heritage from their homelands to the museums and private collections of the conquering nations: aesthetic desire, religious fervor, military vengeance, international competition, commercial greed, imperial arrogance, cultural ig-

norance, and racial intolerance (Greenfield 1996, Messenger 1999).

During the last half of the twentieth century, the demand for antiquities increased dramatically. The clandestine looting of protected archaeological sites, the widespread theft of antiquities, and the illicit trade in cultural property aroused the conscience of officials and politicians, as well as the anger of indigenous people, archaeologists, and private citizens. Efforts to control this looting of world cultural heritage resulted in the 1970 UNESCO Convention on the Means of Prohibiting and Preventing the Illicit Import, Export, and Transfer of Ownership of Cultural Property. Several major art-importing countries, for example, the UK, are not yet signatories to it.

The twentieth century also saw the breakup of the colonial empires and the emergence of new nations eager to recover the physical symbols of their cultural identity. Tribal groups with long memories and functioning cultures began to seek the return of their cultural property. Those politically conscious groups that had already lost their cultural heritage attempted to recreate it through repatriation.

Tourism became a critical component of world economy and tribes and nations alike wanted to have their cultural symbols available for tourists to see. The great European and American museums that displayed the shared global heritage were reluctant to disperse their collections through repatriation. Some of the objects involved, such as the Elgin or Parthenon Marbles in the British Museum and the Venus de Milo in the Louvre, had become 'national treasures' of great importance to the heritage of their new homelands.

3. *International Repatriation*

The return of cultural property at the international level has been carried out for a long time. After Alexander the Great conquered the Persians in 331 BC, he forced them to return the Tyrannicides sculptures they had removed when they defeated Athens in 480 BC. The idea that cultural property was not the booty of conquest was present in European treaties of the twentieth century. It was established in international law when France was required in 1815 to return the art taken by Napoleon. The laws of different countries are often contradictory, making the resolution of claims complex and difficult, but the main problems are not so much legal as moral. Who owns the past? Who curates the cultural property? Who controls the cultural heritage?

Despite these problems, many repatriations have been carried out internationally, especially since World War II. At least a dozen countries have returned cultural property to more than 30 nations, from the Icelandic literary treasures returned to Iceland by Denmark, to the human skeletons turned over to the Australian Aborigines by the University of Edinburgh, and the Lydian Hoard sent back to Turkey by the Metropolitan Museum of Art. However, other international claims continue to be controversial and unresolved, the most celebrated of which is the Greek effort to retrieve the Parthenon Marbles from the British Museum (Greenfield 1996, pp. 42–90).

4. *Domestic Repatriation*

Although restitution of cultural property at the international level attracts a great deal of attention, the most important developments in repatriation have taken place in domestic or internal national contexts. Beginning in the 1960s, indigenous people in the USA, Canada, and Australia documented losses of cultural heritage and made claims for the return not only of their cultural property but also of the bones of their ancestors. They made these claims using the legal and political processes available to citizens in their countries.

In the USA, the creative tension between the national culture and minority groups is not unlike that between the heritage of all humanity and that of ethnic groups. Just as in the global situation, special groups of American citizens take part in the national culture without giving up their cultural identity. American Indian activists had two additional advantages. First, after World War II, the USA began to recognize the rights and needs of its minority citizens and to celebrate its cultural diversity. Second, American Indians are not just another minority group. They have a special status as domestic sovereign nations within the American legal system, continuing the British colonial practice of dealing with Indian tribes on a government-to-government basis. Building on this special status, Indian leaders in the Native American Rights Fund and the National Congress of American Indians insisted that Indian citizens had a right to the control of their cultural heritage and that the bones of their ancestors deserved proper and respectful treatment.

During the national debate on these issues, many tribal groups provided examples of past mistreatment and injustices, the loss of cultural property through warfare, theft, and deceit, and especially the indignities caused by the storage of the bones of their ancestors in museums. In general, the members of the public, increasingly aware of the many injustices suffered by Indian people, especially during the westward expansion of the nation, supported their claims. In 1990, the Congress, influenced by this public support, passed the Native American Graves Protection and Repatriation Act (NAGPRA).

A complex and comprehensive piece of legislation, NAGPRA requires that museums receiving federal assistance inventory their American Indian holdings, report them to appropriate tribes, consult with the tribes, and respond to requests for repatriation of

human remains and associated grave goods, sacred objects, and objects of cultural patrimony. Some museum professionals feared that NAGPRA would cause them to lose all of their Native American materials. Some archaeologists predicted the end of research on American Indian cultures. Biological anthropologists expressed concern that skeletons would be reburied without scientific analysis. This concern is a continuing problem because lineal descent is often difficult to demonstrate and cultural affiliation for remains from remote time periods cannot be assumed solely on the basis of geography. In practice, the requirement that museums and Indian tribes consult with one another has been a powerful mechanism for building positive relationships. The consultations have led to highly productive new working arrangements characterized by considerable good will and cooperation. In sum, although there are still some issues to be resolved, NAGPRA is working reasonably well, as are comparable laws in other countries (Bray and Killion 1994, Mihesuah 2000, Swidler et al. 1997).

5. *Basic Principles*

The various historical, political, legal, and emotional contexts in which the removal and redistribution of cultural property have taken place make each case unique. Case by case analysis is necessary to establish ownership, the method of acquisition, the nature of the cultural property, and the significance of it to the claimants.

Concepts of property differ from one culture or country to another. However, answering the ownership question is often difficult, especially in international situations, because of the secrecy surrounding transactions, the lack of regulation of the art marketplace, and the questionable practices frequently employed. Purchase of an object 'in good faith' from a 'rightful owner' is meaningless unless the place of origin, the means of acquisition, and the subsequent history of the object are known. Fraudulent documentation designed to legitimize stolen and smuggled objects, not only clouds the ownership but also compounds the ethical issues (Messenger 1999).

NAGPRA has been successful because it eliminates problems of acquisition and ownership. The NAGPRA focus is on the nature of the object and its significance to the claimant. If an object fits the statutory definition, is culturally identified with the claimant group, and is significant to that group, it is eligible for repatriation. The question is not the legality of current ownership, but the appropriateness of the proposed repatriation.

Although there may always be problems of definition, cultural affiliation, legality, and even morality, the basic principles for dealing with repatriation claims are clear. Experience shows that these principles work reasonably well, especially when applied in a climate of good will and mutual trust and respect. The USA has facilitated the international return of cultural property through innovative use of its criminal law and has transformed domestic repatriation by creating a new legal framework that allows its American Indian citizens to reclaim with dignity the cultural and spiritual traditions of their heritage.

The right of a people to possess their symbolic cultural property and to control their cultural heritage underlies the very concept of repatriation. Repatriation is neither an effort to go back to the past nor an attempt to redress past injustices. Rather, it is an opportunity to start again in a new relationship based on mutual trust and respect. Such relationships are now the norm in the southwestern USA, where American Indians, anthropologists, and curators have been in close contact for more than 100 years. They have discovered how to talk and listen to one another, how to disagree and compromise, and how to learn from one another. Out of this experience has come the kind of trust and respect that has made possible meaningful consultations, ethical decisions, and successful repatriations.

Bibliography

Bray T L, Killion T W (eds.) 1994 *Reckoning with the Dead: The Lausen Bay Repatriation and the Smithsonian Institution*. Smithsonian Institution Press, Washington, DC

Greenfield J 1996 *The Return of Cultural Treasures*, 2nd edn. Cambridge University Press, Cambridge, UK

Messenger P M (ed.) 1999 *The Ethics of Collecting Cultural Property: Whose Culture? Whose Property?* 2nd edn. University of New Mexico Press, Albuquerque, NM

Mihesuah D A (ed.) 2000 *Repatriation Reader: Who Owns American Indian Remains?* University of Nebraska Press, Lincoln, NE

Swidler N, Dongoske K E, Anyon R, Downer A S (eds.) 1997 *Native Americans and Archaeologists: Stepping Stones to Common Ground*. Altamiva Press, Walnut Creek, CA

R. H. Thompson

Cultural Psychology

1. *Cultural Psychology: What Is It?*

A major aim of the discipline called cultural psychology is to document variations in modes of (and ideals for) 'normal' psychological functioning across ethnic communities. One assumption of the discipline—the premise of non-uniformity—is that there is a substantial portion of the psychological nature of human beings that is neither homogeneous nor fixed across time and space. Cultural psychology seeks to document the protean cultural aspects of human

psychological nature. The discipline can be defined as the study of the distinctive mentalities of particular peoples (Balinese Hindus, Satmar Hasidim, Chinese Mandarins) (see Boesch 1991, Geertz 1973, Greenfield and Cocking 1994, Markus et al. 1996, Miller 1997, Shweder 1991, Shweder and LeVine 1984, Shweder et al. 1998).

Cultural psychology can thus be distinguished from the study of general psychology, which is the study of mental structures and processes that are so widely distributed as to characterize the normal psychological functioning of all human beings (or perhaps even non-human primates as well). Research in cultural psychology has, for example, systematically corroborated the special status accorded to the defense of female 'honor' in the mentality of many Southern American white males (Nisbett and Cohen 1995). Research in cultural psychology has authenticated the claim that there is a positive sense of empowerment and a feeling of virtue associated with 'modesty' and the attitude of 'respectful restraint' in the psychology of women in some regions of the contemporary non-Western world (Menon and Shweder 1998). Such feelings of power and goodness associated with 'modesty' contrast with ideas about (and ideals for) psychological functioning constructed in the contemporary Anglo-American cultural region. Research findings in cultural psychology thus raise provocative questions about the integrity and value of alternative forms of subjectivity across cultural groups.

A second assumption of cultural psychology is that many mental states (and some mental processes) are best understood as by-products of the never-ending attempt of particular groups of people to understand themselves and to make manifest their self-understandings through social practices (Bruner 1990, Geertz 1973, Wierzbicka 1993, 1999). That might be called the premise of self-reflexive social construction. Whether one studies Inuit Eskimos or Anglo-American middle-class urban liberals, the aim in cultural psychology is to spell out the implicit meanings (the goals, values, and pictures of the world) that give shape to psychological processes (Briggs 1970, White and Kirkpatrick 1985). The aim is to examine the patchy or uneven distribution of those meanings on a global scale and to investigate the manner of their social acquisition, for example by means of participation in the symbolic practices, including linguistic practices, of this or that tradition-sensitive cultural group.

2. *The Mental States of 'Others'*

One of the several aims of cultural psychology is to develop a language for the comparative study of mental states that makes it possible to understand and appreciate the mental life of 'others.' 'Others' refers to members of some different cultural community who by virtue of life-long membership in that group ascribe meaning to their lives in the light of wants, feelings, values, and beliefs that are not necessarily the same as one's own.

Cultural psychologists are interested, for example, in cultural variations in the degree to which feelings are constructed as 'emotions.' Anna Wierzbicka (1999) has suggested that while all normal human beings have 'feelings' (e.g., pleasure and pain, arousal and serenity) many of the 'emotions' lexicalized in the English language are not universally available in the mental life of people around the world. There has been much research in cultural psychology into the character of the particular emotions (Ifaluk 'fago,' American 'happiness') that are salient or important in different types of social worlds (Lutz 1988, Kitayama and Markus 1994, Shweder and Haidt 2000).

Cultural psychologists are interested as well in population-based variations in social cognition, moral judgment, and the sources of personal fulfillment or life-satisfaction in different social groups. For example, they have studied the origin, significance, and place of 'filial piety' and the 'social motivation to achieve' in some East Asian populations (Yang 1997). They have investigated the self-empowering aspects of ascetic denial and other forms of sacrifice among high-caste women in South Asia (Menon 2000). They are concerned to document the divergent meanings and distinctive somatic and affective vicissitudes of such experiences as 'loss' (or 'success,' or of 'not getting the things you want,' etc.) for members of different cultural communities. It is in the pursuit of such research questions that they have discovered replicable cultural differences in reports about the quality of the experience of 'loss' (and 'gain'). In comparison to majority populations in northern Europe or the USA, majority populations in Samoa and China are more likely to react to apparent 'loss' with feelings such as headaches, backaches, and other types of physical pain than with feelings such as sadness or dysphoria (Levy 1973, Kleinman 1986). Cultural psychologists also seek to document differences in modes of thought (e.g., analytic vs. holistic) (Nisbett et al. in press), in self-organization (e.g., interdependent vs. independent) (Markus and Kitayama 1991) and in moral judgment (e.g., reliance on an ethics of autonomy vs. an ethics of community vs. an ethics of divinity) across different types of groups (Shweder et al. 1997).

In most of its research the field of cultural psychology has been pluralistic in its conception of 'normal' psychological functioning and interdisciplinary in its conception of how to go about studying the origin, meaning, and social role of particular mental states on a worldwide scale. The field draws together anthropologists, psychologists, linguists, biologists, and philosophers in its study of the diverse, yet potentially effective, modes of psychological functioning that have been produced, and socially endorsed, in different cultural traditions, and have

made those cultural traditions possible. Indeed, cultural psychology is sometimes described as the study of the way culture and psyche 'make each other up.'

3. The Tenet of Psychic Pluralism

Taken together, the premise of non-uniformity and premise of self-reflexive social construction sum up to a principle that is at the heart of the discipline of cultural psychology. This is the tenet of psychic pluralism, which states that the study of 'normal psychology' is the study of multiple psychologies, and not just the study of a single or uniform fundamental psychology for all peoples of the world. Cultural psychology is thus the study of the way the human 'mind' (understood to consist of an inherently complex, heterogeneous collection of abstract and/or latent schemata) can be transformed, and made functional, in a number of different ways, which are not equally distributed across time or space. Hence the slogan, 'one mind, many mentalities: universalism without the uniformity' (Shweder et al. 1998) which is meant to give expression to goals of a discipline aimed at developing a credible theory of psychological pluralism.

Cultural psychology can also be understood as a project designed to critically assess the limitations and incompleteness of all uniformitarian versions of the idea of psychic unity. Alternatively put, cultural psychology is the study of ethnic and cultural sources of diversity in emotional and somatic functioning, self-organization, moral evaluation, social cognition, and human development. It is the study of population differences in the things people know, think, want, feel, and value, and hence are customarily motivated to do, by virtue of life-long membership in some recognizable group that has a history and a conception of its own destiny. 'To be a member of a group,' the eighteenth-century German romantic philosopher Johann Herder argued, 'is to think and act in a certain way in the light of particular goals, values and pictures of the world; and to think and act so is to belong to a group' (Berlin 1976). Although the field of cultural psychology has many ancestral spirits (Giambattista Vico, Wilhelm Wundt, Wilhelm Dilthey, Edward Sapir, Ruth Benedict) and some very prominent contemporary advocates (Jerome Bruner, Michael Cole, Clifford Geertz, Jacqueline Goodnow, Patricia Greenfield, Shinobu Kitayama, Hazel Markus, Richard Nisbett, Hazel Markus, Anna Wierzbicka), Johann Herder is justly claimed as one of the original cultural psychologists, although he was probably not the first. For an account of the historical development of the field, see Jahoda (1991) and Cole (1996).

4. Culture and the 'Custom Complex'

One of the contributions of cultural psychology is to revive a conception of 'culture' that is both symbolic and behavioral. In the history of twentieth-century anthropological thought, the idea of culture was variously defined, either behaviorally (as patterns of behavior that are learned and passed on from generation to generation) or 'symbolically' (as the categories, beliefs, and doctrines that organize, rationalize, and justify a way of life). In research on the cultural psychology of a particular cultural community, the notion of 'culture' usually refers to community-specific ideas about what is true, good, beautiful, and efficient that are made manifest in behavior. To qualify as 'cultural' those ideas about truth, goodness, beauty, and efficiency must be socially inherited and customary; and they must actually be constitutive of different ways of life, and play a part in the self-understanding (including the self-criticism) of members of the community.

Alternatively stated, the concept of 'culture' as used in cultural psychology refers to what Isaiah Berlin called 'goals, values and pictures of the world' that are made manifest in the speech, laws, and routine practices of some self-monitoring group. These are sometimes also called 'cultural models' (D'Andrade 1995, Shore 1996). Cultural psychologists thus engage in the interpretive, symbolic, or cognitive analysis of behavior. They assume that actions speak louder than words and that 'practices' are a central unit for cultural analysis. Thus, what John Whiting and Irvin Child (1953) once referred to as the 'custom complex' is a natural unit of analysis or starting point for a study in cultural psychology.

Whiting and Child introduced the idea of a 'custom complex' in 1953 but the basic idea was not really taken up and carried forward until the rebirth of cultural psychology in the 1980s and 1990s. According to Whiting and Child, a custom complex 'consists of a customary practice [for example, a family meal, arranged marriage, animal sacrifice, a gender identity ceremony involving circumcision] and of beliefs, values, sanctions, rules, motives and satisfactions associated with it.' There are many labels that have been placed on this type of unit of analysis (a 'custom complex,' a 'life space,' a 'habitus'). Whatever the label, this is the type of unit of analysis that makes it possible to conceptualize cultural psychology as the study of the way culture (ideas about what is true, good, beautiful, and efficient made manifest in practice) and psyche (what people know, think, feel, want, value, and hence choose to do) afford each other realization, and thus 'make each other up.'

5. An Example of a 'Custom Complex': Genital Surgeries in Africa

A rather dramatic but highly illuminating example of a 'custom complex' is the 'circumcision ceremony' or 'genital surgery' that is customary for both boys and girls in many East and West African ethnic groups. In

the countries of Sierra Leone, Mali, the Gambia, Ethiopia, Somalia, the Sudan, and Egypt genital surgeries are culturally endorsed and popular for both males and females, and receive high approval ratings from the vast majority of men and women in the general population. Human rights advocacy groups in Europe and in the USA have criticized the practice. Nevertheless, the most comprehensive and rigorous review of the medical and demographic literature on African genital surgeries suggests that widely publicized claims about the severe and negative consequences of the custom for health, sexuality, and childbirth are quite inaccurate, and deserve to be discounted (Obermeyer 1999). In a recent manuscript on the cultural psychology of male and female circumcision among the Kono people of Sierra Leone, Fuambai Ahmadu, who is both a 'native' and an anthropologist, has written as follows (Ahmadu 2000, p. 301).

> It is difficult for me—considering the number of these ceremonies I have observed, including my own—to accept that what appear to be expressions of joy and ecstatic celebrations of womanhood in actuality disguise hidden experiences of coercion and subjugation. Indeed, I offer that most Kono women who uphold these rituals do so because they *want* to—they relish the supernatural powers of their ritual leaders over against men in society, and they embrace the legitimacy of female authority and, particularly, the authority of their mothers and grandmothers.

Among the various goals, values, and pictures of the world (the cultural psychology) that make this practice meaningful and satisfying for those men and women for whom it is a 'custom complex' are the following:

(a) A culturally shared belief that the body (especially the genitals) of both males and females are sexually ambiguous until modified through surgical intervention. According to this picture of the world the foreskin of a boy is viewed as a feminine element and masculinity is enhanced by its removal. Similarly the clitoris is viewed as an unwelcome vestige of the male organ, and Kono females (for example), as described by Ahmadu, seek to feminize and hence empower themselves by getting rid of what they perceive as a dispensable trace of unwanted male anatomy.

(b) A culturally shared aesthetic standard in terms of which the genitals of both women and men are viewed as ugly, misshapen and unappealing if left in their 'natural' state. For many African men and women the ideal of beauty is associated with a sexual anatomy that is smooth, cleansed and cleaned (shaved) and free of all 'fleshy encumbrances.'

There is of course much more to be said about the beliefs, values, sanctions, rules, motives, and satisfactions associated with female (and male) genital surgeries in Africa. Nevertheless, in the light of these and other culturally endorsed reasons a genital surgery is experienced as an improvement of the body in many East and West African ethnic groups (Ahmadu 2000).

The cultural psychology of the culturally endorsed practices of any particular community is likely to result in a depiction of 'other minds' that is unsettling, astonishing, or at least surprising for those whose goals, values, and pictures of the world are characteristic of the cultural psychology of some other group. Radical divergence in the moral evaluation of particular 'custom complexes,' such as 'circumcision,' is itself an important topic for research in cultural psychology. Indeed, the cultural psychology of moral evaluation is currently an active research area. Moral concepts (e.g., human rights, justice as equality) privileged in some sections of society and in some regions of the world are not necessarily the moral concepts that are most salient and important in other sectors of society and regions of the world (Shweder et al. 1997).

6. *Cultural Psychology: What It is Not*

The tenet of psychic pluralism and the emphasis on 'goals, values, and pictures of the world' as a source of psychological differences between cultural communities distinguishes cultural psychology from other fields of study (such as 'cross-cultural psychology' and 'national character studies') with which it should not be confused.

6.1 *It is Not Cross-Cultural Psychology*

Research in cultural psychology proceeds on the assumption that psychological diversity is inherent in the human condition, and that culture and psyche are interdependent and 'make each other up.' It should be noted in passing that any theory of psychological pluralism would lack credibility if it staunchly denied the existence of any and all universals. Indeed, as should be obvious, cultural psychology presupposes many psychological universals (e.g., feelings, wants, goals, ideas of good and bad, of cause and effect, of part–whole relationships, etc.). However, the search for and the privileging of things that are uniform across all peoples is a project that goes under other names, for example, general psychology or perhaps even 'cross-cultural psychology.' (See Schweder 2000 for a discussion of cross-cultural psychology, partly recapitulated here.)

Some of the goals of cross-cultural psychology (which I am contrasting with cultural psychology) have been described by Segall et al. (1998). One goal is 'to generate more nearly universal psychology, one that has pan-human validity' and to attain 'a universally applicable psychological theory.' A second closely related goal is to 'keep peeling away at the onion skin of culture so as to reveal the psychic unity of mankind at its core.' It is for that reason that cross-cultural psychology (not to be confused with cultural psychology) can be viewed as a vigilant cousin of

general psychology; for they both share the same uniformitarian goals. Such goals give a distinctive character to cross-cultural psychology, and they help explain why all of the following kinds of activities are typical of research by cross-cultural psychologists and distinguish it from research by cultural psychologists.

Thus one of the aims of cross-cultural psychology is to determine the boundary conditions for generalizations generated in Western laboratories with Western (mostly college student) subjects, generalizations which, prior to critical examination by cross-cultural psychologists, have been presumptively interpreted as fundamental or 'natural' and universalized to the whole world. The aim here for the cross-cultural psychologist is not to represent the distinctive cultural psychology of particular peoples. This aim does not result in research focused on differences in the way members of different communities perceive, categorize, feel, want, choose, evaluate, and communicate that can be traced to differences in salient community-based 'goals, values, and pictures of the world.' Rather the aim for the cross-cultural psychologist is to make sure that the hoped-for universal psychology is truly universal and to throw out any claim that only holds in the Anglo-American world. This is an extremely useful corrective for the tendency of Western psychologists to overgeneralize their findings, but it is not the same as undertaking a project in cultural psychology.

A second aim of cross-cultural psychology is to establish comparability or equivalence for measuring instruments across different populations. Often the goal here is to try to show that people in different cultures really are alike, and that any reported differences in performance were due to noise, or inappropriate measuring instruments, or bad translations, or misunderstandings about the way to ask and answer questions. The instincts of a cultural psychologist run in quite a different direction. For a cultural psychologist (not to be confused with a 'cross-cultural psychologist') the 'noise' is interpreted as a signal about true differences in cultural meanings (goals, values, and pictures of the world), and not as something to eliminate or overcome. Indeed, cultural psychologists are likely to worry if one's measuring instruments travel easily and well from university classroom to university classroom around the world and display the same psychometric properties here and there. They may suspect that one has not really landed in a truly different culture at all. This is because 'peeling away the onion skin of culture so as to reveal the psychic unity of mankind at its core' is not what cultural psychology is about.

A third major aim of cross-cultural psychology is to focus on 'independent variables' of the cultural environment (for example, nucleation of the family, literacy vs. non-literacy) that are thought either to promote or to retard psychological development. In such research 'development' is almost always defined in terms of universal norms for cognitive, emotional, or social functioning (e.g., Piaget's notion of 'formal operational' thinking or Ainsworth's notion of healthy 'attachment'). Cultural psychology is primarily concerned with the elaboration and discovery of alternative or plural norms for successful psychological development, which is another way it can be distinguished from cross-cultural psychology (LeVine 1990).

6.2 *It is Not National Character Studies*

Attempts to characterize whole populations in terms of generalized dispositions (e.g., authoritarianism, Apollonianism, high need for achievement) went out of fashion in psychological anthropology in the late 1950s and early 1960s. This happened for several reasons but largely because looking for variations in types of personality traits to explain differences in cultural practices or custom complexes (and vice versa) turned out to be something of a dead end. It was discovered that if one tries to describe individuals within and across cultural communities in terms of general dispositions or traits of character, then within-group variations typically exceed between-group variations. It was discovered that hypothesized 'modal personality' types typically characterize only about a third of the population in any particular cultural group.

A major insight, although a fragile one, of recent work in cultural psychology is that it is better to represent and interpret human behavior the way sensible economists do rather than the way personality trait theorists do. That is to say, it is better to think about behavior as emanating from 'agency,' and to analyze it as the joint product of 'preferences' (including goals, values, and 'ends' of various sorts) and 'constraints' (including beliefs, information, skills, material and social resources, and 'means' of various sorts). This avoids the hazards of dispositional approaches in which behavior is interpreted as the by-product of mechanical forces pushing both from inside (in the form of 'personality traits') and outside (in the form of 'situational pressures'). Ultimately, a fully successful piece of research in cultural psychology must avoid nominal dispositional categories such as 'holistic vs. analytic' and render behavior intelligible in terms of the particular goals, values, and pictures of the world that motivate and inform the 'domain-specific' behaviors and routine practices of specific intentional agents. To do otherwise is to reify cultural stereotypes and fall into some of the traps of the past.

7. *The Future: Going Indigenous*

The field of cultural psychology that has re-emerged on the North American and European scene during the past 20 years is quite similar to an intellectual

movement that has grown up in the non-Western world and is increasingly known as 'indigenous psychology.' One of the most eminent theoreticians of this movement, Kuo-shu Yang, the Taiwanese psychologist, lists several ways to 'indigenize' psychological research. Here are four of Professor Yang's virtues for the aspiring indigenous psychologist of China (Yang 1997):

(a) 'Give priority to the study of culturally unique psychological and behavioral phenomena or characteristics of the Chinese people.'

(b) 'Investigate both the specific content and the involved process of the phenomenon.'

(c) Make it a rule to begin any research with a thorough immersion into the natural, concrete details of the phenomenon to be studied.

(d) Let research be based upon the Chinese intellectual tradition rather than the Western intellectual tradition.

Those are some of the virtues that define 'cultural psychology' as well, although it remains to be seen how many of us can live up to such demanding standards. Even today with the rebirth and renewal of cultural psychology not all research actually begins with fieldwork or with 'a thorough immersion into the natural concrete details of the phenomenon to be studied.' All too often research still starts with a published finding from some Western laboratory, which is then subjected to critical examination by means of various attempts at replication with populations from other societies. But things are changing. Given the increasingly international and interdisciplinary character of collaborative scholarship in cultural psychology one looks forward to more and more research that keeps faith with Kuo-shu Yang's high ideals.

See also: Cross-cultural Psychology; Cultural Studies: Cultural Concerns; Culture and Emotion; Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Culture: Contemporary Views; Culture in Development; Culture, Sociology of; Ecology, Cultural; Female Excision: Cultural Concerns; Indigenous Conceptions of Aging; Indigenous View on Human Development: West African Perspective

Bibliography

Ahmadu F 2000 Rites and wrongs: An insider/outsider reflects on power and excision. In: Shell-Duncan B, Hernlund Y (eds.) *Female 'Circumcision' in Africa: Culture, Change and Controversy*. Lynne Rienner, Boulder, CO, pp. 283–312

Berlin I 1976 *Vico and Herder*. Hogarth, London

Boesch E E 1991 *Symbolic Action Theory and Cultural Psychology*. Springer, Berlin

Briggs J L 1970 *Never in Anger: Portrait of an Eskimo Family*. Harvard University Press, Cambridge, MA

Bruner J S 1990 *Acts of Meaning*. Harvard University Press, Cambridge, MA

Cole M 1996 *Cultural Psychology: A Once and Future Discipline*. Harvard University Press, Cambridge, MA

D'Andrade R G 1995 *The Development of Cognitive Anthropology*. Cambridge University Press, New York

Geertz C 1973 *The Interpretation of Culture*. Basic Books, New York

Greenfield P, Cocking R 1994 *Cross-cultural Roots of Minority Child Development*. Erlbaum, Hindsdale, NJ

Jahoda G 1991 *Crossroads Between Culture and Mind: Continuities and Change in Theories of Human Nature*. Harvard University Press, Cambridge, MA

Kitayama S, Markus H R 1994 *Emotion and Culture: Empirical Studies of Mutual Influence*. American Psychological Association, Washington, DC

Kleinman A 1986 *The Social Origins of Distress and Disease*. Yale University Press, New Haven, CT

LeVine R A 1990 Infant environments in psychoanalysis: A cross-cultural view. In: Stigler J, Shweder R, Herdt G (eds.) *Cultural Psychology: Essays on Comparative Human Development*. Cambridge University Press, New York, pp. 454–74

Levy R 1973 *Tahitians: Mind and Experience in the Society Islands*. University of Chicago Press, Chicago

Lutz C 1988 *Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Change to Western Theory*. University of Chicago Press, Chicago

Markus H R, Kitayama S 1991 Culture and the self: Implications for cognition, emotion and motivation. *Psychological Review* **98**: 224–53

Markus H R, Kitayama S, Heiman R 1996 Culture and 'basic' psychological principles. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology: Handbook of Basic Principles*. Guilford, New York, pp. 857–913

Menon U 2000 Does feminism have universal relevance? The challenges posed by Oriya Hindu family practices. *Daedalus: Journal of the American Academy of Arts and Sciences* **129**: 77–100. [Special issue on *The End of Tolerance: Engaging Cultural Differences*]

Menon U, Shweder R A 1998 The return of the 'white man's burden': the encounter between the moral discourse of anthropology and the domestic life of Hindu women. In: Shweder R A (ed.) *Welcome to Middle Age! (And Other Cultural Fictions)*. University of Chicago Press, Chicago, pp. 139–88

Miller J G 1997 Theoretical issues in cultural psychology. In: Triandis H C (ed.) *Handbook of Cross-cultural Psychology*, 1: *Theory and Method*. Allyn and Bacon, Boston, MA, pp. 85–128

Nisbett R E, Cohen D 1995 *The Culture of Honor: The Psychology of Violence in the South*. Westview, Boulder, CO

Nisbett R E, Peng K, Choi I, Norenzayan A (in press). Culture and systems of thought: Holistic vs. analytic cognition. *Psychological Review*

Obermeyer C M 1999 Female genital surgeries: The known, the unknown, and the unknowable. *Medical Anthropology Quarterly* **13**: 79–106

Segall M H, Lonner W J, Berry J W 1998 Cross-cultural psychology as a scholarly discipline. *American Psychologist* **53**: 1101–10

Shore B 1996 *Culture in Mind: Cognition, Culture and the Problem of Meaning*. Oxford University Press, New York

Shweder R A 1991 *Thinking Through Cultures: Expeditions in Cultural Psychology*. Harvard University Press, Cambridge, MA

Shweder R A 2000 The psychology of practice and the practice of the three psychologies. *Asian Journal of Social Psychology* **3**: 207–21
Shweder R A, Goodnow J, Hatano G, LeVine R, Markus H, Miller P 1998 The cultural psychology of development: One mind, many mentalities. In: Damon W (ed.) *Handbook of Child Psychology*, 1. John Wiley, New York, pp. 865–937
Shweder R A, Haidt J 2000 The cultural psychology of the emotions. In: Lewis M, Haviland J (eds.) *The Handbook of Emotions*, 2nd edn., Guilford, New York, pp. 397–414
Shweder R A, LeVine R A 1984 *Culture Theory: Essays on Mind, Self and Emotion*. Cambridge University Press, New York
Shweder R A, Much N C, Mahapatra M, Park L 1997 The 'big three' of morality (autonomy, community, divinity), and the 'big three' explanations of suffering. In: Rozin P, Brandt A (eds.) *Morality and Health*. Routledge, New York, pp. 119–69
Valsiner J (ed.) 1997 The legacy of Ernst E Boesch in cultural psychology. *Culture & Psychology* [Special Issue] **3**: 243–7
White G M, Kirkpatrick J 1985 *Person, Self and Experience: Exploring Pacific Ethnopsychologies*. University of California Press, Berkeley, CA
Whiting J W M, Child I 1953 *Child Training and Personality*. Yale University Press, New Haven, CT
Wierzbicka A 1993 A conceptual basis for cultural psychology. *Ethos* **21**: 205–31
Wierzbicka A 1999 *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge University Press, New York
Yang, K 1997 Indigenizing Westernized Chinese psychology. In: Bond M H (ed.) *Working at the Interface of Culture: Twenty Lives in Social Science*. Routledge, London

R. A. Shweder

Cultural Relativism, Anthropology of

Relativism, which may be taken in most general terms as the awareness of the presence and importance of the perspectival in judgement and understanding, is generally understood as a philosophical problem and a hobgoblin of the pretensions to generalization and universalization of thought characteristic of the intellectual classes. But much if not all that has been taken up and intellectualized by these classes was first found among the folk as part of their lore. Anthropologists and folklorists can readily confirm this observation, for in practically any book of proverbs will be found axiomatic observations which are relativistic in import. People are usually aware from their domestic everyday experience of the difference of perspective and the relativity in understanding between men and women, the old and the young, the parent and the child, the slow and the quick. We might observe that it is philosophers and theologians who have lived among absolutes and have worked to impose them, while in 'the practice of everyday life' the common people characteristically live among relatives. Lately, philosophers have been increasingly taking cognizance of that practice and its implicit perspectivalism if not relativism.

In philosophy, to be sure, there has been a long tradition of relativistic thought. The Sophists' cultivation of the contestatory rhetoric of community life and community organization is often taken as relativistic in implication and action, and Hume's (1975) skepticism about general observations including any universal moral principles is surely relativistic in spirit. Kant's view of the constructed nature of the metaphysical and Hegel's historicism are instrumental in the development of the more radical relativisms found in Marx and Nietzsche. Modern-day pragmatism continues this perspectivalism, whether in the works of Foucault (1979), Derrida (1993) or Rorty (1989). These scholars focus, respectively, on knowledge as conditioned by power, knowledge as conditioned by rhetorical or textual construction, and knowledge as conditioned by social solidarity. Such views have steadily and perhaps definitively been undermining overarching claims to metaphysical universals. It has certainly given rise, in respect to moral philosophy, to a 'rational relativism.'

Anthropology, a discipline mainly emergent since the mid-nineteenth century, has not been immune to these philosophical developments and the crisis of representation and legitimation they usher in. Indeed, it has become, through its various arguments for cultural and linguistic autochthony, a central player in the relativism/universalism wars. The interest of anthropologists, particularly in the issue of cultural relativism, derives first from ethnographic field experience of the variable proverbial wisdom of local life lived on the ground, as it were, or from the fact that anthropologists usually work comparatively across cultures and, therefore, across cultural and language perspectives, or from the long interest in anthropology in the difficult problem of universal human rights. Philosophy, understandably, has reacted by exercising pointed critique of the *incoherence* of anthropological relativism—it is inescapably self-refuting and it undermines its own truth claims—and the *perniciousness* of anthropological relativism—it undermines philosophical commitment to the improvement of thought.

The anthropology of Franz Boas is generally seen as the fountain head of cultural relativism in American anthropology. Although originally a physicist, Boas was increasingly cautious both about universalization of things human and the hierarchization of human groups which often accompanied such practice. His chief intellectual opponent was classic evolutionism, with its invidious ranking of societies on a scale from savagery to civilization. His chief popular opponent was racism, or in its benign form, the Civilizing Mission. Indeed one of Boas' few popular books, in an otherwise vast oeuvre—*The Mind of Primitive Man* (Boas 1938)—attacked 'racial determinism,' that is,

the association between biology and 'personality' (by which he meant behavior in the broadest sense). He did this by focusing on the shaping force of culture upon behavior and understanding (that is, upon 'mind'). In this book, therefore, itself preceded by any number of lectures and papers dating from the late-nineteenth century, Boas early laid the ground work for the 'cultural relativism' of his students, although this was not a term he employed to any extent.

There are several other classic stimuli to relativistic views in the Boas oeuvre. In 'The study of geography' (Boas 1887), while resisting geographical determinism, he stimulated thinking on methodological relativism by pointing to two different and justifiable approaches to field data, the nomothetic or law-seeking and the idiographic or descriptive, each of which, differently attuned to local circumstance, gave different yet necessary understandings of the field data. It may be said, particularly in postmodern times, that although increasingly convinced by the 'structuring power of tradition,' Boas in his emerging relativism never abandoned entirely the universalist hope of the scientific mission and method and the seeking, however cautiously, of lawful generalizations.

Cultural relativism in its strong form is identified with two of Boas' foremost students, Ruth Benedict and Melville J. Herskovits. In her widely read *Patterns of Culture*, Benedict (1934) took particular aim at universalist notions of the normal and abnormal. This argument was made effective through its arresting portraits of three cultures, Pueblo, Dobu, and Northwest Coast Indians, whose characteristic behaviors, abnormal from a Western perspective, are portrayed as normal from the local perspective.

Benedict repeatedly resisted ethnocentric and provincial views of the normal and argued a 'Great Arc of human possibility' (Benedict 1934, p. 34) of which no culture could take full advantage and on which every culture had to make its particular choices. It was Melville Herskovits, however, who most directly framed the issue of 'cultural relativism' in his general text *Man and His Works* (Herskovits 1948), and whose collected arguments could be posthumously published as *Cultural Relativism* (Herskovits 1972).

None of the Boasians, including Herskovits, were out and out relativists in the epistemological sense that they gave up on the possibilities of discovering universals beyond the fact of cultural relativity. In all their writings there appears the promise of those eventual discoveries. They are all to be considered, rather, methodological and moral relativists whose main aim was to resist ignorance of and moral superiority towards other cultures, and exploitative intolerance of them. The aim was to increase mutual respect and tolerance in the world. Herskovits argued, however, that such cultural relativism was not derived from any kind of moral philosophy, but rather from the empirical anthropological observations on diversity by field-working anthropologists. Nevertheless, he struggled with many dilemmas of philosophical type in his 'cultural relativism': negotiating the universal and the particular, the practical and the ethical, the 'is' and the 'ought,' anthropological science and anthropological humanism, the cross-cultural and the within-cultural view, the applied and the philosophical. His strategy was to convert these dilemmas, which arise from the fate of the field-working man or woman of science finding himself or herself inescapably situated in the fold of philosophy, into the revelatory paradoxes of anthropological practice (Fernandez 1990).

The heyday of 'cultural relativism' in the USA—in the sense of seeking to define a durable doctrine in the face of persistent assault from various moral philosophies on the one hand and philosophies of science on the other—were the decades of the mid-twentieth century, the 1930s through the 1960s. Subsequently, however, some notable contributions to this enduring issue have come forth together with a contemporary challenge to relativistic thought: the human rights movement. Marshall Sahlins (1976), although not espousing 'cultural relativism' in its paradoxical sense of moral obligation to defend diverse local moral orders, nevertheless conducted a compelling argument against the culture-free universalization, mainly among economists and political economists, which accepts utility interests as the dominant principle in practical reasoning. In the 1980s, a notable argument developed over relativism and the role of culture in determining reality principally among three anthropologists: Renato Rosaldo, Clifford Geertz and Melville Spiro.

Rosaldo (1989) addresses the impact on older notions of the stable truths and dispassionate objectivities of classic social analysis made by awareness of both the anthropologist and his or her informants as 'positioned (and repositioned) subjects' recurrently engaged in an ongoing dramatic process of 'relational understanding.' In such a process truth becomes a very moving target only perspectivally to be apprehended, if at all, and in any case best interpreted by narrative analysis firmly set within and not beyond cultural contexts. While not espousing a straight-line cultural relativism, he notes and resists the return of a 'nostalgia' for the authoritative, often imperious, truths of objective social science—a nostalgia often expressed in an attack on a stereotyped notion of relativism.

This strategy of skepticism has also been that of Geertz, the central player in the interpretive approach that puts the play of perspectives at the center of anthropological inquiry. In 1984 he offered an intricate argument, exemplified in the double negative of the title of the article, 'Anti–anti relativism' (Geertz 1984). He takes to task the resurgence of absolutist interests in affirming some metaphysical or, what is the same, very physical, that is to say biological, realities behind the influence of cultural and language differences. This search for 'nature's own vocabulary,' and the security

of objective biological truths on this side of culture, truncates what anthropology is all about, and Geertz questions the kind of 'hearty common sense' that simply ignores the diversity of lifeways plentifully evidenced in the anthropological archive. While written in defense of that broadly rather than narrowly humanizing knowledge of diversity laboriously acquired by anthropology, Geertz neither directly espouses in any direct way the moral commitments present in an earlier 'cultural relativism' nor denies the probability that on the other side of culture there may be some universals to discover.

In defense of the possibility of scientific, which is to say universal, measurability and explanation across cultural boundaries, Spiro (1986) develops a useful contrast between three types of cultural relativism: descriptive (that there are important empirical differences between cultures), normative (that evaluative judgement of the norms expressed in these differences are always relative to cultural contexts and hence not generalizeable) and epistemological (that culture determines what we know and how we know it). Spiro accepts descriptive relativism and the perspectivalism it reflects, and the cautions of field methods it enjoins. He rejects the strong form of normative relativism, which argues against any possibility of cross-cultural judgement of human practices, as well as epistemological relativism in general, which would argue both for ontological incommensurability between cultures and for the irrelevance or perversity of the discovery of biological and psychological commonalities in humankind. It has been the tendency, he argues, in the fin-de-siècle age of textual hermeneutics, to espouse an immoderate form of epistemological relativism; something that the earlier cultural relativists, although committed to descriptive and moderate normative relativism, never did. The consequence is simply to deny the possibilities of an explanatory science of humans anchored in a valid and verifiable logic of discovery, with all the implications for professional authority that such a denial of principled comparative inquiry entails.

In the 1990s, an important area of challenge to cultural relativism arose in the argument for universal human rights. At first glance, cultural relativism would seem to be an impediment to such universalization of a value. Indeed, oppressive states proclaiming state sovereignty over their citizens often claim justification in 'cultural relativism.' This defense begs the question for all concerned of cultural boundaries within states, which is to say the question of the definition of culture and its boundaries in a pluralizing, globalizing world. That the relation between cultural relativism and universal human rights need not be taken as inevitably oppositional is seen in recent efforts on the part of the Human Rights Commission of the American Anthropological Association, to posit a species-wide disposition to create culture and cultural difference, and the corresponding right to such creation and such difference as long as it does not impose on that right in others (American Anthropological Association 1998, Turner 1997). There may remain problems with this enlightenment argument, as theory and practice, but it effectively moves the debate away, as for example does Geertz' argument before it, from a simple and fruitless opposition in which one or another pole of the debate is confined to a demonic reading. Also, it preserves the practical and methodological aspects of cultural relativism as a posture of critical inquiry.

In reviewing this century-long argument over relativism in the social sciences, we may note several principal points. First, there is the tendency for stereotypic views of the opposite argument to emerge, with the consequence that the opponent's views are trivialized and simplified. Second, strong and moderate forms of cultural relativism in its various aspects (descriptive, normative and epistemological) must be distinguished if the argument is to obtain to any clarifying power. Third, the methodological aspect of cultural relativism and its importance, in its descriptive and normative aspect, as a posture of inquiry must be recognized for those, mostly but not exclusively anthropologists, who have chosen to work in other cultures. Fourth, the discussion of cultural relativism is inevitably related to one's definition of culture and particularly cultural boundedness. Fifth, it must be recognized that the logics of inquiry differ significantly between the sciences and the humanities. Hence the discussion of relativism is itself relative to what disciplinary faculty is anchoring and being evoked in the discussion. Sixth, there is a politics of relativism which is inevitably present—the Boasian relativists, for example, were pronouncedly egalitarian, pluralist, and anti-authoritarian in their values. Relativism has been a recurrent irritation, a valuable provocation, and a useful restraint on imperious inquiry in the academy, and has produced over this long period some of the most interesting arguments about just how individual scholars and researchers with different agendas working in different cultures with differing worldviews are to come together and engage in productive and fair-minded inquiry. Without the gadfly of relativism, the horsemen of social science, galloping through their agendas of investigation, might well be more roughshod, less discriminating, less insightful, and less empathetic beings.

See also: Benedict, Ruth (1887–1948); Mead, Margaret (1901–78); Postmodernism: Philosophical Aspects; Relativism: Philosophical Aspects

Bibliography

American Anthropological Association 1998 *Code of Ethics of the American Anthropological Association*. Web Document: http://ameranthassn.org/committees/ethics/ethcode.htm

Benedict R 1934 *Patterns of Culture*. Riverside Press, Cambridge, UK

Boas F 1887 The study of geography. *Science* **9**: 137–41
Boas F 1938 *The Mind of Primitive Man*. Macmillan, New York
Derrida J 1993 *Aporias: Dying–Awaiting (One Another at) the 'Limits of Truth' (Mourir–s'Attendre aux 'Limites de la Vérité')* [Dutoit T trans.]. Stanford University Press, Stanford, CA
Fernandez J W 1990 Tolerance in a repugnant world and other dilemmas in the cultural relativism of Melville J. Herskovits. *Ethos* **18**: 140–64
Foucault M 1979 Passim. In: Morris M, Patton P (eds.) *Foucault: Power, Truth, Strategy*. Feral, Sydney, Australia
Geertz C 1984 Distinguished lecture: Anti–anti relativism. *American Anthropologist* **86**: 263–79
Herskovits M J 1948 *Man and His Works*. Knopf, New York
Herskovits M J (ed.) 1972 *Cultural Relativism; Perspectives in Cultural Pluralism*. Random House, New York
Hume D 1975 (1777) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Clarendon Press, Oxford, UK
Rorty R 1989 *Contingency, Irony and Solidarity*. Cambridge University Press, New York
Rosaldo R 1989 *Culture and Truth*. Beocen Press, Boston, MA
Sahlins M D 1976 *Culture and Practical Reason*. University of Chicago Press, Chicago
Spiro M E 1986 Cultural relativism and the future of anthropology. *Cultural Anthropology* **1**: 259–86
Turner T 1997 Human rights, human difference: Anthropology's contribution to an emancipatory cultural politics. *Journal of Anthropological Research* **53**: 273–91

J. W. Fernandez

Cultural Resource Management (CRM): Conservation of Cultural Heritage

Cultural resource management is the practice of conserving and preserving cultural heritage. Cultural resources include archaeological, historical, architectural, and engineering properties, and associated artifacts and documents. Also included are museums, archives, archaeological parks, shipwrecks, cultural landscapes, religious sites, as well as traditional lifeways and languages. (Knudson 1986, pp. 400–2, King 1998, pp. 5–9). These resources represent the cultural heritage of individual peoples and nations, and collectively, they represent the heritage of the world. Cultural resources reflect finite periods of time, or ephemeral cultural traditions. Once destroyed, they are gone forever. Economic development, looting, international trafficking in stolen artifacts, and natural erosion destroy thousands of cultural resources every year (Cameron 1997). For this reason, the primary management goal is preservation *in situ*.

The term 'cultural resource management' is used mostly in the United States by archaeologists working for federal and state agencies or for private firms contracting with such agencies (Fowler 1982, Green and Doershuk 1998). Heritage management is a similar term used in many parts of the world (McManamon and Hatton 2000). Heritage management and cultural resource management can be used interchangeably.

1. *Value of Cultural Resources*

People value cultural resources for present and future uses (Lipe 1984). They are tangible links to the past allowing people to experience and relate directly to their history, and hold an evocative power sometimes lacking in documentary sources. Heritage properties have symbolic values for nations and ethnic groups because they represent a common history and identity, as well as cultural and spiritual norms. The process of deciding which cultural resources to preserve and monumentalize is therefore frequently contentious and politicized (Tunbridge and Ashworth 1996).

The aesthetic value of cultural resources, especially architectural properties, provides incentive for the reuse of historic buildings long after original uses are no longer economically viable. The preservation of the historic core of nineteenth-century business districts in the United States is a priority because of the perceived aesthetic value of early buildings. The aesthetic appeal of architectural ruins around the world is a primary reason for their continued preservation.

Cultural resources have informational value. Archaeological sites provide the only record of human achievement for much of history. At best, written documents cover events of only the past 5,000 years and only in a few parts of the world. In some parts of North America written records represent a scant 200 years. Government and business records, newspapers, diaries, etc., are fragile and subject to loss due to natural events, deterioration, neglect, and intentional destruction. Also, official records provide information primarily about dominant cultures and leaders. For these reasons the physical evidence in historic and archaeological sites may be the only record of many past human activities, even in relatively recent times, and their destruction results in an immeasurable loss of information.

Cultural resources have economic value resulting primarily from their associative and aesthetic values. People want to visit heritage sites. Tourism is one of the largest industries in the world, and heritage tourism is one of the largest sectors of the industry. For many countries, such as the United Kingdom and Jordan, tourism is a major part of the economy. In the United States, cultural resources support a billion-dollar industry in the State of New Mexico alone.

2. *Legal Protection*

Cultural resources are the tangible links to the cultural heritage of the world. For this reason, UNESCO and the International Council on Monu-

ments and Sites (ICOMOS), a private organization advising UNESCO, have produced charters establishing standards for preservation (US/ICOMOS 1999). Two foundational documents guiding heritage management at the international level are the 1964 Venice Charter adopted by ICOMOS, and the Convention Concerning the Protection of World Cultural and Natural Heritage signed by 158 nations, and adopted by UNESCO in 1970. UNESCO also manages the World Heritage List, which recognizes cultural and natural sites of world significance, and the World Heritage Fund, providing financial assistance for various preservation projects.

Despite international interest in preserving heritage resources, most cultural resource management is conducted under the laws of individual nations or local divisions within nations. Laws have been established to protect cultural resources threatened by economic development projects, urban sprawl, vandals, looters, and neglect. In addition, nations must make a multitude of complex decisions regarding these issues. Governments must determine the ownership of individual artifacts and collections; oversee excavation permits and arrange for the use and care of resulting artifacts and records; mediate controversial issues such as the excavation of graves and human remains; allocate responsibility for funding archaeological excavations or restoration projects in advance of construction projects; and determine the extent of involvement of various interested parties, such as indigenous peoples, in the processes of decision-making. Not all countries of the world have laws and regulations dealing with these issues, and some issues are more important in one country than another. Legal traditions and history invariably affect the law. For example, in Mexico the State asserts ownership of all prehistoric artifacts. United States law reflects strong traditions safeguarding private property. Only artifacts located on federal land are federally owned, and private landowners retain the ownership of materials found on their land.

3. *Management Requirements*

Knowledge is crucial in managing cultural resources. Agencies need to know where the resources are, which are significant, why they are significant and to whom. Plans for their preservation must be developed. If sites are threatened by dams, roads, water systems, or buildings, planning statutes usually require agencies or various units of government to address the impact of such projects. Nongovernmental corporations or organizations are frequently required to assess the effects of their projects on cultural resources or pay to have this work done. Accurate information is also crucial to protecting sites from looting and vandalism. Cultural resource inventories and evaluation studies should be conducted by appropriately trained specialists, such as archaeologists, architectural historians, historians, and anthropologists. Thus, an inventory of significant cultural resources is the primary basis for management decisions.

Many countries have developed criteria to determine which cultural resources are significant, and these criteria vary from country to country. In the United States, properties as young as 50 years old are evaluated, while in Jordan only properties older than AD 1750 need be considered. The range in kind and type of significant properties also varies from place to place. Some countries consider only archaeological and historic properties, while others include sacred and traditional cultural sites, and even contemporary expressions of culture such as festivals and dances. Government agencies usually make decisions concerning significance, but even if a country does not have a formal evaluation procedure, someone must decide which resources are ultimately important and worthy of mitigation or preservation (Thompson 1982).

Once an inventory is obtained, it is necessary to determine which resources are threatened and need immediate consideration. It is a sad fact, but not all sites determined to be culturally significant can be preserved. Developing priorities for action is usually a process of give and take between those with preservation interests and others primarily working toward economic development. When properties can not be preserved, appropriate mitigation measures must be developed. The type of mitigation adopted depends on the type of cultural resource effected. Archaeological sites are usually excavated, or sometimes buried for protection. Architectural and engineering properties are documented through photography, photogrammetry, measured drawings, and the salvage of individual architectural features. When feasible, architectural properties can be moved to new locations. It is difficult to develop satisfactory mitigation measures for the religious or sacred properties often important to traditional communities. Hard decisions have to be made about whether such controversial projects will be completed as planned, or moved to another location.

Cultural resources in national parks and other protected areas may not be threatened by economic development projects, but they still need active management to prevent deterioration. The site of Petra in Jordan is a good example. Petra is famous for massive sandstone architecture and for water-control systems built by the Nabataean Arabs before Roman occupation. These famous architectural elements are continually eroded by wind-driven sand and rain and will disappear if not conserved. Jordanian conservation scientists are actively developing methods to preserve the site. Around the world, even in protected areas, limited financial resources dictate preservation priorities.

Another important element in the management of cultural resources involves the care of archaeological

and museum collections, archives, and documents. Great expense is involved in providing state-of-the-art, climate-controlled, and safe environments for storage, processing, and use of such collections. The conservation of paper, metal, and other perishable artifacts is a science in itself. Heated debates arise around the world about the appropriate distribution of artifacts. Many countries such as Guatemala and Turkey have increasingly strict laws about the transport and use of artifacts, and many European countries with colonial traditions and large archaeological collections from the Middle East and elsewhere are facing various questions of repatriation (Herscher 1998). The establishment and funding of curation facilities are part of this debate, as well as being crucial to the availability of knowledge and resources to future generations.

Traditions concerning historic and prehistoric cultural artifacts and human remains differ widely among the cultures of the world. Such belief systems must increasingly be taken into consideration in the formation of cultural resource management policies, as well as in national and international law. In some countries indigenous peoples have insisted on the repatriation of human remains and associated funerary objects, and other objects of cultural patrimony (Jones and Harris 1998). A satisfactory balance between the interests of the broader society in cultural resources and the needs of individual communities has not yet been reached, and these are among the most controversial issues facing heritage managers today (Suagee 1999). Conflicts are exacerbated in countries with recent histories of social injustice, and otherwise routine decisions are fraught with emotion. Animosities can be allayed as governments take responsibility for the legal protection of various rights and as peoples are forced to deal amicably and respectfully with each other's goals and beliefs. Consultation between government bodies and local communities is essential to the success of preservation efforts.

4. *Tourism and CRM*

The management of archaeological parks, museums, and historic districts for tourism is a significant part of cultural resource management. Tourism offers many opportunities for the preservation and restoration of heritage properties around the world, and because of the economic benefits of tourism, many individual communities and national governments are developing properties for tourists. The correct manner of development (Boniface 1995), the authenticity and quality of interpretative presentation (Bruner 1994), and the type and use of restoration methods are hotly debated issues. The tourist infrastructure, as well as interpretive facilities like trails and museums, can negatively impact on the very cultural resources and landscapes tourists want to visit. The plan to build a gondola to transport tourists to Machu Picchu in Peru is a case in point. In addition, when the interests of local communities are not considered, conflicts can be expected.

5. *Public Support and Education*

The preservation of cultural resources is dependent on the support of the general public. The charters of UNESCO and ICOMOS stress that the success of local and national preservation initiatives are dependent on programs educating the public about the value of cultural resources. Heritage education in the public schools is required in many countries to foster a preservation ethic and pride in national heritage. Governmental organizations and professional societies often cooperate in publicity campaigns to combat looting. In those parts of the world where sites of general interest represents a cultural heritage different from that of the contemporary inhabitants, education about the values and economic uses of these resources is especially important. For example, Roman and Byzantine ruins in modern Islamic countries and Native American sites in the United States are often more valued by foreign visitors and national governments than by local communities.

The varieties of cultural resources in the world and the multiplicity of values attached to them make cultural resource management a complex practice. The preservation of cultural resources in the face of economic development, population growth, looting, and world conflicts requires integrated legal tools and protection policies at all levels of government, as well as trained professionals to implement them.

See also: Aboriginal Rights; Archives and Historical Databases; Cultural Property and Ancestral Remains, Repatriation of; Ecotourism; Environmental and Resource Management; Environmental Archaeology; Environmentalism: Preservation and Conservation; Tourism; Tourism, Geography of; Tradition, Anthropology of; Traditions: Social

Bibliography

Boniface P 1995 *Managing Quality Cultural Tourism*. Routledge, London

Bruner E M 1994 Lincoln, Lincoln, as authentic reproduction: A critique of postmodernism. *American Anthropologist* **96**: 397–415

Cameron C M 1997 The loss of cultural heritage—an international perspective. *Nonrenewable Resources* **6**: 71–76

Fowler D D 1982 Cultural resources management. In: Schiffer M (ed.) *Advances in Archaeological Method and Theory*. Academic Press, New York, Vol. 5, pp. 1–50

Green W, Doershuk J F 1998 Cultural resource management and American archaeology. *Journal of Archaeological Research* **6**: 121–67

Herscher E 1998 Many happy returns? New contributions to the repatriation debate. *American Journal of Archaeology* **102**: 809–13

Jones D G, Harris R J 1998 Archeological human remains: Scientific, cultural, and ethical considerations. *Current Anthropology* **39**: 253–64

King T F 1998 *Cultural Resource Laws and Practice*. Alta Mira Press, Walnut Creek, CA

Knudson R 1986 Contemporary cultural resource management. In: Meltzer D J, Fowler D D, Sabloff J A (eds.) *American Archaeology: Past and Future*. Smithsonian Institution Press, Washington, DC

Lipe W D 1984 Value and meaning in cultural resources. In: Cleere H (ed.) *Approaches to the Archaeological Heritage*. Cambridge University Press, Cambridge, UK

McManamon F P, Hatton A 2000 *Cultural Resource Management in Contemporary Society*. Routledge, London

Saugee D B 1999 Human rights and the cultural heritage of Indian tribes in the United States. *International Journal of Cultural Property* **8**: 48–76

Thompson R 1982 Archeological triage: Determining the significance of cultural resources. In: Wilson R L, Loyola G (eds.) *Rescue Archeology*. The Preservation Press, Washington DC

Tunbridge J E, Ashworth G J 1996 *Dissonant Heritage: The Management of the Past as a Resource in Conflict*. John Wiley & Sons, Chichester

US/ICOMOS 1999 ICOMOS Charters and other International Doctrinal Documents. *US/ICOMOS Scientific Journal* **1**: 1

T. J. Green

Cultural Rights and Culture Defense: Cultural Concerns

Although cultural rights are sometimes regarded as the least developed category of human rights, they are unquestionably a part of international human rights law. Cultural rights have been highly controversial because their protection sometimes requires that other human rights be limited. There is substantial debate about the proper interpretation of cultural rights, about which individuals are entitled to invoke them, and about what obligations governments have to enforce these rights.

Cultural rights are significant because culture is an intrinsic part of individual and group identities. Not only does culture exert a crucial influence on the formation of the identity of individuals, but it also helps perpetuate ethnic groups. It is important to appreciate the status of cultural rights in law and politics.

1. The term 'Culture'

In order to interpret cultural rights, it is first necessary to define the term culture. There are many different notions of culture including 'high culture' which refers to museums, operas, and the like, 'popular' or 'mass' culture which means comic books, films, pulp fiction, and 'traditional' culture, which can be construed as a way of life. While some commentators who analyze cultural rights have in mind 'high culture,' most of the discussion focuses on so-called 'traditional culture.' Of the multitude of definitions of culture, a particularly useful one was devised by the Canadian Commission for UNESCO, the United Nations agency responsible for the protection of culture:

> Culture is a dynamic value system of learned elements, with assumptions, conventions, beliefs and rules permitting members of a group to relate to each other and to the world, to communicate and to develop their creative potential (1997).

Culture differs from society inasmuch as culture is an abstraction, whereas society is the collection of individuals in the community.

When culture is analyzed in the context of cultural rights, it is usually used as a synonym for a way of life. Culture encompasses many aspects of life including attire, child rearing practices, diet, marriage customs, politics, religion, death rituals, and so on. In short, culture is constituted of the many and diverse folkways of ethnic groups. One attempt to delineate the many facets of culture is found in the Human Relations Area Files compiled at Yale University. Although some contend that projects which aim at documenting cultural practices risk 'freezing' them in a particular moment in time, cultural theorists invariably acknowledge that cultures are dynamic systems.

2. Cultural Rights

A major question has been whether cultural rights should be conceptualized as individual rights or group rights. Theorists, many of whom come from individual rights traditions, associate group rights with utilitarianism and worry that recognizing group rights may lead to the denial of individual rights. In reality, however, rights which have historically been considered bona fide individual rights such as religious freedom require the existence of a group. While this issue has generated much debate, there is no reason why the right to culture cannot be conceived of both as an individual and as a group right. Insofar as cultural rights include rights to language and religion, it would be difficult to deny the essential group nature of such rights. Moreover, even if the right to culture were an individual right, it could still come into conflict with other individual rights.

Ordinarily, individuals and peoples exist in cultures without having to think self-consciously about the protection of their traditions, most of which are not controversial. It is only when their way of life is threatened by a dominant political system that ethnic minorities are compelled to assert a 'right to' culture.

In this sense cultural rights represent a legal means by which minorities can shield themselves from intervention by the dominant political apparatus, which is normally the state. Governmental intrusion is a manifestation of cultural hegemony, or the imposition of the dominant culture by elites on ethnic minority groups.

Intervention by governments is not an infrequent occurrence as there is a widespread presumption that cultural minorities should become assimilated. When the folkways of a minority group clash with the standards of the dominant system, the expectation is that the members of the minority group should discard their customs. This is sometimes expressed in the form of a proverb: 'When in Rome, do as the Romans do.' However, shedding their folkways may not be a simple matter as individuals are often strongly motivated to follow their cultural traditions. Sometimes they may face a dilemma as failure to comply with the folk law can lead to divine retribution or social ostracism. These individuals either violate the state law and face the consequences in the official legal system or else comply with state law and face dire consequences within their customary law systems.

3. *The Cultural Defense*

If individuals do, in fact, have cultural rights, this implies that there is a duty on the part of governments to allow their cultural practices. This obligation is not an absolute one, though, and cultural rights will be allowed unless a more important, countervailing reason requires that they not be protected. The challenge, naturally, is to distinguish which are valid reasons for overriding traditions and which are not.

Most of the time governments decide that cultural traditions are not permissible because they threaten public health, safety, or morals. If individuals refuse to discontinue the practices, asserting that they have a right to follow their traditions, they may face criminal prosecution. In these circumstances, ethnic minority defendants may invoke what has been called the 'cultural defense.' Such a defense, while not formally recognized by any legal system, sometimes operates to acquit or to mitigate punishment. The core idea behind the defense is that individuals who act under cultural imperatives should be considered less culpable under the law than those who act in the absence of such an imperative. In order for judges to evaluate the extent to which a defendant's cultural background affected his or her behavior, they must consider cultural evidence relating to the offense. The chief difficulty in many cases is the tendency of courts to exclude cultural evidence as irrelevant.

Supporters of the cultural defense favor its adoption in order to ensure the consideration of evidence and to demonstrate genuine commitment to cultural pluralism. The cultural defense would also guarantee more equal treatment for cultural minorities in a legal system whose standards are based on the values of the dominant culture. Advocates contend that the cultural defense would result in condign punishment for ethnic minorities.

Opponents worry that if culturally motivated offenses go unpunished, this will undermine the deterrent function of the law. They often contend that anarchy will reign supreme if each individual can decide with which particular laws he or she will comply. The defense is criticized for giving special rights to minorities. This puts the victims of culturally motivated crimes at a disadvantage, and is therefore unjust.

The presentation of cultural evidence does not commit the court to accepting the argument. The evidence might not establish the existence of the tradition, that the ethnic group had this tradition, or that the defendant was influenced by the tradition when violating the law. The defense merely ensures that the information is treated as admissible in court. To the extent that cultural rights are guaranteed as part of human rights, the argument is that individuals should be entitled, at the very least, to present evidence about their cultural background in a court of law.

4. *Scholarship on Cultural Rights*

The early scholarship on cultural rights was sponsored by UNESCO through symposia designed to clarify the meaning of culture and cultural rights. In 1950 UNESCO published *Freedom and Culture* which was its first attempt to develop the idea of cultural rights. In 1970 UNESCO sponsored a meeting of experts on cultural rights in Paris to discuss the government responsibilities for sacred places, monuments of cultural significance, and language preservation programs. As a great deal of emphasis was placed on cultural property, the meeting dealt with the relationship between individuals and cultural objects.

The Interdisciplinary Institute of Ethics and Human Right at the University of Fribourg, Switzerland held a colloquium on cultural rights in 1991. The Council of Europe held the Eighth International Colloquy on the European Convention on Human Rights in Budapest in 1995 on the theme 'Cultural rights: universal, individual, and legally enforceable individual rights.' In 1998 the International Conference on Cultural Policies of Development met in Stockholm, Sweden to evaluate governmental programs designed to ensure that individuals in all communities can exercise the rights guaranteed by Article 27.

In academic circles two parallel scholarly literatures emerged on the subject of cultural rights, one in political theory, and the other in international law. Political theorists in the Western tradition analyzed the question of whether cultural rights are compatible with liberalism. For the most part political philosophers worry that the acceptance of cultural rights will lead to 'balkanization,' thereby undermining the unity

of the state. The leading proponent of the view that cultural rights can be reconciled with liberalism is Will Kymlicka, a Canadian scholar. His writings underscore the importance that belonging to a community represents to most individuals.

Despite the fact that his work is generally considered to offer the staunchest support for cultural rights, his defense of cultural rights has been criticized for being too limited. First, Kymlicka advocates the protection of cultural rights only in societies which are said to be structured along liberal lines. His theory does not provide cultural rights for immigrants who have chosen to move and thereby give up their culture; only minorities long residing in a state and indigenous peoples can legitimately make claims to cultural rights. His interpretation of cultural rights concentrates mainly on language rights and representation rights. Critics such as Chandran Kukathas have noted that his theory provides minimal protection for cultural rights.

Other leading theorists such as Bhikhu Parekh have attempted to create theories which afford greater protection to cultural rights. In his elaborate scheme advocating 'a principle of dialogical consensus,' Parekh proposes having a minority spokesman who will engage in a dialog with representatives of the majority about cultural practices which clash with 'operative societal values.' It is the duty of the spokesperson to explain how the tradition is authoritative, central to the way of life of the ethnic minority group, and in general, desirable. If it is not feasible to persuade the majority representatives of the value of the tradition under debate and it offends an 'operative societal value,' the ethnic minority must give it up.

Many feminist theorists reject the proposition that cultural rights deserve legal protection because of their concern that protecting these rights will undermine gender equality. Although their work is often framed as a critique of multiculturalism, e.g., Ayelet Shachar, these analysts are essentially denying the validity of the notion of cultural rights.

Generally speaking, most of the commentary on cultural rights in Western political theory is concerned with ways to limit the scope of application of the rights. When cultures do not imprison those who seek to exit, particularly women wishing to flee from oppressive customs, there appear to be fewer objections to the protection of cultural rights. Their objections do not appear to apply to the numerous traditions which have no connection to gender.

In the field of international law, James Anaya has provided a comprehensive analysis of the cultural rights of indigenous peoples. A more general theoretical treatment is that of the late Sebastian Poulter whose brilliant scholarship is extraordinarily important. In Poulter's view, a cultural tradition which violates human rights should not be permitted. According to Poulter, however, if a government, by prohibiting a tradition, is violating human rights, then that tradition should be permitted. His framework has been criticized for presuming the validity of international human rights standards as a basis for determining the validity of traditions and for evading the question of when cultural rights, which are human rights, are superseded by other rights.

5. *Formal Guarantees in Legal Systems*

Few national legal systems officially guarantee a right to culture. The extent to which cultural rights are protected depends on whether or not general principles of law such as religious freedom, equal protection, and freedom of association are interpreted to encompass cultural traditions. Because lawmakers are elites from the dominant culture, they tend not to construe abstract principles in such as way as to benefit ethnic minorities. One sign of the preference for monoculturalism is the widespread preference for single official language policies in many countries.

From time to time, minorities challenge public policies which they regard as unwarranted intrusions into their way of life. For example, in France the Muslim community objected strenuously to the ban against wearing the *hijab* or headscarf in French public schools. The national controversy was known as *L'Affaire des Foulards*. The policy was justified because it ensured the separation of church and state, the integration of foreigners into French society, and gender equality. Some women's groups in France thought the Muslim religious garb, the scarf, symbolized the subordination of women. Eventually, after litigation, the Education Minister relented, allowing the children to wear their religious attire to school.

In the United States indigenous peoples have challenged development projects which would undermine their way of life. In *Lyng vs. Northwest Territory* 1987, tribes argued that the construction of a highway through Chimney Rock would desecrate sacred sites. As part of an environment impact study, the government hired an expert who advised against building the road because of its devastating effect on the tribes. Despite the admonition, the government proceeded with the project, undeterred. The US Supreme Court ruled that the burden on the religion was incidental, not the purpose of the development, and that religious minorities have to accept some burdens.

Difficulties in domestic litigation generally have to do with the question of what counts as the 'authentic' culture or whether a specific practice is 'traditional.' In order to determine the authenticity of cultural traditions, there is the related issue of who is the legitimate representative of the group. Sometimes, indigenous people find offensive the practice of consulting 'experts' or scholars rather than leaders of their communities. Even when the groups themselves are the source of information, it is not always obvious which

person from within the group speaks for the group. Two examples show the practical difficulties in litigating cultural rights claims.

In France the government made an exception to its animal slaughter law to permit Orthodox Jews to slaughter animals in accordance with religious law so that they could consume the meat. France empowered the Central Rabbinical Committee to handle the implementation of the policy. However, one group of Orthodox Jews, concerned that the method employed for the ritual slaughter of animals was not guaranteeing that the meat was sufficiently pure or *glatt*, sued France, challenging its delegation of authority to the Central Rabbinical Committee. After French courts ruled against the group, it appealed to the European Court of Human Rights, where it lost as well.

The Hindmarsh Bridge Island litigation in Australia also shows the dynamics of intragroup tension. A group of aboriginal women objected to a proposed bridge between the Australian mainland and Hindmarsh Island because they claimed it would jeopardize their way of life. The island was ostensibly used for secret women's business, and their customary law supposedly prevented them from divulging information about the nature of the business. During the litigation another group of aboriginal women held a press conference denying the veracity of this claim, saying there was no such 'secret women's business.'

In contrast to domestic litigation which relies on abstract principles interpreted to apply to culture, in international law explicit protections for culture can be found. Among them are the International Covenant on Economic, Social, and Cultural Rights, particularly Article 15, the Declaration on the Rights of Persons Belonging to National or Ethnic, Religious, and Linguistic Minorities (1992), the Framework Convention for the Protection of National Minorities (1995), and the European Charter for Regional or Minority Languages (1992). In addition, there are related documents such as the Draft Declaration on the Rights of Indigenous Peoples and the Declaration on the Elimination of All Forms of Intolerance and of Discrimination Based on Religion or Belief (1981). Cultural rights are covered not only by human rights treaties, but also by instruments concerned with safeguarding cultural heritage such as Convention Concerning the Protection of the World Cultural and Natural Heritage (1972).

To date, the most important cultural rights provision is Article 27 of the International Covenant on Civil and Political Rights (ICCPR), a treaty ratified by virtually all countries in the world. This provision is not subject to any reservation (except for France's), which enables a government to avoid obligations under the article. Article 27 provides:

> In those states in which ethnic, religious, or linguistic minorities exist, persons belonging to such minorities shall not be denied the right, in community with the other members of their group, to enjoy their own culture, to profess and practice their own religion, or to use their own language.

However, Article 27 cannot serve as a guide to determine the substantive meaning of cultural rights under international law because it has been interpreted in only a few situations. As of 2001, the provision has been applied mainly to protect sacred land, the right to hunt animals, and language rights.

Although it is widely referred to as a 'minority' rights provision, it is often invoked in cases involving indigenous peoples, many of whom do not consider themselves to be minorities, subordinate to a state. Some members of indigenous groups object to its application in their cases. Article 27 has also been criticized for being formulated as an individual right and as a negative right. These criticisms may be unwarranted: first, the provision can be invoked by individuals or groups; second, if communities only need to invoke a juridical right to culture when governments threaten to halt their practices, it makes sense that the provision was expressed in negative terms: 'governments shall not deny … .' Moreover, because there was some question about the nature of government obligations for the protection of cultural rights, the Human Rights Committee, which enforces the ICCPR, issued a policy statement interpreting Article 27, known as General Comment 23(5), stipulating that governments take affirmative steps to enforce cultural rights.

One of the best-known disputes in international law in which cultural rights were invoked was the celebrated Lovelace decision which was based on Article 27 of the ICCPR. Sandra Lovelace was a Maliseet Indian who married outside the tribe, thereby losing her tribal status (if a male Maliseet Indian married someone not belonging to the tribe, he would not lose his status). After her divorce, she was unable to rejoin the band as a fully-fledged member. She sued unsuccessfully in the Canadian courts and subsequently filed a complaint with the Human Rights Committee. Although she challenged the sexist membership rule, alleging that it violated human rights law and the tribe defended itself by invoking the right to self-determination (Article 1 of the ICCPR), the Human Rights Committee decided that she had been denied the right to enjoy her culture in her community. Interestingly, this case stands for the proposition that women also have a right to culture!

In the final analysis, the main challenge in international human rights jurisprudence is how to evaluate the relative importance of cultural rights vis-à-vis other human rights. This is known as the hierarchy problem. There are rights conflicts between treaty regimes, for instance between Article 27 and provisions of the Convention on the Elimination of All Forms of Discrimination Against Women which require governments to abolish all traditions that interfere with gender equality. The UN appointed Mrs

Halima Embarek Warzazi as Special Rapporteur on Traditional Practices Affecting the Health of Women and Children. Her studies recommend the elimination of traditions with adverse consequences for women and children. For the most part, fundamental rights conflicts within the United Nations system have not been resolved to date.

6. Theoretical Issues

Regardless of whether or not domestic and international legal systems afford adequate protection to cultural rights, the right to culture may be important because abstract principles require it. Appeal can be made to many principles to justify the recognition of cultural rights; among these are identity, freedom, and tolerance.

If culture is intrinsically linked to the formation of identity, then any general notion of individual or group rights requires that culture be protected. To the extent that democratic systems are committed to the principle of noninterference by the government, intervention in the absence of a serious threat would be illegitimate. The notion of freedom would appear to compel the acceptance of cultural rights. Finally, the idea of tolerance is widely considered an important one in liberal democracies. Although tolerance has somewhat pejorative connotations, i.e., 'putting up' with practices with which one disagrees, allowing the flourishing of diverse cultural traditions is consonant with this principle.

Even if the right to culture is established as a fundamental right, it remains to be seen how much importance it has vis-à-vis other human rights. The hierarchy question is likely to occupy the attention of scholars interested in the nature and scope of cultural rights.

See also: Aboriginal Rights; Civil Liberties and Human Rights; Cultural Policy; Cultural Property and Ancestral Remains, Repatriation of; Cultural Studies: Cultural Concerns; Ethnonationalism: Cultural Concerns; Fundamental Rights and Constitutional Guarantees; Hegemony: Cultural; Human Rights, Anthropology of; Human Rights, History of; Human Rights in Intercultural Discourse: Cultural Concerns; Human Rights: Political Aspects; Multiculturalism and Identity Politics: Cultural Concerns; Rights; Rights: Legal Aspects; Transnational Citizenship: Cultural Concerns

Bibliography

Addis A 1991 Individualism, communitarianism, and the rights of ethnic minorities. *Notre Dame Law Review* **67**: 615–76

Albanese F 1992 Ethnic and linguistic minorities in Europe. *Yearbook of European Law* 313–38

Anaya J 1996 *Indigenous Peoples in International Law*. Oxford University Press, New York

Canadian Commission for Unesco 1977 A working definition of 'culture.' *Cultures* **4**(4): 78–83

Cha'are Shalom Ve Tsedek vs. France. European Court of Human Rights, Application No. 27417/95). June 27, 2000 http://www.echr.coe.int/eng/Judgments.htm

Das V 1994 Cultural rights and the definition of community. In: Mendelsohn O, Baxi U (eds.) *The Rights of Subordinated Peoples*. Oxford University Press, Delhi, pp. 117–58

Gaspard F 1995 *Le Foulard et la Republique*. Editions la Decouverte, Paris

Goulet D 1981 In defense of cultural rights: technology, tradition, and conflicting models of rationality. *Human Rights Quarterly* **3**: 1–19

Human Rights Committee 1994 (April 26) General Comment No. 23 (5) (art. 27) United Nations, Human Rights Committee, Geneva, CCPR/C/21/Rev.1/Add.5

Kukathas C 1992a Are there any cultural rights? *Political Theory* **20**: 105–39

Kukathas C 1992b Cultural rights again. A rejoinder to Kymlicka. *Political Theory* **20**: 674–80

Kymlicka W 1992 The rights of minority cultures. Reply to Kukathas. *Political Theory* **20**: 140–6

Kymlicka W 1995 *Multicultural Citizenship: A Liberal Theory of Minority Rights*. Clarendon Press, Oxford, UK

Levy J T 1997 Classifying cultural rights. In: Shapiro I, Kymlicka W (eds.) *Ethnicity and Groups Rights*. New York University Press, New York and London, pp. 22–66

Lyndall R (ed.) 1996 Secret women's business: the Hindmarsh Island Affair. *Journal of Australian Studies* **48** (Special Issue)

Lyng vs. Northwest Territory 1987 485 US 439

McGoldrick D 1991 Canadian Indians, cultural rights and the Human Rights Committee. *International and Comparative Law Quarterly* **40**: 658–69

Meyer-Bisch P (ed.) 1993 *Les droit culturels – une categorie sou-developpee de droits de l'homme*. Editions Universitaires, Fribourg

Niec H (ed.) 1999 *Cultural Rights and Wrongs*. UNESCO Publishing, Paris; Institute of Art and Law, London

O'Brien S 1987 Cultural rights in the United States: a conflict of values. *Law and Inequality* **5**: 268–358

Parekh B 1996 Minority practices and principles of toleration. *International Migration Review* **30**: 251–84

Poulter S 1986 *English Law and Ethnic Minority Customs*. Butterworths, London

Poulter S 1987 Ethnic minority customs, English law and human rights. *The International and Comparative Law Quarterly* **36**: 589–615

Poulter S 1997 Muslim headscarves in school: contrasting legal approaches in England and France. *Oxford Journal of Legal Studies* **17**(1): 43–74

Poulter S 1998 *Ethnicity, Law and Human Rights: The English Experience*. Clarendon Press, Oxford, UK

Prott L V 1988 Cultural rights as peoples' rights in international law. In: Crawford J (ed.) *The Rights of People*. Clarendon Press, Oxford, UK, pp. 92–106

Renteln A D 2000 Raising cultural defenses. In: Connell J G III, Valladares R L (eds.) *Cultural Issues in Criminal Law*. Juris Publishers, New York

Shachar A 1999 The paradox of multicultural vulnerability. In: Joppke C, Lukes S (eds.) *Multicultural Questions*. Oxford University Press, Oxford, UK, pp. 87–111
Stavenhagen R 1995 Cultural rights and universal human rights. In: Eide A, Krause C, Rosas A (eds.) *Economic, Social and Cultural Rights*. Martinus Nijhoff Publishers, Dordrecht, The Netherlands, pp. 63–77
Symonides J 2000 Cultural Rights. In: Symonides J (ed.) *Human Rights: Concepts and Standards*. pp. 175–227
Szabo I 1974 *Cultural Rights*. A.W. Sijthoff, Leiden, The Netherlands
Tahzib B G 1996 *Freedom of Religion or Belief: Ensuring Effective International Legal Protection*. Martinus Nijhoff, The Hague
Tonkinson R 1997 Anthropology and Aboriginal tradition: The Hindmarsh Island Bridge Affair and the politics of interpretation. *Oceania* **68**(1): 1–26
UNESCO 1970 *Cultural Rights as Human Rights*. UNESCO, Paris
de Varennes F 1996 *Language, Minorities and Human Rights*. Martinus Nijhoff, The Hague
Weiner J F 1999 Culture in a sealed envelope: the concealment of Australian Aboriginal heritage and tradition in the Hindmarsh Island Bridge Affair. *Journal of the Royal Anthropological Institute* **5**(2): 193–210

A. D. Renteln

Cultural Studies: Cultural Concerns

This article attempts to describe that complex contemporary social formation: Cultural Studies (CS). A single thesis organizes the argument. Contemporary CS exists within competing fields of discourse. This discourse, as Carey (1997b) argues, is moving in several directions at the same time (see Frow and Morris 2000, Striphas 1998). This has the effect of simultaneously creating new spaces, new possibilities, and new formations for CS, while closing down others.

Thus, a process of containment is occurring, even as this interdisciplinary field expands globally (*Public Culture* 1988, Frow and Morris 2000, Alasuutari et al. 1998, Hartley 1998). There are those who would marginalize CS, equating it with Marxist thought, and chastising it for not paying adequate homage to Sociology's founding fathers, including Weber and Durkheim (Long 1997). Others would seek a preferred, canonical, but flexible version of the project (see Grossberg 1997a, 1998). Within this framework (see Grossberg 1997a), there are attempts to establish a set of interpretive practices fitted to specific projects. For Grossberg (1997a, pp. 248-261) these practices, or interpretive principles involve a self-reflexive, interdisciplinary project which always detours through theory. This version of CS maintains a commitment to political praxis and radical contextualization, including anti-reductionist, anti-essentialist ontologies. Still others would ironically equate CS with identity politics and critical readings of popular culture. Some would critique the formation from within, distinguishing semiotic, political economy, empiricist, and material approaches to the field's subject matter (Fiske 1994). Still others challenge CS to take up the problems of feminism, gender, racism, colonialism, and nationality (Ferguson and Golding 1997).

1. *The Origins and Definitions of CS*

Ferguson and Golding (1997) observe that 'Few would dispute that CS' 'myths of origin' were made in Britain, and that their 'founding fathers' were Raymond Williams, Richard Hoggart and E. P. Thompson, and subsequently Stuart Hall.' The legacies of these figures is enduring, and includes a commitment to a Marxism without guarantees, a rejection of positivism and functional social theory, and a conception of culture that is political. But on beginnings and origins, Hall (1980) is quite firm, 'there are no absolute beginnings and few unbroken continuities … What is important are the significant breaks.'

The culture in CS is not aesthetic, or literary, it is political, and located in the domain of the popular or the everyday (Storey 1996). The object of study is how culture, as a set of contested interpretive, representational practices, embraces and represents 'a particular way of life, whether of a people, period or a group' (Williams 1976). Popular, everyday cultural practices are treated as social texts. It is understood that nothing stands outside textual representation. Texts, however, involve material practices, structures, flows of power, money, and knowledge. Popular culture is conceptualized as a site of constant negotiation, consent, and resistance. Meaning is always contextual, structural, and anchored in historical processes (Hall 1996b). Cultural processes, in turn, embody class, gender, and racial divisions and relationships. These relationships involve the exercise of power.

1.1 *The Circuits of Culture*

A complex interpretive process shapes the meanings things have for human beings. This process is anchored in the cultural world, in the 'circuit of culture' (du Gay et al. 1997) where meanings are defined by the mass media, including advertising, cinema, and television. This process is based on the articulation or interconnection of several distinct and contingent processes (du Gay et al. 1997). In the circuits of cultural meaning five interconnected processes, representation, identification, production, consumption, and regulation, mutually influence one another (du Gay et al. 1997).

Objects and experiences are represented in terms of salient cultural categories. These categories are connected directly to social and personal identities. These identities are attached to representations of family, race, age, gender, nationality, and social class. These objects and identities are in turn located in an ongoing

political economy. A political economy is a complex, interconnected system. It structures the production, distribution, and consumption of wealth in a society. It determines the Who, What, When, Where, Why, and How of wealth and power in everyday life; that is, who gets what income, at what time, in what places, for what labor, and why?

This economy regulates the production, distribution and consumption of cultural objects. It does so by repeatedly forging links between cultural objects (cars, clothing, food, houses), their material representations, and the personal identities of consumers as gendered human beings (see discussion below). To summarize, an understanding of these five processes leads researchers to examine how cultural objects are represented in the media and in everyday popular culture. Scholars then move from this level of analysis to studies of those social identities that are attached to the cultural object, asking, at the same time, how is this object 'produced and consumed, and what mechanisms regulate its distribution and use' (du Gay et al. 1997). An instance is given in those moments when a cultural fad, a style of dress, or a hairstyle seemingly sweeps across an age or gender group. New and old identities are attached to the use of the object that is produced for consumption by a small number of manufacturers who regulate price and distribution.

2. *The Narrative Turn*

This concern for language and meaning has been called the narrative or discursive turn in CS. This turn implies greater interest in language, discourse, discursive practices, and the argument that meaning is contextual (Hall 1996a). This narrative turn moves in two directions at the same time. First, CS scholars formulate and offer various narrative versions or stories about how the social world operates. This form of narrative is usually called a theory, for example, Freud's theory of psychosexual development. Second, scholars study narratives and systems of discourse, arguing that these structures give coherence and meaning to the world. A system of discourse is a way of representing the world. A complex set of discourses is called a discursive formation (Hall 1996a). The traditional gender belief system in American culture, with its focus on patriarchy and a woman's place in the home, is an instance of a discursive formation. Discursive formations are implemented through discursive practices, for example, patriarchy and the traditional etiquette system.

Systems of discourse both summarize and produce knowledge about the world. These discursive systems are seldom just true or false. In the world of human affairs, truth and facts can be constructed in different ways. Consider this question, 'Are those Palestinians who are fighting to regain a home on the West Bank of Israel freedom fighters or terrorists? (Hall 1996b). The very words that are used to describe these individuals prejudge and evaluate their activity. Freedom fighter or terrorist are not neutral terms. They are embedded in competing discourses. As such they are connected to struggles over power, that is who has the power to determine which term will be used? As Hall notes, 'It is the outcome of this struggle which will define the "truth" of the situation.' Often times it is power 'rather than facts about reality, which makes things "true"' (Hall 1996b).

Power produces knowledge (Foucault 1980). Regimes of truth can be said to operate when discursive systems regulate relations of power and knowledge (Hall 1996b). The traditional gender belief system, which regulates the power relations between men and women in this culture is such a regime. In these ways discursive systems affect lives.

2.1 *Experience and its Representations*

Of course it is not possible to study experience directly, so CS researchers examine representations of experience, interviews, stories, performances, myth, ritual, and drama. These representations, as systems of discourse, are social texts, narrative, discursive constructions. The meanings and forms of experience are always given in narrative representations. These representations are texts that are performed, stories told to others. Bruner (1986) is explicit on this point, representations must 'be performed to be experienced.' In these ways researchers deal with performed texts, rituals, stories told, songs sung, novels read, dramas performed. Paraphrasing Bruner (1986), experience is a performance.

2.2 *Assessing Interpretations*

The politics of representation is basic to the study of experience. How a thing is represented often involves a struggle over power and meaning. While scholars have traditionally privileged experience itself, it is now understood that no life, no experience can be lived outside of some system of representation (Hall 1996c). Indeed, 'there is no escaping from the politics of representation' (Hall 1996c).

This narrative turn suggests that researchers constantly are constructing interpretations about the world, giving shape and meaning to what they describe. Still, all accounts, 'however carefully tested and supported are, in the end "authored"' (Hall 1996a). CS explanations reflect the point of view of the author. They do not carry the guarantee of truth and objectivity.

The narrative turn leads scholars to be much more tentative in terms of the arguments and positions they put forward. It is now understood that there is no final or authorized version of the truth. Still, there are criteria of assessment that should be used. Researchers are 'committed to providing systematic, rigorous, coherent, comprehensive, conceptually clear, well-

evidenced accounts, which make their underlying theoretical structure and value assumptions clear to readers...(still) we cannot deny the ultimately interpretive character of the social science enterprise' (Hall 1996a).

3. Diversity and Constraint

The open-ended nature of the CS project creates a perpetual resistance against attempts to impose a single, umbrella-like paradigm over the entire project. There are multiple CS projects, including: the articulation model of the Birmingham school, and the work of Stuart Hall, Paul Gilroy, Lawrence Grossberg, and their associates (du Gay et al. 1997, Grossberg 1997a, 1997b); Black British CS (Baker et al. 1996); the conjunctural, resistance perspective of Fiske (1994); feminist, ethnographic projects connected to the British and Frankfurt schools; American CS versions located in communication studies (Ferguson and Golding 1997), which work through pragmatism back to the arguments of McLuhan, Innis, Ong, and Dallas Smythe (Carey 1997a, 1997b, Denzin 1995); the African-American, prophetic, postmodern, neopragmatic marxism of West (1992); the empiricist, neofunctional American cultural sociology project which draws on the work of Bourdieu and Parsons (see Schudson 1997); a Chicana/o, Latina/o model that elaborates and then departs from the Birmingham approach (Chabram-Denersesian 1999); a critical, CS pedagogy (McCarthy 1998); Canadian and Australian CS models which focus on policy and public culture and attempt to de-Anglocize the British model (Blundell et al. 1993, Frow and Morris 1993); a Black feminist cultural criticism (Hooks 1990); an American-based, ethnographic, critical CS model centered on a resistance postmodernism; and more recently a transnational CS focused on the critical, ethnographic study of public culture and the flow of cultural forms and cultural representations from one site to another (*Public Culture* 1988).

The generic focus of each version of CS involves an examination of how the history people live is produced by structures that have been handed down from the past. Within these traditions, culture is treated as a verb, a process, and a place, always in motion, where meaning and situated identities connected to race, class, and gender are created and performed (du Gay et al. 1997). A shared emphasis on international cultural and social processes unites these various programs. This shared focus moves in four directions: (a) the 'detour through theory' (Hall 1992, Grossberg 1997a), linked (b) to the politics of representation, the circuits of culture (Hall 1996a), and the textual analyses of the media, literary, and cultural forms, including their production, distribution, and consumption; (c) the ethnographic, qualitative study of these forms in everyday life and the analysis of the social and communication processes that they shape and define; (d) the investigation of new pedagogical practices that interactively engage critical cultural analysis in the classroom (Grossberg 1997a). Each of these versions of CS is joined by a threefold concern with cultural texts, lived experience, and the articulated relationship between texts, materiality, and everyday life. Within the cultural text tradition some scholars (Fiske 1994) examine the mass media and popular culture as sites of resistance where history, ideology, and subjective experiences come together. These scholars produce critical ethnographies of the active audience in relation to particular historical moments. Other scholars read texts as sites where hegemonic meanings are produced, distributed, and consumed. In the feminist ethnographic tradition, there is a postmodern concern for the social text and its production (see Denzin 1997).

These models see culture as a series of ongoing interactional practices, conversations, talk, ways of acting, and representing the meanings of experience. Any cultural practice is significant because it is an instance of a cultural practice that happened in a particular time and place. This practice cannot be generalized to other practices, its importance lies in the fact that it instantiates a cultural practice, a cultural performance (story telling), and a set of shifting, conflicting, cultural meanings (Fiske 1994).

The researcher strategically selects sites for interpretation that constitute the intersection of texts and interacting individuals. Interactional specimens are extracted from these sites, written off from the conversations and actions that occur within them. This model always work upward and outward from the concrete to the larger set of meanings that operate in a particular context. It offers glimpses of culture in practice, setting one set of practices and meanings off against others that may compete in the same situation (Fiske 1994). The concept of structure is critical. In interpretive CS structure is a set of generative (often hegemonic), interactional, and cultural practices that organize meanings at the local level. Critical emic inquiry guides this process, the researcher seeks to understand a subject (or class of subjects) within a given historical moment. Sartre's (1963) progressive–regressive method is employed. A variant on the pragmatic emphasis on the consequences of acts, this method looks forward to the conclusion of a set of acts. It then works back to the conditions, interpretations, and situations that shape this decision. By moving forward and backward in time subjects and their projects are located within culture as a set of interpretive practices.

4. In Conclusion: The Calling of CS

Frow and Morris (2000) observe that contemporary versions of CS have been shaped by encounters between diverse feminisms, ethnic and critical race studies, gay, lesbian, and queer studies, postcolonial

and diasporic research, and with indigenous people's scholarship. These interactions produce a sensitivity to culture in its multiple forms, including the aesthetic, political, anthropological, performative, historical, and the spatial. Thus, there are collections and essays in Australian-Asian CS, Asian-Pacific CS, Latin-American, Mexican, and Chicana/o CS, as well as Black-British, Irish, British, Spanish, Italian, Nordic, and African CS.

At the same time, students of CS wrestle with the multiple meanings of such key terms as identity, place, globalization, the local, nationhood, and difference. These terms are debated constantly in the media, played out in arenas defined and shaped by the new information and communication technologies. CS scholars examine how meanings move between and within various media formations. This yields studies of Madonna, Elvis, the Gulf War, Anita Hill, the memory-work of museums, studies of tourism, shopping malls, and so on. In such work it becomes clear that culture is a contested, conflictual set of practices bound up with the meanings of identity and community.

The disciplinary boundaries that define CS keep shifting, and there is no agreed standard genealogy of its emergence as a serious academic discipline. Nonetheless, there are certain prevailing tendencies including: feminist understandings of the politics of the everyday and the personal; disputes between proponents of textualism, ethnography, and autoethnography; continued debates surrounding the dreams of modern citizenship.

But of course CS is a discursive category, its meanings are established in and through rhetorical and material practices. Nonetheless, the *sine qua non* of CS is irrefutable. A core belief defines the project, the understanding that scholars have a commitment and a responsibility to record and analyze, and register the meanings of 'the great events of our time for the benefit of future generations' (Nelson and Gaonkar 1996). This is the calling of CS.

Bibliography

Alasuutari P, Gray A, Hermes J 1998 Editorial. *European Journal of CS* **1**(1): 5–11

Appadurai A 1996 *Modernity at Large*. University of Minnesota Press, Minneapolis, MN

Baker H A Jr, Diawara M, Lindeborg R H (eds.) 1996 *Black British Cultural Studies: A Reader*. University of Chicago Press, Chicago

Blundell V, Sheperd J, Taylor I (eds.) 1993 *Relocating Cultural Studies: Developments in Theory and Research*. Routledge, London

Bruner E M 1986 Experience and its expressions. In: Turner V M, Bruner E M (eds.) *The Anthropology of Experience*. University of Illinois Press, Urbana, IL, pp. 3–30

Carey J W 1997a Afterword: The culture in question. In: Munson E S, Warren C A (eds.) *James Carey: A Critical Reader*. University of Minnesota Press, Minneapolis, MN, pp. 308–39

Carey J W 1997b Reflections on the Project of (American) Cultural Studies. In: Ferguson M, Golding P (eds.) *Cultural Studies in Question*. Sage, London, pp. 1–24

Chabram-Denersesian A 1999 Introduction: Chicana/o Latina/o Cultural Studies: Transnational and transdisciplinary movements. *Cultural Studies* **13**: 173–94

Denzin N K 1995 Information technologies, communicative acts, and the audience: Couch's legacy to communication research. *Symbolic Interaction* **18**: 247–68

Denzin N K 1997 *Interpretive Ethnography*. Sage, Thousand Oaks, CA

Ferguson M, Golding P 1997 Cultural Studies and changing times: An introduction. In: Ferguson M, Golding P (eds.) *Cultutal Studies in Question*. Sage, Thousand Oaks, CA, pp. xiii–xxvii

Fiske J 1994 Audiencing: Cultural practice and cultural studies. In: Denzin N K, Lincoln Y S (eds.) *Handbook of Qualitative Research*. Sage, Thousand Oaks, CA, pp. 189–98

Foucault M 1980 Power/knowledge: Selected interviews and other writings 1972–1977. In: Gordon C (ed.) (Trans. Gordon C, Marshall L, Mepham J, Soper K). Pantheon, New York

Frow J, Morris M (eds.) 1993 *Australian Cultural Studies: A Reader*. University of Illinois Press, Urbana, IL

Frow J, Morris M 2000 CS. In: Denzin N K, Lincoln Y S (eds.) *Handbook of Qualitative Research*. Sage, Thousand Oaks, CA

du Gay P, Hall S, James L, Mackay H, Negus K 1997 *Doing Cultural Studies: The Story of the Sony Walkman*. Sage, London

Grossberg L 1997a *Bringing it all back Home: Essays in Cultural Studies*. Duke University Press, Durham, NC

Grossberg L 1997b *Dancing in Spite of Myself: Essays on Popular Culture*. Duke University Press, Durham, NC

Grossberg L 1998 The cultural studies crossroads blues. *European Journal of Cultural Studies* **1**: 65–82

Hall S 1980 Cultural studies: two paradigms. *Media, Culture and Society* **2**: 57–72

Hall S 1992 Cultural Studies and its theoretical legacies. In: Grossberg L, Nelson C, Treichler P (eds.) *Cultural Studies*. Routledge, New York, pp. 277–94

Hall S 1996a Introduction. In: Hall S, Held D, Hubert D, Thompson K (eds.) *Modernity: An Introduction to Modern Societies*. Blackwell, Cambridge, MA, pp. 3–18

Hall S 1996b Gramsci's relevance for the study of race and ethnicity. In: Morley D, Chen K-H (eds.) *Stuart Hall: Critical Dialogues in Cultural Studies*. Routledge, London, pp. 411–44

Hall S 1996c What is this 'black' in black popular culture? In: Morley D, Chen K-H (eds.) *Stuart Hall: Critical Dialogues in Cultural Studies*. Routledge, London, pp. 465–75

Hartley J 1998 Editoral (with goanna). *International Journal of Cultural Studies* **1**(1): 5–10

Hooks B 1990 *Yearning: Race, Gender, and Cultural Politics*. South End Press, Boston

Long E 1997 Introduction: engaging sociology and cultural studies: disciplinarity and social change. In: Long E (ed.) *From Sociology to Cultural Studies*. Blackwell, Malden, MA, pp. 1–32

McCarthy C 1998 *The Uses of Culture*. Routledge, New York

Nelson C, Gaonkar D P 1996 Cultural studies and the politics of disciplinarity. In: Nelson C, Gaonkar D P (eds.) *Disciplinarity and Dissent in Cultural Studies*. Routledge, New York pp. 1–22

1988 Editor's comments. *Public Culture* **1**: 1–5
Sartre J-P 1963 *Search for a Method*. Knopf, New York
Schudson M 1997 Cultural politics and the social construction of 'social construction': Notes on teddy bear patriarchy. In: Long E (ed.) *From Sociology to Cultural Studies*. Blackwell, Malden, MA, pp. 379–98
Striphas T 1998 Introduction: The long march: cultural studies and its institutionalization. *Cultural Studies* **12**: 453–75
Storey J 1996 CS: An introduction. In: Storey J (ed.) *What Is Cultural Studies? A Reader*. Arnold. London, pp. 1–13
West C 1992 The postmodern crisis of the Black intellectuals. In: Grossberg L, Nelson C, Treichler P (eds.) *Cultural Studies*. Routledge, New York, pp. 689–716
Williams R 1976 *Keywords*. Fontana, London

N. K. Denzin

Cultural Studies of Science

'Cultural studies of science' marks a loosely amalgamated, multidisciplinary research field drawing from history, anthropology, feminist theory, sociology, and philosophy of science. Its participants emphasize that scientific practices are historically situated, meaningful patterns of interaction with the world. Cultural studies offer an interpretive, critical engagement with scientific practices, rather than an explanation of their outcome. Cultural studies are situated within the 'cultures of science' that they investigate, and are responsive to what is at stake in the ongoing development of scientific practices.

1. The Place of Cultural Studies within Science Studies

Cultural studies of science first emerged in the late 1980's and early 1990's in response to debates between proponents of internalist history and philosophy of science, and adherents of the 'Strong Program' in the sociology of science. These debates concerned how best to explain the content of scientific knowledge, either by appeal to rational norms or as the outcome of particular social interactions. The stakes in these debates over explanatory strategies seemed to concern the authority of scientific claims, their relation to the natural world, and the normative stance of the explanation itself. With respect to scientific authority, philosophical accounts typically vindicated the acceptance of scientific claims or research programs as rationally warranted or reliable, whereas constructivist sociological work emphasized the contingency of scientific work and the impossibility of rationally foreclosing alternative interpretations. Meanwhile, scientific realists argued that scientific claims were accountable to the world as independent of human language, experience, or social practice, while antirealists (including both sociological constructivists and philosophical empiricists and historical rationalists) insisted that the sciences could never reach beyond our categories, practices, and norms. Finally, classically normative accounts of science aspired to assess the adequacy of scientific practices according to philosophical norms, while philosophical and sociological naturalists proposed primarily to describe (and perhaps explain) how the sciences actually proceed.

Work in cultural studies proceeded differently, in ways that rejected assumptions shared by most participants in these debates between internalists and social constructivists (assumptions that are also implicit in classic works of social history, such as Forman 1971 or Shapin and Schaffer 1985, that are otherwise thematically congenial to cultural studies). First, cultural studies are strongly antiessentialist about the sciences: there are no goals, methods, products, or norms belonging to science as such. Such antiessentialism undercuts the explanatory projects of both internalist philosophical histories and of a programmatic social constructivism, which are committed to accounting for the sciences as knowledge-producing activities (where 'knowledge' is a distinctive kind of product, either rationally warranted, causally reliable, or socially authoritative belief). The sciences are historically evolving practices, such that past and present practice cannot fix their nature, significance, aim, or 'content.'

Cultural studies also find no clear boundaries between scientific content and its social context. Both philosophers and sociologists have identified the shared commitments and norms of scientific communities as the locus of their divergent explanatory projects. Cultural studies have instead emphasized both an 'internal' heterogeneity that belies any presumption of community within a scientific field, and an active traffic in meanings and material across any supposed boundaries between scientific communities and their social or cultural contexts. This absence of community-defining consensus also exemplifies a more fundamental issue: cultural studies take the subjectivity of scientific knowers to be part of what needs to be understood rather than as an explanatory resource. Neither the natural body, the rational faculties, or the social identity of a scientist can be taken for granted as the locus or source of agency and belief. Prosthetic embodiments, narrative re-enactments, discursive interactions, and the affordances of material surroundings instead constitute knowing subjects through scientific practices, rather than accounting for practices as the doings of human agents/believers.

The opposing positions in the realism debates thereby dissolve. The question whether human knowers can ever get beyond the confines of their mental or social representations in order to be accountable to natural objects presupposes the autonomous intelligibility of mental or social representa-

tions. Cultural studies take scientific practices to be always already materially engaged and discursively articulated, so that there is no confinement within experience, language, theory, or social world to be transcended. On the contrary, experience, language, theory, and social forms themselves only take on determinate form through ongoing patterns of 'intra-action' in the material world (Barad 1996 coins this term to avoid the connotation that 'interactions' involve agencies definable apart from their interaction). But the same is true of 'nature,' whose objects and capacities only acquire determinate articulation in phenomena, ongoing patterns of 'material-semiotic' intra-action (Barad 1996, Rouse 1996, Haraway 1999).

The dispute between normative and naturalistic (descriptive) approaches to scientific practices also turns out to share false presuppositions. Cultural studies agree with naturalists that there can be no standpoint 'outside' of ongoing scientific practices from which to articulate norms to which they are accountable. Yet they disagree that science studies must therefore simply describe (and perhaps explain) the norms that function within those practices. The normative accountability of scientific practices is not a determinate fact or regularity about them (or about scientific 'communities'), but is instead constitutive of their intelligibility to practitioners and interpreters alike.

Scientific practices only make sense in terms of what is at issue and at stake in those practices. Yet these issues and stakes are themselves contested within the practice, and its interpretation. Cultural studies are not neutral descriptions of what scientists do, but interpretive construals of those doings that may thereby partially reshape the field of subsequent practice. In that respect, cultural studies are continuous with scientific work, which likewise reinterprets what is at issue and at stake through re-enacting the practices it participates in (Rouse 1996).

Cultural studies thereby articulate dynamic, expressive conceptions of meaning, knowledge, and power, which contrast sharply with the standard approaches to these phenomena within philosophy and social theory (Rouse 1996, 1999). On such accounts, meaning is not a property of utterances or actions; the term 'meaning' instead articulates the ways in which such performances inferentially draw upon and transform the field of prior performances in which they are situated. Knowledge is, likewise, not a theoretically coherent kind of representation; the term 'know' is used expressively in ongoing practices of assessing, attributing, relying upon, and contesting understanding and justification. Power, in turn, is not a substantial component of social interaction, something possessed and exercised by some agents, and lacking for others. The term expresses the ways in which one action aligns with others over time to affect the field of subsequent possible activity: 'power' articulates the connections between causal efficacy and intelligibility. Cultural studies of science thus expressively articulate the semantic, epistemic, and effective roles of scientific practices.

2. *Prominent Themes*

The preceding account offers a broadly philosophical interpretation of the aims and commitments embodied in the wide range of historical, anthropological, and sociological investigations that exemplify cultural studies of science. Typically, however, these commitments are expressed through detailed studies of particular practices, performances, or institutions rather than as explicit philosophical theses. A representative list of prominent work in cultural studies of science would include work directed toward particular laboratories, field sites, and other settings (e.g., Haraway 1989, Heath 1997, Traweek 1992, Biagioli 1993), images, models, and metaphors (e.g., Bono 1995, Haraway 1992, Keller 1995, Martin 1994, Daston and Park 1998), and machines and procedures (e.g., Barad 1998, Galison 1997, Rheinberger 1997, Daston and Galison 1992), all of which involve patterns of exchange between scientific settings and other sites of meaning-articulation. These topics have been marked as productive foci of inquiry not because they are relatively localized or self-contained, but because they have been productive nodes in the circulation and transformation of meaning, knowledge, and power around potent and often contested issues.

'The body' has been among the most prominent of these issues taken up in detailed work by cultural studies scholars, perhaps in part because of their insistence upon the embodied, materially interactive dimension of scientific practices. Attention has been directed toward bodily disciplines and skills (both as constitutive of scientific knowing, and as impositions upon who or what is to be known); toward prosthetic extension or reorganization of bodily engagement with the world (including the immunological internalization of the boundary between bodily self and external others); toward bodies as raced and gendered, and otherwise differentially marked (Heath 1997, for example, takes up the distinctive bodily markings and significations of Marfan's Syndrome); toward animal bodies, microbial 'bodies,' and their complex material-semiotic relations with human embodiment; and toward various processes and apparatuses of bodily reproduction. Along with 'the body,' other meanings partially constitutive of scientific knowing have also become issues for cultural studies of science. Authorship and authority, objectivity in its multiple contested deployments, the construction of professional and disciplinary identities within and around the sciences, and the affective dimensions of scientific understanding (e.g., wonder, anxiety, love, empathy, and various forms of humor) have been among the epistemically

significant issues that have provided a focus for cultural studies of science. Cultural studies have also been attentive to many of the ways in which the objects of scientific understanding have been meaningfully articulated. Genes, organisms, cyborgs, blood, information, life, matter, viruses, origins, secrets, foundations, environment, and ultimately nature in its significant differentiation from culture, society, artifact, and history have been prominent themes in both scientific practices and cultural studies of science.

In rejecting global legitimations or critiques of science as justified knowledge, social consensus, empirically reliable methods, cultural tradition, or approximately true theories, cultural studies of science have attempted to set aside the project of accounting for what science really (essentially) is, or how science is actually done, in favor of participatory engagement with specific scientific practices. The sciences are not regarded as optional practices (or belief-systems), of which it could even make sense to be for or against, credulous or skeptical, as a whole. Scientific practices are integral to the formation of a field of intelligible possibilities for action and self-understanding, partially constitutive of any conceptualization of who we are and what our situation is. To engage successfully in cultural studies of science is to clarify, and thereby partially transform, that field of possibilities and our involvement within it. Ultimately, the stakes in cultural studies of science are located in the notion of responsibility. To whom, to what, and in which respects are the sciences and science and technology studies responsible? How might such responsibility be exercised, and even partially fulfilled, through scientific research in particular settings, and through cultural studies of specific scientific or technological practices? What criteria or concerns should govern the assessment of responsible and irresponsible science and science studies? With these questions in mind, cultural studies become not simply a specifiable and in-principle exhaustible domain of academic inquiry, but an ongoing project of responsible engagement within the cultures of science.

Bibliography

Barad K 1996 Meeting the universe halfway. Realism and Social Constructivism without Contradiction. In: Nelson L H, Nelson J (eds.) *Feminism, Science and Philosophy of Science*. Kluwer, Dordrecht, The Netherlands, pp. 161–94

Barad K 1998 Getting real: technoscientific practices and the materialization of reality. *Differences* **10**: 87–126

Biagioli M 1993 *Galileo Courtier: The Practice of Science in the Culture of Absolutism*. University of Chicago Press, Chicago

Biagioli M 1999 *Science Studies Reader*. Routledge, New York

Bono J 1995 *The Word of God and the Languages of Man*. University of Wisconsin Press, Madison, WI

Daston L, Park K 1998 *Wonders and the Order of Nature, 1150–1750*. MIT Press, Cambridge, MA

Daston L, Galison P 1992 The image of objectivity. *Representations* **40**: 81–128

Forman P 1971 Weimar culture, causality and quantum theory 1918–1927. In: McCormmach R (ed.) *Historical Studies in the Physical Sciences*. University of Pennsylvania Press, Philadelphia

Galison P L 1997 *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, Chicago

Haraway D J 1989 *Primate Visions: Gender, Race, and Nature in the World of Modern Science*. Routledge, New York

Haraway D J 1992 The promises of monsters. In: Grossberg L, Nelson N, Treichler P A (eds.) *Cultural Studies*. Routledge, New York, pp. 295–337

Haraway D J 1999 Situated knowledges: The science question in feminism and the privilege of partial perspective. In: Biagioli M (ed.) *The Science Studies Reader*. Routledge, New York, pp. 72–88

Heath D 1997 Modest interventions. In: Downey G L, Dumit J (eds.) *Cyborgs and Citadels: Anthropological Interventions in Emerging Sciences and Technologies*. School of American Research, Santa Fe, CA, pp. 67–82

Keller E F 1995 *Refiguring Life: Metaphors of Twentieth Century Biology*. Columbia University Press, New York

Latour B 1992 One more turn after the social turn. In: McMullin E (ed.) *The Social Dimensions of Science*. University of Notre Dame Press, Notre Dame, IN, pp. 272–94

Martin E 1994 *Flexible Bodies: Tracking Immunity in American Culture from the Days of Polio to the Age of Aids*. Beacon Press, Boston

Rabinow P 1996 *Essays on the Anthropology of Reason*. Princeton University Press, Princeton, NJ

Rheinberger H-J 1997 *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford, CA

Rouse J 1996 *Engaging Science: How to Understand its Practices Philosophically*. Cornell University Press, Ithaca, NY

Rouse J 1999 Understanding scientific practices. In: Biagioli M (ed.) *The Science Studies Reader*. Routledge, New York, pp. 442–56

Shapin S, Schaffer S 1985 *Leviathan and the Air Pump: Hobbes, Boyle, and the Experimental Life: Including a Translation of Thomas Hobbes, Dialogus Physicus de Natura Aeris by Simon Schaffer*. Princeton University Press, Princeton, NJ

Traweek S 1988 *Beamtimes and Lifetimes: The World of High Energy Physicists*. Harvard University Press, Cambridge, MA

Traweek S 1992 Border crossings. In: Pickering A (ed.) *Science as Practice and Culture*. University of Chicago Press, Chicago, pp. 429–65

J. Rouse

Cultural Variations in Interpersonal Relationships

Interpersonal relationships refer to all kinds of social interactions. In the social sciences, and particularly in psychology, the positive side of interpersonal relation-

ships has been studied under such rubrics as attachment, friendship, couple and marital relationships, close relationships, and social networks. The negative side of interpersonal relationships has been highlighted through studies of aggression, conflict, loneliness, alienation, and so forth. This article focuses on positive interactions and relationships.

1. *Commonalities and Differences Across Cultures*

It has been universally documented that in every culture, human beings need other people from the cradle to the grave, because they are social by nature. Humans develop interpersonal relationships with multiple significant others, such as family members, friends, teachers, romantic partners, and respected persons throughout their lives. Moreover, research indicates that through interpersonal interactions, people receive emotional support, enjoy their lives, and learn. That is, it has been found that interpersonal relationships not only enhance the quality of life, but also are essential for human beings to maintain a basic level of functioning.

At the same time, it is widely observed that the nature of at least some interpersonal relationships varies with cultures and subcultures, because each culture in which people participate encourages and sometimes discourages certain social behaviors. For example, in some cultures, for instance, among modern, white, upper-middle-class males, articulating the need to have close relationships is suppressed; these men are encouraged to be emotionally independent of others. In this culture, in comparison with females, men are inhibited from expressing the need to have close relationships.

2. *Attachment Relationships in Infancy*

The onset of interpersonal relationships is clearly observed in the infant–caregiver affective bonds, that is, attachment relationships. Around the first birthday, infants everywhere are upset when they are separated from their primary caregiver, usually the mother. Bowlby (1969/1982), who proposed attachment theory, observed that many of the human infants' behaviors are organized for maintaining proximity to a caregiver, and hypothesized as an evolutionary principle that infants are biologically predisposed to stay close to their mothers for survival. According to Bowlby and his colleagues, an attachment is a special type of affective bond with a specific person who is not interchangeable with anyone. Attachment theorists view an attachment as a highly evolved behavioral system that produces the felt-security state in the infant, that is, the belief that an attachment figure will efficiently respond if he or she asks for help (see *Attachment Theory: Psychological*).

The quality of attachment to the mother among infants and young toddlers is assessed by a series of episodes in a laboratory, known as the Strange-Situation procedure (Ainsworth et al. 1978). The basic rationale of this procedure is that the quality of attachment is observed when the infant is placed under mild stress. As the procedure was originally designed for mother–infant pairs from American white middle-class families based on their child-rearing practices, this mild stress is induced by encounters with an adult stranger and by being left alone in the unfamiliar laboratory. By the procedure, infants are classified into one of three main patterns of attachment: secure (B-type, 67 percent in the original sample); anxious/avoidant (A-type, 21 percent); or anxious/resistant (C-type, 12 percent).

This procedure, constructed in a particular society, has been applied without adaptation to a wide variety of cultures on almost every continent, and sheds a unique light on cultural diversity in child-care practices. Based on meta-analysis of 2,000 classifications in eight different countries, van IJzendoorn and Kroonenberg (1988) identified samples whose ABC distributions were different from those reported by Ainsworth and others. These unexpected findings, all from outside North America, opened the door to further research aimed at understanding the differences.

For example, in Bielefeld, a city in northern Germany, a majority of toddlers were identified as A-type (52 percent), although they had no apparent psychopathological difficulties (Grossmann et al. 1985). In contrast, both in Sapporo, a city in northern Japan (Takahashi 1986), and in a traditional Israeli kibbutz (Sagi et al. 1985), a larger number of the toddlers were classified into C-type (32 and 34 percent, respectively) than in the original American sample. Takahashi (1990) pointed out that the appropriate level of stress in the procedure must vary with everyday child-care practices. That is, as the Japanese have long shared a belief that any stress for young children disturbs their healthy development, young children are virtually never left alone. In that culture, the procedure, especially the infant-alone episode, was so stressful that some of the Japanese children and also their mothers were pushed from B-type to C-type by the procedure. However, what triggered strong stress in kibbutz-reared toddlers was the appearance of a stranger. Sagi and his colleagues maintained that everyday kibbutz child-care practices that were very different from those of conventional families must have combined to produce the difference. The Grossmann et al. suggested that the toddlers in northern Germany were so strongly encouraged to be autonomous and discouraged from staying close to the mother that the children avoided interacting with the mother even in the experimental procedure.

3. Interpersonal Relationships in Lifespan Development

Attachment theorists have focused mainly on the dyadic affective relationships between infants and their primary figure, usually the mother. However, the relationships highlighted by the attachment are too limited to describe real-life experiences in interpersonal interactions, for at least the following two reasons (Takahashi and Sakamoto 2000).

First, it is often observed that, in every culture, humans from birth interact with not only the focal, attached figure, but also with other persons, such as the father, siblings, grandparents, and peers. It is well documented that through expansion of social arenas from family to other social institutions, i.e., day-care centers, schools, and workplaces, humans continuously expand and change their interpersonal relationships. Second, humans need others who fulfill different kinds of psychological functions, not only those crucial for survival, which have been focused on by attachment theorists, but also peripheral but enjoyable functions, such as receiving information and knowledge from, sharing experiences and emotions with, and giving nurture to, others. It may reasonably be assumed that human beings need multiple significant others concurrently, because it is difficult for even the most important figure to fulfill all of the psychological functions required by individuals.

In fact, there is plenty of evidence that shows the differentiation of psychological functions among multiple significant others. The multiple interpersonal relationships are described under the rubrics of social networks, social support networks, or personal frameworks of social relationships (see *Social Relationships in Adulthood*).

Kahn and Antonucci (1980) asked their subjects to map important persons in their lives and to classify them into three concentric circles, because they had hypothesized a hierarchical social network based on the image of a convoy that surrounds a person over time. This hierarchical network questionnaire, applied to every culture, has universally revealed that people, from young children to elderly persons of both genders, designate about 10 figures in their convoy. Moreover, in every culture, the percentage of family members mapped in the inner circle is higher than that of nonfamily members, and the figures who are placed in the inner circle provide higher proportions of all kinds of support, such as confiding, reassurance, and caring, whereas figures in the outer circle lend mostly peripheral support. Until the end of life, human beings maintain their own convoys, although the size of the social network decreases and the main figure changes with age. These studies have successfully made a rough but clear sketch of the universal nature of social relationships consisting of multiple figures.

At the same time, there have been many studies that indicate that social contexts affect interpersonal behaviors and relationships. However, careful, methodological examinations are needed before we can accept that they reveal real cultural differences in interpersonal relationships. First, differences in cross-cultural research may reflect response styles more than the differences in human relationships across cultural groups. For example, contrary to popular notions of interdependence, it is reported that American people at every stage of life from adolescence to old age exhibit their affective needs towards others more strongly than do their Japanese counterparts (Takahashi and Ohara 1997). In Taiwan, people are more reluctant to describe interpersonal relationships positively to an investigator, a third person, than are their white American counterparts (Miller et al. 1997). Second, observed differences could be due to people's different interpretations of terms and tasks. The concept of friend is a good illustration (Krappmann 1990). A cross-cultural comparison of friend relationships suggested that the connotation of the term 'friend' among Japanese elementary-school children was wider than that among German peers. The German children selectively nominated nine friends on average, whereas the Japanese students named most classmates as friends (mean = 23), and refrained from excluding any agemates from their group of friends if they had no appropriate reasons to be excluded. This attitude towards agemates among the Japanese children was attributed to the culture in the schools, where children were trained to establish 'nice' relationships with all of their agemates (see *Cross-cultural Psychology*).

To overcome these response biases, in a cross-cultural study between Japan and the USA, our group made relative comparisons between the two cultures. Instead of direct comparisons of the scores of each of five significant figures (i.e., mother; father; friend; romantic partner/spouse; and child), distributions of the dominant figures (that is, those who were assigned the highest score by each individual) were compared between the cultures. Both commonalities and diversities emerged. In both countries, the parents were rarely assigned the dominant role from adolescence on; rather, a same-sex friend—and later a romantic partner—was consistently and highly rated. Also, in both cultures, parents rated their child very highly when the child was grown-up. However, the number of family-dominant participants was larger among the Japanese than their American counterparts.

4. Individualism vs. Collectivism

Social and behavioral scientists may be tempted to attribute the observed cultural variations in interpersonal relationships to more global cultural differences in values and orientations, such as individualism vs. collectivism, or independence vs. interdependence (e.g., Markus and Kitayama 1991). In

such a dichotomous framework, cultures in the West, typically in North America, are supposed to encourage people to be more independent of others than cultures in the East, typically in Japan. However, some researchers have questioned the validity of such a dichotomy. For example, Takano and Osaka (1999) reanalyzed 15 empirical studies performed in the framework of individualism vs. collectivism, and concluded that the research did not provide sufficient evidence for the dichotomous classifications of cultures.

5. Conclusion

Thus, cultural and subcultural comparisons that respect the native group's perspective advance our understanding of not only differences among people's social relationships, but also the universal nature of human beings and their social relationships with others. Moreover, longitudinal research within each culture will help our understanding of how social and historical factors cause changes in social relationships. Another important challenge for future researchers is to investigate social relationships beyond close and intimate others. Globalization highlights the growing necessity for research concerning positive social relationships with unfamiliar people.

Bibliography

Ainsworth M D S, Blehar M C, Waters E, Wall S 1978 *Patterns of Attachment: A Psychological Study of the Strange Situation*. Erlbaum, Hillsdale, NJ

Bowlby J 1969/1982 *Attachment and Loss: Volume 1. Attachment*. Basic Books, New York

Grossmann K, Grossmann K E, Spangler G S, Suess G, Unzner L 1985 Maternal sensitivity and newborns' orientation responses as related to quality of attachment in Northern Germany. *Monographs of the Society for Research in Child Development* **50**(1–2): 233–56

Kahn R L, Antonucci T C 1980 Convoys over the life course: Attachment, roles, social support. In: Baltes P B, Brim O B (eds.) *Lifespan Development and Behavior*. Academic Press, New York, Vol. 3, pp. 253–68

Krappmann L 1990 Friendship conception and friendship performance of six- through fifteen-year-old children. Paper presented at the Fourth European Conference of Developmental Psychology

Markus H R, Kitayama S 1991 Culture and self: Implications for cognition, emotion and motivation. *Psychological Review* **98**: 224–53

Miller P J, Wiley A R, Fung H D, Liang C-H 1997 Personal storytelling as a medium of socialization in Chinese and American Families. *Child Development* **68**: 557–68

Sagi A, Lamb M E, Lewkowicz K S, Shoham R, Drvir R, Estes D 1985 Security of infant–mother, –father, –metapelet attachments among kibbutz-reared Israeli children. *Monographs of the Society for Research in Child Development* **50**(1–2): 257–75

Takahashi K 1986 Examining the strange-situation procedure with Japanese mothers and 12-month-old infants. *Developmental Psychology* **22**: 265–70

Takahashi K 1990 Are the key assumptions of the 'strange situation' procedure universal? A view from Japanese research. *Human Development* **33**: 23–30

Takahashi K, Ohara N 1997 Are the Japanese more independent?: A cross-cultural perspective on social relationships. Paper presented at the meeting of the Society for Research in Child Development, Washington, DC

Takahashi K, Sakamoto A 2000 Assessing social relationships in adolescents and adults: Constructing and validating the affective relationships scale. *International Journal of Behavioral Development* **24**: 451–63

Takano Y, Osaka E 1999 An unsupported common view: Comparing Japan and the U. S. on individualism/collectivism. *Asian Journal of Social Psychology* **2**: 311–41

van IJzendoorn M H, Kroonenberg P M 1988 Cross-cultural patterns of attachment: A meta-analysis of the strange situation. *Child Development* **59**: 147–56

K. Takahashi

Culture and Economic Development: Cultural Concerns

Economic development presupposes not just the existence of formal institutions like property rights and a rule of law under which buyers and sellers can exchange goods in markets, but also certain norms or social values that promote exchange, savings, and investment. Thus, there is a cultural dimension to economic behavior. Culture has been given a wide variety of definitions, but it will be used here to signify the informal shared values, norms, meanings, and behaviors that characterize human societies. Just how important the cultural component of development is has long been a subject of controversy within the social sciences.

1. The Controversy over Culture

Modern neoclassical economics tends to downplay the importance of culture to development. Economists make the simplifying assumption that human beings are rational utility-maximizing individuals, and that such maximizing behavior is largely invariant across different human societies. The standard economic growth model pioneered by Robert Solow looks only at inputs of capital and labor; more recent so-called 'endogenous' growth models emphasize the role of

technology. From this perspective, culture constitutes at most a kind of residual factor that one appeals to when other explanations fail.

Sociologists, on the other hand, have tended to believe that cultural norms pervade economic life and that the latter cannot be understood apart from them. (Granovetter 1985) Emile Durkheim, for example, argued that the markets described by economists themselves presuppose shared norms (e.g., that people exchange goods rather than trying to rob and murder one another), and that sociology therefore looks to a deeper level of causation than economics. Max Weber (1930) pointed out that the economist's assumption that raising the piece-rate would increase output had the opposite effect in certain peasant societies, since workers, valuing leisure over accumulation, would be able to quit work earlier in the day. His life's work centered on showing how the emergence of the modern economic world depended on a prior shift in cultural values having to do with Protestantism. He in effect stood Marx on his head by arguing that what the latter labeled ideological 'superstructure' was in fact a key shaper of the means of production.

Today, the impact of culture on development tends to be studied more by economic sociologists than by economists. The one important exception to this generalization are development economists, that is, economists who specialize in the problems of growth in less-developed, premodern, or non-Western societies. They tend to see that many behaviors that can be taken for granted in the US or Western Europe, e.g., that public officials will not be grossly corrupt or incompetent, are not necessarily prevalent in other societies (Harrison 1992).

2. *History of the Development of Economic Sociology*

2.1 *The Weber Hypothesis*

The *locus classicus* of studies of culture and development is Weber's *The Protestant Ethic and the Spirit of Capitalism*, first published in 1904. Weber argued that the Calvinist doctrine of predestination, rather than inducing quietism, led believers to seek to demonstrate their status as elect by engaging in commerce and worldly accumulation. According to Weber, Puritanism created a work ethic—that is, the valuing of work for its own sake rather than for its results—and demolished the Aristotelian–Catholic doctrine that one should acquire only as much wealth as one needed to live well. In addition, it enjoined on its believers moral behavior outside the boundaries of the family, which was crucial in creating a more impersonal system of social trust.

The Weber thesis was controversial from the moment it was published, with figures like R. H. Tawney, Ernst Troeltsch, and Talcott Parsons weighing in. Various scholars argued that Weber was wrong empirically about the superior economic performance of Protestants over Catholics; that Catholic societies had started to develop modern capitalism long before the Reformation; that it was the Counterreformation rather than Catholicism *per se* that led to economic backwardness; and that the shift towards modern notions of accumulation was a more rational process than Weber believed. In the end, however, most scholars believe that Weber essentially was correct, given the delayed development trajectory of virtually all Catholic countries when compared with their Protestant counterparts.

2.2 *The Decline of Cultural Explanations*

At the same time, Weber's work gives us reason to be cautious about applying cultural explanations for economic development too readily. The German sociologist's other great work on culture and development was his study that was translated as *Confucianism and Taoism* (Weber 1951). Weber was seeking to answer the question why capitalism arose first in the Protestant West rather than in other societies. He argued that Confucianism created an environment hostile to capitalist development: while it was a rational ethical system, it emphasized the 'sib' or kinship as the primary source of social relatedness and thereby promoted economically inefficient nepotism. Japan, according to Weber, was even less suited than China to produce modern capitalism. In light of the astonishing economic performance of Japan, China, and other Confucian societies in East Asia after World War II, one is forced to conclude that the obstacles to development in that part of the world had much more to do with politics and inappropriate institutions than any specific cultural factor. Weber was in fact correct in pointing to the centrality of kinship as an organizing principle in Chinese society—businesses have been and continue to be family centered—but greatly overestimated the impact this would have on aggregate growth.

Following Weber, there were a number of studies up through the 1950s that emphasized cultural variables in economic development. Werner Sombart in *The Jews and Modern Capitalism* (1911) argued that Jews possessed the same cultural characteristics pointed to by Weber that were conducive to capitalism. A similar argument was for Japan by Robert Bellah, who argued that Buddhist sect founded by Ishida Baigan in the sixteenth century supported a functional equivalent of the Protestant work ethic.

Towards the middle of the twentieth century, authors including Everett E. Hagen, W. Arthur Lewis, and David C. McClelland all argued in different ways that certain less-developed societies lacked cultural characteristics (such as McClelland's 'achievement

orientation') that constituted obstacles to development. The 'modernization theory' popular in American social science during the early postwar period tended to see traditional cultures in a negative light, and to regard contemporary Western societies as models to which they needed to aspire.

This kind of argument fell out of favor during the next generation for a number of reasons. The first was the charge of ethnocentrism: many culturalist arguments tended to assume uncritically that social norms rather than institutions or structural conditions were responsible for lack of development, and that Western cultural values were superior to non-Western ones. This period saw the rise of neo-Marxist doctrines like *dependencia* theory that pointed to the structure of the world economy as the source of underdevelopment.

But a second strand of criticism came from orthodox economics. Cultural factors are, methodologically, very difficult to measure and to disentangle from other kinds of variables. As the Weber example above indicates, Asian underdevelopment could be explained not just by culture, but also by political conditions, poor economic policy, weak institutions, global economic conditions, and a host of other factors. Those promoting culturalist interpretations usually had no empirically convincing way of demonstrating that cultural factors were indeed as important as they claimed. Given that culture tends to change relatively slowly, it would seem very implausible that Asia's rapid rise in the second half of the twentieth century should be due primarily to cultural factors.

2.3 The Revival of Cultural Interpretations

In the 1980s and 1990s there has been something of a revival of culturalist interpretations of development. Part of the impetus for this came from within economics itself, with the rise of the subdiscipline of the so-called 'new institutional economics' associated with economic historian Douglas North. The new institutionalists recognized the importance of norms in economic life; according to North (1990), 'institutions' (i.e., formal or informal rules) were critical in reducing transaction costs and thereby promoting economic efficiency. Without agreed rules concerning property rights, for example, there could be no modern economic world as such since innovators would not have the incentive to take risks or make investments. While they often seek to give rational, maximizing accounts of the origins of institutions, institutional economists as a group are much more aware of the importance of history, culture, tradition, and other so-called 'path dependent' factors in shaping economic behavior.

The second reason for the revival of interest in culture and development was the experience of transitional economies in the late 1980s and 1990s, as well as observation of the so-called 'Asian Miracle.' Following the collapse of communism, many Eastern European and Soviet successor states set up formal market institutions, along with other institutions associated with developed market democracies such as democratic constitutions and electoral systems. Countries like Poland, Hungary, and the Czech Republic made the transition from centrally-planned to market-oriented economies relatively easily. Others, like Russia and Ukraine, had a much more difficult time; institutions were weak and levels of corruption were high. It seemed to many people unlikely that the difference in transition outcomes was simply the result of differing designs of formal institutions, particularly since the outcomes largely followed what one would have predicted on the basis of cultural factors (i.e., the degree of Westernization of each country prior to communism). As a result, many of the economic policy practitioners in international financial institutions like the World Bank and the IMF began to look towards cultural factors as key variables explaining successful transition strategies.

Similarly, rapid economic growth in Asia spawned many theories of how development there was driven by uniquely Asian cultural characteristics like a work ethic or deference to state authority. As noted above, Asian growth is better explained by institutional factors, and much of this cultural theorizing was undercut to some extent by the Asian economic crisis of 1997–1998.

3. Culture and Economic Behavior

In the large literature on culture and development, cultural factors are said to affect economic behavior in at least four ways: through its impact on organization and production, through attitudes towards consumption and work, through the ability to create and manage institutions, and through the creation of social networks.

3.1 Culture and Production

There is a large literature on 'organizational culture,' which studies norms and behaviors characteristic of individual organizations or even subunits within organizations. Organizational culture varies substantially across national cultures (e.g., the organization of a Japanese factory compared with a British one; see Dore 1973), but also between regions, sectors, or even companies within a single sector. Firms can be more or less hierarchical, open to outside influences, flexible in decision-making, and the like, for reasons having less to do with their formal organization than the informal norms that characterize them.

Social classes are also characterized by distinct social norms that give them differing characteristics with respect to class solidarity, attitudes towards money making and education, and the like. One argument for Britain's relative economic decline in the twentieth century, for example, had to do with the absorption by the English middle class of upper-class values that disdained practical education and technology.

3.2 Attitudes towards Consumption and Work

In addition to the classic literature on the work ethic, there has been substantial writing on the so-called 'culture of poverty,' a phrase coined by Oscar Lewis in the 1960s to characterize the failure of the poor in Latin America to take advantage of work opportunities. This term was used extensively in the debate over poverty and welfare in the US, where observers like Daniel P. Moynihan and William Julius Wilson argued that poverty was the result not just of structural problems in the economy, but of dysfunctional social behaviors that had taken on a life of their own among the 'underclass.' Other studies have focused more on the culture of the rich rather than the poor, observing that different elites can place different relative valuations on work and leisure, leading to conspicuous consumption in some cases and frugality and reinvestment in others.

3.3 Culture and Institutions

Many economists correctly point out that the differences in economic performance from one society to another are better explained by differences in institutions and in the policies undertaken by those institutions than by cultural factors. This is not, however, the end of the story, because culture also affects the ability of societies to create and properly manage institutions. For example, in the postwar period Japan, South Korea, and other East Asian countries employed industrial policies, in which the state rather than the market allocated credit to national industries, to encourage economic growth. Not all societies, however, are capable of building East Asian-style economic planning bureaucracies. Such institutions are especially vulnerable to rent seeking and capture by narrow societal groups; to make industrial policy work, planning bureaucracies have to be shielded from undue political influence and overt corruption. Similar institutions created in Latin America, Africa, and other parts of the world proved much less effective than their East Asian counterparts. While effective industrial policy is partly a matter of institutional design, it is also influenced heavily by culture—for example, expectations of corrupt behavior on the part of public officials, bureaucratic professionalism and esprit de corps, levels of education, etc.

3.4 Social Networks

The final broad category of ways in which culture is said to affect economic behavior is in the formation of social networks. In a short essay titled '*The Protestant Sects and the Spirit of Capitalism*,' Weber pointed out that in addition to the work ethic proper, sectarian Protestantism of the sort found in the US encouraged commerce by promoting networks of trust among the members of each sect. Early Protestantism enjoined its members to behave morally not just towards fellow believers, as was the case with many other religions, but towards all human beings. This kind of moral universalism, combined with the propensity of sectarian denominations to organize themselves congregationally (i.e., from the bottom up rather than hierarchically), meant that business could be transacted across a much broader range of people than in other cultural systems. The differing impact of cultural values on networks of social relations is the basis for the concept of social capital, discussed in the following section.

4. Social Capital

One of the most important headings under which the issue of culture and development has been discussed in the 1980s and 1990s has been that of 'social capital' (Putnam 1993). Social capital consists of norms or values shared among a group of people that promote cooperation and trust among them; like physical and human capital, social capital is a source of wealth. While the term 'social capital' was put into general circulation by sociologist James S. Coleman in the 1980s (Coleman 1988), the concept underlying it has a long intellectual history. Alexis de Tocqueville noted in *Democracy in America* that the American propensity for civil association was key to the success of American democracy, since it permitted the society to organize itself without the help of centralized, hierarchical authority. Edward Banfield in his classic ethnographic study *The Moral Basis of a Backward Society* coined the term 'amoral familism' to connote the pathological condition he found in a small town in southern Italy, where people were only able to trust members of their immediate nuclear family and behaved opportunistically towards everyone else. The case described by Tocqueville was one of abundant social capital outside the family, which created the basis for successful democracy and economic development. Banfield's village, by contrast, was characterized by a dearth of social capital except within a narrow circle of kinship, a situation widely held responsible for the underdevelopment and political corruption characteristic of southern Italy.

More recent studies of social capital have shown that it is very important in understanding the flow of information in an economy. Social networks propa-

gate information on job opportunities, relative prices, and a host of other economic information, but they can also obstruct the flow as well. In the 1990s the World Bank has experimented with so-called 'microlending' projects that seek to use social networks to extend very small retail loans to poor customers in Africa and other regions. The success of microlending depends on adequate information about creditworthiness that is best captured through informal rather than formal information channels.

The concept of social capital has been criticized by both economists and sociologists for either being overly vague, or else being simply a repackaging of already-familiar concepts. One of the key weaknesses of the concept of social capital is lack of an agreed-upon method for measuring it, as exists for physical and human capital. Without a good metric, it is difficult to incorporate social capital into econometric models, and therefore to measure the impact of cultural factors relative to other types of variables.

5. *Directions for Future Research*

As the previous discussion indicated, the biggest challenge in studying culture and development is to find a way to incorporate cultural factors into theoretical and empirical models already in use by economists. Many cultural explanations of economic behavior tend to turn into detailed ethnographic studies, in which causal relationships become so complex that they are not generalizable beyond the particular group being studied. Economists, on the other hand, tend to favor abstract universal models of behavior that fail to take into account many of the complex contextual factors that often prove critical in the real world. It may be the case that these extremes cannot be reconciled, for example, because there is simply no reasonable empirical way to quantify cultural variables, or because causality is too multivariate and complex. On the other hand, the renewed interest in concepts like social capital may lead to the development of new data sources that will permit greater interaction between the ethnographic and model-building sides of the social sciences.

See also: Art and Culture, Economics of; Cultural Policy; Culture as a Determinant of Mental Health; Culture: Contemporary Views; Culture, Production of; Culture, Sociology of; Globalization and World Culture; Poverty, Culture of

Bibliography

Coleman J S 1988 Social capital in the creation of human capital. *American Journal of Sociology Suppl. S.* **94**: S95–S120

Dore R P 1973 *British Factory, Japanese Factory*. Allen and Unwin, London

Granovetter M S 1985 Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* **91**: 481–510

Harrison L E 1992 *Who Prospers? How Cultural Values Shape Economic and Political Success*. Basic Books, New York

North D C 1990 *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, New York

Putnam R D 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ

Smelser N J, Swedberg R 1994 *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ

Weber M 1930 *The Protestant Ethic and the Spirit of Capitalism*. Allen & Unwin, London

Weber M 1951 *The Religion of China*. Free Press, New York

F. Fukuyama

Culture and Emotion

In the history of emotion research in the modern social and behavioral sciences, the pendulum has swung repeatedly between universalist positions and social constructionist positions. This perhaps reflects the fundamental fact that emotions are both biologically grounded and culturally shaped. The debate continues, however, because theorists differ in their views on how large a role biology and culture play in the processes of emotion. For example, basic emotions theorists have suggested that culture modulates emotional expressions through culture-specific display rules, but that emotions themselves are innate and shielded from culture. In contrast, many social constructionists have argued that sociocultural processes participate directly in the formation of emotions themselves. However, with the advent of increasing cross-cultural data, a satisfactory integration of the divergent positions is very much needed. This integration should be consistent with the fact that the human is a species that has accomplished biological adaptation by inventing and elaborating cultural systems. Thus, there must be a cultural mode of biological adaptation, and emotion is no exception to this general principle (see *Culture in Development*).

1. *Cultural Perspective on Emotion*

1.1 *Culture, Biology, and Emotion: Definitions*

A cultural perspective on emotion assumes that humans are biologically prepared with a variety of physiological, neurological, and psychological components of emotion. A reasonable rule of thumb may be that humans are equipped with emotional competence and potential roughly comparable to those of

their immediate evolutionary kins such as primates and great apes (Tomasello 1999). Physiological drives of hunger, thirst, and sex, as well as attendant processes of homeostatic maintenance, will surely make this list. So will rudimentary responses of fight–flight and attachment. Further, basic brain processes of pleasure, pain, and arousal, as well as the corresponding psychological dimensions of valence and intensity of affect, are also part of the significant evolutionary heritage of human beings (see *Emotions, Evolution of*).

However, by themselves, these processes are not emotions. Emotions are higher-order mental organizations formed around certain core elements such as pleasure and pain, approach and avoidance, or fight and flight (Ortony and Turner 1990). Hence, many other processes and elements have to be combined and accorded their divergent functions and forms through social and cultural processes by which individuals try to accomplish, collectively and personally, a form of adaptation and adjustment to their own sociocultural environment (see *Cultural Landscape in Environmental Studies*; *Cultural Psychology*). It is important to note that as a human-made part of the environment, culture is fundamentally symbolic and meaning-ridden. In other words, culture includes a variety of tools, images, knowledge, conventions, and practices, which as a whole define the context of adaptation for each and every individual. The cultural perspective assumes that each individual becomes an agent who thinks, feels, and acts adaptively and functionally within a given sociocultural context both by incorporating cultural meaning systems into his or her individual memory and by coordinating his or her responses with the contingencies of external events that are defined and often taken for granted in the context. This applies to emotion as well. Thus, various innate components of emotion are likely to be configured differently, sometimes dramatically so, across different cultures and historical periods.

1.2 Emotions and Categories of Emotion

One significant contribution of the cultural perspective is to have raised a question on a commonsensical notion that emotions causally follow certain antecedent events (Lazarus 1995). In the emotion research of the past, it was often assumed that a cognitive appraisal of an attendant situation as, say, 'unfair' and 'demeaning,' leads causally to an emotion of 'anger.' Although the notion seems familiar, resembling closely our folk understanding of mental life, it is not at all obvious that the appraisal can be so neatly separated from, let alone causally placed prior to, the emotion itself. One may wonder what will really be left in, for example, anger if an appraisal of 'demeaning offense' is taken out. Perhaps, anger is made up of 'demeaning offense' and the actual social act that is readily interpreted as such. If this is the case, the appraisal and the attendant pattern of social interaction may more reasonably be seen as the integral components of emotional experience.

It is important to note that a large number of studies have documented cross-culturally stable associations between appraisals and emotions (Scherer 1997; see *Emotions, Psychological Structure of*). In these studies, subjects are asked to provide ratings on a variety of appraisal dimensions after having imagined an incident where they felt certain common categories of emotion such as joy, anger, and sadness. Although the degree of cross-cultural similarities observed in these studies is impressive, cautions are due for two reasons. First, there also exist small, but systematic cross-cultural differences in a predicted fashion (e.g., Fischer et al. 1999). Second, exactly what the documented cultural similarities might mean is far from clear. To the extent that appraisals are part of the meaning included in the emotion categories, the studies do show that people have cross-culturally and linguistically translatable folk categories of emotion. But do they equally experience the same set of emotions? Are these emotion categories embodied equally? That is, do they neatly correspond to the structure of online subjective experience, behavioral responses, and bodily reactions? The ability to understand each other's experience in terms of a common set of categories (i.e., translatability) surely points to a common human heritage of some sort. However, as some cultural anthropologists have suggested (e.g., Lutz 1988), the translatability is hardly equivalent to the universality of the emotion process (Parkinson 1995). In fact, with better methods, it has become possible to document considerable cross-cultural differences, as well as similarities, in emotional experience and process (Mesquita and Frijda 1992).

Evidence for the thesis that human emotions are biologically prepared, but socioculturally configured and conditioned is still sparse, limited largely to a few domains. Most notably, these domains include (a) culture and happiness; (b) culture of honor, insult, and violence; and (c) culture and facial expressions of emotion. This evidence is reviewed below.

2. Sociocultural Shaping of Emotion: Some Recent Findings

2.1 Culture and Happiness

Cultures vary considerably in terms of the conditions under which happiness is experienced. In North America and many European cultures, self-esteem is shown to be a major predictor of happiness and subjective well-being. Although the association between self-esteem and happiness can also be documented in many other cultures, it is considerably weaker outside of the European and American cul-

tures (Diener and Diener 1995; see *Well-being (Subjective), Psychology of*). In these cultures, appraisals of the self that is separate from other aspects of life carry an especially large weight in the judgment of the quality of life in general.

Cultures outside of Europe and North America tend to be more interpersonally and socially oriented. Instead of the ethos of independence (which is predominant in the West) the ethos of interdependence permeates various aspects of life in non-Western cultures (Markus and Kitayama 1991, Triandis 1989). So far, much has been learned about East Asian cultures (see *Cross-cultural Psychology*). Overall, in these cultures an assortment of interpersonal and social factors come out as more significant predictors of happiness.

For example, Kitayama et al. (2000) asked both Japanese and American undergraduates to report how frequently they experienced a variety of emotions. These emotions included three different types of positive emotions. Some positive emotions such as friendly feelings and respect are interpersonally engaging in that they subjectively result from a success in tasks of interdependence of the self with others in relationship. Some other positive emotions such as pride and feelings of superiority are perceived to have been caused by a success in tasks of independence and autonomy of the self. Finally, still other positive emotions such as happiness, relaxed feelings, and elation are more general in that other perceived antecedents are possible.

The results suggested that happiness is associated with very different experience in the two cultures. In the US, the reported frequency of experiencing the general positive emotions (e.g., happiness) was more closely associated with the reported frequency of experiencing the disengaging positive emotions (e.g., pride) than with the reported frequency of experiencing the engaging positive emotions (e.g., friendly feelings), thus conceptually replicating the earlier finding regarding the central significance of self-esteem in producing happiness among Americans. In contrast, in Japan the reported frequency for the general positive emotions (e.g., happiness) was much more closely associated with the reported frequency for the engaging positive emotions (e.g., friendly feelings) than with the reported frequency for the disengaging positive emotions (e.g., pride). In short, this and other related studies show that different types of people are likely to be happier in different cultures (Kwan et al. 1997). Thus, whereas in North America individuals with high self-esteem are most likely to be happy, in East Asia those with good social relations are most likely to be happy.

A similar point may apply to the question of when individuals are likely to be happy. In a recent study Mesquita and Karasawa (in press) used an experience sampling method to examine this issue. Both American and Japanese undergraduates were asked to record different aspects of their emotional experience four times a day for a period of several days. It was observed that whereas American subjects were likely to experience pleasant feelings when they thought that they had succeeded in tasks of independence, Japanese subjects were most likely to experience pleasant feelings when they thought that they had succeeded in tasks of interdependence. This social nature of happiness is likely to generalize to other Asian cultures. Thus, in an ethnographic work of Indian women, Menon and Shweder (1998) conclude that the well-being of these individuals is immersed in the notion of social harmony, cyclic continuity of the extended household and lineage, and associated duties and obligations.

The foregoing analysis has an important methodological implication. So far, this literature has been based on an assumption that happiness is a positive emotion that is attributed to the self. This assumption is required to justify the operational definition of happiness as an endorsement of a statement like 'I am happy.' However, cultures may vary in the degree to which emotions are defined and, hence, experienced to be shared in a social relationship or interpersonal space. The social sharedness of emotions may be especially pronounced in non-Western, collectivist, or interdependent cultures.

Finally, evidence indicates another curious cultural difference in the association between positive and negative emotions (such as happiness and unhappiness). Bagozzi et al. (1999) have proposed that Asian, especially Confucian, cultures hold a relatively holistic and dialectic construal about emotions. This construal, which asserts that positivity and negativity should coexist as yin and yang of the same process, may be contrasted with a more prototypically Western construal that assumes an opposition between positivity and negativity. These cultural assumptions seem to be related to the actual experience of emotions. Thus, there is considerable evidence in the West that positive and negative affect are statistically either independent or, otherwise, negatively correlated. Thus, feeling negative emotions does not influence or, otherwise, does decrease the likelihood of simultaneously feeling positive emotions. Recently, however, it has been shown that this is not the case in China (Bagozzi et al. 1999) and Japan (Kitayama et al. 2000). Consistent with the dialectic construal of positivity and negativity, in these Asian cultures the frequency or intensity of feeling negative emotions is often positively correlated with the frequency or intensity of feeling positive emotions.

2.2 *Culture of Honor, Insult, and Violence*

A remarkable degree of culture-dependence has also been demonstrated for another kind of emotion, namely, honor-related anger and aggression in the

American South. Nisbett and Cohen (1996) have suggested that there is a culture of honor in the American South and that this culture is responsible for a higher homicide rate in the South than in the North. A culture of honor is characterized by an ethos of individual power, masculinity, and self-protection (Pitt-Rivers 1965). That is to say, one reacts violently to even subtle cues of insult and attack. A resulting reputation of honor and proneness to violence is likely to preempt any actual insults or attacks on the person, hence serving the function of self-protection. Nisbett and Cohen hypothesize that the honor culture that is held in place in the contemporary American South has its origins in the herding and pastoral cultures of Celts who settled in the southern states of the US in the eighteenth century. They suggest that cultures of honor are likely to arise whenever no effective policing power is available for the protection of private properties. This surely applies to herding economies. Once established, however, a culture of honor often becomes self-perpetuating because many different aspects of the culture reinforce each other to form a sociopsychological system and, furthermore, some of those aspects are likely to be conventionalized (e.g., socialization practices) and institutionalized (e.g., laws that tolerate honor-related violence) (see *Emotions, Sociology of*).

Drawing on this general theoretical framework, Nisbett, Cohen, and colleagues have marshaled evidence from three different layers that constitute the cultural system of honor. This work is exemplary of a cultural approach to emotion. It uses multiple methods; it focuses on different interlocking aspects of a culture; and, as discussed above, it addresses historical origins of the psychocultural complex.

First, the researchers demonstrated attitudinal differences between the two regions of the US. Thus, as compared to Northern men, Southern men are more approving of violent responses when personal insults or other threats on one's properties such as houses and women are involved. Second, the authors have also shown that honor-related violence is sustained in part by social institutions. For example, personnel managers of many firms in the South show a much more tolerant attitude toward a potential employee who allegedly killed someone after having been insulted by him.

Will the honor-based system of culture and attendant attitudes and personal beliefs penetrate a much deeper layer of spontaneous emotional reactions? A third line of evidence, based on a series of ingenious experiments, addresses this issue. These experiments demonstrate that spontaneous emotional responses of Southern men are shaped in part through participating in the culture of honor. Specifically, in one condition of these experiments, when male subjects were arriving for the study, they had to pass another person (an experimental stooge) who was sorting files in an open file drawer. When the subject passed by, he inevitably interrupted the stooge because of the narrowness of the hallway. Further, because of the interruption, the stooge 'accidentally' dropped some files on the floor. Seemingly upset, the stooge audibly called the subject an asshole. There was no such insult manipulation in a control condition. Immediately afterward, a variety of measures were taken including levels of cortisol and testosterone (physiological measures of arousal and proneness to aggression). Half of the subjects were from the US South and the remaining half were from the Northeast or Midwest. The researchers found that the insult produced a visible increase of physiological arousal and behavioral readiness for aggression among Southerners. None of the effects were found among Northerners.

2.3 Culture and Facial Expressions of Emotion

How universal is people's ability to recognize emotions on the face? Earlier work had suggested that recognition of facial emotions is quite culture-specific (Bruner and Tagiuri 1954). In the 1960s, however, this conclusion was challenged by Ekman, Izard, and others, who proposed that several basic emotions (e.g., joy, anger, sadness, disgust, and fear) are innate and part of the built-in hardware of a human animal. The most impressive evidence was that posed faces of Americans corresponding to each of the basic emotions were correctly categorized even by New Guinea tribe men who had minimum contact with the West, and vice versa. The current hypothesis is that cultures vary in terms of intensity of perceiving the basic categories of emotion, but the emotion categories themselves are universal (Matsumoto et al. 1999). Although there is a kernel of truth in the universalist position that emotional expressions are biologically prepared, several reservations have been raised (see *Emotion and Expression*; *Facial Expressions*).

First, there are some methodological weaknesses in the cross-cultural studies by Ekman (1992), Izard (1994), and colleagues. Although minor by themselves, as a whole they may have led the researchers to overstate their case. In particular, Russell (1994) has argued that the available evidence is consistent with a much simpler assumption that people in all cultures categorize facial emotions in two dimensions of pleasantness and arousal. Thus, emotion categories such as joy, anger, and sadness may not be part of a universal hardware of emotion. Second, Wierzbicka (1986)—a linguist from Poland—has suggested that the categories of emotion that are claimed to be basic are relatively specific to the vocabulary of the modern English. Indeed, some cultures do not have any corresponding emotion categories for some of the emotions. The question she raises is apt: What would the list of basic emotions have looked like if it had been compiled by a psychologist who did not speak English? Third, as implied by the cultural perspective described

here, emotions may be much more social in both experience and origins.

Importantly, facial expressions of emotion appear to have different subjective consequences across cultures. Levenson et al. (1992) have shown that once the face is posed to mimic expressions of different emotions, there typically result (in North American subjects) both systematic changes in different aspects of autonomic activation, such as blood pressure, galvanic skin responses, and heart rate, and corresponding changes in subjective experiences of emotion. This suggests that the inner sensations caused by the contraction of facial musculature in, say, a grimace or smile, are often enough to generate an emotional state of anger and happiness. Yet, in a Sumatran village where the researchers conducted the same study, they failed to find comparable changes in the subjective experience of emotional states, although they replicated the physiological changes that were contingent on posed facial expressions. The researchers commented that in Asian cultural contexts interpersonal context serves as an important element of emotional experience, so much so that without the provision of an appropriate interpersonal context, the inner sensations are not sufficient to produce any distinct emotional experience.

3. Conclusions and Future Directions

Advances in culturally informed research on emotion towards the end of the twentieth century have highlighted the notion that emotions are both biological and cultural. Species-specific predisposition, potential, and competence are differentially configured and integrated into culture-dependent forms of communication and social life, which in turn are dependent on the historical, ecological, economic, and political conditions of different cultural groups and regions. Accordingly, all mental organizations, including emotional processes and structures, are likely to be significantly shaped and constituted by sociocultural processes. In turn, these mental organizations become integral parts of the sociocultural processes themselves. Much has been learned about the culture-dependence of emotional responses and processes. Yet, much more remains to be learned. It is especially important to develop better theories and methods that can be used to explore mutually constitutive interactions between sociocultural processes and mental organizations. In all likelihood, this will define one major agenda for research in the cultural psychology of emotion during the first decade of the twenty-first century.

See also: Cross-cultural Psychology; Cultural Psychology; Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Culture as Explanation: Cultural Concerns; Culture: Contemporary Views; Culture in Development; Culture, Sociology of; Emotion and Expression; Emotion: History of the Concept; Emotion, Neural Basis of; Emotions, Evolution of; Emotions, History of; Emotions, Psychological Structure of; Emotions, Sociology of

Bibliography

Bagozzi R P, Wong N, Yi Y 1999 The role of culture and gender in the relationship between positive and negative affect. *Cognition and Emotion* **13**: 641–73

Bruner J, Tagiuri R 1954 The perception of people. In: Lindzey G (ed.) *Handbook of Social Psychology*. Addison-Wesley, Cambridge, MA, Vol. 2

Diener E, Diener M 1995 Cross-cultural correlates of life satisfaction and self-esteem. *Journal of Personality and Social Psychology* **68**: 653–63

Ekman P 1992 An argument for basic emotions. *Cognition and Emotion* **6**: 169–200

Izard C E 1994 Innate and universal facial Expressions: Evidence from development and cross-cultural research. *Psychological Bulletin* **115**: 288–99

Kitayama S, Markus H R, Kurokawa M 2000 Culture, emotion, and well-being: Good feelings in Japan and the United States. *Cognition and Emotion* **14**: 93–124

Kwan V S Y, Bond M H, Singelis T M 1997 Pancultural explanations for life satisfaction: Adding relationship harmony to self-esteem. *Journal of Personality and Social Psychology* **73**: 1038–51

Lazarus R 1995 Vexing research problems in cognitive mediational theories of emotion—and some solutions. *Psychological Inquiry* **6**: 183–96

Levenson R W, Ekman P, Heider K, Friesen W V 1992 Emotion and autonomic nervous system activity in the Minangkabau of West Sumatra. *Journal of Personality and Social Psychology* **62**: 972–88

Lutz C 1988 *Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Challenge to Western Theory*. University of Chicago Press, Chicago

Markus H R, Kitayama S 1991 Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review* **98**: 224–53

Matsumoto D, Kasri F, Kooken K 1999 American–Japanese cultural differences in judgments of expression intensity and subjective experience. *Cognition and Emotion* **13**: 201–16

Menon U, Shweder R A 1998 The return of the 'white man's burden': The moral discourse of anthropology and the domestic life of Hindu women. In: Shweder R A (ed.) *Welcome to Middle Age* (*and Other Cultural Fictions*). University of Chicago Press, Chicago

Mesquita B, Frijda N H 1992 Cultural variations in emotions: A review. *Psychological Bulletin* **112**: 179–204

Mesquita B, Karasawa M (in press) Different emotional lives. *Cognition and Emotion*

Nisbett R E, Cohen D 1996 *Culture of Honor: The Psychology of Violence in the South*. Westview Press, Boulder, CO

Ortony A, Turner T 1990 What is basic about basic emotions? *Psychological Review* **97**: 315–31

Parkinson B 1995 *Ideas and Realities of Emotion*. Routledge, London

Pitt-Rivers J 1965 Honor and social status. In: Peristiany J G (ed.) *Honour and Shame: The Values of Mediterranean Society*. Weidenfeld & Nicholson, London, pp. 18–77
Russell J 1994 Is there universal recognition of emotion from facial expression? A review of cross-cultural studies. *Psychological Bulletin* **115**: 102–41
Scherer K R 1997 Patterns of emotion-antecedent appraisal across cultures. *Cognition and Emotion* **11**: 113–50
Tomasello M 1999 *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA
Triandis H C 1989 The self and social behavior in cultural contexts. *Psychological Review* **96**: 506–20
Wierzbicka A 1986 Human emotions: Universal or culture-specific? *American Anthropologist* **88**: 584–94

S. Kitayama

Culture and the Self (Implications for Psychological Theory): Cultural Concerns

The self may be defined as the individual's understanding and experience of their own psychological functioning. As such, it is basic to human experience, while also being culturally constituted in fundamental ways. The present article examines the processes contributing to the cultural grounding of the self and considers implications for understanding variation in psychological processes and outcomes.

Discussion begins with a conceptual argument for recognizing the self as cultural. This is followed by a brief overview of research that highlights the existence of cultural differences in the self and considers implications of such variation for related areas of psychological functioning. Finally, attention turns to a consideration of challenges to be met in future work on culture and the self and to the potential contributions of this work to psychological inquiry.

1. *Theoretical Premises*

The recognition of the self as cultural is premised on core assumptions regarding the nature of psychological functioning and of culture. These assumptions include a conception of the person as self-reflexive and agentic, as well as a view of culture as forming a symbolic medium for human development.

A core assumption held about the self is that universally individuals maintain some awareness of their mental activity, of themselves as existing in time and space, and of their acting in the world. As Hallowell argues:

> it seems necessary to assume self-awareness as one of the prerequisite psychological conditions for the functioning of any human social order, no matter what linguistic and culture patterns prevail … the phenomena of self-awareness in our species is as integral a part of human sociocultural model of adaptation as it is a distinctive human level of psychological structuralization (Hallowell 1955, p. 75).

Along with this premise of self-awareness, the assumption is also made that individuals are agentic in actively contributing meanings to experience (Bruner 1990). From this perspective, that emerged with the cognitive revolution in psychology, individuals are understood to go beyond the information given in making sense of experience, rather than merely passively to absorb externally given information.

Also of critical importance to the view of self as culturally grounded has been the understanding of culture as a symbolic medium for human development, that encompasses shared meanings which are embodied in artifacts and practices (Geertz 1973, Shore 1996, Shweder and Levine 1984). Cultures are seen as encompassing knowledge of experience and rules for conduct as well as objects and events that are created, in part, by cultural definition. Importantly, this reality-creating role of culture is recognized to be broad, extending not only to the establishment of institutions (e.g., marriage), roles (e.g., bride), and artifacts (e.g., wedding ring), but also to the creation of psychological concepts and epistemological categories (e.g., commitment, time).

From the present perspective, it is understood that, whereas an involuntary response may proceed without cultural mediation, all higher order psychological processes depend, at least in part, on cultural input for their emergence. As Geertz asserts:

> We are … incomplete or unfinished animals who complete or finish ourselves though culture—and not through culture in general but through highly particular forms of it (Geertz 1973, p. 49).

Thus, for example, whereas involuntary physiological reactions may be elicited by situational events, whether they become interpreted and experienced in emotional terms depends, in part, on such input as culturally based theories regarding the nature, causes, and consequences of emotions, cultural routines for responding to emotions, natural language categories for defining emotions, as well as a range of other sociocultural processes. The present considerations lead to the expectation that qualitative differences in self-conceptions and in related modes of psychological functioning may be emphasized among individuals from cultural communities characterized by contrasting self-related sociocultural meanings and practices.

2. *Impact of Cultural Variation in the Self on Basic Psychological Processes*

The present section considers representative examples of empirical studies that embody this core insight regarding the cultural grounding of psychological

processes, an insight that is critical to the newly re-emerging theoretical perspective of cultural psychology (e.g., Cole 1990, 1996, Markus et al. 1996, Miller 1997, Shweder 1990, Shweder et al. 1998). Although the overview presented here is necessarily highly selective and incomplete, it serves to illustrate ways in which cultural research on the self is offering new process explanations of psychological phenomena as well as identifying fundamental variability in the forms that psychological phenomena assume.

2.1 Social Attribution and Cognition

In early groundbreaking research, Shweder and Bourne (1984) challenged the completeness of contemporary psychological theories of social attribution. It was documented that, as compared with European-Americans, Oriyan Indians place significantly greater emphasis in person description on actions as compared with abstract traits, with their person descriptions more frequently making reference to the context. Thus, for example, their research revealed that whereas European-Americans are more likely to describe a friend by saying 'she is friendly,' Oriyan Indians are more likely to describe them by saying 'she brings cakes to my family on festival days.' This type of cultural difference, it was observed, was not explicable in terms of the types of factors that had been emphasized in previous studies, such as variation in schooling, literacy, socioeconomic status, linguistic resources, or capacities for abstract thought. Rather, the trends were demonstrated to reflect the contrasting cultural conceptions of the person and related sociocultural practices emphasized in Hindu Indian as compared with European-American cultural communities. Subsequent cross-cultural developmental research on social attribution documented that cultural considerations impact as well on the direction of developmental change, with European-American children showing an age increase in their references to traits (e.g., 'she is aggressive'), and Hindu Indian children showing an age increase in their references to the social context (e.g., 'there are bad relations between our families') (Miller 1984). More recently, this type of work has been extended to understanding the development of individuals' conceptions of mind, with cultural work calling into question claims that theory of mind understandings develop spontaneously toward an endpoint of trait psychology and providing evidence that they proceed in directions that reflect the contrasting epistemological assumptions of local cultural communities (Lillard 1998).

In other lines of work on psychological implications of cultural variation in the self, research is calling into question the universality of various attributional and cognitive tendencies long assumed to be basic to all psychological functioning, such as motives to maintain self-consistency, or to emphasize dispositional over situational information. Thus, for example, it has been shown that Japanese college students tend to maintain weaker beliefs in attitude–behavior consistency than do Australian college students (Kashima et al. 1992), while being less prone than are Canadian college students to show cognitive dissonance biases, i.e., tendencies to distort cognitions to make them more congruent with behavior (Heine and Lehman 1997). Also, relative to European-Americans, various East Asian populations have been documented to display greater sensitivity to situational information in object perception and less vulnerability to the fundamental attribution error, a tendency to treat behaviors as correspondent with dispositions (Ji et al. 2000).

New lines of research in this area are also linking cultural views of the self and related cultural practices to variation in fundamental styles of cognitive processing, such as tendencies to privilege analytic as contrasted with dialectical epistemological stances. In one illustration of such a cultural difference, experimental research has demonstrated that American undergraduates tend to treat information in a polarized manner, as seen in their considering scientific evidence as more plausible when it is presented alone rather than in conjunction with contradictory information. In contrast, Chinese undergraduates tend to process information in ways that involve greater acceptance of opposing viewpoints, as seen in their considering scientific evidence as more plausible when it is presented in conjunction with contradictory information rather than alone (Peng and Nisbett 1999). These types of differences appear linked, at least in part, to contrasting cultural practices, such as the presence of traditions of debate in certain Western cultural settings as compared with their absence in certain East Asian cultural communities.

2.2 Self-esteem and Well-being

In the area of the self concept, cultural work is calling into question the widely held assumption that enhancement of one's self-image is fundamental to psychological functioning and necessarily yields adaptive advantage. It has been observed, for example, that whereas the open-ended self-descriptions of American adults emphasize positive attributes (Herzog et al. 1998), those of Japanese emphasize either weaknesses or the absence of negative self-characteristics (e.g., 'I'm somewhat selfish,' 'I'm not lazy'). It has also been observed that the scores of Japanese on measures of self-esteem tend to be at or slightly below the scale midpoint, an indication of a tendency to view the self as similar to others (Diener and Diener 1995). This contrasts with the tendency for those of Americans to be higher than the scale midpoints, an indication of a tendency for marked self-enhancement.

Challenging the cross-cultural relevance of the psychological construct of self-esteem, cultural research reveals that self-enhancement tendencies do not play as central a role in social functioning in various

Asian cultural populations as observed typically among European-Americans (Heine et al. 1999). For example, whereas self-esteem is found to correlate more strongly with life satisfaction in cultural settings that emphasize individualistic as compared with collectivist meanings and practices, a concern with relationship harmony shows the reverse pattern of correlations (Gabrenya and Hwang 1996).

Highlighting the role of sociocultural institutions and practices in creating and sustaining such self processes, it has been further documented that European-American behavioral practices tend to be viewed, by both Japanese and European-American populations, as leading to self-esteem enhancement, whereas Japanese behavioral practices tend to be viewed as giving rise to a more self-critical stance (Kitayama et al. 1997). Related types of differences in everyday cultural routines have been documented in research on child socialization. For example, ethnographic work reveals that the stories told by Chinese mothers to their children are more likely to focus on the child's transgressions whereas those told by European-American mothers are more likely to focus on the child's accomplishments (Miller et al. 1996). It has likewise been found that both Chinese as well as Japanese mothers tend to be more critical of their children's academic performance than are American mothers, despite the tendency for their children to show levels of academic achievement superior to those of Americans (Crystal and Stevenson 1991).

2.3 Emotion and Motivation

In regard to emotions, cultural work is providing insight into respects in which affective states, while biologically grounded, are nonetheless constituted, in part, by cultural meanings and practices. For example, somatic experiences that tend to be given a psychological interpretation as emotions by European-Americans are frequently understood and responded to primarily as somatic events within certain Hispanic as well as Asian cultural populations (Shweder et al. 1997). Such effects are closely linked to the tendencies within many Hispanic and Asian cultural communities to interpret social deviance and distress as disorders that originate in problems in interpersonal relationships and that require nonpsychological forms of intervention. In another illustration, research reveals that situationally based physiological arousal is not treated as a sufficient condition for experiencing emotions among men from the Minangabau, a matrilineal agrarian cultural group living on the island of Sumatra in western Indonesia (Levenson et al. 1992). Reflecting their culturally based views of self that treat social relations as integrally related to psychological states, Minangkabau men, unlike American men tend not to interpret their physiological arousal in a socially impoverished laboratory situation in emotional terms.

In turn, studies of culture and motivation are highlighting cultural variation in the experience of agency and challenge the widely held assumption that individuals experience most satisfaction when they are acting autonomously. Recent research has shown that, reflecting their culturally based views of self as inherently social, Hindu Indians report comparable satisfaction when acting to meet role-related interpersonal expectations as compared with when acting more spontaneously (Miller and Bersoff 1994). In striking behavioral findings, it has also been demonstrated that, whereas European-American children display highest levels of intrinsic motivation when they have chosen an activity on their own rather than are undertaking one that was chosen for them by their mothers, Asian-American children experience greater satisfaction when their mothers have chosen the activity (Iyengar and Lepper 1999).

3. Implications

3.1 Challenges

In future research, challenges exist in developing approaches to culture that have greater ethnographic validity and that are more dynamic. In this regard, it is important to go beyond dichotomous frameworks for understanding cultural differences, such as the global dimensions of individualism/collectivism, interdependent versus independent cultural views of self, or of Eastern versus Western approaches. While powerful in certain respects, these types of frameworks embody stereotypical generalities that fail to capture the complexity of individual cultural systems and that portray cultures in ways that are overly static, uniform, and isolated.

Equally, to fully appreciate the impact of cultural views of the self on psychological processes, greater effort must be made to enhance the cultural informativeness of the constructs brought to bear in psychological inquiry. This involves not merely being concerned with the extent to which methodological approaches are familiar and meaningful for different cultural populations. Rather, it requires as well working to insure the cultural inclusiveness of the theoretical constructs embodied in psychological research. In this regard, it must be recognized that many present psychological methods are informed by the cultural premises of the middle-class European-American populations that to date have been subject to most study by psychologists and remain insufficiently sensitive to the perspectives and concerns of other cultural populations.

Finally, there is a need for future research on culture and the self to become increasingly interdisciplinary, with investigators taking into account the conceptual and methodological insights of psychological, anthro-

pological, and sociolinguistic research traditions and avoiding the present insularity, which results either from ignoring or dismissing work from different disciplinary viewpoints. It must be recognized that ethnographic work and quantitative empirical modes of inquiry are interdependent and need to inform each other.

3.2 Contributions

In conclusion, work on culture and the self is succeeding in contributing not only to an understanding of diversity in psychological functioning but is highlighting the frequently implicit cultural dependence of many existing psychological constructs and theories. Work in this area is contributing new process explanations to psychological phenomena, through highlighting the dependence of many existing psychological theories on culturally specific views of the self and associated practices. It is also contributing to the broadening of existing psychological constructs and theories through its identification of alternative normative endpoints for human psychological functioning. Rather than leading to an extreme relativism that precludes comparison, work in this area holds the promise of leading to the formulation of models of human development that are increasingly culturally inclusive and theoretically insightful.

See also: Cultural Psychology; Cultural Relativism, Anthropology of; Culture and Emotion; Culture: Contemporary Views; Person and Self: Philosophical Aspects; Personality and Conceptions of the Self; Psychology of the Real Self: Psychoanalytic Perspectives; Self: History of the Concept; Self: Philosophical Aspects

Bibliography

Bruner J 1990 *Acts of Meaning*. Harvard University Press, Cambridge, MA

Cole M 1990 Cultural psychology: A once and future discipline? In: Berman J J (ed.) *Nebraska Symposium on Motivation: Cross-cultural Perspectives*. University of Nebraska Press, Lincoln, NE, Vol. 38, pp. 279–335

Cole M 1996 *Cultural Psychology: A Once and Future Discipline*. Harvard University Press, Cambridge, MA

Crystal D S, Stevenson H W 1991 Mothers' perceptions of children's problems with mathematics: A cross-national comparison. *Journal of Educational Psychology* **83**: 372–76

Diener E, Diener M 1995 Cross-cultural correlates of life satisfaction and self-esteem. *Journal of Personality & Social Psychology* **68**(4): 653–63

Gabrenya W K Jr, Hwang K K 1996 Chinese social interaction: Harmony and hierarchy on the good earth. In: *The Handbook of Chinese Psychology*. Oxford University Press, Hong Kong, pp. 309–21

Geertz C 1973 *The Interpretation of Cultures*. Basic Press, New York

Hallowell A I 1955 *Culture and Experience*. University of Pennsylvania Press, Philadelphia, PA

Heine S J, Lehman D R 1997 Culture, dissonance, and self-affirmation. *Personality & Social Psychology Bulletin* **23**(4): 389–400

Heine S H, Lehman D R, Markus H R, Kitayama S 1999 Is there a universal need for positive self-regard? *Psychological Review* **106**(4): 766–94

Herzog A R, Franks M M, Markus H R, Holmberg D 1998 Activities and well being in older age: Effects of self-concept and educational attainment. *Psychology & Aging* **13**(2): 179–85

Iyengar S S, Lepper M R 1999 Rethinking the value of choice: A cultural perspective on intrinsic motivation. *Journal of Personality & Social Psychology* **76**(3): 349–66

Ji L J, Peng K, Nisbett R E 2000 Culture, control, and perception of relationships in the environment. *Journal of Personality & Social Psychology* **78**(5): 943–55

Kashima Y, Siegal M, Tanaka K, Kashima E S 1992 Do people believe behaviours are consistent with attitudes? Towards a cultural psychology of attribution processes. *British Journal of Social Psychology* **31**(2): 111–24

Kitayama S, Markus H R, Matsumoto H, Norasakkunkit V 1997 Individual and collective processes in the construction of the self: Self-enhancement in the United States and self-criticism in Japan. *Journal of Personality & Social Psychology* **72**(6): 1245–67

Levenson R W, Ekman P, Heider K, Friesen W V 1992 Emotion and autonomic nervous system activity in the Minangkabau of West Sumatra. *Journal of Personality and Social Psychology* **62**: 972–88

Lillard A 1998 Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin* **123**: 1–32

Markus H R, Kitayama S, Heiman R J 1996 Culture and 'basic' psychological principles. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology: Handbook of Basic Principles*. Guilford Press, New York, pp. 857–913

Miller J G 1984 Culture and the development of everyday social explanation. *Journal of Personality & Social Psychology* **46**(5): 961–78

Miller J G 1997 Theoretical issues in cultural psychology. In: Berry J W, Poortinga Y H, Pandey J (eds.) *Handbook of Cross-cultural Psychology: Theory and Method,* 2nd edn. Allyn & Bacon, Boston, MA, Vol. 1, pp. 85–128

Miller J G, Bersoff D M 1994 Cultural influences on the moral status of reciprocity and the discounting of endogenous motivation. *Personality and Social Psychology Bulletin: Special Issue: The self and the collective* **20**(5): 592–602

Miller P J, Fung H, Mintz J 1996 Self-construction through narrative practices: A Chinese and American comparison of early socialization. *Ethos* **24**: 237–79

Peng K, Nisbett R E 1999 Culture, dialectics, and reasoning about contradiction. *American Psychologist* **54**(9): 741–54

Shore B 1996 *Culture in Mind: Cognition, Culture and the Problem of Meaning*. Oxford University Press, New York

Shweder R A 1990 Cultural psychology—What is it? In: Stigler J W, Shweder R A, Herdt G (eds.) *Cultural Psychology: Essays on Comparative Human Development*. Cambridge University Press, New York, pp. 27–66

Shweder R A, Bourne E J 1984 Does the concept of the person vary cross-culturally? In: Shweder R A, Levine R A (eds.) *Culture Theory: Essays on Mind, Self, and Emotion*. Cambridge University Press, New York, pp. 158–99

Shweder R A, LeVine R A 1984 *Culture Theory: Essays on Mind, Self, and Emotion*. Cambridge University Press, Cambridge, UK
Shweder R A, Much N C, Mahapatra M, Park L 1997 The 'big three' of morality (autonomy, community, divinity) and the 'big three' explanations of suffering. In: *Morality and Health*. Routledge, New York, pp. 119–69
Shweder R A, Goodnow J, Hatano G, LeVine R A, Markus H, Miller P 1998 The cultural psychology of development: One mind, many mentalities. In: Damon W (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. 1, pp. 865–937

J. G. Miller

Culture as a Determinant of Mental Health

1. Introduction

Parallel to the revolution occurring with regard to the development of psychopharmacological therapies and biological theories of mental disorders, the health sciences and professions are undergoing a less visible change in theory and practice that promises to radically alter traditional notions about the nature of mental health (Kleinman 1988). The main underlying tenant of this occurrence is the belief that human behavior and its social determinants are critical variables for understanding the etiology, treatment, and prevention of many mental disorders previously attributed to biological causes. Known as cross-cultural psychology, it is defined as the study of behavior and experience as it occurs in different cultures, is influenced by culture, or results in changes in existing cultures. Cross-cultural psychology is not in itself a separate branch of psychology, but rather a method, a point of view, applicable in principle to all other areas of psychology (Dasen et al. 1988).

As a nation of immigrants, the USA has evolved into a multicultural society. The demographic changes that are occurring due to immigration and high fertility rates among the non-white population will continue to enhance the country's cultural diversity. Most researchers, however, address race and ethnicity as demographic characteristics rather than as distinct predisposing cultural and social environmental orientations. This approach has resulted in research and analyses that ignore the contributory role of sociocultural factors to mental health behaviors. Research is needed to understand how these factors contribute to individual and group coping and adaptive mechanisms, in alleviating the distinct socioeconomic and psychological disadvantages of categorical membership. Culture should be conceptualized as a potential individual and group resource, providing psychological, social, and personal identification and group connectedness for racial minorities in particular, and humans in general (Jackson et al. 1995).

In general, culture consists of those aspects of life that people construct and share within a specific reference group (e.g., race, ethnic, social class, age cohort, gender, nationality, and profession). The term 'culture' is so widely used and multidimensional in its applications (e.g., learned as opposed to instinctive behavior; tools humans create; and values and traditions) that its precise meaning varies from one situation to another (MacLachlan 1997). Culture can be seen as the collective heritage of a people handed down from generation to generation. It includes language, religious beliefs, customs, rules of etiquette, and ideas used by people to organize and interpret their lives and existence. Thus, culture refers to the way a people live, the rules of behavior, and conduct they set for themselves.

The interplay between culture and mental health is complex and multidimensional with respect to antecedents, perceptions, expressions, and treatments for numerous mental illnesses (Marsella and Dash-Scheuer 1988). Successful social adjustment, in effect, is the process of learning the habits, norms, and ways of thinking essential for fitting into one's society. This chapter investigates the cultural determinants of mental health, from the standpoint that culture provides a general design for living and a pattern for interpreting reality that impacts mental health. The emphasis here is on the pragmatic role of culture that is especially pertinent to mental health. The likelihood that different cultures encourage different worldviews that buffer or predispose its members to certain disorders is examined. In particular, the role of culture in psychotherapy is discussed in terms of preventing and treating 'non-dominant' cultural minority members within the USA: African Americans, Latinos, Asian Americans, and Native Americans. In sum, the specific focus is on how ideas, values, and assumptions about life guide both client and practitioner.

2. Coping and Prevention

Each culture transmits itself through the process of socialization. Throughout this process, what is conditioned are both general and specific behavioral patterns for promoting human survival and growth. What are learned are acceptable standards and idealized norms that reflect the hard-won wisdom of previous generations. Through culture, biological adaptive capacities are extended to deal with environmental demands. By understanding the ways that individuals from various cultural traditions cope, health researchers and practitioners can gain unique insights into the promotion of human health and prevention of human illness (Marsella and Dash-Scheuer 1988).

For non-dominant minority group members, the problems of living are not only the products of

biological dysfunctions, aberrant intrapsychic development, and familial dysfunctions, social forces also contribute to the onset and course of mental illness (Kleinman 1988). Political, social, economic, and historical and present day relationships (e.g., slavery, legal discrimination and segregation, immigration exclusion and restrictions, internment, poverty, and forced removal to reservations) with the dominant group in the USA, contributes to mental illness among minorities (Murray 1998). In effect, the larger society has played a major role in the developmental etiology of certain mental health disturbances (Wilson et al. 2000). How do the majority of people of color, in spite of ethnic and racial exploitation, cope, adapt, and succeed?

To answer this question, culturally defined strengths indicative of ethnic minority group life should be illuminated. While racial and ethnic groups of color vary widely in the content of cultural beliefs about time, harmony with nature, reciprocity, use of traditional medical substances, and community healers, certain common themes emerge (Jackson et al. 1995). These include deep and abiding community respect for elders, extended kin network, predominant family commitment, cultural aversion to institutional care, and a clear expectation for respect and dignity in the delivery of health care. Some researchers have specifically commented on the elaborate support networks, extended family relationships, and role flexibility of African Americans; the strong sense of communalism, generosity, and harmony characteristic of Native Americans; the socialization of children to respect the entire family system and put family needs over individual needs in Asian American communities; the socializing toward cooperativeness, and the comfort of their Catholicism in times of stress found in Latino communities (Wilson et al. 2000). These group characteristics facilitate coping and mental health, and lessen the nature and expression of disease and chronic conditions, and are necessary to consider for successful care and treatment.

While there are group commonalties, intragroup differences exist in life-course environmental conditions, degree of acculturation, and cultural traditions and beliefs. Thus, there is an indisputable need for mental health care providers to treat each client as an individual, while gathering information on the degree to which cultural coping strategies may facilitate mental health. It is imperative to conduct cross-cultural research on ethnic minorities within the US to identify the relationship between culture-specific coping skills and mental health outcomes.

3. *Cross-cultural Differences in the Perception of Mental Illness*

Although most, if not all cultures, have a notion of madness, ethnographic studies demonstrate that concepts of emotions, self and body, and illness categories differ so significantly cross-culturally that it can be said that each culture's beliefs about normal and abnormal behavior are distinctive (Kleinman 1988). The *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*, APA 1994) recognizes the importance of cultural factors in diagnosing mental illness. Therapists who work with culturally different clients are instructed to take into account (a) the predominant characteristics of manifesting disorders (e.g., possessing spirits, nerves, fatalism), (b) perceived causes or explanatory models, and (c) preferences for professional and indigenous interventions (Sue and Sue 1999). The *DSM-IV* (APA 1994) describes culture-bound syndromes, which are patterns of abnormality only expressed in a particular group, including ghost sickness (a preoccupation with death and the deceased appearing among Native Americans), nervous (symptoms associated with distress and somatic disturbances, such as brain aches and tingling sensations reported by Latinos), and rootwork (illnesses attributed to hexes, spells, and voodoo by African Americans).

The prevalence and course of mental illness also varies cross-ethnically. Recent research comparing African Americans to Whites indicated no difference between the two groups with regard to panic disorders. However, African Americans exhibit a greater frequency of reported tingling in the extremities and a higher mean number of symptoms (Draguns 2000). A comparison of Mexican Americans and Whites, Mexican Americans presented higher rates of generalized anxiety disorder, agoraphobia, and simple phobia than Whites. While there is a dearth of research comparing Asian Americans to Whites, Lin and Endler (1993, cited in Al-Issa and Oudji 1998) reported that anxiety levels were higher among Chinese immigrants. Their findings were attributed to the stress of adapting to a new environment, which resulted in a sense of insecurity and inadequacy.

Another classical difference ignored by Western practitioners is the finding that bodily states and psychological experiences are monitored, perceived, assessed, and reported differently by members of different cultural groups. A consistent finding is that body complaints predominate over psychological complaints in depressive and anxiety disorders among members of non-Western societies, ethnic minorities in the US, and among less educated members of the lower socioeconomic classes (Kleinman 1988).

Culture may influence how depression is experienced and communicated. Some cultures perceive the physician as a healer of physical ailments rather than psychological ailments. The urgency surrounding physical as opposed to psychological ailments may vary across cultures, as may the degree to which psychological weakness is stigmatizing. The biology of depression and anxiety disorders underwrites the inner form of these disorders, but cultural beliefs and values shape the expression of the illness (Kleinman 1988).

4. Etiological Beliefs and Folk Approaches

All healing systems are shaped by the values, worldviews, and self-views of the cultures in which they develop and are utilized. Therefore, all healing systems are 'ethnomedical.' Rules and principles guiding the various ethnomedical models form discreet and different genre altogether, making cross-ethnic comparisons difficult. However, while distinction among the non-Western groups are difficult to delineate, very salient differences exist between various ethnomedical models and traditional Western methods in psychotherapy. For instance, Western psychotherapy is traditionally a one-on-one activity that encourages clients to discuss the most intimate aspects of their lives. In Latino culture, clients rarely describe their complaints as the main step in an intake process, expecting the healer to already know the client's complaints (Koss-Chioino 2000). Due to experience with oppression, African Americans are hesitant to self-disclose. This may serve a protective function against possible physical and psychological harm (Sue and Sue 1990), but individuals who do not self-disclose may be seen as resistant, defensive, or superficial by therapists. Many minorities who are reluctant to self-disclose are judged to be mentally unhealthy and, in the case of African Americans, are diagnosed as 'paranoid' (Murray 1998).

In spite of the differences in each group's healing practices, cultural commonalties do exist across the four ethnic groups described in this paper. Specifically, the commonalties consist of (a) a widespread belief that the causes of mental illness exist outside the person, (b) a group orientation to healing (i.e., family and/or community) as opposed to an individual orientation to treatment, (c) an emphasis on morality as a cause and condition for recovery (Koss-Chioino 2000), and (d) an emphasis on short range goals, as opposed to long range goals (Sue and Sue 1990). These similar healing practices among non-Western groups differ from Western therapeutic practices.

Latinos, Asians, Native Americans, and African Americans are all heterogeneous groups including many specific traditions that differ from each other in profound ways. Divination (i.e., beliefs in the supernatural and practices to connect with the spirit world) was originally part of each of these groups' healing practices. Unlike Western cultural groups who have discarded such practices, divination to some extent is still part of the healing practices of many, but not all of the subcultural groups within each of the broader categories. Among Native Americans, for example, traditional healing practices vary along tribal lines, and are largely shamanic (spiritual intervention to heal the sick). One of the major shamanic healing entities is the Native American Church, which is a pan-Indian religion based on a ritual that facilitates direct contact with the Great Spirit (also God) through peyote, pipe and cigarette smoking, drumming, and songs and prayer (Koss-Chioino 2000). African American healing practices differ in beliefs, practices, and types of healers based on locale (e.g., north vs. south, and places of special tradition such as the Sea Islands of South Carolina). Overall, the underlying belief system of African Americans is a blend of medical beliefs of an earlier day, African beliefs about voodoo, European folklore regarding the supernatural world, and fundamentalist Christianity. Etiological notions are divided into natural and unnatural categories, varying from unhealthy lifestyles and nonproductive worry to beliefs in evil influence or God taking punitive action. Physicians or herbalists can cure the first two causes, but the latter two causes demand intercession of a religious healer or a contract with God.

Among Latinos, divination may involve healer's use of shells or tarot cards as a guide to see into the spirit world to identify the cause of the client's problem. The healer describes the client's problems (somatic distress, bad feelings, and persistent interpersonal or social problems), and then the client confirms (most frequently) or denies. Very little epidemiological data are available on the utilization of ethnomedical treatments by Latinos in the USA. The existing data, however, indicate that the utilization of divination varies across Latino ethnic populations. Within the USA divination appears to be a flourishing mental health alternative in Miami, Florida, among the Cuban population. Estimates among Mexican Americans vary from 7 percent in San Diego to 54 percent in South Texas (see discussion in Koss-Chioino 2000). It is important to note that some studies separate the belief in traditional healing practices from reports of actual use, so discrepancies may not be so extreme.

Variations are especially salient among Asian Americans (e.g., Chinese, Japanese, Vietnamese, Koreans, and tribal people such as the Hmong or Lao). Among the Chinese, shamanic healing (e.g., divination) has not been reported in the USA. The Chinese-American ethno-healers emphasize a balanced and disciplined life to maintain physical and psychology health. In addition, physical/bodily manipulation, acupuncture, massage, and breathing exercises are common. In contrast, Southeast Asians in the USA see shamans (e.g., folk healers, witch doctors, etc.) when they are available (see discussion in Koss-Chioino 2000). This propensity to visit shamans is motivated by the belief that spirits are omnipresent throughout nature and can cause or heal illness. For example, there occurs among adult Southeast Asians, particularly male Hmong refugees living in the USA, a 'Sudden Death Syndrome.' Autopsies produce no identifiable cause of the death: A person in apparently good health went to sleep and died without awakening. Often, the victim displayed labored breathing, screams, and frantic movements just before death. Prior to their deaths, these men reported that they were experiencing spirit visitations, and that the spirits were unhappy and were punishing them. Clearly, these

deaths do not appear to have a primary biological basis, whereas the deaths do appear to involve psychological factors, including a belief in the imminence of death—either by a curse (e.g., voodoo suggestion) or some form of punishment (Sue and Sue 1999).

Western-trained therapists too often dismiss these belief systems and impose their own explanations and treatment upon culturally different clients. Without an understanding of the underlying cultural beliefs, there exists the tendency for the therapist to overestimate the degree of pathology. In the case of Southeast Asian men who report spirit attacks, a Western traditional therapist would most likely diagnose these persons as paranoid schizophrenic, suffering from delusions and hallucinations (Sue and Sue 1999). Such a diagnosis would most likely lead the therapist to prescribe a powerful antipsychotic drug and perhaps institutionalization. Studies indicate, however, that shaman techniques cure these symptoms (Sue and Sue 1999). The effectiveness of such treatment may lie in the client's belief in the power of the doctor (shaman) and the treatment (divination) received.

While the four groups presented here vary in their degree of homogeneity concerning beliefs in spiritual attacks, a sizeable number of each group does hold these beliefs. Western medicine does not embrace indigenous or alternative healing approaches and often rejects such approaches as unscientific. Using a culturally inappropriate approach may result in harm or even death (e.g., as in the case of Sudden Death Syndrome). In an attempt to insure cultural responsiveness, the mental health practitioner must either incorporate the legacy of ancient wisdom that may be contained in indigenous models of healing, or refer the culturally different client to a culturally appropriate healer.

5. *The Impact of Ignoring Culture in Therapy*

Culture (i.e., systems of symbolic meaning) impacts both mental health and therapeutic intervention. Specifically, culture influences our understanding of the major stages of the clinical process: patient and practitioner explanatory models of sickness; perception, labeling, and presentation of symptoms; health-seeking behavior; doctor–patient communication; curing and healing; and lay and health professional evaluations of therapeutic outcomes. Culture exerts its most fundamental and far-reaching influence through the categories we employ to understand and respond to sickness. As previously discussed, different societies and ethnic groups within one society often affix different sickness labels to the same syndrome (Kleinman 1988). When patient and practitioner explanatory modes reflect different cultural backgrounds, the clinical realities they define can be, and frequently are, quite disruptive to the treatment process. When those explanatory models are in conflict, the results may be client dropout, noncompliance and dissatisfaction, and/or practitioner missed diagnosis, inappropriate treatment, and poor quality of care. All of these reportedly exacerbate existing problems and even create new issues for the client (Sue and Sue 1990, Turner and Kramer 1995).

Although ethnic minorities are increasing in the USA, only superficial interest has been paid to training mental health professionals to deliver culturally sensitive services (Turner and Kramer 1995). Furthermore, the practice of substituting a model stereotype (i.e., the most frequent behaviors of middle class White males) for the real world, while disregarding cultural variations in a dogmatic adherence to some universal notion of truth culminates in a technique-oriented definition of the therapeutic process. The results are that a lack of culturally responsive forms of treatment is the single most important reason given for ethnic minorities being underserved, inappropriately served, and/or their significantly higher drop-out rate (Sue and Sue 1990).

Another significant aspect of cross-cultural awareness is the acknowledgment that disadvantaged people often experience oppression, discrimination, and racism. Instead, a lack of sensitivity on the part of mainstream therapists results in further abuse of the minority ethnic client (Murray 1998). On every index, minorities receive fewer and poorer quality services. Specifically, African Americans and Latinos, in contrast to Whites, holding symptoms constant, were less often given individualized psychotherapy, more often jailed instead of hospitalized, and more often treated with pharmacotherapy instead of talk-therapy (Lawson 1999). Many studies have demonstrated that clinicians given identical test protocols tend to make more negative prognostic statements and judgments of greater maladjustment when the individual was said to come from a lower rather than a middle-class background, or was an ethnic minority member (Lawson 1999).

When ethnic minority clients are met with insensitivity or outright racism on the part of the therapist, poorer outcomes are likely to result (Kurasaki et al. 2000). Thus, the need for changes in mental health delivery is real. For mental health professionals to effectively deliver services to members of a negatively stereotyped group, they must first become aware of their own culture and assumptions about human behavior, values, biases, preconceived notions, personal limitations and so forth. Secondly, they must actively attempt to understand the worldview of their culturally different clients in terms of the etiology, coping strategies, and course of mental disorders (Murray 1998). Cross-cultural sensitivity has been accomplished when the therapist has: (a) an understanding of his or her own culture and achieved self-awareness; (b) an understanding of the client's culture; (c) an understanding of the client's relationship to the sociopolitical system; (d) an understanding of the ways in which generic counseling and therapy may

limit the potential of the client; (e) an array of therapeutic techniques at his or her disposal; and (f) a willingness to exercise institutional intervention on behalf of the client when appropriate (Sue and Sue 1990).

6. Conclusions

In conclusion, culture plays a role in every aspect of mental health. Cultural influences vary greatly from individual to individual. Many mental health problems have culture-common, culture-specific, or both components. In the case of minority ethnic members, too often the larger society's role in the developmental etiology and maintenance of certain mental health disturbances is ignored. However, the literature does suggest that racial attitudes that poison our society and are internalized by mental health professionals can hamper the efforts of people of color to maintain or regain their mental health (Turner and Kramer 1995). In addition, despite the fact that coping is fundamental to our understanding of human health, mental health professionals often do not acknowledge characteristics of minority cultures that are adaptive and facilitate coping within a stressful environment. Every culture is a unique solution to the myriad forces imposed on its members. In fact, cross-cultural differences exist in perceptions of mental illness in terms of beliefs about the causes, the prevalence and course, and the experience of symptoms. Within the therapeutic environment, therapist/client communication, diagnosis, intervention strategies and techniques, client use of health facilities, and the course of mental disorders have all been impacted by culture. Therefore, mental health care providers must be trained to identify the role culture plays in mental illness. The culturally sensitive professional goes beyond the qualities of concern, empathy, and credibility. They have self-knowledge, and a willingness and ability to take the cultural background of clients into consideration. Without self-knowledge and cultural sensitivity on the part of the therapist, the mental health needs of ethnic minorities will remain underserved and inappropriately served at best, and at worst ethnic minorities will continue to be misdiagnosed and victimized under the guise of therapeutic intervention.

See also: Culture as Explanation: Cultural Concerns; Culture: Contemporary Views; Culture in Development; Health and Illness: Mental Representations in Different Cultures; Psychiatric Disorders: Culture-specific

Bibliography

Al-Issa I, Oudji S 1998 Culture and anxiety disorders. In: Kazarian S, Evans D R (eds.) *Cultural Clinical Psychology: Theory, Research, and Practice*. Oxford University Press, New York, pp. 127–51

American Psychiatric Association (APA) 1994 *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*, 4th edn. American Psychiatric Association, Washington, DC

Dasen P R, Berry J W, Sartorius N 1988 Introduction. In: Dasen P R, Berry J W, Sartorius N (eds.) *Health and Cross-cultural Psychology: Toward Applications*. Sage, Newbury Park, CA, pp. 11–21

Draguns J G 2000 Psychopathology and ethnicity. In: Aponte J F, Wohl J (eds.) *Psychological Intervention and Cultural Diversity*. Allyn and Bacon, Boston, MA, pp. 40–58

Jackson J S, Antonucci T C, Gibson R C 1995 Ethnic and cultural factors in research on aging and mental health: A life-course perspective. In: Padgett D K (ed.) *Handbook on Ethnicity, Aging, and Mental Health*. Greenwood Press, Westport, CT, pp. 22–46

Kleinman A 1988 *Rethinking Psychiatry: from Cultural Category to Personal Experiences*. Free Press, New York

Koss-Chioino J D 2000 Traditional and folk approaches among ethnic minorities. In: Aponte J F, Wohl J (eds.) *Psychological Intervention and Cultural Diversity*. Allyn and Bacon, Boston, pp. 149–66

Kurasaki K S, Sue S, Chun C, Gee K 2000 Ethnic minority intervention and treatment research. In: Aponte J F, Wohl J (eds.) *Psychological Intervention and Cultural Diversity*. Allyn and Bacon, Boston, pp. 234–50

Lawson W B 1999 The art and science of ethnopharmacotherapy. In: Herrera J M, Lawson W B, Sramek J J (eds.) *Cross Cultural Psychiatry*. Wiley, Chichester, UK, pp. 67–73

MacLachlan M 1997 *Culture and Health*. Wiley, New York

Marsella A J, Dash-Scheuer A 1988 Coping, culture, and healthy human development: A research and conceptual overview. In: Dasen P R, Berry J W, Sartorius N (eds.) *Health and Cross-Cultural Psychology: Toward Applications*. Sage, Newbury Park, CA, pp. 162–76

Murray C B 1998 Racism and mental health. In: Friedman H S (ed.) *Encyclopedia of Mental Health*. Academic Press, San Diego, CA, pp. 345–57

Sue D W, Sue D 1990 *Counseling the Culturally Different: Theory and Practice*. Wiley, New York

Sue D W, Sue D 1999 *Counseling the Culturally Different: Theory and Practice*. Wiley, New York

Turner C B, Kramer B M 1995 Connections between racism and mental health. In: Willie C V, Rieker P, Kramer B M, Brown B S (eds.) *Mental Health, Racism and Sexism*. University of Pittsburgh Press, Pittsburgh, PA

Wilson M N, Kohn L P, Lee T S 2000 Cultural relativistic approach toward ethnic minorities in family therapy. In: Aponte J F, Wohl J (eds.) *Psychological Intervention and Cultural Diversity*. Allyn and Bacon, Boston, pp. 92–109

C. B. Murray

Culture as Explanation: Cultural Concerns

Culture is a shared interpretive scheme. The subject of this article is the process by which culture is created. Without being able to make comparisons, what can be said about culture? The theory to be described began as a method for comparing cultural bias (Douglas 1970). Twenty years later it had become a theory about political change (Thompson et al. 1990), with two explicit assumptions.

Assumption number one: people living together try to control one another. A culture develops in response to the need to persuade and to mobilize opinion; principles have to be argued, justifications advanced, and opposing opinions attacked. For sheer survival an organization depends on its members' efforts to keep it in being; it would dissolve if there were no will to maintain it. Members of a community put each other under continual pressure to show loyalty and give support. It follows from this that a culture protects and promotes a way of organizing, whether it be a family, a shop, or a kingdom.

Assumption number two: a culture is inherently adversarial (Thompson and Schwartz 1990). A culture grows stronger and more coherent by opposition. In one community there will normally be several cultures self-defined against the others; a progressive culture, a traditionalist culture, a culture of individualism, and no doubt always some members who are just not involved. The researcher needs to trace the alignments in any particular community, know what is at stake, and hear members justifying their own preferences and trying to delegitimate opponents. In any major public policy debate, contending parties tend to align themselves in opposed cultures. They invoke different theories of justice which fuel their conflict and keep the intercultural struggle alive (Davy 1997).

1. Four Types of Organization

The method provides an abstract model of organizational forms, a spreadsheet on which the biases of a group of human subjects can be located. The scheme is necessarily abstract, since it is intended to apply to any kind of social context or conflict. The intention is to show four incompatible social organizations. It represents a range of social constraints by which individuals are confronted. By choosing to align with one type, the person automatically invites opposition from the others.

Why four types? Because 500, or even 20, would be too difficult to manage. Another answer is that the two axes generate four types well known in Western social science writing, so the number fits well to the contemporary discourse. Another answer is that four is enough. That is, it is enough for demonstrating the thesis that culture provides a store of justifications which are used to protect or to attack the organization of social life. Another answer is that the four that are shown in Fig. 1 are clearly incompatible with each other. Anyone who uses the scaling possibilities can multiply the number of subtypes at will.

Why are they adversarial? The four cultural types in question are two groups, (complexly structured or unstructured groups), and two cultures of individuals; the culture of individualism, and the isolates. When the principles of their organization are examined, an inevitable rivalry between the four types appears.

Strong grid *Weak group boundary* Isolates	*Strong grid* *Strong group boundary* Hierarchy
Individualist culture *Weak grid* *Weak group boundary*	Egalitarian group *Weak grid* *Strong group boundary*

Figure 1
The four types of control

2. Hierarchy and Individualism

Different principles of organization make contrary demands on the use of time and space. For example, a hierarchy is organized in complementary-ranked compartments which subordinate the individual member to the whole. Its priorities will be reflected in the grandiose design of public spaces, such as a magnificent cathedral precinct, museum, and municipal park, sheer size contrasting with small individual dwellings stacked in tidy little groups; the pattern of the streets is confusing, new patterns built over the old ones without obliterating them, narrow winding lanes still known by their original names. In the individualist culture, old streets can be torn up to make room for wide, straight motorways, numbered for the sake of efficiency, an overall pattern of grand private mansions and tawdry public places, disorderly hovels. As to the use of time, a hierarchy is not in a hurry, it expects its institutions to last for ever, but the individualist culture puts pressure on the use of time as part of its commitment to efficiency. Such incompatible demands uphold the antagonism between these cultures.

To introduce Fig. 1, let us continue to contrast the idea of a hierarchy, quite fairly caricatured in accounts of bureaucracy, with the idea of an individualist culture. Hierarchy assigns status by birth, which gives age, gender, and place as further principles of identification. Hierarchy controls individual behavior and severely limits the scope of negotiations. In the individualist culture, merit, not heredity, determines status, the individualist has to negotiate everything. These are two kinds of organization, one organized by status, in which positions are assigned, the other organized by competition. To map these differences on a graph, start with the horizontal axis (called 'group' or 'group solidarity'). This indicates incorporation

into groups. Along that line the individual members of society have to be located according to how much their behavior is controlled by a group boundary. No group belongs at zero on the left corner, an environment favoring individualism, the strong group is towards the right. Some strong groups, such as sects or monasteries, strictly control their boundaries. Elsewhere, individuals can be members of several groups; a choir, a political club, or a sports club, the local neighborhood watch; none is exclusive, group control is weak, boundaries are weak. It is not difficult to find indices for these features.

3. *Four Types of Control*

These are shown in Fig. 1. The two axes combined provide a range of possible social environments. Strong measures for both grid and group correspond to the organization of hierarchy; weak grid and weak group correspond to the environment of an individualist culture.

There are other impersonal constraints that cut across group boundaries, which structure behavior and in this context are referred to as 'grid' or 'structure.' 'Grid' is an impersonal set of rules that prevents free communication and limits transactions, it imposes ceilings on endeavor and barriers to entry. A class system is a grid factor, immigrant status entails restrictions, so does the difference between wealth and its absence. Limitations on the rights of ethnic groups, prohibition on land owning, sumptuary laws, all these draw lines between how members of the same society can transact with each other. Restrictions on nonconformist religions in England used to prevent their access to universities. Such constraints are shown on the vertical axis. The restrictions cited are selective, but others fall on everybody in the community to the same degree; curfews in time of war, restrictions according to age for drinking alcohol, restrictions on road traffic, control of prices or quality of merchandise, health and safety regulations. At the bottom of that vertical line is a possibility of perfect freedom to wheel and deal, at the top, no options.

4. *Egalitarian Groups and Isolates*

Two corners of the diagram are still not described. One is the relatively rare social environment made by a group which encloses its members within strong boundaries but imposes no structure on them. The one constraint that matters is the external boundary; within it every transaction has to be negotiated. The founding principles of some sects lay down that all the members are free and equal, but in practice it takes a lot of regulation to create perfect equality (Rayner 1988). Where it differs from hierarchy is that this regulation controls all the members in the same way, like rules of the road; it does not create internal divisions.

Environmental or other activist groups as well as religious sects often observe egalitarian ideals. So long as the group is unstructured (a kibbutz, for example), they belong in what is often called (in contradistinction to hierarchy) the sectarian corner of the diagram. Of course, if a group were to maintain a hierarchical system, such as the Mormons or certain groups of environmentalists after having acquired political power, they should be placed along with other hierarchies on the diagram.

There are organizational reasons for why many protesting dissident groups (secular or religious) espouse a doctrine of equality. By withdrawing from the big, complex, mainstream society, the group has problems in maintaining itself. The leaders find they lack authority, they are afraid of being accused of free-riding, decision making becomes difficult. The doctrine of equality is often adopted as a solution, henceforth no-one will be superior to any other member, decisions will depend on consensus. But the solution entails other difficulties. It does not help decision making, or protect authority (Sivan 1995). The group expects to ensure its own survival, but it has no authority to stop its children from leaving. The danger of defection focuses attention on the boundary that defines the insiders. Inside/outside becomes the central cultural issue. Then the other cultural traits associated with sectarianism may come into play, a dichotomizing tendency in judgments, the world seen in black and white, outsiders as evil or ungodly, intransigence in negotiation.

The other is the environment of the isolates, dropout, noncommitted, unaligned individuals. They are by definition not involved when the members of the other three cultures argue strenuously about the right and the good. If they were, they would either have to join up with like-minded souls and be classed as members of a group, or they would have to speak as independents, joining the individualists on the diagram. The diagram shows up well the extreme types of organization, selected on the basis of their incompatibility with the rest. The human subjects have to be fitted into their slots according to agreed measures. Rayner and Gross (1985) have shown how to develop measures with a toy example of a conflict involving a group of environmentalists, town residents, an industrial corporation, and government officials. Theoretically, the four conflicting cultures are potentially available at any time, usually one dominant, each in competition with the others.

5. *Information*

Group boundaries make a hurdle that new information has to cross. The first test of reliability is the source; insiders are more to be trusted than outsiders,

but certain outsiders are privileged. If the group is divided into factions or into administrative compartments, each internal boundary constitutes another barrier.

Blocks to the flow of information abound in a hierarchy; each compartment has its own secrets. It would subvert the structure of the organization to make pressure to disclose records. Knowledge is dispensed according to 'the need to know.' Secrecy tends to be acceptable and normal, information is classified and stored systematically. As the organization has institutions for sifting and storing it, information from outside is not automatically unacceptable, but it is put under political scrutiny. Consequently, it travels slowly and often gets stuck.

An egalitarian commune is in worse case, since it lacks any formal structure for testing and classifying the news. Mistrust is the drawback when two strong group cultures try to negotiate with each other; they mishear what the other side is saying. Within an egalitarian group there may be so much mistrust that official envoys to a peace conference may not be fully informed.

Like the culture, information has to compete to be accepted. Individualism is not strong on protecting the commons against spoliation, it justifies its ways on grounds of individual human rights. Information can flow freely in the absence of boundaries, but public institutions for classifying or organizing it are weak. Instead of information there is uninterpreted noise; individuals may have to pay for special knowledge for serious decisions.

The isolates are in a similar pass. Their isolation means that they are not likely to be influential in councils, their news comes to them haphazardly, and they lack incentives to sort it out. They tend to have disconnected, idiosyncratic opinions. It does not matter because in that environment there is no-one else to argue with: they are isolates.

6. *Storage and Retrieval*

A hierarchy justifies itself by appeal to tradition, and by appeal to the well-being of the whole community. To strengthen its justifications it tends to take its origins back to the beginning of time. Since each subdivision uses the same kind of appeals and justifications as the center, hierarchy develops a characteristic sense of history, extending far into the past, and subdivided at every point that a separate subdivision came into existence. The calendar of feasts revives public memory as it celebrates each point of junction or division. The organization that is strong enough to protect its own security gives guarantees of stability to its members. Assuming the system will endure, they act to assure that it will. Partly because it has no advantage in looking for a sudden end to time, and partly because it is confident enough to allow loans to be repaid over a very long-term future (for example, leased property for 900 years), its sense of future time is also very long.

The individualist culture differs in all these respects. It does not justify itself by reference to history, nor does it regard any of its present ranks or subdivisions as unalterable, so that divisions in past time are not built into the organization. Its festivities are more apt to celebrate individual prowess than past events. Its sense of future time is short. An egalitarian group, by contrast, has no internal divisions whose annual celebrations endow their memories with structure. Their sense of time is millennial; the past will have started apocalyptically with the foundation of the group, the moment which is the point of reference for claims and justifications. Their attacks on outsiders and their justifications of their own record refer constantly to a coming catastrophe for the enemies of the group. The difficulties of organization invite covert competition for leadership; as there is a deficit of power, there are no grades for downgrading, no regular system of punishment, conflicts can only be solved by expulsion or by splitting the group. These structural weaknesses mean that though the movement may have its roots in a long-distant past, each community is relatively new: its history is short.

The fourth cultural type, the isolates, gives minimum opportunity for negotiation, either because of demographic sparsity (nobody to negotiate with), or because other people have restricted their freedom to negotiate, typically in the lower echelons of a big bureaucracy. The culture of isolates is politically apathetic, and limited to the short term. While the isolates themselves are not necessarily unhappy, their sense of a communal past is practically non-existent, and such individuals have little in the way of public benchmarks to help in remembering their private pasts.

In this perspective, only the hierarchy has the possibility of recalling a differentiated past with many twists and turnings, but it is certain that its history will be strongly biased in favor of itself.

7. *Decision Making*

Hierarchy's characteristic decision procedure is based on historical precedent and seniority. What is once decided can be made effective because the system protects authority. Individualist culture is also effective in decision making, because it reflects the balance of power at the time. By contrast, decision making in an egalitarian community is beset with difficulties; authority is weak and the group risks destruction from its own internal dissension. The isolates, knowing that they have no influence, realistically adopt a culture of fatalism, which militates against consistent decisions.

8. Conclusion

We could introduce the theory from almost any situation of conflict, but we have chosen to focus on selection of information, because, after all, it is about argument and disagreement on values. It can explain why information does not always travel, and how it sometimes gets blocked. Some cultures preserve information about a long historical past, some have a store of techniques for projecting the future, others live in the present and are not interested in the past or the future. The variation is due to the way that organizations act as filters and pointers (Douglas 1987). These examples illustrate how this type of analysis can trace different sources of failure to mobilize support for a given set of principles. It has been widely applied to disagreements about risk. In the late 1990s the scope has widened to include environmental politics of the great waterways, the Ganges and the Rhine (Gyawali 1998, Verweij 2000), traffic control in city planning (Hendriks 1999), conflict in recent British Labour party history (Bale 1999), kinds of control used in public administration (Hood 1998), and even to conflicting ideas about personal privacy (Perri 1998), and market research (Karmasin and Karmasin 1997).

Culture sets up the terms on which individual calculations of interest are based. Organization entails coordination, which entails classification and logic. As they responsibly calculate and negotiate, the rational beings are creating, changing, or sustaining their culture. The method is sometimes criticized for focusing on organizations and paying too little attention to the role of individuals in shaping culture, but far from neglecting the individual, the theory sets the individual person in a normal framework of interaction. At center stage, each rational person, assumed to be interested in the costs or benefits of subscribing to other people's opinions, is partaking in a process of collective choice.

See also: Cultural Critique: Anthropological; Cultural Evolution: Overview; Culture: Contemporary Views; Culture in Development; Culture, Production of; Culture, Sociology of; Gender Ideology: Cross-cultural Aspects; Sexual Orientation: Biological Influences

Bibliography

Bale T 1999 *Sacred Cows and Commonsense: The Symbolic Statecraft and Political Culture of the British Labour Party*. Gower, Ashgate, UK

Davy B 1997 *Essential Injustice, When Legal Institutions Cannot Resolve Environmental and Land Use Disputes*. Springer, New York

Douglas M 1970 *Natural Symbols*. Penguin

Douglas M 1987 *How Institutions Think*. Syracuse University Press, Syracuse, NY

Douglas M, Wildavsky A 1983 *Risk and Culture, an Essay on the Selection of Technological and Environmental Dangers*. University of California Press, CA

Gyawali D 1998 Patna, Delhi and environmental activism: Institutional forces behind water conflict in Bihar. *Water Nepal* **6**(1): 67–115

Hendriks F 1999 *Public Policy and Political Institutions, the Role of Culture in Traffic Policy*. Edward Elgar, Cheltenham, UK

Karmasin H, Karmasin M 1997 *Cultural Theory, Ein Neuer Ansatz für Kommunikation*. Marketing und Management, Austria

Perri 1998 *The Future of Privacy*. Demos, London

Rayner S 1998 The rules that make us equal. In: Flanagan J G, Rayner S (eds.) *Rules, Decisions and Inequalities in Egalitarian Societies*. Avebury, pp. 20–42

Rayner S, Gross J 1985 *Measuring Culture*. Columbia University Press, New York

Rayner S, Malone E 1998 *Human Choice and Global Climate Change*. Battelle Institute, Columbus, OH

Sivan E 1995 The enclave culture. In: Marty M (ed.) *Fundamentalism Comprehended*. Chicago University Press, Chicago, pp. 11–68

Thompson M, Schwartz M 1990 *Divided We Stand, Redefining Politics, Technology and Social Choice*. Harvester Wheatsheaf

Thompson M, Ellis R, Wildavsky A 1990 *Cultural Theory*. Westview, Boulder, CO

Verweij M 2000 *Transboundary Environmental Problems and Cultural Theory, the Environmental Protection of the Rhine and the Great Lakes*. Berghahn, London

M. Douglas

Culture: Contemporary Views

1. Introduction: Kroeber's and Kluckholn's Prediction

' … few intellectuals will challenge the statement that the idea of culture, in the technical anthropological sense, is one of the key notions in contemporary American thought.' That prediction was made in 1952 by the American anthropologists A. L. Kroeber and Clyde Kluckholn in the introduction to their monumental book *Culture: A Critical Review of Concepts and Definitions* (1952). In the years immediately following World War 2 many scholars and intellectuals in the United States were quite confident and pleased that the concept of 'culture' was deeply entrenched in the human sciences. Kroeber and Kluckholn even began their famous treatise proclaiming the idea of culture comparable in explanatory importance to the idea of gravity in physics, disease in medicine, and evolution in biology. They ended by adducing a unified (albeit ponderous) definition that became the mantra for cultural anthropologists who

came of scholarly age in mid-century. 'Culture' they wrote, 'consists of patterns, explicit and implicit, of and for behavior acquired and transmitted by symbols, constituting the distinctive achievement of human groups, including their embodiments in artifacts; the essential core of culture consists of traditional (i.e., historically derived and selected) ideas and especially their attached values; culture systems may, on the one hand, be considered as products of action, on the other hand as conditioning elements of further action' (1952 p. 357).

Kroeber and Kluckholn were students of intellectual history and brilliant culture theorists but they were not prophets. Little did they know that during the 50 years following the publication of their book, the idea of 'culture,' in its mid-century anthropological sense, would be frequently debated, doubted, distrusted, and scorned, that the discipline of cultural anthropology itself would be 'rethought,' 'remade,' 'recaptured,' and 'reinvented' time and time again. They did not foretell the many types of humanists and social scientists (cognitive revolutionaries, structuralists, poststructuralists, sociobiologists, feminists, skeptical postmodernists, postcolonialists, subalterns, globalization theorists) who would associate the concept of 'culture' with a variety of supposed sins. Sins such as 'essentialism,' 'primordialism,' 'representationalism,' 'monumentalism,' 'reification,' 'idealism,' 'positivism,' 'functionalism,' 'relativism,' 'sexism,' 'racism,' 'colonialism,' 'Orientalism,' and just plain old-fashioned 'stereotyping' (see for example, Abu-Lughod 1991, Asad 1973, Clifford and Markus 1986, Denzin 1996, Fox 1991, Freeman 1983, Hymes 1972, Kuper 1999, Markus and Fisher 1986, Rabinow 1983, Reyna 1994, Rosaldo 1989, Said 1978, Sangren 1988, Scheper-Hughes 1995, Spiro 1986, Wikan 1996).

Nor did Kroeber and Kluckholn anticipate the ironic fate of the concept of 'culture' at the beginning of the twenty-first century. The irony is that today the idea of culture is once again a key concept in many of the social science disciplines, yet is viewed with great suspicion in some quarters of cultural anthropology. The irony is that after being reviled, pummeled, and rejected by one new wave intellectual movement after another an idea of 'culture' very much like the one recommended by Kroeber and Kluckholn in 1952 remains useful and defensible in social science research and public policy debates. The concept not only survives it thrives (see for example, Harrison and Huntington 2000, Huntington 1996, Landes 1998, Prentiss and Miller 1999, Kitayama and Markus 1994, Wierzbicka 1997). The contemporary discipline of anthropology continues to be a scene for various kinds of 'anticultural' or 'postcultural' critiques. Nevertheless, many anthropologists rehearse and recite some definition of culture, and make good use of it in their scholarship (see for example, D'Andrade 1995a, Dumont 1970, Geertz 1973, LeVine et al. 1994, Sahlins 1995, 1999, Shore 1996, Shweder 1991, Shweder and LeVine 1984). It remains to be seen whether and just how soon the concept regains its former popularity in anthropology. That discipline was historically identified with 'studies of culture.' Tripling the irony, in recent years some anthropologists have abandoned the idea of 'culture' and subjected it to intense critique under the banner of an intellectual movement known as 'cultural studies' (Denzin 1996).

2. *The 'Standard View' of Culture in North American Anthropology*

Kroeber's and Kluckholn's seminal definition of culture—which I shall refer to as the standard view from the perspective of North American cultural anthropology—was cumbersome, in part because it was so inclusive. It called on anthropologists to study not just other people's beliefs (their ideas of what the world is like) but also other people's normative standards (their ideas of what is good and what is right). It called on anthropologists to study not just the explicit 'ethnosciences' and doctrinal moral and religious codes of the members of a community but their tacit, implicit, or intuitive understandings as well. It sought a middle course between the *Scylla* of a purely behavioral definition of culture and the *Charybdis* of a purely ideational one. According to the standard view, 'culture' should be defined in such a way as to avoid the hazards of both behaviorism and idealism. On the one hand, Kroeber and Kluckholn suggest that culture is more than just social habits or 'patterns of behavior that are learned and passed on from generation to generation.' On the other hand, it is not just a system of categories, doctrines, propositions, or symbols *per se*. Thus in the 1952 definition, culture is defined as the ideational side of social action or social practice, and anthropologists are called upon to view cultural analysis as the interpretative study of behavior, although rather little is said about what particular theory of interpretation should guide the analysis.

Useful definitions deserve to be expressed in elegant terms, and Kroeber's and Kluckholn's definition of culture is cumbersome, to say the least. But it is not the only expression of the standard view. The most exquisite and straightforward formulation is Robert Redfield's definition: 'shared understandings made manifest in act and artifact.' Another variation, perhaps the most famous definition of culture since the 1950s, is the one proposed by Clifford Geertz (1973, p. 89). He puts it this way: '… the culture concept … denotes an historically transmitted pattern of meanings embodied in symbols, a system of inherited conceptions expressed in symbolic form by means of which men communicate, perpetuate, and develop their knowledge about and attitudes towards life.' In 1984 D'Andrade, writing very much within the North American anthropological tradition, defined 'culture' as 'learned systems of meanings, communicated by

means of natural language and other symbol systems, having representational, directive and affective functions, and capable of creating cultural entities and particular senses of reality' (D'Andrade 1984, p. 116).

The definitions proposed by Kroeber and Kluckholn, Redfield, Geertz, and D'Andrade call out for specification and clarification. Nevertheless, those definitions are a good reference point for understanding current debates about the values and dangers associated with the very idea of 'culture.' One can summarize the standard view by saying that 'culture' refers to community-specific ideas about what is true, good, beautiful, and efficient. To be 'cultural' those ideas about truth, goodness, beauty, and efficiency must be socially inherited and customary. To be 'cultural' those socially inherited and customary ideas must be embodied and/or enacted meanings; they must actually be constitutive of (and thereby revealed in) a way of life. Alternatively stated, the standard North American anthropological view of 'culture' refers to what the British philosopher Isaiah Berlin (1976) called 'goals, values and pictures of the world' that are made manifest in the speech, laws, and routine practices of some self-monitoring and self-perpetuating group. A cultural account spells out those 'goals, values, and pictures of the world.' A cultural account thus assists us in explaining why, the members of a particular cultural community say the things they say and do the things they do to each other with their words and other actions. These 'goals, values, and pictures of the world' or 'ideas about what is true, good, beautiful, and efficient' are sometimes referred to as 'cultural models' (D'Andrade 1995, Holland and Quinn 1987, Shore 1996).

3. *Fault Lines in Contemporary Anthropology*

The standard North American anthropological view of culture was synthesized and defined by Kroeber and Kluckholn before the discipline went through a series of revolutions and movements that fractured the field and divided it on the basis of somewhat different visions of its mission. Although it may be hazardous to propose a map of the current intellectual camps within cultural anthropology, such a map may also be helpful in understanding the various types of anticultural, postcultural, and procultural positions that have emerged within anthropology over the past 50 years.

I think it is fair to say that contemporary cultural anthropology is divided into at least these four conceptions of the field.

3.1 *Identity Politics*

The first is a conception of anthropology as a platform for moral activism in the battles against racism, sexism, homophobia, and neocolonialism and as a forum for identity politics in the fight against exploitation, discrimination, and oppression. Advocates of this conception of anthropology have several concerns about the idea of 'culture.' They argue that the idea of 'culture' is an excuse for the maintenance of authoritarian power structures and permits despots and patriarchs around the world to deflect criticism of their practices by saying 'that is our custom' or 'that is the way we do things in our culture' (see Abu-Lughod 1991, Said 1978, Scheper-Hughes 1995, Wikan 1996). The claim by liberal 'first world' feminists that 'multiculturalism is bad for women' (Okin 1999) is an expression of this view, which tends to associate 'culture' with the idea of patriarchal domination (Haynes and Prakash 1991, Reheja and Gold 1994). This conception of the mission of anthropology is closely allied with a global human rights movement that has a firm sense of what is objectively and universally right and wrong.

However, not all moral activists in anthropology want to dump the idea of 'culture' and some have found ways to put the idea of 'culture' to work in the service of their own political aims. Anthropologists who are active in the identity politics movement find the idea of 'culture' politically and strategically convenient in their egalitarian battles on behalf of 'oppressed peoples.' They find the idea useful as one of several ways to mitigate various types of invidious comparisons between groups (for example, comparisons in terms of wealth, occupational success, or school performance). There are generally three ways to mitigate invidious comparisons between groups: (a) deny that any real differences exist, (b) attribute all differences to a history of oppression or discrimination, or (c) celebrate the differences as 'cultural.' In the identity politics movement 'culture' is becoming a code word for 'race' and it may not be too long before the expression 'people of color' is replaced by a new shibboleth, 'people of culture.'

3.2 *Skeptical Postmodernism*

The second conception of the mission of anthropology is a conception of the field as a deconstructive discipline and as an arena for skeptical postmodern critiques of all ethnographic representations and so-called 'objective' knowledge (see, for example, Clifford and Markus 1986, Latour and Woolgar 1979, Rabinow 1983, also Culler 1982, Foucault 1973, Rosenau 1992). Advocates of this conception of anthropology call for a deeply skeptical reading of all anthropological representations of 'others,' especially of those accounts that make claims about some 'primordial' or essential or core cultural identity which members of some group are supposed to share. The skeptical postmodernists raise doubts about the reality and existence of groups. They are critical of all attempts to draw a portrait of 'others' that represents

them with any characteristic face. They are suspicious of the very idea of boundaries and borders and loyalties to ones 'tribe.' They view the idea of a 'culture' as a fiction, the goal of objective representation as misguided, and the products of ethnography as largely 'made up' or constructed in the service of domination.

One of the many ironies of contemporary anthropology is that for a while members of the first two camps of anthropology (the identity politics/moral activists and the skeptical postmodern/deconstructivists) thought they were allies. Indeed they had an imagined common enemy. In particular the hegemonic heterosexual 'first world' white males, such as Kroeber, Kluckholn, Redfield, and Geertz, who historically had defined the mission of cultural anthropology.

The alliance, however, was short-lived. Identity politics requires a robust notion of 'identity' and group membership. Moral activism requires a good deal of conviction that some things are 'objectively wrong.' Skeptical postmodernism is intellectually incapable of lending support to either of those metaphysical notions and is readily put to use deconstructing the 'woman' of 'Women's Studies,' the imagined common identity of the ethnic group, and all supposed objective moral foundations for any political cause. If 'groups' and collective identities are so easily dissolved so too are claims about group rights and affirmative action.

3.3 Neopositivism

A third conception of the mission of anthropology is a conception of the field as a pure 'positive' science (see for example, D'Andrade 1995a and 1995b, Sperber 1985, Romney et al. 1986). The positive scientists view anthropology as a value-neutral and nonmoralizing discipline. Their preferred aims for the discipline are to reliably and validly represent the law-like patterns in the world and to develop universal explanatory theories and test specific hypotheses about objectively observable regularities in social and mental life. Advocates of this conception want to protect anthropology from identity politics and skeptical postmodern critique by accurately recording rather than judging and condemning other peoples practices, and by developing objective or scientific standards for evaluating the truth of ethnographic evidence. This is a laudable aim, although one that has been contested by skeptical postmodernists, and there has been much useful work in neopositivist fields such as 'cognitive anthropology' representing the content, structure, and degree of sharing of 'cultural models' (see for example, D'Andrade 1995a, Holland and Quinn 1987, Romney et al. 1986). Nevertheless, the positive scientists in anthropology thereby tend to beg a critical question close to the heart of all great social theorists. Is this or that social order really a moral order? Is this or that social order a way of living that might appeal to a rational and morally decent person, and if not how can we make it become so? When it comes to evaluating what is truly desirable or really 'good' in social life the neopositivists are very much like the skeptical postmodernists—both turn radically subjective or relativistic and believe there is no scientific or objective foundation for value judgments.

3.4 Romantic Pluralism

A fourth conception of the mission of anthropology is a conception of the field as a romantic discipline designed to test the limits of pluralism. Pluralism is the idea that things can be different but equal, and that diversity can be good. It is a measure of some of the tensions within contemporary anthropology that while the 'ethnography of difference' is viewed with suspicion by some of the anticulturalists, it is universally embraced by romantic pluralists. Anticulturalists worry that any description of cultural difference merely sows the seeds of invidious comparison and ethnic conflict, and thus should be disavowed. For the romantic pluralists, however, the recognition and appreciation of cultural differences is one of the major aims of ethnography in particular and cultural anthropology in general.

The intellectual inheritance of the 'romantic' tradition most relevant to this camp of anthropology is a conception of culture as an extension of the creative imagination, which is said to be a distinctive intellectual capacity of human beings (see Geertz 1973, Sapir 1963, Sahlins 1995, Shweder and LeVine 1984, Shweder 1991). According to romantic pluralists, a 'genuine' culture is a reality-binding product of the human mind that is not dictated by either logic or direct (meaning-free) experience. There is thus plenty of room within the limits of logic and experience for cultural variety, and for the historical creation of different lived conceptions of what it means to be a rational and morally decent human being. According to this view social and cultural realities are neither logically deduced nor simply found in direct experience but are rather constructed by, and for, more or less rational agents. The human creative imagination has the capacity to fill in, and give definition to, a vast discretionary space that stretches in between the necessary truths of formal logic and the uninterpreted evidence of the senses. Advocates of this conception of anthropology are dedicated not only to the project of accurate ethnographic representation, but also to the cognitive and moral defense of different ways of life, frames of reference, and points of view. They write books or articles about Azande Witchcraft (Evans-Prichard 1937) or Balinese conceptions of the 'person' (Geertz 1973) or Oriya Hindu family life and gender relations (Menon and Shweder 1998) which portray

the ideas and practices of others as different but equal to our own, in the sense that such ideas and practices are represented as meaningful and imaginative yet supportable within the broad limits of scientific, practical, and moral reason.

4. *A Fifth Camp Within Anthropology? The Return of Cultural Developmentalism and the 'First World's' Burden*

Increasingly these days, as the world 'globalizes,' the concept of 'culture' gets used to explain differences in the economic, social, political, educational, and moral accomplishments of nations, groups, or peoples. An 'evolutionary' or 'developmental' view of culture has returned to the intellectual scene. Along with it comes the claim that some groups have the wrong models, the wrong values, the wrong patterns of behavior, and that is the reason that their economies are poor, their governments corrupt, and their people unhealthy, unhappy, and oppressed. The cultural developmental view of cultural differences was quite popular at the very beginning of the twentieth century, and is associated with the 'civilizing project' or the 'white man's burden' to uplift those who are ignorant, superstitious, primitive, savage, and poor. Quite remarkably the cultural developmental view is increasingly popular at the beginning of the twenty-first century as well, especially outside of anthropology, for example, in economics and political science (Harrison and Huntington 2000, Landes 1998). In development economics (for example, at the World Bank), the view that 'culture counts' or that 'culture matters' is now popular in part because it is a discrete way of telling 'underdeveloped' nations (either rightly or wrongly) that the 'Westernization' of their cultures is a necessary condition for economic growth. Cultural developmentalists want to convert others to some preferred superior way of living. Their aim is to eliminate or at least minimize the differences between peoples rather than tolerate or appreciate them as products of the creative imagination. This viewpoint has returned, at least implicitly, in anthropology as well, especially among moral activists.

Relatively few anthropologists would actually describe themselves as cultural developmentalists. Nevertheless, that stance is far more common in anthropology than many admit, especially when the topic concerns gender relations and family life practices, for example, polygamy, purdah, arranged marriage, bride-price, female circumcision, and the association of femininity with domesticity and the production of children. So along with the international human rights movement and other agents and agencies promoting Western-style globalization, there are anthropologists these days who now take an interest in other cultures mainly as objects of moral scorn. The up-from-barbarism theme of (certain versions of) Western liberalism has once again become fashionable on the anthropological scene, at least among those anthropologists who are the most 'politically correct.'

5. *Culture Theory: Some Classic Problems*

A detailed description of the core assumptions of each of the 'camps' within cultural anthropology is not possible here. Nevertheless, any review of this topic should include at least brief reference to some classic questions that are always addressed (although answered somewhat differently) by the scholars in each of the camps.

5.1 *The Problem of 'Difference': A.K.A, the Problem of the 'Other'*

One tension inherent in all anthropological interpretation is the problem of 'difference,' what to make of it and what to do about it. This is also called the problem of the 'other,' although the term 'other' is used variously in the anthropological literature. It is sometimes used to connote difference *per se* without any initial judgment of relative worth. It is sometimes used to connote unbridgeable differences. It is sometimes used to connote a solipsistic gap between self-knowledge and a mysterious or 'spectral other' whose identity can never be truly inscribed. It is sometimes used to connote the representation of 'others' as so different as to be less than or other than human, or as different in ways that condemn them to inferior status and/or justify their domination. Here I use the term to connote difference *per se*.

The problem of 'difference' inherent in anthropological interpretation is not just a problem for anthropology. It arises whenever members of different groups (for example, Jesuit missionaries and Native North American Indians; British traders and Hindu Brahmans; Western feminist human rights activists and Islamic fundamentalist women) or members of different social categories (for example, gay men and heterosexual men) encounter each other. Someone finds the encounter disturbing, puzzling, strange, or astonishing because of some apparent difference between self and 'other,' and wants to know what to make of it and (if they have the power) what (if anything) to do about it.

In the history of anthropology the apparent differences mostly concerned differences in the ideas and practices of members of different groups. What should one make of and what should one do (if anything) about such ideas as 'witchcraft,' 'ancestral spirit attack,' 'reincarnation,' or 'menstrual pollution' and practices such as 'polygamy,' 'animal sacrifice,' 'infanticide,' 'purdah,' 'child betrothal,' 'suttee,' or 'adolescent circumcision'? Such encounters are obviously hazardous and fraught with dangers of many kinds, intellectual, ethical, and political. Who is that 'some-

one' who finds 'difference' problematic and wants to know what to make of it and what if anything to do about it? Who is doing the representing of the 'other' and to which audience and to what end? Who has 'voice' and authority in such encounters? Who ought to have 'voice'?

Confronted with apparent differences between other people's ideas and practices and one's own, anthropologists have historically reacted in one or the other of three ways, which are instructive to keep in mind when surveying the fault lines in cultural anthropology today. Some, the universalists, have sought ways to minimize or erase the appearance of difference or to deny that any significant differences exist, and to treat 'otherness' as an illusion. Some, the developmentalists, have perceived in the encounter between cultures a story about a 'civilizing process.' They have argued that the more evolved and progressive cultures (those that are enlightened, scientific, ethical, educated, and rational) bear the ('white man's) burden of lifting 'others' up out of ignorance and superstition. They have argued that 'developed' cultures have an obligation to intervene if necessary to bring a halt to the monstrous or barbaric practices of other lands. Still other anthropologists, the pluralists, have argued that cultures can be different but equal, and have cautioned against cultural imperialism, suggesting that one must beware of (ethnocentrically) confusing ones own local cultural evaluations with universal scientific, practical, or moral reason.

5.2 Globalization

The narrowest definition of 'globalization' refers to the linking of the world's economies (e.g., free trade across borders) with the aim of promoting aggregate wealth and economic growth. Yet it readily expands to also include the free flow of capital and labor. A new cosmopolitan economic order gets imagined, which consists entirely of global economic organizations (the IMF, the World Bank), multinational corporations, and multicultural states with open borders. According to this rather utopian vision of a 'borderless capitalism,' goods, capital, and labor ought to be freely marketed on a worldwide scale for the sake of global prosperity. For those who adopt such a perspective any desire for an ancestral homeland or for a national identity based on religion, ethnicity, 'race,' or 'tribe' with associated restrictions on residence, affiliation, and trade is viewed as 'illiberal' and disparaged as a form of retrograde or irrational 'apartheid' or 'ethnonationalism.'

Fully expanded, however, the idea of 'globalization' actually becomes a hypothesis about human nature and an imperial call for 'enlightened' moral interventions into other ways of life in order to free them of their supposed 'barbarisms,' 'superstitions,' and 'irrationalities.' This expansive 'globalization hypothesis' makes three related claims: (a) that Western-like aspirations, tastes, and ideas are objectively the best aspirations, tastes, and ideas in the world; (b) that Western-like aspirations, tastes, and ideas will be fired up or freed up by economic globalization; and (c) that the world will/already has and/or ought to become 'Westernized.' Western-like aspirations include the desire for liberal democracy, free enterprise, private property, autonomy, individualism, equality, and the protection of 'natural' or universal 'rights' (the contemporary human rights movement is in many ways an extension of an expansive 'globalization' movement). They include the modernist notion that all social distinctions based on collective identities (ethnicity, religion, gender) are invidious. They include as well the notion that 'individuals' should transcend their 'tradition-bound' commitments and experience the quality of their lives solely in secular and ecumenical terms, for example, as measured by wealth, health, or years of life.

The true connection between 'globalization' narrowly conceived ('free trade') and 'globalization' expansively conceived (Western values, culture, and institutions taking over the world) has yet to be firmly established. Nevertheless, the picture of a cosmopolitan world of individuals without groups, in which meanings are detached or abstracted from communities and traded on a free market of ideas, has influenced the thinking of some postcultural theorists. Whether that picture is realistic remains to be seen. It is quite possible that other cultures and civilizations do not need to become just like the United States to materially benefit from participation in an emergent global economy. Modern technologies (e.g., television, cell phones, computers, weapons) and economic institutions (e.g., private property) seem to have effectively served many interests, including the interests of communitarians and religious fundamentalists all over the world. It is quite possible that a genuinely successful global political economy will not emerge, or will fail to sustain itself, or that efforts to globalize values and culture will be effectively resisted (in some cases for very good reasons), or that the world will go to war. That is how the last big push to globalize the world came to an end, with World War 1. Nevertheless, the idea that the rich nations of North America and Northern Europe have an obligation to use their economic and military power to civilize and develop the world is no less popular today than it was 100 years ago when the empire was British rather than American.

6. 'Culture': Popular Objections and Common Misattributions

Within anthropology there have been many critiques of the idea of culture. Some are associated with a fear of 'the ethnography of difference.' Some with doubts about the grounds and authority of 'ethnographic

representation.' Some with a renewed interest in, or fear of, a universal civilizing project. Some with claims about the emergence of a cosmopolitan capitalist economy. Many reasons have been advanced for doubting the usefulness of the culture concept. But are they persuasive or decisive reasons? Those who continue to embrace some variety of the Kroeber and Kluckholn definition of 'culture' tend to believe that their idea of culture does not carry most of the implications that are the supposed grounds for various anticultural critiques.

For example, the Kroeber and Kluckholn definition of 'culture' does not really imply that 'whatever is, is okay.' It is important to recognize that valid social criticism and questions of moral justification are not ruled out by the 'standard view' of 'culture.' Nothing in the Kroeber and Kluckhohn formulation suggests that the things that other peoples desire are in fact truly desirable or that the things that other peoples think are of value are actually of value. Consensus does not add up to moral truth. In other words, a definition of culture *per se* is not a theory of the 'good.' From a moral point of view, one need not throw out the idea of culture just because some tyrant puts the word 'culture' to some misuse, or because at times some ethnic groups enter into geopolitical conflict.

The idea of culture also does not imply passive acceptance of received practice or that human beings lack 'agency,' a common claim among anticulture theorists. Indeed, many proculture theorists find it astonishing to see the idea of 'agency' or 'intentionality' used as synonyms for 'resistance to culture' in the discourse of 'anticulture' theorists. Even fully rational, fully empowered, fully 'agentic' human beings discover that membership in some particular tradition of meanings and values is an essential condition for personal identity and individual happiness. Human beings who are 'liberationists' are no more agentic than 'fundamentalists,' and neither stands outside some tradition of meaning and value.

The idea of culture also does not imply the absence of debate, contestation, or dispute among members of a group. Nor does it necessarily imply the existence of within group homogeneity in knowledge, belief, or practice. Every cultural system has experts and novices; one does not stop being a member of a common culture just because cultural knowledge is distributed and someone knows much more than you do about (e.g.) how to conduct a funeral or apply for a mortgage. One does not stop being a member of a common culture just because there are factions in the community. The claim that there are between group cultural differences has never implied the absence of within group differentiation or that there is no variation around the mean. The idea of 'culture' does not imply that every item of culture is in the possession or consciousness of every member of that culture. The idea of culture merely directs our attention to those ideas about what is true, good, beautiful, and efficient that are acquired by virtue of membership in some group. Not everything has to be shared for a 'culture' to exist. Members of a cultural community do not always agree about this or that, but they do take an interest in each other's ideas about what is true, good, beautiful, and efficient because those ideas (and related practices) have a bearing on the perpetuation of their way of life, and what they share is that collective inheritance. Since the standard view does not assume that a culture is a well-bounded, fixed, and homogeneous block, the critique of the concept of 'culture' that starts with the observation of internal variation and ends 'therefore there is no cultural system' should have been a nonstarter.

The idea of culture also does not imply that other kinds of peoples are 'other,' in the sense of being less than human or possessing qualities that entitle us to intervene in their way of life. We live in a multicultural world consisting (as Joseph Raz has put it) 'of groups and communities with diverse practices and beliefs, including groups whose beliefs are inconsistent with one another.' The aspirations (a) not to lose your cultural identity, (b) not to assimilate to mainstream pressures, (c) not to be scattered throughout the city, country or world, (d) not to glorify the Diaspora, and (e) not to join the highly individualistic and migratory multinational, multiracial but (in many ways) monocultural cosmopolitan elite are real and legitimate aspirations, and those aspirations cannot be properly understood by treating them as illusions. They are certainly not the only legitimate aspirations in a multicultural world; there is much that can be said in favor of a liberal cosmopolitan life. But they are legitimate aspirations. Even in a 'global' world, cultural communities and ethnic groups are not going to disappear. We cannot avoid the question, what form does and should multiculturalism take in our emerging postmodern society (see Daedalus 2000)? Perhaps that is one reason that so many social scientists and public policy analysts look to anthropology for a useful concept of 'culture,' not for no concept of culture at all.

See also: Cultural Psychology; Cultural Relativism, Anthropology of; Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Culture as Explanation: Cultural Concerns; Deconstruction: Cultural Concerns; Ethnography; Globalization and World Culture; Identity in Anthropology; Identity Movements; Pluralism; Positivism: Sociological; Postmodernism: Philosophical Aspects

Bibliography

Abu-Lughod L 1991 Writing against culture. In: Fox R (ed.) *Recapturing Anthropology: Working on the Present*. School for American Research Press, Santa Fe, NM

Asad T 1973 *Anthropology & the Colonial Encounter*. Ithaca Press, London

Berlin I 1976 *Vico and Herder*. Hogarth Press, London
Clifford J, Markus G 1986 *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA
Culler J 1982 *On Deconstruction: Theory and Criticism After Structuralism*. Cornell University Press, Ithaca, NY
Daedalus 2000 *The End of Tolerance: Engaging Cultural Differences*. Special issue, Fall, **129**(4) American Academy of Arts and Sciences, Cambridge, MA
D'Andrade R 1984 Cultural meaning systems. In: Schweder R A, Le Vine R C (eds.) *Cultural Theory: Essays on Mind, Self and Emotion*. Cambridge University Press, New York
D'Andrade R 1995a *The Development of Cognitive Anthropology*. Cambridge University Press, New York
D'Andrade R 1995b Moral models in anthropology. *Current Anthropology* **36**: 399–408
Denzin N 1996 Epistemological crisis in the human disciplines. In: Jessor R, Colby A, Shweder R A (eds.) *Ethnography and Human Development: Context and Meaning in Social Inquiry*. University of Chicago Press, Chicago
Dumont L 1970 *Homo Heriarchicus: An Essay on the Caste System*. University of Chicago Press, Chicago
Evans-Pritchard E E 1937 *Witchcraft, Oracles and Magic Among the Azande*. Clarendon Press, Oxford, UK
Foucault M 1973 *The Order of Things: An Archeology of the Human Sciences*. Vintage Books, New York
Fox R (ed.) 1991 *Recapturing Anthropology*. School of American Research, Santa Fe, NM
Freeman D 1983 *Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth*. Harvard University Press, Cambridge, MA
Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York
Geertz C 2000 *Available Light: Anthropological Essays on Philosophical Topics*. Princeton University Press, Princeton, NJ
Harrison L E, Huntington S P 2000 *Culture Matters: How Values Shape Human Progress*. Basic Books, New York
Haynes D, Prakash (eds.) 1991 *Contesting Power: Resistance and Everyday Social Relations in South Asia*. Oxford University Press, Delhi
Holland D, Quinn N (eds.) 1987 *Cultural Models in Language and Thought*. Cambridge University Press, New York
Huntington S P 1996 *The Clash of Civilizations and the Remaking of the World Order*. Simon and Shuster, New York
Hymes D (ed.) 1972 *Reinventing Anthropology*. Pantheon Books, New York
Kitayama S, Markus H R 1994 *Emotion and Culture: Empirical Studies of Mutual Influences*. American Psychological Association, Washington, DC
Kroeber A L, Kluckhohn C 1952 *Culture: A Critical Review of Concepts and Definitions*. The Museum, Cambridge, MA
Kuper A 1999 *Culture: The Anthropologists' Account*. Harvard University Press, Cambridge, MA
Landes D S 1998 *The Wealth and Poverty of Nations: Why Some Are So Rich and Some Are So Poor*. W.W. Norton, New York
Latour B, Woolgar S 1979 *Laboratory Life*. Sage Publications, Beverly Hills, CA
LeVine R, Dixon S, LeVine S, Richman A, Leiderman P H, Keefer C H, Brazelton T B 1994 *Child Care and Culture: Lessons from Africa*. Cambridge University Press, Cambridge, UK
Markus G E, Fischer M J 1986 *Anthropology as Cultural Critique*. University of Chicago Press, Chicago
Menon U, Shweder R A 1998 The return of the white man's burden: the moral discourse of anthropology encounters the family life of Oriya women. In: Shweder R A (ed.) *Welcome to Middle Age! (And Other Cultural Fictions)*. University of Chicago Press, Chicago
Okin S M 1999 *Is Multiculturalism Bad for Women?* Princeton University Press, Princeton, NJ
Prentice D, Miller D 1999 *Cultural Divides: Understanding and Overcoming Group Conflict*. Russell Sage Foundation Press, New York
Rabinow P 1983 Humanism as nihilism: the bracketing of truth and consciousness in American cultural anthropology. In: Bellah R, Rabinow P, Sullivan W (eds.) *Social Science as Moral Inquiry*. Columbia University Press, New York, pp. 33–51
Raheja G G, Gold A G 1994 *Listen to Heron's Words: Reimagining Gender and Kinship in North India*. University of California Press, Berkeley, CA
Redfield R 1941 *The Folk Culture of Yucatan*. University of Chicago Press, Chicago
Reyna S P 1994 Literary anthropology and the case against science. *Man* **29**: 555–81
Rosaldo R 1989 *Culture & Truth: The Remaking of Social Analysis*. Beacon Press, Boston
Rosenau P 1992 *Post-Modernism and the Social Sciences*. Princeton University Press, Princeton, NJ
Romney A K, Weller S, Batchelder W H 1986 Culture as consensus: a theory of culture and informant accuracy. *American Anthropologist* **88**: 313–38
Said E 1978 *Orientalism*. Pantheon, New York
Sahlins M 1995 *How Natives Think, About Captain Cook, For Example*. University of Chicago Press, Chicago
Sahlins M 1999 Two or three things that I know about culture. *Journal of the Royal Anthropological Institute* **5**: 399–421
Sangren S P 1988 Rhetoric and the authority of ethnography. Postmodernism and the social reproduction of texts. *Current Anthropology* **29**: 405–35
Sapir E 1963 *Selected Writings of Edward Sapir in Language, Culture and Personality*, Mandelbaum D (ed.). University of California Press, Berkeley, CA
Scheper-Hughes N 1995 The primacy of the ethical: propositions for a militant anthropology. *Current Anthropology* **36**: 409–30
Shore B 1996 *Culture in Mind: Cognition, Culture and the Problem of Meaning*. Oxford University Press, New York
Shweder R A 1991 *Thinking Through Cultures: Expeditions in Cultural Psychology*. Harvard University Press, Cambridge, MA
Shweder R A 1996 True ethnography: the lore, the law, and the lure. In: Jessor R, Colby A, Shweder R A (eds.) *Ethnography and Human Development: Context and Meaning in Social Inquiry*. University of Chicago Press, Chicago
Shweder R A, LeVine R A 1984 *Culture Theory: Essays on Mind, Self and Emotion*. Cambridge University Press, New York
Sperber D 1985 *On Anthropological Knowledge*. Cambridge University Press, Cambridge, UK
Spiro M E 1986 Cultural relativism and the future of anthropology. *Cultural Anthropology* **1**: 259–86
Wierzbicka A 1997 *Understanding Cultures Through Their Key Words: English, Russian, Polish, German, Japanese*. Oxford University Press, New York
Wikan U 1996 *Culture, power and pain*. Unpublished manuscript. University of Oslo, Norway

R. A. Shweder

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Culture in Development

At present, the term 'culture' generally is used to refer to the entire body of socially inherited past human accomplishments that serves as the resources for the current life of a social group ordinarily thought of as the inhabitants of a country or region (D'Andrade 1996). Although there is evidence of the rudiments of culture in nonhuman species (Tomasello 1999), human beings are unique in their dependence upon the medium of culture and the forms of organism–environment interactions that culture supports to sustain and reproduce themselves.

1. *Defining Culture*

For purposes of thinking about culture as it relates to development, it is useful to begin by tracing the concept of culture as it has evolved since entering the English language from Latin many centuries ago (see *Culture: Contemporary Views*). Modern conceptions of culture originate in terms that refer to the process of helping things to grow: 'Culture, in all of its early uses, was a noun of process: the tending of something, basically crops or animals' (Williams 1973, p. 87). From earliest times, this notion of culture included a general theory for how to promote growth: Create an artificial environment in which young organisms could be provided optimal conditions to develop. Such tending required tools, both material (hoes) and mental (the knowledge that one does not plant until winter is over). These tools are perfected over generations and designed for the special tasks to which they were put.

2. *Culture in Development*

Combining the historical notion of culture as a process of growing things with the modern conception of culture as social inheritance of prior generations' accomplishments, the study of culture in development can be seen to focus on the way in which biologically immature human beings are incorporated into the cultural 'designs for living' that are their social heritage by more mature humans who have already been enculturated. Although the role of culture begins even before conception (for example, by beliefs and customs that restrict who the parents are likely to mate with), it is sufficient for our purposes to begin with the period of prenatal development and provide examples of culture's role in development through infancy, childhood, and the transition to adulthood referred to in many societies as 'adolescence.'

2.1 *Prenatal Development*

The cultural environment influences the growing fetus both indirectly and directly. In the category of indirect influences, mediated by the mother's physiology, are dietary laws in some societies that limit pregnant women's access to nutrients deemed important to later development (Mead and Newton 1967). In others, the use of substances such as cigarettes and alcohol, or the presence of industrial pollutants, can materially damage the fetus, depending upon such factors as timing and degree of exposure (Cunningham et al. 1997).

An important example of a direct influence of the cultural environment comes from evidence that during the last trimester, despite distortion introduced by filtering through the mother's body, fetuses are capable of hearing the language spoken in the mother's immediate environment (Lecanuet and Schaal 1996). Moreover, when they are born they show a preference for their native language, the sound stream of which is specific to the culture in which it is used (Moon et al. 1993). This universal sensitivity to language and culture-specific preference for one's own language provides the starting point for the crucial process of learning to extract meaning from the sound stream.

2.2 *Birth*

When babies emerge from the birth canal and the umbilical cord is cut, their automatic supply of oxygen and nutrients comes to an abrupt halt. Following birth, even essential biological processes occur *indirectly*—they are mediated by culture and other human beings. In order to survive in an environment mediated by culture, the baby must act on the nurturing environment in a qualitatively different way than was true before birth. While the effects of fetal activity *in utero* are minimal, once the baby is born it begins to make urgent, vocal demands upon its caregivers. Babies become social actors who re-order the social relationship among the people around them.

At the same time, babies become cultural objects for the older members of their community; their biological characteristics are interpreted in terms of the community's beliefs to a significant degree. For example, studies which present young infants wearing pink or blue diapers to adults to interact with (pink and blue being the symbolic representations of female and male gender in the society in question) show that the adults interact with them and interpret their behavior in terms of their *symbolic* gender, not their *biological* sex. Regardless of their biological sex, babies dressed in pink diapers are treated in a gentle manner and interpreted as physically pretty and sweet tempered, while babies dressed in blue diapers are bounced vigorously and have 'manly' attributes attributed to them (Rubin et al. 1974).

These forms of adult behavior illustrate a universal feature of the way in which culture mediates adult behaviors toward children: children are behaved toward in the present in terms of culturally prescribed theories that bear no necessary relationship to the

biological child with whom adults are interacting at the time.

2.3 Infancy

The role of culture in development manifests itself in a wide variety of ways during infancy. Perhaps the most direct, and certainly universal, influence of culture is in the way it shapes the development of the system of coordination between infant and caretakers known as 'getting on a schedule.' For example, rural Kenyan (Kipsigis) infants sleep with their mothers and are allowed to nurse on demand. During the day they are strapped to their mothers backs, where they often nap as their mothers go their daily work rounds. By contrast, the children of urban, middle-class North American parents generally sleep separately from their mothers and must adjust to a parental schedule which is governed by the clock (Super and Harkness 1997). As a consequence, within a few weeks of birth, North American children have begun to bunch their sleeping spells in longer and longer night episodes that approximate the adult pattern. Kipsigis children, by contrast, show comparatively slow development of long sleep spells, and remain more flexible in their sleeping habits even as adults.

Whereas the example of getting on a schedule highlights the role of contemporaneous constraints in shaping infant development, studies of mother–child interactions that incorporate objects indicate the way that cultural beliefs concerning future 'proper' behaviors influence current maternal behaviors. One such study of mothers and their 5-month old infants in Tokyo and New York found that the Japanese mothers were more responsive when their infants responded to them, while US mothers were more responsive when their infants oriented to the objects. At this early age, there was no difference in the infants' behavior at all. But at a later age, the Japanese and US children begin to differ in a manner that matches the differential parental behaviors observed at 5 months, although their language and play skills are judged to be of equivalent developmental level (Bornstein et al. 1991).

An area of development which has received special attention with respect to the role of cultural factors is the development of emotional attachments between babies and their parents (Crittenden and Clausen 2000). In a wide range of societies it has been observed that between 6 and 9 months of age children begin to display wariness of strangers, a heightened degree of distress when they are separated from their primary caretakers, and a marked tendency to cling to their parents in unfamiliar situations. The frequency and intensity of these behaviors increase for several months and then begin to wane as children approach the end of infancy at $2\frac{1}{2}$–3 years of age. This universal trend is accompanied by apparent cultural variations in the patterns of distress that children display, a result which is often interpreted as a reflection of cultural differences in the development of attachment relationships (see *Attachment Theory: Psychological*; *Cultural Psychology*; *Cross-cultural Psychology*).

For several decades, a major means of assessing qualitative differences in attachment behaviors has been to confront infants and young children with a standardized set of experiences referred to as the 'Strange Situation.' In this procedure, infants are first left in a room with their primary caretaker, who then leaves the child with a stranger and eventually returns (see Ainsworth 1967 for a more precise delineation of procedures). Children's responses to this sequence of events are coded into three major categories: *Anxious–avoidant* children turn away or run away when their caretaker returns instead of seeking closeness and comfort. *Securely attached* children go to their caregivers, calm down quickly after their early upset, and soon resume playing. *Anxious–resistant* children are often upset even when their caretaker is with them just as a result of being in the strange environment; they become excessively upset when their caretaker leaves and both reject and seek comfort when the caregiver returns.

The Strange Situation has been widely used in cross-cultural research. In middle-class US families that dominant pattern is one of secure attachment; anxious–avoidant behaviors are second most frequent. In Germany, an Anxious–avoidant pattern has been found to be dominant, while in Japan and Israel, the level of secure attachment approximates that of the US but there are high percentages of Anxious–resistant children. Psychologists are divided on the significance of these findings. Some believe that they mirror basic cultural differences (for example, reflecting a German emphasis on early independence). Others believe that the findings reflect more about the relative strangeness of the Strange Situation itself, and reject it as a measure of socioemotional development (see Cole 1999 for a summary of the issues).

Patterns of language socialization have provided a rich source of evidence on cultural universals and similarities. In societies throughout the world, caregivers speak differently with children than with adults, but there are wide variations in the specific forms of interaction. In a great many, if not all societies, adults use a special speech register (something akin to a 'baby-talk' mode with its high pitch, isolation of simplified words, and simplified grammar) when speaking to young babies. In many cultures, adults deliberately seek to teach vocabulary, style of address, and other linguistic features. The Kaluli of Papua New Guinea, for example, hold their young infants facing away from them and caregivers speak for them (Schieffelin and Ochs 1986). There is no evidence that any one of these practices enhances the rate of language acquisition in general although it certainly produces differences in the speech forms that children come to master.

Following infancy, there are wide cultural variations in the forms of experience that adults begin to arrange for their children. In rural, agrarian societies, even children as young as 3–4 years of age will be given work tasks to carry out and their older brothers and sisters will be assigned to watch over them. This form of socialization of children by children produces a distinctive pattern of interpersonal characteristics that mix nurturance and authoritarianism. By contrast, in societies where young children live in nuclear families and spend time in adult-supervised settings, interpersonal behavior is marked more by patterns of attention seeking and playful sociability (Whiting and Whiting 1975).

Even in societies where preschools are the dominant out-of-family settings for young children, the form of that experience can differ dramatically. For example, US preschool teachers are shocked and disapproving when they observe Japanese preschool classrooms with 30 children for each teacher. But Japanese teachers equally are shocked and disapproving when they observe American classrooms with one-third the number of children per adult. Tobin et al. (1989) found that these differences in both classroom organization and evaluation reflect different cultural conceptions of the optimal environment for creating a productive (Japanese or American) adult. The Japanese teachers believed that the children's development is most fully realized in their ability to feel a part of, and to coordinate with, their social group. The US teachers sought to foster independent initiative in the children reflecting the dominant norm in North American cultures. As in several previous examples, the *anticipated future* of the children, based on cultural norms and ideals, played a significant role in how adults structure the environments of children.

2.4 *Middle Childhood*

One of the most pervasive changes in the cultural organization of children's lives is the new social arrangements that societies institute when their children reach the age of 6–7 years (Sameroff and Haith 1996). Even in societies in which there is no schooling, marked changes in children's activities are likely to occur. For example, among the Ngoni of Malawi in Central Africa in the middle of the twentieth century (Read 1960), the boys, who had been living and socializing with other children of both sexes and their mothers, left the protection of the women, stopped playing childish games, and moved into dormitories where they submitted to male authority and began to engage in at least rudimentary forms of adult work. The girls lived among the women and participated in their work practices.

When we contrast the experiences of children who spend several hours a day, five days a week, attending formal schools with comparable children who remain at home helping their parents, several prominent characteristics of the classroom experience stand out (Serpell and Hatano 1997) (see *Academic Achievement: Cultural and Social Influences*; *Cross-cultural Study of Education*).

(a) The settings in which schooling occurs are removed from contexts of practical activity.

(b) There is a distinctive social structure to formal schooling, in which a single adult interacts with many (often as many as 40 or 50, sometimes as many as 400) students at a time. Unlike most other settings for socialization, this adult is unlikely to have any familial ties to the learner.

(c) There is a peculiar value system associated with schooling that sets educated people above their peers and that, in secular education, values change, and discontinuity over tradition and community.

(d) There is a special mediational skill—writing—that is essential to the activity of schooling. Writing is used to represent both language and nonverbal systems (e.g., mathematics).

(e) Language is used in quite distinctive ways in formal schools (Mehan 1978).

A wide variety of evidence indicates that extensive exposure to schooling brings about changes in the intellectual skills, values, and uses of language which it promotes. The degree to which these experiences engender general changes in development remains unclear. At present, it appears that the intellectual changes associated with schooling are rather narrowly tied to particular contents and interactional formats and do not infuse thinking in everyday life. However, changes in values and cultural norms appear to generalize well beyond the school to influence the cultural organization of the community in which the school is located. This is seen, for example, in the way that the experience of formal schooling increases the use of complex language and problem-solving strategies when mothers interact with their children (Le Vine et al. 1996).

2.5 *Adolescence: Stage or Transition?*

By and large, the various periods of development we have used to organize this discussion take for granted the reality of the periods named, reflecting broad agreement across cultures in the basic age periodization of development. When we arrive at the transition from childhood to adulthood, the situation changes (see *Youth to Adulthood across Cultures*, *Transition from*).

An interesting fact about adolescence is that in many societies, there is no word corresponding to this presumed universal stage of life. This linguistic fact raises an interesting psychological question: there may be a universal transition from childhood to adulthood, but is there a distinctive stage of adolescence independent of cultural or historical circumstances?

What is indisputable is that some time around the end of a decade of life (the exact onset time depends

greatly on nutritional and other factors), a cascade of biochemical events begins that will alter the size, the shape, and the functioning of the human body. The most revolutionary of the changes that will occur is the development of the entirely new potential for individuals to engage in biological reproduction (Katchadourian 1989). These biological changes have profound social implications for the simple reason that reproduction cannot be accomplished by a single human being. As their reproductive organs reach maturity, boys and girls begin to engage in new forms of social behavior because they begin to find other individuals sexually attractive. According to many psychologists, some combination of biological changes in the brain and changed social circumstances give rise to new cognitive capacities. Consistent with this line of reasoning, some scholars argue that adolescence is indeed a universal and necessary stage of development, with its own special characteristics (Schlegel and Barry 1991)

However, the universality of adolescence as a unified stage is by no means clearly established on the basis of these data. The evidence from other cultures may support the idea that the transition to adult status is universally fraught with anxiety and uncertainty, but it provides equally strong evidence that adolescence, as the term is used in modern industrialized societies, exists only under particular cultural circumstances (Whiting et al. 1986).

Among the Inuit Eskimos of the Canadian Arctic at the turn of the nineteenth century, for example, special terms were used to refer to boys and girls when they entered puberty, but these terms did not coincide with Western notions of adolescence (Condon 1987). Young women were considered fully grown (adult) at menarche, a change in status marked by the fact that they were likely to be married by this time and ready to start bearing children within a few years. Young men were not considered fully grown until they were able to build a snowhouse and hunt large game unassisted. This feat might occur shortly after the onset of puberty, but it was more likely for boys to achieve adult status somewhat later because they had to prove first that they could support themselves and their families. In view of the different life circumstances of these people, it is not surprising that they developed no special concept corresponding to adolescence that applied to boys and girls alike; such a concept did not correspond to their reality.

When we consider the actual organization of life in terms of the role of culture in development, we are reminded that the process of biological reproduction by itself is insufficient for the continuation of our species. It must be complemented by the process of cultural reproduction (education, broadly conceived), which ensures that the designs for living evolved by the group will be transmitted to the next generation. According to this view, adolescence will exist as a distinctive period of life only when there is a gap of several years between the biological changes that mark the onset of sexual maturity and the sociocultural changes that confer adult status (such as the right to marry without parental consent or to run for elective office).

Matters can be expected to be different in those societies in which there is little or no gap between the beginning of sexual maturity and the beginning of adulthood (Whiting et al. 1986). These are often societies in which the level of technology is relatively low and in which biological maturity occurs relatively late by our standards. By the time biological reproduction becomes possible, which is about the age of 15 in many nonindustrial societies, young people already know how to farm, weave cloth, prepare food, and care for children. Although they still have a good deal to learn (a fact that their elders impress on them, often with considerable force), they are capable of the basics of cultural reproduction. In such societies, there may be no commonly acknowledged stage of development equivalent to adolescence.

2.6 Development during Adulthood

So long as development is thought of essentially in biological terms, adolescence marks the 'high-water mark' of development; the species, imperative to reproduce is achieved and psychological processes remain more or less unchanged for a decade or so until they begin to deteriorate as a consequence of biological aging. However, for those who accord an essential role to culture in development, developmental change is a lifelong process. In fact, owing to the fact that cultural experience continues to accumulate, at the same time that biological capacities wane, serves to highlight a lifelong dynamic of bio-cultural change which is somewhat obscured so long as both biological and cultural capacities are increasing in tandem with each other (Labouvie-Vief 1981, Smith and Baltes 1999).

When the interplay of biological, cultural, and social factors which interweave to produce development are examined during adulthood and later life, one can see complex and shifting mosaics in which gradual change and predictable experiences are mixed with sudden, unexpected events—new insights, conceptual reintegration, and developmental triumphs occur side-by-side with loss of power and decline in a process which Baltes (1997) has referred to as the 'gain–loss' dynamic.

The precise contour of this dynamic depends importantly on the sociohistorical circumstances of development. In modern, industrial societies characterized by rapid change and rigidly enforced bureaucratic rules which remove adults from productive labor and the social supports such labor provides, the ensuing isolation and loss of status can combine with biological decline to hasten the decline or power and block further development. In nonindustrialized

societies where there is no social institution comparable to what we refer to as retirement, and technological change is less rapid, old people are likely to continue to play active and valued roles: among the Ashanti of West Africa in the middle of the nineteenth century, for example,

> The grandparents … on both sides are the most honored of one's kinfolk. Their position and status are of very great importance to the social system. In Ashanti it is the grandparents who are the prototypes of person and institutions commanding reverence and submission to norms of tradition (Fortes 1950, p. 276).

As early in life, the course of development during adulthood and old age depends crucially on the cultural organization of people's experience. If elderly people are placed in settings where they are confronted with new and unfamiliar events over which they have little control, they may simply give up on life (Rodin 1986). However, if special care is taken to provide experiences that increase social connectivity and are enabling and satisfying, healthy development remains a lifelong potential.

3. *Future Directions*

There is urgent need for increased research on the role of culture in development. At the same time, cross-cultural research is hampered by difficult methodological problems such as those encountered in the study of attachment. Future progress will depend upon the development of more adequate methods to ensure the validity of researchers' conclusions.

See also: Cultural Diversity, Human Development, and Education; Developmental Psychology; Historical Change and Human Development; Indigenous View on Human Development: West African Perspective; Lifespan Development, Theory of; Psychological Development: Ethological and Evolutionary Approaches

Bibliography

Ainsworth M D 1967 *Infancy in Uganda: Infant Care and the Growth of Love*. Johns Hopkins University Press, Baltimore, MD

Baltes P B 1997 On the incomplete architecture of human ontogenesis: Selection, optimization, and compensation as foundations of developmental theory. *American Psychologist* **52**: 366–81

Cole M 1999 Culture in development. In: Bornstein M H, Lamb M E (eds.) *Developmental Psychology: An Advanced Textbook*. Erlbaum, Mahwah, NJ,

Condon R G 1987 *Inuit Youth*. Rutgers University Press, New Brunswick, NJ

Crittenden P M, Claussen A H (eds.) 2000 *The Organization of Attachment Relationships: Maturation, Culture, and Context*. Cambridge University Press, New York

Cunningham F G, MacDonald P C, Gant N F 1997 *Williams Obstetrics*, 20th edn. Appleton and Lange, Stamford, CA

D'Andrade R 1996 Culture. In: Kuper A, Kuper J (eds.) *Social Science Encyclopedia*, 2nd edn. Routledge, New York, pp. 161–3

Fortes M 1950 Kinship and marriage among the Ashanti. In: Radcliffe-Brown A R, Forde D (eds.) *African Systems of Kinship and Marriage*. Oxford University Press, London, pp. 252–84

Katchadourian H A 1989 *Fundamentals of Human Sexuality*, 5th edn. Holt, Rinehart and Winston, Fort Worth, TX

Labouvie-Vief G 1981 Proactive and reactive aspects of constructivism: Growth and aging in lifespan perspective. In: Lerner R M, Busch-Rossnagel N A (eds.) *Individuals as Producers of their Own Development: Lifespan Perspectives*. Academic Press, New York, pp. 276–301

Lecanuet J P, Schaal B 1996 Fetal sensory competencies. *European Journal of Obstetrics, Gynecology, and Reproductive Biology* **68** (1–2): 1–23

Le Vine R A, Miller P, Richman A, Le Vine S 1996 Education and mother–infant interaction: A Mexican case study. In: Harkness S, Super C (eds.) *Parents' Cultural Belief Systems: Their Origins, Expressions, and Consequences*. Guilford, New York

Mead M, Newton N 1967 Cultural patterning of perinatal behavior. In: Richardson S, Guttmacher A (eds.) *Childbearing: Its Social and Psychological Aspects*. Williams and Wilkins, Baltimore, MD

Mehan H 1978 *Learning Lessons*. Harvard University Press, Cambridge, MA

Moon C, Cooper R P, Fifer W P 1993 Two-day-olds prefer their native language. *Infant Behavior and Development* **16**: 495–500

Read M 1960 *Children of their Fathers: Growing up among the Ngoni of Malawi*. Holt, Rinehart and Winston, New York

Rodin J 1986 Aging and health: Effects of the sense of control. *Science* **233**: 1271–5

Rubin J Z, Provenzano F J, Luria Z 1974 The eye of the beholder: Parents' view on sex of newborns. *American Journal of Orthopsychiatry* **44**: 512–9

Sameroff A J, Haith M (eds.) 1996 *The Five to Seven Year Shift: The Age of Reason and Responsibility*. University of Chicago Press, Chicago

Schieffelin B, Ochs E 1986 *Language Socialization across Cultures*. Cambridge University Press, New York

Schlegel A, Barry H 1991 *Adolescence: An Anthropological Inquiry*. Free Press, New York

Serpell R, Hatano G 1997 Education, schooling, and literacy. In: Berry J W, Dasen P R (eds.) *Handbook of Cross-cultural Psychology, Vol. 2: Basic Processes and Human Development*, 2nd edn. Allyn and Bacon, Boston, pp. 339–76

Smith J, Baltes P 1999 Lifespan perspectives on development. In: Bornstein M H, Lamb M E (eds.) *Developmental Psychology: An Advanced Textbook*, 4th edn. Erlbaum, Mahwah, NJ

Super C M, Harkness S 1997 The cultural structuring of child development. In: Berry J W, Dasen P R (eds.) *Handbook of Cross-cultural Psychology, Vol. 2: Basic Processes and Human Development*, 2nd edn. Allyn and Bacon, Boston, MA, pp. 1–39

Tomasello M 1999 *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA

Whiting B B, Whiting J W 1975 *Children of Six Cultures: A Psychocultural Analysis*. Harvard University Press, Cambridge, MA
Whiting J W M, Burbank V K, Ratner M S 1986 The duration of maidenhood across cultures. In: Whiting J W M, Burbank V K (eds.) *School-age Pregnancy and Parenthood: Biosocial Dimensions*. Aldine de Gruyer, New York
Williams R 1973 *Keywords*. Oxford University Press, Oxford, UK

M. Cole

Culture in Nonhuman Organisms

Whether or not nonhuman animals can be said to have 'culture' is a contentious issue that hangs on definitions of culture and the complexities of animal social behavior. However, it is well established that numerous animals have the ability to acquire from others skills, calls, and information concerning predators, mates, or resources. In some instances this information transmission is sufficient to propagate the diffusion of novel behavior patterns through animal populations, and to maintain distinct behavioral traditions between different populations of the same species. Such traditions are sometimes regarded as analogous to human culture in many respects, and have been described as 'proto-culture.' However, the use of this term indicates little more than that the behavior concerned has been learned socially, and does not necessarily imply that the animals concerned are unusually intelligent, or that they have exceptional linguistic or cognitive capabilities.

1. Social Learning in Animals

Underlying culture in nonhuman animals is a capacity for 'social learning.' The term social learning refers to learning that is influenced by observation of, or interaction with, another animal (typically of the same species) or its products. Social learning is frequently contrasted with individual or asocial learning, in which animals learn exclusively on the basis of their own personal experience, without recourse to social cues or guidance. In any instantiation of social learning, the learner or receiver of information is generally described as the 'observer,' while the transmitter of information is called the 'demonstrator.' Social learning occurs when the communication of information between two animals results in the observer learning from the demonstrator. Frequently the observer learns through social interaction to perform the same behavior as that exhibited by the demonstrator and, if repeated amongst other animals, this process can result in the spread of a particular behavior pattern and behavioral conformity in a population. Less frequently, animals may learn by observation to perform a different behavior from that exhibited by the demonstrator.

Social learning among animals has been of interest to scientists from a number of different disciplines. Psychologists traditionally have been interested in animal learning. Contemporary animal learning theory describes much learning in animals as resulting from associations formed between two external stimuli (stimulus-stimulus, or classical conditioning), or between the behavior of an animal and an external stimulus (response-reinforcer, or operant conditioning). Arguments abound as to whether there will be anything different or special about social, compared with asocial learning, and whether social learning can be adequately explained in conditioning terms. Many psychologists have suggested those more complex forms of social learning might be indicative of the animal concerned possessing unusual or sophisticated psychological capabilities (Heyes 1994). For instance, it has been suggested that for an animal to imitate the motor patterns of another animal it must be conscious, must be capable of taking the other animal's perspective, must be unusually intelligent, or must be capable of intentional action. All such claims are hotly disputed.

At the same time, social learning has been of interest to ethologists and behavioral ecologists because it seems to allow animals to learn about their environments rapidly and efficiently, without having to engage in potentially costly or hazardous learning trials, or expend considerable time and energy exploring the environment. Animals can learn which foods to eat, acquire food processing skills, learn to identify predators, learn which members of the opposite sex to mate with, or develop songs and calls, by exploiting the knowledge base of more experienced conspecifics (Heyes and Galef 1996, Zentall and Galef 1988). Social learning can be regarded as a short cut to learning about the environment, but may incur a cost if the acquired knowledge is inappropriate or outdated.

Evolutionary biologists, biological anthropologists, and archeologists have also studied animal social learning because it is accorded a prominent role in the evolution of human culture. Although definitions of human culture vary, one central feature is the transmission of acquired information between individuals through social learning processes. Many researchers believe that an understanding of the ways in which social learning operates in animal populations will generate insights into the evolutionary roots of human culture. Social learning is regarded as pivotal to a number of social intelligence hypotheses, which maintain that the evolution of high intelligence or large brains was driven by a capacity for social learning and for proto-culture (Byrne and Whiten 1988).

2. History of Research into Animal Social Learning

The study of social learning in animals dates back to Darwin, and flourished as a result of attempts to demonstrate mental continuity between humans and animals in order to prove that humans had evolved. For many Victorian scientists, demonstrations of imitative learning in animals was seen as providing evidence of an evolutionary origin of the higher mental faculties of humans. Imitation was regarded as intermediate in psychological complexity between animal instinct and human reasoning (Boakes 1984). The verb 'to ape' and the phrase 'monkey see monkey do' are manifestations of the then commonly held view that simian primates are frequent imitators of each other's behavior.

While numerous observations of imitative behavior in primates and other animals were reported in the late nineteenth century, attempts to demonstrate imitation proved problematic. It became apparent that there are many processes that can result in two animals expressing similar behavior. Researchers began to realize that anecdotal stories concerning animals in uncontrolled environments would not provide satisfactory evidence to address this issue. In the first decades of the twentieth century, Edward Thorndike, a leading psychologist, attempted to introduce some scientific rigor into investigations by carrying out controlled experiments exploring imitation in cats, dogs and monkeys (Galef 1988). Thorndike's experiments led him to the controversial conclusion that there was no evidence that animals could 'from an act witnessed learn to do an act.' Out of the controversy there emerged two distinct research traditions in the study of social learning: a laboratory experimental tradition, dominated by experimental psychology, and a second tradition that carried out observations of natural populations of animals, dominated by ethologists and primatologists. The laboratory approach has focused on investigating the psychological mechanisms underlying social learning. In contrast, field studies have reported behavior patterns spreading through animal populations, or cultural traditions. Recent years have seen the development of experimental approaches to the study of the spread of learned information through animal populations, including transmission chain and diffusion studies (Laland et al. 1993). In addition, mathematical approaches have explored the conditions under which social learning is expected to evolve, and the evolutionary consequences of social learning. (See Galef 1988 for more on the history of research into animal social learning.)

3. Laboratory Studies of Animal Social Learning

Modern laboratory experiments investigating social learning in animals commonly train animals (called demonstrators) to perform a target behavior (e.g., to push a lever to get a food reward), pair these demonstrators with untrained experimental subjects (called observers) that watch the demonstrator perform the behavior, and then test the observers alone to establish whether they have learned to perform the target behavior (Galef 1988). Frequently, the rate at which the subjects learn to perform the behavior is compared with control subjects, which have received different forms of observational experience. Such experiments have established that there are a number of ways that animals can learn socially, the variants reflecting the nature of the observer's learning and the type of information transmitted from demonstrator to observer. For instance, in the lever-pressing example, the observers might learn at an accelerated rate relative to controls, (a) because they learned of the presence or significance of the lever, (b) because they learned of the presence of food, (c) because their observational experience allowed them to form an association between the depressed lever and the appearance of food, or (d) because they formed an association between the action of pressing and the appearance of food. In addition, there are other nonsocial-learning processes that could result in an elevation in the observers' performance, for example, if observation changes its motivational state.

Such experiments have engendered a fissioning of imitative phenomena, resulting in a multitude of terms to describe the different processes that can result in social learning and spawning a host of classificatory schemes. At present there is little consensus as to the use of terms, and little agreement as to which classification is superior. Whiten and Ham (1992) distinguish four classes of social learning. 'Stimulus enhancement' occurs when the observer learns from the demonstrator to which object or location its behavior should be oriented. For instance, British birds learned to peck open milk bottles and drink the cream by having their attention drawn to milk bottles by other feeding birds (Hinde and Fisher 1951). (The term 'local enhancement' is sometimes used to describe similar phenomena). 'Observational conditioning' occurs when the observer learns to what circumstances a behavior should be a response. For example, infant rhesus monkeys learned to respond to the presence of a snake with fear and flight behavior through observations of the screams and avoidance behavior of adults. 'Imitation' occurs when the observer learns from the demonstrator to perform a specific motor pattern in a particular context. For instance, orang-utans have been reported to go through the motions of washing clothes in the presence of human caretakers that were doing the laundry. Finally, 'goal emulation' refers to instances where the observer learns from the demonstrator the goal to pursue or the result that can be achieved. For example, juvenile chimpanzees that observed a stick being used to rake in out-of-reach food appeared to learn that the food could be gained using the stick but developed their own means of doing

so. Earlier Galef (1988) identified seven processes that could result in social learning. Other researchers (e.g., Heyes 1994) have argued that social learning can be best described by using the same conceptual tools as are employed by psychologists to describe asocial learning. In general, there remain considerable doubts as to whether any of the current classification schemes are comprehensive, or whether each of a given classification's categories is mutually exclusive. The plethora of competing terms and usage reflect an ignorance of the biological basis of social learning. At present there is little understanding of the genetic, neural or physiological processes that underlie the processing of sensory information, the associationist or transformational organization of this information, and the production of motor patterns, during social learning. Nonetheless, there is little doubt that the laboratory experimental approach has greatly enhanced understanding of the processes that can result in social learning, and developed the methodologies and conceptual tools to further that understanding.

4. *Field Transmission Chain and Diffusion Studies*

In a population of Japanese monkeys, a young female began washing sweet potatoes provisioned on a beach before eating them. This previously unobserved behavior was subsequently recorded to increase in frequency in the troop, with 90 percent of individuals eventually adopting the potato-washing behavior, presumably by learning from each other. There are hundreds, perhaps thousands, of such reports of the spread of novel behavior patterns in populations of animals. Well-known examples include the transmission of dietary preferences by rats attending to cues on each other's breath, the transmission of predator recognition in birds, mammals, and fish, the learning of vocal dialects in song birds, feeding on pine cones among black rats, and tool-using traditions in primates (see Bonner (1980) for a readable overview). Probably all vertebrates and many invertebrates will eventually be found to exhibit some form of social learning. In the main, these reports are observations of acquired behavior patterns spreading through single populations. In other cases, distinct cultural variations may characterize the behavioral repertoires of different populations of the same species. For instance, much birdsong is thought to be learned from older males, and frequently regional song dialects characterize neighboring populations. Undoubtedly the most impressive case of local cultural traditions is provided by the different tool-using repertoires of populations of common chimpanzees. For instance, termiting (using a stalk to fish for termites) is reported in some but not all chimpanzee populations and is thought to be maintained within populations through social learning processes. A recent study documented 39 different behavior patterns, including tool usage, grooming, and courtship behaviors, that were habitual in some chimpanzee communities and absent in others, where ecological explanations had been discounted (Whiten et al. 1999).

These field reports are of value in documenting the breadth of animal cultures and throwing light on the contexts in which social learning occurs. However, it is difficult in the field to establish reliably whether or not a behavior genuinely is socially learned (as opposed to learned independently by each animal), to determine whether local traditions reflect different cultural repertoires or varying ecological resources, or to investigate the psychological and social processes underlying the diffusion.

More recently there have been attempts to investigate animal cultural traditions experimentally, under controlled conditions. 'Diffusion studies' are characterized by the introduction of a trained demonstrators into captive populations, and a careful monitoring of the increase in the target behavior over time (Lefebvre and Palameta 1988). 'Transmission chain studies' involve the gradual replacement of a founder population of trained demonstrators with untrained animals, and explore whether the target behavior remains in the population when the founders have been removed (Laland et al. 1993). These methods have been valuable in determining how processes of social interaction impede, or propitiate, the diffusion of information. It has become apparent that the structure of animal populations (dominance relations, spatial positioning, sex, age, and size differences) strongly influences the pathways by which learned information spreads, a phenomenon known as 'directed social learning' (Coussi-Korbel and Fragaszy 1995). For instance, novel skills often spread at different rates among juveniles and adults, among the two sexes, or within and between family groups.

Typically the new behavior pattern found in an animal culture will have been invented by a single individual. There has been comparatively little investigation of the process of innovation. It is not clear whether innovation should be characterized as a personality trait (associated with clever, creative, or nonconformist individuals), a state-dependent variable (e.g., hunger may drive foraging innovation), or resulting from environmental changes (such as a drought). Preliminary studies suggest that the adage 'necessity is the mother of invention' may be appropriate, with innovation appearing at elevated frequencies in hungry, low status, and small individuals. Across primate species, incidence of innovation co-varies with incidence of social learning (also with tool use and deception), suggesting that innovation and social learning may be manifestations of a general capacity for behavioral plasticity. However, within a given species it is not known whether innovative individuals tend to be good, average, or poor social learners.

5. *The Adaptive Value of Animal Social Learning*

Theoretical analyses have explored the circumstances under which natural selection will favor reliance on social learning, as opposed to asocial learning or evolved nonlearned behavior (see Laland et al. 1996, for an overview of this literature). A consensus has been reached that the issue hangs, in part, on patterns of variability in the environment. In an environment that is changing comparatively slowly, or that exhibits relatively little spatial heterogeneity in resources, populations are able to evolve appropriate behavior patterns through natural selection, and learning is of little adaptive value. In contrast, in rapidly changing or highly variable environments asocial learning pays. Here natural selection cannot track environmental fluctuations quickly enough, while social learning is unreliable because it may lead individuals to acquire outdated or locally inappropriate behaviors. It is intermediate rates of change and patterns of spatial heterogeneity that favor social learning, and within this window vertical transmission of information (social learning of offspring from parents) is thought to be an adaptation to slower rates of change than horizontal transmission (social learning among unrelated individuals of the same cohort). Although these findings are theoretically robust, a number of questions remain. For instance, do contemporary populations of animals capable of social learning switch between social and asocial learning depending on the pattern of environmental variability they experience in pertinent resources? Which other factors affect reliance on social learning? For instance, is social learning more likely when individuals are confronted with an unfamiliar problem, or a threatening environment?

6. *Social Learning and Social Intelligence*

Social learning has been central to a number of social intelligence hypotheses, which suggest that complex social behavior drove brain evolution and intelligence. Allan Wilson's (1985) concept of 'behavioral drive' is the hypothesis that most explicitly emphasises social learning. Wilson noted a positive relationship between relative brain size and rate of evolution in vertebrates. He explained this by arguing that big-brained species exhibit greater behavioral plasticity, are more innovative, and are more capable of social learning than small-brained species. Cultural traditions, he argued, allow populations of animals to open up new niches, and exploit novel resources, thereby exposing themselves to new selection pressures, and accelerating the rate of fixation of mutations. Wilson's hypothesis is supported by recent evidence that there is a significant correlation between relative brain size and incidence of both innovation and social learning in primates.

See also: Animal Cognition; Comparative Neuroscience; Comparative Psychology; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Primate Behavior: Significance for Understanding Humans; Primates, Social Behavior of

Bibliography

Boakes R 1984 *From Darwin to Behaviourism*. Cambridge University Press, Cambridge, UK

Bonner J T 1980 *The Evolution of Culture in Animals*. Princeton Science Library, Princeton, NJ

Box H, Gibson K 1999 *Mammalian Social Learning. Comparative and Ecological Perspectives*. Cambridge University Press, Cambridge, UK

Byrne R W, Whiten A 1988 *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Oxford University Press, Oxford, UK

Coussi-Korbel S, Fragaszy D M 1995 On the relation between social dynamics and social learning. *Animal Behaviour* **50**: 1441–53

Galef B G Jr 1988 Imitation in animals: History, definition, and interpretation of data from the psychological laboratory. In: Zentall T R, Galef B G Jr (eds.) *Social Learning. Psychological and Biological Perspectives*. Erlbaum, Hillsdale, NJ

Heyes C M 1994 Social learning in animals: Categories and mechanisms. *Biological Reviews* **69**: 207–31

Heyes C M, Galef B G Jr 1996 *Social Learning in Animals. The Roots of Culture*. Academic Press, San Diego, CA

Hinde R A, Fisher J 1951 Further observations on the opening of milk bottles by birds. *British Birds* **44**: 393–6

Laland K N, Richerson P J, Boyd R 1993 Animal social learning: towards a new theoretical approach. *Perspectives in Ethology* **10**

Laland K N, Richerson P J, Boyd R 1996 Developing a theory of animal social learning. In: Heyes C M, Galef B G Jr (eds.) *Social Learning in Animals. The Roots of Culture*. Academic Press, San Diego, CA

Lefebvre L, Palameta B 1988 Mechanisms, ecology, and population diffusion of socially-learned, food-finding behavior in feral pigeons. In: Zentall T R, Galef B G Jr (eds.) *Social Learning. Psychological and Biological Perspectives*. Earlbaum, Hillsdale, NJ

Whiten A, Goodall J, McGrew W C, Nishida T, Reynolds V, Sugiyama Y, Tutin C E G, Wrangham R W, Boesch C 1999 Cultures in chimpanzees. *Nature* **399**: 682–5

Whiten A, Ham R 1992 On the nature and evolution of imitation in the animal kingdom: Reappraisal of a century of research. *Advances in the Study of Behavior* **21**: 239–83

Wilson A C 1985 *Scientific American* **253**: 148–57

Zentall T R, Galef B G Jr 1988 *Social Learning. Psychological and Biological Perspectives*. Earlbaum, Hillsdale, NJ

K. N. Laland

Culture, Production of

The production of culture perspective focuses on the ways in which the content of symbolic elements of culture are significantly shaped by the systems within

which they are created, distributed, evaluated, taught, and preserved. Practitioners of this perspective are most likely to focus on the fabrication of expressive-symbol elements of culture such as art works, scientific research reports, religious celebrations, legal judgments, television programming, and newsmaking which, in turn, embody, modify, and give concrete expression to the 'norm,' 'value,' and 'belief' elements of culture. The perspective has also been successfully applied to a range of quite different situations where the manipulation of symbols is a by-product rather than the goal of the collective activity (Crane 1992, Peterson 1994).

Two early studies show the promise of what later came to be called the production perspective. First, in their study of the emergence of impressionist art in nineteenth-century France, Harrison and Cynthia White (1965) found inadequate the then prevalent theories that said changes in art are associated with revolutionary changes in society or with the emergence of persons of genius. Instead they found that the older Royal academic art production system survived the economic turmoil and ideological changes of the French Revolution but collapsed a generation later because of the emerging art market created by Parisian art dealers and critics who fostered unconventional artists such as the Impressionists.

Second, noting the set of nineteenth century American novels that traced the battles of lone individuals against nature, novels ranging from the frontiersman stories of James Fenimore Cooper to Herman Melville's saga of the titanic struggle of a man against a white whale, Moby Dick. Literary critics and Americanists had long asserted that these works evidenced America's distinctive culture of individualism and entrepreneurship. They contrasted such works with the novel of manners produced by English authors which were held to represent English culture. In her careful analysis, Griswold (1981) put this facile generalization in doubt. She found that the English novels of manners sold better in the United States, as well as in England, than the man-against-nature novels. Significantly, before 1909 American publishers preferred works by English authors because they could be sold in the US without paying author's copyright fees, while American authors had to be compensated. To sell their work, American authors turned to several specialty topics, including the man-against-nature theme. With the changed copyright law of 1909, American and English authors were put on the same footing, and a rapidly increasing number of American authors successfully published 'English' style novels of manners. This fact, that was once read as evidence of 'the closing of the frontier,' Griswold sees as an important result of changes in copyright law and its ramifying influences on the literary production process.

Research in the production perspective draws freely on theories, methods, and concepts developed in other branches of sociology. It is, however, distinctive in focusing on the consequences of social activities for culture. Thus, for example, the 'newsmaking' studies of the 1970s exemplify the production perspective because their authors were not satisfied to simply deconstruct the content of news or to trace the organizational and occupational dynamics of news organizations. In addition they highlight the consequences of these for the formation and transformation of 'news' (cf. Molotch and Lester 1974, Tuchman 1978, Gans 1979).

1. *Establishing the Production Perspective*

The production perspective emerged in the 1970s largely in response to the failure of the earlier dominant idea that culture and social structure mirror each other. This 'mirror' or 'reflection' view had been developed by anthropologists who studied small preliterate groups, where it seemed credible that a self-consistent set of social institutions would be reflected in a cognate coherent set of cultural symbols. The idea that there is a symbiotic relationship between a singular functioning social system and its coherent overarching culture was freely embraced by most theorists of contemporary society as well, including in the 1950s and 1960s most Marxists and functionalists such as Talcott Parsons. The former asserted that those who controlled the means of producing wealth could shape culture to fit their own class interests. The latter believed that a set of monolithic abstract values were more long-lasting and determined the shape of social structure.

The reflection view proved to be of limited use, however, in understanding the relationship between social structure and culture in industrial societies and was also being widely questioned by anthropologists studying preliterate societies. Breaking with the older mirror view, the production perspective—like most of the other contemporary perspectives in cultural sociology—views both culture and social structure as an emergent patchwork and the relationship between them as problematic (Peterson 1979).

Publication in 1976 and 1978 of anthologies entitled *The Production of Culture*, edited by Peterson and Coser, respectively, signaled the emergence of the production perspective as a coherent and self-conscious approach in culture studies, claiming to be applicable across a wide range of specific fields and genres. Many of the early researchers in the perspective had been trained in the sociology of organizations, industry, and occupations and brought their skills in the analysis of material production to the fields of symbol production. Others came to the perspective with a love of or commitment to a particular form of cultural expression such as opera, rock, dance, television, or literature. A number of developments since the 1970s show the progressive establishment in

sociology of the production of culture perspective. Two textbooks and an anthology have appeared with the title *The Production of Culture* (Crane 1992, Ryan and Wentworth 1998, Gay 1997), and the perspective receives full chapters in recent texts in the sociology of art (Zolberg 1990) and in the sociology of culture (Hall and Neitz 1993).

From the outset, the perspective promised not only a way to better understand the shape and form of any particular cultural expression but also to facilitate three kinds of comparisons: (a) comparisons across broad cultural production realms including art, popular culture, religion, science and law; (b) comparisons across the division between 'elite' and 'common' that tends to emerge in each realm; and (c) comparisons across societal, institutional, or organizational, and microlevels of analysis (Peterson 1976). Progress toward achieving these goals has been uneven (Ryan and Wentworth 1998).

We will review progress and problems in working toward these three goals and point to several emerging lines of inquiry that link the production perspective to other approaches in cultural studies, citing studies that exemplify the perspective, even when the authors have not formulated their work in production terms. We begin by showing the utility of the perspective by reviewing a few specific lines of research. This review will also briefly note, as appropriate, cognate work in other realms.

2. *The Production System in the Cultural Diamond*

Griswold (1994) has argued that to be complete, the study of any cultural phenomenon, be they books, paintings, legal systems, religious precepts, or news stories, must take into account four classes of phenomena. In addition to a close analysis of the cultural works themselves, they also focus on the situation in which people create works, the nature of the audience, and the relevant aspects of the society at large. She sees these as four points on a diamond with society at the top, the cultural work at the bottom, the creators on the left point, and the audience on the right. With this model in mind, the production system can be seen as a square overlaying the diamond, the square being large enough so that its corners intersect each of the sides of the diamond. This addition makes clear that elements of the production system intervene between each of the four facets of the diamond, so that the influence of each on the others is mediated by the production system.

To illustrate the workings of the production system in shaping culture, Peterson's (1990) analysis of why rock music became the dominant form of popular music will be used. In the early 1950s, the music industry was blind to the large and growing unsatiated demand for greater variety in music and was deaf to the efforts of musicians who might have satisfied that demand. The music industry was financially as well as aesthetically committed to the big band-crooner style of popular music of the time, and, because of its oligopolistic control of the production, distribution, and marketing of new music, was able to thwart the marketing of alternative styles. Yet in the mid-1950s the power of the music industry oligarchy was broken and the rock of Chuck Berry, Elvis Presley, Bill Haley, and Little Richard became the defining force of music in the United States and much of the world. In the middle of the quiescent Eisenhower years, no major political or economic upheaval could account for such change.

Six production factors working together are identified as making possible the rapid cultural change. These include law and regulation, technology, industrial structure, organizational structure, occupational careers, and the consumer market. In brief, with the transfer of network radio programming to the emerging new medium, television, radio turned to playing records as the cheapest effective form of programming. The granting of many new radio station licenses, the arrival of cheap transistor radios, and the development of the top 40 radio-as-jukebox format meant that the listening audience was exposed to a much larger number and far wider range of music. Using the new durable 45 rpm records, and taking advantage of the developing network of independent record distributors, numerous independent record companies experimented with a wide range of new sounds in an effort to tap the unsatiated market demand. In a matter of 24 months between late 1954 and early 1957, rock was forged in this caldron of entrepreneurial creativity (Peterson 1990, Ennis 1992). In the decades since, the large media conglomerates have gradually reconstituted their oligopolistic control of the industry, and the market has been divided into an ever larger number of increasingly homogeneous market segments (Lopes 1992, Negus 1999).

Using the same six variables employed in the analysis of the emergence of rock music, the production perspective can bring some order to a welter of different trends taking place in commercial music at the turn of the millennium. On the one hand, a few artists are accounting for an ever larger proportion of diminishing aggregate record sales, and at the same time there is an explosion in the number and diversity of artists who are making and independently distributing CD recordings. To increase their popularity with the masses of sometime music buyers, megastars such as Michael Jackson, Garth Brooks, Celine Dion, the Spice Girls, and Shania Twain make their work ever more safe and predictable. This aesthetic shift is encouraged by the marketing needs of their record companies who are aggregating into ever larger multinational media conglomerates, dependent for promotion on radio stations whose ownership is rapidly consolidating as well. These large firms can

work efficiently only if they sell millions of units of a few recordings. The majors have been able to control the market by effectively controlling the channels of marketing and distribution.

The hegemony of this star-marketing system is being challenged by a new system of promotion and distribution that is coming into being. Since the late 1980s the technology has been available to make possible the digital recording of music and, therefore, the storage, reproduction, and transmission of sound without any decay in quality. Given the increasing power of Internet technology, its maintenance as a free highway of information, and the availability of inexpensive software, making it easy for any musician to create a 'home page' and transmit sound. In this emerging field independent music producers can make, market, and distribute their work independent of the major media conglomerates.

If this system develops successfully, it can increase the power of artists, reduce the power and profitability of the large firms, foster an explosion in the number of artists and firms able to reach wide audiences, and significantly increase the levels of experimentation in popular music, creating new markets for an explosion in the range of music regularly on offer. It is as likely that, through the elaboration of copyright and censorship law, together with control of the patents for the technology necessary for the new system, the major firms will keep control of developments in popular music.

3. *The Promise of Comparative Analysis*

One of the features of the production perspective is that it facilitates comparative analysis of three sorts: comparisons across symbol-producing realms, comparisons between elite and popular forms within each realm, and comparisons across societal, institutional, and microlevels of analysis. To date, most studies in the production perspective have focused on particular fields, and progress toward fulfilling the promised benefits of comparison has been uneven, as we will show.

3.1 *Comparisons Across Symbol-producing Realms*

Mid-twentieth century functionalists argued that cross-realm comparisons are fruitless because each has its own unique core value. Essentially Parsons (1961) argued that science deals in experimentally established fact, law deals in reason, art deals in aesthetics, and religion concerns faith in the supernatural. The nominalist stance of the production perspective, however, problematizes these differences, making the assertion of uniqueness a residual to be asserted when no similarities can be established.

The expectation that the same range of processes are present in all realms of culture production, be they religion, art, science, law, etc. does not mean that new discoveries in physics are produced in the same way as is new religious doctrine. Rather it means that the differences can be explained, in large part, by the conditions of production in each of the realms (Crane 1976). For example, knowledge in physics today is largely controlled by a single subsidized and self-perpetuating worldwide 'priesthood' that works to allocate resources and rewards in ways that perpetuate its views; at the same time, religious doctrine varies greatly from country to country, and, particularly in the United States, a wide range of religious specialists mu compete with each other in an open market for the lay consumer's attention and for resources, loyalties, and rewards.

These differences, however, are not inherent in the nature of the subject matter of these two realms, as can be seen by reference to earlier times when the production systems of science and religion were very nearly reversed. In the eighteenth century, English scientists supported themselves and their work by periodic shows of spectacular findings that induced laymen to give them further support, and in the seventeenth century religious scholars worked with the support and protection of an Established Church orthodoxy and defied it only in mortal peril of their lives. Likewise, eighteenth-century French painters worked with the protection of an all-powerful Academy, while today they work in conditions of near perfect competition. In each of these instances, the movement from one sort of production system toward the other has had profound consequences for every aspect of the art world, including even what is considered to be art (White and White 1965, Becker 1982).

As recent research on the roles of women in each of the symbol producing realms has shown, much less dramatic changes in the nature of the production process may have major consequences for a particular class of producers. Systematic comparison across realms will highlight the structural and symbolic underpinnings of such institutional discrimination.

3.2 *Distinctions Between High and Low*

An invidious distinction between artifacts, practices, and practitioners considered reverentially and those who are considered to be profane is found in each of the culture-producing realms, and this fact furnishes a great opportunity for comparisons both within and across production realms. To date, however, the distinction has done more to inhibit such comparisons because the truth value vested in high (art, science, religion, law) symbols and the honor status accorded to those in the fields where they are produced is often

seen to render these fields incomparable with each other and also with the low counterparts in their own realms (Gans 1974).

Working from the nominalist stance of the production perspective, one can cut through these problems by asserting that any of the invidious high–low distinctions are not inherent in the cultural objects, but rather they involve a social, or better, a cultural construction of reality. At issue are the questions: who gets to make the designations, on what basis are they made, and within what range are they accepted. There are a number of excellent case studies touching on these issues, but there has been very little thinking about the conditions supporting one system of evaluation or another across all the realms of cultural production. The most systematic efforts in this direction can be found in the works of Pierre Bourdieu and his colleagues. In the paragraphs that follow, two issues are considered: the mobility between high–low categories, and the differentiation and de-differentiation of high and low.

3.3 Cultural Mobility

In the arts, particular works and artists, and even entire genres, experience aesthetic mobility over time (Becker 1982, Griswold 1986). DeVeaux (1997) details the sociopolitical and music industry conditions that caused the aesthetic mobility of jazz from folk to popular to fine art music. In the religious realm in the United States, there has historically been a cycle in which the established church organization becomes increasingly rigid and out of touch with people's felt needs, and a number of reformist charismatic groups form to serve the demand. As the established churches loose power the surviving reform groups become ever more established and church-like.

3.4 Changing Criteria of Prestige

The distinction between popular culture and high art that was such a fixture of American society by the middle of the twentieth century (Gans 1974) only began to be clearly articulated in the later decades of the nineteenth century. The differentiation of high–low was driven by the class interests of rising status groups, and the criteria for claiming prestige has changed in recent decades (Crane 1992). There is evidence that the highbrow exclusiveness that was at the core of cultural capital at mid-century has been replaced by a cultural capital based on cosmopolitan, even omnivorous, tastes. Volume 25, numbers 2 and 3 of *Poetics* are devoted to describing this changing representation of status through taste displays.

4. Micro, Institutional, and Societal Levels of Analysis

The production of a culture model has been formulated at the level of personal interaction, relations among organizations, and the workings of entire societies. Ideally, work at each of these levels complements work at the other two, and while this may happen in the best individual studies, the theoretical questions and research techniques employed at the three levels are different enough that research has tended to develop separately, and practitioners working at different levels of analysis do not learn all they might from each other.

The great strength of studies that explicitly take into account societal level conditions is their view of the culture industry as an expression and reinforcer of the sociopolitical system of the larger society and the attendant systems of domination and subordination, including social class, ethnicity, religion, and gender. Excellent studies can be cited. The danger in looking from the societal level is that the workings of the production process may be seen to follow automatically from the society-level constraints. This need not be the case, as exemplified in a number of empirical works which show the way in which societal values and beliefs are mediated by the culture industries.

Production studies that focus primarily on the organizational or institutional level take advantage of the decades of studies of particular organizations, occupations, industries, and the theoretical developments in these research specialties. Because researchers want to remain value-neutral, they tend to bracket their own aesthetic and political views and use uncritically the aesthetic judgments and ideological stances implicit in the symbolic field being studied. Thus the work of these researchers helps to perpetuate the established categories of the systems they study. At the same time, their reports expose for public review the workings of the systems which create and perpetuate self-serving classifications of a given field (cf. Lamont and Fournier 1992).

Production studies that focus primarily on the microlevel of interaction between persons draw from the rich tradition of research in symbolic interactionism, as well as from the social-psychological elements of the sociologies of occupations and organizations. The most wide-ranging and influential work in this tradition is Howard S. Becker's *Art Worlds* (1982). He shows that by bracketing societal and institutional levels of analysis, one can illustrate how groups of symbol producers create and are shaped by the conventions of aesthetic judgment and collaboration in culture production. In effect such microlevel studies focus less on the production of culture and more on the culture of production (Fine 1992).

Some microlevel studies can be criticized for unreflexively assuming the perspective of the group or occupational category under review, and thus obtain-

ing an incomplete and misleading reading of the field. This 'going native' as it were is cognate to the problem of becoming an unwitting apologist for the institution under study, and the research strategy of comparison across subfields is one of the best ways of guarding against unreflexively adopting the perspective of just one of the groups in a production system.

5. *The Production Perspective in Cognate Areas*

The production perspective has proved a useful model for organizing ideas and research in a number of areas where the production of culture is itself not the conscious aim of the activity.

5.1 *Reception and the Autoproduction of Culture*

Reception by the reader, listener, viewer, or purchaser has always been a central concern in cultural studies. The term 'reception' suggests the end of a communication process. However, reception is an active process of selection, interpretation, and recombination of elements that can be seen as the production of a symbolic world which is meaningful for 'consumers.' A prime question here is how people go about selecting cultural elements, and in turn, the meanings that they attach to them. This production of a lifestyle for oneself and one's group is best referred to by the Latin word for self—'auto' thus the autoproduction of culture. The autoproductions of marginalized groups such as youth, ethnics, eccentrics, and the like have received much scholarly attention, but all individuals and collectivities create lifestyles that display their patterning of cultural choices (Peterson 1983). A growing number of studies focus on how people of different sorts produce meaning out of mass-mediated symbols.

In the 1940s a 'mass culture' was seen as the unreflexive response to the seductions of the mass media (Rosenberg 1957), dissolving the social class-based hierarchy of class cultures by which wage worker, tradesperson, business person, and farmer could readily be distinguished. While the extreme elitist alarm over massification has fewer proponents now (Gans 1974), much of the debate since has focused on the question of the degree of fit between social class, defined in relationship to production, and culture class defined in relationship to consumption. Gans (1974) and Bourdieu (1984) assert that the fit between social class and culture class (groupings of people sharing patterns of consumption) is quite close. However, their perceptive analyses of the United States and France respectively reveal numerous autoproducing status groups that cannot easily be fit into a linear hierarchy.

More recent studies of autoproduction of culture suggest that most people of high status (in the United States at least) are now little concerned with the old class markers of cultural capital that centered on the appreciation of the fine arts. They no longer shun as inferior all forms of popular or mass culture (Goffman 1951). In their study of the musical tastes of the full range of occupational status groups, Peterson and Kern (1996) suggest that the cultural symbols of class position are undergoing a profound transformation. While the top strata were earlier characterized by an elitist and exclusive interest in the fine arts and avoidance of other cultural expressions, now persons and groups show their high status as cultural omnivores, consuming not only the fine arts but also showing a knowledgable appreciation of many, if not all, forms of cultural expression. Near the bottom of the status hierarchy, in sharp contrast to the old conception of the undiscriminating couch potato of the mass audience, they find groups who are univores, that is, people involved with a single form of cultural expression.

The systematic study of the autoproduction of culture is going forward rapidly along several fronts: (a) ethnographic studies of distinctive groups tease out the processes by which group members appropriate and incorporate cultural symbols for their own purposes; (b) other studies show the re-appropriation of the mass media-disseminated pop culture by youth and other specific taste groups; and (c) population surveys with increasingly sophisticated culture questions are becoming available, making possible a much richer understanding of patterns of cultural choice.

5.2 *Tradition, Collective Memory, and Boundary Formation*

Early anthropologists, folklorists, and sociologists developed the idea of 'tradition' as the repository of all that is ancient and virtually unchanging. This fitted the needs of both romantics seeking the 'noble savage' and imperialists needing a rationale for ethnic and racial exploitation. In recent decades this view has been attacked on all sides as factually incorrect and self-serving. For example, Hobsbawm (1983) in a study showing that much of the British monarchial tradition was in fact purposefully created in the middle of the nineteenth century, has coined the felicitous term 'invented tradition' to describe this kind of production of culture.

The term 'collective memory' is now often substituted for 'tradition' in recognition that the past is continually reinterpreted to fit the changing needs of the present. Some studies point to specific individuals or groups of memory 'doctors' and 'cultural entrepreneurs.' One of the most exciting applications of the idea of the mutability of collective memory has been in the study of the manipulation of memory in the service of making status distinctions along the lines of gender,

race, class, etc. Gusfield (1963), in one of the earliest works of this sort, studied the role of status politics in the efforts of the (old-family Protestant) Women's Self-Temperance League to stigmatize the alcohol-use norms of immigrant Catholics. A wide array of applications of this sort of culture-production study can be found in Lamont and Fournier's (1992) anthology, *Cultivating Differences*.

6. Prospects

Research perspectives tend to come and go over the years, so what are the prospects for the production perspective? The numerous emerging developments described above show that the production perspective continues to gain utility for significant research and theory development. The perspective has proved an essential part of any full analysis of cultural phenomena, although some practitioners in cultural studies still do not take into account the influence of production factors. The globalization of production and the high rate of technological development as well as changes in censorship law and intellectual property rights make it all the more imperative in the twenty-first century to take into account the changing influence of production factors. Whether under the distinctive designation 'production of culture' or not, future scholars will find it rewarding to ferret out the many ways in which the system of culture production influences what culture is and what it can become.

See also: Art and Culture, Economics of; Art History; Art, Sociology of; Arts Funding; Collective Identity and Expressive Forms; Cultural Expression and Action; Cultural Policy; Cultural Policy: Outsider Art; Expressive Forms as Generators, Transmitters, and Transformers of Social Power; Leisure and Cultural Consumption; Markets: Artistic and Cultural; Popular Culture; Symbolic Boundaries: Overview

Bibliography

Becker H S 1982 *Art Worlds*. University of California Press, Berkeley, CA

Bourdieu P 1984 [1979] *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA

Crane D 1976 Reward systems in art, science and religion. *American Behavioral Scientist* **19**: 719–34

Crane D 1992 *The Production of Culture*. Sage, Newbury Park, CA

Crane D, Kern R 1996 Changing highbrow taste: from snob to omnivore. *American Sociological Review* **61**: 900–7

DeVeaux S 1997 *The Birth of Be Bop: A Social and Musical History*. University of California Press, Berkeley, CA

DiMaggio P 2000 The production of scientific change: Richard Peterson and the institutional turn in cultural sociology. *Poetics* **28**

Ennis P H 1992 *The Seventh Stream: The Emergence of Rocknroll*. Wesleyan University Press, Hanover, NH

Fine G A 1992 The culture of production: Aesthetic choices and constraints in culinary work. *American Journal of Sociology* **97**: 1268–94

Gans H J 1974 *Popular Culture and High Culture*. Basic Books, New York

Gans H J 1979 *Deciding What's News*. Pantheon, New York

Gay P du (ed.) 1997 *Production of Culture: Cultures of Production*. Sage, London

Goffman E 1951 Symbols of class status. *British Journal of Sociology* **2**: 294–304

Griswold W 1981 American character and the American novel. *American Journal of Sociology* **86**: 740–65

Griswold W 1994 *Cultures and Societies in a Changing World*. Pine Forge Press, Thousand Oaks, CA

Gusfield J 1963 *Symbolic Crusade: Status Politics and the American Temperance Movement*. University of Illinois Press, Urbana, IL

Hall J R, Neitz M J 1993 *Culture: Sociological Perspectives*. Prentice-Hall, Englewood Cliffs, NJ

Hobsbawm E 1983 Introduction: Inventing traditions. In: Hobsbawm E, Ranger T (eds.) *The Invention of Tradition*. Cambridge University Press, Cambridge, UK, pp. 1–14

Lamont M, Fournier M (eds.) 1992 *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago

Lopes P D 1992 Innovation and diversity in the popular music industry, 1969–1990. *American Sociological Review* **57**: 56–71

Molotch H, Lester M 1974 News as purposive behavior. *American Sociological Review* **39**: 101–12

Negus K 1999*Music Genres and Corporate Structures*

Peterson R A 1976 The production of culture: A prolegomenon. In: Peterson R A (ed.) *The Production of Culture*. Sage, Beverly Hills, CA, pp. 7–22

Peterson R A 1979 Revitalizing the culture concept. *Annual Review of Sociology* **5**: 137–66

Peterson R A 1983 Patterns of cultural choice: prolegomenon. *American Behavioral Scientist* **26**: 422–38

Peterson R A 1990 Why 1955? Explaining the advent of rock music. *Popular Music* **9**: 97–116

Peterson R A 1994 Culture studies through the production perspective. In: Crane D (ed.) *The Sociology of Culture*. Blackwell, Oxford, UK, pp. 163–190

Peterson R A 2000 Two ways culture is produced. *Poetics* **28**

Ryan J, Wentworth W M 1998 *Media and Society: The Production of Culture in the Mass Media*. Prentice-Hall, New York

Tuchman G 1978 *Making News*. Free Press, New York

White H C, White C A 1965 *Canvases and Careers*. Wiley, New York

Zolberg V L 1990 *Constructing a Sociology of the Arts*. Cambridge University Press, New York

R. A. Peterson

Culture-rooted Expertise: Psychological and Educational Aspects

Introduction

The gaining of expertise is usually characterized cognitively, as a process of enhancing one's competence in a target domain by accumulating experience

of problem solving, understanding, and task performance in that domain. However, expertise is almost always rooted in culture in the following three respects:

(a) Many forms of expertise occur through participation in practice, in other words, under sociocultural constraints.

(b) Some forms of expertise are culturally valued; thus, practice for them is culturally organized.

(c) Consequences of expertise are sociocultural as well as cognitive: experts share particular evaluative criteria for assessing performances in their domain of expertise; their acquired knowledge and skills have a sociocultural origin; and an increase in cognitive competence is often accompanied by a fuller participation in the community of practitioners.

These culture-rooted aspects of expertise will follow and be elaborated on, and their implications for schooling and instruction will be discussed.

1. *The Acquisition of Expertise Occurs through Culturally Organized Practices or under Sociocultural Constraints*

In many domains of expertise, knowledge and skills are acquired through participating in culturally organized practices that are interesting and/or significant to the participants (Goodnow et al. 1995). Such cultural practices can be characterized as a set of sociocultural constraints that help participants produce culturally valued products and performances. There are a number of built-in factors and conditions that serve to control and direct a participant's attention or that serve to eliminate in advance a large number of logically possible procedures and interpretations. Although such constraints are to support participants' performance, rather than to contribute to their acquisition of expertise, repeated participation enhances the skills and knowledge needed to perform well in these practices.

External sociocultural constraints include (a) other people with considerable prior experience and (b) physical, symbolic, and social tools. When interacting with other people, human cognition and learning are constrained in that what they observe is not randomly selected out of almost infinite pieces of information, but is directed by joint attention to specific items. What they try to do is not a randomly chosen chain of responses from their repertory, but is often triggered by imitation, and how they interpret a set of observations is influenced by guided comprehension activity. Thus, people tend to be more competent when they collaborate than when they work alone. Therefore, when the mastery of skills requires a lot of experience, the opportunity to gain such experience is often provided through apprenticeship. Novices or newcomers are able to participate in practices as apprentices, with little or no pretraining, because they are initially assigned to relatively simple and peripheral parts of the practice, and come to fulfill gradually increasingly difficult and central roles as they gain better skills and an understanding of the total practice. For example, Lave (1988) reported that apprentice tailors in Liberia are given such jobs as fixing buttons, which are simple and repairable tasks, but constitute authentic and essential parts of dressmaking. To generalize, when the master can manage to maintain a balance between the goals of increasing productivity and (the apprentice's) learning, and when the apprentices are committed to enhancing their skills, this learning arrangement can be effective and enjoyable.

Tools or shared artifacts also serve to direct the user's attention and narrow the range of procedures and interpretations they consider. For example, a pedometer draws the user's attention to how far they have walked. A vending machine helps the user find how to convey their request for an item and pay for it, because the machine does not respond to operations performed out of their proper sequence. A set of tools can function as enabling constraints to help participants to perform a task successfully, and also to acquire knowledge and skills. For example, it is often much easier to actually perform steps in cooking than to state the steps verbally or to comprehend them, because human actions are highly restricted by the shared artifacts that are available (e.g., utensils and appliances).

That novices become experts by incorporating culture may be a misleading phrase, because what really occurs is the individual (re)construction of knowledge and skills under sociocultural constraints, with the resultant cognitive products being analogous to the knowledge and skills found in the culture. The emphasis on culture is compatible with the notion of the human construction of knowledge and the internal (cognitive) constraints imposed on it.

2. *Some Kinds of Expertise are Culturally Valued*

Most domains for expertise are apparently optional in technologically advanced societies; that is, individuals are allowed to choose the domains in which they pursue expertise. However, this does not mean that all domains are equally respected and valued in each culture. On the contrary, each culture contains a wide variety of 'cognitive values' (Goodnow 1990), which differentiate domains of expertise in terms of their significance. Some areas of knowledge, for example, are described as 'basic' while others are described as 'extra.' For the former type, people growing up in the culture have almost no choice than to seek mastery in them. National sports constitute a good example of such, almost obligatory, domains of expertise. Many Fijian children are interested in playing rugby, pri-

marily because their culture denotes, though implicitly, that being good at rugby is more significant than expertise in other sports.

Cultures organize practices, the products of which are valued in a particular culture, so that a large number of people can acquire the knowledge and skills needed. The process of cultural imposition need not take a compulsory form. More often than not, developing individuals tend to choose those domains of expertise that are valued by their culture, because the culture provides them with many opportunities to participate in relevant practices, and because people close to them tacitly encourage them to do so.

Similarly, each sub-culture of a given domain or each community of practitioners has its particular set of values. For example, abacus operators are socialized in terms of the values they hold regarding the importance of abacus skills and their status in general education, as well as a respect for the speed of calculation (Hatano 1997). In fact, the community of abacus educators and skilled operators constitutes a strong pressure group in the world of education in Japan. Skilled abacus operators calculate speedily, but they are motivated to perform their operations even faster.

What kinds of experience or participation encourage the development of particular values or forms of privilege? Experts' values are undoubtedly forms of culture in the mind, acquired through internalization. They serve as the source of motivation for experts to excel at the target domain. Although knowing little about the way in which these values emerge, it can be assumed that they are gradually internalized in the course of gaining expertise, through conversing with other practitioners, acquiring sophisticated evaluative criteria, and spending time and effort in and for the domain.

3. *Consequences of Expertise are Sociocultural as well as Cognitive*

Gaining expertise is far from purely cognitive. It is a social process (Lave and Wenger 1991); more specifically, a process in which beginners become fully qualified members of the community of practitioners. The case of gaining expertise in abacus operation exemplifies this sociocultural nature of expertise (Hatano 1997). Abacus school pupils are usually first enrolled by their parents while in elementary school. The parents may believe that the exercise there will foster children's diligence and punctiliousness as well as enhance their ability to calculate and to make numerical estimations. Pupils are first motivated to learn abacus skills to receive parental praise, especially by passing the qualification exam. Like many other out-of-school domains of learning in Japan, abacus learning has an elaborate qualification system and frequent qualification exams. However, the pupils' motivation changes when they join an abacus club at school or become a representative of the abacus school. At this point, they compete in matches and tournaments, like tennis or chess players. Also like these players, abacus club members not only engage in deliberate practice (Ericsson et al. 1993) at least a few hours every day, but also think about how to improve their skills. Their learning is strongly supported by the immediate social context of the club and the larger one of the community of abacus operators. Formal and informal relationships with an instructor, oldtimers, and peer newcomers organize their ways of life, and sanctions from other players regulate their daily activities. Moreover, they may participate in the community of abacus operators by taking an administrative role in players' organizations or by serving as an examiner or judge at matches and tournaments. They begin to increase their participation, by increasing their significant responsibilities within the community.

Consequences of expertise are sociocultural in another sense. The knowledge and skills acquired in the course of gaining expertise have a sociocultural origin, in that they are shared by mature members of the community, before being incorporated by individuals who are gaining expertise. In other words, such knowledge and skills constitute built-in constraints in the relevant practices. Even novices have learned, to a considerable extent, the methods and concepts used in practice that are needed to perform accurately. They have also learned, though to a lesser degree than experts or intermediates, to use the methods and concepts readily and skillfully. Most, if not all, of the knowledge and skills of experts have a sociocultural origin, although the experts may have added something new to their domains. Experts' performance is often more standardized than novices.' For example, wine experts are different from non-experts in their ability to readily classify a set of different kinds of wine almost uniformly in terms of the type of grape and region of production (Solomon 1997). Moreover, they have acquired a complex set of evaluative criteria by which they can aptly judge the quality of innovative as well as conventional products or performances in the domain (Oura and Hatano 1989).

4. *Educational Implications*

What educational implications can be derived from the aforementioned culture-rooted aspects of expertise? To the extent that the acquisition of subject-matter knowledge in educational practices is considered analogous to the process of gaining expertise, the above discussion is relevant. Educators might organize classroom activities, so that students are

supported by sociocultural constraints, at least in the initial phase of learning. They might try to establish a local community within the school in which the target subject-matter knowledge is highly valued. Also, educators might invite students to be involved in designing learning activities and evaluation of their products. In fact, these principles have been incorporated into the 'community of learners' approach to school reform (e.g., Brown 1994).

See also: Education: Cultural and Religious Concepts

Bibliography

Brown A L 1994 The advancement of learning. *Educational Researcher* **23**(8): 4–12

Ericsson K A, Krampe R T, Tesch-Römer C 1993 The role of deliberate practice in the acquisition of expert performance. *Psychological Review* **100**: 363–406

Goodnow J J 1990 The socialization of cognition: What's involved? In: Stigler J W, Shweder R A, Herdt G (eds.) *Cultural Psychology: Essays on Comparative Human Development*. Cambridge University Press, Cambridge, UK, pp. 259–86

Goodnow J J, Miller P J, Kessel F (eds.) 1995 *Cultural Practices as Contexts for Development*. Jossey-Bass, New York

Hatano G 1997 Learning arithmetic with an abacus. In: Nunes T, Bryant P (eds.) *Learning and Teaching Mathematics: An International Perspective*. Psychology Press, Hove, UK, pp. 209–31

Lave J 1988 *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge University Press, New York

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK

Oura Y, Hatano G 1989 *Development of Aesthetic Evaluation of Music Performances: Effects of Age and Experience*. Paper presented at the Tenth Biennial Meeting of International Society for the Study of Behavioral Development, Jyväskylä, Finland

Solomon G E A 1997 Conceptual change and wine expertise. *Journal of Learning Sciences* **6**: 41–60

G. Hatano and Y. Oura

Culture Shock

Anyone who has spent time in a culture significantly different from his or her own has more than likely experienced a sense of disorientation and perplexity at life in the new culture. After an initial period of excitement about the novelty of the new culture and perhaps pleasure in the absence of some of the constraints of one's own society, the traveler may be plunged into a period of anxiety and frustration at the difficulty of understanding and making himself or herself understood. The visitor may feel confused by differences in expectations and in interpersonal relationships, abhorrence at some cultural practices or perceived dispositions of the people, followed by intense homesickness, depression, and perhaps regression to unaccustomed bouts of annoyance and labile feelings. Frequent manifestations include food concerns—yearning for familiar foods or anxiety about the safety of the drinking water.

Such reactions on encounter with a new culture have been referred to in anthropology as 'culture shock,' a term which has been in currency since 1950, when Kalervo Oberg gave a talk of that title to the women's club of Rio de Janeiro—a group of wives of American businessmen and diplomats. Bronislaw Malinowski, although he did not use the term, gave one of the classic descriptions of culture shock in 1922, in the introduction to *Argonauts of the Western Pacific*:

> I well remember the long visits I paid to the villages during the first weeks; the feeling of hopelessness and despair after many obstinate but futile attempts had entirely failed to bring me into real touch with the natives, or supply me with any material. I had periods of despondency, when I buried myself in the reading of novels, as a man might take to drink in a fit of tropical depression and boredom.

Culture shock has been noted frequently in anthropological fieldwork. It has become a regular part of anthropological thinking about culture, especially in the USA. Yet discussion of its manifestations has been comparatively rare in anthropology, only one major book being devoted to it. Some more specialized spinoff concepts, such as Toffler's 'future shock,' have been more extensively written on. Culture shock has been best described, perhaps, in some fictional works by anthropologists and others—Laura Bohannan's *Return to Laughter* (see Bowen 1954); E M Forster's *A Passage to India*—or in briefer introspective works by anthropologists such as Mary Catherine Bateson (1968) and Jean Briggs (1970). George Devereux's *From Anxiety to Method in the Behavioral Sciences* (1968) still remains one of the best theoretical explorations of its significance.

The earliest writings on culture shock treated it as a kind of pathology, an obstacle to smooth operation in the culture one is entering and to communicating with its people. Kalervo Oberg listed a number of 'symptoms' of this 'transitory disorder,' including: compulsive handwashing, emotional lability and excessive preoccupation with cleanliness. The psychoanalyst Gertrude Ticho, in a 1971 article, expands this list to include

> increased rigidity, fatigue, withdrawal anxiety, depression and elation, distressed obsessive neurotic-like symptoms such as over concern with the cleanliness of food and water, and hysterical reactions.

She goes on to add that

> at times, after a period of enchantment with and enthusiasm for the new culture, a hostile reaction to the people of the foreign culture ensues. This reaction becomes endowed with more global prejudices, such as 'they are lazy and un-

trustworthy, rigid, treacherous and sabotage progress.' The previously open-minded, outgoing 'foreigner' avoids contact with the people of the host country and withdraws to 'his own people,' if possible, thus creating an enclave within the foreign country.

'If this pattern is not reversed,' she concludes, 'the gap between the different cultures may be widened and result in a vicious circle' (Ticho 1971, pp. 324–5).

More recently, culture shock has been seen as an essential part of adapting to an unfamiliar culture, of coming to understand ways of living and of seeing the world that are different from one's own. Recent anthropological writers have viewed the process in different ways:

(a) as a loss of one's own culture, with a resulting process of grieving or mourning the loss;

(b) as a 'rite of passage,' a kind of 'initiation' marking the transition from membership in one culture to at least being conversant with another;

(c) as a 'resocialization' into another culture; and

(d) as a process like therapy or psychoanalysis for reorganizing one's personality, resolving old conflicts along new lines (see Kracke 1987).

The first of these, 'mourning the loss of one's own culture,' along with the old view of culture shock as a 'transitory psychological crisis,' may be grouped together as a 'negative' view of the phenomenon, emphasizing the crisis of moving *out of* one's previous 'culture of orientation.' The last three—'rite of passage,' 'resocialization' and the analogy with change in psychotherapy—are 'positive' viewpoints, emphasizing the transition *into* the new culture, the learning of new ways of thinking and reacting.

1. *Phases of Culture Shock*

While everyone's experience of cultural transition is unique, depending on one's own personality and past experiences, and on the cultures involved—the entry into another culture tends to develop in common stages. Unless the entry into the culture is under duress (capture, exile, or fleeing from a natural or political disaster at home), one's first experience tends to be one of enchantment or delight. One may also experience shock at aspects of the new culture like poverty, dirt (or antiseptic cleanliness), or noise and confusion (for a rural person arriving for the first time in the city); but at first, one is likely to feel curiosity about exploring the new culture. For a while, one may take pleasure in the absence of some of the constraints of one's home culture, at the warmth of emotions people express (if you come from a reserved culture), or the willingness of people to help you. After a time, frustrations begin to mount as one realizes, for example, that the directions people give may be more based on desire to be helpful than on a real knowledge of the information asked for, or that invitations are not to be taken at face value. One becomes aware of not knowing the cues for judging situations or people's intent, and finds oneself in awkward situations, or making gaffes. After a certain period, which varies from a few days to a few months, one begins to feel a certain 'culture fatigue' (as George Guthrie has termed it (Szanton 1966)—an irritation at the accumulating aggravations which may mount to an exasperation with the culture itself.

At this point, if one does not give in to the inclination to leave, one begins the process of learning the new culture: what William Caudill (1961) calls 'resocialization.' The phases of distress and learning are reiterated in cycles as one is more and more deeply socialized into the culture, gradually approaching a phase of being more at home in it. The process is to some extent that of reconstructing one's identity, or constructing an alternate identity for one's sojourn in the culture. Over years of living in a new culture, the process may culminate in the formation of a new identity, as Garza-Guerrero (1974) suggests, combining patterns and values from one's culture of origin with patterns and values of the new one.

2. *You Can't Go Home Again*

An unexpected feature of culture shock is the disorientation one experiences on *returning* to one's culture of origin—a phenomenon which has been designated 'reverse culture shock' (Hertz 1977). Many people feel even more difficulty in readjusting to their *own* culture after returning from abroad than they did adjusting to the foreign culture or cultures they have lived in. Why is this? In part, it is the phenomenon which the novelist Thomas Wolfe identified in *You Can't go Home Again*: You have been changed by your encounter with another culture, and you no longer 'fit' in the same place you did before going. You have an experience you do not share with your compatriots who remained at home. It is perhaps exacerbated by the very unexpectedness of it: you have had dreams of coming back home where everything will at last be unproblematic, and suddenly it is not.

Reverse culture shock reveals something about the nature of the culture shock experience itself: it measures the depth that one is changed by the encounter with another culture. We come back a different person from what we were; we have taken something of the other culture into ourselves, and can never again be fully at home in any one culture.

3. *Aspects of Culture Shock: The Epistemology of Knowing Other Cultures*

Culture shock may be an important, if not essential, part of the process of learning another culture, the central task of social or cultural anthropology. In a brilliant article, Mary Catherine Bateson (1968) discusses her own experience of culture shock in a very personal way. Culture shock, she suggests, is, among other things, a useful tool for recognizing where the

new culture differs from one's own. Where something about the new culture annoys or angers you, disgusts you, or arouses anxiety or a sense of disorientation, that is a point where one must look for fundamental differences in assumptions about life or in conceptualizing the world. In our western culture, for example, we view death and mourning as something highly private, and avoid talking with the bereaved person about the loss for fear of intruding on their privacy. Filipino culture, which regards bereavement as something that must be worked through with others, so that friends talk openly with the bereaved about the loss, helping the bereaved person to re-experience the loss and the grief in a context of concerned support. Initially, she was shocked at the direct way in which Filipinos ask the bereaved person about the loss, often provoking tears; by the time she came to experience a deep loss of her own in the Philippines, she was prepared to find 'deeply healing' the opportunity to articulate and vent her feelings of loss.

Another view of the nature of cultures leads George Devereux (1968) to formulate a deeper view of the source of culture shock. In a psychoanalytic viewpoint, he suggests that different cultures have different 'distribution of repressions': some feelings and fantasies are acceptable and may be talked about and accepted in oneself, others are repugnant and should not be acknowledged. The latter, in general, will be repressed, or at least disavowed, by most people in that culture. If something that is regarded as deeply disturbing and repugnant in one's own culture is considered acceptable and openly talked about in the culture you are entering into contact with, that may lead to considerable anxiety on your part. On the other hand, if something you consider normal and not anything to be ashamed of is regarded as repugnant in the culture you are going into, that may lead you to do or say things that will cause great distress to others in that culture; you may inadvertently arouse anxiety and distress, without at first understanding why.

4. Culture Shock and Mental Illness

Culture shock has received more attention in psychiatric and psychoanalytic literature, where the propensity of cultural estrangement to precipitate emotional problems or mental illness has been widely recognized. A number of articles have been written by psychiatrists and psychotherapists on the need for psychotherapeutic treatment for people whose culture shock has led them to emotional crisis. The Nigerian psychiatrist T. Adeoye Lambo has discussed certain psychiatric problems of African students abroad. Prakash Desai and George Coelho have written about the adaptation problems of Indian immigrants in the USA. Juana Antokoletz and many others have written on problems of Mexican and other Latin American immigrants in the USA. One of the most thorough discussions of culture shock was published by the expatriate Cuban psychoanalyst Garza-Guerrero (1974). Recently, the anthropologist George Spindler has developed a 'cultural therapy' to deal with problems of culture shock and intercultural communication.

Why is the encounter with another culture so unsettling? Perhaps it is because each encounter with a different way of life puts in question our own beliefs and values, confronted as we are with alternative ways of looking at reality or at social relationships, with different rules to guide our conduct. Each culture we come into contact with raises these issues anew, challenging new aspects of our world view and ethical standards. The depth to which these experiences disturb us is an indicator of how deeply the cultural framework we grow up in shapes our view of the world and of relations with others. It may be that this experience, the discomfort we feel in encountering another culture, is what most moves our curiosity about other cultures: the stimulus to anthropological study.

See also: Alienation: Psychosociological Tradition; Assimilation of Immigrants; Cultural Assimilation; Cultural Contact: Archaeological Approaches; Diversity and Disagreement; Ethnic Identity/Ethnicity and Archaeology; Ethnic Identity, Psychology of; Migration and Health; Migration: Sociological Aspects; Multiculturalism; Multiculturalism, Anthropology of; Multiculturalism: Sociological Aspects; Racial Relations

Bibliography

Bateson M C 1968 Insight in a bicultural context. *Philippine Studies* **16**: 605–21

Furnham A, Bochner S 1986 *Culture Shock: Psychological Reactions to Unfamiliar Environments*. Methuen, London

Bowen E S, (Bohannan L) 1954 *Return to Laughter*. Harper & Row, New York

Briggs J 1970 Kapluna daughter. In: *Never in Anger: Portrait of an Eskimo Family*. Harvard University Press, Cambridge, MA

Caudill W 1961 Some problems in transnational communication: Japan–United States. In: *Group for the Advancement of Psychiatry Symposium Number 7: Application of Psychiatric Insights to Cross-cultural Communication*.

Devereux G 1968 *From Anxiety to Method in the Behavioral Sciences*. Mouton, The Hague, The Netherlands

Garza-Guerrero A C 1974 Culture shock: Its mourning and the vicissitudes of identity. *Journal of the American Psychoanalytic Association* **22**: 408–29

Grinberg L, Grinberg R 1989 *Psychoanalytic Perspectives on Migration and Exile*. Yale University Press, New Haven, CT

Hertz D G 1977 The problem of 'reverse' culture shock. In: Boroffka A, Pfeiffer W M (eds.) *Fragen der transkulturell-vergleichende Psychiatrie in Europa*. Westfälische Wilhelms-Universität, Münster, Germany

Kracke W 1987 Encounter with other cultures: psychological and epistemological aspects. *Ethos* **15**: 56–79

Malinowski B 1922 Introduction: the subject, method and scope of this enquiry. *Argonauts of the Western Pacific*. Dutton, New York, pp. 1–25
Read K E 1965 *The High Valley*. Scribner, New York
Szanton D 1966 Cultural confrontation in the Philippines. In: Textor R B (ed.) *Cultural Frontiers of the Peace Corps*. MIT Press, Cambridge, MA
Ticho G 1971 Cultural aspects of transference and counter transference. *Bulletin of the Menninger Clinic* **35**: 313–34

W. Kracke

Culture, Sociology of

The word 'culture' has completely different meanings in the social sciences and in current language. In the latter, the meaning of the term has for long been mainly influenced by the humanistic conception of culture, which derives from the Roman *cultura animi*—a term that Cicero seems to have invented in order to translate the Greek concept of *paideia*. This is the classical sense of the word, based upon high productions of mind and art. Referring to the intellectual, esthetic, and ethical aspects of human beings (for which the German language uses the term *Bildung*), culture thus involves the legitimacy of value judgments and of the idea of a hierarchy between societies as much as inside them.

In the social sciences, on the contrary, the term has generally a widely inclusive meaning, which is actually aimed against the humanistic normative conception of culture, since almost anything may be considered as 'culture.' This conception, which opens an almost unlimited field to the term 'culture,' has been the basis of the ensemble of theories which may be termed 'culturalist trend.' It is characterized by a systematical relativism, strongly backed up by the fact that the overall conception of 'culture' is in itself a powerful relativistic tool. The popularization of the relativistic use of the term 'culture' in postmodern societies since the late 1900s poses a serious danger for culture.

1. *The Term 'Culture' in the Social Sciences*

This term has received, first in anthropology and afterwards in sociology, a very specific meaning, that prevailed despite the reluctance of very important authors, such as Morgan, Durkheim, and Radcliffe-Brown to use it in this way.

1.1 *The Tylorian 'Everything is Culture' Conception*

At the basis of the concept of culture as it is mostly used in the social sciences stands the well-known founding definition of the English ethnologist Edward Barrett Tylor (1832–1917), in his book *Primitive Culture* (1871). He borrowed much of the sense he gave to the term from Bismarckian Germany, whose conception of *Kultur* included any kind of expression of the national character. According to the German usage, Tylor takes culture to be *synonymous with civilization*, and therefore gives to that word what he calls 'a wide ethnographic sense.' For him, culture is 'that complex whole' produced by people's historical experience, of which he gives several examples: 'knowledge, belief, art, morals, law, custom.' But this enumeration is not a closed one. Tylorian culture also 'includes … any other capabilities and habits acquired by man as a member of society.' This definition is characterized by the huge breadth of its application.

Derived from the German term *Kultur*, the idea of a broad ethnographic sense of the word 'culture' perfectly suited Anglo-Saxon ethnologists from German or Austrian origin. The German-born US anthropologist Franz Boas (1858–1942) and the British anthropologist Bronislaw Malinowski (1884–1942), who was born in Poland as an Austrian subject, both took an outstanding part in the victorious settlement of the term 'culture' in the very core of the social sciences.

Although since Tylor many other definitions of culture have been formulated, his 'everything is culture' conception remained a basic creed for nearly all anthropologists during the twentieth century, and it still prevails. In its anthropological sense, culture includes religion, law, art, and literature, as well as all of everyday life in its often prosaic and even trivial aspects, such as cooking, sexual habits, or sanitary practices.

In the early 1900s, Emile Durkheim, Marcel Mauss and other French sociologists were reluctant to use the term 'culture' in its Tylorian sense, and preferred to go on using the word 'civilization.' Yet, the anthropologists' broad use of the term 'culture' has progressively spread to sociology. Nevertheless, during the 1950s, many American intellectuals, like Hannah Arendt, did resist the idea of mass culture, which they deemed to be no culture at all, but a deadly threat to it. In the early 1960s, American sociologists still regarded mass culture as a subject of debate, and the mass media as a most important field for sociological studies. Meanwhile, it was a French sociologist, Edgar Morin (1962) who took mass culture as an essential sociological field of study.

1.2 *Radcliffe-Brown's Attempt to Resist Tylorian Culture*

The founder of social anthropology, the British anthropologist Alfred Reginald Radcliffe-Brown (1881–1955) based his theory upon the study of 'social structures,' along the lines of Durkheim's sociological concept of function. In order to keep social anthropology completely separated from cultural anthropology, for some 20 years Radcliffe-Brown and his followers avoided using the word 'culture' as much as possible. Then, in 1952, he defined culture as 'the

process by which a person acquires, from contact with other persons or from such things as books or works of art, knowledge, skill, ideas, beliefs, tastes, sentiments.' So culture is for him the process of intellectual, esthetic, and ethical human development. Culture is the process by which one becomes a cultured person.

Though an anthropologist, and although his functionalist theory is as much relativistic and holistic as culturalism is, Radcliffe-Brown has a classical and humanistic conception of culture. This is because the term has no part in his system, not being his own relativistic tool. He states (1952) that 'the total social structure of a society together with the totality of social usages in which that structure appears and on which it depends for its continued existence' are constituent of a social system, whereas most anthropologists would have regarded all that as culture. Since he had decided to take as basic the social structure instead of the 'culture patterns' of cultural anthropology, Radcliffe-Brown had no need of the Tylorian sense of culture. In fact he had even every reason to refuse that use of the word, which was the basis of a powerful rival theory. Hence his humanistic formulation. Furthermore, Radcliffe-Brown's functionalism derived from Morgan, Durkheim and the whole French sociological school. As they had not admitted the Tylorian use of the term 'culture,' Radcliffe-Brown was naturally induced to follow their example.

However, Malinowski, the other great name of anthropological functionalism, regarded a broad conception of culture as natural. Posthumously published in 1944, his most important book is entitled: *A Scientific Theory of Culture and Other Essays*. Thus, in the 1930s and 1940s, anthropological functionalism was divided into two rival schools: 'social structure' against 'culture,' Radcliffe-Brown's school versus Malinowski's school, one of their motives of disagreement being opposing ideas about the concept of culture. In the UK, this quarrel was considered as opposing social anthropology to cultural anthropology. But, at the same time, nearly all US anthropologists were adherents of cultural anthropology since studying 'cultures.'

As a result of that, in the USA, since the end of the 1940s, the functionalist theory has been taken over, in an somewhat attenuated way, by sociologists, chiefly Robert King Merton and Talcott Parsons, who adhered to the social anthropologists' point of view. Accordingly, they stressed society and not 'culture' as the subject of their research. US sociologists thus stood opposed to US cultural anthropologists. So that the past rivalry of the functionalists schools of Malinowski and Radcliffe-Brown had become a professional quarrel opposing US anthropologists and sociologists, since the former claimed to deal solely with 'culture,' and the latter with 'society.'

Yet, in the early 1950s, some of Radcliffe-Brown's followers began to use the term 'culture' with a very broad meaning: for Raymond W. Firth, it included both the way of life and the content of social relations (1951), and for Meyer Fortes, both social structure and organization (1953). Afterwards, in 1958, American anthropologists and sociologists made an end to their quarrel by means of a peace treaty signed by A. L. Kroeber and Parsons, who mutually acknowledged the validity of one another's field of study: society as well as culture. The 'everything is culture' conception had won.

2. *The Culturalist Trend*

The Tylorian 'omnibus conception of culture' (Singer 1968) has been the basis of the set of theories known as 'culturalism,' and it remains the basis of the culturalist way of thought which still reigns over anthropology and a large part of sociology.

In the first half of the twentieth century, US cultural anthropology has been principally influenced by Tylor, Boas, and through him by the German diffusionists, like Fritz Graebner, who had introduced the idea of *Kulturkreis* (culture area). In the 1930s, US cultural anthropologists began to develop the idea that each 'culture' was an individual unity, well-organized and functioning as a system. This *holistic* and *systemic* conception was clearly formulated by Malinowski in the article on 'culture' that he wrote for the *Encyclopædia of the Social Sciences* (1931). A few years after, Margaret Mead (1935), Ralph Linton (1936) and Abram Kardiner (1939) emphasized the idea of a 'culture' being an integrated totality. The aim of ethnologists and anthropologists became to show the *inner coherence* of each 'culture' functioning as an organized whole, in which any component element is linked to all the others.

For the ethnology and cultural anthropology of the 1930s–1950s, the organization of culture systems consisted of a complex architecture of 'patterns of culture,' which Ruth Benedict had endeavored to describe and classify (*Patterns of Culture*,1934). Hence the definition formulated in 1952 by Alfred L. Kroeber and Clyde Kluckhohn:

> 'Culture consists of patterns, explicit and implicit, of and for behavior, acquired and transmitted by symbols, constituting the distinctive achievement of human groups … '

Each society is regarded as a distinctive cultural whole, characterized by its peculiar system of values. The beliefs existing in a given society are thought to be a product of 'culture.' Following Linton, several culturalist authors even affirmed that the personality structure of the individuals in a particular group was determined by the group's 'culture.' The most extreme advocate of this opinion was Kardiner, with his idea that each sociocultural system shaped a basic personality, the 'I' just being a 'cultural precipitate.' Therefore, culturalists deem socialization, the social process through which the main values of a society are

transmitted from one generation to the next, to be of the utmost importance, and they believe it to be a very deterministic process (see *Action, Theories of Social*; *Socialization, Sociology of*). On the whole, those theories led to an extremely mechanic conception, which attributes to 'culture' the production of beliefs, and to inculcation and conditioning the power to put them into people's minds and to determine their behaviors. So the US anthropologist Leslie A. White, the leader of the culturalist school called 'culturology,' went so far as writing (1968) that 'man is, and remains, a puppet of his culture.'

Culturalism has rightly been reproached for its belief in the ability of social systems to work in such a mechanical way that people so much believe what they have to believe that they feel the need to do what they have to do. It has also been established that the culturalists' holistic conception of a system of values common to a whole society was generally untrue, except perhaps in the very simplest ones. Hence the notion of 'subculture' used by culturalists for their study of complex modern societies (see *Subculture, Sociology of*). As for the culturalist principle of an inner coherence of culture systems, which comes from a counterhistorical method, it maybe partly exist in the most simple societies, but certainly not in complex ones. Against this assumed idea of a whole inside linkage of each 'culture', the French sociologists Raymond Boudon and François Bourricaud also observed (1982) that the 'everything is culture' conception was based upon an erroneous obviousness, so that it would be more adequate to use the term 'culture' only when speaking of artifacts and mental products.

A few anthropologists have dared to criticize the mechanical conditioning model of culturalism. It is chiefly the English anthropologist Robin Horton (1973) and the American anthropologist Melford E. Spiro (1987) who disputed the creed according to which belief and behavior did not result from reasons but were just the product of cultural conditioning. Nevertheless, since Boas the culturalist point of view does represent the mainstream in anthropology. And it has also influenced a great deal of sociology, and even philosophy. For instance, Richard Rorty asserts in a perfectly culturalist way that the idea prevailing in democracies of totalitarianism being evil is just the result of cultural conditioning (1989). The culturalist trend always works for the benefit of relativism, and the mere use of the term 'culture' in an 'everything is culture' sense does just the same.

3. *The Comprehensive Sense of 'Culture' : A Relativistic Tool*

As an evolutionist, Tylor naturally regarded nineteeth-century England as the pinnacle of civilization. But since he intended 'to treat mankind as homogeneous in nature,' he felt the need of a concept synonymous with 'civilization,' but more applicable to everybody and every society, past as well as present. Hence his widely inclusive definition of culture, which permitted him to find evidences of 'culture' among all societies and human groups, 'though placed in different grades of civilization.' Since everything was asserted to be culture, it became possible to find 'culture' everywhere. The Tylor revolution granted a 'culture' even to the most primitive peoples.

Yet, the most famous ethnologist in the 1870s and 1880s, the American evolutionist Lewis H. Morgan (1818–81), did not adopt the Tylorian meaning of the term 'culture.' In his book *Ancient Society* (1877), he went on using the notion of 'society'. Instead of Tylor's stages of 'culture', he prefered to delineate 'ethnical periods,' which he considered to be developmental stages leading, through barbarism, from savagery to civilization. But Morgan died and Tylor prevailed, and in the USA, at the end of the 1800s, Boas, being of German origin, naturally used the Tylorian sense of culture.

It was also Boas who threw out of anthropology the very idea of grades of civilization determining a hierarchy of societies and human groups. Rejecting the evolutionist theory, he declared the total relativity of 'cultures.' The 'Boas revolution' closely followed the Tylorian one. However, through Boas, *cultural relativism* actually proceeded from Nietzsche, who had brought all traditional values to trial, and announced that they were now delusive, since deprived of basis by the decline of dogmatic creeds. As it is well known, Nietzsche sought to destroy the very idea of objective reality, so that he is the father of modern relativism. The influence of Nietzschean ideas accounts for Boas's basic postulate of the relativity of values in the different 'cultures' and his complete rejection of the value judgments of nineteenth-century evolutionism.

In the 1940s, cultural relativism received, through the American anthropologist Melville J. Herskovits's *Cultural Anthropology* (1948), the support of Ernest Cassirer's German historical idealism (1942). According to Herskovits, exactly following Cassirer, human beings are living in a symbolic universe built by them: the universe of 'culture,' and so know no reality but the symbolic, that is, cultural. Herskovits asserts that all human experience of the physical world as well as of society are determined by enculturation, and accordingly culturally mediated. In this theory, all perceptions, evaluations, and judgments are a function of the cultural system to which one belongs, and are relative to it. Since Herskovits as well as Cassirer consider that all 'reality' is perceived through the screen of 'culture,' this one is deemed to be the measure of all things.

The Tylorian sense of 'culture' has been very useful for cultural relativists, as it permitted them to dispense with the word 'civilization' in order to avoid the

closely related ideas of hierarchy and comparative judgment. Being based upon the determinist assumption that values are a function or product of each 'culture,' cultural relativism does infer that all value judgments are necessarily culturally conditioned or determined. Therefore they are deemed to have no validity outside the cultural context from which they proceed. Value judgments concerning alien 'cultures' are therefore strictly forbiden by cultural relativism, which reproaches those judgments with being the product of a preference for one's own 'culture' and a prejudice against the other ones: namely ethnocentrism, which twentieth-century relativistic anthropology has regarded as the supreme offense, as much as evolutionism itself. Since the relativistic postulate of comparative judgment having no validity is asserted to be accurate, the corollary cannot but be the principle of an equivalent validity of all value systems. So that the American anthropologist David Bidney could observe (1968):

> 'Reverence for cultural values, rather than reverence for life, becomes the absolute virtue advocated by the cultural relativist.'

The term 'culture' used in a comprehensive sense has been the main vector of the *egalitarianist assumption* ensuing from cultural relativism, according to which all value systems must be considered to be equally legitimate units. If the meaning of the word 'culture' is taken to include ways of life as well as social relations, there is 'culture' everywhere. Whereas the humanistic idea of culture stands opposite to the absence of culture, or even the barbarity, of certain individuals, groups, or societies, the very fact of absence of culture or barbarity is denied by using the term 'culture' about these individuals, groups, or societies. Since everybody belongs to a social structure and a value system, everybody is regarded as belonging to a 'culture' or a 'subculture.' Sociologists as well as anthropologists consider each society and each social group to have its 'culture,' and they usually strictly obey the relativistic prohibition of comparative evaluations of values, which would lead to classifications showing the factual existence of various grades of civilization. The widely inclusive conception of 'culture' of the social sciences is used with the purpose of eliminating the very idea of uncultured groups. It allows indeed the social scientist to report the presence of 'culture' in the working class or in a primitive tribe as well as in the most learned circles. The 'everything is culture' use of the word 'culture' works just like an optical illusion. It brings about that fallacious value of equality of all societies and groups which is imperatively ordered by cultural relativism. The term 'culture' taken in a comprehensive sense is a necessity of cultural relativism, of which it is the most indispensable tool as well as the very basis.

The theory according to which all reality is symbolic is in no way cogent. As has been noticed by Raymond Boudon and François Bourricaud (1982), if that means that all human experience is mediated by a symbolic system like articulate speech or science, it is merely a commonplace statement. And if symbolic is taken as synonymous with ficticious, it is an erroneous idea. Likewise, the rejection of value judgments actually lacks logical validity. As a matter of fact, the assumption of an equivalent value of all 'cultures' is itself nothing but a value judgment, hidden behind the virtuous looking façade of cultural relativism. The prohibition of value judgments thus results from a value judgment. As emphasized by Allan Bloom (1987), the assertion of values and 'cultures' being relative is by no means proved: it is just a philosophical postulate, which stems from a political intention. And when cultural relativism disclaims freedom of judgment and rational analysis in order to prevent the development of any prejudice against alterity, it is just a way to avoid subjecting our own prejudices to logical criticism. Hence, the social sciences have the utmost need, as David Bidney (1968) had already stressed, 'to transcend the limitations of both cultural relativism and ethnocentrism through the pursuit of scientific truths concerning facts and values.'

However, the relativistic trend prevails, particularly in anthropology. Since nearly all twentieth-century anthropologists have repudiated evolutionism, Leslie A. White (1968) and his school being the main exception, anthropology has been, through the twentieth century and at the present time, essentially relativistic. Despite the endeavors of some authors like Melford E. Spiro (1987) and Robin Horton (1973), or like the British anthropologist Robin Fox (1989) who has turned to sociobiology (see *Biology's Influence on Sociology: Human Sociobiology*), anthropology and cultural relativism may almost be taken as synonyms. The US social anthropologist Clifford Geertz did affirm (1984) that cultural relativism was justified by the factual obviousness that value systems are of course disparate. According to Geertz (1984), relativism has such an obvious validity that it is impossible to resist it. Yet, what is most obvious is that relativism is the core and basis of culturalism, and that twentieth-century anthropology works as a *relativistic machine*. Accordingly, the relativists' fancy for anthropology is in no way surprising. They have often stated anthropology to be the queen of sciences. The cognitive relativist Feyerabend even asserted (1975) that it was the sole genuine science.

The use of a broad sense of the word 'culture' as a relativistic tool is also present in the other social sciences, especially in sociology. Such is the case in the neomarxist functionalist theory of the French sociologist Pierre Bourdieu (1979). In the 1960s and 1970s, Bourdieu emphasized the statement that culture (in the classical and humanistic sense of the term) was nothing but a 'culture' expressing the beliefs and behavior of the 'dominant class.' According to him, this 'culture' is just a 'culture' amongst many others,

but it is imposed as the only legitimate one by school, universities, and other cultural institutions, in order to bring about what he terms 'social reproduction.' Since they have been brought up in other 'cultures,' working-class children are considered by Bourdieu to be handicapped when facing at school a 'culture' which is not their own one, while children from the dominant class have all the means to be successful with it. So, stresses Bourdieu, the domination of the same upper class can be reproduced from generation to generation. Bourdieu deems culture to have in itself no real superiority, but just to be useful for the members of the dominant class in order to distinguish themselves from the rest of the society. He regards artistic or literary value, which he believes to have no actual existence in itself, as merely resulting from the preferences of the dominant class, and hence as being an illusion. Therefore, the combination of Bourdieu's Marxist ideology with his perfect relativism leads to an absolute *cultural nihilism*.

In sociology, the culturalist holistic idea of values being the product of closed wholes has probably been taken in the most absolute way by ethnomethodologists. According to them, 'cultures', subcultures, and any other groups are so much closed units that their values have solely an inside meaning. And they even assert that observers can understand nothing of the cultural products of a group in which they are not immersed. So ethnomethodologists deny the validity of outside observers' studies, affirming that their supposed understanding would be distorted by the fallacy of sociocentrism. Although ethnomethodologists do not consider, as most culturalists do, social actors to be 'puppets of their culture,' ethnomethodology is certainly an extremely relativistic sociological theory, based upon an 'everything is culture' conception.

Several streams of sociology explain beliefs and behavior by the mechanical model of inculcation and conditioning, just as does cultural anthropology. So does the neomarxist functionalism of Bourdieu (1979), with the theory or *habitus* thought as a class-conditioned behavior. But the idea of inculcation and conditioning by means of socialization is also present in the whole Durkheimian trend, and even to a certain extent in Parsons 'functionalism', since it particularly stresses social integration. And, of course, behaviourists following B. F. Skinner's positivistic line regard moral instincts as a product of social conditioning.

As they have to study the diversity of 'cultures,' and as this term has long been linked to the idea of relativism, anthropologists have a professional slant to take cultural relativism as granted. But the prevailing of relativism in the social sciences is very much strengthened by the general *relativistic atmosphere* which characterizes postmodern societies. Relativism has invaded literary criticism, with the deconstructionistic theory of the French Heideggerians Roland Barthes and Jacques Derrida, and the French Heideggerian Michel Foucault has initiated a relativistic interpretation of history. As Allan Bloom (1987) and Raymond Boudon (1990) have pointed out, all those theories mainly derive from the Nietzschean assertion that there are no facts, but only interpretations: a famous phrase that Barthes, Derrida, Foucault and their followers wrongly separated from its context, so that they unduly took it literally. These same authors' theories have been used as a justification by the relativistic movement known as *political correctness*, which since the 1980s has dominated US universities. Relativism has even been epistemologicaly theorized (see *Relativism: Cognitive*), chiefly by Thomas S. Kuhn (1962) and Paul K. Feyerabend (1975). Feyerabend's anarchist cognitive theory, which does not differentiate science from archaic myths or from magic, is clearly a counter-epistemology. With Feyerabend's well-known saying, 'anything goes,' relativism leads to nihilism.

The lack of validity of these theories has been pointed out by Raymond Boudon (1990, 1995), who demonstrated that cognitive relativism was based upon an epistemological archaism. He also observed, about the assertion of an obvious legitimacy of relativim, that since Geertz (1984) considered all conceptions of the world to emanate from 'cultures,' relativism itself was one of these conception. Therefore, in a culturalist perspective, relativism in no way can be a methodological rule that is above suspicion, but just a plain 'cultural product' of postmodern societies. In fact, it is because relativism does fit in with them that it is at the present time so fashionable. Here lies its strength, which is extremely dangerous for culture.

4. *Cultures versus Culture*

The 'everything is culture' conception is no longer confined to anthropological and sociological books and journals, but has spread beyond, so that the broad sense of 'culture' now belongs to everyday language. It is daily used by the media, and has become popularized.

There is even in several countries a ministerial department entrusted with 'culture'—an idea proceeding from the USSR. Such a department was created in France as early as 1958, and departments were established in the 1980s and 1990s in some other Western democracies, such as Belgium, Spain, Italy, and Germany. The French Culture Ministry has adopted since the 1970s, and still more since 1981, a comprehensive sense of 'culture.' In the sociological research ordered by this department, the different ways of filling spare time, and all kinds of fun and entertainment, are called 'cultural habits.' In a political purpose, the 'wide ethnographic sense' of culture is officially acknowledged by the French government.

The term 'culture' as it is used in postmodern societies, particularly in the media, has mostly nothing to do with genuine culture, but is more frequently used in order to refer to beliefs, behaviors, ways of life, customs, habits, and even the whole of society: 'culture' is now often used instead of the term 'identity.'

Having a relativistic effect, a broad use of the term 'culture' puts everything on the same level, so that fun and entertainment are regarded as equal to genuine culture. As pointed out by Allan Bloom (1987), the failure of culture becomes a 'culture.' Though the present democratic societies formally worship 'culture', the popularization of the relativistic use of the term actually destroys in the whole society the very idea of the eminent value of culture. Here is indeed the main danger for culture.

Raymond Boudon (1999), following Tocqueville, has shown that relativism is the very theory congruent with democratic societies, as it conciliates their egalitarian passion (see *Egalitarianism, Sociology of*) with the variety of the opinions. Hence, through the popularized relativistic use of the term 'culture', it is democratic egalitarianism that stands as an adversary of culture.

See also: Civilizations; Cultural Relativism, Anthropology of; Culture: Contemporary Views; Ideology, Sociology of; Taste and Taste Culture; Values, Sociology of

Bibliography

Arendt H 1961 *Between Past and Future*. Viking Press, New York

Bell D 1976 *The Cultural Contradictions of Capitalism*. Basic Book, New York

Bénéton P 1975 *Histoire de mots: culture et civilisation*. Presses de la Fondation Nationale des Sciences Politiques, Paris

Bidney D 1968 Cultural relativism. In: Sills D L (ed.) *International Encyclopædia of the Social Sciences*. Macmillan, New York, pp. 543–7

Bloom A 1987 *The Closing of the American Mind*. Simon and Schuster, New York

Bosch J 1992 *Cultura y contracultura*. Emecé, Buenos Aires, Argentina

Boudon R 1990 *L'art de se persuader des idées douteuses, fragiles et fausses*. Fayard, Paris

Boudon R 1995 *Le juste et le vrai*. Fayard, Paris

Boudon R 1999 *Le sens des valeurs*. PUF, Paris

Boudon R, Bourricaud F 1982 *Dictionnaire critique de la sociologie*. 1st edn. Presses Universitaires de France, Paris

Boudon R, Clavelin M (ed.) 1994 *Le relativisme est-il résistible*? 1st edn. Presses Universitaires de France, Paris

Bourdieu P 1979 *La distinction*. Editions de Minuit, Paris

Bourdieu P, Passeron J C 1966 *Les héritiers*. Editions de Minuit, Paris

Bourdieu P, Passeron J C 1970 *La reproduction*. Editions de Minuit, Paris

Feyerabend P 1975 *Against Method*. NLB, London

Fox R 1989 *The Search For Society, Quest For a Biosocial Science and Morality*. Rutgers, University Press, New Brunswick, NJ

Geertz C 1984 Distinguished Lecture: Anti anti-relativism. *American Anthropologist* **86**(2): 263–78

Harouel J L 1984 *Essai sur l'inégalité*. PUF, Paris

Harouel J L (1994) 1998 *Culture et contre-cultures*, 1st edn., Presses Universitaires de France, Paris

Horton R 1973 Durkheim and the scientific revolution. In: Horton R, Finnegan R (eds.) *Modes of Thought*. Faber, London

Malinowski B 1944 *A Scientific Theory of Culture and Other Essays*. The University of North Carolina Press

Morin E 1962 *L'Esprit du temps*. Grasset, Paris

Singer M 1968 The concept of culture. In: *Sills D L (ed.) International Encyclopædia of the Social Sciences*. Macmillan, New York, pp. 527–43

Spiro M E 1987 *Culture and Human Nature*. University of Chicago Press, Chicago

Steiner G 1971 *In Bluebeard's Castle*. Faber, London

White L A 1968 Culturology. In: Sills D L (ed.) *International Encyclopædia of the Social Sciences*. Macmillan, New York, pp. 547–51

J.-L. Harouel

Curiosity, Psychology of

Curiosity is an important motivator of animal and human behavior. It is a critical motive underlying educational achievement, entrepreneurship, and scientific discovery. Less positively, curiosity has been found to play a key role in such nonsanctioned behaviors as early sexual experimentation, voyeurism, and illicit drug use.

Positioned at the crossroads of cognition and emotion/motivation, curiosity is not only of practical importance but of theoretical interest. As Malone (1981, p. 363) notes, curiosity 'is evoked by the prospect of modifying higher level cognitive structures … [and] … can be thought of as a desire to bring better "form" to one's knowledge structures.' Hebb (1955) and Nissen (1954) both argued that curiosity is a manifestation of the brain's natural tendency toward cognitive processing. According to Nissen (1954, p. 300), 'the requirements of all tissues is that they perform their normal functions … It is positively painful to deny any organ the exercise of its usual function … It is the function of the brain to perceive and to know.'

Research on curiosity has been stymied by divergent interpretations of the concept. Perhaps the most important demarcation lies between 'specific' and 'diversive' curiosity (Berlyne 1960). Specific curiosity refers to the desire for specific information, as exemplified by one's desire to know the perpetrator in the Jon Benet Ramsey case. Diversive curiosity refers to

unfocused exploratory behavior as exemplified by a rat's initial exploration of a new cage. In addition to the different types of curiosity, researchers have examined curiosity-like motives without clearly specifying their relationship to curiosity. For example, the 'need for cognition' (Cohen et al. 1955) is defined as 'a need to … to understand and make reasonable the experiential world.' Likewise, research on 'interest' seeks to understand what particular topics people find interesting and why.

1. Two Waves of Curiosity Research

Curiosity research has unfolded in two waves. The first peaked around 1960 and addressed the question of curiosity's underlying cause. The second, which peaked in the late 1970s but continues into the early twenty-first century, has focused on the problem of measuring curiosity.

1.1 The Underlying Cause of Curiosity

One early account of curiosity, articulated by William James, then extended by McDougall, viewed curiosity as closely related to fear, in the sense that both are produced by similar stimuli. To illustrate the close connection between curiosity and fear, James cited the behavior of an alligator he had observed swimming toward a man seated on the beach, 'gradually drawing near as long as he kept still, [but] frantically careering back as soon as he made a movement' (James 1950. An alternative perspective proposed by Freud characteristically viewed curiosity as derivative of the sex drive (Aronoff 1962, Voss and Keller 1983). Somewhat later, Bühler and colleagues (Bühler et al. 1928) proposed that curiosity could be interpreted as a manifestation of the 'investigatory reflex'—the label that Pavlov (1927) gave to the tendency of his canine subjects to orient themselves toward novel stimuli.

The first well-researched and coherent account of curiosity was offered by Berlyne (1960). Struck by similarities between curiosity and drives such as hunger, Berlyne viewed curiosity as a type of drive. In support of this view, he argued that curiosity exhibits drive-like qualities: it has strong motivational force, intensifies over some interval, and produces arousal if not 'satisfied' (see Loewenstein 1994, p. 81).

In contrast to Berlyne's drive theory, White (1959) argued that curiosity results from the motivation to master one's environment. Reinterpreting the phenomena that Berlyne had cited in support of his perspective, White argued that curiosity has none of the characteristics usually associated with physiological drives such as hunger. When one becomes satiated from eating one type of food, for example, almost all food decreases in appeal. In contrast, satisfying one's curiosity in one area has a negligible impact on curiosity in other areas (see Fowler 1965). White's analysis was extended by Deci, who proposed to subsume curiosity under 'the more general realm of all intrinsically motivated behaviors' (Deci 1975, p. 53).

Yet a third theoretical interpretation of curiosity was proposed by Hunt (1963), who saw curiosity as a response to violated expectations or perceived environmental incongruity, both of which signal an incomplete understanding of the environment. Hunt's theory of curiosity drew on the writings of Hebb and Piaget, and others, which had emphasized the importance of violated expectations as interrupters or motivators of behavior. Hunt's incongruity account of curiosity was refined by Kagan (1972) in a classic paper on motivation that postulated four basic human motives: the motive to resolve uncertainty, sensory motives, anger, and hostility, and the motive for mastery. The first he saw as synonymous with curiosity. Kagen extended Hunt's analysis in part by pointing to sources of uncertainty not discussed by Hunt: incompatibility between ideas, incompatibility between ideas and behavior (both based on Festinger's cognitive dissonance theory), and the inability to predict the future.

Most recently, Loewenstein (1994) proposed an 'information gap' account of curiosity that integrates insights from incongruity theories, Gestalt psychology, and behavioral decision theory. According to this theory, curiosity arises from the perception of a gap between what one knows and what one wants to know. The theory attempts to make sense of three characteristics of curiosity: its intensity (particularly as compared to the often-subdued pleasure one derives from the satisfaction of curiosity), its transience, and its association with impulsive behavior.

1.2 Measurement of Curiosity

A variety of paper-and-pencil scales have been developed to measure an individual's state or trait curiosity (e.g., Day 1971, Naylor 1981, Spielberger 1975). Other attempts to measure curiosity have relied on teacher evaluations, peer evaluations, and self-evaluations (Maw and Maw 1972), as well as the tendency to explore curiosity-evoking items (Coie 1974).

Research on the measurement of trait and state curiosity has produced a number of interesting findings. For example, positively worded items that measure curiosity typically load on a separate, and largely uncorrelated, dimension from negatively worded items. These results have been interpreted as indicating that curiosity (measured by positively worded items) and boredom (measured by negatively worded items) do not lie on a unidimensional continuum, but rather reflect different underlying phenomena.

The use of curiosity scales to measure individual levels of curiosity or to study group differences is, however, inherently problematic. Inevitably, to measure a person's curiosity, one has to measure their curiosity about something, but different people are curious about different things. Thus, if older people register as highly curious on a particular scale, it may be because that scale includes items that evoke curiosity in the elderly. The near intractability of this problem may help to explain why studies of group differences in curiosity have yielded contradictory relationships between curiosity and age, gender, and other demographic factors.

2. *Two Gaps in the Curiosity Literature*

With the current focus on individual and group differences, two aspects of curiosity have received inadequate attention. First, little research has examined curiosity's situational determinants. Second, few investigations have studied curiosity's neural underpinnings.

The lack of attention to situational determinants is unfortunate because insights in this area could have much greater practical ramifications than measures of individual or group differences. Most people are curious some of the time—often intensely so—but at other times completely lacking in curiosity. Knowledge of the situational determinants of curiosity could aid in the development of educational interventions to stimulate curiosity or interventions to reduce youthful curiosity about harmful drugs.

Educational psychologists have made the greatest progress on situational determinants in their attempts to stimulate curiosity in the classroom. For example, Lowry and Johnson (1981) found that group interactions which promoted epistemic conflict stimulated students' curiosity. In another investigation of situational determinants, Loewenstein (1994) tested and confirmed the main prediction of his information gap account of curiosity—that curiosity to learn a specific piece of information increases with the accumulation of background information.

The second major void in the curiosity literature is investigation of its neural underpinnings. In one of the few discussions of this topic, Panksepp (1998) argues that the mesolimbic dopamine system, an area where local application of electrical stimulation evokes exploratory and search behavior, plays a critical role in curiosity. Studies that have measured the response of neurons in this system find that they are activated by novel stimuli that elicit orienting reactions, but that for most stimuli this activation only lasts for a few presentations (Schulz et al. 1997). One study also found that electrical stimulation of a subregion of this area produced a subjective response in one human subject that was reminiscent of curiosity (Heath 1963). Writing of an electrical stimulus repeatedly chosen by this subject, Heath noted that it

> did not, however, induce the most pleasurable response; in fact, it induced irritability. The subject reported that he was almost able to recall a memory during this stimulation, but he could not quite grasp it. The frequent self-stimulations were an endeavor to bring this elusive memory into clear focus.

It is also worthy of note that the same region plays a central role in Berridge's account of 'wanting' and 'liking' (see Berridge 1999). Traditional accounts of motivation assume that people want to engage in actions because they expect to like the consequences of those actions. However, Berridge and co-authors have found that wanting and liking often diverge, and they attribute such divergences to the separation of the two systems in the brain. Analogous to Panksepp's view that the mesolimbic dopamine system is the seat of 'seeking' behavior, Berridge argues that it plays a central role in 'wanting.' The idea that curiosity may be centered in the wanting, but not in the liking system, is of interest given Loewenstein's (1994) argument that one of curiosity's salient features is that the intensity of the motivation to possess information—the wanting—is often vastly out of proportion to the pleasure—i.e., liking—that one experiences upon obtaining the information.

Positioned at the junction of motivation and cognition, the investigation of curiosity has the potential to bridge the historical gulf between the two paradigms. The two most promising avenues for future research are the situational determinants and neural underpinnings of curiosity. An enhanced understanding of situational determinants could produce significant practical benefits and could also provide evidence supporting or challenging some of the alternative theories that have been proposed (which differ in terms of the situational determinants they identify). A better understanding of the neural underpinnings of curiosity could help to clarify issues such as whether exploratory behavior and specific curiosity are closely related phenomena in the sense of sharing underlying neural mechanisms.

Bibliography

Aronoff J 1962 Freud's conception of the origin of curiosity. *Journal of Psychology* **54**: 39–45

Berlyne D E 1960 *Conflict, Arousal, and Curiosity*. McGraw-Hill, New York

Berridge K C 1999 Pleasure, pain, desire, and dread: Hidden core processes of emotion. In: Kahneman D, Diener E, Schwarz N (eds.) *Well-Being: The Foundations of Hedonic Psychology*. Russell Sage Foundation Press, New York

Bühler C, Hetzer H, Mabel F 1928 Die Affecktwirksamkeit von Fremdheitseindrücken im ersten Lebensjahr. *Zeitschrift für Psychologie* **107**: 30–49

Cohen A R, Stotland E, Wolfe D M 1955 An experimental investigation of need for cognition. *Journal of Abnormal and Social Psychology* **51**: 291–4

Coie J D 1974 An evaluation of the cross-situational stability of children's curiosity. *Journal of Personality* **42**: 93–116

Day H I 1971 The Measurement of Specific Curiosity. In: Day H I, Berlyne D E, Hunt D E (eds.) *Intrinsic Motivation: A New Direction in Education*. Holt, Rinehart and Winston of Canada, Toronto, Canada
Deci E L 1975 *Intrinsic Motivation*. Plenum, New York
Fowler H 1965 *Curiosity and Exploratory Behavior*. Macmillan, New York
Heath R G 1963 Electrical self-stimulation of the brain in man. *American Journal of Psychiatry* **120**: 571–7
Hebb D O 1955 Drives and the C.N.S. (Conceptual Nervous System). *Psychological Review* **62**: 243–54
Hunt J McV 1963 Motivation inherent in information processing and action. In: Harvey O J (ed.) *Motivation and Social Interaction Cognitive Determinants*. Ronald Press, New York
James W 1950 *Principles of Psychology*. Holt, New York, Vol. 2, Chap. 24, p. 429
Kagan J 1972 Motives and Development. *Journal of Personality and Social Psychology* **22**: 51–66
Loewenstein G 1994 The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin* **116**: 75–98
Lowry N, Johnson D W 1981 Effects of controversy on epistemic curiosity, achievement, and attitudes. *Journal of Social Psychology* **115**: 31–43
Malone T W 1981 Toward a theory of intrinsically motivating instruction. *Cognitive Science* **54**: 333–69
Maw W H, Maw E W 1972 Differences between high- and low-curiosity fifth-grade children in their recognition of verbal absurdities. *Journal of Educational Psychology* **63**: 558–62
Naylor F D 1981 A state-trait curiosity inventory. *Australian Psychologist* **16**: 172–83
Nissen H W 1954 The nature of the drive an innate determinant of behavioral organization. In: Jones M R (ed.) *Nebraska Symposium on Motivation* **2**: 281–320
Panksepp J 1998 *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, New York
Pavlov I P 1927 *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Clarendon Press, Oxford, UK
Schultz W, Dayan P, Montague P R 1997 A neural substrate of prediction and reward. *Science* **275**: 1593–9
Spielberger C D 1975 The measurement of state and trait anxiety: Conceptual and methodological issues. In: Levi L (ed.) *Emotions: Their Parameters and Measurement*. Raven Press, New York, pp. 713–25
Voss H G, Keller H 1983 *Curiosity and Exploration: Theory and Results*. Academic Press, New York
White R W 1959 Motivation reconsidered: The concept of competence. *Psychological Review* **66**: 297–333

G. Loewenstein

Curriculum as a Field of Educational Study

The word 'curriculum' appears to be quite straightforward in meaning. Most people would identify with the definitions in *Webster's Third New International Dictionary*: 'The whole body of courses offered by an educational institution; all planned school activities including besides courses of study organized play, athletics, dramatics, clubs, and home-room program.' Laymen might well assume that specialists making of curriculum a field of study would devote their attention primarily or exclusively to describing, analyzing, and evaluating the nature and effectiveness of such institutional programs.

A cursory review of the vast body of literature produced by these specialists reveals considerable attention to institutional programs in place but as much or more to a larger domain implied by definitions such as 'all the activity and experience that life unaided would not provide' (see Tanner and Tanner 1995, Chap. 5 for a comprehensive array of definitions). Specialists point out that there are distinctively different genres of curriculum: the courses of study mandated by nations, states, and school districts; the course offerings listed in the schedules and catalogs of educational institutions; the activities actually arranged and conducted by teachers; and what students experience. The scope of the field becomes virtually that of education itself: anthropological, economic, existential, historical, linguistic, philosophical, political, psychological, sociological (see Vandenburg 1990). The present paper seeks to sort out from this wide domain what appear to be three dominant sectors of curriculum inquiry, two of which acquired their legitimacy in the second half of the twentieth century.

1. *From Clear to Unclear Borders*

Curriculum praxis—that is, the development and conduct of schedules of study, work, and other human activity—appears to be an old endeavor. It has ranged from tribal rites of passage to programs for a nation's schools. Curriculum praxeology—the study of practice—is a relatively new endeavor. The book identified as probably the first in the West addressed to curriculum as a field of study was published in 1695 (Schubert 1980). Chapters in contemporary books that address 'early beginnings' rarely include references prior to the nineteenth century. Presumably there were, centuries ago, narratives of debate over what to teach the young, for example, that today would be considered part of the epistemology of curriculum as a field of study. Surely the twelfth-century Chinese philosopher Chu Hsi would be seen currently as a curriculum theorist for his neo-Confucian views on education as a moral imperative and the school curriculum as a vehicle for its implementation (see deBary and Chaffee 1989).

From the beginnings of formal schooling in Europe during the sixteenth and seventeenth centuries, the dominant curricular issue has been whether the needs and interests of children or the 'essentials' of subject matter should be of first priority. Few contemporary

treatises addressing the tension between so-called child-centered and subject-centered curricula omit the contribution to twentieth-century thought of seventeenth- and eighteenth-century reformers and thinkers such as Jean Jacques Rousseau (on children following their own interests as depicted in the volume *Émile*), Johann Heinrich Pestalozzi (on teaching directed to the natural powers of the individual), and Herbert Spencer (on the nature of knowledge, with particular attention to his essay, 'What knowledge is of most worth?').

It is reasonable to conclude that the roots of curriculum as a field of study, with all the usual accompaniments of inquiry, professional organizations, conferences, and publications, are to be found in the subject vs. child debate. At first embraced primarily in psychological and philosophical discourse, the issue broke out of the boundaries of these disciplines to constitute the very core of a field now virtually incapable of defining its own borders. The program of the 1999 conference of the American Educational Research Association (AERA, enjoying a large international membership) lists more than 60 sessions, each with four or five papers presented, for Division B entitled 'Curriculum Studies.' The following selection of session titles suggests the incredible scope of what those in the field consider to be legitimately within it: social, political, and cultural contexts of curriculum renewal in Europe; interpretive research paradigms: ethnography, community and transformation; unpacking in public: exploring social/racial identities in an urban university; gender, education, and citizenship: an international feminist dialogue; masculinities at school: gender relations and representatives of boys, men, and masculinity; pedagogy and remembrances of historical trauma; issues in mathematics and science curriculum; ideology and curriculum revisited: alienation, privilege, and dispossession in the college classroom; the future of the research university and the professoriate; education reform in historical perspective. There are some titles here that might have appeared decades ago, when members of the AERA were primarily white males. The newer themes appear to require long, double titles, separated by colons, to establish their legitimacy.

2. *A Focus on Practice*

Had there been a comparable gathering in 1900, most of the above titles would have been missing. The gathering would have focused primarily on the school subjects, with accompanying argument regarding priorities with respect to topics within them. There probably would have been presentations and discussions regarding the general education of all children in elementary schools and of the priority subjects for the few going on to secondary schools. The most controversial paper would have been on the issues of whether the subject oriented curriculum of the past had any relevance for the increasing numbers of children coming to school and staying there longer, and whether the mismatch was having a negative psychological impact on the young.

World War I stimulated attention to the world beyond national borders. To the old question of how much the curriculum should be child or subject oriented was added that of the extent to which it should be society oriented. This broadening of interest in sources of curriculum development focused much more on technical matters of construction than on bringing into the classroom social issues such as the then budding feminist movement (Rugg 1926).

The central stream of curriculum inquiry throughout the first half of the twentieth century was on what Joseph Schwab (1970) referred to as commonplaces: aims, purposes, and objectives; subjects and the balance of attention to them; scope and sequence of subjects and the curriculum as a whole; and measurement of student progress. The 'pure' study of curriculum was of these. But two domains of inquiry in their own right increasingly come into curriculum discourse.

One of these pertained to the impact of the public context on institutional curricula, attracting the scholarly attention of economists, historians, philosophers, and political scientists, many of whom eschewed identification with the curriculum field but influenced it, nonetheless. The other spilled over quite naturally out of psychological inquiry into learning and the scientific ordering of what was to be learned. Major contributors identified with their disciplines, but their names are embedded in the scientific tradition of the curriculum field: Pavlov, Thorndike, Watson, Piaget, Skinner, Bruner, and more.

The name most commonly identified as contributing significantly to all three streams of inquiry is John Dewey. Arguably, a wider impact on curriculum making was made by Edward L. Thorndike, whose focus on the hierarchical ordering and measuring of subject matter profoundly influenced the structure of school textbooks. During the concluding decades of the twentieth century, one would as likely find—in curriculum conferences and journals at least—as many papers on the social context and the scientific processes of curriculum making and measurement as on the description and analysis of curricular commonplaces in practice.

3. *Theory*

A peculiar aspect of the focus on the ends and means of curricular practice was the general absence of studies from which principles to guide and predict action and its consequences might be deduced—in other words, the absence of theory building. The conference held at the University of Chicago in 1950, designed to explore what the foundations of such theory building might be, appears to be the first of its

kind and one of few to date (see Herrick and Tyler 1950). It received little attention then and has been little cited since.

Arguments for curricular relevance, on the one hand, and for efficiency, on the other, stimulated advocacy rather than inquiry. Proposals for curriculum practice, even when argued from principles such as of child development, rarely were accompanied by studies of implementation and effects. Marketing rhetoric triumphed over empirical evidence. A kind of meta-theorizing about alternative advocacies came to characterize inquiry labeled curriculum theory from the 1950s to the 1970s.

The field of study that centered on practice during the first half of the twentieth century shifted in the second half to discourse about proposals for practice. It became increasingly divisive, noncumulative, and unclear in the criteria for claiming scholarly membership. Two consequences followed: the interest and attention of the constituency in curriculum practice declined; and streams of thought not previously identified with the traditional mainstream declared themselves curricular. What occurred was more a takeover than a joining. Much of what had been the curriculum field was pushed aside and declared by some to be moribund.

4. *Three Streams in Search of a River*

A better fate for an unprecedented study conducted in the 1930s might have altered somewhat the post-World War II development described above. The great depression that engulfed much of the world from 1929 into the war years accelerated decline in the value of children to the workplace, particularly in the most industrialized countries, increasing both the growth of secondary schools and their custodial function. The issue of the suitability of a narrow academic curriculum for all became critical. The seeds for innovation were planted, inviting both advocacy and the study of advocacy in institutional settings. But the state of scientific inquiry in curriculum was primitive.

The study generated during the 1930s in the USA, responding to a provocative question, was groundbreaking. Should not the ends and means of the secondary school reflect the democratic ideals of the nation? Arguably the first of its kind, the Eight-Year Study was born: a comparison during their college years of graduates from 30 schools left free to reconstruct in their own ways, and graduates of control schools not similarly engaged. Uniquely, the inquiry required documentation of ongoing changes in the experimental schools and the development of measures for assessing a wide range of student behaviors in both groups.

By the time five books reporting the work were published in 1942 and 1943, the country was deeply into war. The small number of volumes published, purchased primarily by libraries, virtually ensured that lessons learned were lodged in the minds of few other than the participants (see Thomas 1990, Aiken 1942). There were in the study the makings of a disciplined field of inquiry focused on the commonplaces of practice that, in turn, it is reasonable to assume, constitute the source of a legitimate mainstream of curriculum inquiry.

In his role in the Eight-Year Study, Ralph Tyler had brought into educational measurement the concept of evaluation: determining the values embedded in the educational ends sought, developing techniques for determining the presence of these values in the conditions established and in the outcomes obtained, and then constructing hypotheses and explanations regarding cause and effect. Tyler's conception forced attention not just to outcomes but to the whole of curricular ends and means. During the second half of the 1940s, at the University of Chicago, Tyler developed a graduate course on curriculum making entitled *Basic Principles of Curriculum and Instruction* (Tyler 1949). Never intended to be more than a syllabus free of scholarly references, never revised, and reprinted over and over into the present, this little monograph earned him a firm place in curriculum as a field of study.

Given Tyler's role in the research methodology of the Eight-Year Study, it is puzzling that he turned the research questions of the soon-forgotten study into a procedural tool for curriculum development: what educational purposes should the school seek to attain? What educational experiences can be provided that are likely to attain these purposes? How can these educational experiences be effectively organized?

How can we determine whether these purposes are being attained? A slight rewording changes each of these into a research question, as with the first: what educational purposes *is* the school seeking to attain? Had Tyler constructed the syllabus around his four questions phrased to guide research on practice, using the Study as illustrative, those who took his course, the many more who read his monograph, and curriculum as a field of curriculum inquiry, might have proceeded differently in the postwar years. Instead, the Tyler Rationale (as it came to be known) took on a life of its own in the meta-theorizing that accompanied advocacy. The study of practice languished.

From the 1960s to the 1970s, the dominant stream of curricular thought, part historical and part philosophical, focused on whose interests were being served by the schools, challenging the conventional wisdom that established practices nourished democratic principles. Meanwhile, evaluation (influenced by Tyler, sometimes referred to in the literature as the father of evaluation) was establishing itself as a stream of curriculum thought and practice in its own right. Some scholars engaged in critical inquiry into the values embedded in curriculum policy viewed those in evaluation as mere technicians; some evaluators per-

ceived the reconceptualists, as they came to be called, as inventing the truth divested of evidence. The two groups invigorated and frequently dominated dialogue at curriculum conferences, but there were those in the traditional domain of studying the commonplaces of practice who viewed both as house guests overstaying their welcome.

5. *The River Curriculum*

What Tyler accomplished in a short space was the identification and alignment of the essential components of curriculum construction—the commonplaces. For many critics, the synthesis was too tidy, too linear, or too little attentive to their specialties. Indeed, the Tyler Rationale lent itself to conversion into a sloppy algebraic formula: P (purposes) + L (learning experiences) + O (organization) + E (evaluation) = C (curriculum). Nonetheless, for decades many specialists found it necessary to take a stand with respect to Tyler's work in advancing their own positions. What Tyler contributed was a systematic arrangement of a language of curriculum—a lexicon.

This became clear at the International Seminar for Advanced Training in Curriculum Development and Innovation, supported by UNESCO and the International Association for the Evaluation of Educational Achievement (IAEEA) and directed by Benjamin Bloom, in Gränna, Sweden, during the summer of 1971 (Bloom 1971). Teams of four to six persons from 23 countries of South America, Africa, Asia, and Europe attended lectures by Tyler most mornings and read his monograph. There were fears in the subgroups, each comprised of one member from each of the teams, that cultural differences would deny conversation around common issues. They soon discovered, however, that the lexicon of ordered curriculum commonplaces of the Tyler Rationale provided common ground.

Schooling, especially at the elementary and secondary levels, is a highly political endeavor that connects only loosely and sporadically with scholarly inquiry. The support and use of studies of existing curricula are rare. The seminal studies of Torsten Husén and his colleagues decades ago that influenced practice in Sweden appear to have suffered the fate of the Eight-Year Study; there is no mention of them in the massive *Handbook of Research on Curriculum* (Jackson 1992). The several rather comprehensive inquiries into practice in the USA conducted in the 1980s were supported by private philanthropy and their recommendations obliterated by publicly supported nonresearch reports of state and federal commissions. Had Benjamin Bloom and his associates been successful in securing financial support for creating the International Curriculum Organization of national centers such as those represented at the Gränna conference, an organization conceived as a companion of the IAEEA, the international perspective and exchange thus gained might have served to offset the bureaucratic monitoring role most specialists and centers have been required to perform.

With the macro-issues of curriculum practice dominated by political machinations, today's students of curriculum are limited largely to fine-grained micromatters. Given the current state of scientific inquiry in curriculum, this is society's loss. A robust future for curriculum as a field of study devoted to educational renewal depends on it being recognized and supported for its potential benefit to humankind.

See also: Curriculum Evaluation; Curriculum Theory; Educational Institutions and Society; Educational Planning Worldwide; Educational Policy: Comparative Perspective; Educational Research and School Reform; Educational Research for Educational Practice; Hidden Curriculum in the Classroom

Bibliography

Aiken W M 1942 *The Story of the Eight-Year Study*. Harper & Row, New York

Apple M W 1979 *Ideology and Curriculum*. Routledge & Kegan Paul, London

Bloom B S 1971 *The Report of the International Seminar for Advanced Training in Curriculum Development and Innovation: Gränna, Sweden, July 5–August 14*. International Association for the Evaluation of Educational Achievement, Stockholm, Sweden

DeBary W T, Chaffee J W (eds.) 1989 *Neo-Confucian Education*. University of California Press, Berkeley, CA

Dewey J 1990 *The School and Society* [orig. 1900] and *The Child and the Curriculum* [orig. 1902]. University of Chicago Press, Chicago

Early M J, Rehage K J (eds.) 1999 Issues in curriculum: A selection of chapters from past NSSE yearbooks. *Ninety-Eighth Yearbook of the National Society for the Study of Education*, Part II. University of Chicago Press, Chicago

Giroux H A 1981 *Ideology, Culture & the Process of Schooling*. Temple University Press, Philadelphia, PA

Goodlad J I 1979 The scope of the curriculum field. In: Goodlad J (ed.) *Curriculum Inquiry: The Study of Curriculum Practice*. McGraw-Hill, New York

Herrick V E, Tyler R W (eds.) 1950 *Toward Improved Curriculum Theory*. University of Chicago Press, Chicago

Jackson P W (ed.) 1992 *Handbook of Research on Curriculum*. Macmillan, New York

Lewy A (ed.) 1985 Curriculum. In: Husén T, Postlethwaite T N (eds.) *International Encyclopedia of Education*. Pergamon, Oxford, UK

Rugg H (ed.) 1927 The foundations of curriculum-making. *Twenty-Sixth Yearbook of the National Society for the Study of Education*. Part II. Arno Press, New Yok

Schubert W H 1980 *Curriculum Books: The First Eighty Years: Content, Commentary and Bibliography*. University Press of America, Lanham, MD

Schwab J J 1970 *The Practical: A Language for Curriculum*. National Education Association, Washington, DC

Tanner D, Tanner L 1995 *Curriculum Development: Theory into Practice*, 3rd edn. Merrill, Englewood Cliffs, NJ

Thomas B 1990 The school as a moral learning community. In: Goodlad J I, Soder R, Sirotnik K A (eds.) *The Moral Dimensions of Teaching*. Jossey-Bass, San Francisco
Thorndike E L 1913 *Educational Psychology*. Vol. 2, *The Psychology of Learning*. Columbia University, New York
Tyler R W 1949 *Basic Principles of Curriculum and Instruction*. University of Chicago Press, Chicago
Vandenburg D 1990 *Education as a Human Right: A Theory of Curriculum and Pedagogy*. Teachers College, New York

J. I. Goodlad

Curriculum Evaluation

The term 'curriculum evaluation' has historically been used to refer to several different but inter-related concepts that have not been well delineated in the research literature. Some writers have used 'curriculum evaluation' to refer to curriculum product evaluation; others employ the term in reference to curriculum program evaluation. Curriculum products such as textbooks or natural curricula are evaluated prior to large-scale implementation. These evaluations are conducted using prespecified external criteria, outcome data from field trials, or both. Curriculum programs refer to the instruction that takes place within specific field contexts.

1. *Curriculum Product Evaluation*

Curriculum product evaluation centers on the evaluation of curriculum products by their developers or by those who make election or adoption decisions. Curriculum product evaluation is conducted in a variety of settings. Tamir (1985) categorized these as: (a) part school-based curriculum development; (b) national curriculum study organizations; (c) within the framework of national curriculum development centers; and (d) development, by university teams. Curriculum product evaluation is also conducted by local, state, or national textbook selection committees, which also consider other instructional materials.

1.1 *Evaluation Based on External Criteria*

One type of curriculum product evaluation employs specified external criteria intended to describe appropriateness. The work of Tyler and Klein (1976) offers an excellent early example of curriculum product evaluation employing prespecified criteria. Their evaluation procedure uses characteristics such as: specification of instructional objectives on which the material is based; appropriateness of the materials given the skills, background knowledge, age, ethnicity, and socioeconomic background of the intended students; and adequacy of the teachers' manual for classroom application and for providing an explanation of the content selection, sequence, and presentation. Later, Gall (1981) presented a listing and elaboration of 39 criteria for analyzing curriculum materials.

These evaluation procedures, particularly Tyler and Klein's, reflect a behaviorist orientation to evaluating curriculum products. Brophy and Alleman's work (1991), derived more heavily from research on teaching, provided a constructivist framework for the evaluation of instructional activities. The principles they proposed for evaluation include goal relevance, appropriate level of difficulty, feasibility, cost effectiveness, multiple goals, motivational value, topic currency, whole-task completion, higher order thinking, and adaptability. This constructivist view led to the formation of new curriculum evaluation standards in California and many other states.

Standards typically originate when a policy-making body invites colleges, universities, professional organizations, school districts, and other agencies to nominate distinguished professionals with expertise in a given field to serve on a panel. The standards do not require all products or programs to be alike except in their quality, which assumes different forms in different environments. Judgments as to whether standards are fulfilled are made by professionals who have been trained in their interpretations.

The use of standards for product evaluation is particularly relevant for the evaluation of textbooks. Procedures for textbook evaluation have been most thoroughly systematized in the larger states that practice state-based textbook adoption. California and other large adoption states in the United States are fine tuning selection procedures and standards for each subject area in order to obtain a more tailored product from textbook publishers. In these procedures, criteria are provided for evaluating textbooks to be considered for adoption. The criteria are well specified, further elaborated, and weighted in a separate rating document. A minimum total score cut-off creates the possibility that none of the textbooks will be deemed acceptable at the first review.

1.2 *Evaluation Based on Outcome Data from Field Trials*

A second type of curriculum product evaluation relies on field data for judging the adequacy of curriculum products, specifically the product's actual impact on students. In this sense curriculum evaluation is an examination, or validation, of the impact of a newly developed product.

Curriculum product evaluation based on field data may be performed both formatively (to shape or modify the curriculum under development) or sum-

matively (to validate the curriculum prior to its release). The type of products included are usually more limited than for curriculum product evaluation using external criteria. However, the two procedures are sometimes used concurrently (e.g., when a state evaluation proposed textbook is asked to provide descriptions of the product development process, including field testing data).

2. Curriculum Program Evaluation

The term 'curriculum program evaluation' refers to a complex set of interactions between a given curriculum/instructional program and its setting. Curriculum evaluation activities in this broader context examine particular curricula, instructional programs, courses of study, or national programs within their situational context. Theorists whose work is relevant to curriculum evaluation are mainly those who are known for their work in evaluation of educational or social programs.

Alkin and Ellett (1985) have noted that models of evaluation are 'prescriptive' rather than 'descriptive.' A descriptive model is a 'theory' consisting of empirically derived statements containing generalizations that might describe, predict, or explain evaluation activities. The models in evaluation instead prescribe which evaluation activities are good or bad, right or wrong, adequate or inadequate, rational or irrational, just or unjust. These models claim that evaluation should be conducted in prescribed ways, indicate evaluator's obligations and responsibilities and provide advice, recommendations, and pitfalls in conducting an evaluation.

Three features define the unique characteristics of different evaluation models: (a) the collection and analysis of data (methodology); (b) the ways in which data are valued or judgments made (values); and (c) the purposes for providing evaluation information (uses). All evaluations involve some kind of methodology for description and explanation, all necessitate the valuing of data acquired within the particular context, and all evaluations are conducted with some intended use in mind. Thus, the distinction between evaluation models based on these three dimensions is not exclusionary; for example, it is not claimed that one model only represents the necessity of employing a methodology and others do not. Rather the classification system for evaluation models presented here is based on the relative emphasis within the various models. In essence, the issue is: when evaluators must make concessions, what do they most easily give up and what do they most tenaciously defend?

The methodology-oriented models may be viewed in three subsets: the measurement outcome-oriented models, the research and/or quantitative methodology-oriented models, and models that focus on naturalistic methods. The values-oriented models presented here are expertise-oriented valuing, process-oriented valuing, and valuing based on principles of social justice. 'Uses' will be examined in two categories directed toward decision-makers and primary users.

2.1 Measurement Outcome-oriented Evaluation

The Tylerian evaluation model has been most closely associated with the term 'curriculum evaluation' (see Tyler 1942). This model is one of the earliest oriented toward measurement outcomes as a governing basis for the conduct of evaluation; its ultimate purpose is to determine whether a given program (or curriculum) achieved its intended purposes. For those using this model, or more recent variations of it, the important questions ask how the objectives are to be determined, behaviorally defined, and measured. Other important considerations are short- and long-term objectives, and the use of traditional and nontraditional forms of measurement. The important evaluative focus is on discrepancies between measured outcomes and intended objectives.

2.2 Research and/or Quantitative Methodology-oriented Evaluation

Other methodology-oriented evaluators perceive evaluation as nearly synonymous with experimental research methods. Thus, an emphasis on the experiment or quasi-experiment as the appropriate methodology for social science research/evaluation was lauded by Campbell and Stanley (1963) and embraced by a generation of evaluation researchers. The evaluation text by Rossi et al. (1999) most closely reflects this viewpoint.

Postpositivism represents an extension of this evaluation model category. Greene and McClintock (1991 p. 14) noted that postpositivists retain 'a preference for the quantitative methodology of experimental science. This preference includes a continued emphasis on causal explanation.' However, they are more willing to allow the use of multiple instruments for a given phenomenon, multiple analysis procedures, and critiques from alternative value frameworks.

2.3 Naturalistic Evaluation Methods

A number of theorists have developed their views partly in reaction to what they viewed as a dominance of measurement and other quantitatively oriented evaluation procedures. These evaluation writers have acquired 'naturalistic' evaluation points of view, which emphasize the richness that qualitative and descriptive data can bring to the evaluation process. Naturalistic

evaluation methods typically rely on data from several sources, respect the variety of perspectives, allow understanding to emerge from the process, and portray multiple realities.

Most influential of the group of theorists associated with this category is (see McLaughlin 1991). His early writing provided a framework for viewing the evaluation process more broadly. His later emphasis on the case study as an appropriate method for conducting evaluations served to expand professional perspectives at that time.

2.4 Expertise-oriented Valuing

Values-oriented evaluation theorists have 'values' as their primary concern but differ on their positions concerning who should make the value judgments about programs, and how such judgments should be reached. Some values-oriented theorists maintain that the evaluator should be responsible for making value judgments based primarily on his or her own personal background, knowledge, and experience. This point of view we label 'expertise -oriented valuing.'

Educational connoisseurship and criticism are the cornerstones of Eisner's (1993) approach. He maintained that, like the art connoisseur, the educational connoisseur is capable of deep appreciation, combining an understanding of formal qualities, socio-historical-cultural contexts, and influences. Not unlike wine connoisseurs, educational connoisseurs have attained refined perceptual capabilities acquired in part through previous relevant experience. While connoisseurship is the art of appreciation, criticism concerns itself with the art of disclosure.

Another view of expertise-oriented valuing is presented in Scriven's writings. The expertise of Scriven's evaluator is founded on several bases: (a) the objectivity for unbiased valuing provided by the use of the goal-free technique, the evaluator being 'blind' to the program's goals (Scriven 1973); (b) the broad education of the evaluator (e.g., formal education, human development, academic disciplines) providing the basis for recognizing and defining objective values; and (c) the empirical, statistical, and logical skills, and a sense of perspective, provided by training as a philosopher of science. To these, he adds 'a little toughening of the moral fiber' as the final justification for his evaluator as the person having the expertise to value data.

2.5 Process-oriented Valuing

Other approaches that focus on the valuing of evaluation data do not rely primarily on the expertise of the evaluator but instead depend on a well-defined process for performing the valuation task. One such process-oriented approach involves the employment of standards. Standards may be used for conducting evaluation of programs as well as for evaluation products. In this process, review teams visit and examine a program in terms of a set of process criteria consisting of attributes of programs presumed to be indicators of quality. That is, there is a presumption that these processes will lead to quality outcomes. Most typical of such procedures are the various types of program accreditation involving self-appraisal and site visit teams guided by defined process standards.

The constructivist-hermeneutic model—what Guba and Lincoln (1989) called 'fourth generation evaluation'—emphasizes another approach that focuses on negotiation as the key dynamic of evaluation. Here, the construction of meaning and the valuing of that meaning are the evaluation's principal foci, and a central purpose of the evaluation is to create outcomes that appropriately reflect all participants' values. A principal assumption underlying this approach is that 'realities are not "out there" simply to be discovered,' but are created through human interactions and that there are many possible coexisting realities created out of those interactions.

According to Guba and Lincoln, the evaluation process is necessarily value-laden, as all participants bring with them not only expectations as to what the purpose of the evaluation might be, but personal values that are the products of each individual's social, psychological, and physical context. The evaluation is designed to allow for the participation of competing value-systems, and is organized to produce, through the processes of negotiation, a consensus that takes into account all perspectives.

2.6 Valuing Based on Principles of Social Justice

Sometimes valuing occurs through the application of commonly accepted external principles. Ernest House the theorist has been guided by a concern for democratizing public decision-making (House 1999). In addressing this purpose, he has recognized that evaluation, as a political tool, becomes a means for determining 'who gets what' (distributive justice) in a societal program. He noted that many people's well-being depends on evaluations that provide a 'fair/just' distribution of benefits and responsibilities. Thus, the most important role of the evaluator is applying explicit theories of justice in valuing data. A key aspect is the development of a 'fair evaluation agreement' in which all participants are guaranteed to have their 'real interests' identified and addressed. In rendering judgements on programs, the social justice evaluator must be guided by the following principles: moral equality, moral autonomy, impartiality, reciprocity, and utility.

2.7 Decision Maker-oriented Evaluation

Unlike the previous evaluation models which have a more general audience (e.g., the research community, stakeholders generally, consumers, government),

models oriented toward decision-makers focus on specific individuals and thus emphasize use. The first of the decision-maker models, the CIPP model (developed in the early 1960s and presented in revised form in Stufflebeam 1983) focused on evaluation as a systematic procedure designed to provide information that would 'enlighten' decision-making. Thus, the focus of these evaluation efforts were decision-makers—usually interpreted to mean program administrators, and most typically those who contracted for the evaluation. Stufflebeam described the attributes of four different evaluation types (Context, Input, Process, and Product) each designed to 'serve,' that is, provide information for a different kind of management decision.

2.8 User-oriented Evaluation

Patton (1997) extended the concern for information provision beyond specified decision-makers and presented a model in which evaluators are obligated to identify likely 'primary users.' Thus, the evaluator must seek out those specific stakeholders for whom the evaluation is most likely to be of value in actually effecting change. Patton noted that this means finding strategically located people who are committed and competent and who are able and willing to use information. Patton also recognized a dynamic relationship existing between evaluators and users, a relationship in which prespecified decision concerns have been replaced by interactions, thus illuminating the more incremental nature of decision-making as it applies to educational programs. Alkin's writing also reflects a user-oriented approach to evaluation (see McLaughlin 1991).

3. Other Models: 'Collaborative'

Recently, various kinds of 'collaborative' evaluations have received substantial prominence. These include illuminatory evaluation (Partlett and Hamilton 1976), participatory evaluation (Cousins and Earl 1992), empowerment evaluation (Fetterman et al. 1995), and emancipatory and critical action research (Carr and Kemmis 1988). Some have programmatic interests, philosophic interests, or political interests. These collaborative evaluation approaches are not categorized separately. Indeed they would be placed into existing categories based on the intent and purpose of each evaluation type.

For example, participatory evaluation is generally concerned with participation as a means to increase the utilization of evaluation and would be included in category 2.8 above. Likewise, emancipatory and critical action research is more directly concerned with achieving 'emancipation' from the dictates of injustice, alienation, and unfulfillment, clearly falling into category 2.6 above, as would empowerment evaluation.

4. Current Issues in Curriculum Evaluation

We will briefly comment on several current issues in curriculum evaluation. These relate to the nature of appropriate measures for design considerations and the use of discourse analysis.

4.1 Measurement

One current issue relates to the reliance on standardized tests as the instrument for assessing the impact of curriculum. Although information from such tests is relatively inexpensive to collect, easy to aggregate, and somewhat reliable, standardized tests seldom match national and state goals. These tests are an inadequate measure of what is learned (they fail to sample content adequately) and are limited in assessing the depth of student understanding (no allowance for multiple forms of representation of student's knowledge). Black (1998) has drawn attention to the difficulty of finding tests that meet requirements of validity (determining that the inferences based on the test results are justified) and reliability (showing that the results are reproducible on different occasions and with different questions).

Hence, curriculum evaluation must be based on more than standardized test scores and includes assessing long-term consequences, a wide range of source reports/interviews, classroom observation, work samples, inventories, portfolios, records, exhibits, and the like. Information from such sources is also necessary for determining how the curriculum is implemented, the processes of student learning, and identifying factors that facilitate or restrain implementation.

A related issue concerns the interactions that take place between curriculum evaluation at different levels. Political leaders and central administrators favor standardized testing for the quantifiable data that will guide education policy and control and monitor schools and classrooms. Professional educators and those engaged in classroom instruction and the guidance of individual students favor consumer-oriented evaluation that integrates assessment with learning, such as continuous assessment of student's regular work and self-assessment, using the results formatively as feedback and diagnostically for guidance to reinforce learning and to encourage learners to take responsibility for their own learning as well as to suggest more appropriate curriculum. The challenge

for the next decade is to overcome the polarization between those who put standards and testing to the forefront in the interest of accountability (control) and those who see evaluation as a tool for the promotion of learning.

4.2 Critical Discourse Analysis

Finally, critical discourse analysis is emerging as a powerful tool for showing how curriculum works, how it operates politically, how it constructs and regulates social relations, and how it positions students, teachers, and others. Allan Luke (1995) has used critical discourse analysis to show how the curriculum of the home and preschool orients children toward family, property, social practices, and ownership; how the language of the primary school teacher constructs identities of children and guides them for participation in different content areas; and how textbooks influence students to reproduce, naturalize, and accept particular forms of cultural logic and social identity under the guise of transmitting neutral skills. Similarly, Shirley Grundy (1994) has deepened understanding of the contrasting public curriculum policies in Australia and Canada using this analytic tool.

Essentially, evaluators using discourse analysis study and interpret the language and language patterns as found in both written and oral contexts. The analysis reveals how the curriculum: (a) relates to the students' life trajectories—the kinds of futures portended (field); (b) instills a view of authority and degree of individual freedom, teaching students where authority lies, when and how one can speak, what can be said, and constructing social relations (tenor); and (c) defines a given subject matter—its purposes and functions, modeling what it means to learn the subject, what counts as knowledge in the field, and defining what the students' relations to the subject should be (mode).

See also: Competencies and Key Competencies: Educational Perspective; Curriculum as a Field of Educational Study; Curriculum Theory; Educational Evaluation: Overview; Educational Assessments: Major Developments; Program Evaluation

Bibliography

Alkin M C, Ellett F S 1985 Evaluation models: development. In: Husen T, Postlethwaite N (eds.) *The International Encyclopedia of Education*, lst edn. Pergamon, Oxford, UK

Black P J 1998 *Testing: Friend or Foe? The Theory and Practice of Assessment Testing*. Falmer Press, London

Brophy J, Alleman J 1991 Activities as instructional tools: A framework for analysis and evaluation. *Educational Researcher* **20(4)**: 9–23

Campbell D T, Stanley J C 1963 Experimental and quasi-experimental designs for research on teaching. In: Gage N L (ed.) *Handbook of Research on Teaching*. Rand McNally, Chicago

Carr W, Kemmis S 1988 *Becoming Critical: Education, Knowledge and Action Research*. Falmer, London

Cousins J B, Earl L M 1992 The case for participatory evaluation. *Educational Evaluation and Policy Analysis* **14(4)**: 397–418

Eisner E W 1993 *The Educational Imagination: On the Design and Evaluation of School Programs*, 3rd edn. Macmillan, New York

Fetterman D A, Kaftarian S J, Wandersman A (eds.) 1995 *Empowerment Evaluation*. Sage, Beverly Hills, CA

Gall M D 1981 *Handbook for Evaluating and Selecting Curriculum Materials*. Allyn and Bacon, Boston

Greene J, McClintock C 1991 The evolution of evaluation methology. *Theory Into Practice* **30(1)**: 13–21

Grundy S 1994 Which way toward the year 2000? Contrasting policy discourses in two education systems. *Curiculum Inquiry* **24(3)**: 327–47

Guba E G, Lincoln Y S 1989 *Fourth Generation Evaluation*. Sage, Newbury Park, CA

House E R 1999 *Values in Evaluation and Social Research*. Sage Publications, Thousand Oaks, CA

Luke A 1995 Text and discourse in education: an introduction to critical discourse analysis. *Review of Research in Education* **21**: 3–48

McLaughlin M W (ed.) 1991 *Evaluation and Education: At Quarter Century*. The National Society for the study of Education, Chicago

Partlett M, Hamilton D 1976 Evaluation as illumination: A new approach to the study of innovatory programs. In: Glass G (ed.) *Evaluation Studies Review Annual*. Sage, Beverly Hills, CA, Vol. 1

Patton M Q 1997 Utilization-focused evaluation. *The New Century Text*, 3rd edn. Sage, Thousand Oaks, CA

Rossi P H, Freeman H E, Lipsey M W 1999 *Evaluation: A Systematic Approach*, 6th edn. Sage, Thousand Oaks, CA

Scriven M 1973 Goal free evaluation. In: House E (ed.) *School Evaluation: The Politics and Process*. McCutchan, Berkeley, CA

Stufflebeam D L 1983 The CIPP model for program evaluation, In: Madaus G F, Scriven M, Stufflebam (eds.) *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*. Kluwer-Nijhoff, Boston

Tamir P 1985 The potential and actual roles of evaluator in curriculum development. In: Tamir P (ed.) *The Role of Evaluators in Curriculum Development*. Croom Helm, London

Tyler L L, Klein, M F 1976 *Evaluating and Choosing Curriculum and Instructional Materials*. Educational Resource Associates, Los Angeles, CA

Tyler R W 1942 General statement on evaluation. *Journal of Educational Research* **35(7)**: 492–501

M. C. Alkin and J. D. McNeil

Curriculum Theory

The curriculum is defined as programs of teaching and learning which take place in formal settings. Examples of formal settings are schools, colleges and universities. A curriculum may refer to a system, as in a national

curriculum, an institution, as in the school curriculum, or even to an individual school, as in the school geography curriculum. The four dimensions of curriculum theory are aims or objectives, content or subject matter, methods or procedures, and evaluation or assessment. The first dimension refers to the reasons for including specific items in the curriculum and excluding others. The second dimension is content or subject matter and this refers to the knowledge, skills, or dispositions which are implicit in the choice of items, and the way in which they are arranged. Objectives may be understood as broad general justifications for including particular items and particular pedagogical processes in the curriculum, or as clearly defined and closely delineated outcomes or behaviors, or as a set of appropriate procedures or experiences. The third dimension is methods or procedures and this refers to pedagogy and is determined by choices made about the first two dimensions. The fourth dimension is assessment or evaluation and this refers to the means for determining whether the curriculum has been successfully implemented.

1. Aims or Objectives

Any curriculum is always a selection. Sets of justifications for curricular inclusion may be divided broadly into four types: logical delineations between domains of knowledge, distinctive mental or cognitive operations, cross-cultural social distinctions, and deliberative activity about the ideal society.

An example of logical delineations is Hirst's (1974) forms of knowledge and experience: logico-mathematical, empirical, interpersonal, moral, esthetic, religious, and philosophical. Each of these forms has distinctive kinds of concepts and distinctive ways of determining truth from falsehood. Hirst claimed, therefore, that each has a separate logical form. An example of the second type of justification is Gardner's (1983) seven forms of intelligence: language or linguistic intelligence, logical–mathematical analysis, spatial representation, musical analysis, bodily kinesthetic thinking, interpersonal knowledge, and intrapersonal knowledge. His justification for inclusion of these forms of intelligence is psychological: individual learners have cognitive or mental modules which are separate and act separately from other mental modules. Individuals have been shown to differ in their capacity to perform these different types of operations. A third set of justifications moves us out of the mind and focuses on the culture we inhabit. Lawton (1989) argues that all societies have cultural subsystems: sociopolitical, economic, communication, rationality, technology, morality, belief, esthetic, and maturation. Because these are universal and cross-cultural, Lawton concludes that curriculum developers should seek to represent the forms of knowledge which underpin them. A fourth set of justifications focuses on the arrangements made by particular societies, both currently and in the future. The normative needs of society are given precedence over logical, cultural, or psychological justifications. This is best exemplified in relation to the debate about the respective merits of academic and vocational curricula.

2. Content or Subject Matter

2.1 Relations Between Items

A further concern of curriculum theorists is the relations between these different items of knowledge. The two most important types of relations are the degree of integration between different knowledge domains and progression within the domain itself. The first of these is the degree of integration. A curriculum may be understood as strongly or weakly classified and as strongly or weakly framed. A strongly classified curriculum is defined by Bernstein (1990) as clearly delineated domains of knowledge with strong boundaries between them; conversely, a weakly classified curriculum is understood as having weak boundaries between the different knowledge domains. A strongly framed curriculum, on the other hand, is defined as a program of study in which teacher and student have limited control over the selection of items and the way it is organized in respect of the pedagogical relationship. A weakly framed curriculum is characterized by greater control by teacher and student over the selection of content, the way it is organized and its pacing. Bernstein, in relation to the first of these principles, classification, identifies two types of curricula: 'collection codes' where strong boundaries between domains are present, and 'integrated codes' where weak boundaries are in evidence. Fogarty (1991) has identified 10 models of curriculum integration and these range from strongly classified and strongly framed curricula, as in the traditional approach, to weakly classified and weakly framed networked approaches to curriculum planning. Between the two extremes, traditional and networked approaches, she identifies eight other points on the continuum: connected, nested, sequenced, shared, webbed, threaded, integrated, and immersed.

The second dimension is progression within the domain, or, with integrated curricula, domains. Underpinning the notion of progression is a rationale for teaching some aspects of the knowledge domain before others and a belief that a subject can in fact be arranged in a reliable hierarchy. Adey (1997) argues that it is possible to do this and develops a three-dimensional model comprising conceptual complexity, breadth, and extent. Using only the last of these two dimensions leads to a naive view of learning. For Adey, a measure of conceptual complexity is also

needed to provide a fully developed model of curriculum progression. Examples of these frameworks are Piaget's (1971) schema comprising progression from concrete operational to formal operational thinking and Kohlberg's (1976) stages of moral thought, where the subject progresses from premoral and conventional rule conformity levels to the acceptance of general rights and standards, and even to adopting individual principles of conduct. These hierarchies are based on empirical investigation. The other way of establishing knowledge hierarchies is through some form of logical ordering, where complexity comprises both a progressive development of more items of knowledge and the making of more complicated connections between these items of knowledge.

2.2 Modeling Objectives

However, curriculum theorists have also focused on the way in which those aims and objectives are expressed. Three models have been developed: curriculum as product, curriculum as process, and curriculum as content. The first model specifies the development of behavioral objectives, and this approach has been discussed in a number of classical commentaries on curriculum (Tyler 1949). Curriculum-making is understood as a linear process which starts with the development of clear objectives or goals, proceeds through to the selection of content which is specified in behavioral terms—that is, its acquisition must be an observable or testable process—and finishes with the evaluation of that process to see if those objectives have been met. This behavioral objectives model has been criticized (Elliott 1998) for the following reasons: first, complex and important learning outcomes of any educational program may be neglected at the expense of the more trivial and less important, because it is easier to describe them in behavioral objective terms. Second, the prespecification of behavioral goals may encourage an inflexibility of approach within the classroom, and learning outcomes which may incidentally flow from classroom interaction will be deliberately underexploited. Third, there is a danger of assuming that if something cannot be measured, then it cannot be assessed and therefore it should not be a part of the learning process. Fourth, lists of intended behaviors do not adequately represent the way individuals learn: this is because logical order cannot be conflated with pedagogic process.

In opposition to behavioral objectives, it has been suggested that curriculum theorists should designate appropriate processes which learners need to go through—the second model. This avoids the problems inherent in the designation of prespecified behavioral objectives noted above and builds into the curriculum itself a more active and influential role for the teacher.

The third model, curriculum as content, comprises the transmission of a body of knowledge from the teacher to the learner and is underpinned by a symbol-processing view of cognition (see below) and a positivist/empiricist view of knowledge. The principles which underpin this view of knowledge are a belief in realism, a belief that there is a correct way of describing reality, a belief that scientific theories are superior to commonsense understandings of the world, a belief in the cumulative and progressive nature of scientific knowledge, a belief that it is possible to make a distinction between observation and theory, and that as a result it is possible to collect facts about the world, a belief that language is a transparent medium, that is, words have fixed meanings and concepts can be defined unambiguously, and finally, a belief that the methods which are appropriate to the natural sciences are equally appropriate to the social sciences. These assertions have been challenged in the twentieth century by interpretivists, critical theorists, and postmodernists. New knowledge constructions have implications for, and are reflected in, new ways of understanding learning and therefore new forms of pedagogy.

3. Methods or Procedures

Pedagogy is defined as the mode of delivery of the curriculum. A number of approaches have been developed. The first of these is imitation. The learner seeks to copy the actions of the teacher and in the process incorporate into their behaviors the observed characteristics of the teacher. Much learning of an informal nature, especially in childhood, takes this form. Didacticism, on the other hand, comprises instruction where the teacher inculcates the learner into either a knowledge, skill or affective domain by informing them about what they should do and how they should behave. This view of pedagogy has been disputed by constructivists and situationists who understand learning as contextualized, and thus as incapable of being understood without reference to the environment in which it takes place. In its place they would want to substitute a notion of intersubjective interchange in which the learner constructs knowledge in the light of the experiences they have in the classroom, and in the process creates meanings for themselves and others. The fourth pedagogic approach is apprenticeship, in which learners are supported in their attempts to gain access to the culture of the society in which they are being educated.

The two most important learning theories, symbol-processing and situated approaches, allocate distinctive roles to learning styles, assessment, and meta-cognition. Symbol-processing approaches understand the learner and the environment as separate; learning takes place within the human mind as the individuals process information they receive through

their senses, assimilate that information and create new ways of understanding. This theory understands the individual as a passive recipient of environmental influences. It separates out mind from body, language from reality, and the individual from society. Situated cognition or environmentally embedded learning understands the relationship between the individual and the environment in a different way. Situated learning approaches view the person and the environment as mutually constructed and as mutually constructing. Bredo (1999) suggests that this relationship should be viewed actively and as involving dynamic modification rather than static matching. The learners act with and on the environment, shaping or modifying themselves and at the same time shaping or modifying the environment. Situated cognitive approaches stress active, transformative, and relational dimensions to learning; indeed, situated cognitionists understand learning as contextualized. This new approach to learning has implications for the last of the four dimensions of curriculum, evaluation, or assessment.

4. *Assessment or Evaluation*

Assessment usually refers to the judgments made by individuals about their learning or by others about that learning. Evaluation refers to judgments about systems, institutions, or curricula, and the same tensions which can be found in formal assessments are present here. Notions of reliability and validity which have informed most of the traditional theories about assessment are now being reformulated in response to a number of criticisms (Gipps 1994). The first criticism is that formal assessments focus on the performance of the individual at a set point in time and in controlled conditions, and not to the levels of competence reached by the learner. The second criticism focuses on the social dimension of formal assessment. Formal assessments are designed to produce accurate representations of what learners can do and what they know. Situated cognitivists, on the other hand, argue that the information collected about an individual is always context- and culture-bound. Items in a formal test are chosen by examiners and this process of choosing reflects their own understanding of the world and the set of constructs they possess to order it. This set of constructs may not be shared by the individual taking the test or being assessed. The third problem with formal assessments is that they do not, and cannot, act formatively.

Learning, whether formal or informal, is closely related to assessment. Learners want and need to know what they have learnt in order to make a judgment about what and how they need to learn next. Learning is therefore always self-reflective and indeed self-reflexive, in that learning transforms the individual's sense of how they understand themselves and the world. This is why the word 'formative' is used to describe assessments which are useful to the formation of the individual through learning. The relationship between learning and assessment of that learning is complicated and depends in part on the way summative and comparative assessments—that is, imposed descriptions by other people on the individual concerned—impact on the learning situation. If these are given priority, then the process of formative assessment, indeed the process of learning itself, is impoverished.

See also: Curriculum as a Field of Educational Study; Curriculum Evaluation; Education: Phenomena, Concepts, and Theories; Educational Evaluation: Overview; Educational Innovation, Management of; Educational Institutions and Society; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values; Socialization and Education: Theoretical Perspectives

Bibliography

Adey P 1997 Dimensions of progression in a curriculum. The *Curriculum Journal* **8**: 367–92

Bernstein B 1990 *The Structuring of Pedagogic Discourse, Volume IV: Class, Codes and Control*. Routledge, London

Bredo E 1999 Reconstructing educational psychology. In: Murphy P (ed.) *Learners, Learning and Assessment*. Sage, London, pp. 23–45

Elliott J 1998 *The Curriculum Experiment: Meeting the Challenge of Social Change*. Open University Press, Milton Keynes, UK

Fogarty R 1991 *The Mindful School: How to Integrate the Curriculum*. Skylight Publishing, Pallantine, IL

Gardner H 1983 *Frames of Mind*. Basic Books, New York

Gipps C 1994 *Beyond Testing: Towards a Theory of Educational Assessment*. Falmer Press, London

Hirst P 1974 *Knowledge and the Curriculum*. Routledge and Kegan Paul, London

Kohlberg L 1976 Moral stages and moralization: The cognitive-developmental approach. In: Lickona T (ed.) *Moral Development and Behaviour*. Holt, Rinehart and Winston, London, pp. 36–54

Lawton D 1989 *Education, Culture and the National Curriculum*. Hodder and Stoughton, London

Piaget J 1971 *The Science of Education and the Psychology of the Child*. Routledge and Kegan Paul, London

Tyler R W 1949 *Basic Principles of Curriculum and Instruction*. University of Chicago Press, Chicago

D. Scott

Customer Satisfaction

Simply stated, customer satisfaction is a customer's evaluation of their purchase and consumption experience with a product, service, brand, or company. Policy researchers have long debated the value of

using satisfaction as a subjective measure of just how well a market system is able to meet customer needs (Ölander 1977). More recent interest in satisfaction stems from its role in affecting customers' repeat purchase decisions and subsequent company profits (Johnson and Gustafsson 2000). As the link to profitability has become clear, customer satisfaction is now a prominent business performance metric. As a result, satisfaction research has become an important component of an organization's stakeholder management, social accounting, and reporting. This article provides a broad overview of customer satisfaction, its meanings, the streams of research that have emerged in recent years, and key methodological issues that surround it. The article is not meant to be a review of what has become a vast literature (see Yi 1991). The first part of the article describes two distinct yet complementary streams of customer satisfaction research. After a brief overview of methodological issues surrounding customer satisfaction modeling, future research directions are discussed.

1. *Two Views of Satisfaction*

Customer satisfaction research has developed around two different types of evaluations: transaction-specific satisfaction and cumulative satisfaction. Transaction-specific satisfaction is a customer's evaluation of their experience with and reactions to a particular product transaction, episode, or service encounter. Early transaction-specific satisfaction research explored the cognitive-psychological antecedents of satisfaction, while more recent research has focused on the effects of positive and negative emotions on satisfaction (Oliver 1997).

Since the early 1990s, satisfaction research has grown to include an emphasis on cumulative satisfaction, defined as a customer's overall experience with a product or service provider to date. This definition is more consistent with those in economic psychology and welfare economics, where customer satisfaction is synonymous with the concept of consumption utility. It is cumulative satisfaction upon which the prominent national satisfaction, indices, such as the American Customer Satisfaction Index (ACSI; Fornell et al. 1996), are built.

These two streams of satisfaction research are complementary rather than competing. Transaction-specific research focuses on satisfaction with a product or service on a given occasion or transaction, or over a limited period of time. In cumulative satisfaction studies, the period of evaluation is left open. In the ACSI survey, for example, customers consider all of their experiences to date when evaluating their satisfaction with a product or service. The distinction is important as it affects how one views and models the evaluations. Whereas transaction-specific models provide a rich understanding of the dynamics of product and service encounters or episodes, cumulative satisfaction provides a stable basis for determining the drivers of satisfaction and its consequences (customer retention and subsequent economic performance).

1.1 *Transaction-specific Customer Satisfaction*

Oliver's extensive contributions to satisfaction research over the past 20-plus years illustrate how transaction-specific research has evolved over time (see Oliver 1997). Oliver posited that customers form preconsumption expectations, observe product performance, compare this performance with expectations, and then combine this information with expectation levels to judge satisfaction. When performance exceeds expectations (positive disconfirmation), satisfaction increases. When performance falls below expectations (negative disconfirmation), satisfaction decreases. The disconfirmation of expectations model has dominated the literature for years; a simple version of it is presented in Fig. 1. The direct effect of expectations on satisfaction is the result of a framing or assimilation effect; the expectations establish a baseline or anchor that directly influences the resulting satisfaction judgments. Thus attribute performance, expectations of performance, and disconfirmation of expectations may all have direct effects on evaluations of satisfaction.

The disconfirmation model has been augmented over the years to include the effects of equity, attribution, and emotions on satisfaction evaluations. The concept of reciprocity, and more specifically customer perceptions of equity and fairness, lies at the heart of marketing as an exchange process. Thus customers should be more satisfied when they are treated more fairly or equitably in a transaction, especially when there is a prospect for future transactions with the same party. Oliver and Swan (1989) posited and showed that equity is an important driver of satisfaction. Attribution theory has also been used to explain transaction-specific satisfaction from a cognitive perspective. The attribution of favorable outcomes to oneself and unfavorable outcomes to others systematically affects perceptions of performance and satisfaction (Folkes 1990).

However, Westbrook, Oliver, and others have recognized that these more cognitive perspectives on transaction-specific satisfaction do not completely capture the more affective basis of satisfaction (Oliver 1997, Westbrook and Oliver 1991). Customers form both positive and negative affective states that systematically influence satisfaction. For example, on the positive side there is joy and interest, and on the negative side is anger, disgust, and contempt, which affect satisfaction as expected.

1.2 *Cumulative Customer Satisfaction*

Cumulative satisfaction gained momentum with the advent of national customer satisfaction indices.

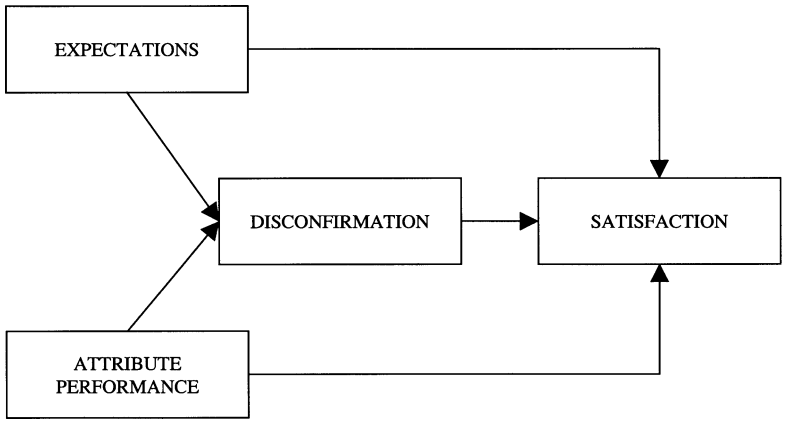

Figure 1
The disconfirmation of expectations model

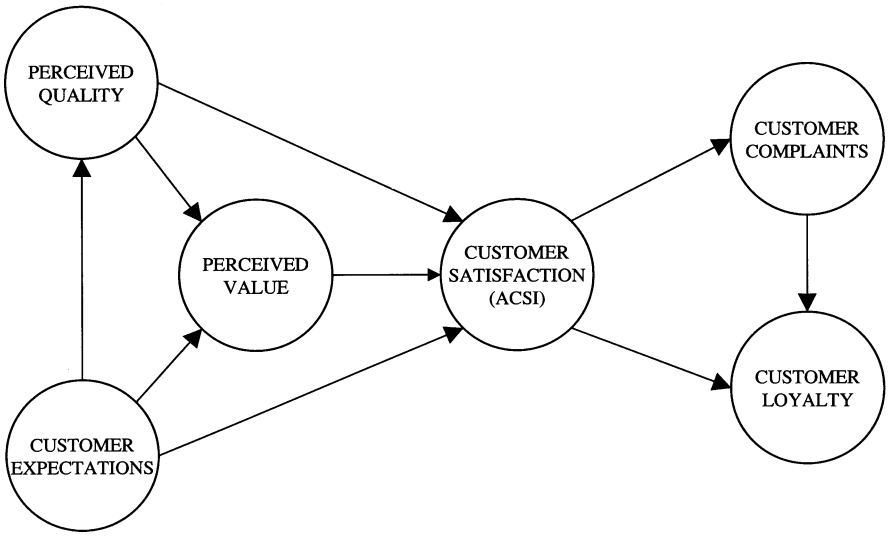

Figure 2
The American Customer Satisfaction Index (ACSI) model

Johnson and Fornell (1991) speak specifically to the problem of constructing broad-based satisfaction measures. They argue that cumulative satisfaction, as an overall evaluation of the consumption experience, should be viewed as a theoretical or latent variable (similar to an attitude). As a latent variable, satisfaction can be empirically measured and meaningfully compared as a weighted-average or index of satisfaction indicators.

The ACSI model exemplifies the measurement and modeling of cumulative customer satisfaction. The model is presented in Fig. 2 (see Fornell et al. 1996). There are three main drivers of satisfaction in the model: quality, value, and customer expectations. Satisfaction, in turn, should reduce the incidence of complaints and increase customer loyalty. Most of the survey questions are rated on 1 to 10-point scales and the measurement variables are specified as reflective indicators of the latent or abstract constructs (circles) in the model.

Of particular interest is the overall satisfaction index, which is a weighted average of three survey ratings: (a) an overall rating of satisfaction, (b) the degree to which performance falls short of or exceeds expectations, and (c) a rating of overall performance relative to the customer's ideal good or service in the category. The three measures provide a highly reliable and stable index of satisfaction. Notice that factors that may be conceptually and empirically distinct drivers of satisfaction in a transaction-specific sat-

isfaction model become multiple measures of a more overall evaluation in a cumulative satisfaction model. Consider the differential role that disconfirmation plays in transaction-specific vs. cumulative satisfaction research. From transaction to transaction or episode to episode, one would expect some variation between performance and expectations. Service providers have a particularly difficult job of conforming to specifications on any given occasion. When these transactions or episodes are accumulated and evaluated over time, they become easier to predict. Thus significant disconfirmation, when it occurs, is more likely when studying a particular transaction or episode than when studying consumption experience to date. Because cumulative satisfaction is a more overall or molar-level evaluation, disconfirmation of expectations is one of several valuable proxies by which customers evaluate their overall experience.

Cumulative satisfaction is systematically related to various financial and business performance measures. Research using the national indices demonstrates the positive impact that cumulative satisfaction has on both market value and accounting returns (Fornell et al. 1996). It is no wonder that customer satisfaction has become an important part of companies' overall accounting, stakeholder management, and business performance measures. Cumulative satisfaction research has also explored the nature of customer expectations and how satisfaction varies across industries. For example, whereas price expectations tend to be more rational, performance expectations are more adaptive; they do not adjust as quickly to changes in performance. Fornell et al. (1996) report that satisfaction is highest among goods industries, lower among services, and lower still among public and government agencies.

2. *Methodological Issues*

In customer satisfaction modeling, satisfaction plays a key role in linking customer perceptions of a product's price and quality with customers' behavioral intentions and subsequent behaviors. Companies use two essential outputs of a satisfaction model to set priorities for improvement. First, how well do products and services perform on their attributes, the benefits they provide, and the customer satisfaction and loyalty that results? Second, what impacts do the attributes and benefits have on satisfaction and subsequent loyalty? That is, how important are the attributes and benefits to customers? This performance and importance information is combined to set priorities for improvement. From a cost–benefit standpoint, the essential benefits to improve are those where importance or impact is relatively high and performance is relatively low.

As latent variables, performance levels on abstract benefits (such as the 'cleanliness' or 'convenience' of a retail store) and overall satisfaction can be constructed using weighted averages or indices of more concrete attributes or proxies. There is a debate, however, as to how to determine attribute and benefit importance measures. Many satisfaction measurement systems rely heavily on customers' direct ratings of importance. The primary advantage of direct importance ratings is that they are easy to collect and analyze. Yet direct measures assume that customers understand what is meant by 'important,' just what attributes are important to them and that they are willing to tell you. As a result, direct importance measures often result in socially acceptable or status quo answers and poor discrimination. More objective estimates of importance or impact can be obtained using statistical estimation. This statistical estimation must accommodate the fact that a satisfaction model is a network of cause and effect relationships (which predicts a pattern of relationships) that contains latent (abstract) variables. Partial least squares (PLS) and principal-components regression (PCR) are particularly well suited to estimating a satisfaction model and deriving importance measures (see Johnson and Gustafsson 2000).

3. *Future Research*

Transaction-specific satisfaction research recognizes that customers may have very complex psychological reactions to a product's performance on a given occasion or over a given time period. Cumulative satisfaction recognizes that, when forming intentions or making decisions regarding their behavior, customers rely on their entire experience to date with a product or service provider. Yet there has been little in the way of an integration of these two perspectives. Going forward, it will be important to better understand just how satisfaction-related constructs and their relationships change when customers move from transaction to cumulative-based evaluations. There is also a need to better understand the psychological and behavioral consequences of satisfaction, specifically customer loyalty (an intention) and retention (a behavior).

Company or brand image (reputation) and relationship commitment (involvement or strength), when combined with overall satisfaction, should explain much greater variation in loyalty and retention than does satisfaction alone. Finally, research in the quality improvement and product development areas must strive to 'bridge the gap' between internal (organizational) quality measures and processes and external (customer) perceptions of quality and satisfaction. It is not sufficient to simply understand what attributes and benefits drive satisfaction. Companies must understand how to break customer satisfaction down into its means of accomplishment within an organization.

See also: Advertising: Effects; Consumer Culture; Consumer Economics; Consumption, Sociology of; Taste and Taste Culture

Bibliography

Folkes V S 1990 Conflict in the marketplace: Explaining why products fail. In: Graham S, Folkes V S (eds.) *Attribution Theory: Applications to Achievement, Mental Health, and Interpersonal Conflict*. Erlbaum, Hillsdale, NJ, pp. 143–60

Fornell C, Johnson M D, Anderson E W, Cha J, Bryant B E 1996 The American Customer Satisfaction Index: Nature, purpose and findings. *Journal of Marketing* **60**: 7–18

Johnson M D, Fornell C 1991 A framework for comparing customer satisfaction across individuals and product categories. *Journal of Economic Psychology* **12**: 267–86

Johnson M D, Gustafsson A 2000 *Improving Customer Satisfaction, Loyalty and Profit: An Integrated Measurement and Management System*. Jossey-Bass, San Francisco

Ölander F 1977 Can consumer dissatisfaction and complaints guide public consumer policy? *Journal of Consumer Policy* **1**: 124–37

Oliver R L 1997 *Satisfaction: A Behavioral Perspective on the Consumer*. McGraw-Hill, New York

Oliver R L, Swan J E 1989 Equity and disconfirmation perceptions as influences on merchant and product satisfaction. *Journal of Consumer Research* **16**: 372–83

Westbrook R A, Oliver R L 1991 The dimensionality of consumption emotion patterns and consumer satisfaction. *Journal of Consumer Research* **18**: 84–91

Yi Y 1991 A critical review of customer satisfaction. In: Zeithaml V (ed.) *Review of Marketing* 1990. American Marketing Association, Chicago, pp. 68–123

M. D. Johnson

Cyborg

In its first definition, a cyborg—literally a contraction of 'cybernetic organism'—was a machine/human hybrid with skills and survival abilities not available to human beings alone. This definition remains widespread both in science and technology studies, and in science fiction. More recently, the hybrid character of the cyborg has been emphasized in additional ways. In particular, it has been defined as an active mixture of fact and fiction, intended as a political tool to transform desirable fictions into facts.

1. *Scientific, Technical, and Military Origins*

The word 'cyborg' was coined by Manfred Clynes and Nathan Kline (1960) in a paper presented to a symposium on the *Psychophysiological Aspects of Space Flight* sponsored by the US Air Force School of Aviation Medicine.

In the context of the challenges posed to human survival in space, and reflecting on the unsatisfactory temporizing needed to provide air and other support systems, Clynes and Kline wrote: 'For the exogenously extended organizational complex functioning as an integrated homeostatic system unconsciously, we propose the term "Cyborg."' They went on to write: 'The purpose of the Cyborg … is to provide an organizational system in which … robot-like problems are taken care of automatically and unconsciously, leaving man free to explore, to create, to think, and to feel.' They listed a series of possible enhancements including both available techniques and imaginative possibilities for enhancing wakefulness, radiation resistance, reduced metabolic demands, oxygen transport, fluid balance, and cardiovascular control.

The 'cyborg' reflects the development of the discipline of cybernetics in the USA in the Cold War. The central concepts of the discipline were developed in an influential series of annual conferences between 1943 and 1954 funded by the MACY Foundation. The major theme of the early MACY conferences was the primacy of information over the material. The aim was to develop a theory of communication and control applicable to humans, animals, and machines alike. The crucial move was to treat organisms and the new machines, computers, as self-controlled and self-regulating information-processing systems. Key figures in this work were Norbert Wiener, John von Neumann, Warren McCulloch, and Claude Shannon.

During the 1950s and 1960s, cybernetics was closely related to work on defense and space exploration. Cybernetically informed research and conferences on bionics, human augmentation, teleoperators, and biotelemetry were funded by the US military and NASA (1963). The concern was with the close coupling of man and machine in weapons systems, the control of human behavior in extreme situations, and augmenting human sensing, intelligence, and decision making under conditions such as combat flying and space travel. This context led to the development of the cyborg as an analytic concept and as a possible socio-techno-scientific assemblage.

2. *Cultural Contexts*

Cyborg-like creatures also proliferated in popular culture where they were used to explore the visionary potential of human–nonhuman hybrids and the mixture of science fact and fiction.

Much 1950s and 1960s science fiction and film was modernist, being either utopian or dystopian. Often it either celebrated or problematized modernity and multinational capitalism, with their futuristic and/or threatening visions of space exploration, nuclear energy, and genetic engineering. It also explored the social, personal, and political threats of invasion by machinelike, coldly rational, and alien others. At the

same time, more 'postmodern' themes were also visible. Boundary disputes around the human body were explored, together with issues of otherness and difference in human/nonhuman encounters. Since 1970, and especially in film, difference has become less threatening, sometimes coming close to dissolution. Aliens and cyborgs have often been pictured more like friendly relatives than hostile invaders. Sometimes, as in Ridley Scott's 1992 *Blade Runner*, they appear more human or humane and less alienated than humans.

At the same time, formerly exotic technologies (cyberspace, virtual life, artificial intelligence, genetic engineering, organ transplants, and xenobiology) have been commodified and domesticated. This is also reflected in cyberpunk, the science fiction sub-genre or genre that grew up in the 1980s (see William Gibson's *Neuromancer* 1984). The cyberpunk literature, often cast in a melancholy mood, wavers between euphoria and nostalgia. It explores the dissolution of human/nonhuman distinctions through electronic technologies, genetic engineering, and the consumer culture controlled by multinational business. Overall, this hugely popular branch of contemporary science fiction has circulated science facts and fictions to such a degree that the cyborg characters of its novels, films, and television series are familiar to an almost global audience.

At the same time, the traffic between science and science fiction is two-way. An early example of the fantasies inhabiting science is found in *The World, the Flesh, and the Devil* (1929), a novel by radical British scientist J. D. Bernal. Since Bernal, science fiction has fuelled developments and informed research and development priorities. Science fiction influences can also be found in contemporary techno-science work on cyborgs: in immersive virtual life applications such as 'cybersuits'; in the bio-robots of artificial life labs; in the development of the bionic chip; and in the diagnostic DNA technology, the 'GeneChip.'

3. *Appreciation*

In *The Communist Manifesto* Karl Marx and Friedrich Engels notoriously observed that 'all that is solid melts into air.' This observation of the erosive character of nineteenth century capitalism prefigures a trope and movement that accelerated during the twentieth century. This is a tendency towards what might be described as 'ontological softening,' the erosion of distinctions between categories and realities that were previously taken to be stable and discrete. Systems thinking, cybernetics, the growth of models of reality as codes of information, and the increasing use of prostheses all tend to problematize previously more stable distinctions. At the same time, they treat future versions of those distinctions as temporarily stabilized expressions of larger, system-based, relations. In both popular culture and techno-science, the cyborg has participated in this process of ontological softening. It has blurred and rendered problematic the distinctions between people and machines, fact and fiction, mind and body, and real and virtual.

At the same time, in the more recent literatures of both social science and science fiction, it has also reflected postmodern skepticism about the limits of organization, the coherence of systems, and the centered, rational, and disembodied picture of human subjectivity implicit in the earlier work of authors such as Clynes and Kline. Instead, reasoning, consciousness, and conscience are seen as more or less unstable or incoherent effects reflecting the mobile and sometimes multiple identities of high modernity. The cyborg has thus become a vehicle for political critique. Nowhere is this more obvious than in the writing of feminist techno-science author Donna J. Haraway, although it is also to be found by others writing within women's studies/feminist theory, postcolonial/cultural studies, science studies, and cyborg anthropology.

Haraway (1985) offers a radical political analysis of the dystopic, gendered, and military character of the 1960s cyborg and the individualist and frontier oriented society from which it emerged. This critique is linked to a vision of the social world as simultaneously non-coherent and exploitative. The cyborg becomes a liberatory set of partial connections between human and nonhuman on the one hand, and a political reality and politically liberatory myth on the other. Seen as a creature located in and produced by partial incoherences, it becomes an ironic figure which participates in and interferes with the current 'state of disorder' to enact better alternative futures. This feminist cyborg thus both extends and rejects earlier work. On the one hand, it continues to soften ontology, but on the other hand, it no longer locates this within a single control-related cybernetic logic. Partially coherent and situated, cyborgs (see *Situated Knowledge: Feminist and Science and Technology Studies Perspectives*) are radical political images that seek to take us beyond system on the one hand and individualism on the other.

4. *Conclusion*

The cyborg concept is a creature of the dissolving character of modernity as it yields space to postmodernity. In this, the human is no longer necessarily central and is replaced by a set of heterogeneous functions. The particular way this dissolution has taken place reflects its circumstances, and in particular US military concerns (command, control, centralization, self-regulation) and US individualism (anarchism, frontier imagery, the lone, self-reliant (usually male) person against both the system and anarchy). These

two circumstances are in tension, as is visible in the differences between cybernetics and cyberpunk. But they are also in tension with a further vision of the cyborg as a form of post-individualist political interference. This resonates with a particularly US version of radical politics which attempts to avoid both holism and individualism in an area of ontological softening and mobile identities.

See also: Artificial Intelligence in Cognitive Science; Feminist Epistemology; Information Technology; Intelligence: History of the Concept; Science, Technology, and the Military; Situated Knowledge: Feminist and Science and Technology Studies Perspectives

Bibliography

Bernal J D 1929 *The World, the Flesh, and the Devil*. Kegan Paul, Trench, Trubner, London

Clynes M, Kline N 1960 Cyborgs and space. *Astronautics*, Sept 26–7, 74–5. In: Gray C H (ed.) 1995 *The Cyborg Handbook*, Routledge, New York

Downey G L, Dumit J (eds.) 1997 *Cyborgs and Citadels. Anthropological Interventions in Emerging Sciences and Technologies*. School of American Research Press, Santa Fe, NM

Featherstone M, Burrows R (eds.) 1996 *Cyberspace/Cyberbodies/Cyberpunk*. Sage, London

Gibson W 1984 *Neuromancer*. Harper Collins, London

Gray C H (ed.) 1995 *The Cyborg Handbook*. Routledge, New York

Haraway D J 1985 Manifesto for cyborgs: Science, technology and socialist feminism in the 1980s. *Socialist Review* **80**: 65–108 *Simians Cyborgs and Women: The Reinvention of Nature*, FA Press, London

Hayles N K 1999 *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics*. Chicago University Press, Chicago

Kirkup G et al. (eds.) 2000 *The Gendered Cyborg. A Reader*. Routledge and Open University Press, London

Levidov L, Robins R (eds.) 1989 *Cyborg Worlds: The Military Information Society*. Free Association Books, London

McCaffery L (ed.) 1991 *Storming the Reality Studio*. Duke University Press, Durham, NC

NASA 1963 *The Cyborg Study: Engineering Man for Space*. Final Report NASw-512. In: Gray C H (ed.) 1995 *The Cyborg Handbook*. Routledge, New York

Sobchack V 1987 *Screening Space: The American Science Fiction Film*. Ungar, New York

J. Law and I. Moser

Cycles: Social

The concept of cycles in the social sciences is multifaceted. Economic cycles, cycles of problem solving in organizations or small groups, or cycles of societal value change are but a few of its forms. Moreover, cycles as a concept are—often simultaneously—used as philosophical heuristics, theoretical explanation as well as analytical technique.

1. Cycle as a Philosophical Concept

Since ancient times the cycle has been believed to be an ideal concept (Weininger 1920). It signifies completeness; all points on a cycle re equal. The wheel of time, eternally going round, is one of the powerful symbols of this kind of thought. Cycles seem to reflect nature quite elegantly: the cycle of life (in modern sociological terms 'life cycles'), the cycle of four seasons (in moderate climates), and the cyclical movements of the sun and stars around the earth (in a Ptolemean view). It is thus not surprising that it took quite a long time for evidence-based science to overrule cyclical concepts so tightly connected to esthetics like Ptolemean worldview.

Christianity marks a counter-philosophy to a strict eternally cyclical worldview: Christ is a singularity, after him nothing will be as it was before. The spell of the eternal wheel of time hence is broken. Christian rebirth is, in contrast to other religions, not a cycle; it is but one step up to an end state. Some philosophers thus believe Christian noncyclical thinking to be among the elements of social and individual progress (Guardini 1946). However, Christian thinking does not exclude cyclical concepts. Christ could be thought of as being a singularity, separating one big cycle from the next.

Another strong symbol of societal and human dynamics is the straight, often directed, line. Linear, noncyclical dynamics of societal developments is one of the basic concepts of modernity and progress. Finally, a sequence of stages is also among the influential philosophical concepts, e.g., Kohlberg's or Piaget's stages in the ethical development of children, adapted by Habermas for societal development (Habermas 1989, Kohlberg 1981, Piaget 1970).

All three ideal concepts, cycle, linear trend, and stage, represent directed dynamics in contrast to stochastically, nondirected fluctuations. Thus they can be used to explain a given societal situation as being embedded in a specific movement towards an end or with a 'known' outcome (next phase going up, going down, coming back). They allow the sequence of historical events to be given an inherent and consistent meaning. Linear dynamics is inherent in socialist and capitalist ideologies, but not necessarily part of social realities (cf. Häder and Mohler 1995 on cyclical worldviews in the GDR). Modern cyclical philosophies, like those based on Friedrich Nietzsche or Oswald Spengler, hence can be understood as counter-attacks on modernism and its Christian roots by conservative intellectuals (A. Mohler 1972 Wendorff 1980).

2. Cycles in Sociological Theory

Sociology is, by definition, a modern science. Ideal societal cycles are seemingly concepts of premodern thinking. Thus it would be hard to find any modern sociologist who explicitly subscribes to cyclical movements. The term is almost absent in modern textbooks. However, reading through influential books on modernizing organizations and companies one finds quite often fourfold sequences of development, i.e., organizational dynamics (Senge et al. 1994). Whatever the labels used are, these fourfold tables resemble quite strongly Bales' concept of problem-solving sequences in social groups (Bales 1953, Bales and Strodtbeck 1953) which in turn relates to Schumpeter's classification of business cycles (1939). Based on empirical observation, Bales proposed a four-stage dynamic for small group problem solving: expressive, adaptive, instrumental, and integrative. Parsons adapted this scheme and transformed it into his famous AGIL scheme (Adaptive, Goal Attainment, Instrumental, Latency; Parsons 1971). Thus one of the most influential modern sociological theories takes advantage of a cyclical model in explaining societal developments. However, other than Bales for small groups, Parsons did not present strong empirical evidence for his assumption of similar dynamics on the societal level. While the AGIL scheme has been taught to many thousands of students during the latter half of the twentieth century, its assumed cyclical societal dynamics has had little influence on actual social research.

3. Cycles in Economic Theory

By contrast, in economics, cyclical dynamics has a long pedigree and was widely discussed (Forrester 1977, Gordon 1952, Gulap 1989, Kleinknecht et al. 1992, Metz 1984, Wallerstein 1980). However, one should differentiate between 'cyclical movements' and 'cycles.' The former refers to fluctuations but not to ideal, recurring cycles. It is this rather ambiguous concept which is mostly used in the literature and the media.

Among the most prominent economic cycles, in the sense of an ideal concept, are Kondratieff's long waves in economic life (1926). These were understood by the Soviet government as a counterattack on the Marxian linear progress concept. Consequently they sent Kondratieff to the Gulag. Schumpeter (1939) in turn developed a complex system of interlaced and superpositioned cycles with Kitchin cycles as the shortest (average three years), followed by Juglar cycles of seven to 11 years, and Kondratieff cycles with 49 to 60 years wavelength.

Today ideal cycles in the Kondratieff and Schumpeter tradition are discussed in relatively small economic circles. Sources for information on cyclical socioeconomic thinking are: Long-Wave Theories of Development (http://faculty.washington.edu/krumme/207/development/longwaves.htm) and the Longwave and Social Cycles Resource Centre (http://www.1-888.com/longwave/lwintro.html).

However, the relative success of cyclical concepts in economics compared to sociology reflects observed ups and downs in economic development coupled with long empirical observations of macroeconomic characteristics like price indices, average growth rates, interest rates, etc. It is also arguable that the prevalence of cyclical thinking in economics is itself cyclical. In a phase of economic growth, linear concepts are prevalent (accepting few setbacks) without seriously investigating the next turning point. As soon as the economy slides into recession cyclical concepts are 're-invented' by predicting the next turning point into an upward development.

4. Critical Issues in Identifying Cycles

Ideal concepts like cycles require sophisticated theoretical arguments to convince skeptics, and special techniques to extract cycles from the usual mass of data. Critical issues in defining cycles thus are: identifying periods with homogeneous dynamics (i.e., no singularity disrupting a cyclical movement), interlacing and/or superposition of cycles of different wavelength, combination of cycles and linear trends, identification of wavelengths and their respective measurement.

Identifying periods of identical wavelength is crucial for cyclical theories. Otherwise there would be eternal cycles, which are counter-evident even to everyday observations (cf. the change in the generational sequence due to higher life expectancy in the last century). Whether a certain historical event is a singularity or just a very prominent event within an otherwise homogeneous cycle is an important issue considering long social cycles. For instance, did the end of WWI and the beginning of WWII represent singularities separating distinct cycles? If so, there would be no time for a full Kondratieff cycle of at least 48 years length between the two events. How does one interpret the interlacing or superposition of different cycles? Are the cycles really a fine-tuned system of small and big cycles, as Schumpeter assumed? What happens if big cycles (long waves) are not multiples of small cycles? If uneven movements due to nonsynchronous multiple cycles in organizations, cohorts, institutions, business, the arts, education, science, welfare, etc. are at work, can one in the end identify the grand cycle? Or, will it be that due to the individual cyclical dynamics the overall outcome will result in unique cyclical superposition at each and every point in time (cf. the discussion between P. Ph. Mohler 1987, 1989, and Namenwirth and Weber 1989). Finally, as is the case with a straight line plotted through a cloud of

data points, a cycle is an analytical construct. Stochastic fluctuations are filtered and smoothed to get as close as possible to the ideal concept of a pure sine wave (i.e., cycle, Namenwirth and Weber 1987, Allen 1997, Drucker and Wesche 1999). Because of the strong underlying, non-modern theoretical assumptions, models using cycles as a heuristic are much more controversial than the linear models used in everyday statistics.

5. *An Example of Empirical Analyses of Cycles in Societal Development*

Instead of the numerous and often less theoretically anchored economic analyses, a strict sociological approach is presented here as an example of identifying cyclical dynamics. Namenwirth and Weber showed in their seminal work on *Dynamics of Culture*, how to use sophisticated methods parsimoniously to establish an empirical model of the real world (1987).

It is obvious that a direct translation of small group dynamics into societal dynamics must fail due to the different time scales at work (Luhmann 1973, Wendorff 1980). Moreover, Bales' work is based on laboratory experiments only. Such experimental settings make the standardizing of situations and filtering of noise much easier than is the case in observing real situations. Thus any empirical investigation into the validity of Parson's assumption that the dynamics of society as a whole is similar to that of the small group has to overcome serious technical and methodological obstacles.

Any observation of a cycle requires at least a full cycle; better, a series of cycles. If societal dynamics works on a time schedule similar to that observed for economic cycles, then to observe one full long Kondratieff cycle requires time series of about 60 years. Considering that systematic sampling-based interviews only came into existence in the 1920s and assuming that enough comparable data is available, early twenty-first century sociologists have barely enough data to analyze one full Kondratieff cycle. However, Namenwirth and Weber found even longer cycles in their data (about 120 years), thus it will be the middle of the twenty-first century before their results can be verified with robust survey data. Moreover, for such long time series, the society observed should keep its basic identity over the period of observation, i.e., the time series should not be interrupted by a singularity.

As Namenwirth showed as early as 1973, there are indeed both time series and societies at hand, which allow for an empirical test of Parson's assumption. However, as we will show, all observed dynamics are a superposition of both linear tends (modernism) and cyclical dynamics. Namenwirth's solution to the data problem was to take textual indicators measured on a group level instead of individualistic interviews. In doing so, he freed long-term sociological investigation from the regression to interviews as its only source of information. He also avoided the individualistic fallacy of inferring from individual level data to societal level dynamics. As a data source he used party platforms of the two leading American parties since 1844. He found two cycles, one of about 152 years (actually longer than the observed period of 120 years), and one of about 48 years (Namenwirth and Weber 1987). Weber refined Namenwirth's technique by separating linear trends and cyclical dynamics in his analysis of 'speeches from the throne 1689–1972' (statements of the government read by the monarch).

In additional to cycles similar to those found by Namenwirth in the United States, his data indicated a break/singularity in the dynamics at about 1795—note the coincidence with the French Revolution. Thus he found two periods with different cyclical dynamics which he named 'mercantilist' and 'capitalist.' However, studies using equivalent measures and similar approaches did not reveal cyclical dynamics (Eisner 1991 for Switzerland, and P. Ph. Mohler 1987 for Germany). Finally in a reanalysis of Sorokin's famous data on cultural changes, Klingemann et al. (1982) found strong evidence for long-term trends plus long waves of philosophical dominant thinking. In accordance with the above observation of the general absence of cyclical theories in the social sciences, the data showed an extremely strong (linear) trend for empirical philosophical theories since 1720, i.e., since then the predominant development in philosophical thinking has been the linear increase of a specific of philosophy over a long period of time.

The techniques developed by Namenwirth and Weber for identifying cyclical dynamics are non-standard in time series analysis. Other than the rather complex mathematical models used there, they use two straightforward strategies: first, they check for a linear trend in the time series of a single variable; second, they test for a short cyclical dynamic in the residuals of the de-trended data; and third, they search for a long trend in the residuals of the residuals of the de-trended data. Because the target is the wavelength, they do not attempt to estimate the magnitude of the cyclical dynamics; thus the amplitude is always set to '1'.

In essence, this approach is a special form of residual analysis. Translated into theoretical concepts: first, a linear dynamic is assumed, which is, second, underpinned by one or two cyclical dynamics. These cycles can be interpreted within an AGIL-like pattern. Moreover, Weber's observation of a break between mercantilist and capitalistic dynamics may be taken as an indicator of a singularity in a society (nothing will ever be like it was), which allows differentiation between epochs like Bales might differentiate between different social groups.

In doing so, this approach re-specifies non-directional theories (like Parsons,) into directed ones.

It also bridges the gap between linear modernism and cyclical conservatism. Modernity is the linear (but not infinite) dynamic in societies, while the cyclical elements of the moving process point to basic laws of problem solving (assuming that problems have to be solved, if at all, in a finite period of time). Singularities, like the one between mercantilism and capitalism in 1795, make us aware that there is no infinite continuous movement in societies.

6. Conclusion

Cycles and straight lines are two idealizations of directed movement (dynamics). Linear trends, be they directed upwards or downwards, have predominated in sociological thinking to the end of the twentieth century. Cyclical theories seem to be absent in the grand social theories, but they are not. They are ever present in schemata like the famous Parsonian AGIL scheme. Cycles are, often without mentioning the concept, inherent in almost all contemporary approaches dealing with organizational and group changes. This indicates the robustness and validity of Bales' early findings. Survey data will not be available for a long time as a base for analyzing societal dynamics, be they linear, curved, or even cycles. Thus one must rely on other data to model the dynamics of a given society.

In economics, long waves are a dated concept. The concept of business cycles, however, is widely used to identity regularities within the fluctuations of economic indictors.

See also: Fashion, Sociology of; Pareto, Vilfredo (1848–1923); Parsons, Talcott (1902–79); Schumpeter, Joseph A (1883–1950); Social Change: Types; Social Changes: Models; Time, Anthropology of; Time, Chronology, and Periodization in History

Bibliography

Allen D S 1997 Filtering permanent cycles with complex root units. Federal Reserve Bank Working Paper 97-0014A, St Louis, MO

Bales R F 1953 The equilibrium problem in small groups. In: Parsons T, Bales R R, Shils E (eds.) *Working Papers in the Theory of Action*. Free Press, New York

Bales, R F, Strodtbeck F L 1953 Phases in group problem solving. *Journal of Abnormal and Social Psychology* **46**: 485–95

Drucker M Wesche K 1999 European business cycle: New indices and analysis of their synchronity. Federal Reserve Bank Working Paper 99-019A St Louis, MO

Eisner M 1991 Politische Sprache und sozialer Wandel. Eine quantitativ-semantische Analyse von Neujahrsleitartikeln in der Schweiz von 1840–1987. Dissertation, Zurich

Forrester J W 1977 Growth cycles. *De Economist* **125**(4): 525–43

Guardini R 1946 *Der Heilsbringer, in Mythos, Offenbarung und Geschichte*. Thomas, Zurich, Switzerland

Gulap H 1989 The stages and long-cycles of capitalist development. *Review of Radical Political Economics* **21**(4): 83–92

Habermas J 1989 *Moral Consciousness and Communicative Action*. Polity Press, Cambridge, UK

Häder M, Mohler P Ph 1995 Zukunftsvorstellungen der Menschen als Erklärungsvariable für die Krise der DDR und die gegenwärtige Situation in Ostdeutschland. *Politik und Zeitgeschichte* **27**: 19–27

Kleinknecht A, Mandel E, Wallerstein I (eds.) 1992 *New Findings in Long-Wave Research*. Macmillan, London

Klingemann H D, Mohler P Ph, Weber R P 1982 Cultural indicators based on content analysis: A secondary analysis of Sorokin's data on fluctuations of systems of truth. *Quality and Quantity* **16**: 1–18

Kohlberg L 1981 *The Philosophy of Moral Development: Essays on Moral Development*. Harper and Row, San Francisco, Vol. 1

Kondratieff N D 1926 Die langen Wellen der Konjunktur. *Archiv für Sozialwissenschaft und Sozialpolitik* **56**: 573–609 [first English partial translation 1935: The long waves in economic life. *Review of Economic Statistics* **17**: 105–115. Full English translation 1979: The long waves in economic life. *Review* **2**: 519–62]

Luhmann N 1973 Weltzeit und Systemgeschichte. In: Ludz P C (ed.) Soziologie und Sozialgeschichte. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*. Special Issue **16**: 81–97

Mohler A 1972 *Die konservative Revolution*, 2nd edn. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany

Mohler P Ph 1987 Cycles of value change. *European Journal of Political Research* **15**: 155–65

Mohler P Ph 1989 Cyclical ideology: A response to Namenwirth & Weber. *European Journal of Political Research* **17**: 121–4

Namenwirth Z J 1973 The wheels of time and the interdependence of value change. *Journal of Interdisciplinary History* **3**: 649–83

Namenwirth Z J, Weber R P 1987 *Dynamics of Culture*. Allen & Unwin, Boston

Namenwirth Z J, Weber R P 1989 Cycles of the fourth kind: A reply to Mohler's cycles of values. *European Journal of Political Research*. **17**: 118–20

Parsons T 1971 *The Social System*. Free Press, Glencoe, IL

Piaget J 1970 Piaget's theory. In: Mussen P H (ed.) *Carmichael's Manual of Psychology*. Wiley, New York

Schumpeter J A 1939 *Business Cycles*. McGraw-Hill, New York

Senge P M, Kleiner A I, Roberts Ch, Ross, R B, Smith B J 1994 *The Fifth Discipline Fieldbook. Strategies and Tools for Building a Learning Organization*. Currency, New York

Weininger O 1920 *Über die letzten Dinge*, 6th edn. Braumüller, Vienna

Wendorff R 1980 *Zeit und Kultur*. Westdeutscher Verlag, Opladen, Germany

P. P. Mohler

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

D

Dance, Anthropology of

Dance has been studied in anthropology for over a century, also in classical works, but mostly in passing as an element in ritual and ceremony. This neglect is likely to have come about because of the elusive nature of dance, in combination with the fact that dance has the potential of releasing emotional and erotic forces that are kept at bay most of the time in everyday life. There are many examples of political and religious control of dance, not least in colonial contexts. This has often produced resistance, expressed through the creation or revitalization of ethnic and national dances.

In an article about the beer dance among the Azande, E. E. Evans-Pritchard (1928) was among the first anthropologists to show that the study of dance had a wider analytical potential and that dance needed to be addressed from a contextual perspective. But it was not until the 1960s and 1970s that the anthropology of dance emerged as a subdiscipline, evoking discussions about the origins and definition of dance that became complicated by the cross-cultural perspective. As Adrienne Kaeppler (1985, p. 93) has remarked, the concept 'dance' is a Western category which often fails to distinguish among different cultural activities. This concern is also paralleled, however, in the anthropology of music and the arts. Drawing on her research in Tonga where movement around ceremonial pig presentations, *kava* mixing, and choreographed group speeches revealed hierarchy and social solidarity, Kaeppler suggested 'structured movement systems' as an alternative concept to that of dance. Another anthropological strategy has been to identify dance events as units of study, rather than to look for particular dances in isolation. On the whole, dance anthropologists seem to converge on a consideration of bounded rhythmical movements that are performed during some kind of altered state of consciousness, an elevation or even trance.

The pioneering dance anthropologists had to deal with the ethnocentric assumptions prevalent among the then mostly autodidact dance historians, about dance as a universal language and 'primitive dance' (often constructed in terms of a generalized 'African dance'), as an early phase in a Western scheme of dance evolution entailing a hierarchicalization where ballet was regarded as a superior dance form. This prompted Joann Kealiinohomoku (1983) to write an article where she argued for a view of ballet as an ethnic dance, as one dance form among others. This is a crucial anthropological contribution to dance scholarship, but Kealiinohomoku's idea that ballet reflects Western cultural traditions is perhaps a bit exaggerated, especially in the long run. Admittedly Western, ballet speaks to a limited historically dominant segment of the West, by contrast with the shifting multicultural landscape of contemporary Europe and North America. To state that ballet is a form of ethnic dance also disregards the traditional difference between ethnic dance as a participatory ritual practice and ballet as a stage art performance separated from the audience, although ethnic and folk dance are now frequently moved from secular and sacred ritual to the realm of art in stage dance. The genres are further blurred by contemporary stage dance experiments with the audience taking part in the performance—even being invited to go on stage.

The central quest and *raison d'être* for the anthropology of dance has been to find out what dance says about its society, informing about social or cultural circumstances that cannot be sufficiently expressed in any other way. This does not only come about through direct reflection, but also through social commentary and critique in the form of parody like (gender) role reversal or suggestions about alternative social scenarios and even predictions of political events. Dance has increasingly been recognized as having an impact on social change (Hanna 1988, Royce 1977). But there is also the cathartic view of dance as a safety valve, a structural outlet, which will keep the social order and its inequalities intact. Dance is moreover often a feature of socialization. One of the functional aspects of dance is that of transformation: the classical examples are rites of passage when people are moved from one stage in the life cycle to the next one, such as initiation rites, wedding celebrations, and funerals.

Like all subdisciplines, the anthropology of dance has followed main theoretical currents in its parent discipline. Dance has thus been analyzed through structuralism, symbolic anthropology, semiotics, and linguistic theory, generating studies of the deep structures of dance and dance as nonverbal communication. Taking these theoretical approaches as points of departure, some of the first dance anthropologists worked out their own elaborate models and classificatory schemes for movement analysis. Drid Williams' (1976) theory of semasiology, the meaning of human movements as action signs, for instance, has

been developed further by Brenda Farnell (1995) in a study of Plains Indian sign language.

There are basically three kinds of anthropological studies of dance. One kind has dealt with meaning, function, and the cultural context of dance, another kind has concentrated on the choreographic form. The majority of anthropological studies of dance, however, have combined these approaches by taking both meaning, function, and context *and* choreographic form into account, as well as including textual descriptions of the movements and/or illustrations such as photographs or simple movement transcriptions consisting of circles and arrows that indicate directions and patterns of dance.

In the United States, research on dance in its cultural context has dominated, just as in the United Kingdom where John Blacking was a key figure in the emerging interest in the anthropology of dance. In continental Europe, dance ethnologists, also called choreologists, who came out of a long nationalist folklore tradition, documented folk dances in their own societies, for preservation and revival, stressing formal characteristics along the lines of music research there. During the Communist era in Eastern Europe, dance (and other folklore) researchers found themselves in ideological conflict over state manipulations of their objects of study and control of their investigations (Giurchescu and Torp 1991).

In the 1980s a new generation of dance anthropologists joined the growing interdisciplinary area of dance and movement studies. Engaging with feminist and gender critiques, notions of the body, identity politics, and literary and cultural studies, scholars of dance and movement studies became more visible in the academic world, also because there was an increasing number of them. Dance historians have moved towards ethnography and social and cultural theory redefining their field as critical dance studies, while dance anthropologists have added Western stage dance and culture, as well as European folk dance, to their previous focus on non-Western dance and movement. Most dance scholars, including dance anthropologists, have been trained as modern and/or classical dancers. The anthropological study of dance usually requires participation in the dancing, at least some kind of bodily knowledge of the steps that are investigated. The fact that dance anthropologists learn sequences of steps and movement that are completely different from their earlier movement experiences is something they often report on in a humble spirit, while it actually deserves recognition not least as an important methodological strategy. Some dance anthropologists have thought of their own dancing in the field as a part of the recent reflexive turn in anthropology, which has also had an impact on the anthropology of dance. The presence of the ethnographer in the field seems to have been more of a problem for neighboring dance scholars, however, who have worried that this would entail undue subjective influence, than for those who have been trained in anthropology.

1. *Moving Bodies, Gender, and Sexuality*

In the upsurge of studies in the human and social sciences on the body as culturally constructed, the *moving* body has nevertheless continued to be marginal (Reed 1998), except in recent dance scholarship including dance anthropology where it has been prominent often in tandem with issues of gender and sexuality. Such issues, also including gay dancers, were early identified by Hanna (1988) in a cross-cultural review. Since then a few monographs that combine attention to moving bodies, gender, and sexuality have appeared. Jane Cowan's (1990) work on the body politics of gender in a town in Northern Greece focuses on dance events, including for example lively wedding dances which unite a contradictory ideology of inequality and visions of closeness in the dancing bodies. For the women, dance is an ambivalent pleasure since they have to balance their sexual expression carefully between release and control; unlike that of the men, women's dancing demeanor always runs the risk of being interpreted as indecent.

This is also a problem for the professional female entertainers, singers, and belly dancers in Cairo that Karin van Nieuwkerk (1995) writes about, who see their job as 'a trade like any other.' Performing at celebrations such as births, engagements, and weddings, these Egyptian entertainers create much happiness and are in fact crucial for the success and prestige of these occasions. Nevertheless, they are regarded as dishonorable in the context of Islam and class in Egyptian society because they make a living out of disclosing their bodies in public, thereby tantalizing men. In Cynthia Novack's (1990) ethnography of contact improvisation, the American communal art-sport that emerged in the 1960s, gender relations are quite differently featured in line with the prevailing ideology of egalitarianism. Although the body is constructed as nongendered in contact improvisation, the performers' touching and supporting of each other's weight while moving together, usually in duets, sometimes produce sexual and sensual feelings both among them and the observers.

Novack's study spans more than two decades, showing that the movements of contact improvisation, and thus the construction of these moving bodies, and the arts organizations around them, followed changes in American culture so that this dance form gradually diminished together with the egalitarian agenda of the 1960s and 1970s. Body and space are in focus in Sally Ann Ness's (1992) interpretation of the sinulog danced in Cebu City in the Philippines. Ness connects dance movements and space to everyday movements like walking and the urban landscape of Cebu City. A Philippine ballet, *Igorot*, which Ness (1997) discusses, produces neoethnic bodies in a structure of post-

colonialism and transnationalism. J. Lowell Lewis (1992) also links everyday movement and dance movement in Brazilian *capoeira*, the martial art and dance with roots in Brazilian slavery and, further back, in Africa. Lewis analyzes *capoeira* as a cultural style through the metaphor of play: bodily, musical, and verbal.

2. *Ethnicity, Nationalism, and Transnationality*

Ethnicity and nationalism are often negotiated through dance, but in different ways depending on historical and political circumstances. J. Clyde Mitchell's (1956) small classic monograph on the Kalela dance in colonial Northern Rhodesia is an early instance of ethnic dance. The Kalela dance was performed every Sunday by labor migrants who had moved from the countryside to the town of Luanshya. While describing colonial society, the dance was also a marker of a 'tribal' group identity called for in contrast with other groups in the town. The European folk dances which European dance ethnologists have long been documenting can be categorized as ethnic dances. Yet it was the growth in studies of nationalism in the social sciences generally that inspired dance anthropologists to investigate ethnic and national dance, for example, by relating Irish traditional dancing and Javanese court dance to the wider escalating concern with questions of representation, authenticity, and appropriation. In a study of the *devadasis,* female temple dancers in South India, Joan L. Erdmann (1996) points out that it is time to rewrite the history of Indian dance without orientalism such as the Western label 'oriental dance,' in order to make room for indigenous categories. This is a completely different situation than that in post-revolutionary Cuba which Yvonne Daniel (1995) discusses in her study of how the government replaced ballet with rumba as the Cuban national dance. Rumba was associated with the working class and the African heritage, which suited the socialist state ideology. Although most dance anthropologists seem to agree that all dance forms are more or less mixed, with increasing global contacts, ethnic and national dance forms are becoming even more mixed than before. *Igorot*, the Philippine ballet which is a mixture of classical ballet and traditional Philippine dance, has been described as an instance of cultural hybridity (Ness 1997). Helena Wulff (1998) discusses homogeneity and diversity in the transnational world of ballet, drawing on Howard Becker's (1982) notion of art worlds.

3. *Transcription, Technology, and Property Rights*

To convey movement through the text is an obvious ambition for dance anthropologists. In her little book on tango and the multidimensional national Argentine identity, accurately entitled *Paper Tangos*, Julie Taylor (1998) has inserted small black-and-white photographs in the margin on every page. A quick flipping through the pages thus creates an image of live movement, of a film as it were, of couples dancing the tango.

In addition to simple figures illustrating the text, some dance anthropologists have learnt advanced standardized notation systems for transcription of movements, such as Labanotation, which is the most widely used, or Benesh or Eskhol–Wachmann notation systems. As it takes about two years to master any of these systems completely, it is like learning another language. In the Western dance world, trained notators, or choreologists, document ballet and dance productions through these systems, and they have also been important in the medical rehabilitation of disabled people. There is, however, a degree of interpretation involved in the notating process, which means that different notators may describe the same movement somewhat differently. There has also been a discussion in dance anthropology about the applicability of these Western notation systems to non-Western dance and movement forms, although there seems to be an agreement that they can be used for such purposes as well. Certain conventions have developed to deal with cultural translation problems, so that the particular use of Labanotation, for example, may be explained in a 'key' for readers.

Photography and film are other, often complementary techniques, for conveying dance and movement. There is, however, an element of cultural selectivity in their use, as in video recording which has increased substantially during recent years. Video has also become a significant aid in the mechanical work with drawing the symbols for the scores of Labanotation and the Benesh notation system. The expanded use of technology in dance research has accentuated questions of copyright and ownership which are further complicated in a transnational context. The area of dance and technology (also including television, CD-ROM, and choreography software computer programs) is generating a growing amount of new research.

Future research in the anthropology of dance is likely to include more systematic reception and audience research. The anthropological inquiry into dance is already being extended to popular dance genres such as rave, dance shows and dance in musical videos, and to world dance, both in Western and non-Western contexts. Unlike ethnomusicology which attracts mostly men, the majority of dance anthropologists have been women. This may have played a role in the relatively modest position of the subdiscipline. Another circumstance that works against an accumulation of knowledge that would produce major growth in the anthropology of dance is that anthropologists often write one book on dance and then move on to other research interests.

Still, the anthropology of dance contributes to central debates in the human and social sciences, especially to an understanding of culture in terms of process and diversity. Importantly, dance and movement are not only shaped by society—dance and movement also shape society. This occurs in increasingly complex ways as the genres of dance and movement in society shift and grow, inviting more studies of the anthropology of dance.

See also: Body: Anthropological Aspects; Cultural Relativism, Anthropology of; Entertainment; Evans-Pritchard, Sir Edward E (1902–73); Religion: Nationalism and Identity; Ritual; Ritual and Symbolism, Archaeology of

Bibliography

Becker H S 1982 *Art Worlds*. University of California Press, Berkeley, CA

Cowan J K 1990 *Dance and the Body Politic in Northern Greece*. Princeton University Press, Princeton, NJ

Daniel Y 1995 *Dance and Social Change in Contemporary Cuba*. Indiana University Press, Bloomington, IN

Erdmann J L 1996 Dance discourses: Rethinking the history of the 'oriental dance'. In: Morris G (ed.) *Moving Words*. Routledge, London

Evans-Pritchard E E 1928 The dance. *Africa* **1**: 446–62

Farnell B M 1995 *Do You See What I Mean?* University of Texas Press, Austin, TX

Giurchescu A, Torp L 1991 Theory and methods in dance research: A European approach to the holistic study of dance. *Yearbook for Traditional Music* **23**: 1–10

Hanna J L 1988 *Dance, Sex and Gender*. University of Chicago Press, Chicago

Kaeppler A L 1985 Structured movement systems in Tonga. In: Spencer P (ed.) *Society and the Dance*. Cambridge University Press, Cambridge, UK

Kealiinohomoku J 1983 [1970] An anthropologist looks at ballet as a form of ehnic dance. In: Copeland R, Cohen M (eds.) *What Is Dance?* Oxford University Press, Oxford, UK

Lewis J L 1992 *Ring of Liberation*. The University of Chicago Press, Chicago

Mitchell J C 1956 *The Kalela Dance*. Institute by the Manchester University Press, Manchester, UK

Ness S A 1992 *Body, Movement and Culture*. University of Pennsylvania Press, Philadelphia

Ness S A 1997 Originality in the postcolony: Choreographing the Neoethnic body of Philippine dance. *Cultural Anthropology* **12**(1): 64–108

van Nieuwkerk K 1995 *A Trade like Any Other*. University of Texas Press, Austin, TX

Novack C 1990 *Sharing the Dance*. University of Wisconsin Press, Madison, WI

Reed S A 1998 The politics and poetics of dance. *Annual Review of Anthropology* **27**: 503–32

Royce A P 1977 *The Anthropology of Dance*. Indiana University Press, Bloomington, IN

Taylor J 1998 *Paper Tangos*. Duke University Press, Durham, NC

Williams D 1976 Deep structures of the dance. *Journal of Human Movement Studies* **2**: 123–44

Wulff H 1998 *Ballet across Borders*. Berg, Oxford, UK

H. Wulff

Darwin, Charles Robert (1809–82)

1. Darwin's Life

Charles Robert Darwin was born in Shrewsbury, England, on February 12, 1809. His Father, Robert Waring Darwin (1766–1848) was a prosperous physician. His mother, Susannah Wedgwood (1765–1817), was the daughter of the industrialist Joshiah Wedgwood II (1769–1843). She died on July 7, 1817. This early bereavement is often invoked as having affected Darwin's health, both physically and mentally.

Although he developed an extracurricular interest in the natural sciences, the young Darwin was considered a mediocre student. At the age of 16 he was sent to Edinburgh to study medicine. He interacted with scientists there and furthered his interests in natural history, but he had no desire to become a physician and realized he would not need to earn a living. His father however insisted upon his having a respectable profession, and it was decided that he would prepare to become a clergyman. He therefore enrolled at Cambridge University on October 15, 1827, and remained there until 1831. His academic performance was respectable, though not outstanding. But his scientific interests were encouraged by the faculty, especially the botanist John Stevens Henslow (1796–1871).

Darwin was preparing for some scientific travel when he was invited to join a naval surveying expedition on HMS Beagle as unofficial naturalist and gentlemanly companion to the captain. The Beagle departed on December 27, 1831, circumnavigated the globe, and returned on October 2, 1836. This long period of field research laid the foundation for Darwin's later career. Working in isolation no doubt forced him to think for himself, and he had much opportunity for reflecting on the many books he had with him, especially the *Principles of Geology* (1830–2) by Charles Lyell (1797–1875). Darwin became one of Lyell's major supporters. While in South America he developed a theory of coral reefs which he verified later on during the voyage. It was his geological work that first made him famous. However he also studied

zoology and botany and assembled a superb collection of specimens.

Darwin did not become an evolutionist until after his return to England, though he began to take the idea seriously toward the end of the voyage. Back in England he began to sort out his specimens and put them in the hands of specialists. Expert opinion, especially on the birds of the Galapagos Archipelago, convinced him (ca. early March, 1837) that evolution was a viable idea, and he set out to develop a theory. That included a search for a mechanism. His notebooks indicate a vast amount of reading and thinking about many topics, including psychology.

Darwin's discovery of natural selection occurred around September 28, 1838, when he read the *Essay on the Principle of Population* (Malthus 1826). Its author, who went by the name of Thomas Robert Malthus (1766–1834), was the first professor of economics. Darwin had already changed his conception of the natural economy to a competitive one but he had not made the crucial connection. The insight came from understanding what happens to human reproduction as resources get harder to come by with an increasing number of mouths to feed. It then became evident that some individuals would reproduce more than others. How much economics Darwin actually knew is questionable, but the general notion of individualistic competition was clearly available from the economics of his day. When, in the 1850s, Darwin developed his ideas about biodiversity he was to some extent helped by his reading of the French zoologist Henri Milne-Edwards (1800–85), who applied the idea of the division of labor to physiology and anatomy. Darwin realized that specialization would make for more efficient use of resources. Evolving populations would seize upon different places in the economy of nature (Eltonian niches, as they are now called).

Darwin married his cousin Emma Wedgwood on January 29, 1839; their first child was born on December 27 of that year. They set up house in London, but his health began to deteriorate and he suffered from mysterious ailments for the rest of his life. In 1842, the family relocated to a country residence near Down, south of London. Nonetheless Darwin was highly productive. A book on his voyage (Darwin 1839) was a popular as well as a scientific success. He edited the zoology of the voyage and, in addition to minor papers, published three volumes on his geological research: (Darwin 1842, 1844, 1846). Development of his ideas on evolution continued, and he wrote a sketch of a book on it in 1842, and a draft in 1844.

In 1846, Darwin began some research on barnacles which led him to spend eight years revising the group and writing a vast taxonomic monograph (Darwin 1851, 1854) which in effect was his first book on evolution. He then proceeded to work up his materials on evolution. A large book entitled *Natural Selection* was in the process of composition when, on June 18, 1858, a manuscript from Alfred Russel Wallace (1823–1913) arrived in which the same basic mechanism was proposed. A joint paper was read before the Linnean Society on July 1, and on July 20, Darwin began to write an 'abstract' that ultimately became the *Origin of Species* (Darwin 1859). The first of six editions was published on November 24, 1859.

Darwin brought out a series of books that dealt with a broad range of evolutionary topics. These books all included results of his ongoing empirical research, much of it on the sexuality, physiology, and behavior of plants. A book on orchids (Darwin 1862) was followed by one on climbing plants that first appeared as a journal article (Darwin 1865). What had been a section of *Natural Selection* became *The Variation of Animals and Plants Under Domestication* (Darwin 1868). It was the second of what are considered Darwin's 'major' evolutionary works. The third was *The Descent of Man, and Selection in Relation to Sex* (Darwin 1871). *The Expression of the Emotions in Man and Animals* (Darwin 1872) was followed by a series of botanical works (Darwin 1875, 1876, 1877a, 1877b, 1880). The last of his books dealt with earthworms, especially their behavior (Darwin 1881). Darwin died at his home at Down on April 19, 1882. He was buried at Westminster Abbey on April 26.

2. *Darwin's Scientific Contribution*

Darwin brought about the most important revolution in the history of the human intellect. For that very reason it has taken a long time for his ideas to become assimilated and the process is far from complete. Although he was able to convince the scientific community that evolution had in fact taken place, he was far less persuasive with respect to natural selection, which only became the majority view towards the middle of the twentieth century. Furthermore, evolution as Darwin and his successors have come to understand it was something quite different from the 'evolution' of his predecessors and many of his contemporaries. This is true of both evolution and its mechanisms.

Darwin's early geological research was crucial to his intellectual development. It gave him experience in thinking historically, dealing with real events that happened to concrete particular things. The fossil record, however, was still very poorly understood, and one of Darwin's major contributions was showing how incomplete it was. It was biogeography, not paleontology, that convinced Darwin that evolution had occurred and provided his most compelling arguments for it. But that required thinking about groups of organisms not merely changing but also having common ancestors. Darwin's predecessor Lamarck (1744–1829) had published a branching diagram, but his basic conception was that of multiple

origins of simple organisms by spontaneous generation, and their descendants becoming increasingly complex as independent lineages. It was a more historical version of the traditional idea of a *scala naturae*, with organisms ranked as higher or lower according to how closely they resembled human beings.

When Darwin's predecessors attempted to explain the transformation of species they invoked the kinds of causation that they were familiar with in other areas of experience. Supernatural influences played a diversity of roles in such thinking, often important ones. Change in lineages was largely thought of in terms of analogies with the development of embryos. A kind of 'evolutionism' was possible in which simple organisms had been created followed by a long succession of generations before the result that had been foreseen from the beginning—Man—was realized. Alternatively, the Creator had ordained laws of nature, such that matter would organize itself into increasingly complicated forms with Man, again, the preordained result. These unknown laws of nature were often conceptualized as similar to those that govern the formation of crystals.

In Darwin's theory evolution was subject to the laws of nature, but there was no law or other cause that drove things in any particular direction. Organisms evolved the properties that they did because of historical contingencies, and Darwin's critics often equated such contingency with 'chance.' Attempting to find explanations for evolution in terms of 'law' rather than 'chance' has been a recurrent theme in efforts to find alternatives to Darwinism.

The term 'orthogenesis' is often applied to notions of evolution occurring along predetermined lines, especially when there is some sort of analogy with the formation of an embryo. Such quasi-evolutionary ideas began to receive attention from social scientists before the publication of the *Origin of Species* (Darwin 1859). The stage theory of Auguste Comte (1798–1857), with its theological, metaphysical, and positivistic levels, is one important example. Closer to what we moderns think of as evolution was the philosophy of Herbert Spencer (1820–1903). Spencer analogized both biological and socio–economic evolution with the embryology of Karl Ernst von Baer (1791–1876). In Spencer's 'synthetic philosophy,' the increasing specialization of the parts in developing organisms was analogized with the specialization and diversification that occurs in society.

The 'super-organism' view of social and biological evolution has been firmly rejected by Darwin's followers. However, a more limited form of 'holism' that affirms the reality and importance of supra-organismal groups has become increasingly popular in the philosophy of systematics. Ontologically, species and other taxonomic groups are now considered not to be classes but individuals in the broadest sense: particular, concrete things that have a definite position in space and time. Like individuals in the familiar sense of organisms, they can change indefinitely, they have no definitions, and there are no laws of nature for any one of them.

Species are now understood to be supra-organismal wholes that are held together by sex. Darwin however never had a satisfactory species concept. But he did emphasize the interaction between organisms and the reproductive competition among them as the fundamental cause of change. As he saw it, the competition for such resources as food is important only insofar as it affects reproductive success. The crucial role of sexual reproduction is underscored in Darwin's theory of sexual selection, a major topic in *The Descent of Man* (Darwin 1871). The 'secondary' sexual characters, such as gaudy plumage in birds and the antlers of stags, were interpreted by Darwin as the result of purely reproductive competition, based upon monopolizing mates, and unrelated to survival or procreation pure and simple.

Misunderstandings about competition have always been a major impediment to the assimilation of Darwin's theory. In economics, competition is thought of as a market condition that is antithetical to monopoly. The lay person, however, tends to confuse competition with other phenomena, such as physical combat, and to treat it as if it were the opposite of cooperation. For many, even teachers who grade on a curve, it is something antisocial, if not downright immoral. Of course political considerations have added to the confusion but it also reflects a deeper, metaphysical problem. The pre-Darwinian mentality has generally tended to assume that the world was created by God and that, in the words of Alexander Pope, 'Whatever is, is right.' Short of rejecting the basic premise, there are two popular alternatives. One is to endorse the conditions that exist. The other is to deny their existence. Various forms of Social Darwinism have opted for the former. More tender-minded social thinkers, from Prince Kropotkin to Stephen Jay Gould, have opted for the latter, portraying the natural economy as one in which competition is unimportant. Darwin of course rejected the basic premise. That perhaps bothered his contemporaries more than any other aspect of his theory. It allowed the phenomena of life to be explained in purely natural-science terms, without recourse to teleology in any but a trivial sense. Everything came into existence as a matter of historical contingency in the course of a vast sex contest.

Partly because of his experience with barnacles Darwin became increasingly aware of the importance and pervasiveness of sex. The barnacles, although generally hermaphroditic, turned out to crossfertilize. Turning his attention to plants, Darwin found out that they cross far more often than anybody had realized. Indeed, there were widespread adaptations furthering outcrossing. He undertook extensive empirical investigations, with striking results that revolutionized floral

biology. His first book after the *Origin of Species* (Darwin 1862) dealt with the pollination biology of orchids. It documented in great detail the often complex adaptations that enabled the orchids to have insects transfer pollen from one flower to another. These adaptations were often quite bizarre and the book can at least be read as a sort of satire on traditional natural theology. It also counteracted the morphological tradition that had tended to dismiss function. The book evoked a debate between Darwin and his supporter the Harvard botanist Asa Gray (1810–88). Darwin firmly resisted Gray's efforts to find God superintending the overall trend of evolution. In the orchid book and other botanical publications, Darwin showed that inbreeding could be highly deleterious and that plants have both anatomical and physiological adaptations that further outcrossing.

In *The Descent of Man, and Selection in Relation to Sex* (Darwin 1871), Darwin laid the foundation for a great deal of work on the comparative biology of reproduction. His theory of sexual selection followed from his basic idea that evolution results from reproductive competition between individuals. Such features as the antlers of stags and gaudy plumage in male birds he interpreted as adaptations used in monopolizing mates, and not as contributing to survival or otherwise giving an advantage in competing with organisms of other species.

Individualistic competition also has a great deal to do with another important topic in that book: the evolution of morals. Darwin was quite clear on the point that organisms generally act in their own interest. He even suggested that an adaptation in one species that existed solely for the good of another species could not be explained in terms of his theory. He likewise ruled out adaptations that would benefit a species as such. There was as he saw it no good evidence that such adaptations exist. However, he still had to account for the existence of self-sacrifice upon the part of social organisms, including apparently altruistic acts by human beings. His discussions of such topics include basically everything that was considered by late twentieth-century sociobiologists but without the genetics. He seems to have understood kin selection, insofar as he pointed out that in artificial selection traits could be favored by what Darwin called 'breeding from the same stock.' A relative of an animal with a trait such as flavorful meat could be selected for breeding instead of that particular animal. He seems not to have considered kin selection very important, however. Neither did he emphasize reciprocity, though he recognized its existence. Rather, he stressed familial selection, which was important in his treatment of the neuter casts of social insects such as ants and termites. He treated families as individuals, with the relevant competition going on between nests of ants or hives of bees. In dealing with human morality Darwin invoked selection at the level of the tribe or nation. His reasons for invoking such group selection, rather than, say, reciprocity, are not explicitly stated in his writings.

To what extent did Darwin go along with the 'social Darwinism' of his times? This term is by no means monolithic and merely saying that he endorsed some of it hardly answers the question. Social Darwinism was basically a political doctrine and it generally had a basis that was if not religious, at least teleological. Interfering with the struggle for existence would tend to prevent the desirable end result from happening. It would be like interfering with the growth and maturation of an organism. Darwin did not endorse such claims. However, he did maintain that free competition favored both biological and economic progress, the two being interrelated insofar as the human intellect is a source of innovation. Lessening the struggle would, he thought, tend to be dysgenic.

There was also a psychological aspect to Darwin's thinking about morals, one that often focused upon the 'moral sentiments,' as Adam Smith had called them. He tried to derive those sentiments from a feeling of sympathy. He considered morality highly serviceable in the struggle for existence, but the selective context was crucial. He suggested that in bees, in which the drones are killed by the workers, it would be appropriate for such fratricide to be a matter of conscience. Some of his readers were scandalized. Selection favoring instinctive behavior was important here, and it contrasts with much of the thinking of his contemporaries. Herbert Spencer had attributed moral sentiments to a more Lamarckian mechanism. There was a kind of racial memory of the sufferings of our ancestors. The British empiricist tradition had sought to explain ethics in terms of personal experience. Spencer made it racial experience. Darwin replaced experience with reproductive success.

The Descent of Man, and Selection in Relation to Sex (Darwin 1871) was followed by its sequel, *The Expression of the Emotions in Man and Animals* (Darwin 1872), which further elaborated upon Darwin's evolutionary psychology. Darwin rejected mind–body dualism and thought that the functioning of the nervous system and the rest of the body were very tightly linked. He used facial expression and other body movements as a means of gaining access to the inner workings of the nervous system. It also provided examples of evolutionary continuity. As he saw it, emotional expressions might come to function in communication, but in keeping with his general anti-teleological position they did not originate that way. Rather, they could be traced back to historical antecedents having a different functional significance. Darwin's interest in psychology can be traced back to his undergraduate years and was a major topic for his empirical research. Two features of this work need to be stressed at the outset. First, although he owed much to the extensive literature on such topics, with which he was quite familiar, much of his contribution came from rejecting the assumptions that had often been

made. For example, he rejected the (teleological) notion of Charles Bell that there are anatomical adaptations for expressing emotions. Second, his work in that area rested upon a solid basis in experience, including laboratory experiments as well as natural history observations, with much attention paid to the underlying physiological mechanisms. This at a time when psychology was just beginning to disassociate itself from philosophy and become a laboratory science.

Much of Darwin's behavioral research was done on plants. One naturally hesitates to call such work 'ethology' or 'physiological psychology,' but the lack of muscles and nerves in plants does not prevent them from moving and responding to stimuli. Bridging the gap between plants and animals was part of Darwin's effort to overturn traditional metaphysics. The book on insectivorous plants does more than just allude to that point. In his work on climbing plants Darwin found a morphological continuity between the organs used in climbing and less modified precursors. Tendrils were the modified petioles of leaves, for example. He also found that he could trace the circling movements with which the plants located objects upon which they climbed to a circling movement that had no such function. That led him to investigate the manner in which plants respond to directional stimuli. His research on that topic was first-rate plant physiology and gave rise to a vast literature. His follower George John Romanes (1848–94) extended that sort of reasoning to simple marine animals and thereby founded an important tradition in the study of animal behavior.

Darwin also concerned himself with the ontogeny of behavior, especially that of human beings. He kept detailed notes on the behavioral maturation of his own children. Some of this material was incorporated in *The Expression of the Emotions in Man and Animals* (Darwin 1872). A paper containing excerpts from the notes (Darwin 1877a) is considered one of the classics of child psychology. Darwin had several reasons for studying behavioral ontogeny. Not the least is the light it may cast on the role of inheritance vs. learning. If there has been no opportunity for learning, and behavior appears spontaneously in the course of maturation, then a strong case can be made for inheritance.

Another reason for studying the ontogeny of behavior was Darwin's basic position on the relation between development and evolution in general. For him, the modification of ontogeny was a fundamental evolutionary process, and he routinely approached evolutionary problems from the point of view of embryology. Darwin believed that the processes responsible for the development of a a fertilized egg into an adult are something that themselves evolve. This was in itself an important innovation. Perhaps more importantly he showed how natural selection could modify the structure and function of organisms by affecting the various stages of development. By implication, the history of life could be approached by asking how the underlying developmental processes had been modified. His great follower Fritz Müller (1821–97) developed Darwin's approach by suggesting that the history of certain crustaceans could be interpreted as the result of stages being added at the end of the life cycle. As a consequence, the most modified animals would repeat, during their life cycles, the series of adult ancestors. Another of Darwin's followers, Ernst Haeckel (1834–1919), took Müller's special case and treated it as the general rule, and enunciated what he called the fundamental biogenetic law, according to which ontogeny recapitulates phylogeny.

Darwin's views on the relationship between development and evolution are discussed at length in *The Variation of Animals and Plants under Domestication* (Darwin 1868). The variations that play an indispensable role in natural selection are here shown to occur in a more or less orderly fashion, evidently as a result of the modification of existing developmental mechanisms. Variation was thus physiologically constrained, and evolution by natural selection was by no means affected only by external conditions as is so often maintained in caricatures of Darwinism. Because this book contains toward its end an effort to explain heredity as well as development, the work is often misinterpreted as mainly about genetics. Darwin attempted to explain the inheritance of acquired characteristics (and much else as well) in terms of particles that control development. These particles could be affected in the somatic tissues, not just the germ. The particles interested some of Darwin's followers and, with considerable modification in the hands of Hugo De Vries and others, they evolved into the 'genes' of later authors.

3. *Darwin's Relevance for Today*

Darwin was a remarkably creative thinker, as attested by the surprise with which his discoveries were so often met. One reason for this was the sort of wide-ranging curiosity and erudition that has become increasingly difficult to emulate as knowledge becomes ever more specialized. He was a very good observer, but insisted (at least in private) that this was because his observations were guided by theory. On the other hand, he was also an empiricist, and did not sympathize with the 'deductive' approach as he called it of such authors as Herbert Spencer. He confronted his theories with a mass of evidence and modified them in the course of investigation.

Darwin was not constrained by academic fashion or pedantry in seeking material that could be put to scientific use. He unearthed a wealth of information in the writings of plant and animal breeders and appealed

to gardeners and others to provide additional materials. He also carried out extensive series of observations and experiments, often opening up new areas for investigation. We should mention his studies of competition between plants as pioneering contributions to experimental ecology. Darwin's experimental work is very highly regarded by plant physiologists and students of animal behavior. However, his originality so far as methodology is concerned has more to do with comparison than with experiment. In Darwin's hands, comparative biology became linked to the revolution in systematics, a discipline which was now based upon genealogical relationships and therefore fundamentally historical. Hypotheses could be developed and tested by such procedures as plotting data onto a phylogenetic tree. Phylogeny was especially important with respect to adaptation, which was now conceived of as the outcome of an historical process.

The shift to historical narrative is an aspect of the rejection of traditional, teleological modes of thinking in which states of adaptation had implied purposefulness, even conscious design. Although structures might give many of the appearances of having been constructed by an intelligent being, it was no longer legitimate to treat ours as the best of all possible worlds. By taking an historical approach and understanding the actual processes and events that have brought about the present properties of organisms we can avoid the teleology. Instead of asking what an organ is 'for' we ask what has brought it into being. It then becomes apparent that the world abounds in maladaptations and historical vestiges.

Contrary to what has sometimes been maintained, natural selection is not capable of explaining just about any conceivable property of organisms. Indeed the mechanism places strong restrictions upon the actual course of evolutionary change. For that very reason, it provides a powerful tool for evaluating hypotheses, especially where adaptation is involved. There is nothing that might reasonably be considered analogous to foresight when evolution occurs by natural selection. Any 'preparation' for future conditions of existence has to be the result of coincidence, and populations remain adjusted to present circumstances, and organisms repeat that which furthered their ancestors' reproductive success.

Because natural selection depends upon individualistic competition, there are strong constraints upon the kind of adaptation that can occur above the level of the organism. As already mentioned, Darwin clearly recognized that the individuals that reproduce differentially can be families as well as organisms, and on that basis he was able to explain the evolution of neuter castes in social insects. How readily that can happen has been a serious topic of controversy ever since. Darwin did not believe that differential reproduction of species could produce adaptations that exist 'for the good of the species,' though the possibility has subsequently been taken seriously by some authors. 'Group selection' was widely invoked in the early twentieth century. A critical attitude toward it did not become common until the latter half of that century, when in fact there would seem to have been something of an over-reaction. From the point of view of Darwin's contribution, however, what matters is not the ultimate answers that may emerge, but rather the manner of framing the question. One needs to ask precisely how the world is structured and how the components might interact, formulate explicit hypotheses and then test them.

Such issues of course are particularly important in the study of social behavior. The modern discussion of what sociobiologists call 'altruism' is a continuation of Darwin's pioneering work in that area. It is important, however, to recognize that 'altruism' in this sense is a technical term without ethical connotations and means self-sacrifice pure and simple. But people habitually slip back into older ways of thinking and tend to attach value judgments to statements of fact on the assumption that what is 'natural' is therefore good. Even where the teleology is repudiated explicitly it may persist in the form of implicit assumptions. It is clearly an over-reaction, however, to reject the study of adaptation in favor of so-called proximate causes. 'Teleology' in the limited sense of homing behavior and the capacity for foresight and understanding on the part of human beings and some other animals is by no means inconsistent with the expectations of selection theory.

Darwin recognized that natural selection is capable of accounting for the phenomena of progress. He reasoned that in a competitive economy innovations that are effective in dealing with the environment would tend to accumulate. On the other hand, he rejected the views of many of his predecessors and contemporaries whose conceptions of progress remained teleological. Notions about evolution, whether social, economic, or biological, based upon analogies with the growth of crystals or the development of embryos implied a kind of inevitability that was out of line with Darwin's theory. But they appealed to those who were looking for some kind of long-term goal to it all. Darwin witnessed a great deal of technological progress in his day, and like the overwhelming majority of scientists then and now, he obviously believed in scientific progress as well. Although the idea of progress has recently come under attack by philosophers and social scientists, this may just be another episode in the ongoing process of coming to grips with Darwin's contribution.

See also: Darwinism; Darwinism: Social; Emotion: History of the Concept; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionism, Including Social Darwinism; Malthus, Thomas Robert (1766–1834); Natural Selection;

Population Ecology; Sociobiology: Overview; Sociobiology: Philosophical Aspects; Spencer, Herbert (1820–1903)

Bibliography

Browne J 1995 *Charles Darwin: Voyaging: Volume One of a Biography*. Alfred A. Knopf, New York

Darwin C 1839 *Journal of Researches into the Geology and Natural History of the Various Countries Visited by H.M.S. Beagle, under the Command of Captain Fitzroy, R.N. From 1832 to 1836*. Henry Colburn, London

Darwin C 1842 *The Structure and Distribution of Coral Reefs*. Smith and Elder, London

Darwin C 1844 *Geological Observations on the Volcanic Islands Visited during the Voyage of H.M.S. Beagle, together with some Brief Notices of the Geology of Australia and the Cape of Good Hope*. Smith, Elder, London

Darwin C 1846 *Geological Observations on South America*. Smith, Elder, London

Darwin C 1851 *A Monograph on the Sub-class Cirripedia, with Figures of all the Species. The Lepadidae; or, Pedunculated Cirripedes*. Ray Society, London

Darwin C 1854 *A Monograph on the Sub-class Cirripedia, with Figures of all the Species. The Balanidae (or Sessile Cirripedes); the Verrucidae, Etc., Etc., Etc*. Ray Society, London

Darwin C 1859 *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London

Darwin C 1862 *On the Various Contrivances by which British and Foreign Orchids are Fertilised by Insects, and on the Good Effects of Intercrossing*. John Murray, London

Darwin C 1865 On the movement and habits of climbing plants. *Journal of the Linnean Society of London (Botany)* **9**: 1–118

Darwin C 1868 *The Variation of Animals and Plants under Domestication*. John Murray, London

Darwin C 1871 *The Descent of Man, and Selection in Relation to Sex*. John Murray, London

Darwin C 1872 *The Expression of the Emotions in Man and Animals*. John Murray, London

Darwin C 1875 *Insectivorous Plants*. John Murray, London

Darwin C 1876 *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom*. John Murray, London

Darwin C 1877a A biographical sketch of an infant. *Mind* **2**: 285–294

Darwin C 1877b *The Different Forms of Flowers on Plants of the Same Species*. John Murray, London

Darwin C 1880 *The Power of Movement in Plants*. John Murray, London

Darwin C 1881 *The Formation of Vegetable Mould, through the Action of Worms, with Observations on their Habits*. John Murray, London

Goldie P, Ghiselin M T 1996 *Darwin Multimedia CD-ROM*, 2nd edn. Lightbinders, San Francisco

Lyell C 1830–32 *Principles of Geology, being an Attempt to Explain the Former Changes of the Earth's Surface, by Reference to Causes now in Operation*. Murray, London

Malthus T R 1826 *An Essay on the Principle of Population*. Murray, London

M. T. Ghiselin

Darwinism

'Darwinism' can properly be dated from 1859, the year in which English biologist Charles Robert Darwin (1809–1882) published his magisterial evolutionary work: *On The Origin of Species by Means of Natural Selection*. In order to understand the magnitude of what Darwin achieved in that work, and the nature of the movement to which he gave his name, it is perhaps best to recognize that there are three different kinds of issue at stake here (Ayala 1985). First, there is what one might speak of as the very fact of evolution itself: the claim that all organisms, living and dead are the result of a developmental process governed by natural law. Then second, one has the question of paths of evolution, or as these are known technically, of phylogenies. Here one is concerned about the routes along which evolution proceeded. Third, there is the question of causes or mechanisms. The work of the early nineteenth century French biologist, Jean Baptiste de Lamarck (1809), who argued that organisms change thanks to the inheritance of acquired characteristics, speaks to this particular issue. Here one is concerned with the driving forces behind the evolutionary process.

Although Charles Darwin was not the first evolutionist—in the eighteenth century, at length in rhyming couplets, his own grandfather Erasmus Darwin had celebrated the joys of organic change—it was in the *Origin* that Darwin established once and for all the fact of evolution. He did this through making appeal to a unificatory argument or what he would have thought of as a 'consilience' (Ruse 1979). Darwin argued simply that if evolution were true, then one could explain facts in a wide range of areas. For instance, one can show why it is that the fossil record is roughly progressive. Likewise, one can show why it is that organisms are distributed as they are around the world. One can explain the facts of embryology, why it is that organisms are very different forms when adult (like the man and the dog) but have embryos which are almost identical. Why it is that the facts of anatomy are as they are? Why do we find skeletal isomorphisms (known technically as 'homologies') between the forelimbs of organisms, despite these limbs having different functions? The bat uses its forelimb for flying, the mole for digging, the whale for swimming, the man for grasping, and yet the bones are very similar. One can explain this, argued Darwin, through the fact of evolution. All of these phenomena, and many more, can be explained by evolution. Conversely, they point backwards up to the truth of the fact of evolution.

In the *Origin*, Darwin had relatively little to say about the paths of evolution, although it was he who tied evolution firmly to the metaphor of the tree of life. Unlike some early evolutionists, Darwin always stressed the extent to which all life is related and descends 'from one or a few forms,' branching as time unfurls to the vast diversity which exists today. But, paradoxi-

cally in the nineteenth century it was the non-evolutionists—particularly Lamarck's contemporary and fellow countryman Georges Cuvier, who was both the father of comparative anatomy and an accomplished paleontologist—who first started to work out the actual course of life's history (Ruse 1996). In such labors, Cuvier was backed by German-style biologists—particularly the Swiss-American ichthyologist Louis Agassiz—who saw analogies between life's history and the development of individuals. Darwin himself relied on both the fossil evidence and on embryology and its supposed analogies when around 1850 he performed a massive taxonomic study of barnacles. It is clear that although he did not draw explicit inferences about evolutionary paths, he had some strong convictions and suspicions (Richmond 1988). But none of these found their ways into the *Origin*, and later in the century it was Darwin's young supporters like Thomas Henry Huxley who were most interested in particular paths. In the USA there was much activity in this direction, thanks particularly to the magnificent fossil findings, especially those of the horse and of dinosaurs (Bowler 1996).

1. Natural Selection

It was when it came to causes or mechanisms that Darwin himself felt that he really was able to make significant scientific advances. He proposed his well-known mechanism of evolution through natural selection brought on by a struggle for existence, starting from the claims made by the English clergyman Thomas Robert Malthus (1826), that population numbers will always outstrip food supplies and space availability. Generalizing this to the animal and plant world, Darwin suggested that there will always be, at the same time, naturally occurring variation within populations: his studies of barnacles convinced him strongly that this was true. Darwin therefore concluded that in the struggle, success would be at least in part a function of the differences occurring naturally within organisms. There will thus be a natural form of selection, corresponding to the artificial selection practiced so successfully by animal and plant breeders in the first part of the nineteenth century in the UK.

But more than just this. Darwin believed that natural selection would lead to change of a particular kind, namely towards adaptive advantage. Darwin had read and been much impressed by the teleological arguments expressed by Archdeacon William Paley (1802), a work which Darwin had to read as an undergraduate. Paley argued that the eye is very much like a telescope, and that just as a telescope has a telescope maker, so the eye has an eye maker. That is to say the Great Optician in the Sky. Darwin rejected Paley's conclusions, but he completely accepted Paley's premises about the design-like, teleological nature of the organic world. It was Darwin's belief that natural selection speaks to this aspect of organic life, providing a naturalistic—that is to say law-bound—mechanism of explanation.

2. Huxley's Program

After the *Origin* was published, evolution as fact was never again seriously challenged as a matter of science. But the mechanism of which Darwin was so proud got relatively little attention. Darwin's great supporters, Huxley particularly, had other ends for evolution than a functioning professional science with selection at its core (which had surely been Charles Darwin's own aim). Huxley and his fellows were busy creating a secular society in Victorian times: one which would have open education for all without religious qualifications; one which would have a professionally trained civil service; not to mention other things like a military which did not simply depend on one's social status and personal wealth. With some good reason, the Huxleyites saw the established church as being very much part and parcel of the powers that they wished to overthrow. Hence, they were eager to find some alternative secular substitute (Ruse 1996).

Evolution fitted this need entirely. It was not so much that evolution was being pushed as something in direct violation of Genesis or anything like that. Rather, evolution could function as a kind of secular religion or ideology, on which one could hang all sorts of hopes of progress and ethical change, and the like. It is no surprise to find that, as the nineteenth century drew to an end, the evolutionist *par excellence* was less Charles Darwin, and more his fellow countryman Herbert Spencer (Richards 1987). The latter, in his massively written 'Synthetic Philosophy,' was providing a world picture, which would be all encompassing, right up to and including, our own species. Naturally, as a prospective religion, evolution was used to promote moral and social causes. Notoriously, Spencer and his supporters—prominent in this respect were the American businessmen John D. Rockefeller the First and Andrew Carnegie—endorsed a kind of *laissez faire*, arguing that as with the struggle for existence and survival and selection in nature, so also should there be such a struggle and selection in society. It is paradoxical that this movement became known as social Darwinism, for although there are certainly hints of such ideas in Darwin's own writings (especially in his work on our species, *The Descent of Man*, 1871), primarily this was a movement to which he contributed little and which altogether took evolutionism in directions he rather regretted.

3. The Synthetic Theory

Darwin's ideas as science only provided part of the picture. The whole question of mechanisms hung fire

rather until the twentieth century. Then came the development of an adequate theory of heredity explaining how and why it is that organic characteristics are transmitted from one generation to another, and why new variations appear occasionally, spontaneously as it were. When first this was done, many thought that the new theory of heredity, genetics, was in opposition to any kind of selection mechanism. But by the 1930s, a number of mathematically gifted biologists—notably R. A. Fisher (1930) and J. B. S. Haldane (1932) in the UK and Sewall Wright (1931, 1932) in the USA—showed that one can put together Darwinian selection with Mendelian genetics in a smooth synthesis (Provine 1971). Finally, an adequate mechanism for the evolutionary process was developed and accepted. However, much more work needed to be done. It fell to the next generation of evolutionists, notably people like Julian Huxley (1942) (the grandson of Thomas Henry Huxley) in the UK, and Theodosius Dobzhansky (1937), the Russian-born American geneticist and his associates to put the empirical flesh on the mathematical skeletons of the 'population geneticists.' Thus, around 1950, one had a full theory, which came to be known as the 'synthetic theory of evolution' or 'Neo-Darwinism': Darwin's theory of natural selection spliced in with Mendelian genetics, offering the way forward for a full and satisfying causal picture of evolutionary change.

4. *Sociobiology*

The years since around 1950 have been amongst the most exciting in evolution's history. Although no one now wants to challenge the ultimate picture, there have been both extensions and challenges. Most notably the former has involved the pushing of a Darwinian selective approach into the realm of social behavior. Darwin himself always thought that such behavior was part and parcel of the evolutionary picture, but he recognized that it was extremely difficult to study. However, by the 1960s evolutionists felt sufficiently confident to extend their work to such organisms as the Hymenoptera (the ants, the bees, and the wasps), as well as to mammals and even up to and including our own species. This development of the evolution of social behavior from a Darwinian perspective—so-called 'sociobiology'—has been highly controversial. The leaders in the field, people like William Hamilton (1964) and John Maynard Smith (1982) in the UK, and Edward O. Wilson (1975) in the USA, not to mention the great popularizer Richard Dawkins, author of *The Selfish Gene* (1976), have been subject to a great deal of critical attention. But now I think it is generally realized that one can profitably take a Darwinian selection-based approach into the social realm, dissecting and showing in great detail how the behavior of organisms is just as much a function of their evolutionary past as is their physical nature.

Still highly controversial has been some of the work by human sociobiologists, trying to take a selection-based approach to human nature and behavior. By way of example, one might mention the Canadian researchers Martin Daly and Margo Wilson (1988). They show that, in many respects, family patterns are tightly controlled by evolutionary considerations. Why it is for instance that violence in families almost always comes from the hands of men: usually not by biological fathers, but by stepfathers. Remarkably, Daly and Wilson show that stepfathers are 100 times more likely to murder their children than are biological fathers. This they relate to straightforward Darwinian principles, namely that from a biological point of view, survival and reproduction may be improved if one can clear the way (as it were) for one's own offspring, but that damaging or harming one's own biological offspring is to go completely against the whole Darwinian evolutionary process.

5. *Critics of Adaptation*

On the other side, there are the critics of adaptationism, most notably the American paleontologist Stephen Jay Gould. Like Richard Dawkins, a very popular writer, Gould has put forward his theory of so-called 'punctuated equilibria,' arguing that if we look at evolution over long time scales, we see that selection as a force for constant, unbroken change falls away dramatically (Eldredge and Gould 1972, Gould 1982). Rather, what we get are long periods of inactivity (so-called 'stasis') broken or punctuated by very rapid changes from one form or blueprint (Gould likes the German term *Bauplan*) to another. Gould argues that although selection certainly perfects organic characteristics, tuning them up as it were, by and large the basic patterns of organisms are part of a general morphologically constrained non-selective picture. A vertebrate having four limbs, for instance, rather than six like many invertebrates, has nothing whatsoever to do with natural selection.

There are many other developments going on, particularly those involving molecular biology. For instance, it is suggested that perhaps at the molecular level, beneath the level at which selection operates, one gets a whole different dimension to evolution: perhaps a dimension where random factors are very significant. (This is the theory of 'molecular drift,' Kimura 1983.) Perhaps also in other spheres there are major factors yet to be discovered: possibly viruses carry units of heredity ('genes') from one organism to another of a very different kind. Also, there is much work going on at the moment of related questions such as the origin of life. Certainly, no one has a firm solution in this direction, although once again major advances have been made and we know a great deal more than we did in the 1980s or 1990s.

6. *The Meanings of 'Darwinism'*

So, in conclusion, trying to tie together strands from history and from the present, how then are we to define and characterize 'Darwinism,' or a 'Darwinian' meaning someone who accepts the doctrine? Darwinism certainly commits one unambiguously to the fact of evolution. The only people today who challenge this seriously are those driven by religious commitments namely the extreme Christian evangelicals known as 'fundamentalists,' finding that evolution clashes with their prior commitments to the Bible (especially Genesis) taken literally (Johnson 1991). There are some Darwinians who think that their position also demands of you, belief in some kind of philosophical materialism, claiming that any kind of supernatural belief is false or impossible (Dennett 1995). However, this is surely not a universal view, and there are many (and not all believers) who think that although evolution holds exclusively in the scientific realm, there may be more to life—certainly there may be more to the after-life—than strict science allows (Ruse 2001).

As far as evolution as a path is concerned, although it is true that one does not have to be a Darwinian to believe in a tree of life, it is still surely true that Darwinism itself commits one to some kind of tree of life. But Darwinism does not specify the details as such. The much-debated question, for instance, as to whether the birds are descended from the dinosaurs or whether from reptiles that were not dinosaurs has no bearing on Darwinism as such. What about the above-mentioned speculations that life's history may have involved such things as the moving of genes from one group to another very different through the agency of viruses and the like? Hence the tree of life might be a bit more net-like than one would have suspected. Certainly this is something which was not envisioned by Darwin himself or indeed by many successors who would have called themselves Darwinians. But I doubt that there would be much concern or felt need to repudiate Darwinism. After all, things like this are more extensions than major denials of the traditional picture.

Evolution as cause is as always the crucial issue. Darwinism has and continues to make natural selection absolutely central. If one does not accept that natural selection is the major cause of evolutionary change and that adaptation is the defining characteristic of organisms, then one is no Darwinian. This applies to humans as well. Interestingly, the very first speculations we have of Darwin about natural selection was a discussion about human evolution and how selection leads to bigger brains and greater intelligence. As far as other points are concerned, then perhaps as with paths there is more flexibility. Darwin himself always stressed that selection works for the benefit of the individual over the group. This is very much the stance of sociobiologists in the twenty-first century, and is indeed the basis of Dawkins's metaphor of the 'selfish gene.' But there have always been those who have yearnings towards a group based form of selection—one of the first indeed was the co-discoverer of natural selection, Alfred Russel Wallace. Perhaps therefore with something like this there is room for difference within the fold (Sober and Wilson 1997).

It was never the claim of Darwin or of any subsequent Darwinian that natural selection is the only mechanism of evolutionary change. Darwin himself accepted Lamarckism. This mechanism is today totally discredited, but there are other putative causes of change, and one can certainly allow them some force and still be a Darwinian. There are many debates about how much of an allowance one can give and still is subscribing to Darwinism. To a certain extent it all depends on how far one is modifying traditional views and how far one is standing outside of, or beyond them. Controversial is the question of 'genetic drift,' the claim by Sewall Wright that evolution can go by chance, even when selection is operating (Coyne et al. 1997). Really strict ultra-Darwinians like Fisher would have very little tolerance for something like drift, thinking it unimportant at best. Others (including most of the American founders of the synthetic theory) found some drift to be quite compatible with their Darwinism. Kimura himself felt (perhaps with good reason) that he was a good Darwinian because he did not propose drift at the physical level but at the molecular level. Darwin himself had said nothing about molecules and so in proposing molecular drift there was no reason to think that one was being non-Darwinian.

7. *Social Darwinism*

Finally, what about social Darwinism? Few want to use that label given its unfortunate connotations with such things as the ruthless business practices of the last century. But still there is a thriving industry among those who want to derive more from evolutionism than mere science (Ruse 1999). Julian Huxley (1927) was one person in this century who wanted his evolution to be the basis of a 'religion without revelation' and more recently Edward O. Wilson (1998) is another. In one of his prize-winning books, *On Human Nature* (1978), Wilson is explicit in his seeing evolution as more than mere science, as a 'myth,' which will absorb or push aside conventional religion:

> 'Make no mistake about the power of scientific materialism. It presents the human mind with an alternative mythology that until now has always, point for point in zones of conflict, defeated traditional religion. Its narrative form is the epic: the evolution of the universe from the big bang of 15 billion years ago through the origin of the elements and celestial bodies to the beginnings of life on earth.'

And so on and so forth up to humans. Little wonder that:

> 'Theology is not likely to survive as an independent intellectual discipline.'

Wilson even finds place for an evolution-based morality—not the *laissez faire* of yesterday but a call to biodiversity and species preservation, else we humans decline and fall in a word impoverished of living nature.

It is certainly not necessary for a scientific Darwinian to be a social or philosophical Darwinian, but a remarkably large number are. The important conclusion is that however defined, Darwinism has a rich tradition, and although it has never been without its critics it continues in the twenty-first century as an important and ever developing testament to human reason and creativity.

Bibliography

Ayala F J 1985 The theory of evolution: Recent successes and challenges. In: McMullin E (ed.) *Evolution and Creation*. University of Notre Dame Press, Notre Dame, France, pp. 59–90

Bowler P J 1996 *Life's Splendid Drama*. University of Chicago Press, Chicago

Coyne J A, Barton N H, Turrelli M 1997 Perspectives: A critique of Sewall Wright's shifting balance theory of evolution. *Evolution* **51**: 643–71

Daly M, Wilson M 1988 *Homicide*. De Gruyter, New York

Darwin C 1859 *On the Origin of Species*. John Murray, London

Darwin C 1871 *The Descent of Man*. John Murray, London

Dawkins R 1989 *The Selfish Gene*, 2nd edn. Oxford University Press, Oxford, UK

Dennett D C 1995 *Darwin's Dangerous Idea*. Simon and Schuster, New York

Dobzhansky T 1937 *Genetics and the Origin of Species*. Columbia University Press, New York

Eldredge N, Gould S J 1972 Punctuated equilibria: An alternative to phyletic gradualism. In: Schopf T J M (ed.) *Models in Paleobiology*. Freeman, Cooper, San Francisco, pp. 82–115

Fisher R A 1930 *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford, UK

Gould S J 1982 Darwinism and the expansion of evolutionary theory. *Science* **216**: 380–7

Haldane J B S 1932 *The Causes of Evolution*. Cornell University Press, New York

Hamilton W D 1964 The genetical evolution of social behaviour. *Journal of Theoretical Biology* **7**: 1–32

Huxley J S 1927 *Religion without Revelation*. Ernest Benn, London

Huxley J S 1942 *Evolution: The Modern Synthesis*. Allen and Unwin, London

Johnson P E 1991 *Darwin on Trial*. Regnery Gateway, Washington, DC

Kimura M 1983 *Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK

Lamarck J B 1809 *Philosophie zoologique*. Dentu, Paris

Malthus T R 1826 *An Essay on the Principle of Population*, 6th edn. John Murray, London

Maynard Smith J 1982 *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK

Paley W [1802]1819 *Natural Theology (Collected Works: IV)*. Rivington, London

Provine W B 1971 *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago

Richards R J 1987 *Darwin and the Emergence of Evolutionary Theories of Mind and Behavior*. University of Chicago Press, Chicago

Richmond M 1988 Darwin's study of the Cirripedia. In: *The Correspondence of Charles Darwin*. Cambridge University Press, Cambridge, UK, vol. 4, pp. 388–409

Ruse M 1979 *The Darwinian Revolution: Science Red in Tooth and Claw*. University of Chicago Press, Chicago

Ruse M 1996 *Monad to Man: The Concept of Progress in Evolutionary Biology*. Harvard University Press, Cambridge, MA

Ruse M 1999 *Mystery of Mysteries: Is Evolution a Social Construction?* Harvard University Press, Cambridge, MA

Ruse M 2001 *Can a Darwinian be a Christian? The Relationship between Science and Religion*. Cambridge University Press, Cambridge, UK

Sober E, Wilson D S 1997 *Unto Others: The Evolution of Altruism*. Harvard University Press, Cambridge, MA

Wilson E O 1975 *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA

Wilson E O 1978 *On Human Nature*. Cambridge University Press, Cambridge, MA

Wilson E O 1998 *Consilience*. Knopf, New York

Wright S 1931 Evolution in Mendelian populations. *Genetics* **16**(2): 97–159

Wright S 1932 The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics* **1**: 356–66

M. Ruse

Darwinism: Social

The term 'social Darwinism' was coined in Europe in the 1880s and rapidly spread through the UK and the USA. It negatively refers to the theories and doctrines which turn the social laws into the aftermath of the laws of nature.

1. The Term 'Social Darwinism'

The first persons ever to use the term—among others, Emile Gautier in France and Achille Loria in Italy—criticized more the Spencerian philosophy of the 'struggle for life' and the 'survival of the fittest' than Darwin's theory of natural selection. Emile Gautier explains that the selection law does not apply to human societies. He describes this extrapolation as 'social Darwinism' and contrasts it with the true social Darwinism which implies reforming society. All this initial confusion has still not been entirely clarified, as is evident with eugenics, a notion mingled with social

Darwinism while being distinct from it. Coined by Francis Galton, Charles Darwin's cousin, and derived from the Greek words *eu* meaning good and *genos* meaning stock or race, the term refers to

> 'the science of improving stock, which is by no means confined to questions of judicious mating but which, especially in the case of man, takes cognisance of all influences that tend in however remote a degree to give to the more suitable races or strains of blood a better chance of prevailing speedily over the less suitable than they otherwise would have had (Galton 1883, 1951, p. 17)

The starting point of eugenics is that natural selection no longer plays its role in modern society, thus leading to the degeneracy of the human species. Therefore an artificial selection, fully recognized, must be substituted for the failing natural selection. The means that are advocated either refer to 'negative' eugenics or 'positive' eugenics. 'Negative' eugenics aims at preventing people with hereditary diseases and serious deformities from getting married while restoring to means ranging from segregation to sterilization. 'Positive' eugenics aims at encouraging 'gifted' or 'strong' individuals, that is to say the carriers of favorable physical and intellectual characteristics, to reproduce themselves as much as possible.

2. *Was Darwin a Social Darwinist?*

While social Darwinists use the term 'struggle' in the literal sense and make of the survival of the fittest its legitimate sanction, Darwin points out that he used the term 'struggle for life' as a metaphor including different levels of opposition and dependency. The word 'struggle,' borrowed from the geologist Lyell, means making efforts in difficult conditions and at the same time facing an opponent. In Darwin the 'struggle for life' is not defined as a direct clash involving individuals but as a set of solidarity mechanisms and clashes precariously balanced, and taking into account the milieu in which individuals and species live and fight. Darwin only used the term 'survival of the fittest' as late as the fifth edition of *The Origin of Species* in order to refer to the process of natural selection and not the struggle for life. Then in *The Descent of Man* (1871), he derived the emergence of morality in humans from animal social instincts. But at the same time, ambiguously enough, Darwin agrees with Galton's eugenic ideas and claims that the development of social instincts should not eradicate competition and the struggle for existence.

3. *Unity and Diversity of Social Darwinism*

Social Darwinism can be defined as an ideology, a philosophy of history or a sociology which not only postulates an analogy between the laws of nature and the laws of society but also affirms that the laws of nature are the 'survival of the fittest,' the 'struggle for life,' and the laws of heredity and that there should be no opposition to this for the welfare of society. At the risk of reducing the great diversity of its forms, historians agree to distinguish two major types of 'social Darwinism':

(a) a liberal and individualistic social Darwinism which makes the struggle between individuals the fundamental law of society;

(b) a selectionist and interventionist social Darwinism which can be defined as a sociology of struggle and emphasizes the struggle between races.

3.1 *Liberal Social Darwinism*

The first social Darwinism, that of Herbert Spencer (1820–1903) appeared before *The Origin of Species* and *The Descent of Man*. It can be defined as an evolutionarist philosophy without any connection to biology. Spencer's concept of the 'survival of the fittest' combines natural selection and economic competition between individuals. It does not deal with struggles between groups and races.

Whereas Spencer, who was not a Darwinist, referred to the population laws, he did not draw the same conclusions as Malthus: the survivors are those who are endowed with a capacity of self-preservation superior to those who could not participate in life's banquet. Economic competition is thus part of the laws of nature and nothing should hamper them because the survival of the fittest is the only process capable of ensuring human progress.

The population's growth not only stimulated competition but also is the driving force behind civilization:

> 'This constant increase of people beyond the means of subsistence, causes, then, a never-ceasing requirement for skill, intelligence and self-control—involves, therefore, a constant exercise of these and gradual growth of them' (Spencer 1865–7, pp. 498–99)

The demographic pressure is meant to come to an end when its task has been carried out, bringing the process of civilization to its end. That was not the case in the last third of the nineteenth century.

Even if the changes in English society and increasing state intervention did not correspond to his expectation, Spencer wrote again about the notion of the survival of the fittest in the *Principles of Sociology*, published from 1876 to 1880. from then he distinguished between two types of modern societies—military societies characterized by an increasing part played by the state, the growing number of civil servants, and faith in authority, and those he thought to be superior forms of society—so-called industrial societies—characterized by the decline of the state and the blooming of personal freedom through the play of competition between individuals. Liberal and indi-

vidualistic social Darwinism mainly spread through the Anglo-Saxon countries. In the USA the economist William Graham Sumner, the major theoretician of this type of social Darwinism explained that the differences between individuals merely reflect natural inequalities: people are not equal in the state of nature, and this should be taken in account.

Sumner was opposed to any type of state intervention, protectionism, and socialism whose goal is to save the individuals of any society from the difficulties or hardships of the struggle for life and vital competition through state intervention.

This type of social Darwinism, which was very popular at the time when liberal ideas gained support in the Anglo-Saxon countries, proved less successful on the European continent. Opponents did not fail to underline the contradictions and weakness of liberal social Darwinism. In the 1890s sociologist Lester Ward explained that economic competition did not result in individual advancement but in big new corporations whose power might need to be controlled and he criticized the social Darwinists' 'fundamental error' that 'the favors of the world are distributed entirely according to merit.' This first type of social Darwinism should be distinguished from that of the end of the nineteenth century which is based on a sociology of struggle between groups, collectivities, and races in order to call for state intervention and the control of human procreation.

3.2 *Selectionist Social Darwinism and Eugenics*

At the end of the nineteenth century this form of social Darwinism, especially in the UK and in the USA has been understood as a eulogy of colonization and imperialism inasmuch as the strongest nations were supposed to be the fittest ones.

Ludwig Gumplowicz developed the concept of race struggle: 'Under the term race struggle we shall designate the struggles of the most varied and heterogeneous ethnic and social units, groups and communities. They are the essence of the historical process' (Gumplowicz 1893, p. 193). The social Darwinism is also mixed up with eugenics because of similarities and converging views between the two ideas where the postulate that races and individuals are unequal or the role of heredity are concerned.

In the preface to the first translation into French of *On the Origin of Species* Clémence Royer justified the natural inequalities between people and was opposed to any restriction to competition: 'Men are unequal by nature: this is where we should start from …' (Royer 1862). Georges Vacher de Lapouge, the theoretician who first introduced eugenics in France, underlined the hereditary character of inequity:

> There are superior peoples and inferior peoples, and among these peoples there are elements more perfect and more imperfect than others. Unequal the classes, unequal the families. Inequality is everywhere and so is differentiation. Individuals are not only unequal but their inequality is hereditary, classes and nations and races are not only unequal but each of them cannot undergo absolute improvement (Vacher de Lapouge 1888).

Vacher de Lapouge explained that in modern societies natural selection no longer played its role and was replaced by what he called *social selections* leading to the eradication of the fittest and which he qualified as military, political, religious, moral, economical, professional, and urban selection. Consequently an artificial selection must be substituted for it so as to prevent society and civilization from being weighed down by the degenerate and inferior people. He therefore advocated *selectionism* as the only remedy for 'senile democracy' and combined it with socialism.

The German Otto Ammon, a theoretician of anthroposociology like Vacher de Lapouge, developed a less pessimistic vision of the evolution of society inasmuch as the prevailing social order expressed and reproduced the natural inequalities between people. If, due to 'defects associated with an exclusively intellectual culture and a sedentary life,' the superior classes are threatened with eradication for the benefit of the inferior classes, the latter can better themselves through education and food, thus contributing to the renewal of the superior classes: The existence of social classes promotes the working of natural selection, restricts also 'panmixia', i.e., the pairing of individuals without previous selection, and favors a much more frequent production of highly endowed individuals (Ammon 1895).

Karl Pearson was the major advocate for selectionist social Darwinism in the UK. Though he was a socialist, he was opposed to revolution and he was in favor of gradual reforms allowing society to strengthen and face international competition. In his eyes, eugenics and socialism could be separated—natural selection must give way to artificial selection in order to prevent the unfit from replacing the fit ones. But all these similarities cannot make one forget that many eugenicists were not Darwinist but rather neo-Lamarckians.

4. *Social Darwinism Today*

The decline of social Darwinism, which was already felt after World War I, has on the one hand been ascribed to the use that has been made of it, especially to justify imperialism, and on the other hand to the atrocities committed by the Nazis in the name of eugenics (mass extermination, euthanasia).

Even if nowadays social Darwinism is not explicitly referred to any more, it is still part of the intellectual scene as the rise of sociobiology shows. The latter is a scientific discipline founded by the zoologist Edward O. Wilson and gave rise to serious ideological controversies in the late 1970s and early 1980s. Wilson

(1975, p. 4) defined it as 'the systematic study of the biological basis of all social behavior.'

The biological basis is provided by the modern theory of evolution as it has developed from Darwin's work and contemporary genetics. Social behavior is the result of evolution—it has been selected and is therefore determined. In a sense sociobiology is a form of biological determinism without being a form of reductionism at the same time since Wilson contends that humans are the product of the joint influences of heredity and environment. To Wilson *biological* only meant that there are genetic predispositions contributing to the shaping of social behavior. To his opponents *biological* meant that social behavior was *biologically determined*. In the wake of the question of animal social instincts dealt with by Darwin in *The Descent of Man*, sociobiology strove to shed new light on the problem of altruism.

While natural selection is a selfish process which endows certain individuals with a superior capacity of reproduction, how can one explain the kind of altruism which is noticed among social insects and how can natural selection favor deeds harmful to those who do them? The answer given by sociobiology is that altruistic deeds toward relatives could favor the survival and the dissemination of the altruists' genes which the parents have in common with them. In the same way, the mother instinct and the care given to the child are explained by social biologists as being the best possible way of arranging gene transmission from one generation to the other.

See also: Darwin, Charles Robert (1809–82); Darwinism; Eugenics as an International Movement; Eugenics, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionary Social Psychology; Evolutionism, Including Social Darwinism; Spencer, Herbert (1820–1903)

Bibliography

Ammon O 1895 *Die Gesellschaftsordung und ihre natürlichen Grundlagen*. Fischer, Jena, Germany

Bagehot W 1867 *Physics and Politics*. Longmans, Green, London

Bernardini J-M 1997 *Le darwinisme social en France (1859–1918). Fascination et rejet d'une idéologie*. CNRS, Paris

Clark L 1984 *Social Darwinism in France*. University of Alabama Press, AL

Darwin C 1859 *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. Murray, London

Darwin C 1871 *The Descent of Man in Relation to Sex*. Murray, London

Drouard A 1999 *L'eugénisme en questions L'exemple de l'eugénisme 'français'*. Ellipses, Paris

Hofstadter R 1983 *Social Darwinism in American Thought*. University of Pennsylvania Press, Philadelphia, PA

Galton F 1883, 1951 *Inquiries into Human Faculty and its Development*. Macmillan, London

Gautier E 1880 *Le darwinisme social*. Dervaux, Paris

Gumplowicz L 1893 *La lutte des races Recherches sociologiques*. Librairie Guillaumin, Paris

Novicow J 1910 *La critique du darwinisme social*. Alcan, Paris

Pearson K 1912 *Darwinism, Medical Progress and Eugenics*. Dulau, London

Royer C 1862 *Préface de la traduction de L'Origine des espèces*. Guillaumin, Paris

Spencer H 1865–7 *The Principles of Biology*. Williams and Norgate, London

Spencer H 1868 *Social Statics or the Conditions Essential to Human Happiness Specified and the First of Them Developed*. Williams and Norgate, London

Spencer H 1876 *Principles of Sociology*. Williams and Norgate, London

Sumner W G 1914 *The Challenge of Facts*. Yale University Press, New Haven, CT

Vacher de Lapouge G 1888 De l'inégalité parmi les hommes. *Revue d'anthropologie* **3**: 11

Vacher de Lapouge G 1896 *Les sélections sociales*. Fontemoing, Paris

Ward L 1907 The establishment of sociology. *American Journal of Sociology* **12**: 581–87

Wilson E O 1975 *Sociobiology The New Synthesis*. Harvard University Press, Cambridge, MA

A. Drouard

Data Archives: International

Data archives are professional institutions for the acquisition, preparation, preservation, and dissemination of social and behavioral data. These data commonly are stored as electronic records. Evolving in the 1950s with the growth of an ethic of data sharing among social scientists, they have enabled thousands of researchers to undertake a wide variety of analyses of data for which they were not the primary data collectors. This leverages by many times, the original investments in data collection. Many archives also provide training and data services, including statistical analysis and consultation.

1. The Contents of Data Archives

Most data archives offer numerically coded data, but those data are not necessarily quantitative. Numbers may represent either quantitative measurements, such as dollars or years of school completed, or merely serve to label categories, such as sex and occupation. Textual, sound, visual, geographic information systems (GIS) data, and physiological data are also now found in some archives. Two main kinds of data are offered.

Microdata are measurements taken on many distinct units of observation, such as persons, families, companies, governmental units, transactions, events,

conflicts, and dyadic and group relationships. Each unit of observation, for example, a particular person, generates one or more 'records,' each of which contains measurements on a body of variables pertaining to that person, such as occupation or income. It is common for data records to contain hundreds of variables, some measured directly by observation and some derived from other measurements (such as the prestige score of the respondent's occupation). Microdata are based typically upon censuses and sample surveys (see *Statistical Systems: Censuses of Population*), although direct observation, administrative records, news articles, regulatory instruments, and experiments also are used and occasionally are combined with survey data. Microdata generally are rendered anonymous by removing both direct identifiers (names and addresses) and indirect identifiers that might permit sound guesses as to the identity of a respondent, whether a person or an organization.

Macrodata are aggregated microdata measurements in the form of matrices, tables, and charts. They are the result of statistical operations, such as a count of the persons of a certain ethnic heritage in a geographical area or the sum of an entire industry's revenue. This form of data presentation is used widely for the reporting of official data at fine levels of geographic identification, as it permits concealing the identities of individual respondents. However, because it is not possible to disaggregate macrodata and perform a different analysis on the underlying microdata, macrodata provides considerably less analytical power to the researcher. Macrodata may be merged with microdata to facilitate contextual analyses.

The archives also contain the documentation that is required to use these data accurately. Most important are 'codebooks,' documents that give conceptual meaning to numerical codes and describe the process of measurement in detail. Increasingly, electronic codebooks are linkable by software to statistical analysis programs such as SPSS, SAS, and STATA, relieving the user of programming tasks that are intellectually arid but fraught with potential error.

Archives make their greatest investments in the preparation of data and its documentation in order that analyses can be undertaken later, even much later, without personal consultations with the original data collector. This preparation involves identifying and resolving undocumented codes, correcting technical errors in data entry, adding substantive information about measurement processes and response categories, and preparing full sampling descriptions, all in consultation with the original data collector. Major changes in format may be made for the convenience of data users.

As the size of archival collections grew, it became essential to create finding aids of various kinds (e.g., catalogs and indices), now on the World Wide Web for interactive access. In recent years, the catalogs of some of the data archives worldwide have been made interoperable, enabling an interactive search of the holdings of several data archives in one operation. Some archives' collections also include ancillary materials, such as detailed records of fieldwork, basic frequency distributions, statistical analyses of the data, methodological notes, and publications based on the data.

As data archives developed, a specialized profession of data librarian/archivist emerged, blending knowledge and skills from information technology, librarianship, statistics, and the social sciences. These professionals staff most national and many campus data archives. The International Association for Social Science Information Service and Technology (IASSIST) is their central professional organization. Among the organization's goals are the promotion of archiving and the development of standards for the preservation of data.

2. *The Establishment of Data Archives*

The intellectual leaders of what came to be known as the 'data archives movement' came from both the academy and the private sector. From the academy came the founder of modern comparative politics, Stein Rokkan (Norway); two of the pioneers of the empirical study of political behavior (see *Databases, Core: Political Science and Political Behavior*), Philip E. Converse and Warren E. Miller (United States); and a leader of comparative research on social structure (see *Human–Environment Relationship: Comparative Case Studies*), Erwin K. Scheuch (West Germany). Converse and Miller sought to share with their colleagues the election data they were collecting and also to use the new model of the National Election Studies to foster the collection of data on the re-emerging European democracies. Rokkan and Scheuch sought to describe and understand the movement toward a united Europe. All recognized the great potential of data archives to facilitate comparative and time-series (see *Time Series: General*) research as well as to aid the social and behavioral sciences to become more transparent and self-correcting.

The private sector saw its polling data as integral to the functioning of modern democracies and held that understanding public opinion was crucial for a free society. One of the founders of modern polling (see *Polling*), Elmo B. Roper, organized the first of the social science data archives before the 1950s when he decided to share with social scientists and others the data from his polls. These collections of survey data were under development by 1947 at Williams College (Massachusetts), based upon holdings that Roper had been accumulating for 10 years. Williams College formally established the Roper Center for Public Opinion Research in 1957. Roper was joined in

contributions by two other leaders of public opinion polling in the US, George H. Gallup and Archibald Crossley, and by many of the other leading survey firms.

A succession of archives followed. The Central Archive for Empirical Social Research (ZA), based at the University of Cologne in the former West Germany, was next in 1960, followed in 1962 by the Steinmetz Archive in Amsterdam, the Netherlands, and the Inter-university Consortium for Political Research (ICPR), at the University of Michigan, Ann Arbor. By the beginning of the 1970s several more European countries had data archives. By 1964 the International Social Science Council (ISSC) had sponsored a second conference on Social Science Data Archives and had a standing Committee on Social Science Data, both of which stimulated the data archives movement. By the beginning of the twenty-first century, most developed countries and some developing countries had organized formal and well-functioning national data archives. In addition, college and university campuses often have 'data libraries' that make data available to their faculty, staff, and students; most of these bear minimal archival responsibility, relying for that function on a national institution.

Data archives usually are administratively housed in research universities or in government science or information agencies. Funding for these operations varies widely in scale and stability. The financially largest of the data archives, the Inter-university Consortium for Political and Social Research (formerly ICPR), had in 2000 an annual budget of almost $7 million that was about equally divided between (a) income from grants and contracts and (b) income from tuition and annual dues paid by more than 350 institutions. Many archives have both long-term governmental support and income from user fees; others subsist on small annual grants or support from individual universities. Since 1977 the archives have been organized loosely under the International Federation of Data Organizations (IFDO), which is an associate member of the ISSC. IFDO provides for cooperation among the archives and has been of material assistance in assisting nascent archives, particularly by providing technical and organizational assistance to new archivists in developing countries. European archives are organized more closely under the Council of European Social Science Data Archives (CESSDA), which has provided a basis for cooperative planning and execution of projects, technical assistance, and training. There is no comparable North American organization, although in the 1960s there was the Council of Social Science Data Archives. The archives are largely national in funding, organization, and substantive scope; there is no single international archive of social science data to which all social and behavioral scientists can turn, nor is there as yet a complete 'union catalog.' There was discussion in the late 1950s of organizing central data archives for all of Europe, but the present nationally based system emerged in its stead.

Because archives' holdings have now accumulated to encompass decades of observation in some societies, it may be anticipated that future historians will be among the heaviest users of data archives. So also will sociologists and political scientists involved in comparative and time-series research. Urban anthropologists and humanists are finding in data archives information about values, knowledge, and mass culture that was previously unavailable to them. There is also a substantial level of usage by journalists, policy analysts, and citizens. Usage can be expected to increase substantially as broader categories of researchers realize the value of archived resources and as technological progress eases access to data.

3. *The Ethic of Data Sharing and its Benefits to Science*

Both natural and social scientists have long recognized the value of sharing data with their contemporaries and of preserving data for their successors. Records of astronomical observations have been kept for several thousand years. During the 1957 International Geophysical Year, the natural sciences established a system of World Data Centers that still exists today. Historical parish, census, and tax records, although not originally designed as scientific data collections, were nevertheless a foundation for the development of actuarial science and demography in the eighteenth century. From the nineteenth century onwards, cultural anthropologists carefully documented and preserved their observations, some of which were incorporated into the Human Relations Area Files (see *Databases, Core: Anthropology and Human Relations Area Files (HRAF)*.

Data archives have made much possible: using their resources social scientists ask diverse research questions that go well beyond those posed by the original data collectors, address new questions to old data, apply newly developed statistical techniques to data gathered long before the methods were developed, construct time series and repeated cross-sections, and perform meta-analyses. Investigators have used data archives in replication research to correct errors and omissions in the published literature and even to ferret out fraud. Archived data have been the foundation for thousands of monographs, research articles, theses, dissertations, and reports. The agencies that support research have benefited because the archives have enabled researchers to avoid needless duplication of data collection. As social surveys are relatively expensive by the modest standards of the social sciences, many social scientists have been enabled to do research

that would otherwise have been impossible. Without the archives, education at the graduate level would be seriously deficient, for it is rare that a graduate student can afford to undertake primary data collection on a large scale; for at least 40 years, analytical theses and dissertations have been based upon archived data. Undergraduate training in the social sciences has increasingly come to resemble the laboratory-oriented training of the natural sciences, with archived data serving in place of biological specimens and chemicals.

Underlying this system of data archives is the ethic of data sharing: scholars should share with other scholars a crucial part of their intellectual property, the data that they have collected. Increasingly, this ethic is codified in formal rules, such as the requirement at the US National Science Foundation that publicly funded data be made available in a public archive in due time after the end of data collection. Journals in economics, political science, and sociology often require that the data supporting an article be made publicly available. The ISSC has stated 'Full and open sharing of the full suite of datasets for all social scientists is a fundamental goal.' The International Council for Science (ICSU) recommends '… as a general policy the fundamental principle of full and open exchange of data and information for scientific and educational purposes.'

One of the several reasons for such norms is the high value that science puts upon the ability to perform independent replications of published analyses. In 1901 Sir Francis Galton (quoted in Stigler 2000, p. 657) observed:

> I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of all his data, in some place where it should be accessible, under reasonable restrictions, to those who desire to verify his work

Today much of the social and behavioral science community shares Galton's view. Indeed, when the raw data are not available, there can arise a cloud of suspicion (or worse) about a researcher's published work. Had Cyril Burt's raw data from his study of the inheritance of intelligence among separated twins not been burned by his secretary upon his death, Burt's reputation might have survived, or at least he might have escaped the most serious charge, that he knowingly fabricated his data.

4. *Concerns for Privacy and Confidentiality*

Data archives have been governed historically by a high professional standard for the protection of the privacy and confidentiality of the individual subjects of research. Today, they are also governed by both laws and restrictions imposed by data collectors. The laws arise from societal concern, over protecting the privacy and confidentiality of individuals and organizations. The restrictions arise from researchers' legitimate concerns with protecting their intellectual property.

Most data archives require a prior commitment by the user of the data to use the data solely for statistical purposes and never to seek to use the data for purposes of identifying a person or an organization. Such pledges could be electronically signed on the Internet, permitting instant access to the data. However, this easy procedure is not the practice worldwide and is unlikely to be the practice in the near future. There are indications that access to data may become more, not less, difficult, primarily because of privacy concerns that arise with particular force now that technology permits extremely rapid and broad dissemination of data. In addition, there is concern about the growing ability to combine information about individuals from several sources of information. Governments differ greatly in data protection practices and laws. The US system provides open access to microdata that has been properly made anonymous, but this is the exception rather than the rule. Some governments consider demographic, microeconomic, and macroeconomic data to be matters of national security or heritage. Public opinion polls and data collected under academic auspices are more often made available for data sharing than are official microdata. Stringent data protections, sometimes involving criminal penalties, exist in many countries. In 2000 there was an ongoing dispute between the European Union and the United States over the protection of personal data collected from European citizens over the Internet.

The most common restriction arising from intellectual property concerns is the provision of a reasonable length of time for the original data collector to have exclusive access for the data, often one year. Other restrictions can effectively block access. Some archives will not accept data deposited with onerous restrictions, while others are forced by law or custom to do so. Several kinds of restrictions may be imposed: (a) the original investigator will permit the data to be used for scientific research and teaching, but not for publication until a 'special release' is personally obtained from the original investigator; (b) the depositor will decide whether to grant permission for specific scientific research and teaching uses of the data on the basis of requests; and (c) the depositor may require prior review (and perhaps censorship) of any publications. Such conditions run counter to the underlying ethic favoring the sharing of data and work against the ready conduct of research. Further inter-archival cooperation in addressing such issues is warranted. Moreover, the archives must strive to ensure that privacy concerns are not inappropriately permitted to limit the kinds of questions that researchers ask.

5. *New Ways of Protecting Privacy and Confidentiality*

Concern over the protection of individual privacy and confidentiality, particularly of official data, has led to the emergence of secure data enclaves. In such enclaves the researcher does not deal directly with the data but instead has highly controlled access to the results of analyses. The raw data never leaves the enclave or passes through the researcher's hands. One of the oldest of these arrangements is that of the Luxembourg Income Study (LIS), which supports comparative research on the economic status of the populations of more than twenty countries. The stringent data protection laws of the Grand Duchy of Luxembourg presume that all personal data, privately or publicly collected, are sensitive. The law further requires the licensing of computer systems that will be used to process personal data. Under these circumstances many governments have agreed to provide access through LIS to data that they would never archive for public dissemination.

The US National Institute on Aging and National Science Foundation are now providing support to data enclaves in the United States, permitting a level of access to detailed Census and other data that was previously unknown. Controls are physical as well as technical and involve searching the contents of materials entering and leaving highly secure data facilities, screening of computer analyses for intentional or inadvertent disclosure of individual information, and limits on the detail of information that may be extracted from the data files. It is unclear what role data archives will play in constructing and operating secure data enclaves, as their aim has always been to allow researchers considerable flexibility in their data analyses. Nevertheless, there may be economic efficiencies to be obtained from embedding secure data enclaves in data archives.

6. *Technological Changes in Data Archives*

The information technology (see *Information Technology*) infrastructure of data archives has undergone the same remarkable transformation as was experienced recently in the worlds of business and government. When data archives began, the principal data storage medium was the punched card, also known as the Hollerith card from its development by Herman Hollerith for use in processing the 1890 US Census. This technology persisted basically unchanged for 60 years. Magnetic media (such as low-density tapes) were introduced into data processing in 1951, when the US Census Bureau installed 'the world's first electronic general purpose data processing computer,' the UNIVAC I. Tapes were in extensive use by data archives by 1965, and their ever-increasing capacity (due to higher recording density) made them practical for storing and copying large-scale data collections that punched cards could never have accommodated. By the mid-1980s the desktop computer was beginning to have the capability of running complex statistical software, with high-speed central processing units, large random access memory, and hard disks capable of handling such large files as those from the International Social Survey Program and the General Social Survey. Risking obsolescence, most archives took many years to adjust fully to the new computing environment. However, major archives now provide services directly to desktop computers through client-server computing environments. In doing so, they are often creating service relationships with individual data users that previously existed only with campus-wide data libraries. The growth of cheap computing has encouraged some data archives to provide new exploratory analytical services from their own servers, as well as the ability to extract variables or units of observation. One of the most advanced of the new efforts is the Networked Social Science Tools and Resources (NESSTAR) project, a collaboration of the United Kingdom Data Archive, the Norwegian Social Science Data Service, and the Danish Data Archive. It enables users to tabulate and visualize the data on-line and to obtain the data over the Internet, as well as to access detailed information. Perhaps equally importantly, NESSTAR can locate multiple data sources across archives and national boundaries, creating what could become a virtual global data archive. Related software developments are underway at the Center for Survey Methods, University of California, Berkeley; the Integrated Library and Survey-data Extraction Service (a project of several Dutch, German, French, and Irish institutes and the ZA); and the Virtual Data Center of Harvard University and the Massachusetts Institute of Technology. Archives' clientele has shifted from being almost exclusively experts in data analysis to including less trained users, and this kind of service is particularly beneficial to these new users.

Archives also had to respond to a second great technological revolution: the telecommunications (see cross-reference) revolution of 1990 onwards. In the era of punched cards, transmittal of data sets to users was exceedingly cumbersome and slow. Magnetic tapes relaxed those problems to a degree, but even as late as 1995, the delivery of one copy of the Public Use Microdata Samples from the US Census involved dozens of high-density tapes. The telecommunications revolution saw the virtually complete elimination of tapes as dissemination media, although they remain valuable as archival media. Internet access, coupled with CD-ROM physical media, took their place. It will soon be possible to assume that high-speed telecommunications are available in much of the world. Fiber optic and wireless technologies have ample capacity to transmit the entire contents of the Library of Congress in only a few seconds; social science data will be trivial

passengers. Thus, Converse's vision of 1964 can be realized:

> … the day when an investigator at one of the far-flung California campuses can, in an afternoon's work at a satellite of his local computer facility, learn just what data exist in the total network bearing on the hypothesis he wishes to check out, order by telecommunication either a statistical analysis of those data from an East Coast repository (or the raw data themselves if he has more extended use in mind), and have his output in time for dinner, all at minimal immediate cost (Converse 1964, p. 280).

This was the idea of the virtual archive: a number of data archives that are 'coordinated' so as to offer an investigator the ability 'to know both that they (data) exist and where they exist.' Archives still require coordination of their finding aids, which in Helen Crossley's vision of 1958 (Riley 1958) would have involved a 'common system of indexing' so that users could find resources across archives. The archives may be nearing that goal with an international project known as the Data Documentation Initiative, an effort to formulate an interarchival standard for describing data.

7. *Conclusion*

In 1944, President Franklin D. Roosevelt asked Vannevar Bush a set of questions about science after the conclusion of the war. Among those was 'With particular reference to the war of science against disease, what can be done now to organize a program for continuing, in the future, the work which has been done in medicine and related sciences?' In the response, *Science—The Endless Frontier*, Bush observed 'The striking advances in medicine during the war have been possible only because we had a large backlog of scientific data accumulated through basic research in many scientific fields in the years before the war.' Perhaps some future scientist will observe that the striking advances in the social and behavioral sciences of the first decades of the twenty-first century were similarly due to the 'large backlog of scientific data' preserved by the data archives. These relatively young sciences have yet to accumulate the centuries of observations of most natural sciences, but the archives are prepared to facilitate that accumulation. Progressively sounder sciences may be built upon this accumulation of observations.

See also: Databases, Core: Sociology; Rokkan, Stein (1921–79)

Bibliography

Clubb J M, Austin E W, Geda C L, Traugott M W 1985 Sharing research data in the social sciences. In: Fienberg S E, Martin M E, Straf M L (eds.) *Sharing Research Data*. National Academy Press, Washington, DC

Converse P E 1964 A network of data archives for the behavioral sciences. *Public Opinion Quarterly* **28**(2): 273–86

Hyman H H 1972 *Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities*. Wiley, New York

International Council for Science 1996 ICSU/CODATA Group on Data and Information. General Assembly Resolution

International Social Science Council 1994 General Assembly Social Science Data Management Policy

Riley J W Jr 1958 Proceedings of the thirteenth conference on public opinion research. *Public Opinion Quarterly* **22**(2): 169–216

Rockwell R C 1997 Using electronic social science data in the age of the Internet. In: Dowler L (ed.) *Gateways to Knowledge*. MIT, Cambridge, MA

Rokkan S 1966 *Data Archives for the Social Sciences*. Mouton, Paris

Stigler S M 2000 The problematic unity of biometrics. *Biometrics* **56**(3): 653–8

R. C. Rockwell

Data Collection: Interviewing

The social science interview is a situated social interaction the purpose of which is to transfer information from one person to another. Information or data are obtained from one person, the respondent or subject, as they answer questions posed by another, the interviewer or researcher. The interview in its myriad forms is a primary data-gathering tool of social scientists, and, therefore, it is important to understand the interview process since this will affect the nature and quality of the data gathered or generated.

1. *Situating the Interview*

Interviews may be conducted in a variety of ways and in a variety of situations. Inevitably they are socially and methodologically embedded in larger contexts. Relatively little research has been done either on the impact of different social situations and settings on the conduct of interviews, or the impact of varying situations on the reliability and validity of interview results. The social setting may range from the privacy of the home to the office to the public 'person on the street,' and the social context may range from a highly private, anonymous, or confidential one-on-one interview to one involving the presence of strangers, or intimate family and friends as co-respondents. Research has shown that in contrast to the one-on-one interviewer/respondent situation, the presence of others tends to produce more socially normative responses (Lever 1981). Much more research is needed on the impact of these varying social contexts.

We must also situate interviews in the different methodologies or styles of research used by social scientists. For heuristic purposes we can identify four

major styles of research—experimental, survey, fieldwork, and nonreactive research. The interview, though found in all four methods, is particularly important in survey and field research and these may be thought of generally as representing quantitatively oriented vs. qualitatively oriented interview styles, respectively. The quantitative/qualitative distinction itself encompasses a variety of comparisons that may range from more philosophical and epistemological issues such as objective vs. subjective knowledge, to more grounded and specific issues of questionnaire construction and closed vs. open answer formats. A central distinction running throughout these comparisons is the question of the nature of control over the interview format and content.

2. *A Note on Control Over the Interview—Form and Content*

Different interview formats may be arrayed on a number of different continua but of particular significance is the question of varying degrees of control. At one extreme the interviewer exerts full control over the situation and at the other the interviewer seemingly minimizes control, relinquishing more control over the situation to the respondent. There is a zero-sum quality of control in the relationship between interviewer and respondent. The case of maximal interviewer control is exemplified by the 'forced choice' survey interview where not only the format of the question is standardized (stimulus) but the possible answers of the respondent are standardized as well (response). The interviewer ensures a constancy of routinized control in order to generate highly reliable data. Concerns over such control are often associated with surveys where the interview is used as a tool to generate standardized data from a sample of numerous units for the purposes of aggregate analysis that may be generalized to some larger population.

The other extreme, where the interviewer exercises minimal control, is seen for example in the Rogerian psychoanalytic interview. Here the respondent is 'in control' as to the content of the interview and the interviewer is trained to be 'nonreactive.' It should be noted, however, that even here the interviewer is exercising control over the format or form of the interview, and that both the forced choice survey interview and the nonreactive interview (Rogers 1945) are interaction situations far removed from the 'natural' form and content of everyday personal interaction. The field interview of participant observation is, by contrast, a situation often closer to the give and take of everyday 'normal conversation' (Seidman 1998). The respondent exercises a degree of control over both the content and the form, but the interviewer, through judicious use of probes, can often guide the interview into topic areas of interest to the researcher. This is the case of the 'focused' or 'topical' interview (Merton 1956). From these considerations one may construct a typology of control which the researcher exercises in the interview situation:

Typology of interviewer control in different styles of research interviews

	Control over content: Hi	Control over content: Low
Hi	survey questionnaire	nondirective psychoanalytic
Low	semifocused in-depth	conversational fieldwork

These different interview formats are often associated with survey vs. fieldwork but extend to a variety of methods. A multimethod strategy which emphasizes different strengths and weaknesses of different methods or styles of research nonetheless values the interview as a central data collection tool that can simultaneously provide quantitative and qualitative data, and even be used in multimethod experimental strategies. For example Schuman et al. (1985) incorporated an experimental design into a survey by having interviewers systematically vary the format of questions asked of different respondents.

3. *The Theory of the Spring*

Some have suggested that all measurement involves interaction (Babbie 1998)—interaction with the external world in some form or another in order that information about that external world may be transferred to an observer. Theories of measurement take into account a detailed exposition of the measurement process itself as constitutive of the data generated. In the prosaic act of weighing a bunch of bananas in the produce section of a grocery store, for example, one is operating on an assumption about how the spring in the scale responds to varying masses of bananas subjected to the law of gravity in moving the dial in a consistent and credible way. One operates on the assumption that some scientist or engineer has worked out in detail the physical elasticity of the steel spring at the heart of the scale's measurement process. In the interview this interaction is particularly noteworthy since it is *social* interaction—usually between two people—and the corresponding theories of measurement most applicable are theories of social interaction itself. In this section we will briefly apply four distinct theories of social interaction to the interview to see the varying perspectives they provide on this all-important data-gathering tool. The four theories of social interaction are: (a) role theory, (b) exchange theory, (c) conflict theory, and (d) symbolic interaction.

4. *Role Theory*

Role theory begins with a set of normative expectations that are presumed to define particular positions or statuses in social structure and their corresponding

roles or behaviors in interaction with others. The roles or statuses most clearly central here are those of 'interviewer' and 'respondent' themselves. How well one knows these roles, that is, knows the normative expectations for behavior associated with the respective roles, is a function of one's prior experience and knowledge gained either first hand or through vicarious observations—in the media, through cartoons, conversations, or the classroom. To be 'trained' as an interviewer is to learn a set of normative expectations about how one should interact with a respondent. The simplest norms are of ten those that are 'taken for granted' such as who asks the questions and who gives the answers. Sometimes there are widely accepted norms defining the ideal interviewer role such as—one should not reveal personal information that might thereby bias the respondent's responses (Gordon 1975). However, even these commonplace widely shared norms may come into question as there may be conflict among the different goals for the interview. For example, one goal may be to get as much revelatory information as possible vs. another goal which is to not bias the respondent's answers. The different goals of the interview (unbiased data vs. more fully revelatory data) may come into conflict and suggest different contradictory norms, for example, about how much personal information the interviewer should reveal to the respondent in the interview situation.

Role theory asserts that the norms governing interaction are there for the purpose of realizing specified goals out of the interaction. In short, normative behavior is goal oriented. The norms governing the interviewer are likewise oriented to maximizing certain qualities (or values) of the data—for example, norms that call for a common stimulus to achieve validity across respondents, or other norms that emphasize supportive interaction to produce full and complete responses.

A common application of role theory to the interview situation is a concern with bias introduced by other social statuses in a person's status set—the most obvious and frequently researched being the ascribed and highly visible social statuses of race, gender, and age. These statuses are often considered to be especially important when questions in the interview touch on topics closely related to them. Much research has been conducted on the effects of these statuses on interview outcomes precisely because they are considered to be master ascribed statuses that are both readily observed and ubiquitous in all interactions. They are often attributed even in telephone interviewing and not just in face-to-face situations.

One of the earliest findings of interviewer effects by race, dating to World War II, was that answers to questions about race relations were strongly impacted by the races of the interviewer and respondent (Hyman et al. 1954). African-Americans reported lower levels of satisfaction with race relations to African-American interviewers than to Caucasian interviewers. Caucasian respondents reported lower levels of acceptance of African-Americans to Caucasian interviewers than to African-American interviewers. These results have been demonstrated repeatedly since then in studies by Schuman et al. (1985) and Anderson et al. (1988). The latter study found that blacks interviewed by whites were much more likely to express warmth and closeness toward whites than blacks interviewed by blacks. It is important to note that more limited effects of race of interviewer and respondent have been observed on studies of other topics, unrelated to race.

Another interviewer status examined in some detail has been gender. Earlier research has generally shown that the gender of the interviewer has no effect on responses to survey questions except when the content of the questions relate to sexual behavior or gender related issues (Clark 1952, Hyman et al. 1954). Research by Kane and Macaulay (1993) found that both male and female respondents express more egalitarian gender related attitudes or greater criticism of existing gender inequalities to female interviewers. Furthermore, male respondents offer significantly different responses to male and females interviewers on questions dealing with gender inequality.

Age, as the third of the most visible ascribed status characteristics, has also been found to have only limited interviewer effects. School age subjects have shown different responses to older vs. younger interviewers for questions about peer influences (Ehrlich and Riesman 1961). Younger interviewers obtained slightly more peer-oriented and less adult-oriented answers than older interviewers. With respect to age of respondents, research has shown that telephone surveys tend to under-represent older respondents, but response distributions do not vary by telephone vs. face-to-face interviews across age categories (Herzog et al. 1983). As with race and gender, age appears to have little effect on interview responses except when the topic being covered is directly related to the status itself.

5. *Exchange Theory*

The central assumption of exchange theory is that all interaction is based on ego's attempt to realize a 'social profit' by minimizing the cost of one's own actions and maximizing the benefits from alter's actions. The operating principle is that for interaction to continue, the norm of reciprocity must be followed such that the ratio of two parties' costs to benefits should be equal. If we view the interview as an exchange, it is relatively easy to define the benefits to the researcher; they are the data provided by the respondent which are his or her costs in time and effort in answering the questions. It is often more nebulous to define what the benefits to the respondent are, that is, the costs incurred by the interviewer that directly benefit the respondent. Of

course when respondents are being directly paid by the interviewer for their time, the benefits are clear and unambiguous. Market researchers, for example, sometimes establish panels of randomly selected respondents who are paid for their time in a variety of repeated surveys. In short, they are 'professional respondents' for whom the exchange is a pure labor market transaction. Exchanges in kind may also be provided such as a lunch or dinner, a free gift, or a chance at a drawing. More often than not, however, the benefits are assumed to be some nebulous internal psychic reward that can include everything from the opportunity to vent one's opinions on a particular topic, that is, make use of a willing listener, to an increased sense of self-worth engendered by volunteering to help someone out, especially when the interviewer may have already appealed to one's 'specialness' by indicating that they have been 'selected' to participate. Altruistic appeals also redound to one's sense of self-worth as when one is solicited as a respondent to provide data that may be useful in developing new policies and programs or evaluating existing ones even if one is not seen to be a direct beneficiary of them. Even the most abstract altruistic appeal—that by responding one is 'contributing to knowledge'—may at times work as a sufficient psychic reward.

What the exchange perspective highlights is that 'time spent' by the respondent is a very real but variable cost that may influence their willingness to be interviewed in the first place, and their loquaciousness even if an interview is granted. The preciousness of 'time spent' doing an interview is differentially valued from one potential respondent to another (for example, the working mom vs. the elderly retired nursing-home resident). Such calculations are highly likely to influence differential response rates in a selective or biased way. Not only are these calculations likely to vary from one set of statuses to another, but they will likely vary even for the same individual under different circumstances and in different situations, such as time of day or day of week or season of the year, or at home, shopping, or workPol 1992).

The cost of an interview is not simply one-sided, for the interviewer's time is also a very tangible cost, and one that must be taken into consideration often in a direct monetary way as when one is hiring interviewers or relying on a survey supplier. Such costs may have direct bearing on the nature and quality of the data gathered, for they can affect the length of the interview forcing real choices in research design as to sample frame and size, or number of topics explored and questions asked.

In short, from an exchange perspective the interview is a cost/benefit transaction. The calculus of these varying costs to the interviewer/researcher must be weighed against the value of the data to be generated. In the design of surveys, for example, a distinction is often made between open-ended vs. closed-ended questions, and the former generally prove to be more difficult and more expensive to code and analyze than those from closed-ended questions. However, Geer (1991) found in an experiment that the open-ended comments are very effective in measuring the important concerns or 'salient' issues of respondents. In short, it appears to be cost considerations as much as validity issues that have led to the dominance of forced choice questions in survey research.

Costs and benefits of the interview also form a part of the calculus of the ethical considerations involving human subjects. Costs to the respondent slide over into a variety of 'risks' that may result from the disclosure of certain data and the loss of privacy. Anonymity or confidentiality may be required to forestall those costs. The idea behind 'informed' consent is that this calculus should be made by the subject/respondent based on a full disclosure of risks and benefits. Research in which there are substantial risks can only be justified by substantial benefits that outweigh these costs, and above all it is up to the respondent to make this determination.

We should be aware that the utilities (value) of a respondent's cost/benefits may differ from those of the researcher and one is confronted with incomensurability which may work to the advantage of the researcher. For example, an elderly retired single respondent may have a surplus of free time which is of low cost to give in answering a lengthy questionnaire, while to the highly trained interviewer this time is a valuable costly commodity in the research budget. Such incommensurability may at times work to the detriment of the interviewer when the inability to reach a common understanding of the relative value of information or time leads to an unwillingness to participate in what is considered to be a 'worthless' activity. (We note that this is one of the points of the symbolic interactionist perspective outlined below which involves negotiating not only a common value but a common understanding of the meaning as well as the value of various social goods.)

6. *Power Theory*

The power or conflict perspective directly addresses the insight that much social interaction—and also the interview—involves differences in power among the parties. To be sure, power or conflict theory may be closely linked to either role theory (as when, for example, one talks about the differences in power that are present between the status of a male and a female in interaction) and also to exchange theory (as when the costs to participating in the interview are different between the interviewer and the respondent, or the benefits that one can provide to the other are different and one or the other party uses those differences in resources to extract what they want from the other). What the power or conflict perspective stresses is the

reality that interviewer and respondent often have different degrees of power relative to one another and these may influence the nature and quality of the data generated in the interview.

Let us first address situations in which the interviewer has more power. We will use power in its broadest meaning to encompass the most subtle uses of prestige to the blatant uses of physical coercion. The latter is rare if ever present but the former is quite ubiquitous. The imprimatur of an individual's title 'Doctor' or 'Professor,' or an institution's stature 'X University,' or 'Y Agency' are often used to impress potential respondents as to the importance of the research and the desirability of their participating. In the actual conduct of the interview itself power may be reflected in who makes the effort to initiate the contact or who travels to whom to conduct the interview. Being on one's home turf would seem to imply more power and control over the situation. A notable exception is the case in which elites may avoid being interviewed in certain situations (coming to their homes) in order to maintain a heightened degree of privacy. Researchers have sometimes strategically relied on the interviewer shifting out of the role of interviewer to becoming an observer and recording the presence or absence of such things as a piano, types of pictures and paintings hung on walls, and the use of curtains or drapes (Laumann and House 1970). This is an example of combining methods in a multimethod strategy (Brewer and Hunter 1990) to provide validity checks surrounding lifestyle and status and class variables.

At times respondents may sense a threat from the interview situation that reflects a potential power differential between them and their interviewers. Potential threats may apply to divulging information concerning content areas that include socially undesirable and even illegal behaviors and could lead to not reporting or under-reporting various forms of deviance and even criminal behavior ranging from drug use to official 'cheating' (Bradburn et al. 1979). Implicit threat could also lead to over-reporting socially desirable behavior such as voting or charitable giving. There is some limited evidence that interviewer characteristics are related to threat. Respondents are willing to report socially undesirable behavior such as drug use to interviewers who are perceived to be more like themselves (Johnson et al. 2000).

The most widely used procedure to reduce threat is to eliminate the interviewer entirely or as far as possible from the data gathering process. In some cases this means group administration of the interview in a public or quasipubic setting but where responses are private, such as a school classroom setting. Even if the interviewer is present, it is possible that the threatening questions may be self-administered. Maximum privacy is obtained when the threatening questions are on a computer and the respondent reads questions from a screen. The respondent enters the answers directly into the computer and the interviewer is completely unaware of what answers are given (Tourangeau and Smith 1996). It must be recognized that the use of self-administered methods reduces but does not eliminate respondent misreporting due to the power differentials in the interview situation and the potential threat of disclosure. It is precisely why the use and assurance of anonymity or confidentiality are not only ethical considerations but themselves may have a direct and positive impact on the validity of data generated through the interview.

Power is implicit in the very asymmetry of who asks questions and who answers them. The benefit is seen to be asymmetrical as the powerful ask questions of, glean information from, and acquire data about the less powerful. Much more rarely does information flow in the reverse direction—from the powerful to the powerless. Knowledge is power and its differential distribution is a means of obtaining and maintaining that power (Hunter 1993).

Interviewing of elites is a situation in which the respondent often has more power than the interviewer. Such control may be reflected first in limiting the ability to gain access to begin with and continue through the interview situation with control over the content, format, recording, and even the use of interview data (Thomas 1993). Although interviewers in their comportment should, of course, show respect to all respondents, an added degree of deference might be shown to those having more power or higher status than the interviewer. This deference could easily affect the likelihood of the interviewer's continuing to probe in finer and finer detail into the lives of the respondents, thereby limiting the quality and depth of the data.

The power differential enters as well into ethical considerations of the interview, and this is why interviews must be voluntary, since the counter case involving any form of coercion, no matter how mild, may be deemed unethical. For example, if a welfare recipient is being interviewed for research purposes, they might agree to participate in the interview if it is closely linked to the intake interview used to determine welfare elgibility; or, a person in a public setting may agree to be interviewed out of a sense of peer pressure (e.g., students in a classroom setting). It is imperative that researchers be fully cognizant of and sensitive to the power relations involved in the interview as these may not only affect the quality of the data but reflect larger social processes of which the interview itself is a microcosm.

7. *Symbolic Interaction*

The interview situation is not one which is necessarily clearly defined and unambiguous, especially on the part of respondents. Therefore, one of the key activities of the good interviewer is to negotiate with the respondent in defining the appropriate behaviors and

what can be expected in the interview situation. Drawing on the symbolic interactionist framework (Goffman 1959), the interview may be considered to be a situation in need of mutual definition through the interaction of the interviewer and respondent. At a minimal level is the idea that the interviewer is supposed to ask questions and the respondent is to give answers. But the fine nuances of this somewhat 'artificial' interaction are a negotiated reality in which the interviewer has the task of 'educating' the respondent into the appropriate behavior. For example, a too loquacious respondent might be interrupted in the course of an answer and asked to respond more simply—'strongly agree, agree, neutral, disagree, or strongly disagree.'

The negotiated definition of the situation draws on the past experiences of both interviewer and respondent. The idea of interviewer training focuses on behaviors used to aid in an appropriate definition of the situation such as what to say to gain access, how to present self for gaining agreement to participate, how to develop and maintain rapport and how to terminate an interview, and of course, how to avoid biasing respondents (Billiet and Loosveldt 1988). Some respondents may be familiar with the interview situation from having participated in them previously, while others may have been exposed to the situation vicariously through the media. Some may in fact be 'professional respondents,' that is, those paid to participate in ongoing panel studies in which they have been more thoroughly trained or educated into the respondent's role.

Expectations of the interviewer about the interview process have also been found to influence the nature of the respondents' participation. Some interviewer expectations are formed before the interviewer conducts any interviews, for example, expectations about how difficult or threatening certain questions may be to respondents; while other expectations may develop in the course of an interview as the interviewer begins to form a picture of the respondent through their prior responses, and yet other expectations may emerge as the interviewer completes multiple interviews and begins to form a patterned set of expectations for emerging 'types' of respondents. For example, interviewers who expected that questions about threatening behavior would be easy to ask received higher rates of responding than interviewers who thought the questions would be difficult to ask (Bradburn et al. 1979). It would appear that interviewer behavior in some way reflects expectations and that behavior calls forth different responses on the part of subjects.

Ambiguity in social interaction is a key problematic in the symbolic interactionist framework. Much interaction is seen to be specifically oriented to the reduction of ambiguity in an attempt to arrive at a mutually agreed upon definition of the situation. The issue of ambiguity is also significant in the interaction between interviewer and respondent. If a respondent should ask quizzically, what is meant by a question, often interviewers are simply instructed to repeat the question without further elaboration or clarification of the question so as no to differentially bias one respondent vs. another. On the other hand, if a respondent gives an answer that is limited, partial, or ambiguous, the interviewer is often instructed to 'probe' to get greater clarification. In a classic laboratory experiment, Hyman et al. (1954) demonstrated that interviewers faced with an ambiguous response coded the answer on the basis of previous answers in the questionnaire. Respondents were asked a series of questions about whether the United States should actively intervene or should isolate itself from world affairs. When the response was ambiguous, those who had previously given unambiguously interventionist answers were coded as giving an interventionist answer, while those who had previously given isolationist answers were coded as giving an isolationist answer. In this experiment interviewers were not allowed to probe the ambiguous answer for clarification.

The first phase of an interview involves negotiation in securing the willingness of the respondent to participate. By negotiating a formal 'commitment' from respondents in the form of a signed 'pledge,' to answer 'fully' and 'truthfully' has been found to increase the degree of cooperation by respondents and the completeness of their responses (Cannell et al. 1981). A further mechanism for increasing cooperation is to flexibly tailor the appeal to participate to each individual respondent rather than using a standardized approach. Such individually tailored appeals generate higher levels of cooperation and also experienced interviewers are more likely to obtain such higher rates of cooperation (Groves and Couper 1998).

8. *A Multimethod Strategy*

Interviews in their various forms and formats will continue to be a primary tool in the social scientist's toolkit for generating reliable and valid data. Direct comparisons of the interview to other tools for data generating, such as content analysis of media, or conversational analysis, the use of systematic observation of behavior, or the analysis of archives can only further the understanding of the strengths and limits of the interview in a multimethod strategy. Some but far too little research has been concerned with such comparisons. One example is Janet Lever's (1981) multimethod study of children's play. She found that interviews with kids produced more gender stereotyped responses to the nature of boy's vs. girl's play than observations of their behavior by the researcher or diaries kept by the kids themselves. Throughout the above discussion the multimethod strategy has been mentioned at various points and briefly noted in passing. Bringing the multimethod strategy to the foreground whether narrowly focused on measure-

ment or more broadly on the research process as a whole (Brewer and Hunter 1990) can only lead to a more refined use of the interview in social science research. The shape and function of the interview as a tool can more clearly be delineated as it is compared to these other tools in the same research project. The centrality and the ubiquity of the interview requires much more systematic research fully exploring the social, situational, and methodological dynamics of the interview process. Whether viewed from a post-modern perspective of a greater need for reflexivity within the social sciences or a positivist perspective of understanding the sources of bias and the need to reduce measurement error, much more research on the interview itself is needed. A few lines of inquiry and understanding have been developed as noted, but a great abyss of ignorance remains to be plumbed.

At various points throughout this discussion I refer to data being either 'gathered' or 'generated.' I fully recognize the implications of these two distinct concepts as the former privileges an older classical positivism while the latter speaks to a more agency oriented social constructionist view of science. The distinction will itself become part of our nuance discussion of the interview.

See also: Clinical Assessment: Interview Methods; Field Observational Research in Anthropology and Sociology; Observational Methods in Human Development Research; Personality Psychology: Methods; Questionnaires: Assessing Subjective Expectations; Questionnaires: Cognitive Approaches; Sample Surveys: Cognitive Aspects of Survey Design; Social Psychology: Research Methods; Social Survey, History of; Surveys and Polling: Ethical Aspects

Bibliography

Babbie E 1998 *The Practice of Social Research*, 8th edn. Wadsworth, Belmont, CA

Billiet J, Loosveldt G 1988 Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opinion Quarterly* **52**: 190–211

Bradburn N, Sudman S et al. 1979 *Improving Interview Method and Questionnaire Design*. Jossey-Bass Publishers, San Francisco

Brewer J, Hunter A 1990 *Multimethod Research: A Synthesis of Styles*. Sage Publications, Newbury Park, CA

Cannell C, Miller P, Oksenberg L 1981 Research on interviewing techniques. *Sociological Methodology* **12**: 389–437

Clark R A 1952 *The Projective Measurement of Experimentally Induced Levels of Sexual Motivation*. *Journal of Experimental Psychology* **44**: 391–9

Geer J 1991 Do open-ended questions measure 'salient issues'? *Public Opinion Quarterly* **55**: 360–70

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday/Anchor, New York

Gordon R 1975 *Interviewing: Strategy, Techniques, and Tactics*. Dorsey Press, Homewood, IL

Groves R, Couper M 1998 *Nonresponse in Household Interview Surveys*. John Wiley and Sons, Inc., New York

Herzog A R, Rodgers W, Kulka R 1983 Interviewing older adults. *Public Opinion Quarterly* **47**: 405–18

Hunter A 1993 Local knowledge and local power: notes on the ethnography of lcoal community elites. *Journal of Contemporary Ethnography* **93**: 36–58

Hyman H, Cobb W, Feldman J, Hart C, Stember C 1954 *Interviewing in Social Research*. University of Chicago Press, Chicago

Johnson T, Fendrich M, Shaligram C, Garcy A, Gillespie S 2000 An evaluation of the effects of interviewer characteristics in an RDD telephone survey of drug use. *Journal of Drug Issues* **30**: 77–101

Kane E, Macaulay L 1993 Interviewer gender and gender attitudes. *Public Opinion Quarterly* **57**: 1–28

Laumann E, House J 1970 Living room styles and social attributes. *Sociology and Social Research* **54**: 321–42

Lever J 1981 Multiple methods of data collection: a note on divergence. *Urban Life* **10**: 199–213

Merton R 1956 *The Focused Interview: A Manual of Problems and Procedures*. The Free Press, New York

Pol L 1992 A method to increase response when external interference and time constraints reduce interview quality. *Public Opinion Quarterly* 356–60

Rogers C 1945 The non-directive method as a technique for social research. *American Journal of Sociology* **51**: 143–52

Schober M, Conrad F 1997 Conversational interviewing. *Public Opinion Quarterly* **61**: 576–602

Schuman H, Steeh C, Bobo L 1985 *Racial Attitudes in America: Trends and Interpretations*. Harvard University Press, Cambridge, MA

Seidman I 1998 *Interviewing as Qualitative Research*, 2nd edn. Columbia University Teachers College Press, New York

Suchman L, Jordan B 1990 Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association* **85**: 2322–53

Suchman L, Jordan B 1991 Validity and collaborative construction of meaning in face-to-face surveys. In: Tanur J (ed.) *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*. Russell Sage Foundation, New York, pp. 241–67

Sudman S, Bradburn N 1974 *Response Effects in Surveys: A Review and Synthesis*. Aldine Publishing Company, Chicago

Thomas R 1993 Interviewing important people in big companies. *Journal of Contemporary Ethnography* **22**: 80–96

Tourangeau R, Smith T 1996 Asking sensitive questions. *Public Opinion Quarterly* **60**: 275–304

A. Hunter

Database, Core: Linguistics and CHILDES

1. History of Data Sharing in Child Language

The Child Language Data Exchange System (CHILDES) was first formally discussed in the summer of 1983, at a meeting of the John D. and Catherine T. MacArthur Foundation funded Network on the Transition from Infancy to Childhood. The MacArthur Foundation had funded this Network, which consisted of five geographically centered nodes of activity, to promote innovative and collaborative work on the age period 1 to 3 years. Within the

network, about a dozen individuals most directly involved in the study of child language attended a working session chaired by Brian MacWhinney (then of the University of Denver) and Catherine Snow (Harvard Graduate School of Education) to plan an effort to promote collaboration by sharing child language transcript data. The level of interest expressed in that initial meeting was sufficiently high that MacWhinney subsequently submitted a proposal to the MacArthur Foundation for funding to establish the data-sharing system.

Child language transcript sharing had been initiated by Roger Brown in the first large longitudinal study of child language in the audiorecording era. Brown's study of children he called Adam, Eve, and Sarah, which established procedures and basic measures that would become standard in the field, had been started in the early 1960s at Harvard. In addition to weekly or biweekly recordings of the children at home, Brown established a seminar of graduate students and colleagues who met to study the newly produced transcripts. To facilitate this effort, the Brown corpus was produced using mimeograph masters. One full (though somewhat faded) set of these transcript copies was still available in 1983, and indeed formed one of the first large corpora entered into the CHILDES system (after being photocopied to improve contrast, scanned, then checked and reformatted extensively to match the CHILDES-prescribed format).

The history of Brown's generous sharing of the Adam, Eve, and Sarah transcripts was crucial to the ready acceptance of the CHILDES proposal by the field of child linguists as well, since it had demonstrated the potential for multiple waves of analysis of a wide variety of phenomena using a single corpus. Dozens of articles and books had been published using the Brown corpus, focusing on topics as diverse as the development of reference to the past (Cromer 1968), maternal teaching techniques (Moerk 1977), conversations about kinship (Goldfield and Snow 1992), and semantic and morphological development (Brown himself 1973). These provided evidence of the utility, efficiency, and feasibility of data sharing.

2. *Design of CHILDES*

Initial funding from MacArthur provided for the development of the basic components of the system. These included a set of transcription conventions, known collectively as CHAT (Codes for the Human Analysis of Transcripts) (see MacWhinney 2000a). The CHAT guidelines were imposed *ex post facto* on important corpora, such as Brown's, that had already been collected, but of course much more importantly came to define one set of expectations for how transcripts should be produced. While considerable discussion concerning the advantages of various transcript formats occurred during the 1980s, and while limitations of the CHAT system for representing conversational features easily are clear, nonetheless the convergence by a large portion of the child language community on a single transcription system had important advantages. The CHAT system was elaborated with suggested procedures for phonetic, morphosyntactic, speech act, and error coding. It is important to note, though, that the vast majority of the CHAT guidelines are merely advisory; a transcript that conforms in certain minimal ways to the CHAT format (minCHAT) can be incorporated into the CHILDES database and analyzed using the CLAN programs. The major inducement for adopting CHAT guidelines was that transcripts that conformed to CHAT were susceptible to analysis by CLAN (Child Language Analysis) programs, a set of programs designed to automate various analytic procedures that had become standard in the field of child language. The earliest CLAN programs produced automatic counts of morphemes, words, utterances, and turns, to generate measures such as Mean Length of Utterance, Mean Length of Turn, and Type-token Ratio. Key word searches were also made possible (KWAL), as were procedures for Boolean searches (Combo). Subsequent expansions added programs to select certain marked portions of the transcript for further analysis (GEM), to calculate contingencies between codes in adjacent turns (Keymap), to code transcript lines for overlap with preceding lines (CHIP), and so forth. In addition, morphological tagging programs have been made available for a number of languages. These various programs are designed in most cases to analyze many files simultaneously and to generate output in a format directly usable by statistical analysis programs.

The CHILDES database is a constantly growing electronic repository of transcript data (see MacWhinney 2000b). Corpora included in the database as of January 2000 represented the acquisition of 26 languages as well as a number of language pairs and triplets (in the Bilingual section). In addition, data from children with language disorders and from narrative elicitation procedures are included in separate sections. Access to files in the database is free to anyone who agrees to use the data in accordance with the guidelines specified, primarily to cite the sources and the CHILDES manual appropriately. Files can be accessed through the website or via CD-ROMs distributed in conjunction with the CHILDES manuals.

3. *Funding History*

After the initial period of funding from the MacArthur Foundation, support for the CHILDES system was provided to Brian MacWhinney and Catherine Snow through the National Institute of Child Health and Human Development (NICHD) and subsequently to MacWhinney from NICHD and the National Science Foundation Linguistics Program.

4. Beyond CHILDES

Since the CHILDES enterprise was designed to promote collection, use, sharing, and analysis of transcript data, it did not optimally accommodate the inclusion of audio or video data. As researchers increasingly turn to such data, and seek methods for annotating, analyzing, and sharing them, a successor system to CHILDES called TalkBank has been founded, with support from the National Science Foundation. TalkBank (*http://www.talkbank.org/*) will provide standards, tools, and support for creating, searching, and publishing video-based language interactions using electronic networks.

See also: Databases, Core: Anthropology and Museums; Databases, Core: Linguistics, Online Grammars, and Dictionaries; Databases, Core: Sociology

Bibliography

Brown R 1973 *A First Language: The Early Stages*. George Allen & Unwin Ltd, London

Cromer R 1968 The development of temporal reference during the acquisition of language. Ph.D. thesis, Harvard University.

Goldfield B, Snow C E 1992 What's cousin Arthur's daddy's name?: The acquisition of knowledge about kinship. *First Language* **12**: 187–205

Moerk E 1977 *Pragmatic and Semantic Aspects of Early Language Development*. University Park Press, Baltimor

MacWhinney B 2000a *The CHILDES Project: Tools for Analyzing Talk, Volume I: Transcription Format and Programs*, 3rd edn. Lawrence Erlbaum Associates, Mahwah, NJ

MacWhinney B 2000b *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database*, 3rd edn. Lawrence Erlbaum Associates, Mahwah, NJ

C. E. Snow

Databases, Core: Anthropology and Human Relations Area Files (HRAF)

The Human Relations Area Files (HRAF) databases are collections of full texts that have been complexly indexed to facilitate worldwide comparative studies of human behavior, society, and culture. The motivating idea was to foster comparative research on humans in all their variety so that social scientists could discover which principles and explanations are universally valid, not culture bound. What may be true for one society may not be true for others. The organization known as HRAF was founded to provide data for the cross-cultural testing of hypotheses. The HRAF databases are also useful to researchers, instructors, and students who wish to compare the ethnography and archaeology of particular cultures, traditions, and regions.

Human Relations Area Files, Inc. is the nonprofit membership organization that produces and distributes the HRAF databases. It also organizes and edits encyclopedias and sponsors the quarterly journal *Cross-Cultural Research: The Journal of Comparative Social Science*. HRAF is an international consortium, with nearly 200 member institutions in 25 countries; it is also a financially autonomous research agency of Yale University, where it was founded in 1949. The HRAF databases, which are produced now only in electronic format, are accessible only through HRAF member institutions. The first and larger database is the HRAF *Collection of Ethnography*, which has been building in paper, microfiche, and electronic format for more than 50 years. It includes indexed full-text materials on more than 365 cultures around the world. The electronic version, *eHRAF: Ethnography*, has been building since 1994 and as of 2000 covers a geographically stratified random sample of 60 cultures, plus assorted ethnic groups. The second database is the HRAF *Collection of Archaeology*, which has been building solely in electronic format since 1999. As of 2000, *eHRAF: Archaeology* includes indexed full-text materials on 20 randomly selected major archaeological traditions, and many more subtraditions and sites, around the world.

1. Uniqueness of the HRAF Databases

The task of finding particular kinds of information in the HRAF databases is facilitated by the unique indexing system that HRAF has developed and refined since the 1950s. The HRAF indexing system (the *Outline of Cultural Materials* or OCM) helps the user find information on more than 750 subjects, covering just about anything that might be described in an ethnographic or archaeological document. Searching by one of the categories goes right to the relevant passages, rather than to a pointer to the passages. The efficiency of the indexing derives from the fact that a particular subject may not be identified unambiguously in a source, and even if there is an index (if the source is a book) there may not be an entry for the kind of information sought. And, of course, articles and dissertations do not usually have indexes at all. So a particular kind of information may not be easy to find just by looking through a document, because different documents may refer to what is being looked for in different ways, using different words or under different headings. The HRAF indexing system solves the problem of searching for data in texts that do not employ a standard vocabulary. Using the OCM index categories enables one to find just about any kind of information in seconds, because the categories bring the user to the relevant passages even if the original authors used different words and headings. There is

also an index to the OCM categories; this large list of words points to the relevant OCM categories for finding particular kinds of information.

For example, if you were interested in assessing the degree to which various cultures depend on stored foods, you might search the index to the OCM, where you would find 'Preservation, of food, 251.' This directs you to OCM category 251, 'Preservation and Storage of Food.' Searching the database by OCM 251 will bring you to all of the passages that describe dried, smoked, pickled, refrigerated, frozen, canned, and irradiated foods, and whatever other ways are used to store or preserve food. The analysts at HRAF, who have read through and indexed every page of every text that goes into the HRAF files, have made it possible to find the relevant information, even when you cannot anticipate the particular words (including untranslated native words) the original authors may have used.

The HRAF collections are indexed by subject, but not coded by variable. This is deliberate. There are many kinds of potential comparisons and therefore no one system of coding would suit every user. The OCM indexing system gets the user directly to the text passages that are relevant, whatever the comparative purpose and whatever the information of interest. The eHRAF collections can also be searched by the words that actually appear in the texts. In eHRAF, word searches (including proximity and Boolean searches) are possible with or without specifying OCM subject categories. In any search, the context of the hits is never lost. The sections before and after the retrieved paragraphs can be read easily.

The OCM indexing system is a flexible search tool. The user is always in control, because it is the user who decides exactly how to search and exactly how to deal with the information retrieved. Although it is frequently believed that the qualitative nature of most ethnographic description precludes coding and quantitative measurement, this is not true. It is not difficult, after a little practice, to develop at least ordinal scales that can allow the user to code the words of ethnography into measures, and once done it is easy to use available software to test hypotheses, and compare, combine, and model rival hypotheses. Needless to say, the indexed texts in HRAF are also amenable to qualitative cross-cultural comparisons.

Installment 48 of the HRAF *Collection of Ethnography* (available in 2000) is the sixth annual installment of *eHRAF: Ethnography*. It completes the updating and electronic conversion of the 60-culture Probability Sample Files (PSF) that were built in the 1970s and 1980s. Raoul Naroll planned how the PSF was built. Quality control considerations were used to select a large list of eligible preindustrial and peasant cultures, and then one culture from each of 60 culture areas was randomly selected. Thus, researchers can use the PSF to test hypotheses on a fairly large and unbiased sample of the world's preindustrial and peasant cultures. Correlations and other statistical results are likely to be trustworthy and functional, not due to duplications in the sample because of random diffusion or common ancestry. From 2000 on, *eHRAF: Ethnography* will grow by adding other randomly selected cases (as well as additional contemporary ethnic groups in the USA and elsewhere). The cases will be drawn by simple random sampling from a new sampling frame for the world that has been constructed by HRAF with the aid of external experts on the various regions. The new sampling frame will eventually include national cultures as well as immigrant cultures all over the world.

The results of comparative studies using ethnography may suggest causality, but social scientists would like to go beyond cross-sectional validation. In particular, they would like to see if temporal sequences validate causal theories. A cross-cultural study using ethnography may be able to discover predictors, but it cannot by itself confirm the temporal sequence assumed in a causal theory. Of course, one could try to measure the ethnographic cases for two different points in time. Measuring each case twice may be increasingly possible as the ethnographic record grows and includes more and more cultures that have been re-studied. But for many questions about cultural evolution, the ethnographic (or ethnohistorical) record is unlikely to provide the required time-series data. This dilemma particularly applies to the classical questions about human cultural evolution. How could we test answers to questions about the emergence of agriculture, the rise of social inequality and the first cities, and the origins of the state?

Now there is a research tool that allows us to escape this dilemma. Investigators of cultural evolution can use *eHRAF: Archaeology* to study and model causal sequences. Cross-cultural (comparative ethnographic) studies can provide archaeological indicators of cultural and other (e.g., physical and social environmental) features. Using those indicators, researchers can test many causal ideas about the major events in cultural evolution on the time-series data in the archaeological record. Thus, the data in *eHRAF: Archaeology* can allow researchers to determine whether evolutionary patterns in one region are repeated in others, and to determine whether the presumed causal factors in one region are also important, and antecedent, in other world regions. Comparative ethnography can tell us about cultural statics, what is associated with what in the ethnographic record. Comparative archaeology can tell us about cultural dynamics, what precedes what in the archaeological record.

Compared with conventional library research, retrieval of information from the electronic (eHRAF) collections is like a day compared to a month. When comparing or measuring things about a number of cultures, it is not necessary to devote weeks to building bibliographies for each of your sample cases; to search

for the relevant books and articles and dissertations, which might have to be obtained by interlibrary loan; to read every page of a source that does not have an index to find all of the passages with relevant information; or to bring any sources back to the library. If the researcher's institution belongs to the HRAF consortium, and the culture(s) of interest are included in eHRAF, it is easy to access the text passages that contain the particular kinds of information sought.

2. HRAF and Useful Social Science

The usefulness of social science, the possibility of social engineering or applying knowledge to solve social problems, depends on the validity of social science findings and theories. If a finding or theory is not true under some circumstances, it would be foolhardy to think of applying it to real world situations and problems. Hence there seems to be a growing awareness that social science risks being useless, even parochial, if it does not apply to many times and places. The more a finding or theory fits the world generally, the better for understanding and application. Culture really does make a difference. People in different times and places behave differently at least partly because they have different repertoires of customary ideas and practices. So cultural variation cannot be ignored by social scientists if they want to discover principles and relationships that are universal, which is probably why introductory textbooks in psychology, sociology, political science, and international relations, as well as in anthropology, are now paying more attention to the results of cross-cultural research.

In addition, more and more types of investigators—not just anthropologists—are now doing worldwide cross-cultural research. Psychologists, political scientists, sociologists, and evolutionary biologists now use HRAF to test universal hypotheses. And researchers are now using HRAF to test hypotheses about many new kinds of human variable—not just the standard cultural observables such as where couples typically live after they marry or how people get their food. The new kinds of variable include: psychological variables that are measured from projective materials such as folklore; population (biological) variables that are measured using information found in ethnographic and archaeological reports; and comparative archaeological variables that are measured on the basis of ethnographic data that have archaeological correlations. There are many kinds of information in the ethnographic and archaeological records that can be used to construct and test theories.

Of all the research strategies for testing theory, worldwide cross-cultural research is the strategy that most ensures the generalizability of results. To be sure, not all theories can be tested on data from the ethnographic and archaeological records. If something is not widely described, we cannot measure it. Hence social scientists also have to do comparative field studies, cross-historical and cross-national studies, and simulations to test theory. But if we want to maximize the effectiveness of applied social science, we need to try to see if our theories survive worldwide cross-cultural tests. Why should theories that have been cross-culturally tested and supported be more likely to suggest effective solutions to social problems? The logic of statistical inference tells us. A theory that seems to fit a particular region or a sample of nations may not be true for human societies generally. There is no way to tell without cross-cultural testing. A study in one or a few countries may suggest an explanation or application. So may computer simulations and modeling. But ultimately we have to see if the data in the real world are consistent with what we hypothesize. This is why HRAF was invented in the first place, to enable scientists to test their ideas about cultural variation and evolution on a worldwide basis. How often could an explanation that is not universally valid suggest an application or solution that is generally effective?

See also: Databases, core: anthropology, museums; Databases, core: linguistics, on-line grammars and dictionaries; Databases, core: linguistics, corpora; Databases, core: sociology

Bibliography

Ember C R, Ember M 1998 Cross-cultural research. In: Bernard H R (ed.) *Handbook of Methods in Cultural Anthropology*. AltaMira, Walnut Creek, CA, pp. 647–87

Ember C R, Ember M 2001 *Cross-cultural Research Methods*. Alta Mira, Walnut Creek, CA

Ember M 1997 Evolution of the Human Relations Area Files. *Cross-Cultural Research* **31**: 3–15

Ember M 2000 *HRAF at the Millennium: Blueprint for the Future*. Human Relations Area Files, Inc., New Haven, CT

Ford C S 1970 *Human Relations Area Files: 1949–1969, a Twenty-Year Report*. Human Relations Area Files, Inc., New Haven, CT

Murdock G P, Ford C S, Hudson A E, Kennedy R, Simmons L W, Whiting J W M 1987 *Outline of Cultural Materials*, 5th rev. edn. Human Relations Area Files, Inc., New Haven, CT

Naroll R 1967 The proposed HRAF probability sample. *Behavior Science Notes* **2**: 70–80

M. Ember

Databases, Core: Anthropology and Museums

Anthropological museum collections—including both ethnographic and archaeological materials—comprise an essential source of data for understanding the dynamic uses of material culture, relationships be-

tween human behavior and its material correlates, long-term social processes, and documentation of culture history and change. With the rapid spread of western material culture and technologies, museum holdings have assumed a greater prominence for studies of non-Western lifeways and cultural forms. Because museum holdings provide both longitudinal data documenting cultural change through time, and also synchronic data showing diversity over space, their potential for social scientific research is considerable. This article provides a brief overview of collections and related core databases in anthropology museums, and their associated problems and prospects.

Many different kinds of museums hold significant anthropological collections; these institutions vary in terms of mission, authority, constituency, and level of physical resources available to support collections care and access. Most museum collections were acquired over a considerable period of time, informed by a variety of interests and acquisition strategies, and supported by varying levels of documentation and conservation. While this diversity insures that collections are not limited by the scope of collecting strategies or research interests dominating a particular discipline or period, it also results in a multiplicity of cataloguing systems, descriptive terms, and standards for care, access, and interpretation.

No reliable estimate of the scale of the resource exists, but available collections are enormous, and growing in both volume and significance. The Ethnography Department of the British Museum holds some 350,000 catalogued items from around the world, Vienna's Museum für Völkerkunde 200,000 items, and the National Museum of Ethnography (Minpaku) in Japan some 221,000 items. The scale of archaeological collections is still larger; the US Department of the Interior alone lists more than 30 million items (most archaeological) in its holdings, while the Canadian Museum of Civilization maintains some 300,000 ethnological artifacts and more than 2.5 million archaeological specimens. Extensive resources focus on specific areas of the world. Often these are based on the museum's immediate region, such as Mesa Verde National Park's two million specimens, most from the Wetherill Mesa Archaeological Project, or the Bishop Museum's 1.2 million Pacific and Hawaiian cultural objects. In other cases, however, regionally focused collections are half a world away. Brussels' Koninklijk Museum voor Midden-Afrika, for example, holds some quarter million artifacts from central Africa, along with extensive archival material.

As yet, no single, standard nomenclature has been adopted for categorizing anthropological resources. Most museums categorize collections by ethnic origin (e.g., organized by Murdock's *Ethnographic Atlas* or *Outline of World Cultures*), material type, intended function (e.g., the Chenhall nomenclature for man-made objects), geographic locality (e.g., by site) or time period (e.g., Bartoloni's *Dizionari Terminologici de Materiali dell'eta del Bronzo finale e della prima eta del Ferro* of 1980). Historically, the ability of scholars to access collection information was constrained by cataloguing terminology and data model limitations intrinsic to each approach. Consider the Chenhall nomenclature, for example. A search for the functional set 'skiing gear' would require searches within 'sports equipment' (snow skis), 'clothing—footwear' (ski boots), 'LTE land transportation equipment—human powered' (ski poles), and 'personal gear' (goggles) (Welsh 1990). Conversion of catalogues to electronic form has in some cases lessened these limitations, but introduced new concerns regarding hierarchical terminologies, data integrity, and long-term structural implications and limitations of extant data models. In many areas of the world core catalogue data remains incomplete or inaccessible; AFRICOM, for instance, is an international initiative to document collections in African museums using a common set of descriptors, and to develop common resources and standards for collection documentation and care (see also Roberts 1990).

The availability of information regarding collections varies greatly. Collections at the Chaco National Historical Park are listed in both a published guide and a computerized database available to researchers listing some 65,000 catalogue records, representing more than 1.5 million catalogued items. Such levels of access are, however, the exception rather than the rule. Lacking meaningful data standards, access to collections usually requires either consultation with appropriate staff at each institution, or use of published guides to available collections, often specific to individual collections within a single institution. A draft standard for anthropological metadata (data about data, or a meaningful data model standardizing usage across data sets) has been proposed through the Council for the Preservation of Anthropological Records (CoPAR), although databases using this standard are still in preparation.

Online resources and other emerging technologies promise some relief. Australian resources, for example, are summarized by online directories of museum collections (Australian Museums Online), archives (both the Directory of Archives in Australia and the Register of Australian Archives and Manuscripts, supplemented by the item-level database ANGAM II), and social science resources relating to Asia and the Pacific (through the efforts of the Australian National University's Research School for Social Sciences and Research School for Asia and the Pacific). In practice, however, these are very broad summaries that provide limited access to key collections data.

Other kinds of anthropological resources also support social science research. Many national or regional authorities maintain records documenting the number, location, age, and character of archaeological

sites within their jurisdiction. Denmark's Kulturhistoriske Centralregister, for instance, has recorded more than 140,000 archaeological sites, with a schedule of 55,000 monuments (not including buildings), and the US State of New Mexico currently lists more than 100,000 recorded sites in its automated ARMS site registration system. In addition, international organizations promote the documentation of sites and monuments (e.g., ICOMOS, the International Council on Monuments and Sites), and have taken first steps toward establishing metadata standards.

Field records, notes, and papers—valuable for adding research value to the catalogued physical objects and related documentation held by museums—are often not preserved as archives, although there are exceptions. The British Library holds the papers of many leading anthropologists, including Edward B. Tylor and James G. Frazer (although the extent of anthropological holdings is not recorded separately), the American Philosophical Society maintains some 6.25 million manuscripts with strengths in American ethnology and linguistics, including the papers of Franz Boas, and the National Anthropological Archives, part of the Smithsonian Institution, preserve more than 7,000 linear feet of primary documents, papers, and photographs for anthropological studies worldwide. Many universities maintain archives of the papers of past faculty, although these are rarely integrated or cross-referenced with museum holdings. Photographs and other images have become increasingly valuable in many areas of research. The Musee de l'Homme in Paris holds more than 400,000 black and white photographs, Leiden's Rijksmuseum voor Volkenkunde 300,000 images and related materials, Harvard's Peabody Museum some 500,000 images of all kinds, and the Smithsonian's Human Studies Film Archives more than eight million feet of film and video. Surveys of extant resources (e.g., Dutton 1999), while valuable, are limited in scope and uneven in geographic coverage.

While cultural resource management (CRM) represents one of the most rapidly growing sources of anthropological collections, most data generated from heritage management or CRM studies are reported in unpublished, limited-distribution 'gray literature' reports to regulatory agencies. Attempts are being made currently to standardize access to this literature; the National Archaeological Database (NADB), for example, provides bibliographic control for some quarter-million archaeological planning and management documents for projects within the US, and the Archaeology Data Service provides access to a series of related databases on archaeology and cultural resources in the British Isles, including the 48,000 records in the Excavation Index for England. Metadata standards for the online publishing of field reports are in development (e.g., UCLA's Digital Imprint initiative), but databases based on these standards are not yet available.

For the most part, the value of anthropological collections depends on the preservation of the objects themselves and the associated contextual data necessary to place them within a particular social, temporal, and spatial setting. Threats to the long-term integrity of anthropological collections include unaddressed conservation needs, changing legal and ethical requirements affecting access and treatment, security issues, data integrity concerns, and the cost (in terms of both infrastructure and cash outlay) of adequately housing and curating ethnological and archaeological materials.

The management and conservation of anthropological collections is difficult, however, as many items of human material culture are complex artifacts including different materials requiring different kinds of conservation regimes for their long-term stability. Physical threats to collections come from a number of sources, including: (a) inherent vice (including flaws and instability due to the material from which artifacts were made); (b) improper storage materials or conditions, including acidic or otherwise reactive storage media and off-gassing of cabinets; (c) improper environmental controls, particularly for humidity and temperature; (d) physical damage through neglect, vandalism, and effects of warfare or natural disaster; (e) theft; and (g) damage from pests. In some cases addressing these concerns leads to new ones. In the past some collections were treated with arsenic and other toxins to protect them against pests; these treatments now represent a threat to both the collections and to museum staff and visitors.

In other cases consolidants or other conservation treatments used to stabilize items may now interfere with more recent chemical studies of artifacts and residues. In addition, inadequate curation or collections management practices can result in loss of artifacts, associated documentation, or the link between the two. A more immediate threat to collections care is the relatively high cost in money, space, and time required to adequately curate and conserve anthropological materials, leading to the selective curation of collections and disposal of 'less valuable' objects or sets of objects.

Stewardship remains the most important responsibility of professional museum staff; this is particularly true in anthropology, where curators hold collections in trust, not only for posterity, but in many cases for specific descendant or ethnically-related groups as well. In addition to physical care and monitoring by conservators, many items require other kinds of maintenance mandated by native peoples. The Museum of Indian Arts and Culture in Santa Fe, New Mexico does not fully seal cases containing kachina dolls, because they 'need to breathe,' and the storage facilities for the National Museum of the American Indian, part of the Smithsonian Institution, include a smoking room for the blessing of material objects by native groups. In some cases the intent of the makers

is at odds with physical conservation concerns; the celebrated Zuni War Gods—that were created to be exposed to the elements and gradually decay—are one example.

Many kinds of collection remain subject to claims for repatriation, either by countries of origin or ethnic groups. Several countries have laws requiring repatriation of certain classes of material to native or aboriginal groups. Whether this represents a long-term threat to the integrity of these collections, or merely a long-overdue reallocation of rights and control to native peoples, remains to be seen.

Despite these concerns, there are many encouraging trends in museum anthropology. These include innovative research projects harnessing old data to address new questions, increasing professionalization of the field, recruitment of native peoples to museum staffs, especially as curators, widespread recognition of the need for common metadata standards and improved access to collections information, substantive research on appropriate collections management and conservation procedures, and a growing recognition of both the value of anthropological materials as a unique and irreplaceable resource, as well as an appreciation that they are a fragile legacy requiring the highest standards of management and care.

Anthropological museum collections remain one of the most valuable and underutilized sources for social scientific and behavioral research. As one of the most important means of documenting the material remains, histories, and social trajectories of cultures both extant and extinct, anthropological museum collections represent a comparative database of unparalleled potential—and one of steadily increasing scientific and social significance.

See also: Archival Methods; Archives and Historical Databases; Databases, Core: Sociology

Bibliography

Bartoloni G 1980 *Materiali dell'eta del Bronzo finale e della prima eta del Ferro. Dizinionari Terminologici 1*. Centro Di, Florence, Italy

Dutton L S 1999 *Anthropological Resources: A Guide to Archival, Library and Museum Collections*. Garland, New York, Vol. 864

Hansen H J 1993 European archaeological databases: Problems and prospects. In: Andreasen J, Madsen T, Scollar I (eds.) *Computing the Past: Computer Applications and Quantitative Methods in Archaeology*. Aarhus University Press, Aarhus, Denmark

Hansen H J, Quine G (eds.) 1999 *Our Fragile Heritage: Documenting the Past for the Future*. National Museum of Denmark, Copenhagen, Denmark

Larsen C U 1992 *Sites & Monuments*. National Museum of Denmark, Copenhagen, Denmark

Roberts D (ed.) 1990 *Terminology for Museums: Proceedings of an International Conference*. Maney, Leeds, UK

Welsh P H 1990 An approach to terminology from the perspective of archaeology, anthropology and ethnography. In: Roberts D A (ed.) *Terminology for Museums: Proceedings of an International Conference*. Maney, Leeds, UK

A. W. Barker

Databases, Core: Demography and Genealogies

Genealogy, according to the Oxford English Dictionary, is defined as an account of one's descent from an ancestor or ancestors by enumeration of the intermediate persons; the investigation of family pedigrees as a department of study or knowledge; lineage, pedigree, family stock; and progeny. In this article, as indicated by its title, the term genealogy is particularly used to refer to the written record of family or lineage members and the discussion is focused on genealogical records as a data source for demographic research.

1. Genealogies

Constructing genealogies has a long history. In the early days of civilization and before written records were made, oral transmission of genealogical information already existed, although this may include only a list of names and the relationships between these people.

The invention of writing and the improvement in keeping written records allowed genealogies to be easily made and better preserved. During early history, however, they were largely compiled for royal and noble lines. Surviving genealogies concerning past ruling and upper class families can still be found in many countries. This is related to the fact that in a society of hereditary monarchy or aristocracy, genealogies were frequently used in determining one's inheritance of office or land, and in establishing a link to a superior social status.

In China, for example, compiling genealogies reached one of the high points of its development during *Wei Jin Nan Bei Chao* (220–589 AD), a period when genealogical information was widely consulted in recruiting government officials and arranging marriages. In addition to the genealogical activities of the individual family or lineage, government genealogical bureaus were set up and responsible for preserving, editing lineage registers, and constructing multilineage genealogies (Pan 1929).

This upper class practice, nonetheless, had a trickle-down effect. The number of genealogies compiled by or for ordinary families or lineages increased during the last few centuries. This was particularly the case in the society where lineage organizations formed an important social institution and the worship of

ancestors was widely encouraged, and in the new frontier where a large number of migrants settled.

The twentieth century has witnessed a continuous increase in genealogical activities throughout the world. Many genealogical societies have been formed and numerous genealogies have been compiled. Recent development in computer technology has brought about a further surge in the study of family history. This has led to an explosion in the number of genealogical databases accessible through the Internet, with some of them providing biographical information for millions of people. Genealogies have increasingly become an important data source for scientific research.

2. Genealogies and Some Other Demographic Data Sources

Like population and vital registers, census and survey data, and church records, genealogies can also provide valuable information for demographic research. However, they differ from the above data sources considerably.

People recorded in genealogies normally have common ancestry and are connected to each other through family relationships, while those included in other data sources are generally not. This gives genealogies a unique character and makes them extremely useful in the study of population genetics, heterogeneity, and social mobility of the family. Indeed, genealogies have already been used by historians, demographers, epidemiologists, and geneticists. Their endeavor has produced some very interesting results (Hollingsworth 1964, Knodel 1988, Bean 1990, Liu 1992).

Population and vital registers and census and survey data are often obtained from a precisely defined territory. Church records also tend to be concerned with religious groups of a certain area, whereas people recorded in genealogies are less, or not at all, restricted by their residence. Accordingly, genealogies can reveal geographic movements of family members and migration patterns. However, people listed in the genealogies may not be the total population of any given region.

Most available parish records, census and survey data, and vital registrations were produced in the last few centuries, and can be used only for the study of recent demographic history. Population counts of earlier periods have been found in some countries, but their reliability is difficult to assess because of the lack of original enumeration and other relevant materials. Furthermore, aggregate figures of this kind do not permit any detailed examination of past demographic regimes. In this context, genealogies covering a very long period may contain extremely useful information for the study of early population history. Some Chinese genealogies, for instance, consist of family or lineage members of more than 80 generations living over a period of more than 2000 years. This gives scholars a rare opportunity to investigate long-term demographic patterns.

Noticeable differences also exist in the process of producing these records. Vital registrations and parish registers are often made soon after certain demographic events take place. Censuses and surveys are also conducted within a well-defined time frame. In collecting these data, rigid rules and instructions are usually followed. In contrast, constructing genealogies can be a series of activities undertaken by people of many generations under different circumstances over a long period. The rules of compiling genealogies or the implementation of such rules may change from time to time. The interval between the time when a certain social demographic event actually happens and when it is recorded by the genealogy compiler can be very long. These characteristics, though to some extent are shared by retrospective and longitudinal surveys, make the genealogies different from other demographic data sources.

3. Descendant Genealogies and Ascendant Genealogies

Conceptually, genealogies can be constructed in two ways. Accordingly, they can be divided into descendant genealogies and ascendant genealogies (Oeppen 1999).

Constructing a descendant genealogy starts with a person (or a group of people) who may be regarded as the founder (or founders) of the family or lineage. Then, descendants of this person (or these people), if there are any, are added into the genealogy generation by generation. Compilation of the genealogy is not conditional on the availability and the survivorship of the members of the descendant generation. When genealogies are produced in this downward manner, the process of accumulating records can be very long. The genealogy may form a primary data source that is derived from direct enumeration.

In contrast, compiling an ascendant genealogy starts with a person of a later generation (the person is often the genealogy compiler). Then his or her ancestors are traced retrospectively generation by generation and entered into the genealogy. When genealogies are constructed in this upward fashion, the recorded family and lineage are those having survived to that later generation. The entries of the ancestors are most likely to be the outcome of genealogical research based on searching church records, vital registrations, or other data sources. The genealogies, therefore, can be viewed as secondary data sources.

Although such a distinction can be made and an ascendant genealogy can be easily identified, descendant genealogies recorded completely in a downward manner are less common. The majority of the

genealogies are more likely to be a combination of the two. Whoever starts to construct a family or lineage genealogy, naturally, would like to trace his or her own family history and accordingly compiles an ascendant genealogy. Thereafter, the compiler may add his or her children, grandchildren, and descendants of the earliest identifiable family or lineage member into the record. Alternatively, the work may be continued by the future generation. As a result, the genealogy will increasingly become a descendant one.

The above distinction is important nevertheless. The extent to which a genealogy is retrospectively constructed and to which the lateral kin are included has a significant impact on whether the recorded family or lineage members can be regarded as an unbiased sample of the general population or a clearly defined substratum of the population.

4. *Genealogies as a Data Source for Demographic Research*

Genealogies have been used in population studies for many years, but they remain a data source of both promise and disappointment. On the one hand, the long time coverage of the record, the richness of the information, and the great potential of integrating genealogical, demographic, and genetic studies make the genealogy a unique research source. On the other hand, biases and under-registration problems make it difficult to use conventional demographic techniques to analyze genealogical data (Willigan and Lynch 1982, Telford et al. 1983, Fogel 1993).

Whether a genealogy is biased for demographic research is affected by the manner of its construction. Descendant genealogies are generally free from the impact of selection. But ascendant genealogies, especially those including only family members of the direct line or the succession, tend to exhibit lower mortality and higher fertility than the general population (Oeppen 1999). This arises from the fact that an ascendant genealogy requires each generation to have at least one member surviving to adulthood and producing at least one child. It is biased towards demographically successful families or lineages, which have a better than average chance of extending their lines to the future generation. For the same reason, nuptiality can also be biased if marriage serves as a precondition for childbearing. Thus, although an ascendant genealogy can document the demographic history of a family well, it may not be an unbiased data source for investigating of the demographic experience of the entire population.

Another difficulty in using genealogies for demographic research comes from under-registration. The amount of information contained in genealogies varies considerably. Some genealogies provide detailed records for all family or lineage members, while others show only names and the relationships between the recorded people. Even within the same genealogy, the completeness of the record can be very different.

One reason leading to such under-registration problems is that genealogies, regardless of the manner of their compilation, are usually constructed for purposes other than demographic research. Because of that, certain family members, for example, children who died young, female descendants who were often regarded as less important to the family in a patrilineal society, and those who brought disgrace to the family, are frequently excluded from the genealogies (Zhao 1994).

In addition to the deliberate omission of certain family members, genealogies suffer from distortions of other kinds. Genealogical records accumulated over a long period could have been affected by changes in the rules and practice of compiling genealogies. While some records may have been damaged or lost, other items and information could have been edited out. When records are gathered from other sources, they could inherit the biases existing in these data and the nominal linkage process is also subject to error. Producing complete and accurate genealogies, therefore, becomes even more difficult.

Knowing where biases come from and what have been or may have been missing is the first step in estimating their impact and using incomplete data. Although these deficiencies create great difficulties in using genealogical data for demographic research, they do not make it impossible. Indeed, considerable progress has been made in recent years. Demographic analysis of genealogical records has and will continue to provide further insights into past demographic process and contemporary population changes.

See also: Databases, Core: Demography and Registers; Family and Kinship, History of; Genealogy in Anthropology; Historical Demography; Kinship in Anthropology

Bibliography

Bean L L 1990 The Utah population database: Demographic and genetic convergence and divergence. In: Adams J, Lam D, Hermalin A, Smouse P (eds.) *Convergent Issues in Genetics and Demography*. Oxford University Press, New York

Fogel R W 1993 New sources and new techniques for the study of secular trends in nutritional status, health, mortality, and the process of aging. *Historical Methods* **26**: 5–43

Hollingsworth T H 1964 *Demography of the British Peerage*. Supplement to *Population Studies* **18**(2)

Knodel J E 1988 *Demographic Behavior in the Past*. Cambridge University Press, Cambridge, UK

Liu T 1992 *Lineage Population and Socio-economic Changes in the Ming-Ch'ing Periods*. The Institute of Economics Academia Sinica, Taipei

Oeppen J 1999 Genealogies as a source for demographic studies: some estimates of bias. *Paper presented at the Workshop on Genes, Genealogies and Longevity*. Max Planck Institute for Demographic Research, Rostock

Pan G 1929 Zhong guo jia pu xue lue shi [A brief history of Chinese genealogical studies]. *Dong Fang Za Zhi [The Eastern Miscellany]* **26**(1): 107–20

Telford T A, Thatcher M P, Yang B P N 1983 *Chinese Genealogies at the Genealogical Society of Utah: An Annotated Bibliography*. Ch'eng Wen ch'u pan she, Taipei

Willigan J D, Lynch K A 1982 *Sources and Methods of Historical Demography*. Academic Press, New York

Zhao Z 1994 Demographic conditions and multigenerational households in Chinese history. Results from genealogical research and microsimulation. *Population Studies* **48**: 413–25

Z. Zhao

Databases, Core: Demography and Registers

Population registers are continuous records of the residential context, socioeconomic status, and demographic behavior of the members of a community. They exist for a number of European and Asian locations from as early as the seventeenth century. Though some countries still maintain them, because of confidentiality issues only data from before the middle of the twentieth century appear to be available for public use. Population registers combine features of most of the other sources commonly used in studies of population, family, and society in the past. In contrast with many of these sources, they are especially well suited to the application of event history techniques to study the determinants of demographic and other outcomes. Though they have some limitations, they are nevertheless an important source for the study of social demography, family history, and social mobility in the past.

1. Contents

Population registers, also referred to as household registers, are longitudinal, nominative records of the households and individuals that make up a community. They follow households through time, recording changes in their composition and other characteristics. Registers also follow individuals, recording when and by what means they enter and exit the household or community. They usually record their relationship to the head of the household. In most cases, they measure socioeconomic status at either the household or individual level. European registers, for example, often record the occupation of the household head. In a rural setting they may provide details on landholding, including type of tenancy.

Authorities originally established population registers for diverse reasons. In China, where systematic population registration has perhaps the longest history, successive governments used registers for centuries as tools for administration. In parts of Europe during the nineteenth century, civil authorities also introduced them as a tool for administration. In Japan, the state originally established a nationwide system of local population registers in the late seventeenth century to prevent the spread of Christianity, though of course it quickly found the registers useful for more routine tasks (Cornell and Hayami 1986). In some places, religious authorities maintained population registers. For example, parish priests in Sweden during the nineteenth century compiled catechetical registers every year that listed households and their members to identify candidates for literacy exams. In Italy, churches compiled *status animarum* every year that listed all of the households in the parish along with their members (Manfredini 1996).

Even though registers had diverse origins and purposes, they are broadly similar in terms of format. Almost all of them are organized in one of two ways. Many of the systems established in continental Europe during the nineteenth century were continuous. A register was created that contained a separate entry for each household in a community. Once created, often from the results of a census, it was updated on a continual basis. As events such as births, marriages, occupation changes, migrations, and deaths occurred in a household, annotations were made to its entry. Once annotations accumulated to the point where further updates were impractical, the register was retired and a new one was created containing only current information.

In the remaining locations, mainly in Asia but also in some parts of Europe, households in a community were recorded in census-like fashion at fixed intervals, most commonly annually, but in some cases less frequently. Copying information from the previous register and updating it produced a new register. Entries commonly included additional annotations indicating whether or not demographic events such as birth, migration, marriage, or death had occurred since the last update. Since households and their members were usually listed in roughly the same order in successive updates, entries for individuals and households are easily linked from one register to the next. Demographic events and other changes in status not specifically annotated can be inferred by comparison of entries in successive registers.

Historical studies that use population registers often supplement them with data from other sources. In most European settings, it is possible to use names, ages, and other identifying information to link to baptism, marriage, and burial records in parish registers and obtain additional information about the timing of demographic events. In some locations, links can be made to vital registration data, including birth, death, and marriage certificates. Such linkage can turn up vital events that the population registers missed. In some cases, it is also possible to link to entries in land

and tax registers, yielding richer detail on socioeconomic status.

Population registers combine features of the sources most commonly used to study population and society in the past. Like both genealogies and family reconstitutions from parish registers, population registers are inherently longitudinal. They follow individuals through time, allowing for the reconstruction of life histories. Like parish registers, population registers usually provide a nearly complete record of the vital events occurring in a community. Like censuses, registers describe entire communities, identifying who was present and at risk of experiencing a demographic event or other transition. They also provide details on household characteristics, including composition, socioeconomic status, and members' relationships to each other.

Registers accordingly overcome many of these sources' limitations and allow for new avenues of inquiry. In contrast with genealogies, they cover entire communities, not selected families. Whereas genealogies are usually retrospective, compiled based on the recollections of family members, registers are prospective. In contrast with family reconstitutions and genealogies, registers describe residential context, not just kinship. Finally, unlike most censuses, registers allow for the direct calculation of demographic and other rates because they are longitudinal, and include information on demographic events and other transitions.

2. *Applications*

Registers may be used to examine the determinants of marriage, childbearing, remarriage, and death because like panel studies they allow use of event-history techniques (see *Event History Analysis: Applications*) that measure the effects of time-varying community, household, and individual characteristics on the chances of these and other events (Alter 1988). Registers can also be used to illuminate the links between economic conditions and demographic behavior. Whereas previous studies have been limited to correlating prices and aggregate vital rates at the regional or national level, application of combined time-series (see *Time Series: General*) and event-history techniques to register data allows comparison of the price responses of different subgroups of the population, yielding new insights underlying causal mechanisms (Bengtsson 1993).

Register data lend themselves to the study of household dynamics. Classic studies of historical households were limited to examinations of size and structure by their reliance on cross-sectional census data. Registers, by contrast, follow changes in the structure and composition of households over time. Indeed, one of the first applications of register data was in a study of household dynamics in mid-nineteenth century Belgium (Van de Walle 1976). From register data it is possible to compute the rates at which households with different characteristics dissolve, disappear, or change structure.

Topics such as migration, adoption, and occupational mobility that are difficult or impossible to study with traditional sources for historical demography and family history are ideal for analysis with registers. Studies of adoption have been carried out in several Asian locations, including Japan and Taiwan (Wolf and Huang 1980). As for migration, studies have been carried out in both Europe and Asia. They usually focus on the determinants of the timing of family migration and, especially in Europe, individual home-leaving. Examination of occupational mobility is also possible because through record linkage, it is straightforward to examine the relationship of children's statuses as adults to the statuses of their parents, grandparents, and other kin.

3. *Limitations*

Like all other historical sources, registers suffer from limitations. One of the most common is that they may fail to record short-term movements into and out of a community. While such problems are most common in systems that record the population at fixed intervals because they miss those who enter and exit again between updates, they also exist in continuous registration systems. In-migrants may never be recorded if they are only in the community for a few weeks or months because they leave or die soon after arriving. Similarly, those who reside in the area but who migrate out for a very short time may never be identified as having been away. Infants who die soon after they are born may leave no record of their existence, leading to underestimates of the level of mortality in the first months of life.

Where service migration was part of the life cycle, older adolescents and young adults are often poorly covered from the time they leave their parent's household to the time they marry and form one of their own. If they remain in the area, they are difficult to link to their natal households. Therefore, it is hard to tell whether young persons who leave their natal households are still in the area but living and working in other households, or have left the area completely. For this reason, studies in settings where service migration was common often exclude from consideration older teenagers and young adults who have never married.

Demographic and other events that occur to residents when they are outside the community are unlikely to be recorded. This is not a serious problem, however, because population registers also allow individuals who have left the community to be excluded from consideration starting from the time of their departure. The caveat remains that if the individuals who left the community were selected on some characteristic that affected the chances of the event of

interest, rates computed for those who remained might differ from the ones that would be observed if no one left.

4. Availability

In Asia, historical registers for China, Korea, and Japan are available. Within China, there are at least three types of registers. In Taiwan province, Japanese colonial authorities established a registration system that generated data of extremely high quality from 1895 to 1945 (Wolf and Huang 1980). In northeast China from the mid-eighteenth century to the beginning of the twentieth, triennial registers recorded the residents of state farm systems in great detail (Lee and Campbell 1997). Registers from what is known as the *baojia* system exist for other parts of the country in the eighteenth and nineteenth centuries, but they appear to be of very poor quality (Skinner 1987). In Japan, there are annually updated population registers for many villages from the mid-seventeenth century to the mid-nineteenth century (Cornell and Hayami 1986). In Korea, triennial registers remarkably similar in format to the northeast Chinese ones exist.

In Europe, registers exist for parts of Belgium, Italy, The Netherlands, Sweden, and elsewhere during the eighteenth and nineteenth centuries. Belgian registers are continuous (Alter 1997), as are those from The Netherlands, where there is now an effort underway to use them to compile a nationally representative sample of individuals (Mandemakers 1994). Italian registers, consisting of annual *status animarum*, exist for a number of parishes (Manfredini 1996). In at least one Italian City, Venice, detailed data from a continuous registration system that operated during the middle of the nineteenth century also exist (Derosas 1989). In Sweden, catechetical registers exist for the nineteenth century. Some of these data have already been transcribed into databases at the Demographic Database at Umeå University.

See also: Databases, Core: Demography and Genealogies; Historical Demography; Statistical Systems: Censuses of Population

Bibliography

Alter G 1988 *Family and the Female Life Course: The Women of Verviers, Belgium, 1849–1880*. University of Wisconsin Press, Madison, WI

Bengtsson T 1993 Combined time-series and life event analysis. The impact of economic fluctuations and air temperature on adult mortality by sex and occupation in a Swedish mining parish, 1757–1850. In: Reher D, Schofield S (eds.) *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK

Cornell L L, Hayami A 1986 The shumon-aratame-cho: Japan's population registers. *Journal of Family History* **11**: 311–28

Derosas R 1989 A database for the study of the Italian population registers. *Historical Social Research/Historische Sozialforschung* **52**: 59–65

Lee J Z, Campbell C D 1997 *Fate and Fortune in Rural China: Social Organization and Population Behavior in Liaoning, 1774–1873*. Cambridge University Press, Cambridge, UK

Mandemakers K 1994 Historical sample of the population of the Netherlands (HSN). Backgrounds, objectives and international context. In: Marker H J, Pagh H (eds.) *Yesterday. Proceedings from the 6th international conference Association of History and Computing*, Odense 1991. Odense University Press, Odense, Denmark

Manfredini M 1996 L'utilizzo degli Status Animarum nellerico-struzioni nominative: miglioramenti informativi qualitativi e quantitativi. Il caso di Madregolo (1629–1914) [The use of Status Animarum in nominative reconstruction: improvements from qualitative and quantitative information]. *Bollettino di Demografia Storica*. **24–25**: 113–29

Skinner G W 1987 Sichuan's population in the nineteenth century: lessons from disaggregated data. *Late Imperial China* **8**: 1–79

Van de Walle E 1976 Household dynamics in a Belgian village, 1847–1866. *Journal of Family History* **1**: 80–94

Wolf A P, Huang C H 1980 *Marriage and Adoption in China, 1845–1945*. Stanford University Press, Stanford, CA

C. Campbell

Databases, Core: Linguistics, Online Grammars, and Dictionaries

Before discussing online dictionaries and grammars, we need to define what these are and how they are produced (and used) in their 'non-online' form. A dictionary is a reference book used (a) to look up the *meaning* of a word or a phrase, in a monolingual dictionary, or (b) to look up the *translation* of a word or phrase in another language. For example, if you want to know the meaning of the word *blockbuster*, you would consult an English dictionary; alternatively, if you want to know the translation of this word into French, you would consult an English–French dictionary. A grammar is a description of a language. For the purpose of this article we shall discuss descriptive (or reference) grammars, i.e., books that describe the grammatical system of a language: the syntax, morphology, and phonology—but much of what we shall say also applies to teaching grammars, i.e., books that are intended as learning tools.

Grammars and dictionaries have a number of things in common. They try to explicate aspects of a language (meaning of words, explanations of the structure of the language) in a written form. They also refer to other parts internally; typically, dictionaries contain

cross-references to other parts of the same book to further elaborate the meaning, or aspects of the meaning of a word or a phrase. Reference grammars typically abound in cross-references to related issues discussed elsewhere in the book.

1. *Problems with Traditional Grammars and Dictionaries*

Because of their intended use as reference tools, dictionaries and grammars must be easily accessible. In the case of dictionaries, this is not a problem. After all, they list all the words alphabetically, so users will usually be able to find what they need. But sometimes searching for what you want is a bit more involved, especially looking up the meaning of an expression. For example, if you want to know the meaning of an expression such as *pull your socks up*, do you look under *pull* or under *socks*? In dictionaries of English, the entry for *pull* is quite big, so searching there is quite a task, so you might first want to look under *socks*. But if the expression is not listed there, you need to go through the columns of text where all meanings and uses of *pull* are described.

Another problem for explanatory monolingual dictionaries is that, in many cases, a word is explained more easily by a picture. Also most dictionaries contain several pictures (indeed, there are so-called picture dictionaries, in which virtually all entries are illustrated with a picture) but pictures also take up a relatively large amount of space, so the publisher needs to be as economical as possible and strike a balance between using words and pictures to illustrate words.

Grammars rely crucially on a good table of contents and a good index. After all, reference grammars are used to look things up: they are not intended to be read from cover to cover. Finding what you want is often a challenge in a grammar: via the table of contents and the index, users search and find the topic they wish to know more about, and are typically sent to other parts of the book as well via cross-references.

A practical point about both dictionaries and grammars is that, apart from giving the meanings of words and explanations about the structure of a language, they also try to give an idea how those words and sentences sound. Many dictionaries include, apart from the meaning or the translation of words, indications of how those words sound, and how they should be pronounced. Although there is a standard symbolic system for representing sounds (the alphabet devised by the International Phonetic Association (IPA)), few dictionaries actually use that, and instead use a system of the publisher's (or the editor's) invention. This is a strain on users, as they need to familiarize themselves with the symbols used; and as most dictionaries use different symbols to represent the same sounds, the situation can become quite confusing.

Grammars have this problem as well. Most descriptive grammars contain an account of the sound system of the described language—not only an inventory of the vowels and consonants, but also how these are combined. Academic grammars are usually standardized to use the IPA alphabet, but even such descriptions can be quite daunting.

Another practical problem that publishers of dictionaries and grammars face is that languages change: new words enter a language, others become obsolete; and words may change their meaning. Apart from this, errors need to be corrected and changing insights incorporated. This means that new editions of a dictionary need to be produced periodically. All this also applies to grammars. It is not only the words of a language that change, but the structure of every language—its syntax, morphology, and phonology—changes as well. And these changes, too, necessitate new editions of grammars.

Most of the problems outlined above can be overcome by online dictionaries and grammars, to which we now turn.

2. *Online Grammars and Dictionaries*

By 'online' we mean the World Wide Web (or 'the web,' for short), as this is the best-known online medium that most people will be familiar with. ('Online' can also be taken to include CD-ROM, but increasingly, publications on CD-ROM are identical to those on the web.) Documents can be presented on the web in two forms, as HTML (hypertext markup language) files, or as PDF (portable document format) files. The differences between these two can be ignored here, as it is the similarities that are important: both allow the same facilities, namely, searching, following built-in links, and playing sounds and displaying images. These three facilities are dealt with below. In the rest of this section, dictionaries and grammars will be discussed separately: dictionaries in Sect. 2.1, and grammars in Sect. 2.2.

2.1 *Online Dictionaries*

The main activity of a dictionary-user is searching for a word or an expression, and often, following up cross-references in the entries found, or looking up the meaning of a word in a description. For example, in the Concise Oxford Dictionary (COD), one of the explanations of the meaning of the entry *leg* is 'obeisance made by drawing back one leg and bending another'. If you do not know what *obeisance* means, you have to look up this word elsewhere in the dictionary. (Professional dictionary-users, such as

translators, benefit greatly from online dictionaries for precisely this reason.)

In an online dictionary, searching an entry is straightforward by using the search facilities offered by the browser, and following up references or looking up unknown words in an explanation (such as *obeisance* in the example given above) is speedy.

Online dictionaries shine in their multimedia capabilities. For example, dictionaries typically contain a phonetic representation that indicates how a word should be pronounced. (This is especially the case in English dictionaries as pronunciation is so unpredictable in English.) For example, the COD contains the entry **ăntėnā′tal**, and from the introduction to the dictionary we learn that ă sounds like the 'a' in 'fat'; ė like the 'i' in 'pill'; and ā like the 'a' in 'fate' (the prime symbol, ′, indicates stress). The circularity here, naturally, is that such explanations assume that the user knows how to pronounce 'fat', 'pill', and 'fate'. In an online dictionary, it is possible to link a sound file to each entry, such that when users click on that entry, they will hear the word pronounced. Such a system is exploited in the Interactive Introduction to Linguistics (http://www.hueber-vfd.de). In this CD-ROM-based system, the user can click on a word to hear how it is pronounced.

Especially when there are small differences in pronunciation, explanations in words can be tedious and ineffectual. A language such as English, which has little correspondence between spelling and pronunciation, can benefit from 'spoken' explanations. But other languages can profit from modern technologies as well. For example, Chinese is a language with lexical tone, which means that words with the same spelling can mean up to four totally different things, depending on the tone used. Tone is usually indicated using special symbols, but can be exemplified much better with the use of sound.

Another way in which the multimedia capabilities of the web can be exploited is in its use of pictures. We mentioned earlier that words in a language are often illustrated more easily with a picture than with words, but that publishers have to strike a balance between the use of pictures and words, mainly for reasons of space. With online dictionaries, however, space is hardly an issue. A publisher, then, can decide to add as many pictures as necessary.

2.2 Online Grammars

As is the case with dictionaries, online grammars are used for reference purposes, and therefore need to be easily accessible. We mentioned that traditional grammars depend heavily on the table of contents and the index. But even when the contents and the index are of good quality, finding what you want to know involves a lot of searching. Online grammars, meanwhile, are searchable very easily, thanks to the basic search facilities offered by all browsers. Online grammars, like traditional grammars, typically contain a table of contents and an index, but these are hyperlinked with the body of the text, so that users can jump immediately from the table of contents to the relevant chapter or section. Similarly, when the index entries are hyperlinked to the relevant occurrences in the text, these can be accessed easily through the hyperlinks. And finally, cross-references in the text can be looked up more easily.

Academic descriptive grammars (but non-academic ones as well) often contain a chapter on the phonology (the sound system) of the language, and one or more texts. The audio capabilities of the web make it possible to have the sounds reproduced through the computer's sound system, and can include the spoken version of the texts included in the grammar.

2.3 Other Advantages

Other advantages of online reference works over their 'paper' versions are of a more practical nature. As was mentioned above, languages change all the time, with new words entering the language and others becoming obsolete. With printed dictionaries, this means that new editions need to be produced, which is time-consuming and expensive: not only for the producer, but also for the user. Conversely, online dictionaries are relatively easy to update and maintain.

The relative ease of production is especially beneficial for grammars, which usually take a long time to write. With online grammars, material can be published earlier to make it available to the academic world. In the course of time, more material can be added. Naturally, this only applies to academic grammars, which can be of scientific importance even when not completely finished.

2.4 Disadvantages

Are there disadvantages to online dictionaries and grammars? There are, but these can be overcome. The first issue is quality. As most traditional grammars and dictionaries are published by reputable publishers, there is in most cases a guaranteed degree of editorial control and therefore a degree of quality. On the web, however, this is not necessarily the case: anyone can publish anything on the web, so if the material found here is not endorsed by a reputable publisher, users must be aware of the possible limited reliability.

Second, documents on the web can be updated literally every minute. We mentioned this earlier as an advantage, which in many ways it is, but there is a downside as well. The ease with which web documents

can be updated also makes them less stable. This is a problem especially for academic purposes, where everything that is cited in a publication is accounted for by a full reference (year of publication, edition, page numbers, etc.) so that other people can verify the correctness of cited material and cited claims. With texts on the web being subject to change, it becomes difficult to check the accuracy of a publication.

2.5 Where to Find Online Dictionaries and-Grammars

A good example of an online grammar can be found at http://doma.bicos.de/elevine/YuGrHTML/home.html, a grammar of Yukaghir (a language spoken in Russia) which is being developed on the World Wide Web. An interesting example of an online dictionary project can be found at http://coombs.anu.edu.au/WWWVLPages/AborigPages/LANG/GAMDICT/GAMDICT.HTM, a dictionary of Gamilaraay, a language spoken in Australia. The biggest online dictionary is the Oxford English Dictionary, at http://www.oed.com/public/publications/online.htm. And a useful start is http://www.facstaff.bucknell.edu/rbeard/index.html, which is an index of grammars and dictionaries available on the web.

See also: Computer Networking for Education; Human–Computer Interaction; Information Society

Bibliography

Butler C S 1992 *Computers and Written Text*. Blackwell, Oxford, UK

Corbeil J-C, Archambault A 1987 *The Facts on File: English/French Visual Dictionary*. Fact on File Publications, New York

Crystal D 1995 *The Cambridge Encyclopedia of the English Language*. Cambridge University Press, Cambridge, UK

Dalby A 1998 *A Guide to World Language Dictionaries*. Library Association Publishing, London

Sinclair J 1991 *Corpus, Concordance and Collection*. Oxford University Press, Oxford, UK

P. Kahrel

Databases, Core: Political Science and Political Behavior

1. Data and Databases

Data is information on properties of units of analysis. These units can be individuals (micro), organizations like business firms or political parties (meso) or nations (macro), they can be geographical divisions like cities, counties or states, or they can even be relations like elite networks. Often, units of analysis and units of observation are the same. However, sometimes units of observation, like all family members in a household, can be aggregated if the household is the unit of analysis.

Data in political science come about through a variety of means. They can be process-produced in an administrative environment like macro- or microcensuses for the purposes of societal self-observation and informed policy formation, they are created by scientists according to (more or less) methodologically sophisticated approaches of questioning individuals or other social units, or they are derived from inobtrusive techniques like content analyses of printed or video material as well as group observations. Information on a particular collection of units of observation or analysis placed in time and space is called a dataset. Correspondingly, databases are general or focused collections of datasets, as held, e.g., by data archives.

2. The Beginnings

In his history of the discipline of political science, Almond (1996) has pointed out that in the context of the professionalization of the field in the twentieth century, already in the first decades the empirical study of political phenomena had blossomed, most noteworthy in the Chicago School which is associated with such eminent scholars like Charles E. Merriam, Harold Gosnell, and Harold D. Lasswell. In this article it is emphasized, though, that a major breakthrough in the theory, methodology, and practice of empirical political science research was accomplished after the end of the Second World War. During the war, many of the later leading figures in the social sciences were employed in a variety of US government or related public and private organizations with the task to find answers to questions of critical military and civilian importance in the war context. One case in point is the series of seminal studies of *The American Soldier* directed by Samuel A. Stouffer.

The above war context in the long haul also led to the foundation of the Institute for Social Research (ISR) at the University of Michigan in Ann Arbor by scholars like Angus Campbell, George Katona, Rensis Likert, and Robert Kahn. In 1952, Campbell and Kahn published a small book on the US presidential election of 1948, and this is worth mentioning on various accounts. First, the first election study by the Michigan Survey Research Center (SRC) was one of the very few if not the only one which got it right in that its data showed Truman's win over Dewey. It was innovative in that it was national in scope, used a fairly sophisticated method for selecting respondents, applied a pre–post panel design and operated with a majority of standardized, but also some open-ended questions.

This study is singled out not only because it very well reflects the beginning theoretical and methodological innovations in research on political behavior, but also because it paved the way for one of the most demanding ventures in database construction: the American National Election Studies (NES).

Warren E. Miller (1994, pp. 249–50), in his recapitulation of the NES history, points out that since the outcome of the 1948 presidential election resulted in 'a public embarrassment to the national community of public opinion pollsters' (because of their falsely picking Dewey as the winner), the successful Michigan study gained a remarkable public and scholarly visibility, resulting in a grant to the Social Science Research Council (SSRC) which in turn commissioned the Survey Research Center to also conduct a study of the 1952 presidential election. With the intellectual success of this study (Campbell et al. 1954), the ground was prepared for a continuous series of Michigan-based election studies which finally were transferred in 1977 into the format of the national political science research resource of the American National Election Studies (Miller 1994, pp. 260–2) funded by the National Science Foundation.

The intellectual pull of the Michigan school of electoral research was felt not the least by European political scientists who especially in the 1950s and 1960s came to Ann Arbor to learn about the study of elections and about advanced methods in social research. It is because of this connection that in a fair number of European countries national election studies have by now become institutionalized (for a detailed account of this see Thomassen 1994). What makes this development so pertinent for the topic of this article is that the Michigan–European interaction helped to prepare the ground for the most important development in creating political science databases: the emergence of academic data archives.

3. *Data Archives—A Core Infrastructure for Political Science*

There is agreement in the literature that the oldest general-purpose survey data archive in the world is the Roper Center for Public Opinion Research (Bisco 1966, pp. 96–7). It was founded in 1946 at Williams College and is now located at the University of Connecticut in Storrs. While in many ways it has served as an example for other data archives founded later, the impact of the Inter-University Consortium for Political and Social Research (ICPSR; initially ICPR) at the ISR on the development of data archives in Europe has been more pronounced because it could build on an identity of substantive research interests between American and European scholars (see *Data Archives: International*).

In his recollection of how the ICPR came about, Miller (1989, pp. 152–3) mentions the intersection of two important strands of thought regarding the improvement of political research. The first was the positive impact on the development of at least parts of political science in the US from the Michigan election studies and, more concretely, from two summer seminars held at the SRC in 1954 and 1958 which were sponsored by the SSRC Political Behavior Committee.

'The attraction of the courses was the ability to analyze the Survey Research Center election surveys with state-of-the-art counter-sorters which were not available at most other campuses.' (Weisberg 1987, p. 282)

Second, in the interaction between the SRC and Stein Rokkan, one of the great European social scientists and comparativists, the SRC took up ideas from a Rokkan report to the Ford Foundation in 1957 proposing the establishment of survey data archives particularly to stimulate comparative research (Rokkan 1964, p. 1966).

The establishment of the ICPR in 1962 was preceded in time by that of the Zentralarchiv für Empirische Sozialforschung (ZA) at the University of Cologne, Germany. It was founded in 1960 by Guenther Schmoelders, a professor of public finance, and later directed by Erwin K. Scheuch, a major figure in international cooperation in the social sciences and in comparative research. Both ICPR and ZA were pioneers in academic data archiving and cooperated closely from their early days on. This innovation owed a lot also to the network of the Committee on Political Sociology of the International Political Science Association (IPSA) and of the International Sociological Association (ISA) which brought together scholars like Campbell, Philip E. Converse, Juan Linz, Seymour Martin Lipset, Richard L. Merritt, and Miller from the American and Hans Daalder, Mattei Dogan, S. N. Eisenstadt, Arend Lijphart, Giovanni Sartori, Scheuch, Otto Stammer, and Jerzy J. Wiatr from the European side.

The emergence of academic social science data archives reflects very much the state of empirical political science in the 1950s and 1960s. First, the methodology of social research and particularly survey research had been advanced to a point where it was regarded as a reliable, fruitful way of collecting information on a broad variety of social and political topics found interesting by the research community, but also by social and political elites as well as by the mass media (for example, see Lazarsfeld and Rosenberg 1955, Hyman 1955). However, the time period for which survey studies were then available was still quite short, and therefore one-shot studies, usually of a cross-sectional and not a longitudinal nature, dominated the field. In addition, high-quality surveys were expensive, funding opportunities were limited, and therefore access of the academic community to such information was difficult.

The resulting scarcity of data plus the growing insight that primary investigators were not capable of

completely exhausting the analytical potential of a given dataset paved the road for considerations to make data which were often paid for by public money and therefore in principle a collective property available also to other than the primary investigators after those had had their first good shot at data analysis and publication.

A further consideration is also relevant. Often, data collected by primary researchers under a specific conceptual scheme can be reanalyzed under a vastly different scheme and may lead to new insights. Thus, the logic of secondary analysis (Hyman 1972) of existing data was an important addition to the arsenal of empirical political and social research and triggered an increasing demand for such data.

In sum, data archives were the answer to many concerns resulting from the data needs of empirical political science. Thus, it is little wonder that already in 1966 Ralph L. Bisco (1966, p. 93) could speak 'of a growing movement within the social sciences: the establishment of repositories of social science data.'

4. *The Historiographic Dimension of Longitudinal Data*

In one of his many insightful contributions to the social sciences, Paul F. Lazarsfeld [1972 (1964)] put himself into the role of a 1984 historian and asked about the obligations a 1950 pollster has vis-à-vis this historian. One can carry this torch much farther now in observing that especially data from survey research have opened up a new dimension in historic analysis along the lines of social history. With the advent of representative sample surveys of national population or special subgroups in society the potential for the *ex post* understanding of microphenomena of an attitudinal as well as of a behavioral nature has been vastly enhanced, and examples of excellent historical analysis based on findings from longitudinal sample surveys are increasing in number (see for instance the analysis by Schildt in 1995 on the modernization process in postwar West Germany).

But there is much more to be gained analytically from data which reach back into the past, and this is even more true when the researcher's interest is in more than one country. In the late 1980s, after heated controversies especially in Europe about the future of democratic government, many political scientists felt that the time had come to take stock of the sociopolitical changes in the 1960–1990 period in the nations of Western Europe which could be observed among mass publics. Such a stocktaking took place between 1989 and 1994 through the Beliefs in Government project funded by the European Science Foundation (ESF) and directed by Max Kaase and Kenneth Newton (Kaase and Newton 1995; Abramson and Inglehart 1998). Clearly, this project would not have been possible without systematic recourse to existing databases held in European data archives. Still, at the same time a number of shortcomings and weaknesses in the availability, equivalence (van Deth 1998), and quality of the data showed when becoming longitudinal and at the same time comparative in perspective.

Pitfalls in comparative research can be more easily avoided when the data are collected through primary research, like in the ongoing (1999) project on the Comparative Study of Electoral Systems (CSES) running in 40–50 countries of the world, and in the planned European Social Survey (ESS). This is a Europe-wide extension of the concept of General Social Surveys as developed first by James A. Davis of the National Opinion Research Center (NORC) at the University of Chicago in 1972 and since then conducted almost annually in the US and also, though less frequently, in some European countries. An additional advantage of both surveys will be that they will be amended by systematic macroinformation on political and social arrangements (like the respective electoral laws) in the participating nations, thereby permitting micro–macro (multilevel) analyses.

5. *Core Databases in Social Sciences*

By the turn of the twentieth century, the initial data scarcity in the social sciences has been replaced by a situation of data richness. There are many reasons for this. With economic growth especially in the OECD countries since the end of the Second World War, the number and/or size of educational institutions has risen dramatically with larger student populations (measured by shares in birth cohorts), and with this also the number of researchers both within and outside of institutions of higher learning. This development has, of course, embraced the social sciences too, and with more research money around the number of academically driven data gathering research projects has increased substantially. Also, a large number of empirical social science studies is regularly conducted on behalf of government agencies, the media, and for private business, often with advice from the academic community. Therefore, quite a few of these data also find their way into academia, and this means mostly through data archives. In addition, new survey research methods, and here especially the rise in telephone surveys, has made data production less expensive than in face-to-face interviewing studies. This data richness, however, does not automatically justify the conclusion that these data also satisfy most of the information needs in political science.

A specification of which core databases in political science presently exist has to be taken according to explicit criteria. For the purposes of this article, the main selected criteria are emphasis on microdata stemming from representative sample surveys, longitudinality and international scope; aggregate political

science databases like the one for the World Handbook of Political and Social Indicators (Taylor and Jodice 1983) and the Comparative Party Manifestoes Project (Budge et al. 1987, Klingemann et al. 1994) will be covered in a separate article in this Encyclopedia. In addition, one has to keep in mind that longitudinal national or comparative multipurpose surveys like the General Social Survey (GSS), the International Social Survey Programme (ISSP), and the biannual Eurobarometer surveys in the member states of the European Union (EU) as well as the Latinobarometro for South America regularly contain a variable mix of political and nonpolitical items, thereby preventing an unambiguous classification as political science surveys. They will, though, be included in the following listing, with the first two entries bearing more emphasis on a national than on an international focus. (See Databases Core: for several additional entries covering other disciplines.)

5.1 The National Election Studies

For political scientists the most interesting data come from the various national election studies, in particular from the American NES. They aim at the heart of the democratic political process—elections—and are longitudinal in nature. The opportunity to look at the electoral process comparatively is provided by these data to the extent that there are enough at least functionally equivalent questions asked across nations and different institutional and political settings to make an international comparison meaningful at all (for an inventory of election studies in nine European nations see Mochmann et al. 1998).

5.2 The General Social Surveys (GSS) and Other Surveys

As the analyses by Putnam (1998) on the decline of social capital in the United States show, the data from the GSSs can be and are put to an excellent longitudinal use by political scientists. In addition, there are many national datasets, cross-sectional and/or longitudinal, which are relevant for political scientists, but cannot be specified in detail here because of their number and their numerous national specialities.

Moving on from a more national to a distinctly comparative perspective, the following databases are of core interest to political scientists:

5.3 The Eurobarometers, the New Democracies Barometer, and the Latinobarometros

The Eurobarometers are representative sample surveys with usually about 1000 respondents conducted twice a year since 1974 in all countries of the European Union. As the number of EU member countries has grown over the years, so has the number of nations covered by the Eurobarometer. While many of the questions are concerned with special needs of the European Commission, the data contain enough continuous information on politics to make them very attractive for political scientists (for example, see Kaase and Newton 1995). Also, since the demise of communism, similar surveys have been regularly conducted in Central and Eastern Europe (Eurobarometer East). All Eurobarometer data are accessible through data archives like ICPSR and ZA.

For research on the transition from communist to democratic rule, it is important to know that in addition to the Eurobarometer East, comparable longitudinal representative sample surveys for 11 Central and East European countries have also been conducted in the context of the New Democracies Barometer organized on behalf of the Austrian Paul F. Lazarsfeld Society. Between 1991 and 1998 five such surveys were fielded (Rose et al. 1998). An extension to 20 countries with surveys in 2000 and 2002 is envisaged. Presently, the data for waves one, two, and three (1991, 1992, 1993–94) and wave five (spring 1998) are freely available through the Austrian data archive Wiener Institut für Sozialwissenschaftliche Dokumentation und Methodik (WISDOM). Furthermore, the Centre for the Study of Public Policy at the University of Strathclyde under the direction of Richard Rose has set up a New Russia Barometer since 1992 with seven surveys until the summer of 1998 and a New Baltic Barometer in the three Baltic states since 1993, with three surveys as of the summer of 1999. For both, a continuation is planned. The data are not yet freely accessible (for other pertinent data on the transition topic see also Sect. 5.5).

A Latinobarometer has been modeled after the Eurobarometer and has been conducted annually since 1995 in countries of Central and Latin America (1995: eight countries; since 1996: 17 countries). The Latinobarometro is anchored at the Centró de Estudios de la Realidad Contemporánea in Santiago, Chile, and directed by Marta Lagos (Lagos 1997). The data of this survey as of now (autumn 1999) are not freely available through an academic data archive.

5.4 The International Social Survey Programme (ISSP)

ISSP originated in 1982 from an American–German collaboration in devoting a small part of their GSSs to a common set of questions. It then developed to a four-nation participation in 1985–1986 and by 1998 had grown to 31 participating nations in all parts of the world. ISSP is often attached to a GSS, uses different data collection modes, and in each year

focuses on a particular topic which is replicated from time to time. For example, in the ISSP surveys in 1985–1986, 1990–1991, and 1996/1997/1998, the topic was role of government. These data are, like the Eurobarometers, accessible through data archives.

5.5 The European/World Values Surveys

In 1981, the European Values System Study Group started with surveys in 10 West European societies; this survey was replicated soon in 14 other countries outside of Europe to become the World Values Survey. Its second wave of surveys was coordinated by Ronald Inglehart between 1989 and 1993, the third wave ran between 1995 and 1998 in 53 countries, and the fourth wave of surveys has presently (1999) begun. Not all of the data at the time of writing is accessible; in the long run this study is, however, one of the most interesting for comparatively working political scientists because of its potential for macro–micro analyses (for a major comparative analysis of these data see Inglehart 1997).

Bibliography

Abramson P R, Inglehart R 1998 Comparing European Publics. *American Political Science Review* **92**: 185–90

Almond G A 1996 Political science: the history of the discipline. In: Goodin R E, Klingemann H-D (eds.) *A New Handbook of Political Science*. Oxford University Press, Oxford, UK

Bisco R L 1966 Social science data archives: a review of developments. *American Political Science Review* **60**: 93–109

Budge I, Robertson D, Hearl D (eds.) 1987 *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge University Press, Cambridge, UK

Campbell A, Kahn R L 1952 *The People Elect the President*. University of Michigan, Ann Arbor, MI

Campbell A, Gurin G, Miller W E 1954 *The Voter Decides*. Row Petersen, Evanston, IL

van Deth J W (ed.) 1998 *Comparative Politics. The Problem of Equivalence*. Routledge, London

Hyman H 1955 *Survey Design and Analysis. Principles, Cases and Procedures*. Free Press, Glencoe, IL

Hyman H H 1972 *Secondary Analysis of Sample Surveys. Principles, Procedures and Potentialities*. Wiley, New York

Inglehart R 1997 *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ

Kaase M, Newton K 1995 *Beliefs in Government. Beliefs in Government Series*. Oxford University Press, Oxford, UK, Vol. 5

Klingemann H-D, Hofferbert R I, Budge I 1994 *Parties, Policies and Democracy*. Westview Press, Boulder, CO

Lagos M 1997 Latin America's smiling mask. *Journal of Democracy* **8**: 125–38

Lazarsfeld P F 1972 The obligations of the 1950 pollster to the 1984 historian. In: Lazarsfeld P F (ed.) *Qualitative Analysis. Historical and Critical Essays*. Allyn and Bacon, Inc., Boston, MA

Lazarsfeld P F, Rosenberg M (eds.) 1955 *The Language of Social Research. A Reader in the Methodology of Social Research*. Free Press, Glencoe, IL

Lipset S M 1963 *Approaches Toward Reducing the Costs of Comparative Survey Research*. Social Science Information 2: 33ff

Miller W E 1989 Research life as a collection of intersecting probability distributions. In: Eulau H (ed.) *Crossroads of Social Science: The ICPSR 25th Anniversary Volume*. Agathon Press, New York

Miller W E 1994 An organizational history of the intellectual origins of the American National Election Studies. *European Journal of Political Research* **25**: 247–65

Mochmann E, Oedegaard I C, Mauer R 1998 *Inventory of National Election Studies in Europe 1945–1995: Belgium, Denmark, France, Germany, Great Britain, Hungary, The Netherlands, Norway and Sweden*. Edwin Ferger Verlag, Bergisch Gladbach, Germany

Putnam R 1998 Democracy in America at the end of the twentieth century. In: Rueschemeyer D, Rueschemeyer K, Wittrock B (eds.) *Participation and Democracy. East and West. Comparisons and Interpretations*. M E Sharpe, Armonk, NY

Rokkan S 1964 Archives for Secondary Analysis of Sample Survey Data. An Early Inquiry into the Prospects for Western Europe. *International Social Science Journal* **16**: 49–62

Rokkan S (ed.) 1966 *Data Archives for the Social Sciences*. Monton, Paris and The Hague

Rose R, Mishler W, Haerpfer C 1998 *Democracy and its Alternatives. Understanding Post-Communist Societies*. Johns Hopkins University Press, Baltimore, MD

Schildt, A 1995 *Moderne Zeiten: Freizeit, Massenmedien und "Zeitgeist" in der Bundesrepublik der 50er Jahre*. Christians, Hamburg, Germany

Taylor Ch L, Jodice D A 1983 *World Handbook of Political and Social Indicators. Vol. II: Political Protest and Government Change*, 3rd edn. Yale University Press, New Haven, CT

Thomassen J 1994 The intellectual history of election studies: Introduction. *European Journal of Political Research* **25**: 239–46

Weisberg H 1987 The Silver Summer: The ICPSR Summer Training Program's 25th Anniversary. *PS* **20**: 281–7

M. Kaase

Databases, Core: Sociology

A database consists of one or more data sets, such as a sequence of decennial censuses. Databases reach core status for several reasons. One is sheer frequency of use. In the USA, the General Social Survey (GSS), a series of cross-sectional surveys beginning in 1972, is used by more than 200,000 students a year and by hundreds of researchers. The data are freely available to anyone via the Internet (http://www.icpsr.umich.edu/GSS99/, (which gets more than two million hits a year) and are well documented and easily accessible. Other databases

become influential because they yield fundamental results and serve as a research model. The two 'Occupational Change in a Generation' surveys in the US produced benchmark estimates of social mobility and spawned similar mobility surveys around the world (Blau and Duncan 1967, Featherman and Hauser 1978). Still other databases are influential because they provide basic data on social structure (e.g., income shares) which then play into other research. For example, many countries have an annual sociodemographic survey that yields income and employment statistics, which in turn drive research on poverty. To an increasing degree, the most influential databases are those which are based on systematic data collection over both space and time. Repeated data collection provides a fundamental tool for attacking some of sociology's defining problems such as macro–micro linkages and the evolution of social structure. They also provide a framework in which one-time data collections can be understood. This article will argue that over time the core databases in sociology will be those which permit historical and cross-cultural comparisons.

1. *The Impact of Data on the Discipline*

Sociologists have argued for many years about the extent to which the physical sciences, particularly physics and chemistry, provide a model for doing sociology (Lieberson 1985). An argument can be made that geology is perhaps a better model. In that field, the scientific revolution known as the theory of plate tectonics resulted in large part from the study of the sedimentary record. Using deep cores and other tools, geologists have been able to reconstruct the physical history of the earth, along with its climate and biology, going back many millions of years. At the beginning, the data were collected in a somewhat a-theoretical way but as the picture became inductively clear, more focused data collection resulted (Van Andel 1985). This kind of interplay between data collection and theory development characterizes other fields as well. In biology, Darwin's theoretical achievements were based on years of data collection and the theory of natural selection drew heavily on the geological record.

In sociology, the data accumulated over many centuries of human activity are somewhat analogous. While the time scale differs by several orders of magnitude, one can think of accumulated data from human activity in the form of government, religious and commercial records, and more recently from surveys, as the sediments that form the basis for social science research. As these records have become computerized the volume of data, and the ease with which it can be accessed, have increased dramatically. Of course, researchers have used surviving hard-copy data for a long time. Historical demographers have used parish registries to study family formation and economic historians have used commercial and government records to study price levels and other economic phenomena. But until very recently the availability of data was, to a large extent, accidental, based on nature's whim and historical chance. Coverage of topic, region, and time period was erratic and inconsistent. But in the last 50 years of the twentieth century, the information revolution has been laying the basis for a databased social science that is far broader and systematic than has been the case in the past.

There are several aspects of the revolution. First, much primary data are now being created directly in electronic form, thus increasing accessibility. Second, vast amounts of data can be stored at very low cost. Third, with the emergence of the Internet, data can be moved about the world almost at will. Finally, software for accessing the data and analyzing it is becoming easier to use, although much remains to be done. The revolution dates back to the early 1960s. Although there are certainly examples of machine-readable data collected before then, the flood began with the availability of second-generation computers to process data and machine readable tapes on which to store and ship it. Since then, a torrent of data has become available and the breadth and depth of the river is increasing daily. Two processes are going simultaneously. First, hard-copy data from old records and texts are being retrieved and converted to electronic form. For example, in the USA very large samples of decennial census records from 1850 to 1950 have been drawn from the original manuscript data and are available to be used in conjunction with contemporaneously created sample files from 1960 onward. More focused data conversion is also going on. To cite just one example from dozens available, the study of slavery in North America is being revolutionized by the conversion of old records such as courthouse registers and slave ship manifests (Eltis 1999, Hall 1999). Second, various data resources—not just surveys but electronic records of all kinds—are slowly accruing over time. For example, tables of live births by age of mother and race are available for the USA from 1933 to the present.

To see the impact of what is happening, imagine that the information revolution had begun a 100 years earlier, i.e., about 1860 rather than 1960. Suppose that the electronic information that social scientists now work with routinely—surveys, macroeconomic data, administrative records of various kinds, detailed geographic information—had begun to emerge at the beginning of the American Civil War and in the midst of the Industrial Revolution. Events such as the Great Depression would be seen in the light of a number of other economic events for which similar data would exist. In the USA, the 70-year period beginning with the 'Depression' and followed by more than half a century of almost continuous

economic growth would be seen in the light of a far more variable economy in the period from 1860 until 1930. In Europe, the effects of the breakup of the Soviet Union could be studied in light of similar political and social transformations that occurred after World War I. Of course, it cannot be assumed that the data would be perfect or complete, any more than it can be assumed that current systematic data collection will proceed unperturbed in the face of economic upheaval, civil strife, or environmental catastrophe. But even with gaps in coverage and quality, with 140 years of data it would be possible to do a kind of social science which is now only beginning to be visible. Gradually, as the database for social science research grows more sophisticated, sociology and related fields will undergo a databased revolution that will redefine both substance and method.

2. *An Overview of the Emerging Database*

Although any classification is somewhat arbitrary, it is helpful to think in terms of five main categories of databases.

2.1 Cross-sectional Micro Level Social, Economic, and Demographic Data

Many countries conduct national demographic surveys, either as censuses or based on samples, and in some cases both. These surveys tend to focus on core demographic data: family formation, fertility, mortality, and related matters. Increasingly, however, they deal with education, income, health, and other basic social variables. If not full censuses, most of the surveys have relatively large samples sizes. It can be expected that more and more countries will conduct such surveys as the benefits of the data become obvious, and as the technology for doing the survey work becomes more widespread. The 'Integrated Public Use Micro Sample' (IPUMS) project provides access to US Census from 1850 to the present at www.ipums.umn.edu. IPUMS will soon be offering access to other national censuses, with plans to include multiple time points from 21 different countries. The US Current Population Survey is available back to the mid-1960s from www.unicon.com. Many other national statistical agencies carry out similar surveys. The Luxembourg Income Study, begun in 1983, provides somewhat similar data for more than 25 countries (www.lis.ceps.lu).

2.2 Cross-sectional Omnibus Social Surveys Focusing on Social Behavior, Attitudes, and Values

The GSS is perhaps the premier example of an omnibus survey. The International Social Survey Program (ISSP), which currently contains data from 31 countries, is linked closely to the GSS. Data and documentation can be obtained at www.icpsr.umich.edu/GSS99/. An important feature of the GSS/ISSP is its careful replication of measurement and methodology both over time and across countries. A somewhat parallel survey series, known as 'Eurobarometer' has been carried out between two and four times a year since 1974 (see http://europa.eu.int/comm/dg10/epo/org.html). Eurobarometer studies tend to be more topical and less focused on replication than the GSS.

2.3 Macro Level Economic and Social Data

Macro data consists of two kinds; data aggregated up to the micro level such as time and country-specific counts of crimes, and nondisaggregatable data such as balance of payments records. Among other organizations, the World Bank, the United Nations, the International Monetary Fund, and the Organization for Economic Cooperation and Development all provide data files and tables, varying in focus and coverage as do various academic institutions such as The University of Pennsylvania via its Penn World Tables (http://pwt.econ.upenn.edu/).

2.4 Micro Level Longitudinal Surveys of the Life Course

Since the mid-1960s, a number of studies in various countries have followed individuals longitudinally over long periods of time. The article on Economic Micro Databases (*Microdatabases: Economic*) discusses a number of these studies. They are important because they provide data on long periods of the life course within a particular historical context. Increasingly, the studies are being replicated, and one can study the flow of lives under differing social and economic conditions. For instance the US Health and Retirement Survey (www.umich.edu/~hrswww/) provides data on the later life course for cohorts born after about 1920, with new cohorts entering the measurement sequence every five years.

2.5 Administrative Databases

Administrative records maintained by government and other organizations such as insurance companies, are potentially incredibly valuable resources. For example, a number of researchers in the US have been able to analyze survey data from studies of aging in conjunction with information from Health Care Financing Administration (Medicare) records (e.g., Wolinsky et al. 1995). These databases are usually not constructed with research in mind, and thus it is often

time consuming to make effective use of them, but the accumulated data in various administrative databases can offer research opportunities available in no other way.

3. Issues to be Resolved

Thus far, this article has been written as though there are no impediments to a new world of data. However, this is not the case.

3.1 Gaining Access to Data

Access to micro level surveys, particularly those conducted by national statistical bureaus, is often difficult. Some countries refuse access entirely and others sharply restrict it. When data files are released, confidentiality concerns often lead to sharp limitations on the availability of potentially identifying information, particularly geographic data. Researchers can often negotiate access to specific data files, but gaining simultaneous access to multiple data sets in multiple countries is often impossible. There are several approaches to the problem. The US Bureau of the Census has created several Research Data Centers, where closely controlled on-site access to restricted data is available to qualified researchers. Another approach is to devise disclosure limitation methodologies to protect individual identities in publicly available files. A third possibility is to make documentation public but require analyses to be run remotely on the agency's computer with agency control over the analyses. Despite these initiatives, public concerns about confidentiality and privacy continue to impact legislation on data access in many countries. If the promise of the information revolution is to be fully realized, efficient ways of giving researchers access to data while satisfying privacy concerns will have to be found.

3.2 Ease of Use

Anyone who has used a complex data set knows that the cost of getting into the data is high. Even the best data sets tend not be documented in such a way that outside researchers find it easy to use them. The problem has been exacerbated by the use of computer-assisted interviewing. Although computer-based interviews have major advantages in data collection, one of which is to allow very complex survey instruments, they tend not to produce easily used documentation and the very complexity of the questionnaires makes the data difficult to use. Because the 'questionnaire' often consists of a series of computer screens, there is often no traditional hard-copy questionnaire, and producing one after the fact is difficult and time consuming. The research community has recognized the problem and there are various meta data projects under way. One example is the 'Data Documentation Initiative,' an international project involving many social science organizations, which has developed mark up standards for machine-readable documentation (http://www.icpsr.umich.edu/DDI/codebook.html). An application of DDI can be found at FASTER (http://www.nesstar.org/), a joint project of a number of European data and statistical organizations. There is also a good deal of work in progress to produce software that combines data collection, documentation, and analysis tasks. One example, which is freely available, is EpiInfo a software package for public health researchers developed at the US Centers for Disease Control (http://www.cdc.gov/epiinfo/). There are also several competing commercial products under development at this point.

3.3 Design

As noted, with important exceptions, the accumulation of data is largely accidental. Within sociology, there is no grand plan to dictate what kind of data gets collected at what intervals. It is quite tempting to design studies to react to major political and social events, such as the breakup of the Soviet Union. Sometimes there is even ample time to anticipate an event, as was the case with welfare reform in the USA. However, it is becoming increasingly clear that slow, steady, systematic data collection offers the best chance of capturing the effects of unanticipated events and in any case provides an enormously powerful tool. The National Election Studies in the USA (www.umich.edu/nes/) provide a sequence of surveys going back to 1948. As the name implies, the principal use of the data has been to study national elections, but as a 50-year sequence of replicated surveys, the data also offer an almost unparalleled chance to study change in American society. With the exception of census data, it is very hard to find examples of systematic micro data series that go back further. It is very difficult for researchers anywhere to find the resources to maintain these kinds of studies over an extended period of time. But it seems clear that the steady accrual of systematic data while costly and by definition slow, is the only way to be sure that one has the right kind of data at the right time. Trying to spring into the field on short notice with the ideal design to study a particular topic or event is risky. This is not to say that short-term studies are valueless but rather to argue that short-term studies are most effective when a long-term framework of data using standard protocols and methods can support them.

3.4 Replication and Comparability

Perhaps the aspect of design that requires the most attention is measurement. Studying a given topic over the long term using multiple databases requires a certain amount of consistency of measurement. The

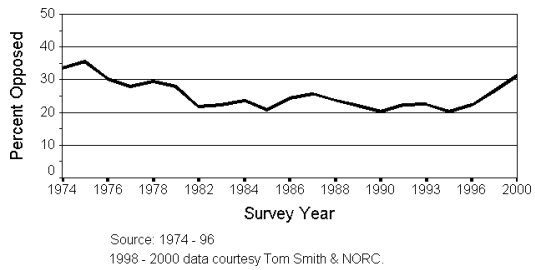

Figure 1
Percent opposed to capital punishment, US population age 18 and above. Source: General Social Survey

GSS, for example, has maintained consistency of its basic items since its inception, and some of those items were taken from previous surveys that go back 20 years prior to the start of the GSS. Figure 1 shows GSS results from the US between 1974 and 2000 for the item 'Do you favor or oppose capital punishment for murder.' As the graph shows, opposition to capital punishment has increased since 1995, but is only now reaching levels that were common in the mid-1970s. The trend is too systematic to be accidental, and there are many possible explanations. It is tempting to speculate on the link between this trend and the fact that crime rates, particularly homicide, have fallen in the US for the past several years. The only point to make is that this comparison would have been difficult to impossible, had the wording of the item changed. Unfortunately, some items literally wear out as changing tastes and social conditions make them archaic or irrelevant, if not offensive, and some change is inevitable.

Unfortunately, this kind of measurement and design consistency is rare, particularly for cross-cultural research. Researchers who wish to pool data from multiple surveys are usually faced with enormous inconsistency both within surveys over time and across surveys. Even within a given field, one finds differences in standard items across surveys. Often, there is no particular reason for the inconsistency; researchers just don't think as much as they might about replicability. However, important progress is occurring. As noted, the ISSP has made great strides in obtaining comparable data across countries. There are several other examples such as the SF-36; a scale designed to measure health that is now in use in more than 45 countries (www.sf-36.com). A few fields of sociology, particularly social mobility, have made great progress in dealing with these issues. Ganzeboom et al. (1991) review mobility studies from more than 30 countries showing that it is possible to develop regression models that maintain at least rough comparability.

Despite the obvious advantages of consistency, there is a certain paradox here. Suppose that we find that when one operationalizes a set of constructs in exactly the same way, using exactly the same measurement protocols and survey design, one gets closely comparable results across multiple surveys. Is the similarity due to the 'methods effect' or does it represent true equality of results? In other words, one might like to show that the basic results replicate despite differences in measurement, sampling, etc. The answer to this question depends in large part on how closely one wants to compare results. To answer the question 'have the economic returns to schooling changed over time?' requires a great deal of consistency because answering the question requires exact comparison of regression coefficients. On the other hand, answering the question 'are the relative effects of achievement and ascription the same in most industrial societies?' might be answered, at least approximately, with something less than exact comparability.

3.5 Data Archiving and Preservation

An important aspect of the data revolution is the existence of data archives in many countries of the world. The idea of a data archive developed in the early 1960s when most data sets were in the form of decks of electronic data processing cards, a technology now virtually unknown. The card files were expensive to maintain and quite fragile, so data archives maintained master copies and provided duplicates to researchers. As the computing world moved to storing data on tapes, archives were sill needed as central repositories of data because the task of maintaining the data and making it available to users in a form suitable for their local computing facility, required technical skill and resources. A web-based world map of most of the major academic archives can be found at http://www.nsd.uib.no/cessda/other.html. The InterUniversity Consortium for Political and Social Research (ICPSR) (www.icpsr.umich.edu) is the world's largest academic archive with literally thousands of data sets, but there are a great many specialized archives around the world. A very thorough guide to archives and related facilities is maintained by the Social Sciences Data Collection at the University of California San Diego (http://odwin.ucsd.edu/idata/) where one can find links to hundreds of electronic resources.

The new world of the Internet is leading to radical changes in the nature of data archives. Currently, many data sets are made available directly on the web by the originating investigator or agency. This has obvious advantages in terms of accessibility, but perhaps less obvious disadvantages. Although Internet search tools are powerful, it is far easier to search the holdings of a few data archives than it is to search the whole Internet for data on a given topic. A more important issue is permanency. Archives are committed to maintaining data and related documentation in perpetuity. Data maintained on local web sites is

subject to all sorts of vicissitudes—funding disappears, investigators move or lose interest, servers crash, and technology changes sometimes rendering data unusable. Indeed, archives have to worry about the fact that data on any storage medium deteriorates over time, as many users of compact disks know (see *Data Archives: International*).

4. The Future

Although there is no question that social science will be affected profoundly by the availability of masses of data, a great deal must be done if the data are to be used to maximum advantage. Three issues seem to be the most important. First, the research community needs to pay closer attention to comparability and replication. The Internet has made that task vastly easier because data documentation can be moved about so easily. But the quality of the documentation varies a great deal and finding the information that one needs to know in order to either analyze the data or replicate the study, can be difficult to impossible. The development of efficient and effective electronic documentation and meta data is in the early stages and much remains to be done. Second, means of giving qualified researchers access to data must be worked out. In the current climate of serious concern about privacy, this will be a difficult task. On the other hand, the power of computing, when used properly, makes it easier to guarantee privacy rather than harder. Data can be encrypted, modified for public consumption to prevent disclosure, maintained at a single site to restrict access to qualified people, etc. There are slowly emerging methods and standards for how to do these things, but many issues remain to be resolved. Finally, and most importantly, there is the intellectual task of coming to terms with the stream of data that will be available. Until very recently, most researchers were content to focus their activities on a single data set. But as data accrue, for almost any question one is interested in, there will be multiple data sets available. Not only will there be multiple data sets across time and place, but there will be corresponding macro level data which might be relevant, as in the case of the GSS data on capital punishment, which can only be understood in light of societal level social and political factors.

See also: Censuses: Comparative International Aspects; Censuses: Demographic Issues; Censuses: History and Methods; Data Archives: International; Databases, Core: Demography and Registers; Databases, Core: Political Science and Political Behavior; Microdatabases: Economic; Survey Research: National Centers

Bibliography

Blau P M, Duncan O D 1967 *The American Occupational Structure*. Wiley, New York

Eltis 1999 *The Trans-Atlantic Slave Trade: A Database on CD-ROM*. Cambridge University Press, Cambridge, UK

Featherman D L, Hauser R M 1978 *Opportunity and Change*. Academic Press, New York

Hall G M 1999 *Databases for the Study of Afro-Louisiana History and Genealogy, 1699–1860: Computerized Information from Original Manuscript Sources*. Louisiana State University Press, Baton Rouge, LA

Lieberson S 1985 *Making It Count: The Improvement of Social Research and Theory*. University of California Press, Berkeley, CA

Ganzeboom H B G, Treiman D, Ultee W C 1991 Comparative intergenerational stratification research: Three generations and beyond. *Annual Review of Sociology* **17**: 277–302

Van Andel T H 1985 *New Views on an Old Planet: Continental Drift and the History of Earth*. Cambridge University Press, New York

Wolinsky F D, Stump T E, Johnson R J 1995 Hospital utilization profiles among older adults over time: Consistency and volume among survivors and decedents. *Journal of Gerontological B Psychological Sciences and Social Sciences* **50**(2): S88–100

R. T. Campbell

De Finetti, Bruno (1906–85)

Bruno de Finetti was born in Innsbruck, Austria of Italian parents on June 13, 1906, and died on July 20, 1985 in Rome. He enrolled in the Milan Polytechnic in 1923, and then after discovering that his interests lay more in science, switched to the Faculty of Mathematics of the University of Milan.

He first became interested in probability as a result of reading an article in a popular magazine concerning Mendel's laws of genetics, which stimulated him to write an article of his own on the subject. His article attracted the attention of Corrado Gini and eventually led to a position as head of the Mathematical Office of the Instituto Centrale di Statistica. Later he accepted a position as actuary in the Assicurazioni Generali insurance company in Trieste, where he was concerned with automation of all actuarial, statistical, and bookkeeping operations and calculations. He became a lecturer in mathematics, actuarial mathematics, and probability at the Universities of Trieste and Padova, and later full professor in Trieste and finally at the University of Rome, from which he retired in 1976.

In 1957 he spent a quarter at the University of Chicago at the invitation of Professor Leonard Jimmie Savage (1917–71). They became friends and each stimulated the other in the development of his ideas.

Although there are differences in the presentations of subjective probability by de Finetti and Savage, their basic outlook was the same, and they stood in sharp contrast to other figures in the field, with the exception of Frank Ramsey (1903–30), whose important article *Truth and Probability* (1926) was very much in the same spirit. De Finetti and Savage were the major figures of the twentieth century in the area of subjective probability, and legitimate heirs of the two giants, Émile Borel (1871–1956) and Henri Poincaré (1854–1912), who also viewed probability theory as a serious subject with numerous important real-world applications, rather than as a mathematical exercise. Both de Finetti and Savage were infuriated by the trivializations of the subject as such an exercise, which dominated the field during most of their lifetimes.

Early in his career de Finetti realized that there was something seriously lacking in the way probability and statistics were being developed and taught. The famous remark 'Everybody speaks about Probability, but no one is able to explain clearly to others what meaning probability has according to his own conception,' by the mathematician Garrett Birkhoff (1940), was indicative of serious problems even with the definition of probability. De Finetti found many of the standard assertions confused and meaningless. In his University of Rome farewell address (1977), concerning the contrast between himself and Savage (a mathematician who had been trained in the standard non-Bayesian theory) he said: 'I, of course, was in quite the opposite situation: a layman or barbarian who had the impression that the others were talking nonsense but who virtually does not know where to begin as he cannot imagine the meaning that others are trying to give to the absurdities that it is imposing as dogma.'

De Finetti then attempted to formulate an appropriate meaning for a probability statement. His knowledge of philosophy and logic as well as mathematics, both pure and applied, enabled him to do so. The areas of his work that will be discussed here are his theory of exchangeability, theory of coherency, and theory of finitely additive probability. For him theory and practice were inseparable, and he emphasized the operational meaning of abstract theory, and particularly its role in making probabilistic predictions.

De Finetti rejected the concept of an objective meaning to probability and was convinced that probability evaluations are purely subjective. In his article *Probability and my Life* published in Gani (1982) he wrote: 'The only approach which, as I have always maintained and have confirmed by experience and comparison, leads to the removal of such ambiguities, is that of probability considered—always and exclusively—in its natural meaning of "degree of belief." This is precisely the notion held by every uncontaminated "man in the street." De Finetti then developed a theory of subjective probability, corresponding to its use as an expression of belief or opinion, and for which the primary feature is coherency.

The starting point for his theory is the collection of gambles that one takes on, whether knowingly or unknowingly. These can be either ordinary gambles or conditional gambles, which are called-off if the contingency being conditioned upon does not occur. The simplest type of gamble is one on an event E, in which one receives a monetary stake S if E occurs and nothing otherwise. If a person values such a gamble at the sure amount $p \times S$, then p is defined to be the probability of E for that person. In this evaluation one uses small stakes S so as to avoid utility complications, which distort probabilities evaluated in this way. One can then take linear combinations of such simple gambles to generate gambles in which there are a finite number of possible outcomes. Similarly, a called-off gamble on the event E given F, is such that one receives a monetary stake S if both events occur, nothing if F occurs but not E, and the gamble is called off if F does not occur. The value of such a conditional gamble for a person is called his or her conditional probability for E given F and is denoted by $P(E|F)$.

In the work of de Finetti (1970, 1972, 1974–5) a primary aim is to avoid the possibility of a Dutch book, that is, to avoid sure loss on a finite collection of gambles. Thus suppose one has taken on a finite collection of gambles of any type whatsoever, including such things as investments, insurance, and so on. De Finetti proved that a necessary and sufficient condition for the avoidance of sure loss of a positive amount greater than some $\varepsilon > 0$ on a finite number of unconditional gambles, consists in consistency with the axioms of finitely additive probability theory. These axioms are the usual ones, apart from the fact that the condition of finite additivity replaces that of countable additivity in the third axiom. Let $P(E)$ denote the numerical probability for the event E. Then the axioms are:

1. $P(W) = 1$, where W is the sure event.
2. For any event E, $0 \leqslant P(E) \leqslant 1$.
3. The probability of a finite union of disjoint events is the sum of the probabilities.

Further, if conditional gambles are also included, then to avoid a Dutch book, conditional probabilities must be evaluated in accord with the usual definition:

$$P\{A|B\} = P\{A \text{ and } B\}/P\{B\},$$

provided that $P\{B\} > 0$. See Hill (1998) for further discussion.

In the subjectivistic conception a probability is essentially a price that one will pay for a certain gamble, with the understanding that this price can and will differ from one person to another, even apart from utility considerations. This conception is very close to that of the common person who speaks of probabilities for various real-world events based upon his or her

own experience and judgment. In this case there is no need to confine attention to trials such as arise in games of chance. On the other hand, it is extremely interesting to see what the subjectivist interpretation of such trials might be. By considering this question, de Finetti developed his celebrated exchangeability theorem.

The problem that de Finetti posed and largely solved was that of justifying from a subjective point of view the importance of statistical frequencies in the evaluation of probabilities. Suppose that we have a sequence of events, E_i, for $i = 1,\ldots, N$, each of which may or may not occur, for example where E_i is the result of the ith flip of a coin. For convenience, we can assign the value 1 to heads and 0 to tails. It is natural sometimes to view the order of the flips as irrelevant, so that any permutation of the same sequence of outcomes is given the same probability. This leads to the notion of exchangeability, which is a symmetry condition, specifying that joint distributions of the observable random quantities are invariant under permutations. De Finetti proved a representation theorem which implies that if one regards a sufficiently long sequence of successes and failures as exchangeable, then it necessarily follows that one's opinions can be represented, approximately, by a mixture of Bernoulli sequences, that is, of sequences for which there is a 'true' (but typically unknown) probability P of success, and such that conditional upon $P = p$, the observations are independent with probability p of success on each toss. The proportion of successes, as the number of trials goes to ∞, converges almost surely to P, and the initial or a *priori* distribution for P is also called the de Finetti measure for the process.

This representation in terms of Bernoulli sequences is exact for infinite sequences, and a slight modification holds also for finite sequences, the difference amounting to that between sampling with and without replacement from an urn. It also follows mathematically that provided a sufficiently long sequence has been observed, to a good approximation one will be led to base opinions about a future trial on the past empirical frequency with which the event has occurred, and this constitutes de Finetti's justification for induction. In his work de Finetti emphasized prediction of observable quantities, rather than what is usually called inference about parameters. In other words, he was primarily interested in the posterior distribution of future observable random variables, rather than in the posterior distribution for the parameter P. This point of view has become known as predictivism in modern Bayesian statistics.

The exchangeability theorem of de Finetti allows one to analyze the relationship between subjective probabilities and the frequency theory of probability that developed from the work of Jakob Bernoulli. The fundamental stochastic process is known as a Bernoulli sequence. Such a sequence is by assumption such that on each trial there is a fixed objective probability p for success (1) and $1-p$ for failure (0), whose value may or may not be known, with the trials independent in the probabilistic sense. It is then a mathematical fact, known as the law of large numbers, that as the number of trials grows large the proportion of successes will stabilize at the 'true' probability p in the sense that the proportion of successes will converge almost surely to p; that is, the sequence converges except for a collection of outcomes that has probability 0.

Critics of this theory, such as de Finetti, argue that there are no real-world applications in which either p or the limiting proportion of successes is known to exist. The empirical support for the objectivist theory is merely that there are certain processes, such as coin flipping, dice throwing, card playing, draws from an urn with replacement, and so on, for which in finite sequences of trials the empirical proportions of successes appear to be in conformity with what the objectivist theory suggests. For example, if there are two similar balls in an urn, one red and one white, and if draws are made with replacement after substantial shaking of the urn, then experience suggests that the proportion of red balls in n draws, when n is sufficiently large, is around .5. Of course it is impossible to perform an infinite number of trials, so the limiting proportion, if it exists at all, can never be known. But in finitely many trials not only does the proportion of red balls drawn often seem to settle down to something like .5, but in addition, if say the n trials are repeated I times, then the proportion of these I trials in which there are exactly j red balls in n trials appears also to settle down to $\binom{n}{j}(1/2^n)$, more or less as the theory suggests. One must, however, exclude situations where such settling down does not appear to happen, and these are usually interpreted as meaning that the shaking was not adequate, or the cards not sufficiently well shuffled, and so on. Nonetheless, it appears that when sufficient care is taken, then the proportions appear to be roughly in accord with the objectivistic theory.

Does this imply that there exists a true $p = \frac{1}{2}$ which is the objective probability, for example, that the ball drawn is red, and hence also that the proportion of red balls drawn must converge almost surely to .5? No one knows to this day. However, de Finetti's exchangeability theorem provides an alternative way to describe our experiences with such random events. A probability as before is merely a price that one would pay for a certain gamble. Under the assumption of an equal number of red and white balls, and that the urn is well shaken, symmetry suggests a price of .5 for the gamble in which you receive a dollar if a red ball is drawn, nothing otherwise. Similarly, if you were to be offered a gamble in which you would receive the dollar if the first two draws were red, then again symmetry together with the assumption that the urn is well shaken, suggest that a fair price would be .25. Continuing in this way we see that your prices are in

accord with the usual probability evaluations for such an urn.

It is then a logical consequence of the exchangeability theorem of de Finetti, that you must evaluate your probability that the sample proportion of red balls converges to .5 as unity. Note that this requires no circular definitions or mysterious conventions. There are, however, two important caveats. First, it is necessary that utility considerations be ignored, which necessitates that the prizes are not too large, and for this reason the stake S was taken to be one dollar. Alternatively, utility can be dealt with after the fashion of Ramsey or Savage. Second, the cases where people take this evaluation most seriously are those in which there have already been many draws in the past, indicating no cheating or violation of the apparent symmetry, so that in effect your probabilities are posterior probabilities based upon Bayes' theorem. In other words: just as dice may be loaded so can other games be fixed, and a prudent bettor would not make the above evaluations merely from apparent physical and time symmetry alone, but rather in addition from experience that the game is not rigged and that the perception of symmetry is not mistaken. At any rate, one sees that it is not necessary to introduce the concept of an objective probability, but merely the experience that suggests an evaluation of .5.

The case of a Bernoulli sequence, whether interpreted according to the objectivist or subjectivist theory, is of course a very special example of an exchangeable sequence of random variables. For a Bernoulli sequence with $p = .5$, that is, a degenerate *a priori* distribution for P concentrated at the value .5, the primary difference between the two theories is that the objectivist theory objectifies the .5 which in the subjectivist theory is merely the opinion of the bettor, and similarly the .25 for drawing two red balls, and so on. In circumstances where prudent bettors have learned from experience to bet in accord with symmetry, the objectivist school states that the .5 and the .25 and so on, are the objective or true probabilities, not merely the values that the bettor chooses. Thus the objectivist theory implies that the proportion of successes necessarily converges to .5, while the subjective theory implies that the bettor must believe, with probability 1, that such convergence takes place.

It should be noted that countable additivity, that is, the modification of the third axiom to include also denumerable unions of disjoint events (which corresponds to a denumerable number of gambles) has never been demonstrated within the rigorous approaches to probability theory and, as argued by de Finetti, is merely a particular continuity condition that one may or may not wish to adopt in a particular context, just as with the use of continuous functions in physics. The relaxation of the condition of countable additivity to finite additivity allows one greater freedom in the representation of degrees of belief. For example, the use of adherent mass distributions (which were introduced by de Finetti and are closely related to Dirac functions) allows one to obtain solutions for Bayesian non-parametric inference and prediction. See Hill (1993).

The work of de Finetti played a major role in the resurrection of the classical Bayesian approach to probability, statistics, and decision theory, as founded by Bayes and Laplace and developed by Gauss, Poincaré, and Borel. His work continued the theory of induction of the British empiricists such as David Hume, and provided mathematically sound foundations for the modern-day subjective Bayesian approach.

See also: Probability and Chance: Philosophical Aspects; Probability: Interpretations; Statistics, History of

Bibliography

Birkhoff C 1940 *Lattice Theory*. American Mathematical Society Publication no. 25

De Finetti B 1970 *Teoria delle probabilitá*. Einaudi, Turin, Italy

De Finetti B 1972 *Probability, Induction and Statistics*. Wiley, London

De Finetti B 1974–5 *Theory of Probability*, Vols. 1 and 2. Wiley, London

De Finetti B 1977 Probability: beware of falsifications. In: Aykac A, Brumat C (eds.) *New Developments in the Applications of Bayesian Methods*. North-Holland, Amsterdam, pp. 347–79

Gani J (ed.) 1982 *The Making of Statisticians*. Springer-Verlag, New York

Hill B M 1993 Parametric models for A_n: Splitting processes and mixtures. *Journal of the Royal Statistical Society, Series B* **55**: 423–33

Hill B M 1998 Conditional Probability. *Encyclopedia of Statistics*. Wiley, New York

Kolmogorov A 1950 *Foundations of the Theory of Probability*. English translation. Chelsea Press, New York

B. M. Hill

Death and Dying, Psychology of

1. Attitudes Towards Death and Dying

In the literature, coping with one's own finitude is often understood as a sequence of stages that people pass through in part or completely. Kubler-Ross (1969), for example, distinguishes between an initial phase of denial followed by phases of anger, depression, bargaining, and—finally—acceptance. Though the generalizability of the stages model is debated, an important conclusion that can be drawn from the work of Kubler-Ross and others is that

attitudes toward death and dying can be understood as an outcome variable, reflecting the effectiveness of individual attempts to cope with the last borderline situation of human existence.

Previous research on factors that might influence attitudes toward death and dying focused extensively on fear of death as the variable to be predicted. Other outcome measures in recent studies are the desire to prolong life (e.g., Mutran et al. 1997), frequency of thoughts about death (e.g., Rao et al. 1998), and valuation of life (Lawton et al. 1999). The literature suggests that age, gender, and ethnicity may be variables associated with fear of death. The likelihood of fearing death appears to be slightly higher for younger persons, for women, and for African-Americans compared with Whites (Mutran et al. 1997).

1.1 The Role of Religious Activity

Religious activity has been a prominent topic in previous research. Studies have shown that people who engage into devotional activities experience less fear of death (Musick et al. 1998). Trying to explain this relationship, some authors focus on processes of primary appraisal, emphasizing that looking forward to an afterlife that is better than the current situation makes religious people less likely to see their disease as harmful. Focusing on processes of secondary appraisal, others have suggested that feeling close to God is seen as a source of power and strength that can be used in times of trouble. According to Filipp (1999), when processing bad news, religious beliefs may enable people to construct an interpretative reality, i.e., to ascribe subjective meaning and to re-establish a better world to live in. From the perspective of the social cohesiveness hypothesis (Idler 1987) religion can benefit individuals by providing support from clergy or fellow church members. However, studies show that the effects of religious activity are not the same for African-Americans as for Whites. To account for this difference it has been argued that, because the former face greater discrimination and barriers to institutions and status, they must find alternative avenues to success and well-being, and that one such alternative avenue is the church (Hummer 1996). Proceeding from the assumption that church and religious activity serve a special place in the lives of African-Americans, religious activity should have a greater effect for them than for Whites.

1.2 The Role of Social Support

Given that Kubler-Ross emphasized that processes of coping with death and dying can be assisted by social interaction and that the relationship between social support and health in general is the subject of numerous studies, it is noteworthy that until recently only a few studies investigated the relationship between supportive networks and attitudes toward death and dying.

Mutran et al. (1997) investigated the potential mediating and moderating effects of social support on the desire to prolong life, an aspect of attitudes toward death and dying that is related to, but not identical with, fear of death. In this study, two alternative models were tested in a hospitalized sample of 212 persons with an average life expectancy of six to 12 months. The first model proceeds on the assumption that those who wish to prolong life are characterized by unfinished business and fear of death, two stressors that are influenced by a number of background variables (e.g., education, gender, health, ethnicity). According to this model, the two stressors lead to the mobilization of family support, i.e., increased frequency of contact with family members and increased family salience, two variables which in turn mediate the impact of background variables on the desire to prolong life. In the second model it is hypothesized that contact with family members moderates the relationship between the two stressors and the desire to prolong life only in the case of family salience, a variable which in turn is associated with background variables like education, gender, health, or ethnicity. The results support the second model, indicating that family salience and family contact contribute to the definition of the context in which the stressors of unfinished business and fear of death prompt desires to prolong life, and that family contact has a different impact on attitudes toward death and dying for different ethnic groups.

1.3 Finitude and Valuation of Life

Rao et al. (1998) emphasize the influence of low spirits and social factors on attitudes toward death and dying. In a sample of community residents aged 81 years and older, those who thought about death more frequently were more likely to be unmarried and to manifest a greater severity of depressive symptoms. In interpreting their results the authors suggest a closer examination of attitudes toward death and dying would be useful in the assessment of depression and suicidal risk. Similarly, from the study of self-definition and time horizon in old age there is evidence that only in a minority of older people are feelings of finitude and finiteness predominant in their conceptualization of their future time perspective (Smith and Baltes 1999, Thomae 1981). An increase in feelings of finitude and finiteness is observed only in cases where life is subjectively supposed to be determined by limits and losses (Kruse 2000).

Lawton et al. (1999) proposed that the relationship between quality of life and years of desired life is mediated by an intervening cognitive-affective schema,

'valuation of life,' that may be defined as 'the extent to which the person is attached to his or her present life, for reasons related to a sense not only of enjoyment and the absence of distress, but also hope, futurity, purpose, meaningfulness, persistence, and self-efficacy' (p. 407). Since the usefulness of this concept was affirmed in an empirical study of 600 healthy and chronically ill older people aged 70 and older, further research is required in order to understand how health, cultural, social, and psychological factors become transformed into strong or weak valuations of life.

2. *Assisting Coping Processes in Institutional Settings*

In the last few years, the field of palliative medicine and care has developed new insights and concepts that could be used to improve treatment and care of dying people. However, the currently existing care systems are not sufficient.

Empirical research on factors influencing coping with death and dying underscore the demand for giving more attention to (a) pain therapy (most dying persons could be free from pain or at least be relieved from severe pain when therapy is offered), (b) activating care that enables people to continue to lead an independent and responsible life, even in the final stage of disease, and (c) ambulant care services that offer instrumental and emotional support for the family (Kruse 1991).

Giving assistance for coping with one's own finitude is complicated when the dying person is suffering from psychiatric disorders. This is especially the case since in the so-called fourth age (from 75–80 years on) vulnerability to chronic physical and psychological diseases increases, and changes in life expectancy go together with a steadily growing number of people dying from long-lasting illness instead of an acute health crisis. As a consequence, institutional care for the dying should be able to rely on psychiatric and psychological expertise. In this context it should be noted that psychiatric disorder (as a kind of serious illness) and the psychological reactions associated with the consciousness of finitude, death, and dying must be differentiated. It is important to distinguish between states of anxiety, despondency, resignation, and despair resulting from the confrontation with one's own finitude and similar emotional states resulting from psychiatric disorders since the diagnosis of a psychiatric disorder might suggest medical treatment. In contrast, the diagnosis of psychological reactions to the confrontation with one's own finitude suggests nonmedical treatment aimed at optimizing active and self-reflexive coping processes. Since coping with one's own finitude is an existential challenge for human beings, the temporary psychological reactions that occur during the process of active coping must not be interrupted by medical treatment. However, practitioners are obliged to provide adequate medical treatment when a changing of emotional state helps the patient's coping process. For adequate medication, differentiated knowledge about the development of psychological symptoms is needed. In order to be able to gain such differentiated knowledge, medical practitioners have to rely on information from caregivers. For this reason, there is a need for cooperation between medicine and care in institutional services trying to give assistance for dying people.

3. *Maintaining Self-responsibility*

The treatment and care of dying people suffering from changes of consciousness that prevent them from leading a self-responsible life should be based on the preferences and values articulated or supposed to be held by the patient before the occurrence of changes in consciousness. Medical ethics insist on enhancing the possibility for patients to lead a self-responsible life as long as possible. Self-responsibility can be interpreted in different ways; and at least five aspects of self-responsibility should be distinguished (Kruse 1991).

3.1 *Self-responsibility as Independence in Performing the Activities of Daily Living*

This aspect is too often neglected in discussing the possibilities for helping people cope with their own finitude, since the former is centered around illness and the corresponding possibilities and limits of therapy. However, Kruse (1987) found a close relationship between independence in performing the activities of daily living and the ability to cope with one's own finitude.

3.2 *Self-responsibility as Self-determination in Organizing Everyday Activities and Realizing Personally Meaningful Interests*

It should be considered that even in the process of dying, most people look for positive emotions and experiences that correspond with a feeling of participation from which they gain support for coping with dying. Studies show that religious activities can be helpful in this context (Musick et al. 1998).

3.3 *Self-responsibility as Autonomy in Choosing Social Contact and Managing Social Relationships*

Even if most people wish to die at home, it should be noted that what is to be regarded as an adequate amount of closeness to or distance from family members and relatives can be determined only by patients themselves (Baltes 1996).

3.4 Self-responsibility as the Ability to Understand the Development of Illness and the Effect of the Chosen Therapy

The quality of the interaction between the patient and the medical practitioner is an essential factor for the effectiveness of pain therapy (Lohmann et al. 1998).

3.5 Self-responsibility as the Opportunity to Decide Where to Die

Individuals cope best when they decide which kind of therapy or care to take and—especially when therapy leads to an extension of the process of dying—to terminate life-sustaining measures.

4. The Problem of Surrogate Decisions Concerning Life-sustaining Measures

In medical practice the question often is not whether prolonging life is possible but whether it is justified to provide life-sustaining measures for older people. When the patient is no longer competent, the offspring have to make the decision. According to Cooper-Kazaz et al. (1999) such surrogate decisions often do not reflect the patient's autonomous wishes, but rather the current feelings of the decision-maker. This is often an emotional reaction which sometimes is not even consistent with the offspring's own ethical outlook. In a study by Sonnenblick et al. (1993), approximately 50 percent of the offspring believed they knew their parent's wishes, but most of them did not comply with the previously stated wishes of their parent. Cooper-Kazaz et al. (1999) argue that attitudes toward the dying patient and the treatment appropriate are influenced by cultural, intuitive, and emotional factors. Moreover, exposure to the critical event (when parents become terminally ill) seems to have an impact on the children's judgement about life-sustaining measures. Those with terminally ill parents were significantly more likely to request the initiation of resuscitation and less likely to prefer that the decision be made by the physician than those who had no terminally ill parents. We therefore need more studies that investigate solutions not only of theoretical dilemmas, but of processes of decision-making under real-time conditions as well as more research on the factors which might influence the decisions of offspring.

5. Future Directions

Recent research suggests that the scope of the stages model is too narrow to account for individual differences in experiencing and coping with death and dying. It could be shown that coping processes depend on a number of personal and environmental factors, e.g., age, gender, ethnicity, social support, and religious activity. However, more research is needed to understand the complex interaction between these factors. Moreover, previous research has concentrated much on explaining interindividual differences in fear of death. However, advances in medical technology in particular point to the significance of other outcome variables that have been neglected too long. A broader perspective on attitudes toward death and dying should be helpful to integrate research on attitudes toward death and dying, thus assisting patients to cope in institutional settings, and to maintain self-responsibility and addressing the problem of surrogate decisions concerning life-sustaining measures. This integration is seen as an essential challenge for the psychology of death and dying.

See also: Caregiving in Old Age; Death and Dying, Sociology of; Death, Anthropology of; Differential Aging; Ecology of Aging; Mortality of the Oldest Old; Old Age and Centenarians

Bibliography

Baltes M M 1996 *The Many Faces of Dependency in Old Age*. Cambridge University Press, New York

Cooper-Kazaz R, Friedlander Y, Steinberg A, Sonnenblick M 1999 Longitudinal changes in attitudes of offspring concerning life-sustaining measures for their terminally ill parents. *Journal of the American Geriatrics Society* **47**: 1337–41

Fillip S-H 1999 A three-stage model of coping with loss and trauma: Lessons from patients suffering from severe and chronic disease. In: Maercker A, Schützwohl M, Solomon Z (eds.) *Post-traumatic Stress Disorder*. Hogrefe, Seattle, WA

Hummer R 1996 Black-white differences in health and mortality: A review and conceptual model. *The Sociological Quarterly* **37**: 105–25

Idler E 1987 Religious involvement and the health of the elderly: Some hypotheses and an initial test. *Social Forces* **66**: 226–38

Kruse A 1987 Coping with chronic disease, dying and death—a contribution to competence in old age. *Comprehensive Gerontology* **1**: 1–11

Kruse A 1991 Caregivers coping with chronic disease, dying and death of an aged family member. *Reviews in Clinical Gerontology* **1**: 411–15

Kruse A 2000 Zeit, Biographie und Lebenslauf (Time, biography, and life course). *Zeitschrift für Gerontologie und Geriatrie* **33**: 90–7

Kubler-Ross E 1969 *On Death and Dying*. Macmillan, New York

Lawton M, Moss M, Hoffman C, Grant R, Ten Have T, Kleban M 1999 Health, valuation of life, and the wish to live. *Gerontologist* **39**: 406–16

Lohmann R, Heuft G, Kruse A, Schneider G 1998 Pain, coping, and psychological well-being in late life. *European Journal of Pain* **2**: 43–52

Musick M, Koenig H, Hays J, Cohen H 1998 Religious activity and depression among community-dwelling elderly persons

with cancer: The moderating effect of race. *Journal of Gerontology* **53B**: 218–27
Mutran E, Danis M, Bratton K, Sudha S, Hanson L 1997 Attitudes of the critically ill toward prolonging life: The role of social support. *Gerontologist* **37**: 192–9
Rao R, Dening T, Brayne C, Huppert F A 1998 Attitudes toward death: A community study of octogenarians and nonagenarians. *International Psychogeriatrics* **9**: 213–22
Smith J, Baltes P B 1999 Trends and profiles of psychological functioning in old age. In: Baltes P B, Mayer K U (eds.) *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, New York
Sonnenblick M, Friedlander Y, Steinberg A 1993 Dissociation between the wishes of terminally ill parents and decisions by their offspring. *Journal of the American Geriatrics Society* **41**: 599–604
Thomae H 1981 Expected unchangeability of life stress in old age. A contribution to a cognitive theory of aging. *Human Development* **24**: 229–39

A. Kruse and E. Schmitt

Death and Dying, Sociology of

The sociology of death and dying is concerned with the role of mortality in social life. Many social thinkers believe that death poses the fundamental threat to the integrity and continuity of society. Social structures must provide a 'symbolic canopy' which guards against the terrors of extinction by providing meaning to human activity. Societies respond to this threat in a wide variety of ways ranging from the collectivist, death-accepting, rebirth orientation of the East, to the personal, death-defying, hedonism of the West. The sociology of death and dying, therefore, provides a unique perspective to investigate the core values of a culture. The discipline has been greatly influenced in content as well as methodology by psychiatry, anthropology, and social history. A main focus has been on changes in experience, attitudes and roles of the dying, the bereaved and their professional helpers. (See *Death*, *Anthropology of*; *Shame and the Social Bond*; *Suicide*; *Social Identity, Psychology of*; *Life Course: Sociological Aspects*; *Historical Sociology*.)

1. *The Broken Connection*

Traditional societies are characterized by widely shared and explicit norms of social behavior, in general, and of mortuary ritual, in particular (see *Death*, *Anthropology of*; *Sociology: Overview*). People knew how to behave when faced with death and dying. In many modern societies, individuals no longer have such clear social rules to guide their behavior. They often feel confused or uneasy in these circumstances. Although the roots of this change are complex, it appears related to a fundamental change in the traditional nature of society as composed of *both* the living and the dead. 'If the structure of African cultures is that of a dialogue (between the living and the dead), then the structure of Western society is that of a monologue—but the monologue of the dead' (Bastide 1968, p. 104). Lifton (1978) proposed the term 'the broken connection' to describe how the link between the living and the dead has been severed in contemporary society. Conceptually, the issue can be put in terms of a question: Where are the dead? 'In the West, post- death conceptions typically involve the integrity and continuity of one's personal self. In the East, the ultimate goal is often an undifferentiated and impersonal oneness with the universe' (Kearl 1989, p. 27). The literal and symbolic responses to this question, therefore, continue to provide key insights into the deep structure of society.

2. *Impact of Recent Social Trends*

Since the mid-twentieth century, there have been a series of dramatic changes in how death and terminal illness are perceived in the West. These changes reflect recent social trends such as secularization, medicalization, and more recently, death awareness (see *Secular Religions*; *Medical Sociology*; *Theory: Sociological*; *Modernity*).

2.1 *Secularization*

From a religious perspective, dying was the 'defining moment' in a person's life (Seale 1998). Typically, a religious authority would assist at this feared rite of passage to the 'next world.' With the decline of religious belief, in general, and fear of damnation, in particular, death lost much of its sacred or theological importance. Instead death became a time of separation from loved ones at the ending of a human life (Kearl 1989) while dying increasingly became a private act. The shift from religious to secular is illustrated by changes in the image of a 'good death.' Modern secularists wished for an 'unconscious death,' a painless sudden demise during sleep, which in the Age of Belief was considered an 'accursed death' since it left no opportunity for the deceased to prepare to meet his or her Maker (Aries 1974). Gorer (1965), in a widely cited monograph, argued that death had replaced sex as the modern form of pornography, an embarrassing taboo topic. Becker (1974) claimed that although fear of death was universal, modern society was characterized by the pathological 'denial of death' (see *Death, Anthropology of*; *Cognitive Dissonance*; *Psychoanalysis in Sociology*).

Secularism also contributed to the rise of lay professionals as part of a funeral industry, especially in the USA, which exploited the denial of death. The

resulting commercialism contributed to enormous and seemingly unnecessary costs which was brought to international attention by the popular critique, *The American Way of Death* (Mitford 1963).

2.2 Medicalization

As the majority of people began to die in hospitals, death and dying gradually came under medical control. The resulting impersonal bureaucratic care led to their physical and emotional isolation, characterized by 'the loneliness of the dying' (Elias 1985). The dramatic rise of life support technology, moreover, created situations in which a person was neither fully alive nor completely dead. As a result, medicalized death lost its status as a defining moment, but became instead an arbitrary time when the machines were finally turned off. 'Death in the hospital is no longer the occasion of a ritual ceremony, over which the dying person presides amidst his assembled relatives and friends. Death is a technical phenomenon obtained by the cessation of care' (Aries 1974, p. 88). This situation also created a need for new definitions of death, the most widely accepted being that of brain death or, more controversially, the demise of the cerebral cortex, associated with the irreversible loss of personhood. Brain death, which opened up the possibility for the transplantation of organs, was widely accepted in the West but remained highly controversial in Japan (Seale 1998).

2.3 Death Awareness

The denial of death ultimately led to 'the revival of death' (Walter 1994) in the form of a movement for death awareness. As an ideology, it argued forcefully for openness and a life-affirming attitude in the face of death. As a social movement, it led to creation of new institutions for caring for the dying and their families. It is significant that the two best known and charismatic figures of the movement were women: Elizabeth Kubler-Ross, an Austrian-born psychiatrist who worked in a large county hospital in Chicago, and Cicely Saunders, a social worker, nurse, and committed Christian who retrained as a physician in order to be 'taken seriously' (Cline 1996). Although both were physicians, they sought to demedicalize dying by reasserting the 'feminine values' of mutuality, equality, and the 'affirmation of love through intimate talk' (Seale 1998, p. 176). 'Coming to terms with my own death or with bereavement is increasingly seen as an inner psychological process in which others are marginal, except to provide "support" for me as I go through a "natural" process' (Walter 1994, p. 189). This psychologization of death and dying often means that those who accompany the dying and bereaved are caring professionals and members of self-help groups who have undergone similar losses and function in place of the traditional community (see *Community, Social Contexts of*; *Community Sociology*).

3. Sociology of the Dying

A landmark in the sociological study of dying patients was the work of Glaser and Strauss (1965, 1968) who developed two frameworks for analyzing the interaction between the dying person and the medical staff: contexts of awareness and dying trajectories. 'Contexts of awareness' referred to the degree of shared knowledge between patient and doctor concerning the fatal prognosis which ranged from 'closed,' if only the doctor knew, through intermediate stages ('suspicious,' 'mutual pretense') to 'open' when both parties could freely discuss the terminal illness. At the time of the original study, closed awareness was the norm, and it remains common in certain countries like Japan or Italy (Walter 1994, Seale 1998).

3.1 Dying Trajectories

Dying trajectories are 'perceived courses of dying rather than the actual courses' (Glaser and Strauss 1968, p. 6). The trajectories were based on the 'time for dying' (Glaser and Strauss 1968) as expressed in a pair of dichotomies: expected/unexpected, sudden/lingering. There were four basic dying trajectories: sudden unexpected, most commonly seen in the Emergency Room; sudden expected, common in intensive care units; lingering unexpected as in the uncertain course of an AIDS patient; lingering expected, as in the final stages of cancer. Each hospital ward was thus geared to a specific dying trajectory and used specific strategies to maintain its routine or 'sentimental order.' Medical personnel and family members experienced most distress when an actual trajectory deviated from its predicted course. Glaser and Strauss's work contributed to changes in the training of doctors and nurses. The social history of death and dying (Aries 1974) provided an important comparative critical perspective of contemporary 'invisible death' and one more voice calling for a return to 'death with dignity.'

4. Hospices

Death in traditional society was often characterized by sudden unexpected dying trajectories, as is still common in the Third World. In modern societies, the lingering death of the elderly from chronic diseases became the norm. Their death was less disruptive since

they had undergone a 'social death' in which they were disengaged from their previous social roles and obligations. At the same time, hospitals with their 'fiction of probable recovery' (Glaser and Strauss 1965, 1968) were increasingly criticized as inappropriate places for the care of the dying (Kubler-Ross 1969, Walter 1994).

The first hospice, an institution for the special needs of the dying and their families, was founded by Cicely Saunders in 1967 (Cline 1996). Its name deliberately recalled 'a much older medieval tradition when hospices were wayside places of rest and refuge for pilgrims' (Seale 1998, p. 113). The dying were thus implicitly considered individuals on a spiritual quest. The hospice ideology was holistic, providing physical, emotional, social, and spiritual care; egalitarian, with expertise based on knowledge and skills rather than status or formal qualifications; and conceived of the entire family, not merely the patient, as the 'unit of care.' It had a strong sense of mission and charismatic leadership. This ideology differed radically from the specialized, hierarchical, disease orientation of the 'medical model,' in which death was often experienced by the physician as a personal or technological failure (see *Ideology, Sociology of*; *Medical Sociology*). Instead the dying were encouraged to undergo a psychological and spiritual journey toward final reconciliation (Kubler-Ross 1969) and personal authenticity (Becker 1973). The 'home care' option allowed some individuals to die at home. 'A major reward for those dying an aware death was that of emotional accompaniment involving the affirmation of a secure social bond … If I hadn't known, how could I have helped him …' (Seale 1998, pp. 175–6).

The hospice as an institution has undergone phenomenal growth in its first 30 years to number over 5,000 worldwide at the end of the twentieth century (Seale 1998, p. 114). With growth has come a predictable trend away from charismatic leadership toward greater routinization, professionalization, and 'managed death' (Seale 1998) (see *Charisma: Social Aspects of*; *Organizations, Sociology of*). Hospice and palliative care will remain important for future investigation.

5. *Stages of Dying*

Based on interviews with cancer patients, Elizabeth Kubler-Ross devised an influential schema of five stages to describe the emotional response of terminal patients: denial ('Not me!'), anger ('Why me!'), bargaining (trading a pious deed for more time), depression (mourning the lost opportunities) and final reconciliation (a peaceful, if not happy, 'OK, it is me'). Kubler-Ross did state that not everyone goes through all five stages nor necessarily in this order, but rather the process of dying is highly individualistic. Nevertheless, these stages did represent a breakthrough at the time for conceptualizing the ideal emotional journey of the dying. Parkes (1986) devised a similar set of stages for the grief process: numbness, pining, disorganization, despair, and finally, reorganization and recovery. From a sociological perspective, the communal aspects of dying are notably absent from these stage theories. Rather, they reflect the social shift to individualism. Each dying person is encouraged to embody the 'Sinatra syndrome' (Walter 1994) and be able to say, in words of his well-known song, 'I did it my way.' The emphasis on 'the discourse of therapy' in place of shared ritual makes the dying person in a sense 'the chief mourner' (Seale 1998). Despite methodological and other criticisms, these stage theories remain highly influential.

6. *Euthanasia*

Euthanasia derives from the Greek, meaning literally 'good death' but in practice refers to right of a dying person to commit suicide alone or with the help of a physician. Hospice ideology supports 'passive euthanasia,' the withholding of treatment which may result in death but firmly rejects 'active euthanasia,' any active measures designed to end life, which is legal only in The Netherlands. Proponents of active euthanasia place great emphasis on autonomy and the senselessness of pain and suffering. In a broader sense, they desire to 'preserve an intact narrative of self-identity up to the point of death' (Seale 1998, p. 8). Opponents of euthanasia usually believe that the sanctity of life places limits on individual choice and fear coercive use by family members or unethical authorities as occurred in Nazi Germany. The right to die will remain an important topic for future research and public debate.

7. *Sociology of the Dead: Funerals—Ritual Uniformity*

In contrast to the widespread interest in the dying, much less attention has been given to the sociology of the dead. Despite the social changes noted above, funeral rituals themselves have remained largely unchanged (Walter 1994). This certainly appears true in the USA where, despite the tremendous hetrogeneity in terms of religion, ethnicity, and social class, the basic structure of funerals remain remarkably uniform: rapid removal of the corpse to the funeral parlor, embalming (which many individuals believe, incorrectly, is required by law), institutionalized viewing of the cosmetically restored corpse, and disposal by burial, with a reliance upon professional or religious experts (Kearl 1989). Likewise in the UK, where a majority of the dead are cremated, atheists, Christians, and Hindus undergo similar treatment in

crematoria (although Hindus are likely to stay to watch the burning body). Except for the practice of dispersing ashes in places of personal significance, only a tiny minority perform alternative 'do it yourself' ceremonies (Walter 1994).

8. AIDS and the Imagined Community of Mourners

The AIDS epidemic has refocused interest on 'the way we die' (Nuland 1994) and the social meaning of such deaths (Ariss 1997). The shared fate of HIV infection has consolidated a local sense of community and at times a larger global one. One of the most poignant exemplars of such a community of living and the dead is the 'AIDS Memorial Quilt,' begun in 1987 by gays and their friends to commemorate the lives of individuals who died of AIDS. By 1992, the Quilt had more than 20,000 memorial panels. 'The quilt … shows how death of the most fearful kind can draw people together' (Walter 1994, pp. 194–5). Likewise the dramatic sudden death of high profile individuals (e.g., President Kennedy, Princess Diana) may temporarily recreate the solidarity of an 'imagined community' of mourners who vicariously participate in traditional ceremonies via television and the Internet. This analysis recalls the seminal work of the great French sociologist, Emile Durkheim (1995) who showed how mourning ritual reinforced the cohesiveness and core values of the social group. While the Internet is becoming an important focus of investigation, many traditional sociological topics (gender, ethnicity, social class) remain relatively neglected (Cline 1996).

A central task for the sociologist is to explore the dialectic of how death can threaten the basis of society or, alternatively, enhance its solidarity.

See also: Aging and Health in Old Age; Death and Dying, Psychology of; Death Penalty; Euthanasia; Mortality, Biodemography of; Mortality of the Oldest Old

Bibliography

Aries P 1974 *Western Attitudes Toward Death*. Marion, London

Ariss R M 1997 *Against Death: The Practice of Living with AIDS*. Gordon and Breach, Amsterdam

Bastide R 1968 Religions africaines et strutures des civilisations. *Presence Africaine* **66**: 82–105

Becker E 1973 *The Denial of Death*. Free Press, New York

Cline S 1996 *Lifting the Taboo: Women, Death and Dying*. Abacus, London

Durkheim E 1995 [1912] *The Elementary Forms of the Religious Life*. Free Press, New York

Elias N 1985 *The Loneliness of the Dying*. Blackwell, Oxford, UK

Glaser B, Strauss A 1965 *Awareness of Dying*. Aldine, Chicago

Glaser B, Strauss A 1968 *Time for Dying*. Aldine, Chicago

Gorer G 1965 *Death, Grief and Mourning in Contemporary Britain*. Cresset, London

Kearl M 1989 *Endings: A Sociology of Death and Dying*. Oxford University Press, Oxford, UK

Kubler-Ross E 1969 *On Death and Dying*. Macmillan, New York

Lifton R J 1978 *The Broken Connection: On Death and the Continuity of Life*. Simon and Schuster, New York

Mitford J 1963 *The American Way of Death*. Simon & Schuster, New York

Nuland S 1994 *How We Die*. Chatto & Windus, London

Parkes C M 1986 *Bereavement: Studies of Grief in Adult Life*. Penguin, Harmondsworth, UK

Seale C 1998 *Constructing Death: The Sociology of Dying and Bereavement*. Cambridge University Press, Cambridge, UK

Walter T 1994 *The Revival of Death*. Routledge, London

H. Abramovitch

Death, Anthropology of

All human cultures struggle to deal with the inevitability and mystery of death. The anthropology of death explores how human societies around the world respond to death. It is concerned with both the conceptual and organizational aspects, that is, what people believe about death and the afterlife, as well as what they actually do when faced with the crisis of death. Two main methodological approaches used are the ethnographic and the comparative. The ethnographic approach examines how single cultures cope with the demise of members, while the comparative approach tries to make sense of the enormous cultural variation in issues such as the disposing of the corpse, the expected behavior of the bereaved, and the ongoing relations between the living and their dead (see *Anthropology*; *Ethnography*; *Ethnology*).

Although elaborate death rituals are probably a defining aspect of human culture, comparative studies have revealed few, if any, universal practices. Even the widespread practice of crying at a funeral is actively discouraged, for example, among the Balinese. Nevertheless, anthropologists have uncovered a number of key metaphors, which help to make sense of the enormous diversity of mortuary rituals. The anthropology of death takes as its task to understand the phrase: 'All humans die,' yet in every culture, each dies in their own way (see *Death and Dying*, *Sociology of*; *Ancestors*, *Anthropology of*).

1. Centrality of the Death Ritual

In the West, death and mourning rituals are often considered to be private, and Western ethnographers have often reported a reluctance to intrude into this

seemingly intimate realm (Palgi and Abramovitch 1984; see *Fieldwork in Social and Cultural Anthropology*). In many traditional cultures, however, death rituals stand at the center of social life. Most societies use expensive and elaborate mortuary ritual as a way of demonstrating status and power, so that the expense incurred is often an enormous financial burden. The atmosphere is not always sad or somber, but may even take on a festive atmosphere so that one may speak of 'celebrations of death' (Huntingdon and Metcalfe 1991).

The impact of the death is related directly to the social status of the deceased. Death is likely to be seen as particularly disruptive when it strikes persons who are most relevant for the functional and moral activities of the social order: 'At the death of a chief, or man of high standing a true panic sweeps over the group' (Hertz 1960, p. 76). The elaborate funeral rituals and pyramids of Ancient Egypt are outstanding examples of this ethos. Likewise, the death of a spouse often leads to a long period of taboos and restricted activity. 'On the contrary, the death of a stranger, slave or child will go almost unnoticed; it will arouse no emotion, occasion no ritual' (Hertz 1960, p. 76). In regions of high mortality, the death of an infant who is not yet considered a 'social person' may have no formal ritual or mourning. Bereaved parents often experience their loss privately, without ceremony, in what has been called 'death without weeping' (Scheper-Hughes 1992). The simple funerals of some hunter–gatherers, such as the Baka Pygmies of the Central African rain forest, are notable exceptions to the widespread pattern of elaborate and complex mortuary rituals (Bloch and Parry 1982; see *Hunting and Gathering Societies in Anthropology*).

2. *Death as an Event vs. Death as a Process*

In the West, despite recent legal and medical controversies (see *Euthanasia*) death is considered as occurring at a specific identifiable moment, symbolized by the 'time of death' on the death certificate. This punctual view of death (Bloch 1988) in which a person is thus either alive or dead is not shared by many cultures. Physiological demise and social death do not necessarily coincide. Instead, death is seen as part of a long, complex, and even dangerous process, with no sharp boundary between life and death. Cultures express the degree of alive/dead using different metaphors. The Merina of Madagascar, for example, use the image of wet/dry to indicate degrees of alive/dead. A newborn infant is very wet and thus very alive; in contrast, a shriveled elderly person, almost totally dry, is mostly dead. The process of drying, and hence dying, will be only completed long after biological death when the bones are dug up and placed in communal tombs (Bloch and Parry 1982).

3. *Theoretical Approaches and Key Metaphors*

'Few areas of contemporary anthropological inquiry are still so dominated by *fin de siècle* thinking as is the study of death and mortuary ritual' (Palgi and Abramovitch 1984, p. 387). The three traditional paradigms are: French sociological school (Durkheim, Hertz), early British functionalists (Malinowski, Radcliffe-Brown) and 'rite of passage' approach (Van Gennep) (see *Anthropology*; *Structuralism*; *Functionalism in Anthropology*; *Liminality*; *Personhood, Anthropology of*; *Ritual*;).

3.1 *Hertz's Analysis of First and Second Funerals*

The single most influential text in the anthropology of death was written by a brilliant disciple of Durkheim, Robert Hertz, who was killed on the battlefield of Verdun in World War I. Many of the central ideas in the anthropology of death: That death is widely perceived as a process and not as an instantaneous event, that death is the starting point of a ceremonial process whereby a dead person becomes an ancestor, and that death is akin to an initiation into a social afterlife and hence resembles a kind of rebirth, all derive from his seminal monograph (Hertz 1907). His study focused on the widespread custom of second funeral, or more correctly, secondary treatment of the remains. One of the main ideological reasons for a double ceremony, often separated by years, has to do with the dual nature of the person as composed of both body and spirit (or soul). Dead bodies or corpses are usually associated with ritual pollution and sorrow. The first funeral, in a literal and symbolic sense seeks to get rid of these polluting and sorrowful aspects. The second funeral is focused on the initiation of dead people's spirits into the realm of the ancestors, where they will continue to serve as a cultural resource, watching over and guiding the living. This structuralist perspective highlighted parallels between the decomposition of the corpse and the fate of the soul.

3.2 *Death Rituals and Core Values*

In many societies, the occurrence of a death disrupts social life severely, so that 'it is stricken in the very principle of its life in the faith it has in itself' (Hertz 1960, p. 78). A functionalist perspective stressed how mortuary rites served to resolve the disruptive tendencies that operate at times of social crisis. Malinowski demonstrated how ceremonies counteract the centrifugal forces of fear, dismay, and demoralization associated with death, and provide a powerful means of reintegration of the group's shaken morale (Palgi and Abramovitch 1984). These rituals must provide an answer to the meaning of life for the community at a

time when it is most threatened. As a result, the study of such death rituals provide a unique opportunity for studying the core values of any culture. The functionalist perspective emphasized the problem of death for society, and especially the issues of inheritance, redistribution of rights, and statuses, as well as the reintegration of mourners into day-to-day life.

3.3 Death as a Journey or Rite of Passage

Most cultures conceptualize death as a transition, or rite of passage. In many cultures, this transition is seen as a journey to an ultimate destination that may culminate in rebirth, ancestral abode, reunion with nature or Divinity, or indeed total oblivion. Death rituals, like all rites of passage, have a three-part structure, first delineated by Van Gennep: Separation, liminality, and reincorporation (Huntingdon and Metcalfe 1991). The spirits of dead people must be separated from their social roles as members in the community of the living and enter an undefined 'in-between' state, finally being reincorporated into a new status as the end of the 'journey.' Often, the fate of the soul is modeled on the fate of the body. Just as the soul of the deceased is formless and repulsive, as in Borneo when the dead are left to ferment in large jars, so the soul of the deceased is seen as homeless and the object of dread. Only when the bones alone are left can the ritual journey be completed (Huntingdon and Metcalfe 1991). Likewise, the ritual status of the mourners is in turn linked to the fate of the corpse in its process of decay, and the soul in its progress on its journey. Some cultures do have a specific tie-breaking ritual, which symbolizes both the end of mourning for the bereaved, and the end of the journey for the soul or spirit of the deceased. The living are often required to nurture the deceased at each stage along the journey through the afterlife and this conception helps to explain why the dead take up so many economic, emotional, and spiritual resources.

3.4 The Danger of the Unincorporated Dead

From this rite of passage perspective, it is possible to understand why the unincorporated dead, trapped permanently in the liminal realm, are often considered as dangerous. These wandering spirits, for whom no rites were performed, may act as hungry ghosts. They yearn to be reincorporated into the world of the living, and since they cannot be, they behave like hostile strangers who lack the means of subsistence which other dead find in their own world, and consequently must find at the expense of the living. Many cultures use elaborate strategies to confuse the spirit of the deceased so that it will not return to the realm of the living. Illness, misfortune, and associated healing rituals are often attempts to incorporate these lost souls (see *Health: Anthropological Aspects*; *Shamanism*; *Spirit Possession, Anthropology of*; *Magic, Anthropology of*; *Witchcraft*). The 'tomb of the Unknown Soldier' in modern states, provides a resting-place for unincorporated spirits of the war dead (see *Collective Memory, Anthropology of*).

3.5 Parallels to Other Life Cycle Rituals

Anthropologists have also noticed how death rituals often parallel other life cycle rituals, notably births and weddings. Conceptually, death is often seen as a rebirth into another world, while birth is viewed as the literal rebirth of a deceased ancestor (see *Birth, Anthropology of*). Likewise, for weddings, just as a bride may leave her family, so too the soul of the deceased departs for another world. In rural Greece, for example, the same lyrics are sung at both weddings and funerals (Danforth 1982).

3.6 Emotional Response of the Mourners

The comparative study of how mourners respond to loss and bereavement has attracted considerable attention (Parkes et al. 1997). Almost all cultures allow, or even encourage, the expression of crying, as well as fear and anger by mourners. Women tend to cry and self-mutilate, while men express anger, or direct aggression away from themselves. Yet most societies have developed mechanisms and institutions to control the anger of the bereaved, especially via the use of ritual experts (Rosenblatt et al. 1976). In many cultures, there is no cultural category of a 'natural death'; rather, each death is akin to murder in that a culturally sanctioned cause of death must be uncovered. Almost all societies concern themselves with the spirits of the dead, which probably give expression both to the 'unfinished business' between the living and the dead, and the fact that death does not end a relationship. The more complicated the relationship and unfinished business in which the dead are dispossessed in favor of the living, the greater will be the dread and concern about ghosts (see *Psychological Anthropology*).

3.7 Good Death vs. Bad Death

Recent interest has focused on cultures' conception of what constitutes a 'good death' or a 'bad death.' A good death represents a cultural ideal that enacts a symbolic victory over death, and the regeneration of life (Bloch and Parry 1982). A bad death does the opposite, leaving survivors despairing, helpless in the face of meaninglessness, evil, or nothingness. Unpredictability, violence, or intentional harm are widespread attributes of a bad death. Archetypal examples of bad death include suicide, homicide, and death in traffic accident. A good death is the mirror image of a bad death: An expected, painless death of an elderly individual with a multitude of descendants in at-

tendance. But a good death has culture-specific aspects, such as the setting, timing, physical posture, thoughts and actions at the final moment, and other 'near death experiences,' as well as a preferred cause and manner of death (Abramovitch 1999). An Indian Hindu ideally should die in the open air, as an act of self-sacrifice, abandoning life to the sound of the chanting of the names of God, for his thoughts at that moment may determine his subsequent fate and rebirth (Parry 1994).

3.8 Future Directions

The anthropology of death traditionally focused relatively narrowly on the death of single individuals in small-scale societies. Future research, in contrast, will need to investigate how local cultural traditions interact with the wider collective and global context. The study of mass death, in the form of epidemics, natural disasters, or political violence, provides one stimulating area of investigation (see *Genocide: Anthropological Aspects*; *Holocaust, The*; *Collective Memory, Anthropology of*; *Terrorism*; *Disasters, Sociology of*). The diverse impact of new technologies may provide another intriguing area for understanding how conceptions of death are changing. Web pages on the Internet provide for new forms of commemorating the dead. Organ donation allows a 'bad death' to be transformed into a 'good death' as a posthumous act of self-sacrifice, literally dying so that others may live. Cloning awakens in some the 'myth of eternal life'; while artificial life support has created a new state in which persons are neither alive nor dead.

Death will always remain a mystery; nevertheless, the cross-cultural investigation of death will continue to provide insights into how humans cope with that mystery.

Bibliography

Abramovitch H 1999 'Good death' and 'bad death.' In: Malkinson R, Rubin S, Witztum E (eds.) *Traumatic and Nontraumatic Loss and Bereavement*. Psychosocial Press, New York

Ariès P 1974 *L'Homme devant la Mort*. Editions du Seuil, Paris [Republished, trans. Weaver H 1982 *The Hour of Our Death*. Knopf, New York]

Bloch M 1988 Death and the concept of person. In: Cederroth S, Corlin C, Lindstrom J (eds.) *On the Meaning of Death: Essays on Mortuary Rituals and Eschatological Beliefs*. Uppsala Studies in Cultural Anthropology, 8. Acta Universitatis Upsaliensis, Uppsala, Sweden

Bloch M, Parry J (eds.) 1982 *Death and the Regeneration of Life*. Cambridge University Press, Cambridge, UK

Danforth L 1982 *The Death Rituals of Rural Greece*. Princeton University Press, Princeton, NJ

Hertz R 1907 Contributions à une étude sur la representation collective de la mort. *Année Sociologique*. **10**: 48–137 [Republished, trans. R Needham, C Needham 1960 *Death and the Right Hand*. Free Press, New York]

Humphreys H, King H (eds.) 1981 *Mortality and Immortality: The Anthropology and Archeology of Death*. Academic Press, London

Huntingdon R, Metcalf P 1991 *Celebrations of Death: The Anthropology of Mortuary Ritual,* 2nd edn. Cambridge University Press, Cambridge, UK

Palgi P, Abramovitch H 1984 Death: A cross-cultural perspective. *Annual Review of Anthropology* **13**: 385–417

Parkes C, Laungani P, Young B (eds.) 1997 *Death and Bereavement Across Cultures*. Routledge, London

Parry J 1994 *Death in Benares*. Cambridge University Press, Cambridge, UK

Rosenblatt P, Walsh R, Jackson A 1976 *Grief and Mourning in Cross Cultural Perspective*. Human Relations Area Files Press, New Haven, CT

Scheper-Hughes N 1992 *Death Without Weeping: The Violence of Everyday Life in Brazil*. University of California Press, Los Angeles

H. Abramovitch

Death Penalty

1. The Death Penalty and Democracy

Almost alone among so-called Western, liberal democracies the USA continues to use the death penalty (Zimring and Hawkins 1986, White 1991). As a result, researchers are increasingly directing their attention not only to the efficacy of capital punishment but also to its compatibility with democratic values and its impact on the cultures in which it is used (see Zimring 1999). This research is animated by the recognition that the right to dispose of human life through sovereign acts was traditionally thought to be a direct extension of the personal power of kings (Foucault 1979). With the transition from monarchical to democratic regimes, one might have thought that such a vestige of monarchical power would have no place and, as a result, would wither away. Yet, at least in the USA, the most democratic of democratic nations, it persists with a vengeance.

How are we to explain this and what are its consequences for the political system and the values and norms of the cultures which practice state killing? What judgments can we make about the political and cultural consequences of capital punishment?

It may be that the persistence of state killing in the USA is paradoxically a result of a deep attachment to popular sovereignty (Sarat 1995). Where sovereignty is most fragile, as it always is where its locus is in 'the People,' dramatic symbols of its presence, like capital punishment, may be most important. The maintenance of capital punishment is, one might argue, essential to the demonstration that sovereignty could reside in the people. If the sovereignty of the people is to be genuine, it has to mimic the sovereign power and prerogatives of the monarchical forms it displaced and about whose sovereignty there could be few doubts.

The death penalty is miraculously transformed from an instrument of political terror used by 'them' against 'us,' to our instrument allegedly wielded consensually by some of us against others (Connolly 1999). Thus capital punishment may be a key to understanding modern mechanisms of consent (Aladjem 1990, Dumm 1987).

Not only does the USA cling tenaciously to capital punishment long after almost all other democratic nations have abandoned it, but both the number of people on death row and the number executed grows steadily. In addition, while for a brief period of time after *Gregg v. Georgia* (428 US 153 1976) reinstated capital punishment there was tight legal supervision of the death penalty and great restraint in its use, that period is now long gone (Weisberg 1983). The pressure is on to move from merely sentencing people to death and then warehousing them to carrying out executions by reducing procedural protections and expediting the legal process (Amsterdam 1999).

As any American who lived through the 1970s, 1980s, and 1990s knows, the politics of law and order have been at center stage for a long time. From George Wallace's 'law and order' rhetoric to Bill Clinton's pledge to represent people who 'work hard and play by the rules,' crime has become such an important issue that some now argue that we are being 'governed through crime' (Simon 1997). In the effort of politicians to show that they are tough on crime the symbolism of capital punishment has been crucial (Connolly 1999). It has also been crucial in the processes of demonizing young, black males (Tonry 1995) and using them in the pantheon of public enemies to replace the 'evil empire.' No US politician today wants to be caught on the wrong side of the death penalty debate.

The politics of capital punishment is crucial in an era of neoliberal withdrawal from other kinds of state intervention. It helps to shift the symbolic capital of the state to an area where, whatever one's doubts about the capacity of government to govern effectively, there can be demonstrable results (Connolly 1999). As Aladjem (1992, p. 8) puts it, 'It is significant that just when there is doubt as to whether rational democratic institutions can generate a public sense of justice, there has been a resurgence of pre-democratic retributivism and a renewed insistence that it is a "rational" response to an increase in violent crime.' A state unable to execute those it condemns to die would seem too impotent to carry out almost any policy whatsoever (see Sarat 1997).

2. *The Policy Agenda*

Because the death penalty is the ultimate sanction, and because the maintenance and use of capital punishment in and by the criminal justice system has been so controversial, social scientists have long gravitated to it as a subject for policy relevant, empirical research. In an earlier era, much of the agenda of social science research on the death penalty was designed to speak to issues that were thought to be important in the ongoing campaign against capital punishment (see Munger 1993, pp. 5–6, Zimring 1993, pp. 10–11). Among the topics traditionally covered by death penalty researchers three have been most important, namely the question of whether the death penalty deters, whether it is compatible with contemporary standards of decency, and whether it has been administered in a racially neutral or in a discriminatory manner.

2.1 *Deterrence*

With respect to the first question, despite what commonsense suggests, most social science research on capital punishment finds no evidence that the death penalty deters (Klein et al. 1986). To say that there is little or no evidence that the death penalty deters may seem strange, but it becomes more comprehensible when one realizes that the comparison made in the deterrence calculus is between a death sentence and a sentence of life in prison, not between a death sentence and no punishment. Moreover, because many murders are crimes of passion they may not be deterrable through any criminal justice policy (see *Deterrence*; *Deterrence: Legal Perspectives*; *Compliance and Obedience: Legal*).

Studies of the deterrent effect of capital punishment have used many different methodologies. Some have examined murder rates before and after well-publicized executions; others have compared murder rates in adjoining states, one of which has capital punishment, the other of which does not (Sellin 1959). Some of the most controversial research on the deterrent effect of capital punishment was done in the 1970s by the economist Isaac Ehrlich (1975). This research was controversial because unlike most other scholarship it showed strong and statistically significant deterrent effects, and because it used sophisticated statistical techniques to control for other confounding variables. If Ehrlich was correct execution did indeed save lives. Indeed Ehrlich (1975, p. 398) claimed that 'On the average the tradeoff between the execution of an offender and the potential victims it might have saved was on the order of magnitude of 1 to 8 for the period 1933–67 in the United States.' But, since Ehrlich's research was published it has been the object of sustained and rather persuasive criticism (see, e.g., Zeisel 1976, Glaser 1977, Lempert 1981, Klein et al. 1986), most of which has focused on the particular statistical techniques that he used. Today, despite Ehrlich's research, few social scientists believe that the death penalty has distinct and discernible deterrent effects (see Gibbs 1986, p. 348). Thus if the justification for capital punishment is to be found it must be found elsewhere.

2.2 Public Opinion and Evolving Standards of Decency

The most powerful alternative justification suggests that even if the death penalty does not deter it is justified on retributive grounds or as a way of expressing society's legitimate moral condemnation of heinous criminality (see Berns 1979, Von den Haag 1975). Social science research has tried to tap into the retributive dimensions of capital punishment and to determine whether the death penalty is compatible with contemporary standards of decency through research on public opinion (see Vidmar and Ellsworth 1974). Does the public support the death penalty? Do people view it as a legitimate expression of society's justifiable search for vengeance? (See *Public Opinion: Political Aspects*; *Public Opinion: Microsociological Aspects*; *Legal Issues: Public Opinion*.)

At first glance evidence on the first question seems to be quite clear. Social scientists have repeatedly documented strong public support for capital punishment (see, e.g., Bohm 1991, Fox et al. 1990–91). Moreover, support for the death penalty is highly correlated with retributive attitudes, suggesting that it is a particular view of justice rather than the death penalty's utility that helps explain why it persists in the USA (Ellsworth and Gross 1994). Yet research has also demonstrated the malleability of this public embrace of capital punishment and its retributive justification.

In the 1970s Sarat and Vidmar (1976), taking up a hypothesis first advanced by Justice Marshall in *Furman v. Georgia* (408 US 238 1972), found that the more that people know about the death penalty, such as how it works, and the evidence concerning deterrence, the less they support it. More recently Bowers (1993) has shown that public support for capital punishment decreases dramatically when respondents in surveys are presented with alternative forms of punishment and asked to choose which they prefer. Bowers (1993, also Steiner et al. 1999) found that people tend to accept the death penalty because they believe that currently available alternatives are insufficiently harsh or meaningful, but that when asked whether they preferred the death penalty or life without parole combined with a restitution requirement, 'expressed death penalty support plummeted' (Bowers 1993, p. 163).

2.3 Discrimination

The third area which traditionally has engaged the interest of social scientists is one which itself raises questions about the compatibility of capital punishment and democratic values. This area involves the question of the fairness of the way capital punishment is administered and more particularly whether it is administered in a racially discriminatory manner (see Zeisel 1981, Wolfgang and Riedel 1973). Prior to *Furman*, research tended to focus on race of the offender effects. In studies of rape cases (Wolfgang and Reidel 1975) and of homicide cases (see Gross and Mauro 1984), social scientists documented such effects, showing, for example, that between 1930 and 1967 almost 50 percent of those executed for murder were black (Gross and Mauro 1984, p. 38). Yet this early research was unable to disentangle the impact of race on capital sentencing from the impact of other 'legitimate' factors. Merely showing that the death penalty was more likely to be imposed on black defendants could not, in itself, establish that that difference was the result of racial discrimination (see *Crime and Ethnicity (Including Race)*; *Discrimination: Racial*; *Discrimination*; *Race and the Law*; *Inequality*).

In the mid-1980s Professor David Baldus and his colleagues (1990) undertook research in Georgia which was designed to remedy this defect. Using sophisticated multiple regression techniques and a large data set, they set out to isolate the effects of race on capital sentencing. They found several things that were highly illuminating about the question of fairness and whether capital punishment is compatible with democratic values. First they found no evidence of discrimination against black defendants because of their race, in the period after *Furman*. On the other hand, they found strong race of the victim effects. Taking over two hundred variables into account Baldus concluded that someone who killed a white victim was 4.3 times more likely to receive the death penalty than the killer of a black victim. Juries, or so it seemed, even after the sustained efforts of the Supreme Court to prevent arbitrariness or racial discrimination in capital sentencing, still valued white life more highly than the lives of nonwhites and were, as a result, more likely to sentence some killers to die on the basis of illegitimate considerations of race (see also Baldus et al. 1999, Paternoster 1983, Gross 1985). (See *Juries*; *Courts and Adjudication*.)

3. Beyond the Policy Agenda

The Baldus study was in some sense the highpoint of policy-relevant, empirically rigorous social science research on capital punishment. It spoke directly to issues which the Supreme Court had put at the heart of its death penalty jurisprudence, and it did so using the best social science methods and expertise. However, it did not move the Court. While the Court accepted the validity of the Baldus study, it concluded that such statistical evidence could not prove discrimination in any individual case (*McCleskey v. Kemp*, 481 US 279 1987). This rejection stunned those who believed that social science could help shape policy in the area of capital sentencing. Following *McCleskey*, the Court has moved in an increasingly conservative direction,

embracing with ever greater intensity the death penalty and rejecting what in earlier years would have seemed to be persuasive challenges rooted in social science research. The result has been that 'after more than two decades of conducting research guided by the hope of influencing the Supreme Court's death penalty decisions, the change signaled by … McCleskey has forced scholars to chart new courses. … scholars have responded with research that is more deeply critical, more theoretically informed, and more broadly concerned about the culture and politics of the death penalty' (Munger 1993, p. 6).

3.1 *The Death Penalty Process*

Some of this new research has been focused on analyzing the processing of capital cases, with special attention being given to the effort to explain how and why actors in the death penalty process behave as they do. Among the most important actors in that process are jurors in capital cases, the ordinary citizens who must decide not only questions of guilt or innocence, but whether those who are found guilty of capital murder should be sentenced to life in prison or to the death penalty. Research on jurors in capital cases has examin how they process the information that is provided to them (Haney et al. 1994, Bowers and Steiner 1999), how they understand their responsibilities in capital cases (Eisenberg and Wells 1993, Hoffman 1995, Haney and Lynch 1997), how they function in an atmosphere surrounded by the fact of violence (Sarat 1995), and how they balance their folk knowledge with the legal requirements in capital cases (Steiner et al. 1999). (See *Juries*.)

Other research recently has been focused on lawyers and the lawyering process in capital cases (Sarat 1998), with special attention being given to the use of stories as persuasive devices and to the narrative strategies deployed in capital cases (Doyle 1996, Sarat 1996). Still other work has examined capital trials as events in which law attempts to put violence into discourse and to differentiate its violence from the extralegal violence which it opposes (Sarat 1993).

3.2 *'The Culture and Politics of the Death Penalty'*

But as Munger (1993, p. 6) points out, research on capital punishment has in recent years been broadened to examine 'the culture and politics of the death penalty,' in particular its compatibility with democratic politics and culture. Scholars have increasingly been asking about the political meaning of state killing in a democracy. Does it strengthen or weaken the values on which democratic deliberation depends? (See *Legal Culture and Legal Consciousness*.)

Capital punishment, some suggest (see Connolly 1999), is the ultimate assertion of righteous indignation, of power pretending to its own infallibility. By definition it leaves no room for reversibility. In so doing it expresses either a 'we don't care' anger or an unjustified confidence in our capacity to recognize and respond to evil with wisdom and propriety (see Connolly 1999). Moreover, increasingly capital punishment has been used as a kind of constituency politics, with new capital offenses being added to existing legislation in a manner designed to convey symbolic recognition on various social groups (Simon and Spaulding 1999). Democracy, scholars argue, cannot well coexist either with such anger or with the cavalier attitude toward human life conveyed by the recent use of capital punishment for narrow political purposes. For democracy to thrive it demands a different context, one marked by a spirit of openness, of reversibility, of revision quite at odds with the confidence and commitment necessary to dispose of human life in a cold and deliberate way (see, e.g., Burt 1994).

Moreover, 'The death penalty,' as Terry Aladjem (1990, p. 36) writes, 'strains an unspoken premise of the democratic state,' that may variously be named respect for the equal moral worth or equal dignity of all persons. Democratically administered capital punishment, punishment in which citizens act in an official capacity to approve the deliberate killing of other citizens, contradicts and diminishes the respect for the worth or dignity of all persons that is the enlivening value of democratic politics. And a death penalty democratically administered, a death penalty that enlists citizens to do the work of dealing death to other citizens, implicates us all as agents of state killing.

But beyond its impact on democratic values, recent research on capital punishment has tried to examine its pedagogical effects, or its broader effects on the culture at large. Here scholars have been influenced by Garland's (1991, p. 193) argument that 'Punishment helps shape the overarching culture and contributes to the generation and regeneration of its terms.' Punishment, Garland (1991, p. 195) notes, is a set of signifying practices that 'teaches, clarifies, dramatizes and authoritatively enacts some of the most basic moral-political categories and distinctions which help shape our symbolic universe.' Punishment teaches us how to think about categories like intention, responsibility, and injury, and it models the socially appropriate ways of responding to injury done to us. And what is true of punishment in general is particularly true for the death penalty.

Punishment, in general, and the death penalty in particular, lives in culture as a set of images, as a marvelous spectacle of condemnation (see Sarat 1999a). The semiotics of punishment is all around us, not just in the architecture of the prison, or the speeches made by judges as they send someone to the penal colony, but in both 'high' and 'popular' culture iconography, in novels, television, and film. Punishment has traditionally been one of the great subjects

of cultural production, suggesting the powerful allure of the fall and of our prospects for redemption. But perhaps the word 'our' is inaccurate here since Durkheim (1933) and Mead (1917), among others, remind us that it is through practices of punishment that cultural boundaries are drawn, that solidarity is created through acts of marking difference between self and other, though disidentification as much as imagined connection (see *Criminal Justice, Ethics of: Aspects of Human Dignity*; *Punishment, Comparative Politics of*; *Punishment: Social and Legal Aspects*).

Traditionally the public execution was one of the great spectacles of power and instruction in the mysteries of responsibility and retribution (Madow 1995). Yet the privatization of execution has not ended the pedagogy of the scaffold: it has redirected it into the medium of the capital trial. Execution itself, the moment of state killing, is still today an occasion for rich symbolization, for the production of public images of evil or of an unruly freedom whose only containment is in a state-imposed death, and for fictive recreations of the scene of death in popular culture (Sarat 1999a).

The cultural politics of state killing has focused on shoring up of status distinctions and distinguishing particular ways of life from others. Thus it is not surprising today that the death penalty marks an important fault line in our contemporary culture wars (Connolly 1999). To be in favor of capital punishment is to be a defender of traditional morality against rampant permissivism; it is to be a defender of the rights of the innocent over the rights of the guilty; it is to be a defender of state power against its anarchic critics. To oppose it is to carry the burden of explaining why the state should not kill the killers, of producing a new theory of responsibility and of responsible punishment, and of humanizing inhuman deeds (Sarat 1999b).

Yet all of this may miss the deepest cultural significance of state killing. As Baudrillard suggests (1993, p. 169), in regard to capital punishment 'the thought of the right (hysterical reaction) and the thought of the left (rational humanism) are both equally removed from the symbolic configuration where crime, madness and death are modalities of exchange ….' In this symbolic configuration, research on capital punishment is essential if we are to understand not only what capital punishment does for the societies in which it continues to be used but also what it does *to* them and to democratic values.

See also: Crime and Delinquency, Prevention of; Crime and Ethnicity (Including Race); Criminal Justice, Ethics of: Aspects of Human Dignity; Criminal Law and Crime Policy; Deterrence; Deterrence: Legal Perspectives; Discrimination; Discrimination: Racial; Law and Democracy; Legal Issues: Public Opinion; Legal Process and Social Science: United States; Public Opinion: Microsociological Aspects; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Race and the Law; Race Relations in the United States, Politics of

Bibliography

Aladjem T 1990 Revenge and consent: The death penalty and lockean principles of democracy. Unpublished manuscript

Aladjem T 1992 Vengeance and democratic justice: American culture and the limits of punishment. Unpublished manuscript

Amsterdam A 1999 Selling a quick fix for Boot Hill: The myth of justice delayed in death cases. In: Sarat A (ed.) *The Killing State: Capital Punishment in Law, Politics, and Culture*. Oxford University Press, New York, pp. 148–83

Baldus D C, Woodworth G, Puloski Jr C A 1990 *Equal Justice and the Death Penalty: A Legal and Empirical Analysis*. Northeastern University Press, Boston

Baldus D et al. 1999 Arbitrariness and discrimination in the use of the death penalty: the Philadelphia story. *Cornell Law Review* **98**: 1638–1762

Baudrillard J 1993 *Symbolic Exchange and Death*. Sage, London

Berns W 1979 *For Capital Punishment: Crime and the Morality of the Death Penalty*. Basic Books, New York

Bohm R 1991 American death penalty opinion, 1936–1986: A critical examination of the Gallup Polls. In: Bohm R. *The death penalty in America: Current research*. Anderson, Cincinnati, OH, pp. 113–43

Bowers W 1993 Capital punishment and contemporary values: People's misgivings and the court's misperceptions. *Law & Society Review* **27**: 157–75

Bowers W J, Steiner B D 1999 Death by default: An empirical demonstration of false and forced choice in capital sentencing. *Texas Law Review* **77**: 605–717

Burt R 1994 Democracy, equality and the death penalty. In: Shapiro I (ed.) *The Rule of Law*. New York University Press, New York, pp. 80–90

Connolly W 1999 The will, capital punishment, and culture war. In: Sarat A (ed.) *The Killing State: Capital Punishment in Law, Politics, and Culture*. Oxford University Press, New York, pp. 187–205

Doyle J M 1996 The lawyer's art: Representation in capital cases. *Yale Journal of Law and the Humanities* **8**: 417–50

Dumm T L 1987 *Democracy and Punishment: Disciplinary Origins of the United States*. University of Wisconsin Press, Madison, WI

Durkheim E 1933 *The Division of Labor in Society*. Simpson G trans. Free Press of Glencoe, New York

Ehrlich I 1975 The deterrent effect of capital punishment. *American Economic Review* **65**: 397–417

Eisenberg T, Wells M 1993 Deadly confusion: Juror instructions in capital cases. *Cornell Law Review* **79**: 1–17

Ellsworth P C, Gross S R 1994 Hardening of the attitudes: Americans' views on the death penalty. *Journal of Social Issues* **50**: 19–52

Foucault M 1979 *Discipline and Punish: The Birth of the Prison*. Sheridan A trans. Vintage Books, New York

Fox J et al. 1990–91 Death penalty opinion in the post-*Furman* years. *New York University Review of Law and Social Change* **18**: 499–528

Garland D 1991 Punishment and culture: The symbolic dimension of criminal justice. *Studies in Law, Politics, and Society* **11**: 191–224

Gibbs J 1986 Punishment and deterrence: Theory, research, and penal policy. In: Lipson L, Wheeler S (eds.) *Law and the Social Sciences*. Russell Sage, New York, pp. 319–68
Glaser D 1977 The realities of homicide versus the assumptions of economists in assessing capital punishment. *Journal of Behavioral Economics* **6**: 243–68
Gross S 1985 Race and death: The judicial evaluation of evidence of discrimination in capital sentencing. *University of California, Davis Law Review* **18**: 1275–1325
Gross S, Mauro R 1984 Patterns of death: An analysis of racial disparities in capital sentencing and homicide victimization. *Stanford Law Review* **37**: 27–153
Haney C, Sontag L L, Costanzo S 1994 Deciding to take a life. Capital juries, sentencing instructions, and the jurisprudence of death. *Journal of Social Issues* **50**: 149–176
Haney C, Lynch M 1997 Clarifying life and death matters: An analysis of instructional comprehension and penalty phase closing arguments. *Law and Human Behavior* **21**: 575–95
Hoffman J L 1995 Where's the buck?: Juror misperception of sentencing responsibility in death-penalty cases. *Indiana Law Journal* **70**: 1137–60
Klein L et al. 1986 The deterrent effect of capital punishment: An assessment of the estimates. In: Blumstein A et al. (eds.) *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*, National Academy of Science, Washington, DC, pp. 336–60
Lempert R 1981 Desert and deterrence: An assessment of the moral bases of the case for capital punishment. *Michigan Law Review* **79**: 1177–1231
Madow M 1995 Forbidden spectacle: Executions, the public, and the press in nineteenth century New York. *Buffalo Law Review* **43**: 461–562
Mead G H 1918 The psychology of punitive justice. *American Journal of Sociology* **23**: 577–602
Munger F 1994 From the editor. *Law and Society Review* **28**: 3–5
Paternoster R 1983 Race of the victim and location of crime: The decision to seek the death penalty in South Carolina. *Criminal Law and Criminology* **74**: 754–85
Sarat A 1993 Speaking of death: Narratives of violence in capital trials. *Law & Society Review* **27**: 19–58
Sarat A 1995 Violence, representation, and responsibility in capital trials: The view from the jury. *Indiana Law Journal* **70**: 1103–35
Sarat A 1996 Narrative strategy and death penalty advocacy. *Harvard Civil Rights-Civil Liberties Law Review* **31**: 353–81
Sarat A 1997 Vengeance, victims and the identities of law. *Social and Legal Studies* **6**: 163–90
Sarat A 1998 Between (the presence of) violence and (the possibility of) justice: Lawyering against capital punishment. In: Sarat A, Scheingold S (eds.) *Cause Lawyering: Political Commitments and Professional Responsibilities*. Oxford University Press, New York, pp. 317–48
Sarat A 1999a The cultural life of capital punishment: Responsibility and representation in *Dead Man Walking* and *Last Dance*. In: Sarat A (ed.) *The Killing State: Capital Punishment in Law, Politics, and Culture*. Oxford University Press, New York, pp. 226–56
Sarat A 1999b Recapturing the spirit of *Furman*: The American Bar Association and the new abolitionist politics. *Law and Contemporary Problems* **61**: 5–24
Sarat A, Vidmar N 1976 Public opinion, the death penalty and the eight amendment: Testing the Marshall hypothesis. *Wisconsin Law Review* **1976**: 171–206
Sellin T 1959 *The Death Penalty*. American Law Institute, Philadelphia
Simon J 1997 Governing through crime. In: Friedman L, Fisher G (eds.) *The Crime Conundrum: Essays on Criminal Justice*. Westview Press, Boulder, CO, pp. 171–89
Simon J, Spaulding C 1999 Tokens of our esteem: Aggravating factors in the era of deregulated death penalties. In: Sarat A (ed.) *The Killing State: Capital Punishment in Law, Politics, and Culture*. Oxford University Press, New York, pp. 81–116
Steiner B et al. 1999 Folk knowledge as legal action: Death penalty judgments and the tenet of early release in a culture of mistrust and punitiveness. *Law & Society Review* **33**: 461–506
Tonry M 1995 *Malign Neglect: Race, Crime, and Punishment in America*. Oxford University Press, New York
Vidmar N, Ellsworth P 1974 Public opinion and the death penalty. *Stanford Law Review* **26**: 1245–70
Von den Haag E 1975 *Punishing Criminals: Concerning a Very Old and Painful Question*. Basic Books, New York
White W S 1991 *The Death Penalty in the Nineties*, University of Michigan Press, Ann Arbor, MI
Weisberg R 1983 Deregulating death. *Supreme Court Review* **1983**: 305–95
Wolfgang M, Riedel M 1973 Race, judicial discrimination and the death penalty. *Annals* **407**: 119–35
Wolfgang M, Riedel M 1975 Rape, race, and the death penalty. *American Journal of Orthopsychiatry* **45**: 658–68
Zeisel H 1976 The deterrent effects of the death penalty: facts vs. Faits. *Supreme Court Review* **1976**: 317–43
Zeisel H 1981 Race bias in the administration of the death penalty: The Florida experience. *Harvard Law Review* **95**: 456–68
Zimring F 1993 On the liberating virtues of irrelevance. *Law and Society Review* **27**: 9–18
Zimring F 1999 The executioner's dissonant song: On capital punishment and American values. In: Sarat A (ed.) *The Killing State: Capital Punishment in Law, Politics, and Culture*. Oxford University Press, New York, pp. 137–47
Zimring F, Hawkins G 1986 *Capital Punishment and the American Agenda*. Cambridge University Press, Cambridge, UK

A. Sarat

Deception, Psychology of

1. Introduction

This article describes how, despite of being an everyday life event, people are generally poor at detecting lies. Several reasons for this inability to detect deceit are given and techniques that might improve people's lie detection ability are discussed.

2. Definition of Deception

Deception is 'a successful or unsuccessful deliberate attempt, without forewarning, to create in another a belief which the communicator considers to be untrue' (Vrij 2000). Lying is an intentional act. The schizophrenic man who believes he is Napoleon is not lying

in his claims. He believes his own story and so does not have the intention of deceiving other people. Moreover, people are only lying when they do not inform others in advance about their intentions to lie. Magicians are therefore not lying, as people in the audience expect to be deceived.

3. *Lying is a Daily Life Event*

People tell on average 1.5 lies a day (DePaulo et al. 1996). The frequency of lying depends on the personality of the liar, the person to whom the lie is told and the situation. Extraverted people and people high in Machiavellianism lie more often than introverted people do and people low in Machiavellianism (Kashy and DePaulo 1996).

People tell fewer lies to people whom they feel emotionally closer to (Anderson et al. 1999). However, the lies they tell are relatively serious. People probably would like to be honest to their romantic partners, but believe that this is not always possible, especially when the truth might harm the relationship. Therefore, the adulterous husband will not inform his wife about his mistress, and the wife addicted to shopping will hide several purchases.

How often people lie also depends on the situation: 83 percent of students would lie to get a job (Robinson et al. 1998) and 90 percent of students are willing to lie on their first date in order to make a favorable impression (Rowatt et al. 1998).

4. *The Human Lie Detector*

In the scientific lie detection studies, observers are typically given videotapes or audiotapes and asked to judge whether each of a number of people is lying. The alternatives to choose from are 'yes' or 'no,' resulting in a 50 percent chance of being correct. A review of 39 studies (Vrij 2000) revealed an average accuracy rate of 56.6 percent. Males are no better (or worse) than females and professional lie catchers (police officers and customs officers) are generally no better than lay persons in detecting deceit (Vrij 2000). Even studies investigating the ability to detect lies in children rarely exceeds an accuracy rate of 60 percent (Vrij in press).

5. *Why are People Poor Lie Detectors*

There are numerous reasons why people are poor at detecting lies and some of these reasons are discussed.

First, conversation rules prevent lie detectors from analyzing an accused liar properly. Conversation rules prescribe that conversation partners look each other into the eyes. Eye movements, however, do not give reliable information about deception (Vrij 2000).

Second, differences between liars and truth tellers are usually very small (Vrij 2000).

Third, there is no behavior or set of behaviors that all liars exhibit (Vrij 2000). Obviously, the fact that generic deceptive behavior does not exist makes it difficult for observers to decide what to look for.

Fourth, observers often fail to take individual differences in behavior into account (Ekman 1992). Naturally, some people make many movements, whereas others exhibit a lot of gaze aversion, and so on. This hampers lie detection. One way of detecting deceit is comparing someone's behavior with his or her baseline truthful behavior. People are more accurate at detecting lies when they are aware of someone's baseline truthful behavior (Feeley et al. 1995).

Finally, people often look at the wrong cues to detect deceit. Although there is nothing like Pinocchio's nose, some behaviors are more likely to occur than others (Vrij 2000). It is perhaps remarkable how little observers know about deceptive behavior. For example, although observers often believe that liars increase their movements, research has shown that liars often decrease their movements (Vrij 2000). Liars tend to control their behavior, both in order to avoid giving possible nonverbal indicators of their deception and to enhance the credibility of the impression they make on others. Paradoxically, deceivers' very attempts to control their behavior serve as cues to deception. The controlled behavior will appear as planned, rehearsed, and lacking in spontaneity (DePaulo and Kirkendol 1989). Also, it is sometimes more difficult to fabricate a plausible and convincing lie that is consistent with everything the observer knows or might find out than it is to tell the truth. Evidence has demonstrated that people engaging in cognitively complex tasks make fewer hand and arm movements (Ekman and Friesen 1972), possibly because a greater cognitive load results in a neglect of body language, reducing overall animation.

6. *Reasons for Misconceptions about Deceptive Behavior*

There are several reasons why observers have misconceptions about deceptive behavior (Vrij 2000). It might be a matter of experience. Almost everybody can think of somebody at some time who acted nervously during deception, and who was subsequently caught out because of this nervousness. Lie catchers may then wrongly assume that all liars behave this way.

People often lack feedback to find out how good they are at detecting lies (DePaulo and Pfeifer 1986). For example, although customs officers gain feedback from those travelers that they stopped and searched, they will almost never find out whether the travelers they did not search were smuggling goods, and they therefore cannot learn from these cases.

Also, people do not know how they behave when they are lying (Vrij et al. 1996, 2001). In the experiment of Vrij et al. (1996), liars actually showed fewer movements while lying, although they thought that they had made more movements during deception. Hence, while detecting lies in others, people are looking for cues, which they think (incorrectly), reveal their own lies.

Several lie detection techniques have been introduced to tackle some of the problems raised above, including the 'baseline technique' and the 'indirect lie detection technique.' They address the issues of idiosyncrasy in people's natural truthful behavior and the tendency to look at the wrong cues, respectively.

6.1 Baseline Method

Knowing someone's baseline truthful behavior would facilitate lie detection, as comparisons can be made with the behavior under investigation. During a videotaped real-life police interview a man was asked to describe his activities during a particular day (Vrij and Mann 2001). The murder suspect gave descriptions of his activities during the morning, afternoon and evening. Detailed analyses of the videotape, however, revealed a sudden change in behavior as soon as he started to describe his activities during the afternoon and evening. He spoke slower, added more pauses, and made fewer movements. It made the impression that he had to think hard. One possible reason for this may have been that he was lying. Evidence supported this view. Police investigations could confirm his story about his morning activities, but revealed that his statement about the afternoon and evening were fabricated. In reality, he met the victim and killed her later on that day. Apparently, lying about the afternoon and evening required mental effort which was revealed by his change in behavior.

Crucial in the use of the baseline technique is that the correct parts of the interview are compared. One should not compare apples with pears. Unfortunately, that happens often in police interviews (Moston and Engelberg 1993). Small talk at the beginning of the interview is used to establish a baseline, which is then compared with the behavior shown in the actual interview. This is an incorrect way of employing the technique as small talk and the actual police interviews are totally different situations. Not surprisingly, both guilty and innocent people tend to change their behavior the moment the actual interview starts (Vrij 2000). In the case of the murderer, we were able to make a good comparison. There are no reasons why different behaviors would emerge while describing the morning or the afternoon and evening. Interestingly, the question on which we based the baseline method 'What did you do that particular day?' could be asked in almost every police interview.

6.2 Indirect Lie Detection Method

People know more about deception than it appears when they are asked directly whether they think someone is lying (DePaulo 1994). In some studies, after watching a truthful or deceptive story, participants were asked to detect deception both in a direct way (i.e. 'is the person lying?') and in an indirect way (i.e. 'Does the speaker sincerely like the person (s)he just described?'). These studies found greater accuracy on the indirect measures (see Vrij in press). It might be the result of conversation rules, which regulates politeness. Observers are often unsure as to whether someone is lying to them. In these cases it will be impolite or for other reasons undesirable to accuse someone of being a liar (e.g. 'I do not believe you'), but it might be possible to challenge the words of a speaker more subtly (e.g. 'Do you really like that person so much?'). Alternatively, people might look at different cues when detecting lies than when applying an indirect method. In a study by Vrij et al. (in press), police officers could distinguish between truths and lies, only by using an indirect method (by judging whether truth tellers and liars had to think hard instead of by judging whether they were lying). Moreover, only in the indirect method they paid attention to the cues that actually discriminated between the truth tellers and liars, such as a decrease in hand movements.

7. Final Conclusion and Future Directions

People are generally poor at detecting lies. However, recently it has been discovered that improvements could be made by using the baseline technique or the indirect lie detection technique. Both techniques need to be assessed further, both inside the laboratory and in real life.

See also: Animal Cognition; Behavioral Genetics: Psychological Perspectives; Comparative Psychology; Confessions: Psychological and Forensic Aspects; Deceptive Methods: Ethical Aspects; Facial Expressions; Forensic Psychology; Primate Behavior: Significance for Understanding Humans; Psychological Development: Ethological and Evolutionary Approaches; Sensitive Periods in Development, Neural Basis of; Sociality: Anthropological Aspects

Bibliography

Anderson D E, Ansfield M E, DePaulo B M 1999 Love's best habit: Deception in the context of relationships. In: Philippot P, Feldman R S, Coats E J (eds.) *The Social Context of Nonverbal Behavior*. Cambridge University Press, Cambridge, UK, pp. 372–409

DePaulo B M 1994 Spotting lies: Can humans learn to do better? *Current Directions in Psychological Science* **3**: 83–6
DePaulo B M, Kirkendol S E, Kashy D A, Wyer M M, Epstein J A 1996 Lying in everyday life. *Journal of Personality and Social Psychology* **70**: 979–95
DePaulo B M, Kirkendol S E 1989 The motivational impairment effect in the communication of deception. In: Yuille J C (ed.) *Credibility Assessment*. Kluwer, Dordrecht, The Netherlands, pp. 51–70
DePaulo B M, Pfeifer R L 1986 On-the-job experience and skill at detecting deception. *Journal of Applied Social Psychology* **16**: 249–67
Ekman P 1992 *Telling Lies: Clues to Deceit in the Marketplace, Politics and Marriage*. W. W. Norton, New York
Ekman P, Friesen W V 1972 Hand movements. *Journal of Communication* **22**: 353–74
Feeley T H, deTurck M A, Young M J 1995 Baseline familiarity in lie detection. *Communication Research Reports* **12**: 160–9
Kashy D A, DePaulo B M 1996 Who lies? *Journal of Personality and Social Psychology* **70**: 1037–51
Moston S, Engelberg T 1993 Police questioning techniques in tape-recorded interviews with criminal suspects. *Policing and Society* **3**: 223–37
Robinson W P, Shepherd A, Heywood J 1998 Truth, equivocation/concealment, and lies in job applications and doctor–patient communication. *Journal of Language and Social Psychology* **17**: 149–64
Rowatt W C, Cunningham M R, Druen P B 1998 Deception to get a date. *Personality and Social Psychology Bulletin* **24**: 1228–42
Vrij A 2000 *Detecting Lies and Deceit: Psychology of Lying and the Implications for Professional Practice*. John Wiley and Sons, Chichester, UK
Vrij A in press Deception in children. A literature review and implications for children's testimony. In: Westcott H, Davies G, Bull R (eds.) *Children's Testimony*. Wiley, Chichester, UK
Vrij A, Edward K, Bull R 2001 People's insight into their own behavior and speech content while lying. *British Journal of Psychology* **92**: 373–89
Vrij A, Edward K, Bull R in press Police officers' ability to detect deceit: Differences between direct and indirect detection measures. *Legal and Criminological Psychology*.
Vrij A, Mann S 2001 Lying when the stakes are high. Deceptive behavior of a murderer during his police interview. *Applied Cognitive Psychology* **15**: 187–203
Vrij A, Semin G R, Bull R 1996 Insight in behavior displayed during deception. *Human Communication Research* **22**: 544–62

A. Vrij

Deceptive Methods: Ethical Aspects

Deception refers to the study of the reactions of subjects who are purposely led to hold false beliefs. Deception was widely accepted prior to 1970, especially in social psychology. The subsequent rise of research ethics committees and federal regulations caused researchers to rethink their methods. However, some behavior (e.g., aggression, conformity, helping) is difficult or impossible to study without deception, hence deception remains permissible and widely used.

1. *Justifiable Uses of Deception*

Deception may be justifiable in order to: (a) achieve stimulus control or random assignment of subjects; (b) study response to low-frequency events (e.g., finding money); (c) avoid risk to subjects (e.g., by using a confederate to study aggression, who will not escalate hostilities beyond the level needed for the research); or (d) obtain information otherwise unobtainable due to subjects' defensiveness, embarrassment, or fear of reprisal. Milgram's (1974) study of obedience to unjust authority illustrates achievement of these objectives: Milgram sought to study the conditions under which apparently decent people obey orders to harm others. He used newspaper ads to invite the public to participate in a 'learning experiment'—in reality a study of obedience. Volunteers were designated as teacher, and instructed to teach lists of words to learners (confederates) and to give 'learners' increasingly severe electric shocks at every mistake. The 'learners' were wired to a bogus device that looked like a shock generator; they feigned failure to learn and pain at being shocked. If the experimenter gave instructions and left, the 'teacher' did not escalate shock beyond a tingle. But if a high status experimenter remained and insisted that orders be followed, 66 percent of subjects escalated 'shock' to lethal levels. (A panel of psychiatrists had predicted that less than 1 percent would obey!) Subjects were visibly shaken, and relieved to be debriefed.

Although Milgram met all four of the above criteria justifying deception, he was criticized severely. He had employed many forms of unacceptable deception.

2. *Aspects of Deception and Ethical Issues Raised*

Deception can take many forms, each with ethical implications.

2.1 *Failure to Inform*

Of the five kinds of failure to inform, two are highly objectionable. No informing typically means that subjects do not know they are participating in a study. False informing means lying or omitting vital details. Both approaches deny self-determination and invade privacy; the researcher extracts information that subjects might not wish to reveal.

The other three approaches somewhat respect subjects' autonomy and privacy. Informed consent to participate in one of several specified conditions

without knowing which condition will actually occur, means subjects can decide whether all of the possible conditions are acceptable hence whether to participate. Consent to deception means subjects know deception may occur. Waiver of informing involves no explicit warning about possible deception. Each of these approaches involves telling subjects what they will experience, and providing debriefing afterward.

2.2 Who Deceives?

The researcher is not necessarily the only deceiver. Third party deception may occur when social interaction is studied, as in research on therapeutic privilege where physicians are observed lying to patients when the truth is deemed harmful. Self-deception occurs when subjects fail to realize that the researcher views them from a different, critical, or unsympathetic perspective.

No informing and no consent, false informing, self-deception, and third-party deception, especially in field experiments, violate privacy, self-determination, and interpersonal trust. Worse, some deception is so convincing that subjects may reject debriefing. Alternatively, if deceived once, why should subjects believe the next thing they are told? However, in laboratory experiments employing such contrivances as bogus equipment and confederates, it usually is easy to demonstrate the deception and hence to debrief convincingly.

2.3 Devices of Deception

Three major devices of deception are: technical deception in which objects, equipment, or procedures are misrepresented; role deception or use of confederates; and implicit deception in which the research aim is so unlikely that subjects generate their own misleading assumptions. Such devices, *per se*, are harmless; they are commonly used in magic tricks and other forms of entertainment. But when combined with certain other features of deception research they may create distrust and cynicism.

2.4 The Research Topic

The most powerful determiner of unacceptability of deception is the degree of immorality and privacy of the behavior studied, the strength of induction, and the degree of upset caused by the combination of these factors.

(a) If the behavior studied is regarded as positive or neutral, deception is less objectionable than if the behavior is regarded as negative or immoral.

(b) The study of public behavior is more acceptable than the study of private or covert behavior.

(c) The study of naturally occurring or mildly induced behavior is more acceptable than use of powerful induction.

(d) To the extent that the research causes upset, it is likely to be considered objectionable.

Upset to subjects may arise from what they thought happened, or from debriefing that leaves them feeling foolish, or vulnerable. Public upset may arise from concern that science condones such procedures. Investigators, assistants, and confederates may be upset about their involvement.

To illustrate the importance of these factors, contrast Milgram's study of obedience with Isen and Levin's (1972) study of kindness. Their study involved no informing, no consent, no debriefing, role deception, and implicit deception. Yet generally it is regarded as a positive intervention, not as deception.

Isen and Levin (1972) studied the effects of positive versus neutral moods on by-stander intervention. Students studying in the library were selected randomly to be given a cookie or not, then given an opportunity to help another student (confederate) who 'accidentally' dropped an armful of books. Isen and Levin showed that people who feel good because of a lucky occurrence or the kindness of some passerby will return the kindness (to someone else) at the next opportunity, while control group members are more likely to ignore the minor mishap of another.

3. Indefensible Uses of Deception

An indefensible use of deception is to trick people into participation that they would find unacceptable if they correctly understood pertinent details of the research. Each culture or population has its own beliefs, values, and sensitivities; these must be respected. When full disclosure to subjects would invalidate the research, the investigator must find other ways to evaluate acceptability to the research population. For example, the researcher might ask surrogate subjects (i.e., peers of the subjects), to consider all relevant details of the research including its sponsorship and likely applications, identify any harmful, worrisome, or objectionable elements, and decide whether participation would be acceptable to them.

A flawed rationale for deception in laboratory settings is to promote spontaneous behavior. Laboratory subjects, especially college students, typically assume that deception is employed and engage in deceptive behavior of their own (Geller 1982). Subjects are respectful of consent procedures that request permission to withhold some information until after participation.

When the value of the research does not justify the risks and wrongs of deception (i.e., when valid knowledge can be obtained otherwise, the research design is flawed, or the research is trivial) deception is indefensible.

4. Debriefing: Dehoaxing and Desensitizing

Debriefing refers to the removal of the deception after research participation; it involves two processes: *dehoaxing* refers to removing any misunderstanding and demonstrating how the deception occurred. If technical deception was involved, an examination of the bogus device is in order. Role deception is dehoaxed by introducing subjects to the real person. Implicit deception is more difficult to dehoax since subjects essentially misinformed themselves; such subjects should be assured that anyone would misinterpret the situation and that the research was designed to produce misinterpretation. The explanation should enable subjects to feel that the research was reasonable, and that appropriate steps were taken to ensure subjects' privacy and safety. Subjects should be given the opportunity to ask questions and discuss the study, and to withdraw their data from the research, especially if private behavior was involved.

Desensitizing means detecting undesirable emotional consequences of the research and restoring subjects to a frame of mind at least as positive as before participation. For example, Milgram's subjects were induced to reveal information about themselves that upset them; it inflicted unflattering self-insight. Milgram desensitized his subjects by explaining that their behavior had been powerfully induced by a demanding supervisor, at a prestigious institution, etc., and that others in the same circumstances behaved similarly. Subjects were offered free psychotherapy sessions in which to examine their concerns and restore their self-assurance and sense of dignity. Since desensitizing efforts do not always succeed (Holmes 1976) studies should not damage self-esteem in the first place.

When studying persons whose typical behavior generally is regarded as negative, desensitizing is unnecessary and dehoaxing harmful. For example, young children typically take things they want. Researchers studying stealing in young children, e.g., by making coins available for stealing, should not tell the child the study was about stealing. The researcher's goal is not to shame, or to accelerate the child's moral development, but only to study how children behave. The researcher must also be mindful of the feedback given to parents. If parents are informed of their child's response, it must be in the context of information about typical behavior of children at that age, and about appropriate parental responses to stealing by children of that age.

5. Alternatives to Deception

This article began with four research objectives that might best be achieved using deception. These same objectives also sometimes can be achieved via simulation (in which subjects act as if the situation were real), or via ethnographic methods.

5.1 Simulations

In game simulations, subjects take roles under a particular set of rules. For example, mock trials have been convened using real jury candidates, real judges, and real cases, and they are highly realistic. Field simulations lack firm rules, use highly realistic staged settings, and encourage subjects to believe they are participating in a natural event. These can be so realistic that in simulation involving harmful behavior subjects may actually suffer harm. Role playing is, perhaps, the most practical simulation substitute for deception. The role player knows what illusions are created but is asked to act as if the situation were real. Geller (1982), using Milgram's equipment, had subjects role play the various conditions of Milgram's obedience experiment. He replicated Milgram's results!

5.2 Ethnographic or Participant Observation Methods

These methods, borrowed from cultural anthropology, coupled with self-report, are now often used to study sensitive behavior of people in their communities. The key to valid ethnography is not technical cleverness, but trust, rapport, knowledge, empathy with the culture studied, and good ethnographic interviewing skills.

Despite recent developments, heightened ethical sensitivities, and government regulation, deception continues to be used. Since 1969, lying to subjects first declined then increased. Human confederates have been partly replaced by computers. Little else has changed (Sieber et al. 1995). Whether use of deception will further tarnish the reputation of science probably depends on whether scientists attend creatively to respecting autonomy, providing appropriate debriefing, and selecting the most valid and least objectionable deception methods (see *Mentoring: Supervision of Young Researchers*; *Privacy of Individuals in Social Research: Confidentiality*; *Research Ethics: Research*).

See also: Deception, Psychology of; Ethical Behavior, Evolution of; Ethics Committees in Science: European Perspectives; Participant Observation; Placebos, Use of: Ethical Aspects; Privacy: Legal Aspects; Psychiatry: Informed Consent and Care; Research Conduct: Ethical Codes; Research Subjects, Informed and Implied Consent of

Bibliography

Geller D 1982 Alternatives to deception: why, what, and how? In: Sieber J (ed.) *The Ethics of Social Research: Surveys and Experiments*. Springer-Verlag, New York

Holmes D 1976 Debriefing after psychological experiments: effectiveness of post experimental desensitizing. *American Psychologist* **32**: 868–75
Isen A, Levin P 1972 Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology* **21**: 384–8
Milgram S 1974 *Obedience to Authority*. Harper & Row, New York
Sieber J 1982 Deception in social research. I: Kinds of deception and the wrongs they may involve. *IRB: A Review of Human Subjects Research* **1–2**: 12
Sieber J 1983 Deception in social research. II: Factors influencing the magnitude of potential for harm or wrong. *IRB: A Review of Human Subjects Research* **1–3**: 12
Sieber J, Iannuzzo R, Rodriguez B 1995 Deception methods in psychology: have they changed in 23 years? *Ethics and Behavior* **5**: 67–85

J. E. Sieber

Decision and Choice: Economic Psychology

1. Introduction

Economic psychology is best defined by what economic psychologists do. Typically, they are interdisciplinary social scientists who recognize that: (a) the economy powerfully influences individuals' lives, at the psychological level of feelings, thoughts, and behavior; (b) individuals' feelings, thoughts, and behavior are what make up economic life; and (c) there are many problems, both academic and practical, to which economics and psychology can both contribute despite their very different approaches and explanatory styles.

2. Relation of Economic Psychology to Other Research Areas

Thus economic psychology is not a branch of psychology, but the interdisciplinary study of the interface—or sometimes the gap—between economics and psychology. One of the most powerful and successful ways of looking at that interface has been through the study of economic decisions (see *Decision Research: Behavioral*), particularly in the areas of risky choice (see *Risk: Empirical Studies on Decision and Choice*), limitations on rationality (see *Bounded and Costly Rationality*), and intertemporal choice. So an article on economic psychology is fittingly placed in this section, alongside discussions of numerous other aspects of decision theory. But it is equally strongly linked to articles in this Encyclopedia devoted to economics (for example, to developmental, social, personality, and motivational psychology (for example, *Leisure, Psychology of* or *Poverty and Child Development*), to organizational and management studies (for example, *Customer Satisfaction* or *Marketing Strategies*), and indeed to others.

A few very distinguished researchers have moved freely between economics and psychology, and are recognized and respected in both disciplines. But many of them would be more likely to identify themselves as decision theorists than as economic psychologists: examples include Herbert Simon, Amos Tversky (see *Tversky, Amos (1937–96)*), Richard Thaler, and George Loewenstein. It is true in general that many more people do economic psychology than call themselves economic psychologists. Very similar approaches are taken not only in decision science, but also under other interdisciplinary names, such as behavioral economics, consumer science, or socioeconomics, and within particular branches of both psychology and economics (see *Behavioral Economics*; *Consumer Economics*; *Experimental Economics*). Because so many of these areas of research are represented elsewhere in this Encyclopedia, the remainder of this article will largely discuss the work of those who have explicitly thought of themselves as economic psychologists.

The first use of the phrase 'economic psychology' is thought to be in the work of the influential French social psychologist Gabriel Tarde (1902). But the modern development of economic psychology started from the work of the Hungarian-American Gestalt psychologist George Katona, summarized in his *Psychological Economics* (1975). Recent surveys of current research include Earl and Kemp (1999) and Webley et al. (2001).

3. Economic Behaviors of Individuals

Economic behaviors include, for example, work, buying, saving, giving, and gambling. Several of these have been subjects of intense research within psychology. Work, for example, is the subject of an entire branch of psychology, variously known as occupational psychology, industrial psychology, or work psychology, and a correspondingly large part of economics, labor economics. It is no accident, therefore, that many economic psychologists have concentrated on economic behaviors such as saving, giving, and paying (or avoiding) tax, which are left aside by the established subdisciplines. When economic psychologists look at topics such as work and buying, it is usually with the aim of integrating them with other aspects of economic life, rather than seeing them in isolation.

The study of saving and debt has been one of the most popular research areas for economic psychologists, for a number of reasons. First, it is macroeconomically important. Katona's consumer surveys focused on how much of their incomes consumers intended to spend rather than save, since quite small

changes in this 'propensity to consume' can tilt a modern economy into recession or inflation. Second, the problem of optimal intertemporal choice is theoretically interesting, and difficult: economists have often had to appeal to (usually speculative) psychological principles to reconcile their theories with some obvious macroeconomic facts. Third, it is important in the lives of individuals: making adequate provision for future life changes such as children's education or one's own retirement is a difficult and complex matter, while getting involved in serious debt or bankruptcy can be a personal disaster. Finally, the problems have proved tractable to psychological investigation. Wärneryd (1999) has summarized extensive research that has given us a thorough understanding of the psychological processes leading some people to save more than others. The psychology of debt has only come under investigation more recently, but it is already clear that, though debt is formally just the opposite of saving, the psychological processes involved are quite different.

The psychology of giving has attracted a smaller research literature, but a great deal of popular attention. Gifts illustrate in a vivid way a general principle: every gift is both an economic and a social transaction. Within any society there are well-understood, though usually unstated, social rules about what may be given to whom and on what occasions. For example, in the UK money is not an acceptable birthday gift from a young adult to his or her mother. Interestingly, the objections are felt more strongly by the potential donor than by the potential recipient.

Almost all economic behaviors involve money in some way. A central problem for economic psychologists has therefore been the psychology of money itself. The problem is the more acute because modern 'fiducial' money is inherently a psychological creation: it is valuable only because we all believe it is. Economic psychologists have investigated what makes 'good' money, from the kinds of coins that look and feel valuable to people, to the kinds of electronic payment that do not leave people feeling they are out of control of their spending. They have also investigated some of the unexpected behaviors people show towards money, for it seems to be a general rule that money attracts a certain amount of behavior that is disproportionate to its actual economic usefulness. Furnham and Argyle (1998) have collected much of the recent literature on the psychology of money.

4. *The Impact of the Economy on Individuals' Psychology*

At any moment, within any given economy, different individuals occupy different positions, from millionaire to beggar; even the most stable economy undergoes unpredictable macroeconomic changes, from growth to recession; some economies undergo massive transformations, as for example the 1990s transition from state socialism to market capitalism in Eastern and Central Europe and the former Soviet Union; within economies, individual industries and industrial sectors wax and wane; and across the globe there are huge variations between countries in people's average earnings, spending power, and standards of living. These many different dimensions of economic variation must have profound psychological consequences, yet the economy is traditionally ignored as an independent variable within psychology. Economic psychologists have sought to repair this omission.

A popular methodology has been to look at how children learn to function within the particular economy they are born into. This is known as the study of 'economic socialization' (Lunt and Furnham 1996). Many earlier studies took a cognitive stance, investigating how children's understanding of economic concepts improves with age. More recent work has tended to look at children's actual economic behavior, and has led us to realize that, as well as interacting directly with the adult economy, children have at least two other kinds of economic life. They are potent influences on the economic behavior of their entire households, both through their unarguable needs, and through their ability to influence their parents' consumer and occupational choices. In consequence, children are the target of major advertising and marketing campaigns, and the legitimacy and effectiveness of these has been keenly debated. Second, however, children live in an entire economic world of their own, involving exchange transactions that are at best tangentially related to the adult economy. They may show much more sophisticated economic understanding in these transactions than they reveal in the traditional studies of the development of economic understanding.

Another widespread approach to considering the economy as an independent variable has been comparative study, either cross-sectional or (more rarely) longitudinal. Cross-sectional studies include a series of impressive collaborative cross-national investigations, for example, of economic socialization (Leiser et al. 1990), of tax evasion (Webley et al. 1991), and of attitudes to the new, transnational Euro currency (Pepermans et al. 1998). The many institutional differences between nations mean that such comparisons can give a good insight into the effective economic variables producing differences between people.

Where such formal comparisons are not feasible, 'impact studies' have been popular: people in a particular economic position are interviewed or given psychometric tests to try to understand how they differ from more typical populations. Such studies have, for example, looked at the impact of the tourist economy on its host populations (Ap 1992), or at the kinds of coping behavior elicited by living in poverty within a wealthy society (Kempson et al. 1994).

5. The Vexed Question of Rational Self-interest

Economics as a discipline is dominated by a single theoretical idea: rational self-interest. Much of economics is theoretical, and much of its theory consists of working out what a rationally self-interested individual would do in a given situation. Psychology, by contrast, has no dominant theory, and its dominant research paradigm is to submit theoretical assertions to empirical test.

Unsurprisingly, two such different disciplines have made uncomfortable bedfellows. Furthermore, the empirical bias of psychologists has led them, as soon as they start to look at economic behavior, to question the economic assumption of rational self-interest. This means that a great deal of research in economic psychology has consisted of investigations by psychologists of the empirical accuracy of theories developed by economists. A smaller amount has consisted of the incorporation into economic theory of psychological principles.

Economic rationality is a protean concept; some have argued, indeed, that it is not a theory at all, more of a general language within which specific theories can be expressed. Economists are not necessarily impressed when psychologists find that, in artificial experiments, some percentage of individuals show a percentage variation from the truly rational behavior pattern: they protest that they are interested in predicting the broad trends of behavior in the great mass of people, so as to explain the performance of the economy as a whole—and they argue that many empirical deviations from rationality are second-order effects of little or no macroeconomic consequence.

Recognizing the various difficulties associated with a frontal assault on rationality assumptions, modern economic psychologists mostly do not test rationality assumptions directly, but rather try to build empirically valid models of the causation and consequences of economic behaviors. Popular causal variables include personality differences, attitudes, socialization experiences, and psychological disorders. All these variables can potentially cause different people, exposed to apparently identical economic situations, to react in different ways. They thus allow psychologists to account for some of the individual variation in behavior that remains when the obvious economic variables have been taken into account.

Cumulatively, of course, such studies can still constitute a powerful attack on the economic theory of rational self-interest. There are many ways of acting rationally in any given situation, depending on your knowledge, abilities, values, and goals. If most of the variation between people's economic behavior actually depends on which rational behavior people choose, and if this has to be predicted from psychological properties of the person, the role of economic psychology has to increase, while the role of rational theory as such has to decrease. How far that process will go remains uncertain; what unites economic psychologists is a belief, based on the fruitfulness of research so far, that it has to go further than it yet has.

See also: Bayesian Theory: History of Applications; Bounded and Costly Rationality; Decision Research: Behavioral; Heuristics for Decision and Choice; Psychology and Economics; Risk: Empirical Studies on Decision and Choice; Utility and Subjective Probability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies

Bibliography

Ap J 1992 Residents' perceptions on tourism impacts. *Annals of Tourism Research* **19**: 665–90

Earl P E, Kemp S 1999 *The Elgar Companion to Consumer Research and Economic Psychology*. Elgar, Cheltenham, UK

Furnham A, Argyle M 1998 *The Psychology of Money*. Routledge, London

Katona G 1975 *Psychological Economics*. Elsevier, Amsterdam

Kempson E, Bryson A, Rowlingson K 1994 *Hard Times? How Poor Families Make Ends Meet*. Policy Studies Institute, London

Leiser D, Roland-Lévy C, Sevón G 1990 Introduction. *Journal of Economic Psychology* **11**: 467–8

Lunt P, Furnham A 1996 *Economic Socialization: The Economic Beliefs and Behaviours of Young People*. Elgar, Cheltenham, UK

Pepermans R, Burgoyne C B, Mueller-Peters A 1998 Editorial: European integration, psychology and the Euro. *Journal of Economic Psychology* **19**: 657–61

Tarde G 1902 *La Psychologie Économique*. Alcan, Paris

Wärneryd K-E 1999 *The Psychology of Saving: A Study on Economic Psychology*. Elgar, Cheltenham, UK

Webley P, Burgoyne C B, Lea S E G, Young B M 2001 *The Economic Psychology of Everyday Life*. Psychology Press, Hove, UK

Webley P, Robben H S J, Elffers H, Hessing D J 1991 *Tax Evasion: An Experimental Approach*. Cambridge University Press, Cambridge, UK

S. E. G. Lea

Decision and Choice: Paradoxes of Choice

Choice paradoxes are cases where people make decisions that systematically violate the implications of a theory that is considered (at the time of the 'paradox') a viable representation of human decision-making. In the days of the classic paradoxes, a rational person's choices were thought to be normative and so paradoxes of choice also influenced theoreticians' ideas of what is rational. Results considered paradoxical from the viewpoint of one theory are explained by newer theories proposed to provide more accurate representations of decision-making. Modern paradoxes are violations of clearly stated premise or implications of a descriptive model.

1. The Classic Paradoxes

Before 1738, it was considered that a rational person should choose between gambles according to their expected values. The expected value (EV) of a gamble with n mutually exclusive and exhaustive possible outcomes can be written:

$$EV = \sum_{i=1}^{n} p_i x_i \tag{1}$$

where p_i and x_i are the probability and prize for outcome i.

1.1 St Petersburg Paradox Refutes EV

The St Petersburg paradox presents a case where scholars insisted that they would and should prefer a small amount of cash to a gamble with much higher EV. Consider the following gamble: a fair coin will be tossed and if the outcome is heads, the prize is \$2 and the game ends; however, if it is tails, the coin will be tossed again. If the second toss is then heads, the prize is \$4; if tails, the coin is tossed again, and the prize for heads doubles for each successive tails that occurs before heads. When heads occurs, the prize is awarded and the game ends. The EV of this St Petersburg gamble is infinite,

$$EV = \sum_{i=1}^{\infty} p_i x_i = \tfrac{1}{2}2 + \tfrac{1}{4}4 + \tfrac{1}{8}8 + \cdots = \infty$$

Therefore, if one prefers higher EV, then one should prefer this gamble to any finite amount of cash. Instead, most people say that they would prefer a small sum (e.g., \$15) rather than play this gamble once.

1.2 EU Explains St Petersburg Paradox

This paradox, or contradiction between EV and human judgment, was explained by Bernoulli (1738), who showed that if utility is a nonlinear function of wealth, then the expected utility (EU) of the gamble might indeed be less than the utility of a finite sum of cash. In EU theory:

$$EU = \sum_{i=1}^{n} p_i u(x_i) \tag{2}$$

Bernoulli went on to show that EU theory could explain why a pauper might be willing to sell a gamble to a wealthy person for less than its EV and why both would consider the exchange rational. He also explained the purchase and sales of insurance.

When von Neumann and Morgenstern (1947) developed axiomatic foundations of EU and when Savage (1954) developed subjective expected utility theory (SEU), generalizing EU to cases of uncertainty where objective probabilities are unspecified, violations of EV no longer seemed paradoxical. Risk-averse behavior (preferring sure cash over a gamble with the same or higher EV) and risk-seeking behavior (preferring the gamble over the EV in cash) could be explained by the shapes of the utility functions for different cases. It was not long, however, before new paradoxes were discovered that confounded EU and SEU.

1.3 The Paradoxes of Allais Violate EU

Allais suggested two-choice problems in which no utility function could be constructed that would explain choices (Allais and Hagen, 1979). The constant consequence paradox is illustrated with the following two choices:

A: \$1 million for sure

B: 0.01 probability to win \$0
0.89 probability to win \$1 million
0.10 probability to win \$5 million

A': 0.89 probability to win \$0
0.11 probability to win \$1 million

B': 0.90 probability to win \$0
0.10 probability to win \$5 million

Choices A versus B and A' versus B' differ only by changing the common consequence on a .89 chance to win \$1 million in both sides of the first choice to \$0 in the second choice. Many people preferred A to B and also preferred B' to A', contrary to EU.

Allais also proposed a constant ratio paradox, which can be illustrated as follows:

C: \$3,000 for sure

D: 0.20 probability to win \$0
0.80 probability to win \$4,000

C': 0.75 probability to win \$0
0.25 probability to win \$3,000

D': 0.80 probability to win \$0
0.20 probability to win \$4,000

Note that the probability to win is one-fourth in C' versus D', compared to C versus D. Many people persisted in choosing C over D and D' over C', contradicting EU theory.

1.4 SEU and the Ellsberg Paradoxes

Consider an option with n mutually exclusive and exhaustive events, where each event, E_i, leads to consequence C_i with subjective probability $S(E_i)$. Savage's (1954) SEU replaces probability, p_i, in Equation 2 with subjective probability, $S(E_i)$.

Although SEU has two subjective functions and may therefore seem hard to test, Ellsberg (1961)

Table 1
Choices illustrating Ellsberg paradox

Option	Payoffs for drawing a ball of each color		
	Red	*Blue*	*Green*
F	\$100	\$0	\$0
G	\$0	\$100	\$0
F′	\$100	\$0	\$100
G′	\$0	\$100	\$100

devised paradoxes contradicting SEU theory. Ellsberg's paradox can be illustrated as follows. Suppose there is an urn with 90 balls, 30 of which are *Red*, and 60 of which are either *Blue* or *Green*, in unknown proportion. Now consider Table 1.

A person who prefers F to G should also prefer F' to G', since the only difference is the (constant) consequence for a Green ball, which is the same within each choice. If a person prefers F to G, the theory implies that $S(Red) > S(Blue)$, and if that person prefers G' to F', then $S(Red) < S(Blue)$, a contradiction. Indeed, many people exhibited this paradoxical choice pattern even when confronted with this argument. One interpretation is that people are averse to ambiguity, as well as to risk. Others suggested that the decider might distrust that the urns are identical in both choices.

1.5 Paradoxical Risk Attitudes

In *EU* theory, the shape of the utility function determines risk attitudes. For example, for $x > 0$, if $u(x) = x^b$, then the person should be risk-averse if $b < 1$, and risk-seeking if $b > 1$. However, many people are both risk-seeking, when p is small, and risk-averse, when p is moderate to large. Furthermore, many people show risk-aversion for small probabilities of heavy losses (they buy insurance) and they accept risks to avoid certain or highly probable losses.

Whereas Allais considered paradoxical choices 'rational', and theory to be wrong, Savage considered paradoxical choices to be human 'errors' that should be corrected by theory. Many psychologists considered the contradiction between theory and behavior to mean that descriptive theory need not be rational. In this purely descriptive approach, a choice paradox is merely a clear contradiction between theory and human behavior.

Paradoxical risk attitudes and the Allais paradoxes can be described by a theory in which decision weights are a function of probabilities (Edwards 1954, Kahneman and Tversky 1979). Prospect theory (Kahneman and Tversky 1979) described many of the empirical phenomena known by the 1970s. However, this theory was restricted to gambles with no more than two non-zero payoffs and it included a number of seemingly *ad hoc* editing rules to avoid implications that were considered both irrational and empirically wrong.

2. Modern Theories and Paradoxes

During the 1980s, a number of investigators converged on an approach that used a weighting function of probability but did not have the problems of prospect theory (See reviews by Quiggin 1993, Luce 2000). This class of models includes rank-dependent expected utility (Quiggin 1993), rank- and sign-dependent utility theory (Luce 2000), and cumulative prospect theory (Tversky and Kahneman 1992, Wakker and Tversky 1993, Wu and Gonzalez 1998), among others.

2.1 Rank-dependent Expected Utility

For gambles with nonnegative consequences, the rank-dependent expected utility (RDEU) of a gamble can be written as follows:

$$RDEU(G) = \sum_{i=1}^{n}\left[W\left(\sum_{j=i}^{n} p_j\right) - W\left(\sum_{j=i+1}^{n} p_j\right)\right]u(x_i) \quad (3)$$

The consequences are ranked such that $0 \leqslant x_1 < x_2 < x_3 < \ldots < x_n$; and W is a strictly monotonic weighting function with $W(0) = 0$ and $W(1) = 1$. This representation (and its extensions to the cases of negative outcomes and uncertain events) can handle the classic paradoxes of Allais and Ellsberg.

However, new paradoxes were soon created to test if rank-dependent models are descriptive of choices that people make. The new paradoxes can be analyzed as the result of combinations of simpler properties.

2.2 Transitivity, Monotonicity, Coalescing, and Branch Independence

Transitivity holds if for all gambles, $A \succ B$ and $B \succ C$ implies $A \succ C$, where $A \succ B$ denotes A is preferred to B.

Monotonicity assumes that, if one consequence of a gamble is improved, holding everything else in the gamble constant, the gamble's utility is improved.

Coalescing holds that, if two (probability-consequence) branches of a gamble yield the same consequence, then such branches can be combined by adding their probabilities without affecting the gamble's utility. For example, (\$0, 0.8; \$100, 0.1; \$100, 0.1) ought to have the same utility as (\$0, 0.8; \$100, 0.2).

Branch independence is the assumption that if two gambles have a common consequence for a given state of the world with known probability, then the value of

the consequence on that common branch should have no effect on the preference order induced by other components of the gambles. Branch independence is weaker than Savage's (1954) 'sure thing' axiom because it does not presume coalescing. It was assumed as an editing rule by Kahneman and Tversky (1979).

Let $G = (x, p; y, q; z, r)$ represent the three-outcome gamble to win x with probability p, y with probability q, and z otherwise ($r = 1-p-q$), where $0 < x < y < z$. For such gambles, restricted branch independence implies,

$$(x, p; y, q; z, r) \succ (x', p; y', q; z, r)$$
$$\text{if and only if}$$
$$(x, p; y, q; z', r) \succ (x', p; y', q; z', r) \qquad (4)$$

where the branches are distinct, the probabilities are non-zero and they sum to 1. The term 'restricted' refers to the constraint that the probabilities and number of outcomes are fixed in all of the gambles. Although branch independence is implied by *EV*, *EU*, SEU, and certain other theories, it can be violated by RDEU (equation 3) if the common consequence (z or z') changes rank. However, if the common consequence maintains the same rank (i.e., the same cumulative probability), the case is termed comonotonic branch independence, which is implied by Eqn. 3.

One can view the constant consequence paradox of Allais as a violation of (non-comonotonic) restricted branch independence, coalescing, and transitivity. Thus, a theory that violates either branch independence (such as RDEU) or coalescing can explain this Allais paradox.

2.3 Paradoxical Violations of Branch Independence

RDEU (Eqn. (3)) accounts for the observed pattern of risk-seeking for small probabilities to win big prizes and risk-aversion for medium to large probabilities to win modest positive consequences (Tversky and Kahneman 1992). It also accounts for Allais paradoxes if the cumulative weighting function, W in Eqn. (4), has an inverse-S shape (Wu and Gonzalez 1998) in which a probability change near zero or 1 has a greater effect than the same change near 1/2.

However, Birnbaum and McIntosh (1996) reported a pattern of violation of branch independence opposite to the prediction of this inverse-S weighting function. This pattern can be illustrated as follows:

S: 0.80 probability to win \$2
0.10 probability to win \$40
0.10 probability to win \$44

R: 0.80 probability to win \$2
0.10 probability to win \$10
0.10 probability to win \$96

Choices S' and R' differ only in the consequence on the common branch of probability 0.8.

S': 0.10 probability to win \$40
0.10 probability to win \$44
0.80 probability to win \$110

R': 0.10 probability to win \$12
0.10 probability to win \$90
0.80 probability to win \$110

According to the inverse-S weighting function, RDEU implies that violations of branch independence should follow a pattern of $R \succ S$ and $S' \succ R'$. However, the opposite pattern occurs significantly more often than this pattern predicted by the weighting function of RDEU (see review in Birnbaum 1999). Although violations are compatible with Eqn. 3, the observed pattern of violations is the opposite of what was assumed in order to fit Allais paradoxes.

2.4 Configurally Weighted Utility Models

Birnbaum (1997) reviewed configural weight models that he and his colleagues had used in previous research. These models satisfy neither restricted branch independence nor coalescing. They reconcile data that seemed to imply the inverse-S and data that seemed to imply the opposite. Birnbaum's models and the RDEU class of models (Eqn. (3)) are related by being different special cases of the following configural weight model:

$$CWU(G) = \sum_{i=1}^{n} w(x_i, G)u(x_i) \qquad (5)$$

where $w(x_i, G)$ is the configural weight of consequence x_i in gamble G, and $u(x)$ is the utility function. In the transfer of attention exchange (TAX) model, lower-valued consequences 'tax' weight from higher valued ones. A simple version of this model, with $u(x) = x$ for $0 < x <$ \$150 has given a good approximation to empirical choices by undergraduates. In this model, the weight of a distinct branch (probability-consequence pair) is proportional to $p^{.7}$, which is modified by the configural tax. In two-outcome gambles, the lower branch takes 1/3 of the weight of the higher valued branch; in three-outcome gambles, any lower branch takes 1/4 of the weight of any higher-valued branch. Birnbaum (1997) proposed tests between this model and RDEU.

2.5 Paradoxical Violations of Stochastic Dominance and Event-splitting

Birnbaum (1997) proposed the following choice:

I: 0.05 probability to win \$12
0.05 probability to win \$14
0.90 probability to win \$96

J: 0.10 probability to win \$12
0.05 probability to win \$90
0.85 probability to win \$96

Gamble *I* stochastically dominates *J* because the probability of getting a higher prize than *t* given gamble *I* exceeds or equals that of gamble *J* for all *t*. *EV*, *EU*, and RDEU imply that people should choose *I* over *J*. The TAX model, fit to previous data, implies people will choose *J*.

Birnbaum and Navarrete (1998) tested this prediction and found that about 70 percent of college students violated stochastic dominance on this and similar choices, in violation of RDEU. RDEU implies coalescing, monotonicity, and transitivity, so it satisfies stochastic dominance. Configural weight models violate coalescing, and violate stochastic dominance for these cases.

According to the TAX model, it should be possible to eliminate the violations of stochastic dominance by event-splitting, as in the following choice:

IS: 0.05 probability to win \$12
0.05 probability to win \$14
0.05 probability to win \$96
0.85 probability to win \$96

JS: 0.05 probability to win \$12
0.05 probability to win \$12
0.05 probability to win \$90
0.85 probability to win \$96

Note that *IS* and *JS* are simply a split versions of *I* and *J*. According to coalescing (and thus RDEU), choices between *I* and *J* and between *IS* and *JS* are the same. More than half of the undergraduates tested chose *J* over *I and IS* over *JS*, contrary to coalescing (Birnbaum, 1999).

Wu (1994) found violations of tail independence, also called ordinal independence, which can be derived as a combination of coalescing, transitivity, and comonotonic branch independence.

2.6 Violations of Lower and Upper Cumulative Independence

To characterize the paradoxical relationship between evidence that appeared to imply an inverse-S weighting function and evidence that appeared to contradict it, Birnbaum (1997) deduced two new paradoxes, which he called lower and upper cumulative independence. Consider the following choice:

S: 0.80 probability to win \$2
0.10 probability to win \$40
0.10 probability to win \$44

R: 0.80 probability to win \$2
0.10 probability to win \$10
0.10 probability to win \$98

Now increase the common prize of \$2 to \$10 on the 0.8 branch of both gambles and coalesce it to the \$10 branch in *R*. Now, increase the consequence of \$40 to \$44 and coalesce it with the \$44 branch on the left. If a person preferred *S* to *R*, they should definitely prefer S'' to R'':

S'': 0.80 probability to win \$10
0.20 probability to win \$44

R'': 0.90 probability to win \$10
0.10 probability to win \$98

However, more people switch preference from *S* to R'' than from *R* to S'', contrary to monotonicity!

The following choice illustrates upper cumulative independence:

S': 0.10 probability to win \$40
0.10 probability to win \$44
0.80 probability to win \$110

R': 0.10 probability to win \$10
0.10 probability to win \$98
0.80 probability to win \$110

If a person prefers R' to S', then they should prefer R''' to S''':

S''': 0.20 probability to win \$40
0.80 probability to win \$110

R''': 0.10 probability to win \$10
0.90 probability to win \$98

Note that the common branch has been reduced in value to \$98 in both gambles, but the consequence of \$44 has been reduced to \$40 in gamble S''', which should make that gamble relatively worse. More people switched from R' to S''' than the opposite (Birnbaum 1999, Birnbaum and Navarrete 1998).

These modern paradoxes, which violate RDEU, were predicted by configural weight models. Undoubtedly, new paradoxes will be developed to confound new theories. Each new finding represents a phenomenon that must be explained by newer theories. Although some are pessimistic that any transitive theory will be able to explain all of the paradoxes, others continue the search for a theory that can explain the widest domain of behavior.

See also: Bayesian Theory: History of Applications; Decision and Choice: Random Utility Models of Choice and Response Time; Decision Biases, Cognitive Psychology of; Decision Making, Psychology of; Decision Research: Behavioral; Decision Theory: Bayesian; Decision Theory: Classical; Game Theory; Probability and Chance: Philosophical Aspects; Rational Choice Explanation: Philosophical Aspects; Subjective Probability Judgments; Utility and Subjective Probability: Empirical Studies

Bibliography

Allais M, Hagen O (eds.) 1979 *Expected Utility Hypothesis and the Allais Paradox*. Reidel, Dordrecht, The Netherlands

Bernoulli D 1738 Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolianae*, Vol. 5, pp. 175–92. English translation: Bernoulli D (1954). Exposition of a new theory on the measurement of risk. *Econometrica* **22**: 23–36
Birnbaum M H 1997 Violations of monotonicity in judgment and decision-making. In: Marley A A J (ed.) *Choice, Decision, and Measurement: Essays in honor of R. Duncan Luce*. Erlbaum, Mahwah, NJ, pp. 73–100
Birnbaum M H 1999 Testing critical properties of decision making on the Internet. *Psychological Science* **10**: 399–407
Birnbaum M H, McIntosh W R 1996 Violations of branch independence in choices between gambles. *Organizational Behavior and Human Decision Processes* **67**: 91–110
Birnbaum M H, Navarrete J B 1998 Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty* **17**: 49–78
Edwards W 1954 The theory of decision making. *Psychological Bulletin* **51**: 380–417
Ellsberg D 1961 Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics* **75**: 643–9
Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263–91
Luce R D 2000 *Utility of Gains and Losses: Measurement-theoretical and Experimental Approaches*. Erlbaum, Mahwah, NJ
Quiggin J 1993 *Generalized Expected Utility Theory: The Rank-dependent Model*. Kluwer, Boston, MA
Savage L J 1954 *The Foundations of Statistics*. Wiley, New York
Tversky A, Kahneman D 1992 Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* **5**: 297–323
von Neumann J, Morgenstern O 1947 *Theory of Games and Economic Behavior*, 2nd edn. Princeton University Press, Princeton, NJ
Wakker P, Tversky A 1993 An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty* **7**: 147–76
Wu G 1994 An empirical test of ordinal independence. *Journal of Risk and Uncertainty* **9**: 39–60
Wu G, Gonzalez R 1998 Common consequence conditions in decision making under risk. *Journal of Risk and Uncertainty* **16**: 115–39

M. H. Birnbaum

Decision and Choice: Random Utility Models of Choice and Response Time

Introduction

The task of choosing a single 'best' option from some available, potentially infinite set of options R has received considerable study in psychology and economics. Random utility models of choice account for the stochastic variability underlying these choices (i.e., the same choice is not necessarily made on repeated presentations of the same set of options) by assuming that there exist a random variable $U(x)$ for each option x and a joint probability measure for these random variables such that the probability of choosing a particular option x from the set of available options is equal to the probability that $U(x)$ takes on a value greater or equal to the values of all other random variables (e.g., see *Luce's Choice Axiom*; *Utility and Subjective Probability: Contemporary Theories*). The basic choice paradigm is extended here by considering, in addition to the option chosen, the point in time that a choice is made from the set of available options. The random utility model can be tailored to cover this situation by replacing the utility variable $U(x)$ by $V(x) = \phi[U(x)]$ (ϕ some monotonically decreasing transformation), where $V(x)$ can be interpreted as decision time for choosing option x, and by replacing the maximum utility rule by a minimum decision time rule: the option chosen is the one which happens to be associated with the minimum choice (or decision) time with respect to all options in the available set. This model will be referred to as 'horse race' random utility model.

A number of problems arise in the study of 'horse race' random utility models for choice and response time that have found only partial solutions: (a) What conditions on the observable choice probabilities and decision times are necessary and sufficient for a random utility representation? (b) What are the consequences of assuming stochastic independence between the time of choice and the identity of the option chosen, and which assumptions on the joint distribution function do imply this independence? (c) What are possible generalizations to other choice paradigms? The presentation here is partly based on Marley and Colonius 1992 and Marley 1992, 1989. Some related results (not explicitly referred to in the following) can be found in Bundesen 1993, Robertson and Strauss 1981, and Vorberg 1991.

1. 'Horse Race' Random Utility Models

Let $R = \{x, y, \ldots\}$ be a finite set of potential choice options (for an extension of the theory to infinite choice sets, see Resnick and Roy 1992); the subset X of R containing at least two elements is the currently available choice set. $T(X)$ is a random variable denoting the time at which a choice is made; and $C(X)$ is a random variable denoting the element chosen from X. For $t \geqslant 0$

$$P_X(x; t) = Pr[C(X) = x \cap T(X) > t]$$

is the probability that options x is chosen from X after time t, and $\{P_X(x; t)\colon x \in X \subseteq R\}$, or (R, P) for short, is called joint structure of choice probabilities and response times. (R, P) is called complete if it is defined for all subsets of X of $R$with $|X| \geqslant 2$.

A complete joint structure of choice probabilities and response times (R, P) is said to satisfy a '*horse race*' random utility model if for any $X \subseteq R$ with

$|X| \geqslant 2$ there exists a probability measure $Pr_X[\ ]$ such that for the collection of nonnegative random variables $\{V(x)\colon x \in X\}$

$$P_X(x;\, t) = Pr_X[t < V(x) < V(y)\colon y \in X \backslash \{x\}] \quad (1)$$

For simplicity, all distributions are assumed to be absolutely continuous, i.e., they must possess a density function. However, as observed in Resnick and Roy 1992, absolute continuity is not necessary for the results cited here. Note that with

$$\begin{aligned} P_X(x) &= \lim_{t \downarrow 0} Pr_X[t < V(x) < V(y)\colon y \in X \backslash \{x\}] \\ &= Pr_X[V(x) < V(y)\colon y \in X \backslash \{x\}] \end{aligned} \quad (2)$$

the collection of choice probabilities $\{P_X(x)\colon x \in X \subseteq R\}$ constitutes a system of choice probabilities with

$$P_X(x) = Pr_X[V(x) = \min\{V(y)\colon y \in X\}] \quad (3)$$

satisfying a random utility representation with min replacing the usual max.

In this article the problem of finding conditions on the—at least theoretically—observable choice probabilities and response time distributions that are necessary and sufficient for the existence of a particular representation in terms of a collection of underlying random variables is referred to as a characterization problem.

Given a (complete) joint structure of choice probabilities and response times, what conditions on the (survival) functions $P_X(x; t)$ are necessary and sufficient for the existence of a (possibly dependent) 'horse race' random utility representation for that joint structure?

A complete answer to this problem is still open. However, a number of results in special cases have been found.

1.1 Representation by Independent Random Variables

A partial answer is given by the following

Theorem 1 (Marley and Colonius 1992).

(a) For a collection of choice probabilities and response times $\{P_X(x;\, t)\colon x \in X\}$ on a fixed set X with $|X| \geqslant 2$, there exist independent random variables $V(x)$, $x \in X$, with unique distributions, such that Eqn. (1) holds provided $P_X(x;\, t)$ is absolutely continuous and positive for all $t \geqslant 0$.

(b) A complete joint structure of choice probabilities and response times (R, P) can be represented by an independent 'horse race' random utility model, with unique distributions, if the conditions of (a) hold and

$$\left[\sum_{y \in X} P_X(y;\, t)\right]^{-1} (\mathrm{d}/\mathrm{d}t) P_X(x,\, t) \quad (4)$$

is independent of $X \subseteq R$ for all $t \geqslant 0$.

Thus, given the above regularity conditions on the response time distributions, any set of choice probabilities and response times on a fixed finite set X can be represented by independent random variables such that the 'horse race' Eqn. (1) holds. Following Berman 1963 the proof of (a) is based on setting up a differential equation with the following unique solution in terms of the (survival) distributions of independent random variables

$$\begin{aligned} G_x^X(t) &\equiv Pr[V(x) > t] \\ &= \exp\left\{\int_0^t Pr[T(X) > s]^{-1}\, \mathrm{d}P_X(x;\, s)\right\} \\ &= \exp\left\{\int_0^t \left[\sum_{y \in X} P_X(y;\, s)\right]^{-1} \mathrm{d}P_X(x;\, s)\right\} \\ &= \exp\left\{\int_0^t \left[\prod_{y \in X} G_y^X(s)\right]^{-1} \mathrm{d}P_X(x;\, s)\right\} \end{aligned} \quad (5)$$

A related result, in the context of parallel-serial processing analysis, appears in Townsend 1976. Some of the above distributions, but not all, may be improper, i.e., they may have positive measure at infinity implying an infinite decision time for that alternative, even if the distribution of response time $T(X)$ is finite almost surely (see Dzhafarov 1993 for a related interpretation and further results under weaker conditions in a different context). Note that Theorem 1 (a) does not say that there always exists a unique set of independent random variables representing a complete joint structure of choice probabilities and response times irrespective of the subset X. The latter postulate is in fact paraphrased in Theorem 1 (b). Thus, the general characterization problem is still unsolved leaving open the possibility that for some joint structures no independent representation over all subsets X exists.

1.2 Hazard Rate Reformulation

In the theory of competing risks (e.g., David and Moeschburger 1978) the 'event' associated with each $x \in X$ is reinterpreted as a cause (e.g., of failure or death), and thus the selection $x \in X$ then corresponds

to the cause associated with x being the first to occur when X contains the possible causes. The instantaneous rate of failure from cause x when all causes in X act simultaneously, called cause-specific (or crude hazard rate), is defined as

$$h_x^X(t) = \lim_{\delta \downarrow 0} Pr[t < T(X) \leqslant t+\delta \cap C(x) = x | T(X) > t| \quad (6)$$

In the choice context this can be interpreted as the instantaneous rate of choosing option x from X at time t given that no choice was made before t. Provided that $Pr[T(X) > t] \neq 0$, going over to the unconditioned probability in Eqn. (6) gives

$$h_x^X(t) = [(-\mathrm{d}/\mathrm{d}t)P_X(x; t)] \Big/ \left(\sum_{y \in X} P_X(y; t) \right) \quad (7)$$

Note that the right hand side of Eqn. (7) agrees (except in sign) with that given in the statement of the Marley-Colonius Theorem (b). Thus, for a fixed set X, the crude hazard rate leads directly to the desired random variable representation by setting

$$Pr[V(x) > t] = \exp\left[-\int_0^t h_x^X(s)\, \mathrm{d}s \right] = G_x^X(t) \quad (8)$$

and for a common representation on all subsets X of R, h_a^X must not depend on X (as in part (b) of Theorem 1).

$P[T(X) > t]$ is commonly referred to as the overall survival distribution (for set X) with a corresponding overall hazard rate $h^X(t)$; i.e., by definition

$$h^X(t) = -\mathrm{d}/\mathrm{d}t\, Pr[T(X) > t]/Pr[T(X) > t] \quad (9)$$

and so from

$$Pr[T(X) > t] = \sum_{y \in X} P_X(y; t)$$

and Eqn. (7), it follows that

$$h^X(t) = \sum_{x \in X} h_x^X(t) \quad (10)$$

Thus, under the stated regularity conditions, the independent random variables $V(x)$, $x \in X$, constructed above can be used to generate the survival distributions $P_X(x; t)$. In competing risks theory this is a well-known result, often referred to as the 'non-identifiability problem.' It means that any set of competing risk data can be explained by some independent risk model which, in most empirical situations, appears to be unrealistic. One way to establish identifiability of a dependent model is to assume that the $V(x)$ follow some (flexible) parametric family, to estimate the parameters of the multivariate distribution, and to test for independence (Tsiatis 1975).

1.3 Independence of Option Chosen from Time of Choice

The following condition is often made in the competing risks theory and it turns out to be equivalent to a condition specifying a large family of random utility models. A joint structure of choice probabilities and response times (S, P) satisfies the proportional hazard rate (PHR) condition if and only if for any subset X of R there are constants $C_X(x)$ such that

$$h_x^X(t) = C_X(x) h^X(t) \quad (11)$$

where $h^X(t)$ is the overall hazard rate for $T(X)$ and $h_x^X(t)$ are the cause specific hazard rates. After some algebra,

$$P_X(x; t) = C_X(x) Pr[T(X) > t]$$

In particular, because $P_X(x) = \lim_{t \downarrow 0} P_X(x; t)$,

$$P_X(x; t) = Pr[C(X) = x] Pr[T(X) > t] \quad (12)$$

Reversing the steps leads to the following result (see Kochar and Proschan 1991).

Theorem 2. The proportional hazard rate condition (PHR) is equivalent to assuming stochastic independence between $T(X)$ and $C(X)$ for any $X \subseteq R$.

Intuitively, proportionality of the hazard rates means that the time of choice gives no information as to the identity of the element chosen, and vice versa. This seems like a rather strong assumption in an empirical context, and the following result concerning Luce's choice model (see *Luce's Choice Axiom*) supports this view. Luce's choice model holds for a system of choice probabilities $\{P_X(x)\colon x \in X \subseteq R\}$ provided there is a ratio scale v on R such that for $P_X(x) \neq 0, 1$

$$P_X(x) = \frac{v(x)}{\sum_{y \in X} v(y)} \quad (13)$$

For simplicity in the following, we assume that all choice probabilities are nonzero. The result can be

generalized when this is not the case by adding a connectivity and a transitivity condition (Luce 1959, Theorem 4, p. 25).

Theorem 3 (Marley and Colonius 1992). Consider a (complete) independent 'horse race' random utility model (R, P) where for each $x \in X \subseteq R$ with $|X| \geqslant 2$, $P_X(x; t)$ is positive and absolutely continuous for all $t > 0$. If $T(X)$ is stochastically independent of $C(X)$, then the choice probabilities satisfy Luce's choice model.

Given the known limited empirical validity of Luce's choice model in empirical preference situations, this result implies that one must study dependent 'horse race' random utility models as long as one is not willing to drop the assumption of independence between $T(X)$ and $C(X)$.

1.4 Generalized Stable Survival Functions

A large class of (possibly) dependent 'horse race' random utility models based on extreme value distributions can be generated from the concept of a generalized stable survival function.

For simplicity, in this section assume $R = \{x_1, \ldots, x_n\}$. We call

$$P_R(t_1, \ldots, t_n) = Pr[V(x_1) > t_1, \ldots, V(x_n) > t_n)$$

a (multivariate) survival function. P_R is called generalized stable if there is a strictly monotone decreasing function η and a constant $\mu > 0$ such that for all $\alpha > 0$, $t_i \geqslant 0$, $i = 1, \ldots, n$,

$$(\eta \circ P_R)(\alpha t_1, \ldots, \alpha t_n) = \alpha^{\mu}(\eta \circ P_R)(t_1, \ldots, t_n) \quad (14)$$

where $\circ$ denotes concatenation of functions. Letting $G_R = (\eta \circ P_R)$ this amounts to

$$G_R(\alpha t_1, \ldots, \alpha t_n) = \alpha^{\mu} G_R(t_1, \ldots, t_n)$$

i.e., G_R is homogeneous of degree μ. Furthermore, a survival function P_R is called a strictly monotone transform of a generalized stable survival function Q_R if there is a strictly monotone increasing function γ (of the random variables associated with Q_R) with $\gamma(0) = 0$, $\gamma(\infty) = \infty$ such that for $t_i \geqslant 0$, $i = 1, \ldots, n$,

$$P_R(t_1, \ldots, t_n) = Q_R(\gamma(t_1), \ldots, \gamma(t_n))$$

Theorem 4 (Marley 1989). Any 'horse race' random utility model that is generated by a strictly monotone transform of a generalized stable survival function is such that $T(X)$ is stochastically independent of $C(X)$ for any $X \subseteq R$.

As shown by way of an example in Resnick and Roy 1992 (Sect. 5) the converse of this result does not hold in general, i.e., independence between $T(X)$ and $C(X)$ does not lead back to the class of generalized stable survival functions. Robertson and Strauss 1981 show that the converse of Theorem 4 is true if the survival function P_R (or some strictly monotone transformation of it) belongs to the generalized Thurstone class: for $t_i \geqslant 0$, $i = 1, \ldots, n$,

$$P_R(t_1, \ldots, t_n) = P(u(x_1)t_1, \ldots, u(x_n)t_n)$$

where $u(x_i) \geqslant 0$ and P is a survival function that is independent of R.

2. General Feature Model

Marley 1989 presents a general class of models which lends itself easily to a process interpretation for the choice paradigm. With $R = \{x_1, \ldots, x_n\}$ suppose each option x_i has m (random) components. Thus, there is a matrix of random variables $(Z_{ik})_{i=1,\ldots,n}^{k=1,\ldots,m}$, and it is assumed that the previous random variables $V(x_i)$ are given by

$$V(x_i) = \min_{k=1,\ldots,m} Z_{ik}$$

In the process interpretation Z_{ik} corresponds to the 'time of occurrence' of feature (or dimension) $k$for option i, and that option is selected which has the earliest such event over all features (or dimensions). Thus, the survival function is

$$\begin{aligned} P(t_1, \ldots, t_n) &= Pr[V(x_1) > t_1, \ldots, V(x_n) > t_n] \\ &= Pr\left[\bigcap_{i=1}^{n} \{V(x_i) > t_i\}\right] \\ &= Pr\left[\bigcap_{i=1}^{n} \left\{\min_{k=1,\ldots,m} Z_{ik} > t_i\right\}\right] \\ &= Pr\left[\bigcap_{i=1}^{n} \bigcap_{k=1}^{m} \{Z_{ik} > t_i\}\right] \end{aligned} \quad (15)$$

For the survival function P to be generalized stable, restrictions must be placed on the distribution of the random variables Z_{ik}. However, no independence of $(Z_{i1}, \ldots, Z_{im})$ for distinct i, $1 \leqslant i \leqslant n$ is assumed as one might be able to give a 'process' interpretation of such dependencies.

Example: Multivariate exponential distribution (Marley 1989, Pickands 1981).

Let $A = (a_{ik})$ be a $n \times m$ matrix with non-negative and finite entries such that there is at least one positive value in each row, and no two values are identical in any column. Also, $Z_1, \ldots, Z_m$ are mutually independent random variables with standard exponential distri-

bution, i.e., $Pr[Z_k > t] = \exp(-t)$, and $V(x_1), \ldots, V(x_n)$ have components

$$V(x_i) = \min_{k=1,\ldots,m} \frac{Z_k}{a_{ik}} \tag{16}$$

Inserting this into Eqn. (15) yields the survival function

$$P(t_1, \ldots, t_n) = \exp - \left[\sum_{k=1}^{m} \max_{i=1,\ldots,n} a_{ik} t_i \right]$$

or,

$$-\ln P(t_1, \ldots, t_n) = \sum_{k=1}^{m} \max_{i=1,\ldots,n} a_{ik} t_i \tag{17}$$

With $G_R \equiv (-\ln \circ P)$ this clearly satisfies the condition of Theorem 4. Marley 1989 demonstrates how this model easily explains the standard 'bicycle/pony' or 'red bus/blue bus' examples that Luce's choice model cannot adequately represent.

3. Subset Choices

The above development can be extended to cases where one is allowed to choose a subset of 'acceptable' options from the set of available options rather than a single element (e.g., approval voting). Extending the previous notation, let $Q_X(Y; t)$, $t \geqslant 0$, be the probability that the nonempty subset $Y \in X$ is selected after time t, and assume that

$$Q_X(Y; t) = Pr[C(X) = Y \cap T(X) > t] \tag{18}$$

where (as before) $T(X)$ is a random variable denoting the time at which the subset choice is made, and $C(X)$ is a random variable denoting the subset chosen. Such a collection $\{Q_X(Y; t)\colon \emptyset \in Y \subseteq X \subseteq R\}$, or $\{R, Q\}$ for short, will be called a joint structure of transition probabilities and response times, and (R, Q) will be called complete if it is defined over all subsets of R, as before. A (complete) joint structure of transition probabilities and response times (R, Q) satisfies a random (subset) advantage model provided there exists for any $X \subseteq R$ a probability measure $Pr_X[\ ]$ such that for the collection of nonnegative random variables $\{V(Y)\colon \emptyset \subset Y \subseteq X \subseteq R\}$

$$Q_X(Y; t) = Pr_X[t < V(Y) < V(Z)\colon \text{ for all } Z \text{ with } \emptyset \subset Z \subseteq X, Z \neq Y] \tag{19}$$

If the $V(Y)$, for a fixed X, are independent, then the random (subset) advantage model is independent. The idea here is that some 'event' is associated with each non empty subset of the available choice set, and that the time of choice and the subset selected are determined by the occurrence of the first such event. As before, assuming absolute continuity,

$$Q_X(Y) = Pr_X[V(Y) < V(Z)\colon \text{ for all } Z \text{ with } \emptyset \subset Z \subseteq X, Z \neq Y] = \lim_{t \downarrow 0} Q_X(Y; t) \tag{20}$$

is the probability of selecting the subset Y of X. The following is an immediate result paralleling Theorem 1 (a).

Theorem 5 (Marley and Colonius 1992). A collection of transition probabilities and response times $\{Q_X(Y; t)\colon \emptyset \subset Y \subseteq X\}$ can be uniquely represented by an independent random (subset) advantage model provided $Q_X(Y; t)$, $\emptyset \subset Y \subseteq X$, are absolutely continuous and positive for all $t \geqslant 0$.

4. Structures of Transition Probabilities

Subset selection as defined in the preceding section might only be an intermediate stage in the determination of a single 'best' element by repeatedly selecting subsets from the subset determined in the previous subset selection. The class of Markovian transition structures has the recursive property

$$P_X(x) = \sum_{Y \subseteq X} Q_X(Y) P_Y(x) \tag{21}$$

However, Eqn. (21) is compatible with any set of choice probabilities, e.g., define

$$Q_X(Y) = \begin{cases} P_X(x) & \text{if } Y = \{x\} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

Thus, some restrictions must be put on the transition probabilities. The following *proportionality* condition leads to Tversky's 1972a, b elimination-by-aspects model (EBA):

$$\frac{Q_X(Y)}{Q_X(Z)} = \frac{\sum_{J \cap X = Y, \emptyset \subset J \subset R} Q_R(J)}{\sum_{J \cap X = Z, \emptyset \subset J \subset R} Q_R(J)} \tag{23}$$

for $\emptyset \subset Y, Z \subseteq R$, provided the denominators are both positive, and if one denominator vanishes, so does the other.

To formulate a generalization of Theorem 3 in the context of subset choices, the following plausible min-consistency constraint on the response time distri-

bution is needed which, however, does not constrain the distribution of the $V(Y)$ to have any specific form:

$$Pr_X[V(Y) \leqslant t] = Pr_R[\min\{V(J) | J \cap X = Y, \emptyset \subset J \subseteq R\} \leqslant t] \quad (24)$$

Theorem 6. (Marley and Colonius 1992). Consider an independent random (subset) advantage model (R, Q) where for each $\emptyset \subset Y \subseteq X \subseteq R$, $Q_X(Y; t)$ is absolutely continuous and positive for all $t \geqslant 0$ and $Q_X(Y)$ is nonzero. If the model satisfies min-consistency, and if the subset chosen, $C(Y)$, is independent of the time of choice, $T(Y)$, then the transition probabilities satisfy proportionality.

Thus, independence between $C(Y)$ and $T(Y)$ in this subset choice context leads to the transition probabilities of the EBA model paralleling the result (Theorem 3) for the single-element choice context with the special case of the EBA model, i.e., Luce's choice model. Finally, it has been shown (Marley 1989) that the general feature model can easily be adjusted such that it contains the EBA model as a special case with a generalized stable survival function where

$$-\ln P(t_1, \ldots, t_n) = \sum_{J \subseteq R} V_R(J) \max_{i \in J} t_i \quad (25)$$

where $V_R(J) \geqslant 0$ for all $J \subseteq R$.

5. Conclusions

The introduction of the notion of response (or decision) time into the random utility approach to modeling choice behavior has been very fruitful. It has brought about new insights into existing stochastic choice models and their characterizations, and it has generated new questions, many of which have only found partial solutions up to now.

See also: Bayesian Theory: History of Applications; Dynamic Decision Making; Luce's Choice Axiom; Stochastic Dynamic Models (Choice, Response, and Time); Utility and Subjective Probability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies

Bibliography

Berman S M 1963 Note on extreme values, competing risks and semi-Markov processes. *Annals of Mathematical Statistics* **34**: 1104–6

Bundesen C 1993 The relationship between independent race models and Luce's Choice Axiom. *Journal of Mathematical Psychology* **37**: 446–71

David H A, Moeschburger M L 1978 *The Theory of Competing Risks*. Griffin, London

Dzhafarov E N 1993 Grice-representability of response time distribution families. *Psychometrika* **58**: 281–314

Kochar S C, Proschan F 1991 Independence of time and cause of failure in the multiple dependent competing risks model. *Statistica Sinica* **1**: 295–9

Luce R D 1959 *Individual Choice Behavior*. Wiley, New York

Marley A A J 1989 A random utility family that includes many of the 'classical' models and has closed form choice probabilities and reaction times. *British Journal of Mathematical Statistical Psychology* **42**: 13–36

Marley A A J 1992 A selective review of recent characterizations of stochastic choice models using distribution and functional equation techniques. *Mathematical Social Sciences* **23**: 5–29

Marley A A J, Colonius H 1992 The 'horse race' random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology* **36**: 1–20

Pickands J III 1981 *Multivariate Negative Exponential and Extreme Value Distributions*. Manuscript, Department of Statistics, University of Pennsylvania, Philadelphia, PA

Resnick S I, Roy R 1992 On min-stable horse races with infinitely many horses. *Mathematical Social Sciences* **23**: 119–45

Robertson C A, Strauss D J 1981 A characterization theorem for random utility variables. *Journal of Mathematical Psychology* **23**: 184–9

Townsend J T 1976 Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology* **14**: 219–38

Tsiatis A 1975 A non-identifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Science, USA* **72**: 20–2

Tversky A 1972a Choice by elimination. *Journal of Mathematical Psychology* **9**: 341–67

Tversky A 1972b Elimination by aspects: A theory of choice. *Psychology Review* **79**: 281–99

Vorberg D 1991 *Characterizations of Independent Random Utility Models Which Imply the Choice Axiom*. Technical Report, Technische Universität Braunschweig, Braunschweig

H. Colonius

Decision Biases, Cognitive Psychology of

Normative analyses of rational decision making dictate *a priori* how decisions ought to be made. Descriptive analyses of decision making are based on experimental studies and emphasize the role of information processing in people's decisions. The empirical evidence indicates that people's decisions are often at odds with the assumptions of the rational theory. The following describes some of the cognitive mechanisms that underlie decision behavior and cause it to depart form the normative benchmark, yielding systematic decision biases.

1. Normative and Descriptive Analyses

The study of decision making is an interdisciplinary enterprise involving economics, political science, and psychology, as well as statistics and philosophy. One

can distinguish two approaches to the analysis of decision making: normative and descriptive. Central to many of the social sciences is the normative approach, which assumes a rational decision maker with well-defined preferences. Every option is assumed to have a subjective 'utility' for the individual, a well-established position in his or her preference ordering (von Neumann and Morgenstern 1947). Because preferences are clear and stable they are expected to be invariant across normatively equivalent assessment methods ('procedure invariance'), and across logically equivalent ways of describing the options ('description invariance'). This conception, known as the rational theory of choice, is based primarily on *a priori* considerations and on normative assumptions rather than on empirical observation. Because the normative treatment specifies clear and compelling principles of rational behavior, it serves as a benchmark against which descriptive studies of decision making are compared.

The descriptive approach to individual decision making is based on empirical observation and experimental studies of choice behavior, which emphasize the role of information processing in people's decisions. The empirical evidence is that people's decisions often depart from the normative benchmark in systematic ways which can be viewed as decision biases. In essence, preferences appear to be formed, not merely revealed, during their elicitation, and their formation is often influenced by factors that ought not to matter from a normative perspective. The formation of preferences appears to depend on subtleties such as problem frame, method of elicitation, and the valuations and attitudes that these trigger. In what follows, selected empirical findings about individual decision making are presented and several cognitive psychological principles that underlie the decision-making process, and lead to systematic bias, are described.

2. *Diminishing Sensitivity, Risk Attitudes, and Frames*

Decisions are often 'risky' in the sense that their outcomes are not known with certainty. Presented with a choice between a risky prospect that offers a 50 percent chance to win $200 (otherwise nothing) and an alternative of receiving $100 for sure, most people prefer the sure gain over the gamble, although the two prospects have the same expected value. (Expected value is the sum of possible outcomes weighted by their probability of occurrence.) Preference for a sure outcome over a risky prospect of equal expected value is called *risk averse*; indeed, people tend to be risk averse when choosing between prospects with positive outcomes. The tendency towards risk aversion can be explained by the notion of diminishing sensitivity, first formalized by Daniel Bernoulli in 1738. Just as the impact of a candle is greater when it is brought into a dark room than into a room that is well lit so, suggested Bernoulli, the utility resulting from a small increase in wealth will be inversely proportional to the amount of wealth already in one's possession. It has since been assumed that people have a subjective utility function, and that preferences should be described using expected utility instead of expected value. According to expected utility, the worth of a gamble offering a 50 percent chance to win $200 (otherwise nothing) is 0.50 * u($200), where u is the person's concave utility function. (A function is concave or convex if a line joining two points on the curve lies entirely below or above the curve, respectively.) It follows from a concave function that the subjective value attached to a gain of $100 is more than 50 percent of the value attached to a gain of $200, which entails preference for the sure $100 gain and, hence, risk aversion.

Consider now a choice between losses. When asked to choose between a prospect that offers a 50 percent chance to lose $200 (otherwise nothing) and the alternative of losing $100 for sure, most people prefer to take an even chance at losing $200 or nothing over a sure $100 loss. This is because diminishing sensitivity applies to negative as well as to positive outcomes: the impact of an initial $100 loss is greater than that of the next $100. This results in a convex function for losses and a preference for risky prospects over sure outcomes of equal expected value, called *risk seeking*. With the exception of prospects that involve very small probabilities, risk aversion is generally observed in choices involving gains, whereas risk seeking tends to hold in choices involving losses.

These risk attitudes form part of prospect theory (Kahneman and Tversky 1979), an influential descriptive theory of choice, which incorporates an S-shaped value function that is concave for gains and convex for losses. This function is also steeper for losses than for gains: Thus, a loss of $X is more aversive than a gain of $X is attractive, which is known as *loss aversion*. In addition, prospect theory is defined on gains and losses rather than total wealth, which captures the fact that people typically treat outcomes as departures from a current reference point, rather than in terms of final assets. These properties seem compelling and unobjectionable, yet their combination can lead to problematic consequences.

In one example (Tversky and Kahneman 1986), respondents are asked to assume themselves $300 richer and invited to choose between a sure gain of $100 or an equal chance to win $200 or nothing. Alternatively, they are asked to assume themselves $500 richer, and made to choose between a sure loss of $100 and an equal chance to lose $200 or nothing. In accord with the properties described above, most subjects choosing between gains are risk averse and prefer the certain $100 gain, whereas most subjects choosing between losses are risk seeking, preferring the risky prospect over the sure $100 loss. The two

problems, however, are essentially identical: When the initial $300 or $500 wealth is added to the respective outcomes, both problems amount to a choice between $400 for sure vs. an even chance at $300 or $500. This notwithstanding, those choosing between what are framed as gains exhibit a preference for the former outcome, whereas those facing what are framed as losses prefer the latter. This is known as a *framing effect*. It occurs when alternative descriptions of what is essentially the same decision problem give rise to discrepant attitudes and to predictably different choices. Framing effects have been observed in dozens of studies involving economic, medical, and consumer decisions, among others.

3. *Loss Aversion and the Status Quo Bias*

A fundamental fact regarding people's reaction to outcomes is loss aversion: The loss of utility associated with giving up a good is greater than the utility associated with obtaining it (Tversky and Kahneman 1991). An immediate implication of loss aversion is that people will not accept an even chance to win or lose $X, because the loss of $X is more aversive than the gain of $X is attractive. Indeed, people are generally willing to accept an even-chance prospect only when the gain is substantially greater than (about twice as large as) the loss. Loss aversion entails that the loss of utility associated with giving up a good that is in our possession is generally greater than the utility gain associated with obtaining that good. This yields 'endowment effects,' wherein the mere possession of a good (thus viewing it as a potential loss) can lead to higher valuation of it than if it were not in one's possession (Kahneman et al. 1990).

A closely related manifestation of loss aversion is a general reluctance to trade, which is illustrated in a study (Knetsch 1989) in which subjects were divided into two groups: Half of the subjects were given a decorated mug, and the others were given a large bar of Swiss chocolate. Later, each subject was shown the alternative gift, and offered the opportunity to trade their gift for the other. Because the initial allocation of gifts was arbitrary and transaction costs minimal, economic theory predicts that about half the subjects should exchange their gifts. On the other hand, if losses loom larger than gains, then most participants will be reluctant to give up the gift in their possession (a loss) in order to obtain the other (a gain). Indeed, only 10 percent of the participants chose to trade their gifts. This contrasts sharply with the 50 percent predicted by standard economic analysis in which the value of a good does not change when it becomes part of one's endowment.

Loss aversion entails a strong tendency to maintain the status quo, because the disadvantages of departing from it loom larger than the advantages of its alternative (Samuelson and Zeckhauser 1988). A striking framing effect which relies on people's tendency to maintain the status quo has been observed in the context of insurance decisions, when New Jersey and Pennsylvania both introduced the option of a limited right to sue, which entitles automobile drivers to lower insurance rates. The two states differed, however, in what they offered consumers as the default option. New Jersey motorists had to acquire the full right to sue (transaction costs were minimal: a signature), whereas in Pennsylvania the full right to sue was the default. Presented with the choice, only about 20 percent of New Jersey drivers chose to acquire the full right to sue, while approximately 75 percent of Pennsylvania drivers chose to retain it. The difference in adoption rates due to the different frames had financial repercussions estimated at around $200 million dollars (Johnson et al. 1993).

Loss aversion promotes stability rather than change by inducing people to maintain their current position. Among other things, the reluctance to change induced by loss aversion can hinder the negotiated resolution of disputes. If each side to a dispute evaluates the opponent's concessions as gains and its own concessions as losses, then agreement will be hard to reach because each side will perceive itself as relinquishing more than it stands to gain.

4. *Compatibility and Reversals*

An important cognitive principle that relates the characteristics of stimuli to the ways in which people code the information and produce a response is the principle of compatibility. This principle has long been recognized by students of perception and motor control. It states that when stimuli and responses are mentally represented, the weight of a stimulus attribute is enhanced to the extent that it is compatible with the required response. Alternative methods of eliciting preference can thus lead to differential weighings of attributes and, consequently, to different choices.

In line with compatibility, for example, setting the price of a gamble is likely to emphasize payoffs more than probabilities because both the price and the payoffs are in monetary units (e.g., dollars). Consequently, a gamble's potential payoff will be weighted more heavily in a pricing task than in a choice between gambles. Consistent with this is the preference reversal phenomenon (Slovic and Lichtenstein 1983), wherein subjects choose a lottery that offers a greater chance to win over another which offers a higher payoff, but then price the latter higher than the former. This pattern has been observed in numerous experiments, including one involving professional gamblers in a Las Vegas casino (Lichtenstein and Slovic 1973), and another offering the equivalent of a month's salary to respon-

dents in the Peoples' Republic of China (Kachelmeier and Shehata 1992).

For another example, consider having to choose one of two options or, alternatively, having to reject one of two options. The two tasks are logically equivalent: If people prefer the first they should reject the second, and vice versa. However, because it is natural to select options based on their strengths, and to reject options based on their weaknesses, compatibility suggests that the strengths of options will loom larger when choosing, whereas options' weaknesses will be weighted more heavily when rejecting.

In one study (Shafir 1993), subjects were presented with pairs of options, one of which-–the 'enriched' option—had more strengths and more weaknesses than the other, 'impoverished' option. Because options' strengths are weighted more heavily when choosing than when rejecting, and their weaknesses are weighted more heavily when rejecting than when choosing, the enriched option was chosen and rejected relatively more often than the impoverished option. Contrary to procedure invariance, logically equivalent tasks give rise to predictably different choices.

5. *Independent and Comparative Evaluation*

Some procedures, such as pricing or rating, involve an independent evaluation of each option, whereas others, such as choice, are inherently comparative in nature. Independent vs. comparative procedures tend to highlight different aspects of options and can thus alter decision. As it turns out, people feel that some attributes (e.g., safety) are more important than others (e.g., cost). There is evidence that the attribute that is judged as more important looms larger in choice, where it can act as a tie-breaker, than in independent evaluation, where its comparative advantage is less prominent (Tversky et al. 1988). This can lead to systematic violations of invariance.

One study, for example, elicited people's support for interventions addressing environmental problems (Kahneman and Ritov 1994). One intervention concerned human health; for example, free medical checkups for farm workers facing the risk of skin cancer. The other concerned environmental protection: the contribution to a fund to provide safe breeding areas for Australian mammal species nearly wiped out by hunters. One group of respondents were asked to choose which of the two interventions they would rather support; a second group were presented with these one at a time and asked to determine the largest amount they would be willing to pay for each intervention. When asked to evaluate each intervention separately, respondents, who were predictably more moved by the hunted animals' plight than by mundane checkups, were willing to pay more, on average, for the safe breeding of the Australian mammals than for skin cancer checkups. However, when faced with a direct choice between these options, most subjects favored checkups for humans over safe breeding for animals. As expected, the issue viewed as more important (cancer in humans rather than breeding of animals) acquired greater prominence in choice, which involves direct comparison between issues, than in separate evaluation, where other considerations prove influential absent a direct comparison.

A related pattern emerges when attributes are difficult to gauge in isolation. Hsee (1996), for example, presented subjects with two second-hand music dictionaries, one with 20,000 entries and a slightly torn cover; the other with 10,000 entries and a cover that was like new. Subjects had little notion of how many entries to expect in a music dictionary. Under separate evaluation, they were willing to pay more for the dictionary with the new rather than the torn cover. When these were evaluated concurrently, however, most subjects obviously preferred the dictionary with twice as many entries, despite its inferior cover.

6. *Preferences Constructed in Context*

People do not always have clear and well-ordered preferences: Instead, they approach decisions as problems that need to be solved, and construct preferences that are influenced by the context of decision. Among other things, decision is thus biased by the ways in which the options are described and by the methods through which preference is elicited. Other aspects of decision that have been found to yield bias include the tendency to delay decision when choices are hard (Tversky and Shafir 1992), the regret anticipated when another option could prove superior (Bell 1982), the influence exerted by costs already incurred (Arkes and Blumer 1985), the effects of temporal separation on future decisions (Loewenstein and Elster 1992), the role of reasons and justification (Shafir et al. 1993, Tetlock 1992), and the occasional inability to predict future tastes or to remember past satisfaction (Kahneman 1994).

Decision biases have been documented in real settings and with experts; for example, in physicians' decisions regarding patient treatment (McNeil et al. 1982, Redelmeier and Shafir 1995), academics' retirement investment decisions (Benartzi and Thaler 1995), and taxi drivers' allocation of working hours (Camerer et al. 1997). Because these biases emerge from underlying cognitive mechanisms, they form an inherent part of the ways in which we process information and make decisions.

See also: Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision and Choice: Paradoxes of Choice; Decision and Choice:

Random Utility Models of Choice and Response Time; Decision Making (Naturalistic), Psychology of; Decision Making: Nonrational Theories; Decision Making, Psychology of; Decision Research: Behavioral; Decision Theory: Classical; Group Decision Making, Cognitive Psychology of; Heuristics for Decision and Choice; Luce's Choice Axiom; Risk: Empirical Studies on Decision and Choice; Utility and Subjective Probability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies

Bibliography

Arkes H R, Blumer C 1985 The psychology of sunk cost. *Organizational Behavior and Human Performance* **35**: 129–40

Bell D E 1982 Regret in decision making under uncertainty. *Operations Research* **30**: 961–81

Benartzi S, Thaler R 1995 Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics* **110**(1): 73–92

Camerer C F 1995 Individual decision making. In: Kagel J H, Roth A E (eds.) *Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ, pp. 587–703

Camerer C, Babcock L, Loewenstein G, Thaler R 1997 Labor supply of New York City cab drivers. One day at a time. *Quarterly Journal of Economics* **112**(2): 407–42

Hsee C K 1996 The evaluability hypothesis: An explanation of preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes* **67**: 247–57

Johnson E J, Hershey J, Meszaros J, Kunreuther H 1993 Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty* **7**: 35–51

Kachelmeier S J, Shehata M 1992 Examining risk preferences under high monetary incentives: Experimental evidence from the People's Republic of China. *American Economic Review* **82**: 1120–41

Kahneman D 1994 New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics* **150**(1): 18–36

Kahneman D, Knetsch J L, Thaler R 1990 Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* **98**(6): 1325–48

Kahneman D, Ritov I 1994 Determinants of stated willingness to pay for public goods: A study in the headline method. *Journal of Risk and Uncertainty* **9**: 5–38

Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263–91

Knetsch J L 1989 The endowment effect and evidence of nonreversible indifference curves. *American Economic Review* **79**: 1277–84

Lichtenstein S, Slovic P 1973 Response-induced reversals of preference in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology* **101**: 16–20

Loewenstein G, Elster J (eds.) 1992 *Choice Over Time*. Russell Sage Foundation, New York

McNeil B J, Pauker S G, Sox H C, Tversky A 1982 On the elicitation of preferences for alternative therapies. *New England Journal of Medicine* **306**: 1259–62

Redelmeier D, Shafir E 1995 Medical decision making in situations that offer multiple alternatives. *Journal of the American Medical Association* **273**(4): 302–5

Samuelson W, Zeckhauser R 1988 Status quo bias in decision making. *Journal of Risk and Uncertainty* **1**: 7–59

Shafir E 1993 Choosing versus rejecting: Why some options are both better and worse than others. *Memory and Cognition* **21**(4): 546–56

Shafir E, Simonson I, Tversky A 1993 Reason-based choice. *Cognition* **49**(2): 11–36

Slovic P, Lichtenstein S 1983 Preference reversals: A broader perspective. *American Economic Review* **73**: 596–605

Tetlock P E 1992 The impact of accountability on judgment and choice: Toward a social contingency model. *Advances in Experimental Social Psychology* **25**: 331–76

Tversky A, Kahneman D 1986 Rational choice and the framing of decisions. *Journal of Business* **59**(4): 251–78

Tversky A, Kahneman D 1991 Loss aversion in riskless choice: A reference dependent model. *Quarterly Journal of Economics* **106**: 1039–61

Tversky A, Sattath S, Slovic P 1988 Contingent weighting in judgment and choice. *Psychological Review* **95**(3): 371–84

Tversky A, Shafir E 1992 Choice under conflict: The dynamics of deferred decision. *Psychological Science* **3**(6): 358–61

von Neumann J, Morgenstern O 1947 *Theory of Games and Economic Behavior*, 2nd edn. Princeton University Press, Princeton, NJ

E. Shafir

Decision Making (Naturalistic), Psychology of

Naturalistic decision-making (NDM) is the process by which people use their experience to make decisions in complex real-world environments. Many of these environments involve high levels of risk and time pressure, are dynamically changing, and have significant consequences for both the decision maker and others. They include domains such as firefighting, aviation, medicine, and nuclear power operations. Knowledge gained from experience on the part of the decision maker plays a key role in the decision process, because that knowledge is the basis for recognizing situations that require decisions to be made, in determining what information is relevant to the decision, and in deciding what would be an appropriate course of action.

Naturalistic decision-making is a new field. The first studies were conducted in 1985 by Klein (1998), who examined decision-making by urban firefighters. This work was part of a new program initiated by the Army Research Institute because of the Army's dissatisfaction with decision aids that had been developed on the basis of traditional decision theories. The dominant decision theories at that point grew out of economic models and emphasized rational choice to select the *optimal* decision option. Decision aids based on these models were laborious, requiring military

commanders to make judgments of probability and value associated with a set of options. This process was incompatible with the thought processes typically used by experienced commanders and so was rejected. The Army decided that there must be a better way.

The 'better way' was to examine what experts do naturally when they make effective decisions under challenging conditions and to determine the basis for the quality of their decisions. Traditional research has sought to determine why human decision makers typically fall short of optimality and fail to use fully rational strategies. In contrast, NDM researchers recognized that experts in many fields, such as nurses, pilots, and military commanders, usually make decisions that achieve their goals, even if they are not optimal from a traditional perspective. The focus of NDM research is on understanding how experts make good decisions under difficult conditions, an understanding that can then serve as the basis for training and aiding novices.

1. *How is Naturalistic Decision-making Research Conducted?*

Because of the NDM investigators' interest in experts' performance in real settings, the nature of the research settings and methods typically differ from those used in traditional decision research.

1.1 *Research Settings*

Harking back to its origins in military concerns, NDM research has been conducted in environments that share some common features that are quite distinct from the laboratory environment. They include:

1.1.1 Ambiguous cues. The need to make a decision is prompted by cues that signal that a problem exists, such as a fire, a malfunction in an aircraft system, or a deterioration of a patient's medical condition. These cues are sometimes ambiguous or subtle, or conflict with other cues, suggesting opposing courses of action. Thus, the nature of the problem may be difficult to determine, and expertise is needed to 'read' the cues and 'size up' the situation.

1.1.2 Dynamic conditions. Conditions faced by decision makers in natural environments often change over time, and the human operator must update his or her understanding of the situation in order to respond appropriately. As conditions shift, goals also may shift, sometimes from maintaining the status quo to coping with an emergency. For example, when an urban fire commander arrives at a fire, his initial goal may be to extinguish the fire. However, when it appears that the roof is about to collapse, his goal shifts to evacuating all the firefighters.

1.1.3 High risk. In most domains of concern to NDM researchers, the consequences of making a poor decision involve high costs, including possible loss of life or property. If an emergency room doctor misdiagnoses the presenting symptoms of an unconscious patient, the patient may die. Pilots' misjudgments about the severity of weather in the vicinity of their destination airport can result in a crash. Firefighters' failure to assess correctly the source and nature of a fire can result in application of the wrong fire suppressor, exacerbating rather than quelling a fire.

1.1.4 Time pressure. Many decisions in NDM settings must be made quickly before a negative consequence occurs. Fires rage, patient conditions deteriorate, enemy aircraft or missiles approach swiftly. From the initial appearance of cues that signal a problem to the onset of serious consequences may be a matter of minutes. Often there is no time to conduct a full rational analysis of all possible response options, their probabilities, and their outcomes, which can take considerable time and computational power.

1.1.5 Team and organizational contexts. Teams, rather than solo decision makers, are frequently the rule in NDM settings. While a single individual may have primary responsibility for the decision, other players provide critical information, alternative perspectives, and critiques of the decision. Teams usually function within an organizational context that establishes goals, norms, and procedures for operations, sets constraints on what is possible, and determines what counts as a good decision.

Because NDM researchers are concerned with real-world settings that embody many of the above features, the methods they use also differ considerably from those employed by traditional decision researchers working in the lab.

1.2 *Methods*

Naturalistic decision-making research tends to be conducted in the field rather than in the laboratory. Thus, the predominant methods involve observations

and interviews. Models of decision processes are built from the ground up, based on rich descriptions derived from multiple observations in field settings. Participants are interviewed about what they did and why. Laboratory experiments, on the other hand, typically involve top-down testing of hypotheses derived from theories. NDM research is young and has not yet been formalized to the point at which laboratory studies are common. However, there have been a number of studies to test predictions of NDM models.

A technique that has come to be associated with the NDM approach is called *cognitive task analysis*. This technique involves interviewing experts about critical decisions to find out what cues signaled the problem, what information was actually used, why particular options were considered and perhaps rejected, and why others were preferred. Highly structured interview protocols have been developed to assure reliability in procedures and responses. Several analyses of data collected using this method have shown it to be sound and reliable. Other approaches involve process tracing, which seeks to make explicit the knowledge used as experts work on a task and make decisions. In a team context, participants' communication can provide insight into how they shape the problem, what they think is important to attend to, their situation awareness, their plans and strategies, and often their rationales for choosing a course of action. These techniques may be applied to videotapes of individuals performing in real or highly realistic simulated situations, such as in high-fidelity aircraft simulators.

2. *The Recognition-primed Decision Model*

Perhaps the most significant accomplishment within the NDM approach has been the development of a new decision process model. Based on his extensive investigations of fire fighters in both urban and wildfire situations, of neonatal nurses, and of military personnel, Klein (1998) proposed a model known as the Recognition-Primed Decision model (RPD). This model emphasizes the importance of recognizing cues in a situation that signal a particular type of problem. This recognition triggers retrieval of a response that has been associated in the past with a similar cue pattern and has led to a successful resolution.

In contrast, traditional analytic decision theories would require the decision maker to generate all possible options in that situation and then to evaluate each option in a concurrent manner, selecting the best option among them. In the recognition-primed model, the decision maker retrieves only one response option at a time and evaluates how well it satisfies the current goals. If the option is inadequate, then another option is retrieved and evaluated. Thus, options are considered in a serial rather than a concurrent manner. Note that many possible options are ignored in this strategy. It reflects what Herbert Simon called 'satisficing,' or choosing the first option that satisfies one's goals and existing constraints, rather than searching for the best of all possible options.

The RPD model highlights several other findings that distinguish the naturalistic approach.

2.1 *Importance of Situation Assessment*

The initial process involved in the RPD model is cue recognition and interpretation. A major distinction between naturalistic and traditional analytic decision approaches is that the NDM approach includes both a situation assessment process and choice of a response option, whereas traditional analytic approaches focus exclusively on the second component. In real-world situations, it is critical to recognize and understand the cues that signal a problem before choosing a response. If a problem has been assessed incorrectly, it is unlikely that an appropriate decision will be made. For example, pilots with an engine malfunction mistakenly thought the right engine was failing and shut it down. In fact, it was the left engine that had failed, and the plane crashed. In traditional laboratory studies, the situation assessment component of the process typically is minimal: the participant simply responds to the problem as designed and presented by the experimenter. A fixed set of options is presented from which the participant makes a choice.

2.2 *Basis for Evaluating Options*

Once the situation is understood and a response option has been retrieved, the option needs to be evaluated. In traditional analytic models, mathematical computation is used to combine preferences and probabilities associated with each option to choose the option that is mathematically optimal. For example, if you are buying a car, you may choose from many models. A variety of factors may influence your choice, e.g., cost, size, make, fuel consumption, and convenience features such as a sunroof. You need to decide how important each feature is to you and then evaluate how each car stacks up on each feature. The car you select presumably maximizes the combined values across all of the features. In naturalistic situations, experts typically use a different and less cumbersome process. The retrieved option is evaluated to determine if it will satisfy the decision maker's current goal. If it does, it is chosen. If it does not, another option is retrieved and evaluated.

The basis for option evaluation in the NDM approach is mental simulation–the decision maker projects forward what is likely to happen if the candidate option is adopted. This includes identifying difficulties that may be encountered, likely results, and

future decisions that may be required. If no significant impediments are foreseen, then the option is accepted. If limitations are found, a new option is retrieved or the original one is modified to get around the obstacle.

2.3 Role of Knowledge

The findings just described reflect the importance of the decision maker's knowledge to the decision-making process. In traditional laboratory studies of decision-making and other cognitive processes such as memory, knowledge has been treated as an unwanted source of variability. As such, efforts were made to design pared-down tasks that bypassed whatever knowledge the participants might have, thus eliminating the possibility that knowledge could give one participant an advantage over another. Hence, early studies of memory involved nonsense syllables or words such as *ang* or *bwap*. In naturalistic decision studies, knowledge is seen as the basis for making a decision, so effort is directed toward assessing participants' knowledge and how that knowledge is used.

For example, knowledge is critical to situation assessment. Early studies have shown that performance differences between experts and novices can be traced to differences in their ability to recognize cue patterns, but not to differences in general problem solving or decision-making skills. For instance, when chess pieces on a board form a realistic pattern from a game, chess experts can remember the location of the pieces much better than novices following only a brief inspection. However, when the same pieces are randomly placed on the chessboard, the experts perform no better than novices. This illustrates that pattern recognition reflects experience with the game; it provides the expert with a basis for making more effective decisions about the next moves.

Likewise, knowledge plays a critical role in retrieval of appropriate options. One might question the serial option-evaluation strategy described above if the decision maker were randomly retrieving and evaluating all possible options. In fact, studies have shown that experts typically retrieve highly appropriate options first, based on past success in similar situations. For example, the quality of chess moves selected by both chess masters and intermediate players during a game was assessed by grandmasters. The players were asked to indicate the moves they thought of first, as well as subsequent moves they considered. More experienced players consistently thought of good moves as their first ones; subsequent moves were typically worse. This was not the case for novices, whose order of retrieval was less systematic.

Knowledge also serves as the basis for experts' evaluation of options. Without knowledge of outcomes based on prior experience, one cannot know what outcomes are likely. Studies have documented that decision strategies used by experts in a field are quite different from those used by novices. Klein found that relatively inexperienced tank commanders frequently used analytic concurrent option-evaluation strategies, whereas more experienced commanders more often used a recognition-primed sequential option-analysis strategy. There is one exception to this finding, however: experts may revert to more general-purpose strategies if a situation is unfamiliar and they lack specific knowledge relevant to the decision.

3. Value of the Naturalistic Approach

One may ask whether the recognition-primed decision model is appropriate to all circumstances or is limited to special cases. Basically, it is appropriate in any domain in which the decision maker has considerable experience, when time is limited and conditions are changing dynamically, and when problems are ill defined and goals are shifting, conditions that make an analytic approach difficult or impossible. An RPD approach represents an efficient use of knowledge and experience. It is less effortful and time consuming than most analytic approaches.

Analytic approaches are useful under certain conditions. These include situations in which the decision maker has little knowledge pertaining to the decision domain. Adequate time must be available in order to do a thorough evaluation of many options, when selecting the optimal solution is important. Analytic approaches are appropriate when conditions are stable, the parameters of the problem are clear, and all options are known. Analytic approaches yield a reliable, thorough outcome that can be justified to participants who may have competing goals and interests in the problem. For example, analytic approaches have been used to make decisions about where to site nuclear power plants. Analytic approaches are general-purpose tools that anyone can use when conditions permit.

4. Relations to Other Research Disciplines

Naturalistic decision researchers have borrowed techniques from several other disciplines to conduct their studies. They are closely aligned with much European research on work analysis, process tracing, and ecological psychology. Observational techniques have been drawn from ethnography, along with interview techniques for understanding the meaning of actions in context. Efforts to understand the role of expertise are grounded in cognitive psychology studies of experts, which involve verbal protocol analysis and process tracing.

In some cases the goals are similar but the approaches are different across related disciplines. For

example, researchers from both the NDM and the judgment and decision-making (JDM) communities are interested in determining what cues are used to make decisions and their relative importance. While JDM researchers seek to determine cue weights, NDM researchers seek to know how the cues support situation assessment.

5. Methodological Issues

Because they deviate from traditional laboratory methods of science, NDM efforts have been challenged on several grounds. Most notable has been a challenge to the scientific rigor of field-based methods, namely, observations and interviews. NDM researchers maintain that a thorough description of the phenomena *in situ* is required in order to understand the role of contextual factors in decision processes and how knowledge is used. Field studies support the discovery phase of science in which new phenomena are identified and documented. Laboratory experiments may be suitable at a later point when specific hypotheses or model tests are desired.

Second, NDM research has not yet yielded formal models. Investigations are continuing in many domains and local models have been developed, but a formal general process model has not yet emerged. In part this reflects the newness of the field; in part it reflects the goals and interests of the investigators. Formal models typically are developed so that studies can be done to disconfirm the models and advance the theory. At the present stage, NDM researchers are still trying to understand the nature of phenomena of interest.

Third, one might question the use of expert models as the basis for prescribing training and decision aids. Experts do not always agree and experts are fallible, so how can models based on experts be used as a basis for training novices? First, the expert models are generalized, not based on a single expert. Second, NDM researchers emphasize the importance of building thorough domain knowledge in order to make better decisions, rather than training specific expert strategies. Training of novices is most often done via context-based scenario training, which emphasizes that (a) good decisions cannot be made on the basis of incomplete or incorrect situation assessment, and (b) option evaluation must take into account future projections, present constraints, and potential pitfalls, both of which depend on domain knowledge.

6. Future Directions in Theory and Research

At the beginning of the twenty-first century, efforts in the NDM community are moving in two directions. One is to formalize the theory so that predictions may be made and tested. The other is to expand the scope to explore issues related to adaptive decision-making and critical thinking. These include study of problem identification and structuring of decision problems, especially in dynamic conditions when goals shift and conflict. Ill-structured and unfamiliar problems may require analogical reasoning or creative problem solving, processes that go beyond the RPD approach. Finally, the roles of organizational and other contextual features on decision processes need to be better understood.

7. Relevant Readings

To achieve a fuller understanding of the origins and development of the field of naturalistic decision-making, the reader is directed to the volumes resulting from a series of workshops that have been held since NDM research began. The first workshop was held in 1989 and yielded the first book on naturalistic decision-making (Klein et al. 1993). Since then, four more workshops have been held. Each proceedings volume includes a critique by a researcher representing a traditional decision-making approach. Klein (1998) is a popular book that provides a deeper understanding of NDM and its practical applications, and Flin (1996) describes decision-making challenges to incident managers in high-risk industries like offshore oil production.

See also: Decision Biases, Cognitive Psychology of; Decision Making: Nonrational Theories; Decision Making, Psychology of; Decision Research: Behavioral; Heuristics for Decision and Choice

Bibliography

Flin R 1996 *Sitting in the Hot Seat*. Wiley, New York
Flin R, Salas E, Strub M, Martin L (eds.) 1997 *Decision Making Under Stress: Emerging Themes and Applications*. Ashgate, Aldershot, UK
Friman H (ed.) 2000 *How Professionals Make Decisions*. CD-ROM. National Defense College, Stockholm
Klein G A 1998 *Sources of Power*. MIT Press, Cambridge, MA
Klein G E, Orasanu J, Calderwood R, Zsambok C E (eds.) 1993 *Decision Making in Action: Models and Methods*. Ablex, Norwood, NJ
Zsambok C, Klein G (eds.) 1997 *Naturalistic Decision Making*. Erlbaum, Mahwah, NJ

J. M. Orasanu

Decision Making: Nonrational Theories

The term 'nonrational' denotes a heterogeneous class of theories of decision making designed to overcome problems with traditional 'rational' theories. Nonrational theories have been denoted by various terms,

including models of bounded rationality, procedural rationality, and satisficing. Although there is as yet no agreed-upon definition of 'nonrational,' rational and nonrational theories typically differ on several dimensions, discussed below. The term 'decision making' is used broadly here to include preference, inference, classification, and judgment, whether conscious or unconscious.

Nonrational theories of decision making should not be confused with theories of irrational decision making. The label 'nonrational' signifies a type of theory, not a type of outcome. In other words, the fact that nonrational theories postulate agents with emotions, limited knowledge, and little time—rather than postulating omniscient 'rational' beings—need not imply that such agents fare badly in the real world.

1. *Historical Background*

A few historical remarks on rational theories help to set the stage for nonrational theories. In the mid-seventeenth century, the calculus of probability replaced the ideals of certain knowledge and demonstrative proof (as in mathematics and religion) with a more modest vision of reasonableness. What may be called the first rational theory of decision making, the maximization of expected value, emerged at this time. According to the theory, an option's expected value is the sum of the product of the probability and the value of each of its consequences; a rational decision maker chooses the option with the highest expected value.

The notion of defining an option's reasonableness in terms of its expected value soon ran into problems, because in some situations (e.g., the St. Petersburg problem), its prescriptions conflicted with educated intuition. Mathematician Daniel Bernoulli therefore proposed to replace the concept of expected value with that of expected utility. For instance, the utility of a monetary gain (say, of $1,000) can be defined as a logarithmic function of its dollar value and the agent's current wealth, assuming that the utility of an additional dollar diminishes as the value of the gain and current wealth increase.

The fact that rational decision making can be defined in more than one way—for example, as maximization of expected value or expected utility—has been interpreted both as the weakness and the strength of the program. This ambiguity was one of the reasons why, by 1840, most mathematicians had given up attempting to define a calculus of reasonableness (Daston 1988). With a few exceptions, rational theories of decision making largely disappeared until their revival in the 1950s and 1960s. Only then did the major species of rational theories, the maximization of subjective expected utility and Bayesianism, become influential in the social and behavioral sciences. At about the same time, some psychologists and economists—most notably Nobel laureates Herbert Simon and Reinhard Selten—criticized the assumptions about the human decision maker in modern rational theories as empirically unfounded and psychologically unrealistic, and called for alternative theories.

2. *Optimizing vs. Nonoptimizing Theories*

Rational theories rest on the ideal of optimization; nonrational theories do not. Optimization means the calculation of the maximum (or minimum) of some variable across a number of alternatives or values. For instance, according to a rational theory known as subjective expected utility (SEU) theory, an agent should choose between alternatives (e.g., houses, spouses) by determining all possible consequences of selecting each alternative, estimating the subjective probability and the utility of each consequence, multiplying the probability by the utility, and summing the resulting terms to obtain that alternative's subjective expected utility. Once this computation has been performed for each alternative, the agent chooses the alternative with the highest expected utility. This 'subjective' interpretation of SEU has been used to instruct people in making rational choices, but was criticized by decision theorists who argue that preferences are not derived from utilities, but utilities from preferences. In this 'behavioristic' interpretation, no claims are made about the existence of utilities in human minds; SEU is only an attempt to describe the choice. People choose *as if* they are maximizing SEU (see Sect. 3).

Nonrational theories dispense with the ideal of optimization. For instance, Simon (e.g., 1956, 1982) proposed a nonrational theory known as *satisficing*, in which an agent is characterized by an aspiration level and chooses the first alternative that meets or exceeds this aspiration level. The aspiration level (e.g., characterization of what would constitute a 'good-enough' house) allows the agent to make a decision without evaluating all the alternatives.

There are several motives for abandoning the ideal of optimization. First, in many real-world situations, no optimizing strategy is known. Even in a game such as chess, which has only a few stable, well-defined rules, no optimal strategy exists that can be computed by a human or a machine. Second, even when an optimizing strategy exists, it may demand unrealistic amounts of knowledge about alternatives and consequences, particularly when the problem is novel and time is scarce. Acquiring the requisite knowledge can conflict with goals such as making a decision quickly; in situations of immediate danger, attempting to optimize can even be deadly. In social and political situations, making a decision can be more important than searching for the best option. Third, strategies

that do not involve optimization can sometimes outperform strategies that attempt to optimize. In other words, the concept of an optimizing strategy needs to be distinguished from the concept of an optimal outcome. In the real world, there is no guarantee that optimization will result in the optimal outcome. One reason is that optimization models are built on simplifying assumptions that may or may not actually hold. An example of a nonoptimizing strategy that performs well in the repeated prisoner's dilemma is 'tit for tat,' a simple heuristic that cooperates on the first move and thereafter relies on imitation, cooperating if the partner cooperated and defecting if the partner defected on the previous move. In the finitely repeated prisoner's dilemma, two tit-for-tat players can make more money than two rational players (who reason by 'backward induction' and therefore always defect), although tit for tat only requires remembering the partner's last move and does not involve optimization.

3. *Normative vs. Descriptive Theories*

Nonrational theories are descriptive, whereas rational theories are normative—this common distinction is half-true. Indeed, nonrational theories are concerned with psychological plausibility, that is, the capacities and limitations of actual humans, whereas rational theories have little concern for descriptive validity and tend to assume omniscience. But nonrational theories have sometimes been interpreted as normative as well. For instance, if an optimization strategy is nonexistent, unknown, or dangerous to perform because it would slow decision making, a simple heuristic— such as copying the behavior of others—can be the best decision-making strategy.

Rational theories typically do not assume that agents actually perform optimization or have the knowledge needed to do so. Their purpose is not to describe the reasoning process, but to answer a normative question: what would be the best strategy for an omniscient being to adopt? In economics, psychology, and animal behavior, however, the answer to this question is sometimes used to predict actual behavior. In this way, a rational theory can be descriptive of behavioral outcomes yet mute about underlying processes. For instance, optimal foraging theory assumes that animals select and shift between food patches *as if* they had perfect knowledge about the distribution of food, competitors, and other relevant information, without claiming that real animals have this knowledge or perform optimizing computations. Instead, it is assumed that animal behavior has evolved to be close to optimal in specific environments. The question of what proximal mechanisms produce this behavior is a different one. These mechanisms may be heuristics, habits, or forms of social imitation that are the topic of theories of nonrational decision making.

To summarize, nonrational theories aim to describe both the process and the outcome of decision making. In certain situations, they can be seen as characterizing the best an organism with limited time and knowledge can do. Rational theories are primarily normative, although often for omniscient rather than real beings. They are often seen as descriptive in the sense of predicting behavior, but not as models of underlying processes.

4. *Search vs. Omniscience*

Search is a central part of nonrational theories. In many rational theories, in contrast, search plays no role; it is instead assumed that all relevant information is already available to the agent (hence the agent's omniscience). Actual humans, however, have to search for information, either in memory or in outside sources, such as on the Internet or in encyclopedias. Information search can cost time and money.

One class of rational theories, known as 'optimization under constraints,' models limited search but retains the ideal of optimization. These theories posit an optimal stopping rule that requires the organism to stop search when the costs of further search exceed its benefits. This rule has been criticized because it can lead to infinite regress: To compute the optimal trade-off between the costs and benefits of search itself carries costs, which must be factored in to a 'meta' analysis of the costs and benefits of computing the costs and benefits of search, and so on. A second point of criticism is that accurate estimates of benefits and costs—such as opportunity costs, incurred by foregoing the benefits of activities other than continuing search—can be troublesome to obtain and may even demand knowledge close to omniscience. Therefore, optimization under constraints can lead to models that are descriptively even more unrealistic than rational theories that ignore search.

Nonrational theories model search with simple stopping rules rather than optimal stopping rules. Search can concern either of two kinds of information, *alternatives* (such as houses and spouses) or *cues* (such as reasons to choose a given house). Two different classes of nonrational theories deal with these types of search: aspiration level theories with the search for alternatives and fast and frugal heuristics with the search for cues.

4.1 *Aspiration Level Theories*

Aspiration level theories assume that an agent has an aspiration level, which is either a value on a goal variable (e.g., profit or market share) or, in the case of

multiple goals, a vector of goal values that is satisfactory to the agent. When choosing among a large (possibly even infinite) set of alternatives, agents search until they find the first alternative that meets or surpasses their aspiration level, at which point search stops and that alternative is chosen. For instance, agents might set a lower limit on the price at which they would be willing to sell their shares in a company (the aspiration level). In this satisficing model, the agent makes no attempt to calculate an optimal stopping point, in this case, the best day on which to sell. The aspiration level need not be fixed, but can be dynamically adjusted to feedback. For instance, if the investor observes that the share price is monotonically increasing rather than fluctuating over time, they might conclude that there is some stable trend and adjust the limit upward. Thus, aspiration level theories model decision making as a dynamic process in which alternatives are encountered sequentially and aspiration levels stop search. The challenge is to understand where aspiration levels come from in the first place (Simon 1982, Selten 1998).

4.2 *Fast and Frugal Heuristics*

In a different class of problems, the set of alternatives is given (i.e., need not be searched for), and the agent needs to search for cues that indicate which alternative to choose. For instance, an employer might want to decide which of three job applicants to hire or a bettor to predict which of two soccer teams will win a game. *Fast and frugal heuristics* employ simple stopping rules to make such inferences with little computation ('fast') and information ('frugally'). For instance, the 'take the best' heuristic bases its inference solely on the best cue on which the alternatives differ and ignores the rest. Such 'one-reason' decision making allows agents to make decisions quickly and, counterintuitively, often as or more accurately than the 'optimal' linear model (multiple regression), which looks up and integrates all available cues. Other types of fast and frugal heuristics include ignorance-based decision making (see below) and elimination heuristics. Thus, fast and frugal heuristics model decision making as a dynamic process in which cues or reasons are sequentially searched for—in memory or in the outside world—and simple stopping and decision rules determine inferences. The challenge here is to understand what the class of heuristics is, how a heuristic is selected, and in which environments it is successful (Gigerenzer et al. 1999).

5. *Ecological Rationality vs. Internal Consistency*

A classical criterion for rational choice is internal consistency or coherence. Numerous rules of consistency have been formulated, for instance, transitivity and additivity of probabilities. These rules, which are the building blocks of rational theories, have been used to test the limits of human rationality, beginning with the work of Jean Piaget and Bärbel Inhelder. Nonrational theories, in contrast, place less weight on internal consistency; for instance, some fast and frugal heuristics can violate transitivity. Instead, nonrational theories emphasize performance in the external world, both physical and social. Measures of external performance include the accuracy, speed, frugality, cost, transparency, and justifiability of decision making. Note that internal consistency does not guarantee that any of these external criteria are met. For instance, the statement 'there is a 0.01 probability that cigarette smoking causes lung cancer and a 0.99 probability that it does not' is internally consistent in that the probabilities add up to 1, but according to relevant research, it is not accurate.

How can heuristics be simple and accurate at the same time? Two major answers have been proposed: They can exploit the structure of environments, and simplicity can entail robustness.

5.1 *Structure of Environments*

The term 'ecological rationality' refers to the match between a heuristic and the structure of the information in a particular environment. The more closely a heuristic reflects important aspects of this structure, the more likely it is to succeed in that environment. Simple heuristics can succeed by exploiting the structure of information in an environment. In other words, the environment can do part of the work for the heuristic. For instance, consider the problem of predicting which of two soccer teams will win a game, which of two cities is larger, or which of two colleges provides the better education. Assume a fairly ignorant agent who has heard of only one of the two teams, cities, or colleges. He can use the 'recognition' heuristic, which infers that the recognized object will win the game, have the larger population, or provide the better education (Gigerenzer et al. 1999). Such ignorance-based decision making works well in environments where ignorance (e.g., lack of name recognition) is not random but systematic, as in competitive environments where the sequence in which the names of objects are first encountered is correlated with their performance, power, or size. The structure of such environments does part of the work in the sense that it allows the recognition heuristic to glean information from ignorance. If the correlation between recognition and the criterion is sufficiently large, a counterintuitive result is observed: Less knowledge leads to more accurate predictions than more knowledge, because people who recognize both alternatives cannot use the recognition heuristic.

Cricket, baseball, and other sports in which players need to catch a ball provide a second illustration of how heuristics can exploit the structure of environments. Suppose that we want to build a robot that can

catch a ball—a thought experiment because none exists so far. One approach to building robots in artificial intelligence is to endow them with a complete representation of the environment and a supercomputer that can draw inferences from this information. Taking this approach, which is consistent with rational theories, one would feed the robot parabolic functions describing all the trajectories that the ball might follow, and equip it with instruments for measuring the ball's initial velocity and angle to select the right parabola. Further instruments would be needed to measure the speed and direction of the wind at each point of the ball's flight, as well as myriad other factors, such as spin, in order to calculate the ball's deviation from the parabolic course. These measurements would be analyzed by a computer that would then compute where the ball will land. Professional athletes, in contrast to this hypothetical robot, seem to use a simple heuristic for catching (McLeod and Dienes 1996). They start running with their eyes fixed on the ball and adjust their speed to keep their angle of gaze constant (or within a certain range). Using this heuristic, a robot can intercept the ball without searching for the information that a 'rational' robot requires. It pays attention to only one variable—the angle of gaze, which does all the work. The 'heuristic' robot does not calculate where the ball will land, but it will be there when it does.

These examples illustrate how heuristics can succeed by exploiting structures of environments. These heuristics are 'domain-specific' rather than 'domain-general,' that is, they work in a class of environments in which they are ecologically rational. Heuristics do not provide a universal rational calculus, but a set of domain-specific mechanisms similar to the parts of a Swiss army knife, and have been referred to collectively as the 'adaptive toolbox' (Gigerenzer and Selten 2001).

5.2 Robustness

A second reason why a simple heuristic can make accurate predictions is robustness. To understand this point, it is necessary to distinguish between data fitting (i.e., determining the best-fitting parameter values for a model given a specific body of data) and prediction (i.e., using these parameter values to predict new data). For data fitting, it generally holds that the more parameters one uses in a model, the better the fit; for prediction; however, there can be a point where less is more (Foster and Sober 1994). For instance, if one records the air temperature on all 365 days in a year, one can fit the resulting jagged curve increasingly well as one adds more free parameters to the model. However, if one wants to predict air temperature during the coming year, the model that best fits the past data may result in less accurate predictions than a simpler model with fewer parameters and a worse fit. Similarly, one can fit one ball's trajectory through the air to arbitrary degrees of precision, but this may be of little help in predicting the next ball's trajectory.

More generally, in noisy environments only part of the available information generalizes to the future. The art of building a good model is to find this part and to ignore the rest. The more noise in the environment, the more that models with many free parameters tend to 'overfit,' that is, to reflect the noise. Overfitting can become a problem when overly powerful mathematical models, such as neural networks with numerous hidden units and multiple regression with many predictors, are used to fit and then predict behavioral data (Geman et al. 1992). Simplicity can reduce overfitting and thereby produce robust decision strategies. An alternative route to robustness is to use the computational power of modern computers to search through large numbers of models to find one that is robust by a given criteria. But one often does not have the time and knowledge to proceed this way.

Heuristics can be fast, frugal, and accurate by exploiting the structure of information in environments and by being robust. But these are not the only reasons why strategies that work with limited knowledge and time can be successful. For instance, there is evidence that language learning fails to occur in fully formed neural networks, but successfully develops in networks that begin with limited working memory and gradually mature. This finding has been interpreted to mean that young children's memory limitations may be an important precondition for rapid language acquisition (Elman 1993). Thus, cognitive limitations might not always be regrettable deficits, but actually enable learning by starting small.

6. Emotions, Imitation, and Social Norms

Like rational theories, most nonrational theories rely on cognitive building blocks, such as aspiration levels, search heuristics, and stopping rules. However, *Homo sapiens* is not only the most intelligent, but also the most emotional and social species—one of the very few in which unrelated members of the same species cooperate. Theories of decision making have often neglected emotions, and sometimes even cast them as the very opposite of rationality. However, emotions can aid decisions making just as fast and frugal cognitive heuristics do. For instance, falling in love can be seen as a powerful stopping rule that ends search for a partner and strengthens commitment to the loved one. This emotion guarantees commitment more effectively than would a nonemotional mind that tried to optimize partner quality and dropped the current partner each time a more promising one came along. Similarly, feelings of parental love, triggered by the presence or smile of one's infant, can prevent cost–benefit computations, so that the question of whether it is worthwhile to endure all the sleepless nights and other challenges of baby care simply never arises. Instead, love keeps parents focused on the

adaptive task of protecting and providing for their offspring. To take another emotion as an example, disgust can limit the choice set of potential foods and help one to avoid becoming ill from spoiled food. Finally, emotions such as fear and anger can speed decision making to the point that there is no decision to be made.

Like emotions, social norms and social imitation can function as decision-making guidelines that keep individual learning and information search to a minimum (Gigerenzer and Selten 2001). Social heuristics such as 'eat what other conspecifics eat' or 'prefer mates preferred by others' can guide behavior without much information gathering and bring benefits such as a reduced likelihood of food poisoning and social disapproval. These forms of social rationality can be found throughout the animal world. For instance, female guppies prefer the mates that other females prefer, and such social imitation can even reverse their prior preferences for one male over another. In humans, media-fueled mate preferences, and occasionally even academic hiring, follow similar heuristics, such as 'If they want him, then we want him!' Custom, not optimization, governs much of life, even in the economic and intellectual worlds.

Social systems foster not only individual but distributed intelligence. That is, by cooperating with one another, myopic individuals can exhibit collective rationality. Communities of social insects are one example of such intelligent 'superorganisms,' as is division of labor in human industry and politics. The intelligence of a superorganism can be seen as the emergent property of a few adaptive heuristics of its members. Honey bees, for instance, make intelligent collective decisions about where to build a new hive that seem to emerge from individual bees' application of a few simple, well-adapted heuristics. Complex phenomena need not to be modeled in terms of complex knowledge and computation.

7. *Summary*

Nonrational theories take account of what we know about humans' and other species' capacities rather than assuming unlimited knowledge, memory, time, and other resources. They model heuristics—cognitive, emotional, and social—that exploit the structure of information in real environments. Nonrational theories conflict with the ideal of *Homo economicus* and other visions of humans developed in the image of an omniscient God, but provide us with a more realistic picture of decision making when knowledge is scarce, deadlines are rapidly approaching, and the future is hard to predict.

See also: Bayesian Theory: History of Applications; Bounded and Costly Rationality; Bounded Rationality; Decision and Choice: Random Utility Models of Choice and Response Time; Decision Biases, Cognitive Psychology of; Decision Theory: Bayesian; Heuristics for Decision and Choice; Luce's Choice Axiom; Rational Choice Explanation: Philosophical Aspects; Subjective Probability Judgments; Utility and Subjective Probability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies

Bibliography

Anderson J R 1990 *The Adaptive Character of Thought*. L. Erlbaum Associates, Hillsdale, NJ

Braitenberg V 1984 *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge, MA

Daston L 1988 *Classical Probability in the Enlightenment*. Princeton University Press, Princeton, NJ

Dawes R M 1979 The robust beauty of improper linear models in decision making. *American Psychologist* **34**: 571–82

Elman J L 1993 Learning and development in neural networks: The importance of starting small. *Cognition* **48**: 71–99

Foster M, Sober E 1994 How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal of Philosophical Science* **45**: 1–35

Geman S, Bienenstock E, Doursat E 1992 Neural networks and the bias/variance dilemma. *Neural Computation* **4**: 1–58

Gigerenzer G, Selten R (eds.) 2001 *Bounded Rationality: The Adaptive Toolbox*. MIT Press, Cambridge, MA

Gigerenzer G, Todd P M, ABC Research Group 1999 *Simple Heuristics That Make us Smart*. Oxford University Press, New York

McLeod P, Dienes Z 1996 Do fielders know where to go to catch the ball or only how to get there? *Journal of Experimental Psychology: Human Perception and Performance* **22**: 531–43

Piaget J, Inhelder B [1951] 1975 *The Origin of the Idea of Chance in Children*. Norton, New York

Rivest R J 1987 Learning decision lists. *Machine Learning* **2**: 229–46

Selten R 1998 Aspiration adaptation theory. *Journal of Mathematical Psychology* **42**: 191–214

Selten R, Buchta J 1999 Experimental scaled bid first price auctions with directly observed bid functions. In: Budescu D, Erev I, Zwick R (eds.) *Games and Human Behavior: Essays in Honor of Amnon Rappoport*. L. Erlbaum Associates, Mahwah, NJ, pp. 79–102

Simon H A 1956 Rational choice and the structure of environments. *Psychological Review* **63**: 129–38

Simon H A 1982 *Models of Bounded Rationality*. MIT Press, Cambridge, MA

Wimsatt W C in press *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press, Cambridge, MA

G. Gigerenzer

Decision Making, Psychology of

Human decision making has been studied by a variety of disciplines including economics, philosophy, psychology and statistics. Behavioral decision making, as the field is generally known in psychology, is being

studied in all branches of psychology. The starting point of much work on human judgment and decision making is rational choice theory. The most influential explanatory concept in the analysis of decisions under uncertainty is the subjective expected utility (SEU) of available alternatives. SEU theory is a normative theory and is often compared to actual decisions. Deviations that violate the rational principles underlying SEU-theory systematically are termed anomalies. Anomalies of decision making will be briefly discussed in this section, followed by descriptive theories of decision making, and a brief discussion on how to improve the quality of decision making.

1. Normative Theories of Decision Making

Traditionally choices are assumed to be consequential, and options are evaluated using prior assessments of beliefs and values. Generally, tasks in this field of research therefore have a probability component or a value component, or both.

Edwards (1954) provided the first major review of research on human judgment and decision making. He argued that normative and prescriptive models based on economic and statistical theory should also be relevant to psychologists interested in human judgment and decision making. Edwards introduced SEU theory, which decomposes decisions or choices in probabilities and preferences, and provides a set of rules for combining beliefs (probabilities) and preferences ('values' or 'utilities'). SEU maximization is considered an appropriate normative rule for decision making under risk or uncertainty. The theory is normative or prescriptive because it specifies how decisions should be made. If one accepts the axioms upon which it is based, then the most rational choice or decision is the one specified by the theory as having the highest subjectively expected utility.

The axioms of SEU theory are all based on the work of Von Neumann and Morgenstern (1944). Their work showed that if a person's choices follow certain rules or axioms, it is possible to derive utilities in such a way that a preference for a specific alternative implies that its expected utility is greater than that of the other alternative(s). Their approach assumes that it is possible to associate a real number with each possible consequence of the relevant alternatives. This number (e.g., a number between one and 10) represents the utility of that consequence for the individual. The expected utility of a specific alternative is then the sum of the numbers associated with each possible consequence weighted by the probability that each consequence will occur. Thus a positive or negative outcome will be weighted by the probability of its occurrence; for example a certain outcome will be multiplied by 1.0, an outcome with a probability of 0.5 will be multiplied by 0.5, etc.

To illustrate this let us compare the two jobs in terms of three possible outcomes or consequences. These are initial salary, career opportunities and atmosphere. Initial salary is higher at firm A than at firm B, and receives ratings of eight and seven respectively on a scale ranging from one to 10. Career opportunities seem better at firm B but this element cannot be assessed with certainty. Let us assume that career opportunities get ratings of eight (firm A) and nine (firm B), but that you are less certain of your assessment of firm A (subjective probability of 0.7), than of firm B (subjective probability of 0.9). Finally, the atmosphere at firm B seemed to be better than at firm A; your scores are eight and seven, respectively. You spent quite some time at firm B and are certain about your judgment of the atmosphere (it seemed a great team of colleagues). You are less certain about your verdict of the atmosphere at firm A because you spent less time there (probabilities of 1.0 (B) vs. 0.9 (A)). This leads to the following scores: Firm A $(1.0*8+0.7*8+0.8*7=20.0)$, Firm B $(1.0*7+0.9*9+1.0*8)=23.2$. Thus, according to this analysis, the job offer of firm B is more attractive than that of firm A.

As argued by Dawes (1988) the central idea of the Von Neumann and Morgenstern theory is that it provides an important tool to reach a rational decision. SEU theory and related theories have dominated research on decision making since the mid-1940s, and especially the validity of theories such as SEU-theory as adequate descriptions of human choice and decision making has been a dominant theme in this research area. This is partly based on the idea that the study of human decision making should focus on the perceptual and cognitive factors that cause human judgment and decision making to *deviate* from the predictions of normative models such as SEU theory.

2. Violations and Anomalies

Research in the past decades has resulted in a wide variety of illustrations of systematic violations of principles of rationality. A clear example concerns the violation of the sure thing principle: If one prefers one option to another no matter what happens—or has happened—one should prefer that option without knowing what happens, or has happened. Tversky and Shafir (1992) gave an example of violating this principle. Their respondents were asked to consider a gamble with a 50/50 chance of winning $200 or loosing $100. Generally, a minority of about one third of respondents offered that gamble accept it. This preference pattern changes dramatically if respondents are asked to imagine they had already played the gamble and were now being offered the opportunity to play it a second time. In case they were asked to imagine they had won the first time, nearly 70 percent

wanted to play again. If they had been asked to imagine they had lost the first time nearly 60 percent wished to play it a second time.

Another violation is related to so-called sunk costs. It is Saturday afternoon and you are on the brink of leaving for the railway station in order to get to the stadium. You bought your ticket for the match some time ago. It is pouring with rain, you do not feel that well and standing in the rain for 2 hours is not going to help. In sum, you prefer to spend the afternoon at home. Nevertheless, if you decide not to go, you will have 'wasted' the money you paid for the ticket. The nonrefundable money already paid for the ticket is termed a sunk cost. Strictly speaking there appears to be no rational argument to do things you would prefer not to do in the interest of honoring the sunk cost. However, people often honor sunk costs. Examples include trivial anecdotes such as the one described above, but also large-scale enterprises such as the development of the Concorde passenger aircraft and the Gulf War.

Other violations concern principles such as transitivity. If people prefer option A over option B and also option B over option C, they should also prefer option A over option C. Violations of transitivity may arise from evaluating alternative options in terms of different dimensions. On some dimensions 'A' might be preferred over 'B' and 'C' while on others 'C' might be preferred over 'A.' If the salience of the dimensions varies, preferences could differ and one could violate transitivity. Another way in which transitivity may occur is through regret (see below).

The way in which alternative options are presented can also lead to anomalies. It turns out that people often make different choices as a function of the way alternatives are presented or framed, even though the alternatives remain unchanged. One example of framing is to present actions in terms of positive as opposed to negative consequences or vice versa. For example, a 50 percent chance of a successful outcome of surgery may appear to someone to be a more desirable outcome than a 50 percent chance of failure. Several studies provide support for this framing effect. Framing effects have been extensively investigated in decision problems in which the consequences of options are compared to psychologically salient reference points (e.g., a zero point, the status quo). Kahneman and Tversky (1979) found that when asked to choose between \$15,000 for sure or a 50/50 chance of \$10,000 or \$20,000 most people choose the certain option. If one has received the \$20,000 and was asked a second choice between returning \$5000 or a 50/50 chance of giving back nothing or \$10,000 most people choose the latter option. This pair of choices is contradictory. Similarly, when presented with a choice between a sure loss of \$100 or a 50/50 chance of a loss of \$200 and a 50/50 chance of no loss at all, most people will prefer the uncertain, risky alternative. This risk-seeking preference is in sharp contrast with preferences when the choice is between a sure gain of \$100 or a 50 percent chance of a gain of \$200 and a 50 percent choice of no gain at all. In this case most people will prefer the certain, risk-avoiding alternative. According to SEU-theory the two options in each problem are equally (un)attractive. Thus, inducing people to adopt a gain frame when choosing between a gamble and a sure thing tends to result in a risk-avoiding preference (the certain alternative). Inducing people to adopt a loss frame, on the other hand, tends to result in a risk-seeking preference (the risky alternative).

A final anomaly concerns violations of response invariance. Choice can also be influenced by the method by which preferences and choices are expressed. For instance, valuing something with money versus making a choice is frequently associated with preference reversals. Another example concerns the effect of choosing versus rejection of alternatives that have both bad and good characteristics. Shafir (1993) argues that a 'choose' orientation leads decision makers to search for good qualities of each option, while a 'reject' orientation makes them focus on bad qualities. Thus, if one alternative has more clear-cut positive and negative characteristics, and another alternative is relatively bland, the former could both be chosen and rejected depending on the question. Shafir (1993) demonstrated this in the context of awarding custody in which the question was either which parent to award custody or which parent to deny sole custody of the child.

3. *Heuristics*

The study of heuristics tends to focus on systematic errors in human decision making and these heuristics often help to understand anomalies of inferring expectations from evidence (see *Heuristics in Social Cognition*).

Three heuristics that deal with probabilistic thinking have received considerable attention: (a) availability, (b) representativeness, and (c) anchoring and adjustment. The availability heuristic refers to the tendency to assess the probability of an event based on the ease with which instances of that event come to mind. This heuristic has been investigated in a variety of domains and relates probability estimates to memory access. Generally people overestimate the probability of an event if concrete instances of that event are easily accessible in memory. Generally, ease of recall and frequency of occurrence are correlated. A number of factors that affect memory are, however, unrelated to probability. For example, vivid images are easier to recall than pallid ones. Thus, having been involved in a serious car accident is likely to be better remembered than annual statistics about the frequency of (types of) traffic accidents. The former is likely to have more impact on probability estimates than the latter. Dawes (1994) argued that the salience of negative and

relatively extreme exemplars of drug addicts can bias policy-makers' perceptions of the entire group and result in negative attitudes toward programs such as the provision of clean needles to prevent a further spread of the AIDS virus.

The representativeness heuristic refers to the tendency to assess the probability that a stimulus belongs to a particular class by judging the degree to which that event corresponds to an appropriate mental model. Kahneman and Tversky (1973) reported a well-known example of how ignoring prior probabilities can affect judgment. In their study, respondents were provided with brief personality sketches, supposedly of engineers and lawyers. They were asked to assess the probability that each sketch described a member of one profession or the other. Half the respondents were told the population from which the sketches were drawn consisted of 30 engineers and 70 lawyers, the remaining respondents were told that there were 70 engineers and 30 lawyers. Findings showed that the prior probabilities were essentially ignored, and that respondents estimated the probability of class membership by judging how similar each personality sketch was to their mental model of an engineer or a lawyer.

Anchoring and adjustment refers to a general judgment process in which an initially given or generated response serves as an anchor, and other information is insufficiently used to adjust that response. The anchoring and adjustment heuristic is based on the assumption that people often start their judgmental process by focusing on some initial value that serves as an anchor. The biases related to this heuristic stem from two distinct aspects. First, one could use irrelevant anchors, second one could insufficiently adjust up or down from an original starting value or anchor.

4. Descriptive Approaches to Decision Making

People often tend to solve problems in a way that satisfices (Simon 1955) in terms of avoiding computational complexity and/or in terms of having good reasons for one's decision. Descriptive theories of human judgment and decision making attempt to explain why people violate rational choice theory. These descriptive frameworks also view decisions as more reasonable and adaptive than previously thought. In this section a number of these frameworks will be briefly discussed.

4.1 Prospect Theory

Kahnemann and Tversky (1979) developed prospect theory to remedy the descriptive failures of SEU theories of decision making. Prospect theory attempts

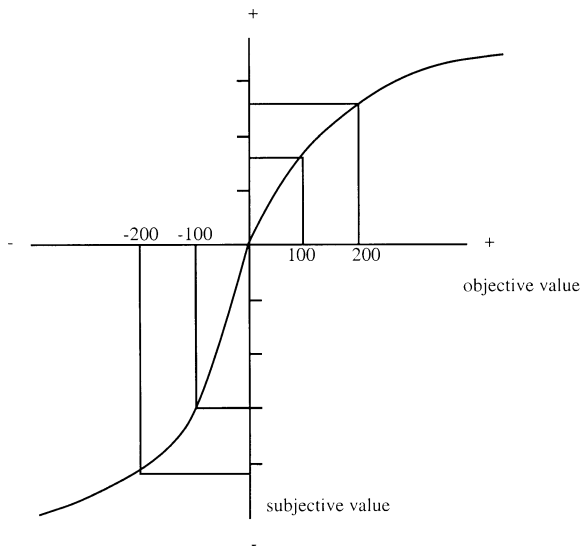

Figure 1
Prospect theory: hypothetical value function

to describe and explain decisions under uncertainty. Like SEU theories prospect theory assumes that the value of an option or alternative is calculated as the summed products over specified outcomes. Each product consists of a utility and a weight attached to the objective probability. Both the value function and the probability weighting function are nonlinear. The two functions are not given in closed mathematical form but have a number of important features. The most important feature of the probability weighting function is that small probabilities are overweighted, and large probabilities are underweighted. The probability weighting function is generally not well behaved near the end-points. Extremely low probability outcomes can be exaggerated or ignored entirely. Similarly, small differences between high probability and certainty are sometimes neglected, sometimes accentuated. According to Kahneman and Tversky this is so because people find it difficult to comprehend and evaluate extreme probabilities.

The value function is defined in terms of gains and losses relative to a psychologically neutral reference point. The value function is S-shaped; concave in the region of gains above the reference point, convex in the region of losses (see Fig. 1). Thus, each unit increase in gain (loss) has decreasing value as gain (loss) increases. In other words, the subjective difference between gaining nothing and gaining $100 is greater than the difference between gaining $100 and gaining $200. Finally, the value function is steeper for losses than for gains. This implies that losing $100 is more unpleasant than gaining $100 is pleasant.

Prospect theory can help to explain both the violation of the sure thing principle and some of the

framing effects discussed above. Thus, for respondents who won the first gamble in the Tversky and Shafir (1992) study, the average of the value of \$100 and \$400 may well be greater than the value of the \$200 they already won. After losing, the negative value for $-\$100$ is less than the average of $-\$200$ and $+\$100$. When people do not know whether they won or lost they will compare the possible outcomes with the zero-point, and in that case the gamble is not very attractive. Prospect theory can also help to explain the framing effects discussed above. The probability function and the value function will confirm the higher attractiveness of the risky option (in case of losses) and the risk-avoiding option (in case of gains).

Most other descriptive approaches assume that people rely on a variety of strategies or heuristics for solving decision problems. Experience will affect the availability of these strategies, and strategy choice will also be affected by the expected advantages (benefits) and disadvantages (costs) of the chosen strategy. For many decisions an exhaustive analysis such as prescribed by SEU-theory simply is not worth the trouble. Thus, for many problems people aim for an acceptable solution and not necessarily the optimal solution due to the costs (time, effort) of 'calculating' the best possible option.

4.2 Other Descriptive Approaches

Beach (1990) proposed 'image theory' which stresses the intuitive and automatic aspects of decision making. In this theory the decision-maker's goals or values (images) play a crucial role and much attention is paid to simplified strategies. The theory emphasizes that individuals make judgments about the compatibility of an alternative with one's image. This assessment is assumed to be rapid and intuitive. More analytical and elaborate decision-making processes such as these assumed by SEU theory are expected to be evoked only in specific circumstances (e.g., decisions with possible severe consequences). Beach refers to so-called noncompensatory strategies that are often used in real-life decision making. Since cognitive overload provokes a need for simplification, especially when one has to combine substantial amounts of relevant information about probabilities and values, it can be expected that people use decision rules that require less cognitive effort than normative theories such as SEU-theory.

Examples of simplifying decision-rules mentioned in the literature are: *The dominance rule* (alternative A should be chosen over B if A is better on at least one attribute and not worse than B on all remaining attributes), *the conjunctive decision rule* (any alternative that does not meet a minimally required value on one or more attributes is dropped from the list of remaining possible alternatives), and *the lexicographic decision rule* (choose alternative which is most attractive on the most important attribute. If two alternatives are equally attractive in terms of the most important attribute move to the next attribute). Finally, *the elimination by aspects rule* is often interpreted as a combination of the lexicographic rule and the conjunctive rule. First, the most important attribute is selected. All alternatives that fail to meet the criterion on this attribute are eliminated. This procedure is repeated for each of the remaining attributes.

These decision rules require considerably less cognitive effort than the decision rule required by SEU models (a compensatory decision rule which allows negative scores on one attribute to be compensated by positive scores on another attribute). These simplifying rules can provide adequate short cuts in complex decision environments. Payne et al. (1992) use the term contingent decision making in which the cognitive costs and benefits of the various strategies people might use determine their choice of strategy. Basic questions in this choice concern the balance between cognitive effort and accuracy, but also factors such as decisional avoidance and accountability influence strategy selection.

4.2.1 *Rule following*

Many decisions are based on rules that convey information about who we are and how we interact with others. They may convey a personal or moral identity, but also a social or group identity. They may also express habits such as routine purchases at the grocery. Generally, rules minimize effort and saves time. For instance, Prelec and Herrnstein (1991) discuss cases in which people avoid cost-benefit analyses and rely on prudential rules such as moral considerations and self-control (see also Mellers et al. 1998). Adopting a rule is especially useful for controlling behavioral practices with adverse consequences that are felt only with repetition or with a delay between benefits and costs. Examples are dietery behavior and alcohol consumption.

4.2.2 *Reason-based choice*

Sometimes rules do not apply and fail to single out the best action. In such cases decision makers may search for reasons to guide their choices. Reasons may be lists of pros and cons, but they can also take the form of stories. Pennington and Hastie (1992) investigated this in the context of jury decision making and presented evidence to respondents either as stories or issues and found that story organization resulted in stronger and more confident decisions. People often wish to have 'good reasons' for what they do. Reasons can also help to explain the violation of the sure thing principle discussed above. Tversky and Shafir (1992) argued that in the case of accepting the gamble for a second play the simple fact that having won the first time and hence having 'nothing to lose' would be a good reason to play a second time.

Other descriptive approaches introduced added elements that play a role in judgment and decision making. One of these approaches is *Regret Theory* (Bell 1982, Looms and Sugden 1987). The basic idea underlying this approach is that the value placed on the outcome of a specific choice or gamble also depends on the outcome that would have been received if one had chosen an alternative choice or gamble. Thus winning $100 is likely to be affected by the knowledge that one could have won $500 (vs., e.g., $10) if one had selected another alternative. Basically regret theory posits that people anticipate the possibility of regret and that this anticipation affects their decision making. People not only want to maximize utility, they also want to minimize the possibility of regret.

5. *Formulas vs. Intuitive Decision Making*

One of the aims of decision-making research is to improve our understanding of how decision makers search for information and of how this information is combined or processed. These processes of information integration are often called clinical judgment, and compared with the use of a formula or model. The latter can be based on experts' decision rules, or empirically assessed relationships between predictors (e.g., the presence of symptoms, scores on tests) and outcomes (e.g., having a specific disease or the ability of a job candidate). A substantial amount of research has shown that judgments are generally better if they are made using a formula (Dawes 1988). This applies especially to diagnostic judgments in which a limited number of indicators can lead to adequate prediction.

Payne et al. (1992) note that two issues are of importance in this context. First, what factors influence the use of a statistical, automated, decision procedure? Second, how can we reach a situation in which clinical and automated decision making complement one another, rather than compete? The still modest use of automated decision procedures could be related to limited knowledge about their benefits, experienced difficulties in applying the rules to individual decisions, and overly optimistic beliefs in the accuracy of clinical judgment.

6. *Conclusions*

Nearly five decades after the seminal work of Edwards (1954), decision making research is becoming more prominent in psychology textbooks and a clear and separate research area has emerged, generally referred to as Behavioral Decision Research. An important characteristic of this field of inquiry is that it relies on concepts, models and methods from a variety of disciplines including economics, statistics, and social and cognitive psychology. Research on decision making pays considerable attention to the discrepancies between normative models and actual behavior, although this is an issue that requires more conceptual and empirical work. Most of the research attempting to account for these discrepancies has focused on the information processing strategies, or heuristics, that people use when making judgments or decisions. A variety of descriptive approaches to human decision making has improved our insight into the cognitive processes underlying decisions, and should also help to improve the quality of decision making. Tasks for the future include determining when various heuristics are most likely to be used and the role of emotion in the decision-making process.

See also: Automaticity of Action, Psychology of; Bayesian Theory: History of Applications; Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision and Choice: Paradoxes of Choice; Decision and Choice: Random Utility Models of Choice and Response Time; Decision Biases, Cognitive Psychology of; Decision Making (Naturalistic), Psychology of; Decision Making: Nonrational Theories; Decision-making Systems: Personal and Collective; Decision Research: Behavioral; Decision Theory: Classical; Dynamic Decision Making; Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Heuristics for Decision and Choice; Heuristics in Social Cognition; Luce's Choice Axiom; Markov Decision Processes; Multi-attribute Decision Making in Urban Studies; Organizational Decision Making; Risk: Empirical Studies on Decision and Choice; Risk: Theories of Decision and Choice; Sequential Decision Making; Social Insurance: Legal Aspects; Subjective Probability Judgments; Unfolding and Vector Models; Utility and Subjective Probability: Empirical Studies

Bibliography

Beach L R 1990 *Image Theory: Decision Making in Personal and Organizational Context*. Wiley, New York

Bell D E 1982 Regret in decision making under uncertainty. *Operations Research* **30**: 961–81

Dawes R M 1988 *Rational Choice in an Uncertain World*. Harcourt Brace Jovanovich, San Diego

Dawes R M 1994 AIDS, sterile needles, and ethnocentrism. In: Heath L, Tindale R S, Edwards J, Posavec E J, Bryant F B, Henderson-King E, Suarez-Balcazar Y, Myers J (eds.) *Applications of Heuristics and Biases to Social Issues*. Plenum Press, New York, pp. 31–44

Edwards W 1954 The theory of decision making. *Psychological Bulletin* **51**: 380–417

Kahneman D, Tversky A 1973 On the psychology of prediction. *Psychological Review* **80**: 237–51

Kahneman D, Tversky A 1979 Prospect theory: Analysis of decision under risk. *Econometrica* **47**: 263–91

Looms G, Sugden R 1982 Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal* **92**: 805–24

Mellers B A, Schwartz A, Cooke A D J 1998 Judgmental decision making. *Annual Review of Psychology* **49**: 447–77
Payne J W, Bettman J R, Johnson E J 1992 Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology* **43**: 87–131
Pennington N, Hastie R 1992 Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology* **62**: 189–206
Prelec D, Herrnstein R J 1991 Preferences or principles: Alternative guidelines for choice. In: Zeckhauser R J (ed.) *Strategy and Choice*. MIT Press, Cambridge, MA, pp. 319–40
Shafir E 1993 Choosing versus rejecting: Why some options are both better and worse than others. *Memory and Cognition* **21**: 546–56
Simon H A 1955 A behavioral model of rational choice. *Quarterly Journal of Economics* **69**: 99–118
Simon H A 1957 *Models of Man: Social and Rational*. Wiley, New York
Tversky A, Kahneman D 1973 Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* **5**: 207–32
Tversky A, Shafir E 1992 The disjunction effect in choice under uncertainty. *Psychological Science* **3**: 305–9
Von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

J. van der Pligt

Decision-making Systems: Personal and Collective

The two best-known models for systematic personal decision-making are the maximization of utility and satisficing. The first was invented by Jevons and paradigmatically formulated by Wicksteed, with many formal refinements by succeeding authors; the second, which exemplifies a 'bounded rationality' approach, has been advocated by Herbert A. Simon. Neither translates easily into a model for systematic collective decision-making. Even as a personal model, maximization strictly conceived makes impossible demands for information. One would not expect it to fare any better collectively. Indeed it is hard to find a collective analogy even for its personal use in adjusting expenditures in any two categories so as to get the same marginal utility in each. It offers at best an idealized hypothetical basis for criticizing omissions in deliberation. But the ideal survives on the collective level in the confusion between being able to point out painfully omitted considerations after the event and being inclined to demand that all considerations be attended to beforehand, which is impossible.

Satisficing does not make this demand on either level; it asks only that a few threshold criteria (say, location, price, floor space) be laid down and that the first option that meets the criteria be chosen, though if this requires only a brief search through options, the next time the aspiration level of the decision-maker will rise, the thresholds will be set higher, and more criteria may be added. Unfortunately, at the collective level, unless an autocrat or a perfect bureaucracy (with a hierarchy that divides up all the issues that come up and assigns responsibility for deciding each of them under a coordinated plan) is authorized to act for the collectivity, a problem arises about different criteria with different thresholds being invoked at the same time by different people. Which are to guide the decisions of the collectivity?

This is a problem for social choice, and if the orderings of personal preferences for social criteria vary every whichway, aggregating the orderings runs into the Arrow theorem, before orderings for policy options themselves are considered. The theorem also applies to any attempt to decide directly on the options themselves. No system of aggregating personal orderings into social choices will meet at once all of Arrow's apparently very weak conditions about having the choices reflect the orderings. One of Arrow's conditions, typical in its weakness, is that if an option rises in one or more personal orderings without falling in any, then the option should not fall in the aggregated social ordering. The aggregated orderings can meet this condition and some others only at the cost of always being liable to fail because some three alternatives exhibit a cycle rather than a true ordering: for example, a majority prefer A to B, and a majority prefer B to C, but a majority prefer C to A. (An analogue of the theorem appears at the personal level, if the criteria in the set of criteria give very different orderings for the options [Thomas Schwartz 1970].)

One way of dealing with the Arrow theorem is to rely on the enterprise of politicians to invent a further option when attempts to aggregate personal orderings break down in cyclical voting. However, social choice theory raises an even worse problem (laid out in papers by R. D. McKelvey (1976) and Norman Schofield (1978), respectively) that a majority can be contrived for any option whatsoever if options are paired for decision, one pair after another, across the whole field of options. One way of dealing with this is to rely on the inertia of political institutions and on their divided jurisdictions to make it difficult to force immediate reconsideration of policies once chosen. Another way is to hope for widespread political agreement on criteria, taken together with a very limited scope for politics. A society of like-minded subsistence farmers (which was perhaps approximated in parts of New England in the eighteenth century) might get along very well without transacting much political business. The business need not be more than plausible departures from the status quo, satisficing departures that are also incremental, in important features of institutions no more than small changes; and the departures might fall into recurrent patterns answering to variations in the environment. This

arrangement could combine with a limited use of the market, either a perfectly competitive market or a market with a robust place for bargaining, allowing for trade in the products of crafts and small industry, operating with only a negligible draft on nonrenewable resources. It could even combine with a market in which modest technological innovations continually arise.

In either case, the market could serve, as Hayek (1948 following Adam Smith) maintained, to elicit details of information too multifarious to be easily dealt with by a deliberated plan, and give effects to these details in subtly coordinated features of the outcomes of economic activities. The market could serve in the same way in much more complex societies, and in many connections unobjectionably. It is thus itself a collective decision-making device furnishing from moment to moment a great variety of instructive information for producers and consumers. A purely competitive market does this without offering any room for bargaining; but a market in which bargaining is widespread can serve, too, and may likewise reach acceptable economic results that could not easily have been reached by any program for deliberate planning. The market with bargaining, moreover, is a device whose virtues can be carried over to some extent to political bargaining, in which contending interest groups reach one way or another, not always in ways that involve threats and counterthreats, what Lindblom has called 'partisan mutual adjustment,' in respect, say, to a presence of religion in the schools or on abortion policies. Bargaining will often, of course, reflect inequalities of power. The inequalities may be unjustified, and exploited in unconscionable ways. But Lindblom is supposing that partisan mutual adjustment, including hard bargaining, takes place in a context in which other systems or processes of decision, among them polyarchy, control through at least minimally democratic procedures, will operate to check oppression. Indeed, for him and Dahl, systems of decision-making come in variable combinations of four processes: the market; bargaining; polyarchy; and hierarchy (as in bureaucracies). Each does some things well that the others cannot, and all have a substantial presence in current democratic societies.

That is no ground for complacency. The present position of collective decision-making in the advanced industrial countries is one in which those taking part are compelled to make incremental decisions both because of inertia (institutional resistance to change) and because the risks of going wrong are so high. Incremental decisions, since they put only so much in present institutions at risk as can be repaired, are appropriate ones for agencies with very imperfect information, and that is the situation for all agencies during rapid social change. Prudence would not license doing more. Nevertheless, it may be suggested, things are going wrong in a very big way. Technological advances that combine with ever-increasing population to use up nonrenewable resources by leaps and bounds may be outracing attempts to combine limited scope for politics with the operation of the market. Think of deforestation; global warming; the collapse of commercial fisheries; the proliferation of the automobile; the marketing in the United States of firearms. In all these cases, collective decision-making seems to many knowledgable people to have faltered because of inattention in the market, inherently, with its shortrun preoccupations; and in politics, because attention there has been intermittent and inconclusive. (This is not to say that government can be counted on to choose better).

Moreover, with powerful political forces arrayed against taking due account of the troubles (which would mean, among other things, big interferences with the market, private profits, and personal incomes), it is difficult to see how the circumstances of decision-making can be made safe for satisficing and incrementalism. Things would not necessarily be better if corruption and special interests miraculously vanished from the system. The quickly multiplying evils cited are all public evils, and remedies for them would be public goods, from which no one could be excluded with respect to benefits and to which no one (taken by himself?) would make an indispensable contribution. So, as Mancur Olson showed in a basic application of rational choice theory (the utility-maximization theory of consumer behavior generalized to personal political choices), everyone would have a motive to escape making a contribution and to free-ride on the contributions of other people. Everyone could see as much by considering just what benefit he would get from the public good in question and how that benefit would not depend on his making a contribution.

This difficulty arises before any policy is adopted to produce the specific public goods that would remedy the evils in question. It arises about taking one's part in the system of collective decision-making. Why should a rational agent vote? There is a public good in having a polyarchal system in which the votes of the multitude exert some control over the politicians and thus check their greed and their hunger for power. It is desirable that the control extends to correcting for inequitable or shortsighted results from political bargaining and the operation of the market. Indeed a polyarchal system of decision-making will work as intended by its advocates only if this control in all its aspects is effective (an empirical condition) and a large proportion of the population participates (a definitional condition, often linked with an empirical hypothesis about the characterological benefits of participating). But no system for decision-making will work unless the people who have parts to play in the system actually take part; and polyarchy is peculiarly vulnerable on this point. Here, besides a troubling question about whether people will play their parts well, there is an equally troubling question about

whether people will play their parts at all. Every voter has a motive to avoid the bother of voting herself, since her contribution to maintaining the system will be effort thrown away either if a sufficient number of other people vote (which would make her vote redundant) or a sufficient number do not vote (which would make her vote quixotic). The same thing is true of voting for a specific policy or candidate.

Nevertheless, people do vote. In spite of complaints about turnout, they vote in large numbers even when there are no ward-heelers acting for political machines to turn them out. A common explanation is that they get entertainment value from voting. But 'entertainment' as invoked in this connection must, awkwardly, stretch to embrace an unexpected assortment of things, including the conscientious performance of duty, something like rooting for a horse to win, and perhaps something like prayer. Here the media, much berated for treating election campaigns as horse races, play a vital part. Even when they trivialize the issues, they make politics entertaining. At their best they induce people to take sides on important topics and consequently become anxious about the outcome. Voting is a ritual that discharges the anxiety and brings about closure for weeks, maybe months, of obsession with campaigns. Perhaps it is paradoxical that the maintenance of a system of decision-making should depend on such considerations, extraneous, it would seem, to the mechanics of recording and aggregating votes, and outside the program of narrow self-interest that rational choice theory starts with. But maybe the paradox lies with being attached to that program regardless of the wayward complications of human motivations.

It is true, rational choice theory can accommodate almost any considerations by accepting the considerations as having utility, even decisive utility, for the agents in question. If it keeps doing this, adding considerations after the event, it falls into tautology as an explanation of human decisions: humans decide on the basis of greater utility; but what has greater utility is to be inferred from what they decide. The tautology is easily escaped, however, even with the most diluted notions of utility: at one stage, consulting the agent and presenting her with experimental choices, drawing up the agent's utility map in advance. Then predicting her decision on the basis of the map is an empirical hypothesis, true or false as the facts turn out.

Moreover, the case of voting gives at least as much comfort to the theorists of narrow self-interest as disappointment. Large numbers of people eligible to vote do vote; but then large numbers do not, in spite of having the same stake in keeping up a democratic system and polyarchal control. In other connections, the free-riders form an even greater proportion of the public, sometimes enough to jeopardize the public good's being produced at all. Only something under 10 percent of its listeners make even small contributions to keeping the classical music radio station in Austin, Texas, on the air. Some of them are enticed (as Olson would expect) by private goods offered by the station as premiums to those who pay members' dues. Are the rest of those who decide to contribute just dupes? They may belong to that minority of the human race 'too rarely found to be presumed on' who out of 'generosity' act conscientiously and disdain to act otherwise (*Leviathan*, Chapter XIV, at the end).

In special respect to voting, turn outs in the millions, even if that is only half, maybe only a third, of the people eligible, suggest that Hobbes underestimated the proportion of conscientious people in the multitude. But even if they number enough to keep a democratic (or polyarchical) system going, will they be able to induce the system to make good decisions? However conscientious they may be, can they know enough about the complex issues of modern politics to tell good decisions from bad ones? There is a style of theorizing about decisions that throws great doubt upon the voters' capacities in this respect. It is just the style that treats social decisions as maximizing or satisficing and in either case having to strike a subtle balance between many different considerations. Doubt that the run of voters can join to do anything of the sort is reinforced by repeated empirical findings that most voters lack even what seems the most elementary political information (like which party is most eager to curtail welfare expenditures).

Yet some issues, even important ones, may lend themselves to forms of presentation that make the choices for voters easy. Among these may figure the grand absurdities of US politics, including the lack of gun control and the soaring rate of imprisonment; here a stark contrast between the status quo and a status quo ante where the problem did not exist, could elicit widespread agreement on the criterion of restoring the status quo ante so far as possible. A similarly stark and useful contrast may be drawn on occasion between the status quo in one's own country and the status quo in another, more successful in dealing with crimes or drugs or health insurance. In both sorts of cases sometimes the contrast can be brought before the public with the question, 'Shall we move to the better status quo?'; and if a simple remedy (which costs less and clearly promises to reduce or eliminate the trouble) is available, the public might reasonably be expected to make the appropriate decision in a referendum on the remedy.

In the absence of a referendum, the public might keep up pressure on suitably representative institutions for the decision that a referendum would lead to; but such pressure is difficult to sustain, and cross-purposes among the institutions, aggravated by corruption, may dissipate the impulse to reform. Tom Paine thought that it was desirable to have arrangements that made it possible to appeal from the people drunk to the people sober, and this may be applied as a caution about the unhappiness that may come from resorting to a referendum; but there may be occasion

for equal unhappiness with the decisions emerging from a complex of representative institutions. In any case, omitting to present the easy cases in the forms that make them easy, omitting to make them salient, and omitting to keep them salient until they have been resolved is not an argument against the capacities of voters or the use of polyarchy as a system of decision-making.

See also: Bounded and Costly Rationality; Bounded Rationality; Decision and Choice: Economic Psychology; Hierarchies and Markets; Market and Nonmarket Allocation; Markets and the Law; Rational Choice and Organization Theory; Rational Choice in Politics

Bibliography

Arrow K J 1963 *Social Choice and Individual Values*, 2nd edn. Wiley, New York

Braybrooke D, Lindblom C E 1963 *A Strategy of Decision: Policy Evaluation as a Social Process*. The Free Press, New York

Dahl R A, Lindblom C E 1953 *Politics, Economics, and Welfare: Planning and Politico-Economic Systems Resolved into Basic Social Processes*. Harper, New York

von Hayek F 1948 *Individualism and Economic Order*. University of Chicago Press, Chicago

Lindblom C E 1965 *The Intelligence of Democracy*. The Free Press, New York

McKelvey R D 1976 Intransitivities in multidimensional voting models and some implications for agenda control. *Journal of Economic Theory* **12**: 472–82

Olson M Jr 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA

Schofield N 1978 Instability of simple dynamic games. *Review of Economic Studies* **45**: 575–94

Schwartz T 1970 On the possibility of rational policy evaluation. *Theory and Decision*, **I**(1): 89–106

Simon H A 1957 *Models of Man: Social and Rational*. Wiley, New York

Wicksteed P H 1910 *The Common Sense of Political Economy*. Macmillan, London

D. Braybrooke

Decision Research: Behavioral

Rational choice theory has been, and continues to be, the dominant theoretical framework in the social sciences. The theory is enormously influential. Yet despite its popularity, the theory is based on untenable assumptions about human nature. Decision makers, or omniscient agents with stable and well-defined tastes, satisfy internal consistency and coherence by making choices that maximize their expected utilities.

Behavioral decision researchers have taken a descriptive approach to human decision making by exploring the actual judgments and decisions that people make as individuals, groups, and organizations. Researchers have identified numerous ways in which people violate the assumptions of rationality. The field has provided psychological insights about utilities and beliefs and has offered descriptive accounts of judgment and choice. This article discusses some of the empirical violations of rational choice theory, as well as alternative descriptive accounts of actual choices. Finally, it concludes with some comments about future directions.

1. Utility-Based Violations of Rational Choice

Behavioral decision researchers have identified a number of factors that influence utilities, but are either not addressed in rational choice theory or are inconsistent with it. A few of these factors are presented below.

1.1 Loss Aversion

In rational choice theory, decision makers are assumed to consider the utility of monetary outcomes as final states. They are sensitive to total wealth, not changes in wealth. In descriptive theories such as prospect theory (Kahneman and Tversky 1979), utilities (referred to as values) apply to gains and losses relative to the status quo. Furthermore, losses have greater impact than gains of equivalent magnitude. This assumption is known as loss aversion.

Loss aversion has been offered as an explanation for a well-known finding called the endowment effect, a result that cannot be predicted by rational choice theory. In a typical experiment on endowment effects, half of the participants are randomly assigned a gift, such as a university coffee mug. These participants are sellers, and the others are buyers. Sellers state their minimum selling prices for the mugs, and buyers give their maximum buying prices. Then an experimental market is conducted. Since buyers and sellers are determined randomly, rational choice theory predicts that approximately half of the mugs should be bought and sold. However, very few mugs are ever exchanged. Furthermore, selling prices are typically larger than buying prices by a factor of two or more. The psychological explanation for this effect is loss aversion; the pain of losing the object is greater in magnitude than the pleasure of gaining that object.

The way in which an object is endowed can also influence the value that people attach to it. People who

are rewarded with an object due to exemplary performance tend to value the object more highly than people who obtain the same object based on either chance or poor performance. These results tie in nicely with other findings that show that windfall gains, such as unexpected tax rebates, lottery winnings, or inheritances, are spent more readily than other assets, presumably because they are valued less.

1.2 Framing Effects

Endowment effects demonstrate how shifts in the status quo can influence the value attached to an object. Framing effects demonstrate how shifts in the *perception* of the status quo can influence choices between otherwise identical options. Framing effects were initially demonstrated by Tversky and Kahneman (1981). A now classic example of framing effects is the 'Asian disease' problem. Participants are told, 'Imagine that the US is preparing for the outbreak of an unusual Asian disease which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: With Program A, 200 people will be saved. With Program B, there is a 1/3 chance that 600 people will be saved, otherwise no one will be saved.' Participants are asked to select a program, and the majority pick Program A.

Another group of participants is told the same story, except the descriptions of the programs are changed from lives saved to lives lost. They read, 'With Program A, 400 people will die. With Program B, there is a 1/3 chance that no one will die, otherwise 600 people will die.' The majority select program B. Although Program A and B are identical in the two problems, preferences between programs reverse depending on the framing of options. Preferences are risk averse in the gain domain and risk seeking in the loss domain.

Framing effects go far beyond laboratory demonstrations. Johnson et al. (1993) found that preferences for automobile insurance in New Jersey and Pennsylvania varied greatly with shifts in the status quo. Both states offered coverage at similar costs, and both states offered reductions in cost if drivers would give up the right to sue other drivers if they were involved in a collision. In New Jersey, the default coverage had no right to sue, although a driver could purchase that right at additional cost. In Pennsylvania, the default coverage included this right, although drivers could decline it to reduce costs. Different 'status quos' led to large differences in choice; only 20 percent of New Jersey drivers, but as many as 75 percent of Pennsylvania drivers, purchased the right to sue. Johnson et al. estimated that if limited tort had been the status quo in Pennsylvania, drivers would have saved approximately $200 million in annual insurance costs.

1.3 Contextual Effects

Another psychological variable that influences choices but is not addressed in rational choice theory is the surrounding context or the set of options under consideration. Effects of the surrounding options can be distinguished in terms of the local context and the global context. The local context refers to one particular choice set, and the global context refers to the context that is built up after evaluating multiple choice sets. The power of both contexts to influence choice is remarkable.

The most well-known local contextual effect was demonstrated by Huber et al. (1982). They asked decision makers to choose between options A and B, each described by two attributes. Then they calculated the preference for B relative to A. At a later point, decision makers were offered a choice among options A, B, and C, where C was selected to be worse than B on both attributes, and worse than A on only one attribute. B looked better in the presence of C, and preferences for B relative to A increased with C in the choice set. This result was a violation of a normative principle known as regularity. The preference for an option should never increase with the addition of other options to the choice set. This effect is robust across choice options, order of choices, and types of experimental designs (i.e., within-subject designs, where the same subject makes both choices, and between-subject designs, where different subjects make each choice).

Global contextual effects refer to the implicit comparisons that participants make across choice sets. One example is the overall range of attribute levels. Several researchers have found that a fixed change in an attribute has greater impact when it appears in a narrow attribute range than in a wide attribute range. For example, suppose students judge the attractiveness of a set of apartments based on information about monthly rent and distance to campus. A $50 change in rent has greater impact on attractiveness ratings when the overall range of rents is narrow (e.g., $200 to $400) than when the overall range of rents is wide (e.g., $100 to $1,000). Some researchers have argued that range effects influence the relative importance of an attribute, whereas others claim that range effects stretch or shrink the utilities of the attribute levels. Evidence favors the utility interpretation.

1.4 Uneven Effects of Time on Utilities

Behavioral decision researchers are also interested in the psychological effects of time. In the normative framework, delayed outcomes are assumed to be discounted at a constant rate over time. However, numerous studies suggest that discounting varies with the decision domain, the reference point, the riskiness of the outcomes, and the sequence of outcomes. When investigating the effects of outcome sequences,

Loewenstein and Prelec showed that, when people are offered choices between sequences of dinners at restaurants, they prefer to save the nicer and more expensive dinner until last. Presumably they derive pleasure from anticipating the nicer dinner, as well as the experience of that dinner. Despite this seemingly reasonable desire, the preference pattern is inconsistent with rational choice theory. Rational choice implies that, when all else is held constant, better outcomes should be preferred sooner rather than later.

1.5 Inaccurate Forecasts of Future Preferences

Kahneman (1994) has recently pointed out yet another problem with rational choice theory. He claims that people are not always able to accurately forecast their own future preferences. To make his argument, Kahneman distinguishes among different types of utilities that he refers to as decision utilities, predicted utilities, experienced utilities, and retrospective utilities. Decision utilities reflect the satisfaction gained from outcomes in a choice context. Predicted utilities are the anticipated hedonic experiences of the outcomes, and experienced utilities are the actual hedonic experiences. Finally, retrospective utilities are the remembered experiences.

Kahneman claims that remembered utilities can systematically differ from experienced utilities because retrospective utilities are surprisingly insensitive to the duration of an experience. For example, Redelmeier and Kahneman examined moment-to-moment and retrospective evaluations of the pain experienced by patients undergoing diagnostic colonoscopy. Patients rated their discomfort every 60 seconds and then made an overall evaluation of the painfulness of the entire experience. The duration of the procedure, which ranged from four minutes to 69 minutes, did not predict retrospective evaluations. Instead, a peak-end rule, representing the average of the worst moments and the final moments of the experience, gave a reasonable description of evaluations.

If people base their choices on predicted utilities, and predicted utilities are guided by inaccurate retrospective utilities, forecasts of preferences will surely be inaccurate and choices could be suboptimal. To illustrate, Kahneman and his colleagues exposed participants to two painful experiences that involved immersing their hand in cold water. In the short experience, people immersed their hand in water that was 14 degrees Celsius for 1 minute. In the long experience, they put their hand in the same 14 degree water for 1 minute and kept it immersed for an additional minute as the temperature gradually rose to 15 degrees. People evaluated the longer experience as less painful, consistent with a peak-end rule, despite the fact that the duration of pain was longer. When later faced with a choice between the same two experiences, people preferred the longer experience over the shorter one. Results of these studies, though quite controversial, suggest enormous possibilities for decision engineering.

2. Belief-based Violations of Rational Choice

In rational choice theory, agents are assumed to weight the utilities of the outcomes by their beliefs in different states of the world. Subjective probabilities or beliefs should obey the rules of probability. Numerous studies suggest otherwise.

2.1 Posterior Probabilities

Edwards was one of the first to investigate whether people updated their beliefs according to Bayes Theorem. (See also *Bayesian Theory: History of Applications*.) He found that participants shifted their judged probabilities in the appropriate direction, but not to the right extent, a result that he called conservatism. Tversky and Kahneman (1982) later challenged his claim on the basis of the results of studies that used stories such as the 'cab problem.' Participants are told the following story.

> A cab was involved in a hit-and-run accident at night. Two cab companies, the green and the blue, operate in the city. Eighty-five percent of the cabs in the city are green and 15 percent are blue. A witness identified the cab as blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80 percent of the time and failed 20 percent of the time. What is the probability that the cab involved in the accident was blue rather than green?

Although some have argued that the cab problem is not adequately specified to determine a normative solution (Birnbaum 1983), many say the Bayesian solution to it is 41 percent. The majority of participants report that there is an 80 percent chance of the cab being blue. This value is sensitive to the 80 percent accuracy of the witness, but not the 15 percent base rate for blue cabs. This result and others like it led Tversky and Kahneman to argue that people often neglect base rate information and are not Bayesians at all. Hundreds of base-rate studies followed, and these studies have identified numerous boundary conditions on base rate neglect.

In a recent review, Koehler (1996) points out that, in fact, people frequently use base rates, depending on task structure and task representation. Within-subject designs, direct experience, frequency representations, unambiguous sample spaces, and random sampling are all factors that increase the degree to which people are sensitive to base rates when judging posterior

probabilities. People are not necessarily Bayesians, but there are many situations in which they are sensitive to base rate information.

2.2 Conjunctive Probabilities

Systematic errors in subjective probabilities of conjunctive events have also been identified by behavioral decision researchers (Tversky and Kahneman 1983). In another story, Tversky and Kahneman told participants about a woman named Linda who was described as 31 years old, single, outspoken, and very bright. She majored in philosophy and cared deeply about issues of discrimination and social justice. She also participated in anti-nuclear demonstrations. Then participants were asked to rank the likelihood of various statements, including 'Linda is a bank teller' and 'Linda is a bank teller and a feminist.' Participants report that the statement, 'Linda is a bank teller and a feminist' is more probable than 'Linda is a bank teller.' Tversky and Kahneman argued that these responses were violations of the conjunction rule, according to which the judged probability of the intersection of two events cannot exceed the judged probability of either single event. They claimed that people base their beliefs on the similarity of the target description to the category prototype, a strategy known as representativeness.

Gigerenzer and his collaborators have challenged this claim and have further explored the use of frequency representations to reduce, and even eliminate, conjunction effects. (Gigerenzer and Hoffrage 1995). They have had remarkable success at reducing base rate neglect and conjunction errors with frequency formats. This topic is yet another area of considerable controversy within the field.

2.3 Disjunctive Probabilities

Behavioral decision researchers have argued that judged probabilities are often inconsistent with the disjunctive rule, according to which the judged probability of the union of two events cannot be less than the judged probability of either single event. Tversky and Koehler (1994) recently showed that people systematically estimate disjunctive probabilities as less than the component event probabilities. For example, when one group of participants estimates the number of seven-letter words that end with 'n,' and another group estimates the number of seven-letter words ending in 'ing,' the latter group generates more examples than the former group, even though the latter group's task was a subset of the former group's task. Why? It is easier to think of seven-letter words ending in 'ing' than seven-letter words ending in 'n.'

Tversky and Koehler offered a theoretical account of these effects called support theory. People assign subjective probabilities to hypotheses about the world. These probabilities can psychologically stretch or increase as the hypotheses are 'unpacked' into more explicit disjunctions. The theory explains conjunction fallacies, hypothesis generation, decisions under uncertainty, fault-tree errors, as well as choice under uncertainty, by appealing to perceptual changes in likelihood with elaboration of plausible details.

Tetlock and Belkin (1996) have examined the effects of 'unpacking' past events, as well as future events. By imagining the likelihood of retrospective events that become increasing specific and detailed, people often judge the probabilities of those events as smaller. Tetlock and Belkin suggest the 'unpacking' of past events as a remedy for the hindsight bias, or the tendency to shift one's beliefs in the direction of an outcome once it is known.

3. Psychological Frameworks

Actual decisions depend on the individual, the context, and the environment. The combined effects of these factors are often described in terms of metaphors. Early metaphors focused on the ability of decision makers to cope with uncertainty, draw inferences, and test hypotheses about the world. Decision makers were treated as intuitive scientists, statisticians, and economists. Their judgments and choices were compared to optimal methods, and deviations were studied carefully. More recently, Tetlock (1992) has suggested some alternative metaphors. People also act as intuitive politicians balancing social pressures from competing sources, intuitive prosecutors who want accountability, retribution, and justice, and even intuitive theologians who protect sacred values from contamination. 'Mistakes' and 'errors' from one perspective may be perfectly reasonable from another.

3.1 Following Social and Cultural Rules

March (1994) offered a social and cultural metaphor for decision making called rule following. Rule following may have been based on tradeoffs initially, but eventually, these decisions become 'generic' applications of norms or conventions. Rule following conveys personal information about who we are, what we eat, and when we wake up in the morning. This strategy for decision making might also reflect a desire for self-control. Rule following is especially useful for controlling one's behavior when the impact of the act requires repetition (e.g., smoking a cigarette), has prolonged delays between costs and benefits (e.g., dieting), or has long-term benefits that are harder to imagine than short-term benefits (e.g., spending versus saving). This decision strategy can often minimize effort and permit the decision maker to avoid difficult or painful tradeoffs.

Fiske (1992) maintains that all social decisions boil down to four basic rules—communal sharing, authority ranking, equality matching, and market pricing. Communal sharing stresses common bonds

among group members, as found with families, lovers, and nations. Authority ranking focuses on inherent asymmetries in relationships; some people have higher rank, privilege, or prestige than others. Equality matching stresses reciprocity. Examples include babysitting cooperatives or car pooling, where one ought to give back whatever one takes. In market pricing, decisions are governed by supply and demand, expected utilities, or tradeoffs between costs and benefits.

What happens when a 'rule-following' decision maker applies the wrong rule to a social situation? Tetlock and his colleagues call these cases 'taboo tradeoffs.' Proposals to place monetary values on things usually thought of as priceless, such as children, body organs, or votes, do not just trigger cognitive confusion—they activate moral outrage, contempt, and a desire for retribution. The strength of people's reactions underscores the fact that taboo tradeoffs can threaten personal identities, community ties, and even social orders.

3.2 Reason-based Decisions

When rules conflict, people look for reasons to make choices (Shafir et al. 1993). Reasons may be lists of pros and cons, or they may take the form of stories. Pennington and Hastie (1993) argue that jurors construct stories to explain the facts. They present evidence to subjects either as stories or issues and find that story organizations result in stronger and more confident jury decisions than issue organizations. Reasons allow us to justify decisions to ourselves. Shafir et al. (1993) noted some interesting decision problems that arise when people have reasons to choose an act X if an event A occurs and different reasons to choose X if A does *not* occur. But if the outcome of event A is unknown, there are no reasons *of either type* to choose X, so X is rejected.

3.3 Emotion-based Decisions

How do emotions influence choices? Is the pleasure associated with an outcome well described by the utility of that outcome? Affective evaluations of outcomes are always relative, and different emotional experiences have been described in terms of different counterfactual comparisons. Some researchers have focused on the consequences of anticipated regret. Regret occurs when the obtained outcome is worse than it would have been under another choice. Other studies have focused on anticipated disappointment. Disappointment occurs when the obtained outcome is worse than an outcome under another state of the world. Both counterfactual comparisons have important influences on hedonic experiences.

There are interesting similarities between counterfactual reference points and the status quo, a critical reference point in the assessment of value. Like value shifts around the status quo, hedonic shifts around counterfactual reference points have asymmetric effects. Imagined losses have greater emotional intensity than imagined gains. Imagined losses can be quite general. Even gains can be imagined losses if one was expecting an even larger gain. The hedonic effects of these counterfactual anchors and others may be even more ubiquitous than the powerful effects of the status quo.

Mellers et al. (1999) proposed a theory of anticipated pleasure called decision affect theory. In decision affect theory, anticipated pleasure depends on obtained outcomes, counterfactual outcomes, and expectations. The theory implies that the anticipated pleasure associated with an outcome can differ from the utility of that outcome. For example, the utility of a larger win is always greater than that of a smaller win. But a smaller win can produce greater pleasure if it was completely unexpected or if a large loss was avoided in the process.

Mellers et al. also offered a link between anticipated pleasure and choice. They proposed that, when making choices, people anticipate the pleasure they might experience with all possible outcomes as described by decision affect theory, weigh those feelings by their subjective probabilities, and select the option with greater subjective expected pleasure. This theory differs from subjective expected utility theory because anticipated pleasure can differ from utility. Maximizing anticipated pleasure does not necessarily imply an egoistic variant of hedonism. It can arise from the senses, from acts of virtue, and from the relief of pain. Likewise, pain can arise from the cessation of pleasure, injustice, or the frustration from not achieving a goal.

4. Future Directions

Behavioral decision researchers have demonstrated important limitations on judgment and choice and have conveyed the message that people violate rationality. Furthermore, they have become much more circumspect about labeling behavior as 'irrational.' Although behavioral decision researchers disagree sharply among themselves about the extent and the implications of the violations, the general message that people systematically violate the assumptions of rational choice theory has had widespread effects in the social sciences. Behavioral assumptions are now being recognized in game theory, finance and legal decision making.

Demonstrations of human errors and foibles have also raised the level of debate about what constitutes rational behavior. If people repeatedly violate normative principles, at what point should the normative principles lose their elevated status? Irrational decisions can only occur when correct decisions are clearly defined. In the real world, this condition is often violated. More than one normative solution may apply. Problems may lack sufficient detail. Further-

more, decision makers may have other things in mind, especially when tasks have flat maxima or accuracy is more important than internal consistency. People may also want to avoid regret or conform to social and institutional norms. The integration of these goals with theories of behavioral decision making will provide fertile ground for future research.

For additional reading, see Kahneman and Tversky (2000) for theory and applications of behavioral decision theory, Kagel and Roth (1995) for experimental economics, Dawes (1998) for coverage of errors and biases, and Wright and Ayton (1994) for recent research on beliefs and subjective probabilities. For reading about decision technology, see Shafer (1997) for introductory material and Pearl (1988) for more advanced material.

See also: Bayesian Theory: History of Applications; Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision and Choice: Paradoxes of Choice; Decision and Choice: Random Utility Models of Choice and Response Time; Decision Theory: Bayesian; Decision Theory: Classical; Heuristics for Decision and Choice; Risk: Empirical Studies on Decision and Choice; Risk: Theories of Decision and Choice; Utility and Subjective Probability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies

Bibliography

Birnbaum M H 1983 Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychiatry* **96**: 85–94

Dawes R M 1998 Judgment, decision making, and interference. In: Gilbert D, Fiske S, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, MA, pp. 497–548

Fiske A P 1992 The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review* **99**: 689–723

Gigerenzer G, Hoffrage U 1995 How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* **102**: 684–704

Huber J, Payne J W, Puto C 1982 Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research* **9**: 90–8

Johnson E J, Hershey J, Meszaros J, Kunreuther H 1993 Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty* **7**: 35–51

Kagel J H, Roth A E (eds.) 1995 *The Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ

Kahneman D 1994 New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics* **150**: 18–36

Kahneman D, Knetsch J K, Thaler R H 1991 The endowment effect, loss aversion, and the status quo bias. *Journal of Economic Perspectives* **5**: 193–206

Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263–91

Kahneman D, Tversky A (eds.) 2000 *Choices, Values, and Frames*. Cambridge University Press, Cambridge, UK

Koehler J J 1996 The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences* **19**: 1–53

March J 1994 *A Primer on Decision Making*. Free Press, New York

Mellers B A, Schwartz A, Ritov I 1999 Emotion-based choice. *Journal of Experimental Psychology: General* **128**: 332–45

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, San Mateo, CA

Pennington N, Hastie R 1993 Reasoning in explanation-based decision making. *Cognition* **49**: 123–63

Shafer G 1997 *Probabilistic Expert Systems*. SIAM, Philadelphia, PA

Shafir E, Simonson I, Tversky A 1993 Reason-based choice. *Cognition* **49**: 11–36

Tetlock P E 1992 The impact of accountability on judgment and choice: Toward a social contingency model. *Advances in Experimental Social Psychology* **25**: 331–76

Tetlock P E, Belkin A (eds.) 1996 *Counterfactual Thought Experiments in World Politics*. Princeton University Press, Princeton, NJ

Tversky A, Kahneman D 1981 The framing of decisions and the psychology of choice. *Science* **211**: 453–8

Tversky A, Kahneman D 1982 Evidential impact of base rates. In: Kahneman D, Slovic P, Tversky A (eds.) *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK

Tversky A, Kahneman D 1983 Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* **90**: 293–315

Tversky A, Koehler D J 1994 Support theory: A nonextensional representation of subjective probability. *Psychological Review* **101**: 547–67

Wright G, Ayton P (eds.) 1994 *Subjective Probability*. Wiley, New York

B. A. Mellers

Decision Support Systems

1. Introduction

A decision is a choice made by some agent between competing beliefs about the world or between alternative courses of action to achieve the agent's goals. The ability to reason flexibly about problems and make good choices are fundamental to any kind of intelligence and the need to do them well is highly valued in society. Consequently, there is a growing demand for *decision support systems* which can assist human decision makers make important choices more effectively. This demand has stimulated research in many

fields, including the behavioural and cognitive sciences (notably psychology, artificial intelligence, and management science) as well as various mathematical disciplines (such as computer science, statistics, and operations research).

Behavioral scientists have traditionally drawn a distinction between two kinds of theory that address problems in decision making. *Prescriptive* theories set out formal methods for reasoning and decision making, and criteria for 'rational' inference. *Descriptive* theories typically seek to explain how people make decisions and account for the errors they make that may result in violations of rational norms. Countless empirical studies of human judgement, in the laboratory and in real-world settings, have demonstrated that human decision making is subject to a variety of systematic errors and biases compared with prescriptive decision models (see *Decision Research: Behavioral*).

There are many reasons for this. Mistakes are clearly more likely when someone is tired or overloaded with information, or if a decision maker lacks the specialist knowledge required to make the best choice. Even under ideal conditions however, with full information, people still make 'unforced errors.' One of the major reasons for this is that we do not seem to be very good at managing uncertainty and complexity.

In the 1970s and 1980s cognitive scientists came to the view that people are not only rather poor at reasoning under uncertainty, but that they also revise their beliefs by processes that bear little resemblance to formal mathematical calculation, and it is these processes that seem to give rise to the characteristic failures in decision making. Kahneman and Tversky developed a celebrated explanation that people use various heuristics in making decisions under uncertainty, as when they judge things to be highly likely when they come *easily to mind* or are *typical* of a class rather than by means of a proper calculation of the relative probabilities. Such heuristic methods are often reasonable approximations but they can also lead to systematic errors (see also *Heuristics for Decision and Choice* and *Subjective Probability Judgments*).

Psychological research on 'deductive reasoning,' has looked at categorical rather than uncertain inferences, as in syllogistic reasoning but with similar questions in mind, namely: how do people carry out logical tasks? And how well do they do them by comparison with prescriptive logical models? (see also *Problem Solving and Reasoning: Case-based*). Systematic forms of error and failures to recognize and avoid logic fallacies have also been found, and this is probably because people do not arrive at a conclusion by applying inference rules the way a logician might, but they appear to construct a concrete 'mental model' of the situation, and manipulate it in order to determine whether a proposition is supported in the model (see also *Mental Models, Psychology of*).

In summary, many behavioral scientists now take the view that human cognition is fundamentally flawed (e.g., Sutherland 1992). Good reviews of different aspects of research on human reasoning and decision making can be found in Kahneman et al. (1982), Evans and Over (1996), Wright and Ayton (1994), Gigerenzer and Todd (1999) and Stanovich and West (2000).

2. *Mathematical Methods and Decision Support Systems*

If people demonstrate imperfect reasoning or decision making then it would presumably be desirable to support them with techniques that avoid errors and comply with rational rules. There is a vast amount of research on *decision support systems* that are designed to help people overcome their biases and limitations, and make decisions more knowledgeably and effectively. If we are to engineer computer systems to take decisions, it would seem clear that we should build those systems around theories that give us some appropriate guarantees of rationality. In a standard text on rational decision making, Lindley summarizes the 'correct' way to take decisions as follows:

> … there is essentially only one way to reach a decision sensibly. First, the uncertainties present in the situation must be quantified in terms of values called probabilities. Second, the consequences of the courses of actions must be similarly described in terms of utilities. Third, that decision must be taken which is expected on the basis of the calculated probabilities to give the greatest utility. The force of 'must' used in three places is simply that any deviation from the precepts is liable to lead the decision maker to procedures which are demonstrably absurd. (Lindley 1985, p. vii)

The above viewpoint leads naturally to Expected Utility Theory (EUT) which is well established and very well understood mathematically. If its assumptions are satisfied and the expected utilities of alternative options are properly calculated, it can be argued that the procedure will reliably select the best decision. If people made more use of EUT in their work, it is said, this would result in more effective decision making. Doctors, for example, would make more accurate diagnoses, choose better treatments, and make better use of resources. Similar claims are made about the decision making of managers, politicians, and even juries in courts of law.

3. *Limits to Theory*

> As living and moving beings, we are forced to act … [even when] our existing knowledge does not provide a sufficient basis for a calculated mathematical expectation. (John Maynard Keynes, quoted by Bernstein 1996).

Many people think that the practical value of mathematical methods of decision making like SEU

and the 'irrationality' of human decision makers are overstated. First of all, an expected-utility decision procedure requires that we know, or can estimate reasonably accurately, all the required probability and utility parameters. This is frequently difficult in real-world situations since a decision may still be urgently required even if precise quantitative data are not available. Even when it is possible to establish the necessary parameters, the cost of obtaining good estimates may outweigh the expected benefits.

Furthermore, in many situations a decision is needed before the decision options, or the relevant information sources, are fully known. The complete set of options may only emerge as the decision making process evolves. Neither logic nor decision theory provide any guidance on this evolutionary process. Lindley acknowledges this difficulty:

> The first task in any decision problem is to draw up a list of the possible actions that are available. Considerable attention should be paid to the compilation of this list [though] we can provide no scientific advice as to how this should be done.

In short, the potential value of mathematical decision theory is limited by the frequent lack of objective quantitative data on which to base the calculations, the limited range of functions that it can be used to support, and the problem that the underlying numerical representation of the decision is very different from the intuitive understanding of human decision makers.

There are also many who doubt that people are as 'irrational' as the prescriptive theories appear to suggest (see also *Decision Research: Behavioral*; *Utility and Subjective Probability: Contemporary Theories*). Skilled professionals may find it difficult to accept that that they do not make decisions under uncertainty as well as they 'should' and that under some circumstances their thinking can actually be profoundly flawed. Most of us have no difficulty accepting that our knowledge may be incomplete, that we are subject to tiredness and lapses of attention, and even that our abilities to recall and bear in mind all relevant factors are imperfect. But we are less willing to acknowledge an irremediable irrationality in our thought processes ('people complain about their memories but never about their judgement').

A more optimistic school of thought argues that many apparent biases and shortcomings are actually artefacts of the artificial situations that researchers create in order to study reasoning and judgement in controlled conditions. When we look at real-world decision making we see that human reasoning and decision making is in fact far more impressive than the research suggests. Herbert A. Simon observes that 'humans, whose computational abilities are puny compared with those of modern super-computers or even PCs, are sometimes able to solve, with very little computation, problems that are very difficult even by computer standards—problems having ill-defined goals, poorly characterized and bounded problem spaces, or which lack a strong and regular mathematical structure.' (Simon 1995)

Shanteau (1987) has investigated 'factors that lead to competence in experts, as opposed to the usual emphasis on incompetence,' and he identifies a number of important positive characteristics of expert decision makers: they know what is relevant to specific decisions, what to attend to in a busy environment, and they know when to make exceptions to general rules. Second, experts know a lot about what they know, and can make decisions about their own decision processes: they know which decisions to make and when, and which to skip, for example. They can adapt to changing task conditions, and are frequently able to find novel solutions to problems. Classical deduction and probabilistic reasoning do not capture these metacognitive skills. Consider medical decision making as an example. Most of the research in this area has viewed clinical decision making primarily in terms of deciding what is wrong with a patient (determining the diagnosis) and what to do based on the diagnosis (selecting the best treatment) (see *Problem Solving and Reasoning, Psychology of*; *Bounded and Costly Rationality*).

In practice a doctor's activities and responsibilities are extremely diverse. They include:

(a) recognizing the possibility of a significant clinical problem
(b) identifying information that is relevant to understanding the problem
(c) selecting appropriate investigations and other procedures
(d) deciding on the causes of clinical abnormalities and test results
(e) interpreting time-dependent data and trends (e.g., blood pressure)
(f) setting out clinical goals
(g) formulating treatment plans over time
(h) anticipating hazards
(i) creating contingency plans
(j) assessing the effectiveness of treatment

Each of these tasks involves different types of patient information and requires many different types of knowledge in order to interpret the information and act upon it appropriately. (For related issues in other areas see *Bounded and Costly Rationality*.) We must understand this complexity if we are to develop theories that are sufficiently sophisticated to capture the diversity or to build decision support systems which can emulate and improve upon human capabilities.

It has also been strongly argued that people are actually well adapted for making decisions under the adverse conditions caused by time pressure, lack of detailed information, and knowledge, etc. Gigerenzer, for instance, has suggested that human cognition is

rational in the sense that it is optimized for speed at the cost of occasional and usually inconsequential errors (Gigerenzer and Todd 1999).

4. *Effects of Tradeoffs on Effectiveness of Decision Making*

> … cognitive mechanisms capable of successful performance in a real world environment do not need to satisfy the classical norms of rational inference: the classical norms may be sufficient, but are not necessary, for a mind capable of sound reasoning. (Gigerenzer and Goldstein, 'Reasoning the fast and frugal way,' *Psychological Review*, 1996)

Tradeoffs simplify decision making and in practice may entail only modest costs in the decision maker's accuracy and effectiveness. This possibility has been studied quite extensively in the field of medical decision making. In the prediction of sudden infant death, for example, Carpenter et al. (1977) attempted to predict death from a simple linear combination of eight variables. They found that weights can be varied across a broad range without decreasing predictive accuracy.

In diagnosing patients suffering from dyspepsia, Fox et al. (1980) found that giving all pieces of evidence equal weights produced the same accuracy as a more precise statistical method (and also much the same pattern of errors). Fox et al. (1985) developed a system for the interpretation of blood data in leukemia diagnosis, using the EMYCIN expert system software. EMYCIN provided facilities to attach numerical 'certainty factors' to inference rules. Initially a system was developed using the full range of available values (-1 to $+1$) though later these values were replaced with just two: if the rule made a purely categorical inference the certainty factor was set to be 1.0 while if there was *any* uncertainty associated with the rule the certainty factor was set to 0.5. The effect was to *increase* diagnostic accuracy by 5 percent.

In a study of whether or not to admit patients with suspected heart attacks to hospital, O'Neil and Glowinski (1990) found no advantage of a precise decision procedure over simply 'adding up the pros and cons.' Pradhan et al. (1996) carried out a similar comparison in a diagnosis task and showed a slight increase in accuracy of diagnosis with precise statistical reasoning, but the effect was so small that it would have no practical clinical value.

In a recent study of genetic risk assessment for 50 families the leading probabilistic risk assessment software was compared with a simple procedure made up of *if … then …* rules (e.g., *if* the client has more than two first-degree relatives with breast cancer under the age of 50 *then* this is a risk factor). Despite the use of a simple weighting scheme for each rule, the rule-based system produced exactly the same risk classification for all cases as the probabilistic system (Emery et al. 2000).

While the available evidence is not conclusive a provisional hypothesis is that, at least for certain kinds of decision such as clinical diagnosis and patient management decisions, the strict use of quantitatively precise decision-making methods may not add much practical value to the design of decision support and artificial intelligence systems.

5. *Nonclassical Methods for Decision Making*

In artificial intelligence (AI) the desire to develop versatile automata has stimulated a great deal of research in new methods of decision making under uncertainty, ranging from sophisticated refinements of probabilistic methods such as 'Bayesian networks,' and nonprobabilistic methods such as fuzzy logic and possibility theory. Good overviews of the different approaches and their applications are Krause and Clark (1993) and Hunter and Parsons (1998). (See also *Artificial Intelligence: Uncertainty*.) These approaches are similar to probability methods in that they treat uncertainty as a matter of *degree*. However, quantitative approaches in general have also been questioned in AI because they require much data and do not capture varied human intuitions about the nature of 'belief,' 'doubt,' and natural justifications for decision making.

Consequently, interest has grown in the use of nonnumerical methods for reasoning under uncertainty that seem to have some 'common sense' validity. Attempts have been made to develop qualitative approximations to quantitative methods, such as *qualitative probability* (Wellman 1990). In addition new kinds of logic have been proposed, including:

(a) nonmonotonic logics which express the everyday idea of changing one's mind, as though the probability of some proposition being true is 1.0 at one point but at some later point the probability becomes zero;

(b) default logic, a form of nonmonotonic logic which formalises the idea of assuming that something is true until there is reason to belief otherwise;

(c) defeasible reasoning in which one line of reasoning can 'rebut' or 'undermine' another line of reasoning

Cognitive approaches, sometimes called 'reason-based' decision making, are gaining ground, including the idea of using informal 'endorsements' for alternative decision options and logical formalizations of everyday strategies of reasoning about competing beliefs and actions based on 'argumentation.' Models of argumentation are reviewed by Krause and Clark (1993) and the role of argumentation techniques in decision support systems is surveyed by Girle et al. (2001) (see also *Problem Solving and Reasoning: Case-based* and *Linear Algebra for Neural Networks*).

Some advocate an eclectic approach to the formalization of uncertainty that sanctions the use of different representations under different circumstances. Fox

and Das (2000, Chap. 4) discuss the possibility that many techniques for reasoning under uncertainty, from quantitative and qualitative probability to default logic and representation of uncertainty in natural language, capture different intuitions about 'belief' and decision making but can all be viewed as different technical specializations of the informal notion of argumentation.

Finally, as noted above, human decision makers often demonstrate 'metacognitive' capabilities, showing some ability to reason about the nature of the decision, the relevant information sources, and applicable forms of argument, and so forth. Metacognition may be at the heart of what we call 'intelligence.' Designers of decision support systems now have available a range of formalisms based on mathematical logic which can emulate such 'meta-level reasoning' (Fox and Das 2000) while conventional algorithmic decision procedures seem to be confined to reasoning about the input data, not the decision problem itself.

It appears from the foregoing that the claim that decision support systems should not be modeled on 'irrational' human cognitive processes is less compelling than it first appeared. This offers greater flexibility for the designers of decision support systems; they can adopt different decision theoretic frameworks for different applications. Indeed, they may also be able to apply various metacognitive strategies, as people appear to do, to select the most effective representations and reasoning methods in light of the demands and constraints of the current task.

See also: Bayesian Theory: History of Applications; Decision Research: Behavioral; Decision Theory: Bayesian; Decision Theory: Classical; Expert Systems in Cognitive Science; Expert Systems in Medicine; Heuristics for Decision and Choice; Multi-attribute Decision Making in Urban Studies

Bibliography

Bernstein P 1996 *Against the Gods*. Wiley, New York, p. 185

Carpenter R G, Garnder A, McWeeny P M, Emery J L 1977 Multistage scoring system for identifying infants at risk of unexpected death. *Archives of Disease in Childhood* **52**: 606–12

Emery J, Walton R, Murphy M, Austoker J, Yudkin P, Chapman C, Coulson A, Glasspool D, Fox J 2000 Computer support for interpreting family histories of breast and ovarian cancer in primary care: Comparative study with simulated cases. *British Medical Journal* **321**: 28–32

Evans J St B, Over D E 1996 *Rationality and Reasoning*. Erlbaum, London

Fox J, Barber D C, Bardhan K D 1980 Alternatives to Bayes? A quantitative comparison with rule-based diagnosis. *Methods of Information in Medicine* **10**(4): 210–15

Fox J, Das S K 2000 *Safe and Sound: Artificial Intelligence in Hazardous Applications*. American Association of Artificial Intelligence and MIT Press, Cambridge, MA

Fox J, Myers C D, Greaves M F, Pegram S 1985 Knowledge acquisition for expert systems: Experience in leukaemia diagnosis. *Methods of Information in Medicine* **24**(1): 65–72

Gigerenzer G, Todd P M 1999 *Simple Heuristics that Make us Smart*. Oxford University Press, Oxford

Gigerenzer G, Goldstein D G 1996 Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* **103**: 650–69

Girle R, Hitchcock D, BcBurney P, Verheij B 2001 Decision support for practical reasoning: A theoretical and computational perspective. In: Reed C, Norman T (eds.) *Proceedings of the Bonskeid Symposium on Argument and Computation*

Hunter A, Parsons S (eds.) 1998 *Applications of Uncertainty Formalisms*. Lecture Notes in Computer Science 1455. Springer, Berlin

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Heuristics and Biases*. Cambridge University Press, Cambridge, UK

Krause P, Clark D 1993 *Representing Uncertain Knowledge*. Kluwer Academic Publishers, Norwell, MA

Lindley D V 1985 *Making Decisions*, 2nd edn. Wiley, London

O'Neil M J, Glowinski A J 1990 Evaluating and validating very large knowledge-based systems. *Medical Informatics* **15**(3): 237–51

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufmann. Palo Alto, CA

Pradhan M 1996 The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial Intelligence Journal* **85**: 3636–97

Shanteau J 1987 Psychological characteristics of expert decision makers. In: Mumpower J (ed.) *Expert Judgement and Expert Systems*. NATO ASI Series, Vol. F35

Simon H A 1995 Artificial intelligence: An empirical science. *AI Journal* **77**(1): 95–127

Stanovich K E, West R F 2000 Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences* **22**(5)

Sutherland N S 1992 *Irrationality: The Enemy Within*. Constable, London

Wellman M P 1990 Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence Journal*. **11**(2): 145–72

Wright G, Ayton P (eds.) 1994 *Subjective Probability*. Wiley, Chichester, UK

J. Fox

Decision Theory: Bayesian

1. Overview and History

Decision theory studies the logic and the mathematical properties of decision making under uncertainty, and is applied to a broad spectrum of human activities. Much of it makes use of ideas, approaches, and formalisms that can be categorized as Bayesian. This discussion focuses on a specific, statistically oriented, definition of decision theory: the investigation of decision making when the state of the world is uncertain, yet further information about it can be obtained through experimentation. We discuss the

approach that stems from the theory of Savage (1954), and was developed in depth in the seminal texts of Raiffa and Schleifer (1961) and DeGroot (1970).

At the core of Bayesian decision theory is the principle of maximization of expected utility. A Bayesian decision maker proceeds by assigning a numerical utility to each of the possible consequences of an action, and a probability to each of the uncertain events that may affect that utility. The weighted average of the utility with respect to the probability is then used as the criterion of choice. In statistical decision theory, this plan can be applied to actions that represent inferences, predictions, or conclusions to be drawn from experimental evidence. Decision theory enables statisticians to bring basic principles of rationality to bear on both the development and the evaluation of statistical methodologies.

Decision making using probability, utility, and the expected utility principle has deep roots. The idea that mathematical expectation should guide rational choice under uncertainty was formulated and discussed as early as the seventeenth century. During that time, an important problem was finding the fixed amount of money that would be fair to exchange for an uncertain payoff, as when paying an insurance premium. Initially, the prevailing thought was that the fair fixed amount would be the expected payoff. The St. Petersburg game (Jorland 1987) revealed an example where the expected payoff is infinite but where, in the words of Bernoulli (1738), 'no reasonable man would be willing to pay 20 ducats as equivalent.' Bernoulli's view was that 'in their theory, mathematicians evaluate money in proportion to its quantity while, in practice, people with common sense evaluate money in proportion to the utility they can obtain from it.' Using this notion of utility, Bernoulli offered an alternative approach: he suggested that the fair sum to pay for a game of chance is the moral expectation, which effectively is an example of expected utility.

Another important element of Bayesian decision theory that can be traced to the Age of Enlightment is the notion that, when weighing evidence provided by data, one should often take explicitly into account the consequences of the actions that are to be taken using the data. Condorcet (1785), for example, reasoned that 'The probability that a convicted person is guilty should be in proportion to the probability that an acquitted person is innocent, as is the inconvenience of convicting an innocent in proportion to that of acquitting a culprit.' Later, the use of optimality principles in the evaluation of inferential procedures was present in some of the key developments in mathematical statistics, such as estimation efficiency (Fisher 1925), and power of statistical tests (Neyman and Pearson 1933), an approach that explicitly recognized the link between decision making and hypothesis testing. The broader connection between rational decision making and statistical inference was eventually formalized in great generality by Wald (1949), who is considered the founder of statistical decision theory.

The use of probability to quantify a decision maker's knowledge about uncertain events originates from the work of Ramsey (1931) and de Finetti (1937) on subjective probability. The subjectivist approach provides a coherent foundation to probability statements about a broad typology of events, and can be integrated in a natural way into individual decision making. A strength of the subjectivist view in statistical decision theory is in the use of a single measure of uncertainty for both experimental variation and uncertainty about the state of the world. This provides a simple, general, and self-consistent formal mechanism for modeling the effects of accruing experimental evidence on knowledge. See also *Probability: Interpretations*.

Justifications for taking the principle of maximization of expected utility as a general guide to rational statistical inference have also been the subject of extensive investigation. A fruitful approach has been to translate this principle into a set of intuitive axioms in terms of preferences among actions that are satisfied if and only if one's behavior is consistent with expected utility. The critical contribution in this regard is that of von Neumann and Morgenstern (1944), who developed the first and fundamental axiom system. Savage (1954) used the subjectivist approach, together with some of the technical advances provided by von Neumann and Morgenstern and Wald, to develop a general and powerful theory of individual decision making under uncertainty, which provides the logical foundation of current Bayesian decision theory. A sketch of what would become Savage's development is outlined by Ramsey (1931).

2. *The Principle of Maximization of Expected Utility*

In the prototypical formulation of decision making under uncertainty, an individual decision maker (DM) must choose one among a set of actions, whose consequences depend on some unknown state of the world. The set of actions will be denoted here by $\mathscr{A}$, with generic action a; the set of consequences by $\mathscr{Z}$, with generic consequences z, and the set of states of the world by Θ, with generic element θ. Objects in Θ and $\mathscr{Z}$ can be arbitrary. The basis for choosing among actions is a quantitative evaluation of the utility of the consequences. An action's consequences depend on the unknown state of the world, however, and each action yields a certain consequence corresponding to each state of the world. This correspondence defines an action, that is an actions is a function a from Θ to $\mathscr{Z}$.

Utilities of consequences are measured by a real valued function u and probabilities of states of the world are represented by a probability distribution π.

The expected utility $\mathscr{U}_\pi$ of action a is then the weighted average of the utilities of the possible consequences of the action, weighted by the probability of the corresponding state of nature. Formally,

$$\mathscr{U}_\pi(a) = \int_\Theta u(a(\theta))\pi(\theta)\,\mathrm{d}\theta \tag{1}$$

An optimal action, or 'Bayes' action a_{B}, is one that maximizes the expected utility, that is,

$$a_{\mathrm{B}} = \operatorname{argmax} \mathscr{U}_\pi(a) \tag{2}$$

Careful specification of the functions u and π is critical to sound decision making. Techniques are studied in decision analysis. See also *Elicitation of Probabilities and Probability Distributions*.

Following Wald (1949), the statistical literature often formulates decision problems in terms of the opportunity loss (or 'regret') associated with each pair (θ, a) by defining a loss function

$$L(\theta, a) = -u(a(\theta)) + \sup_{a(\theta)} u(a(\theta)) \tag{3}$$

where the sup is taken over the consequences of every possible action for fixed θ. The loss $L(\theta, a)$ is the difference between the utility of the consequence of the action for state θ, and the utility of the consequences of the best action for that state (or the lowest upper bound if the best action cannot be achieved). The expected loss is

$$\mathscr{L}_\pi(a) = -\mathscr{U}_\pi(a) + S_\pi \tag{4}$$

where $S_\pi = \int_\Theta \sup_{a(\theta)} u(a(\theta))\pi(\theta)\,\mathrm{d}\theta$ is independent of a. Therefore, the action minimizing the expected loss is again the Bayes action, and the two formulations are equivalent from this standpoint. In statistical analyses, loss functions are often stated directly, without reference to underlying consequences or utility functions.

In statistical decision theory the main alternative criterion to choosing a Bayes action is to choose a minimax action a_{M}, defined as

$$a_{\mathrm{M}} = \operatorname{argmin} \max_\theta L(\theta, a) \tag{5}$$

In words, to find the minimax action the DM considers each action in turn and computes the largest loss under any possible state of the world, for that action. The DM then chooses the actions with the smallest maximum loss. This approach does not requires specifying a probability distribution π over the states of the world. For this reason it is favored by the Frequentist school and was adopted in Wald's original formulation of decision theory. The Minimax action is not necessarily the same under the utility and loss formulation. Berger (1985) discusses the relationships between the Bayes and Minimax approaches, and provides conditions on the distribution π such that the two approaches lead to the same action. See also *Decision Theory: Classical* and *Game Theory and its Relation to Bayesian Theory*.

3. *Axiomatic Foundations*

Axiomatic expected utility theory has been concerned with identifying axioms in terms of preferences among actions, that are satisfied if and only if one's behavior is consistent with expected utility, thus providing a foundation to the use of the Bayes action. The fundamental axiom system is that of Savage (1954). A similar and relatively simple formulation, that provides insights into the most important assumptions of expected utility for statistical purposes, was proposed by Anscombe and Aumann (1963). Critical discussions and analytical details are found in Kreps (1988), whose formulation is followed here, and Schervish (1995).

For simplicity, assume that Θ and $\mathscr{Z}$ are finite: n_Θ is the number of elements in Θ. A simple action is a function a from Θ to $\mathscr{Z}$. Although interest is often in simple actions, Anscombe and Aumann build their theory in terms of more general objects that allow for a state of the world θ to determine the consequence of an action only up to a known probability distribution. In their theory an action is a function $\boldsymbol{a}$, such that a $(\theta, \cdot)$ is a known probability distribution on $\mathscr{Z}$. The boldface type distinguishes these general actions from simple action $a(\theta) = z$, corresponding to the special case in which, for every θ, there is a z such that $\boldsymbol{a}(\theta, z) = 1$.

The theory requires a rich set of possible actions. Specifically, for any two $\boldsymbol{a}$ and $\boldsymbol{a}'$ from $\mathscr{A}$, the composite action $\alpha\boldsymbol{a} + (1-\alpha)\boldsymbol{a}'$, given pointwise by the convex combinations $\alpha\boldsymbol{a}(\theta, z) + (1-\alpha)\boldsymbol{a}'(\theta, z)$, with $\alpha \in [0, 1]$ is also in $\mathscr{A}$. Also, the state θ is said to be null if the DM is indifferent between any two actions that differ only in what happens if θ occurs.

Choosing actions according to the principle of expected utility is equivalent to holding preferences among actions that satisfy the five axioms of Table 1.

The weak-ordering axiom requires that any two actions can be compared and in a transitive way. It effectively permits representation of the preferences as one-dimensional. The Archimedean axiom bars the DM from preferring one action to another so strongly that no reversal of preference by creating a composite action that includes one of the two is possible. This axiom carries most of the weight in representing the preference dimension as the real line. The Independence axiom requires that two composite lotteries should be compared solely based on the component that is different. It carries most of the weight in guaranteeing the 'expected' in the expected utility

Table 1
Axioms of the Anscombe–Aumann expected utility theory

1. Weak order	$\succ$ on $\mathscr{A}$ is a preference relation
2. Independence	If $\boldsymbol{a} \succ \boldsymbol{a}'$ and $\alpha \in (0, 1)$, then for every $\boldsymbol{a}'' \in \mathscr{A}$ $\alpha\boldsymbol{a} + (1-\alpha)\boldsymbol{a}'' \succ \alpha\boldsymbol{a}' + (1-\alpha)\boldsymbol{a}''$
3. Archimedean	If $\boldsymbol{a} \succ \boldsymbol{a}' \succ \boldsymbol{a}''$, then there are $\alpha, \beta \in (0, 1)$ such that $\alpha\boldsymbol{a} + (1-\alpha)\boldsymbol{a}'' \succ \boldsymbol{a}' \succ \beta\boldsymbol{a} + (1-\beta)\boldsymbol{a}''$
4.	There exist $\boldsymbol{a}$ and $\boldsymbol{a}'$ in $\mathscr{A}$ such that $\boldsymbol{a}' \succ \boldsymbol{a}$
5. State Independence	For all $\boldsymbol{a} \in \mathscr{A}$ and any two distribution $\boldsymbol{p}$ and $\boldsymbol{q}$ on Z If $(\boldsymbol{a}(1, \cdot),\ldots, \boldsymbol{a}(\theta-1, \cdot), \boldsymbol{q}, \boldsymbol{a}(\theta+1, \cdot),\ldots, \boldsymbol{a}(n_\Theta, \cdot)) \succ (\boldsymbol{a}(1, \cdot),\ldots, \boldsymbol{a}(\theta-1, \cdot), \boldsymbol{q}, \boldsymbol{a}(\theta+1, \cdot),\ldots, \boldsymbol{\alpha}(n_\Theta, \cdot))$ for some state θ, then for all non-null θ' $(\boldsymbol{a}(1, \cdot),\ldots, \boldsymbol{a}(\theta'-1, \cdot), \boldsymbol{q}, \boldsymbol{a}(\theta'+1, \cdot),\ldots, \boldsymbol{a}(n_\Theta, \cdot)) \succ (\boldsymbol{a}(1, \cdot),\ldots, \boldsymbol{a}(\theta'-1, \cdot), \boldsymbol{q}, \boldsymbol{a}(\theta'+1, \cdot),\ldots, \boldsymbol{a}(n_\Theta, \cdot))$

principle. Axiom 4 is a structural condition requiring that not all states be null.

The first three axioms guarantee that there is a linear function $f: \mathscr{A} \to \mathscr{R}$ that represents the preference relation $\succ$. It also follows that the first three axioms are necessary and sufficient for the existence of real valued functions $u_1,\ldots, u_k$, such that:

$$\boldsymbol{a} \succ \boldsymbol{a}' \Leftrightarrow \sum_\theta \sum_z u_\theta(z)\boldsymbol{a}(\theta, z) > \sum_\theta \sum_z u_\theta(z)\boldsymbol{a}'(\theta, z). \quad (6)$$

In this representation, the utility of consequences is specific to the state of the world. Also, each utility is defined up to a state-specific change of scale. This representation is consistent with expected utility maximization, but does not provide a unique decomposition of preferences into probability of states of the world and utility of consequences. Such decomposition is achieved by the State Independence axiom, which asks the DM to consider two comparisons: in the first the two actions are identical except that in state θ one has probability distribution over consequences $\boldsymbol{p}$ and the other $\boldsymbol{q}$. In the second the two actions are again identical, except that now in state θ' one has probability distribution $\boldsymbol{p}$ and the other $\boldsymbol{q}$. If the DM prefers the action with probability distribution $\boldsymbol{p}$ in the first comparison, this preference will have to hold in the second comparison as well. So the preference for $\boldsymbol{p}$ over $\boldsymbol{q}$ is independent of the state. This axiom guarantees separation of utility and probability. The main theorem about representation of preferences via utilities and probabilities is then:

Axioms 1–5 are necessary and sufficient for the existence of a nonconstant function $u: Z \to \mathscr{R}$ and a probability distribution π on Θ such that

$$\boldsymbol{a} \succ \boldsymbol{a}' \Leftrightarrow \sum_\theta \pi(\theta) \sum_z u(z)\boldsymbol{a}(\theta, z) > \sum_\theta \pi(\theta) \sum_z u(z)\boldsymbol{a}'(\theta, z) \quad (7)$$

Moreover, the probability distribution π is unique, and u is unique up to a positive linear transformation.

In particular, for simple actions the summation over z drops out and the expectations above correspond to $\mathscr{U}_\pi(a)$ and $\mathscr{U}_\pi(a')$.

Axiomatic systems such as Savage's or Anscombe and Aumann's provide the logical foundation for decision making using expected utility in both decision analysis and statistical decision theory. The appropriateness and generality of these axiom systems are at the center of a debate which is critical for statistical practice. Kadane et al. (1999) investigate the implications and limitations of various axiomatizations, and suggest possible generalizations.

4. *Bayesian Statistical Decision Theory*

Statistical decision theory studies applications in which there is the possibility of observing the outcome x of an experiment that has probability distribution $p(x \mid \theta)$, conditional on state of the world θ. Observing x can give information about θ, and therefore assist in decision making. Bayesian decision theory offers (a) approaches for making efficient use of the information provided by x in choosing an action and (b) approaches for measuring the value of an experiment and for deciding whether or not to perform it when it comes at a cost. All decision theoretic Bayesian approaches can be characterized by specifying $\pi(\theta)$, $p(x \mid \theta)$, $L(\theta, a)$, and possibly the cost of experimentation.

Approaches to (a) are based on the notion that the distribution used in computing $\mathscr{U}$ at any stage in the decision process should make use of all knowledge available at the time of the decision, and that knowledge is incorporated by conditioning, that is using Bayes rule. Specifically, after observing data x, the

prior distribution $\pi(\theta)$ is updated to obtain a posterior distribution,

$$\pi(\theta \mid x) = \frac{\pi(\theta)p(x \mid \theta)}{\int_\Theta \pi(\theta)p(x \mid \theta)\,\mathrm{d}\theta}. \tag{8}$$

Then, as in definition (2), the Bayes action is found by maximizing expected utility. The difference now is that the relevant distribution for computing the expectation will not be $\pi(\theta)$ but $\pi(\theta \mid x)$. Formally, the data-dependent Bayes rule is

$$a_{\mathrm{B}}(x) = \operatorname{argmax} \mathscr{U}_{\pi_x}(a) = \int_\Theta u(a(\theta))\pi(\theta \mid x)\,\mathrm{d}\theta \tag{9}$$

where π_x is shorthand for the posterior distribution. The action that minimizes this expression will depend on x. So this procedure implicitly defines a decision function, that is a mapping from data to actions.

This approach has profound implications on statistical practice (Berger 1985, Bernardo and Smith 1994). Here we review some of the salient points.

Generality. Bayesian decision theory can be used to address statistical inference problems irrespective of data type, context, likelihood function, and current knowledge. Guidance is based on basic principles that are simple to state, understand, and refute if deemed inappropriate. Obstacles to implementation can occur, and simplifying assumptions are almost always needed in practice. However, in statistical applications, difficulties seldom arise because the conceptual framework is not sufficiently powerful.

Openness. All elements needed to define a statistical approach are explicitly and quantitatively described as prior, likelihood, loss, and cost of experimentation. This description is explicit, and common to all applications.

Likelihood principle. In a Bayesian decision theoretic approach, observed experimental evidence affects decision making only to the extent to which it is captured in the posterior $\pi(\theta \mid x)$, or equivalent by the likelihood $p(x \mid \theta)$. If there exist sufficient statistics, they alone need to be recorded for decision making purposes. Berger and Wolpert (1988) elaborate on this principle, often called the likelihood principle, or the sufficiency principle. A controversial example of experimental evidence that does not affect the likelihood, but does affect frequentist inference, is provided by stopping rules that depend on the data but are independent of the state of nature (Berger 1985, Bernardo and Smith 1994). See also *Likelihood in Statistics* and *Statistical Sufficiency*.

Nuisance parameters. In statistical inference nuisance parameters, that is parameters that are required to specify a realistic population model, but are not the primary object of inquiry, have created substantial controversy and have hampered the development of general frequentist approaches (for example the so-called most powerful tests are not generally applicable with nuisance parameters). From a decision theoretic standpoint, a nuisance parameter is a parameter that appears in the sampling distribution but not in the loss function. For example, let $\theta = (\gamma, \eta)$, where η is the nuisance parameter, and γ the parameter entering the loss function, that is $L(\theta, a) = L(\gamma, a)$ for every θ. The expected loss can be rewritten as

$$\begin{aligned}\mathscr{L}_{\pi_x}(a) &= \int_\Gamma L(\gamma, a(x)) \int_{\mathrm{H}} \pi(\gamma, \eta \mid x)\,\mathrm{d}\eta\,\mathrm{d}\gamma \\ &= \int_\Gamma L(\gamma, a(x))\pi(\gamma \mid x)\,\mathrm{d}\gamma. \end{aligned} \tag{10}$$

The nuisance parameter η is handled by marginalization of the posterior distribution. While this is a trivial corollary of the expected utility principle, it provides a completely general and unified approach to statistical analyses in presence of nuisance parameters.

Sequential decisions. In multistage decision problems, successive updating of beliefs using Bayes' rule permits modeling of accruing evidence, and also prediction of future states of knowledge. This, in conjunction with backwards-induction algorithms, has constituted the basis for developing complex decision-making strategies that adapt to new evidence as it accrues (DeGroot 1970). See also *Sequential Statistical Methods*.

5. *Standard Statistical Problems as Decision Problems*

5.1 *Hypothesis Testing*

Hypothesis testing problems have traditionally been among the earliest to be approached as decision problems (Neyman and Pearson 1933). The action space usually includes endorsing the null hypothesis, specified as a subset of Θ, endorsing an alternative hypothesis, and possibly no endorsement or the requirement that more experimentation be carried out.

In the simplest formulation, Θ consists of only two points: θ_0 (the null hypothesis), and θ_1 (the alternative hypothesis), with prior probability $\pi(\theta_0) = \pi_0 = 1 - \pi(\theta_1)$. The action space includes action a_0: 'endorse the null hypothesis' and a_1: 'endorse the alternative hypothesis.' This formulation, in both Bayesian and frequentist decision theory, is in contrast with Fisher's significance testing, which is concerned with measuring the extent to which data support a null hypothesis, separate from the consequences of rejecting that hypothesis.

Table 2
Loss function for testing a simple hypothesis against a simple alternative

	a_0	a_1
θ_0	0	L_0
θ_1	L_1	0

All loss functions in this case can be described by two-way tables as in Table 2. The zeroes on the diagonal are because loss functions are constructed so that there is at least one action with zero loss for every state of the world.

The expected losses of the two actions are:

$$\mathscr{L}_\pi(a_0) = L_1\pi(\theta_1 \mid x)$$
$$\mathscr{L}_\pi(a_1) = L_0\pi(\theta_0 \mid x)$$

and the Bayes rule is a_0 whenever

$$L_1\pi(\theta_1 \mid x) < L_0\pi(\theta_0 \mid x) \tag{11}$$

and a_1 otherwise, a formal restatement of Condorcet's opinion, discussed in Sect. 1.

Rewriting the inequality above via Bayes' rule gives

$$\frac{p(x \mid \theta_0)}{p(x \mid \theta_1)} > \frac{L_1}{L_0}\frac{\pi_1}{\pi_0}. \tag{12}$$

The left-hand-side is the so-called Bayes factor in favor of the null hypothesis. It measures the strength with which the experimental evidence supports the null hypothesis as compared to the alternative, and it summarizes the data for the purpose of this decision. The Bayes rule is to accept the null hypothesis when the Bayes factor is sufficiently large. The cutoff is determined based on the prior probability of the null hypothesis and the relative disutilities of the two possible incorrect decisions.

In this simple formulation, there is a close analogy between the decision function induced by the Bayes rule and the most powerful tests (Neyman and Pearson 1933), which are in fact of the form (12). In frequentist formulations, however, the cutoff is chosen by controlling the significance level, that is the probability of incorrectly endorsing the hypothesis in repeated experiments. In practice, this is all too frequently set to 0.05 without consideration of the consequences of incorrect calls; however, original formulations of the theory demanded this consideration.

More pronounced differences between Bayesian and frequentist hypothesis testing emerge in less stylized formulations. For example, if θ is a real valued parameter and the null hypothesis is $\theta \in \Theta_0$, say $\Theta_0 = (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$, the Bayes rule becomes

$$L_1(1 - \pi(\Theta_0 \mid x) < L_0\pi(\Theta_0 \mid x) \tag{13}$$

or

$$\frac{\int_{\Theta_0} \pi(\theta)p(x \mid \theta)\,\mathrm{d}\theta}{\int_{\Theta\setminus\Theta_0} \pi(\theta)p(x \mid \theta)\,\mathrm{d}\theta} > \frac{L_1}{L_0} \tag{14}$$

which preserves the general flavor of Condorcet's view, but is not consistent with typical frequentist approaches, except in very special cases.

5.2 Binary Classification

Binary classification problems (Duda et al. 2001) consider assigning an individual to one of two categories, by measuring a series of attributes. An example is medical diagnosis for a single medical condition (say disease vs. no disease) based on a battery of tests. Again the action space has two elements: (say a_1: 'diagnosis of disease' and a_0: 'diagnosis of no disease'), and the structure of the loss function is analogous to Table 2.

A typical situation occurs when the DM observes data on the attributes x and correct disease classification y of a sample of individuals, and wishes to classify an additional individual, randomly drawn from the same population. If x^* are the observed features of the new individual to be classified, and y^* is its unknown disease state, then the Bayes rule is to choose a diagnosis of no disease if

$$L_1\pi(y^* = 1 \mid y, x, x^*) < L_0\pi(y^* = 0 \mid y, x, x^*) \tag{15}$$

or

$$\frac{\int_\Theta p(y^* = 0 \mid x^*, \theta)\pi(\theta)p(y, x \mid \theta)\,\mathrm{d}\theta}{\int_\Theta p(y^* = 1 \mid x^*, \theta)\pi(\theta)p(y, x \mid \theta)\,\mathrm{d}\theta} > \frac{L_1}{L_0}. \tag{16}$$

This classification rule incorporates inference on the population model, uncertainty about population parameters, and relative disutilities of misclassification errors.

5.3 Point Estimation

Statistical point estimation assigns a single best guess to an unknown parameter θ. This assignment can be

viewed as a decision problem in which the space of actions is $\mathscr{A} = \Theta$. The loss function quantifies the consequences of estimation errors. The most commonly used, primarily for its mathematical convenience, is the squared error loss $L(\theta, a) = (\theta - a)^2$. The expected loss is of an action a is

$$\int_{\Theta} (\theta - a)^2 \pi(\theta \mid x)\, \mathrm{d}\theta = \mathrm{Var}(\theta \mid x) + (a - E(\theta \mid x))^2 \tag{17}$$

so that the Bayes rule is $a_{\mathrm{B}} = E(\theta \mid x)$, that is the mean of the posterior distribution. The expected loss associated with it is the variance of the posterior distribution.

Other commonly used loss functions include the absolute error loss $L(\theta, a) = |\theta - a|$, leading to the median of $p(\theta \mid x)$, and, for discrete Θ, the loss $L(\theta, a) = I_{\theta = a}$, assigning the same loss to all errors, and leading to the mode of the posterior distribution.

6. *The Value of Information*

One of the strengths of the expected utility approach is the simple answer provided to the problem of assessing the value of prospective information. From the point of view of decision making, information is anything that may enable a better decision, that is a decision with a higher expected utility. The expected change in utility resulting from experimentation can therefore be used as a quantitative measure of the worth of an experiment (Ramsey 1990, Raiffa and Schleifer 1961, DeGroot 1984). See also *Experimental Design: Bayesian Designs*.

In its simplest formulation, experimentation can be modeled as a two-stage decision process: in the first stage the DM decides whether to perform the candidate experiment or not; in the second the DM chooses an action a, possibly making use of the experimental outcome. This decision problem is most effectively solved in reversed order, or by backwards induction (DeGroot 1970), by first determining the Bayes action $a_{\mathrm{B}}(x)$ and its expected utility $\mathscr{U}_{\pi_x}(a_{\mathrm{B}}(x))$ as a function of the data, for every possible outcome, and then stepping back and comparing the two options: 'experiment' and 'no experiment.'

Specifically, under 'no experiment,' the DM would choose action a_{B} immediately, with expected utility

$$V(\pi) = \sup_a \mathscr{U}_{\pi}(a) \tag{18}$$

Under 'experiment,' if $a_{\mathrm{B}}(x) = a_{\mathrm{B}}$ for every outcome x, the experiment does not contribute to the solution of the decision problem at hand. An experiment has value only if it may affect subsequent decisions. How large this value is depends on the utilities assigned to the consequences, and the frequency with which the observations lead to a change of decision. Formally, under 'experiment' the expected utility, averaged with respect to the states of the world and the yet unobserved experimental outcomes, is

$$\int_{\mathscr{X}} \sup_a \mathscr{U}_{\pi_x}(a)\, \mathrm{d}x = \int_{\mathscr{X}} V(\pi_x)\, \mathrm{d}x. \tag{19}$$

The difference

$$V = \int_{\mathscr{X}} V(\pi_x)\, \mathrm{d}x - V(\pi) \tag{20}$$

is the amount of utility the DM expects to gain from the experiment. V provides a pragmatic, decision-based way of measuring the value of information and subsumes a variety of specific statistical approaches as special cases. If the experiments comes at a cost, measured by c in the utility scale, it is optimal to perform it whenever $V > c$. This logic can be extended to an arbitrary number of experimental stages (DeGroot 1970), although computational obstacles arise rapidly as the number of stages increases. See also *Sequential Statistical Methods*. The function V is convex, that is

$$V(\pi) = V\left(\int_{\mathscr{X}} \pi_x\, \mathrm{d}x\right) \leqslant \int_{\mathscr{X}} V(\pi_x)\, \mathrm{d}x. \tag{21}$$

This insures that, in expectation, experimentation can never decrease expected utility.

7. *Risk and Admissibility*

Statistical decision theory following frequentist approaches focuses on the performance of decision functions over repeated experiments, for a fixed state of the world, as measured by the risk function

$$R(\theta, a) = \int_{\mathscr{X}} L(\theta, a(x) p(x \mid \theta)\, \mathrm{d}x. \tag{22}$$

For example in point estimation using squared error loss, the risk is the so-called mean squared error, decomposable as the sum of the variance of the decision rule and the square of the bias. See also *Decision Theory: Classical*. The expectation of R with respect to a prior π is often termed Bayes risk. Minimization of the Bayes risk, under very general conditions, leads again to the Bayes rule (Berger 1985).

A critical rationality condition for the evaluation of decision rules from a frequentist standpoint is admissibility. The basis of admissibility is a dominance relation among decision rules. Specifically, decision

rule a is R-better than a' if $R(\theta, a) \leqslant R(\theta, a')$, for every θ, and $R(\theta, a) < R(\theta, a')$ for at least some θ. A decision rule a is admissible if there is no R-better rule. Restricting to admissible rules is critical, although it will often leave many rules to choose from, and the DM will need to commit to a more stringent principle to reach a decision.

Extensive statistical investigation has focused on exploring the admissibility properties of various statistical strategies. Several commonly used approaches, such as maximum likelihood and minimax decision rules, can lead to inadmissible estimates in some circumstances (Stein 1955, Berger 1985). From the point of view of Bayesian decision theory, the interesting connection is that, under relatively general conditions, Bayes rules are admissible. For example, any unique Bayes estimator is admissible (Schervish 1955). Also, in some specific applications, all admissible rules are Bayes rules with respect to some prior distribution (as in the case of finite Θ), or limiting cases of families of Bayes rules.

8. Conclusions

Bayesian statistical decision theory studies the optimal use of experimental information in decision making. It can be applied to specific decision problems as well as statistical methodology as a field, as most statistical analyses can be viewed as decision problems. In this discussion we have reviewed history, axiomatic foundations, and statistical implications of Bayesian decision theory. For further readings, Lindley (1985) presents the key concepts without complex mathematical formalisms, while Berger (1985), Robert (1994), and Bernardo and Smith (1994) provide more recent overviews and comparisons with alternative approaches.

Bibliography

Anscombe F J, Aumann R J 1963 A definition of subjective probability. *The Annals of Mathematical Statistics* **34**: 199–205

Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Springer, New York

Berger J O, Wolpert R L 1988 *The Likelihood Principle*, 2nd edn. Institute of Mathematical Statistics, Hayward, CA

Bernardo J M, Smith A F 1994 *Bayesian Theory*. Wiley, New York

Bernoulli D 1738 Specimen theoriae novae de mensura sortis. *Commentarii academiae scientiarum imperialis Petropolitanae* **5**: 175–92

Condorcet M J A N d C 1785 *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix*. Imprimerie Royale, Paris

de Finetti B 1937 La prévision ses lois logiques, ses sources subjectives. *Annales de l' Institût Henri Poincaré (Paris)* **7**: 1–68

DeGroot M H 1970 *Optimal Statistical Decisions*. McGraw-Hill, New York

DeGroot M H 1984 Changes in utility as information. *Theory and Decision* **17**: 287–303

Duda R O, Hart P E, Stork D G 2001 *Pattern Classification*. Wiley, New York

Fisher R A 1925 Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**: 700–25

Jorland G 1987 The St Petersburg Paradox 1713–1937. In: Kruger L, Daston L J, Heidelberger M (eds.) *The Probabilistic Revolution: Volume 1, Ideas in History*. MIT Press, Boston, MA

Kadane J B, Schervish M J, Seidenfeld T 1999 *Rethinking the Foundations of Statistics*. Cambridge University Press, New York

Kreps D M 1988 *Notes on the Theory of Choice*. Westview Boulder, CO

Lindley D V 1985 *Making Decisions*. Wiley, New York

Neyman J, Pearson E S 1933 On the problem of the most efficient test of statistical hypotheses. *Philosophical Transaction of the Royal Society, Series A* **231**: 286–337

Raiffa H, Schleifer R 1961 *Applied Statistical Decision Theory*. Harvard University Press, Boston

Ramsey F P 1931 *Truth and Probability*. Keegan Paul, London

Ramsey F P 1990 Weight or the value of knowledge. *British Journal for the Philosophy of Science* **41**: 1–4

Robert C P 1994 *The Bayesian Choice: A Decision Theoretic Motivation*. Springer-Verlag, New York

Savage L J 1954 *The Foundations of Statistics*. Wiley, New York

Schervish M J 1995 *Theory of Statistics*. Springer-Verlag, New York

Stein C 1955 Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the 3rd Berkeley Symposium on Mathematica, Statistics and Probabilities* **1**: 197–206

von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Wiley, New York

Wald A 1949 *Statistical Decision Functions*. Wiley, New York

G. Parmigiani

Decision Theory: Classical

The term 'statistical decision theory' is a condensation of Abraham Wald's phrase, 'the theory of statistical decision functions' which occurs, for example, in the preface to his monograph, Wald (1950). Wald viewed his 'theory' as a codification and generalization of problems of estimation of the theory of tests and confidence intervals already developed by Neyman, often in collaboration with E. Pearson. For clear statements of Wald's view see the last two paragraphs of the introduction to Wald's pivotal paper, Wald (1939), and Sect. 1.7 of his monograph, cited above.

Because of this history the term 'statistical decision theory' can be understood as referring to an important group of specific theorems due to Wald and later extended and refined by others, as discussed below; or it can be understood in a much broader sense as a description of the frequentist outlook inherent in the work of Neyman, Pearson, and Wald, among others.

This article focuses on the former point of view, which is undoubtedly the more widely held interpretation. For a discussion of the latter point of view see Brown (2000).

1. Key Elements

1.1 Basic Concepts

The key elements that set formal decision theory apart from the remainder of statistical theory are the notions of 'loss' and 'risk' and the consequent concepts of 'admissibility' and 'minimaxity.' Consider a statistical problem with a sample space, Ξ and a family of possible distributions, $\{F\colon F \in \Phi\}$. These elements should be part of any carefully specified statistics problem. Often Φ is parametrized so that it can be written as $\Phi = \{F_\theta\colon \theta \in \Theta\}$. Θ is then called the 'parameter space,' problems in which Φ cannot be nicely parametrized by a subset of Euclidean space are often referred to as nonparametric.

In addition, Decision Theory requires specification of a set of possible decisions or actions, Δ, to be taken after observation of the data. Possible types of decisions include estimates of the parameter, hypothesis tests related to subsets of parameters, confidence intervals, or predictions about potential future observations. A 'loss function' describes the loss—or negative advantage—of taking a particular action, $d \in \Delta$, when $F \in \Phi$ is the true distribution. The resulting function can be written as $L(F, d)$, or as $L(\theta, d)$ when the 'states of nature' in Δ have been given a parametric representation. Usually, $L \geqslant 0$. A nonrandomized procedure is a map $\delta\colon \Xi \to \Delta$. (Randomized procedures can also be defined. They are useful for developing parts of the theory, even if they are almost never used in practice.) The 'risk,' R, is the expected loss of δ under F

$$R(F, \delta) = E_F(L(F, \delta(X)) \tag{1}$$

These basic notions lead naturally to a comparison of potential decision procedures that defines δ_1 as 'better than' δ_2 if

$$R(F, \delta_1) \leqslant R(F, \delta_2) \quad \textit{for all } F \in \mathscr{F}, \quad \textit{with strict inequality for some } F \tag{2}$$

The admissible procedures are those that are minimal in the sense of this partial ordering; thus, δ is admissible if there is no other procedure that is better.

1.2 Example: Inference about a Binomial Proportion

Inference about a binomial proportion is an important problem, in terms of both basic theory and statistical practice. We only discuss here the most basic version of this problem. In this the statistician observes the value of a binomial variable reflecting an unknown proportion, p.

One common objective involves estimation of p. (In terms of the preceding formal definitions, $\Xi = \{0, 1, \ldots, n\}$ and $\Delta = \{d\colon 0 \leqslant d \leqslant 1\}$ and $\Phi = \{F_p\colon 0 \leqslant p \leqslant 1\}$ with F_p indicating the binomial distribution with sample size n and proportion p.) For this objective a conventional loss function is squared error loss, defined by $L(p, d) = (d-p)^2$. See Sect. 3.2 for some further discussion of this example.

Another common objective involves testing a null hypothesis about p. The null hypothesis may be, for example, that $p = p_0$ for some fixed, pre-specified value p_0. Then the two possible decisions correspond to the statements 'do not reject H_0' and 'reject H_0.' The conventional loss function here ascribes loss 0 to a correct decision and 1 to an incorrect one. (See *Longitudinal Data*.) A related problem concerns the construction of confidence intervals for p. There is an intimate relation between this and the testing problem. See, for example, Lehmann (1986) or Robert (1994 Chap. 5).

1.3 Example: Estimation of Normal Means

Another important paradigm in both theory and practice involves the estimation of one or more normal means. In the simplest version of this problem the statistician observes the sample mean, $\bar{X}$ of a normal sample. In variants of the problem the standard deviation of the population may either be known or unknown. The first variant is easier to treat mathematically but the second is more common in applications. If the objective is to estimate the unknown population mean, say, μ, then a conventional loss function is squared error loss given by $L(d, \mu) = (d-\mu)^2$, for an estimate d being any real number. Testing and confidence interval problems are also of interest for such a family of distributions.

Many generalizations of this formulation are of practical importance, and their theory has been extensively researched. The most prominent situation includes the normal linear model. This setting includes such common statistical topics as the Analysis of Variance. In Sect. 3.1 of this article we discuss in more detail the situation that arises when one looks at several independent normal means simultaneously.

1.4 Discussion

For some authors, the law of large numbers interpretation of the expectation in (1) has suggested that the statistician contemplates a situation in which the identical experiment will be independently repeated an arbitrarily large number of times. Moreover, for the law of large numbers to apply one needs formally also that the same value of F appears in each of these repetitions. Such an interpretation has been on occasion criticized as impractical, and as invalidating the

premises of decision theory. However this interpretation is not required as a logical basis for the theory. It was also apparently not contemplated in the work of Neyman and Pearson that motivated Wald's development.

One can instead consider a collection of independent experiments, each having its own structure. A statistician who uses a collection of procedures many of which are inadmissible will not fare well in the long run by comparison to a statistician who uses a better procedure for each individual problem. One may find early reflections of this attitude in Clopper and Pearson (1934) or in Neyman and Pearson (1933). (It can be argued further that it would be still better to adopt a coherent collection of admissible procedures in such a situation. See for example, Savage (1972 [1954]), who proposes subjective Bayesian methods as the means toward this end.) (See *Bayesian Statistics*.)

Typically, in applications much care is given to realistic specification of the family of distributions in this formulation. When this cannot be done with satisfactory precision then robust methods are often recommended. Such methods can be placed directly within the above framework by considering a suitably large set for Φ. Huber (1981) has a careful treatment of this nature and proposes an often-used robust procedure with important decision theoretic properties.

The action space, Δ, is also generally subject to careful and realistic specification. The loss function, on the other hand, is not so often specified with the same degree of precision and attention to the practical situation at hand. It is common to adopt certain conventional loss functions, most prominently squared error or normalized squared error in real valued estimation problems or squared Euclidean distance in vector valued ones. In hypothesis testing problems a zero-one loss is similarly conventional and predominantly used. It has sometimes been argued that statisticians should expend more care on realistic specification of the loss or on a robust treatment of families of losses. See Robert (1994, Sect. 2.6) for more discussion of this issue.

2. Classical Themes

2.1 Characterization Theorems

A major objective of classical decision theory is to characterize the admissible procedures, or special subsets of them, in a mathematically tractable fashion. Many of these characterizations involve the notion of a Bayes procedure. Corresponding to any 'prior' probability measure, G, on the parameter space a Bayes procedure can be defined either as a procedure that minimizes the average risk with respect to G or, equivalently, as one that minimizes the conditional risk given the observation, for almost every possible observation.

A basic result in Wald's work, later developed further by others, is that a Bayes procedure is automatically admissible if it is essentially uniquely determined. It is also admissible under mild additional conditions guaranteeing continuity of risks if the prior is supported on the entire parameter set. Wald also proved a fundamental converse result: if the parameter space is finite then every admissible procedure is Bayes. He, and others, also extended this result to classes of settings where the parameter space is bounded. See Robert (1994, Sect. 6.3) for a general discussion of such results.

Here, as elsewhere, Wald acknowledged his debt to earlier work on hypothesis testing by noting that these facts about Bayes procedures could provide an alternate interpretation and an alternate proof for the Neyman–Pearson Lemma, and for various generalizations to multiple classification problems. As Wald also frequently noted, these estimators are Bayes only in a formal mathematical sense in that the decision theoretic motivation for their use is frequentist in nature (via the risk function) rather than fundamentally Bayesian (via the conditional expected loss given the data).

If the parameter space is not compact then one cannot expect that all admissible procedures will be Bayes for some prior. However there are a variety of characterizations in such settings of admissible procedures as types of limits of Bayes procedures. In some cases—usually involving exponential families—these characterizations can be extended to prove that all (or most) admissible procedures can be represented as generalized Bayes procedures. That is to say, they can be represented as formally calculated posterior conditional expectations with respect to non-negative measures that are not necessarily finite.

In some special problems more can be proven. The most notable case is the problem of estimating under squared error loss the mean of a multivariate normal distribution having known covariance matrix. There it is also possible to give an explicit characterization of the generalized priors that correspond to admissible procedures. Such a result relies in part on a method for proving admissibility due to Blyth and developed further by Stein. Lehmann and Casella (1998, Chap. 5) discuss many of these results and provide additional references.

2.2 Minimaxity

Minimaxity is another fundamental decision theoretic property defined and treated in Wald's seminal work. A procedure δ^* is minimax if

$$\sup_{F \in \mathbf{F}} R(F, \delta^*) = \inf_{\delta} \sup_{F \in \mathbf{F}} R(F, \delta).$$

Examination of this formal definition shows that minimaxity is a very cautious, conservative criterion.

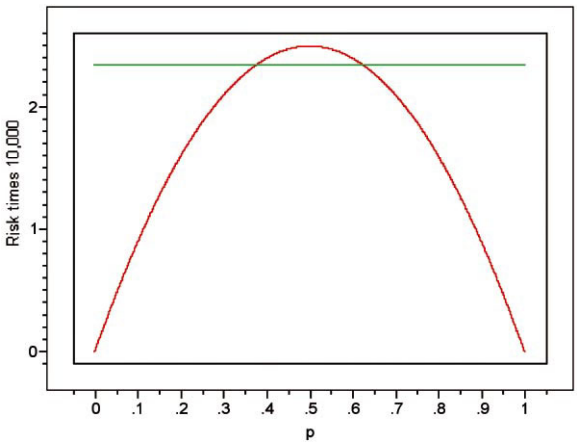

Figure 1
Comparative risks for $n = 1{,}000$ of usual and minimax estimators of a binomial proportion. (The constant line is the risk of the minimax estimator)

A minimax procedure is the one that does best *if* the true $F \in \Phi$ turns out to be the worst possible from the statistician's perspective. In some special problems the minimax procedure provides a broadly satisfactory choice for a decision rule. In others it turns out that minimax rules are not admissible, or are not satisfactory because other rules have more appealing risk functions. A simple classical example may make clear how this can happen. Consider the binomial estimation problem described in Sect. 2.2. The 'usual' estimate of p is the sample proportion $\hat{p} = X/n$. It can be shown that this is not minimax under squared error loss. Instead, the minimax estimator has the somewhat more complicated formula $p^* = (X + \sqrt{n}/2)/(n + \sqrt{n})$. The respective risks of these estimators can be plotted. See Fig. 1, which plots these risks for the moderately large value, $n = 1{,}000$. Examination of these plots makes clear that $\hat{p}$ would generally be preferable to p^* in this setting, since its risk is usually smaller and often much smaller. (The exceptions would be situations where the statistician was nearly convinced that p was quite near 0.5.)

Even in cases where the minimax procedure itself may not be satisfactory, comparison with the minimax risk can provide a useful starting point for judging the worth of proposed procedures. The essays in Brown (1994) and Strawderman (2000) comment on the role of minimaxity as a benchmark for evaluating and developing alternate procedures.

2.3 Sequential Decision Theory

At the same time as he was developing statistical decision theory, Wald was also building a theory of sequential decision procedures. Wald (1947), his monograph on the topic, appeared three years before his monograph on statistical decision functions. It can be argued that there is nothing intrinsically decision theoretic in the idea of sequential stopping that is the basis for this theory. However the structure that Wald created was so thoroughly imbued with the spirit of decision theory that it has since usually been viewed as a part of it. Sequential procedures allow the experimenter to conduct the experiment in stages with a decision to stop the experiment or to continue made after each stage. In a decision theoretic version the size of the stopped sample is included as part of the overall loss function.

The most dramatic result of Wald's research here is the optimal property of the Sequential Probability Ratio Test (SPRT) in Wald and Wolfowitz (1948). The proof of this optimality property relies on a representation of the SPRT as a Bayes procedure, and uses much of the decision theoretic machinery elsewhere developed by Wald. The most prominent modern embodiment of this sequential idea undoubtedly lies in the area of sequential clinical trials. Here the spirit of Wald's work remains as an important foundation, even if the subject has mostly gone in more numerical and pragmatic directions. See Siegmund (1994) for a contemporary overview of sequential decision theory and Jennison and Turnbull (1999) for a recent comprehensive treatment of clinical trials. (See also *Sequential Statistical Methods*.)

2.4 Hypothesis Testing

The basic theory of hypothesis testing was largely developed by Fisher and by Neyman and Pearson in the decades before Wald began his codification of decision theory. It can be argued that Fisher's view of testing was not fundamentally decision theoretic. Fisher's formulation requires a specification of the distributions in the null hypothesis and of a test statistic, but it does not require specification of the possible distributions in the alternative. Neyman and Pearson proceeded further to require these to be specified, and this enabled them to study the competing requirements of high power under the alternative and low power under the distributions in the null hypothesis. (This power should be bounded by the significance level, α.) The Neyman–Pearson theory is clearly decision theoretic and, as noted above, can be cast exactly into the mold created by Wald. Subsequent work of Lehmann, Stein and others has more fully developed the theory. Lehmann (1986) contains a definitive high-level treatment of the theory and selected applications.

2.5 Invariance

Statistical invariance is the embodiment of an intuitive demand. The invariance principle applies when the decision theoretic elements of the problem are in-

variant under the action of a group of one-to-one transformations. (Some authors use the term 'equivariant' for parts of this specification.) In that case the principle demands that the decision procedure adopted also be invariant (or equivariant) under the action of this group. Although the principle seems reasonable, and the resulting procedures often have intuitive appeal, it turns out that they need not be admissible. See below for a brief discussion of the Stein phenomenon, a situation where appealing invariant procedures are inadmissible.

There are structural conditions on the group that can guarantee that there do exist invariant procedures that are minimax. The first such theorem of this type is contained in an unpublished manuscript by Hunt and Stein. In the ideal case where there is a best invariant procedure these conditions then guarantee that this best invariant procedure is minimax. See Kiefer (1957) and Brown (1986a) for more information about invariance and the Hunt–Stein theorem. (See *Invariance in Statistics*; *Analysis of Variance and Generalized Linear Models*.)

2.6 Asymptotics

The basic concepts and tools of decision theory have also been carried into the domain of asymptotic statistical theory. Le Cam was the leader in this activity by formulating and developing such concepts as asymptotic sufficiency, contiguity, local asymptotic normality, and local asymptotic minimaxity. In Le Cam (1953) the seminal paper beginning his development, he included a careful treatment of the concept of asymptotic superefficiency, and a proof in parametric problems that it was limited to sets of Lebesgue measure zero. See Le Cam and Yang (2000) for a presentation of many of these topics, and for further references.

3. Contemporary Developments

3.1 Bayesian Decision Theory

Stein (1956) described the inadmissibility of the usual estimator of a three or more dimensional multivariate normal mean. This initially surprising discovery spurred activity in decision theory, changed many statistical attitudes and preconceptions, and continues to influence current statistical methodology. For estimating the vector of means from a multivariate normal sample under squared-error loss the usual procedure, $\bar{\mathbf{X}}$, is intuitive and has various nice theoretical properties, such as those of being best invariant and minimax. It is also the formal Bayes procedure corresponding to the uniform, Lebesgue measure as the formal prior. Note that Lebesgue measure is also Jeffreys prior for this problem, and hence corresponds to a standard 'noninformative' Bayesian prescription. See Jeffreys (1939) or Robert (1994, Sect. 3.4). This usual estimator is surprisingly not admissible when there are three or more normal means involved.

James and Stein (1961) intensified Stein's earlier discovery by describing an estimator that dominates the usual one by a significant amount. Brown (1971) later demonstrated that all admissible procedures in this problem are either Bayes or formal Bayes, and also characterized which of these priors leads to admissible estimators. Strawderman (1971) subsequently used a hierarchical Bayes construction to derive formal Bayes estimators for this problem that are both minimax and admissible. When the dimension is at least 5 then some of these are actually proper Bayes. Efron and Morris (1973) explained how Stein's estimator could be derived via an empirical Bayes approach. A combined effect of these developments is to demonstrate the importance of a careful choice of prior or formal prior for such multivariate problems and to provide some guidelines for the form of such priors. Ensuing research has led in the directions of empirical Bayes methods, of robust Bayes methods, and also of methods involving careful specification of hierarchical priors. The influence of Stein's decision-theoretic discovery on contemporary Bayesian analysis is discussed at length in such monographs as Berger (1985) and Robert (1994, especially Chaps. 6 and 8).

3.2 Unbiased Estimates of Risk

Stein's discovery has also led to other methodological and substantive advances in decision theory. The most important of these undoubtedly revolve around Stein's (1981) unbiased estimate of the risk. For generalizations and other related results see Brown (1986b, Chap. 4). Where it applies, as it does in the preceding multivariate normal problem, the unbiased estimate of the risk provides a powerful tool for finding procedures that dominate previous proposals and for suggesting new, satisfactory estimates.

3.3 Nonparametric Function Estimation

Much current research in nonparametric function estimation involves a synthesis of several aspects of statistical decision theory, including asymptotic decision theory, minimax theory, and results and methods related to the James–Stein phenomenon described above. A typical problem involves the observation of a nonparametric regression. Suppose for example that $\{x_i : i = 1, \ldots, n\}$ is a given subset of predictors, e.g., $x_i = i/n$, and $Y_i \approx N(m(x_i), \sigma^2)$ are independent normal variables with mean $m(x_i)$. The goal is to estimate the unknown function, m, under a loss function such as $\sum_i (\hat{m}(x_i) - m(x_i))^2$. The problem is nonparametric in that minimal assumptions are made about the form of m. For example, one might assume only that these

functions are in a given ball in a Sobolev space or a Besov space. Using a variation of Le Cam's asymptotic decision theory various papers have recently shown that several apparently distinct problems of this character are actually asymptotically isomorphic in a decision-theoretic sense. Consult Johnstone (2000) for references.

As a sort of culmination of much research in this area, Donoho et al. (1990) proved a minimax theorem for linear estimators in such a problem. These results, and later extensions, relate the minimax risk under a restriction to linear procedures to the usual, unrestricted minimax risk for the problem. In turn, this result can be used to describe the asymptotic behavior of the unrestricted minimax value. One consequence is a determination of the minimax asymptotic rate of convergence for the risk of estimators in this problem. Imbedded in the proofs is also a practical methodology, essentially due to earlier work of Ibragimov and Hasminskii, for deriving the minimax linear procedure.

3.4 Adaptive Estimation

A new variation of the minimax concept has come forward here in order to try to overcome the inherently conservative and global nature of minimaxity. This is the notion of adaptive minimaxity. Adaptive minimaxity is defined relative to a collection, X, of subsets of the possible distributions, Φ. A procedure is adaptively minimax to within a factor K if

$$\sup_{C \in \mathrm{F}} \frac{\sup_{F \in C} R(F, \delta^*)}{\inf_{\delta} \sup_{F \in C} R(F, \delta)} \leqslant K. \tag{3}$$

Procedures built from wavelet bases have been shown by Donoho, Johnstone and others to possess desirable asymptotic adaptive minimax properties. It should also be noted that both James–Stein estimation and Stein's unbiased estimate of risk play a key role here, and in further refinements in this area. Johnstone (2000) discusses many of these issues and provides additional references. Donoho and Johnstone have also developed a new form of decision-theoretic tool, the 'oracle inequality,' that has been instrumental in papers written in the early twenty-first century such as these.

Bibliography

Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Springer, NY

Brown L D 1971 Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Annals of Mathematical Statistics* **42**: 855–903 (Correction 1973 *Annals of Statistics* **1**: 594–96

Brown L D 1986a Commentary on paper [19]. *J. C. Kiefer Collected Papers, Supplementary Volume*. Springer, New York, pp. 20–27

Brown L D 1986b *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics. Hayward, CA

Brown L D 1994 Minimaxity, more or less. In: Gupta S S, Berger J O (eds.) *Statistical Decision Theory and Related Topics*. V. Springer, New York, pp. 1–18

Brown L D 2000 An essay on statistical decision theory. *Journal of the American Statistical Association* **95**: 1277–81

Clopper C J, Pearson E S 1934 The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**: 404–13

Donoho D L, Liu R C, MacGibbon B 1990 Minimax risk over. hyperrectangles, and implications. *Ann. Statist*. **18**: 1416–37

Efron B, Morris C 1973 Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association* **68**: 117–30

Huber P J 1981 *Robust Statistics*. Wiley, New York

James W, Stein C 1961 Estimation with quadratic loss. *Proc. Fourth Berk. Symp. Math. Statist., and Prob*. **1**: 311–19

Jeffreys H 1948,1961 [1939] *The Theory of Probability*. Oxford University Press, Oxford, UK

Jennison C, Turnbull B W 1999 *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall, Boca Raton, FL

Johnstone I M 2000 Function estimation in Gaussian noise: Sequence models. http://www-stat.stanford.edu/~imj/

Kiefer J 1957 Invariance, minimax sequential estimation, and continuous time processes. *Annals of Mathematical Statistics* **28**: 573–601

Le Cam L 1953 On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics* **1**: 277–330

Le Cam L, Yang G L 2000 *Asymptotics in Statistics: Some Basic Concepts*, 2nd edn. Springer, New York

Lehmann E L 1986 *Testing Statistical Hypotheses*, 2nd edn. Springer, New York

Lehmann E L, Casella G 1998 *Theory of Point Estimation*. Springer, New York

Neyman J, Pearson E S 1933 The testing of statistical hypotheses in relation to probabilities *a priori*. *Proceedings of the Cambridge Philosophical Society* **29**: 492–510

Robert C P 1994 *The Bayesian Choice: A Decision Theoretic Motivation*. Springer, New York

Savage L J 1972 [1954] *The Foundations of Statistics*. Dover Publications, New York

Siegmund D 1994 *Sequential Analysis*. Springer, New York

Stein C 1956 Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**: 187–95

Stein C 1981 Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**: 1135–51

Strawderman W E 1971 Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics* **42**: 385–88

Strawderman W E 2000 Minimaxity. *Journal of the American Statistical Association* 1364–8

Wald A 1939 Contributions to the theory of estimation and

testing hypotheses. *Annals of Mathematical Statistics* **10**: 299–326
Wald A 1947 *Sequential Analysis*. Wiley, New York
Wald A 1950 *Statistical Decision Functions*. Wiley, New York
Wald A, Wolfowitz J 1948 Optimal character of the sequential probability ratio test. *Annals of Mathematical Statistics* **19**: 326–39

L. D. Brown

Declarative Memory, Neural Basis of

Memory is a broad topic that has its roots in both biology and psychology. Many important aspects of the problem of memory involve systems-level questions that must be addressed at a relatively global, neuropsychological level of analysis. Two long-standing and inter-related questions have been of central interest since the 1960s. The first is whether there is more than one kind of memory. The second concerns what brain structures and pathways are involved in memory. Scientific work on these issues has introduced the term declarative memory to neuroscience and identified a number of parallels between the neural organization of memory in humans and animals.

It has turned out that the brain is organized such that memory is a distinct and separate cognitive function, which can be studied in isolation from perception and other intellectual abilities. Significant information about how memory is organized has come from two experimental approaches. First, the study of patients with memory disorders (amnesia) has been enormously useful. Second, information about the anatomy of memory has come mainly from the kind of systematic and cumulative work that is possible only in experimental animals.

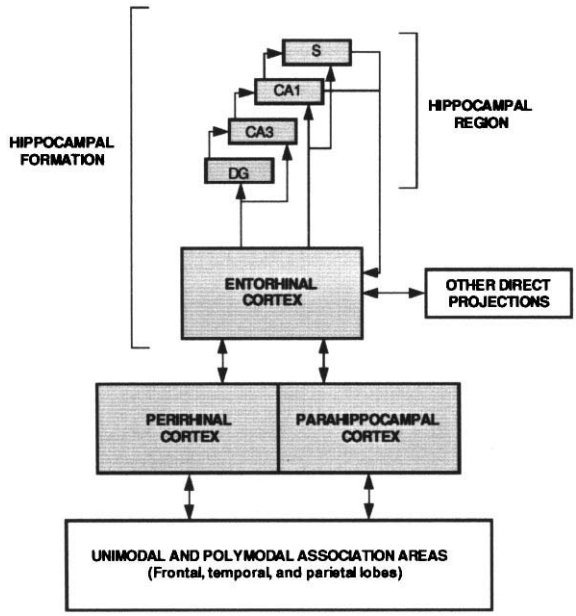

Figure 1
Schematic view of the medial temporal lobe memory system. The perirhinal and parahippocampal cortices receive projections from unimodal and polymodal areas in the frontal, temporal, and parietal lobes. In turn, the perirhinal and parahippocampal cortices account for nearly two-thirds of the cortical input to the entorhinal cortex, the primary source of cortical input to the hippocampal region (i.e., the cell fields of the hippocampus proper, the dentate gyrus, and the subicular complex). The hippocampal region together with the entorhinal cortex make up the hippocampal formation. The entorhinal cortex also receives other direct inputs from orbital frontal cortex, cingulate cortex, insular cortex, and superior temporal gyrus. All these projections are reciprocal

1. Amnesia and the Medial Temporal Lobe Memory System

Amnesia refers to difficulty in learning new information or in remembering the past. Amnesia can occur as a strikingly circumscribed memory impairment in the absence of other cognitive deficits (Scoville and Milner 1957). Patients with this disorder can have intact intelligence test scores, intact language and social skills, and an intact memory for a great deal of information from the remote past, especially childhood. In addition, immediate memory is intact. Indeed, patients can hold information in mind for several minutes, provided there is no distraction and the material can be rehearsed. Their intact immediate memory explains why amnesic patients can carry on a conversation and can appear quite normal to casual observation. The difficulty for individuals with amnesia arises when information must be recalled after a distraction-filled interval, or when recall is attempted directly after the presentation of an amount of information that exceeds immediate memory capacity (e.g., a list of eight or more items). In this situation, normal individuals will usually recall more items than amnesic patients.

An animal model of human amnesia was established in the monkey in the early 1980s (Mishkin 1982, Squire and Zola-Morgan 1983). Systematic and cumulative work during the next 10 years, using the animal model, succeeded in identifying the system of structures in the medial temporal lobe essential for memory (Squire and Zola-Morgan 1991) (Fig. 1). This system includes the hippocampal region (i.e., the hippocampus proper, the dentate gyrus, the subicular complex), the entor-

hinal cortex, and the adjacent perirhinal and parahippocampal cortices. A number of studies in monkeys have been carried out, comparing performance on memory tasks for monkeys with damage to different components of the medial temporal lobe. This work has shown that the severity of memory impairment depends on the locus and extent of damage within the medial temporal lobe memory system. Damage limited to the hippocampal region causes significant memory impairment, but damage to the adjacent cortex increases the severity of memory impairment. It is important to note that the discovery that larger medial temporal lobe lesions produce more severe amnesia than smaller lesions is compatible with the idea that structures within the medial temporal lobe might make qualitatively different contributions to memory function. This is because anatomical projections carrying information from different parts of the neocortex enter the medial temporal lobe memory system at different points (Suzuki and Amaral 1994).

Another important brain area for memory is the diencephalon, However, the critical regions in the diencephalon that when damaged produce amnesia have not at the time of writing been identified with certainty. The important structures appear to include the mediodorsal thalamic nucleus, the anterior nucleus, the internal medullary lamina, the mammillothalamic tract, and the mammillary nuclei. Because diencephalic amnesia resembles medial temporal lobe amnesia in many ways, these two regions together probably form an anatomically linked, functional system.

At about the same time that an animal model of amnesia was established, detailed case studies of human amnesia provided extensive neuropsychological information about memory impairment, as well as detailed neuropathological information about the locus and extent of damage in each patient's brain. New techniques for imaging brains have also made it possible to detect some kinds of neuropathology in living patients. The available neuroanatomical information from humans is consistent with the findings from monkeys. These lines of work have greatly clarified the description of memory impairment and its anatomical basis, and have made it possible to pursue parallel work in simpler animals such as rats and mice. As a result, one can now study memory in rodents and have some confidence that what one learns can be relevant to the human condition.

2. Declarative Memory

An important finding of neuropsychological studies of memory is that memory is not a single entity but is composed of several separate systems (Squire 1992, Schacter and Tulving 1994). Figure 2 illustrates how long-term memory can be divided into several parallel memory systems. The major distinction is between conscious knowledge of facts and events, referred to as declarative or explicit memory, and other nonconscious knowledge systems that provide for the capacity of skill learning, habit formation, the phenomenon of priming, and certain other ways of interacting with the world, collectively referred to as nondeclarative or implicit memory. Nondeclarative memory is not itself a brain-systems construct. Rather, this term encompasses several different kinds of memory. Nondeclarative forms of memory have in common the feature that memory is nonconscious. Memory is expressed through performance and does not require reflection on the past or even the knowledge that memory is being influenced by past events.

Declarative memory (explicit memory, relational memory) is a brain-systems construct, referring to the capacity for conscious recollections about facts and events that is dependent on the integrity of the hippocampus and anatomically related structures in the medial temporal lobe and diencephalon. It seems to be especially good at associating the various aspects of a context that are present at a particular time and place, thereby creating a memory of an episode. Declarative memory is also well suited for connecting the pieces of information needed to acquire a new fact (e.g., the capital of North Dakota is Bismarck). It is sometimes pointed out that declarative memory allows one to model the external world, and it is either true or false; nondeclarative memory is about performance, and it is neither true nor false.

Declarative memory is the kind of memory impaired in amnesia. However, the concept of declarative memory is not defined solely in terms of what amnesic patients can and cannot learn. Other characteristics of declarative memory have also been identified. For example, work with amnesic patients has emphasized the notion of conscious recollection, though this description is unhelpful when considering learning and memory in animals. Other characteristics have emerged from the studies of experimental animals. Declarative memory is fast, it is not always reliable (i.e., forgetting and retrieval failure can occur), and it is flexible in the sense that it is accessible to multiple response systems. It is especially suited for one-trial learning, and forming and maintaining an association between two arbitrarily different pieces of material.

2.1 Recognition Memory

A particularly good example of declarative memory involves the capacity to identify a recently encountered item as familiar, a capacity termed recognition memory. Recognition memory is impaired in amnesic patients, and it is impaired in monkeys and rats following damage to the hippocampal region. In the monkey, the assessment of recognition memory has

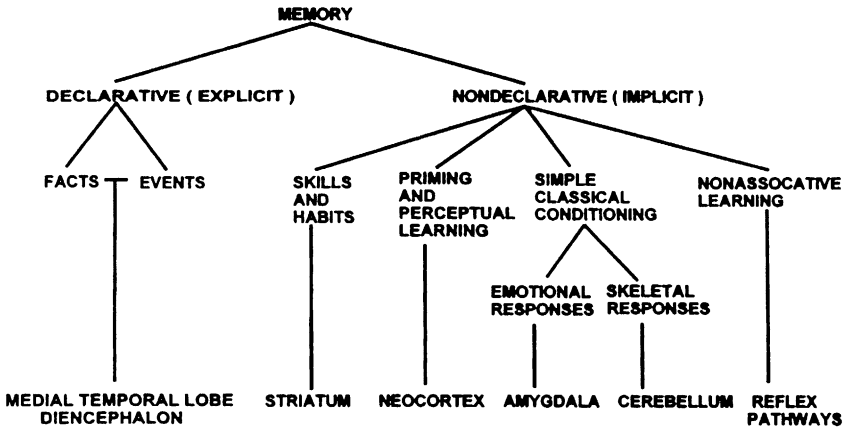

Figure 2
A taxonomy of long-term memory systems together with specific brain structures especially important for each system (reprinted, with permission, from Squire and Knowlton 1994)

depended mainly on two tasks. In delayed nonmatching to sample, two objects are presented, a new one and one that was presented earlier. The monkey must choose the new object. In the visual paired-comparison task, as adapted for the monkey, a new picture and a recently presented picture are presented side by side, and the monkey's spontaneous tendency to look at the new picture is measured. When the monkey shows a preference for viewing the new object, the monkey is demonstrating that it remembers the old object.

At the time of writing (2000), a total of six studies from different laboratories have been carried out on the effects of lesions of the hippocampal region on memory in monkeys. The lesions were made by an ischemic procedure, by radiofrequency, or ibotenic acid. Postmortem neurohistological analysis of the lesions is available in all six studies. Five of the studies have found impaired recognition memory. The sixth study, which differed from the others in potentially important ways, found no impairment (see Zola et al. 2000 for a review of these studies).

The evidence that the hippocampal region is essential for normal recognition memory is entirely consistent with current ideas about the role of the hippocampus in declarative memory and with the view that the hippocampus is essential for acquiring information about relationships, combinations, and conjunctions among and between stimuli (Eichenbaum 1997). The recognition test asks whether an item that had been presented recently now appears familiar. A recognition (or familiarity) decision requires that the stimulus presented in the retention test be identified as what was presented during learning. At the time of learning a link must therefore be made between the to-be-remembered stimulus and its context or between the stimulus and the animal's interaction with it. It is this associating and the ability to retain relational information across time that many have supposed is at the heart of declarative memory and in turn is the function of the hippocampal region in both humans and nonhuman animals.

2.2 *Episodic and Semantic Memory*

Episodic memory refers to the capacity for recollecting an autobiographical memory of events that occurred in a particular spatial and temporal context. Semantic memory refers to the capacity for recollecting general knowledge and facts about the world. Both types of memory are declarative in the sense that subjects are retrieving information explicitly, and they are aware that stored information is being retrieved. There is good agreement that episodic memory is impaired in amnesia following damage to the medial temporal lobe memory system. However, there is less agreement about semantic memory and its relationship to the memory system. By some accounts, amnesic patients are proportionately impaired in both episodic and semantic memory. By other accounts, the capacity for semantic memory is spared, or partially spared in amnesia relative to episodic memory ability, especially if the damage is limited to the hippocampal region.

Evidence for the second view comes especially from two kinds of data. First, there has been interest in cases where amnesia occurred in early childhood, before much of an individual's semantic knowledge has been acquired (Vargha-Khadem et al. 1997). Second, there have been studies of amnesic patients where the ability to accomplish fact learning and event learning has been directly compared. Taken together, the data tend to favor the view that episodic and semantic memory are similarly impaired in amnesia (Squire and Zola 1998). Other work suggests that episodic memory

additionally depends on the frontal lobes, which provide a source of autobiographical involvement with the recollected material. This orientation allows for a fundamental distinction between episodic and semantic memory, but without depending on differences in how the two kinds of memory depend on medial temporal lobe structures.

3. Overview

Work with amnesic patients and experimental animals has converged to characterize the organization of declarative memory. Declarative memory depends on the integrity of the medial temporal lobe, which works in concert with the neocortex. Declarative memory may develop gradually from early in life. Human infants six months of age successfully perform the visual paired-comparison task, and performance on this task is impaired by hippocampal lesions in humans and monkeys. The view that some declarative memory is available early in life is also consistent with demonstrations of long-term, recall-like memory abilities in human infants.

Other work with amnesic patients has suggested that declarative memory is equivalently involved in the two fundamental processes of recall and recognition. Work in humans and animals has shown that the medial temporal lobe memory system is required at the time of learning if an enduring and retrievable long-term declarative memory is to be established. Studies of retrograde amnesia in amnesic patients and in experimental animals suggest that the brain system supporting declarative memory has a time-limited function. Medial temporal lobe lesions spare very remote memories. Thus, as time passes after learning, a process of consolidation and reorganization occurs such that permanent storage does not require the medial temporal lobe. Permanent memory is thought to be stored in the neocortex where information is received and processed. Information is still accumulating about how memory is organized, what structures and connections are involved, and what functions they support. The disciplines of both psychology and neuroscience continue to contribute to this enterprise.

See also: Amnesia; Episodic and Autobiographical Memory: Psychological and Neural Aspects; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Long-term Depression (Hippocampus); Long-term Potentiation (Hippocampus); Memory, Consolidation of; Memory: Synaptic Mechanisms; Recognition Memory (in Primates), Neural Basis of; Short-term Memory: Psychological and Neural Aspects; Working Memory, Neural Basis of

Bibliography

Eichenbaum H 1997 Declarative memory: insights from cognitive neurobiology. *Annual Review of Psychology* **48**: 547–72

Mishkin M 1982 A memory system in the monkey. *Philosophical Transactions of the Royal Society London—Biological Sciences* **298**: 85–92

Schacter D 1996 *Searching for Memory*. Basic Books, New York

Schacter D L, Tulving E (eds.) 1994 *Memory Systems*. MIT Press, Cambridge, MA

Scoville W B, Milner B 1957 Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, Psychology* **20**: 11–21

Squire L R 1992 Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological Review* **99**: 195–231

Squire L R, Kandel E R 1999 *Memory: From Mind to Molecules*. Freeman, New York

Squire L R, Zola S M 1998 Episodic memory, semantic memory, and the medial temporal lobe memory system. *Hippocampus* **8**: 205–11

Squire L R, Zola-Morgan S 1983 The neurology of memory: The case for correspondence between the findings for human and nonhuman primate. In: Deutsch J A (ed.) *The Physiological Basis of Memory*. Academic Press, New York

Squire L R, Zola-Morgan S 1991 The medial temporal lobe memory system. *Science* **253**: 1380–6

Suzuki W A, Amaral D G 1994 Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. *Journal of Computational Neurology* **350**: 497–533

Tulving E, Craik F I M (eds.) 2000 *The Oxford Handbook of Memory*. Oxford Press, New York

Vargha-Khadem F, Gadian D G, Watkins K E, Connely A, Van Paesschen W, Mishkin M 1997 Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**: 376–80

Zola S M, Squire L R, Teng E, Stefanacci L, Clark R E 2000 Impaired recognition memory in monkeys after damage limited to the hippocampal region. *Journal of Neuroscience* **20**: 451–63

S. M. Zola and L. R. Squire

Deconstruction: Cultural Concerns

Deconstruction is a mode of philosophical and literary analysis, derived from the work of the French philosopher Jacques Derrida, which questions the most basic philosophical categories or concepts. In the 1970s the term was applied to analyses, by Derrida and others, of philosophical and literary writings. In the 1980s it came to designate more loosely a range of radical theoretical enterprises in diverse areas of the humanities and social sciences, including law, psychoanalysis, architecture, theology, feminism, gay and lesbian studies, political theory, historiography, and film theory. What these enterprises shared was a critical dismantling of the conceptual oppositions that had previously been regarded as fundamental to the

disciplines in question. In popular usage (e.g., Woody Allen's film *Deconstructing Harry*), the term has come to designate a critical dismantling or undoing; and in polemics about late twentieth-century thought, deconstruction is sometimes used as a pejorative synonym for 'poststructuralism' or 'postmodernism'—one which suggests nihilism or frivolous skepticism.

1. Deconstruction and Philosophy

Deconstruction undertakes a radical critique of fundamental categories of Western thought by exploring how these categories have been constructed in and by the discourses that rely on them. It is a mode of analysis that examines not only the arguments of philosophical texts but also the rhetorical procedures and devices they employ and the tensions or contradictions between what is claimed or assumed and what the texts themselves must do in order to support such claims. A major focus of deconstruction has been the traditional binary oppositions that have structured Western thought since the time of the Greeks: for example, the oppositions between inside and outside, mind and body, literal and metaphorical, speech and writing, presence and absence, nature and culture, intelligible and sensible, form and meaning. Each of these oppositions is hierarchical, in the sense that one of its terms has been taken to be primary and fundamental, and the other secondary and derivative (Derrida 1981). Thus, nature is logically prior to culture; writing is seen as merely a way of representing speech, which is taken to be the basic form of language; meaning is what comes first and is then given expression by form.

A deconstruction of these oppositions asks how the philosophical enterprise and thought in general have relied upon the privileging of one term and invites us to consider whether in fact such hierarchies should not be revised. In particular, a deconstruction questions these hierarchical oppositions by showing how they have been constructed and, in the process, undoing that construction and establishing a new relation between the terms. It can be shown, for example, that the idea of 'nature' is a product of culture: conceiving something as prior to culture is a specific cultural operation whose import and function needs to be assessed. In the opposition between nature and culture, the primacy of nature cannot, therefore, be taken for granted. What we have, rather, is an opposition elaborated within culture; what counts as nature in any historical moment will be a fact about that culture. To argue in this way is to invert and restructure the opposition (to 'displace' it) in a way characteristic of deconstruction. Instead of a primary nature and a secondary culture, we discover a variable distinction between nature and culture within culture.

For Derrida, the most telling and pervasive opposition has been the one that treats writing as secondary or derivative with respect to speech. According to this opposition, in speech the ideas and intentions of the speaker are immediately present; it is a direct and authentic form of language, whereas writing is merely a graphic representation of the spoken word (a sign of a sign) and hence marked by absence and possibilities of discrepancies between form and meaning. (Derrida 1976). By setting aside writing as a secondary and derivative, a mere representation open to misunderstanding, accounts of language have taken as their object an idealized form of speech, where the linguistic form is a direct expression of what the speaker, as we say, 'has in mind.' But Derrida argues that linguistic forms can function as signs only to the extent that they can be repeated in different contexts, in the absence of any particular speaker's presence or intention (Derrida 1988). Speech is only possible, in other words, to the extent that it has the qualities assigned to writing, such as absence, difference, and the possibility of misunderstanding. One mark of this, Derrida has shown, is the frequent recourse, in attempts to describe speech, of examples and metaphors drawn from writing. In effect, speech has been described as a form of writing, even when the claim has been that writing is derivative from speech (Derrida 1976). This deconstruction of the traditional hierarchical opposition between speech and writing argues not that there are no differences between speech and writing but that the traditional opposition is untenable and that both speech and writing are forms of a general writing (*archi-écriture*), which is the condition of possibility for any system of representation whatsoever.

Derrida argues that treating writing as secondary to speech is part of what he calls the 'logocentrism' of Western culture: the presumption of an order of truth or thought prior to its representation by signs (Derrida 1976). Logocentrism tries to treat representation as inessential rather inextricably involved in the structure of phenomena, but an attentive reading of the texts of the philosophical tradition shows that they tell a different story. Derrida's work attempts to demonstrate that discourses that treat writing as secondary in fact must make use of notions linked to writing when they characterize speech and that thus speech can be seen as a version of a generalized writing, which is the condition of language and thought in general. This is an instance of the deconstruction of concepts seen as fundamental, such as 'presence,' 'truth,' 'origin,' and 'identity.'

Deconstruction may be conceived in the first instance as a critique of Western philosophy. Philosophy, Derrida argues, has been founded on a theory of 'presence,' in which such notions as truth, being, and reality are determined in relation to an ontological center, essence, or origin, based on the repression of absence and difference. But since deconstruction works not by erecting an alternative theory but by exploring how philosophical discourses have produced their concepts through rhetorical stratagems, decon-

struction was welcomed by some students and teachers of literature as a powerful practice of reading which would illuminate how texts—whether philosophical or literary—implicitly put in question what they explicitly maintain or what they appear to assume. Barbara Johnson, a leading US practitioner, calls deconstruction 'a careful teasing out of warring forces of signification within a text' (Johnson 1980, p.5).

2. *Deconstruction and Literary Studies*

In the 1970s and 1980s deconstruction played a major role in the animation and transformation of literary studies by what was called simply 'theory,' for short. Briefly stated, 'theory' is concerned with questions about the nature of language, the production of meaning, or the relations between literature and the numerous discourses that structure human experience and its histories. In the USA, deconstruction was associated in particular with a group of scholars at Yale University, where Derrida taught regularly as a visiting professor. In the hands of Paul de Man, J. Hillis Miller, and Barbara Johnson, for instance, deconstruction became a powerful practice of rhetorical reading, attentive to what is at stake in texts of all sorts and to how texts engage those stakes.

Deconstructive literary criticism was often compared, especially by its critics, to an earlier Anglo-American New Criticism, which sought to explicate and appreciate verbal works of art as complex constructions and emphasized the role of paradox and irony in these complex artifacts. However, deconstructive readings, by critics such as those mentioned above, differ in several ways (Culler 1982). They treat works of art not as the harmonious fusion of form and content but as instances of the intractable conflicts between modes of meaning. They generally examine the individual work not as a self-contained artifact but a product of relations with other texts or discourses, literary and nonliterary. Finally, they are especially interested in the ways in which the works may offer implicit critiques of the categories critics use to analyze them.

3. *Deconstruction Beyond Philosophy and Literature*

Critics of deconstruction have tended to attack it either as obscurantist wordplay, excessively invested in supposedly minor or marginal rhetorical features of texts and cast in language that obfuscates rather than illuminates, or as a nihilistic denial of authority, certainty, and foundations, for which 'anything goes' (Ellis 1989). But despite attacks, the impact of deconstruction on modern thought, not just literary studies, has been extensive. Most generally, it has encouraged the questioning of the hierarchical oppositions that structure any field of inquiry and an interest in whether and how these fundamental oppositions are subverted by the phenomena that they are used to describe. Deconstruction is thus a powerful version of the general critique of supposedly scientific metalanguages—sets of terms or concepts used to analyze a domain and which are regarded as external to the phenomena they purport to analyze. Deconstructive analyses explore the ways in which the supposedly neutral or purportedly scientific metalanguage is deeply affected by the phenomena it is supposed to analyze. Thus, it might show, for instance, that the terms deployed by theory of metaphor or figurative language generally are themselves scarcely free of metaphor or explore how a psychoanalytic theory is itself affected by the mechanisms of repression and wish-fulfillment it purports to analyze. Since within literary and cultural studies it was above all structuralism (see *Structuralism*) which sought to elaborate metalanguages for analyzing cultural phenomena as a series of languages or signifying systems, the deconstructive critique of the possibility of a neutral metalanguage was seen as a step beyond structuralism and hence as a case of 'poststructuralism.'

There are two main aspects to the influence of deconstruction outside of literary and philosophical studies. On the one hand, granting primacy not to what a text 'says' but, rather, to how what it 'does' relates to what it says, deconstruction emphasizes that rhetorical structures are centrally at work in language of all sorts and that discourse has an important 'performative' dimension, structuring experience in particular ways. It thus seconds and invigorates the constructivist tendencies of numerous disciplines: the attempt to show how that the phenomena studied by a discipline are not simply given to experience but are produced by conceptual frameworks and discursive practices. On the other hand, as a critical exploration of fundamental oppositions and an attempt to intervene and alter the value attributed to particular terms, deconstructive thinking has affected not just how texts are read but the sense of what a discipline's goal might be. In psychoanalysis, for instance, deconstructive readings of texts by Sigmund Freud and such successors as Jacques Lacan, have done several things:

(a) focused on the structuring role of language in the formation of the psyche;

(b) shown how psychoanalytic case studies are themselves structured by the kinds of psychic mechanisms that they purport to analyze (so Freud's or Lacan's writings are themselves organized by the processes of repression, condensation, and displacement that they theorize); and

(c) questioned the complicity of psychoanalysis with a logocentrism which believes in originary principles or an ultimate reality (Derrida 1987).

Deconstructive thinking in psychoanalysis may seek not the origin of a trauma but the stakes of a belief in origin.

Some strands of feminist thinking espouse a deconstruction of the opposition between men and women and a critique of essentialist notions of identity. Judith Butler's (1990) *Gender Trouble: Feminism and the Subversion of Identity*, for instance, not only challenges the claim that a feminist politics requires an identity for woman (for Butler, identity is the product or result rather than the source of action); it develops a 'performative' concept of gender and sexual identity, modeled on the way in which linguistic acts (such as promising) work to bring into being the entities (the promise) to which they refer. This perspective has been influential in the realm of gay and lesbian studies or 'queer theory,' as the academic avant-garde linked to movements of gay liberation has styled itself.

In the field of law there has been considerable deconstructive analysis within the movement known as 'critical legal studies,' which has focused on conflicts between principles and counterprinciples within the system of legal doctrine. There are critiques of the oppositions such as public vs. private, essence vs. accident, substance vs. form, which are fundamental to the realm of law, and demonstrations that legal doctrine and argument are attempts to paper over contradictions which nonetheless reassert themselves (Kelman 1987). Attacks on critical legal studies, like attacks on deconstruction in philosophy, often accuse it of undermining confidence in a system without offering an alternative model.

Finally, the impact of deconstruction has spread beyond the humanities and social sciences to the field of architecture, where deconstructive architecture has come to mean design that flaunts, often in a parodic way, the elements from which it is constructed.

In its diffusion in the humanities and social sciences, following its greatest moment of fame in the 1970s and 1980s, deconstruction in the broadest sense has become a critique of categories taken as natural and a drive to pursue the analysis of the logic of signification in a given area as far as one can, even if the result is disquieting—an exacerbation of questions or problems rather than their resolution. It has thus joined with other strands of postmodern and post-structuralist thinking to inspire suspicion of established categories and canons and skepticism about objectivity. Although opponents of deconstruction have often sought to link its critique of established values to an amoral nihilism, as a critique scientistic pretensions in the humanities and social sciences (among other things) deconstruction has included a strong element of ethical reflection, particularly about the conditions of possibility of ethics or justice (Critchley 1992, Cornell et al. 1992).

Bibliography

Brunette P, Wills D (eds.) 1994 *Deconstruction and the Visual Arts: Art, Media, Architecture*. Cambridge University Press, Cambridge, UK

Butler J 1990 *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York, NY

Cornell D, Rosenfeld M, Carlson D G, (eds.) 1992 *Deconstruction and the Possibility of Justice*. Routledge, New York, NY

Critchley S 1992 *The Ethics of Deconstruction: Derrida and Levinas*. Basil Blackwell, Oxford

Culler J 1982 *On Deconstruction: Theory and Criticism. After Structuralism*. Cornell University Press, Ithaca, NY

Derrida J 1976 *Of Grammatology*. Johns Hopkins University Press, Baltimore, MD

Derrida J 1981 *Positions*. University of Chicago Press, Chicago, IL

Derrida J 1987 *The Post Card: From Socrates to Freud and Beyond*. University of Chicago Press, Chicago, IL

Derrida J 1988 *Limited Inc*. Northwestern University Press, Evanston, IL

Derrida J 1994 *Spectres of Marx: The State of the Debt, the Work of Mourning, and the New International*. Routledge, New York

Elam D 1994 *Feminism and Deconstruction: Ms en Abyme*. Routledge, New York, NY

Ellis J M 1989 *Against Deconstruction*. Princeton University Press, Princeton, NJ

Game A 1991 *Undoing the Social: Towards a Deconstructive Sociology*. Open University Press, Milton Keynes

Johnson B 1980 *The Critical Difference*. Johns Hopkins University Press, Baltimore, MD

Kelman M 1987 *A Guide to Critical Legal Studies*. Harvard University Press, Cambridge, MA

Taylor M C 1982 *Deconstructing Theology*. Crossroad Pub. Co., New York

J. Culler

Deductive Reasoning Systems

Reasoning is one of the fundamental capabilities associated with intelligence as manifested in humans. It is a mental activity by which we generate (propositional) knowledge not available prior to the act of generation. Experience shows that knowledge generated by reasoning on the basis of already available knowledge enables humans

(a) to predict states of the world not experienced before and to act accordingly; and

(b) to explain phenomena experienced in the world so that we can understand the reasons why they happened, and make them happen again if they are desirable, or otherwise avoid them in the future.

The combination of prediction and explanation is a fundamental and characteristic ingredient of the human way of life, which demonstrates the importance of reasoning.

Experience also shows that humans have a sense for the correctness of reasoning chains. In other words, we are not predicting arbitrary knowledge nor giving arbitrary explanations. Rather, there is an experienced correlation between the correctness of reasoning and its usefulness for predictions and explanations in the

real world. On the other hand, humans are prone to mistakes in their reasoning. In coping with this and other weaknesses lies the motivation for the mechanization of reasoning in the form of *reasoning systems*.

This article describes the technical basis for reasoning systems, points out a number of issues of particular importance for the design of such systems, and briefly surveys their numerous applications.

1. Abstractions

In order to model the complex phenomena involved in reasoning various abstractions are needed, which are now discussed.

1.1 Formal Languages

Humans express their knowledge in natural language which, however, is too complex to be useful as a starting point for the mechanization of reasoning. A first important abstraction is therefore the representation of human knowledge in some formal language. It should be rich enough to be able to express all of our (propositional) knowledge, but simple enough to be useful for the manipulations needed in modeling reasoning, and concise enough to avoid ambiguities and redundancies. There are a number of formal languages meeting these criteria. They differ in the level of abstraction taken for the representation of knowledge, and in the comparative emphasis given to the various concepts involved.

The language of *propositional logic* takes the highest level of abstraction by formalizing knowledge at the level of entire propositions. The language of *first-order logic* is less abstract and much more expressive. It is probably the most popular formal language for knowledge representation, especially in the context of reasoning systems. Many knowledge representation formalisms (like the one used in Aristotelean syllogisms, semantic nets, description languages, etc.) are first-order formalisms in disguise, but lay the emphasis, for instance, on the concept of sets (as in the Aristotelean language) rather than on the functional (or Fregean) relationships among objects emphasized in first-order logic. Modal, transitional, and higher-order languages including the lambda calculus extend the expressiveness of first-order languages even further. For the purpose of this article, we restrict our discussion to the case of first-order logic. Most of our considerations generalize in some way to the other, more general, languages.

1.2 The Inferential Relationship

Reasoning takes place with a given body of knowledge, say K, already available. In the predictive mode of reasoning we want to determine what facts, say P, could be true in the future. Correct reasoning distinguishes between facts possible on the basis of K, and those which are impossible. In other words, there is a relation which associates K and P in the positive case. Following Frege, this relation is formally denoted by $\models$. That is, $K \models P$ is short for expressing that P is a correct reasoning prediction on the basis of K. In the *predictive mode* of reasoning we thus want to determine P such that $K \models P$ holds, or test for a given P whether it holds.

Human reasoning is mostly done in time, and to a significant extent also concerns knowledge about actions and changes taking place in time. On the other hand, a large part of the inferential relation $\models$ is, so to speak, 'timeless.' To simplify further, most logical studies have abstracted from the issues concerning time and change. Only in the last few decades of the twentieth century did those issues become of great interest within Intellectics, the field covering both Artificial Intelligence and Cognitive Science. The result of those recent studies is that the timeless part of $\models$ is already expressive enough to model also the part involving time, and thus the entire relation $\models$. Insofar this abstraction from time is without restriction of generality. Under this abstraction, the predictive mode of reasoning is called the *deductive mode* of reasoning. Because of what was stated above, deductive reasoning at the surface is timeless reasoning, but may as well be regarded as predictive reasoning in general.

In the explanatory mode of reasoning based on K, we start with some observation O and basically look for some explanation E so that $K, E \models O$. This mode of reasoning is called abductive or inductive depending on the circumstances concerning O and E. Since $\models$ is a relation, the difference with deductive reasoning is minor; instead of the second argument here just (part of) the first one is to be determined. The relationship is even closer. In inductive reasoning we employ additional meta-knowledge, say M, which guides the process of generalizing from given observations. Taking this into account, inductive reasoning actually boils down exactly to deductive reasoning, namely in the form $K, M, O \models E$.

In summary, all prevailing modes of human reasoning may actually be regarded as deductive reasoning. This explains why deductive reasoning systems have such a great importance for modeling inference in all its generality.

1.3 Syntactic Inference

The fundamental observation already made more than 2000 years ago by Aristotle and his predecessors is that reasoning (for the most part) is a matter of form rather than of content. This became apparent once knowledge was represented in a *form*alized way. The knowledge 'apes are mammals'—in the sense of 'whenever the property of being an ape holds then also the mammal property holds,' or $A \rightarrow M$, and 'mammals

have hair,' $M \rightarrow H$, allows us to conclude that 'apes have hair,' $A \rightarrow H$. The inference is valid independently of the meaning of the words apes, mammals, and hair, or of the symbols A, M, H, for that matter. Rather, it is exclusively the form of the knowledge which sanctions the validity of such an inference. This fundamental property of reasoning enables machines, which are ignorant of the meaning of words or symbols, to perform correct inferences.

The example demonstrates a fundamental rule of inference: $A \rightarrow M$ and $M \rightarrow H$ implies $A \rightarrow H$. It is called (a restricted form of) resolution or (a slightly generalized form of) *modus ponens*. As we see, it produces new knowledge by eliminating the intermediate atom M; the rule may also be seen as implementing a form of transitive closure.

In order to program computers to search for chains of inferences we first need rules of inference such as the one just discussed, in addition to the formal languages mentioned earlier. The combination of a formal language, along with a set of rules is called a (formal or logical) calculus. If F is some formula i.e., a propositional expression in the formal language under consideration, viz. first-order logic according to our convention, and G is a formula obtained from F by application of the rules of the calculus, then we write $F \vdash G$. Since $\vdash$ is dependent on the calculus c—which, as in the case of formal languages, can be chosen from a great variety of alternatives—we should rather write $\vdash_c$, but omit the index whenever the calculus is clear from the context. We want to design calculi in such a way that $F \vdash G$ holds if, and only if, (iff) $F \models G$, in which case we speak of *correct* (if) and *complete* (only if) calculi. From now on we assume that the rules are chosen such that the resulting calculi are always correct and complete.

We go one step further in the abstraction process described in this entire section. According to the well-known deduction theorem, it holds that $F \vdash G$ iff $\vdash F \rightarrow G$. By applying this theorem iteratively we can always achieve that there are no formulas to the left of $\vdash$. Most deductive systems in use are therefore systems which test whether, for some given calculus and some given formula F, it holds that $\vdash F$ (affirmative proof), or dually, that $\vdash \neg F \equiv \mathit{false}$ (proof by contradiction).

The field in which such systems are developed, studied and applied is called automated deduction (AD). It has applications in any field where reasoning of any kind is relevant, which actually is most of science. AD's first application was to Mathematics, for which historical reason the field is also known as automated theorem proving (ATP).

2. *Issues in AD*

Deductive systems have been developed for numerous different calculi with different formal languages and different rule sets. A very popular language is a version of first-order logic, called *clausal logic*. It uses sets of clauses like the three clauses in our apes example (which actually form four clauses, since presenting the set in the form $\neg F$ logically splits the last clause into the assumption $\rightarrow A$ with an empty premise and the denial of the conclusion, i.e., $\neg H$ or, equivalently, $H \rightarrow$ with an empty conclusion). In general, a clause may contain a conjunction of atoms in the premise and a disjunction of atoms in the conclusion. A correct and complete calculus based on clause sets is given by (essentially) the single rule of resolution in its general form. Resolution allows the elimination of intermediate literals even if they are not exactly identical but can be made identical by replacing variables in the atoms by functional terms, a process known as unification.

A deductive system based on resolution works as a saturation process. It takes the given clause set, and iteratively applies resolution until the so-called empty clause $\rightarrow$ (no premise, no conclusion, representing *false*) is obtained which forms the criterion for the validity of the inference represented by the input. The advantage of this type of system is the simplicity of the data-structures involved (clause sets) and the operations performed at each iteration (resolution). The problem is that, without additional constraints, the system produces relevant and irrelevant intermediate results alike, and for hard problems is often drowned in up to billions of generated clauses. An irrelevant intermediate result in our earlier example might be that 'apes have eyes,' on the basis of the additional knowledge that 'mammals have eyes,' when our only interest is in the question about hair. Strategies (like the so-called set-of-support strategy) to some extent help to reduce the amount of irrelevant clauses generated.

Alternatives to this saturation-based approach are the goal-oriented tableaux-based deductive systems. Human proofs, i.e., chains of inferences, may be regarded as tree-like structures, or (closed) tableaux, which in each branch start out from an obviously valid fact. The basic data-structure of tableaux-based systems is a (not necessarily closed) tableau. Starting with the tableau consisting of the (denial of the) given formula, the tableau is extended at each iteration with the goal of closing it. The data-structure (tableau) and the operations (extension of tableau) are more complicated than those in the saturation-based systems, which complicates the system development. But the approach allows a more goal-oriented fine-tuning of the search for the final proof and therefore is able to avoid some of the redundancy involved in resolution systems. At the time of writing, both approaches have reached about the same performance levels.

Systems based on the connection method may be regarded as a compactly coded form of tableaux systems. In this method, all information coded by a tableau is attached to the parts of the given formula in an extremely condensed form as structural information. Starting with an initial structure of this kind, only

the structure is modified at each iteration until a certain final structure is reached signaling validity of the formula; the latter itself is not changed at all in the process.

Research in AD pursues all three kinds of approach, with the focus in each case being on eliminating redundancy. This can be achieved in various ways. One way consists in refined calculi in terms of the rule sets used. For instance, the application of resolution may be restricted in many ways without sacrificing completeness (e.g., linear resolution). Another way to reduce redundancy consists in applying various strategies which guide the rule application. A third consists in performing several small steps as a single global step (e.g., hyperresolution) which again narrows the search space. A fourth involves global steps sanctioned by theories involved (such as equality theory). Finally, ever-improved software engineering techniques and faster machines contribute to a continuous enhancement of system performance.

3. *Applications and References*

Mathematics is a primary field of application for AD. Deductive systems have been applied to virtually all subdisciplines in Mathematics, and tens of thousands of theorems have been proved automatically. New proofs have been found automatically for known theorems as well as for conjectures never proved before by any human. For instance, a system called Otter/EQP in 1996 settled an important mathematical problem which had been open for sixty years.

AD techniques are further used in programming. For instance, the programming language PROLOG basically works as a deductive system. Systems which synthesize programs from descriptive specifications, like the system KIDS, involve strong deductive components. Another important application of AD techniques to programming consists in the verification of software or hardware. Entire chips as well as compilers have been proved correct this way.

AD techniques have spread out into all fields of science by way of so-called knowledge-based systems which encode knowledge of specific disciplines and allow the manipulation of such knowledge in a deductive way. In the CYC system, an entire encyclopedia of knowledge is encoded in this way.

Bibel (1993) is a special textbook on all aspects of deductive methods and systems. Bibel and Schmitt (1998) is a reader covering all modern deductive techniques, a number of very successful systems, and several important applications. Gabbay et al. (1991) is a handbook which includes deductive reasoning systems and other topics. Robinson and Voronkov (2000) is a handbook covering only deductive methods, techniques and systems. Ongoing research is reported in the *Journal of Automated Reasoning* (Kluwer Academic Publishers) and in the proceedings volumes (published by Springer Verlag) of the International Conferences on Automated Deduction, as well as in other journals and conferences.

See also: Artificial Intelligence in Cognitive Science; Artificial Intelligence: Uncertainty; Informal Reasoning, Psychology of; Intelligence: History of the Concept; Knowledge Representation; Logic and Cognition, Psychology of; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning; Scientific Discovery, Computational Models of

Bibliography

Bibel W 1993 *Deduction: Automated Logic*. Academic Press, London

Bibel W, Schmitt P H 1998 *Automated Deduction—A Basis for Applications*. Applied Logic Series. Kluwer, Dordrecht, Vol. I–III, The Netherlands

Gabbay D, Hogger C, Robinson J A (eds.) 1991 *Handbook of Logic in Artificial Intelligence and Logic Programming*. Oxford University Press, Oxford, UK

Voronkov A, Robinson J A (eds.) 2001 *Handbook of Automated Reasoning*. North Holland, Amsterdam

W. Bibel

Defense, Economics of

This entry treats the nature of defense economics in applying economics to warfare. It first discusses the role of the state in providing the resources to wage war. Second, it presents the Richardson model of arms races. Third, it considers how nuclear proliferation affects the probability of nuclear war. Fourth, it treats the economics of guerrilla warfare. Finally, it reflects on the changing nature of warfare in the twenty-first century.

1. *The Political Economy of War*

Logistics has long been important to warfare. As railroads became more important in the latter part of the nineteenth century, moving men and materials became more complicated and it became necessary to create schools to train officers to formulate and to implement elaborate plans that depended on railroads. These plans may have reached their most refined level in the mobilization plans of the French, German, and Russian armies prior to World War I. These plans were so delicate that any delay in implementation would lead to defeat since the first army to complete its

mobilization was expected to have a decisive advantage. These plans and the strategic environment they created are an example of what later became known as 'crisis instability' during the Cold War. Many historians believe that this requirement to mobilize without delay led to a war that could have been avoided. The plans and the strategic environment they created exemplify a case where maximizing behavior on the part of agents and technology can lead to unexpected or undesirable consequences. In the 1970s and 1980s such cases became a matter of concern among economists working on defense issues.

It can be argued that waging war is one of the more important reasons why states exist. The preamble of the United States Constitution states this explicitly: 'We the people of the United States in order to form a more perfect Union, establish justice, insure domestic tranquility, *provide for the common defense* … do ordain and establish this Constitution for the United States.' (Emphasis added)

Historically, an important function of the state has been to provide the resources to wage war and this became an important issue in political economy. The organization of the state reflects, in part, its need to raise resources for military purposes. A very good example of such a phenomenon and its evolution is the organization of the military in England. William the Conqueror, Duke of Normandy, defeated King Harold at the Battle of Hastings in 1066 and became King of England. He did not invade England in his capacity as Duke of Normandy, however, but rather as the head of a mercenary army motivated primarily by the possibility of enrichment. The barons of Normandy had refused to follow William to England in their corporate capacity and, as a result, the army of William was composed of individuals who were following him for 'land or for pence.' William was the warlord of an army that is estimated to have been about 10 to 12,000 mounted men and about 12 to 20,000 foot soldiers.

After William conquered England, he divided the land among his followers. These followers were expected to provide military service in return for their land. This manner of landholding was known as 'military tenure' and was based on a personal relationship between a man and his lord. Approximately 4500 knights were due this distribution of land. What William and his followers developed created an instrument by which a few thousand men could control a population of approximately one and half million people. The granting of land to tenants-in-chief and subinfeudation by military tenure was an effective mechanism for both supporting an army of occupation and providing a local political presence to control the country. A knight with a smallholding could support himself while not in service and also provide stability to the countryside.

By the early reign of Henry II (1154–89) England's military needs had changed and were no longer primarily internal. Henry's interests were primarily in France, and they required the services of mercenaries rather than the feudal levy of knights. The service of the English feudal levy was not very useful for these purposes. The term of service was such that it would have ended before the force could be employed. Further, feudal hosts were untrained for siege warfare. The mercenaries, on the other hand, were trained professional soldiers who would serve as long as they were regularly paid.

The other historical force that helped to end military tenure was the period that has come to be known as the Great Anarchy. At the death Henry I, in 1135, William the Conqueror's grandson, Stephen of Blois was crowned king of England rather than Henry's daughter, the Empress Matilda, who had been recognized as his successor by his barons. Stephen's coronation was followed by a long period of warfare between his supporters and those of the Empress Matilda. This long war ended with the treaty of Winchester in 1154. A key element of the treaty was that tenure in land became a right guaranteed by the king, based on legitimate possession in 1135. The holding of land was no longer a relationship between a man and his lord. Money replaced military service as a means of payment. Holding of land in what is now known as 'fee simple' replaced military tenure, and land could be bought and sold.

One of the consequences of land being held in fee simple is that it created a class of small landholders who were free men, yeoman, who were able to acquire skill with a longbow. Since this class did not exist in France, the English armies were able to deploy a weapon that was not available to French armies. The value of these weapons was dramatically demonstrated at Agincourt and Crecy when the ratio of casualties between the French and English were on the order of 1000 to 1.

England became primarily a seapower after the sixteenth century. The middle class was the source of officers for the Navy. English naval officers went to sea as young boys, acquired technical skills and were promoted mostly on merit. This approach contrasted with those of the French and Spanish navies where there was a sharp distinction between the aristocrats who commanded the ships and the sailing masters who had the technical skills. As a result, 'Britannia ruled the waves', and seapower became the keystone of the British Empire.

The lesson that can be inferred from this example of England is that institutions can be created and can evolve to meet military needs, but these institutions themselves create both opportunities and constraints for military technology and tactics.

The importance of the relationship between technology, military needs, and political institutions must be kept in mind as we face three military problems of the twenty-first century. First is the question of weapons of mass destruction. Second is the need to use

force to maintain civil order under circumstances where there is weak external political support. Third is the support of guerrilla warfare by nonstate actors. These are all problems that will require creative solutions with respect to military technology and institutions.

2. *Richardson's Equations and the Arms Race*

In the period from the end of World War II in 1945 to the breakup of the Soviet Union in 1991, the most important issue that demanded the attention of economists and political scientists working on defense issues was the arms race between the USA and the Soviet Union. The main goal was to develop models that could be used to address two very important questions: Were arms races stable? Did they lead to war?

The Richardson arms race model was the starting point for most of this research. It was the first formal model of an arms race and one of the most influential formal models in all of the international relations literature. (Richardson (1960), Isard (1988), Brito and Intriligator (1995).) The Richardson model is summarized by two differential equations that describe the rate of change over time of weapon stocks in each of two countries. According to the Richardson model, arms increases can be described as the sum of three separate influences. First is the 'defense term', where the accumulation of weapons is influenced positively by the stock of weapons of the opponent, representing the need to defend oneself against the opponent. Second is the 'fatigue term', where the accumulation of weapons is influenced negatively by one's own stock of weapons, representing the economic and administrative burden of the arms race. Third is the 'grievance term' representing all other factors influencing the arms race, whether historical, institutional, cultural, or derived from some other source. In the Richardson model, these terms are independent, additive, and linear, resulting in two coupled linear differential equations and linear reaction functions that determine the equilibrium levels of weapons.

In the early 1960s stock adjustment models, such as those of Boulding (1962) and Intriligator (1964) were developed. Each country determines a desired stock of weapons, and the rate of change of weapon stocks is assumed to be proportional to the discrepancy between desired and actual weapon stocks. Thus, each side acquires additional stockpiles in order to offset and overcome a perceived deficiency in its stockpiles of weapons, the deficiency given by the gap between the desired and actual levels of weapons.

In general, the desired stocks of weapons depend on the levels of weapons in both countries. If, as one example, each country desires a certain base level of weapons and, beyond that, to match increments of the other side's levels according to a fixed ratio, then the desired stocks are linear functions of the levels of weapons held by the other side. This results in two coupled differential equations as in the Richardson model. Conversely, the Richardson model can be interpreted as a stock adjustment model, with desired stocks a linear function of the weapons held by the opponent.

The problem with these models is that they are mechanical and there is no maximizing behavior. Brito (1972) formulated a model in which the arms race was a result of optimizing behavior. Countries were assumed to choose between investment in weapons or consumption (guns or butter). Weapons were assumed to be inputs into a public good, 'defense', that produces a positive externality for the country that produces them and a negative externality for the other country. The result of this process was a differential game and the equations that described the investment in weapons resulting from this differential game formed a nonlinear version of the Richardson model.

Brito was able to show that there exists an equilibrium level of armaments and to derive sufficient conditions for the equilibrium to be stable. Finally, he showed that in a dynamic economic model total disarmament is unstable except under very special circumstances and that behavior very similar to 'Prisoner's Dilemma' would be observed.

A criticism that can be lodged against the Brito model is that it ignores strategic considerations, in particular, the roles of weapons in both deterring and conducting a war. Intriligator (1975) developed a model that took such factors into account, connecting the acquisition of weapons to their use in both peace (deterrence) and war (war fighting). The political authorities will generally require the military authorities to justify their proposed budgets, in particular, any request for increases in the level of weapons. They may also question the existing levels of weapons in term of their danger and/or expense. The military authorities would typically seek to justify both their budgetary request for weapon acquisitions and their current inventories of weapons in terms of national security considerations by showing their deterrence or war fighting. Intriligator showed that there were combinations of weapons stocks that would make it advantageous for one or both parties to launch a first strike and result in crisis instability. At the same time, there were combinations of weapons stocks that eliminated the advantage of a first strike for both parties and thus were stable in a crisis. Low levels of weapons stocks on both sides were associated with crisis instability.

3. *Nuclear Proliferation and the Probability of War*

An important question that has been the subject of controversy is the influence of additional nuclear weapons states on the probability of nuclear war.

While the usual argument is that proliferation will increase the probability of nuclear war, some have argued that proliferation may reduce rather than increase this probability, as in Gallois (1961), Wentz (1968), and Waltz (1981). Gallois argued that additional nuclear weapon states, such as France at the time he was writing, would so raise the stakes of a potential conflict that such proliferation would make nations more cautious and thus reduce the probability of nuclear war. Wentz argued that it might be in US interest to promote selective nuclear proliferation. Waltz argued that proliferation was, in general, stabilizing.

Intriligator and Brito (1981) showed that there generally are qualitatively different effects of the acquisition by an additional nation of nuclear weapons on the probability of nuclear war, depending on the existing number of nuclear nations. In that paper we introduced the possibility of a nuclear war as a result of an accident or other inadvertent behavior to show that eventually more nuclear powers increase the probability of nuclear war. As additional nations acquire nuclear weapons, it becomes more likely that there will be other nuclear powers prepared to exploit any postwar weakness of the initiating power, further reinforcing general deterrence and thus enhancing stability against war outbreak. The probability of a deliberate initiation of a war thus decreases as the acquisition of nuclear weapons restrains the existing nuclear nations.

Increasing the number of nuclear nations implies that a nation that initiates a war would be relatively worse off in the post-war environment, both in the case where the other nuclear nations are belligerents and in the case where they remain neutral. Increasing the number of nuclear nations also increases the uncertainty as to how other nuclear powers will react during and after a war. As the number of nuclear nations increases, however, there is a rapid increase in the number of potential nations or counter coalitions with which any initiating nuclear power would have to contend both during and after a war. As a result, the probability of a calculated attack may fall. A related treatment is that of Bueno de Mesquita and Riker (1982) that uses an expected utility model to show that the probability of a nuclear war declines as the number of nuclear nations increases.

Brito and Intriligator (1996) demonstrated, however, that a formal proof of this proposition is beyond the scope of existing models since they cannot satisfy the necessary conditions. Formal propositions about the effects of additional nuclear powers on the probability of a deliberate nuclear war require more than just qualitative assertions about the change in the probability that an individual nation will initiate a nuclear war as the number of nuclear powers increases. To make such predictions it is necessary that the model makes quantitative predictions about the magnitude in the change in these probabilities that are beyond the capability of most models in economics and political science, so the formal debate on this issue should be viewed with caution.

An implication of this analysis, however, is that of reorienting policy away from nonproliferation *per se* and toward control over accidental or inadvertent war. While proliferation of nuclear weapons may either increase or decrease the probability of war, control over accidental or irrational war must reduce the chance of nuclear war. In fact, policies and actions, both technical and political, which would reduce the chance of accidental or inadvertent war, could overcome many of the problems associated with nuclear proliferation. The nuclear nations as a group should probably be as concerned about accidental or inadvertent war as about nuclear proliferation. This policy conclusion is further strengthened when considering the relative numbers of weapons involved. There can be considerably more value in controlling the accidental or inadvertent detonation of even a few of the thousands of warheads in the current nuclear nations than in preventing the acquisition of a small number of weapons in a new nuclear nation. There may be, for example, merit in sharing technology to control nuclear weapons with the other nuclear nations or even with all potential nuclear weapons nations. Such policies, which would have the effect of lowering the probability of accidental war, can significantly offset the potentially destabilizing effects of nuclear proliferation.

4. *The Economics of Guerrilla Warfare*

Guerrilla wars are usually considered to be driven by political issues. The mention of guerrillas can conjure up images such as Spanish peasants fighting Napoleon, Lawrence of Arabia leading the Arabs in the desert against the Turks, Fidel Castro fighting in the Sierra Madre against Batista, Vietnamese peasants fighting the forces of the USA, and the Mujahedeen fighting the Afghan government. Whether or not one agrees with the political motivation of these guerrillas, it is hard to dispute the contention that their motivation was primarily political. This political motivation of the guerrillas may, however, be in the process of changing. The motivation for nations to sponsor guerrilla wars has declined dramatically since the breakup of the Soviet Union and the end of the Warsaw Pact.

As guerrilla movements lose the possibility of finding an external sponsor between antagonists in the old bipolar world, a new symbiosis may be developing between guerrilla movements and narco-traffickers. Guerrilla wars appear to be evolving from ones based on political and ideological movements to ones based on political–economic forces that appear to be supported by the international traffic in drugs. Such an

evolution is occurring in Colombia, for example, as guerrillas have become active in protecting the drug trade. At the same time, the drug lords also corrupt government forces. The result is a three-person game in which the drug lords can influence the outcome of the conflict between the government and the guerrillas by transferring resources to one side or another.

Che Guevara divided guerrilla wars into two types (Guevara 1987). He suggested that the first type is an armed group against an established government. Examples of such wars include the Mexican Revolution, the wars in Indochina and Algeria against the French, the Communist insurgencies in Greece and Malaya, and the Cuban Revolution. The second type consists of actions in support of conventional forces in the field. Examples include the Peninsular Campaign against Napoleon, the Arab revolt in World War I, and partisan actions against German and Japanese forces in World War II. Unlike the first type, the motivation of these forces is not unambiguous.

Mao Tse-tung had characterized a guerrilla war as having three stages (Mao 1954). In the first stage, the primary task of the guerrillas is to establish a force. The second stage is then one of stalemate in which the guerrillas slowly accumulate resources and establish a geographical base. The third stage, finally, is marked by the metamorphosis of the guerilla forces into conventional forces and the possible subsequent collapse of the government. Mao's observation was that '… the people are the sea and the guerrillas are the fish.' To extend his metaphor, the government forces are the sharks.

Brito and Intriligator (1989) developed an economic model of guerrilla warfare which combined two classic models from economics, the Solow growth model and the Ricardian model of economic rents, with two classic studies of guerrilla warfare, T. E. Lawrence's *Seven Pillars of Wisdom* (1936) and Mao's *On the Protracted War* (1954). The result was a model of the transition of a guerrilla war from Mao's first stage to his second. The successful transition of a guerrilla force from the first to the second stage would depend on several economic characteristics of the terrain, such as the economic resources available to the guerrillas and the government forces and the 'porosity' of the country, defined as the guerrilla's ability to obtain aid from abroad.

This transition was characterized by a stable equilibrium. A crucial element in determining whether or not a guerrilla force can make such a successful transition, however, is the existence of an outside sponsor. This characteristic of the model is supported by the history of guerrilla warfare. Most successful guerrilla wars, including the Peninsular Campaign, the revolt in the desert, and the Vietnamese war, were characterized by an external sponsoring power that supported the insurgent forces. The one possible exception in recent history is Cuba, where, according to Chapelle (1960), some 85 percent of Castro's arms and ammunition were obtained from Batista's forces in one way or another.

The primary difference is that the problem should now be treated as a three-person game. The players are the guerrillas, the government, and the drug lords. As Stackelberg leaders, the drug lords want to construct a reward schedule that will induce the government and the guerrillas to react according to their desires.

The guerrillas start, as in Mao's first stage, by holding the economically marginal sections of the country, such as mountains and jungle, which are easier to defend. As they gain ground, as in Mao's second stage, they capture increasingly developed and more open parts of the country with a transportation infrastructure that gives the advantage to the offense.

The guerrilla's goal is to maximize the present discounted value of its future consumption subject to its constraints. The government's goal is to maximize the present discounted value of its future consumption subject to its constraints. These goals reflect the assumption that, regardless of political goals, a fundamental objective of the government is that of controlling resources, whether for redistribution, for the provision of public goods, or for other possible less altruistic goals, e.g., in the case of an autocratic government.

Brito and Intriligator (1992) treated the subject of narco-terrorism, where the production of narcotics is assumed to take place in three stages. First, the raw material, poppies or coca plants, are grown on land that is not under the control of the government. Second, the raw material is gathered and processed. Third, the drug is transported to foreign markets. The first stage depends on the degree of control exercised by the guerrillas. The second stage depends on support from the guerrillas and the 'protection' they furnish, permitting the drug lords to purchase the coca leaves or poppies from the peasants and to operate the processing laboratories. The third stage depends on the support from the government in permitting access to airfields and other means of transportation. The government and the guerrillas both furnish generalized support in not attacking the crops. Clearly, the output of the raw material would be maximized if the most productive lands were to be devoted to the cultivation of coca leaves or poppies. Given that the government can be corrupted, an explanation is needed as to why these crops can only be grown in areas which the government does not control. This is equivalent to the question of why a country cannot legalize drugs and export them. This would, other things being equal, be a profitable enterprise. After all, the UK exported Scotch whisky to the USA during prohibition. A possible explanation is that, unlike alcohol, there is an international consensus about the dangers of cocaine and heroin, so that a country openly sponsoring traffic in these drugs would face severe international sanctions. Alternatively, the targeted importing countries such as the USA, would not face severe international

sanctions if they would intervene unilaterally in a country that openly sponsors the drug traffic. Thus, the drug trade can exist only under the apparent inability of the government to control the drug producing areas.

The drug lords will maximize the present value of profits subject to the dynamics that describe the evolution of the guerrilla war. The choice of strategy on the part of the drug lords, which is reflected in the transfer functions, must satisfy the incentive compatibility constraint that requires that the maximization by the guerillas and by the government yield precisely the allocation described by the drug lords. Resources can be transferred from the drug lords to either the government or the guerrillas. In fact, the drug lords currently play a role similar to that played earlier by governments, in particular, the governments of the superpowers that would try to influence the outcome of various guerrilla wars.

5. *Conclusion: The Changing Nature of War*

We now return to the three military problems of the twenty-first century. First was the question of the proliferation of weapons of mass destruction. According to the Brito–Intriligator Cardinality Theorem, however, the question of how the proliferation of nuclear weapons changes the probability of war is one that cannot be answered. Thus, any research that makes claims on this issue must be viewed with skepticism as having been designed to support a particular cause or viewpoint.

Second was the need to use force to maintain civil order under circumstances where there is weak external political support. A major problem in the twenty-first century will be the need to deploy military forces in remote areas of the world, where there are few important national interests. Such conflicts create great difficult political problems for the governments of the major powers of the USA and the European Union. The legacy of the colonial era makes direct intervention very difficult. Furthermore, there is little political support in these countries to intervene and risk lives in such a remote conflict. Armies are institutions that are very difficult to build, and they reflect the society from which they come, as the USA came to learn in Vietnam.

Third is the support of guerrilla warfare by nonstate actors. An example is the conflict in Colombia, where the raw economic motives of nonstate actors play a major role, but is not the only example of such behavior. Another example is the current conflict in various parts of Africa, where warlords are exploiting local conflict to control natural resources such as diamonds or oil. Technological progress will only make the political problems more difficult, with decreasing marginal cost technology becoming more common as computer-controlled machines are substituted for labor in most manufactured goods.

Another problem facing the developing world is the fact that raw materials are becoming less valuable. In general, technology is substituting away from metals. Silicon (sand) has replaced copper in many electronic uses and may soon replace silver in photography. Carbon fiber and similar material are replacing steel and aluminum in many structural uses. Food is in excess supply and biotechnology will continue to reduce employment in agriculture. Recent developments in thin film solar cell suggest that solar energy will be competitive with hydrocarbon fuels in the near future. If the developed world becomes economically independent of the developing world, military intervention will become politically very difficult. The end of the Cold War resulted in decreased interest in the problems of the developing world. Economic disengagement will further compound the problem.

Markets are powerful entities and will satisfy wants if these wants can be translated into demands with resources. Competent armies that fight for profit have been effective throughout history. Henry II found mercenary troops an effective solution to his problems in France in the thirteenth century, and European mercenaries were successful in Katanga until the United Nations intervened. The French Foreign Legion, although not a mercenary army, demonstrates that an effective army can be created from the dregs of humanity if it has good officers. Mercenaries supported by the major powers or multinational corporations will likely fight the wars of the twenty-first century. This process has already started. While mercenary armies seem unattractive, the alternative may be chaos and massacres, such as were witnessed recently in Africa. Clausewitz said that war was not merely a political act, but the continuation of political relations by other means, and money may be the means by which the major powers wage war in the twenty-first century. The emergence of this market may be a most interesting phenomena for economists to study.

See also: Arms Control; Cold War, The; Military Geography; Military History; Military and Politics; National Security Studies and War Potential of Nations; Peace; Terrorism; War, Sociology of

Bibliography

Boulding K E 1962 *Conflict and Defense; A General Theory*. Harper, New York

Brito D L 1972 A dynamic model of an armaments race. *International Economic Review* **13**: 359–75

Brito D L, Intriligator M D 1989 An economic model of guerrilla warfare. *International Interactions*. **15**(3/4): 203–13

Brito D L, Intriligator M D 1992 Narco-traffic and guerrilla warfare: A new symbiosis. *Defense Economics* **3**: 263–74

Brito D L, Intriligator M D 1995 Arms races and proliferation. In: Hartley K, Sandler T (eds.) *Handbook of Defense Economics*. Elsevier, Amsterdam, The Netherlands
Brito D L, Intriligator M D 1996 Proliferation and the probability of war: A cardinality theorem. *Journal of conflict Resolution* **40**(1): 206–214
Bueno de Mesquita, Riker W H 1982 An assessment of the merits of selective nuclear proliferation. *Journal of Conflict Resolution* **26**: 283–306
Chapelle D 1960 How Castro won. In: Osanka F M (ed.) *Modern Guerrilla Warfare*. Free Press of Glencoe, New York
Evron Y 1994 *Israel's Nuclear Dilemma*. Routledge, London
Frankel B (ed.) 1991*Opaque Nuclear Proliferation: Methodological and Policy Implications*. Frank Cass, London
Gallois P 1961 The balance of terror. Houghton Mifflin, Boston
Guevara E 1987 *Che Guevara and the Cuban Revolution: Writings and Speeches of Ernesto Che Guevara*. In: Deutschmann D (ed.) Pathfinder/Pacific and Asia, Sydney, Australia
Hersch S M 1991 *The Samson Option: Israel's Nuclear Arsenal and American Foreign Policy* (1st edn.). Random House, New York
Intriligator M D 1964 Some simple models of arms races. *General Systems Yearbook* **9**: 143–64
Intriligator M D 1975 Strategic considerations in the Richardson model of arms races. *Journal of Political Economy* **83**: 339–53
Intriligator M D, Brito L 1981 Nuclear proliferation and the probability of nuclear war. *Public Choice* **37**: 247–60
Isard W 1988 *Arms races, arms control, and conflict analysis*. Cambridge University Press, New York
Mao T 1954 *On the Protracted War* 1st edn. Foreign Languages Press, Peking, China
Richardson L F 1960 *Arms and Insecurity*. Boxwood Press, Pittsburgh, PA
Waltz K N 1981 *The Spread of Nuclear Weapons: More May be Better*. Adelphi Papers, no. 171. International Institute of Strategic Studies, London
Wentz W B 1968 *Nuclear Proliferation*. Public Affairs Press, Washington

D. L. Brito and M. D. Intriligator

Defense Mechanisms

Defense mechanisms are patterns of feelings, thoughts, or behaviors that are relatively involuntary. They arise in response to perceptions of psychic danger or conflict, or to unexpected change in the internal or external environment, or in response to cognitive dissonance (American Psychiatric Association 1994). They obscure or diminish stressful mental representations that if unmitigated would give rise to depression or anxiety. They can alter our perception of any or all of the following: *subject* (self?), *object* (other person), *idea*, or *emotion*. In addition, defenses dampen awareness of and response to sudden changes in reality, emotions and desires, conscience, and relationships with peoples. As in the case of physiological homeostasis, but in contrast to so-called coping strategies, defense mechanisms are usually deployed outside of awareness. Similar to hypnosis, the use of defense mechanisms compromises other facets of cognition.

General Description of Defense Mechanisms

Phenomenology and pathophysiology both play an important role in our understanding of disease. Whereas nineteenth century medical phenomenologists viewed pus, fever, pain, and coughing as evidence of disease, twentieth century pathophysiologists learned to regard such symptoms as evidence of the body's healthy efforts to cope with physical or infectious insult. In parallel fashion, much of what modern psychiatric phenomenologists classify as disorders can be reclassified by those with a more psychodynamic viewpoint as homeostatic mental mechanisms to cope with psychological stress.

The use of defense mechanisms usually alters perception of both internal and external reality. Awareness of instinctual 'wishes' is usually diminished. Alternatively, as is the case with reaction formation, antithetical wishes may be passionately adhered to. Mechanisms of defense keep affects within bearable limits during sudden changes in emotional life, such as the death of a loved one. They can deflect or deny sudden increases in biological drives, such as heightened sexual awareness and aggression during adolescence. Defenses also allow individuals a period of respite to master changes in self-image that cannot be immediately integrated. Examples of such changes might be puberty, an amputation, or even a promotion. Defenses can mitigate inevitable crises of conscience (e.g., placing a parent in a nursing home). Finally, defenses enable individuals to attenuate unresolved conflicts with important people, living or dead. In each case the individual appears to deal with a sudden stress by '*denial*' or '*repression*.' But used indiscriminately, these two terms lose their specificity. For example, the denial of the conflict laden realization 'I hate my father' might become transformed by defense to: 'My father hates me' (maladaptive projection); 'I love my father so much' (compromising reaction formation); or 'I beat my father at tennis' (adaptive sublimation).

Defense mechanisms involve far more than simple neglect or repression of reality. They reflect integrated dynamic psychological processes. Like their physiological analogues of response to tissue injury—rubor, calor, dolor, and turgor—defenses reflect healthy, often highly coordinated responses to stress rather than a deficit state or a learned voluntary defensive strategy. Thus, a defense mechanism has more in common with an opossum involuntarily but skillfully playing dead than with either the involuntary paralysis of polio or the consciously controlled evasive maneuvers of a soccer halfback.

Implicit in the concept of defense is the conviction that the patient's idiosyncratic defensive response to

stress shapes psychopathology. For example, physicians no longer classify a cough as disease, but as a coping mechanism for respiratory obstruction. There is increasing evidence that choice of defensive styles makes a major contribution to individual differences in response to stressful environments (Vaillant 1992). For example, some people respond to danger or loss in a surprisingly stoic or altruistic fashion, whereas others become phobic (displacement) or get the giggles (dissociation) or retreat into soothing reverie (autistic fantasy). These responses can be differentiated by assigning different labels to the mechanisms underlying the responses. While cross-cultural studies are still sorely needed, socioeconomic status, intelligence, and education do not seem to be causal predictors of the maturity of an individual's defenses (Vaillant 1992).

Despite problems in reliability, defenses provide a valuable diagnostic axis for conceptualizing psychopathology. By including defensive style as part of the mental status or diagnostic formulation, clinicians are better able to comprehend what is adaptive as well as maladaptive about their patients' unreasonable behavior. By reframing patients' irrational behavior in terms of defense mechanisms, clinicians learn to view qualities that initially seemed most unreasonable and unlikeable about their patients as human efforts to cope with conflict.

Taxonomy of Defense

Unfortunately, there is as much disagreement about the taxonomy of defense mechanisms as about the taxonomy of personality types, but progress is being made. First, adaptation to psychological stress can be divided into three broad classes. One class consists of voluntary cognitive or coping strategies which can be taught and rehearsed; such strategies are analogous to consciously using a tourniquet to stop one's own bleeding. The second class of coping mechanisms is seeking social support or help from others; such support seeking is analogous to seeking help from a doctor in response to bleeding. The third class of coping mechanisms are the relatively involuntary *defense mechanisms*. Such coping mechanisms are analogous to depending on one's own clotting mechanisms in order to stop bleeding.

Anna Freud (1936), Sigmund Freud's daughter, was the first investigator to introduce a formal taxonomy of defenses. By the 1970s due to the protean and ephemeral nature of defensive mental phenomena, other competing taxonomies arose (e.g., Kernberg 1975, Thomae 1987, Beutel 1988, Lazarus and Folkman 1984).

Since the 1970s, several empirical studies have suggested that it is possible to arrange defense mechanisms into a hierarchy of relative psychopathology (e.g., projection (paranoia), to displacement (phobia), to sublimation (art)), and also place them along a continuum of personality development (Vaillant 1977, Perry and Cooper 1989). With the passage of decades, a sexually abused child's defenses could mature from *acting out* (e.g., rebellious promiscuity to *reaction formation* (?joining a convent where sex is bad and celibacy is good) to altruism (as a nun in midlife counseling pregnant teenage mothers). The most recent American diagnostic manual (American Psychiatric Association 1994) provides a glossary of consensually validated definitions and arranges defense mechanisms into seven general classes of relative psychopathology (Table 1).

All classes of defenses in Table 1 are effective in 'denying' or defusing conflict and in 'repressing' or minimizing stress, but the mechanisms differ greatly in the severity of the psychiatric diagnoses assigned to their users and in their effect on long-term psychosocial adaptation. At level I, the most pathological category, are found denial and distortion of external reality. These mechanisms are common in young children, in our dreams, and in psychosis. The definition of such *denial* in Table 1 is a far more narrow but more specific use of the term than many writers who use the term *denial* interchangeably with *defense mechanisms*. Level I defenses rarely respond to simple psychological intervention. To breach them requires altering the brain by neuroleptics or waking the dreamer.

More common to everyday life are the relatively maladaptive defenses found in levels II–V. These categories are associated with adolescence, immature adults, and individuals with personality disorders. They externalize responsibility and allow individuals with personality disorders to appear to refuse help. Such defenses are negatively correlated with mental health. They profoundly distort the affective component of human relationships. Use of such defenses cause more immediate suffering to those in the environment than to the user.

Defenses in these categories rarely respond to verbal interpretation alone. They can be breached by confrontation—often by a group of supportive peers. These maladaptive defenses can also be breached by improving psychological homeostasis by rendering the individual less anxious and lonely through empathic social support, less tired and hungry through rest and food, less intoxicated through sobriety, or less adolescent through maturation.

Level VI defenses are often associated with anxiety disorders and with the psychopathology of everyday life. Level VI include mechanisms like repression (i.e., deleting the idea from a conscious emotion), intellectualization (i.e., deleting the emotion from a conscious idea), reaction formation, and displacement (i.e., transferring the emotion to a more neutral object). In contrast to the less adaptive defenses, the defenses of neurosis are manifested clinically by phobias, compulsions, obsessions, somatizations, and

Table 1
Defensive function scale

I. Level of Defensive Deregulation. This level is characterized by failure of defensive regulation to contain the individual's reaction to stressors, leading to a pronounced break with objective reality. Examples are
- delusional projection (e.g., psychotic delusions)
- psychotic denial of external reality
- psychotic distortion (e.g., hallucinations)

II. Action Level. This level is characterized by defensive functioning that deals with internal or external stressors by action or withdrawal. Examples are
- acting out apathetic withdrawal
- passive aggression
- help-rejecting complaining

III. Major Image-Distorting Level. This level is characterized by gross distortion or misattribution of the image of self or others. Examples are:
- autistic fantasy (e.g., imaginary relationships)
- splitting of self-image or image of others (e.g., making people all good or all bad)

IV. Disavowal Level. This level is characterized by keeping unpleasant or unacceptable stressors, impulses, ideas, affects, or responsibility out of awareness with or without a misattribution of these to external causes. Examples are:
- denial
- projection
- rationalization

V. Minor Image-Distorting Level. This level is characterized by distortions in the image of the self, body, or others that may be employed to regulate self-esteem. Examples are:
- devaluation
- idealization
- omnipotence

VI. Mental Inhibitions (Compromise Formation) Level. Defensive functioning at this level keeps potentially threatening ideas, feelings, memories, wishes, or fears out of awareness. Examples are
- displacement
- dissociation
- intellectualization
- isolation of affect
- reaction formation
- repression
- undoing

VII. High Adaptive Level. This level of defensive functioning results in optimal adaptation in the handling of stressors. These defenses usually maximize gratification and allow the conscious awareness of feelings, ideas, and their consequences. They also promote an optimum balance among conflicting motives. Examples of defenses at this level are
- anticipation
- altruism
- humor
- sublimation
- suppression

Source: Adapted from American Psychiatric Association *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. 1994 pp. 751–3 (APA Press Washington, DC); Used with permission.

amnesias. Such users often seek psychological help, and such defenses respond more readily to interpretation. Such defenses cause more suffering to the user than to those in the environment.

High adaptive level (VII) defenses still distort and alter awareness of and emotional response to conscience, relationships, and reality, but they perform these tasks gracefully and flexibly—writing great tragedy (i.e., sublimation), for example, is financially rewarding, instinctually gratifying, and sometimes life saving for the author. The 'distortion' involved in stoicism (suppression), humor, and altruism seems as ethical and as mentally healthy to an observer as the immature defenses seem immoral and antisocial.

Historical Background

That emotions were significant to humans had been known since ancient times, but it was difficult to develop a concept of defense until psychology could

conceptualize the interaction and/or conflict between ideas and emotions. For example, in their influential textbooks the two greatest psychologists of the nineteenth century, Wilhelm Wundt (1894) and William James (1890) discuss a psychology that consisted almost entirely of cognition, not emotion.

Appreciation of the likelihood that it was distortion of ideas *and* emotions via unconscious mechanisms of defense that led to psychopathology originated with Sigmund Freud (1894). Late in the nineteenth century neurologists had obtained a clearer understanding of the sensory and motor distribution of the nerves. They could appreciate that it was impossible for an individual to have 'glove anesthesia,' and certain common paralyses due to neurological damage. Thus, it became possible for neurological science to appreciate that patients could experience anesthesia to pin prick based on their mind's 'defensively' or 'adaptively' to distort sensory reality. Trained in both neurology and physiology, Freud was able to appreciate the importance of the phenomena to psychopathology. He suggested that upsetting emotions and ideas could be cognitively distorted in a manner similar to pain and motor behavior. Freud (1894) observed not only that emotion could be 'dislocated or transposed' from ideas (by involuntary mechanisms like dissociation, repression, and intellectualization), but also that emotion could be 'reattached' to other ideas (by the mechanism of displacement). Later, Freud expanded his theory to suggest that no experience 'could have a have a pathogenic effect unless it appeared intolerable to the patient's ego and gave rise to efforts at defense' (Freud, 1906/1964, p. 276). In other words, much mental illness like coughing and fever might be due to psychological efforts of homeostasis.

Over a period of 40 years, Freud described most of the defense mechanisms of which we speak today and identified five of their important properties. (a) Defenses were a major means of managing impulse and affect. (b) Defenses were involuntary. (c) Defenses could be distinguished from each other. (d) Defenses reflected states, not traits. They were dynamic and reversible. (e) Finally, defenses could be adaptive as well as pathological. Freud conceived of a special class of defense mechanisms—'sublimations'—that could transmute psychic conflict not into a source of pathology, but into culture and virtue (1905/1964, pp. 238–9).

From the beginning, however, defenses posed a problem for experimental psychology. First, there is no clear line between character (enduring traits) and defenses (shorter-lived responses to environment). For example, what is the difference between paranoia and projection. There is no clear line between symptoms of brain disease and unconscious coping processes. For example, sometimes one's obsessions are due to genetic factors alleviated by serotonin reuptake inhibitors and sometimes obsessions are efforts at conflict resolution via the defenses of intellectualization, displacement, and reaction formation. Sometimes the symptoms of neuropathology are wrongly imputed to be due to conflict-driven adaptive aberrations of a normal brain. In addition, behaviors associated with level VII defense can arise from sources other than conflict. For example, altruism can result from gratitude and empathy as well as from conflict.

Thus, in the 1970s both the demands for rater reliability and advances in neuropsychiatry led to psychodynamic explanations for psychopathology falling from favor. As was true in the pre-Freudian era, phobias and paranoia were again attributed more to brain pathology than to adaptive displacement and projection of conflict. Very recently, the limitations of such a purely descriptive psychiatry has led at least American psychiatrists to return defenses to their diagnostic nomenclature (Table 1).

Since the 1970s, several empirical studies well reviewed by Cramer (1991), Vaillant (1992), and Conte and Plutchick (1995) have clarified our understanding of defenses with experimental and reliability studies. This led to a tentative hierarchy and glossary of consensually validated definitions of defense being included in *DSM-IV* (American Psychiatric Association 1994), the American psychiatric manual for diagnosis.

Future Research

A major danger in using psychodynamic formulation as an explanatory concept in psychopathology remains unwarranted inference and subjective interpretation. The hierarchy of defenses presented in Table 1 is no more than a system of metaphors, and is not readily susceptible to experimental proof. But several empirical strategies permit placing defense mechanisms on a firmer footing than say handwriting analysis, phrenology, and astrology.

In the future, psychology needs to take advantage of recent advances in neuroscience to achieve further reliability, objectivity, and validity for the study of defenses.

First, we can identify defenses much as we measure the height of mountains that we cannot climb—by triangulation. We can contrast a patient's *symptoms* (or a healthy person's *creative product*) with his own self-report or *mental content* and with someone else's *objective report*, biography, or medical record. In this way an unconsciously motivated slip of the tongue can be distinguished from a simple mistake, a true paralysis can be distinguished from paresis due to psychogenic illusion, and a paranoid delusion can be distinguished from true political persecution.

The second method is predictive validity. Phrenology and astrology fail to predict future behavior assessed by independent observers. In contrast, choice of defense mechanisms robustly predicts future behavior (Vaillant 1992). Third, the increasing use of

videotape to evaluate complex human process offers the same opportunity for reliability in assessing defense mechanisms (Perry and Cooper 1989).

In the future, there is a need to study defense choice cross-culturally and to clarify the social basis for defense choice and familial transmission. As techniques for brain imaging become perfected, functional magnetic nuclear resonance (FMNR) also offers promise. So does genetics, for there is increasing evidence that personality disorders, which are often characterized by a dominant mechanism of defense, run in families. The demonstration of genetic influences upon defense choice will offer another advance that may make defenses more tangible. Finally, integrating advances in cognitive psychology and evolutionary psychology with advances in our understanding of neural networks may offer still another potential means of studying defense mechanisms with greater scientific rigor.

See also: Bleuler, Eugen (1857–1939); Conscious and Unconscious Processes in Cognition; Freud, Sigmund (1856–1939); Jung, Carl Gustav (1875–1961); Personality Theory and Psychopathology; Psychiatry, History of; Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis, History of; Psychoanalysis in Clinical Psychology; Psychoanalysis: Overview

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders* 4th edn. American Psychiatric Press, Washington, DC

Beutel M 1988 *Bewältigungsprozesse bei chronischen Erkrankungen*. VCH, Munich, Germany

Conte H R, Plutchik R 1995 *Ego Defenses: Theory and Measurement*. Wiley, New York

Cramer P 1991 *The Development of Defense Mechanisms*. Springer Verlag, New York

Freud A 1936 *The Ego and the Mechanisms of Defense*. Hogarth Press, London

Freud S 1905/1964 Three essays on the theory of sexuality. In: Strachey J (ed.) *Standard Edition*. Hogarth Press, London, Vol. 7, pp. 130–243

Freud S 1906/1964 My views on the part played by sexuality in the etiology of the neuroses. In: Strachey J (ed.) *Standard Edition*. Hogarth Press, London, Vol. 7, pp. 271–79

Freud S 1894/1964 The neuro-psychoses of defense. In: Strachey J (ed.) *Standard Edition*. Hogarth Press, London, Vol. 3, pp. 45–61

James W 1890 *The Principles of Psychology*. Henry Holt, New York

Kernberg O F 1975 *Borderline Conditions and Pathological Narcissism*. Jason Aronson, New York

Lazarus R S, Folkman S 1984 *Stress, Appraisal and Coping*. Springer, New York

Perry J C, Cooper S H 1989 An empirical study of defense mechanisms. *Archives of General Psychiatry* **46**: 444–52

Thomae H 1987 Conceptualization of responses to stress. *European Journal of Personality* **1**: 171–92

Vaillant G E 1977 *Adaptation to Life*. Little, Brown, Boston (reprinted by Harvard University Press)

Vaillant G E 1992 *Ego Mechanisms of Defense: A Guide for Clinicians and Researchers*. American Psychiatric Press, Washington, DC

Wundt W 1894 *Human and Animal Psychology*. Macmillan, New York

G. E. Vaillant

Deforestation–Forestation

Deforestation is a major global concern. Despite increased attention to various facets of this issue, however, considerable uncertainty rests in estimates of the rates, the impacts, and the driving forces of deforestation. This article reviews these issues. It focuses primarily on tropical deforestation because of its greater pace and strong social and environmental implications. Three levels of causes are discussed: the agents or proximate causes, the immediate causes, and the underlying causes. The central thesis is that deforestation occurs because it is profitable to some individuals and groups.

1. Past and Present Deforestation

1.1 Rates and Definitions

About one half (47 percent) of Earth's forest cover of 8,000 years past is gone (World Resources Institute 1998), largely the impress of human activity. Yet forests continue to occupy about one quarter of the land surface, or 3.5 million hectares (ha). Fifty-five percent are located in developing countries, primarily in the tropical world. From 1980 to 1995, developing countries lost some 200 million ha (10 percent) of their forest, while developed countries gained about 20 million ha (FAO 1999).

Deforestation is defined as the permanent transition from forest to all other land categories. *Forests* in developing countries, according to the FAO, are ecosystems with a minimum of 10 percent crown cover of trees and/or bamboos covering a minimum of 0.5 ha. Estimates of global deforestation, however, are normally based on national forest inventories which use self-defined measures and data of varying quality and commonly differ from estimates derived from satellite data. Many countries use simple extrapolations or model-based estimates. The uncertainty of deforestation estimates, therefore, must be emphasized.

Deforestation is distinguished from forest degradation or a reduction in tree density and/or increased disturbance resulting in loss of forest products and

forest-derived ecological services. Selective logging—practiced properly—does not qualify as deforestation, but might lead to forest degradation. In addition, roads constructed for this purpose make forests accessible and may attract migrants, precipitating deforestation.

1.2 The Forest Transition

Deforestation in developing countries and reforestation in developed ones form part of a global pattern in land use (see *Land Use and Cover Change*) that some envision as a long-term, temporal sequence. The forest transition captures this sequence, referring to a pattern of forest use initially leading to deforestation during extractive and agrarian episodes and subsequently to reforestation as economic activity shifts to industry and services (Grainger 1995, Mather and Needle 1998, Rudel 1998). Most developed regions of the world incurred rapid deforestation in the past, closely linked to the growth in their populations. With industrialization and urbanization, however, forest land subsequently increased beginning in the nineteenth century in Europe and the twentieth century in the United States.

This decoupling of population growth and deforestation is more complex than changing sectoral composition of the economy. Improved transport networks stimulated commercial agriculture, permitting specialized production based on local conditions. Improved agricultural technologies shifted major production to areas suitable for mechanized cultivation and adjusted agriculture to land capability, permitting the abandonment of marginal lands. Industrialization pulled people off the land and shifted energy supplies from wood to coal and later to other fossil fuels. In the latter half of the twentieth century, forest plantations have contributed to the forest transition, although they contribute only 3 percent to the world's forests at the beginning of the twenty-first century. Finally, political and cultural changes played a role, primarily through state environmental management and policy. The role of forests in stabilizing water supply and runoff was recognized; forests became more than timber.

2. Impacts of Deforestation

2.1 Appropriate vs. Inappropriate Deforestation

Not all conversion of forests to agriculture or other uses is inappropriate. Kaimowitz et al. (1998) use three criteria to define inappropriate deforestation: ecological (high adverse local or downstream effects, high biodiversity and biomass), economic (high forest value to local people and low agricultural potential), and political (none or few stakeholders gain). Appropriate deforestation, therefore, involves relatively benign consequences environmentally and on current forest uses but significant gains to a significant portion of the stakeholders.

Underlying such terms as '(in)appropriate' or 'excessive' deforestation is the concept of 'optimal levels of deforestation'—balancing all relevant costs and benefits of alternative land uses. Besides the practical data and measurement problem such an analysis entails, the concept raises critical methodological and philosophical issues. How are future costs balanced against the current ones (discounting)? How is the enormous uncertainty related to the (costs of) climate change and biodiversity loss taken into account? How are costs and benefits occurring to different groups balanced? Nevertheless, there is convincing evidence for the current pace of deforestation being too high: land managers usually fail to incorporate the downstream or long-term environmental effects of forest conversion; government policies frequently stimulate forest clearing in the service of powerful interest groups, and business may ignore or attempt to circumvent indigenous forest land rights.

2.2 Climate Change

Global concern for tropical deforestation has been strongly linked to the issue of climate change and loss in biotic diversity (see *Extinction of Species*), although increasingly ecosystem services have been invoked as well. Forests are reservoirs (storage in biomass and soils), sinks (increasing area and/or growth), and sources (release through burning, biomass decay, and soil disturbance) of carbon (FAO 1999). The Intergovernmental Panel of Climate Change (IPCC) estimates that net CO_2 emissions from land-use changes, mainly tropical deforestation, contributed about 20 percent of the global anthropogenic CO_2 emissions during the 1990s (IPCC 2000). Deforestation emits two other greenhouse gases, methane (CH_4) and nitrous oxide (N_2O), but no reliable global emission estimates exist. The main emissions of these other gases related to deforestation are not from forest clearing and burning *per se*, but rather from subsequent uses of the deforested land (Houghton 1993).

Reduced deforestation, improved land management, and more agroforestry and forest plantations can reduce the net emissions from tropical forests. The IPCC estimates that over the first half of the twenty-first century such measures could—under a reasonable scenario—amount to 12–15 percent of fossil fuel carbon emissions.

2.3 Biodiversity Loss

Tropical rain forests are the biologically richest terrestrial biome on the planet, hosting 50–90 percent of all plant species. Tropical deforestation is, therefore, the single largest threat to maintaining the planet's flora and fauna diversity. Estimates of species loss—

whether 10 or 150 species/day—are at best educated guesses, but many scientists believe that the rate of extinction at the end of the twentieth century constitutes a mass extinction on the order of those driven by nature in the geological past (see *Extinction of Species*).

High biological diversity is worth preserving for several reasons (e.g., Botkin and Talbot 1992). On utilitarian grounds, forest products provide direct benefits to humans. A large proportion of agricultural plants originates in forests, and new plants with potentially high commercial value may yet be discovered. High genetic diversity also enhances plant breeding and productivity.

In the 1990s, attention has shifted to the functioning and value of ecosystems, and the value of high biodiversity within these systems (Kremen et al. 2000). Change in ecosystem sensitivity to perturbations and loss of *resilience* (the ability of an ecosystem to absorb disturbances) follows from changes in diversity and function and can have serious consequences for local communities (Perrings et al. 1995).

Forests and the biodiversity they contain also have an aesthetic value, and the threat of extinction of 'charismatic fauna' has raised public concerns over deforestation. Finally, as stated in the UN General Assembly World Charter for Nature of 1982, species have a moral right to exist.

2.4 Forest Products

Forestry contributes about 2 percent of gross domestic product globally, and 3 percent of international merchandize trade (FAO 1997). Wood production from natural forests is shrinking due to the 'mining' of precious timber (e.g., tropical hardwoods) and deforestation, but also because of conservation (i.e., parks, biospheres, reserves). Production is shifting to intensively managed forests (i.e., plantations) and other tree media outside forests (FAO 1999).

The supply of fuelwood and charcoal, the sole or dominant source of domestic energy for more than two billion people, is threatened by the loss of forest. The most alarming reports of the 1970s and the early 1980s on the extent of fuelwood crises have passed. Yet FAO (1997) estimates about one billion people face fuelwood shortages.

Non-wood forest products (NWFPs) have drawn great attention due to their role in both rural households and national economies, and due to their potential to combine income generation and forest conservation objectives. At least 150 NWFPs are significant in international trade. Among these, plants used in pharmaceuticals, nuts, ginseng roots, cork, and oils each have an annual trade value of more than US$300 million (FAO 1997).

2.5 Local Communities

Indigenous people living close to or within the forest use a wide range of forest products. Forests provide food, fiber, fuel, medicinal products, and wood for shelter and tools. Close to 50 million of the world's 300 million indigenous people live within or depend on tropical forests for these necessities. Indian tribes in the Amazon and Dayak groups in Borneo, for example, have been put at risk by large government development projects and colonization schemes that remove natural forests. Tropical deforestation therefore threatens cultural diversity.

3. Driving Forces

The process of deforestation involves at least three levels of causes: agents, immediate causes, and underlying causes (Angelsen and Kaimowitz 1999, Contreras-Hermosilla 2000). The various agents of deforestation and their actions are the proximate sources of deforestation. Their land-use decisions are based on their individual characteristics and factors external to them, such as prices, access to technology, information, services, and infrastructure. These factors determine the choice set and the incentives for different land uses, and form the immediate (direct) causes of deforestation. In turn, they are determined by broader forces—the underlying causes: for example, markets and institutions which are strongly influenced by political decisions.

3.1 Agents of Deforestation

The agents of deforestation are those that cut the trees, including smallholders, large commercial landowners, plantation companies, and infrastructure construction agencies. The key actors vary significantly by world region. Most deforestation in Latin America links to medium and large-holder actors, especially cattle ranchers. Deforestation in Africa is largely related to smallholders converting forest and woodlands to short fallow cropland. Asia, in contrast, represents a mix between small- and large-holder operations, in which agro-industrial companies play a significant role, for example, in converting large tracts of forest to oil palm plantations in Indonesia.

Shifting cultivators have been held responsible for more than half of tropical deforestation, but their share remains controversial. Part of the controversy stems from ambiguous use of the terms 'shifting cultivation' (some include any type of smallholder farming) and 'deforestation' (temporary vs. permanent clearing). Moreover, agents often interact (e.g., logging companies provide farmers with chainsaws or

access roads) or different activities can take place *sequentially* (e.g., small farmers occupying abandoned timber concessions). Often 'there are no clear guilty or innocent parties' (Kaimowitz et al. 1998, p. 303).

3.2 Immediate Causes

The farmers or companies deforest because it is the most profitable alternative and they have the necessary means to do so. Thus, the immediate causes are those that make deforestation more profitable and possible. Higher output prices and technologies that increase yield or reduce input costs make expansion more attractive. Better access to credit and lower interest rates pull in the same direction. Higher wages, reflecting the costs of hiring labor or the best alternative use of family labor, work in the opposite direction. Reduced access costs (new or better roads) also provide a great stimulus for deforestation.

These linkages are empirically founded. A survey of more than 140 economic models of deforestation finds a broad consensus on three immediate causes of deforestation: higher agricultural prices, more/better roads, and low wages and shortage of off-farm employment (Angelsen and Kaimowitz 1999). But the effect of factors such as technology, property rights, and poverty is ambiguous.

Easy and cheap access to improved technologies has contradictory effects on deforestation (Angelsen and Kaimowitz 2001). Farmers might shift to more intensive systems and reduce their total demand for land. Yield-increasing technologies will also boost supply and put downward pressure on prices, depending on the scale of adoption and the size of markets. But agricultural land expansion might also become more profitable. The history of tropical deforestation, for example, includes a long litany of situations where new crops led to massive deforestation: bananas, cocoa, coffee, oil palm, and sugar cane in various parts of the world (Barraclough and Ghimire 1995).

Insecure property rights, in principle, reduce the incentives for long-term investments. This can make farmers opt for extensive production systems and avoid investments in forest management. But forest clearing itself is an investment as it commonly gives or strengthens farmers' rights to land. In this context, more secure land rights can actually stimulate 'land races' and deforestation.

Finally, poverty is claimed to exacerbate deforestation because poor households have short time horizons and few alternatives, making the forest frontier the employer of last resort. But poor families also lack the capital to deforest large areas, and they must be present in large numbers to have a significant impact. Their lack of alternative sources of livelihood may make the poor more concerned with the local effects of deforestation and forest degradation.

3.3 Underlying Causes

The links between underlying causes and deforestation are highly complex, difficult to establish, and tend to be location-specific, making generalizations difficult. Indeed, the empirical foundation for much of the conventional wisdom of these causes is surprisingly weak (Angelsen and Kaimowitz 1999).

Excessive deforestation can—as with other environmental problems—be seen as a problem of market failure. The existence of negative externalities (alternatively, the public goods' characteristics of many ecological services from forests) creates a conflict between individual and collective rationality. The land managers have little incentive to consider the implications of their actions on, for example, global warming. Although sketchy, calculations of such non-market values indicate that by far the largest potential lies in the carbon storage service of tropical forests (in the range of US$1,000–3,600/ha), comparable to the timber value. Interestingly, existing estimates of the pharmaceutical and ecotourism potential yield very small monetary values per ha (Pearce 1996).

The term policy failure parallels market failure in reference to government action (or inaction) leading to excessive deforestation. The reasons behind policy failures are many: inadequate government capacity to enforce regulations, and policies with unintended and unforeseen effects or with known effects yet lack of concern about them. Examples include agricultural subsidies and promotion of export crops leading to agricultural encroachment of forests. Moreover, a strong political economy dimension may reside in policy failure. Policies are often designed to benefit elite interests at the expense of marginal groups and the environment. Governments have also used forest rents to buy support from powerful groups or to gain popular support.

Historically, deforestation has been closely linked to population growth. 'Men are a foe to the woods, in proportion as their numbers increase,' argued J. Robertson (1808). But as the forest transition notion shows the link is conditioned by markets, technology, institutions, and policies. Globalization of markets will make international demand rather than local needs increasingly determine the rate and scale of forest clearing.

Another line of argument relates deforestation to declining terms of trade and international debt in developing countries, and their subsequent structural adjustment policies and economic liberalization. Countries are forced to exploit their forests to service their debt or compensate for a shortfall in export income, with environmental impacts determined by the characteristics of this use. The debt crisis and economic reforms have also led to changes in government investments and expenditures (e.g., less road building), whose net effects on forests are context specific. Angelsen and Kaimowitz (1999, p. 91) con-

clude that 'any general claims about the relations among economic liberalization, structural adjustment, and deforestation is misleading.' If anything, the evidence suggests that higher agricultural output and timber prices which normally follow as part of economic liberalization will lead to more pressure on forests.

Overall, the relative emphasis of different explanations has changed over the past decade, reflecting both changing political winds and new research, but also changes on the ground. Population (and poverty) related explanations have lost ground. 'Policy failure' remains a persuasive explanation, but probably less so now compared to the (late) 1980s and early 1990s. Many of the deforestation-inducing policies (e.g., subsidies for colonialization of the Brazilian Amazon) have been weakened or removed. There has also been a tendency to blame governments for everything that went wrong, including the destruction of rainforests, failing to recognize that some government interventions—although bad for the economic performance—probably slowed down deforestation (e.g., high taxation of agriculture). In summary, market and political economy explanations appear to be the more powerful for tropical deforestation presently. Deforestation occurs because it is profitable to some agents, and some groups have the capability to influence the policies to create or enlarge the economic opportunities for themselves.

4. *Future Challenges*

The forces driving tropical deforestation seem to outweigh those favoring forest conservation, a situation unlikely to change in the near future. One strategy to reduce losses in biotic diversity would be to concentrate conservation efforts on the most valuable sites from a biodiversity perspective. Sayer et al. (2000) argue that a set of about 100 strategically located sites or 'hot spots' covering 3–5 percent of the tropical forests might save most tropical biodiversity (also Myers et al. 2000).

The failure and impracticability of land managers to include all the costs in their decisions is the fundamental reason behind excessive deforestation. Payment for one of the most important environmental services standing forests provide—carbon storage—should serve as a strong incentive for forest conservation. The Kyoto Protocol of 1997 allows for carbon trade under the Clean Development Mechanism (CDM), including changes in greenhouse gas emissions due to direct human-induced land-use change and forestry activities. But many technical and political problems remain to be sorted out (e.g., Richards and Moura Costa 1999, IPCC 2000). It remains to be seen whether carbon trading related to forests will become the mechanism that slows the destruction of tropical forests, another form of neo-colonialism where rich countries buy large areas of tropical land and exclude local communities, or whether practical and political problems and high transaction costs will prevent beneficial effects on the fate of tropical forests.

See also: Climate Change, Economics of; Climate Policy: International; Desertification; Environment and Development; Environmental and Resource Management; Environmental Planning; Environmental Security; Environmental Vulnerability; Land Degradation; Land Reclamation; Latin American Studies: Geography

Bibliography

Angelsen A, Kaimowitz D 1999 Rethinking the causes of deforestation: Lessons from economic models. *World Bank Research Observer* **14** (1 (February)): 73–98

Angelsen A, Kaimowitz D (eds.) 2001 *Agricultural Technologies and Tropical Deforestation*. CAB International, Wallingford, UK

Barraclough S L, Ghimire K B 1995 *Forests and Livelihoods, The Social Dynamics of Deforestation in Developing Countries*. St. Martins Press, New York

Botkin D B, Talbot M 1992 Biological diversity and forests. In: Sharma N P (ed.) *Managing the World's Forests*. Kendall/Hunt, Dubuque, IA

Contreras-Hermosilla 2000 *The Underlying Causes of Forest Decline*. Center for International Forestry Research (CIFOR), Bogor, Indonesia

FAO 1997 *State of the World's Forests 1997*. Food and Agriculture Organization of the United Nations (FAO), Rome

FAO 1999 *State of the World's Forests, 1999*. Food and Agriculture Organization of the United Nations (FAO), Rome

Grainger A 1995 The forest transition: an alternative approach. *Area* **27**: 242–51

Houghton R A 1993 The role of the world's forests in global warming. In: Ramakrishna K, Woodwell G M (eds.) *World Forests for the Future: Their Use and Conservation*. Yale University Press, New Haven, CT

IPCC 2000 *Land Use, Land-use Change, and Forestry*. International Panel on Climate Change (IPCC), Cambridge University Press, Cambridge, UK

Kaimowitz D, Byron N, Sunderlin W 1998 Public policies to reduce inappropriate deforestation. In: Lutz E (ed.) *Agriculture and the Environment*. World Bank, Washington, DC

Kremen C, Niles J O, Dalton M G, Daily C, Ehrlich P R, Fay J P, Grewal D, Guillery R P 2000 Economic incentives for rain forest conservation across scales. *Science* **288**: 1828–32

Mather A, Needle C L 1998 The forest transition. A theoretical basis. *Area* **30**: 117–24

Myers N, Mittermeier R A, Mittermeier C G, da Fonseca G A B, Kent J 2000 Biodiversity hotspots for biodiversity conservation. *Nature* **403** (24 (February)): 853–8

Pearce D 1996 Global environmental value and the tropical forests: Demonstration and capture. In: Adamowitz W L, Boxall P C, Luckert M K, Phillips W E, White W A (eds.) *Forestry, Economics and the Environment*. CAB International, Wallingford, UK

Perrings C, Mäler K G, Folke C, Holling C S, Jansson B O (eds.) 1995 *Biodiversity Loss: Economic and Ecological Issues*. Cambridge University Press, Cambridge, UK
Richards M, Moura Costa P 1999 Can tropical forestry be made profitable by 'internalizing the externalities'? *Natural Resource Perspectives* **46** (October): 1–6
Robertson J 1808 *General View of Agriculture in the County of Inverness*. Board of Agriculture, London
Rudel T K 1998 Is there a forest transition? Deforestation, reforestation and development. *Rural Sociology* **63**(4): 533–52
Sayer J, Ishwaran N, Thorsell J, Sigaty T 2000 Tropical forest biodiversity and the World Heritage Convention. *Ambio* **29**(6): 302–9
World Resources Institute 1998 *World Resources 1998–1999: A Guide to the Global Environment*. World Resources Institute (WRI), Washington, DC

A. Angelsen

Deindividuation, Psychology of

According to deindividuation theory, the psychological state of deindividuation is aroused when individuals join crowds or large groups. The state is characterized by diminished awareness of self and individuality. This in turn reduces an individual's self-restraint and normative regulation of behavior. In social psychology, deindividuation is a major theory of group behavior: it provides an explanation of collective behavior of violent crowds, mindless hooligans, and the lynch mob. In addition, deindividuation has been associated with other social phenomena such as genocide, stereotyping, and disinhibition in other settings such as computer-mediated communication. Below, the theoretical evolution of deindividuation is sketched, followed by a brief impression of the empirical support for this theory. Finally, recent research is discussed, which argues for a reconceptualization of deindividuation: It appears deindividuation is not a loss of individual identity, but may be better construed as a transition to a social identity.

1. *Theoretical Evolution of Deindividuation*

Deindividuation theory is rooted in the earliest works of social psychology, in particular Gustave Le Bon's (1895/1995) crowd theory. In his book '*The Crowd*' Le Bon vividly describes how the individual in the crowd is psychologically transformed. He proposes that the psychological mechanisms of anonymity, suggestibility and contagion transform an assembly into a 'psychological crowd.' In the crowd the collective mind takes possession of the individual. As a consequence, a crowd member is reduced to an inferior form of evolution: irrational, fickle, and suggestible. The individual submerged in the crowd loses self-control and becomes a mindless puppet, possibly controlled by the crowd's leader, and capable of performing any act, however atrocious or heroic.

Although many have criticized Le Bon's theory and his politics—the two are closely linked— the influence of '*The Crowd*' in science and society is difficult to underestimate. The sales of this book are among the highest of scientific texts. In science, his analysis has inspired many theories of collective behavior, for example those of Freud, McDougall, Blumer, and Allport. In society, Le Bon's influence has been remarkably strong and more questionable: His theory informed figures such as Mussolini and Hitler, whose ideologies resonated with Le Bon's emphasis on racial factors (Nye 1975).

The field of social psychology revived Le Bon's ideas in the 1950s. They were couched in more scientific terms as a theory of 'deindividuation.' Initially, it was argued that deindividuation occurs when individuals in a group are not paid attention to as individuals (Festinger et al. 1952). Thus, being unaccountable in a crowd or group has the psychological consequence of reducing inner restraints, and increasing behavior that is usually inhibited. The causes of deindividuation were extended from anonymity in groups to other contextual factors, such as reductions of responsibility, arousal, sensory overload, a lack of contextual structure or predictability, and altered consciousness due to drugs or alcohol (Zimbardo 1969).

Thus, although Le Bon's ideas were the blueprint for the deindividuation concept, the situations that cause deindividuation were gradually expanded to include other settings than the crowd. Deindividuation theory differs from Le Bon's theory in one further important respect: The consequence of deindividuation is not that the loss of individuality is replaced by a collective mind that guides the individual's actions. Rather the loss of individuality leads to a total loss of control, and releases a person from internalized moral restraints to produce emotional, impulsive, irrational, regressive and intense behavior.

In the 1970s, deindividuation theory became a popular focus of scientific research, encouraged and inspired in part by the collective disorders characteristic of that period. However, the empirical support for deindividuation theory was not especially forthcoming. In particular, there was inconsistent evidence regarding the influence of anonymity in groups, and there was virtually no evidence for the psychological state of deindividuation (Diener 1980). Indeed, it appeared that this psychological state, which by all accounts appeared to be easily induced, was more elusive than initially conceived.

In order to come to terms with the empirical obstacles, deindividuation theory was gradually modified over the course of the 1980s. In particular, the psychological underpinnings of deindividuation have

been narrowed down. In the original version of the theory, deindividuation was closely bound up with the lack of accountability that accompanies submergence in the group. Hence, empirically the focus often was on anonymity in the group as the most important factor to produce deindividuation. Contemporary formulations of the theory focus on the psychological process of reduced (private) self-awareness as the defining feature of deindividuation (Diener 1980, Prentice-Dunn and Rogers 1982). Self-awareness entails a person's becoming the object of one's own attention. This attention for (private) aspects of oneself, such as one's attitudes and norms, increases the capacity for self-regulation. Classical and contemporary views agree, however, on the main thrust of the deindividuation hypothesis: The psychological state of deindividuation brings about antinormative and disinhibited behavior, see Postmes and Spears (1998) for a review.

2. *Deindividuation Research*

Several influential studies were conducted to illustrate the force of deindividuation. For example, Zimbardo (1969) reports a study that inspired much subsequent deindividuation research. In one study, participants were rendered anonymous by clothing them in oversized lab coats and hoods, compared with normal clothes and name tags in the control condition. The participants' task was to shock a confederate in a situation similar to the classic Milgram studies on obedience. In a first experiment using groups of female students, Zimbardo demonstrated that anonymous participants shocked longer (and therefore more painfully) than identifiable participants, in confirmation of his theory.

Aware of the limitations of Zimbardo's research in terms of external validity, Edward Diener and colleagues devised more naturalistic contexts in which to test the theory. In some studies, groups of participants were required to 'test' a so-called pacifist, trained to remain non-responsive, by hitting the pacifist with foam swords and so forth. The researchers claimed this provided a measure of aggression that was more ecologically valid for crowd behavior. Results were inconsistent, however. In one study, for example, isolated individuals displayed more aggression than groups did, see Postmes and Spears (1998) for a review.

In another study, Prentice-Dunn and Rogers (1982) induced deindividuation by instructing participants repeatedly to focus attention outward. In addition to attentional focus, participants were seated in a dimly lit room with loud rock music playing, verbal interaction was encouraged and the groups played exciting video games. In contrast, participants in an internal attention focus condition were told not to interact, performed individual tasks and played non-arousing games in a well-lit quiet room. Results showed that when attention was focused outward, higher levels of aggression were obtained.

3. *Support for Deindividuation Theory*

These examples illustrate that empirical investigations of deindividuation theory resorted to more and more extreme conditions in search of evidence for deindividuation. Although some studies found support for deindividuation theory's propositions, many failed to support it, and many reported contrary results. Even Zimbardo's (1969) original paper reports that a replication of the study mentioned above obtained the exact opposite results.

Thus, evidence for deindividuation theory appears to be mixed. A recent meta-analysis of the deindividuation literature confirms this impression. This study examines 60 tests of deindividuation theory and concludes that there is insufficient support for deindividuation theory (Postmes and Spears 1998). Disinhibition and anti-normative behavior are not more common in large groups and crowded anonymous settings. Moreover, there is no evidence that deindividuation is associated with reduced self-awareness, or even that reduced self-awareness increases disinhibition. Overall, then, deindividuation theory does not receive sufficient empirical support.

4. *Reconceptualizing Deindividuation*

One likely reason for the lack of support for the deindividuation concept is that the theory is based on Le Bon's analysis of the crowd, which is so oversimplified as to be fundamentally flawed. According to Le Bon, collective behavior is always irrational: The individual in the crowd loses cognitive control. Deindividuation theory accepts this notion. For example, Diener argued that crowd members are 'similar to the stimulus-response organism of early behaviorism, with reduced conscious mediation' (1980, p. 230). This negative and disorderly view of collective behavior stands in stark contrast to historical analyses of the crowd, which stress the crowd's capacity for restraint and orderly behavior (e.g., Rudé 1964). Moreover, these historical analyses accentuate that the crowd often acts on the basis of moral principles and popular consensus (see also Turner and Killian 1972). Thus, a compelling case can be made that collective behavior is not unrestrained (as argued by deindividuation theory) but rather restrained by normative processes. The implication is that collective behavior is under conscious control, and far from irrational and unrestrained.

In recent social psychological research, a normative analysis of collective behavior is supported. In deindividuation studies, for example, people are more likely

to follow local group norms if they are 'deindividuated' (Postmes and Spears 1998). The idea that behavior could be the result of local group norms was considered explicitly by Johnson and Downing (1979) who varied the manipulation of anonymity developed by Zimbardo. Participants were made anonymous by means of mask and overalls reminiscent of the Ku Klux Klan (similar to Zimbardo's uniforms) or by means of nurses' uniforms. Although compared to the control condition participants shocked somewhat more when dressed in the Ku Klux Klan uniforms, they actually shocked less when dressed as nurses. This finding illustrates that groups are sensitive to normative cues associated with the social context.

Researchers working within the framework of social identity theory have attempted to reconceptualize deindividuation in order to account for the crowd's normative regulation. They argue that deindividuating settings do not cause a loss of self, but they shift a person from an individual identity to a collective identity as a member of the group or crowd. For example, being in a demonstration will accentuate the unity of participants if they rally behind their cause—seemingly oblivious to strong divisions and disagreements that may exist between them. This common group membership may become accentuated if an opposing group (such as the police) acts upon the crowd as if it were one, for example by blocking its way or attempts to contain it. The strong adherence to this common group membership, which emerges over the course of such events, explains why normative regulation within the crowd may occur (Reicher 1987).

In sum, whereas deindividuation theory argues that the crowd causes a loss of identity, reverting the individual to irrationality, it seems more productive to reconceptualize deindividuation as a shift from a personal identity to a social identity, shared by members of the crowd. This may explain the normative restraint within the crowd as observed by historians, as well as the normative restraint observed in experimental studies of deindividuation.

See also: Social Psychology, Theories of

Bibliography

Diener E 1980 Deindividuation: The absence of self-awareness and self-regulation in group members. In: Paulus P B (ed.) *The Psychology of Group Influence*. Erlbaum, Hillsdale, NJ, pp. 209–42

Festinger L, Pepitone A, Newcomb T 1952 Some consequences of de-individuation in a group. *Journal of Abnormal Social Psychology* **47**: 382–89

Johnson R D, Downing L L 1979 Deindividuation and valence of cues. Effects on prosocial and antisocial behavior. *Journal of Personality and Social Psychology* **37**: 1532–38

Le Bon G 1995 *The Crowd: A Study of the Popular Mind* (Original work published in 1895). Transaction Publishers, London

Nye R A 1975 *The Origins of Crowd Psychology: Gustave Lebon and the Crisis of Mass Democracy in the Third Republic*. Sage, London

Postmes T, Spears R 1998 Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin* **123**: 238–59

Prentice-Dunn S, Rogers R W 1982 Effects of public and private self-awareness on deindividuation and aggression. *Journal of Personality and Social Psychology* **43**: 503–13

Reicher S D 1987 Crowd behaviour as social action. In: Turner J C, Hogg M A, Oakes P J, Reicher S D, Wetherell M S (eds.) *Rediscovering the Social Group: A Self-Categorization Theory*. Blackwell, Oxford, UK, pp. 107–202

Rudé G 1964 *The Crowd in History: A Study of Popular Disturbances in France and England, 1730–1848*. Wiley, New York

Turner R H, Killian L M 1972 *Collective Behavior*, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ

Zimbardo P G 1969 The human choice: Individuation, reason, and order vs. deindividuation, impulse and chaos. In: Arnold W J, Levine D (eds.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE vol. 17, pp. 237–38

T. Postmes

Deinstitutionalization, Psychiatry of

The term 'deinstitutionalization' is a somewhat cumbersome shorthand for the process that has occurred in North America and Western Europe of shifting the locus of psychiatric care from the large dedicated psychiatric hospitals to the districts in which people live. In order to appreciate the impetus behind deinstitutionalization, it is first necessary to review the reasons for the establishment of asylums. The rundown and closure of the asylums raises fundamental questions about the attitudes of the public to the mentally ill and the nature of 'the community.'

1. The Rise of the Asylum

One of the oldest psychiatric institutions in the West is the Bethlem hospital, founded as a priory in 1247 and used to house the mentally ill from 1403 to the present day. Its name, corrupted to Bedlam, attracted the stigma that became an inescapable feature of such institutions and was used to vilify the psychiatrically ill as bedlamites. Similar hospitals were established throughout Europe and North America in the ensuing centuries but the massive expansion of psychiatric hospitals, or asylums as they became known, did not occur until the nineteenth century. In the UK this was as a result of government policy to provide each county with at least one public asylum, as legislated in the County Asylum Act of 1808. Large sums of money were made available for the asylum building program,

of which the Victorians were very proud. Considerable effort went into the design of the buildings, with particular attention to the circulation of air throughout the asylum and to the external decoration.

Asylums were usually sited some miles outside of towns or cities: far enough away to avoid disturbing the public, but near enough to facilitate the transport of patients. Friern Hospital was constructed on a rural site north of London and the nearest railway line was extended to make it accessible. The hospital was opened in 1851 by Prince Albert, the husband of Queen Victoria, and was designed to house a thousand patients. It rapidly filled up and its population continued to grow way beyond the original limit, reaching a peak of 2,400 a century after its inauguration. This rapid expansion was a general feature of asylums. Once a place of containment became available, the opportunity was seized to expel from society people with a wide variety of problems who were seen as undesirable. These included people with developmental disabilities, who were sometimes sent to segregated institutions and sometimes mixed indiscriminately with the mentally ill. There were even instances of women bearing illegitimate children who, when pregnant, were sent to such institutions where they spent the rest of their life.

As their populations grew steadily, the asylums became increasingly overcrowded, until they reached the point where the beds were so close together that patients had to climb over others' beds to reach their own. Overcrowding generated squalor, and since the staff were unable to cope with the large number of disturbed patients, it was almost inevitable that abuses occurred. The public scandals that ensued hastened the demise of the asylums, but before charting their decline it is worth considering how they functioned in their heyday.

The typical asylum was either a very large building or a collection of villas set in extensive grounds which were surrounded by a high wall with a single entrance guarded by a gate lodge. Within the walls were vegetable gardens and often a farm with cows and sometimes pigs. Some asylums had their own artesian wells and many were virtually self-sufficient or even produced a surplus. Powick hospital in the west of England during the 1920s was selling one thousand bushels of wheat and 1,800 pounds of beef per year. The patients provided the labor to work the farms, as well as helping in the hospital kitchens and the laundry. The use of unpaid patient labor was viewed by the trade unions in the West as exploitation and it ceased in most hospitals in the mid-twentieth century.

When patient labor was replaced by a paid professional workforce, activities for the patients had to be organized. These took the form of recreational activities, designated as occupational therapy, and work-like activities, often termed industrial therapy. Industrial therapy ranged from simple repetitive tasks such as assembling darts to complex procedures such as making furniture designed by the staff. The industrial therapy unit at Friern Hospital, at its peak, employed 120 people, some of whom commuted daily from their homes since nothing comparable existed in the community.

2. *The Turning Point*

In the UK the number of psychiatric beds climbed steadily for more than a century, reached a peak of 149,000 in 1954, and since then has fallen progressively. A similar pattern occurred in the USA, where the number of residents in state psychiatric hospitals declined from 186 per 100,000 in 1969 to 33 per 100,000 in 1992. Psychiatric bed occupancy has also been falling in most Western European countries since the 1970s but the rate differs by country. In Italy the number of psychiatric hospital beds rose from 36,000 at the beginning of the century to a peak of 188,000 in 1971. Since then there has been a steady decline to 7,500 in 1999, including both state and private hospitals. In Israel the number of psychiatric hospital beds has fallen from more than 11,000 to 6,000 in 1999. Four factors can be identified that prompted this change: the introduction of specific antipsychotic drugs, abuse of patients by staff that provoked public scandals, critiques of institutions by sociologists and others, and a change in the attitudes of psychiatrists.

It has been claimed that the key factor was the advent of antipsychotic drugs, but that cannot be the only answer since chlorpromazine, the first of the genre, came into general use in 1955, whereas in some pioneering psychiatric hospitals in the USA and UK the bed numbers began to fall in the late 1940s. By controlling disturbed and unacceptable behavior, the drugs certainly facilitated the process of deinstitutionalization, but they did not initiate it. Similarly, public scandals gave a high profile to the problems of institutions and helped to mobilize a search for alternative forms of care, but did not start the process.

Academic books criticizing psychiatric institutions began to appear at the end of the 1950s and certainly exerted some influence on the attitudes of psychiatrists and other staff. The first was by a psychiatrist, Russell Barton, and appeared under the title *Institutional Neurosis* in 1959. The term he invented was inappropriate since the great majority of patients adversely affected by institutional living suffered from psychotic illnesses. However he accurately described the apathy, inertia, lack of motivation and affective poverty which characterised long-stay patients in institutions. This was followed by a particularly influential book by Erving Goffman (1961) entitled *Asylums*. Goffman spent a year in the late 1950s observing interactions in a US psychiatric hospital. He described the humiliating rituals which were inflicted on new entrants and

saw their main purpose as being to sever all links with the everyday world outside the institution. He coined the term 'total institution', which he defined as 'a place of residence and work where a large number of like-situated individuals, cut off from the wider society for an appreciable period of time, together lead an enforced, formally administered round of life.' Asylums certainly provided accommodation and work, but in addition many supplied every amenity likely to be required in everyday life, including a place of worship, a patients' club, hairdressers, and shops, so that there was no need ever to leave the confines of the hospital.

These books, however seminal, appeared in print too late to explain the fall in beds in the pioneering asylums. For example, Dr. Duncan Macmillan, who occupied the unique position of Medical Officer for Mental Health for Nottingham, UK, established an outpatient clinic in the local general hospital in 1932. Later he became Physician Superintendent of the psychiatric hospital, Mapperley, and by 1957 the senior members of the hospital's medical staff were spending over 50 percent of their time in 34 clinics in the community. As a consequence of these developments, the number of inpatient beds began to fall in 1949.

World War II seems to have provided psychiatrists with experiences that led to innovative experiments in therapy. This entailed the recognition of the legitimacy of fear in the face of appalling danger, and affected soldiers were treated with psychological approaches. Military psychiatrists saw healthy young men break down with severe psychiatric syndromes under stress, and then recover when treated in group settings. In these military centers every aspect of the regime was designed to counteract the development of chronicity. At the end of the war, the army psychiatrists, imbued with optimism about the potential for recovery, took up posts in the psychiatric hospitals. They included British pioneers such as Maxwell Jones (1952) who founded the Henderson hospital as a therapeutic community for people with psychopathy, and Tom Main, who took over and transformed the Cassell hospital. As directors of psychiatric hospitals, these charismatic men had the power to make their vision a reality, unimpeded by bureaucratic inertia. However, they had to fight the resistance of other members of staff, who were often more institutionalized than the patients.

In the USA the impetus for attitude change also arose from the war, but from a surprisingly different source. Conscientious objectors were obliged to do public service in psychiatric hospitals. They brought into these institutions a humanitarian attitude which had not been dulled by the hospital regime. Furthermore, as they had no professional training, they treated the patients as equals and as valued individuals. The ferment they introduced into the system worked its effects on the professional staff.

3. *Decline and Fall of the Psychiatric Hospital*

Once the antipsychotic drug, chlorpromazine, became available, the handful of pioneering hospitals were joined by the rest in discharging patients in large numbers, and a steady decline in the number of occupied beds occurred in the USA and in England and Wales. This was not observed in Scotland, which had a separate health administration. The decline in England and Wales was so steady over a number of years that Tooth and Brooke (1961) predicted that the existing long-stay population of psychiatric hospitals would be eliminated through discharges and deaths by the year 1986. The reasons their prediction proved to be wrong are of considerable import. First, some patients admitted to acute wards remained in hospital for over a year. They were termed the 'new long-stay' in distinction to the old long-stay who were a relic of custodial care. Of course the distinction between the two is arbitrary depending on the point in time when observations are made. The significance of this group is that their problems are clearly not created by the institution but are a consequence of their psychiatric illness. Hence closure of the institutions will not abolish the 'new long-stay.' The second group of patients that confounded Tooth and Brooke's prediction are those whose behavior is so disturbed and disturbing that they cannot be moved to homes in the community. These 'difficult-to-place' patients pose a major challenge to deinstitutionalization and will be considered in more detail later.

Despite these problems, the run-down of the asylums has progressed to the point where many of them have been closed. In the USA the number of state psychiatric hospitals decreased from 277 in 1970 to 231 in 1996. There has been an even more dramatic change in England and Wales where 100 of the 130 psychiatric hospitals have closed since 1980. In Italy the closure of the institutions was achieved by revolution rather than evolution. In 1978 the Italian parliament passed Law 180 which prohibited the admission of patients to psychiatric hospitals thenceforth. This legal constraint on admissions did not directly affect the patients already resident in the asylums, and it took many years to improve their conditions of life.

Law 180 reminds us that changes in psychiatric services cannot be carried through by staff alone. They require the backing of the government of the day to provide the financial resources necessary to create and develop new services. The Italian law was pushed through parliament by the Communist party, whose support was recruited by Franco Basaglia, a charismatic psychiatrist and founder of the Democratic Psychiatry movement. In the UK there was no political polarization over deinstitutionalization. The closure of psychiatric hospitals was heralded by the Conservative Minister of Health, Enoch Powell, in a famous speech in 1961 in which he referred to the

asylums as 'isolated, majestic, imperious, brooded over by the gigantic watertower and chimney combined, rising unmistakeable and daunting out of the countryside.' Successive governments, whether of the right or the left, continued to endorse the policy, but with different agendas: the left supported patients' rights for ideological reasons, while the right hoped to save money by closing the old asylums.

4. *Alternatives to the Psychiatric Hospital*

Where were patients to live when they were discharged from an asylum? Wherever the process of discharge has been studied it has emerged that the least disabled patients were selected to move out first. Many of these patients were able to cope in independent accommodation. As the more dependent patients were considered for discharge, it became evident that they needed supervision in their community homes. For the most disabled patients, 24-hour supervision was required. The closure process for two London psychiatric hospitals, Friern and Claybury, was studied by the Team for the Assessment of Psychiatric Services (TAPS) over 13 years (Leff 1997). TAPS found that 78 percent of the long-stay patients were discharged to staffed homes. In the USA, Eileen Rothbard and her colleagues (1999) monitored the closing of Philadelphia State Hospital and found that 75 percent of the long-stay population were living in residences with 24-hour staff supervision. Intensive staffing invariably carries a high cost, since over 80 percent of the cost of care in the community is contributed by staff salaries.

The obvious choice might seem to be to resettle patients with their families, but this is rarely achieved for the long-stay patients. In the Friern and Claybury programs only 4 percent of them returned to their family home. This is partly due to loss of contact with relatives over the course of decades in an institution, which had occurred to 75 percent of the Friern and Claybury patients. The other reason is that a high proportion of patients are experiencing active delusions and hallucinations which are not controlled by medication and require specialized care.

In many of the resettlement programs, the more disabled patients live in small groups in ordinary houses, indistinguishable from neighboring buildings. This has an obvious advantage in easing reintegration into the community. However there are barriers other than distinctive architecture, which became apparent from the TAPS evaluation. Nearly 700 long-stay patients were followed up for five years after discharge to homes in the community. Although there was no amelioration of their psychiatric symptoms or problems in social behavior, they did acquire skills in caring for their homes and in using community amenities. Compared to the hospital wards, they enjoyed much greater freedom in their homes, which they greatly appreciated. When asked where they wanted to live, 84 percent preferred to stay in their current home, and almost all those who wanted to move wished to live independently. Only a handful desired to return to the hospital.

Beneficial outcomes for long-stay patients discharged from psychiatric institutions have been reported by other researchers in the UK and the USA, although few have studied a large sample of patients over a long period of time. An exception is Courtney Harding and her colleagues (1987) who conducted a 32-year follow-up of 269 patients discharged from Vermont State Hospital, USA in the 1950s as part of an experimental rehabilitation program. The findings were that over half the patients were functioning much better than expected.

One drawback to life outside the asylum that TAPS revealed was the failure of most patients to develop social relationships with their neighbors. Undoubtedly the impairment of social skills by schizophrenia was contributory, but it is not the whole story. Most of the patients were resettled in cities, where there is not a strong sense of community and even healthy neighbors have little to do with each other. Geoffrey Wolff and his colleagues (1996) mounted an experimental study of an education campaign in a street in which long-stay patients from a south London asylum were resettled. The campaign was successful in reducing neighbors' fear of the mentally ill and in increasing their social contact with the new residents.

Another problem is the lack of employment prospects for people with a long history of psychiatric hospitalization. When the national unemployment rate is high, it is almost impossible for someone with such a history to obtain a job. One solution is to establish cooperatives, or 'social firms' as they are sometimes known. These are particularly well-developed in Italy, and a comparison of people with schizophrenia between Boulder, Colorado, and Bologna, Italy (Warner et al.1998) showed that the Italian patients worked for more hours and earned more than the US patients. The development of employment opportunities for mentally ill people in the community needs far more attention than it has received to date, as it is one of the prime means of social integration.

5. *Is Community Care Affordable?*

It is not difficult to provide cheap community care. There have been several low cost schemes to house discharged patients in the USA and the UK. In the USA, 'board and care' was an inexpensive way to provide patients with accommodation, but very little else. The proprietors of these homes ran them as commercial enterprises and had little or no instruction

in the care of the patients. It is hardly surprising that patients were treated extremely badly in some cases, giving rise to public scandals. In the UK, patients were placed in seaside boarding houses, many of which are empty during the winter and would otherwise have closed until the summer season. Hence the offer of permanent residents all the year round was attractive to their owners. However, no provision was made for the patients' work and leisure needs, and television documentaries showed pathetic scenes of patients wandering along deserted windswept promenades in the driving rain. Patients placed in these seaside towns were even further from their districts of origin than when in the asylums.

The crucial question is how a *good* system of community care compares with hospital care in terms of cost. There have been very few attempts to make this comparison. The TAPS project involved a collaboration with a team of health economists who carried out a detailed costing of all services used by the long-stay patients when in Friern and Claybury hospitals and then when they had been in the community for one and five years. As this was done prospectively, the costs for specific individuals can be compared over time, adjusting for inflation. For all the discharged patients, the cost in the community after the first year was slightly less than their cost when in hospital, although the community costs rose slightly by five years (Leff 1997). However, not all the long-stay patients were discharged to community homes. A sizeable group were considered to be too disturbed to move to the community. When the hospital closed these difficult-to-place patients were transferred to highly staffed facilities, mostly on other hospital sites. The high staffing levels made their care very expensive, so that when they were added into the cost equation, community care became more expensive than care in the hospital had been.

By contrast, a costing exercise by Eileen Rothbard and her group (1999) for patients discharged from Philadelphia State Hospital showed that community care was considerably cheaper than the estimated cost of hospital care. Cost comparisons on their own are insufficient: what is needed are cost-benefit analyses across sites and countries.

6. *Critiques of Deinstitutionalization*

Criticisms have come from within the psychiatric profession and from outside, mainly from the media, from disgruntled family members, and from citizens who object to sheltered homes being established in their neighborhood. One of the most stringent professional critics has been Harold Lamb (1976) in the USA who has pointed to the large number of severely mentally ill people who are in prison. He argues that patients have been transferred from a relatively benign institution, the psychiatric hospital, to a punitive one, prison, and has dubbed this process 'transinstitutionalization.' Two related issues are raised by his critique: does community care inevitably lead to the imprisonment of the mentally ill, and are the mentally ill more dangerous than the average citizen? In the TAPS study, of the 670 patients followed up for five years after discharge only five spent time in prison, while 13 (2 percent) committed assaults on others. In the study by Eileen Rothbard and colleagues (1999), 2 percent of the 321 long-stay patients followed up for three years were involved in the criminal justice system. Thus it appears that a well-organized and funded community program does not lose many patients to prison, and furthermore that the great majority of mentally ill people are not dangerous. Unfortunately the media prefer to highlight the few tragic homicides that do occur, hence reinforcing the public stereotype of the 'mad axeman.' It is this irrational fear that fuels the 'not in my backyard' attitude of some members of the public. Geoffrey Wolff and his colleagues (1996) have shown that a localized educational campaign can reassure neighbors about ex-hospital patients coming to live in their street.

The other concern of the media and the public is that many patients discharged from long-term care end up on the streets. While the number of homeless mentally ill has undoubtedly increased, several surveys have found that very few of them have spent much time in psychiatric institutions. In support of this observation, less than 1 percent of the TAPS sample became homeless over five years and only 2 percent of the Philadelphia State Hospital ex-patients over three years. While no mentally ill person should be without a home, the problem stems from the inadequacy of aftercare for those coming out of admission units and not from the long-stay patients who are resettled in sheltered residences.

This raises the question of the balance of resources to be invested in community care programs. The asylums represent a large investment of capital, in the land and buildings, and of revenue, in staff salaries. The value of the site is often considerable since urban expansion has incorporated the once distant asylums. Various financial strategies have been used to release the capital and revenue bound up in the asylum, so that they can be reinvested in providing a good quality service for the long-stay patients. But patients with long-term disabilities continue to arise in a population and need the kind of rehabilitation facilities which used to exist in the psychiatric hospitals but have generally not been re-provided in the new services. Additionally, the care of acutely ill patients is becoming increasingly difficult with the ready availability of street drugs. It seems to be difficult for funders to invest in acute and rehabilitation services at the same time as providing for the long-stay patients coming out of the institutions. Within North America and Europe, only Finland seems to have achieved a

reasonable balance between these two sectors. Without such a balance, it is not surprising that family members are discontented with the service offered to their acutely ill relatives, and that deinstitutionalization has become the target of criticisms that should justly be directed at the failure to provide a *comprehensive* psychiatric service of a uniformly high standard.

Bibliography

Etzioni A 1961 *A Comparative Analysis of Complex Organizations*. Free Press of Glencoe, New York

Goffman E 1961 *Asylums: Essays on the Social Situation of Mental Patients and other Inmates*. Anchor Books, Doubleday, New York

Harding C M, Brooks G W, Ashikaga T, Strauss J S, Breier A 1987 The Vermont longitudinal study of persons with severe mental illness, II: Long-term outcome of subjects who retrospectively met *DSM-III* criteria for schizophrenia. *American Journal of Psychiatry* **144**: 727–35

Jones M 1952 *Social Psychiatry: A Study of Therapeutic Communities*. Tavistock Publications, London

Lamb H R 1976 *Community Survival for Long-Term Patients*. Jossey Bass Publications, San Francisco

Leff J (ed.) 1997 *Care in the Community: Illusion or Reality?* John Wiley, Chichester, UK

Rothbard A B, Kuno E, Schinnar A P, Hadley T R, Turk R 1999 Service utilization and cost of community care for discharged state hospital patients: A 3-year follow-up study. *American Journal of Psychiatry* **156**: 920–7

Tooth G C, Brooke E M 1961 Trends in the mental hospital population and their effect on future planning. *Lancet* **I**: 710–13

Warner R, de Girolamo G, Belleli G, Bologna C, Fioritti A, Rosini G 1998 The quality of life of people with schizophrenia in Boulder, Colorado, and Bologna, Italy. *Schizophrenia Bulletin* **24**: 559–68

Wing J K, Brown G W 1970 *Institutionalism and Schizophrenia: A Comparative Study of Three Mental Hospitals 1960–1968*. Cambridge University Press, Cambridge, UK

Wolff G, Pathare S, Craig T, Leff J 1996 Public education for community care: A new approach. *British Journal of Psychiatry* **168**: 441–7

J. Leff

Deixis

Deixis is a type of reference constituted by the meaning of a linguistic sign being relativized to the extralinguistic context in which the sign is used. The semiotic nature of this kind of reference, its exact communicative prerequisites and functions, its acquisition by children, and its processing have long puzzled linguists, philosophers, psychologists, and anthropologists. This article presents an introduction to some of the research that has focused on deictic signs and meanings and their phenomenology.

1. Introduction

It is one of the fundamental design features of human language that the interpretation of linguistic utterances may strongly depend both on the linguistic and the nonlinguistic context. This context dependence of linguistic reference is known as *indexicality* (cf. Silverstein 1976). It has been argued by philosophers that in fact without some such underlying indexicality in all *referring* expressions, no successful reference to the world would be possible (Putnam 1975). Take, for example, the utterance *She brought this flower for me yesterday*. The noun *flower* has a meaning that can be defined independently of context (e.g., 'the reproductive organs of a plant and their colored (nongreen) envelope'), and it can be used to refer to real or imaginary flowers that fulfill this definition. Yet, in order to refer successfully to a plant as a *flower*, more than mere knowledge of the meaning of *flower* is required: speaker and addressee have to agree on the plant being identifiable by that term (botanists, for example, use the term differently—they do not require the envelope to be colored).

However, there are expressions that point to the context in their very meaning, such that they cannot be used to refer to anything before the relevant information from the context is retrieved. In the above example, the pronoun *she* takes up a referent of feminine gender that must have been introduced in the preceding stretch of discourse (if the speaker announced in the preceding utterance that he is planning to marry his girlfriend Helen, then it will be inferred that it was Helen who brought him the flowers). This illustrates *anaphoric* reference. *Deictic* reference occurs whenever a linguistic sign receives part of its meaning from the *extra*linguistic context. For example, the pronoun *me* refers to the speaker—it has a different meaning depending on who utters it. The demonstrative *this* selects a referent in the speaker's proximity—*this flower*, as opposed to *that* one over *there* (at least in its most simple spatial use).

The verb *bring* designates transport to a deictically defined location (*here*); this could be the location at which the conversation takes place, or the speaker's home (there are in fact many possibilities). The past tense of *brought* indicates that the flower arrived at this location prior to the time of utterance, and the adverb *yesterday* restricts this time interval to the day before the day of utterance. So in order to know what exactly is meant by *She brought this flower for me yesterday*, and whether this statement is true, one first needs to know who uttered it, on what day, and where.

It can be argued that reference to most objects (including people), places, and times in the real world (to be precise, to all those that neither have a proper

name nor a unique status, such as celestial bodies) ultimately requires some form of deictic anchoring. To understand this claim, one may select an object at random from the environment and try to make a statement about it that avoids any form of deixis. If the claim is correct, it follows that language could not be used to talk about the real world (other than in generic statements) without deictic reference.

2. Deictic Expression vs. Deictic Use

The class of linguistic expressions that can be *used* deictically is much larger than the class of linguistic signs with an inherently deictic meaning. Relational expressions seem to have a particular propensity of being used deictically or anaphorically. Relational terms in the domain of spatial orientation are often used deictically, even when their 'arguments' are expressed nondeictically, because they imply a particular observer perspective. Thus, *The rock is left of the tree* means that the rock is to the left of the tree as projected from an observer, as the tree has no inherent left, and *The rock is in front of the tree* means it is between the observer and the tree, as the tree has no inherent front; the perspective will usually be understood as that of the speaker, the addressee, or both. However, the same terms refer nondeictically when used with respect to people (which do have an inherent left side) or, for instance, buildings (which do have an inherent front), respectively (Fillmore 1997).

3. Deixis and Indexicality

Definite descriptions are indexical. Thus, *the chair* refers to a chair that the speaker assumes to be uniquely identifiable to the addressee, e.g., 'the aforementioned chair.' Therefore, definite descriptions can be used deictically, provided the uniqueness condition is fulfilled in the context: *Move the chair over!* will be felicitous in a situation in which there is only one free chair in the room and the addressee is aware of that. Anaphoric expressions generally admit deictic use, while anaphoric use of inherently deictic terms is usually more restricted. True time-deictic expressions apparently cannot be used anaphorically at all, and neither can genuine first or second person pronouns. The English demonstratives *this* and *that* can both be used anaphorically, though under slightly different conditions (see Fillmore 1997), and so can the proximal and distal place adverbs *here* and *there*. In contrast, in Yukatek Maya, only the distal demonstrative forms and adverbs can be used anaphorically (Hanks 1990). The combined impact of the rather weak deictic anchoring of demonstratives in some languages and the pervasiveness of their nonspatial uses has led some researchers to propose nonspatial analyses of the underlying meanings. Similarly, Wilkins and Hill (1995) show that *go* and its equivalents in other languages only acquire a deictic reading pragmatically, through the contrast with a truly deictic *come*.

4. Transposed Deixis and Text Deixis

Textual deixis occurs when (part of) an utterance or discourse is itself the referent of a deictic expression, as in *The preceding sentence contains 27 words*. It has been suggested that some form of textual deixis is present in every anaphoric reference, to the extent that anaphors direct the addressee's attention to an earlier mention of the referent. As long as it is clearly discernible as deixis, textual deixis will always have a *metalinguistic* reference. But as soon as any part of the *meaning* of an utterance becomes the target, textual deixis becomes increasingly indistinguishable from anaphoric reference. An intermediary case is constituted by reference to propositions, facts, or events (cf. Lyons 1977).

Transposed deixis is constituted by an *imagined* situation replacing the actual speech context as the 'indexical ground' of a deictic form. When the imagined situation is itself described in discourse, this may have a quasi-anaphoric effect. Consider *He arrived in Paris in early May. Now he finally had the time to explore this great city*. The deictic forms in the second sentence refer to the time and place introduced in the first sentence. However, they are relativized to the character perspective and therefore function as true deictics. This perspectivizing effect was actually one of the primary concerns of Bühler (1934), who laid the foundations of modern research on deixis.

5. The Semiotics of Deixis

Many philosophers of language have pondered the nature of deictic expressions. Perhaps the most widely known account is within the semiotics of C. S. Peirce. Signs are classified in this framework based on the relationship between their form and the object they represent into *icons* (constituted by similarity; e.g., pictures, blueprints, maps), *indexes* (constitputed by an 'existential' relationship; e.g., a pointing arrow, or smoke as a sign of fire), and *symbols* (constituted by convention; e.g., hammer and sickle as the symbol of Communism). Like all linguistic signs, deictics are fundamentally symbolic on this account; that is, their form does not stand in any *natural* relationship with the object they designate. However, deictic expressions are then often considered 'indexical symbols' (e.g., Tanz 1980) which somehow combine symbolic meaning and indexical reference. Notice, though, that it cannot be the expression itself, as a *type*, that bears the 'existential' relation in Peirce's sense (which in the case

of deictics is always a relation of spatio-temporal contiguity), but only its *use* in a particular context, as a *token* (Silverstein 1976), as an acoustic or graphic gesture, as it were (often accompanied by a manual gesture, see below). This 'token-indexicality' evidently underlies *all* deictic reference, irrespective of whether the form used deictically is itself a deictic or a nondeictic expression. What, then, is the symbolic meaning of deictic signs?

6. *The Semantics of Deixis*

As illustrated with the example *flower* above, the conventional type meaning of a nonindexical term can be conceived of as a set of criteria that has to be fulfilled by all possible referents of the term. This is called an 'intension' in the tradition of Frege and Carnap. Indexical signs do not have such intensions. Perhaps the most striking feature of indexical *meanings* is that they cannot be completely reduced to nonindexical meanings (although the indexical components of an utterance may of course be replaced by nonindexical expressions that happen to have the same *referents*). Philosophers (notably Reichenbach) have attempted to paraphrase an utterance like *I am writing this article* as 'The person who writes the sentence that contains the reference to the person at the time of writing the sentence is writing an article that contains the sentence.' This would in fact eliminate all deictic reference, but it would also fail to identify any article or person in particular, and would therefore make little sense.

D. Kaplan has suggested that the meaning of indexical expressions is not an intension, but rather a rule that directly determines a referent in the context in which the expression is used (e.g., Kaplan 1990). For example, the rule associated with the pronoun *I* may be stated as 'The referent is the person using the pronoun.' It is not the rule that contributes to the propositional content of the utterance, but merely the referent. The rule belongs to a fundamentally different type of meaning, which Kaplan calls *character*.

It is presumably due to their lack of intensionality that no more than a few distinctions of deictic reference are made in each of a limited number of highly general semantic domains in every language. Deictic expressions single out persons as participant roles with respect to the speech act (speaker, addressee, nonparticipant, and, in some languages, nonaddressed participant); they may encode the relative social distance between these (an example are honorific pronouns such as French *vous* or German *Sie*; cf. Brown and Levinson 1987); they refer to objects in space (demonstratives), to locations (place adverbs and motion verbs; cf. Rauh 1982), and to times (tenses and time adverbials). No language has been attested to have a genuine event deictic, but many languages (including Yukatek) have manner indexicals (translating 'like this/that') that are used in reference to events.

7. *The Pragmatics of Deixis*

Indexical elements have, arguably, three functions in linguistic utterances. First, they represent the referent in the utterance, as a variable of sorts. Second, they specify what may be called a *search domain* for the referent in the context. Third, they direct the addressee's *attention* to the referent. The latter two functions constitute the 'character' of the indexical sign.

The complexity of deictic search-domain distinctions varies considerably across languages (see Anderson and Keenan 1985, Weissenborn and Klein 1982). Where the English demonstratives *this* and *that* make just a two-term distinction of distance relative to the speaker, the Bantu language ChiBemba is said to distinguish proximity to the speaker, proximity to the addressee, proximity to both, distance to both, and relative proximity to the speaker *with respect* to the addressee (other languages have been claimed to make even more fine-grained distinctions, but more recent research has failed to confirm such analyses). Whereas English demonstratives distinguish essentially *relative* distance, such that *this* and *that* may be used *contrastively* within a space close to the speaker, the Yukatek system mentioned earlier operates on an *absolute* proximal–distal distinction. Similarly, whereas English tenses merely distinguish past, present, and future with respect to the utterance time, many other languages distinguish e.g., a past of the same day from a 'yesterday' past, a 'tomorrow' future from a more distant future, and so on. ChiBemba distinguishes four such 'degrees of remoteness' in the past and four in the future (Givón 1972). For an example of a complex pronominal system, see Foley (1997, pp. 112–18) on Tagalog.

The conditions for felicitous deictic reference in space may depend on whether the addressee's attention is already on the referent or not. Some languages even have a formal contrast that matches this distinction. Thus, in Yukatek, the regular demonstrative forms translating 'this' and 'that' are expanded by deictic place adverbs ('this here,' 'that there') when the addressee's attention is not on the referent. It has been argued that what is traditionally considered the 'mid-distal' demonstrative of Turkish really has a purely attention-calling function and contrasts with the proximal and distal forms only in this domain, not in terms of distance from speaker (Özyürek 1998). Pointing gestures may support attention direction, and in addition serve to narrow down the search domain. Moreover, in case more than one potential referent of the same kind occurs in the search domain selected by the deictic form, pointing gestures may be used to disambiguate the referent. There are, however, search domains which can be referred to without any additional gesture, since they are uniquely identifiable to the addressee. This is true of *here* when used to refer to the speaker's location, and also of *there*, when used to designate the addressee's location,

e.g., during a phone conversation (Fillmore 1997). The extent to which the use of spatial deictics requires accompanying gestures seems to vary both with the context of use and with the language-particular term. The attention-directing demonstrative of Turkish cannot be used without a gesture at all. In addition to pointing gestures, iconic and conventional gestures may occur with deictic terms, such as extending the hand, open palm facing up, with presentatives (*Voila!*), or iconically indicating an extension in combination with *this big* or a manner of motion with *like this*. The relationship between such gestures and the linguistic reference act is different from the function of pointing gestures.

Directing the addressee's attention to an object in space is subject to the perceptual accessibility of the object. Many languages exclude the use of certain demonstratives in reference to objects that are not visible, or provide special forms for this purpose. Yukatek has a presentative form for referents of which there is perceptual (e.g., acoustic), but not visual, evidence (Hanks 1990). In the Wakashan language Kwakwa'la of British Columbia, every noun phrase is marked for whether its referent is visible to the speaker or not (Boas 1947).

8. The Genesis of Deixis

It has been argued that pointing gestures are a proto-form of reference, and that accordingly deixis should be the earliest form of verbal reference both in the phylogenetic rise of language in human prehistory (Rolfe 1989) and ontogenetically in child language acquisition (Clark 1978). Despite their plausibility, both proposals meet with a major difficulty: the perspectivizing effect of deixis requires highly non-trivial cognitive skills. Thus, there is no conclusive evidence that free-ranging primates use pointing communicatively, and it has been shown that chimpanzees in captivity produce what are commonly considered pointing gestures irrespective of whether they are seen by the addressee or not; in other words, they seem to lack awareness of the interlocutor's mental state of attention (Povinelli et al. 1997). Likewise, it is not the case that the first referring expressions acquired by children are deictics. The first deictic forms acquired by English-learning children are personal pronouns, and the adult-like use of these does not occur before the third year of life. Perspectivizing expressions such as *in front of* are consistently used nondeictically at first, and deictic usage does not come in until the fifth year (cf. Tanz 1980).

9. Conclusions

Deictic reference plays a particularly important role in language: it serves to 'hook up' linguistic representations to the world. Deictic expressions directly point the addressee's attention to a referent given in the situation in which the utterance is made, often in combination with gestures. By relativizing the utterance to the particular context in which a particular speaker uses them, they relativize the utterance to this speaker's perspective. The cognitive demands imposed on language processing by this perspectivizing effect are in sharp contrast with the apparent semiotic primitiveness of deictic reference and with its pervasiveness in verbal communication.

See also: Pragmatics: Linguistic; Semiotics; Word Meaning: Psychological Aspects

Bibliography

Andersen S, Keenan E 1985 Deixis. In: Shopen T (ed.) *Language Typology and Language Description*. Cambridge University Press, Cambridge, UK, Vol. III, pp. 259–308

Boas F 1947 Kwakiutl grammar, with a glossary of the suffixes. *Transactions of the American Philosophical Society* **37**: 203–377

Brown P, Levinson S C 1987 *Politeness: Some Universals in Language Use*. Cambridge University Press, Cambridge, UK

Bühler K 1934 *Sprachtheorie* [*Language Theory*]. G Fischer, Jena

Clark E 1978 From gesture to word. In: Bruner J S, Garton A (eds.) *Human Growth and Development*. Clarendon Press, Oxford, UK, pp. 85–120

Fillmore C J 1997 *Lectures on Deixis*. CSLI Publications, Stanford, CA

Foley W A 1997 *Anthropological Linguistics: An Introduction*. Blackwell, Malden, MA

Givón T 1972 Studies in ChiBemba and Bantu grammar. *Studies in African Linguistics* **3**(Suppl): 1–247

Hanks W F 1990 *Referential Practice: Language and Lived Space among the Mayas*. University of Chicago Press, Chicago

Kaplan D 1990 Thoughts on demonstratives. In: Yourgrau P (ed.) *Demonstratives*. Oxford University Press, Oxford, UK, pp. 39–49

Lyons J 1977 *Semantics*. Cambridge University Press, Cambridge, UK, Vol. 2

Özyürek A 1998 An analysis of the basic meaning of Turkish demonstratives in face to face conversational interaction. In: Santi S, Guaitella I, Cave C, Konopczynski G (eds.) *Oralité et gestualité: Communication multimodale, interaction* [*Orality and Gesturality: Multi-Modal Communication, Interaction*]. L'Harmattan, Paris, pp. 609–19

Povinelli D J, Reaux J E, Bierschwale D T, Allain A D, Simon B B 1997 Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. *Cognitive Development* **12**: 327–45

Putnam H 1975 The meaning of meaning. In: Gunderson K (ed.) *Language, Mind, and Knowledge*. University of Minnesota Press, Minneapolis, MN, pp. 131–93

Rauh G 1982 On coming and going in English and German. *Papers and Studies in Contrastive Linguistics* **13**: 53–18

Rolfe L H 1989 Gesture and deixis. In: Wind J, Pulleyblank E G, De Grolier E, Bichakjian (eds.) *Studies in Language Origins*. John Benjamins, Amsterdam, pp. 25–32

Silverstein M 1976 Shifters, linguistic categories and cultural description. In: Basso K H, Selby H A (eds.) *Meaning in Anthropology*. University of New Mexico Press, Albuquerque, NM, pp. 11–56

Tanz C 1980 *Studies in the Acquisition of Deictic Terms*. Cambridge University Press, Cambridge, UK
Weissenborn J, Klein W (eds.) 1982 *Here and There: Cross-Linguistic Studies on Deixis and Demonstration*. John Benjamins, Amsterdam
Wilkins D P, Hill D 1995 When GO means COME. *Cognitive Linguistics* **6**: 209–59

J. Bohnemeyer

Delegation of Power: Agency Theory

1. The Role of Delegation

Delegation occurs when some people ask others to perform tasks on their behalf. For most of us, delegation is an important part of daily life. We delegate to grocers the task of acquiring various foods, we delegate to doctors the task of repairing broken bones, and we delegate to news agencies the task of informing us about local and world events. Delegation is beneficial because we can use it to overcome personal shortcomings. This benefit is important because each of us has limited time, energy, and talents. When the people to whom we delegate devote their time, energy, and talents to our needs, delegation increases the number of tasks that we can accomplish.

In addition to being an important part of daily life, delegation is a central concept in the conduct and the study of politics. Governments great and small use delegation to increase the range of services that they can provide. The governments of large nations, for example, delegate to defense ministries the task of maintaining national security and they delegate to finance ministries the task of managing their nation's economy. Indeed, the modern nation-state could not exist without delegation—for without delegation, lawmakers would be forced to implement and enforce every single law that they pass. With delegation, by contrast, national governments can address a wide range of social issues simultaneously.

In most polities, the most prominent form of delegation is from lawmakers (e.g., parliaments) to government agencies (e.g., economic, security, and public welfare agencies). Such agencies house the persons responsible for enforcing laws and tend to employ the nation's leading experts in the agency's policy area (e.g., economic agencies employ leading economists). In many places, these agencies are known collectively as the bureaucracy. In some places, people who work within these agencies are called bureaucrats or civil servants.

While delegation allows lawmakers to benefit from the expertise and abilities of others, it can also be hazardous. The hazards arise from the fact that delegation entails a transfer of power. For example, every time lawmakers delegate to bureaucrats, they give away a portion of their authority to govern.

The great danger of delegation is that the people to whom power is delegated will abuse the power they receive. This possibility is why some people equate delegation to abdication (e.g., a presidential resignation or a king giving up his throne). If this fear is realized, then those entrusted with the responsibility of governing lose control of what is done in the government's name and anarchy can ensue.

2. The Study of Delegation

The scholarly study of delegation occurs primarily in the fields of Political Science and Economics. This research examines when delegation confers the benefits of expertise to social decision-makers and when it leads to anarchy. It also helps scholars and practitioners around the world answer one of the most fundamental questions in politics, 'Why do governments do what they do?' For this reason, delegation has been a central focus for social scientists throughout the twentieth century.

2.1 Early Studies

In the first half of the twentieth century, scholars sounded a pessimistic tone about the consequences of delegation. Weber's (1946 p. 232) warning is among the best-known:

> Under normal conditions, the power position of a fully developed bureaucracy is always overtowering. The 'political master' finds himself in the position of the 'dilettante' who stands opposite the 'expert,' facing the trained official who stands within the management of administration. This holds whether the 'master' whom the bureaucracy serves is a 'people,' equipped with the weapons of 'legislative initiative,' the 'referendum,' and the right to remove officials, or a parliament, elected on a more aristocratic or more 'democratic' basis and equipped with the right to vote a lack of confidence …

Michels (1915) was also skeptical of delegation. In his iron law of oligarchy, he asserts that democratic organization inevitably becomes subordinated to the interest of their leaders and that the leadership becomes an entrenched oligarchy. Like most of his contemporaries and many of the scholars who followed him, Michels believed in the tendency of people to whom governing power is delegated to use their power against the interests of those who delegated such powers.

2.2 Recent Work: The Rise of Agency Theory

Modern scholarship has produced more precise insights about when delegation benefits those who delegate. Many scholars now adopt the language of

principal–agent models (i.e., agency theory) to describe the logic of delegation.

The *principal* in principal–agent theories represents someone who delegates. The *agent* represents someone to whom authority is delegated. When a lawmaker delegates authority to an agency, for example, the lawmaker is the *principal* and the agency is the *agent*.

Scholars use principal–agent models to determine when agents do (and do not) act in their principals' interests. *Agency loss* provides a common metric for their distinctions. Agency loss is the difference between the consequences of delegation for the principal and the best possible consequence. Agency loss is zero when the agent takes actions that are entirely consistent with the principal's interests. As the agent's actions diverge from the principal's interests, agency loss increases. When the agent does things that are bad for the principal, agency loss is high.

Research on delegation (see, e.g., Lupia and McCubbins 1998) shows that agency loss is minimized when two statements are true. The first statement is that the principal and agent *share common interests* (Niskanen 1971, Romer and Rosenthal 1978). In other words, the principal and agent desire the same outcomes. The second statement is that the *principal is knowledgeable* about the consequences of the agent's activities. In other words, principals know enough about their agents' actions to determine whether or not these actions serve their interests.

If either of these two statements is false, then agency loss is likely to arise. Agency loss arises when the agent and principal do not have common interests because the agent gains an incentive to act against the principal's interests. If a liberal lawmaker delegates to conservative civil servants, for example, then the agents have less of an incentive to pursue the lawmaker's favored policies than they would if they were also liberals. Agency loss arises when the principal lacks knowledge about an agent's activities because the agent can act against the principal's interests without the principal being aware of the indiscretion. If, for example, lawmakers asks the police to enforce a restriction on public drunkenness, but lack information about agency activities, then the police may not fear retribution if they choose not to enforce the restriction (see Gerber et al. 2001 for examples).

Agency loss is even more likely if neither of the two statements is true. In other words, when principals have conflicting interests with agents whose activities they cannot know, delegation is likely to be abdication. The threats are greatest in these situations because agents have both an incentive and an opportunity to act against the principal's interests without fear of retribution.

When comparing the main themes of modern and early work on delegation, we see that many early scholars reached their dour conclusions because they presumed that the two statements described above were false. More recent work disputes this presumption, uncovering the many ways in which principals can induce their agents to have common interests and learn about their agents' activities. While the new work does not totally contradict the idea that delegation can become abdication, it is more precise about when this negative outcome occurs.

A central theme in recent literature shows how principals design political institutions to affect the incentives and future actions of their agents. McCubbins et al. (1987), for example, examine how legislative decisions about the structure of political institutions affect the extent of agency loss. Drawing on a wide array of actual administrative procedures, they show that lawmakers can and do set rules for agents that reduce agency loss by ensuring that people who share their policy interests are able to influence what agents do. For example, lawmakers often not only require agents to file extensive reports on their activities, but also invite interested parties from outside the agency to do the same. These interested parties are chosen because they are known to share the lawmakers' policy preferences—including them in the process makes agents more attentive to the lawmakers' policy desires. Many studies also show how lawmakers attempt to reduce agency loss by choosing agents who share their interests (i.e., by making civil service appointments on a strictly political basis: see Epstein and O'Halloran 1999 and Laffont and Tirole 1993 for reviews of such findings).

3. *Future Trends*

Emerging trends in world politics make understanding delegation more important than ever. Many nations delegate increased responsibility to international organizations such as the European Union and the World Trade Organization. While such actions can produce lower trade barriers and increase social prosperity, they also require national governments to transfer power to others. Many people fear that such transfers will leave nations too weak to combat critical problems. As a result, the future value of research on delegation will be tied to how well it provides political leaders with the theory and evidence they need to make wise choices when delegating to others. Fortunately, today's scholars are responding to this challenge in creative and influential ways.

An emerging trend in delegation research clarifies which ways of delegating are most effective and efficient for political principals. Lupia and McCubbins (1998), for example, use principal–agent models and laboratory experiments to distinguish the types of institutional designs that are most effective in helping principals overcome problems associated with their lack of information about their agents' activities. Their results show when a wide range of political principals (i.e., voters, legislators, and jurors) can (and cannot) learn what they need to know about their agents.

Elsewhere, Epstein and O'Halloran (1999) use substantial theoretical and empirical research to distinguish when legislatures can (and cannot) benefit from narrowly constraining the agencies to which they delegate (see Huber and Shipan (2000) for a review of recent studies that adopt this perspective).

Works such as these will likely increase the extent to which real-world politicians learn the most effective and efficient ways to delegate. If they do, then delegation research will continue to give governments large and small the ability to cure broader ranges of social ills—a product of great value in an increasingly complex world.

See also: Authority: Delegation; Authority, Social Theories of; Bureaucracy and Bureaucratization; Organizations: Authority and Power; Organizations, Sociology of; Power in Society; Power: Political; Weber, Max (1864–1920)

Bibliography

Epstein D, O'Halloran S 1999 *Delegating Powers: A Transaction Cost Politics Approach to Policy Making Under Separate Powers*. Cambridge University Press, New York

Gerber E R 2001 *Stealing the Initiative: How State Government Responds to Direct Democracy*. Prentice-Hall, Upper Saddle River, NJ

Huber J D, Shipan C R 2000 The costs of control: Legislators, agencies, and transactions costs. *Legislative Studies Quarterly* **25**: 25–52

Laffont J-J, Tirole J 1993 *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA

Lupia A, McCubbins M D 1998 *The Democratic Dilemma: Can Citizens Learn What they Need to Know*. Cambridge University Press, New York

McCubbins M D, Noll R G, Weingast B R 1987 Administrative procedures as instruments of political control. *Journal of Law, Economics and Organization* **3**: 243–79

Michels R 1915 *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. Free Press, Glencoe, IL

Niskanen W A 1971 *Bureaucracy and Representative Government*. Aldine Atherton, Chicago

Romer T, Rosenthal H 1978 Political resource allocation, controlled agendas, and the status quo. *Public Choice* **33**: 27–44

Weber M 1946 Economy and society. In: Gerth H H, Mills C W (eds.) *From Max Weber: Essays in Sociology*. Oxford University Press, New York

A. Lupia

Deliberation: Political Aspects

In its political signification 'deliberation' refers to a process of reasoning about political choices. Such choices may include, among others, the selection of political leaders; the development of the laws, public policies, or regulations that govern a state; or the formation of alliances or agreements with other nations. Political deliberation may occur both within governing institutions and among the citizenry more generally. Despite an ancient lineage, the concept of political deliberation was largely ignored by political scientists for much of the twentieth century. This inattention was dramatically reversed in the last decades of the century as political theorists and experts on governmental institutions focused increasingly on the processes of deliberative decision making and on 'deliberative democracy.'

1. Origins

The first systematic treatment of political deliberation is found in Aristotle's ethical and political works from the fourth century BC. In the *Nicomachean Ethics* Aristotle defined deliberation as the reasoned consideration and investigation of 'things that are in our control and are attainable by action' (Aristotle 1934, p. 135). In political life such deliberation is a necessary element of prudence, or practical wisdom, the preeminent virtue of the statesman. Political deliberation, which draws heavily on experience (Aristotle 1934, p. 351), concerns itself with 'things that are good and bad for human beings' (Aristotle 1934, p. 337). Unlike 'execution [of laws and policies, which] should be swift,' 'deliberating takes a long time' (Aristotle 1934, p. 353).

In his *Politics* Aristotle argued that all regimes have three basic parts: one which 'deliberate[s] about common matters,' one which administers the government through various offices, and one which adjudicates disputes among citizens, and between citizens and the state (Aristotle 1984, p. 139). He called the deliberative part—which is responsible for matters of war and peace, foreign alliances, laws, the choice of public officials, and judicial cases that carry the most severe penalties (Aristotle 1984, p. 139)—the 'authoritative element of the regime' (Aristotle 1984, p. 182). In ancient democracies there were various modes for public participation in political deliberation; most popular was the assembly in which the citizens of the city-state (the free adult males) came together periodically to deliberate on and decide the great issues that faced the polity.

2. The Rise of Modern Democracies

In modern representative democracies the people in their collective capacity choose their lawmakers but normally do not otherwise engage in authoritative deliberation and decision making on matters of war and peace, laws, and foreign alliances. Many of the founders of modern democracies shared the belief, as

stated by Montesquieu, that 'the great advantage of representatives is, their capacity of discussing public affairs. For this the people collectively are extremely unfit, which is one of the chief inconveniences of a democracy' (Montesquieu 1949, p. 154). This preference for political deliberation by the people's representatives rather than by the people themselves was shared by the framers of the US Constitution. As Alexander Hamilton wrote in one of his essays defending the proposed Constitution of 1787, '[i]n the ancient republics, where the whole body of the people assembled in person, a single orator, or an artful statesman, was generally seen to rule with as complete a sway as if a scepter had been placed in his single hand.' In such assemblies (as well as in very large legislative bodies) '[i]gnorance will be dupe of cunning, and passion the slave of sophistry and declamation' (*Federalist* 58 in Hamilton et al. 1999, p. 328).

The desire to promote deliberative lawmaking also explains, in part, the development of bicameral (two-house) legislatures as a common feature in representative democracies. Those who designed the US Congress, for example, believed that one relatively large branch, whose members served short terms so as to keep them close to public opinion, should be balanced by a smaller branch with longer terms, which would function, as James Madison said about the US Senate, 'with more coolness, with more system, and with more wisdom, than the popular branch' (Farrand 1966, Vol. 1, p. 151).

As democratic institutions spread in the eighteenth century, many came to believe that elected representatives should defer to the desires of their constituents on policy matters. The issue was whose deliberations should rule: those of the lawmakers, as formed by their participation in legislative debates, or those of the people themselves? British statesman Edmund Burke gave classic expression to the case for the independent judgment and deliberation of legislators in his famous 'Speech to the Electors of Bristol' (1774). Delivered to the voters who had just elected him to parliament, Burke's speech attacked the growing sentiment that the people had a right to instruct their representatives authoritatively on legislative matters. A legislator, Burke argued, was duty bound to exercise 'his unbiased opinion, his mature judgment, [and] his enlightened conscience' even if these led to conclusions inconsistent with public desires. 'Government and legislation,' he maintained, 'are matters of reason and judgment'; and that judgment must be formed by the debate and discussion that occur within lawmaking bodies. 'Parliament is a *deliberative* assembly of *one* nation, with *one* interest, that of the whole; where … the general good, resulting from the general reason of the whole [ought to guide]' (Burke 1949, pp. 115–16). Although Burke's view was not universally accepted at the time, it became the dominant understanding of theorists and statesmen for many decades hence.

3. *Twentieth Century Developments*

In the twentieth century two powerful alternative interpretations of decision making within governing institutions pushed deliberation to the background. One was the development of the group theory, or pluralist interpretation, of democratic politics that gained prominence in the 1950s and 1960s. This view identified bargaining among group interests, rather than collective deliberation about a common good, as the decisive characteristic of governmental, and especially of legislative, decision making. The second development was the rise of rational-choice theories that posited that legislative decision making could be best understood as a process of aggregating the personal goals and ambitions of political actors. Rational-choice theorists maintained that legislators do not argue and reason together about what constitutes, or how to achieve, the public good but rather find ways through the legislative process to achieve their personal preferences.

Interest in political deliberation was rekindled in the 1980s as a result of (a) the empirical limitations of the alternative paradigms (they could not account for much of what happens in legislatures); (b) a renewed appreciation of the normative consideration that if democracies are to function well they need the capacity to engage in sound deliberation about matters of common concern (both within the governing institutions and among the populace more generally); and (c) a revival of interest in participatory politics, especially among democratic theorists.

4. *Characteristics of Deliberation*

In a process of political deliberation the actors reason together about some matter requiring a political or governmental decision. In so doing they must be open to the facts, arguments, and proposals that come to their attention, and must share a willingness to learn from their colleagues and others. Although such processes may take a variety of forms, deliberative decision making by its nature involves three essential elements: information, arguments, and persuasion. Information here refers to the substantive facts pertinent to some policy issue; arguments connect facts with desirable goals; and persuasion refers to the change or development in the minds of deliberators that results from the reasoned consideration of facts and arguments. As commonly, if not universally, understood, political deliberation is centrally concerned with the merits of public policy. It is, then, distinct from political strategizing, from reasoning about purely self-interested considerations, and from political bargaining—none of which necessarily involve a reasoned consideration of the merits of policy (Bessette 1994, pp. 46–55).

5. Deliberation and Governing Institutions

Within the governing institutions of liberal democracies deliberation is the particular, although not the exclusive, province of the legislative branch (especially in separation of powers regimes). One reason that legislatures in liberal democracies typically are relatively large collegial institutions is to promote the serious consideration of a variety of opinions on policy issues. Moreover, the internal structure (especially committee systems) and the formal rules that govern legislative procedures are designed, at least in part, to promote the careful and (generally) unhurried review of policy options. By contrast, hierarchical and centrally organized executive branches are structured less to promote deliberation than quick, decisive, and sometimes secret, decision making and action (see especially *Federalist* 70 in Hamilton et al. 1999, pp. 391–9). Members of executive branches must, nonetheless, deliberate about such matters as appointments, legislative proposals, the use of military force, and the conduct of foreign policy. The problem is that hierarchy, secrecy, and the need for quick decisions may inhibit sound deliberation.

6. Democratic Deliberation

Despite the renewed interest in the deliberative capacities and functions of governmental institutions, more scholarly attention has focused on deliberation among a democratic citizenry than among its elected representatives. Although the term 'deliberative democracy' is the one most widely used by scholars to characterize 'the idea that legitimate lawmaking issues form the public deliberation of citizens' (Bohman and Rehg 1997, p. ix), some scholars use instead such terms as 'communicative democracy,' 'discursive democracy,' and 'unitary democracy.'

Proponents of deliberative democracy tend to share the belief that policy deliberations by an energized populace will both make for better citizens and help to ensure that national policy will conform to reasonable public desires. Scholars disagree, however, as to whether and how intelligent public deliberation can occur within the large nation-state. Some look to smaller political subunits as the more likely locus of such deliberation (Mansbridge 1983). Others raise concerns about the power of government officials to manipulate public deliberations (by, for example, withholding vital information) or about the ability of the media to shape debates and outcomes (Page and Shapiro 1992, pp. 355–82, Page 1996). Another proposes, as a surrogate for national public deliberation, choosing a random sample of the populace to come together to deliberate (Fishkin 1991). Yet other scholars remain confident that new institutions and procedures (such as neighborhood assemblies, televised town meetings, office-holding by lot, and a multistage national initiative and referendum process) can promote sound public deliberation on national issues (Barber 1984). The spread of the Internet has raised new possibilities for informing and educating the citizenry, and for assessing public reaction and opinion (so-called 'electronic democracy' or 'teledemocracy').

The key issue surrounding political deliberation in liberal democracies remains, as was the case in the time of Edmund Burke, the extent to which national policy should be decided by the independent deliberations of elected officials or whether the formal decision makers should defer to, or at least take direction from, the direct deliberations of the citizenry.

See also: Aristotle (384–322 BC); Congress: United States; Democracy; Electronic Democracy; Legislatures: United States; Participation: Political; Political Representation; Rational Choice in Politics; Referendums; Representation: History of the Problem

Bibliography

Aristotle 1934 *Nicomachean Ethics* [trans. Rackham H]. Loeb Classical Library, Harvard University Press, Cambridge, MA

Aristotle 1984 *Politics* [trans. Carnes]. University of Chicago Press, Chicago

Barber B R 1984 *Strong Democracy: Participatory Politics for a New Age*. University of California Press, Berkeley, CA

Bessette J M 1994 *The Mild Voice of Reason: Deliberative Democracy and American National Government*. University of Chicago Press, Chicago

Bohman J, Rehg W 1997 (eds.) *Deliberative Democracy: Essays on Reason and Politics*. Massachusetts Institute of Technology Press, Cambridge, MA

Burke E 1949 *Burkes Politics: Selected Writings and Speeches of Edmund Burke on Reform, Revolution, and War* [ed. Hoffman R, Levack P]. Knopf, New York

Elster J (ed.) 1998 *Deliberative Democracy*. Cambridge University Press, Cambridge, UK

Farrand M 1966 (ed.) *The Records of the Federal Convention of 1787*, 4 Vols. Yale University Press, New Haven, CT

Fishkin J 1991 *Democracy and Deliberation*. Yale University Press, New Haven, CT

Granstaff B 1999 *Losing our Democratic Spirit: Congressional Deliberation and the Dictatorship of Propaganda*. Praeger, Westport, CT

Gutmann A, Thompson D 1996 *Democracy and Disagreement*. Belknap Press of Harvard University Press, Cambridge, MA

Hamilton A, Madison J, Jay J 1999 *The Federalist Papers* [ed. Rossiter C R]. New American Library, New York

Mansbridge J J 1983 *Beyond Adversary Democracy*. University of Chicago Press, Chicago

Montesquieu C 1949 *The Spirit of the Laws*. Hafner, New York

Page B I 1996 *Who Deliberates? Mass Media in Modern Democracy*. University of Chicago Press, Chicago

Page B I, Shapiro R Y 1992 *The Rational Public: Fifty Years of Trends in Americans Policy Preferences*. University of Chicago Press, Chicago
Uhr J 1998 *Deliberative Democracy in Australia: The Changing Place of Parliament*. Cambridge University Press, Cambridge, UK

J. M. Bessette

Delinquency, Sociology of

1. Sociology of Delinquency

The purpose of this article is to elucidate the major ideas and questions that have animated sociological theories of delinquency over the course of the twentieth century. Current trends and the direction that future research is likely to take in the twenty-first century are also examined. First, the study of delinquency is placed in its definitional, historical, and intellectual context.

2. What is Delinquency?

Delinquency is generally thought to mean criminal behavior committed by juveniles under the legal age of adulthood. A widely used definition was proposed by Hirschi (1969, p. 47), '[D]elinquency is defined by acts, the detection of which is thought to result in punishment of the person committing them by agents of the larger society.' Although the notion of juvenile delinquency is now ubiquitous, prior to the late nineteenth century few legal distinctions were made between juveniles and adults. Even childhood itself is a relatively modern concept.

Breaking from the tradition of treating children as 'miniature adults,' the start of the twentieth century witnessed a group of reformers in Chicago create the first juvenile court. The motivating idea was that children had special needs, and that to treat them like adults violated American ideals of fairness. The juvenile court was to act as a 'parent' for troubled youth whose own families could not, or did not, control them. The doctrine of '*parens patriae*' led to profound changes in how juveniles were treated, not only for the commission of crimes, but also for a range of so-called 'status' offenses (truancy, use of profanity, sexual precocity, tobacco use, running away, and curfew violations). Although most theories of delinquency focus on criminal rather than status offenses, versatility is the hallmark of adolescent misbehavior.

3. Multifactor Correlates of Delinquency

In the early part of the twentieth century, disciplines such as biology, psychiatry, and psychology were major players in delinquency research, along with sociology. In what was then known as a 'multifactor' approach, research proceeded in a purely empirical fashion by identifying all possible correlates of delinquency. Many factors were identified, such as poverty, being male, lack of parental discipline and supervision, poor school performance, peer delinquency, large family size, parental criminality, and immigrant status. Scholars also established a sturdy array of individual characteristics associated with delinquency, including hyperactivity, early conduct disorder, impulsiveness, and low measured intelligence. Moreover, research showed that a small group of juvenile delinquents became repeat offenders and adult 'career criminals,' reinforcing the belief that the State needed to identify and treat troubled youth at an early age. The basic correlates of delinquency have held remarkably consistent over time, as has the belief in a small group of chronic offenders that commits the majority of delinquent acts (Laub 2001).

4. Classic Sociological Theories of Delinquency

Beginning in the 1920s, the sociological study of delinquency began to take a different tack than other disciplines. Against the multifactor approach emerged a shift in attention away from the individual offender and toward the social milieu thought to produce the delinquent. Indeed, as Matza (1964) has argued, the two major contributions of sociology to the understanding of delinquency are that deviance has a history in particular locales and that delinquency best flourishes when it has group support. The 'Chicago School' provided the intellectual breakthrough that spawned this twofold contribution.

4.1 Social Disorganization

Steeped in the tradition of urban sociology pioneered by Park and Burgess at the University of Chicago, Shaw and McKay (1942) proposed a community-level approach to explaining delinquency. In their classic work, *Juvenile Delinquency and Urban Areas*, Shaw and McKay argued that three structural factors—low economic status, ethnic heterogeneity and residential instability—led to community social disorganization. In turn, delinquent gangs and the age-graded transmission of delinquent traditions ensued, accounting for the ecological concentration of delinquency. Delinquent traditions and high crime rates persisted in the same communities over time, Shaw and McKay maintained, regardless of the racial or ethnic groups passing through. Social disorganization was later reconceptualized by Kornhauser (1978) as the inability of a community to realize the common values of its residents and maintain effective social controls. The social-disorganization approach is noted for its focus

on the explanation of rates of delinquency rather than on factors that distinguish offenders from nonoffenders.

4.2 Differential Association

A related intellectual development was linked to one of the most dominant theories of the twentieth century—differential association. In 1934, Sutherland's *Principles of Criminology* was published (Sutherland 1934). As described by a leading student of the theory, 'The balance of learned criminal and anticriminal definitions determines whether one will be conforming or deviant with respect to a given legal code. If the balance of definitions is favorable to abiding by the law, the outcome is conformity; if violative definitions are in excess criminal behavior is the result' (Akers 1998, p. 27). Burgess and Akers (1966) revised the theory to argue that an individual is more likely to commit a violation when he/she differentially associates with others who commit, model, and support law violation; the violation is differentially reinforced; he/she is exposed to more deviant than conforming behavior; and his/her own learned definitions are favorable to committing deviance (Akers 1998, p. 51). Differential association theory thus turns on the idea that delinquency is learned, and that exposure to delinquent definitions (the ratio of definitions favorable to law violation over definitions unfavorable) is the key to explanation.

4.3 Anomie and Subculture

Merton's seminal article, *Social Structure and Anomie*, (1938) set the stage for another dominant strand of sociological inquiry. Merton claimed that there was a disjuncture in American society between the materialist goals to which all aspired (the 'American Dream') and the opportunities to reach them. In a stratified or unequal society coupled with a common value system that stresses material gain, strain, or 'anomie,' is inevitable. Merton developed a typology of adaptations to structurally induced strain; crime, he reasoned, was the innovative response to blocked opportunities. Merton took it for granted that anomie and crime were greatest in the lower socioeconomic strata of American society.

It fell to Cloward and Ohlin (1960) to carry Merton's theory to its logical conclusion for youth, especially adolescent males in lower-class areas of American cities. Finding routes to conventional success blocked, Cloward and Ohlin argued that deprived adolescents formed delinquent subcultures. But the opportunities for crime also varied, with some youth joining criminal subcultures, others conflict (fighting) subcultures, and still others found themselves in retreatist subcultures characterized by drugs and alcohol. Another class-based subcultural theory was proposed by Miller (1958), whereby the focal concerns of lower-class culture—especially an emphasis on 'toughness'—were argued to spawn a generating milieu of juvenile delinquency.

Cohen (1955) also wrote an influential treatise on delinquent subcultures, but his focus was on the status frustration of lower-class boys in an educational system dominated by middle-class standards. The motivation to join delinquent subcultures was not economic in origin, according to Cohen, but frustration borne of educational inadequacy and humiliation in school. The result was 'negativistic' delinquency (e.g., vandalism).

4.4 Control Theory

Rather than looking at 'push' factors to delinquency, such as blocked opportunities and subculture, control theory reverses course and sees delinquency as a lack of restraint. At the heart of control theory is the assumption that delinquent acts result when an individual's bond to society is weak or broken. Delinquency is viewed as a natural outcome of unrestrained human motives, and hence a 'pull' factor. Although there are several versions of control theory that share this basic viewpoint, by far the most influential has been that of Hirschi in *Causes of Delinquency* (1969). Hirschi posed the now-classic question about delinquency, 'Why don't we do it?' His answer is that most juveniles are bonded to society through attachment, involvement, commitment, and belief. Hirschi (1969) showed, however, that these elements of social control varied widely across adolescents and were strongly related to lower delinquency regardless of social status or class background.

4.5 Labeling and Societal Reaction

Unlike theories that explain the origin of delinquency, the innovation of labeling theory was to focus on official reactions to delinquent acts. In Becker's famous dictum, deviance is not a quality of the act a person commits, but rather a consequence of the application by others of rules and sanctions. 'The deviant is one to whom that label has successfully been applied; deviant behavior is behavior that people so label' (Becker 1963, p. 9). The insights of labeling theory led to two strands of research, one on the sources of juvenile justice sanctions, and the other on the consequences of labeling. The latter is most relevant for understanding delinquency causation in the traditional sense. Lemert (1951) maintained that

stigmatizing reactions to primary (or original) deviance created what he termed secondary deviance. In other words, whereas primary deviance arises for a wide variety of reasons, Lemert argued that secondary deviation is an adaptation to problems of adjustment created by official societal reactions. The original causes of deviation thus give way to the disapproving, degradational, and isolating reactions of society, leading to exclusion from conventional opportunities and increased contact with deviant subgroups, and finally to secondary patterns of delinquency. According to Becker (1963 pp. 24–39), delinquent careers are sustained by official labeling and sanctioning processes.

5. *Recent Currents in Delinquency Research*

Social disorganization, differential association, anomie, subcultural, control, and labeling theories have generated a plethora of empirical studies that are beyond the scope of this review. It is fair to state that in some form or another, all these theories remain viable competitors in the current explanation of variations in delinquency. However, a number of developments have taken place to complicate the theoretical story laid down by the classic perspectives.

5.1 *Measuring Delinquency*

Debates have raged for the last few decades on the proper way to measure delinquency and test the classical theoretical perspectives. There are three major sources of data that have been used, self-reports of delinquent behavior, victimization surveys, and official accounts (e.g., arrests, court records). Unfortunately, these sources of data do not always agree. In general, survey-reports of delinquency show weaker associations between social status (e.g., poverty, race, gender) and delinquency than official records. Whether or not the differential correlations of social factors with official records and self-reports reflect bias by the State or methodological weaknesses in survey research is still subject to debate. There has emerged a consensus, however, that the two sources of data yield reasonably similar patterns when the object of inquiry is serious and persistent delinquency (Hindelang et al. 1981).

5.2 *Peer Influence*

One of the sharpest disagreements in the sociology of delinquency centers on the fundamental insight (Matza 1964) that (a) most delinquency is group behavior, and (b) most delinquents have delinquent friends. Differential association theory claims, of course, that peers influence delinquency. But the hypothesis that 'birds of a feather flock together' (Glueck and Glueck 1950) was offered as a direct challenge to this interpretation, and there has been dispute ever since. It is not an exaggeration to say that after decades of research on the matter, we do not know for sure whether peers cause delinquency or delinquent adolescents choose delinquent friends. Although peers remain one of the strongest correlates of delinquency, empirical attempts to identify the causal mechanism thus continue to be paramount. One of the ways in which the mechanism is likely to be identified is through social network analysis, whereby the delinquency patterns of peer groups are independently assessed through sociometric methods.

5.3 *Routine Activities*

The 'routine activities' perspective (Cohen and Felson 1979) was built on the insight that predatory crime involves the intersection in time and space of motivated offenders, suitable targets, and the absence of capable guardians. Because illegal activities feed on the spatial and temporal structure of routine legal activities such as transportation, work, and shopping, delinquency cannot be understood apart from the ecology of everyday life. For example, the increased availability of lightweight, durable goods (radios, cell phones) and the location of stores, entertainment, subway stops, and unsupervised play spaces near schools are opportunities hypothesized to induce delinquency independent of the pool of motivated offenders. Routine activity theory represents a major shift in thinking that shapes current research.

5.4 *Age Invariance and Self-control*

With much controversy, Gottfredson and Hirschi (1990) introduced the concept of 'self control' and argued that the age–crime curve (with delinquency peaking in late adolescence) is invariant across social conditions. They proposed that low self-control causes crime at all ages; self-control differs between individuals but remains constant over the lifetime of a person. In this view, delinquent behavior is generated by heterogeneity across individuals in an underlying propensity to pursue immediate self-gratification. Because differences in self-control are alleged to emerge early in adolescence and remain stable thereafter, Gottfredson and Hirschi (1990) argue that the distinction between delinquency and adult crime is not meaningful theoretically. Attempts to test age-invariance and low self-control represent a major stream of research.

5.5 *Life-course Approaches*

Sociologists have typically addressed the problem of delinquency by studying why some adolescents commit crimes and not others, leading to between-individual analyses and attribute-based correlates

(e.g., sex, family background, personality). The life-course framework proposes a different strategy. Taking a developmental view on changes within persons, life-course criminology is motivated by an apparent paradox: although adult criminality is nearly always preceded by antisocial behavior in childhood, most antisocial children do not become criminals as adults. There is marked variability in adult outcomes even among serious and persistent juvenile delinquents. For example, Sampson and Laub (1993) found that none of the static variables measuring family background, such as poverty, parental criminality, and child supervision, predicted trajectories of adult offending among formerly incarcerated delinquents. Measures of childhood extroversion, egocentricity, aggressiveness, and difficult temperament also failed to distinguish persistent offenders from desisters. Looking forward from youth thus reveals troubled adolescents who desist, yet, apropos the paradox noted above, looking back over the careers of adult criminals exaggerates the picture of stability.

In short, background variables are surprisingly weak in the prospective explanation of trajectories of delinquency over the life course. Sampson and Laub's (1993) life-course theory argues that the reason is because within-individual changes in delinquency are mediated by proximate and time-varying social processes grounded in life transitions, situational interactions, routine activities, and turning points. They propose an age-graded theory of informal social control, where the manifestations of social bonds vary by developmental phases of aging. According to this perspective, theories limited to time-stable factors are incapable of unpacking the temporally variable pattern of individual trajectories of delinquency. Efforts to explain stability and change in delinquency over the life course, along with controversy over whether a small group of persistent offenders comprises the heart of the delinquency problem, dominate recent research and policy debates in delinquency (Laub 2001).

5.6 *Community Variations and Social Process*

A final move concerns the social mechanisms that account for the association of community-level background characteristics with rates of delinquency. A renewed tradition of ecological research has focused on the increasing concentration of poverty in cities worldwide and its implication for delinquency. If 'neighborhood effects' of concentrated poverty on delinquency exist, as hypothesized by Shaw and McKay (1942), presumably they are constituted from processes that involve collective aspects of community life (e.g., social disorganization, reduced collective efficacy). Yet studying these processes has proven difficult for a host of reasons, such as reciprocal causation (crime can cause concentrated poverty) and the differential migration of individuals into and out of communities. Moreover, neighborhood-level research has been dominated by the cross-sectional study of poverty and other demographic characteristics drawn from government statistics that do not provide information on explanatory mechanisms and dynamic processes. A growing body of research in criminology has thus committed to analyzing both structure and process at the community level (Sampson et al. 1997).

6. *Whither Sociological Delinquency?*

There is good reason to believe that the multilevel study of dynamic social processes will be at the forefront of future sociological inquiry in delinquency. Harnessed to the rich tradition of sociological theory on structured inequality, the prospects appear bright for better understanding of social processes of stability and change, within both individuals and communities.

See also: Anomie; Control: Social; Crime and Class; Crime and Gender; Crime: Sociological Aspects; Crime, Sociology of; Neutrality: Axiological; Subculture, Sociology of; Youth Gangs

Bibliography

Akers R L 1998 *Social Learning and Social Structure: A General Theory of Crime and Deviance*. Northeastern University Press, Boston

Becker H 1963 *The Outsiders: Studies in the Sociology of Deviance*. Free Press, New York

Burgess R, Akers R 1966 A differential association-reinforcement theory of criminal behavior. *Social Problems* **14**: 128–47

Cohen A 1955 *Delinquent Boys: The Culture of the Gang*. Free Press, Glencoe, IL

Cohen L, Felson M 1979 Social change and crime rate trends: a routine activity approach. *American Sociological Review* **44**: 588–608

Cloward R, Ohlin L 1960 *Delinquency and Opportunity: A Theory of Delinquent Gangs*. Free Press, Glencoe, IL

Glueck S, Glueck E 1950 *Unraveling Juvenile Delinquency*. Commonwealth Fund, New York

Gottfredson M, Hirschi T 1990 *A General Theory of Crime*. Stanford University Press, Stanford, CA

Hindelang M, Hirschi T, Weiss M 1981 *Measuring Delinquency*. Sage Publications, Beverly Hills, CA

Hirschi T 1969 *Causes of Delinquency*. University of California Press, Berkeley, CA

Kornhauser R 1978 *Social Sources of Delinquency*. University of Chicago Press, Chicago

Laub J 2001 A century of delinquency research and delinquency theory. In: Dohrn B, Rosenheim M, Tannenhaus D, Zimring F (eds.) *A Century of Juvenile Court*. University of Chicago Press, Chicago

Lemert E M 1951 *Social Pathology*. McGraw-Hill, New York

Matza D 1964 *Delinquency and Drift*. Wiley, New York

Merton R K 1938 Social structure and anomie. *American Sociological Review* **3**: 672–82

Miller W B 1958 Lower class culture as a generating milieu of gang delinquency. *The Journal of Social Issues* **14**: 5–19

Sampson R J, Laub J H 1993 *Crime in the Making: Pathways*

and Turning Points Through Life. Harvard University Press, Cambridge, MA
Sampson R J, Raudenbush S W, Earls F 1997 Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science* **277**: 918–24
Shaw C, McKay H 1942 *Juvenile Delinquency and Urban Areas*. University of Chicago Press, Chicago
Sutherland E H 1934 *Principles of Criminology*. J.B. Lippincott, Philadelphia

R. J. Sampson

Demand for Labor

Demand for labor denotes the number of workers and the intensity of effort sought by employers—companies, governments, and others that hire workers. It is derived from employers' desires to profit or otherwise to satisfy customers or clients by selling or providing products or services whose creation requires human inputs in some form. The study of labor demand focuses on the determinants of these inputs into production, how they are affected by other economic variables, and how government policies alter them. The demand for labor helps to generate and partly determines the monetary payments that provide about three-quarters of national income in most industrialized economies.

1. *Static Theory*

We can imagine an employer using technology, described by a production function, that links output Y to inputs such as labor, capital, materials, etc., all represented by L and a vector of other inputs X:

$$Y = F(L, X) \tag{1}$$

If we assume that the employer minimizes the cost of production, the demand for each input is determined by its own price (in the case of labor, the wage, w), other inputs' prices, p_x, the product's price, p, and the scale of the enterprise Y:

$$L = L(w, p_x, p, Y) \tag{2}$$

Throughout this article I subsume in w all costs of employment, including wages, nonwage monetary benefits paid to workers, and also nonmoney benefits. In studying the demand for inputs we typically set the product price equal to 1 , so that the input prices are adjusted for inflation. This labor-demand function then focuses on changes in employers' labor inputs in response to changes in the wage rate (the own-price effect), prices of other inputs (cross-price effects), and total output (the scale effect). All of these responses are determined solely by the underlying technology summarized in the production function. In the end the demand for labor is a technological relationship, but one that is generated by cost-minimizing and often profit-maximizing behavior. Its derivation from underlying economic motivations is similar to that of the consumer's choice of goods and services (see *Consumer Economics*).

A major focus since Marshall (1920) has been on the own-wage elasticity of labor demand—the percentage change in labor demanded in response to a 1 percent increase in the wage rate. Marshall offered four ' laws ' describing the determinants of this elasticity. It is higher: (a) the more readily substitutable other inputs are for labor; (b) the easier it is for customers to switch away from the product when its price rises after an increase in labor cost; (c) the greater share that labor cost represents in the enterprise's total cost; and (d) the more readily the supply of other inputs adjusts when a higher wage leads employers to attempt to switch away from labor and toward other inputs.

Measuring the wage elasticity of labor demand and the underlying technology that generates it has been the central empirical focus of the study of labor demand. The measurements answer the crucial question: if the cost of labor is raised, do employers cut the amount of labor they use, and by how much? The representation of technology that we use to infer the magnitude of this elasticity has become increasingly general over the years. The Cobb–Douglas function (Cobb and Douglas 1928) imposed the requirement that the own-wage elasticity equals $(1–s)$ at each constant output level, where s is the share of labor in total cost. The Constant Elasticity of Substitution function (Arrow et al. 1961) allowed this elasticity to assume any constant value, but it required the cross-price elasticities of labor with respect to each other input to stand in the same ratio as the other inputs' shares in costs. When we attempt to infer this elasticity today we either estimate it directly from labor-demand functions like Eqn. (2) or use generalized representations of technology, such second-order approximations as the translog function (see, e.g., Berndt and Christensen 1974). These approaches allow estimates of the own-wage and cross-price elasticities to take on any value indicated by available data on input quantities, their prices, and the scale of production.

The discussion thus far treats labor as if it were homogeneous, which is obviously wrong. One could expand Eqn. (1) to include labor disaggregated by any set of criteria of interest, for example, sex, race/ethnicity, age, educational attainment, occupation, immigrant status, and others. In each of these cases Eqn. (2) expands into a system of equations, one for each group of workers, each containing the wage rates of each type of worker, other input prices, and scale Y. These expanded representations of technology allow researchers to examine how em-

ployers substitute among workers of different types. They generate estimates of cross-wage elasticities that indicate how the demand for one type of worker changes when the wage costs of others rise or when the prices of nonlabor inputs, particularly the cost of capital services, change. These may be positive, indicating that the two types of workers are p-substitutes, or negative, indicating that they are p-complements.

Equations (1) and (2) and the expanded versions that include several types of workers assume technology is unchanging. The existence of technical progress means that the demand for workers of different types may change even without changes in wage costs. Technical progress may, as seems to have been true in industrialized countries in the 1980s and 1990s, enhance the productivity of skilled relative to less-skilled workers and thus increase the relative demand for skill (see *Innovation and Technological Change, Economics of*). With a limited supply of skilled workers this in turn raises their relative pay, partly choking off employers' desires to employ more of them (see *Wage Differentials and Structure*). The rate and nature of technical progress mean that the underlying technological parameters that determine own- and cross-wage elasticities are changing, so that the parameters themselves are not constant over time.

2. *Inferring the Structure of Labor Demand*

Moving from the theory to inferring the sizes of the responses requires ensuring that we measure employers' reactions rather than other behavior, particularly changes in the supply of labor (see *Labor Supply*). The issue is one of identification—inferring from simultaneous changes in such outcomes as wages and employment (either of all labor, or of particular groups of workers) which ones reflect employers' responses. Consider Fig. 1, which shows two data points, A and B, describing combinations of wage rates and labor employed in a market observed at two different times, or in two different labor markets. Empirical economists might connect these two points, assume that they are traced out by changes in the supply of labor from S0 to S1, and infer that they lie along the demand curve for labor D. The percentage change in L relative to the percentage change in w between these points could then be used to measure the own-wage elasticity along this demand curve.

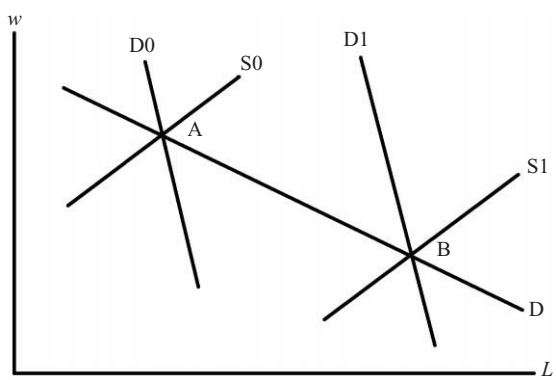

Figure 1
The labor market, illustrating the difficulty of identifying the labor-demand elasticity from combinations of wages (w) and labor (L)

Making this inference might be wrong and misleading. Points A and B could have been generated by a shift in supply from S0 to S1 that was simultaneous with a shift in demand from D0 to D1. Labor demand might have shifted because product demand increased, because technology became more labor-using, or because the price of some p-complementary input to labor fell or that of some p-substitute rose. In that case the true demand elasticity (along D0 or D1) is smaller than the estimate one would obtain by ignoring the possibility that demand shifted. One must adjust for those factors that we believe shift demand (p_x and Y in Eqn. (2)). Most importantly, one needs to find an exogenous source of variation in supply, so that we can be sure that the variations we observe in the combinations (w, L) are generated by shifts that are beyond employers' control. After substantial concern until the 1960s about properly identifying labor-demand responses, it is only since the early 1990s that research on labor demand has refocused researchers' attention on identifying changes in (w, L) combinations as reflecting movements along a demand curve.

3. *Estimates of Wage Elasticities*

3.1 *Sizes of Responses*

An immense body of empirical research has been devoted to measuring the own-wage elasticity of demand for homogeneous labor. Using data on employment, wages, and scale (output) of entire countries, major industries or small industries examined over time, or on individual employers at a point in time, researchers have attempted to pin down this central parameter in the theory of labor demand. Others have used similar data for other research purposes, with estimates of this parameter being a by-product of their investigations. Estimates have been produced for most industrialized countries and for many developing countries.

The central conclusion of the empirical literature on the demand for labor validates the principle of substitution. At a given scale of output, a rise in labor costs leads employers to substitute away from labor and toward other inputs whose cost has not risen. This is true in nearly all the over 100 studies summarized by Hamermesh (1993, Chap. 3) that were unconcerned

about problems of identification. It is also true in the newer literature that pays special attention to ensuring that the slope of the demand curve is identified (e.g., Angrist 1996). With the many different sets of data and methods, the estimates of elasticity vary widely. Nonetheless, the bulk of estimates cluster in the range -0.15 to -0.75, with a ' best-guess ' meta-analytic estimate being -0.30. At a given level of output, a rise in labor cost of 10 percent will eventually reduce employment by 3 percent as employers substitute toward technologies that use less of the now more-expensive labor and more of the now relatively cheaper other productive inputs.

The constant-output own-wage elasticity tells us the degree of substitution between labor and other inputs. That knowledge is useful, for example, for inferring how much extra employment is generated at a given level of output when employment is subsidized, or how much might be lost when payroll tax rates are raised. In order to infer the full effect on employment of labor taxes or subsidies, or of changes in other input prices, in an economy with some available labor (one that is not at full employment) we also need to add the scale effect of changes in labor costs. The evidence suggests that the scale effect is fairly substantial, probably larger than the substitution effect, so that the total elasticity of labor demand for the average employer may be as large as -1.0. Interestingly, this estimate, based on a very large body of empirical studies, is exactly what is implied by the simplest representation of technology, the Cobb–Douglas production function.

When Eqn. (2) is expanded into a system of demand equations describing several types of labor, our ability to draw fairly tight inferences about the sizes of the own- and cross-wage elasticities is diminished by the sheer number of possible disaggregations. The current literature says nothing definite about the absolute sizes of particular own- and cross-wage elasticities, and making progress on this front is high on the agenda for research in the next decade. Nonetheless, the literature does offer some secure conclusions about such parameters for skills defined generally. There is very strong empirical support for the notion of capital-skill complementarity (Griliches 1969)—cuts in the cost of capital services not only lead to more capital being used, but also lead employers to use relatively more skilled and fewer unskilled employees at each output level.

One might array workers by their level of skill and also use the empirical literature to infer the relative sizes of the own-wage demand elasticities for skill. Examples include rankings from most to least formal education, from longest to shortest tenure with their employer, or from technical/professional worker to laborer. Arraying workers by skill yields the inference that own-wage demand elasticities are more negative—larger responses of labor demand to the same percentage change in labor costs—the less skill is embodied in the group of workers. This conclusion is supported by a wide variety of studies within individual economies, but it also seems valid in comparing own-wage demand elasticities for homogeneous labor across economies at different stages of economic development.

3.2 *Usefulness for Policy*

Estimates of own- and cross-wage elasticities are central to debates over a large variety of social and economic policies. If, for example, we wish to determine how imposing a higher payroll tax to finance national health insurance might affect employment, knowledge of the own-wage demand elasticity for labor is central to predicting employers' responses. Many governments encourage employers to add particular types of workers (typically low-skilled workers) by subsidizing their employment cost using targeted wage subsidies. The success of such policies depends on the own-wage demand elasticity for the targeted group. Many payroll taxes have ceilings on the employee's earnings on which the employer must pay the tax. Others exempt some initial amount of earnings from the tax. A rise in the payroll tax ceiling implicitly raises the relative cost of higher-wage workers, leading employers to substitute toward lower-wage workers, with the extent of substitution determined by the cross-wage elasticity of demand between high- and low-wage workers. A rise in the level of initial earnings exempt from payroll taxes induces a similar substitution toward low-wage workers.

Many countries attempt to stimulate job creation by offering employers tax credits to add new equipment and buildings. These policies are often targeted toward geographic areas with high unemployment or unusually high rates of poverty. They may increase employment through a scale effect, as the cut in capital costs leads firms to expand their demand for both capital and labor. But capital-skill complementarity suggests that most of the employment gains will be by more skilled workers. If there are few skilled workers in the area, Marshall's Fourth Law suggests that even the scale effect is unlikely to generate much increase in employment.

Other labor-market policies impose restrictions on the price at which labor can be sold. Minimum wages are the most well-known example and have been extensively studied in numerous countries. The central labor-market issue is the extent to which higher minimum wages alter employment (although the central policy issue may be whether and to what extent they raise total earnings of low-skilled workers). Their effect on employment depends on the own-wage elasticity of demand for the low-skilled workers whose employment opportunities are likely to be affected when the minimum is raised, an elasticity that is quite large. It also depends on the number of workers whose

wages are so low that they might be affected by a higher minimum. Even with a high own-wage demand elasticity, raising the minimum has very small employment effects if there are few such workers. Thus job losses when minimum wages are raised are higher in countries that have maintained a high minimum wage relative to the average wage (historically, for example, France) than in those where the minimum wage has been relatively low (historically, for example, the United States).

4. *Demand for Workers and Hours*

Labor inputs consist of workers (E) and the intensity at which they work. One measure of intensity is hours worked (H) per time period (typically per week in the data we use to examine this issue). The term L in Eqn. (1) can be written as the function:

$$L = L(E, H) \tag{3}$$

A pair of demand equations expanding on Eqn. (2) for workers and hours can be derived from this expansion of Eqn. (1) and includes the prices of each. The difficulty lies in specifying the components of the prices of an extra worker and an extra hour of labor. A higher hourly wage rate raises the prices of both an extra worker and an extra hour. An increase in the payroll tax rate per hour worked, for example, also raise the prices of both. A mandate that employers provide health insurance coverage for all workers, regardless of how many hours they work, raises the price of an extra employee with no impact on the price of an additional hour per employee. These latter two examples demonstrate that determining how a particular increase in labor costs affects the relative prices of workers and hours depends on institutional specifics—changes in nonwage costs can affect both prices and do so differently depending on institutional structures (Hart 1987).

That an increase in the price of an employee, for example, an increase in hiring costs, induces firms to use fewer workers and lengthen workweeks is strongly supported by nearly all empirical research on the issue. Consensus estimates of the magnitudes of the own- and cross-price elasticities for these two facets of labor are not available. Within the range of weekly hours observed in industrialized economies (where issues of worker fatigue are typically unlikely to be crucial), however, employers appear able to substitute fairly readily between workers and hours as their relative costs change.

The major interest in the demand for workers and hours has been for evaluating policies that penalize overtime work. For many years in the United States employers have been required to pay a 50 percent overtime penalty for weekly hours beyond standard hours of 40; many Western European countries impose standard hours below 40, although often at fairly low penalty rates. Overtime penalties are imposed in the hope of generating additional jobs by shifting employers from more intensive use (longer hours) of a few workers to shorter hours for more workers. Early evidence that ignored identification (summarized in Hamermesh 1993, Table 3.11) suggests that this substitution does occur when the penalty is raised, a finding confirmed by research which ensures that the relative cost changes are exogenous to employers (Hamermesh and Trejo 2000). However, when one factors in the scale effect that arises as employers cut back labor demand generally because the higher penalty raises labor costs, the ability of higher overtime penalties to generate additional jobs appears quite small.

Reduced standard hours have been introduced in many countries as a way of spreading work. This policy change has subtle effects. Some hours of work that previously did not require overtime penalties suddenly become liable for penalties, thus reducing employers' demands for hours and leading them to substitute toward more workers. It also increases the cost of employees generally, leading to a decline in the demand for labor and substitution toward nonlabor inputs. The evidence, including (Hunt 1999) who accounts for the identification problem, suggests that mandated cuts in standard hours lead to cuts in actual hours, but not quite on a one-for-one basis. Total employment rises, but by a smaller percentage than hours fall, so that total labor input and presumably total production drop.

5. *New Employers, New Jobs*

Labor-demand functions like Eqn. (2) are based on the behavior of a firm that lasts forever; but most companies do not survive even 10 years. A line of research that accelerated in the late 1980s pointed out the need to abandon the notion that all labor demand is by existing firms. That research has focused on accounting for gross flows of jobs. It has demonstrated that about one-third of new jobs that are added in developed economies each year are in new firms, and about one-third of jobs that end disappear when firms close down (Hamermesh 1993, Davis and Haltiwanger 1992).

Research has hardly begun to study how job creation is affected by wages and changes in product demand as they lead new firms to open and existing companies to close. Evidence on these issues is important for analyzing the impact on employment of such policies as tax holidays that local governments offer to firms. The ideal would be estimates of opening- and closing-wage elasticities of labor demand that indicate how drops (rises) in labor costs move potential (existing) companies across the margin of opening (closing) to generate changes in employment.

6. Hiring, Firing, and Employment Dynamics

The majority of theoretical, empirical, and policy research on labor demand has focused on the level of labor input demanded, not its time path. Understanding the dynamics of employment and hours, and of the flows of hires (h) and separations (s) that are linked to net employment change (ΔE) by the identity $\Delta E \equiv h - s$ is important for understanding cyclical behavior in labor productivity (Y/L) since changes in labor demand may lag behind changes in production. It is also important for evaluating the impacts of policies that alter hiring and firing costs, such as training subsidies, mandated advanced notice of layoffs, and mandated severance pay.

We view employers as maximizing the expected present discounted value of their future profits (or minimizing the expected present discounted value of their future costs) by choosing at each time a plan for the future paths of their inputs, including labor and its components. As in the static case of Sect. 1, the path is lower when product demand and the prices of substitutes for labor are expected to be lower, and when wages are expected to be higher. It is also affected by costs of adjusting labor inputs, including internal disruptions to production that occur when new workers enter or experienced workers leave, and such external costs as advertising for new workers or offering severance pay and unemployment benefits to laid-off or fired workers (Oi 1962).

The size and structure of these adjustment costs alter the path of employment by changing the incentives that employers face when labor demand is shocked by changes in product demand, wages, or other input costs. With larger adjustment costs per worker, a shock to product demand that is viewed as temporary will be met by changing hours rather than altering employment, since employers will seek to avoid incurring hiring or layoff costs. Larger adjustment costs cause changes in employment and perhaps even hours per worker to lag behind those in output, thus generating procyclical changes in labor productivity. The path of employment demand may be smooth if adjustment costs are structured so that it pays employers to spread hiring or firing over several time periods. Hiring and firing may be bunched (lumpy) if adjustment costs are structured so that the marginal cost of an extra hire or fire in a particular period of time is decreasing in the size of the adjustment.

The costs of hiring workers affect both hires and layoffs: employers who know that laying off someone in bad times will create a vacancy that must be filled in good times by incurring hiring costs will be more hesitant to lay off workers. Conversely, imposing restrictions (costs) on layoffs reduces hiring, since employers know that hiring a worker when demand is high will generate additional costs in the future when reduced demand leads to that worker's layoff. One might imagine an extreme employment protection policy in which any layoff is punishable by the employer's death. No layoffs will occur; but employers will be sure to hire only those workers whose jobs might be justified even in slack times. This *reductio ad absurdum* illustrates the general point that higher adjustment costs reduce the overall demand for labor. Like any other increase in labor costs, higher adjustment costs lead employers to substitute away from labor.

Empirical research on dynamic labor demand has lacked the explicit links to the underlying theory and the credible identification of employers' responses that characterize empirical research on own-wage demand elasticities and input substitution. Linking empirical work more closely to theory is high on the research agenda. A number of results are strongly supported by fairly large bodies of literature, however, and are probably reliable. Hours of work adjust more rapidly in response to shocks to labor costs and product demand than does employment. Most important, the more skill–human capital in a group of workers, the slower employers are to hire or fire them when demand conditions change (Rosen 1968). This result has the important implication that employment of less-skilled workers and production workers generally is more highly cyclical than that of skilled and nonproduction workers.

Policies that alter hiring and firing costs have been blamed for causing slow employment growth in Europe in the 1980s and 1990s, and this concern has generated a substantial amount of empirical research trying to gauge their impacts. Much of this research has consisted of broad-brush cross-country comparisons that suffer from problems of identification like those discussed in Sect. 2. The one certain conclusion from this literature is that employment adjustment is faster in North America than in Europe; but whether this is because of differences in the extent of employment protection is unclear. Some researchers have tried to examine how the detailed characteristics of policies that alter adjustment costs have affected the path and level of employment. In some cases, where the policy changes are both unexpected and large (as in the example of Peru studied by Saavedra and Torero 1998), the research has generated evidence that is not confounded by suppliers' behavior (see discussion in Sect. 2). These studies show that higher adjustment costs (severance pay in particular) reduce both the speed of employment adjustment and levels of employment.

Bibliography

Angrist J 1996 Short-run demand for Palestinian labor. *Journal of Labor Economics* **14**: 425–53

Arrow K, Chenery H, Minhas B, Solow R 1961 Capital-labor substitution and economic efficiency. *Review of Economics and Statistics* **43**: 225–50

Berndt E, Christensen L 1974 Testing for the existence of an aggregate index of labor inputs. *American Economic Review* **64**: 391–404
Cobb C, Douglas P 1928 A theory of production. *American Economic Association, Papers and Proceedings* **18**: 139–65
Davis S J, Haltiwanger J 1992 Gross job creation, gross job destruction, and employment reallocation. *Quarterly Journal of Economics* **107**: 819–65
Griliches Z 1969 Capital-skill complementarity. *Review of Economics and Statistics* **51**: 465–8
Hamermesh D S 1993 *Labor Demand*. Princeton University Press, Princeton, NJ
Hamermesh D S, Pfann G A 1996 Adjustment costs in factor demand. *Journal of Economic Literature* **34**: 1264–92
Hamermesh D, Trejo S 2000 The demand for hours of labor: direct evidence from California. *Review of Economics and Statistics* **82**: 38–47
Hart R 1987 *Working Time and Employment*. Allen and Unwin, Boston
Hunt J 1999 Has work-sharing worked in Germany? *Quarterly Journal of Economics* **114**: 117–48
Marshall A 1920 *Principles of Economics*, 8th edn. Macmillan, New York
Oi W 1962 Labor as a quasi-fixed factor of production. *Journal of Political Economy* **70**: 538–555
Rosen S 1968 Short-run employment variation on class-I railroads in the U.S., 1947–63. *Econometrica* **36**: 511–29
Saavedra J, Torero M 1998 *Labor Market Regulation and Employment in Peru*. GRADE, Lima, Peru

D. S. Hamermesh

Dementia: Overview

Dementia is a medical condition which is characterized by a generalized mental deterioration. The word has Latin roots, 'de' meaning separation, cessation or contraction, and 'meme' meaning 'mind.' Therefore, in dementia there is a cessation or contraction of the mind.

The usage of the term dementia can be traced back to the Roman writer and encyclopedist, Aulus Cornelius Celsus in the first century AD. Celsus distinguished dementia, a chronic 'contraction of the mind,' from acute conditions such as delirium. The first medical identification of aging with dementia can be traced to Aretaeus of Cappadocia in the second century AD. He distinguished a condition which began in old age and continued until death, characterized by 'a torpor of the senses, and a stupefaction of the gnostic and intellectual faculties' (Adams 1861). Galen, perhaps the most influential physician of all time, also identified old age as a cause of dementia in the second century AD. He defined 'morosis' as a condition in which 'the knowledge of letters and other acts are totally obliterated, indeed they can't even remember their own names' (Galen 1821–33 translation). One condition associated with morosis was old age.

No further advances in the clinical description of dementia appear to have been made for more than 1500 years. In 1793, the eminent American physician, Benjamin Rush, a signer of the United States Declaration of Independence and author of the first American textbook of psychiatry, a man who is credited by the American Psychiatric Association with being the father of American psychiatry, added to the clinical description of dementia. He noted that:

> it would be sufficiently humbling to human nature if our bodies exhibited in old age the marks only of a second childhood, but human weakness descends even lower. I met with an instance of a woman between 80 and 90 who exhibited the marks of a second infancy, by such a total loss of her mental faculties as to lose all consciousness in discharging her alvine and urinary excretions. In this state of the body, a disposition to sleep succeeds the wakefulness of the first stages of old age' (Rush 1793).

The accuracy of Benjamin Rush's description has become increasingly apparent over the past two centuries.

The modern term, 'senile dementia,' is derived from Esquirol (1838), who described this condition in a French textbook of psychiatry. He noted that this is a condition in which there occurs a weakening of the memory for recent experience and a loss of drive and will power. He noted that senile dementia (in French, *démence senile*), appears gradually and may be accompanied by emotional disturbances.

In modern usage, the term dementia refers to either: (a) any condition associated with a generalized mental deterioration, or (b) a progressive generalized mental deterioration, frequently occurring in later life. Both usages are commonly applied.

1. *Progressive Dementia*

The most important progressive dementia is Alzheimer's disease, which plays a role in a majority of cases (Tomlinson et al. 1970). Other progressive dementias include cerebrovascular dementia, Lewy body dementia, and the frontotemporal dementias. With the exception of the frontotemporal dementias, these conditions increase in occurrence with increasing age. Consequently, the modern increase in lifespan in most regions throughout the world has been accompanied by a dramatic increase in the occurrence of these conditions (Henderson and Jorm 2000).

1.1 *Alzheimer's Disease*

Alzheimer's disease (AD) is one of the most devastating medical conditions of contemporary times.

Table I
Global Deterioration Scale (GDS) for age-associated cognitive decline and Alzheimer's disease[1]

GDS stage	Clinical characteristics	Diagnosis and prognosis[2,3]
1	**No Subjective complaints of memory deficit.** No memory deficit evident on clinical interview.	Normal Adult
2	**Subjective complaints of memory deficit,** most frequently in following areas: (a) Forgetting where one has placed familiar objects; (b) Forgetting names one formerly knew well. No objective evidence of memory deficit on clinical interview. No objective deficit in employment or social situations. Appropriate concern with respect to symptomatology.	Age associated memory impaired (sometimes termed normal aged forgetfulness or aged-associated cognitive decline) 15% develop dementia within 4 years[2]
3	**Earliest clear-cut deficits.** Manifestations in more than one of the following areas: (a) Patient may have gotten lost when traveling to an unfamiliar location. (b) Co-workers become aware of patient's relatively poor performance. (c) Word and/or name finding deficit become evident to intimates. (d) Patient may read a passage or book and retain relatively little material. (e) Patient may demonstrate decreased facility remembering names upon introduction to new people. (f) Patient may have lost or misplaced an object of value. (g) Concentration deficit may be evident on clinical testing. Objective evidence of memory deficit obtained **only with an intensive interview.** Decreased performance in demanding employment and social settings. Denial begins to become manifest in patient. Mild to moderate anxiety frequently accompanies symptoms.	Mild cognitive impairment Two-thirds develop dementia within 4 years[2]
4	**Clear-cut deficit on careful clinical interview.** Deficit manifest in following areas: [a]Decreased knowledge of current and recent events. [b]May exhibit some deficits in memory of one's personal history. [c]Concentration deficit elicited on serial subtractions. [d]Decreased ability to travel, **handle finances,** etc. Frequently no deficit in following areas: (a) Orientation to time and place. (b) Recognition of familiar persons and faces. (c) Ability to travel to familiar locations. Inability to perform complex tasks. Denial is dominant defense mechanism. Flattening of affect and withdrawal from challenging situations.	Mild Alzheimer's disease Mean duration: 2 years
5	**Patient can no longer survive without some assistance.** **Patient is unable during interview to recall a major relevant aspect of their current life,** e.g.: [a]Their address or telephone number of many years. [b]The names of close members of their family (such as grandchildren). [c]The name of the high school or college from which they graduated.	Moderate Alzheimer's disease Mean duration: 1.5 years

Table 1 Continued

GDS stage	Clinical characteristics	Diagnosis and prognosis[2,3]
	Frequently some disorientation to time (date, day of the week, season, etc.) or to place. An educated person may have difficulty counting back from 40 by 4s or from 20 by 2s. Persons at this stage retain knowledge of many major facts regarding themselves and others. They invariably know their own names and generally know their spouse's and children's names. They require no assistance with toileting or eating, but may have difficulty choosing the proper clothing to wear.	
6	May occasionally forget the name of the spouse upon whom they are entirely dependent for survival. Will be **largely unaware of all recent events and experiences in their lives.** Retain some knowledge of their surroundings; the year, the season, etc. May have difficulty counting by 1s from 10, both backward and sometimes forward. **Will require some assistance with activities of daily living:** (a) May become incontinent. (b) Will require travel assistance but occasionally will be able to travel to familiar locations. Diurnal rhythm frequently disturbed. Almost always recall their own name. Frequently continue to be able to distinguish familiar from unfamiliar persons in their environment. Personality and emotional changes occur. These are quite variable and include: [a]delusional behavior, e.g., patients may accuse their spouse of being an imposter; may talk to imaginary figures in the environment, or to their own reflection in the mirror. [b]obsessive symptoms, e.g., person may continually repeat simple cleaning activities. [c]anxiety symptoms, agitation, and even previously non-existent violent behavior may occur. [d]cognitive abulia, e.g., loss of willpower because an individual cannot carry a thought long enough to determine a purposeful course of action.	Moderately severe Alzheimer's disease Mean duration: 2.5 years
7	All **verbal abilities are lost over the course of this stage.** Early in this stage words and phrases are spoken but speech is very circumscribed. Later there is no speech at all—only unintelligible vocalizations. **Incontinent; requires assistance toileting and feeding.** **Basic psychomotor skills** (e.g., ability to walk) **are lost with the progression of this stage.** The brain appears no longer to be able to tell the body what to do. Generalized and cortical neurologic signs and symptoms are frequently present.	Severe Alzheimer's disease Mean time to demise: 2–3 years Potential for survival: 7 or more years

1. Reisberg B, Ferris S H., de Leon M J, & Crook T. 1982 The global deterioration scale for assessment of primary degenerative dementia. *American Journal of Psychiatry* **139**: 1136–1139. Copyright 1983 by Barry Reisberg, M.D. 2. Kluger A, Ferris S H, Golomb J, Mittelman, M S, Reisberg B 1999 Neuropsychological prediction of decline to dementia in nondemented elderly. *Journal of Geriatric Psychiatry and Neurology* **12**: 168–79. 3. Reisberg B, Franssen E, Shah M A, Weigel J, Bobinski M, Wisniewski H M 2000 Clinical diagnosis of dementia: A review. In: Maj M, Sartorius N (eds.) *Dementia,* WPA Series, Evidence and Experience in Psychiatry, Wiley, Chichester, UK, Vol. 3, pp. 69–115.

For example, in the United States, the number of persons in nursing homes and related institutions with AD exceeds the total number of persons in all hospitals and related institutions. Consequently, the institutional burden of AD in the United States is approximately the same as the institutional burden of all other illnesses combined. Worldwide, AD is estimated to afflict more than 15 million persons, and the prevalence continues to increase with increasing life expectancy of the world's people.

The precise mechanism of the cause of AD is unknown. However, particular genotypes which predispose to early onset AD have been identified. These genetic 'defects' are β-amyloid precursor protein mutations located on chromosome 21, presenilin 1 mutations located on chromosome 14, and presenilin 2 mutations which are located on chromosome 1. Some investigators point out that each of these mutations is associated with increased production of β-amyloid protein, one of the neuropathologic hallmarks of AD (Selkoe 1997). However, the presenilin mutations in particular, have been related to a variety of other pathologies apart from increased β-amyloid protein. Further evidence for the role of β-amyloid in the etiopathogenesis of AD is the observation that persons with Down's syndrome develop Alzheimer's pathology early in life and carry an extra copy of the β-amyloid precursor protein gene resulting in excess β-amyloid production.

The occurrence of the common late-life form of AD has also been associated with particular genotypes. The most clearly identified gene which influences the occurrence of late life, also known as 'sporadic' Alzheimer's disease, is the apolipoprotein E (apoE) gene (National Institute on Aging /Alzheimer's Association Working Group 1996). Caucasians who carry the apoE4 allele are at increased risk for AD in later life. Caucasians who are homogenous for apoE4 have as much as eight times the risk of AD as similarly aged persons who do not carry the apoE4 allele. Other genes and gene interactions which influence the occurrence of the common, late-life form of AD have not yet been fully elucidated.

Although the mechanism by which AD occurs is not known, a characteristic clinical course of AD has been described (Reisberg et al. 1982, 2000). This characteristic clinical course is accompanied by characteristic neuropathologic manifestations and a characteristic neuropathologic progression.

Clinically, AD can be described on a continuum with the changes in what is now termed normal aging, progressing to mild cognitive impairment (MCI) and subsequently to AD. This clinical continuum is illustrated in terms of the cognitive, functional and emotional progression of aging and AD in Table 1. Functionally, the progression of brain aging, MCI and AD can be described in even greater detail, in terms which are universally clear to both lay persons and professionals. This characteristic functional progression is familiar to many observers because it reverses the sequence of functional acquisition in human development, a phenomenon which has been termed 'retrogenesis' (Table 2) (Reisberg et al. 1999, 2000)

The hallmark neuropathologic manifestations of AD are extracellular senile plaques which contain the β-amyloid protein, and intracellular neurofibrillary tangles which contain the protein tau (see *Alzheimer's Disease, Neural Basis of*). Very strong relationships have been found between the clinical progression of AD and the progression of pathology in the hippocampus, a brain region which has been notably associated with memory. Progressive pathologic manifestations in the hippocampus include neurofibrillary pathology, volume loss, and neuronal cellular loss. Although the hippocampus is a brain region which manifests early and progressive AD pathology, AD is a generalized brain disease with, for example, progressive decrements in brain glucose utilization (energy metabolism) and progressive slowing of brain wave electrical activity on the electroencephalogram, as well as neuropathologic manifestations in numerous brain regions, which become particularly evident as AD progresses.

Neuropathologic stages of AD have been described (Braak and Braak 1991). These are transentorhinal stages I and II, limbic stages III and IV, and neocortical stages V and VI. Each of these stages of Braak and Braak are based upon the evolution and spread of neurofibrillary pathology in the brain of the AD patient.

Braak and Braak (1996) and Reisberg et al. (1992, 1999) have described a common basis for the neuropathologic and clinical manifestations of AD. This is the progressive involvement of the most recently, and therefore the most thinly, myelinated brain regions, followed by progressively more thickly myelinated, and therefore, better protected, brain regions. The most thickly myelinated brain regions subsume the functions and associated cognitive skills, which are the first to be acquired in the course of human development. Conversely, the most thinly myelinated brain regions subsume the most recently acquired functions and associated cognition. These recently myelinated regions are the most vulnerable to AD pathology. Consequently, cognitive and functional skills which are the most recently acquired are the most vulnerable to AD pathology, and those cognitive and functional skills which are the first to be acquired in normal human development are the last to be lost with the progression of AD. In brief, last in, first out. This retrogenesis phenomenon, which is illustrated for functional skills in Table 2, also applies to cognition and many aspects of brain physiology, including neurologic reflexes and apparently, to at least some extent, brain electrophysiology and metabolism.

Very recently it has been demonstrated that the β-amyloid, the major putative toxic pathology in AD is destructive to the oligodendroglia, the brain cells

Table 2
Retrogenesis: functional landmarks in normal human development and Alzheimer's disease (AD)

Normal Development Approximate Total Duration: 20 years			Alzheimer's Degeneration Approximate Total Duration: 20 years			
Approximate age	Approximate duration in development	Acquired abilities	Lost abilities	Alzheimer stage	Approximate duration in AD	Developmental age of AD
Adolescence						
13–19 years	7 years	Hold a job	Hold a job	3 – Incipient	7 years	19–13 years: Adolescence
Late childhood						
8–12 years	5 years	Handle simple finances	Handle simple finances	4 – Mild	2 years	12–8 years: Late childhood
Middle childhood						
5–7 years	2.5 years	Select proper clothing	Select proper clothing	5 – Moderate	1.5 years	7–5 years: Middle childhood
Early childhood						
5 years	4 years	Put on clothes unaided	Put on clothes unaided	6a – Moderately severe	2 years	5–2 years: Early childhood
4 years		Shower unaided	Shower unaided	b		
4 years		Toilet unaided	Toilet unaided	c		
3–4.5 years		Control urine	Control urine	d		
2–3 years		Control bowels	Control bowels	e		
Infancy						
15 months	1.5 years	Speak 5–6 words	Speak 5–6 words	7a – Severe	7 years or longer	15 months to birth: Infancy
1 year		Speak 1 word	Speak 1 word	b		
1 year		Walk	Walk	c		
6–10 months		Sit up	Sit up	d		
2–4 months		Smile	Smile	e		
1–3 months		Hold up head	Hold up head	f		

Copyright © 1984, 1986, 1996, 2000 by Barry Reisberg, M.D. All rights reserved.

which produce the myelin (Xu et al. 2001). ApoE, the AD risk-determining factor, is a molecule associated with cholesterol transport. ApoE has been shown to affect neuronal repair mechanisms and the apoE3 allele appears to have a greater neuroprotective effect than the apoE4 allele (Poirier et al. 1993; Buttini et al. 1999). Therefore, a common basis for the molecular, pathologic, and clinical manifestations of AD can be postulated although many elements of the precise pathogenic mechanism remain unknown.

1.2 Vascular Dementia

Formerly, this entity was designated multi-infarct dementia because of its strong relationship to strokes, both large and small (Hachinski et al. 1974). Although originally considered by Hachinski to be a stroke-like entity with an abrupt onset and stepwise pattern of deterioration, this deterioration pattern is rarely observed. Vascular dementia is now thought to be a condition marked by progressive decline in cognition and functioning in which cerebrovascular, i.e., stroke risk factors, play a major role. These risk factors include hypertension and cardiovascular disease, as well as more overt evidence of cerebrovascular disease such as transientischemic attacks and focal neurologic signs and symptoms. In most cases, vascular dementia occurs in association with Alzheimer's brain pathology, a condition which is sometimes termed 'mixed dementia.' Much less frequently, vascular dementia appears to occur without evidence of concurrent AD. Consequently, many investigators view vascular dementia as on a continuum with AD. This continuum concept has become particularly compelling in recent years, since virtually all of the elements associated with vascular dementia have been shown to be risk factors for so-called 'pure AD.'

In accordance with the concept of an additive morbidity of vascular dementia and AD, the course of vascular dementia has been found to be generally more rapid than AD. Vascular dementia is generally considered to be the most common dementia entity after AD.

1.3 Lewy Body Dementia

Lewy bodies are spherical inclusions located in the cytoplasm of neurons which characteristically contain a protein known as α-synuclein. Originally Lewy bodies were considered to be a prominent feature of Parkinson's disease and were not associated with late life dementia. However, new staining techniques in the 1990s indicated that Lewy bodies in the brain stem and the cortex commonly occur in dementia patients studied postmortem. Estimates indicate that as many as 15 to 25 percent of dementia patients manifest Lewy bodies at the time of demise (McKeith et al. 1996). However, in the great majority of these cases, Lewy bodies exist together with neuropathologic manifestations of AD and/or vascular dementia. When this mixed neuropathology occurs, the clinical manifestations of the dementia are those of AD and/or vascular dementia, and there is no distinctive Lewy body dementia clinical syndrome.

In approximately 4 percent of all dementia cases coming to autopsy, Lewy bodies occur in the brain in the absence of other dementia disorders. In these cases, a classical Lewy body dementia (LBD) clinical syndrome occurs. This syndrome is marked by three core features, any/or all of which may be present: (a) a relatively fluctuating clinical course, with variability in cognition; (b) the presence of vivid, well formed, visual hallucinations; and (c) the presence of Parkinsonian features. Another important aspect of LBD is sensitivity to neuroleptic (antipsychotic) medications, particularly those known to be associated with Parkinsonian side effects (Ballard et al. 1998). In general, the onset of LBD is more acute than the onset of AD, and the course of LBD appears to be shorter than that of AD.

1.4 Frontotemporal Dementias

These are a diverse group of disorders for which various classification schemata have been developed. For example, biomolecular categorizations have been developed recently which are extremely useful in categorizing one subgroup of frontotemporal dementias, those with frontotemporal dementia with Parkinsonism linked to chromosome 17 (FTDP-17) (Foster et al. 1997). The FTDP-17 dementias are characterized by mutations which alter tau, the protein responsible for neurofibrillary changes such as those found in AD. Another frontotemporal dementia type marked by abnormal tau accumulation is Pick's disease. In Pick's disease there is an intracellular inclusion known as the Pick body which contains tau accumulations as well as other substances. Other forms of frontotemporal dementia are classified mainly on the basis of their clinical presentation. They include semantic dementia (see *Dementia, Semantic*), progressive aphasia, and motor neuron disease dementia.

The core clinical features of the frontotemporal dementias have been described in part as disinhibition, decreased insight, apathy, disorganization, lack of personal hygiene, mental rigidity, and hyperorality (Brun et al. 1994, Kertesz 2000). In addition, aphasia and other language disturbances are frequently present. Like the basic pathologic disturbances, these behavioral manifestations are very heterogenous and occur on a clinical spectrum.

In general, the frontotemporal dementias occur in adults at a younger age than AD. Some believe that

these conditions are very common in patients who are institutionalized for mental illness.

1.5 Other Progressive Dementias

Various other diverse conditions produce progressive dementia. In general, these other conditions are believed to be considerably less common than the dementias which have been reviewed more specifically in the preceding sections.

Perhaps the most notable of these other dementias are the prion dementias, of which the most frequent is Creutzfeldt-Jakob disease (CJD). Prion dementias are believed to result from a self-propagating conformational change in a protein, known as the prion protein. This prion, a proteinaceous infective agent, can be acquired either spontaneously or through transmission of biologic materials containing the prion, including brain, pituitary glandular substance, and corneal lens transplants. Spontaneous CJD occurs in approximately one person per million population worldwide. Recently, a form of transmissible CJD arose in Britain as a result of transmission of a prion disease in cattle to humans. This 'epidemic' which has affected dozens of persons in Britain and, more recently, several persons on the European continent, has resulted in the destruction of the entire British cattle herd and a change in dietary habits of tens of millions of persons in Europe and elsewhere. Consequently, although the prion dementias remain rare conditions in humans, the threat of the spread of these proteinaceous infections, has caused somewhat justified alarm, in particular in Europe, but also throughout the world.

Other relatively uncommon progressive dementias include progressive supranuclear palsy, and corticobasilar degeneration. These conditions, like some of the frontotemporal dementias, are marked by abnormal tau protein and resultant neurofibrillary accumulations.

There are also many progressive dementias associated with broader pathologic spectra, of which dementia is only one element. Among these are the dementia of Down's syndrome, the dementia of normal pressure hydrocephalus (NPH), and the dementia of Huntington's disease. Some of these dementias are very closely related to AD dementia. For example, persons with Down's syndrome universally develop an Alzheimer's type neuropathologic and clinical picture at a relatively early age in comparison with persons with classical AD. Similarly, brain biopsy studies have recently demonstrated that many elderly persons with NPH who manifest a dementia component to the clinical syndrome, show Alzheimer's-type pathology, specifically neuritic plaques, upon brain tissue examination (Golomb et al. 2000). As is the case for AD, the percentage of NPH dementia patients with biopsies demonstrating neuritic plaques increases with dementia severity, rising from about 20 percent of patients with MCI to about 75 percent of patients in stage 6 and 7 of the global deterioration scale (see Table 1). These studies indicate that the gait disturbance which is characteristic of NPH appears to be associated with dilated cerebral ventricles (i.e., expansion of the fluid filled cavities in the brain), whereas the dementia component of NPH is probably related to concomitant AD.

2. Dementias with Variable Prognoses

Numerous diverse physiologic disturbances can produce dementias which may be acute or chronic, reversible or permanent. These diverse conditions include infectious conditions, toxins, tumors, brain trauma, endocrine and metabolic disturbances, nutritional disturbances, and medications. For example, infectious conditions which produce encephalitis may cause dementia of viral or bacterial origin. A relatively common example is herpetic encephalitis. Syphilis and acquired immune deficiency syndrome (AIDS), have been associated with dementia. Heavy metal toxicity, such as aluminum toxicity from dialysis, can cause dementia. A space-occupying brain lesion may produce dementia, such as brain metastases resulting from neoplasms. Brain trauma resulting in diffuse brain injury is a source of dementia. Thyroid disturbances, hyponatremia, and other endocrine disorders are known to cause dementia. Vitamin B_{12} (cobalamin) deficiency, which is associated with demyelination, is a cause of dementia, as is folate deficiency. Numerous medications can produce dementia or exacerbate a preexisting dementia. Some of the conditions which can produce dementia independently are now believed to be risk factors for AD as well. Examples of the latter include major depression and depression dysphoric disorder, and Vitamin B_{12} deficiency.

See also: Alzheimer's Disease: Behavioral and Social Aspects; Alzheimer's Disease, Neural Basis of; Dementia, Semantic

Bibliography

Adams F 1861 *The Extant Works of Aretaeus, the Capadocian*. Syndenham Society, London

Ballard C, Grace J, McKeith I, Holmes C 1998 Neuroleptic sensitivity in dementia with Lewy bodies and Alzheimer's disease. *Lancet* **351**: 1032–3

Braak H, Braak E 1991 Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica* **82**: 239–59

Braak H, Braak E 1996 Development of Alzheimer-related neurofibrillary changes in the neocortex inversely recapitulates cortical myelogenesis. *Acta Neuropathologica* **92**: 197–201

Brun A, Englund E, Gustafson L, Passant U, Mann D M A, Neary D, Snowden J S 1994 Clinical and neuropathological criteria for frontotemporal dementia. *Journal Neuro Neurosurg Psychiatry* **57**: 416–8

Buttini M, Orth M, Bellosta S, Akeefe H, Pitas R E, Wyss-Coray T, Mucke L, Mahley R W 1999 Expression of human apolipoprotein E3 or E4 in the brains of Apoe mice $^{-/-}$: isoform-specific effects on neurodegeneration. *Journal of Neuroscience* **19**: 4867–80

Esquirol J E D 1838 *Des maladies mentales*. Balliere, Paris

Foster N L, Wilhelmsen K C, Sima A A F, Jones M Z, D'Amato C, Gilman S 1997 Frontotemporal dementia and Parkinsonism linked to chromosome 17: consensus conference. *Ann Neurol* **41**: 706–15

Galen 1833 De symptomatum differentis liber (Translation). In: Kuhn C G (ed.) *Opera Omnia*. Knobloch, Leipzig, Germany, Chap. VII, pp. 200–1

Golomb J, Wisoff J, Miller D C, Boksay I, Kluger A, Weiner H, Salton J, Graves W 2000 Alzheimer's disease comorbidity in normal pressure hydrocephalus: prevalence and shunt response. *J Neurol Neurosurg Psychiatry* **68**: 778–81

Hachinski V C, Lassen N A, Marshall J 1974 Multi-infarct dementia: A cause of mental deterioration in the elderly. *Lancet* **2**: 207–10

Henderson A S, Jorm A F 2000 Definition and epidemiology of dementia: A review. In: Maj M, Sartorius N (eds.) *Dementia*, WPA Series, Evidence and Experience in Psychiatry. Wiley, Chichester, UK, Vol. 3, pp. 1–33

Kertesz A 2000 Behavioral and psychological symptoms and frontotemporal dementia (Pick's disease). *International Psychogeriatrics* **12**(Suppl. 1): 183–7

McKeith I G, Galasko D, Kosaka K, Perry E K, Dickson D W, Hansen L A, Salmon D P, Lowe J, Mirra S S, Byrne E J, Lennox G, Quinn N P, Edwardson J A, Ince P G, Bergeron C, Burns A, Miller B L, Lovestone S, Collerton D, Jansen E N H, Ballard C, de Vos R A I, Wilcock G K, Jellinger K A, Perry R H 1996 Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB). *Neurology* **47**: 1113–24

National Institute on Aging/Alzheimer's Association Working Group 1996 Apolipoprotein E genotyping in Alzheimer's disease. *Lancet* **347**: 1091–5

Poirier J, Baccichet A, Dea D, Gauthier S 1993 Cholesterol synthesis and lipoprotein reuptake during synaptic remodelling in hippocampus in adult rats. *Neuroscience* **55**: 81–90

Reisberg B, Ferris S H, de Leon M J, Crook T 1982. The global deterioration scale for assessment of primary degenerative dementia. *American Journal of Psychiatry* **139**: 1136–9

Reisberg B, Pattschull-Furlan A, Franssen E, Sclan S G, Kluger A, Dingcong L, Ferris S H 1992 Dementia of the Alzheimer type recapitulates ontogeny inversely on specific ordinal and temporal parameters. In: Kostovic I, Knezevic S, Wisniewski H, Spilich G (eds.) *Neurodevelopment, Aging, and Cognition*. Birkhauser, Boston, MA, pp. 345–69

Reisberg B, Franssen E H, Hasan S M, Monteiro I, Boksay I, Souren L E M, Kenowsky S, Auer S R, Elahi S, Kluger A 1999 Retrogenesis: Clinical, physiologic and pathologic mechanisms in brain aging, Alzheimer's and other dementing processes. *European Archives of Psychiatry and Clinical Neuroscience* **249**(Suppl. 3): 28–36

Reisberg B, Franssen E, Shah M A, Weigel J, Bobinski M, Wisniewski H M 2000 Clinical diagnosis of dementia: A review. In: Maj M, Sartorius N (eds.) *Dementia*, WPA Series, Evidence and Experience in Psychiatry. Wiley, Chichester, UK, Vol. 3, pp. 69–115

Rush B 1973 An account of the state of mind and body in old age. In: Rush B (ed.) *Medical Inquiries and Observations*. Dobson, Philadelphia, PA, Vol. 2, pp. 311

Selkoe D J 1997 Alzheimer's disease: Genotypes, phenotype and treatments. *Science* **275**: 630–1

Tomlinson B E, Blessed G, Roth M 1970 Observations on the brains of demented old people. *Journal of Neurological Science* **11**: 205–42

Xu J, Chen S, Ahmed S H, Chen H, Ku G, Goldberg M P, Hsu C Y 2001 Amyloid-β peptides are cytotoxic to oligodendrocytes. *J Neurosci* **21**, RC118: 1–5

B. Reisberg

Dementia: Psychiatric Aspects

Dementia has many aspects, and neurobiological, clinical, and family related or socioeconomic factors have been investigated. Here neuropsychiatric aspects are presented i.e., the signs and symptoms of dementia diseases which cause the disturbance in behavior, cognitive functions, and subjective experience of the patients, which are classified as emotional problems or personality change.

1. *Definition of Dementia*

In history, the term dementia has been subject to several changes in meaning. Today, it is defined as acquired, severe loss of mental functions due to structural brain diseases.

In dementia, the information processing capacity has deteriorated to a critical degree. Deficits are found predominantly in cognitive, but also in emotional and intentional domains. By definition, these deficits severely compromise the social and cultural activities of a person. There must be a loss of mental functions which had been acquired before (i.e., the long-standing cognitive deficiencies in mental retardation are excluded). The underlying pathology is a destruction or degeneration of brain structures, therefore dementia may not be caused by functional organic disorders such as psychoses or transient brain dysfunction like in delirium. Psychiatry is concerned primarily with the detection and diagnosis of dementia syndromes, and concerned secondly with the identification and treatment of certain dementia diseases. Dementia syndromes and dementia diseases have to be differentiated (Wells and Whitehouse 1996).

1.1 *Dementia Syndromes*

Characteristic clinical profiles emerge, patterns of cognitive, affective, and intentional disturbances because in dementia diseases, the destruction or degen-

eration takes place with a predominance of certain brain regions or brain systems; the disease does not impair all brain structures to the same extent.

1.1.1 Dementia syndrome of Alzheimer type. What is the clinical presentation, the pattern of deficits which characterize the syndrome of a dementia patient who suffers from Alzheimer's disease? This issue has been clarified in large longitudinal studies (e.g., Berg et al. 1990). A second question addresses the reasons for failures in activities of daily living and the need for care. The syndrome can be differentiated into cognitive signs and not primarily cognitive signs.

(a) *Cognitive neuropsychological signs*

(i) *Amnestic signs*. Early in the time course of the dementia syndrome, deficits in episodic memory are noticed (Brandt and Rich 1995). Recall of word lists, stories, second word of paired associates, or the reconstruction of figures-drawings, which were encoded several minutes earlier, becomes increasingly impaired. The immediate recall of digits (digit span), however, remains typically unimpaired. The patients can immediately repeat a list of words but forget the words within several minutes, e.g., after a disruption by intermittent attention to other topics.

Furthermore, errors of orientation for time and space are hallmarks of this dementia syndrome. Demented persons get lost easily—at first in unfamiliar, and then even in familiar surroundings. The forgetting of data of major impact or the forgetting of actions with dangerous consequences affects activities of daily living. These deficits are characteristics of anterograde amnesia and are a consequence of mediotemporal brain damage, including the hippocampus and entorhinal cortex. Degeneration within the basal forebrain, however, may also explain a large part of the symptoms. Later retrograde amnesia starts and the recollection of data which were well known in former times becomes impaired. In later stages of the disease, demented persons cannot remember anything but episodes from the first decades of their life.

To summarize, a dementia syndrome in persons with Alzheimer's disease presents initially with amnestic signs. Episodic memory and orientation, with respect to time and place, deteriorate, and a retrograde amnesia occurs.

(ii) *Language disturbance, aphasic signs*. Word finding is impaired early in the time course of the dementia syndrome of Alzheimer's type, but these and other signs of language impairment usually occurs later than memory impairment. Difficulties in word finding are the most important feature of this language disturbance. Word finding is clinically observed mostly in the communication situation and in fluency tests.

As demented persons become increasingly incapable of finding correct words, in later stages of the disease, not much content is communicated: a phenomenon which has been called empty speech. Phonematic paraphasias (errors in the phonematic composition of words) or disordered syntax, which characterize various forms of aphasia after cerebrovascular strokes, are, however, not considered prominent signs in this dementia syndrome.

A lowered performance in fluency tasks is often assessed in clinical diagnosis, for example, by asking the patient to name as many animals as possible. It has been argued that fluency is not only a language test, but also a test of retrieval from semantic memory, and it is known that effective retrieval depends on nonlanguage functions. Patients with frontal lobe lesions cannot retrieve words in fluency tasks, presumably because of inefficient retrieval strategies and a defective suppression of overlearned alternative word-finding routines. Furthermore, demented persons cannot retrieve semantic features of perceived objects because of a disturbance in semantic memory (Martin and Fedio 1983).

Later in the time course, language comprehension difficulties additionally impair the communication with a demented person. In summary, there are important language problems in the dementia syndrome of Alzheimer's type which are predominantly caused by deficits in word finding and language comprehension.

(iii) *Apraxia signs*. Persons with the dementia syndrome of Alzheimer's type show errors in performing purposive movements. These errors occur although there are no obvious motor problems, such as weakness of the limb or sensoric deficits in guiding movements. It is thought that the central control of movements is defective. Apraxia in the narrow sense describes errors in specific action sequences, for example, using tools. Clinically, errors in dressing properly are encountered. Tests of apraxia include the call for imitation of limb movements or gestures.

In diagnostic schedules for dementia, constructional apraxia is tested quite often. This apraxia comprises deficits not only of the execution of movement schemata, but also of the integration of visuospatial information into appropriate action (Heilman and Rothi 1997). Copying visual designs or drawings of well-known objects, such as a clock, are defective and are performed laboriously and slowly. In addition, many demented subjects show an apractic type of writing disorder, where the action sequences of writing the individual letters can no longer be carried out correctly.

To summarize, subjects with Alzheimer's dementia show errors in performing motor tasks without weakness or paresis of the limbs, resembling central motor control deficits. Later in the time course of the disease, the impairments lead to incapability to perform any practical task in the household or in self-care.

(iv) *Disorders of judgement*. Judgement errors and inadequate evaluation are especially relevant for the

every day functioning of demented persons and with respect to legal responsibility. In the narrow definition, erroneous or inadequate judgements, with full understanding and appreciation of the problem, can be related to the decision process in itself. This specificity is difficult to ascertain in clinical neuropsychological investigation. Therefore, more generally, inadequate judgements about the own person and the disease are considered for diagnostic evaluation.

Self-related judgements: demented persons often cannot correctly evaluate the cognitive deficits they suffer from. Some patients describe grossly inadequate future plans. At times, not only disorder of judgement, but also the psychopathological phenomenon of denial has to be considered in such situations.

Formal assessment of logical decisions, as part of the capacity for adequate judgment, is done usually by intelligence tests which, demand, for example, the understanding of a logical rule within the sequence of symbols, and finding incorrect items.

In summary, inadequate or wrong judgements in self-related or logical problems are found commonly in the dementia syndrome of Alzheimer Type and may lead to legal consequences or intensify the demand of care for the patient.

(v) *Other cognitive neuropsychological signs*. Demented persons with Alzheimer's diseases often show deficits in executive functions. The term 'executive functions' denotes the planning and the execution of goal directed activities, including preparations and the monitoring of already reached partial goals. Some researchers also included self-regulation and volitional processes in the domain of executive functions. Brain structures for executive functions are at least in part localized in the dorsolateral prefrontal lobe.

Mental calculations are often tested in the clinical diagnosis of a dementia syndrome. They do not, however, test calculation function only. For these tasks, the attentional demands, and additionally, in serial calculations, the demands on working memory are high. Therefore, the reason for individual test deficits is difficult to attribute precisely.

Agnosia is mentioned in some of the dementia criteria. The classical form of object agnosia, however, is very rare. In object agnosia, the patient perceives an object—and is even able to copy the form—but is not able to grasp its meaning, to name it, or describe it's attributes or functions. Demented persons often fail to identify faces of well-known friends or relatives in later stages of the disease, but memory problems contribute to this deficit. More often, however, a perceptual deficit concerning visuospatial features of figural information can be documented by neuropsychological tests. Furthermore, the appreciation of the relation of many objects within a complex pictorial scene is often disturbed in dementia.

In addition to cognitive deficits, there are other dimensions of the dementia syndrome of Alzheimer's type.

(b) *Noncognitive part of dementia syndrome of Alzheimer's type*

Patients with a dementia syndrome of Alzheimer type suffer from affective symptoms or delusions. Noncognitive signs of dementia are traditionally described by psychopathology. They are not specific for dementia. Several of these signs are commonly observed in other psychiatric diseases, for example, anxiety or restlessness. Although these signs are not as relevant for diagnosis as cognitive deficits, they present major problems for the caregivers (Cummings et al. 1996) and are often the reason for institutionalization of demented patients.

(i) *Signs in the affective domain*. Early in the development of the dementia syndrome depression is observed. Later, affect control becomes impaired. Patients may cry after minor adverse events or be aggressive or inadequately cheerful. Often the emotional state lasts often only for a short period of time. The perceived emotional experience may not be as dramatic as the affective expression suggests. This phenomenon has been termed affective lability.

Affective states like aggression may also be expressed as an increased intensity. Such behaviors may often present as behavior which cannot be tolerated and may lead to institutionalization.

(ii) *Signs in the intentional and attentional domain*. The cerebral information processing for motivation and attention is impaired in late stages of Alzheimer's dementia. A state of 'apathy' is described quite often. A person, for example, sits in a chair for most of the time and does not move, and has no intention to start any action. Many of the noncognitive signs of the dementia syndrome are traditionally described as organic personality change. A further component of this is, as a consequence of frontal lobe disturbance, a disinhibition of action tendencies, which are usually supressed because of cultural constraints. In addition, paranoid delusional thinking and hallucinations occur quite often within the time course of a dementia syndrome of Alzheimer's type. These also may induce aggressive behavior.

In summary, signs in multiple domains of brain functions are observed. The signs which characterize the dementia syndrome of Alzheimer's type are predominantely from the cognitive domain. However, disorders in the emotional and intentional domains have some impact for the resulting individual syndrome and for caregivers and relatives. The primary sensory and motor functions, however, are relatively spared. Accordingly, the spread of the Alzheimer's pathology in the brain does not include the primary sensory and motor areas.

1.1.2 Frontotemporal dementia syndrome. The dementia syndrome of frontotemporal type is less common than the dementia syndrome of Alzheimer's

type and should be described here in outline only (Brun et al. 1994). The major feature of this syndrome is a change in personality. Features of the dementia syndrome of Alzheimer's type, like predominantly amnestic and apractic signs, or a decrease in intellectual functions, are not detectable at early stages of the frontotemporal dementia syndrome.

The personality change comprises changes in the way of experiencing certain situations, in the way of reacting in certain situations, or in the way a person was used to acts spontaneously.

(a) *Lack of expected reactions or actions*. The patient is aspontaneous and does not react as they did before the onset of the dementia syndrome. At times, a lack of drive is noted. As there is no emotional symptom of depressed mood and despair, this has been named pseudodepression.

(b) *Disinhibition*. A lack of control of intentions, especially with respect to social norms or cultural restraints, is observable.

(c) *Perseveration and stereotyped behavior*. The patient reiterates actions and is inflexible in the way of reacting.

(d) *Lack of insight and disorder of judgement*, as described above (Sect. 1.1.1(a))

In addition to the psychopathological symptoms described above, formalized neuropsychological testing reveals deficits in planning, in shifting mental set, and in working memory. The frontotemporal dementia syndrome can be identified clinically only in early stages. In late stages, mostly an unspecific terminal dementia syndrome is clinically observed.

As for the dementia syndrome of Alzheimer's type, also in frontotemporal dementia there is a heterogenous etiology with mostly genetically determined underlying diseases.

1.2 Dementia Diseases

Dementia has two sides—the individual pattern of neuropsychological or psychopathological impairments on the one side, and the damage or degeneration within the brain on the other side.

A dementia disease is a class of pathophysiological processes which result in structural brain changes that are underlying the clinical signs of the dementia syndromes (Wells and Whitehouse 1996, McHugh and Slavney 1998). There are many diseases of the brain which lead to a dementia syndrome.

1.2.1 Primary dementia diseases. Alzheimer's dementia disease, Pick dementia disease, or Lewy body dementia are degenerative brain diseases which up to now inevitably lead to a progressive dementia syndrome. They cause no other symptoms except symptoms of the dementia syndromes.

1.2.2 Neurological diseases with associated dementia syndrome.

(a) In some diseases the dementia outcome is obligatory. In Huntington's chorea, for example, a movement disorder precedes the progressive dementia syndrome, which regularly develops later in the time course of the disease.

(b) In other diseases, the dementia outcome is facultative. These diseases are not dementia diseases *per se*. Vascular dementia, which is the second most frequent cause of dementia cases, is in some of its variants not associated with a dementia syndrome. Prevention of subsequent strokes, furthermore, may stop the cerebrovascular disease and prevent the development of vascular dementia.

The brain behavior relationship is interindividually variable and even the distribution of pathological changes within one disease is varying. Because in many cases the dementia disease cannot be identified reliably before neuropathological post mortem examination, the clinical dementia syndrome guides the medical decisions about treatment and secondary prevention.

1.2.3 Dementia disease etiology. Alzheimer's disease is genetically heterogenous. Many different mutations on several genes are known to cause a presenile Alzheimer's dementia. The same is true for frontotemporal dementia. Additionally, multiple causal factors are identified in senile dementia of Alzheimer's type, and it is not known whether an interaction of several factors is responsible for senile dementia cases. In some dementia cases, etiologically distinct causes are known, like in dementia following traumatic brain lesion, or in a dementia patient with a series of strokes within strategic regions of the brain.

2. Diagnosis of Alzheimer's Dementia

The diagnosis of a dementia disease will be described for the most frequent dementia, Alzheimer's.

(a) The confirmation of the disease can only be given by a neuropathological investigation of the brain after death. The microscopic analysis will reveal the amount and distribution of neuropathological features like neuritic plaques and neurofibrillary tangles.

(b) There is some uncertainty as to the clinical diagnosis of Alzheimer's dementia. There are several lists of criteria proposed, e.g., by diagnostic systems, such as International Classification of Diseases (ICD) and by consensus committees (e.g., the classification of probable or possible Alzheimer's dementia according to clinical criteria). However, the criterium lists differ considerably, and the criterium variance results in major differences in the estimated prevalence rates of population-based studies (Erkinjunnti et al. 1997).

Prospective studies of dementia syndromes improved the specificity of clinical diagnostic methods. However, at present technicians cannot distinguish between dementia diseases such as Lewy body dementia and Alzheimer's dementia with sufficient reliability.

2.1 Problems of Diagnosis of Mild Dementia in Old Age

The factors mentioned add to the diagnostic uncertainty with respect to differential diagnosis. Most dementia cases are seen in old age. The differentiation from normal aging and organic psychiatric disorders, like mild cognitive impairment or amnestic syndrome, is for many cases not easy or, especially in early stages, not possible at all.

The early detection of dementia is difficult in old age, because age in itself is associated with some cognitive decline (Reischies 1998). Furthermore, mild cognitive deficits have to be differentiated from low talent domains of a person, and personality changes have to be diagnosed considering the premorbid personality.

2.2 Dementia Case

If the severity of the cognitive impairments of the dementia syndrome of Alzheimer's type is sufficiently high, and the severity of impairment in everyday functioning is high as well, a dementia case can be diagnosed. For these decisions, cut-off values of assessment instruments have been applied (e.g., the Mini Mental State Examination and the Global Assessment of Functioning scale).

The diagnosis of a dementia case without specification of the kind of probable dementia disease is often a clinical reality. With the advent of better diagnostic methods, however, the diagnosis of the dementia disease may be made before a clinical dementia syndrome has developed.

3. Time Course of Alzheimer's Dementia

There is a characteristic time course of Alzheimer's dementia over about 10 years. Interindividual variation of the time course with plateaus and progression phases are common.

Within a preclinical stage of Alzheimer's disease, cognitive decline takes place which can be assessed by serial neuropsychological testing. Especially memory decline can be found. With memory deficits as the only clinical sign, no formal diagnosis of dementia can be made. The early phase of Alzheimer dementia thus represents an amnestic stage.

In the second stage, a clinical dementia syndrome of Alzheimer's type is detectable, which exhibits more or less the complete features of the dementia syndrome, as described above. In this stage, there are disturbances in activities of daily living, but they may still be minor.

In the stage of a severe dementia syndrome, memory loss, aphasia, apraxia, and the noncognitive symptoms have increased, and global functioning is deteriorated such that normal activities of daily living are critically compromised or no longer possible. Additionally, the noncognitive disorder disturbs the relation between the patient and the environment. Institutionalization is common.

The last stage is a terminal or vegetative stage, where no communication is possible, the patient is mostly bedridden, and has to be cared for completely.

4. Treatment, Professional Care, and Prevention

An increasing number of new drug treatments for Alzheimer's dementia are being investigated. However, at present only limited relief can be provided. At best, the time course can be stopped or retarded for a short time only. Training of activities of daily living, with the goal of adaption to the specific environment, can be quite successful in order to aid care of subjects with an advanced dementia syndrome.

Major efforts must be directed to prevention, detection of risk factors, and early detection of dementia syndromes. A certain success of prophylaxis by controlling risk factors for artherosclerosis has been documented in the field of vascular dementia.

See also: Aging Mind: Facets and Levels of Analysis; Alzheimer's Disease: Behavioral and Social Aspects; Alzheimer's Disease, Neural Basis of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Cognitive Aging; Dementia: Overview; Lifespan Theories of Cognitive Development; Memory and Aging, Cognitive Psychology of; Memory and Aging, Neural Basis of

Bibliography

Berg L, Miller J P, Morris J C, Rubin E H, Smith D S, Storandt M, Coben L A 1990 Mild senile dementia of the Alzheimer type: 3. Longitudinal and cross-sectional assessment. *Annals of Neurology* **28**: 648–52

Brandt J, Rich J B 1995 Memory disorders in the dementias. In: Baddeley A D, Wilson B A, Watts F N (eds.) *Handbook of Memory Disorders*. J. Wiley, Chicester, UK, pp. 244–70

Brun A, Neary D, Englund B, Gustafson L, Mann D M A, Passant U, Snowdon J S 1994 Clinical and neuropathological criteria for frontotemporal dementia. *Journal of Neurol Neurosurgery and Psychiatry* **57**: 416–18

Cummings J L, Fiorello T, Gornbein J, Mega M S 1996 The spectrum of behavioral changes in Alzheimer's disease. *Neurology* **46**: 130–35
Erkinjuntti T, Østbye T, Hachinski V, Steenhuis R 1997 The effect of different diagnostic criteria on the prevalence of dementia. *The New England Journal of Medicine* **337**: 1667–74
Heilman K M, Rothi L J G 1997 *Apraxia—The Neuropsychology of Action*. Psychology Press, Hove, UK
Martin A, Fedio P 1983 Word production and comprehension in Alzheimer's disease: The breakdown of semantic knowledge. *Brain and Language* **19**: 124–41
McHugh P R, Slavney P R 1998 *The Perspectives of Psychiatry*. Johns Hopkins University Press, Baltimore, MD
Reischies F M 1998 Age related cognitive decline and the dementia threshold. In: Lomranz J (ed.) *Handbook of Aging and Mental Health. An Integrative Approach*. Plenum Press, New York, Chap. 19, pp. 435–48
Wells C E, Whitehouse P J 1996 Cortical dementia. In: Fogel B S, Rao S M, Schiffer R B (eds.) *Neuropsychiatry*. Williams & Wilkins, Baltimore, MD, pp. 871–89

F. M. Reischies

Dementia, Semantic

1. Introduction

Semantic dementia, which results from a degenerative brain disease, is a cognitive syndrome whose hallmark is a progressive deterioration of semantic memory. Semantic memory is the label given to the component of human memory that encompasses knowledge about concepts, objects, people, facts, and the meanings of words. It was once a matter of considerable controversy whether something so wide-ranging in form and content as semantic memory could constitute a separable cognitive subsystem, in the sense that damage to a specific brain region would disrupt this aspect of cognition and no other (Fodor 1983). However, the debate seems largely to have been resolved with a positive answer.

Semantic dementia (SD), given its current syndrome label by Snowden et al. (1989), was first described around 1900 in some individual case reports by neurologists intrigued by the fact that degenerative diseases could yield such striking and selective cognitive impairments (Girling and Berrios 1997: translation of Pick 1904). The modern history of SD began in the 1970s, with two papers cataloguing the cognitive syndrome in impressive detail (Warrington 1975, Schwartz et al. 1979). Further substantial interest in SD and related conditions has been stimulated in part by the recent dramatic increase in both availability and sophistication of structural brain imaging (particularly Magnetic Resonance Imaging, MRI) and techniques for analyzing and quantifying the brain abnormalities in individual MRI scans.

A specific cognitive impairment combined with a clearly identified locus of neuroanatomical lesion, documented across a large-ish number of patients, has traditionally been considered to licence the assignment of that cognitive function to that brain region. On this logic, the *sine qua non* of semantic memory is the anterior, inferior, lateral region of the temporal lobe, because this is consistently the location of the earliest and most severe atrophy in patients with the syndrome of SD (Mummery et al. 2000). Initially, this atrophy may be measurable just on one side, or at least is often very asymmetrical, but invariably there is bilateral temporal involvement with disease progression. There is also the important caveat, emphasized by recent studies of functional brain imaging, that areas which do not themselves reveal significant structural abnormalities may nonetheless fail to function adequately, presumably because of reduced input from the damaged areas to which they are normally connected. The first activational PET (Positron Emission Tomography) study of SD (Mummery et al. 1999) identified such an area in the posterior left temporal lobe. The consistent locus of atrophy in the antero-lateral temporal lobe for these patients only establishes that this region is one essential component of the brain's network for semantic memory.

2. The Core Syndrome of Semantic Dementia

2.1 Memory

SD is one of the main presentations of fronto-temporal dementia (see Sect. 5) that together constitute the second commonest cause of dementing illnesses with an onset before age 65. The most prevalent cause of dementia, certainly in later life but even before 65, is Alzheimer's disease (AD) (see *Dementia: Overview*). In typical presentations of AD, semantic memory is eventually also disturbed; the earlier and more prominent cognitive consequence of AD, however, is a severe impairment of the capacity for new learning, or episodic memory. A sharp distinction between episodic and semantic memory, originally proposed by Tulving (1972), is almost certainly over-simplified; but it does seem to capture a salient difference between two types of long-term memory. One's semantic knowledge of (for example) the Egyptian pyramids, comprising some facts about their appearance and about how, when, where, for what and by whom they were built, seems to stand in sharp contrast to one's episodic or autobiographical memory of the experience of visiting the pyramids. One typically cannot retrieve much in the way of episodic or contextual details associated with learning the semantic facts about these structures, whereas such details of time, place, sights, sounds, etc., seem the very essence of

memory for the autobiographical episode. An appropriate, though approximate, summary is that a patient with SD would remember at least some aspects of a trip to Egypt to see the pyramids, but would be unable to produce almost any conceptual knowledge about them.

There are at least two ways in which the notion of preserved episodic memory in SD should be qualified. First, there is solid experimental evidence that these patients have good ability to encode new memories of experiences based on perceptually rich information. When asked to look at photographs of paintings (Warrington 1975) or of objects like telephones, bicycles and zebras (Graham et al. 2000), followed by a larger set of such photos incorporating the earlier set plus some new ones, SD patients display good recognition memory for the items seen earlier. This is true even if the objects in the photos are ones on which the patients perform poorly in tests of conceptual knowledge. However, suppose that the recognition memory test employs perceptually altered materials: in the experiment with photos of objects, for example, the picture in the test phase would be of the same object type, such as a telephone, as that seen earlier, but represented in the test by a different exemplar of a telephone. In this case, the SD patients' episodic memory is significantly impaired relative to normal control subjects. This result suggests that stimulus meaning normally supports new learning, especially when the perceptual characteristics of the material to be recognized do not, on their own, readily reinstate the earlier event. The same principle helps to explain SD patients' poor recognition memory for written words, which constitute rather perceptually impoverished stimuli.

A second qualification on this issue concerns memory age. Probed for information about world events or about significant autobiographical episodes in their own lives, SD patients have revealed good memory for events and people prominent over the last year or so, but degraded and vague knowledge of earlier events (Graham and Hodges 1997). This pattern is consonant with the neuroanatomical facts of SD. Extensive research with humans and other animals suggests that structures in the medial areas of the temporal lobe (particularly the hippocampal complex) are essential for establishing new memories and for keeping them retrievable for some period of weeks or months. This is why amnesia-inducing head injuries, which often affect hippocampal regions, tend to disrupt memory of events just preceding the accident. It also explains why patients with Alzheimer's disease, to which medial temporal structures are especially vulnerable, are so poor at remembering recent events. The process whereby memories 'consolidate' in the brain seems to involve lasting changes in the more lateral structures of the temporal lobe, with decreasing reliance on the hippocampus for memory retrieval. SD is associated with significant atrophy in these lateral regions, but less so in the medial areas, thus fitting the observed relative preservation of very recent autobiographical memory. This constitutes a qualification on a sharp semantic/episodic distinction because, while semantic memories are typically 'old,' and thus likely to be disrupted by the disease process of SD, autobiographical memories may be old or new, and their status in SD according to this account will be determined more by this factor of age than by the content/quality of the memory itself.

2.2 Other Cognitive Features

Unlike patients with AD, those with SD are well oriented in time and space and find their way around with ease in familiar or even recently encountered (e.g., hospital) environments. They have good reasoning and problem-solving skills provided that the tasks do not require knowledge of specific concepts. Within the domain of language, they speak in well-formed sentences consisting of properly articulated words. The initial deficit that brings the patient into the clinic is almost always anomia: a difficulty in naming people, places, objects, and concepts. This early word-finding difficulty is typically followed, in both disease progression and prominence, by impaired word comprehension. Deterioration in both the production and comprehension of content-word vocabulary (nouns, verbs, adjectives) is significantly modulated by word frequency or familiarity, with more common words being more resilient. In SD speech, specific content words are replaced by more general and higher-frequency catchall terms: nouns such as 'thing' and 'place,' verbs such as 'make' and 'do.' One patient could only refer to each of the significant people in her life (husband, daughter, sister, etc.) as 'my person.'

These prominent twin problems with producing and understanding the content words of speech might imply that semantic dementia is a deficit restricted to words. On the other hand, in studies where patients with SD have been rigorously assessed on semantic tests involving no verbal stimuli or responses (such as pairing up objects that are standardly found and used together, or photos of famous people that are linked, or demonstrating the conventional use of familiar objects, Hodges et al. 2000), the impairment almost invariably extends well beyond language-based tasks. This suggests that SD is a general conceptual deficit that disrupts knowledge of objects and people as well as the meanings of words. The abilities to produce and comprehend words are probably especially vulnerable even to mild degradation of underlying conceptual knowledge because of the arbitrary relationship between meaning and word form. As discussed further in Sect. 3 and Sect. 4, the robustness of knowledge seems to be governed in part by the extent of consistency or systematicity between representations at different levels of the system. Names of things suffer because

they are devoid of systematicity: nothing about the word 'tiger' gives a clue to the nature of the object that it labels, or to the similarity of that object to a lion.

3. *The Cognitive and Neuroanatomical Progression of Semantic Dementia*

The disruption to conceptual knowledge in SD is eventually profound. A patient with severe SD (Hodges et al. 1992), during a test of yes/no questions about category membership and other semantic features of various familiar objects and animals, was asked 'Is a zebra an animal?' She replied 'Is a zebra an animal? I wish I could remember what an animal was.' At earlier stages of the disease, however, this broad level of object categorization is one of the better-preserved aspects of conceptual knowledge. This is indicated by a large range of cognitive tests, including semantic judgments about pictures or names of objects at different levels of specificity. When patients are studied longitudinally, the conceptual deficit emerges earliest for specific details such as the stripes on a zebra or where it lives, while success in classifying a zebra as an animal is maintained longest. The same pattern is revealed by the patients' responses in picture naming tasks, where they fail to name the particular exemplar but often provide either the general category name 'animal' or else a very high-frequency prototypical animal name such as 'dog' or 'horse' (Hodges et al. 1995). In another revealing anecdote, a patient responded to a picture of a zebra by saying 'It's a horse, isn't it?' Then, as she pointed to the stripes, she queried 'What are these funny things for?' An influential interpretation of this pattern of conceptual deterioration in SD, with important implications for the way in which such information is structured in the normal system, is the notion of disruption to elements in a similarity-based sub-symbolic connectionist network (McClelland et al. 1995). Features that are characteristic of many similar concepts (such as four legs on an animal) are more securely supported by connection weights across the network and are therefore more robust under damage than idiosyncratic features like the stripes of a zebra.

Another notable longitudinal feature of SD concerns the severity of the patients' anomia in relation to the specific pattern of neuroanatomical progression. As noted earlier, although the infero-lateral temporal atrophy characteristic of SD seems invariably to be bilateral beyond the earliest stages of the disease, it is often highly asymmetrical. The fact that semantic memory is disrupted by a predominance of either left or right damage is most consonant with the assumption that the semantic network is distributed bilaterally across both temporal lobes. Recent research demonstrates, however, that for approximately equated degrees of semantic deficit, the anomia is more severe in patients with more prominent left temporal atrophy than for those with greater right-sided abnormality (Lambon Ralph et al. 2001). The left-hemisphere elements of the distributed semantic network may be more crucial for activating the phonological system for speech production, which is typically strongly left-lateralised in the human brain.

4. *Collateral Symptoms*

Knowledge of word meanings is clearly central to the abilities to produce and comprehend informative connected speech, and also to the ability to name objects and people; these aspects of language competence therefore progressively deteriorate in patients with SD. By contrast, there are a number of more constrained language skills, involving the transformation of a word from one modality or form to another, that seem largely independent of word meaning. Two examples of a modality transformation are reading a word aloud (a transformation from orthography to phonology) or writing a word to dictation (phonology → orthography); and an example of a simple change in word form is generating the past tense of a verb from its present-tense form. If word meaning is not significantly implicated in these language skills, then patients with SD should perform well on these tasks. Do they? The answer is yes, and no. Their level of success depends on two aspects of the stimulus word to be transformed: whether it is a common or uncommon word, and whether the relationship between the stimulus and response form is consistent or inconsistent with that for other similar words in the speaker's vocabulary. In the English language and in the domain of reading aloud, an example of an inconsistent or irregular word is PINT, because the relationship between its spelling pattern and its pronunciation deviates from other similar and more typical words like PRINT, MINT and HINT. Likewise, in English and in the domain of past-tense verb morphology, an atypical word is THINK; unlike most verbs, and even unlike the phonologically similar verb BLINK, the past-tense of THINK is not THINKED but THOUGHT.

For reasonably high-frequency words with a typical relationship between stimulus and response, patients with SD achieve excellent reading aloud, spelling to dictation and transformation of a verb from its present to past tense. Relative to normal age- and education-matched subjects, the patients show a significant but not too severe decrement for words that are either high frequency but irregular, or regular but lower frequency. On low-frequency irregular words, however, their average success is a dramatic 30–50 percent below normal on the three tasks (Patterson 2000). The patients' errors to these difficult items reflect the well-learned generalizations: the overwhelmingly typical error is to pronounce a written word like 'pint' as if it rhymed with 'mint;' to spell a spoken word like 'cough' as 'coff;' and to turn the spoken present-tense verb

'grind' (in a sentence context like 'Today I grind' the coffee; yesterday I '—— the coffee') into grinded. To the extent that parallel experiments have been performed in languages other than English, patients with SD have revealed precisely comparable phenomena. For example, the generalizations and inconsistencies of the relationship between orthography and phonology in the non-alphabetic writing systems of Japanese are quite different from those in alphabetic English; but Japanese written words do have varying degrees of consistency in this relationship, and Japanese patients with SD succeed in reading aloud the high-frequency consistent words, and fail on the low-frequency atypical words (Patterson et al. 1995).

How is this pattern of performance to be understood? One view of word processing, based on networks of parallel distributed processing, includes the premise that semantic representations of words interact with phonological and orthographic representations in all forms of lexical processing (Plaut et al. 1996). It further predicts that the interaction with semantic knowledge should have a somewhat more influential role precisely for this set of words that the patients find difficult. The prediction regarding word frequency stems from the fact that, in the networks responsible for these tasks, the strength of connections is heavily influenced by the frequency with which particular patterns are encountered and processed. The prediction regarding typicality of relationship between stimulus and response stems from the fact that the very same distributed network that is sensitive to individual word knowledge and frequency also captures similarities and generalizations about stimulus-response relationships across words in the vocabulary. In this framework, if a word benefits from neither strong connections in its own right (by virtue of being commonly encountered), nor from the strong shared connections that apply to words with typical mappings, its processing comes to depend more on additional activation from other parts of the language system, particularly word meaning. According to this theory, then, the SD patients' pattern of successes and failures in these relatively 'non-semantic' language tasks can be explained in terms of the reduced activation available from semantic representations.

5. *Semantic Dementia in the Context of other Related Syndromes*

The book by Snowden et al. (1996) entitled *Fronto-Temporal Lobar Degeneration* distinguished three major syndromes that result from neurodegenerative diseases selectively affecting the frontal and/or temporal lobes of the human brain. How does SD relate to the other two syndromes, behaviorally and neuroanatomically?

Progressive aphasia, first described in the modern era by Mesulam (1982), is a disorder restricted to the domain of language, at least until late in disease progression. SD itself has sometimes, perhaps confusingly, been called progressive aphasia, because of its profound impact on both language production and comprehension; but this alternative syndrome of progressive aphasia is very different. SD gradually obliterates content word vocabulary, leaving fluent but anomic speech with good phonology and syntax. By contrast, the hallmark of nonfluent progressive aphasia, as it is more informatively called, is disrupted phonological and syntactic processing (Hodges and Patterson 1996). Progressive aphasic patients of this type lose the ability to produce connected sentences; their speech is increasingly restricted to short phrases or even to single content words, and these often contain phonological distortions. For example, one patient named a picture of an elephant as 'feleftent:' recognizable if the referent is known but obviously very distorted from the correct phonological word form. Neuroanatomically, nonfluent progressive aphasia seems to be associated with a less focal pattern of atrophy than SD. Both *in vivo* structural imaging and post-mortem analyses reveal atrophy in the left hemisphere, including both superior temporal and posterior frontal regions around the sylvian fissure.

The third member of the triumvirate of conditions described by Snowden et al. (1996) is frontotemporal dementia, often called just frontal (or frontal lobe) dementia. Although an adequate understanding of frontal lobe function in human behavior and cognition is yet to be achieved, it seems clear that much of what can be called human problem solving, judgment, decision making, emotional and social appropriateness, etc., crucially depends on intact frontal lobes; and these are precisely the capacities that are eroded in frontal dementia. Although profound semantic disruption is characteristic of SD and not of frontal dementia, SD often does, especially with disease progression, begin to reveal some of the behavioral disturbances that are the hallmark of frontal dementia. This is perhaps not surprising in light of the fact that, after initial consistent abnormalities of the anterior temporal lobe, atrophy in SD typically spreads not only posteriorly within the temporal lobe but also anteriorly to inferior frontal regions (Mummery et al. 2000). The behavioral changes include a strong preference for fixed daily routine, often to the point where it virtually becomes a demand for ritual.

6. *The Neuropathology of Semantic Dementia*

The underlying pathological cause of many neurodegenerative conditions can only be identified by analysis of brain tissue, which typically happens (if at all) only after the patient's death. Hodges et al. (1998) summarized the findings for all 13 cases to that date which met two criteria: (a) the patient had been studied in sufficient detail in life to enable identification of a clear

pattern of SD; (b) brain tissue was available for post-mortem analysis. In this set, there were no cases with the neurofibrillary tangles and neuritic plaques that constitute the pathological features of Alzheimer's disease. All 13 cases had either the specific intraneuronal inclusions that typify Pick's disease (Hodges 1994) or nonspecific degeneration/neuronal loss (see *Dementia*: *Overview*). It is worth mentioning, however, that Alzheimer's pathology has recently been identified post-mortem in a few patients who had profound semantic disruption, even if not the most prototypical presentation of SD (Galton et al. 2000). In these cases, the brunt of the AD pathology was found in the left anterior temporal lobe—atypical for AD but exactly where atrophy is centered in semantic dementia. It seems clear that cognitive deficits are determined more by the neuroanatomical locus of pathology than by its precise nature.

Bibliography

Fodor J A 1983 *The Modularity of Mind*. MIT Press, Cambridge, MA

Galton C J, Patterson K, Xuereb J, Hodges J R 2000 Atypical and typical presentations of Alzheimer's disease: A clinical, neuropsychological, neuroimaging and pathological study of 13 cases. *Brain* **123**: 484–98

Girling D M, Berrios G E 1997 On the symptomatology of left-sided temporal lobe atrophy (translation of Pick A 1904 Zur Symptomatologie der linksseitigen Schafenlappenatrophie. *Monatschrift fur Psychiatrie und Neurologie* **16**: 378–88). *History of Psychiatry* **8**: 149–59

Graham K S, Hodges J R 1997 Differentiating the roles of the hippocampal complex and the neocortex in long-term memory storage: Evidence from the study of semantic dementia and Alzheimer's disease. *Neuropsychology* **11**: 77–89

Graham K S, Simons J S, Pratt, K H, Patterson K, Hodges J R 2000 Insights from semantic dementia on the relationship between episodic and semantic memory. *Neuropsychologia* **38**: 313–24

Hodges J R 1994 Pick's disease. In: Burns A, Levy R (eds.) *Dementia*. Chapman and Hall, London, pp. 739–53

Hodges J R, Garrard P, Patterson K 1998 Semantic dementia. In: Kertesz A, Munoz D G (eds.) *Pick's Disease and Pick Complex*. Wiley-Liss, New York, pp. 83–104

Hodges J R, Graham N, Patterson K 1995 Charting the progression in semantic dementia: Implications for the organization of semantic memory. *Memory* **3**: 463–95

Hodges J R, Patterson K 1996 Nonfluent progressive aphasia and semantic dementia: A comparative neuropsychological study. *Journal of the International Neuropsychological Society* **2**: 511–24

Hodges J R, Patterson K, Oxbury S, Funnell E 1992 Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain* **115**: 1783–806

Hodges J R, Bozeat S, Lambon Ralph M A, Patterson K, Spatt J 2000 The role of conceptual knowledge in object use: Evidence from semantic dementia. *Brain* **123**: 1913–25

Lambon M A, McClelland J L, Patterson K, Galton C J, Hodges J R 2001 No right to speak? The relationship between object naming and semantic impairment: Neuropsychological evidence and a computational model. *Journal of Cognitive Neuroscience* **13**: 341–56

McClelland J L, McNaughton B, O'Reilly R 1995 Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–43

Mesulam M M 1982 Slowly progressive aphasia without generalized dementia. *Annals of Neurology* **11**: 592–8

Mummery C J, Patterson K, Wise R J S, Vandenberghe R, Price C J, Hodges J R 1999 Disrupted temporal lobe connections in semantic dementia. *Brain* **122**: 61–73

Mummery C J, Patterson K, Price C J, Ashburner J, Frackowiak R S J, Hodges J R 2000 A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology* **47**: 36–45

Patterson K 2000 The impact of semantic degradation on 'non-semantic' language abilities. *Higher Brain Function Research* **20**: 85–98

Patterson K, Suzuki T, Wydell T N, Sasanuma S 1995 Progressive aphasia and surface alexia in Japanese. *Neurocase* **1**: 155–65

Plaut D C, McClelland J L, Seidenberg M S, Patterson K 1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115

Schwartz M F, Marin O S M, Saffran E M 1979 Dissociations of language function in dementia: A case study. *Brain and Language* **7**: 277–306

Snowden J S, Goulding P J, Neary D 1989 Semantic dementia: A form of circumscribed cerebral atrophy. *Behavioral Neuroscience* **2**: 167–82

Snowden J S, Neary D, Mann D M A 1996 *FrontoTemporal Lobar Degeneration*. Churchill Livingstone, New York

Tulving E 1972 Episodic and semantic memory. In: Tulving E, Donaldson W (eds.) *Organization of Memory*. Academic Press, New York

Warrington E K 1975 Selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology* **27**: 635–57

K. Patterson and J. R. Hodges

Democracy

Democracy is universally understood as a form of government involving 'rule by the people,' which has essentially been its meaning since the term and practice were introduced in ancient Greece around 500 BC. At about the same time, a system of popular rule was also created in Rome in the form of the Roman Republic. Although efforts have sometimes been made to distinguish democratic from republican types of government, the only valid distinction is linguistic: democracy from Greek (*demos*, people, *kratia*, rule) and republic from Latin (*res*, thing, *publica*, public).

In any case, democracy as it came to be understood in the twentieth century was shaped not only by the ideas and practices of Greece, Rome, and the city-state

republics of medieval and Renaissance Italy, but perhaps even more importantly by the development of representative institutions in northern Europe and Britain.

No standard description of democracy exists, and the term has been used in many different and sometimes even contradictory ways. However, for an adequate account we need to describe democracy as a goal, aspiration, or ideal that a good political system would satisfy, and also as a set of actual political institutions, practices, or processes that are required in order for the criteria to be satisfied, within limits set by historical circumstances, human capacities, and other realities.

1. *Democracy as an Ideal*

To understand democracy as an ideal or goal, it is helpful to imagine an association composed of members who view one another as basically equal in their competence to participate in governing their association. In short, they consider themselves to be political equals. In order for the association to be governed in a manner consistent with this assumption, it would need to satisfy at least four criteria. Before a policy is adopted by the association, all the members should have equal and effective opportunities for making their views known to the other members as to what that policy should be. Next, when the time arrives at which the decisions about policy will finally be made, every member should have an equal and effective opportunity to vote on the final decisions of the association, and the vote of each should be counted as equal. Third, within reasonable limits as to time, each member should have equal and effective opportunities for learning about the relevant proposals under consideration, their alternatives, and their likely consequences. Finally, in order to insure that the process required by the three preceding criteria is never closed and the association's policies are always open to change by its members, they should also have the exclusive opportunity to decide how proposals are to be placed on the agenda for decisions, and, if they choose, what proposals should be on the agenda.

Each of these criteria implies the existence of a right: the rights of members to participate in decisions, to equal votes, to explore the issues, to decide on the agenda. Thus the criteria not only prescribe a process; they also prescribe a set of rights to which members of a democratic association are entitled.

Needless to say, these democratic criteria are so demanding that probably no association ever fully achieves them. Yet they are useful for judging how far and in what ways an association is undemocratic and for designing, introducing, and improving democratic governments.

Although democratic criteria can be, and often are, applied to many different kinds of association, their most important application is to the government or the state. Because a state possesses extraordinary power, exercises control over means of coercion to enforce its rules, including imprisonment and death, and has enormous capacities for achieving both harm and benefit, advocates of democracy from the Greeks onward have held that the most desirable state is one governed by democratic procedures.

1.1 *Democracy and the State*

Democracy cannot be applied to a state, however, without an answer to fundamental and often difficult question. To whom should the assumption of political equality be extended? Among all those subject to the government of a state, which persons should be entitled to all the rights and opportunities of membership described earlier? Who, in other words, should be full citizens?

That these rights and opportunities should be confined to adults, variously defined, has never been seriously in question. However, from the introduction of democratic ideas and practices in Greece and Rome until the twentieth century, the assumption that only a minority of adults was qualified for the rights of full citizenship all but universally prevailed. Thus all women were excluded, and so too were many males: slaves, the poor or propertyless, illiterates, and others. Consequently, throughout most of its long history democratic and republican government was not, in fact, government by 'the people' who were subject to it. 'Democracy' in governing a state was understood instead to mean government by a minority of adult males—sometimes, indeed, a very small minority.

During the first quarter of the twentieth century a fundamental change in democratic ideas and practices occurred in most 'democratic' countries then existing: women gained the franchise and other political rights. Democratic ideas and practices shifted away from exclusion toward inclusion. This shift was amplified in the last quarter of the century when other groups that had been denied the rights and privileges of citizenship, such as African-Americans in the southern states of the USA and long-settled immigrants in Germany, were granted those rights and privileges.

During the twentieth century, then, a fifth democratic criterion gained wide acceptance: a state could not be democratic unless all or almost all adults subject to it were full citizens entitled to all the rights and opportunities implied by basic democratic criteria. In effect, then, even if adults were unequal in other respects, it was now assumed that they should be treated as political equals in governing a state.

2. *Democratic Institutions*

The actual political institutions that might be required to achieve these democratic criteria (within the limits of the real world) can vary depending on the actual

circumstances of time and place, and historically they have varied in important ways. One aspect of the real world that greatly influences both democratic theory and practice is the size or scale of the unit to be governed by democratic processes. The two most relevant dimensions are the number of members (in a state, citizens) and area. Units consisting of a small number of members associating in a small, compact area not only tend to be more homogeneous than larger units with a great many members, but small units can also provide greater opportunities for direct participation in political life. In a very small and compact unit, such as a committee or a town, all the members can meet face-to-face, and within reasonable limits of time every member can engage in discussion and deliberation with the others. Consequently, other things being equal, smaller units are potentially more capable of meeting democratic criteria at a higher level than larger units.

It is not altogether surprising, then, that democratic practices (limited of course to males) first took root in towns and cities, as they did, for example, in ancient Greece and Rome, in some Italian city-states during the late Middle Ages and early Renaissance, in mountain villages in central Switzerland, and among small bands of Vikings during the eleventh and twelfth centuries. In these circumstances it was feasible, indeed quite natural, for the free adult males all to assemble in one place, at a site perhaps marked by traditional boundaries, and there to make their collective decisions. The most basic democratic institution was thus the citizen assembly. 'Democracy,' then, meant the direct democracy of the citizen assembly or town meeting.

Yet for many purposes small units are inadequate. For one thing, a city-state can be conquered by a larger imperial state, as Athens was in 323 BC by Macedonia, and later by Rome. Moreover, the important problems that human beings face, including the need to obtain food, water, or myriad other wants, typically require actions beyond the boundaries of a small unit. Consequently, the more interdependent human beings become, the less satisfactory small units become. At the very heart of democratic theory and practice, then, lies the dilemma of scale: the smaller a unit, the greater are the opportunities it can provide its members for participating in governing themselves; but the lower is the capacity of its government for dealing satisfactorily with all the important problems its members face. Conversely, the larger the unit the greater is the capacity of its government for dealing satisfactorily with problems that transcend the boundaries of the smaller units within it; but the more reduced are the opportunities it can provide its members for participating in its decisions.

Since the Middle Ages, city-states and other small units have been almost universally incorporated into nation-states or countries. The dilemma of scale has been met, not always satisfactorily, by granting some autonomy on some matters to smaller units within a country—towns, cities, states, provinces, regions, and so on—and maintaining some control, even exclusive control, by the national government.

To provide a satisfactory level of democracy in the larger units required a set of political institutions which, taken in their entirety, were historically novel.

2.1 The Institutions of Large-scale Democracy

Modern large-scale democracy in units the size of a country, even a small country, required the creation of new political institutions. These can be described at several levels of specificity. At the most general level one can specify the basic political institutions that are essential in order for a country to satisfy democratic criteria: to become and remain a democratic country. At a more specific level, the main varieties of constitutional arrangements that have existed in different democratic countries can also be described—for example, presidential or parliamentary systems, federal or unitary, two-party or multi-party, and so on. Finally, at the most specific level one might describe a country's concrete political institutions and practices, which are likely to form a unique pattern.

Descriptions of the second and third kind are beyond the scope of this article. At the most general level, however, the basic political institutions required in a democratic country are these:

(a) Elected officials. Control over government decisions about its laws and policies is constitutionally vested in officials elected by citizens. Thus modern, large-scale democratic governments are representative.

(b) Free, fair, and frequent elections. Elected officials are chosen in frequent and fairly conducted elections in which coercion is comparatively rare.

(c) Freedom of expression. Citizens have a right to express themselves without danger of severe punishment on political matters broadly defined, including criticism of officials, the government, the regime, the socioeconomic order, and the prevailing ideology.

(d) Access to alternative, independent sources of information. Citizens have a right to seek out alternative and independent sources of information from other citizens, experts, newspapers, magazines, books, telecommunications, and the like. Moreover, alternative sources of information should actually exist that are not under the control of the government or any other single political group attempting to influence public political beliefs and attitudes, and these alternative sources are effectively protected by law.

(e) Autonomous associations. To achieve their various rights, including those required for the effective operation of democratic political institutions, citizens also have a right to form relatively independent associations or organizations, including independent political parties and interest groups.

(f) Inclusive citizenship. No adult permanently residing in the country and subject to its laws can be denied the rights that are available to citizens and are necessary to the five political institutions listed above.

3. Beyond Polyarchal Democracy

It is not difficult to show why each of the institutions described in Sect. 2.1 is necessary in order for a country to satisfy ideal democratic criteria at a minimally acceptable level. Without freedom of expression, for example, it would be impossible for citizens to make their views known, hear the views of others, learn what they need to know about relevant proposals, or control the agenda.

Because the combination of institutions described above was developed fully only in the twentieth century, some political scientists wishing to distinguish such modern systems more clearly from other forms of democracy that exist or had existed in smaller units, such as the city-states of Athens or the town meetings of New England, referred to it as polyarchy or polyarchal democracy. In any case, during the twentieth century, political scientists and other observers tended to agree that a considerable number of countries adequately maintained the political institutions described above, and therefore should be regarded as democratic countries. Typical examples were France, India, Japan, the Scandinavian countries, the United Kingdom, the USA, and most other European and English-speaking countries.

In practice, however, the minimally acceptable levels for satisfying democratic criteria have shifted upward over time. Following his visit to the USA in 1831 Alexis de Tocqueville famously described that country as a democracy, as did his contemporaries in Europe and America. Yet an observer employing twentieth-century standards might hesitate to designate a country as a democracy as long as it excluded women from the suffrage, supported slavery, forcibly expelled its indigenous peoples from their territories, and denied them citizenship.

Because the ideal democratic criteria set extraordinarily high and perhaps unattainable standards, it is altogether possible that an increasing number of citizens in democratic countries might conclude that the institutions of polyarchal democracy are inadequately democratic. If so, the acceptable level for meeting democratic criteria might continue on the upward trajectory traced during the twentieth century.

See also: Democracy, History of; Democracy: Normative Theory; Democratic Theory; Democratic Transitions; Planning, Politics of; Political Culture; Political Discourse; Political Thought, History of; Voting: Compulsory; Voting, Sociology of; Voting: Tactical; Voting: Turnout

Bibliography

Berg E 1965 *Democracy and the Majority Principle: A Study in Twelve Contemporary Political Theories*. Akademiforlaget, Goteborg, Sweden

Burnheim J 1985 *Is Democracy Possible? The Alternative to Electoral Politics*. University of California Press, Berkeley, CA

Dahl R A 1998 *On Democracy*. Yale University Press, New Haven, CT

Dahl R A 1989 *Democracy and its Critics*. Yale University Press, New Haven, CT

Diamond L J 1999 *Developing Democracy: Toward Consolidation*. Johns Hopkins University Press, Baltimore

Ely J H 1980 *Democracy and Distrust: A Theory of Judicial Review*. Harvard University Press, Cambridge, MA

Mishkin J S 1991 *Democracy and Deliberation: New Directions for Democratic Reform*. Yale University Press, New Haven, CT

Gutmann A, Thompson D F 1996 *Democracy and Disagreement*. Belknap Press of Harvard University Press, Cambridge, MA

Hansen M H 1991 *The Athenian Democracy in the Age of Demosthenes: Structure, Principles, and Ideology*. Blackwell, Oxford, UK; Cambridge, MA

Isaac J C 1998 *Democracy in Dark Times*. Cornell University Press, Ithaca, NY

Lijphart A 1977 *Democracy in Plural Societies: A Comparative Exploration*. Yale University Press, New Haven, CT

Lijphart A 1999 *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*. Yale University Press, New Haven, CT

Lindblom C E 1988 *Democracy and Market System*. Norwegian University Press, Oslo, Norway; Distributed worldwide excluding Scandinavia by Oxford University Press, New York/London

Linz J J, Stepan A (eds.) 1978 *The Breakdown of Democratic Regimes*. Johns Hopkins University Press, Baltimore

Sartori G 1987 *The Theory of Democracy Revisited*. Chatham House Publishers, Chatham, NJ

R. A. Dahl

Democracy, History of

In a democracy, political authority is based on the will of the people. From its origins in ancient Greece, democracy has been what the British philosopher W. B. Gallie called an 'essentially contested concept,' that is, a concept whose meaning is permanently open to debate. (Gallie 1955–6) The history of democracy is the history of these debates, debates between democrats and their opponents and, no less vigorously and persistently, among the advocates of different sorts of democracy. Three overlapping sets of issues have dominated disagreements about democratic theory and practice. First and most obviously is the question of whom to include in the demos. If democracy means government by the people, who are the people? Everyone? All Adults? All male adults? All male adults

with property or education? Second is the matter of democracy's prerequisites. What rules, institutions, and values are necessary for democracy to succeed? Are correct procedures sufficient to ensure democratic government or does genuine democracy require a particular kind of society or culture? Is there one road to democracy or many? And finally, there is the issue of what might be called democracy's institutional scope. Is democracy primarily a political system of self-government or is it equally important, or equally necessary, in other aspects of life, such as the family, the workplace, and the schoolroom? In order for there to be a 'true' democracy must its principles and aspirations be realized throughout the social order?

1. *Classical Models*

The openness and complexity of democracy's history were already apparent in the fifth and fourth centuries BC when a number of Greek-speaking city states scattered across the eastern Mediterranean adopted various forms of popular self-government. From the beginning, there was an ambiguity in the meaning of demos which reflected the Greeks' uneasiness about democratic government. For democracy's admirers, the people's will confirmed legitimacy: decrees passed by popular assemblies often began with the words, 'the demos has decided.' But for democracy's critics, the people were the mob, unreliable and easily swayed by passion and self-interest. This is why Plato, in his famous attack on democracy in the eighth book of The Republic, argues that the excesses of democracy will inevitably lead to tyranny.

The ancient city's distinctive blend of religion and politics, the small scale of its public life, the superiority of the spoken over the written word in its political deliberations, and its emphasis on military service and war, reflected a distinctive set of historical conditions that would never be repeated. Nevertheless, the Greek experience continued to play a role in Europe's political imagination. Pericles's eloquent defense of Athenian democracy, recorded (or perhaps composed) by Thucydides in his history of the Peloponnesian War, was persistently cited as a statement of democratic virtue, just as Plato's and Aristotle's criticisms of democracy were used to underscore its dangers. Like the elements of classical buildings, which architects adapted over and over again to serve their own purposes, Greek ideas about democracy were repeatedly borrowed by European thinkers, who often imbedded them in structures quite distant from their original meaning and purpose.

There is no institutional continuity between ancient Greece and modern politics. The democratic practices that Pericles praised and Plato decried did not survive the destruction of the polis by internal discord and foreign enemies. To be sure, the Roman state based its legitimacy on the popular will: its armies fought in the name of 'The Senate and People of Rome.' But after the fall of the republic, first the substance and then the form of popular sovereignty disappeared from European history for almost a millennium. When they reappeared in the late middle ages, republican institutions were limited to relatively small entities, usually city states, in which political power was in the hands of a narrow, exclusive elite. There were also a few self-governing rural communities in medieval and early modern Europe, but they were small and scattered. None of these polities thought of itself as a democracy. The concept, revived with the rest of the Greek legacy in the late middle ages, was a matter for political theory and scholarly debates, not the everyday conduct of public affairs.

2. *The Age of Democratic Revolution*

The history of modern democratic practice begins with a series of political revolutions in the final decades of the eighteenth century. But before erupting into the realm of practical politics, democratic ideals were nourished by three currents of enlightenment thought: first was a new concept of the individual, whose identity was defined by his (and for a few, also her) own talents and desires, rather than his place in the social order; second was a new concept of equality, which accorded to each person (definitions of which varied greatly) the same rights and responsibilities; and finally there was a new concept of social and intellectual activity, which imagined a sphere of free competition among people, commodities, and ideas. Many, perhaps most, of the thinkers who formulated these ideas did not think of themselves as democrats; some of them expressly condemned democratic institutions in terms that Plato would have recognized. Nevertheless, enlightenment notions of individualism, equality, and society helped to undermine the ideological foundations of the old regime and to open the way for democracy's long and circuitous march to political power.

Among enlightenment political theorists the most significant for the history of democracy was Jean Jacques Rousseau, whose ideas inspired the revolutionaries' assaults on the old regime and at the same time reflected democracy's deepest ambiguities. Rousseau is correctly regarded as a spokesman for individual liberty and, no less accurately, as the advocate of the individual's subordination to the 'general will.' He attempted to reconcile these two convictions with a program of education that aimed at freeing individuals from the corruptions of existing society and a program of political reform that was supposed to create institutions appropriate for the products of this education. The result would be a genuinely free society in which individuals want to choose what is best for them and for their community. Because this vision of total harmony between individual desire and social imperatives set a standard that no government could match, among Rousseau's

legacies was the temptation to force people to be free, that is, to use political power to shape their desires to match the alleged needs of the community.

The first step from democratic theory to democratic practice was taken on the far fringes of the European world, in North America, where the struggle for independence from colonial rule was carried out in the name of 'the people's' natural and inalienable right to self-government. Of course the new American democracy was qualified in a number of important ways: the demos excluded women, African-American slaves, and, in some places, men without sufficient property. Nevertheless, the creation of a republic in the new world established a regime whose institutions and legitimacy depended on the people's will.

The democratic experiment in America aroused widespread interest and some enthusiasm, but no imitators among enlightened Europeans. But when revolution engulfed France, Europe's greatest power and the recognized center of fashion and culture, its wider significance could not be ignored. In the summer of 1789, the French National Assembly solemnly proclaimed 'the rights of man and the citizen,' which promised social equality and political liberty. Sovereignty was now firmly rooted in the popular will: 'No body and no individual may exercise authority which does not emanate directly from the nation.' But who spoke for the nation? At each stage of the revolution there was a different answer to this question as power moved from the nation's elected representatives, to the mobilized people of Paris, to the Committee of Public Safety, to an elite of propertied men intent on restoring order, until it finally came to rest in the hands of an heroic young general, Napoleon Bonaparte. As the revolution moved from moderation to terror to reaction to military dictatorship to imperial monarchy, each successive regime invoked the nation, each claimed to rule in the people's name. Beyond France's borders, the revolution's friends and enemies shared one essential conviction: once the forces of democratic government had been unleashed, the world would never be the same.

For more than a century after the revolution was defeated, its historical memory lived on, a common source of inspiration and anxiety. People drew from the revolution's grand narrative the lessons appropriate for their political hopes and fears. To a democrat like Jules Michelet, the revolution was the truest expression of France's commitment to liberty, the most glorious chapter in the 'people's' struggle to realize their historical destiny. Reactionaries like Hippolyte Taine drew a different message from the revolution's course: popular sovereignty, he argued, will always lead first to anarchy, then despotism. But the revolution's most astute analyst was neither a democrat nor a reactionary: Alexis de Tocqueville saw it as part of a larger historical process through which the ideals and institutions of old-regime Europe were destroyed by the forces of social and political democracy that were deeply woven into the fabric of modern life. Like many of democracy's critics, Tocqueville feared it could easily lead to dictatorship; only in America was there a chance that democracy's true promise could be fulfilled (Tocqueville 1988).

3. *The Nineteenth Century*

In 1815, many Europeans hoped or feared that democracy had been defeated for ever. The monarchical principle seemed triumphant; republican institutions survived in Switzerland and a precarious handful of city states. Although monarchy remained the most common form of government in Europe for another century (France after 1870 was the most important exception), almost everywhere representative institutions gradually began to play an important role in public life. The right to vote, always the most obvious marker of democracy's progress, was granted to an increasing number of European males in the course of the century; political parties, interest groups, trade unions, and a variety of newspapers and other publications provided an infrastructure that supported political participation. By 1900, therefore, the distinction between monarchical and democratic governments, which conservative statesmen had struggled to maintain, was hopelessly blurred.

Every political movement had to come to terms with democratic pressures. During the first half of the century, most European liberals, who were the direct heirs of the Enlightenment's ideals of individualism, equality, and freedom, were suspicious of the masses—all the more so after the collapse of the French republic into terror and dictatorship. But in the course of the century, the left wing of the liberal movement was reconciled to democracy, supported universal suffrage (for men) and tried, with varying degrees of success, to win their share of the newly enfranchised voters. Conservatives, who had begun the century as democracy's most consistent enemies, were also unable to resist being pulled into the orbit of democratic politics. Following the example of the Emperor Napoleon III, who had consolidated his position with popular plebiscites after being elected president of France in 1848, conservative statesmen such as Bismarck in Germany and Disraeli in England, set out to use democratic means in support of conservative ends. No less remarkable than these 'Tory democrats' were the defenders of the Roman Catholic church, another of democracy's former enemies, who turned to electoral politics to defend their religious interests against the intrusions of the secular state. Here too we find evidence of the inexorable expansion of democratic ideals and institutions.

Some nineteenth century democrats insisted that political reform was not enough; authentic self-government, they believed, required social revolution. Far and away the most influential version of this belief

was formulated in the 1840s by Karl Marx and Friedrich Engels, who argued that a proletarian revolution was required to establish a truly free and equal society. To Marx, Engels, and their more orthodox disciples, democratic politics was always a means to an end: parties, unions, elections had no value beyond helping the proletariat obtain the power they needed to transform society. In this transformed society, organized political activity would become superfluous because class conflict would disappear. But as so often happens, the socialist parties' democratic means began to shape the ends for which they fought. Eventually the political organizations they built took on their own value, winning elections became an end in itself, democratic politics, a way of life. By the early twentieth century, the European labor movement, most prominently the Social Democratic Party in Germany, had become the most consistent and resolute defender of democratic reform.

In the nineteenth century, the most significant extension of democratic aspirations was not to the social order but rather to the national community. The French revolutionaries had insisted that because people did not belong to a ruler, they could not be traded back forth like pieces of property; nations, like individuals, had the sacred right to determine their own character and destiny. In the course of the century, self-determination came to mean the right of nations to have their own state, a fateful combination of democracy and sovereignty that made the national state the only legitimate way of organizing political space. In the middle decades of the century, Italians and Germans formed new nations; nationalities like the Polish and Irish struggled unsuccessfully to follow their example. In fact, nationality was an elastic and elusive quality, easy enough to invoke, but extraordinarily difficult to use as the foundation for a state. Even in the West, where strong states had had centuries to impose a kind of cultural homogeneity, ethnic minorities remained, while in the multinational empires of the East, the complexities of ethnic settlement meant that any potential nation state would have to contain substantial linguistic and religious minorities. Self-determination, the British political scientist Ivor Jennings remarked, 'seemed reasonable enough: let the people decide. It was in fact ridiculous because the people cannot decide until someone decides who are the people' (Anderson 1996, p. 40).

As difficult as the question of self-determination might be within Europe itself, its problems became even greater when applied to the rest of the world, much of which remained either directly or indirectly under European control. But if the Germans and Italians had a right to their own nation, if the Irish and the Polish could aspire to nationhood, what about the populations of India, Africa, and Asia? In the nineteenth century, even progressive Europeans would have responded that these peoples were not yet ready for independence because they were barbarous. John Stuart Mill, for instance, believed that 'independence and nationality, so essential to the due growth and development of a people further advanced in improvement, are generally impediments' to uncivilized peoples (Mill 1859, p. 252). But by the early twentieth century, advocates of national independence had challenged European influence in the ancient empires of the Manchus and Ottomans, while the ideals of democratic self-determination slowly began to attract followers throughout the colonized world. The global movement against colonization, which would be central to the history of the twentieth century, had many sources and took various forms, but its origins and eventual triumph cannot be understood without the democratic tools that had been forged within Europe itself.

4. *The Twentieth Century*

The historical period that stretched from the Congress of Vienna in 1815 to the outbreak of the First World War in 1914 began with democracy's defeat and ended with its apparently irresistible, if still notably incomplete spread throughout Europe. The period from 1914 to 1945 was characterized by a series of near fatal assaults on the integrity and viability of democratic institutions. There were moments during this dark epoch when thoughtful people could readily imagine that the ideals that had inspired Rousseau and the eighteenth-century revolutionaries, having been abandoned by their advocates and defeated by their enemies, would be consigned to history's dustbin. Who in 1940 could have predicted that, at the century's end, it would be democracy's enemies who would everywhere be in disarray?

That the years after 1914 would be so full of peril for democracy was not immediately apparent. In fact, the First World War destroyed many of democracy's most obdurate opponents, especially the Russian empire, which collapsed into revolution in 1917. Moreover, in most of Europe and America the war accelerated the extension of voting rights to women, an issue that had become increasingly contentious after the turn of the century. Finally, the war ended with the victory of the Western powers, allied since 1917 with the USA, whose president, Woodrow Wilson, proclaimed that the purpose of the war was to make 'a world safe for democracy.' Wilson's League of Nations, a keystone of the treaties signed in the spring of 1919, was designed to create a society of democratic states, in which self-government at home would nurture goodwill abroad and thus purge the world of the twin evils of tyranny and aggression.

The Wilsonian peace was supposed to be the fulfillment of the belief, held in the eighteenth century by men like Immanuel Kant and in the nineteenth by the followers of Jeremy Bentham, that democracy and peace inevitably went together. 'There are properly no

other criminals than the heads of nations,' Bentham had declared, 'the subjects are always honest' (Hinsley 1963, p. 86). Always a triumph of hope over experience, this belief was especially illusory in the years after 1919, when, having been brutalized by the war and frustrated by its outcome, the peoples of Europe were not necessarily peace-loving, while their elected leaders—including those in the USA—were not prepared to cooperate in a common defense of democratic institutions.

In retrospect, therefore, it is clear that two other events in the spring of 1919 were more significant for democracy's immediate future than Wilson's treaties. In March, Benito Mussolini formed his first fascist organization, an inchoate band of the alienated and discontented, which, within three years, was able to take advantage of Italy's postwar crisis and become the basis for a new government. In one state after another, parliamentary democracy collapsed in the twenties and thirties, and nowhere more consequently than in Germany, where Mussolini's admirer, Adolf Hitler, created a dynamic regime based on race, radical nationalism, and anticommunism. Mussolini's fascism and Hitler's National Socialism were at once the enemies and the perversions of democratic politics. Both of them blended dictatorial practice with the trappings of participatory politics—a mass party, rallies and demonstrations, and plebiscites.

A few weeks after Mussolini formed his fascist organization in Milan, Lenin established the Third International in Moscow. Lenin's hopes for world revolution, still very much alive in 1919, were soon disappointed. Contrary to Marxist theory and his own expectations, Lenin was faced with the need to build a socialist society in Russia. This required the violent repression of dissent by the revolutionary regime and then, under Lenin's successor, Joseph Stalin, the ruthless use of state power to create a new social order from above. Tens of millions died in the doomed attempt to turn the USSR into a workers' paradise. Like fascism, which claimed to rule on behalf of the nation, Soviet Communism was a perversion of nineteenth century democratic aspirations which sought to conceal a brutal tyranny behind a facade of cynical democratic rhetoric and empty participatory practice. That so many intelligent people outside of the USSR were drawn to communism in the 1920s and 1930s was a symptom of the political malaise that afflicted democratic regimes everywhere in Europe and America.

The second global war of the twentieth century ended in 1945 with the defeat of the fascist powers. The interwar political spectrum, in which democratic regimes had faced enemies from both the left and the right, was now dramatically simplified; radical right-wing parties were discredited, outlawed, and, in some places, vigorously repressed. Almost all of the regimes that emerged from the war's wreckage declared themselves to be democratic. In the East, 'people's democracies' were formed under the auspices of the USSR and, with the exception of Yugoslavia and Albania, with the assistance of Soviet armies of occupation; in the West, parliamentary regimes were allied with and, to varying degrees, influenced by the USA. This bipolar division of the continent lasted until 1989, when the creation of a reformist regime in the USSR opened the way for democratic movements throughout Eastern Europe. With remarkable speed and, except in Romania, without violence, the communist regimes were driven from power and new, authentically democratic governments were elected to take their place. With the collapse, first of the Soviet imperium and then of the USSR itself, a new era in the history of democracy—with a new set of problems and possibilities—clearly began.

Among the casualties of the Second World War was European colonialism. The principle of national self-determination, proclaimed by Wilson at the end of the First World War and reaffirmed by Franklin Roosevelt and Winston Churchill at the beginning of the second, was not applied to the non-European world without great delay and difficulty. Obviously Japan lost its overseas possessions when it lost the war. While neither Britain nor France, who were on the winning side in 1918 and 1945, wanted to abandon their colonies, they lacked both the ability and the will to maintain them in the face of indigenous resistance, the hostility of both the USA and the USSR, and the reluctance of their own peoples to support a vigorous defense of their empires. Sometimes peacefully, sometimes after violent wars of liberation, Europe's former colonies achieved independence; by the 1980s virtually no residue of direct imperial rule remained. Although the struggle against colonialism was carried out in the name of national self-determination, it proved to be extremely difficult to plant democratic institutions in the new states. With a few notable exceptions, most significantly India, the world's largest democracy, most of Europe's former colonies, especially in Africa, fell under authoritarian regimes, often based on the power of the military. For many of the world's population, therefore, the promise of democracy remained unfulfilled.

5. *Conclusion*

Despite democracy's disappointing record in much of Africa and Asia, there are more functioning democracies in the world at the beginning of the twenty first century than at any other time in history. Furthermore, alternative sources of political authority have apparently disappeared; even dictatorial regimes usually feel obliged to exercise power in the people's name. Elections, even when they are manifestly fraudulent, have become the essential rituals through which authority is legitimized.

Although the alternatives to democracy have largely disappeared, at least in theory, debates about democracy's meaning continue. Some of these debates are new versions of old contests about the composition of the demos, the procedures best suited to guarantee democratic representations, the social and cultural preconditions of popular rule, and the proper scope of democratic politics. Some of the contemporary debates over democracy's meaning reflect new sources of conflict, such as the question of how—or whether—ethnic minorities should be shielded from the impact of majority rule? Others concern the extension of democratic principles into families, schools, religious institutions, healthcare systems.

These debates, as painful and occasionally destructive as they sometimes are, reflect the continued vitality of the democratic tradition. A more serious danger to democracy is a growing skepticism about whether democratic institutions can solve the most significant problems people face in their everyday existence. To the rich, democracy sometimes seems superfluous, an unnecessary distraction from the challenges and opportunities of private life. In the USA, for instance, less half of those eligible turn out to vote in national elections. To the poor, democracy can seem irrelevant, a woefully inadequate antidote to the poisonous power of poverty, violence, and political corruption. Over the past two centuries, the democratic vision has survived defeat and outlived its historic enemies; if it is to survive and flourish in the new century, it must recapture—and in some places, attain for the first time—its hold on ordinary men and women's political imagination.

See also: Aristotelian Social Thought; Citizen Participation; Democracy; Democracy: Normative Theory; Democratic Theory; Democratic Transitions; Participation: Political; Political Parties, History of; Political Representation

Bibliography

Anderson M 1996 *Frontiers*. Polity Press, Cambridge, UK

Bobbio N 1989 *Democracy and Dictatorship: The Nature and Limits of State Power*. Polity Press, Cambridge, UK

Bryce J 1921 *Modern Democracies*. Macmillan, New York

Dahl R A 1989 *Democracy and its Critics*. Yale University Press, New Haven, CT

Duverger M 1974 *Modern Democracies*. Dryden Press, Hinsdale, IL

Finley M I 1985 *Democracy Ancient and Modern*. Rutgers University Press, New Brunswick, NJ

Gallie W B 1955–6 Essentially contested concepts. *Proceedings of the Aristotelian Society New Series* **56**: 167–98

Held D 1996 *Models of Democracy*, 2nd edn. Stanford University Press, Stanford, CA

Hinsley F H 1963 *Power and the Pursuit of Peace*. Cambridge University Press, Cambridge, UK

Marshall T H 1950 *Citizenship and Social Class*. Cambridge University Press, Cambridge, UK

Mill J S 1859 *A Few Words on Non-Intervention, 1873 Dissertations and Discussions*. Holt, New York, Vol. 3

Ober J, Hedrick C (eds.) 1996 *Demokratia. A Conversation on Democracies, Ancient and Modern*. Princeton University Press, Princeton, NJ

Ostrogorski M 1902 *Democracy and the Organization of Political Parties*. Macmillan, New York

Putnam R D 1993 *Making Democracy Work*. Princeton University Press, Princeton, NJ

de Tocqueville A 1988 *Democracy in America*. Perennial Library, New York

J. J. Sheehan

Democracy: Normative Theory

Normative democratic theory deals with the ethical foundations of democracy and democratic institutions. It is distinct from descriptive democratic theory. It does not offer in the first instance a scientific study of those societies that are called democratic. It aims to provide an account of when and why democracy is ethically desirable as well as ethical principles for guiding the design of democratic institutions. Of course, normative democratic theory is inherently interdisciplinary and must call on the results of political science, sociology, and economics in order to give this kind of concrete guidance. This brief outline of normative democratic theory focuses attention on three distinct issues in recent work. First, it outlines some different approaches to the question of why democracy is morally desirable at all. Second, it explores the problem of the rationality of participation in large democratic societies given the inevitably small impact of participants on the outcomes. It also discusses blueprints of democratic institutions for dealing with this. Third, it surveys different accounts of the proper characterization of equality in the processes of representation. The last two parts display the interdisciplinary nature of normative democratic theory.

1. Foundational Questions

To fix ideas, the term 'democracy' refers very generally to a method of group decision making characterized by a kind of equality of input among the participants. That equality may be the mere formal equality of one person, one vote in an election for representatives to an assembly, or it may be more robust, including equality in the processes of deliberation and coalition building. 'Democracy' may refer to any of these political arrangements. Normative democratic theorists attempt to determine which one of these kinds of arrangements is desirable, ethically speaking, and in

what contexts. For instance, Joseph Schumpeter (1956) argues, with some force, that the first highly formal kind of democracy is highly desirable while the second is dangerous. On the other hand, Jean-Jacques Rousseau (1967) is apt to argue that the first kind of democracy is akin to slavery while only robustly egalitarian democracies have political legitimacy. To evaluate their arguments we must decide on the merits of the different principles and conceptions of humanity and society from which they proceed.

2. *The Justification of Democracy*

We can evaluate democracy ethically along at least two different dimensions. We can evaluate a group decision-making method in terms of the outcomes of using it and compare the values of those outcomes with those of other possible methods of political decision making. We can also evaluate democracy in terms of qualities that are inherent in the method itself. We can ask whether there is something inherently fair about making democratic decisions on matters on which people disagree.

3. *Instrumental Values*

The instrumental values often attributed to democracy can be broadly classed into two categories: the quality of the resulting legislation and the impact of democracy on the characters of the participants. John Stuart Mill (1991) argued that democratic ways of deciding tend to be superior to non-democratic ways on account of the fact that they produce superior legislation in three distinct ways. Strategically, democracy is thought to produce superior outcomes to other forms of rule just because it forces decision-makers to take into account the interests, rights and opinions of most people in society. Since democracy gives some political power to each, more people are taken into account than under aristocracy or monarchy.

Epistemologically, democracy is thought to be the best decision-making method on the grounds that it is generally more reliable in helping participants discover the right decisions. Since democracy brings a lot of people into the process of decision making, it can take advantage of many sources of information and critical assessment of laws and policies. Democratic decision making tends to be more informed than other forms about the interests of citizens and the causal mechanisms necessary to advance those interests. Furthermore, the broad-based discussion typical of democracy enhances the critical assessment of the different moral ideas that guide decision makers.

Regarding its effects on character, many have noted with Mill and Rousseau that democracy tends to make people stand up for themselves more because it makes collective decisions depend on them more than monarchy or aristocracy do. Hence, in democratic societies individuals are encouraged to be more autonomous. In addition, democracy tends to get people to think carefully and rationally more than other forms of rule because it makes a difference whether they do or not. Finally, some have argued that democracy tends to enhance the moral qualities of citizens. When they participate in making decisions, they have to listen to others, they are called upon to justify themselves to others and they are forced to think in part in terms of the interests of others. Some have argued that when people find themselves in this kind of circumstance, they come genuinely to think in terms of the common good and justice. Hence some have argued that democratic processes tend to enhance the autonomy, rationality, and morality of participants. Since these beneficial effects are thought to be worthwhile in themselves, they count in favor of democracy and against other forms of rule.

Some argue in addition that the above effects on character also tend to enhance the quality of legislation. A society of rational, autonomous, and moral decision makers is more likely to produce good legislation than a society ruled by a self-centered person or small group of persons who rule over slavish and unreflective subjects.

More detailed knowledge of the effects of political institutions can be used to discriminate in favor of particular kinds of democratic institutions or modifications of them. For instance, in the United States, James Madison (1788) argued in favor of a fairly strong federal government on the grounds that local governments are more likely to be oppressive to minorities. Of course, the soundness of any of the above arguments depends on the truth or validity of the associated substantive views about justice and the common good as well as a causal theory of the consequences of different institutions.

It is also the case that many theorists argue against any kind of democracy on instrumental grounds. Hobbes (1968) argues that democratic institutions are inferior to monarchical government on the grounds that they are more likely to breed destabilizing dissension among subjects. Plato (1974) argues that democracy is inferior to various forms of monarchy, aristocracy, and even oligarchy on the grounds that democracy tends to undermine the expertise necessary to properly governed societies. Some recent thinkers have argued in favor of giving over complete control over society to the market, on the grounds that democracy tends to produce serious economic inefficiencies. More modest versions of these arguments have been used to justify modification of democratic institutions.

4. *Instrumentalism*

Instrumentalists argue that these instrumental values of the democratic process are the only bases on which to evaluate democracy or compare it with other forms

of political decision making. The standards one uses depend on one's conception of a just and good society. For example, libertarian instrumentalists such as Friedrich Hayek (1960) argue that democracy is desirable only to the extent that it protects liberty and private property.

5. *Noninstrumental Values*

Few theorists deny that political institutions must be at least in part evaluated in terms of the outcomes of having those institutions. Some argue, in addition, that some forms of decision making are morally desirable independent of the consequences of having them. A variety of different approaches have been used to show that democracy has this kind of intrinsic value. The most common of these come broadly under the rubrics of liberty and equality.

6. *Liberty*

Some argue that the basic principles of democracy are founded in the idea that each individual has a right to liberty. Democracy, it is said, extends the idea that each ought to be master of his or her life to the domain of collective decision making. First, each person's life is deeply affected by the larger social, legal, and cultural environment in which he or she lives. Second, only when each person has an equal voice and vote in the process of collective decision making will each have control over this larger environment. Thinkers such as Carol Gould (1988) conclude that only when some kind of democracy is implemented will individuals have a chance at self-government. Since individuals have a right to self-government, they have a right to democratic participation. This right is established at least partly independently of the worth of the outcomes of democratic decision making. The idea is that the right to self-government gives one a right, within limits, to do wrong. Just as an individual has a right to make some bad decisions for himself or herself, so a group of individuals have a right to make bad or unjust decisions for themselves.Here we can see the makings of an argument against instrumentalism. To the extent that an instrumentalist wishes to diminish a person's power to contribute to the democratic process for the sake of enhancing the quality of decisions, he or she is committed to thinking that there is no moral loss in the fact that our power has been diminished; but if the liberty argument is correct our right to control our lives is violated by this.

One major difficulty with this line of argument is that it appears to require that the basic rule of decision making be consensus or unanimity. If each person must freely choose the outcomes that bind him or her, then those who oppose the decision are not self-governing. They live in an environment imposed on them by others. So only when all agree to a decision are they freely adopting the decision.

The trouble is that there is rarely agreement on major issues in politics. Indeed, it appears that one of the main reasons for having political decision-making procedures is that they can settle matters despite disagreement. Hence it is hard to see how any political decision-making method can respect everyone's liberty.

7. *Democracy as Public Justification*

One distant relative of the self-government approach is the account of democracy as a process of public justification defended by, among others, Joshua Cohen (1998). The idea behind this approach is that laws and policies are legitimate to the extent that they are publicly justified to the citizens of the community. Public justification is justification to each citizen as a result of free and reasoned debate among equals. Citizens justify laws and policies to each other on the basis of mutually acceptable reasons. Democracy, properly understood, is the context in which individuals freely engage in a process of reasoned discussion and deliberation on an equal footing. The ideas of freedom and equality provide guidelines for structuring democratic institutions.

The aim of democracy as public justification is reasoned consensus among citizens. However, a serious problem arises when we ask about what happens when disagreement remains. Two possible replies have been suggested to this kind of worry. It has been urged that forms of consensus weaker than full consensus are sufficient for public justification and that the weaker varieties are achievable in many societies. For instance, there may be consensus on the list of reasons that are acceptable publicly but disagreement on the weight of the different reasons. Or there may be agreement on general reasons but disagreement about particular interpretations of those reasons. What would have to be shown here is that such weak consensus is achievable in many societies and that the disagreements that remain are not incompatible with the ideal of public justification.

8. *Equality*

Many democratic theorists have argued that democracy is a way of treating persons as equals when there is good reason to impose some kind of organization on their shared lives, but they disagree about how best to do it. On one version, defended by Peter Singer (1973), when people insist on different ways of arranging matters properly, each person in a sense claims a right to be dictator over their shared lives. However, these claims to dictatorship cannot all hold up. Democracy embodies a kind of peaceful and fair compromise

among these conflicting claims to rule. Each compromises equally on what he or she claims as long as the others do, resulting in each having an equal say over decision making. In effect, democratic decision making respects each person's point of view on matters of common concern by giving each an equal say about what to do in cases of disagreement.

One difficulty is that this view relies on agreement much as the liberty views described above. What if people disagree on the democratic method or on the particular form democracy is to take? Are we to decide these latter questions by means of a higher order procedure? And if there is disagreement on the higher order procedure, must we also democratically decide that question? The view seems to lead to an infinite regress. Another egalitarian defense democracy asserts that it embodies the equal consideration of the interests of the citizens of a society when there is disagreement about how best to organize their shared life. On this account, equal consideration of interests requires that individuals' judgments be taken into account equal when there is disagreement. Respect for each citizen's judgment is grounded in the fallibility of everyone's judgment as well as the generally recognized tendency of the moral judgments of persons to be biased towards their own interests. Each citizen would have good reason to think that his or her interests were not being given the same weight as others if he or she had less decision-making power than the others. On this account, since each person has a pre-eminent interest in being recognized as an equal member of the community, only equality in decision-making power is compatible with equal consideration of interests. The principle of equal consideration of interests also implies limits to what can be up for democratic control and so the infinite regress noted above is avoided.

9. *Democratic Citizenship*

A vexing problem of democratic theory has been to determine whether ordinary citizens are up to the task of governing a large society. There are three distinct problems here. First, Plato argued that some people are innately more intelligent than others and that those persons ought to rule. Second, others have argued that a society must have a division of labor. If everyone were engaged in the complex and difficult task of politics, little time or energy would be left for the other essential tasks of a society. Conversely, if we expect most people to engage in other difficult and complex tasks, how can we expect them to have the time and resources sufficient to devote themselves intelligently to politics?

Third, since individuals have so little impact on the outcomes of political decision making in large societies, they have little sense of responsibility for the outcomes. Some have argued that it is not rational to vote since the chances that a vote will affect the outcome of an election are nearly indistinguishable from zero. Worse still, Anthony Downs (1957) has argued that almost all of those who do vote have little reason to become informed about how best to vote. On the assumption that citizens reason and behave roughly according to the Downsian model, either the society must in fact be run by a relatively small group of people with minimal input from the rest or it will be very poorly run.

These observations pose challenges for any robustly egalitarian or deliberative conception of democracy. Without the ability to participate intelligently in politics one cannot use one's votes to advance one's aims, nor can one be said to participate in a process of reasoned deliberation among equals. So, either equality of political power implies a kind of self-defeating equal participation of citizens in politics or a reasonable division of labor seems to undermine equality of power.

Plato and Hobbes argued against democracy in any form on these grounds. Some modern theorists of democracy, called elite theorists, have argued against any robustly egalitarian forms of democracy on these grounds. They argue that high levels of citizen participation tend to produce bad legislation designed by demagogues to appeal to poorly informed and overly emotional citizens. They look upon the alleged apathy and uninformedness of citizens evidenced in many empirical studies in the 1950s and 1960s as highly desirable social phenomena. Joseph Schumpeter's assertion that the 'democratic method is that institutional arrangement for arriving at political decisions in which individuals acquire the power to decide by means of a competitive struggle for the people's vote' (Schumpeter 1956, p. 269) still stands as a concise statement of the elitist view. In this view, the emphasis is placed on responsible political leadership. Political leaders are to avoid divisive and emotionally charged issues and make policy and law with little regard for the fickle and diffuse demands made by ordinary citizens.

An approach that is in part motivated by the concerns above but that attempts to avoid elitism is the interest group pluralist account of politics. Robert Dahl's early statement of the view is very powerful.'In a rough sense, the essence of all competitive politics is bribery of the electorate by politicians… The farmer… supports a candidate committed to high price supports, the businessman…supports an advocate of low corporation taxes… the consumer…votes for candidates opposed to a sale tax' (Dahl 1959, p. 69). In this conception of the democratic process, each citizen is a member of an interest group with narrowly defined interests that are closely connected to their everyday lives. On these subjects citizens are supposed to be fairly well informed and interested in having an influence. On this account, democracy is not rule by the majority but rather rule by coalitions of minorities.

The account ensures that individuals can participate roughly as equals to the extent that it narrowly confines the issues with which each individual is concerned.

A third approach inspired by the kinds of problems above may be called the neoliberal approach to politics favored by public choice theorists such as James Buchanan and Gordon Tullock (1965). Against elite theories, they contend that elites and their allies will tend to expand the powers of government and bureaucracy for their own interests and that this expansion will occur at the expense of a largely inattentive public. They argue against the interest group pluralist theorists that the problem of participation occurs within interest groups more or less as powerfully as among the citizenry at large. Only some interest groups are likely to succeed in organizing to influence the government. Hence only some interest groups will succeed in influencing government and they will do so largely for their own benefit while spreading the costs to others.

Neoliberals argue that any way of organizing a large and powerful democratic state is likely to produce serious inefficiencies. They infer that one ought to transfer many of the current functions of the state to the economic market and limit the state to the enforcement of basic property rights and liberties. These can be more easily understood and brought under the control of ordinary citizens.

But the neoliberal account of democracy must answer to two large worries. First, citizens in modern societies have more ambitious conceptions of social justice and the common good than are realizable by the minimal state. The neoliberal account thus implies a very serious curtailment of democracy of its own. More evidence is needed to support the contention that these aspirations cannot be achieved by the modern state. Second, the neoliberal approach ignores the problem of large private concentrations of wealth and power that are capable of pushing small states around for their own benefit and imposing their wills on populations without their consent. The assumptions that lead neoliberals to be skeptical about the large modern state imply equally disturbing problems for a neoliberal society.

A considerable amount of the literature in political science and the economic theory of the state are grounded in the assumption that individuals act primarily and perhaps even exclusively in their self-interest narrowly construed. The problem of participation and the accounts of the democratic process described above are in large part dependent on this assumption. While these ideas have generated interesting results and have become ever more sophisticated, there has been a growing chorus of opponents. Against the self-interest axiom, defenders of deliberative democracy and others claim that citizens are capable of being motivated by a concern for the common good and justice. They also claim, with Mill and Rousseau, that such concerns are not merely given prior to politics but that they can evolve and improve through the process of discussion and debate in politics. They assert that much debate and discussion in politics would not be intelligible were it not for the fact that citizens are willing to engage in open-minded discussion with those who have distinct morally informed points of view. Empirical evidence suggests that individuals are motivated by moral considerations in politics in addition to their interests. Accordingly, many propose that democratic institutions be designed to support the inclination to engage in moral and open minded discussion with others.

Once we approach the idea of citizenship from a moral point of view and we recognize the importance of a division of labor, the question arises of what the appropriate role for a citizen in a democracy is. If we think that citizens are too often uninformed, we should ask two questions. What ought citizens have knowledge about in order to fulfill their role? What standards ought citizens' beliefs live up to in order to be adequately supported? Some have proposed that citizens know about their particular sectors of society and not others. We have seen that this view has a number of difficulties. Christiano (1996) proposes, along with others, that citizens must think about what ends the society ought to aim at and leave the question of how to achieve those aims to experts. This kind of view needs to answer to the problem of how to ensure that politicians, administrators, and experts actually do attempt to realize the aims set by citizens. And it must show how institutions can be designed so as to establish the division of labor while preserving equality among citizens. But if citizens genuinely do choose the aims and others faithfully pursue the means to achieving those aims, then citizens are in the driver's seat in society.

It is hard to see how citizens can satisfy any even moderate standards for beliefs about how best to achieve their political aims. Knowledge of means requires an immense amount of social science and knowledge of particular facts. For citizens to have this kind of knowledge generally would require that we abandon the division of labor in society. On the other hand, citizens do have first-hand and daily experience with thinking about the values and aims they pursue. This gives them a chance to satisfy standards of belief regarding what the best aims are. Still the view is not defensible without a compelling institutional answer to the question of how to ensure that others are genuinely pursuing the means to achieve the aims specified by citizens.

10. *Legislative Institutions*

A number of debates have centered on the question of what kinds of legislative institution are best for a democratic society. What choice we make here will

depend heavily on the kind of underlying ethical justification of democracy, our conception of citizenship, and our empirical understanding of political institutions and how they function. The most basic types of formal political representation available are single-member district representation, proportional representation, and group representation. In addition, many societies have opted for multicameral legislative institutions. In some cases, combinations of the above forms have been tried.

Single-member district representation returns single representatives of geographically defined areas containing roughly equal populations to the legislature and is present most prominently in the USA and the UK. The most common form of proportional representation is party-list proportional representation. In a simple form of such a scheme, a number of parties compete for election to a legislature that is not divided into geographical districts. Parties acquire seats in the legislature as a proportion of the total number of votes they receive in the population as a whole. Group representation occurs when the society is divided into nongeographically defined groups such as ethnic or linguistic groups or even functional groups such as workers, farmers, and capitalists and returns representatives to a legislature from each of them.

Many have argued in favor of single member district legislation on the grounds that it has appeared to them to lead to more stable government than other forms of representation. The thought is that proportional representation tends to fragment the citizenry into opposing homogeneous camps that adhere rigidly to their party lines and that are continually vying for control over the government. Since there are many parties and they are unwilling to compromise with each other, governments formed from coalitions of parties tend to fall apart rather quickly. The post-War experience of governments in Italy appears to confirm this hypothesis. Single-member district representation, in contrast, is said to enhance the stability of governments by virtue of its favoring a two-party system of government. Each election cycle then determines which party is to stay in power for some length of time.

Charles Beitz (1989) argues that single-member district representation encourages moderation in party programs offered for citizens to consider. This results from the tendency of this kind of representation towards two-party systems. In a two-party system with majority rule, it is argued, each party must appeal to the median voter in the political spectrum. Hence they must moderate their programs to appeal to the median voter. Furthermore, they encourage compromise among groups since they must try to appeal to a lot of other groups in order to become part of one of the two leading parties. These tendencies encourage moderation and compromise in citizens to the extent that political parties, and interest groups, hold these qualities up as necessary to functioning well in a democracy.

In criticism, advocates of proportional and group representation have argued that single-member district representation tends to muffle the voices and ignore the interests of minority groups in the society. Minority interests and views tend to be articulated in background negotiations and in ways that muffle their distinctiveness. Furthermore, representatives of minority interests and views often have a difficult time becoming elected at all in single-member district systems, so it has been charged that minority views and interests are often systematically underrepresented. Sometimes these problems are dealt with by redrawing the boundaries of districts in a way that ensures greater minority representation. The efforts are invariably controversial since there is considerable disagreement about the criteria for apportionment. In proportional representation, by contrast, representatives of different groups are seated in the legislature in proportion to citizens' choices. Minorities need not make their demands conform to the basic dichotomy of views and interests that characterize single-member district systems so their views are more articulated and distinctive as well as better represented.

Another criticism of single-member district representation is that it encourages parties to pursue dubious electoral campaign strategies. The need to appeal to a large, diverse, and somewhat amorphous sector of the population can very often be best met by using ambiguous, vague, and often irrelevant appeals to the citizens. Thus, instead of encouraging reasonable compromise, the scheme tends to support tendencies towards ignorance, superficiality, and fatuousness in political campaigns and in the citizenry. It encourages political leaders to take care of the real issues of politics in back rooms while they appeal to citizens by means of smoke and mirrors. Of course, those who agree in the main with the elitist-type theories will see nothing wrong in this; indeed, they may well champion this effect. Proportional representation requires that parties be relatively clear and upfront about their proposals, so those who believe that democracy is ethically grounded in the appeal to equality tend to favor proportional representation.

Advocates of group representation such as Iris Marion Young (1993) have argued that some historically disenfranchised groups may still not do very well under proportional representation. They may not be able to organize and articulate their views as easily as other groups. Also, minority groups can still be systematically defeated in the legislature and their interests may be consistently set back even if they do have some representation. For these groups, some have argued that the only way to protect their interests is legally to ensure that they have adequate and even disproportionate representation.

One worry about group representation is that it tends to freeze some aspects of the agenda that might be better left to the choice of citizens. For instance, consider a population that is divided into linguistic

groups for a long time, and suppose that only some citizens continue to think of linguistic conflict as important. In the circumstances, such schemes may tend to be biased in an arbitrary way that favors the views or interests of those who do think of linguistic conflict as important.

See also: Democracy; Democracy, History of; Democratic Theory; Democratic Transitions; Law and Democracy; Participation: Political; Political Discourse; Political Representation

Bibliography

Beitz C 1989 *Political Equality: An Essay on Democratic Theory*. Princeton University Press, Princeton, NJ

Buchanan J, Tullock G 1965 *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. University of Michigan Press, Ann Arbor, MI

Christiano T 1996 *The Rule of the Many: Fundamental Issues in Democratic Theory*. Westview Press, Boulder, CO

Cohen J 1998 Deliberation and democratic legitimacy. In: Bohman J, Rehg W (eds.) *Deliberative Democracy*. MIT Press, Cambridge, MA

Dahl R 1959 *A Preface to Democratic Theory*. University of Chicago Press, Chicago

Downs A 1957 *An Economic Theory of Democracy*. Harper and Row, New York

Gould C 1988 *Rethinking Democracy: Freedom and Social Cooperation in Politics, Economics and Society*. Cambridge University Press, New York

Hayek F 1960 *The Constitution of Liberty*. University of Chicago Press, Chicago

Hobbes T 1968 MacPherson C B (ed.) *Leviathan* Penguin-Harmondsworth, UK

Madison J 1788 The Federalist no 10. In: Hamilton A, Jay J, Madison J (eds.) *The Federalist Papers*. Random House, New York, pp. 59–62

Mill J S 1991 *Considerations on Representative Government*. Prometheus, Buffalo, NY

Plato 1974 *The Republic*, 2nd edn. Revised translation by Lee D. Penguin, Harmondsworth, UK

Rousseau J-J 1967 *The Social Contract and Discourses*. Trans. Cole G H D. Dent, London

Schumpeter J 1956 *Capitalism, Socialism and Democracy*. Harper and Row, New York

Singer P 1973 *Democracy and Disobedience*. Oxford University Press, Oxford, UK

Young I M 1993 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

T. D. Christiano

Democratic Party

The Democratic party in the United States claims Thomas Jefferson and Andrew Jackson as its founders. It was the first modern political party, and it remains the oldest political party in continuous existence in the world. For more than two centuries, the Democratic party and its predecessor, the Democratic-Republican party, have been the mainstays in two-party competition that has pitted them sequentially against the Federalists, the Whigs, and the Republicans.

1. Democratic-Republican Prelude

The distant origins of the Democratic party can be found in the governmental faction led by Secretary of State Jefferson at the outset of the new national government established by the 1787 Constitution. Serving under President George Washington, who persistently decried the emerging partisanship in his administration, Jefferson found himself consistently at odds with Secretary of the Treasury Alexander Hamilton.

1.1 Jeffersonians vs. Hamiltonians

Their rivalry was rooted in personal ambition, policy, region, constituency representation, and ideology. Each aspired to inherit Washington's mantle of national leadership. On behalf of his rural, agrarian, common constituents, Jefferson opposed Hamilton's policy initiatives that favored the urban, commercial, affluent interests in society. Jefferson's understanding of the proper role of the national government was much more limited than Hamilton's expansive, energetic one.

More specifically, Jefferson and his followers tended to distrust executive and judicial power, even as the Hamiltonians, or Federalists, as they called themselves, embraced it. Rather, the Jeffersonians exalted the legislature, with its clearer popular links, within the separation of power structure of the federal government, in contrast with the more elitist Federalists. As such, their pioneering partisan institution was the legislative caucus, a gathering of the congressional representatives loyal to Jefferson. The caucus plotted legislative strategy and assumed responsibility for presidential nominations.

When Washington, partisan disclaimers notwithstanding, generally embraced Hamilton's policy initiatives, Jefferson extended his partisan opposition beyond the legislative caucus to the electoral arena in an effort to assume control of the national government. His organizational vehicle was the Democratic-Republican party. Meanwhile, this initiative encouraged Hamilton similarly to organize his Federalist faction within the government into a political party.

1.2 Democratic-Republican Dominance (1800–24)

As the nineteenth century dawned, Jefferson's election to the presidency, accompanied by victories of supportive majorities in both houses of Congress, signaled

the ascendancy of the Democratic-Republican party, and the decline of its Federalist opposition. Under the Virginia Dynasty of Presidents Jefferson (1801–9), James Madison (1809–17), and James Monroe (1817–25), the Democratic-Republicans thoroughly dominated the political landscape.

1.3 Democratic-Republican Disarray (1824–8)

However, the looming presidential election of 1824 revealed serious divisions in the coalition of forces that comprised the Democratic-Republican party. The breakdown of consensus undermined the legitimacy of the legislative caucus' presidential nomination and led to a split that produced a new partisan alignment and the appearance of the Democratic party. One faction flocked to the banner of Andrew Jackson in his unsuccessful 1824 presidential campaign and the two succeeding ones in which he prevailed. This Jackson faction became known as the Democrats, distinguishing themselves from Jackson's rivals, including John Quincy Adams and Henry Clay, whose reconstituted factions became the Whig party.

2. The Jacksonian Democracy

The Democratic party, while clearly rooted in the old Democratic-Republican party, nonetheless demonstrated noteworthy shades of difference pertaining to region, class, religion, and ideology. It had strong appeal in the new southern and western states admitted to the Union. The Democrats also attracted support from urban workers whose heightened presence reflected increasing industrialization in American society. This constituency brought into the party coalition large numbers of Catholic immigrants. The Democrats more enthusiastically promoted the value of popular participation in the political arena. In 1832, they supplanted the discredited caucus with the presidential nominating convention. Following Jackson's lead, they tended to be more comfortable with the exercise of executive prerogative, linking the presidency with the public.

2.1 Democrats vs. Whigs (1828–56)

In this second distinctive era of party competition, lasting over a quarter-century, the Democrats were clearly dominant on the national scene. Their Whig opponents united around their common opposition to the person and policies of Jackson, but they were divided otherwise. In addition to Jackson (1828, 1832), the Democrats also elected Martin Van Buren (1836), James K. Polk (1844), Franklin Pierce (1852), and James Buchanan (1856) to the presidency for single terms. On two occasions in the 1840s, they lost presidential elections when the Whigs nominated military heroes: William Henry Harrison (1840), and Zachary Taylor (1848). Unfortunately for the Whigs, both standard-bearers died before completing a single term, and the Democrats reclaimed the presidency in the ensuing elections.

For the most part, Democratic control of Congress was also the norm. On only one occasion, 1841–3, did the Democrats lack majorities in both houses. Twice they lost a single chamber: the Senate, 1843–5; and the House, 1847–9. Otherwise, from the late 1820s until the mid-1850s, the Democrats were clearly in charge on Capitol Hill. However, the intensifying slavery controversy and rising sectionalism proved divisive and debilitating for both major political parties in the decade of the 1850s. Indeed, the unstable Whig coalition disintegrated, and a new party, the Republicans, arose to challenge the fragmented Democrats for national dominance, winning control of the House in 1854 and 1858.

3. Democrats vs. Republicans I: Democrats in the Minority (1860–96)

With its Northern and Southern wings competing electorally, the Democrats were trounced in the 1860 elections, which produced a Republican president, Abraham Lincoln, Republican congressional majorities, and a new, third era of party competition, in which the Democrats found themselves in an unaccustomed position as the minority party. During the Civil War, the Democrats floundered following the departure of their traditionally strong Southern wing to secession. They held on as the major party alternative to the now-dominant Republicans with their support in the urban strongholds of the Northeast. After the war, the return of the solidly Democratic Southern states to the Union revived party fortunes, but the Democrats continued to suffer nationally from the stigmas of their longtime accommodation of slavery and acquiescence to secession.

For more than three decades following the outbreak of the Civil War, the Democrats typically lost presidential elections, albeit by narrow margins. Only Grover Cleveland, in 1884 and 1892, was able to break the Republican lock on the White House. In turn, only two Congresses in these decades featured Democratic majorities in both chambers: the 46th (1879–81), and the 53rd (1893–5). The Democrats were much more competitive in the lower house, winning control on six additional occasions: 1874, 1876, 1882, 1884, 1886, and 1890.

3.1 Democratic Factions

During this era, the Democrats became even more fragmented. Reinforcing their traditional sectional

division, with its attendant urban–rural dimension, they separated into persistent dueling factions that featured Protestants vs. Catholics, protectionists vs. free traders, wets and drys, populists and progressives vs. reactionaries and states-righters, and hard money and easy money advocates.

4. *Democrats v. Republicans II: Democratic Downturn (1896–1932)*

The critical election of 1896 marked a severe downturn in Democratic party fortunes. Their presidential nominee, William Jennings Bryan, represented the ascendant Southern and Western, rural, Protestant, populist, easy money, free trade, dry factions within the party. He went down to a crushing defeat at the hands of Republican William McKinley. Renominated in 1900 and 1908, Bryan continued to lead the party to electoral disasters, as the new pattern of national electoral competition solidified. In this fourth party system, lasting until 1932, the Democrats maintained their base in the South, and did relatively well in the West, but were usually overwhelmed by the Republicans in the more populous Northeast and Midwest, despite their enduring urban strongholds there.

Their entrenched minority status was interrupted in the twentieth century's second decade. In 1910, the Democrats won a majority in the House. Two years later, division within the dominant Republican party allowed nominee Woodrow Wilson to capture the presidency, and the Democrats to control both houses of Congress for the first time in two decades. Wilson's narrow 1916 re-election was the first by an incumbent Democrat since Jackson's in 1832, but the Democrats relinquished both congressional chambers in the midterm elections of 1918. The 'return to normalcy' was completed by the loss of the White House to the resurgent Republicans in 1920. The Democrats suffered convincing electoral defeats in the remaining national elections of that decade.

5. *Democrats vs. Republicans III: Democratic Resurgence (1932–52)*

However, the onset of the Great Depression in 1929 signaled a revival in party fortunes, catapulting the Democrats back to majority status nationwide. Franklin D. Roosevelt, their successful presidential nominee in 1932, played the leading role in this electoral realignment, justifying his inclusion with Jefferson and Jackson in the top echelon of party heroes. Roosevelt assembled a formidable electoral coalition under the party umbrella. It embraced white Southerners, urban workers, farmers, racial and ethnic minorities, and ideological liberals, who united behind the activist governmental response to economic and social problems embodied in Roosevelt's New Deal. Roosevelt effectively reconciled Hamilton's energetic founding vision for the Federalists with the Democrats' traditional, antielitist identity as the party of the people.

This coalition elected Roosevelt president an unprecedented four times, and kept his successor, Harry S. Truman, in the White House for another full term, encompassing an epochal two decades that featured not only the Depression, but also World War II and the onset of the Cold War. In addition, the Democrats took control of the House of Representatives in 1930 and the Senate in 1932. They held both through 1946, and recaptured the Congress in 1948. Thus, the Roosevelt coalition, mobilized in support of his New Deal policies, dominated this fifth distinctive era of party competition,

6. *Democrats vs. Republicans IV: Fluctuating Fortunes (1952–)*

In the five decades since Truman's 1948 triumph, the national electoral fortunes of the Democrats have fluctuated dramatically and generally deteriorated. They lost the White House and Congress as well in 1952, when the Republicans nominated General Dwight D. Eisenhower, a World War II hero. However, they recaptured control of Congress in 1954. Having done so, they controlled the House for an unprecedented 40 consecutive years and the Senate for 26.

This sustained Democratic dominance of Congress masked a much more vulnerable position in presidential elections. After two terms for Eisenhower, the Democrats regained the presidency under John Kennedy in 1960 and retained it for his successor Lyndon Johnson in 1964. They surrendered the White House in 1968, when Republican Richard Nixon won his first of two terms. Scandals associated with the Nixon presidency enabled Democrat Jimmy Carter to defeat Nixon's replacement, Gerald R. Ford, in 1976. However, Carter met defeat in his 1980 reelection bid against Ronald Reagan. That election also featured the Democrats' loss of the Senate for the first time in over a quarter-century.

Throughout Reagan's two terms, the Democrats held on to their majority in the House, and they took back the Senate in 1986. Nevertheless, their weakness in presidential contests was unmistakable. In 1988, Reagan's vice president, George Bush, won the presidency, marking the first time since 1928 that the Democrats lost three consecutive presidential elections.

Uniting behind nominee Bill Clinton in 1992, the Democrats finally recaptured the White House; but two years later, they lost both houses of Congress for

the first time in four decades. In 1996, Clinton won reelection, the first Democrat since Roosevelt to do so. In turn, the Republicans have maintained their control of Congress at least through 2000, marking the longest stretch the Democrats have constituted the minority since the 1920s.

6.1 Erosion of the New Deal Coalition

The past five decades of volatility in Democratic party fortunes reveal serious erosion in the New Deal coalition. Most notably, the Democrats have lost the allegiance of the white Southerners who had been a party bulwark since the days of Jefferson and Jackson. When the national Democrats embraced the cause of civil rights in the late 1940s, they set the stage for a regional realignment that was clearly underway by the mid-1960s and has continued through the century's end.

In addition, even as their social and economic reforms became widely accepted, the Democrats suffered electorally from rising public distrust of the national government and its capacity to solve social problems. Moreover, the advent of the Cold War led to criticism of the Democrats as 'soft' on communism, undermining their voter support on foreign policy issues. Intraparty conflict over the Vietnam War and the attendant cultural divisions it spawned further weakened the Democrats. Further, both major parties struggle with the disinclination of contemporary voters to develop enduring partisan attachments.

7. Organization and Leadership

The Democratic Party is organized according to the federal principle, in keeping with the constitutional structure of government under which it developed. Thus, at every level of government, national, state, and local, there is a corresponding level of party organization. Traditionally, decentralization prevailed within the party organization, with the state and local parties maintaining noteworthy autonomy. Party reforms beginning in the late 1960s ushered in an era of unprecedented nationalization, as well as an enhancement of the role of primaries in presidential nominations. The organization of the Democratic party, at any level, typically features several distinct institutions: convention, committee, chair, and headquarters staff.

Traditionally, the party's presidential nominee is designated its titular leader. When that nominee occupies the presidency, party leadership is indisputable. Otherwise, it may also be claimed by the congressional chieftains, the national party chair, and, occasionally, incumbent governors of major states.

See also: Congress: United States; Executive Branch, Government; Party Identification; Party Responsibility; Political History (History of Politics); Political Machines; Political Parties; Political Parties, History of; Presidency: United States; Republican Party; Third Parties: United States; Whig

Bibliography

Anderson K 1979 *Creation of a Democratic Majority, 1928–1936*. University of Chicago Press, Chicago

Cunningham N E 1957 *The Jeffersonian Republicans: The Formation of Party Organization, 1789–1801*. University of North Carolina Press, Chapel Hill, NC

Cunningham N E 1963 *The Jeffersonian Republicans in Power: Party Operations, 1801–1809*. University of North Carolina Press, Chapel Hill, NC

Goldman R M 1986 *Dilemma and Destiny: the Democratic Party in America*. Madison Books, Lantham, MD

Kurian G T, Schultz J D (eds.) 1997 *The Encyclopedia of the Democratic Party*, 2 vols. M E Sharpe Reference, Armonk, NY

McCormick R P 1966 *The Second American Party System: Party Formation in the Jacksonian Era*. University of North Carolina Press, Chapel Hill, NC

Parmet H S 1976 *The Democrats: The Years After FDR*. Macmillan, New York

Radosh R 1996 *Divided They Fell: The Demise of the Democratic Party, 1964–1996*. Free Press, New York

Shafer B E 1983 *Quiet Revolution: The Struggle for the Democratic Party and the Shaping of Post-Reform Politics*. Russell Sage Foundation, New York

Silbey J H 1977 *A Respectable Minority: The Democratic Party During the Civil War Era, 1860–1868*. Norton, New York

Sundquist J L 1983 *Dynamics of the American Party System: Alignment and Realignment of Political Parties in the United States*, rev. edn. Brookings Institution, Washington, DC

Ware A 1985 *The Breakdown of Democratic Party Organization, 1940–1980*. Oxford University Press, Oxford, UK

H. F. Bass, Jr.

Democratic Theory

Democratic Theory is a body of normative and empirical claims focused on at least five related questions: (a) Whose preferences count? (b) What kinds of preferences or opinions count? (c) What decision rule is employed to aggregate those preferences? (d) Through what kinds of institutions are those results represented? (e) How can the democratic decision process avoid bad or unacceptable outcomes such as 'tyranny of the majority'?

1. Who is included?

Early democratic institutions, such as those in ancient Athens, enfranchised only a tiny portion of the population. But the 2500 year history of democracy can be viewed as one of progressive inclusion—through the gradual elimination of restrictions based

on narrow definitions of citizenship, on property and/or educational qualifications, on race and ethnicity, and on gender. Most modern developed democracies approach universal enfranchisement of adult citizens (Dahl 1989). Of course, many citizens theoretically entitled to participate do not actually do so. Except for a few countries such as Australia and Brazil that have successfully implemented compulsory voting, most democracies leave participation to individual decision with variations in terms of how much of the burden of voter registration falls on the individual. Countries such as the USA or Switzerland that are distinctive for offering many opportunities for public consultation routinely have low turnout rates. In the United States, only about one-half of the citizen adult population actually votes, even in the most important elections, such as for the US Presidency.

Much of the modern dialogue about adapting democracy to the large-scale nation-state began with the famous debate over the American Constitution between Federalists and Anti-Federalists. The American Founders did not even use the term 'democracy' to apply to the 'Republic' that they wished to establish. The term 'democracy' was reserved, for example, by James Madison (author of many of the key Federalist papers and one of the chief architects of the US Constitution) for small-scale city-states where everyone could gather together. The danger in direct democracy, Madison and the Founders thought, was that passions or interests might motivate 'factions' endangering some people's rights (Madison et al. 1788).

2. *What is Public Opinion?*

The American Founders wanted institutions that would give expression not just to any public views, but to those opinions that had been 'refined' or 'filtered' so as to produce 'the deliberative sense of the community.' Representatives, they believed, serve to 'refine and enlarge the public views by passing them through a chosen body of citizens' as Madison argued in *Federalist* No. 10. 'The public voice' pronounced by representatives under such a regulation 'will be more consonant to the public good than if pronounced by the people themselves convened for the purpose.' Small deliberative bodies, such as the US Senate or a Constitutional Convention, allow representatives to come to a better determination of the public good than one would get just by bringing the people together and asking them. There is a difference, in other words, between the deliberative or thoughtful public opinion one can find in representative institutions, at least at their best, and the uninformed and unreflective preferences commonly found in the mass public.

A central problem in democratic theory is how to reconcile the aspiration for thoughtful and informed preferences with another key aspiration—counting everyone's preferences equally. Deliberative bodies may represent highly informed and competent preferences, but those preferences are often shared only by an elite. Direct consultation of mass preferences will typically involve counting uninformed preferences, often simply reflecting the public's impressions of sound bites and headlines. Hence the hard choice between politically equal but unreflective mass preferences and politically unequal but relatively more reflective elite views.

This basic dilemma is reinforced by the tendency of ordinary citizens in the large-scale nation-state to have little incentive to become well informed. Each voter can understand that his or her individual vote or individual opinion is only one out of millions and is, as a result, unlikely to make much difference to the outcome of an election or to public policy. As a result, individual citizens have little reason to become well informed about complex political or policy issues. Anthony Downs coined the term 'rational ignorance' for this situation in which voters find themselves (Downs 1957). A host of empirical studies confirm that the mass public in most modern societies is not well informed about politics (see Delli Carpini and Keeter 1996, for an overview of both American and comparative data).

3. *The Design of Institutions*

We seem, in other words, to face a forced choice between elected elites and uninformed masses. But there are institutional experiments with deliberation among representative microcosms of the mass public. These experiments, ranging from so-called 'citizens juries' to Deliberative Opinion Polls, take random samples of the public and subject them to situations where they are effectively motivated to get good information, hear balanced accounts of competing arguments, and come to a considered judgment. These experiments show that ordinary citizens are capable of becoming informed and dealing with complex policy matters. And it is possible to get an input to policy-making that is representative of the mass public while also embodying deliberation (Fishkin 1995).

Once opinions are solicited by some democratic process, how should they be aggregated? There is a long-standing claim for majority rule as a procedure that counts everyone's views equally. Modern work in rational choice theory has shown that majority rule is the only decision rule that satisfies a number of appealing normative criteria (May 1952). But majority rule, like other decision rules, is subject to 'instability'—to cycles, violating transitivity (the requirement that if A is preferred to B and if B is preferred to C, then A should be preferred to C). This problem is part of a more general account of problems in social choice theory developed in the modern era by Kenneth Arrow in Arrow's so-called 'impossibility theorem' (Arrow 1951, 1963). For our purposes, it is worth noting that the conundrums of social choice theory,

including the violations of transitivity, presume 'universal domain'—they presume that all logically possible preference orderings may occur. One way out is to limit the range of preference orderings that occur in a given decision-making process. If preferences are, for example, organized according to some overarching dimension so that they are 'single-peaked' (Black 1958) then violations of transitivity will not occur. Some political theorists have suggested that if citizens deliberate, they will come to some shared understanding of the questions at issue and achieve a structure for their preferences that tends to satisfy single peakedness (Miller 1992). To the extent this turns out to be the case, then deliberative preferences could avoid cycles that undermine the claim to rationality for democratic decisions.

Most democratic decisions in the large-scale nation-state do not conform to the simplified situation of citizens voting directly on policy alternatives. Of course, they do fit that picture in initiatives and referendums. And citizens do elect representatives through a variety of electoral systems. However, these systems vary greatly in how well they translate majority preferences into seats in legislatures and into public policies. Robert Dahl, in a classic work (Dahl 1956), described two competing pictures of democracy—'Madisonian' and 'Populistic.' The former offers significant impediments to majority rule (in the interests of avoiding tyranny of the majority). The latter offers a high degree of correspondence between majority preferences and public politics. Most systems, however, are somewhere in between. For example, the United States, Dahl argued, was a 'hybrid' of the two that actually was best understood as operating not on 'majority rule' but on 'minorities rule'—a system in which intense minorities tend to get their way. People who feel most strongly tend to participate actively to make sure their voice is heard. People who do not care tend not to participate.

4. *Tyranny of the Majority?*

A remaining central problem of democratic theory is how to avoid 'tyranny of the majority.' A decision may count all or most people's preferences, it may be supported by the public's considered judgments, it may be aggregated by an appropriate decision rule such as majority rule, and it may, nevertheless, conflict with justice or impose unacceptable consequences on some portion of the population. In other words, the people may, democratically, decide to do bad things. For 2500 years, the image of mob rule in Ancient Athens did much to discredit democracy (Roberts 1994). The fact that the Athenians had killed Socrates worried the American Founders.

Attempts to address the problem of majority tyranny usually rely on answers to the questions addressed above: Whose preferences count? What kinds of preferences are considered? What decision rule is employed? What is the design of democratic institutions? First, the spread of the franchise to most or all adult citizens has been an important factor in making sure that the interests of those citizens are considered. The spread of voting rights across racial, ethnic, religious, gender, and class divisions has proven to be an important factor in creating institutional incentives for addressing the problems of those who were previously disenfranchised. But such a solution is imperfect. A group can have its votes counted, but outvoted, its interests considered but neglected—or even despised.

Hence, the second issue bears on majority tyranny: what kinds of preferences do people bring to the democratic process? The fear of majority tyranny has long focused on 'passions' or 'interests' that might motivate 'factions' adverse to the rights of some members of the community, to use terms that James Madison made famous in the *Federalist Papers*, especially No. 10. One of Madison's strategies for insulating democracy from factions is to 'filter' preferences through deliberation that 'refines and enlarges' the public's views. The idea is to get the 'cool and deliberate sense of the community' rather than the kinds of passions that might motivate faction. Hence democratic theory has long recognized that the kinds of preferences brought to the democratic process may influence whether majority tyranny results.

Third, some theorists have looked to the decision rule itself to limit majority tyranny. Supermajority requirements, specifying a percentage greater than 50 percent plus one, but usually less than unanimity, can make it harder to get new measures passed. Such decision rules bias decision making toward the status quo, as it becomes harder to get changes passed. Hence if there arc bad outcomes that result from inaction, from the failure to pass new measures, then those bad outcomes are made more, not less, likely by such a decision rule. Hence supermajority requirements are an imperfect protection against the system producing unacceptable outcomes.

Fourth, there are institutional designs that can protect, to some degree, against tyranny of the majority. One strategy is to specify a list of rights that are immune to majority decision, and that are protected by judicial review. However, the bad outcomes that might be considered or adopted by democratic decision making can never be fully anticipated. And the empowerment of judges through judicial review, when it is broad enough to provide protection against tyranny of the majority, also threatens democracy in another way. It seems undemocratic for the judiciary to overturn decisions of the public expressed through their representative institutions. Hence, judicial review that can set aside democratically passed legislation, or a bill of rights that takes certain issues off the democratic agenda, protects democracy from itself, by limiting the purview of democratic institutions.

Democratic theory provides us with conflicting visions rather than some uniquely authoritative answer to the questions posed here. Questions about democracy remain unsettled. But unlike any other period in its history, the authority of the basic democratic idea is virtually unchallenged in the modern era. Democracy (in some sense) triumphs even as disagreements proliferate as to what it is, or might be.

See also: Citizen Participation; Citizenship, Historical Development of; Citizenship: Political; Civic Culture; Democracy, History of; Democracy: Normative Theory; Electoral Systems; Public Opinion: Political Aspects; Voting, Sociology of

Bibliography

Arrow K J 1951, 1963 *Social Choice and Individual Values, 2nd edn*. Yale University Press, New Haven, CT

Black D 1958 *The Theory of Committees and Elections*. Cambridge University Press, Cambridge, UK

Dahl R A 1956 *A Preface to Democratic Theory*. University of Chicago Press, Chicago

Dahl R A 1989 *Democracy and Its Critics*. Yale University Press, New Haven, CT

Delli Carpini M X, Keeter S 1996 *What Americans Know About Politics and Why it Matters*. Yale University Press, New Haven, CT

Downs A 1957 *An Economic Theory of Democracy*. Harper and Row, New York

Fishkin J S 1995 *The Voice of the People: Public Opinion and Democracy*. Yale University Press, New Haven, CT

Madison J, Hamilton A, Jay J 1788, 1987 *The Federalist Papers*. Penguin Books, Harmondsworth, UK

May K O 1952 A set of independent necessary and sufficient conditions for majority rule. *Econometrica* **20**: 680–4

Miller D 1992 Deliberative democracy and social choice. *Political Studies* **40**: 54–67

Roberts J 1994 *Athens on Trial*. Princeton University Press, Princeton, NJ

J. S. Fishkin

Democratic Transitions

The study of democratic transitions focuses on the critical step in the history of democracy when a country passes a threshold marked by the introduction of competitive elections with mass suffrage for the main political offices in the land. Though a concern of thinkers since the late eighteenth century, it blossomed as a distinct area of scholarly research in the context of the wave of democratic transitions that swept through every corner of the world in the last quarter of the twentieth century.

1. The Subject Matter

Research on democratic transitions is part of the broader field of democratic theory that gains its distinctiveness from a sharply defined focus on elections. Indeed, the status of democratic transitions as a distinctive field of research is given by an undeniably Schumpeterian approach to democracy, which emphasizes the procedures that regulate the access to power. This delimitation of the subject matter did little to spur interest at the time university-based research was expanding dramatically in the 1960s and 1970s. Not only did the realities of world politics appear to devalue this line of research, but the Schumpeterian conception of democracy was widely out of favor. Even though some landmark studies on democratic transitions were published as early as 1960 (Lipset 1960, Rustow 1970), interest in democratic transitions took a back seat to other, more pressing or more valued concerns.

The status of research on democratic transitions, however, changed quite considerably thereafter. First and most important, the wave of democratization beginning in 1974 made the subject matter immediately relevant. In addition, the influential writings of Robert Dahl (1971) helped foster widespread acceptance in the social sciences of the erstwhile disparaged Schumpeterian view of democracy. Finally, the seminal work of Guillermo O'Donnell and Philippe Schmitter (1986) did much to set the initial terms of the debate and give a strong impetus to research on the topic. With the boom in research on democratic transitions in the 1980s and 1990s, by the turn of the century research on democratic transitions had attained the status of an established field.

2. Research and Findings

The focus on a sharply delimited subject matter has a definite advantage, in that it has allowed research to concentrate on a fairly clear question: why have some countries had democratic transitions while others have not? Thus, an impressive basis of knowledge was generated quite rapidly through a succession of studies. Following in the wake of pioneering work on the transitions in southern Europe and Latin America in the 1970s and early 1980s (O'Donnell et al. 1986), major cross-regional analyses were conducted comparing Latin America to East and South East Asia (Haggard and Kaufman 1995), and southern Europe and Latin America to Eastern Europe and the former Soviet Union (Linz and Stepan 1996, see also Huntington 1991). Excellent region-based studies were produced, focusing on Africa (Bratton and van de Walle 1997), Eastern Europe (Beyme 1996, Offe 1997, Bunce 1999), as well as the three major regions of the developing world (Diamond et al. 1989a, 1989b, 1989c). Finally, impressive efforts were made to put the transitions of the last quarter of the twentieth

century in historical perspective, through cross-regional analyses of Europe and Latin America ranging across the nineteenth and twentieth centuries (Rueschemeyer et al. 1992, Collier 1999).

The richness of this literature is undeniable. Indeed, it offers a wealth of ideas on the causes of transitions, a great amount of nuanced data on very complex processes, and some fruitful comparative analyses that have generated a number of important and surprising findings. This literature has demonstrated that, contrary to the longstanding conventional view, level of economic development is not a good predictor of democratic transitions. It has also shown that democratic transitions do not occur through a single process but rather through multiple paths defined by factors such as the power and strategies of elites and masses and the source of impetus for political reform. Moreover, an important degree of cumulation of knowledge has been achieved through efforts to codify these distinct paths of democratic transition. A closely related finding involves the growing evidence that the path toward democracy that a country follows is strongly influenced by its type of old, nondemocratic regime. Thus, depending on the type of old regime, pacts may be a necessary condition for a successful transition to democracy, and nationalist movements may weaken, rather than enhance, the prospects of democracy. Finally, this literature has shown that issues of regime change are closely linked with those of the state, conceived in Weberian terms. In this regard, a key finding is the 'no state, no democracy' principle, which holds that processes of regime change that lead to state decay or state collapse reduce the prospects of democracy.

3. *Explanatory Variables and Theoretical Integration*

The considerable accomplishments of this body of literature notwithstanding, the agenda of research on democratic transitions faces some considerable challenges. One such challenge concerns the need for greater theoretical integration. The evolution of the literature on democratic transitions has been characterized by the frequent introduction of new causal factors considered critical to an explanation of why democratic transitions occur. These new explanatory variables sometimes reflect the experience of new cases of transitions to democracy, which have brought to light factors that had not seemed important in the cases until then considered. In other instances, the focus on new variables has been driven more by an effort to rescue insights from older bodies of literature. The downside of this theoretical fertility, however, is that the introduction of a large set of explanatory variables has given a somewhat disorderly quality to theoretical debates.

As challenging as the task of theoretical organization and integration is likely to be, it is facilitated somewhat because theoretical debates have evolved around a number of central axes. One main axis contrasts short-term factors and the choices made by actors (O'Donnell and Schmitter 1986, Przeworski 1991) to medium-term factors, such as the characteristics of the old regime (Linz and Stepan 1996), and long-term, more structural factors, such as the mode of production or the model of development (Rueschemeyer et al. 1992, Castells 1998). Another axis of debate contrasts elite-centered explanations (Dogan and Higley 1998) to mass-centered explanations, which focus either on class actors (Rueschemeyer et al. 1992, Collier 1999), social movements (Foweraker 1995, Chap. 5, Tarrow 1995), or ethnic groups (Offe 1997, Chap. 4). Yet another axis contrasts political to economic determinants of transitions (Haggard and Kaufman 1995, Przeworski and Limongi 1997). And one more critical axis of debate opposes domestic factors to international factors (Whitehead 1996), an axis along which one might also locate explanations centered on stateness and nationality issues that might be labeled as 'intermestic' (Linz and Stepan 1996). Thus, the further development of a theory of democratic transitions is likely to hinge on efforts to build upon this implicit organization of the debate and to explicitly formulate a framework that integrates and synthesizes these diverse explanatory factors.

4. *Data Collection and Comparative Analysis*

Another challenge concerns the collection of data and the conduct of comparative analysis. The literature on democratic transitions has evolved in a way that has given primacy to selective rather than encompassing comparisons. The reason for this strategy is obvious, in that the comparison of a small number of cases can provide a sufficient basis for introducing new ideas into the debate and for doing so rapidly. Hence, this strategy has been well suited for the initial stages of theorizing. But this strategy has also had its downside. The key problem is that data has not been collected on all the explanatory variables across all the relevant cases. Indeed, though this literature has generated a great amount of nuanced data, researchers have not always been systematic enough in terms of gathering the same data for all the cases they analyze, let alone all the cases that are discussed in this literature as a whole. As a consequence, the ability of researchers to test their theories systematically and draw strong conclusions has been somewhat constrained.

The ability of scholars to engage in systematic comparative analysis, however, is hampered by a serious practical limitation. A review of the existing literature on democratic transitions shows that there is

a very steep trade-off in the level of nuance of data and hence the explanatory arguments tested as one moves from the literature based on intensive but relatively narrow comparisons of a small set of cases to the statistical literature based on a large number of cases. Thus, concerning the task of data collection and comparative analysis, a multitrack approach is highly recommended.

One such track would consist of the collection of data that can be used for statistical studies. In this regard, the challenge is to collect data, on a large number of cases, on factors other than the standard economic and institutional ones which have long been the staple of statistical analyses (e.g., Przeworski and Limongi 1997). Indeed, the full benefits of statistical tools are unlikely to be felt in the debate on democratic transitions until data sets are generated with information on the kind of actors involved in the process of democratic transitions, the choices these actors make, and the sequence of events whereby democratic transitions unfold. This is a formidable task but one with important payoffs and thus well worth pursuing.

Another track would consist of the application of the research strategy which has contributed so much to this debate, the intensive analysis of a small to medium number of cases, to some relatively unexplored questions. Though some significant works offer a historical perspective on the democratic transitions that have been the focus of the debate, those occurring in the last quarter of the twentieth century, much remains to be learned by cross-time comparisons as well as by a reanalysis of older cases of transitions in light of new theories. In addition, though some thought has been given to the differences and similarities between the explanations currently used to account for democratic transitions and the earlier explanations of democratic breakdowns (O'Donnell 1973, Linz and Stepan 1978, Collier 1979), very little has been done to pursue this potentially illuminating comparison in a systematic manner. Finally, a whole new perspective that needs to be developed concerns the cases of failed transitions. Indeed, while scholars of democratic transitions have learned a lot by focusing on positive cases, where transitions to democracy did occur, an important question that remains to be fully answered is: why did many countries that saw the collapse of old regimes during the last quarter of the twentieth century have transitions that did not lead to democracy? The continued use of qualitative methods, thus, is likely to be highly rewarding.

5. *The Need for Disaggregation*

As the new century begins, scholars of democratic transitions have a lot to build upon but also a lot to do. Thus, it is absolutely essential to devise ways of making this research agenda manageable. Many different options in this regard are available. One option that bears highlighting is suggested by the very definition of a democratic transition offered at the outset of this article. This definition consists of three elements which countries may attain at different paces and in different orders: competitive elections, mass suffrage, and elections for all main political offices. Thus, even though democratic transitions are often portrayed as a single event, they may actually be studied as a series of events, a new perspective that has important implications. If research shows that countries follow different paths, characterized by different sequences of events, as Dahl (1971, pp. 4–6) has suggested, then the challenge facing scholars of democratic transitions might be disaggregated and hence broken down analytically and greatly facilitated. Indeed, if the challenge of explaining transitions could focus on the more manageable task of accounting for a number of influential patterns, the problem of theoretical integration would be simplified and data collection could be organized around a set of priorities. A fascinating but somewhat daunting agenda of research would thus be transformed into a fascinating and manageable agenda.

See also: Democracy; Democracy, History of; Eastern European Studies: Politics; Latin American Studies: Politics; Schumpeter, Joseph A (1883–1950); Soviet Studies: Politics

Bibliography

Beyme K von 1996 *Transition to Democracy in Eastern Europe*. St. Martin's Press, New York

Bratton M, van de Walle N 1997 *Democratic Experiments in Africa. Regime Transitions in Comparative Perspective*. Cambridge University Press, New York

Bunce V 1999 *Subversive Institutions. The Design and the Destruction of Socialism and the State*. Cambridge University Press, New York

Castells M 1998 The crisis of industrial statism and the collapse of the Soviet Union. In: Castells M (ed.) *The Information Age: Economy, Society and Culture Vol. 3. End of Millennium*. Blackwell Publishers, Cambridge, UK, pp. 4–69

Collier D (ed.) 1979 *The New Authoritarianism in Latin America*. Princeton University Press, Princeton, NJ

Collier R B 1999 *Paths Toward Democracy: Working Class and Elites in Western Europe and South America*. Cambridge University Press, New York

Dahl R 1971 *Polyarchy*. Yale University Press, New Haven, CT

Diamond L, Linz J J, Lipset S M (eds.) 1989a *Democracy in Developing Countries. Africa*. Lynne Rienner, Boulder, CO

Diamond L, Linz J J, Lipset S M (eds.) 1989b *Democracy in Developing Countries. Asia*. Lynne Rienner, Boulder, CO

Diamond L, Linz J J, Lipset S M (eds.) 1989c *Democracy in Developing Countries. Latin America*. Lynne Rienner, Boulder, CO

Dogan M, Higley J 1998 *Elites, Crises, and the Origins of Regimes*. Rowman and Littlefield, Lanham, MD

Foweraker J 1995 *Theorizing Social Movements*. Pluto Press, London

Haggard S, Kaufman R 1995 *The Political Economy of Democratic Transitions*. Princeton University Press, Princeton, NJ

Huntington S 1991 *The Third Wave. Democratization in the Late*

Twentieth Century. University of Oklahoma Press, Norman, OK
Linz J J, Stepan A (eds.) 1978 *The Breakdown of Democratic Regimes*. Johns Hopkins University Press, Baltimore, MD
Linz J J, Stepan A 1996 *Problems of Democratic Transition and Consolidation: Southern Europe, South America and Post-Communist Europe*. Johns Hopkins University Press, Baltimore, MD
Lipset S M 1960 *Political Man: The Social Bases of Politics*. Doubleday/Anchor Books, New York
O'Donnell G 1973 *Modernization and Bureaucratic Authoritarianism: Studies in South American Politics*. Institute of International Studies, University of California, Berkeley, CA
O'Donnell G, Schmitter P 1986 *Transitions from Authoritarian Rule. Tentative Conclusions about Uncertain Democracies*. Johns Hopkins University Press, Baltimore, MD
O'Donnell G, Schmitter P, Whitehead L (eds.) 1986 *Transitions from Authoritarian Rule. Prospects for Democracy*. Johns Hopkins University Press, Baltimore, MD
Offe C 1997 *Varieties of Transition. The East European and East German Experience*. MIT Press, Cambridge, MA
Przeworski A 1991 *Democracy and the Market. Political and Economic Reforms in Eastern Europe and Latin America*. Cambridge University Press, New York
Przeworski A, Limongi F 1997 Modernization: theories and facts. *World Politics* **49**(2): 155–83
Rueschemeyer D, Stephens E H, Stephens J D 1992 *Capitalist Development and Democracy*. University of Chicago Press, Chicago
Rustow D 1970 Transitions to democracy: Toward a dynamic model. *Comparative Politics* **2**(3): 337–63
Tarrow S 1995 Mass mobilization and elite exchange: Democratization episodes in Italy and Spain. *Democratization* **2**(3): 221–45
Whitehead L (ed.) 1996 *The International Dimensions of Democratization: Europe and the Americas*. Oxford University Press, New York

G. L. Munck

Demographic Analysis: Probabilistic Approach

1. Random Variation in Demographic Behavior

Demography is one of the many fields of science where statistical theory can be put to good use. Typical events of interest to demographers are births, deaths, the formation and disruption of marital and nonmarital unions, and geographical migration. The quintessential determinants of demographic behavior are sex, age, and cohort or calendar period, and other typical individual-level factors are race, social and family background, ethnicity, religious orientation, labor-force participation, and educational attainment. There may also be contextual determinants, such as institutional settings, laws and regulations (including public policies), and other collective features that individuals face. Aspects of aggregate behavior in the same population may also be important. (When many people already cohabit in nonmarital unions, individuals may find it easier to form such a union themselves. When divorce is common, dissatisfied spouses may find it easier to dissolve their own marriage.) To handle this multitude of features, individual demographic behavior is described in terms of event-history models where individuals move between predefined statuses under the influence of personal or group characteristics, of their own personal history, of contextual factors, and of period influences in the population they belong to.

Since any human population has finitely many members, there is an amount of random error in analytic procedures based on individual-level data. If life histories are selected in some random manner from a larger population, as in a survey sample, then some identifiable part of the error may be due to the sampling procedure, but another part is caused by the fundamental randomness of the individual process paths. Stochastic variation does not only arise from sampling variation in survey data, or from deliberate randomization of experiments. Whether the data were obtained for a sample or for the whole population segment in question is a separate issue. The size order of the random variation involved usually depends on the number of contributing individuals, not on whether the data comes from a sample survey or not. As was expounded by Westergaard (1880) more than a century ago and repeated many times over by people like Udry et al. (1979) and by Brillinger (1986) and his discussants, random variation is intrinsic to data on demographic behavior, just as it is for data from other social and behavioral sciences.

Demography may be special in that some of its data come from populations large enough to permit the investigator essentially to disregard randomness in the corresponding vital rates. In such cases, the challenge is to explain any irregularities in plotted curves of vital rates as local variations of substantive interest, or as caused by systematic registration errors. Significance tests and similar statistical tools may then become less important. This does not preclude random variation from being prominent in data from small populations and from smallish segments of national populations.

2. The Event-history Approach

In event-history analysis, demographic events are seen as transitions between states in a suitable state space. The factors that influence the transitions may be fixed or change over time, they may be exogenous to the individual history or endogenous to it, recorded in an available set of individual-level data, or unobserved causes of observable selectivity. Early behavior on a

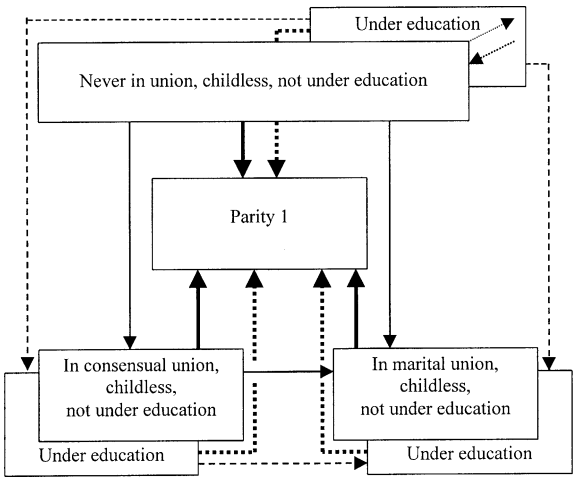

Figure 1
A state space for the analysis of first-birth intensities under the influence of civil status and enrollment in education

process may impact on later behavior on the same process, as when age at first birth is shown to influence an individual's later childbearing behavior; this is one of the most persistent findings in family demography. Individual behavior in one arena may interact with that in another arena, as when a woman's educational attainment and labor-force participation are seen as influencing her childbearing while conversely having children determines her chances of improving her educational level or maintaining her ties with the labor market.

The states between which transitions are made are demographic (and other) statuses like 'never married,' 'in a consensual union,' 'married,' 'childless,' 'of parity n' (having had precisely n live births), 'under education,' 'in the labor force,' and so on. When the state space is not too complex, it can sometimes be represented graphically as in Fig. 1, where the boxes represent the various statuses in question and the arrows indicate possible transitions for an analysis of first births before and in first marital or nonmarital unions. There are moves in three dimensions in Fig. 1, namely (a) changes in civil status (never married, married, in consensual union), (b) school enrollment (under education or not in school), and (c) parity (in this case childless or of parity 1). Arrows are formatted to indicate transitions of prime interest, competing risks, and other transitions. As is typical for such diagrams, some states and transitions have been left out to simplify the graphic, namely (a) the moves in and out of education for individuals living in partnerships, and (b) the disruption of existing partnerships. In empirical analyses one would account for (though possibly not analyze) such transitions as well.

A focus of attention is the *intensity* or *hazard* $\lambda_{ij}(t; x(t))$ of transition from some state i to some other state j at time t, for a suitable set of pairs (i, j) and a suitable vector $x(t)$ of determinants. The basic time variable t typically is the individual's own age or else the duration since a given event (duration of partnership, age of youngest child). Theories about individual demographic behavior are reflected in the specification of these elements, including the choice of t, the selection of covariates in the vector $x(t)$, and the form of the intensities λ_{ij}. For instance, an investigation of stepfamily fertility may address the hypothesis that for comparable partnerships, the rate of arrival of their first common child should be independent of the number of children that each partner had before the union was formed (*the union-confirmation hypothesis*). This would be reflected in a specification of a first-birth (or first-conception) intensity, say $\phi(t; n_1, n_2, y(t))$, that at each union-duration t would depend on the parities n_1 and n_2 of the partners at union formation and on some set of fixed and time-varying covariates collected in the vector $y(t)$. Confirmation of the hypothesis would consist in demonstrating that the intensity does not depend on n_1 and n_2.

In general the intensity λ_{ij} is the probabilistic counterpart of the empirical rate of transition from i to j. In the simplest case the latter is an occurrence/exposure rate $\hat{\lambda}_{ij}(g) = D_{ij}(g)/R_i(g)$ (also called a *vital rate of the first kind*) of some population group g in a given range for t during a particular period. The occurrences $D_{ij}(g)$ and exposures $R_i(g)$ are recorded for the group g in question. Under such circumstances, $\hat{\lambda}_{ij}(g)$ is the maximum-likelihood estimator for the intensity, it is asymptotically normally distributed with the intensity as its mean and an asymptotic variance that can be estimated by $\hat{\lambda}_{ij}(g)/R_i(g)$, and different such rates are asymptotically independent of each other. This means that there exists a statistical theory for conventional demographic rates, a theory which can be used for the purpose of estimation, testing of hypotheses, the computation of confidence intervals, and so on, in the normal manner of statistical analyses.

Various ways of computing demographic occurrence/exposure rates are discussed in *Demographic Techniques: LEXIS Diagram*. Keiding (1990) has given a probabilistic interpretation of the Lexis diagram in epidemiological research. Many other demographic practices can similarly be given a useful underpinning in statistical theory, as we show in what follows.

3. *Individual-level Demographic Models*

For many events, demographers typically compute vital rates for a sequence of age groups and thereby produce a *schedule* $\{\hat{\lambda}_x : x = 1, 2, \ldots\}$ *of age-specific rates*

(of the first kind). The next step is often to fit some parametric mathematical function to this schedule, for instance Thiele's (1871) mortality function

$$\lambda_x = a_1 \exp\{-b_1 x\} + a_2 \exp\{-\tfrac{1}{2}(x-c)^2/\sigma^2\} + a_3 \exp\{b_3 x\}$$

with $a_1 + a_2 + a_3 = 1$. Fitting such a formula to the $\{\hat{\lambda}_x\}$ is an exercise in what actuaries call analytic graduation, for which a statistical theory also exists (Hoem 1972) and is useful for the selection of fitting criteria, variance computations, testing of hypotheses, and so on. When a formula has been established, it is often called a *demographic model*. Occasionally the model is developed for a corresponding schedule of binomial probabilities $q_x = 1 - e^{-\lambda_x}$, as in $q_x = \phi_x/(1+\phi_x)$ with

$$\phi_x = A^{(x+B)^C} + D \exp\{-E(\ln x - \ln F)^2\} + GH^x$$

This three-part expression, which was developed by Heligman and Pollard (1980), is based on considerations similar to those of Thiele's mortality formula. In an unorthodox approach, Ansley Coale (1971) fitted a model for first-marriage formation to first-marriage rates 'of the second kind,' i.e., to rates where the number of marriages formed at a certain age is divided by the *total* person-years at that age, including person-years lived after first-marriage formation, and not only divided by the person-years exposed to risk as in a rate of the first kind. (See *Demographic Models* and *Demographic Techniques: Rates of the First and Second Kind*.) The statistical theory for rates of the second kind is useful for the legitimacy it gives them in empirical analyses and for the further developments it has led to (Finnäs 1980; Borgan and Ramlau-Hansen 1985).

Another class of procedures that can benefit from being brought into the realm of statistical theory is what demographers know as Brass's *relational models* (Brass 1974, etc.). They can be described in elementary probabilistic terms as follows. Given a 'standard' probability distribution function $H(x)$, a whole family of further distribution functions $\{H_{ab}(x); -\infty < a < \infty,\ b > 0\}$ can be generated by the 'relational transformation' $\lambda\{H_{ab}(x)\} = a + b\lambda\{H(x)\}$, where for $0 < y < 1$ $\lambda(y)$ is some function with a positive derivative everywhere. Brass chose $\lambda(y) = \ln\{y/(1-y)\}$ and $\lambda(y) = -\ln(-\ln y)$. He used the former for a relational transformation of $H(x) = 1 - \lambda_x$ and the latter for a corresponding transformation of $H(x) = F(x)/F(50)$, where $F(x)$ is a set of age-specific fertility rates cumulated up to age x. If we let $\Psi(x) = \lambda^{-1}(x)$, then $\Psi(x)$ is another probability distribution function and $\lambda(y)$ is its percentile function. Brass's two choices of $\lambda(y)$ give $\Psi(x) = 1/(1+e^{-x})$ and $\Psi(x) = \exp(-e^{-x})$, a logistic distribution and an extreme-value distribution, respectively. We can rewrite the generating relation as

$$H_{ab}(x) = \Psi\{a + b\Psi^{-1}[H(x)]\}$$

Once it has been established that we are dealing in parameters in a class of statistical models, then the usual considerations from statistical theory can be brought to bear on the issues involved. If H and Ψ are given, then a and b can be estimated in the usual way for statistical parameters, and their statistical properties may be known. If one does *not* want to use an outside 'standard' distribution for H, this function can perhaps be estimated from the data by a method whose statistical properties one knows. In either case, statistical principles are useful. One of these may be that $\mu_{ab}(x) = h_{ab}(x)/\{1 - H_{ab}(x)\}$ is more sensitive to subgroup differences than the distribution function $H_{ab}(x)$ itself is. (With $h_{ab}(x) = \mathrm{d}H_{ab}(x)/\mathrm{d}x$, $\mu_{ab}(x)$ is the hazard function of H_{ab}.) Optimal fitting may not be what one strives for, but when it is, one may be well advised to use the hazard as a basis for model fitting, model testing, and so on, as usual. In this connection it may pay to heed the warning that Feller gave against fitting models to sigmoid curves as far back as in 1940 (Feller 1940).

Demographic models for schedules of age-specific rates (and other similar schedules) are valuable in the analysis of inaccurate or incomplete data and as tools for population forecasting and other kinds of 'technical' computations. They have a place in data-quality assessment, though such usage should be tempered by the fact that the demographic models themselves may be inaccurate representations of reality. (This is demonstrated by the perennial need to revise model life tables and by amusing report titles asking why we get such lousy fits; see Lockridge and Broström 1983.) Demographic models are also useful for smoothing age-specific demographic schedules, and some such efforts may be for the purpose of removing random variation. In most cases, the probabilistic approach provides useful guidelines and helpful operational procedures. In one case, it even makes certain demographic procedures redundant. The whole theory of increment-decrement life tables may be replaced by parts of the theory of time-continuous Markov chains and some elements of numerical analysis (Hoem and Funck Jensen 1982).

For more information about these issues, see *Demographic Models*, *Multistate Transition Models in Demography*, and *Demographic Techniques: Rates of the First and Second Kind*.

4. Standards and Standardization

Brass's use of a 'standard' probability distribution H_x follows an old tradition in demography. Finding and using a 'standard' schedule plays a prominent role in much demographic analysis of rate schedules $\{\hat{\lambda}_x\}$. Such a standard is often sought outside of the current data set and is used to highlight special features of the data in hand. Once the relation to statistical theory has been established for an empirical procedure, the need for an outside standard may be removed and all

parameters may be estimated from the current data, as was demonstrated by Xie and Pimentel (1992) for the celebrated Coale-Trussell formula (1996) for marital fertility. (See *Demographic Models*.) This approach can sometimes provide solutions to problems that have pestered the discipline for a long time. For instance, the competition between *direct* and *indirect standardization* has been settled in favor of the latter (Hoem 1987), in the sense that when standardization is fully legitimate there exists an improved form of indirect standardization based on a maximum likelihood estimator which has minimal variance. This result relieves the user of the onerous job of demonstrating the suitability of a selected outside standard population for use in direct standardization, which is suboptimal in any case. Here is a sketch of the basic ideas.

Suppose that for each of a number of subpopulations, indexed by k, one has a schedule $\{\hat{\lambda}_{xk}: x = 1, 2, \ldots\}$ of occurrence/exposure rates, $\hat{\lambda}_{xk} = D_{xk}/R_{xk}$. In *direct standardization* a 'standard' population age distribution $\{P_x\}$ is specified and a quantity like $\hat{v}_k = \Sigma_x \hat{\lambda}_{xk} P_x / \Sigma_x \hat{\lambda}_{x1} P_x$ is used as a measure of the risk level of subpopulation k relative to that of subpopulation 1. (Note that $\hat{v}_1 = 1$.) For *indirect standardiz ation*, a 'standard' age schedule $\{\tilde{a}_x\}$ of risk is specified instead of $\{P_x\}$ and a quantity like $\hat{w}_k = D_{\cdot k}/\Sigma R_{xk}\tilde{a}_x$ measures the relative risk. The dot symbol in the subscript denotes summation, so $D_{\cdot k} = \Sigma_x D_{xk}$. (In mortality studies, $\hat{w}_k$ would be called a *standardized mortality ratio*.) Demographers, epidemiologists, and surely others agonize a lot over the choice of a standard population $\{P_x\}$, and they should perhaps agonize as much over $\{\tilde{a}_x\}$, though normally one simply chooses $\tilde{a}_x = D_{x\cdot}/R_{x\cdot}$. Correspondingly, one could choose $P_x = R_{x\cdot}$. With these latter choices, no outside standard is involved, of course.

In event-history analysis, $\hat{\lambda}_{xk}$ would be regarded as an estimator of an underlying model parameter λ_{xk}, in the spirit of our account toward the end of Sect. 2 above, and λ_{xk} would be a level-parameter in a piecewise-constant transition intensity. Suppose that $\lambda_{xk} = a_x b_k$ with $b_1 = 1$, which means that schedules are essentially proportional to each other and that $\{a_x\}$ operates as a baseline schedule, since $\lambda_{x1} \equiv a_x$. (Standardization is usually considered fully legitimate only when such a multiplicative specification holds.) Then both $\hat{v}_k$ and $\hat{w}_k$ are estimators of the proportionality factor b_k, but both are suboptimal, for one can find maximum-likelihood estimators $\{a_x^*\}$ and $\{b_k^*\}$ from the relations

$$a_x^* = \frac{D_{x\cdot}}{\sum_k R_{xk} b_x^*} \quad \text{and} \quad b_k^* = \frac{D_{\cdot k}}{\sum_x R_{xk} a_x^*}, \quad \text{with } b_1^* = 1$$

using an iterative numerical procedure to provide a solution. Breslow and Day (1975) have shown that conventional indirect standardization appears as the first step in one such procedure and thus may be regarded as a first approximation to the optimal solution. The approximation is usually excellent. Note how $\hat{w}_k$ has the same form as the second of the maximum-likelihood relations.

5. Population-level Models

Beside the focus on individual-level behavior, a large part of demographic analysis is concerned with the dynamics of entire populations, collected in the theory of stable populations and its ramifications. The purpose is to investigate how the consequences of individual-level processes work their way up to the population level. In this connection as in others, a probabilistic approach gives a more satisfactory basis for the mathematical derivations and opens the way for results that cannot be formulated, let alone proved, in deterministic theory. Extinction theorems and stochastic ergodicity theorems in stable population theory fall in this category (Cohen 1979, 1982, Tuljapurkar 1989, and others). The probabilistic approach has inspired a new treatment of uncertainty in population forecasts (Cohen 1986). Keiding and Hoem (1976) have directly addressed long-standing concerns of demographers in a probabilistic vein.

For more information about *population dynamics*, see *Population Dynamics: Theory of Stable Populations*, *Population Dynamics: Classical Applications of Stable Population Theory*, *Population Dynamics: Momentum of Population Growth*, *Population Dynamics: Theory of Nonstable Populations*, *Population Dynamics: Mathematic Models of Population, Development, and Natural Resources*, and *Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems*. See also *Population Cycles, Formal Theory of*. For material about *population forecasts*, see *Population Forecasts* and also *Demographic Techniques: Small-area Estimates and Projections* and *Families and Households, Formal Demography of*.

6. Conclusions

The present Encyclopedia article has been devoted to the merits of the probabilistic approach to demographic analysis. Such an approach is generally accepted in other fields of application (biostatistics, epidemiology, insurance mathematics), but in demography there is still considerable tension between those who feel that probabilistic notions are essential to the field's methodology and others who largely regard them at best as the icing on a cake. It is our contention that it does not pay to develop demographic procedures of analysis in isolation from statistical principles any more than it pays in other disciplines. The probabilistic

approach is the way to a deeper understanding of inferential techniques as well as to further methodological development beyond intuitive reasoning.

See also: Event History Analysis: Applications; Event-history Analysis in Continuous Time

Bibliography

Borgan Ø, Ramlau-Hansen H 1985 Demographic incidence rates and estimation of intensities with incomplete information. *Annals of Statistics* **13**(2): 564–82

Brass W 1974 Perspectives in population projection (with discussion). *Journal of the Royal Statistical Society, Series A* **137**(4): 532–83

Breslow N E, Day N E 1975 Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *Journal of Chronic Diseases* **28**: 289–303

Brillinger D R 1986 The natural variability of vital rates and associated statistics (with discussion). *Biometrics* **42**: 693–734

Coale A J 1971 Age patterns of marriage. *Population Studies* **25**: 193–214

Coale A, Trussell J 1996 The development and use of demographic models. *Population Studies* **50**(3): 469–84

Cohen J E 1979 Ergodic theorems in demography. *Bulletin of the American Mathematical Society (New Series)* **1**(2): 275–95

Cohen J E 1982 Multiregional age-structured populations with changing rates: Weak and stochastic ergodic theorems. In: Land K, Rogers A (eds.) *Multidimensional Mathematical Demography*. Academic Press, New York, pp. 477–503

Cohen J E 1986 Population forecasts and confidence intervals for Sweden: A comparison of model-based and empirical approaches. *Demography* **23**(1): 105–26

Feller W 1940 On the logistic law of growth and its empirical verifications in biology. *Acta Biotheoretica* **5**(2): 51–66

Finnäs F 1980 A method to estimate demographic intensities via cumulative incidence rates. *Theoretical Population Biology* **17**: 365–79

Heligman L, Pollard J H 1980 The age pattern of mortality. *Journal of the Institute of Actuaries* **107**(1): 49–80

Hoem J M 1972 On the statistical theory of analytic graduation. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* **1**: 569–600

Hoem J M 1987 Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review. *International Statistical Review* **55**(2): 119–52

Hoem J M, Funck Jensen U 1982 Multistate life table methodology: A probabilist critique. In: Land K, Rogers A (eds.) *Multidimensional Mathematical Demography*. Academic Press, New York, pp. 155–264

Keiding N 1990 Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London, Series A* **332**: 487–509

Keiding N, Hoem J M 1976 Stochastic stable population theory with continuous time. *Scandinavian Actuarial Journal* **1976**: 150–75

Lockridge K A, Broström G 1983 Coale and Trussell's 'm' visits Sweden, or, why do we get such lousy fits? In: Lockridge K A *The Fertility Transition in Sweden: A Preliminary Look at Smaller Geographic Units, 1855–1890*. Umeå University, Demographic Data Base, Report No. 3, Appendix I

Thiele T N 1871 On a mathematical formula to express the rate of mortality throughout the whole of life, tested by a series of observations made use of by the Danish Life Insurance Company of 1871. *Journal of the Institute of Actuaries and Assurance Magazine* **16**: 313–29

Tuljapurkar S 1989 An uncertain life: Demography in random environments. *Theoretical Population Biology* **35**(3): 227–94

Udry J R, Teddlie C, Suchindran C M 1979 The random variation in rates based on total enumeration of events. *Population Studies* **33**(2): 353–64

Westergaard H 1880 An exposition and evaluation of recent investigations on the mortality of the various classes of society (in Danish). University of Copenhagen Prize Paper. Published 1882 as *Die Lehre von der Mortalität und Morbilität: Anthropologisch-statistische Untersuchungen*. Gustav Fischer, Jena

Xie Y, Pimentel E E 1992 Age patterns of marital fertility: Revising the Coale-Trussell method. *Journal of the American Statistical Association* **87**: 977–84

J. M. Hoem

Demographic Data Regimes

Demography is the study of human populations, including size, structure, distribution over space, socioeconomic characteristics, households and families, migration, labor force, and vital processes (birth, death, formation and dissolution of marital, and other unions). A central feature of this study is the collection, compilation, and presentation of demographic statistics. The most important single instrument is the census, an effort to enumerate the number of inhabitants and, historically, often households and/or families. In recent centuries, more detail has been collected than simple counts, namely such information as age, sex, race, nativity, language, marital status, occupation, industry, fertility, mortality, income, residence, and family and household structure. Censuses may either be complete counts or statistical samples of the population. The other mainstay of demographic data gathering is the continuous registration of vital events (births, deaths, marriages, divorces, and sometimes migration). This has been achieved through the establishment of parochial (church) or civil systems for the collection of such events. In recent decades, a number of nations have developed population registration systems that simply keep a continuous record of demographic events. A census can be taken by enumerating the register at any point in time. These modes of record-keeping have been augmented by sample surveys, especially where the cost and difficulty of taking a census are large, such as in many developing nations. Surveys also provide a means to calculate vital rates via indirect methods of estimation. Other administrative instruments have been used to study populations, including tax lists, military muster and conscription rolls, budgetary surveys, wills and bequests, and insurance data.

1. Prestatistical Regimes

The ancient civilizations of Egypt, Babylon, China, Palestine, and Rome took censuses. These were used for administrative purposes, such as determining fiscal, labor service, and military obligations. Heads of households, males of military age, adult citizens, and taxpayers usually were counted. Women and children were often omitted. Few of these records survive. The Bible mentions several, including one of the Israelites at the time of the Exodus and one by King David around 1017 BC. Rome conducted its censuses for about 800 years, first counting citizens and property in Italy for tax and military purposes, and later extending this to the entire empire. Joseph and Mary traveled to Bethlehem in response to such a census. The ancient empires of China took periodic censuses, the first relatively accurate one being in AD 2. They counted 'doors' (households) and 'mouths' (persons). These counts have allowed us an approximate reconstruction of China's population over the last 2,000 years. The Inca empire of ancient Peru is said to have taken censuses, but the *quipu* recording system has never been deciphered. Similarly, the pre-Columbian Mexicans are known to have collected tribute censuses, but no codices recording this have survived.

More recently, the Domesday Inquest of England by King William I in 1086 provides us with a population benchmark for England. Such periodic fiscal and administrative surveys have functioned as quasicensuses for demographic historians, other examples being the French hearth tax lists of 1328, the English poll tax list of 1377, and the English hearth tax returns of 1690.

In England, a parish registration system of baptisms, burials, and marriages was instituted in 1538 for the new Church of England. France began such a system piecemeal, but the registers appear to be of acceptable quality by about 1670. In 1563 the Council of Trent decreed that all Catholic countries should keep parish registers. Other European countries followed suit. Ancient China created a system of family registers that was adopted in Japan and made official in 1721. National totals were compiled from the registers in Japan every 6 years up to 1852. Korea also instituted the register system, and compilations were made with some frequency from 1395.

2. Demographic Regimes of the Eighteenth and Nineteenth Centuries

The modern regime of demographic data collection on a national basis had its origins in the Scandinavian countries (Sweden, Denmark, Norway, Iceland, and, at times, Finland). Parochial registers were begun in the seventeenth century, using the clergy of the national, Lutheran churches. In Sweden, registration was made compulsory in 1686, in Norway in 1685, in Denmark in 1645/46, and in Finland in 1628. Annual data on births, deaths, and marriages were not published, however, until the eighteenth century, beginning in 1735/36 for Norway, Denmark, and Sweden. These nations also created census-like results from these registers, but actual household enumerations of a modern type were undertaken in 1749 in Sweden and in 1769 in Denmark and Norway.

Although the Scandinavian nations were forerunners, censuses had been taken elsewhere on more local level. In New France and Acadia in North America, 16 enumerations were made between 1665 and 1754, and a comprehensive system of parish registration was set up from the onset of French colonization (1608). For the British colonies in America, no fewer than 124 censuses were undertaken between 1623 and 1776. Some parish registers were set up, particularly in New England. Several Italian principalities (e.g., Sicily, Naples) enumerated their populations in the seventeenth and eighteenth centuries.

Later in the eighteenth century, some European nations took irregular and often rather incomplete censuses, including Spain (1768/69, 1787, 1797) and Portugal (1768). Regular enumerations truly began with the United States, whose Federal Constitution (1788) mandated a decennial census for the apportionment of seats in the House of Representatives. The first census was taken in 1790, and since then 22 national censuses have been taken. A much more general trend to regular national-level enumerations began in the early to mid-nineteenth century, both for governmental and general scientific purposes. This required the improvement of the bureaucratic mechanisms for collecting, tabulating, and publishing the results. There was also an increase in the types and detail of the types of information collected. More detail by age and sex was taken, and information on economic activities (e.g., occupation, industry) was also collected. Special censuses of agricultural and manufacturing units came to accompany the population inquiries. Among the nations commencing or continuing such activities in the early nineteenth century were the United Kingdom (England and Wales (1801), Scotland (1801), Ireland (1821)), France (1801), the Netherlands (1816), Portugal (1801), Prussia (1816), Bavaria (1816), Switzerland (1837), and Belgium (1831). As new nations were created, they began data collection for the new consolidations of former states: the German Empire took its first census in 1871 and the new United Kingdom of Italy in the same year. As the century progressed other nations joined: Spain (1857), Serbia (1834), Greece (1856), Russia (1897), Romania (1899), Austria (1869), Hungary (1857). Overseas nations also began the process: Canada (1851), Japan (1872), New Zealand and parts of Australia (1851), and many of the new nations of Latin America (e.g., Argentina (1869), Brazil (1872)). For a number of these nations, early

efforts at enumeration were only partly successful. Colonial authorities took censuses in various other areas, especially in the Caribbean and in Asia (e.g., India (1881)). Many of these enumerations were seriously incomplete. Africa was the least well measured (and remains so).

One important issue for the promotion of demographic data sources was the creation of national statistical offices. In many cases, statistical collection functions were distributed among a variety of venues. In the United States, the census initially was collected by a temporary agency created for each census. The United States Bureau of the Census did not attain permanent status until 1902. Many European nations had statistical offices responsible for the collection and publication of demographic data by the middle of the nineteenth century (e.g., Statistics Sweden in 1858), and some had created institutions for similar functions for colonial areas. Many national-level annual statistical yearbooks were being published by these offices by the 1870s and 1880s.

An important impetus in this direction was the creation of a series of international statistical congresses, the first being in 1853. A series of recommendations were made concerning the need for regular censuses meeting basic standards. These resolutions were modified and expanded in subsequent congresses and were reinforced by the creation of the International Statistical Institute (1853). This work was continued by such bodies as the International Labor Office (1920), the Committee of Statistical Experts of the League of Nations (1920), the Inter-American Statistical Institute (1940), the World Health Organization (1948), and the Population Division of the United Nations (1947).

Vital registration did not expand as quickly. The countries of Scandinavia, Britain, France, many German states, and other European nations had the advantage of existing parochial registration systems. Some results from these systems were aggregated and published, such as in Scandinavia and England, or as 'Bills of Mortality' elsewhere, These systems gradually were converted to civil registration and thus became the basis for modern vital statistics collection. France secularized its parish registration under the Napoleonic Code in 1804. England and Wales instituted civil registration in 1837 and commenced published tabulations in 1838, but the system was only likely reasonably complete by the 1860s. The United States, a forerunner in regular enumerations, was a laggard in vital registration. The process was left up to individual states and, consequently, was accomplished unevenly. The state of Massachusetts was the first to have statewide vital registration (1842), but a partial national Death Registration Area was not even set up until 1900 (with 10 states and the District of Columbia, covering about 25 percent of the population). A Birth Registration Area was not created until 1915, and the system did not cover the entire nation until 1933. Canada experienced similar problems with provincial registration, and the system there was not complete until 1926. Although the German Empire immediately created an Imperial Statistical Office in 1871, the various state statistical offices were only able to furnish basic counts immediately. Deaths by age, sex, and cause were not available for the whole empire until 1901.

Additionally, many nations established controls of the movement of persons across their borders, especially in the late eighteenth and in the nineteenth centuries. The consequence was the collection of international migration statistics, especially about arrivals but subsequently (and partially) on departures too. For instance, the United States was a major recipient of European migrants before World War I and began collecting data on arrivals in 1819 but on departures only from 1906. Also, publication of details was often uneven.

3. *The Twentieth Century*

In the period 1855–64, 24 sovereign nations took population censuses, 17 of them in Europe. During 1895–1904, 44 nations or colonies took a census, 23 in Europe and seven in North America and the Caribbean, but also five in South America and seven in Asia and Oceania. This progress continued. For the period 1945–54, 65 nations or colonies took censuses (28 in Europe, 12 in North America and the Caribbean, eight in South America, 15 in Asia and Oceania, and two in Africa). At that point, basic population counts were available for about 80 percent of the estimated world population, though detail was much more limited. For example, population classified by single years of age and sex was reported for only about 46 percent of the estimated world population. By the end of the twentieth century, almost all of the 233 countries in the United Nations had taken a census within the last two decades of the century. In some cases, nations with advanced continuous population registers (which collect information on a wide variety of social and demographic events) have supplemented or replaced censuses with estimates from their registers (e.g., the Netherlands, Sweden, Norway, Denmark, Belgium).

In terms of vital statistics, only about 53 percent of the estimated world population was covered in 1955, varying from 98 percent in North America, 94 percent in Europe, 43 percent in Asia, to 37 percent in Africa. Many of the systems in developing countries seriously under-registered both births and deaths. Sample surveys, such as the recent Demographic and Health Surveys, have supplemented knowledge of vital processes in all regions. As of 1992, of the 48 countries in Europe, all but three very small entities were categorized by the United Nations as having complete birth and death registration. In contrast, only 13 of the 57 countries of Africa were classed as having complete

registration of deaths and only 10 as having complete birth registration. For Asia the proportions were 20 out of 52 for birth registration and 18 for death registration. Of the 14 countries of South America, seven were listed with acceptable registration for both births and deaths. The United Nations standard is rather liberal: to be classified as complete a system must have achieved a coverage of at least 90 percent.

A considerable advance in the collection of mortality statistics was made in 1900 when a conference of the International Statistical Institute approved the first international list of causes of death (*ICD-1*). This has been revised a number of times since then and is now under the auspices of the World Health Organization; the revision being operated is the ninth [*ICD-9* (1979)]. As medical and scientific knowledge has improved, there has been a movement away from symptomatic cause attribution and toward etiological, organic, and epidemiological diagnosis.

Overall, although there have been substantial changes in demographic data regimes, especially in the twentieth century, the world is far from having a reliable and complete coverage. Recent political and economic tendencies do not indicate that the situation will improve. Sample surveys and indirect estimation of vital rates based on census and survey data would seem to be the answer for the immediate future for many developing nations, particularly in sub-Saharan Africa.

See also: Censuses: Comparative International Aspects; Censuses: Demographic Issues; Censuses: History and Methods; Databases, Core: Demography and Registers; Government Statistics; Statistics, History of

Bibliography

Gille H 1949/50 The demographic history of the northern European countries in the eighteenth century. *Population Studies* **3**: 3–65

Hollingsworth T H 1969 *Historical Demography*. Cornell University Press, Ithaca, NY

Linder F 1959 World demographic data. In: Hauser P M, Duncan O D (eds.) *The Study of Population: An Inventory and Appraisal*. University of Chicago Press, Chicago

Shryock H S, Siegel J S et al. 1971 *The Methods and Materials of Demography*. US Government Printing Office, Washington, DC, chaps. 2–3

United Nations 1949 *Population Census Handbook*, provisional edn. United Nations, Lake Success, NY

United Nations 1953 *Principles for a Vital Statistics System*. ('Statistical papers,' Ser. M, No. 20.) United Nations, New York

United Nations 2000 *Demographic Yearbook: Historical Supplement 1948–1997*. United Nations, New York

Willigan J D, Lynch K 1982 *Sources and Methods of Historical Demography*. Academic Press, New York

M. R. Haines

Demographic Measurement: General Issues and Measures of Fertility

1. Introduction

1.1 General Overview

The phenomena measured in demographic analysis are events, states, attributes, and cumulative experience. Classically, the core demographic activity consists in estimating the frequency of vital events, births and deaths, and of related events—migration, marriage (or, in more recent times, the formation of sexual unions), and divorce (or the dissolution of such unions). Other events studied in demography include conception, pregnancy, spontaneous and induced abortion, household formation and dissolution, family changes, and widow(er)hood. States of demographic interest include marital status, both legal and informal, and type of family or household to which individuals belong. The classic attributes of demographic interest are age and sex, along with ethnicity and socioeconomic characteristics such as educational level and social class. Cumulative experience is examined using proportions ever having experienced an event such as marriage, or estimates of the average number of a specified event experienced in a lifetime or by a specific age.

Events can be thought of as changes in state and, correspondingly, states can be considered to result from the occurrence and nonoccurrence of events. For example, the event of first marriage changes a person's marital status from single to married, and a person who has married and has not yet experienced the dissolution of their marriage is in the state of being currently married. In addition, the concepts of 'cumulative experience' and 'state' are to some extent interchangeable, for example, a woman who has had n live births (cumulated experience) is said to be of parity n (state). Measures of cumulative experience that do not refer to individual states are also used in demography, and these will be considered presently. The current entry does not deal with methods of collecting and classifying demographic data, but with the use of data once it has been collected and processed, ready for analysis.

Demography is concerned in general with aggregate phenomena at the population level, where populations vary from global to local in extent. The units whose events and states are aggregated can be individuals, couples, families, or households. Some demographic measures are, however, strictly aggregate in type, for example, a population growth rate, population density per unit area, or a population dependency ratio all have meaning at the aggregate level only. Demographic measures focus on level, timing, and distribution: the frequency of events or prevalence of states (level), the age at or time between events

(timing), and distributional aspects such as age structure.

The need for demographic rates and other measures is readily stated. Absolute numbers, whether population size or numbers of vital events, are certainly of demographic interest, particularly in an applied context, and are frequently the key input or output for policy purposes. However, population analysis is conducted in terms of the underlying phenomena—mortality, fertility, migration, and so on—that determine population change. For example, population projections are usually carried out by the component method, applying assumptions about fertility, mortality, and migration *rates*, rather than by extrapolating birth, death, or population *numbers* using mathematical models. Rates, proportions, and other indices are required also for comparative purposes, whether tracking time trends or making comparisons between (sub)populations. Spatially and temporally, populations vary in size and also in structure, and so measures that abstract from size and structure are required to compare the demographic metabolism, so to speak, of different populations.

Demographic rates and measures are numerous. The variety stems both from the widely varying forms in which data are available and from the nature of the phenomena themselves. Data may come from vital registration systems, from censuses and surveys, from parish registers, or from administrative records. These sources may differ in the kinds of information recorded, in questions asked, and in details published. Fertility is the area in which measures are probably most numerous, a result of the complexity of the phenomenon: births are repeatable events, they can occur to women inside and outside of marital or cohabiting unions, their order in a birth history can be of significance, and they can be associated with several dimensions of personal time—age, union duration, duration since previous birth, and so on.

1.2 Rates, Probabilities, Proportions, and Ratios

A demographic rate expresses the number of events occurring relative to person years at risk of the event in a defined population for a specified time and place. It may be expressed per person year, per 1,000 person years, per 100,000 person years at risk, and so on. The denominator, person years at risk, is often estimated by means of the mid-year population or by the average of two end-year population figures, each of these being an approximation to the average population during the year. Francophone demographers distinguish between type 1 rates (*taux de première catégorie*), in which the denominator of a rate is confined to those who have not yet experienced the event in question, and type 2 rates, where no such restriction is made (*taux de deuxième catégorie*) (see *Demographic Techniques: Rates of the First and Second Kind*). No special terminology is commonly accepted for this purpose in English-language demography, the distinction between such rates being implicit in the specification of the rate, but the English version of the French terminology will be used occasionally for brevity and precision. A probability is the likelihood that an individual will experience an event. It is estimated by the number of events occurring during a defined period or at a particular age to a specified group divided by the numbers of individuals present at the start of the period or age. A proportion is defined in the usual way, as the number with a given attribute at a given time point divided by the total population in question at that time. Finally, demographic measures also take the form of ratios, for example, the sex ratio at birth is the ratio of the numbers of male to female births. Ratios generally refer to aggregates, although anthropometric measurements in ratio form relating to individuals are also found in the medical demographic literature.

1.3 Crude Rates and Degree of Specificity

Demographic measures vary in their level of detail. The simplest are crude rates, expressed as the number of events per 1,000 (or other multiple) of the total population, without any further specificity. The two most common are the crude birth rate (CBR)—the number of births per 1,000 population in a year—and the crude death rate (CDR)—the number of deaths per 1,000 population in a year. They give basic information but do not allow refined analysis. Crude rates are of value in setting out the basic demographic parameters of a population and are used for descriptive purposes, for instance, when data needed for more detailed measures are either unavailable or unreliable.

Demographic rates and other measures are influenced by the composition of the numerator in respect of any factor by which the frequency of the phenomenon under study varies. Most importantly, crude rates are influenced by the age structure of a population, because of the pervasive association between age and demographic event rates, and also by sex composition, since demographic event rates usually vary by sex. Accordingly, specific rates may be calculated by restricting the numerator or the denominator, or both. Specificity may be introduced for one or more factors. For example, age-specific rates may be influenced by composition in respect of marital status, duration in a particular state, parity (number of children already born to a woman), urban/rural residence, and so on. So, the analyst might choose to calculate rates specific for a number of dimensions. For some purposes, variation according to specific factors may itself be the focus of interest. In

other contexts, variation with respect to, for example, age is taken as given and is not the subject of study. If so, the analyst will wish to remove the compositional effect from the comparison to be made. Traditionally, one of two procedures is used to remove the influence of such (nuisance) factors. Rates may standardized for the factors concerned or may be disaggregated progressively so as to arrive at rates specific for groups with greater (or in theory perhaps complete) internal homogeneity. Standardization may be through the conventional direct or indirect methods or by calculating a synthetic indicator of some kind (see Sect. 2.1.3 below).

Both procedures have disadvantages. Standardization, whether by the conventional direct or indirect methods, or by constructing synthetic indicators, is well known to be valid only where there are no interactions between the factors for which the standardization is carried out, or when there are no interactions between them and the categories, populations, etc. to be compared. Since such interactions are often found, straightforward standardization is inappropriate in many instances. Progressive disaggregation of rates has the disadvantage of producing potentially large numbers of rates that are not readily summarized and perhaps not readily interpretable. Modern methods of model fitting can in many instances provide a more general and rigorous solution to the pervasive need for standardization in demographic analysis, and can offer a considerable advance on the progressive disaggregation approach. Hoem (1987) and his predecessors have shown that indirect standardization can be seen as a first step in an iterative estimation procedure for intensity regression based on a multiplicative hazard model.

2. *Measures of Fertility*

2.1 *Level*

2.1.1 Crude and specific rates. The definition of the most common fertility measures is set out in Table 1. The CBR is widely used in cross-national comparison, particularly since the data required (an estimate of total births and of total population) are generally available. The general fertility rate (GFR) is more specific, in that it restricts the denominator to those at risk of experiencing a birth, that is, women of childbearing age (most commonly taken as ages 15–44). Age-specific fertility rates (ASFRs) introduce a further level of detail by age and have importance both in indicating the pattern of childbearing by age and as the basis for calculating the total fertility rate (see below). Age-specific fertility rates may be disaggregated further to obtain age-specific marital fertility rates and age-specific nonmarital fertility rates. Fertility rates may also be calculated specific for duration of partnership or marriage, in place of or jointly with age.

In the last few decades, fertility rates that are specific by order of birth have come into prominence. Such rates are particularly informative since time trends may vary by order of birth, particularly in modern contracepting populations or in populations where the use of contraception is on the increase. They are of two distinct types. Less detailed are rates in which the numerator only is disaggregated by order of birth, giving, for example, age-order-specific fertility rates. These are demographic rates of type 2 (see *Demographic Techniques: Rates of the First and Second Kind*), since the denominator includes all women of a given age, although the numerator is specific by order of birth. They should be distinguished from the more detailed age-parity-specific fertility rates in which the numerator consists of births of order i to women aged x and the numerator of women of parity i–1 who are aged x. (These latter are rates of type 1; note that the term 'order' relates to *births* and that 'parity' relates to *women*.) In the case of first births, both order-specific and parity-specific rates may also be specific by duration since marriage (or start of union), either in place of or combined with age-specificity; for second and higher order births, correspondingly, duration since previous birth may be specified in place of or in addition to age. Computing measures of fertility that are specific by order of birth from vital registration data can present data problems because the order of birth as recorded in a particular system may exclude a woman's premarital births and/or births occurring in previous marriages, and so may not represent true birth order over the woman's entire childbearing history.

With the exception of the crude birth rate, which applies only to a time period, all of these fertility rates may be calculated either on a calendar-period or cohort basis.

2.1.2 Parity progression ratios. Fertility rates may vary by parity, and in modern contracepting populations always do so. Because any given overall level of fertility may be reached through differing patterns of parity-specific birth rates, measures known as parity progression ratios (PPRs) were developed in the early 1950s to summarize the lifetime outcome of such variation. The parity progression ratio of order i is defined as the proportion of women of final parity i and above who, by the end of childbearing, have had at least $i+1$ births. Here $i = 0, 1, 2 \ldots m$–1, where m is the maximum number of births to any woman in the population. These ratios express, in other words, the probability that a woman who has had a given number of live births will have at least one further birth. In their original form, parity pro-

Table 1
Selected fertility measures

Measure and *time reference*	Definition	Notes
Level		
Crude birth rate (CBR) *period*	$\frac{B}{P} \times 1000$	B = number of births in a given year or period P = total person-years lived[1] during the year or period
General fertility rate (GFR) *period; cohort*	$\frac{B}{\sum_{15}^{49} W_x} \times 1000$	W_x = number of woman years lived[1] at age x during a year or period[2]. An upper age limit of 44 is sometimes used
Age-specific fertility rates[3] (ASFR) *period; cohort*	$F_x = \frac{B_x}{W_x} \times 1000$	B_x = births to women aged x or in age group x during a year or period[2] W_x = woman years lived[1] at age x or in age group x during a year or period[2]
Age-specific marital fertility rates[3] *period; cohort*	$\frac{B_{x,m}}{W_{x,m}} \times 1000$	$B_{x,m}$ = births to married women aged x during a year or period[2] $W_{x,m}$ = woman years lived[1] while married and at age x or in age group x during a year or period[2]
Age specific nonmarital fertility rates[3] *period; cohort*	$\frac{B_{x,u}}{W_{x,u}} \times 1000$	$B_{u,x}$ = births to unmarried women aged x or in age group x during a year or period[2] $W_{u,x}$ = woman years lived[1] while unmarried at age x or in age group x during a year or period[2]
Illegitimacy ratio[3] *period; cohort*	$\frac{B_u}{B} \times k$	B_u = births to unmarried women during a year or period[2]. The multiplier k may be 100 or 1000
Age-order specific fertility rates[3] *period; cohort*	$F_{x,i} = \frac{B_{x,i}}{W_x} \times 1000$	$B_{x,i}$ = births of order i to women aged x or in age group x during a year or period[2]
Age-parity specific fertility rates[3] *period; cohort*	$\frac{B_{x,i}}{W_{x,i-1}} \times 1000$	$W_{x,i-1}$ = woman years lived[1] while of parity $i-1$ and age x or in age group x during a year or period[2]
Total fertility rate[3] (TFR) *period; cohort*	$\sum_{15}^{49} F_x$ The upper limit of the summation may be 44 rather than 49	F_x is the age-specific fertility rate defined above. When rates are for five-year age groups, the sum is multiplied by 5. The cohort TFR is simply the mean family size of women in the cohort
Gross reproduction rate[3] (GRR) *period; cohort*	$\sum_{15}^{49} F_x^f$	F_x^f is the age specific rate of female births. The sum is multiplied by 5 if five-year age groups are used
Net reproduction rate[3] (NRR) *period; cohort*	$\sum_{15}^{49} F_x^f L_x^f$	L_x^f is the average number of person-years aged x in the stationary female life table, with a radix of 1.
Parity progression ratio[3] *Cohort; period version differs: see text*	$a_i = \frac{N_{i+1}}{N_i}$	N_i = number of women who have i births

Table 1
Continued

Measure and *time reference*	Definition	Notes
Timing		
Mean age at birth[3] (crude) *Period; cohort*	$\frac{\sum_{15}^{49} xW_x^b}{\sum W_x^b} + 0.5$	W_x^b = the number of women aged x last birthday having a birth during a year or period[2]
Standardized mean age at birth[3] (crude) *Period*	$\frac{\sum_{15}^{49} xF_x}{\sum_{15}^{49} F_x} + 0.5$	
Median age at birth[3] (crude) *Period; cohort*	Median age of women having a birth in a period or, within a cohort, median age of women at childbirth	
Standardized median age at birth[3] *Period*	Median of the age-specific fertility distribution	
Mean age at i-th birth[3] (crude) *Period; cohort*	$\frac{\sum_{15}^{49} xW_{x,i}^b}{\sum_{15}^{49} W_{x,i}^b} + 0.5$	$W_{x,i}^b$ = women aged x last birthday having a birth of order i during a year or period[2]
Standardized mean age at i-th birth[3] *Period*	$\frac{\sum_{15}^{49} xF_{x,i}}{\sum_{15}^{49} F_{x,i}} + 0.5$	$F_{x,i}$ = age-order specific birth rate of order i, during a year or period defined above (type 2)
First birth interval *Cohort; for synthetic period version, see text*	Duration from marriage (or start of union) to first birth	Less useful where moderate to high proportions of first births occur outside of a union
Second, third, etc. birth interval *Cohort; for synthetic period version, see text*	Duration from first to second birth, second to third birth and so on	Caution: cannot be calculated from the mean ages at births of each order, unless specifically for women with i+births; see text.

1 In the specification of demographic rates, person-years lived, woman-years lived, or man-years lived (whether overall, or at a particular age or in a given state) is the true denominator. For practical purposes, person-years lived in a period is estimated in various ways by different statistical agencies. In the UK and USA, it is usually estimated by the mid-year population (overall, or of a particular age, etc.). See also *Demographic Techniques*: *Rates of the First and Second Kind*.
2 Where the rate relates to a year or period, events and person-years lived are those occurring during the year or period in question. A cohort rate is based on events experienced and person-years lived by the cohort concerned.
3 These quantities can be defined analogously for men, but in most cases the male versions are rarely used.

gression ratios are calculated on a cohort basis, although they are meaningful for any group of women aged 45 plus. Methods for estimating these for women who have not reached the end of childbearing have also been proposed (Brass and Juarez 1983).

2.1.3 Synthetic or hypothetical cohort indicators. Since there are many specific fertility rates of any particular type, it is convenient, indeed a practical necessity, to have single-figure indices that summarize them. Such indices are also of value in producing estimates of lifetime experience on the basis of a single period's data, both because cohort data are often not available and because cohort data represent cumulative past rather than current experience. Synthetic or hypothetical cohort indicators are used for this purpose. A synthetic summary indicator is obtained by combining the specific rates of a calendar period to obtain an overall figure representing the lifetime fertility outcome that would result if a cohort of women experienced the period rates in question throughout their childbearing years. The component rates are combined in one of two ways: either multiplicatively, as in a life table, or by adding the age-specific rates at each age. In the first type of procedure, the denominators include only 'survivors,' that is those who have not yet experienced the event by the age or duration in question (rates/probabilities of type 1). The second type of procedure is used when the denominator of each age-specific rate includes both those who have and those who have not yet experienced the event (rates of type 2). The exact details depend on the type of rate in question and are mentioned below in each case.

The most widely used such summary indicator in the fertility arena is the classic period total fertility rate (TFR), obtained by adding the age-specific fertility rates (i.e., type 2 rates) of a given year or period across ages 15 to 44 or 15 to 49. The result can be interpreted as the average number of children that women would bear, if they were to experience throughout their childbearing years the age-specific rates of the period

in question. (For a modern underpinning of this classical interpretation, see Borgan and Hoem 1988.) It can also be seen as the average number of children per woman if the age-specific rates in question were to remain in force over a long period. Closely related to the classic total fertility rate are the gross reproduction rate (GRR) and the net reproduction rate (NRR). The GRR is simply the TFR confined to female births, and so represents the average number of *daughters* a woman would have if she experienced the age-specific female birth rates of a particular period. The NRR is obtained by modifying the GRR to take account of the probability that a newborn female may die before reaching reproductive age (see Table 1); it represents thus the average number of daughters who would reach reproductive age per woman of reproductive age that would result if a period's age-specific female fertility and mortality rates were to remain fixed. In a stable population, in which vital rates are and have been fixed over a lengthy period, the NRR represents the extent to which generations replace themselves. In these circumstances an NRR of $< 1, = 1$, or > 1 means that successive generations are, respectively, declining, stationary, or growing in numerical size, and so that overall population numbers are changing correspondingly. The GRR is roughly half the size of the TFR; the NRR is less than the GRR, with the gap between them depending on the level of female mortality at ages under 45 or 50: the higher the mortality level, the greater the disparity between the GRR and the NRR. The TFR, GRR, and NRR can also be specified on a male basis, that is using male age-specific fertility rates, and male survivorship, and the results will in general differ from the female values. Female rates being more widely available and more reliable, they are in general preferred to male rates.

An advantage of the TFR, the GRR, and the NRR is that they are standardized for age, the standard population being one imaginary woman proceeding through 30 or 35 years of childbearing. Synthetic or hypothetical cohort indicators in general—whether of fertility, nuptiality, mortality, and so on—can be criticized for the fact that they express period events in a metric of lifetime experience, and that this is conceptually at variance with the usually short period of calendar time whose demographic processes they represent (Ní Bhrolcháin 1992). Two points can be made in answer to this difficulty. First, the discrepancy between time series of summary cohort indicators and corresponding period synthetic summary indicators such as the TFR is well known, and the origins of the discrepancies between them have been widely discussed. Second, synthetic indicators can be thought of simply as convenient statistical summaries of the specific rates obtaining in a particular period—as a rough summary of the overall level of a phenomenon.

The major and well known source of discrepancy between cohort and period time series of the TFR, originally identified by Hajnal (1947) in relation to the GRR and NRR, is that time trends in the period TFR are influenced not only by the level of fertility but also by changes in the timing of childbearing. Where successive cohorts are moving to an earlier pattern of childbearing, the period TFR will be inflated, and where cohort fertility is slowing in pace the period TFR will be deflated, relative to the cohort series.

A related source of inaccuracy in these summary measures is that the additive (although not the multiplicative) method of combining a period's rates takes no account of past experience, at earlier ages or durations, of that event, particularly in the case of fertility. Past experience of events is likely to influence current type 2 rates, for example, if proportionately more of a given age group have already had a first birth, the (type 2) age-specific first birth rates can be expected to be lower than if proportionately fewer of the group have already begun childbearing. Furthermore, in a population where age-specific rates vary from year to year—almost always the case in real situations—the past experience of age groups in successive calendar periods (or, equivalently, successive cohorts) will differ. Thus, any additive summary indicator will be influenced by the composition of each age group with respect to past experience.

A final difficulty is that additive synthetic indicators may sometimes produce results that are impossible if the summary is interpreted as relating to a real cohort: for example, the sum of period age-specific first birth rates may exceed one, which appears to imply more than one first birth per woman. This last weakness is not present in multiplicative summary indices.

One solution to the difficulties inherent in the classic TFR is to use cohort fertility series to supplement period information. Hajnal (1947) suggested that the fertility experience of successive cohorts be deployed to reflect long-run trends, and that period fertility be employed to reflect recent and current trends, a practice that has become firmly established in demographic analysis since that time. Another alternative has become available more recently: synthetic or hypothetical cohort indicators calculated on a multiplicative basis. Such measures correct in part for differences in composition resulting from past experience, and have the advantage also that they do not produce results that would be impossible in a true cohort. The best known of these in current use in the fertility arena is the period parity progression ratio (PPPR), devised originally by Henry (1953). PPPRs are constructed from the parity-and age- and/or duration-specific probabilities obtaining in a period, yielding one summary indicator of the level of progression to each order of birth in a period. The (type 1) probabilities at each age or duration are essentially assembled as a life table, with age or duration as the measure of elapsed time, and the proportion having experienced the event (i.e., not 'surviving') by some specified age or duration estimated in the conventional

way (see *Life Table*). This proportion represents the parity progression ratio that *would* result *if* a cohort were to experience during their lifetime the parity- and age- and/or duration-specific birth probabilities of a particular period. For the specification of these indices see Feeney and Yu (1987), Ní Bhrolcháin (1987), and Rallu and Toulemon (1994). The PPPRs can be used to obtain an estimate of the period TFR that is alternative to the standard TFR, generated as

$$TFR^{pppr} = a^p_{0,t} + a^p_{0,t} \times a^p_{1,t} + a^p_{0,t} \times a^p_{1,t} \times a^p_{2,t} + \ldots + a^p_{0,t} \times a^p_{1,t} \times a^p_{2,t} \ldots a^p_{m-1,t} \quad (1)$$

where $a^p_{i,t}$ is the period parity progression ratio of order i for period t, and m is the maximum number of births occurring. (A standard notation is not yet in use for these quantities.)

This version of the TFR has the advantage over the conventional (additive) TFR that its components, the PPPRs, standardize to some extent for past experience and remove part of the timing influence, but it shares with the classic TFR the weakness that a single-figure summary may conceal parity-, age-, or duration-specific variation in probabilities that go beyond pure differences of level.

The issue of how best to summarize specific birth rates to obtain period time series is the subject of current research and debate (Ní Bhrolcháin 1992, Bongaarts and Feeney 1998). The classic TFR is recognized as too crude in many instances, and period parity progression ratios have been accepted as offering considerable clarification, especially in low-fertility societies and in societies in the course of fertility decline. One question at issue is to identify correctly the precise nature of time trends or cross-national differences, so that explanatory inquiry can be accurately focused on these, rather than on a potentially misleading summary indicator. Another objective is to make the best use of existing information in order to evaluate past and likely future trends. Further developments may be anticipated in this area in coming years.

2.2 Timing

Measures of fertility timing are somewhat less numerous than indicators of fertility level. The mean and median age at birth are the most basic ones. Somewhat more precise is the mean or median age specific by order of birth: mean or median age at first birth, at second birth, and so on. Mean and median ages may be obtained from either period or cohort data, and as with measures of the level of fertility, time series of these will generally differ. Whether or not specific by order, the period mean or median age at birth can be calculated in crude or standardized form, and published sources do not always indicate which of these is presented. The period crude mean age at birth (or at first birth, second birth, etc.) is obtained simply as the arithmetic mean age of all women having a (first, second, etc.) birth in a given period. The crude mean age at birth is widely used but has the disadvantage, in a period context, of being influenced by the age structure of women of childbearing age, which can vary through time and especially from one country to another. Far better for the purposes of period analysis, although more demanding of data since population denominators are required, is the standardized mean age at birth, or at first birth, second birth, and so on. Standardization is unnecessary in a cohort framework, which is self-standardizing, although care should be taken in comparisons of cohorts widely differing in mortality during the childbearing years. The standardized mean age at birth is obtained by weighting by the relative period age-specific fertility rates at each age (see Table 1) and so is, in fact, the mean of the age-specific fertility distribution. Similarly, the standardized mean age at i-th birth uses as weights the relative age-order-specific fertility rates (type 2, *not* type 1 rates; see above). The standardized mean age at birth is not a pure measure of fertility timing since it is influenced by the overall level of fertility. Two populations could have identical (standardized) mean ages at births of each order but the standardized mean age at birth could be later in one than in the other because proportionately more women in the first population go on to have births of higher orders. A measure of fertility timing encountered mostly in the historical demographic literature, the mean age at last birth, is used as an indirect indicator of the presence of birth control. This measure can be calculated without bias only for women who have reached the end of reproduction (age 45 or 50).

Another set of indicators of fertility tempo is the mean or median duration of birth intervals. The first birth interval is the time from marriage (or start of informal union) to the first birth. (This interval may be of limited value where a large proportion of births occur outside of formal or informal marital unions.) Later intervals are obtained as the duration from first to second birth (second birth interval), from second to third birth (third birth interval), and so on. Birth intervals calculated from the ages of a group of women at births of successive orders are incorrect because the ages at successively higher orders of birth are based on women of differing ultimate family sizes. The problem is that women who at the end of childbearing have larger families are usually younger at births of any given order than are those with smaller completed families. The i-th birth interval is calculated instead as (a) the time from birth $i-1$ (or start of union when $i = 1$) to birth i among women with i or more births, or equivalently as (b) the age at birth i minus the age at birth $i-1$ (or start of union when $i = 1$) among women with i or more births. Where individual level data are available, means or medians may then be calculated. If aggregate data only are available, median intervals

cannot be calculated accurately by differencing successive median ages at birth within parity groups. The median may be preferred to the mean birth interval, since birth interval distributions are typically positively skewed. Synthetic or hypothetical mean or median birth intervals for time periods may be obtained from the life tables used in constructing PPPRs, thus giving a period measure of birth timing analogous to, and supplementing, the standardized mean age at birth (Ní Bhrolcháin 1987).

Of importance in stable population theory is the mean length of a generation, which is the time it takes a stable population to grow by the factor *NRR*. It is represented in stable population theory by the symbol T and is given by $T = (\ln NRR)/r$, where r is the intrinsic growth rate of a stable population (see *Population Dynamics: Theory of Stable Populations*).

See also: Demographic Techniques: Data Adjustment and Correction; Demographic Techniques: Family Reconstitution

Bibliography

Bongaarts J, Feeney G 1998 On the quantum and tempo of fertility. *Population and Development Review* **24**: 271–91

Borgan O, Hoem J M 1988 Demographic reproduction rates and the estimation of an expected total count per person in an open population. *Journal of the American Statistical Association* **83**: 886–91

Brass W, Juarez F 1983 Censored parity progression ratios from birth histories. *Asia and Pacific Census Forum* **10**: 5–13

Feeney G, Yu J-Y 1987 Period parity progression measures of fertility in China. *Population Studies* **41**: 77–102

Henry L 1953 *Fécondité des mariages. Nouvelle méthode de mesure*. INED, Paris (English translation 1980 *Fertility of Marriage: A New Method of Measurement*. UN/ESCAP Population Translation Series No. 3)

Hoem J M 1987 Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review. *International Statistical Review* **55**: 119–52

Ní Bhrolcháin M 1987 Period parity progression ratios and birth intervals in England and Wales, 1941–1971. *Population Studies* **41**: 103–25

Ní Bhrolcháin M 1992 Period paramount? A critique of the cohort approach to fertility. *Population and Development Review* **18**: 599–629

Preston S, Heuveline P, Guillot M 2000 *Demography. Measuring and Modelling Population Processes*. Oxford University Press, Oxford, UK

Rallu J-L, Toulemon L 1994 Period fertility measures: The construction of different indices and their application to France, 1946–89. *Population (An English Selection)* **6**: 59–93

Shryock H S, Siegel J S 1976 *The Methods and Materials of Demography. Condensed Version*. Academic Press, New York

M. Ní Bhrolcháin

Demographic Measurement: Nuptiality, Mortality, Migration, and Growth

In the article *Demographic Measurement: General Issues and Measures of Fertility*, we saw that demographic measures are required for comparisons through time and space, and between subpopulations. Well-specified measures are needed because the rate of demographic events varies particularly by age, and with other dimensions of personal time, and because populations vary in size and in composition with respect to age and other factors influencing demographic event rates. *Synthetic* or *hypothetical cohort* measures were introduced, their purpose being to summarize an array of specific rates pertaining to a period and to estimate the (hypothetical) lifetime consequences of the set of rates obtaining in a calendar period. Synthetic cohort measures may be calculated either additively or multiplicatively and the merits and demerits of these procedures are presented in the aforementioned article.

Let us now consider some further types of demographic behavior.

1. Marriage, Cohabitation, and Union Disruption/Divorce

Measures of marriage and divorce are much less numerous than are those of fertility. We consider them under the headings of level and timing.

1.1 Level

The most basic of the 'flow' indicators are the *crude marriage rate*, defined as the number of marriages per 1,000 population, and the *general marriage rate*, defined as the number of marriages per 1,000 population aged 15 (say) and above (those eligible to marry; the legally allowable minimum age at marriage varies between countries). These rates are influenced both by the structure of a population across age and sex (since marriage rates vary by both age and sex), and by the proportions already married. To deal with the second of these, a general marriage rate may be defined per 1,000 *unmarried* persons. Age–sex variation in rates is handled in the conventional way by defining marriage rates specific by age and sex. Age–sex specific marriage rates may be refined by restricting the denominator to the unmarried. Further detail may also be introduced by calculating marriage rates specific by order of marriage. Such rates can be either of type 1 or 2, as outlined in *Demographic Measurement: General Issues*

and Measures of Fertility. (This terminology is not in widespread use in English.) Thus, the numerator of age–sex specific first marriage rates is confined to first marriages and the denominator may include either single (i.e., never-married) persons (type 1) or all persons (type 2) in that age–sex group. Remarriage rates are generally of type 1, i.e., they express the number of remarriages (by age and sex) per 1,000 divorced and widowed persons, but they too can be calculated in type 2 form (remarriages per 1,000 persons in an age–sex group, irrespective of marital status). All rates may be calculated on either a period or a cohort basis. The crude and general marriage rates can be standardized for age, using conventional standardization methods.

As an adjunct to, or sometimes in place of, period or cohort marriage rates, the cross-sectional distribution of marital status by age and sex may be used to depict the nuptiality of a population. Such figures have the advantage that they can be obtained from any census or survey in which individuals' current marital status is recorded and so do not require vital registration information or annual population estimates by age and sex. They reflect the past marriage, divorce, and widow(er)hood rates of the various age groups and so do not represent current propensities in the population concerned. However, if detailed marriage histories are collected in a survey, or even only the date of first marriage, marriage rates in preceding years may be reconstructed for what will usually be relatively short time series. Retrospective reports may, of course, be inaccurate.

Measuring the frequency of nonmarital cohabitation and of other types of informal partnerships, such as visiting unions, is less straightforward and hitherto there are no widely-agreed standardized methods of doing so (see Murphy 2000 for a discussion of some methodological issues). Because such partnerships are, by their very nature, not recorded in vital registration systems, current status information and retrospective questions in cross-sectional censuses and surveys are a primary source of information. In developed nations particularly, where the prevalence of cohabitation has increased substantially in the last three decades or so, and in areas such as the Caribbean, where visiting unions have long been common, the absence of good quality information on informal unions limits the accurate description of the *de facto* union status of a population. Among the indicators commonly used to evaluate the frequency of informal unions are: the proportion (of all or of the unmarried) currently cohabiting, the proportion of those married (or married at some time) who cohabited with their marriage partner (or with any partner) prior to marriage, and the proportion of informal unions among all unions, formal and informal—all of these being more useful if classified by age and sex. Where suitable data are available, the proportion ever having lived in an informal union by specified ages may be calculated. Analogous measures may be specified for visiting unions.

The *crude divorce rate* is analogous to other crude rates, being defined as the number of divorces per 1,000 persons in a population. It is influenced not only by the age–sex structure of a population but also by the proportions married by age and sex, and also by the distribution of marriages by duration, since divorce rates are closely tied to duration of marriage. Where data are available, a slightly more detailed measure is to be preferred: divorces per 1,000 existing marriages in a population. (Note that this is not the same as the number of divorces occurring in a year per 1,000 marriages *occurring* in a year, a measure that is neither recommended nor used by demographers.) As with other events, divorce rates are more informative if specific by age and sex (divorces in an age–sex group per 1,000 married persons in that age–sex group). However, in specifying divorce rates, duration of marriage is usually preferred to age since divorce risk varies substantially by marriage duration. Divorce rates may, if data allow, be specific by both age at marriage and duration of marriage, as well as by sex, since the risk of divorce is generally higher among those who marry at younger ages.

Where birth cohort information is available, the cumulative proportions ever having married by specified ages are useful indicators of both the level and the pattern of marriage within and across generations. Birth cohort proportions ever having divorced or ever having remarried can be obtained for the same purpose, though in this case the figures may be influenced by the cohort proportions ever having married or ever having divorced, respectively. To deal with this issue, it may be more appropriate in some circumstances to calculate the proportions ever having divorced (by duration of marriage) for marriage cohorts, rather than birth cohorts, and to obtain the proportions ever having remarried for divorce cohorts.

1.1.1 Synthetic or hypothetical cohort measures. The *total marriage rate* is often used to summarize the overall level of marriage for each sex in a population. It is obtained by adding the (type 2) age–sex specific marriage rates and would, if age–sex specific marriage rates were constant through time, represent the average number of marriages a woman or man would experience in a lifetime. The *total first marriage rate* is defined similarly as the sum of the age–sex specific (type 2) first marriage rates and can be interpreted as the proportion of women or men who would eventually marry, if age–sex specific first marriage rates were to remain fixed. Both of these summary measures have the same strengths and weaknesses as the conventional (additive) total fertility rate (TFR), to which they are analogous (see *Demographic Measurement: General Issues and*

Table 1
Selected measures of marriage and divorce

Measure and *time reference*	Definition	Notes
Level Crude marriage rate *Period*	$\frac{M}{P} \times 1000$	M = number of marriages during a specified year or period. P = total person-years lived[a] during a year or period.
General marriage rate *Period*	$\frac{M}{P_{15+}} \times 1000$	P_{15+} = person years lived[a] at ages 15 and above during a year or period[b].
Age–sex specific marriage rate (type 2) *Period; cohort*	$N_x^f = \frac{M_x^f}{P^f} \times 1000$	M_x^f = women marrying at age x during a year. P_x^f = woman-years lived[a] at age x or in age group x during a year or period[b]. Analogous definition for men.
Age–sex specific first marriage rate (type 2) *Period; cohort*	$\frac{M_{x,1}^f}{P^f} \times 1000$	$M_{x,1}^f$ = single women aged x or in age group x marrying during a year or period[b]. Analogous definition for men.
Age–sex specific first marriage rate (type 1) *Period; cohort*	$\frac{M_{x,1}^f}{P_{x,u}^f} \times 1000$	$P_{x,u}^f$ = woman years lived[a] while single and aged x in age group x during a year or period[b]. Analogous definition for men.
Total marriage rate *Period*	$\sum_{15+} \frac{M_x^f}{P_x^f}$	Where age is in five-year groups, the sum is multiplied by five. Analogous definition for men.
Total first marriage rate *Period*	$\sum_{15+} \frac{M_{x,1}}{P_x^f}$	Analogous definition for men.
Gross nuptiality *Period; cohort*	$1-\prod_{x=15+}(1-n_x^f)$	n_x^f = probability that a single woman aged x exactly will marry before $(x+1)$th birthday. Age range used may vary. Analogous definition for men.
Net nuptiality *Period; cohort*	Cumulative proportions ever marrying in a double decrement life table with marriage and mortality as decrements	Age range used may vary. Defined for males and for females.

Measures of Fertility). On the one hand, they are standardized for age and so are an improvement on the crude marriage rate. However, like the conventional TFR they are influenced by timing changes and they do not take account of past history, since they are not standardized for marital status. As a result, a period-specific total first marriage rate can, and has been observed to, exceed one first marriage per person—an impossible result, if it is interpreted as applying to a cohort.

Again, multiplicative summary indicators are available to overcome the deficiencies of the additive

Table 1
Continued

Measure and *time reference*	Definition	Notes
Timing		
Mean age at marriage (crude) *Period; cohort*	$\frac{\Sigma_{15+} x M_x^f}{\Sigma M_x} + 0.5$	Analogous definition for men.
Mean age at marriage (standardized) *Period*	$\frac{\Sigma x N_x^f}{\Sigma N_x^f} + 0.5$	Analogous definition for men.
Singulate mean age at marriage (SMAM) *Period*	$\frac{5 \sum_{0(5)5}^{45} S_x^f - 50 S_{50}^f}{1 - S_{50}^f}$ or assuming no marriage prior to age 15, $15 + 5 \frac{\sum_{15(5)5} S_x^f - 35 S_{50}^f}{1 - S_{50}^f}$	${}_5S_x^f$ = the proportion of women single in the five-year age group x to $x+5$. S_{50}^f is estimated by the mean of the proportions single in age groups 45–49 and 50–54. Analogous definition for men.

Notes: a In specifying demographic rates, person-years lived, woman-years lived and man-years lived (whether overall, or at a particular age or in a given state) is the true denominator. For practical purposes, person years lived in a period is estimated in various ways by different statistical agencies. In the UK and USA, it is usually estimated by the mid-year population (overall, or of a particular age etc.). See also *Demographic Techniques: Rates of the First and Second Kind.* b Where the rate relates to a year or period, events and person years lived are those occurring during the year or period in question. A cohort rate is based on events experienced and person years lived by the cohort concerned.

measures. Two such summaries of the level of marriage in a population are *gross nuptiality* and *net nuptiality*. Gross nuptiality is so called because it assumes that all survive to the latest age considered (usually 50, but sometimes an arbitrarily chosen later age). It is obtained, for each sex, by assembling the age-specific first marriage probabilities (type 1) of the unmarried as a life table and computing the proportions ever marrying (not 'surviving') by 50 (or an alternative later age), according to normal life table procedures (see *Life Table*). Net nuptiality, on the other hand, takes account of the fact that death may intervene before marriage, and thus that marriage and death while single are competing events for single persons. It is obtained by constructing a double-decrement life table, with marriage as one source of decrement and death while single as a second source of decrement (see *Life Table*). Gross nuptiality may be interpreted as the proportion who would marry if they were to experience at each age the age-specific probabilities of marriage pertaining to a given period and if there were no mortality before age 50 (or by the upper age limit chosen). Net nuptiality, on the other hand, may be taken to represent the proportion of a cohort who would marry by 50 (or some other specified age), if they were to experience, while single and at each age, the probabilities of marriage and of death pertaining to the single in a particular period. A practical difficulty with both gross and net nuptiality is that they require more detailed information than is needed to obtain the total marriage rate—namely first marriages by age and sex together with population estimates by age, sex, and marital status in the case of gross nuptiality, and additionally, mortality rates by sex, age, and marital status in the case of net nuptiality. As with life table proportions 'surviving' or 'not surviving' in general, they are standardized for age structure.

The level of divorce in a population may also be summarized through hypothetical cohort indicators. Analogous to the *total marriage rate* is the male or female *total divorce rate*, defined as the sum of age–sex specific type 2 divorce rates and representing the average number of divorces per person if the age–sex specific divorce rates of the period in question were to remain fixed or were to be experienced throughout a person's lifetime. As before, a multiplicative life table based indicator is to be preferred to the additive measure. The most common indicator used for the purpose is a life table estimate of the proportions who would divorce by a specified duration of marriage, at the rates current at each marriage duration, if these divorce rates were to be experienced by a couple throughout their marriage. As with analogous multiplicative indicators, this standardizes for the distribution by marriage duration and also takes some account of past history. This is probably the best available summary measure of the divorce propensity implied by current divorce rates, but it is demanding of data and is as subject to biases resulting from timing changes as are analogous measures of fertility and marriage. If successive marriage cohorts are moving to an earlier pattern of divorce, the high divorce rates of

early-divorcing couples may be combined with the high divorce rates of later-divorcing couples who had lower divorce rates at early durations, to produce an overall hypothetical proportion ever divorcing that is higher than that of any real cohort. At present, there is no specific name in use for this summary indicator.

1.2 Timing

Age at marriage may be calculated from period data in either crude or standardized form, though sources do not always indicate which measure is presented and, as with births, the terms 'crude' and 'standardized' are not in universal use in this context. The *crude mean age at marriage* is simply the mean age of those marrying in a period. The *standardized mean age at marriage* is analogous to the standardized mean age at birth and is the mean of the schedule of age-specific marriage rates (type 2). A crude and standardized *median age at marriage* may also be obtained. It is often useful to distinguish marriage age by the order of the marriage since the proportion of remarriages among all marriages can vary substantially through time or cross-nationally. Thus, we have the *mean* or *median age at first marriage* or *at remarriage*, either crude (of all those marrying for the first time or for the second or later time) or standardized (using age-specific first or remarriage rates of type 2). Equivalent averages can be calculated for birth cohorts and, as usual, do not require standardization. Mean and median ages at marriage can also be obtained from either gross or net nuptiality tables. Care is needed since the e_x column of a nuptiality table does not give the expected age at or years to marriage. Rather the column gives the expected number of years lived *while single* at and after age x. Since some never marry, this is not the same as the expected number of years to marriage among those who marry (see *Life Table*).

Average age at marriage can be estimated from cross-sectional proportions single (never married) by age, by an indirect procedure developed by Hajnal (1953). It gives what is termed the *singulate mean age at marriage* (SMAM). Based on the hypothetical cohort principle, the procedure is appropriate where marriage rates by age have been reasonably stable and assumes that neither migration nor mortality rates are associated with marital status. Should the proportions single be found to increase over any part of the age range, this indicates that one or more of the assumptions does not hold and so the calculation is not valid. Details are given in Table 1.

Mean and median ages at divorce may be calculated. Of particular interest in relation to divorce is the mean/median etc. duration of marriage at which the divorce occurs. A distinction should be made between the duration of marriages in general (those terminated by either divorce or death) and the duration of those marriages that end in divorce. Crude or standardized versions of these indicators may be obtained; for most purposes, standardized indicators are to be preferred.

2. Mortality

The *crude death rate* (CDR)—deaths per 1,000 population—is widely used as a basic measure of the level of mortality (see Table 2 for selected mortality measures). It is particularly strongly influenced by the age composition of a population because of the wide variation in death rates by age. *Age-specific mortality rates*—deaths at a particular age per 1,000 population of that age—are at the next level of detail, usually specific by sex. Where good quality information is available, rates that are specific for single years of age are produced, in preference to five-year age groups. Rates specific by single years of age are of particular importance at the younger and older parts of the age range, since in these age ranges differences in rates between adjacent single years of age may be sizable. Age may be grouped in a variety of ways, according to the refinement required in analysis and the nature of available data. However, even where five-year groups are used, the age groups zero to one and one to four are often distinguished, because of the relatively high death rate of infants, even in low mortality countries.

The measurement of mortality is simplified to some extent by the fact that death occurs ultimately to all and occurs only once. These twin aspects mean that crude and age-specific death rates are necessarily of type 1 (see *Demographic Measurement: General Issues and Measures of Fertility*): the risk of death is universal and, death occurring only once in a lifetime, nobody currently at risk of death can have experienced the event previously. Death rates are so closely and systematically associated with age, and to a lesser extent sex, that specificity by age and sex is adequate for most demographic purposes. Other dimensions of personal time are required much less often in specifying death rates. However, mortality rates may also be specific by social class, race/ethnicity, income, area of residence (urban/rural or subnational region), marital status, and other such socioeconomic characteristics. Analysis by such factors is of interest in examining social and economic variations in health and for actuarial and public health purposes. With such subspecifications, death rates may also be of type 2.

The *infant mortality rate* (IMR) is reported and published extensively, partly as a demographic indicator but also as an indicator of socioeconomic development. It is defined as the number of deaths of infants (children under the age of one year) per 1,000 births in a given year. It is, thus, not a rate as normally specified since the numerator is not an estimate of the person-years at risk of the event and will be inaccurate for this purpose to the extent that there are year-on-

Table 2
Selected mortality measures

Measure	Definition	Notes
Crude death rate (CDR) *Period*	$\frac{D}{P} \times 1000$	D = the number of deaths occurring during a particular year/period. P = person years lived[a] during the year or period.
Age-specific death rate *Period; cohort*	$M_x = \frac{D_x}{P_x} \times 1000$	D_x = deaths at age x during a year or period[b]. P_x = person years lived[a] at age x or in age group x during a year or period[b].
Infant mortality rate (IMR) *Period; cohort*	$\frac{D_0}{B}$	D_0 = deaths of infants under 12 months during a year or period[b]. B = births during a year or period[b]
Infant death rate *Period; cohort*	$\frac{D_0}{P_0}$	P_0 = person years lived[a] at ages under one year during a year or period[b].
Maternal mortality ratio (sometimes referred to as the maternal mortality rate) *Period; cohort*	$\frac{D_m}{B}$	D_m = maternal deaths during a year or period[b].
Maternal mortality rate *Period; cohort*	$\frac{D_m}{P^f_{15-49}}$	P^f_{15-49} = woman years lived[a] at ages 15–49 during a year or period[b] (age limits may vary).

Notes: a In specifying demographic rates, person-years lived, woman-years lived and man-years lived (whether overall, or at a particular age or in a given state) is the true denominator. For practical purposes, person years lived in a period is estimated in various ways by different statistical agencies. In the UK and USA, it is usually estimated by the mid-year population (overall, or of a particular age etc.). See also *Demographic Techniques: Rates of the First and Second Kind.* b Where the rate relates to a year or period, events and person years lived are those occurring during the year or period in question. A cohort rate is based on events experienced and person years lived by the cohort concerned.

year fluctuations in births. It is, however, a convenient measure since it can be obtained from simple counts of vital events and does not require population estimates by age. It has a correctly specified counterpart in the *infant death rate*, the number of deaths of infants (under the age of one year) in a year per 1,000 person-years lived under age one. Several mortality rates are distinguished also during the first year of life, both because of the steep decline in death rates over the first 12 months of life and because of the changing role of endogenous (genetic, intrauterine, perinatal) and exogenous (environmental, external) causes of death. Endogenous causes predominate in the earliest post-partum period, with exogenous factors growing in importance thereafter. The *perinatal mortality rate* is defined as the number of late fetal deaths plus the number of deaths within one week of birth per 1,000 total late fetal deaths plus live births. The *neonatal mortality rate* is defined as the number of deaths within one month (28 days) of birth per 1,000 live births, and the *post-neonatal mortality rate* as deaths between 28 days of birth and one year of age per 1,000 live births. The probability of dying between birth and the fifth birthday is widely referred to in the current demographic literature as the *child mortality rate*, though the latter is more correctly defined as the number of deaths of children under the age of five years per 1,000 person-years in the age group. The probability definition (also sometimes termed *child mortality risk*) has become widespread because under-five mortality is of particular interest in high-mortality, less-developed societies, for which the indirect methods (see *Demographic Techniques: Indirect Estimation*) used to evaluate child mortality estimate a cohort probability rather than a rate. High-mortality societies do not usually have good vital registration statistics and it can be difficult to formulate accurate assumptions about the average number of years lived by those who die, and thus to obtain an estimate of person-years lived.

Maternal deaths are those that occur during pregnancy or within 42 days of the end of the pregnancy, due to a cause related to the pregnancy or a condition

aggravated by pregnancy; the term thus includes abortion-related deaths. Maternal mortality is most often represented by the *maternal mortality ratio*: the number of maternal deaths per 100,000 live births in a period. This indicator is referred to by many authors and in many sources as the *maternal mortality rate*, although it is not in fact a true rate. Strictly, the *maternal mortality rate* is the number of maternal deaths per 100,000 women of reproductive age in a period, and some authors employ the term in this sense. It should be clear from the context which definition is in use.

Mortality specific by cause of death, of which maternal mortality is an instance, is the principal way in which mortality indicators are disaggregated beyond age and sex. The International Classification of Disease, currently produced and revised from time to time by the World Health Organization, is the standard scheme by which cause of death is classified. Measures of mortality by cause of death are of two kinds. Cause-specific mortality rates refer to the number of deaths from a particular cause during a period per 100,000 population, and may be specific also by age and/or sex. A multiplier larger than 1,000 is normally used for the purpose, since deaths from any one cause will usually be relatively small in number. Cause-specific mortality ratios, on the other hand, relate to the percentage of all deaths that result from a particular cause, either overall or by age and sex, and are of value where information on population denominators is inaccurate or unavailable. Proportional mortality analysis is used particularly in the study of occupational mortality.

2.1 Age Standardization

Because of the substantial variation in death rates by age (and in age structure across time and place), cross-national, areal, or time comparisons of the level of mortality require that death rates are standardized for age. Direct standardization, in which a standard age distribution is employed, yields a *directly standardized death rate*. Indirect standardization, in which a standard set of age-specific mortality rates is chosen, yields two indicators: the *standardized mortality ratio* (SMR) and the *indirectly standardized death rate*. The SMR gives an index figure expressing the level of mortality in the index population relative to that of the standard population, set usually to 100: it is simply the ratio of observed to expected deaths in the index population, with expected deaths based on the standard population's age-specific rates. Multiplying the SMR by the CDR in the standard population gives the *indirectly standardized death rate* in the index population. Direct standardization requires that the age-specific mortality rates are known and precisely estimated in the populations to be standardized. Where the age-specific rates are either unknown or subject to sizable errors, the indirect method must be used. Statistical theory singles out a version of indirect standardization as optimal when a multiplicative model holds for the force of mortality.

2.2 Life Expectancy and Life Tables

Life expectancy at birth, e_0, can be considered both an indicator of the 'timing' of mortality and also a summary of the overall level of mortality in a population. It measures timing in that it represents the expected or average number of years lived by a person or group subject, throughout their lifetime, to a given set of age-specific mortality rates. Because death occurs universally and removes an individual permanently from the population, any increase in death rates at any age is necessarily reflected in a shorter life expectancy, while a decline in death rates at any age is necessarily accompanied by an increase in life expectancy. The expectation of life is influenced both by the level of mortality in a population and by the pattern of mortality by age. As a result, two populations with the same overall life expectancy may differ in their patterns of age-specific death rates. Life expectancy is calculated from a life table (see *Life Table*), and so is standardized for age. The life table in question may be based on the age-specific mortality rates obtained in a calendar period or those of a birth cohort. Where based on a period life table, the expectation of life and other life table functions are synthetic cohort measures and share the shortcomings of all synthetic multiplicative measures (see *Demographic Measurement: General Issues and Measures of Fertility*).

The life table is a central analytical technique in demography, having a large variety of uses. It is, however, fairly demanding of data since it requires information on age-specific death rates and thus on the distribution of deaths by age as well as population estimates by age. Because detailed data of this kind are often unavailable for historical populations and for less developed nations, sets of model life tables have been developed corresponding to varying levels and patterns of mortality (Coale et al. 1983, United Nations 1982). Such compilations allow estimates of the level and/or age pattern of mortality to be made from information that is insufficient in itself to determine these (see *Demographic Techniques: Indirect Estimation*).

3. Migration

Along with births and deaths, migration is the final component of population change. Migration is classified according to whether it is into an area (*in-migration*) or out of an area (*out-migration*). The sum of in- and out-migration is termed *gross migration* and

the difference between them, in-migration minus out-migration, is known as *net migration*. For many purposes it is the absolute values of these quantities that are of interest. Crude and specific rates are also used. The *crude in-migration* (or *immigration*) *rate* is the number of in-migrants per 1,000 population and the *crude out-migration* (or *emigration*) *rate* the number of out-migrants per 1,000 population. The *crude gross migration rate* is the sum of these and the *crude net migration rate* is the difference between them. Like other demographic events, migration is patterned by age and so rates of all four kinds may be specific by age. For example, the *age-specific out-migration rate* is the number of out-migrants of a given age per 1,000 population of that age during a specified period. Cumulative migration experience is represented by whether a person has ever migrated from their place of birth (lifetime migration) or during a specified time (e.g., the last year, the last five years). Migration can be decomposed further according to types of area between which the move was made: for example, between or within urban and rural areas.

4. Population Growth

The difference between the crude birth rate and the crude death rate is known as the *crude rate of natural increase* (CRNI) and represents the rate at which a population is growing or declining in a year purely as a result of vital events. Since migration also influences population growth, the overall population growth rate is CBR–CDR $\pm$ the crude net migration rate. Where population growth rate is inferred from the change in total population during a period of t years, the growth rate is estimated as $r = \ln (P_t/P_0)/t$. The *intrinsic rate of natural increase* is a theoretical quantity, referring to the growth rate of a stable population (see *Population Dynamics: Theory of Stable Populations*) having a fixed set of age-specific mortality and fertility rates. It can be calculated for a real population, but should be interpreted with care since it represents the growth rate that would be obtained in the hypothetical circumstance that current age-specific rates were to remain fixed and that the population were closed to migration. The difference between the actual population growth rate and the intrinsic rate of natural increase is between the observed growth rate and the hypothetical growth rate that would apply if current vital rates were to obtain in a stable population. Further indicators relevant to the theoretical analysis of population growth are outlined in the article *Demographic Measurement: General Issues and Measures of Fertility*: the gross and net reproduction rates (GRR and NRR; see also *Population Dynamics: Theory of Stable Populations*). While the GRR and NRR may be calculated for actual (empirical) populations, an important proviso must be borne in mind: since the NRR accurately reflects growth prospects only in populations that are stable (see *Population Dynamics: Theory of Stable Populations*), when calculated for an actual population it takes no account of prospective growth/decline resulting either from migration or from future changes in mortality, nor does it take account of the growth potential inherent in population age structure (see *Population Dynamics: Momentum of Population Growth*). The same is true of the *total fertility rate* (TFR). A TFR of 2.1 in low-mortality countries is widely referred to in the demographic literature as 'replacement level fertility,' that is, the level of fertility necessary for population replacement in the long run. But this is potentially misleading. Low-mortality populations with a TFR of below 2.1 can and do continue to grow because of in-migration, declines in mortality, and/or population momentum. Equally, low-mortality populations with a TFR of 2.1 or more can and do decline because of out-migration, changes in mortality, and/or population momentum.

See also: Cohort Analysis; Migration, Theory of

Bibliography

Alderson M R 1982 *Mortality, Morbidity and Health Statistics*. Macmillan, Basingstoke, UK

Coale A J, Demeny P, Vaughn B 1983 *Regional Model Life Tables and Stable Populations*, 2nd edn. Academic Press, New York

Hajnal J 1953 Age at marriage and proportions marrying. *Population Studies* **7**: 111–36

Murphy M 2000 The evolution of cohabitation in Britain, 1960–95. *Population Studies* **54**: 43–56

Preston S, Heuveline P, Guillot M 2000 *Demography. Measuring and Modelling Population Processes*. Oxford University Press, Oxford, UK

Shryock H S, Siegel J S 1976 *The Methods and Materials of Demography*, condensed edn. Academic Press, New York

United Nations 1982 *Model Life Tables for Developing Countries*. Population Studies Series No. 77, United Nations, New York

M. Ní Bhrolcháin

Demographic Models

The term 'demographic models' can have two meanings, one broad and one narrow. In its broad meaning, demographic models refer to all mathematical, statistical, forecast, and microsimulation models that are applied to studies of demographic phenomena. In its narrow meaning, demographic models refer to empirical regularities in age patterns of demographic events. Demographic models in the broad definition can be found in various entries related to demography (see, among others, *Demographic Analysis: Probabilistic Approach*; *Multistate Transition Models in*

Demography; *Event History Analysis: Applications*; *Population Dynamics: Theory of Stable Populations*; *Population Forecasts*; *Microsimulation in Demographic Research*). This article is concerned with demographic models in the narrow definition.

1. Age Patterns of Demographic Events

In a classic statement, Hauser and Duncan (1959) defined demography as 'the study of the size, territorial distribution, and composition of population, changes therein, and the components of such changes' (p. 2). In the Hauser–Duncan definition of demography, the study of population changes goes hand in hand with the study of the components of population changes. This is necessitated by the need to decompose a population into components and then study changes in the components before arriving at an overall understanding of the changes in the population. The most elementary, and also the most important, form of population decomposition is by sex and age. Since men are only indirectly involved in reproduction, demographic analysis is often simplified by focussing on women. Such simplification is called 'one-sex' modeling (see *Population Dynamics: Two-sex Demographic Models*), as is commonly seen in fertility and nuptiality models. In mortality models, however, men and women are always kept distinct, and there are substantial mortality differentials in favor of women.

The treatment of age is of significant concern in demographic research. Without exception, the occurrence of all demographic events is age-dependent. Here, the correct interpretation of age-dependency is one of life-course, that is, the likelihood of the occurrence of an event changes as a person (or a cohort) ages. This is true even though most demographic methods and models use cross-sectional data, capitalizing on age-gradients of vital rates in any given population. The use of period-based data usually is necessitated by the lack of cohort-based data.

Demographic models of age schedules are developed on the observation that age patterns of demographic events often show some regularity. Two cautionary notes are in order. First, age regularity is not universal, either across space or over time. Second, all that is assumed is empirical regularity; theoretical reasons behind such regularities are typically neither well established nor well understood. Nonetheless, demographic models capitalizing on empirical regularities are very useful in practice and may provide the basis for theoretical work. A brief discussion of the main uses of demographic models is given as follows.

2. Use of Demographic Models

Demographic models are intended to summarize empirical regularities in age patterns of demographic events, ideally in simple mathematical formulas. Such models can prove very useful in demographic research. One type of use of demographic models concerns data quality. For example, demographic models may be used to detect and correct faulty data, impute missing data, and allow researchers to infer from partial data. Not surprisingly, this use of demographic models is often found in historical demography and research in less developed countries, where quality data are scarce.

The second type of use is substantive and commonly it is found in research with a comparative focus, be it over time (trends) or across societies, regions, or subpopulations. Age-schedule models are also useful for actuarial calculations of life-insurance premiums and premium reserves and for the computation of national or regional population forecasts (e.g., Lee and Tuljapurkar 1994). In demographic applications, cross-group variations in age patterns are typically parameterized as functions of two components: a shape component to capture age effects and a modification component reflecting group membership. Ideally, we would like the second component to be as parsimonious as possible. When demographic models are fully parametric, the two components are integrated, with the functional form being the shape component and the parameter values being the modification component. However, parametric models are not the norm in demography. Parametric models and semi-parametric models will be discussed separately below, with concrete examples taken from research on mortality, nuptiality, and fertility. Coale and Trussell (1996) give a more detailed account of some of the models.

3. Parametric Models

Gompertz (1825) is accredited for the discovery that mortality rates in human populations increase nearly as an exponential function of age. This regularity holds true only after early adulthood (ages 25–30). According to 'Gompertz's law,' the force of mortality (or hazard) is a parsimonious loglinear function of age

$$\ln(\mu_x) = A + Bx \tag{1}$$

where μ_x denotes the force of mortality at age x. Naturally, A is the parameter characterizing the level of mortality at early ages and B is the parameter characterizing the rate of increase in mortality with age. While very parsimonious, the Gompertz model does not always fit empirical data well. Other researchers have modified the model either by adding additional terms or by changing the functional form. For example, Makeham (1867) altered the simple loglinear relationship of Eqn. (1) by adding a time-invariant term intended to capture cause-specific 'partial forces of mortality.' Heligman and Pollard's (1980) three-component model is another such extension.

In studying age patterns of first marriages among women, Coale (1971) found a basic form that is common in different populations. Through trial-and-

error, Coale was able to characterize the typical age pattern of first marriage by a double-exponential curve

$$r_x = (0.174)\exp[-4.411\exp(-0.309x)] \quad (2)$$

where r_x denotes the hazard rate (or risk) of first marriage at age x. This mathematical function was shown later to be approximated closely by the density function corresponding to the sum of distributions representing demographic processes (Coale and McNeil 1972).

To fit observed age patterns of fertility, Brass (1960) proposed a model using polynomial functions. Although Brass' parametric model is very flexible, it requires the estimation of four unknown parameters. Since fertility rates typically are given in a limited number of 5-year intervals, this method leaves very few degrees of freedom for evaluating goodness-of-fit. An evaluation of various parametric models based on empirical data is given by Hoem et al. (1981).

There are three notable problems with parametric models. First, parametric models often do not fit observed phenomena. Demographers have dealt with such empirical deviations through (a) restricting the applicable age range and (b) allowing for further parameterization. Both solutions have surfaced in research using the Gompertz model.

The second major disadvantage associated with parametric models is the lack of behavioral interpretations for key parameters. This problem arises from the fact that almost all parametric models have resulted from exercises of curve-fitting. Parametric models may reproduce observed age patterns of demographic events, but theoretical interpretation of involved parameters is often unclear.

Finally, parametric models are not always convenient to use for comparing populations or subpopulations even though this was one of the motivations for developing them in the first place. This problem is apparent, for example, in the case of a polynomial model. When several parameters in a polynomial function differ between two populations, it is difficult to characterize one population as having higher or lower rates than the other population.

4. Semiparametric Models

In response to these problems with parametric models, semiparametric models have been developed. Semiparametric models are similar to parametric models in specifying parsimonious mathematical functions but differ from parametric models in allowing age-dependency to be unconstrained and subject to empirical estimation, that is, semiparametric models do not impose any global constraint limiting the age pattern to the rigidity of a parametric mathematical function. Instead, the age pattern is estimated freely and empirically from observed data or calculated from external sources. One manifestation of the semiparametric approach is the use of model schedules or model tables. While allowing for the flexibility in a common age function, model tables place constraints on the variations in the age pattern across populations or subpopulations. Such constraints are often motivated by substantive knowledge of demographic phenomena.

In the area of mortality studies, for example, model life tables have been in wide use, primarily as a tool for correcting faulty data and estimating missing data. In essence, a model life table allows for a typical age schedule of mortality shared by a set of populations that differ mainly in their levels of mortality. The age pattern is flexible and empirically determined over all ages but constrained across populations; and cross-population variability lies in the overall level of mortality. The earliest model life table was developed by the United Nations (1955) for all national populations. Subsequently, Coale and Demeny (1966) added more flexibility by identifying four regional model life tables based on distinctive age patterns of mortality and refined the technique for constructing model life tables.

As an alternative, Brass (1968) developed a system of relational life tables, in which a standard mortality schedule is modified by two parameters through a logit equation

$$\text{logit}(p_{xj}) = \alpha_j + \beta_j \,\text{logit}(p_x^*) \quad (3)$$

Here p_{xj} is the probability of survival to age x in population j, p_x^* is the probability of survival to age x in a selected standard table, and α_j and β_j are parameters characterizing population j. In the Brass system, α determines the level of survival, whereas β changes the shape of survival parametrically.

In fertility research, Henry (1961) posited the existence of natural fertility with a peculiar age pattern. According to him, for physiological and social factors, different populations not exercising fertility control may have different levels of fertility but exhibit the same age pattern. Coale and Trussell (1974) later formalized this idea into the *Mm* method for studying age patterns of marital fertility. Let the marital fertility rate for the xth age of the jth population be denoted by r_{xj}. The Coale–Trussell *Mm* method parsimoniously describes the cross-population variation in age patterns of marital fertility with

$$r_{xj} = n_x M_j \exp(m_j v_x) \quad (4)$$

Here n_x is the standard age pattern of natural fertility; v_x is a typical age-specific deviation of controlled fertility from natural fertility, and M_j and m_j measure the jth population's fertility level and control, respectively. Eqn. (4) then states that marital fertility can be modeled as a product of natural fertility and a factor reflecting fertility control. The former is represented by $n_x M_j$; the latter by $\exp(m_j v_x)$. Coale and Trussell estimated n and v and suggested that these age schedules be taken as known in all applications. This approach has been superseded by Xie (1990) and Xie

and Pimentel (1992), who suggested that n and v be regarded as parameters to be estimated along with M and m from any set of empirical fertility rates r_{xj}.

5. Statistical Demography vs. Mathematical Demography

The construction and derivation of parametric and semiparametric demographic model age schedules have been accomplished through mostly mathematical and sometimes graphical methods. From the outset demographic modeling has been regarded as part of mathematical demography. Since the 1980s, however, a mainline statistical approach has played an increasingly important role in the development of demographic models (e.g., Clogg and Eliason 1988, Hoem 1987, Xie 1990, Xie and Pimentel 1992; see also *Demographic Analysis: Probabilistic Approach*) for a number of good reasons.

First, the advancement of demography has brought with it more, richer, and better data in the form of sample data, and the use of sample data requires statistical tools. Treating sample data as exactly known quantities induces the danger of contamination by sampling errors. Second, while the method of disaggregation commonly used in mathematical demography may easily lead to inaccurate estimation due to small group sizes, statistical methods more efficiently utilize covariates to examine group differences. Third, because observed data may sometimes be irregular or simply missing, statistical models can help smooth or impute data. Conversely, the strength of empirical techniques developed around demographic models (e.g., indirect estimation, age patterns of fertility, model life tables, etc.) is that they provide descriptions of age patterns that can be utilized to improve statistical analysis.

Bibliography

Brass W 1960 The graduation of fertility distributions by polynomial functions. *Population Studies* **14**: 148–62

Brass W 1968 *The Demography of Tropical Africa*. Princeton University Press, Princeton, NJ

Clogg C C, Eliason S R 1988 A flexible procedure for adjusting rates and proportions, including statistical methods for group comparisons. *American Sociological Review* **53**: 267–83

Coale A J 1971 Age patterns of marriage. *Population Studies* **25**: 193–214

Coale A J, Demeny P 1966 *Regional Model Life Tables and Stable Populations*. Princeton University Press, Princeton, NJ

Coale A J, McNeil D R 1972 The distribution of age of the frequency of first marriage in a female cohort. *Journal of the American Statistical Association* **67**: 743–49

Coale A J, Trussell T J 1974 Model fertility schedules: Variations in age structure of childbearing in human populations. *Population Index* **40**: 185–258

Coale A J, Trussell T J 1996 The development and use of demographic models. *Population Studies* **50**: 469–84

Gompertz B 1825 On the nature of the function expressive of the law of human mortality and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London* **115**(Part II): 513–85

Hauser P M, Duncan O D 1959 *The Study of Population: An Inventory and Appraisal*. University of Chicago Press, Chicago

Heligman L, Pollard J H 1980 The age pattern of mortality. *Journal of the Institute of Actuaries* **107**: 49–80

Henry L 1961 Some data on natural fertility. *Eugenics Quarterly* **8**: 81–91

Hoem J M 1987 Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review. *International Statistical Review* **55**: 119–52

Hoem J M, Madsen D, Nielsen J L, Ohlsen E M, Hansen H O, Rennermalm B 1981 Experiments in modeling recent Danish fertility curves. *Demography* **18**: 231–44

Lee R D, Tuljapurkar S 1994 Stochastic population forecasts for the United States: Beyond high, medium, and low. *Journal of the American Statistical Association* **89**: 1175–89

Makeham W M 1867 On the law of mortality. *Journal of the Institute of Actuaries* **13**: 325–67

United Nations 1955 *Age and Sex Patterns of Mortality: Model Life Tables for Under-developed Countries*. United Nations, New York

Xie Y 1990 What is natural fertility? The remodelling of a concept. *Population Index* **56**: 656–63

Xie Y, Pimentel E F 1992 Age patterns of marital fertility: Revising the Coale–Trussell method. *Journal of the American Statistical Association* **87**: 977–84

Y. Xie

Demographic Techniques: Data Adjustment and Correction

Demographic data suffer to varying degrees from errors of coverage and of classification. Such errors are more common or more serious in populations with lower levels of education and economic development. Errors in the classification of an individual's age and parity (number of live-born children) are the most common and most problematic for demographic measurement and estimation. Methods of data evaluation and correction rely on mathematical formulae or on the relationship between demographic processes (birth, death, and migration) and the age distribution.

1. Errors of Coverage

In censuses, errors of coverage comprise the underenumeration or occasionally overenumeration of the population. Such errors usually vary by age and sex and possibly by other variables. For example, young children are often omitted from enumeration in some countries, as are young unmarried adults temporarily absent for work. Less frequently, migrants may be enumerated twice, at both source and destination. The

completeness of a census enumeration may be assessed through a postenumeration survey. Incomplete coverage is also a common deficiency of vital registration data. Since year of registration rather than time of event is the usual basis of compilation, late reporting adds to the inaccuracies involved. Death registration, where it occurs, is usually fairly timely since registration is often a legal requirement for burial, though in some populations such requirements are often ignored, notably in rural areas.

Nonresponse to individual questions in censuses and surveys also raises issues of representativeness and should be minimized. Responses may sometimes be imputed from the respondent's other characteristics during data editing. For most variables, nonresponses are omitted from analysis, only their extent being reported. For age, however, nonresponses are often distributed on a *pro rata* basis for each sex because of the need to maintain population size in demographic measurement.

2. *Errors of Classification*

Errors of classification may affect all demographic variables and arise from incorrect reporting on the part of the respondent and/or enumerator. Errors may also arise during coding and editing. Errors in the reporting of age (age misreporting) include digit preference (or age heaping) and under- and overstatement.

2.1 *Age*

Digit preference occurs where real age is not known and its estimation involves the 'rounding' of age (or year of birth) to end in a preferred digit such as 0 or 5. Digit preference is best detected graphically (see Fig. 1). Various indices have been devised to measure the extent of digit preference, but they do not offer a means of correction (US Bureau of the Census 1973). Whipple's index measures the preference or avoidance of the digits 0 and 5 over the age interval 23 to 62, though other multidecennial intervals can be used. Myers' blended index measures preference or avoidance of all 10 digits. Since neither index is theoretically precise, Bachi derived a composite index without this deficiency. These indices assume rectangularity or linearity within quinquennial age groups. Later indices, developed by Carrier and by Ramachandran, take account of age structure and variable digit preference over the age range. In practice, the Bachi, Carrier, and Ramachandran indices were laborious and are used less frequently.

Age misstatement may occur with or without digit preference and generally involves a directional tendency often related to cultural factors. Age understatement may occur among women of lower parity than average for their age, especially in cultures with high fertility. Age overstatement often occurs at older ages, particularly among males, in cultures where status is gained with age. Sex-specific age misstatement may also occur, for example, young female children generally are reported as younger than their male peers. The inaccurate reporting of age affects data on both population and vital events.

In many cases, it is impossible to distinguish between age misstatement, differential coverage by age, and digit preference. Further, one error may offset another. The pattern of overall error may be seen by examination of age ratios and sex ratios for 5 or 10-year age groups. These may be compared with ratios from suitable model life table populations. An indication of the overall accuracy of age reporting is given by the United Nations age-sex accuracy index (or 'joint score') (United Nations 1955). Irregularities arising from such factors as past changes in fertility or age-sex-specific migration should be taken into account.

Error may be assessed through comparison with expectation based on external data using demographic relationships. If accuracy permits, a previous enumeration can be 'updated' using data on births, deaths, and net migration to derive an expected age distribution. Comparison of the cumulated age distribution with model stable distributions identifies distortion due to misreporting (United Nations 1983; see *Population Dynamics: Theory of Stable Populations*). The completeness of birth and death registration may be assessed by indirect estimation techniques (see *Demographic Techniques: Indirect Estimation*).

The effects in demographic measurement and estimation of both digit preference and age misstatement can be reduced by the use of wider (than standard 5-year) age intervals, but this is not always practicable and involves greater approximation. Similarly, the effects on standard quinquennial age groups can be reduced by adopting nonstandard age groups, the optimum choice of which is based on Myers's index (US Bureau of the Census 1973), but these are generally inconvenient. For some purposes it may be appropriate to smooth or graduate the age distribution. For indirect estimation involving simultaneous estimation and detection of error (see *Demographic Techniques: Indirect Estimation*), or when data on population and vital events are obtained from a single source (e.g., a vital statistics survey), unadjusted data are required.

Methods of smoothing an age distribution include the use of graphs (manual smoothing), moving averages, and curve-fitting. Mathematical graduation is used to smooth narrower age groups within broader age groups. The following Carrier–Farrag and Karup–King–Newton formulae treat all age groups the same, whereas the Arriaga formula treats the first and last differently (Arriaga et al. 1994). In the following

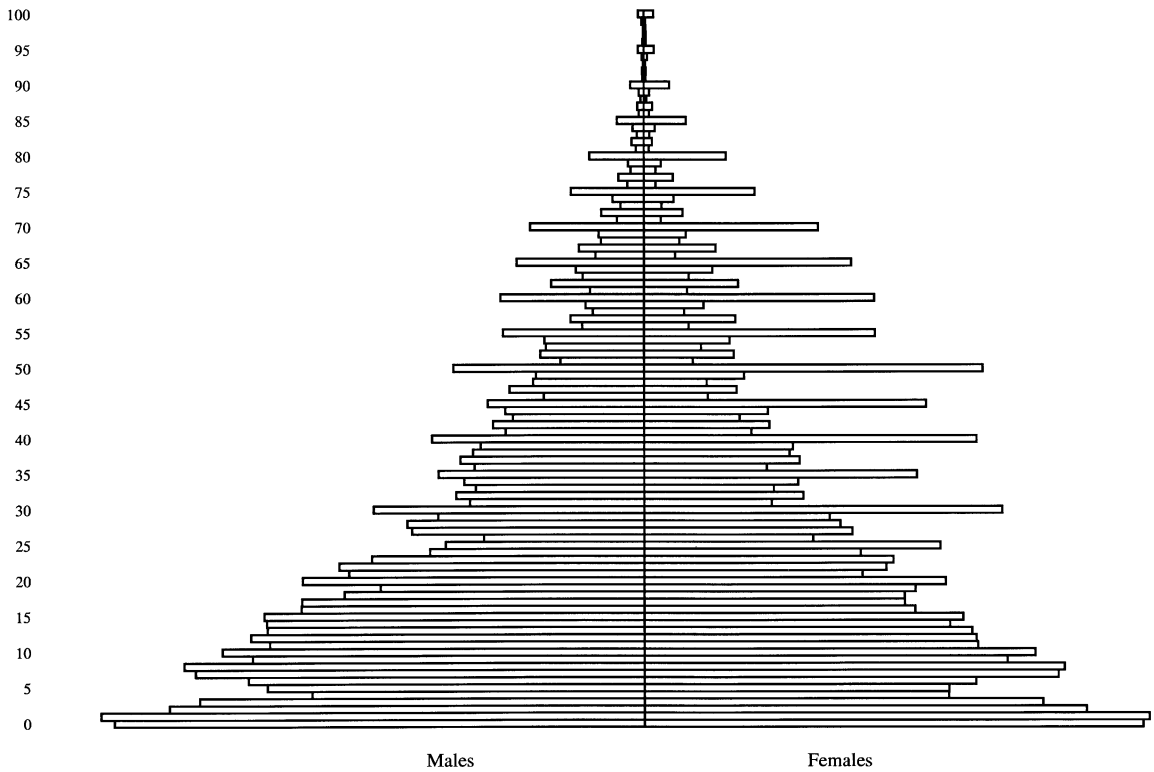

Figure 1
Digit preference in the 1950 Census of Albania

formulae, ${}_nP_x$ represents the population aged x to $x+n-1$, x and n being integers. The Carrier–Farrag ratio method is based on the assumption that the relationship of a 5-year age group to its constituent 10-year age group is an average of such relationships in three consecutive 10-year age groups and is

$${}_5P_{x+5} = {}_{10}P_x/[1+({}_{10}P_{x-10}/{}_{10}P_{x+10})^{1/4}]$$

The Karup–King–Newton formula assumes a quadratic relationship between three consecutive 10-year age groups, that is ${}_5P_x = {}_{10}P_x/2+({}_{10}P_{x-10}-{}_{10}P_{x+10})/16$. The Arriaga formulae are based on second-degree polynomials passing through the midpoints of three consecutive 10-year age groups. Then

$${}_5P_{x+5} = (8\ {}_{10}P_x+5\,{}_{10}P_{x+10}-{}_{10}P_{x+20})/24$$

when the 10-year age group is the first,

$${}_5P_x = (-{}_{10}P_{x-20}+5\ {}_{10}P_{x-10}+8\ {}_{10}P_x)/24$$

when it is the last, and

$${}_5P_{x+5} = (-{}_{10}P_{x-10}+11\ {}_{10}P_x+2\ {}_{10}P_{x+10})/24 \text{ otherwise.}$$

In all cases, the second 5-year age group is derived by subtraction.

When age misreporting is more severe, the totals of all age groups need potential modification through averaging. The United Nations formula lightly smooths 5-year age groups, based on five consecutive 5-year groups, and gives the smoothed population aged x to $x+4$ as

$${}_5P'_x = (-{}_5P_{x-10}+4\ {}_5P_{x-5}+10\ {}_5P_x\ {}_5P_{x+5}-{}_5P_{x+10})/16.$$

For heavier smoothing, three consecutive 10-year age groups are used to derive the central smoothed age group as

$${}_{10}P'_x = ({}_{10}P_{x-10}+2\ {}_{10}P_x+{}_{10}P_{x+10})/4.$$

Neither formula smooths the youngest and oldest age groups. Since smoothing changes overall population size, the smoothed age groups subsequently are adjusted proportionally to achieve their overall enumerated total. The Arriaga (or other such) formulae are then applied to smoothed and adjusted 10-year totals to derive smoothed 5-year age groups.

A single-year distribution may be derived from grouped data by adopting the within-group distribution of a suitable unabridged life table population, but discontinuities at age-group junctions are inherent in this approach. A smoother distribution is obtained

by mathematical interpolation, either osculatory or polynomial. Osculatory interpolation gives a smoother distribution, especially for irregular data, because of the constraint of continuity at junctions. The widely-used Sprague multipliers employ fourth- and fifth-difference osculatory interpolation to smooth single-year data within 5-year age groups without changing the totals. They are less satisfactory for the first and last two age groups and where there is marked curvature. Sprague multipliers can be employed with any continuous data including rates, for any age partition (e.g., tenths), and with unequal age intervals. A blended distribution can be derived by averaging smoothed distributions obtained by osculatory interpolation within the five possible age groupings (with terminal digits 0–4, 1–5, 2–6, 3–7, 4–8). Spline interpolation is a type of osculatory interpolation. It involves extensive iteration requiring modern computers and gives a smoother distribution (US Bureau of the Census 1973).

2.2 Parity

Errors in the reporting of parity often involve the omission or understatement of live births. Where adoption is common, overstatement may occur due to both the natural and adoptive mother reporting the birth. The reporting of stillbirths as livebirths also results in slight overstatement. Omissions may be detected and corrected by indirect estimation (see *Demographic Techniques: Indirect Estimation*). Further error arises when childless women are recorded as 'parity not stated' due to the ambiguous use of the dash symbol (–) by interviewers. If these apparent not-stated cases are excluded, average parity is overestimated. If all not-stated cases are included and treated as childless, average parity is underestimated. The El-Badry correction estimates the proportion of women with parity not stated who are in fact childless (El-Badry 1961). Based on the observed high correlation between true proportions childless, $Z^*(i)$, and proportions with parity not stated, $NS(i)$ where $i = 1$ to 8 represents the age groups 10–14 to 45–49, El-Badry assumes $NS(i) = \alpha\ Z^*(i) + \beta$, where α is the proportion of truly childless women who are erroneously recorded with parity not stated and β is the constant true proportion not stated. Since the reported proportions childless, $Z(i)$, equal $(1-\alpha)\ Z^*(i)$, then $NS(i) = [\alpha/(1-\alpha)]\ Z(i) + \beta$, and $\alpha/(1-\alpha)$ and β can be estimated by ordinary least squares based on the first four or five age groups. Then, β is used to estimate the number of women in each age group for analytical purposes, and true proportions childless are estimated by $Z^*(i) = Z(i) + NS(i) - \beta$. This correction is based on all women. If data on parity are available only for ever-married women (in which case all never-married women are generally assumed childless), the procedure can be applied on the assumption of a linear relation between the proportions of ever-married women who are childless and who have parity not stated.

3. Other Adjustments

The smoothing of mortality rates or of the survival function can be achieved through use of a device like the Brass logit life-table system, which involves fitting observed data to a suitable standard such as a model life table (Brass et al. 1968; see *Demographic Models*; *Demographic Analysis: Probabilistic Approach*). Similarly, the smoothing of fertility rates is done through use of the Brass relational Gompertz model with an appropriate standard such as the Booth standard (Brass 1974, Booth 1984; see *Demographic Models*).

Other adjustments include shifting the population to refer to a specified date. This is achieved by linear or exponential interpolation of each age group between two dates for which enumerations are available; or by adjusting the total population exponentially (the growth rate is required) and assuming the same age-sex structure as at the enumeration. The structure of the open-ended age group can be derived using a suitable model stable population (see *Population Dynamics: Theory of Stable Populations*). When demographic events in a time reference period are reported retrospectively, errors of timing occur. For births, such errors are corrected through indirect estimation (see *Demographic Techniques: Indirect Estimation*).

Bibliography

Arriaga E, Johnson P D, Jamison E 1994 *Population Analysis with Microcomputers. Presentation of Techniques*. US Bureau of the Census, Washington, DC, Vol. 1, Chap. 2.

Booth H 1984 Transforming Gompertz's function for fertility analysis: The development of a standard for the relational Gompertz function. *Population Studies—Journal of Demography* **38**: 495–506

Brass W 1974 Perspectives in population prediction: Illustrated by the statistics of England and Wales. *Journal of the Royal Statistical Society A* **137**: 532–83

Brass W, Coale A J, Demeny P, Heisel D F, Lorimer F, Romaniuk A, van de Walle E 1968 *The Demography of Tropical Africa*. Princeton University Press, Princeton, N J

El-Badry M A 1961 Failure of enumerators to make entries of zero: Errors in recording childless cases in population censuses. *Journal of American Statistical Association* **56**: 909–24

United Nations Department of Economic and Social Affairs 1955 *Manual II. Methods of Appraisal of Quality of Basic Data for Population Estimates*. ST/SOA/Series A/23. United Nations, New York

United Nations 1967 *Manual IV. Methods of Estimating Basic Demographic Measures from Incomplete Data*. ST/SOA/-Series A/42. United Nations, New York

United Nations 1983 *Manual X. Indirect Techniques for Demographic Estimation*. ST/ESA/SER.A/81. United Nations, New York

US Bureau of the Census 1973 Shryoc H S, Siegel J S, Larmon E A (eds.) *The Methods and Materials of Demography* US Bureau of the Census, Washington, DC, Chaps. 8, 22

H. Booth

Demographic Techniques: Family Reconstitution

1. The Technique

Family reconstitution is the technique of linking records of demographic events, usually of an ecclesiastical nature, within and between individual lives, in order to recreate individual life histories and the histories of families. While genealogists have always pursued such linking, the intent of demographers is not simply to record chains of descent and marriage but rather to compile information on the *demographic rates* pertaining to the *population* of which the individuals and families were a part.

Family reconstitution revolutionized the study of historical population processes, allowing estimation of demographic rates that was virtually impossible by other means. Demographers would ordinarily estimate vital rates by comparing counts of vital events to counts of the population at risk. For example, the general fertility rate could be estimated by dividing the number of births in a year by the number of women aged 15–49. However, accurate censuses contemporary with records of vital events are rare in the past. Family reconstitution not only overcame this difficulty by counting both events and persons, but also afforded a methodological alternative that is superior in many ways to estimation from census data, because the reconstitution data are on *individuals*, so that the characteristics of persons at risk are more specifically known. For example, estimates of fertility can be specific for unmarried, married, first-married, and remarried women, can include information on duration of marriage as well as age, and so on. It would be very rare to find information tabulated in such detail in published censuses, even in modern times. The technique of family reconstitution also led to important insights. For example, it showed (in contrast to the expectations of Notesteinian theory) that the decline of fertility in some places occurred well before economic modernization and before any decline in mortality.

1.1 History and Scope of the Technique

Family reconstitution as a demographic technique was developed initially by Swedish demographers (Edin 1915, Hyrenius 1942, 1958) and advanced by the French (Fleury and Henry 1985, Gautier and Henry 1958, Henry 1956, 1967, Henry and Blum 1988), then by English scholars (Wrigley 1966, 1969, 1972, 1973, 1997, Wrigley and Schofield 1973, Wrigley et al. 1997). It linked by *names of persons* the parish records of baptisms, burials, and marriages of individuals (hence the technical phrase, *nominal data linkage*). Genealogists had always used the linking of disparate records by names of individuals. Its large-scale use by historical demographers was an innovation. It is of course now commonplace, especially using unique identifiers, in the management of medical, insurance, taxation, police, and other data, especially by computer. In the hands of historical demographers, the ecclesiastical ceremonies noted in parish records served as *proxies* for the underlying births, deaths, and onsets of cohabitation. It must always be remembered that the events in parish records are not the same as the underlying demographic events, e.g., a baptism is not a birth. With that caveat, a family reconstitution should provide information on the dates of birth, marriage, widowhood, remarriage, dates of the birth of children, and of death for individuals. It is possible in principle to continue such linkages across succeeding generations, or across collateral lines. Historians and demographers can examine family reconstitutions for patterns of fertility, mortality, and nuptiality, and to a limited extent those of migration in historical populations. The technique can be extended to the study of any other type of social or biological phenomena supported by the particular dataset or other data such as medical, notarial, or tax records that can be linked to it, e.g., heritability of cause of death, marriage patterns between social groups, fertility by social class, etc. For example, family reconstitution can be used in investigation of genetic characteristics (Bean 1990), or the relationship between fertility patterns and socioeconomic and cultural characteristics (Hammel 1995), or in patterns of fertility change within and between family lines (Hammel and Goldstein 1996).

2. Related Methods and Expansions

2.1 Genealogies

Genealogies are records of the linked lives of persons related by consanguinity and affinity and can, like reconstituted families, be used to extract demographic information. They are usually very selective, however, and unlikely to be representative of the broader population. Aristocratic and wealthy families and lineages are more likely to be recorded than those

more characteristic of the population. Within genealogies, persons who do not leave descendants are more likely to be forgotten. Males are more likely to be remembered than females. Females may be less well remembered if they have fewer male children of consequence with progeny, so that events linked through females are likely to be lost (especially if the females were in informal liaisons with the males of the lineage). We might expect such biases to shift in matrilineal societies, but no data on such differences have been brought to light.

2.1.1 Ortssippenbücher. An important exception to some selection biases of genealogies are the *Ortssippenbücher* of some German regions, compiled by local genealogists from parish registers primarily to demonstrate the ethnic purity of families. These are genealogies of the common folk (but of course selective as their purpose indicates), and have been used for demographic estimation (Knodel 1988, Knodel and Shorter 1976).

2.1.2 Qing Dynasty genealogies. A second exception to some of these biases are the imperial genealogies of the Qing Dynasty in China, which were continuously compiled for all descendants of the founder and have also been used for demographic estimation (Lee et al. 1993). Females are relatively well recorded in these data.

2.2 Civil Vital Event Registers

Civil registers of vital events have the same potential as parish registers. They are intended to cover all faiths in a civil jurisdiction, but they may under-report some population segments such as vagrants or the poor. In some regions, civil registers can serve as a continuation of earlier parish registers and are especially valuable if secularization progresses and observation of ecclesiastical ceremonies declines.

2.3 Population Registers

Historically, civil and ecclesiastical population registers were periodic listings of the inhabitants of residential units. They have the advantage of preserving the identity of the residential unit (usually 'households') within which the demographic events occurred, although just what that unit was is itself often problematic, since actual and legal residence might be confused, and even long-term out-migrants might sometimes be continued in the listings. (Such continuation is not unknown in some parish records as well, e.g., births to out-migrant former inhabitants.) Population registers accomplish what is most difficult in family reconstitution, namely the nominal data linkage within and between individual lives, because local recorders presumably knew the identities of the persons and would not ordinarily confuse one John Smith with another John Smith. Thus, the families discernible in population registers are not so much reconstituted as preconstituted. The occurrence of vital events is implied through changes in household membership. However, the longer is the cycle according to which the registers were updated (usually at best annually), the more uncertain are the precise *dates* of the events leading to changes in membership, each update being only a *terminus ad quem*, and thus the more uncertain the exact periods of risk.

2.3.1 Civil population registers. Perhaps the most celebrated historical population registers are the household registers of Tokugawa Japan c. 1650–1868 (Hayami 1969, Smith et al. 1977) and later of Japanese-occupied Taiwan c. 1904--45, but registers were also kept in some European countries, e.g., Belgium (Gutmann and van de Walle 1978) and Italy (Kertzer and Hogan 1989). Records in such registers may sometimes be linked to independently kept vital event registers, both civil and ecclesiastical. (The Tokugawa registers actually had an ecclesiastical purpose, namely surveillance of the spread of religious sects, i.e., Christianity.)

2.3.2 Libri status animarum. *Libri status animarum* ('books of the state of souls') were ecclesiastical population registers that often contained not only registration information (e.g., which members of a household had been catechized, or which had been confirmed) but also the dates of these events and of baptism, marriage, and burial. Known by various terms in Catholic and Protestant regions, they vary in continuity and the diligence with which they were kept. Among the best are the Swedish catechitical and similar records (Lext 1984).

3. Methods of Family Reconstitution

The work of family reconstitution was from its inception strongly dependent on the paleographic and historical skills and intuition of the scholar. The earliest efforts were done by hand, starting with a marriage record, then attempting to find the burial and any remarriage records of the spouses, then the baptismal records of the children, then the marriages and deaths of those children. Such data were entered on a 'marriage fiche' or 'family reconstitution form' for later collation and analysis. Account had to be

taken of the literacy of scribes, variations in spellings of the same names, variations in standardization of surnames, and of course the completeness of the records themselves. Later attempts were assisted by computers, but final judgments often still depended on intuition in deciding between two or more equally plausible linkages. Efforts have been made to completely automate the linkage procedure, including the use of mathematical programming methods to resolve ambiguities, although it is not clear whether the algorithms themselves may not simply substitute new for old biases (Wrigley and Schofield 1973, Welford 1989, Shannon et al. 1989, Skolnick 1973, Schofield 1992, Nault and Desjardins 1989, Bouchard 1992, Bengtsson and Lundh 1991, 1993). It is difficult to know when the correctness of the nominal links in a dataset has been maximized, especially when data are scanty or contain errors. For example, if surnames are often omitted in the parish documents, and especially if the stock of first names is small (e.g., 'John married Mary, June 1, 1712'), the number of alternative linkages can be very large because many Johns may have married Marys on that date. In principle linear programming methods can be used to obtain the best solution, one that maximizes the density or quality of linkage, but that solution may not be historically correct. Estimates of central tendency (e.g., the general fertility rate) will not be biased by such presumably random error, but variances will be, as will be any associations with other variables if guesses are more likely to be bad for some classes of persons. The results of nominal record linkage may be compared to population registers, but population registers also have problems of accuracy. Such problems of optimizing linkages are of course less when data are detailed (e.g., Swedish parish and population register records) or contain ethnographically validated clues to linkage (e.g., interfamilial patterns of godparenting in some South Slavic contexts wherein children of a marriage tend to have the same godparents, who also serve as marriage witnesses, etc.).

4. *Limitations of the Technique*

Although a powerful tool, family reconstitution has important limitations caused by missing data. Some stem from the nature of the sources.

4.1 *Data Quality and Coverage*

While some parish records (e.g., Florence, Istria) go back to the fifteenth century, most do not commence until the sixteenth, and many are not reliable until the eighteenth century. Civil event and population registers are generally no older than the nineteenth, but ecclesiastical registers are often as old as the corresponding event registers. *Libri status animarum* were only irregularly kept in many areas, and all ecclesiastical records were subject to variations in quality depending on the abilities and dedication of the recording priest, attempts by higher authorities (both ecclesiastical and civil) to collect sufficient and standardized information, and the vagaries of copying and preservation. Ecclesiastical records are known for Catholic and Protestant populations but seldom for Orthodox Christians (not to mention Muslims, Jews, or others). Thus in areas of mixed religion, for example, in nominally Catholic Croatia or parts of Bosnia, large non-Catholic proportions of the population would not be included in the data and investigation of regional patterns would be hampered. Remarkable discontinuities in the data can occur in periods of shifts in religious adherence, as for example sharp swings between conformity and nonconformity in English parishes, or the widespread conversion of Catholics to Islam in the western Balkans. Similar problems would arise in using missionary records outside of Europe and the Americas. *Civil* event and population registers are by their nature relatively independent of religious bias, although they may miss subpopulations whose members for other reasons are under-registered.

4.2 *Statistical Biases Caused by Sample Selection*

Probably the most serious problems for family reconstitution are selection bias and truncation, issues addressed early by L. Henry (1956, 1967).

4.2.1 Selection by data quality. Where data are in general of acceptable but varying quality, there is an inherent tendency to want to analyze records that are only of the best quality. The socioeconomic condition and demographic regime of a part of a population that has good records may be quite different from those of another part that has poor records. Analysis limited to only the best data may give a misleading picture of the whole, either with respect to entire parishes or to subpopulations within parishes. This kind of selectivity bias is similar to that experienced in the analysis of genealogical data.

4.2.2 Selection by migration. Because the work of linkage is so tedious, even with computer assistance, most reconstitutions have been limited to single parishes. Out-migration from such parishes will ordinarily go unremarked in parish records (and even sometimes in household registers). Since the kinds of

persons who complete their lives within a parish may be quite different from the kinds who leave it, building a demographic picture of a parish based on completed lives may be misleading. These and similar problems (see below) can be partly overcome by analysis of a set of contiguous parishes, especially if migration tended to be short-range. Migration effects may be powerful where migration flows were strong (viz. Ruggles 1992, Kasakoff and Adams 1995, cf. Wrigley 1994). Reconstitution data have been used to examine migration *per se* (Bideau and Brunet 1993).

4.2.3 Truncation and censoring. The time actually spent in a parish by a person who was not born in it or did not die in it may be unknown, and thus the time at risk of some demographic event may not be known. Many demographic measures cannot be properly estimated if time at risk is not known. For example, a woman may have had recorded births but no recorded death. The amount of time she was actually at risk of having the next birth cannot then be accurately estimated. Thus, the fertility of women at her achieved parity or at ages higher than that of her last recorded birth cannot be estimated accurately. For example, estimating the total fertility rate for the population would demand knowledge of the age-specific fertility rates by the usual five-year age groups or by single years, and the loss of information on some women who died without record within such spans would deflate the denominators of the rate calculations.

In general, observation on individuals at risk of some event is said to be truncated if the size of the cohort at risk to which the individuals belong is unknown and censored if the number of individuals at risk is known but the events are unobserved or if observed events are recorded only in time intervals instead of accurately (see *Event-history Analysis in Continuous Time*). Examples are: (a) death records lacking age so that date of birth is unknown and the number of persons in the birth cohort cannot be recovered from birth records (truncated on the left); (b) death records with age recorded but birth records unavailable so that information is obtained only on persons who died at each age, not on those who survived that age (truncated on the right); (c) no marriage records but death records with age information showing survival and marital status so that the birth cohort size is known but the date of marriage is not (censored on the left); (d) birth and marriage records but no information on death so that observation ceases on the date of marriage (censored on the right). If statistical methods take censoring into account, censoring induces no bias in the computation of rates or probabilities of occurrence of an event, or in event-history analyses (see *Event History Analysis: Applications*). By contrast, truncation does induce such bias. As a general rule, the onset or cessation of observation of an individual at risk should not be correlated with the event of interest.

Solutions to the problem of missing data in sets of life histories are various but problematic. They include using the occurrence of another event to the same individual (e.g., next births if interest is in mortality), random censoring, occurrence of an event to a close family member (thus implying continued residence and at least potential registration of an event), or statistical imputation of the censoring point, etc. (Hammel 1993, Trussell and Guinnane 1993).

4.2.4 Bias from delays in recording. Bias may also be created by delays in the recording of demographic events or in the ceremonies that recognize them. An individual may die or move before an event is recorded and therefore be missed. Such delay is especially problematic for births and infant deaths. If the date of recording is accepted as the date of birth, long delays will make infants seem younger than they are, biasing any age-related computations. If an infant dies before it can be baptized, there will be no record of the birth, and often there will be no record of the death since unbaptized infants often received no formal burial. If the death is recorded and an age at death is given, the date of birth can be imputed and thus recovered. If death is unrecorded, mortality may be underestimated. Where both birth and death are unrecorded, infant mortality will also seem to be lower than it really is. Since in many historical populations up to a third of infants died in their first year, and half of those died in their first month, such statistical biases can be very important.

5. *Prospects*

The most promising outlook for family reconstitution and related techniques include the extension and elaboration of computer-based mathematical programming methods for rigorous and testable nominal data linkage, the combination of multiple data sources such as parish event registers, population registers, tax and notarial records, etc. to increase the accuracy and scope of data, and the use of event-history techniques to overcome problems of missing event dates.

6. *Summary*

Despite difficulties imposed by the quality and coverage of the historical data and the possibility of biased inference, family reconstitution has given demographic and historical scholarship a remarkable window on the demographic past, unachievable by any other means. In so doing it provides information for

much longer spans of time than do modern data and often for periods before any effects of industrialization or modernization have been felt. It thus allows examination of longer term trends under broader socioeconomic conditions, and like all historical enterprises allows scholarship to escape the constraints imposed by modern conditions.

See also: Demographic Measurement: General Issues and Measures of Fertility; Demographic Techniques: LEXIS Diagram; Event History Analysis: Applications; Event-history Analysis in Continuous Time

Bibliography

Bean L L 1990 The Utah Population Database: demographic and genetic convergence and divergence. In: Adams J, Lam D A, Hermalin A, Smouse P E (eds.) *Convergent Issues in Genetics and Demography*. Oxford University Press, New York, Oxford

Bengtsson T, Lundh C 1991 Evaluation of a Swedish computer program for automatic family reconstitution. *Lund Papers in Economic History* 8. Lund University, Department of Economic History, Lund, Sweden

Bengtsson T, Lundh C 1993 Name-standardisation and automatic family reconstitution. *Lund Papers in Economic History* 29. Lund University, Department of Economic History, Lund, Sweden

Bideau A, Brunet G 1993 The construction of individual life histories: application to the study of geographical mobility in the Valserine Valley in the nineteenth and twentieth centuries. In: Reher D, Schofield R (eds.) *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK

Bouchard G 1992 Current issues and new prospects for computerized record linkage in the province of Quebec. *Historical Methods* **25**: 67–73

Edin K A 1915 Studier i Svensk fruktsamhetsstatistik. *Ekonomisk Tidskrift* **9**: 251–304

Fleury M, Henry L 1985 *Nouveau manuel de dépouillement et d'expoitation de l'état civil ancien*, 3rd edn. Institut National d'Etudes Demographiques, Paris

Gautier E, Henry L 1958 *La population de Crulai, paroisse Normande au XVIIe et au XVIIIe siècles*. Presses Universitaires de France, Paris

Gutmann M P, van de Walle E 1978 New sources for social and demographic history: the Belgian population registers. *Social Science History* **2**: 121–43

Hammel E A 1993 Censored intervals in family reconstitution: a sensitivity test of alternative strategies with historical Croatian data. In: Reher D, Schofield R (eds.) *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK

Hammel E A 1995 Economics 1: culture 0. Fertility change and differences in the northwest Balkans 1700–1900. In: Greenhalgh S (ed.) *Rethinking Reproduction: Culture and Political Economy in Demographic Behavior*. Cambridge University Press, Cambridge, UK

Hammel E A, Goldstein J 1996 Dick Easterlin in Slavonia: patrilineal fertility decline on the Croatian frontier 1700–1900. Paper presented to the Population Association of America, New Orleans, May

Hayami A 1969 Aspects démographiques d'un village japonais, 1671–1871. *Annales E.S.C.* **3**: 617–39

Henry L 1956 *Anciennes familles Genevoises: Étude démographique, XVIe–XXe siècle*. Presses Universitaires de France, Paris

Henry L 1967 *Manuel de démographie historique*. Droz, Paris

Henry L, Blum A 1988 *Techniques d'analyse en démographie historique*, 2nd edn. Institut National d'Etudes Demographiques, Paris

Hyrenius H 1942 *Estlandssvenskarna. Demograkska studier*. Lund University, Lund, Sweden

Hyrenius H 1958 Fertility and reproduction in a Swedish population without family limitation. *Population Studies*. **12**: 121–30

Kasakoff A B, Adams J W 1995 The effect of migration on ages at vital events: a critique of family reconstitution in historical demography. *European Journal of Population* **11**: 199–242

Kertzer D, Hogan D P 1989 *Family, Political Economy, and Demographic Change: The Transformation of Life in Casalecchio, Italy, 1861–1921*. University of Wisconsin Press, Madison, WI

Knodel J E 1988 *Demographic Behavior in the Past. A Study of Fourteen German Village Populations in the Eighteenth and Nineteenth Centuries*. Cambridge University Press, New York, Cambridge, UK

Knodel J E, Shorter E 1976 The reliability of family reconstitution data in German village genealogies (Ortssippenbücher). *Annales de Démographie Historique* **13**: 115–54

Lee J, Campbell C, Feng W 1993 The last emperors: an introduction to the demography of the Qing (1644–1911) lineage. In: Reher D, Schofield R (eds.) *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK

Lext G 1984 *Studier i svensk kyrkobokföring 1600–1946*. Göteborg University, Göteborg, Sweden

Nault F, Desjardins B 1988 Recent advances in computerized population registers. *Historical Methods* **21**: 29–33

Nault F, Desjardins B 1989 Computers and historical demography: the reconstitution of the early Quebec population *Collection de Tires a Part*, No. 266. Université de Montréal, Département de Démographie, Montreal, PQ, Canada

Reher D, Schofield R (eds.) 1993 *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK

Ruggles S 1992 Migration, marriage, and mortality: correcting sources of bias in English family reconstitutions. *Population Studies* **46**: 507–22

Schofield R 1992 Automatic family reconstitution: the Cambridge experience. *Historical Methods*. **25**: 75–9

Shannon H S, Jamieson E, Walsh C, Julian J A, Fair M E, Buffet A 1989 Comparison of individual follow-up and computerized record linkage using the Canadian mortality data base. *Canadian Journal of Public Health* **80**: 54–7

Skolnick M 1973 The resolution of ambiguities in record linkage. In: Wrigley E A (ed.) *Identifying People in the Past*. Edward Arnold, London

Smith T C, Eng R Y, Lundy R T 1977 *Nakahara: Family Farming and Population in a Japanese Village, 1717–1830*. Stanford University Press, Stanford, CA.

Trussell J, GuinnaneT 1993 Techniques of event-history analysis. In: Reher D, Schofield R (eds.) *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK

Welford J A 1989 *Nominal Record Linkage: the Development of Computer Strategies to Achieve The Family-Based Record Linkage of Nineteenth Century Demographic Data*. University Microfilms International, Ann Arbor, MI

Wrigley E A (ed.) 1966 *An Introduction to English Historical Demography from the Sixteenth to the Nineteenth Centuries*. Basic Books, New York

Wrigley E A 1969 *Population and History*. McGraw-Hill, New York
Wrigley E A 1972 Some problems of family reconstitution using English parish register material: the example of Colyton. *Proceedings of the Third International Conference of Economic History, Munich 1965, Section VII, Demography and Economics*, La Haye, Paris
Wrigley E A (ed.) 1973 *Identifying People in the Past*. Edward Arnold, London
Wrigley E A 1994 The effect of migration on the estimation of marriage age in family reconstitution studies. *Population Studies* **48**: 81–97
Wrigley E A 1997 How reliable is our knowledge of the demographic characteristics of the English population in the early modern period? *Historical Journal* **40**: 571–95
Wrigley E A, Davies R S, Oeppen J E, Schofield R S 1997 *English Population History from Family Reconstitution, 1580–1837*. Cambridge University Press New York, Cambridge, UK
Wrigley E A, Schofield R S 1973 Nominal record linkage by computer and the logic of family reconstitution. In: Wrigley E A (eds.) *Identifying People in the Past*. Edward Arnold, London, pp. 64–101

E. A. Hammel

Demographic Techniques: Indirect Estimation

Noted Princeton University demographer Ansley Coale jokingly described indirect estimation as 'making bricks without straw.' Straw, in this case, consists of the traditional data sources for producing basic descriptive measures of demographic processes: civil registration for numbers of events and population censuses for exposure to risk. Many countries, particularly in the developing world, lack complete or accurate traditional data. Events may not be registered, components of the population may not be enumerated in the census, or information about key demographic characteristics such as age, may be reported with substantial error. Indirect estimation techniques were developed to try to obtain valid estimates of demographic measures in situations where the traditional data were not directly usable. Key characteristics of such techniques are that they be simple to implement and that they be robust to standard patterns of reporting error. The father of such techniques was William Brass. It was he who realized that simple demographic models could be used to simulate population processes, and provide a basis for converting indirect indicators into standard demographic measures. Although numerous authors have developed or refined indirect methods, they all owe an intellectual debt to Brass. The methods are only outlined here, but many of them are described in detail in United Nations (1983).

For conceptual purposes, it is useful to divide estimation techniques into two broad families. One family consists of true indirect estimation techniques. The strategy underlying indirect techniques is to find a readily measurable indicator that is largely, but not entirely, determined by the parameter of interest. Simulations developed using demographic models are then used to explore the relationship between the indicator and the parameter of interest, and to develop a simple approach to adjusting the indicator for other confounding factors, to provide a pure estimate of the parameter itself. An example of such an indicator is the proportion of dead among children ever borne by women of a given age group. This proportion is an indicator of the level of child mortality, but it is also affected by the age pattern of childbearing.

The second family is made up of consistency checks. Different indicators of the same demographic process are compared, either using models or formal demographic relationships. Differences are interpreted as indications of data error. Assumptions about the nature of data errors can sometimes justify adjustment of one set of indicators for presumed error in order to arrive at an improved overall estimate of the underlying process.

1. *Indirect Methods*

Reverse projection can be used to estimate recent aggregate fertility levels from the population age structure by allowing for mortality. The approach has a long history, but has not proved particularly useful. First, a common error in age distributions is for the population under age five years, to be under-recorded. This error causes recent fertility estimates based on reverse projection, to be too low. Second, the method is sensitive to the assumption of no (or not known) migration. Third, mortality rates have to be known at least approximately. Fourth, the fertility indicator estimated is the birth rate or General Fertility Rate; no detail about age patterns of fertility is provided.

This last shortcoming is avoided by the 'Own Children' variant of reverse projection (Cho et al. 1986). Children are matched to a likely mother within each household using criteria such as age and parity. Information on age of child provides a time period for the fertility measure, and in combination with age of putative mother provides the age of the mother at the time of the birth. Unmatched children are distributed proportionately, and both children and women are reverse-projected using appropriate mortality estimates. A suitably-matched age distribution thus provides estimates of age-specific fertility rates for calendar years prior to the census or survey. Results remain sensitive to age reporting errors, however, and the other problems of standard reverse projection.

Fitting stable population models to an observed age distribution in combination with one other indicator

such as the population growth rate or probability of dying by age five years represents another way of indirectly estimating fertility. Stable population analysis is of limited value today, because widespread declining fertility has affected age distributions worldwide. Preston (1983) has developed an elegant approach based on a generalized equation of age structure and a logit transformation of a standard life table survivorship function. The reciprocal of the proportion of the population of a particular age a, $c(a)$, inflated by the exponentiated sum of population growth rates (adjusted if necessary for migration) up to age a, $\exp(R(a))$, is linearly related to the logit of the standard survivorship function to age a, $(1-p^s(a))/p^s(a))$. The intercept of the straight line is the reciprocal of the current birth rate, $1/b$, and the slope is an indicator of the actual mortality level relative to that of the standard survivorship function, itself divided by the birth rate, (γ/b). Thus

$$1/\{c(a)\exp(R(a))\} = 1/b + (\gamma/b)\{(1-p^s(a))/p^s(a)\} \quad (1)$$

This method is an excellent example of how the mathematics of population age structures can be used to improve demographic estimation techniques.

The best-known and most widely applied indirect estimation method is the Brass (1964) method for estimating child mortality from women's reports of the number of children born alive and the number that have died. It is argued that these reports, which involve no reporting of dates of events, are likely to be accurate in most populations, even those with substantial levels of illiteracy. The key indicator is the proportion dead of children ever born: the higher the level of child mortality, the higher the proportion dead, other things being equal. Brass' contribution was devising a method to make other things equal. First, the data are classified by age of mother as a crude control for the average duration of exposure to risk of dying of a mother's children. However, the age pattern of childbearing still affects average exposure time. For women of a given age group, average exposure time of the children will be longer in an early childbearing population than in a late childbearing population. Brass used models of fertility and child mortality to evaluate adjustment factors $k(i)$ for each five year age group of mother i needed to convert proportions dead $PD(i)$ into life table probabilities of dying in childhood, ${}_nq_0$, for a range of age locations of childbearing. In a particular application, the age location of childbearing was identified by the ratio of average parity for 15–19 year olds to that of 20–24 year olds, $P(1)/P(2)$, and appropriate values of $k(i)$ were obtained by interpolation from a standard table.

The original Brass method has undergone numerous refinements. Perhaps the most crucial development, however, was the insight by Feeney (1980) that under conditions of changing mortality, approximate reference dates could be put on estimates from different age groups of mother: the older the mothers, the further in the past the child mortality reflected by the survival of their children. Methods were developed not only to estimate levels of child mortality indirectly but also to estimate the recent time trend. Figure 1 shows the application of the method to data from the 1989 census of Kenya, together with direct estimates from the 1993 Kenya Demographic and Health Survey.

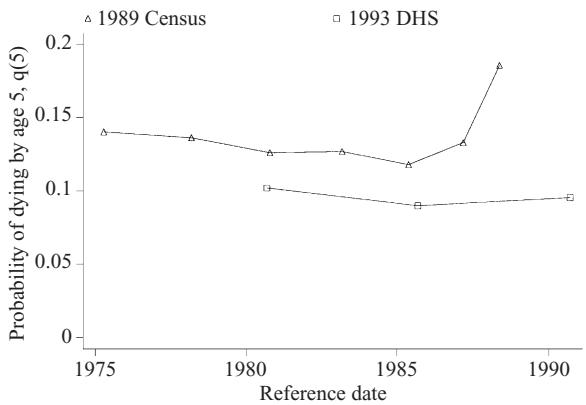

Figure 1
Application of Brass method to child survivorship data, Kenya 1989 Census

The Brass method and the developments of it have been very widely applied. It revolutionized knowledge of mortality conditions, particularly in sub-Saharan Africa but also in Latin America. It also gave rise to a concerted effort to find indirect estimation methods for adult mortality and migration. As such, it has proved one of the most influential ideas of demography.

The Previous Birth Technique is another indirect method for child mortality estimation. Where information is available or can easily be collected concerning the survival of the birth immediately preceding a current delivery, the proportion dead among the preceding births can be calculated (Brass and Macrae 1984). The proportion dead turns out to be a good indicator of the probability of dying by age 2, ${}_2q_0$. Some information on average birth intervals permits refinement of this equivalence. The method is very easy to apply, but suffers from concerns about selection bias since reports are typically obtained from health facilities.

A number of indirect estimation methods have focused on adult mortality. One group relies on information on survival of close relatives. Adult survivorship can be estimated from information on survival of mothers or fathers (Brass and Hill 1973). The age of the respondent is the exposure time, the age distribution of parents at the time of the births of a

group of respondents, the confounding variable that needs to be controlled. An ingenious approach to the estimation of maternal mortality has been the development of the Sisterhood method (Graham et al. 1989). Data are collected on survival of sisters with additional information for reported deaths about the timing of death relative to pregnancy. Despite the ingenuity of these approaches, questions remain about their accuracy.

A second group is made up of various forms of intercensal survival methods. Females aged 20–24 in 1990 will be aged 30–34 in 2000, so in the absence of net migration the ratio of females aged 30–34 in a 2000 census to females aged 20–24 in a 1990 census approximates the life table survivorship ratio ${}_5L_{30}/{}_5L_{20}$. Age distributional effects will be modest, but age misreporting and changes in census coverage can have large effects on such ratios, which as a result tend to be of little value in most developing countries. Generalized age structure equations involving age-specific growth rates have been used to minimize the effects of local age misreporting (Preston and Bennett 1983), but the method does not reduce effects of systematic age bias or coverage change.

The measurement of migration is problematic in many situations, because even sophisticated recording systems miss events. Net migration is often estimated as a residual from intercensal survival. In the example above, if the survivorship ratio ${}_5L_{30}/{}_5L_{20}$ is known, a predicted number of females aged 30–34 in 2000 can be calculated by multiplying the number of 20–24 year olds in 1990 by this ratio. The difference between the observed and expected number of females aged 30–34 in 2000 is then an estimate of net intercensal migration by the 20–24 cohort. An alternative estimation procedure (Hill 1987) is available using the Demographic Balancing Equation: the change in size of an age group between two censuses is equal to the difference between intercensal entries (birthdays across the younger boundary of the age group plus in-migrants of the age group) and intercensal exits (birthdays across the upper boundary of the age group, deaths of the age group and out-migrants of the age group). Net migration can then be isolated and estimated from the other components. This latter approach has the advantage of producing estimates for age groups rather than cohorts. Both approaches, however, estimate migration as a residual, and are thus very sensitive to even quite small errors. Unlike most indirect estimation techniques, the indicator variable (change in population size) may be determined more by the confounder (mortality) than by the parameter of interest (migration). Results are therefore not robust.

Efforts have been made to estimate emigration from information on residence of relatives, particularly of children as reported by mothers and of siblings. These methods have not achieved widespread acceptance, however.

2. *Consistency Checks*

The best-known consistency check and adjustment technique is the Brass (1964) P/F Ratio Method for fertility estimation. Records of births in a year by age of mother provide age-specific fertility rates. Information on children ever born by age of mother, provides measures of lifetime fertility. If fertility is not changing over time and the data are consistent, cumulated age-specific fertility to a particular age group will be the same as lifetime fertility or parity for that age group. Brass developed a simple way for cumulating current age-specific fertility rates to approximate parity equivalents ($F(i)$) in order to check consistency with parities ($P(i)$) for each age group i. However, he also went further, to hypothesize that typical errors in records of numbers of births might be proportionately constant by age of mother, so that the age-specific fertility rates would have the correct shape, but at a potentially incorrect level. Typical errors of reporting of children ever born, however, were hypothesized to vary by age, but to be smallest for younger women. Ratios of $P{:}F$ for successive age groups, thus not only reflect consistency but, for younger women, represent potential adjustment factors for the age-specific fertility rates.

The assumption of unchanging fertility can be relaxed if information is available on average parity for two points in time (Zlotnik and Hill 1981). Parity changes by cohort between the two sources can then be cumulated to obtain a period-specific parity distribution that can be compared with cumulated period-specific fertility rates. If the other assumptions hold, adjustment can be made even under conditions of changing fertility.

Consistency checks with the option for adjustment in the measurement of adult mortality have largely focused on comparisons of the age distribution of

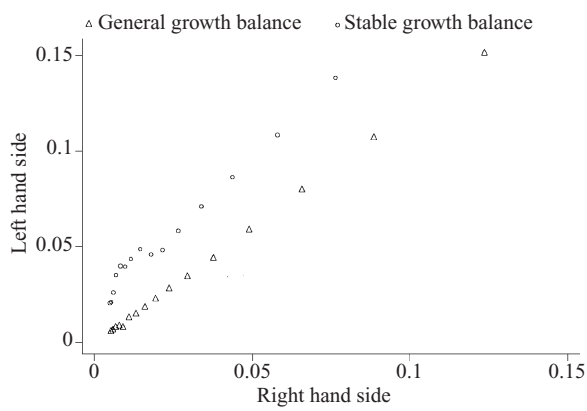

Figure 2
Application of stable and general growth balance methods to data for China females

deaths with the age distribution of the population. The first such method was proposed by Brass (1975). In any population with negligible migration, the entry rate into any open-ended age segment $a+$ is equal to the growth rate of the segment plus the death rate in the segment. In a stable population, the growth rate r is constant at all ages, so if $b(a+)$ is the entry rate and $d(a+)$ is the exit rate,

$$b(a+) = r + d(a+) \tag{2}$$

If deaths are recorded with a completeness c (assumed constant at all adult ages) relative to the population, the recording of which is also assumed to be of constant completeness at all ages, and $b(a+)$ is estimated from the population age distribution alone,

$$b(a+) = r + (1/c)d^o(a+) \tag{3}$$

where $d^o(a+)$ is the observed death rate over age a. There is thus a linear relationship between the entry rate and the observed death rate, the intercept of which is the stable growth rate and the slope of which is the reciprocal of the completeness of death registration.

The assumption of stability can be relaxed if two successive population counts are available and growth rates over ages a can be measured directly (Hill 1987). If census coverage is constant within, but not between, the censuses,

$$b(a+) - r^o(a+) = k + (1/c)d^o(a+) \tag{4}$$

where $r^o(a+)$ is the observed growth rate over age a and k is the natural log of the change in coverage from the first to the second census. Figure 2 shows the application of both growth balance methods to Chinese census data for females, using the 1990 age distribution and deaths by age for the stable version, and 1982 to 1990 population change together with the distribution of deaths by age for the general version. The improved smoothness of the general version for this destabilized population is evident.

Other equations for population age structure can also be used as the basis for similar consistency checks. In an influential paper, Bennett and Horiuchi (1981) use the exponentiated sum of age specific growth rates above age a to expand the deaths at each age to equate deaths above age a to population age a. Though error patterns are likely to affect these methods in slightly different ways, there is no consensus as to which should be preferred in different situations.

3. Future Directions

Indirect estimation is an interim solution to data problems. Data quality will improve, rendering these techniques unnecessary. Birth histories have shown that reasonably accurate direct estimates of fertility and child mortality can be obtained by survey methods. However, for the foreseeable future, indirect estimates of child mortality based on census data will remain the only viable approach to studying small-area differentials in many parts of the world, and indirect estimates of adult mortality and migration still remain important approaches even at the national level.

See also: Demographic Techniques: Data Adjustment and Correction; Demographic Techniques: Small-area Estimates and Projections; Measurement Error Models; Nonsampling Errors

Bibliography

Bennett N, Horiuchi S 1981 Estimating the completeness of death registration in a closed population. *Population Index* **47**(2): 207–21

Brass W 1964 *Uses of census or survey data for the estimation of vital rates*. Paper prepared for the African Seminar on Vital Statistics, Addis Ababa, 14–19 December 1964

Brass W 1975 *Methods for Estimating Fertility and Mortality from Limited and Defective Data*. Occasional Publication of the Laboratory for Population Statistics. University of North Carolina, Chapel Hill, NC

Brass W, Hill K 1973 Estimating adult mortality from orphanhood. *Proceedings of the International Population Conference, Liège*. IUSSP, Liège, Belgium

Brass W, Macrae S 1984 Childhood mortality estimates from reports on previous births given by mothers at the time of a maternity. *Asian and Pacific Census Forum* **11**(2): 5–8

Cho L-J, Retherford R, Choe M 1986 *The 'Own Children' Method of Fertility Estimation*. East-West Center, Honolulu, HI

Feeney G 1980 Estimating mortality trends from child survivorship data: A demographic approach. *Population Studies* **34**(1): 109–28

Graham W, Brass W, Snow R 1989 Estimating maternal mortality: The sisterhood method. *Studies in Family Planning* **20**(3): 125–35

Hill K 1987 Estimating census and death registration completeness. *Asian and Pacific Population Forum* **1**(3): 8–13

Preston S 1983 An integrated system for demographic estimation from two age distributions. *Demography* **20**(2): 213–26

Preston S, Bennett N 1983 A census-based method for estimating adult mortality. *Population Studies* **37**(1): 91–104

United Nations 1983 *Manual X: Indirect Techniques for Demographic Estimation*. United Nations, New York

Zlotnik H, Hill K 1981 The use of hypothetical cohorts in estimating demographic parameters under conditions of changing fertility and mortalily. *Demography* **18**(1): 103–22

K. Hill

Demographic Techniques: Inverse Projection

1. The Character of the Method

Inverse projection is a logical inversion of conventional projection techniques. The method uses crude data—annual totals of births and deaths, and an

Table 1
Eight inverse projection algorithms for estimating population dynamics from birth and death series

Author, year	Age structure (and size) of population				Age structure of mortality	Crisis mortality	Sexes: unisex or two	Comment
	Method	Initial	Final	Net migration				
Lee 1971, 1974	forward quinquennial	model or empirical	—	population at t1 − t0 + births–deaths	adjacent model life table Lxes	—	unisex	First inverse projection program: projected quinquennial estimates for both sexes combined; used adjacent life table functions and large open-ended age group (55+); estimated life expectancy poorly for low mortality populations
Brunborg 1976	forward quinquennial	model or empirical	—	population at t1 − t0 + births–deaths	bounded higher/ lower model life table Lxes	—	unisex	Extended age groups to 80+; used high/low model life tables to improve life expectancy estimates; calibrated method using Norwegian data, 1735–1974; suggested significant technical improvements for implementing forward inverse projection method
Wrigley and Schofield 1981	back quinquennial	—	empirical required	polynomial estimation	Coale and Demeny Region North	—	unisex	Projected English population dynamics back from 1871 to 1541; migration estimated by polynomial methods

Table 1
Continued

Author, year	Age structure (and size) of population				Age structure of mortality	Crisis mortality	Sexes: unisex or two	Comment
	Method	Initial	Final	Net migration				
McCaa and Perez Brignoli 1989, 1991	forward quinquennial or annual	model or empirical	optional	population at t1 − t0 + births–deaths	bounded high/low model life table functions or empirical rates	threshold, empirical death rates by age or model functions	female, two, or unisex	Projects female and male populations separately or combined; allows interactive comparisons in a single session; facilitates the incorporation of a wide range of models or empirical data for any interval in a projection; interpolates missing data; requires minimal inputs; widely tested
Bonneuil 1990, 2000	trend, optimized annual	—	empirical required	closed population	Lederman single parameter life table functions	threshold	unisex	Estimates underregistration of deaths; permits special treatment of crisis years; newer version uses simulated annealing methodology to produce a stochastically optimized admissable reconstruction of available data

Table 1
Continued

Author, year	Age structure (and size) of population				Age structure of mortality	Crisis mortality	Sexes: unisex or two	Comment
	Method	Initial	Final	Net migration				
Oeppen 1993	general annual or other	either or both		polynomial estimation	empirical or model life table functions	—	female, two, or unisex	Maximizes fit to a wide range of available data or hypothetical constraints using non-linear optimization techniques
Bertino and Sonnino 1995	stochastic annual	empirical required		population at t1 − t0 + births– deaths	empirical or model life table functions	empirical death rates by age or model functions	partial two sex	Uses nonhomogeneous Poisson microsimulation methods to estimate standard errors; forward and back versions permit sensitivity tests; uses initial age structure for males and females, but deaths by sex not required
Rosini 1996	differentiated annual	empirical required		population at t1 − t0 + births–deaths	empirical or model life table functions	empirical death rates by age	two or unisex	Applies age-specific mortality rates, when available; projects female population independent of male

estimate of initial population size—to infer refined demographic statistics—life expectancy, gross and net reproduction ratios, and even population age structures. Instead of deriving counts from age-specific rates, as with conventional cohort projection, inverse projection estimates age-specific rates from counts. Where authentic age details for the initial population or age pattern of mortality are unavailable, model data often yield acceptable results. The technique is particularly suited for studying populations of the past where age details are scarce.

Surprisingly accurate demographic statistics can be computed with the inverse projection method. As with most demographic techniques, the better the data, the better the estimates. What is startling about inverse projection is how little data is required to simulate the demographic history of places, large and small, over long periods of time within a range of a few percentage points (Lee 1985, McCaa 1989, McCaa and Vaupel 1992). Recent innovations (see Table 1) fine-tune the method to offer several different algorithms, including a stochastic microsimulation process which estimates not only a wide range of age-dependent demographic rates and measures but most importantly their standard errors (Bertino and Sonnino 1995).

Inverse projection is a method for estimating basic demographic statistics where the vital or parish registration system is relatively efficient, but where population censuses or surveys are lacking, infrequent, or unreliable. The technique requires a count of births and deaths year by year, and the size of the initial population. In the absence of reliable empirical age data, model age structures of mortality, fertility, migration, and even the initial population may be relied upon. In contrast to family reconstitution (see *Demographic Techniques: Family Reconstitution*), inverse projection requires only aggregate data of births and deaths, tallied by year. Thus, the demographic history of large areas is easily reconstructed with inverse methods. Where family reconstitution is labor intensive, inverse projection demands little effort to tally and computerize annual counts of births and deaths. Finally, family reconstitution's stringent methodological rules exclude a sizeable fraction of births and deaths from the calculations. The reconstitutable minority is unlikely to be representative of the majority of the population. Inverse projection applies powerful nonlinear models to data for the entire population. It is superior to stable population methods because no assumption of stability is required.

The inverse method has been used to construct population histories of states (England 1541–1871, Norway 1735–1974, Sweden 1750–1875, Denmark, the Philippines, Valencia 1610–1899, Italy 1750–1911, Chile 1855–1964, Bulgaria, Costa Rica, and Cuba 1900–59), regions (Northern Italy 1650–1881, the Canary Islands 1680–1850, Sardinia 1862–1921, and Scania 1650–1760), cities, parishes, and even missions (Colyton 1545–1834, Pays de Caux 1530–1700, Lucerne 1700–1930, Berne 1720–1920, Amsterdam 1680–1921, Velletri 1595–1740, and California). This list, although incomplete, includes studies which test a wide range of demographic hypotheses, from models of population change in industrializing Europe to demographic responses to epidemics.

2. *History and Scope*

Since Lee first created inverse projection in 1971, seven additional implementations of the method have been developed (Table 1). In each case, macrodemographic data for a parish, city, region, nation, or other geographical or administrative unit are required. The algorithms differ in the use of models to supplement missing or unknown parameters. In recent decades, as computing power increased, so has the sophistication and computational demands of the various methods.

Lee's classic forward inverse algorithm (1974) remains the simplest and most straightforward. Total population is required to initiate the projection. Counts of births for each period (quinquennia in Lee's first program) are added and deaths subtracted to obtain the population at the end of each interval. Where a count of the total population is known for a subsequent period, the projected total is subtracted from the observed to obtain an estimate of net migration, which is apportioned equally among the intervening periods.

The initial age structure of the population may be obtained from either a census or a model. Lacking an empirically derived age structure, the demographically naive might stop at this point, but Lee (1974, 1985) and others have shown that even an arbitrarily chosen model age structure yields remarkably robust estimates (Brunborg 1976, Wachter 1986, McCaa 1989, McCaa and Vaupel 1992, McCaa 1993). Moreover, the bias of a poor choice rapidly disappears as the stream of vital rates molds the projected age structure period by period with births added to the base of the pyramid at age zero and deaths subtracted from the slope, age by age.

'Vital rates, not initial states' is the inverse projectionist's mantra, but this is not a matter of faith. The theoretical foundations of inverse projection lie in the 1920s with the work of H. T. J. Norton on the weak ergodicity theorem, and in the 1960s with its rediscovery by Alvaro Lopez (Wachter 1986). The method rests on what Kenneth Wachter has called 'the cornerstone of formal demography': the notion that, in the long run, vital rates, not initial states, determine what a population will look like. Lee has shown that population age structures are quickly forgotten, even when punctuated by demographically critical events such as epidemics, wars, baby booms, and baby busts

(Lee 1985). The weak ergodicity theorem teaches that when it comes to age distributions, populations have no memory. As populations are projected through time, the initial age structure is displaced by consequences of birth, death, and migration rates. For all practical purposes, populations subject to the same initial population size and identical streams of births, deaths, and migration will converge, even though they may begin from radically different age structures. A century into a projection erases even the most peculiar initial age structure. On the other hand, population size remains a powerful constraint. If it is mis-specified and the population growth rate is negative, distortions will mushroom. Where the population is growing, the effect of this bias diminishes over time (McCaa and Vaupel 1992).

The ingenuity of Lee's method is how, in the absence of data, deaths are apportioned by age (Lee 1974). As a first approximation, an arbitrary set of age-specific death rates from a single-parameter family of life tables is applied for the quinquennium. The product of these rates and the age structure at the beginning of the interval yields an estimate of the number of deaths for each age group. Adding over all groups and dividing by the total number of observed deaths yield an adjustment factor called the normalized death ratio. This factor is a measure of the discrepancy between the observed number of deaths and the number projected given the age structure and level of mortality from model life-table rates (see *Demographic Models Demographic Models*). As a second and final approximation, the normalized death ratio is used to compute the exact number of deaths at each age. These numbers will add up to the total fror the entire interval. This is accomplished directly by, first, subtracting age-specific rates between the model life table used in step 1 and an adjacent table from the same family. The result is a domain of mortality variability at each age. This domain is used to interpolate an adjusted age-specific death rate. This is accomplished by multiplying the normalized death ratio and the mortality domain at each age and adding the result to the first approximation. This yields a final estimate of deaths for each age group. Added up, they equal the total deaths for the period. Net migration by age is apportioned crudely as a function of age-specific rates and a level parameter for each period. Where migration is minor relative to mortality, this solution—required due to lack of data—produces acceptable results. Fertility statistics are by-products of inverse projection, and are wholly independent of age estimation and projection. The number of births for each time interval is determined by the stream of vital events. What is lacking is a summary fertility statistic which takes into account the age structure of the population. A solution common to all inverse projection algorithms is the estimation of the gross reproduction ratio from the number of births, the age structure of the population, and a normalized age pattern of fertility.

3. *Expansions*

Four important expansions of Lee's inverse method emerged in the last decades of the twentieth century: a full range of age groups, annual projections, crisis mortality, and two-sex models. To reduce computation costs, Lee's 1971 algorithm crudely lumped the oldest ages into a group for ages 55 and over. Extending the number of age groups beyond age 55 and over so that the final, open-ended group accounts for only a small fraction of total deaths was the first, most significant, and least controversial innovation in inverse projection methods. Brunborg (1976) quickly discovered that this shortcut introduces a substantial bias, particularly where a large fraction of total deaths occurs in the final age group. As long as life expectancy at birth is less than 30 years, the effect of lumping is small, but where life expectancy is moderately good, highly variable, or improving, error becomes substantial. Brunborg's algorithm and all subsequent ones allow the user to apply a more reasonable upper-age interval, such as 75, 90, or even 100 years and older.

A second important enhancement of the inverse method was the development of single-year projections in place of five- or 10-year periods. Anticipated by Lee and Brunborg, annual projection cycles were first implemented by Bonneuil's trend method (1993). Initially, there were objections to annual projections as spurious precision until Bonneuil demonstrated that aggregation into five- or 10-year periods can introduce serious distortions. Aggregation flattens mortality variation, and thereby flattens the population age structure whenever a mortality crisis occurs. With quinquennial or decennial projections, mortality domain interpolation dampens variation throughout the population age structure. Annual mortality spikes are transformed, at best, into waves which continue to undulate through the age pyramid over subsequent cycles. Now that computation costs are no longer a constraint, annual projections are becoming common, although some inverse projection algorithms still permit quinquennial or even decennial projections as an option.

Crisis mortality, due (say) to epidemics, famine, or other demographic catastrophes, is poorly approximated by model life tables. For the inverse projectionist, the alternative is either to apply a special set of age-specific death rates (empirical or model) for the crisis years (Bertino and Sonnino 1995) or to fix a crude death-rate threshold to mimic the impact of crises on the population age structure (Bonneuil 1993, McCaa 1993). In crisis years, mortality domain interpolation will assign too many deaths to age groups in which the spread is greatest, principally to infants, children, and the oldest or open-ended age group. The optimal solution is to use empirically derived age-specific death rates peculiar to the type of crisis. Less desirable, although widely used due to the scarcity of age-specific data for these situations, is to set a mortality threshold,

say a crude death rate of 40 per thousand population (Bonneuil 1993, McCaa and Vaupel 1992, McCaa 1993). Deaths below the threshold are apportioned according to standard inverse methodology. Any excess is then apportioned at a flat rate for each age group. Although this is an arbitrary fix, the effect is to produce more robust estimates of life expectancy and population age structure than that provided by the standard method.

Single-sex projections were the norm with the first inverse projection programs, produced when computer time was scarce. Now that this is no longer a constraint, two-sex projections are becoming frequent where births, deaths, and population size are readily available by sex. Precision is an obvious benefit, and so is the ability to model gender differences in mortality and migration.

4. Related Methods

Back projection was developed as a means of dealing with the closure problem. Wrigley and Schofield used back inverse projection for their massive reconstruction of the population of England, 1541–1871 (Wrigley and Schofield 1981, Oeppen 1993b). In the long run, most populations are not closed, and that of England is no exception. While forward inverse projection takes into account authentic migration data, Lee's method cannot produce migration flows (Lee 1993). Back projection seeks to generate migration estimates from a terminal age structure by backcasting the population against the flow of births and deaths. While it might seem commonsensical to begin with the best, i.e., most recent, data and project backward, Lee argues that this ignores the weak ergodicity theorem. In Lee's words (1993), 'the problem originates with the attempt to resurrect people who have died into the oldest age group, an attempt that is, in practice, hypersensitive to error.' The Cambridge team used iteration procedures to settle upon a single series of best estimates.

Generalized inverse projection is a response to Lee's criticisms of back projection, and broadens the method into an analytical system which exploits whatever data are available as well as a broad range of assumptions or constraints, including components derived from back projection (Oeppen 1993a, 1993b). Generalized inverse projection uses a standard method of demographic accounting and standard nonlinear optimization algorithm to overcome a range of empirical and theoretical problems. Among these is the temptation to estimate migration endogenously by means of smoothing parameters.

Other flavors of inverse projection were developed, on the one hand, to extend projections back to the first years of parish registration, where either the birth or death series begins somewhat earlier than the other (Bonneuil 1993, 2000), and, on the other, to maximize the use of data available in specific countries, such as population censuses or *status animarum* containing age details (Rosini1996). The most innovative recent algorithm is stochastic inverse projection, a microsimulation technique which seeks to assess the variability in a population's evolution (Bertino and Sonnino 1995). Instead of a single deterministic solution as with forward inverse projection, the stochastic method offers many solutions by simulating individual mortality events as a nonhomogeneous Poisson process. Three additional assumptions are required: deaths must be uniformly distributed over the year, they must be independent of each other, and every person must have the same mortality function as contemporaries at each point in time.

5. Calibration and Limitations

Lee first calibrated the method against family reconstitution results for the parish of Colyton, 1545–1834 (Lee 1974). In addition to confirming the general findings of an excellent family reconstitution study of Colyton, Lee was able to gain new insight on periodic shifts in fertility and mortality that are not readily discernible with the family reconstitution method. For example, he placed life expectancy at birth for the plague years 1645–9 at less than nine years. A second important calibration was Brunborg's reconstruction of the population history of Norway, which he compared with statistics derived from highly accurate age-specific data over the course of more than a century (Brunborg 1976). The results were remarkably consistent, tracking the official life expectancy at birth figures from a low of 43.7 years in the 1830s to a high of 74 years in the late 1960s. Average error was roughly two years. What most surprised Brunborg was how little difference the various mortality suppositions made on the estimates of life expectancy and age structure.

The most important lesson from Brunborg's experiments was that the quality of data is the critical factor for successful inverse projections. Later his findings were confirmed by a series of wide-ranging tests by a skeptical demographer, who challenged an inverse projection enthusiast to a double-blind experiment, indeed to a dozen such experiments (McCaa and Vaupel 1992). Vaupel used conventional cohort component projection techniques to simulate a wide range of demographic histories and challenged McCaa to reproduce the results for each from the bare minimum required for inverse projection: the birth and death series, and initial population size. The results were spectacularly successful; so much so that they led to the discovery of a flaw (bug) in the benchmarking program.

Vaupel's challenge led to the discovery that the required model age data could be ascertained directly by minimizing the normalized mortality ratio. This statistic can be used as a measure of goodness of fit where model age data are required, such as for the

initial population and the age structure of mortality (McCaa 1993). The k ratio is an indicator of the age-standardized intensity of mortality relative to the initial age structure of the population and to the pair of age-specific mortality schedules used to project the population by age to the next period. The sum of squares of this ratio is a useful measure of goodness of fit when the birth and death series and initial population size are held constant. In other words, where two inverse projections differ only in terms of model age structures of mortality or initial population, the squared ratio can be used to identify the better-fitting models. Best-fitting models have the lowest squared ratios. However, if the statistic is to be a reliable guide, the stream of counts and initial population size must remain the same to compare projection results.

Vaupel's challenge confirmed that data quality remains the most important constraint for successfully applying the method. Careful attention should be given to spikes in mortality or surges in migration. They may confound the normality assumption of how these events should be distributed over the age structure.

6. *Prospects*

The inverse method requires good vital registration data. For historical studies, this limits its usefulness primarily to the European cultural area. Nevertheless, even here, only a small number of potential applications have been attempted. Lee's challenge that the method be extended to the modeling of marriage and legitimate fertility has gone unanswered. The inverse method could also be extended to other demographic measures where authentic scale parameters are available and age patterns can be modeled, as in the case of parity progression ratios (Feeney 1985). The most important contribution of the method is its illumination of demographic processes in the past. In the case of England, for example, thanks to inverse projection we now know that the great surge of population in the late eighteenth century was due to a rise in fertility, rather than to a fall in mortality, as had long been supposed (Wrigley and Schofield 1981).

See also: Population Forecasts; Population Size Estimation and Capture–Recapture Methods

Bibliography

Bertino S, Sonnino E 1995 La proiezione inversa stocastica: Tecnica e applicazione. *Le Italie Demografiche. Saggi di demografia storica*. University of Udine, Rome

Biraben J N, Bonneuil N 1986 Population et Société en Pays de Caux au XVIIe siècle. *Population* **6**: 937–60

Bonneuil N 1993 The trend method applied to English data. In: Reher D S, Schofield R (eds.) *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK, pp. 57–65

Brunborg H 1976 The inverse projection method applied to Norway, 1735–1974. Unpublished ms.

Feeney G 1985 Parity progression projection. In: *International Population Conference, Florence 1985*. International Union for the Scientific Study of Population, Vol. 4, pp. 125–36

Galloway P R 1994 A reconstruction of the population of north Italy from 1650 to 1881 using annual inverse projection with comparisons to England, France, and Sweden. *European Journal of Population/Revue Européenne de Demographie* **10**: 223–74

Lee R D 1974 Estimating series of vital rates and age structures from baptisms and burials: A new technique with applications to pre-industrial England. *Population Studies* **28**: 495–512

Lee R D 1985 Inverse projection and back projection: A critical appraisal and comparative results for for England, 1539–1871. *Population Studies* **39**: 233–48

Lee R D 1993 Inverse projection and demographic fluctuations: a critical assessment of new methods. In: Reher D S, Schofield R (eds.) *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK, pp. 7–28

Leeuwenn M H D, Oppen J 1993 Reconstructing the demographic regime of Amsterdam 1681–1920. *Economic and Social History in The Netherlands* **5**: 61–102

McCaa R 1989 Populate: A microcomputer projection package for aggregative data applied to Norway, 1736–1970. *Annales de Démographie Historique*. 287–298

McCaa R 1993 Benchmarks for a new inverse population projection program: England, Sweden, and a standard demographic transition. In: Reher D S, Schofield R (eds.) *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK, pp. 40–56

McCaa R, Perez Brignoli H 1989 *Populate: From Births and Deaths to the Demography of the Past, Present, and Future*. University of Minnesota, Minneapolis, MN

McCaa R, Vaupel J W 1992 Comment la projection inverse se comporte-t-elle sur des données simulées. In: Blum A, Bonneuil N, Blanchet D (eds.) *Modèles de la démographie historique*. National Institute for the Study of Demographics, Paris, pp. 129–46

Oeppen J 1993a Back projection and inverse projection: Members of a wider class of constrained projection models. *Population Studies* **47**: 245–67

Oeppen J 1993b Generalized inverse projection. In: Reher D S, Schofield R (eds.) *Old and New Methods in Historical Demography*. Clarendon Press, Oxford, UK, pp. 29–39

Rosina A 1996 *IPD 3.0: Aplicazione automatica dell'inverse projection differenziata (passo annuale e quinquennale)*. University of Padua, Padua, Italy

Wachter K W 1986 Ergodicity and inverse projection. *Population Studies* **40**: 275–87

Wrigley E A, Davies R S, Oeppen J, Schofield R S 1997 Reconstitution and inverse projection. In: *English Population History from Parish Reconstitutions*, Cambridge University Press, Cambridge, UK

Wrigley E A, Schofield R S 1981 *The Population History of England, 1541–1871: A Reconstruction*. Harvard University Press, Cambridge, MA

R. McCaa

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Demographic Techniques: LEXIS Diagram

The quintessential building block of demographic analysis is a set of occurrence/exposure rates computed by dividing the number of occurrences of a selected demographic event in a (sub) population into the corresponding number of person-years under risk. The event may be repeatable, which means that it may occur more than once in the life course of an individual, such as geographical moves and the birth of children. Alternatively the event occurs only once in a lifetime, such as a first birth or a first marriage, or at the end of life, namely death. For repeatable events, the occurrences recorded for each age group are divided by the total number of person-years observed in the same age group, and one gets quantities like births per 1000 woman-years at ages 30 to 34. For non-repeatable events, the occurrences in the age group are divided into the number of person-years observed among population members who have not yet experienced the index event, and one gets quantities like first births per 1000 person-years among childless women. Demographic rates may be computed for a calendar period or for a birth cohort, by age attained at the time of occurrence or at the end of the calendar year of occurrence. Demographers keep track of this plethora of rates by means of a simple graphical device that has become known as the LEXIS diagram. The purpose of the present Encyclopedia article is to explain this diagram and show how it can be of help in describing usage of rates in each particular analysis.

Demographic rates are computed according to different conventions in different countries, and some countries have changed their practice over time. This makes it difficult to compare series of rates across countries and sometimes over time. To enable analysts to improve comparability, we will also provide conversion formulas that can be used to compute selected types of rates from the ingredients originally intended for the computation of other rates. We also provide methods to eliminate biases produced when standard procedures are applied to unsteady flows of births caused by major political events like wars. Such biases only affect certain birth cohorts, but in the disturbed cohorts all ages are involved.

1. The LEXIS Diagram

The LEXIS diagram provides a visual representation of occurrences and exposures across age and time. Since age and time increase in lockstep, a basic ingredient in the diagram is the lifeline of a person, showing age plotted against time on a 45-degree line starting at the time t_0 of birth and at age zero. If an event occurs to the person at a given time t, the age $t-t_0$ of that person at the time of occurrence is the ordinate of abscissa t on the lifeline.

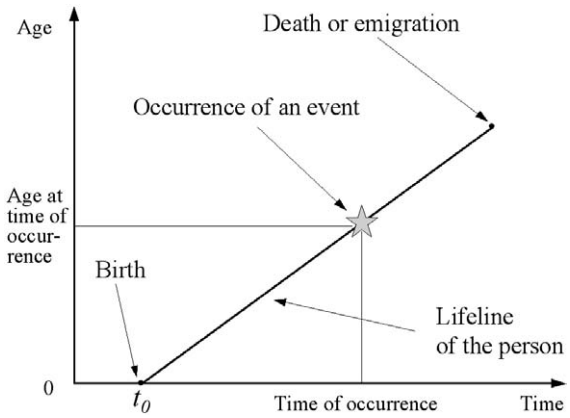

Figure 1
Lifeline of a person born at time t_0

The LEXIS diagram pertains to the residents of a given territory. The start of a lifeline may be the birth on the territory. In that case, the lifeline begins at the horizontal axis. When a person immigrates at age x at time t, the lifeline starts at the point (t, x). The end of the lifeline corresponds to the death or outmigration by the person.

Major vital events (births, deaths, marriages, divorces) are typically recorded through a vital registration system, which provides breakdowns of events by the age of the persons concerned. Residents are periodically enumerated through censuses, which record the date of birth (or the completed age on the census day) of the respondents. At least at the national level, censuses are sometimes updated every year on a fixed date, most often January 1, so that the distribution of residents by single year of (completed) age is usually available annually. The number of persons whose completed age at time t is i is the number of lifelines that cross the vertical segment $(t, [i, i+1])$. These persons were born between times $t-i-1$ and $t-i$.

Two events that occur simultaneously are located on the same vertical line in the LEXIS diagram. Two non-simultaneous events that occur to persons having the same age at the time of occurrence are located on a horizontal line. Two non-simultaneous events that occur to persons born at the same time are located on a 45-degree line.

The events that occur during a given period (e.g. a given calendar year) are located in a vertical strip, between the vertical lines corresponding to the beginning and the end of the period (Fig. 2). Those that occur to persons whose age at the time of occurrence belongs to the same age group are located in a horizontal strip. Those that occur to persons born during the same period (i.e., who belong to the same birth cohort), are located in a 45-degree strip.

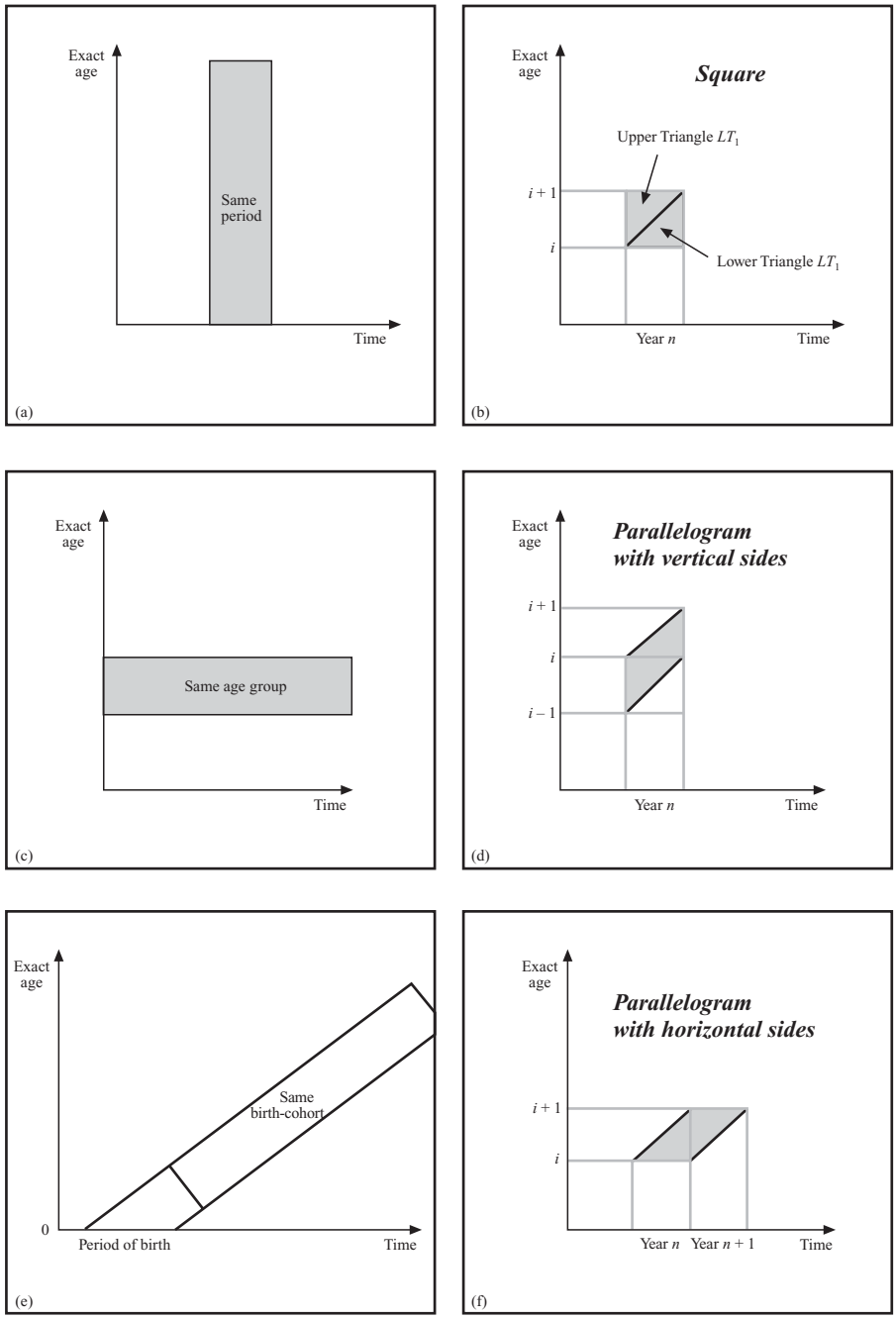

Figure 2
Regions in the LEXIS diagram. (a) Events occurring during the same period to persons of various ages. (b) Events occurring during the same year to persons who had the same age in completed years at the time of occurrence (members of two cohorts). (c) Events occurring to persons whose age was in the same age group at the time of occurrence. (d) Events occurring during the same year to persons who attained the same birthday during that year (members of the same cohort). (e) Events occurring to persons who belong to the same cohort. (f) Events occurring during two consecutive years to members of the same cohort who had the same age in completed years at the time of occurrence

2. Age

The exact age of a person on a given date t is a continuous variable. It is equal to the exact duration from the date of birth to date t. In most connections and in common language, exact age is grouped to take only integer values, however, and this makes age a discrete variable. We then refer to it as age at last birthday, or age in completed years, or completed age. The distribution of residents by age on a given date is a breakdown of the population by completed age. If the date is January 1 of year n, persons then aged i in completed years are the residents born during the year $n-i-1$.

Tables classifying the events that occurred to residents during year n by age, may be presented either by completed age at the time of occurrence or, based on the year of birth, by the age that has been or will be attained during year n. The events of year n that correspond to completed age i (as measured at time of occurrence) are located in a square of the LEXIS diagram. Those that correspond to the same year of birth i.e., to the same age i attained during the course of year n or, in other words, to the same completed age i on December 31 of year n, are located in a parallelogram with vertical sides. A third possibility that combines the two consists in retaining completed age but distinguishing persons according to whether the birthday they celebrated in year n fell before or after the occurrence of the event. When the latter distinction is made, the numbers of events refer to triangles in the LEXIS diagram, the lower triangle of the square concerns persons whose birthday in year n was before the occurrence of the event, the upper triangle those for whom it was after. Note that persons for whom an event was recorded during year n at completed age i were half a year older than those who had it recorded at attained age i on 'average' at the time of occurrence. The former had an 'average' exact age equal to $i+\frac{1}{2}$, the latter an 'average' exact age equal to i. (What we mean by 'average' here will be made clear later.) This half a year difference may be of some importance. For example, teenage fertility, defined as fertility below age 20 years, may differ by as much as 25 percent if age is defined in terms of completed age instead of age attained during the year.

3. Rates of the First and Second Kind

Suppose that one's interest focuses on a given demographic (or other) event. For each region in the LEXIS diagram where the data are available, the analyst may record the number of occurrences of the event in question. For repeatable events, one may also add up the lengths of all lifelines that intersect the given LEXIS region (or their projections on the time axis) to get the total person-years of exposure to the risk of occurrence of the event in question. The ratio between the number of occurrences and the latter exposures is the corresponding occurrence/exposure rate.

For non-repeatable events one can compute two kinds of rates. The rate of the first kind (or occurrence/exposure rate) is computed using lifelines interrupted at the time of event occurrence. Their denominator is again the 'total person-years exposed to risk.' A corresponding rate of the second kind (or occurrence/population rate) uses the same occurrences, but the denominator is computed in the same manner as for repeatable events i.e., lifelines are not interrupted at the time of event occurrence. If the non-repeatable event always terminates a lifeline, the rates of the first and second kind coincide and there is a single occurrence/exposure rate for the given region. This is of course the case if the event is death or out-migration.

Rates of the second kind are sometimes computed because the aggregate data do not provide the person-years exposed to risk but only the total number of person-years for the LEXIS region selected i.e., the person-years at risk plus the person-years for members of the population who are no longer at risk. A case in point is when first marriages observed in the relevant population segment are referred to the total resident population in the same segment, instead of only to the person-years for the never-married, perhaps because the latter exposures are not available by sex and single years of age on an annual basis. Another example arises when births of a given order are divided by the total number of person-years lived in the same population segment instead of only by the corresponding number of person-years for persons of the previous childbearing parity. This article is concerned with rates of the first kind. Rates of the second kind are discussed in a companion article, as is the connection between the two kinds of rates (see *Demographic Techniques: Rates of the First and Second Kind*).

To avoid using many decimals, the rates are normally multiplied by a convenient factor like 1000, 10,000 or 100,000 before tabulation. If the factor is 100,000, such a rate is said to be computed 'per 100,000 person-years,' where for rates of the first kind the latter are years of exposure to risk.

4. A Mathematical Framework of the Study of Occurrence/Exposure Rates

4.1 Repeatable Events

Let us consider a person exposed between ages α and ω to the occurrence of a *repeatable* event E, and let us reason as if the person is not exposed to any other event (in particular, no migration and no death). Let $\lambda(x)\,\mathrm{d}x$ be the probability that E occurs to this person

between exact ages x and $x+\mathrm{d}x$. Outside the interval (α, ω), $\lambda(x)$ is zero. Then function λ is called the risk function of the repeatable event E.

By the properties of a Poisson process, the expected total number of occurrences of E to this person is $\int_\alpha^\omega \lambda(x)\,\mathrm{d}x$. If we let $f(x) = \lambda(x)/\int_\alpha^\omega \lambda(u)\,\mathrm{d}u$, we can regard this as the density of a probability distribution with mean $m = \int_\alpha^\omega xf(x)\mathrm{d}x$ and variance $V = \int_\alpha^\omega (x-m)^2 f(x)\,\mathrm{d}x$. If the field of study is fertility and the event E is the birth of a child to a woman, the integral $\int_\alpha^\omega \lambda(x)\,\mathrm{d}x$ is called the cohort-completed fertility (CCF) and is the average number of children per woman. The mean age at occurrence is called the mean age at childbirth.

4.2 Non-repeatable Events

If the event E can happen only once in a lifetime, then $\lambda(x)\,\mathrm{d}x$ is the probability that E occurs to a given person between exact ages x and $x+\mathrm{d}x$, provided it has not occurred by exact age x. According to standard single-decrement life-table theory (see *Life Table*), the probability that the event E does not occur to a given person before age $x > a$ is $\ell(x) = \exp(-\int_\alpha^x \lambda(u)\mathrm{d}u)$. The probability that event E occurs between ages x and $x+t$, given that it has not occurred before age x, is

$$ {}_tq_x = 1-\exp\left(-\int_x^{x+t} \lambda(u)\,\mathrm{d}u\right) $$

The probability of ever experiencing E is $1-\ell(\omega) = 1-\exp(-\int_\alpha^\omega \lambda(x)\,\mathrm{d}x)$. Note that this probability is less than one if and only if $\int_\alpha^\omega \lambda(x)\,\mathrm{d}x < \infty$.

The density of the age at which E occurs, given that it ever does, is $\lambda(x)\ell(x)/[1-\ell(\omega)]$. The mean age at occurrence is

$$ m = \alpha + \int_\alpha^\omega \frac{\ell(x)-\ell(\omega)}{1-\ell(\omega)}\,\mathrm{d}x $$

The higher moments of this distribution are defined similarly. If the event E is death, considered from birth (which means that $\alpha = 0$ and ω is the maximum human life span) i.e., if the field of study is overall mortality, then the mean age at occurrence is called the life expectancy at birth. The quantity we have denoted $\lambda(x)$ is called the force of mortality at age x and $\ell(x)$ is the survival function at age x. If the event E is death considered only after a given age $a > \alpha$, the difference between the mean age at occurrence and a is called the life expectancy at age a; it is the average number of years remaining to a person still alive at age a.

If event E is first marriage below completed age 50, $1-\ell(50)$ is the proportion ever married by age 50 and $\ell(50)$ is called the final celibacy. The mean age at occurrence is the mean age at first marriage (below 50).

If the event E is the birth of a child of order k ($k \geqslant 2$) to a woman and time is recorded not from the birth of the woman but from the date of birth of her preceding child, $1-\ell(\omega)$ is the proportion who reach parity k among women reaching parity $k-1$. This is also called the parity progression ratio from order $k-1$ to order k. For $k = 1$, time is recorded from the birth of the woman or from any other suitable origin (e.g., the moment of marriage if the field of study is marital fertility) and $1-\ell(\omega)$ is the proportion of women who have at least one child. The frequency of childlessness is $\ell(\omega)$.

5. Annual Record of Events for Both Time of Occurrence and Age at Occurrence

5.1 Rates

Tables classifying events that have occurred during one calendar year, say year n, by single year of age of the person concerned, may be established by squares or by vertical-sided parallelograms in the LEXIS diagram. We let the current value of age completed be the integer i. In both cases, we have to estimate the probability of occurrence of event E within a doubly annual figure. In the mathematical framework presented above, we face three difficulties. *First*, we do not have a single person to consider but a discrete cohort born during calendar year $n-i$ in the case of a vertical-sided parallelogram, or two consecutive discrete cohorts (born in years $n-i-1$ and $n-i$) in the case of a square. In the parallelogram, each member of the cohort is exposed to risk during exactly one year (unless the person dies or migrates), but in the square the members of the two cohorts are exposed to risk during periods of variable duration ranging from 0 (for those born close to January 1 of year $n-i-1$ or close to December 31 of year $n-i$) to one year (for those born close to January 1 of year $n-i$). Furthermore, the risk function $\lambda(x)$ may now well vary with the time of birth g, in which case we should rewrite the function as $\lambda(x, g)$.

Second, the statistical data available may not be the full individual-level information, which would consist of the date of birth of each person and the various dates of occurrence of event E if it happens to occur to that person. We typically just know the *total number* of events inside the parallelogram or the square and the corresponding exposures.

Third, since our mathematical theory was developed for a single event E, migration and death (or only migration if event E is itself death) complicate the situation. The total number of events recorded includes those occurring to in-migrants or to persons who die or out-migrate subsequently to an occurrence inside the region in the LEXIS diagram; migration and deaths are duly taken into account in the total number of residents classified by age at both year-ends. But in

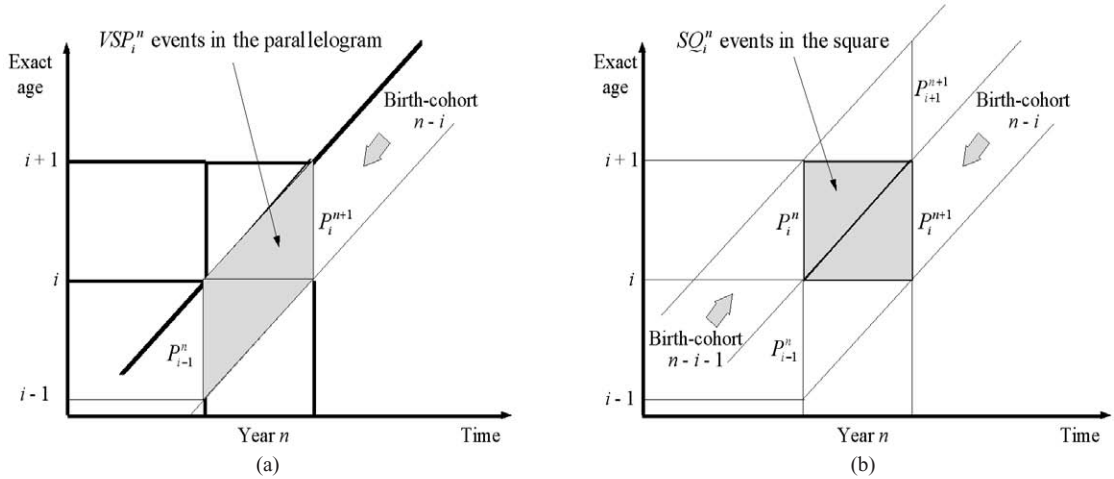

Figure 3
(a) Events observed in a vertical-sided parallelogram. (b) Events observed in a square

aggregate data the exact date of death, in-migration or out-migration of each person exposed to risk in the territory considered is not known. For these reasons, we will now reformulate our mathematical framework. We do it in two steps: first a simple one and second by a more sophisticated one.

5.2 First Simple Assumption: Constant Risk Function

Our simple reformulation holds for any kind of event, whether repeatable or not, and is as follows: Let us assume that the risk function λ is constant inside the age bracket concerned (exact ages i to $i+1$ in the case of a square, $i-1$ to $i+1$ in the case of a vertical-sided parallelogram). The probability that event E occurs is $\lambda\,\mathrm{d}g\,\mathrm{d}x$ for a resident who is born between times g and $g+\mathrm{d}g$ and whose exact age is between x and $x+\mathrm{d}x$, for every g and x inside the LEXIS-diagram region. Then the occurrence/exposure rate (of the first kind) is the maximum likelihood estimate of λ in the case of squares as well as of parallelograms. Let us furthermore assume that the number of entries due to in-migration, and exits due to death and out-migration, lead to a uniform balance: the net number of entries of persons born between g and $g+\mathrm{d}g$ when their exact age is between x and $x+\mathrm{d}x$ is $s\,\mathrm{d}g\,\mathrm{d}x$, where s is a constant.

As we already stated, net 'in-migrants' are supposed to have the same risk λ as the non-migrants. The quantity s can be approximated by $s_i^n = P_i^{n+1} - P_{i-1}^n$ inside a parallelogram with vertical sides on the basis of the numbers of residents at year-ends by single year of age. Here P_i^n is the number of residents of completed age i on January 1 of year n. (If event E is death, s_i^n is equal to $P_i^{n+1} - P_{i-1}^n - D_i^n$, where D_i^n is the number of deaths occurring during year n of residents born in year $n-i$.) If events are observed in squares, then s_{i+1}^n is an approximation to s_U in the upper triangle of the square, while s_i^n is an approximation to s_L in the lower triangle.

5.2.1 Vertical-sided parallelograms. When events are observed in vertical-sided parallelograms, the P_{i-1}^n residents on January 1 of year n, born in year $n-i-1$, who attain their ith birthday during year n, are all at risk during one year each, and the s net 'in-migrants' are at risk during half a year on average, so that the sum of exposures to risk is

$$R_i^n \approx P_{i-1}^n + \tfrac{1}{2}\,s = \tfrac{1}{2}(P_{i-1}^n + P_i^{n+1})$$

The estimate $\hat{\lambda}$ of λ is the observed occurrence/exposure rate $\hat{\lambda} = VSP_i^n / R_i^n$, where VSP_i^n is indicated in Fig. 3(a).

5.2.2 Squares. When events are observed in squares, we estimate the exposures to risk in each triangle of the square. We consider the upper and the lower triangle, one at a time. In the upper triangle, there are P_i^n persons in the population on January 1, or rather $P_i^n + \frac{1}{2}s_U$ persons who attain their $(i+1)$th birthday during year n and who are exposed between 0 and 1 year. If we introduce $m(G)$ as the average location inside year n of the birthdays attained during that year by persons born in year G, their average exposure to risk is $m(n-i-1)$. Total exposures equal $(P_i^n + \frac{1}{2}s_U)\,m\,(n-i-1)$. However, the $s_U/2$ net entries

have an average duration at risk of $\frac{1}{3}$ year. The corresponding exposures must be deducted since they are unduly included in the previous total. Thus the total exposure in the upper triangle is

$$\left(P_i^n + \frac{s_U}{2} \right) m\,(n-i-1) - \frac{s_U}{6} \approx P_i^n\, m\,(n-i-1) + \frac{s_U}{12}$$

Since $m(G)$ is anyhow close to $\frac{1}{2}$ and s is small compared to P_i^n. In the lower triangle, $P_i^{n+1} - \frac{1}{2}s_L$ persons attain their ith birthday during year n. They are exposed during an average period of $1-m(n-i)$, and there are $s_L/2$ net entries with an average duration of $\frac{1}{3}$ year, leading to total exposures of

$$\left(P_i^{n+1} - \frac{s_L}{2} \right) [1-m(n-i)] + \frac{s_L}{6}$$

$$\approx P_i^{n+1}\,[1-m(n-i)] - \frac{s_L}{12}$$

Adding up the exposures in the triangles, we get total exposures in the square of

$$P_i^n\; m(n-i-1) + P_i^{n+1}[1-m(n-i)] + \frac{s_U - s_L}{12}.$$

If the two consecutive cohorts born in years $n-i-1$ and $n-i$ are similar in size, if the mean locations of birthdays in each year for these cohorts are both close to $\frac{1}{2}$, and if migration balances are similar in the two triangles, then the exposures are close to $(P_i^n + P_i^{n+1})/2$ and the constant risk function λ is estimated by

$$\hat{\lambda} = \frac{SQ_i^n}{(P_i^n + P_i^{n+1})/2}$$

where the occurrences SQ_i^n are indicated in Fig. 3(b).

On the other hand, if cohorts born years $n-i-1$ and $n-i$ are quite dissimilar in size or if the mean locations of birthdays are quite different in these two cohorts, the latter formula may prove inaccurate. This is the case when a major disruption in the birth rate such as that provoked by the outbreak of a war occurs during the course of either of the 2 years. The consequence is then a multiplicative bias of approximately

$$\frac{B(n-i)[\frac{1}{2}-m(n-i)] - B(n-i-1)[\frac{1}{2}-m(n-i-1)]}{[B(n-i)+B(n-i-1)]/2}$$

in the above estimate of λ. Here $B(G)$ denotes the absolute number of births recorded in year G. The bias is approximately a constant percent of λ. For example, rates calculated for Belgian and French cohorts born in 1915 and 1916 are systematically biased upwards by almost 15 percent. This bias has almost the same magnitude for any kind of event (fertility, nuptiality, or mortality) and for any age. Rumanian cohorts born in 1966 and 1967 show a bias of a similar size caused by the huge increase in births subsequent to a change in the legislation on abortion in October 1966.

To avoid such biases, one should observe demographic events in vertical-sided parallelograms rather than in squares when the age-pyramid shows irregularities due to major historical events. Contrary to common language, 'statistical' age at the time of occurrence is preferably age attained during the calendar year rather than completed age at the occurrence of the event.

Note that if E is a non-repeatable event, the probability of occurrence of E during a one-year period, given that it has not occurred before the beginning of the period, is $1-e^{-\lambda}$ under the assumption of a constant risk function equal to λ, and if the observed rate r is small, the probability is estimated by $1-e^{-r} \approx 1/(1+\frac{1}{2}r)$.

The value of $m(G)$ hardly changes over years n for a given cohort G; mortality and migration do not alter substantially the characteristics of the distribution of birthdays of the cohort through time, even when these characteristics are exceptional. These characteristics can then be approximated by their value at the time of birth i.e., on the basis of the monthly distribution of births that occurred during year G. They can also be estimated from census data i.e., on the basis of a breakdown of the resident population by year and month of birth.

5.3 *More Sophisticated Assumption: Variable Risk Function*

In our second step we now turn to more refined approximations. We still assume that the risk function is independent of the time g at birth but we now allow (soft) variations with age x. The risk of occurrence of the event E at an exact age comprised between x and $x+\mathrm{d}x$ among the $N(g)\mathrm{d}g$ residents born between g and $g+\mathrm{d}g$ is $\lambda(x)\,\mathrm{d}x\,\mathrm{d}g$. The risk function λ will be assumed to be a quadratic polynomial inside several adjacent regions of the LEXIS diagram in the case of a repeatable event, and a linear polynomial in the case of a non-repeatable event. In the latter case, especially if E is death, we may alternatively assume that it is $\log[\lambda(x)]$ that varies linearly over an interval of a few years of age. This corresponds to the Gompertz law of mortality, which is generally valid at higher ages (say at ages over 30 or 40) below the oldest old. The density s of net entries due to migration (or to deaths if E is not death itself) is still assumed to be constant inside each vertical-sided parallelogram in the LEXIS diagram.

For the following reason we need to estimate the integral of $\lambda(x)$ over a 1-year interval, say between a and $a+1$, where a is either i or $i-\frac{1}{2}$. For repeatable events the mean number of occurrences per person be-

tween ages a and $a+1$ is $\int_a^{a+1}\lambda(x)\,\mathrm{d}x$. For non-repeatable events the probability of recording E between ages a and $a+1$ among those who have not recorded it before age a is $1-\exp(-\int_a^{a+1}\lambda(x)\,\mathrm{d}x)$. In both cases the integral plays a central role.

If the derivatives λ' and λ'' are not too large (i.e., if λ indeed has 'soft' variations), the integral can be estimated on the basis of available estimates of λ, λ' and λ'' as follows. We get

$$\int_i^{i+1}\lambda(x)\,\mathrm{d}x \approx \lambda(i+\tfrac{1}{2})+\frac{\lambda''(i+\frac{1}{2})}{24} \approx \lambda(i)+\frac{\lambda'(i)}{2}+\frac{\lambda''(i)}{6},$$

and

$$\int_{i-1/2}^{i+1/2}\lambda(x)\,\mathrm{d}x \approx \lambda(i)+\frac{\lambda''(i)}{24}$$
$$\approx \lambda(i+\tfrac{1}{2})-\frac{\lambda'(i+\frac{1}{2})}{2}+\frac{\lambda''(i+\frac{1}{2})}{6}$$

5.3.1 Vertical-sided parallelograms. Under the above assumptions, the number of events occurring in the vertical-sided parallelogram for a repeatable event is

$$VSP_i^n = N[\lambda+\lambda'(m-\tfrac{1}{2})+\frac{\lambda'}{2}V(\tfrac{1}{3}-m+m^2+V)]+s\frac{\lambda'}{6}$$

and for a non-repeatable event it is

$$VSP_i^n = N\left[\,1-\mathrm{e}^{-\lambda-\frac{\lambda'}{2}(1-2m)}\,\right]\mathrm{e}^{\lambda m-\frac{\lambda'}{2}(m^2+V)+\frac{\lambda^2}{2}V}$$
$$+\frac{s}{2}\left[\,2-\mathrm{e}^{\frac{\lambda}{3}-\frac{\lambda'}{6}+\frac{\lambda^2}{36}}-\mathrm{e}^{\frac{\lambda}{3}-\frac{\lambda'}{6}+\frac{\lambda^2}{36}}\,\right]$$

In these formulae, N is the number of persons who reach their ith birthday during year n. (This is estimated by the mid-value of P_{i-1}^n and P_i^{n+1} in the case of a repeatable event, and by the mid-value of P_{i-1}^n and P_i^{n+1} plus half the difference between the number of occurrences of E in the upper triangle and in the lower triangle of the parallelogram, in the case of a non-repeatable event.) Furthermore, λ, λ', and λ'' are the values of the risk function and of its two first derivatives at exact age i, s is the density of net entries inside the parallelogram, and m and V are the mean and variance of the locations of birthdays in the cohort born in year $n-i$. (If the distribution of birthdays in the cohort is *uniform*, m is equal to $\frac{1}{2}$, V to $\frac{1}{12}$, and the third moment μ is zero. When the actual distribution of birthdays differs from uniformity only due to usual seasonal variations in the number of births, there is no need to take care of non-uniformity because m is then close to $\frac{1}{2}$ and V to $\frac{1}{12}$.)

If we know N, m and V, and if we estimate λ' and λ'' on the basis of the values of λ at neighboring ages, the appropriate formula above yields an estimate for λ. This is done iteratively and the iteration process is stopped when λ has stabilized for all i. In the first iteration, λ' and λ'' are set equal to zero. In subsequent iterations, $\lambda'(i)$ may be estimated by $\frac{1}{2}[\lambda(i+1)-\lambda(i-1)]$ and $\lambda'(i)$ by $\lambda(i+1)+\lambda(i-1)-2\lambda(i)$, where λ values are those obtained in the previous iteration at adjacent ages $i-1$, i, and $i+1$.

5.3.2 Squares. The corresponding equations in the case of squares are for a repeatable event

$$SQ_i^n = N_1\left[\,\lambda m_1+\frac{\lambda'}{2}(m_1-m_1^2+V_1)\,\right]$$
$$+\frac{\lambda''}{2}\left(\frac{m_1}{4}-\frac{m_1^2+V_1}{2}+\frac{m_1^3}{3}+m_1V_1+\frac{\mu_1}{3}\right)$$
$$+N_2\left\{\,\lambda(1-m_2)-\frac{\lambda'}{2}[m_2-m_2^2-V_2]\right.$$
$$\left.+\frac{\lambda''}{2}\left[\frac{1}{12}-\frac{m_2}{4}+\frac{m_2^2+V_2}{2}-\frac{m_2^3}{3}-m_2V_2-\frac{\mu_2}{3}\right]\right\}$$
$$+\frac{s_2-s_1}{6}\left(\lambda+\frac{\lambda''}{40}\right)$$

and for a non-repeatable event it is

$$SQ_i^n = N_1\left[\,\mathrm{e}^{\lambda m_1+\frac{\lambda'}{2}[m_1(1-m_1)-V_1]+\frac{V_1}{2}\lambda^2}-1\,\right]$$
$$+N_2\left[\,1-\mathrm{e}^{-\lambda(1-m_2)+\frac{\lambda'}{2}[m_2(1-m_2)-V_2]+\frac{\lambda^2}{2}V_2}\,\right]$$
$$+\frac{s_1}{2}\left[\,1-\mathrm{e}^{\frac{\lambda}{3}+\frac{\lambda^2}{36}}\,\right]+\frac{s_2}{2}\left[\,1-\mathrm{e}^{-\frac{\lambda}{3}+\frac{\lambda^2}{36}}\,\right]$$

In the two latter formulae, $_1$ refers to the upper triangle (or to the cohort born in year $n-i-1$), $_2$ to the lower triangle (or to the cohort born during year $n-1$), N_1 is the number of persons reaching their $(i+1)$th birthday during year n, N_2 the number of persons reaching their ith birthday during year n, λ, λ' and λ'' are the risk function and its two first derivatives at exact age $i+\frac{1}{2}$, s_1 and s_2 the densities of net entries inside the parallelogram containing the upper and the lower triangle, respectively, m_1, m_2, V_1, V_2, μ_1, and μ_2 the means, variances and third centred moments of the locations of birthdays in the cohorts born in years $n-i-1$ and $n-i$, respectively.

If we know N_1, m_1, V_1, μ_1, N_2, m_2, V_2, and μ_2, and estimate λ' and λ'' from the values of λ obtained in the previous iteration at neighboring ages (with λ' and λ'' set equal to zero in the first iteration), it is possible to derive λ from the number of events in the square using the appropriate formula. The iteration process is again stopped when λ has stabilized for all i.

On the basis of the above formulae, it is possible to estimate the integral of λ over a one-year range (between exact ages i and $i+1$ and also between exact ages $i-\frac{1}{2}$ and $i+\frac{1}{2}$), whatever the LEXIS diagram region

in which occurrences were observed (squares, vertical-sided parallelograms, or triangles). This leads to an estimate of the risk regime at all relevant ages during any year of observation. These formulae make it possible to reach international comparability in spite of the diversity of national practices. (Some countries observe events in vertical-sided parallelograms, others in squares.) For countries where the national statistical office has changed from one system to the other, it also becomes possible to compute a complete time-series with comparable terms.

5.4 The Special Case of Mortality at Ages 0 and 1

The above formulae hold for fertility, nuptiality, and mortality at any age, except for mortality at age 0 and, to a lesser extent, at age 1. The reason is that the risk function $\lambda(x)$ of mortality decreases very sharply from birth i.e., from exact age 0. This makes the polynomial approximation invalid over the age-interval (0, 1) and even the interval (1, 2). We now provide a solution even for infant mortality.

Let us denote by D_L^0 and D_U^0 (with $D_L^0 + D_U^0 = SQ_0^n$) the numbers of deaths in the lower and the upper triangles of the first square (i.e., between exact ages 0 and 1), D_L^1 and D_U^1 the similar numbers of deaths in the triangles of the second square (i.e., between exact ages 1 and 2), D_L^2 and D_U^2 those in the triangles of the third square (i.e., between exact ages 2 and 3).

Under the assumption that there is no migration inside the first square and that distributions of birthdays within cohorts born in year $n-1$ and year n are identical (in practice: both departing from uniformity only due to 'usual' seasonal variations), the infant mortality rate $IMR^n = 1 - \exp(-\int_0^1 \lambda(x)\mathrm{d}x)$ can be estimated by

$$IMR^n = 1 - \left(1 - \frac{D_U^0}{P_0^n}\right)\left(1 - \frac{D_L^0}{B^n}\right)$$

where P_0^n is the population of completed age 0 on January 1 of year n and B^n is the number of live births of the corresponding sex during year n. If net migration is not zero inside the first square, we rewrite the preceding formula as

$$IMR^n = 1 - (1-a)(1-b)$$

Two estimates of each term a and b can be provided. The term a refers to the upper triangle. The first estimate assumes that migration during year n within the cohort born in year $n-1$ takes place exclusively on January 1 of year n. The second estimate assumes that migration during year n within the cohort born in year n takes place exclusively on December 31. We then get the following estimates:

$$a_1 = \frac{D_U^0}{P_1^{n+1} + D_U^0 + D_L^1} \quad \text{and} \quad a_2 = \frac{D_U^0}{P_0^n}$$

The term b refers to the lower triangle, and our first estimate assumes that migration during year n within cohort born during year $n-1$ takes place exclusively immediately after birth. Our second estimate assumes that migration during year n within cohort born in year n takes place exclusively on December 31 This gives

$$b_1 = \frac{D_L^0}{P_1^{n+1} + D_L^0}, \quad b_2 = \frac{D_L^0}{B^n}$$

It follows that a set of lower and upper bounds of IMR^n are

$$1 - [1 - \min(a_1, a_2)][1 - \min(b_1, b_2)]$$

and

$$1 - [1 - \max(a_1, a_2)][1 - \max(b_1, b_2)]$$

If migration balances are small during year n both inside the vertical-sided parallelogram corresponding to the cohort born in year $n-1$ ($s_U = P_1^{n+1} - P_0^n + D_U^0 + D_L^1$) and inside the lower triangle ($s_L = P_0^{n+1} - B^n + D_L^0$), the two bounds are naturally close to each other. For example, the lower and upper bound of the estimated infant mortality rate were 5.30 and 5.36 per 1000 male live births and 4.16 and 4.13 per 1000 female live births, respectively, in France in 1997. For most practical purposes, the differences are negligible in contemporary Europe.

A satisfactory estimate of the infant mortality rate requires that the numbers of deaths are available for triangles in the LEXIS diagram. If the information available is limited to the total number of deaths SQ_0^n inside the first square, a rough estimate of the infant mortality rate is made possible nonetheless via a preliminary estimate of the breakdown of SQ_0^n into D_U^0 and D_L^0. In contemporary Europe, D_L^0 is approximately 90 percent of SQ_0^n.

The same procedure can be applied to the estimate of the probability of death between exact ages 1 and 2. The lower and upper bounds of that probability are

$$1 - [1 - \min(a_1, a_2)][1 - \min(b_1, b_2)]$$

and

$$1 - [1 - \max(a_1, a_2)][1 - \max(b_1, b_2)]$$

with

$$a_1 = \frac{D_U^1}{P_2^{n+1} + D_U^1 + D_L^2}, \quad a_2 = \frac{D_U^1}{P_1^n},$$

$$b_1 = \frac{D_L^1}{P_1^{n+1} + D_L^1}, \quad \text{and} \quad b_2 = \frac{D_L^1}{P_0^n - D_U^0}$$

The lower and upper bounds of the probability of death between exact ages 1 and 2 are usually even

closer to each other than the two estimates of the infant mortality rate. For example, the lower and upper bounds were 4.839 and 4.843 per 10,000 male live births and 4.102 and 4.106 per 10,000 female live births, respectively, in France in 1997. In contemporary Europe, the share of the lower triangle in the second square (between exact ages 1 and 2) is only slightly over 0.5 (around 52 to 55 percent).

5.5 Literature

The LEXIS diagram got its name from the German statistician Wilhelm Lexis, who published a treatise on the subject some 125 years ago (Lexis 1875). Like so many methodological innovations, the device is really older (Vandeschrick 2000, Keiding 2000). Most straightforward computational formulas above have a similar ancestry, but the more sophisticated ones were developed by the author in a series of publications between 1984 and 1999.

See also: Age Structure; Cohort Analysis; Demographic Techniques: Rates of the First and Second Kind; Event Sequence Analysis: Methods

Bibliography

Calot G 1984 *La mesure des taux en démographie*, Cahier No. 104, *Travaux et Documents de l'INED*. Presses universitaires de France, Paris

Calot G 1999 L'analyse démographique conjoncturelle. In: Kuijsten A, de Gans H, de Feiter H *The Joy of Demography*. Edited in honor of Dirk J. van de Kaa. Thela Thesis, Amsterdam

Calot G 2000 Explanatory note on concepts and calculation methods. In: *Recent Demographic Developments in Europe*, 2000 issue of the Council of Europe Report. Council of Europe Publishing, Strasbourg, France

Calot G 1999 Statement on fertility made at the Round Table organised in Wiesbaden on the 25th anniversary of the BIB, Heft 93. Bundesinstitut für Bevölkerungsforschung, Wiesbaden, Germany

Calot G et al 1998 *Two Centuries of Swiss Demographic History*. Swiss Federal Statistical Office, Neuchatel

Calot G et al. 1999 *Un siglo de Demografía en Andalucía—La población desde 1900*. Instituto de Estadística de Andalucía, Sevilla, Spain

Keiding N 2001 *Graphical Representations in Mortality Measurement*: Knapp, Zeuner, Becker, Lexis. University of Copenhagen, Denmark

Lexis W 1875 *Einleitung in die Theorie der Bevölkerungsstatistik*. Trübner, Strasbourg, France

Sardon J-P 1992 La fécondité: Mesure et Évolution dans la Communauté Européenne. Collection *Population et conditions sociales. Série D, Etudes et Analyses*. Office for Official Publications of the European Community, Luxembourg, p. 169

Vandeschrick C 2001 The Lexis diagram, a misnomer. *Demographic Research* (in press)

G. Calot†

Demographic Techniques: Rates of the First and Second Kind

National vital statistics normally contain 100 age-specific death rates for each sex every year, namely for single years of age from age 0 through 99. There are some 30–35 age-specific birth rates for women aged 15 through 44 or 49. Corresponding rates of first marriage, when available, are also for men and women separately. Occurrence/exposure rates for nonrepeatable events (like first birth and first marriage) are called *rates of the first kind* (or of *type 1*) if the denominator of the rate is the appropriate number of person-years at risk of the event in question. For instance, a first-marriage rate of the first kind has the number of person-years recorded for the never married in the denominator. In a corresponding *rate of the second kind* (or of *type 2*), the denominator also includes the number of person-years for married and previously married members of the same population. In a first-marriage rate of the second kind, a first marriage is treated counterfactually as if it were a repeatable event. This hybrid is sometimes computed when it is impossible to separate out the person-years for the never married from the person-years of the whole population of each age. It is useful because an aggregate of age-specific rates of the second kind over ages for a birth cohort has a meaningful interpretation, as we shall see in the following.

1. Summary Measures for a Cohort

As in *Demographic Techniques: LEXIS Diagram*, let members of a given birth cohort have an intensity or risk function $\lambda(x)$ at exact age x for an event of interest. If the event is repeatable (like the birth of a child), the integral $\int_a^b \lambda(x)\,\mathrm{d}x$ would be the mean number of such events between exact ages a and b to a cohort member if there were no mortality and out-migration (see *Demographic Analysis: Probabilistic Approach* for an explanation of such a 'partial' quantity). This interesting integral can be estimated by adding up the age-specific occurrence/exposure rates over ages from a through $b-1$ in the cohort in question.

For a nonrepeatable event with the risk function $\lambda(x)$, the quantity ${}_{b-a}q_a = 1-\exp\{-\int_a^b \lambda(x)\,\mathrm{d}x\}$ would be the corresponding probability of experiencing the event before exact age b for a cohort member who had not experienced the event by exact age a. The integral $\int_a^b \lambda(x)\,\mathrm{d}x$ plays a central role for the analysis of nonrepeatable events as well, but its function is quite different from the corresponding integral for a repeatable event.

The probability ${}_{b-a}q_a$ may be estimated in two different manners, as follows.

(a) If a sequence of age-specific rates of the first kind is available for a cohort, adding them up for ages a

through $b-1$ produces an estimate $\hat{I}$ of $\int_a^b \lambda(x)\,\mathrm{d}x$, from which an estimate ${}_{b-a}\hat{q}_a = 1-\exp\{-\hat{I}\}$ of ${}_{b-a}q_a$ can be computed.

(b) If a sequence of age-specific rates of the *second* kind is available for the cohort, then adding them up over ages a to $b-1$ produces a sum that is an estimate ${}_{b-a}q_a^*$ directly. This simple fact had been understood intuitively for many years before it was proved in great generality in a sequence of papers by Hoem (1978), Finnäs (1980), and Borgan and Ramlau-Hansen (1985).

While we have described summary measures by aggregation over age for birth cohorts, similar aggregation is equally meaningful for other kinds of cohorts when age is replaced by duration since the formation of the cohort. For example, marital fertility rates can be computed by marriage duration, and adding them up over duration produces corresponding summary measures for marriage cohorts. It will not be equally meaningful to aggregate marital fertility rates across *age* groups, for not all members of a marriage cohort are at 'risk' of childbearing from the lowest age at marriage, simply because some marry at a later age.

2. Summary Measures for a Period

2.1 The Total Fertility Rate

Formally, summary measures can be computed for a calendar year, say, by aggregating over age in the same manner as for a cohort. The resulting period index is often said to be computed for a synthetic cohort (see *Demographic Measurement: General Issues and Measures of Fertility* and *Demographic Measurement: Nuptiality, Mortality, Migration, and Growth*). For a repeatable event the index is called a total rate, such as the total fertility rate, which is computed by adding up age-specific fertility rates for a calendar year over all ages.

As was pointed out by Hajnal (1947) already, it is important not to misunderstand the meaning of such a period index. The total fertility rate of year n is expressed in the same units as a cohort completed fertility (CCF), namely as an average number of children per woman, but it is *not* even some sort of average of the CCFs of the cohorts who are at fertile ages during year n. In fact the TFR^n for year n may even lie outside the range of the corresponding CCFs. To see what TFR^n actually is, suppose that the rates are computed by age attained in vertical-sided parallelograms (see Fig. 3(a) in *Demographic Techniques: LEXIS Diagram*). Let P_i^n be the number of women at age i at the beginning of year n, let R_i^n be the number of person-years recorded for women who attain age i during the year, and let B_i^n be the number of liveborn children to these women during the year. (R_i^n must perhaps be approximated by $(P_{i-1}^n + P_i^{n+1})/2$.) Then the corresponding birth rate is $f_i^n = B_i^n/R_i^n$ and $B^n = \sum_i B_i^n$ is the total number of liveborn children to women of all ages. A trivial calculation shows that

$$B^n/TFR^n = \bar{P}^n,$$

where

$$\bar{P}^n = \sum_i \frac{f_i^n}{TFR^n} R_i^n$$

is a weighted average of the mean sizes of the cohorts of women who are at fertile ages during year n. Thus, $TFR^n = B^n/\bar{P}^n$, which means that as of a fashion the total fertility rate can be seen as the number of births in year n per woman in the cohorts who are at fertile ages during that year.

The value of the total fertility rate has been around 1.4 children per woman for the European Union as a whole over a number of recent years. This means that annual cohorts of newly-born children in the EU have been one-third smaller than the mean total numbers in generations of young female adults during the same years (for $1.4/2.1 = 2/3$). If we disregard the effect of female mortality before age 30 (which has become very low today), the total number of adults at age 30 in about 30 years will be

(a) one-third lower than the mean total number of young adults in the baseline period ('today') if there is no net migration in the meantime;

(b) equal to the mean total number of young adults in the resident population today if net in-migration leads to a percentage of in-migrants one-third higher than the corresponding proportion of out-migrants born in the given area;

(c) lower than, equal to, or larger than the mean total number of young adults today according to the net percentage of in-migrants in the resident population.

In other words, compared to the present situation there will be a reduction in the total number of adults within a generation if net migration leads to a percentage of in-migrants less than approximately the rate of out-migration plus one-third.

2.2 The Total First-marriage Rate

The total first-marriage rate (TFMR) has a similar meaning. It is computed as the sum of age-specific first-marriage rates of the second kind for a given calendar year. A TFMR of 0.5 is close to the values commonly observed in Europe nowadays, and it means that the number of single women who get married during the year is half the average size of female cohorts who are at the ages of first marriage. An observed low value, say, of the TFMR may result from a *low* average level of nuptiality during the year,

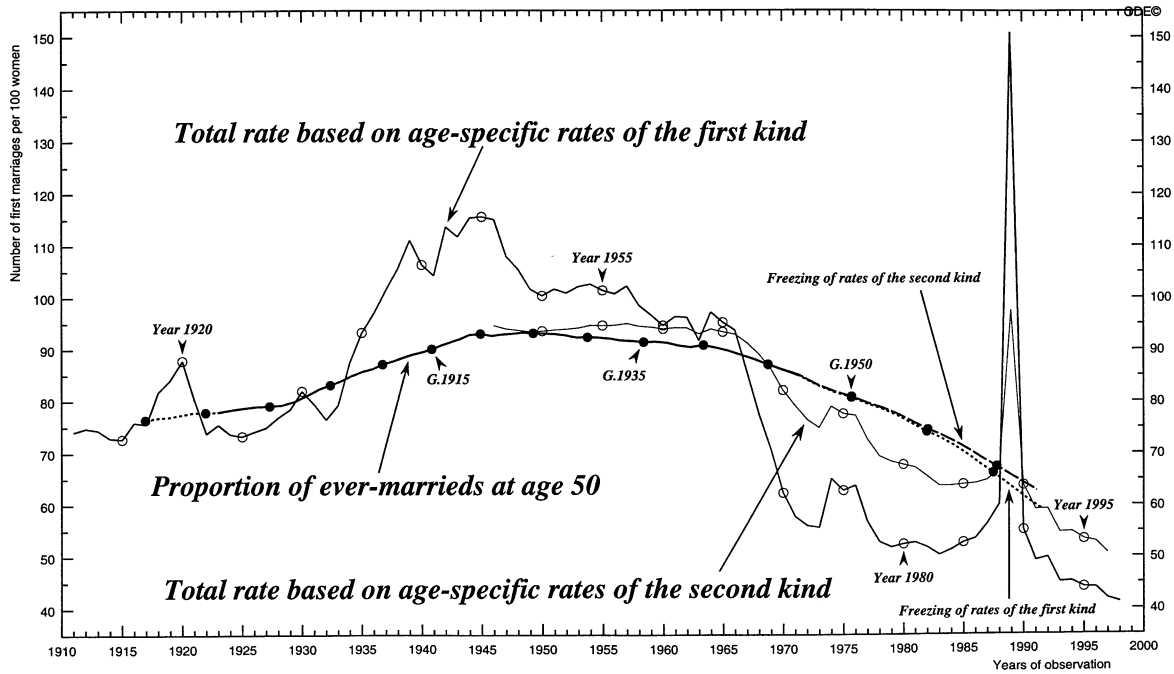

Figure 1
Sweden 1911–98. Female total first marriage rates based on age-specific rates of the first and second kind and cohort-based proportions ever married by age 50, shifted by the period-based mean age at female first marriage. Dotted lines: cohort proportions estimated by freezing rates at their values observed in 1998. Circles indicate years of observation (empty circles) or birth cohorts (filled circles) that are multiples of five

or alternatively from a *high* nuptiality in corresponding cohorts in previous, and possibly in subsequent, years.

The (period-based) total first-marriage rate computed from rates of the second kind may well exceed 1, despite the fact that no one can enter first marriage more than once. For cases in point, see for instance Vogt (1964). The corresponding total first-*birth* rates were above 1 in many countries during the baby-boom years after World War II. See Whelpton (1954) for the United States, for instance. A corresponding *cohort*-based aggregate rate can never exceed 1, nor can the single-decrement computation of the proportion ever married (or of the proportion who ever get a first birth) in a synthetic cohort based on first-marriage rates of the first kind for a calendar year ever produce such a value. The rates of the first kind for a calendar year may add up to more than 1, but this only means that the corresponding probability ${}_{\omega-\alpha}q_{\alpha}$ is larger than $1-e^{-1} \approx 0.63$ (when $\lambda(x) > 0$ only for $\alpha < x < \omega$).

Figure 1 contains total rates of first marriage (based on rates of the second kind) for women in Sweden for calendar years 1946 through 1977 along with corresponding proportions ever married by age 50 for cohorts born in 1890 through 1960. (For cohorts born after 1948 marriage rates at higher ages have been estimated.) The plot also contains aggregated first-marriage rates of the *first* kind (similarly called corresponding 'total rates' on the graph) for years 1911 through 1998. None of the total rates of the *second* kind exceed 1 in this draft, but many of the 'total rates' of the *first* kind do and often indicate a period-based value of ${}_{35}q_{15}$ in excess of 2/3. (We have used $\alpha = 15$ and $\omega = 50$ in these computations.) There was a huge peak in Swedish nuptiality in 1989 caused by a change in the regulations concerning public widow's pensions that led a substantial number of couples to marry before the end of the year, particularly among cohabiting partners (Hoem 1991).

3. *Conversion of Rates of the First Kind into Rates of the Second Kind, and Vice Versa, for Birth Cohorts*

It is useful to be able to convert cohort rates of one kind into corresponding rates of the other kind, both ways. Consider therefore a nonrepeatable event like first marriage, and let p_i^n be its rate of the second kind and π_i^n the rate of the first kind, both observed at age i during year n. Suppose that the population is closed and that there is no death and no migration in the age

bracket $[\alpha, \omega]$, α being the minimum legal age at marriage for the sex under consideration and ω being age 50 as before. Moreover, to be concrete let us assume that these rates are expressed in terms of age attained during the year.

The proportion never married at the beginning of year n among persons then of completed age $i-1$ can be computed in two equivalent ways, and we get

$$1-\sum_{j=\alpha}^{i-1} p_j^{n-i+j} = \prod_{j=\alpha}^{i-1}(1-\pi_j^{n-i+j}). \tag{1}$$

If we write out the same relation for a year later and subtract the latter from Eqn. (1) we obtain the following formula to convert π rates into p rates:

$$p_i^n = \prod_{j=\alpha}^{i-1}[\pi_i^n(1-\pi_j^{n-1+j})]. \tag{2}$$

If we divide the initial identity for year $n+1$ by that for year n and reorganize, we get, conversely, the following formula to transform p rates into π rates:

$$\pi_i^n = \frac{p_i^n}{1-\sum_{j=\alpha}^{i-1} p_j^{n-i+j}}. \tag{3}$$

The total rate for year n, based on rates of the first kind is

$$1-\prod_{i=\alpha}^{\omega-1}(1-\pi_i^n) = \prod_{i=\alpha}^{\omega-1}\left[\frac{1-\sum_{j=\alpha}^{i} p_j^{n-i+j}}{1-\sum_{j=\alpha}^{i-1} p_j^{n-1+j}}\right],$$

which differs from the similar total rate for year n based on rates of the second kind, namely,

$$\sum_{i=\alpha}^{\omega-1} p_i^n = \sum_{i=\alpha}^{\omega-1}\left\{\prod_{j=\alpha}^{i-1}\left[\pi_i^n(1-\pi_j^{n-i+j})\right]\right\}.$$

This does not prevent the aggregate rate of occurrence of the event in question for the cohort born during any year G to be the same for either kind of rates:

$$\sum_{j=a}^{\omega-1} p_j^{G+j} = 1-\prod_{j=\alpha}^{\omega-1}(1-\pi_j^{G+j}),$$

as in Eqn. (1) with $i=\omega$ and $G=n-i$.

The derivation of Eqn. (3) is a simplification of material already in the literature (Shryock and Siegel (1976 p. 292, formula 32), Finnäs (1980), Borgan and Ramlau-Hansen (1985)).

In the equations above, we have disregarded migration and mortality. In an open population, these relationships hold only if disturbing phenomena (migration and death) are independent of the event under study. For example, if the event is first marriage, this assumption essentially means that the proportion ever married among in-migrants arriving at age x is the same as in the nonmigrant population of the same age, and that never-married immigrants arriving at age x have the same propensity to marry according to age as nonmigrants in the host country. Moreover, it presupposes that only a negligible number of in-migrants who arrived as never-married marry in their home country while residents of the host country and have their marriage go undetected by the vital registration system of the host country. In some cases these latter issues are unproblematic (Finnäs 1980).

See also: Age Structure; Cohort Analysis

Bibliography

Borgan Ø, Ramlau-Hansen H 1985 Demographic incidence rates and estimation of intensities with incomplete information. *Annals of Statistics* **13**: 564–82

Finnäs F 1980 A method to estimate demographic intensities via cumulative incidence rates. *Theoretical Population Biology* **17**: 365–79

Hajnal J 1947 The analysis of birth statistics in the light of the recent international recovery of the birth-rate. *Population Studies* **1**: 137–64

Hoem J M 1978 Demographic incidence rates. *Theoretical Population Biology* **14**: 329–37

Hoem J M 1991 To marry, just in case…: the Swedish widow's pension reform and the peak in marriages in December 1989. *Acta Sociologica* **34**: 127–35

Shryock H S, Siegel J S et al. 1976 *The Methods and Materials of Demography* [Condensed Edition by E. G. Stockwell]. Academic Press, San Diego, CA

Vogt J 1964 Component parts of the number of birth. *Statsøkonomisk Tidsskrift* **78**: 287–307

Whelpton P K 1954 *Cohort Fertility: Native White Women in the United States*. Princeton University Press, Princeton, NJ

G. Calot†

Demographic Techniques: Small-area Estimates and Projections

'Small area' refers to an administrative unit that is part of a larger area, normally a country. Frequently, more detailed information is available for the large area. Survey data or administrative records are used to estimate the corresponding totals and sex-and-age distributions for the small area via statistical modeling.

In demographic forecasting small areas differ from larger areas primarily because of the greater role of migration. With the development of the geographic information systems (GIS), small area estimation and forecasting will no longer be limited by administrative boundaries.

1. Smallness

Small area is usually administratively defined. It can be a county, province, city, or municipality within a state or country, for example. In some cases one is interested in a subset of the large population defined by criteria other than geographic affiliation, such as age, sex, income, educational level, health status, etc. The latter subpopulations are referred to as *domains*.

The 'smallness' of a subpopulation can be understood either in terms of its size or in terms of the data available for it. In the case of municipalities or city blocks, populations can be small in the sense that they may vary considerably over time. Domains of a particular educational or health status may be especially variable. This poses challenges for the evaluation of educational policies or for environmental epidemiologic studies, for example. On the other hand, large cities or counties may have populations in the millions, so the populations of interest can be stable over time.

Frequently direct population estimates are available for the large area only, and survey data or administrative records must be used to estimate the population totals for the small areas. When survey data are used, the sample size available for each small area is often small and sampling variability is high. Independently of the size of the target population, estimation is then uncertain. If administrative records or other symptomatic data from the small area are used, then the biases of such data and of the statistical models used, may also be major sources of inaccuracy.

From an administrative point of view, all such estimates may have a similar status. This is the case in the allocation of funds, or apportionment of political representation, based on population size, for example.

2. Data for Small Area Estimation

Direct estimates of small area populations are continuously available for countries with high-quality population registers. In principle, direct estimates of population by educational level and subarea may similarly be obtained from computerized registers in real time. Primary problems in the statistical analysis of small area data are then caused by changes in administrative classifications, the timeliness of information updates, and quality of registration. Moreover, there are typically domains that cannot be distinguished in the register data. An example is the distribution of population by health status. It is also possible that the register based information refers to a legally resident (*de jure*) population, whereas the information needs involve the population that is *de facto* present in the area.

In countries that rely on censuses for population estimation, demographic bookkeeping, administrative records, or sampling techniques are used for small area estimation during intercensal periods. Data on births, deaths, and migration can be used for population estimation. Migration may be estimated from symptomatic data, such as school enrollment, for example.

Population estimation can also be based on *housing units* and estimates of persons per unit. The latter may be further decomposed into a product of an occupancy rate and persons per occupied housing unit. Estimates of the number of housing units can be difficult to obtain and update if there are many unauthorized additions, demolitions, or conversions. Natural disasters may also wipe out housing units. In addition to housing units, it is possible to use the number of registered passenger cars, income tax returns, etc. to estimate the population. Statistical regression analysis can be applied to census data to estimate the relationship between population or population change, and such symptomatic variables.

If the completeness of a population register (or a census) is in doubt, then the so-called *capture–recapture* (or *dual systems*) methods can be used to evaluate the level of *undercount* for small areas. It is possible to recount the population for a sample of the small areas and deduce the number missed by the registration system (or the census) in demographic, ethnic, or social domains. If not corrected, differential undercount may create biases if population data are used to allocate funds to small areas, for example.

3. Methods for Survey Based Estimation

Small area estimation may involve many other types of totals besides population. It can be income, unemployment, crop and live stock production, volume of wood in forests, etc. Such estimates may have similar administrative uses as population figures. Data on income or unemployment can be used in the allocation of funds, for example. On the other hand, for crop and forest statistics there may be additional data sources such as images taken by satellites.

In the so-called *synthetic* estimation method it is assumed that small areas have the same relevant characteristics as the large area. For example, suppose the average income is accurately known by age and sex for the large area from a sample survey, but the sample is too small for reliable estimation for small areas.

Assume that the population size is known by age and sex for a small area. Then, the synthetic estimate of the total income in the small area assumes that the area has the same average income by age and sex as the large area. The estimator has a low variance, but it can be severely biased if income varies between areas. Sample average from the small area is an unbiased estimate of income, but it has a high variance. The *trade-off between bias and variance* is the basic problem of small sample estimation.

Composite or *combined* methods try to alleviate the biases inherent in synthetic methods by using weighted averages of estimators derived according to different principles. A natural approach is to try to find weights that would *minimize the mean squared error* of the estimator. One problem in the use of such weights for administrative purposes is whether or not the same weights can be used for all small areas. Weights that have been tailored for each area separately may be statistically unstable and their use may be difficult to defend in a political process involving monetary allocations, for example.

Better accuracy is available if realistic statistical models can be developed to represent the relationship between the variable of interest (such as income) and possible auxiliary variables (such as age and sex). Possible spatial correlations across small areas may also be accounted for. Regression models that include *random effects* are typically used.

In the so-called *Empirical Bayes* method the bias of the regression model is represented by an area-specific random effect. An optimal weighted average of the regression estimator and a direct sample based estimator is, in principle, obtained by weights determined by the variance of the random effect and the sampling variance of the direct estimator. Both must be estimated from the data so exact optimality is difficult to achieve. The estimation of the variance of the random effect is particularly challenging. Different approaches have been proposed that use both analytical approximations and simulation techniques, such as *bootstrap*. The computationally intensive *Hierarchical Bayes* methods solve the problem in a unified probabilistic framework that offers conceptual advantages.

Depending on the spatial setting, complex interrelationships between the small areas are possible. Several layers of random effects may be warranted, and the effects need not be hierarchical. Such models are also called *variance component* models or *multilevel* models.

4. *Forecasts of Small Area Populations*

Migration typically has a bigger effect on forecasts of small areas than on forecasts of national populations. International migration is frequently concentrated on certain small areas so its relative impact may be large. Internal migration related to labor markets can be highly volatile. In countries without a population register, direct data on both types of migration are often lacking. Estimates must then be based on symptomatic techniques of the type discussed above. In consequence, the uncertainty of small area forecasts is typically higher. Probability theory provides a way to describe the uncertainty to forecast users in a realistic manner.

Should one forecast the populations of the small areas separately, and then aggregate to the national level, or should one forecast the national population first and then disaggregate to small areas? If the former method is applied independently for each small area by a local authority, say, then the sum of forecasts is often too high, because administrative units are reluctant to forecast declines in population for fear that it might become a self-fulfilling prophecy. On the other hand, if the forecasts are prepared by a central authority, there are many technical difficulties. The simplest approach is synthetic: one assumes that changes in the populations for small areas are a fixed proportion of a forecasted national change. At the other end of the spectrum is a multistate cohort-component model that describes migration in terms of full origin–destination migration flows by age and sex. The difficulty with the latter models is that the number of migration flows can be large. In a system of n areas it is $n(n-1)$. A compromise is to forecast out-migration for each area separately, and to pool the out-migrants. In-migrants are then forecasted synthetically as fractions of the pool for each receiving region. As in other applications, the synthetic method may lead to biases.

5. *From Administrative Units to User Defined Areas*

Much of the interest in small area estimation and forecasting comes from administrative uses such as allocation of funds. However, small area data are also used in social and medical research to measure the effects of interventions, for example. Administrative boundaries are frequently not well suited to such analyses. Domains of interest may intersect several areas, or they may be too small for their effect to be discernible in small area data. A new development in small area estimation and forecasting is the introduction of the geographic information systems (GIS). When all housing units are given geographic coordinates, the data user may define new aggregates based on the coordinates, irrespective of any existing administrative boundaries.

Estimation methods from spatial statistics will then become an essential part of small area estimation. Special care is needed when user defined regions are used to locate disease clusters, for example. Both the random variation inherent in small populations and

the selection effects must be properly handled. Existing techniques of forecasting must similarly be modified because information about migration may be more difficult to describe. On the other hand, user defined regions provide opportunities for more meaningful estimates of the population at risk.

See also: Bayesian Statistics; Censuses: Demographic Issues; Censuses: History and Methods; Demographic Techniques: Indirect Estimation; Geographic Information Systems; Population Forecasts

Bibliography

Bogue D J, Arriaga E E, Anderton D L (eds.) 1993 *Readings in Population Research Methodology*. Social Development Center, Chicago, Vol. 5, Chaps. 17, 18, 20

Ghosh M, Rao J N K 1994 Small area estimation: an appraisal. *Statistical Science* **9**: 55–76

Harala R, Tammilehto-Luode M 1999 GIS and register-based population census. In: Alho J (ed.) *Statistics, Registries and Science*. Statistics Finland, Helsinki, Finland, pp. 55–72

National Research Council 1980 *Estimating Population and Income in Small Areas*. National Academy Press, Washington, DC

Pittenger D B 1976 *Projecting State and Local Populations*. Ballinger, Cambridge, MA

Platek R, Rao J N K, Särndal C E, Singh M P (eds.) 1987 *Small Area Statistics*. Wiley, New York

Rogers A 1995 *Multiregional Demography*. Wiley, New York

Schenker N (ed.) 1993 Undercount in the 1990 census. *Journal of the American Statistical Association* **88**: 1044–6

J. M. Alho

Demographic Transition, Second

The term Second Demographic Transition was introduced into demographic literature in the mid-1980s (Lesthaeghe and van de Kaa 1986b; van de Kaa 1987). By then it had become clear that the demographic developments observed in the countries of Northern and Western Europe during the two preceding decades followed a trajectory which would not lead to a new equilibrium between the components of growth as the classical transition theory predicted. It seemed, instead, that dramatic changes in family formation and unprecedented low levels of fertility would engender population decline.

1. The Components of Growth

The decline of fertility to below replacement level is an integral part of the Second Demographic Transition, but the second transition does not consist of this fertility decline alone. [JMH: No need to be defensive.] It is constituted by a series of interlinked demographic changes resulting from a quite purposeful shift in behavior by men, women, and couples, a shift that occurred with striking simultaneity in many countries. As was the case during the first transition the demographic changes were the combined effect of structural, cultural, and technological factors. In both transitions changes in value systems played a crucial role (Lesthaeghe and Meekers 1986a). The essential change took place in people's mind. What one sees is a 'translation of cultural representations.' Again, in both transitions the three determinants of population growth moved in conjunction with each other (Coleman 1996), as follows:

(a) Europe became a region of immigration; the early recruitment of guestworkers may be regarded as the initial manifestation of that new international constellation. It was followed by family reunification, by the inflow of refugees and asylum seekers, and by immigrants seeking to enter illegally.

(b) The decline in mortality, especially at higher ages, and the quite surprising increase in life expectancy during the mid-1970s that resulted, may be interpreted as the lagged response to the behavioral choices men and women made at least a decade earlier. Eating healthier food, giving up smoking, and taking more exercise obviously affect mortality only after a certain delay. National and regional population projections now customarily assume that women will soon experience a life expectancy at birth of 85 years or more.

(c) In the area of nuptiality and fertility, the availability of new, highly effective contraception, commonly followed by increased access to abortion and sterilization, played a catalytic role. Even in the official perception of things, having sexual relations no longer presupposed being married. Male dominated contraception was replaced by means dominated by the female partner. Preventive contraception was replaced by self-fulfilling conception. The changes in fertility and family formation appear to have followed a distinct sequence, partly, no doubt, because of their logical ordering; but also because of the systematic way in which choices were made from available options (van de Kaa 1997).

In the typical pattern, a decline in fertility at the higher ages of childbearing was followed immediately by the avoidance of premarital pregnancies and of 'forced' marriages. Initially, the age at first marriage may have continued to decline, but it rose once the postponement of childbearing within marriage became more common. The, lower-order birth rates declined, and cohabitation gained ground. And, while cohabitation may have begun as a statement of defiance by a few trendsetters, and may have remained deviant for a while after that, it increasingly came to be looked upon as an alternative to marriage. Nonmarital fertility rose, voluntary childlessness became increas-

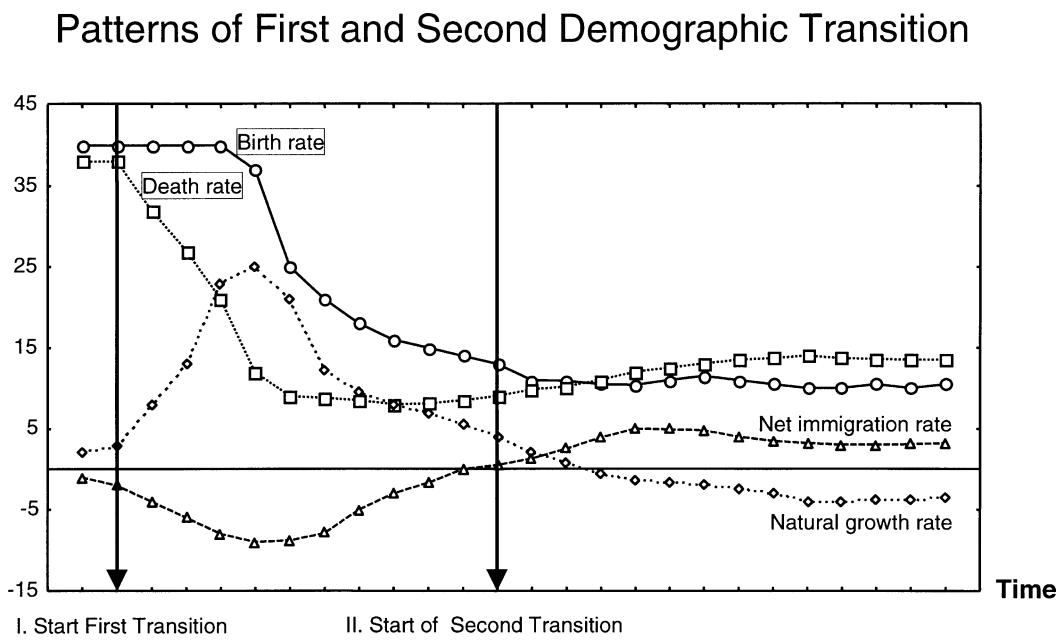

Figure 1
A schematic representation of trends in the components of population growth showing the onset of the successive demographic transitions of the early nineteenth century and the mid-1960s.

ingly significant, and union disruption became dominated by informal unions rather than marriages.

A common feature was that the postponement of childbirth away from younger ages drove total fertility down to well below replacement level. And, while a compensatory effect may occur through increased childbearing at higher ages, the upper part of the reproductive life span appears insufficient to bring cohort fertility up to replacement level in most countries. [JMH: The Nordic countries appear to be an exception, cf. recent findings by Calot and Tomas Frejka, now in the process of being published.] Ultimately the crude death rate is expected to exceed the crude birth rate. In the absence of international migration, population decline would then become inevitable.

Figure 1 illustrates how the shift from first to second transition appears to have evolved. The sinuous lines of the classical transition pattern are followed by the flat lines of the second. During the second transition the migration process is not expected to develop freely until saturation occurs. Immigration is assumed to remain controlled, but it will compensate only part of the shortfall between births and deaths.

2. *Interpretation*

This sketch fits the patterns observed in northwestern Europe quite well. An important and interesting question obviously is whether the Second Demographic Transition will prove to be as universal and ubiquitous as the first. Demographers disagree on this point. Some even doubt that the concept of a second transition is appropriate and prefer to see the developments since the mid-1960s as a temporary disturbance in the demographic transition process which began in the nineteenth century (Cliquet 1991, Vishnevsky 1991, Roussel and Chasteland 1997).

Much depends on the interpretation given to the events observed (Lesthaeghe 1995). For a while it seemed as if the second transition could be compared to a cyclone irresistibly sweeping south from Scandinavia and gradually engulfing the South of Europe before turning East and, most probably, to other parts of the developed world. It is now less evident that such a metaphor is appropriate. Three observations should be made to clarify that point. It is, first, obvious that at any point in time each region or country has its own demographic heritage and cultural endowment. The reaction to the diffusion of innovative forms of behavior will depend partly on how well new ideas can be incorporated into existing patterns and traditions.

Since the European countries represent a harlequin's mantle of experience and conditions, some specificity in reaction is to be expected. Second, It would appear that the Second Demographic Transition has a strong ideational dimension. In the 'translation' of cultural preferences and represent-

ations into behavior, deeply rooted cultural differences between large regions will make their presence felt. That, so far, southern Europe has not embraced cohabitation as an alternative to marriage and does not show wide acceptance of nonmarital fertility may well reflect a fundamental difference between the 'protestant' North and the 'catholic' South. It should, third and finally, be recognized that if socioeconomic and political conditions are unfavorable to these novel patterns of behavior, a diffusion process may be much delayed.

This probably partly explains the difference in demographic patterns (trends?) between the previously socialist countries of Eastern and Central Europe and the rest of the developed world. Conceivably, such differences and lags may disappear very rapidly once crisis conditions abate, the economic situation improves, and generations exposed to mass culture from the West during their adolescence reach the age at which union formation begins. The Czech Republic has shown extremely rapid shifts in the handful of years between 1992 and 1996, and this may signal developments elsewhere in years to come.

Whether all countries of the world will, ultimately, undergo a Second Demographic Transition is much too early to tell. If Inglehart (1997) is correct in his conclusion that at a certain stage of their social and economic development countries reach an inflection point, and then leave the trajectory of modernization to start moving on the trajectory of postmodernization, a universal second transition is not at all unlikely. He sees postmodernization as a '… pervasive change in worldviews.' 'It reflects a shift in what people want out of life. It is transforming basic norms governing politics, work, religion, family, and sexual behavior' (Inglehart 1987, p. 8). Measuring postmodernization (in demographic behavior?) is not particularly easy (van de Kaa 1998).

It essentially implies that at an advanced stage of social and economic development people begin to question the meta-narratives which underpinned the process of modernization for several centuries (e.g., the belief in progress, in the value of working hard, in religion, in obedience to—secular—authority). This means that, just as has been well documented for countries of the European Union after the mid-1960s, concerns about law and order, physical security, and the standard of living give way to higher order needs. The quality of life becomes of prime importance and is placed above material assets in people's perceptions. People seek personal freedom and gender equality, and emphasize the right to self-realization and self-expression. They express support for emancipatory movements and express concern about the maintenance of human rights, biodiversity, and the ability to sustain modernization style developments (more than even before?).

The probable association of the Second Demographic Transition with long term processes of development in Europe's societies and/or civilization needs to be researched very rigorously before firm evidence can replace plausible conjectures.

Bibliography

Cliquet R L 1991 *The Second Demographic Transition: Fact or Fiction?* Council of Europe Population Studies No. 23, Strasbourg

Coleman D (ed.) 1996 *Europe's Population in the 1990s*. Oxford University Press, Oxford, UK

Inglehart R 1997 *Modernization and Postmodernization. Cultural, Economic, and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ

Leridon H 1999 Les nouveaux modes de planification de la famille en Europe [The new manner of family planning in Europe]. In: Van de Kaa D J et al. (eds.) *European Populations. Unity in Diversity*. Kluwer Academic Publishers, Dordrecht/Boston/London, pp. 51–77

Lesthaeghe R 1995 The second demographic transition: An interpretation. In: Mason K O, Jensen A-M (eds.) *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK, pp. 17–62

Lesthaeghe R, Meekers D 1986a Value changes and the dimensions of familism in the European Community. *European Journal of Population* **2**(3–4): 225–68

Lesthaeghe R, Van de Kaa D J 1986b Twee demografische transities? [Two demographic transitions?]. In: Van de Kaa D J, Lesthaeghe R L (eds.) *Bevolking: Krimp en Groei*. Van Loghum Slaterus, Deventer, pp. 9–25

Roussel L, Chasteland J-C 1997 Un demi-siècle de démographie dans les pays industriels. Quelques réflections sur un bilan [Half a century of demography in industrialized countries. Some reflections on a balance sheet]. In: Chasteland J-C, Roussel L (eds.) *Les Contours de la Démographie au Seuil de XXX Siècle. Actes du Colloque International un Démie Siècle de Démographie. Bilan et Perspectives, 1945–1995*. Ined-PUF, Paris, pp. 9–29

Rychtarikova J 1999 Is Eastern Europe experiencing a second demographic transition? *Acta Universitas Carolinae, Geographica* **1**: 19–44

Van de Kaa D J 1987 *Europe's Second Demographic Transition. Population Bulletin*, **42**: 1–57

Van de Kaa D J 1997 Options and sequences: Europe's demographic patterns. *Journal of the Australian Population Association* **14**(1): 1–30

Van de Kaa D J 1998 *Postmodern Fertility Preferences: From Changing Value Orientation to New Behaviour*. The Australian National University, Canberra

Vishnevsky A 1991 Demographic revolution and the future of fertility; a systems approach. In: Lutz W (ed.) *Future Demographic Trends in Europe and North America*. Academic Press, London, pp. 257–80

D. J. van de Kaa

Demography, History of

Demography is the study of processes of change in human populations. Statistical measures of fertility and mortality and their dynamic inter-relationship are

at its core. Demography consists of three inter-related intellectual activities: the collection of appropriate evidence, analysis using a body of mathematical and statistical techniques which avoid certain pitfalls, and the interpretation of the relationship between demographic processes and the political, social, economic, cultural, and natural environment.

As an organized academic subdiscipline it is a recent arrival in the social sciences. There were only a handful of established academic posts in demography, mostly privately funded in US universities, before the Second World War. Yet, by virtue of its obvious importance in terms of *raison d'état*, enumeration of populations for fiscal, military, administrative, and colonial purposes dates back to the world's ancient civilizations (and the Domesday Book demonstrates its sporadic occurrence in the medieval period, too). Furthermore, compilation of family genealogies and maintenance of vital registers of births, marriages, and deaths for religious and for legal purposes both have venerable histories, now proving to be of value for demographic historians of early modern Europe, North and South America, Japan, and China. Proper demographic study requires two kinds of information: accurate and regular census details (now taken decennially in most countries) on the stock of the defined population (numbers of each sex alive at each year of age); and information on the continuous flow of vital events, the numbers of each sex born alive, and the numbers dying at each exact age. This permits calculation of the age- and sex-specific rates of fertility and mortality necessary for genuinely comparable measures of a population's principal dynamic demographic characteristics, while also allowing for the effects of net in- or out-migration. Many further kinds of demographic analysis can be undertaken if the appropriate information is collected.

The study of populations of any size presupposes an official apparatus of data collection of some considerable size and sophistication and demography is therefore unique among the social sciences in the extent of its intimate involvement with the priorities of governments and officials and the wider influences of politics upon the state. Until recently only government bureaucracies could command the clerical resources required for the statistical analysis. Hence, government census bureaux provided the initial demand for automated punch-card tabulating machines and later for the production of the first nonmilitary electronic computers (the US Census Bureau's Univac, commissioned for the 1950 census). The subsequent proliferation of cheap computing resources has undoubtedly been a necessary precondition for the expansion of demography as an academic discipline with numerous independent, nonofficial practitioners.

Demography has therefore had a close historical association with the rise of nation-states throughout the entire modern period. But demography's early development has also been strongly influenced at times by the needs of smaller, subnational polities, such as the municipalities of Britain since the seventeenth century or the federated states of eighteenth-century Germany. Furthermore, the institutional and intellectual character of demography's rapid disciplinary growth since World War II has been much influenced by explicitly transnational institutions such as the UN and the World Bank and major independent (though mostly American) philanthropic organizations, such as the Population Council, the Rockefeller and the Ford Foundations. Nevertheless, national governments remain the principal collectors of continuous series of demographic data, so that the nation remains the typical 'population' addressed.

Modern demography's origins are conventionally traced to the practitioners of 'political arithmetic' in Restoration England. From the sixteenth century onwards it was increasingly the practice of European cities to compile bills of mortality—weekly death records—as an early warning system for epidemics, particularly of plague. In London the existence of this data provided the immediate context for the first recorded efforts at demographic analysis by John Graunt (1629–74), 'the father of modern demography.' In his *Natural and political observations made upon the bills of mortality* (1662) he also used records of London christenings to develop a rudimentary Life Table. This is the fundamental technique for the correct analysis of mortality because it divides the population into age sections and gives statistics for the probability of dying at each age, which can then be arithmetically summed to a single average life expectancy at any age, enabling populations with vastly different age structures or mortality regimes to be compared. Graunt had thus grasped the two fundamental demographic insights: that the age and sex composition of a population powerfully influences the incidence of vital events; and this means that population processes are all significantly interdependent.

Life table techniques developed very gradually over the ensuing century, with notable work by Sir Edmund Halley (1656–1742), the Dutchman Willem Kersseboom (c.1690–1771), the German Johann Sussmilch (1707–67), the Frenchman Antoine Deparcieux (1703–68), the Swiss-born Leonhard Euler (1707–83), and a host of British contributors to the so-called 'population controversy' (over whether or not the kingdom's population had increased since the Glorious Revolution of 1688–9, when the civil servant Gregory King had compiled, from tax returns, the only detailed estimate of a European nation's population made in the early modern period). Finally, the first technically correct empirical life table was published in 1815, for commercial actuarial purposes, by Joshua Milne (1776–1851), using accurate census and vital registration data collected for public health purposes in Carlisle during the years 1779–87.

During the nineteenth century the principal vehicle for the further development of demographic analysis

was the General Register Office (GRO) in London. This British lead was an unlikely development, in that the British census was not even established on a permanent basis (it had to be invoked by a separate Act of Parliament every 10 years), whereas the Scandinavian countries of Iceland (1703) and Sweden (1749) had already instigated regular census-taking, while the new, expanding American federation had also committed itself to regular censuses, as integral to its process of political representation. The French and Belgians, with Condorcet, Laplace, Duvillard, Fourier, and Quetelet led the world in the relevant mathematical fields of probability and the application of statistics to human affairs.

The British lead that opened up was due to a chain of chance occurrences. The GRO was created by statute to oversee from 1837 a national civil registration system of births, deaths, and marriages (created not for public health or demographic purposes but as a liberal legal reform granting dissenters a record of their vital events independent of the registers of the Established Church). As a temporary expedient this small department of government then became vested in 1840 with responsibility for taking the nation's fifth census of 1841 (the first was in 1801). The Registrar-General's statistical assistant, recently appointed through Court patronage in the as yet unreformed civil service, was an unknown man of humble origins, medically trained and a self-taught statistician (partly on a visit to Paris). A representative member of the European-wide generation enthused by the promise of a statistical science of society inspired by the Belgian astronomer, Adolphe Quetelet (1796–1874), William Farr (1807–83) immediately seized the opportunity presented by his responsibility for processing the two essential sources required for demographic analysis. He launched a four-decades long career of innovative demographic and comparative epidemiological work. Britain was at this time in the throes of a profound mortality crisis in its industrial cities (in addition to rural Ireland's fatal famine) and Farr saw his work as a public health campaign.

Thus, while it was in a Parisian publication that the term 'demography' was first coined in 1855—by Quetelet's countryman, Achille Guillard—it was in Britain that it was most intensively practiced and developed as 'vital statistics.' Significant work was certainly done in other official statistical offices. Indeed, the most important subsequent technical innovation was the Lexis diagram invented by the German statistician, Wilhelm Lexis, in 1875 for representing the relationship between cohorts (or generations) and period effects (events).

The GRO's lead was surprising in another sense in that the single English name most famously associated with the origins of demography is not the civil servant William Farr with his primary medical focus on preventable mortality, but the cleric and academic Thomas Robert Malthus (1766–1834), whose focus was on the economy and preventable fertility. His conservative polemic, *An essay on the principle of population*, published anonymously in 1798, enunciated that 'Population, when unchecked, increases in a geometrical ratio. Subsistence increases only in an arithmetical ratio.' Malthus, who held the first ever Chair in Political Economy in Britain and regarded Adam Smith as his principal mentor, thereby applied the fundamental economic problem of scarcity to demography and produced an analysis which famously stimulated Darwin to apply a similar logic to the problem of the origin of species, with profound intellectual results.

Malthus was the seminal influence on demography, in terms of the third intellectual activity outlined above, interpretation. His forthright axioms and policy implications demanded a response from the intellectuals of every country and secured a permanent place in the social and biological sciences for the discussion of population problems. In the much revised and extended editions of his *Essay* he continued to feed the debate, further analyzing and exploring the manner in which aggregate population change, mortality and migration, and individuals' and families' reproductive behavior is integrated into the encompassing natural, social, and economic environment, while using examples drawn from across the world to show how profoundly they were also influenced by customs, beliefs, and politics. He notoriously drew the wrath of Marx for his support for the radical reduction of the nation's welfare net with the amendment of the Poor Law in 1834, on the grounds that its 'generosity' was discouraging prudence among the prolific poor, who married too young and so had too many children.

Throughout the remainder of the nineteenth century most of those studying demographic phenomena were gentlemen scholars in local and national statistical societies, medical men, municipal and some state officials scattered all across Europe and North America. From mid-century they met periodically at International Congresses of Statistics, then of 'Hygiene and Demography' and eventually of 'Eugenics.' But it was not until the interwar decades of the twentieth century that demography began to emerge with an independent disciplinary and professional identity, marked initially by the formation in 1928 of the International Union for the Scientific Study of Population, followed in 1931 by the foundation of the Population Association of America.

However, this formative era from the late nineteenth century onwards had also witnessed the rise of nationalist and imperialist rivalries and associated racist, social Darwinist ideologies. This came to exert a dominant (though not uncontested) influence over the practice of demography during the interwar years. French demography was perhaps the first to experience this overtly nationalist and xenophobic current, in the form of a defensive pronatalist popular reaction

to defeat in the Franco-Prussian war of 1870–1. This was partly ascribed to her extremely low birth rate, the French being the first nation in the world to have practiced widespread family limitation. The titles of the works of France's leading demographers, such as Arsène Dumont, Jacques Bertillon, and Adolphe Landry repeatedly referred to 'Dépopulation' and 'decadence' while her great novelist, Zola, writing in exile in London after the Dreyfus affair, mourned France's lost *Fécondité* (1899).

The study of population problems in the interwar period became increasingly subsumed within this social Darwinist agenda. With its continuing necessary association with the state, demography was inevitably heavily implicated in the work done under the Fascist and Nazi regimes with their totalitarian ideologies. For this reason international conferences in both Rome and Berlin in the 1930s were partially boycotted and disrupted. But even in more liberal Britain, USA, or social democratic Sweden, demographic analysis became obsessed with class-differential fertility, 'problem families,' and fears of a 'flight from parenthood' associated with a succession of illiberal movements for national efficiency, eugenics, and 'social hygiene.' Only in the USSR was population analysis relatively free from this association (until Stalin's own pronatalist drive during the war) but this was primarily because of the absence of demographers, due to Marx's historic demonization of Malthus as the ultimate apologist of the bourgeoisie and their 'dismal science' of political economy. Given the bitter interwar political legacy in Europe, it was only in France and Britain that institutions for demographic research, L'Institut National d'Etudes Demographiques under Alfred Sauvy (1898–1990) and the Population Investigation Committee of the London School of Economics under David Glass (1911–78), positively flourished in the immediate aftermath of the war, both founding their respective journals in 1946, *Population* and *Population Studies*.

The decade of the Second World War marks a significant shift in which the center of gravity of the discipline of demography migrated from Europe to the USA. It was symbolic of this coming change that in work published between 1907 and 1939 a European-born but American-naturalized mathematician, Alfred Lotka (1880–1949), devised the proofs for the equations of renewal processes which underlie stable population theory (showing that populations always come to exhibit a constant age structure and growth rate if their fertility and mortality rates are unchanging). This has proved to be the most important mathematical advance in demography since the perfection of the Life Table technique (which applied to nonrenewal processes). Intriguingly, unknown to Lotka, Euler had already solved these equations in the eighteenth century. Indeed, Lotka's work remained largely unexploited until after the Second World War, both facts which demonstrate that technical advances within disciplines do not in themselves necessarily cause change but are dependent on the institutional context and on 'demand' factors external to the discipline.

After the Second World War, demography rapidly responded to the 'demand' of the growing global preoccupation with planned economic growth, propelled in this direction by US strategic interests in the context of the Cold War rivalry with the USSR for client states among the many decolonizing peoples of the less developed world. Consequently, the second half of the twentieth century has become a relatively clearly demarcated third era in the history of demography when, in some senses, it has returned to the original Malthusian agenda of a preoccupation with economics and with policies to promote fertility restraint, especially among the world's poor. The idea of 'demographic transition' has dominated the discipline's approach throughout this era. This is a general historical model, promoted and adapted by Frank Notestein (1902–83), the highly influential and well-connected first Director, 1936–59, of the Princeton Office of Population Research (OPR), the most important institutional center for the new American-led demography. Demographic transition 'theory' proposes that in order for a nation to achieve sustained long-term economic growth it must reduce both its mortality and, especially, its fertility levels to those attained by the 'modernized,' liberal democratic Western countries. The model has surprisingly little consistent support in either the historical or contemporary record but it has survived repeated revisionist challenges and has undoubtedly served the discipline well as an overarching ideological position which has successfully elicited a continuous flow of funds for both research and family planning programs.

It was in this policy context that the demand for various applications of stable population theory arose. Within the dominant postwar liberal consensus of 'modernization theory,' economic growth was seen as a generic process, such that less developed countries could be engineered to repeat the experience of the developed countries. Fertility restraint and mortality reduction were important population policy inputs to permit the occurrence of incremental capital:output ratios, considered essential for industrialization. This would be jeopardized by too great a demographic dependency burden (too many young and invalid old in the population) diverting income into current consumption instead of into savings and capital accumulation. Stable population theory enabled model life tables to be devised by A. J. Coale (Notestein's successor as Director of OPR) and indirect estimation techniques to be formulated by W. Brass (1921–99) of the London School of Hygiene, which permitted the measurement of demographic parameters in many of the less developed countries where primary information, particularly vital regis-

tration, was defective. A further application, for the same reason, has been its use in back projection and generalized inverse projection in historical demographic work. Along with the use of ground-breaking work by Louis Henry, this has enabled E. A. Wrigley and his colleagues at the Cambridge Group for the History of Population to utilize parish registers to reconstruct the detailed demographic record of the first industrial nation, England, back to the mid-sixteenth century, thereby, incidentally, demonstrating the invalidity of the demographic transition model in this important case.

Demography has achieved a considerable degree of disciplinary and institutional success, in many senses enviable by the standards of the other social sciences. Since 1945 training institutions have proliferated, primarily under the aegis of the Population Council, so that demographers are now trained in many centers in all the continents. By plausibly reducing, for certain purposes, the complexity of human affairs in matters of reproduction, morbidity and mortality, migration and mobility to tractable mathematical models, demography has achieved considerable analytical power. Like classical and neoclassical economics, conventional demographic analysis embraces a positivist epistemology. It also starts from the liberal and 'democratic' foundational fiction that each individual can in principle be legitimately treated as an equivalent accounting unit (a presumption that could only make sense in a post-Lockean world, where a working man could be deemed to count for the same as a prince, hence demography's first text by John Graunt came from the hand of an exact contemporary of John Locke's, another survivor of the modern world's first overthrow of a supposedly divinely-appointed sovereign). The resulting accounting system has provided the epistemological basis for an increasingly elaborate range of useful technical developments (such as, most recently, multistate and multiregional demography), for which the separate entries on 'Demography' may be consulted.

Inevitably, however, there is an intellectual cost for the preservation of demography's analytical rigor. Exceptional as it is, this has rather cut off demography from the rest of the social sciences and humanities, such that the entire 'linguistic turn' has gone largely unremarked in this discipline. Like all scientific activities, demography socially and ideologically constructs its objects of study but there has been little explicit acknowledgment of the implications of this. Thus, for a discipline which has devoted much energy to refining, for instance, calibration of the proximate (biological and physiological) causes of fertility variation, there has been much less interest in understanding how the demographers' own concept of fertility itself has arisen, and this applies to other social categories, such as 'population,' 'family,' or 'household.' This disinclination to engage in critical reflexive practice has undoubtedly had significant and, for such an applied and policy-oriented discipline, real consequences. For many decades vast resources have been devoted to 'family planning' programs aiming to transplant 'modern,' 'western' contraceptive practices into an enormous variety of diverse cultures. More policy effectiveness might have been forthcoming if both fertility and birth control had been initially approached as culturally variable entities, whose meanings had to be researched in any population before attempting to modify them.

The future of demography, established as it now is in all the world's major cultures, is likely to include much more active discussion of the relationship between its inherited inventory of 'Western' social categories and those of other peoples, something also likely to follow from the increasing sophistication of historical demography. It is likely to be less consumed with the problems of engendering economic growth and more with social equity and public health issues, including those of longevity and 'aging,' along with ecological problems of the uses and distribution of material wealth. This would represent a return, on a global scale, to the issues which so preoccupied the mid-nineteenth century demographic observers of the human disruption wreaked in Britain and Europe by the world's first great wave of rapid economic growth. At the same time greater reintegration of demography into the intellectual concerns of the other social sciences seems long overdue and likely to be mutually beneficial. In particular, with its technical core in an abstract accounting approach, demography has remained relatively gender-neutral in its analyses, in the sense that it has singularly failed to attempt to grapple with the linguistic and epistemological complexities that arise from acknowledging a negotiated and dialectical relationship between the two sexes in all that they do. A genuinely two-sex demography is likely to constitute the next fundamental advance.

See also: Databases, Core: Demography and Genealogies; Databases, Core: Demography and Registers; Demography: Twentieth-century History; Families and Households, Behavioral Demography of; Families and Households, Formal Demography of; Global Population Trends; Historical Demography; Malthus, Thomas Robert (1766–1834); Population Aging: Economic and Social Consequences; Population Cycles and Demographic Behavior; Population Dynamics: Theory of Stable Populations; Population Forecasts

Bibliography

Anderson M J 1988 *The American Census. A Social History*. Yale University Press, New Haven, CT

Desrosières A 1998 *The Politics of Large Numbers: A History of Statistical Reasoning*. Harvard University Press, Cambridge, MA

Dupâquier J 1984 *Pour la démographie historique*. P.U.F. Paris
Dupâquier M, Dupâquier J 1985 *Histoire de la démographie*. Perrin, Paris
Eyler J M 1979 *Victorian Social Medicine. The Ideas and Methods of William Farr*. Johns Hopkins University Press, Baltimore, MD
Foucault M 1991 Governmentality. [tr. Pasquino R]. In: Burchell P, Gordon C, Miller P (eds.) *The Foucault Effect: Studies in Governmentality, with two lectures by and an interview with Michel Foucault*. Harvester, Hemel Hempstead, UK, pp. 87–104
Girard A 1986 *L'institut national d'études démographiques. Histoire et developpement*. INED, Paris
Glass D V 1973a *The Development of Population Statistics*. D C Heath, Farnborough, UK
Glass D V 1973b *Numbering the People. The Eighteenth-Century Population Controversy and the Development of Census and Vital Statistics in Britain*. D C Heath, Farnborough, UK
Greenhalgh S 1996 The social construction of population science: An intellectual, institutional and political history of twentieth-century demography. *Comparative Studies in Society and History* **38**(1): 26–66
Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK
Hauser P M, Duncan O D (eds.) 1959 *The Study of Population. An Inventory and Appraisal*. University of Chicago Press, Chicago
Hodgson D 1983 Demography as social science and policy science. *Population and Development Review* **9**(1): 1–34
Hodgson D 1991 The ideological origins of the Population Association of America. *Population and Development Review* **17**(1): 1–34
Ipsen C 1996 *Dictating Demography. The Problem of Population in Fascist Italy*. Cambridge University Press, Cambridge, UK
Kreager P 1988 New light on Graunt. *Population Studies – Journal of Demography* **42**(1): 129–40
Kreager P 1992 Quand une population est-elle une nation? Quand une nation est-elle un état? La démographie et l'emergence d'un dilemme moderne, 1770–1870. *Population* **6**: 1639–56
Notestein F W 1982 Demography in the United States: A partial account of the development of the field. *Population and Development Review* **8**(4): 651–87
Ryder N B 1964 Notes on the concept of a population. *American Journal of Sociology* **69**: 447–63
Schweber L in press *Disciplining Statistics: Demography and Vital Statistics in France and England, 1835–1885*
Soloway R A 1990 *Demography and Degeneration. Eugenics and the Declining Birthrate in Twentieth-century Britain*. University of North Carolina Press, Durham, NC
Szreter S 1991 Special issue: The General Register Office and the Public Health movement 1837–1914. *Social History of Medicine* **4**: 435–63
Szreter S 1993 The idea of demographic transition and the study of fertility change: A critical intellectual history. *Population and Development Review* **19**(4): 659–701
Szreter S 1996 *Fertility, Class and Gender in Britain, 1860–1940*. Cambridge University Press, Cambridge, UK
Teitelbaum M, Winter J 1985 *The Fear of Population Decline*. Academic Press, New York

S. Szreter

Demography: Twentieth-century History

Achille Guillard first used the word 'demography' in 1855 to describe a new science of population dynamics, by which he meant the statistical study of the size, condition, and movement of populations. Malthus, a half century earlier, had initiated a period of speculation on population growth's constraining effect on prosperity, but only by mid-century had the collection of population statistics become extensive enough in Europe and the USA for a 'science' of population to be imagined. The rise of industrial capitalism in Europe and its overseas extensions increased the pace of social, demographic, and economic change, stimulated the development of state structures aimed at directing these changes, and made obvious the usefulness of accurate population statistics. As India's first census in 1868 reflects, such usefulness extended to the administration of colonies as well. In 1882 the fourth International Conference on Hygiene (Geneva) included a section on demography. By its next meeting in 1884 (The Hague) the conference included demography in its title and sections on 'the science of statistics applied to the social well-being of people' were presented. Although the growing body of population statistics was being discussed at international meetings, as the nineteenth century came to an end a unified discipline of demography, with a distinct body of theory and method, had yet to emerge.

For an intellectual activity to become a discipline certain things are needed: a significant interaction among those performing the activity; an organization that directs its investigation and teaching; and entities outside itself who value and use its findings. Using this definition, demography lacked the status of a discipline for much of the twentieth century. The century began with a diverse group of individuals considering themselves population specialists: statisticians, public health officials, biologists, economists, sociologists, geographers, historians, lawyers, political activists, and politicians. There was haphazard interaction among this group and the teaching of demography had yet to find a clear place within the academy. While basic population statistics had an undeniable utility for those engaged in pragmatic planning activities, there had yet to develop a consensus about the larger significance of modern population changes. In fact, while consensus was developing about the causes of modern population changes, the twentieth century began with growing dissension about their societal consequences. Those engaged in examining population trends warned of very different population problems and recommended conflicting population policies. With such disparate voices speaking about population, a harmonizing core discipline was difficult to detect. The lack of a unifying theoretical perspective that explicated why population mattered, and the presence of a fractured population policy agenda were

two factors that obstructed the development of a distinct discipline of demography; these factors continued to inhibit disciplinary development throughout much of the twentieth century.

During the nineteenth century social movements aimed at influencing population trends arose that attempted to provide the unifying theoretical perspective and clear policy agenda for the study of population statistics. Neo-Malthusianism, based on the belief that growing populations are a major cause of poverty and that lowering fertility by making contraception more accessible will facilitate prosperity, originated in Great Britain early in the nineteenth century and spread gradually throughout Europe. The British Malthusian League was established in 1877, followed by neo-Malthusian leagues in Holland (1885), Germany (1889), and France (1895). But by the time the leagues united at the first International Neo-Malthusian Conference (Paris) in 1900, fertility decline had so spread among urbanizing and industrializing societies that the crisis aura surrounding population growth and large families largely had abated. For countries like France, which were beginning to face the specter of depopulation in the context of an increasingly quarrelsome Europe, it had become less clear that population growth was a social evil or that small families were a social good.

The fertility decline experienced in more developed societies generated new population movements as the twentieth century began. The upper and middle classes were embracing the small family norm more quickly and more fully than other classes. The rediscovery of Mendel's research and the power of Weismann's work on an inviolable germ plasm had convinced many that heredity played a transcendent role in molding human destiny. The educated of the period were much inclined to use social class position as a surrogate measure of biological quality, and they bemoaned the 'race degradation' that resulted from the 'less fit' lower classes contributing an increasing proportion of future generations.

As early as 1891 Sir Francis Galton, addressing the demography section of the International Conference on Hygiene and Demography in London, was urging that the issue of heredity and fertility be dealt with in the 'practical politics' of member states. As the twentieth century began, eugenics movements were forming in a number of countries. In the USA's an active immigration restriction movement allied itself with those expressing eugenic fears. Restrictionists railed against the flood of 'inferior' Eastern and Southern European immigrants who were 'degrading' the high quality of the largely Anglo-Saxon and Northern European 'stock'; they even claimed that the increased number of the 'inferior' was a major cause of fertility decline among the 'superior.'

The eugenists and restrictionists feared changes in population composition and quality, not population quantity and growth. The trend that neo-Malthusians praised, fertility decline, was the same one that they bemoaned. Most eugenists fought to keep access to contraceptives restricted, contending that only the 'more fit' higher classes were sufficiently disciplined to limit their family size. Into this amalgam of population movements Emma Goldman, Margaret Sanger, and Marie Stopes attempted to add 'birth control' movements in the USA and Great Britain. During a period when high-ranking politicians were reminding educated women publicly of their patriotic duty to marry and have children, they began mobilizing citizens to legalize a woman's access to contraception.

With most academic population experts at the time producing works that fed the eugenic-based fears of elites, the task of birth controllers was to develop convincing counter-arguments as to why legalizing contraception was socially beneficial. The socialist arguments that Emma Goldman presented—that working-class women could improve their class' negotiating position with capitalists by restricting the production of new workers—had limited public appeal. Margaret Sanger attempted to attract the allegiance of experts and a wider public by deftly weaving together eugenic and neo-Malthusian themes: restrictive laws do not keep contraception out of the hands of the educated classes and only serve to slow the adoption of birth control among the less motivated 'inferior' classes, an adoption that would benefit both the individual and the society. During the 1920s the pace of fertility decline increased in developed societies, fears of actual depopulation grew, and additional controversy swirled around population trends.

Many consider demography a discipline that is either defined by its specialized techniques or delimited by its particular area of study: the causes and consequences of population change, especially the mortality and fertility transitions that have so dramatically swept the world since the 1750s. But for the first third of this century, for a variety of reasons, those engaged in both areas of research found themselves enmeshed in population controversies that made them less able to focus their energies on disciplinary development. For instance, when Louis Dublin and Alfred Lotka developed 'intrinsic' vital rates in 1925 that controlled for the age structure's influence on crude birth and death rates, they applied their technique to the study of US vital rates and dramatically announced that the average American woman in 1920 was having only half a child more than was needed to maintain a stationary population. Dublin then was moved to write articles identifying excessive birth control as a national problem and attacking Margaret Sanger for fostering a one-sided socially harmful movement. In the late 1920s the technical demographer was hard to distinguish from the movement advocate.

Similarly, experts studying the causes and consequences of modern population movements found the 1930s to be an era in which nationalistic chauvinism made discussion of these trends fraught with contro-

versy. The International Union for the Scientific Investigation of Population Problems was launched in 1928, an outgrowth of the 1927 World Population Conference sponsored by Margaret Sanger in Geneva. The IUSIPP (later to become the IUSSP) planned to hold its first meeting in Rome in 1931. Mussolini's rise to power and hints that the meeting would be used to promulgate his racial theories caused the leadership of the IUSIPP to 'dis-establish' the Rome meeting and convene a hastily planned counter conference in London. The IUSIPP's next official meeting was Berlin in 1935, and was even more controversial. Several national committees, correctly fearing that the conference would be used as a platform from which to spread Nazi racial theories, boycotted the meeting. The first two IUSIPP meetings were exercises in divisiveness, and certainly not venues for establishing a new discipline.

Even the academic study of the causes of modern population dynamics proved controversial during the interwar period. By the early 1930s a number of students of population (Warren Thompson in the United States, Adolphe Landry in France, and A. M. Carr-Saunders in Great Britain) arrived at a quite similar summation of modern population dynamics: as agrarian societies become industrial there is a shift from high to low vital rates during which mortality declines more quickly than fertility, producing a period of population growth. Northern and Western Europe and the United States had largely experienced this 'demographic revolution' (Landry's term), Eastern and Southern Europe and Japan were in the middle of their expansion stage, and much of the rest of the world had just begun the revolution.

Although demographic transition theory was to play a central role in the institutionalization of demography 20 years later, its development in the 1930s proved awkward. Both Germany and Japan were using *Lebensraum* rationales for their imperialist moves into their neighbors' land, claiming that their growing populations needed room to expand. Any summation of modern population dynamics that seemed to confer scientific legitimacy on these claims lost attractiveness to a number of Western policy makers. Explicating the causes of modern population changes during this era did little to establishment a firm foundation on which to build a disciplinary infrastructure for demography.

The end of World War II, however, began a most propitious period for establishing such a foundation. War-torn nations, both allied and axis, entered into a period of state-planned national reconstruction. The production and analysis of population statistics were vital for such planning. The National Institute of Demographic Studies (INED) was established in France in 1947 as a planning arm of the Ministry of Social Affairs and National Solidarity. Although not engaged in the teaching of demography, INED did much to enhance the reputation of the discipline, as did the publication of the first volume of *Population* (Paris) in 1946. One year later *Population Studies* made its debut in London. The establishment of two scholarly journals that focused exclusively on publishing analyses with a distinct demographic focus did much to solidify the boundaries of demography as an academic subject. On a more functional level, the Population Division of the UN Secretariat was established in 1946 and worked to standardize demographic measures and promulgate a common set of techniques for demographic analysis.

The war's conclusion also resolved many of the more fractious issues that had fueled earlier population controversies, and the link between population discourse and divisive controversy weakened considerably. Knowledge of the enormity of the holocaust led to the eugenics movement quickly dissipating. A widespread post-war baby boom dispelled many countries' worries of depopulation. The removal of German and Japanese military threats broke the association between transition accounts of modern population dynamics and population-pressure rationales for imperialism. In fact, in the postwar world the transition framework would provide Western policy makers with a needed policy tool: a way of interpreting the unprecedented demographic changes arising in the world's underdeveloped regions during the 1950s and 1960s.

The rapid decolonization of the world that followed World War II produced a 'Third World' of newly independent governments that became a Cold War battleground where the USA and the Soviet Union fought for supremacy. Unprecedented mortality decline resulting from the use of newly developed antibiotics and the application of effective methods for combating malaria produced rapid population growth throughout much of the developing world. Such growth was especially problematic when considered from the perspective of the post-war versions of demographic transition theory promulgated by US demographers working at Princeton University's Office of Population. If the economic strains associated with rapid population growth prevented the transformation of traditional agrarian societies into modern industrial ones, then rapid population growth might forestall the very socioeconomic changes that would induce fertility decline. Without fertility decline, the Third World's period of population expansion would come to an end with mortality rising as starvation and disease increased. Starvation, economic stagnation, and growing poverty were judged propitious for the spread of communism. Inducing fertility decline in still agrarian societies came to be seen as the only way to humanely resolve the Third World's population dilemma; and, additionally, it would protect the geopolitical interests of the Free World.

This vision of the postwar global population situation was promulgated by Princeton demographers during the decade from 1945 to 1955, especially by

Frank W. Notestein and Kingsley Davis. During the early 1950s John D. Rockefeller III and the leadership of the Ford and Rockefeller foundations accepted the validity of this vision and worked to establish a neo-Malthusian movement with a global focus. Their goal was to set up family planning programs throughout the Third World, lower fertility, and lessen population growth. They recognized that only governments could implement effective family planning programs, and their immediate task became converting government leaders, in both developed and Third World countries, to the position that high fertility was a major social problem in need of state intervention. They determined that a dramatic increase in academic research on international population issues was a necessary first step in this conversion process. During the next two decades they expended millions of dollars to develop demographic research centers that focused on international population issues.

These expenditures on demographic research had a profound impact. In the USA in 1950, for example, courses in demography could be found at the graduate level in only three universities. Between 1951 and 1967 major population research centers were established at 16 US universities; all owed their existence to foundation funding, largely from the Ford Foundation. Similar expenditures helped established internationally oriented population centers at a number of major universities in Europe and Australia. Funding from the Population Council, a foundation established by John D. Rockefeller III in 1952 to provide a leadership role for the international population control movement, established UN regional centers for demographic training and research in Bombay, India (1957), Santiago, Chile (1958), and Cairo, Egypt (1963).

Additionally, its fellowship program brought hundreds of Third World students to major population research centers in developed countries for graduate training in demography. The research generated by the rapidly expanding number of demographers found its way into print in the many demographic and family planning journals that foundation funds were instrumental in launching: *Studies in Family Planning* (1963), *Demography* (1964), *Family Planning Perspectives* (1969), *Population and Development Review* (1975), and *International Family Planning Perspectives* (1975). This conscious cultivation of demographers engaged in addressing Third World population issues had the desired effect. While no government leader thought that lowering the fertility of couples was a legitimate state function in 1950, a substantial number did so 20 years later. As governments adopted fertility control policies and implemented family planning programs, additional funds for demographic research became available, as did additional employment opportunities for the alumni of population research centers.

The vast expansion in resources made available for demographic research by leaders of the international population control movement from 1950 to 1975 was largely responsible for establishing demography as an academic discipline worldwide. All the prerequisites for an intellectual activity becoming a discipline—a significant interaction among those performing the activity, an organization that directs its investigation and teaching, and entities outside itself who value and use its findings—were deliberately assembled by this leadership. The decision to fund graduate level demographic training programs at leading Western research universities assured that there would be a significant and continuing interaction among those engaged in demographic research. It also assured that academic programs in demography would be established and that standards for the teaching of the discipline would be developed. The many private and government contracts for fertility control research generated a demand for this new discipline's findings by outside entities, a demand that no doubt impressed university administrators.

These resource flows, however, did not simply impact the disciplinary development of demography in positive ways. There were costs as well. The neo-Malthusian impetus behind the institutionalization of demography caused the discipline itself to be tinged with an ideological aura. During the 1950s and 1960s Soviet and Chinese population specialists consistently questioned the objectivity of Western analyses of Third World population issues. The major premise of the international population control movement was that rapid population growth had deleterious effects on development, and this was certainly the premise that guided the vast bulk of orthodox demographic research. It was, however, more hypothesis than fact. Although it is not uncommon among the social sciences to have certain perspectives represented within a discipline that are based on premises not subject to empirical verification, the discipline itself is usually composed of a multiplicity of such perspectives. Except in France, where a pronatalist tradition among French demographers made many suspicious [?] of neo-Malthusianism, significant counter-perspectives did not develop within Western demography until a stream of 'revisionist' literature began to develop in the late 1970s. Until that lively debate over the relationship between population growth and economic development developed, demography's close association with the international population control movement worked to politicize it. At occasions such as the World Population Conference at Bucharest in 1974, demographers found not only the validity of their research findings questioned but also their motives.

The neo-Malthusian impetus for much foundation and government funding of demographic research also made it subject to the shifting fortunes of that movement. Population research centers were establish on the basis of grants with time limits and lacked departmental status. Many of their positions were

'soft money' nontenured ones, although all had core faculty with permanent positions in established departments, usually sociology, economics, or public health. During the 1950s and 1960s few universities were willing to subsidize population research centers internally, and the survival of demographic training programs required the continual influx of outside funds. Foundation support for the international population control movement remained strong during the 1950s and 1960s but declined dramatically during the 1970s.

Northern government support of the movement tended to be significant but did vary with elections; the end of the Cold War, however, has dissipated a powerful source of Northern concern over problems associated with Southern population growth. Academic demography has responded to the funding vagaries caused by these movement shifts by becoming leaner, more embedded within the university framework, and less oriented to an international fertility control agenda. In the North, where the discipline still finds its center, every country faces a significant aging of its population and many countries face problems associated with long-term below replacement fertility. These domestic population concerns increasingly are the disciplinary focus of Northern demographers and population research centers. A similar transformation is taking place in many Southern countries where fertility is approaching replacement levels. Only in those Southern countries whose fertility levels remain high does demographic research remain synonymous with fertility reduction research. International migration, not fertility control, is increasingly the population trend garnering the attention of both Northern and Southern demographers. In a world with a global economy and with great disparities in national wealth, the mass movement of individuals across national borders can only increase in intensity.

At the end of the twentieth century the most notable development in the field of population has been that a discipline of demography, much like that envisioned by Achille Guillard in 1855, finally has emerged. The study of the causes and consequences of population dynamics is now a widely recognized intellectual activity that is investigated and taught within an academic setting by individuals with academic credentials who are paid for their specialized knowledge and who reproduce themselves with new generations of demographers. The institutionalization of this discipline in the second half of the twentieth century was largely the result of a confluence of events happening at mid-century, the most notable of which was the rise of an international population control movement.

Demography still lacks an overarching theory capable of linking individual and societal demographic action, but currently possesses a significant disciplinary cohesiveness that arises from having a well defined subject matter and, increasingly, from having a common policy focus. Because most populations have neared completion of their mortality and fertility transitions, a common nexus of population trends is present: population aging in nearly every region; below replacement fertility in much of the North; declining population growth in much of the South. This particular set of demographic trends has so far produced few divisive debates over population policy. The twenty-first century begins, therefore, with demography's immediate disciplinary future assured.

See also: Age Structure; Censuses: Demographic Issues; Databases, Core: Demography and Registers; Demographic Models; Demography, History of; Eugenics as the Basis of Population Policy; Families and Households, Formal Demography of; Family Theory: Competing Perspectives in Social Demography; Fertility Control: Overview; Fertility: Political and Political–Economic Perspectives; Human–Environment Relationship: Carrying Capacity; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources.

Bibliography

Barrett D 1995 Reproducing persons as a global concern: The making of an institution. Ph.D thesis, Stanford University

Caldwell J C 1996 Demography and social science. *Population Studies* **50**(3) (November): 305–33

Caldwell J C, Caldwell P 1986 *Limiting Population Growth and the Ford Foundation Contribution*. Pinter, London

Chasteland J-C, Roussel L (eds.) 1997 *Les Contours de la Demographie: Au Seuil du XXIe Siècle*. Institut National d'Etudes Demographiques, Paris

Davis K 1945 The world demographic transition. *Annals of the American Academy of Political and Social Science* **237**: 1–11

Demeny P 1988 Social science and population policy. *Population and Development Review* **14**(3) (September): 451–79

Donaldson P J 1990 *Nature Against Us: The United States and the World Population Crisis*. University of North Carolina Press, Chapel Hill, NC

Finkle J L, Crane B B 1975 The politics of Bucharest: Population, development, and the new international economic order. *Population and Development Review* **1**(1): 87–114

Greenhalgh S 1996 The social construction of population science: An intellectual, institutional, and political history of twentieth-century demography. *Comparative Studies in Society and History* **38**(1) (January): 26–66

Harkavy O 1995 *Curbing Population Growth: An Insider's Perspective on the Population Movement*. Plenum Press, New York

Hauser P M, Duncan O D (eds.) 1959 *The Study of Population: An Inventory and Appraisal*. University of Chicago Press, Chicago

Hodgson D 1988 Orthodoxy and revisionism in American demography. *Population and Development Review* **14**(4) (December): 541–69

Lorimer F 1971 The role of the international union for the scientific study of population. *Milbank Memorial Fund Quarterly* **49**(4) (October): 86–96

McNicoll G 1992 The agenda of population studies: A commentary and complaint. *Population and Development Review* **18**(3) (September): 399–420

Notestein Fran W 1971 Reminiscences: The role of foundations, of the Population Association of America, Princeton University and the United Nations in fostering American interest in population problems. *Milbank Memorial Fund Quarterly* **49**(4): 67–84

Preston S H 1993 The contours of demography: Estimates and projections. *Demography*. **30**(4) (November): 593–606

Sauvy A 1966 *Théorie Générale de la Population*. Presses Universitaires de France, Paris (*General Theory of Population*. Basic Books, New York, 1969)

Szreter S 1993 The idea of demographic transition and the study of fertility change: A critical intellectual history. *Population and Development Review* **19**(4) (December): 659–701

D. Hodgson

Dental Health: Psychosocial Aspects

Dental health is a topic of both practical and theoretical interest. Practically speaking, the pain and loss of dentition that can be a result of dental ill health is a cause of much distress and disability. Although dental ill health is rarely life threatening, it can have severe and long-lasting effects on quality of life. Theoretically speaking, the ways in which people cope with anxiety and seek to maintain oral health can provide many opportunities to test the accuracy of theories of human distress and motivation. In this article, the ways in which psychologists and sociologists have sought to understand dental anxiety and to improve oral health are examined.

1. *The Term 'Dental Health'*

Dental health can be defined and measured both objectively and subjectively. Objectively, dentists often use the decayed, missing, and filled teeth (DMFT) index as an indicator of dental health, but this may be more appropriately seen as a measure of dental history rather than health. Subjective assessments given by patients may not correspond to professionally developed objective measures. For example, Heyink et al. (1986) found only weak associations between dentists' and patients' appraisals of dentures. For dentists, clinical indicators such as fit, stability, and bite force were important, but for patients, denture quality depended on how well they functioned in practical everyday terms. It is useful to recall the World Health Organization's (1980) distinction between impairment (the objective pathology such as tooth loss), disability (limitation on activities such as not being able to eat certain foods as a result of tooth loss) and handicap (inability to perform social roles, such as being reluctant to accept invitations to dinner because of the potential social embarrassment of not being able to eat the food that might be served). Although it is possible to define dental health in any of these ways, it is useful to consider impairment primarily as an objective measure of health, while disability and handicap are more subjective in nature. Seen in terms of disability and handicap, dental disease is widespread. Cushing et al. (1986) found that 26 percent of adults had experienced dental pain, 20 percent had difficulty in eating and 15 percent had associated problems of communication within the previous year.

Psychologists and sociologists have been keen to understand the causes of dental ill health, partly with a view towards reducing disease and partly with a view towards using the dental setting to test and develop theories of health behavior. Two areas which have received considerable attention are (a) the development and alleviation of dental anxiety and (b) the exploration of why some individuals engage in regular preventive dental care whereas others place less emphasis on this aspect of their physical well being.

2. *Dental Anxiety*

Research in this area has addressed three questions. First, there is the issue of the nature of dental anxiety (see *Anxiety Disorder in Children*). Like other forms of anxiety, it has behavioral, somatic, and cognitive components. Second, there is the issue of etiology of dental anxiety. Despite the development of modern anesthetics, it is widespread: depending on how it is measured, 3–5 percent of the general population can be said to have a debilitating level of dental anxiety. Third, there is the issue of treatment. Several treatment approaches show consistently good results.

2.1 *Components of Dental Anxiety*

Historically, the behavioral aspects of dental anxiety have received the most attention. This is probably due to the consistent finding that anxiety is a powerful barrier to the receipt of professional care. For example, Curson and Coplans (1970) interviewed patients in an emergency clinic and 38 percent reported that they did not make regular visits to a dentist because they were too afraid to do so. Of these, only 12 percent made and kept further appointments for a course of treatment at the clinic. The extent of behavioral difficulties has been measured through both observational and self-report methods. Melamed et al. (1975) used an observational schedule which contained a series of behavioral indices of anxiety in children (crying, refusal to open the

mouth, verbal complaints), rating these activities at 3 min intervals during the appointment. Kent et al. (1996) designed a self-report scale to measure the extent to which dental phobics avoided and were troubled by dental stimuli. Physiological measures of anxiety are less easily taken, but they too show a consistent relationship with self-reported anxiety. Patients are more likely to show increased heart rate at certain points of an appointment, especially at the beginning of treatment. Perhaps the most commonly used self-report measure of anxiety is Corah et al.'s (1978) dental anxiety scale, which asks people to indicate how relaxed or tense they are in the dental setting.

An interest in the cognitive aspects of dental anxiety has developed in the late 1990s. As in other forms of anxiety, anxious patients are more likely to engage in catastrophic thinking. They are more likely to ruminate on the worst possible outcomes of an appointment, of an 'Any moment now the dentist will get the nerve and then it will really hurt' kind. For many anxious patients, it is not a question of *if* they will experience pain, but *when*. De Jongh et al. (1995) have developed the dental cognitions questionnaire, which contain such items as 'They will find something terribly wrong with me' and 'They might drill too deep.'

2.2 The Etiology of Dental Anxiety

The research strongly supports the notion that, especially for children, dental anxiety is a classically conditioned response to dentistry (ter Horst and de Wit 1993). Given that modern anesthetics are thought by dentists to be very effective in alleviating discomfort during dental treatment, and given that the need for invasive treatment has declined for many people owing to the introduction of fluoride in the water supply of many areas, it is surprising to find that levels of dental anxiety have not declined to the expected extent in the latter part of the twentieth century. There seems to be several reasons for this. First, it appears that many more patients experience discomfort during treatment than is often assumed. Depending on how it is measured, up to 77 percent of patients report that they feel some pain during their visits to the dentist. Second, it has been argued that the notion of 'preparedness' is particularly relevant here. Just as humans might be innately prepared to become anxious about small animals such as spiders and rats because they posed a realistic threat in our evolutionary past, so too might sitting in the dental chair, with an adult placing sharp instruments in our mouths be one in which anxiety is quick to develop. It might take only the slightest degree of discomfort for anxiety to occur. Although classical conditioning is generally seen as the prime mechanism for the onset of dental anxiety, other learning modalities may also contribute. There is evidence that dental anxiety can be modeled by hearing about someone else's discomfort (Corkey and Freeman 1994). Finally, the role of biological factors merits attention. Dentally anxious individuals score higher on personality questionnaires measuring emotional lability or neuroticism, and children who have a 'difficult child' type of temperament are more likely to be dentally anxious. Such research suggests that dental anxiety might be more likely to develop in individuals who have other psychological difficulties (Aartman et al. 1997).

One of the most important methodological issues in this area concerns the retrospective nature of almost all studies. Although it is likely that negative experiences contribute to onset, it is also likely that current levels of anxiety affect interpretations of dental experiences. Owing perhaps to hypersensitivity, an anxious patient might interpret sensations felt in the dental chair as painful, whereas a nonanxious patient may interpret such sensations differently. Certainly, highly anxious patients expect to experience more pain than patients with low anxiety, even if their actual experiences do not differ. Memory processes may also play a role. Kent (1985) asked patients after their appointment to indicate the amount of pain they felt. Three months later, they were asked to indicate the level of pain they remembered having experienced. Whereas those with low anxiety gave reports that corresponded closely to their immediate postappointment ratings, those with high anxiety recalled much higher levels of discomfort. In fact, their reports corresponded most closely to their preappointment expectations. Such studies indicate that retrospective studies are open to severe criticism.

2.3 Treating Dental Anxiety

Interventions designed to assist people to deal with their dental anxiety are similar to those designed to help people with other types of anxiety. There is the fundamental importance of graded exposure to the dental setting in combination with the provision of coping strategies (see *Stress Management Programs*). Before discussing these interventions, however, it is helpful to outline some of the research pertaining to the prevention of dental anxiety.

2.3.1 Prevention. Two factors stand out as critical in reducing the likelihood that dental anxiety will develop, particularly in children. The first concerns the alleviation of pain, in that a reduction in the need for invasive treatments is likely to reduce anxiety. One study that has not been performed is a comparison of the levels of dental anxiety between areas that do and do not have fluoridated water supplies. The second factor concerns the nature of the relationship between the patient and the dentist. A number of studies have indicated that warm and nurturant dentists can help patients to overcome

their dental anxiety when it has been acquired but also that they can reduce the probability that anxiety will develop, even after a painful experience (Bernstein et al. 1979).

2.3.2 Interventions. As noted above, most interventions for the alleviation of dental anxiety rely on traditional methods of graded exposure to the feared stimulus. There is much consistency between individuals concerning the most feared aspects of dental care (usually extractions followed by drilling and injections) and fear hierarchies can usually be constructed without difficulty. There is considerable evidence that systematic desensitization, relaxation, hypnosis, and modeling can all have positive effects (Kent and Croucher 1998). As in other areas of psychological intervention, ~60 percent of patients report substantial improvement in their difficulties, regardless of the type of approach used. This may be because all types of intervention serve to enhance patients' sense of control over the situation. In this context, the use of stop signals—where the patient is invited to indicate a wish to take a break from treatment—can be a simple yet effective method for increasing a sense of control over the dentist's behavior. Pharmacological treatments—especially using nitrous oxide—can not only be effective for the alleviation of anxiety in the short term but may also have longer-term beneficial effects. The alleviation of uncertainty and the provision of information are particularly relevant to dental anxiety. Because it is closely associated with the presence of catastrophic thinking, realistic information can not only dispel some irrational fears about the nature of professional care, but can also provide patients with knowledge about when invasive treatment might cause discomfort. An interesting theoretical and practical issue here concerns the amount of information provided. There are indications that patients who have a blunting coping style show greater distress when given detailed information, whereas those with a monitoring coping style are most likely to receive benefit (Auerbach et al. 1976).

3. *Maintenance of Oral Health*

The maintenance of oral health is largely a function of behavior (see *Health Behaviors*). Although fluoridated water supplies have led to a substantial reduction of dental caries in some areas, most dental ill health could be prevented if people took the appropriate actions. These actions are mainly of two types: a reduction in the frequency of eating sugar to reduce caries, and an increase in the use of brushing, dental floss, and regular dental visits to reduce the amount of plaque (which leads to gingivitis and tooth loss through the destruction of supporting tissue).

3.1 *Educational Approach*

The notion that education could improve dental health is based on the findings concerning the generally low levels of relevant knowledge in the general population. Many people are unaware that the occurrence of caries is most closely associated with the frequency of eating sugar, rather than the absolute amount. Similarly, many people do not realize the importance of reducing plaque for the maintenance of healthy gums. Theories of health behavior (Connor and Norman 1996) predict that patients' levels of knowledge will have an effect on what they do to maintain good health. According to these models, people will begin to make changes to their behavior—or at least move from one stage of preventive care to another—only when they appreciate the importance of regular dental care and have a correct interpretation of symptoms.

Horowitz et al. (1976) provide a good example of an experiment on intensive dental education. One group of children was given 10 sessions of 30 min on plaque removal, being taught in small groups by a dental hygienist. They were informed about plaque, how to identify it with disclosing agents, and how to remove it. Plaque removal was practiced under supervision every day for the next 6 months. The comparison group of children were not given any of this extra information or practice. At the baseline, the two groups were similar on both measures. At the 8 month assessment, there was a significant decline in gingivitis scores in the education group compared with the comparison group, but at 12 months there was no difference. There was no long-term effect, and there were never any differences between the two groups in caries levels or amount of plaque. On the basis of such typical results, it seems that educational programs have little or no long-term effects, but it should be noted that most training has been given in school and there has been little attempt to integrate what children have learned in the school environment with what they do at home. Another possible reason for the lack of an effect is that the education approach is only part of the answer to changing behavior. Educational approaches may make some difference by increasing knowledge, but they may not provide sufficient motivation to put this new-found information into practice. In order to accomplish this, a motivational or behavioral approach may also be needed.

3.2 *Motivational Approaches*

In terms of the stages of change model, people will move from the contemplative and preparation stages to the action and maintenance stages only when they can see some kind of clear benefit in doing so (see *Health: Self-regulation*). Furthermore, the benefit needs to be apparent in the short term rather than some years in the future. From the viewpoint of

behavioral analysis, it is crucial to identify actual and potential reinforcers in the person's environment. One of the important difficulties facing those who wish to change patients' behavior is that there are many positive reinforcements maintaining undesired behavior (e.g., the sweetness of foods) and few negative ones which might lead to desired changes (e.g., caries might develop months or years in the future).

This issue is particularly relevant for children who will have difficulty in seeing the relevance of regular brushing in the present to ensure that they will not lose their teeth through gingivitis decades in the future. Thus, behavioral programs to encourage preventive dental behavior have focused on the use of immediate reinforcements to increase the likelihood that people will monitor their sugar uptake, brush daily, and floss regularly. There are several evaluations of such programs, most showing positive results (Kent and Croucher 1998). Monetary reinforcers have been used for both children and their parents, but often stars on a calendar can be similarly effective.

See also: Anxiety Disorder in Children; Behavioral Medicine; Health Behavior: Psychosocial Theories; Health Behaviors; Health Education and Health Promotion

Bibliography

Aartman I W, De Jongh A, van der Muelen M J 1997 Psychological characteristics of patients applying for treatment in a dental fear clinic. *European Journal of Oral Sciences* **105**: 384–8

Auerbach S M, Kendall P C, Cuttler F et al. 1976 Anxiety, locus of control, type of preparatory information and adjustment to dental surgery. *Journal of Consulting and Clinical Psychology* **44**: 809–18

Bernstein D A, Kleinknecht R R, Alexander L D 1979 Antecedents of dental fear. *Journal of Public Health Dentistry* **39**: 113–24

Connor M, Norman P (eds.) 1996 *Predicting Health Behavior*. Open University Press, Milton Keynes, UK

Corah N L, Gale E N, Illig S J 1978 Assessment of a dental anxiety scale. *Journal of the American Dental Association* **97**: 816–9

Corkey B, Freeman R 1994 Predictors of dental anxiety in six-year-old children: findings from a pilot study. *Journal of Dentistry for Children* **61**: 267–71

Curson I, Coplans M-P 1970 The need for sedation in conservative dentistry. *British Dental Journal* **128**: 19–22

Cushing A M, Sheiham A, Maizels J 1986 Developing socio-dental indicators—the social impact of dental disease. *Community Dental Health* **3**: 3–17

De Jongh A, Muris P, Schoenmakers N, ter Horst G 1995 Negative cognitions of dental phobics: reliability and validity of the dental cognitions questionnaire. *Behaviour Research and Therapy* **33**: 507–15

Heyink J, Heezen J, Schaub R 1986 Dentist and patient appraisals of complete dentures in a Dutch elderly population. *Community Dentistry and Oral Epidemiology* **14**: 323–6

Horowitz A M, Suomi J D, Peterson J K et al. 1976 Effects of supervised daily dental plaque removal by children. *Journal of Public Health Dentistry* **36**: 193–200

Kent G 1985 Memory of dental pain. *Pain* **21**: 187–94

Kent G, Croucher R 1998 *Achieving Oral Health*, 3rd edn. Wright, Oxford, UK

Kent G, Rubin G, Getz T, Humphris G 1996 The development of a scale to measure the social and psychological effects of severe dental anxiety. *Community Dentistry and Oral Epidemiology* **24**: 394–7

Melamed B G, Weinstein D, Katin-Borland M, Hawes R 1975 Reduction of fear-related dental management problems with use of filmed modeling. *Journal of the American Dental Association* **90**: 822–6

ter Horst G, de Wit C 1993 Review of behavioral research in dentistry 1987–1992: dental anxiety, dentist–patient relationship, compliance and dental attendance. *International Dental Journal* **43**: 265–78

World Health Organization 1980 *Health Promotion. A Discussion Document on the Concept and Principles*. World Health Organization, Copenhagen, Denmark

G. Kent

Dependency Theory

Dependency theory is a school of thought in contemporary social science which seeks to contribute to an understanding of underdevelopment, an analysis of its causes, and to a lesser extent, paths toward overcoming it. It arose in Latin America in the 1960s, became influential in academic circles and at regional organizations, spread rapidly to North America, Europe, and Africa, and continues to be relevant to contemporary debate. This article examines the history of its chief concept, describes its evolution over time, analyzes its influence, and evaluates its validity and prospects for social science in the future.

1. The Dependency Concept

The facts to which dependency refers have existed for a long time, not only in connection with the emergence, consolidation, and evolution of the present-day historical-social system of capitalism, but also with former world systems, all of them asymmetrical: comprised of units located at their core, their semiperipheries, and their peripheries. They are linked by domination–subordination relationships, which are structural because they stem from and are driven by the principles which govern the system as a whole, and hence, the location of the different units.

For the current system, those principles are anchored in the accumulation of capital and in its varying forms, generating modes of international division of labor that comprise and shape the relations among the center, semiperiphery, and periphery. The classical political economy of Adam Smith, David Ricardo, and others in the eighteenth and nineteenth centuries recognized the worldwide nature of capitalism, though

they saw the international division of labor among the system's units (countries) as a source of the wealth of nations, operating through Ricardo's theory of comparative advantages. Marx's critique of classical political economy viewed that division as a source of inequalities and asymmetries among peoples, and therefore posed the need to construct a true social model of the whole system in the long term (Wallerstein 1974, pp. 388 ff.). Though his imitators 'froze' Marx 's model and stripped it of its historical nature, his successors continued to focus on the evolution of inequalities and asymmetries. This was especially the case of the theorists of imperialism (Lenin, Bukharin, Hilferding, Luxemburg) and of unequal and combined development (Trotsky and his followers).

Subsequent to the interpretations based on Marxism and imperialism theory, efforts were made in the political economy to understand the nature of 'backwardness' and the causes of poverty in many societies; this line is represented by Arthur Lewis and Gunnar Myrdal, and later by Albert Hirschman and Amartya Sen, to name only a few significant authors (Rist 1997, pp. 69).

Oddly, the dependency concept did not acquire a central theoretical status in either of those two areas of analysis, but was accepted in its most rudimentary form: Certain economies depend on others to buy or sell raw materials and manufacture. It was only with the appearance of studies on the subject of development in Latin America and the Caribbean that 'dependency' became a social-scientific term, though a very controversial one.

2. *External Dependency*

The term 'dependency' appears for the first time in the 'developmentalism approach,' so called because its key concept (and chief concern) was development, pursued in order to overcome 'underdevelopment' (a term coined at that time), which emerged from the UN Economic Commission for Latin America and the Caribbean (ECLA). Following the proclamation and adoption of the United Nations' Charter, in addition to the Security Council, responsible for finding peaceful ways to resolve conflicts among nations, the Economic and Social Council (ECOSOC) was created to coordinate activities in the UN's other main field of action—promoting the development of peoples. To help the ECOSOC fulfill its mission, economic commissions were designed for the five continents. Despite the initial resistance to its institutionalization by the USA, ECLA was founded in 1948 and received that country's blessing in Point Four of President Truman's January 1949 State of the Union address, referring to the need to overcome underdevelopment (Rist 1997, p. 70). It became the most intellectually creative of all the commissions, and the most active one in designing and applying development policies (Sonntag 1994, p. 228).

Developmentalism's diagnosis begins with a critique of the classical (Ricardian) theory of comparative advantages: The economic relations between the core and the periphery tend to reproduce the conditions of underdevelopment and increase the distance between the developed and underdeveloped countries. Paraphrasing Cardoso, the invisible hand of the market appeared, not as a mother but as a stepmother, intensifying inequalities rather than correcting them. The reason for that outcome is that productivity grows faster in the central countries' manufacturing industries than in the peripheral countries' production, and that growth cannot spread to primary (agricultural or mineral) production through equalization of the prices of both sets of products, because of union pressures and the presence of oligopolies in the developed countries. The end result is the trend toward deteriorating terms of trade and a growing concentration of technical progress in the manufacturing or core countries (Sonntag 1988, pp. 24–5).

Therein lies the key: with this center–periphery axiom ECLA posits a discontinuous worldwide development of capitalism. It identifies the (structural) place of the economies in which capitalist techniques of production are first adopted as the core, and that of the economies which are further behind in technological and organizational terms as the periphery. This dichotomy is constantly reproduced by the dynamic of the international division of labor, to the point where the periphery tends to transfer part of the fruits of its technical progress to the core while the latter retains those of its own progress (CEPAL/ECLA 1969, 160 ff.).

This latter assertion poses the concept of external dependency. Dependency lies in the peripheral economies' reliance on the determinant dynamic of the world system of which they are part: that of the core countries, simultaneously reinforced by the transfer of part of the peripheral countries' surplus to them. An odd feature of this school of social science thought is that, in spite of ECLA's clear analysis of external dependency, in its strategic proposals to achieve development it fell into a voluntaristic confidence in the utility of foreign investment, the industrialized countries' cooperation, and a transformation of international trade.

3. *Structural Dependency*

A debate began at ECLA during the first half of the 1960s on the results of the development policies grounded in developmentalism and intensively applied in most Latin American countries. That also implied a theoretical reorientation in two distinct directions: a renewal of developmentalism itself, and the emergence of dependentism (Sonntag 1988, p. 31 1994, p. 230).

This new approach not only drew on the debates over developmentalism, but also reflected: (a) the rediscovery of Latin American authors from earlier

decades with critical ideas (Sergio Bagú, Caio Prado Jr., José Carlos Mariátegui, etc.); (b) a renewal of Marxism during the 1950s and 1960s, through Baran's *Political Economy of Growth* and Baran and Sweezy's *Monopoly Capital*; (c) studies by authors such as A. Gunder Frank on the capitalist nature of underdevelopment; and (d) Latin American interpretations inspired by a less orthodox Marxism than that of the previous period, advanced by authors in different countries. There is considerable controversy about the significance of these influences: while Cardoso (1976, p. 4) estimates them less, Dos Santos (1998, pp. 103) and Rist (1997) attribute far greater importance to them.

The first texts of what was rapidly to become a new paradigm of thought on underdevelopment in Latin America (and also in Africa, where it contributed to Samir Amin's school of worldwide accumulation of capital, in the developed countries, and to a lesser extent in Asia) initially circulated as mimeographed discussion papers among the members and associates of the Latin American Institute of Economic and Social Planning (ILPES), since 1961 ECLA's think-tank and cadre's school. Two of these papers, those of Cardoso and Faletto (1976a) and Quijano (1977), were the sources of dependentism. Those that followed drew on these initial efforts, some opposing their theoretical proposals and political conclusions, others assimilating them and incorporating them into their own studies.

The first dependentists started with a critique of the modernization theory introduced by ECLA in the late 1950s, and posited an integrated analysis of development combining the study of social change processes in the economic sphere with that of transformations of structures of classes, social sectors, and groups, plus that of the changes occurring within the system of domination. The accent is on sociological interpretation in the strict sense, that is, on the transformations of the social structure and system of domination, since they are the epicenters of the dynamic of development. The latter is the outcome—always subject to change—of the attempts by the different social actors, alone or in alliances, to achieve their goals through their collective actions (Sonntag 1988, p. 63).

The dependentists hold fast to the center–periphery axiom, but redefined the concept of dependency, positing that it 'directly alludes to the conditions and the existence and operation of the economic system and the political system, showing the linkages between the two in both the internal dimension of countries and the external one.' The aim is 'to seek to demonstrate that…the mode of national economies' integration into the international market implies defined and differing forms of interaction by the social groups of each country, among themselves and with external groups' (Cardoso and Faletto 1976a, pp. 24, 28). Accordingly, the relations of political struggles among groups and classes, on the one hand, and the history of the economic-political structures of internal and external domination, on the other, are the key questions. So, they are interested in inquiring into the meaning of the basic structural relations and their phases of development in their dual determination: within the local systems of domination, and their relations with the international order (Cardoso and Faletto 1976b, p. 1). Thus dependency becomes a framework which influences the fundamental tendencies of development in the subordinated societies, by leading to and producing change, impregnating and modifying all aspects of the social structure, social psychology, and cultural institutions, including those of sectors which had been relatively untouched by sociocultural influences prior to the beginning of dependency (Quijano 1977, p. 23).

The dependency concept thus acquired a dialectical content. The relations implied are not perceptible in terms of 'structures,' but rather in terms of collective actors and their practices stemming from their interests and motivations. Change processes are not automatic mechanisms but the outcomes of those practices. That is, dependentism—in its first formulations—tries to restore the subjective actor(s) to history and the dynamic dimension(s) to social-scientific interpretations. It is, once again, a 'global analysis' (Cardoso and Faletto 1976b, p. 4) which stresses the multiple mediations among actors and structures.

A division soon appeared within the field of dependentism. One group, led by Dos Santos (1970) and Marini (1969), claimed that it had achieved the status of a theory of dependency, while another, led by Cardoso and Faletto, insisted on dependentism as a methodological approach. While the former were known as neo-Marxists (Dos Santos 1998, p. 105), the latter asserted that the dependentist approach only implied 'an empirical analytical effort to reconstruct a "concrete whole."' For Dos Santos, Marini, and their followers dependency as a 'theory' implied that a formal and testable body of propositions was inherent to the concept, while Cardoso and his group feared that thus the approach would be formalized (Cardoso 1976, p. 14).

The debate between the two groups was fiery, long, and sometimes bitter. It addressed metatheoretical issues, but focused chiefly on political ones: whether capitalist development was viable or not under conditions of dependency. The 'theoreticians' asserted the impossibility of any such development, while Cardoso, Faletto, and others would not agree with that idea and its assumption that, to improve the quality of analysis, dependency theory had to be formalized (Cardoso 1976, p. 24). They argued thus against the mechanical-formal style of those who believed in the ultimate ends of history, guaranteed by the necessary structural inability of dependent capitalism to expand and reproduce itself.

Dependentism, in either of its two versions, oriented social scientific analysis from the 1960s through part of the 1980s and fostered—as the proponents of both schools agree—a far deeper, more complete and more dynamic understanding of the social reality of underdevelopment.

Special mention should go to what Cardoso (1976, p. 8) has called 'the consumption of dependency theory in the United States.' There interest centered on first, the historical debate over the supposed Latin American feudalism; second, the relations among the dependent countries' social, economic, and political structures and the international capitalist system; third, research on and denunciation of the forms of 'foreign aid,' the CIA, and the Machiavellian actions of the multinational corporations, and so forth. This concern was politically legitimate and drew attention to real aspects of the contemporary process. But little by little it ended up restoring the priority of the external over the internal (Cardoso 1976, p. 13).

Criticisms of dependentism is as old as the approach theory itself. Especially scholars of industrial societies emphazised (a) the absence of an analysis of political power in the relations between core, periphery, and semiperiphery (political scientists), and (b) dependentism's blame of all problems of underdevelopment to external factors (economists). Both criticisms are dubious. On the one hand, the question of power was in the very center of dependentism's emphasis on the structures of internal–external domination. On the other, the insight of the dialectics between external and internal factors was essential, at least to dependentism as approach.

4. *The Future of Dependentism*

Studies inspired by the dependency theory or approach are part of the constantly renewed effort to restore a tradition based on an analysis of the structures of wealth and domination, which does not undermine the historical process as the movement derived from the permanent struggle among groups and classes. It does not accept the existence of a given course of history, but conceives it as an open process. Hence, if structures limit the range of oscillation, the imagination and actions of human beings reinvigorate and transform them, even replace them by others that are not predetermined. However, these studies had a special characteristic: Instead of being limited to the abstract historical dimension and the analysis of narrowly defined problems, they used the 'nonvulgar' historical-structural method to analyze concrete situations and sought—by returning to the subject of development—to pose questions of importance for national policy and the relations among the central capitalist economies and the dependent and nonindustrialized periphery (Cardoso 1976, p. 7). In this sense, contemporary criticism to dependentism as limited to the nation-state as unit of analysis (Grosfoguel 1996, p. 146) is scarcely justified.

The dependency approach and even theory is today a part of historical sociological analysis as it is understood by Wallerstein (1974). But within it an important function may be performed: since the latter tend to give priority to the systemic interpretation of the world-system, dependentism could focus on studying the behavior of concrete peripheral (and semiperipheral) societies in the context of that system, precisely because it stresses the collective actors, their interests, their practices, and the outcomes within them (as well as their repercussions on the system as a whole). So understood, dependentism would not disappear into the broader context of historical sociology, but would combine with it to form a complementarity that would permit more incisive approaches to each society's particular characteristics within the framework of its place in the world-system, in this era of globalization (Dos Santos 1998, Sonntag 1988, 1994).

To achieve that goal, 'it is necessary to have a sense, not so much of proportion as of being ridiculous, and avoid the simplistic reductionism so common among the modern butterfly collectors who abound in the social sciences': they classify in history types of dependency, modes of production, and laws of development, in the misconception that their discoveries may free history of its ambiguities, speculations, and surprises (Cardoso 1976, p. 25). It is, on the contrary, necessary to have the patience for disciplined research, to arrive at a dialectic which is neither lazy nor limited to the construction of general and abstract formulas as if they were synthetic.

In this assertion lies perhaps the challenge not only to this school of thought but also the entire sociology/social science of underdevelopment, particularly in these times of 'globalization,' uncertainty, and complexity.

See also: Capitalism: Global; Center–Periphery Relationships; Colonization and Colonialism, History of; Deprivation: Relative; Development and the State; Development, Economics of; Development: Social; Finance, Geography of; Financial Institutions in Economic Development; Globalization, Anthropology of; Globalization: Political Aspects; Imperialism, History of; Imperialism: Political Aspects; International Business; Modernization, Sociological Theories of; Multinational Corporations; World Systems Theory

Bibliography

Cardoso F H 1976 El consumo de la teoría de la dependencia en EE.UU. (The consumption of dependency theory in the US). In: Various authors *Crisis económica y nacionalismo en Latinoamérica* (*Economic crisis and nationalism in Latin America*). CENDES, Caracas, Venezuela

Cardoso F H, Faletto E 1976a *Dependency and Development in Latin America*. University of California Press, Berkeley, CA
Cardoso F H, Faletto E 1976b Preface. In: Cardoso F H, Faletto E (eds.) *Dependency and Development in Latin America* (2nd ed.). University of California Press, Berkeley, CA
CEPAL/ECLA 1969 *El pensamiento de la CEPAL* (*The thinking of ECLA*). Editorial Universitaria, Santiago de Chile
Dos Santos T 1970 La crisis de la teoría del desarrollo y las relaciones de dependencia en América Latina (The crisis of development theory and dependency relations in Latin America). *Cuadernos de Estudios Socio-Económicos*, **11**: 18–?49
Dos Santos T 1998 La teoría de la dependencia: un balance histórico y teórico (Dependency theory: A historical and theoretical balance). In: López Segrera F (ed.) *Los retos de la globalización: ensayos en homenaje a Theotonio Dos Santos* (*The Challenges of Globalization: Essays in Honor of Theotonio Dos Santos*) (Vol. I, pp.93–151). UNESCO, Caracas, Venezuela
Grosfoguel R 1996 From cepalism to neoliberalism: A world-systems approach to conceptual shifts in Latin America. *Review* **29**: 131–54
Marini R M 1969 *Subdesarrollo y revolución* (*Underdevelopment and Revolution*). Siglo XXI Editores, Mexico
Quijano A 1977 *Dependencia, urbanización y cambio social en Latinoamérica* (*Dependency, Urbanization and Social Change in Latin America*). Mosca Azul Editores
Rist G 1997 *The History of Development from Western Origins to Global Faith*. Zed Books, NJ
Sonntag H R 1988 *Duda. Certeza. Crisis. La evolución de las ciencias sociales de América Latina* (*Doubt. Certainty. Crisis. The Evolution of the Social Sciences of Latin America*). UNESCO, Nueva Sociedad, Caracas, Venezuela
Sonntag H R 1994 The fortunes of development. *International Social Science Journal* **46**: 227–45
Wallerstein I 1974 The rise and future demise of the world capitalist system: Concepts for comparative analysis. *Comparative Studies in Society and History* 16, **4**: 387–415

H. R. Sonntag

Depression

This chapter describes how over the past ten years a solid neurobiological database has emerged that has left no doubt that depression is a complex clinical condition where interaction between genetic endowment and exogenous factors determines the phenotype. The various hypotheses for central nervous system dysfunction that may precipitate depression are examined and the ways in which studies related to stress hormone regulation have led to hypothesis-driven drug discovery strategies are discussed. In the future, the descriptive behavioral aspects of psychiatry will be replaced by more medically oriented schemes where abnormal behavior and mood are viewed as the functional readout of impaired gene environment interaction. Thus, traditional diagnostic schemes, presently still entirely based on behavioral and demographic data, will be modified by the new and exciting database that is emerging from functional genomics. Notably, most types of behavior depend on the interplay between environmental factors and multiple genes.

Based upon the complete genome sequences of humans, the clinical management of depression will become more individualized on the basis of functional phenotyping with laboratory tests and genotyping, once appropriate maps for polymorphisms have been established. The future knowledge base that will include not only behavioral, but also functional and genetic information about the patient will be so powerful that we will wonder how we ever practiced clinically without it.

1. *The Socioeconomic Burden of Depression*

Depression is one of the most common medical conditions worldwide. A national survey conducted in the United States of America has shown that the average 12-month prevalence in the general population is 10.3% with a 1-month prevalence ranging between 1.5% and 2.5%, amounting to a lifetime prevalence of 17% (Kessler et al. 1994). New research has also led to the abandonment of the view that depression is a byproduct of stressful life styles in industrialized countries. In fact, studies conducted under the auspices of the World Health Organization (WHO) indicate that the situation may be even worse in poor countries where malnutrition and infection make the brain more susceptible to mental disorders. Notably, five out of the ten leading causes of disability worldwide are psychiatric conditions. Using a sophisticated method for assessing the duration and severity of a disability, disability-adjusted life years (DALY), it has been calculated that depression will become a leading cause of disability in the coming decades, second only to ischemic heart disease (see Table 1).

It is also abundantly clear that one consequence of depression is a fourfold risk of developing ischemic heart disease and further, that patients who develop depression following a myocardial infarction, have a 15% higher mortality rate within the first six months following the attack than those without depression.

One of the most dramatic sequelae of depression is suicide. Official suicide rates drastically underestimate the actual level of suicide by varying extents. According to conservative estimates suicide claims over a million deaths per year, and among the younger population suicide is a leading cause of death second only to accidents. Given these facts and focusing on prevalence rates, it needs to be emphasized that these rates represent only diagnosed cases, and therefore only a fraction of the total prevalence because de-

Table 1
Changes in the rank order of disease burden for 10 leading causes measured in disability-adjusted life years (DALYs[a])

1990	2020 (Baseline scenario)
1. Lower respiratory infections	1. Ischemic heart disease
2. Diarrheal disease	2. Unipolar major depression
3. Conditions arising during perinatal period	3. Road traffic accidents
4. Unipolar depression	4. Cerebrovascular disease
5. Ischemic heart disease	5. Chronic obstructive pulmonary disease
6. Cerebrovascular disease	6. Lower respiratory infection
7. Tuberculosis	7. Tuberculosis
8. Measles	8. War-associated casualties
9. Road traffic accidents	9. Diarrheal disease
10.	10. Human immunodeficiency virus

a One DALY is one lost year of healthy life

pression is frequently overlooked but precedes suicide in the majority of cases. It is estimated that only approximately 50% of the population with depression in Europe and the United States is correctly diagnosed, while in Japan the rate is much lower, not exceeding 20%.

As a consequence, depression is still widely undertreated and remains one of medicine's most costly illnesses. When combining direct costs such as inpatient and outpatient care and pharmaceutical or talk therapies with the indirect costs of lost productivity, missed work days and loss of lifetime earnings, it becomes clear that depression accounts for a major share of the overall health care expenditure. These costs were calculated to amount to US$ 43 billion in 1990 in the United States alone, and a realistic reflection of the economic size of the burden imposed by depression worldwide underscores the pressing need to increase efforts on all fronts, ranging from public awareness to pharmaceutical research. In the light of these figures and the fact that we possess the therapeutic tools to treat depression successfully it remains one of the major tragedies that depression is still underdiagnosed and undertreated.

2. *The Genetics of Depression*

After more than a century of family, twin, and adoption studies we can now state with certainty that almost all forms of behavior, including depressive psychopathology, are influenced by multiple genes. Such genes are called quantitative trait loci (QTLs) because they are likely to result in quantitative (continous) distributors of phenotypes that underlie a predisposition to depression. Specifically, this means that there is no single locus that by itself causes psychiatric syndromes: instead multiple alleles (gene variants) that interact to produce vulnerability are found at multiple loci within the genome. Genetic research in psychiatry has left little room for the supposition that a single gene at a given locus determines any given psychiatric syndrome. In other words, in different families, different gene variants are required to confer increased vulnerability. In the light of the manifold possibilities of gene–gene interaction, genetic research has provided a particularly interesting demonstration that genes do not act alone.

Genes must interact with nongenetic environmental factors to convert vulnerability into illness. This is an important aspect as much research in the past has been devoted to the understanding of the development of depression based entirely on an individual's biography, making depression one of the last hermeneutic resorts in medicine. Many studies have shown that there is little reason to doubt that major stressful experiences, in particular those that occur in early life or even during pregnancy, render an individual more liable to develop depression (Heim et al. 2000). But it needs to be recognized that the conversion of life events into severe psychopathology requires the presence of the appropriate genetic endowment. An additional level of complexity has been introduced by the studies of Kenneth Kendler, who showed that individuals with a high genetic risk of depression are more likely to expose themselves to depression-triggering life events (Kendler et al. 1997). Indeed, several studies have supported this surprising proposition, that genetic factors may indirectly influence the liability to major depression by predisposing an individual to an aversive environment.

This shows that genetic and nongenetic environmental factors are much too inextricably intertwined to be dissected with high precision. Only when functional genomics have ultimately allowed us to establish a broad enough database on the functions of individual genes and gene–gene interactions, will the influence of environmental nongenetic factors on

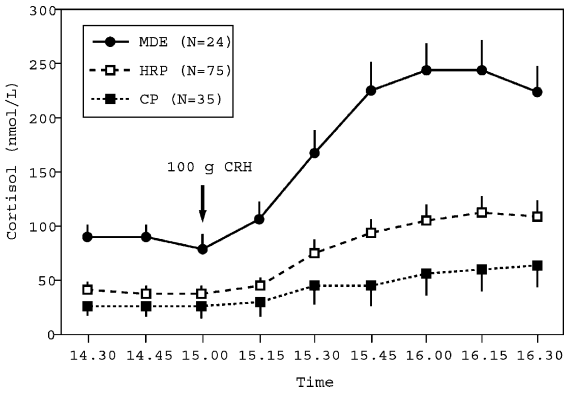

Figure 1
The combined dexamethasone/CRH test 1.5 mg dexamethasone orally at 23.00 hrs and 100 μg CRH at 15.00 hrs the following day was administered to patients with a major depressive episode (MDE) to probands who had never been ill but belonged to families with a high genetic load for depression (HRP) and to matched healthy controls (CP). The plasma cortisol secretion in probands genetically at risk for depression show cortisol increases that lie between patients and controls.

vulnerability to, and the onset and course of, depressive syndromes become available to analysis.

Definition of the depressive phenotype for genetic studies is extremely difficult and findings from these tudies have been largely inconclusive. Another reason for the paucity of genetic studies of unipolar major depression is the psychosocial adversity toward studies that call the stress diathesis into question. The use of association studies, which allow detection of small effects, have for methodological reasons (the need for the whole genome scan) been limited to candidate genes. The most intriguing of these studies is that of Klaus Peter Lesch, who investigated—together with colleagues from the National Institute of Mental Health, Bethesda, MD—the serotonin transporter gene, located on chromosome 17 (Lesch et al. 1996). This serotonin transporter is of particular interest as the majority of antidepressants act on this protein located in the cell membrane of presynaptic nerve terminals. By blocking the reuptake of serotonin molecules these drugs, of which Prozac® is the best known, enhance serotonergic neurotransmission.

Lesch submits that a polymorphism in the transcriptional control region of the serotonin transporter may result in a long and a short version of the promoter polymorphism. The occurrence of the short allele has been shown to be associated not only with decreased transcriptional efficiency resulting in reduced serotonin transporter expression, but also with those personality traits that are likely to be relevant to vulnerability to depression. Most studies aimed at replicating this intriguing finding have failed to do so, but this does not invalidate the experimental approach. Rather, it suggests that the means we currently have to phenotype individuals are perhaps not suited to genetic studies. Researchers have questioned the usefulness of diagnostic algorithms such as the Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association (DSM-IV) or the International Classification of Disease of the World Health Organization (ICD-10) for studies of causality because the psychopathological symptoms that constitute psychiatric diagnoses may be the very remote effects of genes. Therefore, it has been suggested that it would be more fruitful if genetic studies focused on more direct measures of brain function.

Such a strategy is exemplified in the Munich Vulnerability Study conducted at the Max Planck Institute of Psychiatry, where individuals with no history of illness but, being members of families with a high genetic load for depression, are investigated not only according to psychopathology but also according to laboratory measurements and other symptoms that can be objectified. Among these laboratory assessments are neuroendocrine function tests and sleep electroencephalographic-activity recordings. As illustrated in Fig. 1, the hormonal response to a neuroendocrine function test specifically designed to identify impairments in stress hormone regulation is altered in probands with a high familial risk for depression. Their stress hormone release is significantly increased compared with matched controls. What is not yet known is to what extent premorbid stress hormone dysregulation may predict an increased likelihood of developing overt clinical depression.

The idea behind the strategy is to identify patients with impaired stress hormone regulation across diagnostic boundaries and to study their genetic commonalities. Of course, other symptoms or symptom clusters, e.g., psychotic symptoms, need to be considered regardless of the diagnostic category to which the individual patient may belong. The proposed breakdown of the diagnostic description of complex behavior into simpler components including specific symptoms, together with results from neurobiological investigations (e.g., functional brain imaging, electroencephalography, drug treatment response, neuroendocrine assessment, etc.) deserve more attention in future genetic research than do traditional diagnostic attributions.

Bipolar depression, where depressive episodes alternate with manic episodes characterized by elevated mood, ideas of grandiosity, flight of thought, sleeplessness, etc., is much more easily phenotyped. Attempts to understand this disorder in terms of biographic events have also not been very successful, thus facilitating a push towards genetic research. Morbid risk in first-degree relatives is four to six times higher than the population prevalence of 1% for both men and women, making the case for genetic susceptibility

particularly obvious. Enormous efforts in the United States and Europe have allowed the study of families large enough to discover a linkage between bipolar disorder and certain areas on several chromosomes, notably 12q, 18p, 18q, 21q, 4p, and X1.

Of particular interest is a study by Nicholas Barden from Quebec, Canada, who completed a genome screen in a large French Canadian pedigree with over 100 individuals (Morissette et al. 1999). This investigator showed a linkage to chromosome 12q 23–24 in some, but not all branches of the pedigree and concluded from his data that more than one locus may be segregating in this isolated population. A smaller pedigree from the same geographic region also confirmed the linkage between bipolar depression and chromosome 12. Similar conclusions were drawn from another investigation into a family where major affective disorder was segregating with Darier's disease. The gene, whose mutation for Darier's disease encodes a calcium ion pump, was recently identified but screening for mutations in a sample of bipolar probands without Darier's disease did not identify any gene mutation, rendering it unlikely that the susceptibility gene for bipolar disorder is in close proximity to the Darier's disease gene.

This example illustrates that, despite promising results and a fairly clear phenotype (although it may well be that a bipolar case is diagnosed very late because a manic episode occasionally only becomes clinically apparent in later life), we are still far away from the identification of susceptibility genes for bipolar disorder. One conclusion to be drawn is that there are likely to be numerous susceptibility genes that influence the risk of unipolar or bipolar depression. Supported by progress in biotechnology, critical regions for each susceptibility gene will be narrowed down and candidate genes will be tested much faster than in the past. These opportunities, together with new techniques allowing the dissection of gene–gene and gene–environment interactions, will ultimately allow the identification of susceptibility genes. The future role of the clinician scientist will be to improve diagnosis-independent phenotyping and to conduct studies focused on the elucidation of potential pathophysiological roles of candidate genes (see *Mental Illness, Genetics of*).

3. *Hypotheses on the Causality of Depression at the End of the Twentieth Century*

3.1 *Hormones and Neuropeptides*

Depression is associated with a number of changes in hormonal secretion and conversely, certain endocrine disorders (i.e., hypothyroidism or Cushing's disease) are associated with excessive psychiatric morbidity, most frequently depression. Because the peripheral hormonal readout is governed by hypothalamic-releasing factors, which are themselves subject to complex regulations involving biogenic amines (e.g., serotonin, norepinephrine) and amino acids (e.g., gammaaminobutyric acid, GABA, and N-methyl-D-aspartat, NMDA) many researchers in the 1970s and 1980s had hoped that the neuroendocrine system might provide a 'window to the brain.' For example, growth hormone secretion from the pituitary was found to be blunted in patients with depression, following a variety of pharmacological interventions including the α-adrenoceptor system. Thus, the decreased growth hormone release following clonidine, an α-adrenoceptor agonist, was suggested as an indicator of desensitized adrenoceptor systems in these patients. Whereas the neuroendocrine approach that attempted to identify central neurotransmitter disturbances through measurements of hormonal responses to various pharmacological probes proved to be only of limited value in explaining the pathophysiology of depression, a new hypothesis emerged, linking stress hormone regulation to the development of depression and its course during drug treatment.

The fundament of this hypothesis is that elevated central neuropeptides accounting for increased peripheral stress hormone (corticotropin and cortisol) secretion in patients with depression, also mediate a number of psychopathological cardinal symptoms of depression. The neuropeptide in question is corticotropin-releasing hormone, CRH, that, when secreted from the hypothalamus, reaches the pituitary through portal vessels and elicits corticotropin from the pituitary and consequently cortisol from the adrenal cortex. CRH is also present in other brain areas, the frontal cortex, the amygdala, the hippocampus, and the locus coeruleus, all of which are implicated in the generation of stress-related psychopathology such as anxiety, cognitive function, sleep disturbance, appetite and sexual arousal, autonomous regulation and other features.

This has led to research activities from which clear evidence has emerged that CRH not only accounts for peripheral stress hormone secretion, but is also a key mediator of anxiety and depressive symptoms (Holsboer 1999). The question why CRH is elevated in some brain areas and not others is not yet clear, but some evidence has accumulated that a disturbed corticosteroid receptor function exists that prevents elevated cortisol levels from appropriately regulating CRH synthesis. Notably, the regulation of CRH gene expression by steroid activated corticosteroid receptors is very complex and varies across brain areas. The corticosteroid receptor hypothesis represents a paradigmatic shift inasmuch as that a therapeutic strategy has been developed from a clinical observation (Holsboer 2000). Blocking the action of CRH requires knowledge of the receptors through which this neuropeptide exerts its psychopathological effect. Molecular

Table 2
Altered norepinephrine function in depression

Norepinephrine
Increased β-adrenoceptor density in membranes of cortical neurons of depressed suicide victims
Reduced growth hormone response to α-adrenoceptor stimulation (e.g., with clonidine)
Increased stress hormone secretion after α-adrenoceptor antagonism (e.g., with yohimbine)
Reduced levels of norepinephrine transporters in the locus coeruleus
Elevated levels of tyrosine hydroxylase in the locus coeruleus

pharmacology and genetic research have helped us to identify the CRH receptor subtype (CRH_1 receptor) responsible, and the discovery of small molecules that selectively antagonize CRH actions at this receptor have allowed us to test this CRH hypothesis.

Basic studies in animals and first clinical studies support the notion that CRH_1 receptor antagonists are capable of improving signs and symptoms of stress-related clinical conditions, and that they are of potential value in treating major depression. This concept is frequently burdened by two serious misunderstandings that need to be addressed: the clinical efficacy of CRH receptor antagonists is not limited to patients showing peripheral hypercortisolemia. This endocrine symptom reflects hypothalamic–pituitary activity, a neuroanatomical axis, certainly not the primary locus of depressive pathology. CRH may be elevated in the amygdala, the hippocampus or the locus coeruleus, with or without concurrent elevation of hypothalamic CRH, peripheral corticotropin or cortisol. Second, hypersecretion of stress hormones does not reflect the emotional stress associated with depression. It is of note that many studies have shown a time-course pattern where an elevation of stress hormone release preceded the overt depressive syndrome (see *Endocrinology and Psychiatry*). Moreover, a refined neuroendocrine test, the combined dexamethasone/CRH test, has proved to be of value in identifying patients not showing depressive symptoms but being at risk of relapsing into a depressive episode within six months. This time pattern rejects the possibility that exaggerated stress hormone release is secondary to the emotional stress associated with depression (see *Peptides and Psychiatry*).

3.2 Biogenic Amines

The traditional pathogenetic concepts of depression emerged from the discovery of the pharmacological properties of current antidepressants. Investigators in the United States led by Joe Schildkrauth (1965) formulated the catecholamine hypothesis of depression, submitting that depression is associated with a deficiency of catecholamine activity (particularly norepinephrine) at functionally important adrenoceptor sites in the brain, whereas mania is associated with a relatively excessive catecholaminergic neurotransmission. Norepinephrine-containing neurons project from the locus coeruleus, a brain stem nucleus, into almost all brain areas. Notably, CRH-containing nerve terminals innervate the locus coeruleus. Examples of altered norepinephrine activity are listed in Table 2.

A number of these findings are in agreement with a norepinephrine deficit. In particular, postmortem studies have shown that the locus coeruleus of depressed suicide victims contains increased levels of tyrosine hydroxylase, the rate-limiting enzyme of norepinephrine biosynthesis. This finding is in accordance with either premortem overactivity of adrenergic neurons or a deficit in the biosynthetic product norepinephrine. Similarly, in the same brain nucleus the levels of the presynaptic norepinephrine transporter are decreased, which can also be interpreted as a compensatory downregulation of this protein in response to decreased norepinephrine levels. However, what is puzzling is the failure to induce an aggravation of depressive symptomatology in depressed patients after treatment with a drug that inhibits norepinephrine biosynthesis.

The great acceptance of selective serotonin reuptake inhibitors (SSRIs), reflected by their major (over 70%) share of the total antidepressant market, has rekindled the interest in the serotonin hypothesis of depression originally submitted by Alec Coppen (1969) in the United Kingdom. This hypothesis proposed that decreased serotonergic activity accounts for the development and course of depression. These changes in serotonergic neurotransmission may be attributable to a decreased serotonin release, fewer serotonin receptors (of which exist more than 14 different subtypes), or impaired serotonin receptor-mediated signal transduction. Evidence for a serotonin deficiency, however, is conjectural and relies on (a) decreased 5-hydroxyindoleacetic acid (a principal metabolite of serotonin) concentrations in the cerebrospinal fluid, (b) a neuroendocrine function test using the pharmacological serotonin releaser fenfluramine, which induces blunted prolactin response in depressed patients, and (c) postmortem studies where indicators for serotonergic activity were found to be reduced (Delgado et al. 1994).

It is of note that serotonergic activity and stress hormone activity are closely intertwined. For example, if CRH is elevated in the hippocampus, the stress-elicited serotonin release is reduced and if CRH_1 receptor capacity is reduced, either by targeted mutation of the CRH_1 receptor or by CRH_1 receptor antagonists, serotonin response to stress is elevated, suggesting that blocking CRH-mediated effects through CRH_1 receptor antagonists restores impaired serotonergic activity.

4. Treatment of Depression

The front line of therapy for severe major depression is drug treatment with antidepressants. The effect of these agents is assisted by psychotherapy that needs to be pragmatically adjusted to the patients' needs. The antidepressant drugs available for treatment all work through modulating the neurotransmission of biogenic amines (see above). Some of them are fairly selective such as SSRIs (e.g., fluoxetine, paroxetine or citalopram), or norepinephrine reuptake inhibitors (e.g., reboxetine), while other drugs combine both pharmacologies (e.g., venlafaxine, nefazodone). Because of their pharmacological similarity, it is not surprising that only subtle differences in the clinical profiles of these drugs exist, and that antidepressants are consequently chosen by physicians according to their safety and tolerability profiles, provided their choice is not deterred by the higher costs in countries where reimbursement is limited.

While the superiority of antidepressants over placebo is undisputed among clinician scientists, the exact figures are still a matter of discussion (Quitkin et al. 2000). This is mainly related to the fact that most studies designed to provide the requisite statistical significance for monitoring side-effects and confirming efficacy use large outpatient samples that often contain only moderately depressed patients, have high dropout rates, and include the administration of suboptimal drug dosages. All these factors favor the placebo response, frequently found to be close to 50%, whereas the active drug often does not exceed a 75% response rate. Notably, response in these studies is defined as an improvement of 50% over baseline, which is clinically unsatisfactory and certainly not the desired treatment goal. Thus, the impression emerges from such studies that antidepressants improve the clinical condition by not more than 25%, which is unrealistic. A possible consequence of these figures comes from a study by the WHO which revealed that in patients who were diagnosed with depression, only 31.9% were prescribed an antidepressant, 25.5% received tranquilizers (benzodiazepines), and 19% received 'miscellaneous' drugs such as herbal remedies.

The latter includes St. John's Wort, a drug that is prescribed to half the treated population in Germany, where psychotherapy is also much more widely available than in other countries. This reflects how diverse attitudes toward the pharmacological treatment of depression can be. A current study sponsored by the National Institute of Mental Health, will hopefully elucidate whether the widespread use of St. John's Wort is scientifically justified.

The most important impediments in the pharmacotherapy of depression are (a) the lack of differential therapeutic aids, i.e., there is no tool available to select the appropriate drug for the individual patient; (b) the lack of objective laboratory signs and symptoms to monitor adequacy of treatment; (c) the protracted onset of action, which can take weeks or even months; and (d) side-effects, which are particularly disturbing and often lead to noncompliance after the patient has begun to respond, but still needs treatment to consolidate the beneficial therapeutic effect.

Some of these unmet needs no longer seem to be unsurmountable (Holsboer 2001): (a) differential therapy will be guided by genotyping of patients, helping the physicians to identify which drug is optimally suited, ultimately leading to customized therapies; (b) neuroendocrine function tests such as the combined dexamethasone/CRH test have proved useful in providing feedback information about ongoing neurobiological pathology independently from acute psychopathology ratings. Its more widespread use holds the promise of shortening the course of treatment and hospitalization; (c) onset of action can possibly be shortened by enforced normalization of stress hormone excess via CRH_1 receptor antagonists; and (d) side-effects can be partly reduced by customized drug therapies or by a shift to drugs that target neuropeptides instead of biogenic amine neurotransmitters. Neuropeptide hypersecretion primarily reflects a pathological neuronal hyperactivity, and therefore, neuropeptide receptor antagonists are effective only if neuropeptide hypersecretion exists (see *Ethics and Psychiatry*).

5. Summary and Outlook

The lifetime risk of a major depression is high and, once suffering from depression, the morbidity for other clinical conditions, ranging from common infection to ischemic heart disease, escalates considerably. Thus, depression is already a major cause of disability and will become the most costly disease in the coming decades. In the light of this, any effort to improve the prevention and treatment of depression is justified. The evidence implicating genes as factors predisposing to depression is impressive but complex inheritance has undermined efforts so far to identify those susceptibility genes that interact with environmental factors to precipitate the disease. Despite the current excitement about the possibilities of genetics to identify ultimately a set of gene variants at many different loci that confer a predisposition to depression, I believe that the first revolutionary step will emerge from pharmacogenomics.

This research will elucidate inherited differences in sensitivity to drugs, allowing the physician to predict which particular drug would be best and which harmful for an individual patient. This possibility would be an enormous advantage over today's trial and error strategy and, once feasible, would reduce the time before the right drug at the right dosage is administered. Another forthcoming major change that will not be unanimously embraced by all psychiatrists is the departure from diagnostic categories. These

have been a stumbling block for psychiatric research in the past and will be even more so in a future that is dominated by functional genomics. Therefore, clinical research is challenged to phenotype patients more appropriately on psychopathological and neurobiological levels that cut across traditional diagnostic boundaries. The enormous potential of an interdisciplinary approach encompassing psychopathology, neurobiology, pharmacology, and genetics together, with progress in biotechnology, will lead to a body of knowledge so powerful that not only will better treatments become available, but ultimately, the causal factors underlying depression will also be exposed.

See also: Schizophrenia, Treatment of

Bibliography

Coppen A J 1969 Biochemical aspects of depression. *International Psychiatry Clinics* **6**: 53–81

Delgado P L, Price L H, Miller H L, Salomon R M, Aghajanian G K, Heninger G R, Charney D S 1994 Serotonin and the neurobiology of depression: Effects of tryptophan depletion in drug-free depressed patients. *Archives of General Psychiatry* **51**: 865–74

Heim C, Newport D J, Heit S, Graham Y P, Wilcox M, Bonsall R, Miller A H, Nemeroff C B 2000 Pituitary–adrenal and autonomic responses to stress in women after sexual and physical abuse in childhood. *Journal of the American Medical Association* **284**: 592–7

Holsboer F 1999 The rationale for corticotropin-releasing hormone receptor CRH-R antagonists to treat depression and anxiety. *Journal of Psychiatric Research* **33**: 181–214

Holsboer F 2000 The corticosteroid receptor hypothesis of depression. *Neuropsychopharmacology* **23**: 477–501

Holsboer F 2001 Prospects for antidepressant drug discovery. *Biological Psychiatry*, in press.

Kendler K S, Karkowski-Shumann L 1997 Stressful life events and genetic liability to major depression genetic control of exposure to the environment? *Psychological Medicine* **27**: 539–47

Kessler R C, McGonagle K A, Zhao S, Nelson C B, Huges M, Eshleman S, Wittchen H U, Kendler K 1994 Lifetime and 12-month prevalence of DSM-III-R psychiatric disorder in the United States. *Archives of General Psychiatry* **51**: 8–19

Lesch K P, Bengel D, Heils A, Sabol S Z, Greenberg B D, Petri S, Benjamin J, Muller C R, Hamer D H, Murphy D L 1996 Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science* **274**: 1527–31

Morissette J, Villeneuve A, Bordeleau L, Rochette D, Laberge C, Gagne B, Laprise C, Bouchard G, Plante M, Gobeil L, Shink E, Weissenbach J, Barden N 1999 Genome-wide search for linkage of bipolar affective disorders in a very large pedigree derived from a homogeneous population in Quebec points to a locus of major effect on chromosome 12q23-q24. *American Journal of Medical Genetics* **88**: 567–87

Nathan K I, Musselman D L, Schatzberg A F, Nemeroff C B 1995 Biology of mood disorders. In: Schatzberg A F, Nemeroff C B (eds.) *The American Psychiatric Press Textbook of Psychopharmacology*. American Psychiatric Press Inc, Washington, London, pp. 439–77

Quitkin F M, Rabkin J G, Gerald J, Davis J M, Klein D F 2000 Validity of clinical trials of antidepressants. *The American Journal of Psychiatry* **157**: 327–37

Schildkraut J J 1965 The catecholamine hypothesis of affective disorders: A review of supporting evidence. *The American Journal of Psychiatry* **122**: 509–22

F. Holsboer

Depression, Clinical Psychology of

1. Introduction

The word 'depression' in everyday parlance covers a wide range of emotional states that range in severity from transient moods of sadness to major psychotic episodes accompanied by increased risk of suicide. Depression in the form of a brief sad mood is a universal experience; it is a normal part of living that accompanies the losses, frustrations, failures, and disappointments that all of us face. Clinical depression in contrast, is a syndrome, or constellation of co-occurring psychiatric symptoms, that affects about 20 percent of the population. Major Depressive Disorder, the psychiatric label for clinically significant depression, is characterized by at least a two-week period of persistent sad mood or a loss of interest or pleasure in daily activities, and four or more additional symptoms, such as marked changes in weight or appetite, sleep disturbance, psychomotor agitation or retardation, fatigue, feelings of guilt or worthlessness, and concentration difficulties.

People diagnosed with Major Depressive Disorder show marked impairment in their social and occupational functioning; they also have an elevated risk of death from a number of causes. In fact, the World Health Organization Global Burden of Disease Study recently ranked depression as the single most burdensome disease in the world in terms of total disability-adjusted life years among people in the middle years of life (Murray and Lopez 1996). To give but one example of the consequences of this disorder, a recent economic analysis of depression in the workplace estimated that the annual salary-equivalent costs of depression-related lost productivity in the United States exceeds $33 billion (Greenberg et al. 1996). Moreover, because this figure does not take into account the impact of depression on such factors as the performance of coworkers, turnover, and industrial accidents, it is likely to be an underestimate of the overall workplace costs of depression.

People who become depressed rarely experience only a single episode; depression is a highly recurrent

disorder. In fact, over 80 percent of depressed patients will have more than one depressive episode during their lifetime. The interval between episodes is generally short: over half of depressed patients will have at least one relapse within two years of their recovery from a depressive episode. This pattern of recurrence also appears to intensify over time: individuals with three or more previous episodes of depression may have a relapse rate as high as 40 percent within only three to four months after recovery. The high recurrence rate in depression points to the existence of factors that serve to increase people's risk for developing this disorder. In attempting to understand this vulnerability to depression, clinical psychologists have focused on two broad areas in which depressed individuals seem to experience particular difficulties: cognitive functioning and interpersonal functioning. In Sects. 2 and 3, therefore, we describe the results of research conducted to examine the cognitive and interpersonal functioning of depressed persons and of individuals who are at risk for becoming depressed.

2. *Cognitive Functioning in Depression*

In many ways, depression is a disorder of thought. Depressed individuals typically think negatively about themselves, believing that they are useless, deficient, and unattractive people. Depressed persons also express bleak views of their immediate surroundings and of their future prospects. Given the pervasiveness of these dysfunctional cognitive characteristics, scientists have theorized about the importance of cognitions in depression and have conducted considerable research examining how depressed individuals perceive and evaluate information from their environment.

The cognitive theory of depression that has had the greatest impact on research and treatment is that formulated by Aaron Beck (1976), who ascribes the onset of depression to dysfunctional cognitive processes. Beck hypothesizes that individuals who experience loss or adversity in childhood develop a set of negative expectancies, or 'schemas,' concerning loss, failure, or abandonment. These schemas serve as filters through which stimuli and events in their environment are perceived, evaluated, attended to, and remembered. The negative schemas of depressed persons lead them to perceive and evaluate neutral stimuli as negative, guide their attention to negative aspects of their environment, and facilitate their memory for negative experiences. Perhaps most importantly, Beck believes that these negative schemas characterize not only currently depressed persons, but also nondepressed people who are at a high risk for developing depression in the future. Indeed, the negative schemas are the risk factor. In these individuals, the negative schemas remain inactive until the person encounters a relevant stressful event or experience. The stressful experience serves to activate the negative schema, which leads the individual to process information in a negative manner, in turn leading to ineffective coping, culminating in a depressive episode.

Thus Beck's theory, and other similar cognitive theories of depression, predicts that depressed individuals (and individuals who are at risk for becoming depressed) are characterized by negative schemas that lead them to be more attentive to negative than to positive stimuli in their environment and to have better memory for these stimuli. Early studies attempted to examine the cognitive schemas of depressed persons by assessing their responses to self-report questionnaires measuring the negativity of their beliefs. In fact, a large body of literature has now documented that depression is associated with high levels of self-reported dysfunctional attitudes and a negative explanatory style.

Although these studies were important in describing the cognitive styles of depressed persons, their reliance on self-report measures of cognitive schemas is problematic. Perhaps most importantly, it is unlikely that self-report questionnaire measures, which require subjects to make conscious, deliberate, and thoughtful responses, are able to assess the existence and functioning of cognitive schemas, which are hypothesized to be activated automatically and to operate outside of individuals' awareness. Put more simply, because people are typically not aware of the content of their schemas, self-report measures cannot be used to assess them. To address this problem, clinical psychologists have adapted methodologies from research in experimental cognitive psychology to assess the schematic functioning of depressed individuals. Using such information-processing tasks as the emotion Stroop task and the dot-probe task, investigators have provided empirical support for Beck's cognitive theory of depression, demonstrating that depressed individuals are characterized by negative cognitive biases in both attention and memory (see Gotlib and Neubauer 2000).

Whether measured by self-report or by information-processing tasks, there is now little question that depression is associated with negative cognitive biases, a conclusion that supports a major tenet of cognitive theories of depression. Importantly, however, the causal status of these negative biases is still an open question. Cognitive theories of depression also hypothesize that negative schematic functioning is a trait-like characteristic that increases an individual's risk for developing depression. There are very few empirical studies in which investigators have attempted to predict subsequent depression from earlier measures of schematic functioning. Moreover, studies of formerly depressed persons have generally found that cognitive functioning normalizes with recovery from depression. A number of researchers have demonstrated, however, that the presence of negative mood in formerly depressed individuals can reinstate

negative cognitive processing which was not observable in the absence of negative mood (Haaga et al. 1991). If this proves to be the case, it is not difficult to see how stressful life events could generate a negative mood state, which, among vulnerable individuals, also activates a set of dysfunctional cognitions that lead to depression. Thus, although much more work remains to be done, it is clear that depressed individuals are characterized by negative schemas, and it is possible that this schematic functioning plays a causal role in the onset of depression.

3. *Interpersonal Functioning in Depression*

As is the case with cognitive functioning, depressed individuals have consistently been found to experience difficulties in their interpersonal relationships. While there is no single guiding theory of the social functioning of depressed persons that is analogous to Beck's cognitive theory, empirical research has nevertheless implicated interpersonal dysfunction as important in understanding the etiology and maintenance of depression, as well as relapse of this disorder. Early behavioral formulations of depression viewed the depressed state as resulting from a lack of environmental reinforcement, stemming in part from an inability of depressed persons to elicit reinforcement from others (e.g., Lewinsohn 1974). Indeed, numerous studies have demonstrated that depressed persons exhibit social skill deficits during interactions with others. For example, compared to nondepressed individuals, depressed persons have been found to speak more slowly and with less volume and voice modulation, to show longer pauses in their speech patterns, to take longer to respond to others, and to engage in less eye contact. They are also more self-centered in interactions and steer conversations to more negative content. Finally, and understandably, depressed persons seem to induce feelings of depression, anxiety, and hostility in those with whom they interact. It is not surprising to learn, therefore, that people ultimately withdraw from interactions with depressed individuals (Segrin 2000).

Although this negativity seems to permeate the full range of depressed individuals' social interactions, it is especially pronounced in intimate relationships. In this context, a number of investigators have focused on the nature of the relationships between depressed persons and their spouses and children. There is good reason for concern in these areas. For example, the divorce rate among individuals who have been diagnosed with depression is nine times higher than the rate in the general population (Merikangas 1984). Moreover, children of depressed parents are two to three times more likely experience a psychiatric diagnosable disorder at some point in their lives than are offspring of nondepressed parents (Gotlib and Goodman 1999). In attempting to understand the processes underlying these findings, researchers have systematically observed the interactions between depressed people and their spouses, and between depressed parents and their children.

The marital interactions of depressed persons are characterized by high levels of anger, conflict, and negative affect. Depressed spouses derogate themselves and their partners, and both spouses escalate their negative affect and behaviors over the course of the interactions. Interestingly, expressions of sad affect in the depressed spouse appear to have the effect of suppressing anger and aggression in the partner, suggesting that depression may play a functional but maladaptive role in the marriage. With respect to their children, depressed individuals report that they find it difficult to be warm and consistent parents, that they do not derive satisfaction from their children, and that they feel inadequate in their parenting role. Consistent with these self-reports, in interactions with their children, depressed mothers display sad and irritable affect and are either unresponsive or intrusive. In short, depressed parents appear to be difficult or inadequate social partners for both their spouses and their children, although it is important to also recognize the likely reciprocal nature of these problematic relationships.

Thus there is consistent evidence that depression is associated with deficits in interpersonal functioning. There is less support for the position that difficulties in social functioning serve as risk factors for the onset of a first depressive episode, although some investigators have reported that the lack of a supportive intimate relationship may leave individuals particularly vulnerable to the effects of life stress. Importantly, however, many of the interpersonal problems that characterize people while they are depressed have been found to persist following recovery. For example, in contrast to other disorders such as alcoholism, the marital relationships of people who have been depressed continue to be strained even after they have recovered from their depressive episode. In fact, like currently depressed individuals, people who have recovered from depression also report having fewer friends and close relationships than do individuals who have never been depressed. Interestingly, recovered depressed individuals do not show deficits in the overall number of their network contacts, suggesting that the interpersonal difficulties of formerly depressed persons—like those of currently depressed individuals—are most pronounced in intimate relationships. Finally, and further underscoring the importance of focusing on intimate relationships in understanding the role of interpersonal deficits in depression, the degree of hostility or criticism expressed by spouses about their depressed partners is a strong predictor that the partner will relapse into another episode of depression (Hooley and Gotlib 2000).

As is apparent from this brief review, individuals who are experiencing clinically significant depression

are characterized by negative cognitive biases in their perceptions and evaluations of their environment and by impaired interpersonal functioning. Consistent with depressed persons' difficulties in these areas, two of the most effective psychological treatments for this disorder are cognitive therapies and interpersonal therapies. In the following sections we describe these two approaches to the treatment of depression.

4. *Cognitive Treatments for Depression*

Cognitive therapies are based on the formulation that individuals' affect and behavior are largely determined by how they interpret events in their world. As we have seen, clinically depressed individuals are characterized by negative perceptions of the self, the world, and the future. Therefore, a critical premise of cognitive therapies for depression is that replacing maladaptive thoughts with more positive cognitions will result in a significant reduction in depressive symptoms. Cognitive therapies often incorporate behavioral interventions to help in altering the patient's cognitions; for this reason, these therapies are often referred to as cognitive-behavioral treatments.

Cognitive therapy for depression is time-limited, rarely exceeding 30 weekly sessions and ranging most typically from 15 to 25 sessions (see also *Cognitive Therapy*). The therapist and patient work together in a process referred to as 'collaborative empiricism' (Beck et al. 1979) to identify, evaluate, and change the patient's negative thoughts. The therapist teaches the patient behavioral coping strategies, such as problem-solving skills and assertiveness training, to help him or her bolster new positive cognitions. Often homework is assigned to help the patient identify negative thoughts and beliefs and practice adaptive behaviors that the patient has learned with the therapist. Finally, because depression has a high relapse rate, in the final sessions of cognitive therapy, the therapist works with the patient to anticipate the kinds of life stressors that he or she may encounter in the future to prevent the recurrence of depression.

Cognitive therapy and cognitive-behavior therapy for depression have been evaluated in over 80 controlled studies (American Psychiatric Association 2000). The results of these clinical trials are consistent: cognitive therapy is an effective treatment for depression. In fact, cognitive therapy is considered by the American Psychological Association to be a 'well-established treatment' for depression (Chambless and Hollon 1998). Although there have been exceptions, cognitive therapy has generally been found to be as effective as (and sometimes more effective than) alternate forms of treatment for depression, including antidepressant medication. Moreover, several studies have shown that depressed patients who received cognitive therapy are less likely to relapse following treatment termination than are patients treated with medications. Finally, there is now evidence that cognitive therapy can prevent or delay the onset of depression in persons who are at elevated risk for the disorder (Hollon et al. in press).

5. *Interpersonal Treatments for Depression*

Interpersonal psychotherapies were developed from broad interpersonal theories of psychopathology, such as those formulated by Adolph Meyer, Harry Stack Sullivan, and John Bowlby. These theories posited that the quality of current interpersonal relationships was critical in contributing to the development and persistence of many forms of psychopathology. From this perspective, therefore, the therapist's primary goal is to change the patient's patterns of self-defeating interpersonal interactions. The most systematic and widely used therapy in this area is Interpersonal Psychotherapy (IPT), developed by Gerald Klerman and Myrna Weissman for the treatment of depression (Weissman et al. 2000) (see *Cognitive and Interpersonal Therapy*: *Psychiatric Aspects*. Like cognitive therapy, IPT is conducted most frequently as a short-term therapy (approximately 16 weekly sessions), but it has also been modified for use as a 'maintenance therapy' for the longer-term treatment of patients with recurrent or chronic depression. A major goal of IPT is to improve the patient's interpersonal functioning by encouraging more effective expression of emotions, clearer communication with significant members of their entourage, and increased understanding of the patient's behavior in interpersonal interactions. The rationale underlying the use of IPT is that by solving interpersonal problems in therapy, the patient will improve their life situation and simultaneously decrease the number and intensity of depressive symptoms.

IPT for depression is divided into three phases. In the initial phase of treatment (the first 3 sessions), the therapist conducts a diagnostic evaluation for depression, educates the patient about depression, and evaluates the patient's current interpersonal relationships. The therapist then establishes which of four interpersonal problem areas are most closely related to the patient's current depressive episode: grief, interpersonal role disputes, role transitions, or interpersonal deficits. In the middle phase of treatment (sessions 4–13), the therapist uses specific techniques (outlined explicitly in a treatment manual) that are designed to address these four problem areas. Although these techniques may involve changing the patient's cognitions or having the patient engage in new behaviors, there is a clear and strong tie to the patient's current interpersonal relationships. The final phase of treatment (sessions 14–16) focuses on consolidating the changes that the patient has made through therapy and helping the patient recognize and counter depressive symptoms should they arise again in the future.

In the 1980s and 1990s, IPT has been carefully evaluated in numerous research protocols and has been demonstrated to be effective in treating depression in a variety of populations, including adolescents, postpartum depressed women, married spouses, the elderly, and patients in primary medical care facilities. In fact, IPT was one of two forms of psychotherapy (the other was cognitive therapy) tested against antidepressant medication in a large-scale treatment study of 250 depressed patients (Elkin et al. 1989). The results of this study indicate that IPT is more effective than placebo, and as effective as medication and cognitive therapy in reducing depressive symptoms. In addition, IPT was found to be more effective than cognitive therapy in treating more severely depressed patients (Elkin et al. 1989). Importantly, IPT has been shown to prevent or delay the onset of relapse episodes of depression, demonstrating its utility as a maintenance therapy.

6. *Conclusions and Future Directions*

In the 1980s and 1990s, we have made significant progress in our understanding of psychosocial aspects of depression and in the treatment of this disorder. It is now clear that depressed persons are characterized by maladaptive cognitive schemas that guide their attention to negative, as opposed to positive, environmental information, and enhance their memory for this negative information. Depressed persons also have marked impairments in their interpersonal functioning, particularly in more intimate relationships. Despite this progress, however, there are several issues that remain unresolved.

In closing, therefore, we offer three recommendations for future research examining cognitive and interpersonal functioning in depression. First, much of the research that has been conducted in this area has been cross-sectional, essentially comparing samples of depressed and nondepressed participants at one point in time. We urgently need prospective, longitudinal studies in order to clarify the causal relation of cognitive and interpersonal dysfunction to the onset, maintenance, and relapse of depression. Second, while both cognitive and interpersonal therapies have been demonstrated to be effective in the treatment of depression, there are anomalies that need to be addressed. For example, cognitive therapy is not universally effective in the treatment of depression and research must begin to identify characteristics that differentiate treatment responders from nonresponders. Similarly, despite its focus on interpersonal functioning, IPT has been found to be no more effective in improving social functioning immediately following treatment than are medications or cognitive therapy. Therefore, for both cognitive therapy and IPT, research is needed to clarify the mechanisms by which the treatment reduces patients' levels of depression. Finally, we have discussed cognitive and interpersonal functioning in depression as though they occur independent of each other. Clearly, this is not the case. It is only by explicitly integrating data concerning the cognitive and interpersonal functioning of depressed persons, and relating that data to their biological functioning, that we will be able to build a comprehensive understanding of depression allowing us to develop even more effective programs for the prevention and treatment of this disorder.

See also: Childhood Depression; Cognitive and Interpersonal Therapy: Psychiatric Aspects; Cognitive Therapy; Depression; Depression, Hopelessness, Optimism, and Health; Information Processing and Psychopathology; Interpersonal Psychotherapy; Psychological Treatments, Empirically Supported; Social Cognitive Theory and Clinical Psychology; Suicide

Bibliography

American Psychiatric Association 2000 Practice guideline for the treatment of patients with major depressive disorder (revision). *American Journal of Psychiatry* **157**(Suppl. 4): 1–45

Beck A T 1976 *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York

Beck A T, Rush A J, Shaw B F, Emery G 1979 *Cognitive Therapy of Depression*. Guilford Press, New York

Chambless D, Hollon S D 1998 Defining empirically supported therapies. *Journal of Consulting & Clinical Psychology* **66**: 7–18

Elkin I, Shea M T, Watkins J T, Imber S D, Sotsky S M, Collins J F, Glass D R, Pilkonis P A, Leber W R, Docherty J P, Fiester S J, Parloff M B 1989 National Institute of Mental Health treatment of depression collaborative research program: general effectiveness of treatments. *Archives of General Psychiatry* **46**: 971–82

Gotlib I H, Goodman S H 1999 Children of parents with depression. In: Silverman W K, Ollendick T H (eds.) *Developmental Issues in the Clinical Treatment of Children*. Allyn and Bacon, Boston, pp. 415–32

Gotlib I H, Neubauer D L 2000 Information-processing approaches to the study of cognitive biases in depression. In: Johnson S L, Hayes A M, Field T M, Schneiderman N, McCabe P M (eds.) *Stress, Coping, and Depression*. Erlbaum, Mahwah, NJ, pp. 117–43

Greenberg P, Kessler R, Nells T, Finkelstein S, Berndt E R 1996 Depression in the workplace: an economic perspective. In: Feighner J P, Boyer W F (eds.) *Selective Serotonin Reuptake Inhibitors: Advances in Basic Research and Clinical Practice*. Wiley, New York, pp. 327–63

Haaga D A, Dyck M, Ernst D 1991 Empirical status of the cognitive theory of depression. *Psychological Bulletin* **110**: 215–36

Hollon S D, Haman K L, Brown L L (in press) Cognitive behavioral treatment of depression. In: Gotlib I H, Hammen C L (eds.) *Handbook of Depression: Research and Treatment*. Guilford Press, New York

Hooley J M, Gotlib I H 2000 A diathesis-stress conceptualization of expressed emotion and clinical outcome. *Journal of Applied and Preventive Psychology* **9**: 135–51

Klerman G L, Weissman M M, Rounsaville B J, Chevron E S 1984 *Interpersonal Psychotherapy of Depression*. Basic Books, New York

Lewinsohn P M 1974 A behavioral approach to depression. In: Friedman R J, Katz M M (eds.) *The Psychology of Depression: Contemporary Theory and Research*. Wiley, New York, pp. 157–85

Merikangas K R 1984 Divorce and assortative mating among depressed patients. *American Journal of Psychiatry* **141**: 74–6

Murray C J L, Lopez A D (eds.) 1996 *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020*. Harvard University Press, Cambridge, MA

Segrin C 2000 Social skills deficits associated with depression. *Clinical Psychology Review* **20**: 379–403

Weissman M M, Markowitz J C, Klerman G L 2000 *Comprehensive Guide to Interpersonal Psychotherapy*. Basic Books, New York

I. H. Gotlib and J. Rottenberg

Depression, Hopelessness, Optimism, and Health

Depression is a psychological syndrome encompassing a variety of symptoms, including loss of energy and motivation; loss of the capacity to experience happiness or enjoyment; feelings of sadness, despair, and hopelessness; and sometimes a sense of guilt, shame, or worthlesssness. Depression is a debilitating problem in its own right for people who experience it. Depression can also have adverse implications for physical health and well-being.

Depression develops in people who initially were not depressed. Though the experience can be long-lasting or repeated, it does not usually last for the person's entire lifetime. Certain personality traits, however, have overtones of qualities that are part of depression. Consider the dimension of optimism vs. pessimism. Optimists are people who expect good things to happen to them; pessimists expect bad things to happen. Pessimism thus has a link to depression. Unlike depression, however, pessimism is an enduring personality trait. The personality dimension of optimism vs. pessimism has been studied a good deal in health psychology. Some of that research has found that this aspect of personality has consequences for physical well being.

The concept of hopelessness has links to both depression and pessimism. Sometimes hopelessness is used to refer to a state; sometimes it is used to refer to personality qualities. One might view the concept of hopelessness as representing a conceptual bridge between the (usually temporary) state of depression and longer-lasting traits of personality such as pessimism. People who are experiencing depression often feel hopeless about the future; people who are pessimistic are more broadly and consistently lacking in hope for the future.

This article describes some of the research that links depression, hopeless, and pessimism to health problems. As background to that research, the next section describes the conceptual basis for examining how these variables may be involved in health.

1. Pessimism, Hopelessness, and Disengagement

The concepts of optimism and pessimism are rooted in centuries of folk wisdom about the consequences of people's expectations for the future. Contemporary research on optimism and pessimism also has its roots in people's expectations for the future. In adopting that starting point, this research connects to more than a century of psychological theory (Carver and Scheier 1998, 1999). That is, the expectancy construct links this research to a collection of theories of motivation and action that are called expectancy-value theories. This link, in turn, provides a rationale for examining the way in which the traits of optimism and pessimism are expressed in motivation, behavior, and feelings. These expressions of optimism and pessimism represent a set of potential pathways by which this personality disposition may relate to health.

Expectancy-value models begin with the idea that behavior is aimed at attaining desired goals (values). Without a valued goal, there is no reason to act. The other element is expectancy: confidence or doubt about being able to attain the goal. If a person lacks confidence, again there will be no action. Only with sufficient confidence do people exert efforts and remain engaged in those efforts. These principles pertain to specific values and focused confidence, and they also apply to the more general traits of optimism and pessimism. In optimism and pessimism the 'confidence' involved is simply broader in scope.

These principles provide the basis for many predictions about optimists and pessimists: when confronting a challenge, optimists should take a posture of confidence and persistence, even if progress is difficult or slow. Pessimists should be more doubtful and hesitant. This divergence should be amplified under conditions of adversity. Optimists are likely to believe the adversity can be handled successfully, pessimists are likely to anticipate disaster. These differences in orientation can lead to differences in persistence, differences in active efforts to take precautions, differences in many decisions and actions that pertain to health risks.

People's behavioral responses to adversity are important, but overt behavior is not the only response that occurs when people confront adversity. People also experience emotions in such situations. Difficulties elicit a mixture of feelings, and variables such as optimism influence the balance among the feelings. Again, predictions are easily derived from theory. Optimists expect to have good outcomes, even when things are difficult. This confidence should yield a

more positive mix of feelings. Pessimists expect bad outcomes. This should be related to more intense negative feelings—anxiety, sadness, or despair. Though these emotional responses are not the focus of this article, a good deal of research has found evidence of such emotional differences between optimists and pessimists (Carver and Scheier 1999).

The experiences of intense distress, hopelessness, and disengagement from effort have physical concomitants; so do positive emotions and engagement in positive efforts. A number of psychologists believe that these physiological responses can play a role in health outcomes. Such processes may influence a person's probability of getting a disease in the first place, they may influence the progression of diseases, they may even influence mortality. There is a good deal of evidence that personality differences in optimism–pessimism relate to differences in health-related outcomes. Research on the precise pathways of influence is at an early stage, but enough evidence has emerged to suggest that the questions are worth examining further.

Researchers have studied the effects of pessimism and hopelessness in two ways. In some studies the variables are measured as generalized personality dispositions. In other studies they are measured as attitudes toward particular contexts or particular health outcomes. Studies that examine both sorts of pessimism typically find a moderate correlation between the two, but sometimes the relation is much weaker (raising interesting questions about the organization of personality that go beyond the scope of this article). Whether deriving from a generalized or a context-specific approach, however, evidence suggests that pessimism and hopelessness have adverse implications for health.

In the review that follows, some studies used generalized measures, others used more focused measures. A few of the studies did not measure pessimism or hopelessness directly, but rather the psychological responses they are believed to induce, such as disengagement vs. active effort.

2. *Pessimism, Hopelessness, Depression, and Health*

How do pessimism, hopelessness, and depression relate to health outcomes? This selective review of studies is organized as follows. First are examples of how behavioral reflections of optimism vs pessimism, in the form of effort vs disengagement, can influence health risk and thus well being. Then comes a more extensive section dealing with disease processes themselves—disease promotion, progression, and mortality—and their relations to pessimism and depression.

2.1 *Optimism and Health Related Actions*

Sometimes being persistent in one's goal related effort is itself an important determinant of health and well being. An example of the relevance of behavioral persistence comes from work on active health promotion (see *Health Education and Health Promotion*). Optimists, who are confident about attaining and maintaining the goal of good health, should be more likely than pessimists to continue pursuit of health promotion activities even if it entails a good deal of effort. Evidence from several studies supports this hypothesis.

One study looking at potential individual differences in health promotion followed a group of heart patients who were involved in a cardiac rehabilitation program (Shepperd et al. 1996). Optimism among participants was related to increases in exercise levels across the study period. Optimism also related to greater success in lowering levels of saturated fat, lowering body fat, and lowering an index of global coronary risk.

Another study followed coronary artery bypass patients during the period immediately after surgery (Scheier et al. 1989) and then re-examined them five years later (described in Scheier and Carver 1992). The early part of the project focused mostly on patients' responses to the surgery as an event in their lives; the later part of the project investigated their lifestyles five years postsurgery. In the early part of the project, optimists gave several indications of focusing their efforts on getting back on their feet quickly. They were viewed by the hospital staff as recuperating faster, and they reported concentrating on the rehabilitation process. Pessimists, in contrast, were more likely to be avoiding even thinking about the future. Five years later, optimists were more likely than pessimists to be taking vitamins, to be eating low-fat lunches, and to be enrolled in a cardiac rehabilitation program. Thus, both initially and later on, pessimists were less persistent than optimists in health promoting behaviors.

Heart disease is one aspect of health to which behavioral efforts are relevant. Another health risk that is related to behavior, though in a different manner, is HIV infection. Infection can occur from engaging in sexual activities with partners who are not known to be disease free (see *Sexual Risk Behaviors*). By taking steps to avoid such sexual practices—for example, avoiding sex with unknown partners—people reduce their risk of infection. One study of HIV-negative gay men found that optimists reported having fewer anonymous partners than pessimists (Taylor et al. 1992). This finding suggests that optimists were making efforts to reduce their risk, thereby safeguarding their health (see *Sexually Transmitted Diseases: Psychosocial Aspects*).

Another study relevant to this issue concerned health behavior and intentions among a group of hospital employees at risk for skin cancer (Friedman et al. 1995). Optimists were significantly more likely

than pessimists to report intentions to engage in skin cancer-relevant health prevention behaviors (e.g., regular sunscreen use). In addition, among those identified at screening as having suspicious lesions, those who were more optimistic were more likely to comply with recommended follow-up care.

Other studies have examined the health-related habits reported by groups of people with no particular salient health concerns. At least two such projects have found associations such that optimists reported more health-promoting behaviors than did pessimists (Robbins et al. 1991, Steptoe et al. 1994). In sum, these various studies suggest that optimism is associated with behaviors aimed at promoting health and reducing health risk.

It is important to note, in this regard, that optimists are not simply people who stick their heads in the sand and ignore threats to their well-being. Rather, they display a pattern of attending selectively to risks—risks that both are applicable to them and also are related to serious health problems (Aspinwall and Brunhart 1996). If the potential health problem is minor, or if it is unlikely to bear on them personally, optimists show no elevated vigilance. Only when the threat matters does it emerge. This fits the idea that optimists scan their surroundings for threats to their well-being, but save their active coping efforts for threats that are truly meaningful.

2.2 Pessimism and Disengagement

This discussion derives from the idea that optimism leads to effort and pessimism leads to giving up. Thus far the focus has been on persistence. What about giving up? Several kinds of health-relevant behaviors seem to reflect a giving up tendency. One of them is alcohol consumption, which at high levels is a health risk. Excessive alcohol consumption is often seen as occurring in the service of an escape from one's problems. If so, pessimists should be more vulnerable than optimists to alcohol abuse.

Results from at least two studies fit this picture. A study of women with a family history of alcoholism found that pessimists were more likely than optimists to report drinking problems (Ohannessian et al. 1993). Another study (Strack et al. 1987) examined people who had already been treated for alcoholism and were now entering an aftercare program. This study found that pessimists were more likely to drop out of the aftercare program and to return to drinking than optimists. These two studies converge in showing that pessimists display a form of disengagement—alcohol consumption—more than optimists.

People can give up in many ways, of course. Alcohol dulls awareness of failures and problems. People can also turn their backs on their problems by distracting themselves. Even sleeping can help people escape from situations they do not want to face. Sometimes, though, giving up is more complete. Sometimes people give up not just on specific goals, but on all the goals that define their lives, by committing suicide. Pessimism has been found in several studies to be a key indicator of suicide risk (Beck et al. 1985, Fawcett et al. 1987, Petrie and Chamberlain 1983). Although most people might not think of suicide as a 'health' outcome, suicide clearly does interfere with the maintenance of good health.

2.3 Disease Susceptibility

The examples just described all concern how people varied in their efforts to protect themselves against health risks, or how hard they tried to maintain good health, or the extent to which they engaged in behaviors reflecting a disengagement from such efforts. Such behavioral factors are important underpinnings to disease susceptibility. That is, people who take steps to avoid health risks will generally be healthier in the long run than people who do not. But are there more direct associations? Are there cases in which people differ in their susceptibility to diseases apart from their actions?

It is difficult to study initial vulnerability to disease, because doing so requires recruiting participants who are all healthy, and waiting until some of them become ill. An alternative strategy that some have employed is to recruit persons who are all in the same category of risk for a disease. One study using this procedure produced results suggesting that people's orientation to life can play a role in whether they develop cervical cancer. This study (Antoni and Goodkin 1988) examined women who were undergoing testing after an abnormal PAP smear. Some of the women later proved to have cancer, some had a precancerous condition that might eventually develop into cancer, and others showed little or no abnormality. Before testing, these women completed measures of tendencies toward hopelessness and despair about the future. Women whose abnormality ultimately proved to be more severe were the women who had reported more hopelessness.

Another study examined potential associations of pessimism with blood pressure elevation (Räikkönen et al. 1999). This study involved three days of ambulatory monitoring of blood pressure during normal activities. Results indicated that pessimists were more likely to have elevations in blood pressure than optimists. Although this study should not be taken as clear evidence of disease initiation, it suggests that further examination of the role of pessimism in hypertension is warranted.

2.4 Health Following Medical Procedures

A little research has examined pessimism and the initiation of disease, but a good deal more has

examined differences in well-being that follow the initial diagnosis or treatment of disease. Some of this work looks at responses to medical procedures. For example, the study by Scheier et al. (1989) of men undergoing coronary bypass surgery (which was described a little earlier) included measures of the patients' progress through the surgery itself. Two kinds of indirect evidence suggested that pessimists fared more poorly than optimists even on the operating table. Specifically, pessimists were more likely to display two markers during surgery that are widely taken as indicants of myocardial infarction (MI). Thus, pessimists may have been at greater physical risk during the bypass surgery itself.

Scheier et al. (1999) have more recently collected data on a second sample of bypass patients. The first analyses of data from that project focused on a problem that often arises after major surgery: the need for rehospitalization, due either to infection from the surgery or to complications from the disease underlying the surgery. Scheier et al. (1999) found that optimistic persons were significantly less likely to be rehospitalized either for problems related to post-surgical infection or for problems related to the coronary artery disease, even after controlling for medical variables and other personality measures.

Another research group has explored how positive expectations before heart transplant surgery related to later medical adherence, and health (Leedham et al. 1995). In this case, optimism vs. pessimism was assessed at several levels of abstraction. Patients completed measures of optimism prior to surgery; they were then evaluated at discharge and at 3 and 6 months postsurgery. Initial scales assessed patients' confidence about the efficacy of treatment, their expectations about their future health and survival, and broader expectations for the future. Confidence about the future predicted better adherence to the postoperative medical regimen, and strongly predicted nurses' ratings of the patients' physical health at six months after surgery. There was also a tendency for positive expectations to predict longer delays before development of infection (which is a near-universal side effect of heart transplantation).

2.5 Disease Progression

The successful management of disease can be reflected in many ways. In addition to examining responses to medical procedures, researchers have also looked at the process of disease progression. Many chronic, or incurable, diseases produce a sequence of new symptoms. An example is HIV infection, which initially has no observable symptoms, but eventually produces symptoms of a variety of types, which become increasingly debilitating.

At least two studies have examined the development of symptoms in men who were HIV-positive but symptom-free, as a function of psychological variables related to hopelessness and pessimism. In one study (Reed et al. 1999), the men completed a measure which included an index of stoic acceptance. This index reflects a fatalistic mental preparation for the worst and acceptance of the inevitable, and has been characterized as reflecting pessimistic disease-specific expectancies (Reed et al. 1994). In this sample of men who were initially free of symptoms, responses to this measure indicating greater pessimism predicted earlier symptom onset (Reed et al. 1999).

Another study (Ironson et al. 1994) examined the effect of the giving-up response that was linked to pessimism in earlier sections. This project followed asymptomatic HIV-positive men for two years after informing them of their positive serostatus. Measures of coping reactions (including a measure of disengagement—the giving-up response) were collected 5 weeks after diagnosis. Those who at the postdiagnosis assessment reported greater disengagement were more likely to have developed AIDS symptoms two years later.

2.6 Recurrence and Death

Another body of work moves the discussion from management of disease and progression of disease to disease recurrence and even mortality. Greer and coworkers have for some time studied individual differences among breast cancer patients (see *Cancer: Psychosocial Aspects*). Some patients react to their diagnosis and surgery with what has been termed 'fighting spirit,' a focused engagement with the struggle of living and regaining strength. Others react instead with stoic acceptance (fatalism) or feelings of hopelessness and helplessness. Patients who reacted to the initial cancer with fighting spirit were more likely to have recurrence-free survival at five-year follow-up than women who reacted with hopelessness and helplessness (Greer et al. 1979). This pattern held up at both 10-year follow-up (Pettingale et al. 1985) and 15-year follow-up (Greer et al. 1990).

One project bearing on psychosocial variables and survival tested a six-week group intervention in a sample of malignant melanoma patients (Fawzy et al. 1993). Among measures collected was a measure of active-behavioral coping with the illness. Active-behavioral coping was defined as trying to alter the course of the disease by activities such as exercise, relaxation, and frequent consultations with physicians. People with this coping pattern appeared to be staying engaged with the struggle to stay alive, consistent with the picture of optimism described earlier. Fawzy et al. found that higher initial levels of this pattern predicted lower rates of recurrence and death.

In the previous section a study was described in which stoic acceptance of bad outcomes had adverse effects on symptom development in HIV patients. Further evidence of the ill effects of that attitude

comes from research on survival time among patients whose disease had already progressed to AIDS (Reed et al. 1994). By the end of that study, 82 percent had died from complications related to the disease. Higher scores on the index of stoic acceptance (reflecting negative disease-specific expectancies) related to shorter survival times. It was as though individuals with this attitude were preparing to die, and death then came to them more quickly.

Another project bearing on the role of doubt and disengagement in disease progression examined a sample of recurrent cancer patients (Schulz et al. 1996). Patients were followed for 8 months, by which time approximately one-third had died. All had earlier completed a measure of pessimism about the future. Controlling for levels of symptoms at baseline, greater pessimism predicted shorter survival time. Further, the findings were specific to pessimism. Depression did not predict mortality.

One last study, conducted in Finland (Everson et al. 1996), examined the relationship between a sense of hopelessness about the future and mortality in a sample of over 2,000 middle-aged Finnish men who had been treated for cancer or heart attacks. Hopelessness was assessed by two items: 'I feel that it is impossible to reach the goals I would like to strive for' and 'The future seems for me to be hopeless, and I can't believe things are changing for the better.' These men were followed for 6 years. Those who had reported higher degrees of hopelessness had greater disease-specific mortality—and all-cause mortality—than men with less hopelessness.

2.7 *Associations between Depression and Health Outcomes*

Most of the foregoing has dealt with pessimism or hopelessness. What about depression and health outcomes (see Scheier and Bridges 1995 for broader review).

Links between depression and health are best documented for various kinds of cardiac-related events. For example, Ladwig et al. (1991) gathered psychological data from over 500 male survivors of MIs during the third week after the MI, including a measure of depression. By six-month follow-up, 12 of these patients had died from cardiac events. Death was predicted significantly by post-MI depression levels. When the data were adjusted for medical variables, the association between depression and cardiac death was weakened, but remained marginally significant.

Stronger findings relating depression to post-MI mortality have been reported by Frasure-Smith et al. (1993), using a more standardized measure of depression. They interviewed over 200 men and women who had been hospitalized for MI, during the week after the MI. Depression was assessed using an interview that gathers information used to make psychiatric diagnosis. In addition, a large number of medical variables were collected. By six months after the initial interview, 12 cardiac-related deaths had occurred, and depression was a significant predictor. Patients meeting criteria for clinical depression during the week after their MI were five times more likely than nondepressed patients to have died by the six-month follow-up. The effect of depression remained significant even after the data were adjusted for all clinical predictors of death.

Frasure-Smith et al. (1999) have more recently reported a larger (896 patients) and longer-term (one-year) study of men and women who had been hospitalized for MI. Analysis of survival data indicated that both men and women who were depressed while in the hospital were more than three times as likely to die of cardiac causes than were otherwise comparable patients who were not depressed. Again, these effects were independent of medical risk factors.

As noted, much of the evidence that depression has adverse health effects bear on cardiac disease. Some data also hint that depression may hasten HIV disease progression, though other findings have failed to confirm this (see Scheier and Bridges 1995).

Is depression a risk factor for premature death more generally? Wulsin et al. (1999) reviewed a wide range of studies in an effort to address that question. They concluded that the evidence for mortality effects is most compelling for cardiac disease, in line with studies described earlier. They also concluded that most research bearing on the question failed to include important control variables, such as degree of physical illness at baseline, smoking status, and alcohol abuse. All of these are known correlates of depression and potential predictors of premature death. This obviously is an area of investigation in which more information is needed.

It is tempting to view the evidence on depression and premature death as reflecting physical manifestations of a loss of desire to continue living. Perhaps the depressed MI patients in the studies reviewed above simply gave up the struggle to go on. This would represent a kind of disengagement response, albeit one that is quite different from the overt actions represented by suicide or alcohol abuse.

3. *Potential Moderators*

The focus of this article has been on the direct associations between health and the psychological variables of depression, hopelessness, and pessimism. However, some attention should also be paid to limitations on—or moderators of—such effects. One limitation may be extent of disease. It is possible that the effects of the psychological variables under discussion (and others) are greater among persons with disease that is less severe than among persons whose disease is advanced. That is, perhaps advancing disease affords less opportunity for psychological variables to gain leverage over the impact of the disease.

Another potential moderator is the age of the person whose health is under study. In their review of the literature, Scheier and Bridges (1995) noted that several studies had found that variables such as personality and depression had stronger influences on health among younger than older persons. For example, Schulz et al. (1996) found that pessimism predicted increased mortality among a group of recurrent cancer patients, but only among patients under age 60.

Indeed, every age difference that Scheier and Bridges (1995) encountered was one in which stronger associations between personality and health outcome occurred among younger than older persons. This observation led Scheier and Bridges to suggest that biological factors may become more important in determining disease outcomes as people grow older, and psychological factors correspondingly less important. This is another issue that should receive greater attention in the years to come.

4. Closing Comment

This article reviewed some associations that have been observed between pessimism, hopelessness, depression, and subsequent health outcomes. Evidence of associations between these psychological variables and health is fairly robust. On the other hand, it would be wrong to conclude that the mechanisms underlying these effects are well understood. In some cases the mechanism is obvious: persisting in a health-promoting exercise activity, taking steps to control one's weight, acting to lower one's health risks—all these are behavioral ways in which optimism and hopefulness can enhance health outcomes through overt behavior. In other cases, though, the mechanism is far more speculative, involving potential neuroendocrine and immune responses as mediators. The role of such mediators is presently under active investigation, but there remain many uncertainties. Despite the many questions remaining, though, the pattern seems consistent with a picture in which continued optimistic engagement with life goals promotes better health outcomes, and in which the loss of such engagement—giving up—promotes poorer outcomes.

See also: Control Beliefs: Health Perspectives; Explanatory Style and Health; Personality and Health; Self-efficacy and Health; Vulnerability and Perceived Susceptibility, Psychology of

Bibliography

Antoni M H, Goodkin K 1988 Host moderator variables in the promotion of cervical cancer. I: Personality facets. *Journal of Psychosomatic Research* **32**: 327–38

Aspinwall L G, Brunhart S N 1996 Distinguishing optimism from denial: Optimistic beliefs predict attention to health threats. *Personality and Social Psychology Bulletin* **22**: 993–1003

Beck A T, Steer R A, Kovacs M, Garrison B 1985 Hopelessness and eventual suicide: A 10-year prospective study of patients hospitalized with suicidal ideation. *American Journal of Psychiatry* **142**: 559–63

Carver C S, Scheier M F 1998 *On the Self-Regulation of Behavior*. Cambridge University Press, New York

Carver C S, Scheier M F 1999 Optimism. In: Snyder C R (ed.) *Coping: The Psychology of What Works*. Oxford University Press, New York

Everson S A, Goldberg D E, Kaplan G A, Cohen R D, Pukkala E, Tuomilehto J, Salonen J T 1996 Hoplessness and risk of mortality and incidence of myocardial infarction and cancer. *Psychosomatic Medicine* **58**: 113–21

Fawcett J, Scheftner W, Clark D, Hedeker D, Gibbons R, Coryell W 1987 Clinical predictors of suicide in patients with major affective disorders: A controlled prospective study. *American Journal of Psychiatry* **144**: 35–40

Fawzy F I, Fawzy N W, Hyun C S, Elashoff R, Guthrie D, Fahey J L, Morton D L 1993 Malignant melanoma: Effects of an early structured psychiatric intervention, coping, and affective state on recurrence and survival 6 years later. *Archives of General Psychiatry* **50**: 681–89

Frasure-Smith N, Lespérance F, Juneau M, Talajic M, Bourassa M G 1999 Gender, depression, and one-year prognosis after myocardial infarction. *Psychosomatic Medicine* **61**: 26–37

Frasure-Smith N, Lespérance F, Talajic M 1993 Depression following myocardial infarction. *Journal of the American Medical Association* **270**: 1819–25

Friedman L C, Weinberg A D, Webb J A, Cooper H P, Bruce S 1995 Skin cancer prevention and early detection intentions and behavior. *American Journal of Preventive Medicine* **11**: 59–65

Greer S, Morris T, Pettingale K W 1979 Psychological response to breast cancer: Effect on outcome. *Lancet* **ii**: 785–87

Greer S, Morris T, Pettingale K W, Haybittle J L 1990 Psychological response to breast cancer and 15-year outcome. *Lancet* **i**: 49–50

Ironson G, Friedman A, Klimas N, Antoni M, Fletcher M A, LaPerriere A, Simoneau J, Schneiderman N 1994 Distress, denial, and low adherence to behavioral interventions predict faster disease progression in gay men infected with human immunodeficiency virus. *International Journal of Behavioral Medicine* **1**: 90–105

Ladwig K H, Kieser M, Konig J 1991 Affective disorders and survival after acute myocardial infarction. *European Heart Journal* **12**: 959–64

Leedham B, Meyerowitz B E, Muirhead J, Frist W H 1995 Positive expectations predict health after heart transplantation. *Health Psychology* **14**: 74–9

Ohannessian C M, Hesselbrock V M, Tennen H, Affleck G 1993 Hassles and uplifts and generalized outcome expectancies as moderators on the relation between a family history of alcoholism and drinking behaviors. *Journal of Studies on Alcohol* **55**: 754–63

Petrie K, Chamberlain K 1983 Hopelessness and social desirability as moderator variables in predicting suicidal behavior. *Journal of Consulting and Clinical Psychology* **51**: 485–87

Pettingale K W, Morris T, Greer S 1985 Mental attitudes to cancer: An additional prognostic factor. *Lancet* **i**: 750

Räikkönen K, Matthews K A, Flory J D, Owens J F, Gump B B 1999 Effects of optimism, pessimism, and trait anxiety on ambulatory blood pressure and mood during everyday life. *Journal of Personality and Social Psychology* **76**: 104–13

Reed G M, Kemeny M E, Taylor S E, Visscher B R 1999 Negative HIV-specific expectancies and AIDS-related bereavement as predictors of symptom onset in asymptomatic HIV-positive gay men. *Health Psychology* **18**: 354–63

Reed G M, Kemeny M E, Taylor S E, Wang H-Y J, Visscher B R 1994 Realistic acceptance as a predictor of decreased survival time in gay men with AIDS. *Health Psychology* **13**: 299–307

Robbins A S, Spence J T, Clark H 1991 Psychological determinants of health and performance: The tangled web of desirable and undesirable characteristics. *Journal of Personality and Social Psychology* **61**: 755–65

Scheier M F, Bridges M W 1995 Person variables and health: Personality predispositions and acute psychological states as shared determinants for disease. *Psychosomatic Medicine* **57**: 255–68

Scheier M F, Carver C S 1992 Effects of optimism on psychological and physical well-being: Theoretical overview and empirical update. *Cognitive Therapy and Research* **16**: 201–28

Scheier M F, Matthews K A, Owens J F, Magovern G J, Lefebvre R C, Abbott R A, Carver C S 1989 Dispositional optimism and recovery from coronary artery bypass surgery: The beneficial effects on physical and psychological well-being. *Journal of Personality and Social Psychology* **57**: 1024–40

Scheier M F, Matthews K A, Owens J F, Schulz R, Bridges M W, Magovern G J Sr., Carver C S 1999 Optimism and rehospitalization following coronary artery bypass graft surgery. *Archives of Internal Medicine* **159**: 829–35

Schulz R, Bookwala J, Knapp J E, Scheier M F, Williamson G M 1996 Pessimism, age, and cancer mortality. *Psychology and Aging* **11**: 304–09

Shepperd J A, Maroto J J, Pbert L A 1996 Dispositional optimism as a predictor of health changes among cardiac patients. *Journal of Research in Personality* **30**: 517–34

Steptoe A, Wardle J, Vinck J, Tuomisto M, Holte A, Wichstrøm L 1994 Personality and attitudinal correlates of healthy lifestyles in young adults. *Psychology and Health* **9**: 331–43

Strack S, Carver C S, Blaney P H 1987 Predicting successful completion of an aftercare program following treatment for alcoholism: The role of dispositional optimism. *Journal of Personality and Social Psychology* **53**: 579–84

Taylor S E, Kemeny M E, Aspinwall L G, Schneider S G, Rodriguez R, Herbert M 1992 Optimism, coping, psychological distress, and high-risk sexual behavior among men at risk for Acquired Immunodeficiency Syndrome (AIDS). *Journal of Personality and Social Psychology* **63**: 460–73

Wulsin L R, Vaillant G E, Wells V E 1999 A systematic review of the mortality of depression. *Psychosomatic Medicine* **61**: 6–17

C. S. Carver

Deprivation: Relative

A person is relatively deprived if he or she does not enjoy material or symbolic goods compared with the norm or with other people that enjoy them. Relative deprivation and the notions that are directly or indirectly linked to it constitute a strongly structured conceptual set. The related propositions make up what Merton terms a 'middle range theory,' that is, a set of propositions which combines hypotheses and empirical regularities proper to a specific domain of validity and without which these regularities would be neither accounted for, nor cemented.

Even if the concept of relative deprivation was first used by Stouffer et al. (1949) in *The American Soldier*, this idea can be found in the writings of several classical sociologists such as Tocqueville (1852), Durkheim (1893, 1897), and Weber (1904). It also features in the works of contemporary non-sociologists such as Duesenberry (1949) and Easterlin (1973). These authors underscore the importance of the notion of relative income compared to absolute income and thereby that of the comparison between two statuses of the same individual at two different time intervals. This accounts for certain paradoxical phenomena that the standard economic theory is unable to explain. Starting with the well-known study of Stouffer et al. (1949), the theory of relative deprivation has been gradually developed, particularly by Merton and Kitt (1950), Runciman (1966), Boudon (1977), and Hirschman (1981). Its application expands to various areas, mostly to those where the relations between individual contributions and retributions are crucial, just like the problems of justice and equity or revolution. This theory is also interpreted from different standpoints that are discussed later.

1. *Themes and Theory of Deprivation*

Relative deprivation is often used in *The American Soldier* to account for puzzling results whose exhaustive study was conducted by Merton and Kitt (1950). One of these results is that the most favored groups seem to be inclined to criticize the system of promotion. Thus, Military Police soldiers belonging to a unit where promotion is scarce are satisfied with the promotion system governing their military career. In contrast, men in the Air Corps, although belonging to a body where promotion is frequent, are dissatisfied with their opportunities for promotion. To explain this paradox, Stouffer et al. (1949) introduced the notions of relative deprivation and reference group: 'if at the same age, level of education, and seniority, the majority of my colleagues and I have been promoted, I will consider my promotion as normal. If, however, I have not been promoted, I will go through a feeling of injustice, which can translate into criticism of the mobility system. If the promotion is relatively scarce in the group to which I belong or refer to in my evaluations, I will be very pleased with my promotion. However, if I am not promoted I will not experience any sense of injustice since all my peers have not been promoted.' Deprivation therefore occurs only under specific conditions:

(a) The individual retains some knowledge of the system of social promotion, mobility or retributions.

(b) He or she does not get an expected good, for example promotion.

(c) This expectation is based on an evaluation and a comparison that the individual conducts between his or her position, contributions and retributions on the one hand, and those of a large number of his or her peers, who make up his or her reference group on the other hand. This condition expresses *inter alia* the principle of interdependence among social actors.

(d) He or she has a feeling of inequity, which translates into criticism of the mobility system.

Thanks to particular models illustrating the idea of deprivation, several heterogeneous paradoxical phenomena have been accounted for by the founders of sociology as well as by contemporary sociologists. In *The Old Regime and the Revolution*, Tocqueville (1852) underscores the following surprising positive ecological correlation: the higher the economic prosperity, the quicker the laws of the old regime disappear and political liberalization gains ground in certain regions, the stronger the support of these regions to the Revolution. He points out that:

> While prosperity (...) in France is growing, general attitudes seem to be less content and more anxious; public dissatisfaction is turning sour; hatred for all the old institutions is increasing. The nation is clearly heading towards a revolution. Moreover, the parts of France which should have been the major sources of this revolution are precisely those parts where progress is being shown most .... It is not always by going from bad to worse that we fall into revolution.

Durkheim (1893, 1897) also develops two partly complementary ideas linked to deprivation. In the *Division of Labor*, he maintains that happiness is not an increasing monotonic function of wealth. The relation between satisfaction and abundance of wealth is a parabolic function with a maximum corresponding, according to him, to the happy medium, to the *aurea mediocritas* of the common saying: 'happiness lies between the two extremes of wealth.' Besides, the French sociologist devotes the chapter on 'the division of constrained labor' of the same book to the analysis of the relation between deprivation resulting from the violation of meritocratic rules and the questioning of social order. Equal opportunities and congruence between distribution of talents and jobs are, in his view, the conditions of legitimacy and stability of social order. In *Suicide*, it is the pure model of anomie that expresses deprivation. When aspirations and expectations grow faster than the means of realization and when these are not set back or brought to an end, a state of anomie occurs with numerous individual and social consequences. The absence of social regulation is a source of social pathology. With no clear limits to the legitimate needs, and free from any social pressure, men like Goethe's Faust will experience the infinite dimension of desire, frustration, and torment and can go as far as committing suicide.

Hirschman (1981), as an example of contemporary authors, explains tolerance to inequality in the developing countries by his 'tunnel effect' which is directly tied to the theory of deprivation and reference group. Let us take the example of a car driver in a two-lane tunnel with two lanes going in the same direction. Suppose that, after a short period of time during which no car in either lane has moved forward, the cars on the other lane start moving. He will feel relieved and hopeful that the cars on his lane will also move forward. However, upon realizing that only the cars on the other lane keep moving, he will feel disappointed and later furious. He might end up tempted to readjust the inflicted injustice by switching to the other lane, thus violating traffic regulations. The same applies to inequalities. An improvement of the social positions of my neighbors or my peers gives me hope. In case of disappointment, I will feel deprived and try to change the order of things.

2. *Generating Mechanisms of Deprivation*

One can wonder, following Boudon (1977), if all these phenomena are not likely to be generated by social mechanisms. He sets up simple models stemming from game theory, which bring about and explain relative deprivation. In his view, this deprivation can be interpreted as the product of the structures of interaction in which individuals end up. Deprivation entails the notion of competition for good and that of objective opportunities for promotion (Stouffer et al. 1949).

Let us suppose a group of N players who can win the lot $B1$ for a stake $C1$ or the lot $B2$ for the stake $C2$ with:

$$B1 > C1,\ B2 > C2,\ B1 > B2,\ C1 > C2,\ B1 \text{ and } C1 > 0,\ B2 \text{ and } C2 > 0$$

$N1$ and $n2$ are the number of winners of the lots $B1$ and $B2$ respectively with $n1+n2 = N$.

Being rational, the players will bet on $C1$ rather than $C2$ if the expected net profit by betting on $C1$ is superior to that corresponding to $C2$.

If $x1$ players ($n1 < x1 < N$) bet on $C1$, the expectation of net profit $E1$ ($x1$) of a player betting on $C1$ is

$$\begin{aligned} E1\ (x1) &= (B1-C1)\ n1/x1 \\ &\quad +(B2-C2)\ (x1-n1)/x1 \\ &= (B1-B2)\ n1/x1+B2-C1 \end{aligned}$$

On the other hand, the expectation of net profit $E2$ of a player betting on $C2$, when $x1$ ($> n1$) players bet on $C1$ is,

$$E2\ (x1) = B2 - C2$$

Thus, $x1(> n1)$ players bet on $C1$, a player is advised to bet on $C1$ rather $C2$ if

$$E1\ (x1) > E2\ (x1)$$

Or

$$(B1 - B2)\ n1/x1 + B2 - C1 > B2 - C2$$

i.e.,

$$(B1 - B2)/(C1 - C2) > x1/n1$$

One can analyze the relationship between the game structure and deprivation which is the proportion of individuals betting on $C1$ (the highest price) but winning only $B2$ (the lowest price). For that matter, one should examine the different values of the ratio $(B1 - B2)/(C1 - C2)$ that is the ratio of the possible additional gain and the additional cost on one hand, the percentages of players $100\ x1/N$ and of the winners $100\ n1/N$ on the other hand.

The simulation shows how complex the nature of relations is between deprivation and the objective opportunities of promotion. Let us consider the following general cases:

(a) The rate of deprivation increases with the values $(B1 - B2)/(C1 - C2)$. This rate increases with the marginal rate of profitability or return on investment $\{(B - B2) - (C1 - C2)\}/(C1 - C2)$ as shown by Kosaka (1986) who systematizes the propositions of Boudon (1977) and deduces new ones, and makes 'critical points' emerge.

(b) The highest levels of deprivation correspond to cases where a small number of players win the lot $B1$, the highest one.

(c) Deprivation rate is moderated as long as the moderated value lots are distributed to players in a limited number (the case of the Military Police in Stouffer's study). This is no longer the case when the number of winners increases (the case of men of the Air Corps).

(d) We get equal deprivation rates when only a small number of players win important lots or when a large number of players win less important ones.

(e) In general, the curves corresponding to identical relative deprivation levels are complex with different plateaux (ceilings) and critical points.

The simulation confirms the idea according to which the game structure determines the deprived players and their number. It is equally possible to simulate the situations in which the players are not equal, that is when the group is stratified and each social class has its different resources and behaviors toward risk. These simple game models particularly show the conditions in which the mere fact of increasing everyone's promotion opportunities can result in further increasing promotion expectations. These models also demonstrate that the propositions of Tocqueville, Durkheim, Stouffer are not unconditionally valid: when equal opportunities increase among individuals, deprivation tends to grow under certain conditions; but it can decrease under others. The trend of the relation between promotion or social mobility depends on the structure of interdependence among individuals.

3. *Deprivation and Collective Violence: A Case Study*

Revolution theorists generally agree on the assumption that revolutions occur during a period of economic growth and political liberalization. This paradox seems to cast doubt on the hypothesis of the rationality of human behavior. One can wonder why people revolt at the very moment of prosperity. Deprivation theorists' respond that an improvement in objective living conditions tends to generate higher expectations. As the expectations grow faster than the improvement in objective conditions, a wider gap comes up and translates into a higher degree of deprivation. This model is similar to the Durkheimian theory of anomie, as shown by Besnard (1987) and Coleman (1990). The revolution theory developed particularly by Davies (1962), and Gurr (1968) is different. In their view, revolutions occur when a long period of socioeconomic prosperity is immediately followed by a sudden downturn, that is a sharp decrease in gratification accompanied by a constant increase in expectations. The rapidly widening gap between expectations and gratification generates revolution. This proposition commonly translates in an inverted J curve.

Even though the notion of deprivation in this theory is not as precisely and explicitly defined as with Stouffer, it is easy to demonstrate that it also supposes, a meaningful difference deems unacceptable between the expectation and the realization at the individual level and an interpersonal comparison. First, with regard to intra-personal comparison, the difference is at the same time relative and absolute. It is absolute if at the $t1$ moment we compare expectation with realization. It is relative if at the $t2$ moment, the difference between expectation and realization deems unacceptable or unbearable compared to what it was at the $t1$ moment. Besides, individuals with threatened social positions or subject to downward social mobility compare themselves to those with stable positions or upward mobility.

Several varieties of deprivations can be constructed. The first one relies on the theory of relative deprivation

and shows up in the works of Gurr (1968) particularly for whom 'a psychological variable, relative deprivation, is the basic precondition for civil strife of any kind, and that the more widespread and intense deprivation is among members of a population, the greater the magnitude of strife in one or another form.' The relation between deprivation and civil strife is however mediated by societal variables such as coercive potential that can deter strife, institutionalization (social and political association may provide nonviolent means for expressing discontent), facilitation (i.e., means that facilitate civil strife) and finally legitimacy of the regime.

The second variety is that of the status inconsistency that can be found in the founding father's works such as those of Tocqueville (1852), Weber (1904), and in such contemporary research as that of Lenski (1954). The rich bourgeois and the well-off farmer who, thanks to the liberalization of the rules governing the acquisition of land, end up as owners of vast lands, remain for the French aristocrat mere plebeians carrying visible stigma of their original social class. If, as Tocqueville states, the bourgeois and the aristocrats become unknown to each other, it is mainly because they evaluate each other according to values located on different scales. Whereas the bourgeoisie has economic quasi monopoly and looks forward to assuming political power and enjoying honors, the bourgeoisie can not gain access to these institutions because the noblesse get in the way. The bourgeoisie feels compelled to opt for the revolutionary solution. Conversely, in the nineteenth century, as Weber points out, while the Prussian Junker maintains his monopoly over the army and the bureaucracy, the industrialists buy lands to acquire nobility titles and gain access to political positions. Politically, as Weber states, this type of industrialist is reactionary. These two examples illustrate two special cases of the multiple effects of the status inconsistency whose complete taxonomy Cherkaoui (1995) attempts to construct.

The collective action theorists, especially those of mobilization, question the theory of revolution based on deprivation. As revolution is a collective action seeking to realize a public interest, the rational actor will face problems specific to this type of action, particularly the temptation of free-riding. Following Olson (1965), it is believed that the rational individual belonging to a latent group as a social class seeking to satisfy collective goals would rather turn to see the sacrifices accepted by the others. He would not be interested in participating in a revolutionary enterprise whose costs can be heavy. Eventually, as pointed out by several analysts of revolutionary movements as Cochin (1924) and Brinton (1965), the revolutions are the product of some élites of organized and almost unanimous conspirers toppling weak governments at periods of social disorganization. For the theorists of mobilization and social movements as Oberschall (1973), and Tilly (1978), a psychological explanation of the revolutionary phenomenon is unable to account for the shift from disenchantment to the mobilization against authorities and in particular against the state. Besides, the theory of rational choice reduces the moral and solidarity incentives, which play a key role in the emergence of collective action. This is sometimes a moral obligation for the individual who belongs to collectivities where solidarity bonds are predominant. From this perspective, it is rather the lack of participation that will be costly. Oberschall shows that the lack of bonds and the dismantling of social networks are major setbacks to mobilization.

As can be seen by trying to turn the theory of relative deprivation into a maid of all work, we put it through risks that hang over any theory whose domain of application has been excessively expanded. But despite some criticism, its underlying principles remain solid. To mention but one, the principle of interdependence is the center of its axiomatic structure that the formalization put forward by Boudon has clearly highlighted. To explain and understand, *inter alia*, the inflationary spiral of demands, the perception of inequality, the individual conception of justice and of equity, or the degree of satisfaction, the theory of relative deprivation reminds us of an obvious fact that the sociological tradition has taken for granted since the days of Montesquieu whereby the structures of interdependence between social actors are often decisive.

See also: Action, Collective; Conflict Sociology; Durkheim, Emile (1858–1917); Game Theory: Noncooperative Games; Inequality; Inequality: Comparative Aspects; Revolutions, Sociology of; Revolutions, Theories of; Social Movements: Resource Mobilization Theory; Social Movements, Sociology of; Status and Role, Social Psychology of; Status and Role: Structural Aspects; Tocqueville, Alexis de (1805–59); Violence: Public

Bibliography

Besnard Ph 1987 *L'Anomie*, 1st edn. Presses Universitaires de France, Paris

Boudon R 1977 *Effets Pervers et Ordre Social*. Presses Universitaires de France, Paris [*The Unintended Consequences of Social Action*. Macmillan, London, 1982]

Brinton C 1965 *The Anatomy of Revolution*, rev. and expanded edn. Vintage Books, New York

Cherkaoui M 1995 Mobilité sociale et comportement politique: taxinomie des modèles de relations, *Revue Française de Sociologie* **36**: 171–83

Cochin A 1924 *La Révolution et la Libre Pensée*. Copernic, Paris, 1979

Coleman J S 1990 *Foundations of Social Theory*. The Belknap Press of Harvard University Press, Cambridge, MA

Davies J C 1962 Towards a theory of revolution. *American Sociological Review* **27**: 5–19

Duesenberry J S 1949 *Income, Saving and the Theory of Consumer Behavior*. Harvard University Press, Cambridge, MA
Durkheim E 1893 *De la Division du Travail Social*. Alcan, Paris [*The Division of Labor in Society*, Free Press, New York, 1947]
Durkheim E 1897 *Le Suicide*. Alcan, Paris [*Suicide*, Free Press, New York, 1951]
Easterlin R 1973 Relative economic status and the American fertility swing In: Sheldon E B *Family Economic Behavior*. Lippincott Philadelphia
Gurr T 1968 A causal model of civil strife. *American Political Science Review*, **62**: 1104–24
Hirschman A O 1981 *Essays in Trespassing: Economics to Politics and Beyond*. Cambridge University Press, Cambridge, UK
Kosaka K 1986 A model of relative deprivation, *Journal of Mathematical Sociology*, **12**(1): 35–48
Lenski G 1954 Status crystallization: a non vertical dimension of social status. *American Sociological Review* **19**: 405–13
Merton R K, Kitt A S 1950 Contributions to the theory of reference group behavior. In: Merton R K (ed.) 1968 *Social Theory and Social Structure*. Free Press of Glencoe, New York
Oberschall A 1973 *Social Conflict and Social Movements*. Prentice-Hall, Englewood Cliffs
Olson Jr M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
Runciman W G 1966 *Relative Deprivation and Social Justice*. Routledge and K Paul, London
Sheldon E B 1973 *Family Economic Behavior*. Lippincott, Philadelphia
Stouffer S A, Suchman A, DeVinney L C, Star S A 1949 *The American Soldier*. Princeton University Press, Princeton, NJ
Tilly C 1978 *From Mobilization to Revolution*, Addison Wesley, Reading, MA
de Tocqueville A 1852 *L'Ancien régime et la Révolution*, in *Œuvres Complètes*. Edition Gallimard, Paris [*The Old Regime and The Revolution*]
Weber M 1904 *Agrarstatistische und sozialpolitische Betrachtungen zur Fiedeikommissfrage in Preussen*. In: Weber M (ed.) *Gesammelte Aufsätze zur Soziologie und Sozialpolitik*. JCB Mohr, Tübingen, Germany
Weber M 1924 *Gesammelte Aufsätze zur Soziologie und Sozialpolitik*, JCB Mohr, Tübingen, Germany

M. Cherkaoui

Desertification

1. Introduction

Despite considerable uncertainty about its meaning, the importance of desertification for the peoples of the world's drylands should not be underestimated. First, the term, for all its problems, envelops a set of processes, sometimes listed as drought, desiccation and degradation, which pose very real and severe challenges to drylands, and have strong knock-on effects on society. Second, desertification language is deeply entrenched in policy circles and in the current UNCCD (United Nations Convention to Combat Desertification), and will inevitably continue to inform the financing of development interventions in drylands.

1.1 History and Definitions

The 'Agenda 21' document of the 1992 United Nations Conference on Environment and Development (UNCED) defined desertification as 'land degradation in arid, semi-arid, and dry, subhumid areas resulting from various factors, including climatic variations and human activities.' The processes of degradation were said to be soil erosion, nutrient depletion, crust formation, salinization, reduction in pasture and agricultural productivity, loss of biodiversity, and reduction in vegetation cover in susceptible drylands. Of course, uncertainties abound, the circumstances fostering reversible or irreversible change, the resiliency of different semi-arid environments, the different roles of natural and anthropogenic driving forces, and the synergies among these elements.

Scientists were exercised as early as the 1920s with the 'the advance and spread of the desert' in Africa. It was then that France began the first studies of the process in West Africa, finding evidence of mobile sand dunes, human fossil remains in now-dry areas, and declining annual rainfall. It was concluded that the Sahara had grown, and was still growing, owing to poor land management, which had worsened under the colonial regime.

These early judgments preceded what environmental historians are now discovering to have been a much more traumatic event, the American 'Dust Bowl' of the 1930s. Sears' *Deserts on the March* (1949) evoked accusations of widespread anthropogenic degradation. He, and others, believed they had evidence of widespread environmental degradation in the drylands of the Old World, whose cause, as in the mid-western states, was also thought to be mismanagement. With these narratives very much in mind, E. P. Stebbing, a forester, identified the causes of degradation in British West Africa to be shortened agricultural fallow periods, shifting agriculture, and overgrazing (Stebbing 1935). An Anglo–French Forestry Commission toured the Niger–Nigeria border in 1936–7, and their conclusions were far more circumspect. They saw degradation as place-specific and treatable. Yet a member of this Commission, the influential French botanist Auguste Aubréville, held to the 'desert advance' hypothesis, and first used the term 'desertification' in 1949 (Warren 1996).

In dryland Africa, severe drought and famine in the 1970s, following decades of good rainfall, again revived the desertification debate, with depressing reports of desert advance appearing in scientific publications and the media. This alarm stoked the

United Nations Conference on Desertification (UNCOD), held in Nairobi in 1977. UNCOD, and the National Plans of Action agreed to at that meeting, still viewed people as the main agents of desertification.

1.2 Current Research

Research into the links between climate, management and degradation has taught us a great deal since the 1970s. Since then, many detailed studies of human—environment interactions on the local scale have been carried out in the drylands, spurred on by advances in climatology, soil science, botany, geoinformatics, agronomy, and the social sciences.

This research, conducted by numerous scholars from the natural and social sciences, offers a longer-term view of what happens as rainfall fluctuates, or when soils are lost and gained. More generally it has examined the complexity of the linkage between resources and people in the drylands, and the role of adaptive local resource management. Above all, it is now possible to challenge notions about an equilibrium state or 'carrying capacity' in these environments and from there to challenge the desertification narrative with better, if still imperfect evidence (see also *Human–Environment Relationship: Carrying Capacity*). Yet laying the desertification discourse to rest is difficult. At least six themes remain essential elements of the desertification debate.

1.2.1 Advance of desert. Large-scale Saharan expansion has not been proven within the period of the historical record. Desert advance in the Sudan has been asserted since the 1930s, but Helldén (1991) found no such evidence from satellite imagery, settlement histories, sand movements, and degradation conditions around boreholes. The current view is that the Sahara expands and contracts periodically as rainfall varies. Desert advance (as in the severe drought of 1984), and localized degradation, is usually short-term (Tucker et al. 1991, Nicholson et al. 1998).

1.2.2 Resilience of dryland ecosystems and land-use systems. Degradation is, therefore, usually localized and ephemeral (see also *Land Degradation*), largely a function of the complexity of relationships that exist between humans and natural systems. Semi-arid ecosystems, far from being fragile, exist in a range of semi-permanent 'states' dictated by disturbance, drought, fire or insect attack; they are well adapted to these forces. Similarly many indigenous management systems redistribute nutrients, plants, grazing pressure and water to create 'enriched' patches in the landscape (see also *Ecology, Cultural*).

1.2.3 Influence of grazing and livestock. The UNEP *Atlas of Desertification* (1992) asserted that 58 percent of soil erosion in dryland Africa was the result of overgrazing by livestock, lending support to Garret Hardin's 'tragedy of the commons' hypothesis (see also *Environmental and Resource Management*). Such figures gave way to more conservative estimates in the latest version of this *Atlas* (Middleton and Thomas, 1997). There is now greater confidence about the heterogeneity and patch dynamics of pastoralist grazing strategies so critical to many drylands. Less vegetation is not necessarily worse vegetation. It is difficult to overgraze in a dynamic nonequilibrium system, dominated by annual grasses, where the external forces such as drought are more powerful than the internal ones such as animal numbers.

1.2.4 Effects of increased population in rainfed, dryland agricultural systems. No necessary Malthusian link exists between elevated population levels and resource degradation. Indeed, because there is often an incentive for rural people to invest in antidegradation measures in these circumstances, more people (and labor) may mean less erosion, and the initiation of compensatory risk management strategies (see also *Agricultural Change Theory*; *Indigenous Knowledge and Technology*).

1.2.5 Soil erosion and fertility decline. Although topsoil erosion does occur in drylands and soil is blown or washed away, these processes do not themselves necessarily result in degraded landscapes. Soils are often replaceable or may be very deep, and can accumulate down-slope or down-wind where they may be more valuable to local people. Indigenous dryland soil conservation systems are impressive (see also *Indigenous Knowledge and Technology*). Yet even after decades of research, soil erosion and the values placed upon it by society are very difficult to measure.

1.2.6 Climate change. There are two issues here. Charney and Stone (1975) suggested that devegetation might induce regional drought—bare soils are cooler because they reflect more solar energy and cool surfaces discourage rainfall. This hypothesis remains disputed. Current research provides strong evidence that changes in surface characteristics, including vegetation, affect local and regional rainfall patterns significantly. Some climatologists however, find little evidence in this claim as registered in the coarser scale climate record (Williams and Balling, 1996). In addition, potential global warming and the El Niño-Southern Oscillation will have differential, and as yet uncertain effects on certain drylands, with

the majority of current models suggesting increased rainfall and temperature variability in Africa (see also *Climate Impacts*).

2. Desertification Policies

Research, policymaking and interventions are strongly linked. For example, when scientists pinpointed advancing deserts, the policy response, was 'plant green belts'! Roosevelt commissioned a study of a massive belt between Texas and North Dakota after the Dust Bowl. Stebbing (1935) recommended much the same for Africa, and UNCOD proposed a circum-Sahara green belt. The Algerian Government actually began planting trees, employing conscripts to do so. Where scientists believed that herdsmen or farmers were irrationally causing degradation, policy-makers and government officials prohibited goats, tree cutting or grass burning, and destocked herds, or, on a much more massive scale, they enforced soil conservation programs.

Antidesertification measures have been international in scope. A suite of international institutions materialized after the great Sahelian drought of the late 1960s and early 1970s. First, the UN Sudano-Sahelian Office (UNSO) was charged with combating drought, later taking on desertification issues. It was later joined by the UN Environment Programme and the FAO (Food and Agricultural Organisation) in these roles. Scientific panels and advisors produced a 'veritable sandstorm of literature' (Mortimore 1998) on the complexities of drylands. In general, the prognosis was gloomy: 10 percent of the earth's surface was human-made desert; and the process would continue. Desertification became enshrined in international policy.

African states absorbed some of UNCOD's recommendations from the late 1970s, although they were held back by poor funding. Projects were initiated, most designed to arrest soil erosion. Some were implemented without adequate local consultation or participation. Many, like the GERES project in Burkina Faso where bulldozers were used to create earthen bunds across many slopes, actually increased erosion rates. Fuelwood supply projects and the drastic destocking of indigenous pastures failed for the lack of local complicity.

Unforeseen ramifications followed in the post-UNCOD, post-Sahelian drought era. In Niger, proposals for green belts resorted to authoritarian measures for implementation, for in local eyes, as in the eyes of scientists today, they were deeply irrational. In similar vein the military rulers of Mali banned burning, counter to local know-how about the enriching powers of fire on pastures. In many countries, the cutting of fuelwood was unnecessarily proscribed (see also *Deforestation–Forestation*). Most disturbingly, the characterization of African pastures as 'overgrazed' came to infest writing about them. Attempts were thus made to reduce the numbers of livestock, to privatize grazing lands, and even to replace herding the commons with systems based on European and North American models (Warren 1996). The desertification threat was used to justify authoritarian control over natural resources and land-use practices.

A broadened, perhaps softened conception of desertification suggests a new range of policies that, given sensitive application, may prove more successful because they are more trusting of dryland populations and more aware of ecological diversity. New initiatives are based on a better understanding of the three main issues in dryland management: drought, desiccation and degradation. Drought is common currency among dryland folk. It may not be welcome, but people clearly have a portfolio of coping strategies to deal with it, and reflexive management that constantly adapts to an ever-changing situation (see also *Environmental Risk and Hazards*). These indigenous systems are now being supported, rather than replaced or obstructed. Desiccation is known to have been frequent in the history of human occupation of drylands. Large-scale shifts in livelihood practices occur under progressive aridity, but Sahelian drought illustrates that relief, restocking and resettlement are necessary when this happens. Degradation requires a range of responses. Rich repertoires of conserving indigenous techniques exist, some—like permeable contour bunds—already improved through hybrid experimentation between farmers and scientists. Drylands and dryland peoples and their economies are heterogeneous. Policies increasingly aim to protect this diversity, rather than erase it. Greater clarity of land tenure security is needed, and rangeland areas require special forms of collective management (see also *Environment and Common Property Institutions*). In brief, livelihood security is essential to good conservation.

3. The Convention to Combat Desertification

UNCED, held in Rio in 1992, breathed new life into the moribund desertification concept (see also *International and Transboundary Accords, Environmental*). Fortunately, Chapter 12 of UNCED's Agenda 21 document is far less alarmist in its tone than UNCOD's recommendations of 15 years earlier, adopting some of the findings and views of the new research reported above. For example, climate change is now seen as a major contributor to desertification, and drought preparedness, education, and strong local participation are now stressed.

The CCD came into force on December 26, 1996 (see http://www.unccd.ch/ or http://www.iisd.ca/linkages/desert/ for full details and current negotiations). Although there was skepticism about this new international effort, nongovernment organizations and other actors from developing countries had a

relatively strong involvement in drawing up the Convention. The various CCD components are legally binding on the parties, and include annexes for Africa, Asia, Latin America, the Caribbean, and the northern Mediterranean. Almost 160 countries had ratified the Convention by September 1999. The focus has moved to implementation, despite significant questions about who pays, coordinates, and monitors the activities. By late-1999 National Action Plans were drawn up by most dryland nations, but progress is hampered since there is no international CCD implementation fund available.

4. Conclusion

In the absence of well-adapted technologies, life in drylands is harsh. That degradation of these lands follows from social and climatic processes is not in dispute. That degradation amounting to desertification of the magnitude assumed in the early debates, or even in the recent Convention, has been strongly challenged by the research and other communities. Nonetheless, desertification remains a powerful narrative that has captured the imagination of certain policy makers, governments, and scientists, and it has persisted with peaks and troughs of interest since the 1930s. Its power is problematic, since 'the search for accuracy appears to be vulnerable to generalization, over-simplification and distortion for political or funding purposes' (Mortimore 1998). Desertification has certainly served specific constituencies quite well, including national governments seeking to extend their control, international bureaucracies seeking legitimacy, and scientists seeking research funding.

Tackling the various processes that the term desertification subsumes is a major challenge, given that dryland populations suffer the impacts of droughts and desiccation, and need to exploit land, water, and timber resources in ways which often result in degradation. As an explanatory term, however, desertification has been used too broadly and too loosely, and this has been counterproductive.

See also: Climate Impacts; Climate Policy: International; Environment and Development; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Economics; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; Sustainable Development; Water Pollution: Terrestrial and Groundwater; Water Resources

Bibliography

Charney J G, Stone P H 1975 Drought in the Sahara: a biogeophysical feedback mechanism. *Science* **187**: 434–45

Helldén U 1991 Desertification: time for an assessment. *Ambio* **20**: 372–83

Middleton N J, Thomas D S G (eds.) 1997 *World Atlas of Desertification*, 2nd edn. Arnold, London

Mortimore M 1998 *Roots in the African Dust: Sustaining the Sub-Saharan drylands*. Cambridge University Press, Cambridge, UK

Nicholson S E, Tucker C J, Ba M B 1998 Desertification, drought and surface vegetation: an example from the West African Sahel. *Bulletin of the American Meteorological Society* **79**: 815–29

Sears P B 1949 *Deserts on the March*. Routledge and Kegan Paul, London

Stebbing E P 1935 The encroaching Sahara: the threat to the West African colonies. *The Geographical Journal* **85**: 506–24

Swift J 1996 Desertification. In: Leach M, Mearns R (eds.) *The Lie of the Land: Challenging Received Wisdom on the African Environment*. International African Institute in association with James Currey, Oxford, UK, pp. 73–90

Tucker C J, Dregne H E, Newcomb W W 1991 Expansion and contraction of the Sahara Desert from 1980 to 1990. *Science* **253**: 299–301

UNEP 1992 *World Atlas of Desertification*. Arnold, London

Warren A 1996 Desertification. In: Adams W M, Goudie A, Orme A R (eds.) *The Physical Geography of Africa*. Oxford University Press, Oxford, UK, pp. 342–55

Williams M A J, Balling R C Jr 1996 *Interactions of Desertification and Climate*. Arnold, London

S. Batterbury and A. Warren

Deserts, Philosophical Concept of

In everyday discourse, the language of desert is prominent and pervasive. It is common to speak of certain conduct (or persons) as deserving of praise or blame, reward or punishment, and of goods or ills that are visited upon people as deserved or undeserved. The subjects of desert claims may vary: most commonly they are persons or person-related entities (such as paintings that deserve prizes or cities that deserve their reputations), but insofar as inanimate or natural objects (such as rock formations) are evaluable in some way, there may be no barrier to speaking of them also as having deserts of certain kinds. The objects of desert claims are almost always seen as goods or ills. *Inter alia* a person may deserve success or failure, a good or bad reputation, good or bad luck, a greater or lesser portion of some social resource, compensation, or fairness.

1. Desert and Merit

The terms 'desert' and 'merit' are often used interchangeably. Some writers, however, have insisted on their conceptual distinction. If, for example, the ancient tie between diligence and desert (*deservire*: to

serve zealously) is focused upon, the language of merit may be used to capture those outcomes warranted by other forms of value. Milne, for example, argues that

> a natural beauty spot is *worthy of*, or *merits*, or *rates* inclusion in a published guide to places of natural beauty, but it does not make sense to say that the inclusion in the guide is something *deserved* in the way of reward. The classification of human beauty (or intelligence, or skill) is a parallel case: the naturally beautiful person merits a certain rating, but he or she does not deserve it any more than Niagara Falls is *rewarded* by its rating as a tourist spectacle. Only insofar as effort is expended in producing the achievement or contribution, or in acquiring a skill, is there the basis of a desert-claim: a person cannot be deserving of praise or reward without trying (Milne 1986, p. 240).

The point is expressed generally by Pojman. Merit (or demerit), he claims, is the more inclusive term, referring to 'any feature or quality that is the basis for distributing positive [or negative] attribution,' whereas desert is a species of merit, 'typically or paradigmatically connected with action, since it rests on what we voluntarily do or produce' (Pojman 1999, p. 86). He then distinguishes a secondary, negative, compensatory or rectificatory use of desert, to accommodate cases in which—relatively faultlessly—a person suffers some evil and recompense is appropriately called for.

Although a distinction of sorts can thus be made, it is at best a soft one, since a great deal of current debate about desert (see below, Sect. 11) has concerned the enhancement of arbitrarily possessed advantages through diligent effort or the exacerbation of disadvantages, including some relating to effort.

2. *Preinstitutional and Institutional Deserts*

In recent philosophical discussion, a distinction has frequently been drawn between preinstitutional and institutional deserts—between deserts that are said to be 'naturally' implicated by virtue of some characteristic/activity of their subject, and those that are attributable only within the framework of some institutionalized arrangement. A person is said—preinstitutionally—to deserve his good fortune, or—institutionally—to have deserved to win a particular prize. The former is often said to be prior to the latter: 'desert is a "natural" moral notion (that is, one which is not logically tied to institutions, practices, and rules)' (Feinberg 1963, p. 70). In many of its attributions to humans, preinstitutional desert is tied to 'reactive attitudes—the sense we have of others' agency, of their sharing with us at the deepest human level' (Strawson 1962). And the rejection of preinstitutional desert is often associated with a deterministic naturalism (Scheffler 1992).

Those who contend that only institutional desert claims have a legitimate place in our understanding claim that talk of preinstitutional deserts assumes, contrary to fact, that the characteristics on the basis of which such claims are made are fairly distributed. But, such critics of preinstitutional desert argue, since people naturally and arbitrarily differ in the talents, capacities, and social environments that ground what preinstitutional desert claims would allocate to them, they cannot really be said to deserve such things, or their fruits (cf. Rawls 1971, pp. 104–6, 312–4).

Leaving these latter claims aside (though see Sect. 11), it may nevertheless turn out that the distinction between preinstitutional and institutional deserts is one of degree more than one of kind. The good weather which—it might be said—*A* deserves for her vacation makes sense in part because the social institution of providing vacations has a point (recreation) to which 'good weather' intelligibly relates. Where no social institution is in place, it is much more difficult to make a specific attribution of desert. Thus, given the institution of punishment, it is easier to argue that those who deserve it are wrongdoers, than to argue—*ab initio*—that what wrongdoers deserve is punishment. It may well be that doing good has positive deserts, and doing ill has negative deserts; but the specific character of those deserts, whether good or ill, may not be determinable independently of particular social institutions (that, admittedly, are themselves open to scrutiny). It is—in part—the difficulty in establishing what, preinstitutionally, people deserve that creates some of the moral space for contemporary debates about restorative justice. Certainly wrongdoing deserves some form of negative response: but should that response take the form of punishment (narrowly understood as fines, imprisonment), or instead as confrontation, blame, reparation, apology, penitence, and so forth?

3. *Desert and Entitlement*

But even if the distinction between preinstitutional and institutional deserts is soft, it should not lead to a conflation of desert with entitlement. '*A* deserves *p*' is not equivalent to '*A* is entitled to *p*.' Deserts seem to function comparatively or incrementally—as more or less; entitlements tend to be absolute—either we are or we are not entitled to something. Because of his talent and the effort he has put into training, *A* may be the most deserving when it comes to winning the race. But if he pulls a muscle and has to withdraw, he is not entitled to the prize, nor even to win. The purpose of a race may be to determine who is the fastest runner, and *A* may be the fastest runner and therefore deserve to win. But the rules of the race—which formalize the process of determining who is the fastest runner—will determine who is entitled to win. The entitlement rules may not always succeed in establishing the most deserving, and sometimes they may even serve to exclude the most deserving.

4. *Moral and Nonmoral Deserts*

Desert often functions as a moral notion, especially where fair distributions of social benefits and burdens are involved. But it need not do so. If a person deserves success, or deserves to win a prize, or deserves good weather, we need not be talking about moral desert. True, it may seem to align with a certain cosmic fairness, but this need not be moral in character. Even if deserts are paradigmly moral, they shade off into the nonmoral. In asserting that the Niagara Falls is deservedly famous or that the Mona Lisa deserves its reputation or even that a person's efforts deserved a better result, we are not making a moral claim. Desert is a normative notion, not necessarily a moral one.

5. *The Structure of Desert Claims*

It is generally agreed that desert claims have the implicit structure: '*A* deserves *p* in virtue of *x*,' where *A* is the subject of the desert claim, *p* its object, and *x* its basis.

For the most part, the subjects of desert claims will be persons/personal conduct (individual or collective) or people-related artifacts/constructs (such as art objects, cities, and so forth), although common parlance sometimes extends the language of desert to environmental and other natural objects (the sequoia, for example, may be said to deserve preservation or its reputation).

The range of deserved objects is vast: to those mentioned earlier may be added *inter alia* jobs, wages, grades, respect, wins, and losses. Linking all such objects is their perception as goods or ills, as benefits or burdens, as valued or disvalued. Although there may be marginal cases in which one deserves to be treated neutrally rather than well or ill (Sverdlik 1983, p. 320–1), such cases are parasitic on a context in which the distribution of goods and ills is primary. Sometimes—as in the case of *A*'s good weather—what seems to be deserved is a state of affairs rather than some mode of treatment by others or some achievement of one's own.

There is an ambiguity in the negative claim '*A* does not deserve *p*.' It may mean, '*A* does not deserve *p* (but deserves *q* instead)' or '*A* does not deserve *p* (or *q*).' In the former case we may speak of *A* being nondeserving, in the latter case of *A* being undeserving.

6. *Desert Bases*

Desert claims are more than expressions of personal assessment. They are subject to public scrutiny and criticism. In particular, some proportionality between the desert basis and deserved outcome is presupposed, though the principle of proportionality stands in need of clearer articulation than it is generally given. So, although we might be inclined to say that a person who was injured in an automobile accident—because he foolishly refused to use a safety belt—deserved the injuries he suffered (though not to be killed), we would be reluctant to say of a person who failed to wear a safety belt that he deserved to get injured (though we might want to say that if he gets injured he will deserve it).

Most writers on desert consider that desert bases must not lie in the future, but are instead constituted by some past behavior or characteristics already possessed by the deserving object (Kleinig 1971, p. 73, Sher 1987, pp. 150–93). Appeals to future consequences as bases for distributing benefits and burdens may sometimes be justified, but cannot be justified as deserts. Or, at least these consequences cannot be directly justified as deserts. As indirect desert bases, reference to future-oriented considerations may be more plausible. For example, it may be argued that though a person may deserve *p* only in virtue of his past diligence, diligence may be valued as a virtue (in part) because of its consequential/productive significance.

A few writers, however, have recently argued that a desert base might lie directly in the future (New 1992, 1995, Feldman 1995, p. 69–71). The examples used to sustain this thesis refer to some known future facts (say, death and suffering), on the basis of which its subjects are said to deserve some recognition now. It is not clear, however, that the examples show what they are intended to (Smilansky 1994): either there is some past or present commitment that is the basis for the desert or the language of desert is better replaced by that of distributional fairness.

7. *Locating Desert Bases*

It has been suggested above, and argued at length by others, that the nature of a desert base can be established only by going outside the concept of desert to external goals and values that inform the activity in relation to which *p* is said to be deserved (Cummisky 1987, Lamont 1994). That is, merely by looking at the concept of desert, we cannot tell—except in the most general terms—what a particular act or course of conduct deserves. Moreover, because the goals and values of the activities in question may differ, depending on the context, disagreements about deservedness can become intractable if this wider context is forgotten. Disagreements may also be chronically unresolvable if there is no clarity about the goals and values implicit in the activities or institutions in relation to which attributions of desert are made. In addition, of course, there may be a further debate about the appropriateness of the goals and values to the activities or institution in question, and that debate may also impact on a discussion of the appropriateness

of a particular desert basis or on the deservingness of those who are the recipients of its benefits or burdens.

Consider, for example, the institution of paying wages or a salary for work. It may not be determinable—given only a person's expenditure of productive effort outside the context of an employment contract—what she deserves: whether she should receive further opportunities, praise, good fortune, sustenance, free housing, or a wage/salary. But, given that we have in place the work-related institutions that we do, we might wish to argue that *A*'s work deserves—at a minimum—a living wage. If the work that *A* does is especially demanding and/or highly valued, we might want to say that it should be more highly rewarded. Even so—to simplify the illustration—we might view *A*'s work in at least two ways, and these may be considered separately or in combination. We may focus on the demandingness of the work, on its productivity, or on some combination of the two. If the former constitutes our focus, work that manifests great effort, even if minimally productive, will be seen as most deserving of remuneration (whatever we do about entitlements). If the latter is more significant for the particular institution, then productivity will provide the basis for remunerative desert, even if the achievement reflects native talent more than effort. Given that, for many social institutions, productivity and effort will tend to be correlated, we may decide to take both into account in determining what a person's remunerative deserts should be—though if we are not sure for how much each is to count, there may be interminable disputes about a person's deserts. In such cases, rules governing entitlements will enable the institutions to function.

Consider now the salaries earned by sports and pop stars. At one level, we might wish to claim that Michael Jordan and Michael Jackson deserve—in terms of the institutions within which they operated—the status they possess and the remuneration they have received. But it is also open to us to debate the social value our society has accorded to sporting and singing prowess and, if as a result we conclude that they are overvalued, we will also consider that they have received more than they deserved (albeit no less than that to which they were entitled).

Where, as in many human activities, effort and achievement tend to be highly valued as desert bases, it is sometimes asserted that the desert basis—whether it focuses on effort or achievement—must be something for which the deserving person is responsible or over which he or she must have a significant degree of control. So, for example, a person cannot plausibly be said to deserve punishment unless that in virtue of which punishment is said to be deserved is something for which the person can be held responsible.

It is, however, more likely that the issue of responsibility/control is to be settled by reference to the goals and values of the institution to which the desert is related rather than to some feature of desert itself. In the case of impositions such as punishment, the idea that they could be deserved without the triggering behavior being under the person's control would be deeply problematic. On the other hand, there is no great problem involved in thinking that those who deserve to win beauty contests are trading on genetic endowments more than on personal effort. And if inanimate objects—such as the Grand Canyon—can be said to be deservedly famous, the question of responsibility does not even arise.

There are, moreover, institutional contexts in which attributions of desert seem to require that one not have had control over one's circumstances. Consider, for example, deserved compensation. Here the deserving party is not usually seen as being in any way responsible for the misfortune in virtue of which he or she is said to deserve the compensation. Suggesting that the person who caused the misfortune is responsible puts the responsibility in the wrong place, and it is not adequately put in the right place by suggesting that the person who is deserving had to be nonresponsible for putting himself in the way of the misfortune. What is more, the compensation may be deserved even if the cause of the misfortune cannot be held responsible for bringing it about.

Thus the issue of responsibility too is determined by reference to the activity or institution in relation to which the desert claim is made. Prizes may be distributed to reward achievement or some combination of achievement and effort; wages may be given to reward productivity or some combination of productivity, talent, social importance, and effort. Some of these are more closely associated with personal control than others.

Among those who have wished to assert a tight relationship between desert and responsibility—or at least between desert and certain kinds of objects, such as punishments—several subsidiary discussions have ensued.

(a) One arises out of debates about personal identity. It has been argued (notably by Derek Parfit) that since our personal identity is constituted by our beliefs, a significant change in our beliefs will reconstitute our identity as other than it was, with implications for our deserts. Parfit considers the case of a person who, at ninety, after receiving the Nobel Peace Prize, confesses to injuring a police officer when he was twenty. Given the changes he has undergone, he no longer deserves punishment for what he did: 'just as someone's deserts correspond to the degree of complicity with some criminal, so his deserts now, for some past crime, correspond to the degree of psychological connectedness between himself now and himself when committing the crime' (Parfit 1984, p. 326). Although Parfit's account of identity has come in for criticism, his discussion draws attention to an important assumption about the subject of desert claims (though see Fields 1987).

(b) A problem has also surfaced in discussions of moral luck (Richards 1986; or, more generally, Rosebury 1995). If it is simply a matter of good or bad luck that certain consequences come about through *A*'s actions, should this bear on the matter of desert? Given that luck refers to something beyond one's control, it might seem as though it should not. Nevertheless, we often act as though it does. If *A* drives recklessly through the streets and—as a matter of bad luck—an innocent pedestrian happens to step out just as he passes, we will be inclined to see the injury or death as exacerbating the recklessness and making him more deserving of whatever penalty might have been due him. Richards disagrees. Deserts are, he thinks, reflective of character, and the character of an unlucky and lucky reckless driver is the same. He suggests that the difference is largely an epistemic one. We are to judge on the basis of what we have good reason to know about their deserts, and the person whose recklessness (or other failing) luckily results in no harm to others does not provide us with the epistemic basis for judging him as severely as the person who—unluckily, but not surprisingly—injures the other person. 'Luck … affects how we ought to treat him, not by changing what he deserves, but by changing the grounds on which we are obliged to judge' (Richards 1986, p. 200–1). 'To a very large extent people are lucky only in our verdicts, not in what they deserve, and their luck must affect our verdicts because it affects our grounds for reaching them' (Richards 1986, p. 206). Richards' position has not gone unchallenged: recklessness may not deserve what is normally attributed to the driver in the event of an accident (Adler 1987, p. 247); moreover, we may be in an epistemically better position to judge the lucky driver's character because of a record of his speeding tickets (Adler 1987, p. 247). Furthermore, Richards fails to give appropriate regard to the way in which the actual harm exacerbates the situation.

8. *Enumerating Desert Bases*

In the *Nicomachean Ethics*, Aristotle asserts that 'everyone agrees that in distributions the just share must be given on the basis of what one deserves, though not everyone would name the same criterion of deserving: democrats say it is free birth, oligarchies that it is wealth or noble birth, and aristocrats that it is excellence' (Aristotle 1962, pp. 118–19). For a modern reader, this well illustrates the dependence of desert bases on the goals and values of the activities or institutions within which desert attributions are made. In a world that is now characterized more noticeably by institutions and values that give greater weight—on the one hand—to individual responsibility and effort and—on the other—to achievement and social contribution, the bases for desert tend to be similarly characterized. As part of the tension between individual and collective values, a tension is also seen between desert bases that focus on individual responsibility, diligence, and effort and those that are more outcome directed.

8.1 *Effort/Diligence*

Writers who emphasize the importance of effort as a distributive desert base usually limit it in various ways. The effort must be purposeful, in the sense of being directed to the ends of the activity that grounds the desert; it must display a certain level of competence, in the sense of bearing some reasonable relation to achieving the ends of the activity in question; and the activity itself must be normatively acceptable: 'There is no question of rewarding murderers and rapists, no matter how much effort has been expended by them' (Milne 1986, p. 240).

If two people put in the same amount of effort, but one succeeds whereas the other one does not, then it might be said that they are equally deserving in virtue of their efforts. If, however, I offer a reward (or, more correctly, a prize) to the person who succeeds, then only the person who succeeds deserves it. Here the desert has been institutionalized. Our language is blurred between prizes and rewards, reflecting an overlap that sometimes occurs in competitive situations—but it should not lead us to the conclusion that achievement is 'the' proper basis for award claims.

8.2 *Achievement*

Those who focus on the importance of achievement draw attention to the inadequacy of effort unless it embodies competence—that is, is likely to be productive. If brilliant *A* can effortlessly produce the goods that diligent *B* cannot, then *A*, and not *B*, is likely to deserve whatever rewards or benefits are associated with that achievement. However, if it is sheer luck that results in the outcome for *A*, it is not so clear that *A* is more deserving than *B*. Even in cases in which the focus is on achievement, there seems to be some gesture in the direction of purposive effort.

9. *The Force of Desert Claims*

What follows from the claim *A* deserves *p*? Several possibilities have been canvassed: (a) *B* ought to ensure that *A* gets *p*; (b) *A* ought—all things considered—to get *p*; (c) *A* has a right to *p*; (d) There is a reason for giving *p* to *A*; and (e) *A* would be presumptively justified in getting *p*.

Most of these options are implausible. (a) It would be a rare desert claim that imposed a duty on some particular party. If *A* deserves to be punished, it does not follow that the state or some particular person should inflict it; if *A* deserves to succeed it is unlikely to fall to any particular person to ensure that she does so.

(b) Even if desert claims imply that someone ought to get something, they do not provide an overriding reason for it. Suppose, for example, that *A* deserves to be punished. There may, nevertheless, be countervailing—and overriding—reasons why *A* ought not to be punished: perhaps mercy ought to be shown, or the social consequences of punishment would be disastrous. The same would hold of good deserts: other considerations of social justice may take priority. (c) Is also much too strong. (d) Seems too weak—not just because there may be many other considerations (besides deserts) that constitute reasons for giving *p* to *A*—but because deserts seem to be quite strong reasons for giving *p* to *A*. (e) Represents most closely the force of a desert claim. Such claims are presumptively strong enough to justify *A*'s getting *p* even if, in a particular case, some countervailing consideration would defeat the presumption. In practice, the situation is more complicated. If, for example, *p* refers not to some preinstitutional generality, but legal punishment or a competitive prize, other conditions may be necessary if the *p* is to be justified. In such cases, a formalized desert-determining procedure (an entitling procedure) will take precedence over the nonformal assessment of deserts.

10. Desert and Justice

One way of understanding desert claims is as claims about justice. John Hospers, for example, writes, 'Justice is getting what one deserves. What could be simpler?' (Hospers 1961, p. 433). Desert claims then would have the force of claims about justice. But many have thought that this way of connecting desert with justice is too simple (Feinberg 1963, p. 90, Sher 1987, p. 49, and Slote 1973, p. 333). Even so, these writers have retained some sort of conceptual link between desert and justice. After all, when people do not get what they deserve, our immediate response is to think that they have suffered unfairly or unjustly. It is one thing to say, however, that a person's deserts constitute a factor—albeit an important one—in deciding what justice requires, and another to say—with Hospers—that deserts are the sole determinant of justice. The justice of social institutions may affect what people deserve no less than their deserts may constitute a factor in determining how they are to be justly treated (Sher 1987, p. 206–10).

11. The Rejection of Desert

Although I have argued that desert bases need not refer to conduct for which people are held responsible; nevertheless, for some distributions desert bases need to be factors for which people can be held responsible. Punishment, I noted, can hardly be said to be deserved if the conduct for which it is said to be deserved is not conduct for which the person to be punished can be held responsible.

Rawls has turned this point into a general thesis against the justice of desert-based distributions. Rawls argues that the factors that are usually adduced as desert bases in the distribution of social resources are not really things for which people can be held responsible. Those who have greater abilities and hence greater achievements have them because of their 'greater natural endowments,' not because of something for which they can be held responsible. Even a person's 'character depends in large part upon fortunate family and social circumstances for which he can claim no credit' (Rawls 1971, p. 104). It is 'one of the fixed points of our considered judgments that no-one deserves his place in the distribution of native endowments, any more than one deserves one's initial starting place in society' (Rawls 1971, pp. 103–4). Not even effort will get around this. 'Even the willingness to make an effort, to try, and so to be deserving in the ordinary sense is itself dependent upon happy family and social circumstances' (Rawls 1971, p. 74).

Although many have interpreted Rawls as taking a deterministic position on effort and achievement—seeing all effort and achievement as the inevitable outcome of native endowment and social opportunity—Rawls's rejection of desert is actually less categorical. He does not argue that we are only what we are born and bred to be, but that the task of separating out the various factors that have contributed both to our effort and to our achievement—those for which we are responsible from those that represent our endowment and opportunities—and thus enable an assessment of desert, is 'impracticable' (Rawls 1971, p. 312). Judgments about deserts should therefore be excluded from assessments of social justice. Curiously, Rawls does not draw out the same implications for his briefer discussion of retributive justice.

If we allow that deserts need not depend on things for which we can be held responsible, it might not seem so problematic to maintain both that the traditional desert bases can be used as one possible ground for distribution, and also that these are not things for which we can be held—in any strong sense—responsible. Nozick has argued that 'it needn't be that the foundations underlying desert are themselves deserved, all the way down' (Nozick 1974, p. 225). If I win a race in part because of my natural talent, I may plausibly claim to deserve the prize. My personal desert does not presuppose that my talent was also deserved.

Nozick's critique goes further. Whereas Rawls infers from the arbitrariness of natural talent that such talents should be seen as common assets, Nozick does not see how the arbitrariness and undeserved nature of their natural distribution serves to disqualify those who have them from benefiting from them. Since the holdings that come into the world are already held, is

it not plausible to assume that those who fortuitously have the holdings are 'preinstitutionally entitled' to the benefits that flow from them? Plausible, but not self-evident. We may consider it appropriate to grant the naturally talented and socially advantaged the benefits that flow from their fortuitous position, just as we may consider inherited wealth an entitlement of those who benefit from it. But it is hardly more obvious in the former case than in the latter that this is a matter of desert. If it appears so, this is because it is usually the case that the naturally talented and socially advantaged have to make something of their own of the endowments and opportunities they have received.

In its own way, the contemporary debate about personal desert reflects a much larger social debate about the extent to which our individuality—and that which might be said to flow from that by way of our share of social goods and ills—is viewed simply through the lens of our efforts and achievements, and the extent to which those efforts and achievements are considered as elaborations of a social environment that has fostered and sustained them. Because the answer is unlikely to lie at either extreme, the defense of desert will inevitably be qualified.

See also: Attributional Processes: Psychological; Criminal Justice, Ethics of: Aspects of Human Dignity; Justice and its Many Faces: Cultural Concerns; Justice and Law; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Social Justice

Bibliography

Adler J E 1987 Luckless desert is different desert. *Mind* **96**: 247–49

Aristotle 1962 *Nicomachean Ethics* (trans. M. Ostwald). Bobbs-Merrill, Indianapolis, IN

Cummisky D 1987 Desert and entitlement: a Rawlsian consequentialist account. *Analysis* **47**: 15–9

Feinberg J 1963 Justice and personal desert. In: Friedrich C J, Chapman J W (eds.) *Nomos VI: Justice*. Atherton Press, New York, pp. 69–97

Feldman F 1995 Desert: reconsideration of some received wisdom. *Mind* **104**: 63–77

Fields L 1987 Parfit on personal identity and desert. *Phil. Q.* **37**: 432–41

Hospers J 1961 *Human Conduct: An Introduction to the Problems of Ethics*. Harcourt, Brace & World, New York

Kleinig J 1971 The concept of desert. *Am. Phil. Q.* **8**: 71–8

Lamont J 1994 The concept of desert in distributive justice. *Phil. Q.* **44**: 45–64

Milne H 1986 Desert, effort and equality. *Journal of Applied Philosophy* **3**: 235–43

New C 1992 Time and punishment. *Analysis* **52**: 35–40

New C 1995 Punishing times: reply to Smilansky. *Analysis* **55**: 60–2

Nozick R 1974 *Anarchy, State, and Utopia*. Basic Books, New York

Parfit D 1984 *Reasons and Persons*. Oxford University Press, New York

Pojman L P 1999 Merit: Why do we value it? *Journal of Social Philosophy* **30**: 83–102

Pojman L P, McLeod O (eds.) 1999 *What Do We Deserve: A Reader on Justice and Desert*. Oxford University Press, New York

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Richards N 1986 Luck and desert. *Mind* **95**: 198–209

Rosebury B 1995 Moral responsibility and 'moral luck'. *Philosophical Reviews* **104**: 499–524

Scheffler S 1992 Responsibility, reactive attitudes, and liberalism in philosophy and politics. *Philosophy and Public Affairs* **21**: 299–323

Sher G 1987 *Desert*. Princeton University Press, Princeton, NJ

Slote M A 1973 Desert, consent and justice. *Philosophy and Public Affairs* **2**: 323–47

Smilansky S 1994 The time to punish. *Analysis* **54**: 50–3

Strawson P F 1962 Freedom and resentment. *Proceedings and Addresses of the British Academy* **48**: 1–25

Sverdlik S 1983 The logic of desert. *Journal of Value Inquiry* **17**: 317–24

J. Kleinig

Design and Form: Organizational

Organization design refers to the formal structures, practices and processes through which organizations seek to accomplish organizational goals. Interest in organizational design stems from two intellectual trajectories (see Bolman and Deal 1997). Industrial analysts such as Frederick Taylor, Henri Fayol, and Lyndall Urwick were interested in using 'scientific management' to design organizations to create maximum efficiency. This work produced an interest in issues such as the specialization of tasks and the nature of hierarchy, authority, and responsibility. A second trajectory emerged through the work of Max Weber who was interested in how, as part of the increasing rationality of society, organizations were being designed according to bureaucratic principles such as impersonal authority, hierarchy, and systematic controls.

For some writers, an organization design 'emanates from an overall vision for the organization … (which) will give people the skills, direction and motivation to do the work necessary to achieve the strategic objectives' (Nadler and Tushman 1997). For other writers, 'structural designs are really rhetorical devices … (used by managers which) … are not so much descriptions but ideas about what kind of behavior is desirable in certain circumstances' (Eccles and Nohria 1992). Still other writers note that some organizations try to avoid the use of formal designs and organization charts on the grounds that these can limit the range of behaviors that may be necessary to accomplish organizational goals (Mills 1993).

Many writers, implicitly or explicitly, conceive of designs as problem-solving devices (Eccles and Nohria 1992). For example, in moving beyond a position that there is one best design, contingency theorists have directed their attention over the years to identifying which designs are best suited to solving a variety of 'problems' such as the nature of the organization's environment, strategy, technology, age, and size.

1. *Classic Designs*

Traditionally, much attention has focused on the prevalence of two organizational designs. A functional design is one that is organized into departments on the basis of the skills which individuals bring into an organization (marketing skills, production skills, accounting skills, etc). Miles and Snow (1984) suggest that identifying such a design served as a major breakthrough at the turn of the twentieth century in that it enabled organizations to grow and specialize, in terms of either particular markets or products. Such a design is suited to solving the 'problem' of stable environments in which the same goods or services are required repeatedly (e.g., the day-in-day-out delivery of a railway service). The functional design is usually hierarchically organized and focuses on ways of increasing the efficiency of delivering repetitive operations. Bedeian and Zammuto (1991) suggest that a number of advantages accrue to such organization designs. These include: retaining like interests and skills in specific departments, enabling simple communication among specialists, simplifying training and the efficient use of resources, measuring functional outputs, and enabling top management to retain strategic control over the whole organization. Disadvantages include: coordination across different functional areas, departmental rather than organizational orientation of staff, lack of development of broad management and business skills, interdepartmental conflict, low client satisfaction (as no one department has responsibility or accountability for the service or product), and the difficulty of organizational change if environmental or strategic circumstances change (since the organizational orientation is towards stability).

A divisional design—sometimes called an M-form (Multidivisional) design—is one suited to accommodate organizational growth, and market and product differentiation (Miles and Snow 1984). It is organized by divisions each of which is responsible for a particular output, be it product, market, or geographic area. The same function (e.g., marketing, research and development, and manufacturing) is likely to be found in more than one division so that each becomes specialized in orientation towards the particular product, market, or geographic area that it serves. Such a design is one which is suited to solving the 'problem' of more dynamic environments since an organization can, in theory, add a new division oriented towards a new product or service, or close down or sell nonperforming divisions as conditions change. Other advantages of the divisional design include: easy identification of divisional performance, greater coordination among functional areas as each has a similar orientation, greater attention to client needs, flexibility, and a broader business orientation. Disadvantages include: duplication, less communication among similar functional staff who are spread across different divisions, and a potential failure of staff to keep abreast of professional developments in their particular functional field (Bedeian and Zammuto 1991).

2. *Design Variations*

The matrix design breaks a classic organizing principle underlying the first two designs. This principle is what Fayol (1949) referred to as the scalar chain of command in which a single chain links the highest authority to the lowest rank. The matrix design breaks this principle by invoking dual lines of authority in which both functional and product managers share formal power and authority. Functional managers allocate staff into various project groupings, for which a product manager is responsible. At the end of a particular product or project, staff are reallocated into new groupings as the need arises. Staff, therefore, have dual reporting lines, one to their product manager and one to their functional manager. This design is one that is intended to solve the 'problem' of dynamic environments in so far as new project teams can be established or dissolved as required. The advantages of the matrix structure include the ability to meet new market demands, and to coordinate interdependent functional skills and requirements. Disadvantages are that some staff may have trouble coping with the insecurity and instability associated with such a structural arrangement which, in turn, 'creates considerable confusion, stress, and conflict, and requires from its participants highly developed interpersonal skills and considerable tolerance for ambiguity' (Mintzberg 1983).

In addition to these designs, Mintzberg (1983) identified three others. The simple structure typically is found in small organizations and is characterized by direct supervision and minimal technological systems with a dominating owner or entrepreneur having a wide span of control. The professional bureaucracy characterizes organizations such as accounting firms or hospitals that operate traditionally in relatively stable environments and utilize predetermined 'solutions' to client problems. The adhocracy often draws upon advanced technological systems and uses innovative, multidisciplinary teams to produce novel solutions to client problems.

In practice, organizations may use a combination of designs. For example, a hybrid model may combine

elements of functional and divisional designs where an organization retains central control over human resources, strategy, etc., but decentralizes to different divisions' control over marketing, production, sales, etc. Also, the classic work of Lawrence and Lorsch (1967) showed that within well-performing organizations, departments may be structured differently depending upon the nature of their contact with the environment. For example, in both stable and dynamic environments production departments may be relatively insulated and organized on mechanistic lines. However, research and development departments, being more closely oriented to the market and the environment, may be structured more flexibly in dynamic environments, compared with those operating in stable environments. They used the term differentiation to refer to the extent to which departments are designed and structured differently within the one organization, and integration to refer to the process of unifying departmental activities to accomplish organizational goals.

3. *Internal Control and Coordination Mechanisms*

Thompson (1967) argued that the coordination of activities for a given design varies according to the interdependence of work components or departments. He used the term, pooled interdependence, to refer to situations where each part of the organization is relatively independent. Where, in serial fashion, one part needs to act in order for another part of the organization to perform its actions, then this is termed sequential interdependence. Reciprocal interdependence occurs when an output for one part becomes an input for another part, whose work then becomes the input for the first part.

Adapting the work of March and Simon (1958), Thompson (1967) argued that different types of coordination are required in each instance. Standardization (internally consistent rules and routines which direct the activities of each section towards the outcome of the whole) is appropriate to pooled interdependence. Planning is a useful coordination mechanism for sequential interdependence. Mutual adjustment occurs through communication and feedback devices and is most likely to be found where situations are novel and unpredictable and where reciprocal interdependence is required.

More generally, Galbraith (1974) argued that as task uncertainty increases so does the need for lateral design mechanisms for achieving coordination. In order of increasing uncertainty, he pointed to the use of direct contact among staff (where problems are solved jointly), the use of specialized liaison roles (which can assist communication between departments), task forces representing different departments (in order to address a common, specific problem), cross-functional teams (which are longer-term in nature and consider a variety of problems common to different departments), integrating roles (such as program or product managers who liaise across different functional departments to achieve required outcomes), and managerial linking roles (which have formal position and budgetary power attached to them).

4. *Contemporary Challenges to Classic Designs*

The contingency notion that the structural form should reflect key contingencies underlies a still-growing literature that is advocating the replacement of classic structures by 'new organizational forms.' The core argument is that new ways of organizing are necessary if organizations are to perform effectively in 'hypercompetitive environments' (D'Aveni 1994) characterized by rapid and unpredictable environmental change, intensified competition, and short product life cycles.

In such an environment it is argued that organizations need to be flexible, fast moving, and innovative. However, classic modes of organizing commonly are associated with a number of characteristics that are represented as being inconsistent with these demands. Formalization, that involves the specification of formal rules and procedures, inhibits flexibility. Centralization that involves an emphasis on the location of decision-making authority at the higher levels of the hierarchy, reduces discretion at the 'customer interface.' Standardization of procedures may be inconsistent with the needs of diverse and fragmented markets. Size is an important variable because decisions about growth traditionally have been driven by an assumption that bigger is better. This is appropriate where economies of scale are all-important but can become an impediment when speed of response and customization become paramount.

Classic modes of organizing also have been characterized by boundaries, both within organizations and between organizations and key players in their environment. Within organizations, an emphasis on departmental or divisional boundaries inhibits the ability to organize around value-delivering processes. Similarly, maintaining a clear boundary between 'inside the organization' and 'outside the organization' may be an impediment when organizations compete using market-based, rather than hierarchy-based relationships.

5. *Characteristics Attributed to New Organizational Forms*

An extensive array of new organizational forms has been advocated as appropriate to hypercompetitive conditions. These include: the network organization, the virtual organization, the intelligent organization,

the boundaryless organization, the reconfigurable organization, the self-managing organization, and the centerless organization.

Despite the diverse nomenclature, certain core elements pervade these prescriptions. These include flatter structures (fewer levels in the hierarchy, wider spans of control), less rigid boundaries between functions, greater use of networks and alliances, and disaggregation into smaller business units.

What remains to be seen is whether the utilization of new forms of organizing involves the replacement of, or coexistence with, characteristics associated with classic designs.

See also: Bureaucracy and Bureaucratization; Competitive Strategies: Organizational; Conflict: Organizational; Corporate Culture; Innovation: Organizational; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Learning: Organizational; Organization: Overview; Organizational Behavior, Psychology of; Organizational Change; Organizational Decision Making; Organizations, Metaphors and Paradigms in; Organizations, Sociology of; Organizations: Unintended Consequences; Strategic Intelligence; Strategy: Organizational; Technology and Organization

Bibliography

Bedeian A G, Zammuto R F 1991 *Organizations: Theory and Design*. Dryden Press, Chicago

Bolman L G, Deal T E 1997 *Reframing Organizations: Artistry, Choice, and Leadership*, 2nd edn. Jossey-Bass, San Francisco

D'Aveni R *Hypercompetition: Managing the Dynamics of Strategic Maneuvering*. Free Press, New York

Eccles R G, Nohria N 1992 *Beyond the Hype: Rediscovering the Essence of Management*. Harvard Business School Press, Boston

Fayol H 1949 *General and Industrial Management*. Pitman, London

Galbraith J R 1974 Organization design: An information processing view. *Interfaces* **4**: 28–36

Lawrence P, Lorsch P 1967 *Organization and Environment*. Division of Research, Harvard Business School, Boston

March J G, Simon H A 1958 *Organizations*. Wiley, New York

Miles R E, Snow C C 1984 Fit, failure, and the hall of fame. *California Management Review* **26**(3): 10–28

Mills D Q 1993 *Rebirth of the Corporation*. Wiley, New York

Mintzberg H 1983 *Structure in Fives: Designing Effective Organizations*. Prentice Hall, Englewood Cliffs, NJ

Nadler D A, Tushman M L 1997 *Competing by Design: The Power of Organizational Architecture*. Oxford University Press, Oxford, UK

Thompson J D 1967 *Organizations in Action: Social Science Bases of Administrative Theory*. McGraw-Hill, New York

I. Palmer and R. Dunford

Determinism: Social and Economic

The concept of social and economic determinism has many meanings. What seems to be common to all of them is the notion that human beliefs, practices, and actions are the results of social and economic causes. However, ambiguities arise from the fact that the boundaries of the concepts of 'social' and 'economic' are not clear and keep changing. Perhaps, more importantly, the concept of causality itself is not well understood. Although major social theorists such as Karl Marx and Emile Durkheim are often regarded as economic or social determinists, once the concept of causality is made clear, it does not seem that they ever actually held deterministic positions, as will be shown below.

1. Causality and Determinism

This idea that human beliefs, practices, and actions are the results of social and economic causes is often confused with the claim that these results are inevitable and predictable. However, results may not be predictable if either the conditions on which they depend are unknown or these conditions, even when they are known, are so complex as to make it humanly impossible to deduce their consequences. Also, although one may believe on deterministic grounds that the future course of events is inevitable, one may also believe in inevitability for mystical or spiritual reasons.

Causal deterministic laws have the form 'under conditions C, effect E will result,' in which C is said to be a sufficient condition for E, that is, in a deterministic law, the relationship between cause and effect is not merely probabilistic or statistical. However, putative causal claims put forth by those who are regarded typically as determinists usually fall short of being deterministic laws since the conditions C are almost never sufficient for E. Sometimes, at best merely necessary conditions are given. At other times, even less than that is given. This happens when a purported causal relationship between certain social or economic conditions and the beliefs and practices to which they are supposed to give rise is in reality a functional relationship, that is, the beliefs and practices in question are merely adaptations to, rather than results of, these conditions.

2. Durkheim and Social Causes

Perhaps the work of Durkheim best illustrates the need to separate carefully causal from functional hypotheses. In *The Rules of Sociological Method* (1895), he insisted that social facts are to be explained by social causes and that these social causes are to be

carefully distinguished from psychological and other sorts of causes as well as from social functions. However, it is not clear that in practice he always followed these methodological prescriptions.

2.1 The Causes of the Division of Labor

For instance, Durkheim's *The Division of Labor in Society* (1893) is often read as having made the deterministic claim that increasing labor specialization results from increasing population density. However, on a closer reading, he actually said that specialization results from an increase in what he called 'social density,' or the degree to which individual human beings form social relations with one another. He also seemed to suggest that population density, along with the rise of cities and means of communication, is merely a symptom or sign of social density. Because he made population and social density each the necessary and sufficient condition for the other, his interpreters have been confused about which is supposed to be the cause and which the effect. Durkheim then went on to explain that specialization is not caused by population or social density but is an adaptation to it, making explicit references to Darwin's principles. Durkheim even mentioned alternative responses to population pressure, such as emigration. Specialization takes place only when the social bonds among individuals are so strong that emigration is out of the question.

Furthermore, nothing Durkheim said rules out the possibility that specialization could be a successful strategy for coping with other sorts of conditions besides social and population density. For example, hunter-gatherer tribes living under low-density conditions have been known to specialize in the natural products available in their territory, trading with other tribes for other things that they need. Hence, population density does not appear to be even a necessary condition for specialization. At best, it may be what Mackie (1974) has called an 'inus' condition, or an insufficient but necessary part of an unnecessary but sufficient condition.

2.2 The Social Causes of the Categories

Later in his career, Durkheim, along with Mauss and others, advanced a sociology of knowledge according to which our basic categories of thought, including space, time, causality, and substance, are to be explained by social causes (e.g., Durkheim 1912, Durkheim and Mauss 1903). For example, they claimed that a culture's units for measuring time reflect the rhythm of its social life and its concept of space is patterned on the distribution of clans at tribal campsites. The very practice of classifying things in nature by a nested hierarchy of genera and species reflects an elementary social structure consisting of clans subsumed under phratries. However, their sociology of knowledge is ambiguous between the causal claim that these concepts result from social life and the functional claim that these concepts make human social life possible.

3. Marx and Economic Determinism

There are two different but related senses of determinism in Marxist thought. The first is historical determinism or materialism, according to which the course of history is to be explained through material or economic causes. The second is Marx' theory that what he calls the economic 'base' of a society, which consists in the modes and relations of economic production, somehow 'determines' a 'superstructure' of laws, politics, religion, and other ideological beliefs (see *Ideology: History of the Concept*).

3.1 Historical Materialism

'Historical materialism' is actually Friedrich Engels', not Marx's term. According to the materialist conception of history, there is an inevitable class struggle between the bourgeoisie and the proletariat. This belief has antecedents in earlier thinkers, most notably David Ricardo. For Ricardo, class conflict is the inevitable result of a population increase leading to the bringing of marginal lands under cultivation, which leads to wages dropping to a subsistence level. For Marx, on the other hand, class conflict was not a result of population growth but of changing modes of economic production. As new modes of production develop, they come into conflict with the social relations that existed under the previous economic system. The older social relations are encumbrances on these new modes of production. Eventually, these older relations give way to new ones that are more in line with the new form of production. Each major change in the social relations of production, from slavery to feudalism, from feudalism to bourgeois capitalism, and from capitalism to socialism, is accompanied by class conflict (see *Political Economy, History of*).

3.2 Base and Superstructure

Marx appears to have thought that there are irreconcilable differences in thought between the two classes due to their different economic positions in society. According to Marx' famous 'Preface' to his *A Contribution to the Critique of Political Economy* (1859):

> In the social production of their existence, men inevitably enter into definite relations, which are independent of their will, namely relations of production appropriate to a given stage in the development of their material forces of pro-

> duction. The totality of these relations of production constitutes the economic structure of society, the real foundation, on which arises a legal and political superstructure and to which correspond definite forms of social consciousness. The mode of production of material life conditions the general process of social, political and intellectual life. It is not the consciousness of men that determines their existence, but their social existence that determines (*bestimmt*) their consciousness.

Although this passage has been read as suggesting that individuals are the mere puppets of economic and social forces, it is not at all clear that Marx meant 'determines' in any causal or mechanistic sense. The German word *bestimmt* means 'determine' in the sense of decide, define, set limits or boundaries to, or fix something (such as a price or time). Furthermore, the very building metaphor implied in the language of base and superstructure suggests that the form of economic life is at best a necessary condition and not a sufficient cause for that which rests upon it. On the mechanistic reading of 'determines,' Marx would have been guilty of an inexcusable *non sequitor* from premises about necessary conditions to a conclusion about sufficient conditions. Finally, Marx objected to the practice of taking the causal powers of human beings and projecting them onto mere abstractions such as god, the state, or capital. In his *Theses on Feuerbach* (Marx 1845), he had criticized the materialists for doing this, and yet this is precisely what the mechanistic reading of the 'Preface' does.

It appears that Marx was attempting to say that the social and economic conditions choose or select rather than cause certain forms of thought. Marx was trying to find the right words in which to express what was then in fact a new form of explanation that Darwin, whom Marx much admired, had introduced in the *Origin of Species*, which was published the very same year Marx wrote this preface. Perhaps the clearest example of such a functional or adaptationist explanation in Marx is provided by his appeal to class interests in his account of the struggle between bourgeoisie and proletariat.

3.3 Class Interests and Functional Explanation

Under the capitalist mode of production, the bourgeoisie must strive constantly to maintain or increase their rate of profit to stay competitive. Since, to put it simply, the rate of profit is equal to the value of a product divided by the cost of its production, perhaps the best way to increase the rate of profit is to drive down the cost of production. However, part of these costs, such as raw materials and machinery, is fixed, according to Marx. Hence, the bourgeoisie must try to decrease the variable costs of labor, either by reducing wages or by replacing workers with machinery. However, as a greater and greater proportion of the costs of production go into the fixed costs of machinery and a decreasing proportion go into the variable costs of labor, it becomes increasingly difficult for the capitalist to find ways to cut costs. The capitalist is forced continuously to reinvest his profits in machinery in order to replace even more workers with machines and the remaining workers are reduced to a subsistence wage. According to Marx' law of capitalist accumulation in *Capital* (1867), the greater the wealth invested in fixed capital, the greater the misery of the proletariat. Class conflict is the inevitable result.

Economic class interests are not necessarily consciously thought by social actors, but rather are ascribed to these actors by the social scientist observing the situation. This is what Marx meant when he referred to them as 'objective' class interests. For example, to say that it is in the class interest of the bourgeoisie to reinvest their profits in new machinery is to say that those capitalists who thus reinvest their profits will remain in the bourgeoisie while those who do not act in this way will drop down into the proletariat. When Marx said that the class interests of the bourgeoisie and the proletariat conflict, what he seems to have meant is that the capitalist remains a member of the bourgeoisie only as a result of acting in a certain way which directly conflicts with the 'objective' interests of the working class. He did not mean to say that being a member of the bourgeoisie causes one to think in a certain way, which is different than the way in which the proletariat thinks; indeed, it is not all that important or relevant what individuals may actually be thinking. Thus, to say that the capitalist replaces workers with machines because it is in his interest to do so is to give a functional rather than a causal explanation (see *Interest: History of the Concept*).

Indeed, the alternative, causal determinist interpretation of Marx makes him out to be an incoherent thinker. If our ideas were really determined by our class membership, what would be the point of writing manifestoes? One might object that his writings were addressed only to the working class. However, in *The Communist Manifesto* (1848/1955) he described—with obvious reference to people such as himself and Engels—how a portion of the bourgeoisie, 'who have raised themselves to the level of comprehending theoretically the historical movement as a whole,' 'cuts itself adrift' and joins with the proletariat. On the determinist reading of Marx, these claims appear embarrassingly *ad hoc*. A functionalist reading suggests merely that people like Marx have chosen a path in life that will not make them rich.

However, as Cohen (1978) points out, orthodox Marxists tend to resist the interpretation that Marx was offering functional explanations due to their tendency to associate such explanations with the functionalist tradition in the social sciences that began with Bronislaw Malinowski. According to this tradition, a society or culture forms an organic unit, every element of which serves a function that helps to sustain the whole (see *Functionalism, History of*). Marxists

regard functionalist social science as resistant to social change and thus as politically conservative. However, as Cohen argues, the use of functional explanations does not necessarily commit one to this brand of functionalism.

3.4 After Marx

Engels, Karl Kautsky, Enrico Ferri, Antonio Labriola, V. I. Lenin, Nikolai Bukharin, and others have been read as posthumously transforming Marx' theories into deterministic ones, but there is insufficient space here to examine all their theories. Nevertheless, Marx' theories were certainly understood deterministically in their role as the official ideology of the Soviet Union (see *Marxist Social Thought*, *History of*).

4. Recent Determinism: The Strong Program

A group who defended a determinist approach with roots in Durkheim and Marx arose in the 1970s, calling themselves the 'Strong Program in the Sociology of Scientific Knowledge.' Bloor (1976/1991), a principal spokesperson for the group, laid out four tenets for this program. These are (a) that explanations of belief and knowledge should be causal; (b) that sociology should be impartial with respect to the truth or falsity, or the rationality or irrationality, of the belief to be explained; (c) that explanation should be symmetrical, in the sense that the same kind of causes would explain both true and false beliefs; and (d) that the sociology of knowledge should be reflexive, in the sense that the same principles of explanation should apply to sociology itself. The Strong Program regarded accounts of the beliefs and practices of scientists in terms of their social interests as satisfying the demand for causal explanation. Their concept of social interests included not only economic or class interests but political, religious, professional, and other interests as well.

The Strong Program was subjected to sharp criticism by philosophers and other social scientists. Philosophers insisted that scientists held the views that they did on the basis of reason and evidence. Furthermore, they argued that the tenets of impartiality and symmetry prevent the Strong Program from arguing for their position on epistemic grounds, that is, these tenets prohibit them from saying that their beliefs are any more well-founded than the alternatives. Among the social scientists who criticized them, some felt that in plumping for causal explanations, the Strong Program was trying too hard to model the social on the natural sciences. Others criticized them for taking the existence of social structure and social classes as given and not recognizing that these are constantly changing social phenomena brought about through complex human interaction. For these critics, social structure and class exist in and through the beliefs and practices of the members of society and are to be explained in the same way as any other beliefs and practices. To explain a person's scientific beliefs in terms of social interests is thus in effect to privilege one sort of belief over another.

Finally, interest accounts are functional rather than causal explanations. To explain scientists' intellectual commitments in terms of their social interests is to say that these commitments have the consequence of helping to maintain these scientists in their careers, whether or not this is their intended reason for holding these beliefs. However, when giving this sort of account, one must avoid the trap of simply pointing to what appear to be the obvious beneficial consequences of holding some belief or following some practice. One should provide actual evidence that these beliefs and practices actually helped the persons in question maintain themselves in their careers.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Causation: Physical, Mental, and Social; Durkheim, Emile (1858–1917); Environmental Determinism; Functionalism in Anthropology; Functionalism in Sociology; Labor, Division of; Marx, Karl (1818–89); Marxism in Contemporary Sociology; Marxist Social Thought, History of; Reduction, Varieties of; Social Categorization, Psychology of; Strong Program, in Sociology of Scientific Knowledge; Technological Determinism

Bibliography

Bloor D 1976/1991 *Knowledge and Social Imagery*. 2nd edn. University of Chicago Press, Chicago

Bottomore T B, Rubel M 1956 *Karl Marx: Selected Writings in Sociology and Social Philosophy*. McGraw-Hill, New York

Cohen G A 1978 *Karl Marx's Theory of History: A Defense*. Princeton University Press, Princeton, NJ

Durkheim E 1893 *De la division du travail social*. Alcan, Paris (1984 *The Division of Labor in Society*. Free Press, New York)

Durkheim E 1895 *Les Règles de la méthode sociologique*. Alcan, Paris [1982 *The Rules of Sociological Method*. Free Press, New York]

Durkheim E 1912 *Les formes élémentaires de la vie religieuse*. Alcan, Paris [1995 *The Elementary Forms of Religious Life*. Free Press, New York]

Durkheim E, Mauss M 1903 De quelques formes primitives de classification. *L'Année Sociologique* **6**: 1–72(1963 *Primitive Classification*. University of Chicago Press, Chicago)

Mackie J L 1974 *The Cement of the Universe: A Study of Causation*. Clarendon, Oxford, UK

Marx K 1845 Theses on Feuerbach. In: Easton L D, Guddat K H (eds.) *Writings of the Young Marx on Philosophy and Society*. Anchor Books, Garden City, New York

Marx K 1848/1955 *The Communist Manifesto*. Harlan Davidson, Arlington Heights, IL

Marx K 1859 Preface to *A Contribution to the Critique of Political Economy*. K. Paul, London (English: www.marxists.org/archive/marx/works/1850/pole-con/preface.htm; German: www.mlwerke.de/me/me13/me13_007.htm)

Marx K 1867 *Capital*. Modern Library, New York

W. Schmaus

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Deterrence

Deterrence means threatening punitive retaliation to prevent a foe from attacking. It differs fundamentally from defense, which means threatening to fight back if attacked and deny an attacker its objectives. The heart of the distinction is whether it would make sense to carry out the threat.

Although the idea of threatening war to prevent war is as old as the balance of power, deterrence was not just a new word for that old idea. The idea of deterrence originated with the atomic bomb. It was first formulated in 1946 by Bernard Brodie, who tried to capture what was revolutionary about the advent of the nuclear era. First, a nuclear warhead was not a weapon, in Brodie's view, but a terrorist device, whose blast and radiation effects made it inherently indiscriminate. Second, the atomic bomb brought home the fact of mutual vulnerability in a way that aerial bombardment had not. Before the nuclear era, a state's armed forces had to be defeated before the state and its centers of population could be held hostage or destroyed at will. Once nuclear warheads were mated to missiles in the late 1950s, the state and its populace were defenseless. The balance of terror differed in fundamental respects from the prenuclear balance of power, according to Brodie. 'The first and most vital step,' he wrote, 'for the age of atomic bombs is to take measures to guarantee ourselves in case of attack the means of retaliation. Thus far the chief purpose of a military establishment has been to win wars. From now on its chief purpose must be to avert them.' These prescient insights became a matter of considerable scholarly discourse. Regrettably they had little influence on nuclear policy.

1. Deterrence in Theory

The logic of deterrence, as Brodie and others recognized, is inherently paradoxical. To deter nuclear attack, it is deemed necessary to threaten nuclear retaliation. Inasmuch as the cost of a nuclear war is prohibitive, neither side would risk it. Yet that rational calculus conveniently overlooks the irrationality of retaliating for a nuclear first strike with a nuclear second strike. Retaliation makes no sense unless the second strike could eliminate the other side's nuclear forces and prevent it from launching another strike of its own. That deadly logic led Brodie to the paradoxical conclusion that what made deterrence work was the possibility that it might fail, making it prudent for neither side to tempt fate. The same logic led other strategic thinkers to conclude that for deterrence to work, it could not be mere bluff. The threatener had to be willing to carry out the threat. For that to happen, nuclear retaliation had to be automatic or mad. Herman Kahn (1960) captured the sense of that line of reasoning in a phrase, the 'rationality of irrationality.' Yet that line of reasoning only posed another paradox: if the deterrer was hotheaded, why would the attacker be prudent?

Even if the threat of retaliation ruled out nuclear war, that seemed to leave the world safe for war waged by other than nuclear means. That was a source of gnawing anxiety for the USA, which was committed by treaty to deter any attack by the USSR on its allies in Western Europe. It did not have the capability to do that by conventional military means alone, at least in the 1950s, so it chose to rely on the threat to initiate nuclear war. That threat was potentially suicidal.

Some strategists insisted that 'extended deterrence,' as that threat came to be called, would work if the USA had nuclear superiority. Thomas Schelling (1966) questioned whether superiority assured that the weaker side must yield to the stronger. It was impossible, he insisted, to raise the risk of a nuclear war for the other side without raising it for oneself. He spoke, instead, of a 'competition in risk-taking.' Extended deterrence still might work if Soviet leaders could not be sure the USA would shrink from first use of nuclear weapons in the event of war. Schelling (1960) called it 'the threat that leaves something to chance.' Such a cosmic bluff prompted a preoccupation with 'credibility' and a willingness to run seemingly irrational risks to shore up an incredible threat. McGeorge Bundy (1983) came up with a less demanding alternative,'existential deterrence.' So long as a state possessed nuclear arms, Bundy argued, it had a latent ability to use them, even if it did not explicitly threaten to do so. The mere existence of these arms and the incalculable costs of a nuclear war exerted a cautionary effect on potential foes.

Other strategic thinkers took a more forceful tack. British strategist Basil Liddell Hart (1946) first formulated the idea of graduated deterrence, carefully calibrated to threaten greater costs to an attacker without generating all-out war. Still others debated whether military forces or civilians were the most appropriate targets for nuclear attack. This line of thought culminated in the baroque escalation ladder of Herman Kahn (1960) and the contention of some of his followers that the USA could deter any aggression if it somehow had 'escalation dominance,' or military superiority at every level of violence, short of all-out nuclear war. Yet the idea of waging limited nuclear war and controlling escalation raised the question of how a nuclear war could stop, short of a dead end.

2. Deterrence in Practice

Deterrence, in short, did not work very well in theory. What happened in practice was in many ways worse. Deterrence did not govern the production, deployment, or military plans for nuclear arms (Ball

1978). The number of Soviet and US warheads vastly exceeded what was needed for deterrence. The warheads themselves were initially deployed in ways that made them vulnerable to attack. Dispersing them on land and at sea reduced, though did not eliminate their vulnerability, and made command-and-control all the more precarious. Because of the chance that the cosmic bluff could be called, each side drew up detailed war plans to destroy the other. Those war plans fully reflect the deadly logic of deterrence: in the event that war breaks out, nuclear arms confer enormous, some say decisive, advantage on the side that strikes first. The most urgent targets were the other side's nuclear forces and especially their command-and-control. Concern about a disarming first strike led the USA to keep some of its bombers airborne at all times and to put others on alert, ready to take off at a moment's notice. It did the same with its missiles. This hair-trigger posture, and the delegation of authority that it necessitated, raised the risk of loss of nuclear control in a crisis. Schelling (1960) called this predicament 'the reciprocal fear of surprise attack.' As first one side and then the other began mobilizing forces, the very steps taken to deter a nuclear war might provoke one. Far from exerting a cautionary effect on preparations for war, that fear led both sides to a massive buildup of arms, conventional as well as nuclear, which did little to calm the fear.

That had important implications for theorizing about deterrence. Although some theorists saw mutual deterrence as a source of stability in international politics, others saw it as potentially unstable in the extreme. The first group emphasized the need to demonstrate the capability and will to wage war, lest a potential aggressor doubt a state's strength or resolve. By this way of thinking, conciliation was dangerous because it might be mistaken for weakness or irresoluteness. A second group of theorists were led to a renewed appreciation of the security dilemma, an idea dating back to Thucydides, who noted that attempts by one side to enhance or demonstrate its military might prove self-defeating. By alarming its rival and leading it to respond in kind, these measures would leave both sides less secure. Such a vicious circle could generate an arms race, or worse, trigger pre–emptive war. These theorists saw the need to couple coercion with conciliation, lest the interaction spiral out of control. To some scholars, the difference between the deterrence and spiral logic depended on each side's perception of the other's intentions.

Where perception matters, so does the danger of misperception. The fear that war is imminent is sometimes irrational. Once it takes hold, it is not amenable to a cool calculation of costs. The possibility of pre–emptive war calls into question two major premises of most theorizing about nuclear deterrence, that statesmen will behave rationally in the heat of the moment and that they can control the operations of their armed forces. This led to critiques of deterrence theory on the grounds that it was psychologically naive and operationally uninformed. Robert Jervis (1974), John Steinbruner (1976), and Scott Sagan (1993) are leading exemplars of these approaches.

One side's strategy depended on the other's. So did its security. Their interdependence was the central feature of *game theory*. Whether the game of chicken or the prisoner's dilemma was the appropriate analogy, many scholars who formulated the logic of deterrence drew heavily on game theory for their insights.

One consequence is that deterrence is largely deductive and only weakly grounded empirically. Another reason for its weak evidentiary base is that nuclear history everywhere remains cloaked in secrecy. Without knowledge of the details of nuclear planning and operations, it is difficult to determine how much nuclear strategy is informed by deterrence theory, if at all. A more fundamental reason why the evidence for deterrence is less than compelling is the epistemological difficulty of proving why something did not happen in order to demonstrate that deterrence worked. The obvious counter to the contention that nuclear deterrence kept the USSR out of Western Europe is that it never intended to invade in the first place.

Because the side that is losing could still destroy the side that is winning, safety in the nuclear era lay in cooperation between enemies. That interdependence made nonsense of traditional strategic thought. It also called into question the very idea of strategy as a rational relationship between means and ends. War, in theory, if not always in practice, 'is controlled by its political object,' Clausewitz had reasoned. Consequently, 'the value of this object must determine the sacrifices to be made for it in magnitude and also in duration.' But what objective could possibly sustain the full measure of sacrifice in a nuclear war? As Lawrence Freedman (1981) concludes, 'The position we have reached is one where stability depends on something that is more the antithesis of strategy than its apotheosis—on threats that things will get out of hand, that we might act irrationally, that possibly through inadvertence we could set in motion a process that in its development and conclusion would be beyond human control and comprehension.' Nuclear strategy, understood this way, was a contradiction in terms.

See also: Diplomacy; Genealogy in Anthropology; War: Causes and Patterns

Bibliography

Ball D 1980 *Politics and Force Levels*. University of California Press, Berkeley, CA

Brodie B 1946 *The Absolute Weapon*. Harcourt, Brace, New York

Bundy M 1983 The Bishops and the Bomb. In: *New York Review of Books* **10** : 30
Freedman L 1981 *The Evolution of Nuclear Strategy*. St. Martin's Press, New York
Jervis R 1976 *Perception and Misperception in International Politics*. Princeton University Press, Princeton NJ
Kahn H 1960 *On Thermonuclear War*. Princeton University Press, Princeton, NJ
Liddell Hart B H 1946 *The Revolution in Warfare*. Faber and Faber Ltd., London
Sagan S D 1993 *The Limits of Safety*. Princeton University Press, Princeton, NJ
Schelling T 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA
Schelling T C 1966 *Arms and Influence*. Yale University Press, New Haven, CT
Steinbruner J D 1974 *The Cybernetic Theory of Decision*. Princeton University Press, Princeton, NJ

L. V. Sigal

Deterrence: Legal Perspectives

The criminal justice system (CJS) dispenses justice by apprehending, prosecuting, and punishing law-breakers. These activities also project a threat of punishment. Deterrence occurs when some would-be law breakers conclude that the price of crime is prohibitive. General deterrence refers to the crime prevention impact of the threat of punishment on the public at large. Specific deterrence refers to the impact of punishment on persons actually punished. The actual experience of punishment may alter an individual's perceptions and opportunities. Also, persons for whom general deterrence has failed may be systematically different than the population at large. For both these reasons specific deterrent effects may be smaller than the subject of this article, general deterrent effects.

1. A Brief Review of Research on General Deterrence

Deterrence research has evolved in three distinctive and largely disconnected literatures—interrupted time series, perceptual, and ecological studies (Nagin 1998).

1.1 Interrupted Time Series Studies

Interrupted time series studies examine the impact of targeted policy interventions such as police crackdowns or implementation of statutes changing penalties. The best-designed studies mimic important features of a true experiment—a well-defined treatment regime, measurement of response before and after treatment and a control group. Two classic examples are Ross' studies of the impact on drunk driving of the British Road Safety Act and of Scandinavian-style drunk driving laws (Ross 1982).

The great proportion of interrupted time series studies have examined the impact of drunk driving laws or of police crackdown, on drug markets, disorderly behavior and drunk driving. A less extensive literature has also examined the impact of gun control laws and ordinances. Excellent reviews of these studies are available from Sherman (1990) and Ross (1982).

Both reviews conclude that interventions are generally successful in generating an initial deterrent effect. One exception may be interventions that increase sentence severity. If judges or juries believe that the penalties are too harsh, they may respond by refusing to convict guilty defendants with the result that the policy increases rather than deters the targeted behavior. Indeed, Ross concludes that efforts to deter drunk driving with harsher penalties commonly fail for precisely this reason. Sherman and Ross are also in agreement that the effect is generally only transitory: the initial deterrent effect typically begins decaying even while the intervention is still in effect. However, in some instances, the decay is not complete even following the end of the crackdown.

1.2 Perceptual Deterrence Studies

1.2.1 Summary of findings. The perceptual deterrence literature examines the relationships of perceived sanction risks to either self-reported offending or intentions to do so. This literature was spawned by researchers interested in probing the perceptual underpinnings of the deterrence process.

Perceptual deterrence studies have focused on examining the connection of illegal behavior to two categories of sanction variables, the certainty and severity of punishment. The certainty of punishment refers to the probability that a crime will result in punishment whereas severity refers to the seriousness of the consequences to the punished individual, such as prison sentence length.

Perceptual deterrence studies have been based on three types of data: cross-sectional survey studies, panel survey studies, and scenario-based studies. In cross sectional survey studies individuals are questioned about their perceptions of the certainty and severity of sanctions and either about their prior offending behavior or future intentions to offend. For example, Grasmick and Bryjak (1980) queried a sample of city residents about their perceptions of the risk of arrest for offenses such as a petty theft, drunk driving, and tax cheating, and also about their intention to commit each of these acts in the future.

In panel survey studies the sample is repeatedly surveyed on risk perceptions and criminal behavior. For example, Paternoster et al. (1982) followed a sample of students through their three-year tenure in

high school and surveyed the students on the frequency with which they engaged in various delinquent acts and their perceptions of the risks and consequences of being caught for each such acts.

In scenario-based studies individuals are questioned about their perceptions of the risks of committing a crime described in a detailed crime vignette and about their own behavior if they found themselves in that situation. Nagin and Paternoster (1993), for instance, constructed a scenario describing the circumstances of a date rape and surveyed a sample of college males about their perceptions of the risk of the scenario male being arrested for sexual assault and also about what they themselves would do in the same circumstance.

The cross-sectional and scenario-based studies have consistently found that perceptions of the risk of detection and punishment have a negative, deterrent-like association with self-reported offending or intentions to do so. Such deterrent-like associations with perceived severity are somewhat less consistent, but when individual assessments of the cost of such sanctions are taken into account, statistically significant negative associations again emerge. Only in the panel-based studies have null findings on deterrent-like effects been found.

In panel-based studies respondents were typically interviewed on an annual cycle. With these data researchers examined the relationship of behavior between years t and $t+1$ to risk perceptions at the outset of year t. Their purpose was to avoid the problem of causal ordering: Is a negative association a reflection of the deterrent impact of risk perceptions on crime or of the impact of criminal involvement on risk perceptions? Generally, these studies have found only weak evidence of deterrent-like associations (Paternoster 1987).

The null findings from the panel-based studies generated a spirited debate on the appropriate time lag between measurements of sanction risk perceptions and criminal involvement. Grasmick and Bursik (1990) and Williams and Hawkins (1986) have argued that ideally the measurements should be made contemporaneously because perceptions at the time of the act determine behavior.

The argument for temporal proximity is compelling but the challenge is its practical achievement. People cannot be queried on their risk perceptions on a real-time basis as they encounter criminal opportunities in their every day lives. The scenario method offers one solution because respondents are not questioned about their actual behavior or intentions but rather are asked about their likely behavior in a hypothetical situation. On average, persons who perceived that sanctions were more certain or severe reported smaller probabilities of their engaging in the behavior depicted in the scenario, whether it be tax evasion (Klepper and Nagin 1989), drunk driving (Nagin and Paternoster, 1993), sexual assault (Bachman et al. 1992), or corporate crime (Paternoster and Simpson 1997). Thus, it seems that a consensus has emerged among perceptual deterrence researchers in support of this position.

This consensus reframes the question of the deterrent effect of sanctions from the issue of whether people respond to their perceptions of sanction threats to the issue of whether those perceptions are grounded in reality. The literature on the formation of sanction risk perception is small and narrow in scope (Nagin 1998). The dearth of evidence on the policy-to-risk-perceptions linkage also leaves unanswered a key criticism of skeptics of the deterrent effects of official sanctions. Even if crime decisions are influenced by sanction risk perceptions, as the perceptual deterrence literature strongly suggests, absent some linkage between policy and perceptions, however imperfect, behavior is immune to policy manipulation. In this sense behavior lacks rationality, not because individuals fail to weigh perceived costs and benefits, but because the sanction risk perceptions are not anchored in reality.

1.2.2 The linkage between formal and informal sanction processes. Arguably the most important contribution of the perceptual deterrence literature does not involve the evidence it has amassed on deterrence effects *per se*. Rather it is the attention it has focused on the linkage between formal and informal sources of social control. Recognition of this connection predates the perceptual deterrence literature. For instance, Zimring and Hawkins (1973, p. 174) observe that formal punishment may best deter when it triggers informal sanctions: 'Official actions can set off societal reactions that may provide potential offenders with more reason to avoid conviction than the officially imposed unpleasantness of punishment.' See also Andenaes (1974) for this same argument.

Early perceptual deterrence studies did not consider the connection between formal and informal sanctioning systems but a review by Williams and Hawkins (1986) prompted a broadening of the agenda to consider this issue. Their position was this: Community knowledge of an individual's probable involvement in criminal or delinquent acts is a necessary precondition for the operation of informal sanction processes. Such knowledge can be obtained from two different sources: Either from the arrest (or conviction or sentencing) of the individual or from information networks independent of the formal sanction process (e.g., a witness to the crime who does not report such knowledge to the police). Williams and Hawkins observe that deterrent effects may arise from the fear that informal sanctioning processes will be triggered by either of these information sources. They use the term 'fear of arrest' to label deterrent effects triggered by the formal sanction process and the term 'fear of the act' to label deterrent effects triggered by information networks separate from the formal sanction

process. The crux of their argument is that all preventive effects arising from 'fear of arrest' should be included in a full accounting of the deterrent effect of formal sanctions.

Much of scenario-based research confirms their argument. This research has consistently found that individuals who report higher stakes in conventionality are more deterred by perceived risk of exposure for law breaking. A salient finding in this regard concerns tax evasion. In Klepper and Nagin (1989) a sample of generally middle-class adults were posed a series of tax noncompliance scenarios. The amount and type of noncompliance (e.g., over-stating charitable deductions or understating business income) was experimentally varied across tax return line items. The study found that a majority of respondents reported a non-zero probability of taking advantage of the noncompliance opportunity described in the scenario. Plainly, these respondents were generally willing to consider tax noncompliance when only their money (and not their reputation) was at risk. They also seemed to be calculating; the attractiveness of tax noncompliance gamble was inversely related to the perceived risk of civil enforcement.

The one exception to the rule of confidentiality of enforcement interventions is criminal prosecution. As with all criminal cases, criminal prosecutions for tax evasion are a matter of public record. Here there was evidence of a different decision calculus; seemingly all that was necessary to deter evasion was the perception of a non-zero chance of criminal prosecution. Stated differently, if the evasion gamble also involved putting reputation and community standing at risk, middle-class respondents were seemingly unwilling to consider taking the noncompliance gamble.

The tax evasion research does not pin down the specific sources of stigma costs, but other research on the impact of a criminal record on access to legal labor markets suggests a real basis for the fear of stigmatization (Freeman 1995, Waldfogel 1994). Freeman estimates that a record of incarceration depresses probability of work by 15 percent to 30 percent and Waldfogel estimates that conviction for fraud reduces income by as much as 40 percent.

1.3 Ecological Studies

Two broad classes of ecological analyses are considered—studies of the deterrent effect of prison and of the police.

1.3.1 The impact of prison population on crime rate. Except for capital punishment, prison is the most severe and costly sanction. Except for a flurry of flawed studies in the 1970s, however, there have been few studies of the impact of this sanction. The paucity of studies is attributable to the problem of untangling cause from effect in statistical analyses of natural variation in crime rates and sanction levels as measured by indicators like prisons per capita or actual time served in prison. For example, all else equal, more crime will increase the incarceration rate. However, all else is not equal because a larger prison population will tend to reduce the crime rate. In statistical parlance, the problem of sorting out these contending impacts is called identification (Blumstein et al. 1978, Fisher and Nagin 1978).

At least one study has plausibly dealt with the identification problem. Levitt (1996) employs a clever identification strategy based on measuring the impacts on crime rates of court orders to reduce prison overcrowding. He estimates that 15 index crimes are averted for each additional man-year of imprisonment.

Can a study such as that conducted by Levitt provide an all-purpose estimate of the impact of prison population on crime rate. The answer, I believe, is no. Specifically, there is good reason for believing that policies producing equivalent changes in the prison population will not result in the same change in crime rate.

Figure 1 is a two-dimensional taxonomy of sanction policies affecting the scale of imprisonment. One dimension labeled 'Type' distinguishes three broad categories. Policies regulating certainty of punishment such as laws requiring mandatory imprisonment, policies influencing sentence length such as determinate sentencing laws, and policies regulating parole powers. The second dimension of the taxonomy, 'Scope,' distinguishes policies that cast a wide net, such as a general escalation of penalties for broad categories of crime, compared to policies that focus on targeted offenses (e.g., drug dealing) or offenders (e.g., 'three-strikes' laws).

The Levitt study is based on US data during a period in which the imprisonment rate more than quadrupled. This increase was attributable to a combination of policies belonging to all cells of this matrix. Parole powers have been greatly curtailed, sentence lengths increased, both in general and for particular crimes (e.g., drug dealing), and judicial discretion to impose nonincarcerative sanctions has been reduced (Tonry 1995). Consequently, any impact on the crime rate of the increase in prison population reflects the effect of an amalgam of potentially interacting policies. By contrast the impact estimated in the Levitt study measures the preventive effect of reductions in the imprisonment rate induced by the administrative responses to courts orders to reduce prison populations. Thus, his estimate would seem only to pertain to policies affecting parole policies.

1.3.2 The impact of police on crime rate. The largest body of evidence on deterrence in the ecological literature focuses on the police. The earliest

		Scope	
		General	Targeted
Type	Certainty	Increasing the number of police	Crackdown on drug dealing
	Severity	Broad-based mandatory minimums	'Three-strikes' laws
	Parole	Parole abolition	No parole for violent offenders

Figure 1
Taxonomy of prison sanction policies

generation of studies on the deterrent effect of police examined the linkage of crime rate to measures of police resources (e.g., police per capita) or to measures of apprehension risk (e.g., arrest per crime). These studies were inadequate because they did not credibly deal with the identification problem. If the increased crime rates spur increases in police resources, as seems likely, this impact must be taken into account to obtain a valid estimate of the deterrent effect of those resources.

A study by Wilson and Boland (1978) plausibly identifies the deterrent effect of the arrest ratio. They argued that the level of police resources *per se* is, at best, only loosely connected to the apprehension threat they pose. Rather the crucial factor is how the police are mobilized to combat crime; Wilson and Boland argue that only proactive mobilization strategies will have a material deterrent effect. In their words (1978, p. 373), 'By stopping, questioning, and otherwise closely observing citizens, especially suspicious ones, the police are more likely to find fugitives, detect contraband (such as stolen property or concealed weapons), and apprehend persons fleeing from the scene of a crime.' They conclude that the arrest ratio has a substantial deterrent effect on robbery.

An important follow-up to the Wilson and Boland study was conducted by Sampson and Cohen (1988). Their key premise was similar to that of Wilson and Boland—'hard' policing of 'soft' crime such as prostitution, drunkenness, and disorderly conduct deters serious criminality—and their findings supported the deterrent impact of this policing strategy.

An important innovation in Sampson and Cohen is that they not only estimate a population-wide deterrent effect but disaggregate this effect across segments of the population—white juveniles, black juveniles, white adults, and black adults. They do this by using arrest rates as surrogate measures of demographic group-specific offense rates. They find a negative deterrent-like association between aggressiveness and arrest rate for all groups but they also find significant differences by race and age in the magnitude of the effect. For robbery, at least, adults seems to be more deterred by police aggressiveness than do juveniles, with black adults seemingly more deterrable than white adults. Because the results for specific demographic groups are based on arrest rates, they must be qualified in a number of obvious ways. Notwithstanding, disaggregation efforts are laudable and should, where feasible, become standard in deterrence studies.

1.4 Summary Observations on Deterrence

At the outset, it was stated that the accumulated evidence on deterrence leads me to conclude that the criminal justice system exerts a substantial deterrent effect. That said, it is also my view that this conclusion is of limited value in formulating policy. Policy options to prevent crime generally involve targeted and incremental changes. So for policy makers the issue is not whether the criminal justice system in its totality prevents crime, rather it is whether a specific policy, grafted onto the existing structure, materially adds to the preventive effect. Here there is a distinction drawn between absolute and marginal deterrence. Figure 2 depicts two alternative forms of the response function relating crime rate to sanction levels. Both are downward sloping, which captures the idea that higher sanction levels prevent crime. At the status quo sanction level, S_1, the crime rate, C_1, is the same for both curves. The curves are also drawn so that they predict the same crime rate for a zero sanction level. Thus, the absolute deterrent impact of the status quo sanction level is the same for both curves. But because the two curves have different shapes, they also imply different responses to an incremental increase in

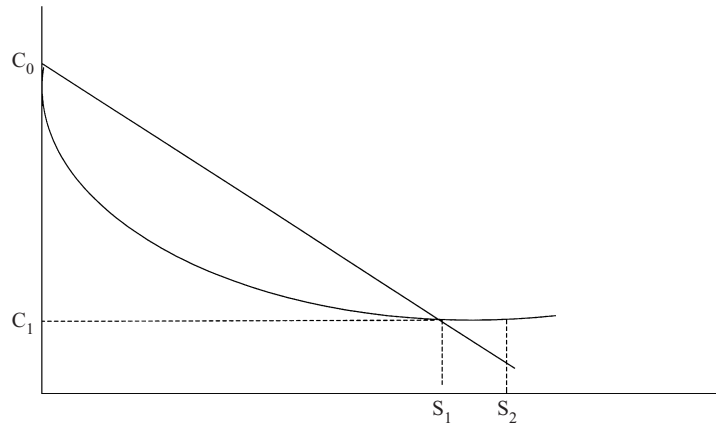

Figure 2
Marginal vs. absolute deterrent effects (curved line is A; straight line is B)

sanction level to S_2. The response implied by curve A is small. Accordingly, it would be difficult to detect and likely not sufficient to justify the change. By comparison, the response depicted in curve B is large, thus more readily detectable, and also more likely to be justifiable as good policy.

While the distinction between absolute and marginal deterrence is useful, it implies an underlying analytical simplicity of the relationship of crime rates to sanction levels that belies the complexity of the phenomenon. Contrary to the implicit suggestion of Fig. 2, no one curve relates sanction levels to crime rates. The response of crime rates to a change in sanction policy will depend on the specific form of the policy, the context of its implementation, the process by which people come to learn of it, differences among people in perceptions of the change in risks and rewards that are spawned by the policy, and by feedback effects triggered by the policy itself (e.g., a reduction in private security in response to an increase in publicly funded security). Thus, while a number of studies have credibly identified marginal deterrent effects, it is difficult to generalize from the findings of a specific study because knowledge about the factors that affect the efficacy of policy is so limited. Specifically, there are four major impediments to making confident assessments of the effectiveness of policy options for deterring crime.

First, while large amounts of evidence have been amassed on short-term deterrent effects, little is known about long-term impacts. Evidence from perceptions-based deterrence studies on the interconnection of formal and informal sources of social control point to a possibly substantial divergence between long- and short-term impacts. Specifically, these studies suggest that the deterrent effect of formal sanctions arises principally from fear of the social stigma that their imposition triggers. Economic studies of the barriers to employment created by a criminal record confirm the reality of this perception. If fear of stigma is a key component of the deterrence mechanism, such fear would seem to depend upon the actual meting out of the punishment being a relatively rare event: Just as the stigma of Hester Prynne's scarlet 'A' depended upon adultery being uncommon in Puritan America, a criminal record cannot be socially and economically isolating if it is commonplace. Policies that are effective in the short term may erode the very basis for their effectiveness over the long run if they increase the proportion of the population who are stigmatized. Deterrence research has focused exclusively on measuring the contemporaneous impacts of sanction policies. Long-term consequences have barely been explored.

The second major knowledge gap concerns the connection of risk perceptions to actual sanction policy. The perceptual deterrence literature was spawned by the recognition that deterrence is ultimately a perceptual phenomenon. While great effort has been committed to analyzing the linkage between sanction risk perceptions and behavior, comparatively little attention has been given to examining the origins of risk perceptions and their connection to actual sanction policy.

For several reasons this imbalance should be corrected. One is fundamental: the conclusion that crime decisions are affected by sanction risk perceptions is not a sufficient condition for concluding that policy can deter crime. Unless the perceptions themselves are manipulable by policy, the desired deterrent effect will not be achieved. Beyond this basic point of logic, a better understanding of the policy-to-perceptions link can also greatly aid policy design. For instance, nothing is known about whether the risk perceptions of would-be offenders for specific crimes are formed principally by some overall sense of the effectiveness of

the enforcement apparatus or by features of the apparatus that are crime specific (e.g., the size of the vice squad or the penalty for firearms use). If it is the former, seemingly targeted sanction policies will have a generalized salutary impact across crime types by heightening overall impressions of system effectiveness. If the latter, there will be no such generalized impact. Indeed would-be offenders may substitute non-targeted offenses for targeted offenses (e.g., committing burglaries in response to increased risk for robbery).

Third, the impact of specific policies—e.g., increasing the number of police—will depend on the form of its implementation across population units. Yet estimates of the deterrent effect of such policies from the ecological literature are commonly interpreted as if they apply to all units of the population from which they were estimated. In general this is not the case. Rather the estimated deterrent effect should be interpreted as the average of the 'treatment' effect across population units. For instance, the deterrent impact of more police in any given city will depend upon a host of factors including how the police are deployed. Consequently, the impact in any given city will not be the same as the average across all cities; it may be larger, but it could also be smaller. Similarly, it is not possible to make an all-purpose estimate of the impact of prison on crime. There are many ways to increase prison population by a given amount, ranging from broad-based policies such as across-the-board increases in sentence length to targeted policies like 'three strikes' statutes. It is likely that the magnitude of the preventive effect will vary materially across these options. The implication is that even though there are credible estimates of average deterrent effects of at least some broad classes of policies, the capacity to translate the average estimates into a prediction for a specific place is limited. This is a major shortcoming in the evidence because crime control is principally the responsibility of state and local government. It is the response of crime to policy in that city or state that is relevant to its population, not the average response across all cities and states.

See also: Crime and Delinquency, Prevention of; Criminal Justice, Sociology of; Deterrence; Deterrence Theory: Crime; Law: Economics of its Public Enforcement; Police, Sociology of; Policing; Prisons and Imprisonment; Prosecution; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Substance Abuse in Adolescents, Prevention of; Time Series Quasi-experiment

Bibliography

Andenaes J 1974 *Punishment and Deterrence*. University of Michigan Press, Ann Arbor, MI

Bachman R, Paternoster R, Ward S 1992 The rationality of sexual offending: Testing a deterrence/rational choice conception of sexual assault. *Law & Society Review* **26**: 401

Blumstein A, Cohen J, Canella-Cacho J 1993 Filtered sampling from populations with heterogenous event frequencies. *Management Science* **37**(4): 886–98

Blumstein A, Cohen J, Nagin D (eds.) 1978 *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. National Academy of Sciences, Washington, DC

Fisher F M, Nagin D 1978 On the feasibility of identifying the crime function in a simultaneous model of crime rates and sanction levels. In: Blumstein A, Cohen J, Nagin D (eds.) *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. National Academy of Sciences, Washington, DC

Freeman R 1995 Why do so many young American men commit crimes and what might we do about it? *Journal of Economic Perspectives* **10**(1): 25–42

Grasmick H G, Bryjak G J 1980 The deterrent effect of perceived severity of punishment. *Social Forces* **59**: 471

Grasmick H G, Bursik R J Jr 1990 Conscience, significant others, and rational choice: Extending the deterrence model. *Law Society Review* **24**: 837–61

Klepper S, Nagin D 1989 The deterrent effect of perceived certainty and severity of punishment revisited. *Criminology* **27**: 721

Levitt S 1996 The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *Quarterly Journal of Economics* **111**: 319–52

Levitt S 1998 Criminal Deterrence Research: A Review of the Evidence and a Research Agenda for the Outset of the 21st Century. In: Michael Tonry (ed.) *Crime and Justice: A Review of Research*.

Nagin D, Paternoster R 1993 Enduring individual differences and rational choice theories of crime. *Law and Society Review* **27**(3): 467–96

Paternoster R 1987 The deterrent effect of the perceived certainty and severity of punishment: A review of the evidence and issues. *Justice Quarterly* **4**: 173–217

Paternoster R, Saltzman L E, Chiricos T G, Waldo G P 1982 Perceived risk and deterrence: Methodological artifacts in perceptual deterrence research. *Journal of Criminal Law and Criminology* **73**: 1238–58

Paternoster R, Simpson S 1997 Sanction threats and appeals to morality: Testing a rational choice theory of corporate crime. *Law and Society Review*

Ross H L 1982 *Deterring the Drinking Driver: Legal Policy and Social Control*. Heath, Lexington, MA

Sampson R J, Cohen J 1988 Deterrent effects of police on crime: A replication and theoretical extension. *Law and Society Review* **22**: 163–89

Sherman L W 1990 Police crackdowns: Initial and residual deterrence. In: Tonry M, Morris N (eds.) *Crime and Justice: A Review of Research*. University of Chicago Press, Chicago, Vol. 12

Tonry M 1995 *Neglect, Race, Crime, and Punishment in America*. Oxford Press, New York

Waldfolgel J 1994 The effect of criminal conviction on income and the trust reposed in the women, *Journal of Human Resources* Winter

Williams K R, Hawkins R 1986 Perceptual research on general deterrence: A critical overview. *Law and Society Review* **20**: 545–72

Wilson J Q, Boland B 1978 The effect of the police on crime. *Law and Society Review* **12**: 367–90

Zimring F E, Hawkins G J 1973 *Deterrence: The Legal Threat in Crime Control*. University of Chicago, Chicago

D. S. Nagin

Deterrence Theory: Crime

1. Introduction

The word *deterrence* has been linked to different topics, including nuclear deterrence (see *Oxford English Dictionary*), but the focus here is on theories about the effects of punishment on crime. Criminal deterrence pertains to the omission or curtailment of crime out of fear of legal punishment, and the first theory can be traced to the utilitarian philosophers Cesare Beccaria (1738–94) and Jeremy Bentham (1748–1832) who believed that people are motivated fundamentally to obtain pleasure and avoid pain. To the extent that potential offenders anticipate pleasure from crime, they can be deterred by increasing the pain associated with it. In particular, potential offenders can be deterred by making legal punishment *certain*, *celeritous*, and *severe*, with *certainty* being the likelihood of being caught and punished for crime, *celerity* being the swiftness of punishment, and *severity* being the amount of punishment (von Hirsch et al. 1999, p. 6). Reduced to one fairly simple statement: 'The rate for a particular type of crime varies inversely [negatively] with the celerity, certainty, and severity of punishments of that type of crime' (Gibbs 1975, p. 5).

2. Complexities

Although that statement is still the basis for contemporary deterrence theory, there are several complexities, one of which has to do with the distinction between *specific* and *general* deterrence (Stafford and Warr 1993). Whereas *specific deterrence* refers to the deterrent effects of directly experiencing a legal punishment, *general deterrence* refers to the deterrent effects of indirectly experiencing it. So when a person is imprisoned for committing a burglary and is then released, the imprisonment deters *specifically* to the extent that the released offender refrains completely from burglary, or at least curtails the amount of burglary that he or she commits, out of a fear of further imprisonment. To the extent that people other than the imprisoned offender learn about the imprisonment and refrain completely or in part from committing burglary out of fear of being imprisoned, those people have been *generally* deterred. The distinction is important because a given punishment might deter *only* specifically or *only* generally instead of both.

Another complexity is implied in the foregoing discussion of specific and general deterrence. Critics sometimes dismiss the possibility of deterrence by arguing that the commission of a crime itself is evidence against deterrence. For example, it sometimes is argued in reference to criminal justice policy to imprison drug traffickers that continued drug trafficking reveals that the threatened imprisonment did not deter potential offenders. However, the argument ignores the distinction between *absolute* and *restrictive* deterrence (Gibbs 1975, pp. 32–4). To the extent that a person refrains completely from committing a crime out of fear of punishment, that person is deterred *absolutely*. However, a person may not be deterred absolutely but, instead, be deterred *restrictively* by curtailing or limiting his or her commission of a crime, as when a person limits the amount of drug trafficking in order to reduce the risk of punishment (perhaps out of a sense of cumulative risk). With restrictive deterrence, the key question is how much crime would occur if there were either no or less threatened punishment for it.

3. Premises of Deterrence Theory

According to Gibbs (1975), there are three premises of deterrence theory:

(a) The greater the actual certainty, celerity, and severity of punishment, the greater the perceived certainty, celerity, and severity of punishment.

(b) The greater the perceived certainty, celerity, and severity of punishment, the less the crime.

(c) The greater the actual certainty, celerity, and severity of punishment, the less the crime.

Or stated diagrammatically: AP $<-->$ PP $>-<$ CR, where AP denotes actual punishments, PP denotes perceived punishments, CR denotes crime; the symbol $<-->$ denotes a positive relationship, and $>-<$ denotes a negative relationship. In contrast to the actual certainty, celerity, and severity of punishment in a given place and time for a given type of crime, perceived punishments pertain to people's perceptions of the certainty, celerity, and severity of punishment (von Hirsch et al. 1999, p. 6). For example, if the actual certainty of arrest for burglary in a particular US state in a given year is 25 percent (law enforcement officials actually make 25 arrests for every 100 burglaries committed), people may perceive the certainty of arrest for burglary to be 25 percent or they may perceive it to be higher or lower than 25 percent.

Deterrence theorists since Beccaria and Bentham have assumed that there is a strong positive relationship between actual punishments and perceived punishments (Premise 1), and that assumption is important

for two reasons. First, 'to the extent that changes in actual penal policies do not alter potential offenders' beliefs about the likelihood or severity [or celerity] of punishment, they cannot generate any … deterrence' (von Hirsch et al. 1999, p. 6). Second, demonstration of a positive relationship between actual and perceived punishments is important to distinguish deterrence from other possible preventive consequences of punishment. To illustrate, suppose there was a strong negative relationship among US states between the actual certainty of imprisonment for auto theft and the auto theft rate. It may not be fear of legal punishment (deterrence) that accounts for the relatively low auto theft rates in states with a high certainty of imprisonment for auto theft; to the extent that imprisonment serves to limit the opportunities to commit auto theft (it is difficult to commit auto thefts in prison), the imprisonment of large numbers of auto thieves can cause a low auto theft rate through *incapacitation* and not fear of legal punishment. There would be evidence of deterrence only to the extent that (1) the people in states with a high actual certainty of imprisonment for auto theft perceived it to be high and (2) there were low auto theft rates in states with high actual and perceived certainty of imprisonment for auto theft.

4. *Additional Complexities*

An additional complexity is that the deterrent effects of legal punishment may be contingent. To begin with, the deterrent effects of a particular property of punishment—say, certainty—may be contingent on the levels of the other punishment properties—say celerity, and severity. For example, Tittle (1969, p. 47) reported in a study of the relationship among US states between the actual certainty and severity of imprisonment and rates for seven crime types (e.g., homicide) that the actual severity of imprisonment acted 'as a deterrent only when there was a high [actual] certainty of punishment.' Similarly, Grasmick and Bryjak (1980) reported in a study of the relationship among individuals between the perceived certainty and severity of punishment and self-reported crime that perceived certainty has stronger effects at higher levels of perceived severity and that perceived severity has stronger effects at higher levels of perceived certainty.

The deterrent effects of legal punishment also may be contingent on *extralegal punishments*, including self-imposed punishments (e.g., guilt, shame) and socially imposed punishments (e.g., divorce or employment loss). William and Hawkins (1986, p. 561) have indicated that extralegal punishment may result from legal punishment (e.g., an offender may experience loss of employment following imprisonment) and, thus, should be 'part of the … deterrence process.' However, extralegal punishment also may be inflicted independent of legal punishment. For example, Grasmick and Bursik (1990, p. 840) have argued that offenders may experience guilt or shame from committing a crime regardless of whether they are caught and legally punished. Hence, the deterrent effects of legal punishment may vary by the likelihood of experiencing swift and severe extralegal punishment for committing crime. For example, the deterrent effects of legal punishment may be greater among those who have little to risk in the way of extralegal punishment for crime because fear of legal punishment may be their only barrier to law-violating behavior. In that connection, Burkett and Ward (1993, pp. 128–9) found that those who believed marijuana use to be sinful were unlikely to contemplate engaging in the behavior; and, hence, the threat of legal punishment for marijuana use was irrelevant for them. Only those who did not consider marijuana use sinful were deterred 'because they appear[ed] to contemplate the act and, by reasonable inference, the probable consequences of their behavior.'

Policymakers need not worry that punishment does not deter all potential offenders as long as *enough* potential offenders are deterred. However, there is evidence that legal punishment sometimes is counterproductive by causing potential offenders to commit *more* crime. In relation to specific deterrence, Sherman and Smith (1992) reported that arrests for domestic assaults produced more subsequent assaults by unemployed offenders because they did not perceive police intervention as legitimate and 'displaced [their anger about being arrested] onto their present or future romantic partners or other citizens' (Sherman 1993, p. 465). According to Sherman (1993, p. 460), legal punishment provokes such 'defiance' (persistent or more serious law-violating behavior) to the extent that offenders (1) perceive the punishment as unfair, (2) are weakly bonded to other community members, (3) perceive the punishment as stigmatizing, and (4) deny the shame associated with it.

5. *Links between Deterrence Theory and other Theories*

The connection between contemporary deterrence theory and the Beccaria-Bentham belief that people calculate the pleasure and pain associated with behavior means that deterrence theory can be linked to other, more general, theories of human behavior. Akers (1990) has argued that deterrence theory is subsumable under more general *social learning* or *behavioral principles*. 'Threat of legal punishment is one source or indicator of aversive stimulus under the general concept of differential reinforcement (balance of rewarding and aversive stimuli)' (Akers 1990, p. 658). Differential reinforcement refers to the overall

balance of rewards and punishments for a behavior, including past, present, and future rewards and punishments. General deterrence is equivalent to observational/vicarious learning in social learning theory, which involves observations of other people and how their behavior is rewarded and punished; and specific deterrence is equivalent to experiential learning, which involves direct, personal experience with the rewards and punishments for a behavior (Stafford and Warr 1993).

Both deterrence and social learning theory can be linked to *rational choice theory* in that both address the question of why people in certain situations *decide* or *choose* to engage in particular behaviors (Akers 1990). However, the meaning of 'rational' is ambiguous. To some deterrence researchers, rationality implies conscious and deliberate calculations about the rewards and punishments for crime (e.g., Bachman et al. 1992). In contrast, others deny that overt mental calculations of any kind are necessarily involved (e.g., Casey and Scholz 1991). Moreover, rational choices are sometimes conceived as willful rather than determined (e.g., Bachman et al. 1992), but Cornish and Clarke (1986, p. 13) have claimed that the 'rational choice perspective is … neutral with respect to the free will-determinism debate.'

Disagreement about the meaning of the word 'rational' stems from the existence of different versions of rational choice theory: old rational choice theory and new rational choice theory, each with different assumptions about rationality (Felson, 1993). Old rational choice theory, as set forth by such utilitarian philosophers as Beccaria and Bentham assumed that people engage in conscious and deliberate cost/benefit analysis in order to maximize the rewards and minimize the punishments for their behavior. The new version of rational choice theory entails weaker assumptions about rationality—that people intuit the rewards and punishments for a behavior and that they are imperfect processors of information and consequently must behave within the limits of their abilities to pursue what they perceive as most satisfying (e.g., Cornish and Clarke 1986). The assumption of 'maximizing' is rejected and replaced by alternative assumptions, such as 'satisficing' and 'bounded' or 'situated rationality.' As for the issue of free will versus determinism, new rational choice theory posits that 'people make choices, but they cannot choose the choices available to them' (Felson 1986, p. 119).

Like new rational choice theory, deterrence theory does not require a strong assumption about rationality. In particular, crime need not be preconceived or carefully planned for legal punishments to deter, only that potential offenders choose between crime and conformity to the law on the basis of their estimates of the rewards and punishments involved. Also like new rational choice theory, deterrence theory does not require that people have free will. Indeed, a negative relationship between properties of legal punishment (e.g., the perceived certainty of punishment) and crime is as compatible with a strictly deterministic conception of human behavior as it is with free will (Gibbs 1975, p. 24).

See also: Bentham, Jeremy (1748–1832); Crime: Sociological Aspects; Crime, Sociology of; Criminal Justice, Ethics of: Aspects of Human Dignity; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Deterrence; Deterrence: Legal Perspectives; Imprisonment: Sociological Aspects; Law: Economics of its Public Enforcement; National Security Studies and War Potential of Nations; Prisons and Imprisonment; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Rational Choice in Politics; War: Causes and Patterns; Warfare in History

Bibliography

Akers R L 1990 Rational choice, deterrence, and social learning theory in criminology: The path not taken. *Journal of Criminal Law and Criminology* **81**: 653–76

Bachman R, Paternoster R, Ward S 1992 The rationality of sexual offending: Testing a deterrence/rational choice conception of sexual assault. *Law and Society Review* **26**: 343–72

Burkett S R, Ward D A 1993 A note on perceptual deterrence, religiously based moral condemnation, and social control. *Criminology* **31**: 119–34

Casey J T, Scholz J T 1991 Beyond deterrence: Behavioral decision theory and tax compliance. *Law and Society Review* **25**: 821–43

Cornish D, Clarke R 1986 Introduction. In: Cornish D, Clarke R (eds.) *The Reasoning Criminal: Rational Choice Perspectives on Offending*. Springer-Verlag, New York, pp. 1–16

Felson M 1986 Linking criminal choices, routine activities, informal control and criminal outcomes. In: Cornish D, Clarke R (eds.) *The Reasoning Criminal: Rational Choice Perspectives on Offending*. Springer-Verlag, New York, pp. 119–28

Felson M 1993 Book review. *American Journal of Sociology* **98**: 1497–9

Gibbs J P 1975 *Crime, Punishment, and Deterrence*. Elsevier, New York

Grasmick H G, Bryjak G J 1980 The deterrent effect of perceived severity of punishment. *Social Forces* **59**: 471–91

Grasmick H G, Bursik R J 1990 Conscience, significant others, and rational choice: Extending the deterrence model. *Law and Society Review* **24**: 837–61

Hirsch Von A, Bottoms A E, Burney E, Wikstrom P-O 1999 *Criminal Deterrence and Sentence Severity: An Analysis of Recent Research*. Hart Publishing, Oxford, UK

Sherman L W 1993 Defiance, deterrence, and irrelevance: A theory of the criminal sanction. *Journal of Research in Crime and Delinquency*. **30**: 445–73

Sherman L W, Smith D A 1992 Crime, punishment and stake in conformity: Legal and informal control of domestic violence. *American Sociological Review* **57**: 680–90

Stafford M C, Warr M 1993 A reconceptualization of general and specific deterrence. *Journal of Research in Crime and Delinquency*. **30**: 123–35

Tittle C R 1969 Crime rates and legal sanctions. *Social Problems* **16**: 408–23

Williams K R, Hawkins R 1986 Perceptual research on criminal deterrence: A critical review. *Law and Society Review* **20**: 545–72

M. C. Stafford and S. D. Goodrum

Deutsch, Karl W (1912–92)

Karl Wolfgang Deutsch belongs to the cohort of the 'brain drain' which, by emigrating in the 1930s from Europe, contributed to the development of the social sciences in the United States. Born in Prague on July 21, 1912, Deutsch escaped from the imprisoned continent in 1939, after having received a law degree from Charles University. Son of the first woman parliamentarian in Czechoslovakia, he moved from one privileged social stratum to another, debarking with a fellowship at Harvard University. He worked for the Office of Strategic Studies during World War Two, began teaching at MIT in 1945, then at Yale University in 1957 and, from there, taught at Harvard University until retiring in 1985. At that time, he was appointed as one of the directors of the Wissenschaftzentrum in Berlin. As is often the case with those who learn a new language after puberty, Karl Deutsch kept his distinctive German accent. He died at the age of 80 in Cambridge, MA, on November 1, 1992.

Karl Deutsch's contributions to the advancement of social sciences can be summed up in five domains:

(a) Building empirically grounded concepts and theories, such as nation-building, social mobilization, national and international integration, center–periphery, distribution of power, and understanding macrosocial processes. His definition of integration of a society is still in use in the literature: the capacity to receive and transmit information on wide ranges of different sectors with little delay. One of his writings appears retrospectively as predictive: 'Cracks in the monolith, possibilities and patterns of disintegration in totalitarian systems' (Deutsch 1954).

(b) Pioneering, cross-national research. Deutsch remains a leading figure of the first post-War generation of comparativists. In the late 1950s and 1960s, he did comparative research, using aggregate data at a time when other great comparativists, such as Gabriel Almond and Sidney Verba, preferred the survey method for comparing countries.

(c) Quantitative methodology. Karl Deutsch advanced the introduction of statistical analysis in political science. Such a vocation could be explained by the fact that, as the son of an optician, early in life, Deutsch considered the possibility of becoming one himself, and consequently studied mathematics. He did quantitative analysis on the size of government, growth of governments, foreign trade as a percentage of GDP, global modeling, two-way: channels of communication between élites and mass. Deutsch was in the forefront of the so-called 'data movement.' His contribution to *Quantitative Ecological Analysis in the Social Sciences* (Deutsch 1969) was considered as a research agenda.

(d) Work on communication theory, particularly cybernetics. He developed Wiener and Neumann's research, focusing on politics and society. The subtitle of *The Nerves of Government*, 'Political Communication and Control,' (Deutsch 1963) paraphrases Wiener's definition of cybernetics: 'Control and Communication in the Animal and Machine.'

(e) Bridges between disciplines, by importing in political science concepts and theories from various disciplines (anthropology, social psychology, economics, and biology). *Nationalism and Social Communication* (Deutsch 1953) is an eminently interdisciplinary work, where Deutsch's intimate knowledge of European history and geography appears in all chapters. Deutsch was a leading contributor to *Advances in the Social Sciences* 1900–80 (Deutsch et al. 1986).

Retrospectively, it can be said that Deutsch's most eminent book is his first, *Nationalism and Social Communication*: a*n Inquiry into the Foundation of Nationality* (1953). This is his most cited work, and has been so since its publication. The subject of this book is related to his personal experience. Born to a Sudeten German family in the Austro-Hungarian empire, a multinational state, 'Deutsch put his emotional and intellectual energy into scholarship focused upon nationalism and formation of large-scale political communities' (Merritt and Russett 1981, p. 6). As a teenager, Deutsch witnessed the collapse of three empires and the emergence of many new independent nation-states in Eastern Europe and the Balkans.

The success of this book can be explained by Deutsch's capacity to interpret statistical data with an intimate knowledge of the historical and social contexts. There are social scientists who write about countries they have only read about in books and who, in spite of their statistical virtuosity, too often offer superficial explanations. They should contemplate the following: 'We will always need the clinician who knows the problem from having seen many cases, who knows the countries, the cultures, the areas where the problem takes place, and who knows the particular syndromes. It is by equipping the clinician with the work of the technicians, by having clinical judgement and technological and scientific evidence side by side, that we are likely to make the most progress' (Deutsch 1980 p. 19).

In a survey of the most commonly cited political scientists, Deutsch's name ranked eighth between 1945 and 1960, fourth from 1960 to 1970, and tenth between 1970 and 1976. After 1980 his name still appears on the list of the three dozen most-cited political scientists. Between 1970 and 1979, his name was cited 1870 times. Between 1996 and 1999 he has been cited 60–90 times per year. Most of them concern the following three books: *Political Community at the International*

Level (Deutsch 1954), *Nationalism and Social Communication* (Deutsch 1953), and *The Nerves of Government* (Deutsch 1963).

It is interesting to note that Deutsch's two most successful books were conceived and written simultaneously, and were published only one year apart, in 1953 and 1954. Undoubtedly the fundamental ideas for these two works were already present in his mind when he escaped from the totalitarian regime. *The Nerves of Government* does not appear to be Deutsch's most popular book. It has been cited less frequently than the other two and, when it has been cited, the citations were in many cases perfunctory, the content not being commented upon. This book was published in 1963, at a time when Deutsch had already been at the front of the stage. Such perfunctory practice has been described by Robert Merton in his *Sociology of Science*.

The high 'mortality' rate of articles in political science, particularly in the field of international relations, is a well-known phenomenon. In several other disciplines, the life expectancy rate of articles is longer. Karl Deutsch published as author, co-author, editor, and co-editor a total of 16 books. In addition, between 1943 and 1980, he published 240 articles in academic journals and chapters in collective works, approximately half of them in collaboration with other scholars. According to the SSCI, which covers as sources only journals, excluding books, some of these 240 contributions were still being cited during the 1990–5 quinquennium. The content of some of Deutsch's articles has become chapters of his books, particularly in *The Nerves of Government*. Other articles, among the best *Towards an Inventory of Basic Trends and Patterns in Comparative and International Politics* (1960) and *Social Mobilization and Political Development* (1961), have become standard information and common knowledge, and, consequently, do not need to be cited. What had been previously innovative has been superseded by more recent studies based on more precise data, analyzed with more refined methods. When writing about microbes, Pasteur need not be cited. When using social indicators, Karl Deutsch, Gabriel Almond, and Daniel Lerner no longer need to be cited. The concept of social mobilization already belongs to the common patrimony of the social sciences.

In the late 1960s and 1970s, Yale's Department of Political Science was one of the most attractive for graduate students, many of whom converged in Deutsch's seminars. One of these seminars inspired the conception of an important book entitled *World Handbook of Political and Social Indicators*, which users of quantitative data had named 'the Yale bible,' by Bruce M. Russett in collaboration with H. R. Alker, K. Deutsch, and H. Lasswell (Russett et al. 1964).

The following are among Deutsch's intellectual progeny: Hayward Alker, Jorge Dominguez, William Foltz, Michael Hudson, Peter Katzenstein, Manfred Kochen, Arendt Lijphart, Richard Merritt, Donald Puchala, Bruce Russett, Charles Taylor, these and others recognized as outstanding political scientists. They honored their mentor in a collective book *From National Development to Global Community*, edited by Richard L. Merritt and Bruce M. Russett (1981).

Stein Rokkan's tribute to Deutsch's intellectual leadership deserves to be cited at length: 'Karl Deutsch taught my generation of social scientists to develop models and to test them against the data of history. My own early research was heavily influenced by Deutsch's ideas. My later work on dimensions of state formation and nation-building was directly inspired by Deutsch's path-breaking study *Nationalism and Social Communication*; he taught me to look out for the decisive characteristics of center-building networks and to study the functions of the printed media' (Rokkan 1981 p. 70).

Deutsch was an interdisciplinary scholar, comfortable in many fields. In his latter days he became interested in 'the conditions favoring major advances in the social sciences.' In 1986, he co-edited a volume focusing on this subject in collaboration with Andrew Markovits and John Platt (Deutsch et al. 1986).

Those who encountered Deutsch were impressed by his encyclopedic erudition. In his book *Nationalism and Social Communication* (Deutsch 1953) some 2000 authors are mentioned. His historical knowledge appears in a particular light in an article published in 1945 on 'Anti-semitic ideas in the Middle Ages,' based on the reading of very old literature (Deutsch 1945). This article merits to be included in any modern anthology of historical sociology.

Soon after the publication of *The Limits to Growth* by the Club of Rome in 1972, a new trend appeared in political science, imitating economists, demographers, and ecologists. Karl Deutsch had alerted the specialists: 'Computers can assist thought, but not replace it. Mathematical thought produced new ideas and techniques, such as game theory and coalition theory, scaling, latent structure analysis, matrix analysis and a succession of increasingly sophisticated models of social and political structures and processes … Some of us have tried to deal with this by over-specialization, risking to fall into the old trap of learning more and more about less and less' (Deutsch 1978). Nevertheless, in the 1980s, he was seduced by the siren of 'world modeling.' Under his directorship at the Wissenschaftzentrum in Berlin, modeling became a priority. The virtue of Deutsch's effort in Berlin was to combine mathematical modeling and solid empirical referent. Abstract mathematical modeling has attracted the attention of a limited number of political scientists outside the circle of the builders of reductionist models.

World trade was one of Deutsch's favored topics, which he discussed in several writings. He measured the ratio of foreign trade to national income for almost

half a century. He presented an impressive amount of statistical data demonstrating a decline in international trade between 1913 and 1957 (Deutsch and Eckstein 1961). But this period was marked by World War One, the Great Depression, protectionism, and autarchy practiced by several great powers in the 1930s, World War Two, and the post-War reconstruction of the devastated European countries and Japan. In the 1950s, only 6 percent of the American GNP depended on foreign trade. Deutsch did not explicitly predict a continuing decline of international transaction. More recent developments do not invalidate Deutsch's historical observation.

One of Deutsch's main ideas was the importance of intertwining economies for political integration. This thesis is still validated: half of France's foreign trade takes place with Germany, and reciprocally. Economic integration has preceded political integration of the European Union, even for Britain.

Since the publication of *Nationalism and Social Communication*, the nations of Western Europe, with the exception of Greece and Ireland, have experienced a rapid decline of nationalist tendencies, reversing a historical trend (Dogan 1993), at a moment when, in contrast, many developing countries in Asia, the Middle East, and Africa have been mobilized by movements of national independence, developing an exacerbated ethnic nationalism. The concepts designed by Deutsch and other scholars—social communication, rapidity of assimilation, rate of mobilization, spread of national consciousness, population clusters, linguistic cleavages, center–periphery—have been transplanted from the continent to which they were first applied to many new nation-states, where Deutsch's theories are more relevant.

Karl Deutsch's professional career was one of exceptional brilliance. He was elected President of the American Political Science Association in 1969 and of the International Political Science Association (IPSA) in 1976, playing an important role in the development of the research committees within IPSA. In collaboration with Richard Merritt, as program director, he organized the IPSA World Congress in Moscow in 1979, building a bridge between East and West for political scientists. A member of the American Academy of Sciences, he lectured in many European universities. Deutsch received honorary doctorates from seven American and European universities and the Commander's Cross Order of Merit from the German Federal Republic.

See also: Communication and Democracy; Comparative Studies: Method and Design; Nationalism: General; Quantification in the History of the Social Sciences; Rokkan, Stein (1921–79); State Formation

Bibliography

Deutsch K W 1945 Anti-semitic ideas in the Middle Ages. *Journal of History of Ideas* **6**: 239–51

Deutsch K W 1953 *Nationalism and Social Communication: An Inquiry into the Foundations of Nationality*. MIT Press, Cambridge, MA; 2nd edn., 1966

Deutsch K W 1954 *Political Community at the International Level, Problems of Definition and Measurement*. Doubleday, Garden City, NY

Deutsch K W 1957 *Political Community and the North Atlantic Area*. Princeton University Press, Princeton, NJ (in collaboration with seven other scholars)

Deutsch K W 1963 *The Nerves of Government: Models of Political Communication and Control*. Free Press, New York; 2nd edn., 1966

Deutsch K W 1969 On methodological problems of quantitative research. In: Dogan M, Rokkan S (eds.) *Quantitative Ecological Analysis in the Social Sciences*. MIT Press, Cambridge, MA, pp. 19–39

Deutsch K W 1978 *Major Changes in Political Science 1952–1977*. Research paper. Internationales Institut für Vergleichende Gesellschaftsforschung, Berlin, p. 28

Deutsch K W 1980 On the utility of indicator systems. In: Taylor C L (ed.) *Indicator Systems for Political Economic and Social Analysis*. Oelgeschlager, Gunn, and Hain, Cambridge, MA, pp. 11–23

Deutsch K W, Alker H, Stoetzel A (eds.) 1973 *Mathematical Approaches to Politics*. Jossey-Bass, San Francisco

Deutsch K W, Eckstein A 1961 National industrialization and the declining share of the international economic sector 1890–1959. *World Politics* **13**: 267–99

Deutsch K W, Edinger L 1959 *Germany Rejoins the Powers: Mass Opinion, Interest Groups and Elites in Contemporary German Foreign Policy*. Stanford University Press, Stanford, CA

Deutsch K W, Markovits A S, Platt J 1986 *Advances in the Social Sciences 1900–1980*. University Press of America, New York

Deutsch K W, Platt J, Senghaas D 1971 Conditions favoring major advances in the social sciences. *Science* **171**: 450–9

Dogan M 1993 Comparing the decline of nationalisms in Western Europe: the generational dynamic. *International Social Science Journal* **136**: 177–98

Dogan M, Rokkan S 1969 *Quantitative Ecological Analysis in the Social Sciences*. MIT Press, Cambridge, MA

Jodice D H, Taylor C L, Deutsch K W 1980 *Cumulation in Social Science, Data Archiving. A study of the Impact of the Two Handbooks of Political and Social Indicators*. Wissenschaftzentrum, Berlin

Merritt L R, Russett B M (eds.) 1981 *From National Development to Global Community*. Allen and Irwin, London

Rokkan S 1981 Variations within Western Europe. In: Merritt L R, Russett B M (eds.) *From National Development to Global Community*. Allen and Irwin, London, pp. 115–44

Russett B M, Alker H R, Deutsch K, Lasswell H 1964 *World Handbook of Political and Social Indicators*. Yale University Press, New Haven, CT

Taylor C L 1968 *Aggregate Data Analysis: Political and Social Indicators in Cross-national Research*. Mouton, Paris

M. Dogan

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Dev

Development and the State

Even though making war and enforcing internal order are the state's classic roles, political survival and internal peace increasingly depend on economic success. States are, therefore, inevitably involved in development. The term 'development' has been used in the post-World War II period as shorthand for the process of socioeconomic change that has shaped contemporary societies, particularly those of Africa, Asia, and Latin America. Development has meant the transformation of agriculture, the growth of industry, and a host of accompanying social, political, and economic changes. The question is: 'When and how does the behavior of the state facilitate or impede development?'

1. Historical Perspectives on the State's Role in Economic Transformation

Calling the poor nations of Asia, Africa, and Latin America 'developing countries' implies that they are undergoing economic and social transformations that are variations of those already experienced by the industrialized countries of the North. Understanding the role of the state in earlier periods of industrialization is, therefore, important to thinking about development and the state.

Two streams of thinking are worth highlighting. First, there is the perspective of economic historians, such as Douglass North (1981), who see the provision of predictable norms and laws, particularly those which define and protect property rights, as the state's primary contribution to development. The second view sees states as actively constructing new institutions to facilitate trade and investment and mobilizing resources in ways that accelerate the growth of new productive capacities. This view is epitomized in Karl Polanyi's (1957 [1944], p. 140) classic statement: 'The road to the free market [in England] was opened and kept open by an enormous increase in continuous, centrally organized and controlled interventionism.'

Alexander Gerschenkron (1962), looking at 'late developers' in Europe, reinforced and extended Polanyi's analysis by suggesting that countries like Germany and Russia faced a more difficult economic problem than England. They had to 'catch up.' By the time they began to industrialize, existing production technology required a scale that exceeded their societal capacities. Lacking both individual capitalists able to assume risks at the scale required by modern technology or private institutions able to spread risk, the state had to serve as investment banker or bear the risks itself, bringing together the necessary funds and encouraging their application in productive activities.

2. Analyzing the State's Role in Third World Development

Perspectives, like those of Polanyi and Gerschenkron, in which development was not the spontaneous result of the activities of private economic actors but had to be explicitly fostered by the state, resonated in the Third World. In Asia, Africa, and Latin America, the gap between local capacities and the institutional scale required to be competitive with firms in the industrialized North was much bigger than the gap Russia and Germany had faced a century earlier. Local elites were too often wedded to maintaining agrarian structures that preserved their privilege at the expense of national productivity. The fact that the United States and Europe were just emerging from a period in which substantial increases in state intervention, first to combat the Great Depression and then to mobilize for World War II, made this perspective even more plausible.

2.1 Development Planning, Import Substitution, and the Entrepreneurial State

The conviction that Third World countries could not develop unless they managed to escape their traditional dependence on exports of agricultural products and raw materials was a key impetus to increased state involvement. Manufacturing products locally rather than importing them was seen as a central part of development and emerging local manufacturers needed protection from established rich-country firms if they were going to succeed at this process of 'import-substituting industrialization.' They also needed public investments in energy production and transportation. Deciding which of these investments made sense required planning. Private investors also faced a collective action problem. Investment in manufacturing would make more sense if other local entrepreneurs were making complementary investments that would provide needed inputs. Active state involvement was a way of giving entrepreneurs confidence that they were part of a general, mutually-reinforcing set of investment decisions.

Without advocating expansion of the state's role, Albert Hirschman was one of the most creative contributors to thinking on these issues. Hirschman carried Gerschenkron's argument a step further, arguing that a dearth of entrepreneurship in the simple sense of 'the perception of investment opportunities and transformation into actual investments' (Hirschman 1958, p. 35) stood in the way of 'late late industrialization' more than the lack of capital. The state could help induce private capital to become more entrepreneurial by providing incentives that made decisions harder to avoid. In Hirschman's view, inventing a national project of entrepreneurship is the key to development and the state can be an important participant in the process of invention.

2.2 *Neo-liberal Critiques of State Intervention*

Import-substituting industrialization and the state involvement that went with it resulted in impressive rates of economic growth during the 1950s and 1960s in countries ranging from India to Brazil, but by the end of the 1970s development strategies in Africa, Latin America, and South Asia were in disarray. Imports were growing faster than exports, creating balance of payments problems and state expenditures were increasing more rapidly than revenues, creating fiscal problems. One of the responses to this disarray was a renewed interest in the state on the part of neoclassical political economists.

Like Marxists, neoclassical political economists saw the state as the instrument of special interests rather than universal ones. Unlike Marxists, they applied the utilitarian logic of individual maximization to actors within the state itself. They saw the state extracting rents and giving entrepreneurs the wrong incentives rather than solving collective action problems and inducing more entrepreneurship. Theorists like Anne Krueger (1974) applied this logic to strategies of import-substituting industrialization, arguing that the desire to gain protection from foreign competition led entrepreneurs to focus on the unproductive quest for state favors rather than on competitiveness. Studies like Robert Bates' (1981) analysis of the perverse effects of state intervention on the development of African agriculture helped build the case that there were serious problems with earlier optimistic visions of the state's role in development.

These critiques were on the mark in many ways. Development planning had indeed presumed an unrealistically benign state, ignoring the self-seeking behavior of state elites. Corruption was rife in developing states. Entrepreneurs in industries like textiles did indeed continue to demand that indulgent states provide them with protection long after they should have become mature and internationally competitive. Nonetheless, these neo-liberal theorists lacked good answers as to how developing countries could overcome the weaknesses of their private entrepreneurial classes. Their theories also failed to acknowledge that there were important variations, even among poor countries, in the behavior and capacities of states.

3. *The Developmental State*

Some states fit the negative neo-liberal image. They are predatory states: aggregations of self-interested elites that extract society's surplus wealth and income, use it for their own benefit, and provide so little in the way of collective goods in return that development becomes impossible. Zaire under the rule of Mobutu in the 1970s and 1980s was a classic example. Under Mobutu, the income of Zaire's citizens shrank at over 2 percent a year, while state officials acquired multiple European mansions and Swiss bank accounts. For most states, however, the neo-liberal view is partial and therefore misleading. The state apparatus may extract, sometimes corruptly, but it also provides essential collective goods without which development could not proceed.

Most interesting is the subset of states that are able to foster long-term entrepreneurial perspectives among private elites, help solve collective action problems, and supply essential public investments in education and infrastructure. The elites that control these states may not be immune from using society's surplus for personal ends, but the nations they govern enjoy faster rates of development because of the collective goods they supply. This category of 'developmental states' is best illustrated by looking at East Asia from the 1950s through the 1980s.

3.1 *The East Asian 'Miracles'*

Japan's successful challenge of the industrial supremacy of Europe and the United States in the 1950s and 1960s was followed by the developmental success of Korea, Taiwan, and Singapore. These countries grew six times faster that the advanced industrial countries had grown in the nineteenth century during the industrial revolution. They are unique in the Third World in the degree to which they have managed to change to their position in the world economic hierarchy. These states are considered developmental, not just because the nations they governed grew, but because the transformation involved in that growth was directly related to the behavior of the state.

3.2 *The Analytical Characteristics of the 'Developmental State'*

Starting with Johnson's (1982) analysis of the Japanese case, a variety of analysts have tried to specify what makes states developmental (see, for example, Akyüz 1999, Amsden 1989, Chang 1994, Wade 1990). The debate continues, but there are surprising elements of

consensus that extend into mainstream development economics (see Meier and Rauch 2000, Chap. 9) and the international policy community (see World Bank 1997).

One key structural feature of developmental states is a capable, coherent economic bureaucracy closely connected to, but still independent of, the business community. The roots of capable bureaucracies lie, in turn, in mechanisms familiar to students of Max Weber, such as meritocratic recruitment and public service careers that offer long-term rewards commensurate with those available in the private sector. These simple indicators predict more rapid economic growth, not just in East Asia, but also across a broad range of developing countries. The qualities that allow such bureaucracies to develop close ties to private elites without being 'captured' by those elites are harder to specify. Nonetheless, the success of developmental states depends on their ability to maintain autonomy from private elites while simultaneously developing close ties to them (Evans 1995).

The policies of developmental states mirrored the behavior Gerschenkron and Hirschman argued was needed in order to promote late industrialization. Investments in essential modern infrastructure, such as telecommunications, helped set the stage for industrial growth. Entrepreneurs were induced to make long-term investments that raised national productivity. Subsidies and selective protection against external competition led to the creation of industries like semiconductors, steel, and shipbuilding that private entrepreneurs would never have embarked on without state support. Exports, overall output, and real wages all grew at unprecedented rates. Public investments made rapid rises in education levels possible and health measures, such as infant mortality levels, improved along with education.

4. *Contemporary Issues and Future Directions*

Current thinking about development and the state has been thrown into question from at least three different sources. The recent evolution of the global political economy has brought into question the capability of individual states to shape developmental trajectories. At the same time, the efficacy of the developmental state has been brought into question by the East Asian financial crisis of 1997–8. In an even more fundamental challenge, the validity of traditional strategies of development is increasingly contested. In combination, this questioning suggests a redefinition of the role of the developmental state.

4.1 *Globalization and the Role of the State*

The reorganization of production, and more crucially finance, around global networks, in combination with the increasing hegemony of market-oriented Anglo-Saxon ideology and legal norms, makes traditional industry-focused state strategies more difficult, even when these strategies combine import substitution with export promotion, as in the case of developmental states. Even though global organizations, like the World Trade Organization, continue to depend on the institutional capacities of individual states to implement global governance rules, the leeway that states have in making their own rules to increase local entrepreneurship has been restricted. Even developmental states have felt the effects of globalization.

4.2 *The Crisis of the Developmental State*

The dominant tenor of first reactions to the 1997–8 Asian financial crisis was that either globalization has turned the developmental state into an anachronism or it was a hoax to begin with. More sober reflection produced a different diagnosis. One of the lessons of the crisis was that states cannot afford to relinquish controls over international financial flows before constructing appropriate mechanisms for regulating national financial markets. Abandoning its previous role, not clinging to it, was the developmental state's failing.

This failing did not result simply from the pressures of globalization. It also reflected a change in the politics and policies of the developmental state itself. Having achieved impressive gains in global markets, local corporate elites were less likely to believe that a coherent, capable state bureaucracy was relevant to their success. Global ideology favoring the dismantling of the state's regulatory capability also became more popular within the state apparatus. Even though financial stability and growth were restored with remarkable rapidity, the crisis revealed that, precisely because of its economic success, the state's alliance with private capital was no longer likely to produce the same developmental results. At the same time, the 'accumulation-oriented' definition of development that had formed the basis of this alliance was increasingly criticized.

4.3 *The Contested Definition of Development*

The strategy of the developmental state, like the strategy of import-substituting industrialization before it, was built on the premise that investment in local manufacturing would produce jobs and that the incomes from these jobs would generate widespread increases in well-being. There have always been critics of this view, but recently they have become more vociferous (e.g., Escobar 1995, Sachs 1992). Critiques argue that as manufacturing becomes more capital intensive, manufacturing investment consumes a disproportionate share of the social surplus relative to the

number of jobs that result. Even when the money incomes of some ordinary workers do increase, improved overall well-being is not necessarily the result. Their ownership of more consumer goods does not compensate for the environmental degradation, increasing inequality, and destruction of intangible social and cultural assets that accompany the accumulation of productive capacity. For too large a proportion of the citizens of poor countries, the new global cornucopia of heavily advertised consumer goods remains beyond economic reach but still devalues prior cultural and social satisfactions. While such perspectives, which ignore the benefits of investment in productive capacity, are one-sided, it is still getting harder to defend development policies premised on the idea that the best way to promote well-being is simply to induce entrepreneurs to invest more.

4.4 Redefining the State's 'Developmental' Role

If development is really about well-being rather than about accumulating industrial capital, then what is the role of the state? A 'well-being-centered' redefinition of development implies a more aggressive focus on the state's traditional role in delivering collective goods and services, like health, education, clean air, water, and sewage disposal, and in preserving the collective environment, both natural and cultural. The kinds of state behavior and structures required by this redefined developmental role are surprisingly consistent with early thinking. Hirschman's notion of inducing decision-making among private economic agents becomes an even bigger challenge if the focus is on the provision and preservation of collective goods. Finding ways to convince entrepreneurs that they should invest in clean production techniques or take into account the social impact of investment decisions requires much more skill and capacity than helping them increase their profits through subsidies and protection. The greater magnitude of the challenge makes the emphasis on a capable, coherent state bureaucracy, which is central to thinking about the developmental state, more important than ever.

The biggest single change implied by redefinition is already foreshadowed by the crisis of the developmental state. State-society relations dominated by close ties to industrial elites at the expense of other social groups will provide neither the political will nor the societal collaboration necessary to execute a 'well-being-oriented' strategy of development. A more broadly defined set of state-society relations is imperative. There are some studies of the role of more broadly-based state-society ties in promoting development (e.g., Evans 1996, Heller 1999), but extending and deepening this kind of analysis is the most obvious challenge confronting contemporary work on development and the state.

See also: Development, Economics of; Development: Socioeconomic Aspects; Globalization: Political Aspects; Multinational Corporations; Nations and Nation-states in History; Nation-states, Nationalism, and Gender; State, History of; State, Sociology of the

Bibliography

Akyüz Y (ed.) 1999 *East Asian Development: New Perspectives*. Frank Cass, London

Amsden A 1989 *Asia's Next Giant: South Korea and Late Industrialization*. Oxford University Press, New York

Bates R H 1981 *Markets and States in Tropical Africa: The Political Basis of Agricultural Policies*. University of California Press, Berkeley, CA

Chang H-J 1994 *The Political Economy of Industrial Policy*. Macmillan, London

Escobar A 1995 *Encountering Development: The Making and Unmaking of the Third World*. Princeton University Press, Princeton, NJ

Evans P B 1995 *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press, Princeton, NJ

Evans P B (ed.) 1996 Government action, social capital and development: Creating synergy across the public-private divide. Special section. *World Development* **24**(6): 1033–1132

Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Belknap, Cambridge, MA

Heller P 1999 *The Labor of Development: Workers and the Transformation of Capitalism in Kerala, India*. Cornell University Press, Ithaca, NY

Hirschman A 1958 *The Strategy of Economic Development*. Yale University Press, New Haven, CT

Johnson C 1982 *MITI and the Japanese Miracle: The Growth of Industrial Policy, 1925–1975*. Stanford University Press, Stanford, CA

Krueger A O 1974 The political economy of the rent-seeking society. *American Economic Review* **64**(3): 291–303

Meier G M, Rauch J E 2000 *Leading Issues in Economic Development*, 7th edn. Oxford University Press, New York

North D C 1981 *Structure and Change in Economic History*. Norton, New York

Polanyi K 1957 [1944] *The Great Transformation: The Political and Economic Origins of our Time*. Beacon Press, Boston

Sachs W (ed.) 1992 *The Development Dictionary: A Guide to Knowledge as Power*. Zed Books, London

Wade R 1990 *Governing the Market: Economic Theory and the Role of Government in Taiwan's Industrialization*. Princeton University Press, Princeton, NJ

World Bank (IBRD) 1997 *World Development Report: The state in a Changing World*. Oxford University Press, New York

P. B. Evans

Development and Urbanization

Economists, sociologists, political scientists, and anthropologists as well as demographers and planners have long theorized a direct connection between

national development, understood economically and/or politically, and the growth of cities. But there is limited consensus about the nature and direction of the relationship, the desirability of analytically linking these processes, and how to define both terms. Reigning ideas have shifted in the face of new theoretical paradigms and as the territorial scope and essential character of markets, states, and cities themselves transform. Unresolved issues include which urban dynamics most articulate with developmental patterns, and how to identify the most appropriate unit of analysis for understanding these articulations.

1. Historical Antecedents

Most scholarship on this subject traces its intellectual origins to writings by nineteenth and early twentieth-century sociologists and economic historians who sought to account for the development of capitalism, the nation-state, and social modernization in early Europe, and who posited a direct and positive role for cities in these fundamental transformations (see *State Formation*; *Modernization*, *Sociological Theories of*; *Capitalism*).

1.1 Cities and Development in the Early Modern World

In early modern Europe cities emerged from and fueled processes of capital accumulation, by virtue of their role as exchange nodes in local, national, and international trade (Pirenne 1936, Weber 1927). Their growth also was tied to the demise of feudal or absolutist orders (Weber 1958) and the rise of the modern nation-state (Tilly 1975). By the late nineteenth century, cities also served as the sites in which the social relations of modernity materialized (Durkheim 1933, Simmel 1950), in turn transforming markets and states.

1.2 Twentieth-century Developments

In the twentieth century when nationalism, war, empire, trade, democracy, and fascism grabbed European scholarly attention, their US counterparts adopted the concern with cities. This was a time when American society demographically tipped its balance to become more urban than rural, and when new waves of postwar black migration combined with renewed streams of international migration to spark interest in the social relations of urban life. Initial research concentrated on the social process and experience within cities, ranging from ghetto life, immigrant social clubs, and gang organization to criminality. Few initially paid attention to the larger political and economic processes that engendered patterns of urbanization in the US, as occurred in Europe; and even fewer tried to link urbanism or urbanization to patterns of national development, understood either politically or economically. It was only when US sociologists began studying cities comparatively and historically, starting in the 1950s, that debates over the relationship between development and urbanization garnered a central place in the social sciences. When the March 1955 (Vol. 60, No. 5) edition of the *American Journal of Sociology* published an entire volume on 'World Urbanism,' the importance of these themes was firmly established.

The desire to examine other parts of the world was partly grounded in a passion for testing ideas drawn from the European-derived grand theoretical narratives of sociology. Practically the only counterfactual cases available for confirming prevailing propositions about the connection between urbanization and developmental progress—whether understood in terms of democratic state formation, capitalist development, and/or 'modern' social values—were the countries of the so-called 'underdeveloped' world of Latin America, East Asia, and Africa. These were countries of the world racked with poverty, limited industrialization, and governed primarily under oligarchic pacts or by despotic or authoritarian governments (see *Authoritarianism*). To the extent that their urbanization patterns also differed across country contexts, scholars were inspired to revisit general claims about development and urbanization, as well as to ascertain how these two processes were articulated in practice.

Working under the influence of the Chicago School, initial scholarship on what were then called Third World cities focused primarily on the folk-urban continuum and urbanization as a social process, including the ways that urbanization produced individualism and severed kinship bases of social organization (see for example Hauser 1957, Redfield 1941, 1953). Over time, interest grew in urbanization as a demographic and spatial process, with scholars examining patterns of population concentration and the growth of cities in a wide variety of comparative and historical contexts (Berry 1973, Hauser and Schnore 1965), occasionally linking them to general propositions about national development. In addition to Hauser and Schnore's seminal *The Study of Urbanization* (1965), the Chicago-based journal *Economic Development and Cultural Change* served as an important outlet for new ideas on the topic.

2. Twentieth-century Dilemmas: Questioning the Nature and Direction of Causality

It was one thing to quantify urban population patterns across time and place (Davis 1969, Hoyt 1962), to identify the extent to which urbanism, as a social

attribute, materialized in different localities around the world (Breese 1966, Mangin 1970), and to elucidate growth processes in these cities (Gugler 1978, McGee 1971). American social scientists were quite good at these tasks, especially the first; and their successes helped catapult the field of demography onto the disciplinary map in the 1960s and 1970s (see for example Goldscheider 1971, Goldstein and Sly 1979). It was quite another to demonstrate a direct relationship between urbanization and patterns of economic development, especially within the confines of the nation-state, the unit that most social scientists were then using for assessing such issues. It was on these latter counts that scholars were least successful, and that longstanding propositions about the positive relationships between urbanization and national development began to generate widespread skepticism.

As Third World cities burgeoned over the 1960s and 1970s, scholars faced the fact that even with high urbanization rates, developmental gains in most of the Third World remained minimal, not just in comparison to the modal patterns established in Europe and the United States, but also as measured in per capita income, GNP, and practically all other standard macroeconomic indicators (Friedmann 1967). With dependency theory taking center stage at about the same period of time, certain attributes of both urbanization and national development were being traced to colonial or mercantile relations that spanned national boundaries rather than to political, social, and economic conditions within the formal confines of the nation-state (see *Dependency Theory*). These transnational relations modified urbanization and development processes, at least when compared to early modern Europe, even as they raised new questions about the best unit of analysis to study the nature and direction of the proposed relationship. Did urbanization husband development, or vice versa; and was it possible that in the late developmental context the relationship between the two sets of processes was more likely to be negative than positive, with one impeding rather than facilitating the other? The answers to these questions shifted over time.

2.1 Development and Urbanization circa 1950: A Positive Synergy

The predominant argument in the 1950s and early 1960s was that the plight of developing nations and their low rates of urbanization owed to their 'backward' social and economic nature, which not only explained why industrialization was minimal, but also why tradition-bound peasants demographically outbalanced 'modern' urban folk. Analytical attention was directed towards how to facilitate one so the other would follow suit, a normative preoccupation pervasive in American social science and quite consistent with US foreign policy goals of the times.

There were disagreements about which domain deserved intervention first. Some urbanists argued that the best way to jump-start Third World development was to shift population and resources out of the countryside and into the city, where industry could expand and modern values flourish (Herrick 1965, Field 1970). Others called for more foreign aid or national investment to buttress urban economies, primarily through industrialization (Richardson 1973), in the hopes that such macroeconomic policies would stimulate employment opportunities and perhaps even produce individuals inculcated with rationality and enough achievement drive to sustain further urban and national gains (see *Industrialization*). Despite policy disagreements, both schools of thought expressed enough faith in the idea of a positive relationship between development and urbanization to keep it alive, albeit embodied in a variety of paradigmatic forms.

Over the 1960s and 1970s, one particularly popular variant was demographic, as those interested in urbanization as a social process looked for linkages between urban growth and national development in fertility patterns as well as family structure and employment. During this period ideas about the demographic transition, at times recast in terms of the 'urban transition' (Friedmann 1975), generated considerable interest among economists and sociologists, who developed expert quantitative skills that sustained a field of research broadly understood as 'population studies.' Those less preoccupied with the social or economic behavior of individuals and more interested in overall patterns of city growth and distribution were more likely to concern themselves with ideas about the proper balance between urban and rural populations (El-Shakhs 1974, Friedmann and Alonso 1964).

During this period, scholars who assumed that the US served as a model of successful economic development paid considerable attention to the extent to which urban systems in developing countries achieved a lognormal rank-size distribution, as opposed to being dominated by one or two large 'primate' cities (Berry 1961), basing their interest on the assumption that rank-size lognormality correlated with economic development. In counterattack, these ideas were criticized as inherently Western-biased, if not drawn directly from the American experience, and thus inapplicable in the Third World (El-Shakhs 1972, McGreevey 1971). While there was no resolution on this contentious issue, in the process the concept of 'over-urbanization' gathered widespread attention, since a city could only be 'too' urbanized if it grew beyond its expected and economically efficient size, which often was calculated on the basis of a rank-size distribution logic.

To the extent that most Third World countries fell into the latter camp, with their cities ballooning in size beyond their fiscal and infrastructural capacities, scholars began to study the internal social and demo-

graphic processes that fueled patterns of over-urbanization, and what Michael Lipton (1977) called 'urban bias,' including rural-urban migration and the national investment decisions that overprivileged a few urban centers at the expense of struggling provincial towns and the impoverished countryside. Still, concerns about city systems died hard, rematerializing in slightly new packaging in the 1980s in the form of arguments about the importance of 'secondary cities' to sustaining national development (Hardoy and Satterwaithe 1986, Rondinelli 1983). This claim, like the rank-size distribution argument, was built on the understanding that (smaller) size and (more equal) distribution of cities mattered; and both arguments confirmed the preoccupation with a relationship between patterns of urbanization and development that had initiated and sustained the field.

But far from merely confirming a connection between cities and national development, these findings laid the foundation for an immanent critique of long-standing claims. Indeed, to the extent that the burgeoning growth of Third World cities was soon seen as an obstacle to economic well-being, there was now evidence that a positive relationship between urbanization and development was effectively reversed on substantive grounds, even if analytically the connection remained. To be sure, there were scholars who conceived of large cities as economically efficient localizers of economies of scale, and they continued to claim a positive relationship between large-scale urbanization and development (Richardson 1973). But by and large, promoters of this idea were arguing against a tide of contrary evidence.

2.2 Development and Urbanization, circa 1975: A Destructive Relationship

The ascendance of ideas about a negative relationship between urbanization and development owed partly to the fact that between 1950 and 1975 conditions in the developing world changed dramatically. This was the period when most Third World cities began to grow enormously, outpacing the capacities of their governments to invest in infrastructure or sustain industrial output in ways that contributed to national development. With large cities peopled by new migrants and un- or underemployed folk with limited education and skills for a newly industrializing economy, they soon were seen as a drain on national coffers. The increasing urban inequality and economic polarization accompanying these trends also sustained urban social movements and overall political coalitions that themselves limited national developmental prospects, by bringing to power governments with restrictions on foreign investment, populist economic policies, and/or protectionist measures that frequently undermined short-term efficiency goals. These social and political problems became so salient in the urban landscape of the developing world that eventually scholars of Third World cities turned their attention directly to them (see Abu-Lughod and Hay 1977, Gugler 1988, Mangin 1970), partially eclipsing the concern with national development. Among the most documented issues were urban housing and employment scarcities, rural-urban migration flows, illegal settlements and urban 'marginality,' and migrant or informal sector politics (Cornelius 1973, Drakakis-Smith 1987, Perlman 1976).

This is not to say that all scholars who looked internally to cities over the 1970s and 1980s failed to examine the larger relationships between urbanization and national development that had consumed their predecessors. Economists at the World Bank and other macroeconomic policymakers loyal to free market ideals continued to study urban economies of scale and the ways urban growth interfaced with industrialization, still seen as the principal route towards prosperity; although they too concerned themselves with questions about housing, illegal settlements, and urban service scarcities more generally (Linn 1983). Likewise, but on an entirely different front, sociologists and political scientists began to examine the national political conditions that made these urban problems so pervasive (Eckstein 1977, Gilbert and Ward 1985, Rabinowitz 1973, Roberts 1978). But this also meant that urban scholars were focusing as much on political as economic developments, linking urbanization patterns to national politics as much as markets (see *Populism*).

When dependency theory kicked in as a dominant paradigm among both resident and foreign scholars of the developing world over the 1970s and 1980s, fewer urbanists began to pose questions about the relations between urbanization and national development unless they were framed within a focus on international economic conditions (Castells and Vélez 1971, Portes and Walton 1981, Walton 1985); and even fewer saw the articulation of urban and national development in positive terms, as had their predecessors. If anything, when urbanization was linked to capital accumulation, the social effects were seen as relatively negative (Armstrong and McGee 1985). By the late 1980s, the main idea in good currency was that capitalist development on a global scale produced and/or reinforced mercantilist relations that fueled the growth of cities, even as they disadvantaged the countryside or the nation as a whole. Scholars focused not just on the ways that unequal, semi-imperial relations between core and periphery brought a few large cities, themselves parasitic with respect to their hinterlands. They also examined the ways these relations produced a form of 'internal colonialism' predicated on extreme income and investment polarization between city and countryside, which further fueled migration, contributed to over-urbanization, and reinforced income and other inequalities within cities themselves (see Timberlake 1985). With world-systems analysts turn-

ing full attention to cities in the 1980s (Smith 1996), claims for a substantively negative relationship between urbanization and sustained or autonomous national development reached new heights of acceptance (see *World Systems Theory*).

2.3 Development and Urbanization, circa 1990: Disarticulated Processes

The growing popularity of world-system theory over the 1980s and 1990s may have strengthened scholars' understandings of the strong connections between urbanization and (under)development; but in the long run it contributed to the analytic disarticulation of cities and nations, thereby casting a mortal blow to studies of relations between these two domains. As scholars joined the world-system bandwagon and bypassed the nation-state as a unit of analysis, few examined national development as analytically distinct from global processes. Those development scholars who did concern themselves with the domestic dynamics of macroeconomic development rarely reserved a place for cities in their analytic frameworks, turning instead to states, classes, and the global economy. The result was that scholars who analyzed Third World cities did so increasingly in the context of global economic conditions, even as most scholars of national development ignored cities by and large.

This analytic disarticulation of urbanization and national development was further fueled by intellectual and political dynamics in the developing world. With many countries in Latin America, Africa, and Asia becoming less democratic and more authoritarian throughout the 1960s and 1970s, their states became stronger and more repressive. These political regimes tended to privilege capital over labor and cities over the countryside, using macroeconomic policy, national investments, and targeted subsidies to accomplish these aims. Resident and foreign observers saw little autonomy for cities, with urban developments mostly dictated by powerful (if not militarized) national forces. This was most evident in the largest cities, home to their nation's most significant industries and investments; but it also was seen as a pattern more generally. The result: in studies of the Third World, urban dynamics were frequently subsumed under national ones. This conceptual sleight of hand may have reinforced views about the direct relationship between city and nation to some extent; but it also chipped away at the analytic distinction between urban and national dynamics. And with these two processes increasingly collapsed into each other, the tendency was to examine one or the other, but not the interactive relationship between the two.

Among the few still fervently examining the direct connections between urbanization and national economic development were neoclassical economists, and World Bank staff and their clients. They worked under the supposition that by facilitating and/or managing cities, national actors could insure productivity, efficiency, and the flow of people and goods sufficiently to maximize overall development. Yet in most of the nonapplied fields, urbanists and developmentalists hardly spoke to one another.

3. A New Articulation (or, Development and Urbanization in the Age of Globalization)

As scholars now widely accept the premise that globalization is a central point of entry for studying changes in the contemporary world, there is a renewed interest in the urbanization-development nexus. But because the units of analysis for theorizing these connections have changed, slightly different claims and different types of research dominate the scholarly agenda. Among them are an increasing preoccupation with large cities, a resurgence of interest in urbanization in advanced countries of the world, and renewed debate about the direction and nature of the relationships between cities, nations, and the global economy

3.1 Globalization and Urbanization

One of the main characteristics of the new scholarship is the fixation with cities of enormous size and importance, frequently labeled world or global cities, whose growth and character owed to the role they play not just within their national borders, but also within global networks of production and consumption (Henderson and Castells 1987, Lo and Yeung 1998, Sassen 1991, 2000). The idea of world cities is not new of course (Friedmann and Wolff 1982); and much current work draws inspiration from earlier writings on overurbanization in Third World cities in which the inflated size and rapid growth of primate cities was linked to unequal relations of exchange between so-called core and peripheral economies. But in the new millennium, world cities are less likely to be seen as fetters on the national development of their host countries, and more likely to be conceptualized as the mechanisms through which global economic integration takes root.

Such views not only challenge past understandings of the destructive impact of global dynamics on cities and nations, they also come with a shift in geographic focus. Over the period 1960–2000 most scholars interested in the urbanization-economic development nexus examined the Third World; but now the US and Europe are back in the center of the conceptual map (King 1990, Sassen 1991), bringing the field full circle in terms of its country origins. The questions being posed also have changed considerably. Given the fact that many of the world cities being studied with

renewed vigor—like New York, London, Tokyo, Paris—are in the most economically advanced nations of the world, fewer scholars care about their potentially negative impact on national economies, since developmental dynamics in these countries are assumed to be impervious. As a result, attention is redirected to urban employment patterns, shifts in sectoral character, and the extent to which these patterns are owed to the globalization of capital and labor (Fainstein et al. 1992, Sassen 1991). There also is renewed interest in the changing 'locations' or economic roles these cities play in a regional, national, or international hierarchy of urban places (see Knox and Taylor 1995). Yet several pressing questions remain about the nature and direction of the relationship between urbanization and globalization.

3.2 Development and Urbanization by way of Globalization

One relatively unexplored issue concerns the impact of globally induced urbanization patterns on aggregate trajectories of economic development, be they understood nationally or on some other scale. On these counts, some headway is being made by those who identify regional dynamics both within and among established nation-states and how they factor into the equation. Saskia Sassen (1991), for example, argues that the globalization of capital and labor flows fuels the growth and economic successes of some cities (e.g., New York) while constraining others (e.g., Detroit), in the process exacerbating regional economic polarization. Taking an entirely different line, contemporary scholars of Europe or even East Asia—where globalization has already increased transnational economic integration in such a way as to form mega-regions with their own supranational governing institutions, and where cities on the receiving end of global investments and labor flows assume greater political and economic significance—are examining the extent to which globally-integrated cities will bypass the nation-state and negotiate directly with one another in larger regional pacts. Both lines of research have direct implications for understanding the dynamics of cities and national development, if only because they underscore the ways that, in an increasingly globalized world, the nation-state either comes under challenge or remains the politically relevant unit for mediating among cities, addressing intra-national regional disparities, and/or coordinating new supranational regional institutions.

To the extent that questions about the interconnected relationships between cities, regions, and nations were once the mainstay of studies of early modern Europe and the Third World, the developmental context where regional analysis reigned supreme (see Friedmann and Alonso 1964, Hardoy and Geisse 1972), scholars can look to the past—and across developmental categories—for insights (Clark 1996). Preliminary results point to a convergence in ideas and research programs, capable of transcending space and time, much like globalization itself, and holding the potential to unite scholars from all parts of the world around a common intellectual project.

See also: Authoritarianism; Capitalism; Dependency Theory; Industrialization; Modernization, Sociological Theories of; State Formation; Urban Geography; Urban Growth Models; Urban History; Urban Studies: Overview; World Systems Theory

Bibliography

Abu-Lughod J, Hay R (eds.) 1977 *Third World Urbanization*. Maaroufa Press, Chicago

Armstrong W, McGee T G 1985 *Theatres of Accumulation: Studies in Asian and Latin American Urbanization*. Methuen, London

Berry B J L 1961 City distribution and economic development. *Economic Development and Cultural Change* **9**: 573–8

Berry B J L 1973 *The Human Consequences of Urbanization*. St. Martin's, New York

Breese G 1966 *Urbanization in Developing Countries*. Prentice-Hall, Englewood Cliffs, NJ

Castells M, Vélez P (eds.) 1971 *Imperialismo y urbanización en América Latina*. Gustavo Gili, Barcelona, Spain

Clark D 1996 *Urban World/Global City*. Routledge, London

Cornelius W 1973 *Political Learning Among the Migrant Poor*. Sage Publications, Beverly Hills, CA

Davis K 1969 *World Urbanization, 1950–1970*. University of California Press, Berkeley, CA

Drakakis-Smith D 1987 *The Third World City*. Methuen, London

Durkheim E 1933 *The Division of Labor in Society*. Macmillan, London

Eckstein S 1977 *The Poverty of Revolution: The State and the Urban Poor in Mexico*. Princeton University Press, Princeton, NJ

El-Shakhs S 1972 Development, primary, and system of cities. *Journal of Developing Areas* **7**: 11–35

El-Shakhs S 1974 *Urbanization, National Planning, and Regional Development in Africa*. Praeger Publishers, New York

Fainstein S, Gordon I, Harloe M 1992 *Divided Cities: New York and London in the Contemporary World*. Oxford University Press, Oxford, UK

Field A (ed.) 1970 *City and Countryside in the Third World: Issues in the Modernization of Latin America*. Schenkman, Cambridge, UK

Friedmann J 1967 *A General Theory of Polarized Development*. Ford Foundation, Santiago, Chile

Friedmann J 1975 *The Urban Transition: Comparative Studies of Newly Industrializing Countries*. UCLA School of Architecture and Urban Planning, Los Angeles

Friedmann J, Alonso W 1964 *Regional Development and Planning*. MIT Press, Cambridge, MA

Friedmann J, Wolff G 1982 World city formation: An agenda for research and action. *International Journal of Urban and Regional Research* **6**: 309–44

Gilbert A, Ward P 1985 *Housing, the State and the Poor: Policy and Practice in Three Latin American Cities*. Cambridge University Press, oxford, UK
Goldscheider C 1971 *Population, Modernization, and Social Structure*. Little Brown, Boston
Goldstein S, Sly D 1979 *Patterns of Urbanization: Comparative Country Studies*. Ordina Editions, Dolhain
Gugler J 1978 *Urbanization and Social Change in West Africa*. Cambridge University Press, New York
Gugler J (ed.) 1988 *The Urbanization of the Third World*. Oxford University Press, New York
Hardoy J, Geisse G (eds.) 1972 *Políticas de desarrollo urbano y regional en América Latina*. SIAP, Buenos Aires, Argentina
Hardoy J, Satterwaithe 1986 *Small and Intermediate Centers: Their Role in Regional and National Developments in the Third World*. Westview Press, Boulder, CO
Hauser P 1957 *Urbanization in Asia and the Far East*. UNESCO Research Centre, Calcutta, India
Hauser P, Schnore L (eds.) 1965 *The Study of Urbanization*. Wiley, New York
Henderson J, Castells M 1987 *Global Restructuring and Territorial Development*. Sage Publications, Beverly Hills, CA
Herrick B 1965 *Urban Migration and Economic Development in Chile*. MIT Press, Cambridge, MA
Hoyt H 1962 *World Urbanization: Expanding Population in a Shrinking World*. Urban Land Institute, Washington, DC
King A 1990 *Global Cities: Post-imperialism and the Internationalization of London*. Routledge, London
Knox P, Taylor P 1995 *World Cities in a World System*. Cambridge University Press, Cambridge, UK
Linn J 1983 *Cities in the Developing World: Policies for their Equitable and Efficient Growth*. Oxford University Press, New York
Lipton M 1977 *Why Poor People Stay Poor: Urban Bias in World Development*. Harvard University Press, Cambridge, MA
Lo F, Yeung Y 1998 *Globalization and the World of Large Cities*. United Nations University Press, Tokyo
Mangin W (ed.) 1970 *Peasants in Cities: Readings in the Anthropology of Urbanization*. Houghton Mifflin, Boston
McGee T G 1971 *The Urbanization Process in the Third World*. Bell, London
McGreevey W P 1971 A statistical analysis of primacy and lognormality in the size distribution of Latin American cities, 1950–1960. In: Morse R (ed.) *The Urban Development of Latin America: 1750–1920*. Stanford University Press, Stanford, CA
Perlman J 1976 *The Myth of Marginality: Urban Poor and Politics in Rio de Janeiro*. University of California Press, Berkeley, CA
Pirenne H 1936 *Economic and Social History of Medieval Europe*. Harcourt, Brace, Jovanovich, New York
Portes A, Walton J 1976 *Urban Latin America: The Political Condition from Above and Below*. University of Texas Press, Austin, TX
Portes A, Walton J 1981 *Labor, Class, and the International System*. Academic Press, New York
Rabinowitz F 1973 *National-local Linkages: The Inter-relationships of Urban and National Politics in Latin America*. Sage Publications, Beverly Hills, CA
Redfield R 1941 *The Folk Culture of the Yucatan*. University of Chicago Press, Chicago
Redfield R 1953 *The Primitive World and its Transformations*. Cornell University Press, Ithaca, NY
Richardson H 1973 *The Economics of Urban Size*. Saxon House, Westmead, UK
Roberts B 1978 *Cities of Peasants: The Political Economy of Urbanization in the Third World*. Arnold Press, London
Rondinelli H 1983 *Secondary Cities in Developing Countries*. Sage Publications, Beverly Hills, CA
Sassen S 1991 *The Global City: New York, London, Tokyo*. Princeton University Press, Princeton, NJ
Sassen S 2000 *Cities in a World Economy*. Pine Forge Press, Thousand Oaks, CA
Simmel G 1950 *The Sociology of Georg Simmel*. Free Press, New York
Smith D 1996 *Third World Cities in Global Perspective: The Political Economy of Uneven Urbanization*. Westview Press, Boulder, CO
Stren R (ed.) 1995 *Urban Research in the Developing World*. Centre for Urban and Community Studies, University of Toronto, Canada
Tilly C 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ
Timberlake M (ed.) 1985 *Urbanization in the World Economy*. Academic Press, New York
Walton J (ed.) 1985 *Capital and Labor in the Urbanized World*. Sage Publications, Beverly Hills, CA
Weber M 1927 *General Economic History*. Greenburg Press, New York
Weber M 1958 *The City*. Free Press, New York

D. E. Davis

Development, Economics of

Economic development is the process through which economies are transformed from ones in which most people have very limited resources and choices to ones in which they have much greater resources and choices. Economic development therefore covers almost all areas of economics, though with modifications to reflect the particular situations of developing countries. Development economics refers to studies of economies with relatively low per capita resources: low- and middle-income economies in Table 1. About 85 percent of humanity currently lives in these economies. Product per capita in low-income economies averages 1.7 percent and in middle-income economies 9.6 percent of averages in high-income economies, with that in Switzerland over 500 times that in Mozambique. Correcting these comparisons for different price structures reduces the variations a lot, but the range still is considerable, with per capita product almost 60 times as high in the United States as in Ethiopia. Human resource indicators (e.g., life expectancies and illiteracy) also are poorer for developing than developed economies, but current differences are much smaller proportionately than are those in product per capita. The question of what causes economies to move from per capita resource levels as in Mozambique and Ethiopia to those as in Switzerland and

Table 1
Population, product per capita, life expectancies at birth, and illiteracy rates by level of development in 1995

Country groups	Population (billions)	Gross national product per capita: United States ($)	Gross national product per capita: Purchasing power parity (United States = 100)	Life expectancy at birth (years)	Adult illiteracy rate (%)
Low-income	3.180	430	1.7–12.1	63	34
Middle-income	1.591	2390	6.6–43.4	68	18
High-income	0.902	24930	42.4–100	77	< 5

Source: World Bank 1991–1997, Table 1.

the United States is the fundamental question that development economics addresses.

1. *Conceptualization of Development in the Mid-twentieth Century*

Prior to the neoclassical developments in economic analysis of the late nineteenth century, most influential economists, for example, Smith, Ricardo, Marx, were concerned substantially with economic development. In the decades prior to the Second World War, in contrast, most neoclassical economists were not concerned with development, but with topics such as behaviors of firms and households, general equilibrium, and efficiency and, starting in the 1930s, short-run macroeconomic equilibria with aggregate demand shortfalls.

In the immediate post-WWII era, mainstream economics continued to focus on the extension of the neoclassical paradigm and on Keynesian macro short-run disequilibria with virtually no attention to development. Development economics as a subdiscipline branched off from what Samuelson called the 'neoclassical synthesis' of mainstream economics because the development of poorer lands was viewed by development economists as different in essence from concerns that dominated the neoclassical synthesis. Development was seen as involving discontinuous shifts from a stagnant low-income traditional equilibrium ('a low level equilibrium trap' or a 'vicious circle of poverty') to a modern higher-income, changing, and growing economy with increasing average per capita income, poverty alleviation and improvements in human resources. Development was viewed as requiring structural change different in kind from the marginalism of neoclassical economics.

The dominant paradigm of the new/reborn discipline of development economics in the mid-twentieth century was based on key perceptions about the way markets, incentives, and other institutions operated in poor countries. These, in turn, had important policy implications.

1.1 *Key Perceptions about Markets, Incentives, and Other Institutions in View of Development in the Mid-twentieth Century*

1.1.1 Surplus labor and capital shortages. Factor proportions in poor countries implied high labor to capital ratios. Family farms/firms with sharing rules based on average products permitted marginal labor products to fall to zero so that labor was in surplus in large traditional agriculture sectors. Surplus labor presented an advantage because, if it could be moved to industry, large gains in product could be obtained and much of the income generated could be reinvested in further capital accumulation since the 'unlimited supply of labor' keeps real wages low. Physical capital was relatively scarce, so that, in Nurske's (1952, pp. 1–2) words, 'the problem of development is largely … a problem of capital accumulation.' But domestic capital accumulation was limited by a vicious circle: low incomes meant low savings so that investment was low that kept incomes low. Human capital was given little emphasis.

1.1.2 Limited role of domestic markets. Markets had limited efficacy because: (a) large population shares, including traditional peasants and governmental officials, were not responsive to market and other incentives. (b) Markets were not well developed because of large subsistence sectors, costly transportation and communication, and the predominance of mechanisms internal to households and firms for transferring resources across time and space. (c) Markets did not incorporate well information about demand complementarities of investments (e.g., Rosenstein-Rodan 1943) and could not provide the right guidance for investments that would change the basic structure of the economy with more than marginal effects. (d) There were important market failures of the types considered in mainstream economic analyses: technological externalities not transferred through markets (e.g., pollution), increasing returns to scale relative to market sizes, and 'public goods' the use of which are not 'rival' so that more consumption by one individual does not mean that less is

available for another entity (e.g., information). (e) Some major actors in markets—rural moneylenders, landlords, and traders—were 'rapacious profiteers' who 'exploited' the uninformed poor by paying them low product prices and charging them high input and credit prices.

1.1.3 Key role of industrialization. Agriculture dominated employment and production. But the development process was in substantial part industrialization. Productivity-enhancing technologies and positive externalities were primarily available for industry, domestic demand would shift towards industry with increased incomes because of income-inelastic demands for food (Engel's law), surplus labor in agriculture could be shifted to productive uses in industry, savings potentials were much greater in industry, and industrialization was essential for maintaining national autonomy (e.g., Mandelbaum 1945, Rosenstein-Rodan 1943, Singer 1950, Mahalanobis 1955, Lewis 1954, Ranis and Fei 1961).

1.1.4 Limited prospects in international markets. Pessimism prevailed regarding international market prospects, in part because of the market collapses in the 1930s. Prebisch (1950) and Singer (1950) argued that international trade transferred most benefits from increased productivity in poor countries to richer countries, the terms of trade of primary producers decline secularly because demands were income-inelastic for primary commodities and the developed countries had monopoly power in markets for manufactured products, and opportunities for developing countries to expand their manufacturing exports were limited because such markets would not grow quickly and these markets were dominated by monopolistic producers in the high-income countries who had succeeded in establishing protectionist policy-determined barriers to expanded manufacturing imports from developing countries. International capital markets would not provide sufficient investable funds for development and they cause exploitation and crowd out domestic investments.

1.1.5 Policy makers have good information, are disinterested and are effective. Policy makers could identify current and future magnitudes of negative and positive externalities, complementarities among production sectors, and responses of various entities to policies so they could and would design and implement good policies.

1.2 Policy Implications of the View of Development in the Mid-twentieth Century

1.2.1 Capital accumulation. Physical capital accumulation was central, as codified by the Harrod–Domar dynamic equilibrium condition, which in the simplest case is that the rate of growth equals the average saving rate divided by the incremental capital output ratio (ICOR). This relation was used to calculate savings requirements to achieve growth targets given ICORs and to justify efforts to increase savings, including shifting resources from private entities to governments under the assumption that governmental savings rates exceeded private ones, and to obtain foreign savings to supplement domestic savings. When exchange rates became overvalued, foreign capital to purchase critical machinery and equipment became of particular concern, as systematized in the 'two-gap' model of Chenery and Bruno (1962) in which the constraint on growth is either savings or foreign exchange. Human capital accumulation was not central in most analysis partly because of the widespread perception of surplus labor.

1.2.2 Import-substituting industrialization with balanced growth. One major mechanism widely used to shift resources to industry, to protect new domestic industries against foreign competitors, and to increase economic independence was high protection, through tariffs and quotas, on manufacturing imports. This established inducements for 'balanced industrial growth,' with important complementarities for intermediate inputs and for final product demands ('backward and forward linkages').

1.2.3 Policy mix—quantitative, direct public-sector production, planning, and irrelevance of macro balance. The state role was very extensive. The dominant policy mode was *dirigiste*. Efforts to improve markets or to use 'market-friendly' taxes/subsidies were relatively rare. Direct quantitative regulations and allocations, including substantial investment in public-sector production, were common because it was thought that the desired ends could be effected with most certainty through such direct actions by knowledgeable, disinterested policy-makers. Public-sector investments included physical capital and human resource infrastructure and many other subsectors in the 'commanding heights' of the economy so that the state could control directly development.

Planning often was viewed as essential because of market limitations, reinforced by international aid agencies that interpreted plans as essential evidence concerning the seriousness of countries' development efforts. Planning agencies were concerned with investment criteria and choice of techniques for deciding on direct public sector investments and for guiding private investments. 'Shadow' ('scarcity') prices were used to evaluate social cost-benefit ratios or internal rates of return to alternative investments, with un-

skilled labor valued at zero because of surplus labor and foreign exchange valued more highly than official overvalued exchange rates. A considerable literature developed methods for calculating general-equilibrium shadow prices. Multisector planning models were developed and widely used to determine sectoral primary factor requirements, physical capital investments, import substitution possibilities and 'noncompetitive imports' (for which domestic substitutes were not available) for target growth rates. These models generally assumed no substitution among production inputs nor among demands, focused on physical capital, and were quite aggregated with little or no variation in qualities of products. They addressed directly possibilities of supply or demand complementarities. Optimizing versions yielded shadow prices that indicated impacts on objective functions of marginal changes in binding model constraints. Often these shadow prices were not very robust to small changes in the model. Moreover, these models did not incorporate most macroeconomic phenomena, so they did not provide guidance concerning such matters as inflation. Furthermore, most of these planning models ignored or only crudely incorporated most of the major phenomena other than the supply and demand complementarities that underlay the critique of the dependence on markets (Sect. 1.1). Subsequent models improved in some important respects, for example, discontinuities were incorporated into integer programming models and price responses in Johansen-type models (1960) that led to more recent computable general equilibrium (CGE) models (Blitzer et al. 1974). But still many of the limitations noted remained.

Policies to establish macroeconomic balance—a major topic in high-income economies—were rejected as irrelevant. Rao (1952) and others argued that the separation between savings and investment that was critical in Keynesian underemployment analysis was not relevant for developing countries given the dominance of family firms. The Latin American 'structuralists' argued that rigidities precluded economic entities from responding quickly to changes so that orthodox currency devaluation or monetary and fiscal policies would reduce output and employment.

2. *The Dominant Paradigm Concerning Development at the Start of the Twenty-first Century*

Aggregate development experience since WWII in many respects has been very impressive, though with considerable variations across economies. Average growth rates in GDP per capita have been high by historic standards and high for developing relative to developed economies. Other indicators of human welfare, such as life expectancy and schooling, have improved even more impressively. *Prima facie*, such success might be interpreted to ratify the post-WWII development paradigm and associated policies (Sect. 1). But the appropriate comparison should be with what could have been achieved given the conditions of the post-WWII era and appropriate development strategies and policies.

One major change in the last three decades or so of the twentieth century was the rapid expansion of systematic quantitative analyses of development, often based on data and approaches that were not available earlier. Many studies explored the underlying assumptions for the dominant initial post-WWII conceptualization of development. Some were undertaken in relatively visible large-scale comparative projects (e.g., Little et al. 1970, Bhagwati 1978, Krueger 1978, 1983, Taylor 1988, Page et al. 1993). Others have been individual country and microstudies that have investigated more detailed aspects of the development experience using new micro data sets with systematic models (e.g., see references in Chenery and Srinivasan 1988–1989 and Behrman and Srinivasan 1995; some more recent examples are Behrman et al. 1997, 1999, Foster and Rosenzweig 1995, 1996a, 1996b, Jacoby 1995, Jacoby and Skoufias 1997, Pitt and Khandker 1998).

These studies question many of the assumed initial conditions in the post-WWII development paradigm and suggest more nuanced characterizations of others. They suggest: (a) traditional agriculture is relatively efficient and traditional institutions such as sharecropping, contracts tied to multiple transactions, attached farm servants, arranged marriages, and migratory patterns are efficient arrangements given imperfect or absent capital and insurance markets. (b) Substantial increases in savings and physical capital investments rates are not sufficient for development because (i) incentives structures affect strongly ongoing productivity growth and (ii) human resource development related to nowledge acquisition and adaptation is critical. (c) Domestic markets work well for many goods and services and returns to market improvements through infrastructure development, better information, and lessening policy barriers to well-functioning markets are often great. Nevertheless some markets—particularly those related to capital, insurance, and information—are imperfect or missing. (d) Industrialization is not equivalent to development and industrialization strategies that penalize other production sectors may retard development. (e) International markets have presented substantial opportunities for developing countries, with the most successful experiences being for those developing economies (initially primarily in East and Southeast Asia) that expanded rapidly their exports and used international markets to induce ongoing productivity improvements as well as sources of foreign investment. (f) Policymakers have limited information and may be rent-seekers rather than knowledgeable and disinterested.

This re-evaluation of initial conditions has substantial implications for policies. Because of the combination of currently perceived much greater responsiveness, limited information (and information asymmetries) faced by governments, limited governmental implementation capacities, much greater variations in quality of goods and services, much more efficient transmission of certain types of information through unhindered markets, important role of markets in inducing greater efficiency, much greater importance of basic macroeconomic stability, and rapid changes in markets and in technology than assumed in the immediate post-WWII paradigm, there has been a sea change in perceptions of appropriate governmental roles and of the best mix policy mix. A broad new consensus had risen by the early 1990s on the nature of development processes and implications for policies. Components of this consensus include the following.

2.1.1 The basic economic environment must be conducive for investments. The two major components are: (a) a stable macroeconomic environment with non-inflationary monetary policy, fiscal restraint, and international balance (e.g., World Bank 1991, Corbo et al. 1992, Page et al. 1993) and (b) institutions that permit investors to have reasonable expectations of reaping much of the gains from their investments. Domestic macrobalance is important because inflationary taxes advocated by many earlier development economists are inefficient, inequitable, erode confidence in policy and in the prospects for the economy, discourage longer-run productive investments by increasing uncertainty, and are difficult to eliminate. Balance in international economic interchange is important so that critical imports of intermediate and capital goods can be maintained, pressures for increased efficiencies are ongoing through exposure to world markets, and access to international finance and technology is maintained. Institutions include basic property rights, legislative and legal systems that resolve disputes fairly, and various mechanisms to share (but not to eliminate) risks.

2.1.2 Human capital is important, additional to physical capital. In the views of some, investment in human resources is in itself development. For most mainstream development economists, in addition to any direct welfare effects of better human resources, there are important productivity effects. Positive externalities through human resource investments may be considerable, moreover, primarily because of externalities to knowledge and to capacities for individuals to find and to process information. Given gender roles and current market/policy failures, the social returns to human capital often may be relatively high for investments in females (King and Mason 2000).

2.1.3 Responsiveness to incentives created by markets and policies are broad and often substantial. Therefore, it is desirable that markets and policies create incentives for behaviors that are socially desirable and not sufficient, as in the earlier paradigm, that policy makers know shadow prices. Economic entities also adjust many or even all elements of their behaviors in response to changes. Therefore, anticipating full effects of policies or market outcomes on behavior is difficult because of such cross-effects. The impact of new technological options, for example, may increase the value of educated women in marriage markets if gender roles dictate that women play major roles in child education (Behrman et al. 1999).

2.1.4 Information problems are pervasive in a rapidly changing world so incentive-compatible institutions and mechanisms are desirable. Information is imperfect in general, information that policy makers/implementers have often is less and more dated than the information that individuals have who are directly involved in production and consumption of particular goods and services, asymmetrical information creates incentives for opportunistic individual behaviors that are likely to be socially inefficient and inequitable, and markets are relatively efficient in conveying certain types of information. Rather than attempting directly to regulate and monitor many micro activities related to the development process, governments more effectively can support institutions that are relatively effective in inducing desirable behavior. The oft-cited claim that an important aspect of the successful East Asian development experience has been outward orientation, for example, may be viewed in part as a mechanism for encouraging increasing productivity through forcing competition on world markets. Likewise, efforts to make markets work better through open and transparent bidding for governmental contracts, better market integration to lessen monopoly power, introducing voucher systems and decentralization of human resource-related services, eliminating governmental regulations on entry, improving communication and transportation and subsidizing or mandating information that has public goods characteristics all may have important effects in increasing efficiency.

2.1.5 Selective global and directed policy interventions that are 'market friendly' where there are market failures are warranted. Policies should: (a) maintain macroeconomic balance, (b) develop institutions, including property rights and related legal frameworks, so that investors have reasonable expectations of gaining returns from good investments, (c) encourage competition to induce productivity increases through an unbiased international trade system and through eliminating legal barriers and lessening transportation and communication barriers to

greater competition, and (d) use 'market-friendly' interventions that are transparent, rule-based, general and nondiscriminatory such as open auctions, vouchers, subsidies, and taxes to attempt to offset market failures (and directing such policies as directly as possible towards the market failures to limit unintended distortion costs). These considerations imply a critical, but also more selective role for government than in the earlier paradigm. Also involvement in these activities is not necessarily best accomplished by governmental direct production rather than taxes/subsidies and regulations concerning disclosure of private information.

3. *Current Controversies and Possible Future Developments*

Despite substantial current general consensus (Sect. 2), there is neither consensus on every dimension nor on how the paradigm is likely to change.

3.1 *Integration into the International Financial System*

International financial movements have expanded rapidly and can swamp foreign exchange movements from international trade. This may provide useful discipline for governments to keep their policies credible. But this also may result in enormous shocks, perhaps fueled by speculative bubbles, which present great difficulties for even the best of governments. The Asian financial crisis of the late 1990s intensified greatly such concerns. Considerable debate continues about the extent to which such problems can be lessened by greater transparency in developing countries finances, more restrictions on capital movements, appropriate sequencing of liberalization efforts with capital flows liberalized relatively late in the process, new systems to assure that developed country investors bear some of the social costs of disruptions originating in large international capital movements, and greater international insurance through international organizations (with attendant moral hazards of encouraging risky behaviors).

3.2 *Demographic 'Windows of Opportunities' and Aging Populations*

Waves of Malthusian overpopulation fears have featured prominently among noneconomists. But statistical analyses of relations between population growth and economic development do not find important long-run associations. However, recently developing countries are claimed to have transitory (but perhaps lasting several decades) opportunities because aging due to fertility reductions means that they have large working-age populations relative to younger and older populations. Some have attributed a substantial part of the high recent economic growth in some Asian economies to such a 'demographic bonus' associated with higher savings and labor force participation and lower health and education expenditures (Asian Development Bank 1997, Bloom and Williamson 1999). There are controversies about what conditions permit developing economies to exploit successfully demographic windows of opportunity to enhance development, rather than being hindered by faster increases in supplies than in demands for labor. Empirical associations for the 1950–1995 period suggest that the conditions embodied in the consensus that is summarized in Sect. 2 increase probabilities of being able to exploit demographic opportunities. Eventually aging populations will imply increased dependency ratios due to a larger number of aged individuals, which will raise issues about intergenerational transfers for retirees and increased health problems associated with aging. Projections suggest that average ages in most currently developing countries will converge substantially to those of developed economies by 2025, so demographic–development interactions will change further.

3.3 *Poverty Reduction vs. Growth*

The dominant view is that the most effective form of sustained poverty reduction is rapid development that exploits comparative advantages (perhaps including relatively large labor forces), though some safety nets are likely to be desirable for individuals who cannot participate or who suffer dislocations in adjustments towards sustained growth (e.g., Page et al. 1993). But others claim that there are substantial tradeoffs between poverty reduction and the dominant conventional wisdom, at least during transition structural adjustment programs (e.g., the UNICEF critique, calling for 'adjustment with a human face,' Cornia et al. 1987). There are inherent productivity-equity tradeoffs in providing safety nets because a sense of entitlement tends to develop that makes such policies hard to change (e.g., political responses to efforts to eliminate food and transportation subsidies) and because they may dampen incentives for work efforts through implying high effective marginal tax rates. Therefore, there is likely to be an ongoing debate about the extent to which such safety nets should be provided and exactly what forms they should take.

3.4 *Social Cohesion and Social Capital*

There is a growing perception that these are important factors in successful development efforts. If so, then

there may be important policy implications from the perspective of the currently dominant paradigm (Sect. 2) because factors that increase social cohesion and social capital are likely to have social returns beyond private returns and thus be under-rewarded by unhindered markets. Establishing the importance of such factors, however, requires the development of more careful integrated modeling/estimation approaches. Suggestive anecdotes and associations are currently available, but not persuasive systematic studies that lead to confident identification of causality in the presence of behavioral choices and unobserved individual and community characteristics.

3.5 *Paternalism/Maternalism, Individual Choice, and Endogenous Preferences*

In the immediate post-WWII dominant paradigm (Sect. 1) policy makers and analysts acted as if they knew what was best for the poor in developing countries. If the poor were perceived to hold the 'wrong' values (e.g., not working hard enough, wanting too many children, saving too little) it was deemed appropriate to ignore or to try to change their preferences. The current dominant paradigm places greater weight on individual preferences, at least in so far as they are manifested through responses to markets and policies (though limited resources limit the effectiveness of the poorest). There also is recognition of difficulties in making welfare comparisons if preferences are changed. Nevertheless, there are some who appear to wish to impose their values on others regarding, for example, appropriate gender relations. Such concerns raise the analytical question that also has been of increasing interest in general mainstream economics of how to model endogenous preferences and to include them usefully in applied analysis.

3.6 *Association between Economic Participation and Political Participation*

An important component of the current economic development paradigm is much greater recognition of active participation of individuals, which, together with discipline of competitive-like markets, provide checks and balances that limit predatory economic behavior. But is an essential component of successful development also a political system with checks and balances? Some argue that some of the most successful East Asian economic development experiences have been under strong autocratic governments (e.g., Korea under Park, China, Singapore). Others note that strong autocratic governments have not been a sufficient condition for successful economic development (e.g., Myanmar, Cuba, the Philippines under Marcos). This controversy about what forms of political systems lead most effectively to development is likely to be ongoing.

3.7 *Extent of Technological Externalities for Human Resources, Population Control, and Environmental Effects and Remedies for those Externalities*

Assertions are widespread that there are substantial positive externalities of human resources and negative externalities of population growth and economic growth on the environment. Yet empirical evidence regarding magnitudes and natures of such externalities is almost nonexistent, at times *a priori* arguments about what constitutes an externality are confused, and means sometimes advocated to address such externalities do not seem directed well towards the supposed sources of externalities nor do they recognize well the importance of incentives, so they may have substantial distortion costs. There would be high returns to persuasive systematic studies that lead to confident identification of the magnitudes of such externalities in the presence of behavioral choices and unobserved individual and community characteristics (e.g., Foster and Rosenzweig 1995).

4. *Conclusion*

The economics of developing countries was an integrated part of most major streams of economic analysis in the nineteenth and early twentieth centuries. But in the immediate post-World War II period, economic development was widely seen as sufficiently different in degree to be different in kind from mainstream economics for the developed economies. Developing economies were seen as having such imperfect markets and limited responsiveness to markets by private entities that effective development could be best pursued by (supposedly disinterested and very knowledgeable) policy makers using direct quantitative policies guided by planning models to industrialize in part by protecting their economies from the many negative effects of international markets. In the last quarter of the twentieth century, there was a radical shift in the dominant paradigm concerning economic development. This was in part due to a number of new systematic quantitative studies of a range of developing experiences, importantly including the unprecedented growth in a number of East and Southeast Asian economies. The dominant paradigm at the start of the twenty-first century emphasizes that there is considerable responsiveness to opportunities in developing countries so the incentives created by markets and policies may be critical, that development is much more than physical capital

accumulation and industrialization, that human capital and knowledge acquisition may be critical in the development process, that policymakers have limited information and may be rent-seekers, and that a much more nuanced understanding of the nature and response to market failures implies an important but much more limited and incentive-sensitive role for policy. There remain a number of current controversies, such as related to the appropriate degree of integration into global financial markets, the importance of some aspects of social cohesion and social capital, the relation between economic and political participation, and the nature and extent of alleged externalities and other market failures. But for the most part they are considered at least by economists within the framework of this paradigm.

Such a summary of decades of the evolving conceptualization of economic development into a few major paradigms, of course, suppresses considerable heterogeneities of perceptions and approaches among the many economists and policymakers concerned with economic development over the years. But it also captures major dominant features of the evolution of thought on economic development that originally was an integral part of the major approaches to economics, was increasingly separated from main-stream economics in the mid-twentieth century, and then has increasingly converged back with mainstream economics by the start of the twenty-first century. This latter convergence has occurred both because development economics has moved closer to mainstream economics with regard to emphasis on responsiveness to incentives, gains from greater market dependence and limitations of policy interventions and because mainstream economics has moved closer to development economics with regard to such issues as market limitations related to information problems and the broader political, social and cultural context in which economic analysis occurs and policies are made.

See also: Development and the State; Development: Socioeconomic Aspects; Development Theory in Geography; Economic Development and Women; Economic Geography; Economic Growth: Theory; Economic Restructuring: Geographic Aspects; Globalization: Geographical Aspects

Bibliography

Asian Development Bank 1997 *Emerging Asia: Changes and Challenges*. Asian Development Bank, Manila

Behrman J R, Foster A, Rosenzweig M R 1997 The dynamics of agricultural production and the calorie-income relationship: Evidence from Pakistan. *Journal of Econometrics* **77**: 187–207

Behrman J R, Foster A, Rosenzweig M R, Vashishtha P 1999 Women's schooling, home teaching, and economic growth. *Journal of Political Economy* **107**: 682–714

Behrman J R, Srinivasan T N 1995 *Handbook of Development Economics*, Vol. 3. North-Holland, Amsterdam

Bhagwati J 1978 *Foreign Trade Regimes and Economic Development: Anatomy and Consequences of Exchange Control Regimes*. Ballinger, Cambridge, MA

Blitzer C, Taylor L, Clark P 1974 *Economy-wide Models in Development Planning*. Oxford University Press, London

Bloom D E, Williamson J G 1999 Demographic transitions and economic miracles in emerging Asia. *The World Bank Economic Review* **12**: 419–56

Chenery H B, Bruno M 1962 Development alternatives in an open economy. *Economic Journal* **72**: 79–103

Chenery H B, Srinivasan T N 1988–1989 *Handbook of Development Economics*, Vol. 1 and 2. North-Holland, Amsterdam

Corbo V, Fischer S, Webb S B (eds.) 1992 *Adjustment Lending Revisited: Policies to Restore Growth*. World Bank, Washington, DC

Cornia G, Jolly R, Stewart F 1987 *Adjustment with a Human Face*, Vol. 1. Clarendon, Oxford, UK

Foster A, Rosenzweig M R 1995 Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy* **103**: 1176–209

Foster A, Rosenzweig M R 1996a Technical change and human-capital returns and investments: Evidence from the Green Revolution. *American Economic Review* **86**: 931–53

Foster A, Rosenzweig M R 1996b Comparative advantage, information and the allocation of workers to tasks: Evidence from an agricultural labor market. *Review of Economical Studies* **63**: 347–74

Jacoby H 1995 The economics of polygyny in Sub-Saharan Africa: Female productivity and the demand for wives in Côte d'Ivoire. *Journal of Political Economy* **103**: 938–71

Jacoby H, Skoufias E 1997 Risk, financial markets and human capital in a developing country. *Review of Economic Studies* **64**: 311–35

Johansen L 1960 *A Multi-sectoral Study of Economic Growth*. North-Holland, Amsterdam

King E M, Mason A 2000 *Engendering Development: Gender Equality in Rights, Resources, and Voice*. Oxford University Press for the World Bank, New York

Krueger A O 1978 *Foreign Trade Regimes and Economic Development: Liberalization Attempts and Consequences*. Ballinger, Cambridge, MA

Krueger A O 1983 Trade and Employment in Developing Countries: Synthesis and Conclusions. University of Chicago Press, Chicago

Lewis W A 1954 Economic development with unlimited supplies of labor. *Manchester School of Economics and Social Studies* **2**: 139–91

Little I M D, Scitovsky T, Scott M 1970 *Industry and Trade in Some Developing Countries*. OECD, Paris

Mahalanobis P C 1955 The approach of operational research to planning in India *Sankhya*. *The Indian Journal of Statistics* **16**: (Parts 1 and 2)

Mandelbaum K 1945 *The Industrialization of Backward Areas*. Blackwell, Oxford, UK

Nurkse R 1952 Some international aspects of the problem of development. *American Economics Review* **42**: 571–83

Page J, Birdsall N, Campos E, Carden W M, Kim C S, Pack H, Sabot R, Stiglitz J E, Uy M 1993 *The East Asian Miracle: Economic Growth and Public Policy*. World Bank, Washington, DC

Pitt M M, Khandker S R 1998 The impact of group-based credit programs on poor households in Bangladesh: Does the gender

of participants matter? *Journal of Political Economy* **106**: 958–96
Prebisch R 1950 *The Economy of Latin America and Its Principal Problems*. United Nations, New York
Ranis G, Fei J 1961 A theory of economic development. *American Economic Review* **56**: 533–58
Rao V K R V 1952 Investment, income and the multiplier in an underdeveloped economy. The Indian Economic Review. (Reprinted in Agarwala A N, Singh G H (eds.) 1963 *The Economics of Underdevelopment*. Oxford University Press, New York, pp. 205–18
Rosenstein-Rodan P N 1943 Problems of indusrialization of Eastern and south-Eastern Europe. *Economic Journal* **53**: 202–11
Singer H 1950 The distribution of gains between investing and borrowing countries. *American Economic Review* **40**: 473–86
Taylor L 1988 *Varieties of Stabilization Experience*. Clarendon, Oxford, UK
Udry C 1994 Risk and insurance in a rural credit market: An empirical investigation in northern Nigeria. *Review of Economic Studies* **61**: 495–526
World Bank 1991–1997 *World Development Report*. Oxford, New York

J. R. Behrman

Development: Organizational

1. Introduction

Organizational development (OD) is a special interdisciplinary field focused on organizational change. It provides conceptions for understanding the logic of change as well as techniques and tools for managing the change process effectively. The major purpose of OD interventions is to overcome resistance to change and to encourage the willingness of organizational members to participate actively in the change process. As a discipline of social sciences OD has been criticized for lacking a sound theoretical basis and for misconceiving the nature of organizational change as exceptional projects in organizational life.

2. Historical Overview

In classical organization theory, the problem of coping with organizational change was not of much concern. Change basically meant reorganization and the critical activity was to determine the best organizational solution to master new situations. Realizing the new solution, if regarded as a problem at all, was seen as a problem of careful preparation and of giving the right instructions. After a brief period of tolerance, all employees were supposed to work according to the new organizational guidelines.

This traditional model, however, rarely proved to be a success. Organizational members notoriously resisted the new organizational design, unforeseen events rendered the design worthless, the change process came to nothing, etc.

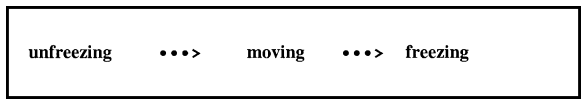

Figure 1
Lewin's triadic episode of organizational change

The typical reaction to these irritations was to blame the plans for having been imperfect and to ask for their refinement. In most cases this actually made the situation even worse; the resistance became stronger, the rejections more sophisticated, etc. Although the problems of organizational change made themselves felt in practice, they did not gain the status of salient question in classical theory.

It was the behavioral sciences, especially the human resources approach, which made organizational change a problem in its own right and developed completely new approaches for its solution.

The impetus to establish organizational change as a field came from empirical studies finding that success in change critically depends on the attitudes of organization members as well as on the general mood towards the new design. This insight originated from studies on 'resistance to change.'

The initiative to study resistance to change, as well as the approaches on how to overcome it, dates back to Lewin (1958). Most salient are his seminal experiments on changing the aversion to specific foods. Group meetings and active involvement of group members proved to be the best way to change food habits. The foundations for what were later to become the 'golden rules' of successful organizational change (Lawrence 1954) have been laid in these studies:

(a) Broad information at an early stage and active participation in the change process.

(b) The group as an important medium (facilitator) for change: Change processes are perceived as less worrying in groups and generally are accepted more easily.

(c) Mutual cooperation among organizational members.

(d) A settling-down period at the end of the change process designed to institutionalize the changes that were carried out.

Lewin (1958, p. 210) later elaborated on these insights and conceived the well-known 'triadic episode' of successful change. The model builds on the idea of homeostasis and suggests a three-step procedure: unfreezing, moving, and refreezing (see Fig. 1).

The most important issue of the model is that organizations need an initial period of irritation in order to break up their habits and routines. Or as Lewin put it, organizations have to become 'unfrozen'; otherwise they are not likely to give up their equi-

librium position and to move to a new level. Changes tend to be short lived if there are no efforts to bring them to permanency. The new patterns, therefore, must be 'frozen,' the system has to settle down again and find a new equilibrium.

3. Organizational Development as a Discipline

Lewin's group experiments and the three-stage model finally boiled down to a special method (Laboratory–Training) designed to cope with behavioral and organizational changes in totally new ways. This new technology of group intervention and observation soon became very popular and built the core of the newly founded National Training Laboratories in Group Development. A series of other approaches emerged and also the facilitation and support of organizational change processes increasingly became a branch for external consultants. All these forces finally led to the birth of a new field, a special discipline in the behavioral sciences, called OD.

Among the other roots of OD, the 'Survey–Feedback' method is one of the most prominent. This approach makes use of attitude surveys and data feedback to facilitate change processes. All organizational members are invited to take part in the change process ('shared power approach'). The intention is to stimulate organizational members' interest in change, that is, to create a motivation to overcome the weaknesses of the present state. Lewin also pioneered this method.

Another, and closely related, historical root was the work of the Tavistock Institute (UK) which is anchored deeply in psychotherapy and group therapy (Bion 1961). The doctor–client consultation framework and the high professional demands which were later put on the change manager originate from this concept. Some prominence was gained from the so-called Tavistock Conferences, that is, leaderless group discussions designed to experience and to explore leadership, authority, and obedience. Tavistock started to transfer the clinical methods to organizational settings in the early 1950s.

Further roots of OD are the approach to job redesign and to the improvement of the quality of working life. Here again, The Tavistock Institute (of Human Relations) played a major role. Frame of reference is the sociotechnical systems approach; that is, a conceptual framework and an action research program which is devoted to meet both individual needs and technical demands (Rice 1958). The emphasis is on changing traditional organizational design towards group-oriented work designs giving organizational members more variety, meaning, and feedback in their work.

Later, the focus of OD was extended significantly. Additional approaches and new issues enriched the field and made it somewhat diverse (Cummings and Worley 2000). At the core of OD, however, still remain change methods and specific interventions at different levels of social systems.

4. Organizational Development: Conception

The field of OD is truly interdisciplinary in both focus and methods. Even though different perspectives exist on what OD is and should be, some overarching characteristics can nevertheless be identified (Cummings and Worley 1993):

(a) *System-wide effort*; that is, OD is being applied to a whole system (company, division, subsidiary, plant, etc.) and addresses comprehensively the change of both structure and behavior. The basic philosophy is that sustainable changes require the involvement of the whole system. Isolated small-scale changes are likely to fail. While OD aims at changing the entire system, the interventions address various levels: the individual, the group, intergroup relations, the whole unit, and interorganizational relations.

(b) *Planned organizational change*; that is, OD conceives change as a planned, systematic effort working on a long-range perspective.

(c) *Based on social sciences* that is, OD methods and procedures are derived from knowledge of various scientific disciplines.

(d) *Full cycle*; that is, OD includes all periods of a change project: its initiation (unfreezing), its realization (moving), and the subsequent stabilization (refreezing).

(e) *Change agent*; that is, OD projects are conducted by external professional consultant facilitators who are trained in the use of sophisticated intervention techniques and clinical methods. It is assumed that organizations often do not know what is wrong and they need professional help in diagnosing the problems and in overcoming the invisible barriers to change.

(f) *Top management support*; that is, commitment of top leadership is considered a precondition of any successful OD project. A commitment is expected in both paying attention to the change problems and giving the OD intervention a sense of urgency.

(g) *Shared power approach*; that is, the OD process is designed to involve all parties affected by the change project and to encourage participants to jointly develop new solutions.

(h) *Dual objectives*; that is, OD projects are intended to reach both organizational effectiveness and fulfillment of individual needs.

From time to time, new themes and fads were included in addition; for example, quality management, learning organization, transformational leadership, organizational culture, strategic vision, intercultural communication. The inclusion of these parts, however, never resulted in a paradigm shift or an essential theoretical revision. The overall framework basically remained the same.

5. Organizational Development Models

Various authors and consultants have developed comprehensive OD programs; among them four models are salient: (a) Survey-guided Development, (b) Process Consultation, (c) Grid Organization Development, and (d) Systemic Intervention.

5.1 The Survey-guided Development

This model is based on a multistep data feedback procedure (Likert 1967, Nadler 1977). After an orientation stage a standardized questionnaire is used to gain a comprehensive picture of the client organization's situation, its major problems, and their causes. The subsequent data feedback is used to determine what changes are required, how they should be reached, and what interventions are appropriate. The data feedback also aims at unfreezing participants and encouraging their willingness to take part in the change process. Follow-up surveys serve to monitor the progress of the OD process and to determine whether the change efforts did work. A somewhat similar abbreviated form of this model is the *Confrontation Meeting* originally developed by Beckhard (1967). Data collection and evaluation are conducted directly and in a condensed form (one day only). A meeting of all those involved is held to identify the major problems of the system. After a joint ranking of the problems the unit develops an action plan and establishes a system of periodical reports on the progress of implementation.

5.2 Process Consultation

This is a nondirective model for supporting the client organization in carrying out change processes, originally developed by Schein (1988). The process consultant helps the client organization to figure out what should be done to improve the situation. A process consultation project typically includes the following stages: Entering and defining the relationship, choice of setting and work methods, data gathering, diagnostic and confrontive interventions (e.g., through the use of feedback or coaching) by the consultant, evaluation of results, and disengagement. Due to its indirect character the results of process consultation are not predictable at the beginning.

5.3 Grid Organization Development

This is a highly standardized intervention program building on one basic conception; that is, an ideal managerial style with high concern for both people and productivity (Blake and Mouton 1969). The development contains six distinct phases: Grid Seminar, teamwork development, intergroup development, developing an ideal model for the whole organization, implementing the model, and examination of the organization's progress toward the plan.

5.4 Systemic Intervention

This is a more recent OD model that builds on family therapy and systems theory. At the core stand paradoxical transactions as invisible barriers to organizational change thereby drawing on the Double Bind Theory of schizophrenia (Watzlawick et al. 1967). The transformation of these ideas into an organizational development approach originally was initiated by the Italian psychotherapist Selvini-Palazzoli and her Milan School. This approach focuses on the hidden rules and games in organizations. In many cases this agenda is pathological in character, its circular reinforcement is likely to consume a lot of the system's energy and to evoke aversions against innovation and change. As a consequence, this approach refrains from any participative change design. Rather, it addresses the hidden agenda and its paradoxical architecture. OD interventions to open the system are conceptualized as a counter paradox (Selvini-Palazzoli 1995). Consultants try to identify the ruling paradox of the pathological system and to dissolve it by implanting counter paradoxes. Systemic interventions mostly result in 'explosions'; that is, the whole architecture of pathological rules breaks apart and unpredictable new patterns emerge.

6. Organizational Development and Beyond

The OD approach in the beginning was received enthusiastically in both theoretical and practical fields but later became increasingly criticized. The most common objection is to the underlying promise of harmony of interests; that is, to meet always both employee needs and the organization's efficiency goals. Critics doubt whether there is always such a win–win solution for interest conflicts. This argument leads into a broader criticism that OD shows a general naïveté towards organizational power structures. It is said that OD can never to be neutral because it always operates within power structures and is likely to become part of micropolitics.

Another objection relates to hidden influences. It is suspected that OD-intervention techniques leave considerable scope for manipulation because 'clients' understand their logic and their implications at best partially. Organizations and their members, therefore, may become objects of uncontrollable influences (Neuberger 1991, McKendall 1993). Also, the financial dependence of the change agents is emphasized. It does not allow them to play the alleged neutral role in the change process. They are paid by corporate

management and, therefore, are expected to pursue the interests of this group.

In addition to these objections a lack of sufficient theoretical foundation has been diagnosed (Beer 1990). The focus is on techniques and instruments but there is no compelling overarching conceptual framework; the empirical evidence is small.

From a practical point of view, OD projects rarely reach the status of a core activity. In most cases, it is considered as an interesting extra-activity, but not as a necessary element of corporate management.

Recently, a new debate on the nature of change came into the foreground of academic and practical interest (Brown and Eisenhardt 1997, Schreyögg and Noss 2000). It goes beyond the criticisms mentioned above and fundamentally challenges the paradigm of organizational change underlying OD models. An in-depth analysis of the OD approach revealed three highly questionable assumptions on the nature of change.

6.1 Organization Change is an Area for Specialists Only

The strong psychological or even psychotherapeutic orientation of OD led to the implicit assumption that organizational change can be achieved by specially trained professionals only. The organization or department is put into the role of a client or even a patient who badly needs a 'doctor' to master the challenge (illness) of change.

This perspective puts managers into a paradoxical situation. They are expected to rationalize and improve procedures and structures, but the important function of putting improved systems into action is said to be an exercise of which they are not capable. Change, however, nowadays is ubiquitous in organizations. It is a challenge for every manager and cannot be reserved for specialists only.

6.2 Organizational Change is a Clear-cut Process

In OD, organizational changes are conceived as clear-cut processes with a clearly defined beginning and end; processes that are separated from the ordinary organizational processes. This view essentially misconceives organizational processes. Problem-solving processes largely overlap and interpenetrate. This is the result of the fact that organizatons are, after all, to be managed in a complex and ambiguous context. Change processes are an integral part of the system's general dynamics; they are not separate entities or events. The definition of a change period is always an artificial construction and does not reflect the (actual) flow of events.

6.3 Organizational Change is an Exception in Organizational Life

In OD change is given a special status, that of an exception. It is conceived of as an extraordinary phase, which is threatening and somewhat chaotic. Changes come close to disturbances likely to undermine the rational order and system of rules. The guiding idea is the stable state of equilibrium. Stability stands for efficiency, rationality, and harmony. Organizational members are assumed to be strongly in favor of order and stability, and to feel helpless and frightened in the face of change. These assumptions fly in the face of nearly all more recent models of *flexible organizations*. Modern organizations are conceived as *inherently restless* and fluid in nature. The most recently discussed conceptions of a *Virtual Organization* or a Modular Organization point in the same direction. All these conceptions fundamentally challenge the view of organizational change as an exceptional state.

7. Outlook

The critique of the basic assumptions showed that a new conceptual orientation is needed which allows for overcoming the episodic view of change and offers a template to understanding change as a *continuous challenge*. The conception of the learning organization seems promising for providing such a reorientation. This concept builds on the idea that organizations operate on the basis of dynamic and not stable elements. The elements constitute and are constituted by communication. Organizational change occurs within streams of learning processes that build the basis of the organization's activities and strategies. By implication, the learning organization no longer experiences the issue of change as an exceptional stage; rather learning and change become common issues in the system's operations. Portraying organizations as self-reflexive learning systems makes it difficult to find reasons why an organization should suddenly cease to learn; that is, stop cognitive enacting, information processing, and knowledge development.

While the basic OD concept of change seems to be questionable, it is not necessary to give up all those well-developed OD methods and experiences. OD methods can be reframed. The mission of OD could be reformulated to help identify communication disturbances and to eliminate learning blockades (Argyris 1985). The purpose of the OD methods would then be to support the processes of learning to learn, and also to guarantee the competency of the system. They would become endogenous system catalysts that can help the ongoing learning processes to function smoothly.

See also: Decision Making, Psychology of; Design and Form: Organizational; Innovation: Organizational;

Institutions; Lewin, Kurt (1890–1947); Organizational Change; Organizations: Authority and Power

Bibliography

Argyris C 1985 *Strategy, Change and Defensive Routines*. Pitman, Boston

Beckhard R 1967 The confrontation meeting. *Harvard Business Review* **45**(4): 149–55

Beer M 1990 *Organization Change and Development: A Systems View*. Goodyear, Santa Monica, CA

Bion W R 1961 *Experiences in Groups*. Basic Books, New York

Blake R R, Mouton J S 1969 *Building a Dynamic Organization through Grid Organizational Development*. Addison-Wesley, Reading, MA

Brown S L, Eisenhardt K M 1997 The art of continuous change: Linking complexity theory and time-paced evolution in relentlessly shifting organizations. *Administrative Science Quarterly* **42**: 1–34

Cummings T G, Worley C G 2000 *Organization Development and Change*, 5th edn. South-Western College, Cincinnati, OH

Lawrence P R 1954 How to deal with resistance to change. *Harvard Business Review* **32**(3): 49–57

Lewin K 1958 Group decision and social change. In: Maccoby E E, Newcomb T M, Hartley E L (eds.) *Readings in Social Psychology*. Holt, Rinehart and Winston, New York, pp. 197–211

Likert R 1967 *The Human Organization: Its Management and Value*. McGraw-Hill, New York

McKendall M 1993 The tyranny of change: Organizational development revisited. *Journal of Business Ethics* **12**: 93–104

Nadler D 1977 *Feedback and Organization Developments: Using Data-based Methods*. Addison-Wesley, Reading, MA

Neuberger O 1994 *Personalentwicklung* (*Human Resources Development*), 2nd edn. Enke-Verlag, Stuttgart, Germany

Rice A K 1958 *Productivity and Social Organization: The Ahmedabad Experiment*. Tavistock, London

Schein E H 1988 *Process Consultation*, 2nd edn. 2 Vols. Addison-Wesley, Reading, MA

Schreyögg G, Noss C 2000 Reframing change in organizations. In: Havlovic S J (ed.) *Best Papers Proceedings of the Annual Meeting*. Academy of Management, Toronto

Selvini-Palazzoli M 1995 *The Hidden Games of Organization*. Pantheon, New York

Watzlawick P, Beavin J B, Jackson D D 1967 *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*. Norton, New York

G. Schreyögg

Development: Rural Development Strategies

Discussions of 'development' often confuse two meanings of the word. Development can be understood as the expansion of capitalist forms of production and exchange—i.e., as capitalist development; it can also be understood as an intentional and organized intervention with specified objectives (Cowen and Shenton 1996). These are two quite different meanings and there is no necessary relationship between them. Development interventions *may* be attempts to foster capitalist development, and they *may* (intentionally or not) have the effect of fostering such capitalist development. But many interventions do not have such intentions or effects, and at the very least often change the particular forms taken by capitalist political economy in specific places.

These observations apply equally to discussions of rural development. That is, some discussions have understood rural development as the further unfolding of capitalism in rural areas; others have focused on policy and project interventions in rural areas that aim to foster socio-economic change and human improvement. Among the latter work, some authors have understood such interventions as attempts to deepen rural capitalism (and particular relationships of power and dependency understood to go with it); others have aimed to understand the conditions under which progressive social change might derive from such interventions. As a result, under the broad banner of rural development studies, one encounters work on: theories of peasant economy informed by Chayanov and Marx (e.g., Harriss 1982); policy analyses informed by both dependency (de Janvry 1981) and neoclassical theory (Eicher and Staatz 1984); critiques of rural development discourses drawing on poststructural theory (Ferguson 1990; Escobar 1995); feminist interpretations of gender bias in the institutions and languages of rural development (Deere and Leon de Leal 1987); analyses of the encounters between development professionals and rural people drawing on both actor-oriented sociology (Long and Long 1992) and pragmatist approaches (Chambers 1983); and so on. In short, there is diversity in both the theoretical approaches taken and the substantive topics pursued.

Still, within this broad field of study, one can discern some pattern. First, writing on rural development has often followed broader trends in social theory. Second, theory and practice have informed each other. Policy innovations have often been informed by particular bodies of theory; they have then gone on to generate new bodies of research and theoretical development as scholars begin to reflect critically on these shifts in practice. Third, debates on theory and practice have been informed by broader shifts in the global economy. This reflects the interactive nature of the relationship between rural development as capitalist development and rural development as intervention. This is a relationship in which political economy clearly patterns practice, and where practice not only reproduces the broad structures and effects of capitalist political economy, but can sometimes change them. The remainder of this article tries to capture

some of these interactions. It focuses on rural development over the last three decades of the twentieth century.

1. A Brief History of Rural Development Policy: 1970–2000

While colonial administrations pursued programs to foster rural economic change (Cooper and Packard 1997) both in their colonies as well as in their own countries, the apogee of rural development interventions came in the 1970s with the invention of 'integrated rural development programs' or IRDPs. An important factor in the expansion of rural development programs during the 1970s was the push by the then President of the World Bank, Robert McNamara, radically to redirect the Bank's finance towards poverty alleviation. As poverty—at that time—was deemed a primarily rural phenomenon, IRDPs became one of the principal means through which the Bank increased its poverty lending. In some cases (especially in Latin America), such initiatives came in the wake of programs of land reform and redistribution, and aimed to increase yields on the properties of new post-reform landowners; in other cases, they were targeted at an already existing peasantry.

While these programs had many components, they aimed above all to increase the physical and economic productivity of smallholder productions systems. Agricultural development was therefore a central theme, and IRDPs often disseminated high yielding crop varieties, promoted agrochemical use, provided credit, built irrigation systems, and in general fostered agricultural modernization. Subsidiary components included support to health and education, rural road building, and—occasionally and marginally—efforts to strengthen rural people's social organizations.

In retrospect, many IRDPs were conceived on the basis of insufficient knowledge of rural conditions: they were too large, designed too hastily, and were often hijacked politically (de Janvry 1981). They became highly bureaucratic, overly complex, and often ineffective affairs (Grindle 1986). Worse still, the era of large scale rural development programs also spawned a series of 'development disasters.' Examples here include programs of small farm development and colonization in the Amazon basin (which often led to deforestation and usurpation of lands from indigenous groups), and a range of programs introducing Green Revolution and other modern crop technologies into environments for which they were not appropriate.

It is important to note that the urge for rural development was not only a phenomenon of broadly capitalist economies. Socialist regimes also launched programs, often linked closely to large-scale and state-led efforts to organize collective forms of agricultural production and marketing (e.g., Soviet kolkhozes, Cuban state farms, Tanzanian ujamaa). While the wider policy and political context of these programs clearly differed from that in capitalist economies, in many respects rural development programs in these two broad contexts shared many similarities: the state played a strong and central role in their implementation; agricultural modernization was a primary objective; and the rural economy was seen as a producer of surplus to be transferred for investment (via the state) in industrial and urban activities. In all cases, these programs worked with highly simplified views of rural economy and society: simplifications that were necessary in order to make rural areas legible and manageable by a centralized state (cf. Scott 1998), even though they meant that the programs would inevitably encounter problems and often engender resistance (Scott 1985).

While this mix of routine problems and stunning failures did not necessarily undermine the broader argument for poverty-reducing interventions in rural areas, it did nourish a body of critical writing aimed at trying to understand why it was that such programs had taken the form that they did, and had such mixed results on the ground (see Sect. 2). Parallel to this more analytical writing, a more applied literature has also developed over time, asking many of the same questions, but with different languages and with the primary goal of rethinking appropriate strategies for rural development (see Sect. 3).

2. Theoretical Approaches to Understanding Rural Development

Just as the 1970s were the high point for large-scale rural development programs, so too was the period of the 1970s to mid-1980s one of particularly vibrant intellectual production on rural development. Much of this work was driven by the empirical observation that the great investments made by governments and international agencies in land reform, integrated rural development, and the Green Revolution seemed to have had little impact on rural poverty, nutrition, and welfare—and that in some cases they had indeed had quite negative effects on rural people and their environments.

This was a period when dependency theory, world systems theory, and modes of production theory were very influential in writing on the 'Third World.' Not surprisingly these same theories were marshaled by many to try to explain this limited impact. Alain de Janvry's (1981) landmark study of rural development in Latin America stands out in this context—indeed, it has had intellectual impacts well beyond Latin America. De Janvry argued that in practice rural development and land reform were rarely intended to do much more than further the maintenance and expansion of dependent capitalist economies. In those

cases where they had more radical objectives, he suggested, they were quickly reined in by dominant national and international social forces. De Janvry argued that the underlying logic of Latin American economies was one of 'functional dualism'—a dualism in which a core of capital accumulation benefited from and required the existence of a rural economy that supplied cheap food and cheap temporary labor. Indeed, this argument had wider resonance, paralleling for instance those made by scholars of Southern Africa who insisted that the South African economy required the existence of labor reserves based in rural areas to provide cheap labor to the urban, mining, and industrial economy. Rural development was, in these frameworks, understood primarily as a mechanism through which these functionally dual relationships were maintained.

In Africa, a similarly formative piece was Henry Bernstein's (1978) notion of the 'simple reproduction squeeze' in which African peasantries are enmeshed. Bernstein suggested that peasant economies in Africa were slowly decapitalized by adverse terms of trade, themselves grounded in exploitative social and political relationships. This implied that interventions that did not address these underlying social relationships would have little effect on rural poverty. A significant debate in this early period was that on the question of 'urban bias.' While scholars such as Michael Lipton argued that rural areas were underprivileged largely because policy was biased to urban areas, more critical scholars insisted that any such policy bias was in turn a reflection of more deeply seated, and far more pervasive, sets of sociopolitical relationships that sustained the privileges of classes (rather than cities) and that constituted the primary obstacle to any significant change in rural areas (Corbridge 1982). Work such as that of Bernstein and Byres was emblematic (and formative) of a broader body of work in this period that linked rural development to theories of peasant economy (Harriss 1982). This work hooked up to the broader debates within peasant studies on how to understand the peasant mode of production (a debate crystallized in the creation of the *Journal of Peasant Studies*).

Neoclassical approaches to rural and agricultural development shared many similar concerns to those of political economic analysis (Eicher and Staatz 1984). Questions of peasant economy (in particular agricultural decision making), policy formulation, rural markets, poverty, and food security were thus prominent in this literature. Perhaps the principal difference was that such approaches were less concerned with the social relationships underlying decision-making processes, market transactions, and policy formulation and implementation. Put another way, they were less concerned with questions of social power, and even less with theories of social power. Analytically, this led to analyses framed in terms such as peasant rationality, or market and policy distortion; practically it led to recommendations that in earlier periods laid considerable faith in government as an autonomous and potentially beneficent social force, and in later periods have emphasized the beneficial role that liberalized markets (of diverse types) ought to play in rural development.

In frameworks such as these, the desirability of rural capitalism is not an issue. Indeed, what is deemed a problem with rural development when viewed within more critical frameworks, is deemed its overall purpose when viewed neoclassically. As they emphasize notions of efficiency over issues of equity, empowerment, or rights, neoclassical approaches have increasingly argued that significant parts of the rural population ought leave agriculture, and in many cases rural areas more generally. The contemporary challenge of rural development when viewed this way is therefore to facilitate this outward movement of people, and to foster a more efficient allocation of resources to those producers able to use them most effectively. The empirical support for such a position appears to come from recent research that suggests that many peasantries are already 'disappearing' (Bryceson et al. 2000) even if much of this research would not invoke a neoclassical interpretation to understand such steady disappearance (Bebbington 1999).

Falling somewhere in between such political economic and neoclassical positions, approaches drawing on institutional economics have become more popular in the 1990s, if more so in academic than policy realms (Carter 1997). Such approaches have also emphasized the need to understand rural markets (and household decision making) as embedded in social institutions, and emphasize the extent to which these institutions structure patterns of access to and participation in such markets. The implication is that rural development strategies ought to work on these institutional questions, a lesson not unlike that of political economy approaches. However, institutional economists insist on using the more formal methodological procedures of economics and econometrics to address these questions, and also hold back from assuming that any particular social theory (of power, class, race, or otherwise) can necessarily explain the particular forms of embeddedness of given markets.

Certain sociological and anthropological approaches have also emphasized such institutional questions, though they hold questions of politics and power at the forefront of their analyses; indeed their concern for these questions predates the emergence of such a concern in economic analyses of rural development. While the 1960s saw work on the management of rural development, it was perhaps only in the 1970s and especially the 1980s that a more coherent body of analytical work on the institutional and organizational dimensions of rural development began to emerge (e.g., Esman and Uphoff 1984). Much of this was also inspired by the question: why had land

reforms and rural development programs been so disappointing? Work drawing on Weberian notions of the state began to demonstrate the effects of bureaucratic structure, incentives, and politics on the implementation and outcome of rural development programs (Grindle 1986, Hart et al. 1989). Other work, drawing more on theories of peasant resistance to development suggested that much of the explanation lay in the ways in which rural people subverted institutional interventions designed to foster forms of modernization that people did not desire (Scott 1985).

Though focusing on different parts of the rural development puzzle, taken together, these two approaches helped open up a body of work that has emphasized questions of human agency in rural development. Some of this work, drawing more on notions of resistance, has emphasized the extent to which rural development programs constitute negotiations between state and citizen, and between different notions of modernity (Long and Long 1992, Peet and Watts 1996). Other authors have linked a concern for social power with the goal of devising improved management systems. Much of this work has focused on the roles of particular types of organization in reworking power relationships between rural people, the state, and market, and in simultaneously initiating and managing rural development programs that might be more poverty oriented and responsive than were the IRDPs. Local organizations (Uphoff 1992), nongovernmental organizations (Farrington and Bebbington 1993) and social movements (Peet and Watts 1996) have figured prominently in this work—linking up discussions of development with broader theoretical debates on civil society, power, and identity.

These concerns for questions of power and identity also underlie a more recent line of analysis focusing on questions of discourse in rural development and using poststructural approaches to address ways in which power is embedded in the very ways in which rural development is talked about and conceptualized (Escobar 1995, Ferguson 1990). Such authors have also focused attention on the often-underemphasized theme of culture in rural development: both by addressing how notions of development are now often constitutive of both rural and institutional cultures, and by interrogating the cultural underpinnings of the very concept of rural development. This approach has linked discussions of rural development up with debates in Foucauldian social science, and has been particularly prominent in anthropology, geography and the subfield of political ecology, which has been one of the most vibrant areas of research on rural development in the 1990s (Peet and Watts 1996). While such poststructurally informed work has been criticized analytically (especially for its weakness in addressing issues of agency and of variation) and practically (for its failure to address how the practice of rural development might be made more effective), it has been the most fruitful and dynamic of the subfields of rural development research in the closing years of the century.

Running across these broad theoretical traditions one encounters a series of more substantive foci. One of these has been an ever-increasing attention to issues of gender and rural development, and the frequent exclusion of women from the fruits of rural development, sometimes as a direct result of the forms taken by projects and policies (Boserup 1970, Carney and Watts 1990). Much work on rural development and peasant economy remained largely silent on issues of gender—a silence that became even more untenable the more it became apparent that the rural economy could only be understood by making explicit the shifting roles of and relations between women and men (Deere and Leon de Leal 1987, Carney and Watts 1990). The central importance of women as rural producers has also become ever clearer as periodic out-migration from rural areas continues to increase. A concern for gender issues has cut across a range of theoretical positions: it has appeared in both neoclassical and Marxian analyses of household decision making and bargaining; in more institutionally focused analyses of 'women-in-development' and of women's social movements; and in poststructurally informed analyses of the ways in which rural development discourses create particular notions of gender, and then act upon these.

Given that rural development is closely linked to production, and that much rural production depends directly on natural resources, environmental issues have also been a recurrent theme in analyses of rural development. Perhaps more significantly, many of the 'disasters' of rural development have involved grave and negative impacts on the environment, spawning many critical studies. More generally, analyses of rural development experiences with irrigation, dams, agrochemical use, soil erosion, and high-yielding crop varieties have inspired work that has made critical contributions to the more general discussion of sustainability and sustainable development (Adams 1990, Peet and Watts 1996).

Finally it ought to be noted that, while much of the analytical (and practical) work on rural development has focused on Africa, Asia, and Latin America, there has also been important work conducted in the USA, Europe, and post-reform Soviet Union. Much of this work has drawn on similar theoretical currents, though, arguably, it has been more descriptive than the often quite theoretical work conducted in the 'Third World.' One of the important contributions of this body of 'First World' work, however, has been to raise questions about the longer-term future of rural societies and environments as agricultural production becomes ever less important within national economies, and as the agricultural labor force continues to decline. This work has addressed some of the ways in which rural landscapes change as they move from

being areas of production to areas of reproduction, leisure, and consumption; it has also addressed the acute poverty of those areas (in the developed world) that have simply become unnecessary to the wider economy. Though apparently peculiar to developed-country settings, these issues will become relevant for other regions of the world. For even if small farmer production becomes increasingly uncompetitive in the global economy, rural areas will not disappear. In some cases, they will be populated by some of the most excluded groups in society, unable to find work in a rapidly modernizing economy—in other cases they will be colonized by retirees, commuters, tourists, conservationists, and new enterprises. Such demographic, economic, and social shifts will continue to raise acute questions for the ways in which societies understand themselves, their lifestyles, and their environments, and, indeed, for the different forms that modernity will take in the post-peasant era. Rural development is likely, then, to continue generating important theoretical and empirical debate that will have relevance beyond the purely 'rural.'

3. *Applied Research and Rural Development Strategies*

Alongside these vigorous theoretical debates, a lively and creative body of applied research has also evolved over the last thirty years, dedicated to addressing issues of rural development strategy. While there are some journals dedicated to applied research in which this material can be published, much output remains in institutional publications, working papers, and gray material produced by nongovernmental organizations, nonprofit and international research centers, some university institutes, and development organizations. Though applied in focus, and often undervalued by academia, such documents have often been at least as important as more formal academic publications in pushing forward debates on themes as varied as definitions of poverty, participatory and populist approaches to development, systems analysis in rural areas, indigenous technical knowledge, sustainability, and empowerment. Furthermore, perhaps because much of this work has been conducted from within nongovernmental organizations, it has also raised important questions about relationships between state, market, and civil society in rural development programs, and as a result about the institutional foundations of sustainability.

Of the various themes that have been dealt with in this work, two merit particular attention. The first has been the discussion of rural development administration. A long-standing discussion, this has helped illuminate the different ways in which state, nongovernmental, for profit, and popular organizations might participate in the management and implementation of rural development. An overarching and recurrent theme in such discussions has been how to deal with the complexities and uncertainties that inhere in rural development, and how to design and structure management systems that can best address these complexities without becoming bureaucratic and top-heavy, as so many IRDPs did. One of the general themes to emerge from this sustained inquiry is that increased popular participation must be central to rural development: both in order to handle these complexities, as well as on its own democratizing merits.

The second body of work worthy of note is that which has attempted to understand the logic of rural production, livelihoods, and ecologies in order to elaborate more appropriate forms of rural intervention. In the 1970s, systems-based analyses of agriculture and rural development occupied a prominent place in this applied research. The more that this work made explicit the complexity of these systems, the more it bled into work on popular participation in rural development. The implication, perhaps more implicit than explicit, was that in the face of such complexity, effective management of rural development necessarily had to involve the knowledge of local populations and other local actors. Thus, while earlier statements suggested that the only way in which development interventions could be adequately fine-tuned was through comprehensive *ex ante* systems analysis, later work concluded that such analysis could never be sufficiently comprehensive. Instead, it has suggested the need to shift emphasis from exhaustive *ex ante* analysis to more systematic participation in the process of managing, implementing, and governing rural development.

These reflections from applied research have thus linked discussions of rural development with a far longer-standing concern in activist social science for questions of critical pedagogy, popular education, and empowerment (Freire 1970, Fals-Borda and Rahman 1991). Indeed, beyond the specific themes addressed in such applied work, perhaps more important are the questions of epistemology (and methodology) that it has raised. Applied researchers have, in effect, suggested that practice generates different forms of knowledge from those that emerge from more basic social research. As such they have echoed questions raised by poststructural approaches to rural development (and knowledge more generally) in suggesting that researchers need to be far more modest in the knowledge claims they make, and that the process of generating knowledge about rural development (as theory or as strategy) must be far more inclusive and participatory than it has traditionally been (Blaikie 2000). The relationship between research and practice, and theory and praxis, in rural development is likely to be a hotly debated, and fruitful, theme for many years to come.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of;

Boserup, Ester (1910–99); Development: Sustainable Agriculture; Rural Geography; Rural Industrialization in Developing Nations and Gender; Rural Planning: General; Rural Sociology; Secular Religions

Bibliography

Adams W M 1990 *Green Development: Environment and Sustainability in the Third World*. Routledge, London

Bebbington A 1999 Capitals and capabilities: a framework for analyzing peasant viability, rural livelihoods and poverty. *World Development* **27**(12): 2021–44

Bernstein H 1978 Notes on capital and peasantry. *Review of African Political Economy* **10**: 60–73

Blaikie P 2000 Development, post-, anti- and populist: a critical review. *Environment and Planning A* **32**(6): 1033–50

Boserup E 1970 *Women's Role in Economic Development*. Allen & Unwin, London

Bryceson D, Kay C, Mooij J 2000 *Disappearing Peasantries? Rural Labour in Africa, Asia and Latin America*. Intermediate Technology Publications, London

Carney J, Watts M 1990 Manufacturing dissent—work, gender and the politics of meaning in a peasant society. *Africa* **60**(2): 207–41

Carter M 1997 Intellectual openings and policy closures: Disequilibria in contemporary development economics. In: Cooper F, Packard R (eds.) *International Development and the Social Sciences. Essays on the History and Politics of Knowledge*. University of California Press, Berkeley, CA, pp. 119–49

Chambers R 1983 *Rural Development: Putting the Last First*. Longman, Harlow, UK

Cooper F, Packard R (eds.) 1997 *International Development and the Social Sciences. Essays on the History and Politics of Knowledge*. University of California Press, Berkeley, CA

Corbridge, S 1982 In: Hariss J (ed.), *Urban Bias, Rural Bias and Industrialization: An Appraisal of the Work of Michael Lipton and Terry Byres*, pp. 94–116

Cowen M P, Shenton R W 1996 *Doctrines of Development*. Routledge, London

de Janvry A 1981 *Land Reform and the Agrarian Question in Latin America*. Johns Hopkins University Press, Baltimore, MD

Deere C D, Leon de Leal M 1987 *Feminist Perspectives on Latin American Agricultural Development*. Westview Press, Boulder, CO

Eicher C K, Staatz J M (eds.) 1984 *Agricultural Development in the Third World*. Johns Hopkins University Press, Baltimore, MD

Escobar A 1995 *Encountering Development: the Making and Unmaking of the Third World*. Princeton University Press, Princeton, NJ

Esman M, Uphoff N 1984 *Local Organizations: Intermediaries in Rural Development*. Cornell University Press, Ithaca, NY

Fals-Borda O, Rahman A (eds.) 1991 *Action and Knowledge: Breaking the Boundary with Participatory Action Rresearch*. Apex, New York

Farrington J, Bebbington A 1993 *Reluctant Partners: NGOs, the State and Sustainable Agricultural Development*. Routledge, London

Ferguson J 1990 *The Anti-politics Machine: Development, Depoliticization and Bureaucratic Power in Lesotho*. Cambridge University Press, Cambridge, UK

Freire P 1970 *Pedagogy of the Oppressed*. Seabury Press, New York

Grindle M 1986 *State and Countryside. Development Policy and Agrarian Politics in Latin America*. Johns Hopkins University Press, Baltimore, MD

Harriss J 1982 *Rural Development: Theories of Peasant Economy and Agrarian Change*. Hutchinson, London

Hart G, Turton A, White B 1989 *Agrarian Transformations: Local Processes and the State in Southeast Asia*. University of California Press, Berkeley, CA

Long N, Long A (eds.) 1992 *Battlefields of Knowledge: The Interlocking of Theory and Practice in Social Research and Development*. Routledge, London

Peet R, Watts M (eds.) 1996 *Liberation Ecologies: Environment, Development, Social Movements*. Routledge, London

Scott J 1985 *Weapons of the Weak. Everyday Forms of Peasant Resistance*. Yale University Press, New Haven, CT

Scott J 1998 *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT

Uphoff N 1992 *Learning from Gal Oya: Possibilities for Participatory Development and Post-Newtonian Social Science*. Cornell University Press, Ithaca, NY

A. Bebbington

Development: Social

If it is true that concepts lack meaning in and of themselves, but rather acquire significance in the context of a set of concepts articulated in what is commonly understood as a theory or a paradigm, then at least as it has been used in the literature the concept of social development is somewhat of an exception. This is true when one compares social development with the notion of economic development, about which there has existed a clear consensus both around a set of powerful theoretical referents (Keynesianism, the neo-classical school, or Marxist radicalism) and around how practical objectives are to be achieved. If the comparison is extended further to encompass political development, the absence of such consensus is mitigated somewhat by the clarity of the goals and objectives to which the concept refers: the construction of the nation state and of democracy. Yet neither of these referents, theoretical nor normative, has played a similar role in the use of the concept of social development.

Social development thus has been a concept with multiple meanings, encompassing at least three connotations developed extensively in the literature pertaining to the social sphere. At its most basic level, social development simply suggests improvement in the conditions and quality of life of the population. Greater levels of wealth, technological advancement, and public policies permit people to live better, to consume more, to feed themselves better, and to get

sick less frequently. This idea of social development is inextricably linked to the idea of economic and material advance of human society. A second approach emerges from considering social development in terms of the distribution of the wealth that societies generate. Here, the economic dimension still constitutes a powerful factor in determining social development, but the latter occupies a space of its own, in the distributive sphere. Finally, a less normative or economistic perspective on social development identifies the latter with the processes of social differentiation and social complexity associated with the capitalist and industrial transformations that gave rise to the modern world.

1. Sketching a Genealogy

In the aftermath of the World War II, and at the height of the period of worldwide decolonization, an important body of literature was produced as a result of concern with economic growth, equity, the establishment of democratic regimes, and the achievement of social order and national autonomy. The basic objective of this literature was to discover appropriate paths through which new nations could advance toward these goals. The term social development originates during this period, and is anchored in an international context characterized by profound transformations of the world order and system of nations.

The notion of social development is as recent as the notion of development itself, and thus stands in contrast to the historical tradition of studies of economic growth and social change carried out primarily in the eighteenth and nineteenth centuries. In the past, attempts to characterize the principal traits of capitalism of that era employed a discourse that relied on different terms. The international system of nations was conceptualized—depending on preferences or discipline—in terms of various binary categorizations: rich and poor countries; industrial or agrarian societies; traditional and modern societies; classic capitalism and late capitalism; or colony and metropolis. Only since the middle of the last century did there emerge an increasing perception that the world was divided in terms of concepts of relative development: of development and underdevelopment. With that shift came the introduction of notions of developing countries. It was during this period as well that researchers elaborated distinctions between center and periphery to understand the insertion of countries into the world economy. This approach was developed subsequently in significant directions by the writings of the Economic Commission on Latin America (ECLA), and contributed to the development of dependency theories (Pinto 1979, Cardoso 1977). Because the concept of social development has its origins above all in the literature produced in, or concerned with, the development of new nations, much of that literature was published in Latin America or in academic and political circles animated by the problems of the great transformation brought about by the economic and political reorganization of the international system of nations. Beginning in the late 1940s, the thinking of ECLA played a notable role in this process. Especially influential were its founding ideas, developed by Raul Prebisch following publication of his classic 1949 article 'Estudio Economico de America Latina.' (See also CEPAL 1989, Gurrieri 1979.)

2. Social Development and Developmentalism

As noted at the outset, it is useful to distinguish social development as a body of concepts and theory as used in academic circles from its prescriptive or normative meaning directed toward action or policy ('planning,' to use the terms of the era). Strictly speaking, there is no theory of social development or similar body of conceptual thinking. The notion of social development formed part of the more general and prevailing idea that economic development was normatively charged. Indeed, this concept of social development originates amid a climate of failure of the assumptions of classical economics regarding social and economic change in late capitalist countries. In particular, it reflected a response to the findings of studies of international trade by such authors as R. Nurkse, R. E. Baldwin, and the Manchester School, or the notion of the 'big push,' rooted in the ideas of H. Ellis (Rosestein-Rodan 1954, Ellis 1961). At the same time, in addition to being a body of concepts articulated in the form of a paradigm, economic and social development was constituted in an ideology known at the time as developmentalism (Hodara 1975). Transforming the objective of development into ideology served to offer a complex system of beliefs that explained social and political arrangements in the world and in society. It thus legitimized and rationalized behaviors and social relations, at the same time that it served as a base for solidaristic and collaborative forms of action motivating individuals to undertake certain kinds of actions. The power of the ideology reflected its capacity to provide meaning to the relationship between circumstances experienced by individuals and those affecting the nation as a whole. Prospects for the eventual improvement of the status of individuals in society could not be separated from the situation and status of the country in the international system.

But if social development had no single meaning from an academic point of view, the elaborations on the concept represent conceptual tributaries of a long tradition of studies and elaborations on social change or social progress. These ideas were central to the thinkers of the nineteenth and early twentieth

centuries, and later would foster formulations of social modernization theory in its structural functionalist versions or in *dependentista* currents.

To a greater extent than the great evolutionary interpretations of Comte, Morgan, or Spencer, or the visions of progress advanced by Condorcet or Saint-Simon, the intellectual currents that situated social change in the context of development were more specific and revived the legacy of a more recent sociological tradition (Nisbet 1966). Especially important was Durkheim's basic concern with the effects of labor and social differentiation on moral density. That relationship was expressed in the tensions of passage from mechanical solidarity to organic solidarity, in the polar structures of community and society of *Gemeinschaft–Gesellschaft*, and in the transformation of small, primitive, traditional, homogenous societies and primary social relations toward their social opposite, as formulated by F. Tönnies, or in the economic sociology of Max Weber on the evolution of the capitalist spirit and the inevitable progress of rationalization and bureaucratization of Western societies; in the process of transformation of authority and autonomization and objectification of the status described by Simmel; or in the sociology rooted in the Marxist tradition and the theorists of conflict. Yet it was Modernization theory, anchored in the work of Parsons, and later in developmentalism, that proved of greatest influence in the analysis of social development in Latin America. Indeed, this is true to the extent that is frequently difficult to distinguish the concept of social development from that of social modernization (see Germani 1963, Moore 1969). In turn, dependency paradigms, particularly in the versions put forth by G. Frank and T. Dos Santos, were centered more in the economic and political spheres than in the realm of social processes.

In its initial formulation, the objective of social development formed part of an optimistic vision of the future of the new nations. Consistent with acceptance of that vision were the two other great goals to be achieved, one at the level of economic development and the other at the level of politics. The perception was that in one way or another, albeit with disjunctures or frictions and with transitory tensions produced by the unequal advance of different dimensions of development, new nations—poor, unequal, repressive, violent, and dependent—would move gradually toward societies characterized by welfare, equity, order, democracy, and autonomy. This evolution was believed likely to follow a path not very different from that of Western societies in Europe and the United States. The literature of the 1960s on social and political modernization was impressive and engaged the talents of diverse students of sociology and political science who shared a common interest in revealing the paths that would permit 'peripheral' societies to overcome their pasts and incorporate themselves into the core (e.g., Almond and Coleman 1960, Lipset 1981, Pye 1962, Deutsch 1963, Almond and Powell 1966). Nevertheless, not all were entirely optimistic, as noted by Huntington (1996). During an initial phase there predominated the view that the objectives of social, political, and economic development were inherently compatible; in a second phase, emphasis was placed on the intrinsically conflictive nature of the three goals; and a third phase was marked by arguments concerning the imperative of generating policies capable of reconciling the goals.

3. *Stages in Latin America*

It is not difficult to distinguish a first phase of the Latin American literature, in which social development was presented as a non-specific set of ideas subordinate to economic development. In the initial writings of ECLA, the terminology 'economic and social development' refers to a single thing. In practice, however, the basic assessment of ECLA stemmed from critical analysis of economic questions and of international trade during the late nineteenth and early twentieth centuries. The criticism of the passive role that classical international trade theory assigned to countries of the periphery was that this effectively denied them a capacity for dynamism and autonomy over the long term (Nurske 1963, Perroux 1958, Baldwin 1956, Barré 1958). Central to this interpretation was the concept of deteriorating terms of trade, which were understood as systematically unfavorable to the periphery. This was seen as the cause of the vulnerability of the productive systems of underdeveloped countries and of the power that traditional groups linked to primary sector exports acquired in those countries. The resulting call for inward-oriented development, based on a push for import-substituting industrialization, productive diversification, and the formation of a modern national entrepreneurial class, was expected to augment national decision-making capacity vis-à-vis the international system. As a natural consequence of these transformations, achievement of several other development goals would be possible: political democratization, cultural independence, and social integration. During this initial phase, social development was understood essentially as improvement in the conditions of life and as a passive dimension of the development process. It is no surprise, then, that the titles of works of the era made reference to 'the social consequences of economic development' or 'social aspects of economic development' rather than to social development *per se*.

Within ECLA itself, the work of J. Medina Echevarria (1955, 1963), of unequivocally Weberian roots, opened important spaces for the consideration of the social as a cause and not merely as a consequence of economic development, particularly with regard to the study of social, cultural, and political prerequisites for economic development and for its emphasis on

linking sociological variables to economic theory. Nevertheless, his ideas developed in parallel to those of the mainstream, and were never incorporated into the core of ECLA thinking.

During a second phase, coinciding roughly with the end of the 1960s and through the 1970s, with the advance of studies in comparative politics and society, doubts began to emerge about the possibility that countries in the periphery would manage to move simultaneously along a path of social, economic, and political development. The notion of compatibility or harmony among these development goals was increasingly questioned. During this phase, research produced in core countries and in other regions of the world had a greater impact on the elaboration of Latin American thought at the same time that ECLA lost its leadership role in thinking about development.

The redefinition of the concept of social development, and above all reconsideration of its role in the general process of development, was influenced by disenchantment with the failure to fulfill expectations. Especially troubling were low levels of economic growth, persistent political instability, the presence of authoritarian regimes, and high levels of indigence and poverty derived from the incapacity to reduce indices of social inequality. The apparent incompatibility of objectives was articulated in the questioning of much of the conventional wisdom about development, in particular with regard to such topics as the relationships between growth and equity (in contrast to the classic thesis of Kuznets; in the controversy over the relationship between democracy and growth (an idea deeply rooted in liberal thought); or about the role of poverty, indigence, and inequality as limits on economic development. The literature on the topic adopted a vision that, if not more pessimistic, advanced at least a more complex and cautious view about the ease of transformation (see Furtado 1996, Olson 1963, Flanigan and Fogelman 1971, Berger 1976, Hewlett 1980, Huntington 1996). At ECLA publication of *El Cambio Social y la Politica de Desarrollo Social en American Latina* (United Nations 1969) reflected renewed attention to the social question, which would become the core of subsequent activities developed by that institution's Division of Social Development.

During a third phase, the temporal boundaries of which are less precise, the term 'social development' became detached from any single current of ideas. In part this was because developmentalist ideology had lost its influence and in part because the confidence deposited in technocratic actions had given way to consideration of alternative logics of an economic, social, and political nature. At least four factors account for these changes. One is the exhaustion of the import-substitution or inward-oriented development model, which proved incapable of sustaining economic growth in the region and inducing the promised social and economic development. Another is the renewed strength of liberal economic orientations that portray the market as the locomotive of history, and that curtail the basic functions of the state expressed in the developmentalist ideology. A third reason is of a different nature, and concerns the sources of discourses about growth. Multilateral financial organizations, and particularly the Inter-American Development Bank and the World Bank, became two of the most important institutions for generating ideas and proposals about the development of peripheral countries, and followed a line consistent with the Washington Consensus and with neo-liberal orientations. Meanwhile, from the United Nations, and in particular from the United Nations Development Program (UNDP), there emerged alternative proposals and the notion of social development. A fourth reason is the competition to which the idea of social development was subjected by a growing array of new themes, concepts, and approaches which broadened and legitimized the inclusion of social goals in development. This competition was reflected in the introduction and increasingly widespread use of such notions as 'sustainable development' and 'human development.'

With regard to the social, there was a contradictory movement of differentiation and fusion. On the one hand, the notion of social development fragmented into multiple sub-goals and in the application of middle-range theories around newly legitimate issues and demands (women, environment, youth, ethnicity, culture, human rights, etc.). On the other hand, the concept of social development—replaced by human development in the UNDP version—implies a return to an integrated vision of the economic and the social. Debates had come full circle, as traditional sets of indicators designed to characterize the evolution of countries once again included economic variables (for example, GDP per capita) alongside social variables (for example, life expectancy at birth). This took place following differentiation between the systems of social indicators and the systems of economic indicators, a distinction that emerged with the 'Scandinavian movement' of the 1960s, and that would be adopted later on by the United Nations Research Institute for Social Development (UNRISD) and ECLA.

See also: Dependency Theory; Development, Economics of; Development: Social-anthropological Aspects; Development: Socioeconomic Aspects; Durkheim, Emile (1858–1917); Inequality: Comparative Aspects; Modernization, Political: Alternate Paths; Modernization, Political: Development of the Concept; Modernization, Sociological Theories of; Parsons, Talcott (1902–79)

Bibliography

Almond G, Coleman J S 1960 *The Politics of Developing Areas.* Princeton University Press, Princeton, NJ

Almond G, Powell G Jr 1966 *Comparative Politics: A Development Approach*. Little, Brown, Boston
Baldwin R E 1956 Patterns of development in newly settled regions. *Manchester School of Economic and Social Studies* **24**: 161–79
Barré R 1958 *Le Développement Economique: Analyse et Politique*. Institut de Science Economique Appliquée de París, Paris
Berger P I 1976 *Pyramids of Sacrifice: Political, Ethics and Social Change*. Doubleday Anchor, New York
Cardoso F E 1977 La originalidad de la copia: La CEPAL y la idea de desarrollo. *Revista de la CEPAL* **2**: 7–40
CEPAL 1989 *Cepal: 40 Years (1948–1988)*. CEPAL, Santiago, Chile
Deutsch K 1963 *The Nerves of Government*. Free Press of Glencoe, London
Ellis H (ed.) 1961 *Economic Development for Latin America*. St. Martin's Press, New York
Flanigan W, Fogelman E 1971 Patterns of democratic development: An historical and comparative analysis. In: Gillespie J V, Nesvold B A (eds.) *Macro-Quantitative Analysis: Conflict, Development, and Democratization*. Sage, Beverley Hills, CA
Furtado C 1996 *Subdesenvolvimento e Estagnaçâo na América Latina*. Civilizaçâo Brasileira, Rio de Janeiro, Brazil
Germani G 1963 *Política y Sociedad en una Epoca en Transición*. Paidós, Buenos Aires, Argentina
Gurrieri A 1979 José Medina Echaevarría; un perfil intelectual. *CEPAL Review* **9**
Hewlett S A 1980 *The Cruel Dilemmas of Development: Twentieth-Century Brazil*. Basic Books, New York
Hodara J 1975 Desarrollismo. In: *Diccionario de Ciencias Sociales*. Instituto de Estudios Políticos, Madrid, Spain
Huntington S P 1996 The goals of development. In: Inkeles A, Sasaki M (eds.) *Comparing Nations and Cultures*. Prentice-Hall, Englewood Cliffs, NJ
Lerner D 1962 *The Passing of Traditional Society*. Free Press, Glencoe, IL
Lipset S M 1981 *Political Man: The Social Bases of Politics*, Expanded edn. Johns Hopkins University Press, Baltimore
Medina Echavarría J 1955 *Las condiciones socials del Desarrollo Económico*. CEPAL, Santiago, Chile
Medina Echavarría J 1963 *Consideraciones Sociológicas sobre el Desarrollo Económico*. Solar/Hachette, Buenos Aires, Argentina
Moore W 1969 A reconsideration of theories of social change. *American Sociological Review* **25**
Nisbet R A 1966 *The Sociological Tradition*. Basic Books, New York
Nurkse R 1963 *Patterns of Trade and Development*. Basil Blackwell, Oxford, UK
Olson M Jr 1963 Rapid growth as destabilizing force. *Journal of Economic History* **23**: 529–52
Perroux F 1958 L'Idée de Pôle et les Ensambles Industriels en Afrique. *Sentier d'Europe* **1**
Pinto A 1979 La internacionalización de la Economía Mundial y la periferia. *Revista de la CEPAL*
Pye L W 1962 *Politics, Personality and National Building*. Yale University Press, New Haven, CT
Rosestein-Rodan F 1954 Las razones del atraso económico. *Panorama Económico* **94**
United Nations 1969 *El Cambio Social y la Política de Desarrollo Social en América Latina*. United Nations, New York

C. H. Filgueira

Development: Social-anthropological Aspects

In the years after World War II, development primarily meant economic development: directed change leading to greater agricultural productivity and industrialization. In later years, development moved from a preoccupation with economic growth to considerations of equity, sustainability, quality of life, and popular participation in decision-making. Development is in most cases funded and directed by external donors, governments, and nongovernmental organizations (NGOs).

1. *Development and Anthropology*

In the late 1940s and early 1950s, the US introduced the 'Point Four Program,' which resulted in the employment of several anthropologists, especially in Latin America. Erasmus (1954) provides a good early account of the desirability of taking into account the needs and capabilities of people, versus the perceptions of planners and administrators.

The British Colonial Service employed several anthropologists to help with administration. Gulliver (1985) gives an honest and revealing account of his period in East Africa as a Government Anthropologist, showing that ethnographic research can be used to inform social policy in a manner that made it more appropriate to local needs, and thus more acceptable to local people.

Here we shall be concerned mainly with the period from the early 1970s, when there was a disillusionment with the development policies of the previous two decades, which had been characterized by huge projects and centralized planning and guided by theories of evolutionary modernization and the 'trickle-down effect.' As Hoben (1982, p. 350) wrote, 'the ethnocentric tech-fix orientation of the dominant development paradigm' was challenged. Many indices—the rural quality of life, infant mortality and life expectancy, per capita GNP, food production—had not only failed to improve, but had often fallen. In addition, it was clear the gap between rich and poor—both between and within countries—was rapidly growing. This resulted in major shifts, in the early 1970s, beginning with two of the main donors, the World Bank and the United States Agency for International Development (USAID), redirecting their development projects more specifically at the poor (or even, 'the poorest of the poor'). USAID introduced a series of measures, under its 'New Directions' guidelines, which required that all projects should have a Social Soundness Analysis, to examine the probable consequences of each proposed project for the supposed 'beneficiaries' (or 'the target population,' an infelicitous phrase that is sometimes used).

Not all administrators welcomed these innovations, complaining that they wasted time and were superfluous; few administrators had training in the social and cultural dimensions of development, which they tended to ignore. The World Bank also introduced several sets of guidelines, e.g., for dealing with resettlement, where those resettled are supposed to receive 'land for land,' not a cash compensation. Host governments ('the borrowers'), powerful players on the development stage, have often been slow to implement the guidelines.

With the new emphasis on development for the poor, who were the people who knew anything about the poor? The development 'experts'—economists, lawyers, and administrators—

> were far more accustomed to working with the aggregated data of national accounts then with the nitty gritty, nonquantitative, imprecise, fuzzy and seemingly subjective data of individual farms, household and small rural communities (Horowitz 1999, p. 12).

So anthropologists, who *do* have a good understanding of poor peoples' lives, began to play increasingly significant roles on the development stage.

2. *The Anthropology of Development*

Anthropology is involved with development in two quite distinct ways: (a) the anthropology of development, 'an anthropological analysis of development as a cultural, economic and political process, rather than anthropology in action' (Gardner and Lewis 1996, p. vii); and (b) anthropology in action, or 'development anthropology.' (Cochrane (1971) was one of the first to use this term in print). Grillo (1997, p. 11) summarizes seven main themes in recent work in the anthropology of development. These include: an increased focus on culture and social relations; 'opposition to the marginalization of indigenous peoples and their knowledge'; emphasis on participation and empowerment; 'cynicism about the aims and practices of development'; and 'advocacy of alternative ways of doing both development and anthropology.'

Anthropologists have provided a critical understanding of the nature of development. This is illustrated by Scudder's and Colson's extensive, longitudinal studies of resettlement, among the Gwembe Tonga of Zambia (Scudder 1999), and by analyses of livestock and pastoralism by Baxter, Dyson-Hudson, Horowitz, Little and others (despite nearly one billion dollars having been spent on pastoral projects in East and West Africa, very little has been achieved, mainly because the planners ignored the detailed knowledge that pastoralists have of their environment, and failed to see the logic behind their transhumance). Similarly, numerous detailed investigations of agriculture have included explanations of the significance of household composition, non-farm income, adaptive strategies, decision-making, and mixed farming.

Anthropological analysis has also had a major impact on forestry. The Community Forestry Unit of the Forestry Policy and Planning Division of FAO has been directed by anthropologists (Marilyn Hoskins and Katherine Warner) for an extended period. It has provided global leadership on incorporating gender analysis, conflict management, participatory processes, and other social dimensions of development into natural resource management and policy. Other anthropologists, such as Michael Dove, have had a major influence on the social analysis of forestry.

The need for long-term research has been successfully demonstrated in a few cases, although there is still a long way to go to persuade all major agencies of the value of such studies. Where such studies have been promoted, they have led to clear and effective policy recommendations. Also, anthropologists are increasingly being appointed as team leaders of multidisciplinary teams, in examples of successful interdisciplinary cooperation. Such appointments have sometimes led to different emphases, more sensitive to the local people.

If they are to be truly effective in development, anthropologists should be involved in *all* stages of the project cycle: project identification, project design, project appraisal, project implementation, project monitoring, and post-project evaluation. Normally project appraisal occurs after the donors and host governments have agreed that the project will go ahead. The anthropologist often becomes involved only at the appraisal stage, and is likely to be seen by administrators as a spoiler, threatening timely implementation by pointing out flaws in the original design.

Despite the above examples, it seems that the major donors have been stronger in rhetoric than in action. For example, in the year 2000, the fraction of World Bank spending devoted to rural development and agriculture, the arenas within which anthropologists have been most active, is down to 7 percent from an earlier high of 40 percent. This hardly signals a clear commitment to 'poverty reduction.'

3. *Should Anthropologists be Involved in Development?*

Grillo (1985, p. 28) notes three perspectives on development. First, there are the 'principled rejectionists,' who think that anthropologists should not be involved—period. For a presentation of this approach, see Escobar (1991). Using a Foucauldian style of analysis of discourse, Escobar argues that development, because of the unequal power relations and control of means of production, is hopelessly biased in favor of western donor nations, and that it does not, and cannot, meet the needs and aspirations of the

poor. Escobar tends to treat development as homogeneous, ignoring the vast diversity that occurs. Little and Painter (1995) provide a good summary of Escobar's views, with an effective rebuttal. Arce and Long (2000, p. 23) state that the postmodernists (like Escobar) argue 'that development is no more than a discourse justifying control and surveillance of people's practices by powerful institutions.' See also Ferguson (1990) who concludes that development is 'a puzzle,' after teasing out the implications of a project in Lesotho. Hobart (1993) provides a measured critique of development; an edited collection of essays (Rahnema and Bawtree 1997) provides vigorously antidevelopment views. Second in Grillo's classification are the 'monitorists,' those who study the anthropology of development, but do not wish to be actively involved. The last category is the 'reformers' (or 'conditional reformers') who are prepared to be directly concerned with development policies and projects.

The view from the academy has varied over the years. Until the 1970s, development anthropology tended to be marginalized, or ignored, at major universities. This attitude arose partly from a general conservative view that the proper business of academics was 'pure' research, and that applied work was somehow tainted. But all development anthropology contains some theory, even if it not explicit. In addition, some scholars shared Escobar's views that development was a spurious neocolonialist plot, and to be avoided; these views hardened after the exposure of some very murky, clandestine anthropological collaboration by anthropologists with the US government, in Latin America and in South-East Asia in the 1960s.

By the 1970s, changes began in university anthropology departments, in part as a response to the failures of conventional development practices, and with the realization that anthropologists indeed had contributions to make. There was also a more general awareness that anthropology had an important part to play in dealing with current contemporary problems, including poverty, human rights, landlessness, environmental degradation, and, particularly, inequality of incomes, influence, and power. By now, most university departments have at least one, often several, faculty members who specialize in development, offer courses on the subject, and are engaged in actual development projects. A subsidiary reason for this change may be that development offers a good chance of employment (either at a university or with a development agency)—an important factor in days of a scarcity of academic appointments.

4. *The Role(s) of Anthropology in Development*

Horowitz (1999, p. 14), who has written extensively on this topic, points out the major contributions that anthropologists have made to our understanding of involuntary resettlement, the informal economy, internal differentiation of rural societies by wealth, age, or gender and by a host of other important factors of rural communities.

Hoben (1982, p. 370), writing from his experience as an anthropologist and as a senior administrator with USAID, claims that

> development anthropology has unique perspectives on development planning … it focuses on culturally patterned perceptions, goals, interests, strategies and organization of the intended beneficiaries including elites and administrators, whose cooperation is a prerequisite to change.

Elites function at the national level, where they may have enough power to influence development aid in their favor; they also appear, albeit in more modest guise, at the local level, where again they have the influence to shape the direction of development projects, in order to benefit themselves.

Grillo (1985, pp. 6–7) presents a useful classification of nine types of development anthropology, ranging from fundamental research to revolutionary or liberation anthropology, and shows the possibilities and problems involved when anthropologists become involved with advocacy, or human rights issues, or when they become brokers.

5. *Donors, Hosts, and NGOs*

Both the multilateral agencies—The World Bank, The United Nations Food and Agriculture Organization (FAO), the United Nations Development Fund (UNDP), The Consultative Group on International Agricultural Research (CGIAR)—and the bilateral ones—USAID, the British Department for International Development (DfID), the German GTZ and the significant Scandinavian organizations—are now employing anthropologists in increasing numbers. The World Bank, for example, employed 64 anthropologists in 1995 (Horowitz 1999, p. 10). Michael Cernea, until recently Senior Sociologist at the World Bank, deserves much of the credit for the hiring of anthropologists (Cernea 1991). In addition to the in-house appointments, many anthropologists are hired on short-term contracts as consultants. However, the numbers are not a true reflection of their influence: most anthropologists are in relatively junior positions; many function more as administrators than as anthropologists, although they may be able to use their anthropological backgrounds to influence decisions; and they often in effect become captured by the culture of the institution. Nevertheless, some anthropologists, in some situations, have been influential in changing the emphases of development programs and policies.

The 'host country governments' are generally unenthusiastic about the involvement of anthropologists

in development, preferring rather to see the funds flow, without any hindrance caused by social or environmental concerns. Professional development 'experts,' both in donor organizations and in host government service, tend to see anthropologists as tiresome and romantic figures, who delay development by their ill-considered wish to preserve a local conservative culture. In addition, anthropologists are accused (with some justification) of not understanding the mechanisms and culture of the institution, of having no regard for time, of being unable to write 'useful' reports, and of being poor team players. Having spelled out some main stereotypes held about anthropologists, it should be said that there are many honorable exceptions—professionals from other disciplines who fully understand the potential contribution of anthropology to development, and who strive to promote it. Administrators also tend to distrust the poor, the supposed 'beneficiaries,' who are often seen as superstitious peasants and the main obstacles to change. There is seldom an appreciation of indigenous knowledge, and of the finely tuned strategies employed by local people, simply in order to survive.

NGOs play an increasingly important part in international development, sometimes acting on their own, often acting on behalf of the major donors. Starting as charity or relief agencies, they have rapidly become effective professional development agencies. As part of this change, the NGOs have made much use of anthropologists, many of whom are employed full-time, with others being engaged as consultants for specific projects. NGOs have at times become involved in advocacy, creating a delicate situation in terms of their legal charitable status. At present, there is some tension between the large 'Northern' NGOs, which have until recently dominated the field, and the increasing number of 'Southern' NGOs, which are challenging the older agencies. The latter seldom use anthropologists, both for lack of funds, and, more particularly, because the leaders claim that as local people themselves, they know their own culture, and have no need of outside advice. This may be questionable, when such officials are often separated by virtue of their education and background from the poor people who are the beneficiaries. However, some indigenous groups, such as the Cree Regional Authority of Canada, now routinely hire anthropologists to work with them on land claims and other concerns.

6. *Domains*

Development anthropology has been concerned with a wide range of topics, especially related to management of natural resources—agriculture, livestock and pastoralism, fisheries, forestry, wildlife, irrigation; settlement and migration; rural-urban linkages and markets; ethnicity and the resolution of social conflict; disaster prevention and relief; health and AIDS. In all these areas, the main emphasis has been on the alleviation of poverty, and on making development more effective, more relevant to people's needs, and more sustainable.

Ethical dimensions are always present when a development anthropologist decides which project(s) s/he should be involved in. Resettlement projects, for example, usually cause tremendous damage to the people resettled, including the 'grieving for a lost home' syndrome. But if an anthropologist is involved, their sufferings may be reduced, their concerns publicly aired. Is this adequate justification for taking part?

7. *Participation*

One of the main trends in development is decentralization, whether in the form of democratization, decentralizing or devolution of services and authority, comanagement arrangements, etc. This all increases the scope for local involvement, yet substantial barriers often exist due to gender, class, caste, race, or other variables. One of the enduring challenges for anthropologists is to widen the scope for participation. Coinciding with increased participation there is often increased conflict, with a growing involvement by anthropologists in conflict management.

Much attention has been paid to the need for local people to participate in development projects that affect them, which ideally means that they should be consulted about the design, implementation, and evaluation of projects. Formidable difficulties present themselves, apart from resistance from administrators, many of whom regard participation as a romantic fallacy. Communities are seldom homogeneous, so who is truly representative?

Pottier (1993) presents a collection of papers that 'illustrate a variety of organizational responses to the discourse of participative development.' Contributors favor a 'client-focused, participative philosophy' rather than a top-down western model. They emphasize the managerial capacity of the vulnerable groups, who lack access to and control over resources. Pottier stresses the need for flexibility and sensitivity; 'the social worlds within which development efforts take shape are essentially fluid.' And what is needed is an ethnography of the total project, not just of one group (Pottier 1993, pp. 7, 31).

8. *Training*

In addition to the standard academic training, anthropologists who wish to work in development need an awareness of a wide range of factors, most of which

will best be gained with hands-on experience. Typically, they will need to have, first, a sound knowledge of the region where the development is proposed. Topical/general expertise must be grounded in detailed knowledge of one people and their area. Inevitably, development anthropologists work in many regions, making up in their broad comparative knowledge what they lack in detailed local information. This has led to their being accused, by more conservative antidevelopment anthropologists, of not being qualified to advise on projects outside their own specialized geographical areas.

Second, they will need an appreciation of the major technical aspects, so that at the least they can communicate with the technical specialists.

Training is also desirable for aid administrators, many agencies having organized successful workshops to heighten sensitivity to basic social issues. Government employees also benefit from such training; for example, FAO has arranged a series of workshops to demonstrate to forestry officials why they should take into account the culture and knowledge of local people, before planning any forestry projects. This means dealing with key issues of poverty, gender, land rights, and tree species preferences.

9. *Accomplishments*

What has development anthropology achieved? Although there is a long road ahead, significant contributions have already been made since the 1970s. Regarding development, anthropologists have succeeded, at least partially, in persuading those who direct development at the major agencies that projects will be more likely to attain their goals if they are planned, implemented, monitored, and evaluated with attention to—and, if possible, with the participation of—those local people who are intended to be the beneficiaries. There is a greater awareness of the strengths of local people, of their knowledge, their adaptive capacities, and the diversity and richness of their cultures. More attention is now paid to poverty and inequality, and to the interplay between local level and global factors. Both development planners and anthropologists now have a clearer and more sophisticated idea of how power relations affect people at all levels. Planners have been encouraged to challenge basic assumptions, and to look at alternative approaches. Institutions like the donor agencies are conservative, so the challenges and alternative approaches must be handled with delicacy, so as not to upset the guardians of tradition in the agencies. Anthropology is by now an integral part of development, both in theory and in practice.

Anthropology itself has also been changed, with a widening of the horizons to include not only particular local communities, but looking also at the macro aspects, seeing how global issues reach out and affect the furthest corners of the modern world. More emphasis is placed on modernity, nationalism, ethnicity, power, and, essentially, on class relations.

10. *Future Directions*

It is difficult to predict what lies ahead for development anthropology in the twenty-first century, or for anthropology itself. Periodically, the major anthropological associations organize conferences to discuss (usually gloomily) the future of anthropology, but past pessimistic forecasts have seldom been proved correct. Whatever future the discipline of anthropology has, an emphasis on development will certainly be a major component. '… the cutting edge of scientific anthropology today is the anthropology that is being practiced by persons who are passionately committed to human rights, cultural pluralism and environmentally and socially sustainable development. For anthropology to survive as meaningful discipline, it must focus on the very issues that engage those of us who identify our work as *development* anthropology. Both academic and practical social anthropologists must understand the process of incorporation of agrarian societies into the world of political economy' (Horowitz 1999, p. 11).

It is more difficult to predict what future directions development will take. In the 1990s, there has been a noticeable lack of enthusiasm for development aid, with many national contributions falling to dangerously low levels. Africa, in particular, which has been the recipient of a large proportion of all development aid, has been dismissed by *The Economist* as a hopeless case—because of the seemingly endless series of intractable wars and conflicts, natural disasters, weak administrations, and corruption on a large scale.

In order to achieve their potential, development anthropologists need to sharpen their methodology, aim for greater analytical clarity, be more rigorous, and be more theoretically sophisticated—which entails proceeding along the paths already used and expanding the vision. 'What is needed is a more systematic way to bring these experiences into an empirically informed version of critical theory that advances our understanding of the human species and its diverse wants and needs' (Horowitz 1999).

Arce and Long (2000, pp. 24, 27) advocate an actor-oriented approach, which emphasizes 'the intricate interplay and joint appropriation and transformation of different bodies of knowledge … the critical task of development anthropology is to develop methodologies and theoretical interpretations of different knowledge interfaces inherent in intervention processes and local/global change.' Researchers should adopt a reflexive approach, to include their own encounters with other actors in the development arena.

Finally, Scudder (1999)—probably the most effective contemporary development anthropologist, as far as influencing policy is concerned—draws from his own views and those of 53 other anthropologists to ask the question: 'The emerging global crisis and development anthropology: Can we have an impact?' Scudder concludes by making 'five proactive suggestions as how we can put our knowledge and discipline to work':

(a) identify the key problems that we have the expertise and responsibility to address—for example, among Scudder's respondents, the most important topics were poverty, globalization, and increasing marginalization;

(b) better integrate development into the curriculum of anthropology departments;

(c) 'bring our expertise to the attention of the public and policy makers'; this should include emphasizing the limits as well as the potential of development anthropology;

(d) 'include more generic research that addresses major global problems and the linkages between them';

(e) cooperate with ecologists in public education, and in influencing local, national and international policy makers. Scudder hopes that 'anthropologists—with their peerless awareness of cultural diversity in time and space, and their understanding of how symbols can influence behavior—(can) work with ecologists to translate what we know into a vision for the future.'

Most writers on development anthropology agree that this is an exciting juncture in the world of development, and that despite the formidable problems posed by the nature of development, anthropologists do have much to offer in this field.

See also: Development, Economics of; Development: Rural Development Strategies; Development: Social; Development: Socioeconomic Aspects; Ethnic Cleansing, History of; Financial Institutions in Economic Development; Globalization, Anthropology of

Bibliography

Arce A, Long N 2000 Reconfiguring modernity and development from an anthropological Perspective. In: Arce A, Long N (eds.) *Anthropology, Development and Modernities: Exploring Discourses, Counter-tendencies and Violence*. Routledge, London

Cernea M M (ed.) 1991 *Putting People First: Sociological Variables in Rural Development*. (2nd edition). Oxford University Press, New York

Cochrane G 1971 *Development Anthropology*. Pergamon, New York

Erasmus C J 1954 An anthropologist views technical assistance. *The Scientific Monthly* **78**: 147–158

Escobar A 1991 Anthropology and the development encounter: The making and marketing of development anthropology. *American Ethnologist* **18**: 658–81

Ferguson J 1990 *The Anti-Politics Machine: Development, Depolitization and Bureaucratic Power in Lesotho*. Columbia University Press, New York

Gardner K, Lewis D 1996 *Anthropology, Development and the Post-Modern Challenge*. Pluto Press, London

Grillo R 1985 Applied anthropology in the 1980s: Retrospect and prospect. In: Grillo R, Rew A (eds.) *Social Anthropology and Development Policy*. Tavistock, London, pp. 1–36

Grillo R, Stirrat R L (eds.) 1997 *Discourses of Development*. Berg, Oxford, UK

Gulliver P H 1985 An applied anthropology in East Africa during the colonial era. In: Grillo R, Rew A (eds.) *Social Anthropology and Development Policy*. Tavistock, London, pp. 37–57

Hobart M (ed.) 1993 *An Anthropological Critique of Development: the Growth of Ignorance*. Routledge, London

Hoben A 1982 Anthropologists and development. *Annual Review of Anthropology* **11**: 349–75

Horowitz M M 1996 Thoughts on development anthropology after twenty years. In: Moran E F (ed.) *Transforming Societies, Transforming Anthropology*. University of Michigan Press, Ann Arbor, MI

Horowitz M M 1999 The sustainability of anthropology and development. *Development Anthropologist* **17**: 10–24

Little P D, Painter M 1995 Discourse, politics and the development process: Reflections on Escobar's 'Anthropology and the Development Encounter'. *American Ethnologist* **22**: 602–9

Long N 2000 Exploring local/global transformations: A view from anthropology. In: Arce A, Long N (eds.) *Anthropology, Development and Modernities*. Routledge, London

Pottier J (ed.) 1993 *Practising Development*. Routledge, London

Rahnema M, Bawtree V (eds.) 1997 *The Post-Development Reader*. Zed Books, London

Rew A 1997 The donor's discourse: Official social development knowledge in the 1980s. In: Grillo R, Stirrat R L (eds.) *Discourses on Development*. Berg, Oxford, UK, pp. 81–106

Scudder T 1999 The emerging global crisis and development anthropology: Can we have an impact? (1999 Malinowski Award Lecture). *Human Organization* **58**: 351–64

D. Brokensha

Development: Socioeconomic Aspects

Introduction and Overview

Although the concept of development first entered the vocabulary of social scientists and policy makers in the 1950s, as we enter the twenty-first century roughly half of the planet's six billion inhabitants remain mired in poverty, lacking access to resources that readers of this encyclopedia would deem basic to human well-being. At the same time that the gap separating rich and poor countries is widening, inequalities are growing within many societies, in North and South alike. Under the circumstances, it is no surprise that economic and social development occupy the attention of social scientists from a number of disciplines, from

anthropology to economics, political science to sociology. Yet despite the intractability of poverty and exclusion—the very circumstances that development is meant to overcome—there remains no consensus concerning precisely what development is or how it should be analyzed. Nor is there agreement about how it might best be achieved in different settings during the twenty-first century, or even whether the concept of development affords a useful lens through which to consider improvements in people's lives. Indeed, at the close of the twentieth century some scholars dismissed 'development' as little more than a discredited remnant of Cold War era policies designed to induce impoverished peoples of the 'Third World' to adopt ways that were often at odds with their interests and with their cultures (Escobar 1995).

Even among social scientists who retain a commitment to development, the term has taken on a variety of meanings, and has motivated inquiries employing vastly different theoretical frameworks. While 'development' may be the concern of macroeconomists modeling the impact of trade liberalization on national growth rates, it may be equally central to ethnographers seeking to reveal linkages between the protection of cultural diversity and the survival of agricultural practices deemed crucial to maintaining biodiversity or to ensuring local production of critical foodstuffs. Despite the semantic overlap, and the likelihood that macroeconomic policies impinge upon the decisions of farmers to plant one crop or another—or to migrate to overcrowded cities that lack the infrastructure to accommodate them—dialogue is at best weak among researchers operating across such broadly different levels of analysis.

Most currents of social scientific research now agree that development refers, ultimately, to improvements in the conditions of life for groups of people experiencing relative deprivation, whether they be located in a particular village, urban metropolis, country, or subnational or transnational region. Efforts to promote development may focus on the entirety of a population, or may be targeted to address needs of specific groups that are deemed especially vulnerable, or whose welfare is perceived as potentially generative of multiplier effects, i.e., whose improved welfare is deemed likely to enhance the life chances of the community as a whole. Thus, peasants, shantytown dwellers, women, or ethnic minorities may merit positions of privilege in considerations of development on grounds that their needs are especially pressing. Similarly, given their potential role in promoting growth or employment, micro-entrepreneurs, managers of competitive firms, or foreign investors may also be central to analyses of development.

Amid continuing disagreement concerning the appropriate scope of analyses of development, there is widespread accord that development concerns trajectories of change, encompassing movement from circumstances of relative hardship and deprivation toward qualitatively improved conditions of life. Those who effect measures to promote development may or may not be the intended beneficiaries of those efforts, but whether development is pursued through interventions by external agents, such as development assistance agencies, or through initiatives of the targeted communities themselves, it is a process that involves social actors. Indeed, to the extent that particular trajectories of economic change are known to correlate strongly with societal characteristics, development increasingly is understood as a set of processes that reflect underlying social dynamics of the territories in which it takes place.

Development implies increasing the volume of available resources, and thus is associated powerfully with economic growth. For many years, widespread scholarly acceptance of the so-called Kuznets curve (Kuznets 1955), according to which inequality expands sharply during initial phases of economic growth and declines steadily upon achievement of substantial industrialization, motivated many development economists to disregard or tacitly advocate economic inequality. More recent research has established a far more complex relationship between growth and distribution, however, suggesting in some instances that a relatively equal distribution of assets is conducive not only to rapid economic growth but also to patterns of growth that reinforce equality (Stallings et al. 2000). Partly in keeping with these findings, and as a result of ethical and pragmatic concern about the plight of impoverished people, at the beginning of the twenty-first century development typically refers not only to economic growth but also to the distribution of income among different segments of the population and, more broadly, to enhancements in quality of life. Thus, alongside the achievement of higher incomes, development is intended to reduce poverty, improve health and nutrition, prolong life expectancy and afford opportunities for education. By some accounts, and for a growing scholarly literature, development encompasses political objectives, such as expanded individual freedoms and greater civic participation, and environmental ones, such as community access to natural resources and protections against ecological degradation.

Of course, these economic, political, and societal goals are not obviously compatible, and as Gereffi and Fonda have noted, '[T]he multi-dimensional nature of development, and the fact that these desired objectives do not necessarily occur together, help to account for some of the … tradeoffs in … development outcomes' (1992). Arguably, this multidimensionality lies at the root of divisions that often limit productive debate about development, but that in some instances, by contrast, have made studies of development especially fertile ground for cross-disciplinary innovation.

The remainder of this article sketches further the evolution of social scientific thinking about development. After tracing movement away from the

conviction that 'all good things would go together' in tandem with economic progress, the article highlights currents of research that illuminate connections between characteristics of social processes and the nature of development outcomes. It concludes by noting promising trends and future challenges for research on development.

1. *From Economy-centered Analyses to Plural Conceptions of Development*

Early twenty-first century preoccupation with the social aspects of development stands in stark contrast to perspectives held widely during the initial decades following World War II. Espousing an underlying idea of development that can be traced as far back in Western thought as Hegel and Marx, who shared a conviction in the inexorable progress of human history toward ever greater levels of perfection, modernization theorists believed that the advanced industrial countries of Western Europe and North America represented the logical end-point toward which the former colonies of Africa, Asia, and Latin America would advance inevitably (Leys 1996, pp. 4–5). Early thinking about development thus posited that properly balanced macroeconomic policies and functioning state institutions would facilitate a relatively seamless evolution of 'backward' nations toward the prosperous and peaceful conditions achieved by economically developed countries.

While the almost universally disappointing experience of postcolonial economies served to undermine the model long ago, confidence in such unilinear progress has disappeared even in the prosperous countries of the North, where deindustrialization and increasing social polarization have eroded the sociopolitical consensus that characterized the period of postwar growth. And despite the often dramatic expansion of industry and rapid urbanization in parts of the former Third World, the economic and political inequalities of the past have not only been slow to disappear but often have actually grown worse in conjunction with these very processes of modernization. These trends have been exacerbated further by growing discrepancies between supply and demand of natural resources needed to assure communities of the ability to satisfy their basic needs.

Dependency theory, the principal alternative to the modernization school as a paradigm for analyzing development, has hardly been proven more prescient. To be sure, the dependency critics were not unjustified in contending that deeply embedded structures of discrimination reinforced the subordinate position of peripheral economies vis-à-vis the so-called core countries. However, they erred in their identification of the particular mechanisms—such as the supposedly deteriorating terms of trade for primary product exporters, emphasized by Raul Prebisch and his followers—through which this process was sustained. Moreover, as evidenced by the significant degrees of industrialization that were achieved in parts of Latin America and Asia, they failed to grasp the degree to which developing economies could advance despite the disadvantageous conditions they confronted (Evans 1979). In particular, dependency analysis was called into question by the considerable success of export-oriented growth in several East Asian countries, which contrasted markedly with the failure of development strategies predicated on isolation from the very global market that *dependentistas* held to be the primary impediment to development in the periphery (Haggard 1990).

Indeed, variations in the development experiences of different parts of the North and South testify to the degree to which the impacts of the international context on particular communities, nations, and regions are contingent on factors that transcend inherited conditions or positions in the world system. Understanding of the strikingly diverse development experiences of apparently similar countries and regions since the 1980s requires analysis of the microfoundations of distinctive trajectories. Structural inequities have by no means disappeared from the scene, but they have taken on new forms, and reflect complexities that are not clarified sufficiently by the sweeping categories of traditional theories of dependency or of the world system.

Moreover, for dependency theorists, as for modernization analyses, development was understood at the level of states and was analyzed through macro-level policies and institutions. The particular societal configurations of developing countries were of little interest to these analyses of developmental impasse, nor were the strategies of social forces or their fate as a result of distinct development trajectories a major concern. As Leys (1996) has pointed out progress, or more frequently stagnation, was deemed a function of underlying structures of interaction operating above and beyond the micro- and meso-level processes that would be of concern to subsequent generations of social scientists concerned with development.

Confidence in the grand theoretical frameworks of modernization and dependency had already weakened by the mid-1970s, when the impact of oil shocks, declining faith in the capacity of economic growth to reach the majority of the world's poor, and debates about ecological limits to growth combined to shift the emphasis in development circles away from large-scale, state-oriented modernization efforts toward localized strategies for meeting the needs of the world's poor (Sutton 1989). The World Bank's 1973 call for targeting development efforts to the 'poorest of the poor' catalyzed a shift in focus away from growth and toward issues of redistribution and equity. Perhaps resigned to the likelihood that significant portions of the globe would remain trapped in low-growth situ-

ations, the concern of policy makers turned to enhancing the well-being of communities that were destined to miss out indefinitely on the postwar vision of modernity and prosperity. This in turn led to increasing emphasis on community participation in development. Whether labeled 'farmer-first,' 'bottom-up,' or 'grassroots' development strategies, the emphasis here was on decentralization, localization, and the satisfaction of basic human needs for food, shelter, health, and security.

Absent expectations of significant increases in wealth, provision of adequate nutrition, sanitation, housing, and health care, and in some instances assurances of minimal incomes and educational opportunities came to be perceived as central objectives of development policies. While such measures as the rate of increase in per capita income retained their traditional appeal for researchers and practitioners, alternative indicators of development gained ever-broader acceptance. The physical quality of life index (PQLI), for example, used by a variety of international agencies to inform their development efforts, sought to measure the overall well-being of particular communities and countries. Incorporating indicators of health, nutrition, and literacy as well as income levels, such measures were acknowledged increasingly as more accurate yardsticks of development than were the more traditional macro-level indicators that focused on economic growth, industrialization, or urbanization.

While the state was sometimes viewed as a potential provider of goods and services required for poverty alleviation, international development agencies increasingly bypassed developing country governments, choosing to rely instead on the mediation of international and local nongovernmental agencies. This trend gained steam during the 1980s, a decade marked by the international debt crisis in Latin America and the implosion of postcolonial states in parts of Africa and elsewhere. Influenced by studies that emphasized the economic distortions effected by rent-seeking elites in command of predatory states (Bates 1981), conventional wisdom increasingly abandoned the notion that central states had the capacity or inclination to ameliorate the hardships faced by most of their populations. Rather, faith was placed in the operation of more-or-less unfettered markets, and in the benefits that market-driven growth could bestow upon societies whose potential for growth had been stymied by inefficient state institutions (Williamson 1993). Accordingly, multilateral institutions and Western governments advocated structural adjustment programs and sharp reductions in the economic role of the state in development countries.

The so-called neo-liberal trend coincided, albeit somewhat awkwardly, with an equally anti-statist notion that development entailed the empowerment of communities and individuals themselves to undertake the development 'project.' The emphasis on local autonomy and grassroots participation provided a useful corrective to top-down development strategies of the past, and engendered significant improvements in efficiency and equity in many settings (Tendler 1997). Yet in most of the periphery, declines in state spending on health, education, and social welfare took a severe toll on living standards, particularly among the poor, as local institutions were given responsibility for tasks without enjoying corresponding increase in their access to resources, or as supply of these public goods was simply transferred from the public to the private sector.

Nor were criticisms of the laissez-faire approach limited to its impact on equity, since a worsening of extreme poverty has been demonstrated to have adverse ecological consequences as well (Martinez Alier 1991). Left solely to the vagaries of the market, limits on pollution, deforestation, soil erosion, and other environmental hazards cannot be ensured. Such considerations emerge as integral to analyses of development that emphasize linkages between the welfare of communities and their access to natural resources. By the end of the 1980s, understanding of these linkages as well as growing scientific and public concern over processes of global environmental change (such as ozone depletion) combined to foster greater emphasis on development strategies that could accommodate the needs of the present generation without jeopardizing the availability of resources for future generations. Much of this work concludes that a high degree of equity and local control over resources are necessary, if not sufficient, preconditions for environmentally sustainable development. In some instances, the argument coincides with an emphasis on empowerment of subordinate groups, ranging from indigenous peoples to women to subsistence farmers, who are portrayed as threatened by a singular focus on economic growth, which is in turn characteristic of the visions of modernity and progress that have provided the justification for much of what we have come to think of as development. It should come as no surprise, then, that proponents of this view often question the desirability of retaining development as a societal objective, and in any event call for 'an alternative modernity,' a modernity fundamentally skeptical of economic growth (e.g., Martinez Alier 1991).

2. *Emerging Trends and Future Challenges*

The need to comprehend the development implications of processes of globalization occasioned a revitalization of development thinking toward the end of the twentieth century. By virtually all accounts, the unprecedented mobility of goods, knowledge, and peoples across national borders, and increasingly transnational flows of capital and information have made the traditional boundaries established by states far more permeable. While some regions, and some

social sectors, appear well situated to benefit from the new circumstances, others face growing hardship, and the boundaries between prosperity and decline often do not coincide with the borders of countries. Research on the changing structure of industries illuminates both the increasingly global scale of economic activity (e.g., Gereffi and Korzeniewicz 1994) and the significance of community dynamics for encouraging cooperative ties conducive to successful adaptation to rapidly changing markets (Piore and Sabel 1984). This literature calls into question the traditional focus on national policies to promote economic growth or to alter prevailing patterns of distribution. Understanding the ways in which different states respond to similar pressures differently (and with varying degrees of success), remains important (Evans 1995), yet state-level analyses of the political economy of development must be supplemented by attention to the ways in which relationships among local and nonstate actors can evolve in such a way as to encourage successful economic adaptation (Doner and Hershberg 1999).

Given strikingly different trajectories of change among apparently similar communities, social scientists have devoted increasing attention to the role of social capital and social institutions in determining prospects of different groups to prosper. Here, the focus has been on the relational dynamics that characterize specific socio-territorial spaces, and that nurture levels of trust without which economic actors cannot work together to insulate themselves from market risks (e.g., Perez Sainz 1996). While a concern with social capital and institutions is shaping a good deal of innovative work in sociology, anthropology, and political science, economists are also advancing influential perspectives on the societal underpinnings of growth. Endogenous growth theorists highlight the mutual connections between rapid economic growth and relatively equal distribution of resources, including such assets as health and education. In this manner, by the end of the twentieth century, equity was increasingly accepted by economists as a foundation of economic growth, and as such has become part of even the most mainstream of development agendas. Similarly, informed by the work of economic historians, researchers in both political science and economics are developing fresh approaches to the study of economic institutions and their relationship to market processes (Williamson 1985).

To the extent that development consists of more than economic growth, and also encompasses movement toward the creation of greater opportunities for political participation, fuller representation of societal interests, and the protection of individual and collective rights, it is crucial to understand how political and economic transformations alter the linkages between state and society in different parts of the world. Whether social actors organize themselves on the basis of class, community, ethnicity, or other principles is of fundamental importance for any comprehensive treatment of development, just as it affects the likelihood for sustaining particular political coalitions, some of which, in turn, appear more likely than others to be able to sustain different strategies of development (e.g., Nelson 1994, Haggard and Kaufman 1995).

The importance of community in considerations of development is underscored by the growing consensus among scholars and practitioners that successful development—in the North as well as in the South—hinges on the strengthening of local institutions, be they community development banks, schools, or health clinics (Ake 1991, Tendler 1997). The importance of neighborhood characteristics—whether for the sociologist studying the experience of growing up in impoverished urban slums, or for the anthropologist concerned with how kinship networks help attenuate the effects of agricultural shortfalls—underscores the degree to which development-related research is concerned with micro-level phenomena. Whether in connection to these sorts of issues or to questions of industrial development and employment (Stallings and Peres 2000), the need to link micro-level processes to the 'macro' level transformations taking place in the global political economy remains an ongoing challenge for students of development in the next decades.

As these comments suggest, many of the topics attracting creative scholarship will continue to reflect challenges that, in hindsight, have been relevant throughout the half century during which development has been the subject of social science inquiry. Yet alongside promising evidence of potential complementarities between economic growth and other indices of well-being, researchers are uncovering disturbing evidence of new forms of immiserizing growth (Kaplinsky 2001). No less daunting are the new forms of exclusion that are emerging, as uneven access to information technologies threatens to reinforce the disadvantaged status of vast segments of the world's population. Perhaps most troubling of all are the development implications for a number of the world's poorest countries, most notably in Sub-Saharan Africa, of the devastation wrought by the AIDS epidemic, the human costs of which are magnified by its likely impact on economic growth for generations to come. It is difficult to conceive of a more striking illustration of the degree to which the economic and social aspects of development are inextricably linked to one another.

See also: Dependency Theory; Development and the State; Development and Urbanization; Development, Economics of; Development: Social; Development: Social-anthropological Aspects; Development Theory in Geography; Economic Growth: Theory; Financial Institutions in Economic Development; Modernization and Modernity in History; Modernization, Political: Development of the Concept; Modernization, Sociological Theories of

Bibliography

Ake C 1991 Rethinking African democracy. *Journal of Democracy* **2**(1): 32–44

Bates R 1981 *Markets and States in Tropical Africa*. University of California Press, Berkeley, CA

Doner R F, Hershberg E 1999 Flexible production and political decentralization in the developing world: Elective affinities in pursuit of competitiveness? *Studies in Comparative International Development* **33**(4)45–83

Escobar A 1995 *Encountering Development: The Making and Unmaking of the Third World*. Princeton University Press, Princeton, NJ

Evans P 1979 *Dependent Development: The Alliance of Multinationals, the State and Local Capital in Brazil*. Princeton University Press, Princeton, NJ

Evans P 1995 *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press, Princeton, NJ

Gereffi G, Fonda S 1992 Regional paths of development. *Annual Review of Sociology* **18**: 419–48

Gereffi G, Korzeniewicz M 1994 *Commodity Chains and Global Capitalism*. Praeger, Westport, CT

Haggard S 1990 *Pathways from the Periphery*. Cornell University Press, Ithaca, NY

Haggard S, Kaufman R 1995 *The Political Economy of Democratic Transition*. Princeton University Press, Princeton, NJ

Kaplinsky R 2001 Gaining from global value chains: The search for the Nth rent. In: Gereff G (ed.) *Who gets Ahead in the Global Economy? Industrial Upgrading and Development in the 21st Century*. (forthcoming)

Kuznets S 1955 Economic growth and income inequality. *American Economic Review* **45**(1): 1–28

Leys C 1996 *The Rise and Fall of Development Theory*. Indiana University Press, Bloomington, IN

Martinez Alier J 1991 Environmentalism and the poor. *Journal of Latin American Studies* **23**(3): 621–39

Nelson J 1994 How market reforms and democratic consolidation affect each other. In: Nelson J (ed.) *Intricate Links: Democratization and Market Reforms in Latin America and Eastern Europe*. Transaction Publishers, New Brunswick, NJ

Perez Sainz J P 1996 *Sarchi: Capital social y artesania*. FLACSO-Costa Rica, San José, Costa Rica

Piore M, Sabel C 1984 *The Second Industrial Divide: Possibilities for Prosperity*. Basic Books, New York

Przeworski A, Alvarez M E, Cheibub 2000 *Democracy and Development. Political Institutions and Well-Being in the World, 1950–1990*. Cambridge University Press, Cambridge, UK

Stallings B, Birdsall N, Clugage J 2000 Growth and inequality: Do regional patterns redeem Kuznets? In: Solimano A (ed.) *Distributive Justice and Economic Development: The Experience of Chile and Developing Countries*. University of Michigan Press, Ann Arbor, MI

Stallings B, Peres W 2000 *Growth, Employment and Equity: The Impact of the Economic Reforms in Latin America and the Caribbean*. The Brookings Institution, Washington, DC

Sutton F X 1989 Development ideology: Its emergence and decline. *Daedalus* **118**(1): 35–58

Tendler J 1997 *Good Government in the Tropics*. Johns Hopkins University Press, Baltimore

Williamson J 1993 Democracy and the Washington consensus. *World Development* **21**(8): 1329–36

Williamson O E 1985 *Markets, Hierarchies and Relational Contracting*. Free Press, New York

Williamson O E 1994 Transaction cost economics and organization theory. In: Smelser N, Swedberg R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ

E. Hershberg

Development: Sustainable Agriculture

Sustainable agriculture presents a challenge to students of development. A respect for sustainability requires us to confront the historical, single-minded commitment to increasing output through the extension and intensification of production. The unprecedented support for research that led to the green revolution contributed to a new model of agriculture that was incorporated into a transnationalized agro-industrial system based on specialized production processes. This coincided with the conviction of many members of the scientific and policymaking establishments that the major problem facing the world in the second half of the twentieth century would be meeting the Malthusian challenge by closing the gap between the geometric advance of population and the slower advance of food output.

Since these changes began to be introduced in the 1950s and 1960s, innumerable studies documented the serious disruptive effects that the new cropping systems have had on traditional rural societies and ecosystems around the world. As evidence accumulated about the problems of declining water tables, soil contamination, salinization, social disruption, and growing spatial, economic, and social disparities, serious questions were raised about the environmental sustainability of these new systems. Even more, a series of allied disciplines attempting to integrate indigenous knowledge with scientific advances in our understanding of the ability of local communities to implement effective programs of ecosystem management have demonstrated the potential for increasing output without occasioning the environmental and social devastation observed in commercial monocropping systems.

At the end of the twentieth century, the paradigm conflicts between these knowledge systems grew even stronger as the revolution in biotechnology promised yet another wave of innovations that are being challenged even more vigorously than in the past. Indeed, the risks of mistakes and tragic accidents are greater as the new modifications involve the introduction of genes from one species into another in order to confer a degree of protection against plagues, to develop resistance to natural phenomena, or to increase productivity. The critics argue that these trans-genetic products open the world to an unknown series of challenges that may threaten the very stability

of existing biological or social systems; the damage wreaked by bovine spongiform encelopathy (BSE or mad-cow disease) in Europe is cited as palpable evidence of such potential.

In view of a broadening commitment to the principles of sustainability agreed upon at the UN Conference on Environment and Development in Rio de Janeiro in 1992, attention is turning to the importance of introducing appropriate production models for agriculture. New approaches call for agriculture to be less disruptive of ecosystems and less dependent on nonrenewable sources of energy to maintain soil fertility, control plagues, and power machinery. The changes in agriculture are proceeding on a number of different fronts simultaneously. In the fields, greater attention is turning to techniques that require less disruption of the land itself, by introducing new soil and water management techniques that take advantage of 'zero tillage' so that fertility and moisture are more jealously guarded. Crop rotations and diversified planting programs are reducing the vulnerability of farming regions to massive infestations that are still a serious concern to commercial farmers. On the market side, a more sensitive consumers' movement has dramatically changed the prospects for organic products, which can be certified as being free of inputs in the production process that might pose a threat to either the purchaser or the environment in which they are cultivated. As this movement has grown, people have also become aware of the importance of insuring that peasant producers and field workers are also correctly paid so that their own poverty does not become a source of degradation.

Sustainable agriculture, however, will not be achieved simply by developing new cultivation techniques or introducing more benign inputs to use in highly specialized production systems. From decades of failure of modernization programs that did not deliver their promised welfare packages, we discovered that many of the uprooted rural peoples are capable of providing for their own needs and those of their regional environments, while making valuable contributions to the needs of the larger society of which they are a part. The invaluable research of innumerable ethnobotanists and agroecologists around the world has demonstrated the productive potential that could be unleashed if these communities were allowed to function freely.

Sustainable agriculture cannot succeed if a broader development framework that encourages the initiatives of rural peoples does not accompany it. At present, peasants and indigenous groups are forced into marginality. They lack the resources to function well in the marketplace and the support to realize how important their communal knowledge and experience is for producing crops and managing their environments and their societies. The complex relations between earthly resources, the changing seasons, and cyclical changes such as El Niño are not new problems, but they were generally integrated into a social and political framework that facilitated their management. Today these traditional productive systems and the people are frequently subjected to institutional controls and bureaucratic supervision that impede their implementation. Instead, small-scale specialized production processes force them to produce single-mindedly for the marketplace rather than take advantage of the varieties of processes and products that the natural world has provided.

If development is to proceed, rural communities must be allowed to overcome their marginality and their poverty. This does not require their full integration into a globalized society. Rather it means that they must be freed from the fetters that restrict them from realizing the full potential that their own sociopolitical systems and ecosystems allow. The drive for sustainability involves a three-pronged program for local advance: the development of autonomy and greater self-governance, the pursuit of self-sufficiency in food production as well as in other areas of life, and the development of new productive activities and products that will contribute to their material well-being, incorporating new community members into activities that will generate new sources of income.

Sustainability is not possible as long as the profound inequalities that characterize most societies in developing countries persist. Social and economic polarization are features of societies in decomposition. Institutions that perpetuate inequality contribute the kind of environmental destruction stimulated by unfettered consumption. These same institutions also perpetuate and rationalize the existence of poverty, and force the poor to contribute to environmental degradation. The growing gap also incubates violence while limiting the possibility of conflict resolution.

Only by encouraging the peasants and indigenous groups to create their own mechanisms to overcome poverty will it be possible to move in the direction of sustainability. They must be allowed to assume responsibility for their own social organization and their local institutions must be legitimized. In the current world order, federal governments cannot continue to try to impose centralized forms of administration and governance on all segments of society. In agriculture, the essence of sustainability is the implementation of a wide variety of production systems, carefully adapted to the demands of the environments, the needs of the communities, and the possibilities of the marketplace. National agricultural policies have generally proved too cumbersome to handle the diversity of situations in which rural producers find themselves.

Self-governance, however, is not a sufficient condition for sustainability. Communities also require some minimum guarantee of well being, of products and services that will offer them some security. This is part of the reason why peoples have so steadfastly adhered to the principles of self-sufficiency throughout history; so powerful is this argument that John

Maynard Keynes saw fit to argue in 1933 that 'national self-sufficiency' must be an overarching facet of economic policy, if a country is to be able to have a modicum of control over its fate. Certainly the same principle applies even more so at the community level, where the vulnerability is even greater and 'natural' disasters often affect the poorest groups most harshly.

But providing for one's basic needs through mechanisms of local autonomy and traditional social institutions cannot offer a means to improve material welfare and confront the challenges of the market economy. Productive diversification is a policy thrust that is fundamental for generating the additional resources required for a community to pull itself from marginality. The guidelines for this process must be solidly grounded in a respect for the ecosystems and the primacy of local provision of basic needs. Many a community has faltered or even been dismembered by well-meaning programs of export-promotion that have destroyed local buffers, be they social, political, or productive. In the world of the twenty-first century, where there is a growing concern for the quality of life, many new products can be derived from conservation and ecosystem enhancement activities, such as the sequestration of greenhouse gases or increasing the rate of recharge of important aquifers. Some suggest that ecotourism offers an ideal match between the consumption drives of the urban population and the needs of rural communities. Certainly the move to expand fair trade channels and organic production is ideally suited to the conditions of rural communities seeking to promote development by implement programs of sustainability.

Sustainable agriculture is now recognized as an important underpinning for any development program. It is clear that poverty and environmental destruction rather than the Malthusian specter threaten humanity at the beginning of the twenty-first century. Implementing a program of locally managed rural production that responds to the dual challenges of self-sufficiency and diversification will contribute in important ways to reducing poverty and improving the distribution of income. A key feature of the three principles enumerated here is that they also remove the onus from rural people of depending exclusively on agriculture for their livelihoods. Although it is crucial that cropping patterns and cultivation techniques become as attuned as possible to the demands of local ecosystems, it is also important to realize the importance of actively encouraging rural communities to initiate a broad range of complementary activities that will spread the burden of supporting the community over a broader base of resources and skills, thereby reducing the pressures on the land and on traditional crops that are no longer able to support the nutritional and income needs of the population.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Development, Economics of; Development: Rural Development Strategies; Development: Social; Environmental Planning; Farmland Preservation; Food Production, Origins of; Food Security; Green Revolution; Greening of Technology and Ecotechnology; Rural Geography; Rural Planning: General; Sustainability Transition: Human–Environment Relationship; Sustainable Development

Bibliography

Altieri M, Hecht S 1990 *Agroecology and Small Farm Development*. CRC Press, Boca Raton, FL

Barkin D 1998 *Wealth, Poverty, and Sustainable Development*. Editorial Jus, Mexico City

World Commission on Environment and Development 1987 *Our Common Future* (The Brundtland Report). Oxford University Press, Oxford, UK

D. Barkin

Development Theory in Geography

1. Overview of Post-World War Two Development Geography

Given the obvious significance of topics falling within the spatial problematic, it is surprising that until the last two decades of the twentieth century the contributions of geographers to development theory in the English-speaking world were comparatively limited. A 1960 report by leaders in the field estimated that only two dozen studies had been published on the issue of economic development between 1949 and 1959, and only half of these were on development in 'underdeveloped areas' (Ginsburg 1960, p. ix). Similarly, a 1967 article by D E Keeble showed that between 1955 and 1964, the journal *Economic Geography* published 251 major articles, only ten of which (under 4 percent) were concerned in whole or in part with issues of economic development, while the topic was addressed in only six out of 242 articles (2.5 percent) in the *Annals of the Association of American Geographers* (Keeble 1967, p. 243). A review of research published in *The New Zealand Geographer* cited a similar neglect of development theory (Chapman 1969).

The theories of economic development employed in this early postwar literature were fundamentally those which had been developed in disciplines such as economics, political science, and sociology. In particular, most work by geographers in the 1960s fell more or less explicitly into the category of 'modernization' studies, adopting the view that all countries could and should follow more or less the same course of development as the advanced industrialized capi-

talist countries, passing through 'stages of growth' which lead from premodern, preindustrial, and relatively stagnant societies to societies which are industrial, modern, and dynamic. Geographers gave modernization approaches a spatial bearing through examination of issues such as urbanization patterns (Berry 1961), regional development (Friedmann 1966), and the development of transportation systems (Gould 1960; Taafe et al. 1963), as well as through statistical analyses of the diffusion of modernization indicators (Gould 1964).

By the time development geographers had embraced modernization approaches, however, modernization theories had already been under serious attack for a decade or more by various 'core-periphery theorists'—including structuralist and dependency authors—who placed the claims of modernization theory up against the realities of development in peripheral regions of the global economy such as Latin America, and found the theory wanting. Structuralists found that the historical legacies of colonialism and imperialism—such as technological backwardness, maldistribution of wealth and income, and a large, underemployed rural labor force—made the development of modern industry in the periphery a much more problematic process than modernization theorists had suggested. Dependency authors more explicitly identified the process of development in core industrialized capitalist countries as an obstacle to development in countries of the global periphery, and more radical dependency and underdevelopment perspectives which became popular by the late 1960s went so far as to suggest the impossibility or unlikelihood of capitalist development in the periphery, precisely because of its relations with the core.

These structuralist and dependency perspectives, even though not originally articulated by geographers, helped bring greater attention to geographic differences in patterns of development and to relationships between development processes in different places. This spatial dimension was naturally attractive to many geographers and spurred a significant amount of core–periphery analysis, including theories regarding revolutionary change in Third World urban centers (Armstrong and McGee 1968), general theories of imperialism (Blaut 1970), and critiques of Eurocentrism in geographical research (Connell 1971). By the middle of the 1970s there were a large number of works which adopted either structuralist (e.g., Brookfield 1975) or dependency approaches (e.g., Santos 1974). This burst of core–periphery analyses opened a period in which attention to development issues in geography became much more sustained. Between 1974 and 1983, for example, articles which addressed development issues rose to 16 percent of all articles in *Economic Geography* and 13 percent in the *Annals*, while articles on development in the more recently founded *Antipode* reached almost 40 percent (Corbridge 1986, pp. 4–5).

Even with this more intensive theorization of Third World development and deployment of the spatial problematic, however, development geography remained more a theoretical borrower than a theoretical leader. By the mid-1970s, dependency approaches were already coming under increasing fire from Marxist scholars for fetishizing geographical relations and miscomprehending class issues, and were thus being forced into refinements which could take account of increasing complexity in the 'world-system,' including the possibility of 'dependent development' in what had formerly been peripheral areas. Moreover, the 1970s had seen the emergence within development studies of important work on gender issues, an emphasis which was still missing from work by development geographers.

By the late 1970s and early 1980s, a more fully-fledged turn in the direction of Marxist theory was evident in the work of many development geographers (Slater 1977), while others had taken note of the importance of world-systems theory (Taylor 1981). Eventually, development geographers also paid greater attention to the claims about dependency put forward by the partisans of 'dependent development' approaches (Forbes 1984, Corbridge 1986), and relatively complex arguments about the interaction of 'external' and 'internal' factors began to appear in works on issues such as regional development (Gore 1984) and urbanization (Armstrong and McGee 1985).

Within development theory in general, however, the issues raised by both dependency theory and its Marxist critics had largely faded from view in the 1980s. For neo-classically oriented economists, the dramatic growth of East Asian newly industrialized countries (NICs) in the 1970s and 1980s both undermined claims of blocked development on the periphery and vindicated assertions that unfettered markets performed better than those in which the state intervened. These arguments were countered by the claims of an emerging 'neo-Weberian' school which presented evidence for the important role of the state in guiding the economic growth of the Asian NICs.

In spite of the fact that theorization of Third World states by geographers has been limited (Glassman and Samatar 1997), emphasis on the political and economic role of the state was by no means foreign to development geography. Indeed, by the mid-1980s a number of development geographers had already begun to take up the theme of state activism in the East Asian growth story (Grice and Drakakis-Smith 1985, Browett 1986). Moreover, development geographers also foregrounded the spatial problematic in discussions of the state by highlighting the state's role in creating and maintaining regional disparities within developing countries (Watts 1983, Samatar 1989, Slater 1989).

Work on issues such as gender and development continued to gain little attention in development geography until the end of the 1980s, when more

studies began to emerge (e.g., Momsen and Townsend 1987). These included analyses of the gendering of immigration processes (Radcliffe 1986) and of the gendered environmental outcomes of development (Rocheleau 1987).

By the 1980s and 1990s, geography as a field could no longer be seen as a laggard and had begun to make what were arguably leading contributions to the formulation of theoretical positions surrounding important contemporary development issues. To some extent, this was less a matter of geography as a field catching up with development studies than it was a matter of development studies coming around to some of the issues which are part of the spatial problematic—in particular, geographic expansion of economic processes, human-environment interactions, and difference and spatiality.

2. *Contemporary Themes in Development Geography*

Geographical issues are now at the forefront of much work in development, a matter which can be highlighted by briefly summarizing two broad themes of contemporary research in development geography which are not only important within the discipline but are either already or potentially of leading significance to theory within development studies more generally.

2.1 *The Spatiality and Unevenness of Development*

While modernization approaches had been generally optimistic about the prospects for capitalist development in developing countries, much of the literature by dependency theorists had been highly pessimistic. Since at least the 1980s, however, many development theorists have been inclined towards a more nuanced and contextual approach which recognizes both possibilities for and limitations to capitalist development. In place of the earlier emphasis on either diffusion or blocked development, there is now much greater emphasis on the general geographical expansiveness of capital and the unevenness with which its expansion proceeds. The important theoretical work of David Harvey (1982) has enriched the conceptual foundations for analyses highlighting uneven internationalization of capital—analyses which have been carried forward by a variety of development geographers.

Much work on uneven development by development geographers has focused on uneven development at the national scale, especially uneven economic growth between regions (Watts 1983; Samatar 1989; Slater 1989). The most basic point to emerge from such studies has been that capitalist development—whether or not it is relatively dynamic in a given context—occurs as the result of differential power relations which have concrete spatial expression. These differential power relations lie behind the spatial unevenness of capitalist development and tend to work at odds with efforts to create greater sociospatial equality in the development process. Indeed, in countries which have been relatively successful at 'late' industrialization, the production of disparity is sometimes accelerated rather than slowed.

Directly related to the phenomenon of national unevenness in the development process are issues surrounding urbanization in developing countries. While it is less fashionable now than it once was to speak of 'over-urbanization,' there is still much emphasis on how regional disparities in growth relate to the national pre-eminence of primate urban centers—these often being former centers of colonial administration and/or commercial centers of long standing. Instead of gradual equilibration of growth rates between these centers and outlying regions, capitalist development has generally increased the dominance of primate urban centers. Indeed, it has been suggested by development geographers studying Asia that continued expansion of primate urban centers has led to a historically unique form of extended metropolitan growth which blurs preconceived boundaries between urban and rural, industry and agriculture (McGee 1991). The evolution of such extended metropolitan regions poses unique new problems for urban management and environmental sustainability in Asian countries.

At the same time, the explosive growth of particular urban centers in the context of global economic transformation has led to increasing investigation of the ways uneven development detaches the trajectories of urban centers from their 'hinterlands' while creating global networks of cities through which there is an intensified flow of money, goods, and information (Friedmann 1986). On this view, the spatial unevenness of development increasingly manifests itself as a bifurcation between the major nodes through which capital moves internationally (large urban centers) and much of the rest of the world, with non-urban areas of the developing world being particularly marginalized.

To present the unevenness of capitalist development in this way, however, might be to risk reification of the spatial dimensions of development at the expense of a spatially grounded analysis of social relations. Geographers have attempted to guard against such 'spatial fetishism'—that is, the view that geographical relations have their own independent dynamics, rather than being the spatial expressions of social processes—by foregrounding issues that include class struggles (Herod 1998) and gender relations (Hanson and Pratt 1995) in relationship to particular spatial dynamics. Uneven development thus implies not only uneven development between places but uneven power relations and opportunities for groups of people, dependent upon such characteristics as their class, gender, race or ethnicity, sexual orientation, and age.

Among development geographers, then, much attention has been paid to the spatialization of disparity in the development process. For example, feminist geographers have examined how transformations in Mexico's accumulation process have involved breaking down the hegemony of the family wage-based, male-dominated, and Mexico City-centered labor regime while substituting a lower-wage and geographically more decentralized regime in which women are employed in much larger numbers (Cravey 1998). They have also examined the migration of women to export processing zones which have been the key to economic growth in newly-industrializing countries of South-east Asia and Latin America (Chant and Radcliffe 1992, Chant and McIlwain 1995). Labor geographers have examined the strengths and weaknesses of attempts to create various forms of solidarity between labor and other popular organizations in developed countries such as the USA, and less developed countries such as those of Latin America (Herod 1995, Johns 1998). Authors writing about 'new social movements' and 'popular development' (Brohman 1996) have examined how such movements attempt to forge international linkages ('jumping scale') in order to strengthen their position. Geographers examining state power have noted both how it is transformed by processes such as 'globalization' and how it continues to play a crucial role in relation to various kinds of social struggles through which 'globalization' occurs (Agnew and Corbridge 1995, Webber and Rigby 1996, Dicken 1997). Finally, development geographers who foreground cultural dimensions of the social process have examined issues such as the importance of ethnic networks to economic growth in southern China (Hsing 1998), and struggles surrounding identity among South Asian communities in Tanzania (Nagar 1997).

All of these types of analyses exemplify certain aspects of the spatial problematic—in particular, issues of sociospatial difference in the process of development—and bring an explicitly geographical dimension to social critique. What is either implicit or explicit in such critique is an emphasis on the way general processes, such as capitalist development, can have different types of manifestations in different locations, because of the relationships between different social groups within and across those locations. Thus, the study of development requires balancing an adequately supple general theory with examination of the specifics of development in particular contexts. Theories of capitalist development must be able to explain why the features of capitalism which are posited as general can lead to different kinds of outcomes—e.g., 'late industrialization' in one setting and 'blocked development' in another; stagnant real wages in one setting and rising real wages in another; increasing marginalization of certain women workers in one setting and new opportunities for empowerment of women in another. This theoretically-informed attention to the details of development processes in specific settings is perhaps one of the most important features of work in development geography, and is particularly evident in work taking place around the second major theme of contemporary geographical work, globalization and human/environment interactions.

2.2 Globalization and Human/Environment Interactions

The internationalization of capital has become so extensive that many people in the social sciences speak of 'globalization'—a qualitatively new phenomenon in which national boundaries are increasingly unimportant to the functioning of the global economy. It can be argued that the literature on globalization frequently overstates the unboundedness of capital, but nonetheless there can be little doubt that the growth of international production, trade, finance, and information flows is reshaping global geography in crucial ways. Indeed, it is perhaps globalization as much as any other single phenomenon which seems to have brought geographical perspectives to the fore in the study of development.

Among the phenomena intensified by globalization are increasingly complex issues of environmental change and degradation. Within development geography, an important literature on issues of political ecology has taken up the task of theorizing environmental change in the context of globalization (Peet and Watts 1996). Political ecology can be seen as combining a focus on details of local ecosystems and social structures with attention to global structural forces. In addition, political ecologists have shown a strong interest in what Peet and Watts (1996) term 'liberation ecologies'—varied struggles to transform environmental conditions in developing countries, which are often approached by political ecologists through conceptual frameworks which are post-Marxist and/or poststructuralist, emphasizing the multiclass and frequently community-based forms of local environmental resistance (Rocheleau and Ross 1995). A number of these political ecologists have brought gender into the center of the research agenda, highlighting, for example, the importance of gender-differentiated control over local and household resources to the economic and environmental outcomes of development projects (Carney 1993). More generally, feminist political ecologists have noted how processes of transformation which are occurring in the global economy may have varied effects on women and men in specific localities, depending upon the social position in which they start and the sorts of resources to which they have (or are denied) access (Rocheleau et al. 1996). This may in turn place women at the core of movements to resist environmental degradation connected with development.

Political ecologists have paid special attention to issues of food and agriculture. Michael Watts' study of uneven development and the food crisis in Nigeria (1983) helped bring the issue of food and access to it firmly onto the agenda, and subsequent research by development geographers has continued to analyze access to and control over agricultural resources. Topics examined in relation to such issues of access and control include the impact of contract farming on local food production (Little and Watts 1994) and the impact of new agricultural technologies on poverty and food security (Yapa 1993, 1996).

Globalization is not a process driven solely from above, but must be taken to include attempts to counteract the unevenness of capitalist development with 'globalization from below'—as, for example, with the attempts at international labor and social movement solidarity mentioned earlier. 'Globalization from below' also includes the attempts of local environmental organizations to 'jump scales' and to make their struggles more national or international. This pertains not only to struggles over environmental conditions and control over resources in rural and agrarian communities but to popular struggles over the urban environment (the 'brown agenda'). More generally, development geographers have become increasingly concerned with attempts to build national and international linkages between various types of 'new social movements'—including women's movements and other movements associated with so-called 'identity politics.' These actors are generally seen as the core of a project to promote 'popular development.'

Popular development literature is closely interrelated with approaches which are postMarxist and postcolonial, yet literature on postcoloniality has as yet made less of a mark on development geography than on some other fields. Even so, it has not been absent from discussions. For example, David Slater (1997) has championed postcolonial approaches as an antidote to Eurocentrism, while a variety of scholars have brought postcolonial approaches to analyses of specific development issues. These approaches are part of a broader 'cultural turn' in geography that is still in the process of making its mark within the field.

3. *The Future of Development Geography*

The various but often overlapping and interconnected strands of research identified here—while all exemplifying the spatial problematic in development geography—need not be anchored in any one approach or research agenda which is putatively constitutive of geography as a discipline. Nonetheless, the proliferation of research topics and agendas also brings eventual attempts to synthesize and to identify unifying principles. This task has been undertaken by authors such as John Brohman (1996), who identifies the connecting threads in various attempts to construct alternatives to neoliberal development models, and Philip Porter and Eric Sheppard (1998), who identify a dialectic of unity-in-diversity running through processes of global change and development. Such syntheses must attempt, in particular, to connect the traditions of political economy of development, political ecology, and the emerging poststructuralist and postcolonial approaches to development studies. Indeed, it is at the conjunction of these approaches, particularly as they address various issues surrounding globalization, that much of the most interesting work may well be done. As the twenty-first century opens, the centrality of globalization, uneven development, and popular responses to these processes by diverse groups seems clear; and much of the specific research agenda confronting development geographers will be rgelated to this central problematic. At the same time, the pace and scale of global change itself makes attempts to predict where development theory in geography will head in the future a daunting task.

See also: Cities: Capital, Global, and World; Cultural Geography; Ecology, Cultural; Ecology, Political; Economic Geography; Economic Transformation: From Central Planning to Market Economy; Environment and Development; Feminist Political Ecology; Food Security; Foreign Investment: Direct; Gender and Environment; Gender and Feminist Studies in Geography; Global Environmental Change: Human Dimensions; Globalization: Geographical Aspects; Human–Environment Relationships; International Trade: Geographic Aspects; Local–Global Linkages: Environmental Aspects; Marxist Geography; Postcolonial Geography; Postmodernism in Geography; Social Movements: Environmental Movements; Social Movements, Geography of; Space and Social Theory in Geography; Spatial Equity; Sustainable Development; Uneven Development, Geography of

Bibliography

Agnew J, Corbridge S 1995 *Mastering Space: Hegemony, Territory and International Political Economy*. Routledge, London

Armstrong W R, McGee T G 1968 Revolutionary change and the Third World city: A theory of urban involution. *Civilisations* **18**(3): 353–77

Armstrong W R, McGee T G 1985 *Theatres of Accumulation: Studies in Asian and Latin American Urbanization*. Methuen, London

Berry B J L 1961 City-size distributions and economic development. *Economic Development and Cultural Change* **9**: 573–87

Blaut J M 1970 Geographic models of imperialism. *Antipode* **2**(1): 65–85

Brohman J 1996 *Popular Development: Rethinking the Theory and Practice of Development*. Blackwell, Oxford

Brookfield H C 1975 *Interdependent Development*. University of Pittsburgh Press, Pittsburgh, PA

Browett J G 1986 Industrialization in the global periphery: The significance of the Newly Industrializing Countries of East and Southeast Asia. *Environment and Planning D: Society and Space* **4**(4): 401–18
Carney J 1993 Converting the wetlands, engendering the environment: The intersection of gender with agrarian change in the Gambia. *Economic Geography* **69**(4): 329–48
Chant S H, McIlwaine C 1995 *Women of a Lesser Cost: Female Labour, Foreign Exchange, and Philippine Development*. Pluto Press, London
Chant S (ed.) 1992 *Gender and Migration in Developing Countries*. Belhaven, London
Chapman M 1969 Geography and the study of development. *Journal of Developing Areas* **3**: 319–38
Connell J 1971 The geography of development. *Area* **3**(4): 259–65
Corbridge S 1986 *Capitalist World Development: A Critique of Radical Development Geography*. Rowman & Littlefield, Totowam, NJ
Cravey A J 1998 *Women and Work in Mexico's Maquiladoras*. Rowman & Littlefield, Lanham, MD
Dicken P 1997 Transnational corporations and nation-states. *International Social Science Journal* **151**: 77–89
Dixon C, Drakakis-Smith D (eds.) 1997 *Uneven Development in South East Asia*. Ashgate, Aldershot, UK
Forbes D 1984 *The Geography of Underdevelopment: A Critical Survey*. Croom Helm, London
Friedmann J 1966 *Regional Development Policy: A Case Study of Venezuela*. MIT Press, Cambridge, MA
Friedmann J 1986 The world city hypothesis. *Development and Change* **17**: 69–84
Ginsburg N S (ed.) 1960 *Essays on Geography and Economic Development*. University of Chicago Press, Chicago
Glassman J, Samatar A I 1997 Development geography and the third-world state. *Progress in Human Geography* **21**(2): 164–98
Gore C G 1984 *Regions in Question: Space, Development Theory and Regional Policy*. Methuen, London
Gould P R 1960 The development of the transportation pattern in Ghana. *Northwestern University Studies in Geography* **5**
Gould P R 1964 A note on research into the diffusion of development. *Journal of Modern African Studies* **2**: 123–5
Grice K, Drakakis-Smith D 1985 The role of the state in shaping development: two decades of growth in Singapore. *Transactions (Institute of British Geographers)* **10**(3): 347–59
Hanson S, Pratt G 1995 *Gender, Work, and Space*. Routledge, London
Harvey D 1982 *The Limits to Capital*. University of Chicago Press, Chicago
Herod A (ed.) 1998 *Organizing the Landscape: Geographical Perspectives on Labor Unionism*. University of Minnesota Press, Minneapolis, MN
Herod A 1995 The practice of international labor solidarity and the geography of the global economy. *Economic Geography* **71**(4): 341–63
Hsing Y-T 1998 *Making Capitalism in China: The Taiwan Connection*. Oxford University Press, New York
Johns R A 1998 Bridging the gap between class and space: US worker solidarity with Guatemala. *Economic Geography* **74**(3): 252–71
Keeble D E 1967 Models of economic development. In: Chorley R J, Haggett P (eds.) *Models in Geography*. Methuen, London
Little P D, Watts M J (eds.) 1994 *Living Under Contract: Contract Farming and Agrarian Transformation in Sub-Saharan Africa*. University of Wisconsin Press, Madison, WI
McGee T G 1991 The emergence of *desakota* regions in Asia: Expanding a hypothesis. In: Ginsburg N, Koppel B, McGee T G (eds.) *The Extended Metropolis: Settlement Transition in Asia*. University of Hawaii Press, Honolulu, HI
Momsen J H, Townsend J G (eds.) 1987 *Geography and Gender in the Third World*. Hutchinson, London
Nagar R 1997 Communal places and the politics of multiple identities: The case of Tanzanian Asians. *Ecumene* **4**(1): 3–26
Peet R, Watts M (eds.) 1996 *Liberation Ecologies: Environment, Development, Social Movements*. Routledge, London
Porter P W, Sheppard E S 1998 *A World of Difference: Society, Nature, Development*. Guilford Press, New York
Radcliffe S 1986 Gender relations, peasant livelihood strategies and mobility: a case study from Cuzco, Peru. *Bulletin of Latin American Research* **5**(2): 29–47
Rocheleau D 1987 Women, trees and tenure: implications for agroforestry research and development. In: Raintree J (ed.) *Land, Trees and Tenure*. ICRAF, Nairobi and Land Tenure Center, Madison, WI
Rocheleau D, Ross L 1995 Trees as tools, trees as text: struggles over resources in Zambrana Chacuey, Dominican Republic. *Antipode* **27**(4): 407–28
Rocheleau D, Thomas-Slayter B, Wangari E (eds.) 1996 *Feminist political ecology: global issues and local experiences*. Routledge, London
Samatar A I 1989 *The State and Rural Transformation in Northern Somalia 1884–1986*. University of Wisconsin Press, Madison, WI
Santos M 1974 Geography, Marxism, and underdevelopment. *Antipode* **6**: 1–9
Slater D 1977 Geography and underdevelopment—II. *Antipode* **9**: 1–31
Slater D 1989 *Territory and State Power in Latin America: The Peruvian Case*. St. Martin's Press, New York
Slater D 1997 Geopolitical imaginations across the North-South divide: issues of difference, development and power. *Political Geography* **16**(8): 631–53
Taafe E J, Morrill R L, Gould P R 1963 Transport expansion in underdeveloped countries: a comparative analysis. *Geographical Review* **53**: 503–29
Taylor P J 1981 Political geography and the world economy. In: Burnett A D, Taylor P J (eds.) *Political Studies from Spatial Perspectives*. Wiley, Chichester, UK
Watts M 1983 *Silent Violence: Food, Famine & Peasantry in Northern Nigeria*. University of California Press, Berkeley, CA
Webber M J, Rigby D L 1996 *The Golden Age Illusion: Rethinking Postwar Capitalism*. Guilford Press, New York
Yapa L 1993 What are improved seeds? An epistemology of the green revolution. *Economic Geography* **69**(3): 254–73
Yapa, L 1996 What causes poverty?: A postmodern view. *Annals of the Association of American Geographers* **86**(4): 707–28

J. Glassman

Developmental Behavioral Genetics and Education

Behavioral genetics is the study of genetic influences on behavior, and it is subdivided into molecular behavioral genetics and quantitative behavioral gen-

etics (Plomin et al. 1997a). Molecular behavioral genetics studies associations between differences in organisms' deoxyribonucleic acid (DNA) and their behavior. Too little is known on how single genes affect human behavior, however, to draw meaningful conclusions for education.

1. Concepts

1.1 Quantitative Behavioral Genetics

Quantitative behavioral genetics accounts for *individual differences* in behavior in terms of several sources of variability, and it is applied to mental disorders as well as to continuously distributed traits such as IQ that are influenced by many genes. This approach is usually traced back to Galton (1869), who was interested in the origins of excellence. Modern quantitative behavioral genetics partitions the observed or phenotypic variance (the mean squared deviation of individual scores from the sample mean) into (a) additive genetic variance or the additive effects of single genes, (b) gene dominance or the interactive effects of the two homologous genes at the same gene locus, (c) epistasis or the interactive effects of genes at different loci, (d) shared environment, defined as contributing to the similarity of persons reared in the same family independent of their biological relatedness, and (e) nonshared environment, defined as *not* contributing to the similarity of persons reared in the same family. In this context, *environment* refers to the effects of all nongenetic sources of variance, including nongenetic biological variables such as infectious diseases and nutrition.

The importance of these sources of variance is inferred from the correlations between relatives, taking their genetic and environmental similarities into account. For example, the correlation between monozygotic (MZ) *twins* reared apart estimates the combined influence of all genetic sources of variance (broad-sense heritability), whereas twice the correlation between a biological mother and her adopted-away offspring estimates the importance of additive genetic variance (narrow-sense heritability) only. Moreover, the reliable differences between MZ cotwins reared together estimate the importance of the nonshared environment, and the correlation between adoptive siblings estimates the importance of the shared environment. Correlations between different kinds of relatives may be analyzed simultaneously, using structural equation modeling (Neale and Cardon 1992).

1.2 Heritability

An important concept in quantitative behavioral genetics is the *heritability* (h^2) of a trait, defined as the ratio of the genetic variance (V_G) to the phenotypic variance (V_P):

$$h^2 = V_G / V_P$$

For example, if the correlation between the trait levels of MZ twins reared apart is $r = 0.50$, it is concluded that 50 percent of the individual differences in that trait reflect effects of different genes, whereas the other 50 percent reflect effects of nonshared environments, implying that the broad-sense heritability of that trait is $h^2 = 0.50$. Note that correlations between MZ cotwins depend on the genetic variability between twin pairs and on the variability of their environments, implying that the heritability statistic may differ between populations and change across time. For example, if impoverished family environments are improved by education to foster equal opportunities for all children, the average level of performance will rise. But as the previous variability in environments will thereby be reduced, the heritability estimates will also rise. Both seem to have happened in industrialized countries in the latter part of the twentieth century (Flynn 1987, Plomin et al. 1997a).

1.3 Mean Levels Versus Individual Differences

Hence influences on mean levels have to be distinguished from influences on individual differences. Heritability may be high because the environmental conditions in that population are homogeneous, leaving genetic factors to account for the individual differences. That, however, does not imply that the mean level cannot be raised by environmental intervention. Rather, by improving environmental conditions for entire populations, mean levels may be raised in spite of high heritabilities of the individual differences. This is not purely hypothetical: Although IQ is substantially heritable within cohorts, massive gains across cohorts in average IQ have been observed in industrialized countries since World War II (Flynn 1987), reflecting secular changes in environmental conditions. Another example is stature: although being highly heritable within cohorts, average height increased substantially during the twentieth century, probably owing to secular trends in nutrition (Vogel and Motulsky 1997).

1.4 Developmental Behavioral Genetics

This is the study of genetic and environmental influences on individual differences in behavioral development, and it combines a behavior–genetic with a developmental approach. A topic for developmental behavioral genetics that can be studied longitudinally as well as cross-sectionally is how heritability estimates are related to the age of the samples under study.

1.5 Genetic and Environmental Continuity and Change

These concepts refer to the extent that observed stability and change in individual differences reflect stability and change of the effects of genes and environments. As they are inferred from the pattern of concurrent correlations and longitudinal cross-correlations between relatives, they require longitudinal studies. These allow for estimates of (a) genetic and environmental influences at time 1, (b) genetic and environmental influences at time 2, and (c) the genetic and environmental continuities between time 1 and time 2. For example, if the concurrent correlations between MZ twins reared apart are higher than their longitudinal cross-correlations (twin A at time 1 × twin B at time 2, and vice versa), this shows lack of continuity of the effects of genes. Correspondingly, if the concurrent correlations between adoptive siblings are higher than their longitudinal cross-correlations, this shows lack of continuity of the effects of the shared environment.

If heritabilities are high but genetic continuities are low, genes are important but their effects change with age. For example, genes that foster high abilities in childhood may have less beneficial effects in adulthood. It is important to realize that, although persons' DNA usually does not change, the effects of their genes may change across time, hence genetic influence does not imply lack of change. This is obvious for the genetically driven physical changes during puberty, but it has also been shown for individual differences in cognitive development that are more strongly synchronized for monozygotic than for dizygotic (DZ) twins (Wilson 1983).

2. Findings

A meta-analysis of the world literature on family resemblance in IQ by Loehlin (1989) yielded broad-sense heritability estimates of about 0.50 for IQ. Moreover, it suggested moderate effects of the shared environment. That genes have an effect on individual differences in IQ is meanwhile established beyond any reasonable doubt. The evidence concerning genetic and environmental influences on personality traits such as extraversion and conscientiousness also suggests substantial heritabilities (Loehlin 1992). The mechanisms that mediate the effects of genes on behavior are largely unknown, however.

2.1 Heritability Estimates and Age

Loehlin's (1989) estimates for IQ were inferred from a meta-analysis of diverse studies that tested participants at various ages, whereas more fine-grained analyses show that heritability estimates depend on age: In a meta-analysis of 103 cross-sectional twin studies, McCartney et al. (1990) focused on relations between twin age and twin similarity. Whereas older MZ twins were more similar in IQ than younger MZ twins, the opposite was found for DZ twins. This suggests that genes become more influential whereas shared environment becomes less influential across the life span.

This finding is confirmed by longitudinal twin and adoption studies: The Louisville Twin Study followed the cognitive development of MZ and DZ twins from 3 months to 15 years. Whereas the correlations between MZ twins increased from $r = 0.66$ to $r = 0.88$, the correlations between DZ twins decreased from $r = 0.67$ to $r = 0.54$, suggesting growing genetic and diminishing environmental influence from infancy to adolescence (Wilson 1983). In the longitudinal Colorado Adoption Project (Plomin et al. 1997b), the IQ correlation between adoptive parents and their adopted children dropped from $r = 0.20$ at age 3 years to $r = 0.00$ in adolescence, suggesting weak and diminishing shared environmental influence. In contrast, the IQ correlation between biological parents and their own children in control families rose from 0.17 at age 3 years to 0.30 at age 16 years, suggesting narrow-sense heritabilities of 0.60 in adolescence.

2.2 Gene–Environment Covariation

The most reasonable explanation of the growing genetic influence on cognitive abilities across the life span is that the environments that contribute to individual differences in cognitive development become increasingly a function of the individuals' genotypes (Plomin 1994). The older children and adolescents become, the more they create and choose their environments, and they probably do so in ways such that their environments fit their genetically influenced dispositions. For example, more intelligent children tend to be more successful in intellectually demanding tasks and thus tend to persist with them, whereas less intelligent children experience more failure in those tasks and therefore turn to less demanding activities. This is sometimes referred to as a smörgåsbord view of the environment: The technical term for that is *genotype–environment covariation* of the evoked (or reactive) and active variety.

3. Implications for Education

The main implication of developmental behavioral genetics for education is that there are ubiquitous genetically influenced individual differences, as has been shown most convincingly for *intelligence*. General intelligence is problem-solving ability, and it is strongly related to speed and potential of learning. Thus individual differences in learning speed and learning potential are also genetically influenced. Indeed, the *dynamic testing* approach (Grigorenko and Sternberg 1998) focuses directly on the potential to learn.

However, environments are also important. Behavior–genetic research shows that at least some of the variability in behavioral traits reflects variability in environments, and this is only the variability between environments in a particular population at a given time. Most behavioral–genetic research is insensitive to secular changes in environmental conditions that affect entire populations in a similar way as is likely to have occurred for IQ (Flynn 1987). And it is also insensitive to *potential* environmental changes that might be induced by purposeful intervention. The latter is the subject of educational, not behavioral–genetic research.

Hence the prime message from behavioral genetics for education is that the uniqueness of human personality and abilities requires different or even individualized intervention programs. It is unlikely that there is a curriculum that fits all children's needs despite their different abilities and interests. Neither can all children be expected to attain the same level of proficiency, nor can they be expected to learn at the same pace, nor can a given child be expected to attain the same level of expertise in all fields. That, of course, makes education not less important but more demanding.

Moreover, there are two important lessons to be learnt from the principle of gene–environment covariation. First, if environmental variables are associated with behavior, we do not know whether they are the cause of that behavior. The sources of the behavior may as well be genetic, the genes being associated with the environmental variables (Plomin 1994). It needs genetically informative designs such as twin and adoption studies to disentangle these influences. Second, the principle of gene–environment covariation suggests that children should be provided with a variety of microenvironmental contexts to select their preferred niches that fit their inherited dispositions (Wachs 1992). The best learning environment allows children to discover and develop their natural interests, proclivities, and special talents. Behavior–genetic findings indicate that they will do so anyway, and it is probably useful to help them to proceed on their way.

See also: Behavioral Genetics: Psychological Perspectives; Economic Status, Inheritance of: Education, Class, and Genetics; Genetic Factors in Cognition/Intelligence; Genetic Studies of Behavior: Methodology; Genetic Testing and Counseling: Educational and Professional Aspects; Genetics and Development; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Intelligence, Prior Knowledge, and Learning

Bibliography

Flynn J R 1987 Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin* **101**: 171–91

Galton F 1869 *Hereditary Genius: An Inquiry into its Laws and Consequences*. Macmillan, London

Grigorenko E L, Sternberg R J 1998 Dynamic testing. *Psychological Bulletin* **124**: 75–111

Loehlin J C 1989 Partitioning environmental and genetic contributions to behavioral development. *American Psychologist* **44**: 1285–92

Loehlin J C 1992 *Genes and Environment in Personality Development*. Sage, Newbury Park, CA

McCartney K, Harris M J, Bernieri F 1990 Growing up and growing apart: A developmental meta-analysis of twin studies. *Psychological Bulletin* **107**: 226–37

Neale M C, Cardon L R 1992 *Methodology for Genetic Studies of Twins and Families*. Kluwer, Dordrecht, The Netherlands

Plomin R 1994 *Genetics and Experience: The Interplay Between Nature and Nurture*. Sage, Thousand Oaks, CA

Plomin R, DeFries J C, McClearn G E, Rutter M 1997a *Behavioral Genetics*. Freeman, New York

Plomin R, Fulker D W, Corley R, DeFries J C 1997b Nature, nurture, and cognitive development from 1 to 16 years: A parent–offspring adoption study. *Psychological Science* **8**: 442–47

Vogel F, Motulsky A G 1997 *Human Genetics: Problems and Approaches*. Springer, Berlin

Wachs T D 1992 *The Nature of Nurture*. Sage, Newbury Park, CA

Wilson R S 1983 The Louisville Twin Study: Developmental synchronies in behavior. *Child Development* **54**: 298–316

P. Borkenau

Developmental Psychology

Developmental psychology is a vast and diverse field of psychology. Some developmental psychologists conduct research on conception, the early breaths of life, and the unfolding maturation of infants. Many others focus on children, their steady growth, accumulating cognitive and social skills, and awakening independence. Development of adolescents is equally of interest: the gravity of pubescence, the search for achievement, and the hovering influences of peers and schools. Developmental psychology is also about launching into adulthood, initiating and nurturing mature relationships, propagating families, and pursuing specializations and careers. Recently, some scholars have addressed concerns about settling into midlife, with shifting constraints and redefined opportunities. Finally, there have been long-standing developmental interests in aging throughout late life, a long period in which many changes are more likely to be losses than gains.

After scouring a wide range of textbooks, handbooks, and journals, one's impression may very well be that developmental psychology is about the sum of all of the above—and more. Indeed, it concerns the normative changes that occur during these phases of life, but it is also about the differences and variations in individuals' lives—the diversity of development—as

well as the biological and cultural conditions that occasion unique developments. Given the breadth of the field, however, the published works in developmental psychology are more selective about the issues, theories, and methods to which they attend. For this reason, a simple but broadly accommodating definition of developmental psychology is offered here. Following the definition, a brief analysis of some of the key concepts in developmental psychology is presented. Subsequently, both the emergence (history) and present structure of the field are sketched.

1. *Definition of Developmental Psychology*

A deliberately simple definition is as follows: developmental psychology is the study of psychological phenomena that change at any point in the life span. This seemingly uncontroversial rendition, however, contains several terms that require elaboration, if not clarification. Notably, all of the key terms of the definition—viz., developmental, study, psychological phenomena, change, and life span—may be interpreted differently and perhaps even contested by notable developmental scholars. Accordingly, the field of developmental psychology may actually be described as containing multiple legitimate clusters of developmental psychologies. Although these clusters share some common assumptions, interests, and themes, they differ—sometimes dramatically so—from one another in these and other respects.

2. *Key Concepts of Developmental Psychology*

Unpacking the seemingly simple definition above into its principal terms provides insight into the present status of developmental psychology. In so doing, both the commonalities and the diversities of contemporary developmental psychologies are exposed.

2.1 *Concept of Development*

The fact that there is commonality and diversity in developmental psychology is embodied in the terms development and change. Most contemporary theorists and approaches would subscribe to an underlying developmental perspective, or the notion that an understanding of the present psychological moment (i.e., behavior, belief, knowledge, performance, phenomenon) requires attention to the conditions and processes that led to it. Concurring that developmental psychology directs attention to the sources and courses of psychological phenomena does not imply, however, that there is agreement on whether all (or, if not, which) changes constitute developmental change.

The 'concept of development' refers to the criteria (if any) for differentiating change from development. This concept varies across approaches to developmental psychology. For example, there is disagreement about whether developmental changes can or should be universal (i.e., occurring to all individuals), irreversible (i.e., not naturally reversible or even immune to intervention), qualitative (i.e., stage-like rather than step-like and gradual), or goal directed (related or unrelated to a possible endstate). Legitimate, but contrasting, developmental psychologies rest on dissimilar conceptual assumptions concerning the extent of the difference, if any, between change and development (e.g., Harris 1957). For example, developmental theories that derive inspiration from models of biological maturation may attend principally to early phases of life (through adolescence) and construct theories that emphasize progressive and even universal growth. In contrast, other theories that are inspired by differential influences of proximal contexts, historical periods, and cultural settings are more likely to emphasize multiple developmental directions and perhaps even strong interactions among levels of influences (e.g., biological and cultural). Vivid contrasts in developmental theories may be viewed in a recent collection of chapters devoted to theoretical issues in developmental sciences (see Elder 1998, Gottlieb et al. 1998, Overton 1998, Schweder et al. 1998).

2.2 *Developmental Research*

The study of psychological development refers to the practices and traditions of scholarly research. A wide range of positions on how psychological development can or should be studied is evident. Methods employed in developmental investigations include virtually all those available in the biological, psychological and social sciences. The reason for this is that the phenomena investigated in developmental studies include the full range of processes or performances targeted by other psychologists, but also the (interactive) influences from biological and sociohistorical realms. From the social to the cognitive, from the macro to the micro, methods for investigating developmental change may be adapted from a wide variety of methods for investigating snapshots and profiles of human performance.

Notably, developmental psychologists pursuing research under the aegis of contrasting models or theories may adopt diverging approaches to the collection, analysis, and interpretation of developmental data. Available methods of developmental research include the large categories of experimental and quasi-experimental procedures, individual and profile difference studies, and qualitative techniques. Each of these categories may feature information garnered from standardized tests, experimental tasks, assays, questionnaires, surveys, interviews, textual interpretations, and combinations thereof. In addi-

tion, procedures are available to accommodate a variety of intentions, including a focus on description (what is changing or for whom?), explanation (why is it changing or what causes the change?), and intervention (can the change be remediated or modified?).

Nevertheless, whatever the research question or procedure, all developmental methods have a common paramount concern. Specifically, in developmental research an estimate of chronological age or developmental time is involved. Thus, typically some variation of cross-sectional (i.e., comparison of different age groups at one point in time) or longitudinal (i.e., examining changes in one or more groups across time) designs is implemented. Whatever one's assumptions about the concept of development, generating inferences about developmental change requires attention to more than one slice of the human life span. Thus, underlying the methodological diversity is the shared developmental goal of assessing the 'how' and 'why' of human change. Moreover, the array of methods may be necessary for understanding complex phenomena of individuals developing psychologically in the contexts of biological processes and social-environmental influences. In this respect, developmental methods associated with research in such areas as person and pattern analyses (e.g., Magnusson and Stattin 1998) and life-span change (e.g., Baltes et al. 1998) embody promising research endeavors.

2.3 Psychological Phenomena and Change

What psychological phenomena merit attention from developmental psychologists? The most inclusive view of the focus of developmental psychology permits the phenomena to include the behaviors and events that are measured by any appropriate method or technique. From this view, the principal restriction is that the phenomenon under investigation will be subject to change, at least at some point in the life span. According to developmental scientists, most psychological phenomena reach this criterion: What aspects of personality, affect, social, interactional, self-related beliefs, cognition, knowledge, motor skills, perception, or neurological events are not subject to change? To many developmental scientists, the answer is that virtually all aspects of human (psychological) phenomena undergo some manner of change at some point in the lifespan. Thus, the broadest view would maintain that most psychological phenomena constitute a legitimate focus of developmental psychology. (More narrow views, in which only selected changes are the foci of developmental psychology are supported by conceptual positions as sketched earlier.)

Naturally, specific theories or models of developmental psychology concentrate on different clusters of phenomena, and in some instances these are mutually exclusive. For example, in the middle part of the twentieth century there was much controversy over whether mechanistic (e.g., behavioral) or organismic (e.g., Piagetian) developmental psychologies implemented methods that identified optimal, and perhaps even acceptable, psychological change phenomena (e.g., Reese and Overton 1970). At present, an appropriate emphasis in evaluating whether phenomena are appropriate for developmental research is on whether the measures and observations can be justified from explicit and coherent perspectives on developmental theory and method.

2.4 Period of Change

Life span is the final term from the definition requiring clarification. The simplest interpretation of this term is that interesting and important psychological changes occur throughout life, but numerous challenging and elucidating implications have been derived from this principle (e.g., Baltes et al. 1999). Developmental psychologists focus their theoretical, methodological, and intervention efforts during one or more common phases of the life span, such as childhood, adolescence, midlife, or late life. Historically, many developmental psychologists have emphasized the roughly 15 percent of the typical life span known as infancy and childhood. A smaller but growing cadre of scholars concentrate on changes occurring during the approximately 10 percent of the life span known as adolescence. The remaining 75 percent of the typical life span constitutes the years of adulthood, ranging from young through the middle to the oldest ages. Unsurprisingly, telling the whole tale of development across the life span in any phenomenon—much less all psychological phenomena—is not the goal of life span developmental scientists. From the perspective that much developmentally interesting phenomena occur at all points of the life span, all developmental scholars contribute to a picture that will eventually be more comprehensive and coherent.

3. Emergence and Status of Developmental Psychologies

Seasoned observers of the field of developmental psychology are keenly aware of its formidable array of active theories, methods, and ideas. Recent publications testify to the vitality of the field, whether the focus is on theoretical or systemic issues (e.g., van Geert and Mos 1991), historical scholarship (e.g., Eckardt et al. 1985, Parke et al. 1994), or substantive empirical research on particular processes (e.g., Bornstein and Lamb 1999, Birren and Schaie 1996).

3.1 Historical Notes

As perhaps a sign of a maturing science, the emergence of developmental psychology received considerable

attention in the latter half of the twentieth century. Perplexing more than one observer is the sheer variety of perspectives on human development. Although distinct in several ways, contemporary theoretical systems and age-related divisions of developmental psychology can be linked to historical precedents in the nineteenth and twentieth centuries (e.g., Cairns 1998, Dixon and Lerner 1999). Although several eighteenth century European scholars (e.g., D. Tiedemann, F. A. Carus, and J. N. Tetens) produced prescient age-oriented sciences, the nineteenth century is often viewed as a crucial period in the emergence of developmental psychology. Numerous influences arising during this era have been cited (e.g., Richards 1987), including (a) the rise of evolutionary perspectives in biology (e.g., Darwinism) and other sciences (e.g., geology), (b) the appearance of historical views in philosophy (e.g., Hegel, Marx), (c) social pressures to accommodate education requirements of growing populations of children, and (d) the growth and institutionalization of psychology as an intellectual entity separate from biology and philosophy (e.g., William James). Arguably, the notion that present phenomena may be linked to past and present conditions is a hallmark of the intellectual life of the nineteenth century (Mandelbaum 1971). By the early twentieth century, psychologists were applying this concept of development to behavioral change, and even differentiating their views of development from one another. For example, the distinctions in theoretical perspectives among James Mark Baldwin (e.g., 1895), John Dewey (e.g., Cahan 1994), Sigmund Freud (e.g., Sulloway 1979), G. Stanley Hall (e.g., 1883), William James (e.g., 1890), Jean Piaget (e.g., 1983), Lev S. Vygotsky (e.g., 1929), and John B. Watson (e.g., 1924) are detectable. The theoretical differences are distributed along the dimensions of such generic and long-standing contrasting positions as origin of development (e.g., nature–nurture) and characteristics of developmental change (e.g., continuity–discontinuity). Rooted in differentiable assumptions, contemporary theories and systems of developmental psychology, may be traced to a variety of sources from ancient philosophical, to the fertile nineteenth century, to more recent influences (Cairns 1998, Dixon and Lerner 1999, Parke et al. 1994).

3.2 Present Status

In an initial effort to clarify the extent of divergent perspectives residing in the field of developmental psychology, two prominent and recent handbooks of development in the early and late phases of the life span were examined. The collections boast authors who are widely known specialists, and contain excellent tutorials on important developmental topics. The topics covered in the volume on early life development (edited by Bornstein and Lamb 1999) include the role of (a) cultural contexts, (b) school and community influences, (c) neurobiological aspects, (d) perceptual changes, (e) cognitive theories, (f) language emergence, (g) personality and temperament development, (h) familial relationships, and (i) peer or friendship patterns. There are also chapters on (j) theoretical/historical issues, (k) methodological characteristics, and (l) life span perspectives. A recent collection of reviews of adult development edited by Birren and Schaie (1996) contain a similar range of topics: (a) theoretical overviews, (b) methodological recommendations, (c) biological, neurological, and genetic influences, (d) gender-related differences, (e) cultural and racial perspectives, (f) sensory changes, (g) cognitive and intellectual aging, (h) psychosocial and personality processes, as well as (i) familial, career, and environmental influences.

Although impressively similar in range of topics, among the hundreds of references in each volume are a scant few appearing in both. Apparently, much work remains to be done in portraying the patterns and causes of development across (artificially) divided phases of the life span. Indeed, linking the descriptive patterns will be more manageable than the explanations. Naturally, the why questions in child development are of the sort: What accounts for the predominant patterns of growth and why are there exceptions? In contrast, the why questions in adult development may be summarized as what accounts for the predominant patterns of decline and are there any systematic exceptions?

The variety in contemporary developmental psychology derives in part from the fact that human development as a whole is fascinatingly (but unforgivingly) complex. This observation applies at any point in the life span, but is perhaps exacerbated by the existence of changes throughout life. At the theoretical level, it derives also in part from the fact that the developmental psychologies of the twentieth century have common sources, but often heterogeneous specific philosophical lineages. Different strands of developmental phenomena are captured adequately by select combinations of theories and methods, but no single all-encompassing developmental metatheory has established unparalleled pre-eminence. There are multiple families of developmental theories, multiple systems of developmental psychology. From this pluralistic perspective, each of the systems is more or less adequate in sponsoring theoretical and empirical work.

4. Contemporary Structure of the Field

Developmental psychologists share interests in human psychological change. The enormous variety of topics and perspectives covered by this common interest may be viewed as complementary, but not often overlapping.

4.1 Organizations

In the latter years of the twentieth century, a variety of institutional support mechanisms appeared. These included organizations and societies, journals and regular publications, training and educational sites, professional and career opportunities, and sources of research funding. With respect to societies, organizations are available that reflect interests in general infant and child development (e.g., Society for Research in Child Development), adolescence (e.g., Society for Research on Adolescence), adult development and aging (e.g., Gerontological Society of America), and even all phases of the life span (e.g., International Society for the Study of Behavioral Development). The American Psychological Association sponsors several active and large divisions with developmental agendas (notably, Division 7: Developmental Psychology; Division 20: Adult Development and Aging). In addition to these societies, numerous more focused associations have appeared and prospered. These include the long-standing Jean Piaget Society, a series of Cognitive Aging conferences, and a new child-oriented Cognitive Development Society.

4.2 Significant Publications

Regarding publications, the number of research journals related to developmental psychology has expanded enormously. Prominent examples include *Developmental Psychology*, *Adult Development and Aging*, *Infant Behavior and Development*, *Child Development*, *Journal of Research in Adolescence*, *Journal of Gerontology: Psychological Sciences*, and *International Journal of Behavioral Development*. In addition, several theoretical journals, such as *Human Development* and *Developmental Review*, have been published successfully for many years.

When initially produced, regular publications, such as handbook or advances series, could reasonably seek to cover broad topics such as child psychology or psychology of aging. More recently, however, the number of handbooks has multiplied, and the topics have become more focused as the field has become more specialized. Today, there are handbooks on such specialized topics as infancy research, infant perception, infant mental health, child language, adolescence, parenting, aging and the family, emotion and adult development, and cognitive aging.

What can be concluded about the structure of the field of developmental psychology? Perhaps primarily that, reflecting the rapidly advancing specialization and knowledge generation that occurred during the second half of the twentieth century, it is variegated but well-organized and copiously documented. For virtually every interest area, there are societies that sponsor meetings, agencies that distribute funding, journals that publish empirical and theoretical reports, edited volumes that present recent reviews of research progress, or, in many cases, handbooks or other compendia. If there is a downside to the knowledge advances and the variegated structure that has resulted, it is that there are precious few points of contact among many of the units in the structure of developmental psychology. Piece-by-piece and year-by-year, however, ambitious developmental psychologists are taking on larger strands of the story of life-span human development.

5. Conclusion

Developmental psychology is a vigorous, vast, and not altogether unified field of psychology. Far from being a handicap, however, its theoretical and methodological diversity has resulted in an immense and complex structure of assumptions, interests, and knowledge about human development. Undergirding its many manifestations is a commitment to portray and understand change, psychological change. This interest includes descriptions and explanations of developmental change, but also implementation of procedures whereby suboptimal change may be remediated or successful development may be facilitated. Thus, like many sciences, it has both basic and applied aspects. Like many sciences, it is exceedingly specialized—hopelessly so, if one wishes to capture the whole story of psychological development. Although the gathering momentum appears to move toward further variegation—and not yet toward unification—it also moves toward even greater achievements, even more knowledge. Like most sciences, developmental psychology is far from finished, but the past century has hosted an especially auspicious beginning.

See also: Adult Development, Psychology of; Change: Methods of Studying; Human Development, Successful: Psychological Conceptions; Lifespan Development, Theory of; Midlife Psychological Development; Plasticity in Human Behavior across the Lifespan

Bibliography

Baldwin J M 1895 *Mental Development in the Child and the Race*. Macmillan, New York

Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology*, 5th edn. John Wiley, New York, Vol. 1, pp. 1029–1143

Baltes P B, Staudinger U M, Lindenberger U 1999 Lifespan psychology: theory and application to intellectual functioning. *Annual Review of Psychology* **50**: 471–507

Birren J E, Schaie K W (eds.) 1996 *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA
Bornstein M H, Lamb M E (eds.) 1999 *Developmental Psychology: An Advanced Textbook* 4th edn, L. Erlbaum Associates, Mahwah, NJ
Cahan E D 1994 John Dewey and human development. In: Parke R D, Ornstein P A, Riesser J J, Zahn-Waxler C (eds.) *A Century of Developmental Psychology*, 1st edn. American Psychological Association, Washington, DC, pp. 145–67
Cairns R B 1998 The making of developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology*, 5th edn. John Wiley, New York, Vol. 1, pp. 25–105
Dixon R A, Lerner R M 1999 History and systems in developmental psychology. In: Bornstein M H, Lamb M E (eds.) *Developmental Psychology: An Advanced Textbook*, 4th edn. L. Erlbaum Associates, Mahwah, NJ, pp. 3–45
Eckardt G, Bringmann W G, Sprung L (eds.) 1985 *Contributions to a History of Developmental Psychology*. Mouton, Berlin
Elder G H Jr. 1998 The life course and human development. In: Lerner R M (ed.) *Handbook of Child Psychology*, 5th edn. John Wiley, New York, Vol. 1, pp. 939–91
Gottlieb G, Wahlsten D, Lickliter R 1998 The significance of biology for human development: a developmental psychobiological systems view. In: Lerner R M (ed.) *Handbook of Child Psychology*, 5th edn. John Wiley, New York, Vol. 1, pp. 233–73
Hall G S 1883 The contents of children's minds. *Princeton Review* **11**: 249–72
Harris D B (ed.) 1957 *The Concept of Development*. University of Minnesota Press, Minneapolis, MN
James W 1890 *The Principles of Psychology*. H. Holt, New York, Vol. 1
Lerner R M (ed.) 1998 *Handbook of Child Psychology*, 5th edn. Wiley, New York
Magnusson D, Stattin H 1998 Person-context interaction theories. In: Lerner R M (ed.) *Handbook of Child Psychology*, 5th edn. John Wiley, New York, Vol. 1, pp. 685–759
Mandelbaum M 1971 *History, Man, & Reason*. Johns Hopkins University Press, Baltimore, MD
Overton W F 1998 Developmental psychology: philosophy, concepts, and methodology. In: Lerner R M (ed.) *Handbook of Child Psychology*, 5th edn. John Wiley, New York, Vol. 1, pp. 107–88
Park R D, Ornstein P A, Rieser J J, Zahn-Waxler C (eds.) 1994 *A Century of Developmental Psychology*, 1st edn. American Psychological Association, Washington, DC
Piaget J 1983 Piaget's theory. In: Mussen P H, Kessen W (eds.) *Handbook of Child Psychology*, 4th edn. Wiley, New York, Vol. 1, pp. 103–28
Reese H W, Overton W F 1970 Models of development and theories of development. In: Goulet L R, Baltes P B (eds.) *Life-span Developmental Psychology: Research and Theory*. Academic Press, New York, pp. 115–45
Richards R J 1987 *Darwin and the Emergence of Evolutionary Theories of Mind and Behavior*. University of Chicago Press, Chicago
Schweder R A, Goodnow J, Hatano G, Le Vine R A, Markus H, Miller P 1998 The cultural psychology of development: one mind, many mentalities. In: Lerner R M (ed.) *Handbook of Child Psychology*, 5th edn. John Wiley, New York, Vol. 1, pp. 865–937
Sulloway F J 1979 *Freud: Biologist of the Mind*. Basic Books, New York
Van Geert P, Mos L (eds.) 1991 *Annals of Theoretical Psychology*. Plenum, New York, Vol. 7
Vygotsky L S 1929 The problem of the cultural development of the child. *Journal of Genetic Psychology* **36**: 415–34
Watson J B 1924 *Behaviorism*. W W Norton, New York

R. A. Dixon

Developmental Psychopathology: Child Psychology Aspects

Developmental psychopathology can be defined in a variety of ways, all having to do with development on the one hand and the resulting set of maladaptive behaviors on the other. Developmental psychopathology focuses on and integrates these two areas, looking at maladaptive processes themselves as well. Underlying much of the study of developmental psychopathology is the principle of predictability. The prediction of maladaptive behavior has been viewed not only as possible, but also as an important feature in the study of psychopathology. With this added feature, we now have a more complete definition: Developmental psychopathology is the study and prediction of maladaptive behaviors and processes across time.

As we can see, the definition of developmental psychopathology involves the merger of two fields of study, that of development and that of psychopathology (see *Developmental Psychology*). Although there is much interest in the etiology of psychopathology, until recently, there has been little research on the development of maladjustment. In part, this was due to the now discredited idea that some adult psychopathologies could not occur in childhood and to the misconception that the adult forms of pathology are the same as those for children. The introduction of a developmental perspective allows for the understanding that the same underlying psychopathology in children and in adults may be seen in very different behaviors. Thus, a developmental perspective is necessary in order to understand psychopathology.

1. *Models of Developmental Psychopathology*

Three models, a *trait model*, a *contextual* or *environmental model*, and an *interactional model* reflect the various views in regard to development and, therefore, to the etiology of psychopathology. Because attachment theory remains central to normal and maladaptive development, it is used as an example in our discussion (see *Attachment Theory: Psychological*). Unfortunately, by describing sharp distinctions, we may draw too tight an image and, as such, may make them caricatures. Nevertheless, it is important to

consider them in this fashion in order to observe their strengths and weaknesses and how they are related to developmental psychopathology.

1.1 Trait or Status Model

The trait or status model is characterized by its simplicity, and holds to the view that a trait of the child at one point in time is likely to predict a trait at a later point in time. This model is not interactive and does not provide for the effects of the environment. In fact, in the most extreme form, the environment is thought to play no role either in effecting its display or in transforming its characteristics. A particular trait may interact with the environment, but the trait is not changed by that interaction.

Traits are not easily open to transformation and can be processes, coping skills, attributes, or tendencies to respond in certain ways. Traits can be innate features, such as temperament or particular genetic codes, or can be acquired through learning or through interactive processes with the environment. However, once a trait is acquired, it remains relatively unaffected by subsequent interactions. The trait model is most useful in many instances, for example, when considering potential genetic or biological causes of subsequent psychopathology. A child who is born with a certain gene or a set of genes is likely to display psychopathology at some later time. This model characterizes some of the research in the genetics of mental illness. Here, the environment, or its interaction with the genes, plays little role in the potential outcome. For example, the work on heritability of schizophrenia supports the use of such a model, as does the lack or presence of certain chemicals on depression.

In each of these cases, the trait model argues that presence of particular genes is likely to affect a particular type of psychopathology. Although a trait model is appealing in its simplicity, there are any number of problems with it; for example, not all people who possess a trait at one point in time are likely to show subsequent psychopathology or the same type of psychopathology. That all children of schizophrenic parents do not themselves become schizophrenic, or that not all monozygotic twins show concordance vis-à-vis schizophrenia, suggests that other variables need to be considered.

This model also is useful when considering traits that are not genetically or biologically based, but based on very early experiences. For example, the attachment model as proposed by John Bowlby and Mary Ainsworth holds that the child's early relationship with its mother, in the first year of life, creates a trait (secure or insecure attachment) which will determine the child's adjustment throughout his or her life. The security of attachment that the child shows at the end of the first year of life is the result of the early interaction between the mother and the child. Once the attachment is established, it acts as a trait affecting the child's subsequent behavior.

Figure 1
Trait model.

Figure 1 presents the trait model. Notice that the interaction of the mother and child at T_1 produces the trait, C_{t1}; in the attachment case, a secure or an insecure child. Although attachment is the consequence of an interaction, once established, it is the trait (C_{t1}) residing in the child that leads to C_{t2}. There is no need to posit a role of the environment except as it initially produces the attachment.

The trait model is widely used in the study of developmental psychopathology; for example, the concept of the vulnerable or invulnerable child. A vulnerable child is defined as one who has some characteristic (trait), acquired early, which makes them vulnerable, even if they show no ongoing problems. Once stress appears, this trait is expressed. As Alan Sroufe has said, 'even when the child changes rather markedly, the shadows underlying that of the earlier adaptation remains [the insecure attachment] and in times of stress, the prototype itself may be clear (Sroufe 1983, p. 74).

On the other hand, the invulnerable child, even when stressed, is less likely to show pathology. Such a model of invulnerability has been used by Norman Garmezy and Michael Rutter (1983) to explain why children raised in highly stressful environments do not all show psychopathology. In the attachment literature, a secure child is more likely to be invulnerable to subsequent stress, whereas an insecure child is more vulnerable to stress.

The trait notion leaves little room for the impact of environment on subsequent developmental growth or dysfunction. Environments play a role in children's development in the opening year of life and continue to do so throughout the life span.

1.2 The Environmental Model

The prototypic environmental model holds that exogenous factors influence development. Two of the many problems in using this model are our difficulty in defining what environments are, and the failure to consider the impact of environments throughout the life span. In fact, the strongest form of the environmental or contextual model argues for the proposition that adaptation to current environment, throughout the life course, is a major influence in our socio-

emotional life. As environments change, so too does the individual. This dynamic and changing view of environments and adaptation is in strong contrast to the earlier models of environments as forces acting on the individual, and acting on the individual only in the early years of life.

In the simplest model, behavior is primarily a function of the environmental forces acting on the organism at any point in time. In such a model, a child does behavior *x* but not behavior *y*, because behavior *x* is positively rewarded by his or her parents and *y* is punished. Notice that, in this model, environmental forces act continuously on the organism, and the behavior emitted is a direct function of this action. Much of our behavior is controlled by an indirect form of environmental pressure. For example, a young child is present when the mother scolds the older sibling for writing on the walls. The younger child, although not directly punished, learns that writing on walls will be punished. Indirect forms of reward and punishment have received little attention, but exert as much environmental influences as direct forms.

There are many other types of environmental pressures. For example, we see an advertisement for a product 'that will make other people love us.' We purchase such a product in the hopes that others will indeed love us. We can reward or punish ourselves by our knowledge of what is correct or incorrect according to the environmental standards that we learn. Poor environmental effects, such as an overdeveloped sense of standards, can lead to high shame and, through it, to other forms of psychopathology.

Because other people make up one important aspect of our environment, the work on the structures of the *social* environment is particularly relevant, and an attempt has been made to expand the numbers of potentially important people in the child's environment beyond the mother. Sociologists and some psychologists, such as Urie Bronfenbrenner, Michael Lewis, and Judy Dunn have argued that other persons, including fathers, siblings, grandparents, and peers, clearly have importance in shaping the child's life. Given these diverse features of environments and the important roles attributed to them, it is surprising that so little systematic work has gone into their study. For the most part, mothers and, to some extent, families, have received the most attention and we, therefore, use them in our examples.

The role of environments in the developmental process has been underplayed because most investigators seek to find the structure and change within the organism itself. Likewise, in the study of psychopathology, even though we recognize that environments can cause disturbance and abnormal behavior, we prefer to treat the person—to increase coping skills or to alter specific behaviors—rather than change the environment. Yet we can imagine the difficulties that are raised when we attempt to alter specific maladaptive behaviors in environments in which such behaviors are adaptive—a point well taken by Thomas Szasz.

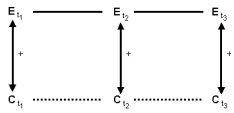

Figure 2
Model of change as a function of the environment.

Our belief that the thrust of development resides in the organism rather than in the environment, in large part, raises many problems. At cultural levels, we assume that violence (and its cure) must be met in the individual—a trait model—rather than in the structure of the environment. The murder rate using handguns in the USA is many times higher than in any other Western society. We seek responsibility in the nature of the individual (e.g., XYY males, or the genetics of antisocial behavior), when the alternative of environmental structure is available. In this case, murders may be due more to the culture's nonpunishment of persons or nonrestriction of handguns. The solution to the high murder rate in the USA might be the elimination, through punishment, of the possession of weapons. Thus, we either conclude that Americans are by nature more violent than Europeans or that other Western societies do not allow handguns and therefore have lower murder rates.

Figure 2 presents the general environmental model. The environment (*E*) at t_1, t_2, and t_3 all impact on the child's behavior at each point in time. The child's behavior at C_{t1}, C_{t2}, and C_{t3} appears consistent, and it is, as long as *E* remains consistent. In other words, the continuity in *C* is an epiphenomenon of the continuity of *E* across time. Likewise, the lack of consistency in C reflects the lack of consistency in the environment. The child's behavior changes over t_1 to t_3 as the environment produces change. Even though it appears that *C* is consistent, it is so because *E* is consistent. Consistency and change in *C* are supported by exogenous rather than by endogenous factors.

Other forms of maladaptive behavioral development have a similar problem. Depressed women are assumed to cause concurrent as well as subsequent depression in children. What is not considered is the fact that depressed mothers at t_1 are also likely to be depressed at t_2 or t_3. What role does the mother's depression at these points play in the child's subsequent condition? We can only infer the answer given the limited data available. The question that needs to be asked is: What would happen to the child if the mother was depressed at t_1 but not at t_2 or t_3? This type of question suggests that one way to observe the effect of the environment on the child's subsequent behavior is to observe those situations in which the environment changes.

The environmental model is characterized by the view that the constraints, changes, and consistencies in

children's psychopathology rest not so much with intrinsic structures located in the child as in the nature, structure, and environment of the child.

1.3 The Interactional or Transformational Model

Interactional models vary; Michael Lewis (1990) prefers to call them 'interactional', while Sameroff (1975) calls them 'transactional.' All these models have in common the role of both child and environment in determining the course of development. In these models, the nature of the environment and the traits of the child are needed to explain concurrent as well as subsequent behavior and adjustment. Maladaptive behavior may be misnamed because the behavior may be adaptive to a maladaptive environment. The stability and change in the child need to be viewed as a function of both factors and, as such, the task of any interactive model is to draw us to the study of both features. In our attachment example, the infant who is securely attached, as a function of the responsive environment in the first year, will show competence at a later age as a function of the earlier events as well as the nature of the environment at that later age.

One of the central issues of the developmental theories that are interactive in nature is the question of transformation. Two models of concurrent behavior as a function of traits and environment can be drawn. In the first, both trait and environment interact and produce a new set of behaviors. However, neither the traits nor the environment are altered by the interaction. From a developmental perspective, this is an additive model because new behaviors are derived from old behaviors and their interaction with the environment, but these new behaviors are added to the repertoire of the set of old behaviors. This model is very useful for explaining such diverse psychopathological phenomena as regression and vulnerability. In its most general form, it has been called by Richard Lerner (1986), the goodness-of-fit model.

In the second model, both trait and environment interact, producing a new set of behaviors which transform themselves. From a developmental perspective, this is a transformational model because the old behaviors give rise to new behaviors and the environment itself is altered by the exchange. The variety of interactional or transactional models is considerable.

2. Future Directions

The understanding of these different models of development directly bear on the type of research on developmental psychopathology. It seems reasonable to assume that different forms of psychopathology have different developmental pathways, some fitting one or the other of these models. The task for the future is to articulate which forms of psychopathology are best explained by what models.

See also: Attachment Theory: Psychological; Child and Adolescent Psychiatry, Principles of; Childhood and Adolescence: Developmental Assets; Childhood Sexual Abuse and Risk for Adult Psychopathology; Coping across the Lifespan; Developmental Psychology; Early Childhood: Socioemotional Risks; Infancy and Childhood: Emotional Development; Infant and Child Development, Theories of; Mental Illness, Etiology of; Mental Illness, Genetics of; Personality Theory and Psychopathology; Psychobiology of Stress and Early Child Development

Bibliography

Ainsworth M D 1973 The development of infant–mother attachment. In: Caldwell B M, Ricciuti H N (eds.) *Review of Child Development Research*. University of Chicago Press, Chicago, Vol. 3, pp. 1–95

Bowlby J 1969 *Attachment and Loss: Vol. 1, Attachment*. Basic Books, New York

Bronfenbrenner U 1979 *The Ecology of Human Development*. Harvard University Press, Cambridge, MA

Dunn J 1993 *Young Children's Close Relationships*. Sage, Newbury Park, CA

Garmezy N, Rutter M 1983 *Stress, Coping, and Development in Children*. McGraw Hill, New York

Lerner R M 1986 *Concepts and Theories of Human Development*, 2nd edn. Random House, New York

Lewis M 1987 Social development in infancy and early childhood. In: Osofsky J D (ed.) *Handbook of Infant Development*, 2nd edn. Wiley, New York, pp. 419–93

Lewis M 1990 Models of developmental psychopathology. In: Lewis M, Miller S M (eds.) *Handbook of Developmental Psychopathology*. Plenum Press, New York, pp. 15–28

Lewis M 1997 *Altering Fate: Why the Past does not Predict the Future*. Guilford Press, New York

Sameroff A 1975 Transactional models in early social relations. *Human Development* **18**: 65–79

Sroufe L A 1983 Infant–caregiver attachment and patterns of adaptation in preschool: The roots of maladaptation and competence. *Minnesota Symposia in Child Psychology* **16**: 41–83

Szasz T S 1961 *The Myth of Mental Illness*. Hoeber-Harper, New York

M. Lewis

Developmental Sciences, History of

Prior to the nineteenth century, most efforts to understand human development were not associated with scientific theories of human development (Dixon

and Lerner 1999). Instead, many models of human development were derived primarily from philosophical, literary, or theological domains. From Aristotle (384–332 BC) through St. Augustine (354–430) to William Shakespeare (1564–1616) and Jean-Jacques Rousseau (1712–78), many important thinkers wrote about the ages or stages of human life, often speculating on their unique needs and purposes. It was not until the eighteenth and the early nineteenth century, however, that these theoretical perspectives were attached to systematic empirical investigations.

The nineteenth century has often been viewed as a formative period in the emergence of scientific models of human development, for example, as in the field of developmental psychology (Baltes 1983, Cairns 1998). A key influence in the development of such models was Darwin's (1959) theory of evolution. For instance, Dixon and Lerner (1999) have argued that historical and evolutionary theories—and particularly Darwinism—provided key intellectual bases for five different theoretical systems in developmental psychology. These systems include, first, the influence of nineteenth-century evolutionary thought on the organismic model; this influence may be traced from the work of G. Stanley Hall, James Mark Baldwin, Pierre Janet, and Arnold Gesell, to the still influential work of Jean Piaget and Heinz Werner. Second, the related psychodynamic model developed in a less direct fashion from the work of Sigmund Freud and Carl Jung to that of Erik Erikson. Third, Wilhelm Preyer and Sir Francis Galton, some of the later aspects of the work of G. S. Hall and John B. Watson, contributed to the development of the mechanistic model. The dialectical model, influenced by G. W. F. Hegel and Karl Marx, was later developed more specifically by Lev S. Vygotsky and Klaus F. Riegel. Contextualism derived from the turn-of-the-century American pragmatic philosophers, most notably William James, Charles Sanders Peirce, John Dewey, and G. H. Mead. Contextualism has evolved—through integration with probabilistic epigenetic ideas (Gottlieb 1997)—to exist now as developmental systems theory (e.g., Ford and Lerner 1992, Gottlieb 1997, Thelen and Smith 1998), for example, as illustrated by developmental contextualism (Lerner 1986).

In addition to the influence of evolutionary theory, several other contributions to the scientific study of human development occurred in the nineteenth century and, as well, in the late part of the eighteenth century. Important contributors include: Dietrich Tiedemann (1748–1803), who, in 1787, produced the first psychological diary of the growth of a young child (see Murchison and Langer 1927); Friedrich August Carus (1770–1808), who endeavored to develop a comprehensive age-oriented science; Carl Gustav Carus (1789–1869), who argued for the application of the developmental method to a broad range of psychological issues (Köhler 1985); Adolphe Quetelet (1796–1874), who produced highly advanced methods to disentangle the multiple influences on the course of human development; and Johann Nikolas Tetens (1736–1807), who argued for natural scientific methods to guide the search for general laws regarding human development from birth to death (Müller-Brettel and Dixon 1990).

In addition, in the nineteenth century numerous 'baby biographies' began to appear. Such ontogenetic observational studies often took the form of diaries written by scientist-parents describing the successive achievements of a child (often their own). For example, Darwin kept such a diary of his infant son's growth during the years 1840–1 (Dixon and Lerner 1999). Many other baby biographies appeared around the turn of the century (e.g., Hall 1891, Sully 1903).

In the early decades of the twentieth century, and continuing through at least the beginning of the 1940s, much of developmental psychology was descriptive and normative. Instrumental in promoting this emphasis was the research of Arnold Gesell (1880–1961). Gesell (e.g., 1946) emphasized that maturationally based changes unfold independent of learning, and his research stressed the need for the careful and systematic cataloguing of growth norms. His work provided the field with useful information about the expected sequence for, and normative times of, the emergence of numerous physical and mental developments in selected groups of infants and children (who were from predominantly European American, middle class families). Conceptually, Gesell's work is related to one side of what has been a continuing debate in the history of developmental psychology: The nature–nurture controversy (Overton 1998). This controversy pertains to a consideration of where the sources of development lie, whether in inborn (or hereditary) mechanisms or in acquired (or learned) processes. By stressing that maturation rather than learning is the prime impetus for developmental change, Gesell (1946) was taking a 'nature', as opposed to a 'nurture', stance. Historically, other terms associated with the nature position are preformationism, nativism, and innateness; some terms associated with the nurture position are learning, conditioning, experience, and socialization.

Although Gesell's (1946) work emphasized the descriptive and normative nature of the field, there was indeed work occurring in other areas of psychology that countered Gesell's emphasis (White 1970). For example, some experimental psychologists stressed the applicability of learning principles to the study of childhood (e.g., Horowitz 1994, Miller and Dollard 1941). One consequence of this activity was to provide evidence that nurture-based learning phenomena, as opposed to nature-based maturational phenomena, could account for some features of children's behavior and development. These learning psychologists emphasized less the facts of development *per se* (e.g., 'What is the age at which an infant sits, stands,

walks, or has a two-word vocabulary?') than the explanation of those facts (e.g., 'What mechanisms—nature or nurture—need to be referred to in order to explain these facts?').

This alternating emphasis between nature and nurture was furthered in the 1940s by events leading up to and including World War II (Lerner 1986). Nazi persecution led many Jewish intellectuals to flee Europe, and many sought refuge and a new start for their careers in the United States. Many of these refugees were able to secure positions in American universities and associated institutions, despite the fact that they often brought with them ideas counter to those that were typical of the American scene (e.g., behaviorism and learning theory). For instance, although Freud himself settled in London (and died there in 1939), many psychoanalytically oriented psychologists, some trained by Freud or his daughter Anna, emigrated to North America (e.g., Peter Blos and Erik Erikson).

One reason that so many psychoanalytically oriented professionals were able to secure faculty positions in America may be related to the fact that the federal government was contributing large amounts of money to universities to support the training of clinical psychologists. Such psychologists were needed to test soldiers for both psychological and physical fitness (Misiak and Sexton 1966). This infusion of psychoanalysis in American academe resulted in the introduction of a nature-based perspective into numerous psychology departments, some of which had been dominated by nurture-based theorists (Gengerelli 1976). Nevertheless, the psychoanalytic orientation represented just one of many different theoretical accounts of human functioning—accounts that stressed either nature or both nature and nurture as sources of behavior and development—that were now making inroads into North American thinking.

For similar reasons, nativistic ideas about perception and learning, introduced by psychologists who believed in the holistic (integrative) aspects of behavior, began to appear more frequently on the North American scene. The gestalt (meaning 'form' or 'shape') views represented by some Europeans (people like Max Wertheimer, Kurt Koffka, Wolfgang Kohler, Kurt Goldstein, and Kurt Lewin) were shown to be pertinent also to areas of concern such as brain functioning, group dynamics, and social problems (Sears 1975). European perspectives relevant to human development also were being introduced. For the most part, these views also emphasized maturational (nature) components and, to some extent, clinical, nonexperimental methods. Werner (e.g., 1948) and Piaget (e.g., 1970) were especially influential.

The outcome of this cross-fertilization of ideas about development was to reinforce, if not redefine, the evolving ideas about development in North America. Any given behavior, pattern of behaviors, or systematic change in behavior could be explained by a number of different theories, and these various theories were advanced by respected advocates often working in the same academic contexts (Gengerelli 1976). Thus, in American departments of psychology, developmentalists were now confronted with a range of perspectives, ones that advanced explanations of development that varied in their reliance on purportedly nature and/or nurture processes.

Indeed, in an early review of the history of developmental science, Bronfenbrenner (1963) noted that from the 1930s to the early 1960s there was a continuing shift away from studies involving the mere collection of data, and a shift toward research concerned with abstract processes and constructs. Some books and essays published during this period epitomized this trend by calling for the study of developmental processes and mechanisms (e.g., Harris 1957). Accordingly, describing the status of the field, Bronfenbrenner (1963) wrote that 'first and foremost, the gathering of data for data's sake seems to have lost favor. The major concern in today's developmental research is clearly with inferred processes and constructs.' For example, this concern with the pluralism of ideas about developmental processes and constructs is exemplified in the now-classic publication, *The Concept of Development*, a collection of chapters pertinent to theoretical and conceptual issues in development that was edited by Harris (1957).

In turn, in a review almost a decade after Bronfenbrenner's analysis, Looft (1972) found a continuation of the trends noted by Bronfenbrenner. Looft's review, like Bronfenbrenner's, was based on an analysis of major handbooks of developmental psychology published from the 1930s to 1972. Looft noted that a shift toward more general concerns about integrating theoretical ideas occurred after World War II, and that the trend continued through 1963 (Bronfenbrenner 1963) to 1972. As a case in point we may note that the editor of the (third) 1970 edition of the *Handbook of Child Psychology*, Mussen (1970), pointed out that: 'The major contemporary empirical and theoretical emphases in the field of developmental psychology, however, seem to be on explanations of the psychological changes that occur, the mechanisms and processes accounting for growth and development.' Mussen's observation may be equally apt today, i.e., the last two decades of the twentieth century served to underscore and extend Mussen's (1970) analysis.

For example, in the fourth edition of the *Handbook of Child Psychology* (Mussen 1983), the first volume was devoted to historical, theoretical, and methodological issues, topics not even juxtaposed in a single section of previous editions. In the most recent (fifth) edition of the *Handbook of Child Psychology* (Damon 1998), the first volume featured developmental theoretical content, including attention to such enduring issues as conceptions of process, how structures function, and how functions are structured over time

(Lerner 1998). Other prominent collections, such as the *Advances in Child Development and Behavior* series (edited by Hayne W. Reese), and the *Life-span Development and Behavior* series (edited by Paul B. Baltes, Orville G. Brim, Jr., David L. Featherman, Marion Perlmutter, and Richard M. Lerner) illustrate similar trends.

The interest in explanations and the conceptual issues pertinent to them generated considerable scholarship during the last three decades of the twentieth century about the different philosophical, or meta theoretical, bases upon which individual development is studied and interpreted (e.g., Lerner 1986, 1998, Overton 1998, Reese and Overton 1970). Following the work of such philosophers as Kuhn (1970) and Pepper (1942), Reese and Overton (1970) identified two major philosophical positions—termed the organismic and the mechanistic models—that provided the basis for many extant assumptions about human development and influenced lower-order theoretical and methodological statements.

The organismic position stresses the qualitative features of developmental change and the active contribution of the organism's processes in these changes. The theories of Piaget (e.g., 1970) and Freud (e.g., 1954) are examples of such organismically oriented approaches. In contrast, the mechanistic position stresses quantitative change and the active contribution of processes lying outside the primary control of the organism (e.g., in the external stimulus environment) as the major source of development. The behavioral analysis theory of Bijou and Baer (1961) is a major example of such mechanistically oriented approaches.

Elaborating on this interest in the philosophical bases of developmental theory, Reese and Overton (1970) initiated a discussion concerning the 'family of theories' associated with models of development. For instance, although there are differences among theories in the organismic family (Freud emphasized emotional development and Piaget emphasized cognitive development), there is greater similarity among the theories within a family (e.g., the common stress on the qualitative, stage-like nature of development) than there is between theories associated with different families (e.g., mechanistically oriented theories would deny the importance, if not the reality, of qualitatively different stages of development).

Since the early 1970s, recognition of the philosophical differences between the family of theories derived from the organismic and the mechanistic models, respectively, resulted in several discussions about the different stances theories from one or another family take in regard to an array of key conceptual issues. Examples are the nature and nurture bases of development (e.g., Lerner 1986, Overton 1998); the quality, openness, and continuity of change (Brim and Kagan 1980); the extent to which theories of development may be inextricably linked to particular content domains (Lerner 1998); the extent to which the developing individual is viewed as a biological, psychological, or sociological entity versus as an integrated aspect of dynamic contexts or systems (e.g., Lerner 1986); the methods that are appropriate to the challenges of studying human development (Baltes et al. 1977); and ultimately, the alternative truth criteria for establishing the 'facts' of development (Overton 1998, Reese and Overton 1970).

From the 1970s to the end of the twentieth century, this awareness of the philosophical bases of developmental theory, method, and data contributed to a consideration of additional models appropriate to the study of psychological development. In part, this consideration developed as a consequence of interest in integrating assumptions associated with theories derived from organismic and mechanistic models (Overton 1998). For instance, Riegel (1976) attempted to apply an historical model of development that included some features of organicism (e.g., active organism) and some features of mechanism (e.g., active environment).

In turn, there was an interest in continual, reciprocal relations between an active organism and its active context (and not in either element *per se*) and with these relations as they exist on all phenomenal levels of analysis. These interests were labeled in different ways in the human development literature. For example, such terms as dialectical (Riegel 1976), transactional (Sameroff 1975), dynamic interactional (Lerner 1986), and person-environment, or holistic, interactional (Magnusson 1996, 1999) have been used. In a similar way, a number of theorists have explored the application of contextual (Rosnow and Georgoudi 1986) and integrative, developmental systems notions about human development (Ford and Lerner 1992, Gottlieb 1997, Magnusson 1996, 1999).

Overton (1998) has explained that these relational theories of human development arose at least in part because of the recognition that differences in approaches to, and interpretations of, behavior, prevalent between World War II and the 1970s, were limited due to their having been framed as Cartesian splits. In addition to the nature-nurture issue the controversies of continuity versus discontinuity, stability vs. instability, and constancy vs. change are cases in point (Lerner 1986). Overton argued that these issues remained controversial precisely because they cast development unnecessarily into a split frame of reference, into a categorical either/or form (Overton 1998). The early formulation and promulgation of such conceptual controversies had the ironic outcome of keeping the field focused on diverse interpretations of behavioral change.

However, more recently, developmental scientists have presented alternatives to such 'split' conceptions. These ideas have stressed the systemic integration, or fusion, of levels of organization ranging from the biological to the sociocultural and historical (e.g.,

Ford and Lerner 1992, Gottlieb 1997, Magnusson 1999, Thelen and Smith 1998). They have encompassed the entire life span (Baltes et al. 1998), and thus have engaged the action of individuals as an active source of their own development (Brandtstädter 1998, Lerner and Busch-Rossnagel 1981).

Indeed, the majority of the chapters in Volume 1 (Lerner 1998) of the above-noted most recent, fifth edition of the *Handbook of Child Psychology* (Damon 1998) reflect such developmental systems notions. By the end of the twentieth century these integrative notions had replaced either mechanistic and reductionistic nature (e.g., sociobiology, behavior genetics) or nurture (e.g., learning or behavior analysis) theories as the preferred frame for research and application in the scientific study of human development (Lerner 1998).

See also: Adolescence, Psychiatry of; Childhood and Adolescence: Developmental Assets; Culture in Development; Development: Social-anthropological Aspects; Development Theory in Geography; Developmental Psychology; Developmental Psychopathology: Child Psychology Aspects; Developmental Sciences, History of; Freud, Sigmund (1856–1939); Genetics and Development; Historical Change and Human Development; Human Development, Bioecological Theory of; Law and Development; Piaget, Jean (1896–1980); Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis: Overview; Psychobiology of Stress and Early Child Development; Psychological Development: Ethological and Evolutionary Approaches

Bibliography

Baltes P B 1983 Life-span developmental psychology: Observations on history and theory revisited. In: Lerner R M (ed.) *Developmental Psychology: Historical and Philosophical Perspectives*. Erlbaum, Hillsdale, NJ, pp. 79–112

Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 1029–44

Baltes P B, Reese H W, Nesselroade J R 1977 *Life-Span Developmental Psychology: Introduction to Research Methods*. Brooks/Cole, Monterey, CA

Bijou S W, Baer D M (eds.) 1961 *Child Development*. Appleton-Century-Crofts, New York

Brandtstädter J 1998 Action perspectives on human development. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 807–63

Brim O G, Kagan J (eds.) 1980 *Constancy and Change in Human Development*. Harvard University Press, Cambridge, MA

Bronfenbrenner U 1963 Developmental theory in transition. In: Stevenson H W (ed.) *Child Psychology*. University of Chicago Press, Chicago, Part 1, pp. 517–42

Cairns R B 1998 The making of developmental psychology. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 419–48

Damon W (ed.) 1998 *The Handbook of Child Psychology*, 5th edn. Wiley, New York

Darwin C 1959 *On the Origin of Species*. John Murray, London

Dixon R A, Lerner R M 1999 History and systems in developmental psychology. In: Bornstein M H, Lamb M E (eds.) *Developmental Psychology: An Advanced Textbook*, 4th edn. Erlbaum, Mahwah, NJ, pp. 3–45

Ford D H, Lerner R M 1992 *Developmental Systems Theory: An Integrative Approach*. Sage, Newbury Park, CA

Freud S 1954 *Collected Works*. Hogarth, London

Gengerelli J A 1976 Graduate school reminiscence: Hull and Koffka. *American Psychologist* **31**: 685–8

Gesell A L 1946 The ontogenesis of infant behavior. In: Carmichael L *Manual of Child Psychology*, 295–331

Gottlieb G 1997 *Synthesizing Nature–Nurture: Prenatal Roots of Instinctive Behavior*. Erlbaum, Mahwah, NJ

Hall G S 1891 Notes on the study of infants. *The Pedagogical Seminary* **1**: 127–38

Harris D B (ed.) 1957 *The Concept of Development*. University of Minnesota Press, Minneapolis, MN

Horowitz F D 1994 John B. Watson's legacy: Learning and environment. In: Parke R D, Ornstein P A, Reiser J J, Zahn-Waxler C (eds.) *A Century of Developmental Psychology*, 1st edn. American Psychological Association, Washington, DC, pp. 233–50

Köhler U 1985 The concept of development and the genetic method of C. G. Carus. In: Eckhardt J, Bringmann W G, Sprung L (eds.) *Contributions to a History of Developmental Psychology*. Mouton, Berlin, pp. 117–26

Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

Lerner R M 1986 *Concepts and Theories of Human Development*, 2nd edn. Random House, New York

Lerner R M 1998 *The Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1

Lerner R M, Busch-Rossnagel N A (eds.) 1981 *Individuals as Producers of their Development: A Life-span Perspective*. Academic Press, New York

Looft W R 1972 The evolution of developmental psychology: A comparison of handbooks. *Human Development* **15**: 187–201

Magnusson D (ed.) 1996 *The Life-span Development of Individuals: Behavioral, Neurobiological, and Psychosocial Perspectives: A Synthesis*. Cambridge University Press, Cambridge, UK

Magnusson D 1999 Holistic interactionism: A perspective for research on personality development. In: Pervin L A, John O P (eds.) *Handbook of Personality: Theory and Research*, 2nd edn. Guilford Press, New York, pp. 219–47

Miller N E, Dollard J 1941 *Social Learning and Imitation*. Yale University Press, New Haven, CT

Misiak H, Sexton V S 1966 *History of Psychology: An Overview*. Grune & Stratton, New York

Müller-Brettel M, Dixon R A 1990 Johann Nicolas Tetens: A forgotten father of developmental psychology? *International Journal of Behavioral Development* **13**: 215–30

Murchison C, Langer S 1927 Tied observation on the development of the mental faculties of children. *Journal of Genetic Psychology* **34**: 205–30

Mussen P H (ed.) 1970 *Carmichael's Manual of Child Psychology*, 3rd edn. Wiley, New York

Mussen P H (ed.) 1983 *Handbook of Child Psychology*, 4th edn. Wiley, New York
Overton W 1998 Developmental psychology: Philosophy, concepts, and methodology. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 107–87
Pepper S C 1942 *World Hypotheses*. University of California, Berkeley, CA
Piaget J 1970 Piaget's theory. In: Mussen P H (ed.) *Carmichael's Manual of Child Psychology*, 3rd edn. Wiley, New York, Vol. 1, pp. 703–23
Reese H W, Overton W F 1970 Models of development and theories of development. In: Goulet L R, Baltes P B (eds.) *Life-span Developmental Psychology: Research and Theory*. Academic Press, New York, pp. 115–45
Riegel K F 1976 The dialectics of human development. *American Psychologist* **31**: 689–700
Rosnow R L, Georgoudi M (eds.) 1986 *Contextualism and Understanding in Behavioral Science*. Praeger, New York
Sameroff A J 1975 Transactional models in early social relations. *Human Development* **18**: 65–79
Sears R R 1975 Your ancients revisited: A history of child development. In: Hetherington E M (ed.) *Review of Child Development Research*. University of Chicago Press, Chicago, Vol. 6, pp. 1–73
Sully J 1903 *Studies of Childhood*. Appleton, New York
Thelen E, Smith L B 1998 Dynamic systems theories. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. I, pp. 563–633
Werner H 1948 *Comparative Psychology of Mental Development*. International Universities Press, New York
White S H 1970 The learning theory tradition and child psychology. In: Mussen P H (ed.) *Carmichael's Manual of Child Psychology*, 3rd edn. Wiley, New York, pp. 657–702

R. M. Lerner

Developmental Sport Psychology

Weiss and Bredemeier (1983) coined the term 'developmental sport psychology' to advocate an appropriate theoretical perspective for studying children's psychosocial experiences through sport. According to Baltes et al. (1977), a developmental theoretical orientation seeks to explain psychosocial and behavioral changes within individuals across the lifespan, as well as differences and similarities in the nature of changes among age groups. Weiss and Bredemeier conducted a content analysis of youth sport research from 1970–81, and found that less than 10 percent of empirical studies employed designs or age selection criteria that considered developmental factors. They strongly encouraged researchers to adopt a developmental perspective to enhance understanding of children's social and psychological outcomes of sport participation.

Adopting a developmental approach means employing designs and methodology that capture the essence of age-related differences in cognitive, affective, or behavioral variables. Weiss and Bredemeier suggested three means by which developmental studies could be designed: (a) select ages of participants based on cognitive or physical developmental criteria; (b) compare age groups at key periods of development; and (c) follow individuals longitudinally on constructs of interest. One could also assess developmental level directly (e.g., cognitive age) and compare individuals who vary in developmental level on the target construct. Since the publication of Weiss and Bredemeier's work (1983), several topics in sport psychology have been investigated using a developmental perspective. The topics focused upon here are observational learning, perceived competence, social influences, and moral development.

1. *Observational Learning*

Observational learning or modeling is a powerful means of transmitting information about attitudes, values, and behaviors. It is not surprising that visual demonstrations are salient for teaching motor skills and influencing psychological responses in youth sport. Because Bandura (1986) specified that attention, retention, production, and motivation processes must be considered for successful modeling, this area is susceptible to a developmental approach (see *Social Cognitive Theory and Clinical Psychology*). Yando et al.'s (1978) developmental theory of modeling emphasized the observer's cognitive-developmental level and motivational system in observational learning of physical and social skills. Cognitive-developmental level includes attention span, memory capacity, rehearsal skills, and physical abilities. Motivational system refers to the intrinsic and extrinsic motives guiding children's attempts to reproduce observed behaviors, such as desire to develop and demonstrate competence, and social and tangible reinforcement. A synthesis of comprehensive reviews of children's observational learning of motor and psychological skills follows (McCullagh and Weiss 2001).

1.1 *Modeling and Motor Performance*

Early research on children's modeling of motor skills adopted a nondevelopmental approach, and focused upon topics such as temporal spacing of demonstrations and model characteristics. Research since the early 1980s employed designs where children were selected based on cognitive-developmental criteria, such as attention and rehearsal capabilities. Results show that a 'show-and-tell model,' one who imparts visual information while verbalizing the task to be completed, was especially helpful for younger children's (8 years or less) successful imitation of motor skills, while older children equally benefited from a visual model or show-and-tell model. These differences were attributed to older children's mature capability for selectively attending to salient task parts and spontaneously engaging in symbolic coding and verbal

rehearsal strategies. Some researchers used kinematic analysis to show that children increasingly match performance form to that demonstrated by the model as a function of observations and practice, with more dramatic increases following trials in which verbal labels are attached to visual stimuli.

1.2 Psychological Effects of Modeling

Peer models maximize similarity between the observer and model, and thus may especially motivate changes in children's psychological responses to activities they perceive as frightening or difficult. Weiss et al. (1998) examined the effect of peer mastery and coping models with children (ages 5–8) who were fearful of swimming. Peer coping and mastery models produced better performance, higher self-efficacy, and lower fear than did control participants. Peer coping models had a stronger effect on self-efficacy than peer mastery models (see *Self-efficacy*). Findings imply that using a peer model along with swim lessons is an effective means of instructing children who are fearful of the water.

Developmental differences in children's use of social comparison are an important consideration in peer modeling of motor and psychological skills. Butler (1989) contends that younger children (less than 8 years) primarily use peers for *observational learning*, while older children use peers for *evaluating one's ability*. It is conceivable that children in the Weiss et al. study primarily used peer models as sources of information for how to perform the skills. Older children who emphasize social comparison over observational learning may respond differently to peer models who demonstrate skills of which they are afraid or see as difficult.

2. Perceptions of Physical Competence

Perceptions of competence refer to evaluation of one's abilities in a particular achievement domain (e.g., athletics) or subdomain (e.g., swimming). Perceived physical competence is a strong correlate of cognitions (e.g., perceived control), affects (e.g., enjoyment), and motivated behaviors. Within Harter's (1978) competence motivation theory, perceived competence is a developmental construct, in that level, accuracy, and sources of perceived competence undergo maturational change. Understanding such changes is important for explaining age-related variations in children's thoughts, feelings, and actions in the physical domain (Weiss and Ferrer-Caja in press).

2.1 Level and Accuracy of Perceived Competence

Research in the academic domain shows an age-related decline in level of perceived competence across middle and late childhood, along with a concomitant increase in the accuracy with which children judge their abilities. Trends in level of perceived physical competence have been mixed, with studies showing a decline, little change, or an increase. In contrast, accuracy of perceived physical competence shows a reliable increase with age. Developmental change in level and accuracy of perceived competence has been attributed to shifts in information sources used to judge one's self-abilities, such as social (e.g., peer comparison), outcome (e.g., performance statistics), and internal (e.g., goal achievement) factors.

2.2 Sources of Perceived Competence

Consistent age-related trends are found for sources of perceived physical competence. Younger children (under age 10) cite adult feedback, game outcome, and sport attraction more frequently than older children and adolescents (ages 10–15), who identify peer evaluation and comparison as more salient sources of judging physical ability. With internalization of a self-reward system and mastery goals, older teenagers (ages 16–18) cite self-referenced criteria as more important than do younger peers, including skill improvement, goal achievement, enjoyment of sport, and internal information (confidence, nervousness). Gender differences in sources of perceived physical competence do not emerge until adolescence (ages 14–18). Females cite social and internal sources as more important than do male cohorts, who prioritize competitive outcomes and speed/ease of learning skills as physical competence sources.

3. Social Influences

Significant adults (parents, coaches) and peers (teammates, nonplaying friends) are central to children's psychological development through sport participation. While parents are primarily responsible for socializing their children into sport at an early age, the child's peer group increases in importance in later childhood and adolescence. Coaches occupy an important role in positively or negatively affecting children's and adolescents' sport experiences, and consequently decisions about whether to continue or drop out from athletic involvement. A rich literature exists on parents' and coaches' influence on youth involved in sport, with peer influence an unfolding area in recent years (Weiss and Ferrer-Caja in press).

3.1 Parents

Parents influence their children's self-perceptions, emotions, and motivation in sport through expressed beliefs, values, expectancies, and behaviors. For example, parents' reported enjoyment of physical ac-

tivity and the degree to which they encourage their child to participate in sport are associated with children's perceived physical competence and attraction toward physical activity. Several studies show that children's perceptions of parents' beliefs about their physical competence are associated with their own self-beliefs, affective responses, and physical activity levels. Still other studies have shown linkages between children's perceptions of parents' beliefs and behaviors with their achievement goal orientations, global self-esteem, and intrinsic motivation.

3.2 Coaches

Several studies implicate the significance of coaches to psychosocial development of children and adolescents. Smith and Smoll (1996) showed that baseball coaches who emphasized praise after successes, encouragement and instruction following errors, and refrained from punishing remarks, were associated with young athletes with more positive profiles. Their athletes reported higher baseball ability, global self-esteem, team cohesion, and enjoyment, as well as lower performance anxiety and dropout rates. These results held regardless of win–loss records.

Coaches' feedback and reinforcement patterns strongly influence adolescent athletes' perceived competence, enjoyment, and motivation. While most results point to a positive relationship between frequency of praise and self-perceptions, emotions, and motivation, Horn (1985) found that greater praise was associated with lower perceived competence and greater criticism with higher perceived competence in adolescent softball players. Horn's analysis showed that coaches' praise was not given contingent to the level of skill demonstrated but rather in a more general way. In contrast, criticism was given in response to specific skill errors, and followed by information on how to improve on subsequent efforts. To bolster this explanation, Horn found that lower skilled players received more praise, while talented players received more criticism and instruction for their efforts. Thus the quality rather than the quantity of praise, instruction, and criticism is a crucial consideration for understanding adult influence on young athletes' psychological development through sport.

3.3 Peers

Sources of physical competence information revealed peer comparison and evaluation to be especially important from later childhood to middle adolescence. Both peer acceptance (being liked by one's classmates or teammates) and close friendship (one-on-one relationship denoted by supportive qualities) are strongly linked to self-beliefs, emotions, motivation, and physical activity behavior during these developmental periods (Weiss and Ferrer-Caja in press). Even as early as middle childhood, children identify being good at sports as an important quality for being popular with one's peers, especially among boys.

Greater attention has been directed toward the developmental importance of children's close sport friendships. Weiss et al. (1996) interviewed children 8 to16 years of age, and uncovered several positive and few negative dimensions of close sport friendships. Younger children (ages 8–12) identified prosocial behavior and physical characteristics as attractive friendship qualities more often than older youth (ages 13–16), who spent more time talking about intimate self-disclosure and positive personality characteristics. Other research has shown that younger children cite companionship and pleasant play more frequently, while older youth rate loyalty, intimacy, similar beliefs and values, and conflict as reflective of sport friendships. Females rated self-esteem enhancement and supportiveness, similar beliefs and values, and loyalty and intimacy higher than males, who rated conflict as occurring more frequently. Close sport friendships and specific friendship qualities (esteem enhancement, companionship) are associated with higher physical self-worth, positive affect, intrinsic motivation, commitment toward sport participation, and physical activity levels.

4. Moral Development

Although we are acquainted with the adage 'sport builds character,' popular media provide more convincing evidence of a negative association between sport participation and ethical behaviors. Widely accepted moral development theories are developmental in nature (e.g., Piaget, Kohlberg), suggesting that individuals proceed through stages of moral reasoning that reflect egocentric, societal, and principled levels (Shields and Bredemeier 1995). Moral reasoning and behaviors should be ultimately linked. Social learning theories suggest that significant others, in particular, exert a strong influence on young athletes' beliefs and behaviors concerning legitimate sport behavior. Structural-developmental theories, in contrast, emphasize the interaction between individual differences in moral reasoning level and the social context in which moral judgments and behaviors are enacted.

4.1 Social Approval and Aggression in Sport

Significant adults and peers model and reinforce beliefs and behaviors that impact upon young athletes' beliefs about the fairness of, and actual use of, aggressive game tactics. For example, adolescent players who selected more violent professional role models received more penalty minutes during games

than did those who chose less violent models. Ice hockey players' perceptions of significant others' attitudes toward illegal violence indicated greater disapproval toward fist-fighting for younger athletes (ages 12–15 years), but approval ratings dramatically shifted for 16–21 year old players. For all age groups, rankings were consistent for significant others' approval of violence: teammates, spectators, self, coach, father, and mother.

Interviews have tapped athletes' perceptions of significant others' approval of legal and illegal violence. Nonplaying peers were viewed as exalting legal and illegal violence during hockey play, with teammates favorable to a lesser degree. Athletes perceived that coaches and fathers approved of legal but hard body-checking, but disapproved of violence prone to penalties. The relation between perceptions of social approval for unsportsmanlike play and moral beliefs and behaviors were examined among fourth–fifth and seventh–eighth grade basketball players. For the younger group, approval of antisocial behaviors by parents, coach, and teammates was associated with lower moral judgment and intention. For older players, perceived approval of unsportsmanlike behaviors by teammates, coach, and parents was associated with lower moral judgment and intent, as well as lower moral reasoning and prosocial behaviors. Mothers for younger players and teammates for older players emerged as the strongest influences on moral variables, lending support for developmental differences in sources of greatest influence.

4.2 Moral Reasoning and Aggression Tendencies

Shields and Bredemeier (1995) established a line of research on moral reasoning and aggression in sport. First, college athletes score lower on daily life and sport moral reasoning than nonathletes. Second, youth (age 12 and older), high school, and college sport participants reason at lower levels for sport than for daily life moral dilemmas. Third, divergence between daily life and sport moral reasoning scores emerge at age 12 and widen as age/competitive level increase. Finally, lower moral reasoning about sport dilemmas is associated with greater legitimacy of and tendencies toward aggressive play. This latter relationship especially holds for male participants at all levels.

Shields and Bredemeier coined the term *game reasoning* to explain the findings of divergence in reasoning about daily life and sport moral dilemmas, athlete–nonathlete moral reasoning differences, and correlations between moral reasoning and legitimacy and expression of aggression. Game reasoning reflects one's viewpoint of sport as set apart emotionally from the broader morality of everyday life, and involves a moral transformation during athletic contests in which an egocentric perspective is considered to be legitimate in the pursuit of winning.

5. Summary and Future Research Directions

Since Weiss and Bredemeier's (1983) paper on the need to adopt a developmental theoretical perspective for studying children in sport, several lines of research were identified. These included observational learning, perceived competence, social influences, and moral development. Still, there is a long way to go. A recent content analysis showed that about 20 percent of more than 600 studies and reviews conducted in 1982–97 on youth sport psychology used a developmental approach. Despite the explosion of studies on children in sport, generally, and the doubled amount of developmental studies, theoretical approaches have still been largely nondevelopmental in nature.

The established lines of research reported here, as well as others, could serve as templates for researchers and educators in the field. Hopefully such efforts will inspire others to follow suit with a series of interrelated investigations that consider developmental factors in psychosocial phenomena, ones that include both between age-group as well as within age-group differences.

If we go back to Baltes et al.'s (1977) definition of a developmental theoretical perspective, the term 'lifespan' was used. While the term 'developmental' has typically referred to infancy through adolescence, the more contemporary view is to consider a lifespan developmental perspective. Research on middle and older adult psychological development through physical activity pales in comparison to youth sport research, and scholars are encouraged to broaden the age groups of their investigations to truly understand 'developmental' sport and exercise psychology.

See also: Sports Performance, Age Differences in; Youth Sports, Psychology of

Bibliography

Baltes P B, Reese H W, Nesselroade J R 1977 *Life-span Developmental Psychology: Introduction to Research Methods*. Brooks/Cole, Monterey, CA

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ

Butler R 1989 Mastery versus ability appraisal: A developmental study of children's observations of peers' work. *Child Development* **60**: 1350–61

Harter S 1978 Effectance motivation reconsidered. *Human Development* **21**: 34–64

Horn T S 1985 Coaches' feedback and changes in children's perceptions of their physical competence. *Journal of Educational Psychology* **77**: 174–86

McCullagh P, Weiss M R 2001 Modeling: Considerations for motor skill performance and psychological responses. In: Singer R N, Hausenblas H A, Janelle C M (eds.) *Handbook of Research on Sport Psychology*. Wiley, New York

Shields D L, Bredemeier B J 1995 *Character Development and Physical Activity*. Human Kinetics, Champaign, IL

Smith R E, Smoll F L 1996 The coach as a focus of research and

intervention in youth sports. In: Smoll F L, Smith R E (eds.) *Children and Youth in Sport: A Biopsychosocial Perspective*. Brown and Benchmark, Madison, WI, pp. 125–41
Weiss M R, Bredemeier B J 1983 Developmental sport psychology: A theoretical perspective for studying children in sport. *Journal of Sport Psychology* **5**: 216–30
Weiss M R Ferrer-Caja E in press Motivational orientations and sport behavior. In: Horn T S (ed.) *Advances in Sport Psychology*, 2nd edn. Human Kinetics, Champaign, IL
Weiss M R, McCullagh P, Smith A L, Berlant A R 1998 Observational learning and the fearful child: Influence of peer models on swimming skill performance and psychological responses. *Research Quarterly for Exercise and Sport* **69**: 380–94
Weiss M R, Smith A L, Theeboom M 1996 'That's what friends are for': Children's and teenagers' perceptions of peer relationships in the sport domain. *Journal of Sport and Exercise Psychology* **18**: 347–79
Yando R, Seitz V, Zigler E 1978 *Imitation: A Developmental Perspective*. Erlbaum, Hillsdale, NJ

M. R. Weiss

Dewey, John (1859–1952)

1. Biography

John Dewey was born in Burlington, Vermont, on October 20, 1859. By the time of his death (June 1, 1952 in New York City), he was regarded by many as America's foremost philosopher. During his boyhood and youth in Vermont, Dewey had been an average pupil. In 1864, Dewey's mother moved with her four boys to Northern Virginia to be close to her husband, a quartermaster with the Union army during the civil war. They did not return to Vermont before 1867, and John as a young boy saw the carnage that war brings.

As a student at the University of Vermont in Burlington, he found a mentor, H. A. P. Torrey, who nourished his interest in philosophy. After graduation in 1879, Dewey spent 2 years teaching at the high school at South Oil City in Pennsylvania, and for one year taught at an elementary school in a small town in Vermont. He continued studying philosophy in private tutorial, and in 1882, enrolled at Johns Hopkins University in Baltimore. There, he gravitated quickly to G. S. Morris, a neo-Hegelian. Hegel's ideas had a liberating effect on Dewey's thought. The religious teachings that he was brought up with were transformed into the Hegelian view of cultural institutions as agents of an objective mind that through them shapes the mental life of individuals. Morris let Dewey teach his undergraduate class, secured a fellowship for him, and offered him a position as instructor at the University of Michigan at Ann Arbor where he had moved in 1884, the year of Dewey's Ph.D. exam. During the following decade, Dewey stayed at the University of Michigan (except for 1888, the year spent at the University of Minnesota), where he was promoted to full professor, and wrote articles on psychology and ethics. In 1886, he married Alice Chipman, a strong personality with a secular humanist background. A formative influence—next to Morris' and to that of William James that mainly came by correspondence—was his friendship with G. H. Mead.

The exchange began in Ann Arbor and continued during the decade in Chicago, lasting until Mead's death in 1931. Mead's scholarship in physiology surpassed that of Dewey, and Mead's concept of psychology—based on a view of the organism that is interacting with its environment—was integrated by Dewey into his own transactional theory.

He accepted the offer to head the department of Philosophy, Psychology, and Pedagogy at the University of Chicago in 1894. The city of Chicago, teeming with European immigrants, full of destitute people as well as of people of means and goodwill, appeared to Dewey like one big social experiment. He worked closely with Jane Addams of Hull House, a social settlement; with Ella Flagg Young, District Superintendent of Chicago Schools; and with Francis Parker, principal of the Cook County Teachers' Training School. In 1896, at the request of parents who were willing to pay for most of the school's maintenance, he founded the Laboratory School that quickly became known as the Dewey School. The school's curriculum centered on 'occupations' that engaged pupils' minds and hands in activities such as sewing, weaving, cooking, building, growing corn, etc. A debate society complemented the syllabus for fifth- and sixth-graders. What made the school unique was not so much the emphasis upon the priority of the manual to the conceptual as the attempt to realize the idea of school as a community, in which both teachers and students were stakeholders.

Dewey's book *School and Society* (1899)—probably his most widely read book, with translations into a dozen languages—consists of talks he gave to raise money for the Laboratory School. In 1897, he had published a brief text, *My Pedagogic Creed*, that not only reflects his idea of the Laboratory School but holds much of his later more explicit writing on education in a nutshell.

Dewey's time in Chicago came to an abrupt end in 1904 when the University's president merged the Laboratory School with the Chicago Institute for Teacher Training, dismissing Dewey's school's teachers and his wife, who acted as principal, without prior notice or consultation with Dewey. He resigned immediately and found the position as professor of philosophy at New York's Columbia University that he would hold from 1904 through his retirement in 1930, until his final resignation of all duties in 1939, at age 80.

Courses he had taught in Chicago were published in New York (*How we Think*, 1910; *Human Nature and*

Conduct, 1922), and his major work on philosophy of education, *Democracy and Education* (1916) is a fruit of Dewey's involvement with schools and philosophy of education in Chicago.

In New York, Dewey was engaged with philosophy at Columbia and education at Teachers' College, and quickly he became involved in various political activities, not only as president of the American Association of University Professors and functions in the American Federation of Teachers, but also as leading contributor to *The Republic*, a magazine promoting liberal views, and as president of the People's Lobby and Chairman of the League for Independent Political Action. His travels and consultant activities in Japan (1919), China (1919–1921), Turkey (1924), and Mexico (1926) extended his influence, and a series of lectures on a variety of topics resulted in major publications (Tokyo University 1919: *Reconstruction in Philosophy*; Carus lectures at Columbia 1923: *Experience and Nature*; Gifford lectures in Edinburgh 1929: *The Quest for Certainty*; Terry lectures at Yale 1934: *A Common Faith*).

John and Alice Dewey had six children, three boys, one of whom had died in 1895 on a sojourn in Italy, and three girls. In 1904, on their second trip to Europe, another boy died, and even though the Deweys adopted later, on that same tour, an Italian orphan who was eight years old like their lost son, they never came to accept these deaths.

Alice died after long illness in 1927; John Dewey was married again in 1946, at age 87, to Roberta Grant, and the couple adopted two children. In 1949, when he was 90 years old, Dewey published a book that he had written together with Arthur Bentley, *Knowing and the Known*, a study in epistemology.

2. *Experience the Connecting Agent of Dewey's Life and Work*

One of the more unusual features for a philosopher's biography is the close connection between Dewey's intellectual and actual life.

In politics, he tried his hand at founding a third party, at editing a newspaper devoted to 'thought news,' and at heading the Outlawry of War Movement ('That war be outlawed and declared a crime'). In his fifties and sixties, he traveled widely not only to spread his ideas, but also to help install systems of education in Turkey, Mexico, the Soviet Union; his sojourn in China lasted over two years. At age 78, he chaired the committee to investigate Leon Trotsky (an investigation at Trotsky's own request, in Mexico, into accusations by Stalin's courts in Moscow).

Looking back at 80, he wrote: 'I reached fairly early in the growth of my ideas a belief in the intimate and indissoluble connection of means used and ends reached. I doubt if the force of the idea in the theory of social action would have come home to me without my experience in social and political movements, culminating in events associated with my membership in the Trotsky Inquiry Commission' (Jane Dewey 1939).

As for education, having been a schoolteacher himself and having run the Laboratory School in Chicago, he points to the interaction between thoughts and actions: 'The idea that lay back of my educational undertaking was a rather abstract one of the relation of knowledge and action. My school work translated this into a much more vital form' (Jane Dewey 1939).

In the field of aesthetics and the fine arts, he had worked with Albert C. Barnes, the millionaire collector who befriended him and introduced him to the art of French postimpressionism; Dewey dedicated his book *Art as Experience* (1934) to him.

As for metaphysics, his concept of unity of the ideal and the actual that he named 'God' (*A Common Faith*), appears to have been backed by personal experience. During his sojourn as a young man in South Oil City, he had experienced a mystic oneness with the universe, and he would later declare: 'I've never had any doubts since then, nor any beliefs. To me faith means not worrying … ' (Dykhuizen 1973).

His first years as a young instructor at the University of Michigan in Ann Arbor had impressed on him the actual attainability of a cooperative community and the experience of personal growth within a democratic atmosphere, with encouragement of freedom and individual responsibility, and the immediate acceptance as an adult responsible member of the faculty. This experience may well have contributed to the formulation of his later visionary concept, presented in *The Public and Its Problems* (1927), of democracy as 'the great community.'

He himself summed up the mutual enforcement of idea and action throughout his experience in the autobiographical notes that he added to his daughter Jane's account of his life: 'My belief in the office of intelligence as a continuously reconstructive agency is at least a faithful report of my own life and experience' (Jane Dewey1939).

3. *Contribution*

The upshot of Dewey's work can be described as the subordination of philosophy to life, a contribution to the project of intellectualization of practice. Since intelligence—the assessment of the possibilities inherent in a given situation—is a tool for the solution of practical problems, Dewey referred to his brand of pragmatism as 'instrumentalism': like the wings of a bird and the fins of a fish, human intelligence is an instrument for survival, but unlike the other creatures' physical features, the human mind is capable through the powers of community to promote life's quality

itself, not unlike a process of evolution directed by intelligent action.

One of the most obvious principles in Dewey's philosophy is his exposure of dualisms as artificial alternatives that have become barriers to the kind of intelligence needed to promote the process of experience, e.g., the separation between science and morals appeared scandalous to him, a roadblock on the path to progressive application of intelligence. The method of scientific inquiry developed and employed successfully in modern times provides a model for inquiry into social affairs to the degree that it is an application of reason against that of myth; analogously, intelligence must be applied in the social realm to develop a methodology for the resolution of the problems of men and to facilitate social reconstruction.

His dictum, that philosophers should turn away from the problems of philosophy and instead should turn to the problems of men, is another way of saying that traditional philosophy is concerned with questions that result from views framed in terms of alternatives which do not relate. As he states in an essay, *The Influence of Darwinism on Philosophy* (1909): 'But in fact intellectual progress usually occurs through sheer abandonment of questions together with both of the alternatives they assume—an abandonment that results from their decreasing vitality and a change of urgent interest. We do not solve them: we get over them. Old questions are solved by disappearing, evaporating, while new questions corresponding to the new attitude of endeavor and preference take their place' (Middle Works 1973–1983).

With regard to politics, Dewey maintained that the majority of men and women could and should become active participants in the decision-making processes that govern and shape their lives. He urged the development of a society that facilitates individual growth through participation in community life. Integration of self-realization and the social to form a common good is the major ethical postulate entwined with such an instrumentalization of the political. In *The Public and Its Problems* (1927), Dewey answered an analysis of the political situation in the USA in the years after World War II by Walter Lippmann, *The Phantom Public* (1925). Lippmann, a democratic realist, had shown that self-government and the idea of participatory democracy fell short of the reality of a population largely manipulated by propaganda, and he proposed the establishment of an elite ruling class to identify and serve the public interest. Dewey, while accepting the factuality of Lippman's analysis, did not subscribe to his proposal, but suggested instead the continued pursuit of the idea of a great community, especially through the means of education. His idea of the fullest possible liberation of the potentials of human experience allowed for no other social framework than that of participatory democracy. And for Dewey, the school as the 'church of democracy' is to provide the actual experience for achieving the 'great community' through its model as 'embryonic society.'

Probably, Dewey is known best for his contributions to educational thought. At the core of his philosophy of education is the concept of experience as an ongoing process with no externally identifiable aim. '… education is a constant reorganizing or reconstructing of experience. It has all the time an immediate end, and so far as activity is educative, it reaches that end—the direct transformation of the quality of experience' (*Democracy and Education* 1916, Middle Works 1973–1983). This concept was opposed to teleological concepts of education dominant in countries with an idealistically inclined tradition such as Germany or Japan, and at the same time provided the foundation for an interchange of participatory democracy and education as its exercise in embryonic form.

In *Experience and Education* (1938), Dewey emphasizes the need for a theory of education. Such theory is not a system or an application of static principles. It is entirely subservient to practice and should, in Dewey's words, be 'a light to the eyes and a lamp for the feet'; it might be described as a framework to facilitate a plan of action, or as a set of principles to interpret an experience's educational function.

Dewey elaborates two such principles of experience, that of continuity and that of interaction. He reminds educators to avoid the opposite pitfalls of overindulging children and of subjecting them entirely to external demands. Proponents of progressive education who promoted a child-centered approach had claimed Dewey as provider of the philosophical foundation for their sentiment, and W. H. Kilpatrick, an influential representative of that school, had declared the project method as Dewey's brainchild. But Dewey, while sympathetic to the idea of a participatory approach to education, would not want to see the curriculum content in terms of knowledge and competence buried underneath project-learning for its own sake. Continuity of experience, as he explained in 1938, requested the educator to transfer 'the vested capital of civilization': 'The mature person, to put it in moral terms, has no right to withhold from the young on given occasions whatever capacity for sympathetic understanding his own experience has given him' (*Experience and Education*, Later Works 1981–1991).

In *The Quest for Certainty* (1929), Dewey gives an account of cultural history that exposes the tradition of bad metaphysics in philosophy which seeks to establish a realm of certainty protected from the precariousness of human experience in a world of unavoidable uncertainty. While early religion attempted an alignment with the powers of destiny, philosophy—its inheritor—sought access to knowledge of the immutable, to the certainty provided by everlasting values and principles. Both religion and philosophy distinguished between skills and technical knowledge that eased the circumstances of everyday

existence, and the superior knowledge connected to the higher order of absolute certainty in terms of truth, goodness, and beauty. The sciences that developed in the seventeenth century effectively replaced the qualitative view of nature with one describing it in terms of mathematical relationships. But instead of using that knowledge as an instrument in the service of liberating experience, scientists engaged in a quest for certainty on their own by redescribing reality in scientific terms, as if 'the findings of science are a disclosure of the inherent properties of the ultimate real' (*The Quest for Certainty*, Later Works 1981–1991). Thereby, another dualism was generated: that between scientific procedure with the resulting quantitative description of relationships in nature, and that of the appreciation of qualities and values in experience. Dewey demonstrated how these dualisms dissolve as one turns away from the fixation on certainty in the name of reason and turns instead to intelligence as a means of assessing the possibilities inherent to a situation.

With *The Quest for Certainty*, Dewey aimed at a destruction of the history of philosophy in order to gain plausibility for democracy as the social arrangement with the greatest potentials for the full realization of society's capacities.

In *Art as Experience* (1934), Dewey stresses the continuity of all experience. The aesthetic is that part of experience that is identical with the form that separates an experience from the ongoing flow of ordinary experience. It is of a satisfying quality for its consistent movement toward an intended end, which makes it self-sufficient and lifts it out of the stream of the ordinary, at the same time. An agreeable meal, an intelligent conversation, a talk professionally delivered, a well-conducted concert, a well-played game, etc. all qualify as an experience. As for art, the artist clarifies and intensifies an object's traits in such a way that a new experience results. He does so in a process that requires his attention to the medium he is working with, and the beholder of the finished work appreciates it only to the degree that he goes through the same processes that the artist went through when making the object. By describing art this way, by identifying the aesthetic as part of experience accessible to all, Dewey claims that the removal of the arts from the ordinary lives of the common people, its relegation to museums and to private collections of the wealthy has to be understood as symptomatic of the dualism that separates the aesthetic from the quotidian and is at the roots of the divisions of society.

Like the aesthetic, the religious offers an aspect of consummatory experience. Dewey had hinted in the final chapter of *Reconstruction in Philosophy* (1920) at a 'natural piety' that transcends the denominational and liberates the potentials of religious belief. In 1934, he delivered a series of lectures to extend and clarify his position, which were published under the title *A Common Faith*. He draws a distinction between the religious as disposition and religions as sets of rituals and dogmatic beliefs; he shows the realm of the religious to be the kind of belief that is not in competition with the kind of knowledge achieved through application of research and reason; organized religion, however, has interfered with the progressive expansion of knowledge. Dewey sketches the contour of a natural piety that takes humans to be part of nature and seeks its cooperation not without a just and dignified attitude. Experience of the religious culminates in the unification of the ideal and the actual, i.e. the realization of our most inclusive moral convictions in a given situation, through which an all-pervasive unity may be glimpsed momentarily that aligns human purpose and nature.

Dewey's studies of various fields are all contributions to the foundation of a framework for social life that facilitates the liberation of human potential. Against the separation of experience and dualistic views, he promotes the unifying moments in experience and the resolution of dualism; against concepts of certain truth and the eternal, he emphasizes the contingent and the present in the situation at hand; instead of the closed-down (e.g., Marxist 'law of the laws of history') and shut-off, he points out the continuous and the transactional; against reifications of language and reason, he demonstrates their instrumentality; and instead of leaning to romantic pessimism, he puts a premium on cautious optimism in matters concerning human beings and the universe.

4. *Relevance for Contemporary Social Sciences*

A renewed interest in Dewey is indicated by the number of doctoral dissertations, which doubled during the 1990s from an average seven at US universities in the early years to an average 15 for each of the later years. At the same time, the tone of studies on Dewey appears to have changed. An observation that might be symptomatic: Westbrook (1991) follows Dewey for the most part, but is still critical of his emphasis on education to reach the goal of participatory democracy; Eldridge (1998) attacks Rorty for misreading Dewey on scientific method, Niebuhr for misconstruing Dewey's metaphysics, and Westbrook for prioritizing political action over cultural development.

The explanation of Russel's and Horkheimer's reproach of Dewey's pragmatism as curious misunderstanding has become a common feature in the literature of the nineties; the fierce Marxist–Leninist critique of Dewey's philosophy during the Soviet era is no longer an issue, other than a footnote, and the neo-Thomist repudiation of Dewey's secular philosophy is no longer even mentioned.

At the same time, several tendencies in the development of philosophy have helped facilitate a renewed interest in Dewey's philosophy. The rejection of 'great tales' in French postmodernist philosophy, the promotion of pluralistic views by proponents of

constructivism, have exposed the untenability of philosophy's claim of a relationship with the immutable, for their part. One of the most influential thinkers of the 1980s and 1990s, Richard Rorty, has combined these sentiments and applied them in his reformulation of the contributions of philosophy. His *Philosophy and the Mirror of Nature* (1979) can be read as mirroring Dewey's *Quest for Certainty*. He claims to be 'Deweyan' and comments explicitly on Dewey's writings. While this claim has been disputed by Westbrook (1991) and Eldridge (1998), Rorty's influence on contemporary thought has contributed to a renewed interest in Dewey's work.

There is renewed interest in Dewey's theory of democracy. Westbrook, in his seminal study (1991), has made explicit the relevancy and actuality of the Lippmann–Dewey debate for American democracy at the end of the twentieth century, and the time of that study's publication and its wide reception coincided with a similar appeal for the systematic pursuit of community by a group of philosophers (e.g., M. Dancer) and sociologists (e.g., R. N. Bellah) referred to as 'communitarians.' Perhaps it is indicative of a renewed influence of Dewey's concept of 'great community' that J. Habermas closed a review of the publication in German, after 70 years, of *Quest for Certainty*, expressing the wish: 'He (Dewey) would be the better patron for the Berlin Republic.'

See also: Darwinism; Democratic Theory; Educational Philosophy: Contemporary Views; Educational Philosophy: Historical Perspectives; James, William (1842–1910); Mead, George Herbert (1863–1931); Pragmatism: Philosophical Aspects; Pragmatist Social Thought, History of

Bibliography

Boydston J A (ed.) 1973 *The Early Works of John Dewey, 1882–1898*, 5 Vols. Southern Illinois University Press, Carbondale, IL

Boydston J A (ed.) 1973–1983 *The Middle Works of John Dewey, 1899–1924*, 15 Vols. Southern Illinois University Press, Carbondale, IL

Boydston J A (ed.) 1981–1991 *The Later Works of John Dewey, 1925–1953*, 17 Vols. Southern Illinois University Press, Carbondale, IL

Boydston J A (ed.) 1974 *Guide to the Works of John Dewey*. Southern Illinois University Press, Carbondale, IL

Dewey J 1939 Biography of John Dewey. In: Schilpp P A (ed.) *The Philosophy of John Dewey*. Tudor, New York

Dykhuizen G 1973 *The Life and Mind of John Dewey*. Southern Illinois University Press, Carbondale, IL

Eldridge M 1998 *Transforming Experience. John Dewey's Cultural Instrumentalism*. Vanderbilt University Press, Nashville, TN

Hickman L A 1990 *John Dewey's Pragmatic Technology*. Indiana University Press, Bloomington, IN

Rorty R 1979 *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ

Thomas M H 1962/1972 *John Dewey: A Centennial Bibliography*. Chicago University Press, Chicago

Thomas M H Writings of John Dewey to October 1939. In: Schilpp P A (ed.) *The Philosophy of John Dewey*. Tudor, New York

Westbrook R B 1991 *John Dewey and American Democracy*. Cornell University Press, Ithaca, NY

H. Schreier

Diabetes: Psychosocial Aspects

Diabetes mellitus is a group of chronic illnesses for which there is no cure, and which requires a complex, life-long, daily regime of management. Diabetes is a demanding chronic illness, behaviorally and psychologically, and psychosocial factors are implicated heavily in all aspects of the disease and its treatment.

1. Diabetes Mellitus

1.1 Background

Diabetes mellitus is characterized by abnormality in the production or utilization of insulin to regulate blood sugar levels. It has a complex etiology and pathogenesis (Mandrup-Poulsen 1998), and affects about five percent of the population, with variation by age, gender, ethnicity, and geographic location. It is grouped crudely into two forms, Type 1 and Type 2, with some rarer forms such as gestational diabetes and maturity onset diabetes of the young. Type 1 diabetes (insulin-dependent diabetes or IDDM) is an autoimmune disease characterized by insulin deficiency, accounting for about 15 percent of diagnoses, typically in people less than 30 years of age. Type 2 diabetes (noninsulin-dependent diabetes or NIDDM) is more complex and is characterized by resistance to insulin-mediated glucose disposal or impaired insulin secretion. It accounts for about 85 percent of diagnoses, typically in people aged 40 or older. The disease has serious long-term consequences caused by elevated blood sugar levels, or hyperglycemia, and can result in blindness, limb amputation, liver failure, and increased risk of stroke and heart disease.

1.2 Management

The diabetic individual must attempt to maintain strict glucose control to avoid low blood sugar levels, or hypoglycemia, which is very aversive, and to minimize long-term complications. Obesity is a major issue in Type 2 diabetes, and management initially focuses on control through diet and exercise, which is usually supplemented with oral medication, and sometimes with insulin, as the disease progresses. Type 1

diabetics, and Type 2 diabetics with little insulin production, must receive insulin through daily self-injection to survive. Treatment also involves dietary change, eating foods containing complex rather than simple sugars and distributing food intake over the day. Regulation of exercise levels and balancing these with food intake is also required to maintain blood sugar levels within a normal range. Management thus involves regular attention to a range of daily self-care tasks, attending to diet, exercise, medication, and blood glucose level monitoring.

Most research into psychosocial factors in diabetes has focused on successful control of blood glucose levels. A major longitudinal study, the Diabetes Control and Complications Trial, established that strict metabolic control can sharply reduce long-term complications from diabetes, and it set standards for glycated hemoglobin levels considered as adequate control. However, strict control is not always achieved as diabetics juggle the demands of the disease against the demands of daily life. This has prompted a move from compliance to adherence (Jenkins 1995) and for educational programs to incorporate local cultural understandings, increase knowledge, motivation and goal-setting, and empower diabetics to self-adjust their regime.

2. *Psychosocial Aspects of Diabetes*

2.1 *Knowledge, Attitudes, and Beliefs*

Understandings of diabetes and attitudes towards it can influence how it is managed. Diabetics with greater knowledge of the disease have been found to deal with issues of glycemic control more appropriately than patients with less knowledge. Knowledge and beliefs about insulin function has also been linked to the pattern of insulin use adopted, and indirectly to control. Hence, knowledge and beliefs are considered important components in diabetic education programs aimed at enhancing adherence. However, their effects on control are generally indirect and mediated by other psychosocial factors.

Comparison of the understandings of diabetes held by patients and health professionals reveal considerable differences between them in responsibility for control and explanations for the disease. Professionals' disease models emphasize pathological and physical aspects whereas patients' models emphasize social aspects and the impact of the disease on their lives, leading to issues of adherence and control being valued differently by these groups.

2.2 *Stress, Coping, and Adjustment*

Stress is widely considered to have effects on diabetes onset and management, but the evidence for this is mixed, and dependent upon the assessment of stress and outcome measures (see *Stress and Health Research*). Cox and Gonder-Frederick (1992) concluded that:

(a) relationships between stress and diabetes are bidirectional, and stress can affect diabetes directly (through hormone release) and indirectly (by disrupting self-care behavior),

(b) with regard to onset, no research has linked stress to onset of Type 2, but there is evidence for increased incidence of stressful life events before Type 1 onset, and

(c) the effects of stress on blood glucose levels after onset are variable. Laboratory studies produce mixed findings and indicate marked individual differences, suggesting that some subgroups are more reactive than others. Field studies generally support a relationship between increased stress and increased blood glucose levels, but differ as to whether this is mediated by self-care behaviors.

Subsequent research provides little evidence for changing these conclusions. Recent research has moved to focus on daily stressors rather than major life events, to use repeated daily records rather than retrospective ratings, and to examine individual variability in stress response. These directions, combined with prospective designs and better controls, may lead to improved models of stress effects in diabetes.

Adjustment to diabetes follows a similar course to other chronic illnesses. Diagnosis can lead to a sense of loss, grieving, and concerns about the future, followed by a period of adaptation to a new life situation and readjustment of goals and life understandings. Successful negotiation of this leads to the development and use of skills to manage the changed situation successfully. Research into diabetic coping strategies is limited and has generally been confined to the postdiagnosis period, and focused on Type 1 diabetes. Most studies demonstrate problem-focused or task-oriented strategies to be better, at least in relation to glycemic control. However, some studies have shown denial or avoidance coping to be adaptive, and this is a continuing area of debate. Gender differences in the use of strategies have also been documented.

2.3 *Social Support*

Social support is a complex construct (see *Social Support and Health*), but generally is found to have positive effects on diabetes management. Research has demonstrated positive effects on adherence and control for both structural support (e.g., family, friends, workers as sources, density of support networks) and for functional support (e.g., diabetes-specific help, communication style, cohesiveness). As much of this research is cross-sectional, directional relationships between support and outcome are not clear. Gender differences have been documented for both level and

type of support and for its relation to outcome. Adolescence has been a particular focus of this research (Burroughs et al. 1997), and has documented the importance of a supportive family, particularly following diagnosis. Functional support, especially involving open, empathic communication within families, is also important for achieving good adherence with adolescents.

Evaluation of interventions designed to enhance support, such as social skills training for diabetics, and diabetic understanding for families, has revealed positive effects on adherence and control. Several trials have demonstrated that group instruction to impart diabetes knowledge and coping skills produces better results than individualized instruction. Support provided through self-help groups (see *Support and Self-help Groups and Health*) or through a mentor (a well-adjusted patient) has been promoted but not researched in the context of diabetes.

2.4 Quality of Life

Diabetes, with daily requirements for self-monitoring and management in order to avoid the short-term consequences of hypoglycemia and the long-term complications associated with hyperglycemia, has a substantial impact on daily life. This has led to considerable interest in diabetic quality of life (QoL), assessed through a wide range of concerns including morale, well-being, depression, and role functioning (see *Chronic Illness: Quality of Life*). The relationship between glycemic control and QoL for Type 2 diabetics is mixed; for Type 1 diabetics, most research finds glycemic control and QoL to be related, cross-sectionally and prospectively. Recent research has also found intensive insulin therapy to be associated with better QoL outcomes, attributed to making management more compatible with the demands of daily living.

Psychiatric morbidity is higher for both Type 1 and Type 2 diabetics than in matched controls, and this is compounded by diabetic complications: those with complications have higher levels of morbidity than those without. Depression has been a specific focus of studies, and prevalence rates between 8 and 27 percent are reported in diabetic patients. Depression appears not to be directly related to glycemic control.

QoL is multidimensional, and only some aspects of it relate to disease-specific outcomes such as control. The development of diabetes-specific measures of QoL (Bradley 1994) provides a means to standardize measurement in this area, and to assess the impact of diabetes on aspects of life other than control.

2.5 Diet and Eating Disorders

Dietary factors are specifically relevant in diabetes: food choice for Type 1 and reducing body weight for Type 2. Limited research here suggests that breaches of dietary self-management show high individual variability, and are associated with reactions to hypoglycemic symptoms and social cues about eating.

Research on dietary issues has focused on younger female diabetics and the relationship between diabetes and eating disorders. Diabetic women are reported to have higher levels of body dissatisfaction and desire for thinness than age-matched healthy controls, and women who omit insulin injections (to lose weight) are found to be more sensitive to external cues to eating. Many reports suggest an increase of eating disorders in Type 1 diabetics, and also an increase of Type 1 diabetes amongst patients with eating disorders. However, recent controlled research does not support this. Intentional insulin admission is, however, potentially indicative of eating disorder in Type 1 patients.

2.6 Sexual Dysfunction

Diabetes is associated with disorders of sexual function, but this varies by gender. Diabetic men report lower levels of erotic drive, sexual arousal, enjoyment, and satisfaction than controls. Up to one-third of diabetic men experience erectile disorder: three times the rate for healthy controls but similar to rates for other chronic diseases. For diabetic women, sexual dysfunction rates appear not to differ substantially from general population rates, although there is limited evidence for loss of libido, enjoyment, and satisfaction. Apart from an association between neurological disorder and erectile dysfunction for men, it appears that psychosocial factors, such as disease acceptance or depression, rather than disease-related factors, such as glycemic control or disease duration, are more predictive of sexual dysfunction, particularly for women (Thomas and LoPiccolo 1994).

3. Future Directions

Psychosocial factors are implicated in almost all aspects of diabetes, but research findings in this area are often inconclusive, fragmented, or limited methodologically. To overcome this, future research should: develop improved theoretical accounts of psychosocial influence; make more systematic use of standard measures such as those described by Bradley (1994); extend research to examine other outcomes alongside glycemic control; conduct more controlled trials and longitudinal studies; and conduct more research, absolutely and comparatively, on the more common Type 2 diabetes.

See also: Behavioral Medicine; Chronic Illness, Psychosocial Coping with; Health Behavior: Psychosocial

Theories; Health Behaviors; Health: Self-regulation; Self-regulation in Adulthood

Bibliography

Bradley C (ed.) 1994 *Handbook of Psychology and Diabetes*. Harwood, Chur, Switzerland

Burroughs T E, Harris M A, Pontious S L, Santiago J V 1997 Research on social support in adolescents with IDDM: A critical review. *The Diabetes Educator* **23**: 438–48

Cox D J, Gonder-Frederick L 1992 Major developments in behavioral diabetes research. *Journal of Consulting and Clinical Psychology* **60**: 628–38

Jacobson A M 1996 The psychological care of patients with insulin-dependent diabetes mellitus. *New England Journal of Medicine* **334**: 1249–53

Jenkins C D 1995 An integrated behavioral medicine approach to improving care of patients with diabetes mellitus. *Behavioral Medicine* **21**: 53–65

Mandrup-Poulsen T 1998 Diabetes. *British Medical Journal* **316**: 1221–25

Thomas A M, LoPiccolo J 1994 Sexual functioning in persons with diabetes: Issues in research, treatment and education. *Clinical Psychology Review* **14**: 61–86

Warren L, Hixenbaugh P 1998 Adherence and diabetes. In: Myers L B, Midence K (eds.) *Adherence to Treatment in Medical Conditions*. Harwood, Amsterdam, The Netherlands

K. Chamberlain

Dialectical Behavior Therapy

Dialectical behavior therapy (DBT) is a comprehensive cognitive-behavioral treatment program designed for complex difficult-to-treat mental disorders. Originally developed to treat chronically suicidal patients, the treatment evolved into a treatment for suicidal patients who also meet criteria for borderline personality disorder (BPD), and has since been adapted for BPD patients with presenting problems other than suicidal behaviors and for other disorders of emotion regulation. DBT combines the basic strategies of behavior therapy (see *Behavior Therapy: Psychological Perspectives*) with Eastern mindfulness and teaching practices, residing within an overall frame of a dialectical worldview that emphasizes the synthesis of opposites. The fundamental dialectical tension in DBT is between an overarching emphasis on validation and acceptance of the patient on the one hand, with persistent attention to behavioral change on the other. Change strategies in DBT include analysis of maladaptive behavior chains, problem solving techniques, behavioral skills training, management of contingencies (i.e., reinforcers, punishment), cognitive modification, and exposure-based strategies. DBT acceptance procedures consist of mindfulness (e.g., attention to the present moment, assuming a nonjudgmental stance, and focusing on effectiveness) and a variety of validation and stylistic strategies.

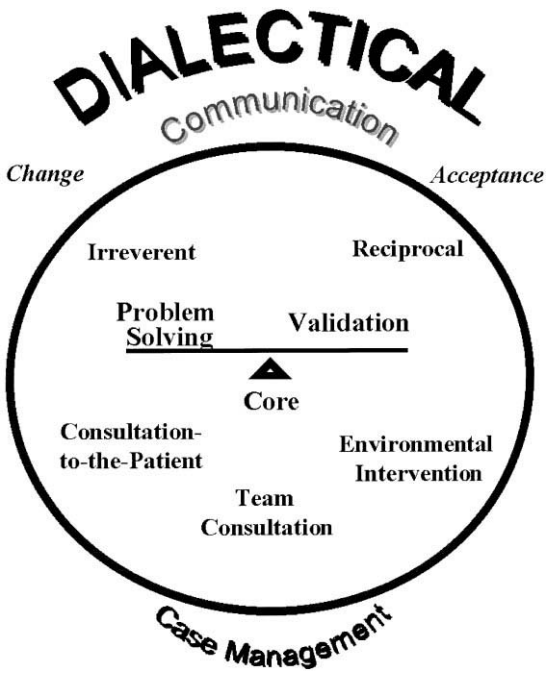

Figure 1
A graphical representation of the key concepts in dialectical behavior therapy

1. *Development of DBT*

DBT, developed by Linehan in the early 1980s, grew out of a series of failed attempts to apply the standard cognitive and behavior therapy protocols of the early 1980s to chronically suicidal patients. Difficulties that emerged included the following. (a) Focusing on client change, either by improving motivation or by enhancing skills, was often experienced as invalidating by individuals meeting criteria for BPD and often precipitated withdrawal from therapy, attacks on the therapist, or vacillation between these behaviors. (b) Teaching and strengthening new skills was extraordinarily difficult, if not impossible, within the context of a therapy oriented to reducing the motivation to die. (c) The rigorous control of the therapy agenda needed for skills training did not allow for sufficient attention to be given to motivational issues. (d) New behavioral coping skills were extraordinarily difficult to remember and apply when the patient was in a state of crisis. (e) Borderline individuals often

unwittingly reinforced therapists for iatrogenic treatment (e.g., a client might stop attacking the therapist if the therapist changed the subject away from topics the client was afraid to discuss) and punished them for effective treatment strategies (e.g., a client may respond with a drug overdose when the therapist refuses to recommend hospitalization stays that reinforce suicide threats).

Three modifications to standard behavior therapy were made to take these factors into account. First, strategies that more clearly reflected radical acceptance and validation of clients' current capacities and behavioral functioning were added to the treatment. The synthesis of acceptance and change within the treatment as a whole and within each treatment interaction led to adding the term 'dialectical' to the name of the treatment. This dialectical emphasis brings together in DBT the 'technologies of change' based both on principles of learning and crises theory and the 'technologies of acceptance' (so to speak) drawn from principles of Eastern Zen and Western contemplative practices. Second, the therapy as a whole was split into several different components, each focusing on a specific aspect of treatment. The components in standard outpatient DBT are highly structured individual or group skills training, individual psychotherapy (addressing motivation and skills strengthening), and telephone contact with the individual therapist (addressing application of coping skills). Third, a consultation/team meeting focused specifically on keeping therapists motivated and providing effective treatment was added as a fourth treatment component.

2. *Treatment Stages and Goals*

DBT is designed to treat patients at all levels of severity and complexity of disorders. In stage 1, DBT focuses on patient behaviors that are out of control. The treatment targets in stage 1 are: (a) reducing high-risk suicidal behaviors (parasuicide acts, including suicide attempts, high-risk suicidal ideation, plans, and threats); (b) reducing client and therapist behaviors that interfere with the therapy (e.g., missing, coming late to sessions, telephoning at unreasonable hours or otherwise pushing a therapist's limits, not returning phone calls); (c) reducing behavioral patterns serious enough to substantially interfere with any chance of a reasonable quality of life (e.g., depression and other axis I disorders, homelessness, chronically losing relationships or jobs); and (d) acquisition of sufficient life skills to meet client goals (skills in emotion regulation, interpersonal effectiveness, distress tolerance, self-management, as well as mindfulness). Subsequent to achieving behavioral control, it becomes possible to work on other important goals including (e) focusing on the ability to experience emotions without trauma (stage 2); (f) enhancing self-respect; (g) addressing problematic patterns in living that interfere with goals (stage 3); and (h) resolving feelings of incompleteness and enhancing the capacity for joy (stage 4). In sum, the orientation of the treatment is to first get action under control, then to help the patient to feel better, to resolve problems in living and residual disorder, and to find joy and, for some, a sense of transcendence.

With respect to each goal in treatment, the task of the therapist is first (and many times thereafter) to elicit the client's collaboration in working on the relevant behavior, then to apply the appropriate treatment strategies described below. Treatment in DBT is oriented to current in-session behaviors or problems that have surfaced since the last session, with suicide attempts and other intentional self-injurious acts and life-threatening suicidal ideation, planning, and urges taking precedence over all other topics. Because therapeutic change is usually not linear, progress through the hierarchy of goals is an iterative process.

3. *Treatment Strategies*

DBT addresses all problematic client behaviors and therapy situations in a systematic, problem solving manner. The context of these analyses and of this solution-oriented approach is that of validation of the client's experiences, especially as they relate to the individual's vulnerabilities and sense of desperation. In contrast to many behavioral approaches, at least as described in print, DBT places great emphasis on the therapeutic relationship. In times of crisis, when all else fails, DBT uses the relationship itself to keep the client alive.

Treatment strategies in DBT are divided into four sets: (a) dialectical strategies, (b) core strategies (validation and problem solving), (c) communication strategies (irreverent and reciprocal communication), and (d) case management strategies (consultation to the patient, environmental intervention, supervision/consultation with therapists). There are, as well, a number of specific behavioral treatment protocols covering suicidal behavior, crisis management, therapy-interfering behavior and compliance issues, relationship problem solving, and ancillary treatment issues, including psychotropic medication management. These are more fully described in the treatment manuals (Linehan 1993a, 1993b). Their relationships to each other, and to acceptance and change, are represented in Fig. 1.

Dialectical strategies are woven throughout all treatment interactions. The primary dialectical strategy is the balanced therapeutic stance, the constant attention to combining acceptance with change. The goal is to bring out the opposites, both in therapy and the client's life, and to provide conditions for synthesis. Strategies include extensive use of stories, metaphor,

and paradox; the therapeutic use of ambiguity; viewing therapy, and indeed all of reality, as constant change; cognitive challenging and restructuring techniques; and reinforcement for use of intuitive, nonrational knowledge bases.

Core strategies consist of the balanced application of problem solving and validation strategies. Problem solving is a two-stage process involving, first, an analysis and acceptance of the problem at hand, and second, an attempt to generate, evaluate, and implement alternative solutions which might have been made or could be made in the future in similar problematic situations. Analysis of the client's problem behaviors, including dysfunctional actions, emotions, physiological responses, and thought processes, requires detailed analyses of the chains of events (both responses of the individual and events in the environment) leading up to the problematic responses. This analysis is repeated for every instance of targeted problem behaviors until both therapist and client achieve an understanding of the response patterns involved. The second stage, which is actually interwoven with the first, requires the generation of new, more skillful responses, as well as an analysis of the individual's capabilities and motivation to actually engage in the new behaviors. This process leads into application of change procedures drawn primarily from innovations in behavior therapy and from the cognitive modification techniques of rational-emotive therapy (see *Behavior Psychotherapy: Rational and Emotive*). Management of contingencies in the therapeutic relationship, training in behavioral skills, exposure techniques with response prevention, and cognitive restructuring anchor the behavioral change end of the primary dialectic in DBT.

Validation, which balances and keeps problem solving moving forward, requires the therapist to search for, recognize, and reflect the validity, or sensibility, of the individual's current patterns of behavior. Validation is required in every interaction and is used at any one of six levels: Level 1, listening to the client with interest; Level 2, reflection, paraphrasing, summarizing; Level 3, articulating or 'mind-reading' that which is unstated, such as fears of admitting emotions or thoughts, but without pushing the interpretation on the client; Level 4, validating the client's experience in terms of past experiences or in terms of a disease or brain disorder model; Level 5, validating the client in terms of present and normal functioning (e.g., high emotional arousal disorganizes cognitive functioning in all people); and Level 6, radical genuineness with clients—treating the client–therapist relationship as an authentic and 'real' relationship. In line with this latter prescription of the treatment, the therapist does not treat clients as fragile or unable to solve problems.

In DBT, the therapist balances two communication strategies which represent rather different interpersonal styles. The modal style is the reciprocal strategy, which includes responsiveness to the client's agenda and wishes, warmth, and self-disclosure of both personal information that might be useful to the client as well as immediate reactions to the client's behavior. Reciprocity is balanced by an irreverent communication style: a matter-of-fact or at times slightly outrageous or humorous attitude where the therapist takes the client's underlying assumptions or unnoticed implications of the client's behavior and maximizes or minimizes them, in either an unemotional or overemotional manner to make a point the client might not have considered before. Irreverence 'jumps track,' so to speak, from the client's current pattern of response, thought, or emotion.

There are three strategies designed to guide each therapist during interactions with individuals outside the therapy dyad (case management strategies). The consultation/supervision strategy requires that each DBT therapist meet regularly with a supervisor or consultation team. The idea here is that severely suicidal individuals should not be treated alone. The consultant-to-the-client strategy is the application of the principle that the DBT therapist teaches the client how to interact effectively with the client's environment, rather than teaching the environment how to interact with the client. When absolutely necessary, however, the therapist actively intervenes in the environment to protect the client or to modify situations that the client does not have the power to influence.

4. *Research*

Data suggest that with severely disordered patients meeting criteria for borderline personality disorder, DBT is more effective than treatments applied as usual in the community in reducing targeted problem behaviors (suicide attempts, substance abuse, treatment dropout, use of psychiatric in-patient hospitalization) and improving both general and social functioning (Linehan et al. 1991, 1994). With a less severe population of BPD patients, DBT appears to produce specific improvements in suicidal ideation, depression, and hopelessness, even when compared to a treatment-as-usual condition that also produces clinically significant changes in depression (Koons et al. 2001). Where follow-up data are available, it appears that these results may hold up to one year post-treatment (Linehan et al. 1993). These findings are generally replicated by both randomized clinical trials with better experimental control as well as quasi-experimental effectiveness studies. From available evidence, it appears that DBT is a step toward more effective treatment for multidisordered patients, particularly those with suicidal behavior.

See also: Behavior Therapy: Psychiatric Aspects; Behavior Therapy: Psychological Perspectives; Bor-

derline Personality Disorder; Cognitive Therapy; Communication and Social Psychology; Psychological Treatment, Effectiveness of; Suicide

Bibliography

Bohus M, Haaf B, Stiglmayr C, Pohl U, Böhme R, Linehan M in press Evaluation of inpatient dialectical-behavioral therapy for borderline personality disorder—a prospective study. *Behaviour Research and Therapy*

Koerner K, Linehan M M in press Research on dialectical behavior therapy for borderline personality disorder. *Psychiatric Clinics of North America*

Koons C R, Robins C J, Bishop G K, Morse J Q, Tweed J L, Lynch T R, Gonzalez A M, Butterfield M I, Bastian L A 2000 Efficacy of dialectical behavior therapy in women veterans with borderline personality disorder. *Behavior Therapy* **32**: 371–90

Linehan M M 1993a *Cognitive Behavioral Therapy of Borderline Personality Disorder*. Guilford Press, New York

Linehan M M 1993b *Skills Training Manual for Treating Borderline Personality Disorder*. Guilford Press, New York

Linehan M M 1997 Behavioral treatments of suicidal behaviors: Definitional obfuscation and treatment outcomes. In: Stoff D M, Mann J J (eds.) *Annals of the New York Academy of Sciences: Vol. 836. The Neurobiology of Suicide: From the Bench to the Clinic*. New York Academy of Sciences, New York, pp. 302–28

Linehan M M 1998 An illustration of dialectical behavior therapy. *In Session: Psychotherapy in Practice* **4**: 21–44

Linehan M M, Armstrong H E, Suarez A, Allmon D, Heard H L 1991 Cognitive-behavioral treatment of chronically parasuicidal borderline patients. *Archives of General Psychiatry* **48**: 1060–4

Linehan M M, Heard H L, Armstrong H E 1993 Naturalistic follow-up of a behavioral treatment for chronically parasuicidal borderline patients. *Archives of General Psychiatry* **50**: 971–4

Linehan M M, Tutek D A, Heard H L, Armstrong H E 1994 Interpersonal outcome of cognitive behavioral treatment for chronically suicidal borderline patients. *American Journal of Psychiatry* **151**: 1771–6

Miller A, Rathus J H, Linehan M M, Wetzler S, Leigh E 1997 Dialectical behavior therapy adapted for suicidal adolescents. *Journal of Practical Psychiatry and Behavioral Health* **3**: 78–86

M. M. Linehan

Dialectical Thinking, Psychology of

Dialectical thinking refers to the mental processes of compromising or synthesizing facts, views, and goals of opposing perspectives (Basseches 1984, Moshman 1995, Peng 1997). Only in the late 1970s did the study of dialectical thinking appear in the mainstream psychological literature as a research topic (e.g., Riegel 1973, Basseches 1980, 1984, Kramer 1983), and even at the start of the twenty-first century little is known about the nature of dialectical thinking and its effects on basic reasoning, judgment, and decision-making processes. Still, a variety of work from psychology and philosophy can be seen as bearing on the issue, including thinkers who rooted philosophical systems in dialectical processes. A number of different levels of analysis have been considered—including dialectical dynamics at the societal level, at the level of teacher–student and opposing debaters, and at the intrapsychic level—each offering insights on how opposite positions interact and how they are reconciled, if at all.

1. *Dialectical Thinking as Methods of Reasoning*

Dialectical thinking as a method of reasoning has a long tradition in Western thought. One of the earliest examples of the dialectical method is Plato's *Dialogues*, in which he sought to study truth through discussion in the form of questions and answers. Aristotle also referred to dialectics as a method of reasoning but used it often as a synonym for the science of logic. The first thinkers who distinguished dialectical reasoning from formal logic were Immanuel Kant and G. W. F. Hegel. Kant maintained that logical reasoning is very effective within the confines of science, but 'all the worse for the beyond' (Kant 1781/1965). He believed that there are ways to go beyond things as experienced ('phenomena') and to seek the ultimate reality of things in themselves ('noumena'); this is Kant's transcendental dialectic. Hegel, meanwhile, believed that the evolution of ideas occurs through a dialectical process—that is, a concept (or thesis) gives rise to its opposite (antithesis), and as a result of this contradiction, a third view, the synthesis, arises. The synthesis is at a 'higher' level of truth than the first two views (Hegel 1812/1967). The Hegelian method of reasoning has since become seemingly synonymous with dialectical thinking in Western cultures.

2. *Marx and Engels' Laws of Dialectics*

One of the most influential thinkers of the nineteenth century, Karl Marx opened new vistas of understanding by applying materialistic and dialectical perspectives to the analysis of civilization and cultures. According to Marx and Engels, there are three basic laws of dialectics: (a) the law of the transformation of quantity into quality and vice versa; (b) the law of the interpenetration of opposites; and (c) the law of the negation of the negation (Engels 1872–82/1940, p. 26). Developed in contrast to the Hegelian idealism that envisioned an end to contradiction in achievement of the absolute ideal, Marxist dialectical thought recognizes the permanence of opposition and contradiction in the real material world, and therefore in thought about reality. According to Marx and Engels, the

process of historical development is endless because the synthesis itself becomes a new thesis and is in turn negated by a new antithesis producing a new synthesis, and so on *ad infinitum*.

3. *Dialectical Thinking as a Form of Argument*

Dialectical thinking in contemporary epistemology is defined as a form of argument (Alexy 1989, Habermas 1990, Goldman 1994) that purposely draws out the views of either participant. In formal logical reasoning, an argument is good if the conclusion either follows deductively from the premises, or receives strong evidential support from them, or has only true premises. However, in dialectical argument or discourse where two (or more) speakers successively defend conflicting positions, both citing premises in support of their respective positions, the judgment of the goodness of the argument is a matter of social epistemology and conversational norms (Goldman 1994). In this form of argument, the opposing perspective is invited for the sake of discourse.

4. *Dialectical Thinking as a Psychological Concept*

In the late 1970s and early 1980s, a school of developmental psychologists, frustrated by the limited relevance of Piagetian theories of human development for adult development, started to look for cognitive development beyond Piaget's adolescent 'formal operations' stage. Klaus Riegel, in his 1973 paper on 'Dialectic operations,' pointed out that, 'It has never been shown convincingly that the highest level of operation, i.e., formal operational intelligence, characterizes the thinking of mature adults' (Riegel 1973, p. 363). Riegel believed that adult thoughts, particularly in creative scientific activities, 'are dominated by playful manipulations of contradictions and by conceiving issues integratively which have been torn apart by formal operational thinking.' Accordingly, the final period of human cognitive development and the nature of mature thought are characterized in part by dialectical operations.

A more empirical approach to dialectical thinking was offered by Michael Basseches (1980, 1984), who abstracted 24 cognitive schemata or patterned movements in human thought as basic components of dialectical thinking. These 'dialectical schemata,' as Basseches called them, include steps in dialectical analyses and ways of introducing and maintaining dialectical movement within one's own thought, such as recognition and description of thesis–antithesis–synthesis development, location and description of contradiction, and assertion and description of reciprocal relationships (Basseches 1980). It has been found that such dialectical thinking becomes increasingly important and common in old age (e.g., Basseches 1984, Kramer and Woodruff 1986, Chandler and Boutilier 1992). Middle-aged and older people are more likely to accept contradiction in reality and synthesize contradiction in their thinking than young people (Kramer and Woodruff 1986).

However, some scholars still believe that formal operations can also account for dialectical forms of thought (e.g., Piaget 1980, Kramer 1983). For example, Piaget (1980) argued that there are two types of contradiction: a logical contradiction ('real contradiction') that results from an error in the reasoning process and a natural contradiction ('pseudocontradiction') that results from disequilibrium in one's knowledge. Dialectical synthesis, according to Piaget, is a re-equilibration process of formal operations in dealing with natural contradiction. A conciliatory proposal was made by Kramer, who suggested that formal operations and postformal operations may actually reflect only two different world views, not necessarily two different stages of human cognitive development. Formal operational thought is an analytic world view that emphasizes independence of variables and underlying stability whereas postformal operational thought is a synthetic world view that assumes interdependence of variables and change as basic to reality (Kramer 1983).

5. *Dialectical Thinking as Sources of Individual Differences*

The construct of conceptual or integrative complexity has been used by scholars in many fields of social science to capture individual differences in the complexity of the cognitive rules that people use to process and analyze information (Harvey et al. 1961, Schroder et al. 1967) or situational determinants of complex information processing (Streufert and Streufert 1978, Tetlock 1983, 1985). Levels of integrative complexity in people's thinking vary. For instance, people's reasoning and decision-making processes could be characterized as low differentiation (lack of awareness of alternative ways of viewing problems), high differentiation with low integration (awareness of alternative ways of viewing problems, but no recognition of relations between viewpoints), or high differentiation with high integration (awareness of multiple perspectives on problems and of interrelations among perspectives) (Tetlock 1985). Dialectical thinking should be reflected in higher levels of integrative complexity.

Some psychological work on negotiation, conflict management, and decision making has suggested that integration of different perspectives improves efficiency and outcomes. A coordinative negotiation strategy that attempts to establish collaboration between both sides was found to be more likely to achieve an agreement than a competitive negotiation strategy

(e.g., Tetlock 1985, Pruitt 1981, Pruitt and Carnevale 1993). In a decision-making process, the method of dialectical inquiry (DI) offers a credible counterplan and conducts a structured debate on both plans. Research suggests such an approach may be preferable to simple devil's advocacy (DA) that promotes conflict by criticizing a prevailing plan without offering any counter-plan (e.g., Mason 1969, Mitroff and Mason 1981, Schwenk 1990).

6. *Cultural Foundations of Dialectical Thinking*

Despite the differences among various conceptions of dialectical thinking, they make similar fundamental assumptions. First, they assume contradiction is a temporary state that will be replaced by integrated or synthesized thoughts (even Marx and Engels' constant negation assumes at least temporary integration). Second, they assume that the progress of reasoning is linear, logical, and moving in one direction—from a contradiction to a synthesis. Third, they assume integration or synthesis are at higher levels of cognitive functioning (e.g., Baltes and Staudinger 1993) usually considered to be more sophisticated, creative, and advanced models of thought (e.g., Benack et al. 1989). In the end, most of these approaches are rooted in the laws of formal logic, particularly the law of non-contradiction.

It has long been suspected that certain cultural groups are more dialectical in their reasoning and thinking about the world, knowledge, and human life. Peng (1997) (also Peng and Nisbett 1999, 2000, Peng et al. 2001) has suggested that Chinese folk epistemology (folk beliefs of knowing) emphasizes three interconnected principles—the principle of change (e.g., reality is a process, it does not stand still but is in constant flux), the principle of contradiction (reality is not precise, but is full of contradictions), and the principle of holism (in reality, nothing is isolated but connected, and the whole is more than the sum of its parts). These principles form the foundation of Chinese dialectical thinking, which is seemingly at odds with three basic laws of formal logic: the law of identity (if anything is true, then it is true; thus $A = A$), the law of non-contradiction (no statement can be both true and false; thus $A \neq A$), and the law of the excluded middle (any statement is either true or false; thus (A v B) and (A & B)). Such epistemological differences were found in folklore of Eastern and Western cultures. For example, Chinese folklore contains more dialectical proverbs than does American folklore (Peng 1997, Peng and Nisbett 1999, 2000). Americans as a whole preferred non-dialectical to dialectical proverbs and the Chinese preferred dialectical to nondialectical proverbs when they were presented with samples of commonly used proverbs from Chinese, American, and Yiddish cultures (Peng and Nisbett 1999).

7. *Effects on Reasoning and Judgment*

These epistemological differences of dialectical thinking imply strong cultural differences in reasoning about contradiction. A number of empirical findings have emerged which reveal important cultural differences in dialectical thinking. For instance, in a study of interpersonal conflict resolution, Peng (1997) found that American participants' resolutions of contradictions in everyday life were noncompromising and polarizing. In contrast, Chinese responses were much more dialectical and compromising. Chinese dialectical thinking may also contribute to the fact that Chinese are less likely than Americans to engage in debates and formal argumentation about truth (Becker 1986). Peng (1997) presented two types of arguments—logical and dialectical—on physics and religion to American and Chinese college students and found that, compared with Americans, Chinese participants preferred dialectical arguments. In a direct test of cultural difference on reasoning about contradiction, Peng (1997) used contradictory research reports as stimuli and asked participants to judge the plausibility of each piece of information. He found that American participants followed a *differentiation* approach, such that, when presented with (weak) evidence against the more plausible statement, they actually judged it to be *more* likely than those who saw no contradictory evidence did. Meanwhile, the Chinese participants showed a *compromising* approach. They judged the more plausible statement and the less plausible statement as equally likely when they presented together. In effect, the Chinese behaved as if they believed that both statements might be (somewhat) true.

8. *Conclusion*

Dialectical thinking is an important part of everyday reasoning and judgment. While it has long played a pivotal role in various philosophical discussions, it is only beginning to be understood as a psychological issue. It seems clear that dialectical thinking exists alongside other modes of thought, its use differing over the life course, between individuals and across cultures. Still, psychologists have much left to discover in order to fully understand the psychology of dialectical thinking.

Bibliography

Alexy R 1989 *A Theory of Legal Argumentation*. Oxford University Press, Oxford, UK

Baltes P B, Staudinger U M 1993 The search for a psychology of wisdom. *Current Directions in Psychological Science* **2**: 75–80

Basseches M 1980 Dialectical schemata: A framework for the empirical study of the development of dialectical thinking. *Human Development* **23**: 400–21

Basseches M 1984 *Dialectical Thinking and Adult Development*. Ablex, Norwood, NJ
Becker C B 1986 Reasons for the lack of argumentation and debate in the Far East. *International Journal of Intercultural Relations* **10**: 75–92
Benack S, Basseches M, Swan T 1989 Dialectical thinking and adult creativity. In: Glover J A, Ronning R R, Reynolds C R (eds.) *Handbook of Creativity*. Plenum Press, New York, pp. 199–208
Chandler M, Boutilier R G 1992 The development of dynamic system reasoning. *Human Development* **35**: 121–37
Engels F 1872–82/1940 *Dialectics of Nature*. International Publishers, New York
Goldman A I 1994 Argumentation and social epistemology. *Journal of Philosophy* **91**: 27–49
Habermas J 1990 *Moral Consciousness and Communicative Action*. MIT Press, Cambridge, MA
Harvey O J, Hunt D E, Schroder H M 1961 *Conceptual Systems and Personality Organization*. Wiley, New York
Hegel G W F 1812/1967 *The Science of Logic*. Harper and Row, New York
Kant I 1781/1965 *Critique of Pure Reason*. St. Martin's Press, New York [trans Kemp Smith N]
Kramer D A 1983 Post-formal operations? A need for further conceptualization. *Human Development* **26**: 91–105
Kramer D A, Woodruff D S 1986 Relativistic and dialectical thought in three adult age-groups. *Human Development* **29**: 280–90
Mason R O 1969 A dialectical approach to strategic planning. *Management Science* **15**: B403–14
Mitroff I, Mason R O 1981 The metaphysics of policy and planning: A reply to Cosier. *Academy of Management Review* **6**: 649–52
Moshman D 1995 Reasoning as self-constrained thinking. *Human Development* **38**: 53–64
Peng K 1997 Naive dialecticism and its effects on reasoning and judgment about contradiction. Doctoral dissertation, University of Michigan, Ann Arbor, MI
Peng K, Ames D, Knowles E 2001 Culture and human inference: Perspectives from three traditions. In: Matsumoto D (ed.) *Handbook of Cross-cultural Psychology*. Oxford University Press, Oxford, UK, pp. 243–63
Peng K, Nisbett R E 1999 Culture, dialectics, and reasoning about contradiction. *American Psychologist* **54**: 741–54
Peng K, Nisbett R E 2000 Dialectical responses to questions about dialectical thinking. *American Psychologist* **55**: 1067–8
Piaget J 1980 *Experiments in Contradiction*. University of Chicago Press, Chicago
Pruitt D G 1981 *Negotiation Behavior*. Academic Press, New York
Pruitt D G, Carnevale P J 1993 *Negotiation in Social Conflict*. Open University Press, Buckingham, UK
Riegel K F 1973 Dialectical operations: The final period of cognitive development. *Human Development* **18**: 430–43
Schroder H M, Driver M J, Streufert S 1967 *Human Information Processing*. Holt, Rinehart and Winston, New York
Schwenk C R 1990 Effects of devil's advocacy and dialectical inquiry on decision making: a meta-analysis. *Organizational Behavior and Human Decision Processes* **47**: 161–76
Streufert S, Streufert S C 1978 *Behavior in the Complex Environment*. V H Winston, Washington, DC
Tetlock P E 1983 Accountability and complexity of thought. *Journal of Personality and Social Psychology* **45**: 74–83
Tetlock P E 1985 Integrative complexity of American and Soviet foreign policy rhetoric: A time-series analysis. *Journal of Personality and Social Psychology* **49**: 1565–85

K. Peng and D. Ames

Dialectology

Dialectology, the systematic study of linguistic variation, began as an autonomous discipline in the second half of the nineteenth century and became subsumed as a branch of linguistics in the second half of the twentieth century, though not without difficulty. Today it has worldwide currency and hundreds of active participants engaged in diverse research.

1. *Beginnings*

Its beginning is conventionally dated as 1876, when the German professor Georg Wenker mailed a list of sentences written in standard German to schoolmasters in the north of Germany and asked them to return the sentences to him transcribed into the regional dialect of their school districts. Wenker made further mailings until 1887 by which time he had blanketed the entire nation. He ultimately sent his list to nearly 50000 schoolmasters, and he received completed forms from 45000 of them, The schoolmasters, with few guidelines and no precedents, provided Wenker with thousands of regional variants of all kinds: lexical, pronunciation, phonological, morphological, and syntactic. Overburdened by data, Wenker was forced to limit his analyses to a few variants found in a closely circumscribed area of north and central Germany, which he then painstakingly transcribed by hand onto two sets of maps bound as *Sprachatlas des Deutschen Reichs* and deposited in Marburg and Berlin in 1881 (Wenker 1881).

Thirty years before Wenker, the Finnish polymath Anto Warelius (1821–1904), approached the study of dialect from an entirely different perspective in a scholarly undertaking that remains virtually unknown outside his own country, but which has a prior claim over Wenker as the first truly systematic study. In 1846, Warelius set out on foot from Hamina in southeastern Finland and walked in a northwesterly line, eliciting and transcribing words and expressions from the villagers he encountered across almost 400 km (Rapolo 1969, Chambers 2000).

Warelius remains an obscure figure in the history of dialectology, but this method of eliciting dialect forms in face-to-face interviews with villagers foreshadowed the predominant method used by professional dialectologists for the next 150 years. The influential prece-

dent for the method of direct fieldworker elicitation was set in 1896 by Jules Gilliéron, a Swiss professor at the Sorbonne in Paris, in the dialect survey of France. Gilliéron trained an obscure grocer, Edmond Edmont, in the craft of phonetic transcription, and dispatched him into the French countryside on his bicycle. Edmont has become a legendary figure in dialectology for his unflagging devotion as he pedaled into 639 villages and interviewed 700 *paysans* in less than four years. He sent his transcriptions to Gilliéron in Paris at intervals, where they were incorporated into a cumulative mapping enterprise, and publication of the maps began in 1901 and ended in 1910 with the thirteenth volume of the *Atlas linguistique de France* (ALF). Gilliéron and Edmont's atlas (1901–10) stands as an early consummation of dialect geography and became the influential precedent for future projects for almost a century.

2. Impetus

The motivation for these methodical, large-scale studies in the second half of the nineteenth century is conventionally held to be a reaction to Neogrammarian philology, one of the great intellectual currents of the day. The Neogrammarians demonstrated law-like regularities in the historical descent of modern languages from classical ones. One of their tenets was *Ausnahmslosigkeit der Lautgesetze* 'sound changes are exceptionless.' Such a hypothesis, bold and admirably empirical, invites refutation, and a likely source would come from dialect data gathered from highly differentiated, relatively isolated villages.

The view that the first dialectologists rose to the Neogrammarian challenge has romantic resonance, but it is not strongly supported by facts. Neither Wenker nor Gilliéron nor any of their successors actually marshaled their data for elucidation, revision or refutation of any overriding linguistic principles, Neogrammarian or otherwise. In practice, the main results of their scholarship were descriptive rather than analytical, and their expressed goals aspired to no more. The *Sprachatlas* became the preferred medium for dissemination of results, with linguistic tokens carefully arrayed on geographic representations, usually with commentary on secondary variants which had to be omitted because they would overcrowd the map, and occasionally with isoglosses, lines drawn between locations where different variants tended to occur.

Rather than Neogrammarianism, the main impetus for dialectology as a systematic discipline appears to have been Linnaean taxonomy, another forceful current of nineteenth century intellectual history. Dialect data lends itself quite naturally to the taxonomic bent, with numerous gradations subcategorizable as alloforms of -emic categories roughly analogous to genera and their species.

Though taxonomy dictated the form that dialect studies took, the ultimate impetus has a much deeper source in human nature. History is replete with observations about dialect, going back at least three millennia in the Western tradition to the myth of Babel in *Genesis* 11: 1–9, whereby the Lord punished his ambitious subjects by 'confusing their speech, so that they will not understand what they say to one another.' Since Babel there has been a continuous tradition of astute comment on dialect differences which makes a long prelude to the institution of dialectology as a systematic classificatory enterprise. The English scholar Alexander Ellis remarked in 1875 that 'collecting country words is looked upon as an amusement, not as laying a brick in the temple of science' (quoted by Chambers and Trudgill 1998, p. 15). Ellis's spirit thrives, and is evident today in word columns in newspapers, tourist pamphlets on the local 'twang,' stand-up mimicry in theaters, news features on slang, and in dozens of other ways. The discipline of dialectology appears to be the scholarly manifestation of this natural human curiosity about language variation.

3. Methods

Until recently, dialectology gave primacy to recovering the oldest extant forms of language. This deliberate bias colored its methodology at every point. Fieldworkers chose their subjects as a judgment sample of NORMs, an acronym for a prototype who was nonmobile, older, rural, and (predominantly) male. For instance, Edmont's 700 subjects, all of whom were rural, included only about 200 educated people and 60 women. The NORM prototype was purposeful. Lack of mobility guaranteed that their speech would be characteristic of their village. Harold Orton, director of the Survey of English Dialects (SED), England's nationwide survey on Gilliéronian principles for which data was gathered from 1950 to 1961, said, 'dialect-speakers whose residence in the locality had been interrupted by significant absences were constantly regarded with suspicion' (Orton 1962, 15–16). Old age guaranteed that their speech would reflect a bygone era. Hans Kurath, director of the *Linguistic Atlas of New England* (LANE), part of the Linguistic Atlas of the USA and Canada that was begun in 1931 and still continues, declared that 'since most of the informants … are over 70 and not a few over 80, it would seem that we shall be able to establish the regionalism of the pre-industrial era of New England' (Kurath 1949, p. 27). Rurality eliminated mobile urban dwellers who were likely to be exposed to other dialects and accents (Chambers 2000). Males were targeted because their speech is less class-conscious than women's. As Orton put it, speaking about England, 'In this country, men speak vernacular more frequently, more consistently, and more genuinely than women' (1962, p. 15).

The basic methodology of traditional (Gilliéronian) dialect geography is to send trained fieldworkers to the locations where the subjects live in order to elicit specific items from the subjects guided by a standard questionnaire. The questionnaires are lengthy (Gilliéron's no less than 1500 questions, the SED about 1200, LANE about 700), with questions usually ordered by subject matter (flora and fauna, the barnyard, chores, etc.). Its purpose is to guarantee comparable data from all subjects. The fieldworker is expected to transcribe only the key word or phrase from the subject's response, no matter how long it might be. Labov summarized the method as 'a long question from the interviewer and a short answer from the subject' (Labov 1971, p. 113).

These methods have been subject to criticism from commentators versed in social-science methodology, especially with respect to the reliability of the population sample and the validity of the data (for instance, Pickford 1956). Obviously, NORMs constitute a rarefied minority when the majority population is exactly the converse—mobile, younger, urban, and female (Chambers 1993, p. 134). But criticisms like these largely miss the point about the antiquarian bias that motivates the methods. Raven I. McDavid, a venerated fieldworker on the American project and an articulate apologist, defended the traditional methods with this succinct capitulation: 'There is a deliberate bias in the choice of communities—in the direction of smaller and often isolated places; there is a deliberate bias in the selection of informants, insistence on local roots and a statistically disproportionate sample of the older and less educated; there is a deliberate bias in the choice of items for the questionnaire, in the direction of the old fashioned rural society, the horsedrawn vehicle, the mule-powered plow and homemade bread. All of these biases are essential to the purpose of the investigation, to push the historical records as far back in time as possible' (McDavid 1981, p. 71). Even as he articulated this defence, he inadvertently exposed the limits of the tradition at a time when modern technology had eliminated horsedrawn vehicles and social forces were inexorably diminishing the ranks of the uneducated and, indeed, rural society.

4. *Structuralist Analyses*

With the advent of Saussurean structural linguistics in 1916, language studies took a decisive turn in the direction of synchronic analysis. Taxonomy fell out of favor and intensionality gained adherents, at least in theory. However, dialectology clung to its autonomy, no more moved by structuralism than it had been by Neogrammarianism. Over the years, linguists made several attempts at integrating dialectology into linguistics. For instance, Uriel Weinreich (1954) proposed a level of analysis called the diasystem, which in effect posited a level above the phonemic where 'diaphones' could be segmented into the actual phonemes in two (or more) specific dialect systems by means of interpretive rules. Weinreich speculated that speakers must use something comparable to a diasystem for comprehension when they interact with speakers of other dialects.

Later, with the development of generative frameworks, there were similar attempts at integrating dialectology. For example, Keyser reanalyzed traditional atlas data from the Atlantic coast of the USA by demonstrating that several vowel distinctions found in the painstaking tabulations for dozens of speakers could be generated simply by positing two phonological rules and 'accounting for the formation of new dialect groups through the geographical dissemination of rules' (Keyser 1963, p. 313). Efforts like these gained few adherents, and they remained peripheral to the linguistic mainstream and had little impact on dialect studies.

More successful were the uses of atlas data in elucidating problems of historical linguistics. For instance, Hall (1943) used data collected in the *Sprach- und Sachatlas Italiens und der Südschweiz* ([AIS] Jaberg and Jud 1928–40) to demonstrate the persistence of the old political boundary of the Papal States on patterns of dialect differentiation in twentieth-century central Italy. The purposes of historical linguistics were of course consonant with the intentions of traditional dialectology, although the exploitation of the latter by the former offered much greater potential than was actually realized. It did, nevertheless, establish a vital, if minor, tradition.

5. *Decline and Rebirth*

As post-Saussurean language studies continued to be dominated by synchronic linguistics and entered a period of considerable currency stimulated by the ideas of Noam Chomsky, the failure of dialectology to become integrated caused it to slip ever further into the periphery of language studies. By the middle of the twentieth century, scholarly activities had slowed and become relatively inconspicuous. By 1980, even dialectologists began to realize that reform was necessary. In the first book-length summation of the field, Chambers and Trudgill stated: 'one might well wonder what the future holds for dialect geography. There is no doubt whatever that the financing of research and publishing for large-scale surveys is a great problem, and there is also no doubt that capable and enthusiastic practitioners of the calibre of those who went before are difficult to find; but it seems most likely that both of these facts follow from another fact, namely, that dialect research has taken a new direction … It is probably safe to say that the future of dialect geography depends upon the ability of its practitioners to embrace and incorporate the concerns

and perhaps the methodology of urban dialectology' (Chambers and Trudgill 1980, p. 23).

Many aspects of dialect geography came under criticism when they were noticed at all, including the utility of the *Sprachatlas* as a medium for disseminating results. 'For whom are these expensive volumes designed?' Macaulay asked, and then he said: 'Far from being a visual aid the maps are probably a barrier to grasping the pattern of distribution and it seems very unlikely that anyone nowadays would attempt this kind of mapping. Given the sorting power of computers, basic information is much better stored in a form that is suitable to machine processing than in conventional map form' (Macaulay 1985, p. 180).

With the founders gone, the rearguard reaction to these criticisms was muted. Starting in the mid-1980s, dialectology underwent a renaissance from which it would emerge as a vital and vigorous discipline with greater worldwide currency than ever before and, not coincidentally, greater intellectual depth. The sources were the antidotes identified in the two critiques above: the embracing of powerful new technology and the adaptation of more inclusive research methods.

6. *Computerization*

From its inception, dialectology had been burdened by an overabundance of data. Dialectology aspired to be, in Saussurean terms, a science of *parole*, that is, the study of language as a communal effusion, apprehended in actual speech events, imbued by and indeed mediating what Sapir called 'the apparent lawlessness of social phenomena' (Sapir 1929, p. 166). But the profusion of data proved daunting, and the methods used in classic dialect geography came into being partly as devices for delimiting the effusion and subjugating its lawlessness, by focusing on a narrow, linguistically conservative population, restricting elicitation to a closed set of items, and transcribing only individual tokens from the flow of speech.

The development of tape-recording as a reliable, affordable technology in the 1950s provided a means for making a permanent record of the flow of speech. However, the immediate impact of tape-recording on dialectology was negligible, coming into use merely for recording traditional elicitation sessions and thus serving mainly to provide archival documentation of the practices of McDavid and other fieldworkers. The most innovative of all regional surveys, the *Linguistic Atlas of the Gulf States* (Pederson et al. 1986–92 [LAGS]), directed by Lee Pederson, made a significant advance by recording all its interviews so they could be transcribed later under laboratory conditions. Meanwhile, beyond the formal confines of dialectology, tape-recording was embraced as a liberating medium by a few linguists genuinely interested in studying vernacular speech. In this way, tape-recording stimulated the rise of sociolinguistics, an urban dialectology with methods and goals that bore little resemblance to dialect geography.

More significant technology for the revitalization of dialectology came with the development of computers as reliable data-handling devices in the late 1960s, and especially as affordable and accessible universal processors in the 1980s. As Kretzschmar and colleagues put it: ' … the development of dialect studies, whether geographical or sociolinguistic, has always been hampered by a superfluity of data … Even smaller surveys have had to settle for selective analysis of their data because the wealth of possibilities for analysis overran the editors' time and the human capacity for holding in the mind only so much information at once. Computers can help overcome these problems: they are wonderful tools for quickly sorting and matching pieces of information and for performing complex calculations on the results, and … they are practically unlimited in their ability to store data' (Kretzschmar et al. 1989, p. v). Many dialectologists recognized the computers' potential for relieving the perpetual problem of overabundance. No less important, many scholars recognized the enormous advantages that would accrue from the computerization of the classic databases of the ALF (Heap 1997), SED (Ramisch 1997, Elmer and Rudin 1997), AIS (Goebl 1993), and the others. Generalizations buried in lists and tables for many decades could be disinterred by judicious computation.

7. *Dialectometry*

A significant development directly attributable to computational processing is dialectometry, literally, the 'measure of dialect.' The word *dialectométrie* was coined by Pierre Séguy, director of the *Atlas Linguistique de Gascogne*, one of the regional surveys comprising the *Nouvel Atlas de la France*. Séguy hoped to replace subjective judgments on the occurrence of dialect regions by quantifying the Gascony data in such a way that the regions would be revealed objectively—in effect, they would reveal themselves. In the sixth and final volume of the Gascony atlas (Séguy 1973), Séguy and his associates—*mes élèves-camarades*, he called them—published a set of maps, which linked each of the survey sites to its nearest neighbors by a number representing the percentage by which the two sites differed with respect to the forms elicited there. Séguy conceived the percentages as an index for measuring the 'linguistic distance' between the two places. The analyst, by inspecting the distance indices, could draw inferences about dialect regions based on degrees of similarity and difference.

Though Séguy's methods now look primitive, restricted as they were at the time to univariate comparisons and hand-made cartography, the numerical gradations across the maps rightly impressed many dialectologists as the first approximation of a visual

analogue for the dynamics of dialect gradation in space. Since then, dialectometry has advanced to multivariate statistical programs and computer-generated multidimensional displays that simultaneously compare all sites to one another. Dialectometry has become a generic term for all kinds of statistical dialect geography linked, as Goebl says, by a common goal: 'The preponderant aim of dialectometry consists in discovering, by the numeric combination of many low ranking patterns, higher ranking patterns which have remained hitherto hidden, in order to obtain a systematic insight into the problem of the basilectal management of space by HOMO LOQUENS' (Goebl 1993, p. 277). Practitioners of dialectometry can be found around the world from Japan to Canada. The numerous research programs remain largely uncoordinated and partly incompatible, but in the next decades the consolidation of methods and goals will undoubtedly crystallize into a powerful branch of dialectology unimaginable in the first century of the discipline.

8. *Sociolinguistic Dialectology*

The other revitalizing impulse for dialectology is sociolinguistics, which had its effective beginning in the 1960s with the dissemination of William Labov's fountainhead studies in Martha's Vineyard and New York's Lower East Side (collected in Labov 1972). Linguists began studying language variation with freer data-gathering methods, using larger and more representative subject samples, and they sought to understand its social significance with quantitative, acoustic and statistical analyses. The intellectual forebears of sociolinguistics, far removed from Linnaean taxonomy, were post-structural linguistics and the social sciences.

Dialect geographers, not surprisingly, did not immediately embrace the new currents, and it took a generation schooled primarily in sociolinguistics to recognize that the two disciplines were both, in essence, dialectologies. With hindsight, the rise of sociolinguistics can be viewed as a natural response to global social forces that have arisen since dialectology came into being, including urbanization, mobility of all three kinds—geographical, occupational and social—embourgeoisement, mass education and mass literacy. Those forces shrank the traditional domains of dialect geographers by reducing geographically isolated regions and socially isolated subjects to statistical insignificance in all the developed nations and beyond.

Observations of linguistic variation became refocused on social class, network, age, sex, ethnicity, and other social sub-groups as well as region, and the systematicization of those observations required new sets of independent variables and new methods for tracking them. In the social context that came into existence a century after Warelius and Wenker, the remaking of dialectology as sociolinguistic dialectology was inevitable. As Chambers and Trudgill said, 'A decade or two ago, it might have been possible to think that the common subject matter of dialectology and sociolinguistics counted for next to nothing. Now we know it counts for everything. Dialectology without sociolinguistics at its core is a relic. No serious perspective on dialectology can grant urban research and variation theory less than a central role' (Chambers and Trudgill 1998, p. 188).

One of the historical consequences of dialectology becoming sociolinguistic is that it has become an integral branch of linguistics. Like other branches of linguistics, its seminal works and formative investigations have been impelled and largely defined by specific issues, that is, by the need to develop or refine or refute particular hypotheses. The defining issues include these (discussed in Chambers 1993), among others: (a) How do the changes dialectologists observe in progress reflect the changes that occur in the history of languages? (b) How, if at all, do the grand concepts of the sociology of language—concepts such as diglossia, language shift, heteronomy, linguistic legislation and planning—affect social interaction, or become encoded emblematically in sociolinguistic variants? (c) How do dialect features diffuse throughout regions, and how does such diffusion affect the linguistic behavior of people in the community? (d) What is the adaptive function of dialect and accent? This last question, in particular, can be posed as a kind of ultimate question for all dialect studies.

There is a long tradition that presupposes dialect and accent to be dysfunctional, starting with the Babelian myth. Dialectology in all its methodological and intellectual guises shows that, on the contrary, in every human society linguistic variation prevails and thrives. It is a profoundly human attribute that is essentially irrepressible. The large, open-ended questions of dialectology can only begin to find their answers in fine-grained, minute research into the social significance of linguistic variation. The answers to those questions will ultimately elucidate the human language faculty. With that as its goal, dialectology constitutes a vital branch of modern linguistics.

See also: Historical Linguistics: Overview; Language and Ethnicity; Linguistic Fieldwork; Regional Identity; Sociolinguistics

Bibliography

Chambers J K 1993 Sociolinguistic dialectology. In: Preston D (ed.) *American Dialect Research*. John Benjamins, Amsterdam/Philadelphia, pp. 133–64

Chambers J K 2000 Region and language variation. *English World-Wide* **21**: 169–99

Chambers J K, Trudgill P 1980 *Dialectology*. Cambridge University Press, Cambridge, UK

Chambers J K, Trudgill P 1998 *Dialectology*, 2nd edn. Cambridge University Press, Cambridge, UK
Elmer W, Rubin E 1997 The survey of English dialects as an electronic database for research in areal and variationist linguistics. In: Thomas A R (ed.) *Issues and Methods in Dialectology*. Department of Linguistics, University of Wales, Bangor, Wales, UK, pp. 234–46
Gilliéron J, Edmont E 1901–10 *Atlas Linguistique de la France*. Champion, Paris, 13 Vols [ALF]
Goebl H 1993 Dialectometry: a short overview of the principles and practice of quantitative classification of linguistic atlas data. In: Kohler R, Rieger B B (eds.) *Contributions to Quantitative Linguistics*. Kluwer Academic, Amsterdam, pp. 277–315
Hall R A Jr 1943 The papal states in Italian linguistic history. *Language* **19**: 125–40
Heap D J 1997 *La Variation grammaticale en géolinguistique: les pronoms sujet en roman central*. Department of Linguistics, University of Toronto: Toronto Working Papers in Linguistics
Jaberg K, Jud J 1928–40 *Sprach- und Sachatlas Italiens und der Südschweiz*. Ringier, Zolfingen [AIS]
Keyser S J 1963 Review of Kurath and McDavid, *Pronunciation of English in the Atlantic States*. *Language* **39**: 303–16
Kretzschmar W, Schneider E, Johnson E (eds.)1989 *Computer Methods in Dialectology*. *Journal of English Linguistics* **22**: special edition
Kurath H 1949 *Word Geography of the Eastern United States*. University of Michigan Press, Ann Arbor, MI
Labov W 1971 Some principles of linguistic methodology. *Language in Society* **1**: 97–120
Labov W 1972 *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia
Macaulay R K S 1985 Linguistic maps: visual aid or abstract art? In: Kirk J, Sanderson S, Widdowson J D A (eds.) *Studies in Linguistic Geography*. Croom Helm, London, pp. 172–86
McDavid R I 1981 On obtaining spontaneous responses. In: Warkentyne H. J. (ed.) *Papers from the 4th International Conference on Methods in Dialectology*. University of Victoria, British Columbia, pp. 66–84
Orton H 1962 *Survey of English Dialects: Introduction*. E. J. Arnold; Leeds, UK (SED)
Pederson L et al. 1986–92 *The Linguistic Atlas of the Gulf States*. University of Georgia Press, Athens, GA, Vols 1–7 [LAGS]
Pickford G 1956 American linguistic geography: a sociological appraisal. *Word* **12**: 211–33
Ramisch H 1997 Dialectological and cartographical features of the computer developed linguistic atlas of England. In: Thomas A R (ed.) *Issues and Methods in Dialectology*. Department of Linguistics, University of Wales, Bangor, Wales, UK, pp. 224–33
Rapola M 1969 *Johdatus Suomen Murteisiin*. Suomalaisuuden Kirjallisuuden, Seura
Sapir E 1929 The status of linguistics as a science. *Language* **5**: 207–14
Séguy J 1973 *Atlas linguistique de Gascogne*. Centre national de la recherche scientifique, Paris, Vol. 6
Weinreich U 1954 Is a structural dialectology possible? *Word* **10**: 388–400
Wenker G 1881 *Sprachatlas des Deutschen Reichs*. MS folio. Marburg University

J. K. Chambers

Diaspora

This article considers the origin and changing use of the notion of diaspora, showing how, despite its specific reference to the history of the Jews, it has now come to be important in understanding many forms of international migration. How modern nation-states respond to the challenges of social and cultural pluralism is discussed. The different types of diaspora and diverse ways in which the concept is now used are also identified.

1. *Origins and Development*

The word 'diaspora' is derived from the Greek word *speiro* (to sow) and the preposition *dia* (over). When applied to humans the ancient Greeks thought of the word as signifying expansion through outward migration and settlement. 'Diaspora' soon acquired a more brutal and catastrophic meaning. In the Greek translation of the Old Testament it was used to describe the forcible dispersion of the Jews. The destruction of the Temple in Jerusalem in 586 BC led to the enslavement and displacement of the key military, civil, and religious leaders of Judah and their exile in Babylon. This fate was held to be predicted in Deuteronomy (28: 58–68) where God had warned that anybody who disobeyed his law would be scattered to all ends of the earth. There they would 'find no peace,' while they would additionally suffer 'an unquiet mind, dim eyes and a failing appetite.' 'Babylon,' then the capital of the Mesopotamian empire, become the code word for Jews (and later for Africans in the New World) signifying a place of loneliness and estrangement, where they were held in captivity.

This catastrophic meaning of diaspora, which may be described as the 'victim tradition,' was one that informed the historical and popular use of the term for nearly two millennia. Gradually, other groups saw themselves as victims of forcible dispersion in a manner similar to the ancient Jews and used the expression 'diaspora' as a self-description of their fate. Armenians, for example, allude to the period 1894–6 when their national movement was violently opposed by the Ottoman Sultan Hamid and some 300,000 Armenians were killed in Turkish Armenia. In 1915–16 'close to one million people' (Melson 1992, p. 146), about half the Armenian nation, were either killed or died of starvation. Armenian scholars and activists have accused the Turkish authorities of genocidal intentions, an interpretation that is generally accepted, though still hotly disputed by the current Turkish government. The subsequent dispersal of the Armenians to the USA, Iran, France, Lebanon, and other countries is now commonly described as a diasporic experience. The brutal Atlantic slave trade for New World Africans, the potato famine for the Irish, and (ironically) the return to Palestine of the Zionist part

of the Jewish diaspora for Palestinians are similarly considered the crucial traumatic events that triggered the creation of these other victim diasporas.

The above examples have an apparent theoretical and experiential coherence. A tragedy perpetrated by a hostile exogenous force generates movement to a number of countries. But because the displacement was forced and the traumatic event so encompassing, collective folk memories are developed, the populations remain psychologically unsettled, and a return movement gains some following. A number of conventional historians would wish to go no further with the idea of diaspora. However, from the 1980s, this relatively well-established meaning began to lose definition at its edges. Many ethnic groups of varying histories defined themselves as 'diasporas,' or were so recognized by scholars or journalists. The decisive intellectual movement for this new pluralism was marked by the launch of a journal called *Diaspora*, issued by the New York office of Oxford University Press in 1991 and edited by a US-based literary scholar of Armenian provenance. His opening manifesto was an impressive call for openness. The journal would be concerned with 'the way in which nations, real but imagined communities, are fabulated, brought into being, made and unmade, in culture and politics, both on the land people call their own and in exile' (Tölölyan 1991, p. 3). By 1998, the journal had published articles centrally identifying 34 ethnic groups as diasporas. Considerably more ethnic groups were more casually referred to as 'diasporas.'

2. *Types of Diaspora*

Natural scholarly caution would invite rejection of this profligacy. Yet, in fact, the pluralists have a good case. The paradigmatic example of the Jews, for example, shows that the communities abroad often arose from voluntary, not forced, movement, while the Babylonian experience can be characterized far more positively—as a place where new creative energies were released and challenging cultural encounters arose over a long period in a largely peaceful setting. Moreover, as has been mentioned, the original Greek meaning of diaspora allowed voluntary movement or even aggressive colonial settlement. Drawing on this etymology it would, for example, be perfectly consistent to allude to the British colonies in New Zealand, Australia, and Canada as comprising part of 'the British diaspora'—of an imperial, rather than victim, type.

In addition to the victim and imperial types, scholars have identified labour, trade, business, professional, religious, cultural, ethno-national, and refugee diasporas, as well as intermediate forms such as incipient, asylum, and 'new' diasporas (Safran 1991, Sheffer 1995, McDowell 1996, Cohen 1997, Van Hear 1998). These subtypes describe the predominant characteristic or characteristics of each diaspora, but cannot represent adequately the full complexity of each group in terms of its history, mixed motivations for emigration, and class diversity. It is apparent, for example, that the 50 million 'Chinese Overseas,' who have only recently been described as a diaspora, include people whose origins lay in semi-free 'coolie labor,' as well as those who migrated as entrepreneurs, traders, and skilled professionals.

Beyond the new pluralism, the key difference between 'diaspora' and more popularly recognized forms of migration is that in the first case cultural, linguistic, religious, historical, and affective ties with the place of origin remain strong. In the second, at least in theory, immigration from 'an old country' involves a one-way ticket, assimilation to the 'new country,' the adoption of a local citizenship and language, and the public acceptance of local ways and customs. As many studies of ethnicity in the USA and elsewhere have shown, in culture and endogamous practices the old ways were more resilient than public ideology proclaimed. Despite this finding, the myth and goal of full integration remain at the heart of the nation-building project and inform the difference between immigration in general and the specific case of diasporic migrant groups.

3. *Responses by Nation-states to Diasporas*

During the period of intense nation-building in the core European, South, and North American states a diasporic consciousness was either not sanctioned or firmly suppressed. The US state's motto '*Ex Pluribus Unum*' provided an example of the prevailing preference for assimilation or the creation of a new national culture. However, under the impact of diverse and accelerating immigration a number of important nation states (Canada, the USA, the Netherlands, Australia, and the UK might be cited) have lost their appetite and perhaps the capacity to assimilate immigrants to a single dominant political culture. By recognizing ideas of multicultural, multifaith, and multiracial societies, their political leaderships have reversed their previous disapprobation of the retention of pre-immigration ethnic identities.

Many social commentators are relaxed about this development and see no incompatibility between a national and an ethnic affiliation. Others, including those from the traditional liberal side of the political spectrum are anxious that the painfully achieved democratic and secular order in societies such as the USA will be undermined by the recognition of diasporas (Dickstein 1993). Even more alarming is the expressed fear that diasporas will become the means to foster religious zealotry, the international drugs trade, political instability, and cross-border terrorism. The involvement of exile groups, like the Tamils, Kurds, Sikhs, or Afghans, in the political life of their homelands or in movements to create nation states, has often created violent conflicts or embarrassing and difficult diplomatic and legal conflicts (Sheffer 1986).

By contrast, Shain (1999) argues that the democratic and liberal ways learnt in a the place of settlement are more likely to spread back into the countries of origin, with benign political effects for all parties.

The attitude of the countries of origin to their diasporic communities abroad has not gained much systematic discussion, though these states are increasingly important players in retaining and amplifying the links between 'home' and 'away.' Historically, countries like China and India have been hostile to their overseas communities, seeing them as painful reminders of colonial days when their labor forces could be captured by labor recruiters or, in the contemporary period, as 'brain-drain' migration. India, China and many other migrant-exporting countries now recognize that their diasporas can provide an important political lobby for advancing their countries' interests (American Jews seeking to influence US attitudes to Israel are cited as a positive example), a bridgehead for economic marketing of the home countries' goods, a means for acquiring transferable skills (for example in information technology) and, above all, a rich source for remittance and investment income. The Philippines has an expertly-run, state-supported set of agencies that promote labor export on a vast scale, help to channel return income, and seek to protect their workers from exploitative employers abroad. A diasporic consciousness is thus often preferred by a particular migrant group, tolerated by many host nation-states, and strongly supported by the country of origin. This is a significant change of policy and practice compared with the situation facing immigrants at the beginning of the twentieth century.

4. *Forms of Diaspora*

In addition to typologizing diasporas by examining their general characteristics, scholars have sought to classify the different 'forms' in which the term has been used, particularly by the different academic disciplines. A generally successful fourfold scheme has been proposed by Vertovec (1997) and modified in Vertovec and Cohen (1999).

The first form, advanced by sociologists and business studies specialists, is 'diaspora as social form,' alluding to the maintenance and development of transnational social organizations and networks, a collective identity, orientations to homeland and particular kinds of economic strategy (Safran 1991, Kotkin 1992, Cohen 1997). The second form is 'diaspora as consciousness.' Cultural studies theorists as well as those working in the fields of ethnic and race relations found in the notion of diaspora a means of breaking free from the 'essentialist' discourse that previously informed the subject matter. Notions of what constituted ethnicity became more fluid, hybrid, and negotiated, and 'diaspora' was a way of mapping how complexity and difference arouse as cultures travelled, interacted, and mutated (Clifford 1992, Brah 1996). The third is 'diaspora as mode of cultural production.' Though closely related to the second, this form is used by anthropologists to show how social and cultural formations, as well as cultural objects, images, and meanings have leaped from their old locales and crossed national lines (Hannerz 1996, Basch et al. 1994). Finally, 'diaspora as political orientation' has been considered by political scientists and international relations theorists. As noted earlier, such specialists concentrate on how the attachment to the politics of homeland can negatively and positively impact on the countries at both ends of the migration chain.

5. *Future of the Concept*

The concept of diaspora has firmly escaped its historical confines in the Jewish experience and at the beginning of the twenty-first century is used in a variety or settings and for a variety of purposes. The enduring logic is that the term conceptually connects home and abroad. The conflicting claims on group loyalties have been given an added emphasis in an age of global networks and the changing role, some would argue the relative decline, of the nation-state as a focus for group identification. 'Diaspora' will remain as an important means of interrogating and exemplifying this tension. It will also provide a useful means of describing the enhanced role for cyclical, transversal, and intermittent international migration.

Beyond this, the future of the term may be imperilled by having been commandeered to serve too many discrepant purposes and its use to describe cognate, yet different, phenomena. The use of 'diaspora' to describe transnational religious communities (as in 'Muslim diaspora') provides one example. In some cases, such as the Jews, Parsis, and Sikhs, religion, ethnicity, minority status, and peoplehood are so closely interpenetrated that the description 'diaspora' is apposite. 'Hindu diaspora' is acceptable, but undervalues the ethnic diversity under the expression 'Hindu.' However, 'Indian diaspora,' which expression is now in common use, misleadingly elides the differences between the three principal faith communities (Hindu, Muslim, and Sikh) and ignores well-organized national minorities (like the Kashmiris), with potentially confusing effects. The idea of a 'Turkish diaspora' presents similar issues. In addressing such problems, there may be some virtue in using the expression 'transnational communities' to characterize a number of different postnational and non-national forms of sociality and association. Examples so categorized can be virtual Internet communities, Christianity, Islam, transnational social movements, and international professional associations. Within the rubric of transnational communities can shelter 'diaspora' as a particular form of ethnically defined community. Some degree of self-restraint may, in

other words, prove necessary if the term is to retain its conceptual purchase.

See also: Assimilation of Immigrants; Creolization: Sociocultural Aspects; Hybridity; Jewish Diaspora; Migrations, Colonizations, and Diasporas in Archaeology; Multiculturalism; Multiculturalism, Anthropology of; Multiculturalism: Sociological Aspects

Bibliography

Basch L, Glick Schiller N, Szanton Blanc C 1994 *Nations Unbound: Transnational Projects, Post-Colonial Predicaments, and Deterritorialized Nation-states*. Gordon & Breach, Langhorne, PA

Brah A 1996 *Cartographies of Diaspora*. Routledge, London

Clifford J 1992 Traveling cultures. In: Grossberg L, Nelson C, Treichler P A (eds.) *Cultural Studies*. Routledge, New York, pp. 96–116

Cohen R 1997 *Global Diasporas: An Introduction*. University of Washington Press, Seattle, WA

Dickstein M 1993 After the Cold War: Culture as politics, politics as culture. *Social Research* **60**(3): 531–44

Hannerz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London

Kotkin J 1992 *Tribes: How Race, Religion and Identity Determine Success in the New Global Economy*. Random House, New York

McDowell C 1996 *A Tamil Asylum Diaspora*. Berghahn, Oxford, UK

Melson R 1992 *Revolution and Genocide: On the Origins of the Armenian Genocide and the Holocaust*. University of Chicago Press, Chicago

Safran W 1991 Diasporas in modern societies: Myths of homeland and return. *Diaspora* **1**(1): 83–99

Shain Y 1999 *Marketing the American Creed Abroad: Diasporas in US and their Homelands*. Cambridge University Press, Cambridge, UK

Sheffer G (ed.) 1986 *Modern Diasporas in International Politics*. Croom Helm, London

Sheffer G 1995 The emergence of new ethno-national diasporas. *Migration* **28**: 5–28

Tölölyan K 1991 Preface. *Diaspora* **1**(1): 3–7

Van Hear N 1998 *New Diasporas: The Mass Exodus, Dispersal and Regrouping of Migrant Communities*. University of Washington Press, Seattle, WA

Vertovec S 1997 Three meanings of 'Diaspora' exemplified among South Asian religions. *Diaspora* **6**(3): 277–99

Vertovec S, Cohen R (eds.) 1999 *Migration, Diasporas and Transnationalism*. Edward Elgar, Cheltenham, UK

R. Cohen

Dictatorship

The term 'dictatorship' is one of Roman derivation, and originally signified the temporary suspension of constitutional constraints on governmental power and its unlimited transfer to one person. Among the Romans, there was provision for a temporary suppression of citizen rights during times of crises, and the equally temporary transfer of unlimited power, to a person chosen by the Senate and appointed by the Consul. At the end of a specified period—usually six months—the person so endowed, the dictator, surrendered power, and citizen rights were restored. The period of emergency stewardship was then duly reviewed, in accordance with constitutional requirements. During instances of emergency, over three centuries of republican constitutional rule, Rome appointed 88 dictators in conformity with law.

Following the rule of Julius Caesar, who exercised dictatorial powers for an unconstitutional length of time, the institution disintegrated and rulers assumed extraconstitutional powers in permanence. Rather than a traditional dictatorship, such rule took on the properties of what might more appropriately be called 'autocracy,' 'despotism,' or, perhaps, 'tyranny.'

Throughout the Middle Ages and more modern times in Europe, as well as during the institutionalization of dynastic rule in East and South Asia, 'autocracies' tended to be regularized through the agency of hereditary right—providing for succession without recourse to violence. Emperors, kings, and sultans could, and frequently would, allocate powers to 'noble' subordinates through a relatively well-articulated, rule-governed process. At times, as was the case in the rights embodied in the Magna Carta, some substantial rights were accorded subordinates.

1. The Modern Period

With the decline of the Middle Ages, it became increasingly common that adventurers would seize power in emerging city-states and principalities, and rule as despots or benevolent autocrats. The term 'dictatorship' was rarely, if ever, employed to identify such rule until the commencement of the modern period.

In regions such as Latin America, the term came into increasing vogue after the liberation from Spanish colonialism, irrespective of the fact that the ideal, throughout Central and South America—influenced by the North American model—was constitutional democracy and political stability. The basic charters of most of the nations of Central and South America had incorporated broad versions of the civil, political, and economic rights of citizens to the point of extravagance (Davis 1961). Unfortunately, governments in Latin America, throughout the remainder of the nineteenth and the beginning of the twentieth century, regularly came to power through violence and were commonly dislodged by revolution. During a 25 year period, from 1930 to 1955, for example, more than 50 violent, mainly military, revolutions took place in the region—to provide an impressive roster of

'military dictatorships' that ruled for variable periods of time (Lieuwen 1960).

In Africa, military dictatorships became the norm in the decades after the decolonization of European holdings that followed the conclusion of the Second World War. The African military—having a monopoly of coercion at its disposal, and often the only institution having logistical reach over the entire territory—acceded to absolute political power over inhabitants. Control, unconstrained by constitutional limitations, was more frequently than not secured by violence or the threat of violence.

In fact, the term 'dictatorship' (as distinct from terms like '*caudillismo*' used in Latin America) came into increased currency in Europe long before its employment elsewhere. At the commencement of the modern period, it came to be used, for example, to designate what was initially understood to be a temporary assumption of full power by what was called the Directorate under Cromwell in England, and in 1793, by the Committee of Public Safety, during the French Revolution.

After the conclusion of the First World War, the term was invoked to distinguish a pattern of rule by 'dictators' such as Benito Mussolini in Italy, Primo de Rivera in Spain, or Mustafa Kemal in Turkey. Actually, what such dictatorships signaled was a development that was far more portentous than simply the identification of the exercise of political power by a single ruler in any particular country. What was emerging was a singular political system, having a relatively specific syndrome of properties, that can be legitimately identified as 'modern dictatorship' (Talmon 1960).

2. *Modern Dictatorship*

The dislocations that attended the end of the First World War, gave rise to a wave of revolutionary ferment that involved Eastern and Southern Europe (Lee 1987). In Russia, revolutionaries first overthrew the Romanov monarchy, and then the provisional democratic government of Alexander Kerensky, in order to impose, in 1917, a 'dictatorship of the proletariat' under the leadership of V. I. Lenin.

According to the tenets of classical Marxism (the Marxism of Karl Marx and Friedrich Engels, the dictatorship of the proletariat was to be a transitional form of minority rule that marked the distance between 'bourgeois democracy'—what Marxists considered to be, in reality, a 'dictatorship of the bourgeoisie'—and communist democracy (see *Communism*). Lenin assumed full power over the emerging political system that followed the 'proletarian' revolution (see *Revolution*) by legitimizing it through an appeal to an ideological rationale (Furet 1999).

At approximately the same time, in 1922, Benito Mussolini led a collection of revolutionaries against the parliamentary and monarchial government of peninsular Italy (De Felice 1965). Initially what appeared to be a transitional government, calculated to deliver the nation from the crises that followed the First World War, the government of Mussolini quickly took on the features of a modern dictatorship. Through reforms commencing as early as 1925, Mussolini's Fascist regime assumed the properties of permanent minority rule under single-party auspices (De Felice 1968).

Lenin's proletarian dictatorship and Mussolini's Fascism provide a syndrome of traits that specifically characterize modern dictatorship (Johnson 1991). In both cases, a charismatic leader exercises almost total control over followers as well as the national community out of which they emerge. Leadership in both cases is epistemarchic in character. That is to say, leadership, in the absence of popular elections, is legitimatized by the leader's putative possession of an ideology (see *Ideology: Political Aspects*) that is an inerrant guide to policy, satisfying the needs of the collectivity (respectively the proletariat, or the nation). The inerrant ideology informs the belief system of an organization of believers, usually identified as a 'vanguard' of the revolutionary truth. In the course of time, that vanguard becomes the 'unitary party' of the political system—to the exclusion of alternatives. The inerrancy of the ideology and its attendant policy, in turn, is used to justify control over any and all political dissidence (which, under the circumstances, could only be in error) and any initiatives independent of the vanguard leadership (Manoilescu 1941).

In general, modern dictatorships maintain a military adjunct to unitary party rule. In the Bolshevik case that adjunct was the Red Army. In the case of the Fascist revolution, that adjunct was the Blackshirt militia. In the case of Adolf Hitler's National Socialism, that adjunct was the Schutz Staffel While military force remains an ultimate recourse of the system, modern dictatorships tend to be populist in disposition. They seek plebiscitary confirmation through controlled elections. Leadership, in modern dictatorships, because it is characteristically 'charismatic,' enjoys life tenure. The Democratic People's Republic of Korea (North Korea), as a case in point, exemplifies such a system. After the Second World War, Castro's Cuba, and Mao's China constituted illustrative instances. The system under the rule of Gamal I Abd' El-Nasser in Egypt, and Ghana under Kwame Nkrumah took on many of the same features. They took on the properties of a modern dictatorship however much their actual practice failed to fully satisfy the criteria for admission into the class.

In most cases, modern dictatorships, whatever their initial character, took on the features of an emphatic nationalism. Having begun as an ideology predicated on class, Bolshevism, as an instantial case, very quickly made recourse to 'patriotism' to sustain popular commitment. All classes and segments of classes in the

Soviet Union were expected to be supportive of what were unmistakably national purposes, however 'proletarian' they were supposed to be.

Similarly, given the traits exemplified by such a class of polities, they all failed to allow for what parliamentary and constitutional systems understand to be voluntary association, formal opposition, free elections, the separation of legislative, judicial and executive powers, as well as popular freedom of expression and individual movement. In almost all cases, the charter constitutions of such systems change with some regularity (the constitutions of the People's Republic of China, for example, have changed with every change of leadership) and their provisions are subject to unconstrained interpretation (the provisions of the constitution of Fascist Italy, for example, were subject to the regular reinterpretation of Mussolini and the Grand Council of Fascism).

3. Modern Dictatorship and Totalitarianism

By the beginning of the 1930s, a subset of modern dictatorships began to evince characteristics that led political scientists to design them a specific identity: totalitarianisms (see *Totalitarianism*). Possessed of an elaborate security and control infrastructure, supported by a factitious or real consensus, and animated by an irrepressible conviction of their own aggressive moral superiority, such politically repressive systems destroyed millions of innocents, either as a measure of control, or in the pursuit of utopian intention (Horowitz 1997) .

The numbers killed by Stalin, Hitler, Mao, and the Khmer Rouge will probably never be accurately known—but the victims of Marxist–Leninist (see *Marxism/Leninism*) regimes probably exceed one hundred million (Courtois et al. 1999). 'Mild' totalitarianisms—like those of Mussolini's Italy, Chiang Kai-shek's China, or Castro's Cuba—neither particularly homicidal nor genocidal—employ incarceration and exile, together with the characteristic denial of voluntary association and freedom of expression, as means of control.

Whatever the subsequent qualifications on the use of the term 'totalitarianism,' throughout the twentieth century an imprecise subset of modern dictatorships have shared a number of reasonably well-defined traits that qualify them for entry into a specific class of modern political systems (Schapiro 1972, Soper 1985, Gleason 1995). How far into the twenty-first century exemplars of the totalitarian subset of modern dictatorships will persist remains a matter of conjecture.

That such systems can evolve into something like a democratic alternative has been established by the transformation of Chiang Kai-shek's Kuomintang on Taiwan into something very much like a vital democracy. The massive reforms on the mainland of China, that followed the death of Mao Zedong, suggest that such systems have the potential to undergo systemic change. How such change is effected remains uncertain.

See also: Authoritarianism; Communist Parties; Latin American Studies: Politics; Maoism; Military and Politics; National Socialism and Fascism; Socialist Societies: Anthropological Aspects; Spatial Data Infrastructure; Totalitarianism: Impact on Social Thought

Bibliography

Courtois S, Werth N, Panne J-L, Paczkowski A, Bartosek K, Margolin J-L 1999 *The Black Book of Communism: Crimes, Terror, Repression*. Harvard University Press, Cambridge, MA

Davis H E 1961 *Latin American Social Thought Since Independence*. University Press, Washington, DC

De Felice R 1965 *Mussolini il rivoluzionario 1883–1920*. Giulio Einaudi, Turin, Italy

De Felice R 1968 *Mussolini il fascista: L'organizzazione dello Stato fascista 1925–1929*. Giulio Einaudi, Turin, Italy

Furet F 1999 *The Passing of an Illusion: The Idea of Communism in the Twentieth Century*. University of Chicago Press, Chicago

Gleason A 1995 *Totalitarianism: The Inner History of the Cold War*. Oxford University Press, New York

Horowitz I L 1997 *Taking Lives: Genocide and State Power*. Transaction, New Brunswick, NJ

Johnson P 1991 *Modern Times: The World from the Twenties to the Nineties*. Harper, New York

Lee S J 1987 *The European Dictatorships 1918–1945*. Methuen, London

Lieuwen E 1960 *Arms and Politics in Latin America*. Praeger, New York

Manoilescu M 1941 *Die einzige Partei als Institution der neuen Regime*. Otto Stollberg, Berlin

Schapiro L 1972 *Totalitarianism*. Praeger, New York

Soper P S 1985 *Totalitarianism: A Conceptual Approach*. University Press of America, Lanham, MD

Talmon J L 1960 *The Origins of Totalitarian Democracy*. Praeger, New York

A. J. Gregor

Dictatorship in History

Originally, the word dictatorship had a limited and unambiguous meaning in that it defined an extraordinary magistrature of the ancient Roman Empire. During the nineteenth and, particularly, the twentieth century, the meaning of the term has been substantially modified. In contemporary parlance, 'dictatorship' is used to define any means of governing and any type of political regime based upon the *unlimited concentration of power* in the hands of one individual (or of a group of people, a party or institution) holding it for an indeterminate time and exercising it without being

controlled by either the state's institutions or by its citizens.

This article will trace the transformation of the concept of dictatorship within the main instances of government to which it has been applied and with regard to the principal interpretations of the phenomenon, that is, the governmental types defined as dictatorships produced during the twentieth century when the term first became the object of specific historical research and theoretical analysis.

1. Roman Dictatorship

1.1 Origins and Functions

Dictatorship was probably introduced into the constitutional order of the Roman Republic at the beginning of the fifth century BC, soon after the end of the monarchy. Today, the original interpretation of dictatorship as attributed to the survival of the monarchy tradition has been abandoned. A more plausible interpretation is that which sees Roman dictatorship as being derived of an already existing institution in other cities in the province of Latium (Sterling 1965). The role of dictator could be conferred for the performance of various public duties, the most important being that of the defence of the city and of its rule in situations of serious danger, in the case of external war (*belli gerundi causa*) or internal turmoil (*seditionis sedandae causa*). In these circumstances, full civil and military powers (*imperium maximum*) were conferred upon the dictator for a maximum period of six months. The nomination was made by one of the two consuls in office (who could not, however, nominate themselves) upon the request of the senate, which assessed the gravity of the situation. The dictator stepped down at the same time as the consul who had nominated him. All other magistrates were subordinate to the dictator. He himself, however, did not have the power to repeal them nor to pass or repeal laws, modify the constitutional order, or nominate his successor. Once the duty for which he had been nominated to fulfill was completed, even before the passing of six months, the dictator resigned power to the ordinary magistrates.

1.2 The Transformation and End of the Roman Dictator

The dictator remained in force until the third century BC. The last dictator to have maximum civil and military powers was nominated in 216 BC during the second Punic war between Rome and Carthage. When reinstated in the first century BC by L. Cornelius Sulla and later by Julius Caesar during the time of the civil wars that led to the end of the Republic, the dictator's function was radically transformed. The dictatorships of Sulla and Caesar, despite the outward appearance of legality, originated in the force of the army, were conferred for an indeterminate length of time, and were used to eliminate political adversaries and to change the order of the Roman State. Roman dictatorship was legally abolished by Mark Anthony in 44 BC, directly following the death of Caesar (*lex Antonia de dictatura tollenda*) and was not reinstated again during the Empire. The term 'dictator' also was not used for many centuries and was not in any case adopted again until the beginning of contemporary times in order to define new exceptional forms of personal power that appeared in the medieval and modern ages in the form of the 'Podestà' of the Italian communes, the Renaissance 'Signoria' or Oliver Cromwell's 'Protectorate' (see *English Revolution, The*; *Middle Ages, The*; *Renaissance*).

2. From Roman Dictatorship to Modern Dictatorship

2.1 The Model of Roman Dictatorship in the Modern Ages

The term 'dictatorship,' referring exclusively to the Republican magistrature, appeared once again in political writings of modern times such as those by Jean Bodin and Thomas Hobbes, who cited it as an example of the fulfillment of the principle of Reason of State and of absolute sovereignty. Niccolò Machiavelli made a decisive contribution to that which we may term the 'apology conception' of dictatorship, understood as exceptional personal government for the purpose of saving the Republic and defending its liberty against internal and external enemies. The value of dictatorship, according to Machiavelli, was the brevity of its duration. As long as this function should be retained, dictatorship '*fece bene e non danno alla Republica romana*' (*Discourses*, I, 10, 33, 35).

This apologetic conception of dictatorship was again proposed in the seventeenth century with the idealization by democratic thinkers of the Enlightenment of republican public spirit (see *Enlightenment*). Jean-Jacques Rousseau maintained that in periods of danger the democratic state could nominate a 'supreme chief who silences all the laws,' fixing a relatively short period for the fulfillment of this chief's role without the capacity for its being prolonged, because only in this way would it be possible to avoid the transformation of dictatorship into tyranny (Rousseau 1997, p. 138).

2.2 Dictatorship and Revolutionary Power

The term 'dictatorship' re-entered political language at the end of the eighteenth century with the democratic revolutions in America and Europe (see *American Revolution, The*; *French Revolution, The*). In France, following the experience of the Public Health Committee, nominated by the National Convention (1793–5) both a new conception and practical expe-

rience of dictatorship emerged. This was identified by the advent of *revolutionary power* as a form of absolute government, exceptional and temporary, yet personal, which claimed its legitimacy on the basis of the principle of popular sovereignty, the wish to create a new constitutional order and a more just society. After the fall of the Jacobins, François-Noël Babeuf theorized the necessity of dictatorship by a small revolutionary minority in order to seize power by means of insurrection with the aim of realizing a 'society of equals.' After the failure of his insurrectionary attempt (1796) the concept of revolutionary dictatorship was taken up by Filippo Buonarotti, a disciple of Babeuf.

3. *Dictatorship in the Nineteenth Century*

3.1 *'Bonapartism' and 'Caesarism' in Europe*

In the course of the nineteenth century, the term 'dictatorship' was used in its original meaning by the military leaders of the national revolutions. Simon Bolivar proclaimed himself dictator in 1813, and Giuseppe Garibaldi assumed the power of dictator in 1860 during the expedition of the Thousand. The idea of dictatorship was often associated, within nationalism, with the idea of exceptional government for national liberation. Most important, for the history of dictatorship and the transformation of the concept were the new forms of personal regimes founded in Europe by Napoleon Bonaparte (1799–1814) and later by his grandson Luigi Bonaparte, becoming Napoleon III (1851–70). Their dictatorship originated in a *coup d'état*. However, they attempted to legitimize their personal power later with a popular plebiscite and with the transformation of the monarchy without formally reneging on the principles of the French Revolution.

These two experiences of government gave rise to the term 'Bonapartism,' from then on used to define the new type of dictatorship imposed by a *coup d'état*, but claiming a democratic legitimacy by means of plebiscites. The introduction of the term 'Caesarism' was also used in the definition of this new type of dictatorship, in that Julius Caesar was considered the prototype for the modern dictator, emerging victorious after a period of serious social and civic conflict and imposing his personal power using military strength, but also with the support of the masses, presenting himself as the pacifier of the classes, above parties, restorer of stability and of the unity of the State, the incarnation and the interpreter of the national will.

3.2 *Dictatorship in Latin America*

A particular demonstration of a South American version of 'Caesarism' can be found in the presidential dictatorships installed in the majority of the Latin American Republics emerging after independence. These dictators were politicians or military men who elected themselves president and perpetuated their personal power for decades while, nevertheless, paying homage to liberal ideology, allowing for the survival of the constitutional order and, above all, imposing the centralization of state power against the *caudillos*, the local leaders who, in the first years following independence, instituted personal control with the help of the army, the personal fascination with virility, and the support of the indigenous and mestizo populations. Other president-dictators, followers of positivism, justified their dictatorship with the formula 'order and progress' as a necessary means for reinforcing the unity and cohesion of the state, as a pedagogical tool for the education of the people, and as a principal factor in the renovation of society and progress. This new version of the apologetic conception of dictatorship remained one of the principal formulas of populist self-legitimation for the dictators that successively dominated South America (see *Populism*).

3.3 *Apologists for and Critics of Dictatorship*

Other versions of the apologetic conception of dictatorship were elaborated in romantic culture, in Catholic reactionary thought, and in positivism: the dictator was variously exalted as the supreme guide of the people, the interpreter of God's will, or the political genius who actuated the laws of history and of progress. Liberal culture, on the other hand, held that the same possibility of conferring absolute power upon a single individual was incompatible with the concept of the modern state founded upon the liberty of citizens, on the sharing of power and on the principle of majority rule. Duclerc and Pagnerre's *Dictionnaire Politique* of 1848 explicitly refuted the treatment of dictatorship as 'an element of modern politics' because 'invoking dictatorship means invoking violence' and annihilating 'the most sacred principle of democracy, that of majority rule' (see *Crowds in History*; *Liberalism*).

3.4 *The Dictatorship of the Proletariat*

In socialist thinking, the concept of dictatorship as revolutionary power was developed. Louis-Auguste Blanqui theorized the necessity of the dictatorship of a revolutionary minority for the realization of communism. This conception was criticized by Karl Marx and Friedrich Engels because their idea of revolutionary dictatorship had to be the *impersonal power* of the working class. The Paris Commune of 1870 represented for Marx the first concrete attempt at proletarian dictatorship. In that in a society divided into classes every type of state is no other than the dictatorship of the dominant class, according to the idea sketched out by Marx in the *Critique of the Gotha Programme* (1875), 'proletarian dictatorship' was un-

derstood as the organization of the working class at the level of the state following the conquest of power in the transition period from capitalist to communist society.

The concept of 'proletarian dictatorship' was first elaborated with the conquest of power in Russia in 1917 by the Bolshevik Party led by Lenin (see *Communism, History of*). Lenin was a convinced and tenacious advocate of the necessity of dictatorship intended as an unlimited, inflexible, and implacable power in the repression of enemies of the proletariat because, in the absence of this power, the transition from capitalism to socialism would be impossible ('The immediate duties of Soviet power,' April 26, 1918, in Lenin 1971). However, against Marx, Lenin claimed that dictatorship of single individuals, endowed with unlimited power for the realization of the socialist revolution was not in fact incompatible with proletarian dictatorship, exercised by the Bolshevik Party as the vanguard of professional revolutionaries and interpreters of the historical role of the proletariat.

The Leninist conception of dictatorship and its practical activation were criticised by other revolutionary Marxists such as Rosa Luxemburg who claimed that proletarian dictatorship should consist of 'a means of applying democracy, not of its abolition,' and 'the work of the *class*, and not of a small minority of leaders in the name of the class' (Luxemburg 1970 pp. 570–94).

4. The Twentieth Century: A Century of Dictatorships

4.1 The Epidemic of Dictatorship in Europe between the Two World Wars

'In the Victorian era,' wrote Winston Churchill in 1930 in the introduction to one of the first attempts at the historical and theoretical analysis of dictatorship (Forst de Battaglia 1930), 'dictatorship was at once a monstrosity and an anachronism.' The triumph of parliamentary democracy in Europe in the aftermath of World War I was ephemeral. In 1922, the 'crisis of democracy' and the 'crisis of parliamentarism' were already being talked of (Guy-Grand 1922, Schmitt 1923), while with the affirmation of absolute power by the Bolshevik Party in the USSR a new type of dictatorship as a regime by a single party was being affirmed for the first time. The fascination with the Bolshevik experiment excited masses of European workers and prompted the birth of communist parties with the aim of violent conquest of power for the establishment of 'proletarian dictatorship.' Nevertheless, no attempt at communist conquest of power was successful. On the contrary, the fear of Bolshevism generated reaction on the part of the bourgeoisie and the middle classes in many countries which led to the creation of authoritarian, nationalistic, and anti-Bolshevik governments. After 1922, almost every year, more or less durable dictatorial nationalist governments emerged in Italy, Spain, Turkey, Poland, Greece, Portugal, Austria, Germany, the Baltic States, and the Balkans.

4.2 The Model of Fascist Dictatorship

At the end of 1922, with Mussolini's coming to government in Italy (28 October), for the first time a nationalist revolutionary movement organized around a new form of party, the 'party-militia,' conquered power in a country governed by a liberal democracy with the declared object of destroying it. This gave rise to the establishment of a regime that antifascists were the first to call a 'one-party state,' 'totalitarian system,' or 'totalitarianism,' defining a new type of dictatorship founded upon the predominance of a single party, the centralization of power in the hands of the party leader, the imposition of a singular ideology with the aim of realizing the subordination of society to the state by means of control by the singular party and its organizations (Gentile 1995).

Until the middle of the 1920s, fascism, and not Bolshevism, was considered the real threat to parliamentary democracy (Martin 1926). Italian fascism effectively became the model for other nationalist movements that proclaimed their dictatorial vocation, openly opposing the ideals of the French Revolution. At the beginning of the 1930s, Europe was invaded by an 'epidemic of dictatorship' (Sforza 1931). Initially, the dictatorial disease was considered to be exclusive to less developed countries (Cambò 1930). This interpretation was proved wrong after the rise of National Socialist dictatorship in Germany (see *National Socialism and Fascism*). In the second half of the 1930s, however, the 'seduction of fascism' seemed to grow (Drabowitch 1934). During World War II, most of Southern, Central, and Eastern Europe was governed by dictatorial systems.

4.3 Totalitarian States

In the period between the two world wars a new apologetic myth of the dictatorial figure as the 'strong man' who emerges in times of crisis to rescue the State and re-establishes order, discipline, and efficiency was diffused (Séché 1924, Béraud 1933, Massis 1939). At the same time, the first attempts at comparative analysis aimed at elaborating a unitary perspective on the phenomenon of dictatorship were begun. Already, by the second half of the 1920s, several analogies between the dictatorship of Bolshevism and that of fascism came to prominence, later compared also to the dictatorship of National Socialism. The classification of these new types of dictatorship under the category of 'totalitarian states' appeared at the end of

the 1920s and was already consolidated by the 1930s and 1940s when the first comparative analyses of fascist, communist, and national socialist regimes were carried out (see Kohn 1935, Hoover 1937, Sturzo 1938 1972, Borkenau 1939, American Philosophical Society 1940, Neumann 1942, Aron 1946).

These studies analyzed the new form of dictatorship from the point of view of its most original aspects, neatly distinguishing it from both Roman and nineteenth-century dictatorship. These aspects included:

(a) the peculiarities of the original phenomena of dictatorship, attributed to the legacy of war, the embitterment of class war, the radicalization of politics, and the incapacity of the parliamentary system to adapt itself to the tumultuous growth of mass society and to provide an effective solution to social and economic conflicts;

(b) the interest of the masses living through the turmoil of a society undergoing transformation and looking for simple and reassuring solutions by subordinating themselves to the command of a 'strong man' who promised to resolve the crisis with brutal yet efficient methods;

(c) the novelty of contemporary dictatorship as a one-party regime, created by a mass movement that used parliamentary channels, violence, and terror to conquer and maintain power, rendering permanent its own dictatorship and institutionalizing it as a fundamental pillar of a new political system;

(d) the personalities of the dictators brought to power by these parties such as Stalin, Mussolini, and Hitler, that is, 'new men' emerging from anonymity and rising to power solely due to their political capacity, exalted and revered like demigods by their followers and by the masses;

(e) the fanaticism and missionary spirit of the new dictators who endowed their ideology and action with the character of a secular religion;

(f) the use of methods of technological and organizational modernization for the mobilization of the masses and the development of an intense and constant activity of collective indoctrination; and

(g) the ambition of subordinating the collectivity and society to the permanent and capillary control of a single party with the aim of creating a new species of human being.

4.4 The Dictatorial Phenomenon After the End of World War II

In the aftermath of World War II, while parliamentary democracy was being re-established in Italy and Germany, liberal democracy was definitively consolidated in Western Europe and, after 1945, there were no new dictatorships in this part of the continent. From among the nationalist dictatorships emerging from the interwar period, only the Salazar regime in Portugal and that of Franco in Spain remained standing, both of which returned to democracy after the death of the two dictators in the second half of the 1970s. The only new dictatorship was that of the Colonels' regime in Greece (1967–1974).

The phenomenon of dictatorship, nevertheless, enjoyed a new and intensified period of expansion in almost all continents. Only North America and Australia remained uncontaminated by the new dictatorial epidemic. Directly after the end of World War II, East Germany, all of Eastern Europe, and the Balkans were put under the control of the new communist dictatorial regimes, the so-called 'popular democracies' (Fejtö 1969), completely subordinated to the political and military hegemony of Stalin's USSR from which only the communist regimes of Yugoslavia and Albania escaped. From 1949, with the victory of the communist revolution led by Mao in China, the world's most densely populated country became governed by a dictatorial regime with its own characteristics, different to the European communist dictatorships and to those of the USSR. During the 'Cold War,' the communist model, with variations original to each country, inspired the construction of several dictatorial regimes, both civilian and military, in Central America, Asia, and Africa with the help of the two largest communist countries, the USSR and the People's Republic of China, becoming, since the 1950s, rival and antagonistic states both for ideological reasons and for those concerning territorial conflict (see *Chinese Revolutions: Twentieth Century*; *Cuban Revolution, The*; *Vietnamese Revolution, The*).

The common characteristics of the communist dictatorships were (a) the adoption of Marxist-Leninism as their official ideology; (b) the supremacy of the Communist Party as the single and dominant party; (c) the exaltation of the party leader resulting in the institutionalization of the 'cult of personality'; (d) the policed repression of any opposition or dissidence, including the recourse to mass terror; (e) the nationalization and planning of the forces of production and of the economy; (f) the submission to control by the Communist Party of all political and social activities; and (g) the development of a particular form of 'communist patriotism,' centered upon the myth of the historical mission of the proletariat.

During the 'Cold War,' anticommunism was the motive and principal ideological justification for the establishment of right-wing, civilian, and, more often, military dictatorships sustained by the USA. These dictatorships, especially the military ones (Nordlinger 1977), when not based exclusively on terrorist repression and the arbitrary use of power in the interests of the dictator, his family, or acolytes, attempted to relate policed repression to a demagogic politics of nationalist populism all the while linked to programs of modernization and development (see *Modernization and Modernity in History*). In the second half of the twentieth century, dictatorship in the Third World became 'the normal form of govern-

ment' (Roy 1977). Almost all the new states emerging in Africa and Asia after the fall of the colonial empires were governed by dictatorial regimes, generally military with a single party created from above, a charismatic leader, and a populist ideology of national socialism with programs of secularization and modernization. In the second half of the 1970s, in several Islamic countries, religious fundamentalism became the matrix and ideology for new dictatorial governments.

The tendency towards the proliferation of dictatorships came to a halt in the last two decades of the twentieth century (Hamon 1982). The number of dictatorships—or governments considered as such—has notably diminished. In Latin America, parliamentary democracy has been reinstated in the grand majority of states previously governed by dictators. The fall of the USSR brought about also that of the communist regimes of Central and Eastern Europe that it sustained (see *Revolutions of 1989–90 in Eastern Central Europe*). The promotion of human rights has acquired new determination, albeit not always effective, in contesting antidemocratic, repressive, and terrorist governments and in the defeat of the various apologetic conceptions of dictatorship (see *Human Rights, History of*). Nevertheless, it would be too optimistic to use the above to declare the definitive death of dictatorship in the contemporary world and the victory of the democracy of human rights (Hermet 1993).

5. *Contemporary Dictatorship: Problems of Definition and Interpretation*

5.1 *Definitions and Typologies*

Contemporary dictatorships are extremely heterogeneous: civilian or military, personal or collective, conservative or revolutionary, traditionalist or modernizing, bureaucratic or charismatic, authoritarian or totalitarian. Moreover, not only do contemporary dictatorships have nothing in common with their ancient predecessor, apart from the concentration of power, they also differ extensively among themselves according to the geographical, historical, and cultural contexts in which they emerge and owing to the economic, political, and social situations that give rise to them. Furthermore, they differ as regards the social origins of their protagonists as well as the latter's political formation, the modes in which power and methods of action are organized, the ideologies they profess and the aims they propose. Finally, differences exist relative to the historical development of their emergence, the results they have achieved, and the legacies they have left.

Given the variety and complexity of the phenomenon of dictatorship in the twentieth century, it is very difficult to elaborate a general, unitary theory of dictatorship. Several fundamental problems therefore remain unsolved, particularly those which concern the specific nature of contemporary dictatorship with respect to other forms of absolute personal power. There are various definitions and typologies for classifying the numerous forms of unlimited concentration of power appearing during the course of the twentieth century, commonly adopting the term dictatorship albeit with a meaning varying widely from the original (see especially Schmitt 1928, Friedrich and Brzezinski 1956, Neumann 1957). There are some scholars, however, who prefer to use the term dictatorship in reference only to its original meaning, finding the term 'antidemocratic regime' more appropriate for defining the modern forms of unlimited concentration of power, be they authoritarian or totalitarian (Linz 2000). The identification of contemporary dictatorship with antidemocratic regimes is, however, only one aspect of a much more ambiguous relationship between contemporary dictatorship and democracy, understood as the affirmation of the principle of popular sovereignty and mass participation in political life. Contemporary dictatorship in fact 'belongs to the postdemocratic era' (Kohn 1935), that is, it presupposes the democratic process yet nonetheless searches for its own formula of legitimation within the principle of popular or national sovereignty (Duverger 1982).

5.2 *Dictatorship, Tyranny, Despotism*

In contemporary speech, dictatorship is often associated with tyranny or despotism. Terms such as dictator, tyrant, and despot are often used as synonyms or alternated between. The first students of the phenomenon of dictatorship in the interwar period, while maintaining that the concept of dictatorship is linked to the Roman model of an extraordinary, yet constitutional and temporary magistrature, preferred to speak of a 'new age of despots' (Spencer 1927), or of a new '*ère des tyrannies*' (Halèvy 1938). In more recent studies also the term 'tyrant' is preferred over that of dictator for defining the numerous ranks of those holding absolute personal power, dominating, in the second half of the twentieth century—and still in some cases dominating—in many countries of the Third World (Rubin 1987, Chirot 1994). In fact, contemporary dictatorship evokes some of the characteristics of Greek tyranny (Stoppino 1976), but there are also substantial differences. Above all, the origin of contemporary dictatorship is not always illegitimate. For example, Hitler came to power in 1933 through legal channels and many presidents who became dictators in Latin America were legally elected. Furthermore, the majority of modern dictatorships have a very complex structure of organization and exercise of power due to the mass dimension of the societies in which they operate, the methods they adopt, and the aims they propose. All the above makes contemporary

dictatorship substantially new and different to tyranny. Therefore, the thesis that holds that the difference between tyranny and contemporary dictatorship as a form of absolute personal power is not substantive, but regards solely the form, technique, and method of government should be forcefully rejected (Hallgarten 1957).

5.3 Dictatorship and the Dictator

Several scholars consider the figure of the dictator an essential and fundamental element in the definition of dictatorship (Cobban 1939). A large part of the general history of dictatorship consists of collections of biographies that sometimes link, in more or less chronological succession, Greek tyranny and the Roman dictators, Cola of Rienzi and Geralomo Savonarola, Cromwell and Robespierre, the two Napoleons, the 'caudillos' and the authoritarian and totalitarian dictators of the twentieth century (Funck-Brentano 1928, Bainville 1939, Hallgarten 1957). Some scholars maintain that governments in which unlimited power is not concentrated in one individual cannot be defined as dictatorships in the proper sense of the term (Sartori 1971).

The identification of dictatorship with the figure of the dictator is certainly justified in the majority of the cases in history. In almost all contemporary dictatorships, the dominant personality of one man emerged at the outset or was subsequently imposed, often surrounded by a charismatic and sacred fascination with the cult of the personality (Schweitzer 1984). At the same time, dictatorial governments exist without the figure of the dictator, for example, the USSR following Stalin's death and the dictatorial regimes governed by military juntas. Moreover, not even in totalitarian regimes that appear to be internally controlled by the personal power of a dominant and charismatic figure, can the nature and character of the dictatorship as a system of power founded upon the single party be identified in the personage of the dictator (Tucker 1972, Schapiro and Lewis 1970, Gentile 1998).

5.4 Conclusion: General Characteristics of Contemporary Dictatorship

The diversity in the experiences of power defined as dictatorship in the contemporary world, the variety of definitions and typologies and the problems remaining controversial regarding the nature of dictatorship do not, nonetheless, impede us from being able to identify at least some common characteristics enabling the definition of the specificity of the phenomenon of dictatorship in the modern era in their formative or stabilizing phases.

According to the definition given at the beginning of this article, all types of government and regime in which a single person, a group of people, a party, or a military or religious institution (a) conquers the monopoly over political power, by legal or illegal means, and maintains it for an indeterminate length of time; (b) suppresses, reduces, or impedes citizens' enjoyment of their civil or political rights; (c) forbids or represses the activities of dissidents or of the opposition with the use of force and with recourse to terrorist means; (d) excludes *de facto* the possibility of the legal and pacific repeal of the rulers by the ruled even while allowing for electoral consultations and plebiscites to take place, may be included under the category of contemporary dictatorship. Contemporary dictatorship is, thus, the antithesis of democracy understood as a representative political system based upon the division and balance of power, a respect for the personal and public rights of citizens, on freedom of speech and of association, and on the periodic election of the government by the electorate, by means of a free and peaceful political contest and according to legal procedures guaranteed by the constitution.

See also: Authoritarianism; National Socialism and Fascism; Sovereignty: Political

Bibliography

American Philosophical Society 1940 Symposium on the totalitarian state. From the standpoints of history, political science, economics, and sociology. *Proceedings of the American Philosophical Society*, Philadelphia, PA

Arendt H 1951 *The Origins of Totalitarianism*. Harcourt, Brace and Company, New York

Aron R 1946 *L'homme contre les Tyrans*. Gallimard, Paris

Aron R 1965 *Démocratie et totalitarisme*. Gallimard, Paris

Bainville J 1939 *Les Dictateurs*. Plon, Paris

Béraud H 1933 *Dictateurs d'aujourd'hui*. Flammarion, Paris

Borkenau F 1939 *The Totalitarian Enemy*. Faber & Faber, London

Cambo F 1930 *Les Dictatures*. Alcan, Paris

Chirot D 1994 *Modern Tyrants: The Power and Prevalence of Evil in Our Age*. Free Press, New York

Cobban A 1939 *Dictatorship: Its History and Theory*. Jonathan Cape, London

Collier D 1979 *The New Authoritarianism in Latin America*. Princeton University Press, Princeton, NJ

Drabowitch W 1934 *Fragilité de la libertè et seduction des dictatures*. Mercures de France, Paris

Duverger M 1956 *De la dictature*. Julliard, Paris

Duverger M (ed.) 1982 *Dictatures et légitimité*. Presse Universitaire de France, Paris

Fejtö F 1969 *Histoire des démocraties populaires*. Seuil, Paris

Ford S G 1935 *Dictatorship in the Modern World*. University of Minnesota Press, Minneapolis, MN

Forst de Battaglia O (ed.) 1930 *Dictatorship on its Trial*. George G. Harp, London

Friedrich C J, Brzezinski 1956 *Totalitarian Dictatorships and Autocracy*. Harvard University Press, Cambridge, MA

Funck-Brentano J F 1928 *Ce qu'il faut connaître de la dictature*. Boivin, Paris

Gentile E 1995 *La vita italiana al totalitarismo. Il partito e lo Stato nel regime fascista*. Carocci, Rome

Gentile E 1998 Mussolini's charisma. *Modern Italy* **3**(2): 219–35
Guy-Grand G 1922 *La Démocracie et l'après guerre*. M. Rivière, Paris
Halèvy E 1938 *L'ère des tyrannies: études sur le socialisme et la guerre*. Gallinard, Paris
Hallgarten W F 1957 *Dämonen oder Retter*. Europäische Verlagsanstalt, Frankfurt, Germany
Hamon L (ed.) 1982 *Mort des Dictatures?* Economica, Paris
Hermet G 1993 *Les désenchantements de la libertè: La sortie des dictatures dans les années 90*. Fayard, Paris
Hoover C B 1937 *Dictators and Democracies*. Macmillan, New York
Kohn H 1935 Communist and fascist dictatorship: a comparative study. In: Kohn H (ed.) 1939 *Revolutions and Dictatorships: Essays is Contemporary History*. Harvard University Press, Cambridge, MA
Lenin V I 1971 *Opere scelte*. Edizioni Progresso, Mosca, Italy
Linz J 2000 *Totalitarian and Authoritarian Regimes*. Boulder, London
Luxemburg R 1970 *Scritti Politici*. Editori Riuniti, Rome
Martin H 1926 *Demokratie oder Diktatur?*. Verlag für Politik und Wirtschaft, Berlin
Massis H 1939 *Chefs*. Plon, Paris
Neumann F L 1957 *The Democratic and the Authoritarian State: Essays in Political and Legal Theory*, Marcuse H (ed.). Free Press, Glencoe, IL
Neumann S 1942 *Permanent Revolution: The Total State in a World at War*. Harper & Brothers, New York
Nordlinger E A 1977 *Soldiers in Politics*. Prentice Hall, Englewood Cliffs, NJ
Perlmutter A 1981 *Modern Authoritarianism: A Comparative Institutional Analysis*. Yale University Press, New Haven, CT
Rousseau J J 1997 *The Social Contract and Other Later Political Writings*, Gourevitch V (ed.) Cambridge University Press, Cambridge, UK
Roy M-P 1977 *Les regimes politiques du Tiers monde*. Libraire générale de droit et de jurisprudence, Paris
Rubin B 1987 *Modern Dictators: Third World Coup Makers, Strongmen, and Populist Tyrants*. McGraw-Hill, New York
Sartori G 1971 Appunti per la teoria generale della dittatura. In: von Beyme K (ed.) *Theory and Politics. Teorie und Politik*. The Hague, The Netherlands, pp. 456–85
Schapiro L, Lewis J W 1970 *Party Leadership and Revolutionary Power in China*. Cambridge University Press, Cambridge, UK
Schmitt C 1923 *Die geistesgeschichtliche Lage des heutigen Parlamentarismus*. Dunker & Humboldt, Berlin
Schmitt C 1928 *Die Diktatur: Von del Anfängen des modernen Souveränitätsgedankens bis zum proletarischen Klassenkampf*, 2nd edn. Duncker & Humbolt, Munich, Germany
Schweitzer A 1984 *The Age of Charisma*. Nelson-Hall, Chicago
Séché A 1924 *Le Dictateur*. Bossard, Paris
Sforza C 1931 *Dictateurs et dictatures de l'après-guerre*. Gallimard, Paris
Spearman D 1939 *Modern Dictatorship*. Jonathan Cape, London
Spencer H R 1927 European dictatorships. *American Political Science Review* **3**: 537–51
Stammer O 1955 *Demokratie und Diktatur*. Gehlen, Berlin
Sterling E 1965 *Der unvollkommene Staat: Studien über Diktatur und Demokratie*. Europäische Verlagsanstalt, Frankfurt, Germany
Stoppino M 1976 Dittatura. In: Bobbio N, Matteucci N (eds.) *Dizionario di politica*. Utet, Turin, Italy, pp. 341–51
Sturzo L [1938, 1953] 1972 *Politica e morale. Coscienza e politica*. Zanichelli, Bologna, Italy
Tucker R 1972 The dictator and totalitarianism. In: Tucker R (ed.) *The Soviet Political Mind. Stalinism and Post-Stalin Change*. Norton, New York

E. Gentile

Differential Aging

Differential aging is a term within lifespan psychology that is intended to communicate the heterogeneity of developmental processes in the later years of adulthood. The concept of differential aging underlies many contemporary theoretical and empirical analyses; however, the introduction of this particular term can be traced back to the work of Thomae in the Bonn Longitudinal Study (Thomae 1979). More recently, investigation of differential aging was a central component of the Berlin Aging Study (BASE), a comprehensive study of over 500 adults between 70 to over 100 years living in Berlin, Germany (Baltes and Mayer 1998).

According to the principle of differential aging, the rate of functional losses and gains over time shows both inter-individual variability and intra-individual variability. Inter-individual variability is shown by the fact that there are differences across individuals in the rate at which changes occur over time. Intra-individual variability is demonstrated by the finding of differences in the rates of change over time within the same individual. The principle of differential aging stands in sharp contrast to the previously held views in developmental psychology that aging is a uniform process of decline, also known as dedifferentiation. Although it is true that death is the inevitable outcome of the aging process, this fact does not rule out the possibility that growth can occur in psychological functions even at the end of life. Plasticity is another component of differential aging, meaning that the potential exists throughout later adulthood for growth and adaptation.

1. *Evidence for Differential Aging*

Support for the concept of differential aging has emerged from a variety of sources within the field of gerontology. One of the strongest arguments used to support differential aging is a meta-analysis of a wide range of empirical studies based on data available through the late 1980s. This analysis yielded evidence of greater variability among older, compared to younger, samples of adults (Nelson and Dannefer 1992). However, without providing explicit tests of the concept of differential aging, many other investigations have led to similar conclusions. The concept of differential aging is becoming a well-established fact

within the field of psychology and aging (Whitbourne 2001). It is no longer unexpected to show heterogeneity across older samples and within areas of functioning in the same individual.

Explanations of differential aging relate to models in lifespan developmental psychology that emphasize the importance of person–context transactions as influences on development (Lerner 1995). Such transactions affect the rate of aging within and across individuals as they are exposed to and select environments that differentially affect their physical, psychological, and social characteristics. Furthermore, rates of aging across and within individuals may reflect exposure to opportunities for prevention and compensation as represented, albeit crudely, by the variable of socioeconomic status. The higher mortality rates of individuals from lower socioeconomic backgrounds may be viewed as a manifestation of these differential opportunities (Lantz et al. 1998). To the extent that higher mortality rates reflect a more rapid acceleration of deleterious changes either directly or indirectly linked to aging, variations among socioeconomic status may be seen as a critical contextual variable within the framework of differential aging.

A concept highly related to differential aging is that of selective optimization with compensation (SOC; Baltes and Baltes 1990). Selection refers to the process of actively or passively reducing the number of goals and domains in which to become involved. The purpose of selection is to conserve energy and motivation for more important goals or to select new goals that will serve particular developmental tasks, such as coming to terms with mortality. Compensation in this model is defined as searching for and using alternate means of reaching a goal once it is no longer possible to use previous methods to achieve that goal. For example, an individual may wish to remain physically active but when limitations occur in mobility, the method of involvement may change accordingly. Optimization is the process of refining the methods and resources needed to reach a goal and thereby achieve success in particular domains. Applications of the SOC model to the processes through which individuals attempt to control and regulate developmental change over the lifespan are providing further elaboration of this model (Heckhausen 1997).

1.1 Application to Physical Aging Processes

The physical aging process is a prime example of the concept of differential aging. Aging of the body occurs across all organ systems, but at different rates within and across individuals. These concepts are expressed in the multiple threshold model (Whitbourne 1996), which postulates a reciprocal relationship between the aging of the systems within the body and the individual's identity or self-conception throughout the years of adulthood. It is known that changes occur within bodily systems at varying rates over time, reflecting both genetic predisposition and the impact of the environment. A threshold is the point at which the individual becomes aware of having experienced an age-related change in an area of functioning. After the threshold is crossed, changes in identity occur as the individual becomes closer to defining the self as 'old.'

Predictions of the multiple threshold model were tested in an investigation of a large sample of adults from 40 to 95 years of age (Whitbourne and Collins 1998). Respondents were asked to indicate the extent to which they had experienced a set of physical and cognitive changes related to the aging process as well as the way that these changes had affected their identities. There were clear individual differences in the areas of functioning identified as significant, as well as variations in the responses that individuals had to these age-related changes. Changes with regard to aging were clearly distinguished within the minds of respondents in this study, who reacted differently to changes in cognitive processes compared to changes in mobility and physical strength.

Another facet of the multiple threshold model is that individuals vary in the preventive and compensatory mechanisms available to them across a wide area of functions. For example, in the case of aerobic capacity, it is well known that regular activity can help to offset the aging process, even if this activity is begun quite late in life. By contrast, individuals can also choose to alter the rate of the aging process by the adoption of harmful health habits, such as cigarette smoking or a sedentary life style. Their decision to engage in these activities may reflect social factors as well as psychological processes. As noted, education and occupational background are further contributing factors, as individuals from higher socioeconomic levels are more likely to be aware of the strategies they need to employ to moderate the rate of their own aging (Lantz et al. 1998).

The extent to which individuals possess resources based on their sensory, cognitive, personality, and social skills provide another factor to be considered in differential aging (Baltes and Lang 1997). People who are low in these personal and social resources have more difficulty negotiating the tasks of daily life in later adulthood. In part, differences between resource-rich and resource-poor older adults occur because these individuals have unequal opportunities to take advantage of compensatory measures that would help offset the negative effects of aging.

One caveat to the notion of increasing variability with age in physical functioning comes from the BASE (Mayer et al. 1998). Among the oldest-old in the sample, i.e., those over 85, there was a decline across a variety of abilities and a tendency for greater similarity across domains of functioning. This finding suggests that individuals in the latest period of life, until

recently not widely available for study, may be undergoing changes that are discontinuous with the earlier years of the lifespan. Changes in biological functioning may become more widespread after this point, affecting a broad range of abilities that depend on the integrity of the body's systems.

1.2 Application to Intellectual Development

Differential aging also appears in research on intellectual development in later adulthood. Researchers in this field recognized by the 1960s that there was a considerable discrepancy between age patterns based on verbal vs. non-verbal tests of intelligence as well as differences between cross-sectional and longitudinal studies. Findings from the Seattle Longitudinal Study (SLS) revealed an overall picture of intelligence in later adulthood as one of stability until the 50s or 60s followed by decline through the oldest age tested. However, there are important qualifications on this conclusion (Schaie 1996). First, although some individuals may show declines in intelligence by the mid-50s, significant losses do not occur overall until the decade of the 70s. A second point is deterioration of functioning does not occur across all abilities, even into the decade of the 80s. Third, most people are able to retain competent performance of familiar skills, particularly those that are of importance to them. Finally, there was evidence for plasticity as indicated by the success of training studies intended to improve the scores of older individuals on tests that had otherwise shown age-related declines.

A fundamental distinction has emerged in studies of intellectual functioning and age indicating that there are different trajectories over the lifespan for the two theorized components of intelligence: 'fluid' abilities, or the mechanics of intelligence, and 'crystallized' abilities, or the pragmatics of intelligence (Baltes et al. 1999). With increasing age in adulthood, losses occur in the functions that underlie the mechanics of intelligence, such as attention, working memory, and psychomotor speed. The pragmatic abilities involving mastery of a body of declarative and procedural knowledge, a well as expertise in specific domains, are maintained into later life. These abilities can allow older individuals to perform at relatively high levels despite the aging of functions that contribute to losses in the mechanics of intelligence. The area of wisdom, or optimal expression of knowledge in the fundamental pragmatics of life, provides a prime example of the growth of abilities in later life even in the face of sensory, perceptual, and cognitive losses (Baltes and Staudinger 1993).

1.3 Application to Personality and Social Functioning

The concept of differential aging, with the implication of variations in gains and losses across and within individuals, is implicit in the theory of socioemotional selectivity, the proposition that throughout adulthood, individuals reduce the range of their relationships to maximize the social and emotional gains and minimize risks (Carstensen 1992). According to this theory, aging is associated with a desire to become more focused on maximizing the emotional rewards of relationships and less involved in seeking information or knowledge through social interactions. The desire to maximize emotional rewards leads adults increasingly to prefer spending time in their relationships with people who are familiar to them rather than seeking out new friends and acquaintances. Although this theory postulates a universal process, that of seeking out the familiar rather than the new in relationships, it also has relevance to the differential aging principle. According to the theory, it is not aging but the recognition of limited time that stimulates the shift in motivation for relationships. Younger individuals will show similar tendencies to prefer spending time with close friends and family rather than strangers when they perceive that an ending is drawing near.

That older adults are able to maximize their feelings of well being by active steps they take to regulate their relationships is also reflected in research on subjective well-being in later adulthood. In one large-scale survey of a national sample of adults (Mroczek and Kolarz 1998) the majority of older adults were found to rate their happiness levels as 'high' or 'pretty high.' However, there were variations in the relationship between subjective well-being and age as a function of gender, personality, and marital status. Furthermore, the results differed according to which component of subjective well-being was investigated. Positive affect was higher in the older age groups, particularly so for women. For men, personality interacted with age in ratings of positive affect. Higher positive affect scores were found only in the older group of introverted men. Men who were extroverts had high positive affect scores throughout the adult age range tested. Only the married men showed a relationship between age and ratings of negative affect. The older groups of married men had lower scores than the younger married men in the sample. Although these data only reflect one study out of many conducted within the field (Diener 1998), they are of particular relevance to the principle of differential aging because they point to the roles of personality and contextual factors as influences on levels of well-being in middle and later adulthood.

2. Suggestions for Future Research

Despite the existence of the principle of differential aging, many researchers within psychological gerontology and lifespan psychology continue to focus their efforts on the search for consistent patterns of age differences that would characterize older adults as a homogenous group. Theories such as the general slowing hypothesis (Salthouse 1996) propose that the

aging process involves a unilateral loss of speed which in turn accounts for reductions in processing abilities. Trait theories of personality propose, in contrast, that there is overall stability of personality throughout the decades of middle and later adulthood (McCrae and Costa 1990). However, even within these frameworks, there are attempts to identify patterns of differences associated with biological or experiential factors. For example, experienced typists may preserve their abilities to perform well-learned tasks by compensating for loss of speed with changes in strategy (Salthouse and Saults 1987). Similarly, in the area of personality, although the search for universal trends continues (McCrae et al. 1999), researchers also attempt to identify patterns of life events that alter levels of personality traits or their expression in behavior (Helson et al. 1995).

3. Summary

The examples provided here of research supporting the concept of differential aging suggest the need for further empirical investigation that employs a multidimensional and multifaceted approach to the study of lifespan processes as these unfold in later life. It is no longer sufficient to document the existence of age differences or changes over time employing standard methods using correlation or analysis of variance. With the increasing availability of structural modeling techniques, lifespan researchers have the opportunity to translate the principle of differential aging into operational terms. Research based on this principle will not only reflect more accurately the nature of the aging process but can also provide the basis for effective interventions to improve the daily lives of older adults.

See also: Aging, Theories of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Cognitive Aging; Memory and Aging, Cognitive Psychology of; Mortality Differentials: Selection and Causation; Mortality, Epidemiological, and Health Transitions; Old Age and Centenarians

Bibliography

Baltes M M, Lang F R 1997 Everyday functioning and successful aging: the impact of resources. *Psychology and Aging* **12**: 433–43

Baltes P B, Baltes M M 1990 Psychological perspectives on successful aging: A model of selective optimization with compensation. In: Baltes P B, Baltes M M (eds.) *Successful Aging: Perspectives from the Behavioral Sciences*. Cambridge University Press, New York, pp. 1–34

Baltes P B, Mayer K U 1998 *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, Cambridge, UK

Baltes P B, Staudinger U M 1993 The search for a psychology of wisdom. *Current Directions in Psychological Science* **2**: 75–80

Baltes P B, Staudinger U M, Lindenberger U 1999 Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology* **50**: 471–507

Carstensen L L 1992 Social and emotional patterns in adulthood: Support for socioemotional selectivity theory. *Psychology and Aging* **7**: 331–8

Diener E 1998 Subjective well being: Three decades of progress. *Psychological Bulletin* **125**: 276–302

Heckhausen J 1997 Developmental regulation across adulthood: Primary and secondary control of age-related challenges. *Developmental Psychology* **33**: 176–87

Helson R, Stewart A J, Ostrove J 1995 Identity in three cohorts of midlife women. *Journal of Personality and Social Psychology* **69**: 544–57

Lantz P M, House J S, Lepkowski J M, Williams D R, Mero R P, Chen J 1998 Socioeconomic factors, health behaviors, and mortality: Results from a nationally representative prospective study of US adults. *Journal of the American Medical Association* **279**: 1703–8

Lerner R M 1995 Developing individuals within changing contexts: Implications of developmental contextualism for human development, research, policy, and programs. In: Kindermann T J, Valsiner J (eds.) *Development of Person-Context Relations*. Lawrence Erlbaum, Hillsdale NJ, pp. 13–37

Mayer K U, Baltes P B, Baltes M M, Borchelt M, Delius J, Helmchen H, Linden M, Smith J, Staudinger U M, Steinhagen-Thiessen E, Wagner M 1998 What do we know about old age and aging? Conclusions from the Berlin Aging Study. In: Baltes P B, Mayer K U (eds.) *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, Cambridge, UK, pp. 475–520

McCrae R R, Costa P T Jr., 1990 *Personality in Adulthood*. Guilford, New York

McCrae R R, Costa P T Jr., de Lima M P, Simoes A, Ostendorf F, Angleitner A, Marusic I, Bratko D, Caprara G V, Barbaranelli C, Chae J-H, Piedmont R L 1999 Age differences in personality across the adult life span: Parallels in five cultures. *Developmental Psychology* **35**: 466–77

Mroczek D K, Kolarz C M 1998 The effect of age on positive and negative affect: A developmental perspective on happiness. *Journal of Personality and Social Psychology* **75**: 1333–49

Nelson E A, Dannefer D 1992 Aged heterogeneity: Fact or fiction? The fate of diversity in gerontological research. *Gerontologist* **32**: 17–23

Salthouse T A 1996 The processing-speed theory of adult age differences in cognition. *Psychological Review* **103**: 403–28

Salthouse T A, Saults J S 1987 Multiple spans in transcription typing. *Journal of Applied Psychology* **72**: 187–96

Schaie K W 1996 Intellectual development in adulthood. In: Birren J E, Schaie K W, Abeles R P, Gatz M, Salthouse T A (eds.) *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 266–86

Thomae H 1979 The concept of development and life span developmental psychology. In: Baltes P B, Brim J O G (eds.) *Life-Span Development and Behavior*. Academic Press, New York, Vol. 2, pp. 281–312

Whitbourne S K 1996 *The Aging Individual: Physical and Psychological Perspectives*. Springer, New York

Whitbourne S K 2001 *Adult Development and Aging: Biopsychosocial Perspectives*. Wiley, New York

Whitbourne S K, Collins K C 1998 Identity and physical changes in later adulthood: Theoretical and clinical implications. *Psychotherapy* **35**: 519–30

S. K. Whitbourne

Differential Diagnosis in Psychiatry

Diagnosis (derived from two Greek words: *dia*, meaning through, and *gnosis*, meaning knowledge) is a key medical activity. Indeed, there is a sense in which it is the most fundamental of all medical skills because it is a necessary preliminary to treatment, and treatment is often carried out by others. The patient consults a doctor who assesses them and makes a diagnosis. Thereafter, depending on the nature of the problem, the patient may be treated partly or entirely by members of a wide range of other professions, including nurses, physiotherapists, clinical psychologists, and occupational therapists. Sometimes the diagnosis is a simple statement: the patient is suffering from disease or disorder X. More commonly, at least initially, it takes the form of a differential diagnosis: the patient complaining of recurrent upper abdominal pain is probably suffering from a hiatus hernia (an anatomical defect allowing regurgitation of stomach contents into the gullet), but the symptoms may be cardiac in origin (angina), or the result of a gastric or duodenal ulcer, or even of a gastric cancer, and it is important to exclude these possible alternative diagnoses before treatment starts. In principle, differential diagnosis in psychiatry is no different from that in any other branch of medicine. In practice, however, there are a number of important differences because of our limited understanding of the causes, or etiology, of most mental disorders.

1. *The Definition of Diseases*

In the seventeenth century all the diseases that were then recognized were defined by their clinical syndromes, that is by the symptoms the patient complained of, the clinical signs elicited by the physician on examination, and the evolution over time of these symptoms and signs. Over the last three or four hundred years, however, increasing understanding of the etiology of most of these diseases has resulted in their being redefined at some more fundamental level. Smallpox is now defined, not by its characteristic rash, high fever, and mortality, but by the presence of the variola virus. Cancer of the stomach is defined, not by the patient's rapidly worsening indigestion and palpable abdominal mass, but by the histopathology (detailed microscopical structure) of the tumor. Thyrotoxicosis is defined, not by the patient's goiter, weight loss, and bulging eyes, but by a raised blood level of the thyroid hormone thyroxin, and so on. As a result, the diagnosis of most diseases depends on the results of laboratory investigations. Although the physician usually has a shrewd idea which diagnoses are likely after examining the patient, the provisional clinical diagnosis needs to be confirmed by the laboratory. A diagnosis of smallpox is only provisional until the variola virus has been identified; a diagnosis of cancer of the stomach has to be confirmed by histological examination of the tumor; and a diagnosis of thyrotoxicosis has to be confirmed by demonstrating a raised blood level of thyroxin.

Because understanding of the etiology of most psychiatric disorders is still rather rudimentary they still have to be defined by their clinical syndromes. This means that diagnosis, and hence differential diagnosis, is determined by the patient's symptoms (abnormal or distressing subjective experiences reported by the patient) and behavior (observed or reported by others), and the temporal evolution of these abnormalities. Apart from age and sex, information of other kinds, particularly from laboratory and radiological investigations, is rarely relevant, and even if it is relevant it is not decisive. The difference between psychiatry and the rest of medicine is not absolute, however. Several neurological disorders, such as migraine and torticollis, are still defined by their syndromes, and so too are some skin disorders, such as pityriasis rosea. Conversely, some psychiatric disorders, such as Alzheimer's disease, vascular dementia, and Creutzfeldt-Jakob disease, are defined by their neuropathology. Their defining characteristics are comparable, in other words, with those of cancer of the stomach. But even here there is an important practical difference. The histopathology of a gastric tumor can be determined by taking a biopsy through a gastroscope, or at operation, whereas the histopathology of the brain can normally only be determined *postmortem*, and this means that the diagnosis on which treatment is founded still has to be based on the clinical syndrome.

At present there are only a few psychiatric disorders which are defined by the presence of a fundamental abnormality which can be identified or excluded during life. Down's and Klinefelter's syndromes are both defined by the presence of chromosomal abnormalities—the former by an additional chromosome 21 and the latter by an additional X chromosome. Huntington's chorea and the Fragile X syndrome are both defined by abnormal genes at specific sites, and phenylketonuria (like Fragile X, an important cause of mental retardation) is defined by a specific biochemical abnormality. Eventually, of course, as understanding of their causation develops, most psychiatric disorders will either be redefined in terms of an underlying abnormality of some kind, or be replaced by new diagnostic concepts defined in this way. At that stage differential diagnosis in psychiatry

will be no different from that in any other branch of medicine. Until then, however, problems of differential diagnosis have to be handled differently. In other branches of medicine diagnostic problems usually lead to more extensive laboratory and radiological investigations. In psychiatry this is rarely appropriate, and the best approach is usually to try to obtain a more detailed account of the development of the patient's illness, often from other informants, and to monitor its course and response to treatment as closely as possible. Because most psychiatric disorders are defined by the patient's symptoms and abnormal behavior, and changes in these over time, differential diagnosis has to be determined by information about these things, and uncertainties can only be resolved by obtaining more accurate or more detailed information, or by following the course of the illness.

2. *The Adoption of Operational Definitions*

Before the 1970s psychiatric diagnoses were alarmingly imprecise. There was broad agreement on the core symptoms of the major syndromes, and this was reflected internationally in textbook descriptions. It was generally agreed, for example, that the characteristic symptoms of schizophrenia were thought disorder, flattening of affect and delusions, and hallucinations of particular kinds, and that some patients also exhibited other typically schizophrenic symptoms such as apathy and stereotypies. There was no agreement, however, on which combinations of these symptoms were adequate to establish a diagnosis of schizophrenia. Indeed, that question was almost never posed, and different schools of psychiatry and individual psychiatrists made different assumptions, and in practice trainee psychiatrists learnt how to diagnose syndromes such as schizophrenia and mania largely by modeling themselves on their teachers. As a result, several studies of the reliability of psychiatric diagnoses in the 1960s found the reliability of the diagnoses even of experienced psychiatrists to be distressingly low, and international studies like the US/UK Diagnostic Project and the World Health Organization's International Pilot Study of Schizophrenia revealed major between-country differences in the usage of key diagnostic terms such as schizophrenia and manic-depressive illness, and even between different cities or teaching centers in the same country.

As soon as this chaotic situation became apparent it was agreed by the clinical research community that operational definitions were required, at least for all the major syndromes (see *Nosology in Psychiatry*). Essentially, an operational, or semantic, definition is a statement or set of statements which spells out precisely which combinations of symptoms and behaviors are adequate to establish the diagnosis in question. The first set of fifteen operational definitions for use in clinical research was published by the Department of Psychiatry in St Louis in the USA (Feighner et al. 1972) and other rival sets quickly followed. A few years later the American Psychiatric Association provided operational definitions for most of the 200 syndromes described in the third (1980) edition of its *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III*) (American Psychiatric Association 1980). It also demonstrated in the field trials conducted before the glossary was published that these operational definitions made it possible to achieve much higher levels of reliability than had previously been attained. Subsequently the World Health Organization provided similar operational definitions for most of the mental disorders listed in the tenth revision of the *International Classification* (*ICD-10*) (World Health Organization [WHO] 1993).

As a result of this general recognition, at least by clinical research workers and epidemiologists, of the need for operational definitions, and their incorporation into the two most widely used glossaries, *DSM-III* and its successors and *ICD-10*, diagnosis and differential diagnosis have been transformed. Reliability is now much higher; and the unrecognized differences in diagnostic criteria between different countries and schools, which had caused so much confusion previously, have either been eliminated or, if different operational definitions are being used, the differences are now overt and explicit. This in turn has transformed differential diagnosis.

3. *The Consequences of Definition by Clinical Syndrome*

There are, even so, still a number of fundamental problems arising from the fact that most mental disorders are defined by their clinical syndromes. The most important is that there is no scientific way of deciding which of two or more alternative operational definitions of a single disorder is correct, or even which is the best. The choice between alternative clinical definitions of Down's syndrome would be straightforward. One would pick whichever definition most accurately identified individuals with Trisomy 21 (an additional chromosome 21 in all the cells of the body), because Trisomy 21 is now the defining characteristic of Down's syndrome. In the case of most psychiatric disorders, however, there is no comparable external criterion to appeal to, because the clinical syndrome is itself the defining characteristic. This is why alternative operational definitions of disorders such as schizophrenia and agoraphobia remain in widespread use, side by side, and it is commonplace for individual patients to meet one of these definitions but not the other. A patient with typical schizophrenic symptoms for the past six weeks, for example, might well meet the

formal criteria for schizophrenia in *ICD-10*, but not the corresponding criteria in *DSM-IV*, because the latter require evidence of continuous signs of illness for at least six months. In practice this means that discussions about diagnosis need to take place in a context in which it is agreed which glossary (i.e., which set of operational definitions) is being used, and also with a clear recognition of the fact that the choice between alternative glossaries or sets of definitions is arbitrary.

4. *Single Versus Multiple Diagnoses*

In medicine as a whole an attempt is generally made to explain all the patient's symptoms with a single diagnosis, and to try to avoid multiple diagnoses. The rationale for this is that several important diseases, including infections such as tuberculosis and syphilis, and collagen diseases such as polyarteritis nodosa, can affect almost every organ in the body, and the correct diagnosis is more likely to be made if the physician consciously strives to account for all the patient's apparently unrelated problems with a single unifying explanation. In fact, of course, people often have several different things wrong with them simultaneously, particularly as they age. Old people are rather like old cars in this respect.

The same attempt to avoid multiple diagnoses is apparent in psychiatric practice, and the underlying rationale is similar. Patients are commonly given two diagnoses simultaneously if one is a lifelong state, such as mental retardation or a personality disorder, and the other an illness such as schizophrenia or a mood disorder with an identifiable onset and a distinctive course. A diagnosis of some form of substance misuse is also commonly combined with a wide range of other diagnoses. There is a deep-seated reluctance, however, to diagnose a neurotic or stress-related disorder in the presence of psychotic symptoms such as delusions and hallucinations, despite the fact that psychotic and neurotic symptoms commonly coexist, and there is a similar reluctance to diagnose any kind of nonorganic psychosis if there is evidence of brain disease. There is an even greater reluctance to diagnose two psychotic disorders simultaneously, or a mood disorder and a neurotic disorder simultaneously, despite the fact that depression and anxiety commonly coexist.

5. *The Hierarchy of Diagnoses*

This reluctance to make multiple diagnoses achieves formal expression as a diagnostic hierarchy. Organic disorders come at the top of this hierarchy. This means that if the patient has clear evidence of a brain disease—a tumor, for example—it is legitimate for them to have a wide range of other psychiatric symptoms, psychotic or neurotic, without any other diagnosis being required. Next in the hierarchy come the so-called functional or nonorganic psychoses. If the criteria for one of these disorders—schizophrenia, for example—are met, the patient may also be depressed or anxious, or have phobic or obsessional symptoms, without any additional diagnosis being required. They must not, however, have evidence of overt brain disease. Neurotic and stress-related disorders come at the bottom of the hierarchy, which means that diagnoses at this level cannot normally be made if psychotic symptoms or evidence of brain disease are also present. It is no accident, either, that this is the sequence in which disorders are listed in contemporary glossaries, and also the sequence of the decisions required by the flowcharts of computer programs for generating diagnoses from clinical ratings.

The requirements of this hierarchy are only relevant if the symptoms in question all appear at much the same time. If two groups of symptoms have a quite different time course from one another, it may be both legitimate and appropriate to make two independent diagnoses. In the past the hierarchy used to be implied rather than explicit. Now it is built into the operational definitions of individual disorders. At the same time it has become more complicated, particularly in situations where schizophrenic and affective symptoms coexist, or follow one another in quick succession.

It is also still uncertain whether the hierarchy is simply an artifact, a device for reducing multiple diagnoses to a minimum, or whether it is a reflection of a fundamental characteristic of psychiatric symptomatology. It is certainly true that a high proportion of people with psychiatric symptoms, but not all, move up and down the hierarchy as they relapse and remit. Phobic and obsessional symptoms almost invariably become more widespread and severe if depressive symptoms develop, for example, and the development of hallucinations and delusions is usually accompanied by affective and neurotic symptoms which do not wane until after the hallucinations and delusions have disappeared.

6. *The Boundary Problem*

There is an old axiom that the art of classification consists in 'carving nature at the joints.' This is an elegant way of saying that a good, or a valid, nosology places the boundaries between adjacent disorders, and between these disorders and health, where there are genuine discontinuities, either in symptomatology or in etiology. If the defining characteristic of the disorder is a fundamental biological abnormality of some kind it may be self-evident that a discontinuity is involved. An additional chromosome or an abnormal gene, for example, clearly involve qualitative deviations from normality. If, however, the defining characteristic of the disorder is still its syndrome, or a quantitative deviation of some kind, such as a raised blood pressure or a raised blood level of thyroxin, it is not self-evident

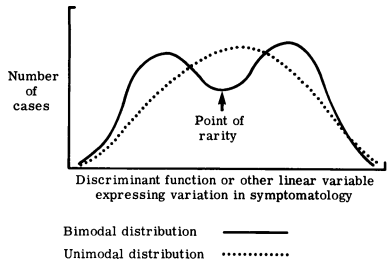

Figure 1
Variation in symptomatology expressed by a linear variable (reproduced by permission of Royal College of Psychiatrists from the *British Journal of Psychiatry* 1980, **137**: 325)

that there is a valid boundary between the syndrome in question and normality, or between this syndrome and others. Indeed, the fact that there are several rival operational definitions for many mental disorders, each of them encompassing slightly different populations of patients, strongly suggests that it is not obvious where the boundaries of those syndromes should be drawn.

Ideally, a point of rarity needs to be demonstrated between each syndrome and its neighbors (Kendell 1975). For example, it is widely assumed that schizophrenic and affective psychoses are distinct conditions. Indeed, the distinction Kraepelin drew at the end of the nineteenth century between what he called dementia praecox and manic-depressive insanity is still regarded as a fundamental feature of our contemporary classification of mental disorders. The fact that patients with a mixture of schizophrenic and affective symptoms—so called schizo-affective patients—are quite common does not in itself prove that there is no genuine boundary between the two syndromes, provided that these mixed forms are less common than the pure forms. The 'greys' must be less common than the 'blacks' and the 'whites'; or to express the issue graphically, the distribution of symptomatology along an axis extending from typical schizophrenia through schizo-affective states to typical affective psychoses must be bimodal rather than unimodal (see Fig. 1). A number of attempts have been made to address this issue using discriminant function analysis and the results are summarized below.

An alternative to demonstrating discontinuities in the distribution of symptomatology is to identify a discontinuity in the relationship between symptomatology and some other important variable, such as prognosis or response to treatment. For example, it is well established that, on average, the outcome of affective psychoses is much better that that of schizophrenic illnesses, both in the short and the long term. The key issue, however, is not whether the two disorders have a significantly different outcome, but whether, when the variation in symptomatology is

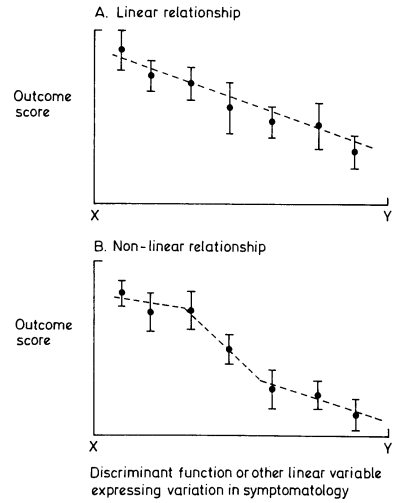

Figure 2
Relationship between symptomatology and outcome when symptomatology is converted to a linear variable (reproduced by permission of Royal College of Psychiatrists from the *British Journal of Psychiatry* 1980, **137**: 325)

portrayed as a linear dimension (as described above) there is a discontinuity at some point in the relationship between symptomatology and outcome, or whether there is simply a linear improvement in outcome as the symptomatology becomes increasingly more affective (see Fig. 2).

To date, most attempts to demonstrate valid boundaries by these means have been concerned with the boundaries between what used to be called endogenous and reactive depressions, between schizophrenic and affective psychoses, and between major depressive disorder and normality. The attempts to demonstrate a valid boundary between endogenous and reactive depressions were unsuccessful; when representative populations of patients were studied the distribution of individual patients along the discriminant function was consistently unimodal rather than bimodal. Partly for this reason, contemporary glossaries such as *ICD-10* (WHO 1993) and *DSM-IV* (American Psychiatric Association 1994) no longer distinguish between these two types of depressive illness. Similar attempts to demonstrate a valid boundary between schizophrenic and affective psychoses, either in symptomatology or in the relationship between symptomatology and outcome, were also unsuccessful, but a bimodal distribution of scores was subsequently demonstrated on a linear function discriminating between schizophrenia and all other psychiatric disorders combined. More recently, an attempt to identify a valid boundary between major depression and normality by demonstrating, in a population of 2,000 twins, a discontinuity in the relationship between depressive

symptomatology and either outcome or concordance between twin pairs ended in failure (Kendler and Gardner 1998).

7. *The Consequences of Failure to Validate Boundaries*

None of these studies proves that there are no valid boundaries between the syndromes in question. For one thing, it is always possible that more extensive data, or a different choice of criterion, might have provided a different result. Cumulatively, however, the failure of most of these attempts to demonstrate valid boundaries between syndromes emphasizes that it is unsafe to assume that there are valid boundaries there to be found. We must take seriously the possibility that we are not yet 'carving nature at the joints,' or even that there are no 'joints' there, and that we are trying to impose an artificial set of discrete categories on what is in reality continuous variation.

This state of affairs holds several important implications for the differential diagnosis of psychiatric disorders. The first is the importance of recognizing that most of the categories of disorder recognized in contemporary nosologies, even those such as schizophrenia with pedigrees dating back to the nineteenth century, have not yet been adequately validated, and that their formal definitions are simply agreed conventions. The second is a consequence of the first. All decisions about diagnosis, and all discussions about differential diagnosis, should take place in a setting in which it is agreed which set of diagnostic categories and corresponding set of operational definitions (in effect, which glossary) are being used, and with a clear understanding that both the categories and the definitions are arbitrary and inadequately validated, and therefore likely to be modified as understanding of the underlying etiology accumulates.

The third implication is a consequence of the likelihood that many existing boundaries, either between syndromes or between psychiatric disorder and health, are not situated at points of rarity in the distribution of symptomatology. To the extent that this is the case, relatively minor changes in the operational definition of a syndrome, which produce corresponding changes in the position of its boundaries, may result in quite large changes in the numbers of people, or patients, who fall within those boundaries. This is almost certainly why small changes in the definition of common disorders such as major depression or generalized anxiety—particularly changes in the number of typical symptoms required, or in their duration—have been associated with major changes in the recorded prevalence of these disorders in community surveys.

Finally, if it is indeed the case that there are no genuine boundaries within large tracts of the territory of psychiatric disorders, and we are imposing an artificial set of discrete categories on what is in reality continuous variation, psychiatrists will have to consider seriously the possibility of replacing their traditional categorical classification of disorders with a dimensional classification. This would never be done lightly, if only because the rest of medicine has always been committed to and content with categorical classifications of its disorders. If a dimensional format ever were to be adopted for mental disorders, however, the consequences for differential diagnosis would be profound. Disputes about which disorder or combination of disorders an individual was suffering from would be replaced by discussions about the precise locus in a multidimensional space at which that individual should be located. These discussions would probably also be calmer than the previous disputes, because in most cases it would be clear that the alternative loci under discussion were all quite close to one another, and not separated by dividing lines implying major differences in treatment or prognosis.

8. *Practical Issues in Differential Diagnosis*

As in other branches of medicine, some differential diagnoses are more critical than others. The most important functions of a diagnosis are to determine treatment and to predict outcome, and the most important differential diagnoses are those involving a disorder requiring urgent and specific treatment. All psychiatrists are keenly aware of the importance of diagnosing, or at least suspecting, cerebral tumors, and other brain diseases such as viral encephalitis and subdural hematoma, which may also present with psychiatric symptoms, as quickly as possible. Because these conditions may all present with depressive symptoms, with apathy or other behavioral changes, or even with psychotic symptoms, and because they all require urgent treatment, they enter into the differential diagnosis of a wide range of psychiatric disorders. Many reputations for diagnostic acumen are, rightly, based on successful detection of relatively uncommon brain diseases of this kind, or of metabolic disorders such as thyrotoxicosis and Cushing's syndrome, which underlie some anxiety states and depressions. It is also important to detect brain diseases such as Creutzfeldt–Jakob disease, which sometimes present with psychiatric symptoms even though there is not yet any treatment capable of altering the course of the disease, because of the implications for prognosis. If someone has an illness which is inevitably going to result in permanent disability or death, the sooner that is established the better, both for the patient and the family.

There are other circumstances in which the critical diagnosis in the differential diagnosis between brain disease and psychiatric disorder is the latter rather

than the former. Dementias are common in old age and most are associated with progressive decline to dependency and death. Sometimes, however, depressive illnesses present with such widespread evidence of intellectual impairment that they closely resemble a dementia. Because depressions are eminently treatable it is vital that these 'depressive pseudodementias' are quickly recognized and treated, and depression is therefore an important part of the differential diagnosis of dementia. Similar considerations apply to rare curable dementias, such as those associated with a deficiency of thyroid hormone or vitamin B12.

Traditionally, much time and intellectual effort have been devoted by psychiatrists to distinguishing accurately between schizophrenic and affective psychoses. To some extent this is appropriate. The two disorders require different treatments and there are major differences in long-term prognosis between them. Some disagreements are due simply to undisclosed differences in diagnostic criteria which would be resolved by the adoption of an agreed operational definition. However, a more fundamental conceptual issue may also be involved. Although there are many instances, including the differential diagnosis between schizophrenic and affective psychoses, where no valid boundary has yet been demonstrated between the two syndromes, psychiatrists often still behave as though they were discussing the distinction between two 'disease entities.' In reality, if the patient's symptoms are 'schizoaffective,' it may not matter much which side of an arbitrary boundary they are deemed to lie, because their therapeutic needs and their prognosis may both be genuinely intermediate between those of the two classical syndromes.

To diagnose schizophrenia is always a grave decision. Although some people develop schizophrenic illnesses which resolve completely, even without treatment, and never recur, for many more a schizophrenic episode is the prelude to a life of invalidism and the death knell of hopes and ambitions. It is understandable, therefore, that many psychiatrists are reluctant to diagnose schizophrenia, particularly in an adolescent, unless the evidence is incontrovertible. If some less doom-laden diagnosis, such as depression, is plausible they will opt for that, and hope that time will justify their decision. Scandinavian psychiatrists sometimes make a diagnosis of psychogenic psychosis with similar motives. Recently, however, psychiatrists have become increasingly convinced by the evidence that early treatment with neuroleptics improves the long-term outcome of schizophrenic illnesses and reduces the risk of a 'defect state.' Although this evidence is not yet buttressed by random allocation trials it is sufficiently strong to swing the balance of advantage in patients with ambiguous symptomatology in favor of early diagnosis. A similar mix of social and therapeutic considerations has an influence on differential diagnosis in other areas too.

Other important issues of differential diagnosis arise when patients with psychiatric disorders present to other medical disciplines, as they commonly do. Depressive illnesses are extremely common at all ages and depressed patients may present with a wide variety of complaints, particularly if they are reluctant to accept that they may have a psychiatric disorder. They may complain of 'feeling off color', of loss of appetite or weight, of difficulty sleeping, or of ill-defined pain almost anywhere in the body. As a result, the correct diagnosis is often missed unless questions are asked about characteristic depressive symptoms such as difficulty concentrating or loss of the ability to enjoy things. Phobic anxiety is another common disorder which often presents with somatic symptoms such as palpitations, or recurring bouts of upper abdominal discomfort or difficulty breathing. Unless specific questions are asked about the environmental circumstances in which these symptoms develop, their phobic nature may easily be missed and the patient may be subjected to a long series of irrelevant investigations. In both cases the fundamental problem is that the doctor concerned does not realize, or remember, that a depressive illness or phobic anxiety should be important elements in the differential diagnosis of the patient's presenting complaints, even though these may be entirely somatic.

Bibliography

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. (DSM-III), American Psychiatric Association, Washington, DC

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. (DSM-IV), American Psychiatric Association, Washington, DC

Feighner J P, Robins E, Guze S B, Woodruff R A, Winokur G, Munoz R 1972 Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry* **26**: 57–63

Kendell R E 1975 *The Role of Diagnosis in Psychiatry*. Blackwell Scientific, Oxford, UK

Kendler K S, Gardner C O 1998 Boundaries of major depression: An evaluation of DSM-IV criteria. *American Journal ofPsychiatry* **155**: 172–7

World Health Organization 1993 *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization, Geneva, Switzerland

R. E. Kendell

Differentiation: Social

Social differentiation is an important attribute of society, especially modern society (see also *Societies, Types of*). The concept of differentiation spread within sociology since Herbert Spencer's evolutionary theory of societal development 'from incoherent homogeneity

to coherent heterogeneity'. Later on, Emile Durkheim, Georg Simmel, Talcott Parsons, and Niklas Luhmann were important proponents of the concept. Other social thinkers, such as Karl Marx and Max Weber, who did not use the term 'differentiation' prominently nevertheless contributed to a proper understanding of the social structures and dynamics it designates. At the start of the twenty-first century, the theoretical and, more and more, empirical debate about social differentiation is continued, especially by American 'neo-functionalists' in their discussion of Parsons' legacy, and some German sociologists who keep up the critical debate about Luhmann's theory of society.

Social differentiation means both a process and a structure. In structural terms, it refers to the fact that a unit of analysis, such as a society, consists of a number of distinct parts. These parts may be of the same kind, such as families as the basic components of tribal societies. Or the parts may be different, such as the subsystems—economy, politics, mass media, etc.—making up modern society. As a process, social differentiation is the dynamic that brings about and changes a given structure of differentiation.

Two fundamentally diverging views of the social differentiation of modern society have been put forward in sociological thought. These views do not strictly contradict, but balance each other in important respects. In the following, the partly opposing but mostly complementary relationship between both views will be sketched, with respect to the features, causes, and consequences of social differentiation. The focus of attention will be put on the differentiation of modern society.

1. *Social Differentiation as Decomposition or Emergence*

For Durkheim and Parsons as the most important representatives of the first view of social differentiation, it is a process of decomposition of a functionally diffuse unit—a role, or an institution—into several, functionally more specific units. This idea is modeled after the division of labor in work organizations, a process which fascinated nineteenth-century thinkers (see also *Labor, Division of*). Durkheim's major illustration is the widespread specialization of occupations in industrial society. In Parsons' more abstract analysis, the theoretical and structural-functional analytical framework of society is composed of four principal subsystems (economy, polity, societal community, and fiduciary system) each of which fulfills one of the four fundamental functional prerequisites of societal reproduction. Social differentiation as a process of social change tends towards a successive decomposition of each of these subsystems into four second-level subsystems (see also *Exchange: Social*). Parsons called this 'functional differentiation,' a term that is still used to describe the type of differentiation of modern society. His analytical perspective, Parsons suggests, is validated empirically over time by the concrete differentiation of social roles and institutions.

In structural terms, the decomposition perspective on social differentiation highlights the advantages of specialization for society, which are basically the same as on the organizational level. Specialization greatly improves performance. For instance, if formal education in schools is differentiated from the family, where education was previously concentrated, both sides profit in terms of functional efficiency and effectiveness—teachers can rely on primary socialization accomplished by parents and are able to focus on secondary socialization, and vice versa.

In contrast to this perspective of social differentiation as a fruitful division of labor, Weber portrayed the birth and composition of modern society as the emergence of a number of autonomous 'value-spheres.' One after another, science, law, art, politics, economics, sexuality, and others freed themselves from their former domination by religious ideas. These value-spheres became societal domains of their own which cultivate their particular central value without concern for the value orientations of other domains. Science, for instance, does not care about economic profits, aesthetic beauty, legality, or the conformity of its truths with religious doctrine. Weber (1919, pp. 27–8) speaks graphically about the 'polytheism' of value-spheres in modern society.

Luhmann formulates this second perspective on social differentiation even more to the point. To him, society consists of about a dozen subsystems in each of which communication is guided by a specific binary code. For example, the science systems binary code is true/untrue, the legal systems legal/illegal. The binary code of each subsystem establishes a highly selective searchlight onto the social world, illuminating certain corners of it and leaving the rest in the dark.

The binary codes generate self-referentially closed universes of communication. Thus, a societal subsystem consists of chains of communications each of which refers to other communications of the same kind. For this fundamental feature, Luhmann picks up the concept of 'autopoiesis' from biological thinking (see also *Biology's Influence on Sociology: Human Sociobiology*). To illustrate, in the social system of science assertions of truth are the specific kind of communication. These assertions, usually made in scientific publications, lead to other such assertions, and—within science—not to political declarations or declarations of love. Any scientific publication, if it provokes any reactions at all, is either affirmed or rejected by later publications. Which of these two eventualities happen does not matter with respect to the reproduction of science as a societal subsystem. In both instances science continues as an open-ended sequence of publications.

Luhmann's conceptualization is better suited than Weber's to grasp the basic dynamics of continual

reproduction of a societal subsystem or value-sphere. For both, each subsystem can take only its own code-determined perspective on social events. In science, nothing but its own concerns are important, and the same holds true for all other subsystems. The overall result is an effective multiplication of society with respect to the communicative meanings of events. An event, such as a train accident, looks totally different from the point of view of the economy, compared to the point of view of science, compared to the point of view of the health care system, compared to the point of view of the mass media, and so on. Borrowing a term from philosopher Gotthard Guenther, Luhmann (1997) calls this the 'polycontexturality' of modern society.

Although all subsystems have, in this sense, no 'window' to the social world outside, this basic autonomy goes along with an openness to their societal environment with respect to resources and functional performance. The science subsystem, for instance, needs financial resources among other things, and gets them from the political subsystem, or the economy. In the other direction, science provides these and other societal subsystems with knowledge. Since all societal subsystems make essential contributions to the reproduction of many others and none of them can substitute any other one, each is a functional prerequisite of modern society. Modern society would break down without economy, but also without mass media, or without a health care system, and the latter could not be replaced, for example, by politics or education.

Despite these manifold mutual dependencies, Luhmann insists that it makes no sense to understand the ensemble of societal subsystems as a division of labor. In a division of labor, the subsystems as parts are teleologically oriented towards the performance and maintenance of the whole society. Thus, the whole comes first, historically and logically, and decomposes itself for its own benefit. By contrast, Weber and Luhmann emphasize that social differentiation consists of a simultaneous birth and liberation of the parts. They emerge and become autonomous from each other and from the whole. Even more, the whole vanishes in the sense that it becomes nothing more than the—often tension-ridden and antagonistic—interrelationship of the parts.

2. *Causes of Social Differentiation*

Both analytical perspectives identify different driving forces of social differentiation. Four kinds of causal factors can be distinguished, improvements of performance, evolution, cultural ideas, and interests of actors.

The first causal factor, improvements of performance by specialization, was especially emphasized by Parsons and structural-functionalist studies of social differentiation (see also *Functionalism in Anthropology*). Such specialization is initiated either by performance deficits of existing structures of differentiation, or by opportunities to improve a satisfactory level of performance. An example of the problem-driven dynamic is the previously mentioned differentiation of the education system. Schools, and a duty of all children to attend them, were established because other societal subsystems, especially the economy, needed a general level of cognitive skills and social discipline within the population that could no longer be provided by the family alone. A differentiation dynamic driven by opportunities was the spur for differentiation of the radio and television sector within the mass media. Entirely new possibilities of public communication opened up by the technological innovations of radio and television, rather than any perceived performance deficits of newspapers, caused the sectoral differentiation of roles and organizations within mass media.

By contrast to a planned division of labor within organizations, on the societal level performance improvements by differentiation occur often as unintended results of actions motivated by quite different reasons. In this sense, Parsons already relied on evolution as the mechanism that brings about performance improvements (see also *Social Evolution, Sociology of*). In contrast, Luhmann gives up all connotations of evolution with progress. Adopting neoDarwinian evolutionary thinking from biology, he conceives of evolution as an interplay of three mechanisms, variation, selection, and retention. Each societal subsystem has its own specific evolutionary mechanisms. For instance, in the science subsystem most publications are variations of the existing body of knowledge. Selection happens whenever one of these publications is cited affirmatively. Retention occurs when this new knowledge is incorporated into overview and review articles and into textbooks for students. As in biological evolution, most variations are not selected, and most of those selected do not reach the stage of retention. This evolutionary cognitive growth and differentiation goes along with a social differentiation of scientific specialties.

Luhmann does not deny that this evolution of the science system may bring about performance improvements usually called 'scientific progress.' But for him, this is no necessary outcome of evolutionary differentiation. Evolution may also lead to dysfunctional differentiation. For instance, the growing quest for interdisciplinary research indicates an over-specialization in more and more scientific fields. An understanding of evolutionary differentiation as progress becomes even more dubious in other societal subsystems, such as art, politics, or the mass media. But even if evolutionary differentiation on the level of each subsystem would result in performance improvements of all of them, the overall result for society would not add up to progress but only to an increase of societal

complexity which manifests itself particularly in a growing destabilization of society. Each evolutionary change within one subsystem is also a change of the societal environment of all other subsystems. For them, often adaptive changes become necessary each of which again changes the societal environment, with new adaptive pressures arising. Thus, social differentiation becomes permanently autodynamic. Until now, these dynamics often brought about increasing differentiation. But Parsons' statement that this is a linear and unlimited social process is clearly wrong as not only Luhmann's theoretical reflections point out but also empirical cases presented especially by the 'neofunctionalists' demonstrate.

A third driving force of social differentiation was already mentioned with respect to Weber: an autodynamic rationalization of cultural ideas. In premodern societies the values that today constitute the guiding principles of the various societal subsystems were closely interwoven with each other and, moreover, integrated into an encompassing order of religious values. The differentiation of the principal subsystems of modern society is understood by Weber as a gradual but irresistible separation of economic, political, scientific, aesthetic, legal, erotic, and other values, first, from the religious context, and second, from each other. This dynamic is triggered and sustained by the compelling intellectual urge to contemplate extensively the implications and consequences of each value for everyday life. When this rationalization reaches the point where each value is rigorously reflected in its own terms, without regard to others, soon a self-referential closure of value-spheres occurs, giving way to a corresponding differentiation of roles, organizations, and subsystems.

The fourth driving force of social differentiation is the general interests of individual, collective, or corporate actors who are involved in constellations with each other (see also *Interests, Sociological Analysis of*). One explanation of an ongoing division of labor mentioned by Durkheim is growing 'social density' resulting in an intensification of competition among actors. This constellation evokes an interest to preserve one's own domain of action, be it an occupation for an individual or a market position for a firm. This interest, in turn, leads many of the respective actors to find new niches for themselves where they are relieved from high competitive pressure. The aggregate effect of such domain-preserving activities is social differentiation on the level of roles and organizations.

Other general interests of actors that often stimulate dynamics of social differentiation are the preservation or extension of one's autonomy, of one's control of other actors, and of one's resource base. Some professions, especially those guided by such interests were strong carriers of the differentiation of certain societal subsystems including science, law, education, and health care (see also *Professions, Sociology of*). Sometimes, a particular kind of social differentiation is the explicit goal of certain groups of actors who believe that this will serve their interests best. In other cases, social differentiation is an unnoticed, unforeseen, or unwanted effect of actors' pursuit of their interests. Outcomes often diverge markedly from intentions because usually the relevant constellation consists of actors in quite different positions: catalyzers and promoters of the respective process of change; followers who join in when a 'critical mass' has been mobilized; defenders of the *status quo* who resist any change; more open-minded defenders with whom compromises are possible; and those who are indifferent or indecisive at the beginning and whom the promoters of change as well as the defenders of the *status quo* try to recruit as allies.

Each of these four kinds of driving-forces is causally relevant in the dynamics of social differentiation, although the combination varies greatly from one case to another. Often these factors are interrelated. Thus, what actors understand as their interests are framed by cultural ideas. Cultural ideas as well as actors' interests are explicit or implicit criteria to judge the performance of a given differentiation structure; and *vice versa*, performance deficits provoke actors' interests. Finally, the evolution of societal subsystems creates new challenges, restrictions, and opportunities for the pursuit of interests and the spelling out of cultural ideas.

3. *Consequences of Social Differentiation*

The differentiation of modern society became a central theme not just of sociological thought but also of public discourse in the nineteenth century, and still is today. This is because social differentiation has far-reaching consequences for society at large, as well as for each individual member of it. These consequences are neither totally positive nor totally negative but a mixed blessing.

To begin with, the modern individual profits from role differentiation and an enormous increase of options in every sphere of life, from economic markets for goods and services to opportunities of political participation, sports activities, lifelong education, and cultural life. The abundance of options is the result of performance improvements in all societal subsystems and of an inclusion of all members of society into the status of customers, clients, and public of the subsystems. In connection with the polycontexturality of modern society, role differentiation and the growth of options allow each person to become an individual in the proper sense of the word: a person who creates a unique biography for himself or herself. However, options are only one side of a person's 'life chances' (Dahrendorf 1979). The other side includes 'ligatures,' meaningful attachments which tie a person to certain institutions, cultural worlds, social groups, or communities. In this respect, the functional differentiation of

modern society seems to neglect important psychic needs of persons. Ligatures are eroded by 'polycon texturality' and individualization. As a consequence, anomy, alienation, and other experiences of meaninglessness are present in modern society, not just as individual misfortunes and temporary phenomena, but as structurally produced permanent vulnerabilities of everyone.

Looking at society at large, social differentiation exhibits the same kind of ambivalence. First, it is obvious that functional differentiation is evolutionary highly successful. All kinds of premodern, traditional societies were overrun by those national societies from Europe and North America that converted to functional differentiation since the sixteenth century. Worldwide, there was no alternative to the adoption of functional differentiation. As shown by the socialist experience, those countries that tried to stop halfway failed. Countries where the political system attempted to control all other subsystems were beaten by the fully developed functional differentiation of Western countries, as Parsons, among others, had anticipated in the 1960s.

However, global supremacy of a fully-fledged functional differentiation implies serious problems. Their common denominator is an endangering of the integration of modern society. More precisely, systems integration, social integration, and ecological integration must be distinguished.

Systems integration seems to be the least problematic aspect. Coordination and cooperation among different subsystems of modern society is obviously possible, despite the fact that each of them is governed by its own self-referential binary code. But there are manifold interrelationships by which subsystems adapt to each other in a decentralized way. Where this does not suffice, representatives of subsystems, such as the respective interest groups, meet and bargain for common solutions in the political arena. Finally, political guidance may support or promote intersystems adaptation and bargaining among subsystems. Thus, problems of systems integration remain fragmented and manageable for modern society.

With social integration, it is different (see also *Integration: Social*). The integration of individual members into modern society may be precarious as a result of the already mentioned erosion of ligatures. More important could be the fact that the increase in individual options might put excessive demands on all subsystems' abilities in the long run. In this respect, social inequality and the subsystems' unlimited drive to improve their performance work hand in hand (see also *Inequality*). In conflicts among those who, as individuals, profit more and those who profit less from a particular subsystem's performance, the former will try to avoid redistribution. One way to achieve this is a growth of the subsystem's performance so that those whose claims were not satisfied can be served somewhat better without those who were better of giving up anything. For instance, children from the lower classes have got access to higher education. But this has been no zero-sum conflict in which the higher classes have lost some of their educational opportunities but a positive-sum constellation because the higher education sector has been extended—which has fitted also the interests of the respective groups of teachers and professors and has been reinforced by the cultural imperative of 'progress.' Such dynamics, driven by complementary interests and backed by powerful cultural ideas could go on forever, were there no limits to subsystems' growth. But because these limits exist, for example with respect to resources, continuing inequalities contain a potential for serious social conflict to which functional differentiation has no ready answer.

Finally, the functional differentiation of modern society is a major cause of the growing problems of ecological integration. Modern society's path away from a 'sustainable' relationship with its natural environment has occurred as a result of two basic characteristics of all societal subsystems, their growth tendencies, and the neglect of ecological aspects by their binary codes. The modern economy is the most obvious example, but the same is true of the family, the subsystem primarily responsible for population growth. No societal subsystem exists now that takes care of the ecology, only social movements address their protest to the political system. But this subsystem, following its binary code of maintaining and increasing power, must listen to many other demands from society and so often disregards ecological concerns.

Thus, the functional differentiation of modern 'world society' is a high-risk outcome of social evolution. No alternative to this type of social differentiation is apparent as an option for the future. But whether modern society can survive in this way or will ruin itself sooner or later is an open question that sociological theories of social differentiation cannot answer. Theorists can only observe what has happened previously and infer warnings from this.

Bibliography

Alexander J S, Colomy P (eds.) 1990 *Differentiation Theory and Social Change*. Columbia University Press, New York

Dahrendorf R 1979 *Life Chances*. University of Chicago Press, Chicago

Durkheim E 1893 *The Division of Labor in Society*. Reprinted (1964) Free Press, New York

Luhmann N 1977 Differentiation of society. *Canadian Journal of Sociology* **2**: 29–53

Luhmann N 1997 *Die Gesellschaft der Gesellschaft*. *(The Society of Society)*. Suhrkamp, Frankfurt, Germany

Parsons T 1971 *The System of Modern Societies*. Prentice Hall, Englewood Cliffs, NJ

Parsons T, Smelser N J 1956 *Economy and Society*. Routledge, London

Schimank U 1996 *Theorien gesellschaftlicher Differenzierung. (Theories of Social Differentiation)*. Leske and Budrich, Opladen, Germany
Simmel G 1900 *Philosophie des Geldes*. Reprinted (1977) Duncker and Humblot, Berlin. Translation (1978) *Philosophy of Money*. Routledge, London
Weber M 1919 *Wissenschaft als Beruf. (Science as Vocation)*. Reprinted (1967) Duncker and Humblot, Berlin

U. Schimank

Diffusion and Random Walk Processes

Random walk and diffusion models form one of the major classes of models in the response time domain. They best apply in situations in which subjects make two-choice decisions that are based on a single, 'one shot', cognitive process, decisions for which response times do not average much over one second. The basic assumption of the models is that a stimulus test item provides information that is accumulated over time towards one of two decision criteria, each criterion representing one of the two response alternatives. A response is initiated when one of the decision criteria is reached. The researchers who have developed random walk and diffusion models take the approach that all aspects of experimental data need to be accounted for by a model. This means that models should deal with both correct and error response times, with the shapes of the full distributions of response times, and with the probabilities of correct versus error responses. It should be stressed that dealing with all these aspects of data is much more of a challenge than dealing with only one dependent variable.

Random walk models have been prominent since the 1960s (Laming 1968, Link and Heath, 1975). Diffusion models, continuous versions of the random walk, appeared in the late 1970s (Ratcliff 1978, 1980, 1981). The random walk and diffusion models are close cousins and not competitors of each other as they are with other classes of models (e.g., accumulator models and counter models, e.g., LaBerge 1962, Smith and Vickers 1988).

The earliest random walk models assumed that the accumulation of information occurred at discrete points in time, each piece of information either fixed or variable in size (e.g., Laming 1968, Link and Heath 1975). The models succeeded in accounting for mean correct response time and accuracy, and were sometimes successful with mean error response times, but they rarely addressed the shapes of response time distributions.

In diffusion models, the accumulation of information is continuous over time. Ratcliff's (1978) diffusion model is illustrated in Fig. 1. One parameter of the model is the mean rate of accumulation of

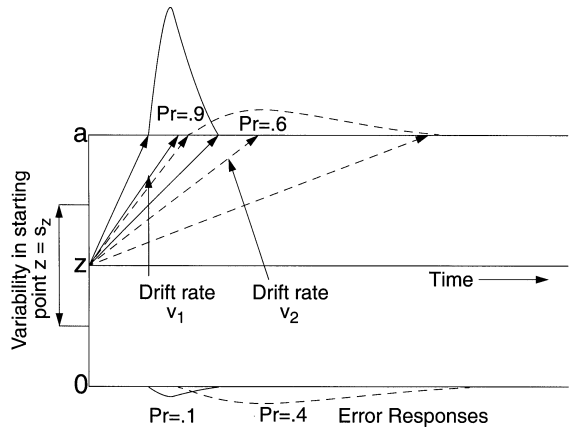

Figure 1
Illustration of Ratcliff's diffusion model

information. In the figure, two processes are shown one with mean drift rate v_1 (solid arrows) and the other with mean drift rate v_2 (dashed arrows). Drift rate varies around the mean as shown by the arrows for the average, the fastest and the slowest processes for the means v_1 and v_2 (these are meant to be illustrative; real processes are highly variable). Besides mean drift rate, the other parameters of the model are the separation of the boundaries (a), the starting point (z), the variability in drift rate across trials (η), and a parameter (T_{er}) that represents all the nondecisional components of response time, such as encoding and response preparation. There is also a parameter representing variability in processing (drift) within a trial (s). This is a scaling parameter; if it were doubled, for example, all the other parameters could be changed to produce the same identical predictions as before the change. s is a fixed parameter(not a free parameter) in fits of the model to data.

The basic diffusion process is used in many areas of research, including heat transfer, Brownian motion, neural processes, financial calculus, and fluid mechanics. As a result, the process has been examined in great detail and many explicit solutions for the distributions of the 'first passage' times ($G(t, \xi)$) and probabilities of reaching a boundary ($P(\xi)$) (where ξ is drift rate) are available.

$$P(\xi) = (e^{-(2\xi a/s^2)} - e^{-(2\xi z/s^2)})/(e^{-(2\xi a/s^2)} - 1)$$

$$G(t, \xi) = P(\xi) - \frac{\pi s^2}{a^2} e^{-(z\xi/s^2)} \times \sum_{k=1}^{\infty} \frac{2k \sin(k\pi z/a)\, e^{-\frac{1}{2}(\xi^2/s^2 + \pi^2 k^2 s^2/a^2)t}}{(\xi^2/s^2 + \pi^2 k^2 s^2/a^2)}$$

But stochastic differential equations are difficult, and numerical solutions are required for examination of

more complex characteristics of the diffusion process, such as the effects of changes in drift rate and boundary positions during the course of information accumulation. Smith (2000) has developed alternative numerical methods to solve for the first passage times above; these are based on recursive expressions for the passage of a process to a response boundary. The idea is that the probability that a process will hit a response boundary at time t is the integral over time t_1 of the product of the probability that it first hit the boundary at time t_1 and the probability that it reaches the boundary starting at time t_1 and ending at time t. This solution allows the calculation of first passage times even when drift rate and boundary positions are not fixed. Another solution is to model the diffusion process with a discrete random walk process with very small step sizes, using Monte Carlo methods to obtain response probabilities and first passage time distributions for correct and error responses.

Ornstein-Uhlenbeck (OU) models are variants of diffusion models in which drift rate is offset by a decay term that tends to push the accumulation of information back towards the starting point. The size of the decay term is a linear function of the distance moved away from the starting point. OU models have been used to describe the behaviors of single neurons, they have been applied to simple reaction time tasks (Smith 1995), and they have been applied to decision-making and choice behavior (Busemeyer and Townsend 1993). Ratcliff and Smith (1999) have found that the OU process mimics the diffusion process, and vice versa, in many situations, but further work is needed to decide whether and how the two kinds of models can be discriminated.

1. *Diffusion Models and Empirical Data*

1.1 *Scaling and Distribution Shape*

The two dependent variables, response time and accuracy, have different scale properties. Response time has a minimum value and its variance increases as response time increases. Accuracy is bounded at probabilities 0.5 and 1.0, and as probability correct approaches 1.0, variance decreases. Diffusion models account for these scale properties automatically as a result of the geometry of the diffusion process. When mean drift rate is high, the probability of a correct response is near 1.0 and decision processes approach the correct boundary quickly with little spread in arrival times (e.g., the processes with mean v_1 in Fig. 1). When mean drift rate is nearer zero (processes with mean v_2 in Fig. 1), variability leads some processes to hit one boundary, other processes to hit the other boundary, and accuracy nears 0.5; the arrival times at boundaries are long and highly variable. In addition to these scale properties, the geometry of the diffusion process also gives response time distributions skewed to the right; as mean response time increases (e.g., from drift rate v_1 to v_2), the fastest responses slow a little and the slowest responses slow a lot.

1.2 *Speed vs. Accuracy*

Human subjects can manipulate their speed performance relative to their accuracy performance. For example, if in some blocks of an experiment they are instructed to respond as quickly as possible, and in other blocks they are instructed to respond as accurately as possible, mean response times can vary between the two kinds of blocks by as much as 500 ms and mean accuracy can vary by 10 percent (Ratcliff and Rouder 1998). Ratcliff's diffusion model can account for these differences with only the boundary separation parameter; moving the boundaries close together produces fast responses for which variability in drift can cause large numbers of errors; moving the boundaries apart produces slower responses that are more likely to be accurate.

1.3 *Error Response Times*

A major problem that restricted the development of random walk and diffusion models was their inability to accurately predict error response times. According to the models, errors and correct responses should have the same mean response times when the correct and error boundaries are equidistant from the starting point. However, experimental data showed that in some situations, errors are faster than correct responses and in some situations they are slower, and both patterns can even occur in different conditions in the same experiment (Luce 1986, Ratcliff et al. 1999, Ratcliff and Rouder 2000, Smith and Vickers 1988). In early attempts to deal with this problem, Laming (1968) added variability to the starting point to produce fast errors, and Link and Heath (1975) used variability in the sizes of the steps toward correct and error boundaries to account for fast errors in some experiments and slow errors in others.

Recently, a more complete solution has been provided by Ratcliff et al. (1999; see also Ratcliff 1981), showing that the complete patterns of error versus correct response times can be explained by combining variability in starting point and variability in drift rate across trials. This solution to the problem was so late in coming because earlier applications of diffusion models did not have enough computer power to automatically search the complete parameter space (e.g. Ratcliff 1978).

1.4 *Time Course of Accumulation of Information*

Diffusion and random walk models are explicitly based on a process that accumulates information over

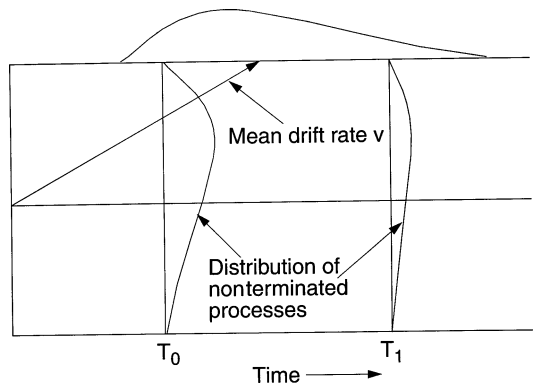

Figure 2
Illustration of the distribution of nonterminated processes within a diffusion process at two times T_0 and T_1. The top curve represents the distribution of finishing times

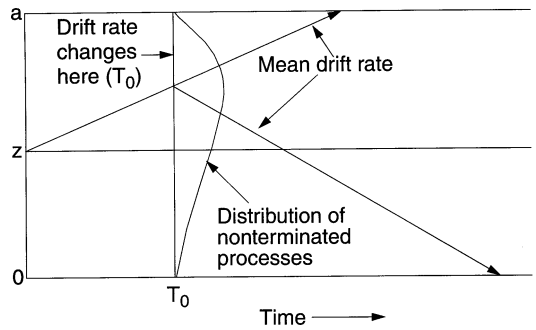

Figure 3
An illustration of the way drift rate might change in the diffusion process. At some point T_0 the drift rate changes and the distribution of nonterminated processes becomes a distribution of starting points for the second phase of the diffusion process

time. Empirical examination of the time course of processing is possible with deadline and response signal procedures in which subjects are asked to respond at experimenter-determined times (a signal is presented after each test item to indicate exactly when a response to the item is to be made). The dependent variable is accuracy. The first diffusion model approach to this kind of data was to assume that there are no boundaries on the diffusion process; when the signal to respond is given, the position of the process is evaluated and, if it is above the starting point, one response is given; if it is below the starting point, the other response is given. This assumption leads to a simple formula for d' (see *Signal Detection Theory*) as a function of time:

$$d'(t) = d'_a / \sqrt{(1 + s^2/(\eta^2 t))}$$

An alternative approach (Ratcliff 1988), more in line with subjects' intuitions, is to assume that responses come from a mixture of two kinds of processes: processes that have not yet reached a boundary, like those just described, and processes that have reached a boundary. Figure 2 shows what happens with a test item for which the mean drift rate is v. The figure shows the distribution of the processes not terminated with this mean drift rate at time T_0. Most of these processes are above the starting point z, and so, if a signal to respond is given at time T_0, most responses from processes not yet terminated at a response boundary will produce the response represented by the top boundary. At time, T_1, responses will be a mixture of processes with mean drift rate v, which have terminated, and other processes which have not.

1.5 Nonmonotonic Functions

With some kinds of test stimuli, it can be shown empirically that the accumulation of information does not proceed at a constant rate. One source of information drives the process early in the time course and a different source of information drives the process later (e.g., Ratcliff and McKoon 1982). If the two sources of information contradict each other, then drift rate changes substantially during the time course of processing. Typically, it is the latter source of information that yields a correct response, and so accuracy first decreases from chance, and then later increases above chance, showing a nonmonotonic function across time. Figure 3 shows how this is handled in Ratcliff's diffusion model (Ratcliff 1980). When the second source of information becomes available, at time T_0, mean drift rate for nonterminated processes changes from a positive to a negative value and the distribution of nonterminated processes at T_0 serves as the starting point for the second process.

1.6 Constraints on Diffusion Models

In many paradigms, accuracy is high and differences in accuracy values across experimental conditions are too small to be statistically significant. The consequence is that the effects of experimental manipulations can be observed only in response time. Data like this do not provide strong tests of diffusion models because they can be fit with any of a number of different parameter values.

Strong tests are provided when the models must account jointly for a broad range of response times (and the shapes of the response time distributions) and accuracy values from experimental conditions that are

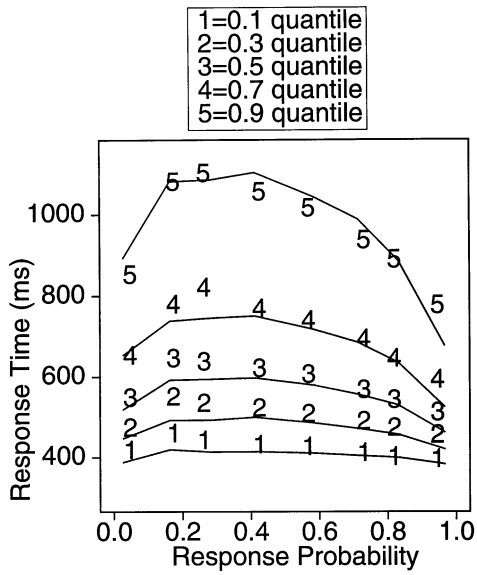

Figure 4
A quantile-probability function. The points represent the quantiles from the data and the solid lines the fit of a model to the data. Errors are to the left of the 0.5 probability point and correct responses corresponding to the errors to the right (mirror images). Each vertical line of quantiles represents a different condition in the experiment

manipulated in such a way that subjects cannot alter response criteria as a function of test item type (e.g., number of repetitions of words in a list where some studied items have few repetitions, some many). The manipulations should yield accuracy values varying from near floor (50 percent correct) to near ceiling (100 percent correct).

The issue arises of how to display this amount of data. Latency-probability functions have been used to simultaneously display mean response times and probabilities, with mean response time plotted on the y-axis and response probability on the x-axis. Different conditions are different points on the function, each condition represented by a pair of points, one for correct responses at probability p and a mirror image for errors at probability $1-p$. A latency-probability function gives a good picture of how accuracy and response time covary, showing, for example, whether response time changes quickly in conditions for which accuracy is high and changes little, or whether errors are faster than correct responses.

However, latency-probability functions do not show information about the shapes of response time distributions. One way to show this information is to plot the quantile response times for each of several quantiles (e.g., fifths) of the distributions as a quantile-probability function. Figure 4 shows an example using five quantiles. The numbers represent the data and the solid lines show a theoretical fit of the diffusion model. The figure shows that the speed of the fastest responses (the 0.1 quantile) does not change much across conditions (i.e., across response probability values). Responses do slow from the conditions with high accuracy to the conditions with lower accuracy (where probability of a correct response is about 0.6), but they slow because the distributions spread out with the slowest responses (in the fifth quantile) increasing the most in response time.

2. *Applications*

Diffusion and random walk models have been applied to a wide variety of tasks. Here, we summarize the most frequent and salient applications.

2.1 *Recognitions Memory*

Diffusion models were first applied in the recognition memory domain (Ratcliff 1978). They provide a good account of correct response times, response probabilities, and the shapes of response time distributions. With the assumptions of variability in starting point and across-trial drift rates, they can also explain the patterns of correct vs. error response times.

2.2 *Signal Detection*

In one class of signal detection experiments, subjects are presented with stimuli and probabilistic feedback to those stimuli (e.g., Ratcliff et al. 1999). The diffusion model provides an extension to signal detection theory to explain the standard accuracy-based results as well as reaction time. Experimental manipulations allow accuracy to be varied from ceiling to floor and the experimental paradigm has served as a useful procedure for testing the diffusion model (Ratcliff and Rouder 1998, Ratcliff et al. 1999). This procedure also serves as one of the main procedures to test and evaluate models of categorization (e.g., Ashby 2000, Nosofsky and Palmeri 1997).

2.3 *Perception*

Ratcliff's diffusion model has been used to study the time course of processing when a stimulus is presented briefly, then masked (Ratcliff and Rouder 2000). In this situation, the amount of information coming into the decision process could either increase during stimulus presentation and then begin to fall during mask presentation, or the information coming into the decision process could be an integration of the

total information available from the stimulus, from its onset until the mask is presented. Only with the latter assumption could the data be well fit by the diffusion model.

2.4 Simple Reaction Time

Simple reaction time tasks require the subject to respond when a stimulus is presented. Smith (1995) used an OU model with assumptions about sustained and transient processing channels to model the effects of stimulus intensity in these tasks. The model was able to simultaneously account for the shapes of response time distributions as a function of stimulus intensity and the time course of the availability of stimulus information (see *Stochastic Dynamic Models (Choice, Response, and Time)*).

2.5 Decision Making

Busemeyer and Townsend (1993) developed a dynamic model of decision making using an OU process to represent the time course of processing for decision making under uncertainty. This model has been further developed to provide an integrative framework for understanding several seemingly diverse empirical phenomena in the area of multialternative preferential choice and also this kind of model has been extended to represent the time course of processing for multi-attribute decision making (see *Dynamic Decision Making*; *Sequential Decision Making*).

2.6 Categorization

Random walk and diffusion have been married to models of categorization to account for response time data in categorization. In exemplar-based models (Nosofsky and Palmeri 1997), each exemplar provides an increment in a random walk. In decision-bound models (Ashby 2000), the drift rate is determined by the distance of the test stimulus from a boundary that separates the categories on the relevant dimensions. These models generalize the earlier categorization models to allow them to account for response times as well as accuracy and similarity measures.

2.7 Neural Processes

Recent data suggest that neural processes can be represented by random walk or diffusion processes (Hanes and Schall 1996). Neurons in the eye movement system in monkeys behave as though they increase their firing rates until a fixed response criterion is reached and this is the signal for the response (in the form of an eye movement) to be made. Diffusion processes are continuous processes and so are not exactly consistent with neural processes that are discrete. But if the diffusion process is replaced with a random walk with one step corresponding to 1 ms (about the same time scale as neural counts), then the approximation between the diffusion process and a discrete random walk (with time steps of 1 ms) is very good. Also, if a diffusion decision process were to be represented neurally, it would be represented by a population of neurons and this would lead to quite a good approximation between a population firing rate and a continuous process.

3. Conclusion

Random walk and diffusion models are under continuous testing. As members of the class of sequential sampling models, they appear to account for experimental data more successfully than any other class of models. Perhaps the most encouraging sign is that fits of the models to data are beginning to show invariances in parameter values across subjects and across tasks, which means that some experimental variables are represented in the model by single parameters. This kind of success is something we do not always see in modeling in psychology.

See also: Categorization and Similarity Models; Categorization and Similarity Models: Neuroscience Applications; Decision and Choice: Random Utility Models of Choice and Response Time; Decision Theory: Classical; Dynamic Decision Making; Letter and Character Recognition, Cognitive Psychology of; Luce's Choice Axiom; Sequential Decision Making; Stochastic Dynamic Models (Choice, Response, and Time)

Bibliography

Ashby F G 2000 A stochastic version of general recognition theory. *Journal of Mathematical Psychology* **44**: 310–29

Busemeyer J R, Townsend J T 1993 Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review* **100**: 432–59

Feller W 1967 *An Introduction to Probability Theory and its Applications*, 3rd edn. Wiley, New York

Hanes D P, Schall J D 1996 Neural control of voluntary movement initiation. *Science* **274**: 427–30

LaBerge D A 1962 A recruitment theory of simple behavior. *Psychometrika* **27**: 375–96

Laming D R J 1968 *Information Theory of Choice Reaction Time*. Wiley, New York

Link S W, Heath R A 1975 A sequential theory of psychological discrimination. *Psychometrika* **40**: 77–105

Luce R D 1986 *Response Times*. Oxford University Press, New York
Nosofsky R M, Palmeri T J 1997 An exemplar based random walk model of speeded classification. *Psychological Review* **104**: 266–300
Ratcliff R 1978 A theory of memory retrieval. *Psychological Review* **85**: 59–108
Ratcliff R 1980 A note on modelling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology* **21**: 178–84
Ratcliff R 1981 A theory of order relations in perceptual matching. *Psychological Review* **88**: 552–72
Ratcliff R 1988 Continuous versus discrete information processing: Modeling the accumulation of partial information. *Psychological Review* **95**: 238–55
Ratcliff R, Mckoon G 1982 Speed and accuracy in the processing of false statements about semantic information. *Journal of Experimental Psychology: Learning, Memory and Cognition*. **8**: 16–36
Ratcliff R, Rouder J N 1998 Modeling response times for two-choice decisions. *Psychological Science* **9**: 347–56
Ratcliff R, Rouder J N 2000 A diffusion model account of masking in letter identification. *Journal of Experimental Psychology: Human Perception and Performance* **26**:127–40
Ratcliff R, Smith P L 1999 Comparing stochastic models for reaction time. Paper presented at the 32nd Annual Meeting of the Society for Mathematical Psychology, Santa Cruz, CA
Ratcliff R, Van Zandt T, McKoon G 1999 Connectionist and diffusion models of reaction time. *Psychological Review* **106**: 261–300
Smith P L 1995 Psychologically principled models of visual simple reaction time. *Psychological Review* **102**: 567–91
Smith P L 2000 Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology* **44**: 408–63
Smith P L, Vickers D 1988 The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology* **32**: 135–68

R. Ratcliff

Diffusion: Anthropological Aspects

In anthropology, diffusion has been taken to be the process by which material and immaterial cultural and social forms spread in space. A number of specialized fields of inquiry in cultural diffusion developed in anthropology throughout the twentieth century. At the beginning, and going back to the late nineteenth century, stood cultural history. After World War I came acculturation and culture contact studies, which continued for several decades. The 1970s marked a growing anthropological interest in world-system studies and questions of cultural imperialism. In the late 1970s and early 1980s, transnational and globalization studies became a subfield of anthropology. This is a rough chronology without claim to comprehensiveness—in particular if one considers that so much of anthropology has been about cultural interaction and change and thus, in some sense, about the distribution of cultural and social forms across space.

The anthropological interest in diffusion began as an argument against evolutionary interpretations of history. Such interpretations had become part of popular culture in late-nineteenth-century Europe and Euro-America. The general idea was that cultural evolution and cumulative reason had taken humankind from a stage of savagery to one of civilization. The title of one of Lewis Henry Morgan's (1818–81) works put it succinctly: *Ancient Society, or Researches in the Lines of Human Progress from Savagery through Barbarism to Civilization* (1877). Evolutionary theoreticians like Morgan, Edward B. Tylor (1832–1917), and Adolf Bastian (1826–1905), held that, while cultural borrowing between social groups certainly had occurred throughout human history, similar cultural and social forms emerged in different places and times, mostly due to the psychic similarity of humans the world over, and to a general uniformity in evolutionary stages.

Friedrich Ratzel (1844–1904) was the first to explicitly reject this explanation for cultural and social similarities. He argued that no culture had evolved in isolation, and that contact and borrowing between groups should be considered to be far more common in human history than instances of social and cultural innovation. Since Ratzel believed that the principal mechanism of diffusion was migration, he arrived at the conclusion that 'migration theory is the fundamental theory of world history' (Ratzel 1882, p. 464). The object of study for Ratzel was the reconstruction of how cultural and social forms had spread from centers of innovation. This reconstructive effort went under the name of cultural history (*Kulturgeschichte* in German). Among the most notable cultural historians were: Bernhard Ankermann (1859–1943), who, unlike the majority of cultural historians who wrote universal histories of humankind, believed that insufficient data only allowed for the reconstruction of diffusion in particular regions; Leo Frobenius (1873–1938), a student of Ratzel who elaborated the latter's migration theory into a theory of cultural areas (*Kulturkreise* in German); and Fritz Graebner (1877–1934), the author of *Methode der Ethnologie* (1911), arguably the most influential theoretical and methodological work of cultural history. At the University of Vienna, Fathers Wilhelm Schmidt and Wilhelm Koppers formulated a theory of primary cultural units, which were said to be at the root of all existing cultures and to have evolved through processes of diffusion from the primeval culture of hunters and gatherers. At University College, London, Grafton Elliot Smith and his disciple William Perry tried to prove that from a single center of innovation, namely ancient Egypt, culture had diffused to the rest of the world:

> towards the close of the period of the New Empire, or perhaps even a little later, a great many of the most distinctive practices of Egyptian civilization suddenly appeared in more distant part of the coast-lines of Africa, Europe, and Asia and also in course of time in Oceania and America (Smith 1916, p. 3).

As chief agents of diffusion, Smith and Perry identified the Phoenicians.

At Cambridge, UK, worked W. H. R. Rivers, who had been a member of the Torres Strait Expedition, and who had emerged as the most influential British anthropologist in the first two decades of the twentieth century. After conducting fieldwork among the Todas of southern India in 1902, Rivers turned to 'ethnological analysis'—that is cultural history (cf. Rivers 1911). In two volumes, Rivers (1914) subsequently reconstructed the history of Melanesia by mapping the process of cultural diffusion in the area.

In the USA, Franz Boas (1858–1942) collected specimens, information in the native language about these specimens, and grammatical information necessary for the interpretation of this information. His ambition was to reconstruct a cultural history of the nonliterate people he studied. In particular, Boas developed his diffusionary analyses in the field of linguistic studies. At the same time, Boas was highly critical of cultural history's reconstruction by way of presumption. Boas thought that too much in cultural history was guesswork. Also, he formulated his own ethnographic method in contradistinction to the logic of museum exhibitions, which, at the time, placed specimens next to one another without concern for the cultural and social context they had once been part of.

Common to the work of all cultural historians was a research orientation toward topics that can be summarized in the following four concepts: (a) cultural origins; (b) material culture; (c) culture areas; and (d) culture strata.

(a) Cultural origins: cultural history shared with evolutionary theory a profound interest in ultimate origins and long-term diachronic developments. It is this preoccupation that put the question of independent innovations versus diffusion into stark relief. Evolutionary theory suggested that the origins of cultural elements should be sought in innovation because cultural similarities indicated similar mental dispositions and similar reactions to similar environmental conditions ('like causes produce like effects'). Cultural history responded that the origins of cultural elements could be found in the diffusion from centers of innovation since it was far more likely that similarity was an indication of contact than of similar dispositions and conditions.

(b) Material culture: methodologically, cultural history relied to a considerable extent on typologizing artifacts. Everything from techniques of mummification and architectural styles to types of domesticated animals, dyeing techniques, weapons, and tools were used as indicators of cultural diffusion. Specimens were taken to be cultural elements, and single artifacts, institutions, and ideas were studied like museum objects detached from their cultural and social context.

(c) Cultural areas (*Kulturkreise* in German) were conceptualized as areas of homogenous culture. Alfred Kroeber (1931, p. 646) considered culture areas to be organized in a center-periphery fashion, with centers 'in which the culture is most typical and intensive and which are to be construed as foci of radiation,' and border areas that intergrade with adjacent areas.

(d) Cultural strata (*Kulturschichten* in German) were considered to be layers of cultural diffusion, that is, superimpositions of cultural elements upon one another originating from waves of diffusion.

In the face of a methodological shift in anthropology, with direct observation during extended times of fieldwork replacing the study of museum specimens and survey studies with the help of questionnaires, the question of cultural origins waned. Consequently, the reconstructive work of cultural history, and its great reliance on artifacts, fell into disrepute in much of anthropology. Cultural interaction, however, continued to be a core issue and was now studied under the headings of culture contact, acculturation, and culture change. Acculturation, the key concept in this context, was defined as follows: 'Acculturation comprehends those phenomena which result when groups of individuals having different cultures come into continuous first-hand contact, with subsequent changes in the original cultural patterns of either or both groups' (Redfield et al. 1936, p. 149). Following World War I, culture contact had become a matter of self-conscious reflection:

> The great world-shattering changes of the past decade, the War, the Russian Revolution, the Treaty of Versailles and the consequences of these events, have led men to question for the first time many of the previously unchallenged assumptions implicit in their own civilization, whilst it has prompted them to regard with real concern and a desire to understand civilizations unlike their own (Pitt-Rivers 1927, p. xii).

Some acculturation studies, like Pitt-Rivers', were explicitly directed at the types of political domination that in so many instances produced culture change during colonialism. Thus, Richard Thurnwald (1932, p. 568f) distinguished between two kinds of acculturation through colonial relations: on the one hand, the biological and cultural mixing between men and women, on the other, formal relations of tutelage. In the one instance, acculturation was a matter of intimate relations, of primary socialization into local traditions in the form of care, which attributed to women as 'principal caretakers' the role of keepers of tradition. In the other case, the social relations were ones of formal education or secondary socialization through 'channels by which the new and foreign flows,' such as 'missions, schools, governmental ordinances, control of courts, etc.'

Soon, however, leading American anthropologists made acculturation studies turn on the issue of why some cultural groups resisted formal colonial tutelage and the 'modernization' it was to bring, and others did not. The cause was sought in the rigidity and closedness of some societies versus the adaptability and openness of others. One suggestion was that it was possible to distinguish 'hard-shelled, vertebrate' from 'soft-shelled, invertebrate' cultural systems (Social Science Research Council Summer Seminar 1954, p. 978). For instance, two Native American groups were studied with regard to how they received tribal members who returned as World War II veterans (Adair and Vogt 1949). One group was described as engaging in 'antagonistic acculturation,' that is, to resist culture change. The other group was described as 'receptive,' 'flexible,' and 'intensely curious about the outside world' (Adair and Vogt 1949, pp. 558–60)—what Margaret Mead (1938) once characterized as an 'importing culture.' These characteristics were attributed to different types of social organization, ritual systems, and collective dispositions. The attention to relations of domination that had been present in studies like those of Pitt-Rivers and Thurnwald had given way to a perspective that, at least implicitly, valued cultures that were open and adaptive to modernization.

Not all studies of culture change, however, were framed in a modernization-friendly perspective. Some classical studies—such as Eric Wolf's (1958) of syncretism (i.e., the synthesis of old and diffused cultural elements) in the case of the Virgin of Guadalupe—attempted to provide an analysis of the specifics of acculturation. Other studies were explicit arguments against modernization-friendly theories. In particular, a critical anthropology of colonialism and neocolonialism developed under the influence of Andre Gunder Frank's (1969) thesis that underdevelopment was not a result of traditionalism, but of colonialist exploitation, and of Immanuel Wallerstein's (1974) synthetic work about the dynamics of European expansion and the center-periphery structure of the world-system.

This type of anthropology tended to treat cultural diffusion as a secondary effect of economic structures and relations. The exceptions to this tendency—such as Michael Taussig's ethnography, *The Devil and Commodity Fetishism in South America* (1980)—are noteworthy. Taussig's work was, in part, an argument against the idea that European, colonial, or capitalist expansion would result in the sort of cultural imperialism that would erase all cultural differences and bring about the 'coca-colonization' of the world. Ethnographers had for some time provided examples of culturally creative responses to capitalism's local encroachments in different parts of the world (see Nash 1981 for a review of this genre). World-systemic approaches also deviated from theories of cultural imperialism's homogenizing effects in that they, in accordance with Wallerstein's interpretation, saw the rise of national ideologies in response to globalization. They did not grant independent cultural logic, however, to the global periphery. Instead, they held the periphery to be open to radical change initiated by the center. Marshall Sahlins (1985, p. viii) has called this 'a confusion between an open system and a lack of system.'

Increasingly, local responses and resistance to colonial and capitalist diffusion became a major research orientation in anthropology and ethnographic genre in its own right. At the theoretical level, work progressed toward an analysis of the channels, types of social organization, and cultural imaginations through which diffusion occurred in the late twentieth century. Much of this effort went under the name of globalization and transnational studies. One of its representatives, Ulf Hannerz (1992, p. 46), identifies four major organizational frameworks within which cultural diffusion now takes place: form of life (i.e., everyday life or life world), market, state, and movement. The frameworks differ with regard to centering and decentering tendencies, power–culture relationships, and their cultural economies. The social organization of cultural diffusion is thus complex, exhibiting great variation in scale, rhythm, encompassment, and types of responses. Another anthropologist of transnational processes, Arjun Appadurai (1991), identifies a number of 'landscapes' of modernity—ethnoscapes, ideoscapes, financescapes, etc.—which replace previous notions of well-bounded, highly localized entities such as ethnic groups, national cultures, and market centers. The landscapes are in themselves diffusionary, so that for example ethnoscapes 'around the world are no longer familiar anthropological objects, insofar as groups are no longer tightly territorialized, spatially bounded, historically unselfconscious, or culturally homogenous' (Appadurai 1991, p. 191). People travel, migrate, seek refuge and work, and live with images and ideas that also travel—and the people know that they do, would be one way of summarizing Appadurai's diffusionary argument.

> In a world of rapid communication technologies, mobile populations, a culturally saturated international economy, and transnational connections that transcend national borders and local lives, anthropologists must reckon with 'the new power of the imagination in the fabrication of social lives,' Appadurai (1991, p. 199).

The kind of diffusionary approach that Hannerz, Appadurai, and theoretically related anthropologists have developed has tended to emphasize certain types of diffusion. Rather than to focus on the diffusion of fairly isomorph cultural and social forms—like bureaucratic, educational, and scientific models—it investigates new forms of cultural diversity that are more likely to emerge in and through market, as well as other, socially more loosely organized networks (at

least where the capitalist market is not a primary source of standardization and cultural isomorphism).

Since the late nineteenth and early twentieth century, the debate in the anthropology of diffusion has shifted from the question of cultural origins to the issue of the future of culture. If cultural historians argued against the idea that the origin of similar cultural elements was to be found in independent innovations due to the psychic unity of humankind, anthropologists of globalization and transnational connections argue against the idea that current processes of diffusion will result in a homogenous world void of cultural diversity. In the 1920s, anthropologists thought they witnessed the end of cultural diversity as colonialism and capitalism increased the levels of diffusion and acculturation even in remote areas. At the beginning of the twenty-first century anthropologists, while knowing that cultural diversity did not end with the rise of colonialism, search for answers about the character and consequences of contemporary forms of diffusion. Diffusion is at the heart of current anthropological debate in so far as it asks a fundamental question: to what extent is anthropology about cultural diversity, and could anthropology survive in a world without it?

See also: Boas, Franz (1858–1942); Cultural Assimilation; Cultural Contact: Archaeological Approaches; Diffusion: Geographical Aspects; Diffusion, Sociology of; Evolution: Diffusion of Innovations; Globalization and World Culture; Globalization, Anthropology of; Innovation and Technological Change, Economics of; Innovation: Organizational; Innovation, Theory of; Technology Transfer

Bibliography

Adair J, Vogt E 1949 Navaho and Zuni veterans: A study of contrasting modes of culture change. *American Anthropologist* **51**: 547–61

Appadurai A 1991 Global ethnoscapes: Notes and queries for a transnational anthropology. In: Fox R (ed.) *Recapturing Anthropology. Working in the Present*. School of American Research Press, Santa Fe, NM

Frank A G 1969 *Latin America: Underdevelopment or Revolution*. Monthly Review Press, New York

Graebner F 1911 *Methode der Ethnologie*. Carl Winter's Universitätsbuchhandlung, Heidelberg, Germany

Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York

Kroeber A L 1931 Culture area. In: Seligman E R A (ed.) *Encyclopaedia of the Social Sciences*. Macmillan, New York, Vol. 4, pp. 646–7

Mead M 1938 The Mountain Arapesh. Part 1: An importing culture. *American Museum of Natural History Anthropological Papers* **36**: 139–349

Morgan L H 1877 *Ancient Society, or Researches in the Lines of Human Progress from Savagery through Barbarism to Civilization*. Barrati Library, Calcutta, India

Nash J 1981 Ethnographic aspects of the world capitalist system. *Annual Review of Anthropology* **10**: 393–423

Pitt-Rivers G H L-F 1927 *The Clash of Culture and the Contact of Races: An Anthropological and Psychological Study of the Laws of Racial Adaptability, with Special Reference to the Depopulation of the Pacific and the Government of Subject Races*. Routledge, London

Ratzel F 1882 *Anthropo-Geographie oder Grundzüge der Anwendung der Erdkunde auf die Geschichte*. Engelhorn, Stuttgart, Germany

Redfield R, Linton R, Herskovits M 1936 Memorandum for the study of acculturation. *American Anthropologist* **38**: 149–52

Rivers W H R 1911 The ethnological analysis of culture. In: Smith G E (ed.) *Psychology and Ethnology*. Trubner, London, pp. 120–40

Rivers W H R 1914 *The History of Melanesian Society*, 2 Vols. Cambridge University Press, Cambridge, UK

Sahlins M 1985 *Islands of History*. University of Chicago Press, Chicago

Smith G E 1916 The influence of ancient Egyptian civilization in the East and in America. *The Bulletin of the John Rylands Library*, January–March, Manchester, UK

Social Science Research Council Summer Seminar 1954 Acculturation: An exploratory formulation. *American Anthropologist* **56**: 973–1000

Taussig M T 1980 *The Devil and Commodity Fetishism in South America*. University of North Carolina Press, Chapel Hill, NC

Thurnwald R 1932 The psychology of acculturation. *American Anthropologist* **34**: 557–69

Wallerstein I 1974 *The Modern World-System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

Wolf E 1958 The virgin of Guadalupe: A Mexican national symbol. *Journal of American Folklore* **71**: 34–9

R. Stade

Diffusion: Geographical Aspects

1. Introduction

From a geographic perspective, diffusion involves the propagation of a phenomenon in a manner such that it spreads from place to place, leading to maps that differ from one time to another. Most commonly the phenomenon is an innovation—a new product, new idea, new technology, new organizational structure, new way of doing things—where 'new' means new to a particular place by way of adoption. Hence, diffusion phenomena cover a wide range that includes transportation modes such as the automobile, farming techniques, family planning, credit cards, broadcast and cable television, shopping centers, production practices, such as the assembly line or just-in-time inventorying, political movements, cultural practices, frontier development, modernization in Third World settings, epidemics, urban ghettoes, and urban areas themselves. The enormous range of phenomena studied under the diffusion rubric is further indicated by

Brown (1981), Gould (1969), Morrill et al. (1988), and Rogers (1995).

While geographers have studied diffusion through most of the twentieth century, the focus has shifted over time, often with marked departures from prevailing norms. Consequently, diffusion studies by cultural geographers earlier in the century may be hardly recognizable as similar in genre to those at the end of the century, focusing on entrepreneurial agents in diffusing products or technologies.

2. *The Cultural Geography Tradition*

Diffusion has been a subject of geographic enquiry for much of the twentieth century. The spread of a phenomenon, idea, or technique through a population or region incorporates basic geographic elements of distance, direction, and spatial variation. Within geography, diffusion studies originated in cultural geography, whose research goal is:

> … to identify environmental features characteristic of a given culture and…discover what role human action…has played in creating and maintaining given geographic features (Wagner and Mikesell 1962, p. 1)

The cultural geography tradition dominated pre-World War II scholarship and focused on change in indigenous cultures, often centuries ago. Culture and the human or cultural landscape are the main concern, and diffusion studies are relevant in clarifying questions concerning culture origins, the spatial configuration of culture areas, and culture-related or culture-influenced features on the earth's surface. Carl Sauer was the central figure in this approach, which tends to be closely aligned with and complementary to some concerns of anthropology (Mikesell 1967, 1978).

Within this framework, many studies aim to assess the relative importance of diffusion vs. simultaneous invention in development of cultural landscapes. They employ an empirical, inductive approach with emphasis on landscape development in prehistory and/or over considerable periods of time. Often investigation considers a single phenomenon, seen as a culture trait, and explains its observed spatial distribution as the outcome of processes unfolding over space and time. A typical procedure is to examine a phenomenon's spatial distribution at different times or its frequency of occurrence in different locations at a single time. Inferences are drawn regarding origins (or cultural hearths), means, and routes of dispersal of both the observed phenomenon (or culture trait) and ideas–techniques–culture for which it is a surrogate. Attention also is given to cultural and environmental conditions necessary for adoption of particular phenomena (Bobek 1962) and to macroscale, big picture phenomena, such as the domestication of plants and animals, their origins, routes of dispersal, and destinations (Sauer 1952).

In stressing the role of diffusion in development of the human landscape, many cultural geography studies proceed further to consider indirect effects of the culture trait innovation on physical and cultural environments. A basic notion is that the diffusing item is both a stimulus to new innovation and itself subject to modification as it spreads. The relation between diffusion, the item being diffused, and the human landscape is therefore complex and subject to continual change.

3. *The Hagerstrand Tradition*

Sauer's counterpart in later diffusion research is Torsten Hagerstrand who, in the early 1950s, initiated a shift towards the generative processes underlying diffusion and modern-day products and practices. Since this focus is consistent with Sauer's conceptualization, why was process less prominent in earlier research? One reason may be that identification of processes controlling locational change is less feasible in the sparsely documented situations studied by cultural geographers. Second, Hagerstrand's main impact was on researchers in economic and urban geography who were primarily concerned with location and locational processes rather than landscape. Illustrative is Haggett's (1965, pp. 10–13) classification of research under the Sauer school as landscapist or ecologist and research under the Hagerstrand school as locationalist.

Hagerstrand's research (1952, 1967) focuses on diffusion of contemporary innovations such as the automobile and bovine tuberculosis control. His first study, *The Propagation of Innovation Waves* (1952), continues the cultural geography tradition by its use of an inductive approach and concern with diffusion pattern rather than process. At a somewhat later date (1953), however, Hagerstrand observed that

> … spatial order in the adoption of innovations is … so striking that it is tempting...to create theoretical models which simulate the process and make predictions achievable (Berry and Marble 1968, p. 370)

This statement indicates a change from description and inductive generalization to a more deductive approach focusing on generative processes. The work marking this shift is Hagerstrand's Ph.D. dissertation of 1953, subsequently translated into English as *Innovation Diffusion as a Spatial Process* (1967). Hagerstrand made three contributions that continue to play a role in diffusion research.

3.1 *Empirical Regularities in Diffusion*

Three regularities characterize diffusion. Over time, a graph of the cumulative level of adoption is expected to approximate an S-shape that rises with the passage of time. In an urban system, diffusion is expected to

proceed from larger to smaller centers, a regularity termed the hierarchy effect. Within the hinterland of an urban center, diffusion is expected to proceed in a wave-like fashion outward from the urban center, first hitting nearby rather than farther-away locations, and a similar pattern is expected in diffusion among rural populations. This regularity is termed the neighborhood or contagion effect.

Hagerstrand (1952) also provides a framework that connects these regularities. For diffusion in an urban system, this posits three phases: a primary stage, during which initial diffusion centers are established; a diffusion stage, during which neighborhood-type diffusion occurs in areas near diffusion centers and secondary diffusion centers are established in lower-order urban places; and a condensing or saturation stage, during which filling-in occurs and diffusion eventually ceases.

This framework thus illustrates the complementarity of hierarchy and neighborhood effects and their temporal positioning within diffusion. Also, if locales of adopters were tabulated, higher-order urban places characterize the left tail of the S-curve of adoption; lower-order urban places characterize the middle, sharply rising portion of the S-curve, along with rural hinterlands in proximity to urban centers; and remote rural hinterlands characterize the S-curve's right tail.

3.2 Hagerstrand's Conceptualization of the Diffusion Process

Hagerstrand's conceptualization of innovation diffusion across the landscape holds that adoption is primarily the outcome of a learning or communications process. Factors related to the effective flow of information are most critical and, therefore, a fundamental step in examining diffusion is identifying spatial characteristics of information flows and resistance to adoption.

In Hagerstrand's conceptualization, information originates from the mass media or past adopters. The destination of personal messages depends on a sender's network of social communications (interpersonal contacts), and configuration of this network reflects various social and terrestrial barriers that impede, divert and/or channel communications. Particular attention is given to terrestrial barriers such as lakes, forests, difficult terrain, and geographical distance separating potential communicants.

The conceptualization further recognizes that resistance levels differ among individuals, and that higher levels of resistance require more information for adoption to occur. This leads to a dichotomy of resistance as either the result of values which are inconsistent with adoption (social resistance) or practicalities which make adoption difficult or impossible (economic resistance).

To address diffusion at different geographical scales, a hierarchy of networks of social communication is posited. One network might operate locally and another regionally. The first would control diffusion among farmers in a single locale, whereas diffusion at the central-place level of aggregation might utilize a regional network of persons in different urban areas that communicate with one another. Superimposition of these two levels provides a comprehensive picture of diffusion in the area represented.

In summary, Hagerstrand's conceptualization proposes transformation of a population from a low proportion of adopters to one with a high proportion, by means of information dissemination through media and interpersonal contact. Spatially, the principal mechanisms of transformation are networks of social communication characterized by biases and distortions that, in turn, effect and are reflected in spatial patterns of innovation diffusion.

3.3 Hagerstrand's Methodological Framework

To implement his conceptualization, Hagerstrand used Monte Carlo simulation. Adoption takes place after a specified number of messages are received, the specified number varying according to an individual's resistance to adoption, which is randomly assigned. Time is treated in discrete units of equal interval, and each past adopter sends a message in every time interval. The destination of messages depends on the probability of contact between tellers and potential receivers. This probability is a function of distance between them unless a water or forest barrier intervenes. A distance decay function is employed to compute the probability of interaction between each of 25 cells in a 5-by-5 cell grid and the center cell of that grid. The resulting grid of probabilities, termed the mean information field (MIF), is employed with a map gridded in a similar manner, but with many more than 25 cells. The MIF is centered over a map cell containing an adopter of the innovation at time t; a random number is then drawn to determine which of the 25 surrounding map cells receives a message to adopt in time $t+l$. If the message passes over a water or forest barrier, it is either nullified or a new random number is drawn, depending on the level of obstruction. This procedure is repeated for every adopter in every time interval. If a previous adopter receives a message, the message is lost, that is, it has no effect and is not repeated. If a message is received by a non-adopter who, owing to resistance level, requires more than one message before adoption, the message is stored until sufficient messages for adoption are accumulated.

This model mechanism is crude by today's standards of computational capability, but in the 1950s it was highly innovative. Accordingly, Hagerstrand's diffusion model played an enormous role in Geography's quantitative revolution of the 1960s and 1970s, as both an example and standard of excellence (Berry and Marble 1968). This observation also applies to social

science overall; the model is included, for example, in books such as Coleman's (1964) *Introduction to Mathematical Sociology*.

4. *The Market and Infrastructure Perspective*

Hagerstrand's formulation, and those of the many other social scientists studying diffusion from the 1950s onward, took an atomistic view of society as individuals with free will, any of whom could adopt an innovation. Thus, the onus of adoption was on individuals, who were classified on a scale ranging from innovator to laggard, with the social meaning each term implies (Rogers 1995). This is termed the adoption perspective, focusing primarily on the demand side of innovation diffusion.

The market and infrastructure perspective focuses on supply or availability of the innovation (Brown 1981). Rather than assuming all have an equal opportunity to adopt, this is seen as purposely and egregiously unequal, so that adoption represents a choice within a constraint set; government, non-profit, and entrepreneurial institutions establish these constraints; and attention is given to the process by which innovations and conditions for adoption are made available to individuals, households or firms. Hence, focus shifts from the individual to the structure within which individuals operate, paralleling realignments in social science overall (Brown 1999). Nevertheless, the market and infrastructure perspective is put forth as complementary, not contradictory, to the adoption perspective. Broad outlines of diffusion are shaped by where (and to whom) innovations are (made) available, but within those outlines, more detailed variation in diffusion patterns reflects adoption behavior.

Central elements of this perspective include the propagator of the innovation, the entity behind its diffusion; diffusion agencies (outlets) through which innovation is distributed to the population at large; and diffusion strategy, which is designed to encourage adoption and includes decisions on agency location, pricing, advertising, population segments targeted by promotional efforts, etc. Hence, locations and temporal sequencing of diffusion agencies determine where and when an innovation will be available and provide the general outline of the diffusion pattern. Diffusion strategy further defines the pattern by creating different levels of access depending on a potential adopter's economic, locational, demographic, and social characteristics. The final pattern occurs via adoption behavior, the last stage of the process. Interacting with these factors are infrastructures like transpotation and utility networks that are independent of the innovation's propagator and diffusion agencies, but distinctly affect adoption.

Hence, the market and infrastructure perspective emphasizes the marketing aspects of diffusion, especially strategies for market penetration. Whether the propagator is profit or non-profit motivated, broad acceptance is a common goal, and a strategy will be designed to accomplish this, that is, to best penetrate the market. This also is true whether propagation involves centralized decision making, decentralized decision making with a coordinating propagator, or decentralized decision making without a coordinating propagator, although such distinctions affect procedures by which diffusion occurs. Similarly, diffusion agencies may be represented by salespeople or agricultural extension agents, but either should be guided by marketing strategy and, accordingly, make selective decisions, regarding which potential adopters will be contacted or cultivated, and to what degree. Considering the role of infrastructure, broadcast television stations were needed before television sets diffused. Today, cable and satellite services vary enormously from place to place, affecting innovation availability. Enhanced accessibility through communication or transportation improvements increases the market areas of diffusion agencies, and renders innovations more attainable to potential adopters. Also, for example, irrigation infrastructure is often critical to the success of high-yielding variety seeds in agriculture.

Finally, this framework calls into question hierarchy and neighborhood effect regularities, because relevant factors from a marketing perspective are often not arranged in that manner. The diffusion of cable television in Ohio, for example, occurred initially in small, more remote places where broadcast signals were weak or nonexistent, because cable was a means of enhancing television sales by appliance stores.

5. *Technological Diffusion*

Agricultural technology has been a major focus of diffusion studies. Beginning in the 1930s, and accelerating thereafter, rural sociologists and agricultural economists, working through agricultural experiment stations and extension agents, carried out an enormous number of studies focused on farm innovations, such as hybrid corn, fertilizers, silage methods, high-yielding varieties, and the like (Rogers 1995). Until recent industrial agriculture, the adopting unit was the farm household, thus joining this effort with diffusion research concerned with consumer innovations. Manufacturing firms that are well beyond household size, also adopt technological innovations, as do service and retail firms. This has been studied extensively, but rarely joined with the body of diffusion research described previously.

Rather than emphasizing communications, studies of firm adoption give more attention to characteristics of the firm and innovation itself. Also significant is the fact that innovation replaces an older technology with a newer one, highlighting comparative costs, benefits, and risks; factors such as relative profitability, in-

vestment, sunk costs, and the like. In this regard, there is greater emphasis on innovation as a continuous process; one that might be tailor-made for a particular firm or firm setting, rather than as a singular, unchanging phenomenon; and one that involves a more intimate and ongoing relationship between the innovation's supplier and user. Issues such as competition among firms and relative factor prices, favoring one technology over another, also play a role.

As these factors indicate, economic perspectives are important to investigating the diffusion of technological innovations among firms (Brown 1981). But spatial considerations enter in two ways. First, there has been a tendency to assume that larger firms in larger urban agglomerations are earlier adopters, giving rise to a self-reinforcing (cumulative and circular causation; Myrdal 1957) hierarchy-effect pattern of diffusion. In fact, earlier adopters are just as likely to be less-than-the-largest in size, which may correlate with aggressiveness and innovativeness of management. There also may be a spatial order to innovative change, such that newer economic regions are favored, leading to regional rotation or spatial succession; Silicon Valley provides an example (Malecki 1997). The second way spatial considerations come into play is that the study of technological innovation has become closely intertwined with regional development and regional economic change (Malecki 1997). Hence, the questions may not be inherently geographic in terms of process, but very geographic in outcome.

6. *Development as a Diffusion Process*

Seen as integral to economic development, diffusion of innovations in Third World settings has received much attention. Examples include improving agricultural production and living conditions; generating entrepreneurial activity and related employment; promoting family planning, and improving infrastructure. Research and policy have largely employed the adoption perspective, but examples of a market and infrastructure approach are also evident (Brown 1981, Rogers 1995).

More broadly, development itself has been treated as a diffusion process. One approach involves choosing an index of development, and tracing its spread across the landscape; this approach is similar methodologically to the culture trait studies of earlier diffusion research. Such indices include agricultural cooperatives, banking, communications (telegraph, telephone), educational facilities, electricity, employment patterns across economic sectors, medical services, police organizations, postal agencies, transportation modes (roads, rail), urbanization, and other artifacts of a market economy (Gould 1976, Pedersen 1975, Riddell 1976, Taaffe et al. 1963).

Frontier expansion also has been viewed in a diffusion framework (Hudson 1969). Substantively, these studies focus on aspects such as capital generation and control; migration, travel, and trade interactions; the role of particular economic sectors such as banking and wholesaling; and transportation linkages (Meyer 1980, Pred 1980)

Finally, formal models of development have a distinct diffusion dimension. Modernization, dual economy, core-periphery, growth center, and stages of growth frameworks put forth the notion that Third World settings are composed of traditional and modern elements, and that development consists of transforming traditional to modern. Dependency and related theories take a similar tack, albeit one that differs in emphasis. These models divide Third World landscapes into dynamic, propulsive, modern cores and traditional peripheries; development is represented by transformation and expansion of the modern; and this occurs through mechanisms such as information and capital flows, migration, human capital developments, and more generally, backwash, trickle-down, and imitation effects (Brown 1991).

7. *Overall Observations*

Diffusion comprises a long tradition of research in geography. From epoch to epoch in social science, it has changed focus and content. Geographical studies of diffusion also represent only one approach in that the topic is highly multi-disciplinary including, at least, contributions from anthropology, communications, economics, marketing, political science, psychology, and sociology. Accordingly, concepts used by geographic studies of diffusion may be found in research by other disciplines; for example, interpersonal interactions, mass media effects, utility-maximization, and marketing strategies. In all these instances, as well as in purely geographic studies, the critical contributions of a spatial perspective to understanding the dynamics of diffusion are apparent.

See also: Cartography; Culture in Nonhuman Organisms; Diffusion: Anthropological Aspects; Diffusion, Sociology of; Place in Geography; Scale in Geography; Spatial Data

Bibliography

Berry B J L, Marble D F (eds.) 1968 *Spatial Analysis: A Reader in Statistical Geography*. Prentice Hall, Englewood Cliffs, NJ

Bobek H 1962 The main stages in socio-economic evolution from a geographical point of view. In: Wagner R L, Mikesell M W (eds.) *Readings in Cultural Geography*. University of Chicago Press, Chicago, pp. 218–47

Brown L A 1981 *Innovation Diffusion: A New Perspective*. Methuen, London

Brown L A 1991 *Place, Migration, and Development in The Third World: An Alternative View*. Routledge, London

Brown L A 1999 Change, continuity, and the pursuit of geographic understanding: Presidential address. *Annals of the Association of American Geographers* **89**: 1–25

Coleman J S 1964 *Introduction to Mathematical Sociology*. Free Press, New York
Gould P R 1969 *Spatial Diffusion*. Resource Papers Series, Association of American Geographers, Washington, DC
Gould P R 1976 Tanzania 1920–63: The spatial impress of the modernization process. In: Knight C J, Newman J L (eds.) *Contemporary Africa: Geography and Change*. Prentice-Hall, Englewood Cliffs, NJ, pp. 423–37
Hagerstrand T 1952 *The Propagation of Innovation Waves*. Lund Studies in Geography, Gleerup, Lund, Sweden
Hagerstrand T 1967 *Innovation Diffusion as a Spatial Process*. University of Chicago Press, Chicago (trans. of *Innovations for loppet ur Korologisk Synpunkt*. Gleerup, Lund, Sweden, 1953)
Haggett P 1965 *Locational Analysis in Human Geography*. Edward Arnold, London
Hudson J C 1969 A location theory for rural settlement. *Annals of the Association of American Geographers* **59**: 365–81
Malecki E J 1997 *Technology and Economic Development: The Dynamics of Local, Regional, and National Competitiveness*. Longman, Essex, UK
Meyer D R 1980 A dynamic model of the integration of frontier urban places into the United Sates system of cities. *Economic Geography* **56**: 120–40
Mikesell M W 1967 Geographic perspectives in anthropology. *Annals of the Association of American Geographers* **57**: 617–34
Mikesell M W 1978 Tradition and innovation in cultural geography. *Annals of the Association of American Geographers* **68**: 1–16
Morrill R, Gaile G L, Thrall G I 1988 *Spatial Diffusion*. Scientific Geography Series, Sage, Newbury Park, CA
Myrdal G 1957 *Economic Theory and Underdeveloped Regions*. Duckworth, London
Pedersen P O 1975 *Urban-regional Development in South America: A Process of Diffusion and Integration*. Mouton, The Hague, The Netherlands
Pred A 1980 *Urban Growth and the Circulation of Information: The United States System of Cities, 1790–1840*. Harvard University Press, Cambridge, MA
Riddell J B 1976 Modernization in Sierra Leone. In: Knight C J, Newman J L (eds.) *Contemporary Africa: Geography and Change*. Prentice-Hall, Englewood Cliffs, NJ, pp. 393–407
Rogers E M 1995 *Diffusion of Innovations*, 4th edn. Free Press, New York
Sauer C O 1952 *Agricultural Origins and Dispersals*. American Geographical Society, New York
Taaffe E J, Morrill R L, Gould P R 1963 Transportation expansion in underdeveloped countries: A comparative analysis. *Geographical Review* **53**: 503–29
Wagner P L, Mikesell M W (eds.) 1962 *Readings in Cultural Geography*. University of Chicago Press, Chicago

L. A. Brown

Diffusion, Sociology of

The diffusion of social practices, beliefs, technologies or moral rules is a subject of interest for both sociologists and anthropologists. In both cases, diffusion process analysis points out how individuals, groups, or communities may incorporate, reject, or adapt practices, rules, or social representations designed by others. For both sociologists and anthropologists, analyzing such processes opens a debate about the construction of existing social situations.

1. *Tarde's Contribution and the Issue of Diffusion*

According to (Tarde [1890] 1979), imitation comprises the integration of initially isolated and individual inventions into collective and social practices. When such inventions disseminate, mix, or interfere with one another, they form the basis of society:

> Renewing initiatives stand as a starting point. They bring new needs and new satisfactions to the world, then spread or tend to do so through forced or spontaneous, chosen or unconscious, quick or slow imitation, always responding to a regular pace, as a light wave or a family of termites (Tarde 1979, p. 3).

The diffusion of inventions assumes that social structures, at least at the time of their appearance, result from actions that are not determined by a certain state of the world: 'Every invention arises as one achieved possibility among thousands of others, as one between other different possibilities' (p. 49). The three processes mentioned by Tarde—imitation, diffusion, and the creation of social situations—are to be found in future works related to the issue of diffusion, even though these processes are progressively rethought.

Thus, it is assumed that new social practices and new cultural features ('novelty') are integrated and transformed at the same time: one should therefore not only speak of 'imitation' (Boas 1949). The impacts resulting from the use of novelty also interfere with the other variables of the social system, knowing that the importance of such impacts has little in common with the restricted integration of a new element: when the Tanalas of Madagascar began to grow rice and, therefore, to build terraces, the entire organization of property and laws as well as the social organization changed together (Linton 1936). In all cases, what is adopted is not a single practice or a specific cultural feature, but the principle which founds them: in the Canadian northern territories, the inhabitants learnt how to harness reindeers while observing how to harness horses.

Tarde's works also tackle the issue of the relationship between individual choices and collective conducts, that is, the progressive construction of norms. The analysis of fashion phenomena belongs perfectly to that sort of approach (Hurlock 1929): fashion is effectively the result of an individual choice, but it also matches a norm in this domain. And the aggregation of individual choices shapes the norm. A study of first-name choices in France during the twentieth century explicitly examines phenomena of that type (Besnard

1979). It also underlines the cyclic dimension of fashions: the choice of new first names is decided by superior social classes, as the renewal of such choices enables them to be distinguished from other social classes, who then imitate them. Fashion therefore fits the process of novelty diffusion and reflects the tension between the will for imitation expressed by some and the search for distinction expressed by others (Simmel [1904] 1956). In addition to this 'vertical diffusion,' there is also a horizontal process: people who work in the media or in show business and, in a wider sense, all those who hold a key position in relational networks, disseminate fashions. The social groups who remain close to fashion are the middle classes and the upwardly mobile.

However, the most important issue raised by Tarde lies in the transformation of social structures through the constant integration of novelty. This perspective is addressed in two different ways: anthropologists observe the construction of society according to a diffusionist paradigm, whereas sociologists focus on social forms and forces that may enable or limit novelty diffusion.

2. *Diffusionist Anthropology*

While contesting the evolutionist idea according to which every society faces different development stages in the field of social, economic, technological, and institutional practices, diffusionist anthropology examines the processes and spaces required for the propagation of these dimensions, as means to design societies. It pays special attention to the migration of practices rather than to their local elaboration. But it often forgets that, in the field of social relationships, a diffusion flow from A to B also involves a reverse movement from B to A, so that practice or belief diffusion also includes interaction diffusion (Sorokin 1964).

Graebner (1911) underlines the existence of 'areas of social practices,' which share similar origins (i.e., which originate from a common central point), and which find expression in comparable forms of collective life, beliefs, or economic practices. It may for instance concern the 'totemic complex' in Australia: this complex can be characterized by housing, arms, funeral rituals, and astral mythology. But local populations seldom incorporate the features of social practices or beliefs entirely. They especially proceed to some selection among the elements of the system, as well as to some alteration or adaptation to local practices. Every cultural circle that is worked out this way regularly increases, until it meets the development of another cultural circle: their encounter leads to the fusion of both areas.

The analysis of the spatial sharing out of cultures therefore comprises the trajectory of their diffusion and, in that sense, their own history. According to Boas, the cultural areas may be continuous or discontinuous. In the latter case, all the features that typically characterize a social practice or a belief (e.g., the themes and sequences of a myth) can be found in the center of the area, whereas the other features remain at the periphery. During their trajectories, practices and beliefs may be enriched or, on the contrary, may lose some of their dimensions, due to encounters with elements coming from other spaces. There are strong similarities between the Eastern Siberian populations' culture and that of British Columbian Indians, whereas Eskimo populations, while not so recent, are less culturally close, even though they live in an intermediate territory.

These perspectives are of double interest. They renew the methodology of anthropology as they contribute to replacing general reflection with empirical enquiries. For instance, Lowie (1934) explains that all known varieties of bananas intended for sale come from new shoots, rather than seeds; it is therefore possible to follow human migrations on the basis of their consumption. Within the northern American tradition, these perspectives result from a reasonable epistemological position, which aggregates internal and external causes of change and gives a key role to the specificity of living conditions and to the nature of the contacts that characterize geographical areas. Boas therefore gives up defining general rules, and turns to questioning the diversity of social and cultural practices within a limited region. However, some of the positions expressed by advocates of diffusionist anthropology seem to result more from their conflict with those who defend natural evolution than from empirically founded research. Thus, one may explain that the majority of civilizations originate from ancient Egyptian culture, which was disseminated by sea. For example, the pyramids erected by Native Americans are considered to be reproductions of those designed by the Cambodians or the Japanese, who would also have taken the Egyptian ones as their model.

3. *The Statistical Form of Diffusion*

When linking the age of cultural features or observed social practices with the number of people that may use them, anthropologists often compare the diffusion process to a logistic, epidemiological 'S' curve (e.g., Pemberton 1936). In so doing, they encounter the works of numerous sociologists (e.g., Mendras and Forsé 1983). At the beginning of the process, a limited number of people are affected by the novelty: the speed of diffusion is therefore rather low; then, the speed regularly increases, and the growing number of informed people leads others to adopt the novelty; finally, the curve decreases, and the number of noninformed people regularly diminishes. This form of diffusion can also be considered as a process of inverted norms: in the first stage, a few pioneers only behave differently from usual; then their behaviors progressively shape new norms and impose themselves

on the 'followers'; during the last phase, 'resistants' may slowly and partially accept the integration of new norms. In both perspectives, the 'S' curve seems to reflect the diffusion process adequately.

However, this kind of curve seldom reflects the real development of the phenomenon, especially when looking at its complexity. Its existence requires specific conditions with regard to the homogeneity of the populations and to the regularity of the process—conditions that real social situations seldom show. Regarding rumors, Boudon and Bourricaud (1982) identify three statistical forms of diffusion. When a rumor appears to be disseminated by word of mouth within a vast and homogeneous population, the increase in the number of people informed at every stage is roughly proportional to the number of people already informed; the diffusion curve is somewhat exponential. Whenever the population is simultaneously homogeneous and limited, the process shapes a logistic 'S' form, which characterizes the epidemiological approach. And if the information is only disseminated through the media (in other words, if the members of a population have few interpersonal relationships), the pace of the increase in the number of informed people is proportional to the number of those not yet informed. Thus, the epidemiological curve is only a rare possibility among other diffusion processes for a novelty, a belief, or a rumor. Referring to the 'S' curve, Sorokin ([1937] 1964), while reading his contemporaries' works, explains that such a phenomenon occurs if and only if the novelty remains of the same kind, and if the civilization, the channels of communication, and the means of support (here including physical strength) it is based on remain similar.

The interest in this kind of analysis resides in the identification of factors, phases, or actors who take part in the achievement of the diffusion process designed to integrate a novelty. Within this perspective, Katz and Lazarsfeld (1955) underline the determining role of influencing networks. As for decision criteria and mechanisms related to Midwest women's practices in food shopping, clothes, political issues, or the cinema, they observe that choices are inspired by two sequences of influence. The first one refers to the media, who transmit information to everyone; the second is undertaken by influential people, who use the information to implement new social practices (in this case, women coming from the same social class). When observing the prescription of a new medicine, Coleman et al. (1966) perfectly underline the heterogeneous dimension of a population and, in that sense, the first complex mechanisms of innovation adoption. The first doctors to prescribe a medicine ('the innovators') 'test' it after it has been marketed, and therefore tend to slow down the diffusion process; then, when they are convinced of its advantages, they speed up the diffusion process by actively influencing the diffusion to other doctors who may have hesitated before. These innovating and influencing doctors are those who have kept a close relationship with the world of hospitals and research; they globally benefit from wider and denser relational networks than others. They play the same role of 'intermediaries' as the minority of farmers who have 'specialized' in links with other regions of the same country, whereas the majority limits its relationships to neighborhood contacts (Hägerstrand 1965).

The diffusion of social practices, cultural features or novelties thus follows a rather unpredictable trajectory: It depends on the kind of interactions that occur. It also depends more on the recognition gained by a novelty when being used than on its own characteristics. With this in mind, the sociological analysis of diffusion is based, at least partially, on the diffusion of rumors; the latter finally create the effective use of novelty. In that sense, the 'self-fulfilling' prophecy (Merton 1949) remains at the top of diffusion issues.

4. *The Conditions of Diffusion*

The novelty diffusion processes may not always be perfectly comparable; however, the conditions required for their existence always relate to the initial action of 'pioneers,' 'innovators,' 'intermediaries,' or 'cosmopolitans' (Merton 1949). These types of actors are the first ones who effectively use it. They benefit from their own capability to behave regardless of the norms, as opposed to the rest of the population, because they benefit from an additional background of cultural references: The 'stranger' character often harmonizes with the diffusion analysis. But the ability to reverse norms, the type of ability which allows a pioneer's behavior to be seen as a normal one, assumes that cultural differences refer to a group's substantial weight (Moscovici 1976); to put it more simply, such an ability requires that the novelty becomes a matter of interest for the rest of the population, the reasons for such an interest being cultural, economic, or strategic.

Such conditions are obviously not always immediately fulfilled. Therefore, the researcher's interest may turn to the analysis of obstacles and resistance faced by the diffusion process. Within such a perspective, Bloch (1935) examines the reasons that may explain the relatively slow diffusion of water mills (a one millennium process), knowing that such mills show marked advantages in productivity increase and labor reduction. The water mill cannot be set up everywhere, as it has to be supplied with numerous and regular streams; in the Middle Ages, the risk of a city being besieged led the lord to keep a significant number of traditional millstones, as these mills operate regardless of the natural environment, which could be controlled by the enemies. Building a water mill also requires that the builders legally possess the supplying stream, although one stream may run through different rural

properties. It also assumes that builders are able to finance the effective erection of the building, whereas their available financial capital is actually low. It finally requires that they make profit on the investments, by dedicating them to a wider community use than the one offered by traditional techniques; but social structures are based on family farming. Nonetheless, such objective dimensions are not the only reasons that may explain the slow diffusion of water mills. From the lords' point of view, being obliged to keep an important workforce available in order to use the millstones also represents an indicator of power as well as an opportunity to keep everybody in their own social position within the existing division of labor. There are symbolic and political dimensions involved in such a process. Despite fines and police force activities to make the use of common mills compulsory, despite the payment of high taxes when millstones are in use, despite courts' decisions in favor of the owners of water mills, despite all these efforts, the use of mills remains an issue.

The factors interfering in novelty diffusion are diverse: legal, symbolic, strategic, economic, and cultural. Their number and their interdependency do not allow for any forecast about the end of the novelty. Within novelty diffusion, the competition between different rationalities may interfere in a positive or in a negative way (Alter 2000). In addition, many novelties are implemented in favor of a restricted minority of the population; the diffusion may be very quick, and it will only concern a fraction of the population (Edmonson 1961). In other words, the price to be paid to adopt innovation may slow down its diffusion, whereas the expected income would speed it up (Griliches 1957). The diffusion analysis therefore comprises a nondiffusion one. It could not be purely diffusion oriented.

Finally, the lessons to be drawn from these different perspectives are threefold. The diffusion of an object, a belief, or a political system does not depend on its own intrinsic qualities but more on the involved social actors' rationalities. Novelty diffusion leads to the transformation of social norms; this transformation is based on the existence of social actors who break the established social order. A novelty does not disseminate without radically changing; it could therefore always be described as an innovation, at least partially.

See also: Collective Beliefs: Sociological Explanation; Fashion, Sociology of; Influence: Social; Innovation and Technological Change, Economics of; Innovation: Organizational; Innovation, Theory of

Bibliography

Alter N 2000 *L'innovation ordinaire*. Presses Universitaires de France, Paris

Besnard P 1979 Pour une étude empirique du phénomène de mode dans la consommation des biens symboliques: Le cas des prénoms. *Archives Européennes de Sociologie* **XX**(2): 343–61

Bloch M 1935 Réflexions sur l'histoire des techniques. *Annales d'Histoire Economique et Sociale* **36**: 538–63

Boas F 1949 *Race, Language and Culture*. Macmillan, New York

Boudon R, Bourricaud F 1982 *Dictionnaire critique de la sociologie*. Presses Universitaires de France, Paris

Coleman J S, Katz E, Menzel H 1966 *Medical Innovation. A Diffusion Study*. Bobbs-Merril, New York

Edmonson M R 1961 Neolithic diffusion rates. *Current Anthropology* **2**: 71–102

Graebner F 1911 *Methode der Ethnologie*. Carl Winter, Heidelberg, Germany

Griliches Z 1957 Hybrid corn: An exploration in the economics of technological change. *Econometrica* **25**: 501–22

Hägerstrand T 1965 Quantitative techniques for analysis of the spread of information and technology. In: Anderson C A, Bowman M J (eds.) *Education and Economic Development*. Aldine, Chicago, pp. 244–80

Hurlock E B 1929 Motivation in fashion. *Archives of Psychology* **XVII**: 432–48

Katz E, Lazarsfeld P F 1955 *Personal Influence. The Part Played by the People in the Flow of Mass Communication*. The Free Press, Glencoe, IL

Linton R 1936 *The Study of Man: An Introduction*. Appleton, New York

Lowie R H 1934 *An Introduction to Cultural Anthropology*. Farrar and Rinehart, New York

Mendras H, Forsé M 1983 *Le changement social*. Armand Colin, Paris

Merton R K 1949 *Social Theory and Social Structure*. The Free Press, Glencoe, IL

Moscovici S 1976 *Social Influence and Social Change*. Academic Press, London

Pemberton H E 1936 The curve of culture diffusion rate. *American Sociological Review* **1**: 547–56

Simmel G 1956 Fashion. *American Journal of Sociology* **LXII**(6): 541–58

Sorokin A S 1964 *Comment la civilisation se transforme*. Librairie Marcel Rivière, Paris

Tarde G 1979 *Les lois de l'imitation. Etude sociologique*. Slatkine Reprints, Paris

N. Alter

Digital Computer: Impact on the Social Sciences

The invention of the computer has been described as the third information revolution—after the advent of writing and the printing press. Computers transformed our world, in reality and fiction: from online library catalogues and the World Wide Web to the vision of machines that will eventually surpass humans in intelligence and even replace us by self-replicating computers divorced from biological evolution. This revolution was difficult to foretell. Howard Aiken, the Harvard mathematician and builder of the Mark I calculator, predicted in 1948 that there would be no commercial market for electronic computers; he esti-

mated that the USA would need only five or six such machines, but no more. As early as the 1960s, electrical engineer Douglas Carl Engelbart had designed the first interactive computer tools, including the mouse, on-screen editing, screen windows, hypertext, and electronic mail. However, at this time, human–computer interaction still seemed science fiction—computers were for processing punched cards, not for interacting with humans. The impact computers had on society and science was difficult to imagine, and we may be in the same position with respect to the future.

The impact goes in both directions: computers and humans coevolve. This coevolution is illustrated by the work of Charles Babbage (1791–1871), the English mathematician who is often credited with the invention of the digital computer.

1. *The Computer as a Factory of Workers*

Babbage's Analytical Engine, a mechanical computer, was inspired by and modeled on a new social organization of work: the large-scale division of labor, as evidenced in the English machine-tool industry and in the French government's manufacturing of logarithmic and trigonometric tables for the new decimal system in the 1790s. Inspired by Adam Smith's praise of the division of labor, French engineer Prony organized the project in a hierarchy of tasks. At the top were a handful of first-rank mathematicians, including Adrien Legendre and Lazare Carnot, who devised the formulae; in the middle, seven or eight persons trained in analysis; and at the bottom, 70 or 80 unskilled persons who performed millions of additions and subtractions.

Once it was shown that elaborate calculations could be carried out by an assemblage of unskilled workers—rather than by a genius such as Gauss—each knowing very little about the larger computation, it became possible for Babbage to conceive of replacing these workers with machinery. Babbage, an enthusiastic 'factory tourist,' explicitly referred to this division of mental labor as the inspiration for his mechanical computer, and he used terms from the textile industry, such as 'mill' and 'store' to talk about its parts. Similarly, he borrowed the use of punched cards from the Jacquard loom, the programmable weaving machines that used removable cards to weave different patterns. Thus, in the beginning there was a new social system of work, and the computer was created in its image.

2. *The Hierarchical Organization of the Mind*

Babbage's Analytical Engine was never completed. Around 1890, the first 'digital calculators' were manufactured, so named because they represented numbers by cogged wheels whose positions corresponded to the *digits* in the numbers denoted. (An abacus was still cheaper and faster.) Analog computers, in contrast, are devices in which continuous quantities, such as electrical potential or mechanical motion, are used instead of discrete (e.g., on/off) states. Until the first third of the twentieth century, digital calculators remained essentially mechanical. In the late 1930s, electromechanical relay switches started replacing gears and cogs in calculators. The electronic computer based on vacuum tubes and containing no movable parts appeared during World War II and was invented three times: in Germany for airplane design, in the USA for calculating artillery tables, and in the UK for breaking German secret codes. Vacuum tubes were eventually replaced by transistors, and finally by silicone chips, which allowed the transition from large mainframes to minicomputers. The first personal computers—complete with keyboard, screen, disk drive, and software—were marketed by Apple in 1977, and by IBM in 1981.

Through these dramatic improvements in hardware and speed, Smith's ideal of the division of labor became the basis for a fresh understanding of the human mind. Herbert Simon and Allan Newell proposed that human thought and problem solving were to be understood as a hierarchical organization of processes, with subroutines, stores, and intermediate goal states that decomposed a complex problem into simple tasks. In fact, a social system rather than a computer performed the trial run for the Logic Theorist, their first computer program. Simon's wife, children, and graduate students were assembled in a room, and each of them became a subroutine of the program, handling and storing information. This was the same as the nineteenth-century French '*bureaux de calculs*' and the Manhattan project, where calculations were done by an unskilled workforce—mostly women, at low pay. Similarly, Marvin Minsky, one of the founders of artificial intelligence, regarded the mind as a society of dumb agents, collectively creating true intelligence. Similarly, anthropologists have begun to use computer analogies to understand how social groups make decisions, such as how the crew on a large ship solves the problem of navigation by storing, processing, and exchanging information. The direction of the analogy thus eventually became reversed: originally, the computer was modeled on a new social system of work; now social systems of work are modeled on the computer.

3. *From Calculation to Computation*

A crucial insight occurred when the early designers of computers realized that a computer could be used not only for calculation but more generally for symbol processing. The switch positions inside the machine

could take on other meanings than numbers; they could stand for symbols representing concepts. This insight fueled the view of *cognition as computation*, that is, symbol manipulation. A computer that merely performed fancy calculations would not have suggested itself as a model of human thought. During the Enlightenment, proficient calculation indeed was seen as the essence of intelligence and even moral sentiment, but after the introduction of the large-scale division of labor around 1800, prodigious mental reckoning had became the mark of the idiot savant rather than the genius.

The notion of computation led to two questions: Can computers think? Is human thought computation? To avoid endless philosophical debate on exactly how to define the first question, Alan Turing (1912–54) suggested the 'imitation game,' now known as the Turing test: A human interrogator must distinguish between a computer and a human subject based on their replies to the interrogator's questions. In a series of such experiments, a computer's ability to think could be quantified as the proportion of interrogators who misidentify the machine as the human subject. Turing asked whether the computer is like a mind: whether it thinks, has free will, and whether one can teach machines to become intelligent using the same psychological principles used to teach children. Those who focused on the second question—is the mind like a computer?—saw the analogy pointing in the other direction. For instance, they tried to teach children using the same principles that had worked for computers. This analogy of the mind as computer fueled the cognitive revolution in psychology during the 1960s and 1970s.

The 'physical symbol system' hypothesis proposed by Simon and other researchers in artificial intelligence assumes that human cognition *is* computation in the sense of symbol manipulation. According to this view, symbols represent the external world; thought consists of expanding, breaking up, and reforming symbol structures, and intelligence is nothing more than the ability to process symbols. Information processing became the key to understanding human thought, and the tools of programming provided the mechanisms of thought: production systems, subroutines, recursion, iteration statements, local naming, interpreters, and so on. This view was not easily reconcilable with prevailing theories of thought, from Freud's theory of mind as manipulating biological energies, to the behaviorists' view that thought is simply silent speech, or to the Gestalt psychologists' view emphasizing the context dependency, interconnectedness, and emergent quality of thought. In essence, the vision of cognition as symbol manipulation revived the Enlightenment view of thought as a combinatorial calculus. For Condillac, d'Alembert, Condorcet, and other Enlightenment philosophers, the healthy mind worked by constantly breaking down ideas and sensations into their minimal elements, then comparing and rearranging these elements into novel combinations and permutations.

The view of thought as symbol manipulation also revived the old dream of a logic of scientific discovery, the existence of which Karl Popper had so vehemently disclaimed. Computer programs such as BACON managed to infer scientific laws from data, and artistic creativity was modeled on the same combinatorial principles. Computer programs were written that composed music, wrote stories, and designed new Palladian villas and prairie houses. Leibniz's dream of a Universal Characteristic that would assign to each concept a number, and reduce each problem to calculation, seemed to have come true. Like Babbage and Pascal, Leibniz had designed one of the first computers.

Through the medium of the computer, the Enlightenment psychology re-entered modern psychology. This is, however, only half the story. Just as George Boole had set out in the mid-nineteenth century to derive the laws of logic and probability from the psychological laws of human thought, researchers such as Simon and Newell tried to enter into their computer programs the results of psychological research about human thought. For instance, the concept of heuristics that speed up search in a problem space proved to be essential for intelligent programs. Heuristics are smart rules of thumb that humans apply in everyday reasoning when optimization, such as maximizing expected utility, is impossible or would cost too much time (see *Decision Making: Nonrational Theories*).

The metaphor of the mind as computer must be distinguished from John von Neumann's (1903–57) metaphor of the brain as computer. Von Neumann, known as the father of the modern computer, was concerned with similarities between the nervous system and computer, neuron and vacuum tube—but added cautionary notes on their differences. Turing, in contrast, thought the observation that both the modern digital computer and the human nervous system are electrical was based on a superficial similarity. He pointed out that the first digital computer, Babbage's Analytical Engine, was mechanical rather than electrical, and that the important similarities to the mind are in function rather than in hardware.

How quickly did the metaphor of the mind as computer come to be accepted by cognitive psychologists? Consistent with the tools-to-theories heuristic—which says that new scientific tools suggest new metaphors of mind, society, or nature but are accepted by a scientific community only after the members have become users of the tool—the broad acceptance of the mind as computer metaphor was delayed until the 1970s and 1980s, when personal computers became the indispensable tool of researchers. Before the advent of personal computers, researchers had little direct contact with the large mainframe computers,

and for those who did, computers were a source of constant frustration. For instance, in an average week in 1965–66, the PDP-4C computer of the Center for Cognitive Studies at Harvard saw 83 hours of use, but 56 of these were spent on debugging and maintenance. A 1966 technical report of the Center was entitled 'Programmanship, or how to be one-up on a computer without actually ripping out its wires.' The metaphor of the mind as computer only became widely accepted after the tool had become part of the daily routine in psychological laboratories.

4. *From Thought Experiments to Simulation*

Thought experiments, from physics to political science, have been a research tool in situations where no data could be obtained, such as counterfactual historical scenarios. The computer created a new species of thought experiments: computer simulation. Jay Forrester's Club of Rome world resource model—which predicted worldwide famine followed by war in the first third of the twenty-first century—is one of the best-known early social science simulations. Robert Axelrod's classic 'prisoner's dilemma' computer tournament demonstrated the success of a simple social heuristic called tit-for-tat and became the model for numerous simulations on game theory, reciprocal altruism, and political science. Simulations force researchers to specify their theoretical ideas precisely, and they have become a tool for discovering new implications of these ideas. In some fields the need to specify assumptions precisely has become a requirement for what constitutes a theory: what cannot be simulated does not deserve to be called a theory.

Computer simulation allows the modeling of dynamic systems, such as the evolution of social structures through the repeated interaction between individuals. Computer simulation is a truly new form of experiment, but social scientists are divided over what role it should have in research. In artificial intelligence, organizational theory, and demography, for instance, simulation has a long tradition. In sociology, in contrast, simulation is still rare and little respected. Simulation tends to be rare in fields where a distinction (or even hostility) between theoretical and empirical research is deeply entrenched, such as sociology, and where computers are heavily used for the analysis of empirical data. Perceived as the instrument of the 'other,' number crunching camp, the computer's theoretical possibilities tend to be forgotten.

5. *A World of Numbers and Statistics*

When the statistician Karl Pearson assembled the famous *Biometrika* statistical tables between 1914 and 1934, he used a 50-lb Brunsviga calculator and numerous clerks. The 1946 ENIAC—which is commonly referred to as the first modern digital electronic computer using vacuum tubes—could calculate 5,000 additions per second, that is, it did the work of a whole factory of clerks. Fifty years later, a tiny and relatively inexpensive Pentium chip could conduct over 500 million operations per second. This explosion of calculating power together with decreasing cost and size has had a dramatic influence on the social sciences, both positive and negative. The emphasis has shifted from collecting qualitative observations to quantitative data, and from *a priori* theoretical thinking to the *post hoc* interpretation of significant correlations.

At the end of World War II, half of the articles in the *American Sociological Review* and *American Journal of Sociology* lacked any kind of quantitative mathematical analysis; thirty years later this number had decreased to about 13 percent. In psychology, as in sociology, the predominant quantitative methods before the advent of electronic computers were descriptive statistics, such as means, variances, percentages, correlations, and cross tabulations. With the advent of fast calculators, the social sciences have witnessed a rapid growth in the usage of computationally expensive statistical methods, including factor analysis, cluster analysis, analysis of variance, and multiple regression. The computer also provided tools that could make data structures visually transparent and help researchers see underlying structures—called exploratory data analysis. Sociologists, demographers, and anthropologists were finally able to analyze large samples in a short time, such as census data and voting polls.

Computerized statistical packages such as SPSS (Statistical Package for the Social Sciences) have not only helped but also have hurt progress. Their ease of application (e.g., compute 'everything by all') has often replaced statistical thinking with statistical rituals. One widespread ritual is null hypothesis testing, in which the researcher mechanically tests whether data significantly deviates from a null hypotheses ('chance'), without ever specifying the predictions of the research hypothesis and of competing hypotheses, and without measuring effects and constructing substantive theories of some depth. As a consequence, researchers tend to skim through computer printouts looking not at the data but at the significance level of the correlations or chi-squares and engage in *post hoc* explanations for whatever comes out significant.

6. *Research and Teaching Assistance*

Next to data analysis, the most common use of computers in the social sciences is that of an efficient and patient research assistant. A computer can serve as a librarian who searches for relevant publications and historical records; as a data bank manager who stores and retrieves narrative information, qualitative

and numerical data; as a secretary who edits and files texts, keeps track of footnotes and references; and as a fast electronic mail system that facilitates contacts with colleagues worldwide. In experimental psychology, computers run entire experiments with humans and animals; they present the experimental stimuli and record the participants' responses.

This interactive use of computers would not have been possible without the invention of the personal computer that replaced the large mainframes. Engelbart's invention of the mouse exemplifies how human abilities such as hand–eye coordination can be exploited to design appropriate tools with which humans can almost intuitively interact with computers. In social science teaching, human–computer interaction can increase the participation of all students, whereas in the traditional classroom small groups of students often tend to dominate the discussion. Chat rooms on the World Wide Web have become a new and cheap data source for anthropologists and psychologists. The Web in turn has generated legal and ethical problems concerning secrecy and data protection, motivating revisions of the copyright law, tax law, and criminal law.

7. *The Computer Revolution*

Has the computer fundamentally changed the social sciences? As we have seen, the kind of impact differs across the social sciences. The invention of the modern computer has increased interdisciplinarity through the creation of *cognitive science*, consisting of artificial intelligence, psychology, philosophy, linguistics, anthropology, and neuroscience. However, sociology is notably absent from that list. Electronic mail and the Internet have made international collaboration easier and faster than ever. Simulation, expert systems, neural networks, and genetic algorithms would be unthinkable without high-speed computers, and so would be statistical analyses of large bodies of data. In these respects, the social sciences today have available a repertoire of powerful new tools and opportunities.

Has the computer revolutionized our conception of mind and society? The answer seems to be yes if we use the original meaning of the term—a re-volution, that is, a return to earlier conceptions. The metaphor of the mind as computer, for instance, has brought back the combinatorial view of cognition of the Enlightenment and Adam Smith's ideal of a division of labor. Thinking, problem solving, creativity, and scientific discovery came to be explained by a hierarchically organized mind that decomposes complex problems into simpler subroutines. In this way, theories of mind and society that had existed before computers, were expanded and perfected.

See also: Computational Approaches to Model Evaluation; Computational Imagery; Computer Networking for Education; Computer-assisted Instruction; Computerized Test Construction; Computers and Society; Metaphor and its Role in Social Thought: History of the Concept; Pearson, Karl (1857–1936); Quantification in the History of the Social Sciences; Science and Technology, Social Study of: Computers and Information Technology; Social Simulation: Computational Approaches; Statistics, History of

Bibliography

Ceruzzi P E 1998 *A History of Modern Computing*. MIT Press, Cambridge, MA

Collins T W 1981 Social science research and the microcomputer. *Sociological Methods & Research* **9**: 438–60

Crevier D 1992 *AI: The Tumultuous History of the Search for Artificial Intelligence*. Basic Books, New York

Daston L 1994 Enlightenment calculations. *Critical Inquiry* **21**: 182–202

Garson G D 1994 Social science computer simulation: Its history, design, and future. *Social Science Computer Review* **12**: 55–82

Gigerenzer G 1993 The superego, the ego, and the id in statistical reasoning. In: Keren G, Lewis G (eds.) *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Erlbaum, Hillsdale, NJ

Gigerenzer G, Goldstein D G 1996 Mind as computer: The birth of a metaphor. *Creativity Research Journal* **9**: 131–44

Langley P, Simon H A, Bradshaw G L, Zytkow J M 1987 *Scientific Discovery*. MIT Press, Cambridge, MA

McCorduck P 1979 *Machines Who Think*. Freeman, San Francisco

Schaffer S 1994 Babbage's intelligence: Calculating engines and the factory system. *Critical Inquiry* **21**: 203–27

Simon H A 1969 *The Sciences of the Artificial*. MIT Press, Cambridge, MA

Todd P M 1996 The causes and effects of evolutionary simulation in the behavioral sciences. In: Belew R, Mitchell M (eds.) *Adaptive Individuals in Evolving Populations*, Addison-Wesley, Reading, MA, pp. 211–24

Turing A M 1950 Computing machinery and intelligence. *Mind*. **59**: 433–60

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

G. Gigerenzer

Diglossia

Although the French term *diglossie* was introduced by the Arabist William Marçais in 1930 (Marçais 1930), it is the late Charles A. Ferguson who is most often credited as the first to introduce the notion of a 'high' (H) and a 'low' (L) variety or register of a language in a classic (Ferguson 1959) article with the now famous one-word title, 'Diglossia.' The gist of his widely

influential essay, which has been reprinted several times, was to demonstrate that the idea of H and L registers best explained the pervasive linguistic distinctions observable in a few speech communities concerning the strict complementary distribution of formal vs. informal usage. For Ferguson, who calqued the term from the French, there were only four 'defining' languages which he considered representative: Arabic, Swiss German, Haitian Creole, and Modern Greek. Since then, many other languages (in actuality, the speech communities which speak them) have been re-evaluated as diglossic, although most, to be sure, are far less diglossic than Arabic is (see further Ferguson 1991). In fact, Ferguson predicted what was to come when he affirmed (Hymes 1964, p. 429): 'it is likely that this particular situation in speech communities is very widespread, although it is rarely mentioned, let alone satisfactorily described.' It is important to emphasize that Ferguson's original conception of diglossia sharply contrasted with bilingualism, since he states that diglossia is different from 'the analogous situation where two distinct (related or unrelated) languages are used side by side throughout a speech community, each with a clearly defined role' (Hymes 1964, p. 429).

I focus in this essay on the case of Arabic, the area of my fieldwork expertise. Although Arabic diglossia is similar to that of other diglossic speech communities, there are some noteworthy differences. Ferguson (Elgibali 1996, p. 63) points out that, until recently, an after-dinner speech in Greek would have had to be given in katharevousa (H), but nowadays, demotiki (L) is acceptable. However, in south India, the Telugu community would never use H that way, and Ferguson further elucidates: 'People write articles, and especially poetry, in it, but it is never used in speaking on formal occasions' (Hymes 1964).

Over the past few years, the notion of the continuum has proven useful in attempts to redefine diglossia as it relates to triglossia, tetraglossia, and multiglossia or polyglossia (Kaye 1994). This latter perspective represents the current linguistic scene, particularly as it applies to the world's most complicated, in my opinion, diglossic situation, viz., Arabic (see, e.g., Kaye 1994 and Hary 1996). As illustrative, consider that 'what do you want (m. sg.)' is *maa(ðaa) turiid* in Modern Standard Arabic (hereafter MSA, a modern version of Classical Arabic), whereas an Egyptian would say in Egyptian Colloquial Arabic (ECA) for the very same thing *Gaawiz* or *Gaayiz eeh*, a Sudanese *daayir šinu*, and a Lebanese *šu baddak*. Diglossia can, however, border on bidialectalism, perhaps even on bilingualism, depending on the situation, since some native speakers of Arabic would be hard-pressed to understand a radio or TV news broadcast in MSA, let alone comprehend the oral rendition of an essay on a modern technological or literary theme. Native oral and written knowledge of and fluency in MSA are entirely different matters, however, which need not concern us here (see Parkinson 1991, 1993). Of course, it is important to recognize that an Arab deprived of an education will have far less knowledge of MSA than a university graduate.

Let us now consider an English parallel (although it is a rough one only) to the classical diglossia of Ferguson 1959. English has a formal style or register: 'To whom do you wish to speak?' corresponds to its informal equivalent 'Who do you want to talk to?' Further, one may also compare the triglossic verbs wed—get married or marry—get hitched, the diglossic verbs 'put out' (say, a cigarette) vs. extinguish, and the different nuances conveyed by the nouns kids vs. children vs. offspring vs. progeny (see Kaye 1991 for details). English even has a difference in H and L prepositional usage. Consider upon vs. on, as in 'Upon arrival, go directly to the baggage claim area.' Further down the continuum ladder in the mesolect, one may utter: 'On arrival …,' or 'On arriving … .' If one were uttering this sentence to a friend (toward the basilect), though, it would come out something like, 'When you get there, go straight to get (or pick up) your luggage.' It is the different vocabulary items in particular which instantaneously mark a speaker as using the H 'classical/standard,' or the L 'colloquial/vernacular/*patois*.' A particularly good example of this phenomenon is MSA *ðahaba* 'to go' vs. the very widespread vernacular *raaħ* or *maša/miši*. It is the utilization of the former which would be interpreted by a native speaker as a symbol of H, whereas the use of the latter would be indicative of L.

Let us now turn to the intricate phenomenon of register mixing/switching/variation. First of all, even Ferguson himself admits that he failed to examine this topic in the 1959 paper (Elgibali 1996, p. 61). It would be considered very odd to mix registers in the wrong fashion, although all sorts of register-dialect mixing regularly occur throughout the Arab world. Thus, *maa(ðaa) Zaawiz* (MSA 'what' + ECA 'want') would be perceived as most awkward—in fact, downright bizarre. It simply does not and would not occur. By a similar token, speaking MSA to one's maid at home or bargaining over the price of an item in a crowded market in MSA would be absurd, even ludicrous (unless one was trying to be facetious, weird, or sarcastic; however, non-Arabs can and often do use MSA in these and related situations, and can get away with it). This violates the 'normal' functional constraints on when and how to mix MSA and a colloquial dialect. On the other hand, ECA *da ma biyustaxdamš* 'this is not used' is perfectly acceptable, with its MSA imperfect internal passive plus the ECA *-š* 'negative.' The aforementioned examples bolster the viewpoint that native speakers possess an overall 'communicative competence' rather than a mere 'grammatical competence.'

Before taking up the details concerning the function, acquisition, grammar, and phonology of MSA and colloquial Arabic dialects, let me state that a phenom-

enal number of works on diglossia stemming from Ferguson (1959) have been published. In fact, one might wonder if any other sociolinguistic topic has generated such a prodigious research effort over the same 40-year time span. There is one book-length bibliography (Fernández 1993), and another monographic compilation appeared at about the same time in the leading sociolinguistic journal (Hudson 1992). Much of this prolific scrutiny has centered around investigating diglossia in other speech communities. Indeed, we possess studies of diglossia in Bengali, Sinhalese, Kannada, Tamil, Telugu, Malayalam, Chinese, Japanese, Persian, Old Russian, Ottoman Turkish, Irish, and Welsh, to mention but a few languages. Some researchers, such as Joshua A, Fishman, have even deemed a very specialized case of bilingualism to be diglossia. He affirms that the H and L varieties in Paraguay are Spanish and Guaraní respectively (Fishman 1971, p. 75, originally formulated in Fishman 1968). Furthermore, Fishman (1971, p. 74) has also claimed that the use of any two varieties of a language in different situations constitutes diglossia. We may refer to this as 'extended' diglossia which, in essence, leads to labeling all speech communities diglossic. At the very heart of the term is really the matter of degree, which is why the notion of the continuum is so useful in the analysis being presented in this essay. Diglossia in Arabic is a different phenomenon from diglossia in English—quantitatively more than qualitatively—especially when one notes the extreme, all-encompassing pervasiveness of the Arabic H-L dichotomy vs. the limited extent of formal vs. informal English stylistic variation. With this in mind, it appears safe to assert that diglossia is holding steady or is on the increase in the English-speaking realm, but declining in Arab countries such as Egypt, where ECA is becoming increasingly more acceptable in written form (e.g., in advertisements) and in some formal oral situations, such as certain political or religious contexts (e.g., the speeches of Abdul Nasser or the Koranic commentary by Sheikh Shagrawi). Egypt is the most populous Arab country, and the one which has the greatest tendency to use its dialect for cultural identification and nationalistic purposes in situations bordering on the unthinkable in other Arab states.

It is important to keep in mind that diglossia was slow to evolve. In the case of Arabic, Blau (1977, p. 190) is of the opinion that diglossia is not pre-Islamic, but began as the Arab armies spread out of Arabia conquering foreign lands such as Egypt in 640 AD. This is a controversial view not shared by all Arabists. Be that as it may, the Koran was instrumental in preserving the classical language in a frozen state, while its contemporary spoken dialects continued to change, like all living languages. It is generally agreed that the most classical (read: 'most elegant' or 'best' variety of H) Arabic today is, as has always been the case, the language of the Koran.

1. *Function*

Ferguson (1959) noted that H is apropos for sermons, all sorts of letters, speeches, university lectures, news broadcasts, and poetry, whereas L is appropriate when talking to servants, waiters, clerks, and workmen, conversing with family, colleagues, and friends, and in soap operas, captions on political cartoons, and folk literature. Thus, it is inconceivable that a husband would tell his wife that he loved her using anything other than L. In a similar vein, a parent scolding a child (or two children conversing) would use L. However, a professor giving a lecture would be using MSA, although there might be some L interspersed (especially in Egypt).

2. *Acquisition*

Modern linguistics distinguishes between language acquisition and language learning. Thus, it is axiomatic that native languages are acquired, whereas superposed varieties (and foreign languages taught to adults in schools) are formally learned. Every Arab grows up speaking his/her native dialect, and is formally taught MSA, given the opportunity. No one speaks MSA as his or her mother tongue. Exposure to MSA has been on the increase throughout the Arab world in the twentieth century, although a few peripheral Arabic-speaking communities have had little or no exposure to it, thus preventing the development of diglossia.

3. *Grammar*

One of the major characteristics of H is a more involved grammar. Ferguson (Hymes 1964, p. 435) calls H 'grammatically more complex.' Basically, MSA may be viewed as a marked system, and colloquial Arabic as unmarked (see Kaye 1972 for details). That is to say, MSA marks many more categories of grammar: nominative, genitive, and accusative cases, duality in the pronoun, verb, and adjective, and so on. No Arabic dialect has retained these grammatical categories, which have been lost over time.

Let us now consider a simple sentence: 'I saw the man going to his house.'

MSA: *raʾaytu rrajula ʾaahiban ilaa baytih.*

ECA: *šuftʾirṛaagil ṛaayiħ beetu.*

In addition to the different lexemes for 'see' and 'go,' MSA marks the noun 'man' with *-a* 'accusative masculine singular,' also observable in the accusative *-an* in the following active participle. Moreover, the preposition *ʾilaa*, absent in ECA, governs the genitive case of the following noun, which would, in its non-pausal form, be *baytihi* (the form *baytih* is pausal, since it occurs in final position, i.e., before pause).

4. Phonology

It is well known that the MSA voiceless uvular plosive corresponds to a glottal stop in ECA, Syro-Palestinian dialects, etc. (Kaye and Rosenhouse 1997 for the details of cross-dialectical comparisons). However, a native knows that *alqur'aan* 'Koran' is to be pronounced only with /q/. Ferguson correctly concluded that the L phonology is thus the basic system (Hymes 1964, p. 435). Thus, it is of great importance to realize that MSA pronunciation is not uniform throughout the Arab world. This is one of the reasons for my claiming that MSA is ill-defined, while ECA and all colloquials are well-defined (Kaye 1972). The most important factor determining MSA phonetics is the nature of the L underlying input (assuming a solid background of schooling in the language). Thus, ECA speakers, such as President Hosni Mubarak of Egypt, pronounce the letter *jiim* as *giim* (a voiced velar stop for the voiced palato-alveolar affricate) because this is the ECA reflex of the *jiim*. There are ECA stress patterns which are carried over to MSA as used in Egypt; e.g., *tajannábataa* 'they (f.) both avoided' (Mitchell 1990, p. 103). As Mitchell succinctly points out: 'Vernacular influenced' is nowhere more clearly in evidence than in the matter of accentuation' (Mitchell 1990).

5. Conclusion

Kaye (1972) concluded that MSA is, to a large extent, unpredictable, and thus ill-defined. There are so many varying microscopic details, since MSA is used in so many different countries by so many people with so many different educational backgrounds. Hary (1996, p. 84) refers to this state of affairs when he affirms: 'In the middle (between H and L: ASK), the mesolect, one finds almost countless variation, or lects, used by native speakers, on different occasions and under various circumstances.'

See also: Area and International Studies: Linguistics; Areal Linguistics; Bilingualism and Multilingualism; Cultural Rights and Culture Defense: Cultural Concerns; First Language Acquisition: Cross-linguistic; Language Acquisition; Linguistic Anthropology; Linguistics: Overview; Second Language Acquisition; Sociolinguistics

Bibliography

Blau J 1977 The beginnings of the Arabic diglossia: A study of the origins of Neoarabic. *Afroasiatic Linguistics* **4**: 175–202

Eid M, Holes C (eds.) 1993 *Perspectives on Arabic Linguistics V*. John Benjamins, Amsterdam

Elgibali A (ed.) 1996 *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said Badawi*. American University in Cairo Press, Cairo, Egypt

Ferguson C A 1959, 1964 Diglossia. *Word* **15**: 325–40. Reprinted in Hymes 1964 pp. 429–39. Reference herein is to the reprinted version

Ferguson C A 1991, 1996 Diglossia revisited. *Southwest Journal of Linguistics* **10**: 214–34. Reprinted in Elgibali 1996, 49–67. Reference herein is to the reprinted version

Fernández M 1993 *Diglossia: A Comprehensive Bibliography 1960–1990 and supplements*. John Benjamins, Amsterdam

Fishman J A 1968 Bilingualism with and without diglossia: Diglossia with and without bilingualism. *Journal of Social Issues* **23**: 29–38

Fishman J A 1971 *Sociolinguistics: A Brief Introduction*. Newbury House, Rowley, MA

Hary B 1996 The importance of the language continuum in Arabic multiglossia. In: Elgibali A (ed.) *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said Badawi*. American University in Cairo Press, Cairo, Egypt, pp. 69–90

Hetzron R (ed.) 1997 *The Semitic Languages*. Routledge, London

Hudson A 1992 Diglossia: A bibliographic review. *Language in Society* **21**: 611–74

Hymes D H (ed.) 1964 *Language in Culture and Society*. Harper and Row, New York

Kaye A S 1972 Remarks on diglossia in Arabic: Well-defined vs. Ill-defined. *Linguistics* **81**: 32–48

Kaye A S 1991 Is English diglossic? *English Today* **7**: 8–14

Kaye A S 1994 Formal vs. informal in Arabic: Diglossia, triglossia, tetraglossia, etc., polyglossia-multiglossia viewed as a continuum. *Zeitschrift für arabische Linguistik* **27**: 47–66

Kaye A S, Rosenhouse J 1997 Arabic dialects and Maltese. In: Hetzron R (ed.) *The Semitic Languages*. Routledge, London, pp. 263–311

Marçais W 1930 'La diglossie arabe'. *L'Enseignement public* **97**: 401–9

Mitchell T F 1990 *Pronouncing Arabic*. Oxford University Press, Oxford, UK

Parkinson D B 1991 Searching for modern Fuṣ'ḥa: Real-life formal Arabic. *Al-Arabiyya* **24**: 31–64

Parkinson D B 1993 Knowing standard Arabic: Testing Egyptians' MSA abilities. In: Eid M, Holes C (eds.) *Perspectives on Arabic Linguistics V*. John Benjamins, Amsterdam, pp. 47–73

Parkinson D B 1996 Variability in standard Arabic grammar skills. In: Elgibali A (ed.) *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said Badawi*. American University in Cairo Press, Cairo, Egypt, pp. 91–101

A. S. Kaye

Dimensionality of Tests: Methodology

A critical concern underlying the claims for validity argument relates to ensuring appropriate representation of the construct to be assessed. It is critical to ensure that the construct is neither over- nor under-

represented, and that the assumption of unidimensionality is not violated (Ackerman 1987, Luecht and Miller 1991, Messick 1989, Tam 1992). This unidimensionality assumption specifies that a score can only have meaning if the set of items measures only one attribute. If the measuring instrument has a number of items that measure differing attributes it is difficult to interpret a total score from this set of items, to make psychological sense when relating variables to this total score, or to make comments about individual differences. For example, if there are 10 spelling and 10 mathematics items in a test then it is not clear what the total score for any individual could mean. For example, two students scoring 50 percent are not comparable as it could be that one student scored correctly on the spelling and the other scored correctly on the mathematic items. Test scores only have meaning when the items measure a single attribute (for a more complete and earlier review see Hattie 1985).

The concept of unidimensionality is a specific instance of the principle of local independence—that is, the existence of one latent trait underlying the data. The principle of local independence requires that responses to different items are uncorrelated when θ (the person's ability) is fixed, although it does not require that items be uncorrelated over groups in which θ varies. Lord and Novick (1968) gave the definition of local independence more substantive meaning by writing that:

> an individual's performance depends on a single underlying trait if, given his value on that trait, nothing further can be learned from him that can contribute to the explanation of his performance. The proposition is that the latent trait is the only important factor and, once a person's value on the trait is determined, the behavior is random, in the sense of statistical independence (p. 538).

A set of items can be said to be unidimensional when it is possible to find a vector of values $\theta = (\theta_i)$ such that the probability of correctly answering an item k is $\pi_{ik} = f_k(\theta_i)$ and local independence holds for each value of θ. This definition is not equating unidimensionality with local independence, because it can further require that it is necessary to condition only on one θ dimension and that the probabilities p_{ik} can be expressed in terms on only one θ dimension.

McDonald (1981) outlined ways in which it is possible to weaken the strong principle of local independence. The strong principle implies that not only are the partial correlations of the test items zero when the latent traits are partialled out, but also the distinct items are then mutually statistically independent and their higher joint moments are products of their univariate moments. A weaker form is to ignore moments beyond the second order and test the dimensionality of test scores by assessing whether the residual covariances are zero (see also Lord and Novick 1968, pp. 225, 544–5). Under the assumption of multivariate normality, the weaker form of the principle implies the strong form, as well as conversely. McDonald (1979, 1981) argued that this weakening of the principle does not create any important change in anything that can be said about the latent traits, though strictly it weakens their definition.

1. *Stout's Evaluation of Essential Unidimensionality*

Stout (1987, 1990) also used this weaker form to develop his arguments for 'essential unidimensionality.' He devised a statistical index based on the fundamental principle that local independence should hold approximately when sampling from a subpopulation of examinees of approximately equal ability. He defined essential unidimensionality as: a test $(U_1,\ldots, U_N)$ of length N is said to be essentially unidimensional if there exists a latent variable θ such that for all values of θ,

$$\frac{1}{N(N-1)} \sum_{1 < i \neq j < N} |Cov(U_i, U_j|\theta)| \approx 0.$$

That is, on average, the conditional covariances over all item pairs must be small in magnitude (for more detail regarding the theoretical developments see Junker 1990, 1991, 1992, Nandakumar 1994, Nandakumar and Stout 1993, Stout 1990). Essential unidimensionality thus replaces local independence and is based on assessing only the dominant dimensions. Stout has developed an empirical notion of unidimensionality to match his weaker form (Stouts T-index). There are many applications of Stout's procedures (Nandakumar 1994, Nandakumar et al., 1998), and simulation studies that have supported the value of his statistic (Hattie et al. 1996). Zhang and Stout (1999a, 1999b) have generalized the Stout procedure via the DETECT method (Dimensionality evaluation to enumerate contributing traits). This method simultaneously assesses the latent dimensionality structure and determines the degrees of multidimensionality present in test data. This procedure has been extended by Kim et al. (Douglas et al. 1998, Stout et al. 1996) to identify that partition P of the available k items into disjoint and exhaustive subsets that best represents the underlying dimensionality structure. The best partition is detected by $\text{DETECT}_{\max}$, and thus is a measure of the amount of departure from unidimensionality. The methods have been generalized to polytomous items (Nandakumar et al. 1998).

Thus, there are many methods for assessing unidimensionality, such as nonlinear factor analysis (NOHARM), full information item factor analysis (TESTFACT), confirmatory factor analysis via struc-

tural equations programs such as LISREL, AMOS, and testing essential dimensionality with DIMTEST, and detection of simple structure multidimensionality (DETECT).

2. The Concept of Unidimensionality

It is critical to note that the assumption of unidimensionality underlying most measurement models does not necessarily imply that only these models can assess the most simple of tasks. The concept of unidimensionality can be considered at various levels of analysis. At the first-order level we can construe of a set of items being combined (by appropriate weighting) to form a total unidimensional score (such as a collection of addition items leading to a total addition score). At the second-order level we can also construe of a set of first-order scores being combined to form a total undimensional score (such as a collection of addition, subtraction, multiplication, and division totals being combined to form a total arithmetic score). In both cases we can conceive of a single underlying trait.

In a different context, we could imagine a single set of items all requiring a similar cognitive process (e.g., successive or simultaneous processing) and a total score being formed—and we could consider this a unidimensional total score—even if the items were across many content domains. All too often, the notion of unidimensional has been conceived in terms of a narrowly defined set of items. The key issue is that the methods outlined to assess unidimensionality do not lead to identifying what the total score represents—this requires other validity methods (such as judgmental analyses). Identifying a unidimensional set of items is not the same as naming it.

3. Constructing Multidimensional Data

One of the problematic issues when assessing dimensionality is that there are differing ways that the dimensions can relate to the construct of interest. Coombs (1954) defined three methods of constructing multidimensional data: the conjunctive model, whereby an excess of one ability, no matter how large, does not compensate for lower ability in other dimensions; the disjunctive model, whereby an examinee will pass an item if he/she is dominant over the item in any one dimension and will fail only if the item dominates him/her in all dimension; and the compensatory model, whereby an examinee's response to an item is a function of a weighted sum of underlying abilities. Other researchers (notably Ackerman 1987, 1992, Ansley and Forsyth 1985, Reckase et al. 1986, Wang 1987, 1988) have demonstrated that with compensatory multidimensional models the univariate calibration of two-dimensional response data can be explained in terms of the interaction between the multidimensional test information and the distribution of the two-dimensional abilities, and various interpretations of the multidimensional item difficulty, multidimensional discrimination, and multidimensional item information have been suggested.

Sympson (1978) proposed a partially compensatory model where a decrease in one ability could only be offset by a large increase in the other ability (this model has often been erroneously called a noncompensatory model). Outside of a relatively narrow ability range, the probability of passing reduces to zero regardless of the value of the stronger ability (see Lord 1984). The probability of a correct response is simply the product of probabilities for each dimension (but see Coughlan 1974, Jannarone 1986).

4. Concluding Comments

A major need is for a clear explication of the concept of dimensionality. Within the item response literature there has been much agreement about its meaning, although it has often been (falsely) assumed that only tests consisting of 'simple' attributes could lead to a unidimensional test. Certainly there are cognitive processing models, whereby it may be possible for a student to solve an item using differing cognitive strategies, thus adding another meaning to the notion of undimensional; the unidimensionality of processes or strategies. Much of the discussion about newer models of item and test analyses is oriented around such notions (Mislevy and Gitomer 1995).

It is critical to ensure that any set of items map the construct of interest, and if this construct is inherently multidimensional than so too should there be multiple unidimensional subtests. At this test score level, we may desire 'lumpy' items (such as reading comprehension, or mathematical problem solving), and thus the items may assess multiple dimensions while the total test still aims to assess a single 'over all' dimension. As Dorans and Lawrence (1999) have demonstrated, it is possible to conceive of situations where analysis at the test score level would reveal a single common factor, while analysis at the subscore level would uncover a few distinct subscore factors. Similarly, Reckase et al. (1988) have demonstrated that sets of items can meet the unidimensionality assumption even though they require more than one proficient for a correct response.

At present, the Stout procedures seem worth pursuing, as they are based on clear theoretical developments, are easy to use, dependably provide indications of the dimensionality, are reasonably robust, and provide a reasonably clear and practical demarcation between one and many dimensions. Further, there is need for further research on developing multidimensional models that allow for differing ways in which items or test scores can be combined, and we need

better diagnostics to understand the complexity of the dimensions involved in item and test scores. There are already many useful heuristics available to interpret multidimensional data using unidimensional calibration methods (Ackerman 1992, Carlson 1987, Junker 1991, 1992, Luecht and Miller 1991, McKinley and Reckase 1989, Reckase et al. 1986, 1988). It is certain that discussion will continue about the concept of unidimensionality, the best forms to assess whether items are essentially unidimensional, and the importance of diagnostics to aid in constructing better unidimensional assessments.

See also: Classical (Psychometric) Test Theory; Factor Analysis and Latent Structure, Confirmatory; Factor Analysis and Latent Structure: IRT and Rasch Models; Indicator: Methodology; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Ackerman T 1987 *The Robustness of LOGIST and BILOG IRT Estimation Programs to Violations of Local Independence*. American College Testing Program, Iowa City, IA

Ackerman T A 1992 A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement* **29**: 67–91

Ansley R A, Forsyth T N 1985 An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement* **9**: 37–48

Carlson J E 1987 *Multidimensional Item Response Theory Estimation: A Computer Program*. American College Testing Program, Iowa City, IA

Coombs C H 1954 *A Theory of Data*. Wiley, New York

Coughlan J R 1974 *A Multidimensional Extension of the Normal Ogive, Logistic and Linear Latent Trait Model*. Unpublished Master's thesis, University of Toronto

Dorans N J, Lawrence I M 1999 *The Role of the Unit of Analysis in Dimensionality Assessment*. (Educational Testing Service Report, No. 99–14). Princeton, NJ

Douglas J, Kim H R, Habing B, Gao F 1998 Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics* **23**(2): 129–51

Hattie J A 1985 Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement* **9**: 139–64

Hattie J A, Krakowski K, Rogers H J, Swaminathan H H 1996 An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement* **20**: 1–14

Jannarone R J 1986 Conjunctive item response theory kernels. *Psychometrika* **51**: 357–3

Junker B W 1990 *Progress in Characterizing Strictly Unidimensional IRT Representations*. Technical Report 498, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA

Junker B W 1991 Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika* **56**: 255–78

Junker B 1992 April *Ability Estimation in Unidimensional Models When More Than One Trait is Present*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA

Lord F M 1984 Conjunctive and disjunctive item response functions. (Educational Testing Services Report, No. 150–520). ETS, Princeton, NJ

Lord F M, Novick M R 1968 *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA

Luecht R M, Miller T R 1991 *Unidimensional Calibrations and Interpretations of Multidimensional Tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL

McDonald R P 1979 The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioural Psychology* **14**: 21–38

McDonald R P 1981 The dimensionality of tests and items. *British Journal of Mathematical Statistical Psychology* **34**: 100–117

McKinley R L, Reckase M D 1983 MAXLOG: A computer program for estimating the parameters of a multidimensional extension of the two-parameter logistic model. *Behavior Research Methods and Instrumentation* **15**: 389–90

Messick S 1989 Validity. In: Linn R L (ed.) *Educational Measurement*, 3rd edn. Macmillan, New York, pp. 13–103

Mislevy R J, Gitomer D H 1995 The role of probability-based inference in an intelligent tutoring system. *User Modeling & User-Adapted Interaction* **5**: 253–82

Nandakumar R 1991 Traditional dimensionality vs. essential dimensionality. *Journal of Educational Measurement* **28**: 99–117

Nandakumar R 1994 Assessing latent trait unidimensionality of a set of items—comparison of different approaches. *Journal of Educational Measurement* **31**: 1–18

Nandakumar R, Stout W 1993 Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics* **18**: 41–68

Nandakumar R, Yu Z Y, Li H H, Stout W 1998 Assessing unidimensionality of polytomous data. *Applied Psychological Measurement* **22**: 99–115

Reckase M D, Ackerman T A, Carlson J E 1988 Building a unidimensional test using multidimensional items. *Journal of Educational Measurement* **25**: 193–203

Reckase M D, Carlson J E, Ackerman T A, Spray J A 1986 *The Interpretation of Unidimensional IRT Parameters when Estimated from Multidimensional Data*. Paper presented at the Psychometric Society Annual Meeting, Toronto

Stout W F 1987 A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika* **52**: 589–617

Stout W F 1990 A new item response theory modelling approach and applications to unidimensionality assessment and ability estimation. *Psychometrika* **55**: 293–325

Stout W, Habing B, Douglas J, Kim H R 1996 Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement* **20**(4): 331–54

Tam S S 1992 *A Comparison of Methods for Adaptive Estimation of a Multidimensional Trait*. Unpublished doctoral dissertation, Graduate School of Arts and Sciences, Columbia, NY

Wang M M 1987 *Estimation of Ability Parameters from Response Data to Items that are Precalibrated with a Unidimensional Model*. Paper presented at the Annual meeting of the American Educational Research Association, Washington, DC

Wang M M 1988 *Measurement Bias in the Application of a Unidimensional Model to Multidimensional Item-response Data*. Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, LO

Yu F, Nandakumar R in press Poly-Detect for quantifying the degree of multidimensionality. *Journal of Educational Measurement*

Zhang J, Stout W 1999a Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika* **64**(2): 129–52

Zhang J, Stout W 1999b The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika* **64**(2): 213–49

J. Hattie

Diplomacy

Diplomacy is the art of obtaining agreement between countries who need one another's cooperation to create or maintain a situation in which each of them has an interest. The needed cooperation may involve refraining from certain kinds of behavior as well as changing the location or condition of people, goods, structures, military deployments, and other assets. The premise of diplomacy is that these behaviors would not occur automatically: they require the positive or negative stimuli of deliberate international (usually intergovernmental) interaction among those with the power to bring about the desired results. It is not diplomacy, for example, when the immigration officials of a country bar a group of refugees from coming into the country, nor when the supporters of the refugees voice their outrage; diplomacy transpires when officials of the affected countries interact with one another to clarify or modify the rules and procedures each of them will be applying to deal with the situation.

1. *The Essential Characteristics of Diplomacy*

Conventionally equated with negotiations to resolve international conflict, diplomacy is widely regarded as an alternative to war—and war as a failure of diplomacy. But diplomacy can also be the essential process in a nonadversarial interaction of friendly countries attempting to obtain the best allocation of costs and benefits in a commonly sought project. The term 'diplomatic' has been imported from the international into the domestic realm, where it is commonly applied to relations among groups or individuals in which polite interaction is artfully employed to avoid otherwise tense confrontations; however, the discussion here will be confined to the characteristics and role of diplomacy in the international realm.

Diplomacy, so defined, refers mainly to agreement-seeking behavior on the part of national governments. Diplomats are the officials who do the actual negotiation. But not all diplomats are engaged in negotiations; some merely represent their governments in foreign lands or international forums, conveying information about their home countries, and reporting back on conditions at their posts. Bargaining and informational diplomacy are practiced as well by international actors other than nation states—namely, the United Nations, the European Union, numerous specialized intergovernmental organizations, and, increasingly, various nongovernmental organizations, particularly those that are accorded quasi-official status in certain international forums.

Although the most important and characteristic aspect of diplomacy is face-to-face bargaining among officials of governments, such bargaining is also frequently conducted through unilateral pronouncements by officials of what they will or will not accept. Diplomacy's principal medium is verbal communication—spoken, written, often electronically transmitted—and therefore is itself basically nonviolent; but the words may threaten, or be delivered simultaneously with coercive action, even massive military attacks, to alter an adversary's assessment of the costs and benefits of accepting or rejecting particular arrangements. Diplomacy's attribute resides in its purpose—international agreement; as such it requires a modicum of nonviolent interaction, even when that interaction is coerced (as in the imposition of surrender terms at the end of a war by the victors on the vanquished). Diplomatic undertakings may fail, in that agreement is not obtained; but, succeed or fail, as long as the purpose of the international interaction is agreement, it is still diplomacy. Without any interaction, however—in particular, verbal interaction to at least explore possibilities for agreement—there can be no diplomacy.

2. *Diplomacy as a Reflection of World Politics*

The rules and rituals of international diplomacy have never been a self-contained system. The picture of the diplomat as an impeccably mannered aristocrat exchanging toasts, repartee, and subtle threats with his counterparts is a caricature derived from the Parisian scene in the court of Louis XIV. Rather, the protocols of diplomacy have always been a subsystem of the more comprehensive system of political relationships prevailing during a particular era among various sets of countries and international organizations. Thus, as world politics have evolved, so have the rituals, forms, and objectives of diplomacy, even though its most basic function—to forge agreements among international actors—has persisted.

2.1 *The Evolution of Diplomacy*

The evolution of diplomacy parallels and reflects the transformation of international relations from a system of maneuvering and secretive deal-making by

kings and their emissaries to the contemporary practice of conducting foreign policy as a continuation of domestic politics through public speech-making and media manipulation. While late twentieth-century diplomacy is unique in many respects, some of its prominent features are inherited from past international systems.

The most significant legacy from previous international systems are the norms that were codified in the Westphalian peace treaties ending the wars of religion in 1648 and which became the cardinal principles of statecraft—at least for Western Europe. No longer regarding their rivals as agents of the devil, the Protestant and Catholic monarchs, re-established the ancient imperial rituals of polite dialogue among governments, their high officials, and plenipotentiaries, even when threatening or engaging in war against one another. Most important, they pledged respect for state sovereignty and the principle of noninterference in one another's domestic affairs. This structurally anarchic yet mutually respectful society of states prevailing in Western Europe from the last half of the seventeenth century to the last quarter of the eighteenth century is often referred to by diplomatic historians as the 'classical' system. The so-called Classical Balance of Power denotes the presumably self-equilibrating characteristic of the system, featuring short-term and shifting alliances through which states, big and small, preserved their sovereignty through balancing one another's military power.

The classical power-balancing diplomacy could work as long as the games of maneuver and alliance making could be played by elites who did not have to answer to the people within their jurisdictions. But as the countries of Europe (and America) evolved into nation–states in which major governmental undertakings required broad cooperation from their commercial, industrial, and intellectual classes, often in the form of parliamentary approval, the conditions allowing for flexible and shifting alliance commitments began to erode. Following the American and French Revolutions, war—now to be fought by citizen armies and industrially produced arsenals, rather than lightly armed mercenary troops—could no longer be credibly threatened or sustained without the consent of the governed. And such support would not be forthcoming unless the nation as a whole could be convinced that the country's vital interests and way of life were at stake. Accordingly, as became evident in World War I, military hostilities once begun became more difficult to stop. War, previously regarded as an instrument of diplomacy, was becoming its antithesis, at least for the democracies. The democratic aversion to war's disruptions of normal life was one of the reasons why Hitler and the Japanese militarists were able to bully the world as much as they did before a balance of power could be mobilized against them. The development and use of nuclear weapons at the end of World War II reinforced these trends.

For many statespersons and citizens, the highest imperative of diplomacy has become to manage international conflicts *without* resorting to war. But for others, the threat or actual use of military force, even the new weapons of mass destruction, would continue to be an instrument of diplomacy. Moreover, the expanding and deepening interdependence of the world's peoples has, paradoxically, moved the issue of military intervention to the top of the world's diplomatic agendas.

2.2 *The New Diplomacy*

The new diplomacy, generated by the events and forces since World War II that have transformed the world polity—the Cold War, decolonization and self-determination movements, the human rights ideology, terrorism, economic globalization, and threats to the planet's ecologies—involves an expansion of the ends and means of foreign policy, and of the functions of international organizations. Political, economic, and environmental conditions in the majority of the world's countries have become the world's business. Accordingly, the new diplomacy features efforts by governments and nongovernmental actors to affect conditions *within* other countries, not only to adjust relations between countries.

Guardians of traditional statecraft are concerned that the norms and structures that have sustained the 'anarchical society' of mutually respectful sovereign countries are being undermined by the new interventionist diplomacy. They worry that the new diplomacy is lending support, whether purposefully or inadvertently, to the forces subverting the capacity of governments to protect their national polities and cultures, and is giving license to unaccountable actors to impose their will on others. The anarchical international *society* is in danger of degenerating into an uncontrolled anarchy in which only the most aggressive thrive. Yet, realistically, there is no going back to the traditional Westphalian system, for the technology-driven mobility of people, things, and information that has been eroding the sovereignty of the territorially-defined nation state is irreversible. At the most, defenders of the authority and power of nation states can attempt to slow the process through protectionist economic measures to control access into their markets and through restrictions on immigration. National governments will still retain their role, into the foreseeable future, as the main authorities for assuring that people remain accountable to larger community standards for how they treat one another, and how they manage the natural environment. But, increasingly, the national government will need to share this role with supranational and subnational authorities. And these political entities, as they assume more responsibility for determining who gets what, when, and how in world

society, will themselves need to engage in diplomatic interaction with the existing nation states and with each other.

3. The Requisites of Effective Diplomacy

Diplomacy in the world polity, like legislation in domestic polities, is directed toward getting parties to agree on a particular course of action (or inaction), so they will not have to be physically forced to accept it. But unlike the domestic legislative process, in which it is sufficient to obtain agreement among a controlling majority for the course of action to take place, in international diplomacy each participating state normally has the recognized unilateral right to accept or refuse what even a substantial majority of the states have agreed upon. (The major exception to this norm is the UN Security Council's authority, in situations threatening international peace and security, to order all states to obey its resolution.) Thus, diplomatic strategies must always be targeted specifically on each state whose adherence to an agreement is sought. Each state must be convinced that its own interests will be better served by agreeing to the arrangements under discussion than in refusing to agree.

Effective diplomacy, accordingly, requires the ability to induce agreement from states that may initially be opposed to a contemplated arrangement, but whose acquiescence is essential for its realization. Skillful diplomats are able to secure favorable terms for their countries by persuasively outlining the benefits and costs their counterparts can expect from alternative arrangements. This requires a detailed and accurate understanding of both the material conditions and the political circumstances (domestic and international) of the countries involved in a particular diplomatic encounter.

States with the material resources to affect their counterparts' preferences for alternative outcomes by credibly offering to provide valued objects ('carrots') or threatening to apply uncomfortable or painful sanctions ('sticks') have traditionally enjoyed an enormous advantage in the arenas of international diplomacy. But increasingly, states, international organizations, and political movements, deficient in the relevant material assets are able to redress such resource–power imbalances through artful appeal to, and manipulation of, the cultural and political values of the parties to a negotiation, other governments, and implicated transnational and subnational groups. Success in regional and global coalition-building (what used to be called 'balance of power' politics), now penetrating beyond the crust of state sovereignty, has become the most important requisite of effective diplomacy, especially the new diplomacy.

See also: Human Rights in Intercultural Discourse: Cultural Concerns; International and Transboundary Accords, Environmental; International Organization; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; State Formation; War: Anthropological Aspects; War: Causes and Patterns; Warfare in History

Bibliography

Aron R 1968 *Peace and War: A Theory of International Relations*. Praeger, New York
Brown S 1994 *New Forces, Old Forces, and the Future of World Politics*. Harper Collins, New York
Bull H 1977 *The Anarchical Society*. Columbia University Press, New York
Gulick E V 1955 *Europe's Classical Balance of Power*. Cornell University Press, Ithaca, NY
Kissinger H A 1994 *Diplomacy*. Simon and Schuster, New York
Mattingly G 1955 *Renaissance Diplomacy*. Houghton Mifflin, Cambridge, MA
Morgenthau H 1978 *Politics Among Nations: The Struggle for Power and Peace*. Knopf, New York
Nicolson 1950 *Diplomacy*. Oxford University Press, London

S. Brown

Disability, Demography of

There are two definitions of 'disability' that have wide acceptance in demographic studies. The first, developed by the World Health Organization, conceives of disability as a restriction or lack of ability to perform an activity normally, as a result of the impact of disease. The second definition, developed by Saad Nagi and adapted by the US Institute of Medicine, is similar, but emphasizes the relevance of disease impact to the ability to perform social roles. This latter definition allows for the possibility of intervention to minimize restriction of social roles, despite disease and consequent physical impairment (Verbrugge and Jette 1994). It has gained in popularity as a framework for thinking about the needs of the world's aging population. More recently, this framework has been applied to understand disability among children, persons entering the labor force, and working-age adults, with an emphasis on supportive services, such as home health and physical therapy, to improve or maintain physical functioning and other aspects of quality of life, and elimination of environmental barriers.

1. History of Disability Measurement

1.1 Work Limitation

Traditionally, demographers relied on census data on the number and characteristics of people, combined with information on death from vital statistical sys-

tems, to project the working age population. A table of working life added age-specific data on labor force participation (Wolfbein 1949). This system proved unworkable, as many persons of working age are limited in their labor force participation and others retire voluntarily. While these causes of being out of the labor force could be handled with multiple decrement life tables, the recognition that people in these situations often re-entered the work force reduced the value of this working life table approach (Schoen and Woodrow 1980).

Despite these practical problems, nations found it essential to be able to project the numbers of workers by age, sex, and educational levels, the numbers of these workers who will be employed under current and alternative projections of the demand for workers, the number of working-age persons who are disabled or limited in labor force participation, and the retired population who would draw on pension systems. This necessitated an individual level of analysis in which more direct data on factors in labor force participation (including health and disability) can be collected, and the use of methods for event history data to identify the demographic social variables involved.

To estimate the scope of disability within countries, scientists have devoted substantial effort to the development of valid measures. Health interview surveys in Europe and the USA collected data on disability in large populations as early as the late 1800s and the early 1900s. Disability was measured as the prevalence of days of sickness per 1,000 persons and the number of people who were sick and unable to work on the day of interview, thus introducing a measurement approach still in use in current large-scale surveys. In addition, measures of labor force participation were expanded to include persons who report they are not working because of disability.

1.2 Activities of Daily Living

In 1949, the Commission on Chronic Illness was established in recognition of the high prevalence of chronic illness in the USA. The Commission called for a classification of difficulty in the performance of activities of daily living in order to measure the disabling impact of chronic illnesses and conditions on people's lives. A number of scales of this type emerged in the 1950s. Among the most valid and replicable is the Index of Activities of Daily Living (ADLs), constructed on the basis of empirical observations of patients' recovery of ability to perform necessary daily tasks after hip fracture (Katz et al. 1963). Standard ADLs include bathing, dressing, toileting, transferring in and out of bed and chairs, eating, and getting around inside the house. Measures more indicative of cognitive than physical impairment, e.g., using the telephone and managing medications, are often included also. This scale and its many hybrids have been in use since it was developed, both to monitor the decline and recovery processes in individuals, and also to measure levels of disability in populations.

Given that ADL scales tap levels of disability most relevant to the very frail, Lawton and Brody (1969) devised the Instrumental Activities of Daily Living (IADL) Scale, designed to measure competence at a more complex level than the basic ADL scale. This scale includes activities more typically problematic for people with chronic conditions who live in the community, such as shopping, housework, and the ability to handle finances. Unlike ADLs, which are mainly applicable to people in long-term care institutions, IADLs are particularly useful in capturing variation in physical functioning among community-dwelling populations. Severity of disability at the population level is typically measured by counts of activity limitations, e.g., the proportion of the population with one or two ADL limitations versus three or more. While in general limitations in ADLs are indicative of a more severe disability than IADLs, empirical research has not identified a strict hierarchical order. Nevertheless, estimates of disability are often framed in terms of the proportion of the population with (any) ADL limitations, and separately, with (any) IADL limitations. More recently, it has become common to ask survey respondents who indicate difficulty in activity performance whether they need help, and whether needed help is received; thus generating estimates of unmet need for services.

In specialized studies of disability during the 1990s these scales have been supplemented by questions that tap a variety of physical impairments, including a series of questions developed by Saad Nagi to measure mobility and flexibility, e.g., ability to walk a quarter mile (Nagi 1979). Such studies generally include measures of sensory impairments as well, including the ability to see, hear, and communicate. The strategy to use measures of daily activity, ability to perform basic tasks, and participation in age and gender-appropriate roles and functional capacities is widely applicable over the life span, in developed nations, and among different socioeconomic and cultural groups. Such measures provide an indication of level of need for medical and supportive services and equipment within given populations. For persons not covered by old age medical insurance plans, information on access to and cost coverage for services for potentially disabling health conditions is essential. In the USA both cost restrictions and lack of access to appropriate services are factors sometimes preventing appropriate medical care and support services for both children and adults.

2. Considerations and Cautions in Disability Measurement

While still in widespread use, the applicability of measurement of limitation in major activities (i.e., work, housekeeping, and going to school) is limited

and interpretation is difficult. Limitation in ability to work is very much determined by what one does for a living, for example. Two people with similar health profiles may report limitation in major activity quite differently if one is a construction worker and the other a teacher. Furthermore, it is often difficult to define one's major activity, particularly for people of retirement age. Thus, while this measure may be helpful in estimating attrition from the work force due to disabling health conditions, it is not a precise measure of the prevalence or severity of disability in populations. Limitation in 'any' or 'other' activities presents even greater challenges to interpretation.

Use of measurement of limitations in ADLs/IADLs has far wider application as a measure of disability in populations, although such measures do have problems too. Differences in question wording, the specific activities included in measurement, and differences in study design can result in variation in estimates of ADL variation across surveys of community dwelling populations.

While ADLs such as bathing and dressing have universal meaning, IADLs such as shopping and cooking have a strong gender role component, which can result in substantial underestimates of disability for men. Thus, ADL measures are more 'pure' measures of physical function than IADLs, while assessment of IADL disability has a greater subjective component.

Other confounding influences in measurement of ability to perform daily living activities include the measurement of capability ('Can you ...') versus actual performance ('Do you ...'), ascertainment of difficulty ('Do you have difficulty with ...') versus dependency ('Do you receive help with ...'), inclusion versus exclusion of human help and/or special equipment in ascertaining difficulty ('Even with help and/or use of special equipment, do you have difficulty ...'), and use of proxy vs. self-reporting. The necessarily subjective nature of disability measurement presents challenges for researchers to achieve comparability across studies within countries. Nuances in languages present even greater challenges to assessing the scope of disability across nations.

3. *Demographic Methods for the Study of Disability*

The recognition that disability is a phenomenon best measured at the individual level has led to great advances in the demographic study of disability. Crucial to this new approach is the recognition of disability as a dynamic process characterized by entries and exits and competing risks. Thus studies of labor force participation among older workers have been designed to take into account potentially disabling medical conditions, type of job and work disability, shift to alternative jobs where the condition is not disabling, and the competing state of retirement, as well as movements into and out of these statuses. Both multidimensional life tables and methods for event history analysis have proved useful (Manton et al. 1992). The use of event history methods has proved to be of particular value because of their ability to incorporate demographic or socioeconomic covariates that affect these processes. The greatest advances have come in the study of the elderly, but these methodologies should prove powerful in the study of disability among children, the transition from adolescence to adulthood, and the situations of adults who have not reached old age. A shortcoming of this research approach as currently practiced is that it tends to focus exclusively on the individual instead of an individual within an environment that may present obstacles or provide assistance, though extension to include contextual variables is possible.

4. *Current Trends in Disability*

4.1 *Activity and Work Limitations*

Worldwide concerns about population aging and the likely increase in rates of disability associated with lengthening life expectancies have escalated research interest in the measurement of population disability trends. In the USA, a variety of cross-sectional and longitudinal national surveys are available to determine current disability rates as well as disability trends over the past several decades. Overall, annual estimates of disability (measured as activity and work limitations) available from the National Health Interview Survey since 1970 indicate that the proportion of the US population with disabilities has increased during that time period.

When disability rates are calculated by age group, it is clear that the elderly experience disability at approximately twice the rate of people between ages 45 and 64, and at four times the rate of the younger working age group (aged 18–44). Thus, the increase in the proportion of the population who are elderly contributes to the increase in rates of disability in the larger population, although actual rates of activity limitations among persons 65 and older remained relatively constant from the early 1980s to 1994, estimated at about 38 percent for men, and 39 percent for women (Kaye et al. 1996).

Activity limitation rates have also remained constant among the older working age group (ages 45–64) since the early 1980s, at about 22 percent for men, and 23 percent for women. Work disability is relatively constant in this population, with 11 percent unable to work and 7 percent limited in amount and/or kind of work. In contrast, disability has increased somewhat among people under 45 years of age, particularly in the early 1990s. From 1990 to 1994, activity limitations among younger men (aged 18–44) increased from 8.7

percent to 10.2 percent, and among younger women from 8.9 percent to 10.3 percent. During the same time period, the proportion of this subgroup that is unable to work increased from 2.9 percent to 3.7 percent.

Similarly, activity limitations among children have increased in the early 1990s, from 5.6 percent to 7.9 percent among boys, and from 4.2 percent to 5.6 percent among girls. These increases are reflected in an increase in the prevalence of asthma and cognitive disorders such as Attention Deficit Disorder and learning disabilities. Thus, increases in disability among younger adults and children, combined with an increase in the proportion of the population who is elderly, are largely responsible for the trend toward higher disability levels in the overall US population (Kaye et al. 1996).

4.2 ADL/IADL Disability Trends

There is a growing consensus that while rates of certain chronic diseases have increased in recent decades, both in the USA and in other countries, rates of disability in daily living activities have actually decreased within the elderly population. Early reports of this decrease were met with skepticism, with some researchers attributing apparent decline to an artifact of question wording or survey design. However, increasing consistency in findings across several longitudinal US surveys provide reliable evidence that disability, has decreased from the early 1980s until the early 1990s particularly in IADLs (Manton et al. 1998). Findings regarding improvement in ADL disability are less consistent.

Researchers have begun to focus their work on identifying factors that account for observed decline. Advances in medical technology and disease management as well as improved access to medical and support services and technical aids that may reduce difficulty in task performance have been credited with at least some of the reduction in measured disability levels. However, recent evidence suggests that education is the most important factor associated with disability decline in the US population, as a less educated cohort of elderly persons is being replaced by a cohort of people raised in more prosperous times. Attribution of disability decline to increases in educational levels suggests that observed decline may continue, albeit at a reduced rate (Freedman and Martin 1999).

The Global Burden of Disease Project has undertaken the massive task of estimating the incidence and prevalence of disability for world regions, for men and women across the lifespan (Murray and Lopez 1996). In this work the expectation of life at birth is divided into years of life without disability and with disability. In the established market economies males can expect a life at birth of 73.4 years, of which 8.1 percent will be spent in disability. Women's expectation of life at birth is 80.5 years, of which 8.3 percent will be spent in disability. While women spend more time than men in the disabled state because of their longer life span, that longer life span itself does not lead to an inevitable increase in the percentage of years in disability. Indeed, in those regions with an expectation of life at birth of less than 60 years, more person years are spent in disability than in nations with lower mortality, with 12 percent or higher of the lifespan lost due to disability. Similar observations apply to calculations for the population at age 60 as well. The best strategy for relieving disability over the lifespan is to begin promoting health among younger adults in order to reduce disability in later middle age and old age.

See also: Accidents, Normal; Disability, Geography of; Disability: Psychological and Social Aspects; Disability: Sociological Aspects; Illness Behavior and Care Seeking; Injuries and Accidents: Psychosocial Aspects

Bibliography

Freedman V A, Martin L G 1999 The role of education in explaining and forecasting trends in functional limitations among older Americans. *Demography* **36**: 461–73

Katz S, Ford A B, Moskowitz R W, Jackson B A, Jaffe M W 1963 Studies of illness in the aged: The index of ADL. *Journal of the American Medical Association* **185**: 914–19

Kaye H S, LaPlante M P, Carlson D, Wenger B L 1996 Trends in disability rates in the United States, 1970–94. In: *Disability Statistics Abstracts 17*. Disability Statistics Center, University of California, CA

Lawton M P, Brody E M 1969 Assessment of older people: Self-maintaining and instrumental activities of daily living. *The Gerontologist* **9**: 179–86

Lawton M, Powell B E 1969 Assessment of older people: Self-maintaining and instrumental activities of daily living. *Gerontologist* **9**: 179–88

Manton K G, Woodbury M A, Stallard E, Corder L S 1992 The use of grade of membership techniques to estimate regression relationships. *Sociological Methodology* **2**: 321–81

Manton K G, Stallard E, Corder L S 1998 The dynamics of age-related disability 1982–1994 in the US elderly population. *Journal of Gerontology*, *Biological Sciences* **53A**: B59–B70

Murray, C J L, Lopez A D 1996 Global and regional descriptive epidemiology of disability: Incidence, prevalence, health expectancies and years lived with disability. In: Murray C J L, Lopez A D (eds.) *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020*. Harvard University Press, Boston, pp. 201–46

Nagi S Z 1979 The concept and measurement of disability. In: Berkowitz E D (ed.) *Disability Policies and Government Programs*. Praeger, New York

Schoen R, Woodrow K 1980 Labor force status life tables for the United States 1972. *Demography* **17**: 297–322

Verbrugge L M, Jette A M 1994 The disablement process. *Social Science in Medicine* **38**: 1–14

Wolfbein S L 1949 The length of working life. *Population Studies* **3**: 286–94

S. Allen and D. Hogan

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Disability, Geography of

The sociospatial causes and consequences of disability, that is, of being socially constructed as negatively different on the basis of mental or physical bodily impairment, has become a significant focus of geographic inquiry. The assumption that only able-bodied women and men occupy and negotiate social space has impoverished our understanding of society and space. Differences in mental and physical abilities affect patterns and degrees of access to urban public spaces, services, and employment opportunities, and help to determine who is most affected by urban and regional changes such as rising poverty and loss of affordable housing. Differences in bodily ability find expression in the geography of cities and regions: in the spatial concentration of disabled persons in poor inner city neighborhoods with a concentration of social services, for instance.

Critical geographers view disablement as a sociospatial process that marks mentally and/or physically impaired bodies as negatively 'other' and works to marginalize disabled persons within social space. Geographies of disability are, thus, as much about ableism as a regime of power based on able-bodied privilege and the sociospatial practices that sustain it, as about disabled persons' experiences of trying to negotiate environments that physically and/or socially exclude and oppress them.

1. *Why a Critical Geography of Disability?*

The recent shift toward critical geographies of disability is a response to changing conditions within society, and new perspectives on the differences disabilities make in how people are situated in society and space (e.g., in terms of access to basic necessities and acceptance in urban spaces). An increasingly vocal disability rights movement, particularly in the West, but also internationally, aging populations with higher incidence of disability, cutbacks in social programs for the disabled, and problems such as grinding poverty and inaccessible cities, have helped to encourage such research. New perspectives on disability issues have also been influential. The growth of disability studies has challenged conventional biomedical models of disability with 'social constructionist' ones, and offered significant insights into what it is like to be disabled in various spaces of everyday life (e.g., Oliver 1990, Morris 1991). These society-centered perspectives view disability as the outcome of social processes such as the economic devaluation of the disabled through lack of employment and low wages, and cultural representations of disabled people as inferior. Disability is thus a social rather than individual condition. This does not necessarily mean that disablement is entirely socially produced. Bodily impairments affect experiences of disablement. Visual impairment may result in difficulty moving within spaces such as sports arenas and negative reactions to such visible differences may help to make disabled persons feel 'out of place.' Similarly, chronic pain sufferers often find it difficult to negotiate spaces of daily life and interact with others. Scholars continue to debate how best to incorporate such bodily realities within conceptualizations of social processes of disablement.

Changes in the discipline have also encouraged critical geographies of disability. These include: greater use of critical social theory, feminist geographic studies of social differences (e.g., gender and sexuality) and marginalization in urban space, recognition of multiple bases of oppression in society and space, and desires to use geographic knowledge to help empower disadvantaged groups. Critical geographers studying disability have promoted related changes in the discipline: such as making conferences more accessible and promoting research through an international network of scholars (i.e., Disability and Geography International). They have challenged the invisibility of disabled women and men in the geographic literature; emphasizing disablement as a sociospatial form of oppression, exclusionary practices within spaces of academia, and the need to put gendered experiences of disability on Geography's research map (e.g., Chouinard and Grant 1995).

What do critical geographies of disability offer? They offer the possibility of geographic research that empowers persons with disabilities (building on the legacy of radical geography). They promise more balanced accounts of the sociospatial production of disability as a place-specific condition of life, not a natural outcome of bodily impairment. They can help us understand how ableist social relations and practices are sustained over space and time, and how more enabling environments might be created. They can shed light on why resistance to ableism varies among places: of the significance, for example, of national differences in human rights law for disability activism. By providing critical causal explanations of the sociospatial production of disabling environments and of resistance to disablement, these geographies can help us create more enabling societies and spaces.

2. *Research Topics and Directions*

Some studies focus on how disability transforms experiences of personal space and individuals' capacities to negotiate private and public spaces. Others examine the sociospatial production of disabling environments through analyses of the recursive relationship between regulatory institutions (e.g., the state and health care system) and disabled persons' access to and experiences of urban space. Other studies investigate exclusionary attitudes and practices that help to isolate, stigmatize, and segregate disabled persons from others through, for instance, neighborhood opposition to group homes. Underlying much of

this research is a concern with how and why inequalities in power give rise to social spaces that disadvantage and exclude disabled persons. This critical stance toward past and present social orders distinguishes contemporary geographies of disability.

Most studies to date emphasize how ableism shapes disabled persons' lives and life spaces. Butler (1999) examines how the importance of visual cues in lesbian women's negotiations of different life spaces disadvantages visually impaired lesbian women. Valentine (1999) discusses how masculine identity may be threatened by acquired physical impairment. Studies of the links between divisions of power within particular social orders, and sociospatial forms of disablement are relatively rare. Gleeson (1999), however, attempts to show how changing class relations and divisions of labor in the transition from feudal to industrial capitalist societies excerbated the sociospatial marginalization of physically impaired people. In a different vein, Chouinard (1999) suggests that phenomena such as sociospatial barriers to effective political action by disabled women can be understood as outcomes of a patriarchal corporeal class system that systematically devalues disabled persons in general and disabled women in particular, marginalizes and exploits bodies (especially women's) which differ from Western racialized norms, and helps to sustain intolerance toward bodily diversity and regional disparities in resources and well-being.

Contemporary geographies of disability vary in analytic approach and geographic scale. Studies of disablement within personal life spaces emphasize the 'micro' scale of everyday life. They tend to be descriptive and use personal narratives to illustrate changing geographies of daily life. Golledge's work on vision-impaired populations, wayfinding and navigational systems for the vision-impaired (1993) adopts a behavioral conceptualization of disability as an individual condition which can be compensated for through assistive devices and environmental design. Recently, Gollege's work has taken a more critical turn; reflecting on the barriers, such as print, which he and other vision-impaired professors face in academic environments (Golledge 1997).

The differences disability makes in women's abilities to negotiate individual life spaces has been the focus of Dyck's (1995) research on women, chronic illness, and healthcare. Dyck has shown how chronic illnesses such as multiple sclerosis translate into shrinking lifeworlds and sociospatial isolation. Her work examines how professional services, practices, and evaluation procedures, particularly in healthcare, affect chronically ill women's capacities to negotiate spaces of daily life. Her recent research investigates how racial and cultural differences affect women's access to medical knowledge and care, and capacities to negotiate spaces outside the home (Dyck 1997a, 1997b).

Feminist theories of the body have helped to inspire geographic research on cultural codings, readings of and responses to disabled bodies in various life spaces. Butler and Bowlby (1997) examine the disciplining of visually impaired bodies in public spaces through practices such as misinterpreting awkward bodily movements as inappropriate behavior in public space (e.g., intoxication). Complementing geographic studies of experiences of disability are studies of how discriminatory practices within the state and design professions, for instance, perpetuate disabling environments. Some also consider disabled persons' resistance to such practices. An early example is Dear's (1981) analysis of the role of bureaucratic power and practices in the sociospatial segregation of the mentally ill within urban areas. Wolch and Dear's (1993) work on homelessness in the United States and Canada shows how policies of deinstitutionalization, community care, and urban renewal resulted in growing problems of homelessness, the ghettoization of social services and service-dependent populations in central cities, and the loss of affordable housing to gentrification and redevelopment. They examine local geographies of homelessness; for example, the fragile support networks that homeless women establish along daily life paths. Imrie (1996) has examined how urban planning and design practices in Britain and the US have contributed to inaccessible urban built-environments. He argues that architects and planners largely have failed to address the access needs of disabled persons and that attempts to legislate more accessible urban and building designs have had limited success. Historical geographic studies of the development of asylums for the mentally ill and challenged (Philo 1989, Parr and Philo 1996), and contemporary studies of psychiatric survivors' struggles to challenge mental health services that are organizationally and spatially fragmented and unresponsive to needs (Parr 1997), have advanced our understanding of changing geographies of care, discipline, disablement, and resistance in institutional and community settings. Interest in disablement in academic environments is increasing, but has yet to receive systematic attention.

Cultural aspects of sociospatial exclusion are addressed in studies of community attitudes toward persons with disabilities and related geographic phenomena such as NIMBY (Not in My Backyard syndrome). Dear et al. (1997) examine how hierarchically structured attitudes of acceptance and rejection influence levels of community opposition to the siting of group homes in US urban neighborhoods. They find that attitudes toward disabled persons labeled as mentally ill, homeless, or AIDS sufferers are especially negative, and result in stronger desires for spatial separation from and greater community opposition to group homes for these persons.

Exciting directions are emerging in recent critical geographies of disability. Cultural codings of public space as able-bodied and ways of contesting such inscriptions is one emerging theme. Kitchin (1999)

examines British disability activists' use of confrontational tactics, such as chaining themselves to inaccessible public transit vehicles, to draw attention to the denial of disabled citizens' rights to mobility within cities. Stanley (1999) is investigating how the use of sports wheelchairs challenges negative cultural images of disability by associating users with athletic identities and activity spaces. Dorn's (1997) historical geographic study of a nineteenth-century medical practitioner examines medical professionals' roles in defining disability (including the moral causes of conditions such as alcoholism) and shaping healthcare practices in the Western US. Stables and Smith's (1997) study of media portrayals of and public attitudes toward disabled parents who rely on their children for care raises issues of cultural representations of disability within the home. They find that, in Britain, people demonize disabled parents who have child care-givers; regarding this arrangement as exploitative and even unnatural. Their work raises interesting questions about how outsiders and the media read departures from normal (i.e., able-bodied) parental and child care-giving relations in the home. In light of drastic changes in social programs for disabled persons, it is surprising that the impacts of state restructuring on disabled women and men has not received greater attention in the critical geographic literature. Few studies consider the interactive effects of multiple social policy changes on disabled persons' lives and life spaces. Systematic studies are needed which will explain related changes in disabled peoples' status as citizens and dependents of the state, and their capacities to struggle for enabling environments from local to global scale. Disabled peoples' access to recreation spaces is another topic in need of critical geographic research. Methodological and ethical issues in studies of disability and space are another emerging theme Researchers are exploring questions such as how the disabled might be empowered through geographic research, how efforts to make research inclusionary are complicated by the places in which research is conducted (Dyck 1997b), and whether ethical aims such as full disclosure of study objectives should be compromised in situations where this furthers the political objectives of the disabled (Wilton 2000).

3. *Future Challenges and Directions*

Where should geographies of disability go from here? One important challenge is explaining how the uneven development and dynamics of global patriarchal capitalism perpetuates social and spatial exclusion of the disabled. Systematic studies of how sociospatial restructuring of state institutions and programs is shaping the geographic boundaries and conditions of disabled peoples' lives in different places, and how different ways of organizing workplaces and labor processes (e.g., differences in accommodating special needs) affect disabled persons' access to paid employment, would help to address such challenges. Geographic variations in the design, production, availability, and consumption of disability aids, and their impacts on disabled peoples' abilities to negotiate life spaces is another topic for future research. Studies of the sociospatial production of disabling conditions; examining, for instance, how unsafe, stressful conditions of work and the hyperexploitation of female labor in places such as export zones influence the incidence and severity of different disabilities, would also help to address this challenge.

Another challenge is understanding the differences that cultural practices and images make in disabled persons' lives. Do negative images of disabled parents restrict their own or child care-givers' access to spaces outside the home? How do conditions of life in particular places influence ideas about where disabled people do and don't belong? Are urban neighborhoods with a history of disability activism less likely to view the disabled as nonproductive and hence out of place in local labor markets? Future studies could explore geographic variations in the political culture of disability struggles. Why do activists in some places adopt radical, confrontational strategies demanding collective rights, while others favor actions such as sponsoring stage appearances by famous disabled 'heroes'; events which reinforce cultural notions that only exceptional disabled persons succeed?

Theoretical challenges include developing dynamic, deeply geographic conceptions of the regulation of bodily differences in space; conceptions which allow for multiple, diverse bases and processes of oppression as disabled women and men move through space and time. How do shifting bases of exclusion, for instance, sexist discrimination within the disability movement and ableism, homophobia, and class differences within the women's movement, constrain disabled women's abilities to promote collective action on issues such as reproductive rights? How important are such processes of exclusion in the lives of disabled women in the Third World; where inabilities to walk to distant paid employment, to farm, or to care for family members often translates into dependency and isolation within the home? We have only just begun to tap the feminist theories of bodily oppression that can help us to understand such geographic processes of disablement.

Methodological challenges include gathering data on neglected aspects of disability such as spatial variations in employment and accommodation practices in workplaces, and developing accessible research formats. Ethical and political challenges include ensuring that research empowers disabled persons.

Perhaps the key challenge is promoting restless geographic imaginations about disabling environments. Such imaginations will be open to surprise about patterns and processes of disablement over space and time, and willing to learn from disabled peoples' knowledges. With creativity and commit-

ment, future geographies of disability can help to create environments that enable all members of society.

Bibliography

Butler R 1999 Double the trouble or twice the fun? Disabled bodies in the gay community. In: Butler R, Parr H (eds.) *Mind and Body Space: Geographies of Illness, Impairment, and Disability*. Routledge, London, pp. 203–20

Butler R, Bowlby S 1997 Bodies and spaces: An exploration of disabled people's experiences of public space. *Environment and Planning D: Society and Space* **15**: 411–33

Butler R, Parr H (eds.) 1997 *Mind and Body Space: Geographies of Illness, Impairment, and Disability*, Routledge, London

Chouinard V 1999 Body politics: Disabled women's activism in Canada and beyond. In: Butler R, Parr H (eds.) *Mind and Body Space: Geographies of Illness, Impairment, and Disability*. Routledge, London, pp. 269–294

Chouinard V, Grant A 1995 On being not even anywhere near the project: ways of putting ourselves in the picture. *Antipode*. **27**: 137–66

Dear M 1981 Social and spatial reproduction of the mentally ill. In: Dear M, Scott A J (eds.) *Urbanization and Urban Planning in Capitalist Society*. Methuen, London, pp. 481–97

Dear M, Gaber S L, Takahashi L, Wilton R 1997 Seeing people differently: The sociospatial construction of disability. *Environment and Planning D: Society and Space* **15**(4): 455–80

Dorn M 1997 The moral topography of intemperance: Dr. Daniel Drake on the health and habits of the American westerner, 1821–52. Paper presented at the annual meeting of the Institute of British Geographers, Exeter, UK

Dyck I 1995 Hidden geographies: the changing life-worlds of women with multiple sclerosis. *Social Science and Medicine* **40**: 307–20

Dyck I 1997a Whose body? Whose voice? Contradictory cultural constructions in the politics of research. *Atlantis* **21**: 54–62

Dyck I 1997b Dialogue with difference: A tale of two studies. In: Jones J P, III, Nast H J, Roberts S M (eds.) *Thresholds in Feminist Geography: Difference, Methodology, Representation*. Rowman and Littlefield, pp. 183–02

Gleeson B 1999 *Geographies of disability*. Routledge, London and New York

Golledge R 1993 Geography and the disabled: a survey with special reference to vision impaired and blind populations. *Transactions: Institute of British Geographers* **19**: 63–85

Golledge R 1997 On reassembling one's life: Overcoming disability in an academic environment. *Environment and Planning D: Society and Space* **15**(4): 391–11

Imrie R 1996 *Disability and the City*. St. Martin's Press, New York

Kitchin R 1999 'To boldly go where others have been before': Spatialities of disabled people's resistance. Paper presented at the annual meeting of the Association of American Geographers, Honolulu, Hawaii

Morris J 1991 *Pride against Prejudice: Transforming Attitudes to Disability*. Women's Press, London

Oliver M 1990 *The Politics of Disablement*. Macmillan, London

Parr H 1997 Mental health, public space, and the city: Questions of individual and collective access. *Environment and Planning D: Society and Space* : 435–54

Parr H, Philo C 1996 A forbidding fortress of locks, bars and padded cells? The locational history of Nottingham's mental health care. *Historical Geography Research Series* No. 32

Philo C 1989 'Enough to drive one mad': The organization of space in 19th-century lunatic asylums. In: Wolch J, Dear M (eds.) *The Power of Geography: How Territory Shapes Social Life*. Unwin Hyman, Boston, pp. 258–90

Stables J, Smith F 1997 Caught in the Cinderella trap: Narratives of disabled parents and young carers. In: Butler R, Parr H (eds.) *Mind and Body Space: Geographies of Illness, Impairment, and Disability*, Routledge, pp. 256–68

Stanley C 1999 Disability and Assistive Technology: The Sportiness of Impaired Bodies. Paper presented at the annual meeting of the Association of American Geographers, Honolulu, Hawaii

Valentine G 1999 What it means to be a man: the body, masculinities, disability. In: Butler R, Parr H (eds.) *Mind and Body Spaces: Geographies of Illness, Impairment and Disability*. Routledge, London, pp. 167–80

Wilton R D 2000 Too close for comfort: Exploring the politics of transgression. *Ethics, Place and Culture*

Wolch J, Dear M 1993 *Malign Neglect: Homelessness in an American City*. Jossey-Bass, San Francisco

V. Chouinard

Disability: Psychological and Social Aspects

Disability can be conceptualized as activity limitations resulting from pathological and biomedical factors such as inflammation of the joint, brain lesion, age, or sensory impairment. However, there is evidence that psychological and social factors can predict and possibly influence disability unexplained by underlying pathophysiology. In addition to the definition, prevalence, and measurement of disability, this article examines explanatory models, psychological and social predictors and consequences of disability, as well as the possibility of reducing disability and its effects through psychological interventions.

1. The Term 'Disability'

Disability is defined by the World Health Organization (WHO) in their International Classification of Impairment, Disability and Handicap (ICIDH) model as 'any restriction or lack of ability to perform an activity in the manner or within the range considered normal for a human being' (WHO 1980). The ICIDH model proposes that disability is the result of impairment and that disability may result in handicap. Impairment is 'any loss, abnormality or failure of psychological, physiological, or anatomical structure or function' and is the direct result of an underlying disease or disorder. Handicap is 'a disadvantage resulting from impairment or disability, that limits or prevents the fulfillment of a role that is normal for that individual' (WHO 1980, pp. 27–9). So, for example, an individual who experienced a stroke (disorder) might have

restricted leg functioning (impairment), resulting in limitations in mobility (disability), and therefore be unable to continue to work (handicap).

However, *disability* is a widely used term both in everyday speech and by individuals in academic and applied disciplines, in addition to medicine, including psychology, sociology, physiotherapy, occupational therapy, health economics, and health services research. The common use of the construct suggests that there are varying underlying representations of the content and cause of disability (Johnston 1997). The term 'disability' has been used to refer to the WHO concepts of either impairment or handicap or a combination of disability and handicap. In some countries, the term disability may simply refer to the conditions that attract welfare support so that an assessment of 100 percent disability would mean that someone is entitled to the maximum benefit rather than being totally disabled.

2. *Prevalence of Disability*

Estimates of the prevalence of disability vary according to the definition or types of disability as well as the threshold of severity that is chosen for inclusion in a study. In the UK, the Office of Population Censuses and Surveys (OPCS) 1988 survey of disability found that 14 percent of adults living in private households have at least one disability ranging from slight to very severe (Martin et al. 1988). Disabling diseases vary in their effects on life expectancy and this also has implications for considering incidence and prevalence data. The most common disabilities in the 1988 survey were locomotion followed by hearing and personal care (for example, dressing or toileting). The most commonly cited causes of disability were musculoskeletal complaints (e.g., arthritis), ear complaints, eye complaints, diseases of the circulatory system, and nervous system complaints (e.g., stroke). Disabilities may arise from genetic disorders, from birth injuries, as a result of accidents, as well as due to disease. Martin et al. (1988) found the overall rate of disability rose with age, with almost 70 percent of disabled adults aged over 60. Of the most severely disabled adults, 64 percent were aged 70 or over. There were more disabled women then men. Even allowing for differing numbers of men and women in the population, the rate of disability among those aged 75 or over was still higher for women than for men, indicating that elderly women were more likely to be disabled than elderly men. Martin et al. (1988) found no evidence of ethnic differences in prevalence of disability.

3. *How Disability is Measured*

A disability rating serves as a key indicator of care requirements, ability to remain independent in the community, and the nature and extent of participation in society. Disability is typically measured based on self-report or observers' ratings of behavior, frequently the level of performance of basic and instrumental activities of daily living (ADL), such as eating, toileting, dressing, bathing, walking, doing housework, and shopping. However, the selection of items, the method of assessing performance, and the scaling of items to give an overall score, show immense diversity. Most measures involve self-report of either what one 'can do' or of what one has actually done. Some assessments require observing an activity to rate the 'normality' of performance. Others concentrate on whether the activity was achieved, rather than the manner of its accomplishment. Measures that elicit performance of movements assume that the disability is a constant, inherent aspect of the individual that can reliably be produced on demand. By contrast, measures of achievement of ADL use reports or performance under ordinary rather than assessment conditions. Recently, mobility limitations have been directly assessed using ambulatory monitoring techniques.

There have been considerable problems in identifying appropriate items and methods of scoring and scaling and variability in the rigor of measures used. Most measures use additive scaling, indicating a concept of equivalence of activities. In these measures, failure on feeding, self-care, or mobility would contribute in a similar manner to the overall score. Other measures use weighted additive scores, where failure on some items counts for more than failure on other items. For example, in the Sickness Impact Profile (Bergner et al. 1981), difficulties in walking a long distance count for less than difficulties in walking a short distance. In developing the OPCS scales, Martin et al. (1988) used a Thurstone scaling approach to create a set of scales for each area of disability, such as locomotion, dexterity, and personal care. For each scale, the value of each item was determined by judges' ratings. The method of combining disabilities was also empirically derived to reflect the judgment of expert raters. It is heavily weighted by the individual's most severe disabilities. This contrasts with averaging or additive measures which carry the implication that many mild disabilities, such as walking more slowly, being unable to walk half a mile, and being unable to climb stairs, may be equivalent to one severe disability; for example, being unable to walk at all. This diversity in the measurement of disability reflects the wide variation in views about the nature of disability.

4. *Explanatory Models*

The three main explanatory models of the nature of disability are the medical model, the social model, and the psychological model. The ICIDH model is a *medical model* since it conceptualizes disability within a predominantly pathological framework, explaining disability as a consequence of an underlying disease or

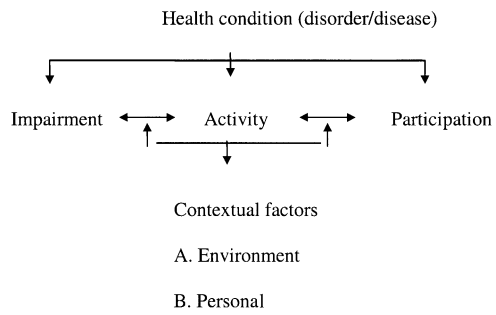

Health condition

Any alteration or attribute of health status which may lead to distress, interference with daily activities, or contact with health services, e.g., disease (acute or chronic), disorder, injury, trauma, pregnancy, aging, stress, congenital anomaly, or genetic predisposition.

Impairment

Loss/abnormality of body structure or of physiological or psychological function.

Activity

Nature and extent of functioning at the level of the person. Activity limitation (formerly disability) is the limitation in the performance of the activity that derives totally from the person. Limitations can be in duration and quality.

Participation

Nature and extent of a person's involvement in life situations in relationship to impairment, activities, health conditions, and contextual factors.

Contextual factors

This is the complete background to the person's life and living. *Environmental factors* are composed of components of the natural environment (weather or terrain), the human-made environment (tools, furnishing, the built environment, social attitudes, customs, rules, institutions, and other individuals). *Personal factors* are composed of features of the person that are not parts of a health condition or disablement, including age, race, gender, education, personality, character style, experiences, aptitudes, other health conditions, fitness, lifestyle, habits, upbringing, social background, profession, and past and current experiences.

Figure 1
International Classification of Impairments, Disabilities and Handicaps (ICIDH-2; WHO 1998)

disorder. However, a *social model* of disability conceptualizes disability in terms of limitations placed on the individual due to constraints in the social and physical environment and proposes that functional limitations are determined as much socially as biologically, for example, by the presence, attitudes, and actions of others (e.g., Oliver 1993). *Psychological models* conceptualize disability as a behavioral construct (e.g., Johnston 1996). These models suggest that disability is determined by psychological factors that influence behavior, for example, the individual's beliefs, emotions, skills, and habits.

Medical, social, and psychological models are not necessarily in conflict. Although some level of disability experienced by an individual may be related to their degree of impairment, there is considerable evidence that much of the variance in disability experienced by an individual remains unexplained by underlying pathology or impairment (Johnston and Pollard 2001). For example, McFarlane and Brooks (1988) found that in people with rheumatoid arthritis, psychological factors predicted more variance in ADL three years later than impairment and disease severity measures. Cognitive and emotional manipulations have also influenced the level of disability experienced by individuals under conditions where impairment was not affected. Additionally, the major task of rehabilitation therapists is to enable patients to overcome disabilities, often without resolution of underlying impairments. Moreover, many social factors are likely to operate via psychological mechanisms to alter the relationship between the components of the ICIDH model; for example, an environment with barriers to mobility or an overprotective social environment may reduce the individual's confidence with the result that their activities become even more limited. An explanatory model of disability which combines medical, psychological, and social factors would appear to be needed.

Indeed, the ICIDH model was recently modified to allow for the possibility that disability may be influenced by factors other than structure and function (WHO 1998). The revisions added complexity by proposing reverse causality, so that impairment may both result in, and result from, limitations in activity which may in turn determine, or be determined by, lack of social participation. It also introduced the possibility that contextual factors, both environmental and personal, can explain the level of disability experienced by an individual. In addition, the use of the terms 'disability' and 'handicap' were dropped in order to 'avoid deprecation, stigmatization and undue connotations' (WHO 1998, p. 19). Disability and handicap were redefined and renamed activity and participation. Activity is defined as the nature and extent of functioning at the level of the person and disability as the activity limitations of the individual person, whereas participation is defined in relation to the social context as the 'nature and extent of a person's involvement in life situations in relationship to impairment, activities, health conditions, and contextual factors.' This amended model of the consequences of disease is depicted in Fig. 1.

While consistent with previous research which has shown a meager relationship between impairment and disability, and that changes in disability have occurred without concurrent changes in impairment, the ICIDH (WHO 1998) model is nevertheless vague about exactly which psychological and social factors contribute to explaining, and thereby predicting, disability.

5. *What Predicts Disability?*

Many studies, including investigations where the effects of impairment have been controlled statistically, and randomized controlled trials where impairment was equally distributed in the control and intervention groups, have shown disability to be predicted and influenced by emotional, cognitive, behavioral, and social factors.

5.1 *Emotional Factors*

How people feel may influence their level of disability. Patients' emotional state is determined partly by their disposition and partly by situational factors such as the threats associated with their condition and treatment. Psychologists have often noted the association between disability and low mood or emotional disorder, and this is usually interpreted as emotional consequences of disability. However, most studies which examine the influence of emotional factors on disability are cross-sectional and observational in design, and it is therefore difficult to disentangle cause and effect. It seems likely that complex feedback loops occur. For example, while disability results in depression, depressed individuals are likely to be more disabled given comparable levels of impairment. Since emotional states are accompanied by physiological changes in the neuroendocrine, immune, and cardiovascular systems, it is also possible that negative emotions may influence disability by exacerbating impairments.

Nevertheless, there is evidence from longitudinal, prospective studies that emotional states, such as anxiety or depression, may influence the level of disability itself. For example, in a longitudinal study of middle-aged and older community residents, Meeks et al. (2000) found that depression predicted functional disability more strongly than disability predicted depression. In addition, changing emotions can have an immediate effect on disability levels. Fisher and Johnston (see Johnston 1997) used mood induction techniques to either increase or decrease anxiety (by asking about good or upsetting events) in two ran-

domly allocated groups of patients with chronic pain. Patients' disability was assessed by a lifting task before and after the experimental procedure, and results demonstrated that mood enhancement was associated with reduced levels of disability, while mood depression resulted in greater disability. It appears that emotional state can influence the level of disability independently of impairment.

5.2 *Cognitive Factors: Attitudes and Beliefs*

What people think or believe may influence their level of disability, particularly their beliefs about how much control they have over behaviors assessed as disability. Control cognitions are part of many theoretical frameworks constructed to predict and explain behavior. Two control cognitions, perceived personal control and self-efficacy, have been particularly associated with disability (see *Control Beliefs: Health Perspectives*).

Perceived personal control or internal locus of control refers to the degree to which persons expect that an outcome is contingent upon their own behavior. It can be contrasted with beliefs in external control, i.e., the degree to which persons expect that the outcome is not under their control, but under the control of powerful others, or is unpredictable. The evidence suggests that patients with high perceived personal control beliefs have better outcomes. For example, patients suffering from a stroke or wrist fracture were found to have greater recovery of function if they had higher beliefs in personal control over their recovery, even allowing for initial levels of disability—a finding which has been replicated in a larger sample of stroke patients six months after discharge from hospital (see Johnston et al. 1999). Härkäpää et al. (1996) found that patients in a back pain program showed greater improvement in functional activity if they had weaker beliefs in control by powerful others. In an experimental study, Fisher and Johnston (1996) manipulated the control cognitions of chronic pain patients by asking them to recall occasions when they had achieved or failed to retain control. They found that increasing perceived control resulted in reduced disability, assessed by the observed performance on a lifting task before and after the manipulation, while decreasing perceived control resulted in increased disability. Given that these were immediate effects, it is clear that the results were not due to changes in impairment, suggesting that control cognitions may play a causal role in activity limitations.

Self-efficacy relates to individuals' confidence in their ability to perform critical behaviors (see *Self-efficacy and Health*). Bandura (1997) proposes that there are four sources of information which are used as a basis for self-efficacy judgments: performance attainments, vicarious experience, verbal persuasion, and physiological state. It has been shown that successful performance leads to increased self-efficacy, whereas repeated failures result in lower self-efficacy. People with high self-efficacy beliefs are more likely to perform the behavior. Kaplan et al. (1984) examined exercise maintenance and found that changes in self-efficacy were associated with changes in walking activity in patients with chronic obstructive pulmonary disease. Changes in self-efficacy have also been associated with changes in disability among participants in arthritis self-management courses (Lorig et al. 1989), in low back pain patients, in patients with fibromyalgia, and in coronary bypass patients (see *Coronary Heart Disease (CHD), Coping with*).

5.3 *Coping Behavior*

What people do in order to cope with the stress imposed by their condition and the practical problems which arise from physical disability may affect their level of disability. For example, adherence to medication, diet, and rehabilitation regimens all may impact on disability. Coping refers to the procedures adopted to minimize stress and may or may not constitute effective management. Coping responses can be problem-focused or emotion-focused. Problem-focused coping includes attempts to modify or eliminate the sources of stress through one's own behavior, i.e., solving the problem. Emotion-focused coping includes attempts to manage the emotional consequences of stressors and to maintain emotional equilibrium. A coping model that has frequently been applied in studies of patients with chronic disease is Leventhal's Self-Regulation Model (see Brownlee et al. 2000). According to this model, the individual develops mental representations of their impairments which in turn determines what they do to cope (see *Stressful Medical Procedures*, *Coping with*). Dominant mental representations which determine coping behavior are: (a) identity of the condition: what they think it is; (b) timeline: the pattern over time (e.g., acute, chronic, or cyclical/remitting); (c) cause: what caused their condition (e.g., if it was due to their own behavior or character); (d) control/cure: if they perceive their condition can be controlled by their own or health professionals' actions; and (e) consequences: what they believe they have experienced as a result of their condition and what they expect the outcome to be. These mental representations are proposed to influence coping with both the objective condition and the emotional reaction to the condition.

However, while some predictive studies have found coping associated with activity limitations (e.g., Evers et al. 1998), other studies have not. For example, Revenson and Felton (1989) found that coping did not predict change in disability in a six-month prospective study of patients with rheumatoid arthritis. Similarly, others have found that coping did not predict and

therefore did not influence disability in people with chronic fatigue. It is possible that these findings are specific to these clinical conditions, where there is no clear action an individual can take to affect outcomes. Nevertheless, it is also possible that coping strategies are the result rather than the cause of disability.

5.4 Social Factors

Disability may be influenced by social factors such as stigmatizing reactions from others, overprotection by immediate family members, or carers' social reinforcement of activity limitations. Evidence of social influences comes from comparisons of disability assessments by different health professionals. For example, nurses rated patients as more disabled than rehabilitation therapists over a 24-hour period (Johnston 1996). While differences in ratings are obviously not due to changes in impairment, they may be due to the different social demands on the patients in the two care contexts: nurses assist patients in achieving tasks, while rehabilitation therapists enable patients to perform the tasks themselves.

Disability may also be influenced by the level and type of formal or informal support available (see *Social Support and Recovery from Disease and Medical Procedures*). While most informal support is provided by spouses—primarily for household assistance (Melzer et al. 1999), a social support network includes all members of the individual's household, family and friends living nearby, as well as more distant contacts. In patients with rheumatoid arthritis, a small network soon after diagnosis predicted a greater decrease in mobility over the following year (Evers et al. 1998). The support can be *practical*, e.g., tangible help and assistance with information, or *emotional*, e.g., that makes the individual feel esteemed and valued. Both have been related to health outcomes and behavior. For example, Weinberger et al. (1986) found that biweekly telephone calls to people with osteoarthritis led to improved functional status six months later and attributed the gains to the effects of enhanced emotional social support.

6. Consequences of Disability

Disability has a wide range of social and emotional consequences. For example, disability can impact on the person's social functioning. Newsom and Schulz (1996) report that older people with impairments have fewer family and friendship contacts and less perceived belonging and perceived tangible aid. Johnston and Pollard (2001) found that disability predicted social participation (handicap) in stroke patients, with greater disability associated with lower social participation.

Disability also has consequences for emotional functioning (e.g., Meeks et al. 2000). Perhaps the most frequently used models in psychological studies of disability explore the impact of disability on emotional functioning. These models include: (a) mental health models where rates of disorder are assessed; (b) life-event models where the emotional response to the life event precipitating in the impairment/disability is examined; and (c) stress models where the disability is seen as a stressor which elicits the strain evident as high levels of distress.

The nature of the underlying pathology, for example, whether a disease has static, fluctuating, or deteriorating patterns, is likely to temper the consequences of disability. Age at onset is also likely to be crucial, with earlier onset affecting education, family, housing, work, and leisure.

There is also evidence that disability has consequences for the people supporting the disabled person, impacting on their social and emotional functioning, as well as their health (see the review by Shulz and Quittner 1998). The effect of disability on caregivers may have further implications for the health outcomes of patients. For example, Elliott et al. (1999) found that the risk of disability and secondary complications increases for care recipients when carers have difficulty in adjusting to their new role demands.

The impact of disability is found to relate more closely to cognitive appraisals of functional limitations and coping resources than to objective indices of impairment and disability. For example, Schulz et al. (1994) proposed a model of the impact of impairment and disability, suggesting that negative affect is primarily due to the perceived loss of control rather then directly from impairment and disability.

7. Interventions

Interventions can be directed at reducing the disability or the impact of disability or both. Interventions directed at reducing the impact are similar to other cognitive behavioral programs that aim to reduce emotional and social consequences of life-events and difficulties: they concentrate on cognitive reframing and stress-management skills. Given the evidence that the individual's perception of limitations, social support, and control appear to be predictors of emotional and quality-of-life outcomes, there is clearly potential to improve these outcomes without reducing impairment.

Interventions directed at reducing disability *per se* are more akin to other programs that aim to achieve behavioral change. These include stress management techniques or elements designed to enhance self-efficacy in order to increase motivation and to develop skills and habits required to achieve the performance of activities. Kaplan et al. (1984) showed improvements in disability for patients with chronic obstructive pulmonary disease following cognitive and behavioral exercise programs. Lorig and col-

leagues have shown similar benefits from arthritis self-management programs designed to enhance self-efficacy (Lorig et al. 1989). Behavioral programs in which social attention is contingent on successful performance rather than on limitations have also been shown to decrease disability, especially in chronic pain patients.

See also: Chronic Illness: Quality of Life; Control Beliefs: Health Perspectives; Coronary Heart Disease (CHD), Coping with; Disability, Demography of; Disability, Geography of; Disability: Sociological Aspects; Law and People with Disabilities; Self-efficacy and Health; Social Support and Recovery from Disease and Medical Procedures; Stressful Medical Procedures, Coping with

Bibliography

Bandura A 1997 *Self-Efficacy: The Exercise of Control*. Freeman, New York

Bergner M, Bobbitt R A, Pollard W A 1981 The sickness impact profile: Development and final revision of a health status measure. *Medical Care* **19**: 787–805

Brownlee S, Leventhal H, Leventhal E A 2000 Regulation, self-regulation, and construction of the self in the maintenance of physical health. In: Boekaerts M, Pintrich P R, Zeider M (eds.) *Handbook of Self-Regulation*. Academic Press, San Diego, CA

Elliot T, Shewchuk R, Richards J S 1999 Caregiver social problem solving abilities and family member adjustment to recent-onset physical disability. *Rehabilitation Psychology* **44**: 104–23

Evers A W, Kraaimaat F W, Geenen R, Bijlsma J W 1998 Psychosocial predictors of functional change in recently diagnosed rheumatoid arthritis patients. *Behaviour Research and Therapy* **36**: 179–93

Fisher K, Johnston M 1996 Experimental manipulation of perceived control and its effect on disability. *Psychology and Health* **11**: 657–69

Härkäpää K, Järvikoski A, Estlander A 1996 Health optimism and control beliefs as predictors for treatment outcome of a multimodal back treatment program. *Psychology and Health* **12**: 123–34

Johnston M 1996 Models of disability. *The Psychologist* **9**: 205–10

Johnston M 1997 Representations of disability. In: Petrie K J, Weinman J A (eds.) *Perceptions of Health and Illness*. Harwood Academic Publishers, London

Johnston M, Morrison V, MacWalter R S, Partridge C J 1999 Perceived control, coping and recovery from disability following stroke. *Psychology and Health* **14**: 181–92

Johnston M, Pollard B 2001 Problems with the sickness impact profile. *Social Science and Medicine* **52**: 921–34

Kaplan R M, Atkins C J, Reinsch S 1984 Specific efficacy expectations mediate exercise compliance in patients with COPD. *Health Psychology* **3**: 223–42

Lorig K, Chastain R, Ung E, Shoor S, Holman H R 1989 Development and evaluation of a scale to measure the perceived self-efficacy of people with arthritis. *Arthritis and Rheumatism* **32**(3): 7–44

McFarlane A C, Brooks P M 1988 Determinants of disability in rheumatoid arthritis. *British Journal of Rheumatology* **27**: 7–14

Martin J, Meltzer H, Elliot D 1988 *The Prevalence of Disability Among Adults*. HMSO, London

Meeks S, Murrell S, Mehl R C 2000 Longitudinal relationships between depressive symptoms and health in normal older and middle-aged adults. *Psychology and Aging* **15**(1): 100–9

Melzer D, McWilliams B, Brayne C, Johnson T, Bond J 1999 Profile of disability in elderly people: Estimates from a longitudinal population study. *British Medical Journal* **318**: 1108–11

Newsom J T, Schulz R 1996 Social support as a mediator in the relation between functional status and quality of life in older adults. *Psychology and Aging* **11**(1): 34–44

Oliver M 1993 Redefining disability: A challenge to research. In: Swain J, Finkelstein V, French S, Oliver M (eds.) *Disabling Barriers—Enabling Environments*. Sage, London

Revenson T A, Felton B J 1989 Disability and coping as predictors of psychological adjustment to rheumatoid arthritis. *Journal of Consulting and Clinical Psychology* **57**: 344–8

Schulz R, Heckhausen J, O'Brien A T 1994 Control and the disablement process in the elderly. *Journal of Social Behaviour and Personality* **9**(5): 139–52

Schulz R, Quittner A L 1998 Caregiving for children and adults with chronic conditions: Introduction to the special issue. *Health Psychology* **17**: 107–11

Weinberger M, Hiner S L, Tierney W M 1986 Improving functional status in arthritis: The effect of social support. *Social Science and Medicine* **23**(9): 899–904

World Health Organization 1980 *The International Classification of Impairments, Disabilities and Handicaps*. World Health Organization, Geneva

World Health Organization 1998 *The International Classification of Impairments, Activities and Participation: A Manual of the Dimensions of Disablement and Health*. www.who.int/msa/mnh/ems/icidh/introduction.htm

M. Johnston and D. Bonetti

Disability: Sociological Aspects

Disability has captured the public's attention globally, stimulated sociological inquiry, and emerged as an urgent priority for policy makers. This is due to the increased prevalence of functional limitations in society as a result of an aging population, impairments caused by war and landmines, technological innovations that prolong the lives of persons with disabling chronic illnesses, unhealthy lifestyles, and broadened definitions of disability.

The social and economic costs of disability continue to rise as exemplified in contentious national debates in most industrialized countries over the solvency and future of social welfare systems. The uneven dis-

tribution of the global burden of disability within and between countries is recognized by the World Bank and the World Health Organization as a major threat to development in the Third World and a cause of inequality in developed countries (Murray and Lopez 1996). The Institute of Medicine (1997) has emphasized that control of disability-generating infectious and chronic diseases is of vital concern to economic stability, social justice, and world peace. The salience of the problem is highlighted by the growing visibility of disabled persons in public spaces as a result of legislation such as the Americans With Disabilities Act of 1990 (ADA) and intense media coverage of wars, landmine accidents, unhealthy working conditions, and natural disasters in countries around the world. Increasingly accessible environments, personal experience with family members, co-workers, and friends who experience the impact of disability on their lives and the public activism of the disability community also contribute to a public awareness of disability.

1. *Definitions, Concepts, and Meanings of Disability*

Disability is defined by a person's difficulty in performing a task or activity due to an underlying physical or mental dysfunction. However, the specific features of a particular disability are varied, and the ways of identifying and evaluating them are diverse as well. While the disability concept has gained general acceptance by researchers, policy makers, and the public, specific applications and interpretations of the concept vary considerably depending on ideology, cultural context, and intended use (Albrecht and Verbrugge 1999).

Early definitions of disability focused on abnormality, deviance, and inability of individuals to fit into society or perform their expected roles. Disability was a defining master status often coincident with being poor, marginal, sick, and jobless. As modern medicine emerged in the nineteenth century to exert power over the definitions of health and illness, disability was conceived as a personal characteristic residing inside the individual due to some organic abnormality resulting from genetics, disease, or injury. According to this medical model, acceptable disabilities were spinal cord injury, blindness, deafness, schizophrenia, and severe depression. Increasing dissatisfaction with a strictly medical view of disability occurred as people increasingly became aware that disability often resulted from, or was compounded by, physical and social isolation. New definitions consequently emphasize the power of forces external to the individual in defining disability and shaping disabled peoples' lives (Albrecht 1992, Usten 1997).

Regardless of conceptual preference, real-world programs and policies operate on predominantly medical terms. Medical professionals concentrate on the organic base of disability in their diagnoses and treatment, assessing disability in terms of functional limitations such as independence in activities of daily living such as walking, bathing, working, and thinking. Governments generally base social welfare programs on medical definitions of disability, but they modify their definitions depending on the intent of the specific program. For example, the definition of disability embodied in the ADA emphasizes impairments that limit a person's ability to perform activities of daily living while the US Internal Revenue code stresses impairments which limit the individual's ability to work. In sum, the US government has 49 separate definitions of disability. This pattern is typical in other nations as well.

2. *Distribution and Causes of Disability*

Recent worldwide assessments indicate that, depending on how disability is measured, about 12–15 percent of the world's population is disabled. This means, with an estimated world population of six billion people in the year 2000, approximately 846 million were disabled. This disability burden is experienced unequally between less developed and more developed countries. In the year 2000, for example, 4.82 billion people (80.4 percent of the world's population) lived in the less developed countries of Asia, India, China, Africa, Latin America, the Middle East, and the Caribbean and Pacific Islands, but included 89.3 percent (755 million) of the world's disabled people. By contrast, the more developed countries of North America, Europe, and Australia, comprising 19.6 percent of the world's population, had only 11.7 percent (91 million) of the world's disabled people. The impact of this unequal distribution of disability on the health and development of less developed nations is amplified by the fact that only 10 percent of the health care dollars in the world are spent in less developed countries, where almost 90 percent of the disability is found.

Disability differs in type and distribution within and between countries (Murray and Lopez 1996). In developed countries, the major causes of disability are ischemic heart disease, cerebrovascular disease, and physical inactivity. In poorer nations, nearly half of the disability is due to infectious and childhood diseases and injuries. Disability is increasingly exported from developed to Third World nations in the form of high-risk work such as that found in chemical plants and the low-cost, high-volume labor businesses such as garment making and athletic shoe assembly. Across the entire world, the leading causes of disability are respiratory infections, diarrheal diseases, and conditions arising in the perinatal period. In the year 2020, they are projected to be cerebral vascular disease, depression, and road traffic accidents. In all circumstances, disability is disproportionately experienced

by the poor and disenfranchised, but will be experienced increasingly by those more privileged as they live longer with chronic diseases.

3. Models for Analyzing Disability

There are three major theoretical models used to analyze disability: (a) the disability as deviance model, (b) the medical model, and (c) the social model. The disability as deviance model arose from an effort to differentiate between who was normal and therefore accepted in society, and who was abnormal and to be treated as different. This distinction underlies current analyses of inclusion and exclusion in society. The distinction is historically based on appearance, gender, race, ethnicity, age, and nationality, as well as the moral stigma attached to the condition (Stiker 1982). This theoretical approach underscores that disability is a physical and/or mental condition interpreted in a historical, social, and moral context.

The medical model gives power to the physician whose task it is to diagnose and treat disability based on organic abnormalities and function identified through scientific procedures. According to this model, disability is regarded as a medical problem residing in the individual and best tackled through medical treatment. The social model of disability switches emphasis from the physician to disabled persons, and from individuals to the social and physical environment in which they live (Shakespeare 1998). Much of the work on the social model of disability was generated by disabled people themselves, who took an activist stance in embracing a minority group approach to disability (Barnes 1991). They de-emphasize the significance of impairments such as paralysis and low vision, and focus instead on socially constructed barriers such as exclusion and discrimination based on ideology and social values.

Disability studies is an emerging field of study based on the social model of disability (Albrecht et al. 2001). It is being developed by a multidisciplinary group of researchers, academics, government officials, and activists intent on shifting emphasis away from a prevention–treatment–rehabilitation paradigm to a social–cultural–political paradigm. This refocusing effort underscores the agency of disabled persons in defining and shaping their own environment by becoming involved in research, political action, and policy making. The intent is to redefine traditional power relationships between disabled persons and other stakeholders.

4. Disability as a Social and Cultural Product

Recent attention has been given to disability as a symbol and vehicle in art, literature, and film. The representation of disability reflects social and cultural values and presents images that shape public opinion. The key questions are: what does disability signify? and what does disability tell us about a culture or group? An analysis of art and literature reveals that disabled people are often portrayed across history as cripples, lunatics, mendicants, and misfits symbolizing their deviancy and marginality. Yet, when Goya painted maimed veterans, he pointed to the heroism of soldiers in combat and devastation of war, questioning the cost of sovereign power. Likewise, the many Vietnam movies such as *The Deerhunter* use disability to symbolize the divisions in society over ambiguous wars and conflicting values.

On inspection, great literature is filled with disabled figures: Shakespeare's mad King Lear, Hugo's Hunchback of Notre Dame, and the Japanese novelist Oe's autistic son. These characters tell us about the conception of disabilities in diverse societies, but also suggest universal themes warranting further investigation. Disability is the consequence of foolish or heroic action. Disability is a lonely yet public and community experience. Disabled people are portrayed in the roles of fools or as impotent, yet they have produced some of the world's great music and art. For example, Beethoven wrote his Ninth Symphony while deaf, and Monet painted his landscapes at Giverny while losing his vision. Disability is an acknowledged part of most cultures, and helps to define the meaning of life, value of the individual, and nature of society.

5. Societal Responses to Disability

The values of a group can be ascertained by the way the community treats its women, children, elderly, and disabled members. As the populations of Western nations become older, and disability is recognized as a major problem inhibiting development in Third World countries, governments are concentrating increased resources on the prevention and management of disability. In many countries of the world, disabled people have been excluded from public affairs by physical barriers which prevent their mobility, and social discrimination which makes school and work difficult at best. For much of the world, out of sight has meant out of mind. Disabled people are often cared for in their homes by their families, or they live on the street as beggars and homeless (Castel 1995). Western European countries have a strong community ethos and social welfare tradition, so disability is addressed through strong social welfare programs focused on keeping people in the community. While there are well-established disability benefits programs in the USA, the emphasis is aimed more at the individual, and institutionalization is a more frequently invoked solution. In either case, disability is becoming an enormously expensive enterprise, testing

the values and taxing the resources of many nations. Conservatives argue for limiting benefits and sharply controlling the numbers of people on the disability rolls while liberals maintain that government has a direct obligation to see that disabled people have the minimum resources to live their lives empowered and in dignity. Funding for these programs will necessarily come at the expense of the military and other social programs or through increased taxes. Ultimately, the inclusion or exclusion of disabled people in society will depend on available resources, national values, and the political activism of disabled people.

6. *Future Trends*

Because of the tremendous cost of providing benefits to an aging population, increased efforts will be made to prevent and better manage disability. Citizens will be encouraged to live with disability in their homes and communities rather than in institutions. While medical technology promises remarkable help in dealing with disability, equal attention will be given to the subjective experience of illness, because people's experience shapes their lives and those of their families. Finally, there is a growing awareness of the role of spirituality, in the broadest sense of the word, in accepting and managing life with a disability.

See also: Disability, Demography of; Disability, Geography of; Disability: Psychological and Social Aspects; Injuries and Accidents: Psychosocial Aspects; Stigma, Social Psychology of

Bibliography

Albrecht G L 1992 *The Disability Business: Rehabilitation in America*. Sage, Thousand Oaks, CA

Albrecht G L, Seelman K, Bury M (eds.) 2001 *The Handbook of Disability Studies*. Sage, Thousand Oaks, CA

Albrecht G L, Verbrugge L 1999 The global emergence of disability. In: Albrecht G, Fitpatrick R, Scrimshaw S C (eds.) *The Handbook of Social Studies in Health and Medicine*. Sage, London, pp. 293–307

Barnes C 1991 *Disabled People In Britain and Discrimination*. Hurst and Company, London

Castel R 1995 *Les Métamphorphoses de la Question Sociale*. Fayard, Paris

Institute of Medicine 1997 *America's Vital Interest in Global Health*. National Academy Press, Washington, DC

Murray C J, Lopez A M 1996 *The Global Burden of Disease*. Harvard University Press, Cambridge, MA

Shakespeare T 1998 *The Disability Reader: Social Science Perspectives*. Cassell, London

Stiker H-J 1982 *Corps Infirmes et Sociétés*. Aubier Montaigne, Paris

Usten B 1997 *Beta Version of the ICIDH Manual*. World Health Organization, Geneva

G. L. Albrecht

Disasters, Coping with

Natural disasters and catastrophic failures of technological systems require considerable adjustment by both primary victims and rescue personnel. Uncertainty, loss, threat to life, exposure to the grotesque, hazardous working conditions, and breakdowns in social support networks are just some of the consequences of exposure to disasters. Coping and adjustment can be a protracted process because disasters continue to affect people long after their initial impact. Stress associated with these disasters and their aftermath affect mental and physical health, but successful coping may have an impact on distress and other symptoms. Distinctions among coping with natural and technological disasters, coping in emergency personnel and primary victims of disasters, measurement of coping, and relationships between coping and health have been important research topics in this area.

1. *Types of Disasters*

Perhaps it is easier to call to mind examples of disasters than it is to precisely define the term 'disaster.' The 1999 earthquake in Turkey was a disaster, as was the crash of two Norwegian trains in early 2000. In the United States, a catastrophic event officially qualifies as a disaster based on the amount of damage done and the amount of assistance required. The strength of the event alone is not sufficient; a powerful earthquake in a remote area of a California desert may not be considered a disaster, whereas an earthquake of the same magnitude in the city of San Francisco would qualify because considerable damage would result. In addition to damage sustained, substantial disruption to people's lives can also be a feature in the definition of 'disaster.' Further components of what constitutes a disaster are explained when two types of disasters, natural and technological, are contrasted.

1.1 *Coping With Natural Disasters*

Natural disasters are caused by intense and powerful natural events such as hurricanes, tornadoes, earthquakes, and floods. They can also occur because of drought and famine. These and other cataclysmic

events are some of the most threatening situations a person can experience. In a sudden and uncontrollable manner, natural forces can take lives and dramatically alter the environment.

Natural disasters vary widely in predictability and impact. Earthquakes are virtually unpredictable, whereas hurricanes can be tracked for days before they hit land. However, consequences such as the extent of physical destruction and disruption of daily life often take victims by surprise, even after the more predictable types of events. Months of cleaning and rebuilding can follow initial rescue work and recovery of human remains. Moreover, drawn-out and complicated insurance, litigation, and financial issues may compound adjustment difficulties following disasters (Green et al. 1990).

Cataclysmic natural phenomena are widespread, and long-term psychological effects of natural disasters are being studied in the realm of trauma research. Reactions to trauma vary among individuals, and research has focused on documenting the effects of disasters by comparing groups who had different degrees of exposure to the event and predicting who copes well and who does not. Adjustment to natural disasters has been studied in response to hurricanes, tornadoes, earthquakes, and floods. Severe and long-lasting consequences of natural disasters may be exacerbated by characteristics of their onset. For example, sudden disasters with lack of warning are especially devastating. Disaster characteristics such as duration and intensity also influence how people cope with such events. Even events of short duration, such as an earthquake, may cause disruption and threat that last months beyond the acute phase of the disaster. Other characteristics of the disaster experience include threat to one's life, witnessing the death of others, injury to oneself or others, loss of social cohesion and resources of the community, loss of possessions and finances, and separation from family.

Anticipation, acute experience, and aftermath of natural disasters all require coping, as does the period of rebuilding. Coping includes strategies or behavior that reduce threat and may take many forms. Direct action involves reducing exposure to the stressor or directly altering the stressor so that it becomes less threatening. For example, evacuating an area where a hurricane is likely to hit can reduce immediate threat to one's life. Information seeking is another form of coping. Keeping track of a hurricane's path and windspeed, for example, may increase a sense of predictability over an impending disaster. Palliative coping includes efforts at reappraising the stressor or reinterpreting the situation to reduce threat. For example, a volcanic eruption could be seen as a way of nature reclaiming land, or a forest fire may be seen as a part of a natural cycle that is necessary for continued growth and differentiation of the forest. Emphasizing the positive aspects of the postdisaster experience, such as increased appreciation for one's health and family, is yet another way of reappraising the situation.

Seeking emotional and practical assistance from others is another common way of recovering from natural disasters (Smith et al. 1986). Community responses directed at restoring the environment are important in regaining control and coping emotionally. Unfortunately, some cataclysmic events cause social support networks to deteriorate at the time they are most needed. Social support is important in both a practical and an emotional manner (see *Social Support and Stress*). Physical assistance in recovery is often necessary for individuals to cope effectively, and emotional assistance is an important factor in recovering from the trauma. However, disasters often weaken or destroy social ties and cause social disruption that makes the kinds of social support that are so helpful in recovery more difficult to obtain. Disaster disrupts both social support and the ability to perform one's role in the family (Kaniasty and Norris 1993) and increased demand for support may be stressful for those providing it. Denial, use of alcohol and other drugs, as well as learning to live with the stressor are additional ways in which people cope with cataclysmic events like natural disasters (Lazarus and Folkman 1984).

1.2 Coping With Technological Disasters

Like natural disasters, technological disasters are caused by events that can be intense and sudden. Examples include bridge collapses, dam failures, and industrial, maritime, and aviation accidents. They also include more drawn-out events resulting from breakdown in technology in which highly toxic substances are used or stored. Such situations would include leaking toxic waste dumps, such as Love Canal, or accidents involving release of radiation, such as the Three Mile Island nuclear disaster. With increasingly widespread prevalence of technological systems there will inevitably be an increase in the potential for loss of control over these systems.

An important difference between natural and technological disasters is that natural disasters often have a 'low point' at which it seems that the worst is over. In contrast to this situation, many technological disasters do not have a low point, a situation that can contribute to prolonged distress (Baum et al. 1983). The visible damage that a natural disaster such as flooding leaves behind can serve as a reminder that the event has passed, perhaps allowing a starting point for recovery. It also provides a focus for postdisaster recovery emphasizing constructive responses associated with rebuilding. The 'low point' may actually be psychologically beneficial because victims know where they stand in relation to the disaster.

For example, once maximum velocity winds of a hurricane have passed, the worst is probably over.

When disasters do not have a low point, these 'safety' signals are not available. The exceptions, of course, include situations like airplane crashes and maritime disasters. Catastrophic breakdowns of technology that result in toxic contamination do not necessarily have a point at which the worst is over. Those exposed can wonder and worry for years about the extent of their exposure and its possible consequences.

As Green (1998) noted, it can be difficult to compare disasters because each one is unique. She suggested a continuum of intent as a means of conceptualizing disasters. Natural disasters would be placed on the low end of intent, technological disasters in the middle, and hijacking and other purposeful acts of violence would be placed at the high end of the continuum. In the case of technological disasters, an implicit social contract between citizens and corporations is violated. The assumption is that corporations will not harm their customers, workers, or members of the community where they make their products. When this contract is violated, anger and rage are added to the range of emotional responses to disasters.

Disasters have also been categorized with respect to perceived control. Natural disasters reflect a lack of control over the environment, but technological disasters reflect a *loss* of control of what was once under control (Baum et al. 1983). A major assumption underlying reliance on technological systems is that they won't break down. That is, ships are not supposed to sink, and toxic substances are not supposed to leak into the environment. Violations of this implicit assumption can contribute to the distress felt by victims and witnesses of this type of disaster.

Like natural disasters, the onset of technological disaster may be sudden or drawn-out. When the Buffalo Creek dam in West Virginia failed in 1979, there was little if any warning of the impending disaster. Very powerful and overwhelming destruction followed, resulting in hundreds of deaths. While it is possible to gain a semblance of control over disasters like this with regular inspection and evacuation plans, the fact remains that technology was not built to fail and it is extremely difficult to predict when it will. Because this disaster was a failure of a human-made dam, the flood that resulted was less predictable than naturally caused floods.

Research on reactions to the Buffalo Creek disaster identified characteristics that contributed to long-lasting mental health consequences (Green et al. 1990). One of these characteristics was that a mining company was considered responsible for the flooding. Anger, rage, and scapegoating of people and agencies that were not directly involved were some repercussions of the flooding being human-caused. In natural disasters, there is rarely someone to blame. Another factor contributing to the impact was the unusual severity of the flood, which was compounded by the isolated location of the community in a valley. Survivors remained homeless sometimes for weeks because removal of debris was delayed and prolonged. In addition, when temporary housing was arranged, it was not conducive to maintaining social support networks among family and friends, an important component of post-trauma recovery.

2. *Types of Victims*

Important distinctions exist among victims of disasters. Some are affected directly because they were present during a disaster event and experienced threat or loss as a result. Others are affected less directly; people who live in an area hit by a tornado, for example, may not have been there when it hit but may have lost property or had family members injured. Another type of victim, the rescue worker, may be exposed to a variety of extreme stressors including hazardous and fatiguing working conditions.

2.1 *Coping in Primary Victims*

Research on disasters and their effects on mental health have identified a range of outcomes associated with this kind of victimization. Predictors of a disaster's effect on people and what factors may reduce or modify this impact have also been studied. As a result, advances in understanding and treating post-traumatic syndromes have been made based on studies of victims of natural and technological disasters. For many years this research was focused on the primary victims of disasters. Differences in the extent of threat experienced during the disaster contribute to differences in their postdisaster response. Research after a wide variety of disasters has suggested that loss of resources, perceived control, adaptational demand, and life threat influence the intensity and duration of distress and stress responding in primary victims. This variation in response, together with the fact that the majority of disaster victims appear to cope effectively, suggests that situational and individual factors work together to produce long-term reactions to disasters.

The study of victims of large-scale traumas started with the documentation of responses to war trauma. Wars can be construed as human-made disasters. After the American Civil War and World War I, descriptions of 'shell shock' and 'battle fatigue' characterized extreme reactions to this type of stress. Following World War II, studies were conducted on the long-term effects of concentration camps and other traumas of war such as the devastation of Hiroshima. Disasters unrelated to war began to be studied intensively in the 1970s. Currently, a wide variety of disasters ranging from earthquakes to nuclear catastrophes are studied with respect to how victims cope over time.

Symptoms of distress following a disaster experience can include shaking, hyperventilation, palpitations,

sleep disturbances, intrusive memories, shortness of breath, sweating, and gastrointestinal reactions. Many of these anxiety symptoms are also associated with exposure to toxic substances, which can escalate panic in individuals in proximity of industrial accidents. The considerable variation in the range of responses and the ways in which people cope with their distress over time underscores the complexity of the factors which constitute the severity of postdisaster stress (Baum 1987). Some people report only a minimal emotional response to a disaster, but typically postdisaster effects include anxiety, depression, helplessness, numbness, anger, guilt, and relief. Most disasters appear to have some impact on the mental health of witnesses and victims, and sometimes these negative effects last for years. The profound effect disasters can have on people is evidenced by many studies on post-traumatic stress disorder (PTSD) in victims of a wide variety of disasters. For example, survivors of the Buffalo Creek disaster exhibited anxiety, depression, withdrawal, numbness, anger, nightmares, and other stress-related physical symptoms for years after the accident (Gleser et al. 1981).

Coping behavior is an important part of the stress response (see *Stress and Coping Theories*). Coping responses to disasters consist of attempts to remove the source of stress or insulate oneself from its effects. If coping is successful, the person returns to normal physiological and psychological arousal as the threat is eliminated or minimized. The victim can adapt cognitively by reducing awareness of the stressor or by changing their perception of threat. The individual can also cope behaviorally by removing or reducing the threat. People may have particular coping styles or habitual ways of responding to stress. For example, people may use problem-focused coping, or do wishful thinking about a stressful situation, or they may seek and use social support, avoid thoughts about the stressor, or use self-blame coping to different degrees.

Lazarus and Folkman (1984) proposed that stress responses can take manipulative or accommodative forms. They may include direct action responses, where the individual tries directly to manipulate or alter his or her relationship to the stressful situation. Changing the stressor by direct action would include solving a problem that was creating distress or leaving the scene. Flexible coping styles, where the person tailors their strategy to the situation and can manipulate and change their coping, can be especially effective in a changing environment following disasters.

Chronic stress following disasters may affect how primary victims cope with additional stressors, and these acutely stressful events may in turn feed into and maintain the chronically stressed state. Schaeffer and Baum (1984) studied stress-related consequences of living near Three Mile Island (TMI), site of a nuclear accident in the United States in 1979. Uncertainty involving the long latency of radiation's effects on the body, the potential for cancer, uncertain locations of safe and unsafe areas and amount of exposure, and the lack of scientific information on the effects of low-level exposure were all characteristics making this disaster chronically stressful. In fact, higher blood pressure and higher levels of stress hormones were observed in residents of TMI compared with a control sample living 80 miles from the site. Major stressors like the TMI disaster can have negative effects on health (Baum 1987) but relatively little systematic and long-term investigation of health-related variables in disasters have been reported. Evidence of blood pressure, hormonal, and changes in immune system functioning that could be considered intermediate to health effects has been found, and sustained changes in these functions could predispose people to illnesses.

2.2 *Coping in Rescue Workers*

Major aspects of disaster-related trauma include threat to life, witnessing death and injury, and making difficult choices like risking one's own life to save that of another. Other stressors operating during disasters include high levels of uncertainty, lack of clear leadership, and lack of control. Lack of warning also contributes to long lasting effects of those exposed, regardless of whether they were primary victims or rescue personnel. There are qualities of responding to disasters that are more specific or unique to rescue workers, however. Emergency personnel respond to a wide range of events, including large-scale catastrophes like earthquakes and airplane crashes. They must search for and help survivors as well as recover human remains and personal effects. These workers face long hours, unpleasant or gruesome working conditions, and exposure to body fluids and other hazardous materials. Exposure to dead or mutilated bodies from a disaster is a potent stressor (Ursano and McCarroll 1994). Studies have investigated emergency workers and their reactions to the stress of body handling and victim identification. These studies agree that disaster workers experience considerable distress but methods and measures used have been different across studies, resulting in findings of varying stress duration and intensity. Rescue and clean-up activities following natural and technological disasters have been associated with symptoms of PTSD, including recurrent intrusive thoughts and distressing dreams. However, it is difficult to predict long-lasting psychological consequences because many rescue workers exposed to trauma do not experience lasting distress.

Predicting long-term distress is difficult partly because of variability in rescue workers' responses. Some studies have reported consistent negative effects of disaster work, and others have failed to find long-lasting distress in rescue workers. After the 1988 Piper Alpha oil rig disaster, Alexander and Wells (1991)

found no differences in distress when comparing police officers with more or less previous experience dealing with human remains. Anxiety was actually reduced after body handling, reflecting that many of the officers found the anticipation to be more distressing than the actual body handling. Other studies using different rescue workers and disasters did in fact find differences in distress depending on prior experience. In a situation that involved extensive body handling, rescue workers who recovered human remains after the crash of Flight 427 showed persistent symptoms of chronic stress and intrusive thoughts related to their experience at the crash scene. Workers who did not expect to see remains but who did in fact come into contact with them were more distressed than those prepared for the recovery activities (Delahanty et al. 1997).

In studies on coping in rescue workers, talking about their experiences to other workers was found to be beneficial for dealing with distressing thoughts and feelings. Following the 1981 Hyatt-Regency skywalk collapse where 144 people were killed and 188 were injured, rescue workers' distress scores were consistently higher but not significantly different from the controls (Miles et al. 1984). Sharing stories as well as the experience of identifying with the dead were forms of coping commonly endorsed by these workers. Less healthy forms of coping included increased use of tobacco, caffeine, and alcohol. In a review of the literature on handling the dead after disasters, Ursano and McCarroll (1994) highlighted the aspects of disaster work that may contribute to stress or help rescue workers cope with stress. They discussed rescue and recovery work in terms of phases, including anticipation, active duty, and postduty experiences. Preparation of workers so that they know what to expect at the disaster site is beneficial. Identification with the victims, personalization, and emotional involvement heighten the trauma and increase perceived threat, however. As a consequence of this type of research, intervention efforts and critical incident debriefing sessions have been started for rescue workers responding to a variety of disasters. However, research on the effectiveness and specific mechanisms behind such intervention efforts is in its beginning stages.

3. *Summary and Conclusions*

Disasters can be defined by their origins as well as their consequences. Interactions among the nature of the stressor, degree of exposure, types of victims, and coping result in the individual's experience of traumatic stress. Natural disasters such as floods and technological disasters such as nuclear accidents and plane crashes are unpredictable and powerful threats that affect generally all of those touched by them. As a threat to life, property, and assumptions people hold about the world, disasters can cause substantial life changes. Cataclysmic events and disasters can have a sudden, powerful impact and almost universally elicit a stress response. These events usually require a great deal of effort for effective coping. Although evidence suggests that disaster stress varies according to the type of disaster experienced, research has shown that responses to disasters in primary victims and rescue workers can share a relatively predictable trajectory. Early shock and anxiety are often followed by combinations of anger, depression, distress, and post-traumatic stress. Coping is part of the stress response and is involved in changing the perceived level of threat from the stressor. Consequently, recovery from the disaster experience is intimately tied to the coping process.

See also: Coping Assessment; Disasters, Sociology of; Military and Disaster Psychiatry; Post-traumatic Stress Disorder; Stress and Coping Theories; Stress and Health Research

Bibliography

Alexander D A, Wells A 1991 Reactions of police officers to body handling after a major disaster: A before and after comparison. *British Journal of Psychiatry* **159**: 547–55

Baum A 1987 Toxins, technology, and natural disasters. In: Vanden Bos G R, Bryant B K (eds.) *Cataclysms, Crises, and Catastrophes: Psychology in Action*. American Psychological Association, Washington, DC, pp. 247–51

Baum A, Fleming R, Davidson L M 1983 Natural disaster and technological catastrophe. *Environment and Behavior* **15**: 333–54

Delahanty D L, Dougall A L, Craig K J, Jenkins F J, Baum A 1997 Chronic stress and natural killer cell activity following exposure to traumatic death. *Psychosomatic Medicine* **59**: 467–76

Gleser G C, Green B L, Winget C N 1981 *Prolonged Psychosocial Effects of Disaster: A Study of Buffalo Creek*. Academic Press, New York

Green B L 1998 Psychological responses to disasters: Conceptualization and identification of high-risk survivors. *Psychiatry Clinical Neuroscience* **52**(suppl): S57–S73

Green B L, Grace M C, Lindy J D, Gleser J C, Leonard A C, Kramer T L 1990 Buffalo Creek survivors in the second decade: Comparison with unexposed and non-litigant groups. *Journal of Applied Social Psychology* **20**: 1033–50

Kaniasty K, Norris F H 1993 A test of the social support deterioration model in the context of a natural disaster. *Journal of Personality and Social Psychology* **64**: 395–408

Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York

Miles M S, Demi A S, Mostyn-Aker P 1984 Rescue workers' reactions following the Hyatt Hotel disaster. *Death Education* **8**: 315–31

Schaeffer M A, Baum A 1984 Adrenal cortical response to stress at Three Mile Island. *Psychosomatic Medicine* **46**(3): 227–37

Smith E, Robins L, Przybeck T, Goldring E, Solomon S 1986 Psychosocial consequences of disaster. In: Shore J H (ed.)

Disaster Stress Studies: New Methods and Findings. American Psychiatric Press, Washington, DC, pp. 49–76
Ursano R J, McCarroll J E 1994 Exposure to traumatic death: The nature of the stressor. In: Ursano R J, McCaughey B G, Fullerton C S (eds.) *Individual and Community Responses to Trauma and Disaster: The Structure of Human Chaos*. Cambridge University Press, Cambridge, UK

T. Y. Schooler

Disasters, Sociology of

Disasters are nonroutine events in societies or their larger subsystems (e.g., regions and communities) that involve conjunctions of physical conditions with social definitions of human harm and social disruption. The sociological study of disasters requires attention to their special character as social problems (Fritz 1961, Hewitt 1983, Stallings 1995, Kreps and Drabek 1996, and Quarantelli 1998). The development of knowledge about disasters before, when, and after they occur (see summaries of the disaster literature by Barton 1969, Dynes 1970, Quarantelli and Dynes 1977, Kreps 1984, Drabek 1986 and Mileti 1999) necessarily draws on theories of social process and structure (see *Action, Collective*; *Action, Theories of Social*; *Functionalism in Sociology*; *Structure: Social*).

1. Sociological Meaning of the Term Disaster

As defined above, disasters are both physical and sociological events and, as such, inherent to all social systems. The phrase 'nonroutine events' distinguishes disasters as unusual and dramatic happenings from everyday issues and concerns. The dual reference to 'physical conditions' and 'social definitions' means that *each* is individually necessary and *both* are collectively sufficient for disasters to occur in social time and space. The designation 'societies or their larger subsystems' means that human harm and social disruption must have relevance for larger social systems (see *System: Social*). The death of an individual and its impact on his or her family, therefore, is not a disaster unless a rare sequence of actions is triggered that has broader societal impact (e.g., consequences of the assassination of a political elite in an unstable nation). Similarly, the fact that thousands of people die on a nation's highways in any given year does not mean that a disaster has occurred. But the explosion of a terrorist bomb in the heart of a city, even if the number killed was significantly lower than the annual highway death toll, would be consistent with the above definition of disaster. The 1979 nuclear power plant accident in the United States (Three Mile Island) was an emergency and certainly a potential disaster. Potentiality became reality in 1986 at the Chernobyl plant in the former Soviet Union. Poverty, hunger, disease, and social conflict are chronic societal concerns. Economic depressions, famines, epidemics, and wars are disasters as defined above. Global warming and ozone depletion are defined objectively and subjectively as environmental hazards or risks (see *Risk, Sociology and Politics of*). The possible disastrous consequences of these hazards, including secondary effects such as sea-level rises in the oceans, remain matters of scientific and public debate. A nuclear attack is a disaster for which there are historical precedents. A nuclear war is a possibility that heretofore has only been simulated (Kreps and Drabek 1996).

The above, precise sociological definition of disaster is important for three reasons. First, the term itself is common parlance (e.g., 'my day was a disaster') and, as such, much too inclusive to be useful. But despite this inclusiveness, the term is part of the nomenclature of those having responsibility for dealing with actual or potential events. These same events are also of interest to sociologists. Because 'disaster' has bureaucratic as well as popular meaning (e.g., disaster declarations), its usage necessarily is central to the mix of social definitions and physical conditions that sociologists must consider. Second, while all concepts in science are nominal (see *Positivism: Sociological*), consensus about objects of inquiry (in this case, what disasters are) is essential to the development of systematic knowledge about them. Third, disaster research emerged as a post-World War II research specialty within the social problems literature in sociology. That literature was and remains very broad (e.g., see *Alienation, Sociology of*; *Crime, Sociology of*). Identifying the special features of disasters is important for examining a broad range of environmental, technological, and sociopolitical events on their own terms and for comparing collective action related to these events with those of social problems generally (Erikson 1976, 1994). Disasters are no more or less important than other social problems, but they are different, and their uniqueness is of substantial importance theoretically.

2. The Special Character of Disasters as Social Problems

A core theoretical issue in the social problems literature (Spector and Kitsuse 1977) is to account for how social problems are constructed through *claims-making activities* (i.e., collective action that asserts the existence of conditions and defines them as problems), and then how their life histories are determined by *response activities* (i.e., collective action directed to addressing these socially constructed problems). The

life histories of disasters can be compared with other social problems because this social structure of claims-making and response activities is the core object of study. However, crime, poverty, hunger, disease, violence, and many other social problems that are studied have an open-ended quality in social time and space. Disasters do not. Indeed, disasters have a special feature that facilitates their analysis as social problems: *the theoretical centrality of a specific event*. Reference to a specific event is critical because, first, any disaster can be demarcated specifically in social time and space, and second, distinctions can be made among pre-, trans-, and postevent claims-making and response activities at the community, regional, national, or international levels.

The fact that disasters are nonroutine events makes them strategic research sites for their analysis as social problems. Prior to an event, claims-making and response activities within localities (as opposed to nation states) focus on future events for which the probabilities of occurrence are small. With the exception of hazards managers and selected others, disaster prevention and mitigation activities at this systemic level have low salience to government officials, economic and political elites, and the public at large. In a sense, this is a rational response (see *Rational Choice Theory in Sociology*) because the annual probability that a disaster will affect given individuals or localities directly and severely is, indeed, quite low. This does not mean that individuals in harms way are insensitive to hazards that threaten them. But it does suggest that they are preoccupied with more immediate problems and concerns of daily living, and that there are any number of disincentives to being proactive (Turner et al. 1986). The resulting social structure of claims-making and response activities about putative events tends to be subtle and is often contentious. Even at societal levels, where annual probabilities of events and aggregated losses are higher, the prominence of disasters relative to other issues can never be assumed.

However, a major transformation occurs when threats are perceived as imminent, or actual events occur. In effect, nonissues become focal public concerns. The more severe the impacts, the more massive the formal and informal mobilization that follows at local, regional, national, and international levels. Existing groups and organizations restructure everyday routines, and emergent structures of various types are socially created. Claims-making and response activities by a panorama of interlocking structural forms are directed to demands that are both objectively and subjectively labeled as acute. Thus, disasters are nonroutine problems because social processes related to them change dramatically, depending on what stage of their life histories is being considered. But whatever the stage (pre-, trans-, or postevent), disasters are catalysts of collective action at smaller to more inclusive systemic levels.

Prior to their occurrence, disasters are socially constructed as *hazards* or *risks* of various types (Clarke and Short 1993). Claims-making and response activities are directed to defining the magnitude of possible impacts and taking specific steps to prevent, mitigate, or prepare for these impacts. As noted above, these steps are constrained by social processes that are as likely to generate organized opposition as support. When disasters occur, terms such as *emergencies*, *crises* or *catastrophes* are used to capture increasing levels of severity. Claims-making activities are no longer at issue—i.e., there is a conjunction of physical conditions and social definitions—and response activities are directed to dealing with the immediate consequences of physical impact and social disruption. After disasters occur, the events become relevant to the evolution of impacted social systems. It must be recognized, however, that any social changes that follow must be interpreted in light of already existing trends (Bates and Peacock 1993, Peacock et al. 1997). Collective action is directed to long-term reconstruction of physical structures and infrastructures and restoration of the social order.

Unraveling the drama of disaster events requires reliance on basic sociological concepts. Thus, past research has drawn heavily on collective behavior theory (see *Collective Behavior, Sociology of*), organizational theory (see *Organizations, Sociology of*; *Networks: Social*), the integration of collective behavior and organizational theories, role theory (see *Status and Role: Structural Aspects*; *Status and Role, Social Psychology of*), and the integration of collective behavior, organizational, and role theories (Barton 1969, Weller and Quarantelli 1973, Perrow 1984, Kreps and Bosworth 1993, Dynes and Tierney 1994, and Webb 1998). The field clearly has become multidisciplinary in more recent decades, with strong intersections among the natural, social, and policy sciences. There remains, however, a very strong sociological core within ongoing research. That core characterizes both major research emphases within mainstream American sociology and a growing body of disaster research internationally.

3. *Disaster Research*

The origins of disaster research can most readily be traced to the US Strategic Bombing Surveys of World War II (Quarantelli 1994). In this historical context, the prototypical hazards or risks were weapons, ultimately nuclear weapons. The prototypical events studied were war-time bombings of cities and regions (primarily in Germany and Japan). The historical context within which these bombings took place was a global war of unprecedented proportions. In general, the war-time studies showed that civilian populations were remarkably resilient in the face of sustained bombing attacks. But with a rapidly developing

nuclear arms race, there was a very practical rationale for the subsequent funding of research on peacetime disasters—to study preparedness for and response to events that parallel war-time situations (i.e., acute impacts of considerable magnitude and scope). Prior to World War II, there was no systematic program of disaster research in sociology. Following the war, the above rationale supported the development of disaster research in the United States during the next 25 years, allowing the field to become an established research specialty within sociology. With funding from defense-related government agencies, studies at the University of Chicago (National Opinion Research Center) and two other state universities (Maryland and Oklahoma) were initiated in the late 1940s and early 1950s. These were followed by a series of studies under the auspices of the National Academy of Sciences—National Research Council from the early 1950s to the mid-1960s. Then in 1963, the Disaster Research Center was established and has been in continuous operation ever since. This Center was located for over 20 years at the Ohio State University (1963–85), and it presently resides at the University of Delaware (1985–present). Disaster research became an established specialty within sociology primarily because a critical mass of professionally trained sociologists in the above and other universities completed most of the studies undertaken prior to the late 1960s. While a public policy rationale was important to those funding the work, the interests of university researchers and the students they trained were theoretical as well. The contemporary attention to systemic structures and processes continues to be grounded in the earlier studies, although considerations of possible war-time applications (Perry 1982) have been supplanted by concerns about a much broader range of societal hazards and risks.

Like the strategic bombing studies, the early to more recent research on peacetime disasters have highlighted the remarkable absorptive capacities of social systems, contradicting conventional notions that during a disaster victims will panic (see *Panic, Sociology of*), that those expected to respond will abandon occupational roles, that community structure will break down, and that antisocial behavior will be rampant. Having systematically debunked such myths about disaster behavior, the more interesting questions to sociologists have related to describing and explaining structural continuity and change before, during, and after an event. The earlier reference to claims-making and response activities of social systems provides a very useful way of framing these questions. Where the majority of studies prior to the 1970s focused on collective action during the immediate emergency period, contemporary research gives more balanced attention to disaster prevention, mitigation, preparedness, and long-term recovery.

Sociology no longer dominates disaster research. There has been growing attention since the late 1960s to studying an expanded range of peacetime hazards and disasters on their own terms. There has been growing involvement of other social sciences (geography, economics, psychology, political science, anthropology, public policy, and public administration) in studying these hazards and disasters. It has become increasingly necessary to wed social science research and findings with those of natural sciences and engineering specialties (seismology, meteorology, structural engineering, and many others). In effect, disaster research has become a multidisciplinary, applied field that seeks practical solutions to putative or agreed upon problems posed by actual or potential events. Sociologists pursuing their own applied or basic research interests cannot do so in an intellectual vacuum.

4. *The International Context of Disaster Research in Sociology*

Just as disaster research has become a multidisciplinary applied field in the United States and internationally, ironically, fundamental sociological questions are being emphasized by scholars outside of the United States (Lars Clausen, Wolf R. Dombrowski, John K. Schorr from Germany; Kitao Abe, Hirotada Hiros, Ritsuo Akimoto from Japan; Tom Horlick-Jones, Barry Turner, Nicholas Pidgeon from England; Charles Chandessais, Claude Gilbert from France; Carlo Pelanda, Bruna De Marchi from Italy; Kenneth Hewitt, Joseph Scanlon from Canada; Boris Porfiriev, Konstantin Popov from Russia; Neil R. Britton from Australia and New Zealand; Uriel Rosenthal from The Netherlands; and many others from both industrialized and developing societies). The involvement of these scholars has been very beneficial. They are raising important theoretical issues about social structure and disaster. They are questioning the dominance of one country (United States) or a single paradigm (see *Functionalism in Sociology* on the evolution of the field. They are concerned about the historical preoccupation with a narrow range of events (rapid as opposed to slow onset disasters). Such issues merit serious attention.

Perhaps the most fundamental issue that continues to be discussed internationally became the title of a recent book edited by a pioneer disaster researcher (Quarantelli 1998): *What is a Disaster?* This volume essentially culminates a multiyear dialogue among social scientists from the United States and other countries on defining the core subject matter for sociology and, hopefully, disaster research as a whole. Despite the wealth of alternative issues and perspectives that have been presented, the precise definition used in this article captures the substantial underlying agreement that exists internationally. Explicitly or implicitly, this agreement is based on two factors. First, there is a strong commitment to using

the tools of science to conceptualize and measure disaster as a sociological construct. Second, there is an effort to extract general knowledge about the relationship between social structure and disaster from studies of particular historical and cultural contexts. The international character of disaster studies therefore provides the comparative approach that is essential for defining the limits to which extant theoretical relationships and models can be generalized. To that end, it is advantageous to maintain the broad range of environmental, technological, and sociopolitical events subsumed by the characterization of disasters as nonroutine social problems.

Bibliography

Barton A H 1969 *Communities in Disaster: A Sociological Analysis of Collective Stress Situations*. Anchor, Doubleday, Garden City, New York

Bates F L, Peacock W G 1993 *Living Conditions, Disasters, and Development: An Approach to Cross-cultural Comparisons*. University of Georgia Press, Athens, GA

Clarke L, Short J F 1993 Social organization and risk: some current controversies. *Annual Review of Sociology* **19**: 375–99

Drabek T E 1986 *Human System Response to Disaster: An Inventory of Sociological Findings*. Springer-Verlag, New York

Dynes R R 1970 *Organized Behavior in Disaster*. Heath Lexington Books, Lexington, MA

Dynes R R, Tierney K J 1994 *Disasters, Collective Behavior, and Social Organization*. University of Delaware Press, Newark, NJ

Erikson K 1976 *Everything in its Path: Destruction of Community in the Buffalo Creek Flood*. Simon and Schuster, New York

Erickson K 1994 *A New Species of Trouble: Explorations in Disaster, Trauma, and Community*. W. W. Norton and Company, New York

Fritz C E 1961 Disasters. In: Merton R K, Nisbet R (eds.) *Contemporary Social Problems*. Harcourt Brace & World, New York

Hewitt K 1983 *Interpretations of Calamity: From the Viewpoint of Human Ecology*. Allen & Unwin, Boston

Kreps G A 1984 Sociological inquiry and disaster research. *Annual Review of Sociology* **10**: 309–30

Kreps G A, Bosworth S L 1993 Disaster, organizing, and role enactment: A structural approach. *American Journal of Sociology* **99**: 428–63

Kreps G A, Drabek T E 1996 Disasters are nonroutine social problems. *International Journal of Mass Emergencies and Disasters* **14**: 129–53

Mileti D S 1999 *Disasters by Design: A Reassessment of Natural Hazards Research in the United States*. Joseph Henry Press, Washington, DC

Peacock W G, Morrow B H, Gladwin H 1997 *Hurricane Andrew: Ethnicity, Gender, and the Sociology of Disasters*. Routledge, London

Perrow C 1984 *Normal Accidents: Living with High-Risk Technologies*. Basic Books, New York

Perry R W 1982 *The Social Psychology of Civil Defense*. Lexington Books, Lexington, MA

Quarantelli E L 1994 Disaster studies: The consequences of the historical use of a sociological approach in the development of research. *International Journal of Mass Emergencies and Disasters* **12**: 25–49

Quarantelli E L 1998 *What is a Disaster: Perspectives on the Question*. Routledge, London

Quarantelli E L, Dynes R R 1977 Response to social crisis and disaster. *Annual Review of Sociology* **3**: 23–49

Spector M, Kitsuse J I 1977 *Constructing Social Problems*. Cummings, Menlo Park, CA

Stallings R A 1995 *Promoting Risk: Constructing the Earthquake Threat*. De Gruyter, New York

Turner R H, Nigg J M, Paz D H 1986 *Waiting for Disaster: Earthquake Watch in California*. University of California Press, Berkeley, CA

Webb G 1998 *Role Enactment in Disaster: Reconciling Structuralist and Interactionist Conceptions of Role*. UMI Dissertation Services, Ann Arbor, MI

Weller J M, Quarantelli E L 1973 Neglected characteristics of collective behavior. *American Journal of Sociology* **79**: 665–85

G. A. Kreps

Disciplines, History of, in the Social Sciences

1. Social Science and Disciplinary History

The social nature of human beings means that there have always been rules regulating the life of a community. Tacit social knowledge exists in any human community. The whole of human history is filled with examples of efforts to give guidance to rulers, sometimes written down in books or manuals. Some contemporary social scientists try to trace a lineage of reasoning reaching back hundreds, if not thousands, of years. Thus political scientists often portray Plato and Aristotle as early representatives of their discipline. Economists and educational researchers may point to the perennial nature of the aspects of human life that constitute their scholarly domains. In the case of legal scholarship a tradition, if only in a weak sense of the word, may be said to obtain from at least the twelfth and thirteenth centuries. The emergence of this tradition is in fact coterminous with the creation of European universities (Kelley 1990). Examples of this kind are interesting in their own right. However none of them amounts to a disciplinary social science history in any real sense of the word. This is so for three reasons (which provide the rationale for the structure of this article).

First, the concept of social science appears only in the 1790s and its use presupposes a meaningful conceptualization of something called a society. This does not occur before the second half of the eighteenth century.

Second, in none of these cases can we talk of a discipline in the sense of a relatively coherent and

delimited program of research and teaching that is consolidated and consistently reproduced in a university environment. This presupposes the existence of a university, but also that the university is seen as a primary vehicle for research activities. However, the research-orientated university does not emerge as a key societal institution until the second half of the nineteenth century.

Third, in none of the cases mentioned above is there an unbroken scholarly tradition that links these early efforts to present-day activities in university settings. Such intellectual and institutional continuities presuppose a degree of disciplinary consolidation that does not occur on a more general level until the early twentieth century. In this respect, developments in the 1930s are crucial although limited to Europe and the Americas. Elsewhere, disciplinary consolidation does not occur until after World War II.

2. *The Rise of the Social Sciences: From Moral and Political Philosophy to Social Science*

The first recorded uses of 'social science' appear in France in the 1790s. Ever since then, these new forms of knowledge have been characterized by an effort to understand the world of modernity. They have tended to describe key features of this new world in terms of processes of industrialization, urbanization, and political upheaval, originating at the northwestern edge of the Eurasian landmass but eventually having global repercussions. In the self-understanding of the social sciences, accordingly, there is a longstanding and predominant view about the formation of modernity, which highlights transformations in the late eighteenth and early nineteenth centuries. It sees these transformations as a conjunction of a technological and a political transformation—the industrial and the democratic revolutions, respectively. This traditional interpretation, however, radically underestimates the deep-seated epistemic transformation that occurred in the same period.

In fact, in this period there is a fundamental transition from earlier forms of moral and political philosophy into social science. This transition is linked to an institutional restructuring not only in forms of political order but also in the forms in which human knowledge is brought forth and claims to validity are ascertained. One feature of this institutional transition is the emergence of a public sphere that gradually replaces arenas of a more closed nature such as aristocratic literary salons (see *Civil Society/Public Sphere, History of the Concept*). Another is the rise of new or reformed public higher education and research institutions that come to replace both the laboratories of wealthy amateurs and the academies under royal patronage and partial control (see *Universities, in the History of the Social Sciences*).

The rise of the social science disciplines must then be cast in terms of the fundamental transformation of European societies that the formation of modernity entailed. One intellectual and cultural transformation in this period pertains precisely to the concepts of society and history and to the new awareness of the structural and constraining nature of societal life. Pierre Manent has put forward the notion that society is a 'postrevolutionary discovery' (Manent 1998, p. 51). It is true that the term 'society' underwent a long conceptual development in the French context in the course of the seventeenth and eighteenth centuries (see Baker 1994)—with a dramatic increase in the utilization of the term in the mid-eighteenth century. However, even if there was a long process of gestation for the modern concept of society, the unique event of revolutionary upheaval requires that discursive controversy and political practice become joined in the formation of a distinctly modern era.

The late eighteenth century witnessed the creation of a political project encompassing the whole world and shattering the existing absolutist order. In this process horizons of expectation to use Koselleck's term opened up that were previously unknown. This sense of openness and contingency also served as a forceful impetus to an examination of the structural conditions of the political body and entailed a passage from political and moral philosophy to social science. This transition required that four key problematics—which today are more acutely open to reinterpretation than they have been for decades—be formulated or at least fundamentally reformulated and enter into the new social-science discourse.

First, the role of historical inquiry becomes crucial. Historical reasoning becomes an integral part of the intellectual transition, and even abstract reason itself becomes historicized in early nineteenth-century philosophy. However, the moral and political sciences break up into a variety of new discourses that in the course of the nineteenth century coalesce and are reduced to a number of disciplines. This means that the stage is set for the divergence between a professionalized historical discipline and the other social and human sciences, a divergence that we still today experience as a major intellectual divide.

Second, interest in language and linguistic analysis enters into all domains of the human and social sciences. One outflow of this is the constitution of textual and hermeneutic modes of analysis. There is also an effort to historicize language and linguistic development itself, thereby providing a crucial link with the historic construction of the notion of different peoples.

This leads to a third problematic: that of constituting new collective identities. Membership in a collectivity could no longer be taken for granted in the life experiences of the inhabitants of a certain village or region. Nor could a relationship of obligation and loyalty between the princely ruler and his subjects

continue to constitute the unquestionable core of the body politic. That, however, meant that even the most basic categories of societal existence were open to doubt. In the western part of Europe, categories such as 'citizen' and 'compatriot' came to play an even more important role at the turn of the eighteenth century.

Fourth, assumptions about what prompts human beings to act and how to interpret their actions within a broader framework are at the very core of any scholarly program in the social and human sciences. At the turn of the eighteenth and nineteenth centuries, the fundamental categories that we still largely draw upon were elaborated and proposed. We might describe these categories as follows:

(a) *Economic-rationalistic*, with a corresponding view of society as a form of compositional collective;

(b) *Statistical-inductive*, with a view of society as a systemic aggregate;

(c) *Structural-constraining*, with a view of society in terms of an organic totality; and

(d) *Linguistic-interpretative*, with a conceptualization of society as an emergent totality.

The transition from a discourse of moral and political philosophy to a social science entailed a decisive shift from an agential—some would say voluntaristic—view of society to one that emphasized structural conditions. In economic theorizing this also entailed a shift away from a concern with moral agency. During the nineteenth century, the context of 'average economic man' became a web of structural properties and dynamic regularities rather than a moral universe of individual action.

Thus, fundamental categories of agency and society that came to be elaborated and refined during much of the rest of the nineteenth and twentieth centuries can be discerned in rudimentary form already during the great transition. So too, however, can some of the features that came to affect these endeavors.

One such tacit but crucial feature concerns the abandonment of the truly universal heritage of the Enlightenment project in favor of forms of representation and endowment of rights based on territoriality or membership in a linguistically and historically constituted and constructed community. Another feature was an emerging and growing chasm between moral discourse and other forms of reasoning about society. Thus, an earlier encompassing conception of the moral and political sciences was gradually replaced by social sciences that marginalized moral reasoning or consigned it to the specialized discipline of philosophy. Third, historical reasoning, which had been at the core of the intellectual transformation at the end of the eighteenth century, became a separate discipline and, toward the end of the nineteenth century, a permanent divide emerged between history and the social sciences.

The end of the eighteenth century was a formative period in the rise of the social sciences in conceptual terms. It is possible to discern, across all confrontations and divergences, a fundamental acknowledgement of the idea that agency, reflexivity, and historical consciousness might help construct a new set of institutions but that this takes place within a complex web of interactions that jointly constitute a society. Thus, there existed a limited number of thematic foci underlying the cultural constitution of a new set of societal macroinstitutions.

3. The Institutionalization of the Social Sciences: The Social Question, the Research University, and the New Nation-states

Relatively early in the nineteenth century, economics (or rather political economy) became differentiated from moral philosophy. It is also in this period that history emerges as a scholarly field with its own canon of rules, but the full disciplinary formation of history is a highly extended process.

However, social science as an institutionalized scholarly activity performed within a series of disciplines is largely a phenomenon of the late nineteenth and early twentieth centuries. It is an activity that directly and indirectly reflects concerns about the wide-ranging effects of the new industrial and urban civilization that was rapidly changing living and working conditions for ever-larger parts of the population in many European nations during the nineteenth century. These changes, often collectively referred to as 'the social question' (*die soziale Frage*), were gradually forcing themselves upon the agendas of parliamentary bodies, governmental commissions, and private reform-minded and scholarly societies (see *Social Question: Impact on Social Thought*). Often the impetus for the search for new knowledge came from modernizing political and social groupings that favored industrialization but advocated more or less far-reaching social reforms. These groupings gradually came to embrace the notion that political action to alleviate 'the social question' should be based on extensive, systematic, empirical analysis of the underlying social problems. In country after country, the political agenda of the nineteenth century was being formed by two macro-projects: the search for a solution to 'the social question' and to the question of national identity and nationhood within new or reformed nation states (Rueschemeyer and Skocpol 1996, Wagner et al. 1991a, 1991b).

Between 1870 and the turn of the century, all signs seemed to indicate that new social science approaches would gain access to the scientific institutions. The thinking of their proponents was widely diffused among the intellectual and political elites. This occurred at a time when traditional liberal economics was undergoing a crisis. Thus traditional political economy was seen to be unable to deal with the 'social question' and the widespread deterioration of living

and working conditions due to urbanization and industrialization.

However, another important, and indeed competing, field of study and training, whose concerns overlapped those of the new social sciences, was that of the legal sciences. The strength of the legal scholar's position in continental Europe arose from two factors. First, the training function of the universities for the state service produced officials and judges. This meant that legal scholarship came to exert a considerable impact on the general outlook of the servants of the state. Second, legal scholars sought to provide a doctrine, a body of concepts that was based on elaborate technical distinctions and would enable lawyers and judges to act with promptness and precision, clarify the deliberations of the law-maker, and bring coherence and order into the legal system (Dyson 1980, p. 112). This doctrine was legal positivism, which was first developed for private law and later for public law (in the latter case known as the legal theory of the state). Legal scholarship remained an important alternative to social science and, in fact, strengthened its position towards the end of the nineteenth century. Thus, the so-called 'state sciences' in the German and Nordic states did not really develop into anything that might be labeled 'political science.' Rather they became increasingly embedded in the legal constitutional scholarship that was expanding in Germany in the wake of German political unification. Conversely, in the Nordic countries a similar tradition of 'state sciences' in the late nineteenth century was characterized by a dual legacy of constitutional legal scholarship and, as in Britain, of studies of philosophy and political history.

By and large, approaches to a social and political science did not succeed in institutionalizing themselves in European academia. In some cases, they tried but failed. In other cases, they did not perceive the relevance of academic institutionalization, but blossomed during a passing period of a favorable political climate and decayed with the changing societal situation. For example, the term 'sociology' had been coined early in the nineteenth century, but the key sociologists of the late nineteenth century (Weber, Durkheim, and Pareto) remained broad social science generalists. Their contributions and professional allegiances traversed a range of fields, including politics, economics, education, history and religion, and the term sociology often referred to a broad historical-comparative study of society. However, it was a type of study that saw itself as scientific and separate from reform-orientated activities of a more general nature.

In the United States, social science research originally had the same characteristics of associational organization and ameliorative orientation as in Europe. For example, the American Social Science Association (created in 1865) embraced the notion that the social scientist was a model citizen helping to improve the life of the community, not a professional, disinterested, disciplinary researcher. During the final cades of the nineteenth century, however, this model was gradually replaced by the emerging disciplinary associations such as the American Economic Association (1885), the American Political Science Association (1903), and the American Sociological Society (1905) (Furner 1975, Haskell 1977, Manicas 1987, Ross 1991). They did not have to face the kind of entrenched opposition or ideologically motivated hostility that many similar efforts in Europe met with (even though it would be misleading to portray these processes as the harmonious unfolding of disciplinary consolidation).

By the turn of the century, a particular pattern of differentiation and professionalization had emerged in the US and proved to be compatible with an active role in government service early on. Economists and psychologists were, for instance, able to play a role in government service as early as World War I, just as demographers were able to do through the Bureau of the Census. At the same time, disciplinary and professional recognition was being achieved within the setting of American universities which started to become ever more orientated towards research undertakings (Geiger 1986, Rothblatt and Wittrock 1993)

This process of successful disciplinary consolidation marked the beginning of a divergence between American and continental European social science. Late nineteenth century American social scientists, many of them German-trained, defined their intellectual projects in a society undergoing a process of rapid transformation: industrialization, urbanization, and concomitantly emerging massive social and political problems. As social scientists with a professional legitimacy, they tried to mark out their own scientific territory and establish their own systems of accreditation. These ambitions entailed the establishment of separate social science disciplines (Manicas 1987). In Europe, on the other hand, the professorate often already had an established position; one writer (Ringer 1969) even uses the term 'a Mandarin class' to describe the situation of the leading German academics at the turn of the nineteenth century. The situation in some other countries, such as Sweden and Norway, was no different.

The social sciences emerged as forms of knowledge about societies undergoing fast and deep societal transformations. The relative success of given research programs largely depended on the intellectual coherence and viability of these programs. The cases of Max Weber and Emile Durkheim are exemplary in this respect, but so is the ability of marginalist economics programs to assert themselves even under political conditions which may have appeared as far from advantageous. However, the institutionalization of the social sciences was also clearly related to two other forms of institutional processes.

First, the late nineteenth century was the period when the idea of the university as a key vehicle for

research became predominant in a number of countries across the globe, from California in the west to Japan in the east. This process created the possibility for the new social sciences, if in a highly uneven way, to find a relatively stable basis for continued research and training.

Second, the late nineteenth century was also a period of deep institutional change in the political and administrative order in a number of countries. Administrative reforms were undertaken and new objectives assigned to the state. Furthermore demands for wider political representation meant that institutional reforms of the state, not least the demand for a culturally coherent nation-state, created demands upon the social and historical sciences to contribute to these reform processes. The particular forms of interaction between research programs, efforts to deal with the social question, and the relationship of that question to the restructuring of universities and of the state, differed greatly across countries. However, in a number of cases, opportunities arose for the successful institutionalization of some of the social science disciplines in ways that came to structure disciplinary developments in these countries for many decades.

4. *Research Programs in the Interwar World*

The end of World War I saw the triumph of liberal democracy and the reorganization of the political order across the European continent as well as a decisive weakening, if not the end, of European global predominance. Soon it became evident that the triumph of democracy was being replaced by its crisis and the emergence of new forms of political and social order with claims to represent the future of humankind in Europe and beyond. It was in this period that preeminent representatives of a range of social science disciplines engaged in a self-critical reflection on the history of their own disciplines. Based on such historical reflection, however, they were able to formulate research programs that came to serve as focal points for scholarly endeavors for decades to come.

Within just a few years in the middle of the 1930s a range of pathbreaking programmatic formulations occurred. At this time economists, most notably Keynes in England but also Gunnar Myrdal and the other members of the so-called Stockholm school in Sweden, took stock of the historical experience of their discipline and formulated a coherent long-term research program. In sociology, Parsons' *The Structure of Social Action* was equally historically and programmatically orientated. Its influence extended far beyond the domain of disciplinary sociology and came to affect developments in a number of other fields, including political science and social anthropology. In statistics, Fisher was able to achieve an encompassing synthesis that became a landmark. In political science and sociology, scholars at Chicago and Columbia opened up new areas of empirical research. In Europe, the early work on electoral geography by André Siegfried in France was complemented by sociological and sociopsychological studies by Paul Lazarsfeld and his collaborators in Austria, and by Tingsten's and Wold's minor classic, *The Study of Political Behaviour* (1937).

The comprehensive philosophical programs outlined by the Continental logical empiricists and their Anglo-Saxon counterparts, despite their relative numerical weakness, came to set agendas for years to come. In this respect, Alfred Jules Ayer's polemical and programmatic volume on *Language, Truth and Logic* (1934) became one of the most noted examples, as did Otto Neurath's *Empirische Soziologie* published at roughly the same time. Meanwhile the French Annales School charted a completely new terrain of research that forever changed scholarly efforts in the historical and the social sciences. This was also the period when the first edition of the *International Encylopedia of the Social Sciences* appeared, as a further testimony to need for reflection, for stocktaking but also for the setting of new scholarly agendas.

Furthermore, research programs emerged in the 1930s that positioned themselves in conscious opposition to disciplinary developments in the social sciences and the humanities. One prominent example of this was the broad synthetic program associated with the so-called Frankfurt school. A very different one was the effort of philosophical phenomenologists, most prominently Husserl and Heidegger, to elaborate a conceptualization of human activity from a point of view at odds with that of both analytical philosophy and dominant forms of empirical social research.

Many of these efforts proved to be of lasting importance. However, in institutional terms, the devastating effects of the European political landscape in these years and the ravages of World War II mean that it is difficult to discern clear institutional continuities in the continental European case. Many of the developments in the interwar years were followed by deep ruptures sometimes tended to make the social sciences after World War II appear as a new phenomenon.

5. *The Internationalization of the Social Sciences: The Age of International Associations, Public Policies, and Mass Higher Education*

The full-blown institutionalization of the social science disciplines on a global scale is largely a phenomenon of the era after World War II. One manifestation of this was the establishment, originally under the auspices of *UNESCO*, of the *International Political Science Association (IPSA)* and the *International Sociological Association (ISA)* in 1949. A process of

professionalization was set in motion and came to exert a truly profound influence worldwide in the wake of the expansion of higher education systems in a range of countries in the 1960s and early 1970s.

The 1960s saw a dramatic expansion of higher education systems across Western Europe and North America but in many other parts of the world as well. In the same countries, sweeping processes of administrative reform also occurred, which often coincided with the coming to power of new political majorities. Major new public policy programs were launched across the board in these countries. In this context, the social sciences came to be finally and firmly entrenched as academic disciplines in university settings. In this same age of great public policy programs, disciplines such as political science and sociology were able to secure a firm basis in a series of European countries, in some cases for the first time; in others (as in Britain, Germany, Sweden, and Italy) in a renewed and greatly expanded form. This expansion was paralleled by a growth of the social science disciplines on a global level that for the first time tended to make the international associations truly international.

On all continents, the full array of disciplines and subdisciplines appeared. There was also an impressive expansion of research methods being utilized. Thus a previously predominant concern for institutions and processes on a national level was gradually complemented by a stronger research orientation towards the study and management of clearly quantifiable processes. In terms of research methods, the 1960s were the period of the breakthrough of the behavioral revolution; a revolution which had been largely foreshadowed by European scholarly efforts in the interwar period. No longer could historical, juridical, and philosophical reasoning alone or in combination be considered sufficient for the analysis of social and political phenomena. Methods and techniques previously elaborated in statistics, sociology, psychology, and economics were now being tapped by social scientists on a vast scale.

This shift in research methodology coincided with the numerical expansion of the social science disciplines and in Europe was often complemented by the introduction of more formalized graduate education programs, normally with compulsory courses in research methodology. At the same time it became possible to see the emergence of an informal 'invisible college' of younger scholars in Europe and in other continents as well, in marked contrast to the much more national orientation of scholars of older generations. In those universities and countries where this shift was most decisively pushed through, there were certainly instances where the older juridical, historical, and philosophical competence was either partially lost or could at least not be developed on a par with the new approaches. Thus it became gradually antiquated and lost touch with what had been the rationale for the use of these traditional methods in the first place.

Apart from external political-societal reasons for this pattern of disciplinary development, a fundamental fact of a long period after World War II was that only in the United States did the social sciences have sufficient size and scope to make widespread international emulation appear feasible. Thus, for example, the American Political Science Association had more than five thousand members in the early 1950s, when corresponding figures for European nations were generally only on a scale of 1 percent or less of that number. Even in the early 1990s, the combined enrolment of West European political science associations was only a third of that of the American association.

Furthermore, only the American version of a science of society and of politics could present a clear institutional lineage dating back to the late nineteenth and early twentieth centuries. In this perspective, the history of the social sciences in a number of European countries constitutes exceptions to this general tendency only in a weak sense of the word. This is perhaps particularly clear in the case of political science. There has been a persistent notion that political science is intimately linked to the particular intellectual and institutional history of the United States. This can be discerned as early as 1923, in Charles Merriam's presidential address to the American Political Science Association. The same theme recurred 30 years later in the presidential address of Pendelton Herring, and again in the one delivered by Samuel Huntington another quarter of a century later. Political science is then seen to constitute an originally American-conceived science that has successfully spread to a range of nations that (by virtue of their effective, if belated, acceptance of a pluralistic political system) had come to qualify as proper recipients of this intellectual heritage.

Even if such a perspective is understandable, it rests on a misperception. It identifies political science inquiry with one particular form of such inquiry—related to the concerns of liberal individualism. In this and many similar cases, a lack of historical perspective threatens to unduly narrow both the range of problems addressed and the theoretical traditions explored and advanced.

6. *Continuities and Reassessments*

Three key features stand out in the development of the social sciences in the late twentieth century. First, their professional consolidation has proceeded. Their institutional position in the modern mass higher education systems seems relatively secure. Never before have there been so many social scientists in the world, and never before have they been so well organized in

professional terms, well trained in technical terms, and internationally linked through journals, networks, and research conferences. Many social science disciplines—political science, sociology, educational research, business administration studies, and public policy research—have to a large extent emerged and evolved as confederations of different practices. They have been held together by a common concern with broad substantive themes rather than by a core of theoretical assumptions.

Second, however, it is remarkable to what an extent truly innovative research contributions have resulted from work in scholarly settings outside of the structure of regular disciplinary university departments. Three types of such environments seem to have been particularly important to the intellectual advancement of the social sciences. First, especially in the major American research universities, the system of so-called organized research units has been of great importance in providing settings outside of the departmental structure designed to be hospitable to research of a transdisciplinary nature. Second, research institutes, also normally of a transdisciplinary nature, have been created both outside and inside of the structure of regular universities. In Germany, the institutes of the Max Planck Society have played a crucial role in this respect. Third, the important role of so-called institutes for advanced study has been steadily increasing during the second half of the twentieth century. Such institutes allow scholars, free from the normal concerns of a university department and coming from a variety of disciplines, to focus on their own research efforts during a limited period, normally one academic year. Whereas initially such institutes tended to be international in their scholarly orientation and national in their own organizational structure, there has recently been an increase in the number of such institutes set up within individual universities. The so-called humanities centers now found in a large number of American universities provide one example of this trend. The recommendation of the Gulbenkian Commission (Wallerstein 1996) that every university create a transdisciplinary institute for advanced study is another indication, a recommendation that now slowly seems to correspond to an emerging institutional reality.

Third, the demands upon the social sciences, from politicians, administrators, and the public at large, to demonstrate the usefulness of their contributions, do not seem to have subsided but if anything to have increased. Sometimes these demands have even resulted in the creation of new proto-disciplines such as social work, criminology, and migration research.

The foremost social scientists of the late nineteenth century had to demonstrate the intellectual and institutional legitimacy of their activities relative to competing forms of scholarship in law, economics, and philosophy. These competitors remain but have been complemented by management consultants, computer experts, and representatives of new transdisciplinary fields such as that of the cognitive sciences.

There is no reason to believe that the social sciences will not be able to defend their current position. It is important, though, to recognize that advances in these sciences depend upon intellectual integrity in the face of an array of demands for immediate usefulness. At the same time, any truly innovative social science inquiry has to be contextually and historically sensitive. This also means that a search for a solution to most of the essential problems of the social sciences has to proceed with a high degree of openness to the contributions from outside of the field of a single discipline. This, however, should be seen as an asset rather than as a liability. In this situation, historical reflection is a necessary component of serious research efforts. Clearly, a number of assumptions concerning our deepest cultural and national identities and their relationship to the basic forms of political and social organization have to be examined anew with the same openness for foundational inquiry as social scientists have shown in earlier periods of fundamental change. In fact, social science scholarship may be crucially important to the possibility of a civilized community of human beings.

See also: Anthropology, History of; Economics, History of; Hermeneutics, History of; History of Science; International Science: Organizations and Associations; Paradigms in the Social Sciences; Political Economy, History of; Psychology: Historical and Cultural Perspectives; Quantification in History; Quantification in the History of the Social Sciences; Social Science, the Idea of; Sociology, History of; Statistical Methods, History of: Post-1900; Statistical Methods, History of: Pre-1900; Universities, in the History of the Social Sciences

Bibliography

Baker K M 1994 Enlightenment and the institution of society: Notes for a conceptual history. In: Melching M, Velema W (eds.) *Main Trends in Cultural History: Ten Essays*. Rodopi, Amsterdam, pp. 95–120

Collini S, Winch D, Burrow J 1983 *That Noble Science of Politics: A Study in Nineteenth Century Intellectual History*. Cambridge University Press, Cambridge, UK

Dyson K 1980 *The State Tradition in Western Europe*. Martin Robertson, Oxford, UK

Fox C, Porter R, Wokler R (eds.) 1995 *Inventing Human Science: Eighteenth Century Domains*. University of California Press, Berkeley, CA

Furner M O 1975 *Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science 1865–1905*. The University Press of Kentucky, Lexington, KY

Geiger R L 1986 *To Advance Knowledge: The Growth of American Research Universities 1900–1940*. Oxford University Press, New York

Goodin R E, Klingemann H-D 1996 Political science: The discipline. In: Goodin R E, Klingemann H-D (eds.) *A New Handbook of Political Science*. Oxford University Press, New York, pp. 3–49
Graham L, Lepenies W, Weingart P (eds.) 1983 *Functions and Uses of Disciplinary Histories*. Reidel, Dordrecht, The Netherlands
Haskell T S 1977 *The Emergence of Professional Social Science*. University of Illinois Press, Urbana, IL
Heilbron J, Magnusson L, Wittrock B (eds.) 1998 *The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850*. Kluwer, Dordrecht, The Netherlands
Jarausch K (ed.) 1983 *Transformations of Higher Learning, 1860–1930*. Klett-Cotta, Stuttgart, Germany
Kelley D R 1990 *The Human Measure: Social Thought in the Western Legal Tradition*. Harvard University Press, Cambridge, MA
Lepenies W 1988 *Between Literature and Science: The Rise of Sociology*. Cambridge University Press, Cambridge, UK
Manent P 1998 *The City of Man*. Princeton University Press, Princeton, NJ
Manicas P T 1987 *A History and Philosophy of the Social Sciences*. Blackwell, Oxford, UK
Pagden A (ed.) 1987 *The Languages of Political Theory in Early Modern Europe*. Cambridge University Press, Cambridge, UK
Ringer F K 1969 *The Decline of the German Mandarins*. Harvard University Press, Cambridge, MA
Ross D 1991 *The Origins of American Social Science*. Cambridge University Press, Cambridge, UK
Rothblatt S, Wittrock B (eds.) 1993 *The European and American University Since 1800: Historical and Sociological Essays*. Cambridge University Press, Cambridge, UK
Rueschemeyer D, Skocpol T (eds.) 1996 *States, Knowledge and the Origins of Social Policies*. Princeton University Press, Princeton, NJ
Wagner P, Weiss C H, Wittrock B, Wollmann H (eds.) 1991a *Social Science and the Modern State: National Experiences and Theoretical Crossroads*. Cambridge University Press, Cambridge, UK
Wagner P, Wittrock B, Whitley R (eds.) 1991b *Discourses on Society: The Shaping of the Social Science Disciplines*. Kluwer, Dordrecht, The Netherlands
Wallerstein E et al. 1996 *Open the Social Sciences: Report of the Gulbenkian Commission on the Restructuring of the Social Sciences*. Stanford University Press, Stanford, CA

B. Wittrock

Discourse and Identity

The human being is inextricably tied to language. As soon as children are born, a biologically rooted, built-in discourse-developing mechanism is activated enabling them to become communicatively competent, to interact with the world and, as will be discussed in this article, to become aware of their role and specificity in society; that is, of their identity. However, not all discursive sources of identity have the same quality and strength. In general, three major discursive sources of identity may be schematized as an inverted triangle in which a broad section at the top represents 'inherited' sources such as nationality, sex, and race/ethnicity. A narrower section in the middle of the triangle would represent 'optional' identity sources, mainly social groups to which the human being chooses to belong. Finally, a very narrow section at the bottom of the inverted triangle would represent the human self as it is developed and shaped through everyday face-to-face conversations.

In the following three sections these three major identity sources will be discussed. A fourth section will deal with contemporary, transnational communication through the Internet, which actually *inverts the inverted triangle*, providing surprising sources of identity.

1. 'Inherited' Sources of Identity

When a child is born, it inherits several distinctive discursive features such as belonging to a community (e.g., a nation), being male or female, and having distinctive racial or ethnic features. All of these inherited qualities have closely related discursive features (consolidated as the child grows up) ranging from living in a speech community to particular sex- or race-related features.

The role which discourse plays in the community (i.e., the nation in a broad sense) as a source of identity varies depending on its current sociopolitical state, but in all cases it is one of the clearest sources of social identification: experiences within our culture(s) and our community shape human beings and their role in society (see Lanehart 1996). This level of identity relates to Turner's (1987, p. 45) intermediate level of *ingroup–outgroup categorizations* between the broad superordinate level of *being human* and the subordinate level of *personal attributes*.

Needless to say, the sense of *belonging* to a community and *sharing* the same language, a major source of identity, goes beyond the political lines as drawn on maps. This is why the speech community, defined as 'any human aggregate characterized by regular and frequent interaction by means of a shared body of verbal signs and set off from similar aggregates by significant differences in language use' (Gumperz 1971, p. 114) is a much better term, since it goes beyond the reductionist view of society as a clear-cut physical space with members speaking the very same language.

On the other hand, places in which a shared language is embraced as part of a broader resistance against other imposed languages, as is the case of Quebec in Canada or Catalonia in Spain, tend to originate very tight links between discourse and identity, whereas in more politically stable linguistic

areas this source of identity may not be felt so explicitly.

Another major identity marker is sex. There is now a huge amount of bibliography dealing with (non)-verbal speech differences of men and women in same-sex and cross-sex conversational interactions. To a certain extent, fitting the expected discursive attributes of one's sex may reinforce personal identity as a *sense of belonging to a social group* (see West and Zimmerman 1985 for a general overview of the topic). General areas of research include vocabulary differences (including the traditional coining of male-connoted words) and the analysis of conversational *genderlects* (systems of linguistic features linked to one of the sexes).

Finally, the relationship between race, discourse, and identity is important since one's race may be a strong link to society as a multifaceted experience covering such aspects as the recognition of biological origins and, on some occasions, also a deep reinforcement of personality against oppressive, racist attitudes. A related issue has to do with (inherited) ethnic differences within the same community. A clear example is the United Kingdom, in which we can find a great deal of variation between standard and nonstandard forms of the English language, several indigenous languages (Welsh, Scots, Gaelic), and non-English languages brought to Britain by immigrants mainly from countries in the Commonwealth. In all cases, language is raised to a basic source of personal specificity against 'default' speaking correctness.

2. *'Optional' Sources of Identity*

Next in the inverted triangle of identity sources, a middle section of optional sources may be isolated. As human beings grow up, they tend to get together in medium-sized social groups providing an essential sense of belonging and often involving particular discursive features which work as inherent sources of intragroup identity and also as intergroup differentiation (the so-called jargons). A typical example is scientific discourse. Adolescents also display explicitly narrow codes of vocabulary choices and nonstandard pronunciations only available to those 'in the gang' or the group of peers.

Furthermore, laughing is considered to be an important source of identity display in this age group through reducing the 'others' to some laughable characterization that makes them different to us (Boxer and Cortés-Conde 1997, p. 283), normally to enhance a deeper bonding between group members. Finally, in adolescence the strict code of group membership is extended to other codes such as clothing, so that very fixed combinations define 'urban tribes' and provide a supplementive visual channel of intragroup identity.

An important term in this medium-sized area of optional group membership is social network (Milroy 1978), a way of representing people's social transactions within a speech community. This term is essential for understanding the role that discourse plays in the shaping and developing of one's identity. For example, a key issue in the theory of networks is that social groups with tight links of solidarity (*tight-knit networks*) favor the preservation of their vernacular language, a major identity source as commented upon above. In this sense, it should be stressed that there is a difference between vernacular languages, which are learned at home (thus belonging to the broad 'inherited' source of identity) and argots or special parlances, learned after childhood and typically used on those occasions in which group membership is worth stressing. Closely related to this issue are Le Page's (1986) *acts of identity*, a term which stresses the individual's tendency to try and resemble the linguistic attributes of those groups of people which he or she aims to identify with: 'when we talk we project the universe as we see it on to others as on to a cinema screen in our own images, expressed in the language we consider appropriate at that moment, and we invite others by these acts to share our universe' (p. 23).

3. *Discourse, Identity, and the Self*

The speaker's interactional *self* would be placed at the bottom vertex of the inverted triangle which has been suggested to describe different-sized discursive sources of identity. In this case we refer to the person's identity as shaped and modified by daily and ordinary conversational interactions. It is commonly assumed in research on the pragmatics of language use that interlocutors continuously realign their conversational role and either explicitly display their identity or negotiate a *relational identity* with and through others (Boxer and Cortés-Conde 1997, p. 282). Consequently, one's current (and past) conversational exchanges will shape or affect the relationship between participants and speakers' relational self-identity.

Analysis of the self as developed and affirmed through daily interactions goes back as far as the 1950s, with analyses by researchers of the social attributes of exchanges, the so-called *interactional sociolinguists* (Schiffrin 1994), with Goffman (1959) as one of the main contributors to this trend of research. Goffman argues that the self is a social construction (specifically, an *interactive construction*). This is clearly exemplified by the identity-related notion of *face*, defined as the social value that speakers effectively claim for themselves in the course of a conversational interaction. What Goffman focuses on is the social value of conversational involvement, the way different social and conversational settings involve specific expectations and display of this involvement, which

normally entails the use of certain ritualized forms of address. Schiffrin (1994) also includes Gumperz in the label of interactional sociolinguistics because of his view of language as a culturally constructed system and which can be used either in order to reflect macrolevel social meanings or to create microsocial meanings at the level of ordinary conversational exchanges.

Another trend of research which focuses on the social implications of daily conversational interactions is *ethnomethodology*, with Garfinkel as one of its main representatives. This trend focuses on the cognitive rules by which members of a society assess the significance of actions in everyday life. People's previous experience and their knowledge of the institutions and practices of the world around them act to constrain their interpretations of what they see and hear, in somewhat the same way that grammatical rules constrain their perception of sound sequences (Gumperz 1972, p. 215). It is clear that daily interactions play an important role in the shaping of personal identities within a social background. Under ethnomethodology, conversational scenes acquire an essential importance in the display and shaping of self-identities, which are only knowable through the understandings displayed by the interlocutors themselves: 'membership of a category is ascribed (and rejected), avowed (and disavowed), displayed (and ignored) in local places and at certain times, and it does these things as part of the interactional work that constitutes people's lives' (Antaki and Widdicombe 1998, p. 1).

The main conclusion of this research is that people's self-identity is not a fixed feature of their lives, but a dynamic attribute continuously under negotiation, reshaping, transformation, as people engage in daily ordinary communication with others. When some social or cultural aspects are continuously brought into conversations, they tend to become attached to the identity of interlocutors as part of their background of group membership. For example, Georgakopolou (1999) focuses on identities of the young and how *peer networks* are created in the conversational repetition of specific elements in the discursive (sub)culture of the young. As mentioned above, for young people, symbolic cultural associations such as dress styles and others such as music preferences, activity patterns, etc. are firmly located in their conversational interactions, especially in the unfolding of conversational narratives.

On the other hand, self-identity is not only achieved through the *blending* with others through conversational interactions, but also in the feeling of personal specificity when contrasted with others in the course of a conversation. In this case, Davies and Harre's (1990) term *positioning* as 'discursive production of selves' is very interesting. They agree on the interactive formation of self-identity, and that self-identity is not fixed but constituted and reconstituted through the various discursive practices in which people participate. They symbolize this by saying that the human being is an open question with a shifting answer depending on the positions made available by one's own and others' discursive practices and within those practices. Positioning would be the discursive process through or within which people's selves are foregrounded in conversations as observably and subjectively coherent participants in jointly produced story lines. There can be *interactive positioning* in the sense that what one person says positions another, and there can also be *reflexive positioning* when one simply positions oneself.

Markus and Nurius' (1986) theory of *possible selves* also deals with the way people acquire a sense of self-identity through the conversational interactions in which they participate. These possible selves represent the picture that all interactants have about what they are and what they would like to become. In this sense, possible selves influence people's future behavior and are a constant evaluation of people's current state of their self-identity. People's social and cultural contexts play an important role in this shaping, since the range of possible selves derives not only from people's personal experiences (e.g., prior conversational interactions), but also from categories which acquire prominence in a specific sociocultural and historic context, together with images and symbols provided by (highly influential) media discourses constraining the personal choices of self-identity (Lanehart 1996, p. 323).

4. *Inverting the Inverted: Discourse and Identity on the Internet*

So far, three sources of links between discourse and identity have been isolated and represented as an inverted triangle: at the broad top, macrosocial issues such as the speech community, one's sex, and race provide general inherited factors shaping identity from childhood. A middle section of the triangle dealt with optional linguistic groups whose special argots, parlances, and jargons act as a major source of intragroup identification. Finally, the vertex at the bottom of the triangle addressed the personal self as acquiring identity through conversational interactions.

However, these are far from clear-cut, stable features of human identity, but are subject to constant change depending on the current state of sociopolitical relations among people. Specifically, we are now in a *global village* dominated by an increasing quantity and quality of international communications which affect the way the self is viewed and located in an interconnected (i.e., networked) world beyond physical or political boundaries. Castells (1997) specifically deals with this *self/net* dichotomy and how it influences personal identity. The network is inherent to modern, economically advanced societies. The self is a symbol of human activities oriented towards the reaffirmation

of identities in a world like today's, subject to the constant change and instability of networks. In this environment, identity is viewed as 'the process of construction of meaning on the basis of cultural attribute, or related set of cultural attributes, that is/are given priority over other sources of meaning' (Castells 1997, p. 6). The important conclusion is that people tend to react by reaffirming their historical, cultural, ethnic, and religious roots as a reaction against networked globalization.

The most paradigmatic example of a communicative network is the Internet. It is a centerless network linking people from all over the world by an intricate system of (mostly telephonic) wires. It was initially used for military purposes and exploited by academic researchers, but nowadays is has become an inherent source of identity for millions of people. Actually, due to the ubiquity of the Internet, a tendency towards *inverting the inverted triangle* can be observed nowadays.

First, the formerly broad section of inherited attributes at the top of the triangle is narrowed as a consequence of increasing transnational communication and depersonalization of networked transactions. Nowadays, it is easy to exchange messages with people in different parts of the world simultaneously, and therefore national attributes may become blurred and their identity-shaping effect be weakened or invalidated. Other inherited features such as race and sex may be suppressed from daily communicative exchanges through the Internet and may no longer affect the way people are viewed and shaped as interactants in the network. As Meyrowitz (1985, pp. 143–4) points out, 'many categories of people… were once 'naturally' restricted from much social information by being isolated in particular places. The identity and cohesion of many groupings and associations were fostered by the fact that members were 'isolated together' in the same or similar locations…. Now, however, electronic messages…democratize and homogenize places by allowing people to experience and interact with others in spite of physical isolation. As a result, physical location now creates only one type of information system, only one type of shared but special group experience.'

Second, the middle section of optional sources of identity, mainly through self-socialization in groups shaped by specific linguistic patterns, remains more or less in the same medium-sized section of the triangle. On the Internet, these specific discourse groups with their limited-access jargons are reproduced in parallel forms of organization: the so-called *discussion forums* and *distribution lists* via e-mail, which also exploit the same sense of group-membership language specificity that we can find in real-life social instances. The sense of sharing in this type of community also requires that the participants agree on the ideas around which the group is based; even if they disagree, there needs to be some fundamental common ground. Trust in the shared motivations and beliefs of the other participants—in other words, their social identity—is essential to the sense of community (Donath 1999).

Third, the former narrow vertex of personal self in the inverted triangle has now become very broad due to the multiplicity of selves (often anonymous) that may be used in conversational exchanges through various forms of Internet communication such as *chats* or *multiuser domains* (see Turkle 1995). In the physical world the body provides explicit (and stable) clues of inherited and personal qualities shaping the person's identity. On the Internet, however, these attributes are not explicitly conveyed, and one can have as many electronic identities and selves as one is willing and/or able to create, all of which are presented with no corporeality. As Mitchell (1995) correctly states, the process of mutual construction of identities on the Internet usually gives very little away. Because communication takes place without bodily presence or the sound of users' voices, others who 'know' a user quite well may not realize how he or she looks, and thus may be unable to make the usual inferences from that. On the Internet people can very easily conceal, leave carefully ambiguous, or falsely signal gender, race, age, body shape, and economic status. 'People's representation on the Net is not an inevitability of biology, birth, and social circumstance, but a highly manipulable, completely disembodied intellectual fabrication' (Mitchell 1995).

In some cases, real-life and virtual-life selves may overlap, creating, in extreme cases, problems of self-identity when one is unable to dissociate the physical and the virtual. In normal circumstances, though, our on-line selves become just one more series of selves in our everyday lives that we create in order to deal with a variety of situations. The difference between a cyberspace persona and an ordinary workplace persona is that, unlike a workplace persona, the cyberspace persona does not exist in a direct relationship to survival on any level. It is a *surplus personality* (Newitz 1995) but which might occasionally take over from our real-life self.

The medium of identity construction on the Internet is a special kind of text, a hybrid between the permanence and fixed quality of written texts, and the spontaneity and ephemeral status of oral utterances (Baron 2000, Tella 1992). Beyond the evident problems that text-based interlocutors face in the course of a conversation due to the lack of contextual support, the increasing appeal that *chats* and other interactive online phenomena have nowadays should make us think twice before claiming that the oral medium is undoubtedly superior to the written medium. The profusion of written messages that people exchange through mobile phones nowadays, for instance, is a clue suggesting that, perhaps, people are still willing to construct and communicate identities through language, but perhaps not so much through face-to-face encounters.

See also: Discourse, Anthropology of; Ethnomethodology: General; Identity in Anthropology; Identity: Social; Metaphor and its Role in Social Thought: History of the Concept; Narrative, Sociology of; Narratives and Accounts, in the Social and Behavioral Sciences; Networks and Linkages: Cultural Aspects; Self: History of the Concept; Semiotics

Bibliography

Antaki C, Widdicombe S (eds.) 1998 *Identities in Talk*. Sage, London

Baron N S 2000 *Alphabet to Email*. Routledge, London

Boxer D, Cortés-Conde F 1997 From bonding to biting: Conversational joking and identity display. *Journal of Pragmatics* **27**: 275–94

Castells M 1997 *The Information Age: Economy, Society and Culture. Vol. 2: The Power of Identity*. Blackwell, Oxford, UK

Davies B, Harre R 1990 Positioning: The discursive production of selves. *Journal of the Theory of Social Behavior* **20**: 43–63

Donath J S 1999 Identity and deception in the virtual community. In: Kollock P, Smith M (eds.) *Communities in Cyberspace*. Routledge, London

Georgakopolou A 1999 Doing youth in and through conversational narratives. In: Verschueren J (ed.) *Language and Ideology* (*Selected Papers from the 6th International Pragmatics Conference*). IprA, Antwerp

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, New York

Gumperz J J 1971 *Language in Social Groups*. Stanford University Press, Stanford, CA

Gumperz J J 1972 Sociolinguistics and communication in small groups. In: Pride J B, Holmes J (eds.) *Sociolinguistics*. Penguin, Harmondsworth, UK

Lanehart S L 1996 The language of identity. *Journal of English Linguistics* **24**: 322–31

Le Page R B 1986 Acts of identity. *English Today* **8**: 21–4

Markus H, Nurius P 1986 Possible selves. *American Psychologist* **41**: 954–69

Meyrowitz J 1985 *No Sense of Place: The Impact of Electronic Media on Social Behavior*. Oxford University Press, New York

Milroy L 1978 *Language and Social Networks*. Blackwell, Oxford, UK

Mitchell W J T 1995 *City of Bits*. MIT Press (e-book), Cambridge, MA http://mitpress.mit.edu/e-books/City-of-Bits/contents.html (Accessed 30-7-2000)

Newitz A1995 Surplus identity on-line. *Bad Subjects* **16** (on-line document)

Schiffrin D 1994 *Approaches to Discourse*. Blackwell, Oxford, UK

Tella S 1992 The adoption of international communication networks and electronic mail into foreign language education. *Scandinavian Journal of Educational Research* **36**(4): 305–12

Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster, New York

Turner J C 1987 A self-categorization theory. In: Turner J C, Hogg M A, Oakes P J, Reicher S D, Wetherell M S (eds.) *Rediscovering the Social Group: A Self-Categorization Theory*. Blackwell, Oxford, UK

West C, Zimmerman D H 1985 Gender, language, and discourse. In: van Dijk T A (ed.) *Handbook of Discourse Analysis*. Academic Press, London, Vol. 4

F. Yus

Discourse, Anthropology of

Anthropology has been shaped at various points by what has often been referred to as a 'linguistic turn,' reliance on models of language for theories of culture and society. Franz Boas, for example, used phonetics as a model for cultural processes; he suggested that just as each language selected a small range of sounds from a practically unlimited array and arranged them in unique ways, so culture mapped the infinitely varied range of personal experience onto a limited range of categories. This line of reasoning was extended by Edward Sapir and Benjamin Lee Whorf. In the 1960s and 1970s, Claude Lévi-Strauss drew on Roman Jakobson's and Morris Halle's distinctive feature model of phonology in creating a variety of structuralism that revolutionized anthropology, literary criticism, and other fields.

At first glance, the centrality of the concept of 'discourse' to the work of many anthropologists and other scholars in the 1980s and 1990s might seem to represent another linguistic turn, a new convergence of linguistic and social/cultural models. Nevertheless, the situation, at least until recently, has been the opposite. Practitioners influenced by poststructuralist and postmodern theories have generally followed Michel Foucault in 'treating discourse [not] as groups of signs … but as practices that systematically form the objects of which they speak' (1972, p. 49). Invoking the term thus signaled rejecting semiotic and structuralist analyses that identified formal patterns immanent in cultural codes in favor of discerning the power relations that shaped them. Many linguists, on the other hand, were trying to overcome the formalism of structuralism by seeking a more rigorous empiricism, and the concept of discourse was appealing as a rallying cry for going beyond decontextualized analysis of isolated signs to the analysis of communicative acts as they are embodied in concrete contexts. Although the latter practitioners defined discourse—generally implicitly—in a variety of ways, most associated the term with particular texts, conversations, and electronically mediated messages, and analyses generally focused on details of form, content, and context. The result has been the creation of divergent bodies of literature, both of which carry the banner of discourse, that have had relatively little to say to one another until recently.

This article outlines these opposing approaches at the same time that it points to recent research trends that are not only narrowing the gap between the two

types of perspectives but leading to critical reconceptualizations of discourse, language, culture, and other key concepts and producing new insights into social and political life.

1. Discourse in Poststructuralist and Postmodern Perspective

A number of writers turned scholarship on its head in the 1960s and 1970s. Academics had claimed an epistemologically superordinate position, framing their analytic concepts and methods as privileged devices that were not subject to the political and empirical constraints of the phenomena they illuminated. Poststructuralism challenged such claims, placing scholarship squarely into the middle of the politics of social life. Having played a key role—along with transformations in social and material relations—in the creation of modernity, philosophical and academic texts became central objects of analysis. Jacques Derrida suggested that concepts of language, culture, aesthetics, and so forth were in need of 'deconstruction,' of playful and irreverent analyses that unraveled the powerful illusions that imbue dominant frameworks with authority. While Lévi-Strauss characterized culture as based upon unconscious oppositions, Derrida revealed the magical acts that he and other scholars used in creating these sorts of dichotomies and making them seem real.

The influential writings of Michel Foucault (1972) brought the term 'discourse' and a scholarly practice known as 'discourse analysis' or 'the archaeology of knowledge' to the fore in a wide range of disciplines. He suggests that we must examine how 'effects of power circulate among scientific statements, what constitutes, as it were, their internal régime of power' (1980, p. 112). Foucault does not identify discourse with particular utterances or texts, and he argues that it is not possible to grasp the power of discourse when it is seen as rooted in a particular context, viewed as an isolated and self-sufficient totality. He rather suggests that 'a statement always belongs to a series or a whole, always plays a role among other statements, deriving support from them and distinguishing itself from them' (1972, p. 99). One of his major contributions is demonstrating that institutions form key loci for the production and legitimization of discourses, and he identifies the way they are embodied in 'political technologies' that create schemes of surveillance and regulation in such settings as prisons, clinics, and asylums.

Foucault's work has been criticized for its focus on dominant texts, thereby reproducing the marginalization of contributions by women, people of color, and populations outside of Europe. Particularly in his early work, Foucault seems to go against his disavowal of structuralism and his call to look beyond the dispersion and mutability of elements of discourse in favor of finding of 'a unity … [that] resides, well anterior to their formation, in the system that makes possible and governs' (1972, p. 72). Foucault's archeological approach ultimately construes discursive practices as constituting an autonomous sphere that is epistemologically privileged and possesses causal efficacy.

A complementary point of departure was provided by attempts to reverse the marginalization of rhetoric, a step long advocated by Kenneth Burke. In the course of analyzing narrative emplotment, Hayden White suggests that the way historians construct the past involves two steps—an act of imagination that creates a vision of the social world and a subsequent act of rhetorical magic that renders the imaginary real, convincing audiences that that is just the way it happened. Attending to issues of both rhetoric and power, practitioners have critically examined how anthropology similarly contributes to the imagination of what Michel-Rolph Trouillot (1991) refers to as the 'savage slot,' the continual reproduction of modernity's Others. James Clifford and George Marcus (1986) and other writers dissect 'the poetics and politics' of anthropological texts, the rhetorical strategies ethnographers use in imagining 'the field' and imbuing their constructions with authority and authenticity. When viewed critically as power-laden discourse, ethnography became suspect, particularly in the eyes of scholars in cultural and ethnic studies. Unfortunately, classic deconstructions by scholars of color of the rhetorics adopted by ethnographies that stereotyped racialized minorities (see Paredes 1977) were seldom cited as precedents.

Anthropologists have also detailed the way that complex, multifaceted, and contested relations were discursively transformed into seemingly simple dichotomies, such as savage vs. civilized and traditional vs. modern, through colonial discourses. By revealing the complicity of ethnography in colonial regimes and the colonial underpinnings of contemporary anthropology, practitioners have deepened the critique of fundamental concepts and practices within the discipline. Michael Taussig (1987) draws on the work of the Frankfurt School in revealing the way that the production of racialized images went hand-in-hand with the violent oppression of indigenous Latin Americans. Students of colonialism and others have turned to theorists who come from outside the Euro-American orbit, including the Subaltern Studies group in India (Guha and Spivak 1988) and such Latin American theorists as Néstor García Canclini (1995 [1989]) for points of departure.

Finally, a number of scholars have responded to analyses of the 'free floating' character of postmodern discourse by experimenting with techniques of ethnographic analysis and writing. Anthropologists have sought to disrupt dominant rhetorical strategies and political positions by creating ethnographic and other texts that draw on surreal, fictional, and other tech-

niques. The idea that both culture and ethnography are dialogic, which can be traced to M. M. Bakhtin's conception of discourse and Dennis Tedlock's work on narrative, has influenced many of these innovations. Concern with cultural discourses has thus been translated into attempts to transform anthropologists' own discourses of culture.

2. *Linguistic Approaches to Discourse*

Other scholars have used the concept of discourse to grapple with limitations in existing approaches to language. Structural linguists largely followed Ferdinand de Saussure in constructing languages as closed systems of context-independent signs that stand apart from politics, history, and social differences between speakers. Noam Chomsky pushed many linguists toward even more formal, abstract approaches and further marginalized the study of social heterogeneity by positing an ideal speaker–listener as the model of the subject that should be adopted by linguists. An orientation toward discourse promised to free practitioners from focusing exclusively referential relations as encoded in small, decontextualized units in favor of studying the way that texts, conversations, and entire genres were embedded in social life. Rather than seeking to circumvent 'positivistic' concerns with concrete, empirical data, as was the case with many poststructuralist and postmodern scholars, the concept of discourse was seen by linguistic anthropologists and other scholars as a means of achieving greater empirical scope and depth.

Beginning early in the twentieth century, 'Russian Formalists' focused on such discursive units as stories or poems, detailing how writers played with linguistic forms apart from their semantic content (see Jakobson 1960). The Prague School developed a dynamic approach that emphasized expressive, appellative, and aesthetic functions in addition to the referential function and examined how they intersected in different genres. Russian V. N. Volosinov took on Saussure directly, drawing attention to the capacity of signs to link multiple contexts, voices, and texts and detailing the special role of reporting the words of others in creating this diversity. M. M. Bakhtin (1981) characterized discourse as heteroglossic, suggesting that languages are not closed, stable systems but points of intersection between competing structures, styles, genres, and perspectives. Rather than the sources of social consensus—and national unity—that Saussure had envisioned, languages became sites of struggle. British philosopher J. L. Austin (1962) transformed received understandings of the relationship between word and world by arguing that utterances are performative—that they do not simply refer to real world events but create them; saying, in short, is a form of doing. Linguists have mapped the grammatical and 'discourse level' features of texts and genres and tied them to communicative functions, as in Michael Halliday's systemic-functional approach.

In the 1960s and 1970s, American scholars drew on these insights in challenging dominant structuralist and transformational (Chomskyian) approaches. Dell Hymes (1974) called for an ethnography of speaking that documents cultural patterning of language in its social context. While Saussure and Chomsky banished linguistic variation, the way that forms and meanings vary between speakers and contexts, from serious linguistic study, Hymes and other scholars placed it at the center of an emerging sociolinguistics. A generation of practitioners who obtained their Ph.D.'s in the 1960s and early 1970s conducted ethnographies of speaking that focused on such issues as gender, political and religious ritual, rank and status, and expressive genres (see Bauman and Sherzer 1974). Richard Bauman and other writers have illustrated how this approach can be used as a tool for historical analysis. This ethnographic approach was linked directly to a concern with esthetic and performative dimensions of language in a focus on performances, speech events that are marked by special forms of authority, and formal elaboration (Bauman 1977, Hymes 1975). At the same time, Joshua Fishman (1972), John Gumperz (1982), and William Labov (1972) drew attention to the linguistic and political importance of relationships between dialects, languages, and discourse strategies. The frameworks they presented helped to focus attention on linguistic diversity and language policies in contemporary nation-states, presaging later interest in discursive dimensions of nationalism, power, conflict, and social inequality. Michael Silverstein (1976) drew on the work of Charles S Peirce and Roman Jakobson in developing a 'metapragmatic' approach that complements the study of semantics with rigorous attention to the centrality of indexes that anchor discourse in existing social contexts—and create new contextual parameters. All of these approaches placed questions of communicative functions—the multiple roles of language in shaping social life—alongside those of structure at the center of inquiry.

Other scholars working in the United States sought to draw the study of discourse away from the study of types, rules, norms, and functions; they defined discourse as—and limited objects of analysis to—particular stretches of talk. Sociologists Harold Garfinkel and Erving Goffman argued that discourse should be studied not as an end in itself but as a window on social interaction. Garfinkel's concern with everyday, mundane encounters was placed at the center of the field of conversation analysis, which proclaimed 'an insistence on the use of materials collected from *naturally occurring* occasions of everyday interaction by means of audio- and video-recording equipment or film' (Heritage and Atkinson 1984, p. 2; emphasis in original). Eschewing reference to contextual information that does not appear in the

recording itself, conversation analysts scrutinize transcripts for regularities in the sequencing of conversational features. Many practitioners now focus on institutional contexts, detailing the role of interaction in shaping what takes place in courtrooms, schools, clinics, and other settings. Joel Sherzer proposes a 'discourse-centered' framework that similarly privileges 'naturally occurring and recorded speech' (1983, p. 10) at the same time that he calls for systematic attention to broader social and cultural contexts. These approaches represent an opposite pole from poststructuralist frameworks in that they view discourse in maximally empirical terms and focus on minute analysis of recorded talk.

3. *Converging Paths: The Politics of Discourse*

In recent years, many anthropologists and other scholars have focused on the role of discourse in shaping power and the politics of difference, resulting in a number of important intersections between practitioners who formerly held opposing views of discourse.

Pierre Bourdieu (1991) placed sociolinguistic concepts into a framework that defined types of linguistic competence as 'symbolic capital' in seeking to reveal the role of discourse in processes of 'symbolic domination.' He redefined the concept of communicative competence in such a way as to draw attention to the barriers of class and race that limit access to sites in which such competence can be acquired. Aaron Cicourel, Charles Briggs, Elliot Mishler, and others have focused on interviews as social as well as communicative events, detailing how formal and contextual features shape and are shaped by power asymmetries, thereby creating discourses that scholars use in imagining social life and imbuing their texts with authority and authenticity.

Building on Silverstein's (1979) foundational work, linguistic anthropologists have explored the role of language ideologies, socially situated conceptions of the nature and functions of language, speech, and communication, in shaping both linguistic structure and social life (see Schieffelin et al. 1988). Challenging the notion that linguistics provides a privileged and at least ideally objective understanding of language, investigations of language by linguists, philosophers, missionaries, and others come to be seen as being just as socially situated and politically interested as other social constructions. Bauman and Briggs (2000) argue that language ideologies have played a key role in shaping modernity, having been used in constructing and naturalizing political ideologies and institutions since the seventeenth century. Such ideologies are tied in crucial ways to metadiscursive practices that provide the means for attempting to control the production, circulation, legitimization, and reception of discourse (see Silverstein and Urban 1996).

Anthropologists responded to Jurgen Habermas's work on the construction of a 'public sphere' in which citizens debate discourses that shape social life and political institutions by detailing the way that dominant forms and arenas are contested by resistant voices and counterpublics. Such investigations have helped to illuminate the processes by which limits are imposed on what sorts of things can be said and who can say them. Research on multilingualism and language policy similarly have provided crucial insights into the way that efforts to legislate the dominance of national languages and standard dialects provide politically palatable means of legitimizing forms of social inequality based on class, race, and immigration status (see Hill 1998, Zentella 1997). Work by US based linguists parallels European critical discourse analysis (Wodak and Reisigl 1999) and British concern with language and power, which have also scrutinized the discursive foundations of racism, anti-immigrant rhetorics, and other ideologies of inequality. Approaches have similarly converged in analyzing the discursive underpinnings of science and public health. At the same time that this work has helped to transform Foucault's call to scrutinize how discourses 'systematically form the objects of which they speak,' it has provided tools used by specialists in cultural, ethnic, women's, and science studies in revealing how structures of subordination and marginalization are produced and maintained.

The study of discourse has come under attack of late. Martha Nussbaum criticizes Judith Butler (1997), for example, for 'turning from the material side of life, toward a type of verbal and symbolic politics that makes only the flimsiest of connections with the real situation of real women' (Nussbaum 1999, p. 38). Such criticisms seek to reduce these many types of discourse analysis to explorations of ideological—and therefore ultimately inconsequential—dimensions of social life, thereby suggesting that such approaches cannot illuminate inequalities of power and material resources. Responding effectively to such criticisms will require new syntheses of sociopolitical and linguistic approaches in demonstrating the materiality of discourse and the way it, along with other symbolic (including material) forms, shapes social life.

See also: Communication: Philosophical Aspects; Critical Theory: Contemporary; Discourse and Identity; Ethnomethodology: General; Linguistic Turn and Discourse Analysis in History; Political Discourse; Postmodernism in Sociology; Postmodernism: Philosophical Aspects

Bibliography

Austin J L 1962 *How to Do Things with Words*. Harvard University Press, Cambridge, MA

Bakhtin M M 1981 [trans. Emerson C, Holquist M] In: Holquist M (ed.) *The Dialogic Imagination: Four Essays*. University of Texas Press, Austin, TX, Chap. 4
Bauman R 1977 *Verbal Art as Performance*. Newsbury House, Rowley, MA
Bauman R, Briggs C L 2000 Language philosophy as language ideology: John Locke and Johann Gottfried Herder. In: Kroskrity P V (ed.) *Regimes of Language: Ideologies, Polities, and Identities*. School of American Research Press, Santa Fe, NM, pp. 139–204
Bauman R, Sherzer J (eds.) 1974 *Explorations in the Ethnography of Speaking*. Cambridge University Press, Cambridge, UK
Bourdieu P 1991 *Language and Symbolic Power* [trans. Raymond G, Adamson M]. Harvard University Press, Cambridge, MA
Butler J 1997 *Excitable Speech: A Politics of the Performative*. Routledge, New York
Clifford J, Marcus G E (eds.) 1986 *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA
Fishman J A 1972 *The Sociology of Language: An Interdisciplinary Social Science Approach to Language in Society*. Newbury House, Rowley, MA
Foucault M 1972 *The Archaeology of Knowledge* [trans. Sheridan A M]. Pantheon Books, New York
Foucault M 1980 *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977* [trans. Gordon C et al.]. Pantheon, New York
Garcia Canclini N 1995[1989] *Hybrid Cultures: Strategies for Entering and Leaving Modernity* [trans. Chiappari C L, López S L]. University of Minnesota Press, Minneapolis, MN
Guha R, Spivak G C (eds.) 1988 *Selected Subaltern Studies*. Oxford University Press, New York
Gumperz J J 1982 *Discourse Strategies*. Cambridge University Press, Cambridge, UK
Heritage J, Atkinson J M 1984 Introduction. In: Atkinson J M, Drew P (eds.) *Structures of Social Action*. Cambridge University Press, Cambridge, UK, pp. 1–15
Hill J H 1998 Language, race, and white public space. *American Anthropologist* **100**: 680–9
Hymes D 1974 *Foundations in Sociolinguistics: An Ethnographic Approach*. University of Pennsylvania Press, Philadelphia, PA
Hymes D 1975 Breakthrough into performance. In: Ben-Amos D, Goldstein K S (eds.) *Folklore: Performance and Communication*. Mouton, The Hague, pp. 11–74
Jakobson R 1960 Closing statement: Linguistics and poetics. In: Sebeok T A (ed.) *Style in Language*. MIT Press, Cambridge, MA, pp. 350–77
Labov W 1972 *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, PA
Nussbaum M C 1999 Excitable speech: A politics of the performative by Butler J. *The New Republic* 22 February: 37–45
Paredes A 1977 On ethnographic work among minority groups: A folklorist's perspective. *New Scholar* **7**: 1–32
Schieffelin B, Woolard K A, Kroskrity P V (eds.) 1988 *Language Ideologies: Practice and Theory*. Oxford University Press, New York
Sherzer J 1983 *Kuna Ways of Speaking*. University of Texas Press, Austin, TX
Silverstein M 1976 Shifters, linguistic categories, and cultural description. In: Basso K, Selby H A (eds.) *Meaning in Anthropology*. University of New Mexico Press, Albuquerque, NM, pp. 11–55
Silverstein M 1979 Language structure and linguistic ideology. In: Clyne P R, Hanks W F, Hofbauer C L (eds.) *The Elements: A Parasession on Linguistic Units and Levels*. Chicago Linguistic Society, Chicago, pp. 193–247
Silverstein M, Urban G (eds.) 1996 *Natural Histories of Discourse*. University of Chicago Press, Chicago
Taussig M 1987 *Shamanism, Colonialism and the Wild Man: A Study in Terror and Healing*. University of Chicago Press, Chicago
Trouillot M-R 1991 Anthropology and the savage slot: The poetics and politics of otherness. In: Fox R G (ed.) *Recapturing Anthropology: Working in the Present*. School of American Research, Santa Fe, NM, pp. 17–44
Wodak R, Reisigl M 1999 Discourse and racism: European perspectives. *Annual Review of Anthropology* **28**: 175–99
Zentella A C 1997 The Hispanophobia of the Official English movement in the U.S. *International Journal of the Sociology of Language* **127**: 71–86

C. L. Briggs

Discovery Learning, Cognitive Psychology of

1. Discovery Learning

Like most of the terms and catchphrases that provoke controversy in education, 'discovery learning' is marvelously vague. A plausible one-sentence definition would simply elaborate on the name itself to produce something like the following: discovery learning is a philosophy of education in which students discover ideas and knowledge via exploration, projects, activities, problem solving, and play. As such, discovery learning is usually contrasted with a style of education which might (depending on the background or predilections of the person doing the contrasting) be referred to as instructionist, regimented, orderly, traditional, or efficient—i.e., a style of education focusing on the transfer of knowledge via lessons, textbooks, lectures, and direct instruction.

More realistically, the phrase 'discovery learning' denotes, as much as anything, a style or aesthetic of education. Broadly speaking, advocates of discovery learning tend to invoke visual images of children learning in playgrounds, gardens, laboratories, workshops, and studios; the opposing visual images are those of children seated at rows of desks, listening to a teacher, reciting in unison, or writing in lockstep. These contrasting visual images are important because they encapsulate the concerns, and anticipate the rhetoric, of both the proponents and opponents of discovery learning. Proponents focus on notions such as motivation, depth and authenticity of experience, and respect for individual learning styles; opponents focus on notions such as rapid skill acquisition,

efficiency of learning, and quantitative assessment of learning outcomes. The arguments between the two camps, then, are in large part rooted in a strong division between two venerable notions of education—one which focuses on the cultivation of deep individual interests, and one which focuses on mastering fundamental skills as the basis for productive citizenship.

This article focuses on discovery learning as an object of study in cognitive science. As such it does not, and could not, resolve the philosophical debate over the value of discovery learning; to do so would be to resolve far deeper questions about what the essential purpose of education might be. Nonetheless, a cognitive science perspective can inform this ongoing debate in several ways: in particular, cognitive theory suggests where the strengths or pitfalls of discovery learning may lie, whether there are settings and domains for which it might best be suited, or what sorts of research and development efforts might improve its practice. The following sections explore these issues. By way of preface, Sect. 2 begins by describing the role of discovery learning in the history of science and mathematics education. Section. 3 then focuses on those particular issues for discovery learning that are highlighted by cognitive science research; and Sect. 4 discusses the role of educational technology in the practice (both current and future) of discovery learning.

2. *Discovery Learning in Science Education: From Ancient History to Recent Debates*

Images of discovery learning have an ancient pedigree in educational writing. Indeed, one could argue that Socrates' mathematics lesson in Plato's *Meno*—perhaps the quintessential classical portrait of pedagogical success—is an instance of 'guided discovery learning' in that the slave boy (student) is brought to understand a special case of the Pythagorean theorem through his own responses rather than direct instruction. (Undoubtedly, this instance of discovery learning is very energetically guided.) Elements of discovery learning can be found in the writings of the seventeenth-century Czech religious educator John Comenius:

> Men must, as far as possible, be taught to become wise by studying the heavens, the earth, oaks, and beeches, but not by studying books; that is to say, they must learn to know and investigate the things themselves … (N)o information should be imparted on the grounds of bookish authority, but should be authorised by actual demonstration for the senses and to the intellect. (Rusk 1967, p. 97)

A century later, Jean-Jacques Rousseau's educational classic *Emile* includes similar passages: 'Teach your scholar to observe the phenomena of nature … Put the problems before him and let him solve them himself. Let him know nothing because you have told him, but because he has learnt it for himself. Let him not be taught science, let him discover it' (Rousseau 1992, p. 131). Rousseau's ideas were taken up in the nineteenth century by the Swiss educator Johann Pestalozzi, and by Friedrich Froebel, the innovative German educator famous for his development of the idea of a 'kindergarten' (Brosterman 1997, pp. 18–22, 30–5).

In the past century, the central images of discovery learning are prominently found in the writings of Dewey, particularly in his descriptions of science education (cf. Dewey 1916, pp. 266–76). Piaget's work is likewise often cited as providing a theoretical foundation in cognitive psychology for discovery learning: the Piagetian child is one who (in the apt summary of Gardner et al. 1996, p. 28) 'would first learn about the world directly through spontaneous and natural actions upon the physical world; then, acquiring a set of more complex cognitive structures in a preordained fashion, that child could perform "mental operations" of increasing abstraction and power on the representations that she had constructed.' Perhaps the most eloquent advocate for the role of discovery in education is Bruner, whose essay 'The act of discovery' (Bruner 1979) may be read as a sort of manifesto for the entire discovery learning enterprise. In this essay, Bruner hypothesizes four broad benefits that a student might derive from learning via discovery: an increase in intellectual potency (i.e., development of stronger techniques for problem solving); a shift from extrinsic to intrinsic rewards (i.e., a shift to motivation derived from the act of discovery itself); learning of heuristics of discovering; and improved memory for learned material (since that material was learned in a rich context developed by the student themselves).

After centuries of such energetic and passionate advocacy, one might expect that discovery learning would be a relatively uncontroversial notion in education; but the history of education shows that uncontroversial notions are few and far between. A representative critique of discovery learning may be found in Hirsch (1996, p. 250), who presents a skeptical one-paragraph definition of the concept, noting that '(T)here are two serious drawbacks to preponderant or exclusive reliance on discovery learning. First, students do not always make on their own the discoveries that they are supposed to make … Second, discovery learning has proved to be very inefficient.' This theme of the pedagogical uncertainty of the method is described as well by Driver (1994, p. 42): '(T)he connections that are apparent to a scientist may be far from obvious to a pupil.' Even among educators sympathetic to discovery learning, pitfalls are noted. Papert (1993, p. 16), for example, sounds the important theme of context: 'It is simply double-talk to ask children to take charge of their own learning and at the

same time order them to 'discover' something that can have no role in helping them understand anything they care about or are interested in or curious about.'

3. *Discovery Learning and Cognitive Science*

The preceding paragraphs leave us with a frustrating problem. On the one hand, learning via discovery would seem to be desirable on affective and intellectual grounds; on the other hand, the difficulties of practical implementation of discovery learning are widely acknowledged. One question that we might begin with, then, would be whether there is evidence that children (as opposed to, say, professional scientists) are indeed capable of effective learning via discovery.

One body of relevant evidence comes from recent work in infant cognition. The tasks with which babies are confronted—to learn the grammar of their natural language, to learn the behaviors of objects in the world, to learn about their own and other people's minds—seem, upon examination, to be approached by persistent and at least somewhat systematic inquiry. Gopnik et al. (1999, p. 86) summarize this research, writing:

> Human children in the first three years of life are consumed by a desire to explore and experiment with objects. In fact, we take this for granted as a sometimes exhausting fact of parenting … (F)or our species the dangers of exploration are offset by the benefits of learning. The rapid and profound changes in children's understanding of the world seem related to the ways they explore and experiment.

It appears, then, that successful discovery learning does occur, given the right conditions. But note that infant cognition may be special in a number of respects: first, the domains in which discovery takes place for infants (language, object recognition) are most likely domains in which a great deal of inborn cognitive and perhaps neurological structure already exists prior to exploration. Second, as Gopnik et al. point out, the infant's environment is interestingly poised to reward infant experimentation in the form of adult behavior: adults, for example, ease the learning of language through the use of 'motherese,' with its short, simple sentences and affectionate intonations (Gopnik et al. 1999, pp. 128–32).

The problems with discovery learning in science—the most typical classroom domain—are more pointed, and form an interesting contrast to the successful learning scenarios of infant cognition. Children (and adults) come to scientific learning with many misconceptions, some of them arguably derived from the very same natural explorations they performed as infants (e.g., the notion that inanimate moving objects inevitably 'run down,' Driver et al. 1985). Our everyday environment is not especially well suited to disclosing phenomena that occur in the absence of friction (such as Newtonian mechanics), or over long periods of time (such as evolution), or at high speeds (such as relativistic effects) or small scales (such as quantum effects), making scientific subjects difficult to teach and understand, even for adults. Moreover, as discussed by Kuhn et al. (1988, p. 220), key aspects of sustained scientific inquiry (e.g., maintaining a conscious distinction between evidence and theory) are difficult for both children and adults.

4. *Directions for Discovery Learning*

Despite the difficulties involved in getting it right, the benefits that accrue from providing students with opportunities for discovery seem profound, not least on the affective dimension. Csikszentmihalyi, who has done extensive empirical work on motivation and creativity, writes:

> When people are asked to choose from a list the best description of how they feel when doing whatever they enjoy doing most—reading, climbing mountains, playing chess, whatever—the answer most frequently chosen is 'designing or discovering something new.' … (U)nless enough people are motivated by the enjoyment that comes from confronting challenges, by discovering new ways of being and doing, there is no evolution of culture, no progress in thought or feeling. (Csikszentmihalyi 1996, pp. 108–10)

Csikszentmihalyi's admonition suggests that one promising direction of research in discovery learning is to make it more motivating and enjoyable for students, perhaps by looking to the sorts of factors that encourage discovery and exploration in infants and toddlers. These factors include increased attention to (a) the nature of students' preexisting theories and to their skills and predilections for theory building; (b) the arrangement of learning environments, both physical and social, conducive to enjoyable discovery; and (c) the design of novel artifacts whose purpose is to encourage and reward sustained periods of exploration and discovery.

Many provocative recent examples of research in this last direction may be found in the development of new technologies for education. The advent of numerous computer 'microworlds' (Papert 1980), in which simulations or experimentation can take place under conditions difficult to achieve in the 'real' world of objects, is one notable development in the past two decades. More recent developments in the integration of physical and computational elements to produce (for example) new types of 'digital manipulatives' (Resnick et al. 1998, Eisenberg and Eisenberg 1999) suggest new and uncharted vistas for creating 'discovery-friendly environments.' Such environments, to be pedagogically successful, will need to be designed with an eye both toward cognitive research in students' scientific understanding, and toward supportive social structures (e.g., for adult mentoring, for apprenticeship, for collaboration). Nonetheless, despite the practical obstacles, there is still life in the centuries-old dream of learning via discovery.

See also: Education and Learning: Lifespan Perspectives; Explanation-based Learning, Cognitive Psychology of; Learning Curve, The; Learning from Text; Scientific Reasoning and Discovery, Cognitive Psychology of; Self-regulated Learning; Teaching and Learning in the Classroom; Transfer of Learning, Cognitive Psychology of

Bibliography

Brosterman N 1997 *Inventing Kindergarten*. Harry Abrams, New York

Bruner J 1979 *On Knowing*. Belknap Press, Cambridge, MA

Csikszentmihalyi M 1996 *Creativity*. Harper Collins, New York

Dewey J 1916 *Democracy and Education*. Free Press, New York

Driver R 1994 The fallacy of induction in science teaching. In: Levinson R (ed.) *Teaching Science*. Routledge, London, pp. 41–8

Driver R, Guesne E, Tiberghien A (eds.) 1985 *Children's Ideas in Science*. Open University Press, Philadelphia

Eisenberg M, Eisenberg A 1999 Middle tech: Blurring the division between high and low tech in education. In: Druin A (ed.) *The Design of Children's Technology*. Morgan Kaufmann, San Francisco, pp. 244–73

Gardner H, Torff B, Hatch T 1996 The age of innocence reconsidered: Preserving the best of the progressive traditions in psychology and education. In: Olson D, Torrance N (eds.) *The Handbook of Education and Human Development*. Blackwell, Malden, MA, pp. 28–55

Gopnik A, Meltzoff A, Kuhl P 1999 *The Scientist in the Crib*. William Morrow, New York

Hirsch E D 1996 *The Schools We Need*. Doubleday, New York

Kuhn D, Amsel E, O'Loughlin M 1988 *The Development of Scientific Thinking Skills*. Academic Press, San Diego, CA

Papert S 1980 *Mindstorms*. Basic Books, New York

Papert S 1993 *The Children's Machine*. Basic Books, New York

Resnick M, Martin F, Berg R, Borovoy R, Colella V, Kramer K, Silverman B 1998 Digital manipulatives: New toys to think with. *Proceedings of CHI '98*, Association for Computing Machinery (ACM) Special Interest group on Computer-Human interaction (SIGCHI), ACM, New York, pp. 281–7

Rousseau J-J 1992 *Emile*. Charles E. Tuttle, Rutland, VT

Rusk R 1967 *The Doctrines of the Great Educators*. St. Martin's Press, New York

M. Eisenberg

Discovery: Legal Aspects

Discovery is the legal process for compelling production of information relevant to disputed factual issues in litigation. While discovery can take place in both civil and criminal cases, this article will focus on the civil side. This article briefly outlines the purposes of pretrial discovery. It then describes the basic regimes of pretrial civil discovery in the United States and England, the two leading common law jurisdictions, reviews some data on their performance, and contrasts their approach with analogous civil law institutions.

1. The Aims of Discovery

In common law systems, discovery is viewed as too complex and costly for small claims and ordinarily is unavailable in those settings. In larger cases, discovery normally occurs after the parties have filed their initial pleadings, which make clear the general character of the dispute and, in some cases, the main points of contention. If a claim or defense is legally invalid on its face, the court can resolve the dispute at that point. But if the dispute concerns an issue of fact, the pleadings will not suffice. The ultimate test of disputed facts is a trial, but trials are expensive both for the parties and for the legal system.

Discovery can improve the handling of factual disputes by making each party disclose, before trial, the evidence it plans to offer. Such disclosures can make trial more efficient by clarifying disputed issues, permitting the parties to edit out irrelevant or redundant evidence, and reducing unfair surprise and consequent mistrials or adjournments.

Discovery can also make trials more accurate by compelling production of information that otherwise would not be presented because it is harmful or embarrassing to the person who controls it. Greater accuracy might appear noncontroversial, since it benefits right-doers and harms wrong-doers. But if a class of plaintiffs or defendants tend to control a disproportionate share of relevant information, accuracy-promoting discovery may enable new types of claims and defenses that otherwise would fail outright. For example, although price fixing schemes are unlawful, it may be difficult to prove their existence without compelled discovery of defendants' business records (James et al. 1992, p. 236).

Discovery may also improve the quality of public and private decisions about whether there should be a trial. It may demonstrate that a factual contention is doomed to fail at trial and can be dismissed summarily. The additional information disclosed (and the costs of disclosure) may also influence the likelihood and terms of settlement.

2. The Law of Discovery

The United States and England differ in their approach to pretrial discovery. For ease of exposition, the discussion of American law is based exclusively on the Federal Rules of Civil Procedure, which apply only in federal trial courts but have heavily influenced state court practice. The discussion of English law is based exclusively on the Civil Procedure Rules and Practice Directions, which apply in the High Court and county courts.

2.1 Scope and Limitations

A discovery system which aimed only at preventing surprise could limit its scope to evidence that the

parties intend to present at trial. For example, with respect to testimonial evidence, English law requires only that parties prepare and submit statements for each witness whom they intend to call at trial summarizing the proposed testimony. (Civil Procedure Rules, Part 32.4). A system that aims for improved accuracy needs a broader principle. England adopts such a principle for documentary evidence, setting the scope of discovery as materials 'relating to the matters in question in the action,' which are then defined in terms of their tendency to prove or disprove the claim of any party. (Civil Procedure Rules, Part 31.6.) American discovery extends even further. It reaches 'any matter … relevant to the claim or defense of any party.' But unlike English law, the information sought need not have probative weight, provided it 'appears reasonably calculated to lead to the discovery of admissible evidence' (Federal Rules of Civil Procedure 26 (b) (1)). Thus, under American law, discovery can be used to identify and pursue investigative leads. In England and all other common law jurisdictions, such discovery is rejected as an improper 'fishing expedition' (Jolowicz 2000, pp. 41–5).

Under both English and American law, some relevant information is shielded from discovery. Information within a recognized evidentiary privilege (such as the attorney–client privilege) is immune. Both systems recognize a complete or qualified privilege for materials prepared by attorneys and others in anticipation of litigation. Finally, each system has recently adopted a principle holding roughly that the scope and costs of discovery should be in proportion to the value of the information to be produced. Concerns about commercial confidentiality and invasion of privacy, though frequently raised, rarely operate as a complete bar to otherwise proper disclosure. Instead they are addressed through restrictions on access to and use of the information disclosed.

2.2 *Mechanisms for Obtaining Discovery*

The only significant form of discovery in English law is disclosure of documents. In cases valued at more than £5,000, each party is routinely ordered to provide a list of all relevant documents in its possession, custody, or control. The party must then produce listed documents on request unless the document is privileged or the burden of producing it would be disproportionate. A party may also file a request with the court for additional document discovery either from another party or from a nonparty. The party seeking discovery must specify the documents sought, demonstrate that they are relevant, and show how their discovery will advance the case. Discovery with respect to witnesses is ordinarily limited to the provision of the summaries of testimony noted above. The deposition, an interview of a witness, taken under oath and stenographically recorded, is not available for discovery, but only to preserve the testimony of a witness who cannot attend trial (Civil Procedure Rules, Part 34).

In United States courts, the obligation to make discovery was traditionally contingent upon receipt of a proper demand. But recent amendments oblige each party to provide certain information without request at the outset of the action. A party must identify each person likely to have 'discoverable information that the disclosing party may use to support its claims or defenses.' They must also provide copies or descriptions of documents and other tangible things in its possession, custody, or control that it may use for that purpose. Prior to trial parties must identify trial witnesses and exhibits (Federal Rule of Civil Procedure 26 (a)).

Discovery beyond the initial disclosures can be pursued through multiple devices. Broadly framed document demands and depositions both may be used, without leave of court, to seek any information falling within the scope of discovery against any party. Even more striking, if the document or deposition discovery is properly joined with a subpoena (which is also available without leave of court), these devices may be used against any nonparty as well. Other discovery devices, including interrogatories, court-ordered physical examinations, and requests for admission, are available only against parties. The most important in practice are interrogatories, written questions which must be answered in writing and under oath. Because the response must include all information available through reasonable investigation, preparing an answer may involve significant data gathering activities.

2.3 *Costs*

The costs involved in discovery include lawyers' fees, the time of the responding parties and witnesses, and intangible costs from invasions of privacy and the frictions of adversary litigation. On the public side, costs include judicial time spent in supervising discovery. For out-of-pocket costs such as attorneys' fees, both the English and American systems have generally followed the standard cost allocation rules applicable for civil litigation. In England that involves awarding the winner's reasonable costs of discovery against the loser. In America, costs are normally borne by the party who incurred them, with certain exceptions for costs that result from a party's improper use of, or improper opposition to, a discovery request.

2.4 *Sanctions and Judicial Supervision*

Discovery can fail in various ways. Deliberate withholding of properly requested information and inten-

tional use of discovery to inflict financial or emotional harm are illegal. But separating intentional misconduct from the assertion of legitimate legal rights is difficult. There are often nonfrivolous legal justifications for discovery conduct that deprives the tribunal of critical information or imposes costs on an opponent, and taking incidental advantage of an opponent's lack of perspicacity or resources is a well-accepted feature of adversarial ethics. The resulting strategic interaction may lead to underproduction of information or to an escalating cycle of demands, resistance, and disputes (Brazil 1979). The scope of American discovery has made these issues urgent in the United States.

For much of its modern history, discovery has been largely party initiated and controlled. The judge was to be involved only in the event of a dispute concerning the propriety of a request or the adequacy of a response. Compliance was to be assured by the professionalism of the lawyers involved and the threat of judicial sanctions. Available sanctions included monetary penalties, contempt, orders deeming certain matters to have been established against the offending party, and default judgments.

Many American lawyers and judges have come to believe that this essentially reactive enforcement model allows too much abuse and waste. Changes in the legal profession have reduced the potential for easy cooperation between counsel (Gilson and Mnookin 1994). Judges feel they have more important work than to master the details of discovery disputes. Sanctions are an imperfect tool, not draconian enough to eliminate temptations to withhold critical evidence in high stakes cases, but too cumbersome to address the routine tactical jostling that characterizes much discovery.

These concerns have led to reforms designed to cut back on the scope of discovery and to increase judicial control over its initiation. In both England and the United States, reforms have narrowed discovery's scope, placed renewed emphasis on the concept of proportionality between benefits and costs, and, in England, eliminated some types of discovery altogether. In both countries, the occasions on which advance judicial approval must be sought have been broadened. Finally, in both countries there is an increased emphasis on proactive judicial supervision and participation in the discovery planning process (Marcus 1999, Wolff 1998, Jolowicz 2000).

3. *Effectiveness of Discovery*

For many years there was a widespread view, based largely on anecdotal accounts, that American-style discovery generally tends to be unduly costly in proportion to its social benefits. Repeated empirical studies have demonstrated that this view is at best overly simple (Willging et al. 1998, Kakalik et al. 1998, McKenna and Wiggins 1998). Those studies show that even in the categories of civil litigation where discovery is available and in common use, a substantial minority, approaching 40 percent, of filed cases are resolved without any formal discovery, typically through abandonment or settlement. In cases where formal discovery occurs, it typically involves only a few discovery events and moderate costs (measured primarily in terms of lawyer time expended). Finally, in a small percentage of cases, typically involving greater factual complexity and higher stakes, there are more discovery events and much higher discovery costs. Even in this class of cases, costs remain roughly proportional to the amount in controversy. Lawyers believe that in most cases the discovery process works tolerably well, generating about the right amount of information given the needs of the case and at a level of expenditure that is about right in relation to the stakes (Willging et al. 1998, pp. 550–2).

Within this overall pattern, however, there are more or less severe problems of withholding, delay, or waste in a substantial proportion of cases where formal discovery occurs. By some estimates, nearly 20 percent of discovery costs are attributable to such problems (Willging et al. 1998). Some evidence suggests that problems are disproportionately encountered in high stakes, complex, and contentious litigation (Brazil 1980a, 1980b). But others argue that the higher incidence of reported problems in such cases may reflect the larger number of discovery events and the higher proportion of litigation expenditures devoted to discovery, rather than any underlying pathology (Willging et al. 1998). Attempts to moderate the costs of discovery in such cases through judicial management have thus far met with very limited success (Kakalik et al. 1998).

While these studies show that broad discovery is not clearly a failure, they do not demonstrate its success, for several reasons. First, their focus on lawyer time as the measure of cost leaves out costs to clients, nonparty witnesses, and the court system itself. Second, the focus on formal discovery overstates some marginal effects of the discovery rules and omits others. Some information exchanged in formal discovery would be shared voluntarily even if there were no discovery rules, simply to persuade the opponent to abandon its claim or settle (Hay 1994). Neither the production of this information nor the costs of producing it are due to the discovery rules. On the omission side, the studies ignore effects that *prospective* application of the discovery rules may have on filed and unfiled claims resolved without formal discovery but 'in the shadow' of those rules.

Third, current studies say little about how the marginal effects of discovery rules actually influence litigation outcomes and behavior. Are outcomes more accurate? Or are they less accurate, for example because the prospective costs of discovery discourage prosecution or defense of some claims? (Brazil 1980a,

at 225). What is discovery's impact on the likelihood of settlement and how does it generate that impact? On this issue, the theoretical literature suggests several possibilities (Cooter and Rubinfeld 1994, Hay 1994), but the empirical literature is silent.

Finally, the studies do not explore discovery's impact on real world conduct. The benefits of more accurate adjudication may differ dramatically for different types of cases, and they may be very imperfectly correlated with the practicing lawyer's conventional measures of a dispute's importance, such as the amount in controversy (Hay 1994, Kaplow 1994, Bundy 1994).

4. *Comparison to Civil Law Approaches*

Civil law jurisdictions have much less discovery. The German system is a helpful example, since it is not atypical and has heavily influenced other systems, including Japan, Korea, and Brazil. In the German system, a judge or panel of judges acts as the trier of fact. Party requests for information occur during what amounts to an ongoing hearing and are directed, not to the opposing party, but to the trial judge (Langbein 1985).

German hearings proceed primarily on the basis of the witnesses and documents identified in the parties' preliminary submissions, which generally do not identify much information harmful to a party's case. A party can ask the judge to require the production of additional evidence only within limited categories. A party is entitled to all documents referred to in the opposing party's submissions and to documents to which he has a right of access under the substantive law (typically because of a contract or fiduciary relationship between the parties) (Gerber 1986). Even within these categories, the party must make a stronger demonstration of relevance than in American law: the information must be reasonably likely to influence the court's determination of a disputed factual issue and described with reasonable particularity. Beyond these categories of information, there is no discovery.

Differences in discovery regimes have produced significant conflict between the United States and foreign states with more restrictive discovery practices. Most of the issues have arisen from efforts to take extensive American-style discovery from defendants who are located in civil law jurisdictions. Some of these issues are now the subject of international agreements, such as the Hague Convention on the Taking of Evidence (Hazard 1998).

Differences between common law and civil law discovery reflect contrasting arrangements for trial, conceptions of adjudication, and commitments to party control of litigation. The common law's concern for preventing surprise is not matched in civil law systems, whose professional triers of fact exercise greater control over the presentation of evidence, are less likely to be influenced by surprise, and can more readily adjourn to permit rebuttal evidence (Langbein 1985). American discovery came of age when civil litigation was increasingly seen as a law enforcement device, in which 'private attorneys general' play an important role (Scott 1975). Conversely, civil law's 'relative indifference to the completeness of the evidentiary material submitted to the court' (Damaska 1997, p. 114) sees adjudication more as resolving disputes on the parties' terms. Finally, in America, party initiation and control of discovery at least initially sidestepped the steep administrative and political costs of an intrusive judge-directed inquiry into private conduct. In civil law systems, the expectation that expanded discovery would be channeled through overworked career judges—while consistent with civil law traditions—certainly makes it unattractive on both administrative and political grounds.

See also: Expert Testimony; Legal Process and Social Science: United States; Litigation; Litigation: Effects on Medical Research; Procedure: Legal Aspects; Rules in the Legal Process; Scientific Evidence: Legal Aspects; Statistics as Legal Evidence

Bibliography

Brazil W D 1979 The adversary character of civil discovery. *Vanderbilt Law Review* **31**: 1295–361

Brazil W D 1980a Views from the front lines: observations by Chicago lawyers about the system of civil discovery. *American Bar Foundation Research Journal*: 217–51

Brazil W D 1980b Civil discovery: lawyers' views of its effectiveness, its principal problems and abuses. *American Bar Foundation Research Journal*: 787–902

Bundy S McG 1994 Valuing accuracy: filling out the framework. *Journal of Legal Studies* **23**: 411–33

Cooter R, Rubinfeld D 1994 An economic model of legal discovery. *Journal of Legal Studies* **23**: 435–63

Damaska M R 1997 *Evidence Law Adrift*. Yale University Press, New Haven, CT

Gerber D J 1986 Extraterritorial discovery and the conflict of procedural systems: Germany and the United States. *American Journal of Comparative Law* **34**: 745–88

Gilson R J, Mnookin R H 1994 Disputing through agents: cooperation and conflict between lawyers in litigation. *Columbia Law Review* **94**: 509–66

Hay B 1994 Civil discovery: its effects and optimal scope. *Journal of Legal Studies* **23**: 481–515

Hazard G C 1998 Discovery and the role of the judge in civil law jurisdictions. *Notre Dame Law Review* **73**: 1017–28

James F, Hazard G C, Leubsdorf J 1992 *Civil Procedure*, 4th edn. Little, Brown, Boston

Jolowicz J A 2000 *On Civil Procedure*. Cambridge University Press, Cambridge, UK

Kakalik J S, Hensler D R, McCaffrey D, Oshiro M, Pace N, Vaiana M 1998 *Discovery Management: Further Analysis of the Civil Justice Reform Act Evaluation Data*. Rand, Santa Monica, CA

Kaplow L 1994 The value of accuracy in adjudication: an economic analysis. *Journal of Legal Studies* **23**: 307–401
Langbein J H 1985 The German advantage in civil procedure. *University of Chicago Law Review* **52**: 823–66
Marcus R L 1999 Retooling American discovery for the twenty-first century: toward a new world order? *Tulane International and Comparative Law Review* **7**: 153–99
McKenna J A, Wiggins E C 1998 Empirical research on civil discovery. *Boston College Law Review* **19**: 785–807
Scott K E 1975 Two models of the civil process. *Stanford Law Review* **27**: 937–50
Willging T E, Stienstra D, Shapard J, Miletich D 1998 An empirical study of discovery and disclosure practice under the 1993 Federal Rule Amendments. *Boston College Law Review* **19**: 525–96

S. Bundy

Discrete Mathematics

Some of the most important problems of the social and behavioral sciences are concerned with arrangements, designs, patterns, sequences, schedules, and assignments. These are among the fundamental notions dealt with by discrete mathematics (dm). This article describes some of the tools of dm—digraphs, graphs, signed and marked graphs, partial orders, rankings, combinatorics—and their applications.

1. *What is Discrete Mathematics?*

John Kemeny made the following distinction. Many problems of science deal with quantities so large that it is natural to assume that they are dense, continuously distributed, and that all real numbers can be used to measure them. Centuries of development of 'continuous mathematics' have given us extremely powerful tools for handling these kinds of problems. Other problems are so small that we can deal with all the possible cases by hand. These are truly 'finite' problems. Some of the most important problems, however, fall in between: not big enough to assume density, continuity, etc., but not small enough to allow us to consider all cases. These are, for the most part, the problems with which discrete mathematics (DM) deals. Because increasingly powerful computers are allowing us to replace computations by hand, it is becoming increasingly feasible to deal with problems of DM. This explains, in part, why DM has become perhaps the fastest growing field of modern mathematics.

Many of the basic problems of the physical sciences, dealing with time, mass, velocity, etc., are of the first kind. So are many problems in the biological sciences. However, many problems of the social and behavioral sciences fall in the middle ground described by Kemeny. The tools of DM are especially relevant here. For an elementary introduction to DM, see Kolman et al. (1999) and Roberts (1984).

2. *Digraphs*

A *digraph* is a pair (V, A) where V is a set called the set of *vertices* and A is a set of ordered pairs from V called the set of *arcs*. *Throughout, we shall assume that V is a finite set*. Usually, arcs (x, x), known as *loops*, are not permitted or disregarded, and we shall disregard them. Typically, we represent a digraph geometrically by letting the vertices be points and including a directed line from x to y if (x, y) is in A. The arcs in a digraph are used in social and behavioral science models to represent relations among the vertices. Thus, $(x, y) \in A$ can mean that x is preferred to y, x is louder than y, x influences y, etc.

Suppose (V, A) is a digraph with arcs representing 'preferred to.' Then an *ordinal utility function* can be thought of as a function f that assigns a real number $f(x)$ to each x in V so that for all $x, y \in V$,

$$(x, y) \in A \leftrightarrow f(x) > f(y). \tag{1}$$

Similarly, if arcs represent 'louder than,' then a function satisfying Eqn. (1) might measure loudness. If a function satisfying Eqn. (1) exists, then the digraph is *asymmetric* in the sense that if there is an arc from x to y, then there is no arc from y to x, and *transitive* in the sense that if there are arcs from x to y and y to z, then there is an arc from x to z. It is also *negatively transitive* in the sense that if there is no arc from x to y and no arc from y to z, then there is no arc from x to z. It turns out that asymmetry and negative transitivity of the digraph are together sufficient for the existence of a function satisfying Eqn. 1 (see Roberts 1976). An asymmetric, negatively transitive digraph is called a *strict weak order*.

Digraphs provide precise mathematical language for expressing concepts in the social and behavioral sciences. For instance, let us say that vertex y is *reachable* from vertex x if there is a sequence of vertices $x_1, x_2, \ldots, x_t$ so that

$$(x, x_1), (x_1, x_2), \ldots, (x_{t-1}, x_t), (x_t, y)$$

are arcs. The notion of reachability is important in sociology in the study of *communication networks*. If an arc in the digraph means that x can communicate directly with y (not necessarily a *symmetric* relationship in the sense that $(x, y) \in A \rightarrow (y, x) \in A$), then

reachability means that it is possible for x to initiate a message to y. There is then interest in minimum-size sets of vertices that can reach all others (so-called *vertex bases*). In large digraphs, it is often very important to represent them geometrically in such a way that their structural properties become clear. We might wish to find another digraph on the same vertex set with the same reachability properties and a minimum number of arcs. This idea arises in the area of data analysis called *interpretive structural modeling* (see Warfield 1976). The drawing of digraphs so as to visualize the information they contain in useful ways has become a major research field known as *graph drawing*.

Understanding the structure of communication networks is aided by considering *strong components*, maximal sets of vertices each two of which are reachable from each other. Many large social networks have some huge strong components. This is the case in *telephone call digraphs* where the vertices are telephone numbers and arcs correspond to calls between them, and in *scientific collaboration digraphs* where an arc means that x collaborates with y (a symmetric relationship). Similar results are being found for pages on the World Wide Web, with arcs corresponding to references from one page to another, and the results seem to suggest that totally new models are needed to explain how large social networks are formed. (For a basic introduction to such digraph concepts as reachability, vertex basis, strong component, etc., see Harary et al. 1965 and Roberts 1976. For more on social networks, see Degenne and Forse 1999.)

3. Partial Orders

If (V, A) is a digraph, its *symmetric complement* is the digraph (V, B) where

$$(x, y) \in [B \leftrightarrow (x, y) \notin A \quad \& \quad (y, x) \notin A]$$

If there is an f satisfying Eqn. (1), then the digraph (V, B) is transitive. If A corresponds to preference, then B corresponds to indifference. Arguments against the transitivity of indifference were made by the economist Armstrong in the 1930s and even earlier by Karl Menger in the nineteenth century. Similarly, in *psychophysics*, if A corresponds to 'louder than,' then B corresponds to 'equally loud' and this might not be transitive. Arguments against transitivity of B often involve *thresholds*: x and y might be within threshold and so sound equally loud, y and z might be within threshold and so sound equally loud, but x and z might not be within threshold.

This kind of argument led Luce in 1956 to introduce a different model for utility of loudness judgments:

$$(x, y) \in A \leftrightarrow f(x) > f(y) + \delta \tag{2}$$

where δ is a fixed positive constant measuring threshold or *just noticeable difference*. Luce and also Scott (1958) and Suppes gave necessary and sufficient conditions for a digraph to have a function f satisfying Eqn. (2), conditions defining a *semiorder*. Semiorders are transitive digraphs. They are also very important examples of a second fundamental concept of DM, a partial order, an asymmetric, transitive digraph. (Sometimes this is called a *strict partial order* and the term *partial order* is reserved for the digraphs that are transitive, *reflexive* $[(x, x) \in A$ for all $x]$, and *antisymmetric* $[(x, y) \in A \ \& \ (y, x) \in A \leftrightarrow x = y]$.) Partial orders arise in numerous applications in the social and behavioral sciences (see *Partial Orders*).

If f satisfies Eqn. (2), then consider the real interval $J(x) = [f(x) - \delta/2, f(x) + \delta/2]$. Suppose that when J and J' are two real intervals, $J \succ J'$ means that every point of J is to the right of every point of J'. Then Eqn. (2) can be rewritten as

$$(x, y) \in A \leftrightarrow J(x) \succ J'(y) \tag{3}$$

We can think of $J(x)$ as a range of fuzziness about $f(x)$. If these intervals have different lengths, then the digraphs defined by Eqn. (3) are called *interval orders*, another important class of partial orders. The term interval order is due to Fishburn in 1970, but the concept was anticipated by Norbert Wiener in 1914. For more about semiorders, interval orders, and a variety of partial orders, see Fishburn 1985 and Trotter 1992.

4. Graphs

If a digraph (V, A) is symmetric, we call it a *graph*. Replacing each pair of arcs (x, y) and (y, x) by an unordered pair $\{x, y\}$, we can think of a graph as a pair (V, E) where E is a set of unordered pairs from V. Elements of E are called *edges*. Geometrically, the vertices are represented as points and if $\{x, y\} \in E$, then we connect x and y by an undirected line. The symmetric complement of a digraph, of course, is a graph. The symmetric complement of a semiorder is called an *indifference graph* or a *unit interval graph*. The name comes from the case where A corresponds to preference. The symmetric complement of an interval order is called an *interval graph*. Interval graphs have a huge number of applications in archaeological seriation, developmental psychology, ecology, molecular biology, data retrieval, scheduling, etc. (see Roberts 1976).

Indifference graphs and interval graphs are special cases of *intersection graphs*, graphs whose vertices correspond to sets and which have an edge between two vertices if and only if the corresponding sets have a nonempty intersection. There are many important open research questions concerning intersection graphs. For example, characterizing intersection graphs where the sets are unit squares in the plane, a long-standing open problem, is relevant to judgments of indifference when there are two dimensions to the alternatives being compared.

Social networks are often graphs since the underlying relation defining the arcs is often symmetric. In this case, the strong components are simply called *components*. The components often consist of sets of vertices each two of which are joined by an edge. Such a set of vertices is called a *clique*; if an edge between x and y means that they are friends, then a clique is a set of individuals each two of which are friends. Surprisingly enough, it is computationally difficult to find the largest clique in a graph.

If (V, E) is a graph, its *complement* is the graph (V, F) where $\{x, y\} \in F \leftrightarrow \{x, y\} \notin E$. We often seek to partition the vertices of a social network into cliques. In the complement, this problem corresponds to partitioning the vertices into sets with no edges joining vertices within the same set. Such sets are called *independent sets*. The problem of partitioning the vertices of a graph into independent sets is equivalent to one of the most important problems in graph theory, the *graph coloring problem*. A *coloring* of a graph is a function f that assigns a color $f(x)$ to each vertex x so that $\{x, y\} \in E \rightarrow f(x) \neq f(y)$. The sets of vertices of a given color form independent sets. Graph coloring has many important applications. In a typical application, individuals who dislike each other must be assigned to different tasks or time schedules or groups. Disliking defines the edges while the colors correspond to the tasks, schedules, or groups. The theory of *role assignments* in sociology, sometimes known as the theory of *role colorings*, is closely related. While the subject of graph coloring is an old one, going back to the famous 'four color problem' of the nineteenth century, it is a very lively area of research today with important modern variations on graph colorings motivated by practical applications. An example is the *list coloring problem*, where a list of acceptable colors (tasks, time schedules, groups) is given at each vertex and the color assigned to a vertex must be chosen from its list (see Roberts 1984 for an introduction to graph coloring and Kratochvíl et al. 1999 for the list coloring problem).

5. Signed and Marked Graphs

In some applications of graph theory, it is useful to classify edges as + or –. We then speak of a *signed graph*. One of the most fascinating applications of signed graphs is as a model of *balance* in small group sociology. A group of people is called balanced if it works well together, lacks tension, etc. Cartwright and Harary (1956) and Harary (1954) made this concept precise as follows. Let the vertices of a graph be individuals in the group, an edge mean they have a strong relationship, and the sign distinguish between a positive relationship (such as likes, associates with, agrees with, stimulates) and a negative relationship (such as dislikes, avoids, disagrees with, inhibits). If (V, E) is a (signed) graph, a *circuit* or *cycle* is a sequence $x_1, x_2, \ldots, x_t$ of distinct vertices so that $t \geqslant 3$, $\{x_t, x_1)$ is in E and for $i = 1, \ldots, t-1$, $\{x_i, x_{i+1}\}$ is in E. There is a lot of evidence to show that a small group is balanced if and only if in the corresponding signed graph, every circuit has an even number of – signs. If the latter holds, we call the signed graph *balanced*. The same notion was introduced into the economics literature by Morishima in 1952. Characterization of balanced signed graphs was first given by a 'structure theorem' by Harary (1954). Later, Maybee and Maybee, Harary and Kabell, and Hansen gave linear time algorithms for determining whether a signed graph is balanced. Notions of balance theory have found applications in a variety of areas, including the study of 'social justice,' social inequalities, international relations, and the analysis of multiparty political systems. Since not all systems are totally balanced or totally unbalanced, ways to measure the degree of balance are important and remain a subject of research today (for an introduction to balance theory, see Taylor 1970 or Roberts 1976).

In some applications, it is useful to put signs on the vertices of a graph rather than on the arcs. Such *marked graphs* arise in the study of social networks, where we think of vertices as individuals who either tell the truth (+ sign) or lie (– sign), and where we trace messages through the network. Beineke and Harary called a marked graph *consistent* if every circuit has an even number of – signs, a concept analogous to balance in signed graphs. Surprisingly, it is mathematically very difficult to characterize consistent marked graphs. In spite of a lot of work on this subject, a complete understanding of what marked graphs are consistent remains elusive.

6. Rankings and Combinatorics

Strict weak orders, that is, asymmetric, negatively transitive digraphs, are sometimes called *rankings*. We can rank the vertices from top to bottom, with ties allowed. Rankings arise in the theory of voting, group consensus, and social welfare. There are situations where just knowing a voter's first choice candidate does not give us enough information to find an appropriate winning candidate. It can be useful to have each voter provide a ranking of the candidates. Then we need a rule that provides a winning ranking. If such a rule is defined for every possible combination of rankings, we call it a *group consensus function* or *social welfare function* (SWF).

The theory of SWFs goes back at least to the eighteenth century (in the work of Marie Jean Antoine Nicolas Caritat, Marquis de Condorcet, and of Jean-Charles de Borda), but had its modern impetus in the economic modeling of Kenneth Arrow. Arrow listed some reasonable axioms we might require of an SWF and showed that there was no function satisfying these axioms. *Arrow's Impossibility Theorem* has been modified, replicated, and applied in a myriad of ways since then. A few general references on group con-

sensus are Campbell 1992, Kelly 1988, and Roberts 1976, 1989.

Modern applications of SWFs and notions of group consensus put a great deal of emphasis on computational tractability of rules for determining the winning consensus. Bartholdi, Tovey, and Trick, Wakabayashi, and others showed that there are voting rules that are *computationally intractable*, that is, where it is extremely difficult to compute the winner of an election (see Roberts 1989). They also pointed out that sometimes computational intractability is useful, as when we wish to make it difficult to manipulate the outcome of an election such as by adding voters, declaring candidates ineligible, etc. Computational tractability becomes an issue when there are large numbers of voters or candidates. This is the case in applications of theories of group consensus to *metasearch* on the Internet, that is, combining the results of different search engines, and to such e-commerce applications as *collaborative filtering* where the rankings of different individuals among alternative movies, books, or restaurants are combined to produce an automated ranking as a recommendation to a given user of a system. Important open questions about SWFs arise in such applications, including issues of learning and adapting through repetition.

If a ranking does not allow ties, we sometimes talk about a *strict simple order* or a *permutation*. Strict simple orders are also important in applications such as models of power in voting situations. Suppose different 'players' (such as states in the US Electoral College or owners of shares in a company) have different numbers of votes. Imagine adding players to a 'coalition' one at a time until enough votes are obtained to 'win.' The player whose votes throw a coalition over from losing to winning is called *pivotal*. One measure of the 'power' of a player in such a situation is the probability that the player is pivotal if a coalition is formed at random. This is known as the *Shapley value*. There are a variety of other interesting power measures that, like this one, have come out of the theory of cooperative games (see Fudenberg and Tirole 1991 and Roberts 1976). The Shapley value can be calculated by counting the number of permutations of the players in which a given player is pivotal and then dividing by the number of permutations of the players. In large games, this becomes a computationally intractable problem. To solve it in special cases, one has to solve counting problems.

The *counting problem*, involving finding the number of permutations, combinations, and other mathematical objects, is a fundamental problem of the branch of DM known as *combinatorics*. The Shapley value has been calculated for various governmental bodies such as different legislatures, the UN Security Council, and the US Electoral College. Yet, new computational procedures for calculating it are very much required. This is especially important in light of the many uses of the Shapley value in cost allocation problems such as fairly sharing the cost of telephone systems, airports, libraries, and multicast transmissions of movies.

In conclusion, discrete mathematics is finding many new and unexpected applications in the social and behavioral sciences. At the same time, the social and behavioral sciences are one of the important stimuli for the development of this rapidly growing field of mathematics.

See also: Bayesian Graphical Models and Networks; Graphical Models: Overview; Partial Orders; Ranking Models, Mathematics of; Social Network Models: Statistical

Bibliography

Campbell D E 1992 *Equity, Efficiency, and Social Choice*. Clarendon Press, Oxford, UK

Cartwright D, Harary F 1956 Structural balance: A generation of Heider's theory. *Psychology Review* **63**: 277–93

Degenne A, Forse M 1999 *Introducing Social Networks*. Sage, London

Fishburn P C 1985 *Interval Orders and Interval Graphs*. Wiley, New York

Fudenberg D, Tirole J 1991 *Game Theory*. MIT Press, Cambridge, MA

Harary F 1954 On the notion of balance of a signed graph. *Michigan Mathematical Journal* **2**: 143–46

Harary F, Norman R Z, Cartwright D 1965 *Structural Models: An Introduction to the Theory of Directed Graphs*. Wiley, New York

Kelly J S 1988 *Social Choice Theory*. Springer-Verlag, Berlin

Kolman B, Busby R C, Ross S C 1999 *Discrete Mathematical Structures*, 4th ed. Prentice-Hall, Englewood Cliffs, NJ

Kratochvil J, Tuza Zs, Voigt M 1999 New trends in the theory of graph colorings: Choosability and list coloring. In: Graham R L, Kratochvil J, Nešetřil J, Roberts F S (eds.) *Contemporary Trends in Discrete Mathematics*. DIMACS Series, Volume 49, American Mathematical Society, Providence, RI, pp. 183–97

Luce R D 1956 Semiorders and a theory of utility description. *Econometrica* **24**: 178–91

Roberts F S 1976 *Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems*. Prentice-Hall, Englewood Cliffs, NJ

Roberts F S 1984 *Applied Combinatorics*. Prentice-Hall, Englewood Cliffs, NJ

Roberts F S (ed.) 1989 *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences*. IMA Volumes in Mathematics and its Applications, Vol. 17. Springer-Verlag, New York

Scott D, Sappes P 1958 Foundational aspects of theories of measurement. *Journal of Symbolic Logic* **23**: 113–28

Taylor H F 1970 *Balance in Small Groups*. Van Nostrand Reinhold, New York

Trotter W T 1992 *Combinatorics and Partially Ordered Sets: Dimension Theory*. Johns Hopkins University Press, Baltimore, MD

Warfield J N 1976 *Societal Systems: Planning, Policy, and Complexity*. Wiley, New York

F. S. Roberts

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Discrete State Models of Information Processing

Around the middle of the twentieth century, experimental psychology shifted from a behavioristic to a cognitive perspective, where unobservable mental processes were allowed as theoretical explanations of behavioral data. A dominant type of theoretical explanation analyzed the mind analogously to a computer in the sense that it performs operations like inputting, storing, organizing, maintaining, comparing, and retrieving information. This approach came to be known as information processing (IP) psychology. At about the same time, formal probabilistic IP models were advanced in mathematical psychology in such areas as choice theory, concept identification, learning, memory, and signal detection (see Batchelder 2000, for a review). Many of these formal models fall into the category of discrete state information processing models, the subject of this article. Discrete state models of an IP task assume that manifest (observable) responses in the task are generated probabilistically from one of a finite, and usually small, set of latent (unobservable) cognitive states. The input stimulus of the IP task probabilistically triggers a sequence of one or more cognitive states, that in turn probabilistically generates the response output. In this article, the focus will be on parametric, discrete state models that yield *categorical data*, that is, observations that fall into a discrete set of nonoverlapping response categories.

1. Introduction

In the early history of IP modeling, most discrete state models were of two varieties, signal detection models and *Markov models* for learning tasks. Sections 2 and 3, respectively, will describe these two areas. Section 4 will describe the family of multinomial processing tree models (MPT) in some detail because the family has a well worked out statistical theory and it subsumes most discrete state models for categorical data, as well as the two earlier varieties. Suppose an experimental subject's response to a stimulus must fall into exactly one of a discrete system of categories, say $C_1, C_2, \ldots, C_J$. If there are M different stimulus conditions, then under repeatable and homogeneous conditions, there are theoretically M different observable probability distributions, $p_m = (p_{1m}, \ldots, p_{Jm})$, one for each stimulus condition, $1 \leqslant m \leqslant M$. This situation is typical of many IP tasks, for example, identification, categorization, and yes–no signal detection. In an identification task, there are J different stimuli, and the subject's task is to identify the presented stimulus. The data consist of a $J \times J$, row stochastic *confusion matrix*, whose ijth term is Pr $(R_j | S_i)$, the probability that stimulus i is identified as stimulus j, $1 \leqslant i, j \leqslant J$. In categorization, each stimulus belongs to a particular response category, however, unlike the identification task, several stimuli belong to the same category, so $M > J$. Yes–no signal detection is described in the next section. Many discrete models exist for these and other IP tasks.

2. Signal Detection Models

One of the earliest applications of discrete state models was the development of threshold models for the yes–no signal detection task. The task involves exposing observers to a series of items, some of which satisfy a task definition of 'signal items' (s_1) and the others are 'noise' (s_2) items. Examples include whether or not an acoustic event includes a particular sound signal in an auditory signal detection experiment, whether or not a tested item was on a previously studied list in a recognition memory experiment, or whether or not an x-ray exhibits a particular type of tumor in a medical diagnosis setting. The observer's task is to say 'yes' (r_1) to all items that satisfy the definition of signal, and otherwise say 'no' (r_2).

Discrete state models for the yes–no detection paradigm are characterized by the assumption that observers set detection thresholds. One example is the general high threshold model. The model is motivated by the assumption that the observer sets a threshold for detecting a signal item and another one for detecting a noise item. Both thresholds are set sufficiently high that the observer never detects the wrong type of item; however, if no detection occurs (neither threshold is reached), the observer must guess.

The model postulates three latent states, D_1, D_2 and D_0 for detect signal, detect noise, and detect neither, respectively. The parametric representation of the model is easy to present in terms of two matrices, one giving the probability that the stimulus triggers the various discrete states, and the second the probability distributions over the responses for each discrete state. The stimulus input matrix is given by

$$I(\theta) = \begin{bmatrix} 1-d_1 & d_1 & 0 \\ 1-d_2 & 0 & d_2 \end{bmatrix} \tag{1}$$

where the ijth term is $\Pr(D_j | s_i)$, i = 1, 2 and $j = 0, 1, 2$, and the response matrix is given by

$$R(\theta) = \begin{bmatrix} g & 1-g \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2}$$

where the jkth term is $\Pr(r_k | D_j)$, $j = 0, 1, 2$ and $k = 1, 2$, and where $\theta = (d_1, d_2, g)$ is the parameter vector with parameter space $\Omega = [0, 1]^3$.

Equation (2) gives the three probability distributions over the two response categories, r_1 and r_2, and the output distributions for each item type are easily computed by matrix multiplication as follows:

$$M(\theta) = I(\theta) \cdot R(\theta) = \begin{bmatrix} d_1 + (1-d_1)g & (1-d_1)(1-g) \\ (1-d_2)g & d_2 + (1-d_2)(1-g) \end{bmatrix} \quad (3)$$

Note that Eqn. (3) is in the form of a 2×2 row stochastic matrix, with two degrees of freedom. However, there are three parameters so the model is not useful in analyzing data without further restrictions. For example, if $d_1 = d_2$, the model is known as the double high threshold model, and if $d_2 = 0$, the single high threshold model results. Both models are discussed in Macmillan and Creelman (1991).

Most theorists in perception and memory have rejected these two models because they fail to account for the shape of experimental *receiver operator characteristics* (ROC), sometimes called relative operating characteristics, e.g., Swets (1996). The data for an ROC come from a set of experimental conditions wherein the detection processes are held constant and the response bias processes are allowed to vary. The ROC then becomes a plot of the hit rate, $H = \Pr(r_1 | s_1)$, as a function of the false alarm rate, $F = \Pr(r_1 | s_2)$, where each experimental condition x yields a pair $(F(x), H(x))$. Such plots are obtained experimentally in a variety of ways (see *Signal Detection Theory*).

The general high threshold model implies a linear ROC given by

$$H(x) = d_1 + \left(\frac{1-d_1}{1-d_2} \right) F(x) \quad (4)$$

where x indexes conditions leading to variations in g but not in d_1 and d_2. Experimental evidence is strongly against a linear ROC, and therefore most theorists have abandoned simple high threshold models. Nevertheless, measures of detection, such as $H-F$, consistent with the double high threshold model, are frequently seen both in empirical papers in recognition memory and as a 'correction-for-guessing' formula in psychometric models.

Discrete state models that are based on 'low thresholds' that allow an observer to detect a signal when none is present have had more appeal to theorists. One example is Luce's (1963) low-threshold theory. The theory is difficult to reject from ROC data, unless conditions lead to extreme values of the false alarm rate. There are other, more complex discrete state models in the signal detection area. One example is the *neural quantal model* (e.g., Larkin and Norman 1964). Also, Graham et al. (1987) discuss both continuous and discrete state models for multidimensional stimuli. Finally, Batchelder and Riefer (1990) develop a family of models for the so-called source monitoring paradigm, where signals come from several sources or are of several types.

3. Markov Learning Models

Markov learning models (ML) are a variety of discrete state IP models for learning and memory tasks which were especially popular in the 1960s and 1970s (see *Mathematical Learning Theory, History of*). A typical task requires a learner to acquire simple associations that involve an experimenter assigned correct response pairing to each of a set of stimulus items over a series of discrete learning trials. On each trial, learners receive a series of study events in each of which they are informed of a correct stimulus-response pairing and a series of test events in each of which they are given a stimulus item and required to produce the correct response. The task is continued for a number of learning trials until some performance criteria are met. Typically, the responses for each item are scored as errors (E) or correct responses (C), so the response categories for an item are defined in terms of the possible observable sequences of Es and Cs over trials.

Markov learning models are characterized by several assumptions. First, it is assumed that on each test trial, an item is in one of a (small) finite set of latent cognitive states, each characterizing a possible state of the to-be-learned item–response pair, e.g., an unlearned state, a short-term memory state, a long-term memory state. Second, it is assumed that associated with each state is a probability of a correct response, where this probability is constant over trials. On any trial, the item is assumed to be in exactly one of the latent states and the response is governed by the state's response probability. Third, it is assumed that the effect of study trials is to generate transitions among the learning states described dynamically by a finite state, homogeneous Markov chain. The states of the chain are the cognitive states, and the starting distribution over the states, as well as the state-to-state transition probabilities and the state-to-response probabilities, are each functions of a parameter vector, $\theta = (\theta_1, \ldots, \theta_s)$ where each component parameter taps some latent cognitive process. Technically, most ML models in the literature can be viewed as discrete trial, discrete state hidden Markov chains (e.g., Smyth 1999). The statistical inference for an ML model involves estimating the parameter vector θ from observed frequency counts over the response sequence categories. The mathematical and statistical details are presented in Wickens (1982).

Many examples of Markov learning models are reviewed in Greeno and Bjork (1973). However, Markov learning models ceased to be of major theoretical importance in learning and memory some-

time in the late 1970s, when a variety of more elaborate and detailed IP models of learning began to be introduced such as *connectionist* models (see *Connectionist Models of Concept Learning*; *Artificial Intelligence: Connectionist and Symbolic Approaches*) and global memory models (see *Memory Models: Quantitative*). Nevertheless, ML models have continued to serve as psychometric models for psychological measurement. For example, they have been applied to early development and to aging, where standardized paradigms provide learning sequences for different age groups that can be used to estimate age-related trends in latent storage and retrieval processes (reviewed in Batchelder and Riefer 1990).

4. Multinomial Processing Tree Models

Multinomial processing tree models constitute a large family of discrete state IP models for categorical data. The MPT family was developed and formalized, including its statistical inference, in Riefer and Batchelder (1988) and Hu and Batchelder (1994). Informally, the assumption is that each stimulus presentation triggers one of a number of processing sequences, each of which, in turn, terminates in an observable category. There may be more than one processing sequence that results in a particular category, so knowledge of the response category does not necessarily reveal the underlying processing sequence. Most current and past discrete state IP models, including signal detection and Markov learning models, can be parameterized to fit into a subclass of MPT models called binary MPT models, so that subclass is presented next in detail.

4.1 Definition

There are four aspects to a binary MPT model: observable categories; latent parameters; tree architecture; and computational rules. The first aspect consists of J observable categories, $C_1, C_2, \ldots, C_J$. Second, there are S functionally independent parameters, $\theta_1, \theta_2, \ldots \theta_S$, each with full range, $0 \leqslant \theta_s \leqslant 1$. Each parameter is interpreted as the probability of some latent IP event, e.g., storing an item, retrieving an item, matching an item, clustering items, making an inference, detecting a signal, discriminating between items, or guessing a response.

The tree architecture is a binary tree with a single initial node, intermediate nodes, and terminal nodes. There are two types of continuations after each nonterminal node: parametric probability links and numerical links. Each parametric probability link is associated with a particular parameter θ_s, and the two directed links from the node correspond to the success (θ_s) or failure ($1-\theta_s$) of the corresponding latent event. The numerical links are constructed similarly with a numerical probability instead of a parameter. Finally, each terminal node corresponds to one of the J categories.

The computational component of any MPT model provides the probabilities of each observable category in terms of the parameters. First, each branch in the tree is a sequence of links from the initial node into one of the categories. Branch probabilities are computed as products of link probabilities and have a special form as follows:

$$\Pr(Branch\ k) = c_k \prod_{s=1}^{S} \theta_s^{a_{ks}} (1-\theta_s)^{b_{ks}} \tag{5}$$

where c_k is the product of numerical constants on the links of branch k and a_{ks} and b_{ks} are the number of links on branch k with parameters θ_s and $(1-\theta_s)$, respectively.

Since each branch ends in a particular category, category probabilities p_j are sums of suitable branch probabilities:

$$p_j = p_j(\theta_1, \ldots, \theta_s) = \sum_{k=1}^{I_j} \Pr(Branch\ k \text{ ending in } C_j) \tag{6}$$

where I_j is the number of branches ending in category C_j. Given this structure, it is easy to show that $0 \leqslant p_j \leqslant 1$, and

$$\sum_{j=1}^{J} p_j(\theta_1, \ldots, \theta_S) = 1$$

for all $0 \leqslant \theta_s \leqslant 1, s = 1, 2, \ldots, S$. It is a consequence of the computational rules of an MPT model that if the category counts, $N_1, N_2, \ldots, N_J$, are obtained as independent and identically distributed cases, then the MPT model becomes a specially parameterized multinomial model with probability distribution

$$\Pr(N_j, \ldots, N_J) = N! \prod_{j=1}^{J} \frac{[p_j(\theta_1, \ldots, \theta_S)]^{N_j}}{N_j!} \tag{7}$$

where $N = \Sigma N_j$.

Figure 1 provides an example of an MPT model which is a representation of the general high threshold model in Eqns. (1)–(3). In the representation, there are six branches, $J = 4$ response categories, and $S = 3$ parameters. Psychological paradigms yielding categorical data usually involve more than one system of categories, wherein each category system reflects a type of item. For example, for the general high threshold model, there are two category systems, one for signal items and one for noise items. One of the purposes of the constant links in the earlier definition of MPT models is to combine separate MPT models into a single model.

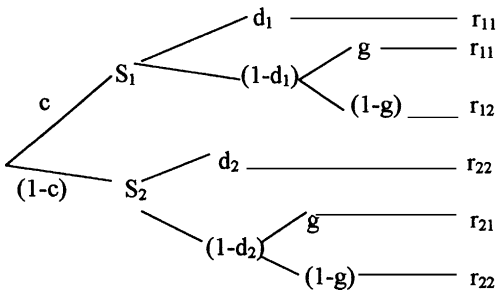

Figure 1
A binary MPT representation of the general high threshold model. The quantity c represents the proportion of signal items, d_1, d_2, and g are the parameters of the model, and r_{ij} is the category of response type j to item type i (1 for signal and 2 for noise)

4.2 Statistical Inference

Hu and Batchelder (1994) provide methods for statistical inference for the entire class of globally identifiable MPT models based on the *Expectation–Maximization algorithm* (EM) (Demster et al. 1977). Global identifiability means that at most one parameter vector generates any particular probability distribution over the J categories. Thus, in principle, knowledge of the category probabilities uniquely determines the ('best fitting') underlying parameter vector θ.

The key to the application of the EM algorithm is to treat the branch frequencies as 'missing data.' Let N_{jk} be the frequency out of N counts of the occurrence of branch k terminating in category j, where $1 \leqslant k \leqslant I_j$, $1 \leqslant j \leqslant J$ and I_j is the number of branches ending in category C_j. Suppose the N_{jk} are known, then it is easy to see that the *maximum likelihood estimators* (MLEs) of the θ_s are given by

$$\hat{\theta}_s = \frac{\sum_{j=1}^{J}\sum_{k=1}^{I_j} a_{jks} N_{jk}}{\sum_{j=1}^{J}\sum_{k=1}^{I_j} (a_{jks}+b_{jks}) N_{jk}} \tag{8}$$

where the a_{jks} and b_{jks} are defined as in Eqn. (5). Thus, the maximization step of the EM algorithm yields a simple closed form expression.

There is a software package described in Hu and Phillips (1999) that enables a researcher to design an MPT model graphically, load in a data file of category frequencies, and obtain point estimates, confidence intervals, goodness-of-fit measures, do hypothesis testing within and between groups, and run Monte Carlo simulation with and without beta distributed hyperparameters.

4.3 Application Areas

It is obvious that the class of possible binary MPT models is large and that most members are uninteresting and devoid of scientific meaning. However, it is a framework where a researcher can construct and experimentally validate interpretable discrete state models. By the end of the twentieth century, many researchers have used models in the MPT class to understand data in specific cognitive paradigms (reviewed in Batchelder and Riefer 1990). They include applications to many areas of human memory including proactive and retroactive inference, associative recall, short-term memory, recognition failure, cued-then-free recall, clustering in free recall, the bizarreness effect, source monitoring, recognition memory, process dissociation, eyewitness memory, and hindsight bias. Outside of the area of human memory, there are applications of MPT models in other areas including object perception, speech perception, letter identification, propositional reasoning, interjudge reliability, social perception, social networks, and ability testing. In addition, MPT models have been developed for anthropological methodology, under the rubric cultural consensus theory (see Romney and Batchelder 1999), and for the area of clinical assessment, for example, Batchelder (1998).

The family of MPT models stands in contrast to most nondiscrete state IP models in that the latter assume much more complex cognitive architecture, more elaborate underlying processes, and in particular, continuous type variables that determine the output. In fact, many theorists view discrete state models as theoretically impoverished. Indeed, this view is common in the signal detection literature, e.g., Macmillan and Creelman (1991), as well as in other areas of cognition, e.g., Mulder et al. (1995). However, it may not be possible to rule out discrete state models on empirical grounds, e.g., Summers and Suppes (1997) or from a computational perspective, e.g., Pylyshyn (1986). In practice, however, MPT models have been most useful as simple, approximate models to measure underlying cognitive processes in particular experimental paradigms rather than providing a deep understanding of cognitive processes across a wide range of phenomena. In this sense, they provide psychometric tools for cognitive psychology.

See also: Markov Decision Processes; Markov Processes for Knowledge Spaces; Mathematical Psychology, History of; Signal Detection Theory

Bibliography

Batchelder W H 1998 Multinomial processing tree models and psychological assessment. *Psychological Assessment* **10**: 331–44

Batchelder W H 2000 Mathematical psychology. In: Kazdin A E (ed.) *Encyclopedia of Psychology*. American Psychological Association, Oxford University Press, New York, pp. 120–3
Batchelder W H, Riefer D M 1990 Multinomial processing models of source monitoring. *Psychological Review* **97**: 548–64
Dempster A P, Laird N M, Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* **39**: 1–38
Graham N, Kramer P, Yager D 1987 Signal detection models for multidimensional stimuli: probability distributions and combination rules. *Journal of Mathematical Psychology* **31**: 366–409
Greeno J G, Bjork R A 1973 Mathematical learning theory and the new 'mental forestry.' *Annual Review of Psychology* **24**: 81–116
Hu X, Batchelder W H 1994 The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika* **59**: 21–47
Hu X, Phillips G A 1999 GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, Instruments and Computers* **31**: 220–34
Larkin W D, Norman D A 1964 An extension and experimental analysis of the neural quantum theory. In: Atkinson R C (ed.) *Studies in Mathematical Psychology*. Stanford University Press, Stanford, CA, pp. 188–200
Luce R D 1963 A threshold theory for simple detection experiments. *Psychological Review* **70**: 61–79
Macmillan N A, Creelman C D 1991 *Detection Theory: A User's Guide*. Cambridge University Press, New York
Mulder G, Sanders A F, van Galen G P (eds.) 1995 Discrete and continuous information processing (Special Issue). *Acta Psychologica* **90**
Pylyshyn Z W 1986 *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press, Cambridge, MA
Riefer D M, Batchelder W H 1988 Multinomial modeling and the measurement of cognitive processes. *Psychological Review* **95**: 318–39
Romney A K, Batchelder W H 1999 Cultural consensus theory. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 208–9
Smyth P 1999 Hidden Markov models. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 373–4
Summer R, Suppes P 1997 Dispensing with the continuum. *Journal of Mathematical Psychology* **41**: 3–10
Swets J A 1996 *Signal Detection Theory and ROC Analysis in Psychology and Diagnosis: Collected Papers*. Erlbaum, Mahwah, NJ
Wickens T D 1982 *Models for Behavior: Stochastic Processes in Psychology*. Cambridge University Press, New York

W. H. Batchelder

Discrimination

The Oxford Dictionary defines discrimination as 'the perceiving, noting, or making a distinction or difference between things.' As modified by descriptors such as 'invidious,' the term 'discrimination' refers to the making of distinctions that ought not to be made. When the distinction is inappropriate because it distinguishes among individuals, entities, or practices that are *legally* entitled to equal treatment, 'discrimination' takes on its legal meaning.

Contemporary antidiscrimination law is predominantly (or at least most controversially) concerned with distinctions drawn (or, in some cases, inappropriately *not* drawn) on the basis of socially defined personal attributes that have been and remain the bases for adverse treatment (such as religion, race, gender, national origin, and disability). The legal concept of discrimination is not limited in its historical origins to group-based distinctions of this sort. As a concept in American constitutional rhetoric and law, the term also designates protectionist distinctions between states and countries engaged in interstate and foreign commerce. For purposes of this article, however, our focus shall be on 'discrimination' of the inter-group sort, in part because it is that form of discrimination that has most interested the social scientific community.

The development of any nation's antidiscrimination law is deeply dependent upon its own history and culture. In the field of race and ethnicity, the need for legal redress for race discrimination arises out of variable histories of immigration and its selective suppression, slavery and indentured labor, colonialism, and genocide. Social movements calling for legal and social change are culturally specific as well, although influenced by past experiences elsewhere in the world. Movements to identify and redress gender discrimination are also shaped by local settings. This is true not merely because the degree to which gender takes center stage depends in part upon whether and how the stage must be shared with the issue of race, but also because the variable social situation of women has much to do with how women will conceptualize their own equality and the spheres of activity in which they will press their legal claims. This essay will focus on efforts by the United States at the federal level, but will conclude with a discussion of leading issues from several other countries.

1. *Federal Antidiscrimination Law in the United States*

Federal antidiscrimination efforts reach a wide array of social and legal arenas. Some of the most important of these are education, employment, voting, public and private housing, public accommodations (e.g., transportation), bank lending, government contracting, marriage and parenting, and the administration of criminal justice.

In light of the American history of slavery, civil war, failed post-civil-war reconstruction of the South, and the mid twentieth-century civil rights movement aimed

at ending the pervasive segregation of African-Americans, race has been the most frequent starting point for American antidiscrimination law. The Fourteenth Amendment of the Constitution of the United States, itself a product of the Civil War experience, contains an Equal Protection Clause that serves as the primary source of federal constitutional antidiscrimination doctrine. Federal equal protection analysis generally assigns different 'levels of scrutiny' to different types of discriminatory classifications. For example, race discrimination is subject to 'strict scrutiny'; gender discrimination to 'intermediate scrutiny'; and age discrimination to 'rationality review,' the same level of review that is given to ordinary economic legislation for a general survey, see Tribe (1978). Governmental classifications rarely survive strict scrutiny and struck down rarely under rationality review, although there are some exceptions. Because the Fourteenth Amendment does not reach discrimination by purely private parties, much of the federal antidiscrimination effort in the United States has been made under the auspices of antidiscrimination statutes.

The American law of discrimination starts with high-sounding commitments in seminal statutes or Supreme Court opinions, but its development rapidly takes the form of the elaboration of highly specific structures of proof for dealing with discrimination claims arising under the various sources of law and in the various areas of social practice surveyed above. One result is that antidiscrimination law has developed a number of distinct specialities, each concerned with a separate jurisprudence of discrimination in its own field. Another result is that, in American antidiscrimination law, the devil is in the details. By shaping proof structures, and by applying civil procedure and evidence rules to discrimination cases, courts shape the likelihood that plaintiffs will succeed in their claims (Albiston 1999).

1.1. Employment

The leading federal antidiscrimination statute in the employment field, Title VII of the Civil Rights Act of 1964 (known as 'Title VII') was stimulated by the violent resistance by the South to court-ordered racial desegregation and to the 1960s civil rights movement. Title VII prohibits discrimination on the basis of race, color, sex, religion, and national origin. Although feminist activists lobbied for the inclusion of sex discrimination, the statute as originally proposed left sex discrimination out. It was included through an amendment proposed by a congressional opponent of the statute, whose confidence that including sex would block passage of the statute proved misplaced. Once enacted, however, the prohibition of sex discrimination has been a robust and creative aspect of statutory enforcement.

Employment discrimination on the basis of age and disability was subsequently addressed in the Age Discrimination in Employment Act (1967) and Americans with Disabilities Act (1993). Legally-focused civil rights movements on behalf of non-racial groups (such as the development of the disability-rights movement and its involvement in the design of the Americans with Disabilities Act) have contributed to these legal developments. Not all groups have met with success: for example, Congress has yet to enact legislation barring employment discrimination on the basis of sexual orientation.

In the early years of race discrimination litigation under Title VII, the Supreme Court developed three basic proof structures: 'individual disparate treatment,' 'pattern and practice,' and 'disparate impact.' None of these terms appear in Title VII itself, which speaks most generally in terms of discrimination 'because of' race, sex, and so forth. The 'individual disparate treatment' and 'pattern and practice' proof structures turn upon proof of the employer's discriminatory intent. The disparate impact proof structure does not. Relying instead upon proof of discriminatory effects of racially-neutral practices.

In the 'individual disparate treatment' case, intent is an inference drawn by a factfinder (judge or jury) from a variety of types of circumstantial evidence aimed at refuting the possible nondiscriminatory reasons for an adverse employment action. While this sounds relatively straightforward, the Supreme Court surprised the plaintiffs' bar in 1993 by holding that merely establishing the required prima facie case and disproving the employer's stated justification for its action does not entitle the plaintiff to a victory (Malamud 1995), and has only begun to clarify the circumstances under which doing so entitles the plaintiff to go to trial.

In 'pattern and practice' cases, plaintiffs (most often a group of plaintiffs using the procedural mechanism of the class action) rely primarily on labor market analysis and statistical evidence to create an inference of intentional discrimination. The theory underlying the use of statistics is that, absent discrimination, each employer's workforce would resemble the racial composition of the relevant labor market. Employers' defense consists chiefly of attacks on the appropriateness of the plaintiffs' labor market definition and on the accuracy of their statistical analysis. Because the pattern and practice case is so reliant on expert testimony, these cases are expensive to litigate and often push the courts beyond their level of technical expertise.

In a disparate impact case, a plaintiff or plaintiff class challenges a facially-neutral employment practice (for example, a test) on the grounds of its disparate adverse effects on members of her group. So, for example, in the case of *Griggs* v. *Duke Power Co.* (401 US 424 [1971]), markedly lower proportions of blacks and Latinos than whites in the relevant labor market

passed standardized aptitude tests and had high school diplomas, leading to diminished access to preferred jobs in the employer's workforce. Absent a showing of 'business necessity' or 'job relatedness,' the Supreme Court held, employers could not use tests and other hiring criteria that placed 'arbitrary headwinds' in the way of the progress of traditionally excluded groups. In 1988 and 1989, the Supreme Court made it markedly harder for plaintiffs to win cases under the disparate impact proof structure; legislation in 1991 rewrote the statute to eliminate some but not all of the effects of those decisions. (See Hunter and Shoben 1998, for a discussion in a comparative context.)

Even after Congress passed clarifying legislation in 1991, elements of the disparate impact proof structure remain uncertain. Because these cases involve comparison-group definitions, questions are raised in every case as to whether the reference group should be employer's existing labor force, the applicant pool, or some other definition of the relevant labor market. The current standard for showing disparity has been criticized for failing to comport with social-scientific definitions of statistical significance (See, e.g., Meier et al. 1984). The precise conceptual content of the phrases 'business necessity' and 'job relatedness' is still subject to doubt, particularly when disparate impact analysis is used to challenge employment practices other than paper-and-pencil tests (where 'job relatedness' has long been tested through psychometric techniques of test validation). In addition, while the use of disparate impact analysis is now provided for by statute in the case of Title VII and the Americans with Disabilities Act, its availability under federal age discrimination legislation is open to serious doubt.

One area in which Title VII law has been relatively stable is the law of affirmative action. Under *United Steelworkers of America* v. *Weber* (442 US 193[1979]), employers are permitted to use affirmative action (if it is consistent with standards laid out in that opinion), and the Supreme Court reiterated more recently that the employer does not need to prove that its own prior policies were discriminatory in order to do so (*Johnson* v. *Transportation Agency of Santa Clara County*, 480 US 616 [1987]). The debate among the justices in these cases was open and virulent, and yet affirmative action in employment has not provoked recent challenge (in contrast to the status of affirmative action under the constitution, discussed below). The Supreme Court's embrace of affirmative action in employment is not without limits, however, and as of 2001 there is a mismatch between current social practice and the Supreme Court's past opinions. The Supreme Court has never held that the desire for greater diversity is a permissible basis for affirmative action under Title VII but, despite that fact, diversity is the major stated justification for many affirmative action programs. In addition, the Court has never held that Title VII permits the use of affirmative action principles in making layoff decisions.

The example of American employment discrimination law shows that the true state of antidiscrimination law cannot be assessed merely by looking at the written laws themselves. One must also be aware of how the statutes are being interpreted by courts and/or relevant administrative agencies; of how generally applicable procedural and evidentiary rules are used in discrimination cases; of the expense of litigation; and of the practical availability of experts with necessary skills and experience. One must also have a sense of trajectory (is the law becoming more or less friendly to plaintiffs over time?) and of symmetry (are similar laws being interpreted in similar or dissimilar ways?).

1.2. Other Areas of American Law

In considering proof structures outside the employment discrimination field, a key factor is the unavailability of the disparate impact cause of action in constitutional cases arising under the Fourteenth Amendment. (The Supreme Courts so held in the case of *Washington* v. *Davis* 26 US 229 [1976]). Although *Washington* v. *Davis* was in fact an employment-testing case, the Court's decision was likely motivated by the wide range of government programs that might be subject to challenge under the Constitution.

Because of the lack of the disparate impact route, showing discriminatory intent is the key to constitutional antidiscrimination litigation. Needless to say, it can be quite difficult to determine the intent of a legislative body. One conspicuous example is in the field of legislative redistricting, in which a string of equal-protection reverse-discrimination suits have been brought by white voters to challenge majority–minority districts created pursuant to the Voting Rights Act. The Court has held that the Equal Protection Clause is violated (under strict scrutiny) when race becomes the predominant factor motivating the redistricting decision. The Supreme Court is still wrestling with how to characterize the intent of the state legislature when, in order to protect incumbents, it creates majority–minority districts that are less compact and/or less contiguous than ordinary districts (see Pildes and Niemi 1993).

In primary- and secondary-school desegregation litigation, close judicial scrutiny was triggered in years past by findings that boards of education were operating their schools in an intentionally discriminatory fashion. Courts retained jurisdiction to administer complex court-ordered desegregation plans, often for decades. Until such time as the districts are declared 'unitary' by the presiding court, practices that were not themselves intentionally discriminatory could be challenged as perpetuating the effects of past intentional discrimination. In recent years, however, the trend has been for courts to conclude that the school districts under their jurisdiction have achieved

'unitary' status and that any continuing inequalities within the system no longer bear an organic connection to the district's past history of intentional discrimination (Orfield and Eaton 1996).

In higher education, affirmative action has been the focus of major political and judicial attention in a variety of fields under constitutional law. The continued constitutionality of affirmative action in higher education under the Supreme Court's decision in *Regents of the State University of California* v. *Bakke* (438 US 265 [1978]) question, and challenges have been brought to affirmative action programs in the undergraduate and professional schools of several major state universities. Whether the Supreme Court will re-commit itself to affirmative action in higher education when the appropriate case comes before it remains to be seen.

Criminal justice in the United States has a significant racial dimension, one that contributes to the gap in perception of the fairness of the system between blacks and whites. Some important efforts to redress racism in criminal law enforcement have met with failure, in part because of the difficulty of persuading courts of discriminatory intent through statistical evidence. For example, in *McCleskey* v. *Kemp* (481 US 279 [1987]), the Court rejected the claim, based on extensive statistical evidence, that the death penalty as practiced by state governments is unconstitutionally discriminatory (Baldus, Pulaski, Woodworth 1983). (See *Death Penalty*.)

2. *International Developments*

As other nations and transnational entities address the discrimination in their own midsts, they learn not only from their own histories but also from the successes and the failures of the American model. A few examples must suffice.

The new constitution of South Africa expressly permits constitutional challenges to be brought under disparate impact theory, by prohibiting both direct and indirect discrimination. The Constitutional Court of South Africa is a court of great sophistication and deep commitment to comparative constitutional analysis. It will thus serve as a counterexample to the United States, which early eschewed constitutional disparate impact theory and has never looked back.

The law of free speech under the First Amendment of the United States constitution has stood in the way of pervasive hate speech regulation in the US. In contrast, out of the historical experience of Germany has grown a strong commitment in German law to prohibiting the group-hate speech exemplified by expressions of anti-Semitism (Stein 1986; see Schauer and Pildes 1999). Notwithstanding these legal prohibitions, skinhead and other white racialist groups in Germany continue to attack perceived outsiders to their idealized homogeneous ethnic German society, just as racist attacks continue to occur in the United States.

In the field of gender discrimination in employment, the theory of 'comparable worth' is an important (if still unclear) part of the gender discrimination law of the European Community and some of its member states (e.g., Ellis 1998, Libeson 1995). In comparable-worth analysis, women use the tools of job and wage analysis to demonstrate that they do not receive the same returns on effort, education, and skill as do occupants of traditionally-male jobs. In contrast comparable worth has not made significant inroads into American law (Blum 1991).

Affirmative action (under its many names) and its surrounding controversies are international phenomena. In Europe, the European Court of Justice has upheld the use of affirmative action in favor of women seeking promotion, but only when they are of 'equal suitability, competence and professional performance' and no 'reasons specific to an individual [male] candidate tip the balance in his favor.' (See Zuleeg 1999, Cox 1998). In contrast, Indian law has 'reserved' scarce positions for members of 'backward castes and classes' for decades, including legislative seats (Galanter 1984). In South Africa, the post-apartheid constitution explicitly permits affirmative action. The slowly-developing case law on the issue has met with controversy as members of the intermediate racial groups in the old apartheid system (coloreds and Indians) find themselves in competition with blacks for scarce positions (e.g., Higginbotham 1999, Adam 1997). See *Affirmative Action: Comparative Policies and Controversies*.

In surveying examples, one must not forget that in many countries, antidiscrimination law is still in its infancy. To choose one example, Japanese employment discrimination law has only recently been used to challenge the pervasive exclusion of professionally educated women from professional-level jobs (e.g., Fan 1999, Yamakawa 1999).

3. *The Role of Social Science in Antidiscrimination Legal Practice*

As should be apparent from the above survey of American antidiscrimination practice, social science has a crucial role to play in proving discrimination. Social science has also been an important reality check, by asking important empirical questions both about the efficacy of current antidiscrimination laws and the variables underlying the development of effective antidiscrimination law. Social science has played the weakest role, however, in the area where it may well have been the most helpful—that is, in developing the law's basic concepts of discrimination.

In the employment discrimination field, expert testimony from social scientists is a necessity in disparate impact and pattern and practice cases. (See

Expert Testimony). At a minimum, several types of social scientific expertise must be brought into a case. In both disparate impact and pattern and practice cases, labor market experts are necessary to formulate and to defend definitions of the relevant labor market, using census data, industry surveys, and other sources. Testing psychologists (psychometricians) play a crucial role in disparate impact cases both in designing employment screening tests in such a way as will allow them to withstand legal scrutiny, and in challenging the designs of tests in litigation. Statisticians are needed in both types of cases to ensure that appropriate data are collected through the court-administered evidence discovery process, to formulate an affirmative case from the data, and to critique the rival statistical story being told by the opposing party. Expertise from a number of disciplines is drawn upon to demonstrate or to refute the extent to which underrepresentation of groups results from their lack of interest in particular categories of jobs (e.g., Schultz 1990). Thus, disparate impact and pattern and practice cases are universally viewed as battles of the experts—which is a problem for the untrained judges and juries who must evaluate the evidence. Expert testimony has also found its way into individual disparate treatment cases. In *Price Waterhouse* v. *Hopkins* (490 US 228 [1989]), for example, expert testimony was admitted in order to demonstrate that certain statements by supervisors were based on sexual stereotypes.

While employment discrimination litigation employs the talents of psychologists, labor economists, and statisticians, the field of voting rights has brought political scientists into the expert witness role. Legislatures employ sophisticated demographic and political models when they redistrict, and the political scientists who created these models have played an important role in defending them and their results in court. Educational desegregation litigators utilize the expertise of specialists not only in education but also in demography, because patterns of housing segregation play an important role in long-term educational segregation.

Social scientists have ascertained the effectiveness of existing antidiscrimination frameworks both in the United States and elsewhere (e.g., Donohue and Heckman 1991, McCrudden et al. 1991, Edelman 1990). Social scientists have also contributed to the legal debate by re-examining legal assumptions about the mechanisms through which discrimination takes place. In the area of comparable worth, for example, US-based researchers have argued that gender wage disparities are caused by organizational practices that are not likely to be remedied by 'putting comparable worth as an overlay on existing compensation systems (Nelson and Bridges 1999). On a more theoretical level, empirical comparative studies in political science can shed light on the social and institutional circumstances that lead constitutional courts in some countries to be particularly willing to use their authority to articulate and to extend anti-discrimination norms (e.g., Epp 1998).

Where social science has played less of a role is in the development of the law's basic concepts of discrimination. Contemporary studies of prejudice in social psychology and elsewhere are critical of the intent-based model of discrimination that permeates American law. Perhaps the lack of penetration by social scientists into the conceptual core of American anti-discrimination law is a problem of origins. American antidiscrimination law grew out of the blatant intentional discrimination practiced against African Americans. In *Brown* v. *Board of Education* (347 US 483 [1954]), social scientific testimony convinced the Court that intentional segregation stigmatizes blacks; by so doing, it fostered the nation's hope that desegregation would allow blacks and whites to put ancient hatreds and mistrusts to rest. Close to fifty years of life post-*Brown* has put that optimism to the test, perhaps to a greater degree than the courts are prepared to admit. When they are not ready to admit the problem, legal actors show little willingness to turn to social science either to redefine it or to solve it.

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on its Effectiveness; Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Civil Liberties and Human Rights; Discrimination, Economics of; Discrimination: Racial; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Gender, Class, Race, and Ethnicity, Social Construction of; Human Rights, History of; Injustice: Legal Aspects; Justice and its Many Faces: Cultural Concerns; Prejudice in Society; Race and the Law; Sex Differences in Pay

Bibliography

Adam K 1997 The politics of redress: South African style affirmative action. *Journal of Modern African Studies* **35**(2): 231–49

Albiston C 1999 The rule of law and the litigation process: The paradox of losing by winning. *Law Society Review* **33**: 869–910

Baldus, Pulaski, Woodworth 1983 Comparative review of death sentences: An Empirical study of the Georgia experience. *J. Crim. L. & C.* **74** 661, 674, n. 56

Blum L M 1991 *Between Feminism and Labor: The Significance of the Comparable Worth Movement*. University of California Press, Berkeley, CA

Cox K 1998 Positive action in the European Union: From Kalanke to Marschall. *Columbia Journal of Gender and Law* **8**(1): 101–42

Donohue J J III, Heckman J 1991 Continuous versus episodic change: The impact of civil rights policy on the economic status of blacks. *Journal of Economic Literature* **29**(4): 1603–43

Edelman L B 1990 Legal environments and organizational governance: The expansion of due process in the American workplace. *AM. J. SOC.* **95**: 1401
Ellis E 1998 *EC Sex Equality Law*. 2nd edn. Clarendon Press, Oxford, UK
Epp C R 1998 *The Rights Revolution*. University of Chicago Press, Chicago
Fan J S 1999 From office ladies to women warriors?: The effect of the EEOL on Japanese women. *UCLA Women's Law Journal*. **10**(1): 103–40
Galanter M 1984 *Competing Equalities: Law and the Backward Classes in India*. University of California Press, Berkeley, CA
Higginbotham F M 1999 Affirmative action in the United States and South Africa: Lessons from the other side. *Temple International and Comparative Law Journal* **13**: 187–230
Hunter R C, Shoben E W 1998 Disparate impact discrimination: American oddity or internationally accepted concept? *Berkeley Journal of Employment and Labor Law* **19**(1): 108–52
Libeson S J 1995 Reviving the comparative worth debate in the United States: A look toward the European Community. *Comparative Labor Law Journal* **16**(3): 358–98
Malamud D C 1995 The last minuet: Disparate treatment after Hicks. *Michigan Law Review* **91**(8): 2229–324
McCrudden C, Smith D J, Brown C 1991 *Racial Justice At Work: Enforcement of the Race Relations Act 1976 in Employment*. Policy Studies Institute, London
Meier P et al. 1984 What happened in Hazelwood. *American Bar Foundation Research Journal* **1**: 139–86
Nelson R L, Bridges W P 1999 *Legalizing Gender Inequality: Courts, Markets, and Unequal Pay for Women in America*. Cambridge University Press, Cambridge, UK
Orfield G, Eaton S E 1996 *Dismantling Desegregation: The Quiet Reversal of Brown v. Board of Education*. The New Press, New York
Pildes R H, Niemi R G 1993 Expressive harms, 'bizarre districts,' and voting rights: Evaluating election-district appearance after Shaw v. Reno. *Michigan Law Review*. **92**(3): 483–587
Schauer F, Pildes R H 1999 Electoral exceptionalism and the First Amendment. *Texas Law Review* **77**: 1803–36
Schultz V 1990 Telling stories about women and work: Judicial interpretations of sex segregation in the workplace in Title VII cases raising the lack of interest argument. *Harvard Law Review* **103**(8): 1749–843
Stein E 1986 History against free speech: the new German laws against the 'Auschwitz' and other 'Lies'. *Michigan Law Review* **85**: 277–328
Tribe L H 1978 *American Constitutional Law*, Vol. 2
Yamakawa R 1999 We've only just begun: The law of sexual harassment in Japan. *Hastings International and Comparative Law Review* **22**: 523
Zuleeg M 1999 Gender equality and affirmative action under the law of the European Union. *Columbia Journal of European Law* **5**: 319–28

D. Malamud

Discrimination, Economics of

1. Discrimination

In the abstract, 'discrimination' refers to distinctions or differentiation made among objects or individuals. In economics, as in ordinary parlance, discrimination carries a pejorative connotation. The term is generally reserved for distinctions that are socially unacceptable and economically inefficient. Economic theories of discrimination usually fall into one of three broad categories: tastes and preferences of economic agents, economic power, and statistical discrimination. Unlike the world of theory, empirical documentation of discrimination is subject to much ambiguity and differences of interpretation.

2. Theories of Discrimination

2.1 Tastes and Preferences

Publication of *The Economics of Discrimination* (1957) by Gary Becker was instrumental in making discrimination a legitimate and popular topic for study by economists. The tastes and preferences approach adopted by Becker was pure neoclassical economics. This approach appeals to tastes for discrimination by utility-maximizing economic agents. Tastes for discrimination are embodied in discrimination coefficients that are susceptible to measurement by the economist's monetary yardstick. Agents' tastes for discrimination are manifested by a desire to minimize economic transactions with members of certain demographic groups. While discriminating agents may be prepared to forfeit income in order to indulge their tastes, the effects of their discriminatory behavior can lead to segregation or reduced incomes for the groups discriminated against. In addition to employer tastes for discrimination, Becker also considered the effects of tastes for discrimination by consumers and fellow workers. Under some circumstances, competitive pressures can ameliorate the effects of discriminatory behavior.

The literature provides support for Becker's contention that labor-market discrimination is lessened by product market competition. Ashenfelter and Hannan (1986) present statistical evidence that within the banking sector the employment opportunities for women are negatively related to market concentration. Peoples and Saunders (1993) show that deregulation of trucking in the USA lowered the white/black wage ratios among both union and nonunion drivers. This competitive effect was identified with the reduction of rents accruing to whites.

One avenue for exploring the effects of discrimination by fellow workers is to examine the effects of unionism on racial wage differentials. An early study by Ashenfelter (1972) showed that the average wage of black workers relative to the average wage of white workers in the USA was higher in the union sector than in the nonunion sector. In particular, Ashenfelter estimated that compared with the complete absence of unionism, the black/white wage ratio was 4.0 percent higher in the industrial union sector, and 5.0 percent lower in the craft union sector. These results suggest that the union sector is less discriminatory than the

nonunion sector, and that industrial unions are relatively more egalitarian compared with craft unions. Peoples (1994) found evidence not only that the returns to union membership are higher for blacks than for whites, but also that racial wage differentials were higher among nonunion workers in the highly concentrated industries.

2.2 Market Power

Prior to Becker, early work by economists on the topic of discrimination focused on pay differentials between men and women. Joan Robinson's (1933) classic on imperfect competition provided an early example of the economic power theory of discrimination. Robinson considered the case of a monopsonistic employer (single employer labor market) of men and women who are equally productive (perfect substitutes in production). She proved that if the labor supply elasticity (responsiveness to wages) were smaller for women, a profit-maximizing employer would equate the incremental employment costs of men and women by offering women a lower wage.

Madden (1975) explored the circumstances under which the labor-supply elasticities of women would be lower, and how male monopoly power insulates men from any monoposony power that might exist in labor markets. The key element was the existence of job segregation which reduced competition from women. It was shown that under these circumstances women would receive lower wages than men, if both worked in equally productive, but different, sets of jobs. Earlier, Bergmann (1971) advanced the occupational crowding hypothesis to explain income inequality between white male workers and otherwise equally capable nonwhite workers. Crowding nonwhites and women into a narrower range of occupations raises the incremental value of white males in their occupations and lowers the incremental value of all workers in the occupations dominated by nonwhites and women.

The ubiquity and self-reinforcement of a broad range of discriminatory behavior in society was examined in Thurow (1969). For example, discrimination in schooling opportunities has the effect of diminishing the subsequent earnings potential of minority workers. This reduces the costs to discriminating employers of paying lower wages to minority workers, as the latter will have fewer skills to begin with. Lower wages reduce the cost of housing discrimination because better housing is beyond the incomes of most minority workers.

2.3 Statistical Discrimination

Statistical discrimination is an approach that relies upon risk and information costs rather than tastes for discrimination or market power to explain income disparities among different demographic groups in society (Aigner and Cain 1977). One example is that of two groups who are equally productive on average, but the productivity signal, for example, test scores, is a noisier signal of productivity for one of the groups. A risk-averse, profit-maximizing employer might offer lower wages or less desirable positions to applicants from the demographic group with the noisier signal. Consequently, employment decisions are made on the basis of group characteristics as well as on the basis of individual characteristics.

3. Wage Decompositions

In the study of wage inequality, the most commonly used methodology for apportioning wage gaps between two demographic groups to productivity and discrimination can be traced to Blinder (1973) and Oaxaca (1973). The basic idea is quite simple. One first estimates wage equations for each of the demographic groups from a sample of workers. These wage equations are mathematical formulas relating wages to observed productivity factors such as education and experience. In the absence of direct labor-market discrimination, equity is characterized as a situation in which both demographic groups are expected to be compensated according to the same wage (salary)-determination formula. In this scenario the only reason for the existence of a wage gap would be that the two groups differ on average with respect to the observed wage (productivity)-determining characteristics. This wage gap is commonly referred to as the 'endowment' or human capital effect.

The difference between the actual wage gap and the endowment effect is often referred to as the 'unexplained gap.' Many practitioners regard the unexplained gap as an estimate of discrimination. Those who prefer the term 'unexplained gap' reason that this estimate is in the nature of a residual and as such may include the effects of productivity factors not explicitly taken into account in the wage-determination formula. Hence, the measure may be a biased estimate of discrimination. From an econometric standpoint, the bias can go in either direction. The counterargument is that there are always factors omitted from any statistical relationship and that there is no reason to suppose that these are anything but random in the present context. In addition, the same sort of methodology is very commonly used to estimate other types of wage gaps, for example union/nonunion, north/south, manufacturing/nonmanufacturing. In these contexts the term 'unexplained gap' is not used.

According to the Blinder/Oaxaca standard decomposition methodology, one of the two demographic comparison groups is designated as the reference group. If, for example, one is interested in male/female wage differentials, one might select males as the reference group. This means that the statistically estimated wage formula for males serves as the

nondiscriminatory wage structure. Operationally, one estimates the average wages for both males and females by applying the estimated male wage formulation to each group's wage-determining characteristics. The gender difference in these predicted average wages is the 'endowment' effect, since the same wage formula is applied to the characteristics of both groups. The difference between the original unadjusted wage gap and the gap attributable to the 'endowment' effect is by definition the result of different wage formulas for males and females. This residual gap is what is termed either the 'unexplained gap' or the gap attributable to direct labor-market discrimination. Therefore, the decomposition technique is a method for decomposing an observed wage gap into an endowment component and an unexplained/discriminatory component.

3.1 Identification of Individual Components of Discrimination

Apart from differences in what researchers call the 'residual gap,' there are other ambiguities in the standard decomposition technique. First, it would seem desirable to be able to identify the contribution of each wage-determination characteristic to the endowment component and the discrimination component. For example, one could identify the contribution of gender differences in a characteristic such as work experience to the overall endowment effect as the gender difference in average work experience multiplied by the pecuniary return to an additional unit of work experience estimated from the male wage equation. Or consider the case of a wage-determining characteristic defined by a set of mutually exclusive categories such as educational levels, for example, 'less than high school,' 'high school,' 'some college,' and 'college graduate.' The estimated return to each included educational category is relative to the return on an arbitrarily selected educational category variable that is omitted from the wage equation. The monetary value of gender differences in the proportion of workers in the sample who are college graduates is simply calculated as the gender difference in the proportion of college graduates weighted by the estimated value of a college education for males relative to the omitted educational category. The sum of all of these endowment valuations equals the overall endowment effect.

Unfortunately, an identification problem can arise when one attempts to estimate the contribution of each wage-determination characteristic to the discrimination component. The problem is initially described in Jones (1983) and further extended in Oaxaca and Ransom (1999). For example, weighting the average amount of work experience for females by the male–female difference in the estimated return to an additional unit of work experience would constitute a measure of the contribution of gender differences in the return to experience to the overall discrimination measure. It turns out that this computation is affected by a simple shift in the experience measure. For example, potential experience is commonly calculated as either schooling-age-five or schooling-age-six. This is a seemingly innocuous shift in the experience measure, yet it will affect the estimated contribution to discrimination of gender differences in the return to experience. Because this change in the experience measure creates an offsetting change in the intercept term in the wage equations for males and females, it does not affect the overall estimate of discrimination. Nor does this change in the experience measure have any effect on the contribution of gender differences in work experience to the endowment effect.

This arbitrariness is more manifest when attempting to estimate the contribution to discrimination of gender differences in the returns to a variable defined by mutually exclusive categories. Returning to the education example, note that the contribution to measured discrimination of the gender difference in the returns to education is calculated as the sum of the gender differences in the estimated returns for each included educational category weighted by the proportion of females who are in each included category. As it turns out, this calculation is affected by the choice of which educational category variable is to play the role of the omitted reference category. The choice of omitted reference category is of course quite arbitrary. Fortunately, the overall estimated discrimination component is not affected by which categorical variables are selected to be the omitted reference categories. Also, the choice of omitted reference group has no effect on the estimated contribution to the endowment effect of gender differences in the values of the categorical variables.

3.2 Nondiscriminatory Wage Structures

To examine a second source of ambiguity in standard wage decompositions, imagine that the choice of reference group is reversed so that female workers are designated as the reference group. Now one estimates the (nondiscriminatory) average wages for both males and females by applying the estimated female wage formulation to each group's wage-determining characteristics. An ambiguity arises in that in general estimates of the endowment and discriminatory wage effects will depend on which demographic group is serving as the reference group. This ambiguity is a manifestation of the familiar index number problem. The researcher is trying to measure 'value' using alternative base weights.

As a measure of discrimination, the estimated residual gap is subject to quite different interpretations depending on the choice of reference group. The situation is one in which the presence of discrimination is manifested as a wage gap based on the fact that in actuality men and women do not face the same wage-

determination process. This can be viewed as either that women are underpaid relative to the male wage structure or else that men are overpaid relative to the female wage structure. The interpretation and estimated magnitude of each of the two components comprising the original unadjusted wage gap depend on which group's wage structure is regarded as the nondiscriminatory standard. To further add to the ambiguity surrounding wage-decomposition analysis, there is the theoretical possibility that which group is being discriminated against depends on whose wage-determination formula is being used as the nondiscriminatory standard.

It is clear that the second type of ambiguity surrounding the standard wage decomposition analysis has to do with identifying an appropriate nondiscriminatory wage standard. Ideally, one would want the wage structure that would prevail in a nondiscriminatory world to serve as the standard by which to measure actual discrimination. Under some assumptions, such a wage structure would be the one that would exist in a world characterized by perfect competition in product markets and labor markets. One would estimate the average wages for both males and females by applying the nondiscriminatory wage formulation to each group's wage-determining characteristics. The gender difference in these predicted average wages is the endowment effect. Again, discrimination is measured by the unadjusted gender wage gap net of the endowment effect. However, it is now possible to further decompose the discriminatory wage gap because the gap is in general the sum of two distinct components. The first component is the difference between the actual average wage of the male workers, and their predicted average wage based on the nondiscriminatory structure. The second component is the difference between the predicted average wage of females based on the nondiscriminatory wage structure, and their actual average wage.

These discriminatory wage components have interesting interpretations (Neumark 1988). The first component is a measure of the male wage advantage relative to the nondiscriminatory norm. This component can be viewed as favoritism toward males. The second component is a measure of the female wage disadvantage relative to the nondiscriminatory norm. This component can be viewed as pure discrimination against female workers. This methodology, in principle, allows researchers to sort out how much of the male–female discriminatory wage gap results from overcompensation of men, and how much stems from undercompensation of women. In the light of this approach, the conventional decomposition method is seen as a special case.

While seemingly attractive as a decomposition method, there is a practical difficulty in making operational the concept of a nondiscriminatory wage structure. One notion is the wage structure that would prevail had discrimination never existed. A slightly less challenging task might be to define the non discriminatory wage structure as the structure that would emerge in the immediate aftermath of a cessation of discrimination. In general, this is not the same as the wage structure that would have prevailed had discrimination never existed. This is because of a likely path-dependence feature in which discrimination permanently alters the evolution of wage structures. While a still formidable enterprise, it is possible to reach useful approximations to a wage structure that would emerge after the cessation of discrimination.

The basic idea underlying this approach is contained in Neumark (1988) and Oaxaca and Ransom (1999). A reasonable presumption is that with the cessation of discrimination between two demographic groups, the resulting wage structure would in some way be related to the preexisting separate wage structures. A natural approximation is the wage structure that is estimated on the basis of a combined sample obtained by pooling the data on the two demographic groups. Such a common wage structure can be shown to be a matrix-weighted average of the separately estimated wage structures. The weights are related to the precision by which each group's wage-equation parameters are statistically estimated. Neumark (1988) applies this methodology to the gender wage gap.

In Oaxaca and Ransom (1994) the methodology is generalized to allow for estimation of a nondiscriminatory wage structure as any matrix-weighted average of the separately estimated wage structures. Without restrictions on the weighting matrices, any common wage structure can be rationalized as a matrix-weighted average of the separately estimated wage structures. Reimers (1983) uses simple averages of separately estimated wage structures when comparing whites, blacks, and Hispanics.

3.3 Residual Decomposition

An important aspect of the study of wage inequality is trying to understand the components of changes or trends in wage inequality. Juhn et al. (1991) introduce a method of decomposing changes in the residual wage gap between two demographic groups into the effects of changes in the average rankings of the two groups in the residual wage distribution of the dominant group, and the effects of changes in the residual wage dispersion (standard deviation) of the dominant group. This approach is referred to as the Juhn–Murphy–Pierce (JMP) decomposition. It follows the practice of interpreting the wage gap left after controlling for group differences in observed characteristics as an unexplained differential.

In Blau and Kahn (1996) the JMP decomposition is adapted to analyzing gender wage gap differences between nine countries and the USA during the 1980s. Earnings are standardized to a full-time (40 hours) work schedule. The earnings models estimated with

Table 1
International gender wage gaps

Country	G	$G-G_{us}$	Quantities	Prices	Rankings	Dispersion
Australia	0.3101	−0.1198	0.0563	−0.0732	−0.0671	−0.0358
Austria	0.3195	−0.1059	0.0575	−0.1573	0.2813	−0.2873
Germany	0.3542	−0.0712	0.0318	−0.0973	0.2421	−0.2480
Hungary	0.4326	0.0129	−0.0567	−0.0287	0.6272	−0.5290
Italy	0.3249	−0.1008	0.0299	−0.0094	−0.0239	−0.0974
Norway	0.3490	−0.0761	−0.0760	−0.0142	0.5829	−0.5688
Sweden	0.2580	−0.1674	−0.0309	−0.0270	−0.0122	−0.0971
Switzerland	0.4377	0.0123	0.1030	−0.0061	0.0683	−0.1530
UK	0.4875	0.0621	0.0108	−0.0411	0.4232	−0.3310
USA	0.4254					

Source: Blau and Kahn 1996, Table 5.
G: The (unadjusted) gender wage differentials (in logs). $G–G_{us}$: The gender wage gap difference between each country and the USA. Quantities: The effects of gender differences in personal characteristics. Prices: The effects of gender differences in the returns to personal characteristics. Rankings: The effects of gender differences in percentile rankings in the wage distribution. Dispersion: The effects of international differences in wage dispersion.

microdata sets control for education, potential experience, union membership, industry, and occupation. Table 1 summarizes the Blau and Kahn findings in which the gender wage gaps in each country are compared with that of the USA. Column 2 reports the raw (log) wage differentials by country. The wage gaps range from 0.258 log points (approximately 26 percent) in Sweden to 0.4875 log points (approximately 49 percent) in the UK. Column 3 reports each country's wage gap minus that of the USA. It is this difference that is decomposed into the effects of international differences in the following: (a) gender differences in human capital characteristics (quantities); (b) the (male) returns to human capital characteristics (prices); (c) gender differences in the percentile rankings in the male wage residual distribution (rankings); and (d) differences in the (male) standard deviation of the wage residuals (dispersion). In six of the nine countries, the male human capital advantage is greater than that of the USA. Without exception, international differences in the returns to human capital and in the dispersion of wage residuals lower the gender wage gap in other countries relative to the USA. In six of the nine countries the less favorable rankings of women in the male wage residual distribution widens the gender wage gap relative to the USA. What emerges from this study is that the gender wage gap is relatively high in the USA because the greater wage residual inequality in the USA more than offsets the greater relative human capital of US women.

4. *Correcting Wage Inequities*

Although economists are presumably interested in measuring the extent to which discriminatory wage gaps represent overpayment of the dominant group versus underpayment of the subordinate group, custom and the legal environment lean more to the dominant group's wage structure as the nondiscriminatory standard. The idea is that inequities should be addressed by raising the wages of the subordinate group, as opposed to lowering the wages of the dominant group. Indeed, the US Civil Rights Act of 1964 prohibits equity wage adjustments that involve lowering any group's nominal wages. Thus, attempts to remove discrimination in the workplace in a way that would correspond to accepted economic principles of efficiency are constrained by legal and employee morale considerations.

4.1 *Econometric Modeling Approaches*

One might properly speculate about how econometric models of the wage-determination process can be used to address a salary inequity issue in the workplace. Examples from legal practice are provided in Finkelstein (1980). Oaxaca and Ransom (1999) report on several approaches on how to use statistical salary models to correct salary inequity within a firm. For illustrative purposes, the examples are in terms of gender inequity, but the methodology extends without modification to racial or ethnic salary inequity. The remedies range from simply compensating each female worker exactly according to the estimated male wage equation to constraining the salary equity adjustment process to award minimum settlements regardless of personal circumstances.

4.2 *Comparable Worth*

Comparable worth is a general policy for eradicating discriminatory wage disparities within firms. The idea

behind comparable worth is that jobs of equal value should pay equal wages. While this sounds like free market economics, at issue is how value is determined (Hill and Killingsworth 1989). Advocates of comparable worth envision a system in which points are awarded to each job based on the skills, effort, and responsibility required. Experts are responsible for evaluating a firm's jobs in a manner that does not bias wage determination against women. The objection against letting the market value the jobs is that occupational segregation dooms women to a subordinate wage position. An important question is whether or not comparable worth would provide employers with the incentives to eliminate job segregation, and whether or not women would switch among occupations that were equally compensated. Critics of comparable worth fear the massive inefficiencies that could emerge from interference with labor-market resource allocation.

5. *Self-selection in the Labor Force*

Estimation of labor-market discrimination in general is complicated by the observation that individual workers in a statistical sample generally do not constitute a random sample of the adult population. Whether or not to participate in the labor force will depend upon the individual's calculation of the net benefits of labor-force participation. While the net benefits are unobserved by researchers, the decision to participate is observed, as well as many determinants of the net benefits. For women, the presence of small children at home would likely be a factor in the decision to be in the labor market. The statistical equation corresponding to labor-force participation is known as a selection equation, since it describes a rule for individuals to select themselves in or out of the labor force.

Sample selection bias can arise when the random statistical errors in the selection equation are correlated with the random statistical errors in the wage equation that is being estimated for a self-selected sample of workers (Heckman 1979). Unless the probability of selection (which depends on the selection equation) is taken into account, the estimated parameters of the wage equation can be biased. The selection probabilities for the selected sample are controlled for in a variable known as the Inverse Mills Ratio. This variable is constructed from the variables and parameters of the selection equation.

A number of studies have attempted to correct for sample selection bias when dealing with occupational segregation or more generally labor-market discrimination (Reimers 1983, Dolton et al. 1989, Neuman and Oaxaca 1998). Neuman and Oaxaca (1998) explore alternative ways of estimating (ethnic and gender) labor-market discrimination from selectivity-corrected wage equations using a sample of Jewish professional workers drawn from the Israeli 1983 Census of Population and Housing.

6. *Progress and Future Directions*

Smith and Welch (1989) chronicle the economic progress of African-Americans over a 40-year period. Black male wages as a percent of the white male wage rose from 43.3 percent in 1940 to 72.6 percent by 1980. The documented wage gains of African-Americans were largely attributed to more and better education, migration, and general economic growth. Some evidence was found that affirmative action policies improved the employment and wage prospects of African-Americans. Data from the US Current Population Survey are consistent with the Smith and Welch findings. The median income of full-time, year-round black workers as a percent of the median income of full-time, year-round white workers rose from 68.1 percent in 1970 to 73.9 percent in 1998.

Empirical findings from gender labor-market studies are quite mixed, variable over time and across countries. Zabalza and Tzannatos (1985) examine the gender employment and wage effects of labor-market legislation in the UK on the employment and wages of women. Panel data, which contain detailed information on individuals over their working lives, have been used to study gender wage gaps in Denmark (Rosholm and Smith 1996). In a shift of emphasis from gender wage discrimination Flückiger and Silber (1999) examine both theoretically and empirically the phenomenon of gender labor market segregation with special emphasis on the Swiss labor market.

Empirical studies of racial-, ethnic-, gender-, and caste-based labor-market outcomes in developing countries may also be found (Birdsall and Sabot 1991, Psacharopolulos and Patrinos 1994). Intercountry comparisons of labor-market inequity between the developed and the developing countries will be the inevitable result of globalization. The implications of labor-market discrimination for international trade and specialization should occupy the attention of social scientists well into the twenty-first century.

See also: Comparable Worth in Gender Studies; Discrimination; Discrimination: Racial; Economic Development and Women; Egalitarianism: Political; Egalitarianism, Sociology of; Equality and Inequality: Legal Aspects; Equality: Philosophical Aspects; Gender and Feminist Studies in Economics; Income Distribution; Income Distribution: Demographic Aspects; Inequality; Inequality: Comparative Aspects; Labor Supply; Poverty, Sociology of; Prejudice in Society; Racism, Sociology of; Sex Differences in Pay; Social Inequality in History (Stratification and Clas-

ses); Social Stratification; Underclass; Urban Poverty in Neighborhoods; Wage Differentials and Structure; Wealth Distribution

Bibliography

Aigner D, Cain G 1977 Statistical theories of discrimination in the labor market. *Industrial and Labor Relations Review* **30**: 175–87

Ashenfelter O 1972 Racial discrimination and trade unionism. *Journal of Political Economy* **80**: 435–64

Ashenfelter O, Hannan T 1986 Sex discrimination and product market competition: The case of the banking industry. *Quarterly Journal of Economics* **101**: 149–4

Becker G 1957 *The Economics of Discrimination*. University of Chicago Press, Chicago

Bergmann B 1971 The effect on white incomes of discrimination in employment. *Journal of Political Economy* **79**: 294–313

Birdsall N, Sabot R (eds.) 1991 *Unfair Advantage: Labor Market Discrimination in Developing Countries*. The World Bank, Washington, DC

Blau F D, Kahn L M 1996 Wage structure and gender earnings differentials: An international comparison. *Economica* **63**(Suppl.): 29–62

Blinder A S 1973 Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources* **8**: 436–55

Dolton P J, Makepeace G H, Van der Klaauw W 1989 Occupational choice and earnings determination: The role of sample selection and non-pecuniary factors. *Oxford Economic Papers* **41**: 573–94

Finkelstein M 1980 The judicial reception of multiple regression studies in race and sex discrimination cases. *Columbia Law Review* **80**: 737–54

Flückiger Y, Silber J 1999 *The Measurement of Segregation in the Labor Force*. Physica-Verlag, Heidelberg, Germany

Heckman J 1979 Sample selection bias as a specification error. *Econometrica* **47**: 153–61

Hill M A, Killingsworth M R (eds.) 1989 *Comparable Worth: Analyses and Evidence*. Industrial Labor Relations Press, Ithaca, NY

Jones F L 1983 On decomposing the wage gap: A critical comment on Blinder's method. *Journal of Human Resources* **18**: 126–30

Juhn C, Murphy K M, Pierce B 1991 Accounting for the slowdown in black-white wage convergence. In: Kosters M H (ed.) *Workers and their Wages*. AEI Press, Washington, DC, pp. 107–43

Madden J F 1975 Discrimination—A manifestation of male market power? In: Lloyd C B (ed.) *Sex, Discrimination, and the Division of Labor*. Columbia University Press, New York, Chap. 6, pp. 146–74

Neuman S, Oaxaca R L 1998 *Estimating Labor Market Discrimination with Selectivity Corrected Wage Equations: Methodological Considerations and An Illustration From Israel*. Centre for Economic Policy Research, London

Neumark D 1988 Employers' discriminatory behavior and the estimation of wage discrimination. *Journal of Human Resources* **23**: 279–95

Oaxaca R L 1973 Male-female wage differentials in urban labor markets. *International Economic Review* **14**: 693–709

Oaxaca R L, Ransom M 1994 On discrimination and the decomposition of wage differentials. *Journal of Econometrics* **61**: 5–21

Oaxaca R L, Ransom M 1999 Using econometric models for intrafirm equity salary adjustments. Unpublished paper, University of Arizona, Tucson, AZ and Brigham Young University, Provo, UT

Peoples J 1994 Monopolistic market structure, unionization, and racial wage differentials. *Review of Economics and Statistics* **76**: 207–11

Peoples J, Saunders L 1993 Trucking deregulation and the black/white wage gap. *Industrial & Labor Relations Review* **47**: 23–35

Psacharopolulos G, Patrinos H (eds.) 1994 *Indigenous People and Poverty in Latin America: An Empirical Analysis*. The World Bank, Washington, DC

Reimers C 1983 Labor market discrimination against Hispanic and black men. *The Review of Economics and Statistics* **65**: 570–9

Robinson J 1933 *The Economics of Imperfect Competition*. Macmillan, London

Rosholm M, Smith N 1996 The Danish gender wage gap in the 1980s: A panel data study. *Oxford Economic Papers* **48**: 254–79

Smith J P, Welch F 1989 Black economic progress after Myrdal. *Journal of Economic Literature* **27**: 519–64

Thurow L 1969 *Poverty and Discrimination*. Brookings Institution, Washington, DC

Zabalza A, Tzannatos Z 1985 *Women and Equal Pay: The Effects of Legislation on Female Employment and Wages in Britain*. Cambridge University Press, Cambridge, UK

R. Oaxaca

Discrimination: Racial

Racial discrimination exists in straightforward actions but also in complex systems of social relations that produce racial inequities. Discrimination is not, as popularly thought, simply the accumulation of individual acts of unfairness between members of different groups. It consists as well of an elaborate web of institutional arrangements that produces group inequalities, a web that becomes racial discrimination when the groups perceive each other as separate 'races.' Studied throughout the social sciences, these inequalities can involve political power, economic resources, and cultural access.

1. Conceptual Complexities

Viewing racial discrimination as an institutional web casts a wide net but leaves room for considerable ambiguity. Disagreement about what constitutes discrimination stems from two sources—one ideological and political, the other empirical. First, because discrimination violates law and professed values in many countries, a judgment that unequal outcomes reflect discrimination is a call for unpopular change and costly remedies. Resistance to such change and

costs activates debate about the extent of discrimination. Second, deficiencies in analysis and evidence limit the ability of scientists to trace precisely the dynamic system of effects triggered by discrimination.

Consider first the political issue. The broadest definitions of discrimination assume that racial minorities have no inherent characteristics that warrant inferior social outcomes. Thus, all inequality becomes a legacy of discrimination and a social injustice to be remedied. Favored by the political left, this view appears to many as too sweeping.

By contrast, political conservatives favor a narrow definition that limits the concept's scope by including only actions intended to restrict a group's chances. Most social science specialists resist this reformulation for several reasons. An intentionality criterion returns the concept to the realm of psychology and deflects attention from restraining social structure. And the invisibility of intentions creates formidable obstacles for substantiating discrimination. Finally, the effects of discrimination are the same whether or not the causal institutional mechanisms were intended to produce group inequalities.

This rejection of the intentionality criterion leads to an important distinction between direct and indirect discrimination (Pettigrew 1985). Direct discrimination occurs at points where inequality is generated, often intentionally. When decisions are based explicitly on race, discrimination is direct. Indirect discrimination occurs when the inequitable results of direct discrimination serve as a basis for later decisions. Hence, discrimination is indirect when an ostensibly nonracial criterion becomes a proxy for race in determining social outcomes. Indirect discrimination perpetuates or magnifies the original injury.

Wage discrimination illustrates the distinction. Direct racial discrimination exists when employers pay equally qualified members of two races different rates for the same work. Indirect discrimination exists when the employers pay the two groups unequally because prior discrimination in employment, education, or housing created apparent differences in qualifications or channeled the groups into better- and worse-paying jobs. This direct/indirect distinction resembles the distinction often made by legal systems between disparate treatment and disparate impact. While intentional direct discrimination may have triggered the causal chain, the original injury is perpetuated and magnified by largely unwitting accomplices.

2. *Empirical Ambiguities*

This conceptual complexity carries over to impede the empirical study of discrimination. Apparently rigorous quantitative analyses often camouflage the crucial issues, as an examination of wage decomposition research reveals.

Assessments of racial discrimination produced by decomposing gross racial differences in social outcomes are common in sociology (Farley 1984) and economics (Gill 1989). This approach first assigns one segment of the gross racial differential to 'qualifications' and other factors deemed legitimate determinants of social rewards. The residual segment not demonstrably linked to 'legitimate' determinants of the outcomes often is presented as the estimate of discrimination. However, without better information than usually available and closer agreement on what constitutes discrimination, no unique estimate of discrimination is possible. The choice of control variables to index 'legitimate' determinants of social outcomes shapes the answers. Any appearance of scientific certitude is an illusion. Not surprisingly, discrimination estimates from this approach vary enormously.

Empirical research on group discrimination must mirror the phenomenon in its variety and complexity. The regression decomposition approach is useful but limited. Regression analyses could provide more pertinent information if based on more homogeneous job groups (Conway and Roberts 1994) and on structural equation models that test reciprocal causation. Most important, if the aim is to guide policy, a framework far more complex than the dichotomous discrimination-or-not approach is required. Research that traces the actual processes of institutional discrimination is essential (Braddock and McPartland 1987). Also needed is attention to victims' perceptions of discrimination and to the changes generated by anti-discrimination efforts.

3. *Racial Discrimination in the United States and Western Europe*

Both direct and indirect forms of racial discrimination can be found throughout the world. Most of the available research, however, focuses on the long-festering racial problems of the United States. In such long-term situations, injuries of indirect discrimination often are more extensive than those of direct discrimination. This conclusion does not imply that direct discrimination no longer exists. Employment complaint records reveal the continued operation of direct forms of discrimination. US women and minorities have filed 1.5 million job discrimination complaints since 1965. Not all such complaints reflect genuine discrimination, but analysts believe that false complaints are minimal in comparison with the amount of actual discrimination that goes unreported and thus undetected. Many major corporations have been found guilty of direct racial discrimination in the 1990s. And employment audits using paired, equally qualified applicants reveal widespread direct discrimination (Reskin 1998). Employers have admitted to interviewers that their stereotypes of minorities, par-

ticularly of innercity black men, lead them to avoid minority hiring (Neckerman and Kirschenman 1991).

Along with direct employment discrimination, indirect discrimination is pervasive. Many employers rely on word-of-mouth recruiting of new employees, especially for unskilled jobs. Thus the current workforce, often disproportionately white, reproduces existing racial inequalities in employment by bringing in white friends and relatives as the next generation of workers. This is just one aspect of 'business as usual' that represents indirect discrimination. Although word-of-mouth recruiting appears to be color blind, it serves to perpetuate racial inequalities (Reskin 1998).

Recent analyses of indirect discrimination in the USA have focused on wealth. Racial inequalities in assets dwarf disparities in other social outcomes. Although the magnitude of inequality in home equity is grave, racial inequality in liquid assets is even more marked. The average white family has been estimated to hold liquid assets 14 times greater than those of the average black family (Conley 1999).

The racial gap in assets results from a history of direct and indirect discrimination. Prohibitions against black self-employment and land ownership in the American South were an early part of the picture. More recently, occupations to which blacks were confined, agricultural and domestic work, were excluded from Social Security provisions. Direct and indirect discrimination by federal housing agencies, banks, realtors, and white homeowners have limited black Americans' opportunity to accrue wealth through home ownership (Conley 1999, Massey and Denton 1993, Oliver and Shapiro 1995).

The cumulation of disadvantage is particularly acute when the focus is assets. Since wealth reproduces itself exponentially, the racial gap in assets in the USA has continued to grow. And assets have important implications for other social outcomes such as education and labor market success. Where parents do not have the means to live in higher status neighborhoods, children do not have access to good public schools. And for most students, attendance at the best colleges and universities depends on parents' ability to provide tens of thousands of dollars per year to cover tuition and expenses. Labor market outcomes reflect educational disadvantage, but assets have other indirect effects as well. Access to networks of information crucial to job-seekers, proximity to lucrative employment, automobile transportation to work—all these are a function of assets, and black Americans are at a pronounced disadvantage (Conley 1999). Innovative strategies for intervention must be devised to break into the spiraling disadvantage produced by black Americans' deficits in wealth.

Analogous dynamics afflict the minorities of other nations. And a research literature on the subject is now emerging in Western Europe. Discrimination against the region's new immigrant racial minorities is pervasive (Castles 1984, MacEwen 1995, Pettigrew 1998). And both direct and indirect discrimination are involved, though the indirect forms are largely unrecognized in Europe.

Investigators have repeatedly uncovered direct discrimination in the UK (Amin et al. 1988). Controlled tests reveal the full litany of discriminatory forms involving employment, public accommodations, housing, the courts, insurance, and banks. Employment discrimination poses the most serious problem. In every European Union nation, minorities have markedly higher unemployment rates than natives.

As in the US, there are many reasons for minority unemployment disparities. The 'last-in, first-out' principle selectively affects the younger minority workers. Typically less skilled, they are more affected by job upgrading. Minorities also are more likely to be in older, declining industries. These patterns are not accidental. Planners put minorities into these industries for cheaper labor precisely because of their decline. Yet these multiple factors offer insufficient explanations for the greater unemployment of minorities. Veenman and Roelandt (1990) found that education, age, sex, region, and employment level explained only a small portion of the differential unemployment rates in The Netherlands.

Indirect discrimination arises from the inability to obtain citizenship. It restricts immigrants' ability to get suitable housing, employment, and schooling. In short, the lives of Europe's noncitizens are severely circumscribed. Castles (1984) contends that the guestworker system that brought many of the immigrants to Europe was itself a state-controlled system of institutional discrimination. It established the newcomers as a stigmatized outgroup suitable for low-status jobs but not citizenship. And racial prejudices add to this stigma. Widespread indirect discrimination, then, was inevitable for these victims of direct discrimination.

4. *Remedies for Racial Discrimination*

The most elaborate array of remedial measures is found in North America. Both Canada and the United States have instituted a battery of laws against many types of racial and other forms of discrimination. Court rulings in the two nations also have proven decisive. The effects of these legal actions are magnified by class action suits—the ability of single litigants to sue for remedy for their entire 'class' or group. Further remedies have involved various forms of 'affirmative action' in which special efforts are made to close group disparities caused by discrimination. All these approaches have led to narrowing racial disparities in social outcomes. But sharp racial inequities remain in both countries, and majority resistance has hardened against the most successful remedies.

Antidiscrimination remedies have been largely ineffective in Europe. Germany guarantees basic rights

only to citizens. So, the disadvantages of noncitizenship include the inability to combat discrimination. There is extensive German legislation to combat anti-Semitism and Nazi ideology, but these laws have proved ineffective in protecting noncitizens. The German constitution explicitly forbids discrimination on the basis of origin, race, language, beliefs, or religion—but not citizenship. Indeed, the Federal Constitutional Court has ruled that differential treatment based on citizenship is constitutional if there is a reasonable basis for it and if it is not wholly arbitrary. And restaurants can refuse service to Turks and others on the grounds that their entry might lead to intergroup disturbances (Layton-Henry and Wilpert 1994).

Few means of combatting discrimination are available in France either. Commentators often view discrimination as 'natural' and universally triggered when a 'threshold of tolerance' (*seuil de tolerance*) is surpassed (MacMaster 1991). Without supporting evidence, this rationalization supports quotas and dispersal policies that restrict minority access to suitable housing.

The Netherlands, United Kingdom, and Sweden have enacted antidiscrimination legislation that specifically applies to the new immigrant minorities. And the Dutch have instituted modest affirmative action programs for women and minorities. Yet this legislation has been largely ineffective for two interrelated reasons. First, European legal systems do not allow class action suits—a forceful North American weapon to combat discrimination. Second, European efforts rely heavily on individual complaints rather than systemic remedies. The UK's 1976 Act gave the Commission for Racial Equality power to cast a broad net, but individual complaints remain the chief tool (MacEwen 1995).

Individual efforts are unlikely to alter discrimination, because they are nonstrategic. Minorities bring few charges against the worst discriminatory firms, because they avoid applying to them. Complaints about job promotion are common, but they are made against employers who hire minorities. Effective anti-discrimination laws must provide broad powers to an enforcement agency to initiate strategic, institution-wide actions that uproot the structural foundations of discrimination.

5. *Two Summary Propositions*

The long-lasting character of racial discrimination means that the effects typically outlive the initiators of discriminatory practices. Discrimination is fundamentally normative; its structural web operates in large part independent of the dominant group's present attitudes or awareness. Hence, models based primarily on individual prejudice or 'rationality,' whether psychological or economic, will uniformly understate and oversimplify the phenomenon.

Discrimination feeds upon itself; it is typically cumulative and self-perpetuating. An array of research on African Americans has shown that neighborhood racial segregation leads to educational disadvantages, then 'to occupational disadvantage, and thus to income deficits, (Pettigrew 1985). To be effective, structural remedies must reverse this 'vicious circle' of discrimination.

Seen in this perspective, discrimination is far more intricate and entrenched than commonly thought. The complexity of discrimination presents major challenges to social scientific attempts to trace its impact. This complexity also precludes any simple correspondence between perpetration and responsibility for remedy. Broad social programs will be necessary if the full legacy of direct and indirect racial discrimination is finally to be erased.

See also: Ethnic and Racial Social Movements; Ethnic Conflicts; Ethnic Groups/Ethnicity: Historical Aspects; Gender, Class, Race, and Ethnicity, Social Construction of; Multiculturalism: Sociological Aspects; Race Identity; Racial Relations; Racism, Sociology of; Residential Concentration/Segregation, Demographic Effects of

Bibliography

Amin K, Fernandes M, Gordon P 1988 *Racism and Discrimination in Britain: A Select Bibliography, 1984–87*. Runnymede Trust, London

Braddock J H, McPartland J M 1987 How minorities continue to be excluded from equal employment opportunities: Research on labor market and institutional barriers. *Journal of Social Issues* **43**: 5–39

Castles S 1984 *Here for Good: Western Europe's New Ethnic Minorities*. Pluto Press, London

Conley D 1999 *Being Black, Living in the Red: Race, Wealth, and Social Policy in America*. University of California Press, Berkeley, CA

Conway D A, Roberts H V 1994 Analysis of employment discrimination through homogeneous job groups. *Journal of Econometrics* **61**: 103–31

Farley R 1984 *Blacks and Whites: Narrowing the Gap?* Harvard University Press, Cambridge, MA

Gill A M 1989 The role of discrimination in determining occupational structure. *Industrial and Labor Relations Review* **42**: 610–23

Layton-Henry Z, Wilpert C 1994 *Discrimination, Racism and Citizenship: Inclusion and Exclusion in Britain and Germany*. Anglo-German Foundation for the Study of Industrial Society, London

MacEwen M 1995 *Tackling Racism in Europe: An Examination of Anti-discrimination Law in Practice*. Berg, Washington, DC

MacMaster N 1991 The 'seuil de tolerance': The uses of a 'scientific' racist concept. In: Silverman M (ed.) *Race, Discourse and Power in France*. Avebury, Aldershot, UK, pp. 14–28

Massey D S, Denton, N A 1993 *American Apartheid, Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA

Neckerman K M, Kirschenman J 1991 Hiring strategies, racial bias, and inner-city workers: An investigation of employers' hiring decisions. *Social Problems* **38**: 433–47
Oliver M L, Shapiro T M 1995 *Black Wealth/White Wealth: A New Perspective on Racial Inequality*. Routledge, New York
Pettigrew T F 1985 New black–white patterns: How best to conceptualize them? *Annual Review of Sociology* **11**: 329–46
Pettigrew T F 1998 Reactions toward the new minorities of Western Europe. *Annual Review of Sociology* **24**: 77–103
Reskin B F 1998 *The Realities of Affirmative Action in Employment*. American Sociological Association, Washington, DC
Veenman J, Roelandt T 1990 Allochtonen: Achterstand en achterstelling. In: Schippers J J (ed.) *Arbeidsmarkt en Maatschappelijke Ongelijkheid*. Wolters-Noordhoff, Groningen, The Netherlands, pp. 241–61

T. F. Pettigrew and M. C. Taylor

Disgust, Psychology of

Disgust appears in virtually every discussion of basic emotions, including ancient Hindu texts and Darwin's *The Expression of the Emotions in Man and Animals* (Darwin 1965). Darwin defined disgust as referring to 'something revolting, primarily in relation to the sense of taste, as actually perceived or vividly imagined; and secondarily to anything which causes a similar feeling, through the sense of smell, touch, and even of eyesight' (p. 253). In the classic psychoanalytic treatment of disgust, Andras Angyal (1941, p. 395) claimed that 'disgust is a specific reaction towards the waste products of the human and animal body' (p. 395). Angyal related the strength of disgust to the degree of intimacy of contact, with the mouth as the most sensitive focus. Sylvan Tomkins (1963) has offered the more general contention that disgust is 'recruited to defend the self against psychic incorporation or any increase in intimacy with a repellent object' (p. 233). Derivative of Angyal's definition, Rozin and Fallon (1987, p. 23) expand on Angyal's definition, and define disgust as: 'Revulsion at the prospect of (oral) incorporation of an offensive object. The offensive objects are contaminants; that is, if they even briefly contact an acceptable food, they tend to render that food unacceptable.' Disgust is related to pollution, as used in the anthropological literature. In particular, Mary Douglas' (1966) view of polluting, meaning matter out of place, has direct relevance to disgust.

Most scholars recognize a special relation between disgust and ingestion; the facial expression of disgust represents an oral and nasal 'rejection,' the word 'disgust' in English and some other languages means 'bad taste,' and nausea, the physiological signature of disgust, is directly related to, and inhibitory of, ingestion.

However, it is also clear that, cross culturally, the emotion of disgust is elicited by a wide range of elicitors, most of which have nothing to do with food (Miller 1997, Rozin et al. 1993). In certain respects, disgust can be thought of as the emotion of civilization. The history and trajectory of disgust, from a food-related emotion to a much more general emotion, is the central concern of this entry.

1. *Disgust as an Emotion*

Disgust is the only basic emotion that is specifically related to a motivation (hunger) and to a particular bodily location (mouth). Along with other basic emotions, as described by Paul Ekman (Ekman 1992, Scherer and Wallbott 1994), disgust has a characteristic behavior (withdrawal), physiological state (nausea), and facial expression. The facial expression of disgust has received most attention from Ekman (Ekman and Friesen 1975), Izard (1971), and Rozin (Rozin et al. 1994). The expression centers on the mouth and nose, and is characterized principally by a wrinkling of the nose, a raising of the upper lip, and a gape. The mental or feeling component of disgust is often described as revulsion; it is typically of shorter duration than most other.

2. *The Meaning of Disgust*

Rozin et al. (1993), based on surveys in Japan and the United States, designate nine domains of disgust: certain foods or potential foods, body products, animals, sexual behaviors, contact with death or corpses, violations of the exterior envelope of the body (including gore and deformity), poor hygiene, interpersonal contamination (direct or indirect contact with unsavory human beings) and certain moral offenses. Disgust has to do with offensive elicitors which are wide in range, but exclude many other types of negative events, such as pain, and those that elicit the emotions of fear, sadness, and anger.

Rozin et al. (1993) propose that disgust originated in animals as a response to distasteful food. These authors hold that in cultural evolution, the output side of disgust (expression, physiology, behavior) remained relatively constant, but the range of elicitors and meanings expanded dramatically. Disgust came to mean offensiveness to the self or soul, rather than bad taste, its origin. Indeed, it was proposed that, in humans, distasteful items do not elicit the full emotion of disgust. The expansion of elicitors and meanings in cultural evolution was attributed to the process of preadaptation; use of a mechanism evolved in one system for a novel use in another system. For example, the mouth, and in particular the teeth and tongue,

evolved for processing food, but these structures were co-opted in human evolution for use in speech articulation. The expansion of the meanings of disgust is proposed to have occurred in four stages (Rozin et al. 1993).

2.1 Core Disgust

The initial form of disgust, that appears first in children, has to do with food, but primarily the nature and origin of the food, rather than its taste (Rozin and Fallon 1987). Core disgust has three components: (a) a sense of oral incorporation (and hence a linkage with food or eating), (b) a sense of offensiveness, and (c) contamination potency. The mouth is the principal route of entry of material things into the body; such things may be essential for life, but may also be very harmful. Disgust has to do with the harmful side, the risks of ingestion. The threat of oral incorporation is illustrated by a widespread belief that 'You are what you eat.'

Almost all elicitors of core disgust are whole animals, or animal body parts or body products. The great majority of food taboos concern animals and their products, and in most cultures the great majority of animals as potential foods are considered disgusting.

The contamination response (e.g., the rejection of a potential food if it even briefly contacted a disgusting entity) appears to be powerful and universal among adults, and is a defining feature of disgust. It is an example of the sympathetic magical law of contagion, which essentially holds that 'once in contact, always in contact' (Rozin and Nemeroff 1990).

2.2 Animal Nature Disgust

Common elicitors of disgust include inappropriate sex, poor hygiene, death, and violations of the ideal body envelope (e.g., gore, deformity, obesity). Contact with death and the odor of death (decay) are potent elicitors of disgust. Individuals who score high on disgust sensitivity also score higher on a fear of death scale (Haidt et al. 1994).

The centrality of death in disgust suggests a more general construal of disgust. While Freud (1953) conceived of disgust as a defense against culturally unacceptable sexual or coprophilic urges, disgust may be better understood as a defense against a universal fear of death. Ernest Becker (1973) has argued that the most important threat to the psyche is the certainty of death. Human beings are the only organisms that know they will die. Disgust, as a withdrawal system, may help to suppress thoughts or experiences that suggest human mortality.

More generally, anything that reminds us that we are animals elicits disgust (Rozin and Fallon 1987); eating, excretion, sex, soiling of the body surface, soft viscera within the body envelope, and death are all properties that humans share with animals. Cultures typically have extensive rules about these matters, including a wide range of prohibitions, in an attempt to humanize or deny our animal bodies. People who ignore these prescriptions are reviled as disgusting and animal-like (Rozin et al. 1993, see also Miller 1997 for a related conception). This conception of disgust links it to the concern of humans, frequently cited by anthropologists, to distinguish themselves from animals.

2.3 Interpersonal and Moral Disgust

Across cultures, there is a sense of disgust at touching most people who are not intimates, of making indirect contact with them, such as by wearing their clothing. Most other persons are contaminating, in this sense, and this sensitivity can serve to reinforce social boundaries. Cultures differ in degree of this sensitivity; it is particularly high in Hindu India, where it serves as an important means of maintaining caste distinctions.

In many, if not all languages, the word for 'disgust' is applied to certain moral violations. And the contamination property of disgust is also extended to the perpetrators of such immoral acts. It is possible that disgust functions, along with anger and contempt, as a basic moral emotion involved in expressing moral outrage at the actions of another (Rozin et al. 1999). Richard Shweder and his colleagues (Shweder et al. 1997) identify three types of moral codes that occur in cultures around the world. Violation of the community code, having to do with issues of hierarchy and duty, may map on to the emotion of contempt. Violations of the autonomy code deal with rights and justice, and may be associated with anger. The third code, divinity, focuses on the self as a spiritual entity and seeks to protect that entity from degrading or polluting acts. Disgust may be the emotion that is evoked when the divinity code is violated (Rozin et al. 1999).

3. Development of Disgust

Distaste, the origin of disgust, is present at birth, as in the innate rejection by neonates of bitter substances. Disgust first appears in North American children between the ages of 4 and 8 years, with the appearance of the contamination response, and rejection of foods because of their nature or origin. Disgust, as opposed to distaste, seems to be uniquely human, and is a product of culture. The contamination response, also uniquely human, is probably, in origin, a means of

protection from microbial contamination; however, it is acquired. Even feces, surely a source of such contamination, and a universal disgust substance among human adults, is not rejected by infants. Toilet training may well be the first disgust-generating experience for humans, but the means by which strong negative affect becomes associated with feces is unknown, as is the way that disgust 'spreads' to a wider range of elicitors.

Presumably parental reactions to disgusting things play a central role in the development of disgust. There is evidence for family resemblance in disgust sensitivity.

4. Disgust, History and Culture

Across cultures and history, the expression of disgust seems to vary rather little, while the elicitors and meanings vary widely. It seems that disgust was relatively infrequent and limited to a narrow range of elicitors in medieval Europe. The full scope of disgust in Western history, and its flowering as both an emotion of sensibility to the animal nature of humans, and a marker of immorality, is presented in William Miller's *Anatomy of Disgust* (Miller 1997).

Many cultures show specific attractions to entities that most other cultures find disgusting. For example, particular spoiled foods (e.g., cheese for the French, fermented fish sauce for Southeast Asians) are highly desired in particular cultures, and there are major differences in the sexual practices and contact with the dead across cultures.

The largest historical and cultural variation in disgust is probably associated with interpersonal and moral disgust. For example, disgust seems to center more on interpersonal and moral matters for Hindu Indians than it does for Americans.

5. Individual Differences

Results from a standardized test of disgust sensitivity, applied in Japan and the United States, indicate that sensitivity is somewhat higher in females and people of lesser education or lower social class (Haidt et al. 1994). As well, disgust has been implicated in two clinical phenomena: phobias and obsessive compulsive disorder.

6. Neural Correlates of Disgust

Very recent work by Sprengelmeyer, Young, Phillips, and others, in Germany and England, has established a link between disgust and the basal ganglia; in particular, damage to the basal ganglia, as in Huntington's disease, is associated with a sharply defined inability to recognize expressions of the emotion of disgust. These patients do not seem to have major deficits in disgust other than in the recognition of its expression.

7. Disgust and Related Emotions

Fear and disgust share a behavioral component, withdrawal based on a threat. Anger and sociomoral disgust share the appraisal of moral condemnation, and the common facial gesture of raised upper lip. Contempt shares the appraisal that someone is base and inferior to oneself, but has a complex and rich set of connotations, as described by Miller (1997). Shame has been described as disgust turned inward, a judgment that the self is disgusting. It is not clear what emotion is 'opposite' to disgust, but perhaps love is the closest opposite.

8. Disgust and the Arts

Disgusting images appear frequently in visual art. Disgust is also a focus of humor, at least in Western culture, and plays a particularly central role in children's humor. Disgust appears to be attractive in situations where it is contained and not too threatening; in this regard, our enjoyment of disgust may be related to our enjoyment of horror movies, sad movies, and amusement park rides. All of these proclivities seem to be uniquely human. In general, disgust plays a major role in socialization, and the sensibilities it promotes are at the heart of what it means to be human and to be civilized.

See also: Emotions, Evolution; Emotion and Expression; Emotion, History of the Concept; Emotion, Neural Basis of; Emotions, Psychological Structure of; Death, Anthropology of; Death and Dying, Psychology of; Death and Dying, Sociology of; Sociology of Emotion; Food in Anthropology; Food Preference; Taste and Smell, Psychology of; Taste and Taste Culture

Bibliography

Angyal A 1941 Disgust and related aversions. *Journal of Abnormal and Social Psychology* **36**: 393–412

Becker E 1973 *The Denial of Death*. Free Press, New York

Darwin C R [1872] 1965 *The Expression of the Emotions in Man and Animals*. University of Chicago Press, Chicago

Douglas M 1966 *Purity and Danger*. Routledge & Kegan Paul, London

Ekman P 1992 An argument for basic emotions. *Cognition and Emotion* **6**: 169–200

Ekman P, Friesen W V 1975 *Unmasking the Face*. Prentice-Hall, Englewood Cliffs, NJ

Freud S [1905] 1953 Three essays on the theory of sexuality. In: Strachey J (ed. and trans.) *The Standard Edition of the*

Complete Psychological Works of Sigmund Freud. Hogarth Press, London, Vol. 7, pp. 123–231
Haidt J, McCauley C R, Rozin P 1994 A scale to measure disgust sensitivity. *Personality and Individual Differences* **16**: 701–13
Izard C E 1971 *The Face of Emotion*. Appleton-Century-Crofts, New York
Miller W I 1997 *The Anatomy of Disgust*. Harvard University Press, Cambridge, MA
Rozin P, Fallon A E 1987 A perspective on disgust. *Psychological Review* **94**: 23–41
Rozin P, Haidt J, McCauley C R 1993 Disgust. In: Lewis M, Haviland J (eds.) *Handbook of Emotions*. Guilford, New York
Rozin P, Lowery L, Imada S, Haidt J 1999 The CAD triad hypothesis: a mapping between three other-directed moral emotions (contempt, anger, disgust) and three moral ethics (community, autonomy, divinity). *Journal of Personality & Social Psychology* **76**: 574–86
Rozin P, Nemeroff C J 1990 The laws of sympathetic magic: a psychological analysis of similarity and contagion. In: Stigler J, Herdt G, Shweder R A (eds.) *Cultural Psychology: Essays on Comparative Human Development*. Cambridge University Press, Cambridge, pp. 205–32
Scherer K R, Wallbott H G 1994 Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology* **66**: 310–28
Shweder R A, Much N C, Mahapatra M, Park L 1997 The 'big three' of morality (autonomy, community, divinity), and the 'big three' explanations of suffering. In: Brandt A, Rozin P (eds.) *Morality and Health*. Routledge, New York, pp. 119–69
Tomkins S S 1963 *Affect, Imagery, Consciousness: The Negative Affects*. Springer, New York, Vol. 2

P. Rozin

Dispute Resolution in Economics

Negotiation is common in many settings. A common feature of all negotiation is an understanding by the parties to the negotiation about the process that will be used to resolve a dispute in the event that the parties fail to reach agreement. Disputes between labor unions and employers can end in a strike, or can be resolved by arbitration. Disagreement over contracts or tort damages can result in civil litigation and a decision by a judge or jury. Disagreement between nations can result in military disputes. The dispute settlement process defines the expectations that the parties have about what will happen in the event that no agreement is reached. And this process, in turn, determines negotiator behavior.

There are two fundamental types of dispute settlement mechanism. The first is the systematic imposition of costs on each side. Strikes and military disputes are examples of this type of mechanism. No determinate end to the dispute is implied, but presumably one or both parties concedes sufficiently to make an agreement possible and end the dispute. The second mechanism is the mandating of a settlement by a third party. Arbitration of labor and commercial disputes, and litigation of civil disputes, are examples of this type of mechanism.

Both types of mechanism (strikes and arbitration) are used in the resolution of labor disputes, and the discussion here focuses on that area. This discussion is necessarily selective and is meant to provide a brief introduction to important concepts and issues. See Hirsch and Addison (1986) for a more extensive survey.

1. *Strikes*

A difficult problem in the economics of labor unions is the modeling of strikes. The difficulty arises because strikes appear to be non-equilibrium phenomena. Both sides would have been better off if they had been able to agree on whatever the outcome was without bearing the costs of a strike. In other words, strikes are Pareto-dominated by the same bargaining outcome without a strike. While research on strikes in the United States has diminished with the level of strike activity in the USA since the mid-1980s, there was an active empirical literature in the 1970s and 1980s.

Ashenfelter and Johnson (1969) developed a non-equilibrium model of strike activity by attributing to the union an arbitrary concession schedule positing an inverse relationship between the minimum acceptable wage and the length of strike. The firm selects the point on this schedule that maximizes profits, the tradeoff being longer strikes involving larger up-front strike losses, but lower labor costs into the future. Ashenfelter and Johnson used this model to motivate an empirical analysis of the aggregate level of strike activity in the United States. Farber (1978) used the same theoretical framework to estimate a structural model of the joint determination of wage changes and strike length using data on the outcomes of a time-series of cross-sections of negotiations for large US corporations.

Tracy (1986) developed an equilibrium model of strikes based on learning in a model of asymmetric information (Fudenberg and Tirole 1981, Sobel and Takahashi 1983). The model is based on the idea that the union does not know the profitability of the firm. The union would like to extract a high wage from high-profit firms and a lower wage from low-profit firms. The key is that high-profit firms have a high cost of bearing strikes relative to low-profit firms. In this case, the union can make a high initial wage demand. High-profit firms will prefer to pay this wage immediately rather than to lose profits during a strike. Low-profit firms will prefer to lose their lower profits during the strike in exchange for a lower wage into the future.

Tracy (1986) used this model to generate the empirical implication that there will be more strikes where there is more uncertainty about the firm's

profitability. He then implemented this model using data on a large number of negotiations and using the volatility of daily stock returns as his measure of uncertainty about profitability. He finds evidence that uncertainty is correlated positively with strike activity. A final point is that the declining concession schedule in the Ashenfelter–Johnson model can be interpreted as a reduced-form expression of the union learning function. As the strike goes on, the union is willing to accept a lower wage as they learn that the firm is lower-profit.

It is quite possible that strikes are not equilibrium phenomena, and that the insistence in the economics literature, both theoretical and empirical, on equilibrium explanations, is inappropriate. Strikes can result from miscalculation. In particular, the parties may be relatively optimistic about the outcome after a strike. For example, the union may expect a wage settlement of \$20/hour after a strike, while the firm expects to pay only \$18/hour after the same strike. If the cost of a strike is \$0.50/hour to each side, then the union will accept no less than \$19.50 to avert the strike, while the firm will be willing to offer no more than \$18.50. The result will be a strike. Of course, the expectations of both sides cannot be accurate, *ex post*. This relative optimism can have a variety of sources. One possibility with a long history in the industrial relations literature, is that the parties selectively overweight arguments that are favorable to their side (Walton and McKersie 1965, Neale and Bazerman 1991). However, relative optimism can be generated even with appropriate weighting of available information. If each side is getting independent (and accurate in expectation) information on the likely uncertain outcome, for some fraction of the time the information will lead to relatively optimistic expectations sufficient to outweigh the costs of a strike. Strikes will occur in these cases.

2. *Arbitration*

In the United States, arbitration is used commonly to determine the terms of labor contracts in the public sector in cases where the parties cannot agree on their own. In conventional arbitration, the arbitrators are free to impose whatever settlements they deem appropriate. A common alternative is final-offer arbitration, where the arbitrator must select one of the final offers submitted by the parties and cannot compromise.

2.1 *Conventional Arbitration*

Consider the following simple model of the negotiation process where conventional arbitration is the dispute settlement mechanism (Farber and Katz 1979). Suppose the parties are negotiating over the split of a pie of size one. The share of the pie going to the union is Y, and the share going to the firm is $(1-Y)$. The arbitrator will impose a settlement of Y_a, and each party has a prior distribution on Y_a. Let Y_a^u and Y_a^f represent the prior expected value to the union and firm, respectively, of the arbitration award. Finally, let C_u and C_f represent the cost to the union and the firm of using arbitration (denominated in pie shares). Assume further that the parties are risk neutral, so that the union will prefer any negotiated share for itself (Y_n) such that $Y_n > Y_a^u - C_u$ and the firm will prefer any negotiated share for itself such that $1-Y_n > 1-Y_a^u - C_f$. These are the reservation values of the parties in negotiation. The result is a *contract zone* of settlements preferred by both sides to arbitration that consists of all shares that fall in the range:

$$Y_a^u - C_u < Y_n < Y_a^f + C_f \tag{1}$$

The size of the contract zone is:

$$CZ = Y_a^f - Y_a^u + C_u + C_f \tag{2}$$

In the simplest model, there will be a negotiated settlement whenever $CZ \geqslant 0$, and there will be an arbitrated outcome whenever $CZ < 0$.

This model provides an important insight into when one might expect a dispute. Relative optimism about the likely arbitration award $(Y_a^u > Y_a^f)$ can be a reason for a dispute. If the relative optimism is sufficiently great to outweigh the total costs of arbitration $(C_u + C_f)$, then there will be no contract zone, and arbitration will be required. Thus, relative optimism is a cause of disputes. Clearly, higher total costs of arbitration imply a larger contract zone. Larger costs make it less likely that any degree of relative optimism will cause there to be no contract zone. Thus, larger total costs of disputing reduce the likelihood of disputes. This latter property is implied by any reasonable model of disputes.

The model also demonstrates how the dispute settlement mechanism determines fundamentally negotiated outcomes. Consider the case where the parties have identical expectations about the arbitration award of $Y_e = Y_a^u = Y_a^f$. The size of the contract zone in this case is determined entirely by the costs of arbitration (Eqn. 2), and there will be no dispute. The negotiated settlement will lie in the range defined by:

$$Y_e - C_u < Y_n < Y_e + C_f \tag{3}$$

In other words, the negotiated settlement will lie in some range around the common expectation of the arbitration award.

One criticism of conventional arbitration as a dispute settlement mechanism in labor disputes is that arbitrators appear to split the difference between the

demands of the parties, resulting in an arbitration award that is close to the average of the demands. However, Eqn. 3 makes it clear that the opposite is likely to be true: The negotiated settlement is likely to lie between the reservation values of the parties. Mnookin and Kornhauser (1979) and Farber (1981) present analyses that make this point.

2.2 Final-offer Arbitration

The perception that arbitrators split the difference in conventional arbitration led to the belief that conventional arbitration 'chilled' bargaining by providing the parties with an incentive to maintain extreme demands. This led to the development of final-offer arbitration, which prevented the arbitrator from compromising (Stevens, 1966). Farber (1980) developed a simple model of final-offer arbitration where arbitrators select the offer closest to what they would have imposed in conventional arbitration (Y_a). In this model, the probability that the union's offer is selected is:

$$P_u = Pr\left(Y_a > \frac{Y_u + Y_f}{2}\right) = 1 - F\left(\frac{Y_u + Y_f}{2}\right) \tag{4}$$

where Y_u and Y_f represent the final offers of the union and the firm respectively and $F(\cdot)$ represents the cumulative distribution function of the parties' common prior on Y_a.

Assuming risk-neutrality, the expected payoff to the union from final-offer arbitration is:

$$V_u = P_u Y_u + (1 - P_u) Y_f \tag{5}$$

and the expected payoff to the firm is:

$$V_f = 1 - [P_u Y_u + (1 - P_u) Y_f] \tag{6}$$

The Nash equilibrium final offers of the parties are the offers such that (a) Y_u maximizes V_u conditional on Y_f, and (b) Y_f maximizes V_f conditional on Y_u. These offers have the property that neither party will desire to change its offer.

Since the parties are uncertain about Y_a, the Nash equilibrium final offers reflect a tradeoff between the value to the party of its offer if its offer is selected, and the probability that its offer is selected. A more extreme offer yields a higher payoff to the party if its offer is selected by the arbitrator, but it lowers the probability that the offer will be selected.

A key implication of this model is that final offer arbitration may not encourage more convergence in negotiation than conventional arbitration because of the desire to present optimal final offers to the arbitrator that may be extreme.

2.3 Empirical Evidence

Ashenfelter and Bloom (1984) present evidence from arbitration of contract disputes for public sector employees in New Jersey, where both conventional and final-offer arbitration are used, consistent with the general model of arbitrator behavior outlined here. They find that arbitrators use a common underlying notion of Y_a to make awards in both conventional and final-offer arbitration. Their evidence is consistent with the arbitrators' decision in final-offer arbitration being to select the offer closest to Y_a. Bazerman and Farber (1986) reach similar conclusions based on evidence from arbitration awards, both conventional and final-offer, made by professional arbitrators in hypothetical cases.

Convincing evidence is scarce on whether final-offer arbitration leads to more settlements and reduces reliance on arbitration. Ashenfelter et al. (1992) present experimental evidence that does not show higher rates of agreement in final-offer arbitration than in conventional arbitration.

Bibliography

Ashenfelter O, Bloom D E 1984 Models of arbitrator behavior. *American Economic Review* **74**: 111–24

Ashenfelter O, Currie J, Farber H S, Spiegel M 1992 An experimental comparison of dispute rates in alternative arbitration systems. *Econometrica* **60**: 1407–33

Ashenfelter O, Johnson G 1969 Trade unions, bargaining theory, and industrial strike activity. *American Economic Review* **59**: 35–94

Bazerman M H, Farber H S 1986 The general basis of arbitrator behavior: An empirical analysis of conventional and final-offer arbitration. *Econometrica* **54**: 1503–28

Farber H S 1978 Bargaining theory, wage outcomes, and the occurrence of strikes: An econometric analysis. *American Economic Review* **68**: 262–71

Farber H S 1980 An analysis of final-offer arbitration. *Journal of Conflict Resolution* **24**: 683–705

Faber H S 1981 Splitting-the-difference in interest arbitration. *Industrial and Labor Relations Review* **34**: 70–7

Farber H S, Katz H C 1979 Interest arbitration, outcomes, and the incentive to bargain. *Industrial and Labor Relations Review* **32**: 55–63

Fudenberg D, Tirole J 1981 Sequential bargaining with incomplete information. *Review of Economic Studies* **50**: 221–48

Hirsch B T, Addison J T 1986 *The Economic Analysis of Unions: New Approaches and Evidence*. Allen & Unwin, Boston, MA

Mnookin R, Kornhauser L 1979 Bargaining in the shadow of the law—the case of divorce. *Yale Law Journal* **88**: 950–97

Neale M A, Bazerman M H 1991 *Cognition and Rationality in Negotiation*. The Free Press, New York

Sobel J, Takahashi I 1983 A multi-stage model of bargaining. *Review of Economic Studies* **50**: 411–26

Stevens C M 1966 Is compulsory arbitration compatible with bargaining? *Industrial Relations* **5**: 38–52

Tracy J S 1986 An investigation into the determinants of US strike activity. *American Economic Review* **76**: 423–36

Walton R E, McKersie R B 1965 *A Behavioral Theory of Labor Negotiations: An Analysis of a Social Interaction System.* McGraw-Hill, New York

H. S. Farber

Disputes, Social Construction and Transformation of

The 'social construction' of disputes refers to the process by which people perceive a wrong or an injustice, articulate a claim against the offending party, and assert that claim publicly before a third party. Disputes are not concrete entities or objective events, but are instead social constructs, the results of subjective feelings and perceptions, and expressed through language and actions. 'Transformation of disputes' refers to the change in form or content of disputes as other participants become involved in the disputing process and add their own perceptions and interests to those of the original parties.

1. *Law and Disputing*

Empirical research on dispute transformation developed in the late 1970s and early 1980s as a result of trends in several social science fields and in law. Anthropologists showed how a focus on 'trouble cases' could provide a useful framework for comparative study of law in different societies. Research reported in Nader and Todd (1978), for example, explored the factors behind parties' choices of different methods for pursuing, or attempting to resolve, disputes in diverse societies. Relations between disputants, the location of the dispute, and the nature of the problem, along with the range of disputing alternatives, could all explain the construction and handling of disputes. The interest of anthropologists in disputant choice and the availability of different alternatives for dispute resolution overlapped with the interest of legal scholars in access to justice.

Expansion of legal services for litigants in civil cases and concern for the role of civil courts in society led analysts to examine the legal needs of citizens. What kinds of cases and what groups of citizens are not being served by the civil justice system? How can access to justice be improved? Questions such as these spurred efforts to describe and count 'unmet legal needs.' It is easy to count cases in court, but counting cases that have not been filed proves more problematic. The problems center on what kinds of cases should be in court, are there too many court cases, are they the wrong ones, and what populations do the courts serve? Through such questions, the very nature of what constitutes a civil court case is called into debate. Comparative studies of legal caseloads in different societies also fueled an interest in improving understanding of the relation between individual and social conflicts and court cases. Such concerns led scholars to focus explicitly on the social construction of disputes and their emergence in court.

Recognition that courts provide only one way to resolve conflicts, and a costly way at that, led some observers to propose alternative processes for dispute resolution. Interest in alternatives to court also led to research on dispute transformation as scholars. Interest in alternatives to court also led to research on dispute transformation as scholars explored the impact of different institutional mechanisms on disputes. Alternative dispute resolution (ADR) provides new processes such as mediation and arbitration as alternatives to adjudication in the courtroom. Those who favor ADR praise its informality, lower cost, speed, and case outcomes particularized to the needs of the disputants. ADR opponents worry about a two-tier legal system with courts reserved for wealther litigants and alternatives for others, and about the lack of legal norms guiding the resolution of disputes outside of court. Scholars on both sides of the debate have sought to bolster their positions by gathering information on how disputes are transformed through negotiation, mediation, arbitration, or adjudication.

Other interest in dispute transformation comes from sociologists studying case processing in the criminal justice system, from psychologists examining conflict resolution, and from political scientists concerned with the role of courts in resolving political issues. Criminal justice officials negotiate the meaning of events and identities of suspects as in the process of plea bargaining. Crimes are defined according to their normal situational and social features, as well as by legal codes. Sociological research shows that criminal defendants, for example, are sometimes offered a plea bargain to a lesser offense in the process of narrowing their cases into conventional categories of crime. Psychological study of negotiation in conflict resolution involves attention not only to how players try to maximize their interests, but also how they seek actively to define the nature of the issues and the scope of their conflict.

Finally, political scientists study dispute transformation as a means of understanding legal mobilization and the role of courts in society. Test case litigation involves the construction of individual lawsuits as vehicles for articulating and advocating collective political interests. In successful litigation, disputants may become important creators of law, and their lawyers are spokespersons for a cause. Court trials may be confined to the individual facts and issues in dispute, or broadened to address the normative implications and legal meaning of the dispute. As a consequence of cause lawyering, interest group activity, or media attention, political trials may result from such broadening of issues in dispute. Parties in

litigation, their lawyers, and court officials may each seek to transform disputes in different ways, for different political ends.

For detailed treatment of these topics and for further references see *Law, Mobilization of*; *Public Interest Law*; *Class Actions: Legal*; *Parties: Litigants and Claimants*; *Legal Culture and Legal Consciousness*; *Mediation, Arbitration, and Alternative Dispute Resolution (ADR)*; *Law and Everyday Life*; *Law as Constitutive*.

2. *The Emergence and Transformation of Disputes*

A hypothetical sexual harassment case can illustrate some of the central issues in the emergence and transformation of disputes. The sections that follow this example elaborate on some of the key contributions of dispute transformation research: the perception of a grievance; the disputing pyramid; the lawyer's role in transformation; the role of language; how courts and institutions shape dispute transformation; and the relation between dispute transformation, politics, and law.

2.1 *Sexual Harassment as an Example*

Consider the situation in which a manager makes sexual overtures toward his secretary. He frequently tells crude jokes, embarrasses her with comments about her looks and dress, and invites her for after-hours dates. She is not attracted to him, but she likes her job. What would be her response? She might simply brush him off and try to live with his unpleasant behavior ('lump it'); she might seek advice from co-workers, neighbors, friends; she might leave her job and seek work elsewhere ('exit'); she might pursue an internal complaint within her place of work, or seek legal counsel outside the job and file a claim of sexual harassment. This suggests how the secretary's dispute with her manager is not a discrete event, but is instead socially constructed by her perceptions, by others, and by her disputing alternatives. The conflict may or may not lead to a legal case.

The emergence of a dispute depends first on recognizing and identifying a problem. For many women employees in the United States until the 1990s—and still for many worldwide—the occasional lecherous boss was part of a normal working environment. Cultural and political changes led to women's recognition of their grievances and the language of federal law introduced a new name for their grievance: 'sexual harassment.' Supporters of an employee—such as co-workers, a women's group, or employees, union—could broaden the scope of the conflict to transform her individual dispute with the offending manager into a larger protest over sexist working conditions. Additional issues could be added to the dispute, such as discrepancies in pay and working conditions between men and women employees. Advice from a civil rights lawyer could lead to a class action lawsuit, which would further broaden the scope of the conflict and add new parties to the case. Alternatively, internal processes of dispute resolution within the organization could depoliticize the dispute, remove any aspect of sex discrimination or sexual harassment, and transform it into a personality conflict between the two parties. In short, both the meaning and the course of the dispute are contingent. They depend on the perceptions, actions, and interests of various parties, and on the alternatives provided by language and institutions. Furthermore, the way in which the dispute is defined—and whether it is even seen as a dispute or not—reinforces certain power relations and ideas, and links law to everyday life. (See *Sexual Harassment: Legal Perspectives*.)

2.2 *The Disputing Pyramid*

Why are some grievances ignored but others become disputes? And why do some disputes in American society end up as court cases but the vast majority do not? There is a 'disputing pyramid,' or 'legal iceberg,' with vast numbers of grievances at the bottom, fewer disputes in the middle, still fewer legal claims toward the top, and only small number of actual cases in court (Galanter 1974, Miller and Sarat 1981). The emergence of disputes depends first on perceiving an injury or problem ('naming'), then attributing the problem to the fault of another ('blaming'), and finally confronting the opposing party and seeking some kind of remedy ('claiming') (Felstiner et al. 1981). Many injuries are never perceived as such, and still others are not acknowledged because of cultural ideas, institutional barriers, or the lack of social support. Not all grievances become claims and not all claims become lawsuits in court. Movement of grievances to the top of the disputing pyramid depends in part on structual characteristics of the dispute, such as the relationship between the parties, the type of claim, and the relation parties have to the legal system. The disputing pyramid is shaped differently for various kinds of cases but demographic characteristics of the parties (e.g., class, education) may not be as significant for disputing as is commonly thought (Miller and Sarat 1981). Institutionalized procedures and remedies for handling conflicts may encourage claiming by lowering costs and making the disputing process more accessible and predictable.

In addition to social and institutional factors affecting dispute transformation, psychological factors play a role. Individuals vary in their claims consciousness and in their propensity to assert claims or pursue legal

rights. This variation is associated with preferences for different modes of conflict resolution and with self-described personality attributes, but appears to be independent of demographic status (Vidmar and Schuller 1987).

Dispute transformation helps to explain not only where disputes end up in the disputing pyramid but also how changes occur in the very nature of disputes. Both the object of the dispute and the normative framework for handling it are matters to be negotiated (Mather and Yngvesson 1981). To succeed in establishing what the dispute is about can go a long way toward resolving it. Changing the normative framework for viewing events or relationships can transform disputes over them. Viewing interpersonal conflict as a cognitive bargaining process has led psychologists to pursue research on conflict frames, that is, on disputants' cognitive interpretations of conflict. Cognitive framing influences both the process and outcomes of disputes, as well as disputant satisfaction (Pinkley and Northcraft 1994).

2.3 The Lawyer's Role in Dispute Transformation

Lawyers play a major role in constructing and transforming disputes. Exactly what the nature of that role is, however, is a matter of considerable controversy. In one view, lawyers escalate conflict, urging their clients to pursue adversarial formal legal solutions to problems that might be more appropriately resolved informally. Thus, lawyers are said to encourage clients to take their claims higher on the disputing pyramid in order to vindicate legal rights and obtain greater compensation. Of course, such vigorous legal advocacy could also enrich the lawyer, through enhanced fees or a percentage of a larger settlement. Although this view has wide currency in the popular press, it has mixed support in empirical research on lawyers and clients.

Studies of lawyer/client interaction, particularly in the areas of divorce, personal injury, and criminal law, suggest that lawyers more frequently dampen conflict than escalate it. Financial incentives discourage lawyers from pursuing extensive formal legal action and encourage quick, informal settlements of cases. Where clients have the resources to support it, it is they—not their lawyers—who may insist on extended legal combat. Thus, O. J. Simpson spared no expense for his defense lawyer, but the vast majority of criminal defendants lack such resources and plead guilty.

In divorce, some of the most bitter and highly contested trials occur between very wealthy spouses who can underwrite the fees of their attorneys. But in most divorces in the United States only one spouse can afford a lawyer and the other is *pro se* or else neither party is represented by counsel. Even in two-lawyer divorces, attorneys typically seek to cool out their clients, to protect the children, and to preserve the long-term relationship between ex-spouses, because cooperative settlement is seen to be cost-effective, or because the law encourages equitable division of property and shared parenting.

An important question about lawyers' interactions with clients is the extent to which clients' stories and objectives are heard by counsel. One thing that lawyers clearly do is translate clients' problems into legal language. Lawyers rephrase auto accidents into torts, domestic violence into assaults, and racist actions into civil rights violations. In lawyers' offices, clients may want to pour out stories in their full social and personal context. But details that clients see as highly relevant are dismissed by counsel as not germane to the legal case. Lawyers are trained to focus on particular issues, leading to frustration in clients (who are not being 'heard') and possibly to overlooking key details. Variation in the lawyers' role in dispute transformation by different areas of law and by different types of lawyers and clients remains a crucial topic for further study.

2.4 Language and Dispute Transformation

As social constructs, disputes depend upon the language used to describe them. Yet language is not neutral. It orders 'facts' and invokes 'norms' in ways that reflect the interests of the parties, or that anticipate the reactions of others (Mather and Yngvesson 1981). By including or excluding certain facts, or by selecting particular words, parties are situating their conflict in a certain normative framework. For example, in a case of sexual harassment, the word '*harassment*' itself invokes a different normative meaning than, say, '*flirting*' or '*being friendly*.' Struggles over language involve negotiation over which framework to apply, and such negotiations are integral to the disputing process. How a dispute is defined has consequences for the identity and relations between the parties, and for whether and how a dispute is resolved. Working-class Americans seeking to obtain legal redress in court often have their disputes redirected away from a legal framework and into a therapeutic framework of mental health or social welfare concerns (Merry 1990).

An analytic framework that examines language and disputing not only applies in modern legal systems, but also to smaller societies and even ancient ones. For instance, research on litigation in ancient Athens has produced powerful insights about Athenian law and democracy. The rhetoric and procedures of litigation there were used to transform broad conflicts into two-party disputes, impose specialized roles and identities, create new political resources, and foreclose other options—thus effecting the reproduction of Athenian democracy (Johnstone 1999).

Anthropological studies in different cultural settings also reveal how particular linguistic definitions of

conflict affect individual disputes and reinforce patterns of social order. Narratives in a Fiji Indian community, for example, aid in the management of social conflict by constructing identities, transforming personal experiences, and constituting knowledge. By 'telling troubles' to various different audiences, the Indian villagers construct their social relationships, privileging some accounts over others (Brenneis 1996).

2.5 How Courts and Other Institutions Shape Dispute Transformation

Institutional structures influence disputing by channeling the ways in which conflicts can be transformed. Courts, for example, often have strict rules of evidence that limit what kinds of information may be included in the construction of a case during trial. But informal negotiations outside of court, such as plea bargaining or settlement talks, have few such limits. When informal plea negotiations over disposition of a first-time drug defendant include discussions of the defendant's social background, family situation, and employment, lawyers may lean toward lenient treatment. But none of this material might be admitted in adjudication and the resulting sentence could be much harsher after conviction at trial. Different institutional procedures transform cases in different ways, producing variation in case outcomes.

In theory, a distinguishing feature of mediation—in contrast to adjudication—is the ability of mediators to range widely over different issues in a conflict. Whereas judges are supposed to narrow their focus to particular legal elements of a case, mediators (who are not bound by law) may encourage the emergence of new issues in their effort to formulate an agreement that is acceptable to both parties. Divorce mediators sometimes comment on the 'strange-looking agreements' that result from such wide-open mediation sessions. Some observers believe that the individualized agreements produced by mediation are more effective in resolving specific issues in dispute and also lead to increased compliance with the outcome. But critics of mediation worry that the agreements come at a high price—that is, through possibly overpowering the weaker party or through the lack of legal norms shaping the final agreement.

Mediators may have more in common with judges than with other mediators in terms of how they actually handle cases, and negotiation is frequently intertwined with adjudication. Instead of assuming differences across processes, a more productive focus centers instead on how processes influence dispute transformation—for example, through specialized or general language, through closed or open dispute hearings, or through linkages between dispute processes and other organizations in society (Ynvgesson and Mather 1983).

One response of large corporations to increased legal regulation of employment in the United States has been to create internal processes of dispute resolution. As a condition of employment, workers may even be prohibited from using courts for their disputes in order to favor internal organizational complaint procedures and avoid external legal liability for the employer. The result of such policies is the transformation of civil rights in the workplace from an issue of public law to an issue of private employee relations (Edelman et al. 1993). The object of dispute shifts from the working conditions provided by the employer to the personal relations among employees. Institutional mechanisms for resolving problems influence the ability of disputants to negotiate over issues and set limits on how legal norms penetrate everyday disputes. Conflict in organizations is increasingly handled by procedures that more closely resemble the imposition of authority than third party facilitation of dispute resolution (Kolb and Putnam 1992).

2.6 Dispute Transformation, Public Order and Law

A focus on the normative framework and definitions of a dispute reveals the political aspects of disputing and links the routine disputing processes to broader questions of power, order, and law. Dispute transformation thus bridges consideration of individual agency and social structure. When a secretary pursues a harassment claim against her supervisor, she is pursuing an individual dispute while simultaneously her case is reproducing the social structure of gender relations in the workplace. By providing the language of law and certain institutions for handling disputes, the social structure makes resources available to challenge and change that very structure. Those resources, however, are then countered by others (e.g., internal complaint mechanisms) that may sustain existing power relationships.

Most disputes are narrowed or rephrased in established categories of discourse. But, on occasion, new meanings are given to disputes. When a dispute is *expanded*—rephrased in terms of a new normative framework—it challenges the prevailing order. 'Expansion of individual disputes is one way that social change is linked to legal change' (Mather and Yngvesson 1981, p. 779). Agents of dispute expansion include the parties themselves, lawyers (especially cause lawyers), interest groups or supporters, third parties, or the audience to the dispute. Judges, or other third parties, may expand disputes in order to further their own political interests. In appellate court decisions, judges sometimes introduce new legal rules or ways of interpreting conflicts. For example, a dispute over obscenity in the lower courts became the famous *Mapp v. Ohio* case of illegal police behavior in the appellate court.

Litigation against tobacco manufacturers illustrates both dispute narrowing and dispute expansion. When ill cigarette smokers initially filed claims against the cigarette industry, their cases were *narrowed* into

established normative frameworks that insulated the industry from attack. Smokers were said to have 'assumed the risks' of tobacco harm, and were thereby responsible for their own illnesses. With the release of damaging internal evidence from the tobacco industry, and with the introduction of new normative frameworks (of fraud, consumer protection, or Medicaid reimbursement), there was some change in cases against tobacco manufacturers. Through dispute *expansion* anti-tobacco lawyers succeeded in transferring responsibility for smoking-related injuries back to tobacco manufacturers. With sizable jury verdicts the normative claims of anti-tobacco advocates assumed greater strength (Mather 1998). Indeed, through litigation and other forms of disputing, the meaning and scope of law are actively constructed in the process of defining individual disputes.

3. *Problems and Prospects for Studying Dispute Transformation*

The dispute transformation perspective can be faulted for its overemphasis on process to the exclusion of dispute outcome, a failure to examine the political significance of disputing, its individualistic focus, the methodological difficulty of identifying or measuring transformations, and its insufficient concern for institutions. By addressing these issues, however, researchers have redefined their work to include data that allows a more complete picture of disputing and law. The concept of dispute transformation is inherently interdisciplinary. Insights and methodological advances, from for example, anthropological linguistics, organizational studies, social network theory, psychology of negotiation, legal mobilization, and judicial decision making have broadened the application of the transformation perspective.

Viewing how disputes emerge and are transformed contributes to understanding both the micropolitics of dispute resolution and the macropolitics of disputing in society. The transformation perspective is useful crossculturally, and across different historical periods, and helps to explain how law is created and maintained.

See also: Conflict: Anthropological Aspects; Conflict: Organizational; Conflict Sociology; Dispute Resolution in Economics; Justice and its Many Faces: Cultural Concerns

Bibliography

Brenneis D 1996 Telling troubles: Narrative, conflict, and experience. In: Briggs C L (ed.) *Disorderly Discourse*. Oxford University Press, New York, pp. 41–52

Edelman L B, Erlanger H S, Lande J 1993 Internal dispute resolution: The transformation of civil rights in the workplace. *Low and Society Review* **27**: 497–534

Felstiner W L F, Abel R L, Sarat A 1981 The emergence and transformation of disputes: Naming, blaming, claiming. *Law and Society Review* **15**: 631–54

Galanter M 1974 Why the "haves" comes out ahead: Speculations on the limits of legal change. *Law and Society Review* **9**: 95–160

Johnstone S 1999 *Disputes and Democracy: The Consequences of Litigation in Ancient Athens*. University of Texas Press, Austin, TX

Kolb D M, Putnam L L 1992 The multiple faces of conflict in organizations. *Journal of Organizational Behavior* **13**: 311–24

Mather L 1998 Theorizing trial courts: Lawyers, policymaking, and tobacco litigation. *Law and Social Inquiry* **23**: 897–940

Mather L, Yngvesson B 1981 Language, audience, and the transformation of disputes. *Law and Society Review* **15**: 775–821

Merry S E 1990 *Getting Justice and Getting Even: Legal Consciousness among Working-Class Americans*. University of Chicago Press, Chicago, IL

Miller R E, Sarat A 1981 Grievances, claims, and disputes: Assessing the adversary culture. *Law and Society Review* **15**: 525–65

Nader L, Todd H F 1978 (eds.) *The Disputing Process: Law in Ten Societies*. Columbia University Press, New York

Pinkley R L, Northcraft G B 1994 Conflict frames of reference: Implications for dispute processes and outcomes. *Academy of Management Journal* **37**: 193–205

Vidmar N, Schuller R A 1987 Individual differences and the pursuit of rights. *Law and Human Behavior* **11**: 299–317

Yngvesson B, Mather L 1983 Courts, moots, and the disputing process. In: Boyum K, Mather L (eds.) *Empirical Theories About Courts*. Longman, New York, pp. 51–83

L. Mather

Dissolution of Family in Western Nations: Cultural Concerns

1. *The Question of Family Centrality*

Apprehensions about the demise of the family reach back to biblical times (Stern 1938). Social historians have amply demonstrated that eras of family change, which are endemic in Western history, are invariably accompanied by fears of family breakdown. The Industrial Revolution introduced a widespread anxiety that the family was losing its centrality—kinship was becoming less salient, elders were losing authority, and functions were being replaced by other institutions in society—although historians continue to debate whether any of these claims are valid (Goldthorpe 1987). The challenge of establishing the breakdown of the family system is complicated.

Comparisons to any period in the past are inevitably tinged with a longing for what the eminent sociologist of the family, William J. Goode, referred to as 'the classical family of Western nostalgia' (Goode 1963). It is a difficult task to sort out ideology from experience when it comes to charting trends in family dissolution.

If dissolution is described as the weakening of kinship systems, marriage, and the significance of family ties because the claims of other institutions have taken on some of the tasks previously carried on within the household, then it can hardly be doubted that the family as a social institution has experience a decline. However, following this definition, we must conclude that the strength of the family has been waning for at least several centuries, along with the church and the community. The creation of a job-based economy, the growth of nation-states, and, more recently, the development of a global economy have eroded locally based institutions. Families are less likely to be fixed in a geographical place, to control job opportunities, and hence to be able to sponsor the next generation; accordingly, they are far less inclined to support the claims of extended family members. Moreover, state welfare systems have lessened the burden of caring for the elderly and distant relatives who might have called upon family assistance in times past. (Elaborations of this general argument can be found in Burch and Matthews (1987) and Thornton and Freedman (1983).)

At the same time, in other arenas the family has taken on added importance. For example, the family has grown as a center of emotional development and perhaps of personal identity as well (see, for example, Gillis 1996). There is little evidence that might suggest that entering parenthood is less important now than it was a century or two ago. It is difficult to establish that parents feel less responsibility for their children today or take on parental duties less deliberately. Indeed, it could be argued that intensity of kinship obligations to children replaced more extensive kinship ties during the twentieth century. Parents assume a greater cost of bearing children as they provide little or no economic return in the form of household labor or care in old age (Caldwell 1982). The cost of education and training has risen and children since the middle of the twentieth century have spent a longer period as economic dependents (Watkins et al. 1987). Thus, kinship has become more concentrated within the nuclear family and close extended and collateral kin.

The rise of the so-called 'isolated nuclear family,' a system based on a strong and durable conjugal tie that is structurally independent of and equally linked to maternal and paternal kin lines, became the ideal family form in all Western nations during the latter part of the nineteenth and first half of the twentieth centuries. Family scholars debated whether this form would spread to developing countries in other regions of the world along with the free market economy. Goode (1963) argued that the functional 'fit' between an advanced industrial economy and the nuclear form would promote the transformation of family systems to Western-style families worldwide.

Clearly, the convergence has not occurred to the extent that functionalists predicted it would, although it may still be too soon to rule out the possibility that complex and varied kinship systems in Asia and Africa will wane as the economy becomes globalized (Thornton and Fricke 1987). However, provisionally, it appears that these Western-centric predictions may have underestimated the capacity of complex kinship systems to adapt to economic change and vice versa.

Moreover, the functional 'fit' of the nuclear family to the economy did not account for the revolution in gender roles that occurred in the second half of the twentieth century, or the strain placed on the conjugal relationship by what Rose Coser once labeled the 'greedy institution of work.' Paradoxically, the 'fit' between the nuclear family and a postindustrial economy worked far less well than was envisioned by social theorists of family change. Thus, the family place as 'a haven in the heartless world'—a description that was applied to the early, postindustrial family of the late nineteenth century—has continued to be an apt description of its current social definition. Indeed, a mountain of evidence from public opinion surveys shows that the family remains the central institution from which individuals derive meaning and authenticity. Individuals continue to form families although the definition of what constitutes a family has broadened beyond the nuclear unit. In large measure, this recent trend toward diversification has come about because the institution of marriage has been redefined.

2. *Reworking The Marriage Contract*

In the last third of the twentieth century, divorce and nonmarital childbearing rose throughout the West, with the exception of a few countries in southern Europe which remained more dependent on local agriculture, such as Italy and Greece. Generally, when sociologists and social critics argue that the family is declining, they mean that the notion of conjugal permanence has diminished and that marriage expectations have changed. Indisputably, the marriage contract has been radically altered since the 1950s (Bumpass 1990). In virtually all Western societies, individuals have been granted more autonomy in choosing marriage partners; parents' ability to influence marriage decisions has all but disappeared. Over the past several decades, partners enter marriage expecting relative equality in decision making. The authority of males has declined as a cultural expectation even when it is still maintained in practice by marriage partners. This transformation was brought about in part because of the erosion in the gender-based division of labor within the family. This shift, in turn, can be traced to the growth of a high-skills labor

market in which women are, more or less, on an equal footing with men. The growth of gender equality has also been linked to education, political activism by feminists, and legislative changes resulting in growing economic and political equity among men and women. These changes have altered the cultural context in which marriage takes place. Marriage ceremonies are less likely to include the pledge of obedience—a symbolic shift that suggests that equality among marriage partners has become the ideal.

3. *Changing Marriage Practices*

The change in the marriage contract has generated considerable strains in the matrimonial institution: both men and women have been forced to adhere to a new set of rules governing everyday household practices. Women are now expected to work outside the home and men to assume a greater share of the domestic chores. This renegotiation of gender roles has been accepted more quickly as an ideal among the better educated. Whether it has been put into practice among more educated couples has been questioned by some researchers (Hochschild 1989), but the preponderance of the evidence suggests that this change in gender roles has been dramatic in the past several decades. At least in childcare, men are assuming more responsibility than they did in previous generations, and the normative support for paternal involvement has grown at a phenomenal pace. Still, it is unusual in most countries to find stay-at-home fathers, with the exception of the Scandinavian countries, Norway and Sweden most notably, where public policies have promoted paternal leave, partly to signal men's equal responsibility to childcare.

No doubt the decline of the gender-based division of labor is partially responsible for the remarkable increase in marital dissolution during the latter third of the twentieth century. It may also be linked to the growing practice of cohabitation and the rise of nonmarital childbearing. Marriage is widely viewed as a more problematic and less permanent arrangement than at any time in the recent past, perhaps especially by women. This wariness about wedlock is both a consequence and a cause of higher levels of marital instability. Couples are inclined to enter marriage more cautiously and deliberately than they did in the 1950s. For example, no longer is premarital pregnancy a cause for marriage as it was in the mid-twentieth century. As couples have become more discerning and planful about entering marriage, they have relied on cohabitation as a way of hedging their bets. Some researchers have argued that cohabitation establishes a pattern of leaving relationships that are not good enough.

What constitutes a 'good enough' marriage has become more exacting as well. Whereas in earlier times couples remained wed even if they were discontent with the relationship, today they are more inclined to leave the relationship unless they are content. A rising standard of marital expectations can be inflationary, creating ever higher standards for couples.

This search for greater contentment also plays out in the sexual dimension of marriage. The more open attitude about sexuality and the greater level of premarital experience places greater demands on achieving sexual satisfaction within marriage. Whereas in earlier times couples might have put up with minimal gratification, it is less likely that they are willing to do so today. Increasingly, then, we have seen a shift from a companionate marriage to an intimate marriage as the standard of success.

Another and perhaps more important source of marital instability derives from women's ability to live independently, albeit generally at an economic lower standard. Economic independence permits women to leave relationships that are deemed to be unsatisfactory and men to justify dissolving unhappy partnerships more easily. The turnover of relationships means that marriage is exposed to norms of the market place. The conditional marriage contract itself contributes to the inflationary expectations for marital contentment, driving individuals to be ever more wary and to keep an eye on their options. In some sense, this process resembles the inflation of political expectations that Tocqueville describes in explaining why revolutions breed discontent.

4. *Marital Dissolution and Children's Welfare*

How these changes in the institution of marriage have affected children's welfare is a matter of considerable discussion and a certain amount of debate among researchers (Cherlin 1981). Yet the range of scholarly disagreement seems to be narrowing. Almost everyone agrees that divorce and nonmarital childbearing leave children with fewer resources owing to the facts that women who earn less than men are typically awarded primary custody and that marriage generally provides more than a single income. Only in Sweden do children from single-parent families retain the same standard of living as children in two-parent families through welfare benefits; in all other countries, the income gap is at least modest and more often extreme as is the case in the Anglo-speaking countries which provide less generous support for single parents. In general, parental marital status is strongly associated with family income and appears to account for a substantial portion of the 'divorce effect' on children (Garfinkel and McLanahan 1986).

It is also widely accepted now that some portion of the difference in the well-being of children in two-parent and lone-parent families results from social selection. Couples who divorce have less human and

social capital and are probably less resourceful, skilled, and committed to maintaining a stable union. In other words, children in such families would presumably not do as well even if their parents maintained a stable union while they were being raised owing to a host of conditions related to the probability of marital discontentment and divorce as well as perhaps parenting practices and the ability of parents to manage their children's success.

Still, most scholars agree that marital instability, or family instability of any type, is associated with developmental disadvantages over and above the costs of growing up with fewer resources or social selection. Children profit from stable, secure family situations where they receive attention, support, monitoring, and sponsorship from parent figures, and two such individuals (in a working collaboration) are surely better than one. A mounting number of studies show that, on average, children do better both in the short-term and as adults when they are raised continuously by their biological parents. Remarriage does not appear to confer similar benefits to children apart from the obvious economic benefits, though much of the research on remarriage fails to distinguish the effects of divorce from the impact of remarriage.

It is still an unsettled question whether parents who provide a stable union to children need to be the biological progenitors and whether formal marriage offers any benefits over informal, contractual unions. These issues are especially thorny ones to examine, for they introduce difficult issues of social selection. Provisional evidence suggests that it is the quality of the partnership and the ability of parents to collaborate successfully that seem to matter more than the gender of parents or whether the union was sanctified by the state or church.

5. *Societal Consequences of Marital Dissolution*

The continuing high levels of nonmarital childbearing and union instability in all Western countries has placed added demands on extended kin, especially in segments of the population where remarriage rates are low. Grandparents, especially, are called upon to lend assistance to divorced or unmarried parents and their children. And even prior to family formation, parents may incur greater obligations as their adult children forestall marriage. In many countries, the postponement of marriage has meant that adult children remain in the home well into their twenties and return to reside with their parents in the event of divorce or nonmarital childbearing.

The pressure on parents at mid-life to maintain responsibilities to their children and grandchildren has certainly increased over several decades. Similarly, adults at mid-life frequently face growing obligations to care for their elderly parents who are surviving in greater numbers. In these respects, the extended family is alive and well. The flow of assistance to children is increasingly downward through the generations, at least until parents are no longer able to provide care for their offspring and support for their grandchildren. In contrast to the past, parents of mature children appear to provide more assistance to their offspring than is given to them.

The ways in which divorce and remarriage have affected the provision of support among close kin has received growing attention by scholars. Evidence is accumulating that divorce seriously disrupts the exchange of support from fathers to their children and from the extended patrilineal kin to their former in-laws and their offspring. Whether the reverse is true when fathers retain custody has not been documented. To a remarkable degree, fathers' contact with their children tends to slacken off after marital dissolution. It is possible that this pattern of attrition in paternal contact and support may be less marked today than it was in the 1970s and 1980s. Men, it seems, shift allegiance to new households, and custodial mothers may actively discourage their former mate's involvement with the children when they acquire new partners. Despite the strong legal support for the maintenance of biological ties, the practical implications of creating harmonious relationships across households of formerly and currently married couples is formidable. Consequently, the cumulative impact of divorce (and nonmarital childbearing, where fathers are even less likely to be an active presence in the child's life) has been to give the kinship system a matrilineal tilt.

Remarriage partially restores the bilateral pattern of the family system by reintroducing paternal kin into the family. However, stepparents do not generally create the same strong kinship ties as biological parents, owing perhaps to the fact that kinship ties are often cemented during childhood. Stepkin may also divide their allegiance to former and current marriage partners and their offspring. Nonetheless, the creation of 'remarriage chains' that link former and current kin is a growing source of connection in Western family systems with high rates of conjugal succession. There is some evidence suggesting that the network of in-laws may expand with remarriage as some relatives from the first (or second) union are retained along with the acquisition of kin from the second (or third).

Possibly, too, the pattern of growing involvement of formerly married men in the lives of their children may also help to maintain kinship ties between children and their paternal kin. The attrition of family bonds between men and their children appears to have lessened in recent years as men have become more involved in childcare both after marriage and divorce. The growing pattern of joint legal, if not physical, custody may encourage continued contact between fathers and their children. If this happens, it is likely that the kin of divorced fathers will also play a more active role in children's lives.

The growing complexity and diversity of kinship networks are likely to be a feature of family life in the twenty-first century. It seems unlikely that life-long monogamy will be a prominent feature of the family as it was, at least for a time, in the previous century. If this prediction proves to be right, then public policy is likely to focus on how to deal with the problems of family life created by marital fragility.

6. *Public Policy Considerations*

Current public policy in the USA and Europe has weighed heavily in favor of promoting paternal involvement, especially in the form of material support. A series of legislative acts in the USA, beginning with the Child Support Act of 1988, have been designed to increase the level of child support payments. This legislation has had modest effects in increasing payments, but has also provided normative pressure on men to retain their paternal responsibilities. Fathers Rights groups have become advocates for more even-handed custody policies, to the dismay of some feminists who believe that children will suffer from an increase in conflict created between custodial and noncustodial parents when joint custody is imposed. It is still too early to draw firm conclusions about the consequences of the new policies promoting greater paternal involvement. However, judging from recent information collected by the Census Bureau on Child Support Payments, it seems unlikely that change in postdivorce family relationships is likely to be dramatic. A substantial minority of fathers are still unwilling or unable to pay the mandated child support, although these official figures do not always reveal informal payments that are passed from fathers to mothers and children. Even taking 'under the table' support into account, these results suggest a high degree of recalcitrance on the part of men to undertake support of their children or, alternatively, that men are giving a greater priority to supporting their current households than their former ones.

The largest segment of nonpaying parents is unmarried fathers whose ties with their children appear to become especially tenuous over time. This is also the fastest growing portion of childbearing couples who do not reside together. In most European countries, the growth of nonmarital childbearing has not implied the absence of cohabitation as it does in many of the Anglo-speaking countries, especially the USA. Although *de facto* marriages are ultimately less stable than *de jure* unions, they appear to share most of the features of nuclear families.

In contrast to most of the Western nations, the characteristic pattern of nonmarital childbearing in the USA is matrifocal, putting the father in a more marginal and tenuous position. This family form, especially common among low-income couples and minority populations, may be linked to economic and social marginality in the larger society. One of the most intriguing questions facing social demographers and household economists is whether a sustained economic boom will reverse this alternative pattern of family formation by making marriage more affordable. It seems that for less privileged segments of the population marriage has been considered a luxury item.

In the USA, and to a lesser degree in other Anglo-speaking countries, concern about family dissolution has generated considerable public discussion about the link between family change and the welfare of the next generation. Americans, particularly, are culturally committed to what might be called a highly 'privatized' family system, in which the State and family are seen as antagonists rather than allies. In countries where the opposite is the case, such as Scandinavia and northern European nations, the family and State are viewed as isomorphic. The State is expected to sustain and support the family, and parents are accustomed to receiving benefits for assuming parental responsibilities.

Some scholars have raised the question of whether the relationship between family and State has implications for the overall welfare of children. In short, are children better off when the State is held at arm's length, or when the State is granted more leeway in allocating resources and mandating requirements? The paucity of cross-national data and studies precludes an answer to this question. Rather little research has been undertaken that compares State and family arrangements or examines the institutions that mediate between the two, such as schools, community organizations, voluntary groups, and the like. Such research provides a fertile ground for understanding both the determinants and consequences of change in kinship systems and the resulting consequences for the welfare of children.

See also: Divorce, Sociology of; Family and Psychiatric Illness/Family Therapy; Gender and Reproductive Health; Gender Role Stress and Health; Love and Intimacy, Psychology of; Marriage and the Dual-career Family: Cultural Concerns; Men's Health

Bibliography

Bumpass L L 1990 What's happening to the family? Interactions between demographic and institutional change. *Demography* **27**: 483–98

Burch T K, Matthews B J 1987 Household formation in developed societies. *Population and Development Review* **13**: 495–511

Caldwell J C 1982 *Theory of Fertility Decline*. Academic Press, New York

Cherlin A 1981 *Marriage, Divorce, Remarriage*. Harvard University Press, Cambridge

Garfinkel I, McLanahan S 1986 *Single Mothers and Their Children: A New American Dilemma*. Urban Institute Press, Washington, DC

Gillis J R 1996 *A World of Their Own Making: Myth, Ritual, and the Quest for Family Values* 1st edn. Basic Books, New York
Goldthorpe J E 1987 *Family Life in Western Societies*. Cambridge University Press, Cambridge, UK
Goode W J 1963 *World Revolution and Family Patterns*. Free Press of Glencoe, New York
Hochschild A 1989 *The Second Shift*. Viking, New York
Stern B J 1938 *The Family Past and Present*. Appleton-Century, New York
Thornton A, Freedman D 1983 The changing American family. *Population Bulletin* **38**: 1–43
Thornton A, Fricke T E 1987 Social change in the family: comparative perspectives from Europe, China, and South Asia. *Sociological Forum* 2 **4**: 746–79
Watkins S C, Menken J, Bongaarts J 1987 Demographic foundations of family change. *American Sociological Review* **52**: 346–58

F. F. Furstenberg

Distance Education

1. *History: From Correspondence Education to Distance Education*

The beginning of what today we call 'distance education' is to be found in the concept of correspondence courses starting in the middle of the nineteenth century, with surface-mail corrections of assignments in, for example, shorthand courses or language-learning courses. Printed material as instructional provision was enriched by tutorial interaction via the mail.

The university extension movement, which enabled students in the colonies to prepare for a final examination, is another root of distance education. Elements such as written assignments and corrections, as well as instructions following the prescribed curriculum of an institution, could be more and more elaborated and diversified as the quality and ease of mail services all over the world developed.

Wherever the mail service was unreliable or wherever socialization processes were predominant, alternative forms of higher education evolved. An example is the former USSR, where until the middle of the twentieth century correspondence education was developed as a special form of higher education for the workforce, with personal contacts, seminars, and examinations offered on the premises of classical education institutions at times when the regular students were on vacation (e.g., in summer periods). Adult education with evening courses and correspondence education with vacation courses increased the efficiency of higher-education institutions: more than one-third of all higher-education examinations in the former USSR came from this kind of correspondence education.

There is also a tradition of distance education in rural regions (where geographic distances have to be overcome) and even for secondary schooling, for example in Australia. Some governments provide their staff and families with opportunities for distance education during stays abroad. During World War II, for example, the US government had a large department of distance education for the military.

As basic communication systems gradually changed in the middle of the twentieth century from surface mail to electronic systems, a gradual change from correspondence education to distance education took place. There was more and more dependency on the efficiency of modern media, leading to today's situation, in which even final examinations in some universities may take place as video conferences.

2. *Definition*

Distance education is an organizational form of education in which instructional provisions, tutorial interactions, monitoring of practice, as well as individual control of learning may take place via media which make the simultaneous personal presence of tutors and students avoidable.

3. *Theory*

The characteristic feature of distance education is the planned and controlled two-way interaction which differentiates it from all self-instructional methods. For a long time distance education was regarded as an inferior form of higher education, lacking all the possibilities of socializing students in an academic tradition.

Meanwhile the large institutions for distance education increase their leading role in national higher-education provision. Sir John Daniel (1996) of the British Open University speaks of 'mega-universities' regularly enrolling more than 100,000 students a year. In Europe at the end of the twentieth century more than 2.2 million students were enrolled in formal systems of distance education—in France alone the Centre National d'Education à Distance, founded in 1939, offers distance education as a form of continuing education for the public services to more than 400,000 students. With the decline of government-controlled higher education in the former socialist regions of Europe, there is also a growing 'private market' selling distance education courses to half a million people in Eastern European countries. Enrolments of more than 100,000 students each year are also reported from China and Korea.

The above-mentioned sense of distance education as inferior is fading away not only because of its success in terms of enrolment numbers but also because of the insight that the processes of planning or better 'designing' instruction, foreseeing instances for communication, including regular assessment possibilities for self-paced learning, and the necessity for constant quality control all have a long tradition in distance education and are seen to be the key factors of modern higher education.

3.1 Dropout Rates

One of the major problems of all distance education institutions is the fact that they have to report dropout rates of mostly more than 50 percent. The reason for such dropout figures is the heterogeneity of the students attracted by this form of higher education.

Many students drop out because they:

(a) just want to own course material for reference to keep up with the development in their subject area (nonstarter);

(b) try to work through the material and decide that the course does not meet their expectations (drawback);

(c) worked through the material, participating in regular assessment, but stopped when saturated (dropout); or

(d) worked through the whole course but decided not to sit for an examination (no-show).

It is clear that for adult learners none of these decisions can be associated easily with academic failure. Only in one of the categories, whenever 'dropout' students fail to meet their own goals, can we speak of 'dropout' in the usual sense of failure. Therefore the dropout rates do not give information about the didactic quality of the program but rather about bureaucratic problems of handling different types of customers with different types of interests.

3.2 Microdidactics for Megasystems

For a long time educational systems have been organized as representing a static concept of learning. Curricula with a plethora of old and new elements recommended as being useful for life or jobs were prescribed for young people—with initiation rites at the end of schooling (Vertecchi 1998). The clearer it became that learning will never end in one's lifetime, the faster and more regularly industry and changing administrations needed new qualifications, the faster educational systems had to adapt to these changing scenarios. One of these modern scenarios is distance education. In the middle of the twentieth century the striking analogy between distance education and the processes and pitfalls of early industrialization was presented by Peters (1983), who coined the expression

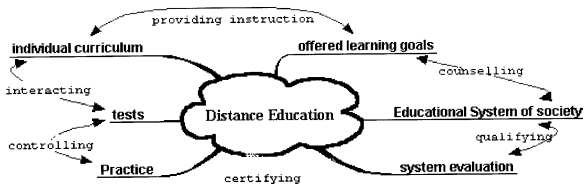

Figure 1
Structure of distance education

'DE as an industrialized form of teaching and learning,' mass production and specialization of the workforce in the production process being the characteristic features of large distance-education institutions. This subsequently led to the discussion of Fordism and labour-market theories in distance education (Rumble 1995). At the turn of the millennium observations in so-called virtual seminars led back to the microdidactic features of distance education in proving that characteristics of school learning would also take place in virtual settings (Fritsch 1997).

The discussions of quality control in the field of education at the turn of the twenty-first century showed that distance education institutions are much more at ease with the requirement of institutional evaluation than traditional institutions, where didactical situations are beyond control of the management because the door of the classroom is usually closed.

3.3 Media as Method

The didactic dimensions of the field of education are also valid for modern higher education using modern media:

(a) The pitfalls of misunderstanding media can easily be traced to the assumption that some areas of interest (content and goals) are easier to transmit via modern media than others.

(b) Media for instruction, especially increasing the number of possible students by mass production of teaching material (and thus the functions and characteristics of teachers), are the reason why distance education has suffered from an inferiority complex: virtual teaching situations are considered less valid than real world ones.

(c) There is a misunderstanding that media must be a separate element (what is the most convenient way of doing this?) in the dimensions of education. In fact, it is the method of teaching which forms this element, not the media. Consequently the didactic field has not changed in its basic structure.

Where media have a role to play in all fields is with regard to the question of interactions between didactic functions: 'qualifying, counselling, instructing, interacting, controlling and certifying.' Many interactions between didactic elements (see Fig. 1)—goals–content,

teacher–student, and practice–society—can also be realized using modern media. 'Non-contiguous forms' (Holmberg 1989) of interaction (transcending the limitations of time and place) in the educational process are in the foreground of distance-education theory.

For example: whenever a university uses campus TV or satellite TV to 'broaden' the scope of its audience, this should not yet be called distance education because the didactical setup remains dependent upon the simultaneous presence of student and teacher at defined places. Instead, this setup belongs to the concept of 'distributed learning' (Bates 1995). However, when a 'lecture' is integrated (e.g., as a video) into a defined setting of didactical considerations it might very well be part of a distance-education program, because the necessity for individuals to attend a certain place simultaneously is no longer a given. This argument does not mean that distance education will never provide face-to-face contact: most distance-education organizations do have provisions for social interaction or tutorial help. The question is which predefined didactical function can be best met by which medium.

3.4 Error-oriented Tutoring

Mass production of teaching material on the one hand is counterbalanced by the need to tutor large numbers of students on the other—as long as tutoring and individual help are carried out in the same way as in regular educational settings, distance education will not have a big economic effect on educational systems. But the possibility of designing tutorial help, via modern media, to include systems of automatic tutorial help makes distance education economically much more efficient. In classroom situations teachers know there are certain points in the learning process at which students always have similar difficulties. Good teachers not only indicate that an error has been made, but give assistance in the form of explanations, additional assignments, etc. Such error-oriented comments, explanations, and additional assignments can be geared automatically to the student exactly when an error occurs—just in time. In the early days of distance education the Open University, for example, had two different forms of assignments: tutor-marked assignments (TMA) and computer-marked assignments (CMA). It is more costly to have many TMAs and the turnaround time of CMAs can be far more effective (even more so if transferred to the World Wide Web for instant control). Fern University at Hagen, Germany, has developed a system of CMAs which provide the student with comments on typical errors made. Empirical research has shown that a very high percentage of previewable errors can be handled in this way. In projects of language learning, too, instant correction of previewed errors has proven to be a very important didactical feature in the use of modern media.

3.5 Transformation

In many cases concentrating on the element of instructional provision and transforming it into modern media is misunderstood as a way of attempting to transform the whole system of education, and neglecting other elements of teaching (tutorial interactions, monitoring of practice, and individual control of learning).

It is much easier to transfer existing courses of distance education to the World Wide Web, if they have been prepared in a way that is typical for high-quality education of this type, with:

(a) learning goals explicitly formulated;
(b) structured curriculum decisions according to individual interests;
(c) instances for monitored practice and tutorial help;
(d) control of learning outcomes; and
(e) evaluation of the system of teaching and learning.

4. State of the Art

Many large universities have found a new market in courses for continuing education and have installed departments for distance education, especially because of the need to cater for students who belong to the regular workforce or live far away and cannot be asked to come to campus for lectures or seminars. These universities are called 'dual-mode universities' because their academic position in the hierarchy of higher-education institutions has been thought to depend on the classical structure of also being a 'regular' university. Even the earliest forms of distance education institutions, such as the University of South Africa, are regarded as dual-mode universities. It was the British Open University which was the first to enter the higher-education sector as a purely single-mode distance-education university, and it has influenced the development of dozens of other universities worldwide. It is the impact of new technologies on both systems, traditional and distance education, which is forcing higher-education institutions to think about their specific potential. The distinction between campus-based and distance learning becomes more and more irrelevant as both learner groups use modes of distributed teaching. Asynchronous communicative acts for learning are no longer specific to the distance-education sector. The boundaries of campus have lost their relevance with the advent of the World Wide Web. Whether distance education will survive in future

depends on whether classical higher-education institutions are prepared to enter into a continuous dialog with each individual learner.

5. *Distance Education Institutions Developing into Virtual Universities*

There has been an explosion of distance-education courses made available through the Web—at the end of the 1990s linked Web descriptions of more than 15,000 different courses could be traced—and this comprises just the major institutions of English-speaking countries, which have a longer tradition in distance education. Global universities (http://www.global-learning.de/g-learn/aktuell/index.html), Globewide Network Academy (http://www.gnacademy.org), and World Lecture Hall (http://www.utexas.edu/world/lecture) are the main addresses of these initiatives. This booming advance has been made possible by the fact that *customers* want to know what they will 'get' for their participation, and internationally accepted credits are the key to this development. Without a functioning credit transfer system like the one students are familiar with in the USA, this development will halt at the borders of traditional tertiary education. As a result the development of a European Credit Transfer System has become a major topic in the European Union. But the time that elapses before such a European-wide system is introduced and accepted will show that this form of education is big business; people interested in continuing education are more interested in 'just-in-time' courses, taken when they need them, than in following a prescribed curriculum. Consequently, one of the goals of the International Council for Distance Education (ICDE), the worldwide association of major institutions acknowledged by UNESCO, which runs biannual conferences (http://www.icde.org), is to provide databases with information about all the courses available. This goal is also being pursued by the International Centre for Distance Learning (ICDL) at the Open University in Milton Keynes, which has produced a structured bibliography of distance learning (http://www.icdl.open.ac.uk/bib/toc.htm).

See also: Adult Education and Training: Cognitive Aspects; Computer Networking for Education; Computer-assisted Instruction; Education in Old Age, Psychology of; Educational Assessment: Major Developments; Educational Evaluation: Overview; Educational Media; Lifelong Learning and its Support with New Media: Cultural Concerns; Soviet Studies: Education

Bibliography

Bates A W 1995 *Technology, Open Learning and Distance Education*. Routledge, London

Daniel J 1996 *The Mega-Universities and the Knowledge Media*. Kogan Page, London

Fritsch H 1997 Witness-learning: Pedagogical implications of net-based teaching and learning. *Staff and Educational Development International* **1**: 107–28

Holmberg B 1989 *Theory and Practice of Distance Education*. Routledge, London, New York

Keegan D 1996 *Foundations of Distance Education*, 3rd edn. Routledge, London, New York

Peters O 1983 Distance education and industrial production: A comparative interpretation in outline. In: Sewart D, Keegan D, Holmberg B (eds.) *Distance Education: International Perspectives*. Croom Helm/Routledge, London, pp. 95–113

Rumble G 1995 Labour market theories and distance education I: Industrialisation and distance education. *Open Learning* **10**: 10–20

Vertecchi B 1998 *Elemente einer Theorie des Fernunterrichts*. ZIFF Papiere 109. FernUniversität, Hagen, Germany (http://www.fernuni-hagen.de/ZIFF/ziff109/109-contents.htm)

H. Fritsch

Distributions, Statistical: Approximations

Every random variable has its own particular distribution. It is, however, very useful if one can use a well-studied distribution as an approximation. This article describes a set of situations in which a random variable with a complicated probability distribution can be approximated well by another, better understood and the studied probability distribution. The first example is the normal distribution which approximates the distribution of many random variables arising in nature. This approximation is based on the central limit theorem. Generalizations of this theorem provide a second set of approximations called the stable laws. A third example arises when one is considering extremes, or the maximum of a set of random variables. Here, the family of extreme value distributions is shown to provide accurate approximations. Moreover, the excesses over a high level can be approximated by a generalized Pareto distribution. The final example involves the distribution of the times between events, such as system failures. Under certain conditions, these times can be well approximated by an exponential distribution. Consequently, the corresponding point process (or event history) can be approximated by a Poisson process.

1. *Normal Approximations*

The normal or Gaussian distribution has a probability density function (p.d.f.) given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \qquad -\infty < x < +\infty$$

where the parameter μ is the mean or median and σ is the standard deviation. The standard normal distribution corresponds to the special case in which $\mu = 0$ and $\sigma = 1$, for which the p.d.f. is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}x^2), \quad -\infty < x < +\infty$$

The normal distribution serves as an excellent approximation to many random variables that arise in nature. The reason for this lies in the central limit theorem, the simplest version of which states:

Theorem (Central limit theorem)
Suppose $\{X_r\}_{n=1}^{\infty}$ is a sequence of independent identically distributed (i.i.d.) random variables with cumulative distribution function F, mean μ, and variance $\sigma^2 < \infty$. Then

$$Z_n = \frac{\Sigma_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}} \Rightarrow \Phi$$

where Φ represents the c.d.f. of a standard normal distribution and $\Rightarrow$ denotes weak convergence.

This theorem says that the standardized partial sums of the $\{X_n\}_{n=1}^{\infty}$ sequence converge weakly to a standard normal distribution.

This theorem has many generalizations which allow for the $\{X_n\}_{n=1}^{\infty}$ to have different distributions (so long as each random variable has finite variance and contributes only a small part to the sum) and to have a dependency structure (so long as the correlation among the variables diminishes rapidly enough). The reader should consult *Distributions*, *Statistical*: *Special and Continuous* for additional information about the normal distribution. The textbook by DeGroot (1986) offers examples of the application of the central limit theorem, some extensions of the basic theorem, and discussions of weak convergence, also known as convergence in distribution. The reader should consult the text by Billingsley (1995) for the precisely stated theorems.

The central limit theorem indicates that suitably standardized sums of independent random variables will have an approximate normal distribution. Thus, any random variable which arises as the sum of a sufficiently large number of small random components can be modeled accurately by a normal distribution. This is especially appropriate for error distributions, where the error term is the sum of a number of different sources of error. The question of how large n must be for the approximation to be accurate depends on two factors: (a) the degree of accuracy required; and (b) the underlying distribution F. A common 'rule of thumb' is that the normal approximation can be safely used when $n \geqslant 30$ as long as F is not highly skewed and a high degree of accuracy is not required. The normal distribution is commonly used as a model for many statistical problems especially in linear models; see *Linear Hypothesis* and *Linear Hypothesis*: *Regression* (*Basics*) for details.

2. *Stable Distributions*

In this section, we present an extension of the central limit theorem. When $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables, each having a finite variance, then the partial sums of these variables can be normalized to converge weakly to a normal distribution. When the underlying random variables do not have a finite variance, then normalization of the partial sums can still lead to weak convergence to a nondegenerate limiting distribution; however, the limiting distribution will no longer be normal, but will instead be a *stable* distribution.

Definition (Stable distributions)
A distribution F is stable if given any three i.i.d. random variables, X_1, X_2, X_3, with distribution F and any constants c_1, c_2, there exists constants a and b (depending upon c_1, c_2) such that $c_1X_1 + c_2X_2$ has the same distribution as $a + bX_3$. If $a = 0$, the F is *symmetric stable*.

This means that any linear combination of two i.i.d. random variables with a common stable distribution can be transformed linearly to have the same stable distribution. This definition can be used to prove that if $\{X_i\}_{i=1}^{n}$ are i.i.d. with a stable distribution F, then there exist constants a_n, b_n such that $\frac{\Sigma_{i=1}^{n} X_i - a_n}{b_n}$ also has c.d.f. F. The symmetric stable distributions are most easily described by their characteristic functions.

Theorem (Characteristic function of a symmetric stable law)
A standard symmetric stable distribution has characteristic function $\phi(s) = e^{-|s|^{\alpha}}$ for $0 < \alpha \leqslant 2$.

Two special cases are the normal (0,1) distribution ($\alpha = 2$) and the standard Cauchy distribution ($\alpha = 1$). The following theorem indicates that the non-degenerate stable laws represent all possible limiting distributions for sums of i.i.d. random variables.

Theorem
Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables with cumulative distribution function F, and suppose that there are sequences of constants, $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ for which $\frac{\Sigma_{i=1}^{n} X_i - a_n}{b_n} \Rightarrow G$, a nondegenerate limiting c.d.f. Then G is a nondegenerate stable law.

The family of stable distributions represents a wider class of possible limits of sums of i.i.d. random

variables. The distributions other than the normal have heavier tails than the normal distribution. Informally, a symmetric distribution with exponent $\alpha < 2$ has finite moments of order less than α. For example, the Cauchy distribution ($\alpha = 1$) does not have a first moment, but all moments less than 1 are finite. The stable distributions have been used to model certain economic variables and the return of financial assets such as stocks or currency exchange rates. See *Distributions, Statistical: Special and Continuous* for additional information on the Cauchy distribution. The reader should consult Billingsley (1995), Embrechts et al. (1997), or Lamperti (1966) for more detailed information on stable distributions and the theorems underlying this more general central limit problem.

3. Extreme Value Distributions

Another type of approximation occurs when one is studying the maximum of a set of random variables. This situation is particularly relevant in the casualty insurance industry, where interest centers on the largest claims. These situations also arise in engineering and environmental problems, where a structure such as a dam or building must withstand all forces, including the largest. Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables with common distribution function F. Suppose that x^* denotes the rightmost support point of F, the smallest x for which $F(x) = 1$, or $+\infty$ if no such x exists. We consider $\{M_n\}_{n=1}^{\infty}$ where $M_n = \max_{\{1 < i < n\}} X_i$, the largest among the first n random variables in the sequence. Fisher and Tippett (1928), a classic paper, showed that if there are sequences of constants $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ for which $\frac{M_n - a_n}{b_n} \Rightarrow H$, a nondegenerate limiting c.d.f., then H takes on one of three possible forms:

(a) The Gumbel distribution:

$$H(x) = \exp(-e^{-x}), \quad -\infty < x < +\infty$$

This limiting form arises when $x^* = \infty$ and the underlying distribution F has a right-hand tail which decreases at an exponential rate. Specifically, if there exists a constant $b > 0$ for which $\lim_{x \to \infty} \exp(x)(1 - F(x)) \to b$, then $M_n - \log(nb)$ converges to the Gumbel distribution as $n \to \infty$.

(b) The Fréchet distribution

$$H(x) = \begin{cases} \exp(-x^{-\alpha}) & \text{if } x > 0 \\ 0 & \text{if } x \leqslant 0 \end{cases}$$

where $\alpha > 0$. This limiting form arises when $x^* = \infty$ and the right tail of the underlying distribution F has a tail which decays at a polynomial rate. Specifically, if $\lim_{x \to \infty} x^{\alpha}(1 - F(x)) \to b$, for some constant $b > 0$ and $\alpha > 0$, then $M_n/(bn)^{1/\alpha}$ converges to the Fréchet distribution as $n \to \infty$.

(c) The Weibull distribution

$$H(x) = \begin{cases} \exp(-(-x)^{\alpha}) & \text{if } x \leqslant 0 \\ 1 & \text{if } x \geqslant 0 \end{cases}$$

where $\alpha > 0$. The above distribution is concentrated on $(-\infty, 0)$, and it is the negative of the traditional Weibull distribution. This limiting form arises when $x^* < \infty$, i.e., the underlying distribution F is bounded above by a constant. Specifically, assume $\lim_{x \to x^*} (x^* - x)^{-\alpha}(1 - F(x)) \to b$ for some constants $b > 0$ and $\alpha > 0$, then $(nb)^{1/\alpha}(M_n - x^*)$ converges to the Weibull distribution defined above as $n \to \infty$. The reader should consult *Distributions, Statistical: Special and Continuous* for additional information on the Weibull distribution.

Fisher and Tippett (1928) proved that these are the only possible limiting forms. It is, however, important to point out that, for some underlying F, there may be no set of constants $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ for which $\frac{M_n - a_n}{b_n} \Rightarrow H$. The conditions under which this convergence occurs are complicated, and the interested reader is referred to Embrechts et al. (1997). Lamperti (1966) also provides a nice treatment of extreme value theory.

The three extreme value distributions given above describe the limiting of the maximum of a random sample. There is a second random quantity that is related to the maximum and is interesting in it own right. Suppose we begin with a random variable X with c.d.f. F and consider the conditional excess distribution of X, $X - u$, given that it exceeds u. Specifically, consider $F_u(x) = P(X - u \leqslant | X > u)$. One could also consider the expected excess, conditional on the random variable exceeding u, $e(u) = E(X - u | X > u)$. This distribution and expectation arise in engineering and insurance problems, where the quantity above a certain high limit is of concern, such as the insurance claim above a large deductible amount.

The standard generalized Pareto distribution is a family of c.d.f.s indexed by a parameter $\theta \in (-\infty, +\infty)$ given by

For $\theta < 0$,

$$G_{\theta}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - (1 + \theta x)^{-1/\theta} & \text{if } 0 \leqslant x \leqslant -1/\theta \\ 1 & \text{if } x \geqslant -1/\theta \end{cases}$$

For $\theta = 0$,

$$G_0(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x} & \text{if } x \geqslant 0 \end{cases}$$

For $\theta > 0$,

$$G_{\theta}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - (1 + \theta x)^{-1/\theta} & \text{if } x \geqslant 0 \end{cases}$$

The family can be enlarged by adding a location and scale parameter.

For a given value of θ, the probability distribution of F_u approaches that of a generalized Pareto distribution as $u \to x^*$. Thus, the generalized Pareto distribution can be used to approximate the probability distribution of excesses over a high limit. The mean excesses can also be modeled using a generalized Pareto distribution.

The reader should consult Embrechts et al. (1997), a comprehensive treatment of extreme value theory and for exact conditions and statements of these approximation results.

4. *Times Between Events in a Point Process*

A random point process is a sequence of events (or points) that occur in time. Common examples of point processes are event histories, the failure times of systems, arrivals of customers to a service facility or the firing of neurons in the human nervous system. Clearly, modeling such point process phenomena can be very complicated and detailed. There is, however, an approximate result that is widely useful in such modeling. Specifically, for an important class of point processes, a reasonable approximation is that the times between events are i.i.d. random variables with an exponential distribution, a p.d.f. given by:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geqslant 0 \\ 0 & \text{if } x < 0 \end{cases}$$

with $\lambda > 0$. The parameter λ measures the rate at which the events occur and is measured in units of events per unit time. Such a point process is called a Poisson (λ) process. Poisson processes have three important properties: (a) the number of events in an interval of length L is a random variable with a Poisson (λL) distribution; (b) the number of events in non-overlapping time intervals are independent random variables; and (c) the probability of exactly one event in an interval of length h, divided by h, converges to λ as $h \to 0$. Hence events cannot occur simultaneously in a Poisson process. See *Distributions, Statistical: Special and Continuous* for additional information on the exponential distribution.

Now, suppose that the point process is composed of the superposition of a large number of independent point processes, each individually with insignificant intensity function. Then the superposed (aggregated) point process converges (as the number of components superposed increases) to a Poisson process. The reader should consult the text by Karlin and Taylor (1975) for a precise statement of this theorem. This situation is very realistic for system failures in which the system is composed of many different parts, the failure of any one of which leads to system failure. Each component has its own failure distribution, and the time to the next failure will be the minimum of these failure times, which will converge to an exponential distribution as the number of components increases. A similar situation occurs with aggregate customer arrival processes, where the aggregate process is the superposition of a large number of different arrival processes. As the number of different arrival sources increases, the limiting aggregate arrival process will become Poisson. The reader should consult *Queues* to learn about the use of the exponential distribution, and the Poisson process as a model for customer arrival processes.

5. *Further Reading*

The reader should consult a basic textbook on probability such as DeGroot (1986) for information on many of the distributions mentioned in this article, for a description of the basic central limit theorem and examples of its use, and for some more general results. Billingsley (1995) provides a more comprehensive and technically detailed treatment of the central limit theorem, the stable laws and the extreme value distributions. Finally, the book by Embrechts et al. (1997) gives a comprehensive treatment of the extreme value distributions and associated approximation theorems.

Bibliography

Billingsley P 1995 *Probability and Measure*, 3rd edn. Wiley, New York

DeGroot M H 1986 *Probability and Statistics*, 2nd edn. Addison-Wesley, Reading, MA

Embrechts P, Kluppelberg C, Mikosch T 1997 *Modeling Extremal Events*. Springer-Verlag, Berlin

Fisher R A, Tippett L H C 1928 Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society* **24**: 180–90

Karlin S, Taylor H 1975 *A First Course in Stochastic Processes*, 2nd edn. Academic Press, New York

Lamperti J 1966 *Probability*. W. A. Benjamin, New York

J. P. Lehoczky

Distributions, Statistical: Special and Continuous

Continuous probability distributions are used widely in probability and statistics when the underlying random phenomenon is measured on a continuous scale. Common examples include the time until an

event occurs, such as the failure time of a component or system, or most human physical dimensions such as height or weight. In some cases, the underlying random quantity has a discrete set of possible values. If that set is large, it may be possible to rescale those values and treat them as continuous. Continuous distributions arise frequently in the social sciences, for example in IQ or educational testing settings in which the scale is designed to yield a normal distribution of performance. Economic measurements such as income levels are often modeled using continuous distributions. Many observed distributions do not fit any of the standard distributions presented in this article (for example, they may be skewed when compared with the normal distribution); however, it may be possible to apply a transformation to the variable in question to make it fit one of the standard distributions better. In this article, we begin by describing univariate probability distributions, then specialize to the absolutely continuous case. The concept of a failure rate function as an alternative way to describe an absolutely continuous distribution is also introduced. The next section introduces special continuous distributions, including the uniform, beta, normal, exponential, gamma, Weibull, Chi-square, Student t, F, and Pareto. The entry concludes with a definition of multivariate distributions, and the bivariate normal distribution is presented.

1. Definitions and General Properties

A univariate random variable X has a probability distribution which is defined by its cumulative distribution function (c.d.f.),

$$F(x) = P(X \leqslant x), \quad -\infty < x < +\infty$$

A c.d.f. has several properties:

(a) $F(x)$ is non-decreasing, i.e., $F(x) \leqslant F(y)$ if $x < y$;

(b) $\lim_{x \to -\infty} = F(-\infty) = 0$;

(c) $\lim_{x \to +\infty} = F(+\infty) = 1$;

(d) F is right continuous, i.e. $\lim_{h \to 0} F(x+h) = F(x)$; and

(e) $P(a < X \leqslant b) = F(b) - F(a)$, $-\infty \leqslant a < b \leqslant +\infty$.

When the c.d.f. is a step function, $\{(x_i, p_i)\}_{i=1}^{\infty}$, with a jump of height p_i at location x_i, $1 \leqslant i < \infty$, then the associated probability distribution is called a *discrete distribution*. If the c.d.f. is a continuous function, the associated probability distribution is called a *continuous distribution*. When the c.d.f. has jumps satisfying $0 < \sum_{i=1}^{\infty} p_i < 1$, the distribution is said to be *mixed*.

For continuous distributions, we can define the derivative, $f(x) = dF(x)/dx$. The function f is called the probability density function (p.d.f.). When the c.d.f. is absolutely continuous, then one can integrate the p.d.f. to recover the c.d.f., that is,

$$F(x) = \int_{-\infty}^{x} f(u)\, \mathrm{d}u, \quad -\infty < x < \infty$$

One can calculate the probability of an event by integrating the density function; for example, if the event is an interval $[a, b]$, then

$$P(a < X < b) = \int_{a}^{b} f(x)\, dx$$

If X has a continuous probability distribution, then $P(X = x) = 0$ for any value $-\infty < x < +\infty$. The interpretation of the p.d.f. is that $P(x \leqslant X \leqslant x+h) \approx f(x)h$ for small $h > 0$. That is, the probability a random variable with an absolutely continuous probability distribution takes a value in a small interval of length h is given by the probability density times the length of the interval.

For absolutely continuous distributions, the c.d.f. F and the p.d.f. f are equivalent in that one can be derived from the other. An absolutely continuous probability distribution might also be described by its failure rate function, also known as the hazard function. This function is often denoted by $\lambda(x)$ and is defined by:

$$\lambda(x) = \frac{f(x)}{1 - F(x)}$$

The failure rate function arises naturally in specifying the reliability of a system or in mortality tables. This function gives the probability that a system will fail (person will die) in the next small interval of time conditional on it not yet having failed (died) up to the present time. The failure rate function satisfies

$$P(X \leqslant x+h \mid X \geqslant x) \approx \lambda(x)h.$$

The c.d.f., $F(x)$, can be derived from $\lambda(x)$ from the formula

$$F(x) = 1 - \exp\left(-\int_{-\infty}^{x} \lambda(u)\, du \right)$$

where $\lambda(x)$ can be any function satisfying $\lambda(x) \geqslant 0$, $\int_{-\infty}^{x} \lambda(u)\, du < \infty$, $-\infty < x < \infty$, and $\int_{-\infty}^{+\infty} \lambda(u)\, du = +\infty$. See *Survival Analysis: Overview* and *Quality Control, Statistical: Reliability and Life Testing* for additional information about the use of the failure rate function.

The c.d.f., p.d.f., or the failure rate function give complete information about the distribution of a random variable having an absolutely continuous distribution. It is, however, often useful to provide a few simple characteristics or features of the distribution. Typically, one specifies the mean and the variance of the distribution. For absolutely continuous distributions, these can be computed from the p.d.f. by

$$E(X) = \int_{-\infty}^{+\infty} xf(x)\,dx,$$

$$\mathrm{Var}\,(X) = \int_{-\infty}^{+\infty} x^2 f(x)\,dx - (E(X))^2$$

2. *Individual Continuous Distributions*

This section describes some of the most important univariate continuous distributions.

2.1 *Uniform Distribution*

This is a basic distribution with two parameters, a and b denoting the endpoints of an interval. The p.d.f. is given by:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leqslant x \leqslant b, \\ 0 & \text{otherwise} \end{cases}$$

If X has a Uniform (a, b) distribution, then for any interval (c, d) satisfying $a \leqslant c < d \leqslant b$, $P(c < X < d) = (d-c)/(b-a)$. The mean and variance are given by

$$E(X) = \frac{(b+a)}{2},\ \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

The Uniform (0, 1) distribution is often referred to as the standard uniform distribution. If U is a random variable with standard uniform distribution, then $X = a+(b-a)U$ has Uniform (a, b) distribution. Thus, any uniform random variable can be constructed from a standard uniform random variable. This property of using standard uniform random variables to construct other random variables with specific distributions is quite general. The method is called the probability integral transform. Specifically, suppose X is a random variable with a c.d.f. given by F. Let $F^{-1}(p)$, $0 < p < 1$ denote the inverse of F defined by

$$F^{-1}(p) = \inf\{(y \mid F(y) \geqslant p\}$$

It can be shown that if U has Uniform (0, 1) distribution, then $X = F^{-1}(U)$ has c.d.f. given by F. This property is especially important in simulation studies in which one must generate a sequence of random variables with some prescribed probability distribution. If F^{-1} can be computed, then one need only generate a sequence of standard Uniform (0, 1) random variables, plug each of them into F^{-1}, and the result will be a sequence with the required distribution. See *Random Numbers* for more details.

2.2 *Beta* (α, β) *Distribution*

The Uniform (0, 1) distribution is a special case of a two parameter family called the Beta (α, β) distribution. It has a p.d.f. given by:

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta} & \text{if } 0 \leqslant x \leqslant 1, \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha > 0$, $\beta > 0$ and $\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}\,\mathrm{d}x$.

The Beta (1, 1) distribution is the same as the Uniform (0, 1) distribution. The mean and variance of a random variable with Beta (α, β) distribution are given by

$$E(X) = \frac{\alpha}{\alpha+\beta},\quad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The Beta distribution offers a very flexible two parameter family of distributions for random variables taking values between 0 and 1. Consequently, it can often be used to model percentage or fractional quantities.

2.3 *Normal* (μ, σ^2) *Distribution*

One of the most important probability distributions is the normal or Gaussian distribution. Its p.d.f. is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right), \quad -\infty < x < +\infty$$

The parameter μ is the mean, while σ^2 is the variance. Since the p.d.f is symmetric about μ, this parameter is also the median.

The standard normal distribution corresponds to the case in which $\mu = 0$ and $\sigma = 1$. It has p.d.f. given by

$$f(x) = \frac{1}{\sqrt{2x}} \exp(-\tfrac{1}{2}x^2), \quad -\infty < x < +\infty.$$

One can construct a Normal random variable X, with mean μ and variance σ^2 from a random variable, Z, with standard normal distribution by using the relationship $X = \mu + \sigma Z$. Consequently, the calculation of

the probability of an event involving X can be reduced to a calculation involving Z. For example,

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$$
$$= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

where $\Phi(x)$ is the c.d.f. of the standard normal distribution. The function $\Phi(x)$ is widely tabled and is given in nearly every basic statistics textbook.

The normal distribution is of great importance, because it is an excellent approximation for the probability distribution of many random variables that arise in nature. The reason for this is based on the central limit theorem. This theorem states that if a random variable arises as the sum of a large number of independent random variables with a common probability distribution having a positive but finite variance, then the distribution of the sum has an approximate normal distribution; see *Probability: Formal* for additional details. The normal distribution is widely used in the social sciences. It is often used to model the distribution of IQ and educational testing measurements within a population. It is commonly used as the distribution of the error term in regression analysis and other linear models problems; see *Linear Hypothesis* and *Linear Hypothesis: Regression (Basics)* for details. Given the importance of this distribution as a basis for much of statistical inference, in cases for which the normal dstribution does not serve as an appropriate distribution for a random variable, one often attempts to find a transformation such that the transformed variable will have an approximate normal distribution.

2.4 *Exponential* (λ) *Distribution*

The exponential distribution arises frequently in problems involving system reliability and the times between events. This distribution is most easily described using the failure rate function, which for this distribution is constant, i.e.,

$$\lambda(x) = \begin{cases} \lambda & \text{if } x \geqslant 0, \\ 0 & \text{if } x < 0 \end{cases}$$

The constancy of the failure rate function leads to the memoryless or Markov property associated with the exponential distribution. Specifically,

$$P(X > s+t \mid X > s) = P(X > t)$$

The interpretation of the above property is easiest to understand when X represents the lifetime of a component or system. The memoryless property indicates that if the component has lasted for s units of time and is still functioning, then the probability that it will last an additional t time units is independent of s. Thus, the likelihood of an additional life of at least t time units does not depend on its current age, hence the term 'memoryless.'

The p.d.f. is given by

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } 0 \leqslant x < +\infty, \\ 0 & \text{otherwise} \end{cases}$$

for $\lambda > 0$.

If X has an exponential (λ) distribution, then

$$E(X) = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

The exponential distribution is the basis for continuous time Markov chain models. This distribution is used to model the sojourn or holding time in a state for a continuous time chain; see *Markov Models and Social Analysis*. Moreover, the distribution is often used as a model for the time between events; for example, in modeling the arrivals of customers to a service center; see *Distributions, Statistical: Approximations* and *Queues*.

2.5 *Gamma* (α, λ) *Distribution*

The Gamma (α, λ) distribution generalizes the exponential distribution. The parameters must satisfy $\alpha > 0$ and $\lambda > 0$. The p.d.f. is given by

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & \text{if } 0 \leqslant x < +\infty, \\ 0 & \text{otherwise} \end{cases}$$

Thus a Gamma (1, λ) is an exponential distribution with mean $1/\lambda$. Suppose α is an integer, and $\{X_i, 1 \leqslant i \leqslant \alpha\}$ are independent random variables, each with a Gamma (m, λ) distribution, then $S = \sum_{i=1}^{\alpha} X_i$ has a Gamma (αm, λ) distribution.

If X has a Gamma (α, λ) distribution, then

$$E(X) = \frac{\alpha}{\lambda}, \quad \mathrm{Var}(X) = \frac{\alpha}{\lambda^2}$$

The Gamma distribution offers a two-parameter family of distributions of positive random variables with many possible shapes. This is particularly useful for modeling event histories in social science studies.

2.6 Weibull (A, λ) Distribution

Another probability distribution that arises in reliability and event history modeling is the Weibull (α, λ) distribution for $\alpha > 0$ and $\lambda > 0$. The p.d.f. is given by

$$f(x) = \begin{cases} \alpha\lambda^{\alpha}x^{\alpha-1}e^{-(\lambda x)^{\alpha}} & \text{if } x \geqslant 0, \\ 0 & \text{otherwise} \end{cases}$$

The failure rate function for the Weibull distribution is given by

$$\lambda(x) = \alpha\lambda^{\alpha}x^{\alpha-1}$$

If $0 < \alpha < 1$, then the failure rate function is decreasing, whereas if $\alpha > 1$ the failure rate function is increasing. The $\alpha = 1$ case results in a constant failure rate which corresponds to the exponential distribution.

If a random variable X has Weibull (α, λ) distribution, then X^{α} has an Exponential (λ^{α}) distribution. Thus, when $\alpha = 1$, this distribution becomes the Exponential (λ) distribution.

The Weibull distribution is often used in reliability or survival analysis (or life testing) studies; see *Quality Control, Statistical: Reliability and Life Testing*. The distribution can exhibit a wide range of tail behaviors, especially heavy tails. The Weibull distribution is one of the three *extreme value distributions*, distributions which arise by normalizing the maximum of a random sample from a distribution; see *Distributions, Statistical: Approximations*.

2.7 Chi-square (n) Distribution

The Chi-square distribution with n degrees of freedom has p.d.f. given by

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)}x^{(n/2)-1}e^{-x/2} & \text{if } x \geqslant 0, \\ 0 & \text{otherwise} \end{cases}$$

This distribution is a special case of the Gamma (α, λ) distribution with $\alpha = n/2$ and $\lambda = \frac{1}{2}$. It arises as a sum of squares of independent standard normal random variables. Specifically, suppose $\{X_i, 1 \leqslant i \leqslant n\}$ are independent standard normal random variables, then $\sum_{i=1}^{n} X_i^2$ has a Chi-square (n) distribution.

The mean and variance of a random variable having a Chi-square (n) distribution are given by

$$E(X) = n, \quad \mathrm{Var}(X) = 2n.$$

The Chi-square distribution is very widely used in statistics, especially in goodness of fit testing (see *Goodness of Fit: Overview*) and in categorical data analysis. The distribution also arises as the distribution for the sample variance estimator of an unknown variance based on a random sample from a normal distribution.

2.8 Cauchy Distribution

The standard version of the Cauchy distribution has a p.d.f. given by

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < +\infty.$$

This distribution is symmetric about 0; however, one can introduce both location and scale parameters which can move the center and change the concentration of the distribution.

The Cauchy distribution has a very heavy tail, comparable to the tail of the Pareto $(1, c)$ distribution. In fact, the tail is so heavy that the distribution does not have a mean value. The Cauchy distribution is a *stable distribution*, see *Distributions, Statistical: Approximations*. In particular, it can arise as a properly normalized sum of independent random variables having a suitably heavy tail. The Cauchy distribution and related stable distributions have been used to model certain economic variables, especially financial asset returns.

2.9 Student $t(n)$ Distribution

The Student t distribution with n degrees of freedom arises in statistical inference problems associated with the mean of a normal distribution when both the mean and the variance are unknown. If Z and Y are independent random variables, with Z having a standard normal distribution and Y having a Chi-square (n) distribution, then the ratio nZ/Y has Student t distribution with n degrees of freedom. It has a p.d.f. given by

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}, \quad -\infty < x < +\infty$$

The Student t distribution with 1 degree of freedom is the same as the Cauchy distribution. For large n, the Student t distribution is well approximated by the standard normal distribution. The Student t distribution is widely used in the formation of confidence intervals for an unknown mean of a normal distribution estimated from a sample of data from a normal distribution with both mean and variance unknown. It also arises in the formation of confidence intervals for regression parameters; see *Linear Hypothesis: Regression (Basics)*.

2.10 $F(m, n)$ Distribution

The $F(m, n)$ distribution arises in statistical inference problems associated with the normal distribution, especially in the analysis of variance to determine the significance of effects. If random variables V and W are independent, each with a Chi-square distribution

having degrees of freedom given by m and n, respectively, then the ratio $(V/m)/(W/n)$ has an $F(m, n)$ distribution. The p.d.f. is given by

$$f(x) = \begin{cases} \dfrac{\Gamma((m+n)/2)\, m^{m/2} n^{n/2}}{\Gamma(m/2)\Gamma(n/2)} \dfrac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}} & \text{if } x \geqslant 0, \\ 0 & \text{otherwise} \end{cases}$$

The F distribution is of great importance in the analysis of variance and the design of experiments. This distribution is used to determine the significance of different effects in an analysis of variance study; see *Experimental Design: Overview*; *Linear Hypothesis*; *Analysis of Variance and Generalized Linear Models*.

2.11 Pareto (α, c) Distribution

This is a two-parameter family of distributions indexed by $\alpha > 0$ and $c > 0$. The p.d.f. is given by

$$f(x) = \begin{cases} \dfrac{\alpha c^{\alpha}}{x^{\alpha+1}} \text{if } x \geqslant c, \\ 0 \quad \text{otherwise} \end{cases}$$

This distribution has a relatively heavy tail which decreases at only a polynomial rate. Consequently, only moments of order less than α exist. In particular, if X has a Pareto (α, c) distribution, then

$$E(X) = \frac{\alpha}{\alpha-1} c, \quad \text{if } a > 1$$

and

$$\mathrm{Var}(X) = \frac{\alpha}{(\alpha-1)^2(\alpha-2)} c^2, \quad \text{if } a > 2$$

If X has a Pareto (α, c) distribution, then the transformed random variable $\log(X/c)$ has an Exponential (α) distribution.

The Pareto distribution arises in contexts in which the tail of the distribution is relatively heavy. This occurs especially in economics; for example, in the distribution of incomes and the returns from various financial assets. The distribution also arises as the excess of a random variable over a high limit. The excess is especially important in the casualty insurance where one is interested in the excess of a claim over a high deductible limit; see *Distributions, Statistical: Approximations*.

3. Multivariate Distributions

It often happens that it is wished to consider two or more random variables simultaneously rather than individually. If $\{X_i, 1 \leqslant i \leqslant n\}$ is a set of n random variables, then each will have its own probability distribution, F_i, $1 \leqslant i \leqslant n$. This individual or marginal information does not offer insight into the relationship among these variables. To capture these relationships, one must consider the n-dimensional random vector $(X_1,\ldots, X_n)$. To do this, one defines the higher dimensional analogs of the c.d.f. and p.d.f. Specifically, the n-dimensional c.d.f. (or joint c.d.f.) is defined by

$$F(x_1,\ldots, x_n) = P(\{X_1 \leqslant x_1\} \cap \ldots \cap \{X_n \leqslant x_n\}), \quad \text{for } -\infty < x_i < +\infty, \quad 1 \leqslant i \leqslant n$$

If the joint c.d.f. is absolutely continuous, then one can define the joint p.d.f.,

$$f(x_1,\ldots, x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_1,\ldots, x_n)$$

In the absolutely continuous case, one can calculate the probability of events by integrating the joint p.d.f. over the values specified by the event. When we consider the joint distribution of two random variables, (X_1, X_2), the associated joint c.d.f. and joint p.d.f. are referred to as the bivariate c.d.f. and bivariate p.d.f. A special case of this occurs when the pair of random variables has a bivariate normal distribution. See *Multivariate Analysis: Overview* for additional detail.

3.1 Bivariate Normal $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ Distribution

The bivariate normal distribution is a special joint distribution associated with two random variables, X_1 and X_2. The joint p.d.f. of (X_1, X_2) is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2} \exp\left(-\frac{1}{2(1-\rho^2)} \right) \left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) + \left(\frac{x_2-\mu_2}{\sigma_2} \right)$$

for $-\infty < x_1, x_2 < +\infty$.

The parameters μ_i, $i = 1,2$ are arbitrary; however, the other parameters must satisfy $\tau_i > 0$, $i = 1,2$ and $-1 < \rho < +1$. When $\rho = \pm 1$, the random variables X_1, X_2 are related linearly.

If a pair of random variables (X_1, X_2) has a bivariate normal distribution with the parameters given above, then each has a marginal normal distribution, i.e., X_i has a Normal(μ_i, σ_i), $i = 1,2$ distribution. The parameter ρ is the correlation coefficient of X_1 and X_2. The bivariate normal distribution has the special property that if $\rho = 0$, then X_1 and X_2 are independent random

variables. Generally, uncorrelated random variables are *not* independent.

One can represent (X_1, X_2) in terms of two independent random variables (Z_1, Z_2), each with a standard normal distribution through the relationship:

$$X_1 = \mu_1 + \sigma_1 Z_1,$$
$$X_2 = \mu_2 + \sigma_2(\rho Z_1 + \sqrt{1-\rho^2} Z_2).$$

The bivariate normal distribution arises naturally in simple regression problems, see *Linear Hypothesis: Regression (Basics)*.

4. Further Reading

A large amount of material on continuous distributions, including statistical inference for unknown parameters of the distribution, is given in the text by Johnson et al. (1994). Multivariate continuous distributions are covered extensively by Johnson and Kotz (1972). Many of the above distributions are discussed in basic probability textbooks; for e.g., the text by DeGroot (1987). Some illustrations of the use of continuous probability distributions in the social sciences are given by Handcock and Morris (1999).

See also: Distributions, Statistical: Approximations; Distributions, Statistical: Special and Discrete; Probability: Formal

Bibliography

DeGroot M H 1987 *Probability and Statistics*, 2nd edn. Addison-Wesley, Reading, MA

Handcock M S, Morris M 1999 *Relative Distribution Methods in the Social Sciences*. Springer-Verlag, New York

Johnson N L, Kotz S 1972 *Continuous Multivariate Distributions*. Wiley, New York

Johnson N L, Kotz S, Balakrishnan N 1994 *Continuous Univariate Distributions*, 2nd edn. Wiley, New York

J. P. Lehoczky

Distributions, Statistical: Special and Discrete

Discrete probability distributions form the basis of much of the statistical methodology for modeling, analyzing, and making inferences about discrete data (sometimes known as count data). Usually each observation (outcome) is an integer; for example, the number of individuals in a family, or the number of telephone calls handled by a switchboard in a unit of time. This article is primarily concerned with these underlying distributions and their properties, rather than their applications. The article first discusses the general properties of univariate discrete distributions and then defines and examines a number of widely used distributions. Ways of combining distributions to form new distributions are briefly considered. Finally multivariate discrete distributions (more than one variable is measured for each outcome) are introduced and the multinomial distribution (which is involved in the analysis of contingency tables) is examined (see *Multivariate Analysis: Discrete Variables (Overview)*).

1. Definitions and General Properties

Formally, a (univariate) discrete random variable (which is used to model the observation) X can take the values $x_1, x_2, x_3, \ldots$. Usually these $\{x_i\}$ are restricted to be the nonnegative integers or some subset of them. For example, an investigator might assume that $\{1, 2, 3, \ldots, 20\}$ includes all the possible different values for family size that he might observe.

The (discrete) distribution of X associates with each possible X-value the probability p_x that it will be the observed value. The distribution must satisfy the two conditions:

$$0 \leqslant p_x \leqslant 1; \quad \sum_x p_x = 1. \tag{1}$$

i.e., the probabilities are all nonnegative real numbers lying between zero (impossibility) and unity (certainty), and their sum must be unity because it is certain that one or other of the values will be observed. It is often convenient to assume that the number of possible outcomes (values of X) is countably infinite, i.e., there is no upper bound to X.

In discrete distribution theory the p_x, are regarded as values of a mathematical function, $p(x)$, that is usually called the probability mass function (pmf). Different functional forms lead to different discrete distributions. For example:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{2}$$

is the pmf of the Poisson distribution, while:

$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \tag{3}$$

is the pmf of the binomial distribution. λ is the parameter of the Poisson distribution; this fact is sometimes highlighted by writing $p(x; \lambda)$ or $p(x \mid \lambda)$ instead of $p(x)$ or p_x.

There are various useful and important constants associated with a distribution. They may be used to summarize various features of the distribution. The most familiar and widely used are the moments, particularly the mean and the variance. The rth (uncorrected) moment of the distribution is defined as:

$$\mu'_r = \sum_x x^x p_x. \tag{4}$$

The mean is μ'_1, the first moment of the distribution; it is conventionally denoted by μ and is used as a measure of the location of the distribution on the X-axis.

The rth moment about the mean is defined as:

$$\mu_r = \sum_x (x-\mu)^r p_x. \tag{5}$$

The variance is μ_2, the second moment about the mean; it is conventionally denoted by σ^2 and is widely used as a measure of the spread of the distribution. μ_3 is sometimes used to form a measure of the degree of asymmetry of the distribution. There is a formula for μ_r in terms of the μ'_i, and vice versa.

There are other related sets of constants, such as the cumulants, the factorial moments, the factorial cumulants, which may be more convenient to use in some circumstances. For details of these constants, and very detailed information generally about univariate discrete distributions, see Johnson et al. (1992).

In principle all the theoretical properties of the distribution can be derived from the pmf. There are, however, several other functions from which exactly the same information can be derived. This is because the functions are all one-to-one transformations of each other, so each characterizes the distribution. One particularly useful function is the probability generating function (pgf), which is defined as

$$G(s) = \sum_x s^x p_x. \tag{6}$$

For example, the Poisson distribution has pgf:

$$G(s) = \exp[\lambda(s-1)] = e^{-\lambda}\sum_x \frac{\lambda^x}{x!}s^x = \sum_x p_x s^x. \tag{7}$$

If any function that is known to be a pgf is expanded as a power series s, then the coefficient of s^x must be p_x for the corresponding distribution. An alternative way of obtaining the probabilities is by repeated differentiation of $G(s)$ with respect to s. If the pmf of a distribution is a complicated function and there are difficulties in computing numerical values directly, it may be possible to deduce from the pgf relatively simple relationships between successive probabilities and thus compute probabilities sequentially.

During the past half-century there has been a great proliferation of univariate discrete distributions. Patil et al. (1984) listed 136 univariate and multivariate distributions; Wimmer and Altmann (1999) list about 750 univariate distributions, but estimate the total number currently known at around 1000. Whilst some are rather thinly disguised versions of others, in general there are important differences between distributions. Some structure can be imposed on this vast array by grouping distributions into families whose members share certain properties.

2. *Individual Univariate Discrete Distributions*

2.1 *Bernoulli Distribution*

This is an extremely simple and basic distribution. It arises from what is known as a Bernoulli trial: a single observation is taken where the outcome is dichotomous, e.g., success or failure, alive or dead, male or female, 0 or 1. The probability of success is usually denoted by p. The probability of failure is $1-p$ and this is often denoted by q. The distribution has pmf:

$$p_x = p^x q^{1-x}, \quad 0 < p < 1, \quad q = 1-p, \quad x = 0, 1. \tag{8}$$

The mean is p, the variance is pq, and the pgf is $(q+ps)$. The Bernoulli is a one-parameter distribution (the parameter is p). Two standard examples of a Bernoulli trial are (a) the toss of a coin, and (b) the random selection of a ball from an urn containing a white and b black balls.

2.2 *Binomial Distribution*

This two-parameter distribution describes the outcome of a sequence of n independent Bernoulli trials, each with the same probability p of success. Then the probability that success is the outcome in exactly x of the trials is:

$$p(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x},$$
$$0 < p < 1, \quad q = 1-p, \quad x = 0, 1, 2, \ldots, n. \tag{9}$$

This defines the binomial distribution. In terms of the two Bernoulli trials examples above, X would be (a) the number of heads in n tosses of a coin, or (b) the number of white balls in a random sample of n balls selected sequentially, *provided* that each ball selected is returned to the urn before the next ball is selected; this is necessary in order that p remains constant and equal to a/n. This latter procedure is an example of sampling with replacement. The mean of the binomial distribution is np and the variance is npq. The pgf is $(q+ps)^n$. Note that this is the Bernoulli pgf raised to the nth power. This illustrates a very useful property of pgfs: if X and Y are two independent random variables whose distributions have pgfs $G_X(s)$ and $G_Y(s)$, then

their sum $Z = X + Y$ has a distribution with pgf $G_Z(s) = G_X(s)G_Y(s)$.

If n is large and p is small, the Poisson distribution (see Sect. 2.3) with $\lambda = np$ may be used to approximate the binomial. If n is large and p is not too close to 0 or 1, the (continuous) normal distribution may be used to construct an approximation.

The well-known chi-squared test for a 2×2 contingency table can be shown to be equivalent to a test of the difference between two binomial proportions (the proportions of 'successes' within the two classes in the table).

2.3 Poisson Distribution

This distribution occupies a central position in discrete distribution theory analogous to that occupied by the normal distribution in continuous distribution theory. It also has many practical applications. The full formal definition of the distribution is:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad 0 < \lambda, \quad x = 0, 1, 2, \ldots . \tag{10}$$

The pgf of the distribution has the very simple form:

$$G(s) = \exp[\lambda(s-1)]. \tag{11}$$

Both the mean and the variance of the Poisson distribution are equal to λ. For small values of λ the distribution is very skewed (asymmetric) but as λ increases it becomes less skewed and is nearly symmetric by $\lambda = 15$.

The Poisson distribution plays an important part in loglinear model theory for contingency tables; the model assumes independent Poisson distributions for each cell in the table (see *Multivariate Analysis: Discrete Variables (Loglinear Models)*). Another reason why the Poisson distribution is important is that it is the distribution of events that occur randomly and independently in space or time. A classic example is the number of particles, X, recorded by a Geiger counter in a fixed time interval. This property of the Poisson means that it can act as a reference standard when deviations from pure randomness are suspected.

2.4 Geometric Distribution

If X is the number of trials *before* the first success in a sequence of Bernoulli trials with constant p, then:

$$p_x = pq^x, \quad 0 < p < 1, \quad q = 1-p, \quad x = 0, 1, 2, \ldots . \tag{12}$$

The probabilities form a geometric progression, hence the name.

The mean is p/q, the variance is p/q^2, and the pgf is $p(1-qs)^{-1}$.

Just as the Bernoulli distribution is a special case of the binomial distribution, the geometric distribution is a special case of the negative binomial distribution (see Sect. 2.5).

Some authors define X to be the number of trials up to and *including* the first success; the pmf is then $p_x = pq^{x-1}$, X ranges over the positive integers, and the pgf is $ps(1-qs)^{-1}$.

2.5 Negative Binomial Distribution

The negative binomial distribution (NBD) is a widely used alternative to the Poisson distribution for handling count data when the variance is appreciably greater than the mean (this condition is known as overdispersion and is frequently met in practice).

The pmf of the NBD is:

$$p_x = \binom{k+x-1}{x} p^k q^x,$$

$$0 < k, \ 0 < p < 1, \ q = 1-p, \ x = 0, 1, 2, \ldots . \tag{13}$$

The mean is kq/p, the variance is kq/p^2, and the pgf is $p(1-qs)^{-k}$.

In the literature, several different parameterizations of the distribution are found. A common one uses $Q = 1/p$ and $P = q/p$. It follows that $P > 0$, $Q = 1 + P$, and:

$$p_x = \binom{k+x-1}{x} \left(1 - \frac{P}{Q}\right)^k \left(\frac{P}{Q}\right)^x. \tag{14}$$

When $k = 1$, the NBD becomes the geometric distribution. When k is a positive integer, the NBD is sometimes known as the Pascal distribution; it can then be interpreted as the distribution of the number of failures before the kth success (i.e., X is the sum of k independent geometric random variables). This kind of model is known as a waiting time distribution.

There are several other models that lead to the NBD. A classic example arises from the theory of accident proneness which was developed during and after World War I. It was assumed that the number of accidents suffered by an Individual was distributed as a Poisson distribution, but that the Poisson parameter λ (interpreted as the individual's accident proneness) varied between individuals in the population under study. If λ is assumed to follow the (continuous) gamma distribution, then the NBD is the resultant overall distribution of accidents per indi-

vidual. This model has been applied to retail purchasing, absenteeism, doctor's consultations, amongst many others. The model reflects a type of heterogeneity in the population. A different type of heterogeneity occurs when there is clustering. If it is assumed that the number of clusters is distributed as a Poisson, but the number of individuals in a cluster is distributed according to the logarithmic distribution (see Sect. 2.6), then the overall distribution is NBD.

The existence of these and other models means that the NBD is used to model many different types of data. The penalty for this ubiquity is the difficulty in inferring which underlying model applies.

2.6 Distributions Without a Zero Class

In some applications the meaningful outcomes do not include $X = 0$. Examples are the size of a family and the size of a cluster of dwellings. Such cases require distributions that start at $X = 1$. One way of achieving this is to take a distribution that is defined for $x = 0, 1, 2, \ldots$ and truncate the term $X = 0$. In order to form a valid distribution from the remaining probabilities, it is necessary to adjust them upwards by a factor of $(1-p_0)^{-1}$ so that the second condition in (1) is satisfied. For example, the zero-truncated Poisson distribution has pmf:

$$p_x = \frac{e^{-\lambda}\lambda^x}{x!(1-e^{-\lambda})}, \quad 0 < \lambda, \quad x = 1, 2, 3, \ldots . \tag{15}$$

Alternatively, there are some distributions for which the pmf is not defined for $X = 0$. One of these is the logarithmic distribution, with pmf:

$$p_x = \frac{\theta^x}{-x\log_e(1-\theta)}, \quad 0 < \theta < 1, \quad x = 1, 2, 3, \ldots . \tag{16}$$

The mean is $a\theta/(1-\theta)$ and the variance is $a\theta(1-a\theta)/(1-\theta)^2$, where $a = -1/\log_e(1-\theta)$. The pgf is:

$$G(s) = \frac{\log_e(1-\theta s)}{\log_e(1-\theta)}. \tag{17}$$

This distribution can model data distributions with very long tails.

2.7 Hypergeometric Distribution

A key feature of the binomial distribution is its representation in terms of a sequence of independent Bernoulli trials with fixed p. In many practical situations, however, p is not constant due to dependence between successive Bernoulli trials. In the derivation of the binomial using the ball in an urn model (b), each selected ball was returned immediately to the urn (sampling with replacement) so p was always equal to $a/(a+b)$. However, if each selected ball is retained and not returned, then p changes after every trial; this is sampling without replacement (see *Sample Surveys: Methods*). The resultant distribution is hypergeometric with pmf:

$$\begin{aligned} p_x &= \binom{a}{x}\binom{b}{n-x}\Big/\binom{a+b}{n} \\ &= \binom{n}{x}\binom{a+b-n}{a-x}\Big/\binom{a+b}{a}, \\ &\quad x = 0, 1, 2, \ldots, \min(n, a). \end{aligned} \tag{18}$$

The mean is $na/(a+b)$ and the variance is $nab(a+b-n)/(a+b)^2(a+b-1)$. The pfg is:

$$G(s) = \frac{{}_2F_1[-n, -a; b-n+1; s]}{{}_2F_1[-n, -a; b-n+1; 1]} \tag{19}$$

i.e., the ratio of two classical Gaussian hypergeometric functions.

An important use of the hypergeometric distribution is in the analysis of 2×2 contingency tables; the probability of a result as extreme as the observed result is given by the tail of a hypergeometric distribution. This is known as Fisher's exact test; it is frequently recommended for use instead of the chi-squared test when the smallest expected frequency is less than 5.

2.8 Beta-binomial (Negative Hypergeometric) Distribution

In Sect. 2.5 heterogeneity of the Poisson parameter λ was allowed for by assuming that λ had a gamma distribution. The resultant Poisson–gamma distribution was in fact the NBD. Analogously, when there is heterogeneity in the binomial parameter p one way of allowing for it is to assume that it follows the (continuous) beta-distribution—a flexible distribution on the interval (0, 1). The resultant beta-binomial distribution has pmf:

$$\begin{aligned} p_x &= \binom{-a}{x}\binom{-b}{n-x}\Big/\binom{-a-b}{n}, \\ &\quad 0 < a, \quad 0 < b, \quad n \text{ an intger}, \\ &= \binom{a+x-1}{x}\binom{b+n-x-1}{n-x}\Big/\binom{a+b+n-1}{n}, \\ &\quad x = 0, 1, 2, \ldots, n. \end{aligned} \tag{20}$$

The mean is $na/(a+b)$, the variance is $nab(a+b+n)/(a+b)^2(a+b+1)$, and the pgf is:

$$G(s) = \frac{{}_2F_1[-n, a; -b-n+1; s]}{{}_2F_1[-n, a; -b-n+1; 1]}. \quad (21)$$

The distribution is also known as the negative hypergeometric or inverse hypergeometric distribution. Like the NBD, it can be interpreted as a waiting time distribution.

3. Combinations of Distributions

New distributions may be created from existing ones. Two major methods have been mentioned in Sects. 2.5 and 2.8. The process of allowing the parameter of one distribution to itself be distributed (as in the derivation of the NBD by letting the Poisson parameter λ be distributed as a gamma variable) is known as mixing (in earlier literature the term compounding was commonly used). Many other mixture distributions appear in the literature.

The process exemplified by the NBD cluster model of Poissonly distributed clusters with logarithmically distributed cluster sizes has been given several names including compounding and generalizing, both of which are ambiguous; a clearer terminology that is frequently used nowadays is to call the resultant distribution a randomly stopped-sum or stopped-sum distribution. If the pgf of the distribution of clusters is $g(s)$ and that of the cluster size distribution is $h(s)$ then the pgf of the resultant distribution has the very simple structure:

$$G(s) = g(h(s)). \quad (22)$$

Various properties of the stopped-sum distribution can be deduced via this relationship.

These processes are one reason why there are so many different discrete distributions in the literature. They often arise from attempts to model processes in applications. Clearly some distributions like the NBD that are mixtures are also stopped sums, but this is not always the case. As mentioned previously, the existence of fundamentally different ways of deriving the same distribution can raise serious problems in practice when trying to make causal inferences from data.

4. Multivariate Discrete Distributions

The move from distributions of one variable to joint distributions of several variables is complicated by the existence of relationships (dependencies) between the variables. This is reflected in the complexity of the structure of the multivariate pmf and related functions. Furthermore, there may be more than one way of representing dependency; more than one bivariate version of a particular univariate distribution may be constructed. Often the multivariate pgf is one of the simpler ways of defining and characterizing a distribution, A bivariate pgf is defined as:

$$G(s_1, s_2) = \sum_{x,y} p_{x,y} s_1^x s_2^y; \quad (23)$$

the definition of a multivariate pgf is a straightforward extension.

The multinomial distribution has pgf and pmf:

$$G(s_1, s_2, \ldots, s_r) = (p_1 s_1 + p_2 s_2 + \cdots + p_r s_r)^n, \quad (24)$$

$$p(x_1, x_2, \ldots, x_r) = \frac{n!}{x_1!, x_2!, \ldots, x_r!} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}. \quad (25)$$

When $r = 2$ it becomes the binomial distribution. All the marginal distributions of the multinomial are binomials. The conditional distribution of $(X_2, X_3, \ldots X_r)$ given X_1 is also multinomial.

One of the first reference texts to be devoted to the multivariate area was Kocherlakota and Kocherlakota (1992). They set out the general theory of bivariate discrete distributions and considered in detail the properties of bivariate versions of the major univariate distributions. Johnson et al. (1997) give a very thorough treatment of the current position in multivariate discrete distributions; the general theory is given and a wide range of distributions is examined.

Johnson et al. (1992) is the standard reference text for univariate discrete distributions; all the univariate distributions above are considered in detail, together with very many more.

Bibliography

Johnson N L, Kotz S, Kemp A W 1992 *Univariate Discrete Distributions*, 2nd edn. Wiley, New York

Johnson N L, Kotz S, Balakrishnan N 1997 *Discrete Multivariate Distributions*. Wiley, New York

Kocherlakota S, Kocherlakota K 1992 *Bivariate Discrete Distributions*. Dekker, New York

Patil G P, Boswell M T, Joshi S W, Ratnaparkhi M V 1984 *Dictionary and Classified Bibliography of Statistical Distributions in Scientific Work, Vol 1: Discrete Models*. International Co-operative Publishing, Fairland, MD

Wimmer G, Altmann G 1999 *Thesaurus of Univariate Discrete Probability Distributions*. Stamm, Essen, Germany

C. D. Kemp

Diversification and Economies of Scope

1. Introduction

Diversification refers to the process by which single product, specialist firms become multiproduct firms.

The industrial landscape in advanced industrial countries is characterized by the prevalence of firms that maintain diversified product lines. However, diversification is not a random process. On the contrary, firms tend to diversify by adding product lines that are 'close in' to their established product lines. Put differently, diversified firms diversify in ways that are coherent. Those that don't tend not to survive. This article focuses on explaining why and how firms diversify, with an eye toward explicating the successful modes of corporate diversification.

Three related research fields have much to say on corporate diversification. The 'resource-based' view of the firm draws attention to the quasi-fungible nature of the firm's assets, and postulates that diversification is a way to use excess capacity. Transaction cost economics helps explain whether and when excess capacity is better used internally (diversification) or contracted out to parties seeking to outsource services and products. Evolutionary economics helps explain how excess capacity comes about. These three intellectual wellsprings help explain the nature, timing, and pattern of observed diversification in advanced industrial nations.

It must be pointed out, however, that there is a largely discredited literature that offers an alternative view. Theories of market strength and monopoly position stress these attributes as drivers of diversification (Montgomery 1994). Studies in this genre have been actively explored empirically, but offer little to inform the diversification process. Measures such as Tobin's Q provide a framework to compare how firm (stock) market evaluation changes with added products and markets. Such measures are useful but do little to explain why firms diversify and the relationship between direction of diversification and organizational factors, such as legal structure. This line of research has provided excellent measurement and context with which to compare diversification outcomes. However, it does not illustrate satisfactorily the decision to diversify or the real constraints on diversification that managers and firms face.

We now turn to two fundamental questions: why do firms diversify, and how do they organize for diversification. The goal in the next section is to communicate the logic and strategy of diversification, motivating the subsequent discussion on organizing for diversification. We do not mean to describe these two activities as serial, but rather interactive. The decision to diversify for profit seeking firms requires them to analyze how they can secure extra profits from the process. We can conclude that successful diversification is the combination of managers making strategic choices as to lines of business and developing appropriate organizational structures to support the resulting diversification. This conclusion has implications for empirical testing, and several scholars have recently begun to undertake an effort to integrate theories of diversification (Silverman 1999).

2. *Why Firms Diversify: Corporate Diversification and the Resource-based View of the Firm*

Although the resource-based view has, since the 1980s, been applied to several areas in firm strategy, the theory has its roots in the study of firm diversification (Teece 1982). We begin by summarizing the major points, and describe the decision to add product lines as a response to the availability of underutilized assets created by prior activity. We then discuss the strategy of diversification as the choice of those areas in which to diversify that maximize the value of the unused assets. This typically results in what has been referred to as corporate coherence (Teece et al. 1994).

Note that corporate diversification issues, as discussed herein, is a superset of multiproduct firm issues. Much of our current attention focuses on multiproduct firms, with less attention on other diversification modes (new customer or geographic markets, for example). However, firm diversification in a broader context includes many of the same analyses that are applied below. That is, diversifying international markets follows much of the same logic and reasoning as adding product lines.

2.1 *The Firm as a Set of Resources*

The resource-based view seeks to understand why firms grow and diversify. The theory grew largely out of Penrose's (1959) study, in which she cites unused managerial resources as the primary driver of growth. Penrose recognized that internal managerial resources are both drivers and limits to the expansion any one firm can undertake. This stream of literature was expanded in the 1970s and early 1980s on the heels of significant diversification and firm expansion (Rubin 1973, Teece 1980, 1982).

The resource-based view advances the importance of firm-specific resources, that is, those resources that maintain value *in the context* of the given firm's markets and other resources that are difficult to replicate by other firms (Wernerfelt 1984). Such resources include managerial ability, customer relationships, brand reputation, and tacit knowledge regarding specific manufacturing process. Resources are not the same as competencies or capabilities. Rather, a firm's access to resources and ability to mobilize and combine these resources in specific ways determine the firm's competence in a given product area. As a firm gathers resources for one business, these resources will, to differing extents, be sufficiently fungible for use in other product lines or markets (Teece 1982). Some of these resources will maintain excess capacities over time, especially since the units required for operations in one area are not necessarily consistent across multiple fields.

In some cases it is knowledge which is the 'excess resource' (Teece 1982, 1986). It is this excess capacity, coupled with profit-seeking behavior, that drives decisions to diversify. Some scholars have noted that such resources are transferable in a limited sense between firms, but such endeavors require organizational interaction beyond market contracting to include acquisitions or hybrid forms of governance (Chi 1994). This point will be considered in Sect. 4.

2.2 The Logic and Coherence of Diversification

The literature on diversification includes analysis of whether specific diversification patterns are associated with improved firm performance (Chatterjee and Wernerfelt 1991). Specifically, scholars have sought to discern whether there exist performance differences between related and unrelated diversification. Many efforts have focused considerably on outcome measures that do little to inform the rationale behind diversifying. The resource-based foundation provides scholars with a motivation for examining the linkages (or 'coherence') between classes of resources within a given firm (Teece et al. 1994). Fungible resources allow for firms to apply managerial practices and reputations to other industries, as discussed in the previous section. However, these assets have varying degrees of value across industries. The resources of firms even within the same industry may not be identical. Moreover, the degree of coherence that one expects to observe among the parts of an organization at a particular point in time depends on the relationship between learning, path dependencies, the technological opportunities available to the firm, and the firm's existing complimentary assets. This suggests that seemingly homogeneous (in terms of primary product outputs) firms may exhibit radically different diversification strategies.

While empirical evidence will be examined in more detail below, existing research suggests that there exists a *logic to diversification*. As firms grow more diverse, they maintain a constant level of coherence between neighboring activities. For example, the sales, distribution, and market understanding maintained in pharmaceutical industries is relevant to a specific set of other industries, such as medical equipment and over-the-counter medicines. The latter is in turn related to candy and confectionery and other drug store items. Note that this result contrasts with the purely market power studies discussed earlier in that the focus of analysis is on firm specific assets as an input to the diversification process, rather than solely on the industries in which a firm ends up.

We now turn to the consideration of how firms organize to diversify, and consider the implications for a combined analysis of diversification strategy and underlying organizational structures.

3. How Firms Organize to Diversify: Corporate Diversification and Transaction Cost Economics

Transaction cost economics argues that economic agents organize activities so as to economize on the costs of transacting. Specifically, the probability of observing a given form of organization is a function of various features of the transaction under consideration and the influence of these features on the relative costs of alternative modes of organizing. While transaction cost theory is most noted for its use in analyzing vertical integration decisions (i.e., make-vs.-buy decisions), its logic has also been applied to issues related to diversification (Teece 1980, 1982).

Early work in this area focused on the so-called 'conglomerate' form of organization. Williamson (1975) argues that large, diversified conglomerates are best understood as a logical outgrowth of the multidivisional form (M-Form) mode for organizing complex economic transactions. Building on work by Chandler (1962), Williamson suggests that the M-form offers a number of advantages in managing separable, related lines of business, such as a group of automobile divisions. Specifically, the M-form allows for improved transaction governance in directing and monitoring the use of assets, where a unified or function-based organization would falter under information asymmetries, bounded rationality, and managerial opportunism. Recognizing these merits, he goes on to argue that the M-form may be similarly advantaged when it comes to managing less closely-related activities (Williamson 1988). While he is careful to note that the management of diverse products is not without problems, Williamson (1988) nevertheless proposes that the advantages of the M-Form of organization carry over to the case of diverse product lines. Thus, early research in organizational economics suggested that it is more efficient to administer diversification in a multidivisional corporation than in other, more centralized forms of internal organization.

More recently, the question of how to most efficiently organize for diversification has been extended to include a direct comparison of specific alternative modes of diversification. Bethel and Liebeskind (1998), for instance, employ transaction cost reasoning to account for the decision of firms to diversify by means of a 'simple corporation,' in which a single legal entity is responsible for all contracting relationships, vs. a 'corporate group' approach, in which a parent corporation owns or partially owns a series of subsidiary corporations. Firms select a particular mode of diversification to minimize a specific type of transaction cost; namely 'protection costs.' Because diversification has a differential effect on the various stakeholders within a firm, disadvantaged stakeholders may take action to protect themselves either by seeking to prevent diversification or by seeking compensation after diversification has taken place. A 'corporate group' form of organization can reduce the

costs associated with these protection measures by reducing the losses from diversification that some stakeholders would otherwise bear. However, the use of 'corporate groups' to manage diversification is not costless. Although 'corporate groups' may reduce or eliminate the negative effects of diversification, they carry with them costs: economies of scope may be lower within a 'corporate group' than within a 'simple corporation.' Managers must explicitly trade off the advantages of economies of scope with the disadvantages associated with protection efforts by corporate stakeholders.

The research in transaction cost theory related to diversification thus attempts to account of how diversification is organized by exploring the nature of the diversification taking place, as well as its impact on the relative costs of different modes of diversification. The theory predicts that the mode of organization (such as vertical integration or joint venture) chosen in a given instance will be that which minimizes the total costs of transacting. By exploring how it is that firms diversify, transaction cost theory complements the work coming out of the resource-based view. Whereas the resource-based view analyzes the direction of corporate diversification, transaction theory allows one to predict which organizational structures are necessary to support diversification efforts.

4. *Integrating Two Views of Diversification*

The above discussion highlights the theoretical similarities and complementarities between the resource-based view and transaction cost theory. Indeed, asset specificity, a key analytical frame in traditional transaction cost economics theory, is analogous to the firm-specific resources that are the engines of diversification in the resource-based view. Recently, researchers have begun to explicitly combine these two approaches.

Among these efforts is Silverman's (1999) examination of how contractual hazards affect the likelihood of corporate diversification. As was noted in Sect. 2, the resource-based view draws heavily on Penrose's view that firm growth is driven by the internal exploitation of excess resources. A result of Penrose's approach is that firms always enter new markets, but never divest. In contrast, transaction cost theory suggests that there are multiple ways in which excess assets may be utilized, including contracting these assets to outside parties and spinning them off into stand-alone enterprises (Teece 1980). The added incentive intensity associated with exploiting excess resources through markets rather than internally provides a rationale for utilizing various contracting approaches. The relevant empirical question with respect to diversification is: under what circumstances are excess resources more suitably exploited through internal diversification than through outside contracting?

Drawing on work by Teece (1986), Silverman suggests that the contractual exploitation of technological knowledge resources (e.g., licensing) is a feasible alternative to diversification except for cases where the knowledge is highly tacit or easily transferred but weakly protected. In the former case, licensing contracts are difficult to write and enforce due to the tacit nature of the resource being contracted for. In the latter case, attempts to negotiate a licensing contract are plagued by problems of appropriating the returns to technology. Silverman therefore hypothesizes that a firm is more likely to diversify into a business when contracting out its technological resources in that business is subject to high contractual hazards. His analysis of diversification decisions for firms entering new SICs between 1982 and 1985 indicates that in fact a firm is less likely to diversify into an industry when viable contractual alternatives exist to exploit its technological resources.

5. *Conclusion: The Future of Diversification Research*

The efforts by Silverman and the others cited in this article to exploit and, where possible, combine the resource-based and transaction cost approaches to diversification are useful to help elucidate the critical aspects of each theoretical approach with respect to corporate diversification. While important first steps, the research examined here by no means provides definitive resolution to the topic of firm diversification. In particular, a number of theoretical and empirical issues remain. On the theoretical side, efforts to recognize the dynamic nature of corporate diversification decisions are needed. The treatment of diversification in this chapter and in much of the existing literature treats the why and how of diversification as two distinct (and serial) activities. In practice, successful diversification is a combination of *both* choosing where to diversify and how to support a diversification strategy.

While resource-based scholars have begun developing a theoretical understanding of corporate coherence, there exist few empirical tests of these predictions. Resource-based explanations for diversification have struggled with how to operationalize the resources that are relevant in diversification decisions. That is, the lack of adequate measures for critical firm resources has limited empirical testing of the theory's predictions with respect to diversification. Finally, greater effort to explicate and test the transaction level factors that are relevant in driving how diversification efforts are organized is needed.

See also: Entrepreneurship; Firm Behavior

Bibliography

Bethel J E, Liebeskind J P 1998 Diversification and the legal organization of the firm. *Organization Science* **9**(1): 49–67

Chandler A D 1962 *Strategy and Structure: Chapters in the History of the Industrial Enterpise*. MIT Press, Cambridge, MA

Chatterjee S, Wernerfelt B 1991 The link between resources and type of diversification: Theory and evidence. *Strategic Management Journal* **12**(1): 33–48

Chi T 1994 Trading in strategic resources: Necessary conditions, transactions cost problems, and choice of exchange structure. *Strategic Management Journal* **15**(4): 271–90

Montgomery C A 1994 Corporate diversification. *Journal of Economic Perspectives* **8**(3): 163–78

Penrose E 1959 *The Theory Of The Growth Of The Firm*. Wiley, New York

Rubin P H 1973 The expansion of firms. *Journal of Political Economy* **81**(4): 936–49

Silverman B S 1999 Technological resources and the direction of corporate diversification: Toward an integration of the resource-based view and transaction cost economics. *Management Science* **45**(8): 1109–24

Teece D J 1980 Economies of scope and the scope of the enterprise. *Journal of Economic Behavior and Organization* **1**(3): 223–47

Teece D J 1982 Towards an economic theory of the multiproduct firm. *Journal of Economic Behavior and Organization* **3**: 39–63

Teece D J 1986 Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy* **15**(6): 285–305

Teece D J, Rumelt R, Dosi G, Winter S 1994 Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior and Organization* **23**(1): 1–30

Wernerfelt B 1984 A resource-based view of the firm. *Strategic Management Journal* **5**(2): 171–80

Williamson O E 1975 *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York

Williamson O E 1988 Corporate finance and corporate governance. *Journal of Finance* **43**(3): 567–91

R. A. Lowe, C. S. Boerner, and D. J. Teece

Diversity and Disagreement

The extraordinary variety of ethical belief and practice that is possible within and across human societies, along with the ethical disagreement to which that variety frequently gives rise, has been a rich source of philosophical reflection. Philosophers have sought to understand the nature and causes of ethical diversity and disagreement, as well as what that diversity and disagreement might reveal about the nature of ethical argument and inquiry. They have investigated the similarities and differences between ethical inquiry and other kinds of inquiry, particularly scientific inquiry, and asked what the relevant similarities and differences can tell us about the nature of rationality. They have also analyzed the implications of ethical diversity and disagreement for ethical practice, and for the foundations of social and political life.

1. *The Nature and Sources of Ethical Diversity and Disagreement*

1.1 *Philosophical Understandings of Anthropological and Sociological Evidence*

Philosophical awareness of ethical diversity is as old as philosophy itself. It is evident in the ancient Greek beginnings of Western philosophy—particularly in the ethical and political thought of Plato and Aristotle. In the early and middle *Dialogues* of Plato, for instance, exchanges between the Platonic Socrates and his interlocutors reveal many of the ethical conflicts and debates, and hence the ethical diversity and disagreement, that helped to shape the social and political life of fourth-century BC Greece. A generation later, Aristotle's sensitivity to the concrete contexts of deliberation and action, detailed in the *Nichomachean Ethics* and the *Politics*, reflected an even richer understanding of social and political diversity, and of the resultant complexity of ethical belief and practice. Neither Plato nor Aristotle believed that ethical diversity and disagreement required us to relinquish a commitment to ethical objectivity. But this stance was in marked contrast to that adopted by some of their philosophical forebears, especially fifth-century BC sophists such as Protagoras and Gorgias, to whom Plato's philosophy was in part a response. Moreover, the objectivist conceptions defended by Plato and Aristotle were met by the emergence of a powerful current of skepticism. This development culminated in the Academic skepticism of Cicero, and especially the second-century AD Pyrrhonian skepticism of Sextus Empiricus, in which ethical diversity and disagreement inevitably raised the specter of ethical relativism.

The sixteenth-century rediscovery and dissemination of Sextus' writings coincided with the beginnings, in the West, of a period of extraordinary exploration and economic expansion. Both developments were important in shaping the rise of early modern Western philosophy in the late sixteenth and early seventeenth centuries. European exploration and expansion produced a body of literature purporting to offer a richer picture of human ethical variety than anything to be found in ancient philosophical texts. These accounts had a profound effect on early modern thinkers who pondered the nature and sources of ethical variety. Thinkers such as Montaigne came to accept the indispensability to philosophy of what we would now call sociological or anthropological explanations of ethical diversity. But, deeply influenced

by the Pyrrhonian skepticism he found in Sextus, Montaigne also insisted that fully appreciating the details of human ethical diversity would always lead a reasonable person to accept ethical relativism. Montaigne thus rejected the claims of philosophers, from Plato onward, who believed that one could render the existence of ethical diversity and disagreement compatible with ethical objectivism.

In the seventeenth, eighteenth, and early nineteenth centuries, debates about the implications of ethical diversity and disagreement—based primarily on the testimony of explorers, settlers, and economic adventurers—continued to shape philosophical reflection about ethics. That reflection took one of its most interesting forms in Montesquieu's *Persian Letters* (1973), in which Montesquieu's potentially dangerous criticisms of ethical and political life in eighteenth-century France are presented to the reader as the contents of letters written home by two fictional Persian visitors. But with the rise of modern anthropology in the late nineteenth and early twentieth centuries, philosophical reflection about diversity and disagreement took on a new tone. Philosophers were gradually persuaded by the claims of anthropology to provide expert testimony—rooted in theoretically sophisticated empirical research, rather than idiosyncratic personal or economic interests—about the nature and extent of diversity and disagreement across cultures and social groups.

Yet the cultural ascendancy of anthropological claims about diversity also reinvigorated the skeptical challenge to ethical objectivism. Ruth Benedict's *Patterns of Culture* (1934) provides an instructive and influential expression of this development. Benedict claimed to have given ethical relativism a scientific foundation for the first time. Some philosophers were prepared to accept this account of the matter and to cede authority for assessing the claims of relativism to empirical social science. But, Benedict's claims of a newly 'scientific foundation' for relativism may be difficult to sustain. Her formulation and defense of relativism seem simply to reiterate Montaigne's argument that we have 'no other test of truth and reason than the example and pattern of the opinion and customs of the country we live in.' Melville Herskovits (1972) went on to argue that modern anthropology finally provided 'tough-minded' empirical support for claims about the 'ethnocentrism' of objectivist analyses of ethical disagreement. But Herskovits's efforts to defend charges of ethnocentrism do not go much beyond Montaigne's contention that 'each man calls barbarism whatever is not his own practice.' The fact that so many familiar defenses of relativism predate the rise of modern anthropology has led some philosophers to wonder whether ethical relativism is simply a reasonable inference from ethical diversity and disagreement, or a pretheoretical assumption which determines the way some theorists characterize the data of diversity.

Yet other philosophers have undertaken detailed investigation to show the value and importance of anthropological study to philosophical reflection about ethics. Some have conducted their own ethnographic studies of the ethical beliefs and practices of Native American groups, in an effort to determine the empirical reliability of anthropological accounts (Brandt 1959, Ladd 1957). Others—spurred on, in part, by the late work of Ludwig von Wittgenstein, especially by the concept of 'forms of life' in his *Philosophical Investigations* (1978)—have interrogated the theoretical underpinnings, as well as the broader philosophical implications, of the social scientist's claims about diversity and disagreement (Wilson 1970, Winch 1972). These efforts introduced important new empirical detail into philosophical discussions of the extent of cultural diversity—details drawn, most famously, from the work of E. E. Evans-Pritchard on the Azande, and from the work of Edmund Sapir and Benjamin Whorf on the conceptual variety embedded in human languages. The resulting debates raise new and controversial philosophical questions about the nature of ethical disagreement—particularly the question whether human ethical systems could ever contain wholly incommensurable ethical (or other) concepts.

Philosophical attempts to understand the origins and causes of ethical diversity and disagreement have frequently relied on generally utilitarian or consequentialist accounts of that diversity. On these accounts, ethical practices (and their associated beliefs) come into existence, and continue in existence, primarily because they serve the aims of the groups that develop them. From the eighteenth century forward, such accounts are usually linked with essentially evolutionary hypotheses about the development of ethical practices. The ethical and political thought of David Hume, for instance, attributes the rise of particular ethical practices to unconscious, or unintended, processes that are vaguely akin to natural selection. But accounts of the processes in question usually have more in common with Herbert Spencer's theory of social evolution than with Darwin's claims about the role of natural selection in biological evolution. William Graham Sumner's *Folkways* (1909)—which combined evolutionary ethics with a forceful defense of ethical relativism—contains what became one of the most influential expositions of just such a view.

1.2 Methodological Concerns about the Idea of Fundamental Disagreement

During the mid-twentieth century, many philosophers accepted the claims of relativist anthropologists that the data of comparative anthropology could firmly ground an argument for metaethical relativism: the view that there is no rational method for adjudicating between the conflicting ethical claims of different cultures. Yet, as many of their critics objected, the

mere facts of ethical diversity and disagreement within and across cultures cannot be sufficient to establish the truth of ethical relativism (in any form). The objection is that merely from the fact of cultural variability we cannot confidently conclude that any resultant ethical disagreements will necessarily be disagreements in fundamental ethical commitments between which no rational adjudication is possible; and establishing this conclusion is essential to establishing the plausibility of ethical relativism.

This objection points to a way in which ethical relativism may, in fact, be fundamentally underdetermined by all available empirical evidence. For it is at least possible, as the psychologist Karl Duncker argued, that principles of ethical valuation might remain 'invariant' across the most diverse cultures, even while the 'situational meanings' of the acts, traits, and states of affairs subject to ethical evaluation could vary dramatically across those same cultures (Duncker 1939). In short, cross-cultural ethical diversity might be properly traced not to differences in fundamental ethical principles but to cultural variation in derivative ethical principles—that is, to principles embodying local understandings of how to apply fundamental ethical principles to particular actions, character traits, and states of affairs. Thinkers who claim to identify fundamental ethical disagreement must, therefore, be able to show that they have identified a conflict between genuinely fundamental ethical principles or beliefs, and not between derivative principles or beliefs intended to apply the more fundamental principles or beliefs. Further, as the philosopher R. B. Brandt once noted, it is very difficult to find two cultures, or even two individual people, who can be shown attribute identical non-moral properties to some action, trait, or state of affairs and who nonetheless give the relevant action, trait, or state of affairs different ethical valuations (Brandt 1959, 1967).

Identifying these difficulties does not prove that it is impossible to find convincing evidence of genuinely fundamental ethical disagreement. Brandt, himself, claimed to have found at least one such disagreement, in spite of the methodological difficulties he believed to attend any effort to do so (Brandt 1959). Moreover, the demand for conceptual clarity in defining a disagreement as 'fundamental,' rather than 'derivative,' is difficult to meet, since distinguishing fundamental from derivative ethical beliefs and principles is no simple matter. In particular, it is not clear how to find a criterion for describing a belief or principle as fundamental or derivative that doesn't presuppose a commitment to some fairly complex elements of an ethical conception—that is, to ethical notions which may themselves be the subject of considerable controversy and disagreement. Yet in coming to appreciate the methodological and conceptual difficulties inherent in attempts to establish fundamental ethical disagreement, the empirical researcher and the philosophical interpreter alike are reminded that there are important obstacles to surmount if one seeks to draw a supportable inference for meta-ethical relativism from the fact of cross-cultural disagreement.

1.3 Individual Uncertainty and the Roots of Ethical Pluralism

Reliance on the evidence of cross-cultural variability is not the only means—or even the most important means—by which philosophers have sought to understand the nature and sources of ethical diversity and disagreement. According to some philosophers, the decision-making processes of an individual agent experiencing ethical uncertainty can sometimes tell us as much (if not more) about the nature and sources of ethical diversity and disagreement than the comparative anthropologist's claims about cross-cultural conflict. On these accounts, when a sincere and reflective agent attempts to choose between equally compelling, but conflicting and even apparently irreconcilable, ethical ends, what that agent typically confronts is the fundamental plurality and diversity (and, on some views, sometimes even the ultimate incommensurability) of available and plausible ethical ends. Pluralists such as Isaiah Berlin and Charles Taylor believe that philosophers like Plato, along with those of his intellectual descendants who would defend the unity of ethical value, simply fail to understand that plurality and diversity are intrinsic features of ethical value. According to the ethical pluralist, ethically compelling ends and purposes cannot all be reduced to a single measure of ethical value; nor can they all be realized in a single life. This means that even the sincere and reflective agent (or, perhaps, especially the sincere and reflective agent) will eventually confront what Thomas Nagel has called the fragmentation of ethical value (Nagel 1979).

Those who understand ethical value as intrinsically plural and diverse thus have a way of explaining diversity in the beliefs and practices of various cultural and social groups in terms that go beyond any available on the standard anthropological account. On a pluralist conception of ethics, cross-cultural ethical diversity is simply an important—and essentially unavoidable—reflection of the fundamental plurality and diversity of ethical value itself. More controversially, some ethical pluralists go on to assert that neither individual human beings, nor cultures and societies, can escape fundamental ethical conflict—conflict, that is, between irreconcilable or incommensurable fundamental ethical principles and concepts. But some of the same difficulties inherent in the effort to defend the thesis of fundamental conflict across cultural groups, particularly the challenge of defining the contrast between derivative and fundamental ethical principles and beliefs, will attend any effort to show that no individual agent can avoid fundamental ethical conflict.

Following the lead of Isaiah Berlin, some pluralists contend that any individual agent who claims to experience no fundamental ethical conflict is in some way rationally deficient, or else deeply self-deceived. Yet there are many conceptions of human flourishing (both secular and religious) which purport to make fundamental ethical conflicts effectively impossible by identifying or articulating a single 'dominant end'—as John Rawls has described it—to which all other human ends must be subordinate, and with respect to which the subordinate ends can be unified. Such conceptions, particularly when combined with a willingness to allow other agents to reject them, have often been thought to embody defensible understandings of rational choice and rational agency. Moreover, some ardent defenders of the view that dominant-end conceptions of human flourishing are in some way deficient or irrational actually acknowledge that these conceptions do not intrinsically violate principles of rational choice. But they nonetheless insist that there remains something ultimately incomplete, or perhaps even irrational, about a life lived in accordance with such a conception (Rawls 1971, Hampshire 1983). In response to such assertions, the question of whether a compelling defense of ethical pluralism requires a commitment to the inescapability of fundamentally irresolvable ethical conflicts—across social and cultural groups, or in individual lives, or both—remains a subject of lively philosophical debate (Gowans 1987, Stocker 1990, Moody-Adams 1997).

2. *Implications of Disagreement for Understanding Ethical Argument and Inquiry*

Historically, philosophers have held quite differing views about what the facts of ethical diversity and disagreement might tell us about the nature of ethical argument and inquiry. The standard views can be divided up in terms of the position they take with regard to the objectivity of ethics—that is, with regard to the question whether there can be at least some right answers to ethical questions, where the rightness of the answer is independent of human beliefs and practices. But there is great variety within both the objectivist and the nonobjectivist positions, particularly with regard to important epistemological questions (questions of whether and how we can have moral knowledge) and serious metaphysical concerns (concerns about the existence or non-existence of objective value entities).

2.1 *Varieties of Ethical Objectivism and Serious Ethical Disagreement*

The most familiar variety of ethical objectivism is *absolutist ethical objectivism*, on which there is at most one right answer to any given ethical question, and according to which the ethical correctness of that answer does not depend upon human beliefs and practices. Plato's *Republic* contains one of the most fully developed and widely influential versions of this conception. The 'Allegory of the Cave,' in Book Five of the *Republic*, richly depicts the underlying epistemological and metaphysical assumptions. According to the allegory, the inhabitants of a cave who remained chained to a wall, seeing only the shadows of real things by the light of a humanly built fire, are deprived of any possibility of rationally understanding what is real, and their beliefs and practices reflect their ignorance. Only the person who is released from ignorance and illusion by being unchained, and allowed to leave the cave, can become capable of a rational appreciation of the real essences of things as illuminated by the genuine light of the sun. The metaphysical commitments associated with this view—in Plato's doctrine of the Forms—are developed in surrounding passages in the *Republic*, and in other Dialogues like the *Symposium* and the *Phaedrus*. On the resulting view, what the relativist takes to be fundamental ethical diversity and disagreement is simply a function of moral ignorance rooted in the tendency to our experience of imperfect copies and shadows of the real as though they might reveal the essence of the ethical.

But absolutist ethical objectivism need not involve this kind of metaphysical realism about value entities existing independent of human belief and practice, or even any kind of realism about independently existing value-entities or antecedently given moral facts. Nor need it involve the potentially problematic epistemological assumptions needed to explain how anyone might come to know the Form of the Good, and why so few seem genuinely capable of doing so. Immanuel Kant held a constructivist conception of ethical objectivity, on which humanly constructed practices (and their underlying beliefs and assumptions) are objective when they have been constructed in accordance with what Kant took to be the objectively valid (and already known) supreme principle of morality, the 'Categorical Imperative.' In its most basic formulation, the Categorical Imperative directs the agent to act only on those policies (reasons) that he or she could will to serve as universal laws of action. What the relativist identifies as 'fundamental' ethical disagreement is for Kant simply evidence that one or more parties to the disagreement has failed to subject the policy underlying a practice or a proposed course of action to the rational scrutiny required by the Categorical Imperative.

Some ethical pluralists believe that ethical objectivism need not be linked with the kind of absolutism found in the work of Plato or Kant. But whether ethical pluralism is indeed compatible with ethical objectivism remains a subject of rich, and ongoing, philosophical debate. How far can the pluralist go in rejecting the absolutist's 'one right answer' approach

and remain an ethical objectivist? Many think that the most plausible pluralist candidates for the label 'ethical objectivist' are forms of ethical pluralism which recognize the possibility of diverse answers to disagreements involving derivative principles, yet retain the 'one right answer' approach to analyse disputes involving fundamental principles. Of course, any such views must confront the problem of reliably distinguishing fundamental from derivative ethical principles. Moreover, while there is no evidence that this problem is not, in principle, capable of a compelling solution, some ethical pluralists avoid the problem altogether by defending a limited and moderate relativism. On these accounts, the relativist rightly posits the possibility—and the frequent occurrence—of at least some fundamentally irresolvable ethical disputes.

2.2 Varieties of Nonobjectivism and Serious Ethical Disagreement

Meta-ethical relativism is, perhaps, the most familiar form of ethical nonobjectivism. Its main claim, as noted above, is a negative thesis about the impossibility of rationally adjudicating (purportedly) fundamental ethical disagreements. But some meta-ethical relativists also attempt to defend a normative ethical relativism that makes a positive claim about the content of morality and about the nature and sources of ethical disagreement. In its bolder formulations, normative ethical relativism asserts the 'equivalence' of all ethical practices and beliefs. This version of the view has some important implications for a relativist approach to ethical practice—most often resulting in a demand to 'tolerate' all ethical practices as 'equally valid.' According to more modest versions of normative ethical relativism, we cannot confidently assert the superiority of one ethical stance over another with which it somehow conflicts, and we should therefore withhold judgment as a recognition of ethical uncertainty.

Metaethical relativism starts from the facts of ethical diversity and disagreement, but some forms of nonobjectivism posit very different explanations of those facts. One such view is ethical subjectivism, on which ethical claims are fundamentally statements about the individual speaker's subjective feelings (not about the practices of a particular social or cultural group). What the relativist identifies as a fundamental ethical disagreement is, for the ethical subjectivist, simply evidence of the deep-rooted feelings of the parties to the dispute. Yet both metaethical relativism and ethical subjectivism allow that ethical statements have some cognitive content. On the subjectivist views, ethical statements tell us something meaningful and informative about the feelings of the speaker—and indirectly about the feelings of any similarly constituted persons. On metaethical relativism, in contrast, ethical statements tell us what members of a particular social or cultural group believe to be good—and, ultimately, about the local social worlds in which ethical practices arise and have their (limited) application (Williams 1985).

It is important that some nonobjectivist views find ethical claims to be cognitively substantive, because there are varieties of nonobjectivism which deny that ethical statements have any cognitive content at all. The most influential and controversial such view is the theory of emotivism, first articulated in the early twentieth century by A. J. Ayer (1946). According to Ayer, fundamental ethical judgments (he called them 'pure judgments of value') are non-rational, non-cognitive expressions of emotion which cannot be true or false. Even the apparently most serious ethical disagreement, for Ayer, is not much different from a round of energetic 'boos' and 'yeas' for the opposing teams at a sporting event. Ayer's critics charged that such claims appear to deny that ethical disagreements ever occur. Ayer's efforts to mollify his critics were problematic: moral disagreements occur, he responded, but only as disagreements 'without formal contradiction.' This hypothesis can be confirmed, he urged, by the tendency for the parties to such disagreements to resort to 'mere abuse' when arguments fail.

3. The Infirmity of Ethics Compared with Science?

3.1 Ethical Disagreement as a Source of Skepticism about Ethical Methods

Yet despite its counter-intuitive claims about ethical disagreement, and its implausibly narrow understanding of what it is to have cognitive content, emotivism was a powerful and challenging embodiment of skepticism about the rationality of ethical argument and inquiry—a skepticism firmly rooted in the frequent intractability of ethical disagreements. There is a special irony in the fact that emotivism seems to deny the reality of the very phenomenon underwriting its skeptical challenge to the methods of ethics. But for many philosophers, even those who would explicitly reject emotivist understandings of ethical language, the emotivist's challenge to ethics retains a certain potency. The continued existence of intractable ethical disagreements, for these thinkers, is evidence (along with the explanatory and predictive success of science), that ethics, in the words of logician and philosopher of science W. V. O. Quine, is 'methodologically infirm' compared with science.

3.2 Alternatives to Skepticism

Yet the conception of science as the arbiter of rationality, capable of supplying standards which must be met by any domain of rational argument and

inquiry, has been challenged. Its critics object that it wrongly presumes that all domains of human inquiry might have the same aim. They maintain that ethics and science are fundamentally different kinds of enterprises—one claiming to explain and predict natural phenomena, the other seeking to help us to lead lives worth living—and that there is no reason to expect similarities between the methods of ethics and the methods of science. To be sure, some ethicists continue to search for fundamental similarities between the methods and aims of ethics and science. But their critics urge that such efforts can only make ethics look like a poor and distant relation of science, and that we distort the aims and purposes of ethics by trying to make it conform to structures of argument in science. Still further, these critics reject the notion that the persistence of disagreement in ethics—even intractable disagreement—is intrinsically a sign of the failure of ethical methods. Might it not be, they ask, that in some domains of human argument and inquiry disagreement is an important catalyst for serious reflection, and a crucial stimulus to avoiding smugness and self-righteousness?

4. *Implications of Ethical Disagreement for Practice*

4.1 *The Merits and Demerits of Tolerance*

The existence of ethical diversity and the persistence of ethical disagreement nonetheless pose important difficulties for ethical practice. How are we to respond to those whose ethical beliefs and practices differ, sometimes quite substantially, from our own? The ethical objectivist will argue that in some or even all ethical disputes, one or more parties to the dispute has made some kind of error in moral reasoning or perception. Ethical relativists, and some pluralists, who argue for tolerance in response to ethical disagreements urge that the objectivist response is too frequently associated with moral smugness and moral self-righteousness, and potentially dangerous and culturally imperialist ethnocentrism. But these associations are not intrinsic, and the demand for tolerance poses its own dangers when it functions (as it sometimes has) to sanction local practices which involve the violent coercion of innocent victims, or murderous intolerance of internal cultural dissent. The demand for tolerance also raises urgent questions about theoretical consistency. If it is intended as a non-relative demand, incumbent on any agent, then it is clearly incompatible with the relativist stance. If it is aimed only at the inhabitants of some societies—say Western liberal societies—then it violates relativist structures in demanding the extension of what may well be a locally compelling principle to a intercultural context. Still further, uncritical tolerance of ethical diversity and disagreement must not be confused with respect for cultural diversity. Many philosophers urge that we sometimes show the greatest respect for diverse ways of life by being ready to engage in open and noncoercive discussion with those who represent them.

4.2 *The Dangers and Incoherence of Ethical Isolationism*

Some relativists are so concerned about the dangers of ethnocentrism and the imperialist over-reaching of some cultures that they insist on an ethical isolationism. According to the ethical isolationist, cultural outsiders can never understand ethical practices and beliefs associated with a genuinely unfamiliar way of life, and thus have no rational basis for any kind of critical reflection on—let alone outright criticism of—those practices and beliefs. But, ethical isolationists typically defend their views by detailing their own special understanding of the practices and beliefs of the unfamiliar 'other,' richly conveying facts about those practices and beliefs in order to convince other cultural outsiders of the limits of their understanding. This tack shows how difficult it is to defend ethical isolationism except in terms which reveal the inconsistency and implausibility of the view.

4.3 *Pluralism and Multiculturalism*

Yet philosophers who reject ethnocentrism, relativism, and isolationism must still confront the persistence of ethical diversity and disagreement and the challenges they pose, especially in the culturally complex societies found in large, modern nation-states. An increasingly influential response, associated with some forms of ethical pluralism, is the doctrine of multiculturalism, which recommends that we act and judge on the presumption that the ethical beliefs and practices of every way of life are in principle valuable and worthy of respect. But compelling philosophical discussions of multiculturalism have construed this presumption as, at best, a 'starting hypothesis,' maintaining that a final verdict on the worth of any practice must always await the results of respectful, but sustained, critical reflection (Taylor 1992). In response, some philosophers have urged adoption of a critical multiculturalism (as a form of objectivist ethical pluralism) which leaves open the possibility that reflection on almost any practice might generate rationally compelling grounds on which to reject the practice as ethically indefensible (Moody-Adams 1997). The question of how to put this critical multiculturalism into practice—how to articulate and apply plausible principles for tolerating some stances, rejecting others, and intervening in practices deemed intolerable—will be a central topic of debate in the normative ethical and political philosophy of the twenty-first century.

See also: Democracy: Normative Theory; Discrimination; Discrimination: Racial; Diversity and Disagreement; Ethics and Values; Multiculturalism; Relativism: Cognitive; Value Pluralism

Bibliography

Ayer A J 1946 *Language, Truth, and Logic*. Pelican, Harmondsworth, UK

Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston

Berlin I 1969 Two concepts of liberty. In: Berlin I (ed.) *Four Essays on Liberty*. Oxford University Press, New York

Brandt R 1959 *Ethical Theory*. Prentice-Hall, Englewood Cliffs, NJ

Brandt R B 1967 Ethical relativism. In: Edwards P (ed.) *Encyclopedia of Philosophy*. Macmillan Press, New York, Vol. 3

Duncker K 1939 Ethical relativity? (An inquiry into the psychology of ethics). *Mind* **48**: 39–57

Gowans C W (ed.) 1987 *Moral Dilemmas*. Oxford University Press, Oxford, UK

Hampshire S 1983 *Morality and Conflict*. Harvard University Press, Cambridge, MA

Herskovits M J 1972 Tender and tough-minded anthropology. In: Herskovits F (ed.) *Cultural Relativism: Perspectives in Cultural Pluralism*, 1st edn. Random House, New York

Ladd J 1957 *The Structure of a Moral Code*. Cambridge University Press, Cambridge, UK

Montaigne M de 1958 *The Complete Essays* [trans. Frame D]. Stanford University Press, Stanford, CA

Montesquieu C de S 1973 *Persian Letter* [trans. Betts C J]. Penguin, Harmondsworth, UK

Moody-Adams M M 1997 *Fieldwork in Familiar Places: Morality, Culture and Philosophy*. Harvard University Press, Cambridge, MA

Nagel T 1979 The fragmentation of value. In: Nagel I (ed.) *Mortal Questions*. Cambridge University Press, Cambridge, UK

Plato 1961 *The Collected Dialogues*. Pantheon Books, New York

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Stocker M 1989 *Plural and Conflicting Values*. Oxford University Press, Oxford, UK

Sumner W G 1907 *Folkways*. Ginn Publishers, Boston

Taylor C 1985 The diversity of goods. In: Taylor C (ed.) *Philosophy and the Human Sciences: Philosophical Papers*. Cambridge University Press, Cambridge, UK, Vol. 2

Taylor C 1992 The politics of recognition. In: Gutmann A (ed.) *Multiculturalism and the Politics of Recognition*. Princeton University Press, Princeton, NJ

Williams B 1985 *Ethics and the Limits of Philosophy*. Harvard University Press, Cambridge, MA

Wilson B (ed.) 1970 *Rationality*. Harper and Row, Evanston, IL

Winch P 1972 Understanding a primitive society. In: Winch P (ed.) *Ethics and Action*. Routledge and Kegan Paul, London

Wittgenstein L 1978 In: Wittgenstein L *Philosophical Investigations*. Basil Blackwell, Oxford, UK

M. M. Moody-Adams

Divorce and Children's Social Development

Parental divorce is a common experience in the lives of children. Each year in the USA, over one million children experience the breakup of their parents' marriage. Researchers are interested, therefore, in determining the effects of this event on child development. Although much of the knowledge about children's adjustment to parental divorce is limited to studies in which children are examined at a single time point after the divorce, a small number of researchers follow children over time, beginning with the divorce, and a few provide a rare look at children before the divorce occurs. The studies tap a variety of variables, including measures of personality, intelligence, mental health, and relationship quality with family and peers. This article summarizes literature across a variety of designs and indicators of children's development. The focus is predominately on work originating in the USA, an emphasis that reflects the fact that the national divorce rate is higher than that of other Western countries.

1. Rise of Divorce as an Area of Study

As compared to earlier decades, the divorce rate in the USA rose rapidly during the 1960s and 1970s. The annual divorce rate in 1940, for example, was nine divorces per 1,000 existing marriages. This figure almost tripled by the end of the 1970s, climbing to an annual rate of over 22 divorces per 1,000 marriages (see Cherlin 1992 for a discussion). Since then, the annual divorce rate has declined slightly, to around 20 divorces per 1,000 marriages, but remains far above the levels of the middle part of the twentieth century. The divorce rate in the USA is far higher than in other industrialized nations, more than 1.5 times the divorce rate in the UK, and more than double that of Germany, France, and Japan. As a consequence, it is estimated that as many as 60 percent of children in the USA will spend some of their childhood or adolescence in a single-parent family. The marked increase in divorce has led researchers to examine how variations in family arrangements affect the development of children.

The most common strategy used to investigate this topic has been to compare children who have experienced parental divorce with children who have not. Although the specific estimate varies across studies (see Amato and Keith 1991a for a meta-analysis), in general, about 20 percent of children experiencing parental divorce exhibit serious problems, compared to about 10 percent of children who have not experienced parental divorce. This finding can be interpreted in different ways to support opposing perspectives of the effects of divorce on

children's adjustment. Because this finding shows that risk of serious problems is doubled, it can be concluded that parental divorce has very negative consequences for children. Alternately, because the finding also shows that 80 percent of children who experience this transition show no serious difficulties, it can be concluded that parental divorce has limited negative consequences. These opposing interpretations, based on simple comparison studies, are evident throughout the history of divorce as an area of study (see Simons et al. 1996 for a discussion).

Over the last several decades, research strategies have evolved to provide a more nuanced picture of divorce and children's development. Improvements in the methods used in these studies, such as including statistical controls, broadening the taxonomy of outcomes assessed, obtaining assessments from multiple perspectives, incorporating observational measures, documenting changes in children's adjustment over time, and obtaining assessments of children and families prior to divorce, have allowed for a more detailed view of parental divorce and its aftermath. These studies have led to a growing awareness that children's responses to parental divorce are highly variable and dependent on the particular set of circumstances faced by families before and after the actual marital dissolution (see Hetherington et al. 1998 for a review). In this view, the experience of divorce brings with it a series of unique challenges and changes and entails a complex mixture of potentially positive and negative consequences. Researchers from this perspective, which is labeled a transactional or mediational model, view marital transitions as involving a cascade of predivorce and postdivorce changes in family relationships and economic/social circumstances. It is these changes, rather than divorce itself, that affect children's adjustment. This transactional model of divorce has become the dominant paradigm for contemporary research (Amato 1993, Emery 1988, Forgatch et al. 1995, Hetherington 1989, Hetherington et al. 1998, Simons et al. 1996).

2. *Adjustment of Children in Divorced Homes*

Compared to children who have never experienced their parents' divorce, those who have score lower on measures of well-being and higher on measures of a wide array of problems (see Hetherington et al. 1998 for a review). Children of divorce are more likely to have academic problems and difficulties in relationships with peers, parents, and siblings. They exhibit higher levels of externalizing problems, such as acting-out behaviors, noncompliance, aggression, hyperactivity, and delinquency. They also exhibit higher levels of internalizing problems, such as depression, anxiety, low self-esteem, and social withdrawal. As adolescents, they have higher rates of school dropout, substance abuse, sexual activity, and teenage pregnancy. As adults, they have lower occupational status and life satisfaction, and higher rates of economic hardship and mental health problems such as depression (see Amato and Keith 1991b for a meta-analysis of adult outcomes). In addition, their own marriages are more likely to be negative and coercive and to end in divorce. Although these adverse outcomes are more common in children from divorced homes than in children who have not experienced divorce, it is important to note that the majority of children experiencing this parental transition escape these problems.

Parental divorce may either initiate new difficulties or exacerbate existing ones. Studies that assess children prior to parental divorce demonstrate that problem behaviors are evident well before the marital dissolution (e.g., Block et al. 1988). In addition, many of the differences between children who have and have not experienced parental divorce are diminished once predivorce functioning has been controlled (Cherlin et al. 1991). Thus, it is possible that some of the difficulties exhibited by children of divorce are a result of factors that have existed prior to the divorce itself. Studies showing higher problem behavior among children of divorce therefore may be a consequence of simply having a greater proportion of more vulnerable individuals represented in the divorced population, a problem known as selection. In support, some studies have demonstrated that genetic factors play a significant role in the risk of divorce (Jockin et al. 1996). It is important to note, however, that several other studies show a significant increase in problems for children of divorce, even after controlling for predivorce factors (e.g., Chase-Lansdale et al. 1995).

Researchers have investigated whether the effects of parental divorce depend on the child's gender, age, or the length of time since divorce (see Hetherington et al. 1998 for a discussion). A clear picture from these studies has yet to emerge. With respect to gender, early studies showed that boys had more difficulty adjusting to divorce than girls. More recent studies cast doubt on these findings by showing that adolescents from divorced families demonstrate higher rates of conduct problems and depression, regardless of gender. In fact, emerging evidence indicates that females may exhibit greater long-term consequences when teen parenthood is involved. This early transition to parenthood is likely to have more adverse effects for females because of their more common role as custodial parents.

Relative to child age at the time of divorce, some researchers have concluded that preschool-age children are at greatest risk for long-term problems. They argue that limitations in cognitive development hinder younger children from understanding the complex nature of divorce, giving rise to fears of parental abandonment and problematic outcomes. Other researchers have countered that the results concerning age are inconsistent and inconclusive. In many studies, for example, the child's age is confounded with the

length of time since parental divorce: the greater difficulties found in younger children may reflect the fact that less time has passed since the divorce occurred.

It is also not clear how children's adjustment to divorce changes over time. Some studies have shown that children exhibit an increase in problem behavior in the immediate aftermath of divorce that dissipates over time, suggesting that at least some of the difficulties are transitory. Other studies have indicated that problem behavior may emerge at a later age such as adolescence, when children face new challenges (e.g., Hetherington 1989). Some long-term follow-up studies document that the effects of divorce on problem behavior persists into adulthood at least for a subset of children by affecting educational and occupational attainment (Amato and Keith 1991b, Chase-Lansdale et al. 1995).

In sum, although parental divorce is associated with an increased risk of problem behavior overall, children's responses to divorce are highly variable, and the particular trajectory of development is affected by the particular circumstances occurring before and after divorce. The environmental factors influencing children's adjustment to divorce are addressed in the next section.

3. *Risk and Protective Factors*

In the transactional model of divorce, changes in family relationships and environmental circumstances before and after divorce, rather than the divorce itself, are the critical determinants of children's postdivorce adjustment problems. These risk and protective factors influence the extent to which children will be adversely affected by parental divorce. Factors that are present prior to divorce include characteristics of the parents themselves (see Capaldi and Patterson 1991). Research indicates that parents who divorce are more likely to have a history of economic problems, alcoholism, depression, and antisocial behavior. Not surprisingly, they also are more likely to have poor parenting skills prior to the divorce. Postdivorce differences in children's adjustment may thus reflect predivorce exposure to inconsistent, inept childrearing practices and chaotic household routines. Moreover, couples who later divorce are more likely to display poorer conflict resolution skills and higher negativity. Through modeling and imitation, children may learn these ineffective strategies and employ them in other contexts. For example, in high conflict homes, children may be drawn into marital conflict as a way for spouses to gain relative advantage over each other, and these experiences may train children to manipulate their parents. Consistent with these notions, research has shown that children living in homes where the marriage is distressed but not yet formally dissolved fare worse than those who experience a relatively amicable divorce (see Hetherington 1999). In addition, divorce appears to improve the adjustment of children for whom postdivorce parental conflict is dramatically reduced over predivorce levels.

Some risk and protective factors involve changes occurring after divorce. For many families, parental divorce introduces a series of social and economic changes that may interfere with the functioning of both parents and children (see Forgatch et al. 1995, Hetherington et al. 1998, Simons et al. 1996 for more detailed discussion). Economic stress introduces deterioration in living conditions, as families move to neighborhoods and schools of lower quality. Custodial parents face the challenge of task overload in managing new household, childcare, and financial responsibilities. Custodial parents may lose friends who were part of the couple's network, resulting in isolation and reductions in social support. Adults dealing with such stressors show higher rates of physical and mental health problems that interfere with their ability to be competent parents. Consistent with this view, custodial and noncustodial parents are less effective and consistent in controlling and monitoring their children, and show higher levels of punitiveness and irritability with their children than nondivorced parents. Moreover, conflict between parents and children becomes more frequent and intense after divorce. Researchers have consistently demonstrated that child adjustment after divorce is explained in part by the effect of stressors experienced by custodial parents on their parenting practices and that high quality custodial parent–child relationships mitigate the negative effect of divorce-related stressors on adjustment problems.

In addition to deteriorations in the quality of the relationship between the custodial parent and child, most children experience a dramatic loss of contact with the noncustodial parent. Although noncustodial mothers are more likely to maintain contact with their children than are noncustodial fathers, simply maintaining contact with the noncustodial parent does not appear to enhance children's postdivorce adjustment. Involvement with a noncustodial parent is most beneficial when parents maintain an amicable postdivorce relationship. Moreover, the quality rather than the quantity of the involvement by the noncustodial parent appears to play a more critical role in children's postdivorce adjustment (Hetherington et al. 1998, Simons et al. 1996).

Following divorce, relationships with siblings and peers change as well. Sibling relationships become more negative and conflictual, and such changes in relationships have been related to increased aggression and other problem behaviors. Moreover, children of divorce are more likely to become involved with antisocial peers, and may experience a loss of positive peer support due to residential changes. These alterations in peer relationships may exacerbate potential

difficulties by increasing the child's exposure to delinquent activities or by removing a potential buffer from conflictual relationships within the family (Hetherington et al. 1998).

4. Future Directions

For a sizeable group of children, the initial divorce is the first in a series of parental transitions. About half of all children whose parents divorce will become part of a stepfamily, and 10 percent of children will experience two divorces of their custodial parent during childhood or adolescence. Many children will join new families through cohabitation rather than remarriage, and there is some evidence that cohabiting families experience even higher rates of dissolution than remarried families. A sequence of multiple family transitions places children at increased risk for problematic outcomes. Research is needed to understand how children respond to their custodial parents' dating, repartnering, and cohabitation, and to identify factors that promote or hinder children's adaptation.

See also: Developmental Psychopathology: Child Psychology Aspects; Divorce and Gender; Early Childhood: Socioemotional Risks; Family Theory: Economics of Marriage and Divorce; Infancy and Childhood: Emotional Development; Marital Interaction: Effects on Child Development; Parenting: Attitudes and Beliefs

Bibliography

Amato P R 1993 Children's adjustment to divorce: Theories, hypotheses, and empirical support. *Journal of Marriage and the Family* **55**: 23–38

Amato P R, Keith B 1991a Parental divorce and the well-being of children: A meta-analysis. *Psychological Bulletin* **110**: 26–46

Amato P R, Keith B 1991b Parental divorce and adult well-being: A meta-analysis. *Journal of Marriage and the Family* **53**: 43–58

Block J H, Block J, Gjerde P F 1988 The personality of children prior to divorce: A prospective study. *Child Development* **57**: 827–40

Capaldi D M, Patterson G R 1991 Relation of parental transitions to boys' adjustment problems: I. A linear hypothesis. II. Mothers at risk for transitions and unskilled parenting. *Developmental Psychology* **27**: 489–504

Chase-Lansdale P L, Cherlin A K, Kiernan K E 1995 The long-term effects of parental divorce on the mental health of young adults: A developmental perspective. *Child Development* **66**: 1614–34

Cherlin A J 1992 *Marriage, Divorce, Remarriage*. Harvard University Press, Cambridge, MA

Cherlin A J, Furstenberg F F Jr, Chase-Lansdale P L, Kiernan K E, Robins P K, Morrison D R, Teitler J O 1991 Longitudinal studies of effects of divorce on children in Great Britain and the United States. *Science* **252**: 1386–9

Emery R E 1988 *Marriage, Divorce, and Children's Adjustment*. Sage, Newbury Park, CA

Forgatch M S, Patterson G R, Ray J A 1995 Divorce and boys' adjustment problems: Two paths with a single model. In: Hetherington E M, Blechman E A (eds.) *Stress, Coping, and Resiliency in Children and Families*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 67–105

Hetherington E M 1989 Coping with family transitions: Winners, losers, and survivors. *Child Development* **60**: 1–14

Hetherington E M (ed.) 1999 *Coping with Divorce, Single Parenting, and Remarriage: A Risk and Resiliency Perspective*. Lawrence Erlbaum Associates, Mahwah, NJ

Hetherington E M, Bridges M, Insabella G M 1998 What matters? What does not? Five perspectives on the association between marital transitions and children's adjustment. *American Psychologist* **53**: 167–84

Jockin V, McGue M, Lykken D T 1996 Personality and divorce: A genetic analysis. *Journal of Personality and Social Psychology* **71**: 288–99

Simons R L, et al. 1996 *Understanding Differences Between Divorced and Intact Families*. Sage, Thousand Oaks, CA

E. R. Anderson and S. Wolchik

Divorce and Gender

Divorce has a significant impact on the lives of both women and men. Men and women have different experiences at divorce, however, and divorce has different consequences for them. Typically, women experience economic decline after divorce and some divorced mothers fall into poverty. This article will examine recent trends in divorce in Europe and the USA, and will then consider the case of the USA in more depth.

While divorce causes women great hardship, it can also be liberating for them. Before divorce became legal, for example, married women who encountered abuse—physical or psychological abuse, or the abuse of alcohol—or who were otherwise trapped in bad relationships, had no way to get out of their marriages. Nineteenth and twentieth century women's movements fought for the right to divorce, which was legalized in the nineteenth century in the USA, and by various European countries throughout the twentieth century. The freedom to divorce is an important right for women, one still denied to wives in some countries (Goode 1993).

Even after the legalization of divorce, however, it took until the second half of the twentieth century for divorce to become socially acceptable in Western countries. States tried to discourage divorce until well into the twentieth century. To keep families together

and to 'protect' women and the family, divorce was restricted through fault-based systems that remained in effect until the 1960s (Goode 1993). Under these systems, divorce could take place only if one party was able to prove the other guilty of adultery, cruelty, or other serious offenses. In the event that a divorce did occur, the wife was to be financially protected by her ex-husband. Since the state put in place no mechanisms to enforce this system, however, it did not in fact protect women. In the United States, most women never received alimony from their ex-husbands nor did most women obtain much property from a marriage, as most couples had little property to divide (O'Connell 1988). Thus, a major contradiction existed between the law's assumption that the wife was dependent and should be protected in her domestic role, and the fact that most women received no such protection.

The last half of the twentieth century saw the stigma against divorce lifted and a rise in divorce rates in all North American and European countries (Goode 1993). The USA has one of the highest divorce rates—50 percent of all marriages now end in divorce (Stewart et al. 1997). Fundamental changes in divorce law in Europe and North America have also taken place. The most notable change has been the introduction of 'no-fault' divorce legislation. Most European countries have provisions for ending marriages without charges of fault, and in the USA by 1985 all states had passed 'no fault' laws, based on the principle of 'irretrievable breakdown' of the marriage, 'incompatibility,' or 'irreconcilable differences,' which permitted a divorce even if only one partner desired it (Kurz 1995). The no-fault system overturned the traditional view of women as dependents, stipulating that men are no longer responsible for the lifelong support of their former wives. In the USA, the new model looks at marriage as a partnership between husbands and wives, much like a business partnership. At divorce, women are expected to support themselves through paid work, supplemented by child support and an equitable division of assets from the marriage. Alimony, or lifelong financial support that in theory was available to a divorced wife, has been replaced by 'spousal support' awards of short duration, typically to the former wife, until she can begin supporting herself through paid work. Similarly women receive no special consideration in the division of property in the no-fault era. Property is to be divided equitably between divorcing spouses, much as one would divide property when dissolving a business.

No-fault divorce legislation with its gender-neutral assumptions has had benefits for women, particularly in overturning the stereotypic view of women as dependents of their husbands and substituting a view of men and women as equal. It has continued to leave women economically vulnerable, however. I will demonstrate below how divorce disadvantages women in the USA. While divorced women everywhere face economic difficulties, the USA provides an instructive case of what happens to women and children in a country that does not have adequate social supports for single mothers or a strong social welfare system that can mitigate the negative impact of divorce. The assumption behind current US policy, that divorced women would be able to support themselves adequately on their salaries and wages and on assets from their marriages, overlooks the realities of women's lives. In the following section, I will examine current divorce policies, how they impact on women's economic situation after divorce, and divorced women's ability to secure assets from the marriage and ongoing support.

1. *Economic Situation of Divorced Women*

In the USA, after divorce mothers' standard of living goes down by at least a third while the incomes of divorced fathers go up by 15 percent (Hoffman and Duncan 1988). Mothers' decreased standard of living affects not only themselves, but also their children, who typically live at the standard of living of their mothers. Why do divorced women have so much difficulty securing an adequate standard of living? First, many women cannot earn enough in the paid labor market to adequately support themselves and their children. Women make only 78 percent of what men do (Blau 1998) and many women are concentrated in low-paying jobs (England 1992).

Second, although women make an enormous life investment in their marriage, they are left on their own with little to show for it. Many divorced women receive no property of any value. The main asset of a couple is often the home, which in some cases is sold at the time of divorce so that each spouse can receive half its value. This usually leaves a woman and her children with smaller quarters and causes the children additional disruption. As well as having little property, only about one-sixth of divorced women have received spousal maintenance awards, most of which are of limited duration, are not large, and only half of which are paid in full (Weitzman 1985).

While divorced women have been unable to secure many assets from their marriages, in theory one resource remains available to the 85–90 percent of the divorced women who have physical custody of their children–child support. Most mothers are entitled to it, nearly all mothers need it, yet relatively few receive it: fewer than half of divorced women receive their full child support award (Roberts 1994). Child support can be critical for divorced mothers and their children, as well as other single mothers, to help them escape poverty. Researchers have established that regular child support payments would improve the standard of living of women and children and would lift a number of them out of poverty (Nichols-Caseboldt 1992).

Despite mothers' demonstrated need for child support, however, the government has not implemented an adequate child-support system. First, child support awards have been very low. As child support is currently calculated and awarded, it covers less than half the costs of raising a child and typically is no more than 10 to 20 percent of a father's income, never more than 30 percent (Weitzman 1985). Thus fathers contribute much less to the support of their children than do mothers. The problem is that child support awards are based on the costs of raising children in a two-income family; they do not take into account that divorced single mothers have many additional costs in establishing a new household. Further, career assets are not recognized as marital property. Thus the primary wage-earner, generally the husband, is permitted to keep most of the assets accumulated during his marriage and does not suffer financially at divorce, while the wife, who has invested in her family and in her husband's career, is deprived of a return on her marital investment.

As for child support enforcement procedures, fortunately the Federal government and states have recently legislated tougher policies. The most important and effective reform for child support enforcement enacted to date is income withholding, a procedure by which an obligator's (typically the father's) payment is withheld from his paycheck by his employer, who then sends the money to the state child support agency (Roberts 1994). This will surely help raise the rate of child support collections. Problems will remain for women in improving their receipt of child support, however. Even with income withholding, estimates are that between 23 percent and 35 percent of women will still not receive child support (Garfinkel 1992). One problem is that there is no child support enforcement system for women whose ex-husbands or former partners are self-employed. Collecting child support in interstate cases, which comprise 30 percent of all child support cases (Garfinkel 1992), is also a problem. There are new enforcement mechanisms for interstate cases; however the effectiveness of these new regulations depends on the ability of state child support agencies to take action on individual child support cases. Unfortunately, at this time state agencies are underfunded and labor under very large caseloads (Roberts 1994).

These problems with the child support system mean that the great majority of divorced mothers are in an economically dependent, vulnerable position. There are few social welfare programs or subsidies in the USA that enable them to live at an adequate standard of living outside of marriage. Some must resort to welfare, which provides a standard of living that is typically below the poverty level. Women who come from communities that have fewer resources and who experience discrimination, including many women of color, face even higher rates of poverty after divorce.

While there are problems with child support collection in many European countries as well, according to Glendon (1987) the legal system and public policies in the USA do less than those in any other Western country to help families cope with the consequences of divorce. By contrast, some European countries have social welfare systems that are typically much more supportive of single mothers and children. In the Scandanavian countries, the state actually pays child support to mothers and collects child support from fathers (Goode 1993). This policy is based on the belief that it is the state's responsibility to ensure that children are adequately provided for.

2. *Social Issues at Divorce*

While divorce brings economic disadvantage for women and children, it can also have positive consequences. US researchers have noted that many men and women are positive about having divorced and report a greater sense of well-being after ending their marriages. Some researchers believe that divorce brings particular contentment to women, who gain an increased sense of competency and control (Riessman 1990). In Kurz's (1995) study of divorced women, women's feelings about their divorce were related to the circumstances of their marriages. The majority of women reported feeling positive about the ending of their marriages, particularly the substantial number of women whose husbands had been violent, had used drugs and alcohol, had been controlling, or had rarely been at home. For these women divorce was a welcome release from a bad marriage, an outcome consistent with the goals of the nineteenth century divorce reformers. The exception to this pattern was those women who had been left by their husbands against their wishes, for whom adjustment to divorce was more difficult.

While a number of couples report their satisfaction with being divorced, however, a certain percentage of them report conflict after divorce. Arendell (1995), in her study of divorced fathers, found that some fathers report harassing their former wives over issues of payment of child support, custody, and visitation. Forty percent of the men in her sample described incidents in which they had threatened explicitly or resorted to acts of violence against their former wives since the end of their marriages. In Kurz's sample, 30 percent of women were fearful at some point during negotiations for child support, with similar percentages reporting fear during negotiations for custody and property. Those women who were fearful were most often those who had experienced violence during their marriages. Some of these women were afraid to demand child support. Some feminists have been concerned that new child support legislation mandating that mothers identify fathers and cooperate with child support enforcement could put some women with violent ex-husbands at risk (Kurz 1995).

These advocates argue that, for this reason, women who have experienced violence should not be forced to cooperate with child support enforcement agencies against their will.

The fear of harm or violence can have other negative consequences for women. In the case of negotiating for custody and child support, some fathers threaten to sue for custody even when they don't really want it, as a way to get mothers to reduce their demands for child support. In response, some mothers forfeit their right to child support in order to secure custody of their children. Recent changes in custody law make these kinds of suits possible. A presumption of joint custody has replaced the maternal presumption in the determination of child custody. Fathers' rights groups in particular have agitated for more rights to participation in their children's lives and many feminists have also hoped that fathers would become more involved with their children after divorce (Kurz 1995).

At the present time, most divorced parents have some type of joint custody agreement, most often joint legal custody mandating that both parents should be involved in major decisions about the child's life and in a small number of cases, joint physical custody, where the child goes back and forth between the mother's and father's residences. While joint custody can promote shared parenting that is good for the child, there has been some debate about whether joint physical custody is best for children (Maccoby and Mnookin 1992). For divorced mothers, however, joint custody can be disadvantageous. In many such custody agreements, children end up spending more time with their mother. Since fathers with joint custody agreements do not pay child support, however, mothers and their children then lose financially in this arrangement (Mason 1999). In addition, in the minority of cases when fathers actually do want custody, or at least sue for it, they have a greater likelihood of getting custody than the mothers (Mason 1999, p. 22).

Conflict is not only stressful for former spouses and financially disadvantageous for most women, but it can also have a negative impact on children. Scholars have disagreed on whether or not divorce harms children. Some argue that it does, while others claim that being a child of divorce is no more difficult than being a child of an intact marriage characterized by high levels of conflict (Johnston 1994). One thing scholars agree on, however, is that conflict between spouses during marriage and after divorce can be harmful for children (Stewart et al. 1997).

3. Conclusion

Gendered laws, procedures, and practices play a critical role in divorce and have a negative impact on the economic situation of divorced women and children. At the same time, the discourse on divorce and the assumptions underlying much divorce law are gender-neutral. Unfortunately, some researchers and policymakers have proposed divorce policies that ignore the gender-based aspects of divorce. They advocate policies to increase commitment to marriage and make it more difficult for people to get divorced—mandatory waiting periods before divorce, mandatory marriage counseling or marriage education, and the reintroduction of fault-based divorce (Popenoe 1996).

Women's rights advocates view such proposals with alarm. They favor strengthening child support enforcement and increasing child support awards. They also argue for supplementing private support with strong public measures such as are found in the Scandanavian countries, measures which provide guaranteed child support. Many also argue that the state should promote mothers' self-sufficiency through job training, higher incomes, subsidized day care and health care, and family allowances. Only such comprehensive measures, they argue, can reduce the high levels of poverty among divorced women and children.

See also: Dissolution of Family in Western Nations: Cultural Concerns; Divorce and Children's Social Development; Divorce, Sociology of; Family and Gender; Family and Kinship, History of; Family as Institution; Family Bargaining; Family Law; Family Theory: Economics of Marriage and Divorce; Family Theory: Feminist–Economist Critique; Fatherhood; Gender, Economics of; Lone Mothers in Affluent Nations; Lone Mothers in Nations of the South; Marriage; Marriage and the Dual-career Family: Cultural Concerns; Motherhood: Economic Aspects; Partnership Formation and Dissolution in Western Societies; Poverty and Gender in Affluent Nations; Religion: Morality and Social Control; Repartnering and Stepchildren; Social Welfare Policies and Gender

Bibliography

Arendell T 1995 *Fathers and Divorce*. Sage, Thousand Oaks, CA

Blau F 1998 Trends in the well-being of American women, 1970–1995. *Journal of Economic Literature* **36**: 112–65

England P 1992 *Comparable Worth: Theories and Evidence*. Aldine de Gruyter, New York

Garfinkel I 1992 *Assuring Child Support: An Extension of Social Security*. Russell Sage, New York

Glendon M A 1987 *Abortion and Divorce in Western Law*. Harvard University Press, Cambridge, MA

Goode W J 1993 *World Changes in Divorce Patterns*. Yale University Press, New Haven, CT

Hoffman S D, Duncan G J 1988 What are the consequences of divorce? *Demography* **25**: 641–5.

Johnston J 1994 High conflict divorce. *The Future of Children: Children and Divorce* **4**(1): 165–82

Kurz D 1995 *For Richer, For Poorer: Mothers Confront Divorce*. Routledge, New York

Maccoby E E, Mnookin R H 1992 *Dividing the Child: Social and Legal Dilemmas of Custody*. Harvard University Press, Cambridge, MA
Mason M A 1999 *The Custody Wars: Why Children are Losing the Legal Battle and What We Can Do About It*. Basic Books, New York
Nichols-Caseboldt A 1992 The economic impact of child support reform on the poverty status of custodial and noncustodial families. In: Garfinkel I, McLanahan S S, Robins P K (eds.) *Child Support Assurance*. Urban Institute Press, Washington, DC, pp. 189–202
O'Connell M E 1988 Alimony after no-fault: A practice in search of a theory. *New England Law Review* **23**: 437–513
Popenoe D 1996 *Life without Father*. Martin Kessler Books, New York
Riessman C K 1990 *Divorce Talk: Women and Men Make Sense of Personal Relationships*. Rutgers University Press, New Brunswick, NJ
Roberts P G 1994 Child support orders: Problems with enforcement. *The Future of Children: Children and Divorce* **4**(1): 101–20
Stewart A J, Copeland A P, Chester N L, Malley J E, Barenbaum N 1997 *Separating Together: How Divorce Transforms Families*. Guilford Press, New York
Weitzman L J 1985 *The Divorce Revolution*. The Free Press, New York

D. Kurz

Divorce, Sociology of

Among the striking features of family life during the latter part of the twentieth century has been the steep rise in divorce. This worldwide trend reflects the profound changes in relationships between men and women and in the laws that govern family stability. It carries new and still untold implications for the future of the family as a social unit. The obvious benefit of divorce is that it enables persons caught in wretched marriages to escape and to improve their lives either by remaining single or by embarking on a happier second marriage. The tragedy is that, in many countries, it leads to widespread impoverishment of mothers and children. At issue also is the long-term emotional and social impact on children's lives and on traditional family functions such as care of the aged.

The ripple effects of the high divorce rate overflow into many social domains. As reported throughout the industrialized countries, these include the diminishing propensity to marry, the rise in cohabitation, an increase in single-parent and remarried families, the greater presence of mothers of young children in the workplace, and the lowering of expectations for long-term commitment in young people who have yet to establish their own families (Cherlin 1992). Thus the accelerated rise in divorce is a major indicator of social and psychological changes and a powerful agent of these changes. It reflects not only the shifting life perspectives of men and women; it impacts powerfully on society in ways that were unexpected and probably unintended by those who led the political battles of the twentieth century to loosen the constraints of marriage.

1. Wide Variation in Divorce Laws and Practices Worldwide

It is impossible to compare divorce in different countries, except perhaps among the industrialized countries of Europe and North America. Nor is it prudent to try to delineate common antecedents or consequences of the breakdown in family life, because reports about divorce and postdivorce sequels are sparse or unavailable for much of the world. Legal constraints on divorce vary widely. In some countries divorce does not confer the freedom to remarry. Waiting periods to obtain a divorce can extend from a few months to many years. Even between contiguous countries within Europe there are widely varying restrictions in laws which reflect diverse cultural and religious attitudes (Goode 1993). But overall, there are similarities in the broad social patterns which show that the rise in the divorce rate began gradually, in different countries, during the late nineteenth and early twentieth centuries; it mounted rapidly throughout the 1970s, and seems to have reached a plateau by the 1990s. Whether this is a pause before another climb, or represents a real leveling off or downward swing, is not predictable at this time (Cherlin 1992).

The rise in divorce is most pronounced in the industrialized countries of North America and Europe, especially in the USA, Scandinavia, and the UK and remaining countries of the British Commonwealth. The fall of the last important bastion of resistance to divorce in Western Europe occurred in 1997, when Ireland adopted the Family Law Act. But divorce is increasingly evident as well in Central and South America, as it is in Eastern Europe (Everett 1997). In Japan and China, where traditionally divorce was not blocked by law—though it was discouraged by community disapproval and, in the case of China, by a burdensome bureaucracy—the divorce rate is rising (Mason et al. 1998). Family change has taken a much slower course in Muslim countries, where Islamic law has long made marital dissolution readily available to men, but has only fairly recently extended limited access to women. As the power of the traditional family and tribe was weakened by civil laws in sub-Saharan African societies, women were finally allowed to initiate divorce under very restricted circumstances (Aderanti 1997). The Moslem communities of South East Asia, on the other hand, have long institutionalized divorce and remarriage as an acceptable recourse for young women in arranged

marriages (Jones 1997). So that, except in fundamentalist Muslim countries, signs of greater freedom for men and women to free themselves from a bad marriage are apparent throughout the world.

2. Social Factors Underlying Divorce

The sheer number of divorces—especially in the USA and the Scandinavian countries, which lead the world in incidence of divorce—is an inaccurate measure of family life, because divorce statistics reflect only the dissolution of legal marriages, and millions of couples today reject marriage altogether in favor of cohabitation (Wallerstein 1998). Divorce in the West may be profoundly influenced by new and greater demands of relationships between men and women. Roussel (1989) and others have proposed that people today conceptualize marriage primarily as a vehicle for personal happiness and individual fulfillment, rather than as a family obligation and an economic and child-rearing contract. Obviously, marriages based on the participants' search for love and companionship are more fragile than those traditionally buttressed by law, church, family traditions, and social mores.

The rise in the world divorce rate is indeed a product of many complexly interacting social, economic, and psychological factors (Fu and Heaton 1995). Economic development, women's greater labor force participation, women's rising educational levels, and lower fertility rates as a result of improved contraception and the availability of abortion—all correlate positively with a country's divorce rate. Goode (1993) and other scholars have pointed to urbanization and industrialization as leading to the general weakening of the family unit as key provider of economic and social resources. As a result, individuals have greater freedom to define personal goals and to make career and marital choices separate from those laid down by traditions that demand conformity and continuity with the past.

Although it has been suggested that legal changes have spawned many divorces, the evidence is that diminishing legal barriers cause a short-term increase, but that over the long haul laws reflect rather than shape community trends. Women's employability has strongly influenced divorce by reducing women's economic dependence on men; in the USA and Europe, where women's emancipation is greatest, a majority of divorces are sought by women. But changes in the market participation of women are not necessarily accompanied by increased status or power, and certainly not by greater economic protection. Nor is there evidence that feminism as a political movement is a major influence in the worldwide rise in divorce.

The correlation between divorce and social class has varied. In earlier times divorce was the province of the affluent, but in recent decades it has been increasingly utilized by the lower social classes (Goode 1993). Divorce rose in the USA, Sweden, and other Western countries alongside the achievement of a higher social status for women. This was not true in the former USSR, where the divorce rate was high but there was no feminist movement of any consequence, and where women were consigned to more menial jobs (Everett 1997).

3. Psychological Effects of Divorce on Families

Although easier access to divorce has produced many benefits for adults locked into troubled marriages, it has carried a high cost in suffering for parents and long-term difficulties for millions of children raised in divorced or remarried families. The breakup of a marriage with children is rarely a mutual decision, and the parent who is left behind is often plunged into crisis. Approximately two-thirds of divorces in the USA and Europe involve minor children, often very young children, since typically divorce peaks during the first decade of marriage. There is mounting evidence that children in divorced families suffer with initial and long-term effects that place them at greater risk than their counterparts raised in intact families for problems throughout childhood and into adulthood.

The issue of whether children suffer as a result of divorce or whether they show the effects of the antecedent or continuing conflict in the family has long been disputed. There are now available a range of long-term studies (Amato and Booth 1997, Wallerstein 1998) which show that persons raised in divorced families have greater psychological problems in adult life, a higher divorce rate, a lower marriage rate, and more difficulty managing the interpersonal skills required for a stable marriage. Whether these sequelae can be changed by psychological or educational interventions with the parents or the children has not been demonstrated. Mediation programs which were expected to safeguard children by reducing the acrimony of the court adversarial process have not influenced psychological outcome among children (Wallerstein 1998).

4. Economic Consequences of Divorce

Women and children are economic casualties of divorce except in the Scandinavian countries, whose governments provide child support when the father fails to pay, and offer generous program child allowances and universal health care. They are somewhat protected in countries like Germany, where the father's wages can be garnished for child support. In most other countries, payments for child support are inadequate and poorly collected (Goode 1993). The court-ordered amount usually falls well below the

costs of raising a child. Moreover, women's employment is heavily gendered in most countries, and many women who were homemakers during the marriage have few marketable skills. In the former USSR, no alimony or regular child support is mandated and state assistance is meager. Many fathers are unable to pay child support and provide for their own needs. In the USA, in 1991, an estimated 40 percent of divorced women and their minor children lived below the poverty level (Teachman and Paasch 1994). Several governments are now putting more effort into collecting child support, but the plight of divorced mothers and children remains serious.

5. *The Continuing Dilemma*

The rise in divorce continues at the beginning of the third millennium to generate passionate controversy. In the USA, 'family values' has become a rallying cry among conservative groups who regard the high incidence of divorce as a symptom of moral decline and seek to reintroduce 'fault' into the 'no-fault' family laws that were adopted in the 1970s. Divorce has led to greater anger between fathers and mothers, as each group struggles for the custody of children and a larger share of economic resources. How to allocate family resources among children of first, second, and even third families is baffling. Concern has mounted about how to maintain the child's connection with both parents after divorce without disrupting the child's life.

Few dispute the importance of divorce as an avenue of escape from a brutal marriage. But many are concerned that divorce no longer reflects a necessary escape from exploitation, but rather promotes the unbridled freedom of the parents to give priority to their own desires and to overlook their children's needs (Glendon 1989). Much of the ongoing debate about the family has been motivated by clashing family ideologies and self-seeking groups. But much of it also derives from a genuine conviction that children need a stable family environment to ensure their proper emotional, educational, and moral development during their growing-up years.

See also: Divorce and Children's Social Development; Divorce and Gender; Family and Kinship, History of; Family as Institution; Family Bargaining; Family Processes; Family Theory: Economics of Marriage and Divorce; Marriage; Marriage and the Dual-career Family: Cultural Concerns

Bibliography

Aderanti A (ed.) 1997 *Family Population and Development in Africa*. Zed Books, London

Amato P R, Booth A 1997 *A Generation At Risk: Growing Up in an Era of Family Upheaval*. Harvard University Press, Cambridge, MA

Cherlin A J 1992 *Marriage, Divorce, Remarriage*, rev edn. Harvard University Press, Cambridge, MA

Everett C A (ed.) 1997 *Divorce and Remarriage: International Studies*. The Hayworth Press, Binghamton, NY

Fu X, Heaton T B 1995 A cross-national analysis of family and household structure. *International Journal of Sociology of the Family* **25**: 126–44

Glendon M A 1989 *The Transformation of Family Law*. University of Chicago Press, Chicago

Goode W J 1993 *World Changes in Divorce Patterns*. Yale University Press, New Haven, CT

Jones G W 1997 Modernization and divorce: Contrasting trends in Islamic southeast Asia and the West. *Population and Development Review* **23**: 95–11

Mason K O, Tsuya N, Choe M K (eds.) 1998 *The Changing Family in Comparative Perspective: Asia and the United States*. East-West Center, University of Hawaii Press, Honolulu, HI

Roussel L 1989 *La Famille Uncertaine*. Odele Jacob, Paris

Teachman J, Paasch K 1994 Financial impact of divorce on children and their families. *The Future of Children* **4**: 63–83

Wallerstein J 1998 Children of divorce: A society in search of policy. In: Mason M A, Skolnick A, Sugarman S D (eds.) *All Our Families*. Oxford University Press, New York, pp. 66–94

Wallerstein J, Blakeslee S 1989 *Second Chances: Men, Women & Children a Decade After Divorce*. Warner Books, New York

J. S. Wallerstein

Doctor–Patient Interaction: Psychosocial Aspects

1. *Introduction*

The quality of the interaction between doctor and patient in the consultation is central to effective healthcare delivery. Studies of doctor–patient interaction have focused on examining how doctors and patients communicate in a range of clinical settings. In addition to describing the nature of and variation in the communication process, many studies have also sought to examine how this interaction influences other healthcare outcomes, such as patient satisfaction and treatment adherence.

The development of audio and video-recording techniques has provided social and behavioral science researchers with a window into the private world of the medical consultation, allowing detailed investigations of the consultation process. From the early research it quickly became evident that there were considerable problems in the process and outcome of many consultations. Many patients reported that they had not been given sufficient information or had difficulty in understanding information which was presented in the consultation. Very often this was also

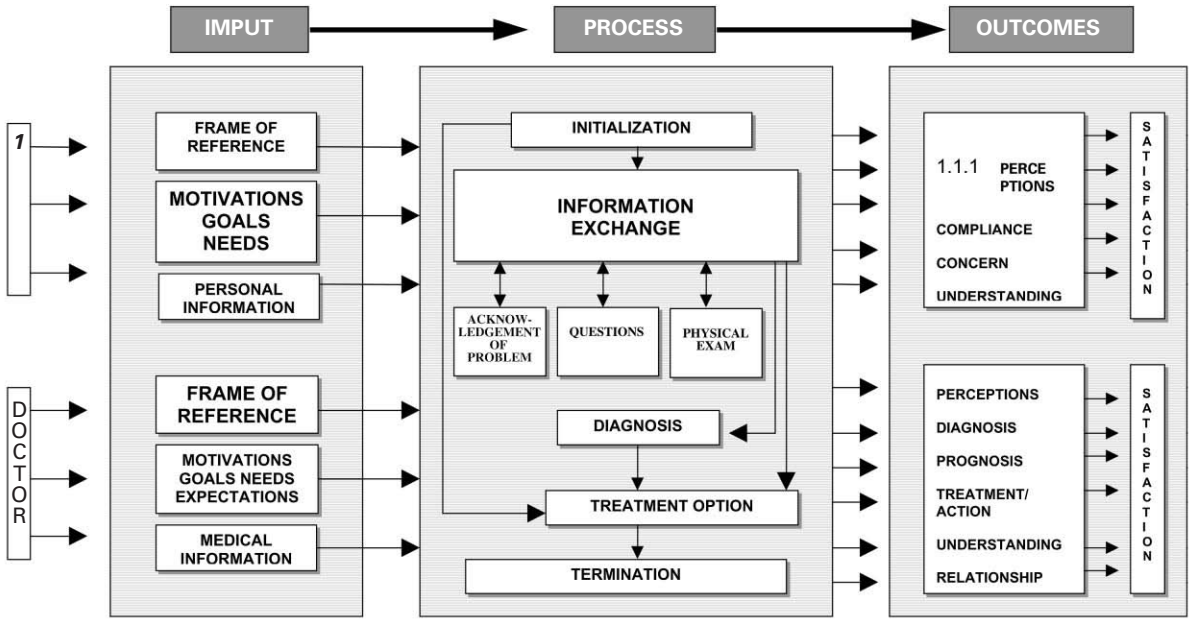

Figure 1
An input, process, and outcome model of the medical consultation (adapted from Frederikson 1993)

found to be associated with a subsequent reluctance or inability to follow the doctor's recommended treatment or advice. For example, in the landmark study of Korsch and Negrete (1972), which involved mothers taking their children to a paediatric outpatient clinic, about 20 percent of mothers were not informed clearly about the diagnosis and nearly 50 percent were uncertain afterwards as to the course of their child's illness. This and other studies have also shown that patients often report feeling that they have not been sufficiently involved in the decision-making which may have occurred in the consultation (Guadagnoli and Ward 1998).

Many studies which have analyzed the process and outcome of the medical consultation have revealed that there is considerable variation in the pattern of doctor–patient interaction and that this is influenced by factors which the doctor and patient bring to the encounter. As a result, current theoretical frameworks for understanding doctor–patient communication tend to focus on the relations between inputs (i.e., the attitudes, beliefs, expectations, etc. which patient and doctor bring to the consultation), process (the nature of the encounter), and outcome (the short and longer term effects on the patient). An example of one such framework is provided by Frederikson (1993) and is shown in Fig. 1, which provides an indication of the range of variables which have been studied. This article will, therefore, follow this structure and focus initially on these three broad areas. One of the practical benefits of this research has been the development of interventions to improve doctor–patient communication, and this work will be summarized in the final section.

2. *Input Factors in Doctor–Patient Communication*

The main input factors in doctor–patient communication reflect variations between and within both doctors and patients. Studies have shown that patients can differ along many dimensions (e.g., age, personality, coping style, expectations, etc.) which can influence the process and outcome of a medical consultation. Similarly, there is growing evidence that, despite having had a common medical training, doctors differ in their attitudes and role perceptions in ways which can have a significant influence on what happens in the consultation. This section will, therefore, review a selection of the evidence on the nature and effects of the two input factors.

2.1 *Patient Input Factors*

A number of studies have shown that patients not only cope with health threats in diverse ways but also show consistent differences in how they want to be involved in the healthcare process. These studies also show that patients differ in the amount of information which they would like to receive about their health problem.

For example, Miller et al. (1987) distinguish between two broad groups of patients whom they have labeled as 'monitors' and 'blunters,' with the former being more inclined to need and seek out information about their problem and treatment whereas the latter group prefer consultations in which relatively limited information is provided.

Patients come into the healthcare setting with different levels of biomedical knowledge, based on their past experience. There is also consistent evidence that patients have differing expectations for specific consultations and, for doctors, an awareness of these can be helpful in understanding not only why they are seeking help at that time but also in being able to respond effectively to their needs. Williams et al. (1995) have shown that the most common preconsultation expectation in primary care patients is for explanation and understanding of their problem, with much smaller numbers wanting tests, diagnoses, or support. These prior expectations can be important in determining outcomes since consultations in which patient expectations are met, have been shown to result in greater satisfaction and an increased willingness to follow advice or treatment.

These differences in patients' expectations are also related to their perceptions of the role of the doctor. A number of studies have shown that patients differ in their perceptions of the role of the doctor and the extent to which it is appropriate to discuss different types of problems, particularly those of a more psychosocial type (e.g., Bower et al. 1999). However, even when patients feel that it is appropriate to discuss psychosocial problems with their doctor, this does not necessarily happen since the doctor's own approach to the consultation may preclude this.

2.2 *Differences Between Doctors*

Doctors have been categorized in various ways according to their role perceptions and the extent to which they concentrate on the technical or more psychosocial aspects of patient care, as well as their beliefs about whether patients should be actively involved in the consultation and in decision-making about the management of the clinical problem (e.g., Grol et al. 1990). Inevitably these broad attitudinal differences are reflected in the way in which the consultation is conducted and in other aspects of professional behavior including decision-making, prescribing and management of clinical problems.

Although doctors undergo a common training and share a common body of knowledge and skills, wide variations in their consulting behavior have been noted. Part of this variation is a by-product of personality differences and can be seen in differences in interpersonal aspects of clinical practice but part is a reflection of differences in the approach to clinical problems. Some of these are between-individual differences whereas others reflect factors which can vary within the same individual, such as changing mood, time pressure, and various other contextual influences. In their overview of clinical decision-making, Schwartz and Griffin (1986) point out that there are often substantial disagreements between doctors when interpreting the same clinical information (e.g., x-rays). They also note that doctors may be inconsistent and disagree with their own previous judgments.

During clinical decision-making, the doctor is often faced with having to process a large amount of information about the patient and the possible diagnoses and treatment options. Since there are limits in what can be attended to and how much information can be held in working memory, doctors develop heuristics, or general strategies, for processing diagnostic and treatment information and these are subject to a range of influences. Kahnemann et al. (1982) have described a number of the heuristics which are used during clinical decision-making. Similarly, the doctor's attitudes and beliefs about the doctor's and the patient's role can influence the weighting given to information from patients (e.g., the relative attention given to physical and psychosocial information). In addition, the doctors' mood can affect information-processing in a number of ways by directly influencing the speed and accuracy of clinical problem-solving (Isen et al. 1991) or possibly by constraining access to mood congruent semantic information, as has been demonstrated in more general studies of the effect of mood on cognition.

3. *The Consultation Process*

There are a range of methods and frameworks for analyzing and describing the process of the consultation. One of the broadest distinctions made has been between consultations which are described as patient-centered and those which are doctor-centered, reflecting the extent to which the doctor or patient determines what is discussed (Grol et al. 1990). Doctor-centered consultations are ones in which closed-questions are used more often and the direction is determined by the doctor, typically with a primary focus on medical problems. In contrast, patient-centered encounters involve more open-ended questions with greater scope for patients to raise their own concerns and agendas.

A number of specific methods have been developed for carrying out detailed analyses of the social interaction between doctor and patient based on audio or videotapes or transcripts of the consultations (see Roter and Hall 1989). Early methods classified the verbal statements of the doctor and patient into those which were task or emotion-focused and then into more specific categories (e.g., asking questions, giving

interpretations, etc.). Attempts have also been made to define a number of more general ways of classifying doctor–patient interactions. For example, one can distinguish between verbal and nonverbal information and, within the verbal domain, six broad categories can be defined (information-giving; information-seeking; social conversation; positive talk; negative talk; partnership building). From a meta-analysis of these broad categories (Roter 1989) it has been found that for the doctor, information-giving occurs most frequently (approximately 35 percent of the doctor's communication) followed by information-seeking (approximately 22 percent), positive talk (15 percent), partnership building (10 percent), social conversation (6 percent), and negative talk (1 percent). In contrast the main type of patient communication consists of information giving (approximately 50 percent) with less than 10 percent involving information seeking.

A fairly recent and quite widely used method for analyzing the consultation process has been developed by Roter et al. (1997). This method classifies the doctor's communication patterns into five distinct groups: (a) 'narrowly biomedical' characterized by closed-ended medical questions and biomedical talk; (b) 'expanded biomedical' similar to the narrowly biomedical but with moderate levels of psychosocial discussion; (c) 'biopsychosocial' reflecting a balance of psychosocial and biomedical topics; (d) 'psychosocial' characterized by psychosocial exchange; (e) 'consumerist' characterized by patient questions and doctor information giving. They found that 'biomedical' approaches were used more often with more sick, older, and lower income patients by younger, male doctors.

4. *Outcomes of Doctor–Patient Communication*

Many studies of doctor–patient communication have attempted to evaluate the effects on a range of outcomes. Three broad groups of outcome have been studied, namely understanding and recall, satisfaction, and adherence, and each of these are now briefly discussed.

4.1 *Patient Understanding and Recall*

Changes in patients' knowledge, understanding, and recall of the relevant information provided in the consultation have been assessed in a number of studies. Ley and co-workers have shown that about half the information provided in the consultation is forgotten by patients a short time afterwards but there is considerable range in what is recalled (see Ley 1997). This variation partly reflects the type of setting and sample used and partly the method which is used to test for recall. There are a number of other factors involved including the content of the information, the patients' prior knowledge, and their level of anxiety. Generally information which is presented early in the consultation is recalled better (the primacy effect) as are statements which are perceived as being important or relevant. Moreover, recall is better for information which has been presented in an organised way based around specific themes (e.g., the nature of the problem, the details of the treatment, etc.).

4.2 *Patient Satisfaction*

Patient satisfaction is concerned with how patients evaluate the quality of their healthcare experience. It is increasingly being assessed in surveys of healthcare settings, as a marker of quality of care, along with other dimensions of quality such as access, relevance to need, effectiveness, and efficiency. Current research (e.g., Fitzpatrick 1997) views patient satisfaction as a multidimensional concept since patients differ in their views about specific aspects of their healthcare, such as the doctor's behavior towards them, the information provided, the technical skills of the doctor, and the access to and quality of the healthcare setting. Generally the most important factor is the behavior of the doctor since this can influence significantly ratings of all the other aspects of healthcare.

There are a number of difficulties in assessing patient satisfaction. Results from most satisfaction surveys reveal very skewed data, apparently indicating very high levels of satisfaction across patient samples, particularly amongst older patients. Part of this problem seems to lie in the reluctance of patients to criticise healthcare services and part of it is due to the structure of the questionnaires which have been used to assess patient satisfaction (Ware and Hays 1988). In addition, the use of more in-depth methods which require patients to describe their experiences of health from their own perspective, typically give rise to a more critical view.

A number of studies have attempted to examine the relation between indices of the consultation process and measures of patient and doctor satisfaction. For example, in a study based on their five distinct patterns of consulting style (see above), Roter et al. (1997) showed that the highest levels of doctor satisfaction were found in those using the consumerist approach and the lowest levels were found in those using the narrowly biomedical approach. In contrast, the highest levels of patient satisfaction were found with those who had seen doctors using the psychosocial communication pattern whereas the lowest satisfaction scores were recorded in those who had experienced either of the two biomedical patterns.

4.3 *Adherence to Advice or Treatment*

The most widely studied behavioral outcome from healthcare consultations is reflected in the extent to which the patient adheres to the advice or treatment offered by the doctor, and this topic is described in

much more detail in another article (*Patient Adherence to Health Care Regimens*). Many consultations result in the prescription of treatment or advice by the doctor and the appropriate adoption of self-care behaviors, including use of medicines, is a key aspect to the self-management of most chronic illnesses. The incidence of reported medication nonadherence varies from 4–92 percent across studies, converging at 30–50 percent in chronic illness (Meichenbaum and Turk 1987). Even patients who have experienced major health problems, such as heart attacks, may show low levels of uptake of rehabilitation programs as well as considerable variation in the adoption of recommended lifestyle changes (Petrie et al. 1996).

In the context of doctor–patient communication, one cause of nonadherence arises from the poor cognitive outcomes outlined above, particularly poor understanding and recall of information presented in the consultation. Many patients lack basic knowledge about their medication but the relationship between this and their adherence is complex. Although some studies have demonstrated a positive association between knowledge and adherence, there are more that had failed to demonstrate a link. Moreover, interventions which enhance knowledge do not necessarily improve adherence.

The quality of communication with the doctor can also influence patients' attitudes or beliefs about their illness or treatment which, in turn, have been shown to have significant effects on levels of adherence. Patients develop their own beliefs about the nature of their illness (Petrie and Weinman 1997) and these have been linked with a range of adherence-related behaviors other than the use of medicines. These include various self-management behaviors, such as dietary control and blood glucose testing in diabetes, attending rehabilitation, and the adoption of various lifestyle changes following myocardial infarction.

In addition to influencing patients' perceptions of their illness, communication with the doctor can affect patients' attitudes to their treatment. Work by Horne and Weinman (1999) has shown that patients develop very specific positive and negative beliefs about their prescribed treatment, both of which can have significant effects on treatment adherence. The positive beliefs are reflected in strong beliefs about the necessity of taking their medication in order to manage the illness whereas the negative beliefs reflect patients' concerns about treatment side-effects or long-term dependence.

5. *Interventions to Improve Doctor–Patient Communication*

The majority of communication training interventions have been aimed at improving the communication skills of medical students or practitioners at various stages of training but a few have been targeted at patients, to enable them to get maximum benefit from a consultation.

Communication skills training is now a core part of the curriculum for medical students. Typically, students are provided with an overview of the basic skills of 'active' listening, which include the importance of developing good rapport and the use of open-ended questions early in the consultation, appropriate eye-contact, and other verbal and nonverbal responses to help the patient talk, together with the ability to summarize and arrive at a shared understanding of the patient's problem (see Kendrick 1997). There is now consistent evidence that this type of training can result in enduring improvements in basic communication skills.

In addition to acquiring basic communication skills for work with all patients, medical students need to learn how to communicate about sensitive or difficult areas of clinical practice, including dealing with distressed patients or relatives and giving 'bad news.' Research in this area has shown that doctors tend to give detailed information rather than find out and respond to patients' concerns and informational needs. As Maguire (1997) notes, this avoidance of patients' worries can have negative effects in the short and longer term. An immediate consequence is that patients may remain preoccupied with their own concerns and fail to take in information or advice. They may also selectively attend to negative phrases or messages and be unresponsive or misconstrue more positive or neutral information.

Training in this area emphasises the value of patient-centered approaches, and exploring what patients know, think, and feel. Encouragement is given to take time to provide the patient with the information which they want, letting patients proceed at their own pace, and allowing information to be assimilated. The other key skills involve developing the ability not only to recognise distress but also to allow this to be expressed (Maguire 1997).

In recent years there have been a number of interesting interventions aimed at patients. Generally these have involved interventions for patients prior to a consultation in order to increase their level of participation, particularly to ensure that their own concerns are dealt with and that information provided by the doctor is clearly understood. A successful development of this approach can be seen in the work of Greenfield et al. (1985) who used a preconsultation intervention lasting 20 minutes for hospital outpatients who were helped to identify their main questions and encouraged to ask these in the consultation. Compared with control patients, these patients participated more actively in the consultation and this was also associated with better long-term health outcomes, including lowered blood pressure in hypertensives and better glycaemic control in diabetic patients.

6. Conclusions

Research on the nature, determinants, and effects of doctor–patient communication has revealed its complexity and its importance in effective healthcare delivery. Audio and videotaped recordings of the consultation have allowed the researcher to examine the nature of the interaction between doctor and patient. Although the tasks of coding and analysing this interaction have proved difficult, many studies have demonstrated that the interaction process can have major effects on a wide range of outcomes. These findings have been used for designing training in communication skills, and it will be a major task for researchers to evaluate the efficacy of these on the quality of healthcare delivery in the future.

See also: Coping across the Lifespan; Emotion in Cognition; Illness Behavior and Care Seeking; Patient Adherence to Health Care Regimens; Psychological Therapies: Emotional Processing; Therapist–Patient Relationship

Bibliography

Bower P, West R, Tylee A, Hann M 1999 Patients' perceptions of the role of the general practitioner in the management of emotional problems. *British Journal of Health Psychology* **4**: 41–52

Fitzpatrick R 1997 Patient satisfaction. In: Baum A, Newman S, Weinman J, West R, MacManus I C (eds.) *Cambridge Handbook of Psychology, Health and Medicine*. Cambridge University Press, Cambridge, UK

Frederikson L G 1993 Development of an integrative model for the medical consultation. *Health Communications* **5**: 225–37

Greenfield S, Kaplan S, Ware J E 1985 Expanding patient involvement in care: Effects on patient outcomes. *Annals of Internal Medicine* **102**(4): 520–8

Grol R, de Maeseneer J, Whitfield M, Mokkink H 1990 Disease-centered versus patient-centered attitudes: Comparison of general practitioners in Belgium, Britain and The Netherlands. *Family Practitioner* **7**: 100–4

Guadagnoli E, Ward P 1998 Patient participation in decision-making. *Social Science and Medicine* **47**: 329–39

Horne R, Weinman J 1999 Patients' beliefs about prescribed medicines and their role in adherence to treatment in chronic physical illness. *Journal of Psychosomatic Research* **47**: 555–67

Isen A, Rosenzweig A, Young M J 1991 The influence of positive affect on clinical problem solving. *Medical Decision Making* **11**: 221–7

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK

Kendrick T 1997 Teaching communication skills. In: Baum A, Newman S, Weinman J, West R, McManus I C (eds.) *Cambridge Handbook of Psychology, Health and Medicine*. Cambridge University Press, Cambridge, UK

Korsch B M, Negrete V F 1972 Doctor–patient communication. *Scientific American* **227**: 66–74

Ley P 1997 Doctor–patient communication. In: Baum A, Newman S, Weinman J, West R, McManus I C (eds.) *Cambridge Handbook of Psychology, Health and Medicine*. Cambridge University Press, Cambridge, UK

Maguire P 1997 Breaking bad news. In: Baum A, Newman S, Weinman J, West R, McManus I C (eds.) *Cambridge Handbook of Psychology, Health and Medicine*. Cambridge University Press, Cambridge, UK

Meichenbaum D, Turk D C 1987 *Facilitating Treatment Adherence: A Practioner's Guidebook*. Plenum, New York

Miller S M, Brody D S, Summerton J 1988 Styles of coping with threat: Implications for health. *Journal of Personality and Social Psychology* **54**: 142–8

Petrie K J, Weinman J (eds.) 1997 *Perceptions of Health and Illness: Current Research and Applications*. Harwood Academic, London

Petrie K J, Weinman J, Sharpe N, Buckley J 1996 Predicting return to work and functioning following myocardial infarction: The role of the patient's view of their illness. *British Medical Journal* **312**: 1191–4

Roter D L 1989 Which facets of communications have strong effects on outcome: A meta-analysis. In: Stewart M, Roter D (eds.) *Communicating with Medical Patients*. Sage, Newbury Park, CA

Roter D, Hall J A 1989 Studies of doctor patient interaction. *Annual Review of Public Health* **10**: 163–80

Roter D, Stewart M, Putnam S M, Lipkin M, Stiles W, Inui T S 1997 The patient–physician relationship: Communication patterns of primary care physicians. *Journal of American Medical Association* **277**: 350–6

Schwartz S, Griffin T 1986 *Medical Thinking: The Psychology of Medical Judgement and Decision*. Springer Verlag, New York

Ware J, Hays R 1988 Methods for measuring patient satisfaction with specific medical encounters. *Medical Care* **13**: 669–79

Williams S, Weinman J, Dale J, Newman S 1995 Patient expectations: What do primary care patients want from the GP and how far does meeting patient expectations affect patient satisfaction? *Family Practitioner* **12**: 193–201

J. Weinman

Documentary and Ethnographic Film

The term documentary came into usage in the nineteenth century as an adjective referring to documentation, hence with connotations of evidence (Rosen 1993, p. 6). With reference to film, it was first employed by John Grierson (in his 1926 review of Robert Flaherty's film *Moana*) to indicate a body of work which documented reality, involving 'a creative treatment of actuality' (Grierson, cited in Barnouw 1974, p. 287). The twin roles of the documentary filmmaker, as a creator with an agenda and as a purveyor of 'actuality,' have been at the root of several debates concerning the truth claims of the documentary. In many accounts, documentary film was regarded as indexically representing a real world, mimetically reproducing pictures of reality with a view to providing knowledge about it, as is the case in any scientific endeavor.

On the one hand, developments in motion picture technology have perfected hardware enabling the unobtrusive documentation of reality; on the other, the very notions of 'documentary' and 'reality' have undergone radical reformulation. In theory and practice, the truth claims of documentary and its status as 'evidence' of a given historical world have increasingly been problematized. Over time, documentary film has come to embody diverse modes of negotiating the complex set of relationships between the author, the film text, and the world.

Ethnographic film comes under the rubric of anthropology, which traditionally involved scientific observation and recording of the lives and cultures of other races, usually 'primitive' indigenous communities in distant lands. Jean Rouch terms anthropology as 'the eldest daughter of colonialism' (cited in Winston 1993, p. 51), thus foregrounding the inherent relationship of power that underpins any encounter between the anthropologist and the other. The premise that ethnographic film is a value-free, indexical scientific record has been subjected to critical interrogation and revision, leading to dramatic changes in its theory and practice.

1. *Demarcating the Terrain of Documentary Film*

1.1 *Documentary vs. Other Nonfiction*

Documentary film has been regarded as a specific body of work under the broader rubric of 'nonfiction' film. The latter includes genres such as the first 'actuality' films of the Lumiere Brothers, the newsreel, the travelogue, and the educational film. These genres, while sharing in common with documentary the use of actuality footage, have been seen as distinct because of the differing context of their production and use (Barsam 1992). Though documentary and ethnographic films occupy distinct terrains within nonfiction film, the issues confronting both share a great degree of commonality.

1.2 *Documentary vs. Fiction*

The fiction film, to which the documentary and other nonfictional genres were opposed, involved an obvious process of construction to offer an illusion of reality. The documentary filmmaker, however, bore the onus of providing a window to the world. The twin positions eschewed by the documentary filmmaker were that of the 'manipulative' creator of fiction, on the one hand, and the didactic educator on the other. The control and omnipotence represented by both these positions, which 'told' audiences about an external reality, was rejected in favor of a position that purported to 'show' its viewers the world.

2. *Towards a Taxonomy*

The wide array of themes, styles, formats, and modes of address employed by documentary filmmakers make it difficult to construct an inclusive typology that accounts for this variety. Barnouw (1974), for instance, chooses to classify documentaries on the basis of the role of the maker (explorer, reporter, advocate, poet, observer, guerilla, and so on), while emphasizing that these roles are not mutually exclusive. However, certain roles have been privileged at certain historical moments, making it possible to construct a typology on this basis. Barsam (1992) follows a classification employing time and space (American/British/European and Asian, at various periods of time).

Nichols (1991) identifies four modes of representation. The expository mode, used by many of the early filmmakers, such as Grierson and Flaherty, as also in current television network news, tends to use 'voice-of-God' commentary or poetic narration to inform or enlighten the viewer about the world outside. The argument of the commentary is what propels the text, the visuals serving more as illustration or counterpoint. The expectation of the viewer is generally to obtain certain, authoritative knowledge about reality, to understand cause and effect, problems and solutions.

The observational mode (also referred to as Direct Cinema) was facilitated by the invention of lightweight equipment with synchronous sound recording capability. It rejects the relatively didactic stance of the expository mode, in order to bring to viewers, life as it unfolds. The filmmaker aspires to the role of a detached observer. The expectation of the viewer is to gain access in an unmediated external world.

The interactive mode (represented by the work of Jean Rouch, Emile de Antonio, Michael Rubbo, among others) also draws on the availability of mobile equipment, but rejects this detachment in favor of a more interventionist role. The filmmaker's relationship with the subjects of the film comes to the fore and is not invisible, as in the observational mode. The focus is on local knowledge; the filmmaker is not confined to creating continuities in time and space, but could employ unusual juxtaposition, archival material, and other devices in constructing the narrative. The viewer's expectations are situated in the realm of the interpretative, since the mode makes no claim to absolute truth and neutral knowledge.

In the reflexive mode, the act of representing the world itself becomes central to the film's narrative. The premises from which the narrative is constructed and the process of construction are available to the viewer. While it bears some resemblance to the interactive mode, the reflexive mode, in addition to problematizing the relationship between the filmmaker and the subjects, also problematizes the relationship between the narrative/filmmaker and the viewer. The camera is not an evidence gathering

apparatus. The conventions of realism, when used, are interrupted and put to question. Broadly speaking, the reflexive mode draws on the poststructuralist critique of the indexical and instrumental view of language and the speaking subject (see *Semiotics*).

3. *A Brief History of Documentary and Ethnographic Film*

The theoretical positions and the narrative strategies within the discourse of the documentary film have undergone fundamental changes, making it impossible to establish continuity across them. In spite of this apparent diversity, they could perhaps be classified in terms of the notions of authorship and viewership that they are, implicitly or explicitly, putting forth or negating. A useful device in establishing intelligibility across the vast corpus of documentary and ethnographic film practice is, therefore, to look at the various authorial stances that inform it. This, by no means exhaustive, historical account explores the work of major figures and movements, which contributed to the development of new conventions, styles, and norms within documentary film.

3.1 *The Author as Reconstructing the Authentic Other—Robert Flaherty*

Flaherty's *Nanook of the North* (1922), which celebrates the heroic struggle for survival of its Inuit protagonist, could perhaps be regarded as the first ethnographic film. Flaherty manipulated reality in order to reveal what he regarded as the 'true spirit' of the culture; at times, this recreation involved danger and discomfort to the subjects of his films. In hindsight, his work has been critiqued for its mythologizing of the 'noble savage' and its elision of the 'real' social and political circumstances of his subjects' lives. Flaherty's work poses questions about the nature of 'truth,' the relation of the documentarist to the 'reality' s/he creatively appropriates, and the politics of representing the other.

3.2 *The Author as Social Educator—Grierson*

Grierson's extensive writings on the place, function, and codes of the documentary regard it as performing the function of educating and involving citizens in democratic civil society. In 1930, Grierson established the EMB (Empire Marketing Board) Film Unit, later to become the GPO (General Post Office) Film Unit. This was a nucleus for the work of documentary filmmakers such as Basil Wright (*The Song of Ceylon*, 1934, *Night Mail*, 1936, co-directed with Watt), Harry Watt (*North Sea*, 1938), Alberto Cavalcanti (*Pett and Pott*, 1934, *Coalface*, 1935) and Humphrey Jennings (*A Diary for Timothy*, 1944). The themes handled included the lives of working people, social problems related to health, housing and unemployment, and wartime propaganda.

3.3 *The Author as Reflexive Roving Eye—Dziga Vertov*

Denis Arkadievitch Kaufman called himself Dziga Vertov, meaning 'spinning top.' His writings and films were produced in the context of the post-Revolution ferment in the Soviet Union. His work, influenced by Futurism, represents a seminal contribution to film theory and practice, repudiating all pre-existing forms of cinema. Vertov regarded the cinematic way of seeing, the 'Kino-Eye' as establishing a radically different relationship with reality, making the 'invisible visible,' turning 'lies into truth,' propagating 'the communist decoding of reality.' His work on the Kino Pravda series, among others, and his film *The Man with the Movie Camera* (1929) created much debate and controversy for its formal illusionism, its subjectivism, its radical play with elements of picture and sound so that a sense of dislocation from reality is achieved. Its reinvention of cinematic language in a mode which reflexively foregrounds the act of construction and directly interpellates the viewer with revolutionary fervor stands as unique and perhaps far ahead of its time.

3.4 *The Author as Propagandist—Leni Riefenstahl*

Leni Riefenstahl is known for her documentaries that mythologized the politics of German National Socialism. While her work is noteworthy for its aesthetic virtuosity, it raises questions of the politics of aesthetics. *The Triumph of Will* (1935) documents the 1934 National Socialist rally in Nuremberg, a spectacular celebration of the myth of fascism. Sontag (1982) discusses how the actuality of a film like *The Triumph of Will* was constructed through carefully choreographed rituals of domination and enslavement, that contrast the hypnotic presence of the omnipotent leader with the faceless regimentation of large masses of ordinary people.

3.5 *The Author as a Fly on the Wall—Direct Cinema*

Direct Cinema represents a significant departure in the American documentary of the 1960s. The term Direct Cinema is sometimes used interchangeably with cinema verité (from which, however, it is distinct). It positions the author as an unobtrusive observer, thereby attempting to reclaim the scientific project of documentary film. The camera as well as the filmmaker occupy neutral positions, endeavoring, 'to convey [to the viewer] the feeling of being there' (Leacock, cited in Winston 1993, p. 43). The invention of lightweight

equipment with synchronized sound was first utilized by Drew Associates to produce *Primary* (1960) in the election campaigns of John F. Kennedy and Hubert Humphrey. Drew collaborated with Richard Leacock, Pennebaker, and Albert Maysles in the making of this film. The filmmakers had themselves perfected the technology and the now legendary shot of the camera following Kennedy through the crowds on to the podium, shot by Maysles, demonstrated the versatility of the new technology.

The then prevailing dominant empirical traditions of the scientific method, which took sides with the 'body,' in an age old debate on its antithetical position to the 'mind,' was at heart of the positioning that the proponents of Direct Cinema aspired to. The dichotomies of mind/body, subject/object held center stage, not only in the debates on documentary cinema, but also in the discourses of the social sciences and linguistics.

The truth claims of the proponents of the American Direct Cinema did not pass uncontested. In response to its critics, the movement of Direct Cinema had to allow for subjectivity, while retaining its claim to authenticity. Frederick Wiseman, who terms his work 'totally subjective,' represents a new denouement in the Direct Cinema debates. While the work of Leacock and Pennebaker focused on portraiture and recording of events, Wiseman's work investigates various institutions such as hospitals, high schools, and prisons, revealing patterns of behavior, which raise disturbing questions about institutional morality and social values. His first film *The Titicut Follies* (1969) was an indictment of an institution for the criminally insane and saw him embroiled in legal controversy. Wiseman does a tightrope walk between the subjectivity of the film-maker and the imperative to provide the viewer with fair evidence so that she/he can make up his/her own mind on the reality that is represented.

3.6 The Author as a 'Fly in the Soup'—Cinema Verité

Henry Breitrose (cited in Winston 1993, p. 53) uses the interesting analogy of a 'fly in the soup' to describe cinema verité (a term borrowed from Dziga Vertov). In contradistinction to Direct Cinema, cinema verité resolves the question of objectivity by situating the film-maker as an *agent provocateur*. The work of Jean Rouch and Edgar Morin typifies this stance. Rouch's *Moi, un Noir* (1959) represents a turning point in ethnographic film; in this film, the role of the camera shifts from a recording device to a precipitator of action. The film presents the inhabitants of Treichville, playing out scenes of their choice for the camera. After viewing themselves on screen, they reflect on the recorded events, reinterpreting them. This complex and novel process of representation explored a new relationship between the film-maker, his subjects, and the narrative. Rouch's project was rooted in an attempt to redefine the relationships of power in ethnographic film-making which had hitherto been a colonial gaze at the other. *Chronique d'une Été* (1961), in a reversal of ethnographic roles, turns the camera on 'the strange tribe that lives in Paris' (Morin cited in Winston 1993, p. 51). With all its reflexivity and radical revision of the project of ethnographic film, Rouch and Morin are unable to escape the hegemony of science and the promise of the image as evidence, if not of an objective world, then of the documentarist making a documentary (Winston 1993).

3.7 The Postcolonial 'Other' as Author

Indigenous documentary film production in the postcolonial context was sometimes state-sponsored (as in Cuba and India) and often projected the dominant agenda of modernity. Over time, the rise of independent documentary production gave voice to marginal groups and looked critically at larger political questions. The work of Fernando Solanas and Octavio Gettino in Argentina (*Hour of the Furnaces*, 1968), Patricio Guzman (*The Battle of Chile*, 1974–9), Santiago Alvares in Cuba (*Now*, 1965), Nana Mahomo in South Africa (*Last Grave at Dimbaza*, 1975), and Anand Patwardhan (*Father, Son and Holy War*, 1998) in India are but a few examples of a vibrant tradition of political engagement. The technology of video and the possibility of broadcast by international television networks have created a market for stories from the South. Many 'independent' documentary film-makers in the South are dependent on these markets for support. The market demand is for images that present the other as exotic, oppressed, or disaster-ridden. Raul Ruiz's *Of Great Events and Ordinary People* (1979) ironically comments on this endless dividing practice that sustains international documentary production and consumption.

The 'linguistic turn' within postcolonial ethnography, which regards the ethnographer as an author constructing narratives, has led to a rethinking of the relationship between the ethnographer and his/her informants. Since this perspective attempts to render apparent the process by which the narrative is produced, the informants in a sense become co-authors. Given the unequal power flows between the ethnographer and his/her informants, various modes of integrating the voices of the other in ethnographic texts have been explored. Some of these have involved the informants taking control of the medium to construct representations of themselves. The Navajo project, initiated in 1966 by Sol Worth and John Adair, was an experiment in teaching the Navajos, in Pine Springs, Arizona, to use the technology of film-making. The attempt was to understand the modes of representation that were at work in the narratives produced by the Navajos (Worth and Adair 1970). The extent to which such projects can redefine power flows is debatable, for the other continues to be the

destination of a scientific experiment, by the ethnographer.

Eric Michaels (1994) explores the space of alternative production and distribution by the Warlpiris. The project, initiated in 1985, involved local production and illegal transmission by the Warlpiri Media Association, at Yuendemu, in Central Australia. Michaels identifies significant differences between aboriginal and European creative practices. These pertain to 'ideological sources and access to inspiration, cultural constraints on invention and imagination, epistemological bases for representation and actuality, indistinctness of boundaries between authorship and oeuvre and restrictions on who makes or views expressive acts' (Michaels 1994, p. 105).

3.8 The Reflexive Author

The reflexive mode, already discussed earlier, gained prominence in the 1970s and 1980s. On the one hand, it was a response to the current debates on language and signification, subjectivity and ideology, involving formal strategies that question the 'materiality of the cinematic signifier' (Wollen, cited in Nichols 1991, p. 65). On the other hand, it arose in the context of movements (feminist, civil liberties, gay and lesbian, revolutionary struggles in the Third World and so on) that questioned the 'materiality of social practice' (Wollen, cited in Nichols 1991, p. 65). The films of Chris Marker (*Sans Soleil*, 1982), Errol Morris (*The Thin Blue Line*, 1987), Michael Moore (*Roger and Me*, 1989), and Jill Godmilow (*Far from Poland*, 1984) are few examples of emerging trends. Nichols (1991, p. 65) points out that reflexivity is not merely a matter of formal narrative strategies, but an engagement with the marginal and alternative ways of seeing and acting in the world. Examples of such politically reflexive films are Georges Franju's *Blood of the Beasts* (1949), Emile de Antonio's *In the Year of the Pig* (1969), Wiseman's *High School* (1968), and Joris Ivens' *The Spanish Earth* (1937).

The early independent women's films (*Union Maids*, 1976, *Joyce at 34*, 1972, and *Janie's Janie*, 1971) were situated within the realist tradition, and made as tools for consciousness raising and mobilization (Kaplan 1982–3). Semiological film theory tends to dismiss the radical potential of these films, a view which is contested by Kaplan, who argues that 'the same realist signifying practices can indeed be used to different ends' (Kaplan, cited in Nichols 1991, p. 68) and that a consideration of the strategies of reception is also important (see *British Cultural Studies*). Feminist films such as Mitchell Block's *No Lies* (1973) and Michelle Citron's *Daughter-Rite* (1978) problematize cinema verité techniques by drawing the viewer into an illusion of reality, only to reveal at the end that the narrative was scripted and the protagonists are actors.

Postfeminist work attempts to rewrite feminist themes to break away from the white/middle-class/heterosexual character of the women's movement in the West. Trinh T. Minh-Ha's *Surname Viet Given Name Nam* (1989), for instance, contests the systems of representation at work in a patriarchal ethnography, rejecting an essentializing of the heroic, suffering Vietnamese woman.

4. Current Issues

Debates regarding documentary and ethnographic film have revolved around various sets of relationships: between the author and the narrative (the question of realism), between the author and the subjects filmed (the question of the ethics and politics of representation), and between the narrative and audiences (the question of effects and audience reception). An issue that has received relatively less attention is the question of how mass media institutional dynamics, the economics of documentary production, and the structures of distribution have influenced the politics and aesthetics of documentary production (see *Media Imperialism*; *Television: Industry*). For instance, the fact that television constitutes a major market for documentaries today sets the agenda for documentary film production and its reception.

With new technologies such as digital image manipulation and inexpensive portable video camcorders that are redefining the realm of the real, the project of recording reality has never been more problematic. A variety of other genres, ranging from reality programming on television to fiction films, have begun to appropriate the idiom of the 'real,' blurring the divide between fictional and nonfictional narratives. When the technology for appropriating actuality in the Direct Cinema dream has become a reality, it is this very actuality that has become elusive, thus making imperative for the documentary and ethnographic film 'to negotiate an escape from the embrace of science' (Winston 1993, p. 57).

See also: Film and Video Industry; Film: Genres and Genre Theory; Film: History; Media, Uses of; News: General; Visual Anthropology; Visual Images in the Media; Visual Methods in the Social Sciences

Bibliography

Barnouw E 1974 *Documentary: A History of the Non-Fiction Film*. Oxford University Press, New York

Barsam R M 1992 *Non-Fiction Film: A Critical History*. Indiana University Press, Bloomington, IN

Crawford P I, Simonsen J K 1992 *Ethnographic Film: Aesthetics and Narrative Traditions*. Intervention Press, Aarhus

Ellis J C 1989 *The Documentary Idea*. Prentice-Hall, Engelwood Cliffs, NJ

Hardy F (ed.) 1966 *Grierson on Documentary*, Rev. edn. Faber, London

Kaplan 1982–3 Theories and strategies of the feminist documentary. *Millenium Film Journal* **12**: 44–67

Michaels E 1994 *Bad Aboriginal Art: Tradition, Media and Technological Horizons*. University of Minnesota Press, Minneapolis, MN
Nichols B 1991 *Representing Reality: Issues and Concepts in Documentary*. Indiana University Press, Bloomington, IN
Rabiger M 1998 *Directing the Documentary*, 3rd edn. Focal Press, Boston, MA
Renov M (ed.) 1993 *Theorizing Documentary*. Routledge, London
Rosen P 1993 Document and documentary: on the persistence of historical concepts. In: Renov M (ed.) *Theorizing Documentary*. Routledge, London
Rosenthal A 1988 *New Challenges for Documentary*. University of California Press, Berkeley, CA
Sobchack V (ed.) 1996 *The Persistence of History: Cinema, Television and the Modern Event*. Routledge, London
Sontag S 1982 Fascinating Fascism. In: Sontag S (ed.) *A Susan Sontag Reader*. Farrar, Straus, Giroux, New York
Tomaselli K G 1994 *Appropriating Images: The Semiotics of Visual Representation*. Intervention Press, Aarhus
Winston B 1993 The documentary film as scientific inscription. In: Renov M (ed.) *Theorizing Documentary*. Routledge, London
Worth S, Adair J 1970 Navajo filmmakers. *American Anthropologist* **72**(1): 9–34

K. P. Jayasankar and A. Monteiro

Domestic Life and Material Culture

1. *Definitions and Nomenclature*

Domestic life and material culture are terms used in both the social sciences and the humanities. The first is widely known; the second less so. The first serves as a covering term for a range of societal activities; the second represents a distinctive type of cultural evidence.

Domestic life includes those familial/societal interactions that take place within the context of a residential environment—be it a single family home, an apartment complex, a tribal lodge, or any other form of human dwelling. Hence the study of domestic life involves understanding child-rearing practices, kinship patterns, life-cycle history, *foodways*, gender divisions of labor, and use of household goods. It is a field (or subfield) for historians of the family and childhood, folklife researchers, economic historians, cultural anthropologists, as well as those who study architectural interiors and decorative arts.

Material culture study boasts an equally expansive research agenda and diversity of scholarly clientele. But, despite the fact that humans were making things before they were writing about them; that is, creating artifacts before alphabets, it is a relatively new term in scholarly discourse. Only in 1875 did British anthropologist A. Lane-Fox Pitt-Rivers, in an essay, *On the Evolution of Culture*, first urge researchers in the emerging social sciences to consider material culture as 'the outward signs and symbols of particular ideas in the mind.'

Anxious to be able to make more precise distinctions in their examination of human behavior as well as to widen the evidential base for their social research, some anthropologists and a few folklorists heeded Pitt-Rivers' injunction. They developed a tripartite division of cultural data. The first component was ideological (evidence found usually in the form of oral or written data). The second division was sociological (evidence documented by fieldwork observation of human behavior, such as parent–child or sibling relations). A third resource was material (evidence found in the work of human hands such as furniture, cooking utensils, and houses).

In both past and present scholarship, 'material culture' is often considered a synonym for terms such as 'artifacts,' 'objects,' or 'things.' Such interchangeability is frequently evident in both professional and popular writing. However, instead of treating these four terms as co-equals, it is more useful to regard them as nouns of increasing specificity of meaning, with 'things' seen as the most general rubric and 'material culture' as the most specific label.

For example, only in its tertiary meaning does the word 'things' connote a sense of inanimate entities distinguishable from living organisms. The word 'objects,' like the term 'things,' is another highly abstract collective noun, whereas 'artifacts' (artefacts in Britain), coming from the Latin *arte*, meaning skill, and *factum*, meaning something done, at least includes an indirect reference to a human being (an artificer) in its meaning. Objects lack this sense of implied human agency and, unlike artifacts, are commonly (and imprecisely) used to describe everything in both the natural and the man-made environment.

By contrast, the general definition of material culture specifically includes the factor of human artifice and also aptly circumscribes the scope of physical data within its research domain. What is useful, therefore, about the term 'material culture' is that it suggests, at least among its current students, a strong inter-relation between physical objects and human behavior. Despite its cumbersomeness, the phrase continually presses the researcher to consider the complex interactions that take place between creators and their culture. In other words, the assumption is that there is always a culture behind the material.

In recent scholarship, the terms 'material' and 'culture' have been extended in academic usage. The concept and content of their meaning now includes the objects of high culture (for instance, Charles Wilson Peale's painting of *The Peale Family*, 1773 or William Rinehart's sculpture of *Latona and Her Children*, 1874) as well as artifacts of popular and vernacular culture (for instance, a Los Angeles bungalow or a Louisiana shot-gun house). The term 'material' has also been

expanded. For example, James Deetz (*In Small Things Forgotten: The Archaeology of Early American Life*, 1977) sees cuts of meat as material culture, since there are many ways to dress an animal and the scientific breeding of livestock is human manipulation of a species. Kinesics (for instance, the body motions involved in performing Aaron Copland's *Appalachian Spring* ballet or a southern Indiana clog dance) and proxemics (for instance, the spatial uses and interactions of people at a Midwest home quilting bee, a Great Plains threshing crew noon dinner, or a courting couple in a Victorian parlor) are other culture expressions that assume, if but for a limited time, physical form, and could be included within the legitimate boundaries of material culture.

What might be an adequate (but not elaborate) working definition of material culture? Here is a brief one that informs this article: material culture is that segment of humankind's physical environment that has been purposely made or modified, consciously or unconsciously, by people according to culturally dictated plans. Material culture entails those artifacts that humankind creates to cope with the physical world, to regulate its social relations, to perform its work, to delight its fancy, and to create symbols of meaning. Material culture research can be best described at present as an expanding movement or coalition of individuals (Deetz calls it 'an emerging research community') working in museums, historical societies, and cultural agencies, as well as in academic departments in colleges and universities, who are intrigued by the idea of studying the (possibly unique) explanatory potential of material evidence as cultural meaning.

What, if anything, do the diverse practitioners of material culture research have in common? In addition to an obvious evidential concern with physical objects as cultural data, the movement works under at least three assumptions. Most studies of material culture usually involve fieldwork research during which artifacts are collected, identified, compared, and categorized either *in situ* or in museums, in private and public collections, and in historical agencies. A historical perspective also characterizes most material culture studies. As several of their names indicate—the history of technology, art history, historical geography—they seek to measure and understand change over time. Finally, in their analyses, most researchers use the culture concept in varying ways. While some see culture only with a capital C, the majority view it as an anthropological construct.

2. *Interconnections and Common Topics*

In what ways have analyses of domestic life intersected with material cultural studies? How has the recent scholarship of the latter contributed to the former? Conversely, how have contemporary trends in domestic life research set certain agendas in material culture inquiry? Finally, what paradigms or interpretive models have been most influential in the interaction of these two approaches to cultural study?

In order to answer, albeit briefly, these questions, three specific areas of common interest—residential spaces, household artifacts, women's roles—to both investigators are summarized here. The format of the three is deliberately bibliographical in order to provide a sampling of recent work. Examples cited are primarily Anglo-American, while the concluding entry bibliography emphasizes broader interpretations.

2.1 *Residential Spaces*

When John Demos devoted part one ('The Physical Setting') of his *Little Commonwealth: Family Life in Plymouth Colony* (1970) to the domestic shelter and artifacts of the seventeenth-century Pilgrim community, he was recognizing a fact crucial to much material culture research: as the elementary unit of humankind is the individual, the elementary artifact on the human landscape is the dwelling. Housing represents social and economic identity; it is a microcosm of domestic culture. Demos explored the domestic implications of architectural evidence (for example, its correlation to economic status, privacy, social-segregation, repression of familial anger and aggression, and child-rearing practices) in his pioneering application of social and behavioral science concepts to the houses of seventeenth-century Massachusetts.

Although Demos' analysis of the material culture of a single seventeenth-century New England colony was brief, inferential, and only a case study, it does represent one type of material culture that has employed housing as significant evidence. Others include Gwendolyn Wright's *Building The Dream: A Social History of Housing in America* (1981) and Richard Bushman's, *The Refinement of America: Persons, Houses, Cities* (1992). Such research eschews the usual approach of the traditional architectural historian in that it avoids mere facadism (interpreting a house primarily through its front elevation and aesthetic style) as well as elitism (researching only structures designed by professional architects).

Housing sites, as well as the housing, has always been an important material culture subfield. Other scholars have investigated the history of urban middle- and working-class neighborhoods through their extant housing stock. Tenement districts, slums, and even alley residences have received attention. Zoning and property controls, residential class segregation, and subsidized public housing have also been topics.

Folklorists studying the construction techniques involved in erecting and furnishing housing often explore changes in a community's economic and social organization of work through building practices. In two pioneering books—*Pattern in the Material Folk Culture of the Eastern United States* (1968) and *Folk Housing in Middle Virginia: A Structural Analysis of*

Historic Artifacts (1975)—Henry Glassie has researched such historical processes in order to understand the persons who employed them and the products that resulted from them. In *Folk Housing*, Glassie approached material culture by means of the paradigm of twentieth-century structuralism, particularly as articulated by Noam Chomsky and Claude Lévi-Strauss.

Glassie, following Chomsky, argued that 'culture is pattern in mind, the ability to make things like sentences or houses.' Rejecting the concept that objects are the simple products of passive minds, Glassie attempted to develop a systematic model that would account for and help analyze the design abilities of an idealized maker of artifacts. As a case study, he investigated the builders of 156 houses constructed between the middle of the eighteenth century and World War I in a 70-square-mile (182-square-kilometer) area of central Virginia. His research objective, using almost exclusively material culture data left behind by anonymous builders, was to discover the unwritten boundaries or 'artifactural grammar' of the creative process as it was exercised in a particular region's domestic architecture.

American material culture structuralism has drawn criticism from various camps. Some historians claim that despite its gestures toward tracing a change in minds, the approach tends only to work in areas of relative cultural stasis. Fellow folklorists have complained that Glassie's subjective system of binary mental opposites (intellect–emotion, internal–external, complex–simple, and 12 others) merely substitutes one kind of interpretive arbitrariness for another in plotting human behavior patterns. Still others remark that the results from such structural analysis are not adequately comparative.

2.2 Household Artifacts

In keeping with the folk proverb that claims that a house is not a home, historians have researched the material life of Americans as revealed in patterns of home furnishings, foodways, clothing, and organization of domestic spaces as ways of gaining insight into the social past of middle-class and working-class culture. For instance, Lizabeth Cohen's interpretation of the material culture of American working-class homes between 1885 and 1915 ('Embellishing a Life of Labor: An Interpretation of the Material Culture of American Working-Class Homes' in Thomas J. Schlereth, ed., *Material Culture Studies in America*, 1982, 1992) is an example of how the study of the domestic material culture environment can contribute to the historical understanding of the familial and residential experience of Americans.

Some material culture researchers in the decorative arts have sought to expand the standard cultural history surveys of American domestic life by studying a whole range of common household items—parlor furniture, mourning pictures, eating utensils, cleaning devices, family photograph albums, kitchen appliances—that are seen as indices of a society's values as important as its elite artifacts or its literary remains. The material culture of nineteenth-century hallways (hat and clothing stands, mirrors, hall chairs, and card receivers), parlors (especially its ubiquitous home parlor organ), and dining rooms (table settings, massive sideboards, and eating accoutrements) have been imaginatively analyzed by Kenneth Ames (*Death In the Dining Room and Other Tales of Victorian Culture*, 1992), while Edward O. Laumann and James S. House ('Living Room Styles and Social Attributes: The Patterning of Material Artifacts in a Modern Urban Community,' *Sociology and Social Research* **54**: 1970) have studied the patterning of material artifacts in the living rooms of twentieth-century working-class families. Still others have attempted to write American social history using wallpaper, silver, ceramics, chairs, and other commonplace domestic artifacts as important evidence in their research. Several of these studies have also employed quantification techniques in their comparative analyses of past material culture. Such data are found in the form of probate inventories, craftsmen's ledgers, auction lists, wills, deeds, and sales records. Historians of Colonial America were the first to turn to inventories to study social and economic behavior and they have subsequently contributed a substantial corpus of material culture research based on inventories.

Photography, an artifact created in the nineteenth century, has also been of immense use to historians interested in American domestic life. Historical photography, when examined closely and in sufficient quantity to ensure a representative evidential sample, provides valuable inferences as to how occupants organized and used space as well as how they interacted with one another. Using photography, much can be learned about increasing compartmentalization of residential spaces such as kitchens, bedrooms, hall passages, pantries, inglenooks, nurseries, servants' quarters, as well as in parlors, living rooms, porches (front and back), plus lawns, patios, and garages.

Strangely, few material culture scholars have paid much research attention to eating, assuredly one of humankind's most common necessities. What research has been done focuses primarily on rural and preindustrial communities. It has also focused largely on artifacts of twentieth-century food preparation, service, and disposal.

In calling for the integration of domestic foodways (and all other relevant material, statistical, and documentary data) into a cultural, historical, and anthropological approach to material culture, James Deetz has argued for a research paradigm that seeks 'the detection and explication of apparently unrelated changes in *all* [Deetz's emphasis] aspects of a people's culture, material and otherwise.' Although not directly

influenced by the devotees of a *mentalité* approach to the past, he shares this school's interest in the study of popular beliefs, attitudes, customs, sentiments, and modes of behavior as well as its commitment to 'thick description.' Deetz, a historical archaeologist long associated with the historical reconstruction of Plimoth Plantation (Massachusetts) developed his theory and method in *In Small Things Forgotten and has now continued it in Flowerdew Hundred: The Archaeology of a Virginia Plantation* (1993).

Of particular interest is the challenge Deetz's paradigm presents in *In Small Things Forgotten* to the traditional interpretive claim of political and diplomatic history that the American independence movement was the major cultural watershed of American colonial history. The domestic material data (house forms, furniture types, ceramics, funerary arts, musical instruments, refuse dumps) that Deetz assembled and integrated suggest that the Revolution actually had little impact on American cultural history; in fact, the general American cultural pattern of involvement with English material culture, both before and after the war, remained remarkably constant.

In some ways, Deetz's *In Small Things Forgotten* can be viewed as an American counterpart to Fernand Braudel's *Capitalism and Material Life* (1973). For example, although much briefer and lacking Braudel's extensive documentation, the Deetz interpretation, like Braudel's, stays close to home for its data. The changing domestic technologies of house building, heating, lighting, plumbing, food preparation, and garbage removal are resources in both histories. Extant artifacts are often the only data by which such past behavior can be reconstructed. For example, utility fixtures have been carefully studied in order to speculate on how innovations in such domestic technology drastically altered the traditional evening orientation (a shared communal space around a single hearth) of the pre-nineteenth-century family, thereby prompting major changes in parent–child and sibling relationships. Others have documented behavior changes in domestic history with the advent of indoor plumbing (especially the appearance of the bathroom) and the installation of central heating.

2.3 Women's Roles

Understandably, the material culture of domestic life has figured in the work of historians of women. A special interest among some such scholars in the decorative arts and in the history of technology has focused on kitchen tools and appliances and their roles in defining, confining, or undermining a 'woman's place.' While various researchers have contributed to this scholarly enterprise, the work of Ruth Schwatz Cowan, Dolores Hayden, and Susan Strasser aptly represent their general concerns.

Cowan's research has moved from the study of a single artifact genre (*A Case of Technology and Social Change: The Washing Machine and the Working Wife* 1974) to kitchen appliances in general ('The "Industrial Revolution" in the Home,' *Technology and Culture*, 1973) to women's interaction with technological material culture throughout American history (*More Work for Mother: The Ironies of Household Technology From Open Hearth to Microwave*, 1983). In an essay titled 'Virginia Dare to Virginia Slims: Women and Technology in American Life,' *Technology and Culture* (1979), Cowan points out that none of the standard histories or bibliographies of America technology contain adequate reference to such culturally significant artifacts as the baby bottle. Here is a simple implement that, along with its attendant delivery system, revolutionized a basic biological process, transformed a fundamental human experience for vast numbers of infants and mothers, and served as one of the more controversial exports of Western technology to underdeveloped countries.

According to Cowan, there is a host of questions that scholars might reasonably ask about the baby bottle. For how long has it been part of Western culture? When a mother's milk could not be provided, which classes of people used the bottle and which the wet nurse, and for what reasons. Which was a more crucial determinant for widespread use of the bottle, changes in milk technology or changes in bottle technology? Who marketed the bottles, at what prices, to whom? How did mothers or different social classes and ethnic groups react to them? Can the phenomenon of 'not enough milk,' which was widely reported by pediatricians and obstetricians in the 1920s and 1930s be connected with the advent of the 'safe' baby bottle? Which was cause and which effect?

Hayden's interests have been more spatial and environmental than technological. Her *Seven American Utopias: The Architecture of Communitarian Socialism, 1790–1975* (1976) details the social and cultural history of several countercultural societies through their extant structures and sites, as well as through their furniture and household technology. Hayden's *The Grand Domestic Revolution: A History of Feminist Designs for American Homes, Neighborhoods, and Cities* (1981), provides the social historian with a useful review of the interaction of domestic feminism, cooperative housekeeping, the community kitchen movement, and projects for social reform.

Susan Strasser's *Never Done: A History of American Housework* (1982) has set the agenda for future research into basic topics such as food preparation, housekeeping and housecleaning, and domestic consumerism. On the last topic, Strausser has followed up on her own work and written books on the material culture of advertising, particularly its orientation toward women consumers and on the history of trash.

How might the nexus of domestic life and material culture appear in the near future? One prediction is that the three subfields outlined above will continue to thrive. In each research area there are scholars

experimenting with new methodologies as well as widening the boundaries of traditional evidential data.

For example, a few individuals in organizations such as the Vernacular Architecture Society (Great Britain) and the Vernacular Architecture Forum (United States) now study prefabricated and manufactured housing, condominiums, and assisted-care residences for the elderly. Some decorative arts historians are venturing beyond museum period-room settings as well as the old prescript maintained by dealers, collectors, and the US Customs office that 'one hundred years doth an antique make,' and are examining post-World War II kitchen remodelings, family rooms, the changing functions of the garage, plus patios and decks. Within the interdisciplinary gender studies movement, investigators have turned their attention to the social and material culture of Tupperware parties, to the current popularity of replica/reproduction nineteenth-century household goods, and the artifacts (computers, fax machines, telemarketing devices) of an increasing number of home offices serving as career centers for a variety of occupations.

Two, relatively recent, scholarly perspectives should influence future work on domestic life research and material culture studies. First, the re-emergence of anthropology, a parent-field for material culture, will strength both areas, particularly in paradigm building and in social science rigor. Material culture sessions at the International Congresses of Ethnological and Anthropological Sciences suggest this trend. It has also been argued that consumerism studies may become the most noticeable cynosure of contemporary material culture research. Perhaps. The twentieth-century data appears omnipresent and continually proliferating. While some basic research has been done over the past several decades, the full impact of door-to-door sales, general/specialty merchandising, department, chain, and franchise stores, shopping centers, mail-order cataloging (perhaps reaching more homes now than in the nineteenth century), suburban and megamalls on domestic life awaits at least a generation of researchers.

Whatever the future configurations of those who explore the intersections between domestic life and material culture, they will most probably share common cause in the vernacular, the populist, the typical; and in the desire to make human experience, as the novelist William Makepeace Thackeray once hoped 'more familiar than heroic.'

See also: Family and Kinship, History of; Gender and Place; Gender and Technology; Household Archaeology; Household in Anthropology; Photography as a Medium; Residential Concentration/Segregation, Demographic Effects of; Residential Environmental Psychology; Residential Segregation: Geographic Aspects

Bibliography

Ames K L 1992 *Death in the Dining Room and Other Tales of Victorian Culture*. Temple University Press, Philadelphia, PA

Braudel F 1973 *Capitalism and Material Life: 1400–1800*. Harper and Row, New York

Bushman R L 1992 *The Refinement of America: Persons, Houses, Cities*. Knopf, New York

Cohen L 1982, 1992 Embellishing a life of labor: an interpretation of the Material culture of American working class homes. In: Schlereth T J (ed.) *Material Culture Studies in America*. American Association for State and Local History, Nashville, TN

Cowan R S 1983 *More Work for Mother: The Ironies of Household Technology from Open Hearth to Microwave*. Basic Books, New York

Demos J 1970 *Little Commonwealth: Family Life in Plymouth Colony*. Oxford University Press, New York

Glassie H 1968 *Pattern in the Material Folk Culture of the Eastern United States*. University of Pennsylvania Press, Philadelphia, PA

Glassie H 1975 *Folk Housing in Middle Virginia: A Structural Analysis of Historic Artifacts*. University of Tennessee Press, Knoxville, TN

Grier K 1988, 1995 *Culture and Comfort: People, Parlors and Upholstery, 1850–1930*. Strong Museum, Rochester, NY

Hayden D 1976 *Seven American Utopias: The Architecture of Communitarian Socialism, 1790–1975*. MIT Press, Cambridge, MA

Hayden D 1981 *The Grand Domestic Revolution: A History of Feminist Designs for American Homes, Neighborhoods and Cities*. MIT Press, Cambridge, MA

Laumann E O, House J S 1970 Living room styles and social attributes: the patterning of material artifacts in a modern urban community. *Sociology and Social Research* **54**: 22–31

Miller D 1987 *Material Culture and Mass Consumption*. B. Blackwell, New York

Schlereth T J 1980 *Artifacts and the American Past*. Altrima Books, Walnut Grove, CA

Schlereth T J (ed.) 1985 *Material Culture: A Research Guide*. University of Kansas, Lawrence, KA

Schlereth T J 1992 *Cultural History and Material Culture: Everyday Life, Landscapes, Museums*. University of Virginia Press, Charlottesville, VA

Strasser S 1982 *Never Done: A History of American Housework*. Pantheon Books, New York

Wright G 1981 *Building the Dream: A Social History of Housing in America*. Pantheon Books, New York

T. J. Schlereth

Domestic Violence: Sociological Perspectives

The issue of 'domestic violence' is an important area of public, political, and academic concern that goes to the heart of the institution of the family and marriage and of gender relations between men and women. It encompasses public and private life as well as social,

political, and economic institutions. At the social level, it involves ideological beliefs and institutional policies and practices. At the individual level, it involves personal attitudes and prejudices as well as individual behavior and conventions of daily life.

The area is characterized by a history of recent, rapid, social change in institutional policy and practice and is steeped in controversy. There have been political, ideological, and academic controversies about the term itself, the definition of the phenomenon, the nature of the problem, research methods used in its study, findings about the nature and occurrence of the problem, and about the focus and effectiveness of social and legal interventions. Since its rediscovery in Britain and the USA in the early 1970s, virtually every aspect of this issue has been the subject of development and the object of debate not only within the academic community but also across a wide spectrum of society as the media, community activists, academic researchers, and policy makers have taken different positions and entered into sustained and sometimes heated debates. The lively nature of this very important social issue warrants further examination of each of the issues touched upon within this characterization which, by virtue of its it brevity, can only introduce each topic and provide a broad view of it.

1. *Definition*

The terms wife beating, battered wives, wife abuse, woman abuse, spouse abuse, partner violence, family violence, and domestic violence have all been used to describe this phenomenon and have themselves been the topic of debate and controversy. Two main issues have been at stake: (a) the terms reflect different conceptions of intimate relationships between men and women, with some terms reflecting the idea of formal marriage while others encompass long-term intimate relationship without the formal sanction of marriage; and (b) the terms reflect different conceptions of the phenomenon, as male violence, mutual violence, or female violence. Both of these general issues were the subject of considerable debate when the problem was rediscovered, described, and named. While the term 'wife abuse' contained a clear imagery of men's violence against a woman partner, it also contained notions of formal marriage as a necessary condition of concern and attention. By contrast, the term 'domestic violence' included the notion of state and nonstate sanctioned relationships but lost the conception of gender asymmetry in the perpetration of the violence. For reasons of popular usage rather than resolution of these debates, the term 'domestic violence' became the term in common usage throughout most of the world.

2. *History*

Very ancient customs, beliefs, and laws dating back to the early Roman empire and many of the first conventions of marriage, allowed a husband to kill his wife for infidelity and serious challenges to male authority, these 'rights' were slowly eroded over time through the development of Canon law and the gradual alteration of legal systems in Europe and Britain to punishments allowing chastisement that did not kill or maim. In 1765, Sir William Blackstone's *Commentaries on the Laws of England* noted that conventional cultural beliefs and common legal practices endorsed a man's right to use 'moderate' chastisement against his wife. Blackstone's Commentaries codified this patriarchal right and thus confirmed a husband's authority over his wife for a considerable period of time. Such ideals and legal views crossed the Atlantic with some of the pioneers who brought familiar social attitudes and legal systems to the newly forming nations of North America. While there were some departures from this position it, nonetheless, constituted the normative position in many European and North American cultures across a fairly long period of time. Indeed, contemporary evidence from a wide range of societies, reveals strong cultural beliefs expressing such ideals, often endorsed by legal practices which illustrate direct support for 'wife-beating' or at best weak reactions to its victims and perpetrators.

In European and North American societies, the history of challenges to and changes in social and institutional responses to the problem, is punctuated by three historical periods in which domestic violence was discovered anew and obtained widespread recognition as a social problem. Recognition eventually resulted in changes in the laws regulating the institution of marriage and the family as well as developments in the responses of agencies of the state, particularly law and law enforcement, but also in economic, political, and social institutions.

In the 1870s, 1910s, and 1970s, the problem of 'wife beating' was discovered by women in Britain and North America who were engaged in more general campaigns for changes in women's marital, economic, political, and/or social status. Over this time period, wider campaigns about changes within the family and marriage (such as, the ability of women to own their own property, or to obtain a divorce for excessive cruelty—1870s), about changes in the political status of women (the right to vote—early 1900s), about changes in the economic status of women (equal pay and equal opportunities within the job market—1970s) and changes in social relations between men and women in general each spawned an interest in the specific issue of physical violence against women within the home, usually by intimate, male partners/husbands within the context of marriage. At each of these junctures, radical community groups (usually

though not always feminist) were formed, there were public protests, public recognition was increased, the agencies of the state were challenged and responded with public hearings, legislation, and various new policies and practices. After the first two periods of recognition and public response, the issue was soon forgotten and was marked by a return to an older, established status quo in which men maintained social, political, and economic authority over women throughout society and particularly within the institution of family and marriage where this right could be exercised legally and morally through the use of violence against women/wives for various real or perceived violations of male authority.

The third period of challenge began in the 1970s, again in Britain and the USA and, rather than being forgotten after a brief period of interest, has instead continued to expand and develop throughout much of the world. Public awareness of the issue and community and state responses to it are now global, resulting in violence against women being defined by the United Nations as an issue of human rights and social justice. At present, many, although certainly not all, societies have experienced challenges to the complacency and social and institutional tolerance that once prevailed and in many countries new laws and institutional practices have been initiated.

3. *The Nature of the Problem, Research Methods, and Findings*

Controversies about the nature of the problem itself have been intense and are intertwined in other controversies about research methods and resulting findings. The debates regarding these issues were apparent at the onset of the rediscovery process and have altered little since, although the amount and type of research grew at a tremendous rate in the last quarter of the twentieth century. Debates about the phenomenon itself center on whether the violence is asymmetrical in nature (men as the most usual perpetrators, with women the usual victims) or symmetrical (women and men as equally violent toward one another). Debates about research methods have focused on whether one method should supersede all others in producing the 'best' or indeed the only findings of value in answering the question of 'who is violent to whom.' One camp has claimed variously that only surveys based on national probability samples produce valid and reliable findings. Research using this approach has focused on questions of the prevalence of this form of violence in the population and on the relative levels of violence of men and women. Others defend the use of a wide variety of research strategies which include historical studies showing continuities and changes in patriarchal beliefs and institutional practices over time and ethnographic studies with an intensive focus upon cultural practices as well as the more traditional sociological approaches of victim surveys, criminal statistics, intensive interview studies, studies of strategic sites, and evaluations of new interventions. Proponents of the broad- based sociological and contextual approaches claim that such studies address more pressing and relevant questions about the problem such as 'what is the exact nature of the violence,' 'how are victims and offenders processed by agencies of the state,' and 'what works at stopping violence.' Such investigations also attempt to examine in detail the contextual dynamics of violent episodes and violent relationships as well as the meaning and consequences of being a victim and/or perpetrator.

Survey research, particularly in the US, has relied primarily on a single measurement tool that only attempts to assess violent acts with little or no concern for the intentions, meanings, and consequences associated with these acts. The resulting findings of these assessments appear to demonstrate symmetry in the perpetration of domestic violence, with women equally or more violent than their male partners. A corollary of the claims of symmetry in domestic violence is that theoretical accounts will be gender neutral, based on the conclusion that gender is not related to differential rates of violence within intimate relationships. Other surveys and more intensive research methods, not based on these purely act-based measurements, whether used in the US or elsewhere, show asymmetry in domestic violence, with men the most likely perpetrators and women the most likely victims. Predictably, these contradictory findings have resulted in considerable controversy with claims and counterclaims regarding the validity and reliability of this single type of measurement and scepticism regarding the veracity of findings produced by this type of research.

Over time, controversy about the simple question of 'symmetry–asymmetry'—who perpetrates the majority of violent acts (men or women)—has become more complex with the addition of further research questions regarding who is most likely to be responding in self-defense, to be harmed, to live in fear, and to experience other consequences of violence as well as additional research efforts to differentiate violent acts from acts of verbal aggression or mere conflicts between partners.

Despite the controversies in the area of domestic violence, there is a growing body of evidence-based literature that identifies strong and consistent patterns across cultures and historical periods. Most researchers and policy makers now agree that the problem of domestic violence is primarily one of men's violence against women in state as well as nonstate sanctioned relationships. Evidence suggests that serious, malevolent violence is perpetrated largely by men and often results in serious injuries. Survey research in several

countries indicates that at least a quarter of all women who have lived in marital and marital-like relationships with men report at least one incident of violence and a considerable proportion of these women report repeated violence. It is now clear that men's violence also includes acts of sexual aggression and rape; research indicates that a fair proportion of women experience forced sexual relations and rape by intimate male partners, and women who experience physical assaults are also likely to experience sexual violence. Homicide statistics further suggest that women are most at risk of lethal violence from male partners; research in Britain and North America reveals that every year 40 to 50 percent of all women who are killed, are killed by intimate partners and ex-partners. By contrast, very few men are killed by their women partners. Men's violence directed at women in the domestic setting also is linked intrinsically to other techniques and strategies of intimidation and control, such as threatening violence and restricting and/or closely monitoring women's activities.

Intensive and extensive research has identified the risk factors associated with lethal and nonlethal violence against women in the home. Although this violence occurs at all levels of society, higher rates are apparent among the economically and educationally disadvantaged. Young women and those living in non-state sanctioned relationships are more at risk than older and married women. Situational risks include high levels of conflict, jealousy, sexual proprietariness among male partners, and alcohol abuse. Risk of serious nonlethal and lethal violence is also associated with separation and women's attempts to leave violent relationships. In the 1970s investigations of policy and institutional practices revealed considerable failure on the part of legal and social agencies to provide support and protection for victims and children and a considerable reluctance to deal with male perpetrators. More recent research has charted the changes in policy and practices and revealed that some developments have resulted in decreased risks of violence and important benefits to women and children.

4. *Theoretical Accounts*

Theoretical explanations of domestic violence predictably have included a plethora of competing, at times contradictory, accounts, ranging from the sociocultural to the individualistic. Sociological, usually feminist, explanatory frameworks stress the historical and contemporary existence of male violence across almost all societies and locate its perpetration within patriarchal ideals and practices, emphasizing male power and control. In these accounts, violence is seen as functional, a resource used in men's attempts to enforce and maintain their domination and control, with conflicts associated with male jealousy, household labor, and economic resources constituting the contexts of violent events. Other sociologically-based gender-neutral accounts emphasize power but note that this social attribute is not skewed toward either gender—men and women possess significant amounts of power—and the symmetrical violence which occurs is an expression of such power. Variants of 'gender-neutral' explanations of domestic violence emphasize the importance of individual psychology and/or social traits that are deemed to have similar effects on the violence of men and women alike, these include, stress, frustration, socialization into (sub)cultures of violence, poor anger control, and impoverished social skills. Evolutionary psychology has challenged the foundations of these sociocultural explanations.

While evolutionary approaches also stress the functional nature of men's violence and locate it within patriarchal structures, explanations of this violence emphasize a universally evolved male psyche involving male sexual proprietariness toward women. This evolved psyche leads men to attempt to control women through various means including violence. Other, individualistic accounts have emphasized the pathological personalities of violent men.

An important strand in explanatory approaches is attempts to understand the predicament and actions of women who live with violent men. Sociological accounts have emphasized how male violence includes a range of intimidating and controlling acts as well as concerted actions to isolate women from family and friends. Continuous forms of intimidation and isolation make it very difficult for women to seek help, although it is now clear that most women suffering domestic violence engage in active efforts to deal with the violence. The vast majority seek help from family and friends and some also seek assistance from agencies of the state. Despite these efforts, women find it difficult to leave and sociocultural explanations stress men's threats of further violence, women's lack of economic and social support, a moral order emphasizing women's obligations to maintain family unity, and, until very recently, the inadequate responses of agencies of the state. Other commentators, particularly in North America, have proposed the importance of personality disorders in women as the primary reason that they are unable to leave violent relationships. According to these theoretical accounts women who do not leave violent relationships suffer from masochism, relationship addiction, 'battered woman' syndrome, and learned helplessness.

By the 1990s and into the 2000s, the enhanced and refined nature of the research questions, along with an increasing concern to evaluate the relative effectiveness of new innovations further expanded the theoretical questions in this area and, consequently, the relevant research strategies. Future research in the area will continue to concentrate on efforts to explain domestic violence and the predicament of the women who experience it as well as more focused questions about the impact of social and legal interventions on the

safety of victims and the violence of men. While much of the earliest research was shaped within sociology, criminology, women's studies, and law, the global expansion of interest in this social problem has been mirrored by an equal expansion of disciplines contributing to its study. In the future research, disciplines such as psychology, healthcare, and medicine will contribute to advances in knowledge about this significant social problem.

See also: Divorce and Gender; Family and Gender; Family Bargaining; Family Law; Family Processes; Rape and Sexual Coercion; Violence and Effects on Children; Violence and Media; Violence in Anthropology; Violence: Public

Bibliography

Bachman R, Saltzman L E 1995 *Violence Against Women: Estimates from the Redesigned Survey*. US Department of Justice, Washington, DC

Campbell J C, Soeken K L 1999 Forced sex and intimate partner violence. *Violence Against Women* **5**: 1017–35

Counts D A, Brown J K, Campbell J C (eds.) 1992 *Sanctions and Sanctuary*. Westview, Boulder, CO

Dobash R E, Dobash R P 1979 *Violence Against Wives*. Free Press, New York

Dobash R E, Dobash R P 1992 *Women, Violence and Social Change*. Routledge, London and New York

Dobash R E, Dobash R P, Cavanagh K, Lewis R 2000 *Changing Violent Men*. Sage, Thousand Oaks, CA

Dobash R P, Dobash R E, Wilson M, Daly M 1992 The myth of sexual symmetry in marital violence. *Social Problems* **39**: 71–91

Gondolf E W 1997 Batterer programs: What we know and what we need to know. *Journal of Interpersonal Violence* **12**: 83–98

Gordon L 1988 *Heroes of Their Own Lives*. Viking, New York

Johnson H, Sacco V F 1995 Researching violence against women: Statistics Canada's national survey. *Canadian Journal of Criminology* **37**: 281–304

Mirrlees-Black C 1999 *Domestic Violence: Findings from a New British Crime Survey*. Home Office, London

Pleck E 1987 *Domestic Tyranny*. Oxford University Press, Oxford, UK.

Romkens R 1997 Prevalence of wife abuse in the Netherlands. *Journal of Interpersonal Violence* **12**: 99–125

Straus M A, Gelles R J (eds.) 1990 *Physical Violence in American Families*. Transaction, New Brunswick, NJ

Tjaden P, Thoennes N 2000 Prevalence and consequences of male-to-female and female-to-male intimate partner violence as measured by the National Violence Against Women Survey. *Violence Against Women* **62**: 142–61

Wilson M, Daly D 1998 Lethal and non-lethal violence against wives and the evolutionary psychology of male sexual proprietariness. In: Dobash R E, Dobash R P (eds.) *Rethinking Violence Against Women*. Sage, Thousand Oaks, CA, pp. 199–230

R. E. Dobash and R. P. Dobash

Dramaturgical Analysis: Sociological

Dramaturgical analysis of social interaction is based on the assumption that social acts are staged, consciously or unconsciously, and thus embody all the elements found in enactments in the theater. Plays on the theater stage highlight all the elements of ordinary social life to present to an audience a new perspective on some aspect of social interaction. Thus, concepts used in theater production can be turned back again for the analysis of the social behavior that they are designed to reflect.

1. The Main Concepts

The main concepts that can be used in dramaturgical analysis center on the stage or action area (Hare and Blumberg 1988). The action area can be divided into two parts; backstage and stage. The backstage area is where the actors prepare for their roles and where special effects are produced to influence the audience. The stage is where the action takes place in full view of the audience. Even if only two people are involved in the interaction, they take turns providing an audience for each other. When one is speaking the other is listening. As an audience, an actor provides consensual reality, cueing, social reinforcement, and continual observation (Sarbin in Allen and Scheibe 1982).

For any action area there may be offstage areas in which persons who have organized the activity (producers), as well as those who have rehearsed the cast (directors) and are providing cues for action remain hidden from the audience. In addition, there may be a person or group who provided the original idea or script for the performance (playwright). This does not complete the list as there may also be persons who cater for the audience.

The meaning of the event binds all the participants together. The overall meaning has been termed the definition of the situation, the frame, and the illusion. In the progress of the development of social interaction over time, agreement on an actable idea consistent with the overall meaning is the first phase. A second phase involves staging by locating or constructing an action area and providing resources such as props, costumes, and other necessary equipment. In the third phase, which may overlap with the second, actors are recruited, if they are not already involved in the development of the actable idea, and the actors are trained for their roles. The fourth phase is the period of enactment when the play is performed. In the final phase, after the performance, new meanings for selves, others, and the situation may be assessed by the actors and audience.

The actable ideas that form the basis for social dramas may be either very general or very specific in their implications. On one end of the continuum is an

actable idea in the form of an image, an emotionally loaded set of words, physical object, or event, for example, ghetto or bacchanalia. An image has a program of action packed into it, much like a symbol in a dream may be a merger of waking events. An image often guides interaction in a small informal or therapy group. The actable idea may be a more complex theme (including a direction of movement, an emotional tone, and a minimal set of roles to be enacted) or a plot (with a detailed scenario, defined roles, and an indication of the phases a group must go through to reach its goal). At the specific end of the continuum, the idea may be as fully developed as a script for a play, with parts for each member of the cast and stage directions to guide the performance. A script is most evident in organizations, especially where the actors must use equipment.

Building on the perspective of (Burke 1968), Duncan (1968) emphasizes the importance of symbols, especially as they reflect authority relationships. For MacCannell and MacCannell (1982), the study of signs and symbols had become so important by the late 1900s that they suggested a semiotic interpretation of modern culture as an approach to the time of the sign. Brissett and Edgley (1990) add to the list variables that follow from the work of Goffman (1959) with a primary focus on the way in which individuals present themselves to others. These variables include: role distance, the case when an individual is playing with a role with full consciousness and thereby has the choice of either modifying the role or ceasing to play it altogether; ecstasy, stepping outside the world of everyday life; and remedial work, when a person attempts to change the meaning of an act from one that is seen as offensive to one that is acceptable.

2. *Academic Fields that Contribute to Dramaturgical Analysis*

Dramaturgical analysis has been used in several ways. One way is used more by sociologists of the symbolic interaction tradition with a focus on role playing (cf. Goffman 1959). The second way, introduced by the philosopher–linguist Burke (1968), is based on the analysis of the act. Another approach, used primarily by anthropologists, is to analyze the function of social dramas that are enacted by members of a society to bring about conflict resolution (Turner 1974), to signify a rite of passage, or to memorialize some important event in the history of the society.

Since role and script are central concepts for drama, the fields of psychology, political science, philosophy, linguistics, psychotherapy and the methods of narrative and qualitative analysis, also include descriptions of similar relationships that can be incorporated in the dramaturgical perspective.

3. *Earlier Definitions of the Dramaturgical Perspective*

Evreinoff, in *The Theatre in Life* (1927), records that, at least since the sixteenth century, the similarity between life and the theater has been noted by Shakespeare and others. However, the dramaturgical perspective was not widely used until it was popularized by Goffman in the 1960s. The 1930–35 edition of the *Encyclopedia of the Social Sciences*, edited by Seligman and Johnson, does not include an entry on the dramaturgical perspective, although there is an entry for drama, covering activity in theaters. Burke's entry on dramatism in the 1968 *International Encyclopedia of the Social Sciences*, edited by Sills, presents Burke's pentad of the five terms: act, scene, agent, agency, and purpose. Sets of concepts are linked in ratios, such as the scene–act ratio, where the scene (the situation) has implicit in it the attitudes and behaviors that will become explicit through action. Burke places his own work clearly in the symbolic interaction tradition with references to Mead (1938).

For Brissett and Edgley (1990), in a second edition of their source book on *Life as Theater*, dramaturgy is the study of how human beings accomplish meaning in their lives through interaction with others. Just as meaning is created through enaction it can be recreated through re-enaction. Moreno observed that 'every second time is a liberation from the first' (Hare and Hare 1996) and developed psychodrama as a method of group therapy based on re-enactment.

In the 1992 *Encyclopedia of Sociology* edited by Borgatta and Borgatta in the entry on Dramaturgy, Warner records that the unique insight of dramaturgy is that daily social interaction consists of performances, that individuals stage and in which they present themselves to others. As with Brissett and Edgley (1990) the emphasis is on the work of Goffman and the symbolic interaction tradition of Mead and the Chicago School.

4. *Methodological Issues*

Brissett and Edgley (1990) provide a list of some of the critiques that have been made of dramaturgy. For some critics, dramaturgy is a pedestrian, nonsystematic form of inquiry that does not possess the properties of formal theory. Brissett and Edgley agree, but indicate that is it linked propositionally to other forms of social thought, such as symbolic interactionism, ethnomethodology, existential sociology, interpersonal psychology, and other humanistic models in the social sciences.

A second criticism is that dramaturgy does not produce universal statements about human behavior. It is said to be an artifact of studying the situational behavior in Western culture. In response, Brissett and Edgley first note that understanding how people

interact in Western society is no small achievement, but go on to point to the fact that the expressive behavior of individuals is well-documented in the anthropological literature.

Dramaturgy, or at least Goffman's version of it, has also been criticized for its methodology. It is said to have no specific, systematic method of testing propositions about the world. Brissett and Edgley claim that there is nothing special about doing dramaturgy. Being sensitive to the expressive dimension of behavior demands no special methodology or observational skills. However, there should be an unnerving and single-minded commitment to the observation of people's doings.

Another criticism is that dramaturgy slights the impact of larger social units, such as institutions, on human behavior. In response, Brissett and Edgley assert that there can be no doubt that people's interaction contexts are circumscribed by structural arrangements. However, rather than dwelling on the limitations, dramaturgy focuses on what people do within the contexts that are available to them. The concept of role is used as a way of accounting for people's connections to one another as well as to the structures and organizations with which they are identified.

Finally, the most prevalent critique has to do with the theatrical metaphor. Critics say that the theater is make-believe and everyday life is real. Brissett and Edgely respond that life is neither theater nor is it different from theater. It is theater like.

5. *Future Directions for Theory and Research*

A first act to begin the analysis of social interaction would be to use the dramaturgical perspective to describe the stage, the various actors who are on or off the stage, the audience, and the overall meaning of the event (definition of the situation). As one turns the spotlight on the various actors in turn, their functions (formal and informal) and the process through which they express themselves in their roles can be described in terms of other perspectives, drawn from the social sciences and humanities, until all aspects of social behavior have been included.

See also: Action, Collective; Action, Theories of Social; Collective Behavior, Sociology of; Functionalism in Sociology; Groups, Sociology of; Interactionism: Symbolic; Social Psychology: Sociological

Bibliography

Allen V L, Scheibe K E (eds.) 1982 *The Social Context of Conduct: Psychological Writings of Theodore Sarbin*. Praeger, New York

Brissett D, Edgley C (eds.) 1990 *Life as Theater: A Dramaturgical Sourcebook*, 2nd edn., Aldine de Gruyter, New York

Burke K 1968 Dramatism. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York

Duncan H D 1968 *Symbols in Society*. Oxford University Press, New York

Evreinoff N 1927 *The Theatre in Life*. Bretano's, New York

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY

Hare A P, Blumberg H H 1988 *Dramaturgical Analysis of Social Interaction*. Praeger, New York

Hare A P, Hare J R 1996 *J. L. Moreno*. Sage, London

MacCannell D, MacCannell J F 1982 *The Time of the Sign: A Semiotic Interpretation of Modern Culture*. Indiana University Press, Bloomington, IA

Mead G H 1938 *The Philosophy of the Act*. University of Chicago Press, Chicago

Turner V 1974 *Dramas, Fields, and Metaphors: Symbolic Action in Human Society*. Cornell University Press, Ithaca, NY

Warner K A 1992 Dramaturgy. In: Borgatta E F, Borgatta M L (eds.), *Encyclopedia of Sociology*. Macmillan, New York

A. P. Hare

Dreaming, Neural Basis of

From the animistic dawn of consciousness to the modern psychoanalytic era, dreams have been central to the mythologies of human cultures. Because of their peculiar formal differences from waking consciousness, dreams have typically been thought to represent cryptic messages emanating from sources outside of the dreamer (such as the godhead and other disembodied spirits) or from within the self (such as unacceptable desires arising from the depths of one's own unconscious mind). Dreams have thus been considered to be proverbial, privileged, and prophetic signals so deeply encoded as to require interpretation by a specialist. The interpretation of dreams has fallen to shamans, to priests, and since 1900, to psychoanalysts. But despite vigorous self-promoting appeals, not even the latter group could make any legitimate claim to a genuine science of dreaming. It is only within the second half of the twentieth century that sleep research has provided an objective, experimental foundation for the study of dreaming. By establishing a solid tie with neurobiology, it has also created the opportunity to understand this fascinating mental state in terms of brain physiology.

In such a young field, the data and the conceptualizations of them are still naturally fragmentary, preliminary, and rapidly evolving. And it should be recognized that while animistic and external agency models can no longer be countenanced, a new psychology of dreaming that emphasizes undisguised personal meaning might find a scientifically legitimate place. In fact, it would appear that one of the important functions of dreaming could well be the reconfiguration of emotionally salient biographical information. If this were true then the interpretation

of dreams might be quite simple: the dream is its own interpreter; its meanings are transparent; and its formal peculiarities are the straightforward expression of the unique information-processing mode of the brain in sleep.

1. *Definition of Dreaming*

Dreaming is a distinctive mental state that occurs periodically in normal human sleep. Typical dream reports include such psychological features as hallucinations, delusions, cognitive abnormalities, emotional intensifications, and amnesia. These five remarkable features of dreaming have invited its comparison to abnormal states of mind that occur during waking in certain clinical conditions, especially delirium and the organic mental syndromes. The exploration of the neurobiological basis of dreaming therefore constitutes not only an aspect of mind-body interaction but also provides a model for understanding the brain basis of mental illness.

An obvious obstacle to the scientific study of dreaming is that the unconfirmable nature of all subjective experience is compounded by difficult access to the mind in sleep. The recent development of sleep laboratory recording techniques has given the study of dreaming an instrumental and systematic character, and the emerging data encourage psychophysiological integration. The constellation of psychological features of dreaming, their concomitance with objectively identifiable physiological states, and the occurrence of the same physiological states in all mammals allows the investigation of the neurobiology of dreaming in animal preparations.

1.1 *Psychological Features of Dreaming*

Dreams are characterized by vivid and fully formed hallucinatory imagery with the visual sensory domain predominant; auditory, tactile, and movement sensations are also prominent in most dream reports. Compared with the common representation of these sensorimotor domains, taste and smell are underrepresented. Reports of pain are exceedingly rare despite the involvement of dreamers in frightening and even physically mutilating scenarios. Table 1 summarizes the formal features of REM (rapid eye movement) sleep dreaming.

Dreaming is properly considered delusional because subjects have virtually no insight regarding the true nature of the state in which they have their unusual sensorimotor experiences. Thus the tendency is great to consider dream events as if they were completely real during the dream even though they are promptly recognized as fabrications when recalled in subsequent waking states. This lack of insight is all the more surprising since uncritical belief in the reality of dream events must overcome high degrees of improbability and even physical impossibility.

The lack of insight that makes dreams delusional is part of a broader set of cognitive disturbances. Dreams are characterized by discontinuities (with unexplained changes of subject, action, and setting); by incongruities (with social inappropriateness and cognitive illogicality); by improbabilities and impossibilities (with defiance of physical law); and by uncertainties (with explicit vagueness). Dream characters and dream objects may be banal or they may be altogether fantastic and impossible collages of existing reality; they may behave normally or indulge in the most absurd, improbable, or impossible actions in settings either familiar or bearing only the faintest resemblance's to those of real life. To explain these unique and remarkable dream features, illogical thought processes such as non sequiturs, *post hoc* explanations, and mythical, metaphorical, and symbolic elaborations are the norm.

Memory undergoes a paradoxical intensification and suppression during dreaming; remote recall is intensified within the dream as characters, scenes, events, and concerns from previous experience are knitted into the fanciful and evanescent fabric of the plot. Dreams can thus be said to be hypermnesic within the state itself. This increased access to memory

Table 1
The formal features of REM sleep dreaming

Hallucinations—especially visual and motoric, but occasionally in any and all sensory modalities
Bizarreness—Incongruity (imagery is strange, unusual or impossible; Discontinuity (imagery and plot can change, appear or disappear rapidly); Uncertainty (persons, places and events often bizarrely uncertain by waking standards)
Delusion—we are consistently duped into believing that we are awake (unless we cultivate lucidity)
Self-reflection absent or greatly reduced—relative to waking
Lack of orientational stability lack—persons, times, and places are fused, plastic, incongruous and discontinuous
Narrative story lines—explain and integrate all the dream elements in a confabulatory manner
Emotions increased—intensified and predominated by fear-anxiety
Instinctual programs—(especially fight-flight) often incorporated
Volitional control—greatly attenuated
Memory deficits—across dream–wake, wake–dream and dream–dream transitions

within the state of dreaming contrasts markedly with the virtual impossibility of recovering the dream product after the state has terminated. On awakening, even from a dream in progress, subjects have difficulty holding the vivid experiences in short-term memory long enough to give a report or transcribe the dream scenario in detail. Thus there is amnesia for the hypermnesic dream. It can be conservatively stated that at least 95 percent of all dream mentation is entirely unremembered.

Emotion fluctuates widely in association with the abnormal and vivid mental content of dreaming; anxiety, elation, and anger are the three affects that undergo most marked intensification during dreaming. Obsessional concerns are common with dreamers focusing their worry on nudity, being lost, missing trains, and a host of other incomplete arrangements. Depressive affects are markedly under-represented with shame and guilt playing a relatively small part in dream emotion. Contrary to the predictions of psychoanalytic dream theory, sexual dreams are disappointingly rare.

Mental experience with some or all of these dream features can occur in any physiological sleep stage but it is most common at sleep onset, in REM sleep, and in late night light NREM sleep. Early in the night, deep NREM sleep is usually accompanied by little or no recallable mentation or by perseverative thinking about the individual's ongoing events and concerns.

2. *The Association of Dreaming with REM Sleep*

In 1953, Eugene Aserinsky and Nathaniel Kleitman noted that the sleep of children began and was periodically interrupted by activation of the electroencephalogram and by bursts of saccadic eye movement, the so-called rapid eye movements, or REMs. William Dement confirmed the hypothesis that these periods of brain activation during sleep were correlated with dreaming. The capacity to recall dreams appeared to be related to the nature of the awakening process: subjects who learned to obtain a fully aroused state without moving increased their recall capacity. Home-based recording methods, such as the Nightcap (see Ajilore et al. 1995) have recently confirmed these laboratory observations. Because they are simpler, less disruptive, and cheaper, they can greatly extend this data and put objective dream research within reach of any scientifically inclined psychologist, anthropologist, or social scientist.

Within each REM period, dream intensity tends to parallel the intensity of phasic physiological events, especially the eye movements; arousal during REM sleep with flurries of eye movement activity yielded reports fulfilling the definition of dreaming given here in 90–95 percent of the cases (some aspects of this physiology are shown in Fig. 1). When scored

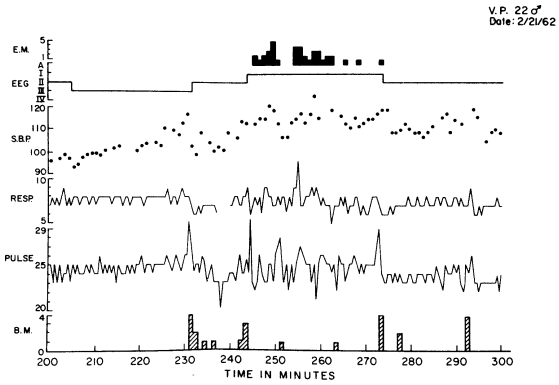

Figure 1
Recording of physiologic activity in one human NREM-REM cycle

for vividness, emotionality, and imagined physical activity, these measures were correlated positively with the quantitative intensity of the eye movement in the REM sleep just prior to awakening.

Estimations of dream duration correlate positively with the time spent in REM sleep prior to arousal. When subjects were aroused after only 5 minutes of REM sleep, they gave shorter reports; after 15 minutes had elapsed, reports were considerably longer and more detailed. Thus it would appear that despite expansion and contraction of temporal duration estimates within individual dream scenarios, the overall correlation between time estimation of dream duration and real time elapsed in REM sleep is quite good. This finding helps to dispel the skeptical notion that dreams occur in the instant before awakening or as part of the awakening process. To test the resistance of memory to dreams, awakenings were performed in the non-REM sleep phase at intervals following the termination of REM. The incidence of reported dreams dropped to non-REM levels within 5 minutes, indicating the extremely fragile state of memory within sleep and highlighting the strong state dependency of recall upon arousal from REM sleep.

3. *A Psychophysiological Theory of Dreaming*

The simplest and most direct approach to the correlation of dream mentation with the physiological state of the brain is to assume a formal isomorphism between the subjective and objective data. By isomorphism is meant a similarity of form in the psychological and physiological domains. For example, the subjective experience of visually formed imagery during dreaming implicates activation of perceptual elements in the visual system in a manner formally similar to that of the waking state. Other details of psychophysiological correlation are assumed to obey the same general law; for example, the vivid

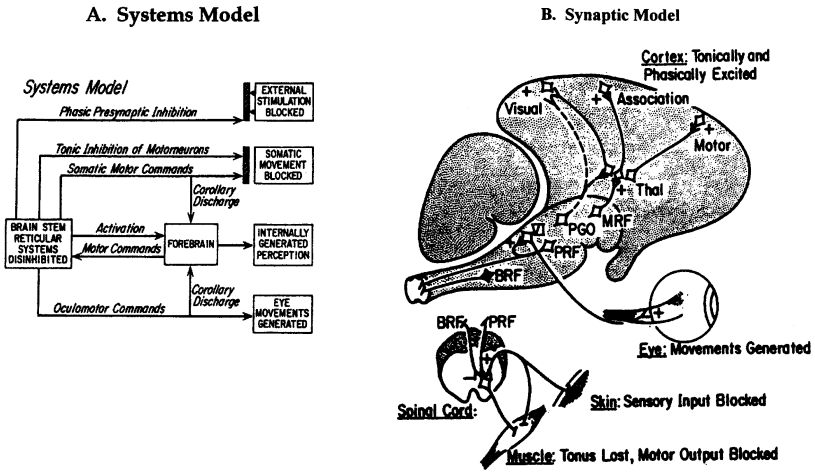

Figure 2
The Activation-Synthesis Model. (**A**) Systems model. As a result of disinhibition caused by cessation of aminergic neuronal firing, brain stem reticular systems autoactivate. Their outputs have effects including depolarization of afferent terminals causing phasic presynaptic inhibition and blockade of external stimuli, especially during the bursts of REM; postsynaptic hyperpolarization causing tonic inhibition of motoneurons which effectively counteract concomitant motor commands so that somatic movement is blocked. Only the oculomotor commands are read out as eye movements because the motoneurons are not inhibited. The forebrain, activated by the reticular formation and also aminergically disinhibited, receives efferent copy or corollary discharge information about somatic motor and oculomotor commands from which it may synthesize such internally generated perceptions as visual imagery and the sensation of movement, both of which typify dream mentation. The forebrain may, in turn, generate its own motor commands which help to perpetuate the process via positive feedback to the reticular formation. (**B**) Synaptic model. Some of the directly and indirectly disinhibited neuronal systems are schematized together with their supposed contributions to REM sleep phenomena. At the level of the brain stem, five neuronal types are illustrated: midbrain reticular neurons (MRF) projecting to thalamus convey tonic and phasic activating signals rostrally; burst cells in the peribrachial region (PGO) convey phasic activation and specific eye movement information to the geniculate body and cortex (dashed line indicates uncertainty of direct projection); pontine reticular formation (PRF) neurons transmit phasic activation signals to oculomotor neurons (VI) and spinal cord which generate eye movements, twitches of extremities, and presynaptic inhibition; bulbar reticular formation (BRF) neurons send tonic hyperpolarizing signals to motor neurons in the spinal cord. As a consequence of these descending influences, sensory input and motor output are blocked at the level of the spinal cord. At the level of the forebrain, visual association and motor cortex neurons all receive time and phasic activation signals for nonspecific and specific thalamic relays

hallucinated sensation of moving is assumed to be related to patterned activation of motor command systems and those central brain structures subserving the perception of position, and changes in position, of the body in space.

To be fully adequate, a psychophysiological hypothesis has to account for the following processes, which are illustrated in Fig. 2 and described in more detail in the entry on *Sleep States and Somatomotor Activity*.

3.1 Activation

The brain has to be turned on and kept internally activated to support dream mentation throughout the REM sleep episode. A possible mechanism for this activation process is disinhibition of sensorimotor circuits (related to cessation of activity in aminergic inhibitory neurons) and internal excitation (owing to the reciprocal increase in cholinergic activation) Aminergic neurons decrease their output by 50 percent in non-REM sleep and virtually cease firing during REM while cholinergic neurons resume their high waking levels of discharge and fire in clusters as the eyes move back and forth.

3.2 Input Blockade

Input from the outside world to the internally activated brain has to be prevented in order for sleep and the illusions of dreaming to be maintained. This is accomplished in at least two ways. One is presynaptic

inhibition of the afferent terminals of the group-Ia afferent nerves, which impedes access of external stimuli to the central nervous system. Presynaptic inhibition has also been recorded at secondary relay nuclei throughout the brain stem and thalamus guaranteeing even more complete insulation of the activated brain from the outside world. The second mechanism for excluding sensory input is the occlusion by internally generated messages that occupy the higher levels of brain sensory and perceptual circuits. Rodolfo Llinas and Denis Paré theorize that the deafferentation of REM sleep is a major determinant of the difference between dreaming and waking consciousness.

3.3 Output Blockade

The internally activated and aminergically deafferented brain must also quell motor outputs to prevent the disruption of sleep by reafferent stimulation and to halt the enactment of dreamed motor commands. This is accomplished by postynaptic inhibition of final common path motor neurons in the spinal cord. The dreamer who fancies him- or herself walking, running, swimming, or flying is, in fact, paralyzed from the neck down.

3.4 Internal Signal Generation

It remains to provide the activated but sensorially disconnected brain with internal signals, which it can then process as if they came from the outside world. This appears to occur in part by a mechanism intrinsic to brain activation: the reciprocal interaction of aminergic and cholinergic neurons in the brain stem. In most mammals including humans, the so-called PGO waves (P, pons; G, (lateral) geniculate; and O, occipital cortex), present themselves as candidates for an internally generated information signal arising at the level of the pontine brain stem. In association with oculomotor activity, strong pulses of excitation are conducted by reticular pathways to the thalamus and radiated by independent pathways to the visual and association cortices.

It is now known that the PGO waves are generated by cellular activity, which faithfully replicates the directional aspects of generated eye movements at the level of the brain stem. Thus not only is internal information generated but the information has a high degree of spatial specificity. According to the activation-synthesis hypothesis of dreaming, the now autoactivated and autostimulated brain processes these signals and interprets them in terms of information stored in memory. Recent evidence indicates that the PGO system is facilitated by cholinergic and restrained by aminergic neuromodulation.

Three important observations concern the origin and fate of motor command signals arising in the brain during REM sleep.

The first observation, suggesting that the motor commands would generate behavior if not inhibited, is the revelation by Michel Jouvet and by Adrian Morrison of elaborate motor behaviors during REM sleep in cats that have lost their capacity to inhibit muscle tone; these behaviors include running, pouncing, attack and defense postures, and intense fear reactions accompanied by hissing and piloerection.

The second observation, suggesting a brain stem origin of the motor patterns generated in REM, is that the locus for the elicitation of locomotion by electrical and chemical stimuli is immediately adjacent to the cellular controllers of REM in the anterodorsal pontine brain stem.

The third observation, offering dramatic proof that brainstem motor commands are actually the building blocks of dream plots, is the emergence of dream content specific movements in human subjects who have lost their ability to inhibit motor activity in sleep owing to degenerative disease of the brainstem.

4. The Neuropsychology of Cognitive Deficits in Dreaming

Although, the basis of the cognitive disturbances occurring in dreaming is not yet well understood, it is possible that the deficits are related to the cessation of activity in aminergic neurons. The REM sleep arrest of aminergic neuronal activity affects the entire brain including the cerebral cortex by depriving it of the tonic modulatory influence of both norepinephrine and serotonin. This tonic aminergic modulatory influence may be essential to waking attentional processes, including the capacity to organize information in a logical and coherent manner and to achieve full self-awareness. In the absence of external cues and deprived of internal modulatory influences by aminergic systems, the activated forebrain thus could interpret its internally generated signals as if they were external stimuli.

By a similar mechanism it may be speculated that the synthesized dream product is unremembered. That is, the activated forebrain circuits which mediate the dream experience are simply not instructed to keep a record of the transaction. The mnemonic instruction is also postulated to be carried out by signals from the aminergic interneurons to their vast postsynaptic domain in waking (when they are active) but not in dreaming (where they are inactive). Current models of learning and memory at the cellular level evoke the intervention of an aminergic interneuron. The attribution of dream amnesia to the loss of aminergic modulation is consistent with these hypotheses.

Table 2
Four features of dream psychology—motor hallucinosis, emotion, bizarreness, and memory deficits—can now be mapped on to the physiological findings

Region	PET studies of activation in REM	Lesions studies of effects on dreaming
Pontine tegmentum	↑	—
Limbic structures	↑	↓
Visual cortex	—	—
Parietal operculum	↑	↓
Dorsolateral prefrontal cortex	↓	—
Mediobasal frontal cortex	—	↓

KEY: ↑ Increase; ↓ Decrease; —No Change.

The complementary results of recent positron emission tomography (PET) imaging and brain lesion studies of dreaming have greatly enriched the experimental study of dreaming in humans. These results are summarized in Table 2.

Two forebrain regions appear to be critical for the maintenance of dreaming. One is the parieto-temporal-occipital-cortex, a region known to be important to visuomotor imagery. The other is a subcortical zone connecting the limbic emotion centers with the frontal cortex. PET studies indicate that both of these regions are selectively activated in REM sleep, and when either is damaged by a brain lesion, subjects report a marked dimunition in the subjective experience of dreaming. The PET studies also indicate that while emotion centers in the amygdala and related limbic cortical regions are turned on, the dorsolateral prefrontal cortex is selectively deactivated in REM. This reciprocal pattern maps on to the intensification of emotion and to the deficiencies in cognition that typify dreaming.

5. *Summary*

The activation-synthesis model of dreaming and the reciprocal interaction theory of sleep cycle control on which it is based are both incomplete and controversial models. They represent linked working hypotheses about the fundamental physiology of sleep and the way in which that physiology may help us to understand unique features of the dream process. The attribution of automaticity to the control system and the feature of randomness in the information generator model should not be taken to exclude the meaningful nature of the synthetic process carried out by the activated dreaming brain. The brain–mind of each dreamer is obliged to make as much sense as it possible of its internally generated signals under the adverse working conditions of REM sleep. Thus, the dream product of each individual may reveal unique stylistic psychological features and emotional concerns that are worthy of scrutiny by individuals interested in reviewing life strategies. But the new theory challenges the psychoanalytic idea that the many meaningless aspects of dream mentation are the result of an active effort to disguise the meaning of unconscious wishes, which are in turn postulated to be the driving force of dreaming. Instead, it ascribes these defective cognitive properties of dreaming to unusual operating features of the internally activated, autostimulated brain during REM sleep.

See also: Circadian Rhythms; Consciousness, Neural Basis of; Electroencephalography: Basic Principles and Applications; Electroencephalography: Clinical Applications; Hallucinations, Neural Basis of; Sleep and Health; Sleep Disorders: Psychiatric Aspects; Sleep States and Somatomotor Activity; Sleep: Neural Systems

Bibliography

Ajilore O A, Stickgold R, Rittenhouse C, Hobson J A 1995 Nightcap: Laboratory and home-based evaluation of a portable sleep monitor. *Psychophysiology* **32**: 92–8

Antrobus J 1991 Dreaming: Cognitive processes during cortical activation and higher afferent thresholds. *Psychological Reviews* **98**: 96–121

Hobson J A 1988 *The Dreaming Brain*. Basic Books, New York

Hobson J A, Pace-Schott E, Stickgold R 2000 Dreaming and the brain: toward a cognition: Neurosciences of conscious states. *Behavioral and Brain Sciences* **23**: 793–842

J. A. Hobson

Dress and Fashion

1. *Introduction*

Before reading and writing there was dress. Styles of dress and patterns of appearance were visual images designed to inform self and other of social responsibility and social rank. These visual images formed a language culled from a body of images accumulating in the group's public memory, a language that became codified as fashion. Past research has concentrated on how fashion, a period's desired appearance, was designed to support the social superiority of those who had acquired wealth in the form of money. By the end of the twentieth century social science was focusing on how fashion was emerging as a means of social criticism that commented on cultural taboos and integrated ethnic ideology into Western dress.

2. Basic Images and Fashion

A group's basic images represent what Emile Durkheim (1893/1947) called the group's 'collective consciousness,' i.e., the group's ideas, beliefs, and values. Lasting images enter into the group's collective memory. When these images are codified, the result becomes a language for social communication. Fashions involving such images as a reshaped head, a lip plug, scissored teeth, or trousers for men and skirts for women convey such a language.

2.1 Collective Memory

Every society has a body of images and meaning that it considers essential. Many images are transmitted from one generation to the next, forming the public memory (Halbwachs 1980). These images become a part of core culture, like time and space, and affect a child's orientation to social realities. Images relevant in a particular period give shape to societal discourse.

A group searching for a visual image to embody its ideas might create a new synthesis from among the elements represented in its public memory—elements that are found in carvings in places of worship, in manuscripts, and in museums and other public places. Mussolini, for example, wanted to create 'a new organic whole' composed of people of all levels of society. He commissioned artistic representations that combined images from classical art of the past with those from traditional folk art. In this genre, men assumed the postures of victory portrayed in ancient Roman times and were seen as victors in agricultural setting. Through this appeal to pride in a shared past, Mussolini gained support for his political program (Malvano 1988).

2.2 Language

Clothes worn by a stranger inform the public of the person's sex, occupation, nationality, and social standing, making it possible for the observer to adjust his or her behavior accordingly, noted Flugel (1930). Writing about Western European societies, Flugel observed that the basic clothing images underlying social interaction are those of *power*, *authority*, *gender distinction*, and *seductiveness*.

Power is identified by images indicating access to human power and superhuman force. The archetypical elements of the image of power are a sword and a crown. Power is also conveyed when a person's appearance is augmented by elements that extend the body's reach, a guard carrying arms, for example. Such elements increase the person's physical ability to control the environment.

Authority is conveyed by body-concealing attire, a vertical continuity of fabric, and a generalized unity from head to toe. The styles traditionally worn by judges, priests, and nuns increase the size and volume of the person. The resulting commanding presence makes it easier for office holders to perform their roles. Conversely, clothing that is too big, too tight-fitting, or too small can dwarf the body and imbue the wearer with a sense of insignificance (Flugel 1930, p. 36).

Gender distinction, as affirmed by the church fathers, was determined by sexual characteristics at birth (*Corinthians* 1, Chap. 11). Men were expected to wear clothes that support activity and work. Women customarily wore clothing alluding to their reproductive role.

Seductiveness, the church fathers claimed, was achieved by such elements as color, ornament, and dress that clung to the body or exposed it. They proscribed such attire. As a result, these elements of image became established as a sexually alluring look (Prusack 1974).

According to this type of linguistic model, clothing 'speech' can be defined as an individual's manipulation of the language of clothing to produce specific utterances characterized by personal intonation and style. People interpret fashion, just as they decode other forms of speech, to ascertain the meaning behind the utterance. To understand the meaning of a cultural expression, research from a variety of fields has suggested that knowing something about the author's purpose, intention, and social situation is essential to valid interpretation (cf. literary critic E. D. Hirsch Jr. (1967), art historian Michael Baxandall (1985), and sociologist Wendy Griswold (1989)).

An example of such a cultural expression is the *Codex Manesse*, a manuscript in the Heidelberg Library illuminated between 1298 and 1304, that teaches morals and manners. One of the scenes involves a classroom. Sitting on a platform and higher than anyone in the room is the teacher, *Magister Henricus*. Imposing in size, and voluminously dressed, he lectures from a manuscript. In his left hand he holds a birch branch. The birch is an emblem of Grammatica, one of the Seven Liberal Arts. It is a symbol of learning, but it also extends the teacher's reach and can inflict punishment. Across the room on a lower level sits the assistant teacher. He is smaller in size and wears a dark tunic not long enough to cover his ankles. His birch branch is smaller. Two students facing him are trembling, as if they had already been hit by the birch rod. The illustration demonstrates that one must listen attentively to the teacher, or the consequence will be punishment and humiliation.

3. Semiotics of Dress

Clothes that indicate expected behavior are called clothing signs. Elements of dress and styles of appearance that reflect the achievement of cultural values are clothing symbols. Clothing tie-signs are clothes

that signify membership in a group and are different from those proffered by mainstream culture. Clothing tie-symbols, on the other hand, represent a preference for a political agenda, a public persona, or a particular social group. It may also reflect an individual's fears, hopes, or dreams. Clothing tie-signs and clothing tie-symbols are forms of publicspeak, the individual's response to existing social arrangements.

3.1 Signs and Tie-signs

Patterns of dress and styles of appearance designated by mainstream culture as identifiers of group membership are clothing signs. Such signs are styles of dress and patterns of appearance required for social participation. They are mandated by a formal code, are instrumental to achieving a goal, and have one meaning. A typical case of sign analysis is dress indicating power, authority, gender, and seductiveness; Flugel's previously identified as clothing signs. Clothing signs are mandated by a formal code, and are instrumental to achieving a goal.

3.2 Symbols and Tie-symbols

Veblen (1899, pp. 118–30) coined the term 'status symbol' to identify images that reflect the achievement of cultural values. Unlike clothing signs, which identify positions within social institutions and are governed by rules of conduct, the wearing of clothing symbols entails personal choice (Goffman 1951).

Every society establishes its own set of values or goals that members of the society should attempt to achieve. The values lead to the norms that guide behavior. In the Amazon, the desire of the indigenous *Yanomamo* for equality leads to norms that forbid the amassing of surpluses and show strong attachment to tradition. For adornment they use sticks, feathers, and paint, which are readily available in their environment.

Social change in Western history encouraged new cultural values. Each of the values was accompanied by the appropriate visual image. Until the fourteenth century *spirituality* was the core value of Christian Europe. Unlike the Romans, who had conceived of the body as a neutral thing placed between nature and society, Christian leaders saw the body as a clearly visible locus of spiritual order. Early Christian writers insisted that what separates Christians from non-Christians is spirituality. Sexual desire and the longing for sensual garments therefore had to be curtailed. Approved clothes were garments that were monochromatic and body concealing.

Wealth as a cultural value emerged in the twelfth and thirteenth centuries. Acquisition of wealth was symbolized by sumptuous fabrics, gold, and costly jewelry. The term 'elegant' referred to the ideal of *harmony*. It entailed creating a costume in which fabrics, colors, and ornaments were coordinated to create a complementary whole.

Fashion is a category of dress that indicates the openness of the society to social change. Fashion in Western society emerged in the fourteenth and fifteenth centuries with the growth of international trade and the resultant spread of affluence to the mercantile class. Fashionableness became the most desired feature of a person's appearance (Bell 1976). It reflected awareness of the novel.

Beauty as a cultural ideal emerged in nineteenth-century England in response to industrialization and urbanization, as strangers were brought together to live and work close to one another. The resulting tension led to a new search for beauty which the ancient Greeks defined as symmetry, harmony, and perfection of the physical form. These qualities were considered to have a calming effect. Beauty as a cultural value was manifested as a preference for better looking servants and employees, for example, and in beautification of gardens and parks (Marwick 1988).

Youth as a cultural value emerged at the end of the nineteenth century as manufacturers chose a younger, more energetic labor force, and desired a population of consumers. The young, having fewer financial burdens, were freer to buy the products of industry. *Health* as a cultural value was informed by increased awareness that illness could be prevented through cleanliness, vaccination, and consumption of healthful food. At the end of the nineteenth century, charts listing appropriate weights for given heights were hung on the bathroom doors of hotels in London.

Since the United States is ethnically and racially diverse—a nation of immigrants and thus strangers—the values of beauty, youth, and health are particularly poignant. They form the fashionable ideal, and provide criteria for social acceptance and self-evaluation. By the end of the twentieth century the desirability of the fashionable ideal was being promoted in the media, drugstore shelves, and establishments offering cosmetics, vitamins, exercise, diet plans, and plastic surgery.

A different category of clothing symbols entails styles of dress and elements of appearance that cover up what Goffman (1963, p. 45) termed '*stigma symbols*,' i.e., an appearance that deviates from cultural values and results in feelings of vulnerability, inadequacy, and fear. Such styles are long and wide to conceal deformities such as crooked spindly legs or a hunchback.

Tie-symbols are clothing that reflect alternative values and suggest a desired association. People often wear tie-symbols to inform themselves and others of their support of groups, political ideas, or social agendas. Examples are T-shirts with messages such as 'Save the Earth' and Spike Lee's black baseball cap with the mark 'X' reflecting his admiration for Malcolm X. Because clothing tie-symbols express

grass-roots sentiments, the wearing of tie-symbols may be the first step in the development of a social movement.

In the 1990s there were about 70 million young adults, teenagers, and school-age children in the United States. The styles they preferred were identified as 'trendy' and included hippie, 'cutting edge,' 'freaky,' and hip-hop. The most popular of these styles was hip-hop. Originating among urban black ghetto youth, the hip-hop style related to a type of music and lyrics; baseball cap turned backwards, pants several sizes too big and other baggy, oversized clothing emblazoned with logos; and to a swaggering manner, all violating the norm of self-restraint. The style was intended to counteract the longstanding perception that blacks and other minorities in America should be invisible. As one rap lyric went, you 'don't see us/but we see you.'

Hip-hop music is used in backing rap, a form of rhythmic speaking in rhyme that is repeated again and again. Rap songs deliver a message that is streetwise and reality-oriented. As a listener noted, 'You listen and you say, yeah that's right' (Farley 1999, p. 56).

The lyrics encourage listening and evaluating how one feels or what one thinks. The music is composed of a collage of excerpts or samples of familiar songs. The rap of the early 1990s was called Gangsta rap because it emphasized crime as a means of acquiring wealth. Gangsta rap, however, also emphasized the drawbacks of such behavior: paranoia, betrayal, revenge, and bloodshed (Reynolds 1999). The required dress was the baggy style that could hide knives, guns, and other implements of attack and self-protection. It was initially worn or advertised by rap musicians, and their youthful audience regarded the attire as empowering (Staples 1996).

By 1999 the rap in hip-hop songs had changed to the 'player style,' celebrating the economic success of the performer, and reveling in the consumption of luxury goods (Reynolds). Hip-hop music and style had become popular in black city ghettos, in affluent white cities like New Canaan, Connecticut, as well as in the dirt-poor Indian reservations of Arizona (Farley 1999, p. 13), and went on to acquire international recognition.

4. *Fashion*

On the European continent up to the twentieth century, the wearing of fashionable attire by the middle and lower classes was regarded as an attempt to violate social order by concealing one's social origins. In nineteenth century England, Thomas Carlyle criticized the wearing of fashionable attire by those not entitled to it, as signifying the pursuit of frivolity rather than the pursuit of spirituality.

Fashion in eighteenth century America was seen by American social critics as representing European decadence, violating Protestant norms of simplicity and hard work. Social crusaders in the nineteenth century, Veblen in particular, argued that the function of fashion is to demonstrate access to wealth. Made of sumptuous fabrics, to order, in a style that made it impossible to engage in productive labor, fashionable attire represented 'conspicuous consumption' and encouraged social inequality and waste.

The American mass production of fashion was found wanting by European social critics. Techniques of mass production 'shatter' the renewal of the human spirit, argued W. Benjamin (1973). Breward (1995) argued that fashion tends to obscure nationality, social status, and gender identity. Coleridge (1988) protested that fashion makes 'bad clothes' look good and in doing so, the fashion industry and fashion magazines assault the notion of beauty.

4.1 *American Fashion*

Fashion in the United States is governed by advances in technology, making possible mass production of clothing at prices affordable to middle class and working class consumers. The main source of style was Parisian fashion. Paris designers used to set the standards, and American manufacturers adapted couture originals for the mass market. This mode of production led to the full-skirted 'New Look' after World War II, the 'sack' in the 1950s, and the space age sleek look of the 1960s notes Agins (1999, p. 9).

Youth revolution of the 1960s led to a decline of Parisian influence as a source of American fashion ideas. In the 1970s name brands became important (Agins 1999, p. 179). Also American designers introduced new themes in to American women's fashion such as *masculine tailoring* and '*sweats*,' fashions based on dance and exercise clothing. Through advertising, editorials, and fashion shows designers communicated directly with the consumer, and designer labels became household names. Styles most emotionally relevant to consumers were copied, and designated the fashion.

Department stores organized their sales floors into brand name boutiques, geared largely to career women of the baby boom generation who needed clothes to wear to work. As brands began to dominate, store buyers were less inclined to try new and unfamiliar labels. Dozens of specialty chains appeared with their own private labels. In the early 1980s the industry, through market research, also geared itself to respond to a new demand. The profitable way of doing business, they insisted, was to find out what the consumer wanted, and make and market it. 'The consumer-is-king' approach became the practice in the United States.

In the fall of 1987, American designers tried, with little success, to take back the authority for the production of fashion. Growth in women's fashions came to a halt (Agins 1999, p. 182). European fashions also failed to find a profitable market in the United States. Styles introduced by couture designers like the

'waif' look, the unkempt grunge look, and the 'fishtail' dress with asymmetrical hemlines were commercial flops.

In the early 1990s, marketing analysts coined the term 'commoditization of fashion.' It described a new era in the move to low-cost factories and offshore production. The trend, which began many years ago with production of cheap lines of clothing, expanded on a vast scale in the 1990s. This time the aim was to manufacture clothes in 'classic' (nontrendy) 'simple chic' and minimalist styles that were good-looking and moderately priced. The large-scale production of attractive, reasonably priced clothes emerged as megabrands.

Traditionally, apparel manufacturers had produced several lines and retailers could reorder styles that were successful. In the trend toward greater economy, megabrand manufacturers produced an initial output that included several styles to be retailed at moderate prices. These, however, could not be reordered. To insure that their output would sell, design houses hired stylists to help create new lines that would respond to emerging consumer preferences. The stylists scouted the streets and clubs to identify new trends. Their influence on new styles emphasized comfort and fun, and some stylists developed into designers. In 1999 about 100 designers—more than ever before—participated in the fashion show '7th-On-Sixth,' organized by the Council of Fashion Designers of America.

4.2 European Fashion

In Europe, the quality of fabric and the integrity of design are paramount. The fashion elite decides on the desirable styles and these styles are found in most stores. The demand for clothing is met by machine-made clothes on a smaller scale than in the United States, and by custom tailoring and dressmaking.

A powerful trade group, Chambre Syndicale de la Couture Parisienne, sets the standards for couture fashion. It holds the right to designate who can be called a couturier, the highest 'occupational' category in the production of fashion for the elite. In the year 2000 only 14 design houses held the title of couturier, which accords social rewards. Although each design house has its own unique style, each is able to draw the audience in, envelop it, and infuse the air with a sense of awe. The trade group is thus the guardian of the expectation that fashion will be an aesthetic experience.

In the 1980s, male and female and European and non-European designers demanded a voice in the production of fashion. Joining together to present their own designs, they formed an international fashion circuit and presented their ideas at fashion shows in Paris, Milan, London, and New York. By the early 1990s there were over 1,500 major showings on the fashion calendar in Paris. The styles they created departed from the traditional criteria of beauty and expressed their own personal experience within a particular cultural milieu. The fashion shows became media events where designers and their sponsors seemed to outdo each other with 'seminude costumes, strange hair and makeup, and gimmicky staging' (Agins 1999, p. 37). TV cameras joined in capturing runway footage for news programs.

Their seemingly 'wild and crazy' fashions conveyed several themes. The *grunge* and the '*heroine chic*' styles reflected the *malaise* that characterized the period. The *seminude costumes and strange hair* told of the need to *restructure gender and sexual arrangements*; the *gimmicky staging* reflected *the split between work and play* that governs everyday life. The fashions presented on the international circuit addressed the problems experienced in contemporary society.

The history of the Western world had been marked by centralized control over clothing and appearance. In the twentieth century these strictures loosened significantly. As a result, fashion evolved to give people of all social classes more opportunities for expression of personal preference in clothing and fashion than ever before.

See also: Collective Behavior, Sociology of; Collective Memory, Psychology of; Consumer Culture; Consumer Psychology; Consumption, Sociology of; Culture, Sociology of; Fashion, Sociology of; Markets: Artistic and Cultural; Status and Role: Structural Aspects; Taste and Taste Culture; Youth Culture, Anthropology of; Youth Culture, Sociology of

Bibliography

Agins T 1999 *The End of Fashion: Mass Marketing of the Clothing Business*. William Morrow, New York

Bell Q 1976 *On Human Finery*. Schoken Books, New York

Blumer H 1969 *Symbolic Interactionism: Perspective and Method*. Prentice-Hall, Upper Saddle River, NJ

Carlyle T 1838 *Sartor Resartus*. Everyman's Library, London

Farley C J 1999 Hip-hop nation. *Time* Feb. 8

Flugel J C 1966/1930 *Psychology of Clothes*. Hogarth Press, London

Goffman E 1951 Symbols of class status. *British Journal of Sociology A* **4**: 294–303

Goffman E 1963 *Stigma: Notes of the Management of Spoiled Identity*. Prentice-Hall, Englewood Cliffs, NJ

Halbwachs M 1980 *The Collective Memory* [trans. Ditter F J, Ditter V Y]. Harper Collins, New York

Jacobsen T 1946 *The Intellectual Adventure of Ancient Man*. University of Chicago Press, Chicago

King S R 1999 Where does generation Y go to shop? 70 million young buyers without a thing to wear. *New York Times* August 28

Malvano L 1988 *Fascismo e politica dell'immagine*. Bollati Boringhieri, Turin

Marwick A 1988 *Beauty in History*. Thames and Hudson, London

Prusack B P 1974 Women: seductive sirens and source of sin. In: Reuther R R (ed.) *Religion and Sexism: Images of Women in*

Jewish and Christian Traditions. Simon and Schuster, New York
Reynolds S 1999 It isn't easy being a superman. *New York Times* Oct. 5
Staples B 1996 Dying to be Black. *New York Times* Oct. 9
Veblen T 1899 *The Theory of the Leisure Class*. Macmillan, New York

R. P. Rubinstein

Drinking, Anthropology of

Since at least the 1940s, alcohol has been the focus of a growing interdisciplinary cadre of researchers. Some have proposed 'alcohology' as an academic discipline of its own (Heath 1987, p. 100). Anthropological data and viewpoints have played a part from the start. Their influence has perhaps been greater than expected, considering the limited number of professional anthropologists involved. The first appeal for anthropological studies of drug dependence came from Edward Glover (1932), a leading British psychoanalyst who needed data for psychoanalytic theory. Ruth Bunzel (1940) offered an early response.

The scrutiny of social and cultural aspects of alcohol consumption can be seen as anthropological in the sense that it benefits from comparison between societies. The most ambitious endeavor is still that of Donald Horton; in his own words, 'an experiment in scientific method' (Horton 1943, p. 202). He examined anthropological statements on drinking from 77 'primitive' societies and concluded that the primary function of alcohol is to reduce human anxiety and fear, a view that has been questioned by later scientists. Horton's aim was to study the 'functions' of alcohol as a starting point for a psychocultural theory. The available data were, however, of varied character and quality. The lack of systematic descriptions of ordinary drinking prevented useful conclusions.

It has been an anthropological aim to focus on relatively small and homogeneous groups and there to indicate the social and cultural structures encompassing and supporting the patterns of drinking. Ethnographic analyses show rich variation in the use of alcohol in the multitude of cultural contexts. Neither the use of alcohol nor its effects and consequences are universal, or in any way anchored in inherent properties of alcohol, or in 'human nature,' or in 'male nature'—for a long time women were not made particularly visible in the literature on drinking, although times are now changing (see, for example, Gefou-Madianou 1992).

A complex, diverse and often paradoxical pattern of belief and behavior has come to surround a simple chemical product. Alcohol is simultaneously considered a stimulant and a sedative, both food and poison. It has a close connection to religious practice, though drinking is often characterized as 'diabolic.' A great many phenomena, from the most positive to the most negative, have been attributed to the use of alcohol, although most people are well acquainted with its allegedly negative effects on human behavior.

1. *Fusing the Symbol*

It is not strange that alcohol holds a special position in human life. Its influence on motor activity and its gradual effect on human consciousness are sufficiently dramatic as a 'sign' that alcohol is something out of the ordinary. Reports from other cultures often repeat that the purpose of drinking is to achieve drunkenness, sometimes to the extent of unconsciousness.

Bunzel (1940, p. 384) notes that the Chamula 'drink for pleasure—direct somatic pleasure of warmth, relaxation and deep untroubled sleep,' a sequence comparable to what Radó (1957) calls 'alimentary orgasm' She declares that alcohol has replaced all other forms of aesthetic expression: 'People are too engrossed in drinking to develop the arts of life. Commitment to one form of gratification has blotted out every other possibility' (Bunzel 1940, p. 385). Bunzel's use of 'commitment' brings up the human ability to fuse, under certain circumstances, with a symbol, a phenomenon that could cause everything else to seem of little importance.

The unique position of alcohol—despite a simple and natural origin—easily identifies it as symbol. As symbol, alcohol has the quality of allowing or even calling upon *fusion* with its human imbibers. The description of a symbol is certainly poor and empty if severed from the human experience for which the symbol is crucial.

The basis of the argument is the possible relationship of commitment between man and symbol. Intoxication through alcohol may be seen as a relation of symbolic commitment. This also implies that the inebriated *use* this possibility for some purpose, for instance to get an outlet for thoughts and feelings otherwise not permitted, or perhaps just hard to express. Commitment to a symbol does not imply cultural consensus. All those committed, even if enjoying social consensus, need not agree on the meaning of drinking. Drunkenness, like religious experience, is abstract in relation to its content.

2. *Alcohol as Ornament*

In a relatively early interpretation, Harold Fallding (1964) places drinking in a wide civilizational context. Civilization, he suggests, may be understood as 'the dignifying of human existence through adornment.'

Human beings—at any cultural level—always use a 'surplus to devise need-gratifications not strictly related to survival' (Fallding 1964, p. 714). Alcohol is such a surplus, such an ornament. Men might be enslaved by excessive devotedness to their ornaments, a main predicament and principle of fetishism. Men, growing dependent on things they have unnecessarily produced themselves, are degraded by their own invented rubbish.

Fallding deprives alcohol of its demonical and divine disguise and exposes it as a dead thing, unnecessary for human existence. In the same breath alcohol is rehabilitated to only *one legitimate function* in society: as symbol for a necessary community. Community is indispensable for the trust which humankind cannot do without for its common survival, its common production.

Fostering of community keeps the 'fount of trust open in the midst of the garden of civilization' (Fallding 1964, p. 714). Trust must be maintained despite society's acceleration in numbers and specialization. Production will develop only if reciprocal trust is there. To the extent that drinking symbolizes an *existing* community it represents 'the most precious support men can experience' (Fallding 1964, p. 717).

The occurrence of intoxication may offer a specific experience of community. To share a drunken moment could produce a feeling of trust and a touch of being a participant in something potentially dangerous. The individual feeling unsafe, or with a position to defend, may disallow himself from becoming intoxicated, even if drinking.

The 'ornamental, community-symbolic drinking' suggests an already existing community. If this drinking appears to be rare in modern society, it might indicate a decline of community. Community and trust must in that case be restored first, since 'nothing could be less sufficient for the genesis of community than to supply the materials by which it could be symbolized, if present' (Fallding 1964, p. 719).

3. *Drinking as a Problem*

Mutual drinking can contribute to the maintenance of social structure. Harriet Kupferer observes that studies asserting that 'drinking together enhances social solidarity and increases sociability' (Kupferer 1979, p. 198) obviously take their starting point in Durkheim and Radcliffe-Brown. Mary Douglas (1987), however, warns that the literature might overemphasize the idea of community as an answer to alcohol problems. Authority, ritual, and solidarity—on a foundation of community and trust—certainly seem to bring the drinking under control. But sometimes the community is aggressively competitive. The drinker following norms in such company risks his health.

It is however necessary, not least in the anthropological context, to distinguish between addiction and the 'nonpathological heavy indulgence in drinking' (Jellinek 1941, p. 777) of which the Bolivian Camba is a good example (Heath 1962). Horton reminds us that 'primitive peoples' are unaware of alcohol 'problems' (Horton 1943, p. 200). They just know that some of their beverages produce certain characteristic sensations and behaviors. Bales (1946) also emphasizes the need to avoid the confusion between the extent to which alcohol is *used* in a given culture and the extent to which alcohol causes problems in that culture.

Comparing Jewish and Irish Americans, Bales presents an illustration to the fact that differences in cultural context influence the incidence of alcohol problems. He suggests that despite the fact that practically all Jews drink some kind of alcoholic drink every week, the number of alcoholic Jews is small (Bales 1946, p. 480). Snyder observes that Orthodox Judaism, as a normative and cultural system, seems to have restrictive influences on its members. 'But this effect may depend in the long run upon the continuing participation of Jews in the ceremonial and ritual activities of the traditional religious community' (Bales 1946, p. 112).

Alcohol abuse among Jews in New York is thought not to depend on their assimilation into American norms of consumption, but rather on their distance from traditional Orthodoxy and from ceremonial participation renewing solidarity and moral authority. Catholicism and Protestantism offer no similar ritual control of personal behavior. The priests advocate soberness or moderation but are as a rule isolated from popular drinking. Church ceremonies do not incorporate drinking other than the Communion.

Kunitz et al. discuss notable differences between Navaho and Hopi in the incidence of cirrhosis (Kunitz et al. 1971). Drinking is customary among both groups, but patterns differ. The authors suggest that drinking in Navaho society is a 'normal pursuit of young men and tends to be outgrown' (Kunitz et al. 1971, p. 715). The individual who drinks is not expelled from the community except in rare cases. Higher prevalence of cirrhosis near the border towns might be a consequence of a freer access to alcohol, and some might develop cirrhosis before they outgrow their drinking. 'Many of those currently abstaining had been very heavy drinkers previously and some no doubt would have been diagnosed as alcoholics. Nevertheless, they stopped drinking with relative ease and without the problems commonly reported by Anglo alcoholics' (Kunitz et al. 1971, p. 716).

The fact that the Navaho drinkers are not sent away from the community, and are not isolated, seems to allow them a kind of shelter. The Navaho possibly retain a traditional pattern for drinking immoderately when there is alcohol. They appear not to see self-control as a social value in the same way as the Hopi do. The Hopi situation is thus different. The Hopi way has emphasized harmony and peace and avoided

'acting out.' The traditional method of social control has included gossip, witchcraft allegations, and final expulsion from the community. As a result the drinker may become isolated and released from restraint.

Differing fatality rates as a result of cirrhosis could depend on the Navaho pattern of heavy drinking at an early age, which is abandoned so early in life that cirrhosis, if it occurs, is reversible. The Hopi pattern, on the other hand, might lead the individual ejected from the community to become isolated, and to continue drinking until illness becomes fatal.

4. *Alcohol and Health*

Some of the value of anthropological study on alcohol lies in the way it contrasts with the segmentation that is otherwise characteristic of alcohol research: there are inquiries into alcohol and violence, alcohol and traffic, alcohol and health, and so forth. Anthropology instead cherishes the idea of studying alcohol in its full social and cultural setting, and anthropologists therefore have a perspective on drinking activities grounded in fieldwork. They do not necessarily treat alcohol as a problem, and seem to find that medicine and sociology may be inclined to some exaggeration in this regard. 'Even in the United States where there is so much concern about alcohol abuse, the most pessimistic estimate is that alcohol-related troubles afflict fewer than 10 percent of those who drink' (Douglas 1987, p. 3).

The anthropological approach challenges the assertion that drinking leads to anomie. It is rather anomie that leads to alcoholism, and the anthropologists are of course ready, adds Douglas, to confront the theoretical problems of the anomie concept. Nor do they find any clear relation between drinking and disposition to aggressiveness and lawbreaking, and they question the idea of drunken misbehavior as an example of alcohol's capacity to slacken cultural restrictions.

An anthropologist focusing on alcohol abuse probably will conclude that the most efficient way to control misuse is through socialization. But there is no way to stipulate 'a cultural level of tolerance for alcohol' against which questions implied in a cultural approach could be asked: 'Do cultural constraints dictating a restricted framework for drinking enable more individuals to absorb more alcohol with less apparent physical effect? Does cultural training which enables individuals to hold their drink more easily protect their livers from damage? Most likely not. Does the individual breaking out of a set of cultural constraints drink more deeply and more dangerously than one whose heavy drinking is culturally expected and approved? What difference does the surrounding culture's attitude to drunkenness make to the prospects of an addict's cure?' (Douglas 1987, p. 5).

Gender differences in alcohol use may imply a greater freedom for men, and contribute to men's collective construction of 'real masculine identities' (Gefou-Madianou 1992, p. 11); a prerogative possibly counterbalanced by women's avoidance of an obvious health risk and conspicuous capacity for 'getting drunk' on conviviality (Gefou-Madianou 1992, p. 15). A pessimistic prediction, not without empirical evidence, is that females will catch up with males primarily in what Fallding characterizes as drinking in 'protest of distrust against the community' (Fallding 1964, p. 721).

Anthropology has no immediate solution to alcoholic health hazard. It is possible but not proven (or even properly studied) that heavy alcohol consumption is less hazardous to health if symbolizing community than if in protest against exclusion. Most people drink alcohol without suffering in any observable way. That reservation is important. The causal connection supposedly established between drinking and illness is not easily confirmed or refuted in the anthropological field, although anthropology, back from field, may profitably 'meet' other branches of learning (Demers et al. 1996).

A fundamental contribution from the anthropological quarter could be the statement that studies from the biomedical or the cultural sphere, as well as true interdisciplinary attempts (Lucia 1963) and consensual policy efforts (Edwards 1994)—though welcome for furthering a deeper understanding of alcohol use—may not provide a solution to the drinking problem. The solution will not come about by way of any research program, but rather, if at all, as the result of a major cultural achievement.

See also: Alcohol-related Disorders; Drug Addiction: Sociological Aspects; Drug Use and Abuse: Psychosocial Aspects; Ritual

Bibliography

Bales R 1946 Cultural differences in rates of alcoholism. *Quarterly Journal of Studies on Alcohol* **6**: 480–99

Bunzel R 1940 The role of alcoholism in two central American cultures. *Psychiatry* **3**: 361–87

Demers A, Kishchuk N, Bourgault C, Bisson J 1996 When anthropology meets epidemiology: using social representations to predict drinking patterns. *Substance Use & Misuse* **31**: 847–71

Douglas M 1987 A distinctive anthropological perspective. In: Douglas M (ed.) *Constructive Drinking: Perspectives on Drink from Anthropology*. Cambridge University Press, Cambridge, UK

Edwards G (ed.) 1994 *Alcohol Policy and the Public Good*. Oxford University Press, Oxford, UK

Fallding H 1964 The source and burden of civilization illustrated in the use of alcohol. *Quarterly Journal of Studies on Alcohol* **25**: 714–24

Gefou-Madianou D (ed.) 1992 *Alcohol, Gender and Culture*. Routledge, London
Glover E 1932 Common problems in psychoanalysis and anthropology: drug ritual and addiction. *British Journal of Medical Psychology* **12**: 109–31
Heath D 1962 Drinking patterns of the Bolivian Camba. In: Pittman D, Snyder C (eds.) *Society, Culture, and Drinking Patterns*. Wiley, New York
Heath D 1987 Anthropology and alcohol studies: current issues. *Annual Review of Anthropology* **16**: 99–120
Horton D 1943 The functions of alcohol in primitive societies: a cross-cultural study. *Quarterly Journal of Studies on Alcohol* **4**: 199–320
Jellinek E M 1941 Immanuel Kant on drinking. *Quarterly Journal of Studies on Alcohol* **1**: 777–8
Kunitz S J, Levy J E, Odoroff C L, Bollinger J 1971 The epidemiology of alcoholic cirrhosis in two southwestern Indian tribes. *Quarterly Journal of Studies on Alcohol* **32**: 706–20
Kupferer H 1979 A case of sanctioned drinking: the Rupert's House Cree. *Anthropological Quarterly* **52**: 198–203
Lucia S P (ed.) 1963 *Alcohol and Civilization*. McGraw-Hill, New York
Radó S 1957 Narcotic bondage. *American Journal of Psychiatry* **114**: 165–70
Snyder C 1958 *Alcohol and the Jews*. The Free Press, Glencoe, UK

G. Johansson

Drug Addiction

1. Drug Addiction

Drug abuse and addiction are among the largest and most challenging problems facing society today. Scientific advances over the past twenty-five years have greatly contributed to our understanding of drug abuse and addiction. We now know that drug abuse is a preventable behavior and that drug addiction is, fundamentally, a treatable chronic, relapsing disease of the brain. We also now have tremendously detailed knowledge of what drugs actually do to an individual's brain and behavior. For example, it is known that drugs work by attaching themselves to a particular class of proteins called receptors. Many different proteins can serve as receptors for drugs. Scientists have identified and cloned the receptors in the brain for every major drug of abuse. They have identified the cellular sites where drugs like cocaine, marijuana, nicotine, and opiates bind to the brain. For example, in the case of cocaine and methamphetamine, they have recognized the dopamine re-uptake transporter as a major site of action. Researchers have not only discovered specific brain circuits involved in drug experiences, like euphoria, but the processes of addiction and drug withdrawal as well. Research also shows that addiction occurs as a result of the prolonged effects of abusable drugs on the brain—and that addiction results in a brain that is changed both structurally and functionally.

2. Why People Take Drugs

To fully understand the phenomenon of addiction, we need to have a thorough understanding of why people take drugs in the first place. Although we know some of the initial reasons why people take drugs—to modify their mood, perception, emotional state, or to self-medicate in an effort to cope with their problems—no one single underlying factor determines which individuals are more or less likely to abuse drugs and which will later become addicted to drugs. An individual's vulnerability to becoming addicted to drugs is dependent on a number of complex factors, including a variety of genetic, biological, emotional, cognitive, and social risk factors.

Researchers have identified many of the risk factors that increase the probability that someone will develop drug problems. These factors can be found at the individual, the family, peer group, and at the broader community level. Researchers have also identified many of the protective or resiliency factors that protect individuals from developing drug-related problems. An individual's vulnerability to becoming a drug user and later an addict is likely to be due to environmental and genetic factors interacting in some, as yet, unknown way. We also know that regardless of the reason for initial drug use, continued use can lead to neuroadaptation in brain reinforcement circuits.

Clear evidence exists demonstrating that almost all drugs of abuse have effects on brain circuitry involved in the control of motivated and learned behaviors. Although each class of drug may have its own specific mode of action, in tracing the pathways through which a variety of drugs exert their effects; certain commonalities for all drugs of abuse have emerged. This evidence originated from studies in animals (Weiss 1996, Wise and Bozarth 1981) and is being confirmed in humans with recent developments in brain imaging techniques. Anatomically, the brain circuitry principally involved in most of the actions of the major addictive drugs is the reward center or the mesolimbic pathway. This pathway consists of cells (mesolimbic neurons) that originate in the ventral tegmental area (located in the middle of the brain) that terminate and signal other cells in distant brain areas.

These cells (like all neurons) do not actually touch each other but are separated by specialized gaps called synapses. Neurons communicate with each other via chemical messengers referred to as neurotransmitters, which cross the synapses. In the case of mesolimbic neurons, the neurotransmitter used in the communication process is dopamine. It is at the synapse that

drugs of abuse disrupt the normal communication process among neurons. Scientists have found that all drugs of abuse cause an increase in dopamine levels in the brain's reward pathway. Activation of this reward system appears to be a common element in the reinforcement of drug-taking behaviors. Increased dopamine levels are associated with the feelings of euphoria and reward that are often associated with drug use. Researchers are also finding that the dopamine system may help play a role in determining what makes some individuals more susceptible to addiction than others (Volkow et al. 1999). Through the use of neuroimaging, researchers at Brookhaven National Laboratory have found that individuals who have higher densities of dopamine receptors before any drug exposure, report having a pleasant response to a mild stimulant; whereas individuals with lower dopamine 2 receptor levels report having unpleasant responses to the same mild stimulant. These findings are in concurrence with studies in laboratory animals indicating that dopamine receptors mediate reinforcing responses to drugs of abuse. Not only does this finding suggest that an individual's resting brain chemistry may predispose how he or she responds to different drugs of abuse, but it potentially helps to predict who will develop drug problems. The finding also serves as a critical piece of information to help facilitate the development of more effective prevention and treatment approaches.

3. *The Transition from Drug Abuse to the Disease of Addiction*

The fact that addiction is tied to changes in both brain structure and function is what makes it fundamentally a brain disease. For many of the individuals who do not have an initially negative reaction to their first drug-administration, continued repetition of 'voluntary' drug taking begins to change—often undetectably—into compulsive drug taking, ultimately to the point where the behavior is driven by an overwhelming craving for the drug. The compulsion to get and use drugs becomes the most powerful motivator in a person's life. It is at this stage that addiction becomes a chronic disease with relapse and remission. We often make the analogy that a methaphorical switch in the brain is thrown that changes an individual user/abuser into a drug addict. The individual changes from a totally voluntary drug user into a compulsive, often uncontrollable drug user.

Although research has significantly advanced our understanding of the behavioral, biological, and molecular factors which underlie initial drug use and the actual state of addiction, we are lacking a thorough understanding of the mechanisms that are involved to explain what literally occurs when an individual transitions from a voluntary drug user to an addict. We know that the full course of a typical drug use disorder evolves through a whole series of transitions over a period of years that begins with the initial onset of drug use, the maintenance or continuation and eventual escalation of use, movement from one drug class to another, escalation within that drug class, and multiple drug use (in many cases). Typically once treatment is provided, the individual discontinues drug use, but often relapses even after long periods of abstinence (Glantz and Hartel 1999). Although this is a frequent pattern for this chronic relapsing disease, we do not know if the changes in the brain are in fact gradual or precipitous in nature. The 'switch' seems to flip at different points for different individuals. That is, for some it may happen slowly or not at all, whereas for others it may happen rapidly. Researchers do know that once the switch is turned on and the individual becomes addicted, it is extremely difficult for most people to stop using drugs without treatment. Understanding this transition is a major area of investigation for researchers.

4. *Long-lasting Changes in the Brain*

Studies in both humans and animals have demonstrated that chronic drug use changes the brain in fundamental ways that persist long after the individual has stopped taking drugs. This addicted brain is different from the nonaddicted brain, as manifested by changes in brain metabolic activity, receptor availability, gene expression, and responsiveness to environmental cues. These induced changes are responsible for the uncontrollable, compulsive drug-seeking and use—that are the essence of addiction. How long these brain changes last is just now being determined. Some of these long-lasting effects may also affect an individual's emotional responses and his or her learning and memory capacity. One of the most recent examples can be seen with the drug methylenedioxymethamphetamine (MDMA) or 'Ecstasy.' MDMA is being used increasingly by teens and young adults who attend organized all night social gatherings or 'raves.' Chronic use of MDMA has recently been found to produce long-term damage to serotonin-containing neurons in the human brain. The neurotoxic effects of this drug have long been known in animals, but now we are seeing its damaging effects in humans as well. Given the important role that the neurotransmitter serotonin plays in regulating emotion, memory, sleep, pain, and higher order cognitive processes, it is easy to see how MDMA use can impair ones memory; and probably generate other behavioral consequences as well (McCann et al. 1998). MDMA can decrease an individual's ability to remove this important neurotransmitter from the synapse, thereby amplifying its effects within the brain. This effect lasts at least three weeks after the individual has stopped using Ecstasy.

Given serotonin's critical role in many behaviors, one can speculate that this change in the serotonin system might be responsible for some of Ecstasy's long-lasting behavioral effects.

5. *Reversing or Compensating for Brain Changes Through Treatment*

Given that the brain fundamentally changes when an individual becomes addicted to drugs, preventing the initial use of drugs is of course the best treatment for preventing a lifetime of drug problems (National Institute on Drug Abuse 1997). If, however, the compulsion to use drugs takes over an individual's life, treatments are available that can help an individual to stop using drugs and maintain a drug-free lifestyle. Many of today's behavioral and pharmacological treatments for addiction can reverse or compensate for the brain changes that occur during prolonged drug use. Given that these changes in brain function do not occur in isolation, and are in fact interacting with environmental factors such as stress and situational triggers as well (O'Brien 1997) addiction is never simple to treat. It is often these external factors that cause an individual to relapse to drug use, even after successful treatment. Researches are working to determine what treatments work best for individuals who are likely to be affected by these environmental factors. Given that the ultimate goal of treatment is to return the individual back to a functioning member of society, it is critical that individuals can cope with these factors without relapsing to drug use.

Just as addiction affects all aspects of a person's life, his or her treatment program must address the 'whole person.' It must address not only the individual's drug use, but also provide him or her with the tools necessary for maintaining a drug-free lifestyle, while also helping with the achievement of productive functioning in the family, at work, and in society. Because addiction is so complex, and can impact so many aspects of the patient's life, effective treatment programs typically must incorporate many components, each directed at a particular aspect of the illness and its consequences.

Fortunately, as with other illnesses, drug abuse professionals have at their disposal an array of useful tools to treat addicted individuals, though admittedly not enough. Treatments vary depending on the characteristics of the patient and the type of drug to which they are addicted. Treatment can include behavioral therapy, medications, or their combination. The best programs provide a combination of therapies and other services to meet the needs of the individual patient, which are shaped by such issues as age, race, culture, sexual orientation, gender, pregnancy, parenting, housing, and employment, as well as physical and sexual abuse (National Institute on Drug Abuse 1999).

Maintenance treatments such as methadone and LAAM (levo-alpha-acetyl-methadol) exist for treating heroin addiction; psychoactive medications such as antidepressants can be particularly useful when patients suffer from a co-occurring mental disorder; and notable behavioral interventions, such as cognitive behavioral therapies and contingency management, can be used effectively to treat both adults and adolescents. Another effective medication for treating some opiate patients, especially those with co-occurring alcohol dependence problems is Naltrexone. For persons addicted to nicotine, nicotine replacement products (such as patch, gum, or nasal spray) or an oral medication (such as bupropion) can be an effective component of treatment.

It bears repeating that the commonly held misperception that two of the most effective medications for treating heroin addiction—methadone and LAAM—are simply substitutes for heroin is wrong. Instead of destabilizing the individual, as heroin does, methadone and LAAM stabilize the patient and facilitate a return to productive functioning (Ball and Ross 1991). It is also important to note that detoxification itself is not a form of drug treatment. At best, detoxification is a first step in beginning treatment, but by itself does little to affect long-term drug use.

Research studies have repeatedly demonstrated that addiction is eminently treatable and is comparable with the results found with other chronic disorders such as diabetes or asthma (O'Brien and McLellan 1996). Drug addiction treatment reduces drug use by up to 60 percent and has been shown to be effective in reducing the spread of HIV, diminishing the health and social costs that result from addiction, and decreasing criminal behavior (Hubbard et al. 1997).

6. *Conclusion*

Although the onset of addiction begins with the voluntary act of taking drugs, the continued repetition of voluntary drug taking begins to change into involuntary drug taking, ultimately to the point that a compulsive craving for the drug drives behavior. This compulsion results from a combination of factors, including in a large part, the dramatic changes in brain function produced by prolonged drug use. This phenomena explains why addiction is considered a brain disease—one with embedded behavioral and social aspects (O'Brien 1997, Leshner 1997). Once addicted it is almost impossible for most people to stop the spiraling of addiction on their own without treatment. Fortunately there are quite a number of effective treatments available to manage addiction.

See also: Alcohol-related Disorders; Alcoholics Anonymous; Alcoholism: Genetic Aspects; Drug Addiction: Sociological Aspects; Drugs and Behavior, Psychiatry

of; Drugs: International Aspects; Neurotransmitters; Reinforcement: Neurochemical Substrates; Synaptic Transmission

Bibliography

Ball J C, Ross A 1991 *The Effectiveness of Methadone Treatment*. Springer, New York

Childress A R, McLellan A T, Ehrman R, O'Brien C P 1988 Classically conditioned responses in opioid and cocaine dependence: a role in relapse? In: Ray B (ed.) *Learning Factors in Substance Abuse, Research Monograph* **84**: 25–43

Glantz M D, Hartel C R 1999 *Drug Abuse Origins and Interventions*. American Psychological Association, Washington, DC

Hubbard R L, Craddock S G, Flynn P M, Anderson J, Etheridge R M 1997 Overview of 1-year follow-up outcomes in the Drug Abuse Treatment Outcome Study (DATOS). *Psychology of Addictive Behaviors* **11**: 261–78

Institute of Medicine 1995 *Development of Medications for the Treatment of Opiate and Cocaine Addictions: Issues for the Government and Private Sector*. Fulco C E, Liverman C T, Earley L E (eds.). National Academy Press, Washington, DC

Leshner A I 1997 Addiction is a brain disease, and it matters. *Science* **278**: 45–47

London E D, Cascella N G, Wong D F, Phillips R L, Dannals R F, Links J M, Herning R, Grayson R, Jaffe J H, Wagner H N Jr. 1990 Cocaine-induced reduction of glucose utilization in human brain. A study of positron emission tomography and [fluorine 18]-flurodeoxyglucose. *Archives of General Psychiatry* **47**: 67–74

McCann U D, Szabo Z, Scheffel U, Dannals R F, Ricaurte G A 1998 Positron emission tomographic evidence of toxic effect of MDMA ('Ecstasy') on brain serotonin neurons in human beings. *Lancet* **352**: 1433–37

National Institute on Drug Abuse 1997 *Preventing Drug Use Among Children and Adolescents: A Research-Based Guide*. Sloboda Z, David S L. NIH publication No. 97-4212

National Institute on Drug Abuse 1999 *Principles of Drug Addiction Treatment: A Research-Based Guide*. NIH publication No. 99-4180.

O'Brien C P 1997 A range of research-based pharmocotherapies for addiction. *Science* **278**: 66–70

O'Brien C P, McLellan A T 1996 Myths about the treatment of addiction. *Lancet* **347**: 237–40

Volkow N D, Fowler J S, Wolf A P, Hitzemann R, Dewey S, Bendriem B, Alpert R, Hoff A 1991 Changes in brain glucose metabolism in cocaine dependence and withdrawal. *American Journal of Psychiatry* **148**: 621–6

Volkow N D, Wang G J, Fowler J S, Logan J, Gatley S J, Gifford A, Hitzemann R, Ding Y S, Pappas N 1999 Prediction of reinforcing responses to psychostimulants in humans by brain dopamine D_2 receptor levels. *American Journal of Psychiatry* **156**: 1140–443

Weiss F 1996 Neurochemical adaptation in brain reward systems during drug addiction. *Institute of Medicine Symposium on Neuroscience Research: Advancing our Understanding of Drug Addiction*. National Academy of Sciences, Washington, DC

Wise R A, Bozarth M A 1981 Brain substrates for reinforcement and drug-self-administration. *Progress in Neuropsychopharmacology* **5**: 467–74

A. I. Leshner

Drug Addiction: Sociological Aspects

The main aim of this article is to understand how sociology has considered the recent social phenomenon of drug addiction. This article will not deal with all the areas concerned with drug problems, notably crime and criminal justice (see Weisheit 1990) or drug control policies (see e.g., Musto 1987) that have become legitimate topics for social scientists. It will relate the history of the principal sociological theories concerning drug addiction. In turn, it will discuss the respective concerns and impact of these theories, as well as their contributions and failures in understanding this type of deviant behavior.

1. *Medical Definitions*

First, it is necessary to define the terminology. Generally speaking, a drug (heroin, crack, ecstasy, cannabis, cocaine, alcohol, or psychoactive medicines) is either a natural or a synthetic substance that can alter the mind and influence mood behavior and perception. But common points of view and modern Western law generally consider drugs to be only those psychotropic substances that are illegal and that can lead to addiction. The concept of addiction remains very controversial and troublesome (Akers 1991), but physicians usually define drug addiction as a personal inclination to use drugs that can provoke dependence. From this medical point of view, drug addicts experience at least one of two kinds of dependence: (a) a physical dependence characterized by withdrawal symptoms that follow abstinence from drug use; (b) a psychological dependence that manifests itself by a strong desire to continue using drugs. But this medical definition does not offer any information and knowledge about what sociology is actually concerned with: the social dimension of the drug user's experience.

2. *The Recent Development of Drug Uses*

People have always used drugs for medical, religious, and artistic purposes (opium was used in Babylon), and those uses have long been culturally regulated and socially integrated in all kinds of traditional rituals (Escohotado 1998, Knipe 1995). But in the middle of the nineteenth century, some drugs formerly used as medicines came to be regarded as dangerous (Bachmann and Coppel 1989, Yvorel 1992): Abuse of certain substances gradually became regarded as pathological. The medical concept and the term 'drug addiction' appeared at the end of the nineteenth century (the term 'alcoholism' was invented by Magnus Huss around 1850), but in those times drug addiction was restricted to only a few soldiers

and officers in the colonies, artists, and some physicians who had experimented with these substances on themselves. Both the use of drugs and of the term 'drug addiction' became more widespread in the 1920s (Bachmann and Coppel 1989), and the stereotypes according to which 'drugs are a demon, leading to social and moral misery,' 'drug addicts' behavior is completely irrational,' 'drug addiction is an epidemic disease,' etc., started to spread in Western countries.

Sociological research began to consider drug addiction as a legitimate object when drug use became very widespread in Western countries at the beginning of the 1950s. Indeed, in the 1950s in the USA and from the early 1960s in Europe, drug use and abuse developed considerably and behaviors appeared not to be as socially and culturally patterned as they were previously considered: chronic and massive individual use started to spread in slum neighborhoods. For that reason, drug addiction was soon viewed as a kind of deviance, and not surprisingly, sociological theories concerning drug addiction, drug experiences, and drug addicts emerged from the sociology of deviance, at least at the beginning: The theoretical issues were then quite similar to those existing in this older sociological field. Apart from the diversity of the results of these studies, they almost all contributed, willingly or not, to the denial of commonly held beliefs and the contradiction of medical conceptions (see e.g., the causal theory developed by Dole and Nyswander 1967 about opiate addiction as metabolic disease) and psychological reasoning (see e.g., the psychoanalytic points of view of the French specialists Bergeret and Fain 1981 which considers drug addiction as basically determined by traumas which have occurred during childhood). All the more so since these points of view quickly dominated the discourse on drug care policies. Indeed, research in social sciences about drug addiction was a lot less developed than biomedical and epidemiological studies (Berridge 1988, Ogien 1992).

3. *Ethnography and Sociology of Drug Use and Addiction*

The theoretical perspective of most of the sociological research of the 1950s and 1960s was heavily grounded in symbolic interaction, labeling, and role theory. An exception is the theory of Cloward and Ohlin (1960), inspired by the work of Finestone (1957), in which drug addiction is considered to be an extreme case of 'retreatism adaptation' (known as the 'double failure' hypothesis), emphasized by the functionalist theory of anomie by Merton (1949) and which attempts to discover a social causality for drug use. After having criticized the perspective of theories (Cloward and Ohlin, Merton) that are primarily concerned with social structure, and having distanced themselves from social psychological research (Jessor and Jessor 1977, Kandel 1980) which tries to discover the etiology of drug addiction in individual dispositions and in the immediate social context of the individual, most of these works focused on the mechanisms of social control (Becker 1963), but above all, on social aspects of drug use and drug addiction experiences as well as on the lifestyles of drug users and addicts (Ogien 1992).

Though now much criticized (McAuliffe and Gordon 1974, Stephens 1991), the first and most famous of these studies are certainly those of Lindesmith (1938, 1947, 1968), instigated by Herbert Blumer of the Chicago School of sociology: After having intensively interviewed about 60 heroin addicts, he concluded that if drug addicts keep on taking opiates, it is not in search of euphoria but to counter the pain suffered by withdrawal. From this perspective, he pointed out, as Becker (1953, 1963, 1967) did later on, the fact that to become addicted, one has to recognize the effects of the drug and to associate these effects with it. Becoming a drug addict is a learning process in group interaction, by which, in contradiction to the ideas of many, the pharmacological properties of the drug are not sufficient in themselves. Thus, Lindesmith suggested, to be an opiate addict, one has (a) to recognize withdrawal symptoms as resulting from abstinence and (b) to take drugs again conscientiously in order to neutralize those symptoms. This theory explains the well-known fact according to which a large number of people who have been prescribed opiates for medical reasons do not become addicted: They would associate the withdrawal symptoms with tiredness or another pathological cause.

Yet another set of studies, a little later, greatly contributed to the understanding of the social conditions of the lives of many drugs users and addicts and helped change common misperceptions about drug addiction. The well-known survey of Preble and Casey (1969) thus depicted drug addicts 'taking care of business' as 'compulsive hardworking business executives,' harassed with many intensive activities (find the dealer, get good dope, avoid the police, etc.) in which the moment of euphoria following absorption is very brief. Another study by Feldman (1968, 1973) conducted in the Lower East Side of New York argued that using heroin can be a means of obtaining the highest status in the street hierarchy (to be a 'stand-up cat' and to avoid being considered a 'chicken,' 'punk,' 'square,' or 'faggot'), and thus ruled out the thesis, described by Cloward and Ohlin, of drug addiction as a retreat strategy (see also Lindesmith and Gagnon 1964). Both studies seem to prove that many drug addicts can manage organized activities and the pursuit of an eventful life (see Adler 1985, Cusson 1998, Johnson et al. 1985, Williams 1989 on user–sellers and drug dealers' activities), and that for these reasons drug addiction cannot be regarded as a way of escaping reality. These studies also suggest that drug addicts are not so different from other social actors, in

that they have to maintain lasting social relationships and develop all kinds of skills and competences (hustling, avoiding getting caught by police, finding a trustworthy dealer, learning how to administer drugs, etc.) to survive in the street and to cope with their passion. For Stephens (1991), street addicts from the inner cities of US large urban areas, in fact, adopt a singular lifestyle which certainly contains many disadvantages but which may give dignity and a sense of belonging to them, while also being full of excitement. Concerning this last point, many studies (see e.g., Johnson et al. 1990) point out that some drug addicts (especially those who are also engaged in drug-dealing and can thus afford a lifestyle that they would not have if working in the legal economy, see Adler 1985), are as much addicted to the lifestyle linked to drug consumption as to the drugs themselves.

Moreover, analysis of the data not only sheds light on the social conditions of drug addicts' lives, but also proves that there are many users who do not necessarily become addicted and go downhill. After having demonstrated that some drug addicts could terminate their addiction and easily achieve abstinence by changing context (such as soldiers in Vietnam who abused heroin but quit it once they returned to the USA), Zinberg (1974), in a very important book, and also Castel and Coppel (1991), have offered evidence that many chronic drug users can continue to have a normal social life while using very addictive drugs such as heroin. Consumption can not only be self-controlled but also regulated in all kind of rituals, and addicts usually learn what is socially acceptable in terms of quantity, frequency, and behavior while consuming drugs (see also the study by Zimmerman and Wieder 1977, which describes how marijuana use in groups is a socially regulated practice); in fact, many regular drug users appear to be able to control their consumption and to remain mild habitual users, and can interrupt their habit for months or years. Finally, drug addiction, these studies suggest, would characterize the situation only of those who have lost control of their habit: becoming a 'greasy dope fiend' is not the destiny of every user. Other well-known studies (Biernacki 1986, Hanson et al. 1985) have come to similar conclusions: There are many regular users unknown to the police, judicial, or hospital authorities (the 'hidden populations') and among whom many belong to the middle class (Granfield and Cloud 1996). These individuals are capable of not becoming addicted and can manage the fact that they are participating both in the world of normality and the world of delinquency or criminality. But these studies also reveal that drug addicts do not necessarily become delinquent (Biernacki 1986), and that the causality between drug addiction and delinquency is not so obvious. Many studies (for a synthesis see Jobart and Fillieule 1999) wish thus to prove that delinquency, for a lot of people, precedes drug use and that drug use has only intensified an old practice of delinquency. But they also explain why the commonly held belief according to which drug use causes delinquency and crime is so widespread: One of the reasons could be the fact that police officers usually arrest drug addicts who are not successful at delinquent activities because they were not socialized in such a culture and social context. As a result, most of the police statistics accredit the idea that drug addiction is a cause of delinquency (Jobart and Fillieule 1999).

More recently, a great deal of work has tried to consider the reasons why individuals first try drugs, become and remain addicted but above all, achieve abstinence. This last entry point (cessation stage) is an interesting one because biological, social pathologies, or psychological theories have difficulty in explaining why many addicts stop using drugs. This type of sociological research takes from interactionism the concept of career, adopts a more rationalist view on the behavior of the addict, is more focused on decision making and choice of the actors. It has certainly produced results that contradict conventional wisdom but also medical, social psychological, and psychological perspectives like psychoanalysis, for instance, which considers drug addicts as basically determined by unconscious psychic forces (Bergeron 1999) and has never taken seriously addicts' reasons for their behavior. Taking into account the fact pointed out by many epidemiological surveys according to which most heroin addicts are free of heroin when they are 35 years old, Castel (1994, 1998), Ogien (1994), inspired by the research of Stimson and Oppenheimer (1985), and also Bennett (1986) and Stephens (1991) have tried to discover the reasons for quitting drugs (e.g., the expense; marriage, social, and medical risks; end of euphoria; experience of violence; imprisonment, etc.). Here they demonstrate that drug addicts have to build or rebuild social networks that can help them on the road to a drug-free life and that many of them can achieve abstinence by themselves (Waldorf and Biernacki 1981), without any professional or medical help. In this perspective, drug addiction is not, as Biernacki (1986) showed, untreatable: It is not a fatal sickness, and the phrase 'Once an addict, always an addict' is no longer relevant.

4. *Conclusion*

Finally, all these studies have tried to destroy both the psychological predisposition theory (psychologically damaged individual) and the social structural weakness of 'poor suburbs' theory and have succeeded in drawing a more complex and diversified image of reality concerning drug addicts and drug experiences. They, of course, do not deny the fact that drug addiction remains for many people a terrible experience, but at the same time they offer evidence that drug addicts are not all doomed to go downhill, that the pharmaceutical properties are not sufficient to

make somebody become an addict, that drug addiction is a learning process in association with others, in which one has to learn how to use drugs and to appreciate their effects, that some can control their consumption even while using very addictive drugs, that drug addiction can be a lifestyle that is chosen by people adopting the role appropriate to this way of life, etc. There is no longer any reason for the exclusive domination of the 'sickness model' (Stephens 1991) in the way societies respond to the drug addiction phenomenon. In this perspective, sociology and political science have studied state reaction to such a social problem and helped discover which models and values determine the public policies of each Western country. Most of the researchers have thus tried to discover specificities of national drug-care policies (see e.g., Bergeron 1999, Cattacin et al. 1996, *Déviance et Société* 1999, Tham 1995), and the conclusion has been the same: Drug control and drug care policies of Western countries are now interdependent and it will be more and more difficult to maintain autonomous views and solutions on this issue (Albrecht and Van Kalmthout 1989). In addition, many sociologists, like Ehrenberg (1998), argue that the now old-fashioned barriers between alcohol, illicit drugs, and psychotropic medicines must be abolished: public health authorities should pay more attention to the way people use a wide range of psychotropic products, legal or illegal, in our modern societies. It is thus clear that policies in Western countries toward addiction problems have to take these research results into account.

See also: Alcohol-related Disorders; Clinical Psychology: Manual-based Treatment; Drinking, Anthropology of; Drug Addiction; Drug Use and Abuse: Psychosocial Aspects; Drugs and Behavior, Psychiatry of; Drugs: Illicit Use and Prevention; Drugs: International Aspects; Health Behaviors; Interactionism: Symbolic; Prostitution; Public Health as a Social Science; Smoking and Health; Smoking Prevention and Smoking Cessation

Bibliography

Adler P A 1985 *Wheeling and Dealing: An Ethnography of Upper-Level Drug Dealing and Smuggling Communities*. Columbia University Press, New York

Akers L A 1991 Addiction: The troublesome concept. *Journal of Drug Issues* **21**(4): 777–93

Albrecht H-J, Van Kalmthout A (eds.) 1989 *Drug Policies in Western Europe*. Criminological Research Reports, vol. 41. Max Planck Institute for Foreign and International Penal Law, Freibourg, Germany

Bachmann C, Coppel A 1989 *Le dragon domestique. Deux siècles de relations étranges entre l'Occident et la drogue*. Albin Michel, Paris

Becker H S 1953 Becoming a marihuana user. *American Journal of Sociology* **59**(3): 235–43

Becker H S 1963 *Outsiders*. The Free Press, New York

Becker H S 1967 History, culture and subjective experience: An exploration of the social bases of drug-induced experiences. *Journal of Health and Social Behavior* **8**(3): 163–76

Bennett T 1986 A decision-making approach to opioid addiction. In: Cornish D B, Clarke R V (eds.) *The Reasoning Criminal. Rational Choice Perspectives on Offending*. Springer-Verlag, New York, pp. 83–102

Bergeret J, Fain M 1981 *Le psychanalyste à l'écoute du toxicomane*. Dunod, Paris

Bergeron H 1999 *L'État et la toxicomanie. Histoire d'une singularité française*. Presses Universitaires de France, Paris

Berridge V 1988 *Drug Research in Europe*. Institute for the Study of Drug Dependence, London

Biernacki P 1986 *Pathways From Heroin Addiction. Recovery Without Treatment*. Temple University Press, Philadelphia, PA

Castel R 1994 Les sorties de toxicomanie. In: Ogien A, Mignon P (eds.) *La demande sociale de drogue*. La Documentation Française, Paris, pp. 23–30

Castel R, Bénard-Pellen M, Bonnemain C, Boullenger N, Coppel A, Leclerc G, Ogien A, Weinberger M 1998 *Les sorties de la toxicomanie*. Editions Universitaires de Fribourg Suisse, Fribourg, Switzerland

Castel R, Coppel A 1991 Les contrôles de la toxicomanie. In: Ehrenberg A (ed.) *Individus Sous Influence*. Editions Esprit, Paris, pp. 237–56

Cattacin S, Lucas B, Vetter S 1996 *Modèles de politique en matière de drogue. Une comparaison de six réalités européennes*. L'Harmattan, Paris

Cloward R A, Ohlin L E 1960 *Delinquency and Opportunity: A Theory of Delinquent Gangs*. Free Press, New York

Cusson M 1998 Le trafic de drogue. In *Criminologie actuelle*. Presses Universitaires de France, Paris

Déviance et Société 1999 Politiques publiques et usage de drogues illicites. *Déviance et Société* **23**(2)

Dole V P, Nyswander M E 1967 Heroin addiction: Metabolic disease. *Archives of Internal Medicine* **120**: 19–24

Ehrenberg A (ed.) 1998 *Drogues et médicaments psychotropes. Le trouble des frontières*. Éditions Esprit, Paris

Escohotado A 1998 *Histoire des drogues. Des origines au XVIIe siècle*. Lézard, Paris

Feldman H W 1968 Ideological supports to becoming and remaining a heroin addict. *Journal of Health and Social Behavior* **9**(2): 131–9

Feldman H W 1973 Street status and drug users. *Society* **10**: 32–9

Finestone H 1957 Cats, kicks and color. *Social Problems* **5**: 3–13

Granfield R, Cloud W 1996 The elephant that no one sees: Natural recovery among middle-class addicts. *Journal of Drug Issues* **26**(1): 45–61

Hanson B, Beschner G, Walters J M, Bovelle E 1985 *Life With Heroin: Voices from the Inner Cities*. Lexington Books, Lexington, MA

Jessor R, Jessor S L 1977 *Problem Behavior and Psychosocial Development: A Longitudinal Study of Youth*. Academic Press, New York

Jobart F, Fillieule O 1999 Action publique sous dépendance. Conditions et effets du changement de paradigme dans la lutte contre la délinquance associée à la drogue en Europe. *Revue française de sciences politiques* **49**(6): 803–34

Johnson B D, Goldstein P, Preble E, Schmeidler J, Lipton D S, Spunt B, Miller T 1985 *Taking Care of Business: The Economics of Crime by Heroin Abusers*. Lexington Books, Lexington, MA

Johnson B D, Williams T, Dei K A, Sanabria H 1990 Drug abuse in the inner city: Impact on hard-drug users and the community. In: Tonry M, Wilson J Q (eds.) *Drugs and Crime*. University of Chicago Press, Chicago, pp. 9–67
Kandel D B 1980 Drug and drinking behavior among youth. *Annual Review of Sociology* **6**: 235–85
Knipe E 1995 *Culture, Society, and Drugs: The Social Science Approach to Drug Use*. Waveland Press, Prospect Heights, IL
Lindesmith A R 1938 A sociological theory of drug addiction. *American Journal of Sociology*. **43**(4): 593–613
Lindesmith A R 1947 *Opiate Addiction*. Principia, Bloomington, IN
Lindesmith A R 1968 *Addiction and Opiates*. Aldine, Chicago
Lindesmith A R, Gagnon J H 1964 Anomie and drug addiction. In: Clinard M B (ed.) *Anomie and Deviant Behavior: A Discussion and Critique*. Free Press, New York, pp. 158–88
McAuliffe W, Gordon R 1974 A test of Lindesmith's theory of addiction: The frequency of euphoria among long-term addicts. *American Journal of Sociology* **79**: 795–840
Merton R K 1949 Social structure and anomie. In: Merton R K (ed.) *Social Theory and Social Structure*. Free Press, Glencoe, IL, pp. 125–49
Musto D F 1987 *The American Disease: Origins of Narcotic Control*. Oxford University Press, New York
Ogien A 1992 Situation de la recherche sur les toxicomanies en Europe et aux Etats-Unis. In: Ehrenberg A (ed.) *Penser la drogue. Penser les drogues*. Éditions Descartes, Paris, pp. 49–85
Ogien A 1994 La morale du drogué. *Revue française des affaires sociales* **2**: 59–67
Preble E, Casey J J 1969 Taking care of business. *International Journal of the Addictions* **4**: 1–24
Stephens R C 1991 *The Street Addict Role. A Theory of Drug Addiction*. State University of New York Press, Albany, NY
Stimson G, Oppenheimer E 1982 *Heroin Addiction. Treatment and Control in Britain*. Tavistock, London
Tham H 1995 Drug control as a national project: The case of Sweden. *The Journal of Drug Issues* **25**(1): 113–28
Waldorf D, Biernacki P 1981 Natural recovery from opiate addiction: Some preliminary findings. *Journal of Drug Issues* **11**: 61–74
Weisheit R (ed.) 1990 *Drugs, Crime and the Criminal Justice System*. Anderson, Cincinnati, OH
Williams T 1989 *The Cocaine Kids*. Addison-Wesley, New York
Yvorel J J 1992 *Les poisons de l'esprit. Drogues et drogués au XIXe siècle*. Quai Voltaire, Paris
Zimmerman D, Wieder L 1977 You can't help it but get stoned. Notes on the social organization of marijuana smoking. *Social Problems* **25**(2): 198–207
Zinberg N 1974 *Drug, Set and Setting*. Yale University Press, New Haven, CT

H. Bergeron

Drug Use and Abuse: Cultural Concerns

In popular and scientific publications, addiction is often referred to as a chronic, relapsing disorder. In support of this claim, individuals who obtain treatment for alcoholism, drug addiction, and smoking more often than not relapse (Thombs 1994). However, if the population of addicts is expanded to include those who do not seek treatment a different picture emerges. Most addicts recover. More generally, when the study of addiction includes epidemiological, ethnographic, and experimental research, as well as clinic studies, the image of the addict as a victim of drug cravings becomes difficult to support. In its stead emerges a more complicated picture in which cravings are but one of several competing influences on drug consumption, and addicts actively choose whether or not to continue using drugs. For instance, the state of being addicted, although debilitating, does not preclude the salutary influences of nondrug activities and goals, such as employment, new relationships, and the desire to be a positive influence on others. A woman who was a regular cocaine user for 16 years describes quitting in terms of economics and parenting (Waldorf et al. 1991, p. 202): 'Oh, for a time my nose opened up when I went out partying and drinking, but I learned how to handle it…Once I stopped selling, I didn't have the money to buy it anymore. I would have literally had to say, "Sorry girls [her daughters], you don't eat this week" to buy some.' This article will flesh out these introductory remarks, focusing on three issues: the criteria for distinguishing drug use from drug abuse, the correlates of addiction, including pharmacological, social, and individual factors, and the correlates of recovery.

1. *Terminology: Substance Dependence, Abuse, and Use*

On the basis of a series of collaborative research projects (e.g., Spitzer et al. 1979), clinicians have crafted a set of diagnostic criteria that reliably distinguish three levels of drug use: 'substance dependence,' 'abuse,' and social or recreational use. An important consequence of this advance in terminology is that the results from different studies can be meaningfully compared, thereby setting the stage for a cumulative science of drug use and addiction.

According to the diagnostic manual of the American Psychiatric Association (1994), the 'essential feature' of *drug dependence* 'is a cluster of cognitive, behavioral, and physiological symptoms indicating that the individual continues use of the substance despite significant substance-related problems.' The physiological symptoms include tolerance and withdrawal. The cognitive-behavioral symptoms include taking the substance in larger amounts than was intended, a persistent desire to curtail drug use, increases in time spent in drug-related activity, and drug-related failures in fulfilling obligations at work and at home. In other words, the technical term 'substance dependence' replaces the more commonly used term 'addiction.'

Substance abuse is similar to dependence in that the central feature is continued use despite adverse consequences. However, other aspects of dependence, such as tolerance, withdrawal, and a persistent desire to control drug use, are absent.

Social or recreational drug use is a level of consumption that produces little or no harm. This is often a matter of learning from negative experiences. For example, the APA manual (1994) points out that when social drinkers experience alcohol-related problems, they learn to moderate their future drinking.

The World Health Organization has published a similar set of diagnostic criteria (1992). For instance, in both accounts the key concept is the relationship between drug use and drug-related aversive consequences.

1.1 Earlier Definitions of Substance Dependence

Earlier definitions of substance dependence put more emphasis on tolerance and withdrawal than do the current APA and WHO criteria. This reflects trends in drug use. Until the late 1970s, heroin dependence was the prototypical addiction, and it is characterized by a well defined set of withdrawal symptoms, including spasms ('kicking the habit'), chills, goose bumps ('going cold-turkey'), muscle aches, diarrhea, and agitation. However, in the late 1970s and early 1980s, cocaine use greatly increased. Cocaine leads to many of the same cognitive-behavioral signs of addiction as heroin, but without any obvious withdrawal syndrome. For example, withdrawal from cocaine may elicit dysphoric feelings but it does not produce somatic, flu-like symptoms, as does withdrawal from heroin. Accordingly, the diagnostic significance of withdrawal symptoms has decreased.

1.2 'Addiction' and 'Substance Dependence'

In everyday speech, heavy drug use is usually referred to as 'addiction' not 'substance dependence.' The different terms identify the same patterns of behavior, however 'addiction' has an additional meaning. It implies that drug consumption is involuntary or against one's will. For example, the Oxford English Dictionary defines *drug addiction* as 'a compulsion and need to continue taking a drug as a result of taking it in the past,' and it defines *compulsion* as '[acting according to an] irresistible impulse…' However, the idea that drug dependence entails irresistible urges is not a matter of fact, but a hypothesis that experts do not agree on. For example, experimental, clinical, and autobiographical reports show that addicts learn to resist drug cravings (e.g., Cohen et al. 1971, Waldorf 1983), especially when offered incentives for abstinence. Thus, the prudent course is to use the term 'substance dependence.' Its meaning is restricted to a pattern of behavior that different observers reliably identify, and it does not imply a mechanism as to why drug use persists. (Although 'addiction' is an ambiguous word, it often fits better into sentences. Accordingly, when the word 'addiction' or 'addict' is used in this article, the referent is a pattern of behavior or a drug user who exhibits that pattern of behavior.)

2. Correlates of Drug Abuse

The transition from drug use to abuse depends on the interplay of three factors: characteristics of the drug, the setting, and the individual.

2.1 Pharmacological Impact

The importance of the drug is illustrated by two comparisons. First, although there are scores of highly reinforcing substances and activities, only a few support addiction, and these few are all drugs (with the possible exception of gambling). Second, changes in drug structure or route of administration can markedly influence the likelihood of addiction. For example, opiates were used for thousands of years with little or no mention of the dangers of addiction. However, in China during the eighteenth century, opium dependence became a national problem and the subject of governmental legislation. The key precipitating event was a change in the manner of consumption. The Chinese began smoking opium instead of taking it orally. Smoking increases the rate at which the psychoactive components of opium (e.g., morphine) enter the circulatory system, and this, in turn, greatly increases brain levels of these substances. There are similar stories for alcohol (the development of distillation), tobacco (the development of cigarettes), and stimulants (the development of the syringe and smokable forms). In each case, increasing the rate and magnitude of the pharmacological effect increased addiction rates.

However, pharmacology was not the only factor in the rise of opiate addiction in China. During the eighteenth century, smoking and opium use were familiar practices in many parts of the world, but only in China did the two combine to produce a significant social problem. Historical studies (e.g., Spence 1975) suggest that members of the bureaucratic classes were especially vulnerable to opiate addiction, and this may have been due to their wealth and available free time.

2.2 A Pharmacological Common Denominator for Dependence?

Much of the research on the pharmacological basis of addiction has been guided by the hypothesis that dopamine, a neurotransmitter that plays a key role in

reward and motor activity, is an essential component of the biology of addiction (e.g., Wise and Bozarth 1987). In support of this hypothesis, laboratory studies with animals show that self-administration of addictive drugs increases the availability of dopamine in areas of the brain that play an important role in motivated behavior, and, conversely, drugs that block dopamine receptors decrease or eliminate drug self-administration in animals. However, other laboratory findings indicate that it is too simplistic to equate addiction with dopamine. For instance, many non-addictive consummatory activities are associated with increases in brain dopamine levels, and in a series of recent experiments, mice self-administered cocaine even though they were missing the protein that mediates cocaine's ability to increase dopamine levels (Caine 1998). Thus, increases in brain dopamine levels are not a sufficient condition for drug dependence or a necessary consequence of drug self-administration.

Although there may not be a common receptor-level event for different addictions, there are common drug-related behavioral effects (Heyman 2001). These include rapid onset of action, few if any inhibiting mechanisms for self-administration (e.g., drugs do not satiate), intoxication (except for cigarettes), and the capacity to undermine the ability to engage in competing, nondrug related activities. For example, intoxication and withdrawal symptoms interfere with most work responsibilities and social conventions.

2.3 Setting

Interviews and questionnaires provide much of the information on the influence of setting on drug use. As these research methods depend on self-report, the results may not be valid. However, the literature presents a coherent account of drug use, with different studies producing quantitatively similar outcomes (e.g., compare Anthony and Helzer 1991 with Warner et al. 1995).

One of the strongest cases for the influence of setting on drug consumption is the story of opiate dependence among American soldiers in Vietnam (Robins et al. 1975). In contrast to conditions at home, the men entered a situation in which opiates, including heroin, were widely available, cheap, and largely free of sanctions.

Among the young enlistees, opiate use increased from less than 5 percent to 43 percent. Among those who used just once, about 60 percent became addicted. Among those who used more than four times, about 90 percent became addicted. However, upon return to the United States, dependence rates declined to near their pre-Vietnam levels (from about 20 percent to 1 percent). Even most of those who were regularly injecting heroin while in Vietnam were not dependent on heroin during the three-year period that they participated in the study.

The researchers were able to provide some information about the specific factors that influenced opiate dependence and recovery. Their results suggest that the increase in dependence was largely due to the increased availability of opiates, and that, conversely, the decline in dependence was largely due to the reassertion of negative drug-related consequences. For example, the veterans reported that they knew where to buy heroin—and 50 percent did at least once—but that they quit because of the 'sordid' atmosphere that surrounded its purchase and use, the higher prices, the criticism from relatives and friends, and the possibility of being arrested.

2.4 Individual Differences

Individual differences play an important role in the etiology of substance dependence. For instance, in studies that compare same sex siblings or neighbors, the adolescent with a history of delinquency is significantly more likely to experiment with addictive drugs, especially heroin, and given experimentation, delinquency predicts dependence. However, it is not yet clear which individual traits are most important in relation to drug use and dependence. The following examples were selected as they are intuitively plausible and well established.

2.4.1 Psychiatric comorbidity. In the early 1980s the US National Institutes of Health sponsored a nationwide survey of psychiatric disorders and their correlates in a representative national sample (the Epidemiological Catchment Area study, Robins and Regier 1991). There were approximately 20,000 subjects, and they were evaluated according to the APA criteria for mental disorders. For those with a history of substance dependence or abuse, the expected likelihood of a nondrug related psychiatric disorder was 22 percent (Regier et al., 1990). The observed frequency was more than twice as great, 53 percent. Moreover, the likelihood of additional psychiatric disorders varied with type of drug abuse. For example, it was 65 percent and 76 percent for opiate and cocaine dependence, respectively. These results are consistent with (but do not prove) the ideas that substance dependence is sometimes a form of self-medication (for the underlying psychiatric disorder), and that addicts with psychiatric disorders are less likely to engage in activities that would displace drug use, such as holding a good job.

2.4.2 Individual differences in pharmacological response. There is also evidence of etiologically significant individual differences in the pharmacological response to addictive drugs. For instance, in a long-

itudinal study, the intoxicating effects of alcohol at age 20 predicted the development of alcoholism at age 30 (Schuckit 1994). Surprisingly, the men who reported the lowest intoxication scores were most likely to become alcoholic. The author reasoned that feeling less intoxicated invites heavier drinking.

2.4.3 Is there a genetic basis for some types substance dependence? The discussion of individual differences can be taken one step further. In general, individual differences reflect both genetic and experiential factors. Thus, there may be a genetic predisposition for substance dependence. This has been studied in regard to the etiology of alcoholism.

The basic finding is that male children of alcoholic fathers are about four times more likely to become alcoholics than male children of nonalcoholic fathers. Twin and adoption studies show that a significant component of this difference is heritable. For instance, in one frequently cited report (Pickens et al. 1991), the correlation for alcoholism was about 60 percent higher in identical twin boys than in fraternal twin boys (concordance rates of 0.59 and 0.36). The research also shows that the degree of heritability varies according to gender and pattern and level of drinking.

But heritability only specifies the extent to which genes influence the likelihood of the trait of interest. For both clinical and scientific purposes, it would be useful to know the processes that the genes influence. Broadly speaking, the possibilities are direct reactions to alcohol, such as tolerance, or behaviors that indirectly influence drinking, such as impulsivity and delinquency. These are not mutually exclusive outcomes, and the available data support both pathways (Cloninger 1987, Schuckit 1994). Also note that gene-influenced behavioral processes do not ineluctably lead to alcoholism. Rather, they establish a bias for alcoholism, just as Vietnam established a bias for opiate addiction.

2.5 On the Prerequisite Conditions for Substance Dependence

In representative populations, unassisted recovery from heroin addiction is common (e.g., Waldorf 1983). As heroin is the most addictive drug, this suggests that pharmacological processes establish a necessary but not a sufficient condition for addiction. The Vietnam experience demonstrates the importance and the limits of environmental influences. With easy access to heroin, there was about a 20-fold increase in addiction, yet 40 percent of those who experimented with this drug did not go on to become regular users. This suggests that there is no setting in which all drug users will become drug abusers. Similarly, the twin data show that individual characteristics matter, but not enough so that pharmacology or setting can be ignored. Strictly speaking, then, a substance is not addicting. Rather some substances are more likely than others to become the focus of an addiction given the presence of other contributing factors. Similar statements can be made about dependence promoting settings and individual characteristics.

3. Recovery: Is Substance Dependence a Chronic Relapsing Disorder?

Clinic outcome studies show that individuals in treatment for substance abuse often relapse back to drug use. However, most of the men who were regular users of heroin in Vietnam stopped using upon return home. Which population is representative of most addicts? There are two large, epidemiological studies of psychiatric disorders that collected information on drug use (Anthony and Helzer 1991, Warner et al. 1995). Both found that most of those who at some point in their life met the criteria for substance dependence no longer did so. And according to the larger of the two studies (Robins and Regier 1991), substance abuse had the shortest duration and highest recovery rate of any other psychiatric disorder.

The common factor in the studies with higher recovery rates is that subjects were selected independently of their treatment history. In contrast, in studies that find high relapse rates, the subjects are drawn from clinic rosters. The clinic data include two biases. First, most addicts do not seek treatment (e.g., Anthony and Helzer 1991). Second, the frequency of psychiatric comorbidity is much greater for the clinic population. For example, in the ECA study, 64 percent of the addicts who were in treatment had an additional nondrug related psychiatric diagnosis, whereas the comorbidity rate for addicts not in treatment was 29 percent (Regier et al. 1990). Thus, higher relapse rates for clinic populations are correlated with higher comorbidity rates.

4. Factors that Influence Recovery

4.1 Values, Employment, and Cultural Milieu

Autobiographical accounts of recovery often include a pivotal event that triggers evaluative judgments about drug use, such as the desire to change in order to please parents and children. In support of these accounts, which may be biased, an ethnography of a population of recovered heroin addicts found that the most frequently cited reason for change was an event that led to feelings of regret about addiction or the junkie lifestyle (Waldorf 1983).

In a series of longitudinal studies on heroin addicts and alcoholics, the strongest correlate of recovery was the ability to hold a job (Vaillant 1992). The author hypothesized that a job weakens drug use by removing

the addict from drug taking cues and by reinforcing patterns of behavior that are incompatible with drug use, such as keeping to a daily schedule.

Since 1964 more than 50 million people have quit smoking in the United States (e.g., Schelling 1992). This decline is correlated with the 1964 publication of the Surgeon General's report on the health risks of cigarette smoking and the subsequent changes in the culture of cigarette smoking. For instance, movies no longer represent the cigarette as a sign of sophistication, and many if not most public places are now smoke free.

4.2 Pharmacological Treatments

Methadone and nicotine replacement therapies increase abstinence rates for addicts in treatment. This does not contradict the research on the situational correlates of recovery. Rather, by attenuating the distracting effects of withdrawal symptoms, the pharmacotherapies increase the likelihood that conventional alternatives to drug use will take hold.

4.3 Correlates of Recovery and the APA Criteria for Substance Dependence

An important theme in the study of recovery is that factors that are important in bringing drug use to a halt play much less of a role, if any, in disorders that are widely acknowledged as entailing compulsions, such as obsessive compulsive disorder or Tourette's syndrome. For example, incentives and social disapproval can persuade addicts to stop using drugs (Higgins et al. 1994, Robins et al. 1975) but, in general, do not markedly decrease symptoms such as repetitive hand washing or motor tics. This comparison suggests that the APA criteria for substance dependence although useful are overstated. Addicts may be less sensitive to the negative consequences of drug use than nonaddicts, but, as shown in laboratory and natural experiments (e.g., Vietnam), they are not immune to them. Put another way, the nature of recovery from addiction implies that it is not a compulsive disorder.

5. Brain, Substance Dependence, and Recovery

In recent brain imaging studies, drug users and control subjects differed in terms of dopamine receptor density and dopamine related neural functioning (e.g., Volkow and Fowler 2000). Interestingly, the differences were in areas of the brain that are believed to play a role in decision making. This work is new, and it is not known if the imaging results reflect drug use, the correlates of drug use (e.g., a pre-existing state), or both. Similarly, it is not known if the neural correlates of drug use influence decision-making capacities or preference for the drug (leaving rationality intact). A further complexity is that the relationships between drug consumption, brain function, and behavior can also be initiated by changes at the 'surface.' Just as drugs may alter the brain in ways that promote future drug consumption, the environmental events and changes in behavior that lead to recovery must modify the brain in ways that decrease the likelihood of future drug consumption. For example, when cognitive-behavioral treatments for obsessive-compulsive disorder are effective they lead to changes in brain structures that are correlated with the patients' symptoms (Schwartz 1998). A number of research groups are beginning to study the complex interrelationships between drug use, decision-making, and brain functioning. Thus, the next generation of addiction studies promises a wealth of new information on the biology and psychology of decision-making as well as on the biology and psychology of substance use and abuse.

See also: Drug Addiction; Drug Addiction: Sociological Aspects; Drug Use and Abuse: Psychosocial Aspects; Drugs and Behavior, Psychiatry of; Smoking and Health; Smoking Prevention and Smoking Cessation

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. (DSM-IV). American Psychiatric Association, Washington, DC

Anthony J C, Helzer J E 1991 Syndromes of drug abuse and dependence. In: Robins L N, Regier D A (eds.) *Psychiatric Disorders in America*. Free Press, New York

Caine S B 1998 Cocaine abuse: Hard knocks for the dopamine hypothesis? *Nature Neuroscience* **1**: 90–2

Cloninger C R 1987 Neurogenetic adaptive mechanisms in alcoholism. *Science* **236**: 410–16

Cohen M, Liebson I, Faillace L, Speers W 1971 Alcoholism: Controlled drinking and incentives for abstinence. *Psychological Reports* **28**: 575–580

Herrnstein R J, Prelec D 1992 A theory of addiction. In: Loewenstein G, Elster J (eds.) *Choice over Time*. Russell Sage Foundation, New York

Heyman G M 2001 Is addiction a chronic relapsing disease? Relapse rates, estimates of duration, and a theory of addiction. In: Heymann P, Brownsberger (eds.) *Drug Addiction and Drug Policy*. Harvard University Press, Cambridge, MA

Higgins S, Budney, A, Bickel W, Foerg F, Donham R, Badger G 1994 Incentives improved outcome in outpatient behavioral treatment of cocaine dependence. *Archives of General Psychiatry* **51**: 568–76

Julien R M 1995 *A Primer of Drug Action*. Freeman, New York

Latimer D, Goldberg J 1981 *Flowers in the Blood: The Story of Opium*. Franklin Watts, New York

Pickens R W, Svikis D S, McGue M, Lykken D T, Heston L L, Clayton P 1991 Heterogeneity in the inheritance of alcoholism. *Archives of General Psychiatry* **48**: 19–28

Regier D A, Farmer M E, Rae, D S, Locke B Z, Keith S J, Judd L L, Goodwin F K 1990 Comorbidity of mental disorders with alcohol and other drug abuse. *Journal of the American Medical Association* **264**: 2511–18

Robins L N, Helzer J E, Davis D H 1975 Narcotic use in Southeast Asia and afterwards. *Archives of General Psychiatry* **32**: 955–61
Robins L N, Regier D 1991 *Psychiatric Disorders in America*. Free Press, New York
Schelling T C 1992 Addictive drugs: The cigarette experience. *Science* **255**: 430–3
Schuckit M A 1994 Low level of response to alcohol as a predictor of future alcoholism. *American Journal of Psychiatry* **151**: 184–9
Schwartz J M 1998 Neuroanatomical aspects of cognitive-behavioural therapy response in obsessive-compulsive disorder. *British Journal of Psychiatry* **173** (Suppl 35): 38–44
Spence J 1975 Opium smoking in Ch'ing China. In: Wakeman F, Grant C (eds.) *Conflict and Control in Imperial China*. University of California Press, Berkeley, CA
Spitzer R L, Forman J B, Nee J 1979 DSM-III field trials, I: Initial interrater diagnostic reliability. *American Journal of Psychiatry* **136**: 815–17
Thombs D L 1994 *Introduction to Addictive Behaviors*. Guilford Press, New York
Vaillant G E 1992 Is there a natural history of addiction? In: O'Brien C P, Jaffe J H (eds.) *Addictive States*. Research publications: Association for Research in Nervous and Mental Disease, Vol. 70. Raven Press, New York
Volkow N D, Fowler J S 2000 Addiction, a disease of compulsion and drive: Involvement of the orbitofrontal cortex. *Cerebral Cortex* **10**: 318–25
Waldorf D 1983 Natural recovery from opiate addiction: Some social-psychological processes of untreated recovery. *Journal of Drug Issues* **13**: 239–79
Waldorf D, Reinarman C, Murphy S 1991 *Cocaine Changes*. Temple University Press, Philadelphia, PA
Warner L, Kessler R C, Hughes M, Anthony J C, Nelson C B 1995 Prevalence and correlates of drug use and dependence in the United States. *Archives of General Psychiatry* **52**: 219–29
Wise R A, Bozarth M A 1987 A psychomotor stimulant theory of addiction. *Psychological Review* **94**: 469–92
World Health Organization 1992 *The ICD-10 Classification of Mental and Behavioral Disorders*. World Health Organization, Geneva

G. Heyman

Drug Use and Abuse: Psychosocial Aspects

The use of psychoactive drugs has existed among various human cultures through the millennia. The abuse of many such drugs has been an important though socially defined, adverse consequence of their use.

1. *Psychoactive Drugs*

The term 'drugs' as commonly used today is a shorthand for 'psychoactive drugs'—compounds or substances which alter the neurological functioning and/or structure of the organism in such a way as to directly influence perception, mood, consciousness, and/or energy levels. There are many compounds that have such effects, ranging from those occurring naturally in plants (e.g., cocaine, caffeine, nicotine, cannabis, opium, psilocybin) and natural process (fermentation of fruit and other plant products containing carbohydrates, yielding alcohol) to compounds that are synthesized (e.g., amphetamines, LSD, 'ecstasy,' and PCP).

2. *Distinguishing Characteristics of Drug-using Behaviors*

There is one central characteristic that distinguishes psychoactive drug use from most other classes of behavior. Drugs influence the neurological system directly and provide pleasure or other forms of reinforcement (Koob and Nestler 1997). This reinforcement occurs without any of the intervening 'adaptive behaviors' which normally would be associated with such reinforcement because, in evolutionary terms, those behaviors that were adaptive evolved as having reinforcing effects due to their importance in the survival of the species. In essence, drug use involves a shortcut to the pleasure centers of the brain—a bypassing of adaptive behavior. Two other characteristics, which derive from this central one, also characterize this domain of behavior. First, because of this shortcut, drug-using behavior can come to interfere with, or displace, many behaviors that are adaptive such as eating, forming appropriate bonds to family and community, working, or performing well academically. Put another way, it can come to assume a higher position in the hierarchy of needs. It can also disrupt fundamental psychological processes essential to performing a wide range of other behaviors or tasks by altering or impeding perception, memory, decision making, motivation, and energy levels. Second, the direct reinforcing properties of psychoactive drugs on the neurotransmitters in the brain can alter the structure and function of the neurological system itself, such that a need is created for the drug in order for the individual to attain what had previously been a normal level of neurological functioning. In other words, the user can become dependent on or addicted to the drug; thus, the underlying motivation for using a drug can shift quite dramatically from what might have motivated use at earlier stages of involvement.

3. *Illicit vs. Licit Drug Use*

There are many psychosocial determinants of the use of psychoactive substances, two of the most important of which are the legal and normative status of particular substance-using behaviors in the society. The primary focus in this article will be on psychoactive substances, the use of which is considered illicit in most societies. Tobacco and alcohol are generally

licit to use and legal to possess; but because they are dealt with at some length in other entries, they will receive only very general treatment here. It should be noted, however, that even the use of alcohol and tobacco by people under certain ages is considered illegal and illicit. Three general social classifications of psychoactive drugs will be distinguished, and the importance of their classification considered: psychotherapeutic drugs, legal drugs, and illegal drugs.

3.1 Psychotherapeutic Drugs

Certain drugs can be either licit or illicit to use, depending on the conditions under which use occurs; and they can be legal or illegal to possess, depending on how they have been obtained. A number of psychoactive drugs—the so-called psychotherapeutic drugs—have medical applications, so their use is considered as legitimate for the medically supervised therapeutic treatment of particular conditions. These drugs include tranquilizers (most of which are benzodiazepines, used to alleviate anxiety), sedatives (most of which are barbiturates, used to induce sleep), amphetamines (used to treat hyperactivity disorders), and various opium derivatives and their synthetic analogues (used to control pain, coughing, and diarrhea). All of these drugs have regulated access, through the requirements of a doctor's prescription and pharmacy dispensation. However, the sale, possession, or use of these substances outside of prescribed medical regimen is considered illicit and illegal, in part because all of these drugs have the potential to produce dependence in the user; and dependence is presumed to be more likely to develop under conditions of unsupervised use.

3.2 The Legal Drugs

Certain psychoactive substances have emerged as legitimate to use in most societies outside of any medical regimen and without any therapeutic purpose, usually because at the time of their emergence, they are presumed to be safe. These include alcohol, nicotine, and caffeine. All of these substances have been in fairly widespread use for some centuries in different cultures. Although there have been some concerns in the past about the health and societal consequences of alcohol and tobacco, such concerns increased considerably over the course of the twentieth century, leading to increased stigmatization of their use and legal controls on their sale and distribution.

3.3 The Illegal Drugs

In the twentieth century most other psychoactive substances have been deemed illegal and illicit to use, possess, or sell in most societies. This illegal status is based largely on their presumed negative impact on the functioning of the individual, family, and/or larger society. Various international treaties have had the effect of making the legal status of various of these drugs (plus most of the psychotherapeutic drugs) fairly homogenous throughout the world, sometimes changing traditional and accepted drug using behaviors in certain societies into illegal and illicit ones (United Nations 1971, 1972). The potential beneficial uses of these drugs, if any, generally are considered to be outweighed by potential negative consequences. Among them are marijuana and hashish, various hallucinogenic drugs, including LSD and PCP, cocaine powder and crack cocaine, methamphetamine, opium and heroin.

The term 'presumed' is used throughout this discussion to describe both the negative effects and the benefits of these drugs, because the presumptions sometimes turn out to be quite erroneous. This was true of the presumed safety of tobacco products and of cocaine, to take but two examples. Also, there remains considerable controversy, and incomplete scientific knowledge, about the consequences of various of the psychoactive substances in use today, particularly the newer ones.

The boundaries of social legitimacy and illegality of the use of various psychoactive substances are themselves subject to change, as knowledge accumulates about the effects of the drugs and also about the effects of the control mechanisms intended to control their use. In fact, there has been a vigorous policy debate over whether reclassifying a previously illegal drug, like marijuana, as a psychotherapeutic drug might also affect the extent to which that drug is used in society outside of its newly legitimated medical regimen. Some argue that such a reclassification would destigmatize the drug and make its use seem less dangerous—both important psychosocial determinants of use, as will be discussed further below.

4. Religious and Cultural Constraints

The fact that many drugs are taken for reasons other than instrumental ones—that is, their use does not have accepted therapeutic purposes and does not facilitate task performance—but rather for pleasure (or other directly reinforcing properties of the drugs) itself stigmatizes their use for many people. Against the Protestant work ethic, such behavior can be seen as slothful, self-indulgent, hedonistic, and therefore inherently immoral. In the Mormon religion alcohol use is explicitly prohibited, and in the Islamic religion the use of alcohol and most other psychoactive substances is banned.

On occasion the use of illegal drugs may take on symbolic meanings that change their acceptability within the society, or at least within certain sectors of

society. A dramatic instance of this occurred in the United States during the Vietnam War era, when the use of marijuana and LSD, in particular, came to be associated with being a part of the counterculture and the antiwar movements of the time (Zinburg and Robertson 1972). Within the context of a major generational division over values, use of these drugs became legitimate, and even positively valued, within a large sector of the youth generation. Use of these drugs at once symbolized defiance of adult norms and a statement of solidarity with those social movements. During that period of generational rift, the normative constraints among youth on the use of these drugs eroded considerably, and thus came to diverge from the prevailing norms in adult society. It was just a few years after this historical period that illicit drug use among adolescents and young adults reached its pinnacle in the country. In the years since, these symbolic meanings have virtually disappeared as the movements themselves have become part of history; and youth norms against use strengthened considerably. Not coincidentally, illicit drug use among young people declined quite substantially between the late 1970s and the early 1990s.

In sum, religious and other cultural considerations, as well as more practical concerns about the costs and benefits of use, and the costs and benefits of various control mechanisms, weigh heavily in determining the legality and social acceptability of using various psychoactive compounds. Legal status and social acceptability, in turn, are among the most powerful psychosocial determinants of use, as is evidenced by the fact that the psychoactive substances that are considered legal and socially acceptable—including alcohol, nicotine, and caffeine—have by far the greatest prevalence of use.

Use of these drugs is widespread not only because their use is legally and normatively accepted, however, but also because their legal status allows those who sell them to engage in massive advertising and promotional campaigns aimed at encouraging and facilitating consumption of the drug. One of the major reasons given by those who oppose the legalization of marijuana or other illegal drugs is that legalizing these drugs would unleash just such powerful marketing forces, which could further dramatically increase use (Inciardi and McBride 1991).

5. *Research Traditions*

Many subareas of research on psychoactive drugs have evolved, particularly in the past forty to fifty years. These include etiology, epidemiology, neurology, psychopharmacology, molecular biology, prevention, and treatment. The psychiatric tradition in epidemiology and treatment has emphasized accurately diagnosing and effectively assigning treatment regimens to individual cases. The other social sciences have looked more at drug use in the normal population, with an emphasis on causes, correlates, and consequences. There have emerged a few large, ongoing epidemiological survey studies of national populations that have as their primary focus the tracking and studying of drug use in the normal population. Two are in the United States: the National Household Survey of Drug Abuse (SAMHSA 2000) and the Monitoring the Future study (Johnston et al. 2000) that began in 1971 and 1975, respectively. More recently a coordinated series of national school surveys, modeled largely after Monitoring the Future, has been launched in more than two dozen countries in Europe under the auspices of the Council of Europe and the government of Sweden (Hibell et al. 2000).

The fact that addiction has long been conceptualized as a disease within certain scientific circles has led to terminology that is medical in nature in defining the subfields of research (e.g., epidemiology, etiology, prevention, treatment). In fact, Alan Leshner, the current director of the National Institute on Drug Abuse, has asserted that drug addiction is a disease of the brain, evidenced by lasting structural and functional changes in the brain induced by drug use, but that drug use short of addiction is a preventable behavior (Leshner 1999).

The importance of the disease concept, for which there is increasing neurological evidence, is that it implies that drug dependence should be responded to as more than just a moral failing. It implies that dependent users have important changes in their neurotransmitter systems that reduce to a considerable degree the extent to which they have reasonable control over their drug-using behavior.

6. *Use vs. Abuse*

One important distinction in psychosocial aspects of drugs has been between use and abuse, though it is a distinction less discussed in the literature today than it was early in the epidemic of illicit drug use. In theory, the distinction lies in whether the drug-using behavior has reached the stage at which it is adversely affecting the normal functioning and well-being of the individual and, therefore, has become 'abuse.' Is the individual's drug use disrupting normal emotional, cognitive, social, or work functioning, including school performance for those in school, or is it endangering physical health? Is it interfering with the user's functioning in key social roles such as spouse, parent, child, friend, student, worker, and citizen? As may be obvious, such a distinction is easier to conceptualize than to operationalize. Beliefs about the adverse effects of drugs influence not only the larger society's view of, and constraints on drug use, but the likelihood that a given individual will engage in using

drugs—a point to which we will return below. Because many of the dangers of a particular drug are not well documented in the early years after it comes onto the scene—witness cocaine use in the late 1970s—'correctly judging whether there are serious health or addiction hazards, say, and therefore whether use constitutes 'abuse,' is not at all straightforward.

Some research on the reasons adolescents give for using various drugs may help to elucidate the theoretical distinction between use and abuse (Johnston and O'Malley 1986). Young people who have used a drug only once or twice usually give as their primary reason curiosity, 'to see what it was like.' Those who have used only a reasonably limited number of occasions, but have gone beyond experimentation, most often mention reasons such as 'to feel good or get high' or 'to have a good time with my friends'—what might be called social and recreational reasons for use. However, those who have used a drug with some frequency start to mention what might be called psychological-coping motives for their use—'to get through the day,' 'to get away from my problems,' 'because of anger or frustration,' 'to relax or relieve tension.' That is not to say, necessarily, that using drugs created these psychological coping needs, although drug use may well contribute to them. More likely it means that youngsters who are having trouble coping, or are dysphoric for whatever reason, receive more relief and, therefore, reinforcement from using the drug. These most likely are the individuals at greatest risk of progressing to abuse or even dependence. Needless to say, the drugs usually offer a false promise of helping them cope with the vicissitudes of life.

7. *Stages of Involvement*

The research literature on substance use has distinguished a number of stages of involvement with a drug (Clayton 1992). In general, the stages of use of a drug can be described as abstention, initiation, continuation, progression, dependence, and relapse. In addition, cessation of use can occur after initiation, progression, dependence, or relapse. Psychiatric researchers have been most concerned with defining, understanding, and intervening in the later stages of this continuum.

Explaining progression in the earlier stages of the continuum of involvement, and trying to intervene in that progression, has fallen largely to the social sciences. Though many social scientists who work in this field are informally called epidemiologists or etiologists, in fact they have tended to come mostly from psychology (including social, clinical, counseling, and community psychology), sociology, social work, anthropology, and, more recently, public health and economics. Two very general fields have evolved here—first, the study of the natural phenomena and second, the development of interventions (e.g., education, prevention, persuasive communication). Because prevention is the subject of another article (see *Substance Abuse in Adolescents, Prevention of*), it is mentioned here only to note that the scientists and program developers in prevention tend to have quite similar disciplinary backgrounds and theoretical underpinnings as those who work on the etiology and epidemiology of substance use. Social learning theory has been widely adopted in all of these subareas, for instance.

The notion of a continuum of involvement applies not only to individual drugs, but also across drugs. The use of certain substances tends to precede the use of other ones with a high degree of regularity (Kandel 1975, Kandel et al. 1992). Marijuana use, for example, is usually preceded by the initiation of cigarette smoking, and the two behaviors remain highly correlated at later ages. Alcohol use also tends to precede the use of marijuana; and marijuana use, in turn, usually precedes use of most of the so-called 'hard' drugs like stimulants, opiates, and hallucinogens. The regularity of this pattern of progressive involvement has led to work aimed at predicting specific details of these transitions.

8. *Attitudes and Beliefs about Drugs*

8.1 *Social Connotations of Use*

One of the many influences on substance use is the symbolic meaning the individual sees it as having within society at large or within various social circles. It is obvious that cigarettes and alcohol have acquired certain associations in peoples' minds as a result of intentional manipulation through advertising. Less obvious perhaps is the fact that even unadvertised substances acquire social connotations, and that these may change considerably over time. The high degree of association of marijuana and LSD came to have with the counterculture already has been mentioned. When cocaine use first started to grow in the late seventies, it was very expensive and thus was used disproportionately by the rich and sometimes famous. As a result it acquired a certain caché that made its use more desirable to many, whereas a number of other drugs, like heroin, were more associated with marginal groups and the less affluent. It seems likely that adolescents come to see inhalants as 'kids' drug' as they get older, and that their use declines as a result.

8.2 *Perceived Risk*

Whether drugs are seen as dangerous to the user or to others can likewise affect how they are seen more generally. The dramatically lowered social acceptability of tobacco over the past two decades illustrates the point. As smoking came to be seen as carrying high

risk of morbidity and mortality, it became a less popular behavior, primarily with adults. And as the evidence grew of the adverse effects on others from their passive exposure to tobacco smoke, smoking in many public places, once an assured right, became a frequently denied privilege. Social norms, as well as regulations and laws, shifted quite dramatically. A not dissimilar thing happened with alcohol, when the many adverse effects of drinking came to be more widely recognized and, in particular, when the consequences of drunk driving were emphasized in the eighties. In that decade drinking and driving declined substantially among adolescents in the United States, and even occasions of heavy drinking declined considerably (O'Malley and Johnston 1999).

A number of the illicit drugs have also had considerable transformations in their social acceptability, as their adverse consequences were documented and made widely known. In the late 1970s, as cocaine use became more widespread, there was little convincing evidence of adverse effects, as 'experts' proclaimed that cocaine was neither lethal nor addictive. The evidence eventually cumulated to prove them wrong, but it took some years because there turned out to be a long lag-time on average between the initiation of use and the development of dependency and other serious problems. Researchers at the University of Michigan have demonstrated the importance of perceived risk as a determinant of trends not only in cocaine use, but also in the use of marijuana and other drugs (Bachman et al. 1998, Johnston et al. 2000). Among other things they have shown that a turnaround in the aggregate-level perceived risk associated with a drug by adolescents often presages a turnaround in actual use a year later.

8.3 Disapproval of Drug Use

As illustrated in the cigarette example above, perceived risk can influence the social acceptability of using a drug, though it is only one such determinant of social acceptance. At both the aggregate and the individual level, personal disapproval of using a drug also has been shown by the Michigan researchers to be an important determinant of use. It is not uncommon for disapproval to show a turnaround a year later than perceived risk, consistent with the notion that risk can influence disapproval.

9. Individual Susceptibility

In addition to attitudes and beliefs about drugs, there are a great many other known individual risk factors for psychoactive drug use (e.g., Clayton et al. 1995, Glantz and Pickens 1992). Some of the strongest among youths and young adults have to do with their degree of attachment to, or integration with, key institutions in society. At the broadest level, deviance from societal norms in general, as reflected by delinquent behavior, is a strong correlate of virtually all forms of illicit drug use (including under-age drinking and smoking). It should not be surprising that one class of illicit behaviors (drug use) correlates strongly with a range of other illicit behaviors. It simply means that people who are willing to violate legal and normative constraints in one domain of behavior are willing to do so in others, as well. One result of this fact is that the correlates and determinants of the various 'problem behaviors' among youth tend to be the same. Among them are attachments to various other institutions, but in particular the family, school, and church (Brook et al. 1990).

Young people who have a low degree of attachment or 'bonding' to school—meaning that they are not performing well, do not like school, and may frequently be absent or truant—are at particular risk of involvement with illicit drug use. The same may be said of those who are frequently out of the parental home in the evening and/or who are having difficulties with their family, and of young people who are not very religious (regardless of religion or denomination) as reflected in low levels of religious attendance and rated importance of religion in their lives (Petraitis et al. 1995).

10. Future Directions

It has become increasingly clear that both licit and illicit substance use must be viewed in terms of age, period, and cohort effects, since all three different types of change have been observed and documented (Johnson and Gerstein 2000, Johnston et al. 2000). The factors explaining each of these three types of change are likely to be different, as is the timing of their action. The continual emergence of new illicit drugs onto the scene will yield new drug-taking phenomena to be understood and influenced. Undoubtedly more will be learned about the neurological action and effects of the many drugs, which may well change the importance of certain psychosocial factors, such as perceived risk, in determining their use.

The field is extremely multifaceted as has been illustrated here. The number of drugs is large and growing, and they have a wide range of psychological and physical effects. There are many levels of social and psychological determinants of this group of behaviors, ranging from legal, moral, normative, and religious, to neurological, psychological, sociological, political, and economic. (Because drugs are commodities, market considerations also influence use.) Further, drug use is the object of many social interventions, from various forms of education and persuasion to treatment. Thus, despite the considerable resources currently going into research on the determinants of drug use, it is clear that there will remain a great deal to be learned about this unique domain of behavior for some time to come.

See also: Drug Addiction; Drug Addiction: Sociological Aspects; Drug Use and Abuse: Cultural Concerns; Drugs and Behavior, Psychiatry of; Drugs, Decriminalization of: Cultural Concerns; Drugs: Illicit Use and Prevention; Substance Abuse in Adolescents, Prevention of

Bibliography

Bachman J G, Johnston L D, O'Malley P M 1998 Explaining the recent increases in students' marijuana use: The impacts of perceived risks and disapproval from 1976 through 1996. *American Journal of Public Health* **88**: 887–92

Brook J, Brook D, Gordon A, Whiteman M, Cohen P 1990 The psychosocial etiology of adolescent drug use: A family interactional approach. *Genetic, Social, and General Psychology Monographs* **11**: 111–267

Clayton R R 1992 Transitions in drug use: Risk and protective factors. In: Glantz M, Pickens R (eds.) *Vulnerability to Drug Abuse*. American Psychological Association, Washington, DC, pp. 15–52

Clayton R R, Leukefeld C G, Donohew L, Bardo M 1995 Risk and protective factors: A brief review. *Drugs & Society* **8**: 7–14

Glantz M, Pickens R (eds.) 1992 *Vulnerability to Drug Abuse*. American Psychological Association, Washington, DC

Hibell B, Andersson B, Ahlström S, Balakireva O, Bjarnasson, T, Kokkevi A, Morgan M 2000 *The 1999 ESPAD Report Alcohol and Other Drug Use Among Students in 30 European Countries*. The Swedish Council for Information on Alcohol and Other Drugs, Stockholm, Sweden

Inciardi J A, McBride D C 1991 The case against legalization. In: Inciardi J A (ed.) *The Drug Legalization Debate*. Sage, Newbury Park, CA, pp. 45–79

Johnson R A, Gerstein D R 2000 Age, period, and cohort effects in marijuana and alcohol incidence: United States females and males, 1961–1990. *Substance Use & Misuse* **35**: 925–48

Johnston L D, O'Malley P M 1986 Why do the nation's students use drugs and alcohol?: Self-reported reasons from nine national surveys. *Journal of Drug Issues* **16**(1): 29–66

Johnston L D, O'Malley P M, Bachman J G 2000 *Monitoring the Future National Survey Results on Drug Use, 1975–1999: Vol. I; Secondary School Students* and *Vol. II; College Students and Adults Aged 19–40*. National Institute on Drug Abuse, Bethesda, MD

Kandel D B 1975 Stages in adolescent involvement in drug use. *Science* **190**: 912–14

Kandel D B, Yamaguchi K, Chen K 1992 Stages of progression in drug involvement from adolescence to adulthood: Further evidence for the gateway theory. *Journal of Studies on Alcohol* **53**: 447–57

Koob G F, Nestler E 1997 The neurobiology of drug addiction. *Journal of Neuropsychiatry and Clinical Neurosciences* **9**: 482–97

Leshner A I 1999 Science-based views of drug addiction and its treatment. *Journal of the American Medical Association* **282**: 1314–6

O'Malley P M, Johnston L D 1999 Drinking and driving among US high school seniors. *American Journal of Public Health* **89**: 678–84

Petraitis J, Flay B R, Miller T Q 1995 Reviewing theories of adolescent substance use: Organizing pieces in the puzzle. *Psychological Bulletin* **117**: 67–86

Substance Abuse and Mental Health Service Administration 2000 *Summary of Findings from the 1999 National Household Survey on Drug Abuse*. Substance Abuse and Mental Health Services Administration, Rockville, MD

United Nations 1971 *Convention on Psychotropic Substances 1971* (United Nations publication, Sales No. E.78.XI.3)

United Nations 1972 *Protocol Amending the Single Convention on Narcotic Drugs, 1961* (United Nations publication, Sales No. E.77.XI.3)

Zinburg N E, Robertson J A 1972 *Drugs and the Public*. Simon & Schuster, New York

L. D. Johnston and P. M. O'Malley

Drugs and Behavior, Psychiatry of

At the beginning of the twentieth century, Western culture had few drugs that were useful to treat central nervous system (CNS) disorders, including mental illness. Ethyl alcohol, used alone or as a component of numerous patent medicines, had use, perhaps as a sedative. Other sedatives, such as chloral hydrate, were available as sleep-inducing agents. More potent sedatives such as nitrous oxide and diethyl ether were used as general anesthetics. Morphine had been isolated from crude opium and was used as an analgesic. Cocaine was used as a local anesthetic. Virtually all of these drugs had abuse potential. By the 1920s, drug abuse had become a significant problem, and this problem continues into the twenty-first century.

This situation continued until the 1950s, at which time a revolution in psychopharmacology began, a revolution that continues into the twenty-first century. Thus, the field of psychopharmacology is really only about 50 years old! In the 1950s and 1960s, significant advances (Julien 2000) included:

(a) Development of the *barbiturates* as sedative-hypnotic drugs. Recognition of their toxicity and abuse potential led to identification and development of the *benzodiazepines* as safer agents for the treatment of anxiety states and insomnia.

(b) Development of the *phenothiazines* for the treatment of schizophrenia. Recognition of the therapeutic limitations and toxicities of these drugs led to development of the *atypical antipsychotics* in the 1980s and 1990s.

(c) Development of several classes of *antidepressant drugs* effective in the treatment of major depression and dysthymia. These included the monoamine oxidase inhibitors and the tricyclic antidepressants. This was followed in the 1980s and 1990s by development of less toxic antidepressants such as the serotonin-specific reuptake inhibitors and 'dual action' antidepressants. As the twenty-first century begins, additional classes of safer and even less toxic antidepressants are being introduced. These include a specific norepinephrine reuptake inhibitor, COMT-inhibitors, a selective, reversible MAO-inhibitor, a

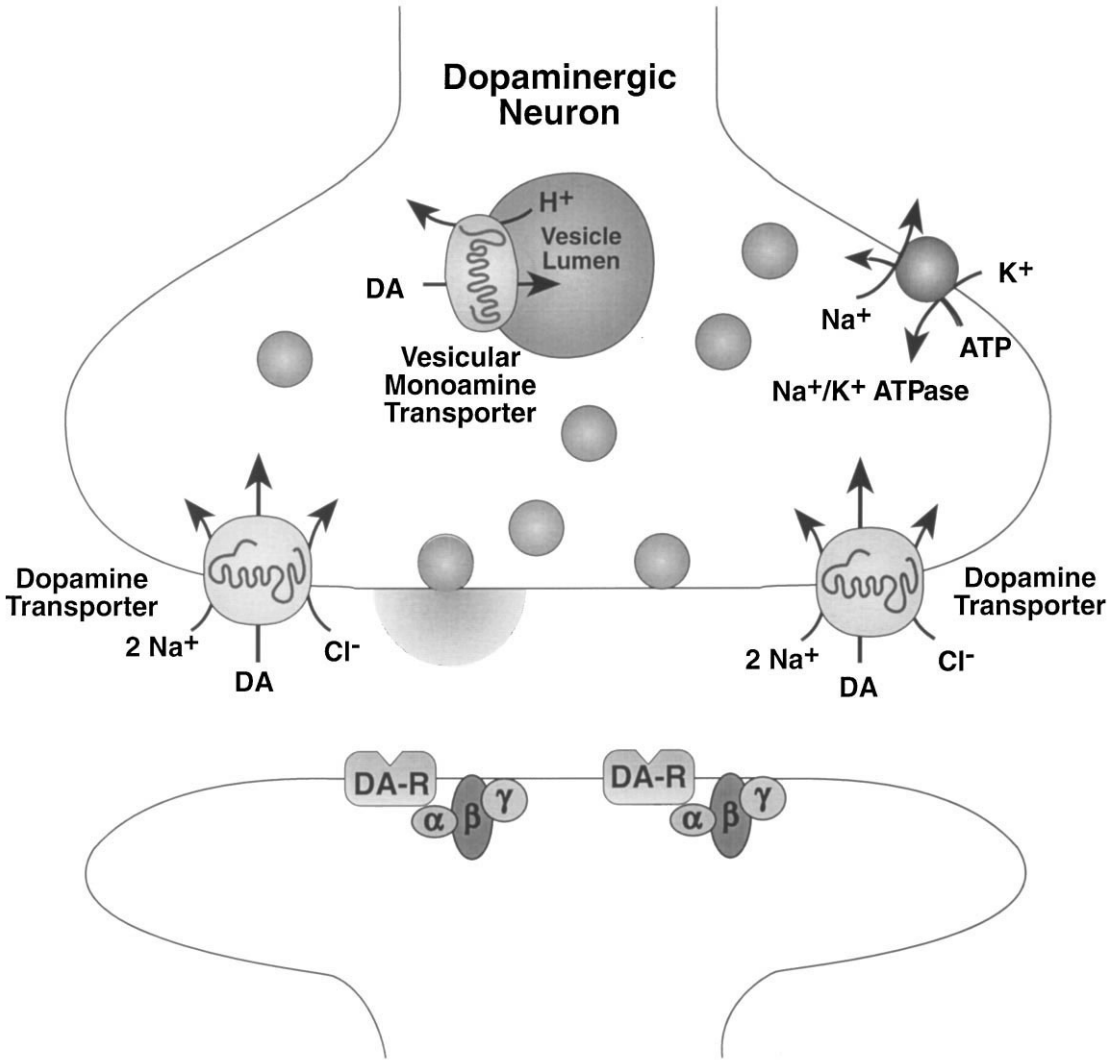

Figure 1
Transporter proteins involved in the active uptake of dopamine (DA). Two transporters are shown. The first is a 'vesicular DA transporter' located in the cytoplasm of the presynaptic neuron, bound to DA-containing storage vesicles. This transporter carries DA from the cytoplasm into storage. The second type of DA transporter is found on the synaptic membrane of the presynaptic neuron and functions to transport DA from the synaptic cleft into the presynaptic nerve terminal, 'recycling' the transmitter and ending the process of synaptic transmission. It is this second transporter that is blocked by cocaine, prolonging the action of DA in the synaptic cleft. Amphetamines function to induce the release of increased amounts of DA from the storage vesicles into the synaptic cleft. The end result is the same: increased amounts of DA at the postsynaptic receptor (DA-R)

neurokinin-1 receptor antagonist, and 'natural' products such as St. John's wort and DHEA (an adrenal androgen).

(d) Identification of the therapeutic efficacy of *lithium* in the treatment of bipolar illness. In the 1990s, safer and less toxic alternatives were developed. These included at least 6 different *antiepileptic drugs*, atypical antipsychotic drugs, and 'natural' products such as omega-3 fatty acids.

(e) Identification of the usefulness of certain of the antidepressant drugs and atypical antipsychotics in the treatment of specific anxiety disorders (e.g., post-

traumatic stress disorder (PTSD), phobias, panic disorder, and obsessive-compulsive disorder), and behavioral disorders with associated symptoms of agitation, aggression, and violence.

(f) Development of *psychostimulant agents* effective in the treatment of childhood, adolescent, and adult attention-deficit/hyperactivity disorder (ADHD), Alzheimer's disease, narcolepsy, and other disorders of attention, thought, and memory.

(g) Development of drugs useful in the treatment of Parkinson's disease. Here we include dopamine replacement agents, dopamine receptor agonists, and a variety of other drugs that have improved the lives of individuals suffering from this disorder.

(h) Development of new anesthetic agents, allowing for the provision of safe anesthesia in patients in any state of health, debility, or surgical requirement. Such agents include ketamine, etomidate, non-toxic inhalation agents, and ultrashort-acting narcotics and muscle relaxants.

Thus, major psychiatric illnesses and other brain disorders have become, for the first time in human history, amenable to pharmacologic treatment; and the search for more effective, more tolerable, and safer drugs continues. As of 2001 at least 102 new medications are in development for treating mental illness. This includes 23 drugs for dementias, 19 for substance abuse disorders, 18 for depression, 15 for schizophrenia, 16 for anxiety disorders, 5 for eating disorders, 4 for ADHD, and 2 for treating PTSD.

It is important to note, however, that psychopharmacology involves much more than the development and testing of new therapeutic agents. Psychopharmacology has also provided powerful tools to study and characterize neurochemical pathways in the brain, and how these pathways may be involved in the pathophysiology and treatment of psychiatric illness (Duman 1999). This work has focused largely on neurotransmitter systems, including the synthesis, release, and metabolism of neurotransmitters and also of the receptor subtypes that control both presynaptic release and postsynaptic actions of neurotransmitters. All of this is to enable science to better understand brain physiology as well as to expand our knowledge of therapeutic targets.

In recent years, the fields of molecular pharmacology and pharmacogenetics have contributed mightily to neuroscience. Many of the enzyme and receptor targets for psychotherapeutic drugs have been identified, sequenced, cloned, three-dimensional structures proposed, and shown to be part of larger families of receptors with multiple homologous members (Julien 2000, Tallman 1999). Both therapeutic actions and side effects of drugs can often be explained by the interactions with multiple different types of receptors. For example, some antidepressants bind to and block the presynaptic transporters for serotonin and norepinephrine as well as postsynaptic receptors for histamine and acetylcholine. Other classes of antidepressants are quite specific for blocking either a single presynaptic transporter or a specific postsynaptic receptor (Figure 1; from Amara and Sonders 1998). This allows one to tailor a specific drug for a specific patient.

With the identification and cloning of different receptor subtypes, whole new generations of drugs specific for each subtype will allow for new advances in the treatment of mental disorders. For example, there are over 15 subtypes of serotonin receptors all of which are potential therapeutic targets. Tallman (1999) states that 'taken together, over 1,000 G-Protein Couples Receptors, over 100 ligand-gated and other ion channel subunits, over 20 reuptake proteins, almost 50 cytokine, 25 nuclear receptor, and several thousand other molecular targets have been cloned, expressed in mammalian, yeast, or insect cells, and are currently available to the pharmaceutical drug discoverer as potential therapeutic targets for agonist or antagonist discovery.'

Classically, a drug is often characterized by a fairly rapid-onset action in a specific model (i.e., it blocks a specific receptor). However, the therapeutic action may take days or weeks of therapy to become evident. This difference in time course has led to the hypothesis that the therapeutic action of many drugs is dependent on adaptations to the acute drug action (Duman 1999). For example, antidepressants acutely block a reuptake transporter, but the clinical relief of depression may take 6 weeks to become evident. Identification of the relevant adaptations, which can be thought of as a form of drug-induced neural plasticity, could occur at several cellular levels, including receptor and second-messenger coupling, modification of intracellular signaling pathways, alterations in nuclear transcription of receptor protein production, or 'remodeling' of intracellular architecture by regeneration, reconfiguration, or modification of intracellular 'neurotropic factors' (Duman 1999, Russo-Neustadt et al. 1999, Brown et al. 1999). Figure 2 (Duman et al. 1997) illustrates a model of such an interaction between acute drug effects and chronic 'plastic' neuroadaptations. The complexity of these adaptations is at least comparable to that of the complex neuronal circuitry in the brain and in the etiology of psychiatric illness. In addition, drug action interacts with genetic and environmental influences to modulate these neuroadaptive influences. For example, a 'Brain-Derived Neurotropic Factor' (BDRF) is poorly expressed in clinical depression and in stressful situations characterized by elevated levels of glucocorticoids (Duman 1999). The combination of exercise and an antidepressant (either a tricyclic-type or a monoamine oxidase inhibitor) exerts a positive and additive effect on BDRF messenger RNA expression in rat hippocampus over a period of 20 days (Russo-Neustadt et al. 1999). This is consistent with enhanced BDRF mRNA expression after antidepressant treatment and increased expression of the tran-

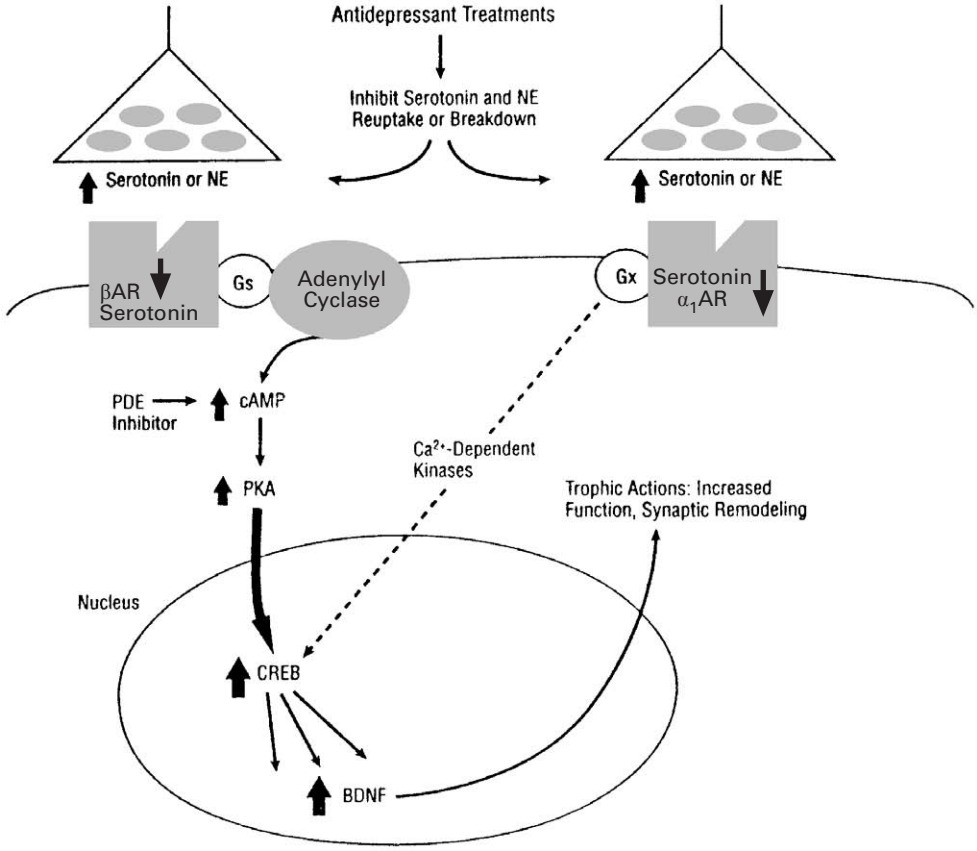

Figure 2
A model for the molecular mechanism of action of long-term antidepressant treatments. Antidepressants induce short-term increases in 5-HT and NE. Longer-term use decreases the function and expression of their receptors, but the cAMP sinal transduction pathway is increased, including increased levels of *adenyl cyclase* and cAMP-dependent protein kinase (PKA), as well as translocation of PKA to the cell nucleus. Antidepressants increase expression and function of the transcription factor cAMP response element-binding protein (CREB), suggesting that CREB is a common postreceptor target for antidepressants. Brain-derived neurotrophic factor is also increased by antidepressant treatment; upregulation of CREB and BDNF could influence the function of hippocampal neurons or neurons innervating this brain region, increasing neuronal survival, function, and remodeling of synaptic or cellular architecture

scription factor cAMP response element binding protein (CREB) in rat hippocampus (Duman 1999, Nibuya et al. 1996, Tao et al. 1998). One implication of this is that it explains the additive effects of medication and behavioral therapies in the treatment of depression. It may also encourage the development of better behavioral treatments aimed at reducing depression and anxiety, and enhanced psychological capacity for dealing with chronic stress (Duman 1999).

Here, we have been discussing the interface between the molecular actions of drugs and their clinical utility in the treatment of disorders of the nervous system. This interface lies within the purview of neuropsychopharmacology; a unique, integrated area of science and medicine that is ideally positioned to ask fundamental questions about drugs, their actions on minute subunits of the nervous system, and how these drugs may be used for therapeutic benefit in human beings (Walker 1999). While molecular biology holds out the promise of giving a complete understanding to all aspects of biology, pharmacology is charged with understanding the actions of a drug at all levels of biology, from its effects on the genetic code or biomolecule to large populations of human persons. Indeed, it is equally important for the pharmacologist to determine the interaction of a drug with its 'receptor,' with the transduction mechanisms that result and lead to cellular action, to integration at the

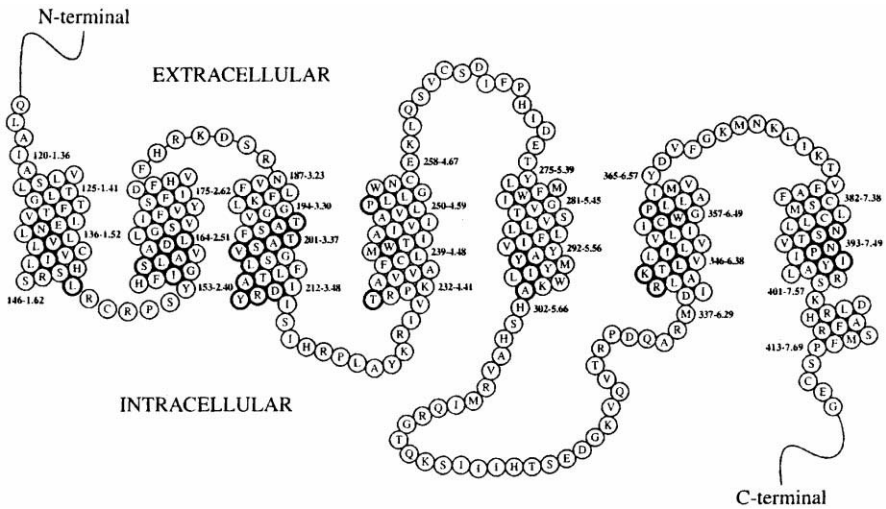

Figure 3
Two-dimensional representation of the anandamide receptor, which is a protein consisting of a chain of more than 450 amino acids (the first letter of each individual amino acid is shown). Like the opioid receptors, the anandamide receptor is a G-protein linked, seven-membrane spanning structure with three extracellular loops

organ or system level, and to the effects on the whole person. Thus, pharmacology offers a complete picture in which a drug is 'mapped' all the way from its initial action at a bio-macromolecule through to effects on a human population with all of the social, economic, and geographic constraints of the latter (Walker 1999). In essence, pharmacologists will translate fundamental genetics and structural biology into drugs with therapeutic value (Burks 1999). To illustrate this characterization of a psychoactive drug, we consider *tetrahydrocannabinol*, the active ingredient in marijuana.

In 1964, delta-9-tetracydrocannabinol (THC) was isolated from marijuana and identified as its pharmacologically active ingredient. About 10 years later, it was hypothesized that THC and other cannabinoids act via a distinct set of receptors. In 1986, THC was found to inhibit the activity of the intracellular enzyme *adenylate cyclase*, and such inhibition required the presence of a newly identified G-protein complex. In 1990, that receptor was isolated, purified, the amino acid sequence determined, and the receptor cloned. The receptor was found to be a specific G-protein-coupled receptor that inhibits adenylate cyclase, binds cannabinoids, inhibits calcium ion flux and facilitates potassium channels (Childers and Breivogel 1998, Felder and Glass 1998). This 'cannabinoid receptor' is a chain of 473 amino acids with seven hydrophobic domains that extend through the cell membrane (Fig. 3). When THC binds, it activates G-proteins that act upon various effectors including the second-messenger enzyme adenylate cyclase and both potassium and calcium ion channels (Fig. 4; from Childers and Breivogel 1998).

The identification of a naturally occurring 'ligand' that might function as a 'natural THC' remained to be demonstrated. In 1992 the arachidonic acid derivative *anandamide* was identified. This substance not only bound to the cannabinoid receptor, but produced cannabinoid-like behavioral, hypothermic, and analgesic effects that parallel those caused by THC. Indeed, anandamide exhibits the essential criteria required for its classification as the *endogenous ligand* at cannabinoid receptors.

In 1999, it was demonstrated that anandamide functions as a 'partial agonist' at hippocampal glutamate-releasing neurons to 'reduce, but not totally block, excitatory transmission' (Shen and Thayer 1999). This may explain not only an antiepileptic action of THC, but THC's 'neuroprotective antioxidant' action as well (Hampson et al. 1998).

A cannabinoid *antagonist* has been synthesized; and mice lacking the cannabinoid receptor were bred and a lack of response to cannabinoid drugs demonstrated (Ledent et al. 1999). These same anandamide antagonists have been used to study dependence on THC; and, in 1999, a 'marijuana abstinence syndrome' was described (Haney et al. 1999a, 1999b).

Anandamide is synthesized within neurons by a condensation reaction between arachidonic acid and ethanolamine, under the regulation of calcium ions and cyclic adenosine monophosphate enzyme. It is broken down by a process of hydrolysis after carrier-mediated neuronal and astrocyte reuptake by an

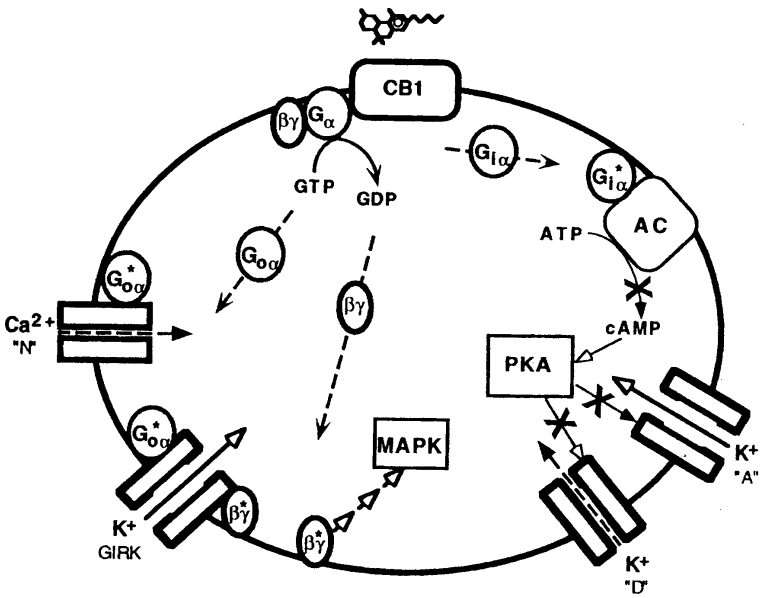

Figure 4
Several proposed signal transduction mechanisms of cannabinoid receptors. After THC (or anandamide) binds (top), the receptor activates G proteins, which in turn act upon various effectors including: adenylate cyclase (AC), calcium (Ca^{2+}), and 'mitogen-activated protein kinase' (MAPC). Inhibition of adenylate cyclase and subsequent decreases in cAMP decreases activation of cAMP-dependent protein kinase (PKA), which leads to decreased potassium (K^+) channel fluxes. Stimulatory effects are shown by open arrows and inhibitory effects by filled arrows. The 'open' or 'closed' states of the channels and 'Xs' over some arrows reflect the final effects of cannabinoid agonists

amidase for rapid hydrolysis (Piomelli et al. 1999). Anandamide, THC, and other anandamide-receptor agonists inhibit the presynaptic release of the excitatory neurotransmitter glutamate in the hippocampus (Shen et al. 1996). Thus, the hippocampus, cerebral cortex, cerebellum, and basal ganglia are major loci of action of THC because these structures are involved in cognition, learning, memory, mood, and other higher intellectual functions, as well as motor functions, all of which are affected by THC.

In the early 1990s, the unique pattern of localization of cannabinoid receptors in the brain was delineated (Herkenham 1992), with large numbers of receptors found in the basal ganglia, cerebellum, and cerebral cortex. The latter receptors for THC mediate the psychoactive effects of the drug. Cannabinoid receptors are also dense in the hippocampus; accounting for THC-induced disruption of memory, memory storage, and coding of sensory input. Brain stem structures do not bind cannabinoids; they do not depress respiration; and they are relatively nonlethal. Perhaps we can now rename the cannabinoid receptor as the anandamide receptor, recognizing the endogenous neurotransmitter for the receptor just as we define other transmitter-activated receptors.

How does this knowledge of THC pharmacology translate into clinical useful information? One of the major effects of marijuana in humans is disruption of memory, presumably due to the reduction in hippocampal activity. Such memory impairments can be attenuated by cannabinoid antagonists, implying that THC- and anandamide-induced memory disruption is mediated by cannabinoid receptors and not by any indirect sedative mechanism (Mallet and Beninger 1998). THC impairs cognitive functions, perception, reaction time, learning, and memory. These impairments have obvious implications for psychomotor functioning, for example in the operation of a motor vehicle (Kurtzthaler et al. 1999).

THC and anandamide are both analgesic (at both spinal and brain stem levels). Cannabinoids produced analgesia, potentiating the analgesic action of morphine, increasing morphine's potency; this action is blocked by cannabinoid antagonists, indicating again that THC is again acting through a cannabinoid receptor mechanism (Smith et al. 1998). In primates, THC decreases aggression, decreases the ability to perform complex behavioral tasks, seems to induce hallucinations, and appears to cause temporal distortions. THC causes monkeys to increase the frequency

of their social interactions. Finally, THC can disrupt appetite regulation, inducing overeating in rats exposed to anandamide; providing evidence for the involvement of a central cannabinoid system in the normal control of eating.

Currently, there is one approved and several possible therapeutic uses for THC and its various derivatives. *Dronabinol* (Marinol), which is synthetic THC formulated in sesame oil, has been available for many years for use as an appetite stimulant in patients with AIDS and for use in the treatment of nausea and vomiting associated with chemotherapy in cancer patients. Other potential uses of dronabinol are to reduce the muscle spasms and pain in multiple sclerosis and reduce the intraocular pressure in glaucoma. Antidepressant and analgesic effects are also claimed, with analgesic actions best described in animal studies (discussed above).

The recent passage of initiatives legalizing the use of marijuana for medicinal uses has reopened this issue of therapeutics. Does marijuana have true therapeutic efficacy, equal to or greater than those of existing agents? The answer is muddied both by smoking as a route of administration and by the use of a crude product containing THC, which is only a weak agonist at cannabinoid receptors. Both of these facts limit the therapeutic use of smoked plant material as a therapeutic choice. The advent of highly potent synthetic analogues (pure anandamide agonists) and specific antagonists may make possible the development of compounds that lack the undesirable side effects of marijuana plant material (Adams and Martin 1996).

A 1999 report (Institute of Medicine 1999) concluded that cannabinoids have potential applicability for some human symptoms. However, the cannabinoids should be delivered by a mechanism other than inhaling smoke; the future may be non-smoked, inhaled molecules, whether they be natural THC from marijuana or totally synthetic compound. This entire area of research, drug development, and social discussion involves the input and opinion from both research and clinical pharmacologists.

See also: Alzheimer's Disease: Antidementive Drugs; Antidepressant Drugs; Drug Addiction; Prophylaxis in Psychiatry; Psychopharmacotherapy, Costs of; Psychopharmacotherapy: Side Effects; Psychotherapy and Pharmacotherapy, Combined; Schizophrenia, Treatment of

Bibliography

Adams I B, Martin B R 1996 Cannabis: Pharmacology and toxicology in animals and humans. *Addiction* **91**: 1585–614

Amara S G, Sonders M S 1998 *Drug and Alcohol Dependence* **51**: 87–96

Brown E S, Rush A J, McEwen B S 1999 Hippocampal remodeling and damage by corticosteroids: Implications for mood disorders. *Neuropsychopharmacology* **21**: 474–84

Burks T F 1999 Pharmacology in the 21st century in the United States of America. *Proceedings of the Western Pharmacology Society* **42**: 171–5

Childers S R, Breivogel C S 1998 Cannabis and endogenous cannabinoid systems. *Drug and Alcohol Dependence* **51**: 173–87

Duman R S 1999 Neuropharmacology in the next millennium: Promise for breakthrough discoveries. *Neuropsychopharmacology* **20**: 97–8

Duman R S, Heninger, Nestler 1997 *Archives of General Psychiatry* **54**: 597–606

Felder C C, Glass M 1998 Cannabinoid receptors and their endogenous agonists. *Annual Review of Pharmacology and Toxicology* **38**: 179–200

Hampson A J, Gromaldi M, Axelrod J, Wink D 1998 Cannabidiol and (−)delta-9-tetrahydrocannabinol are neuroprotective antioxidants. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 8268–73

Haney M et al. 1999a Abstinence symptoms following oral THC administration to humans. *Psychopharmacology* **141**: 385–94

Haney M et al. 1999b Abstinence symptoms following smoked marijuana in humans. *Psychopharmacology* **141**: 395–404

Herkenham M 1992 Cannabinoid receptor localization in brain: Relationship to motor and reward systems. *Annals of the New York Academy of Sciences* **654**: 19–32

Institute of Medicine 1999 *Marijuana and Medicine: Assessing the Science Base*. National Academy Press, Washington, DC

Julien R M 2000 *A Primer of Drug Action*, 9th edn. W H Freeman, New York

Kurzthaler I et al. 1999 Effect of cannabis use on cognitive functions and driving ability. *Journal of Clinical Psychiatry* **60**: 395–9

Ledent C et al. 1999 Unresponsiveness to cannabinoids and reduced addictive effects of opiates in CB_1 receptor knockout mice. *Science* **283**: 401–3

Mallet P E, Beninger R J 1998 The cannabinoid CB1 receptor antagonist SR141716A attenuates the memory impairment produced by delta-9-tetrahydrocannabinol or anandamide. *Psychopharmacology* **140**: 11–19

Nibuya M, Nestler E J, Duman S 1996 Chronic antidepressant administration increases the expression of cAMP response element binding protein (CREB) in rat hippocampus, *Journal of Neuroscience* **16**: 2365–72

Piomelli D et al. 1999 Structural determinants for recognition and translocation by the anandamide transporter. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 5802–7

Russo-Neustadt A, Beard R C, Cotman C W 1999 Exercise, antidepressant medications, and enhanced brain derived neurotropic factor expression. *Neuropsychopharmacology* **21**: 679–82

Shen M, Piser T M, Seybold V S, Thayer S A 1996 Cannabinoid receptor agonists inhibit glutamatergic synaptic transmission in rat hippocampal cultures. *Journal of Neuroscience* **16**: 4322–34

Shen M, Thayer S A 1999 Delta-9-tetrahydrocannabinol acts as a partial agonist to modulate glutamatergic synaptic transmission between rat hippocampal neurons in culture. *Molecular Pharmacology* **55**: 8–13

Smith F L, Cichewicz D, Martin Z L, Welch S P 1998 The enhancement of morphine antinociception in mice by delta-9-tetrahydrocannabinol, *Pharmacology, Biochemistry and Behavior* **60**: 559–66

Tallman J F 1999 Neuropsychopharmacology at the new millennium: New industry directions. *Neuropsychopharmacology* **20**: 99–105

Tao X, Finkbeiner S, Arnold D B, Shaywitz A J, Greenberg M E 1998 Ca^{++} influx regulates BDNF transcription by a CREB family transcription factor-dependent mechanism, *Neuron* **20**: 709–26

Walker M J A 1999 Pharmacology—A unique discipline for unique questions? *Proceedings of the Western Pharmacology Society* **42**: 153–9

R. M. Julien

Drugs, Decriminalization of: Cultural Concerns

1. Introduction

The issue of drug legalization or decriminalization is whether and to what extent the state should use criminal sanctions to regulate drug-related activities such as production, sale, purchase, possession, and use. Analysis of the issue must therefore focus on consideration of the legitimate uses of the criminal law, with special attention to the moral and political rights of citizens to use drugs and to the costs and benefits of criminal prohibition. After considering the definitions of drugs and legalization and the basic limits on the use of the criminal law, this article will discuss the contributions of rights and cost-benefit analyses to the decriminalization debate.

2. Defining Drugs and Decriminalization

Despite common belief to the contrary, there is no uncontroversial definition of 'drug' or 'legalization,' but clarity about both is necessary to understand the legalization debate and to guide public policy.

2.1 Drugs

Virtually all definitions are vague or overinclusive, permitting categorization as a drug of almost any substance that may be consumed. Some definitions are circularly dependent on legal regulations: if a law regulating 'drugs' includes a particular substance within its ambit, it is a drug; if it is not so included, it isn't. Vague, overinclusive, and circular definitions cannot sensibly guide public policy, especially when the state's awesome power to blame and punish for illegitimate use is dependent on the definition.

Definitional problems concerning drugs and legitimacy cannot be entirely avoided by using allegedly scientific or value-neutral medical/therapeutic or nutritional concepts. Concepts like disease and therapy are themselves value-laden and inevitably problematic and controversial. Many substances that have unquestionably legitimate medical or nutritional uses can also be used for nonmedical or non-nutritional purposes that may be of questionable legitimacy. Finally, definitions that define a substance as a drug in terms of its intended use rather than in terms of its inherent properties obscure important questions of legitimate use.

Despite the conceptual problems attending the definition of a drug, a loose but common sense definition is possible for discussing decriminalization, which centrally concerns substances that are consumed recreationally, that is, primarily to produce pleasure, whether or not those substances have other legitimate uses. For the purpose of understanding the current debate, then, recreational drugs can be defined as consumable substances that (a) can affect mood, cognition, and behavior in pleasurable ways; (b) can be used primarily for recreation, including relaxation, excitement, and pleasurable states generally; and (c) can be used so as to endanger the user and others. Although admittedly loose, this definition covers both legal recreational substances such as ethanol (alcohol), nicotine, and caffeine, and substances that are illegal either *per se* or if they are not properly prescribed by physicians such as marijuana, cocaine, opiates, heroin, barbiturates, amphetamines, phencyclidine (PCP), and the like.

2.2 Legalization and Criminalization of Drugs

Legalization (decriminalization) and criminalization are as difficult to define as drug because each includes widely varying legal responses to drug-related activities. The core of decriminalization proposals, however, is the limitation or abolition of present criminal prohibitions of actions related to recreational drug use by competent adults. Some prefer the term legalization to refer to the limitation of criminal prohibitions and penalties. Other, primarily rights-based theorists, believe that state criminal prohibition of recreational drug activities is not *prima facie* justified, so they prefer the apparently more politically neutral term decriminalization. For ease of exposition, the latter term will be used.

The specific details of particular decriminalization proposals depend on the proponent's position on rights and on the costs and benefits of the proposal compared to the current regime of criminal prohibition. Proponents recognize that some criminal law regulation and substantial civil law regulation will probably be necessary to respond to recreational drugs, but all agree that substantially less use of criminal law is desirable and that it should not be the primary mode of regulation.

The core of criminalization proposals is that the criminal law should be the primary means of regulating recreational drugs, but the proper scope of criminalization is disputed and virtually all defenders of criminal prohibitions believe that noncriminal means such as education and treatment should also be used. Virtually no proponent of criminalization argues for expanding the scope of criminalization, say, by prohibiting the recreational use of currently legal substances such as alcoholic beverages or by substantially increasing the criminal penalties for drug offenses. The current debate is about limitation, not expansion, and whether criminal law should be the primary mode of regulation.

3. *The Limits of the Criminal Law: The Harm Principle*

The state's police power to apprehend, prosecute, convict, and punish citizens is one of the most awesome and probably the most afflictive power it exercises. Infliction of such pain requires substantial justification in liberal democracies that seek to protect individual rights and liberty. Afflictive and expensive criminal regulation should be avoided in favor of less intrusive means unless the harm is great, punishment would be deserved for causing it, and criminal prohibition seems necessary to reduce the level of harm and will not unduly impose on other important interests. Criminal law theorists term these limitations on the state's power to criminalize the 'harm principle.' For example, intentionally inflicting emotional cruelty on others is morally despicable, but such behavior is not criminalized because the harm is insufficient to justify criminal punishment. Socialization by families, religious organizations, and schools is reasonably effective to limit such conduct, and regulating such behavior by criminal law might intrude on protected liberties, such as the right to free speech.

There is little debate about the state's justification for prohibiting and punishing traditional crimes against the person, such as homicide and forcible rape, and against property, such as theft and arson that cause grievous harm. Although individual liberties are precious, no citizen has a right to kill, rape, steal, or burn the property of others for private advantage and the criminal sanction seems necessary to curb such behavior. Societies can disagree about the scope of appropriate criminalization. For example, there may be reasonable dispute on how much creation of homicidal risk should be necessary to warrant criminal penalties in addition to civil damages or how severely arson should be punished, but the state's power to criminalize great harm to others is uncontroversial.

The harm principle suggests caution before criminalization, however, when the harm is primarily inflicted on oneself, when the harm is primarily moral and not physical, psychological or economic, or when using criminal law to prevent the harm appears to intrude on important rights or appears unnecessary effectively to reduce the harmful behavior. In liberal societies, the right of the state to criminally prohibit actions that harm primarily the actor (so-called legal paternalism) is more controversial than the right to prohibit harm to others because respect for individual liberty and autonomy generally entails that citizens have greater liberty to harm themselves than to harm others. Such liberty is most arguably extensive and the state's right to regulate criminally is most questionable when the threatened harm to self is apparently solely moral and no other harm to self or to others can be discerned. Most liberal theorists reject the state's power to use criminal law solely for the purpose of enhancing the moral perfection of its citizens—so-called legal moralism. Such goals are not the business of the state and should be achieved by other, less intrusive means.

In political and moral regimes that grant the state more authority to regulate individual liberty, both legal paternalism and moralism might be less controversial. For example, theocracies may coherently use the criminal law more extensively than liberal states to achieve paternalistic and moralistic goals. Even within liberal regimes, the permissible scope of paternalism and moralism may vary, although both require more substantial justification than criminalizing harm to others. Consideration of the state's justifiable use of criminal law to regulate recreational drug use, however, must attend to the appropriate limits of state power within a particular political regime. The remainder of this article assumes that the regime considering decriminalization is liberal.

4. *Recreational Drug Use and Rights*

People often assume that there is no right, no liberty interest to use drugs recreationally and, consequently, that the state has appropriate authority to use criminal law to prevent such use. Such an assumption is seldom supported with cogent argument, however, and it is not easy to give compelling moral and political reasons for rejecting such a right.

Sensible discussion must begin with the recognition that people take drugs recreationally because they desire the pleasurable effects produced by consuming substances that alter mood and cognition. It seems clear, however, that the state has no justifiable interest in criminal prohibition simply because the primary or even sole goal of behavior is pleasure or because the behavior involves consumption of a substance that can alter mood or cognition. Many people might consider it wrong to engage in activities solely for pleasure, especially if alteration of mood and cognition were involved and there were some danger of de-

pendence on the activity, but a liberal society does not interfere with a citizen's right to make autonomous choices to engage in such activities. Citizens surely have a *prima facie* right to seek pleasure for its own sake and it is very difficult to imagine a secular, liberal argument suggesting that such a goal is immoral or harmful *per se*, even if some danger might sometimes be involved.

For example, suppose that a citizen engages in meditation solely for recreational purposes. No persuasive liberal moral or political theory would justify criminal prohibition of meditation, even if some citizens became dependent on meditating. Recreational drug use involves the consumption of a substance, but it is hard to imagine why the source of the recreation alone should make a difference. Currently, illegal drugs produce pleasure by altering mood and cognition, but so do many legal activities, including meditation, and legal drugs such as caffeine, alcohol, and nicotine. The state's right to criminalize alterations of mood and cognition *per se* is questionable because such alterations are not *per se* immoral or harmful.

Citizens appear to have a *prima facie* right to engage in recreational alteration of mood and cognition by drug consumption and the state surely has the burden of justifying criminal prohibition of such recreation by powerful arguments. No right is absolute, however, and the possible right to pursue recreational drug use must yield to a powerful state interest in preventing such use, especially the undoubted state interest in curtailing the threat that drugs will cause substantial harm. Nonetheless, the harm limitation counsels caution before using the criminal law.

Some proponents of decriminalization refer to illegal drug use as a 'crime without a victim' because the person potentially most harmed—the user—in effect consents to the threat. But virtually no action affects only the agent, and recreational drug use sometimes threatens families and communities with economic, psychological, and physical harms. Many of these harms may be paradoxically produced or enhanced by criminal prohibition itself, but many surely are not. Moreover, harms one consents to suffer are nonetheless harms. The question is whether the harms to others or to self that drugs produce are sufficient to trump the individual's *prima facie* right to use drugs and thus to warrant criminalization of drug use.

The next section on cost-benefit analysis considers the harms drugs produce, but it must be remembered that liberal societies often permit dangerous behavior in order to protect the right of citizens to pursue their own visions of how life should be lead. Such dangerous behavior is regulated noncriminally, if at all. For example, knowingly following unhealthy dietary habits can severely harm and even kill, and the harm or death can threaten economic, psychological, and physical harms to others. The enormous costs of unhealthy diets are well known, but it is unthinkable in a liberal society to use the criminal law to regulate individual eating habits. To promote the health of the populace, the state can properly regulate the manufacture and sale of food and can impose a duty on manufacturers to warn consumers of potential health risks associated with particular foods. It can even use the criminal law to police egregious failures to comply with such regulations, but it is unthinkable that it would criminally prohibit unsafe eating habits. For another example, consider the immense costs to many users and to society at large that flow from the consumption of alcohol and nicotine. Again, the state can appropriately regulate many aspects of such substance consumption to reduce the consequent harms, including the use of criminal law for egregiously harmful forms of misuse, such as drunk driving. For many reasons, however, importantly including the right to make one's own choices unencumbered by undue state interference, criminal prohibition of the consumption of alcohol and nicotine by the state is considered unwarranted.

If drugs produce great harms *per se* and no one has a right to use drugs recreationally or if that right is insubstantial at best, then the only question for analysis is whether criminal regulation is justified by the benefits it achieves. But if there is a weighty liberty interest in making choices about how to live one's life, including potentially unwise and dangerous choices about drugs, then the creation of harm is insufficient *per se* to warrant criminalization. Some dangerous behaviors must be permitted to protect liberty and regulated only by civil remedies, including damages when the harm occurs. Any adequate analysis of decriminalization must therefore include consideration of the strength of the liberty interest and it appears that in liberal societies the right is substantial. Empirical analysis of the consequences of various behaviors under various regulatory regimes is necessary to inform decisions about the appropriate scope of criminalization and decriminalization of drugs, but the question of rights cannot be avoided.

5. *Cost-benefit Analysis*

The outcome of a cost-benefit analysis of decriminalization of drugs will inevitably be indeterminate because much of the data are unavailable, unreliable, or fluctuate and because the costs and benefits of a new regime must be speculative. Also, debate too often focuses on worst case scenarios that depend on salient but unrepresentative data. Insufficient information and biased misinformation often suggest the wisdom of maintaining the status quo rather than risk making matters worse by changing public policy. Policies such as the criminal prohibition of drugs that arguably fail and that impose great costs, including the infringement of citizens' rights, should be and are debated and cost-

benefit analysis plays a crucial role in these debates. Unfortunately, historical and cross-cultural comparisons are of extremely limited value because social variables, which vary radically intertemporally within a society and across different societies, immensely affect the consequences of drug use and regulation.

5.1 Harms of Drugs and the Benefits of Criminalization

Most currently illegal drugs would create some potential risks to the user and others, even if they were pure, inexpensive, and legal. For example, misuse of drugs can cause addiction and other extremely serious and well-known health, psychological, and economic effects. People can arguably become addicted to substances other than drugs and to activities such as gambling. Moreover, the concept of addiction is itself problematic. Nonetheless, assume that, for various cultural and biological reasons, most currently illegal drugs have greater addictive or other harmful potential than most other substances and activities.

Much decriminalization debate mistakes the probability and extent of the risks drugs pose. For example, most people who use most drugs recreationally do not become addicted and the moderate recreational use of drugs is not *per se* dangerous if the drug is pure and properly consumed. Nonetheless, the ease of access to drugs following decriminalization would surely increase the prevalence of use, abuse, and addiction and their further harmful consequences. The harms might be especially great in poorer and minority communities, where the prevalence of drug use is no higher than in other communities, but where the effects have been more devastating. Indeed, some responsible observers believe that decriminalization might effectively destroy poorer and minority communities already ravaged by drug abuse and addiction.

The relation between drugs and other criminal activity and health harms is fraught because it is difficult to disentangle the effects of criminalization itself from the independent effects of drug use. Some drugs, especially if used heavily, can substantially impair judgment, facilitate impulsivity, and have other effects that increase the risk of criminal offending and other dangerous behavior such as careless driving or sharing needles for intravenous injection. Increased drug use, especially heavy use, will raise the rates of such undesirable behavior.

Criminalization of drugs surely inhibits use generally by making production, sale, possession, and use more expensive and dangerous and by sending the clearest possible message that drug use is wrong. For example, there may have been both rights-based and consequential reasons to repeal Prohibition in the United States, but the data suggest that consumption and consumption-related diseases such as cirrhosis of the liver both decreased when most alcohol production and sale were illegal. Criminalization also helps prevent drug use by minors. Although minors regrettably have access to and consume drugs in a regime of criminalization, ease of access generally would also increase use and abuse among minors.

The use of criminal law to regulate drug use and related harms is obviously not fully successful by any standard and it is immensely expensive, but the proponents of continued criminalization argue that the current regime provides greater benefits than costs and that the unknown benefits, if any, of decriminalization would pale in comparison to the costs of inevitably increased use and abuse.

5.2 The Costs of Criminalization

Criminal law enforcement, including imprisonment, is an especially intrusive and expensive form of regulation of any behavior. The cost-benefit critique of criminalization argues that such costs are not outweighed by the benefits because criminal law is powerless to make more than a small dent in the use of drugs and because criminalization itself creates many avoidable harms.

The attempt to eradicate drug use by criminal prohibition cannot fully succeed because large numbers of people want recreational drugs for the pleasure or relief they provide and it is widely recognized that the dangers of drugs are often exaggerated. The criminal law's so-called 'war on drugs' in the United States is fighting a fluctuating but always immense number of citizens, most of whom are not addicts or abusers and whose use does not pose grave danger to society or themselves. It is hard to enforce laws that such large numbers of people wish to violate and that have no immediate and direct complaining victim. Furthermore, public hypocrisy about drugs inevitably undermines law enforcement. Western society is not nearly as monolithically antidrug as many claim, as public policy towards alcohol and nicotine indicates.

Given the powerful factors that motivate the desire for drugs, the criminal sanction appears ineffective. The criminal justice system cannot prosecute and imprison more than a tiny fraction of the enormous numbers of people involved in the illegal drug trade unless the justice system massively diverts resources from other, undisputed criminal law needs and abandons civil liberties protections. The criminal courts of the United States are already clogged with drug cases and the ever-expanding prison population includes an enormous proportion of drug offenders. Even spending vastly more money seemingly will provide only limited benefit. Despite billions spent on drug enforcement each year, undoubted law enforcement successes and the huge numbers of people in prison for drug-related offenses, drugs are still freely available and ever-purer and ever-cheaper. Fluctuations in drug usage appear far more related to major social, cultural,

and economic forces than to criminal law enforcement. The limited decrease in use criminal law enforcement produces does not seem to justify the gargantuan costs.

Criminalization also threatens its own effectiveness and creates other problems. For example, criminal prohibitions raise the price on a good or service—the so-called 'crime tariff'—and insure that there will be an endless supply of producers and dealers who seek to realize the monstrous profits to be made through any means necessary, including violence, and that many users will be impoverished and driven to a life of further crime simply to support themselves and to obtain drugs. The immense profits facilitate the growth and power of domestic and international organized crime and can be used to corrupt law enforcement, the criminal justice system in general, and both domestic and foreign politicians. Moreover, the consensual nature of drug offenses insures that law enforcement will unwittingly—and sometimes wittingly—use illegal means to enforce drug laws, thus further corrupting law enforcement and the administration of justice and causing disrespect for both. Criminalization decreases the probability that inevitable drug consumption will occur under conditions safest to health. Finally, primary emphasis on criminal law virtually insures that fewer resources will be devoted to investigating and implementing less intrusive and potentially more effective means to reduce the harms drugs cause.

5.3 *The Benefits of Decriminalization*

Specific decriminalization proposals would produce different benefits, but the potential general benefits of decriminalization are the opposites of the costs of criminalization and can be stated easily in the abstract. They include an increase in personal liberty, probable cost savings, decrease in the crime, violence, and corruption that criminalization produces, increased respect for law enforcement, and increased attention to possibly more effective means such as education, treatment, and civil regulation to reduce use and the harms drugs cause. Criminal law would be used only to prohibit particularly dangerous drug-related activities such as drugged-driving or selling to minors. Such traditional use of the criminal law would receive broad public support.

6. *Conclusion*

Proponents of criminalization sometimes grudgingly concede the problems with the criminal law enforcement approach, but virtually always criticize decriminalization proposals for their abstract or theoretical nature. They fear that decriminalization will produce an administrative, medical, economic, and social tragedy far worse than the current situation, and they ask for highly specific details and guarantees that this will not happen. Proponents of decriminalization recognize that they cannot provide such guarantees, but properly request in return that proponents of criminalization provide the details of how their approach will finally succeed after nearly a century of ever greater expenditures but apparent failure. Neither position appears able to convince the other on theoretical or empirical grounds, but only in recent years has decriminalization been openly debated. Nonetheless, at least in the United States, it appears that the political support to experiment with decriminalization is still weak.

See also: Drug Addiction; Drug Addiction: Sociological Aspects; Drug Use and Abuse: Cultural Concerns; Drug Use and Abuse: Psychosocial Aspects

Bibliography

Adler F, Mueller G O W, Laufer W S 1995 *Criminology*, 2nd edn. McGraw-Hill, New York

Bayer R 1991 Introduction: the great drug policy debate—what means this thing called decriminalization. *The Milbank Quarterly* **69**: 341–64

De Grieff P (ed.) 1999 *Drugs and the Limits of Liberalism: Moral and Legal Issues*. Cornell University Press, Ithaca, NY

Elster J 1999 *Strong Feelings: Emotion, Addiction and Human Behavior*. MIT Press, Cambridge, MA

Falco M 1992 *The Making of a Drug-Free America: Programs That Work*. Times Books, New York

Feinberg J 1986 *Harm to Self*. Oxford University Press, New York

Husak D N 1992 *Drugs and Rights*. Cambridge University Press, Cambridge, UK

Jacobs J B 1990 Imagining drug legalization. *The Public Interest* **101** (Fall): 28–42

Kaplan J 1983 *The Hardest Drug: Heroin and Public Policy*. University of Chicago Press, Chicago

Kleiman M A 1992 *Against Excess: Drug Policy for Results*. Basic Books, New York

Moore M H 1991 Drugs, the criminal law, and the administration of justice. *The Milbank Quarterly* **69**: 529–60

Morse S J 2000 Hooked on hype: addiction and responsibility. *Law and Philosophy* **19**: 3–49

Nadelmann E A 1989 Drug prohibition in the United States: costs, consequences and alternatives. *Science* **245** (1 September): 939–47

Wilson J Q 1990 Against the legalization of drugs. *Commentary* February: 21–8

Zimring F E, Hawkins G 1992 *The Search for Rational Drug Control*. Cambridge University Press, Cambridge, UK

S. J. Morse

Drugs: Illicit Use and Prevention

For the present article, drugs are defined psychoactive chemical agents, i.e., substances that alter cognitive processes. Within this definition two classes of illicit substances emerge. Some legal psychoactive sub-

stances reside in *consumer products* that are widely used, normative approved, and not usually perceived as drugs at all. Examples of these are nicotine in tobacco products, ethanol in alcoholic beverages, and caffeine in coffee or tea. Norms binding consumption of these goods vary in sanctions, but keep the list of permissible users and contexts broad. The next class of legitimate drugs is *medicines*. They can be legally acquired and consumed but only if matching symptoms have been verified by a physician. Drugs belonging to either of the two categories are also abused, i.e., consumed in a problematic or harmful manner. A third group of psychoactive drugs has been declared illicit because of the problematic effects or consequences caused by any of its constituents.

Illicit drugs that are widely used include central nervous system stimulants (such as crack cocaine, cocaine, and amphetamines), central nervous system inhibitors (opiates, heroin and sedative-hypnotics such as benzodiazepines or barbiturates), or hallucinogens (hemp products such as marijuana or hashish, LSD, and phenocyclidine) (AOD 1995). Some goods employed as illicit drugs have legal uses also; examples are home and office products containing solvents or propellants, the fumes of which can be inhaled. Finally, so-called designer drugs are derivatives of illicit substances that technically are not illegal, but produce comparable effects. Experimentation with illicit drugs may lead through various mechanisms involving social, psychological, and pharmacological connections to *dependency* and *abuse* (Des Jarlais and Hubbard 1997).

This article describes ways to combat illicit drug abuse at a national level, including equivalent products, a task that has been characterized as a failure in twentieth century health promotion (Fielding 1999). After the introduction to the situation in the United States and Europe—broadening the perspective is impossible as even authorities have only a vague conception of illicit drug use in many sites (United Nations International Drug Control Programme 1997)—attention turned towards evidence concerning the significance of prevention efforts in communities, schools, and families. Obviously, more is already known than have been put into practice—it is important to disseminate the knowledge—but nevertheless more research is needed on the results of general social policy measures, community and family interventions, and interventions directed to those already marginalized.

1. *Breadth and Risk Factors of Illicit Drug Use/Abuse*

Quantitative information on use exists from industrialized countries including the United Sates (Johnson et al. 1996, NIDA 1997) and most European countries (EMCDDA 1998). Having had at least one occasion of use seems to be a quite variable phenomenon in terms of type of drug, period, region, and ethnic group in the United States. Substantial geographical variability is observed in Europe also, from which a much more limited time series is available.

It is noteworthy that though the countries with statistical information available are usually classified as illicit drug consumers as opposed to producers, this labeling does not describe the situation accurately: drugs are also consumed where they are produced. Development of a global illicit drug use information system is timely.

Statistics on illicit drug use are based on self-reporting. While there are good reasons to have doubts about the accuracy of this information (Hser 1997), many experts (including Preston et al. 1997) have provided evidence about its usefulness. Only surveys are feasible and their results permit reliable comparisons of time points and cultures even though level-of-use estimates based on them are diluted.

United Nations' experts (United Nations International Drug Control Programme 1997) contend that global illicit drug consumption is higher than ever. Surveys indicate accordingly that both experimentation with drugs and their regular use have increased substantially in industrialized countries since World War II. At present illicit drugs involve both genders in the age range from early teens to young adulthood almost equally. In the United States, tenth graders (15–16 years of age; this is the age of the average first user) reported in 1996 the following lifetime use of illicit drugs: marijuana 42 percent, inhalants 18 percent, hallucinogens 11 percent, cocaine 7 percent, crack cocaine 4 percent, heroin 2 percent, stimulants 17 percent, and steroids 2 percent. As a comparison, the daily use of marijuana was admitted by about one in every 10 lifetime users of teen age, and 30 days' consumption of the above-mentioned drugs varied between one third and one sixth of their respective lifetime consumption among US teenagers.

In 1997, at least experimenting with hemp products was admitted by 33 percent of 15–16-year-olds in the UK, 18 percent in Denmark, but only 3 percent in Greece. The corresponding figures for solvents were 20 percent, 12 percent, and 7 percent, and for amphetamines 18 percent, 2 percent, and 4 percent, which represent high, medium, and low ranges of their respective consumption in the 14 European Union countries (EMCDDA 1998). While some calming information exists about the former socialist countries situated in Europe (EMCDDA 1998), this area must be included among those that need a better monitoring system for reliable conclusions (eesv MSDP 1998).

There are numerous risk factors for drug use (Newcomb 1994) involving cultural and structural, interpersonal, psychosocial, or biogenetic factors. Cultural and societal risks include drug use promoting social norms, economic hardships, and drug availability. Among the interpersonal influences are family

use, favorable familial attitudes, family or peer conflict and economic difficulties, and connections to peers using drugs. Psychosocial determinants involve earlier persistent problem behavior, rebelliousness, poor school achievement, early experimentation, and positive drug attitudes. Among the biogenetic causes, genetic susceptibility and psycho-physiological vulnerability to drug effects must be mentioned.

2. *Ways to Control Illicit Drugs*

Illicit drugs are consumer goods subject to economic laws of supply and demand. Increases in world travel, opening of national borders, deregulation of trade and finance, and political instability and conflicts have led into increased supply of these drugs. To reduce the supply, countries have started to network in many ways, including cooperation in such international organizations as the United Nations and World Health Organization. The traditional goal of control politics has been the elimination of illicit drugs from the market. When in fact drug use has increased in spite of efforts to achieve the opposite, some European countries have added harm reduction as a goal for educating youth in general (Uhl and Springer 1998).

International cooperation of law-making and law-enforcing organizations is one of the instruments to control illicit substance use (Van der Stel 1998). Another vital method involves social policy directed to the prevention of marginalization of citizens. Societies form a web of complex social systems in which the role of governmental control varies. Irrespective of their position in that dimension, a few countries have already established health political programs for illicit drug prevention. They, like all health promotion programs (Green 1999), must be implemented in the basic social units, communities.

2.1 *Community Programs to Influence Supply and Demand*

Aguirre-Molina and Gorman (1996) described community programs in drug prevention in the United States, although few of them exist in large numbers (see *Health Interventions: Community-based*). These are comprehensive, empowering, and developmental community-based interventions that target multiple social systems and employ manifold strategies. They aim at all factors in the environment that contribute to the risk or protective factors. Programs already tested and three decades of experience with cardiovascular community programs encourage the continued cooperation of prevention workers and researchers.

Community-based programs usually cannot affect the production and manufacture of drugs (although this may be possible), but they can influence drug availability and access. Second, community organization involves both law enforcement and the local media, and can involve broad-based community change interventions. Furthermore, high-risk groups of people or high-risk neighborhoods need to be focused upon. Third, community-based intervention programs may aim at individuals at different phases of their life cycle.

2.2 *School Programs to Control Demand*

Drug education can be directed to different audiences: in the most general form, it is directed to populations via the mass media in community programs. The next lower level is that of groups—most notably school classes. Finally, education may be executed on a face-to-face basis to individuals. As youth is the subpopulation most at risk and can be addressed in school programs, these have be the main focus of efforts and most of the research (see *Health Promotion in Schools*).

School programs have been largely developed and implemented in the United States (Tobler 1997, Kreft and Brown 1998), but they have also been applied in Europe (White and Pitts 1998, Uhl and Springer 1998). Programs properly tested involve contents such as knowledge (drug effects, social influences, actual use of drugs), emotions (self-esteem, self-awareness, values), refusal skills (drug refusal, public commitment, cognitive behavioral skills, nondrug support systems), more generic or life skills (communication, assertiveness, decision making, coping, social, goal-setting skills), safety skills (own-peer security when drugs are used), extracurricular activities (job, sports, culture, leisure time, community work), or what can be termed prosocial activities (peer counseling, home work, rewards, parent and/or community involvement). Participation of the peer group in the program implementation varies on the dimension of noninteractive vs. interactive. In the list above, the divide is between programs clarifying emotions and teaching refusal skills: programs dealing with knowledge and emotions are noninteractive, the rest are interactive.

Kreft and Brown (1998) present sharp critique directed towards the poorly conducted bulk of illicit drug education studies. Dusenbury and Falco (1997) also note that extra care must be taken when making a purchase among programs available. The review by Tobler (1997), however, enumerates studies defying that criticism and uses sound meta-analytic techniques to yield conclusions, but not, however, verifying whether multilevel modeling important for school studies where students of same class unit have a common environment has been adopted or not. White and Pitts (1998) came to essentially the same results also.

Only interactive school programs of drug education can be deemed effective. Among the interactive programs, the highest effects were observed for general

life-skills programs, followed by programs having either prosocial nonschool activities or social influence as content. Second, the effects when they exist are usually modest in size. As effect size estimates are bound to be subjective, it is helpful to contend that if an effect size of interactive programs were to be observed in a study concerning effects of medicines, a recommendation for use would be made. Third, the effect includes postponing experimentation—often the interactive programs' effects are to be seen for 3 years. Fourth, when programs were implemented in larger groups, the effectiveness attenuated, except among multicultural audiences. This may indicate either poorer administration of larger programs and/or increasing methodological problems posed by larger intervention to internal validity. Fifth, sizeable effects in postponing drug experimentation have been observed in small studies using interactive delivery. Sixth, the effects seem to be equivocal for different drugs.

Educational programs have been targeted at audiences from 8 to 25 years old in a variety of settings including schools and colleges, community settings, the family, medical/therapeutic settings, and mass media (White and Pitts 1998). Tobler and White and Pitts made it clear that most of the available programs are meant for school and for primary and secondary prevention. There is a serious lack of programs to be implemented in other contexts or directed at groups at special risk (White and Pitts 1998).

2.3 Family-based Interventions to Increase Prosocial Bonding and Help with Problems

The community and schools have to be named as important contexts of the above interventions. The possible sites of impact include recreational and religious settings, work, and, importantly, the family, in which the primary socialization takes place. Because of centrality of the family in illicit drug use prevention, this contexts attracts primary interest.

Biglan and Metzler (1998) noted that drug use among young people is associated with engagement in diverse problem behaviors and certain parental practices (lack of effective monitoring, discipline, and positive involvement with children). Three types of family interventions can be discerned: universal to the general population of families, selective to families with high-risk youth, and indicated interventions to dysfunctional families with many risks (Kumpfer 1998). Of these, the most significant from the public health perspective are universal interventions (Catalano et al. 1998).

At the beginning of the twenty-first century, some promising interventions involving the social developmental model, using prosocial bonding family, school, and peers as a protection against the development of conduct problems, school misbehavior, and drug use, are in the evaluation phase. It is hypothesized that strong prosocial bonding to others reduces the risk of substance abuse. It can be expected that the success of this undertaking is dependent on the extent to which prosocial involvement within the family is available to the child, skills necessary in participating in the family interactions, and rewards/punishments provided by parent for conforming/nonconforming with the family expectations and beliefs. As in the case of community research, more studies need to be reported before definite conclusions of their real worth can be arrived at.

3. Conclusion: Towards More Effective Cooperation

Illicit drug abuse prevention continues to be an important task of health promotion, both nationally and internationally. Although all undertakings in prevention may not be easily evaluated (a good example of which is international control measures), scientific rigor should always be aimed at when making the choice. Only school-based programs have been reliably evaluated so far, leading to a recommendation of adopting interactive programs relying on life or refusal skills. Both community and family interventions seem to be important to consider, although they need to be applied and then evaluated further before firm conclusions can be made. At the highest level, national policy goals need to be adapted to changing circumstances in many countries, a global information system should established, and international cooperation at all levels strengthened for the success of prevention and developing its objectives.

See also: Adolescent Health and Health Behaviors; Drug Addiction; Drug Addiction: Sociological Aspects; Drug Use and Abuse: Cultural Concerns; Drug Use and Abuse: Psychosocial Aspects; Drugs and Behavior, Psychiatry of; Drugs, Decriminalization of: Cultural Concerns; Drugs: International Aspects; Health Behavior: Psychosocial Theories; Health Education and Health Promotion; Health Interventions: Community-based; Health Promotion in Schools; HIV Risk Interventions; Smoking Prevention and Smoking Cessation; Substance Abuse in Adolescents, Prevention of

Bibliography

Aguirre-Molina M, Gorman D M 1996 Community-based approaches for the prevention of alcohol, tobacco, and other drug use. *Annual Review of Public Health* **17**: 337–58

A O D 1995 *The Alcohol and Other Drugs Thesaurus: A Guide to Concepts and Terminology in Substance Abuse and Addiction*. NIAA and CSAP, Washington, DC

Biglan A, Metzler C W 1998 A public health perspective for research on family-focused interventions. In: Ashery R S,

Robertson E B, Kumpfer K L (eds.) *Drug Abuse Prevention Through Family Interventions*. US Department of Health and Human Services, National Institutes of Health, Rockville, MD, pp. 430–58
Catalano R F, Kosterman R, Haggerty K, Hawkins J D, Spoth R L 1998 A universal intervention for the prevention of substance abuse. Preparing for the drug-free years. In: Ashery R S, Robertson E B, Kumpfer K L (eds.) *Drug Abuse Prevention through Family Interventions*. US Department of Health and Human Services, National Institutes of Health, Rockville, MD, pp. 130–59
Des Jarlais D C, Hubbard R L 1997 Alcohol and drug abuse. In: Detels R et al. (eds.) *Oxford Textbook of Public Health*. Oxford University Press, London, Vol. 3, pp. 1495–515
Dusenbury L, Falco M, Lake A 1997 A review of the evaluation of 47 drug abuse prevention curricula available nationally. *Journal of School Health* **67**: 1271–32
eesv MSDP 1998 *Phare Project on Drug Information Systems. Final Phase Report on Information Map*. European Commission, Brussels, Belgium
EMCDDA 1998 *Annual Report on the State of the Drugs Problem in the European Union*. European Monitoring Centre for Drugs and Drug Addiction, Lisbon, Portugal
Fielding J E 1999 Public health in the twentieth century: advances and challenges. *Annual Review of Public Health* **20**: xii–xxx
Green L W 1999 Health education's contributions to public health in the twentieth century: a glimpse through health promotion's rear-view mirror. *Annual Review of Public Health* **20**: 67–88
Hser Y-I 1997 Self-reported drug use: results of selected empirical investigations of validity. In: Harrison L, Hughes A (eds.) *The Validity of Self-reported Drug Use: Improving the Accuracy of Survey Estimates*. US Department of Health and Human Services, National Institutes of Health, Rockville, MD, pp. 320–43
Johnson R A, Gerstein D R, Ghadialy R, Choy W, Gfroerer J 1996 *Trends in the Incidence of Drug Use in the United States, 1919-1992*. US Department of Health and Human Services, Public Health Service, National Institutes of Health, Rockville, MD
Kreft I G G, Brown J H 1998 Zero effects of drug prevention programs: issues and solutions. *Evaluation Review* **22**: 3–14
Kumpfer K L 1998 Selective prevention interventions: the strengthening families program. In: Ashery R S, Robertson E B, Kumpfer K L (eds.) *Drug Abuse Prevention Through Family Interventions*. US Department of Health and Human Services, National Institutes of Health, Rockville, MD, pp. 160–207
Newcomb M D 1994 Prevalence of alcohol and other drug use on the job: cause for concern or irrational hysteria? *Journal of Drug Issues* **24**: 403–16
NIDA 1997 Trends in drug use among 8th, 10th, and 12th graders. *NIDA (National Institute on Drug Abuse) Notes 13(2)*. US Department of Health and Human Services, National Institutes of Health, Rockville, MD
Preston K L, Silverman K, Schuster C R, Cone E J 1997 comparison of self-reported drug use with quantitative and qualitative urinalysis for assessment of drug use in treatment studies. In: Harrison L, Hughes A (eds.) *The Validity of Self-reported Drug Use: Improving Accuracy of Survey Estimates*. US Department of Health and Human Services, National Institutes of Health, Rockville, MD, pp. 130–45
Tobler N S 1997 Meta-analysis of adolescent drug prevention programs: results of the 1993 meta-analysis. In: Bukoski W J (ed.) *Meta-analysis of Drug Abuse Prevention Research*. US Department of Health and Human Services, National Institutes of Health, Rockville, MD, pp. 5–68
Uhl A, Springer A (eds.) 1998 *Evaluation Research in Regard to Primary Prevention of Drug Abuse*. Office for Official Publications of the European Communities, Luxembourg (http://www.fed.org.uk/bib/austria.html)
United Nations International Drug Control Programme 1997 *World Drug Report*. Oxford University Press, Oxford, UK
Van der Stel J (ed.) 1998 *Handbook of Prevention. Alcohol, Drugs and Tobacco*. Pompidou Group, Council of Europe and Jellinek Consultancy, Strasbourg, France (http://www.pompidou.coe.int/English/prevenir/manneluk.pdf)
White D, Pitts M 1998 Educating young people about drugs: A systematic review. *Addiction* **93**: 1475–87

A. Uutela

Drugs: International Aspects

Although the term 'drug' has a wider range of meanings, we are concerned here with substances which are psychoactive, that is, which alter the mental state of humans when ingested. Such drugs have been both prized and feared throughout human history. They have long been a substantial item in international trade, and indeed concerning the European colonial systems of 1500–1960 they have been described as the glue of empires. In the twentieth century, they also become the subject of ambitious international control structures. In connection with such control structures, brief consideration is given, too, to performance-enhancing drugs, a category overlapping with psychoactive drugs.

1. *Psychoactive Drugs and How They are Used*

There are a wide variety of psychoactive substances, naturally occurring and synthesized, including tobacco, alcoholic beverages, coffee, tea, chocolate, and some spices, as well as substances legally available only through medical channels such as benzodiazepides, cannabinols (including marijuana), opiates (including heroin), and cocaine. Such substances often have other use values, along with their psychoactive properties (Mäkelä 1983). Users may like the taste, or the image of themselves that the use conveys. Use may be a medium of sociability, or part of a religious ritual. Some substances have other useful properties; alcohol, for example, is a source of calories, and is the solvent in many tinctures and medicines.

Psychoactive drugs differ in their metabolic pathways and mechanisms of action in the human body, in the strength of their effects, and in the states of mind and feelings they induce. In particular, it is easy to become intoxicated with some substances— alcohol is

a prominent example—while it is much less likely with others. But the effects of drug use are also powerfully dependent on the pattern of use, and on set and setting—this is, the expectations of the user and of others present, and the context of use (Zinberg 1984). While the psychoactive effect of tobacco may not even register in the consciousness of a habituated cigarette smoker, in other circumstances the effect of tobacco use may be so strong that the user is rendered unconscious, as early Spanish observers reported concerning native South Americans.

Three social patternings of psychoactive drug use can be distinguished as prototypical: medicinal use, customary regular use, and intermittent use. In many traditional societies, particular drugs or formulations have been confined to medicinal use—that is, to use under the supervision of a healer to alleviate mental or physical illness or distress. For several centuries after the technique for distilling alcoholic spirits had diffused from China through the Arab world to Europe, for instance, spirits-based drinks were regarded primarily as medicines. This way of framing drug use has been routinized in the modern state through a prescription system, with physicians writing the prescriptions and pharmacists filling them. Drugs included in the prescription system are usually forbidden for nonmedicinal use.

Where a drug becomes a regular accompaniment of everyday life, its psychoactivity is often muted and even unnoticed, as is often the case for a habitual cigarette smoker. Likewise, in southern European wine cultures, wine is differentiated from intoxicating 'alcohol'; wine drinkers are expected to maintain the same comportment after drinking as before. We may call this a pattern of 'banalized use': a potentially powerful psychoactive agent is domesticated into a mundane article of daily life, available relatively freely on the consumer market.

Intermittent use—for instance, on sacred occasions, at festivals, or only on weekends—minimizes the build-up of tolerance to the drug. It is in the context of such patterns that the greatest attention is likely to be paid to the drug's psychoactive properties. The drug may be understood by both the user and others as having temporarily taken over control of the user's behavior, and thus to explain otherwise unexpected behavior (Room and Collins 1983). Given the power attributed to the substance, access to it may be limited: in traditional societies by sumptuary rules keyed to social differentiations, in industrial societies by other forms of market restriction.

In industrial societies, a fourth pattern of use is commonly recognized for certain drugs: addicted or dependent use, marked by regular use, often of large doses. Since the pattern of use of the drug in question is not defined in the society as banalized, addiction is defined as an individual failing rather than as a social pattern. While attention is paid to physical factors sustaining regular use, such as use to relieve withdrawal symptoms, most formulations of addiction focus on psychological aspects, including an apparent commitment to drug use to the exclusion of other activities and despite default of major social roles. An addiction concept thus also focuses on loss of normal self-control, but the emphasis is not so much on the immediate effects of the drug as on a repeated or continuing pattern of an apparent inability to control or refrain from use, despite adverse consequences.

Among the attractions of psychoactive substances is that, at least up to a certain point, they potentially enhance one or another kind of performance by the user: For instance, coffee or a cigarette may modestly improve mental performance, a limited amount of alcohol may improve dancing or other expressive activity, amphetamines may improve wakefulness on a long watch. There are also nonpsychoactive substances (steroids, beta-blockers, etc.) which can potentially improve one or another kind of performance. In particular, in the context of competitive sports, international attention has been focused on a wide variety of substances which can improve some aspect of athletic performance.

2. *Problems from Drug Use, and Systems of Control*

From ancient times, it has been recognized that problems can result from drug use. Formerly, the main concerns were with social and political problems seen as associated with drug use. In medieval England, the government's main concern in controlling alcohol sales was with enforcing discipline in the labor force. Violence, disturbances of the peace, and harm to dependent family members would also be commonly recognized, particularly in relation to alcohol consumption. Often it was rebellious or countercultural elements of a society which introduced or became associated with new drugs. Reacting to this, political establishments in Europe tried at first to control coffee houses, and in Turkey and elsewhere imposed severe punishments on tobacco use (Austin 1978).

New social concerns about problems from drug use entered the scene in the modern era. Concerns about the physical health of the drinker were a factor in responses to the English 'gin epidemic' of the eighteenth century. In the early nineteenth century, the idea of addiction became established, at first for alcohol and in the USA (Levine 1978). By the late nineteenth century the idea had spread to other societies, and was applied also to other psychoactive drugs. As safety became identified as a social problem in the late nineteenth century, attention was also directed to the potential effects of drinking in degrading performance of mechanical tasks. In recent decades, as the automobile became a central part of daily life, this concern became much stronger, and was extended too to other drugs.

Social concerns about problems from drug use have resulted in a variety of kinds of efforts to control availability (MacCoun et al. 1996, Room 2000). One option has been a flat prohibition on use, or at least on recreational use. Combined with this, sometimes, has been a system for use of the substance as a medication. The prescription system in modern states is essentially a state-sanctioned form of rationing of availability, with doctors and pharmacists delegated to serve as the gatekeepers on access. A third option has been a system of state control or licensing of sales, such as the systems of alcohol control through state monopolies and licenses in North American states and provinces. Such a structure can be combined with other discouragements for use, such as excise taxes and restricted opening hours. A fourth option, applicable for instance nearly everywhere for tea and coffee, is that the psychoactive substance is treated as an ordinary commodity, with no special controls. For intoxicating substances, choice of the third or fourth option is often accompanied by 'harm reduction' measures (in the terminology of the illicit drug field), since the options involve official acceptance of the fact of recreational drug use.

Where a psychoactive substance is seen as causing great problems, prohibition of use has been a fairly common societal response. Such a prohibition on alcohol had long been imposed on all faithful Muslims by the Koran. Legal prohibition of alcohol, initially on a local basis or for special population categories such as indigenous peoples, became a feature of some Christian societies in the late nineteenth and early twentieth centuries. Prohibition of alcohol sales extended to the national level around the time of World War I in the USA, Canada (briefly), Iceland, Norway, Finland, and Russia. The USA, in particular, had some success in gaining international cooperation with its prohibition through bilateral treaties (Tyrrell 1994). A flourishing worldwide temperance movement succeeded in pushing the European powers with colonies in Africa into adopting treaties prohibiting aspects of the distilled spirits trade in Africa (Bruun et al. 1975). These fell into disuse by the 1940s.

The alternative to prohibition in the early twentieth century was 'state control', and comprehensive systems of alcohol control were eventually adopted, not only in the societies which had tried prohibition, but also in many other nations. These systems have been gradually weakened, particularly since 1950 (Mäkela et al. 1981). On the other hand, in a broad range of societies there has been a steady strengthening of controls over drinking in specific circumstances, in particular over drinking and driving (Gusfield 1981).

In the early twentieth century, the idea of a prohibitory regime was extended from alcohol to other psychoactive substances. In the realm of 'narcotics', which came to refer to opiates, cocaine, and marijuana, the first international treaty was signed in 1906, and extended globally as part of the treaties concluding World War I (McAllister 1992). The agreement on narcotics can be seen as part of an emergent tradition of such regimes for social problems defined as international in scope (Nadelmann 1990). A series of subsequent international agreements were codified in the 1961 Single Convention on Narcotic Drugs (amended 1972). While the system had always included provision for a legal market in narcotic medications under a prescription system, in a 1971 convention its scope was broadened to extend international control to other synthesized medications such as amphetamines, barbiturates, and benzodiazepines. A 1988 convention extended control also to 'precursor chemicals' used in manufacturing controlled drugs. By 1996, more than 240 substances were in principle under the system's control (Room and Paglia 1999).

In the mid-1990s, the international narcotics control system included the United Nations Drug Control Program (UNDCP), with about 200 staff, headquartered in Vienna. The political organ of the system is the Commission on Narcotic Drugs, with an elected membership of 53 nations. The International Narcotics Control Board (INCB) serves as the control board for the legal (pharmaceutical) markets in controlled drugs. Alongside this international effort is a substantial global antidrug effort by the USA, including over 400 Drug Enforcement Agency staff stationed in 50 countries in 1993, and an annual certification process to the US Congress on the extent of cooperation by each other country with antidrug actions.

Although effective action has been taken to diminish use of prescribed drugs like barbiturates, the system has by most accounts been relatively ineffective in suppressing illicit international traffic in drugs. The system's own summation, in the *World Drug Report* (UNDCP 1997), is that the 'outcome is undoubtedly less than satisfactory' in terms of the decreased availability of drugs.

At the end of the 1990s, two other international systems for the control of psychoactive drugs were in formation. In late 1999, in the wake of a World Conference on Doping in Sport, a World Anti-Doping Agency started operation in Lausanne, as a joint venture between governments and the International Olympic Committee (WADA 2000). Prior to this, in 1989 the Council of Europe had put forward an Anti-Doping Convention which by 2000 had been signed by 40 European countries and other countries such as Australia and Canada (COE 2000).

In October 2000, the World Health Organization began negotiations on a global Framework Convention on Tobacco Control. Among topics discussed for inclusion in the Convention were the harmonization of taxes, cooperation against smuggling, elimination of duty-free sales, and control of advertising and sponsorship (Joossens 2000). Unlike the international narcotics control system and the nascent

antidoping system, the tobacco control effort is thus not prohibitory of use, at least by adults.

In the meantime, in 2000 there is no prospect of any international agreement on alcohol control. For tobacco and alcohol, as well as other psychoactive substances not under international control, international treaties have operated in the opposite direction. The World Trade Organization and related treaties, and regional free trade agreements, have been used to weaken national and local alcohol controls (Holder et al. 1998, Jernigan et al. 2000).

3. *Traditions of Social Science Analysis*

Dominant patterns of use of a drug in a society, and shifting conceptions about its use, greatly affect the size and shape of the social science literature concerning that drug. In particular, the very different sociocultural positions of 'illicit drugs', benzodiazepines, alcohol, tobacco, and coffee in most developed societies means that the literatures about these different psychoactive substances are often quite separate.

Novel substances have proved a potent catalyst in many societies in crystallizing free-floating fears about the younger generation and the future. Culturally accepted psychoactive substances, on the other hand, become deeply entrenched in the culture, although often with the potential to again become contested. Social concerns about drugs have thus often ensured substantial funding for social scientists to work in the area. For instance, a substantial part of Polish sociology was financed to work on alcohol problems in the difficult days of martial law there in the 1980s. But while much social science work has been funded and expected to proceed within the parameters of official frames, the work of sociologists and other social scientists has also played a key role in broadening the base of scholarship and understanding.

In the first place, alcohol and drug issues have proved fertile grounds for constructivist analyses by sociologists. Historical and comparative work has identified a number of alternative governing images or models of alcohol and drug problems, and shown that these governing images change over time (Bruun 1971, Gusfield 1967). Constructivist perspectives have often been combined with a positivist concern about measuring and comparing social and health problems associated with drug use, and about distinguishing problems attributable to the drug use from problems that are secondary to societal reactions (e.g., Fischer et al. 1997). Some analyses point to a function for illicit drugs of providing 'suitable enemies' for the modern state, in comparison with the 'unsuitable enemies' of tobacco, alcohol, and pharmaceuticals (Christie and Bruun 1996).

In the second place, there is a long sociological tradition of critical analysis of the concepts of alcoholism or addiction as a disease (Room 1983), concepts which were the dominant discourse in the last half of the twentieth century. The alcohol and drug arenas have thus figured prominently in sociological analyses of the 'medicalization of deviance'— and its demedicalization (Conrad and Schneider 1992). Applying perspectives from Foucault and Rothman, Levine (1978) showed that the addiction concept has a history, first coming to the fore in the early nineteenth-century USA, as an explanation of the mystery of continued drinking by the drunkard, even after pledging to abstinence. Analyses in this tradition point to the centrality to addiction concepts of the focus on self-control (Valverde 1998). The economist's concept of 'rational addiction' has also contributed to the ferment in the late 1990s in thinking about addiction concepts (Elster and Skog 1999).

There are as yet few social science analyses of the burgeoning international systems for control of psychoactive substances. Bruun et al (1975) studied the development of the international drug control system until the early 1970s, but arms-length analysis of the expansion of this system's scope after then has been sparse (Room and Paglia 1999), and analyses of the development of new international instruments and agencies around doping in sports (Houlihan 1999) and around tobacco (Taylor 1996) as yet come primarily from those involved in building the systems. The issues of why the three systems are developing separately, of the nature and practicability of their goals, and of their place in the developing international system, are still important topics for future work.

See also: Drug Addiction: Sociological Aspects; Drugs: Illicit Use and Prevention; Drug Use and Abuse: Psychosocial Aspects; Health Behaviors, Assessment of; Health Policy; Policy Knowledge: Universities; Public Health

Bibliography

Austin G 1978 *Perspectives on the History of Psychoactive Substance Use*. Research Issues Vol. 24. National Institute on Drug Abuse, Washington, DC

Bruun K 1971 Finland: The non-medical approach. In: Kiloh L G, Bell D S (eds.) *29th International Congress on Alcoholism and Drug Dependence*. Butterworths, Australia pp. 549–59

Bruun K, Pan L, Rexed I 1975 *The Gentlemen's Club: International Control of Drugs and Alcohol*. University of Chicago Press, Chicago

Christie N, Bruun K 1996 *Den Gode Fiende: Narkotikapolitikk i Norden, 2nd edn*. Universitetsforlaget, Oslo, Norway

Council of Europe (COE) 2000 Implementation of the Anti-Doping Convention. Council of Europe, Strasbourg. Website: Convention text at: http://culture.coe.fr/Infocentre/txt/eng/econ135.htm

Conrad P, Schneider J 1992 *Deviance and Medicalization: From Badness to Sickness*, 2nd edn. Temple University Press, Philadelphia, PA

Elster J, Skog O-J (eds.) 1999 *Getting Hooked: Rationality and Addiction*. Cambridge University Press, Cambridge, UK

Fischer B, Kendall P, Rehm J, Room R 1997 Charting WHO goals for licit and illicit drugs for the year 2000—are we 'on track'? *Public Health* **111**: 271–5
Gusfield J 1967 Moral passage: The symbolic process in public designations of deviance. *Social Problems* **15**: 175–88
Gusfield J 1981 *The Culture of Public Problems: Drinking, Driving and the Symbolic Order*. University of Chicago Press, Chicago
Holder H D, Kühlhorn E, Nordlund S, Österberg E, Romelsjö A, Ugland T 1998 *European Alcohol Policy and Nordic Integration*. Ashgate, Aldershot, UK
Houlihan B 1999 Anti-doping policy in sport: The politics of international policy coordination. *Public Administration* **77**: 311–34
Jernigan D H, Monteiro M, Room R, Saxena S 2000 Towards a global alcohol policy: Alcohol, public health and the role of WHO. *Bulletin of the World Health Organization* **78**: 491–9
Joossens L 2000 From public health to international law: Possible protocols for inclusion in the Framework Convention on Tobacco Control. *Bulletin of the World Health Organization* **78**: 930–7
Levine H G 1978 The discovery of addiction: Changing concepts of habitual drunkenness in American history. *Journal of Studies on Alcohol* **39**: 143–74
MacCoun R, Reuter P, Schelling T 1996 Assessing alternative drug control regimes. *Journal of Policy Analysis and Management* **15**: 330–52
Mäkelä K 1983 The uses of alcohol and their cultural regulation. *Acta Sociologica* **26**: 21–31
Mäkelä K, Room R, Single E, Sulkunen P, Walsh B et al. 1981 *Alcohol, Society and the State: 1. A Comparative Study of Alcohol Control*. Addiction Research Foundation, Toronto, ON, Canada
McAllister W B 1992 Conflicts of interest in the international drug control system. In: Walker W O (ed.) *Drug Control Policy: Essays in Historical and Comparative Perspective*. Pennsylvania State University Press, University Park, PA, pp. 143–66
Nadelmann E 1990 Global prohibition regimes: The evolution of norms in international society. *International Organization* **44**: 483–525
Room R 1983 Sociological aspects of the disease concept of alcoholism. *Research Advances in Alcohol and Drug Problems* **7**: 47–91
Room R 2000 Control systems for psychoactive substances. In: Ferrence R, Pope M, Room R, Slade J (eds.) *Nicotine and Public Health*. American Public Health Association, Washington, DC, pp. 37–56
Room R, Collins G (eds.) 1983 *Alcohol and Disinhibition: Nature and Meaning of the Link*. NIAAA Research Monograph No. 12. USGPO, DHHS Publication (ADM) 83-126, Washington, DC
Room R, Paglia A 1999 The international drug control system in the post-Cold War era: Managing markets or fighting a war? *Drug and Alcohol Review* **18**: 305–15
Taylor A L 1996 An international regulatory strategy for global tobacco control. *Yale Journal of International Law* **21**: 257–304
Tyrrell I 1994 Prohibition, American cultural expansion and the new hegemony in the 1920s: An interpretation. *Histoire Sociale/Social History* **27**: 413–45
United Nations Drug Control Program (UNDCP) 1997 *World Drug Report*. Oxford University Press, Oxford, UK
Valverde M 1998 *Diseases of the Will: Alcohol and the Dilemmas of Freedom*. Cambridge University Press, Cambridge, UK
World Anti-Doping Agency (WADA) 2000 WADA history. World Anti-Doping Agency, Lausanne. Website: http://www.wada-ama.org/asiakas/003/wada-english.nsf/Home?OpenPage
Zinberg N E 1984 *Drug, Set and Setting: The Basis for Controlled Intoxicant Use*. Yale University Press, New Haven, CT

R. Room

Du Bois, William Edward Burghardt (1868–1963)

William Edward Burghardt Du Bois was born in Great Barrington, Massachusetts on February 23, 1868. Ninety-five years later, having renounced his citizenship of the USA, he died an exile in Accra, the capital of the newly decolonized republic of Ghana on April 29, 1963. An unorthodox and multidimensional career had unfolded between these two events. An unwavering devotion to the cause of human progress underpinned Du Bois' cosmopolitan life that was also characterized by bold shifts in ideological outlook. He moved through pragmatism, elitism, socialism, nationalism, and several versions of Pan Africanism in an idiosyncratic, but always thoughtful sequence that terminated amidst the gloom of the Cold War when the old savant, aged 93 and deprived by his government of the means to travel freely, applied to join the Communist Party of the USA and then fled the land of his birth.

Du Bois had been identified as an outstanding student at an early age. After completing high school in Massachusetts, he continued his studies initially at Fisk University in Nashville. When he entered the college in 1885, it was a haven lodged in a bitterly segregated world still resonating with the destructive energies released by civil war. Life there, confined within what he called the veil of color, was a very different proposition from anything that Du Bois had experienced in the New England environment that had nurtured him. It was there that he discovered and wholeheartedly embraced the black cultures of the American South that he encountered inside the historic college, and discovered in more intense and ambiguous forms while working as a schoolteacher in rural Tennessee during the summer vacations. He concluded his studies at Fisk in 1888, choosing the German Reich's Chancellor Otto Von Bismarck as the topical subject of his oration at the college graduation ceremony.

In fulfillment of a childhood dream, and in recognition of the opportunities it would bring, Du Bois moved on to continue his education at Harvard later the same year. Among other classes, he studied

philosophy with William James, who would become a friend, and read Kant with George Santayana. He took economics under Frank Taussig and history with Albert Bushnell Hart who helped to pilot his career through numerous institutional obstacles. After graduating *cum laude* in Philosophy in 1890, he applied to pursue postgraduate study in Political Science and gained an MA degree in 1891. It was at this point that Du Bois began to seek ways to continue his education in Europe. Financial support was difficult to find but in 1892 he was eventually rewarded with a year-long grant to study in Germany. Later that year, he entered Friedrich-Wilhelm III University in Berlin where his teachers included Gustav Von Schmoller, Adolf Wagner, and Heinrich Von Treitschke. It was also there that he first encountered Max Weber and Georg Simmel.

Du Bois attended several meetings of the Social Democratic Party (SPD) while living in Berlin. He had been working towards a German doctorate in economics but, to his disappointment, was denied the degree on the grounds that he had been registered at the university for an insufficient length of time. After extensive travel in Europe he returned to the corrupting provincialism of 'nigger-hating America' to complete his thesis in Harvard's History department in 1895. Shortly afterwards, he became the first black person to receive the Ph.D. degree from that University, but as a result of this success was hugely overqualified for the limited professional opportunities available to him in a world that could not concede the intellectual capacity of 'the Negro' and consequently viewed him as something of a curiosity.

He eventually found a position at Wilberforce College in Ohio where he taught Classics, Greek, and English. His growing interest in Sociology was manifest in some unsuccessful attempts to have the new discipline added to the curriculum. He married Nina Gomer, who had been a Wilberforce student, in 1896 and, later that year, his first book *The Suppression of the African Slave Trade to the United States of America 1638–1870* (1896) was published to very favorable reviews. This welcome development boosted Du Bois' career prospects and enabled him to move to the University of Pennsylvania, where he was employed to produce a groundbreaking sociological study of Philadelphia's black population. Work on that project would continue after he accepted a Professorship of Economics and History at Atlanta University the following year.

This new post provided an institutional platform from which Du Bois organized an ambitious program of sociological investigations into various aspects of black life in the USA. It resulted in a flow of major publications between 1898 and 1914. His substantial empirical study of Philadelphia was published under the title *The Philadelphia Negro* in 1899; it was given a distinctive character by well-developed interests in economics and comparative history. It is rightly remembered as a landmark publication: the first academic monograph to consider the life of an American black community in depth. The Du Bois household was struck by calamity when Burghardt, their two year-old son, died of diphtheria in May 1899. This tragedy seems to have led to estrangement between Du Bois and his wife. However, it combined with his growing academic and political ambitions to endow an elegiac and sometimes melancholy tone in his next book: *The Souls of Black Folk* published in 1903. This was a remarkable, multivoiced collection of his essays, autobiographical reflections, and fiction. It made him a leader of black Americans.

Du Bois' commitment to resolutely open-ended, liberal, and humanist education which would nourish the needs and responsibilities of a black elite destined for leadership at both local and global levels, brought him into conflict with the more modest and practically-oriented educational proposals that were being advocated by Booker T. Washington and were on show at his Tuskeegee Institute. This national debate, which exposed the molten core of black America's political and economic marginality, resulted in Du Bois leaving Atlanta University to pursue a new para-academic career as a writer, editor, and activist based in the North. In 1910 the newly formed National Association for the Advancement of Colored People (NAACP) hired him as Director of Publications and Research. He was the founding editor of, and a prolific contributor to, the organization's influential and dynamic monthly magazine *The Crisis*, a post he would hold until 1934. His first novel, *The Silver Fleece* was published in 1911, the same year that he attended the Universal Races Congress organized in London by the English Ethical Culture movement.

The death of Booker T. Washington and the advent of the 1914–18 war transformed the political conditions in which Du Bois argued for the equality of black Americans and indicted the catastrophic impact of racism on the American polity stratified by color. Though he never completed a proposed volume on the experience of African-Americans during the war, the immediate aftermath of the conflict provided the impetus for the first of several Pan African congresses, which was held in Paris in 1919. After the war, Du Bois' voice was raised against America's failure to recognize its black citizens and in opposition to the gross violence that greeted the returning soldiers. He was severely critical of the more populist and militaristic forms of Black Nationalism associated with Marcus Garvey's United Negro Improvement Association (UNIA) which was growing in importance as the influence of the NAACP diminished.

Du Bois' cosmopolitan outlook and the breadth of his published work made him a pivotal figure among the younger writers, artists, and thinkers who led the artistic and literary movement known as the Harlem Renaissance. He contributed a memorable essay to *The New Negro* anthology, edited by the philosopher

Alain Locke, that defined this important cultural departure but conflict soon followed with some of the younger writers, who were regarded by Du Bois as insufficiently attentive to their world-historic political responsibilities. *Dark Princess*, a vibrant political novel that concerns a grand planetary alliance of peoples of color against imperialism and colonialism and which was identified by Du Bois as his own favorite among his many books, was published in 1928.

The thirties saw Du Bois renewing his interest in Marxism and developing an original and complex critique of Marx's political theory. This striking synthesis was elaborated in a monumental historical study: *Black Reconstruction In America* (1935). The first of his several autobiographical volumes appeared in 1940. Du Bois rejoined the NAACP as their director of special research in 1944. He was sacked four years later after yet another bout of political disagreements, this time resulting from his intermittent affiliation to communist causes that the utterly respectable and sometimes conservative organization was unable to tolerate. Anti-colonialist and anti-capitalist commitments led Du Bois into further battles with the US government. In 1951 he was arrested and charged with the crime of operating as an unregistered agent of a foreign power. The charges were eventually dismissed but this unhappy experience galvanized him in support for the Soviet bloc. This phase of his life was explored in yet another book *In Battle for Peace: The Story of My eighty-third Birthday* (1952).

1. *Du Bois' Scientific Contribution*

Du Bois was the first sociologist of the American South. He articulated the complexities of African-American life in comparative and historical terms that owed much to his multidisciplinary training. He placed sociological reason at the disposal of black Americans' freedom struggles and sowed the seeds that would bear fruit in the mid-twentieth century legal judgements that cited later sociological studies in support of desegregation. His own timely empirical contributions, particularly *The Philadelphia Negro* (1899) were also important because they replaced top-heavy theoretical speculation with an honest blend of theory with practical innovations in research technique. Guided as much by Schmoller as by the spirit of Charles Booth, Du Bois was also the first American social scientist to make empirical research the pivot of an investigation. His investment in the power of properly scientific sociology as a stimulus for governmental intervention and the pursuit of democracy unmarked by racial categories was considerable; however his faith in progress was neither simplistic nor naïve. Sociology was obviously indispensable but it was also insufficient. Its hard-won truths had to be communicated in several tongues, and made accessible to the widest possible audience. But more than this, Du Bois' achievements suggested that sociology was always at its most eloquent when supplemented and enhanced by the power of poetry.

2. *The Contemporary Influence*

In addition to his accomplishments as a distinguished social scientist Du Bois is still remembered as an essayist, editor, novelist, historian, and political leader. The breadth of his intellectual interests and the skill with which he turned them into a host of compelling written material in several distinct disciplines and styles made him the dominant African-American thinker of the twentieth century. Close to the aftermath of slavery and against the expectations of those who had judged 'Negroes' to be incapable of higher learning, his extraordinary work boldly redefined America's racial hierarchy as intrinsic to the incomplete workings of American democracy. In making that unpalatable argument in potent language which was both scholarly and poetic, Du Bois created a black intelligentsia which understood itself and its historical and sociological predicament in terms he had specified. His exceptional scholarly insights were articulated in a prolific output that continued throughout his long life. All of this work was characterized by a double achievement to match the double consciousness from which it grew—a sense of black intellectuals caught, or as he put it 'imprisoned,' within the folds of European civilization developing interests of their own distinct from those of the black poor.

He transformed understanding of African-American history and society while supplying the intellectual cornerstones for new varieties of critical and oppositional consciousness that could move beyond the task of defending America's black communities as they emerged from the twilight of slavery. In articulating an especially rigorous, energetic, and durable brand of dissident intellectual work, Du Bois not only provided solid foundations for the scholarly study of racism, he enabled his many successors to challenge white supremacy in the world of ideas, and provided important conceptual and historical resources for anticolonial movements worldwide.

See also: Race: History of the Concept; Racial Relations; Racism, Sociology of; Sociology, History of

Bibliography

Broderick F 1959 *W. E. B. Du Bois: Negro Leader In A Time of Crisis*. Stanford University Press, Stanford, CA

Du Bois W E B 1896 *The Suppression of The African Slave Trade To The United States of America 1638–1870* Harvard Historical Studies, no. 1, Longmans Green, New York

Du Bois W E B 1899 *The Philadelphia Negro: A Social Study*. University of Philadelphia Press, Philadelphia
Du Bois W E B 1903 *The Souls of Black Folk: Essays and Sketches*. A. C. McClurg, Chicago
Du Bois W E B 1911 *The Quest of The Silver Fleece*. A. C. McClurg, Chicago
Du Bois W E B 1928 *Dark Princess: A Romance*. Kraus-Thomson Organization, Millwood, New York
Du Bois W E B 1935 *Black Reconstruction In America: An Essay Toward a History of the Part Which Black Folk Played in the Attempt to Reconstruct Democracy in America 1860–1880*. Atheneum Press, New York
Du Bois W E B 1952 *In Battle for Peace: The Story of My Eighty-Third Birthday*. Masses & Mainstream, New York
Levering L D 1993 *W. E. B. Du Bois: Biography of A Race*. Henry Holt, New York
Locke A (ed.) 1925 *The New Negro*. A. and C. Boni, New York
Rampersad A 1976 *The Art And Imagination of W. E. B. Du Bois*. Harvard University Press, Cambridge, MA

P. Gilroy

Dual Task Performance

1. Introduction

Studies on dual task performance belong to the core research areas on divided attention and also have strong applied relevance. In the epoch of automation, it is important to know which tasks can be performed together more or less easily and how they contribute to work load. Intuitively people can do only a limited number of 'things' at once, but much may depend on the level of skill. During the 1970s this wisdom was expressed in terms of the influential conceptual metaphor of the human as a limited capacity processor which was inspired by the early digital computers. Present theoretical formulations emphasize 'cross-talk' and 'specification of action' as the main sources of performance limits in dual task performance. This article presents a concise review of the various theoretical positions. A more complete account can be found in Sanders (1998).

2. Limited Capacity Theory

The limited capacity was thought to be distributed over the tasks, permitting time sharing as long as the joint task demands did not exceed the limits, with the exception of a little capacity which was required for coordination of the tasks or for monitoring the environment. Effects of skill on capacity demands were described originally in terms of more efficient programming of the individual tasks. More recently, the notion of automatic vs. controlled processing has gained currency. In contrast to a controlled process, an automatic process—the ultimate result of practice—was not supposed to demand any capacity, and could therefore be always time shared. Sometimes, capacity was assumed to be constant (Moray 1967, Navon and Gopher 1979), whereas it was thought to vary as a function of arousal on other occasions (Kahneman 1973). A variable capacity has the problem that performance measures of capacity limits are basically unreliable. This was the reason why Kahneman proposed physiological measures with pupil size as favorite. Among others, pupil size was found to increase as a function of cognitive processing demands, e.g., the number of items to be rehearsed in a memory span task. However, such effects are easily overruled by the reaction of the pupil to light. More generally, there is the problem that cognitive information processing and autonomous physiological responses have no common conceptual framework.

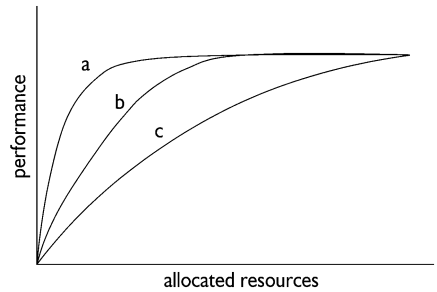

Figure 1
A performance–resource function describes the relation between performance and the number of allocated resources. In an easy task (a) data limits are reached much sooner than in the more difficult tasks (b, c) (from Sanders A F 1998 *Elements of Human Performance*. Erlbaum, Mahwah, NJ)

The most elaborate capacity theory stemmed from economics rather than from the digital computer (Navon and Gopher 1979). A performance–resource function (PRF, Fig. 1) depicts the relation between performance and invested capacity or resources. Performance improves as more resources are invested—the resource-limited part of the function—until reaching a maximum, which signals the start of the data-limited part of the function. As is clear from Fig. 1, a data limit means that performance does not improve further when more resources are allocated. Task difficulty is defined as marginal efficiency of the resources.

Unlike in economics, a behavioral PRF cannot be obtained directly but it can be estimated from a performance operating characteristic (POC) which describes the relation between performances of two

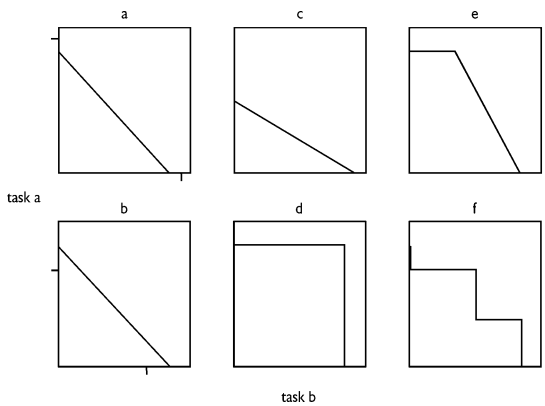

Figure 2
A performance operating characteristic describes the exchange relation between two tasks when performed singly or in combination. (a), (b), and (c) are examples of linear trade-off; (d) illustrates the case of perfect dual task performance; (e) shows initial time sharing followed by a linear trade-off; (f) shows a case in which trade-off is not gradual but occurs in a few rough steps (from Sanders A F 1998 *Elements of Human Performance*. Erlbaum, Mahwah, NJ)

tasks in a dual task setting. Examples of POCs are given in Fig. 2.

Figures 2(a)–(c) have linear exchange relations, suggesting that, when performed alone, either task consumes all resources so that performance suffers when the second task acquires some resources as well. The slope of the POC expresses the relative difficulty of the two tasks. Thus, in Fig. 2(c), task A is more difficult than task B, since a small loss on task A has the effect of a relatively large gain on task B. Figure 2(a) shows a case in which maximum performance on task A, when carried out alone, is somewhat better than dual task performance with an exclusive emphasis on task A. This difference is labeled concurrence costs, which is the opposite of concurrence benefits (Fig. 2(b)) in which case maximum performance on task A profits from the presence of task B. Doing two tasks together without any performance loss on either one (Fig. 2(d)) suggests pronounced data limits in both tasks. The available resources cannot be used by either task, thus enabling maximum joint performance. In Fig. 2(e) maximum performance on task A can still be combined with some activity on task B, followed by a linear exchange relation between both tasks. Presumably task A has some data limits but much less strongly than in Fig. 2(d). Finally, Fig. 2(f) raises the question of continuous vs. a discrete trade-off between the tasks. People may be only capable of dividing resources by way of a few discrete steps.

The main difference between economics- and computer-based capacity models is that the economics view assumes that people always aim at investing all resources. A proper study on dual tasks, therefore, should only be concerned with the resource-limited parts of the PRF. The difference between a difficult and an easy task is that people do more on the easy task but all resources are fully invested in either case. In contrast, computer models have usually defined difficulty in terms of resource demands, a difficult task leaving less spare capacity than a simple task. The two hypotheses have different predictions when capacity shrinks somehow. The economics theory suggests that the easy task will suffer most, since the quantitative effect of removing resources is larger on the easy task. In contrast, the computer model suggests that the difficult task has largest capacity demands and will therefore suffer most. The evidence appears to favor the economics model; as long as both tasks are resource-limited, limiting resources appeared to have more effect on the easy task.

The above capacity concepts are aspecific and assume task invariance. Aspecificity means that any task or activity consume the same resources. Task invariance means that the quantity of performance may vary with resource allocation but the nature of the task remains the same. All models implicitly assume a central executive which is responsible for the allocation policy. The distinction between resource and data limits adds the important suggestion that tasks may be timeshared because each one only needs part of the available capacity to reach its maximum performance. This will usually refer to capacity demands over time. Indeed, one popular measure of workload—timeline analysis—involves a time record of activities; an operator is supposed to be overloaded when task demands overlap in time, which implies the assumption that only a single activity is possible at any moment. It should be noted, though, that this issue—serial vs. parallel processing of demands from various tasks—is not raised in capacity models. The point is that they never specify whether 'capacity demands' refer to a single moment or to an extended period of time.

It is not surprising that models of dual task performance started with the above simple capacity notions They had the advantage of providing a simple approach to the applied problem of mental workload—the more demands, the higher is the load—and of task difficulty. In addition, they give a straightforward definition of selective attention in terms of capacity limits and allocation policy, and they suggested basic rules why two tasks can be sometimes timeshared whereas they are mutually exclusive on other occasions. An aspecific limited capacity was also an appealing metaphor during the 1960s and 1970s and had at least one interesting prediction. Thus the level of dual task performance should not depend on the specific characteristics of the various activities involved. In other words, if a task A suffers more from task C than from task D, task B should also suffer more from task C than from task D (Navon and

Gopher 1979). Perhaps the most compelling failure of this prediction stems from a classical study by Brooks (1968). He found that a task requiring spatial working memory (task A) was performed better in combination with a verbal processing task (task B) than with a spatial processing task (task D), whereas a task demanding verbal working memory (task B) was performed better in combination with spatial processing (task D) than with verbal processing (task B). Other evidence showed that search in verbal short-term memory was more easily combined with manual tracking than with recalling abstract words. However, searching spatial short-term memory was performed better in combination with recalling abstract words than with tracking. Various other examples are discussed by Wickens (1992). They all relate to effects on dual task performance of structural alteration, i.e., changing the structure of a task while keeping its difficulty constant, and to difficulty insensitivity, i.e., conditions in which increasing the difficulty of one task hardly affects the performance level of another. Both effects are at odds with a single aspecific limited capacity.

3. *Multiple Resources*

This was the reason for considering multiple and mutually independent resources (e.g., Sanders 1998), in which a task is characterized by its profile of demands for various resources. Wickens (1992) proposed three resource dimensions, one relating to input modality (visual vs. auditory), one to the relative emphasis in central processing (perceptual/cognitive vs. response) and one to central processing codes (verbal vs. spatial). This last dimension relates to the finding that a spatial task is usually easier to combine with a verbal task than with another spatial task. Similarly, there is evidence that, when two tasks have the same input modality, their joint maximum performance is less than when input modalities differ. Finally, a predominantly motor task, e.g., tapping a regular interval, can be combined more easily with a predominantly cognitive task, e.g., classifying letters, than with another motor task such as tracking. Tapping is insensitive to demands on classification (Donk and Sanders 1989).

Multiple resources have some important theoretical implications. Thus, there will usually be ample spare capacity. Whenever a task is limited to certain resource pools, the demands on those pools may exceed capacity whereas capacity from the remaining pools is not used. This undermines the notion of task difficulty. Indeed, one task may only need and overload one particular resource pool, whereas another task may require all resource pools without imposing a heavy load on any. Which one is harder? Viewed from the perspective of total resource consumption, the last task may be the most demanding. However, the first cannot be completed successfully, which may be decisive in the discussion of workload. This point also bears on assessing workload. In the single-capacity models this was simply a matter of resource consumption. In the case of multiple resources, resource allocation is multidimensional, which prohibits a single-dimensional workload concept. As an example, the proposal of assessing workload by measuring performance on a second task derived from the assumption that a main task leaves aspecific spare capacity which is allocated to the second task. However, from the perspective of multiple resources, the cognitive architecture of the main and second tasks controls the second task performance. The multiple resource view has the obvious merit of recognizing the role of the cognitive architecture of tasks. Finally, Wickens' three dimensions capture some important cognitive differences which are particularly relevant from an applied point of view. For example, when, as in car driving, a task is spatially loading, additional information, e.g., route information, should be processed verbally. In fact, studies showing almost perfect time sharing in complex tasks, such as piano playing and repeating a verbal message (Allport et al. 1972), used tasks that differed in all three of Wickens' dimensions. Thus, piano playing may proceed from visual input to analogue representation to manual output, whereas the verbal message was auditorily presented, verbally encoded, and ended with a vocal output. In addition, the participants in such studies were always highly skilled, which means a large marginal efficiency of resource consumption.

Despite these merits, the heuristic value of resource models has come under heavy attack (Neumann 1987). One reason was that, in line with the enormous processing capacity of the brain, the new generation of computers overruled the notion of limited capacity as a viable metaphor. Second, 'interference among tasks is usually much more specific than predicted on the basis of a limited amount of resources and, on the other hand, there is unspecific interference that seems not to depend on any specific resource' (Neumann 1987, p. 365). For instance, timesharing a spelling task and a mental arithmetic task is better than timesharing two arithmetic or two spelling tasks. However, they would all be subsumed under Wickens' verbal central processing resource. It is obvious that not much is gained by postulating a new resource for each new case in which timesharing is found. In regard to aspecific interference, Gladstones et al. (1989) had participants perform two machine-paced serial reaction tasks, singly and in combination, and found no indication for better shared performance when the stimuli for one task were visual and of the other task auditory in comparison with visual stimuli in both tasks. Performance did not profit either from a difference in effectors, that is, when one task had a vocal and the other a manual response, in comparison with two manual responses. Third, merely distinguishing

among resources runs the risk of conceptualizing information processing in terms of a set of distinct faculties without detailed process descriptions and without an integrative account. An attempt to account for the last point was the assumption of an additional general-purpose resource responsible for attention control in scheduling and timing. The notion of attention control was fostered by evidence that practice not only has the effect of diminishing resource demands but also that attention control profits from practice. Thus, practicing two single tasks until full proficiency does not mean that they can be easily time shared without additional practice in dual task performance (Detweiler and Lundy 1995). The relevance of scheduling and timing is evident from findings that two tracking tasks—which presumably share the same resources—are much easier to do together when their temporal dynamics are the same than when they deviate (Fracker and Wickens 1989).

4. *Interference*

The above objections led to accounts in terms of interference—outcome conflict and cross-talk—among simultaneous tasks, depending on the extent the various task demands compete in an abstract space. There may be cooperation, conceived of as a single more complex action plan for both tasks, or confusion, referring to problems in establishing correct input–output connections (Wickens 1992) As an example, in a dual task consisting of manual tracking and discrete choice reactions, tracking proficiency appears to impair briefly before a choice reaction, suggesting outcome conflict in motor programming. Neumann (1987) suggested two reasons for limits in performance, both concerned with the control of action, namely effector recruitment and parameter specification. The issue is not that the capacity is limited but that one has to choose among competing actions. Effector recruitment refers to the selection of effectors which should become active at a given moment. The main meaning of effector recruitment is to abort competing responses, in which way it may lead to a bottleneck in response selection. Parameter specification refers to getting the proper values, e.g., direction, force, size of a movement, but also selecting appropriate stimuli for completing an intended action. Practice has the effect that default values develop which means that the action can be completed without time-consuming specifications. Effector recruitment and parameter specification at the time of an action prohibit other actions from occurring at the same time. The consequence is that time sharing two actions is hard when they are independent. In contrast, complex skills usually consist of integrated sequences, summarized in chunks and groups of actions which are initiated by a single command. This may account for the finding that time sharing has been found mainly in well-trained complex tasks and not in uncorrelated reactions to successive stimuli, as for instance in the study of Gladstones et al (1989). Time sharing is better as tasks (a) allow processing and acting in larger units, (b) permit preview and advance scheduling of demands, and (c) do not involve similar or competing patterns of stimuli and actions.

Thus, formulating performance limits by way of interference puts full weight on processing architecture and timing of action. Its popularity is at least partly due to the advent of performance models in terms of connectionist networks (Detweiler and Schneider 1991) and to the present interest in the neurophysiology of behavior. A potential problem for the interference models may be a lack of predictive power. Thus far the research has been insufficiently specific about constraints and too occupied with the description of the existing results beyond the concept of limited capacity. Although the connectionist models are quantitative, their detailed predictions are still open to *post hoc* parameter setting.

With regard to future research, patterns of interference should be described as a function of the stage of processing at which such patterns are supposed to arise. This should lead to a more detailed catalog of actions which are liable to output conflict and, hence, to be avoided in task design. There are interesting recent developments into this direction, among which Pashler's research on the extent of parallel versus serial processing in responses to successive stimuli (e.g., Pashler 1994) and the EPIC simulation (Meyer and Kieras 1997) of human performance. In addition, much may be expected from the analysis of brain images while performing a dual task. The new approaches should enable a more detailed account of dual performance and of attention control of multiple tasks than was possible on the basis of capacity notions.

See also: Attention and Action; Attention: Models; Attention: Multiple Resources; Attention, Neural Basis of; Automaticity of Action, Psychology of

Bibliography

Allport A, Antonis B, Reynolds P 1972 On the division of attention: a disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology* **24**: 225–35

Brooks L R 1968 Spatial and verbal components of the act of recall. *Canadian Journal of Psychology* **22**: 349–68

Detweiler M C, Lundy D H 1995 Effects of single and dual-task practice on acquiring dual-task skill. *Human Factors* **37**: 193–211

Detweiler M C, Schneider W 1991 Modelling the acquisition of dual task skill in a connectionist/control architecture. In: Damos D L (ed.) *Multiple Task Performance*. Taylor & Francis, Washington, DC, pp. 247–95

Donk M, Sanders A F 1989 Resources and dual task performance: resource allocation versus task integration. *Acta Psychologica* **72**: 221–32
Fracker M L, Wickens C D 1989 Resources, confusions and compatibility in dual axis tracking: displays controls and dynamics. *Journal of Experimental Psychology: Human Perception and Performance* **15**: 80–96
Gladstones W H, Regan M A, Lee R B 1989 Division of attention: the single channel hypothesis revisited. *Quarterly Journal of Experimental Psychology* **41**: 1–17
Kahneman D 1973 *Attention and Effort*. Prentice Hall, Englewood Cliffs, NJ
Meyer D E, Kieras E 1997 A computational theory of executive cognitive processes and multiple task performance. Part 1. Basic mechanisms. *Psychological Review* **104**: 3–65
Moray N H 1967 Where is capacity limited? A survey and a model. *Acta Psychologica* **27**: 84–92
Navon D, Gopher D 1979 On the economy of processing systems. *Psychological Review* **86**: 214–55
Neumann O 1987 Beyond capacity: a functional view of attention. In: Heuer H, Sanders A F (eds.) *Perspectives on Perception and Action*. Erlbaum, Hillsdale, NJ, pp. 361–94
Pashler H 1994 Dual-task interference in simple tasks: data and theory. *Psychological Bulletin* **116**: 220–44
Sanders A F 1998 *Elements of Human Performance*. Erlbaum, Mahwah, NJ
Wickens C D 1992 *Engineering Psychology and Human Performance*. HarperCollins, New York

A. F. Sanders

Dubos, René Jules (1901–82)

René Jules Dubos, the French-American microbiologist, medical researcher, environmentalist, and Pulitzer-prize-winning author, was born in Saint-Brice-sous-Fôret, France on February 20, 1901. He was the eldest son of Georgesand Adeline DeBloëdt Dubos, proprietors of a series of small butcher shops. Two years after his birth, the Dubos family moved to neighboring Hénonville, a small agricultural village in the Île de France region fifty miles north of Paris. René Dubos spent the next dozen years of his life there exploring the French countryside, reading avidly, and excelling as a student at *l'ecole communale d'Hénonville*, the local one-room primary school. Encouraged by his mother to pursue a scholarly career, he dreamed of attending France's distinguished schools of higher learning. The family moved to Paris in 1914. Dubos won a fellowship in 1915 to attend the *Collège Chaptal*, part of the French secondary school system. His father's untimely death at the conclusion of World War I, his family's concomitant financial difficulties, and his own periodic bouts with poor health resulting from an early battle with rheumatic fever, however, prevented René Dubos from studying to become a historian, his first choice among academic careers. Instead, he gained admission in 1919 to the *Institut National Agronomique* (INA), France's National School of Agronomy.

Although many graduates of the INA pursued careers in scientific agricultural research, Dubos disliked laboratory work. Instead, he accepted a fellowship from the French government of Indochina to study at the *Institut National d'Agronomie Coloniale*, where he prepared for a position as a colonial administrator. However, the colonial government revoked its offer of employment at the conclusion of the training program when Dubos failed a physical examination. Having few alternatives, Dubos accepted an editorial position at *l'Institut International d'Agriculture* in Rome, where he abstracted technical articles for publication in the *Journal of International Agricultural Intelligence*. While in Rome, Dubos experienced a change of heart regarding his career plans. Inspired by an argument made in an article by the Russian agricultural bacteriologist, Sergei Winogradsky, Dubos reconsidered his decision to steer clear of laboratory research. In this article, Winogradsky encouraged his colleagues to reject pure culture techniques in favor of studies that more accurately monitored soil microbes in their natural environments. This more ecological approach to microbial investigation appealed to Dubos because it emphasized themes familiar to him from the training he had received as an agricultural engineer in Paris. Dubos gained the opportunity to study soil microbes himself when, at the 1924 International Congress of Soil Science, Jacob Lipman of the New Jersey Agricultural Experiment Station invited the young Frenchman to study soil microbiology at his institution. Dubos accepted the offer, sailed to the USA, and spent the next three years in a doctoral program under the direction of Selman Waksman at Rutgers University and the New Jersey Agricultural Experiment Station.

Soon after completing his Ph.D. in soil microbiology, Dubos received an unusual and compelling offer to join the research team of Oswald Avery, an immunochemist at the Rockefeller Institute for Medical Research. Few agricultural researchers received an opportunity to apply their skills to medical research in the 1920s, particularly at the nation's most prestigious medical research institution, but Dubos's doctoral research on cellulose decomposition filled a specific need in the Avery laboratory. For years, Avery's laboratory group had tried unsuccessfully to strip the polysaccharide coating from the type III pneumococcus responsible for lobar pneumonia. Avery theorized that the coating protected the bacillus from the human body's natural defenses. Dubos suggested that he could find a soil microbe with the ability to produce an enzyme capable of dissolving the polysaccharide coating. Avery gave the young agricultural scientist a chance.

Thus, in 1927, René Dubos began a nearly fifty-year association with the Rockefeller Institute for Medical Research (later Rockefeller University). Only a two-

year stint as George Fabyan Professor of Comparative Pathology and Tropical Medicine at Harvard Medical School, 1942 to 1944, interrupted his career-long affiliation with the Rockefeller Institute. As a medical researcher, Dubos made numerous contributions to bacteriology and medical science. In the Avery laboratory group, he extracted an enzyme from a soil microbe that proved effective against the polysaccharide coating of the lobar pneumonia bacillus. Although the advent of sulfa drugs precluded the use of the Dubos enzyme in the clinical treatment of pneumonia, the process he used to extract the enzyme proved beneficial in the late 1930s, when he returned to the soil to find a microbe with broader bactericidal properties. In 1939 he earned international recognition by developing gramicidin, the first clinically useful antibiotic. His success suggested the clinical potential of soil microorganisms to researchers around the world. News of Dubos's advances spurred Oxford University's Howard Florey and Ernst Chain to resurrect Alexander Fleming's long-dormant research on penicillin. Dubos's work also inspired the effective development of streptomycin by his graduate school advisor, Selman Waksman. Collectively, these advances marked the beginning of the antibiotic era. Dubos's leading role in antibiotics research earned him the status of full member at the Rockefeller Institute, and, in 1941, election to the National Academy of Sciences.

Although his advances in antibiotics research represented his most celebrated contributions to medical science, they were by no means his only accomplishments. In the late 1940s Dubos gained recognition as an outstanding tuberculosis researcher. With the assistance of Bernard Davis, he developed an effective culture medium for tuberculosis bacilli. This new Tween-albumin medium permitted researchers to study the difference between virulent and nonvirulent bacilli and was credited with catalyzing a 'renaissance in tuberculosis research.' A strong advocate for vaccination against tuberculosis, Dubos also established guidelines for the standardization of BCG vaccines worldwide. Finally, in the 1950s Dubos collaborated with physicians in the Rockefeller Hospital on a study which revealed that prolonged bed rest did not expedite the recovery of tuberculosis patients receiving antibiotic therapy. This discovery had profound implications for sanatoriums, the traditional sites for tuberculosis treatment.

Despite his pathbreaking contributions to the modern pharmacopoeia, Dubos is more often remembered for his provocative ideas on human health. No one contributed more to the hubris of medical science at mid-century than Dubos with his breakthroughs in antibiotics research, but shortly after unveiling gramicidin, he publicized his concerns about the adequacy of antibiotic therapy for handling infectious disease. As early as 1942 he warned that the unmitigated enthusiasm for antibiotic wonder drugs should be tempered by a concern that overprescription of the drugs could lead to the development of more virulent, resistant bacterial strains. He argued that scientists gained an incomplete understanding of infectious disease when they focused solely on infectious agents. His studies of tuberculosis and the loss of his first wife to the disease taught him that, particularly in the case of latent infection, the host's ability to resist a disease is at least as important as the presence of an infectious agent. Pathologic processes, in his opinion, depended upon the total environment of the host, and deserved the serious attention of scientists. This challenge to the specific etiology of disease contributed significantly to the formulation of current conceptions which now include social factors, mental and physical disturbances, and environmental conditions in the understanding and management of disease.

Dubos's reconceptualization of disease, environment, and human health reached a broad audience of medical researchers, social scientists, politicians, and laypeople. He developed his ideas through over twenty books, countless articles in scholarly and popular journals, and an ever-increasing number of public addresses between the 1940s and the late 1970s. His most influential writings on human health include *The White Plague: Tuberculosis, Man, and Society* (1952), *Mirage of Health: Utopias, Progress, and Biological Change* (1987), and his Pulitzer-prize-winning text, *So Human an Animal: How We are Shaped by Surroundings and Events* (1998). In *The White Plague*, a book Dubos co-wrote with his second wife, Jean Porter Dubos, they mined historical accounts of tuberculosis outbreaks to demonstrate that the incidence of tuberculosis increased during times of social strife and decreased during less tumultuous eras. By linking it to poverty, they argued that tuberculosis was as much a social as well as a microbiological disease. In *Mirage of Health*, Dubos again challenged his readers to recognize disease as the indirect outcome of a constellation of circumstances rather than the direct result of a virulent microorganism.

Microbial disease, when it occurred, was due to an imbalance in the host's equilibrium caused by any of a number of factors in the external or internal environment—be it weather conditions, availability of food, working habits, emotional status, or emotional stress. Therefore, he argued, by ameliorating social and environmental conditions, social reformers had been more effective than scientists in reducing mankind's susceptibility to disease. Disease, according to Dubos, could not be eradicated simply by medical treatment. Social conditions must also be managed. His message enjoyed broad appeal among social scientists, many of whom, like Dubos, refocused their scholarly inquiries toward research that revealed how changes in an organism's social environment affected its health. Scientists also utilized Dubos's biopsychosocial perspective on health to construct a theoretical justification for the study of psychosomatic medicine.

By the 1970s, Dubos had become one of the USA's most visible scientists. His message appealed to a large number of Americans interested in the increasingly apparent connection between the environment and human health. Although organizations invited Dubos to speak on the environmental aspects of health and disease as early as the 1950s, his influence among environmentalists grew substantially after the publication of his Pulitzer-prize-winning book, *So Human an Animal*, in 1998. The book appeared when the USA was in the midst of the largest grassroots environmental movement in its history. Humans, like soil microbes, argued Dubos, adapt to their changing environments. Therefore, the quality of the environment in which a human develops holds great significance in the determination of human health and well-being. Although some environmentalists took issue with Dubos' notion that humans could improve on nature, many embraced him as an elder statesman of their movement. They adopted several of his maxims, including 'Think globally, but act locally' and 'Trend is not destiny,' as rallying cries for their cause. Dubos remained active in environmental activities years after his 1971 retirement from Rockefeller University. He co-wrote *Only One Earth* with economist and political scientist, Barbara Ward. This document provided scientific and social guidelines for the First UN Conference on the Human Environment in Stockholm in 1972. In 1977, Dubos lent his name to the New York-based René Dubos Center for Human Environments, an organization charged with the planning of international forums to promote beneficial human interventions into nature.

Dubos traveled an uncommon path from his early days as a French agricultural engineer to his respected position within the Rockefeller University and twentieth-century biomedical research. Although recognized and applauded as part of the scientific establishment by the 1940s, Dubos never hesitated to challenge the methodologies of his colleagues. Because he believed that all living organisms must be understood in the context of their respective environments, he advocated less reductionistic laboratory research, studies that traversed interdisciplinary boundaries, and an ecological understanding of disease. He followed his inner logic outside the confines of scientific safety, where he found diverse and interested audiences willing to engage his provocative message. René Dubos died on his 81st birthday in New York City.

See also: Biomedical Sciences and Technology: History and Sociology; Environmental Health and Safety: Social Aspects; Environmental Stress and Health; Public Health as a Social Science

Bibliography

Benison S 1976 René Dubos and the capsular polysaccharide of pneumococcus. *Bulletin of the History of Medicine* **50**: 459–77

Cooper J 1998 Of microbes and men: A scientific biography of René Jules Dubos. Ph.D. thesis, Rutgers University

Dubos R 1987 *Mirage of Health: Utopias, Progress, and Biological Change*, 2nd edn. Rutgers University Press, New Brunswick

Dubos R 1992 *The White Plague: Tuberculosis, Man, and Society*, 2nd edn. Rutgers University Press, New Brunswick

Dubos R 1998 *So Human an Animal: How We are Shaped by Surroundings and Events*, 2nd edn. Transaction, New Brunswick, NJ

Hirsch J G, Moberg C 1989 René Jules Dubos. *Biographical Memoirs of the NAS* **58**: 132–61

Moberg C L 1996 René Dubos: A harbinger of microbial resistance to antibiotics. *Microbial Drug Resistance* **2**: 287–97

Moberg C L, Cohn Z (eds.) 1990 *Launching the Antibiotic Era: Personal Accounts of the Discovery and Use of the First Antibiotics*. Rockefeller University Press, New York

Piel G, Segerberg O 1990 *The World of René Dubos: A Collection from His Writings*. Holt, New York

J. Cooper

Due Process in Law

The concept of due process emerges from the central paradox in the creation of state authority. Following Thomas Hobbes, the governmental state may be seen as an evolutionary response to the failings of the state of nature. In the war of all against all, no man has an incentive to industry, writes Hobbes in the *Leviathan*, because each lacks security in investment, planning, and the ability to exchange. Each faces the threat of bodily violence, invasion of the home and seizure of property. The response, per Hobbes, is the creation of the state, a body that can impose order, lift mankind from the pitiable war of all against all, and secure collective benefits. The paradox emerges in that the creation of the state also produces a magnified form of the threat that was ever-present in the state of nature. The state represents the largest entity capable of inflicting bodily violence, invading the home, and seizing property. Indeed, all states reserve to themselves the right to do just that to their citizens, and most include the right even to punish by death. What then differentiates the modern state that claims political legitimacy from the state of nature? The answer must be an abiding respect for procedural fairness in the imposition of punishment and the regulation of the conduct of the citizens' lives.

This conception of procedural fairness is at the heart of the concept of the rule of law, and of its more technically exigent offshoot, the guarantee of due process. The typical formulation of due process, as that concept has developed in Anglo-American jurisprudence, is derived from the Magna Carta, issued by King John in 1215 in the face of rebellious barons seeking restraint upon the exercise of royal prerogatives:

> No free man shall be seized or imprisoned, or stripped of his rights or possessings, or outlawed or exiled, or deprived of his standing in any other way, nor will we proceed with force against him, or send others to do so, except by the lawful judgment of his equals or by the law of the land.

The crux of the original formulation was to limit the power of the royal sovereign by granting legal protection to the life, liberty, and property of English freemen. The significance of the Magna Carta's guarantees emerges from the form of the limitation on sovereign authority. The document does not attempt to deny the sovereign the capacity to imprison, outlaw or banish the rights-holding citizenry, or to have restraints imposed on the enjoyment of property or other rights held by citizens. Rather it imposes a condition of procedural regularity in the exercise of the sovereign's power, including central, orderly trial processes and formal proclamation of the applicable legal standards. This notion of procedural regularity is captured in the actual phrase, 'due process of law,' which first appeared when the Magna Carta was issued in an English version in 1354 by King Edward III. Its modern vitality draws from the efforts of Sir Edward Coke in the seventeenth century to revive the Magna Carta as a check on the power of the Stuart monarchs.

Not surprisingly given these origins, due process emerged from medieval Britain primarily as a restraint on the executive power of the sovereign, specifically of the Crown. The concept of due process was transformed when the American colonists adapted the Magna Carta as a limiting principle not only on the use of executive power, but also on the scope of the laws that could be passed. For colonists lacking the rights of participation in the creation of laws, the insult of arbitrary state power occurred whether power was exercised legislatively or by the sovereign. As recast by the colonists, due process became a limitation that extended beyond the power of the crown, to encompass as well local assemblies and, by the time of the revolution, Parliament. Thus, even as England adopted a system of increasing parliamentary sovereignty, the concept of due process as a constraint on the exercise of governmental power retained its vitality in the colonies. As a result, the newly independent Americans established a more limited government, reflected by the Bill of Rights (including a due process clause), as an explicit limitation on the use of state power, including the legislative branch. As articulated by James Madison, 'The legislature, no less than the executive, is under limitations of power,' and hence the rights of the people were to be secured 'against legislative as well as executive ambition' (Madison 1896). The source of the limitation was a constitutional command that stood apart and above the normal operations of state authority.

The American focus on legislative power also highlights the importance of due process as a restraint not just on the penal powers of the state, but on its civil regulatory powers as well. Due process emerges from this historical origin with two central features. First, it protects substantive rights independently secured by positive law by requiring procedural regularity in the exercise of all governmental power. Second, the requirement of procedural regularity may not be altered by the normal operations of political power. As reflected in the American constitutional experience, the guarantees of due process are not directly accountable to the political dictates of either the executive branch of government (the closest in form to the sovereign), or the legislative will of the majority. As a constitutional guarantee, the requirements of due process may not be altered though the customary political processes of legislation. It is these two features of procedural regularity and independence from direct political accountability that define due process as it has emerged most clearly in American constitutional law, which long ago held due process to be a protection of 'those fundamental principles of liberty and justice which lie at the base of all our civil and political institutions' (*Hurtado* vs. *California*, 110 US 516, 535 (1884)). Because the development is most clear in American constitutional law, that will serve as the template for further elaboration.

There are five basic elements of due process that may be adduced from American case development, particularly after the post-Civil War amendments to the Constitution extended the command to the conduct of the States toward their citizens. The two given the greatest significance are that, 'Parties whose rights are to be affected are entitled to be heard; and in order that they may enjoy that right they must first be notified' (*Baldwin* vs. *Hale*, 68 US (1 Wall.) 233 (1863)). The opportunity to be heard and its 'instrumental corollary, a promise of prior notice' (Tribe 1988), according to the Supreme Court, 'seem to be universally prescribed in all systems of law established by civilized countries' (*Twining* vs. *New Jersey*, 211 US 78, 110–11, 1908). The additional elements of due process are further corollaries necessary to make the right to a hearing meaningful: the right to an impartial arbiter, the right to be represented by counsel, and the right to timely resolution of claims.

The articulation of these elements of due process leaves unanswered a central question dating from the time of the Magna Carta: are these due process rights intended as affirmative entitlements of the citizenry, as indispensable restraints on the state, or as instrumental checks against arbitrary, capricious, or simply misguided state conduct? Put another way, the constitutional enshrinement of a principle of due process can be seen as creating a distinct set of rights among the citizenry, or as a constraint on governmental processes. Measured against this template, it is possible to characterize the development of twentieth-century due process law as passing through three central phases.

1. Substantive Due Process

The most forceful introduction of due process into the American constitutional canon came in the closing years of the nineteenth century with the emergence of the substantive due process doctrine. Federal courts gave independent substantive meaning to the Fourteenth Amendment's prohibition on states' restricting the enjoyment of life, liberty, or property without due process of law. Under this view, state incursions upon the affirmative rights of citizens to be secure in their life, liberty, and property were presumptively invalid and subject to exacting judicial scrutiny. The key to this period was an expansive set of presumed individual claims against state regulation. The domain of legitimate state activity was restricted to insuring the public health, protecting against inherently hazardous activities, and protecting those who could not fend for themselves at law. This last group included (at the time) children, the mentally incompetent, and women. The most controversial feature of substantive due process came with the expansive definition of 'liberty' to include a broad range of market freedoms that in turn prompted federal courts to strike down state regulatory measures by the dozens. Most famous no doubt was the Court's 1905 ruling in *Lochner* vs. *New York* (198 US 45, 1905), a case that struck down a maximum hours regulation for the employment of bakers on the substantive due process grounds of 'liberty of contract.' The hallmark of this period was the use of a high threshold of procedural presumptions against the legitimacy of state regulatory initiatives. The period of substantive due process lasted until its repudiation by the Supreme Court in the New Deal period of the 1930s. Nonetheless, the reaction to the perceived hostility of the Court to progressive social legislation earned the '*Lochner* period' notoriety as a constitutional epithet and removed substantive due process from the constitutional lexicon, until its tentative re-emergence at the end of the twentieth century in difficult cases such as the constitutional review of assisted suicide (*Washington* vs. *Glucksberg*, 521 US 702, 1997).

2. Procedural Due Process

The second great phase of due process activism came with the Warren Court of the 1960s. Unlike the *Lochner* period's focus on the expansion of substantive claims that triggered due process, the hallmark of the second phase was a focus on 'procedural due process,' which was understood as an expansion of the process-based requirements for the state to act adverse to the interests of affected citizens. In a series of decisions involving legislatively-conferred benefits, ranging from driver's licenses to welfare benefits, the Court interposed an elaborate series of processes required before such benefits could be withdrawn. The right to a hearing was repeatedly invoked in this period as an affirmative right of participation in governmental processes that inhered in the rights of citizens. The Court did not seek to expand the ambit of positive rights—entitlements were created by the normal processes of government—but instead used the Due Process Clause to enforce heightened procedural hurdles before government could deny a benefit. Instead of the liberty to contract of the early twentieth century, the aggrieved due process claimant of the post-War period would claim a 'right to be heard' prior to adverse governmental action. Critics charged that the elevated procedural steps created a presumption against adverse governmental action almost as certain as that of the substantive due process period. But more problematic, as evidenced in the ensuing case law, was how to reconcile near-categorical requirements of an inherent participatory right to process with a range of governmental conduct that required swift response to emergency situations. The case law struggled to find a mediating principle that would allow emergency responses, as with police searches, the impounding of goods in flight, the embargo on goods threatening public health or safety, and a range of comparable conduct whose successful implementation could not await notice and an adversarial hearing.

3. Due Process Functionalism

The tension in the heightened proceduralism of the Warren Court prompted a scaling back of inherent rights in favor of an instrumental approach to due process. Where both the substantive due process and procedural due process period had read the due process command to vest the citizenry with affirmative rights against the state, the functional period beginning in the mid-1970s saw in due process a check against improper or incorrect governmental decisions affecting life, liberty, or property. Thus, in the watershed case of *Matthews* vs. *Eldridge* (424 US 319, 1976), the Supreme Court analyzed a challenge to the termination of social security disability benefits not in terms of whether due process protections attached, but rather in terms of how much process was due. By contrast to cases of the *Goldberg* vs. *Kelly* period that assumed that process would have to be the full panoply of evidentiary protections, *Matthews* vs. *Eldridge* held that the amount of process required prior to adverse governmental conduct would be determined by a balance of interests heavily weighted toward considerations of cost and likelihood of error:

> [D]ue process generally requires consideration of three distinct factors: first, the private interest that will be affected by the official action; second, the risk of an erroneous deprivation of such interest through the procedures used; and the probable value, if any, of additional or substituted

> procedural safeguards; and finally, the government's interest, including the function involved and the fiscal and administrative burdens that the additional or substitute procedural requirement would entail. (424 US at 334–35)

In the decades since *Matthews* vs. *Eldridge*, this form of due process functionalism has come to command this area of American constitutional law. All the critical elements of due process—from the right to an evidentiary hearing to the right to counsel—are now subject to the balancing of magnitude of loss versus likelihood of error, versus governmental interest in swift action. Some of the applications appear inevitable, as with the ability of police to immobilize illegally parked cars, even in the absence of notice and a prior hearing. Some are more problematic, as with the denial of the right to have counsel provided to an indigent prisoner facing civil proceeding for termination of parental rights. But, nearly 800 years after the Magna Carta, this functional view of due process has firmly taken hold in the most developed of due process legal regimes.

4. *Due Process in Other National Legal Systems*

Broadly speaking, a large number of the fundamental rights accorded to individuals under the American conception of due process are also found in other nations' laws, with one critical caveat. The conception of due process, particularly in civil law systems, primarily addresses rights of criminal procedure—e.g., the French Declaration of the Rights of Man, which explicitly grants the right to be free from arbitrary arrest and detention. Various international agreements—such as the Convention for the Protection of Human Rights and Fundamental Freedoms—contain similar language. Many national constitutions similarly guarantee the rights to notice, choice of counsel, a speedy trial, and appeal.

Even in the sphere of criminal procedure, there are significant differences between the American notion of due process and that found in other countries. Perhaps the most striking example is the absence in many foreign legal regimes of the right to be presumed innocent that is considered so fundamental in the United States. Also absent in many countries is a notion that both sides in a criminal trial have equal opportunity to advocate their positions.

More revealing is a comparison of the American and German conceptions of due process. On one level, the two countries' approaches are similar. Both put a pre-eminent value on human liberty. Thus, Article 1(1) of the German constitution states that 'the dignity of man shall be inviolable.' This notion of human dignity is akin to the American concept of substantive due process. Both are tools that have been utilized by the respective courts to 'discover' hitherto unenumerated rights. In this sense, both countries have a conception of due process that is, to a certain extent, open-ended. As a consequence, both must grapple with an abstract and amorphous concept—often with controversial results.

But perhaps the most fundamental difference lies in their highest courts' approaches to constitutional interpretation. While the German Constitutional Court has interpreted its constitutional mandate to develop a systematic approach to constitutional jurisprudence, the United States Supreme Court, by emphasizing procedural regularity, has exhibited a tendency to act in a piecemeal fashion. The result is that, in some ways, the German view of due process appears even more comprehensive than that of the United States. Most importantly, Germany guarantees the protection of 'personality rights' and 'physical integrity,' which are akin to the American procedural due process. But, unlike the United States Supreme Court, which has often worked to construe due process narrowly, the German Constitutional Court has consistently taken a broad view of due process, viewing as properly furthering the fundamental principle of human dignity—a far more sweeping command than recognized in the initial procedural focus of Anglo-American law.

See also: Balance of Power: Political; Constitutional Courts; Constitutionalism; Fundamental Rights and Constitutional Guarantees; Law: Overview; Rechtsstaat (Rule of Law: German Perspective); Rights: Legal Aspects; Rule of Law

Bibliography

Hobbes T 1994 *Leviathan*. J M Dent, London

Madison J 1896 Report on the Virginia Resolutions. In: Elliot J (ed.) *Debates in the Several State Conventions on the Adoption of the Federal Constitution*. Lippincott, Philadelphia, PA, Vol. IV

Tribe L 1988 *American Constitutional Law*. Foundation Press, Mineola, NY

S. Issacharoff

Durkheim, Emile (1858–1917)

Émile Durkheim's writings, though they cover a vast range of topics, reveal a single, focused, organizing vision of sociology's subject matter and objectives. These went beyond the purely scientific and were also moral and political. He sought, through education, to contribute to the stability and legitimacy of the French Third Republic and, more widely, to find cures for modern society's ills. His writings include several

classics that remain central to the discipline of sociology, through teaching and through successive reinterpretations in which advances continue to occur by confrontation with Durkheim's ideas. These ideas, developed in close collaboration with a school of remarkably talented younger colleagues grouped around the journal, *L'Année sociologique* which flourished before World War I, had a considerable impact within France and later far beyond, within sociology and across several other disciplines.

1. Biography

Born on April 15, 1858 in Épinal in the Alsace-Lorraine region into a close-knit rabbinic family, he abandoned Judaism at an early age. He thereafter treated religion agnostically as a social phenomenon demanding sociological explanation (from the mid-1890s it was increasingly central to his thought) yet maintained that in studying it one is blind without a certain religious sentiment. After his studies at the *École Normale Superieure* in Paris, he taught philosophy in secondary schools. On a government grant he visited Germany, for whose scientific education there was widespread admiration in defeated France. His encounter with German social scientists strengthened his so-called 'social realism.' In 1887 he was appointed to teach social science and pedagogy at the University of Bordeaux, as part of a national drive for a new system of republican, secular education. In 1896 he was appointed the first Professor of Social Science in France and in 1897 founded *L'Année sociologique*. While at Bordeaux he published three major works, *The Division of Labor in Society* (Durkheim 1893), *The Rules of Sociological Method* (Durkheim 1895), and *Suicide* (Durkheim 1897) and his Latin doctoral thesis on Montesquieu. He also lectured and published across a wide range of topics including education, the family, the incest taboo, crime and social health, the evolution of punishment, the state, professional ethics and civic morals, religion, Rousseau, the history of sociology, and the history of socialism—the last in response to his students' political concerns. Averse to the sharpening of class conflict and believing revolutions to be as impossible as miracles, he stressed the need for institutional reforms that would encourage cross-class solidarity around common principles of social justice. He favored an interventionist State offset by a plurality of local, civic, and especially occupational associations, believing that out of this conflict of social forces individual liberties are born (Durkheim 1950). He remained close to the liberal socialism of Jean Jaurès, but in the late 1890s the Dreyfus Affair induced him to break his principle that the social scientist should, as such, keep apart from everyday politics. An active Dreyfusard, he denounced those who defended the army and the Church as sacred national institutions and proclaimed that modern society could only be integrated around the religion of individualism, protecting the sacredness of the individual and individual rights, the 'sole link' still binding citizens to one another.

In 1902 he moved to Paris to teach education and sociology. He continued to lecture and write across a wide range, publishing a study of *Primitive Classification* with Marcel Mauss (Durkheim and Mauss 1903) and his masterpiece *The Elementary Forms of the Religious Life* (Durkheim 1912). He debated with philosophers, historians, and psychologists and exercised a magisterial influence on the institutional development of sociology and on educational thinking and practice in France. On all these fronts his doctrines generated controversy and opposition, often fierce, to which he responded vigorously. Much of his energy went into the editing of *L'Année*, of which 12 volumes had appeared by 1913. These contained monographs and reviews of works in many disciplines and languages, systematically classified and mined for their theoretical and empirical usefulness to a progressing research program. An extraordinary collaborative achievement, they constitute the Durkheimians' ongoing debate with the sociologically relevant literature of their time, distinctive, among other things, in its comprehensive attention to the ethnographic record of the world's remoter peoples. All of this ended with the onset of the World War on whose battlefields several, including the most brilliant and promising of them, including Durkheim's son André, met their deaths. For a second time Durkheim broke his principle of political noninvolvement, contributing to the war effort and writing patriotic pamphlets, but, unable to overcome his grief at his son's death, he died in Paris on November 15, 1917.

2. Major Contributions

The focus of Durkheim's sociological vision was his so-called 'social realism': the view that social phenomena, or as he called them *faits sociaux*, are *sui generis*, with their own characteristics: they are general among, external to, independent of, and constraining upon particular individuals and typically are recognizable by their obligatory force. They constitute a distinct level of reality and are not reducible to the meeting of individual intentions. Hence, his attacks on social contract theorizing and on other forms of methodological individualism that attempted just such a reduction, as practiced for instance by utilitarians, economists, psychologists, and the English sociologist Herbert Spencer. At different stages he focused on *faits sociauz* of different kinds: on 'morphological' phenomena such as population volume and 'dynamic density' and spatial distribution, on types of law (see Cotterrell 1999), on institutions and, increasingly, on what he came to call *représentations collectives* (Durkheim 1924). By this he meant collective ideas and sentiments which could, moreover, be evanescent, like crowd enthusiasms or fashions, or more 'crystal-

lized' or institutionalized, like legal codes or most stable of all, the fundamental categories of thought. *Faits sociaux* were what we would now call emergent phenomena whose properties, resulting from the structured interactions of their constituents, could not be accounted for exhaustively in terms of individuals' presocial properties.

Durkheim advanced this view polemically, as a doctrine, above all in *The Rules*. It has been suggested that this is best understood in its context as the forging of a vocabulary suited to securing French citizens' allegiance to their Republic (Jones 1999). What is certain is that Durkheim's social realist rhetoric generated unproductive controversies and, worse, led to misunderstandings that have, even until today, deflected attention from his explanatory practice, to which his methodological pronouncements are an inadequate guide. Thus he claimed that only social causes can explain social effects and that every time a social fact is explained directly by a psychological fact one could be sure the explanation is false. In various places he uses language that personifies collectivities, making 'society' appear as a subject of experience, with a *conscience commune*, or as an agent with a unified will, requiring conformity, pursuing ends, and imposing ways of acting on its members.

But against such passages one can cite others that stress that society only exists in and through individuals, that neither can exist without the other, and that *représentations collectives* are the product of individuals' activity and, indeed, of an immense cooperation extending across space and time. In fact, Durkheim's actual explanations are by no means confined to the macrolevel: he sometimes seeks to explain the micro by the macro, as in *Suicide*, and the generation of macrophenomena at the microlevel, as in the generation and regeneration of religious faith through collective effervescences and rituals in *The Elementary Forms*. In short, Durkheim's explanatory practice was interactionist, not holistic. Unfortunately, his polemical antimethodological individualism and hostility to psychological explanation precluded him from reflecting on the microunderpinnings of his macro explanations and, in general, on micro–macro connections, so that these aspects of his explanations remain under-theorized. Yet this very inattention may have generated insights. By not seeing certain things, perhaps he focused more intensely on what he did see.

Faits sociaux, he wrote in *The Rules*, should be treated 'as things' ('*comme des choses*'), independent of and external to observers as well as actors, and enabling one to decide between competing theories. This expresses his commitment to scientific method as he saw it, as against philosophical theorizing uncontrolled by empirical evidence (to philosophize, he once told a doctoral candidate, is to think what one wants) and as against the Positivism of Comte, committed to the identification of a small set of transhistorical laws, rather than the systematic study of specific causal mechanisms (he rejected the label 'positivist,' preferring that of 'rationalist'). Social scientific method required the abandonment of pretheoretical preconceptions, the systematic testing of theories by evidence, quantitative where appropriate (as amply illustrated in *Suicide*), the careful selection of indicators, the rejection of teleological and functional explanations (though not the identification of social functions), the use of comparison (sociology's substitute for experiment), and the construction of social typologies. He conceived of types of society within a broadly evolutionary framework, proceeding from the least to the most advanced societies (and citing recent and contemporary tribal societies and those of Antiquity as instances of the former). He further thought that one could distinguish between 'normal' and 'pathological' developments within each type, at each stage of its development. This enabled him, he thought, to diagnose certain aspects of modern industrial societies, notably the prevalence of *anomie*, as pathological. The problem here was that his notion of normality was ambiguous between a statistical definition and one that embodied an implicit theory of social and psychological health or equilibrium relative to a given society's 'conditions of existence.' (How could a pathological phenomenon be both prevalent and abnormal?) Such diagnosis and the elaboration of feasible institutional remedies were, he thought, the chief ways in which the sociologist could contribute to social progress.

The Division of Labor contains (a) the positing of two ideal types of society, which Durkheim took to characterize less and more advanced societies respectively (but which can evidently co-exist), (b) an analysis of the mechanisms claimed to explain the transition from the former to the latter; and (c) a diagnosis of the latter's pathological forms. Mechanical solidarity, held to typify preindustrial societies, based on resemblances, exhibits identical and juxtaposed elements, segmentally structured and with weak social bonds, low population volume and moral density, undifferentiated social organization, repressive or penal law, and collective beliefs and sentiments (which at that stage he called the *conscience collective*) that strongly tie the individual to the society as a whole, while organic solidarity, typical of industrialized societies (in their 'normal' state of functioning), based on the division of labor, the fusion of markets, and the growth of cities, exhibits interdependence and strong social bonds, high volume and density, an increasing differentiation of social functions, restitutive or cooperative law, the gradual withering away of the *conscience collective*, and the emancipation of the individual, allowing space for individual initiative and reflection. The transition from mechanical to organic solidarity, which constituted the 'progress of the division of labor and of civilization,' occurs, Durkheim maintained, 'mechanically' as a consequence of growth in the social mass and its social or dynamic

density, through expanding markets, a 'struggle for existence,' a resultant development of new occupations and of corresponding consumer demand. Durkheim intuitively assumed an equilibrium outcome of increasing division of labor but did not specify the parameters that would generate that outcome. The division of labor exhibits perverse effects and thus abnormal forms where the conditions for the social equilibrium of organic solidarity are not satisfied. *Anomie*—the lack of appropriate rules—is evident in industrial and commercial crises, veering between prosperity and depression, resulting from ignorance and uncertainty of producers concerning demand, in the antagonism between capital and labor, and also in increasingly fragmented scientific specialization, and results in general from a too rapid industrialization in which the appropriate rules have not had time to stabilize relations between social functions and roles. The 'forced division of labor' exists where inequality is seen as illegitimate, as when unequal chances of mobility violate the meritocratic principle. And a third abnormal form, which Besnard calls 'bureaucratic' (Besnard 1987), is found where excessive specialization is combined with weak productivity and poor mutual adaptation of functions.

Durkheim considerably deepened his notion of *anomie* and his diagnosis of capitalism's ills in *Suicide*. *Anomie*is one of the two major modern social causes of suicide: a condition of rulelessness in which individuals lose their moorings: their expectations exceed the possibilities of satisfying them, potentially insatiable desires are no longer regulated within a stable and legitimate normative framework of economic rewards or status ranking or conjugal sexual relations. The other major modern social condition causing suicide is egoism: a breakdown of integration, a lack of attachment of individuals to common values, to socially approved relationships, activities, and purposes, isolating and privatizing their lives. (Excessive regulation, or fatalism, and excessive integration, or altruism, are less prevalent, Durkheim thought, in modern societies.) Suicide is the extreme response of the small, randomly distributed but statistically stable minority of suicide-prone individuals whose social situations expose them to the social causes in question. He expounded this theory by accounting for differential suicide rates (e.g., higher among Protestants than Catholics, lower among married men, higher during economic crises, lower during wars) by characterizing suicide's social causes as 'suicidogenic currents,' thereby suggesting a macro–macro explanation that evaded the microlevel, in line with his anti-individualist methodological stance. In reality his theory of suicide is a macro–micro theory of the social conditions for individual psychological equilibrium, or health (*anomie*being the 'malady of infinite aspiration'), which that stance deflected him from elaborating. *Suicide* has been widely seen as remarkable for its sophisticated use of statistics (albeit questionable in abstracting from suicide's social meanings to which, likewise, Durkheim paid no attention), including the use of multivariate analysis to exclude alternative explanations. But above all, the work exhibits its author's typical boldness in extending sociological explanation to what might seem least amenable to it: the most private and individual of acts.

The same boldness is evident in *The Elementary Forms*,whose introduction and conclusion advance a theory of the fundamental categories of thought, while the rest of the work advances a theory of the origins, meaning and functions of religious beliefs and practices, of magic, and of the relation of both to science. All this he developed with reference to ethnographic evidence of totemism (assumed to be the most primitive form of religion and its evolutionary origin, or at least its earliest known form) as found in 'simpler,' tribal societies in Australia and America. He made the questionable assumption that the essential elements of religious thought and life should be found, at least in germ, in the most primitive religions, regarding the study of a 'very simple' religion as particularly instructive, since the relations between its constituent elements are 'more apparent' there.

Observing that such categories as space, time, number, causality, the person, etc. impose themselves on us inescapably as the framework of our understanding, Durkheim rejected as nonexplanatory Kant's account of them as given *a priori*, arguing that they express the most general relations existing between things which are revealed to us through social action and mutual communication: their necessity has a social basis. Moreover, they are fundamental to commonsense and to science (natural and social) and across cultures. Religion he defined by its universal dichotomizing of the world into the sacred and the profane, uniting believers around the sacred to which they accord a special, awed respect. Starting from the assumption that religious beliefs cannot be illusory (for otherwise they would not persist), he asked to what reality they correspond, answering that, once decoded, they express our understanding of our society and our obscure and intimate relations with it. They require decoding since, until social science, these cannot be apprehended directly: religion is, in this respect, a sort of mythological proto-sociology. (Thus, the idea of the soul is the way religion represents the duality of human nature, part corporeal, part social.)

This exclusively society-oriented interpretation of religious symbolism has been much criticized (it was Durkheim, not the savage, Evans-Pritchard suggested, who made society into a God). A further contentious issue is whether or not accepting it challenges religious belief (Durkheim claimed it did not). In the domain of knowledge science replaces religion but religion continues to symbolize and dramatize social relations, strengthening and re-animating social bonds and believers themselves through the collective enactment of rituals. This aspect of religion survives and would,

he speculated, remain basic to the political cohesion of modern, secular societies as civil religion. Religious beliefs themselves originate in such collective effervescences and are renewed by their re-enactment.

Cognitively, religions are, he held, cosmologies, offering a 'first explanation of the world,' seeking realities behind appearances, teaching people to unite what the senses separate, and thus prefiguring science. Religion and science are alike in that they both seek to render the (natural and social) world intelligible, but only the latter disposes of methods that subject its observations to systematic control. Likewise, magic, which occurs within profane or everyday life, draws on prevailing religious beliefs and seeks to predict and manipulate the world, but without benefit of scientific tests, statistics or experiments, adducing (as scientists also do) auxiliary hypotheses to account for its failures. So for Durkheim, humanity is unitary in its thinking, not, as Lucien Lévy-Bruhlbelieved, divided between prelogical primitives and logical, science-based moderns.

3. *Influence, Impact, and Current Significance*

His collaborators applied and developed Durkheim's ideas in many directions, from economics to Chinese thought (in the work of Marcel Granet) to primitive and ancient law (see Besnard 1983). They also had a considerable contemporary impact on neighboring disciplines in France, for instance, on historians, initially through Henri Berr and his *Revue de synthèse historique* and later through Lucien Febvre, Marc Bloch, and the *Annales* school, on human geography, beginning with Paul Vidal de la Blache, on jurists, notably Léon Duguit, and on the study of comparative mythology in the work of Georges Dumézil. The surviving Durkheimians continued into the 1930s but their research program gradually lost its dynamism and, as Raymond Aron observed, it came to seem increasingly irrelevant to the times. Among those who have made a significant impact on subsequent work, one can single out Robert Hertz's studies of death and of right and left (Hertz 1907, 1909), Maurice Halbwachs's work on collective memory (Halbwachs 1925) and, above all, Marcel Mauss, Durkheim's nephew, pupil, and collaborator, author of several major works on Durkheimian themes, including magic and sacrifice (co-authored with Henri Hubert), seasonal variations among the Eskimos (with Henri Beuchat), the techniques of the body and the category of the person (Mauss 1950), who succeeded in breaking down Durkheim's conceptual dichotomies and methodological rigidities vis-à-vis other disciplines, in the interest of studying the *fait social total*, as in his classic and influential study of *The Gift* (Mauss 1925). In several ways, Mauss rendered Durkheimian sociology more subtle, complex, and forward-looking. Through Mauss, it influenced French structuralism via his impact on Claude Lévi-Strauss (see his introduction to Mauss 1950) and also on Louis Dumont, author of the major work on the Indian caste system, *Homo Hierarchicus*and of parallel studies of Western individualism. Its influence is also evident in the *Revue de MAUSS* (*Mouvement anti-utilitariste dans les sciences sociales*) which since 1993 has published anti-utilitarian articles on various aspects of social life. Among French sociologists, George Friedmann, Pierre Bourdieu, and Raymond Boudon (see Boudon 1999) have incorporated Durkheimian ideas, albeit far from uncritically, the latter insisting upon a convincing interactionist reading of Durkheim. This has, in turn, been developed by Mohamed Cherkaoui (Cherkaoui 1998) who has, moreover, resurrected for close attention Durkheim's sociology of education, in particular his remarkable study of the evolution of French educational thought and institutions, and their interaction, since the early Church (Durkheim 1938), which explores various themes less than evident in his better known works: class power and conflict and the struggle for cultural domination, the role of intellectuals and the sociology of educational knowledge focusing on dominant epistemologies and disciplinary hierarchies. Another important work is the definitive study of the concept of anomie and its career within sociology (Besnard 1987). Besnard also inaugurated the valuable *Études durkheimiennes* in Paris in 1977 as a forum for scholarship on Durkheim and the Durkheimians. It moved to Urbana, Illinois under the editorship of Robert Alun Jones in 1989 and to Oxford in 1995, published by the British Centre for Durkheimian Studies along with a continuing series of important Occasional Papers.

The Durkheim school's impact on British social anthropology has been immense, initially through Radcliffe-Brown but chiefly through his successor E. E. Evans-Pritchard, whom it helped liberate from his predecessor's structural-functionalism. Despite serious reservations, he identified himself with the work of the Durkheimians, praising it as an indispensable theoretical capital, embodying rigorous methods of scholarship (see his introduction to Hertz 1907). In the early 1950s he launched an extensive series of English translations of works by Durkheim and his collaborators, and himself drew on Durkheimian ideas, especially in his work on the Nuer. The impact of Durkheimian ideas is also evident in the work of other Oxford anthropologists, particularly that of Rodney Needham and Nick Allen on symbolic classification, but also in the various writings of Mary Douglas on pollution and symbolic boundaries, and in the work of Robin Horton on African religions.

Durkheim's reception into American sociology began early, influencing, among others, Elton Mayo and W. Lloyd Warner, but it was Talcott Parsons's extended treatment of his work in his *Structure of Social Action* that marked his canonization as a 'grand

theorist,' treated there as a pre-Parsonian proto-structural functionalist, with a distinctive solution to the 'problem of order' through the internalization and institutionalization of consensual, society-wide values.

Substantively, the impact of his ideas was pervasive, influencing innumerable accounts of industrialization, urbanization, social control, social disorganization, and collective behavior. These were, directly or indirectly, extrapolated from Durkheim's picture of advancing social differentiation leading through rapid social change to the breakdown of social ties. According to such accounts, rapid industrialization generates disharmony between life experiences and the normative framework which regulates them, leading to virtually unlimited possibilities for group conflict and thus to a variety of social movements: peaceful agitation, political violence, millennarianism, nationalism, revolution, underground subversion, etc., that in turn attract those suffering most severely under the displacements created by structural change. Examples of this approach are found in the work of Samuel Huntington, Chalmers Johnson, Ted Gurr, and Neil Smelser. Such accounts were criticized by Tilly (1981) for relying on three basic 'Durkheimian' arguments that are held to have been refuted by the historical evidence: (a) that conflict varies directly with the prevalence of differentiation within organization and the scale of groups; (b) that social disorder correlates with accelerated urbanization or industrial growth; and (c) that the various forms of 'disorder,' such as crime, suicide, conflict, and protest, vary together, in the same settings and circumstances. In short, Tilly rejects the 'Durkheimian line' on conflict, protest, and collective action, which he explains instead in terms of group solidarities, common interests, and mobilization. But, first, the 'line' is extrapolated from, not argued by, Durkheim, who never sought to explain so much by so little; it is his American successors, not Durkheim, who posited the strains resulting from rapid industrialization as the catch-all source of so many 'disorders,' which he never treated as equivalent and associated. Second, his ideas about the processes creating and re-animating collective solidarity are surely not irrelevant to the explanation of collective action. And third, his focus on the consequences of rapid economic and social change outstripping the norms that could regulate it has an inescapable relevance for the diagnosis of the distinctive ills of the most advanced forms of contemporary capitalism, in particular the corrosive effects of technical change on social relations and individual lives.

The concept of *anomie* has had an impact in other fields. Robert Merton imported it in his influential paper on *Social Structure and Anomie* (Merton 1949) but in the process mis- or re-interpreted Durkheim's notion as a maladaptation of means to ends, rather than a sociopsychological condition of deregulation (so that the problem, on Durkheim's account, should be seen as lying in the success ethic itself, not in the socially structured inability of some to realize its promises). But this idea, together with the other causes of suicide, and indeed the structure, content and mode of its attempted explanation, have exercised a continuing and productive stimulus within the sociology of deviance and criminology, as often through attempts to develop or revise Durkheim's contribution as through more or less radical refutations of his whole approach.

It is, however, his later 'religious sociology' that has had most impact within sociology during the1990s, applying its concepts, insights, and hypotheses to historical and modern societies. Durkheim and the Durkheimians never did this, and indeed a surprised Mauss could write in the late 1930s, contemplating Nazism and Fascism, of the tragic 'verification' of their theory in the spectacle of large modern societies 'hypnotized like Australians are by their dances' (Ranulf 1939, cited in Lukes 1972). From the 1960s through the 1980s, practitioners of historical sociology were largely focused on the rise and development of capitalism and were class-oriented and state-centered in approach: Marx, Weber, and Tocqueville were the presiding ancestors. Since then a growing interest in cultural analysis has been evident, with an associated interest in Durkheimian ideas. Initially, this development had a Parsonian, consensual emphasis. Edward Shils argued that secular, differentiated societies have sacred 'centers' that explain the allocation of status (Shils 1975), Clifford Geertz interpreted cultural systems in a way that linked sacredness, ritual and the mobilization of solidarity (Geertz 1973), and Robert Bellah published and stimulated studies of civil religion under modern conditions (Bellah and Hammond 1980). More recently, there has been a growing interest in symbolism and ritual and their relation to the institutionalization of power and social control but without the assumption of social consensus. In Jeffrey Alexander's words, a 'late Durkheimian framework' is applied to 'analyze phenomena which are disruptive, contingent, and related to social change,' while Randall Collins, exponent of conflict sociology, has sought to combine Durkheim with Erving Goffman, in exploring how 'focused and emotionally arousing interaction produces group solidarity' (Alexander 1988). Studies along such 'late Durkheimian' lines have included French Revolutionary festivals, ritual and symbolism under Fascism, nationalist movements, mass strikes, and media events.

These are the main lines of Durkheim's impact upon twentieth-century social science. I have not discussed his impact beyond the social sciences (for example, on and through such partly literary figures as Roger Caillois, Michel Leiris, and George Bataille, members of the so-called *Collège de Sociologie* in the late 1930s), or countless unacknowledged influences. More specific and acknowledged social scientific influences could also be cited, for instance, upon the work of Basil

Bernstein on sociolinguistic and educational 'codes,' drawing on Durkheim's ideal-typical polarities, and upon the so-called 'strong program' in the sociology of science, originating in the University of Edinburgh, focusing, in a relativizing vein, on the 'social' origins of 'natural' classifications (see Bloor 1976). Here, as ever, Durkheim's ideas are interpreted in accordance with locally generated concerns. The story of their reception, sketched here, confirms both their continuing attraction and their continuing power to generate alternative and invariably controversial readings.

See also: Altruism and Prosocial Behavior, Sociology of; Altruism and Self-interest; Collective Behavior, Sociology of; Conformity: Sociological Aspects; Determinism: Social and Economic; Differentiation: Social; Disciplines, History of, in the Social Sciences; Evolutionism, Including Social Darwinism; Functionalism, History of; Functionalism in Sociology; Individual/Society: History of the Concept; Industrial Society/Post-industrial Society: History of the Concept; Integration: Social; Labor, Division of; Law: Change and Evolution; Mauss, Marcel (1872–1950); Norms; Positivism: Sociological; Religion: Evolution and Development; Religion, Sociology of; Social Change: Types; Sociology, Epistemology of; Sociology, History of; Solidarity: History of the Concept; Solidarity, Sociology of; Suicide, Sociology of; Values, Sociology of

Bibliography

Alexander J C (ed.) 1988 *Durkheimian Sociology: Cultural Studies*. Cambridge University Press, Cambridge, UK

Bellah R N, Hammond P (eds.) 1980 *Varieties of Civil Religion*. Harper and Row, San Francisco

Besnard P (ed.) 1983 *The Sociological Domain. The Durkheimians and the Founding of French Sociology*. Cambridge University Press, Cambridge, UK

Besnard P 1987 *L'Anomie, ses usages et ses fonctions dans la discipline sociologique depuis Durkheim*. Presses Universitaires de France, Paris

Bloor D 1976 *Knowledge and Social Imagery*. Routledge, London

Boudon R 1999 *Les Formes élémentaires de la vie religieuse:* une theorie toujours vivante. *L'Année sociologique* **49**(1): 149–98

Cherkaoui M 1998 *Naissance d'une science sociale. La sociologie selon Durkheim*. Droz, Geneva, Switzerland

Cotterrell R 1999 *Émile Durkheim. Law in a Moral Domain*. Edinburgh University Press, Edinburgh, UK

Durkheim E 1893 *De la division du travail social: étude sur l'organisations des sociétés supérieures*. Alcan, Paris (1984 *The Division of Labor in Society*. Free Press, New York and Macmillan, London)

Durkheim E 1895 *Les Règles de la méthode sociologique*. Alcan, Paris (1982 *The Rules of Sociological Method*. Macmillan, London and New York)

Durkheim E 1897 *Le Suicide: étude de sociologie*. Alcan, Paris (1951 *Suicide*. Free Press, Glencoe, II and 1952. Routledge and Kegan Paul, London

Durkheim E 1912 *Les Formes élémentaires de la vie religieuse: le système totemique en Australie* (1997 *The Elementary Forms of the Religious Life*. Free Press, New York)

Durkheim E 1924 *Sociologie et philosophie*. Alcan, Paris (1953 *Sociology and Philosophy*. Cohen and West, London and Free Press, Glencoe, IL)

Durkheim E 1938 *L'Evolution pédagogique en France*. Alcan, Paris (1977 *The Evolution of Educational Thought: Lectures on the Formation and Development of Secondary Education in France*. Routledge and Kegan Paul, London and Boston)

Durkheim E 1950 *Leçons de sociologie: physique des moeurs et du droit* (1957 *Professional Ethics and Civic Morals*. Routledge & Kegan Paul, London; 1958, Free Press, New York)

Durkheim E, Mauss M 1903 De quelques formes primitives de classification. *L'Année sociologique*. **6**: 1–72 (1963 *Primitive Classification*. Cohen and West, London and University of Chicago Press, Chicago)

Geertz G 1973 *The Interpretation of Cultures*. Basic Books, New York

Halbwachs M 1925 *Les Cadres sociaux de la mémoire*. Alcan, Paris (1992 *On Collective Memory*. University of Chicago Press, Chicago

Hertz R 1907 Contribution à une étude sur la représentation de la mort. *L'Année sociologique* **10**: 48–137 (1960 *Death and The Right Hand*. Cohen and West, London)

Hertz R 1909 La prééminence de la main droite. *Revue Philosophique* **68**: 553–80 (1960 *Death and the Right Hand*. Cohen and West, London)

Jones R A 1999 *The Development of Durkheim's Social Realism*. Cambridge University Press, Cambridge, UK

Lukes S 1972 *Émile Durkheim: His Life and Work*. Harper and Row. New York and 1973 Allen Lane The Penguin Press, London

Mauss M 1925 Essai sur le don. *L'Année sociologique* **1**: 30–186 (republished in Mauss M 1950/1990 *The Gift*. Routledge, London

Mauss M 1950 *Sociologie et anthropologie*. Presses Universitaires de France, Paris (4th enlarged edn. 1968)

Merton R K 1949 *Social Structure and Anomie* (1968 Merton R K *Social Theory and Social Structure*. Free Press, New York, pp. 185–214)

Pickering W 1984 *Durkheim's Sociology of Religion. Themes and Theories*. Routledge & Kegan Paul, London

Ranulf S 1939 Scholarly forerunners of Fascism. *Ethics* **50**: 16–34

Shils E 1975 *Center and Periphery: Essays in Macrosociology*. Chicago University Press, Chicago

Tilly C 1981 Useless Durkheim. In: Tilly C (ed.) *As Sociology Meets History*. Academic Press, New York, pp. 95–101

S. Lukes

Dynamic Decision Making

Dynamic decision making is defined by three common characteristics: a series of actions must be taken over time to achieve some overall goal; the actions are interdependent so that later decisions depend on earlier actions; and the environment changes both spontaneously and as a consequence of earlier actions

(Edwards 1962). *Dynamic decision tasks* differ from *sequential decision tasks* (see *Sequential Decision Making*) in that the former are primarily concerned with the control of dynamic systems over time to achieve some goal, whereas the latter are more concerned with sequential search for information to be used in decision making. (Dynamic decision tasks also need to be distinguished from dynamic decision processes. Dynamic decision models (e.g., Busemeyer and Townsend 1993), describe the moment-to-moment deliberation processes that precede each action within the sequence of actions.)

Psychological research on dynamic decision making began with Toda's (1962) pioneering study of human performance with a game called the 'fungus eater,' in which human subjects controlled a robot's search for uranium and fuel on a hypothetical planet. Subsequently, human performance has been examined across a wide variety of dynamic decision tasks including computer games designed to simulate stock purchases (Rapoport 1975), welfare management (Dorner 1980, Mackinnon and Wearing 1980), vehicle navigation (Jagacinski and Miller 1978, Anzai 1984), health management (Kleinmuntz and Thomas 1987, Kerstholt and Raaijmakers 1997), production and inventory control (Berry and Broadbent 1988, Sterman 1994), supervisory control (Kirlik et al. 1993), and fire-fighting (Brehmer 1992). Cumulative progress in this field has been summarized in a series of empirical reviews by Rapoport (1975), Funke (1991), Brehmer (1992), Sterman (1994), and Kerstholt and Raaijmakers (1997).

1. Stochastic Optimal Control Theory

To illustrate how psychologists study human performance on dynamic decision tasks, consider the following experiment by You (1989). Subjects were initially presented with a 'cover' story describing the task: 'Imagine that you are being trained as a psychiatrist, and your job is to treat patients using a psychoactive drug to maintain their health at some ideal level.' Subjects were instructed to choose the drug level for each day of a simulated patient after viewing all of a patient's previous records (treatments and health states). Subjects were trained on 20 simulated patients, with 14 days per patient, all controlled by a computer simulation program.

There are a few general points to make about this type of task. First, laboratory tasks such as this are oversimplifications of real-life tasks, designed for experimental control and theoretical tractability. However, more complex simulations have also been studied to provide greater realism (e.g., Brehmer's (1992) fire-fighting task). Second, the above task is an example of a discrete-time task (only the sequence of events is important), but real-time simulations have also been examined, where the timing of decisions

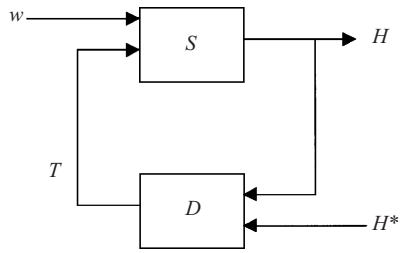

Figure 1
Feedback diagram for a typical dynamic decision-making task. S represents the linear stochastic dynamic system, D represents the decision policy, w is the disturbance factor, T is the control treatment, H is the output state of health, and H^* is the desired state of health

becomes critical (e.g., fire-fighting). Third, the cover story (e.g., health management) provides important prior knowledge for solving the task, and so the findings depend on both the abstract task properties and the concrete task details (Kleiter 1975). Fourth, the stimulus events are no longer under complete control of the experimenter, but instead they are also influenced by the subject's own behavior. Thus experimenters need to switch from a stimulus–response paradigm toward a cybernetic paradigm for designing research (Rapoport 1975, Brehmer 1992).

This health management task can be formalized by defining $H(t)$ as the state of the patient's health on day t; $T(t)$ is the drug treatment presented on day t; and $w(t)$ is a random shock that may disturb the patient on any given day. Figure 1 is a feedback diagram that illustrates this dynamic decision task. S represents the dynamic environmental system that takes both the disturbance, w, and the decision maker's control action, T, as inputs, and produces the patient's state of health, H, as output. D represents the decision-maker's policy that takes both the observed, H, and desired, H^*, states of health as input, and produces the control action, T, as output.

Based on these definitions, this task can be analyzed as a stochastic linear optimal control problem (Bertsekas 1976): determine treatments $T(1), \ldots, T(N)$, for $N = 14$ days, that minimize the objective function

$$F = E\{\Sigma_{t=1,N}\, \alpha[H(t) - H^*(t)]^2 + \beta T^2(t)\} \qquad (1)$$

contingent upon the environmentally programmed linear stochastic dynamic system,

$$H(t+1) = a_1 H(t) + a_2 H(t-1) + b_1 T(t) + b_2 T(t-1) + b_3 T(t-2) + w(t) \qquad (2)$$

Standard dynamic programming methods (Bertsekas 1976) may be used to find the optimal solution to this problem. For the special case where the desired state

of health is neutral ($H^* = 0$), some treatment effect takes place the very next day ($b_1 0$), and there is no cost associated with the treatments (i.e., $\beta = 0$), then the optimal policy

$$T(t) = -[(a_1/b_1)H(t) + (a_2/b_1)H(t-1) \quad (b_2/b_1)T(t-1) + (b_3/b_1)T(t-2)] \quad (3)$$

is the treatment that forces the mean health state to equal the ideal (zero) on the next day. If the cost of treatment is nonzero (i.e., $\beta > 0$), then the solution is a more complex linear function of the previous health states and past treatments (Bertsekas 1976, You 1989).

Dynamic programming is a general-purpose method that can be used to solve for optimal solutions to many dynamic decision tasks. Although the example above employed a linear control task, dynamic programming can also been used to solve many nonlinear control problems (Bertsekas 1976). However, for highly complex tasks, dynamic programming may not be practical, and heuristic search methods such as genetic algorithms (Holland 1994) may be more useful (see also *Artificial Intelligence: Genetic Programming*).

The formal task analysis presented above provides a basis for determining factors that may affect human performance on the task (Brehmer 1992). One factor is the stability of the dynamic system, which for Eqn. (2), depends on the two coefficients, a_1 and a_2. In particular, this system is stable if the roots of the characteristic equation

$$\lambda^2 - a_1\lambda - a_2 = 0 \quad (4)$$

are < 1 in magnitude (Luenberger 1979). A second factor is the controllability of the system, which depends on the three coefficients, b_1, b_2, and b_3 (Luenberger 1979). For example, if the treatment effect is delayed ($b_1 = 0$), then the simple control policy shown in Eqn. (3) is no longer feasible, and the optimal policy is more a complex linear function of the system coefficients.

Consistent with previous research, You (1989) found that even after extensive experience with the task, subjects frequently lost control of their patients, and the average performance of human subjects fell far below optimal performance. However, this is a gross understatement. Sterman (1994) found that when subjects tried to manage a simulated production task, they produced costs 10 times greater than optimal, and their decisions induced costly cycles even though the consumer demand was constant. Brehmer (1992) found that when subjects tried to fight simulated forest fires, they frequently allowed their headquarters to burn down despite desperate efforts to put the fire out. Kleinmuntz and Thomas (1987) found that when subjects tried to manage their simulated patients' health, they often let their patients die while wasting time waiting for the results of nondiagnostic tests and performed more poorly than a random benchmark.

2. *Alternative Explanations for Human Performance*

There are many alternative reasons for this suboptimal performance. By their very nature, dynamic decision tasks entail the coordination of many tightly interrelated psychological processes including causal learning, planning, problem solving, and decision making (Toda 1962). Six different psychological approaches for understanding human dynamic decision-making behavior have been proposed, each focusing on one of the component processes.

The first approach was proposed by Rapoport (1975), who suggested that suboptimal performance could be derived from an optimal model either by adding information processing constraints on the planning process, or by including subjective utilities into the objective function of Eqn. (1). For example, Rapoport (1975) reported that human performance in his stock purchasing tasks was reproduced accurately by assuming that subjects could only plan a few steps ahead (about three steps), as compared with the optimal model with an unlimited planning horizon. In this case, dynamic programming was useful for providing insights about the constraints on human planning capabilities. As another example, Rapoport (1975) reported attempts to predict performance in a multistage investment game by assuming that the utility function was a concave function of monetary value. In this case, dynamic programming was used to derive an elegant decision rule, which predicted that investments should be independent of the size of the capital accumulated during play. Contrary to this prediction, human decision makers were strongly influenced by the capital that they accumulated, and so in this case, dynamic programming was useful for revealing qualitative empirical flaws with the theory.

An alternative approach, proposed by Brehmer (1992) and Sterman (1994), is that suboptimal performance is caused by a misconception of the dynamic system (Eqn. (2)). In other words, a subject's internal model of the system does not match the true model. In particular, human performers seem to have great difficulty in discerning the influence of delayed feedback and understanding the effects of nonlinear terms in the system. Essentially, subjects solve the problem as if it was a linear system that has only zero lag terms. In the case of Eqn. (2), this yields the simplified policy

$$T(t) = -c_1 H(t) \quad (5)$$

where c_1 is estimated from a subject's control decisions. Diehl and Sterman (1993) found that this type of

simplified subjective policy described their subjects' behavior very accurately.

A more general method for estimating subjective decision policies was proposed by Jagacinski and Miller (1978). Consider once again the example problem employed by You (1989). In this case, a subject's treatment decision on trial $T(t)$ could be represented by a linear control model:

$$T(t) = c_0 + c_1 H(t) + c_2 H(t-1) + c_3 H(t-2) + c_4 T(t) + c_5 T(t-1) + c_6 T(t-2) + \text{error} \quad (6)$$

where the subjective coefficients $(c_0, c_1, \ldots, c_6)$ are estimated by a multiple regression analysis. This is virtually the same as performing a 'lens model' analysis to reveal the decision-maker's policy (Slovic and Lichtenstein 1971, Kleinmuntz 1993). This approach has been applied successfully in numerous applications (Jagacinski and Miller 1978, Kleinman et al. 1980, You 1989, Kirlik et al. 1993). Indeed, You (1989) found that subjects made use of information from both lags 1 and 2 for making their treatment decisions.

Heuristic approaches to strategy selection in dynamic decision tasks have been explored by Kleinmuntz and Thomas (1987), and Kerstholt and Raaijmakers (1997). These researchers examined health management tasks that entailed dividing time between two strategies: collecting information from diagnostic tests before treatment vs. treating patients immediately without performing any more diagnostic tests. A general finding is that even experienced subjects tended to overuse information collection, resulting in poorer performance than that which could be obtained from an immediate treatment (no test) strategy. This finding seems to run counter to the adaptive decision-making hypothesis of Payne et al. (1993), which claims that subjects prefer to minimize effort and maximize performance. The information collection strategy is both more effortful and less effective than the treatment strategy in this situation.

Finally, an individual difference approach to understanding performance on complex dynamic decision tasks was developed by Dorner and his colleagues (see Funke (1991) for a review). Subjects are divided into two groups (good vs. poor) on the basis of their performance on a complex dynamic decision task. Subsequently, these groups are compared with respect to various behaviors in order to identify the critical determinants of performance. This research indicates that subjects who perform best are those who set integrative goals, collect systematic information, and evaluate progress toward these goals. Subjects who tend to shift from one specific goal to another, or focus exclusively on only one specific goal, perform more poorly.

3. *Learning to Control Dynamic Systems*

Although human performers remain suboptimal even after extensive task training, almost all past studies reveal systematic learning effects. First, overall performance improves rapidly with training (see, e.g., Rapoport 1975, Mackinnon and Wearing 1985, Brehmer 1992, Sterman 1994). Furthermore, subjective policies tend to evolve over trial blocks toward the optimal policy (Jagacinski and Miller 1978, You 1989). Therefore, learning processes are important for explaining much of the variance in human performance on dynamic decision tasks (Hogarth 1981). Three different frameworks for modeling human learning processes in dynamic decision tasks have been proposed.

A *production rule* model was developed by Anzai (1984) to describe how humans learn to navigate a simulated ship. The general idea is that past and current states of the problem are stored in working memory. Rules for transformation of the physical and mental states are represented as condition–action-type production rules. A product rule fires whenever the current state of working memory matches the conditions for a rule. When a production rule fires, it deposits new inferences or information in memory, and it may also produce physical changes in the state of the system. Navigation is achieved by a series of such recognize–act cycles of the production system. Learning occurs by creating new production rules that prevent earlier erroneous actions, with new rules being formed on the basis of means–ends type of analyses for generating subgoal strategies. Simulation results indicated that the model eventually learned strategies similar to those produced in the verbal protocols of subjects, although evaluation of the model was based on qualitative judgments rather than quantitative measurements.

An *instance or exemplar*-based model was later developed by Dienes and Fahey (1995) to describe how humans learn to control a simulated sugar production task, and to describe how they learn to manage the emotional state of a hypothetical person. This model assumes that whenever an action leads to a successful outcome, then the preceding situation and the successful response are stored together in memory. On any given trial, stored instances are retrieved on the basis of similarity to the current situation, and the associated response is applied to the current situation. This model was compared to a simple rule-based model like that employed by Anzai (1984). The results indicated that the exemplar learning model produced more accurate predictions for delayed feedback systems, but the rule-based model performed better when no feedback delays were involved. This conclusion agrees with earlier ideas presented by Berry and Broadbent (1988), that delayed feedback tasks involve implicit learning processes, whereas tasks without delay are based on explicit learning processes.

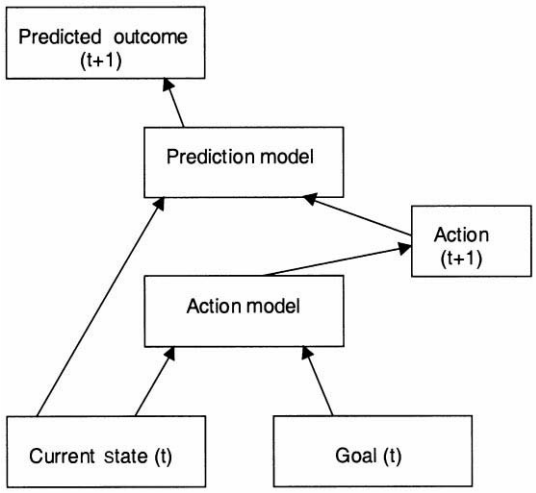

Figure 2
An artificial neural network for learning to control a dynamic system. The current state and the goal feed into a layer of hidden units that learn to produce an action. Next the current state and the action feed into another layer of hidden units that learn to predict the consequence

An *artificial neural network* model was developed by Gibson et al. (1997) to describe learning in a sugar production task. Figure 2 illustrates the basic idea. At the bottom of the figure are two types of input nodes, one representing the current state of the environment, and the other representing the current goal for the task. These inputs feed into the next layer of hidden nodes that compute the next action, given the current state and goal. The action and the current state then feed into another layer of hidden nodes, which is used to predict the consequence of the action, given the current state. The connections from the current state and action to the prediction hidden layer are learned by back-propagating prediction errors, and the connections from the current state and current goal to the action hidden layer are learned by back-propagating deviations between the observed outcome and the goal state (holding the connections to the prediction layer constant; see Jordan and Rumelhart (1992) for more details.) This learning model provided good accounts of subjects' performance during both training and subsequent generalization tests under novel conditions. Unfortunately, no direct comparisons with rule- or exemplar-based learning models were conducted, and this remains a challenge for future research.

Another type of artificial neural network model for learning dynamic systems is the reinforcement learning model (see Miller et al. (1991) for a review). Although this approach has proven to be fairly successful for robotic applications, it has not yet been empirically tested against human performance in dynamic decision tasks.

4. *Laboratory vs. Naturalistic Decision Research*

Klein and his associates (Klein 1998) have made progress toward understanding dynamic decisions in applied field research settings, which they call *naturalistic decision making* (e.g., interview fire chiefs after fighting a fire). Field research complements laboratory research in two ways: on the one hand, it provides a reality check on the practical importance of theory developed in the laboratory; on the other, it provides new insights that can be tested with more control in the laboratory.

The general findings drawn from naturalistic decision research provide converging support for the general theoretical conclusions obtained from the laboratory. First, it is something of a misnomer to label this kind of research 'decision making,' because decision processes comprise only one of the many cognitive processes engaged by these tasks—learning, planning, and problem solving are just as important. Decision making is used to define the overall goal, but then the sequence of actions follows a plan that has either been learned in the past or generated by a problem-solving process. Second, learning processes may explain much of the variance in human performance on dynamic decision tasks. Decision makers use the current goal and current state of the environment to retrieve actions that have worked under similar circumstances in the past. Klein's (1998) recognition-primed decision model is based on this principle, and this same basic idea underlies the production rule, exemplar, and neural network learning models. Third, learning from extensive experience is the key to understanding why naïve subjects fail in the laboratory whereas experts succeed in the field—it is the mechanism that moves performance from one end of the continuum to the other.

See also: Bounded and Costly Rationality; Decision Making, Psychology of; Markov Decision Processes; Sequential Decision Making; Stochastic Models

Bibliography

Anzai Y 1984 Cognitive control of real-time event driven systems. *Cognitive Science* **8**: 221–54

Berry D C, Broadbent D E 1988 Interactive tasks and the implicit–explicit distinction. *British Journal of Psychology* **79**: 251–72

Bertsekas D P 1976 *Dynamic Programming and Stochastic Control*. Academic Press, New York

Brehmer B 1992 Dynamic decision making: Human control of complex systems. *Acta Psychologica* **81**: 211–41
Busemeyer J R, Townsend J T 1993 Decision field theory: A dynamic–cognitive approach to decision-making in an uncertain environment. *Psychological Review* **100**: 432–59
Diehl E, Sterman J D 1993 Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes* **62**: 198–215
Dienes Z, Fahey R 1995 Role of specific instances in controlling a dynamic system. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 848–62
Dorner D 1980 On the problems people have in dealing with complexity. *Simulation and Games* **11**: 87–106
Edwards W 1962 Dynamic decision theory and probabilistic information processing. *Human Factors* **4**: 59–73
Funke J 1991 Solving complex problems: Exploration and control of complex systems. In: Sternberg R J, Frensch P A (eds.) *Complex Problem Solving: Principles and Mechanisms*. Lawrence Erlbaum Associates, Hillside, NJ, pp. 185–222
Gibson F, Fichman M, Plaut D C 1997 Learning in dynamic decision tasks: Computational model and empirical evidence. *Organizational Behavior and Human Performance* **71**: 1–35
Hogarth R M 1981 Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin* **90**: 197–217
Holland J 1994 *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA
Jagacinski R J, Miller R A 1978 Describing the human operator's internal model of a dynamic system. *Human Factors* **20**: 425–33
Jordan M I, Rumelhart D E 1992 Forward models: Supervised learning with a distal teacher. *Cognitive Science* **16**: 307–54
Kerstholt J H, Raaijmakers J G W 1997 Decision making in dynamic task environments. In: Crozier W R, Svenson O (eds.) *Decision Making: Cognitive Models and Explanations*. Routledge, London, pp. 205–17
Kirlik A, Miller R A, Jagacinski R J 1993 Supervisory control in a dynamic and uncertain environment: A process model of skilled human–environment interaction. *IEEE Transactions on Systems, Man, and Cybernetics* **23**: 929–52
Klein G 1998 *Sources of Power: How People Make Decisions*. MIT Press, Cambridge, MA
Kleinman D L, Pattipati K R, Ephrath A R 1980 Quantifying an internal model of target motion in a manual tracking task. *IEEE Transactions on Systems, Man, and Cybernetics* **10**: 624–36
Kleinmuntz D 1993 Information processing and misperceptions of the implications of feedback in dynamic decision making. *System Dynamics Review* **9**: 223–37
Kleinmuntz D, Thomas J 1987 The value of action and inference in dynamic decision making. *Organizational Behavior and Human Decision Processes* **39**: 341–64
Kleiter G D 1975 Dynamic decision behavior: Comments on Rapoport's paper. In: Wendt D, Vlek C (eds.) *Utility, Probability, and Human Decision Making*. Reidel, Dordrecht, The Netherlands, pp. 371–80
Luenberger D G 1979 *Introduction to Dynamic Systems*. Wiley, New York
Mackinnon A J, Wearing A J 1980 Complexity and decision making. *Behavioral Science* **25**: 285–96
Mackinnon A J, Wearing A J 1985 Systems analysis and dynamic decision making. *Acta Psychologica* **58**: 159–72
Miller W T, Sutton R S, Werbos P J 1991 *Neural Networks for Control*. MIT Press, Cambridge, MA
Payne J W, Bettman J R, Johnson E J 1993 *The Adaptive Decision Maker*. Cambridge University Press, New York
Rapoport A 1975 Research paradigms for studying dynamic decision behavior. In: Wendt D, Vlek C (eds.) *Utility, Probability, and Human Decision Making*. Reidel, Dordrecht, The Netherlands, pp. 347–69
Slovic P, Lichtenstein S 1971 Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance* **6**: 649–744
Sterman J D 1989 Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes* **43**: 301–35
Sterman J D 1994 Learning in and about complex systems. *System Dynamics Review* **10**: 291–330
Toda M 1962 The design of the fungus eater: A model of human behavior in an unsophisticated environment. *Behavioral Science* **7**: 164–83
You G 1989 Disclosing the decision-maker's internal model and control policy in a dynamic decision task using a system control paradigm. MA thesis, Purdue University

J. R. Busemeyer

Dynamic Mapping in Geography

Dynamic mapping summarizes various cartographic presentation methods which incorporate the dimension of time into a map. These methods can be differentiated into those used either for static presentations or for dynamic presentations. All presentation methods have their own individual characteristics and are suitable for specific applications (Fig. 1).

1. *The Term and Concept of Dynamic Mapping*

Dynamic mapping is a cartographic concept used to depict dynamic spatial phenomena or to present spatial information in a dynamic way. It is involved with the depiction of spatial processes on maps. Our world is in permanent change. If the changes are of less interest because of their relative unimportance, it is sufficient to present a specific state of our world and of the corresponding date. Yet, if the occurring changes are of great importance, the dynamic aspect of reality—the dimension of time—has to be depicted in maps. Maps, however, are static: they can depict states but they can hardly show dynamic processes. To overcome the static nature of maps, different presentation methods were developed to give an impression of the changes and movements in space (Monmonier 1990, Muehrcke 1992). Recently the concept of dynamic mapping has been expanded by computer technology. Not only is the presentation of spatial changes now a matter of interest, but so too is

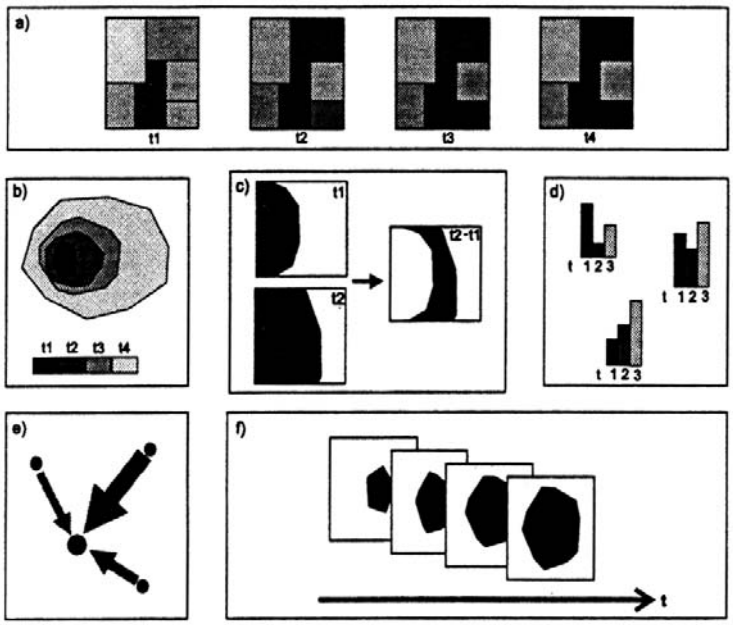

Figure 1
Methods of dynamic mapping: (a) time-series maps, (b) time-composite map, (c) change map, (d) time-related chart map, (e) flow map, (f) cartographic film or animation

dynamic presentation of spatial data, such as the successive display of map objects.

2. *Methods of Dynamic Mapping in Static Maps*

One presentation method for depicting spatial processes are time-series maps. They show a particular area and subject at different points in time in a series of maps. The user has to compare the individual maps and has to understand the changes which have occurred. Time-series maps have to be arranged in a way that enables visual comparison. They have to be placed next to each other, and their content should not be too complex. Map series are applied to show the distribution of particular phenomena at certain given intervals, e.g., changes of population distribution, boundaries, or landforms, etc.

Another method is the time-composite map. Such a map depicts two or more stages of the development of a particular phenomenon in one single map. The distinct stages can be differentiated by color, pattern, isolines, or point symbols. Time-composite maps are suitable for showing the expansion or contraction of areas, linear networks, or other phenomena, e.g., the growth of a city, the development of transportation systems, or the spread of diseases. Time-composite maps succeed in presenting one-directional changes, but they fail if retrogressive developments have to be depicted as an intermediate stage.

Map series and time-composite maps show only selected stages of the entire development. The change itself has to be understood by interpretation. However, some geographical questions emphasize, above all, the change and its extent. These questions require a map which shows explicitly the change that has taken place within a certain time period. For that purpose the change map was developed. Qualitative change maps depict objects which have been added to or removed from a particular area during a relevant time period, e.g., the difference in forest cover in different years. Quantitative change maps present changes in magnitude, the change in a value such as population density or air pollution. They either show the direction of change as increase or decrease, or give the absolute values of magnitude change.

A further method of presenting changes directly is a map with time-related charts which depicts developing phenomena of areas or locations in line or bar graphs, e.g., maps with diagrams of the annual temperature curve for selected stations. This type of map is able to give a detailed impression of development in time but it has its limitations in depicting the spatial distribution of a changing phenomenon.

Geographical processes may take place at a particular location (similar to the mentioned changes of population or temperature) or they may be a sort of movement in space like the movement of people, goods, or winds. To show these spatial movements, vector or flow maps with special symbols such as

arrows or lines are applied. The position and alignment of the arrows and lines can depict the moving objects' routes and directions, and the size of the signatures can denote the objects' values. Typical examples of vector or flow maps are maps of expedition routes, commuter movements, or ocean currents.

3. Cartographic Film and Animation as Dynamic Maps

All of the presentation methods referred to above are able to produce an idea regarding the changes that take place, but they have their limitations: they fail to depict the dynamics of the process. They are snapshots and the changes in the inter-snapshot periods can only be inferred.

For this reason, cartographic films were made. Films are dynamic presentation forms which are not restricted to single stages and can therefore give an impression of the complete process. Cartographic films are created like animated cartoons. They are sequences of slightly varying maps that give the illusion of continuous spatial movement or change in the case of fast presentation (24–30 pictures per second). Early cartographic films had to be made by hand; each map of the sequence was drawn by a cartographer. They were produced for military propaganda to depict frontier changes during World War II, or as instruction material for students. Despite their advantages, cartographic films did not become a widespread presentation form because of their time-consuming and expensive creation.

The cartographic film benefited from computer animation (Thrower 1961). Computer animations are film sequences completely generated by computer. Only some initial maps of the sequence are arranged by a cartographer, the intermediate maps are constructed automatically by the computer. Computer animations have two advantages compared to traditional films. First, computers take over the time-wasting map-making process of traditional films and allow quicker and less expensive production of cartographic animations; second, computer animations can be controlled interactively by the user. At a basic level, interaction means the start, stop, and interruption of an animation. At a higher level, more sophisticated interactions are possible, such as the selection of particular time intervals or the definition of different viewpoints (Peterson 1994).

Because of the advantages of speedy production and interaction, computer animation brought great potential to cartographic visualization and expanded the concept of dynamic mapping. Computer animation is no longer used exclusively for the depiction of spatial dynamic phenomena, it is also employed for any dynamic presentation of spatial data (DiBiase et al. 1992, Dransch 1997). Examples include the successive display of map objects, the alteration of graphic presentation, or the variation of the user's viewpoint. According to this differentiation two types of cartographic animation exist: temporal animation and nontemporal animation. These are able to improve the visualization of spatial data for demonstration purposes, as well as for exploration tasks; support the understanding of spatial processes and relations; increase the users' attention; and assist graphic data exploration.

Computer animations are quite different presentation forms compared to static maps. They are dynamic and they provide plenty of information to the user in a short time. The existing cartographic rules of map-making, therefore, have to be extended and the question of suitable applications, efficient design, and also cognitive aspects has generated recent research topics in the field of dynamic mapping (MacEachren and DiBiase 1991, Monmonier 1992, Buziek 1999, Kraak 1999).

4. Future of Dynamic Mapping

Developments in computer science and technology concerning virtual reality (VR) may further expand the concept of dynamic mapping. VR technology produces virtual environments by computer (Shneiderman 1998). The user will become part of this virtual world and may interact with it via multifarious interfaces. In cartography, spatial processes could be created with VR technology to give the user a vivid and lifelike impression. In addition, VR technology could also be used to build abstract worlds in which to present spatial concepts or data in a dynamic and highly interactive way. In this way, VR technology will add more vividness and interactivity to cartography and dynamic mapping, however, it also changes the traditional cartographic viewpoint. In contrast to the traditional method of looking at a map, using VR means being in the presentation. This entails the danger of getting lost because the spatial overview that is provided by maps is missing.

See also: Cartographic Visualization; Cartography; Cartography, History of; Cognitive Maps; Geographic Information Systems; Geographic Information Systems: Critical Approaches; Planetary Cartography; Tactile Maps in Geography; Thematic Maps in Geography; Time–Space in Geography;

Bibliography

Buziek G 1999 Dynamic elements of multimedia cartography. In: Cartwright W, Peterson M P, Gartner G (eds.) *Multimedia Cartography*. Springer, Heidelberg, Germany, Chap. 22, pp. 231–4

DiBiase D, MacEachren A M, Krygier J, Reeves C, Brenner A 1992 Animation and the role of map design in scientific visualization. *Cartography and Geographic Information Systems* **19**: 201–14
Dransch D 1997 *Computer-Animation in der Kartographie. Theorie und Praxis. Computer Animation in Cartography: Theory and Practice*. Springer, Heidelberg, Germany
Kraak M J 1999 Cartography and the use of animation. In: Cartwright W, Peterson M P, Gartner G (eds.) *Multimedia Cartography*. Springer, Heidelberg, Germany, Chap. 17, pp. 173–80
MacEachren A M, DiBiase D 1991 Animated maps of aggregate data: conceptual and practical problems. *Cartography and Geographic Information Systems* **18**: 221–9
Monmonier M 1990 Strategies for the visualization of geographic time-series data. *Cartographica* **27**: 30–45
Monmonier M 1992 Summary graphics for integrated visualization in dynamic cartography. *Cartography and Geographic Information Systems* **19**: 23–36
Muehrcke P C 1992 *Map Use: Reading, Analysis and Interpretation*. J. P. Publications, Madison, WI
Peterson M P 1994 *Interactive and Animated Cartography*. Prentice-Hall, Englewood Cliffs, NJ
Shneiderman B 1998 *Designing the User Interface* Addison-Wesley, Reading, MA, Chap. 6.8
Thrower N J W 1961 Animated cartography in the United States. *International Yearbook of Cartography* **1**: 20–30

D. Dransch

Dysexecutive Syndromes

At any given moment, your brain—with its tens of billions of neurons—produces a cacophony of neural activity. Indeed, it seems amazing to think that we are able to direct our attention to any single thought or execute any particular action. Yet, somehow we manage to orchestrate this noisy world and effectively control our mind and actions. Psychologists have coined the term 'executive control' to refer to the ability to select, manage, and control one's mental abilities. Without executive control the brain acts like an orchestra without a conductor or business without a boss.

Studies of patients with brain damage have suggested a syndrome in which executive control appears to have run amok. In particular, patients with damage to the frontal lobes, the most anterior part of the brain, have problems in controlling and directing their thoughts and actions. Indeed, psychological studies have suggested that the frontal lobes are important for a variety of executive functions, including paying attention, searching for thoughts, retrieving memories, initiating actions, and even controlling emotions (Shimamura 2000, Smith and Jonides 1999). Problems often arise in these patients when they are confronted with multiple sensations or choices and must decide which one to select and which ones to ignore. In this sense, these patients exhibit a 'dysexecutive syndrome' (Baddeley 1986).

When confronted with multiple thoughts or choices, patients with frontal lobe damage are easily distracted, unable to focus attention, and have difficulty making plans and decisions. By analogy, consider the kind of control needed to direct traffic in a busy intersection. Cars are entering from all directions, and control must be initiated by preventing some cars from entering the intersection while letting others pass through. Without some control—stop sign, signal, or traffic officer—crashes are likely. Now consider the cars to be neuronal impulses that cross literally billions of intersections. Executive control helps select some impulses while inhibiting others.

1. *Paying Attention*

Paying attention requires the focusing of one's thoughts to specific features in the environment. Let's say you are conversing with someone at a party where many other conversations are occurring at the same time. To carry on your conversation, you must focus your attention on the person with whom you are conversing and attempt to block out or suppress the conversations around you. In the laboratory, a task that has been used to measure attention is the Stroop color-word test. In this task, subjects see a list of color words written in different colored ink (e.g., the word 'RED' written in green ink). They are asked to report the ink color of each word and to ignore the word itself. Thus, the task requires subjects to pay attention to ink colors and suppress the natural tendency to read the words. Our ability to suppress reading is not completely successful as it takes much longer to identify the ink color of these words compared to reading color patches (MacLeod 1991).

Patients with frontal lobe damage exhibit particularly long response latencies when asked to name the ink color of stimuli when the stimuli are themselves incongruous color names (Perret 1974). Also, neuroimaging studies (PET, fMRI) suggest that areas in the frontal lobes and areas along the inner or medial surface of the cortex (the anterior cingulate gyrus) are particularly active when individuals perform the Stroop task (Peterson et al. 1999). Thus, based on these findings there is some evidence for a role of the prefrontal cortex in selecting or initiating activation. Damage to this area produces a dysexecutive deficit in the ability to pay attention—that is, control which perceptual features to select and which to suppress (Moscovitch 1994, Shimamura 2000).

2. *Searching for Memories*

Dysexecutive syndromes include problems in the ability to attend to thoughts and memories just as it includes problems in paying attention to environ-

mental stimuli. For example, take a minute to retrieve all of the words that you can think of that begin with the letter 'A.' This task requires you to search your memories and retrieve specific items. Perhaps, you developed a strategy to control what words you retrieved. For example, after trying to come up with just any words that begin with 'A' maybe you started to cue yourself with categories, such as foods that being with 'A' or animals that begin with 'A.' Or maybe you used two-letter cues, such as 'AB,' then 'AC,' then 'AD.' Such strategies are efficient because they facilitate the retrieval of different words and prevent the report of the same words. Patients with frontal lobe damage have difficulty directing their memory searches (Baldo and Shimamura 1998). They may report only five or six different words in a minute and often repeat the same words.

Controlling one's memories is also important for learning. For example, how would you try to learn the following series of words: car, pear, truck, shirt, banana, hat, airplane, shoe, and apple. Maybe you tried to organize the words by making a story involving the words. Perhaps, you noticed that the words could be grouped by the categories: vehicles, fruits, and clothing. In fact, when learning such word lists, many people group the words in terms of category and then report the words by category saying all of the vehicles first, then fruits, and so on. Controlling our thoughts for efficient learning requires the organization of new material so that it can be retrieved easily (Stuss et al. 1994).

Consider your memory as a library. What if every time you added a new book to your library, you simply threw the book in a pile? After dozens of new books, how easy would it be to retrieve any particular book? An efficient library requires a librarian who catalogs new books and organizes them on the shelves for efficient retrieval. Learning new knowledge (i.e., adding new books to your library) requires cataloging and organizing new information. A dysexecutive syndrome is akin to a library without a good librarian. Patients with frontal lobe lesions have problems learning new information because they cannot organize their thoughts and put them in an appropriate context. That is, they do not catalog information into meaningful groupings.

3. *Making Plans and Selecting Appropriate Responses*

A third aspect of executive control concerns the ability to focus on which plans or responses to initiate. If you had many plans for the day, such as meeting friends, shopping, and meals, it is necessary to develop a strategy or scheme for ordering your actions. Some things must be done at specific times (e.g., a lunch date), whereas other actions can be done with less time constraint. Efficient organization of one's actions requires a kind of control similar to that needed to control one's memories. Strategies must be developed for efficient use of time and energy. Dysexecutive syndromes lead to poor planning, impulsiveness, and flighty behavior. Patients with frontal lobe lesions often exhibit problems in handling everyday events, because they cannot seem to control and organize their plans and actions. Instead, they seem to initiate actions that immediately comes to mind, even if the actions may be inappropriate at the time. For example, a patient with severe frontal lobe damage may enter a restaurant, sit next to a stranger having a meal and start eating. Such patients appear to be overcome with immediate needs and motivations rather than controlling and planning actions appropriate for the moment, such as having a host seat you and ordering food from a menu (Miller and Cummings 1999, Milner and Petrides 1984).

4. *Controlling Emotions*

Finally, dysexecutive syndromes can include problems in controlling one's emotions. The following passage describes the rather interesting case of Phineas Gage, a railroad foreman who incurred frontal lobe damage following a bizarre accident in which an iron rod penetrated—harpoon-like—through Gage's skull damaging his frontal lobes (Harlow 1868, MacMillian 2000).

> He is fitful, irreverent, indulging at times in the grossest profanity (which was not previously his custom), manifesting but little deference for his fellows, impatient of restraint or advice when it conflicts with his desires, at times pertinaciously obstinate, yet capricious and vacillating, devising many plans of future operation, which are sooner arranged than that are abandoned in turn for others appearing more feasible…In this regard his mind was radically changed, so decidedly that his friends and acquaintances said he was "no longer Gage" (Harlow 1868).

Patients with severe damage to the frontal lobes often fail to control their emotions. They exhibit emotional outbursts, both anger and glee, that are not often socially acceptable, such as physical aggression or inappropriate advances to the opposite sex. In this domain, it appears that most people are able to suppress their emotions when they are not appropriate to elicit. Patients with frontal lobe damage often fail to suppress their emotions.

5. *A Board of Executives*

Recent research on human frontal lobes have shed light concerning dysexecutive syndromes. Most of these investigations have centered on the role of the frontal lobes in selecting, monitoring, and controlling mental processes. Indeed, psychological studies of patients with frontal lobe lesions and functional

neuroimaging studies of neurologically intact individuals suggest that this brain region is involved in tasks that require many aspects of executive control, such as stimulus selection, working memory, memory retrieval, and set shifting (D'Esposito et al. 1995, Shimamura 2000).

In simplified terms, the prefrontal cortex is your brain's chief executive—it functions to attend to the environment, ignore irrelevant information, select memories, make decisions, and initiate actions. Without it, thoughts are disorganized, more susceptible to interfering information, and more confused. In actuality, there appears to be several supervisors, each one overseeing a particular mental capacity, such as perceptions, memories, emotions, and action. Thus, the frontal lobes acts more like a governing board of supervisors whose job it is to monitor and control many aspects of mental activity. In the case of Phineas Gage, the supervisor in charge of emotional control was dysfunctional and therefore emotions or feelings were blurted out without regard to the situation or social context. In sum, Phineas was not able to monitor and control the expression of his emotions. Other dysexecutive syndromes appear to affect the attention to perceptions and memories rather than emotions. Interestingly, different parts of the frontal lobes seem to control different aspects of mental function.

See also: Action Planning, Psychology of; Attention and Action; Attention-deficit/Hyperactivity Disorder (ADHD); Attention-deficit/Hyperactivity Disorder, Neural Basis of; Attention: Models; Attention, Neural Basis of; Cognitive Control (Executive Functions): Role of Prefrontal Cortex; Somatosensation

Bibliography

Baddeley A 1986 *Working Memory*. Oxford University Press, Oxford, UK

Baldo J V, Shimamura A P 1998 Letter and category fluency in patients with frontal lobe lesions. *Neuropsychology* **12**(2): 259–67

D'Esposito M, Detre J A, Alsop D C, Shin R K 1995 The neutral basis of the central executive system of working memory. *Nature* **378** (6554): 279–81

Harlow J M 1868 Recovery of an iron rod through the head. *Publications of the Massachusetts Medical Society* **2**: 327–47

MacLeod C M 1991 Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin* **109**(2): 163–203

Macmillian M B 2000 *An Odd Kind of Fame: Stories of Phineas Gage*. MIT Press, Cambridge, MA

Miller B L, Cummings J L (eds.) 1999 *The Human Frontal Lobes: Functions and disorders*. The Guilford Press, New York

Milner B, Petrides M 1984 Behavioural effects of frontal-lobe lesions in man. *Trends in Neurosciences* **7**(11): 403–7

Moscovitch M 1994 Cognitive resources and dual-task interference effects at retrieval in normal people: The role of the frontal lobes and medial temporal cortex. *Neuropsychology* **8**: 524–34

Perret E 1974 The left frontal lobe of man and the suppression of habitual responses in verbal categorical behaviour. *Neuropsychologia* **12**: 323–30

Peterson B S, Skudlarski P, Gatenby J C, Zhang H 1999 An fMRI study of Stroop word-color interference: Evidence for cingulate subregions subserving multiple distributed attentional systems. *Biological Psychiatry* **45**: 1237–58

Roberts A C, Robbins T W, Weiskrantz L W (eds.) 1998 *The Prefrontal Cortex: Executive and Cognitive Function*. Oxford University Press, Oxford, UK

Shimamura A P 2000 The role of the prefrontal cortex in dynamic filtering. *Psychobiology* **28**(2): 207–18

Smith E E, Jonides J 1999 Storage and executive processes in the frontal lobes. *Science* **283**(5408): 1657–61

Stuss D T 1991 Interference effects on memory functions in postleukotomy patients: An attentional perspective. In: Harvey E, Levin S, Howard E, Eisenberg M, Benton A (eds.) *Frontal Lobe Function and Dysfunction*. Oxford University Press, Oxford, UK, pp. 157–72

Stuss D T, Alexander M P, Palumbo C L, Buckle L 1994 Organizational strategies with unilateral or bilateral frontal lobe injury in word learning tasks. *Neuropsychology* **8**: 355–73

A. P. Shimamura

Dyslexia (Acquired) and Agraphia

1. Forms of Acquired Dyslexia

It has been known at least since the late nineteenth century that damage to the left hemisphere of the brain can specifically affect a person's ability to read—that is, can result in an acquired dyslexia. However, it was not until the seminal paper by Marshall and Newcombe (1973) that it became clear that acquired dyslexia occurs in a variety of forms, and that one can understand these forms by relating them to a model of the normal reading system. The basic approach here is to propose an explicit model of reading containing a variety of processing components, and then to seek to show how a particular pattern of sparings and impairments of these components would result in a set of reading symptoms that corresponds to the symptoms of a particular observed form of acquired dyslexia.

Marshall and Newcombe (1973) described three forms of acquired dyslexia; at least three other forms have subsequently been identified, and all six will be described here.

1.1 Phonological Dyslexia

Beauvois and Derouesné (1979) were the first to report that in cases of acquired dyslexia the ability to read nonwords aloud could be selectively impaired relative to the ability to read words aloud. They reported six cases of previously literate readers of French for whom

this was the case, and numerous other cases have been reported since then; see Coltheart (1996) for a recent comprehensive review. The most extreme of these cases was WB (Funnell 1983) who scored 0/20 correct on reading aloud very simple nonwords such as *nust* or *cobe*, while achieving scores of 85–90 percent correct in reading aloud long low-frequency morphemically complex abstract words such as *satirical* or *preliminary*.

When the ability to read aloud pseudohomophones (printed nonwords whose pronunciations are real words, such as FAKT) is compared to the ability to read aloud nonpseudohomophonic nonwords such as FANT, some phonological dyslexics show a clear advantage for the pseudohomophonic items while others clearly show no difference here (see e.g., Berndt et al. 1996). These results are not only obviously relevant for any attempts to offer a theoretical interpretation of phonological dyslexia; they also show that this condition is not a homogeneous one. Selective difficulty in reading nonwords aloud must occur for different reasons in different patients, if some patients' nonword reading shows an advantage for pseudohomophones while other patients' nonword reading does not.

1.2 Surface Dyslexia

If, after brain damage, word reading can be spared while nonword reading is selectively impaired (that is, if phonological dyslexia can occur), it is natural to wonder whether the reverse pattern might also arise: worse reading of words than nonwords. Here it is important to distinguish between *regular* and *irregular* words. Regular words are those that respect the standard relationships between letters or letter groups and their corresponding pronunciations; irregular words are those whose spelling–sound relationships violate at least one such correspondence. Thus the word *steak* is irregular because of the way the letter group *ea* is pronounced in that word; in almost all words containing this spelling, the *ea* is pronounced as in *steam* or *freak*, both of which are regular words. On all extant models of the reading system, the procedure people use to read nonwords also allows perfectly accurate reading of regular words: so all theories predict that one would never find a person with acquired dyslexia who could read nonwords but was impaired at reading of regular words.

Irregular words could be a different matter, however. According to at least some models of reading, irregular words cannot be read correctly by the procedure that is used to read nonwords—the application of knowledge about typical relationships between segments of spelling and their most common pronunciations would produce correct reading of nonwords but incorrect reading of irregular words. So one might expect to come across patients with acquired dyslexia who are good at reading nonwords but bad at reading irregular words—and such patients have indeed been described.

This form of acquired dyslexia is known as surface dyslexia, first described by Marshall and Newcombe (1973) and reviewed by Patterson et al. (1985). The most extreme cases so far reported are MP (Bub et al. 1985) and KT (McCarthy and Warrington 1986), both of whom were normal in speed and accuracy of reading regular words and nonwords while severely impaired in reading exception words.

If patients with surface dyslexia are erring with exception words because they are reading these words via the application of knowledge about typical relationships between segments of spelling and their most common pronunciations, then one would expect their reading errors to demonstrate this. Precisely this type of error is indeed very frequent in surface dyslexia, and it is known as the 'regularization error'—for example, reading *broad* to rhyme with 'code,' or *steak* to rhyme with 'meek.' Some patients with surface dyslexia are not perfect at reading regular words: at least some of these patients produce what might be called 'irregularization errors' with regular words, such as reading *hoo*t as if it rhymed with 'foot' or *hea*t as if it rhymed with 'threat' (Patterson et al. 1996).

The deficit in irregular-word reading by surface dyslexics is frequency-sensitive, with performance much worse for low-frequency irregular words than for high-frequency irregular words.

As with phonological dyslexia, there is more than one form of surface dyslexia. Reading errors such as *gauge* → 'That is something about a railway…that's as much as I've got, train,/gɔdʒ/' (Kay and Patterson 1985; patient EST) clearly show that the problem here is not in visual word recognition (since the printed word has been comprehended correctly, it must first have been recognized correctly) but instead is a failure in retrieval of stored information about the correct pronunciation of that particular word. Since the patient is being asked to read aloud— i.e., to produce a spoken response—he can only fall back on the procedure used for reading nonwords aloud, which of course will yield a regularization error with this irregular word. This is 'output surface dyslexia.'

A different form, 'input surface dyslexia,' is demonstrated when patients asked to define printed exception words produce such responses as *bear* → 'a drink, beer' (Coltheart et al. 1985; patient CD). Here the patient has not recognized the word as an orthographic whole (which is necessary for normal reading comprehension) so has had to attempt the reading comprehension task by first translating the word from orthography to phonology via the application of knowledge about typical relationships between segments of spelling and their most common pronunciations. Since *bear* is an irregular word, a regularization error (translating *bear* into the pronunciation of 'beer') will occur here. The use of this regularized phonological form to access

semantics is responsible for the specific comprehension error made here.

1.3 Visual Dyslexia

There has been curiously little research on this form of acquired dyslexia. Its characteristic symptom is the abundance of visual errors in reading aloud—responses which share many letters with the stimulus but are otherwise unrelated to it, such as reading *enigma* as 'image' or *garden* as 'anger.'

1.4 Lexical Nonsemantic Reading

Brain damage sometimes affects the semantic system and so impairs the comprehension of written or spoken words. This can happen as a consequence of head injury, but is more commonly due to a progressive brain deterioration such as dementia of the Alzheimer's type. Schwartz et al. (1980) studied the reading performance of WLP, a patient of this type. At a point in the evolution of her disorder where she could no longer understand written words reliably, she could still read aloud very well—and that included the ability to read aloud exception words such as *blood*, *shoe*, and *sweat*.

According to some models of reading aloud (Hillis and Caramazza 1991), the procedure by which irregular words can be read correctly is via a pathway from orthography to meaning and then a pathway from meaning to pronunciation. If that were so, then a severe semantic impairment would always be accompanied by an impairment in the reading of irregular words. The data from WLP, and from other reports of patients with severe semantic impairment but good irregular-word reading (Cipolotti and Warrington 1995, Coslett 1991, Lambon Ralph et al. 1995), show that this kind of model of reading cannot be correct; and this form of acquired dyslexia has come to be called 'lexical nonsemantic reading.'

1.5 Pure Alexia (Letter-by-letter Reading)

In this form of acquired dyslexia (see Coltheart (1998) for a recent comprehensive review), words and nonwords are read aloud slowly and effortfully, usually by naming each letter in left-to-right order. If all the letters are named correctly, the patient will then usually produce the whole word or nonword correctly. Latency to read aloud increases linearly with the length of the word or nonword, with slopes sometimes as shallow as one letter per five or more seconds, so that a six letter word might require 30 seconds or more before a correct reading response is produced. The term 'letter-by-letter reading' is used since it describes the characteristic reading behavior of these patients. The term 'pure alexia' (or 'alexia without agraphia') is used because writing and spelling is often intact in these patients; this is discussed further below.

Some patients with pure alexia exhibit a phenomenon known as 'covert reading.' When letter strings are presented at durations too short for the patient to be able to read them aloud—two seconds or so, for example—and the patient asked to judge whether the letter string was a word or not, or in other studies whether it was the name of an animal or not, performance on these classification tasks can be well above chance, and even sometimes perfect, even though the stimuli cannot be read aloud and the patient may even claim not to have seen them at all (see e.g., Saffran and Coslett 1998). It is argued that in pure alexia 'the right hemisphere supports performance in covert reading tasks, and…letter-by-letter reading is the product of the left hemisphere operating on information transferred from the right' (Saffran and Coslett 1998, p. 141). Here it is important to appreciate the lesion sites typically associated with pure alexia. These patients generally have damage to two distinct regions of the brain, the left occipital cortex and the splenium of the corpus callosum. The left occipital lesion means that printed words must be visually processed by the right hemisphere's occipital cortex. That in itself would not cause reading to be abnormal, however, since intact readers have no difficulty in reading words briefly presented in the left visual field (must be visually processed by the right hemisphere's occipital cortex). Normal readers read promptly under these conditions by transferring information from right occipital cortex to language systems in the left hemisphere. Pure alexics can't do that, and the reason must be that the splenium of the corpus callosum is the pathway by which such interhemispheric transfer for reading is normally done. In the patients, the splenial lesion means that some other callosal pathways have to be used for this transfer, and the use of these other pathways may impose a slow and serial letter-by-letter transfer process.

1.6 Deep Dyslexia

This acquired dyslexia is reviewed in detail by Coltheart et al. (1980). Its cardinal symptom is the semantic error in reading aloud. When single isolated words are presented for reading aloud with no time pressure, the deep dyslexic will often produce as a response a word that is related in meaning, but in no other way, to the word he or she is looking at: *dinner* → 'food,' *uncle* → 'cousin,' and *close* → 'shut' are examples of the semantic errors made by the deep dyslexic GR (Coltheart et al. 1980). Visual errors such as *quarrel* → 'squirrel' or *angle* → 'angel,' and morphological errors such as *running* → 'runner' or *unreal* → 'real,' are also seen. Concrete (highly-imageable) words such as *tulip* or *green* are much more likely to be

successfully read than abstract (difficult-to-image) words such as *idea* or *usual*. Function words such as *and*, *the*, or *or* are very poorly read. Nonwords such as *vib* or *ap* cannot be read aloud at all.

As noted above, Marshall and Newcombe (1973) proposed that different forms of acquired dyslexia might be interpretable as consequences of specific different patterns of breakdown within a multicomponent model of the normal skilled reading system. That kind of interpretation has also been offered for deep dyslexia, by Morton and Patterson (1980). However, this way of approaching the explanation of deep dyslexia was rejected by Coltheart (1980) and Saffran et al. (1980), who proposed that deep dyslexic reading was not accomplished by an impaired version of the normal skilled reading system, located in the left hemisphere of the brain, but relied instead on reading mechanisms located in the intact right hemisphere.

Subsequent research has strongly favored the right-hemisphere interpretation of deep dyslexia. Patterson et al. (1987) report the case of an adolescent girl who developed a left-hemisphere pathology that necessitated removal of her left hemisphere. Before the onset of the brain disorder, she was apparently a normal reader for her age; after the removal of her left hemisphere, she was a deep dyslexic. Michel et al. (1996) report the case of a 23-year-old man who as a result of neurosurgery was left with a lesion of the posterior half of the corpus callosum. They studied his ability to read tachistoscopically displayed words presented to the left or right visual hemifields. With right hemifield (left hemisphere) presentation, his reading was normal. With left hemifield (right hemisphere) presentation, his reading showed all the symptoms of deep dyslexia. In a brain imaging study, Weekes et al. (1997) found that brain activation associated with visual word recognition was greater in the right than the left hemisphere for a deep dyslexic, but not for a surface dyslexic, nor for two normal readers.

It seems clear, then, that deep dyslexia is unlike all the other patterns of acquired dyslexia discussed here, in that deep dyslexics do not read via some damaged version of the normal (left-hemisphere) reading system, whereas patients with other forms of acquired dyslexia do.

2. *Forms of Acquired Dysgraphia and Their Relationship to Acquired Dyslexia*

Just as damage to the left hemisphere of the brain can sometimes specifically affect reading, so too left-hemisphere damage can sometimes specifically affect spelling and/or writing. Such an impairment of writing or spelling in a previously literate person is referred to as an acquired dysgraphia; and various different forms of patterns of acquired dysgraphia have been reported. There is a selective difficulty in spelling nonwords relative to words, known as 'phonological agraphia' or 'phonological dysgraphia' (e.g., Shallice 1981). There is a selective difficulty in spelling irregular words relative to regular words and nonwords, known as 'lexical agraphia' or 'surface dysgraphia' (e.g., Beauvois and Derouesné 1981). There are cases in which semantic errors are made in spelling to dictation, and spelling of nonwords to dictation is impossible, known as 'deep dysgraphia' (Bub and Kertesz 1982). There is also 'lexical nonsemantic spelling' (Patterson 1986)

Almost all patients with acquired dyslexia also exhibit acquired dysgraphia—the only clear exception here is for pure alexia (letter-by-letter reading), in which writing and spelling are often intact. Patients with acquired dyslexia and acquired dysgraphia offer unique opportunities for learning more about the relationships between the systems intact people use to read and to spell. For example, when one sees a patient who reads *yacht* as '/jætʃt/' and spells 'yacht' as *yot*—a patient with surface dyslexia and surface dysgraphia—it is exceedingly tempting to take this as evidence that the body of orthographic knowledge we use to recognize familiar printed words is the same as the body of orthographic knowledge we use when retrieving familiar spellings. If this were so, then loss of the representation of the word *yacht* from this body of knowledge would force the word to be both read and spelled via the application of knowledge about typical relationships between segments of spelling and their most common pronunciations. This method of reading and of spelling would generate the particular erroneous reading and spelling responses just described. Here one would be arguing against the view that there are distinct bodies of orthographic knowledge for orthographic input (reading) and for orthographic output (spelling and writing) (as proposed by many (e.g., Morton 1980) and for the view that there is just a single body of orthographic knowledge used both for reading and for spelling and writing (as proposed by Allport and Funnell 1981).

But the relationships between the form of acquired dyslexia a person has and the form of acquired dysgraphia accompanying it are unfortunately much more complicated than this. Most people with surface dysgraphia also show surface dyslexia, but not all do: RG (Beauvois and Derouesné 1979) was a surface dysgraphic but a phonological dyslexic. Most people with deep dysgraphia also show deep dyslexia, but not all: MK (Howard and Franklin 1988) was a deep dysgraphic but a surface dyslexic, and JC (Bub and Kertesz 1982) was a deep dysgraphic but essentially normal at reading.

Thus the relationships between acquired dyslexia and acquired dysgraphia are simply not understood at present. But as has been documented above acquired dyslexia itself is rather well understood. Various types of acquired dyslexia have been clearly delineated, as have subvarieties of each type, and we have very clear ideas about how to interpret these patterns of impaired

reading in relation to an information-processing theory of the normal reading system.

3. Computational Cognitive Neuropsychology of Reading

Currently a number of information-processing theories of reading aloud are expressed as computational models—that is, as executable computer programs which turn print into phonology and do so by using the specific information-processing procedures posited by the particular theory. Some of these models are in the connectionist tradition and are built up via connectionist learning algorithms such as backpropagation (Plaut et al. 1996, Zorzi et al. 1998). Others are nonconnectionist, with their architectures specified by the modeler rather than by a learning algorithm (Coltheart et al. 2001). This is an area of much ongoing theoretical development.

One way to test such models is to investigate their ability to simulate different patterns of acquired dyslexia. The way this is done is to try to 'lesion' particular components of the program so as to cause the computational model to exhibit the same kinds of errors as do patients with particular forms of acquired dyslexia—to cause the program to make regularization errors with irregular words, especially low-frequency ones, for example, whilst still being good at reading nonwords (thus simulating surface dyslexia) or to cause the program to be poor at reading nonwords while still being able to read words (thus simulating phonological dyslexia). The computational cognitive neuropsychology of reading is currently being pursued both in the connectionist tradition (Plaut 1997) and in the nonconnectionist tradition (Coltheart et al. 2001)

See also: Dyslexia, Developmental; Dyslexia: Diagnosis and Training; Dyslexia, Genetics of; Reading, Neural Basis of

Bibliography

Allport D A, Funnell E 1981 Components of the mental lexicon. *Philosophical Transactions of the Royal Society of London B* **295**: 397–410

Beauvois M F, Derouesné J 1979 Phonological alexia: three dissociations. *Journal of Neurology, Neurosurgery, and Psychiatry* **42**: 1115–24

Beauvois M F, Derouesné J 1981 Lexical or orthographic agraphia. *Brain* **85**: 411–26

Berndt R S, Haendiges A N, Mitchum C C, Wayland S C 1996 An investigation of nonlexical reading impairments. *Cognitive Neuropsychology* **13**: 763–801

Bub D, Kertesz A 1982 Deep agraphia. *Brain and Language* **17**: 146–65

Bub D, Cancelliere A, Kertesz A 1985 Whole word and analytic translation of spelling-to-sound in a non-semantic reader. In: Patterson K E, Coltheart M, Marshall J C (eds.) *Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading*. Erlbaum, London, pp. 15–34

Cipolotti L, Warrington E K 1995 Semantic memory and reading abilities: A case report. *Journal of the International Neuropsychological Society* **1**: 104–10

Coltheart M 1980 Deep dyslexia: A right hemisphere hypothesis. In: Coltheart M, Patterson K, Marshall J C (eds.) *Deep Dyslexia*. Routledge and Kegan Paul, London

Coltheart M (ed.) 1996 *Phonological Dyslexia*. Erlbaum, Hove, UK

Coltheart M (ed.) 1998 *Pure Alexia (Letter-by-Letter Reading)*. Psychology Press, Hove, UK

Coltheart M, Byng S, Masterson J, Prior M, Riddoch M J 1983 Surface dyslexia. *Quarterly Journal of Experimental Psychology* **35A**: 469–95

Coltheart M, Patterson K, Marshall J C (eds.) 1980 *Deep Dyslexia*. Routledge & Kegan Paul, London

Coltheart M, Rastle K, Perry C, Langdon R, Ziegler J 2001 DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review* **108**: 204–56

Coslett H B 1991 Read but not write 'idea': Evidence for a third reading mechanism. *Brain and Language* **40**: 425–43

Funnell E 1983 Phonological processes in reading: New evidence from acquired dyslexia. *British Journal of Psychology* **74**: 159–80

Hillis A D, Caramazza A 1991 Mechanisms for accessing lexical representations for output: Evidence from a category-specific semantic deficit. *Brain and Language* **40**: 106–44

Howard D, Franklin S 1988 *Missing the Meaning? A Cognitive Neuropsychological Study of the Processing of Words by an Aphasic Patient*. MIT Press, Cambridge, MA

Kay J, Patterson K E 1985 Routes to meaning in surface dyslexia. In: Patterson K E, Marshall J C, Coltheart M (eds.) *Surface Dyslexia: Cognitive and Neuropsychological Studies of Phonological Reading*. Erlbaum, Hove, UK

Lambon Ralph M A, Ellis A W, Franklin S 1995 Semantic loss without surface dyslexia. *Neurocase* **1**: 363–69

Marshall J C, Newcombe F 1973 Patterns of paralexia: A psycholinguistic approach. *Journal of Psycholinguistic Research* **2**: 175–99

McCarthy R, Warrington E K 1986 Phonological reading: Phenomena and paradoxes. *Cortex* **22**: 359–80

Michel F, Henaff M A, Intrilligator J 1996 Two different readers in the same brain after a posterior callosal lesion. *NeuroReport* **7**: 786–88

Morton J 1980 The logogen model and orthographic structure. In: Frith U (ed.) *Cognitive Approaches to Spelling*. Academic Press, London

Morton J, Patterson K E 1980 A new attempt at interpretation, or, an attempt at a new interpretation. In: Coltheart M, Patterson K E, Marshall J C (eds.) *Deep Dyslexia*. Routledge and Kegan Paul, London

Patterson K E 1986 Lexical but nonsemantic spelling? *Cognitive Neuropsychology* **3**: 341–67

Patterson K E, Marshall J C, Coltheart M (eds.) 1985 *Surface Dyslexia: Cognitive and Neuropsychological Studies of Phonological Reading*. Erlbaum, Hove, UK

Patterson K, Plaut D C, McClelland J L, Seidenberg M S, Behrmann M, Hodges J R 1996 Connections and disconnections: A connectionist count of surface dyslexia. In: Reggia J A, Ruppin E, Berndt R S (eds.) *Neural Modeling of Brain and Cognitive Disorders*. World Scientific, New York

Patterson K, Vargha-Khadem F, Polkey C 1987 Reading with one hemisphere. *Brain* **112**: 39–63

Plaut D C 1997 Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes* **12**: 767–808
Plaut D C, McClelland J L, Seidenberg M S, Patterson K E 1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115
Saffran E M, Bogyo L C, Schwartz M F, Marin O S M 1980 Does deep dyslexia reflect right-hemisphere reading? In: Coltheart M, Patterson K, Marshall J C (eds.) *Deep Dyslexia*. Routledge and Kegan Paul, London
Saffran E M, Coslett B 1998 Implicit vs letter-by-letter reading in pure alexia: A tale of two systems. In: Coltheart M (ed.) *Pure Alexia (Letter-by-Letter Reading)*. Psychology Press, Hove, UK
Schwartz M, Saffran E M, Marin O S M 1980 Fractionating the reading process in dementia: Evidence for word-specific print-to-sound associations. In: Coltheart M, Patterson K, Marshall J C (eds.) *Deep Dyslexia*. Routledge and Kegan Paul, London
Shallice T 1981 Phonological agraphia and the lexical route in writing. *Brain* **104**: 413–29
Weekes B, Coltheart M, Gordon E 1997 Deep dyslexia and right-hemisphere reading—A regional cerebral blood flow study. *Aphasiology* **11**: 1139–58
Zorzi M, Houghton G, Butterworth B 1998 Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance* **24**: 1131–61

M. Coltheart

Dyslexia, Developmental

Developmental dyslexia can be characterized as a specific problem with reading and spelling that cannot be accounted for by low intelligence, poor educational opportunities, or obvious sensory or neurological damage. Children with developmental dyslexia often appear to have good spoken language skills and extensive vocabularies. Accordingly, the first theories of developmental dyslexia postulated underlying deficits in visual processing ('word blindness'). Recent evidence from developmental psychology, genetics, and brain imaging increasingly suggests, however, that the central deficit in developmental dyslexia is a linguistic one. In particular, dyslexic children appear to have difficulties in the accurate specification of phonology when storing lexical representations of speech.

1. The Normal Development of Lexical Representations

A schematic view of a lexical representation is shown in Fig. 1. Researchers in child phonology have shown that, in the normal course of development, the phonological aspects of children's lexical representations become increasingly segmental and distinctly specified in terms of phonetic features. When vocabulary size is small, there is no need to represent words in a systematic or detailed manner, and so early word representations are thought to encode fairly global phonological characteristics (e.g., Jusczyk 1993). As vocabulary grows, these holistic representations are gradually *restructured*, so that smaller segments of sound are represented (Metsala and Walley 1998). Developmentally, children appear to first represent syllables (a word like *wigwam* has two syllables), then onsets and rimes (the onset in a syllable corresponds to the sound/s before the vowel, the rime corresponds to the vowel and any subsequent sounds, e.g., *p - it*; *scr - ew*), and ultimately, phonemes (phonemes are the smallest units of sound that change word meaning, e.g., *pit* differs from *sit*, *pot*, and *pin* by the initial, medial, and final phonemes, respectively). According to this 'emergent' view of segmental representation (Metsala and Walley 1998), the phoneme is not an integral aspect of speech representation and processing from infancy onwards (as was once thought), but rather emerges as a representational unit via *spoken language experience*. The degree to which segmental representation has taken place is in turn thought to determine how easily the child will become phonologically 'aware' and will learn to read and write.

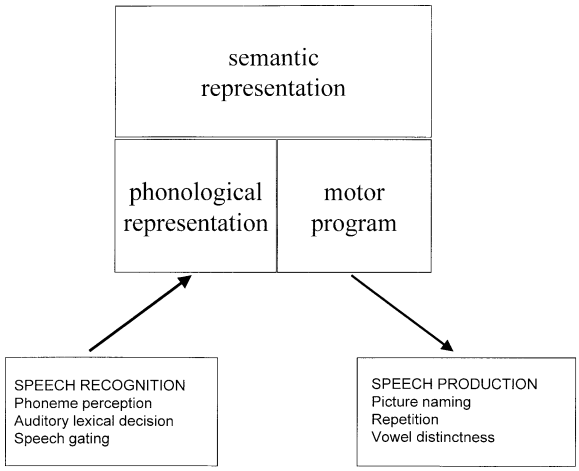

Figure 1
Schematic depiction of a lexical representation, showing cognitive tasks tapping input and output processes

Dyslexic children appear to have specific difficulties in the representation of segmental phonology. Most of the evidence for this 'phonological deficit' has come from 'phonological awareness' tasks, in which children are asked to make judgements about shared phonology or to manipulate phonology (e.g., the *rhyme oddity* task, in which the child is asked to distinguish the odd word out from a triple like *man*, *pat*, *fan*, and

the *phoneme substitution* task, in which the child is asked to substitute the initial phonemes in words, so that Bob Dylan becomes Dob Bylan). A large number of studies carried out in many languages has shown that dyslexic children perform poorly in such phonological awareness tasks (see Goswami in press, for review). Importantly, such deficits are found even when dyslexic children's performance is compared to that of younger children who are reading and spelling at the same level as them (the methodologically important *reading level match* research design, which holds reading level constant rather than chronological age and thus gives a mental age advantage to the dyslexic children). A phonological deficit seems to characterize dyslexia across languages. Whether the child is learning to read a nontransparent orthography like English, or a transparent orthography like German, phonological awareness deficits are consistently found. This suggests that an important causal factor in dyslexia is an underlying difficulty in developing segmentally well-specified lexical representations of speech.

2. *The Abnormal Development of Phonological Representations in Dyslexia*

The cognitive profiles of developmental dyslexics show fairly specific linguistic deficits that are related to phonology rather than semantics. For example, developmental dyslexics have poor phonological memory skills. They find it difficult to remember verbal sequences, and they show poor performance on standardized tests of phonological memory (such as repeating nonsense words like *ballop* and *thickery*). Repeating nonsense (novel) word forms requires the child to set up new phonological representations. Developmental dyslexics are also slower to recognize familiar words on the basis of acoustic input. For example, they require more acoustic-phonetic information than age-matched controls to recognize familiar words in speech gating tasks, in which incremental segments of speech from word-onset are presented through headphones. On the output side, developmental dyslexics show deficits in fine-grained measures of speech production such as the distinctness of vowel articulation. They also show problems in confrontation naming tasks (in which simple pictures must be named on demand), even though they are very familiar with the items in the pictures. Their problems are particularly marked when the familiar names are long and of low frequency, and thus correspond to phonological forms that have been encountered less frequently and that also require the accurate specification of more phonological segments. Developmental dyslexics find rapid 'automatized' speeded naming tasks difficult. These tasks require the child to name familiar items presented in lists (such as digits) as quickly as possible, thereby accessing and outputting what should be well-practiced and so well-specified phonological forms ('nine, two, five,…').

3. *Auditory Processing Problems*

Studies that have attempted to uncover basic auditory processing deficits to explain this apparently selective difficulty in representing phonological forms in memory have so far produced mixed results. Typically, a basic processing task will show dramatic impairments for some dyslexic children, and no apparent impairment for others, despite extremely careful subject selection. Thus, some dyslexic children have great difficulty with 'minimal pairs' tasks that require the discrimination of similar-sounding words (e.g., *smack–snack*), yet others do not (Adlard and Hazan 1998). Some dyslexic children find phoneme categorization and repetition tasks very difficult (e.g., discriminating between or repeating /b/ and /d/), and others do not (e.g., Manis et al. 1997). Some dyslexic children find temporal order judgements very difficult. Typically, such tasks require them to decide whether a high tone preceded a low tone or vice versa. Accordingly, it has been proposed that dyslexia is caused by temporal processing deficits (e.g., Tallal 1980), but this view is quite controversial. For example, Mody et al. (1997) have shown that dyslexic children who find it difficult to discriminate phonemes like ba/da show no deficit with a nonspeech analogue task that makes the same temporal processing demands (sine wave speech). Thus far, most auditory processing deficits that have been found appear to be specific to linguistic stimuli. However, the lack of uniformity across subjects is puzzling. One possibility is that the primary basic processing difficulty has not yet been discovered. A second is that the difficulty lies not in perception *per se*, but in *representing* the segments that can be detected in the input, as discussed above. A third is that sensory deficits in basic auditory processing and phonological deficits are not causally related, but arise from the dysfunction of a neural system common to both (Eden and Zeffiro 1998). Finally, it is logically possible that there may be a variety of causes of dyslexia, with some children's primary problems lying outside the auditory realm.

4. *Phonological Processing Deficits in Dyslexia: Brain Imaging Studies*

Nevertheless, evidence from brain imaging work is consistent with the view that the key difficulty in dyslexia is a weakness in the representation of phonological information. The phonological processing system spans a number of cortical and subcortical regions in the brain, including frontal, temporal, and parietal cortex. Studies of aphasia long ago revealed the importance of Broca's area (frontal cortex) and

Wernicke's area (posterior superior temporal gyrus) for the motor production and receptive aspects of speech, respectively, and the lateralization of language-related processes to the left hemisphere is equally well documented.

Interpretation of functional neuroimaging studies of dyslexia is complicated by the fact that only adult developmental dyslexics can be scanned, which means that the brain's compensatory responses to any initial disorder may account for some of the patterns of task-related activity that are found. Nevertheless, despite seemingly inconsistent results in early studies, localization of phonological mechanisms in the temporo-parietal junction and abnormalities in activation surrounding this junction in adult developmental dyslexics during phonological processing tasks have been documented by a number of research groups (see Eden and Zeffiro 1998 for review). When performing tasks like rhyme judgement, rhyme detection, and word and nonsense word reading, dyslexic subjects show reduced activation in temporal and parietal regions, particularly within the left hemisphere.

5. *Genetics of Developmental Dyslexia*

The heritability of dyslexia has been demonstrated by a number of family and twin studies (e.g., Gayan and Olson 1997). More recently, linkage studies have been used to try to determine where in the human genome the critical genes for dyslexia are located. The most promising findings so far concern the short arm of chromosome 6, and sites on chromosome 15 (e.g., Grigorenko et al. 1997). These studies depend on definitions of the dyslexic phenotype that are based on deficits in phonological awareness tasks and single or nonsense word reading. Thus there is evidence that individual differences in phonological processing are to some extent heritable. Of course, there cannot be a 'gene' for dyslexia in the sense that there is a gene for eye color, as reading is a culturally determined activity. Individuals at genetic risk for dyslexia who develop in a favorable early environment as far as reading is concerned (for example, children who have carers who stimulate their linguistic awareness via language games, nursery rhymes, and so on, and who read books to them and model and encourage literacy activities by reading and writing extensively themselves) may be able to compensate to some extent for their genetic predisposition to dyslexia. Other individuals with a lower degree of risk but relatively adverse early environments may be more handicapped.

6. *Developmental Dyslexia Across Languages*

As noted earlier, the degree to which the segmental representation of phonology has taken place is thought to determine how easily the child will become phonologically 'aware' and will learn to read and write. Phonological awareness is usually defined as the ability to recognize and manipulate subsyllabic units in spoken words. It has been shown to be an important predictor of reading and spelling progress in a vast number of studies of normally developing children in many languages, and follows the developmental sequence syllable–onset/rime–phoneme across all the alphabetic languages so far studied (Goswami in press). Nevertheless, it is also the case that learning to read itself affects the representation of segmental phonology. In particular, attainment of the final developmental stage of full phonemic representation is highly dependent on being taught to read and write an alphabetic language (see Goswami and Bryant 1990 for a review).

For example, adults who are illiterate show the same difficulties with phonological awareness tasks at the phoneme level as prereading children, and Japanese children (who learn to read a syllabary) show relatively poor phoneme level performance in comparison to American children.

Given the close relationship between the development of phonemic awareness and learning to read and write, the difficulties faced by developmental dyslexics would be expected to vary to some degree with linguistic transparency, even if the underlying problem in all languages concerns phonological representation. This is in fact the case. Dyslexic children who are learning to read and spell transparent orthographies like German show the confrontation naming deficit characteristic of English dyslexic children, suggesting that their phonological representations of speech are compromised. However, they do not show accuracy deficits in nonsense word reading tasks, although they do show marked speed deficits (e.g., Wimmer 1993). When nonsense words are carefully matched across languages, German dyslexic children are much better at reading nonsense words than English children (Landerl et al. 1997). This presumably reflects the consistency of grapheme–phoneme relations in German compared to English. The German child, even the child with a linguistic difficulty, is learning a less variable orthographic code.

7. *Subtypes of Developmental Dyslexia?*

Some researchers have argued strongly for two subtypes of developmental dyslexia: 'surface' dyslexia and 'phonological' dyslexia (but see Oliver et al. in press). Phonological dyslexics are thought to have problems in assembling sublexical phonology (reading using abstract spelling-to-sound 'rules'). Surface dyslexics are thought to have problems in using stored lexical knowledge to pronounce words (they seem to have difficulty in directly addressing orthographic representations). This 'double dissociation' model of de-

velopmental dyslexia has been drawn from patterns of adult pathology (the 'acquired' dyslexias).

The subtyping argument has depended critically on the use of regression procedures, in which performance relationships between the use of lexical and sublexical phonology characteristic of normal children are used to derive confidence limits for assessing the same performance relationships in dyslexic populations. These regression procedures typically depend on the use of *chronological* age-matched normal readers. As processing tradeoffs between the reliance on lexical and sublexical phonology may depend on the overall level of word recognition that the child has attained, *reading level* matched controls are more appropriate. Stanovich et al. (1997) demonstrate clearly that when reading level matched controls are used to define processing tradeoffs, then almost no 'surface dyslexic' children are found, even in the original samples used to support the notion of subtypes of developmental dyslexia. They conclude that the 'surface' dyslexic profile arises from a milder form of phonological deficit accompanied by exceptionally inadequate reading experience.

8. *Towards a Phenotype of Dyslexia*

Neurodevelopmental disorders are not discrete syndromes, but most typically have a graded character. They are also seldom highly specific, and rarely involve a single deficit. Although a difficulty in representing segmental phonology appears to be the core symptom of developmental dyslexia, many researchers have noted a number of associated deficits that more or less frequently accompany the phonological deficit. For example, dyslexic children can be clumsy, they can have attentional problems, they may display mild cerebellar signs such as slow motor tapping, and they may have subtle abnormalities in visual processing such as poor motion processing when visual stimuli are of low contrast and low luminance. At present, the most widely accepted phenotype of developmental dyslexia depends on deficits in performance on a battery of cognitive tasks that tap the integrity of the child's phonological representations of speech. This battery most typically includes phonological awareness tasks, single word and nonsense word reading tasks, and rapid 'automatized' naming tasks. Developmental dyslexia is thus currently best conceived as a subtle linguistic deficit that involves an impairment in the representation of the phonological segments of speech in the brain.

See also: Dyslexia (Acquired) and Agraphia; Dyslexia: Diagnosis and Training; Dyslexia, Genetics of; Language Development, Neural Basis of; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Phonology; Reading, Neural Basis of; Spelling

Bibliography

Adlard A, Hazan V 1998 Speech perception in children with specific reading difficulties (dyslexia). *Quarterly Journal of Experimental Psychology* **51A**: 153–77

Eden G F, Zeffiro T A 1998 Neural systems affected in developmental dyslexia revealed by functional neuroimaging. *Neuron* **21**: 279–82

Gayan J, Olson R K 1997 Common and specific genetic effects on reading measures. *Behavioral Genetics* **27**: 589

Goswami U in press Phonological representations, reading development and dyslexia: Towards a cross-linguistic theoretical framework. *Dyslexia*

Goswami U, Bryant P E 1990 *Phonological Skills and Learning to Read*. Erlbaum, Hillsdale, NJ

Grigorenko E L, Wood F B, Meyer M S, Hart L A, Speed W C, Shuster A, Pauls D L 1997 Susceptibility loci for distinct components of developmental dyslexia on chromosomes 6 and 15. *American Journal of Human Genetics* **60**: 27–39

Jusczyk P W 1993 From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics* **21**: 3–28

Landerl K, Wimmer H, Frith U 1997 The impact of orthographic consistency on dyslexia: A German–English comparison. *Cognition* **63**: 315–34

Manis F R, McBride-Chang C, Seidenberg M S, Keating P, Doi L M, Munsun B, Petersen A 1997 Are speech perception deficits associated with developmental dyslexia? *Journal of Experimental Child Psychology* **66**: 211–35

Metsala J L, Walley A C 1998 Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In: Metsala J L, Ehri L C (eds.) *Word Recognition in Beginning Literacy*. Erlbaum, Hillsdale, NJ, pp. 89–120

Mody M, Studdert-Kennedy M, Brady S 1997 Speech perception deficits in poor readers: Auditory processing or phonological coding? *Journal of Experimental Child Psychology* **64**: 199–231

Oliver A, Johnson M H, Karmiloff-Smith A, Pennington B in press Deviations in the emergence of representations: A neuroconstructivist framework for analysing developmental disorders. *Developmental Science*

Stanovich K E, Siegal L S, Gottardo A 1997 Converging evidence for phonological and surface subtypes of reading disability. *Journal of Educational Psychology* **89**: 114–27

Tallal P 1980 Auditory temporal perception, phonics and reading disabilities in children. *Brain and Language* **9**: 182–98

Wimmer H 1993 Characteristics of developmental dyslexia in a regular writing system. *Applied Psycholinguistics* **14**: 1–33

U. Goswami

Dyslexia: Diagnosis and Training

Developmental dyslexia, a childhood disorder characterized by 'unexpected' reading and spelling difficulties, became the subject of serious scientific enquiry in the 1970s. This article discusses how the definition of dyslexia is critical to its diagnosis and considers the implications of research on dyslexia for the assessment and teaching of dyslexic children.

1. The Definition of Dyslexia

Since the landmark Isle of Wight studies, it has become conventional to distinguish children who have specific reading difficulties (dyslexia) from children who have reading difficulties in the context of more general learning problems (reading backwardness). A child is considered to have a specific problem with reading if their reading attainment is significantly below that predicted from their general cognitive ability on the basis of the correlation between reading and IQ in children of the same age (a discrepancy between expected and actual attainment). However, a limitation inherent in the 'discrepancy-definition' of dyslexia is its dependence on IQ because the verbal IQ of poor readers is likely to decline as a consequence of their limited reading experience. Thus, diagnosis using this definition of dyslexia runs the risk of missing true cases of dyslexia either in children whose verbal IQ is lowered, or in those who have overcome their reading problems but may have significant difficulty with spelling and written work.

2. Dyslexia as a Core Phonological Deficit

The most widely accepted current view of dyslexia is that it is part of the continuum of language disorders and reflects phonological (speech) processing difficulties (Share 1995). These include problems with phonological awareness, verbal short-term memory, naming especially in speeded tasks, phonological learning, speech perception and nonword repetition. Current theories suggest that dyslexic children have difficulties first in establishing, and later in accessing, adequate phonological representations (see *Dyslexia, Developmental*). The strong evidence of phonological deficits in dyslexia has led to the definition of dyslexia as a core phonological deficit, irrespective of IQ (Stanovich and Siegel 1994). This definition makes sense in terms of what is known about normal reading acquisition; phonemic skills are critical to the development of the alphabetic principle that allows children to decode words they have not seen before. Dyslexic children typically have difficulty reading words that are not part of their sight vocabulary, such as non-words.

3. Subtypes of Dyslexia

Most systematic studies of individual differences among dyslexics have revealed variations in their reading skills. As well as dyslexic children who have difficulties using phonological strategies for reading and spelling, it is possible to identify others who appear to have mastered alphabetic skills and can use these proficiently; at least when there are no time demands. Such children have been referred to as developmental 'surface' dyslexics. These children rely heavily upon phonological strategies for reading and spelling. Thus, they tend to pronounce irregular words as though they were regular (e.g., island—izland), they have particular difficulty distinguishing between homophones like pear-pair, and their spelling is usually phonetic (see *Dyslexia (Acquired) and Agraphia*).

A number of recent studies have suggested that individual differences in the reading skills of dyslexic readers might be explained by differences in the severity or developmental status of underlying phonological skills. Thus, children with phonological dyslexia appear to have a disorder of reading development whereas surface dyslexics are developmentally delayed in their acquisition of reading skills (Snowling 2000).

4. Dyslexia Across the Life-span

A second advantage of the core phonological deficit hypothesis of dyslexia is that it accounts well for the changing profile of dyslexia across the life span. While reading problems are the key symptom of dyslexia in the early school years, adults with a childhood history of dyslexia often become fluent readers though few overcome their spelling problems. Adult dyslexics also have difficulty decoding words they have not seen before (i.e., nonwords) and with phonological awareness, speeded naming and verbal short-term memory tasks.

At the other end of the age spectrum, preschool children with a genetic risk of dyslexia show phonological difficulties in the context of slower than expected vocabulary development (Gallagher et al. 2000). Indeed, problems of phonological awareness, together with limitations of letter knowledge presage reading difficulties in high-risk children (Byrne et al. 1997, Lefly and Pennington 1996).

5. The Assessment of Dyslexia

The convergence of evidence from studies of the normal development of reading and from dyslexia enables practitioners both to identify children at risk of dyslexia and to provide early intervention. Children who have a family history of dyslexia and children who have a history of speech-language difficulties appear to be at risk, although the nature of the speech-language impairment is an important factor in determining outcome. By school entry it is possible to identify with some certainty, children who will go on to have difficulties with reading by assessing their performance on tests of phonological awareness and letter knowledge. With young children in the age range of four to five, tests of syllable and rhyme awareness seem the most suitable, while from six years onwards, tests requiring awareness of spoken words at the level of the phoneme are the most sensitive.

In spite of the attractions of the core phonological definition of dyslexia, the whole-scale abandonment of general cognitive ability in the diagnosis of dyslexia would be hazardous. For children who have already failed to learn to read, the starting point of the assessment should involve an assessment of general cognitive ability. This form of assessment is important in diagnosing whether dyslexia accompanies a broader language disorder (Bishop 1997) or the syndrome of nonverbal learning disabilities (Rourke 1995). However, the assessment will be most useful if it proceeds to assess the phonological skills that underlie reading development and the reading strategies the child is currently using with a view to prescribing appropriate intervention (Snowling and Stackhouse 1996). From the point of view of diagnosis, a limitation of the core phonological definition of dyslexia is its focus on phonology and the development of decoding skills. Whilst there is a consensus that automatic decoding skills are critical to reading development, proficient reading is the product of both decoding and linguistic comprehension skills. Although dyslexic readers have basic level decoding difficulties, many have normal linguistic comprehension that can afford them a number of advantages; they tend to comprehend what they read relatively well and they can monitor their comprehension processes.

6. *Comorbidity and Compensation*

Some of the variation among dyslexics seen at the behavioural level may be due to comorbidity between dyslexia and another disorder. A substantial body of research suggests that many dyslexic children have low level impairments of the transient visual system (Hogben 1997). Although there is no conclusive evidence that these are causally linked with dyslexic children's reading impairments, such problems may affect the ability of children who have phonological difficulties to compensate. Similarly, some dyslexic children have motor impairments and others problems of attention control (Pennington et al. 1993).

Arguably, the course of reading development followed by a particular dyslexic child will be determined not only by the severity of their phonological processing problem, but also by their other cognitive skills (Snowling 2000). For dyslexic children who have good semantic skills and poor phonology, it is likely that their word recognition will improve in the face of decoding deficits because they rely heavily on context. In contrast, dyslexic children who have milder phonological difficulties or poorer general language skills will tend to show delayed development along normal lines. Furthermore, the environmental variables impinging upon the dyslexic child are critical to eventual outcome. Cross-linguistic studies are beginning to reveal that the manifestations of dyslexia depend on the orthography in which a child learns to read (Frith et al. 1998). More generally, print exposure, a measure of the extent to which a child reads, is an independent predictor of reading attainment.

7. *Interventions for Dyslexia*

There is now a considerable body of literature pointing to the benefits of phonological awareness training to subsequent literacy development (Toija 1999). Training in phonological skills is particularly effective when it is combined with highly structured reading tuition. However, it is also becoming clear that dyslexic children and children at risk of dyslexia respond relatively less well to these techniques than normally developing readers (Byrne et al. 1997). Indeed, initial phonological skill is one of the best predictors of responsiveness to intervention in such training regimes (Hatcher and Hulme 1999). For children who do not respond to such treatments, more intensive therapies may need to be designed and special school placement considered.

See also: Dyslexia (Acquired) and Agraphia; Dyslexia, Developmental; Dyslexia, Genetics of; Reading, Neural Basis of; Reading Skills; Spelling

Bibliography

Bishop D V M 1997 *Uncommon Understanding*. Psychology Press, Hove, UK

Byrne B, Fielding-Barnsley R, Ashley L, Larsen K 1997 Assessing the child's contribution to reading acquisition: What we know and what we don't know. In: Blachman B A (ed.) *Cognitive and Linguistic Foundations of Reading Acquisition and Dyslexia*. Lawrence Erlbaum Associates, Hillsdale, NJ

Frith U, Wimmer H, Landerl K 1998 Differences in phonological recoding in German and English speaking children. *Scientific Studies of Reading* **2**: 31–54

Gallagher A, Frith U, Snowling M J 2000 Precursors of literacy-delay among children at genetic risk of dyslexia. *Journal of Child Psychology and Psychiatry* **41**: 203–13

Hatcher P J, Hulme C 1999 Phonemes, rhymes and intelligence as predictors of children's responsiveness to remedial reading instruction: evidence from a longitudinal intervention study. *Journal of Experimental Child Psychology* **72**: 130–53

Hogben J H 1997 How does a visual transient deficit affect reading? In: Hulme C, Snowling M (eds.) *Dyslexia, Biology and Cognition*. Whurr, London

Lefly D L, Pennington B F 1996 Longitudinal study of children at family risk for dyslexia: The first two years. In: Rice M L (ed.) *Toward a Genetics of Language*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 49–76

Pennington B F, Grossier D, Welsh M C 1993 Contrasting cognitive defects in attention deficit hyperactivity disorder vs. reading disability. *Developmental Psychology* **29**: 511–23

Rourke B P 1995 *Syndrome of Nonverbal Learning Disabilities*. Guildford Press, New York

Seymour P H K 1986 A *Cognitive Analysis of Dyslexia*. Routledge, London

Share D L 1995 Phonological recoding and self-teaching: sine qua non of reading acquisition. *Cognition* **55**: 151–218
Snowling M J 2000 *Dyslexia*. Blackwell, Oxford, UK
Snowling M J, Stackhouse J 1996 Dyslexia. *Speech and Language: a Practitioners Handbook*. Whurr Publishers, London
Stanovich K E, Siegel L S 1994 The phenotypic performance profile of children with reading disabilities: a regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology* **86**: 24–53
Toija G A 1999 Phonological awareness intervention research: a critical review of the experimental methodology. *Reading Research Quarterly* **34**: 28–52

M. J. Snowling

Dyslexia, Genetics of

Dyslexia is a developmental language disorder which mainly affects the acquisition of reading and spelling skills. Dyslexia is an interesting example of the intersection between an evolved behavior (language) and a cultural invention (literacy). While there cannot be genes for reading or other relatively recent cultural inventions (consider agriculture, banking, and football), there can be genetic influences on evolved cognitive and behavioral traits necessary for proficiency in such cultural inventions. Because there is now extensive evidence of genetic influences on individual differences on most domains of cognition and behavior, it is not surprising that there are genetic influences on reading and spelling skills. In what follows, we will first consider how the dyslexia phenotype is defined, then consider genetic studies of dyslexia, and conclude with how these genetic discoveries will help us understand the development of this disorder.

1. *Definition of the Phenotype*

In the case of complex behavioral traits like dyslexia, careful cognitive analysis is often needed before the genetic dissection begins. Without such analysis, the phenotype utilized in genetic studies will be much coarser and more likely to be influenced by numerous genes and environmental influences. Genetic studies of dyslexia have built upon extensive research on the psychology of reading. This research has pinpointed which of the many cognitive components of skilled reading are most important for individual differences in reading proficiency. Some of the cognitive components of skilled reading include eye movements, visual perception, language processing, reasoning, and memory. In neuroimaging studies of skilled reading, many parts of the brain 'light up,' including parts associated with the cognitive components just listed. The key question then is whether all of these cognitive components are equally important for individual differences in learning to read or whether a cognitive explanation of most of reading difficulty can be provided by a much smaller set of components. If the former situation were the case, we would want to divide individuals with reading difficulty into many different cognitive subtypes before undertaking genetic studies. On the other hand, if the latter situation obtains, genetic studies would be more feasible.

Dyslexia was first described 100 years ago by Pringle-Morgan (1896), but real advances in our understanding of its cognitive phenotype have only come in the last 20 years (see Shaywitz 1996 and Snow et al. 1998 for reviews). Earlier theories of dyslexia focused on the reversal errors made by individuals with dyslexia, such as writing **b** for **d** or **was** for **saw**, and postulated a basic deficit in visual processing. Orton (1925) termed this deficit 'strephosymbolia,' which means 'twisted symbols.' The recent advances in our understanding of the cognitive phenotype in dyslexia have built upon basic research focused on both human speech and skilled reading. Although the goal of reading is comprehension, which depends on many higher cognitive processes such as reasoning and memory, it turns out that a very substantial proportion of the variation in reading comprehension can be accounted for by individual differences in the accuracy and speed of single printed word recognition. So, most poor readers have a primary problem with recognizing individual printed words, not with reading comprehension. Not too surprisingly, problems with printed word recognition are almost inevitably accompanied by problems with spelling, since both word recognition and spelling depend on understanding the relations between letters and sounds. Poor readers with these primary problems in word recognition and spelling that are not due to a vision or hearing problem, or a lack of instruction, are called dyslexics. The smaller subset of poor readers with a primary problem in reading comprehension and not in word recog nition are thus not included in the definition of dyslexia.

Single word recognition skill is highly related to phonological coding skill, the ability to pronounce letter strings that have never been seen before (usually measured by having a subject read pseudowords). A reader who can pronounce pseudowords demonstrates an implicit understanding of the quasi-regular correspondences that exist between written letters and spoken sounds. Without this implicit understanding, a reader would have to depend on pure rote memory to read or spell a very large number of seemingly unrelated items (about 100,000 or more in the case of a reader of this encyclopedia). Of course, the pronunciation and spelling of some exception words, such as *yacht*, have to be memorized. The ability to learn word-specific pronunciations and spellings is called

orthographic coding skill. Extensive research has documented that dyslexics are less accurate in reading pseudowords (phonological coding) than even younger, normal readers matched on single, real word reading accuracy. While the majority of developmental dyslexics also have deficits in orthographic coding skill, it is not clear that there is a subtype of dyslexia with only a selective deficit in orthographic coding. So, the vast majority of dyslexics have a deficit in phonological coding that at least partially explains the key defining symptoms of their disorder, slower and less accurate single word recognition and spelling.

But what are these speech sounds that the phonological code represents? A skilled reader knows that the spoken word *bag* is spelled with three letters and that these letters correspond to the three 'speech sounds' in the spoken word. However, a sound spectrograph of this or other spoken words will not reveal discrete speech sounds because of the phenomenon of coarticulation: the articulatory gestures for each speech sound overlap in their production, allowing us to talk much faster than would be the case if we produced each speech sound separately. Linguists call these speech sounds 'phonemes,' explicit awareness of which is called 'phoneme awareness,' which can be measured by tasks which require a subject to count or manipulate the phonemes in a word. Interestingly, one can learn to talk without phoneme awareness, because spoken language can be processed in larger 'chunks,' such as words or syllables. But one cannot learn to read, unless the language one is learning to read is a syllabary, like Japanese Kana. Linguistic theory predicts, and extensive empirical research has confirmed, that the ability to segment these phonemes in spoken speech is a prerequisite for learning to read an alphabetic language like English. If the alphabet is a code and if the letters correspond to phonemes, then how could one understand that code without phoneme awareness? Thus, one might expect deficiencies in phoneme awareness to underlie dyslexia. Indeed, numerous studies have found that dyslexics across the lifespan have deficits in phoneme awareness, which appear to have their origin in sometimes subtle problems with early language development. However, deficits in phoneme awareness are not uniquely predictive of later dyslexia, since other measures of phonological development predict about as well (Snow et al. 1998). Finally, it should be acknowledged that the relation between reading and phoneme awareness is a reciprocal one; while phoneme awareness is a prerequisite for normal reading, reading experience also facilitates phoneme awareness.

In sum, contrary to intuition, individual differences in reading skill have more to do with speech than with vision, and at least as much to do with single word processing as with the processing of connected text. Once we focus on poor readers with primary problems in word recognition, the vast majority are found to have problems in phonological coding and phonological development. So, the cognitive explanation of most of reading difficulty involves many fewer cognitive components than those involved in skilled reading. These facts about the dyslexia phenotype have facilitated genetic studies which have built on this cognitive science understanding of normal and abnormal reading. If these linkage studies of dyslexia had used letter reversals, or even reading comprehension, as the phenotype, it is unlikely that as much progress would have been made.

It is important to emphasize that few of the complex behavioral phenotypes now being tested for linkage have an equally mature cognitive science analysis to build on. For example, other substantially heritable childhood disorders, such as autism and attention deficit hyperactivity disorder, and most adult psychiatric disorders, are less well understood than dyslexia at the cognitive level of analysis. How much of a barrier this will prove to be for genetic studies remains to be seen, but it could be a substantial one.

2. *Genetic Mechanisms*

In the last two decades, our understanding of the etiology of dyslexia has increased considerably due to advances in both behavioral and molecular genetics. For fifty years after it was first described by Pringle-Morgan (1896) evidence for recurrence in families was repeatedly documented in case reports, leading Hallgren (1950) to undertake a more formal genetic epidemiological study of a large sample of families. Besides conducting the first test of the mode of transmission, his comprehensive monograph also documented several characteristics of dyslexia that have recently been rediscovered: (a) the sex ratio is nearly equal, about 1.5 M:F and (b) there is not a significant association between dyslexia and non-right handedness. Hallgren also documented that dyslexia co-occurs with other language disorders; however, the degree and basis of this comorbidity has not yet been fully worked out, although it fits with the view that dyslexia is a developmental language disorder.

Although Hallgren and his predecessors provided considerable evidence that dyslexia is familial, it has taken modern twin studies (see DeFries and Gillis 1993 for a review) to demonstrate that this familiality is substantially genetic and modern linkage studies to actually begin to locate the genes involved. Unlike the situation in Hallgren's time, we now have very strong, converging evidence that dyslexia is both familial and heritable (see Pennington and Gilger 1996 for a review). We can also reject the hypotheses of classic, X-linked or simple recessive autosomal transmission, at least in the vast majority of cases. We also have evidence that dyslexia is genetically heterogeneous. Perhaps most importantly, we have evidence that supports Hallgren's observation that what looks like autosomal dominant transmission occurs in many

dyslexic families. So there do appear to be effects of loci with sizable effects, acting in a dominant or additive fashion, on the transmission of reading problems.

However, we can place several important constraints on Hallgren's hypothesis of a single autosomal dominant gene influencing dyslexia. First, it is very unlikely to be one gene, because of the evidence for genetic heterogeneity. Second, it may not be a gene influencing dyslexia *per se*, since the familiality, heritability, and transmission results for normal variations in reading skill are not clearly different from those for dyslexia. Hence, the same *loci* may be involved in the transmission of both normal reading skill and dyslexia. If this is true, then dyslexics would just have more of the unfavorable alleles at these loci and/or more environmental risk factors, such that their reading scores are pushed beyond the cutoff for dyslexia. In this case, the locus (or loci) is not a necessary 'disease' locus, but is instead better conceptualized as a susceptibility locus. A susceptibility locus, unlike a disease locus, is neither necessary nor sufficient to produce the disorder in question. If a susceptibility locus influences a continuous (as opposed to categorical) trait, then it is called a quantitative trait locus (QTL). Complex behavioral traits are more likely to be influenced by several QTLs than a single Mendelian locus.

So instead of a classic, autosomal dominant 'disease' gene, which is rare in the population and which is by itself necessary and sufficient to produce the disorder of dyslexia, we are likely dealing with several, more frequent quantitative trait loci (QTLs), which are involved in the transmission of both dyslexia and normal variations in reading skill. No one QTL is likely to be necessary to produce dyslexia. Whether one QTL has an effect size sufficient to produce dyslexia is an empirical question which only linkage methods can answer.

In summary, several hypotheses about the transmission of dyslexia can be rejected on the basis of available data reviewed in Pennington and Gilger (1996). (a) Dyslexia is not an X-linked disorder, and there is little evidence of parental sex effects on transmission. There is converging evidence for sex differences in penetrance, which would produce the slight preponderance of males (1.5 M:F) that is observed. (b) Simple polygenic/multifactorial transmission can be rejected because there is a major locus effect in several samples. This major locus effect acts in an additive or dominant, but not in a recessive fashion. (c) A monogenic hypothesis can be rejected because dyslexia is genetically heterogeneous. (d) A necessary 'disease' allele hypothesis can be rejected because there is evidence for a major locus effect on the transmission of normal variation in reading skill. A remaining hypothesis that fits the empirical data considered here is that a small number of QTLs underlie the transmission of both dyslexia and normal variations in reading skill.

Given the strong possibility that the loci contributing to dyslexia are QTLs, and given the evidence for genetic heterogeneity, traditional linkage analysis (of large extended dyslexic families) is not the most appropriate method to identify these loci. Instead, sibling pair linkage analysis is more appropriate. By selecting sibling pairs in which at least one sibling has an extreme score, one can perform linkage analyses which screen for genetic loci influencing extreme scores on a continuous measure.

Using such a method, evidence for a QTL on the short arm of chromosome 6 was found across two independent samples of sib pairs and across two sets of genetic markers (Cardon et al. 1994). Each sample gave significant evidence of a quantitative trait locus located in a two centimorgan (2cM) region of 6p22.3-21.3.

This finding has now been replicated in three other samples and by two other laboratories (Fisher et al. 1999, Gayan et al. 1999, Grigorenko et al. 1997), although a fourth laboratory using a different phenotype definition did not replicate this result (Field and Kaplan 1998). The Grigorenko et al. study also found some evidence of differential genetic effects as a function of which cognitive component of reading was used to define the phenotype. However, Fisher et al. and Gayan et al. both found that the QTL on chromosome 6 was linked to deficits in both of the two main cognitive components of word recognition, phonological and orthographic coding. This result is consistent with behavioral genetic analyses of these two components, which have found that they have a large, although not complete, genetic overlap (Olson et al. 1999). Work is now under way to identify the actual gene in this region, to screen the remainder of the genome to identify other QTLs that influence dyslexia, and to test whether they differentially influence cognitive components of reading, such as phonological and orthographic coding.

3. *Implications of the Genetic Findings*

These exciting linkage findings may eventually allow us to address how much variance in reading scores these genetic loci account for, and eventually how frequently unaffected siblings have unfavorable alleles at these loci. If a similar, sib pair linkage study were conducted using probands selected for extremely high reading scores, we could determine whether different alleles at the same loci influence exceptionally good reading. If so, we could conclude that the same quantitative trait loci are affecting reading scores across the whole distribution. If they are, then our speculation that the same genes are influencing normal and extreme individual differences in reading would be supported. If not, then we would have direct evidence that dyslexia is etiologically distinct. Once we have a better understanding of these genetic mechanisms, we

can also conduct much more revealing studies of environmental factors, both risk factors and protective ones, which are also undoubtedly operative in the transmission of both abnormal and normal reading skill.

These linkage results also permit a direct test of the etiology of the comorbidity of dyslexia with other developmental disorders, such as attention deficit hyperactivity disorder and speech and language disorders. Most importantly, once these genes are clearly identified, we can begin to trace the dynamic, developmental pathway that runs from gene to brain to behavior.

Finally, it is important to discuss how the finding of moderate genetic influence on dyslexia and the approximate location of some of the genes involved influences how we treat this common disorder. Here it is important to emphasize that the issues of etiology and treatability are separate ones. Traits under a high degree of genetic influence can nonetheless be quite malleable because the reaction range of the genotype is broad. Many non-geneticists are resistant to genetic explanations because they fear that a genetic etiology means that the disorder is immutable. Instead, exactly the opposite is often the case; understanding the genetics of a disorder can lead to fundamental advances in its treatment. Although we do not currently know whether the linkage results for dyslexia will lead to changes in its identification or treatment, we already know that the phenotype is malleable and that training can significantly improve the educational outcome of children with dyslexia.

See also: Brain Development, Ontogenetic Neurobiology of; Brain, Evolution of; Developmental Behavioral Genetics and Education; Developmental Psychopathology: Child Psychology Aspects; Dyslexia (Acquired) and Agraphia; Dyslexia, Developmental; Dyslexia: Diagnosis and Training; Information Processing and Psychopathology; Reading, Neural Basis of; Reading Skills

Bibliography

Cardon L R, Smith S D, Fulker D W, Kimberling W J, Pennington B F, DeFries J C 1994 Quantitative trait locus for reading disability on chromosome 6. *Science* **266**: 276–9

DeFries J C, Gillis J J 1993 Genetics of reading disability. In: Plomin R, McClearn G E (eds.) *Nature and Nurture*. American Psychological Association, Washington, DC

Field L L, Kaplan B J 1998 Absence of linkage of phonological coding dyslexia to chromosome 6p23-p21.3 in a large family data set. *American Journal of Human Genetics* **63**: 1448–56

Fisher S E, Marlow A J, Lamb J M E, Williams D F, Richardson A J, Weeks D E 1999 A quantitative-trait locus on chromosome 6p influences different aspects of developmental dyslexia. *American Journal of Human Genetics* **64**: 146–56

Gayan J, Smith S D, Cherny S S, Cardon L R, Fulker D W, Brower A M, Olson R K, Pennington B F, DeFries J C 1999 Quantitative-trait locus for specific language and reading deficits on chromosome 6p. *American Journal of Human Genetics* **64**: 157–64

Grigorenko E L, Wood F B, Meyer M S, Hart L A, Speed W C, Shuster A, Pauls D L 1997 Susceptibility loci for distinct components of developmental dyslexia on chromosomes 6 and 15. *American Journal of Human Genetics* **60**: 27–39

Hallgren B 1950 Specific dyslexia: A clinical and genetic study. *Acta Psychiatrica Neurological Scandinavian* (*Supp*) **65**: 1–287

Olson R K, Datta H, Gayan J, DeFries J C 1999 A behavioral-genetic analysis of reading disabilities and component processes. In: Klein R M, McMullen P A (eds.) *Converging Methods for Understanding Reading and Dyslexia*. MIT Press, Cambridge, MA.

Pennington B F, Gilger J W 1996 How is dyslexia transmitted? In: C H Chase, G D Rosen, G F Sherman (eds.) *Developmental Dyslexia: Neural, Cognitive, and Genetic Mechanisms*. York Press, Parkton, MD, pp. 41–61

Pringle-Morgan W P 1896 A case of congenital word-blindness (inability to learn to read). *British Medical Journal* **2**: 1543–4

Shaywitz S 1996 Dyslexia. *Scientific American* **277**: 98–104

Snow C E, Burns M S, Griffin P 1998 *Preventing Reading Difficulties in Young Children*. National Academy Press, Washington, DC

B. F. Pennington

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

E

Early Childhood: Socioemotional Risks

Socioemotional development in early childhood pertains broadly to the quality of children's evolving participation in relationships, with the parent–child relationship serving as a primary focus in early childhood. Socioemotional development takes into account normative achievements in emotional and cognitive developmental domains, and parental and child characteristics that contribute to individual differences in the quality of children's emotional functioning and social behavior. This article is concerned with individual differences in the quality of young children's socioemotional functioning, and selected risks to such functioning from the vantage point of attachment theory and research. Indeed, security of attachment represents a central means of conceptualizing and assessing socioemotional functioning in early childhood.

1. *Quality of Attachment and Socioemotional Functioning*

The reader is referred to *Attachment Theory: Psychological* for detailed treatment of the emergence of attachment theory and description of attachment classifications in infancy. Mary Ainsworth's (Ainsworth et al. 1978) identification of the three basic attachment patterns in infancy (secure: group B; insecure-avoidant: group A; and insecure-ambivalent: group C), and Main and Solomon's (1986) addition of the insecure-disorganized/disoriented category (group D), significantly revolutionized thinking about the roots of socioemotional functioning in childhood. Whereas sensitive parental care has been found to predict group B classifications, rejecting, intrusive caregiving has been linked to group A classifications, and inconsistent and/or noncontingent caregiving has been predictive of group C attachments (Cassidy and Berlin 1994, Main 1981). Group D attachments have been linked to highly pathological caregiving such as child abuse and other forms of severe parental inadequacy resulting from parental psychopathology or drug and alcohol abuse (Solomon and George 1999). Among the insecure attachment groups, group D may represent the most insecure and the most at risk for the development of serious behavior problems in later childhood. In low-risk samples, 60–65 percent of infants are typically classified as secure, 15–20 percent as insecure-avoidant, 10 percent as insecure-ambivalent, and 10–15 as insecure-disorganized.

2. *Quality of Attachment and Later Relationships*

Central to attachment theory is the development of internal working models of relationships, which emerge with the advent of more sophisticated cognitive and memory capabilities as children move from sensorimotor to representational modes of functioning. Internal working models are believed to exert a powerful organizational influence on the manner in which information from the social environment is processed and interpreted, and on one's behavior toward attachment figures and peers. Significant predictive relations are thus expected between quality of attachment in infancy and the quality of children's subsequent relationships with peers. Secure children, who are likely to experience a history of caregiving empathy, warmth, and sensitivity, should be more likely to manifest such characteristics in interactions with peers and adults than children experiencing a history of caregiver rejection or inconsistency. Further, because of their history with a sensitive, responsive caregiver, secure children will be more likely than insecure children to organize their behavior around caregivers, and to enter into positive, cooperative relationships with them. Indeed, children classified as secure in infancy have later been found as toddlers and preschoolers to be more empathic and socially competent with peers, more compliant and cooperative with mothers, and more sociable with unfamiliar adults than children classified in infancy as insecure. Among the four Ainsworth–Main infancy classifications, the insecure-disorganized category appears to be particularly associated with more severe relationship disturbances in older children. Children classified as disorganized in infancy have been found to show heightened levels of disruptive behavior toward peers, and role-reversing, controlling behavior toward caregivers at age six. These findings are consistent with work finding disorganization in infancy to be linked with significant caregiving pathology. The extent to which early attachment security can predict the quality of later relationships with parents and peers is likely to depend on the stability of parental care over time (Lamb et al. 1985). Few would argue, however, that early security is inconsequential. Indeed, a basic tenet of attachment theory is that children who enjoy a secure relationship with the

primary caregiver are better equipped psychologically to negotiate and resolve subsequent developmental challenges (e.g., establishing a sense of independence and mastery in the preschool years) than children who are insecure.

3. Risks for Insecure Attachment

Given the connection between quality of parenting and security of attachment, it follows that specific threats to the quality of parenting will place young children at risk for insecure attachment and, in turn, for socioemotional difficulties in the wider world. The following section identifies risks as derived from three domains: the parent, the parent's social-ecological niche, and the child. What follows is by no means an exhaustive list. Indeed, it is reasonable to conclude that any stressor that compromises parents' ability to provide adequate care places children at socioemotional risk. The following section highlights factors that have been repeatedly identified either theoretically or empirically as threats to attachment security in early childhood.

3.1 Parental Risks

3.1.1 Nonautonomous working models of attachment. Attachment theory proposes that working models of relationships continue to be revised and updated as older attachment relationships evolve and as individuals develop new attachments over time. The most significant development in the assessment of working models of relationships in adulthood is the Adult Attachment Interview (AAI) (George et al. 1985), a semistructured, autobiographical interview that inquires about adults' perceptions of their early relationships with mothers, fathers, and other attachment figures. The AAI classification system places great emphasis on the manner in which past and present attachment relationships are described and the overall coherency of the narrative. The AAI yields four attachment groupings—autonomous, dismissing, preoccupied, and unresolved—which are viewed as reflecting one's current state of mind with respect to attachment, and which map conceptually onto the Ainsworth–Main infancy classifications of secure, insecure-avoidant, insecure-ambivalent, and insecure-disorganized. Autonomous (secure) adults evaluate attachment relationships as meaningful and important, provide convincing supportive evidence for perceptions of their early attachment relationships, and take an open, balanced, and forgiving stance in portrayals of untoward experiences with caregivers, past or present. Dismissing adults tend to idealize their parents, using very positive descriptors to describe their relationships with them without convincing supportive evidence. Lack of memory, or providing accounts of experiences that belie the very positive descriptors given for parents, is characteristic of dismissing adults. Preoccupied adults are actively or (less commonly) passively preoccupied with their attachment figures. Their narratives are entangled, angry accounts of past and present experiences with caregivers, with excessive focus on perceived parental slights and transgressions, and parental character 'flaws.' Finally, unresolved adults show conspicuous lapses in the monitoring of reasoning and discourse in discussion about loss or abuse, indicative of significant mental disorganization vis-à-vis these experiences.

In a meta-analysis, Van Ijzendoorn (1995) reported that nonautonomous mothers were less sensitive and supportive in interactions with their young children than autonomous mothers. In this same report, 75 percent of infants had an attachment classification that corresponded with the analogous AAI classification of the mother. Autonomous mothers were likely to have secure infants, dismissing mothers were likely to have insecure-avoidant infants, preoccupied mothers were likely to have insecure-ambivalent infants, and unresolved mothers were likely to have insecure-disorganized infants. Thus, autonomous mothers' ability to respond to infants with sensitivity, dismissing mothers' proclivities to reject infant bids, preoccupied mothers' tendencies toward inconsistency, and unresolved mothers' frightened/frightening behavior function to organize infant attachments that are complementary to those of the parent. These findings suggest that parents with nonautonomous working models of attachment may place their children at risk for insecure attachment.

3.1.2 Parental psychopathology. There is a large literature on the debilitating effects of parental psychopathology on children's emotional well-being. Most commonly assessed has been maternal depression, probably because it is so widespread among women with small children, and this work consistently demonstrates that children of depressed mothers have elevated rates of internalizing (e.g., depression, anxiety) and externalizing (e.g., oppositional, 'acting-out' behavior) symptoms than children of nondepressed controls (Gelfand and Teti 1990).

Depressed mothering has been described as hostile, critical, disengaged, and intrusive, and thus it is not surprising that theoretical explanations for the transmission of psychopathology from depressed parent to child target the depressed parent–child relationship. Indeed, depression in mothers is associated with elevated rates of insecure attachment in infants, particularly when the depression is chronic and moderate to severe. Teti et al. (1995) found that 80 percent of infants of depressed mothers were classified as insecure (with 40 percent insecure-disorganized), compared to only 30 percent (10 percent insecure-

disorganized) in the nondepressed group. That high rates of attachment disorganization were found among infants of chronically depressed mothers is consistent with the premise that unpredictable, frightened, or frightening behavior, which has been theoretically linked to disorganization, is more probable among parents with psychiatric disturbance.

3.2 Social-ecological Risks

3.2.1 Marital discord. Much has been written about the debilitating impact of marital discord on the socioemotional well-being of young children. High levels of marital stress are associated with reduced parental sensitivity and responsiveness, and with higher rates of insecure attachment. Presumably, marital discord affects attachment security indirectly, via its impact on quality of parenting, although Davies and Cummings's (1994) emotional-security hypothesis proposes that young children's emotional security with parents is impacted directly by repeated exposure to marital conflict. Recent work has emphasized the importance of distinguishing between marital conflict that is resolved, versus unresolved marital conflict, with the latter being especially problematic for children.

3.2.2 Parental social supports. The quality of support parents receive from their broader social networks also relates to quality of parenting, although evidence for a linkage between social support and security of children's attachment to parents is not as clear as it is for marital quality. Like marital discord, low social support has been associated with unresponsive, insensitive parenting, although its relation to insecure attachment is most clearly discernible in higher-risk samples. For example, Crockenberg (1981), in her study of relations of infant temperament and parental social supports to security of infant–mother attachment, reported an association between low support and insecurity only among temperamentally irritable infants. Social support may function most clearly for parents as a buffer against the ill effects of specific life stressors, such as difficult infant temperament.

3.3 Child Risk: Difficult Temperament

Although there is no universally accepted definition of temperament, most students of temperament would agree that it refers to a set of trait-like predispositions that are genetically based, that present themselves relatively early in infancy, and on the basis of which infants can be reliably distinguished. Typical 'core' constructs of temperament include emotionality (e.g., positive vs. negative mood), activity level, reactivity to stimuli, adaptability to changes in routine, and, possibly, sociability. Intuitively, it is reasonable to expect that infants who are negative in mood and highly reactive and unadaptable to new situations would present unwelcome challenges to parents, and in turn to their developing attachment security with the parents, in comparison with infants typically positive in mood, generally adaptable, and low in reactivity. Despite research since 1980, however, the role of infant temperament in the development of infant–parent attachment remains controversial. Temperament-attachment associations are, at best, modest, suggesting that quality of attachment and temperament are not redundant constructs. It is likely that temperamental predispositions do indeed influence the development of attachment security, but indirectly, via a more direct impact on parent–child relations and parental feelings of efficacy. Parents are likely to find difficult infants more challenging and potentially stressful than easier infants, and those parents who can engage successfully with and adapt to difficult infants are theoretically more likely to foster infants' security than parents who withdraw from or overreact to these difficulties.

4. Attachment Security, Cumulative Risks, and Protective Factors

Studies of the impact of risk factors on development in older children consistently indicate that development is increasingly compromised as the number of risk factors to which a child is exposed increases. Such research, focusing on the socioemotional functioning of young children, is less common. However, Belsky (1996) recently demonstrated such a cumulative relation in a study of attachment security in infancy, which included the risk factors of parental psychological maladjustment, marital discord, difficult infant temperament, low parental social supports, high levels of work and family stress, and low social class. Studies of this kind illustrate the point that development takes place against a backdrop of risks and protective factors, the latter serving to buffer the child against the adverse effects of risks, and that both should be considered together in attempts to predict development. Interestingly, factors that are considered 'protective' are frequently at the opposite end of the same spectra as the risk factors (e.g., nondepressed vs. depressed parent, marital harmony vs. marital discord, easy vs. difficult temperament).

5. Future Directions

The specific manner in which risk and protective factors in early childhood interact to impact parenting and early attachment is poorly understood. However,

such questions are at the forefront of the field. For example, how might an easy temperament in infancy buffer the negative impact of maternal depression on socioemotional functioning? Do an easy baby's positive mood, low reactivity, and adaptability elicit less negative (or more positive) parenting from that parent, which, in turn, increases the likelihood of a secure attachment? Alternatively, does easy babies' generally high adaptability lead them to more adaptable developmental outcomes even in the face of parental inadequacy, within limits? Answering such questions should help to elucidate important mechanisms of influence on development.

See also: Attachment Theory: Psychological; Childhood and Adolescence: Developmental Assets; Conflict and Socioemotional Development; Developmental Psychopathology: Child Psychology Aspects; Infancy and Childhood: Emotional Development; Infant and Child Development, Theories of; Play and Development in Children

Bibliography

Ainsworth M D S 1989 Attachments beyond infancy. *American Psychologist* **44**: 709–16

Ainsworth M D S, Blehar M, Waters E, Wall S 1978 *Patterns of Attachment: A Psychological Study of the Strange Situation*. Lawrence Erlbaum Associates, Hillsdale, NJ

Belsky J 1996 Parent, infant, and social-contextual antecedents of father–son attachment security. *Developmental Psychology* **32**: 905–14

Bowlby J 1982 *Attachment and Loss: Vol. 1. Attachment*. Basic Books, New York

Cassidy J, Berlin L J 1994 The insecure/ambivalent pattern of attachment: Theory and research. *Child Development* **65**: 971–91

Cassidy J, Shaver P R (eds.) *Handbook of Attachment: Theory, Research, and Clinical Applications*. Guilford Press, New York

Crockenberg S 1981 Infant irritability, mother responsiveness, and social support influences on the security of infant–mother attachment. *Child Development* **52**: 857–65

Davies P T, Cummings E M 1994 Marital conflict and child adjustment: An emotional security hypothesis. *Psychological Bulletin* **116**: 387–411

Gelfand D M, Teti D M 1990 The effects of maternal depression on children. *Clinical Psychology Review* **10**: 329–53

George C, Kaplan M, Main M 1985 Adult attachment interview protocol. Unpublished manuscript, University of California

Lamb M E, Thompson R A, Gardner W P, Charnov E L 1985 *Infant–Mother Attachment*. Lawrence Erlbaum Associates, Mahwah, NJ

Main M 1981 Avoidance in the service of attachment: A working paper. In: Immelmann K, Barlow G, Petrinovich L, Main M (eds.) *Behavioral Development: The Bielefeld Interdisciplinary Project*. Cambridge University Press, New York, pp. 651–93

Main M, Hesse E 1990 Parents' unresolved traumatic experiences are related to infant disorganized attachment status: Is frightened and/or frightening parental behavior the linking mechanism? In: Greenberg M T, Cicchetti D, Cummings E M (eds.) *Attachment in the Preschool Years: Theory, Research, and Intervention*. University of Chicago Press, Chicago, pp. 161–82

Main M, Solomon J 1986 Discovery of a new, insecure-disorganized/disoriented attachment pattern. In: Brazelton T B, Yogman M W (eds.) *Affective Development in Infancy*. Ablex, Norwood, NJ, pp. 95–124

Solomon J, George C (eds.) 1999 *Attachment Disorganization*. Guilford, New York

Sroufe L A, Fleeson J 1986 Attachment and the construction of relationships. In: Hartup W W, Rubin Z (eds.) *Relationships and Development*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 51–71

Teti D M, Messinger D, Gelfand D M, Isabella R 1995 Maternal depression and the quality of early attachment: An examination of infants, preschoolers, and their mothers. *Developmental Psychology* **31**: 364–76

van Ijzendoorn M H 1995 Adult attachment representations, parental responsiveness, and infant attachment: A meta-analysis on the predictive validity of the Adult Attachment Interview. *Psychological Bulletin* **117**: 387–403

D. M. Teti

Early Concept Learning in Children

Coverage of concepts in this article is restricted to objects. Concepts are part of our explicit knowledge; they are the ideas we bring to mind when planning, thinking, and recalling; that is, they involve more than just recognizing an object as familiar or knowing how to interact with it. Over the course of a lifetime we accumulate a huge amount of explicit information about objects such as dogs and spoons, including what they look like, what kind of thing they are, what they do or is done with them, locales in which they are found, materials of which they are composed, and so forth. However, the core of a concept of a given object consists of the kind of thing it is, which means that concepts are fundamentally classificatory. For example, the core of a concept of dog is that it is an animal and the core of a concept of spoon is that it is an eating utensil. The broader concepts of animal and utensil are in turn defined in terms of what they do or is done to them. Concepts at this abstract level of animal or utensil begin to be formed early in life.

1. Traditional Views of Concept Formation

Due to the writings of the Swiss psychologist, Jean Piaget (e.g., 1936), for most of the twentieth century it was assumed that children do not begin to form concepts until late in the second year. Piaget claimed that infants know about objects only in terms of sensorimotor activity; they were thought to be able to recognize objects and to know how to interact with them, but not to be capable of classifying them or

thinking about them in their absence. Because this view was so widespread little research on the foundations of concepts in infancy was conducted. Typically, research on early concept formation was carried out on newly verbal children, which meant that concept learning and language acquisition were often confounded.

The most common words for objects in daily speech are at what is called the 'basic level' (Mervis and Rosch 1981), that is, words such as 'dog' and 'spoon' as opposed to 'animal' and 'utensil.' Because these common words for objects are the ones that children most often hear, they are the first object words that children learn. However, it was then assumed that the language reflects fairly closely their level of conceptualization. Thus, children were said first to form *basic-level concepts* such as dog and spoon and only later to form more difficult *superordinate concepts* such as animal and utensil. However, although young children do not often use superordinate words, they often overextend words such as 'dog' to refer to all animals, which does not fit well with the notion that they do not have some concept of animal.

2. Current Theory and Research

2.1 Methods of Study

As preverbal infants have come to be studied, the traditional view of concept formation has begun to change. Infant research not only removes the confound between learning concepts and learning language, it also enables examination of concept formation at a younger age, thus providing information about the earliest foundations of the conceptual system. This research effort is still young and there is a great deal yet to be discovered about the first concepts. However, several useful techniques have been developed. These include the *object-examination task*, in which infants are given a series of little models of objects from one class to examine, for example a series of animals, and then an object from another class, for example, a car.

If infants examine the car longer than still another animal, it indicates the infants think that a car is different. Using this task it was found that by nine months of age infants differentiate animals, vehicles, plants, and furniture, but make relatively few distinctions within these classes. For example, they do not distinguish dogs from cats or rabbits nor tables from chairs or beds (see Mandler 1998 for a summary). Other research showed that even younger infants are sensitive to the differences in physical appearance of different kinds of land animals as well as different kinds of furniture, so the failure to categorize them on the object-examination task indicates a failure to make conceptual divisions within these domains. Thus, even though young infants are able to perceptually categorize dogs as different from cats, they do not at first treat them as if they were different kinds of things.

The object-examination task is suitable for testing infants as soon as they begin to manipulate objects (about six months). Another test, suitable from about twelve months, is the *sequential-touching task*. This task is an instructionless version of the standard classification test in which children are instructed to put together objects that are the same kind of thing. In the sequential-touching task, infants are presented with a number of objects from two classes simultaneously, and encouraged to play with them. The number of times infants touch objects from the same class in a row (run length) is measured and compared to chance run lengths. This test provides data similar to that from the object-examination test, indicating that even in the second year infants are not making many subdivisions in the large classes of animals, vehicles, furniture, and utensils.

A particularly useful test of early concept formation is the *generalized imitation task*, suitable from about ten months of age. Infants are shown an event acted out with little models, such as a dog being given a drink from a cup. They are then encouraged to imitate what they have just seen, but are given the cup and two different objects (for example, a different dog and a cat, or a rabbit and a car). Their choice of object to use to imitate drinking is recorded. This technique, by using the generalizations that infants are willing to make, is useful to test the breadth of early concepts. It is also useful to determine how preverbal infants have interpreted the events they observe. By varying the choices provided for infants to use for their imitations, one can narrow in on what they have understood. For example, if one-year-olds are shown a little model of a dog drinking and then given a different dog and a rabbit, they are equally likely to choose the rabbit as the dog as their first choice of animal to give a drink. On the other hand if the infants are given another dog and a bird, they are more likely to use the dog first for their imitations and only later give the bird a drink. This kind of finding suggests the infants interpreted the dog as a land animal, rather than as a dog or more generally as an animal. However, their willingness to give the bird a drink as a second choice, indicates the breadth of the generalizations they have made about drinking (Mandler and McDonough 1998).

2.2 Some Early Object Concepts

These different techniques converge in showing the global nature of the early concepts that have been studied. By nine months of age (and perhaps earlier) infants make distinctions among the domains of animals, vehicles, plants, and furniture, but within these domains make few distinctions at the basic level. For example, infants generalize any animal behavior to all other animals (including drinking and sleeping

to fish), and generalize car keys as applicable to any vehicle (including forklifts and airplanes). Near the end of the first year they have learned some of the functions of ordinary household objects, but still at a global level. For example, they have learned something about the function of containers (Baillargeon 1995) but they interpret both cups and frying pans as appropriate containers from which to drink (Mandler and McDonough 1998).

Gradually during the second year infants begin to narrow down their concepts of common household objects until they begin to approximate basic-level concepts. In the studies to date, the earliest differentiation of broad concepts into subclasses occurs in the vehicle domain, followed by household artifacts, with differentiation of animals and plants occurring at a much slower rate. This ordering seems likely to be culturally determined, but this remains to be studied. In any case, by the time they are about three, children are more apt to base their generalizations on basic-level classes rather than on larger ones and in that sense they are approaching adult functioning. However, right from the beginning inductive inferences are determined by conceptual meaning, rather than physical appearance. The main difference between these early concepts and those of later years is the sophistication of the theories that underlie them (see Carey 1985).

Overall, the general principle on which the conceptual system is built in the first two years is that broad concepts gradually become subdivided into finer ones (even though there may be a few more individualized concepts from an early age). This principle explains why the conceptual system (in so far as objects are concerned) is strongly hierarchical. Every time a new distinction is learned it is done on the basis of forming a subdivision in a more inclusive class. This order of acquisition may also partly determine the order of breakdown of the conceptual system following brain damage, in which basic-level concepts disappear before more foundational ones (Patterson and Hodges 1995).

2.3 The Nature of the Earliest Concepts

Relatively little is yet known about the kinds of information infants use to form their first concepts. It is known that infants parse the world into coherent objects, and make a distinction between animate and inanimate objects at least from two to three months of age. It seems reasonable to assume that the basis for this distinction lies in some appreciation of the different roles that animate and inanimate objects play in events. Infants are sensitive to the difference between caused and noncaused motion and to the contingent movements that characterize interactions among animates. Hence, the earliest concept of animal is likely to be something like an object that begins to move by itself and that interacts with other objects contingently and/or from a distance. Similarly, the earliest concept of artifact or inanimate thing may consist of an object that either does not move at all or, if it moves, only does so when another object comes in contact with it, and which does not interact with other objects from a distance. This distinction presumably forms the initial conceptual division of the world into two kinds of objects, from which the further division into more detailed concepts proceeds. In this sense the construction of the conceptual system lags behind the perceptual system, which is capable of making very fine distinctions from quite early in infancy. The conceptual system begins with a few crude, and probably event-related, distinctions and only gradually subdivides these into finer categories.

See also: Conceptual Change, Cognitive Psychology of

Bibliography

Baillargeon R 1995 Physical reasoning in infancy. In: Rovee-Collier C, Lipsitt L (eds.) *Advances in Infancy Research*. Ablex, Norwood, NJ, Vol. 9, pp. 305–71

Carey S 1985 *Conceptual Change in Childhood*. MIT Press, Cambridge, MA

Gelman S A, Wellman H M 1991 Insides and essences: early understandings of the nonobvious. *Cognition* **38**: 213–44

Leslie A M 1988 The necessity of illusion: perception and thought in infancy. In: Weiskrantz L (ed.) *Thought Without Language*. Oxford University Press, Oxford, pp. 185–210

Mandler J M 1998 Representation. In: Damon W (ed.) *Cognition, Perception, and Language, Handbook of Child Psychology*. 5th edn. J Wiley, New York, Vol. 2 5th edn, pp. 255–308

Mandler J M, McDonough L 1998 Studies in inductive inference in infancy. *Cognitive Psychology* **37**: 60–96

Mervis C B, Rosch E 1981 Categorization of natural objects. *Annual Review of Psychology* **32**: 89–115

Patterson K, Hodges J R 1995 Disorders of semantic memory. In: Baddeley A D, Wilson B A, Watts F N (eds.) *Handbook of Memory Disorders*. Wiley & Sons, Chichester, pp. 167–86

Piaget J 1936 *La Naissance de l'Intelligence Chez l'Enfant*. [Engl. transl. 1952 *The Origins of Intelligence in the Child*. International Universities Press, New York]

J. M. Mandler

Early Education and Care in the World Regions

Introduction: a New Understanding of Early Childhood

It is hardly too strong to say that societies in the past sometimes behaved as if young children under three did not engage in learning and cognitive activity.

Today, the phrase 'learning begins at birth' (Jomtiem World Conference of Education for All) conveniently sums up a sea-change in public knowledge and attitudes toward young children. There is now an increasing recognition among researchers, educators, social policymakers and the public at large that the early years constitute a critical period in the human development cycle. What was seen almost exclusively as a care period is now recognized as an important moment in the human life-cycle for brain and cognitive development (Lindsey 1998). As a result, public authorities now tend to intervene more readily in the period, either by supporting parents in their child rearing or by investing in early childhood services across the age range 0–6 years.

Countries differ widely, however, in their response to these research findings. In many countries, the care of young children is still seen as a private matter, or, from a gender perspective, considered in practice to be the responsibility of women. Young children may be included with other beneficiaries in generalized health, welfare and labor policies, but little help is given to parents either to improve the quality of private care arrangements or to reconcile work with the responsibilities of family life. In other countries, e.g., the USA, public day-care exists, but it is often of a traditional, social welfare nature, focused mainly on the health, hygiene and protection of children from low-income backgrounds. In yet others e.g., the Nordic countries, explicit policies and systems are in place to cater for all young children as a distinct group with their own specific needs and rights. Excellent early care and education services are available, organized by the public authorities on the assumption that 'professional caregiving and interaction with children can provide an environment that supports the social, emotional and cognitive development of the child,' (Swedish Ministry of Education 1999).

1. *Definition*

In its narrowest sense, early education is defined as the initial stage of organized instruction, designed primarily to introduce young children to a school-type environment' (ISCED 1999). In this definition, it is understood as a center-based (often school-based) program carried out by adequately trained staff, designed to meet the educational needs of children over 3 years of age, and prepare them for school.

For several reasons, the definition is unsatisfactory. With its emphasis on formal education and the age-group 3–6 years, it excludes most programming for young children throughout the world. It is more realistic to define early childhood education and care as any systematic intervention aiming at the development of young children in the age-group from birth to 6 years (the beginning of formal schooling in most countries). Though day-care and preschool programmes immediately come to mind, early childhood programming in this broad sense would also include interventions with strong health and nutrition components, or embrace non-formal programs and play groups with a developmental focus. In this article the term 'early education and care' will be interpreted in this broader sense.

2. *An Overview of Early Education and Care Programs in the World Regions*

2.1 *Minority High-income Countries*

In most of the high-income countries of Europe, North America, Asia, and Oceania, early education and development has been a part of government policy for decades. Provision of educational services for children over 3 years of age has predominated, but today, public or private investment in childcare services for the under-3s is increasing rapidly. The increase is generally attributed to the greatly augmented participation of women in the labor market, and, in some countries at least, to the new awareness of the importance of child development and learning.

Pre-school provision for children over 3 years of age is likewise expanding, and from the age of 4 years, ranged in coverage from 55 percent to 98 percent across the OECD countries (OECD, 1999). In a number of countries, e.g., in Belgium and increasingly in France, entitlement to a place in pre-school begins at the age of 30 months. In the year—and frequently in the 2 years—immediately preceding the age of compulsory schooling, provision in excess of 90 percent is common in high-income countries.

Significant differences exist, however, between countries in terms of how early education is conceptualized. A basic difference concerns the separation or integration of care and education. In countries where separation is practised, e.g., Australia, Greece, Ireland, Italy, Netherlands, Portugal, UK (until recently), and USA, early education predominates, with higher coverage rates, and better training and renumeration of personnel in the education sector. Preschool is considered to be the initial stage of organized instruction, while early care is relatively underdeveloped. Programs for young children are formal and institutional, organized according to the school model, although subject based teaching is generally not encouraged. Formal quality control—registration, regulation, control of curriculum, centralized staff training and inspection—is usual. France and Belgium also fit into this model, although it should be noted that in these countries public child care is well-developed and preschool begins, in general, around the age of 30 months.

In other countries, e.g., the Nordic countries, Spain, and parts of Italy, a more social and integrated

approach to care and early education is adopted. There is greater decentralization of administration and services to regional or municipal levels. Unified policies are generally formulated for children from birth to six years, and there is more mixing of ages within groups than in the education tradition. Personnel are generally trained to a satisfactory level, have adequate career paths and enjoy a professional status. Staff–child ratios are favorable and centers can rely on stable public funding. Emphasis is placed on providing a caring and stimulating environment for children, rather than on preparation for school.

The province of Reggio Emilia in Italy provides a successful illustration of integrated early childhood services. In-service training policies that reach across traditional boundaries, and the creation of centers with integrated age groups, are among the strategies used to bridge education and care. Municipally-funded advisory services, professional training, and staff development are made available to all services, whatever their target group or sponsoring body. The pedagogical approach of Reggio Emilia is even more renowned. It seeks to foster the cognitive and social development of young children through a systematic focus on symbolic representation and expression (language, movement, sketching, painting, mathematics, media…). Young children are encouraged to explore their feelings thoughts and environments through different modes of expression natural to them—'the hundred languages of children' (Edwards et al. 1993). The Reggio approach emphasizes the autonomy and self-direction of children. Within mixed age groups, children choose, negotiate and participate in long-term pedagogical projects that are often extremely challenging. There is also continuous assessment and documentation of the progress of each child, and well-developed strategies for teacher development and parental participation.

In countries in which early care and education is not seen as a right of young children, compensatory programs for young children may receive significant government subsidies, e.g., 'Head Start' in the USA, 'Sure Start' in the UK. Research has shown that early childhood education helps, in particular, children from disadvantaged backgrounds to acquire social skills and integrate effectively the early primary school classes (Schweinart et al., 1993, NICHD 1997). Targeted programs are often based on a mixed rationale that may include: the preparation of disadvantaged children for school; social intervention toward 'at risk' populations; reduction of unemployment among mothers and, hence, savings in direct welfare payments. Despite the real benefits brought by such programs, the approach may have negative effects, e.g., stigmatizing disadvantaged groups, giving the impression that they and their children are problematic and in need of educational 'intervention.' In addition, the compensatory approach may place too great an emphasis on education as an agent of social change. It can be forgotten that poor neighborhoods may need—in addition to education—significant investments in the housing stock, local industry, training and employment policies. In sum, to see early education as a sufficient response to social ills may be to divert attention from other more necessary inputs and from the equity obligations of government.

2.2 Former Soviet Eastern European and Central Asian Republics

In the immediate postwar period, after the Eastern European and Central Asian Republics were absorbed or became part of the Soviet sphere of influence, there was a rapid expansion of fullday day care, kindergarten and preschool facilities. Under communist governments, child care was organized as a state-funded, supply model. Early childhood education was seen as a means of inculcating the new values of communist society and of forming the 'socialist personality.' Enrolment rates in daycare centers and kindergartens soared, reaching, for example in former Czechoslovakia, high levels of coverage. In addition, work leave permissions were given to parents—almost always taken up by women—during the first year or 2 years after the birth of a child. However, since 1990, these countries have gained independence and undergone profound social and economic change.

The transformation has had beneficial effects on children and families, in particular, through the adherence of states both to the European Human Rights Charter and the United Nations Convention on the Rights of the Child. However, it may be suggested that from a gender perspective and the funding of early education services, the trend has been less positive. As economies struggle, unemployment has grown. The response of a number of countries has been to extend parental leave, which, in practice, takes women disproportionately out of the labor market. In addition, a marked reduction of state expenditure on early care and education may be noted. During the transition period, early childhood services were devolved to the municipalities or local authorities, and state budgets to the sector drastically reduced. The reduction has hit hardest early childhood services in rural areas, where the former rural cooperatives no longer function. Thus, in the space of 10 years, Bulgaria went from almost full early education coverage of young children in 1988, to 70 percent coverage in 1998.

Moreover, in the former socialist republics, early-years teachers and child care personnel (almost all women) suffer from low status. Wages are so meagre that many leave the profession or are obliged to seek supplementary earnings elsewhere. From a gender perspective, the low status of the profession indicates an underestimation of child-rearing and of the contribution that women make to society through this

work. In addition, the dependence of services on increased contributions from families has led to inequality of access, the poorest families being no longer able to afford services. Even when the state has waived fees for disadvantaged categories of children, e.g., for minority or Roma children, cultural expectations about education hinder participation by these groups in preschools, and reinforce weak performance in primary school. The situation has led, in Central Europe and the Balkans, to the segregation of many of these children into special education.

Signs of a reversal of the situation are now apparent in several countries. Despite recurring difficulties, Hungary, for example, is now making special efforts in favour of Roma children through the preschool system. Likewise, Romania has actually been able to expand its early childhood services over recent years, and effective efforts are being made to improve the previously disastrous situation of poor and institutionalized children. In the Czech Republic, after an initial down-sizing, the government is now making an important effort to reorganize services for young children and to improve their quality.

2.3 *The Majority World*

In the majority world, expansion is greatest for the age group 4–6 years, especially for the children in the year preceding compulsory school (UNESCO 2000). Yet, the growing need for child care should not be underestimated. Extended family help is diminishing in majority world countries, in a situation where many mothers in the rural areas are fully occupied in the rural economy or in meeting the survival needs of older children. In the cities too, in order to provide for their children, mothers are obliged to seek paid work or engage in exchanges of services, while young women, like their counterparts in the industrial world, see a career as part of their social identity. During their periods of work, often irregular, mothers must entrust the daily care and education of young children to other family members, to informal private arrangements or, where they exist, to community, voluntary, or for-profit services.

In several countries, especially in Latin America, governments now receive significant aid from the international organizations and the major development banks to provide early childhood programs (World Bank Human Development Network 1999). These agencies see early childhood programming as an integrated set of activities (including nutrition, immunization, health, care, development activities, preschool education, parent programming, reproductive health, prenatal and postnatal care, community building) spanning the whole age range from conception to the beginning of compulsory school, and including a strong family outreach. When accompanied by equitable economic and social policies, integrated early childhood programming is perceived as a promising means of tackling the poverty cycle (Sen and Brundtland 1999).

2.3.1 Sub-Saharan Africa. Tackling poverty is also part of the early childhood agenda in sub-Saharan Africa, but the challenges are infinitely greater. Due to national debt repayments and an AIDS pandemic, poverty and illness are pervasive in many African countries. Moreover, governance, macroeconomic management, basic infrastructure and public accountability are weak in many countries, and need to be radically improved before any permanent progress can be made in the educational and social fields (Oxfam 1999). Because of the crisis, early childhood intervention must take the form, in many instances, of food supplementation, primary health, and family support programs.

Yet, some governments and the international donors see the field from a broader perspective, that is, in terms of the holistic development of children and of their basic human rights. Much attention is given in the Sahel countries, for example, to investing in reproductive health care for girls and women, and to making it possible for girls to begin education as young as possible. A few countries have also made remarkable progress in the numbers of children being reached through regular preschool programs. South Africa, for example, progressed from an enrolment rate of 10 percent in 1985 to a 28 percent preschool enrolment rate in 1995, while both Senegal and Kenya have now reached an enrolment rate for preschool children of 40 percent. In addition, Kenya has trained almost half its preschool teachers, and established national training centers for preschool staff in several provinces (UNESCO 2000). Major problems remain, however: acute lack of resources and stable funding; inequitable access to services (rural populations generally do not have services); poor quality of services as many service providers have little understanding of child development; and a general absence of trained personnel.

2.3.2 Arab States. Though enrolment figures continue to rise throughout the region, the rate of coverage across countries remains extremely varied, ranging from 1 percent in Yemen, to 8 percent in Egypt, to 63 percent in Morocco, and over 70 percent in Lebanon. As in the rest of the developing world, provision is concentrated in the urban areas, and only a few governments take a leading role in funding and supporting services. The burden of early childhood programming is still borne by the international donors, the NGO sector and local

communities. In several countries, including in sub-Saharan Africa, religious leaders have successfully developed early childhood programmes, e.g., in parallel with traditional Koranic classes.

2.3.3 Asia and the Pacific. As the region is extremely diverse (with three-quarters of the world's children), generalization are difficult to sustain. Enormous contrasts exist with respect to government investment in early childhood policies. The OECD countries of the region—Japan, Korea, Australia, and New Zealand—and Singapore, have policies, coverage figures and system structures comparable to the Western European countries. In contrast, many other countries are still engaged in broadening or improving their primary education systems, leaving it to the voluntary sector—religious institutions, NGOs, and communities—to engage in early childhood programming. In consequence, great inequalities of access continue to exist. Some notable progress in the region can be seen, however, in the enrolment of 4 to 6-year-olds, especially in Malaysia (90 percent), Vietnam (ca. 70 percent) and in China (ca. 50 percent). The other large, high-population country in the region, India, has an enrolment rate of only 17 percent for children aged 3 to 6 years but a number of far-reaching food supplementation and early stimulation programs, such as the Integrated Child Development Services, exist.

2.3.4 Latin America and the Caribbean. During the last decades in Latin America, initial (preschool) education has been the fastest growing sector in education. A number of countries have now reached over 80 percent enrolment rates for children aged 4 years and upward: Barbados, Costa Rica, Chile, Cuba, Jamaica, Mexico, Peru, Uruguay. Another characteristic of the region is its broad developmental approach to early care and education, which contrasts favorably to the didactic approach found in primary education. Much emphasis is placed on nutrition, health, and the developmental needs of children, e.g., in the countrywide programs of Colombia. The approach has proven to be effective, leading to the prevention of stunting, improved health, better physical and mental performance, and enhanced educational achievement. However, as in much of the majority world, children from better-off families are more likely to be enrolled than children from families with few resources or belonging to minority groups. Urban programs still predominate, and entire rural or indigenous populations, e.g., along the Andes, have, in general, extremely weak access. Poverty, too, is a huge barrier to the meaningful participation of children in early care and education programmes in Latin America. Unlike the major countries in Asia, child poverty has actually increased in the continent over the last decade.

4. Conclusion

In summary, the situation of early childhood programming worldwide may be expressed as follows. Research consistently shows that services for families and young children have positive impacts on child development, in particular, on children from disadvantaged backgrounds. Through early intervention, early disability and learning difficulties are more readily diagnosed and prevented. Again, longitudinal research from over 20 countries shows that regular attendance by young children at organized developmental programs, before they begin compulsory schooling, enhances social and cognitive development, and influences positively later educational outcomes, such as reading ability and retention in grade (OECD 1999).

In terms of volume, early childhood services have expanded steadily in almost all countries in the world in the last ten years. Eastern Europe and Central Asia stand out as temporary exceptions, when, with the economic recession in these regions, early childhood services were reduced. In other regions, and in particular, in the industrialized world, early education and care has emerged from the private, domestic sphere. As the foundation stage of life-long education, it has become a legitimate focus of state intervention. The challenge for the coming years will be to unify care and education, and the conceive appropriate policies for the whole age-group from birth to 6 years.

See also: Child Care and Child Development; Children's Play: Educational Function; Cognitive Development in Childhood and Adolescence; Educational Policy: Comparative Perspective; Infancy and Childhood: Emotional Development; Infant and Child Development, Theories of; Infant Education; Play and Development in Children; Poverty and Child Development; Socialization in Infancy and Childhood

Bibliography

Edwards C, Gandini L, Forman G 1993 *The Hundred Languages of Children*. Ablex, Norwood, NJ

ISCED, International Standard Classification of Education Descriptors 1999 *Classifying Educational Programmes*, Organisation for Economic Co-operation and Development, Paris

Lindsey G 1998 Brain research and implications for early childhood education. *Childhood Education* **75**: 97–101

NICHD 1997 *Mother-child Interaction and Cognitive Outcomes Associated with Early Child Care: Results of the NICHD Study*. NICHD, Bethesda, MD

OECD, Organisation for Economic Co-operation and Development 1999 *Education Policy Analysis 1999*. OECD, Paris, Chap. 2
Oxfam 1999 *Education for All: A Compact for Africa*. Oxfam GB/Action Aid Position Paper, Oxford
Schweinart L J, Barnes H V, Weikart D P 1993 *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27*, High/Scope Press, Ypsilanti, MI
Sen A K, Brundtland G H 1999 *Breaking the Poverty Cycle: Investing in Early Childhood*. Inter-American Development Bank, Washington, DC
Swedish Ministry of Education 1999 *National Curriculum for Pre-schools*. Stockholm, Sweden
UNESCO 2000 United Nations Educational, Scientific and Cultural Organisation *2000 Assessment EFA Country Reports*. UNESCO, Paris
UNESCO 2000 *Global thematic review of early childhood care and development, January 1999*
World Bank Human Development Network 1999 *Education Sector Strategy*. World Bank Group, Washington, DC (See also the World Bank website for a list of countries in which the Bank is making massive investments in early childhood.)

J. Bennett

East Africa: Sociocultural Aspects

After more than a half century of modern scholarship, it is unnecessary to reconfirm the historical and cultural complexity of East African societies now encompassed by the independent nations of Kenya, Tanzania (including Zanzibar), and Uganda. From the maritime civilizations of the Indian Ocean islands and coast which have traded with the external world for at least two millennia, to the more archaic forms of economic adaptation, including the still extant gathering and hunting societies in the interior, this region displays an exceptional range of social and cultural forms. In addition to Swahili as a regional Bantu language, branches of all the other African linguistic families, Nilotic, Cushitic, and Khoisan, are also represented.

1. *Prehistory and History*

Based on finds at Kobi Fora in Kenya and Olduvai Gorge in Tanzania, generally it is now accepted that the earliest hominids emerged in East Africa some six to four million years ago. Although permanent agricultural settlements have existed for at least 10,000 years, historical records on the area extend back only to the first century of the Christian era. The original Greek manuscript, *Periplus of the Erythrean Sea*, describing the trade routes from Alexandria on the Red Sea in Egypt and then south along the East African coast, mentions Rhapta, a settlement assumed to have existed somewhere on the now Tanzanian shore line. Arabic documents from the eighth century onward are more prolific and exact, indicating the existence of the still occupied trading sites of Lamu in southern Kenya and Kilwa in southern Tanzania. This land of 'Zinj' is described as inhabited by a population engaged primarily in a maritime trade in animal skins and horns and some slaves, but particularly ivory, as well as subsisting on agriculture and domesticated herds. Subsequent Arab influence, including Islamization, comes to a halt for a time at least with the arrival of the Portuguese in the late fifteenth century and their partial conquest of the then more numerous city states dotting the littoral. However, these first Europeans were themselves expelled from the area by a resurgent Afro–Arab coalition directed from Oman on the Arabian peninsula which in turn failed to establish suzerainty over the entire area.

Subsequent European incursions, primarily British and German, which included penetration of the interior in the latter half of the nineteenth century hindered the now more extensive slave trade supporting the off-shore, particularly Zanzibari, plantation economies based upon the export of spices. This era culminates in 1885, with the Berlin Treaty that partitioned all of Africa, excluding Ethiopia, among the European powers, and in the case of East Africa between Britain and Germany. The results of World War I brought the entire area under British control until Independence in the early 1960s.

2. *Social Forms*

The existing subsocial formations in recent and contemporary East Africa can be classified in the usual manner as either mercantile or agricultural states, less politically complex agricultural chiefdoms of varying scale, noncentralized agricultural and pastoral societies, and the few remaining gathering-hunting groups. However, these categories are heuristic and ideal as most, if not all, the economies were mixed and interdependent prior to the colonial era. In addition, there was also fluidity in economic forms over time as the archaeological data indicate that some gathering/hunting peoples may have resorted at times to agriculture and then reverted to their previous mode of adaptation on numerous occasions. The same flexibility characterized pastoral peoples. Such adaptive shifts in more recent times have involved simultaneous alterations in ethnic identities that are often labels based primarily upon livelihood. Thus, for example, indigent Maasai who no longer kept cattle were reclassified by others as *il torrobo* (Maasai for the poor or hunters) or Nderobo, while in another instance a whole segment of Maasai who took up agriculture as their primary occupation in the nineteenth century

were redefined as Arusha. Thus, in addition to the complexity of social forms, the fluid basis of ethnic identities, often assumed to be fixed 'tribal' isolates, was a constant feature of the region's social history. Although there were and are cultural differences, in some instances a matter of degree and in others of kind, a more rigid colonial mentality and policies often helped to bring into existence these so-called tribes.

2.1 City States and Kingdoms

Prior to the advent of European incursions, complex economic and political entities existed at both extremes of the East African region. The coast, as noted, was intermittently dotted with numerous city states while the western-most interior, some 800 miles away, featured a series of large-scale polities referred to as the Interlacustrine Bantu Kingdoms due to their clustering around the lakes, such as Victoria and Tanganyika, in the Great Rift Valley. Most notably these included Baganda, Banyoro, Basoga, and Batoro, in contemporary Uganda, as well as Rwanda and Burundi that eventually fell under Belgian colonial control and re-emerged as separate nations with independence. These two regional extremes were linked by a long-distance commerce, including, by the early nineteenth century, some managed by interior middle-men groups, allowing all concerned to expand and flourish with the exchange of valued trade goods. The Bantu kingdoms were early on the subject of academic scholarship. However, in part due perhaps to its less romantic and more complex 'un-tribal' or 'detribalized' condition the coastal region failed to receive serious scholarly attention until after the colonial era.

Although these ancient city states, extending 1,000 miles along the coast north to Somalia and south to Mozambique, never formed a single political unit, they were linked by a common civilization. In addition to a shared maritime mercantilism, marriage alliances, and patrilineal descent, other commonalities included by the eleventh century Islam and a related life-style, as well as Swahili as a *lingua franca*. This term 'Swahili' eventually emerged as a complicated and wide-spread ethnic category on and beyond the coast as the language and Islam spread throughout East Africa before and after the colonial period. The word Swahili itself is derived from the Arabic *sahel* for coast. The language has a modicum of loan words from this source but otherwise it is Bantu in all of its essential grammatical features. As it was adopted as a first or second language by others in the region, many also identified themselves as an *mswahili* or *waswahili* (a Swahili person or people respectively), which implied not only language usage but also a form of modernization involving the adoption of Islam or even Christianity, migration to other areas of the country for a livelihood, and a modified Arab or European lifestyle, including forms of dress and architecture. Among themselves in the towns and cities they inhabit, the *Waswahili* may also admit to another ethnic affiliation reaffirming the plasticity of this sort of identification.

The agricultural bases of the interior states situated in fertile highland areas with the banana (originally from Indonesia) as a staple may have been the historical outcome of an incursion of northern cattle-keeping peoples into an area inhabited for at least a thousand years by Bantu-speaking agriculturalists. The high population densities, in the case of Baganda close to two million people, and social complexity resulted in what has been described as semifeudal states involving unequal control over resources including land and cattle, concomitant social stratification (though no caste system) and in most instances, a series of patron-client relationships spanning economic and political divides. These kingdoms which shared a host of cultural similarities in addition to the economic and social ones, competed with each other for more extensive control over the region. At the apex of the highly centralized political systems stood a king advised by a prime minister and other councilors who appointed chiefs over large districts who in turn delegated authority to headmen in villages. Although often popularly viewed as a despot, it is unlikely that the king could have achieved this level of power and authority until the introduction of firearms in the early nineteenth century via trade from the coast. The stability of the entire political/economic system was further solidified with colonialism and its application of indirect rule that permitted little if any challenges to established authority, or radical changes in the system. Although these were not the only centralized political systems in East Africa, as others of varying degrees in other highland areas existed beyond the coast (such as the Shambaai and Chagga of Tanzania), the vast majority of the population in the less watered lowland savanna and woodlands of the region, shared with pastoral peoples, were organized into smaller political units.

2.2 Agriculturalists and Pastoralists

Among the agriculturalists the principle traditional crops were millet, sorghum, and maize. In the more abundantly watered areas, relying in some instances on complex irrigation systems as among the Sonjo of Tanzania, rice and bananas were also cultivated. Cattle which play such a significant cultural role in the entire area, including among agriculturalists, and smaller live stock such as sheep and goats, were also maintained when and wherever possible. Politically this is the classic 'Tribes Without Rulers' region as termed by one of the classic anthropological volumes of the colonial period and the 'golden age' of East African ethnography. This vision of much of the social

landscape was influenced in no small part by the prior researches of E. E. Evans-Pritchard among the acephalous pastoral Nuer of the neighboring Southern Sudan, an area now racked by a civil strife pitting the Islamic Northern and more traditional Southern peoples against each other. In light of the range of societies in East Africa proper and their variability they are best represented by a limited number of examples which also reflect the interests of the ethnographers at different points in the history of East African scholarship.

The quarter of a million Sudanic-speaking agricultural Lugbara who live along the Nile–Congo divide on both sides of the Uganda–Congo border are characteristic in that they had no political offices until chiefs and subchiefs were appointed by the colonial authorities. This lack of a formal political system was somewhat unusual in light of the relatively high population density which exceeded 100 people per square kilometer in some areas in the early twentieth century but explicable in light of the absence of nucleated settlements. The basic social unit was the homestead of various sizes composed of patrilineally related kin residing in their own dwellings surrounded by their fields. This family cluster recognized the authority of a senior male elder with ritual significance due to his control over shrines dedicated to the lineage ancestors. Thus, at its core this social group had both a descent and a related religious ideology.

The Lugbara themselves conceive of their social structure in terms of clans originally historically associated with specific territories despite the dispersal of their members over time. As such, principles of descent provided an ideological rather than actual form of social organization as is the case for many other East African peoples. These recognized territories or 'tribes' may have a 'rain-maker' with ritual authority in certain matters but as indicated no formal leader. Order was maintained by compromise and conciliation among the resident elders. Disputes among members of different territorial groups were difficult to mediate and as such could result in violence and a state of feud. On the other hand, since clan kin were dispersed among the various Lugbara tribal units and marriages were also effected across these lines resulting in affinal relationships, violence was not entered into easily, especially among contiguous territories. More distant Lugbara tribal sections presented less opportunity for conflict and if a state of feud did exist, it was difficult to act upon as the result of spatial separation. Thus, there was a political order, despite anarchy in the technical sense of the term, among these people who recognized a common moral code supported by the ritual authority of elders and ancestors.

The manner in which social order or stability was maintained among a people lacking a well-defined political system was one of the driving research concerns during the colonial era, while change became a principal interest of postcolonial scholarship. This shift in orientation is exemplified by the situation and description of the Iteso, Nilotic-speaking sedentary cultivators residing in this instance on both sides of the Kenya–Uganda border. By the time of their analysis the Iteso were almost all Christians, and their expansionist tendencies, as well as the former political system which had permitted this process, had been radically altered by the often violent imposition of colonialism. According to historical reconstruction, prior to European intervention the Iteso were organized into a number of territorial sections under the leadership of a single individual or 'war leader.' These units were broken down into a number of subsections that tended to correspond with the residence of patrilineally related males who also recognized a local leader for political as well as ritual purposes. These locations undertook common defense, the peaceful settlement of disputes, and collective religious activity. Typical of Nilotic speakers, the territories were crosscut by an age-grade system including a warrior class which had facilitated Iteso incursions into the land of other ethnic groups. During the colonial period appointed chiefs replaced the former leaders and the territorial units dissolved as neither expansion nor defense were relevant concerns. These aggregations were replaced by diffuse neighborhoods of a smaller social scale composed of cooperating but individual households now engaged in the cash economy. The resulting process generated a typical East African society brought into being by the colonial presence. Similar processes of ethnic consolidation and consciousness among the more numerous Kikuyu and Luo peoples of Kenya during the colonial era had a significant impact on postcolonial politics.

The primarily pastoral Nilotic-speaking peoples inhabiting a grassland expanse in both arms of the vast Rift Valley in eastern Uganda, western Kenya, and north-central Tanzania provide a striking alternative to agricultural societies in more than economic fashion as many of the representatives of this language family have preserved their traditional cultural system to this day. Included in this category are the Jie, Karimojong, and Turkana of Uganda, the Kipsigis, Nandi, and Samburu of Kenya, and the Barabaig of Tanzania. However, the quarter of a million Maa-speakers or Maasai who have occupied the central plains of both Kenya and Tanzania for the past 400 years are best known to travelers, tourists, and scholars of the region for the persistence of their traditional ways despite impositions placed on their life style by both the colonial and independence regimes during the twentieth century.

The enormous expanse of Maasailand is divided into numerous large and politically autonomous named territorial sections in some instances covering thousands of square kilometers including tens of thousands of individuals. Within these sections, social order is maintained at the local level by male elders with either ritual authority derived from membership

in a particular clan, seniority in the age grade, or local community of kin. Although patrilineal in terms of descent principles, kinship ties of various types are a significant factor only at a settlement composed of a number of usually polygynous households that herd their animals together under the direction of a senior male. The most striking feature of the social system is the age grade system bisecting the entire expanse of Maasailand which includes the moran, formerly a warrior class of young men but now primarily responsible for cattle herding.

For the Maasai, cattle are not only the basic foodstuff in the form of milk, meat, and blood but also the most important cultural idiom. The animals are valued in and of themselves as esthetic items, the object of admiration, discussion, song, and dance, but they also mark both existing social relations through their exchange, and the passage of time through inheritance. They are circulated as bride-wealth and blood-wealth, in compensation for violence, and loaned out to establish herding partnerships over wide areas. Finally, these animals are the primary means for communicating with the supernatural world as sacrificial offerings to the creator in religious ceremonies. Thus, this single item in all its capacities links humans both to each other and to their deity in a manner unavailable to cultivators. In more practical economic terms the ability of the Masaai and other pastoralists to sell off some animals for cash to meet the demands of the state in the form of taxation has permitted them to avoid further integration with the wider society and in doing so conserve existing cultural patterns. However, the alienation of their former land by both the independent and colonial regimes for the sake of agriculture and agriculturalists, as well as for the creation of national parks for tourism, has had a severe impact on these people.

2.3 Foragers

Scattered among the pastoralists and agriculturalists in both forests and plains are pockets of hunters and gatherers, such as the well-known Mbuti Pygmies of western Uganda and the Congo. Other foraging groups, though physically distinct from the Mbuti, include the Okiek of Kenya, the Ik of Uganda, and the Sandawe and Hadza of Tanzania. The forest-dwelling Mbuti and related populations generally are considered to be descendants of the first residents of this and other parts of Africa who have adopted the languages of their later arriving and now Bantu neighbors, while the Okiek and Ik speak the related Nilotic languages of their pastoral neighbors. However, Sandawe and Hadza bear some phonetic similarities to the tongues of hunters and gatherers in Southern Africa. Despite the traditional academic classification on the basis of foraging, there are also variations among these peoples including the tendency to mix this adaptive strategy with agriculture and herding when feasible. Thus, a portion of the few hundred Hadza residing in a desolate region of Northern Tanzania unsuitable until recently for either agriculture or herding are interesting in that they maintained a foraging lifestyle in pure form, as well as a significant degree of economic independence, from surrounding groups and the state, even after Independence.

Although seemingly inhospitable to others, the Hadza territory of over a thousand square miles, open to any member of the group as their are no recognized territories, is replete with natural resources harvested on a nomadic and individualistic basis. Although the area is heavily populated with game, the Hadza rely mainly on wild vegetable matter for food in the form of roots, berries, and fruit gathered daily by both males and females, the latter with their small children. These nutrients usually are consumed immediately rather than being collected for later distribution. Traditional bow and arrow hunting, however, is exclusively a male activity. If the kill is small it is also eaten on the spot, while some of the remains of a larger animal may either be transported back to camp or simply abandoned. Thus, there is almost complete economic independence even among the Hadza themselves. As such, it would be misleading to characterize their local aggregations as bands in the sense of groups recognizing a common territory or cooperative bonds. Beyond the nuclear family, based on relatively unstable monogamous unions, the larger and only unit is a camp, occupied for no more than a few weeks, and on average by a dozen or more adults. This aggregation does not have a recognized leader nor form a stable unit, as some members will depart and others join it on a constant basis. At other times it may split off into independent units or merge with other camps for a period of time. This organizational fluidity of foraging societies is perhaps their characteristic feature in both East Africa and other parts of the world.

3. Cultural Concerns

This overview of East Africa in terms of economic, political, and social forms in no way suggests a paucity of equally, if not more, complex cultural systems including beliefs, values, myth, history, and literature both oral and written. These integrated systems of thought, often shared over a wide area, even in terms of particulars, defy simple categorization into a number of ideal types. In general, the peoples of this region, regardless of their form of economic adaptation must be considered highly spiritual, whether in terms of their adherence to the local religions or the world religions that have supplanted them. In turn, these cosmologies have been transformed by local beliefs and practices. As in other parts of the world, the traditional ideological systems took a great interest

in notions of good and evil and thus acceptable standards of interpersonal morality guarded by the living as elders and/or the deceased as concerned ancestors. The 'problem of evil' in society was often represented (by agriculturalists in particular) as a matter of witchcraft or sorcery responded to by recourse to diviners to identify the culprit in order to restore a harmonious social order through confession and reconciliation.

As indigenous philosophies these systems of thought also reflected on the meaning and passage of life and the significance of death, occasions usually marked by elaborate ritual gatherings and activity. Communication with their creator, most often viewed as a more distant and less knowable figure than in world religions, traditionally was maintained through prayer and sacrifice. Indeed as this review of primarily social forms suggests, it was and is often best to view many East African peoples at the local level and in the most meaningful sense as composed of ritual communities concerned with the spiritual and social welfare of their members.

See also: African Studies: Culture; African Studies: History; African Studies: Politics; African Studies: Religion; Age: Anthropological Aspects; Central Africa: Sociocultural Aspects; Colonialism: Political Aspects; Colonization and Colonialism, History of; Ethnicity: Anthropological Aspects; Historiography and Historical Thought: Sub-Saharan Africa; Hunting and Gathering Societies in Anthropology; Pastoralism in Anthropology; Southern Africa: Sociocultural Aspects; Sub-Saharan Africa, Archaeology of; West Africa: Sociocultural Aspects

Bibliography

Århem K 1985 *Pastoral Man in the Garden of Eden*. Uppsala Research Reports in Cultural Anthropology, Uppsala, Sweden

Arens W 1975 The Waswahili. *Africa* **45**: 426–37

Arens W (ed.) 1976 *A Century of Change in Eastern Africa*. Mouton, The Hague, The Netherlands

Beattie J 1971 *The Nyoro State*. Clarendon Press, Oxford, UK

Evans-Pritchard E E 1940 *The Nuer*. Oxford University Press, London

Feierman S 1990 *Peasant Intellectuals*. University of Wisconsin Press, Madison, WI

Ingold T, Riches D, Woodburn J 1988 *Hunters and Gatherers*. Berg, Oxford, UK

Karp I 1978 *Fields of Change among the Iteso of Kenya*. Routledge & Kegan Paul, London

Karp I, Bird C (eds.) 1980 *Explorations in African Systems of Thought*. Indiana University Press, Bloomington, IN

Middleton J 1992 *The World of the Waswahili*. Yale University Press, New Haven, CT

Middleton J, Tait D (eds.) 1958 *Tribes without Rulers*. Routledge & Kegan Paul, London

Oliver R, Mathew G (eds.) 1963 *History of East Africa*, Vol.1. Clarendon Press, Oxford, UK

Phillipson D 1993 *African Archaeology*, 2nd edn. Cambridge University Press, Cambridge, UK

Spear T, Waller R 1993 *Being Maasai*. Currey, London

W. Arens

East Asia: Environmental Issues

1. Human Beings and the Environment

Environment means what surrounds human beings on the outside. It can be divided into the natural and the social environment. The natural environment is the air, water, mountains, the sea, soil, plants, animals, and the earth as a whole. The social environment is the family, the school, the firm, the community, the social stratification, the market, the state, and global society as a whole.

When we say the 'environmental problem,' the central interest is of course on the natural environment. While human beings have, since ancient times, utilized the natural environment as a resource, the degree to which they exploit natural resources increased rapidly after the Industrial Revolution, during which they obtained the industrial technology through which to transform nature as they wished. The other side of the coin of this industrial technology, however, was to break the natural environment of unanticipated consequences. Thus, today human beings are confronted with the environmental problems such as pollution of the air, water, soil, and sea, destruction of the ecosystem of the plants and animals, ozone layer depletion, and global warming.

On the other hand, however, the social environment has always something to do with it. This is because it is the human social action of production and consumption that deteriorates the natural environment, unless it is caused by earthquake, volcanic eruption, or typhoon. In this sense the destruction of the natural environment is connected closely with the ways in which human social actions are organized by the social environment. The social environment is, sociologically speaking, of importance as the socializing agent that brings people to conform to environmental ethics. This means that the degree to which the natural environment is polluted is controllable through the ways of the working of the social environment.

2. The Beginning of the Industrialization in East Asia

The history of the 'environmental problem' is not necessarily new, in the sense that even in the pre-industrial age there was deforestation, land exhaustion, and urban spoilage. However, the development

of modern industrialism brought a new stage of environmental problems from the pollution of land, water, and the air to car fumes, waste disposal, global warming, ozone depletion, and so on. The result is the recent issue of 'sustainability.'

Industrialization was originated by the Westerner in the second half of the eighteenth century. Non-Western countries, more than 100 years later, began to industrialize through cultural diffusion from the West. The beginning of industrialization in the East Asian countries was toward the end of the nineteenth century. In East Asian countries, as the beginning of the industrialization was late, the beginning of the environmental problem was of course late, too. But once industrialization started, it advanced very rapidly, so that the destruction of environment came so intensively.

Among East Asian countries, Japan's start to the industrialization was the earliest. Japan's industrial revolution was between the 1890s and the economic boom after World War I (1914–18). While the Japanese economy of the 1920s was in depression, there was an economic recovery after the Manchurian War (1931–3). However, the defeat in World War II (1941–5) made the Japanese economy degenerate to the level before 1920, so that the Japan's industrialization was interrupted between 1945 and 1955. It was the result of unprecedented high rate of economic growth period (1955–73), in which more than 10 percent in the yearly growth rate continued during 18 years, that Japan reached the economic level of the Western advanced countries in terms of per capita GNP. The other side of the coin, however, of this rapid economic growth was the peak in the unprecedented occurrence of pollution in this period.

The countries or the areas in East Asia that achieved the industrialization after Japan were Singapore, Hong Kong, Taiwan, and Korea (which are known as the 'Four Little Dragons'), as well as the advanced area of China (the coast area along the Yellow Sea and the East China Sea, and the basin area along the River Yangtze-Jiang). Among them the most advanced is Singapore which was a British colony before the World War II and is now the most successful international city state consisting of mainly overseas Chinese. The second most advanced is Hong Kong, which has developed as an international financial center until 1997 as British colony, but is now a part of China. The third is Taiwan, which has achieved success in economic development through an export market strategy since the latter half of the 1970s. Korea, similar to Taiwan, has achieved economic success since the government began to take the development strategy through export market in the 1960s and the 1970s.

The economic development of the Four Little Dragons is important, although they are small countries, because the industrialization of the East Asia was once a 'point' (Japan) but now it spreads to a 'surface' by them. In this surface the three points (Taiwan, Hong Kong, and Singapore) are Chinese in ethnicity, and it is important in that they effectively promoted the industrialization of the advanced area of mainland China. In the 1980s the surface including Japan, the Four Dragons, and China changed the image of the East Asia drastically from stagnation to development. However, from the point of view of the environmental problem, the other side of the development was the peak of the pollution.

3. *Economic Development of Japan and the Problem of Environment*

The first big industrial pollution in Japan was the contamination of rice fields by spill of the slag of copper mines into rivers. There were several such events, among which the biggest was the case of Ashio Copper Mine Co. Ltd. of Tochigi Prefecture beginning from 1885. Slag that spilled into Watarasegawa River in every flood polluted rice fields of the lower villages irrigated from the polluted river, so the victims were farmers whose rice fields became impossible to cultivate. Shozo Tanaka, a member of the Diet represented from the district, having decided to devote himself to the opposition movement of the victim farmers, not only interpellated government in the Diet, but also made a direct appeal to the Emperor Meiji by presenting a petition written by an anarchist Shusui Kotoku, but in vain.

Japan's Industrial Revolution was centered on light industries of silk and cotton spinning, and was followed by heavy industries of steel and machines after World War I. All of them used coal as energy, so that the sky of Osaka and other industrial cities were covered by soot and smoke. As these were regarded as a symbol of economic prosperity, however, there was no recognition that they were harmful for residents' living.

Between war defeat in 1945 and the beginning of high rate of economic growth in 1955 there were no environmental problems because Japanese industries had not recovered from the war damage.

At the starting point of the events of the industrial pollution after the war stood the case of Minamata disease of 1956 caused by methyl mercury included in the waste water of Chisso Co. Ltd. located in Minamata City, Kumamoto Prefecture. The Company threw the waste away into the Shiranuhi sea and polluted fish seriously, so the victims were fishermen who caught the polluted fish. The symptoms of the victims were terrible: paralyzed limbs, convulsed bodies, dribbling, sight and speech impediment, muddled senses, mental derangement; more than half died within 3 months. The cause was not known at the outset but was clarified gradually by the physicians of the medical department of Kumamoto University. Until 1968 the government did not acknowledge that

the liquid waste of Chisso Co. Ltd. was the cause, as there was a difference of opinion between the Ministry of Trade and Industry, which took the side of the Company and the Ministry of Health and Welfare that took the side of the Kumamoto University physicians. Because of this delay it was in 1973, 17 years after the event, that some of the victim's families won compensation by winning the case.

Following the Minamata case, pollution events occurred in many places of Japan during the high rate of economic growth period. In Niigata Prefecture, Niigata Minamata disease occurred in 1965 by the pollution of the Aganogawa River caused again by methyl mercury included in the waste water of Showa Denko Co. Ltd. The victims were fishermen again. In Toyama Prefecture, *itai-itai* disease occurred in 1955–60, caused by cadmium pollution from Mitsui Metal Mine Co. Ltd. into Jinzugawa River. The victims were farmers whose rice fields were irrigated from the polluted river. In Yokkaichi City, Mie Prefecture, once a harbor city with beautiful coast, a large-scale petrochemical complex was built in the latter half of the 1950s. Yokkaichi asthma occurred there, beginning in 1960, caused by sulfur oxide emission into the air from the complex. In Tagonoura Coast, Shizuoka Prefecture, in which the beautiful Mt. Fuji scenery of the Tokaido Line has been well known since the ancient Manyo poem, the coast was polluted by waste fluid discharged from the many paper manufacturing factories into the sea. Polluted sediment accumulated at the coast, was reported by photos in newspapers, and provoked national indignation.

Though these many cases of industrial pollution intensively occurred in the high rate of economic growth period, public opinion that was formerly permissive to the behavior of business firms came to severely watch them. Neighborhood protest movement against industrial pollution came to be organized in many districts. Under these backgrounds of rising public opinion and protest movement, the Basic Law for Environmental Pollution Control was enacted in 1967. In the so-called 'Pollution Diet' in 1970 this law was revised more rigorously and the 14 antipollution Laws were established at a stroke. In 1971 the Environment Agency started. Between 1969 and 1975 a series of environmental standards for sulfur oxide, carbon monoxide, nitrogen dioxide, water pollution, floating particles of dust, factory noise, airplane noise, and Shinkansen noise were settled.

Against the pollution for which industrial enterprises are responsible, these governmental anti-pollution measures under the background of public opinion and protest movement were very effective. After the latter half of the 1970s no enterprise discharged untreated wastewater or soot and smoke into rivers, sea, or air. This fact is also related to the change of the Japanese industrial structure into the 'postindustrial' stage, the border of which was again in 1975.

Thus, after 1975, the character of the environmental problem in Japan changed totally from the problem of pollution by factory to that of the 'global environment' such as the pollution caused by automobile gas and by family garbage for which all the residents in general who were once victims of the factory are responsible. The sort of targets with which the environmental problem must wrestle changed to, for example, saving energy and resources, separating and recycling garbage, regulating automobile gas, preventing global warming, preventing ozone layer depletion, stopping acid rain, and protecting nature.

Confronting these changes in environmental problem, the ways of the environmental administration as expressed in the 1967 Basic Law for Environmental Pollution Control came to be changed. In 1993 the enacted Fundamental Law of Environment altered the new environmental policy. The direction of this policy change was basically from regulating pollution to making a society in which sustainable development is possible.

4. Economic Development of China and the Problem of Environment

The People's Republic of China under the Mao Zedong regime (1949–76) pushed the Soviet Russian type of heavy industrialization until 1957, but after that the industrialization policy was discarded in adopting the People's Commune from 1958 and the Great Proletarian Cultural Revolution from 1966. After the death of Mao Zedong in 1976, Deng Xiaoping took political power and the Cultural Revolution came to an end. Deng Xiaoping, succeeding Zhou Enlai's policy of 'four modernizations,' promoted China's industrialization, on the one hand, by bringing up many small-scale industrial firms in rural villages after the abolition of People's Commune, and on the other hand, by introducing joint companies with Japanese and Western capitals by setting up in big cities of the 'Economic Special Areas' for that purpose.

As the economic development of Taiwan, Hong Kong, and Singapore progressed in this same period, Fujian as the root of Taiwanese, Guangdong as the root of Hong Kong people, and Taiwan, Hong Kong, and Singapore were unified into the 'Chinese People's Economic Area' beyond borders. The impact of this Chinese People's Economic Area, being extended to the whole advanced area of mainland China (the coastal area along the Yellow Sea and the East China Sea including 'Economic Special Areas,' and the basin area along the River Yangtze-Jiang), had the effect of promoting the industrialization of the advanced area of China as a whole. The average figures of GDP growth rates in real terms of 8 years from 1990 to 1997 are: China 9.9 percent, Singapore 8.0 percent, Korea

7.4 percent, Taiwan 5.5 percent, and Hong Kong 4.5 percent (the data are partially deficit for Hong Kong and Taiwan). In sum, China and the Four Little Dragons are now the area of highest economic growth in East Asia.

However, from the point of view of environmental problem, the other side of the coin of the high rate of economic growth in this area is the big deterioration of the environment of the East Asia. Let us look at this problem on mainland China.

The first point is air pollution. As the industrial as well as domestic energy in China depends overwhelmingly on coal, the sky in Chinese cities, especially in winter, is, even in clear weather, always darkened by sulfur dioxide, nitrogen oxide, soot, and smoke. When these melt into moisture, acid rain falls, and seasonal wind carries them to Korea and Japan.

The second point is deforestation. The River Yangtze-Jiang, the longest river in China, was once blue if not clear, but is now of yellow ocher color and there was a big flood in 1998, as the result of deforestation in the upper reaches. The Chinese government, though it announced that deforestation is prohibited, is unable to control offenders. This is the result of deforestation, based on the expanding difference between advanced and underdeveloped areas.

The third point is the problem of water shortage. In the Yellow River, the second longest river in China, water in the dry period disappears halfway, partly by the worsened condition in the upper reaches, but mainly by the increased use of water in the middle districts. Disappearance reaches 427 km and 107 days on the average in the 1990s. This brings big problems in agricultural and industrial production as well as in civil life in the middle and lower reaches.

The fourth point is the shortage of the facilities for the disposal of factory waste water and sewage. While the role of the many small industries in rural villages is important in the economic growth of contemporary China, they are mostly in primitive condition with no facilities of wastewater disposal. In cities, too, drainage and sewage disposal facilities are not enough.

The fifth point is the spread of desert. Mongols live in the dry area as nomadic tribes. Industrialization of China brought higher demand for wool, and the result an increase in breeding sheep. Because of the limited supply of grass in the dry area, the increasing number of sheep brought about the desert. Now it approaches the west side of Beijing.

In China, the environmental policy of the government started late after entering this period of high-rate economic development. The Vice Prime Minister Li Peng announced in 1983 for the first time that the environmental problem must be the basic policy goal. The Environment Protection Law was, although begun in 1979, enacted, after revision, in 1989. The administrative organization of the Bureau of Environment Protection started in 1974 as a section in the Ministry of the State.

In Korea, the environmental policy started earlier than in China. The Pollution Protection Law in Korea was enacted in 1965 and it was revised as the Environment Conservation Law in 1977. As the administrative organization, the Environment Agency started in 1980 and became the Ministry of Environment in 1994.

When the UN Conference on Environment and Development (Earth Summit) was held in Rio de Janeiro, Brazil in 1992, and the Rio Declaration on Environment and Development as well as 'Agenda 21' was adopted, China made 'Agenda 21 of China: White Paper on Population, Environment and Development in the 21st century China' and established the environmental standards in quantitative terms toward the sustainable development, such as the rate of diffusion of sewage disposal facilities in cities, the rate of reutilization of solid industrial waste, the rate of forest to total national land in area, and the rate of nature protection districts to total national land in area.

5. *The Significance of the Environmental Problem in East Asia*

While many of the East Asian countries are now in a stage of high rate of economic growth in industrialization, there are big differences among them in terms of the level of per capita GNP in US Dollars in 1998: Japan 32,350, Singapore 30,170, Hong Kong 23,660, Taiwan 12,265, Korea 8,600, China 750. The figures for Southeast Asian countries are: Malaysia 3,670, Thailand 2,160, Philippine 1,050, Indonesia 640. We can say that Japan and the Four Little Dragons achieved the industrialization, but others are still at the level of developing countries. Unlike Western European countries, in which industrialization is already complete and economic levels are not very different, in Asia the many countries have different industrialization levels.

As was shown in the process of COP3 Kyoto Convention (UN Framework Convention on Climate Change) in December 1997, there is opposition of interest between the advanced and the developing countries around the measures against environmental pollution, as the developing countries want to develop more rapidly. While this is also the case among the Asian countries, in East Asia, not only Japan and the Four Little Dragons, but also the advanced area of the mainland China have already achieved industrialization. In this sense East Asian countries, together with the Western countries, must take the responsibility for the global environmental problems with which now the entire world is confronted.

See also: Air Pollution; East Asian Studies: Economics; East Asian Studies: Politics; Environment and Development; Environmental and Resource Manage-

ment; Environmental Challenges in Organizations; Water Pollution: Marine

Bibliography

Ando J, Fujita Y 1998 *Ningen to Kankyo (Human Beings and Environment)*. Nisshin Shuppan, Tokyo

Chinese Publishing Company on Environmental Science 1990–2001 *Chugoku Kankyo Nenkan (Year Book on Environment)*. Chugoku Kankyo Kagaku Shuppansha, Peijing

Environment Agency 1972–2001 *Kankyo Hakusho (White Paper on Environment)*. Kankyo-cho, Tokyo

Funahashi H, Furukawa A (eds.) 1999 *Kankyo Shakaigaku Nyumon (Introduction to Environmental Sociology)*. Bunkashobo-Hakubunsha, Tokyo

Iijima N 1993 *Kankyo Shakaigaku (Environmental Sociology)*. Yuhikaku, Tokyo

Iijima N et al. (eds.) 2001 *Kankyo Shakaigaku no Shiten (Viewpoint of Environmental Sociology), Koza Kankyo Shakaigaku (Lecture Series: Environmental Sociology)*, Vol. 1. Yuhikaku, Tokyo

Kojima R (ed.) 2000 *Kankyo: Seicho eno Seiyaku to naruka? (Environment: Is it a Restriction to the Growth?). Gendai Chugoku no Kozohendo (Structural Change of Contemporary China)*, Vol. 6. Tokyo University Press, Tokyo

Tominaga K 1990 *Nihon no Kindaika to Shakai-Hendo (Modernization and Social Change in Japan)*. Kodan-sha, Tokyo

Tominaga K 1997 *Kankyo to Joho no Shakaigaku (Sociology of Environment and Information)*. Nikkagiren Shuppansha, Tokyo

K. Tominaga

East Asia, Religions of

Comprising China, the Korean peninsula, Taiwan, and Japan and containing a sizeable proportion of the world's population and home to some of the oldest and most advanced cultural systems in the world, East Asia has produced a number of major religious traditions such as Taoism, Confucianism, and Shinto, as well as developing its own distinctive regional forms of Buddhism. Besides these religions there exists also a pervasive folk religious tradition linked to customs, practices, and beliefs that forms the underlying bedrock to the entire religious structure of the region. Especially since the nineteenth century, too, numerous new religious movements have developed in the region, in response to rapid social change and modernity, but often reasserting many of the themes of folk religion and of the traditional religious environment within new forms adapted to the modern age. The region has also developed its own scholastic traditions which are essential for any understanding of its religious dynamics and that provide important new perspectives and alternative voices from those derived from Western academia.

1. *East Asia in Religious Contexts: an Overview*

East Asia represents a problematic area of analysis because of its size, vast and varied population, historical and social complexity, the differing historical and social experiences of each country, and its multiplicity of religious forms. This, and the existence of religious expressions associated with particular countries of the region, raises questions about how far one can talk about the region as a whole in religious or cultural terms. An example here would be Shinto, which is specific to the Japanese cultural sphere and which raises questions about the use of religious ideology in the formation of nation states, about cultural identity, and about the relationship between religion and politics. However, while Shinto provides meaningful examples relating to such questions, it cannot necessarily be considered as an exemplar of the patterns of East Asian religion in general, given that no similarly identifiable national religious traditions exist in other countries in the region. Similarly, the religious responses to rapid social change and modernity, which have seen the formation of new religious movements in different parts of the region, cannot necessarily be considered as indicative of a region-wide commonality. Many of the new movements that have arisen in Korea, e.g., draw heavily on evangelical Christian themes as well as the shamanic beliefs and practices which are central to the Korean tradition, while Japanese new religions have drawn much more heavily on Japanese cultural traditions based on Buddhism, Shinto, and folk religion. Indeed, scholars from the region have been reluctant to draw many parallels between the countries of East Asia. Japanese scholars, e.g., prefer to discuss the Japanese new religions in the comparative context of new religions in the West rather than in the Asian context. Rather than expressing similarities and commonalities, indeed, the academic traditions of each country in the region tend to emphasize concepts of cultural uniqueness and difference and where they have sought a comparative focus have looked outside the region.

2. *Religious Traditions of East Asia: Buddhism, Confucianism, Taoism*

The East Asian religious spectrum includes historical religions that have influenced the whole region and localized traditions that are influential in some, or one, area but that cannot claim cross-border significance. Three major historical traditions have been influential across boundaries and formed part of the religious constitution of the region as a whole: Buddhism, which although coming from outside, has taken on particular regional characteristics in China, Korea, Taiwan, and Japan, and Taoism and Confucianism, both of which originated in China but have also been influential in Korea, Taiwan, and Japan.

2.1 Buddhism

Buddhism entered East Asia along the major trade routes, supported by the patronage of ruling elites who saw it as a means of strengthening the ethos and power of the state. This was particularly so in Japan, where it was adopted as an arm and pillar of the state from the seventh century onwards. While its growth and development of an extensive monastic system owed much to its promises, expressed in popular Buddhist texts such as the Lotus Sutra, of spiritual emancipation for all, and while it has developed monastic forms and meditative traditions (most famously Zen in Japan) with a primary orientation towards renunciation and spiritual transcendence, Buddhism's main orientation in East Asia has been social, familial, and mundane. As elsewhere, it has used magical symbols and rituals for the amelioration of the living, both individually and in social contexts, and operated as a religion dealing with death and the after-life, offering an eschatological framework of other-worldly salvation through a range of memorial and funerary services which aid the dead spirits in their journey to the other world. It has been through Buddhism that the spirits of the dead have been transformed into ancestors who guard over their living kin and in this way it offers solace to the bereaved while reinforcing the solidarity and continuity of the household.

2.2 Confucianism

While the thoughts of Confucius (551–479 BC) may be regarded as the foundations of a moral and philosophical tradition emphasizing filial piety and the moral duties of rulers and subjects, Confucianism has developed into a broader cultural and religious form that has influenced the development of all East Asian societies. While Confucianism has taken on overt religious manifestations in the ritual veneration of Confucius, its primary influences have been ethical and relating to the affirmation of social codes that underpin the existing order of society. Confucianism, besides emphasizing classical learning and the veneration of tradition, affirms the basic goodness of human nature and asserts that the ways individuals conduct themselves affects the basic social order at familial and wider social levels. This notion of individual spiritual responsibility for one's actions, and of the individual as a microcosm, in spiritual terms, of society at large has been central also to many of the East Asian new religions.

2.3 Taoism

Although originally a mystical tradition of meditation and passive withdrawal from worldly affairs, Taoism became associated with the search for longevity and the overcoming, through ritual and mystical practices, of the physical limitations of the body. Taoist practitioners in China sought, through the study of alchemy, to find the elixir of life, and they formed cults venerating that deities symbolized longevity. Taoism also developed complex systems of divination and fortune telling which were central to its appeal beyond China. Taoist ideas spread across the region in tandem with other Chinese cultural influences, including Confucianism, Buddhism, and the Chinese writing system. In Japan, Taoist influences were subsumed within the broader cultural sweep of continental culture and it never assumed its own distinct religious form there. However, its influences were pervasive, especially in terms of divination and magical practices. The early Japanese Imperial Court, e.g., developed a Bureau of Divination using Taoist techniques to predict and govern state affairs, while popular divination practices derived from Taoism remain widespread in Japan.

3. Indigenous Traditions: The Case of Shinto

Although Shinto, literally 'the way of the gods,' is often termed an indigenous Japanese religion, it is more accurately considered as a localized tradition that developed in conjunction with and in antithesis to the foreign religion of Buddhism which entered Japan from the sixth century onwards. The native religious tradition of Japan prior to the entry of Buddhism centered around the worship of local gods such as clan tutelary deities who were guardians of communities and clan lineages. This loose-knit local tradition acquired a specific identity with the entry of Buddhism. Even the term Shinto first came into existence after the incursion of Buddhism, as a means of differentiating the local from the imported tradition. Shinto is, thus, an example of a tradition that developed in response to (and hence in conjunction with) another, external religion, a point that has been emphasized consistently by influential Japanese scholars such as Kuroda Toshio.

Shinto is based in concepts relating to the *kami*—a term usually translated as god or deity, but implying a spiritual life-giving force. *Kami* may be associated with specific natural phenomena or landscape (for in Shinto, the entire world of Japan is considered to be the creation of the *kami* who give life to and animate it) but they may be spiritual entities that embody particular virtues or capacities and that can, if suitably beseeched and venerated, bestow blessings. The Shinto pantheon contains many thousands of *kami* from localized deities to ones of national influence. *Kami* may also be the spirits of humans of note or special accomplishments, such as Sugawara no Michizane, the ninth century scholar whose spirit is venerated as Tenjin, the *kami* of learning.

3.1 Shinto, Buddhism, and Religious Integration in East Asia

Interaction rather than conflict between Buddhism and Shinto has been a core theme in Japanese religious history. Buddhism's strategies of assimilation and accommodation have been nowhere more striking than in its dealings with Shinto and its *kami*. Since the *kami* were the guardians of the earth and the landscape, Buddhism adopted them as spiritual protectors of Buddhist temples: the striking vermilion Shinto gateway, representing entry to the *kami*'s sacred area, is rarely absent from Japanese Buddhist temple courtyards and, especially up to the nineteenth century, many religious centers were not so much Buddhist temples or Shinto shrines, but shrine–temple complexes combining both traditions. Many deities are also shared between the traditions, while this complementarity is manifest also in the roles Shinto and Buddhism play in the life cycles of individuals and communities. While Buddhism deals with death rituals and the memorialization of ancestors, Shinto is associated with beginnings and growth. The Shinto gods guard over the beginnings of life and are called upon for protection and support when babies are born and as infants grow: the *miyamairi* rite in which newly born children are taken to shrines to be placed under the protection of the Shinto gods, remains widely followed in Japan, as is the *shichigosan* festival in November when children of seven, five, and three are given further ritual blessings.

This functional integration, in which Shinto deals with birth and life and Buddhism with death and the hereafter, has strengthened both traditions in the Japanese system. It is not a pattern limited only to Japan, for a dominant characteristic of the religious structures of the entire region has been that of integration of a number of traditions in the lives of the people. Thus, a state official in premodern China might be involved in Confucian rituals at work, Taoist divination rites and practices in his family life, and Buddhist funerary rituals when someone died, while in Korea, too, similar patterns (Confucian public rituals, the use of shamans and diviners for personal concerns, and Buddhist death rituals) could be discerned. It has only been in recent times, with the emergence of exclusive religious groups, such as the evangelical Korean protestant sects and a number of Japanese new religions that refuse to condone traditional religious practices, that this integrative pattern has been challenged. For the majority, however, it remains dominant.

3.2 Festivals and the Ritual Calendar

Shinto illustrates the close interaction of religion, festivity, and the calendrical cycle—a theme that permeates much of East Asian religion. Its festivals (*matsuri*, meaning both prayer and celebration) were traditionally a central element in the lives of local communities, and in modern Japan they remain a colorful and vibrant element in the calendar, involving a mixture of sacred ritual and spectacular festive entertainment that may include processions, dances, fairs, tourist attractions, and much celebratory eating and drinking. The most widespread of all such festivals is New Year when people make a first visit of the year to a shrine (*hatsumôde*) to thank the gods for blessings received in the past year and for protection and good fortune in the coming year. Many people also visit Buddhist temples at this time. *Hatsumôde* is highly popular and, since the 1970s, has been performed by a growing number of Japanese: numbers have risen from around 30 million in the late 1950s to around 85 million (out of a population of 120 million) in the 1990s. The growth of *hatsumôde* numbers is not isolated: Japanese scholars have noted a growth in participation in customary religious ritual activities between the 1950s and the early 1990s, and this has led them to question the validity of secularization theories and to show that these do not necessarily have universal validity.

However, such growth masks a shifting pattern within Shinto, in which small, rural shrines have declined due to rural depopulation and because people have preferred to visit better known and more identifiably national shrines further afield. Partially this has been a result of the development of an advanced transportation system but also because, as localized community structures have broken down or been transformed due to changing life patterns, people have sought their primary focus of identification in the national community. Hence, while they once would have performed *miyamairi* and *hatsumôde* at local shrines, tying themselves to local communities and deities, they now do these at shrines of national significance, an act that confirms their identification with the wider national community and reaffirms, in an age of internationalization in which the primary forms of comparison in terms of identity are across national rather than regional or local lines, their sense of Japanese identity. Many nationally well-known shrines receive many millions of worshippers at festival times: Meiji shrine in Tokyo, for instance, receives around four million visitors at New Year.

3.3 Shinto, Mythology, National Identity, Nationalism, and the State

The connection between Shinto and national identity indicated by recent patterns of shrine visiting points to a deep-seated characteristic of religion in the region: the close connection of religion, state, and identity. Here, again, Shinto is a striking (if particularized) example. Its myths attribute the creation of Japan and the origins of the Japanese people and its Imperial dynasty to the *kami*. In these myths the Imperial family are ordained divinely to oversee the nation, as

they are descended from Amaterasu, the Sun Goddess, who is venerated at the great shrines of Ise, the most important shrines in Japan. The *kami* are the procreators of Japan and its people, who thus exist in a special, unique relationship to their gods and land. This special relationship is central to Japanese concepts of identity and has been associated closely with Japanese nationalism since the Meiji Restoration of 1868, when Japan developed a modern-style sense of nationhood by using Shinto and its mythological links to the Emperor as a focus of national unity and as a means of emphasizing national uniqueness as a chosen people and land. The association of Shinto and the state helped create a myth of racial superiority which legitimated Japan's colonization of Taiwan and Korea, involved the suppression of other religions and played a major part in Japan's fascism and war mongering in the first half of the twentieth century.

In the postwar era, official links between Shinto and the state have been barred under the Constitution, Japan has become a secular state, and Shinto exists predominantly as a shrine-centered tradition based around life cycle and other rituals and festivals. Nevertheless, there remain many people (especially of right-wing persuasion) and many Shinto priests who aspire to a return to a formal association between Shinto and the state. Many of the discussions of Shinto in the modern age thus center around issues of religion, nationalism, and politics.

4. *Underlying Structures and Common Themes—Folk Religion*

Although the above-mentioned religions are the most visibly identifiable older traditions of the region, they are underpinned by a pervasive folk tradition. It is a common argument of scholars from the region (e.g., Miyake Hitoshi in Japan) that this amorphous folk religion of shared customs, attitudes, beliefs, and practices deeply rooted in everyday lives and concerns of the people, forms the foundational religious structure of the region. Indeed, many of the ideas and themes that are central to the new religions derive directly from these folk religious foundations, while the new religions themselves have been analyzed widely as modern articulations of the folk tradition. The folk tradition has also been the filter through which the historical religions of the region have been adopted by and adapted to local cultures. Those religions that have best developed strategies to accommodate or work with this tradition have been the ones which have transplanted most successfully to or developed in the region: examples include Buddhism, whose strategies and doctrines of accommodation have enabled it to absorb and work with local customs, and (in a reverse case) Christianity, whose inability to accommodate or accept folk religious beliefs and customs, notably relating to the ancestors, has proved a major impediment to its growth in much of the region.

The importance and vitality of this folk tradition is such that it has been a major focus of academic study in the countries of the region. Folk studies (or folk religious studies) are a primary academic discipline in Korea (where shamanism, as the cornerstone of the folk and native religious structure, has received especial attention) and Japan, and is generally accorded greater weight as an academic discipline than in the West. It is not just concerned with the indigenous traditions of folk religion; Japanese scholars such as Sakurai Tokutarô (shamanism), Miyake Hitoshi (mountain religion), and Shinno Toshikazu (pilgrimage and legends) have been concerned to map out and discuss the interactions of folk religion and Buddhism, and to demonstrate that 'folk religion' as manifest in Japan should not be discussed solely in Japanese terms but placed in a wider context. Making use also of Western academic studies of religion, scholars such as Miyake have demonstrated the potential universality of folk religious models, as well as broadening the comparative and potentially universalizing dimensions of Japanese scholarship of religion.

4.1 *Themes of East Asian Folk Religion*

The varying elements associated with the East Asian folk religious tradition(s) include emphases on calendar and ritual cycles relating to individuals and communities, the prevalence of shamanic beliefs and concepts relating to the role of spirits (including the spirits of the dead), spirit possession, healing, and the analysis of misfortune and provision of healing, the veneration of ancestors, and various this-worldly beliefs and customs relating to seeking assistance and amelioration in this world.

All these exist alongside or interact with the major traditions or are manifest in and expressed by them. Concepts relating to the spirits of the dead and the ancestors are deeply rooted in the folk religious arena, but the ritual services related to death and the ancestors are most commonly performed through the medium of Buddhism, whose success in East Asia has depended greatly on its ability to adapt to and accommodate folk beliefs in this area. The importance of the ritual calendar and the notion of the yearly cycle of festivals and rituals, originally based in the folk tradition and centered around agricultural rites, has already been mentioned in the context of Shinto in Japan, but it extends throughout the region in differing forms.

Perhaps the most predominant beliefs in the folk tradition relate to psychic interpretations of causation (of illness, misfortune, and, indeed, good fortune). Beliefs that this world does not exist in isolation but is influenced by other realms, and that those living in this

world can be affected by the actions of spirits, mean that, in folk terms, any problem or event may be seen as having spiritual as well as physical causes. Thus, an illness may have psychic causes (e.g., the actions of an unhappy spirit) as well as physical (e.g., the effects of germs) ones, and hence solutions to problems require spiritual attention (e.g., ritual performances, acts of penance) to deal with its spiritual aspects. A strong shamanic tradition has existed throughout East Asia, in which spiritual healers (very often female, and with claims to be able to either travel between this and the spiritual world, or to make contact with it and find ways of appeasing the spirits who are afflicting the living) play a major role. This shamanic tradition is perhaps the most ubiquitous element in the Korean religious universe, but it also remains important in the rest of the region. Studies of Japanese religious behavior in modern times show that people are far more likely to turn to a local healer or diviner than to a Buddhist priest when they have personal problems. This loose-knit tradition of healers and diviners serves also as a counterbalance to the formal religious systems of Buddhism and Shinto, whose priesthoods are almost entirely male, and provides a mode of religious expression for spiritually inclined females, who have remained marginalized in the formal traditions. By contrast, the vast majority of shamanic spiritual healers in Korea and Japan are female.

5. *Millennial Themes in East Asian Religious Culture*

An important theme in East Asian religion that shows how religious ideas may be expressed across the various traditions and cultures of the region relates to millennial beliefs and expectations of future worldly salvation. These are found throughout the region, articulated through the folk religious traditions, Taoism, Buddhism, and, more recently, the new religions. Expectations of a new dawn, the arrival of a spiritual kingdom on earth, and/or the appearance of a celestial or spiritual savior who will bring justice, redeem the sick, punish the unworthy, and transform or renew the world have been especially potent in China, sometimes with Taoist influences but often associated with Buddhism, but they have emerged in Korea and Japan as well. In China, various Taoist-inspired movements such as the second century CE Way of Great Peace movement have rebelled against the state in the hope of establishing a new paradise on earth, while Buddhist-inspired movements have also been inspired by the concept of imminent salvation on earth, often centered on the future Buddha Maitreya (whose cult also inspired similar millenarian impulses in Japan). In China, too, there have been repeated uprisings of the millennial White Lotus Society, which combined Pure Land Buddhist ideas along with folk, Taoist, and other elements, while the Pai Shan-ti Hui movement, which incorporated Christian ideas and sought to create a heavenly kingdom of eternal peace on earth, caused the bloodiest and most dramatic of all such millennialist uprisings, the Tai Ping rebellion of 1849–64 which claimed millions of lives.

Millennial themes are prevalent in many of the new religions. In Korea, numerous religious movements have fused extreme forms of evangelical Protestantism with dramatic millennial visions. The Tami Church (known also as the Mission for the Coming Days) attracted attention when it announced that the apocalypse would begin on October 28, 1992 and that its members would be raptured up to heaven. While this case of failed prophecy damaged the Tami Church, it has not dampened the emphasis on apocalypticism, which remains a central theme in much of Korean Protestantism, and which forms one of its cardinal attractions for the local populace.

The most striking manifestation of millennialism in Japan has been in the new religions that have arisen since the mid-nineteenth century, initially through the actions of charismatic female religious leaders such as Nakayama Miki (1798–1887) of Tenrikyô and Deguchi Nao (1837–1918) of Ômotokyô, but most recently among the 'new' new religions of the late twentieth century, many of which have emphasized that the world is at a critical turning point in which the old, materialist civilization will be swept away as a new spiritual world dawns. While many of the adherents of earlier millenarian movements were poor and disadvantaged, and the thrust of their movements was initially towards seeking social justice, the beliefs of the latest wave of new religions have been conditioned more by concerns that this world is facing extinction because of nuclear or environmental problems, and that a spiritual revolution is essential in order to avoid destruction. The most dramatic and famous of these later millennial movements was Aum Shinrikyô, which believed that a final war between good and evil was inevitable and in which the vast majority of humanity would be wiped out prior to the advent of a more spiritually oriented world.

6. *New Departures: The New Religions*

Since the mid-nineteenth century massive economic development, the impact of Western influences, and immense social change and population movement have produced recurrent tensions between traditional social, cultural, and religious structures, and the individualizing patterns of modernity. In religious terms, the most striking sign of such tensions and developments has been the emergence of new religious movements throughout the region. This phenomenon has been most pronounced in Japan but is discernible also in the rest of the region.

The new religions have generally been eclectic, drawing on the primary religious inspirations of the

societies in which they have emerged: Korean new religions, e.g., might draw on Buddhism, shamanism, and Confucianism, while Shinto (along with Buddhism and the folk tradition) is a source of inspiration for many Japanese new religions. In very recent times, influences have come from outside the region as well, with evangelical Christian movements especially prominent in Korea and new-age themes appearing in the most recent new religions in Japan.

The new religions have been especially effective in dealing with the tensions between modernity and tradition, through their articulation of many aspects of traditional folk religious belief (e.g., concepts of psychic causation) in ways that fit with the needs of individuals in the modern world. They have flourished mostly in the growing conurbations of East Asia, providing a new sense of belonging and community identity for populations that have moved from rural areas to the cities for work reasons. While providing individual solutions to problems and focusing on understandings and definitions of the self in modern terms, they have equally emphasized tradition, reaffirming many of the social parameters of traditional culture and morality. The solutions they pose to individual problems center around shamanic concepts of spirit possession and magically-charged healing rituals, but combined also with reaffirmations of traditional moral structures and obligations, notably Confucian ethical values of filial piety. A typical example taken from a new religion in Japan might be that a personal misfortune or illness is the result of having neglected to perform appropriate rituals for one's ancestors and of causing their wrath: this breach of ritual propriety is a moral lapse which needs to be rectified through repentance and self-reflection, as well as participating in healing rituals. As such many of the new religions display a pronounced interweaving of magical practices alongside affirmations of morality; as various Japanese scholars have noted (e.g., Shimazono 1992), rather than magical practices decreasing as ethical and moral considerations become more strongly emphasized, the Japanese new religions show that the two often operate in tandem, the latter increasing and being articulated more pronouncedly as magical practices become more widespread.

The region's new religions almost invariably owe their origins to the charisma and inspirational leadership of individuals who claim to have direct contacts with other realms or to bring new revealed truths to the world. While serving as a medium for a newly revealed deity, however, a common pattern is that the charismatic leader gradually displaces the deity as a source of inspiration, authority, and worship, often becoming regarded as living deities and venerated accordingly. Their charisma, however, may continue after death, a point often emphasized by Japanese scholars interested in the study of charisma: authority and succession to the leadership generally (following traditional East Asian familial patterns) passes through the family lineage to a son or daughter, suggesting that charisma is in part related to the concept of blood and lineage, and the deceased founder becomes an even more central figure of worship, with their mausoleum often becoming a pilgrimage center for the religion.

6.1 Aum Shinrikyô: New Religions and Violence

In 1995 world attention was directed towards the millennial Japanese new religion Aum Shinrikyô and its blind charismatic leader, Asahara Shôkô, because of its nerve gas attack on the Tokyo subway. While Aum was not the first or only religious movement in East Asia in modern times to commit acts of violence the sensational nature of the subway attack, Aum's use of chemical weapons, and the subsequent evidence showing that it had engaged in a protracted series of crimes emanating from its religious activities and beliefs, caught the attention of defense and law agencies as well as scholars world wide.

The Aum affair has special significance for the sociological study of religion for many reasons. It showed similarities with patterns of violence that manifest in other religious movements in the West (e.g., the Peoples Temple, the Rajneesh movement, and the Order of the Solar Temple) and appeared, in an extreme form, to fit into a wider pattern of the production of violence within certain forms of newly formed charismatic, world rejecting millennial movements. It provides a dramatic case study for those interested in the relationship between religion and violence, and several of those who have studied Aum have interpreted it as an example of the violence-producing dimensions of religion (e.g., Shimazono 1997, Reader 2000). As an example of how new religions can produce violence primarily because of their doctrinal orientations and internal dynamics, the case of Aum provides a significant counter-example to the arguments widely held by many scholars of new religions in the West, that movements that engage in violence do so primarily in response to excessive external pressures placed on them by a hostile society.

7. Scholarship on the New Religions and the Problem of East Asia

Mention has been made of the scholastic traditions in the region, especially of Japan, and some of the analyses and ideas they have transmitted. The new religions of Japan provide a case in point where work done by Japanese scholars has added to the wider understanding of new religions across the globe. Japanese scholarship generally has examined the 'new' new religions in the context of late twentieth century

urban society and modernity, and as such has thus recognized, either explicitly or implicitly, the parallels that could be drawn between Japan as a first-world, modern, and technological society, and Western countries. Western scholars specializing in new religions have also been drawn to comparative studies of the Japanese new religions because of these recognizably common influences and factors. Much of the best Japanese scholarship on the new religions of Japan has taken account of Western theoretical perspectives, while many of the scholars in Japan have worked closely with Western scholars of new religions, mutually influencing and being influenced by them.

This has placed contemporary Japanese scholarship on the new religions in an international framework, and has made Western scholars more than ever aware of the importance, when seeking to develop comprehensive or universal models and categories in the study of religion, of the Japanese situation. Certainly no sociologist of religion concerned with the rise of new religions in the modern world can ignore the work being done in this field in and relating to Japan, just as no student of religious violence or of millennial movements could afford not to take note of Aum Shinrikyô. However, the corollary of this development has been to separate out Japan from the rest of East Asia in terms of the study of new religions: rather than viewing parallels within the region, e.g., Japanese scholars (and Western scholars working on Japan) have tended to examine Japanese new religions in the comparative context of Western new religions. This, however, means that the (potential) parallels between Japanese new religions and those of the rest of East Asia are focused on less clearly.

In other words, the problematic question raised earlier about how far one can discern or discuss the region in broad terms, is in some areas being answered in practical terms by contemporary scholastic work in and centered on Japan, which associates Japan far more with the industrialized West than, and thereby draws implicit boundaries with, the rest of geographical East Asia. While this does not invalidate the conceptual category or notion of East Asian religions, it does indicate the problematic aspects of talking of the region in any uniform sense. Not only do the traditions themselves take on different guises in the different parts of the region, but the manner in which they may be studied and the lessons that may be drawn from them may also highlight questions of difference rather than commonality. To what extent future studies will further add emphasis to the issue of difference and challenge the viability of treating the region as an entity is a moot question of great interest and significance to the field as a whole.

See also: Buddhism; Charisma and Charismatic; Folk Religion; Folklore; New Religious Movements; Religion: Nationalism and Identity; Religion, Sociology of; Traditions in Sociology; Traditions: Social; Youth Movements

Bibliography

Buswell R E 1992 *The Zen Monastic Experience: Buddhist Practice in Contemporary Korea*. Princeton University Press, Princeton, NJ

Ching J 1993 *Chinese Religions*. Macmillan, Basingstoke, UK

Faure B 1991 *The Rhetoric of Immediacy: A Cultural Critique of Chan/Zen Buddhism*. Princeton University Press, Princeton, NJ

Grapard A G 1992 *The Protocol of the Gods: a Study of the Kasuga Cult in Japanese History*. University of California Press, Berkeley, CA

Hardacre H 1986 *Kurozumikyo and the New Religions of Japan*. Princeton University Press, Princeton, NJ

Hardacre H 1989 *Shinto and the State*. Princeton University Press, Princeton, NJ

Hori I 1968 *Folk Religion in Japan*. University of Chicago Press, Chicago

Jones C B 1999 *Buddhism in Taiwan: Religion and the State, 1660–1990*. University of Hawaii, Honolulu, HI

Kendall L 1985 *Shamans, housewives and other restless spirits: women in Korean ritual life*. University of Hawaii Press, Honolulu, HI

Kuroda T 1975 *Nihon no chûsei no kokka to shûkyô* [*The Medieval Japanese State and Religion*]. Iwanami Shoten, Tokyo

Kuroda T 1981 Shinto in the history of Japanese religion. *Journal of Japanese Studies* **7**(1): 1–21

Kyburz J A 1987 *Cultes ou Croyances au Japon: Kaida, une Commune dans les Montagnes du Japon Central*. Maisonneuve à Larose, Paris

Maspero H 1981 *Taoism and Chinese Religion* (trans. Kierman F A). University of Massachusetts Press, Amherst, MA

Miyake H 1974 *Nihon shûkyô no kôzô* [*The Structure of Japanese Religion*]. Keiô Tsûshinsha, Tokyo

Miyake H 1989 *Shûkyô Minzokugaku* [*Folk Religious Studies*]. Tôkyô Daigaku Shuppankai, Tokyo

Ooms E G 1993 *Women and Millenarian Protest in Meiji Japan: Deguchi Nao and Ômotokyô*. Cornell University, Ithaca

Overmeyer D L 1976 *Folk Buddhist Religion: Dissenting Sects in Late Traditional China*. Harvard University Press, Cambridge, MA

Overmeyer D L 1986 *Religions of China: The World as a Living System*. Harper and Row, San Francisco

Reader I 2000 *Religious Violence in Contemporary Japan: The Case of Aum Shinrikyô*. Curzon and University of Hawaii, London and Honolulu, HI

Sakurai T 1978 *Nihon shûkyô no fukkô kôzô*. Kôbundô, Tokyo

Shimazono S 1992 *Gendai kyûsai shûkyôron* [*The Study of Thisworldly Salvationist Religions*]. Seiyusha, Tokyo

Shimazono S 1997 *Gendai shûkyô no kanôsei: Oumu to bôryoku* [*The Potential of Religion in the Modern World: Aum and Violence*]. Iwanami Shoten, Tokyo

Shinno T 1991 *Nihon yugyô shûkyôron* [*Studies of Japanese Ascetic Religious Travel*]. Yoshikawa Kôbunkan, Tokyo

Smyers K A 1998 *The Fox and the Jewel: Shared and Private Meanings in Contemporary Japanese Inari Worship*. University of Hawaii, Honolulu, HI

Teiser S F 1988 *The Ghost Festival in Medieval China*. Princeton University Press, Princeton, NJ

I. Reader

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

East Asian Studies: Culture

'East Asian Studies' is a well-established area studies program with a long and distinguished tradition of studying the peoples of East Asia and their lives. 'East Asia' appears to be a fairly well demarcated geographical area, consisting at present of the People's Republic of China (mainland China, now including Hong Kong), the Republic of China (Taiwan), North and South Korea, and Japan. However, shifting political boundaries—even at the start of the twenty-first century—and recent epistemological and theoretical questions in various fields have reminded us that 'East Asian Studies' is a highly complex undertaking. Building on excellent studies of the peoples of East Asia and their cultures, the field is now invigorated by fresh insights and perspectives. Historically the Chinese, Korean, and Japanese 'civilizations' have had much in common, due mainly to the diffusion of the Han and Tang civilizations from China to the rest of the area. Linguistically, the Sino-Tibetan languages, all of which are characterized by the tonal system (whereby tones give distinctive meanings to the same sound types), belong solely to the Chinese; the Korean and the Japanese, together with the Ainu and Gilyaks, come from an altogether different, non-tonal linguistic stock. Yet the Chinese writing system was adopted and imposed on the spoken languages of the Korean and the Japanese, although both later added their own system.

1. *Recent Developments—'Non-socialized Culture'*

Given Japan's indebtedness to Chinese civilization, Japan has had a long tradition of Sinology, attracting scholars of China—including Chinese—to study there. Asian peoples themselves, along with scholars, used to view their country primarily in relation to 'the West'. However, at least since the mid- to late nineteenth century, there have been vigorous dialogues on and by Asians and Asianists on the relationships—cultural, political, economic, etc.—among Asian peoples. Representations of Japan, for example, to the rest of the world have, until recently, been made primarily by Western scholars alone, but Chinese, Koreans, and other Asians—and not exclusively the elites as in the past—are now taking part in the study and representation of the Japanese. Conversely, Japanese scholars have greatly shifted their focus to other Asian peoples, especially Koreans (Shima and Janelli 1998). A new Pan-Asianism, distinct from the one imposed by the Japanese during their imperial period, is a product of renewed recognition of cultural, social, economic, and political relationships of long standing among Asian countries and peoples.

Recent changes of emphasis and perspective in East Asian Studies derive from criticisms of, and even some frontal attacks on, the concept of *culture*—a cherished or, some would argue, the most fundamental concept in anthropology. Today, the term culture is in increasingly common use outside of anthropology, and in fact flourishes in Cultural Studies and various other disciplines, while becoming widely used in the mass media. At the same time, anthropology remains divided on this issue. Those who advocate to do away with 'culture' altogether have been accused of not offering alternatives to deal with collectivity, that is, 'society' or 'culture.' Even those sympathetic to the criticisms of the culture concept feel it is hard to discard it altogether, or, '[W]hatever the infirmities of the concept of 'culture' ('cultures,' 'cultural forms' …) there is nothing for it but to persist in spite of them' (Geertz 1995).

We are now increasingly aware that the *concept* of culture, if used as a hermetically-sealed entity, fails to explain the very *dynamics* of culture, always in the process of transforming at the confluence of internal/local historical development and external/global historical forces. Almost all major foundational features of Japanese culture and political economy came from the Asian continent, usually via Korea. They include the 'wet-rice economy' in early times, and later—during the fifth and sixth centuries—the imperial ritual, the writing system, city planning, metallurgy, and a host of other activities. The draft of the Meiji constitution, promulgated in 1889, was drawn up by German and Austrian legal scholars and many of the key articles were adopted almost verbatim. Japan's modern army and navy were modeled after the French and German army and navy. In fact the modernization in Japan, as in other Asian countries, was, to a great extent, Westernization. No society is an island in history. Hybridity is a *sine qua nom* of any culture. Recent emphasis on globalization makes us keenly aware of the basic nature of culture as a dynamic, unending process of the interpenetration between the local and global.

2. *Dynamics Within*

A powerful critique of the notion of culture as a 'totalizing' and 'essentializing' representation of a people or a culture led our studies in several productive directions. One such direction is the further emphasis on *diversity*, multiple voices within an artificially demarcated unit of our study—be it a people, a culture, or a society—resulting in a plethora of research on social groups other than the dominant ones. These groups include not only the so-called minorities but also the homeless, school dropouts, etc. This move is a corrective to representation in the past, for example, of the Chinese by the Han Chinese, or the Japanese as the agrarian Japanese: the emphasis on the latter led to a label, 'rice paddy courses,' refering to introductory courses on Japan. The recent research on minorities places these social groups not as isolated groups but in

relation to dominant social groups, especially but not exclusively in regard to power inequality between the dominant and marginalized social groups.

In the case of Japan, wet-rice agriculture, which originated somewhere as yet unidentified within tropical Asia, reached northern Kyūshūsome time around 350 BC, supplanting the hunter-gatherer economy of the first occupants of the archipelago, who had arrived around 200,000 BC. It laid the foundation for the political economy of the Yamato state and the imperial system. Gradually the segment of the population engaged in rice agriculture—or, more accurately, who reaped the economic/financial crop from wet-rice economy—became the dominant social group, representing 'the Japanese'. The agrarian cosmogony and cosmology became those of the Japanese, whereby the rice agrarian seasonality governed all Japanese and rice paddies came to represent the Japanese space.

There is an increasing awareness now among scholars that the other side of the development of 'Agrarian Japan' was the marginalization of 'the Japanese' in nonagrarian sectors. They included those people whose livelihood derived from the resources of the sea and the mountains. In particular, this included what later became known as the '*hisabetsu burakumin*' (settlement people under discrimination). These were people who engaged in nonagrarian occupations such as the performing arts, architecture, and trading, plus itinerant religious specialists, who, from late medieval times—when the notion of impurity received radical negativity—became increasingly marginalized, until 1582 when they were codified as two categories of 'outcasts.' The marginalization of the Ainu, who continued to engage in hunting, fishing, and plant gathering well into the nineteenth century, was similarly based on the yardstick of agrarian values, which placed emphasis on plant domestication. In addition, the assignment of 'primitiveness' to them was a byproduct of the Japanese, who tried to 'modernize' their nation according to the model of Western modernization (Ohnuki-Tierney 1987, 1993).

A related development is the shifting of focus among scholars from males to females—their voices, their life experiences, and women as social agents who actively carve out or change the courses of their lives. The issue of gender preferences, both institutionalized and culturally sanctioned as well as individual, has also received considerable attention. Gender-related issues often address the question of gender inequalities, which are in turn juxtaposed with the power inequalities between classes within society, on the one hand, and between the colonizers and colonized, on the other, breaking new ground for various disciplines (Watson and Ebrey 1991).

3. *Transnationals, Diasporas*

Another major critique of culture is how it was presented in the past as being spatially tied. The non-isomorphic nature of the cultures, locality, and peoples of East Asia was realized early by scholars of East Asia, who have contributed to the studies of 'overseas communities'—numerous 'Chinatown' studies, Brazilian Japanese, and the like. 'Emigrant communities' are a native category of villages from which people regularly emigrated to overseas communities, while closely maintaining ties with their home villages (Chen 1939, Watson 1975). Building on this distinguished tradition, scholars of 'transnationalism' study emigration, the resultant overseas communities (some diasporic), and especially transnationals with 'flexible citizenship' in the contemporary context of globalization which created a unprecedented flow of population across borders of nation-states (Ong 1999). The recognition of the non-isomorphic relationship between a political unit—such as a nation-state—and a people/culture became more than obvious in East Asia when 'the Chinese' were divided into those in the People's Republic of China, those in Hong Kong—first under the British and then the People's Republic of China—and those in the Republic of China (Taiwan) across the Formosa Strait. Likewise, 'the Koreans' were split into North and South. Japan, which has been stereotyped as homogeneous in terms of demographic composition, has been a multi-ethnic nation throughout history. But, to give one example, in 1940 the then prime minister of Japan, Konoe Fumimaro, delivered a now infamous speech that 'one hundred million Japanese with one soul' be dedicated to the service for the Emperor. This figure included 30,000,000 Koreans, Chinese, and others who resided in Japan and its occupied territories, who were given an equal right to death, that is, from 1944 they were forced to become Japanese with the Japanese soul in order to be drafted to fight for imperial Japan. This is not to mention the diversity within, with the Ainu to the north, and the Okinawans to the south, who were once thought of as distinct from those on the main islands, and 'Westerners,' including the 'white Russians' and Jews.

The political mosaic in East Asia offers a fertile ground for studies of 'identities' and conflicts among social groups within a nation-state and between nation-states. Within a nation-state, the majority vs. minorities are interlaced with the question of what makes a social group 'ethnic' (the Okinawans in Japan 'used to be' ethnic); whereas the two Chinas and two Koreas present a different situation—the question of two nation-states consisting of one people. In other words, East Asia offers a microcosm of what we see in the world in which collective identities form flexible circles, some of which overlap with each other.

4. *Nationalisms, Colonialisms, Imperialisms*

At the level of nation-states, the identity of a people could be expressed through peaceful 'cultural

nationalism' which defines and represents a people of a nation-state as distinct from another, often those who, physically or symbolically, press at the 'border'—figuratively and politically. Thus, when the Japanese began to emerge from their total immersion in the Han and Tang civilizations of China, they began to assert their own culture, switching from embracing the Chinese aesthetics of plum blossoms to that of cherry blossoms as their flower. Since the mid-nineteenth century Japan's nationalism, both cultural and political, has been informed by paradoxes and ambivalence of the Japanese's own identity, which was formed in part by rejecting Western political imperialism—while at the same time eagerly emulating Western civilization—and in part by rejecting their own Asian identity—while seeking it in order to establish a difference from the West.

Research on the complex problem of identities in East Asia will join the now burgeoning field within social science of ethnicity, nationalism, and ethnonationalism. Ethnonationalism is 'the politicization of ethnicity,' related to 'the generation of regional or subnational reactions and resistances to what is seen as an overcentralized and hegemoic state' (Tambiah 1996, p. 16).

Colonialism in East Asia is another genre of research in which interest is growing, and which promises a rich source of comparative work. Studies of Western colonializations have yielded rich fruit for scholars in the past. The Opium War (1839–1942) and the resultant cessation of Hong Kong became the symbol of the power of Western (British) colonialism over Asia, whereas the Boxer Rebellion (1899–1900), crushed by the combined forces of Western and Japanese powers, foreshadowed not only Western but Japanese imperialism. Another growing field of interest concerns the histories and cultures of Japanese colonialism during its military period, as in the colonization of Korea (1910–1945) (Duus 1995), and the building of the puppet state of Manchukuo (1931–1945) (Young 1998). These studies offer important comparative bases for Western and non-Western colonializations in Africa, the Americas, and other parts of the world.

A genre which is similarly regarded as important, although it has not yet received systematic attention, is cultural imperialism, especially of Western European cultures over many parts of Asia, and also of the world—how political domination by other, primarily Western European nations (but also Japan to a lesser degree) nonetheless became the cultural mirror to emulate. Western cultural imperialism has been at the center of modernization processes in Asia, with the often eager adoption of Western science, technology, and capitalism leading to almost everything from the 'West' being revered, including the body esthetics. Even before affluence led the Japanese to seek out designer names such as Gucci, Bally, and Louis Vuitton, American Revlon lipsticks and Ivory soap were far more expensive than their domestic counterparts and were preferred as gift items.

This is an important area of study, not only because it offers us a valuable window onto the other side of political colonialism but also because it constitutes a significant dimension of the current globalization, in which the notion of brain death, organ transplantation, human rights, environmental conservation, and the like, which involve concepts and values that are predominantly Western in origin, are considered outside their areas of origin. Medical high technologies, saluted as the golden achievement of Western biomedicine, brings prestige to those in another society who embrace them as being 'progressive,' although such technologies may require not only radical reconceptualization about the death in relation to the body but also mortuary rituals, with beliefs, values, kinship, and a host of other important dimensions of a people needing to be considered and, where necessary, renegotiated.

5. *Global Forces and East Asia*

Another category of studies that offers an approach to study people not as an enclosed entity is 'global studies,' or studies of transnationality—there are a wide range of approaches within this field. Those at the forefront of the study of globalization consider the twenty-first century as a new era marked by an unprecedented flow of people, goods, and signs. Others urge us to compare the present with the past when spice trades, capitalism, colonialism, features of Western High Culture, etc. spread throughout the world. The present globalization is marked by the spread of high technology and market economy of late capitalism. The global flows of people are studied as research on tourism, diaspora, and transnationals. The flows of goods and signs/symbols are studied through research on global dissemination of youth fashion, fast food, television programs, films, pop music, and a host of other things that are engulfing the globe. Most originate in the West, but some, such as reggae, do not. Reassessments of the claims for 'McDonaldization' or 'CocaCola-ization' are underway, for example, in Watson (1997), which offers a close examination of the fast food industries in Beijing, Hong Kong, Taipei, Seoul, and Japan. These studies of the introduction of McDonald's—a symbol of postmodern fast-lane life in the USA and an epitome of Americana for peoples elsewhere—show a simultaneous transformation of the locale, which in turn transforms the fast food in significant ways.

6. *Cultural, Social, Political Institutions*

East Asian studies continue to produce distinguished scholarship on: kinship and social organiz-

ation; religions (beliefs and rituals); educational systems; arts and performing arts; and healthcare systems. Excellent studies of the institutionalized religions of Buddhism and Confucianism, shared by most Asian societies in variant forms, and of Shintoism in Japan describe how they have impacted on sociopolitical organizations as well as arts and science. They also tell us how people experience these religions in their daily lives. The emphasis on the impact of religion in individual lives receives the focus of attention in the studies of 'ancestor worship' shared by most peoples of East Asia (Ahern 1973, Watson and Rawski 1985, Janelli and Janelli 1982, Smith 1974). 'Ancestor worship' is a misnomer but refers to the practice of taking care of the souls of deceased family members and relatives through a long series of rituals. Despite profound changes that such worship underwent in each society, it remains perhaps the most important aspect of religious beliefs and rituals in these societies.

Shamanism and other religious beliefs and practices that have been regarded not to belong to institutionalized religions are also well studied in relation to the day-to-day lives of contemporary peoples. In contemporary Korea, for example, Shamanism is of exceptional importance and recent studies highlight the role of women in Korean shamanism (Harvey 1985, Kendall 1985). Although their pervasive influence is less recognized, Taoism and geomancy, as practiced in China and spread to Japan, constitute important dimensions of cosmology comprising symbolic oppositions (expressed in *yin-yang*) and remain important in the daily lives of people in East Asia.

In addition to these traditions—large and small—of religions that are considered to be 'indigenous' or of long standing in East Asia (Buddhism originated in India), we witness recent studies on Muslims and Christians in East Asia, again showing the recent view of culture as not localized or hermetically sealed. Indeed, Protestantism is quite a strong force in contemporary Korea, while many influential intellectual and political leaders of Japan since the Meiji 'restoration' in 1868 have been devout Christians, some of whom were presidents of the University of Tokyo, taking on important roles as intellectual and moral leaders.

The academic study of healthcare has a long tradition in East Asia. 'Chinese medicine' in particular has received attention from distinguished scholars such as Joseph Needham, who explicated how their medicine constitutes a science distinct from Western science (Needham 1969), and Manfred Porkert, who later produced a similarly distinguished work (Porkert (1974). A compiled work by Arthur Kleinman (1976) focuses on doctors, patients and their interactions, rather than medicine as a system.

See also: China: Sociocultural Aspects; Colonization and Colonialism, History of; East Asian Studies: Politics; East Asian Studies: Society; Globalization and World Culture; Japan: Sociocultural Aspects; Korea: Sociocultural Aspects

Bibliography

Ahern E M 1973 *The Cult of the Dead in a Chinese Village*. Stanford University Press, Stanford, CA

Chen T 1939 *Emigrant Communities in South China*. Kelly and Walsh, Shanghai, China

Duus P 1995 *The Abacus and the Sword: The Japanese Penetration of Korea, 1895–1910*. University of California Press, Berkeley, CA

Geertz C 1995 *After the Fact*. Princeton University Press, Princeton, NJ

Harvey Y K 1985 *Six Korean Women: The Socialization of Shamans*. West, St. Paul, MN

Janelli R L, Janelli D Y 1982 *Ancestor Worship and Korean Society*. Stanford University Press, Stanford, CA

Kendall L 1985 *Shamans, Housewives, and Other Restless Spirits: Women in Korean Ritual Life*. University of Hawaii Press, Honolulu, HI

Kleinman A (ed.) 1976 *Medicine in Chinese Cultures*. US Government Printing Office for Fogarty International Center, Washington, DC

Needham J 1969 *Science and Civilization in China*. Cambridge University Press, New York

Ohnuki-Tierney E 1987 *The Monkey as Mirror*. Princeton University Press, Princeton, NJ

Ohnuki-Tierney E 1993 *Rice as Self: Japanese Identities through Time*. Princeton University Press, Princeton, NJ

Ong A 1999 *Flexible Citizenship: The Cultural Logics of Transnationality*. Duke University Press, Durham, NC

Porkert M 1974 *The Theoretical Foundations of Chinese Medicine: Systems of Correspondence*. MIT Press, Cambridge, MA

Shima M, Janelli R L (eds.) 1998 *The Anthropology of Korea: East Asian Perspective*. National Museum of Ethnology, Osaka, Japan

Smith R J 1974 *Ancestor Worship in Contemporary Japan*. Stanford University Press, Stanford, CA

Tambiah S 1996 *Leveling Crowds: Ethnonationlist Conflicts and Collective Violence in South Asia*. University of California Press, Berkeley, CA

Watson J L 1975 *Emigration and the Chinese Lineage*. University of California Press, Berkeley, CA

Watson J L (ed.) 1997 *Golden Arches East: McDonald's in East Asia*. Stanford University Press, Stanford, CA

Watson J L, Rawski E S (eds.) 1985 *Death Ritual in Late Imperial and Modern China*. University of California Press, Berkeley, CA

Watson R S, Ebrey P B (eds.) 1991 *Marriage and Inequality in Chinese Society*. University of California Press, Berkeley, CA

Young L 1998 *Japan's Total Empire: Manchuria and the Culture of Wartime Imperialism*. University of California Press, Berkeley, CA

E. Ohnuki-Tierney

East Asian Studies: Economics

The East Asian economies can be divided roughly into three groups—the advanced industrial economy of

Japan, the newly industrialized economies (NIEs) of Taiwan, South Korea, Singapore, and Hong Kong, and the transforming planned economy of China. Dramatic changes have reshaped East Asia as a regional economy since the early 1970s. The rapid development of the NIEs and the early success of China's transforming economy have made this region one of the most dynamic in the global economy. However, the economic difficulties wrought by the economic downturn in Japan in the 1990s, the Asian financial crisis of 1997, and the considerable challenges still facing the Chinese reform effort have shown the vulnerabilities of this regional economy.

1. East Asian Economic Development

In East Asia, the second half of the twentieth century was a period of strong economic development in one form or another. The newly industrialized economies of South Korea, Taiwan, Singapore, and Hong Kong were viewed as the most dynamic economies in the world, having leapt, in the postwar period, from the status of poor underdeveloped to strong industrialized economies (Clark and Roy 1997). The last two decades of the century also saw great success in the economic transformation of the Chinese economy. The patterns and trajectories of change within these nations have challenged conventional wisdom about how developing nations modernize and industrialize, and about the role of the state in these processes of change.

1.1 The Newly Industrialized Economies

The three decades following World War II were a period of enormous enthusiasm about the economic prospects for developing countries. These were the years when a large number of such countries embarked on ambitious projects of agricultural and industrial modernization. Indeed, it was this effort that gave rise to 'development economics,' and more broadly, 'development studies,' as new disciplines in the social sciences. In the halcyon years of development planning, most eyes were focused on the larger or more industrialized countries such as India, Brazil, Mexico, and Egypt. Few expected very much, if anything, from East Asia, large areas of which had been colonized by Japan in one way or another in the first half of the twentieth century, and parts of which seemed hopelessly backward. Taiwan had a minuscule industrial base in 1950, Hong Kong and Singapore were tiny city-states, and South Korea seemed to be sinking in a morass of corruption and inefficiency by the end of the 1950s. Yet, a quarter of a century later, the tables had been turned: East Asia emerged as the economic 'miracle' of the postwar period, witnessing some of the fastest growth rates of the modern epoch, while Latin America and South Asia were mired in political and economic crises. Given this regional growth phenomenon, does it make sense to talk of an 'East Asian Model'? There are some important similarities in development strategies across the cases, enough to persuade many commentators that the appellation is warranted. But there are also interesting differences that are worth noting.

Among the East Asian NIEs, South Korea and Taiwan have attracted the most attention. Over the years, two general views have emerged about the reasons behind their spectacular economic success. One, which was perhaps the dominant one for much of the 1970s and 1980s, held that the two countries were successful because of their fidelity to free-market policies, particularly an open trade regime (Lal 1985, Bhagwati 1988). But this view has come under a great deal of criticism from many quarters, and a new argument has emerged in recent years, which offers a different explanation for Korean and Taiwanese success. In some ways, the two countries achieved their remarkable economic success the 'old-fashioned' way: both started with highly successful land reforms, which immediately set them apart from much of the developing world; both managed to bring about extraordinarily high savings rates, which were then channeled into correspondingly high rates of investment. But the two countries did not simply grow by increasing the rate of accumulation. They also succeeded in acquiring, and then unpacking, technology from advanced industrial countries, so that they were able to move up the production chain into manufacturing and exporting high value-added goods. The involvement of the state was crucial to this process (Gold 1985). First, both governments launched highly effective campaigns to increase literacy rates, which accounted for a highly educated work force (World Bank 1993). Furthermore—and this has been one of the most heralded facts about their experience—the two states played a crucial role in facilitating the acquisition and transfer of imported technology. These technologies were brought in sometimes through foreign direct investment, but mostly through licensing agreements with multinational corporations (MNCs). Agreements were carefully crafted to ensure that foreign partners did not simply hand over obsolete technology, and that local engineers were able to acquire blueprints as well as the training necessary to master the technology. Firms were also encouraged, assisted, and even coerced to produce for export markets, in order to submit them to the rigors of international competition (Amsden 1989, Wade 1990). Success in export markets was used by both states as a benchmark for efficient production. Even firms producing mainly for the domestic market often had to show export success in order to be granted credit, investment licenses, and other state benefits (Aoki et al. 1997). Finally, the state was active in maintaining, for many of the crucial years of take-off, a tightly policed and powerless labor force, which added to the 'flexibility' of production.

While South Korea and Taiwan shared these elements in their development strategies, they also exhibited some important differences. Korea emphasized a reliance on the private sector, which was not only encouraged to grow, but also to become increasingly concentrated. In this, it resembled the Japanese experience: the Korean economy was dominated by a small number of highly diversified business groups, known as the Chaebol, which together accounted for the majority of the country's domestic product. The Chaebol quickly emerged as the pivot on which industrial policy turned, establishing a close working relationship with the ruling regime (Amsden 1989). The Taiwanese state, on the other hand, did not have such a close relationship with private capital. Taiwanese planners relied much more on a large agglomeration of state enterprises as the nucleus for development policy; private capital was influenced—in comparison with the Korean case—more through pricing and procurement policies than by direct negotiation. The private sector was, in turn, not actively encouraged to become highly concentrated. Taiwanese industry was, much more than its Korean counterpart, characterized by small enterprises (Wade 1990). Nevertheless, a kind of quasi-corporatist structure was used to bargain with some sectors. Indeed, in both countries, a similar structure was put to use for this: firms in targeted sectors were organized into sectoral or trade associations, which were given the right to bargain with the state (Fields 1995). The overall complexion of industrial policy, however, remained visibly different in the two cases, with Korea relying on a highly concentrated private sector, and Taiwan more on state enterprises and small firms.

Singapore and Hong Kong are usually regarded as contrasting cases to Korea and Taiwan, in that whereas the latter two are examples of state-led development, the former pair embody a free-market path—or so the story goes (Rabushka 1979). There is some measure of truth to this. Hong Kong, in particular, did not experience anything like the selective industrial policy witnessed in the larger countries. And both of the city-states were much more open to foreign direct investment than either Korea or Taiwan. But there are important respects in which the city-states converged with the experience of their larger neighbors. First, like Korea and Taiwan, both Hong Kong and Singapore actively suppressed labor rights. Second, the state was of some importance in facilitating industrialization. Singapore, despite being more open to foreign investment, relied greatly on state enterprises to facilitate the entry of MNCs by forming partnerships with them, hence reducing the risk of such undertakings; further, the domestic market was not thrown open to imports, and was protected much like that in Taiwan and Korea. Local firms were also offered carefully targeted fiscal incentives, under the guidance of an Economic Development Board. Hong Kong was, of the four, the furthest away from the interventionist model. But here, too, the state played a crucial role in solving a host of collective action problems for local firms: most directly through its considerable expenditure on infrastructure, training programs, and, perhaps most importantly, its massive socialization of wage costs through the construction of an extremely generous welfare state. Between 1949 and 1980, government expenditure in Hong Kong grew by 26 times, and expenditure on welfare by an amazing 72 times (Castells 1994). This was absolutely crucial to the success of Hong Kong exports, in lowering the wage costs to producers.

In sum, despite their differences, the four cases do exhibit certain important similarities. Perhaps the most important of these is the critical role played by the state in fostering the transition to highly productive industrialized economies. The differences lie mainly in the instruments used by the various regimes: selective intervention in the larger two countries, and more macro-level policies in the smaller ones.

1.2 China's Transforming Planned Economy

While the NIEs were executing successful development processes, the first three decades of Communist Party rule in China (1950s through 1970s) were organized around the establishment of a socialist planned economy (also called a 'command' economy). In this system, markets were eliminated and replaced by governmental plans that guided the economic decisions of enterprises and individuals. China's one-party government replaced private ownership with state ownership, established economic targets for industrial enterprises, and set up collective farming in the agricultural sector. Economic agents (whether industrial enterprises or farmers) were instructed to whom they would sell their goods and for what price. Revenues were then collected by the government and redistributed to enterprises and individuals in the economy. This system was thus more than a purely economic system; it was also the basis for the social security system in China (Walder 1986, Whyte and Parish 1984).

Without a competitive marketplace to regulate and guide economic behavior, planned economies, such as China's, are often rife with inefficiencies. In 1978, after a decade of upheaval wrought by the Cultural Revolution, the Ten Year Plan of 1976 was already faltering, and the need for some degree of economic reform was clear. Following a political struggle that pitted reform-minded Deng Xiaoping against the economically conservative Hua Guofeng, Deng Xiaoping took control of the economy in late 1978 at the Third Plenum of the Eleventh Central Committee of the Chinese Communist Party. Once in control of the government, Deng launched China on a path of economic transformation.

One of the central debates in research on transitions from planned to market economies revolves around

the extent to which privatization is a necessary part of this process. Some scholars have argued that without rapid and complete privatization—the so-called 'shock therapy' approach to reform—true economic reform is impossible (Kornai 1990, Sachs and Woo 1997). Other scholars have argued that China is the case that disproves this view: in contrast with the rapid changes that characterized the economic reforms of Russia and Eastern Europe, China's reforms have been introduced in a gradual and stepwise fashion, where the government has experimented with new economic policies and incrementally implemented the successful practices on a wide scale (Naughton 1995). Proceeding in this gradual fashion, the Chinese transition to a market economy has been unquestionably the most successful in the world. (China sustained the fastest rates of growth in the world for the first decade and a half of economic reform. Throughout the 1980s, its real GDP grew at an average annual rate of 10.2 percent, a level that was equaled only by Botswana. From 1990 to 1996, the average annual rate of growth for real GDP was 12.3 percent, the highest rate of any country in the world for that period. It has also had the highest industrial growth rate (an amazing 17.3 percent average annual growth 1990–96) and the second highest growth in services (9.6 percent per annum, 1990–96) in the world.)

In the industrial sector, often regarded as the backbone of the Chinese economy, five key changes have reshaped this system. First, the government gradually removed itself from day-to-day decision-making, effectively placing the responsibilities of economic decisions on the shoulders of industrial managers. In the early years of the reforms, enterprises were still expected to produce a certain amount of goods for the government's plan, but anything they produced above this residual plan they were allowed to sell in China's emerging markets. This transitional system was called the 'dual-track' system (Naughton 1995). Second, a private economy was allowed to emerge to compete with the state sector, a factor that was critical for the creation of a competitive marketplace. Thus, while privatization was the cornerstone of industrial reform in other transforming planned economies, Chinese firms were introduced to market economic behavior without privatization (Rawski 1994), and state firms eventually 'grew out of the plan,' producing an increasing proportion of their goods for the marketplace (Naughton 1995). The success of this transition, however, has varied in accordance with a firm's position in the former command economy: firms under township and village governments, most of them the so-called Township and Village Enterprises (TVEs), have been managed closely by local governments and thus have weathered the reforms much more successfully than those under the central and municipal governments (Walder 1995).

A third change that has defined China's economic reforms is that foreign investors were allowed in the Chinese economy for the first time since the communists expunged the foreign presence in the early 1950s. This factor has been very important for the transfer of technology and the transfer of international management practices. Fourth, the government adopted a decentralized and piecemeal approach to development regionally: in the late 1980s, the government adopted the coastal development strategy, in which the goal was 'state-directed' export-oriented growth (Yang 1991). Finally, Chinese enterprises have been relieved of their responsibilities in the provision of social security benefits to their employees and, perhaps more significantly, they are no longer required to make lifetime commitments to the workers they employ. In the agricultural sector, a similar process of change has unfolded, where farmers have been given the freedom to sell whatever they produce beyond the plan in the marketplace, until eventually the market has overtaken any vestiges of an economic plan in this sector. All of these economic changes have been formalized through the passage of some 700 national laws and more than 2,000 local laws (Pei 1994).

Transforming the planned economy of the world's most populous nation has been a complex process and is far from complete. Nevertheless, the first two decades of transition from plan to market have been impressive and, compared to the reforms in Russia and Eastern Europe, a dramatic success. For the first decade and a half of economic reform, China had the fastest growing economy in the world, sustaining double-digit growth figures for much of the 1980s and 1990s. Where it was a third-world developing economy as late as the early 1980s, by the beginning of the twenty-first century China had the seventh largest economy in the world overall in terms of gross domestic product (GDP), and it is second only to the USA when GDP is adjusted for purchasing power within the country. (For example, if a bushel of rice costs half as much in China as in the USA, then a unit of currency is twice as valuable there. GDP statistics can, when appropriate, be adjusted for 'purchasing power parity' (PPP): a statistic that adjusts for cost of living differences by replacing exchange rates with relative measures of local purchasing power.)

Among the challenges that still lie ahead are the continued reform of the state sector—the leviathan organizations that remain inefficient and insolvent after more than 20 years of reform; the continued reform and clarification of property rights; the liberalization of several critical sectors, such as the telecommunications sector, and the development of the western provinces, which have fallen far behind the development in the eastern coastal region.

2. *National and Regional Economic Crises*

It is, in part, because of the great successes of development in East Asia that recent economic crises have hit the region so hard. Two related, but distinct,

economic crises are worthy of note here. The first is the economic downturn that Japan suffered in the second half of the 1990s. The second is the Asian financial crisis, which gripped much of East and Southeast Asia in 1997 and 1998.

2.1 Japan

Japan was, for most of the postwar period, one of the fastest-growing economies of the industrial world, racking up huge rates of growth year after year until it had achieved its current position as the second largest economy in the world, with an annual GDP of more than $5 trillion. Its economy was the engine for the broader regional development of East Asia, both through its capacity as a market, and as a source of capital and technology. A more intangible, but nevertheless important, role was that it provided the model on which the two Northeast Asian 'tigers'—Taiwan and South Korea—based their economic strategies.

By the end of the twentieth century, however, the Japanese dynamo seemed to be losing its energy. It spent most of the 1990s mired in an economic recession that turned out to be the longest and most severe of its postwar history. The precipitating event behind the crisis was the collapse of the asset bubble of the 1980s, which burst in the beginning of the next decade. The bubble itself had been built up during the 1980s as the Japanese state had initiated a policy to ratchet upwards the domestic rate of investment and make the economy less dependent on external markets. Spurred on by the constant pressure by the USA to raise the value of the yen, the Japanese government moved by the late 1970s to trigger a massive investment drive by raising the value of domestic assets. At one level, it worked very well—as asset prices rose, Japanese firms launched a massive drive to retool and invest in new technology. But the flow of investment soon exceeded anything that actual returns could justify, driving up asset prices steadily into a speculative bubble. Sooner or later all speculative bubbles burst, as did this one in the 1990s. And once the bubble burst and the economy slowed, firms found themselves stuck with massive overcapacity and onerous levels of debt. Through much of the decade, business and residential investment continued to fall, throwing the economy into a downward spiral.

The Japanese downturn was one of the background factors that laid the conditions for the East Asian crisis of 1997–8, which is now regarded as the most significant economic catastrophe the world has seen since the Great Depression. As the economy slowed down, firms compensated by turning outward, in a massive wave of investment in Southeast Asia. This led, in turn, to a process not unlike that witnessed in Japan: corporations went into deep debt to pay for new investments, leading to the build-up of extraordinary industrial capacity; after a point, money also began to pour into real estate and other forms of wealth, driving up their prices until a speculative bubble was created reminiscent of the Japanese bubble.

2.2 The Asian Financial Crisis

Though much of East Asia had bounced back from the Asian financial crisis by the spring of 2000, this crisis revealed some weaknesses in the rapid development of the 'Asian Tigers.' The crisis itself was set off by currency devaluation: in the spring of 1997, as the pace of growth in the Thai economy fell below projections, currency investors began to speculate against the baht in hopes of profiting from the inevitable devaluation of the currency. When the currency devaluation came in July, foreign investors fled from Thai markets, and similar behavior spread quickly throughout the region. Stock prices plummeted, currencies went into free fall, and banks began to call in loans that their customers—many of them large corporations—could not repay. By the end of 1997, the International Monetary Fund had bailed out Thailand, Indonesia, and South Korea with $17 billion, $43 billion, and $57 billion, respectively.

Two explanations have been advanced to account for the crisis. In one view, an over-reliance on foreign capital and international institutions, changes in regional economies—namely China's and Japan's—and reliance on the dollar brought about the crisis (Ahn 2000). First, the strength of the Japanese yen in the 1980s and the dollar in the 1990s led to excessive foreign investment in the strong East Asian economies, particularly for Japanese investors. As local banks sought strong currencies, investment and lending agreements were structured to attract these foreign players, and excessive capital led to high-risk investments, many of which would prove to be insolvent when the crisis stripped the veneer of economic discipline. Second, as China's economy began to expand, it took over a large share of the export market, the centerpiece of development for many East Asian Economies. Third, when European currencies depreciated against the dollar in 1995, dollar-pegged East Asian currencies appreciated—relatively speaking—leading to a decline in exports. Fourth, as the flow of international capital increased through such institutions as hedge funds and short-term loans, speculative behavior on the part of international currency investors in these rapidly expanding financial markets led to market volatility.

A second explanation acknowledges the problems listed above, but places the blame of the crisis on the East Asian economies themselves. This explanation often veers into a cultural argument about the prevalence of 'crony capitalism' in the region. In this view, the East Asian business model, often led by government-directed investment in key industries, is also often built around implicit guarantees from the government for the liabilities of local financial institutions

(Krugman 1998). Thus, the so-called 'soft budget constraint' (Kornai 1994), which was the hallmark of planned economies, some economists argue, was also fundamental to the East Asian business model more generally. Governmental guarantees led to risky behavior on the part of financial institutions, which ultimately led to the financial crisis itself.

3. *The Balance of Power: Regionally and Globally*

Despite the economic crises experienced in the region in the 1990s, overall the second half of the twentieth century saw a dramatic shift in the economic balance of power both intra-regionally and with regard to Asia's position globally. Within the region, while Japan was far and away the economic powerhouse of the region by the end of the 1980s, at the dawn of the twenty-first century the balance of power looked very different. Partly because of the rise of the NIEs, but perhaps more importantly, because of the dramatic growth of China's economy, Japan's economic dominance in the region has unquestionably waned. It is difficult to predict what the exact economic balance of power will be within the region in the coming decades, but it is certainly clear that China will challenge Japan as the center of the Asian economy.

Globally, the economic balance of power has shifted as well. While Japan has been counted among the group of seven (now group of eight) industrialized nations since the organization's inception, during most of the twentieth century Western economies have dominated the global scene. However, with Japan and China both among the top 10 largest economies in the world and four Asian economies among the top 20 (five if we include Australia), Asia is certain to play an increasingly important role in the balance of the global economy.

See also: Commodity Chains; Development and the State; Development, Economics of; Development: Rural Development Strategies; Development: Socioeconomic Aspects; Development Theory in Geography; East Asian Studies: Culture; East Asian Studies: Politics; East Asian Studies: Society; Economic Development and Women; Globalization: Geographical Aspects; Globalization: Political Aspects; Industrialization; World Systems Theory

Bibliography

Ahn S 2000 Reconsidering the Asian financial crisis: A comparative study of Korea, Thailand, and Indonesia. *East Asian Review* **12**(2): 69–86

Amsden A H 1989 *Asia's Next Giant: Korea and Late Industrialization*. Oxford University Press, New York

Aoki M, Hyung-ki K, Okino-Fujiwara M 1997 *The Role of Government in East Asian Economic Development*. Cambridge University Press, Cambridge, UK

Bhagwati J 1988 *Protectionism*. MIT Press, Cambridge, MA

Castells 1994 *Technopoles of the World: Making the Twenty-first Century Industrial Complexes*. Routledge, New York

Clark C, Roy K C 1997 *Comparing Development Patterns in Asia*. Lynne Reinner, Boulder, CO

Fields K J 1995 *Enterprise and the State in Korea and Taiwan*. Cornell University Press, Ithaca, NY

Gold T 1985 *The State and Development in the Taiwan Miracle*. M E Sharpe, New York

Kornai J 1990 *The Road to a Free Economy: Shifting from Socialist System: The Example of Hungary*. Norton, New York

Kornai J 1994 The place of the soft budget constraint syndrome in economic theory. *Journal of Comparative Economics* **26**(1): 11–17

Krugman P 1998 Saving Asia: It's time to get radical. *Fortune* **138**(5): 74–80

Lal D 1985 *The Poverty of Development Economics*. Harvard University Press, Cambridge, MA

Naughton B 1995 *Growing Out of the Plan: Chinese Economic Reform 1978–1993*. Cambridge University Press, New York

Pei M 1994 *From Reform to Revolution: The Demise of Communism in China and the Soviet Union*. Harvard University Press, Cambridge, MA

Rabushka A 1979 *Hong Kong: A Study in Economic Freedom*. University of Chicago Press, Chicago

Rawski T G 1994 Progress without privatization: The reform of China's state industries. In: Milor V (ed.) *Changing Political Economies: Privatization in Post-Communist and Reforming Communist States*. Lynn Reinner, Boulder, CO

Sachs J D, Wing Thye W 1994 Experiences in the transition to a market economy. *Journal of Comparative Economics* **18**(3): 271–5

Wade R 1990 *Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization*. Princeton University Press, Princeton, NJ

Walder A G 1986 *Communist Neo-Traditionalism: Work and Authority in Chinese Industry*. University of California Press, Berkeley, CA

Walder A G 1995 Local governments as industrial firms: An organizational analysis of China's transitional economy. *American Journal of Sociology* **101**: 263–301

Whyte and Parish 1984 *Urban Life in Contemporary China*. University of Chicago Press, Chicago

World Bank 1993 *The East Asian Miracle*. Oxford University Press, Oxford, UK

Yang D L 1991 China adjusts to the world economy: The political economy of China's coastal development strategy. *Pacific Affairs* **64**: 42–64

V. Chibber and D. Guthrie

East Asian Studies: Gender

1. *The Concept of 'Gender'*

Gender is an analytic term that came into common academic usage in the late 1970s and early 1980s in the United States among feminist social scientists. The widely influential historian Joan Scott urged historians to use the term gender to describe women's social and historical differences from men positively (Scott 1986).

By the early 1990s, debate over the gender concept had widely expanded. Judith Butler's theoretical work (1990) established gender to mean the cultural practices that make sexual natures of all kinds appear to be fundamental, heterosexual, and unchanging. Simultaneously a publishing boom of high-profile volumes began that, over the 1990s, has firmly secured the notion of gender into questions of national or political difference. Other volumes (e.g., Sangari and Vaid 1990, Gilmartin et al. 1994) have linked the gender concept to the problem of colonialism in what have become the fields of postcolonial studies, comparative feminism, and gendered area studies.

2. *The Policies of Gender*

Gender is modern political policy. Presently it forms an element of the post-Cold War global governance goals disseminated through the United Nations, so-called nongovernmental organizations (NGOs), and donor or civil society organizations. This post-1989 neo-liberal gender policy is juridical. It argues that all states should adopt the position that women's rights are universal human rights. Gender as state policy, however, is older and more historical than current rhetoric would have it. In the twentieth century, industrialized and late industrializing states have cast citizenship in gendered terms.

2.1 *Contemporary Gender Policy, Global Governance, Women's Human Rights, 1979–*

The importance of gender as policy emerged in the initiatives for global governance over the last quarter of the twentieth century. It is an aspect of the international human rights law movement. Its cornerstone is the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), which seeks to abolish 'gender-based violence' and is a 1979 revision of the 1948 Universal Declaration of Human Rights. The United Nations, donor agencies (like the Ford Foundation), and a newly arisen civil society sector began advancing gender-sensitive programs during the United Nations Decade for Women (1975–85) and during subsequent United Nations Conferences on Women, culminating in the 1995 Beijing Fourth World Conference on Women and its resulting 'Platform for Action.' As a whole, these institutions and the World Bank's investments promote the view that gendered difference is an important axis of inequality. Many advocates in the United States blame non-Western states for gendered inequality. They seek to use gender as a development index to force perceived noncompliant states to redress the relative poverty and vulnerability of female citizens. Loans, investment, education, and outreach into semi-official social service agencies are the instruments of global governance. Though a lack of clarity prevails about whether these attempts are effective or a US form of triumphalism, Charlotte Bunch and Hillary Clinton are the movement's best-known leaders.

3. *Historical Gender Policies of Modernizing States in 'East Asia': The 1840s to the 1980s*

The novelty of contemporary uses of the gender analytics tends to obscure the longevity of gender modernization policies throughout Asia-Pacific. Scholars working in the framework of the 'East Asia' area studies during the 1980s and 1990s have established that much. It can be said that colonialism itself advanced the rationalization of gender by bureaucratizing social modernity or (in some cases) by directly exploiting the bodies of women. During the mid-nineteenth and twentieth centuries, the forces of imperialism, colonialism, and sub-imperialism propelled late industrializing states toward expeditious efforts of political modernization and capital accumulation. States pursued this objective in various ways, although interstate relations meant that events in one state (e.g., Japanese imperialism in Korea, Manchuria, Taiwan) contributed to the shape of gender policy in other states.

3.1 *Colonial Gender Policy in Modern Korea*

Between the 1870s, when its traders arrived in Korea's overwhelmingly agricultural economy, and 1945, when Japanese colonial administration evacuated and the neo-colonial United States occupation began, the peninsula took on the classic colonial configuration of a political economy twisted to serve Japanese imperial expansionism. Japan's colonial authorities reordered Korea's rural economy, instituting modern concepts of ownership to expedite taxation. The new system placed tax gathering and relations of production in the hands of a semi-colonial and semi-feudal native elite, 'landlordizing property rights'. This industry led to a widespread shift from peasantry to tenancy as the primary relation of production. Agricultural productivity rose, but so did rural immiseration. Poverty, occupation, and assimilation policies flowered into a widespread licensed military prostitution industry (see Choi 1997). This, and particularly Japanese family law, burdened Korean gender and class relations and helped to shape a male-centered style of nationalist anti-colonial struggle, involving women and men of all classes (see Park 1998, Fujime 1997).

In divided South Korea after the Pacific War, a state-compelled development that Alice Amsden calls 'late-industrializing' (and argues is a global model) reconfigured Korean social relations of gender (Amsden 1989). The incorporated Cold War state

inherited a Europeanized, Japanese, colonial civil code. As Hyunnah Yang has shown, along with the modern legal regulation of domestic relations the civil code supported an imaginary 'Korean tradition' that naturalized female subordination. Consolidation of Japanese domestic civil statutes and modern family law into Korean jurisprudence in the Korean Republic (during and after the partition of the nation) also ratified a gendered modernity, heavily weighted to the nuclear, patrilineal, small family unit. The now constitutional doctrine of patrilineality triggered waves of feminist intervention into state-sanctioned subordination of women and women's labor power under the US-supported military regime (Yang 1998). Feminist and labor movement successes have led to revisions of the family law in 1962, 1977, and 1989. However, the fact remains that the modern state—more perhaps than the dynastic regime—consolidated domestic property, family jurisdiction, domestic labor, and legal claims in the hands of the patriline and restricted the claims of female kin. Little is known of the gendered state policy of North Korea, but it appears to be modeled on that of the People's Republic of China (PRC) during the Cultural Revolution (1966–76).

3.2 Colonial Modernity, Imperialist Economy, National Gender Policy in Modern Japan

Historians are agreed that the Meiji state (1868–1912) was preoccupied with gender policies from the outset of its drive to modernize Japan on the European model (under threat of Western imperialism) (see Bernhardt 1991, Sievers 1983). The Meiji civil code of 1898 ideologically reformulated the patriarchal family and its relation to the state, emphasizing the political obligation of scientific housewifery, or 'good wife, wise motherhood.' Severe restrictions on women's political rights were coupled with the statist argument that the labor of women should be considered a civil service to the nation. Deliberate gender policies like these persisted, toward varying ends, throughout the twentieth century. First, the new state-sponsored industrialization policy willfully and strategically regendered the household division of labor by moving men out of domestic production into heavy industry. Second, policy affecting men and children led to the official drive in the late nineteenth century to 'reshape conceptions of womanhood' (Uno 1991). Fostering smaller family formations, female education, young women's heavy participation in waged light industrial production, instituted bans on female public political participation, as well as campaigns encouraging feminine virtues in household and reproductive labor affected growing sectors of the national population. A twenty-year effort to bureaucratize the policy on women culminated in the 1898 implementation of Japan's prewar civil code.

Capital accumulation through industrial production and constitutional state formation in Japan rested on colonial expansion onto the Asia mainland, into Korea and Taiwan, and eventually throughout the Asian Pacific. Despite its best efforts, the state had never been able to contain female wageworkers from moving into the labor market. As militarization and colonial administration drained off increasing numbers of men into service, state policy began gendering the notion of national sacrifice. Men were to provide military or colonial service. Militarized policy on women sought to rework the family into a feminine sector of the public sphere. The onset of glorification of national motherhood is datable to this era. It resurfaced (in altered form) in immediate postwar policies.

After 1945, state policy sought to emphasize domestic reproduction and super-exploit male labor in an economic drive to recover Japan's regional supremacy under the umbrella of a US-dominated Cold War containment policy. Adjunct to this was an effort to limit female participation in Japan's national labor market while increasing (after 1956) women's responsibility to undertake domestic consumption responsibilities.

As in the Korea case, the history of gender and the nation sometimes precludes an emphasis on the Japanese women's rights movement, to say nothing of a transnationally influential, Japanese feminist critical theory that arose with colonial modernization and state formation. Nonetheless, as Mariko Asano Tamanoi noted in her study of rural women and nationalist ideology, the term 'national woman' is too simplistic and does not convey the full import of gendered social change in the twentieth century (Tamanoi 1998).

3.3 Sub-imperialism, Colonial Modernity, and Policies of Gender Revolution in Modern China

The Chinese record of state-mandated gender policy is intricate. That is due to its enormous size vis-à-vis other East Asia nation-states, its semi-colonial status after 1842, a full half-century of multiply sited civil wars, and the high priority which the Chinese Communist Party accorded to women's liberation. As early as the compromised Republican era (1911–49), the state was initiating gender and development policies that were later absorbed into influential Party state policies associated with Maoism and the revolutionary mobilization policies of the PRC.

National gender policy first appeared in the form of state-mandated reforms of education, female dress, and foot-binding practices. A juridical modernization indebted to both continental European and Japanese models was instituted in the 1929–31 modern civil code and, under the Nationalist or Republican Party

state (1911–49), mobilized women into participatory forms of citizenship. A further link exists between the weakly instituted republican civil code (organized around a liberal divorce law) and campaigns during the 1940s and 1950s to propagate the Chinese Communist Party's Marriage Law (Bernhardt 1994).

The ethos that infused revolutionary gender ideology came from the early twentieth-century Chinese Enlightenment: an anti-colonial, urban, nationalist movement led by modern intellectuals. The Chinese Communist movement, which incorporated feminist themes, was—relative to contemporary states—hospitable to policies that institutionalized women's liberation. But, like other states, women's liberation became an index of national modernity, particularly, perhaps, under Chinese Communism. What most distinguished the gender policy of the early Chinese communist states was their focus on mobilization as a means of institutionalizing political policy and their organization of gender revolution around a Marriage Law that put the conjugal couple at the center of jurisprudence (Gilmartin 1995).

Land reform and class war in the late 1940s and early 1950s and the state policy of rural communization in the early 1960s were no doubt two of the most transformative revolutions of the twentieth century. These revolutions consolidated the gains of Chinese state feminism by placing laws on the books that governed the social labor of women and granted women the rights to own (however briefly) land in their own names (Johnson 1983.) Socialist development policies, which were aimed at consolidation of state power and primitive accumulation for investment in late, socialist industrialization, rested on the nuclearization of family resources in the urban areas. They also proletarianized women and periodically used female labor as a reserve pool. The mobilization of peasant women into agricultural production furthered primitive capital accumulation as well as mechanisms to extract rural capital gains for use in urban development. Nonetheless, recent studies point to the ways that particularly poor peasants took advantage of the modern state institutions to further personal goals, particularly women seeking divorces. The socialist state's welfare policies had a wide-reaching impact on several issues: the gendering of labor, domesticity, sexuality, and demography; civil rights and access to public space; and avenues of upward political mobility.

4. *Economic Liberalization, the Displacement of 'East Asia' as a Region, Gender Scholarship and Transnational Flexible Economies*

Just as historically 'modernization' is usually linked to Europe and the United States in the nineteenth and twentieth centuries, neo-liberal 'globalization' is linked to Asia-Pacific in the last third of the twentieth and probably much of the twenty-first century. Asia-Pacific is a laboratory of post-Fordist flexible capital accumulation inside a post-Cold War international state system. The Asian Pacific roots of this neo-liberal globalization lie in:

(a) the Chinese economic reforms of the 1970s;

(b) the hyper-development of newly industrialized countries (NICs) and city-states (Taiwan, South Korea, Singapore, Hong Kong);

(c) Japanese capital investment in Malaysia and other South East Asian countries;

(d) the rise of flexible Chinese diasporic capitalism; and

(e) US sponsored neo-liberal policies of the World Bank and the General Agreement on Tariffs and Trade (GATT) system.

Within intra-Asian political economies, in so-called transnational gendered labor flows, and outside it in interregional global trade, policy elites in Asia-Pacific states and national corporations are scaling back modernization strategies based on a state welfare logic (see Higgott 1999, Ong and Nonini 1997).

No longer central to development are the projects that had earlier supported the public ideology of a state or nation-typical, modern woman and domestic relations, although the use of maids among the middle class is stabilizing the elite family structure in some regions. It is premature to speculate what specific gendered configurations in subregions and specific locales will emerge in the future. However, gender is becoming a central concept in writing throughout the Asia-Pacific region and the United States. Scholars globally seek to explain how liberalization of economic life affects populations that are now defined in relation to their consumption power or their capacity to be absorbed into the flexible service sector (e.g., tourism, prostitution, travel) as well as their labor power. Heightening class differentials separating consumers and service providers, surging global inequality explained in cultural terms, ethnic and race-coded discriminations, systematic preference accorded flexibility in labor migration, and capital investment patterns across state lines (but within regional cores) are all having an impact on the way that the concept 'gender' (in Chinese usage 'social gender' or *shehui xingbie*) will be employed. Transnational corporations (TNCs) now recruit and market with the concept in mind. New social movement policy strategists, networks of internationalist progressive scholars, United States feminist human rights lawyers, World Bank lending policy makers, United Nations monitors of prostituted labor and infectious diseases: all are incorporating the policy and analytics of 'gender' into their thinking.

Particularly since the finance capital crisis of 1997–8, it has been increasingly apparent to observers that a new regional production of scholarship is on the rise. Produced by diasporic scholars from Asian locales resettled in the metropole, or writing from 'new

metropolises' of Singapore, Hong Kong, Seoul, Jakarta, Tokyo, Beijing, and Sydney, 'conscientized' scholars in the First World, and expatriate Asians in the white nations of Europe and the US, this scholarship is eroding the boundaries of area studies, properly speaking. Whether we are speaking of scholarship on gender inequality which Women's Federation scholars in Beijing produce in Chinese for national circulation, or studies produced in Tokyo and Seoul in English for international dissemination, or the steady translation of mainline US feminist and gender theory to a wide number of gender studies institutes throughout Asia, the full impact of localized scholarship on US-oriented area studies has yet to be registered. The trend toward globalized scholarship is exemplified in the work of the journal *Positions: East Asia Cultures Critique* which publishes research that questions strict national divisions on which conventional area studies have rested and also injects the gender analytics into scholarship. The appearance of pan-Asian or regional cultural studies journals like *Inter-Asian Cultural Studies* (Taiwan), the *Asian Journal of Women's Studies* (Korea), and many others is also blurring the edges of area studies and area-oriented gender studies.

The effort to grasp the implications over the years 1970–2000 of re-regionalization (as well as the rise of flexible capital and gendered consumer markets) is producing significant lines of innovative global scholarship. The concepts of 'transnational Chinese capitalism' and 'flexible citizenship' allow Aihwa Ong, for instance, to propose the emergence of ethnicized capital flows which are not nationally grounded and which rest on the persistent powers of the state, opportunistically taking advantage of state forms (Ong 1999). Modern Chinese transnationalism, its distinctive strategies of capital accumulation, and gender analytics are all being studied empirically. Specialists consider gender as a significant axis among other axes, such as ethnicity, nationality, sexuality, racialization, and caste status. A special issue of the journal *Positions: East Asia Cultures Critique* (1999) on 'Asian Transnationalisms' has highlighted these dynamics in relation to ethnicity and sexuality.

Advocates and critics of liberal globalization alike recognize the enlarged opportunities for public sexual expression opening up in the new metropolises of Pacific Asia. Sexualized and commodified styles of self-presentation open fresh possibilities for new social movements of minority social groups, such as self-styled queers (Ho 1998). Minority sexual communities—gay, bisexual, non-procreative heterosexual—exist in most of the new metropolises. The centrality of the gender analytics to scholarship about these communities and cultural criticism from within them are very clear in queer social, cultural, cinema, and theoretical studies which presume trans-nationality and track how sexual expression is gendered among specific groups of queers (see Manalansan 1997, Hanawa 1994). It is premature to consider how crosscurrents between queer criticism and a retrenched or renaturalized market heterosexuality will play out. One critic has shown (in a field study of young, heterosexual, urban, upwardly mobile, Chinese women) that claims to sexual pleasure have led to enriched lives. Yet, sexual rights have proved consistent with authoritarian forms of governmentality (Erwin 2000).

Another avenue for innovation in gendered area studies of the re-regionalized Asia-Pacific has been the political economy of ethnic coding in intra-national, transnational and trans-border contexts. Diasporic ethnicities such as the Hmong are trans-Asian minorities whose communities are now globalized. Ethnic and national coding have been significant to studies of minority service work in tourism and in gendered, often sexed, labor flows of subaltern national citizens from the Philippines into the boom economies of Taiwan, South China, and Hong Kong. Ethnographers and theorists are pursuing a combined transnational and ethnic focus that promises to clarify the political complexity of the new Asia Pacific and its new social movements.

5. *Conclusion*

The analytics of gender offers increased penetration of markets as easily as it offers progressive scholars another significant index for gauging relative oppression and appropriately representing those who cannot represent themselves. Certainly advertising, cultural markets, and lifestyle expenditure in the ongoing consumer revolution refiguring the region are naturalizing heterosexual gender norms around anatomical sexual difference and the capacity to consume and produce what is appropriate to one's 'nature.' Increasingly feminist critique has taken up the problems endemic in consumer society as well as extending its traditional concern for poverty, anti-sexism, and freedom from reproductive tyranny and disease. Undoubtedly, scholarship will look increasingly at the conjugal pair, partly in order to queer commercial norms and partly to expose the political culture's reliance on a heterosexual mode of consumption.

Another task for gender analytics in Asia-Pacific area studies is to document the vast and increasing rifts that separate women in the same society, as well as the economic differentials between women and men affecting specific sectors. A related project will be to follow up on Li Xiaojiang's comment (2001) that the struggle for gender equality is a long historical struggle for all sectors and subgroups, classes, and class fragments, and should be studied as such. A further probable area of new research may be the political economy of pan-'Asianism' and the cultural production of women in transition in many of the liberalized state economies (Barlow 1998). How the

new social movements engaged in Asia-Pacific (as elsewhere) proceed in the years 2000–10 may have some impact on how the gender analytics is employed in state policymaking and that, in turn, will affect the problems taken up in academic venues.

See also: East Asian Studies: Culture; East Asian Studies: Politics; East Asian Studies: Society; Nation-states, Nationalism, and Gender

Bibliography

Amsden A 1989 *Asia's Next Giant: South Korea and late industrialization*. Oxford University Press, New York

Barlow T E 1996 Theorizing Woman: Funü, Guojia, Jiating. In: Scott J (ed.) *Feminism and History*. Oxford University Press, Oxford, UK

Barlow T E 1998 Green blade in the act of being grazed: Late capital, flexible bodies, critical intelligibility. *Differences* **10**(3): 119–58

Bernhardt K 1994 Women and the Law: Divorce in the republican period. In: Bernhard K, Huang P, Allee M (eds.) *Civil Law in Qing and Republican China*. Stanford University Press, Stanford, CA

Bernhardt G (ed.) 1991 *Recreating Japanese Women, 1600–1945*. University of California Press, Berkeley, CA

Butler J 1990 *Gender Trouble: Feminisms and the Subversions of Identity*. Routledge, New York

Choi C 1997 (ed.) The comfort women: Colonialism, war, and sex. Special Issue of *Positions: East Asia Cultures Critique* **5**(1)

Erwin K 2000 Liberating sex, mobilizing virtue: Gender, family values, and transnational desires in Shanghai, China. Unpublished dissertation. University of California, Berkeley, CA

Fujime Y 1997 The licensed prostitution system and the prostitution abolition movement in modern Japan. In: Choi C (ed.) *Special Issue of Positions: East Asia Cultures Critique* **5**(1): 135–70

Gilmartin C K 1995 *Engendering the Chinese Revolution: Radical Women, Communist Politics, and Mass Movements in the 1920s*. University of California Press, Berkeley, CA

Gilmartin C K, Hershatter G, Rofel L, White T (eds.) 1994 *Engendering China: Women, Culture, and the State*. Harvard University Press, Cambridge, MA

Grewal I, Gupta A, Ong A (eds.) 1999 Asian Transnationalities. Special Issue of *Positions: East Asia Cultures Critique* **7**(3)

Hanawa Y (ed.) 1994 Circuits of Desire. Special Issue of *Positions: East Asia Cultures Critique* **2**(1)

Higgott R 1999 The political economy of globalization in East Asia: The salience of 'region building.' In: Olds K, Dicken P, Kelly P F, Kong L, Yeung H W (eds.) *Globalization and the Asia Pacific: Contested Territories*. Routledge, London

Johnson K A 1983 *Women, the Family, and Peasant Revolution in China*. Chicago University Press, Chicago

Manalansan M IV 1994 (Dis)Orienting the body: Locating symbolic resistance among Filipino gay men. *Positions: East Asia Cultures Critique* **2**(1): 73–90

Mani L 1990 Contentious traditions: The debate on sati in colonial India. In: Sangari K, Vaid S (eds.) *Recasting Women: Essays in Indian Colonial History*. Rutgers University Press, New Brunswick, NJ

Ong A 1999 *Flexible Citizenship: The Cultural Logics of Transnationality*. Duke University Press, Durham, NC

Ong A, Nonini D M (eds.) 1997 *Ungrounded Empires: The Cultural Politics of Modern Chinese Transnationalism*. Routledge Press, New York

Park H O 1998 Ideals of liberation: Korean women in Manchuria. In: Kim E, Choi C (eds.) *Dangerous Women: Gender and Korean Nationalism*. Routledge Press, New York

Sangari K, Vaid S (eds.) 1990 *Recasting Women: Essays in Indian Colonial History*. Rutgers University Press, New Brunswick, NJ

Scott J W 1986 Gender: A useful category of historical analysis. *American Historical Review* **91**: 1053–76

Sievers S L 1983 *Flowers in Salt: The Beginnings of Feminist Consciousness in Modern Japan*. Stanford University Press, Stanford, CA

Tamanoi M A 1998 *Under the Shadow of Nationalism: Politics and Poetics of rural Japanese Women*. University of Hawaii Press, Honolulu, HI

Uno K 1991 Women and changes in the household division of labor. In: Bernstein G (ed.) *Recreating Japanese Women, 1600–1945*. University of California Press, Berkeley, CA

Li X 2001 From modernization to globalization: Where are Chinese women? Trans. Barlow T. Special Issue on Gender and Globalization. *Signs*

Ho J 1998 Kuier lilun yu zhengzhi. Special issue on queer politics and queer theory. *Working Papers in Gender/Sexuality Studies* Zhongli guoli daxueh, Zhongli, Taiwan

Yang H 1998 Envisioning feminist jurisprudence in Korean family law. Unpublished dissertation. New School for Social Research, New York

T. E. Barlow

East Asian Studies: Geography

East Asian studies seen through the prism of geographic analysis is a subset of the larger interdisciplinary field of area studies which developed during and immediately after World War II as the need for understanding foreign areas became more imperative in the US with its growing role in global affairs. Geography, with its emphasis on the study of people and places and their interrelationships, occupied an important niche, especially in East Asia, a region where extensive military operations and commitments were involved and where major perceived strategic interests of the United States were at stake.

1. Changing Views and Realities of East Asia

The study of East Asia from a geographic perspective in recent years has been strongly influenced by events in the region and the unfolding drama and trajectory of development within the regional units—China, Japan,

and the Koreas—as well as by changing approaches and methods within the social sciences. China is defined here in its greater context of the People's Republic of China (PRC), Hong Kong, Macao, and Taiwan. In addition, the shifting global geopolitical and economic trends and realities have exacerbated the challenges to the region and its national units. Meanwhile, area studies as an approach to knowledge has been undergoing far reaching changes as new conceptual trends in the social sciences and humanities have led to new ways of looking at how nations are structured and cultures operate. Thus, the nature of economic and social behavior within the national units has been reexamined in the context of the remarkable economic growth of Asia during the last 35 years and the emergence of East Asian states such as Japan and China as major global economic and political players with increased power.

Paradoxically, while political and military/security issues have made the most dramatic headlines, geography research and study of the region and its subunits has focused mainly on economic, urban, and regional development studies including trade, regional growth, urbanization, migration, and environmental issues (see, e.g., Yeung 1998). Much if not most of this work has followed a well-established and more traditional paradigm of empirically based, positivist work in the social sciences and has employed conventional quantitative and statistical methods rather than more recent postmodernist and qualitative approaches.

2. *New Methodologies and Technologies*

At the same time geographic analysis employed in regional studies of East Asia has begun to embrace emerging methodologies of GIS (geographic information science), digital cartography, and remote sensing from earth satellite and conventional aerial platforms. These emerging methodologies focus on acquiring spatial data on natural and human patterns and phenomena and storing, manipulating, and analyzing these data digitally and numerically. By such means and methods, much larger sets of data can be reviewed and analyzed, which not only permit greater knowledge and understanding, but may result in the development of new approaches to the study of East Asia's geography and new ways of conceptualizing regional and spatial arrangements.

Owing to the newness of these methodologies and data-acquisition sources, their impact and significance continues to evolve in step with ability to comprehend and harness their remarkable capacities in new and innovative ways. One significant trend associated with these technologies is their use in environmental studies and in the interface between human and environmental relations and linkages. While the effects of this trend are only beginning to be seen, given the power and utility of the new tools and methods it is likely their effect will be to encourage more studies that focus on the environment and the manner in which humans organize their economic, political, and social activities in space to respond to the environmental opportunities and realities that confront them. At the same time, given the enormous environmental changes underway and the growing challenge for local and regional environments to support the massive East Asian populations with food, fiber, water, and other essential elements for development, much remains to be done if geographic study in East Asia is to provide timely and valuable service for the social and behavioral sciences (Smil 1993).

3. *Challenge of Regional Diversity*

Another striking reality about East Asia that affects the manner in which geographic analysis is employed is the range of social and economic systems that operate across what is only a partially coherent cultural region. Thus, Japan operates a relatively open democracy and managed market economy. South Korea has followed a similar path with closer relations between government and major private enterprises or *chaebol* (large multifaceted commercial/industrial conglomerates). China, by contrast, is a massive socialist economy in transition that has undergone far-reaching restructuring, in which it has sought to rely on market mechanisms to increase efficiencies in production and output, and to increase dramatically the productivity of its land, labor, and capital resources while maintaining a high degree of political control and considerable economic control. Much recent work in geography is aimed at seeking better understanding of China and its growth and development.

3.1 *Contributions of Asian Researchers*

One of the salient trends of the last two decades in Chinese geographic study is the emergence of a large cadre of young scholars from the region—China, Taiwan, and Hong Kong—who are aggressively engaged in research and who frequently publish the results of their work in English. These young scholars are scattered through Asia, Oceania, North America, and to a lesser extent Europe, but their contributions have been very significant in bringing a Chinese and Asian view coupled with methodological and substantive training in both Asia and the West. This blending of intellectual traditions paired with training in the latest methods and fashions of Western science and social science provides creative and highly innovative approaches to study. It often results in scholars who are based in Western universities but who maintain very close contact with the East Asian setting

and point of origin, and who bring a special insight to bear on the analysis of social, political, and economic issues in the region (Chan 1994).

4. Theoretical Matters and Substantive Studies

Examples of such geographic research on China have centered on analysis of regional development and economic growth, associated with foreign investment as linked to China's recent economic growth and restructuring. Distinctive among the social sciences, economic geographers have centered their work on spatial issues and locational questions related to regional disparities in income, output, investment, and growth processes (Wei 1999). Generally empirical in nature, this work has added to the broader conceptual debate on China's approach and policy to national development and its regional dimensions (Fan 1995). As much of this work has centered on variations in regional growth, it has evolved to look at issues in local and regional development and has explored some of theoretical points that touch on the special nature of Chinese development. Geographers have been especially active in investigating south-east China and the Pearl River Delta region around Guangzhou and Hong Kong (Lin 1997). This work has moved beyond a focus on economic variables and begun to explore issues at the interface of regional and local politics, society, and economy. In this way, it has drawn from theoretical and conceptual ideas within other social sciences and even the humanities.

For example, research in recent years in other social sciences has examined the social/cultural notion of *guanxi* (Chinese for social relations or connections) and its role in commerce and politics. Much has been written on this topic and its more broad and related subject, what the sociologists Granovetter (1985) and Wank (1999) call the 'embeddedness' of private business in social networks and political power. What is this *guanxi* and how does it work? How does it interplay with commerce and politics and where did it originate? Can it be used and linked ideologically with Chinese and Asian thought to create a larger theoretical framework and context associated with Confucian and Asian values? Do these values and patterns vary in different parts of China, and is there a definable geography or spatial pattern of such matters? Related to this is the larger question of the role of the central state in development versus the role of local players and forces, and geography researchers are writing increasingly on the role of local forces and operatives.

Geographic study on this has largely been empirical, and some scholars have taken the difficult path of seeking to understand the commercial linkages between the PRC and Taiwan that transit the Taiwan Straits (Hsing 1998). On the surface such transactions would appear to be straightforward enough—the desire of Taiwan capitalists to minimize labor costs and thereby remain competitive in world markets for their manufactured goods that have a high labor content. Yet the relationships are complex, even among Chinese confrères who speak the same Fujian regional language. The political and commercial systems between Taiwan and mainland China have operated differently for a century if we include the previous period of Japanese colonial rule of Taiwan (1895–1945), and the subtle outward variation in behavior belies the complexity of culture and its interface with commerce and politics. Geographic research on shoe manufacturers, for example, who seek to bridge the gap between Taiwan and the PRC mainland reveals some of the difficulty and offers probing insight into the spatial complexity and significance of regional variation in what Wank (1999) has called the 'commodification of communism'. Such work confirms the murky and conflated relationships of Taiwan foreign investment in China and the linkages between business and politics in Taiwan as well as in neighboring Fujian on the mainland.

Similar work on the Pearl River Delta of the Hong Kong–Canton Region reveals the significance of a political-economy approach to understanding how intimate and intertwined are relationships among foreign investment, economic development and growth, technology and information transfer, and local and regional politics (Leung 1993). Seeking answers and understanding to such processes of economic, social, and political behavior requires a combination of conceptual approaches that may range from the local ethnographic and qualitative to the statistical and quantitative. The resourceful investigator will employ all that have utility and value in achieving better understanding of the particular processes of social, political, and economic change underway in the time and place of her/his study.

Geographic study of Japan and Korea has paralleled some of the same themes discussed for China while also exploring some topics that relate to economic trends associated with Japan's economic growth and global financial role. Geographers have written extensively about Japan's increasingly global economic ties, and much of this work has focused on the tactics and policies of Japan's foreign investment especially as directed toward North America, Europe, and Oceania. At the same time work has been done on the evolving industrial and corporate networks in Japan with reference to small businesses and the dynamics of local, regional, and national forces that drive location and growth. Such research originates in the past growth and emergence of Japan as a global player in the world's economy, and one whose decisions play a mighty role in determining world patterns of manufacturing. Thus, this work links events and trends in Japan to a larger world space economy, and its thrust comes from scholars who have a perspective that extends beyond a commitment to the analysis of Japan as a regional object of study.

Such research provides a more intellectually viable and catholic framework for contemporary area studies (Watters and McGee 1997).

Urban studies on Japan continue to be an important focus of study, and these include more traditional historic studies of cities and their economic and social processes as well as a number of recent studies of specific cities such as Tokyo and Yokohama (Cybriwsky 1998). Similar work is found in Korea. Seoul, as the nation's primate city, dominates as an object of study where the approach of urban biography has been used. This direction builds on traditional approaches in urban geography and continues an important qualitative dimension in urban related research. Knapp (1990) and Gaubatz (1996) have done extensive work in cultural geography that draws inspiration from the landscape and builds upon the intimacy of human activity and landscapes in China, although this is a minor theme in recent Japanese geographic studies. Related to this and historical geography is continuing work on the history of cartography of China, Japan, and Korea.

5. *Geography and East Asian Studies, a Continuing Successful Blend?*

Is geographic research and analysis on East Asia, as a branch of the social sciences, serving its purpose well, and providing creative and thoughtful answers and expositions on salient and significant issues, and matters that should benefit from spatial and geographic approaches to knowledge? In some specialized areas such as aspects of space economy, urban and migration studies, some specific cultural studies, and human–environmental studies, the answer is yes, and there are many good examples of high-quality work done in recent years in China, Japan, and Korea for evidence. On the other hand, social themes and gender studies are few, and geographers have largely ignored urgent political and security issues in East Asia. Are the current approaches, tools, and interests of contemporary geographers sufficient to meet the intellectual as well as policy challenges of contemporary East Asia as an increasingly important component of the world equation for the twenty-first century? Until geographers confront the key issues of geopolitics as well as political economy, it is likely their influence will be muted among social scientists and beyond in the broader arena of academic and policy oriented scholarship. At the same time, given the new tools of GIS and remote sensing that should enable a much more aggressive approach to environmental studies, geographers would do well to exploit that advantage and blend it with related studies in the social sciences to meet the many new challenges of the twenty-first century, as people everywhere cooperate and compete to make the fullest and best use of their environments and resources.

See also: China: Sociocultural Aspects; East Asia: Environmental Issues; East Asian Studies: Economics; East Asian Studies: Politics; East Asian Studies: Society; Fertility Transition: East Asia; Geography; Japan: Sociocultural Aspects; Korea: Sociocultural Aspects; Regional Geography

Bibliography

Chan K W 1994 *Cities with Invisible Walls: Reinterpreting Urbanization in post-1949 China*. Oxford University Press, Hong Kong and New York

Cybriwsky R 1998 *Tokyo: The Shogun's City at the Twenty-first Century*. Wiley, New York

Fan C C 1995 Of belts and ladders: State policy and uneven regional development in post-Mao China. *Annals of the Association of American Geographers* **85**: 421–49

Gaubatz P R 1996 *Beyond the Great Wall: Urban Form and Transformation on the Chinese Frontiers*. Stanford University Press, Stanford, CA

Granovetter M 1985 Economic action and social structure: The problem of embeddednes. *American Journal of Sociology* **91**: 481–510

Hsing Y 1998 *Making Capitalism in China: The Taiwan Connection*. Oxford University Press, New York

Knapp R G 1990 *The Chinese House: Craft, Symbol, and the Folk Tradition*. Oxford University Press, Hong Kong and New York

Leung C K 1993 Personal contacts, subcontracting linkages, and development in the Hong Kong–Zhujiang Delta region. *Annals of the Association of American Geographers* **83**: 272–302

Lin G C S 1997 *Red Capitalism in South China: Growth and Development of the Pearl River Delta*. University of British Columbia Press, Vancouver, BC

Smil V 1993 *China's Environmental Crisis: An Inquiry into the Limits of National Development*. Sharpe, Armonk, NY

Wank D L 1999 *Commodifying Communism; Business, Trust, and Politics in a Chinese City*. Cambridge University Press, Cambridge, UK

Watters R F, McGee T G (eds.) 1997 *Asia Pacific: New Geographies of the Pacific Rim*. University of British Columbia Press, Vancouver, BC

Wei Y H D 1999 Regional inequality in China. *Progress in Human Geography* **23**: 49–59

Yeung Y M (ed.) 1998 *Urban Development in Asia: Retrospect and Prospect*. Hong Kong Institute of Asia–Pacific Studies, The Chinese University of Hong Kong, Hong Kong

C. W. Pannell

East Asian Studies: Politics

'East Asia' in this article refers to the northeastern section of Asia, known also as 'Northeast Asia,' that embraced the five polities of China, Taiwan, North Korea, South Korea, and Japan in early 2000. For the sake of economy, this article discusses the study of politics in Taiwan as part of Chinese studies and that

of North Korea as part of Korean studies. 'Politics' is a category of human activities concerned with the determination and enforcement of rules governing the assignment of rights and duties to variously classified members of an organization, such as the state. The article discusses, on the basis of a survey of scholarly works published mainly in English, how specialists studied post-World War II politics in the abovementioned five countries.

1. *The Evolution of the Studies of Politics During the Cold War Period (1945–89)*

Several patterns and tendencies, four of which in particular stand out, have characterized the studies of politics in postwar East Asia. First, they have been led by landmark political events and developments, especially in Chinese and Korean studies. Second, American scholars and their scholarship have dominated the field. Third, the majority of the studies have been descriptive rather than theoretical. Moreover, even the few theoretical works have tended to place greater emphasis upon empirical accuracy than theoretical relevance. Fourth and last, studies of politics in one East Asian country have tended to be undertaken without much attention to studies of politics in another country, especially until the mid-1980s. These characteristics are particularly conspicuous in the works published during the Cold War period.

1.1 *The Evolution of Chinese Studies*

For a decade following the end of World War II, there were few social scientists anywhere in the world who were seriously interested in, much less knowledgeable about, modern Chinese politics. In Europe, scholarly interest in China was limited largely to a few British and Dutch philologists and historians, commonly known as 'sinologists.' Chinese politics was a serious concern only to a very small number of social scientists in Japan and the United States.

Japan did have a tradition of social science work on aspects of contemporary Chinese society, including surveys of rural society and economy dating back to the 1930s. The bulk of such work was, however, economic studies by economists in government and academia with only limited and indirect relevance to the study of contemporary Chinese politics. Even in the United States, the amount of scholarly work produced in the early postwar years was quite meager and owed mainly to sinologists and historians, such as John King Fairbank and Karl A. Wittfogel.

Interest in and research on Chinese politics grew rapidly in the subsequent years, propelled by a series of headline-catching events and developments: the unfolding Cold War (1947–89), the Cultural Revolution (1966–9), the emergence and evolution of the Reform regime (1978–present), and the Tiananmen Square incident (1989). These events, especially the Cold War and the Cultural Revolution, gave rise to as highly ideological, and often explicitly partisan, discussions of Chinese politics among social scientists as among policymakers.

In the United States, the Cold War inspired studies with ideological overtones of the doctrinal basis, organizational forms, and political and economic impacts of Chinese communism under Mao Zedong's leadership, often in comparison to Soviet communism. In Japan, ideological concerns divided nearly all China specialists into friends and enemies of the Maoist regime and its policies. In Europe, where the traditional sinological approach was stronger, and in Australia where there was no tradition of East Asian studies, ideological divisions among China specialists were neither as evident nor as serious.

The Cultural Revolution, however, introduced as sharp, or even sharper, ideological divisions among European, especially French, China specialists as among their American and Japanese counterparts. A good portion of American scholarship on the Revolution and related topics was, in fact, deliberately nonideological and analytical. Some works applied concepts associated with modernization theory or elite and faction politics models. These works helped to lay conceptual foundations for future studies of Chinese politics to build on.

The reforms undertaken by the Deng Xiaoping regime in the late 1970s and 1980s led to the 'opening' of China to investigations by foreign scholars, encouraging less ideological and more analytical studies of Chinese politics. This was true for works by European as well as American and Japanese scholars.

Changes in American scholarship were particularly notable. First, the range of topics chosen for investigation vastly expanded. Elite and faction politics continued to attract scholars' attention, and so did work on political culture in the tradition of modernization theory (e.g., Pye and Pye 1985). Such studies, however, were no longer the mainstream of American scholars' work on Chinese politics but were increasingly overshadowed by works on the Chinese bureaucracy and policy making (e.g., Lieberthal and Oksenberg 1988), local politics, rural industrialization, special economic zones, and the military.

Some scholars borrowed terms and concepts from works on Eastern European politics of the period, such as civil society, public sphere, and the state. This practice paved the way for more explicit and systematic comparisons of political and economic reforms in China and Eastern Europe to be undertaken in the 1990s. Also worth noting are the several revisionist studies of Chinese–United States relations in the early phases of the Cold War that relied on newly available archival and biographical sources.

Many of the changes in American scholars' work noted above were replicated in those of Japanese

scholars. Above all, the heavily ideological and partisan overtones that had characterized much of the work done in the heyday of the Cold War dissipated, as more and more Japanese social scientists began not only to visit China for fieldwork but also to collaborate with Chinese scholars.

Even more dramatic and significant was the change in the way Chinese intellectuals thought about and discussed Chinese politics. This was an integral part of the sweeping transformation of the intellectual environment of Chinese society, especially in Beijing and other large cities. Pro-reform, pro-democracy scholars, writers, and journalists set up clubs, salons, and seminars some of which began to publish journals and books, such as the *Journal of Dialectics of Nature* and *Towards the Future* book series, both edited by Jin Guantao and his colleagues (Gu 1999). The book series featured translations of Western social science works by, for example, Max Weber, Robert Merton, Kenneth Arrow, Cyril Black, and Alex Inkeles. The activities of these groups presaged the emergence in the 1990s of a new generation of Chinese students of Chinese politics who were mentored in Western-style social science theories and methodologies.

The process of the democratic transition in China, however, was abruptly interrupted in 1989 by the Tiananmen Square incident. The sensational event generated a spate of studies of the nature and problems of the democratic polity and civil society that had emerged under the Reform regime. In the United States alone, some two dozen books on the subject were published in the decade immediately following the event.

The glare on goings-on in mainland China had left 'the other China,' or Taiwan, nearly totally in the shadow throughout the Cold War period. The only notable exception was a modest amount of work on Taiwan's 'economic miracle.' It was not until the mid-1990s that serious scholarly work on the reform and democratization of Taiwan (which lagged behind mainland China studies by nearly a decade) began to appear in print in the United States and elsewhere.

1.2 The Evolution of Korean Studies

Having been under Japanese colonial rule since 1910, Korea had no viable native academic community at the end of World War II. Moreover, the peninsular state was immediately divided into antagonistic halves amidst the rising tide of the Cold War, each governed by a highly authoritarian government. While a Korean Political Science Association and a Korean Association of International Studies were founded in South Korea in the mid-1950s, it was mainly United States-based scholars who pioneered studies of South Korean politics until the early 1980s.

Most of these scholars focused on topics related to one of three broad subjects: the Korean War and South Korea–United States relations, the authoritarian regime and the democratization movement, and the rapid development of the South Korean economy. The first type of work is represented by Cumings' two-volume study (1981, 1990) of the origins of the war, the second type by Henderson's work (1968), and the third type by Amsden's (1989).

In the 1980s, studies of Korean politics by Korean scholars began to be published in South Korea. Most were historical and normative accounts of Japanese colonial rule and post-independence politics under military rule. The historical and normative emphasis in the studies of politics in the post-independence period reflected the frequent, and almost wanton, amendments of South Korea's constitution and electoral laws. Apparently for similar reasons, Marxist and other radical theories, such as dependency theory, were popular among Korean scholars.

Little was done on politics in North Korea, obviously due to the nearly total absence of reliable information on the subject. Suh's study (1988) of the lifelong leader, Kim Il Sung, was one of the few valiant attempts to understand and explain the politics of the 'hermit kingdom' of this period.

1.3 The Evolution of Japanese Studies

Unlike either China or Korea, Japan entered the post-World War world with a substantial community of social scientists with its origins dating back to the late nineteenth century. The Japanese Political Science Association (JPSA) was founded as early as in 1948. The best known among its early members turned their attention to the causes and consequences of prewar and wartime Japanese 'fascism.' Maruyama's work (1963) exemplifies this line of investigation.

By the mid-1950s, some Japanese political scientists began to study Japanese politics during the Allied Occupation (1945–52) and early post-Occupation periods. The subjects of these studies included the evolution of the Occupation authorities' policies, the Japanese government bureaucracy, parliamentary and local elections, public opinion, and social movements. The first major fruit of these efforts was a symposium published in 1958 under the title *Gendai nihon no seiji katei* (Political Processes in Contemporary Japan).

By comparison, studies of Japanese politics in the early postwar period by American and other foreign scholars were more narrowly focused on particular Occupation policies, such as the revision of the Meiji constitution, the purge of war criminals, and the land reform. By the early 1960s, however, the field of Japanese studies had come to be dominated by works based on and espousing modernization theory, as exemplified by the half-dozen edited volumes published by Princeton University Press in the mid-1960s to the early 1970s under the rubric of the theory. From this point onward, American scholars and their works inspired and guided studies of Japanese politics by Japanese scholars.

Research on Japanese voting behavior, for example, relied largely on theories and methods developed by American political scientists, notably by the so-called Michigan school. The Japan Election Studies Association, founded in 1981, inherited and perpetuated this tradition. The dominant influence of American scholarship was also evident in the study of the Japanese bureaucracy and policy making that gained great popularity following the publication of an edited volume on the subject in the United States in the late 1970s (Pempel 1977).

A major landmark in this line of work was the publication of Johnson's (1982) detailed study of the Japanese industrial and trade policy bureaucracy. This work triggered an international controversy on the nature and role of 'industrial policy' and the 'developmental state,' and inspired a series of case studies on related subjects. These included works on the relevant experiences of other East Asian countries, looked at deliberately and systematically in comparative perspective, thus giving rise to a new field of 'East Asian' studies (e.g., Haggard 1990, Wade 1990).

2. *Studies of East Asian Politics in the First Post-Cold War Decade*

The four tendencies or patterns that characterized studies of politics in East Asian countries during the Cold War period generally continued into the last decade of the twentieth century. The descriptive and interpretive emphasis, however, somewhat diminished and gradually began to yield to a more analytical and theoretical concern, consistent with a trend in American political science in general.

2.1 *The State of Chinese Studies*

There was a global shift in the study of Chinese politics in the post-Cold War decade from ideological and normative to empirical and theoretical interest and concerns. Foreign scholars now had and availed themselves of opportunities to visit and engage in fieldwork in China. Even Russian scholars began to study and write about aspects of Chinese domestic politics and foreign policy, although much of their work remained historical and philological rather than analytical and theoretical. Nonetheless, American scholars and their work not only continued to dominate the field but to do so more thoroughly than before.

American scholars' interest in elite and faction politics at the top of the Chinese political hierarchy held up during the 1990s (e.g., Baum 1994). So did their interest in Chinese-style bureaucratic politics that had emerged in the 1980s. The scope of their interest, however, substantially widened in the 1990s. Detailed and systematic investigations began to touch the politics of economic development and industrialization, the environmental crisis, and the military establishment and security policy.

Particularly noteworthy was the rapid growth of a theoretically informed empirical literature on regional and local politics and political economy. In addition to a number of detailed case studies undertaken by independent scholars, collaborative research projects jointly sponsored by Chinese and American organizations began to appear by the early 1990s. The Shandong Field Research Project of the Shandong Academy of Social Sciences and the Committee on Scholarly Communications with China is a good example. The University of Michigan–Peking University project is another.

The use of large-scale surveys to collect sets of data amenable to systematic statistical analysis was another new and significant development in Chinese studies of the 1990s. The Michigan–Peking joint project mentioned above, for example, involved such a survey and generated such a data set (Jennings 1997, Manion 1996). Moreover, Chinese scholars participated increasingly in the design, as well as the implementation, of a survey project and the analysis, as well as the collection, of the data.

More often than not, 'collaboration' between American and Chinese scholars was apprenticeship of the latter to the former in practice. It occurred in the context of a rapid increase in the number of Chinese scholars and students at American universities and in that of publications co-authored by 'senior non-Chinese and junior Chinese scholars' (Unger 1994, p. 181). By the end of the decade, however, some of these Chinese scholars began to engage in independent research and publish its results in sole-authored works, often in English.

Consistent with one of the four tendencies of East Asian studies in general noted earlier, the reversion of Hong Kong to China in 1997 spawned a spate of works on the event by both Chinese and foreign scholars. Most works predicted a gloomy future for the economy and, especially, politics of the island after the reversion, some based on elaborate theoretical models (e.g., de Mesquita et al. 1996). The reality, however, turned out to be more complex and mysterious than could be predicted by currently available social science models.

Studies of Taiwan politics in the 1990s were also characterized by increasing numbers of native scholars apprenticed to American scholars. Both the mentors and students tended to focus on the marketization of the island's economy and the democratization of its politics and publish increasingly sophisticated studies of the subjects.

2.2 *The State of Korean Studies*

American and other foreign scholars' interest in the Korean War, democratic reform, and economic development in South Korea continued well into the

1990s. In addition, North–South relations and the prospects of reunification became subjects of serious academic interest following Kim Il Sung's death in 1994 and, especially, the election of Kim Daejung as the South Korean president in 1997.

By far the most remarkable development in Korean studies in the 1990s was, however, the emergence of a substantial number of Korean specialists of Korean politics. Even more than its Chinese counterpart, this new generation of Korean political scientists was under the powerful influence of American political science. A 1995 survey of the Korean Association of International Studies' 1300 or so members revealed that nearly half had earned their doctoral degree at North American universities, as compared to about one-third who had done so at Korean universities. By the late 1990s, American-trained scholars had come to dominate Korean political science, including the field of Korean politics, not only numerically but also in setting research agendas and defining what was good research. The American influence was especially great in works with a strong theoretical and/or methodological bent.

North Korea remained as closed to and hidden from scholars' scrutiny even after Kim Jong Ill took over the country's political leadership as it had been under his father's iron-fisted rule. There was no genuinely academic research on its politics, or even its economy and society, inside North Korea itself. There was some outside, especially in Japan and the United States, but it remained, as the reviewer of an edited collection of research papers (Suh and Lee 1998) commented, a practice 'reminiscent of medieval scholasticism or ancient Chinese oracle-bone divining' (C. K. Armstrong 1999 in *The Journal of Asian Studies* **58**: 224). North Korean politics in the last decade of the twentieth century thus remained as inaccessible as before to scholars either inside or outside the country.

2.3 The State of Japanese Studies

Research on the Japanese national bureaucracy and policy making that began to attract concentrated academic attention in the 1980s continued to expand in the 1990s. The objects of detailed investigations were no longer limited to industrial policy, as in Johnson's (1982) study, but expanded to include many others, such as fiscal, construction, telecommunications, health, welfare, science and technology, environment, and defense policies. Attention expanded also to include, in addition to the role of the central and local government, those of labor, the mass media, and ethnic minorities.

Moreover, work by Japanese and foreign scholars became increasingly theory-conscious and methodologically sophisticated. Nearly all such work, however, relied on theories, models, or methods developed by American scholars, such as rational choice, new institutionalism, agency, and path dependence (see *Political Science: Overview*). Among those studying Japanese electoral politics and voting behavior, for example, the use of not only the least squares and maximum likelihood methods but multinominal probit and logit models became increasingly commonplace.

Some American comparative politics scholars with expertise in the field of American or European politics undertook research and published results on one aspect of Japanese politics or another (e.g., Katzenstein 1996). Theoretical and methodological Americanization was thus well advanced even among Japanese political scientists, the overwhelming majority of whom, unlike their South Korean counterparts, not only did not hold a doctoral degree from an American university but, in fact, did not hold one from any university.

2.4 The State of Comparative Studies of East Asian Politics

During the 1990s, the economic and political reforms and transitions in China began to be systematically compared to those in Eastern Europe. Hungary in particular was chosen as China's European counterpart. Some studies attempted also to compare transition processes in Eastern Europe on the one hand and East and Southeast Asia on the other. Yet others compared reforms in mainland China with those in Taiwan and Russia or those in Taiwan and South Korea. One study compared the relationship between mainland China and Taiwan with that between North and South Korea.

Aspects of Japanese politics were often compared to those of American politics. For example, one edited book compared policy processes in the two countries from a rational choice theoretical perspective, while another compared the parallel processes of globalization and decentralization in them. More commonly, however, Japan was subject to multilateral, rather than bilateral, comparisons, especially with other advanced industrial countries. For example, the causes, processes, and consequences of market liberalization and deregulation in Japan were compared in one study with those in the United States and Western European countries and, in another, with those in France, Spain, South Korea, and Mexico. Likewise, Japanese environmental policy, labor politics, and judicial system were compared with those in North America and Western Europe.

Studies of international relations in East Asia became theory- and methodology-conscious more slowly than comparative studies of domestic politics. By the late 1990s, however, there had appeared some work that applied a sociological evolutionary model to Chinese foreign policy, a model of linked international–domestic negotiations, known as two-level

games, to Japanese–United States trade negotiations, network analysis to Japanese relations with its Asian neighbors, and so on. Some work on security relations among East Asian and, more broadly, among Pacific Rim nations used the international regime model that had gained considerable popularity among American international relations scholars in the mid-1980s.

At the turn of the millennium, the impacts of unexpected events on the study of politics in East Asia did not appear to have diminished or to do so in the foreseeable future. Nor did the pace-setting role of American scholars, especially in theoretical and methodological innovations. On the other hand, the pull of the empirical in studies concerned with the development of empirical theory of politics did seem to have diminished somewhat, while the pull of the theoretical had become somewhat stronger, in large measure as a result of the powerful influence of American political science. There was also growing interest in comparative studies of selected aspects of politics in two or more East Asian countries or those in them and in other areas of the world. These features of the field are likely to persist into the first decade of the new millennium.

Bibliography

Amsden A H 1989 *Asia's Next Giant: South Korea and Late Industrialization*. Oxford University Press, New York

Baum R 1994 *Burying Mao: Chinese Politics in the Age of Deng Xiaoping*. Princeton University Press, Princeton, NJ

Cumings B 1981, 1990 *The Origins of the Korean War* (2 vols). Princeton University Press, Princeton, NJ

de Mesquita B B, Newman D, Rabushka A 1996 *Red Flag Over Hong Kong*. Chatham House Publishers, Chatham, NJ

Gu E X 1999 Cultural intellectuals and the politics of the cultural public space in communist China (1979–1989): A case study of three intellectual groups. *Journal of Asian Studies* **58**: 389–431

Haggard S 1990 *Pathways from the Periphery: The Politics of Growth in the Newly Industrializing Countries*. Cornell University Press, Ithaca, NY

Henderson G 1968 *Korea: The Politics of the Vortex*. Harvard University Press, Cambridge, MA

Jacob H et al. 1996 *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT

Jennings M K 1997 Political participation in the Chinese countryside. *American Political Science Review* **91**: 361–72

Johnson C 1982 *MITI and the Japanese Miracle: The Growth of Industrial Policy, 1925–1975*. Stanford University Press, Stanford, CA

Katzenstein P J 1996 *Cultural Norms and National Security: Police and Military in Postwar Japan*. Cornell University Press, Ithaca, NY

Lieberthal K, Oksenberg M 1988 *Policy Making in China: Leaders, Structures, and Processes*. Princeton University Press, Princeton, NJ

Manion M 1996 The electoral connection in the Chinese countryside. *American Political Science Review* **90**: 736–48

Maruyama M 1963 *Thought and Behavior in Modern Japanese Politics*. Oxford University Press, London

Pempel T J (ed.) 1977 *Policymaking in Contemporary Japan*. Cornell University Press, Ithaca, NY

Pye L W, Pye M W 1985 *Asian Power and Politics: The Cultural Dimensions of Authority*. Belknap, Cambridge, MA

Suh D-S 1988 *Kim Il Sung: The North Korean Leader*. Columbia University Press, New York

Suh D-S, Lee C-J (eds.) 1998 *North Korea after Kim Il Sung*. Lynne Rienner Publishers, Boulder, CO

Unger J 1994 Recent trends in modern China studies in the English-language world: An editor's perspective. *Asian Research Trends: A Humanities and Social Science Review* **4**: 179–86

Wade R 1990 *Governing the Market: Economic Theory and The Role of Government in East Asian Industrialization*. Princeton University Press, Princeton, NJ

H. Fukui

East Asian Studies: Society

For most of the past 1,500 years East Asia has consisted of three principal societies—China, Japan, and Korea. Each developed with little impact from outside until the 1840s, except frontier conquerors such as the Mongols and Manchus, who remade the state somewhat but the society very little. Japan and Korea borrowed from China as their model, but interaction was irregular. Distinct domestic forces, including Japanese community orientation and samurai loyalties, evolved independently. With strong state traditions and complex organizations, East Asian societies retained their independence and national identities despite powerful influences of the more developed West in modern times. Except for the interval in the first half of the twentieth century, when Japan imposed colonialism and the 'East Asian Co-Prosperity Sphere' on the rest of East Asia, it was not until globalization in the 1990s that economic, political, and even cultural barriers fell sharply in this region. After Japan reformed under the US Occupation and China borrowed from the Soviet socialist model, both societies turned away from their mentors and even basked in claims of uniqueness. Although such claims hid the universal effects of modernization, state-led efforts to sustain idealized social practices lasted a long time and, to a degree, continue. The past still weighs heavily amidst new commitments to internationalize to meet global competition.

The study of societies in East Asia crosses disciplinary boundaries. Given the importance of state policies, it is not surprising that political scientists often play a large role, treating values through political culture, social change through campaigns and policy initiatives, and civil society through state controls. Historians are notable too; after all, links between past and present figure in many explanations of social

problems and successes. Anthropologists are in the foreground in the many analyses of family and community, often central even in discussions of modernization. While sociologists until recently have counted as but one part of a large mix of researchers, they have gained prominence at the extremes: articulating the models of development that have worked in the region and exposing the serious social problems that threaten further development. Their research, along with historians in the 1960s–1970s, highlighted preconditions for modernization, along with political scientists in the 1980s and 1990s, analyzed social movements and values, and together with anthropologists of late has focused on the social exchange and networks that allow informal ties to prevail over formal institutions.

Transformation in East Asia rarely was imposed from outside or reflected domestic class struggle. Administrative continuity under Confucian principles allowed top-down change to dominate. Over time, three transformative forces began at the top and were redirected from below. Bureaucratization occurred as at first itinerant officials such as Confucius in the sixth century BC offered their services and later powerful emperors or shoguns built multilayered staffs. This process drove the spread of male literacy and examinations in China and Korea, as public service became the source of honor and often wealth. It also transformed feudalism in Japan as vassals learned to serve their lords in administrative tasks. Yet, leaders relied on moral suasion and community self-administration, keeping the size of the bureaucracy down.

Urbanization spread, beginning with Chinese walled cities as outposts of administration. The role of these cities changed as they gradually reorganized within regional networks of cities linked by numerous rural periodic markets and legally unfettered migration. Controls in Japan were tighter, but when *daimyo* accelerated the concentration of samurai in castle towns as warring states gave way to economic competitors, Japan's urbanization after 1600 far overtook that of China. Eased by coastal transportation, a national market ensued. While intra-regional ties as well as global ties were curtailed in East Asia, strong state support and growing commercialization made Edo (Tokyo) and Beijing the world's largest cities in the eighteenth century (Rozman 1974).

Confucianism also was embraced at the top and then spread widely through education and family-based rituals, guiding mass as well as elite social practices in scattered communities. There was no religious hierarchy. Although rice cultivation required village cooperation, the family was relatively free to pursue opportunities in landownership, commerce, or education except for limitations in Japan on social class mobility. In China, the population climbed to 400 million or nearly a third of the world's total in 1850. In Japan, with about one-tenth as many people, decentralized rule through more than 250 domains spurred regional competition, allowing some domains to move ahead in development. Four led in seizing power soon after the US forced open the country and in centralizing Japan under the emperor through massive reforms that resulted in rapid modernization. A social transformation came from part of the old elite inspired by a new nationalism. Studying which of the leading industrial countries excelled in each area, they borrowed widely, abolishing closed social classes, pursuing universal education, and building a national bureaucracy based on a modern university system with rigorous entrance examinations. Free of the rigid imperial Confucianism in China, Japan could draw on tradition flexibly to meet the challenge of the West (Rozman 1991).

As precocious as East Asian societies had been—with China in the forefront of global development from 500–1500 AD, and Japan and Korea gaining in dynamism—they lacked some forces that enabled the rise of the West (Rozman 1981). In long-lasting dynasties, each was ruled for continuity and stability with little impetus from foreign trade or thought, religious minorities, or intellects that questioned Confucianism. Once obliged to change, however, Japan and later China and Korea could tap some promising traditions. Bureaucracies of the educated elite were useful in latecomers to modernization catching up via reforms from the top. Urban and commercial traditions nurtured an entrepreneurial spirit. Family initiative and paternalistic organization reached across the population despite traditions of deference to the state, respect for elders, and subservience of women that limited the scope of change. In Japan, habits of loyalty to lords proved convenient when new leaders pursued reform. In China it took longer to activate and unite challengers to authority.

Japan's transition was fastest, relying on more intense bureaucratization, higher rates of urbanization, and a combination of Confucianism and the samurai code of obedience and honor (Black et al. 1975). While Japanese society experienced most of the changes usually associated with development, there were important differences (Dore 1967). Administrative guidance at crucial times accelerated modernization and left society more vertically organized and dependent. Recitation of the imperial rescript on education inculcated a sense of duty rather than a questioning civil society. The decision in 1898 to model the civil code on samurai family ideals reinforced a patriarchal family intended to restrain deviance. The state took a harsh stand in the face of social problems such as tenant disputes and urban riots, stressing an ideology of harmony and reliance on go-betweens in conflict resolution. It turned to repression to counter the popularity of socialist ideas. As nationalism fueled wars and colonialism, the state asserted new powers, for instance, organizing ward associations for surveillance and mobilization. Loyal conscripts and diligent workers made success possible

in building a centralized, industrial power able to dominate Korea and move ever more aggressively against China. Yet, an emerging middle class was ready for reform.

As the imperial center in Beijing's Forbidden City retarded reform, Chinese society failed to modernize. It took much longer for alternative power centers to emerge, and they were left divided by warlordism across a vast country. After imperial examinations were ended in 1905 literacy rates stagnated despite the appearance of modern schools. Treaty ports and the sudden rise of Shanghai as the leading city gave foreigners high visibility, but failed to spur urbanization. Criticisms of Confucianism as a cause of backwardness reinforced attacks against female foot-binding, elite cultural aloofness, and growing inequality in the villages. Japanese aggression plunged the country into war, exacerbating social problems. Accustomed to regarding their country as the center of the world, Chinese searched for a new source of pride. Communists rode the coat-tails of nationalism and land reform to power.

The period of 1945–55 shifted the direction of social change in each of the East Asian societies. Defeat exposed the falsity in wartime rhetoric to Japanese and brought support for many of the reforms imposed by the US Occupation. Policies to create a democratic society included land reform, coeducation, increased equal rights in the family, decentralization, and freedoms for the society in dealing with the state. Although Japanese ways of implementation soon limited the impact of some of these changes and elements of wartime centralization led in the recovery, the results were far-reaching and favorable for social harmony and national development. China's communists were far more radical, aggressively targeting reputed class enemies in land reform and urban reorganization. They stifled traditional society and sprouts of capitalism as the reach of the state extended downward. Yet, the enormous scale of the society and the limited modern resources available to the state forced compromises. At first, these were promises of gradual application of the Soviet model such as collectivization. Later, even after many excesses of radicalism, they were impatient calls for more extreme changes to compensate for what was missing.

Postwar development started as top down in all of the East Asian states except Hong Kong under British administration. Taiwan started with tension between the newly arrived mainlanders and the local residents who had fared better than Koreans under Japanese rule. Yet, reforms gave scope to a dynamic society as entrepreneurs expanded family enterprises and US-educated professionals brought high-tech exports to the fore. In South Korea a strong state at first stymied the business community, but then colluded with a small number of *chaebols* to promote rapid growth. Of course, the top-down approach went furthest under communism: in China, where the Soviet model was applied in the 1950s, and in North Korea, where more extreme social mobilization and isolation remained unchanged.

The Japanese postwar model of modernization inspired the rest of the region to double-digit growth. Three types of 'Confucian capitalists' rose to the fore. First, were the elite officials chosen after examination competition to the best universities and working long hours while their wives devoted themselves to a supportive home and education of the next generation. Second, were the salarymen whose entry into large corporations rested on a similar background. Companies fostered lifetime employment through seniority wages, small work groups that socialized together, and inculcation of identity to company rather than to profession. Third, were small family enterprises, which even in Japan employed a disproportionate number of people. This model succeeded as long as young people gave up their adolescence for rote memorization, women accepted brief working careers as office girls before marriage, and salarymen gave priority to company over family and leisure (Rohlen 1989, Brinton 1993). It benefited from widespread belief in the fairness of exams for social mobility and in the existence of a middle class seen as embracing 90 percent of Japanese.

As rates of growth slowed and productivity became increasingly dependent on creativity, Japanese lost confidence in their model. Young people became engrossed in pop culture, the burden of one of the world's most rapidly aging populations raised new demands for social welfare, women sought careers, and men preferred leisure. Society was converging with the West. The Japanese media touted the urgency of a third wave of reforms comparable to those of the Meiji era and the Occupation. Yet, vested interests blocked decentralization as well as other means to far-reaching reform. 'Internationalization' remains an elusive ideal for people used to all-encompassing in-groups rather than individualism (Sugimoto 1997).

Under communist rule, life in China was jolted by a quarter of a century of social experimentation. Starting with land reform, all citizens were assigned class labels that determined their opportunities in life and left them vulnerable to periodic campaigns and collective punishment. In the Great Leap Forward, people's communes for a time even threatened family life with collective day care and mess halls. Extreme forms of thought control relying on Mao's little red book and on sending officials and intellectuals to the countryside for re-education accompanied the Cultural Revolution. The result was a dossier society, where criticism and self-criticism in small groups politicized daily life. Yet, despite forced relocation of millions of teenagers from the cities and the havoc in education from attacks on experts, conditions of life and work were mostly frozen. Workers were assigned an organization claiming broad control over all aspects of life with little hope for mobility (Walder 1986). Peasants could not

leave the commune after urban–rural differences became fixed. Some indicators such as life expectancy and literacy improved rapidly, but urbanization stalled. Ideological claims to be building a pure socialist society in contrast to Western capitalism and Soviet revisionism were based on concealment of social problems and arbitrary authority. When Red Guards were unleashed to beat their teachers and denounce their parents, it appeared that campaigns against Confucianism were in fact reversing Chinese traditions. In fact, most people coped with pressures from above by turning to family for support. Without much modernization of life or, in rural areas, support from the state, Chinese retained many traditional attitudes (Parish and Whyte 1978).

From 1978, the Chinese mainland turned to a different model of development already successful in Taiwan, Hong Kong, and Singapore (Orru et al. 1997). To some extent, this resulted from the pragmatism of leaders who had suffered through Mao Zedong's radicalism and were now bent on winning popular support and catching up. Even more, it was a tribute to the resilience of Chinese society. As commune controls were relaxed, villagers reinvigorated the family farm, then seized opportunities to develop township and village enterprises, and finally flocked to the cities as the 'floating population' ready for any onerous job even if the state forced them to accept a temporary status without ordinary benefits. The state favored coastal areas, initially in the Southeast, with special economic zones (Solinger 1999). Not only did inequalities across China widen abruptly, many social groups felt unfairly treated. Personal connections, including bribery, determined life chances for many. Eager to sustain the monopoly of the communist party and hostile to a free press and independent judiciary, China's leadership failed to address problems of corruption. As society became increasing materialistic, it also lost more of its spiritual moorings. While some flaunted their new wealth, others searched for new values in the West or in religion. The state had some success in countering this with nationalism, portraying China as long humiliated by the West and newly contained by US hegemonism.

After social class labels were dropped and millions were rehabilitated, Chinese regained many freedoms despite some significant limitations. With the end of rationing and easier movement, urbanization doubled. Authorities sought to channel part of the influx into small cities or townships. With the return of competitive examinations, a meritocracy re-emerged and bureaucracies became more technocratic; yet power in some localities rested with party leaders who behaved as petty tyrants. Barriers to markets were dropped, and entrepreneurship flourished. Workers could now move from one job to another, but the state continued to provide loans to state-owned enterprises deeply in debt in order to limit layoffs and demonstrations by workers and pensioners whose housing and other benefits depended on these firms. The state sought to channel interest groups, authorizing professional associations and trade unions subject to its influence. Scared by the large demonstrations led by students in 1989 and by the example of the Soviet Union, China's communist party allowed more safety valves in popular culture and materialism while putting the lid on a civil society. Ethnic minorities, which were concentrated mostly in the west and south, were not much assimilated. Outnumbered ten to one by Han Chinese, they were no longer subject to the decimation of the Cultural Revolution and allowed to revive many customs as long as no political organization posed a challenge. Contradictions mounted between rising middle class aspirations in coastal urban areas and local corporatism in which networks linked to officials protected business arrangements.

Hong Kong, Taiwan, and Korea influenced China not only by their own ascent as middle-class societies, but also by their transfer of business practices through direct investment to nearby regions of the mainland. Foreign firms spread new management practices—a sharp shift from the inefficient socialist egalitarianism and clientelism in the factory. For many young migrants, exploitation in the form of long hours and few benefits took its toll quickly, but an endless supply of surplus labor from the countryside was ready to replace them. Multinational corporations spread high-tech business as well, drawing many urban professionals into the information age. After television had spread rapidly and popular culture was bringing Taiwan and Hong Kong together with the mainland in Greater China, the Internet became the new force of internationalization. Yet, in a population nearing 1.3 billion in the year 2000, over 80 percent remained well removed from a modern lifestyle. For most of rural and inland China, state-controlled television provided the main source of internationalization. A wide gulf still separated urban and rural. Despite stringent family planning since the 1970s when the state imposed a one-child policy, later relaxed to permit two children in some rural families, the workforce continued to grow so fast that only a high economic growth rate could meet the need for jobs. A return to household and community control of labor exposed the inferior status of women. So, too, did the struggle to have a son in the countryside. While all over East Asia aging populations and low birth rates were threatening to impose a great cost in social welfare, the costs in China would come later but with even less of an economic cushion from national pension and healthcare systems.

In the year 2000, contrasts abound in East Asia. In North Korea years of famine and total isolation from the outside leave a totalitarian society in desperation. In the Pacific coast megalopolis dominated by Tokyo high incomes bring many of the problems of affluence. Except for Koreans brought to Japan before 1945 and ex-patriots in Hong Kong after it was transferred from British rule, few foreigners live in any of the East Asian

societies. While guest workers on a small and regulated scale have begun to come to Japan, Taiwan, and South Korea to fill the dirtiest jobs, they are outsiders. Disparities slow regionalism as well as globalization. Calls for a common identity based on Confucian heritage or Asian values confront distrust left by Japan's insensitive handling of its legacy of war crimes and China's nationalist pressure to regain Taiwan and expand its power without assurances of democratic evolution.

Recent research explores the impact of accelerated globalization on East Asian societies. On Japan and Korea, domestic and foreign experts warn that a new model of development is required. They show limitations in educational creativity, bottom-up entrepreneurship, and competition free of administrative guidance. On China, foreign specialists are split between those who recognize the unexpected adaptability of a communist-led system still able to achieve high growth rates after the Asian financial crisis and those who point to mounting frustration among farmers and workers in an era of unbridled exploitation over much of the country. Both sides have come to recognize a model of family-based enterprise through 'Greater China' challenged by both the larger-scale and more transparent operations of a new era and the lingering fetters on initiative left in West China and other areas bypassed by the recent dynamism.

On the one hand, borderlessness should bring high-tech urban agglomerations such as Beijing through Seoul to Tokyo or Taipei-Hong Kong-Guangzhou. On the other hand, civil societies remain weak, and state power combines with deep historical scars to make nationalism strong. New-age materialists are succeeding older generations socialized to forego self-gratification in times of far less modernization. As dating and divorce spread, single-child families are challenging traditional solidarity. Organizational paternalism with lifetime employment is fading. Still trying to guide social change, states face a powerful combination from bottom-up social forces inspired by global information and linked to the global economy. While many Confucian elements endure in social life, the challenges to top-down control are greater than ever. The need for deeper research on what occurs behind formal institutional facades is also greater than ever. While scholars have extended empirical studies and are armed with more theoretical models, there is still a dearth of research that reveals the hidden mechanisms in societies still not known for their transparency.

See also: China: Sociocultural Aspects; East Asia, Religions of; East Asian Studies: Culture; East Asian Studies: Economics; East Asian Studies: Gender; East Asian Studies: Geography; East Asian Studies: Politics; Japan: Sociocultural Aspects; Korea: Sociocultural Aspects

Bibliography

Black C E, Jansen M B, Levine H S, Levy M J, Rousky H, Rozman G, Smith H D, Starr F S 1975 *The Modernization of Japan and Russia*. The Free Press, New York

Brinton M 1993 *Women and the Economic Miracle: Gender and Work in Postwar Japan*. University of California Press, Berkley, CA

Dore R (ed.) 1967 *Aspects of Social Change in Modern Japan*. Princeton University Press, Princeton, NJ

Orru M, Woolsey Biggart N, Hamilton G G 1997 *The Economic Organization of East Asian Capitalism*. Sage Publications, Thousand Oaks, CA

Parish W L, Whyte M K 1978 *Village and Family in Contemporary China*. University of Chicago Press, Chicago

Rohlen T P 1989 Order in Japanese society: attachment, authority, and routine. *Journal of Japanese Studies* **15**: 5–26

Rozman G 1974 *Urban Networks in Ch'ing China and Tokugawa Japan*. Princeton University Press, Princeton, NJ

Rozman G (ed.) 1981 *The Modernization of China*. The Free Press, New York

Rozman G (ed.) 1991 *The East Asian Region: Confucian Heritage and Its Modern Adaptation*. Princeton University Press, Princeton, NJ

Solinger D 1999 *Contesting Citizenship in Urban China: Peasant Migrants, the State, and the Logic of the Market*. University of California Press, Berkeley, CA

Sugimoto Y 1997 *An Introduction to Japanese Society*. Cambridge University Press, Cambridge, UK

Walder A 1986 *Communist Neo-Traditionalism: Work and Authority in Chinese Industry*. University of California Press, Berkeley, CA

Whyte M K, Parish W L 1984 *Urban Life in Contemporary China*. University of Chicago Press, Chicago

G. Rozman

Eastern European Studies: Culture

Culture can be treated as a distinct object of study. For instance, literature and art are usually understood as expressions of 'high culture,' while the economy is typically studied as something apart from culture. Culture also is used to differentiate groups. Nations are not only political categories, but they are distinguished by language, ways of life and interpretations of the past. Some argue that to explain social life at all, one must understand meanings in everyday life, or the culture of a place and time. Finally, culture is fundamental in social inquiry itself. It is embedded in the constitution of the objects of our inquiry (Calhoun 1995, pp. 61–6). All of these meanings of culture inform this entry, but the final sense of culture shapes its overall architecture.

The notion of 'Eastern European Studies' should be understood as a regional theme in the more general enterprise called 'area studies,' that interdisciplinary

study of a region's languages, cultures, histories, and institutions. This particular form of scholarship achieved great prominence in the USA during the Cold War, during which a new linkage between Western geopolitics and scholarship was cultivated. The American government sought expertise about other parts of the world in order to aid its struggle against communism's spread. Consequently, it subsidized research and training in area studies across the world's regions.

Combining Western geopolitical aims with scholarly work was relatively more compatible in studies of Eastern Europe and the Soviet Union than of other world regions. During and after World War II, the Soviet Union expanded its borders to include Estonia, Latvia, Lithuania, Western Ukraine, and Belarus, while communists took power in other parts of Eastern Europe—East Germany, Poland, Hungary, Czechoslovakia, Romania, Bulgaria, Yugoslavia, and Albania. Many political émigrés from these countries could be found among the West's experts in Eastern European Studies. Despite the political associations of exile and state sponsorship, Eastern European Studies enjoyed relative autonomy and scholarly growth through its location within various disciplines. After communism's collapse throughout the region, Eastern European Studies no longer faces such a direct and obvious political implication. It does, however, face a challenge of coherence and other problems associated with culture's place in disciplinary practices.

1. *Eastern Europe as Cultural Category*

Eastern European Studies has its antecedents in various forms of civilizational study, most notably 'Orientalism.' Like Orientalism, area studies expertise tends to be organized through networks of knowledge production grounded outside the region. While the Cold War inspired American area studies projects, an East European focus could build on an earlier West European invention. During the Enlightenment, West Europeans understood this region, then including Russia, as neither Orient nor Western, a place filled with contradictions whose coherence a 'European' gaze could not recognize (Wolff 1994, Gal 1991).

This combination of Western proximity and 'backwardness' has shaped much of Eastern European Studies. One enduring debate asks whether Eastern Europe's indigenous social relations or its trade relations explain its economic underdevelopment over the last 400 years (Chirot 1989). Despite its economic trajectory, East European art, music, architecture, and literature have thrived, leading some to identify modernity's disarticulation as a regional distinction (Wandycz 1984). These accomplishments lead others to argue that Europe itself should be understood as much from its east and especially its center as from its west (Sayer 1998).

Eastern European Studies has not only been cultivated in the West. For instance, the Soviet Union had its own complement of area studies. Eastern Europe was particularly important in this array given the role post-World War II Eastern Europe played, until *perestroika*, as a *cordon sanitaire*. Within Eastern Europe, however, the region's area study is not well developed; the region is rarely articulated as a category of cultural imagination. East European diversity has invited much finer breakdowns.

Most significantly, the region's imperial heritage is diverse—with the Ottoman, Habsburg, Russian, and Prussian empires shaping the region's long nineteenth century, with enduring effect. The 'Balkans,' territories of the former Ottoman Empire in Southeastern Europe, have remained an important regional category throughout the twentieth century; with war in the 1990s, the term gained new cultural power (Todorova 1997). Those from the Western-most parts of the region also took the area's diverse imperial legacy as a starting point for their contemporary critique of Soviet domination, arguing that communism ill suited East Central Europe, with its more developed civil societies (Szücs 1988). Some resist Cold War categories altogether and argue that East Central Europe's boundaries should move into Soviet space, to include the lands of the fifteenth- and sixteenth-century Polish-Lithuanian Commonwealth (Magosci 1993).

There is no simple cumulation of scholarly competencies in Eastern European Studies. There has been no regionally hegemonic language, like Russian was in the Soviet Union. Depending on the boundaries one invokes, and the time one imagines, different languages have been more and less dominant and numerous. For example, in 1900, Germans, Poles, and Ukrainians were the largest ethno-linguistic groups in Magosci's (1993, pp. 97–8) East Central Europe, but Romanians, Hungarians, Czechs, Belorussians, Serbs, Greeks, Bulgarians, Croats, Slovaks, Lithuanians, Slovenes, Albanians, Macedonians, Bosnian Muslims, and Montenegrins were all dominant in particular places across the region. Beyond these groups who then had, or who would eventually have, East European states in their name, there were significant numbers of Jews, Turks, Gypsies/Rom, Italians, Carpatho-Rusyns, Szekelys, Friulians, Armenians, and other ethno-linguistic groups.

There also is no principal religious tradition in Eastern Europe, whose status Islam occupies in the definition of Middle Eastern and North African studies. Islamic and Orthodox, Roman Catholic, and Protestant Christian religions dominate various parts of the regional landscape. Judaism was demographically and culturally significant before the Holocaust, and even after genocide, Judaism remains critical to understanding the region's cultural politics. Thus, while Eastern Europe is itself a profoundly important

cultural category, it is not based on the region's cultural homogeneity. Eastern European Studies is, however, profoundly shaped by the culture of nations and nationalism.

2. *The Significance of the Nation*

Cultural analysis in Eastern European Studies draws upon an indigenous scholarship, but one organized through national fields rather than across the region. For instance, Polish sociologies of the intelligentsia, both before and after World War II, had the nation as their field of reference (Chałasinski 1997). Hungarian ethnographies of the countryside, while foundational for Western anthropology's study of Transylvania and elsewhere, were focused on the making of the Hungarian nation (Hofer 1968). Musicology or literary criticism also has been organized within national fields, around leading figures like Béla Bartók or Witold Gombrowicz, whose complicated expressions of their nation through music and literature form leitmotifs of their cultural interpretation. Consequently, the nation is one of the region's most important cultural formations. It shapes other cultural studies and is an object of study itself.

Studies of the nation are themselves framed within larger cultural formations. Many studies of the nation in Eastern Europe are set up within a narrative of civilizational distinctions. For instance, national historiographies tend to depict the Ottoman Empire's expansion through the 1600s and its retreat from the Balkans during the nineteenth and beginning of the twentieth century in terms of the contest between Islam and Christianity, on which more recent stories of civilizational clash draw. Romanian national historiographies identify their civilizational alternatives with Western and more autochthonous (indigenous, native) traditions (Verdery 1991). Estonian social science around ethnicity portrays a 'civilizational' difference between the West and Russia (Kirch 1997).

In contrast to those who demonstrate enduring cultural oppositions, others have emphasized the process of establishing national identity and difference, asking, for example, how intellectuals articulate nations (Suny and Kennedy 1999). Those who focus on 'historic' nations (Walicki 1982) and nations that had to '(re) awaken' (Rudnycky 1987) have different emphases. The latter tend to ask how peoples come to recognize themselves as nations and how they mobilize to realize political ambitions including the formation of nation-states out of imperial states (Hroch 1985). Those who focus on historical nations are more influenced by narratives of loss, whether of sovereignty, territory, or tolerance. For instance, how should one theorize the condition of Jews in a nation-state rather than an empire (Hertz 1988)? How did nationalism learn to hate (Porter 2000)? How do diasporas beyond the homeland's state boundaries figure in the nation's membership (Brubaker 1996)?

Nationalism and its study are not limited to Eastern Europe, but its significance in Eastern European Studies has been elevated by the region's distinctive experience with communist rule, which in turn has come to define the region as a whole, and to organize most of its cultural social science.

3. *Culture in the Social Sciences*

Czesław Miłosz's *The Captive Mind* (1951) is one of the best-known works detailing how communist rule shaped high culture. His portraits were based on particular Polish writers' responses to Stalinism, but they also could be understood as ideal typical portraits of authorial responses to communist rule. Jeffery Goldfarb (1981) developed this concern for cultural freedom, identifying communism's constraining and enabling properties for artistic expression and audience appreciation in Polish theatre and other cultural productions. While political censorship certainly constrained cultural expression, it also made artistic expression broadly valued and independent of culture's commodification.

Many treatments of the intelligentsia were as concerned with power and privilege as cultural expression *per se*. In one of the most influential and controversial works assessing the intelligentsia, György Konrád and Iván Szelényi (1979) argued that communist rule's maturation allowed this class to shape the conditions of distribution in its own favor. Underwritten by a principle of teleological knowledge that privileged their own system of 'rational redistribution,' the intelligentsia was poised to come to power. With the aid of monetarist ideology, Eyal et al. (1998) argued this finally came to pass in communism's collapse. This formulation of the intelligentsia has also been helpful in understanding how communism's institutional politics of redistribution escalated nationalist politics among Romania's cultural producers (Verdery 1991).

Studies of class, stratification, and inequality under communist rule were mostly unconcerned with culture, but Stanisław Ossowski's (1963) emphasis on ideology in inequality's study opened the way for 'subjective' assessments of class and stratification. A number of Polish researchers asked about the popularity of gradational, functional, and dichotomous visions of inequality in the social consciousness (Zaborowski 1986). Class consciousness was complicated by the authorities' appropriation of that discourse, but other sensibilities of work and labor could inform peasant and worker identities (Lampland 1995). Perceptions of 'us' and 'them,' based on political distinctions were found to be quite widespread through survey research, and became especially powerful in Poland after the formation of the Solidarity movement in 1980 (Adamski 1993).

A substantial body of scholarship has been developed to address each period of communism's systemic crisis—1956, 1968, 1980, 1989 (e.g., Ekiert 1997). Rock music has been analyzed as a contributor to these transformations of communism, but the explicitly cultural analysis of political crisis is not so common. Culture rather performed as the obvious background, most notably in terms of a nation's fit with communist rule imposed by the Soviet Union. The Hungarian Revolution of 1956, the Prague Spring of 1968, the Solidarity movement of 1980–81, and the end to communist rule throughout the region in 1989 all implied a mismatch between the nation and communist rule. Studies of political culture were typically organized, in fact, to explain the conditions under which the political cultures of some nations would or would not resonate with communist political culture. More recent studies tend to break down the assumption of opposition between nationalism and communism and rather study how opposition and authorities construct the nation in their own terms. For instance, instead of treating official communist discourse as simply illegitimate and therefore unworthy of serious scholarly engagement, Kubik (1994) explains how Polish communist rulers tried to appropriate the national imaginary. He also explained Solidarity of 1980–81 as 'oversymbolized,' or laden with more images and meanings than that single social movement could manage. This oversymbolization in 1980–81 also helps to explain postcommunist Poland's political divisions (Ekiert and Kubik 1999).

Not every nation developed a political regime under communist rule that could produce easily imagined dichotomies of 'us' and 'them.' Ceausescu's Romania generated a social consciousness organized around duplicity and complicity. With symbolic violence the norm, and physical violence the backup, self-censorship and dissimulation became intrinsic features of the socialist habitus; the public sphere was based on the production of lies and regulation so profound as to destroy the meaning of public/private distinctions (Kligman 1998). Civil society and the public sphere, rejuvenated in the West alongside its mobilization in Eastern Europe (Keane 1988, Cohen and Arato 1992), could not find a home in Romania. It did, however, find an important place, if in different forms, across the region.

The discourse of civil society found modes of expression, and even optimism in various cultural products and forms of social organization—in Vaclav Havel's *Power of the Powerless* (1986), or Adam Michnik's (1977) notion of 'new evolutionism.' Civil society's culture was argued to be tolerant, pluralist, and progressive.

The very spaces opened up in these advocacies could also produce nationalist expressions that threatened others and even led to war (Bunce 1999). This was most clearly seen in the nationalist mobilizations within civil society around Yugoslavia's breakup. Memories of oppression by others, rekindled by polyvalent symbols, could become the cultural stock for civil society's mobilization to hate. Alternative narratives of the war's making seemed to deny the possibility of panoptic stances from which to explain Yugoslavia's dissolution. Different vantage points, shaped by variable appreciation for the survival of a united Yugoslavia and for different national narratives, have organized competing stories about the war's making, and charged the field of their interpretations (e.g., Woodward 1995 vs. Ramet 1999).

4. *Culture in Transition*

Studies of the nation and of nationalism after communism's collapse were not restricted to those in war. Communism's end offered peoples new opportunities to redefine their nation, and this invited substantial scholarly engagement. These transformations were apparent already in 1988, when for instance the polysemy of Béla Bartók was played out in the last summer of communist monopolistic rule in Hungary (Gal 1991). Some scholars shaped claims to national distinction with their own expertise, clarifying the ways in which proximate languages were related to one another and could distinguish nations, leading others, further removed, to explain the complexities and sources of these arguments. Some utilized less culturally reflexive approaches like rational choice or game theory, but the cultural field and historical legacies of places and peoples continued to shape scholarly approaches to national identification (e.g., Laitin 1998). While the nation demanded some attention to historical legacies and cultural formations, other subjects were more tolerant of distancing cultural analysis.

In the first years after communism's collapse, most studies of transition to markets and democracy were rather formal and structural, as more and more scholars without contextual expertise identified Eastern Europe as the natural laboratory for testing their generalizing theories. This led to significant debate, especially in political science, about the relative value of area studies. Much of the contest hinged upon how much one needed to know about culture and history in order to produce good explanation, but it did not attend to the global cultural formation of transition itself (Kennedy forthcoming). Of course methodology also matters in this variable attention.

Ethnographies of transition were more likely to put into question the guiding concepts of transition, especially by noting the hermeneutic gap between transition's key concepts and their everyday meaning (Burawoy and Verdery 1999). For instance, these inquiries help us appreciate money's magical powers and land's elasticity in transitional economies (Verdery 1996), and the difficulty of distinguishing private from public ownership (Stark and Bruszt

1998). Sometimes ethnographies even challenge the framework in which the transition is envisioned. Ought postcommunist capitalism be conceived as merchant, and not bourgeois, capitalism (Burawoy and Krotov 1992)?

A similar range of questions has emerged around gender studies. As in most transition studies, some gender analysts have kept culture at a distance. By treating gender more as a variable—asking, for instance, whether women's employment chances have been disproportionately hurt by economic transition—cultural analysis is less obviously necessary. But when gender serves as a lens for interpreting social change, as Watson (1993) suggests in her characterization of 1989–92 as a 'masculinist revolution,' gender studies engage profound cross-cultural questions. For instance, terrific debates have occurred over the measure of cultural imperialism involved in Western feminism's categories and modes of inquiry in regard to Eastern Europe (Gal and Kligman 2000).

5. *Culture and Discipline in Eastern European Studies*

As a field of inquiry, Eastern European Studies has been organized through culturally constituted geopolitical categories. During the Cold War, the field was driven by questions of fit between nations and communist rule. After communism's collapse, the field has been reconstituted around questions of transition and nationalism's threat. At the same time, however, communism's collapse has also invited a transformation of the field's methodological premises. With open borders to scientific collaboration, social scientists without area expertise can work with indigenous scholars to produce important findings, bracketing culture as extraneous to the scholarly enterprise in ways that area studies could never manage. More open borders have also meant, however, greater room for ethnographies and other direct qualitative fieldwork, and more intimate portraits of everyday life and assessments of its fit with the larger frameworks guiding global visions of Eastern Europe and its trajectory. This opening to methodological variety has transformed Eastern European Studies, simultaneously making more cultural analysis possible, and making culture more invisible through fusions of scientific horizons across the world.

To the extent disciplines treat culture as a variable, and therefore something that may or may not be incorporated in analysis, the meaning of Eastern Europe as a region tends to be lost, and expertise in its history, institutions, languages, and cultures can be devalued. To the extent scholars engage the 'double hermeneutic' in their work, or explore meaning on the ground and in the cultural relationship between scholarship and the places under study, area studies expertise and cultural theory are more likely to be valued. Indeed, the extensive engagement between social science and the design of change in postcommunist Eastern Europe suggests that culture could become more, rather than less important, in social science about Eastern Europe.

Economists advise market transition, political scientists and sociologists debate democracy's making, linguists articulate nations, and historians adjudicate responsibility for past crimes. Not all of these disciplines are so deeply involved in engaging that double hermeneutic, however. For many, as Žižek (1993, p. 200) has observed, Eastern Europe becomes the mirror in which the West's generalizability can be recognized. In that reflection, Eastern European Studies poses little challenge beyond extension for Western social science. But for those who take up the challenge of difference (Calhoun 1995), the category of Eastern Europe and the translation of inquiry across that space remains part of the question with which not only Eastern European Studies, but also the social sciences themselves, could be culturally reconfigured. What would Eastern European Studies look like if the privilege of nations, men, and global integration through markets became the object of reflection rather than the point of departure for social science about, and out of, the region?

See also: Area and International Studies: Development in Eastern Europe; Communism; Eastern European Studies: History; Eastern European Studies: Society; Nationalism, Sociology of; Socialist Societies: Anthropological Aspects

Bibliography

Adamski W (ed.) 1993 *Societal Conflict and Systemic Change: The Case of Poland, 1980–92*. IfiS, Warsaw, Poland

Brubaker R 1996 *Nationalism Reframed*. Cambridge University Press, Cambridge, UK

Bunce V 1999 *Subversive Institutions: The Design and Destruction of Socialism and the State*. Cambridge University Press, Cambridge, UK

Burawoy M, Krotov P 1992 The Soviet transition from socialism to capitalism: Worker control and economic bargaining in the wood industry. *American Sociological Review* **57**: 16–38

Burawoy M, Verdery K (eds.) 1999 *Uncertain Transition: Ethnographies of Change in the Postsocialist World*. Rowman and Littlefield, Lanham, NJ

Calhoun C 1995 *Critical Social Theory: Culture, History and the Challenge of Difference*. Blackwell, Oxford, UK

Chałasinski J 1997 *Przeszłość I Przyszłość Inteligencji Polskiej*. Świat Książki. Warsaw, Poland

Chirot D (ed.) 1989 *The Origins of Backwardness in Eastern Europe: Economics and Politics from the Middle Ages Until the Early Twentieth Century*. University of California Press, Berkeley, CA

Cohen J, Arato A 1992 *Civil Society and Political Theory*. MIT Press, Cambridge, MA

Ekiert G 1997 *The State Against Society*. Princeton University Press, Princeton, NJ

Ekiert G, Kubik J 1999 *Rebellious Civil Society: Popular Protest and Democratic Consolidation in Poland, 1989–93*. University of Michigan Press, Ann Arbor, MI
Eyal G, Szelényi I, Townsley E 1998 *Making Capitalism without Capitalists: Class Formation and Elite Struggles in Post-Communist Central Europe*. Verso, London
Friedman V (forthcoming) Language in Macedonia as an identity construction site. In: Joseph B (ed.) *When Languages Collide: Sociocultural and Geopolitical Implications of Language Conflict and Language Coexistence*. Ohio State University, Columbus, OH
Gal S 1991 Bartok's funeral: Representations of Europe in Hungarian political rhetoric. *American Ethnologist* **18**: 440–58
Gal S, Kligman G 2000 *Politics of Gender after Socialism*. Princeton University Press, Princeton, NJ
Goldfarb J 1981 *On Cultural Freedom*. University of Chicago, Chicago
Havel V 1986 Power of the powerless. In: Havel V *Living in Truth*. Faber, London
Hertz A 1988 *The Jews in Polish Culture*. Northwestern University Press, Chicago
Hofer T 1968 Anthropologists and native ethnographers in Central European villages: Comparative notes on the professional personality of two disciplines. *Current Anthropology* **9**: 311–5
Hroch M 1985 *Social Preconditions of National Revival in Europe: A Comparative Analysis of the Social Composition of Patriotic Groups among the Smaller European Nations*. Cambridge University Press, Cambridge, UK
Keane J (ed.) 1988 *Civil Society and the State*. Verso, London
Kennedy M D (forthcoming) *Cultural Formations of Post-communism*. University of Minnesota Press, Minneapolis, MN
Kirch A (ed.) 1997 *The Integration of Non-Estonians into Estonian Society: History, Problems and Trends*. Estonian Academy of Sciences, Tallinn, Estonia
Kligman G 1998 *The Politics of Duplicity: Controlling Reproduction in Ceausescu's Romania*. University of California Press, Berkeley, CA
Konrád G, Szelényi I 1979 *The Intelligentsia on the Road to Class Power*. Harcourt, New York
Kubik, J 1994 *The Power of Symbols against the Symbol of Power: The Rise of Solidarity and the Fall of State Socialism in Poland*. Pennsylvania State University Press, University Park, PA
Laitin D 1998 *Identity in Formation: The Russian Speaking Populations in the Near Abroad*. Cornell University Press, Ithaca, NY
Lampland M 1995 *The Object of Labor: Commodification in Socialist Hungary*. University of Chicago Press, Chicago
Magosci P R 1993 *Historical Atlas of East Central Europe*. University of Washington Press, Seattle, WA
Michnik A 1985 *Letters from Prison and Other Essays*. University of California Press, Berkeley, CA
Miłosz C 1951 *The Captive Mind*. Random House, New York
Ossowski S 1963 *Class Structure in the Social Consciousness*. Free Press of Glencoe, New York
Porter B 2000 *When Nationalism Began to Hate: Imagining Modern Politics in 19th Century Poland*. Oxford University Press, Oxford, UK
Ramet S P 1999 *Balkan Babel: The Disintegration of Yugoslavia from the Death of Tito to the War for Kosovo*. Westview, Boulder, CO
Rudnycky I L 1987 *Essays in Modern Ukrainian History*. Harvard University Press, Cambridge, MA
Sayer D 1998 *The Coasts of Bohemia: A Czech History*. Princeton University Press, Princeton, NJ
Stark D, Bruszt L 1998 *Postsocialist Pathways*. Cambridge University Press, Cambridge, UK
Suny R G, Kennedy M D (eds.) 1999 *Intellectuals and the Articulation of the Nation*. University of Michigan Press, Ann Arbor, MI
Szűcs J 1988 Three historical regions. In: Keane J (ed.) *Civil Society and the State*. Verso, London, pp. 291–332
Todorova M 1997 *Imagining the Balkans*. Oxford University Press, Oxford, UK
Verdery K 1991 *National Ideology Under Socialism: Identity and Cultural Politics in Ceausescu's Romania*. University of California Press, Berkeley, CA
Verdery K 1996 *What Was Socialism and What Comes Next?* Princeton University Press, Princeton, NJ
Walicki A 1982 *Philosophy and Romantic Nationalism: The Case of Poland*. Oxford University Press, Oxford, UK
Wandycz P 1984 *The Lands of Partitioned Poland: 1795–1918*. University of Washington Press, Seattle, WA
Watson P 1993 The rise of Masculinism in Eastern Europe. *New Left Review* **198**: 71–82
Wolff L 1994 *Inventing Eastern Europe: The Map of Civilization on the Mind of the Enlightenment*. Stanford University Press, Stanford, CA
Woodward S 1995 *Balkan Tragedy: Chaos and Dissolution after the Cold War*. Brookings Institution, Washington, DC
Zaborowski W 1986 Dichotomous class images and worker radicalism. In: Słomczynski K M, Krauze T K (eds.) *Social Stratification in Poland: Eight Empirical Essays*. Sharpe, Armonk, pp. 105–23
Žižek S 1993 *Tarrying with the Negative*. Duke University Press, Durham, NC

M. D. Kennedy

Eastern European Studies: Economics

The changes in the Eastern European economies during the first decade of the transition to capitalism have been, without exaggeration, momentous. Beginning in the late 1980s, the countries in the region emerged from a system based on communist party rule, state ownership, and central planning, and adopted in its place a system based on multiparty democracy, private ownership, and markets. Although the break with the past was radical, and notwithstanding that the region is poised for period of rapidly improving living standards, the first decade of transition was difficult. Most countries suffered deep recessions, experienced massive redistributions of income and wealth, and struggled with the slow process of building supporting institutions for a market economy. And even for the countries that have advanced furthest, the transition is not over. The next

decade will bring further momentous changes as countries prepare for accession to the European Union.

Transition has been a challenge for economists. Never before had industrial countries attempted to adopt abruptly the institutions of capitalism. New tools had to be developed to understand such phenomena as transitional recession, mass privatization, and the political economy of institutional reform. Transition has also provided a great opportunity to understand the role of the many background institutions of capitalism. By observing the long list of institutions that had to be put in place, and by observing how economies performed when these institutions were lacking, the economics profession has attained a deeper understanding of how even established market economies work.

1. Initial Conditions

By the end of the 1980s, a number of communist countries, including the Soviet Union, had gone through a period of *reform socialism* (Kornai 1992). Among other changes, these reforms allowed greater autonomy for state enterprises, permitted some small-scale private enterprise, and increased the use of prices in allocating economic resources. However, these reforms did not alter the one-party rule and a preponderance of state ownership that were the foundation of the communist system. So the pathologies of the *classical socialist system* remained (Kornai 1992). Resources were allocated according to the dictates of central planners rather than household needs. Production methods were inefficient. And lacking the long-term profitability constraint faced by enterprises on market economies, socialist enterprises with *soft budget constraints* retained insatiable demands for inputs. Together with the continuing reliance on price controls, these demands made shortages pervasive.

Reform socialism led, in fact, to a slackening of the financial discipline on enterprises. And this in turn led to an increase in macroeconomic imbalances. At prevailing prices, the aggregate demand for goods and services exceeded the supply. In the years prior to the transition, communist governments dealt with these imbalances in three ways. They engaged in large foreign borrowing to increase the availability of goods and services; they repressed the inflationary pressures with price controls and rationing of the available goods; and they allowed open inflation.

At the outset of transition, then, reforming governments inherited economies with combinations of large debts, long queues, and high inflation rates (Lipton and Sachs 1990). They also inherited a distorted industrial structure, most notably an over-reliance on heavy industry, and a badly developed service economy. From a longer-term perspective, perhaps the most adverse part of the inheritance of all was the absence of key marketing supporting institutions, such as an established code of commercial law or a supervisory apparatus for market-oriented financial institutions. With this mix of macroeconomic imbalances and microeconomic deficiencies, the path to a market economy must have appeared daunting to the first postcommunist governments.

2. Transition

It is common to put the main steps in postcommunist transition under three rather ugly headings: liberalization, stabilization, and privatization. Much of the subsequent work the economics of transition has concentrated on these processes.

2.1 Liberalization

The first, and arguably the easiest, steps in the transition are the removal of most government controls (see Sachs 1993 for an account of the Polish experience). For example, the government simply stops controlling prices, allowing them to jump to their market levels, or ends prohibitions on new private businesses. In a related step, the government decentralizes decision-making authority to enterprises, in effect dismantling the central planning apparatus.

On the external front, the government moves to allow for freer trade and factor movements. Complicated permission requirements for exporting, high tariff and nontariff barriers to imports, and severe restrictions on foreign ownership are relaxed. In addition to gaining access to foreign markets, goods, and capital, opening up the economy to the rest of the world increases competition and exposes the population to best practice in technology and organization.

2.2 Stabilization

As noted before, postcommunist governments inherit severe macroeconomic imbalances. Even with price controls, the inflation rate typically is already high prior to the transition. Price liberalization is associated with an upward jump in prices as they reach market-clearing levels, ending the previous situation of chronic excess demand. These jumps can also have a longer-lasting impact on inflation to the extent that they cause expectations of higher future inflation.

The most important cause of the high inflation, however, is the underlying growth in the money supply. This has two main sources. First, there is the government's budget deficit. Under communism, governments relied heavily on taxes on enterprise turnover to fund their spending. They also delegated much of the responsibility for social spending directly to the enterprises. Transition has been associated with a fall

Table 1
A decade of transition

	Year transition began	Starting date of stabilization program	Real output 1995 % of 1989	Real output 1999 % of 1989	Average inflation 1989–1999	Private sector % of GDP mid-1999
EU Accession Countries (excluding Baltics)						
Bulgaria	1991	Feb-91	75	67	68.4	60
Czech Republic	1991	Jan-91	85	94	7.8	80
Hungary	1990	Mar-90	86	99	19.7	80
Poland	1990	Jan-90	97	128	49.2	65
Romania	1991	Jan-93	81	74	76.1	60
Slovak Republic	1991	Jan-91	84	101	14.3	75
Slovenia	1990	Feb-92	94	105	12.9	55
Baltic Countries						
Estonia	1992	Jun-92	74	78	24.3	75
Latvia	1992	Jun-92	54	56	35.1	65
Lithuania	1992	Jun-92	42	70	41.0	70
Other South Eastern European Countries						
Albania	1991	Aug-92	75	93	33.4	75
Bosnia and Herzegovina	…	…	…	93	13118.0	35
Croatia	1990	Oct-93	84	80	100.0	60
Macedonia, FYR	1990	Jan-94	53	59	75.6	55
Former Soviet Union (excluding Baltics)						
Armenia	1992	Dec-94	37	48	106.5	60
Azerbaijan	1992	Jan-95	35	47	233.2	45
Belarus	1992	Nov-94	54	81	162.4	20
Georgia	1992	Sep-94	17	31	17.9	60
Kazakstan	1992	Jan-94	44	61	77.3	55
Kyrgyz Republic	1992	May-93	43	61	22.3	60
Moldova	1992	Sep-93	42	31	16.5	45
Russia	1992	Apr-95	49	55	88.0	70
Tajikistan	1992	Feb-95	84	29	688.5	30
Turkmenistan	1992	…	63	61	4.9	25
Ukraine	1992	Nov-94	42	35	169.4	55
Uzbekistan	1992	Nov-94	82	97	304.5	45

Sources: International Monetary Fund (2000) and European Bank for Reconstruction and Development (1995 and 1999)

in revenues and increased spending demands on governments. For many countries, the result has been large budget deficits that are financed in part by printing money. Second, there is the financing of enterprise losses. Although the enterprises attain much greater autonomy, budget constraints remain soft. In other words, their losses are bailed out through support from the government or (more likely) 'soft' loans from the (government-controlled) banking system. In turn, the banks' losses are bailed out by 'soft' support from the central bank, which they finance by printing money.

Since a stable price level is an essential feature of a market economy, one of the most urgent tasks for the post-communist government is to bring down inflation toward single-digit levels. Table 1 provides dates for the beginning of major stabilization efforts in various countries. Given the underlying sources of money supply growth, true stabilization requires reducing the size of the government's budget deficit (or at least reducing reliance on monetary financing of those deficits), and the ability to impose hard budget constraints on enterprises. Each step is difficult as it imposes additional hardships on the population. Thus, for many countries, stabilization has been a slow and indeed a continuing process. As Table 1 shows, average inflation rates for the decade of the 1990s were high by established market economy standards.

For economists, the nature of the post-communist stabilization challenge was not especially novel. The nature and causes of high (and even hyper) inflation were well known, with the most recent experience

being the hyperinflation episodes in Latin America following the debt crises of the 1980s. Some of the challenges were more novel, including the massive and often sudden decontrol of prices, and the mass introduction of new currencies in the countries of the former Soviet Union. By and large, however, economists were able to use their accumulated wisdom on such matters as the importance of ending the printing of money to finance budget deficits, and the usefulness of exchange rate pegs to anchor inflationary expectations in the early stages of a stabilization program. (For an overview of research on postcommunist inflation and the stabilization experience, see Boone and Horder 1998.)

2.3 Privatization

In theory it is possible to have a successful market economy while retaining extensive state ownership. In practice, state ownership is associated with the politicization of enterprise decision-making (Shleifer and Vishny 1994). It is easier for politicians to control firms when the state is the owner. Moreover, politicians will use their control to pursue political ends—say, preserving jobs for constituents—rather than to maximize profits. Thus continued state ownership is an impediment to restructuring enterprises and reallocating resources. Political control is also associated with soft budget constraints, as politicians are willing and able to bail out the losses incurred while pursing their political objectives. Success with stabilization can thus depend on success with privatization.

There are a number of methods for achieving privatization: Sales to the highest bidder; restitution to pre-communist owners; voucher privatization; manager–worker buyouts, etc. Whatever the method, it might be thought that privatization is easy politically. After all, the government is transferring assets to the private sector, potentially at a large discount. In practice, groups seeing themselves as losing from privatization, or not gaining as much from the transfer of assets as they could, often have the power to block an unfavorable privatization proposal. For example, the management of state enterprises might thwart attempts to sell the enterprise to outside owners. Where such blocking power exists, the blocking groups may have to be bought out by the government, possibly by buying access to shares at steeply discounted prices.

A good example of such 'buy-offs' is the early stage of Russian privatization. Then, enterprise insiders—managers and workers—received excellent deals on their companies' shares. While this allowed privatization to proceed, and helped to depoliticize enterprise control, it might have led to excessive insider control (see Boycko et al. 1994). When insiders control a firm, they are likely to use their power to achieve objectives that run counter to efficiency, such as prolonging overstaffing. As outsiders recognize insider power to appropriate the valued added of the enterprise, high insider control makes it difficult to raise outside finance, and thus acts as an obstacle to restructuring and new investment (see Blanchard 1997). Moreover, preferential privatization has led to quite unequal distributions of previously state-owned assets, a topic returned to in Sect. 3.2 below.

The theoretical literature has concentrated on the political economy of mass privatization as well as the various channels for efficiency gains when shifting from state to private ownership (see Roland 2000, for a survey.) There is an even larger empirical literature the records the privatization strategies pursued and investigates the impact of the resulting ownership forms on restructuring behavior. Although this literature does not speak with one voice, the balance of evidence suggests that ownership by enterprise outsiders with large ownership stakes leads to most restructuring (for a flavor of the results and the empirical methods used, see the papers by Estrin 1998 and Frydman et al. 1999).

3. The Transition Experience: Pain and Progress

3.1 Transitional Recession

All the transition economies experienced severe and, for many, long-lasting recessions (Layard 1998). Table 1 shows the level of real GDP in both 1995 and 1999, each as a percentage of output in 1989. The table makes for grim reading. In 1995 all countries had measured real output levels that were below levels achieved in 1989. The picture had brightened by 1999, with Poland, the Slovak Republic, and Slovenia surpassing their 1989 output levels, and a number of other countries approaching this benchmark. In the countries of the former Soviet Union, however, the picture remained bleak. In Russia, for example, real output in 1999 was just 55 percent of the 1989 level.

These real output statistics almost certainly overstate the true output contractions. Much of the lost output had little value—and perhaps even negative value—as valuable raw materials were destroyed in irrational production processes. Nonetheless, the recessions were real, with sharp reductions in employment and expansions of unemployment.

Why did the transition economies suffer these transitional recessions? After all, if the communist system was so inefficient, should we not expect that real output would increase as countries adopt market systems? The early hope of experts on communist systems was that the postcommunist economies would move from a situation of generalized excess demand (or shortage) to macroeconomic balance, without tipping into a situation of generalized excess supply (Kornai 1995). Various explanations for the sharp contractions have been given in the literature.

Early explanations focused on the impact of the sharp postliberalization price jumps, and on Keynesian-type explanations of insufficient demand. The sharp jump in prices was thought to have limited real working capital severely, forcing enterprises to cut their production (see, e.g., Calvo and Coricelli 1992). On the demand side, a general lack of consumer and investor confidence, combined with tight government fiscal and monetary policies, was initially thought to be responsible for pushing real output below the economy's potential to produce. As time passed, and the transition economies failed to recover, or recovered very slowly, these explanations were revised. Economists began to look for more systemic explanations. One important explanation focused on the 'disorganization' that results in the no-man's-land between systems (Blanchard and Kremer 1997, Roland and Verider 1999). For a time after a central planning apparatus has been destroyed, but before market coordination is properly established, the ability to coordinate economic activity is severely impaired. Supply disturbances in one part of the economic system tend to cascade through the economic system, impairing severely the economy's potential to produce.

A related explanation focuses on the inherent unemployment bias present in market systems. When unemployment falls below the *natural* (or structural) *rate of unemployment*, inflation tends to rise. Thus, stable inflation is only consistent with what can be a relatively high rate of unemployment. (That this structural unemployment rate can be quite high is demonstrated by the high rates of unemployment in many Western European economies, even in the expansion phases of their business cycles.) In contrast, under communism, unemployment was negligible, and inflation was controlled by fiat. As an economy goes through a postcommunist transition it can no longer sustain its previously high rate of activity without suffering accelerating inflation. Over time it is expected that the superior efficiency of market coordination will lead to superior output performance, but in the short-run the loss of ability to run a high-pressure economy dominates, and output falls.

Lastly, a number of authors have stressed the impact of the large reallocations that must take place during the transition (see, e.g., Gomulka 1998). Some enterprises, and even entire industries, decline rapidly as there is little demand for their products in the newly liberalized economy. The result is a large but temporary decline in output as it takes time for the displaced resources to find productive employment in expanding sectors.

3.2 Large Redistributions of Income and Wealth

Transition is associated with substantial changes in household incomes and balance sheets. Factors affecting the income relativities include the large rise in unemployment and falls in (especially female) participation, the shift in resources from heavy industry to lighter industry and services, and hardships faced by older workers with specific skills in adjusting to the new pattern of labor demands.

In addition to the large income gains and losses for particular households, there has also been a large increase in overall income inequality. For the mid-1990s, Milanovic (1998) reports substantial increases in inequality as measured by the Gini coefficient for income per capita compared with the late 1980s, with particularly large increases for the countries of the former Soviet Union. The Gini coefficient for Russia increased to a staggering 48 in the period 1993–5 from 24 in the period 1987–8. This compares with much smaller increases for more advanced reformers, with the Gini rising to 28 from 26 for Poland, and to 23 from 21 for Hungary for the same periods.

There have also been very large shifts in wealth distribution (although reliable numbers are hard to come by). Under communism, price controls and generalized shortages forced people to accumulate undesired liquid wealth holdings, creating a *monetary overhang* of latent spending power. With price liberalization, people were able to spend down these undesired balances, contributing to the large price jump. However, this price jump also reduced the purchasing power of all (nonindexed) nominal wealth, causing substantial but unequal wealth losses.

Privatization has also affected wealth distribution, although given the low value of much of the communist-era capital stock, the amount of wealth available for distribution should not be exaggerated. The distributional impact of privatization depends on the privatization method pursued. At one extreme, sales at reasonably fair prices (as in Hungary) should have rather minor impacts on wealth inequality, while at the other, extreme giveaways to the politically well-connected of valuable energy-sector assets (as in the second wave of privatization in Russia) must have had a major impact.

If we take a broader view of wealth to include the value of assets such as entitlements to pensions and job incumbency, the impact of transition on wealth distribution has obviously been great.

3.3 Supporting Institutions for Capitalism

For economists and political scientists, one of the unanticipated benefits of the transition experience is that it has given us a deeper understanding of the role of institutions in the capitalist system, and the difficulties involved in their establishment (Roland 2000). An incomplete list of the institutions that facilitate economic activity in a private ownership economy includes the following: Regulatory and supervisory authorities for financial intermediaries and securities

markets; company laws that protect the interests of outside shareholders; a body of contract law that facilitates long-term contracting; an independent judiciary; etc.

Reformist governments can liberalize and stabilize relatively quickly, and even privatization can be well advanced in a few years with the necessary will. Institutional development takes more time. The Czech government provides a good example of a country that pushed ahead quickly with price decontrol, disinflation, and sales of state assets, but the inadequacies of its background institutions were shown in a series of scandals related to insider dealing (see LaPorta et al. 2000). These failures undermined investor confidence, and clearly hurt the performance of the Czech economy.

On the research front, a great deal of attention has been given to the economics and politics of large-scale institutional change. One of the most debated questions is the relative merits of 'shock therapy'—the rapid movement along as many fronts as possible to the desired institutional configuration. Lipton and Sachs (1990) and Sachs (1993) provide early arguments for moving quickly. In the gradualist camp, Roland (2000) outlines models that suggest the merits of a more phased approach, particularly when the outcomes of the reform process are uncertain.

4. *The Road Ahead*

In highlighting the difficulties of transition, the momentousness of the achievement must not be missed. Political and economic freedoms have been restored; the state has largely withdrawn from the ownership and control of productive resources; and markets have been allowed to exist and to flourish. The capitalist political–economic system, though not flawless, has proven its capacity to generate growth and improvements in living standards (see Maddison 1995). In contrast, the communist system brought repression, and eventually stagnation. The future should be bright for those countries that continue to advance down the road to a free economy (see Kornai 1990).

At the outset of transition, Eastern European politicians said their goal was 'to return to Europe' (Sachs 1993). The ultimate expression of such a return would be accession to the European Union (EU). For the more advanced reformers, this goal should be in reach within the first decade of the twenty-first century. The process of accession will be difficult, of course, forcing combinations of further liberalization and the adoption of often-costly EU regulations. Meeting the convergence criteria for joining the Economic Monetary Union (EMU), and thus the right to adopt the euro currency could be especially painful. Moreover, easy immigration to the richer countries to the west could drain the Eastern European of talent needed to rebuild their economies. Nonetheless, accession would be a remarkable achievement, given where these economies started from at the beginning of the 1990s. It would also mark a clear date for the end of transition.

See also: Area and International Studies: Development in Eastern Europe; Democratic Transitions; Eastern European Studies: History; Eastern European Studies: Politics; Eastern European Studies: Society; Economic Transformation: From Central Planning to Market Economy; European Integration; European Union Law; Privatization of the Economy; Revolutions of 1989–90 in Eastern Central Europe; Transition, Economics of

Bibliography

Blanchard O 1997 *The Economics of Post-Communist Transition*. Clarendon Press, Oxford, UK

Blanchard O, Kremer M 1997 Disorganization. *Quarterly Journal of Economics*, **CXII**(4)

Boone P, Horder J 1998 Inflation: Causes, consequences, and cures. In: Boone P, Gomulka S, Layard R (eds.) *Emerging from Communism: Lessons from Russia, China, and Eastern Europe*. MIT Press, Cambridge, MA

Boycko M, Shleifer A, Vishny R 1995 *Privatizing Russia*. MIT Press, Cambridge, MA

Calvo G, Coricelli F 1992 Stabilizing a previously centrally planned economy: Poland 1990. *Economic Policy* **14**

Estrin S 1998 Privatization and restructuring in Eastern Europe. In: Boone P, Gomulka S, Layard R (eds.) *Emerging from Communism: Lessons from Russia, China, and Eastern Europe*. MIT Press, Cambridge, MA

European Bank for Reconstruction and Development 1995 *Transition Report: Investment and Enterprise Development*. EBRD, London

European Bank for Reconstruction and Development 1999 *Transition Report: Ten Years of Transition*. EBRD, London

Frydman R, Gray C, Hessel M, Rapaczynski 1999 When does privitization work? The impact of private ownership on corporate performance in the transition economies. *Quarterly Journal of Economics* **114**(3)

Gomulka S 1998 Output: Causes of decline and recovery. In: Boone P, Gomulka S, Layard R (eds.) *Emerging from Communism: Lessons from Russia, China, and Eastern Europe*. MIT Press, Cambridge, MA

International Monetary Fund 2000 *World Economic Outlook* (October). IMF, Washington, DC

Johnson S, LaPorta R, Lopez-de-Silanes F, Shleifer A 2000 Tunneling. *American Economic Review, Papers and Proceedings* (May)

Kornai J 1990 *The Road to a Free Economy: Shifting from a Socialist System—The Example of Hungary*. W W Norton, New York

Kornai J 1992 *The Socialist System: The Political Economy of Communism*. Princeton University Press, Princeton, NJ

Kornai J 1995 *Highways and Byways: Studies on Reform and Post Communist Transition*. MIT Press, Cambridge, MA

Layard R 1998 Why so much pain? An overview. In: Boone P,

Gomulka S, Layard R (eds.) *Emerging from Communism: Lessons from Russia, China, and Eastern Europe*. MIT Press, Cambridge, MA
Lipton D, Sachs J 1990 Creating a market economy in Eastern Europe: The case of Poland. *Brookings Papers on Economic Activity* **1**
Maddision A 1995 *Monitoring the World Economy*. OECD, Paris
Milanovic B 1998 *Income, Inequality, and Poverty During the Transition from Planned to Market Economy*. World Bank, Washington, DC
Roland G 2000 *Transition and Economics: Politics, Markets, and Firms*. MIT Press, Cambridge, MA
Roland G, Verdier T 1999 Transition and the output fall. *Economics of Transition* **7**(1)
Sachs J 1993 *Poland's Jump to the Market Economy*. MIT Press, Cambridge, MA
Shleifer A, Vishny R 1994 Politicians and firms. *Quarterly Journal of Economics*, **CIX**(4)

J. McHale

Eastern European Studies: History

The collapse of communism in 1989 unleashed a flood of articles and books in political science, sociology, and international relations. The primary aim of this work was to identify and explain the causal forces and teleology propelling the region's 'transition' from communism to free-market democracy. When comparisons with Latin America, Greece, and Iberia are added to the mix, it becomes possible to speak of an entire genre of 'transitionology' whose main source of sustenance has been the history of Eastern Europe in the late twentieth century. While some historians did seek to draw usually problematical parallels between the interwar period and postcommunist developments, as a whole the discipline of history tended to remain immune to the fixation on the present. The wars of Yugoslav succession form the one notable exception, since they spawned substantial literature dedicated to elucidating the perceived connections between nationalism's role among the south Slavs in the past and the present.

And yet the underlying importance of 1989 for history far exceeds its significance for the other social sciences, for it has given birth to the need to reconceptualize Eastern European history in an entirely new framework. 'Never has an ideology and a system of power built on such vast pretensions and still greater human tragedies,' it has been noted in reference to 1989, 'fallen so ignobly as did the red star that sought to unite the lands from the gray shores of the Baltic to the azure coves of the Adriatic within a single allegiance' (Banac 1992, p. ix). Although the degree of *de facto* control varied widely from country to country, one of the communist system's greatest pretensions was its stated intent to impose a uniform way of viewing and interpreting the past.

In the short term, from the historian's standpoint, the single most noteworthy aspect of the collapse of communism in Eastern Europe is the opening of previously closed archives. Unlike in Russia, where the old way of doing things continues to erect obstacles to archival access, the archives of Eastern Europe are now available to researchers with few if any restrictions. Formerly taboo topics from the interwar period can be studied fully. To take Czechoslovakia as an example, the result has been a monograph on the Republican (Agrarian) Party's critical place in the democratic Czechoslovak political system between the wars (Miller 1999), a study of the Slovak nationalist movement's place in the interwar state and society (Felak 1994), as well as the first English-language study of Czechoslovak diplomacy during this period to be based on archival sources (Lukeé 1996). The thorough coverage afforded to Czechoslovak interwar history is a model that could be applied easily to the history of the Baltic States, Bulgaria, and Romania (Hungary, Poland, and Yugoslavia have received better treatment already, especially in their own national historiographies).

Unlike the study of the interwar period, which more often than not entails grappling with traditional themes, the collapse of communism has cleared the way for new approaches to World War II and its immediate aftermath. Instead of simply positing the war as a prelude to the communist seizure of power in Eastern Europe, the attempt is made to draw a more nuanced picture. The gray hues produced first by collaboration with the Nazis and then by retribution against those deemed guilty of this sin make the conventional story of anticommunist struggle more complex and ambiguous (Deak et al. 2000).

Above all, the immediate post-1945 period, when communists seized power throughout Eastern Europe, and the four decades of communist rule itself can now be empirically researched in a manner that is free from ideological distortions. Not surprisingly, these years have attracted the greatest attention from Eastern European historians themselves. There have also been original and meaningful English-language contributions to the study of the post-1945 period (Kenney 1997, Connelly 2000). While there have been no great revelations from the archives that would fundamentally alter our understanding of the communist period, it is now possible to document just how diverse Eastern Europe became while remaining under Soviet influence. When narrating the history of Polish–East German relations, one historian aptly speaks of a 'cold war in the Soviet bloc.' While this is certainly the most extreme such case, the basic model obtains for all of Eastern Europe, where, after 1945, individual states sought to realize their own national aims within a common ideological framework (Anderson 2001). The concept of 'national communism' was first developed

by political scientists to explain the variations that emerged in the Eastern European states following the death of Stalin in 1953. In the aftermath of 1989, this concept has acquired new relevance and salience. Thus, it is also now possible to explain, with the aid of source material, the similarities as well as the differences between the Eastern European communist states during the period of so-called high Stalinism, which stretched from 1948/49 to 1953 and witnessed the creation of something that, at least on the surface, approximated a monolithic, Moscow-controlled bloc. (The best source for English-language material on post-1945 Eastern Europe is the *Cold War International History Bulletin*.)

Independently of the topics addressed above, newer methodological trends have also made their mark on Eastern European history, though, for the time being anyway, they are less prominent in this field than elsewhere (Porter 2000, Todorova 1997).

Over the long term, the lasting impact of 1989 will lie close to the agenda of historians who are interested in the role of language and in the changes experienced by language over time. The collapse of communism has initiated a process that promises ultimately to render the very concept of Eastern Europe passé in any meaningful political and thus contemporary sense. The mix of factors driving the distinction has often changed from period to period, whereby foreign domination and perceived 'backwardness' have always played a central role. But the fundamental distinction between Eastern and Western Europe has been in existence since the Enlightenment, i.e., since the birth of modern notions of nationhood and statehood. As an Enlightenment thinker, Johann Wolfgang Herder stood out because he argued that the rights and prerogatives of modern nationhood should be extended to the stateless and usually 'small' Slavic nations of Eastern Europe.

Reminiscent of Herder's understanding of self-determination, the demise of communism has opened the way for these same nations to define their statehood within the context of an expanding European Union, which represents an unprecedented supranational pooling of resources, as well as an expanding NATO, which brings with it an unprecedented Western security guarantee for Eastern Europe. Furthermore, the extension of state-creating nationalism into Byelorussia, Ukraine, and Central Asia means that the very notion of what constitutes Europe is now confronting a new and novel environment. Since all of the states sandwiched into the vast region between Germany and Russia will, presumably, one day join the European Union, the old East–West dichotomy may remain in place in civilizational and cultural terms for some time to come, but the stark political distinction between the two halves of Europe is already beginning to fade.

What does all of this portend for the field of Eastern European history? First of all, it confronts a growing need to formulate and answer questions that can be fit into a broader European framework. This is a tricky proposition, because it denotes a challenge to previously dominant assumptions about the region's peculiarity. Second, the very concept of Eastern European history will, in all likelihood, fall by the wayside completely. In keeping with the trend towards regionalization within the European Union, the individual subcomponents of Eastern Europe, East-Central and Southeastern Europe, together with the Baltics and Ukraine, will retain their usefulness as broader categories of historical research and thinking. However, the general decline in the number of monographs devoted to Eastern European historical subjects shows no sign of abating, just as interest in area-studies and linguistic training continues to diminish.

See also: Area and International Studies: Development in Eastern Europe; Democratic Transitions; Eastern European Studies: Culture; Eastern European Studies: Economics; Eastern European Studies: Politics; Eastern European Studies: Society; Revolutions of 1989–90 in Eastern Central Europe; Socialist Societies: Anthropological Aspects

Bibliography

Anderson S 2001 *A Cold War in the Soviet Bloc. Polish–East German Relations, 1945–1962*. Boulder, CO
Banac I (ed.) 1992 *Eastern Europe in Revolution*. Ithaca, NY
Connelly J 2000 *Captive University. The Sovietization of East German, Czech, and Polish Higher Education, 1945–1956*. Chapel Hill, NC
Deak I, Gross J T, Judt T (eds.) 2000 *The Politics of Retribution in Europe. World War II and its Aftermath*. Princeton, NJ
Felak J R 1994 *At the Price of the Republic. Hlinka's Slovak People's Party, 1929–1938*. Pittsburgh, PA
Kenney P 1997 *Rebuilding Poland. Workers and Communists, 1945–1950*. Ithaca, NY
Lukeé I 1996 *Czechoslovakia between Stalin and Hitler. The Diplomacy of Edvard Beneé in the 1930s*. New York
Miller D E 1999 *Forging Political Compromise. Antonân Évehla and the Czechoslovak Republican Party, 1918–1933*. Pittsburgh, PA
Porter B 2000 *When Nationalism Began to Hate. Imagining Politics in Nineteenth-Century Poland*. New York
Todorova M 1997 *Imagining the Balkans*. New York

W. L. Blackwood

Eastern European Studies: Politics

Eastern European studies has been the stepchild of Slavic Studies, Sovietology, or a number of research centers and projects that endeavored to study 'actually

existing socialism,' command economies, communist ideology and culture, informal markets, social movements, Marxist–Leninist theories, or other specialized topics in the political economy of the region. Prior to 1989, political scientists were likely to characterize the region as a cluster of satellite states, largely marginal to the political and economic institutions of the West, except as systemic prototypes. The fascinating histories of the region and the debates about the sources and cause of 'backwardness' were not widely read. The 1990s have transformed the intellectual terrain as much as the political.

Since 1989, there has been a veritable avalanche of material—reports, evaluations, diagnoses, prognoses, disputes, re-evaluations, chronicles, and other descriptive accounts. The academic literature has expanded at least as dramatically as the popular literature, and thus far shows few signs of waning. Virtually every major journal in the social sciences has had at least one special issue entirely devoted to the demise of communism; many have had five- and 10-year retrospectives as well. Professional associations have organized special workshops and conferences; foundations have funded seminars both here and abroad; study groups, dissertations, and research projects have proliferated. In due course, these scholarly convocations have translated, edited, and published a host of material. To this day, annual meetings include thematic panels and plenary sessions on the Transition. The sheer amount of intellectual activity and interest in and on Eastern Europe has been overwhelming. Even ignoring work that is journalistic or blatantly ideological, the research archives are flooded.

Area specialists were, of course, deeply involved in this process. Journals that focused on Eastern Europe, on socialism or communist political systems, or on the comparative analysis of institutions were deluged with requests for symposia, reprints, reviews, and manuscripts. Several new journals were established; a surge of newsletters, bulletins, and e-journals entered the scene, providing a wealth of information on current events. It is fair to say that the bulk of the work produced to date has yet to be absorbed or even thoroughly reviewed by regional specialists.

One intriguing aspect of the collapse of the socialist system was the fact that it was largely nonviolent, even harmonious—typically a 'bloodless coup' or a 'velvet revolution.' In academic circles, however, the debate has been more contentious. For Eastern European specialists, the fall of communism meant a dramatic change of government and a host of domestic problems and challenges specific to the new governments. A common response among empirical scholars was to add Eastern European and the Balkan countries to the existing pool of countries that had successfully negotiated regime change and democratization. The temptation to apply carefully crafted theories of the transition from authoritarianism to democracy, based on Latin America or Spain, for example, was nearly irresistible. For area specialists, such generalizations were clearly premature. Democratization was a fragile process, and not a foregone conclusion. Scholars with comparative interests in building democratic institutions were typically newcomers to this region; they could not claim the cultural knowledge and experience of these countries as broad and as deep as that enjoyed by resident scholars or area specialists. Moreover, Eastern Europe had produced a considerable number of homegrown activists and intellectuals as well as political commentators. After decades of repression, the crumbling Wall was greeted with a surge of publications and collaborative projects aimed at the West.

For the vast majority of Westerners, the events of 1989 were completely unexpected. The 'man on the street' was typically euphoric or incredulous. Political scientists were, however, shocked by the 'embarrassing failure' of theory to predict or explain the sudden end to the Cold War. An auxiliary issue became deciding which experts if any might have alluded to the possibility. Efforts to understand and explain the 'revolution' have proliferated, with the most successful involving scholars from the region working with scholars from the West. The work is largely multidisciplinary, rather than interdisciplinary. Few analysts could ignore the implications of combining economic and political reform. Such collaboration seems to have given rise to a whole new field of inquiry, labeled somewhat facetiously 'Transitology' or 'Consolidology.'

As a result, the 'revolution' invited a whole new debate on the most appropriate methods for studying political change. This debate centers on, 'What questions can be fruitfully approached using comparative analysis?' and 'How much emphasis should be placed on unique, country-specific aspects of political change?' These debates have tended to be louder, more vociferous, and far more colorful than is customary in methodological matters.

1. *The Big Questions*

Given the enormous amount that has been written about socialism, democracy, and economic development, revising social theories was a growth industry. After the seismic events of 1989, virtually all perspectives on and approaches to political theory required rethinking. Did socialism fail, and if so, why? Have markets been successfully introduced? Why have they failed to meet popular expectations? Has 'market socialism' proved to be an oxymoron? Is it possible to initiate 'dual track' transformations, modernizing economic and political institutions simultaneously? Or, to paraphrase one text: is it possible to rebuild the ship of state at sea? And where could one find an appropriate blueprint for such an endeavor? Needless

to say, these questions have ignited more heat than light. They do, however, promise to provide fertile ground for future theorizing.

At least since Schumpeter, analysts have argued that modern democracies emerge from capitalist processes. There are authoritarian countries that have market-oriented economies and authoritarian countries that have nonmarket economies; there are countries governed by popular elections with market economies. However, there are no countries that have both popularly elected leadership and state-controlled economies. Although there were undeniable tensions between democracy and capitalism, the conventional wisdom linked them irretrievably.

Comparative work on the transition to democracy was well established. Beginning with Portugal in 1974, a wave of successive democratic upheavals had swept several European and then Latin American countries. Analysis and theoretical work on these countries had produced a number of exemplary comparative studies. Not surprisingly, these authors were keen to add cases and test and validate their previous conclusions with new data from Eastern Europe; to create, in short, a laboratory for studying the processes involved in nation-building, democratizing, converting economies, creating and legitimating the rule of law, and nourishing civil society.

2. *Redrawing the Boundaries*

When territorial boundaries change, maps must of course be redrawn; but in Eastern Europe, the very notion of what constitutes an 'area' or 'region' of the world also needs to be redefined. The collapse of communism created some 15 new nations, and still counting. Although contiguous geographically, the primary quality these countries share is 40 or more years of Soviet domination. Culturally, linguistically, and ethnically, the postcommunist states are quite diverse. Moreover they seem to be differentiating rather than converging. Should the area be subdivided, merged, or treated as separate states. Creating new nations is a difficult process: new governments are formed; new constitutions are written; new treaties are made. For specialists on Europe or for political scientists generally, the number of units to be studied—whether defined as states, localities, urban centers, trading partners, electoral districts, monetary unions, or bilateral agreements—has burgeoned. Such entities, and the divisions among them, constitute the most basic elements of political and social inquiry.

The borders between East and West, both *de facto* and *de jure*, have obviously changed. But even more fundamentally, our conceptual boundaries have shifted and these are much harder to detect or to map. It is not yet clear whether the partially reborn countries of Eastern Europe and the Balkans will join (or rejoin) Europe or fall back into authoritarianism and fevered nationalism.

In almost every subdiscipline of politics, there have been calls for revising the theoretical agenda. Like other fields, International Relations had neither anticipated nor could explain the transformations underway. Both the European Union and NATO expanded. Ideally, the end of the Cold War should have signified a substantial lessening of security concerns. In practice, however, the focus moved to terrorism and subnational groups. Instead of a bifurcated world, the risk today seems to be a proliferation of nuclear weapons and regional arms' races.

Macrohistorical approaches were also suspect. Historical materialism and the Grand Experiment had both fizzled. The world system seemed about to topple and it was all but impossible to identify the matrix of power relations in place. Eastern Europe was liberated from the Soviet Empire, decolonized, and eager to join the banquet of Europe—an altogether different place in the world.

3. *Policy Issues*

At the outset of transition, reforms involved moving toward a market economy and a democratic state. Two competing theories were put forth to assess the speed and effectiveness of economic reform: Gradualism and Shock Therapy. At present, there is still little agreement about which strategy works best. A host of important but less urgent policies were also reviewed. Experts in comparative law, business and management, the role of the media, privatization, nation-building, social safety nets, gender relations, and institutional reform of all sorts descended on these fledgling democracies, to investigate and advise. Specialized studies on poverty and social welfare, migration, pension reform, health care, and unemployment were conducted, some for the first time. Public opinion polling and even focus groups became very frequent occurrences.

Among academics, transitology was flourishing, even if the contours were imprecise. Was transitology a branch of comparative politics or an uncharted new field? Did it apply to all forms of change, or primarily political transitions. Schmitter traced the earliest theoretical propositions in the field to Machiavelli, who argued that introducing any new system implies that all those who profited from the former state of affairs are now enemies, while those who might stand to gain from a change become at best lukewarm allies. The fundamental premise of transition is uncertainty about the outcome; in contrast, democratic consolidation appears to develop in consecutive steps.

What many comparativists did not seem to realize was that the wall separating European Studies from Eastern European studies had disappeared long before

the collapse of the Berlin wall. Political theory and comparative politics had crossed the Oder considerably before the arrival of 'democracy,' let alone transitology or consolidology. Efforts to locate these aspiring democracies in terms of their institutions, their standard of living, or their expectations for the future were certain to be off the mark.

A number of theoretical treatments of the course of transition seem to have promise. In comparative terms, the 'paths' chosen were quite different and not necessarily predictive of success. However, the point of departure tended to be enormously important. Economic and political transitions seemed to be path dependent. They were not universal, however. The legacies of socialism persisted throughout the region; however, they were mediated by unique cultural features specific to each country. Nationalism was widespread in the region; however, its effects and consequences differed enormously. Economically, virtually all of the countries have progressed; politically the future is less clear. Liberal political theory privileges concepts of citizenship that are inclusive if not universal. Yet the politics of transition seems to mean that nationalist figures were more likely to endorse democratic politics, if only for the majority. A free and unfettered market economy meant joining or rejoining the prosperous West, a prospect that European governments either shunned or deferred, a reunified Germany being the obvious exception. In short, a great many theoretical ideas have been redefined.

4. *Gender*

A fairly sizeable literature has developed on questions of gender and political transitions. With respect to politics, the number of females elected or appointed to public office has declined compared to the numbers during the *ancien regime*. Unemployment, in contrast, has climbed relative to the rate for males, excepting only Hungary. Historically, women were better educated than men on average; women were concentrated in the professions and the service sector under communism. Since these are the sectors most in demand, it is ironic that women are being displaced. In medicine, for example, one finds that men are quick to hang a shingle, while women physicians remain in depressed clinics in the public sector.

Gender inequality has become a large issue in the West; unraveling the sources and consequences receives a great deal of academic attention. Historically, communism attempted to insure gender equality through equal rights and labor market participation, although the evidence suggests these efforts were short-lived. The corollary proposition of why socialism failed to 'liberate' women despite the ideology and the rhetoric is still perennially fascinating. Analyzing trends becomes a method for proving or disproving macro-historic issues such as the reasons for segmented employment and gendered markets.

5. *Elites*

A related topic concerns the composition of elites, and in particular how to account for the persistent advantages enjoyed by the *nomenklatura*. Under communism, the top leadership in both the party and state enterprise were carefully screened for ideological consistency and rewarded for loyalty to the regime. A recurring critique was that class stratification under socialism was based on privileges instead of property, but that otherwise little had changed. Whether based on ideology ('Reds') or skills and education ('Experts'), the communist system had certainly not banished inequality.

To date, the research suggests that in Eastern Europe, there have been changes, although perhaps not as dramatic as those that occurred immediately after World War II. In Hungary, the economic elite has changed the most; in the Czech Republic, political power has clearly shifted; and in Poland, the cultural elite is composed of recent recruits. However, in other parts of the Communist bloc, managers became owners with little need for new blood.

6. *Social Protest in Eastern Europe*

Despite peaceful roundtable agreements that initiated transitions, most of the countries of the region had relatively large opposition movements and dissidents who had been more or less silenced. These groups were often involved in deliberations, pact-making, and electoral victories. The origins and activities of groups such as Charter 77 and Solidarity were of considerable interest, even though neither group had been active recently.

Political activity during the early transition period varied by country in terms of both magnitude and duration. Not surprisingly, a variety of groups and organizations contested the policies and austerity imposed by the new regime; moreover, citizens were more acquainted with protest activity than with compromise and negotiation. Comparing the patterns in East Germany, Hungary, Poland, and Slovakia between 1989 and 1993, Ekiert and Kubik (1998) conclude that relative deprivation, whether measured in objective or subjective terms, was unrelated to forms of contention. The extent of protest is best explained by institutional and resource mobilization theories. Moreover, democratic consolidation was not undermined by a high degree of protest activity. In fact, when collective demands are moderate and the strategies orderly, then protest may actually contribute to the development of robust democratic institutions.

An enormous number of research questions have been turned topsy-turvy; many eternal verities have been challenged. Rational choice theorists have been forced to include civil society and social contracts to make sense of political change. Post-communist ballots tend to be long, filled with political parties that come and go; indeed, despite the 'overthrow' of communism, socialists and social democrats are still an organized and powerful force. Predicting winning candidates or votes cast has become difficult. With little variation in economic position, at least at the outset of the transition, it is not surprising that citizens neither recognized nor voted as their economic interests dictated. Parties were not more cohesive after successfully overthrowing the communists; in fact, this was usually the point when political fragmentation began in earnest.

7. *Conclusion*

The political changes in Eastern Europe have clearly opened a wealth of social science theory and data. Thus far, there has been an unprecedented number of productive projects, both theoretical and practical. Area specialists have been indispensable to this work. In the process, they have helped to refine and define the most pressing questions drawn from a venerable literature. To be sure, there are still more questions than answers, but progress is readily apparent.

See also: Communism; Democratic Transitions; Eastern European Studies: Culture; Eastern European Studies: Economics; Eastern European Studies: History; Eastern European Studies: Society; Revolutions of 1989–90 in Eastern Central Europe; Socialist Societies: Anthropological Aspects

Bibliography

Bunce V 1995 Should transitologists be grounded? *Slavic Review* **55**: 111–27

Bunce V 1999 *Subversive Institutions: The Design and the Destruction of Socialism and the State*. Cambridge University Press, New York

Chirot D 1991 *The Crisis of Leninism and the Decline of the Left: The Revolutions of 1989*. University of Washington Press, Seattle, WA

Domanski H, Heyns B 1995 Toward a theory of the role of the state in market transition: From bargaining to markets in post-communism. *European Journal of Sociology* **36**(2)

Ekiert G, Kubik J 1998 Contentious politics in new democracies: East Germany, Hungary, Poland, and Slovakia, 1989–93. *World Politics* **50**: 547–81

Elster J, Offe C, Preuss U K, Boenker F, Goetting U, Rueb F W 1998 *Institutional Design in Post-Communist Societies—Rebuilding the Ship at Sea*. Cambridge University Press, New York

Eyal G, Szelenyi I, Townsley E 1998 *Making Capitalism Without Capitalists: The New Ruling Elites in Eastern Europe*. Verso, New York

Fleron F J, Hoffmann E P 1993 *Post-Communist Studies and Political Science: Methodology and Empirical Theory in Sovietology*. Westview Press, Boulder, CO

Funk N, Mueller M (eds.) 1993 *Gender Politics and Post-Communism: Reflections from Eastern Europe and the Former Soviet Union*. Routledge, London

Gal S, Kligman G (eds.) 2000 *Reproducing Gender: Politics, Publics, and Everyday Life after Socialism*. Princeton University Press, Princeton, NJ

Heyns B, Bialecki I 1991 Solidarnosc: Reluctant vanguard or makeshift coalition? *American Political Science Review* **85**: 351–70

Keddie N R (ed.) 1995 *Debating Revolutions*. New York University Press, New York

Lane D 1996 *The Rise and Fall of State Socialism: Industrial Society and the Socialist State*. Polity Press, London

Levesque J 1997 *The Enigma of 1989: The USSR and the Liberation of Eastern Europe*. University of California Press, Berkeley, CA

Linz J J, Stepan A 1996 *Problems of Democratic Transition and Consolidation: Southern Europe, South America, and Post-Communist Europe*. The Johns Hopkins Press, Baltimore, MD

Nelson J M, Tilly C, Walker L (eds.) 1998 *Transforming Post-Communist Political Economies*. National Academy Press, Washington, DC

Przeworski A 1991 *Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America*. Cambridge University Press, Cambridge, UK

Rose R, Mishler W, Haerpfer C 1998 *Democracy and Its Alternatives: Understanding Post-Communist Societies*. The Johns Hopkins University Press, Baltimore, MD

Schmitter P C, Karl T L 1994 The conceptual travels of transitologists and consolidologists: How far to the East should they attempt to go? *Slavic Review* **53**: 173–85

Staniszkis J 1991 *The Dynamics of the Breakthrough in Eastern Europe: The Polish Experience*. University of California Press, Berkeley, CA

Stark D, Bruszt L 1998 *Postsocialist Pathways: Transforming Politics and Property in East Central Europe*. Cambridge University Press, Cambridge, UK

Verdery K, Burawoy M 1999 *Uncertain Transition: Ethnographies of Change in the Postsocialist World*. Rowman & Littlefield, New York

Weil F D 1992 *Democratization in Eastern and Western Europe: A Research Annual*, Vol 1. JAI Press, Greenwich, CT

Woodward S 1995 *Balkan Tragedy: Chaos and Dissolution After the Cold War*. Brookings Institution, Washington, DC

B. Heyns

Eastern European Studies: Society

'Eastern Europe' is as much a historically shaped sociocultural concept with contents and connotations that have changed markedly over time, as a reality of

recurrently shifting geopolitical divisions. The standing aspiration of the region's nations has been their full integration into a unified Europe driven by Western values and norms. After looking at the perennial societal experiments for reaching this fundamental goal, the article outlines the major dilemmas of closing down the East–West divide in our times, and gives an overview of contemporary social science research in reflecting and assisting these attempts.

1. *Eastern Europe: A Region with Clear Contours, but Opaque Borders*

Customarily, Eastern Europe as a geographical concept is understood to denote the region positioned east from the River Elbe and west from the Ural Mountains. Such a geographical definition meets, however, insurmountable difficulties when translated into societies, nations, and states. As to the western border of the region, the River Elbe runs through Germany. Does this mean that German society is sharply divided into a 'Western European' and an 'Eastern European' segment? Similarly, the eastern border of the region—the Urals—cut through Russia; does this mean that one half of Russia belongs to 'Eastern Europe,' while the other is part of the 'non-European East'? Further, if such faultlines divide one or more countries of the region, what are the implications of their presence: do they simply carry sociocultural differences of historical and ethnographical interest, or, does their persistence endanger the stability of the affected states, and, beyond, that of the continent itself? In short: if not simply a tool of geographical orientation, what then is the function of making the East–West differentiation within Europe?

As is well known, certain sociocultural differences on regional grounds can be observed in all societies. Still, despite the remarkable differences between Northern Italy and Sicily, nobody in our times would question Italian unity, and the belonging of the entire country to the core of the West. Likewise, even the linguistic division of countries like Belgium or Switzerland would not inspire anybody to doubt the integrity and genuinely Western character of these societies. Thus, the ambiguities in identifying the borders between Western and Eastern Europe arise from deeper layers of differentiation. Their essence is neither a neutral geographical classification according to the magnetic poles, nor a definitive and value-loaded gesture of clearly circumscribing 'togetherness' as opposed to 'otherness.' The concept of 'Eastern Europe' represents a mixture of the latter two, and carries the inherent ambivalence of such a combination. With the very same momentum, it signals undivided 'Europeanness,' and also some reservations toward it. Although the societies embraced by the concept all are 'European,' they do not come up 'fully' to the standard of what 'Europeanness' implies. In this ambivalent classification, the implicit norm is the West, and Eastern Europe remains as much Eastern, as it maintains its 'other than Western' character. Thus, the concept has always been deeply imbued with politics. On the part of Eastern Europeans, its inherent political character has manifested itself in continuous attempts to close the East–West divide through adaptation to the Western norms of a given time; on the part of the West, the same politicized attitude expressed itself in the interest-driven recognition/refusal of the Eastern attempts, having immediate consequences on the (re)drawing of the invisible political, economic, and military map of the continent of Europe.

The doubtfulness of the borders—both in the harsh physical and the more subtle sociopsychological sense of the term—has inspired recurrent attempts to refine the geographically founded geopolitical classifications by denoting important and 'meaningful' subunits within the region (Todorova 1997). Probably the most important of such attempts has been the repeated separation of 'Central Europe' from the rest of 'Eastern Europe.' The concept of Central Europe as a more 'Westernized' entity with coherent sociopolitical and cultural meanings on its own appeared mainly in nineteenth/twentieth-century public debates by writers, political thinkers, and philosophers (Kundera 1984), but attempts at the political realization of togetherness and simultaneous separation from the rest of the region's nations also came to the surface among the goals of the sweeping political movements of 1848, 1918, and the late 1980s.

However, the ideas of 'Westernizing' Eastern Europe certainly are not the fresh production of modern times. Various grandiose experiments for their realization date back as far as the time after the fall of the Roman Empire. Since those distant days, the history of the region has been accompanied by heated debates over the 'belonging-issue,' which, in turn, has gradually become a decisive aspect of the self-perception and identity struggle of the people living on its territory. The East–West divide, despite its opaque character, has proved extremely strong during the past 1000 years of European history, serving a variety of social, political, military, and economic interests both within and between the countries of the European continent. With this historically elaborated comprehensive political meaning of the concept, 'Eastern Europe' has been as much a point of departure for Western self-identification as for the shaping of programs to combat those 'historic illnesses' which have hindered certain societies from becoming parts of Europe proper. In other words, up to the time of writing, 'Eastern Europe' has been considered the embodiment of the major historical failures of the European continent. It has symbolized the stone-hard limitations of creating a Europe without any subdivisions, and has also represented the accompanying feelings of discomfort and frustration. In the

course of the century-long failed efforts, these feelings have grown to become self-perpetuating mutual prejudices, which have exhibited an ever stronger influence on shaping politics of the East–West cohabitation. The persistence of these prejudices cannot be ignored when attempting to understand the difficulties that the current program of European integration (see Sect. 3) has to face in the course of its day-to-day realization.

2. *The Historical Roots of Eastern Europe's One-sided Development*

Historians, political thinkers, writers and philosophers seem to be in full agreement on identifying the emergence of 'the three historical regions of Europe' (Szücs 1988) in the aftermath of the Barbarian conquest against the late Roman Empire, which resulted in the division of the occupied territories between Europa Occidens, Byzantium, and a distinct transitional territory in between. This latter region—the historical forerunner of Eastern Europe—occupied a halfway position in all senses of the term. It was neither 'Western,' nor 'Eastern,' but an ever-changing mixture of the two. First, the region's geographic placement exposed it to frequent attacks from both of the neighboring big powers, each of whom regarded themselves the genuine heirs of the late Roman Empire—the Frank Empire and Constantinople. Second, the shaping of the region's decisive religions, traditions, and ethos was put also under dual pressure from the West and the East alike. As pointed out by the influential work of medievalists and historians of political ideas (Gerschenkron 1962, Said 1978, Szücs 1988, Wolff 1994), although the people of Eastern Europe took up Christianity early enough in their history, but the subsequent lasting Mongol, Ottoman, and Russian invasions revitalized a number of 'Eastern' features in their social relations, institutions, faith, and the running of their daily life. Third, as it has been demonstrated amply by recent studies in comparative history and historical anthropology (Gerschenkron 1962, Said 1978, Konrád and Szelényi 1979), the states of the region as potential actors of modernization played an ambivalent role in the premodern era: Though the late Medieval centuries brought about absolutist reigns similar to those in the West of the time, the actual function of the Eastern European absolutist rule was to reinforce the estate-based power of the landed nobility, while their Western counterparts prepared the soil for the take-off of capitalism, democracy and modern bourgeois development. The repeated events of military occupations, wars, mass uprisings, and the subsequent hurried efforts at reconstruction recurrently blockaded the crystallization of fundamental institutions of statehood and administration. Instead, these events usually concluded in hasty and forceful attempts of the conqueror to employ central power to eliminate the 'historic losses,' at all social and political cost. Such attempts at forced and inorganic modernization, however, caused lasting deformations in the sociopolitical order that have had a far-reaching impact on the development of the affected societies. In assessing the potential and limitations of corrective societal actions, the most dramatic consequences shaping both politics and societies in modern times were discussed extensively by sociologists and political scientists in the 1980s and 1990s (Kochanowicz 1997, Garton Ash 1993, Arato 2000). According to the virtually unanimous conclusions of these studies, the historically informed and persistent deformations of Eastern European societies can be summarized in three main features which clearly distinguish them from their contemporaneous Western counterparts. These are: the one-sided development of nationhood and national identity (which led easily to repeated waves of radical populism and irredentist nationalism); the underdeveloped nature of civil society (which recurrently has provided natural justification for the states' inclination to gain oppressive power over people's needs); and, finally, the periodic disappearance, at best, weak institutionalization of democratic governance (which has been the cause of loose public control over administration and economic management, and which also can be seen as being responsible for the lasting symptoms of social disintegration in the region).

These historic deformations became sources of dangerous political conflicts around the beginning of the twentieth century. After the failed revolutions of 1848 which, for a century to come, represented the last heroic attempts of the nations of the region to (re)gain national independence and create their state which would guard civil rights, law and order, and free capitalist development, the societies of Eastern Europe faced exclusion from Europe proper, and reacted accordingly—in a frustrated and hostile manner. As historical research has demonstrated convincingly (Konrád and Szelényi 1979, Schöpflin and Wood 1989), the outbreak of World War I cannot be explained without considering the preceding unhealthy nationalist rivalry within the borders of the increasingly ineffective, but still oppressive Austro-Hungarian monarchy, and the deep tensions between the cultural–political formations of the communities and their governing states in all the surrounding societies. Again, the outbreak of World War II cannot be explained merely in terms of the clashing political and economic interests of the leading powers of the European continent; the picture remains disturbingly incomplete without taking into account the extremist aspirations of Hitler's Germany and its allies to find a 'final solution' for the century-long diseases of confused cultural and national identity, and also without considering their deeply embedded attempts to claim new territories and borders for their historically 'stolen' nation states (Bibó 1991). However, not only

the two world wars, but also the 'invention' of the two, diametrically contrasting ways of totalitarian ruling originate from the same roots. As argued in Arendt's and Bauman's studies that have become classics of postwar social science, one can find in their creation the same forms of traumatized, oblique, frustrated, and confused nationalism, and the unstable institutions built on them (Arendt 1958, Bauman 1989). While the Nazi way of totalitarian rein put the superiority of race and nationhood at the center of its legitimizing ideology, communist totalitarianism—with its aim of radically transcending the past—denied the role of nationhood as such, and attempted to substitute it with values of internationalism and enforced equality. However, despite the viciously different responses of the two forms of totalitarianism, the emotionally heated attitudes toward nationhood and national identity, the conflictive relationship between the state and people's spontaneously formed communities, the weak observance of rights and law, and, consequently, the enduring harm done to individual freedom, remained key problems of Eastern European societies until the end of the twentieth century. True, the modernization of the economies and societies of the region, and the ultimate closing down of the East–West divide belonged to the core of the communist program of the four decades following World War II. However, the domino-like collapse of the Soviet-type regimes in 1989–90 testified sharply to the failures of these attempts, and drove back the societies in question towards the lost track of classical capitalism. Perhaps the principal evidence of the failures is the fact that, with the sweeping events of the late 1980s, the allegedly abolished problems of the interwar years have come to the surface with dramatic intensity: Widespread poverty, intense racism, extreme social marginalization, sharp inequalities of income and wealth, pervasive corruption, and weak legal protection of defenseless communities and individuals are among the most disturbing troubles of Eastern European societies at the time of writing. In order to understand the unfolding of these problems, a brief examination has to be made of the legacy that the communist systems of Eastern Europe have left behind.

3. *Eastern European Societies: From Where to Where?*

As a number of recent studies in social and political history have shown, despite the steps taken throughout the 1970s and 1980s toward economic and political liberalization in different countries of the region, the fundamental features of the communist system of Eastern Europe remained unchanged until the late 1980s (Kornai 1991, Garton Ash 1993, Arato 2000). Backed by the military, political, and economic monopoly of the all-embracing Soviet force in the region, the communist party-states of the affected countries maintained their power over all the segments of the economy and society, including people's most private matters of career choice, marriage, child-bearing, daily consumption, or cultural taste. Aside from the tools of enforcement solely in the hands of the ruling communist parties, the bases for such an extraordinarily extended power were provided by two major deeds: the concentration of 'people's property' in the hands of the party-ruled state, and the institutionalization of the day-to-day control over society through compulsory full employment. As comprehensive studies on the working of the socialist economy have demonstrated (Kornai 1991, Sik 1991), the concentration of all the means of production was intended to induce economic rationality through the centrally prescribed production commands driven by the Plan, while compulsory employment was seen as the base for fair shares from distribution and, simultaneously, as an effective means to combat poverty. However, such a structure turned against the initial goals rather quickly: from the early 1960s onwards, the Eastern European communist regimes, one after the other, had to face recurrent deep economic crises, and an increase in social tension originating from the 'dictatorship over needs' (Fehér et al. 1983) that led to nothing else but the equality of poverty on an unparalleled scale. Thus, the mere maintenance of the regimes required certain concessions: the most important among them were the rehabilitation of the family as a legitimate sphere of people's undisturbed privacy, and the tacit acceptance of the slow evolution of a second, informal economy.

However, these concessions proved powerful tools in societies' hands slowly to erode, then effectively to decompose the oppressive structures above them. Eastern European sociological, social-historical and economic research of the 1970s and 1980s provided a rich literature of these multifaceted processes which, in retrospect, can be regarded as the direct ascendants of the radical social upheavals after 1989–90 (Vajda 1988, Staniszkis 1991, Szalai 1991, Kochanowicz 1997). The move away from classical socialism was unstoppable. First, ever-spreading participation in the informal economy gradually took a lead in determining the standard and forms of consumption for substantial layers of society; thus, the tacit acceptance of people's informal work and their partial withdrawal from party control, even in the strictly regulated socialist workplaces, were also in the immediate interest of those in power—otherwise, the communist parties would have risked unmanageable crises in production and consumption that could have led to the swift fading of the justification of their 'blessing rule.'

Second, the family-based framework of the informal economy provided a forum for learning a number of market-related skills and thus assisting the underground continuation of the interrupted embourgeoise-

ment process. Such an underground preservation of officially denied values, attitudes, skills, and socio-economic arrangements later proved crucial in the post-1989 systemic changes of returning to a capitalist/bourgeois track. Third, the informal economies of Eastern Europe complemented the ever-impoverishing state-run welfare provisions with community-based social services, and with slowly elaborated forms of secondary redistribution. Thus, they had a role in correcting the inadequacies of the distribution of incomes driven from the 'first' economy, and also in developing those market-conforming forms of social protection, which then, effectively, assisted the transition process amid the rapid withering away of state-distribution some decades later.

In short, the informal economy and the social relations around it gradually became a second—mostly hidden—regulator in the functioning of the communist regimes, and an important second pillar in organizing the daily life, well-being, aspirations, and life-careers of substantial groups of Eastern European societies.

However, the emergence of this second system had also serious deficiencies. As sociological research on the malfunctions of the 'socialist' welfare states has evidenced it (Deacon-Szalai 1990, Kochanowicz 1997, Atal 1999, Lessenich 2000), those who, for a range of varied reasons, could not participate, became increasingly dependent on the state that gradually intensified their defenselessness, instead of mitigating it. With the spreading and tacit acceptance of the regulatory role of the informal economy, the nonparticipating groups were designated as 'misusers' of the public resources, and cuts in assistance for them were approved on a mass scale. Thus, the spreading of participation in the informal economy gave rise to serious social divisions which, amid the troublesome economic conditions of the post-1989 years, eventually became an open war between the 'nondiligent' poor and the 'industrious' nonpoor.

A further shortcoming of the dual system was its weak institutionalization: any claims for the formal acknowledgement of informal work-collectives, service-providers, interest-representations, and the like were seen as open attacks against the principles of the prevailing order, and were treated as political crimes. On these grounds, the informal economy became a battlefield for freedom, but at the same time, the political connotations of participation further hindered its institutionalization. The negative consequences of such a development came to light mainly after the collapse of the old rule. As in-depth studies of the democratic potentials of Eastern Europe have pointed out convincingly (Staniszkis 1991, Berend 1993, Arato 2000), despite widespread earlier practices, civil society has proven to be extremely weak, and the new system hardly could count with its provisions, protective associations, and community-run services amid the deep 'transitional crisis' of the 1990s.

The weakness of civil society has given rise to serious crisis phenomena amid the two intertwined major processes of postcommunist transition: the decomposition of the formerly unlimited power of the central states, and the privatization of the earlier anonymous 'people's properties.' Paradoxically, the earlier resistance to the state as oppressor is now quite clearly being reversed, and while the various corporate bodies and interest alliances only clung to the state distribution policy out of fear and defenselessness, they now make angry claims on it. Behind the opposing principles of privatization intended to 'regulate' the plundering of public assets, one finds intensely competing demands for compensation. Having the arguments justifying these claims accepted and embodied in the legislation is a question of raw political force. As comparative economic studies in the region have demonstrated with ample empirical evidence (Róna-Tas 1994, Stark 1996), in this way, privatization and the creation of a bourgeoisie is the direct function of the latent bargaining positions established over past decades. Hence, the process of privatization has led to a jump in uncontrolled and unregulated inequalities of income and wealth, and to the emergence of a new. prosperous elite with little legitimization. At the same time, the process has produced masses of victims at the other end of the scale. Besides depriving them of any access to the newly distributed properties, the easiest compromise to find resources to meet the hunger of the vocal groups has been to cut back sharply on state expenditures on welfare and social services. With these steps, however, Eastern European transition societies have produced vast numbers of chronic poor, whose fate has been extreme social marginalization. Thus, despite the initial hopes of 1989–90, massive poverty and the accompanying syndromes of dramatic social disintegration have not disappeared with the recovery from the crises of the early 1990s, but have become the decisive indicators of a still non-Westernized path of current Eastern European development. The conflicts that arise from this situation have further consequences. The still excessive presence of the state in most of the region's economies, as well as the understandable open aspiration of the earlier overblown state bureaucracy to reinforce its own position of power, at most only help broad strata of societies to reach the recognition that the creation of a private market in the Western sense is a 'foreign' prescription for the process of embourgeoisement in the region—their 'genuine' path still leads through nationalized modernization. Thus, it is hardly surprising that Western and Eastern economists, sociologists, and political scientists analyzing the major sociopolitical processes of postcommunist transition arrived at largely consensual conclusions by the late 1990s (Kochanowicz 1997, Standing 1997, Arato 2000). There seems to be a general agreement that the spreading of anti-West feelings provides a fertile soil

for heated nationalism, aggressive racism, and the re-emergence of authoritarian tendencies; hence, these recent tendencies in the changes of popular attitudes endanger seriously the once vivid hopes of creating a Europe without boundaries. However, these unanimous warnings have little impact on actual politics: at best, social science reflects and informs the ongoing attempts of European integration, but can hardly influence it.

4. *Eastern Europe's (Re)turn to Europe: As Social Scientists See It*

With the collapse of communism in 1989–90, both Western and Eastern social scientists investigating Eastern Europe had to face a number of new challenges. It was not only the 'subject' of their research which suddenly changed its framework, but the formerly prevailing concepts and methods of scientific analysis had also to be reevaluated. Such a simultaneous change of 'subject' and 'approach' is a rather rare development in the history of the social sciences. However, the collapse of communism swept away a sociopolitical reality together with its prevailing ideological reflection: the death of communism put an end to the scientific lead of Marxism, and forced this powerful worldview into the background of the history of philosophy. Marxism as a frame of reference ceased to function as an adequate and applicable conceptual model for postcommunist scientific research.

Among the multifaceted consequences of these changes—that have affected as much the structure and content of higher education, as the composition of the academic community, or the financing of research and its accessibility to the broader public—there are two major aspects of the profound renewal of scholarly work that have to be underlined here. First, as an immediate outcome of the fall of communism, certain areas of research once so important in characterizing Eastern European social science have disappeared completely, while formerly nonexistent approaches and strands of thinking have gained substantial influence. It is enough to refer here to the utter disappearance of 'scientific socialism,' or 'socialist political economy' on the one hand, and the impressive rise of 'political science,' 'constitutional studies,' or 'microeconomics,' on the other. Second, due to the radical switch in the politics of science, one can record a remarkable shift in the topical composition of research, together with a fundamental institutional reorganization of academic work. Due to the preceding monopoly of Marxism as the official stand of the social sciences in Eastern Europe, previously suppressed intellectual strands demanded their well-deserved institutional acknowledgement after 1989–90, bringing to the forefront the 'underground' culture of formerly nonaccessible research on 'forbidden' subjects such as the history of the Gulags, the persistence and reproduction of poverty, the critical analysis of the planned economy, or the sociology of the hidden embourgeoisement process in societies under Soviet rule. With the now formal acknowledgement of these areas of research in the 1990s, marked shifts of emphases as well as an increased pluralism of approaches to issues at the core of social science interest were registered. However, one has to remark: these developments signal more a 'back to normal' state of affairs in research than the appearance of a genuinely new voice that one could call 'Eastern European social science.' It is better to characterize the situation as Eastern Europe's reintegration to the universal world of the academia. True, the paths to this scholarly universe have had markedly different histories in the West and the East. As to the East, the potentials of the current surprisingly quick reintegration have to be identified in the widespread underground movements of 'flying universities,' samizdat publications, and secret mass-events of the so-called 'secondary public' under communism, which not only formed the bases of strong national communities of social scientists, but also embraced a great number of cooperating groups of dissident intellectuals from Moscow to Warsaw, and Prague to Sofia. After the lifting of the political barriers, the preceding long history of such extensive underground experimentation and the simultaneous preservation of the critical stand of social science have provided a rich foundation to a strong theoretical orientation, as well as to the internationally acknowledged high qualities of empirical research in Eastern Europe.

Given the major aim of the postcommunist transition in closing down the East–West divide within the shortest possible timespan, it is understandable that the bulk of social science research in and about Eastern Europe has addressed issues of the greatest political importance of the time: marketization, democratization, issues of institution-building, and Eastern Europe's integration into an undivided Europe are in the forefront of the ongoing investigations. Following predominantly the theory of neoliberal economics, research on the changes in property relations, the institutional conditions of transition from the command ruled economy to a market regulated one, the creation of a banking system from scratch, and Eastern Europe's participation in the processes of globalization have served mainly to provide assistance to the newborn governments in the region, to orchestrate the difficult processes of transformation with the least possible social and political conflict. With such a shift toward applied research, little has yet been done, however, toward the establishment of an original economic theory which would provide ample concepts and explanations to one of the most disturbing puzzles of the late 1990s: though the neoliberal prescriptions of transformation have been followed in practically all of the 28 countries of the former Soviet empire,

their markets rarely resemble those of the Western societies.

The dominance of descriptive empirical studies and applied research is perhaps even more characteristic in other areas of the social sciences. True, the relatively unstable political processes in the region require a close follow-up of party affiliations and social movements, and a regular monitoring of people's political attitudes. Similarly, the success of institution building has to be measured against people's use/misuse of the new institutions. The list of examples can be followed with equally justifiable descriptive studies on the restructuring of people's value orientations and daily consumption, or their evaluation of the old–new institutions of public administration. Although one can hardly question the usefulness of any such research from the point of view of administrative and political decision making, as at the time of writing, it has made little contribution either to the building of a richer theory of capitalist societies, or to the deeper understanding of the still prevailing marked differences between the two halves of Europe.

Some steps toward a more thorough analytical examination of the peculiarities of contemporary Eastern Europe have been made, however, in a few novel comparative studies on state–society relationships, the characteristics of the welfare states, and the features of 'old' and 'new' poverty in both the West and the East. These studies have been initiated mainly by huge international organizations (ILO, OECD, the World Bank, UNESCO), and were run on principles of equal participation by Eastern and Western European scholars. Though it is perhaps too early to arrive at definitive conclusions, the lessons of these recent studies indicate unanimously the persistence of a sharp East–West divide, much in accordance with the sociocultural faultlines that have been explored by the historical studies on earlier centuries. It seems that in their post-1989 development, Eastern European societies have taken a lead toward increasing social disintegration between their adaptive majorities and the extremely marginalized minorities in long-term poverty. The states of these new democracies are squeezed between the contradictory challenges of accomplishing marketization and combating social disintegration at the same time; due to the weakness of their civil societies, they have postponed repeatedly the addressing of the latter problem by creating the required strong welfare systems. As a result, the current social map of Eastern Europe resembles more some Latin American societies than those of the neighboring Western Europe.

5. *Summary*

Though meant also as a geographical classification for denoting the region between the River Elbe and the Ural Mountains, the concept of 'Eastern Europe' is used mainly to express certain historically informed sociocultural differences within the European continent. It carries heavy layers of ambivalence: although the societies embraced by it all are regarded as 'European,' they do not come up 'fully' to the standard of what 'Europeanness' implies.

Such historically informed ambivalence characterizes deeply as well as the current efforts to unite Europe. Besides, they leave their mark on the internal social development of the region's societies, and also have serious consequences for the endurance of their post-1989 as the crisis.

Social science research of the past decades has centered around the historical origins and perpetuating factors of the East–West divide. Nevertheless, following the collapse of communism, it has turned more toward investigating the potentials and limitations of closing it down. The necessary shift toward the dominance of descriptive research in assistance of the politics of integration has slowed down, however, the in-depth exploration of those aspects of recent Eastern European development which carry the potential of repeated diversion from the Western path.

See also: Area and International Studies: Development in Eastern Europe; Communism; Democratic Transitions; Eastern European Studies: Culture; Eastern European Studies: Economics; Eastern European Studies: History; Eastern European Studies: Politics; Revolutions of 1989–90 in Eastern Central Europe; Socialist Societies: Anthropological Aspects

Bibliography

Arato A 2000 *Civil Society, Constitution and Legitimacy*. Rowman & Littlefield, Lanham, MD

Arendt H 1958 *The Origins of Totalitarianism*. Meridian, New York

Atal Y (ed.) 1998 *Poverty in Transition and Transition in Poverty*. Berghahn, New York

Berend I 1993 *Alternatives of Transition?* Center for German and European Studies, University of California, Berkeley, CA

Bibó I 1991 *Democracy, Revolution, Self-determination: Selected Writings*. Atlantic Research and Publications, Columbia University Press, Boulder, CO

Bauman Z 1989 *Modernity and the Holocaust*. Cornell University Press, Ithaca, NY

Deacon B, Szalai J (eds.) 1990 *Social Policy in the New Eastern Europe*. Avebury, Aldershot, UK

Fehér F, Heller Á, Márkus G 1983 *Dictatorship over Needs*. Basil Blackwell, London

Garton Ash T 1993 *The Magic Latern: The Revolution of '89 Witnessed in Warsaw, Budapest, Berlin Prague*. Vintage Books, New York

Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Harvard University Press, Cambridge, MA

Kochanowicz J 1997 Incomplete demise: Reflections on the welfare state in Poland after communism. *Social Research* **64**: 1445–71

Konrád G, Szelényi I 1979 *The Intellectuals on the Road to Class Power*. Harcourt Brace Jovanovich, New York

Kornai J 1991 *Anti-equilibrium: On Economic Systems Theory and the Tasks of Research*. A. M. Kelley, Fairfield, NJ
Kundera M 1984 The tragedy of Central Europe. *New York Review of Books*, 26 April
Lessenich S 2000 The southern image revisited: The dynamics of 'transformation dynamics' in East Central Europe. *ECE/ECE* **1**: 21–37
Milanovic B 1998 *Income, Inequality and Poverty during the Transition from Planned to Market Economy*. World Bank, Washington
Róna-Tas Á 1994 The first shall be last? Entrepreneurship and communist cadres in the transition from socialism. *American Journal of Sociology* **100**(1): 40–69
Said E W 1978 *Orientalism*. Pantheon Books, New York
Schöpflin G, Wood N (eds.) 1989 *In Search of Central Europe*. Barnes & Noble, Totowa, NJ
Sik O (ed.) 1991 *Socialism Today? The Changing Meaning of Socialism*. St. Martin's Press, New York
Standing G 1997 The folly of social safety nets: Why basic income is needed in Eastern Europe. *Social Research* **64**: 1339–79
Staniszkis J 1991 *The Dynamics of the Breakthrough in Eastern Europe: The Polish Experience*. University of California Press, Berkeley, CA
Stark D 1996 Recombinant property in East European capitalism. *American Journal of Sociology* **101**(4): 993–1027
Szalai J 1991 Some aspects of the changing situation of women in Hungary. *Signs* **17**(1): 152–70
Szücs J 1988 Three historical regions of Europe. In: Keane J (ed.) *Civil Society and the State*. Verso, London
Todorova M N 1997 *Imagining the Balkans*. Oxford University Press, New York
Vajda M 1988 East-Central European perspectives. In: Keane J (ed.) *Civil Society and the State*. Verso, London
Wolff L 1994 *Inventing Eastern Europe. The Map of Civilization on the Mind of the Enlightenment*. Stanford University Press, Stanford, CA

J. Szalai

Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder

Two major eating disorders have been defined and operationalized: anorexia nervosa (AN) and bulimia nervosa (BN). In recent years preliminary definitions have been given for a third eating disorder: binge eating disorder (BED). Anorexic and bulimic eating disorders usually begin in adolescence and young adulthood and more than 90 percent of the cases of AN and BN are female. Eating and food intake can have very different functions: eating fulfills the biological necessity of keeping us alive, but it can also be a source of pleasure and joy. Food intake plays an important role when we relax or celebrate with other human beings (festivities). For a person who hungers unvoluntarily this can be bothersome or painful. On the other hand, an anorexic patient who has achieved reducing weight considerably to a pathological level will be proud of the result. Refusal to eat on the level of society can be used as a means of political pressure (hunger strike). Within a family system the refusal to eat of an anorexic girl can—in the lack of other possibilities of expression—be an attempt to find and define herself; it can be a sign of inner resistance and defense against the intrusion of others. The anorexic patient avoids confrontation with the increasing demands of the role of an adult person (sexuality, relationships, profession, achievement) and regresses to an earlier more infantile stage. Eating too much as well as eating too little can be damaging to mental and physical health.

1. Clinical Features (Diagnoses and Differential Diagnoses)

1.1 Anorexia Nervosa

Anorexia nervosa is characterized by a refusal to maintain body weight at or above a minimally normal weight for age and height, an intense fear of gaining weight or becoming fat even though underweight, a disturbance in the way in which one's body weight or shape is experienced with undue influence of body weight or shape on self-evaluation. Frequently the seriousness of current low body weight is denied or hidden. In postmenarcheal females, amenorrhoea is the fourth diagnosic criteria; its usefulness has, however, been questioned by some research groups; anenorrhoea largely is secondary to low body weight and the starvation syndrome. Laboratory tests are usually of little help for diagnosing AN. Diagnosis is made primarily on the basis of a thorough psychiatric examination in which it is a task of the examiner to find out the motives behind being underweight. According to the *DSM-IV* criteria of the American Psychiatric Association (1994) two subtypes of AN can be distinguished: (a) the restricting type and (b) the binge eating/purging type.

AN is frequently associated with social withdrawal, irritability, insomnia, depressed mood, and diminished interest in sexuality. In part, these depressive features are secondary to starvation and disappear with weight restoration. Preoccupation with thoughts about food, collecting recipes, and hoarding food are frequently seen in AN patients. In more than a third of AN cases obsessive compulsive features unrelated to food and eating are also present. Associated with AN can also be an exaggerated need to control one's environment, concerns about eating in public, feelings of ineffectiveness, inflexible black and white (dichotomous) thinking, an overly restrained initiative, and emotional expression. AN patients of the binge eating/purging type (in comparison to restricting anorexics) more frequently have problems in impulse-

Table 1
Laboratory findings and metabolic changes in underweight anorexia nervosa patients

Leucopenia with relative lymphocystosis[c]
Anaemia[c]
Thrombocytopenia[c]
Hypercholesterolemia[c]
Elevated liver function test results[c]
Hyperamylasemia[c]
Hypokaliemia[a]
Hypophosphatemia
Hypochloremia[a]
Hypozincemia
Hypomagnesemia
Metabolic alcalosis (elevated serum bicarbonate)[a]
Metabolic acidosis[b]
Tri-jodothyronine decreased[c]
Hyperadrenocorticism and unsufficiently suppressed DST[c]
Regression of hypothalamo-pituitary gonadal axis[c]
Reduced resting energy expenditure[c]

[a]Resulting from vomiting. [b]Resulting from laxative abuse. [c]Resulting from starvation.

control, abuse of alcohol or other drugs, lability of mood, and they are sexually more active.

Starvation and/or inappropriate compensatory behaviors (vomiting, use of laxatives, etc.) have a variety of effects on most organ systems. Table 1 lists a number of laboratory findings usually found in starving AN patients. Except for hypokaliemia, hypochloremia, and metabolic alcalosis which are secondary to vomiting and metabolic acidosis resulting from laxative abuse, most laboratory findings are secondary to the state of starvation and the peculiar eating behaviors and eating patterns. Clinically the reduced activity of the sympathetic nervous system goes hand in hand with bradycardia, the development of fine downy body hair (lanugo), low blood pressure, constipation, cold intolerance, petechiae in the extremities, and peripheral edema during weight restoration.

1.1.1 Differential diagnosis. Occasionally general medical conditions (gastrointestinal disease, occult malignancies, brain tumors, unclear infectious diseases) associated with weight loss are misdiagnosed as AN. When diagnosing AN it is important not to focus primarily on the emaciated state but rather on the psychological features explicated in the diagnostic criteria for AN. For the physician it is important to search for a psychological motive for reducing weight. In psychiatric disorders such as depression or schizophrenia, weight loss may occur, which, however, is secondary and not resulting from a drive for thinness as in AN.

Compulsive disorder, body dysmorphic disorder, and social phobia concur with AN not infrequently. AN of the binge eating/purging type can be very similar to BN and the differentiation is (rather artificially) based on current body weight.

See *Anxiety and Anxiety Disorders*; *Obsessive–Compulsive Disorder*.

1.2 Bulimia Nervosa

Bulimia nervosa (BN) also occurs like AN predominantly in adolescent girls or young women. Its essential features are binge eating and inappropriate compensatory behaviors to prevent weight gain. In these patients usually self-evaluation is excessively influenced by body shape and weight. According to the *DSM-IV* criteria, a binge is defined on objective grounds (rather than subjective) as 'eating in a discrete period of time an amount of food that is definitely larger than most individuals would eat under similar circumstances.' Continuous snacking (eating small amounts of food throughout the day) which may also occur would not qualify as a typical (objectively defined) binge. Diagnosis of BN according to *DSM-IV* criteria requires that binge eating and inappropriate compensatory behaviors both occur on average at least twice weekly for three months.

Associated laboratory findings are very similar to those reported for binge eating/purging AN. Fluid and electrolyte abnormalities (hypokaliemia, hyponatremia, hypochloremia), metabolic alcalosis with elevated serumbicarbonate resulting from vomiting or metabolic acidosis from laxative abuse, dentrocavities, enlargement of the parotid glands, and rarely but potentially fatal esophageal tears, gastric rupture, and cardiac arrhythmias.

1.2.1 Differential diagnosis of BN. Disturbed eating behavior can be present in certain neurological or general medical conditions such as the Kline–Levin syndrome; however, certain key features of BN such as overconcern with body shape and weight are lacking. Also, a multi-impulsive form of BN has been described in which several impulsive behaviors are observed (impulsive stealing, impulsive drinking, impulsive suicidal acts, impulsive self-damaging behavior). Prognosis for this subgroup seems to be worse than for BN in general.

1.3 Binge Eating Disorder

Binge eating disorder (BED) is defined as being very similar but not identical to nonpurging BN. Main features of BED are binge eating at least twice weekly over a period of six months, associated with marked distress. However, in BED no inappropriate compensatory behavior to prevent weight gain, such as self-induced vomiting, use of laxatives, diuretics or other medication, is present. Many but not all BED patients are overweight. Overweight BED patients have the usual risks associated with obesity (hyper-

tension, orthopedic problems, diabetes, etc.). They do not have the risks of purging behavior observed in binge eating/purging AN and BN (electrolyte imbalance, alcalosis/acidosis from laxative abuse, sialadenosis from vomiting, severe dental problems from vomiting, renal failure). New research indicate that BED has many similarities with BN but that BED is quite dissimilar from AN. Very few anorexics later develop BED and very few BED patients later develop AN.

As in BN, neurological and other general medical conditions associated with bingeing and weight gain must be differentiated from BED.

1.4 Eating Disorders—Not Otherwise Specified

The eating disorders—not otherwise specified (EDNOS) category is for disorders of eating that do not meet the criteria of any specific eating disorder such as AN, BN, and BED. In *DSM-IV* six examples are given. Usually one criteria of a major eating disorder (AN, BN) is not met. One of several examples for EDNOS is: all of the criteria of BN are met, except that the binge eating and inappropriate compensatory mechanisms occur at a frequency of less than twice a week and for a duration of less than three months.

2. Epidemiology, Course and Prognosis

2.1 Anorexia Nervosa

A number of studies indicate that the prevalence of AN has increased over the decades of the twentieth century—especially since the 1960s. Research among females in late adolescence and early adulthood have revealed rates of 0.4–1.1 percent for presentations that meet full criteria for AN according to *DSM-IV*. Less than 10 percent of all treated cases with AN are male. Mean age at onset for AN is 16–18 years. Onset very rarely occurs in individuals over age 40 years. The onset of illness is frequently triggered by life events which increase uncertainty, cause anxieties, and put new demands on the individual. The mortality rate of patients with AN is only second to that of politoxicomanic drug use and dependence and it is much higher than for schizophrenia or depression. Over the long-term course of 20 years, 10–15 percent of patients with AN die from starvation, suicide, or electrolyte imbalance (which can cause cardiac or renal failure (Ratnasuriya et al. 1991). First-degree relatives of patients with AN have an increased risk for that disorder. Mood disorders have also been found to be more prevalent in the biological relatives of individuals with (binge eating/purging type) AN.

2.2 Bulimia Nervosa

Among adolescent and young adult females the prevalence of BN is about 0.8–3 percent. As in AN, the disorder is usually found in pubertal girls or young adult women; less than 10 percent of the cases of BN are men. The onset of BN is usually in late adolescence and early adult life. Although patients with BN at times can be very tough to treat, the course over a time is not as malignant as it can sometimes be in AN. Frequent and excessive dieting may trigger the start of binge eating. In a high percentage of clinical samples the disturbed bulimic eating behavior persists over several years and the course can be chronic or intermittent or with remission. In the last few years several studies have reported on the six- to 10-year course of BN (Fichter and Quadflieg 1997a). The outlook for BN patients is not quite as pessimistic as for AN patients: although about a third of treated cases show a more or less chronic course; the majority at six-year follow-up no longer fulfill the criteria for a major eating disorder. This, however, does not mean that all are free of eating disorder symptoms which in some way interfere with their everyday life.

2.3 Binge Eating Disorder

The prevalence of BED is roughly the same as for BN (1.5 percent among adolescent and young adult females). Although BED is predominantly found in females, the proportion of males is higher than in AN or BN. The short- and medium-range course of BED appears to be equivalent to that of BN (Fichter et al. 1998). Research indicates that besides BED there are other eating disorders associated with overweight not fulfilling criteria for BN or BED. Psychiatrists have carried out very little research on obese patients with atypical binges (continuous snacking throughout the day) and mental problems resulting from getting obese on the basis of a general medical condition (endocrinological disturbance) or medication which increases appetite and body weight (neuroleptics, certain antidepressants, lithium, etc.).

3. Etiology and Pathogenesis

For the pathogenesis of anorectic and bulimic eating disorders the following factors must be taken into consideration: (a) sociocultural factors (mediated through family, school, and mass media), (b) biological factors (genetic, neurochemic), and (c) life events and chronic difficulties.

See *Mental Illness, Etiology of.*

3.1 Sociocultural Factors

Eating disorders are fairly prevalent among young women in industrialized countries where the ideal of body thinness exerts considerable pressure especially on adolescent girls and young women. Most young

women have dieted once, several times, or many times—some successful, many not. In industrialized countries food exists in abundance. Studies in Third World countries have shown that anorexic and bulimic eating disorders are seen rarely there and if they occur the patient was usually from a well-to-do family. Among persons who migrated from a Third World country to an industrialized country and who were thus exposed to ideals and pressures to be slim, anorexic and bulimic eating disorders are seen more frequently than among those who remained in the country of origin.

The ideals of body slimness (especially for women) and the ideal of body fitness (especially for men) are very prevalent in industrialized countries and exert pressures on individuals. We do not know exactly why these ideals exist and how they came about but some interesting explanations exist. From an anthropological–evolutionary perspective, it is important to note that mankind never in its history has had an abundance of food in such a large proportion of the population over such a long period of time. During the evolution of mankind, food at times was scarce and our predecessors would not have survived if they had not developed special biological adaptations to periods of hunger and starvation. When we reduce our food intake and lose body weight many adaptive physiological adaptations take place which help us save energy (bradycardia, hypotension, lowering of T3, hypercortisolism, and temporary loss of reproductive functions). Mankind, however, is biologically very ill-equipped for dealing with food abundance over very long periods of time. The abundance of food appears to be a necessary but not sufficient factor contributing to the etiopathogenesis of these eating disorders. See *Psychiatry and Sociology*.

3.2 Biological Factors

A number of twin studies have shown relatively high concordance rates among monozygotic compared to dizygotic twins especially for AN but also for BN. Currently large studies using the techniques of molecular genetics are underway but the results are not yet available. AN and BN frequently have an onset around puberty, the time when, due to hormonal changes, the body undergoes considerable changes. Starting at puberty, luteinizing hormone (LH) and follicle-stimulating hormone (FSH) are secreted regularly, inducing the secretion of gonadal hormones which have an impact not only on our body but also on our brain. Also, body fat is not distributed the same way in men and women. There exists a sexual dimorphism: the fat deposits of men are predominantly in the abdominal region and those of females are primarily observed on the trunk, buttocks, and the extremities.

The hypothalamus plays an important role in the regulation of eating behavior. Simplifying a highly complex regulation, one can speak of an 'eating center' in the lateral hypothalamus and a 'satiety center' in the medial hypothalamus. Electrical stimulation of the 'eating center' or making a lesion in the 'satiety center' results in excessive eating. Stimulating the 'satiety center' or putting a lesion into the 'eating center' results in loss of appetite, loss of weight, and finally death.

3.2.1 Control of food intake. The following substances secreted in the human body have been shown to reduce hunger or food intake: (a) corticotropin-releasing hormone (CRH) which is secreted in the hypothalamus, (b) certain peptides secreted centrally or peripherally such as cholecystokinin (CCK), glukagon, bombesin, gastrin-releasing peptide, somatostatin, neurotensin, and leptin, and (c) the monoamine serotonin. Other substances increase food intake, such as the neuropeptide Y and the peptide YY, galanin, beta-endorphine, dynorphine, growth-hormone-releasing-hormone, and the monoamine epinephrine. In animal experiments it has been shown that overeating can become a conditioned response to stress.

3.2.2 Hormonal adaptations. Most if not all hormonal changes observed in AN and to some extent also observed in other eating disorders (BN, BED) are a sign of adaptation to a temporarily reduced food intake or low body weight. In emaciated patients with AN (as in healthy starving individuals), hypercortisolism and insufficiently suppressed cortisol response following application of dexamethasone are observed. During starvation the hypothalamo–pituitary–gonadal system is shut down so that reproduction cannot take place. The 24-hour secretory pattern of the luteinizing hormone (LH) and the follicle-stimulating hormone (FSH) regress to a pubertal or infantile pattern. Plasma levels of estradiol and progesterone are practically down to zero in emaciated anorexic patients, resulting in amenorrhoea. Normal-weight bulimic women show ovary dysfunctions. About 50 percent of them have unovulatory cycles and 40 percent show defects in the lutal phase (Fichter 1992). Underweight anorexic patients show increased levels of insulin and glucose following a test meal. When body fat is metabolized during starvation, free fatty acids and ketone bodies are increased in the plasma. They can serve as state markers for a negative energy balance. In emaciated patients with AN, the activity of the sympathetic nervous system is reduced, resulting in bradycardia, hypotension, and dsregulation of body temperature. These are also mechanisms for saving energy during starvation. Emaciated anorexics also appear to have lower levels of serotonin and its metabolites. Starvation in healthy subjects as well as starvation in AN

Table 2
Goals and interventions in the therapy of AN and BN

Areas of deficit	Aims and areas of intervention
Information	**Giving relevant information to patient and relatives (psychoeducation)** ● nutrition: eating wide variety of foods; no 'forbidden foods' ● effects of specific stress/pressures on behavior ● possibilities and limitations of therapy ● consequences of anorexic and bulimic behavior ● importance of taking responsibility oneself (self-help)
Disturbance of interoceptive, proprioceptive, and emotional perception	**Enhancing perception in different ways** ● self-monitoring exercises to improve body perception (movement, touch) ● exercises to enhance enteroceptive perception (satiety, hunger, feeling pressures, pain) ● improving the perception of emotions
Problems in emotional expression and lack of social skills	**Training the expression of emotions** ● understanding the 'antecedents-behavior-consequences chain' ● exercises in catharthis ● how to express emotions appropriately (at the right time to the right person in the overall most effective way) ● social skills training in role play sessions
Dysfunctional, irrational thoughts, beliefs, and values	**Cognitive behavior therapy (CBT)** ● functional analysis of 'antecedents-behaviors/thought-consequences' ● identifying dysfunctional, irrational thoughts, beliefs and values ● 'Socratic dialog' questioning irrational thoughts, values and dysfunctional beliefs ● designing new and more appropriate 'software' (rational thoughts and functional beliefs)
Chronic difficulties or stress in social environment	**Reducing pressures and stress** ● by making changes in the environment (e.g. move to own flat) ● through focusing on interpersonal relationships (IPT) ● Sessions with parents or partner when indicated
Pathological eating behavior	**Nutritional counseling and behavior therapy** ● structure for 3 main meals and 1–2 snacks; no food consumption outside the structured meals ● eat every food (no banned food, no food selection) ● learning about the nonsense of dieting ('weight cycling') ● identify antecedents of binges; eating outside structured meals
Body weight deviations	**Body weight control and management** ● eat when feeling hungry but not when under pressure ● if underweight: stepwise increase of body weight in behavioral weight increase program ● if overweight: slow decrease in food intake and body weight over long period of time (behavioral program); in several cases very low calory diet (VLCD) combined with behavioral program; in very severe cases consider surgery
Passivity, hesitance to take responsibility and make personal decisions	**Activation of patient's own initiative and responsibility** ● encourage patient to make his/her own relevant and necessary decisions ● supporting reasonable self-help groups

Table 2
Continued

Areas of deficit	Aims and areas of intervention
Medication	**When indicated:** ● supplement potassium and other electrolytes and vitamins ● in BN and BED consider prescribing serotonin-receptor-inhibitor especially when fading out (behavioral) psychotherapy
Fear of failure and of relapse following therapy	**Maintenance program** ● find solutions for problems which can be anticipated ● gradual exposure to 'daily hassles' ● Further support necessary (outpatient treatment, self-help group) ● learning to cope efficiently with everyday stress

subjects results in brain atrophy demonstrated in CT and MRT studies.

3.3 External Factors

Life events and chronic difficulties may over time make an individual vulnerable to later develop an eating disorder or may trigger its onset. During the time of adolescence not only bodily changes occur but also the individual is confronted with new demands. They are now expected to have more responsibilities and they are increasingly confronted with contacts and relationships with the other sex. Especially during adolescence, feeling left out or rejected by others can hurt deeply and increase vulnerabilities. Developing an eating disorder is a way of avoiding failures in everyday life because a wide range of important things in life are reduced to a small area of shape and weight concerns. In controlling her weight an anorexic girl can succeed. In some families with an anorexic patient, family relationship interactions cause problems. Research in this area, however, is difficult because what we see in clinical cases can be the reaction of a family to the illness rather than its cause (Russell et al. 1987). More solid research is needed in this area.

4. Treatment of Eating Disorders

4.1 General Aspects

Details on the treatment of eating disorder are given in the practice guidelines of the American Psychiatric Association (2000) which can be obtained through the internet (http://mentalhealth.uca.edu/apa.scpg/eating-disorders). Antidepressants such as serotonin re-uptake inhibitors can be beneficial in the therapy of BN and BED. Their efficacy in AN has not clearly been established.

4.2 Psychological Therapies

There exists a wealth of data on the efficacy of cognitive behavior therapy (CBT) in BN (Agras et al. 2000). CBT has also been shown in some studies to be effective in BED. Although there are several controlled studies on the therapy of AN, research in the past two decades has focused on BN and BED and further controlled studies on the treatment of AN are needed. Table 2 shows major areas of intervention for anorexic and bulimic eating disorders. Most controlled trials on psychological therapy have focused on dealing with dysfunctional, irrational thoughts and beliefs using cognitive behavior therapy or have focused on directly modifying the eating behavior. However, it has also been shown that other forms of therapies such as nutritional management or interpersonal therapy which are completely different can be effective treatments for BN.

See *Cognitive and Interpersonal Therapy: Psychiatric Aspects*.

4.3 Specific Aspects for the Therapy of AN

When a patient is severely emaciated, psychological therapies alone will not lead very far; motivating the patient to eat more and gain weight is then very important. Some patients gain weight when they are encouraged and supported to do so; others need structured behavioral programs with strong incentives for weight gain. In a few patients both these approaches will fail and tube feeding will at least for a certain period of time be necessary. Weight increase of about 150 grams per day is desirable. Other focuses of the treatment of AN are listed in Table 2 (enhancing perception, training of the expression of emotions, enhancing social skills, cognitive restructuring of dysfunctional thoughts and beliefs, normalizing eating behavior, dealing with anxieties, and encouragement to take on new responsibilities).

Most controlled treatment studies with BN patients use cognitive behavior therapy (CBT) with some variation of it. Some studies have shown the usefulness of behavioral management of eating behavior, interpersonal therapy, nutritional counseling, and guided self-help. Little attention has been given by researchers

to assess treatment, focusing on disturbances of interoceptive, proprioceptive, and emotional perception which according to Hilde Bruch (1973) is a central feature of all eating disorders (Laumer et al. 1997). For dysfunctional behaviors such as bingeing it is of importance to identify antecedent events which trigger the dysfunctional behavior of the patient. In the next step when antecedence, behavior, and consequences are clearer, behavioral interventions such as social skill training can be employed to improve coping in future problem situations.

In addition to the other areas listed in Table 2, body weight control and management is of special importance for the treatment of patients with BED if they are overweight or obese.

See also: Body: Anthropological Aspects; Body Image and Gender; Eating Disorders, Determinants of: Genetic Aspects; Hunger and Eating, Neural Basis of; Obesity and Eating Disorders: Psychiatric; Psychological Treatment, Effectiveness of; Somatoform Disorders

Bibliography

Agras W S, Walsh B T, Fairburn C G, Wilson G T, Kraemer H C 2000 A multicenter comparison of cognitive-behavioral therapy and interpersonal psychotherapy for bulimia nervosa. *Archives of General Psychiatry* **57**: 459–66

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*. American Psychiatric Association, Washington, DC

American Psychiatric Association 2000 *Practice Guidelines on Eating Disorders*. http://mentalhealth.ucla.edu/apa.scpg/eating-disorders/page01.html-page81.html

Brownell K D, Greenwood M R C, Stellar E, Shrager E E 1986 The effects of repeated cycles of weight-loss and regain in rats. *Physiology and Behavior* **38**: 459–64

Bruch H 1973 *Eating Disorders: Obesity, Anorexia Nervosa, and the Person Within*. Basic Books, New York

Fichter M M 1985 *Magersucht und Bulimia*. Springer, Berlin

Fichter M M 1990 Psychological therapies in bulimia nervosa. In: Fichter M M (ed.) *Bulimia Nervosa*. Wiley, Chichester, UK, pp. 273–91

Fichter M M 1992 Starvation-related endocrine changes. In: Halmi K A (ed.) *Psychobiology and Treatment of Anorexia Nervosa and Bulimia Nervosa*. American Psychiatric Press, Washington DC

Fichter M M, Pirke K M, Holsboer F 1986 Weight loss causes neuroendocrine disturbances: Experimental study in healthy starving subjects. *Psychiatry Research* **17**: 61–72

Fichter M M, Quadflieg N 1997a Six-year course of bulimia nervosa. *International Journal of Eating Disorders* **22**: 1–24

Fichter M M, Quadflieg N 1997b Six-year course and outcome of anorexia nervosa. *International Journal of Eating Disorders* **26**: 359–85

Fichter M M, Quadflieg N, Gnutzmann A 1998 Binge eating disorder: Treatment and outcome over a six-year course. *Journal of Psychosomatic Research* **44**: 385–405

Garfinkel P E, Lin E, Goering P, Spegg C, Goldbloom D S, Kennedy S, Kaplan A S, Woodside B 1995 Bulimia nervosa in a Canadian community sample: prevalence and comparison of subgroups. *American Journal of Psychiatry* **152**: 1052–8

Harris E C, Barraclough B 1998 Excess mentality of mental disorder. *British Journal of Psychiatry* **173**: 11–53

Laumer U, Bauer M, Fichter M, Milz H 1997 *Therapeutische Effekte der Feldenkrais-Methode 'Bewußtheit durch Bewegung' bei Patienten mit Essstörungen. Psychotherapie, Psychosomatik, Medizinische Psychologie* **47**: 170–80

Ratnasuriya H, Eisler I, Szmukler G, Russell G F M 1991 Anorexia nervosa: outcome and prognostic factors after 20 years. *British Journal of Psychiatry* **158**: 495–502

Russell G F M, Szmukler C 1987 Dare: An evaluation of family therapy in anorexia nervosa and bulimia nervosa. *Archives of General Psychiatry* **44**: 1047–56

M. M. Fichter

Eating Disorders, Determinants of: Genetic Aspects

1. Etiology of Eating Disorders

The etiology of anorexia nervosa (AN) and bulimia nervosa (BN) has been widely assumed to be multiply determined by sociocultural, family environmental, and individual psychological and biological factors. Historically, the ubiquitous sociocultural emphasis on thinness has been thought to figure centrally in the causation of eating disorders by inducing body dissatisfaction, and by encouraging dieting and other unhealthy weight control practices.

Although intuitively appealing, evidence for a central etiological role of sociocultural factors is hardly persuasive. Exposure to thin ideals and dieting are nearly universal in industrialized countries; however, only a relatively small number of young women actually succumb to clinically significant eating disorders. The recent surge in research regarding the contribution of genetic effects to liability to eating disorders offers a more powerful strategy for addressing the perennial question of differential susceptibility to the development of eating disorders in the presence of near universal exposure to a powerful societal risk factor.

This article considers how the application of recent advances in family and twin study methodology, as well as quantitative and molecular genetics, offer a firm new scientific footing for pursuing more refined searches for genetic and environmental factors that contribute to the etiology of eating disorders.

2. Components of Genetic Epidemiology

There are three major research designs in genetic epidemiology that allow for the delineation and quantification of the relative contribution of genes and environment to complex behavioral traits. The first step is to determine whether the disorder aggregates in

families. This hypothesis is addressed by the classic controlled family study which determines whether there is a statistically greater lifetime risk of eating disorders in blood relatives of individuals who have an eating disorder in comparison to relatives of individuals without eating disorders. If no differential risk is observed, then the probability of the disorder being genetically influenced is low. However, because genetic and environmental effects are confounded in the family study design, twin and adoption studies are needed to determine the extent to which the observed familial resemblance for illness is due to genetic or environmental effects. If genetic effects appear to be important in the transmission of the disorder, then linkage and association studies are next employed to determine the precise location, identity, and function of the genes that are implicated. Each of these approaches (with the exception of adoption studies) has been employed in the study of eating disorders.

2.1 Family Studies of Eating Disorders

There are now a series of family studies addressing the fundamental question of 'are eating disorders familial?' Most (Gershon et al. 1983, Hudson et al. 1987, Kassett et al. 1989, Lilenfeld et al. 1998, Strober et al. 1990, 2000) but not all (Gershon et al. 1983, Halmi et al. 1991, Stern 1992) controlled family studies, have found a significantly greater lifetime prevalence of eating disorders among relatives of eating disordered individuals in comparison to relatives of controls. Intriguingly, several studies have found increased rates of both AN and BN (i.e., co-aggregation) in relatives of individuals with AN as well as individuals with BN, compared to rates among relatives of controls (Gershon et al. 1983, Hudson et al. 1987, Kassett et al. 1989, Strober et al. 1990, 2000), suggesting that AN and BN share transmissible risk factors. Moreover, relatives of individuals with AN and BN have also been found to have a significantly increased rate of subthreshold eating disorders compared to relatives of controls (Lilenfeld et al. 1998, Strober et al. 2000), suggesting that the eating disorders do not 'breed true,' but are expressed in families as a broad spectrum of eating-related pathology.

On balance, evidence from family study data suggests there is an elevation in the lifetime prevalence of eating disorders among the relatives of people with eating disorders. Moreover, suggestive evidence of the co-aggregation in families of AN, BN, and milder eating disturbances raises the provocative possibility of shared etiologic factors across these conditions.

2.2 Twin Studies

The goal of the classical twin study is to qualify and quantify similarities and differences between monozygotic (MZ) and dizygotic (DZ) twin pairs in order to identify and delineate genetic and environmental causes for a particular trait. Given that MZ twins, for most intents and purposes, share all of their genes, and DZ twins share, on average, half of their genes, any excess concordance in MZ twins over DZ twins suggests a genetic contribution to liability to the disorder. Conversely, any differences between MZ twins provides strong evidence for the role of environmental influences (Plomin et al. 1994, pp. 171–2), whereas differences between DZ twins can result from genetic and/or environmental effects. More complicated statistical modeling allows one to parse the variance in liability to illness into three sources: additive genetic effects (a^2), shared environmental effects (c^2), and unique environmental effects (e^2). Additive genetic effects reflect the cumulative impact on a trait of many individual genes each of which has a relatively small individual effect on the behavioral phenotype. The presence of a^2 is inferred when the correlation between MZ twins is greater than the correlation between DZ twins. By contrast, common environmental effects reflect etiological influences to which both members of a twin pair are exposed, regardless of zygosity. Examples of such effects include the social class and religious preference of the family of origin. Unique environmental effects, on the other hand, result from etiological influences to which one member of a twin pair is exposed but not the other, and contribute to *differences* between members of a twin pair. Examples include one member of a twin pair being exposed to a traumatic experience not shared with the co-twin.

One of the key caveats to the twin study is the assumption of equal environments (EEA) which posits that MZ and DZ twins are equally correlated for their exposure to environmental influences that are of etiologic relevance to the trait under study (Plomin et al. 1994). If the EEA is incorrect, then the greater resemblance of MZ twins in comparison to DZ twins could actually be due to environmental factors. A violation of the EEA does not necessarily invalidate the results, but may influence the magnitude of the estimated genetic and environmental components. Studies of the EEA with regard to eating disorders suggest that this assumption has not been violated in twin studies (Bulik et al. 2000, Kendler et al. 1993, Klump et al. 2000, Sullivan et al. 1998).

2.3 Twin Studies of Anorexia Nervosa

Beyond isolated case reports, the first systematic study of clinically ascertained twins with AN (Holland et al. 1984, 1988, Treasure and Holland 1989) found that the concordance for MZ twins was substantially greater than for DZ twins. Reanalyses of these data (assuming a population prevalence of AN of 0.75 percent) revealed evidence of familial aggregation with

parameter estimates of 88 percent for a^2, 0 percent for c^2, and 12 percent for e^2. In short, the observed familial aggregation for AN appears to be influenced most strongly by additive genetic effects; however, the estimates were rather imprecise given the small sample size.

Population-based studies of anorexia nervosa are difficult to conduct given the relatively low prevalence of the disorder. There has been one published study that has been able to successfully apply twin modeling techniques. Wade et al. (2000) derived heritability estimates for AN in the context of studying the nature of the comorbid relationship between AN and major depression. The heritability of AN was estimated to be 58 percent, although the authors could not rule out a contribution of shared environment to the liability to AN. On balance, we can conclude from these twin and family studies (see Lilenfeld et al. 1997 for a review) that AN is familial. However, the definitive resolution of the independent contribution of genetic and shared environmental factors to the observed familiality of AN will require more ambitious collaborative efforts to obtain larger sample sizes with sufficient statistical power.

2.4 Twin Studies of Bulimia Nervosa

Twin studies of BN have been more successful given the higher population prevalence of the disorder. Initial case series of twins with BN revealed consistently greater concordance for BN in MZ than DZ twin pairs (Fichter and Noegel 1990, Hsu et al. 1990, Treasure and Holland 1989). Pooling data from these case series for twin modeling and assuming a population prevalence of BN of 2.5 percent revealed evidence of familial aggregation with 47 percent of the variance accounted for by additive genetic effects, 30 percent by shared environmental effects, and 23 percent by unique environmental effects. However, the sample sizes were small and the estimates imprecise.

Population-based studies of BN have been conducted in the US (Bulik et al. 1998, Kendler et al. 1991) and Australia (Wade et al. in press). The studies that have estimated the heritability of BN based on a single occasion of measurement suggest a moderate contribution of additive genetic effects, a negligible contribution of shared environmental effects, and a more substantial contribution of unique environmental effects to liability to BN. Although a marked improvement over the clinical series, these studies still had limited statistical power given the relatively low population prevalence of BN and given that the reliability of the diagnosis of BN tends to be poor, which can lead to underestimation of both a^2 and c^2 (Bulik et al. 1998).

Two studies have augmented statistical power by incorporating more than one occasion of measurement into the twin model (Bulik et al. 1998, Wade et al. in press). This approach controls for unreliability of diagnosis, increases power to detect both a^2 and c^2, and therefore provides the most reliable information regarding the nature and magnitude of genetic and environmental contributions to BN. In short, these two studies reveal a markedly greater contribution of additive genetic effects to the liability to BN (59 and 83 percent, respectively), a negligible contribution of shared environment (0 percent in both studies), and a moderate contribution of unique environmental effects (41 and 17 percent). Although the parameter estimates for c^2 were 0, the confidence intervals did not completely rule out a contribution of shared environment. The results from these two studies confirm the central role of genes in the observed familiality of BN.

In summary, from twin and family studies, we can conclude that BN is familial and that there appears to be a moderate to substantial contribution made by genetic factors and unique environmental factors to liability to the disorder. The contribution of shared environment is less certain, but appears to be of lesser prominence than the effect of genes and unique environment. A reasonable next step for twin studies is to determine the precise nature of the unique environmental effects that increase risk for developing BN.

2.5 Molecular Genetic Studies

Several efforts have been made to identify the map position of individual genes conferring susceptibility to AN. To date, the focus has been on genes involved in serotonergic (5HT) neurotransmission. Collier et al. (1997) conducted an association study of 81 women with AN and 226 controls, finding a modest increase in the frequency of the 5HT2A -1438/A allele in AN compared to controls. These results have since been replicated by Sorbi et al. (1998) and by Enoch et al. (1998). Interestingly, in the Sorbi et al. study the increased frequency of this allelic variant was seen only in the subset of nonpurging restricting anorexics, whereas in the Enoch et al. study this genotype was not found to discriminate BN patients from controls, yet was also more common among patients with obsessive compulsive disease.

Two studies have failed to replicate the 5HT2A finding. Campbell et al. (1998) studied 157 anorexic patients and 150 controls, finding no association of this polymorphism with AN; however, stratification by the presence vs. absence of purging was not performed and the frequency of the 1438A genotype in their controls was higher than in other reported controls samples. Still, Hinney et al. (1998) studied 100 anorexics, with 100 underweight but nonanorexic controls, and 100 obese controls and found no intergroup differences in the frequency of this geno-

type. Ongoing international collaborative studies are continuing the search for susceptibility genes for both AN and BN (Kaye et al. 2000).

3. *Summary*

An increasing body of evidence supports the notion that AN and BN have a strong familial component, whereas twin studies suggest a substantial contribution of additive genetic effects to the observed familiality of both disorders. By contrast, the magnitude of the contribution of shared environmental effects remains unresolved. Moreover, both family and twin data suggest there is at least a moderate familial–genetic correlation between AN and BN, which may both reflect and explain the common presence of weight phobia in both disorders and the not infrequent development of binge eating during the natural course of AN.

That the contribution of additive genetic effects appears to be more prominent than shared environmental effects for both AN and BN is surprising given the historical focus on cultural and social factors in the study of eating disorders. There are several plausible explanations. First, environmental experiences that are shared in common by members of a family may have little bearing on risk to eating disorders. Thus, environmental factors that do influence risk may be those experiences that are unique to the individual, possibly acting to increase the likelihood of body dissatisfaction, or initiating weight control practices. Second, the statistical power of twin data to detect common environmental effects in the presence of genetic influences may be inadequate, although we note once again that the studies with greater statistical power support a greater contribution of genetic than shared environmental effects. Third, if familial factors exert effects in concert with an individual's genetic propensities, the main effects of familial environment will be weak but effects operating in gene–environment interaction might be considerable. Finally, the effects of shared environment may be more robust in childhood, but diminish with age (Klump et al. 2000).

An obvious critical question that ensues from these findings is the nature of a purported inherited diathesis, as it is improbable that specific genes exist that code for body weight phobia or dietary restraint. Indeed, critical to future studies of etiology will be efforts at refining the phenotype or phenotype spectrum, identifying those components of the phenotype that are most heritable, and, through co-variation with symptoms of the disorder, are the most likely candidates for mediation of genetic risk. Incorporating a focus on certain quantitative temperamental or personality traits in future genetic studies of eating disorders may be of great promise in this regard. For example, restricting-type AN in particular is characterized by a remarkably consistent cluster of heritable behavioral traits, including emotional restraint, avoidance of novelty, anxious worry and self-doubt, compliancy, perfectionism, and perseverance in the face of nonreward (Strober 1995). The current focus on candidate genes regulating serotonergic neurotransmission as a risk factor for AN derives, intuitively and empirically, on the one hand from preclinical data linking serotonergic neurotransmission to restraint of reward motivation for exploring novel environments, inhibitory modulation of feeding and sexual behavior, and enhanced sensitivity of neurobehavioral systems to stimulus events (Kaye et al. 1993), and on the other to evidence adduced from clinical studies (Kaye et al. 1991) of impaired regulation of serotonergic activity in AN patients well after restoration to normal weight levels. Thus, a convergence of preclinical and clinical data implicate abnormalities of serotonergically mediated brain and behavioral systems in the pathogenesis of AN by conferring extreme propensities toward behavioral rigidity, obsessiveness, perfectionism, and constraint. However, the contribution to risk of abnormalities in other brain substrates cannot be discounted.

Although the presentation of BN is more variable than that of restricting-type AN, a role for heritable personality and behavioral traits as indices of a biologically mediated etiologic mechanism in the development of this disorder is also heuristically and theoretically defensible. In contrast to AN, women with BN more often display traits of thrill seeking and excitability, and are more dysphoric in response to rejection or nonreward (Strober 1995). These features may predispose to periodic lapses in control in eating and other realms under conditions of stress. In this same vein, possible heritable variations in the sensitivity of brain emotional systems to the reward and motivational properties of feeding behavior may constitute a diathesis for the development and reinforcement of dysregulated appetitive behavior at times of emotional despair or life stress. These models of 'genetic' risk are easily reconciled with prevailing psychological theories of BN which view binge eating as a compensatory modulation of dysphoric states linked to deeply entrenched negative self-schemas.

In closing, we believe that the search for susceptibility genes in eating disorders is an exciting and valid research agenda. It may well hold forth the promise of bringing resolution to the continuing uncertainty of whether AN and BN are truly separate entities, or have risk and vulnerability factors in common. Given the body of research reviewed in this article, future empirical investigation bridging the behavioral and biological sciences and exploiting the rapidly advancing science of molecular and family genetic epidemiology may yield keys to refining our knowledge of the etiology, nosology, and therapeutics of eating disorders.

See also: Body: Anthropological Aspects; Body, Evolution of; Body, History of; Body Image and Gender; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Health and Mortality, Body Stature as a Measure of; Hunger and Eating, Neural Basis of; Obesity and Eating Disorders: Psychiatric; Schizophrenia and Bipolar Disorder: Genetic Aspects; Weight Regulation: Psychological Aspects

Bibliography

Bulik C M, Sullivan P F, Kendler K S 1998 Heritability of binge-eating and broadly defined bulimia nervosa. *Biological Psychiatry* **44**: 1210–8

Bulik C M, Sullivan P F, Wade T D, Kendler K S 2000 Twin studies of eating disorders: A review. *International Journal of Eating Disorders* **27**: 1–20

Campbell D A, Sundaramurthy D, Markham A F, Pieri L F 1998 Lack of association between 5-HT2A gene promoter polymorphism and susceptibility to anorexia nervosa. *Lancet* **351**: 499

Collier D A, Arranz M J, Li T, Mupita D, Brown N, Treasure J 1997 Association between 5-HT2A gene promoter polymorphism and anorexia nervosa. *Lancet* **350**: 412

Enoch M A, Kaye W H, Rotondo A, Greenberg B D, Murphy D L, Goldman D 1998 5-HT2A promoter polymorphism–1438G/A, anorexia nervosa, and obsessive-compulsive disorder. *Lancet* **351**: 1785–6

Fichter M M, Noegel R 1990 Concordance for bulimia nervosa in twins. *International Journal Eating Disorders* **9**: 255–63

Gershon E, Schreiber J, Hamovit J, Dibble E D, Kaye W H, Nurnberger J I, Andersen A, Ebert M 1983 Anorexia nervosa and major affective disorders associated in families: A preliminary report. In: Guze S B, Earls F G, Barrett J E (eds.) *Childhood Psychopathology and Development*. Raven Press, New York, pp. 279–84

Halmi K, Eckert E, Marchi P, Sampugnaro V, Apple R, Cohen J 1991 Comorbidity of psychiatric diagnoses in anorexia nervosa. *Archives of General Psychiatry* **48**: 712–18

Hinney A, Ziegler A, Nöthen M, Remschmidt H, Hebebrand J 1998 5-HT2a receptor gene polymorphisms, anorexia nervosa and obesity. *Lancet* **350**: 1724–5

Holland A J, Hall A, Murray R, Russell G F M, Crisp A H 1984 Anorexia nervosa: A study of 34 twin pairs and one set of triplets. *British Journal of Psychiatry* **145**: 414–19

Holland A, Sicotte N, Treasure J 1988 Anorexia nervosa: Evidence for a genetic basis. *Journal of Psychosomatic Research* **32**: 561–71

Hsu G L K, Chesler B E, Santhouse R 1990 Bulimia Nervosa in eleven sets of twins: A clinical report. *International Journal of Eating Disorders* **9**: 275–82

Hudson J I, Pope H G, Jonas J M, Yurgelun-Todd D, Frankenburg F R 1987 A controlled family history study of bulimia. *Psychological Medicine* **17**: 883–90

Kassett J, Gershon E, Maxwell M, Guroff J J, Kazuba D M, Smith A L, Brandt H A, Jimerson D C 1989 Psychiatric disorders in the first-degree relatives of probands with bulimia nervosa. *American Journal of Psychiatry* **146**: 1468–71

Kaye W H, Lilenfeld L, Berretini W, Strober M, Devlin B, Klump K, Goldman D, Bulik C M, Halmi K A, Fichter M, Kaplan A, Woodside B, Treasure J, Plotnikov K, Pollice C, Rao R, McConaha C 2000 A genome-wide search for susceptibility loci in anorexia nervosa: Methods and sample description. *Biological Psychiatry* **47**: 794–803

Kaye W H, Weltzin T E, Hsu L K G 1993 Relationship between anorexia nervosa and obsessive-compulsive behaviors. *Psychiatric Annals* **23**: 365–73

Kaye W H, Gwirtsman H E, George D T, Ebert M H 1991 Altered serotonin activity in anorexia nervosa after long-term weight restoration. *Archives of General Psychiatry* **48**: 556–62

Kendler K S, MacLean C, Neale M C, Kessler R C, Heath A C, Eaves L J 1991 The genetic epidemiology of bulimia nervosa. *American Journal of Psychiatry* **148**: 1627–37

Kendler K S, Neale M C, Kessler R C, Heath A C, Eaves L J 1993 A test of the equal environment assumption in twin studies of psychiatric illness. *Behavioral Genetics* **23**: 21–7

Klump K L, Holly A, Iacono W, McGue M, Willson L 2000 Physical similarity and twin resemblance for eating attitudes and behaviors: A test of the equal environments assumption. *Behavioral Genetics* **30**: 51–8

Klump K L, McGue M K, Iacono W G 2000 Age differences in genetic and environmental influences on eating attitudes and behaviors in adolescent female twins. *Journal of Abnormal Psychology* **109**: 239–51

Lilenfeld L, Kaye W H, Greeno C, Merikangas K R, Plotnicov K, Pollice C, Rao R, Strober M, Bulik C M, Nagy L 1998 A controlled family study of restricting anorexia and bulimia nervosa: Comorbidity in probands and disorders in first-degree relatives. *Archives of General Psychiatry* **55**: 603–10

Lilenfeld L, Kaye W, Strober M 1997 Genetics and family studies of anorexia nervosa and bulimia nervosa. *Balliere's Clinical Psychiatry* **3**: 177–97

Plomin R, DeFries J C, McClearn G E, Rutter M 1994 *Behavioral Genetics. Third Edition*. W. H. Freeman & Co., New York

Sorbi S, Nacmias B, Tedde A, Ricca V, Mezzani B, Rotella C 1998 5-HT2A promoter polymorphism in anorexia nervosa. *Lancet* **351**: 1785

Stern S L, Dixon K N, Sansone R A, Lake M D, Nemzer E, Jones D 1992 Psychoactive substance use disorder in relatives of patients with anorexia nervosa. *Comparative Psychiatry* **33**: 207–12

Strober M 1995 Family-genetic perspectives on anorexia nervosa and bulimia nervosa. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity*. Guilford Press, New York, pp. 212–18

Strober M, Freeman R, Lampert C, Diamond J, Kaye W 2000 A controlled family study of anorexia nervosa and bulimia nervosa. *American Journal of Psychiatry* **157**: 393–401

Strober M, Lampert C, Morrell W, Burroughs J, Jacobs C 1990 A controlled family study of anorexia nervosa: Evidence of familial aggregation and lack of shared transmission with affective disorders. *International Journal of Eating Disorders* **9**: 239–53

Sullivan P F, Bulik C M, Kendler K S 1998 The genetic epidemiology of binging and vomiting. *British Journal of Psychiatry* **173**: 75–9

Treasure J, Holland A 1989 Genetic vulnerability to eating disorders: Evidence from twin and family studies. In: Remschmidt H, Schmidt M (eds.) *Child and Youth Psychiatry: European Perspectives*. Hogrefe & Huber, New York, pp. 59–68

Wade T D, Bulik C M, Kendler K S 2000 Anorexia nervosa and major depression: An examination of shared genetic and environmental risk factors. *American Journal of Psychiatry* **157**: 469–71

Wade T D, Martin N, Neale M, Tiggemann M, Trealor S, Heath A, Bucholz K, Madden P in press The structure of genetic and environmental risk factors for three measures of disordered eating characteristic of bulimia nervosa. *Psychological Medicine*

C. Bulik and M. Strober

Eclectic Psychotherapy: Psychological Perspectives

Eclectic psychotherapy refers to the practice of psychotherapy that is not limited to one theoretical orientation such as psychoanalysis, behavior therapy, or cognitive therapy. Instead, the therapist can select and use any therapeutic procedure or combination of techniques that he or she believes is best suited for the treatment of the individual patient. In the *American Heritage Dictionary of the English Language* (1971), the word eclectic is defined as follows: 'adj. 1. Choosing what appears to be the best from diverse sources, systems, or styles. 2. Consisting of that which has been selected from diverse sources, systems, or styles' (p. 412). Being an eclectic thus allows the therapist the widest latitude in conducting psychotherapy. At the same time, knowing that a therapist is an eclectic practitioner tells one little about the specific procedures or approach the therapist may use in working with specific types of patients.

1. The Growth of the Psychotherapies and the Popularity of Eclecticism

Although the field of psychotherapy grew at a distinctly slow pace during the first half of the twentieth century, the situation in the second half was decidedly different. Influenced greatly by the large number of psychiatric casualties in World War II, a tremendous increase in both the number of psychotherapists and in the number of different approaches or schools of psychotherapy has occurred. One can even view this surprising phenomenon as a growth industry, not only in the United States, but in many other countries as well. From the more than 60 different orientations or approaches to psychotherapy that I noted in the mid-1960s, the number has grown to over 400 at present (Garfield 1994). What has been particularly interesting (or amazing) has been the theoretical differences among these forms of psychotherapy, some of which appear to be diametrically opposed to each other, and the strong identification that the adherents of these approaches have to their chosen form of psychotherapy.

Despite this proliferation of the psychotherapies, another movement in psychotherapy, eclecticism, was growing rapidly, and in fact became the most popular approach to psychotherapy, particularly among psychologists. In the period from 1960 to 1990, a number of studies were conducted. 'Depending upon the type and size of the subject sample as well as the survey methods and procedures used, the percentage of those who designated themselves as eclectics has ranged from around one-third to well over one-half of the samples evaluated, and eclecticism has always ranked as the number one choice' (Garfield 1994, p. 124). In a recent report on primary theoretical orientations of clinical psychologists over the past three decades, it was also noted that eclecticism 'remained the modal orientation' (Norcross et al. 1997). It is also worth noting that an eclectic approach to psychotherapy has been popular with a wide variety of practitioners including clinical psychologists, counseling psychologists, psychiatrists, social workers, and marriage and family therapists.

2. Diversity among Eclectic Therapists

Although one can state with some degree of assurance that an eclectic approach to psychotherapy is popular among most groups of psychotherapists, this knowledge will not inform a potential patient what procedures a specific eclectic therapist prefers and will actually use. There is a wide diversity among eclectic psychotherapists in how they conduct their therapy. In a study of 154 self-identified eclectic clinical psychologists, such diversity was clearly apparent (Garfield and Kurtz 1977). Each of the eclectic therapists was asked to select the two theoretical orientations that were most characteristic of the therapist's eclectic approach and the two that were least characteristic. Interestingly, whereas 46 of the eclectic therapists indicated that they held a positive view of a psychoanalytic orientation, 41 expressed a negative view. Somewhat similar divergences were apparent with regard to other theoretical views. For example, 36 therapists expressed a positive view of Rogerian or client-centered theories as compared with 25 who gave negative responses. There were also some significantly divergent reactions to some theoretical orientations. Whereas 72 respondents held positive views of learning theory, 26 expressed negative views. In contrast, only 18 respondents indicated a positive response to rational–emotive therapy as compared with 57 who indicated a negative view. Existentialist views were also relatively unpopular with only 17 positive selections and 44 negative indications.

The respondents in this study were also asked to indicate the two theoretical orientations that they combined and considered most characteristic of their actual eclectic approach to therapy. Their responses

again reflected the variability of views expressed by this sample of eclectic clinical psychologists. The most frequent combination, psychoanalytic and learning theory, was indicated by only 21 of the 154 respondents. The next two most frequent combinations were neo-Freudian and learning theory with 16, and neo-Freudian and Rogerian with 9. Altogether, there were 32 different combinations of theoretical views indicated for the total sample. As is apparent, the eclectic orientation in psychotherapy covers a wide variety of viewpoints and procedures.

Eclectics thus are similar in not limiting themselves to just one theoretical orientation, but in their practice they exhibit considerable diversity. The emphasis is on selecting the procedures or techniques that appear to best meet the particular needs of the individual patient, or as some have phrased it: 'Different strokes for different folks'—adjust to the individual patient instead of requiring the patient to adjust to the therapist's strict adherence to one theoretical approach.

The diversity among eclectic therapists was also apparent in later studies and publications. In the *Handbook of Eclectic Psychotherapy* and the *Casebook of Eclectic Psychotherapy* edited by John Norcross (1986, 1987), several categories of eclecticism were discussed and illustrated. These included 'functional eclectic psychotherapy,' 'systematic eclectic psychotherapy,' 'Multimodal Therapy,' 'radical eclecticism,' 'systematic technical eclecticism,' 'theoretically consistent eclecticism,' 'a transtheoretical approach to eclecticism,' 'a structural–phenomonological eclectic psychotherapy,' and my own approach which is now referred to as a 'common factors approach.' As is apparent from the preceding list, diversity and proliferation appear to be characteristic features of the field of psychotherapy at the present stage of development. However, to provide the reader with a more concrete example of eclecticism in psychotherapy, I will make a brief presentation of one eclectic approach to psychotherapy which emphasizes the importance of both common and specific variables in the therapeutic process (Garfield 1995).

3. *Therapeutic Variables: Common and Specific*

Since the 1970s there has been a significant increase in the quantity and quality of research on psychotherapy. Although not all forms of psychotherapy have been systematically evaluated, the findings have generally been positive for those that have been so evaluated. Furthermore, with some exceptions, the comparative studies of these psychotherapies have reported few differences in outcome between them. For example, in one well-conducted study that compared psychoanalytic psychotherapy with behavior therapy—two very different forms of psychotherapy—no significant differences in outcome were secured (Sloane et al. 1975). Both forms of therapy had positive outcomes and their results exceeded those of a wait-list control group. Furthermore, taped sessions of the two therapies clearly reflected the differences in procedures of these therapies so that they could be identified easily. Nevertheless, the results of both were positive and not statistically different. In light of such findings, the possibility of common therapeutic factors among most forms of psychotherapy being responsible for at least a significant portion of the positive change obtained has emerged as a reasonable hypothesis. At the same time, for understandable reasons, the common factors hypothesis has not been warmly embraced by psychotherapists who adhere strictly to one theoretical approach. It would appear that at least a number of therapists prefer to be identified with a specific form of therapy than to acknowledge that most therapies probably have a number of features in common, and that these common factors may account for a significant portion of the positive changes they produce.

4. *Some Obvious Common Therapeutic Variables*

A number of basic similarities among the different forms of psychotherapy can be noted. Even on a surface or superficial level, the following commonalities can be noted:

(a) Each form of psychotherapy requires a client or patient who seeks help for some psychological problem or problems.

(b) Each form of psychotherapy requires a therapist or counselor who supposedly has received or completed a program of appropriate training and has learned certain therapeutic skills.

(c) Some type of verbal interaction takes place between the participants.

(d) Usually some type of interpersonal relationship is also formed.

(e) Most patients also will have some expectations about psychotherapy and the role of the therapist.

(f) A significant percentage of patients, regardless of the type of therapy secured, will drop out of therapy before the therapist indicates it is completed.

(g) A significant number of patients, usually around two thirds, will experience some positive change in their perceptions, cognitions, affects, and behavior.

Although many psychotherapists would agree with this list of possible common aspects of psychotherapy, they most likely would also state that these aspects or factors are not the really important factors that make for change in psychotherapy. The therapists would refer to variables or procedures that are emphasized in their particular approach. Interpretation and transference would be emphasized by psychoanalysts, reflection of feelings by client-centered therapists, and desensitization by behavior therapists. However, the

theoretical processes mentioned can also be viewed or designated differently. What is important in psychotherapy is not what a particular therapist says has been done and expressed in the terminology of a particular system, but what was actually experienced by the patient. 'The therapist's perception of events is merely one view of what has occurred and may not necessarily be the best view' (Garfield 1991, p. 9). Potential therapeutic variables that are not emphasized by a particular orientation will generally not be perceived by practitioners of that orientation. However, if the therapist assumes a more flexible attitude, a number of commonalities can be noted among the psychotherapies.

5. *Important Common Therapeutic Variables*

The most obvious important common variable is the therapeutic relationship that develops in psychotherapy, and it is important for both continuation and outcome in psychotherapy. Even behavior therapists who previously said relatively little about it have now acknowledged its importance (see *Dialectical Behavior Therapy*). For example, Paul Emmelkamp, a distinguished Dutch behavior therapist has stated: 'It is, however, becoming increasingly clear that the quality of the therapeutic relationship may be influential in determining success or failure of behavioral therapies, although well-controlled studies … are rare' (Emmelkamp 1986, p. 432).

Although the therapeutic relationship is a common and necessary therapeutic variable, it is not sufficient alone to account for significant positive change in most cases. Other variables are usually required. For example, one variable that is emphasized by psychoanalytically oriented therapists is termed 'interpretation,' and is offered to provide 'insight' to the patient about the patient's underlying problems. However, one can offer nonanalytic interpretations to patients. I have referred to such a process as providing explanations or understanding and others have referred to the same process as providing the patient with a therapeutic rationale. The point emphasized here is that there is no one 'true insight' or understanding or explanation. Each form of therapy provides its own and usually different explanations or interpretations—and, if they make sense to the patient and are accepted by the patient, they will usually have some positive impact. In other words, this is a common therapeutic variable among diverse forms of psychotherapy.

Additional therapeutic variables that may be common to most forms of psychotherapy are the creation of hope for being helped and the restoration of morale, the opportunity to confide one's difficulties to a sympathetic and sanctioned healer, being helped to confront one's problems, encouraged to try out new behaviors, being given helpful information, and even receiving advice and suggestions. These are all potential common factors that may not be mentioned in the official descriptions of the various forms of psychotherapy, but they do play a therapeutic role in psychotherapy. Reports of both observers of well-known therapists and patients attest to their use and importance. For example, in the Temple University Study (Sloane et al. 1975) referred to earlier, that compared psychoanalytic and behavior therapy, the patients in both groups tended to emphasize the same features of therapy in their appraisals at the end of therapy.

The successful patients in both therapies placed primary importance on more or less the same items. The following items were each termed 'extremely important' or 'very important' by at least 70 percent of successful patients in both groups:

(a) The personality of your doctor,
(b) His helping you to understand your problem,
(c) Encouraging you gradually to practice facing the things that bother you,
(d) Being able to talk to an understanding person,
(e) Helping you to understand yourself.

(Sloane et al. 1975, p. 206).

6. *Specific Factors in Psychotherapy*

In the preceding discussion of common factors or variables in psychotherapy, an attempt was made to abstract processes that have been emphasized within different systems of therapy with variable terminology or that have been neglected in the descriptions of some therapeutic orientations. These have been viewed as basic common variables in psychotherapy. At the same time, there are some specific procedures or techniques that have been developed largely by behavioral or cognitive-behavioral therapists for treating specific clinical problems. These include relaxation procedures, homework assignments, exposure treatment, systematic desensitization, and cognitive restructuring. Whereas common factors can be viewed as basic therapeutic variables of importance in practically all therapeutic endeavors, specific factors of the type mentioned would be selected for use with specific problems or disorders. The choices made would be influenced by the available empirical research. The eclectic therapist thus has a wide choice of procedures, rationales, and techniques that can be used as indicated for the individual case.

See also: Behavior Therapy: Psychiatric Aspects; Behavior Therapy: Psychological Perspectives; Person-centered Psychotherapy; Psychoanalysis in Clinical Psychology; Psychotherapy Integration; Psychotherapy Process Research

Bibliography

Emmelkamp P M G 1986 Behavior therapy with adults. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York, pp. 385–442

Funk C E (ed.) 1971 *American Heritage Dictionary of the English Language*. American Heritage Publishing Company, New York

Garfield S L 1991 Common and specific factors in psychotherapy. *Journal of Integrative and Eclectic Psychotherapy* **10**: 5–13

Garfield S L 1994 Eclecticism and integration in psychotherapy: Developments and issues. *Clinical Psychology Science and Practice* **1**: 123–37

Garfield S L 1995 *Psychotherapy, An Eclectic Approach*, 2nd edn. Wiley, New York

Garfield S L, Kurtz R 1977 A study of eclectic views. *Journal of Consulting and Clinical Psychology* **45**: 78–83

Norcross J C (ed.) 1986 *Handbook of Eclectic Psychotherapy*. Brunner/Mazel, New York

Norcross J C (ed.) 1987 *Casebook of Eclectic Psychotherapy*. Brunner/Mazel, New York

Norcross J C, Karg R S, Prochaska J O 1997 Clinical psychology in the 1990s: Part I. *The Clinical Psychologist* **50** (2): 4–9

Sloane R B, Staples F R, Cristol A H, Yorkston N J, Whipple K 1975 *Psychotherapy Versus Behavior Therapy*. Harvard University Press, Cambridge, MA

S. L. Garfield

Ecological Economics

Ecological economics is a recently developed field, which sees the economy as a subsystem of a larger finite global ecosystem. Ecological economists question the sustainability of the economy because of its environmental impact and its material and energy requirements, and also because of the growth of population. Attempts at assigning money values to environmental services and losses, and at correcting macroeconomic accounting are part of ecological economics, but its main thrust is in developing physical indicators and indexes of sustainability. Ecological economists also work on the relations between property rights and resource management; and model the interactions between the economy and the environment. They study ecological distribution conflicts; use management tools such as integrated environmental assessment and multi-criteria decision aids; and propose new instruments of environmental policy.

1. Origins

The book that came out of the first world conference of ecological economists in Washington DC in 1990 (Costanza 1991) defined the field as 'the science and management of sustainability.' In the late nineteenth and early twentieth century the biologist and urban planner, Patrick Geddes, the *narodnik* revolutionary and physician, Sergei Podolinski, and the engineer and social reformer, Josef Popper-Lynkeus had unsuccessfully tried to promote a biophysical view of the economy as a subsystem embedded in a larger system subject to the laws of thermodynamics (Martinez-Alier and Schlupmann, 1987). By 1850 or 1860 the carbon cycle and the cycles of plant nutrients had been discovered, while the first and second laws of thermodynamics (conservation and transformation of energy, but also dissipation of energy and increase in entropy) had been established. The contrived conflict between the 'optimistic' theory of evolution which explained the diversity of life, and the 'pessimistic' second law of thermodynamics, was a staple of the cultural diet of the early 1900s. The main ingredients for an ecological view of the economy were therefore present long before the birth of a self-conscious ecological economics, which was delayed by the strict boundaries existing between the natural and the social sciences.

The biologist and systems ecologist, Alfred Lotka, born in 1880, introduced in the 1910s and early 1920s the fundamental distinction between the endosomatic and the exosomatic use of energy by humans. The winner of the Nobel Prize in chemistry, Frederick Soddy, born in 1877, also wrote on energy and the economy. He compared 'real wealth' which grows at the rhythm of nature and which, if turned into manufactured capital, is worn down, with 'virtual wealth' in the form of debts which apparently could grow exponentially for ever. Later, four well-known economists, who did not at that time form a school, are seen in retrospect as ecological economists. Kenneth Boulding, born in 1910, worked mainly on general systems analysis. K. W. Kapp, also born in 1910, and S. von Ciriacy-Wantrup, born in 1906, were both institutionalist economists, and Nicholas Georgescu-Roegen, born in 1906, who was the author of *The Entropy Law And The Economic Process* (1971). The systems ecologist, H. T. Odum, born in 1924, studied the use of energy in the economy. Some of his former students were among the founders of the International Society for Ecological Economics in 1987. Other sources of ecological economics are found in environmental and resource economics (i.e., microeconomics applied to environmental pollution and the depletion of natural resources), in human ecology, ecological anthropology, agroecology, and urban ecology. They are also found in the study of 'industrial metabolism' as developed by Robert Ayres, now known as industrial ecology.

After an influential meeting in Sweden in 1982, organized by the ecologist AnnMari Jansson, on the integration of economics and ecology, the International Society for Ecological Economics (ISEE) was launched at a workshop in Barcelona in 1987. This was in precisely the same year as the Brundtland Report on 'sustainable development' was published.

Herman Daly (a former student of Georgescu-Roegen, and today's best known ecological economist) proposed that the word 'development' should mean changes in the economic and social structure, while 'growth' means an increase in the scale of the economy which probably cannot be ecologically sustained. 'Sustainable development' is thus acceptable to most ecological economists, while 'sustainable growth' is not (Daly and Cobb 1994). The first issue of the successful academic journal, *Ecological Economics*, came out in 1989, edited by the ecologist Robert Costanza, who was also the first president of ISEE. The ISEE has affiliated societies in Argentina and Uruguay, Australia and New Zealand, Brazil, Canada, the European Union, India, Russia and the United States.

Outside the United States and Europe, the Japanese 'entropy school' of economic analysis (Tamanoi et al. 1984) studied the environmental services provided by the water cycle and the ancient urban ecosystems of Japan. In India, much work has been done since the 1970s by economists but also by biologists (Madhav Gadgil) on the links between forest or water management and common property rights; which were, in the beginning of the twenty-first century a main focus of interest in ecological economics (Berkes and Folke 1998). Other early ecological economists (whose major works were not in English) are, in France, Rene Passet (1996, [1979]) and Ignacy Sachs who proposed in the early 1970s the notion of 'eco-development'; Roefie Hueting (1980) in the Netherlands and Christian Leipert (1989) in Germany; as well as Jose-Manuel Naredo in Spain. (For general introductions to the field, see Costanza et al. 1997a, Costanza et al. 1997b, and Common 1995).

2. *Scope*

Ecological economists see the economy as an open system. In thermodynamics, systems are classified as open to the entry and exit of energy and materials; closed to the entry and exit of materials though open to the entry and exit of energy, such as the Earth; and isolated systems (without the entry or exit of energy and materials). The availability of free energy and the cycling of materials allows life forms to become increasingly highly organized and complex. The same thing applies to the economy. Dissipated energy and waste are produced in the process. At least part of the waste can be recycled or, if not, the economy takes in new resources. However, if the scale of the economy is too large and its speed is too rapid, then natural cycles cannot produce the resources or absorb or assimilate the residues such as, for instance, heavy metals or carbon dioxide. In ecological economics, the economy is seen as embedded in the ecosystem (or, more accurately, in the historically changing, social perception of the ecosystem). The economy is also embedded in a structure of property rights to environmental resources and services, in a social distribution of power and income, and in social structures of gender, social class, or caste.

In contrast, in conventional economics the economy is seen as a self-sufficient system where prices for consumer goods and services and prices for the services of production factors, are formed. This preanalytic stand is reflected in the category of 'externalities.' Ecological economists (Norgaard, 1990) have disputed the view expressed in the 1960s by Barnett, Krutilla and many other resource economists that, since natural resources are cheap, they must be abundant. Markets are myopic, they discount the future; they cannot see future uncertain scarcities of sources or sinks. Ecological economists sympathize with attempts at 'internalizing' externalities into the price system, and they readily concur with proposals to correct prices by taxes such as 'natural capital depletion taxes' or taxes on pollution. But they deny that there exists a set of 'ecologically correct prices.'

In summary, ecological economics is a new transdisciplinary field which develops or introduces topics and methods such as:

(a) new indicators and indices of (un)sustainability of the economy;

(b) the application of ecological notions of carrying capacity and resilience to human ecosystems;

(c) the valuation of environmental services in money terms, but also discussion on the incommensurability of values and the application of multicriteria evaluation methods;

(d) risk assessment, uncertainty, complexity, and 'postnormal' science;

(e) integrated environmental assessment, including the building of scenarios, dynamic modeling, participatory methods of decision making;

(f) ecological macroeconomics, the measurement of 'natural capital,' the debate between 'weak' and 'strong' notions of sustainability;

(g) relations between ecological and feminist economics;

(h) ecological distribution conflicts;

(i) relations between the allocation of property rights and resource management, and old and new communal institutions for environmental management;

(j) international trade and the environment, the 'ecological debt';

(k) environmental causes and consequences of technological change, relations between ecological economics and evolutionary economics;

(l) theories of consumption (needs, satisfactors) as they relate to environmental impacts;

(m) the 'dematerialization' debate; relations with industrial ecology;

(n) applications in business administration; and

(o) instruments of environmental policy, often

centered on the 'precautionary principle' (or on 'safe minimum standards,' as developed by Ciriacy-Wantrup).

Only some of these points can be developed in the remaining space, doing some injustice by this choice to other work by ecological economists.

3. *Disputes on Value Standards*

The Greek distinction (as in Aristotle's *Politica*) between *oikonomia* (the art of material provisioning of the household) and *chrematistics* (the study of the formation of market prices, in order to make money) seems irrelevant today. In our world, material provisioning appears to be achieved mostly through market exchanges, and there is a fusion of *chrematistics* with *oikonomia*. However, many caring activities in families and in society and many of the services of nature (Waring 1988), remain outside the market. In ecological economics, the word 'economics' is used in a sense closer to *oikonomia* than to *chrematistics*. Ecological economics is not committed to a unique type of value expressed in a single numeraire (defined as the present value, in money terms, of costs and benefits, including, of course, monetarized externalities and environmental amenities).

> The issue is not whether it is only the market place that can determine value, for economists have long debated other means of valuation; our concern is with the assumption that in any dialogue, all valuations or 'numeraires' should be reducible to a single one-dimension standard (Funtowicz and Ravetz 1994, 198).

Ecological economics encompasses money valuation, together with physical appraisals of the environmental impacts of the human economy measured in their own physical 'numeraires.'

Nature provides resources for the production of commodities as well as environmental amenities. As shown by Gretchen Daily, R. de Groot, and others, more importantly, nature gives free, essential life-support services. These include the cycling of nutrients, the water cycle, soil formation, climate regulation, the conservation and evolution of biodiversity, concentration of minerals, the dispersal or assimilation of pollutants, diverse forms of useful energy, etc. Attempts have been made to assign money values to the annual flows of some environmental services, to compare them to GNP in monetary units of account. For instance, the cycling of nutrients (nitrogen, phosphorous) in some natural systems may be given a plausible money value by comparison with the costs of alternative human-made technologies. Could this same methodology (i.e., the cost of alternative technology) be applied consistently to the valuation of biodiversity in a kind of science fiction framework?

For biodiversity, money valuation has taken a completely different tack, namely the small sums exchanged in some 'bioprospecting' contracts, or fictitious, subjective money values in terms of 'willingness to pay' for conservation projects. One such is the so-called Contingent Valuation method favored by environmental economists (though not by most ecological economists). But how can we count the service that nature provides us by concentrating minerals which we disperse? ('Exergy' costs have been calculated by industrial ecologists, but the technology for creating mineral deposits does not exist). The figures obtained for the money values of environmental services provided free by nature are therefore inappropriate, although useful, in stimulating the debate on how to take nature into account.

Ecological economics rests on a foundation of 'weak comparability of values' (O'Neill 1993). For example, let us assume that a new large garbage dump must be built near a city, and one of three possible locations, A, B, C, will be sacrificed to this use. In our example, the three different locations are compared under three different types of value: value as habitat, value as landscape, and economic value. Location A is a most valuable, publicly owned wetland (valuable as habitat or ecosystem because of its richness of species) but a monotonous landscape, much visited by bird watchers and schools (and, as such, of some economic value according to the 'travel cost method'). Location C produces much rent as industrial and urban land, and therefore ranks first in economic value, but ranks only third as an ecosystem or habitat, and comes second as landscape (because of its historical qualities). Location B is an old agricultural area of beautiful derelict orchards and ancient manor houses, which ranks first as landscape, but ranks only third as rent-producing, and second as ecosystem or habitat. Which location should be sacrificed and how is the issue to be decided? Should and could all values be reduced to a super-value, so as to achieve strong comparability, and even strong commensurability (cardinal measurement)? In the example, the economic values (in actual or fictitious markets) of all three locations have been taken into account, but there is no super-value (economic or otherwise, such as, for instance, net energy production, by which the wetland would presumably come out on top). Certainly, the present rankings could be reconsidered. Thus, the landscape value of A could be upgraded and its economic value (like that of B) could be increased by Contingent Valuation. Also, giving more weight to some criteria than to others, or 'veto thresholds' for some criteria such as the 'endangered species' provision in American legislation, or the introduction of more locations or more criteria, would help us to reach a decision. The point of the exercise is merely to show the meaning of 'weak comparability of values.' The decision-making process need not be irrational (by lottery, for instance).

In contrast to such a multicriteria approach (Munda 1995), in cost–benefit analysis the projects to be evaluated are all valued in the same numeraire.

In microeconomics there is strong comparability of values, and indeed strong commensurability when externalities are internalized into the price system. One example is in the definition of a Pigovian tax as the economic value of the externality at optimum pollution level. In macroeconomics, El Serafy's practical proposals to 'green' the GNP (in Costanza 1991)—the results of which will depend on the chosen rate of interest—do not go beyond strong commensurability in money terms.

According to El Serafy, not all receipts from the sale of exhaustible resources ('natural capital') should be included in GNP. Only one part should be included, 'true' income, and the rest should be counted as 'decapitalization' or the 'user cost' of such 'natural capital' which should be invested at compound interest over the period until the resource is exhausted, so as to allow the country to live at the same standard of living even when running out of the resources. This proposal, based on the definition of 'income' by Hicks, and related to Hotelling's rule in resource microeconomics, is based on a notion of 'weak' sustainability only. 'Weak' sustainability allows the substitution of manufactured capital for so-called 'natural capital'—implying, therefore, a common unit of measurement, i.e., money value—while 'strong' sustainability refers to the maintenance of physical natural resources and services. This distinction was introduced by David Pearce and Kerry Turner, c. 1990.

The so-called 'Environmental Kuznets Curve,' an inverted U-curve, relates income and some environmental impacts (Bruyn and de Opschoor 1997). In urban situations, as incomes grow, sulfur dioxide emissions first increase and then decrease, while carbon dioxide emissions increase with incomes. If something improves and something deteriorates, one reaction from the conventional economist might be to put weights or prices on such effects, in pursuit of the commensurability of values. However, the uncertainty and complexity of such situations (sulfur dioxide may counteract the greenhouse effect, for instance), and the fact that the price of externalities would depend on the outcome of social conflict, implies that the economist's accounts would be convincing only for the believers of the same school.

The pattern of use of environmental resources and services and the burdens of pollution depends on changing social structures and on power and income distribution. This leads to the field of political ecology originating in geography and anthropology, and which we define as the study of 'ecological distribution conflicts.' Economic growth leads to increased environmental impact and to increased conflicts (often outside the market sphere). Hence, for instance, the growth of the Environmental Justice movement in the United States. Examples abound of the failure of the price system to indicate such environmental impact, or (to use K. W. Kapp's idea), examples abound of cost–shifting successes. Thus, attempts at using cost–benefit analysis of the increased greenhouse effect (as in reports of the Intergovernmental Panel on Climate Change) are not convincing because of the arbitrariness of the discount rate (Azar and Sterner 1996, see below), and also because many items are not easily measured in physical terms, much less easily valued in money terms.

Moreover, the very pattern of prices in the economy would be different to start with, without the free access to carbon sinks. Should restrictions then be imposed on the 'ecological footprint' of rich economies or on the 'human appropriation of net primary production' (see below)? If equal property rights on carbon sinks are bestowed on everybody, there might still be a tendency for the price of carbon emissions to be low, according to the principle 'the poor sell cheap.' Everybody, except slaves, is the owner of her or his own body and health; however, poor people sell their health cheaply when working for a wage in mines or plantations. The free use of sinks has been modeled in a neo-Ricardian framework by C. Perrings, Martin O'Connor, and others, showing how the pattern of prices in the economy would be different, assuming different outcomes for ecological distribution conflicts.

Some remarks are still needed on the discount rate. Economists explain discounting of the future by subjective 'time-preference,' or because economic growth per capita caused by today's investments will make the marginal utility of consumption lower for our descendants than it is for us today. If discounting arises from the productivity of capital, and if such 'productivity' is a mixture of true increases in production and a lot of environmental destruction, then the discount factor should be the per capita rate of sustainable economic growth, minus the destruction of environmental resources and services. Now, in order to determine the present economic value of destruction caused by economic growth (such as loss of biodiversity, the filling up of carbon sinks, the production of radioactive waste and so on), we must not only put money figures on it (as discussed above), we need also a discount rate. Which is appropriate?

4. *Environmental Indexes of (Un)sustainability*

Because of the shortcomings of money valuation, ecological economists favor physical indicators and indexes in order to judge the overall impact of the human economy on the environment. Therefore, we here leave aside monetary corrections to GNP, such as those of El Serafy (see above), or Hueting. These calculate the economic costs of adjusting the economy to socially negotiated norms or standards of pollution and resource extraction, in a 'cost-effectiveness' approach (meaning the analysis of the cheapest instrument in money terms in order to adjust the economy to such physical norms or standards). We

also leave aside Cobb's and Daly's ambitious Index of Sustainable Economic Welfare (ISEW) (Daly and Cobb 1994), first calculated for the United States, which has inspired work in many countries, and whose end result is a figure in money terms strongly commensurable with GNP though often showing quite a different trend.

The main physical indexes of (un)sustainability discussed at present are as follows:

4.1 HANPP

HANPP (the human appropriation of net primary production) has been proposed by Vitousek et al. (1986). The net primary production (NPP) is the amount of energy that the primary producers, the plants, make available to the rest of living species, the heterotrophs. Of this NPP, humankind 'coopts' around 40 percent in terrestrial ecosystems. It is assumed that the higher the HANPP, the less biomass is available for 'wild' biodiversity. The proportion of NPP appropriated by humans is increasing because of population growth, and also because of increasing demands on land per person for urbanization, for growing foodstuffs, and for growing timber ('plantations are not forests' is a slogan of environmental activists in the tropics).

4.2 Ecospace

Ecospace and ecological footprints. Which is the environmental load of the economy, in terms of space? H.T. Odum posed this question, and later authors developed some answers. Rather than asking what maximum population a particular region or country can support sustainably, the question becomes: how large an area of productive land is needed (as source and sink) in order to sustain a given population indefinitely, at its current standard of living and with current technologies? Computations, not only for cities or metropolitan regions (whose ecological footprint is hundreds of times larger than their own boundaries) but for whole countries, show that some densely populated European countries (assuming per capita eco-footprints of 3 ha.) or Japan or Korea (with per capita eco-footprints of 2 ha.) occupy eco-spaces ten times larger than their own territories. This is the 'appropriated carrying capacity,' from which an 'ecological debt' arises. (For details, Wackernagel and Rees 1996.)

4.3 EROI

The term EROI, which stands for 'energy return on (energy) input,' also originates in H.T. Odum's work. Is there a trend towards an increasing energy-cost of obtaining energy (Hall et al. 1986)? The idea of looking at the basic economics of human society as a flow of energy is well known to ecological anthropologists (through Roy Rappaport's *Pigs for the Ancestors* and similar work). It goes back to Podolinski in 1880. Engels in 1882 exchanged correspondence with Marx on this topic. He denied the relevance of energy accounting for Marxian economics. Clearly, for an economy to be sustainable, the energy productivity of human work (i.e., how much energy is made available per day, by one day of human work) must be higher (or equal, if everybody is working) than the efficiency of the transformation of the energy intake into human work.

The energy productivity of a coal miner (wrote Podolinski) was much larger than that which a primitive agriculturalist could obtain, but this energy surplus from fossil fuels was transitory. Max Weber in 1909 had criticised Wilhelm Ostwald's interpretation of economic history in terms of (a) an increased use of energy, (b) an increased efficiency in the use of energy, because economic decisions on new industrial processes or new products were based on prices, entre preneurs did not pay attention to energy accounts *per se*. (No environmental auditing of firms was yet required.) Max Weber (whose book review against Ostwald was much praised by Hayek in later years), did not question energy prices from the environmental point of view as we would today.

In the early 1970s there were a number of studies on energy flow in agriculture, of which the best known were those of David Pimentel showing a decrease in energy efficiency in maize cultivation in the United Sate, because of the large energy input from outside agriculture itself. A new field was opened up by such studies on the efficiency in the use of energy, such as fuelwood, oil, and gas (Peet 1992), in different sectors of the economy, including the energy sector itself, taking also into account that increased energy efficiency might paradoxically lead to increased energy use, by reducing its cost (the Jevons effect). Such energy analysis has nothing to do, in principle, with the adoption of an 'energy theory of value,' or with the view that sources of energy are more problematic for sustainability than sinks for residues.

4.4 MIPs and DMR/TMR

The indicator of material input per unit service (MIPS) was developed at the Wuppertal Institute by Schmidt-Bleek. It adds up the materials used for production directly and indirectly (the 'ecological rucksack,') such as mineral ores, energy carriers like coal and oil, and all biomass (though not water, which is used in much larger amounts), including the whole 'life-cycle' down to the disposal or recycling phases.

This material input is measured in tons and it is compared with the services provided, sector by sector and, in principle, for the whole economy. For instance, to provide the service of one passenger per km, or to provide the service of living space of so many square metres, what amount of materials is involved, comparing different regions of the world, or historically. MIPS is useful as a measure of the material intensity of production but not as a measure of the toxicity of materials. The MIPS notion has been developed further in statistics published by the World Resources Institute in 1997 on the Direct Material Requirement and the Total Material Requirement (i.e., the aggregate tonnage, including in the TMR—the 'ecological rucksacks') coming into the economies of some countries such as the United States, Germany, The Netherlands, and Japan, both from domestic sources and from imports. This therefore tests the hypothesis of 'dematerialization' of production (Bunker 1996, Cleveland and Ruth 1998).

All the indexes mentioned here are measured in different units. How should a situation be judged in which, for instance, a synthetic indicator or index such as TMR improves while HANPP deteriorates, EROI decreases, and GNP grows? Commensurability would imply reducing such values to an encompassing supervalue but this is not necessary in order to reach reasonable judgments by a sort of macroeconomic multicriteria evaluation or integrated assessment (Faucheux and O'Connor 1998).

5. *The 'Dematerialization' of Consumption?*

In economic theories of production and consumption, compensation and substitution reign supreme. Not so in ecological economics, where diverse standards of value are deployed 'to take Nature into account' (O'Connor and Spash 1999). In the ecological economics theory of consumption, some goods are more important and cannot be substituted by other goods (economists call this a 'lexicographic' order of preferences). Thus, no other good can substitute or compensate for the minimum amount of endosomatic energy necessary for human life. This does not imply a biological view of human needs; on the contrary, the human species exhibits enormous intraspecific socially caused differences in the use of exosomatic energy. To call either the endosomatic consumption of 1,500 or 2,000 kcal or the exosomatic use of 100,000 or 200,000 kcal per person/day a 'socially constructed need, or want would leave aside the ecological explanations and/or implications of such a use of energy. Similarly, to call the daily endosomatic consumption of 1,500 or 2,000 kcal a 'revealed preference' would betray the conventional economist's metaphysical viewpoint.

There is another approach which, as pointed out by John Gowdy, builds upon the 'principle of irreducibility' of needs. This idea was proclaimed by Georgescu-Roegen in the previous edition of the *Encyclopedia of the Social Sciences*, article on 'Utility'. According to Max-Neef (Ekins and Max-Neef 1992) all humans have the same needs, described as 'subsistence,' 'affection,' 'protection,' 'understanding,' 'participation,' 'leisure,' 'creation,' 'identity,' 'freedom'; and there is no generalized principle of substitution among them. Such needs can be satisfied by a variety of 'satisfactors.'

Instead of taking the economic services as given, as in MIPS (passenger per km, square metres of living space), we may ask why there is so much travel, why so much building of houses with new materials instead of the restoration of old ones, and so on. In the late 1990s research was undertaken on the following question: is there a trend to use 'satisfactors' that are increasingly intensive in energy and materials, in order to satisfy predominantly nonmaterial needs? (Jackson and Marks 1999). The expectation that an economy that has less industry will be less resource intensive is perhaps premature. Input–output analysis of household lifestyles (by Faye Duchin and other authors) shows the high material and energy requirements of the consumption patterns of many of those employed in the postindustrial sector.

6. *Carrying Capacity and Neo-Malthusianism*

Many ecological economists have emphasized the pressure of population on resources. Has humankind exceeded its carrying capacity? This is defined in ecology as the maximum population of a given species, such as frogs in a lake, which can be supported sustainably in a given territory without spoiling its resource base.

However, the large differences internal to the human species in the exosomatic use of energy and materials mean that the first question that arises is; a maximum population at what level of consumption? Second, human technologies change at a quick pace. Boserup's thesis (1965) of endogenous technical change according to which preindustrial agricultural systems had changed in response to increases in population density, has long since turned the tables on the Malthusian argument. Third, the territories occupied by humans are not 'given.' Other species are pushed into corners or into oblivion (as the index HANPP implies). Internal to the human species, territoriality is politically constructed through State migration policies. Fourth, international trade, which is similar to horizontal transport in ecology, but which humans can consciously regulate, may imply ecologically unequal exchange, though if one territory lacks a very necessary item which is abundantly present in another territory, then Liebig's law of the minimum would recommend exchange. Then the joint carrying capacity of all

territories would be larger than the sum of the carrying capacities of all autarchic territories. This links up with the proposals of non-governmental organizations for fair and ecological trade.

Because of the shortcomings of 'carrying capacity' as an index of (un)sustainability for humans, the formula I = PAT has been proposed by Paul Ehrlich, where I is environmental impact, P is population, A is affluence per capita, and T stands for the environmental effects of technology. Efforts in the late 1990s were made to operationalize I = PAT. True, popultion remains one important variable. True also, neo-Malthusian population policies in the twentieth century caused many forced sterilizations and large-scale female infanticide in some countries, and they threatened small surviving ethnic groups. However, in the late nineteenth century, another neo-Malthusian movement in Europe and America (as shown by Francis Ronsin and other authors) opposed Malthus' view that poverty was due to overpopulation rather than social inequality, and fought successfully for limiting births by allowing women to exercise their reproductive rights, sometimes also appealing to ecological arguments of population pressure on resources. Human demography is self-conscious or reflective. Though it also follows Verhulst's curve, it is different from the ecology of a population of frogs in a lake.

7. *Final Remarks on Transdisciplinarity*

Ecological economics, based on methodological pluralism (Norgaard 1989) must not follow the reductionist road. Instead it should adopt Otto Neurath's image of the 'orchestration of the sciences,' acknowledging and trying to reconcile the contradictions arising between the different disciplines which deal with issues of ecological sustainability. Thus, how could a history of the industrialized agricultural economy be written taking into account both the viewpoints of conventional agricultural economics (technical progress, growth of productivity), and of agroecology (loss of biodiversity, decreased energy efficiency)? The image of the 'orchestration of the sciences' fits well with the ideas of coevolution and of emergent complexity; implying the study of the human dimensions of ecological change and therefore the study of human environmental perception. This means to introduce self-conscious human agency and reflective human interpretation in ecology. While 'emergent complexity' looks more to the unexpected future, 'coevolution' looks toward history.

Ecological economics as an 'orchestration of the sciences' takes into account the contradictions between the disciplines, it also takes into account the changing perceptions through history of the relations between humans and the environment, and it highlights the limits of the authoritative judgments of any particular expert in a particular discipline. This is not a technocratic or scientistic project. On the contrary, as explained by Funtowicz, Ravetz and other students of environmental risks, in many current problems of importance and urgency, where values are in dispute and uncertainties (not reducible to probabilistic risk) are high, we observe that 'certified' experts are often challenged by citizens from environmental groups. Examples of these are popular epidemiology activists inside the Environmental Justice movement in the United States, or debates on nuclear energy or on the labeling of new biotechnological foods, or arguments based on the practical knowledge of indigenous and peasant populations. This is postnormal science, leading toward participatory methods of conflict resolution and even toward discursive democracy, notions which are dear to ecological economists.

See also: Ecology, Political; Ecotourism; Environment and Development; Environmental and Resource Management; Environmental Economics; Environmental Planning; Environmental Vulnerability; Feminist Economics; Feminist Political Ecology; Lotka, Alfred James (1880–1949); Sustainability Transition: Human–Environment Relationship; Sustainable Development

Bibliography

Azar C, Sterner T 1996 Discounting and distributional considerations in the context of global warming. *Ecological Economics* **19**: 169–84

Berkes F, Folke C (eds.) 1998 *Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience*. Cambridge Univ. Press, Cambridge, UK

Boserup E 1965 *The Conditions of Agricultural Growth*. University of Chicago Press, Chicago

Bruyn S M, de Opschoor J B 1997 Developments in the throughput–income relationship: theoretical and empirical observations. *Ecological Economics* **20**: 255–68

Bunker S 1996 Raw materials and the global economy: oversights and distortions in industrial ecology. *Society and Natural Resources* **9**: 419–29

Cleveland C, Ruth M 1998 Indicators of dematerialization and the materials intensity of use. *Journal of Industrial Ecology* **2**: 15–50

Common M 1995 *Sustainability and Policy: Limits to Economics*. Cambridge Univ Press, Cambridge, UK

Costanza R (ed.) 1991 *Ecological Economics: The Science and Management of Sustainability*. Columbia Univ. Press, New York

Costanza R, Cleveland C, Perrings C (eds.) 1997a *The Development of Ecological Economics*. E Elgar, Cheltenham, UK

Costanza R, Cumberland J, Daly H, Goodland R, Norgaard R 1997b *An Introduction to Ecological Economics*. St. Lucie Press, Boca Raton, FL

Daly H, Cobb J 1994[1989] *For the Common Good. Redirecting the Economy toward Community, the Environment and a Sustainable Future*, 2nd edn. Beacon, Boston

Ekins P, Max-Neef M (eds.) 1992 *Real-life Economics*. Routledge, London

Faucheux S, O'Connor M (eds.) 1998 *Valuation for Sustainable Development*. E Elgar, Cheltenham, UK

Funtowicz S, Ravetz J 1994 The worth of a songbird: ecological economics as a post-normal science. *Ecological Economics* **10**: 197–207
Georgescu-Roegen N 1971 *The Entropy Law and the Economic Process*. Harvard University Press, Cambridge, MA
Hall C A S, Cleveland C J, Kaufman R 1986 *Energy and Resources Quality: the Ecology of the Economic Process*. Wiley, New York
Hueting R 1980 *New Scarcity and Economic Growth: More Welfare Through Less Production?* North Holland, Amsterdam, The Netherlands
Jackson T, Marks N 1999 Consumption, sustainable welfare, and human needs—with reference to UK expenditure patterns between 1954 and 199. *Ecological Economics* **28**: 421–441
Leipert C 1989 *Die heimlichen Kosten des Fortschritts*. S. Fischer, Frankfurt am Main
Martinez-Alier J, Schlupmann K 1987 *Ecological Economics: Energy, Environment and Society*. Blackwell, Oxford
Munda G 1995 *Multicriteria Evaluation in a Fuzzy Environment. Theory and Applications in Ecological Economics*. Physika Verlag, Heidelberg, Germany
Norgaard R 1989 The case for methodological pluralism. *Ecol. Econ.* **1**: 37–57
Norgaard R 1990 Economic indicators of resource scarcity. A critical essay. *Journal of Environmental Economics and Management* **19**: 19–25
O'Connor M, Spash C (eds.) 1999 *Valuation and the Environment*. E Elgar, Cheltenham, UK
O'Neill J 1993 *Ecology, Policy and Politics*. Routledge, London
Passet R 1996 [1979] *L'economique et le vivant* 2nd edn. Payat, Paris
Peet J 1992 *Energy and the Ecological Economics of Sustainability*. Island Press, Washington, DC
Perrings C 1987 *Economy and Environment: A Theoretical Essay on the Interdependence of Economic and Environmental Systems*. Cambridge Univ Press, Cambridge
Podolinski S A 1883 Menschliche Arbeit und Einheit der Kraft. *Die neue Zeit* **1** March–April
Rappaport R 1967 *Pigs for the Ancestors: Ritual in the Ecology of a New Guinea People*. Yale University Press, New Haven, CT
Tamanoi Y, Tsuchida A, Murota T 1984 Towards an entropic theory of economy and ecology—beyond the mechanistic equilibrium approach. *Economie appliquée* **37**: 279–94
Vitousek P, Ehrlich P, Erhlich A, Matson P 1986 Human appropriation of the products of photosynthesis. *Bioscience* **34**: 368–373
Wackernagel M, Rees W E 1996 *Our Ecological Footprint*. New Society Publ, Gabriola Island BC and Philadelphia, PA
Waring M 1988 If *Women Counted: A New Feminist Economics*. Harper & Row, San Francisco

J. Martinez-Alier

Ecological Fallacy, Statistics of

Researchers are said to commit the 'ecological fallacy' when they make untested inferences about individual-level relationships from aggregate data. That practice is called a fallacy because it is based on the problematic assumption that relationships that hold at one level of aggregation also hold at another level of aggregation. Researchers are subject to the ecological fallacy in virtually all the social and behavioral sciences, from history (where individual-level data often are unavailable) to criminology (where, for example, a positive relationship between poverty rate and crime rate across cities might be interpreted as evidence that poor people commit more crime) to epidemiology (where preliminary studies might correlate cancer rates across regions with other regional characteristics to decide what sorts of expensive individual-level data to collect).

This article cites examples of the ecological fallacy, explains why inference across levels of aggregation is sometimes unavoidable in social science research, and overviews solutions that have been proposed for alleviating the problem. Because individual-level data are not available for studying many issues of concern to social and behavioral scientists, the article emphasizes the theoretical conditions under which individual-level and aggregate-level relationships are the same.

1. An Example

One important contemporary example involves the relationship between race and voting behavior in the United States. The assumption is that voters have a propensity to vote for someone of their own race: whites are more likely to vote for a white candidate, blacks for a black candidate, and so on. If there is a propensity for intraracial voting, candidates are handicapped in political districts where their race is different from that of the majority of the voters. A probable outcome is that members of minority races will be under-represented in governing bodies because of their minority status in almost all political districts. One proposed solution is to use race as an explicit criterion for determining political districts, with the goal of drawing district boundaries in such a way that blacks and other minorities are in the majority in some of the election districts.

This issue poses the classic ecological fallacy problem for researchers because use of the secret ballot rules out the best source of data on how individuals actually voted. To make sound policy decisions about the use of race as a criterion for drawing up political districts, policy makers should know the strength of intrarace vote propensity, yet individual-level data typically are not available for estimating that propensity. Voting data come in aggregated form, by precincts. Hence researchers often must rely on correlations for voting data aggregated to the precinct level. In an election pitting a white candidate against a black candidate, for example,

researchers might attempt to use the correlation between a precinct's percentage of white voters and the percentage of votes for the white candidate to draw inferences about the propensity of whites to vote for white candidates.

Those inferences are problematic because relationships at the precinct level need not equal relationships at the individual level. Even the direction of the relationship can be different for individuals and aggregates. The 1968 US Presidential election provides a real-world example. In that election, George Wallace, a third-party candidate who was known primarily for his segregationist position as Governor of Alabama, captured about 10 percent of the national vote. What is relevant here is that in the South he tended to receive a higher percentage of the vote in congressional districts with higher percentages of blacks (the correlation across districts in the South was $r = 0.55$ (Firebaugh 1978)). Because no one believes that blacks were more inclined than other racial groups to vote for George Wallace, this is a compelling example of the danger of using aggregate data to draw inferences about the behavior of individuals.

Why was there a positive correlation between percent black and vote for Wallace in the 1968 election? The answer is that whites in heavily black districts were more likely to vote for Wallace than were whites in districts with fewer blacks. In other words, in this election the effect of one's race on one's voting behavior was affected by one's racial context—an example of a contextual effect (see *Contextual Studies: Methodology*). Subsequently this article shows that relationships differ at the individual and aggregate levels when such contextual effects are present.

2. *The Impact of the Ecological Fallacy on Research Practices*

Although the term ecological fallacy is most often associated with W. S. Robinson, Robinson does not use the term in his classic 1950 article. Instead he uses the term 'ecological correlation,' and he states that his analysis 'provides a definite answer as to whether ecological correlations can validly be used as substitutes for individual correlations. They cannot' (Robinson 1950, p. 357). Robinson uses illiteracy data from the 1930 US census to show the discrepancy between ecological correlations and individual correlations. The correlation between percent black and percent illiterate is $r = 0.77$ across States and $r = 0.95$ across nine larger geographic divisions of the USA. The size of those ecological correlations stands in stark contrast to the size of the individual-level relationship between race (black vs. others) and illiteracy ($r = 0.20$). As a second example, Robinson notes that the ecological correlation of percent foreign-born and percent illiterate across States is large and negative ($r = -0.53$), suggesting that the foreign-born are substantially less likely to be illiterate. Yet the foreign-born actually are more likely to be illiterate, as shown by the positive correlation at the individual level ($r = 0.12$).

Whereas Robinson focused on the danger of using aggregate data to draw conclusions about the behavior of individual people, the ecological fallacy problem in fact applies to inferences across levels of aggregation in general. Individuals need not be people. The terms individual level and aggregate level as used here refer to relative levels of aggregation. To illustrate, the term 'individual' could refer to a single church congregation within a population of congregations. The ecological fallacy problem would arise if a researcher tried to use data on racial diversity and attendance growth rates across church denominations (collections of congregations) to draw conclusions about the effect of racial diversity on local church growth. Because parishioners attend a local congregation, it is racial diversity at the local level, not at the denomination level, that is the natural focus here. One could of course survey local congregations in this instance, but the survey would ask questions about a collective (a congregation) rather than about a person.

It would be difficult to overstate the impact Robinson's article has had on social science research during the second half of the twentieth century. The use of ecological correlations to study individual-level relationships had been commonplace before Robinson's article, and the article sharply curtailed that practice. The article also served to motivate the development of survey research. If aggregate data are not adequate to study individuals, then social scientists need data on individuals. One efficient way to gather data on individuals is to ask them questions. So in this way Robinson's message about the need for individual-level data to study individuals no doubt played a role in the amassing of the large survey data sets that have become standard fare in social science research in the twenty-first century.

It is sometimes difficult or impossible to collect survey data to overcome the ecological fallacy problem, however, for two main reasons. First, individuals might be hard to reach. The extreme case of this occurs in historical research, where the pertinent individuals are dead. For this reason, historical research is especially plagued by the ecological fallacy problem (geography is another discipline especially affected, since geographic research is often based on spatial units). Aggregate data often are the only data available to historians, and it is not possible to go back in time to collect data to determine who voted for the Nazis, for example. Second, individuals might be reachable, but the needed information is sensitive. To protect the privacy of individuals, government agencies in some instances aggregate individual-level data before releasing the data to researchers (see *Aggregation:*

Methodology). But aggregation in turn creates the risk that researchers using the aggregate data might commit the ecological fallacy.

3. *Mitigating the Ecological Fallacy Problem*

No expert disputes Robinson's claim that aggregate-level relationships very often differ markedly from individual-level relationships, but a very active literature since Robinson takes issue with his strict prohibition against the use of aggregate data to study individual-level relationships. Necessity motivates this literature: Given the unavailability of individual-level data for many issues of interest in the social and behavioral sciences, the Robinson prohibition would rule out research in some areas of compelling interest, as illustrated by the earlier example of race and voting behavior.

Writing shortly after the Robinson broadside, Goodman (1953, 1959) observed that aggregation tends to affect correlation coefficients (r) more than it does regression slopes and that, under certain statistical conditions, aggregate-level and individual-level slopes are the same. The challenge has been to translate Goodman's statistical conditions into broadly applicable substantive terms that researchers can use. That challenge has resulted in extensive literatures on the ecological fallacy in sociology, geography, economics, political science, and history. One prominent approach seeks to spell out the conditions under which aggregate-level coefficients provide unbiased estimates of individual-level parameters. A second approach seeks to use a method-of-bounds to derive estimates of individual-level effects. This article concludes by briefly overviewing the two approaches.

3.1 *Equivalence Conditions for Aggregate-level and Individual-level Parameters*

Since the Robinson and Goodman articles in the 1950s, much of the methodological work on the ecological fallacy has been devoted to spelling out the conditions under which aggregate-level coefficients provide unbiased estimates of individual-level relationships. Following Goodman's lead, this work focuses on regression slopes, since the bias tends to be smaller for slopes than for correlation coefficients. Hannan and Burstein (1974) consider the effects of the variable by which the data are grouped (e.g., precincts, schools, States). For the bivariate case, they show that aggregate data yield unbiased estimates of individual-level relationships when the grouping variable is either (a) uncorrelated with the independent variable or (b) uncorrelated with the dependent variable, with the independent variable controlled for. Applied to the example of race and voting behavior, the precinct-level data provide unbiased estimates of individual-level voting behavior when precinct is unrelated to race, or precinct is unrelated to voting behavior, controlling for race. The first condition is improbable, since neighborhoods in the USA tend to be highly segregated by race. The second condition (voting behavior is unrelated to precinct, race controlled for) means that one's propensity to vote for a same-race candidate is unaffected by the racial makeup of one's precinct. Thus whites in majority-black precincts are as likely to vote for the white candidate as whites in majority-white precincts, blacks in majority-white precincts are as likely to vote for the black candidate as blacks in majority-black precincts, and so on. Unless Americans of all races are remarkably oblivious to their social contexts, this assumption is unlikely to hold.

Firebaugh (1978) subsumes the Hannan–Burstein conditions under a single rule. The rule states that the aggregate-level regression slope provides an unbiased estimate of the corresponding individual-level effect of X on Y if and only if aggregated X (denoted $\bar{X}$ since aggregated X is usually a type of mean) is unrelated to Y, controlling for individual-level X. For example, using precinct-level data to estimate the effect of one's race on one's vote will yield a biased estimate of that effect unless the racial context of precincts is unrelated to one's vote. Firebaugh shows that the $\bar{X}$-rule generalizes to the multivariate case where a researcher wants to estimate the effect of more than one independent variable.

Firebaugh's $\bar{X}$-rule suggests one solution to the ecological fallacy problem: eliminate $\bar{X}$ effects. To eliminate $\bar{X}$ effects, researchers must determine why $\bar{X}$ might have an effect on Y beyond the effect of individual X, and then add variables to try to remove the effect. Using the voting example to illustrate: the first step is to ask why it is that the racial composition of one's precinct might be related to one's vote, independent of one's race. The most likely answer is that minorities in largely white precincts differ from minorities in largely nonwhite precincts on a host of factors, *other than race*, that are related to voting—characteristics such as income, education, and home ownership. These non-racial characteristics of the precincts should be entered into the aggregate-level equation as control variables. Here the logic of control is the conventional logic used in nonexperimental analysis: Regressors are included (average income in the precinct, average education level, percent homeowners, etc.) in an attempt to use the least-squares method to 'equate' the precincts on those properties. If the method is successful, the aggregate-level slope for race will reflect the effect of race for voters who live in precincts that are otherwise equivalent—that is, the aggregate-level slope will reflect the effect of race itself on vote.

3.2 *Bounding Individual-level Parameters*

The $\bar{X}$ -rule is not a foolproof solution to the ecological fallacy problem since there is no way (with aggregate

data only) to test the assumption that $\bar{X}$ effects have been eliminated. So methodologists continue to work on the problem. One line of work gains leverage on the problem by applying the insight that marginal frequencies bound the cell frequencies in a contingency table (Duncan and Davis 1953). For example, in a precinct where 500 whites and 100 nonwhites collectively cast 450 votes for candidate A, the number of nonwhites voting for candidate A is bounded by 0 and 100 and the number of whites voting for candidate A is bounded by 350 and 450.

A book by King (1997) entitled *A Solution to the Ecological Inference Problem* combines the bounding principle of contingency tables with probabilistic statistical methods to fashion a sequential procedure for estimating individual-level coefficients and their confidence intervals. In essence the method begins with the deterministic information provided by the bounds for each of the aggregates (e.g., precincts) and then applies likelihood methods to narrow probabilistically within those bounds. A public-domain computer program provides estimates, visual depictions, and diagnostic tests. King's method is the most ambitious attempt to date to provide a general solution to the ecological fallacy problem. Ultimately, the worth of a tool is proven in use, and the most telling test of King's method and its successors will come as researchers apply them.

See also: Aggregation: Methodology; Causal Inference and Statistical Fallacies; Ecological Inference; Probabilistic Thinking; Secondary Analysis: Methodology

Bibliography

Achen C H, Shively W P 1995 *Cross-Level Inference*. University of Chicage Press, Chicago

Duncan O D, Davis B 1953 An alternative to ecological correlation. *American Sociological Review* **18**: 665–6

Freedman D, Klein S, Sacks J, Smyth C, Everett C 1991 Ecological regressions and voting rights. *Evaluation Review* **15**: 673–711

Firebaugh G 1978 A rule for inferring individual-level relationships from aggregate data. *American Sociological Review* **43**: 557–72

Goodman L 1953 Ecological regressions and behavior of individuals. *American Sociological Review* **18**: 663–4

Goodman L 1959 Some alternatives to ecological correlation. *American Journal of Sociology* **64**: 610–25

Hannan M, Burstein L 1974 Estimation from grouped observations. *American Sociological Review* **39**: 374–92

King G 1997 *A Solution to the Ecological Inference Problem*. Princeton University Press, Princeton, NJ

Robinson W S 1950 Ecological correlations and the behavior of individuals. *American Sociological Review* **15**: 351–7

G. Firebaugh

Ecological Imperialism

Ecological imperialism refers to the environmental aspects of the political domination of territorial areas and subjugated peoples. It is concerned with the consequences to imperialism of environmental conditions and processes. Ecological imperialism has been central to the geographical expansion and historical rule of empires, most notably those of Europe in the sixteenth to twentieth centuries, and to their present-day environmental legacies (Crosby 1986). The term also has been applied to the environmental dimension of economic and political domination of poorer countries by global superpowers and institutions (MacKenzie 1990).

Ecological imperialism includes: (a) the geographical introduction of non-native organisms such as disease pathogens, crops, and livestock and, during recent years, the transfer of toxic waste, nuclear contaminants, and other hazardous materials; (b) the ecological interaction of the introduced organisms and materials with human activities that incur environmental changes, often degradation; (c) the environmental role of imperial institutions, historically often those belonging to colonialism; (d) the work of the environmental sciences in the context of imperialism, in some cases for the goal of conservation; and (e) the use and ideas of the environment by persons and groups that resist the rule of empire.

Geographical introduction of nonnative organisms is a facet of ecological imperialism that is based on the dynamics of biogeography or the distribution of plants and animals. The spread of non-native organisms, or 'colonizers,' was a concomitant of the European empires (Crosby 1986). Introduced organisms were brought both by accident and by design. Disease organisms resulted in massive epidemics and mortality among non-Europeans. Mortality due to pathogens that were introduced in imperial exploration and colonialism of the Europeans resulted in the deaths of 25–50 million American Indians, whose immunities did not include such Old World diseases as smallpox, measles, typhus, and influenza (Denevan 1992). Another avenue of ecological imperialism was the introduction of crops, food plants, and livestock. Ecological ease of the transfer of major economic organisms—such as wheat, pigs, cattle, and sheep—was a key to success of the European empires in expanding their rule over the 'Neo-Europes' of temperate North and South America, southern Africa, New Zealand, and Australia (Crosby 1986, Gade 1999). Not least, the ecological colonization of invasive weeds and animal vermin, such as nettles and black rats, actively undermined the living habits, resource use, and production systems and social organization of the subjects of European imperialism.

The impacts of ecological imperialism are influenced by human ecosystems, that is, the ecological inter-

actions among people and their biotic and abiotic surroundings in a given place (i.e., human and cultural ecology). Common interactions are associated with cultivation, land and resource use, and other sorts of environment-related activities that have occurred in the expansion and rule of empires. Dynamics of human ecosystems determined the actual impacts of introduced diseases, crops and livestock, and weeds and vermin. Often these effects were set amid general environmental degradation, such as worsened soil erosion and fertility loss, deterioration of water supplies, and the decline of vegetation resources due to overgrazing and extensive clearing. Some human ecosystems allied to ecological imperialism are thought to have been consonant with environmental conservation. The colonial wheat and livestock-raising customs of the Spanish Empire, for example, may have been environmentally sound in certain places in Mexico (Turner and Butzer 1992, Sluyter 1999).

Ecological imperialism involves the power of particular institutions and social groups, which is a focus of the perspectives of environmental history and political ecology (Grove 1996, Mackenzie 1990). Imperial institutions typically governed the ownership or allocation and use of arable resources such as farmland and irrigation. Also, protection was common under colonial conservation institutions. Conservation policies and resource control were aimed at soil protection and irrigation works as well as nonarable resources such as forests, grazing and rangelands, and game reserves. The motivation for colonial conservation included the use and appropriation of resources for business and the needs of the imperial state; the interests of states and rulers in limiting or localizing degradation; and cultural concerns symbolized by images of Eden and Paradise. Some imperial conservation has been deeply influenced by the concerns and agendas of modern science. For example, the colonial scientists in the employ of European empires in India, southern Africa, and on oceanic islands (e.g., Mauritius, St. Helena, Barbados) were driven by the fear of resource scarcity. This concern in the context of ecological imperialism, or 'Green Imperialism,' contributed to the birth of modern environmentalism (Grove 1996).

Ecological imperialism is also concerned with the resistance and opposition to authoritarian rule. Imperial subjects have both reacted against imposed environmental restrictions and created ecological customs as part of political alternatives. Colonial subjects in Africa, India, and Asia drew on their customs of a moral or affective economy in protesting or refusing to comply with conservation policies involving soil protection, game reserves, and water diversions (Neumann 1998). Slaves and colonial Indians in North and Latin America created customs of food plant use and production with environmental capabilities that enabled their resistance against the administrative fiats of European colonialism (Carney 2001, Zimmerer 1996). These investigations are a broadening of the study of ecological imperialism that extends beyond either moral outrage with denunciation or mere justification and prescription.

See also: Ecological Economics; Ecology and Health; Environment and Development; Environmental Health and Safety: Social Aspects; Environmental Risk and Hazards; Environmental Vulnerability; Land Degradation

Bibliography

Carney J 2001 *Black Rice*. Harvard University Press, Cambridge, MA

Crosby A W 1986 *Ecological Imperialism: The Biological Expansion of Europe, 900–1900*. Cambridge University Press, Cambridge, UK

Denevan W M (ed.) 1992 *The Native Population of the Americas in 1492*, 2nd edn. University of Wisconsin Press, Madison, WI

Gade D W 1999 *Nature and Culture in the Andes*. University of Wisconsin Press, Madison, WI

Grove R 1996 *Green Imperialism*. Cambridge University Press, Cambridge, UK

MacKenzie J M (ed.) 1990 *Imperialism and the Natural World*. Manchester University Press, Manchester, UK

Neumann R 1998 *Imposing Wilderness: Struggles over Livelihood and Nature Preservation in Africa*. University of California Press, Berkeley, CA

Sluyter A 1999 The making of the myth in postcolonial development: Material-conceptual landscape transformation in sixteenth-century Veracruz. *Annals of the Association of American Geographers* **89**(3): 377–401

Turner B L II, Butzer K 1992 The Columbian encounter and land-use change. *Environment* **34**(8): 16–20, 38–44

Zimmerer K S 1996 *Changing Fortunes: Biodiversity and Peasant Livelihood in the Peruvian Andes*. University of California Press, Berkeley, CA

K. S. Zimmerer

Ecological Inference

In nineteenth century Europe, suicide rates were higher in countries that were more heavily Protestant, the inference being that suicide was promoted by the social conditions of Protestantism (Durkheim 1897); also see Neeleman and Lewis (1999). According to Carroll (1975), death rates from breast cancer are higher in countries where fat is a larger component of the diet, the idea being that fat intake causes breast cancer. These are 'ecological inferences,' that is, inferences about individual behavior drawn from data about aggregates.

To continue with Durkheim, the Protestant countries were different from the Catholic countries in many ways besides religion (the problem of 'confounding'). Moreover, his data do not tie individual suicides to any particular religious faith. Confounding must be dealt with in any observational study. But the second problem—that exposure and response are measured only for aggregates rather than for individuals—is specific to ecological studies. If there is no confounding, the expected difference between effects for groups and effects for individuals is 'aggregation bias'; in general, the difference is partly attributable to confounding and partly to aggregation bias.

The ecological fallacy consists of thinking that relationships observed for groups necessarily hold for individuals: if countries with more Protestants have higher suicide rates, then Protestants must be more likely to commit suicide; if countries with more fat in the diet have higher rates of breast cancer, then women who eat fatty foods must be more likely to get breast cancer. These inferences may be correct, but are only weakly supported by the aggregate data.

Ecological studies in epidemiology yielded important insights for Snow (1855), Finlay (1881), Goldberger (Terris 1964), and Dean (1938) among others. However, it is all too easy to draw incorrect conclusions from aggregate data. Greenland and Robins (1994) review the issues. For one example, recent studies of individual-level data cast serious doubt on the link between breast cancer and fat intake (Holmes et al. 1999). Another well-known example, on the sources of popular support for the Nazi party pre-war Germany, is discussed by Lohmoller et al. (1985)

1. Ecological Correlation

Robinson (1950) discusses ecological inference, stressing the difference between ecological correlations and individual correlations. One striking example is the relationship between nativity and literacy. For each of the 48 states in the US of 1930, Robinson computes two numbers: the percentage of the population who are foreign-born, and the percentage who are literate. (The base for each percentage is the number of residents of the state who are 10 or older; data are from the Census of 1930.) The correlation between the 48 pairs of numbers is 0.53. This is an 'ecological' correlation, because the unit of analysis is not an individual person but a group of people—the residents of a state. The ecological correlation suggests a positive association between foreign birth and literacy: the foreign-born are more likely to be literate (in American English) than the native-born. In reality, the association is negative: the correlation computed at the individual level is -0.11. The ecological correlation gives the wrong inference. The sign of the correlation is positive because the foreign-born tend to live in states where the native-born are relatively literate.

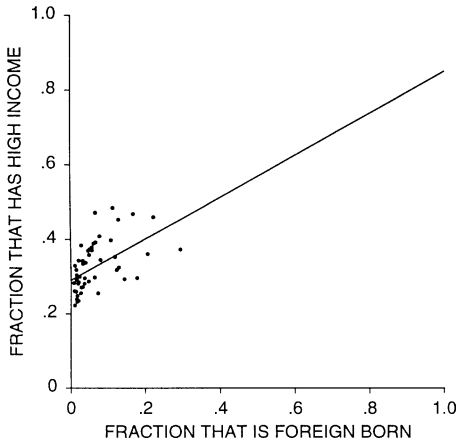

Figure 1
For each state, the horizontal axis shows the fraction of the population aged 25 and over that is foreign-born; the vertical axis shows the fraction having high incomes. The regression line is also shown

The example can be replicated using data from the March 1995 Current Population Survey, restricted to persons age 25 and over. Literacy is not reported, but incomes are. Say that a person has 'high' income if family income in 1994 was $50,000 or more. Figure 1 shows for each of the 50 states the fraction of persons who are foreign-born, and the fraction of persons who have high incomes. The regression line is shown too. The ecological correlation (across the 50 areas) is 0.52, suggesting that the foreign-born have higher incomes on the whole than the native-born. The truth is the opposite. About 35 percent of the native-born have high incomes, compared to 28 percent for the foreign-born. The correlation at the individual level is -0.05. In Fig. 1, there is no state where the fraction of foreign-born approaches 1. However, the same sort of reversal can occur even with a full spectrum of x-values. The issue is the disaggregation, not the range of the data.

2. Ecological Regression

Under certain restrictive assumptions, individual behavior can be inferred from aggregate data using a technique called 'ecological regression.' The unit of analysis in the regression is a group of people, typically defined by geography (as in Fig. 1). The technique has been widely used in voting rights litigation in the US. For discussion from various perspectives, see Dogan and Rokkan (1969), Freedman et al. (1991), Grofman and Davidson (1992), Achen and Shively (1995), and Cho (1998).

It may, for instance, be desired to estimate the support for a particular candidate among Hispanics and non-Hispanics. For each precinct in an electoral contest, suppose we know the fraction x of voters who are Hispanic; we also know the fraction y of votes obtained by the candidate of interest. We would not know the fraction of Hispanic voters who voted for the candidate, due to the secrecy of the ballot.

A regression equation can be fitted to the data:

$$y_i = a + bx_i + \varepsilon_i \tag{1}$$

In Eqn. (1), the subscript i indexes precincts; x_i is the fraction of voters in precinct i who are Hispanic; y_i is the fraction of votes cast in that precinct for the candidate of interest. (More exactly, y_i is the number of votes for that candidate, divided by the number of voters.) The precinct-level disturbance terms ε_i would typically be assumed to be independent and identically distributed with mean 0, although this is seldom made explicit.

The parameters a and b in Eqn. (1) can be estimated by ordinary least squares; call the estimates $\hat{a}$ and $\hat{b}$, respectively. Then $\hat{a}$ would be interpreted as the fraction of non-Hispanic voters who supported the candidate in question: after all, $\hat{a}$ is the height of the regression line at $x = 0$, corresponding to precincts with no Hispanic voters. Likewise, $\hat{a}+\hat{b}$ (the height of the regression line at $x = 1$) would be interpreted as the fraction of Hispanic voters who supported the candidate.

The data are available for groups defined by area of residence; the inference is to groups defined by ethnicity. The statistical logic connecting the inference to the data rests on the 'constancy assumption,' that voting preferences within ethnic groups do not systematically depend on the ethnic makeup of the area of the residence (Goodman 1953, 1959). In Fig. 1, the regression equation is $y = 0.29 + 0.56x$, leading to the inference that 30 percent of the native-born have high incomes, compared to $29 + 56 = 85$ percent of the foreign-born. In reality, as noted above, the percentages are 35 and 28 percent. The ecological inference is in error, because the constancy assumption fails: the incomes of the native-born increase systematically with fraction of foreign-born in the state. Immigration to the US tends to concentrate in richer states—California, Hawaii, and New York rather than Kentucky, Tennessee, and West Virginia.

3. *The Method of Bounds*

In 1995 in the state of Washington, 8 percent of the population was foreign born, while 34 percent had high incomes. These are known. Let p be the fraction of the foreign-born who had high incomes, and let q be the corresponding fraction for the native-born. Investigators making ecological inferences would typically not know the fractions p and q. The two unknowns satisfy an equation:

$$0.08p + (1 - 0.08)q = 0.34 \tag{2}$$

We have one equation with two unknowns: that is the whole problem of ecological inference. However, Eqn. (2) does contain some information, because $q = (0.34 - 0.08p)/(1 - 0.08)$. Since p is bounded between 0 and 1, it follows that q is bounded below by $(0.34 - 0.08)/(1 - 0.08) \doteq 0.26$. Similarly, the upper bound is $0.34/0.92 = 0.37$. This is 'the method of bounds' (Duncan and Davis 1953). Such bounds can be computed for each study area, and then aggregated. However, in typical applications, the bounds are too broad to be informative (also see Goodman 1959).

4. *Ecological Regression with Random Coefficients*

Models with random coefficients have been used to make ecological inferences. By way of illustration, we consider a model for nativity and income. Index the study areas by i. As before, let x_i be the fraction of the population in area i that is foreign born, and y_i the fraction with high incomes. These fractions are known, as is the total population n_i of area i. Let p_i be the fraction of the foreign-born with high incomes, and let q_i be the corresponding fraction for the native-born, so

$$y_i = p_i x_i + q_i(1 - x_i) \tag{3}$$

The problem is to make inferences about p_i and q_i. To solve such problems, King (1997) assumes the pairs (p_i, q_i) to be independent and identically distributed across the study areas. This is his version of the constancy assumption: the statistical behavior of a demographic group is not allowed to depend on area of residence. The parent distribution is taken to be bivariate normal, conditioned to lie in the unit square so that p_i and q_i are both between 0 and 1. The five parameters of this parent normal distribution can be estimated from the data by maximum likelihood, and estimates $(\hat{p}_i, \hat{q}_i)$ can be derived from Eqn. (3). Rates for each demographic group can then be estimated by addition over areas, as $\sum_i n_i \hat{p}_i / \sum_i n_i x_i$ and $\sum_i n_i \hat{q}_i / \sum_i n_i (1 - x_i)$, respectively. More complex models have also been developed.

To show the force of the constancy assumption, we can use the 'neighborhood model' which assumes that behavior is determined by geography not demography. Then $\hat{p}_i = \hat{q}_i = y_i$ for each study area i. (In our example of nativity and income, if 33 percent of the residents of a particular study area have high incomes,

this percentage applies equally to the foreign-born and the native-born in that area.) The neighborhood model turns the constancy assumption on its head. In a variety of test applications where all the data are available, the neighborhood model gives more accurate estimates for demographic groups than ecological regression or random-coefficients models. In some contexts, of course, the neighborhood model proves deficient (Achen and Shively 1995). If the data are incomplete so estimation is needed, it will be unclear which model if any is giving the right answers. The track record to date favors the neighborhood model (Freedman et al. 1991, 1998). For another perspective see Schuessler (1999).

5. *Summary and Conclusions*

Aggregate data are often easier to obtain than data on individuals, and may offer valuable clues about individual behavior. Ecological inferences will therefore continue to be made. The problems of confounding and aggregation bias, however, are unlikely to be resolved in the proximate future.

See also: Causal Inference and Statistical Fallacies; Ecological Fallacy, Statistics of; Ecological Imperialism

Bibliography

Achen C H, Shively W P 1995 *Cross-Level Inference*. University of Chicago Press, Chicago

Carroll K 1975 Experimental evidence of dietary factors and hormone-dependent cancers. *Cancer Research* **35**: 3374–83

Cho W K T 1998 Iff the assumption fits: A comment on the King ecological inference solution. *Political Analysis* **7**: 143–63

Dean H T 1938 Endemic fluorosis and its relation to dental caries. *Public Health Reports* **53**: 1443–52

Dogan M, Rokkan S (eds.) 1969 *Quantitative Ecological Analysis in the Social Sciences*. MIT Press, Cambridge, MA

Duncan O D, Davis B 1953 An alternative to ecological correlation. *American Sociological Review* **18**: 665–6

Durkheim E 1897 *Le suicide*. F. Alcan, Paris [English translation by J A Spalding, 1951, Free Press, Collier-MacMillan, Toronto, Canada]

Finlay C J 1881 The mosquito hypothetically considered as the agent of transmission of yellow fever. *Anales de la Academia de Ciencias Medicas, Físicas y Naturales de Habana* XVIII [English translation by Finlay, reprinted in Buck C, Llopis A, Nájera E, Terris M 1989 *The Challenge of Epidemiology: Issues and Selected Readings,* Scientific Publication No. 505, World Health Organization, Geneva]

Freedman D A, Klein S P, Ostland M, Roberts M R 1998 Review of A Solution to the Ecological Inference Problem. *Journal of the American Statistical Association* **93**: 1518–22; with discussion, 1999, **94**: 352–7

Freedman D A, Klein S P, Sacks J, Smyth C A, Everett C G 1991 Ecological regression and voting rights. *Evaluation Review* **15**: 659–817 (with discussion)

Goodman L 1953 Ecological regression and the behavior of individuals. *American Sociological Review* **18**: 663–4

Goodman L 1959 Some alternatives to ecological correlation. *American Journal of Sociology* **64**: 610–25

Greenland S, Robins J 1994 Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. *American Journal of Epidemiology* **139**: 747–60

Grofman B, Davidson C 1992 *Controversies in Minority Voting: The Voting Rights Act in Perspective*. Brookings Institution, Washington, DC

Holmes M D, Hunter D J, Colditz G A, Stampfer M J, Hankinson S E, Speizer F E, Rosner B, Willett W C 1999 Association of dietary intake of fat and fatty acids with risk of breast cancer. *Journal of the American Medical Association* **281**: 914–20

King G 1997 *A Solution to the Ecological Inference Problem*. Princeton University Press, Princeton, New Jersey

Lohmoller J B, Faller J, Link A, de Rifke J 1985 Unemployment and the rise of national socialism: contradicting results from different regional aggregations. In: Nijkamp P (ed.) *Measuring the Unmeasurable*. Martinus Nijhoof, Den Haag, pp. 357–70

Neeleman J, Lewis G 1999 Suicide, religion and socioeconomic conditions. An ecological study in 26 countries, 1990. *Journal of Epidemiology and Community Health* **53**: 204–10

Robinson W S 1950 Ecological correlations and the behavior of individuals. *American Sociological Review* **15**: 351–7

Schuessler A A 1999 Ecological inference. *Proceedings of the National Academy of Sciences USA* **96**: 10578–81

Snow J 1855 *On the Mode of Communication of Cholera*. Churchill, London. Reprinted by Hafner, New York, 1965

Terris M 1964 *Goldberger on Pellagra*. Louisiana State University Press, Louisiana

D. A. Freedman

Ecology and Health

1. *Ecology and Health*

The term 'ecology' pertains broadly to the interrelations between organisms and their environments (Hawley 1950). From its early roots in biology, the ecological paradigm has evolved within several disciplines (e.g., sociology, psychology, economics, and public health) to provide a general framework for understanding people's transactions with their physical and sociocultural surroundings (e.g., Barker 1968, Bronfenbrenner 1979, Park et al. 1925). The field of ecology and health examines the direct and interactive effects of conditions in the physical and sociocultural environment on the physiological, emotional, and social well-being of individuals and groups (cf., Duhl 1996, Moos 1979, Stokols 1996). Ecological analyses provide more integrative and comprehensive explanations for the joint influence of biogenetic, psychological, behavioral, sociocultural, and physical environmental factors on human well-being, than do narrowly-gauged biomedical, behavioral, and environmental theories.

2. Emergence of Ecological Perspectives on Health and Illness

The emergence of an ecological paradigm for understanding the determinants of health and illness in human populations can be traced to research developments in several scientific fields. During the 1920s and 1930s, a series of correlational analyses were conducted by researchers in the Chicago School of human ecology, and revealed systematic associations between community rates of illness and crime, and physical and social conditions of the urban environment including poor-quality housing and high population densities within residences and neighborhoods (cf., Faris and Dunham 1939, Park et al. 1925). These correlational studies in the fields of human ecology and sociology suggested a strong link between conditions of the urban environment and residents' health, but did not provide an integrative theoretical account of the causal pathways connecting those variables.

An important theoretical analysis of the links between the social environment and illness etiology was provided by John Cassel in the field of public health. Cassel's (1964) conceptualization of *social epidemiology* signaled a shift from 'germ-theory' accounts of health and illness (focusing on the health-impairing effects of specific pathogens once they invaded a human 'host'), toward a more ecologically-oriented, contextual perspective emphasizing the effects of one's social environment (e.g., work overload, sources of interpersonal stress) on immune functioning and host resistance to pathogenic agents. Cassel's formulation of social epidemiology during the 1960s and his hypotheses concerning the impact of social and cultural factors on human health have been corroborated empirically in subsequent studies linking job strain, social isolation, and interpersonal stress to a variety of illness conditions and premature death (cf., Berkman and Syme 1979, Karasek and Theorell 1990, Marmot et al. 1997).

Publication of the *US Surgeon General's Reports* in 1964 and 1979, documenting the deleterious effects of smoking on health, prompted the rapid growth of health promotion and disease prevention research during the 1980s (cf., US Public Health Service 1979). Much of the early research in this field was rooted in the concepts and methods of health education, and focused largely on behavioral change strategies aimed at fostering healthier lifestyles and health habits among individuals (cf., Belloc and Breslow 1972). Targeted efforts to enhance individuals' health behaviors have led to reduced use of tobacco products among adults, lower rates of alcohol-related automobile deaths, moderate gains in the proportion of adults exercising regularly and eating less fatty diets, and reduced employee health costs at the worksite (cf., McGinnis and Lee 1995, Pelletier 1996). At the same time, however, certain health risks, such as exposure to community violence, substance abuse, and lack of medical insurance, remain segmented in hard-to-reach 'pockets of heightened prevalence' (Fisher 1995)—especially among low-income and minority groups in the population (cf., Adler et al. 1994).

Growing recognition of the limitations associated with behavioral change programs has prompted a paradigm shift in the health promotion field, away from individually-focused interventions toward more comprehensive, ecological formulations that address interdependencies among socioeconomic, political, environmental, organizational, biological, and psychological determinants of health and illness (cf., Kickbush 1989, Winett et al. 1989). An explicit emphasis on ecological strategies of health promotion is evident in the design and implementation of broader-gauged community interventions aimed at reducing handgun violence, smoking prevalence, teen pregnancy, and HIV infection in recent years (cf., Breslow and Johnson 1993, Schneider Jamner and Stokols 2000). A common feature of these programs is that they incorporate multiple levels of analysis and intervention components—including political action, legislative reform, media advocacy, as well as behavioral change and environmental enhancement efforts—to address key contextual factors that substantially influence individual and population health.

One other major development that has prompted greater interest in ecology and health in recent years is the discovery of adverse global environmental changes—for instance, the depletion of atmospheric ozone and the trend toward global warming—which have the capacity to elevate population morbidity and mortality rates at worldwide as well as regional and national levels (cf., Stern et al. 1992). Potential health impacts of these global changes include elevated skin cancer rates resulting from ozone depletion and individuals' heightened exposure to ultraviolet rays; and the food shortages, nutritional ailments, and respiratory illnesses associated with global warming (cf., Leaf 1989). The discovery of these global environmental threats to population health has prompted greater international collaboration in the development of ecologically-oriented health promotion programs through the World Health Organization's Healthy Cities Program (cf., Duhl 1996).

3. Core Principles of Ecology and Health Research

Research on ecology and health encompasses several disciplines, including medicine, public health, urban planning, environmental design, public policy, and the behavioral and social sciences. The scientific contours of this research are not easily delimited, due to the interdisciplinary scope of the field (cf., Stokols 2000). The unique concerns of ecology and health can be better understood in terms of the overarching con-

ceptual principles that underlie this field, than by searching for a clearly defined body of research organized around this topic. The core principles underlying ecology and health research are drawn largely from programmatic statements about the 'New Public Health' and the conceptual and methodological assumptions of systems theory and social ecology (cf., Duhl 1996).

3.1 Intellectual Antecedents: Systems Theory and the Ecological Paradigm

The 'New Public Health' outlined in the Ottawa Charter for Health Promotion (1986) called for the development of comprehensive, ecologically-oriented approaches to the study and prevention of disease (cf., Kickbush 1989). For instance, the Ottawa Charter gave explicit attention to social causes of illness, as well as the physical-environmental health threats that exist in many communities. Subsequent analyses of ecology and health extended these programmatic ideas by providing a set of conceptual and methodological principles derived largely from systems and ecological theories, for organizing and evaluating community-based health promotion programs (cf., Green and Ottoson 1999).

The field of ecology and health incorporates a number of concepts derived from systems theory (e.g., interdependence, homeostasis, negative feedback, deviation amplification; cf., Emery and Trist 1972, Maruyama 1963) to understand the interrelations among people and their environments. Systems analyses suggest that the healthfulness of particular settings and the well-being of their participants are jointly influenced by multiple facets of the physical environment (e.g., geography, architecture, technology) and the social environment (e.g., culture, economics, politics). The health status of individuals and groups is also influenced by personal attributes such as genetic heritage, psychological dispositions, and behavioral patterns. From the vantage point of systems theory, efforts to promote well-being should be based on an understanding of the relationships among diverse environmental and personal factors, rather than on analyses that focus exclusively on environmental, biological, or behavioral factors (cf., Moos 1979).

3.1.1 Research findings from ecological analyses of health. The direct and interactive effects of environmental and personal attributes on health have been observed in several lines of research. For example, a substantial body of research highlights the ways in which poverty, inequality of income distribution, and minority status jointly undermine the health of individuals and vulnerable subgroups in the population (cf., Adler et al. 1994, Bullard 1990, Kaplan et al. 1996). Other studies indicate that neighborhood socioeconomic status (percent of households in a geographic area receiving public assistance) is associated with poorer self-reports of health, above and beyond individual socioeconomic status (Malmstrom et al. 1999, Robert 1998). Interactive effects of environmental and behavioral risk factors are exemplified by the findings that environmental exposure to asbestos exacerbates chronic smokers' risks of developing lung cancer (Grunberg 1991), and that an individual's disposition toward exerting internal control over his/her environment buffers the potentially negative effects of low education and income on health status (Lachman and Weaver 1998).

Systems theory also posits that people–environment transactions are characterized by cycles of mutual influence, whereby the physical and social features of settings directly affect occupants' health and, concurrently, the participants in settings modify the healthfulness of their surroundings through their individual and collective actions (or, alternatively, adopt an increasingly passive stance toward their environment). For example, individuals' routine exposure to community violence in some low-income neighborhoods can provoke feelings of disempowerment, helplessness, and depression, which in turn diminish their efforts to adopt health promotive practices related to physical activity and dietary change (cf., Sanders-Phillips 2000). Similarly, the well-being of elderly individuals depends on the degree of fit between their functional abilities, on the one hand, and environmental constraints on their daily activities, on the other. Either or both of these elements can be modified to promote higher levels of person–environment fit (cf., Lawton 1999).

Ecologically-oriented studies of health also incorporate certain conceptual and methodological principles drawn from the field of social ecology. For instance, ecological theory highlights the multidimensional nature of human environments. Environmental settings can be characterized in terms of their physical and social components, their objective (actual) or subjective (perceived) qualities, and their scale or immediacy to individuals and groups (proximal vs. distal). Moreover, the participants in environmental settings can be studied at varying levels ranging from individuals, small groups and organizations, to larger communities and populations. Rather than focusing solely on individuals or aggregates, the social ecological perspective incorporates multiple levels of analysis and diverse methodologies (e.g., medical exams, questionnaires, behavioral observations, environmental recordings, epidemiologic analyses) for assessing the healthfulness of settings and the well-being of individuals and groups.

Social ecological analyses conceptualize human environments as complex systems in which local settings and organizations are nested within more complex and remote regions. Accordingly, efforts to promote human well-being must take into account the

interdependencies that exist among immediate and more distant environments. The occupational health and safety of workplaces at the local level, for example, are directly influenced by state and national ordinances aimed at protecting environmental quality and public health (cf., Stokols 1992).

Another core theme of ecological theory is that physiological, emotional, and social well-being are influenced by the multiple, interrelated settings that comprise *the daily activity systems* of individuals and groups (cf., Michelson 1985). Bronfenbrenner's (1979) ecological model of human development highlights the joint influence of people's families, school, work, and recreational settings on their health and developmental outcomes. Rather than focusing exclusively on person–environment transactions that occur within single settings (*microsystems*, such as home or work environment), Bronfenbrenner's research emphasizes the ways in which functional linkages between two or more settings (i.e., the *mesosystem*, the *exosystem*, and the *macrosystem*) influence developmental processes and outcomes. Mesosystem units are comprised of two or more environmental settings (e.g., family and occupational environments) in which the individual is directly involved (e.g., a parent employed outside the home). Exosystems include an environment in which the developing person is directly involved (e.g., a child in his or her home environment) and one or more other settings in which that person does not participate directly (e.g., the workplaces of the child's parents); but that, nonetheless, affect his or her development and well-being (e.g., stressful experiences at work that impair the quality of parents' interactions with their children at home). Finally, the macrosystem level of analysis encompasses the belief systems, social norms, and institutions of the culture as a whole.

The joint influence of multiple life settings on individuals' health has been observed in several studies. Children's exposure to environmental stressors (such as high levels of spatial density and noise) in their home and elementary school environments, for instance, was associated with both additive and interactive effects of those conditions on their physiological health (e.g., systolic and diastolic blood pressure) and academic achievement (cf., Cohen et al. 1986). In other studies, employees' perceived lack of flexibility in work hours for scheduling children's doctor visits was associated with their underutilization of employer-provided family health benefits (Fielding et al. 1994). Health outcomes associated with the work–family mesosystem have also been documented through studies of the adverse impacts of 'work–family conflict,' as well as the positive effects of spousal support in buffering work-related stressors (cf., O'Neil and Greenberger 1994).

The research findings noted above suggest the value of studying human health from an ecological systems perspective. By considering the joint influence of multiple life settings on individuals' well-being, a broader understanding of how personal and situational factors contribute to health status can be achieved, than by focusing more narrowly on environmental conditions within single settings.

4. *Directions and Challenges of Ecology and Health Research*

The broad conceptual scope of ecology and health research is both an asset and a limitation. On the one hand, the breadth of ecological analyses affords a more comprehensive understanding of the multiple determinants of well-being. For instance, major political, economic, and environmental determinants of health that were neglected in earlier behavioral and medical studies have been targeted in more recent ecological analyses. On the other hand, so many disciplines and levels of analysis are combined in ecological research that it is difficult to achieve parsimonious explanations of health phenomena. Thus, an important challenge facing the field of ecology and health is to develop criteria for identifying the most influential determinants of wellness at particular levels of analysis (intrapersonal, social, organizational, institutional, community, global) and to give greatest priority to these high-leverage variables when designing scientific studies and community health promotion programs (see *Health Education and Health Promotion*).

4.1 *Developing High-leverage Strategies for Effective Health Promotion*

Examples of high-leverage variables and health promotion strategies include: (a) the implementation of strategic corporate and public policies that can enhance population health on a broad scale (e.g., California's Proposition 99 which imposed a $0.25 tax on each pack of cigarettes sold and thereby reduced statewide levels of smoking prevalence (cf. Breslow and Johnson 1993); (b) the modification of 'other-directed' health behaviors enacted by corporate and public decision-makers whose actions directly influence health outcomes among large numbers of their employees and constituents; cf., Stokols (1996); (c) the identification of key life settings (e.g., a stressful commute between home and work, or poorly designed work facilities) that can be modified to reduce behavioral risk factors (e.g., smoking, substance abuse, injurious work routines) associated with poor health; (d) the implementation of defensible space design guidelines to reduce the incidence of violence and crime in urban neighborhoods (cf. Newman 1972); and (e) the development of internet-based 'telewellness' programs designed to enhance health awareness and well-being among large segments of the population (cf., Stokols 1999).

4.1.1 Targeting vulnerable subgroups in health promotion programs. An important direction for future ecological research is to incorporate the above leveraging strategies into the design of community health promotion programs aimed at improving the well-being of impoverished and underrepresented groups. Certain illness risk factors, including environmental racism, exposure to community violence, and lack of medical insurance, have proven to be highly resistant to earlier and more traditional health promotion programs (cf., Bullard 1990, Fisher 1995). These interrelated community problems, borne disproportionately by low-income minority groups in the population, are highly complex and will be reversed only through the design and implementation of strategically leveraged, multilevel intervention programs (Winett et al. 1989)—example, the multisectoral Healthy Cities Programs sponsored by the World Health Organization (cf., Duhl 1996).

5. Conclusions

The field of ecology and health, which grew rapidly during the 1980s and 1990s, highlights the importance of developing integrative, broad-gauged strategies for promoting human wellness. As proponents of ecological analyses have noted aptly, improvements in individuals' health practices 'usually require some combination of educational, organizational, economic, and environmental interventions in support of change in both behavior and conditions of living' (Green et al. 1997, p. 125). The principles of systems theory and social ecology outlined earlier provide a useful framework for establishing more integrative and effective approaches to wellness promotion in the coming decades.

See also: Ecology, Cultural; Environmental Health and Safety: Social Aspects; Environmental Stress and Health; Global Environmental Change: Human Dimensions; Health Education and Health Promotion; Nature, Concepts of: Environmental and Ecological

Bibliography

Adler N E, Boyce T, Chesney M A, Cohen S, Folkman S, Kahn R L, Syme S L 1994 Socioeconomic status and health: The challenge of the gradient. *American Psychology* **49**: 15–24

Barker R G 1968 *Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior*. Stanford University Press, Stanford, CA

Belloc N, Breslow L 1972 Relationship of physical health status and health practices. *Preventive Medicine* **1**: 409–21

Berkman L F, Syme S L 1979 Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda County residents. *American Journal of Epidemiology* **109**: 186–204

Breslow L, Johnson M 1993 California proposition 99 on tobacco, and its impact. *Annual Review of Public Health* **14**: 585–604

Bronfenbrenner U 1979 *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press, Cambridge, MA

Bullard R D 1990 *Dumping in Dixie: Race, Class, and Environmental Quality*. Westview Press, Boulder, CO

Cassel J 1964 Social science theory as a source of hypotheses in epidemiological research. *American Journal of Public Health* **54**: 1482–8

Cohen S, Evans G W, Stokols D, Krantz D S 1986 *Behavior, Health and Environmental Stress*. Plenum, New York

Duhl L J 1996 An ecohistory of health: The role of 'healthy cities'. *American Journal of Health Promotion* **10**: 258–61

Emery F E, Trist E L 1972 *Towards a Social Ecology: Contextual Appreciation of the Future in the Present*. Plenum Press, London

Faris R E L, Dunham W 1939 *Mental Disorders in Urban Areas*. University of Chicago Press, Chicago, IL

Fielding J E, Cumberland W G, Pettitt L 1994 Immunization status of children of employees in a large corporation. *Journal of the American Medical Association* **271**: 525–30

Fisher E B Jr 1995 The results of the COMMIT trial. *American Journal of Public Health* **85**: 159–69

Green L W, Ottoson J M 1999 *Community and Population Health*, 8th edn. WCB/McGraw-Hill, Boston, MA

Green L W, Simons-Morton D, Potvin L 1997 Education and life-style determinants of health and disease. In: Detels R, Holland W, McEwen J, Omenn G S (eds.) *Oxford Textbook of Public Health*, 3rd edn. Oxford University Press, New York, pp. 125–37

Grunberg N E 1991 Cigarette smoking at work: Data, issues, and models. In: Weiss S M, Fielding J E, Baum A (eds.) *Perspectives in Behavioral Medicine: Health at Work*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 75–98

Hawley A H 1950 *Human Ecology: A Theory of Community Structure*. Ronald Press, New York

Kaplan G A, Pamuk E R, Lynch J W et al. 1996 Inequality in income and mortality in the United States: Analysis of mortality and potential pathways. *British Medical Journal* **312**: 999–1003

Karasek R, Theorell T (eds.) 1990 *Healthy Work: Stress, Productivity, and the Reconstruction of Working Life*. Basic Books, New York

Kickbush I 1989 Approaches to an ecological base for public health. *Health Promotion* **4**: 265–8

Lachman M E, Weaver S L 1998 The sense of control as a moderator of social class differences in health and well-being. *Journal of Personality and Social Psychology* **74**: 763–73

Lawton M P 1999 Environmental taxonomy: Generalizations from research with older adults. In: Friedman S L, Wachs T D (eds.) *Measuring Environment Across the Lifespan*. American Psychological Association, Washington DC, pp. 91–124

Leaf A 1989 Potential health effects of global climatic and environmental changes. *New England Journal of Medicine* **321**: 1577–83

Malmstrom M, Sundquist J, Johansson S E 1999 Neighborhood environment and self-reported health status: A multilevel analysis. *American Journal of Public Health* **89**: 1181–6

Marmot M G, Bosma H, Hemingway H et al. 1997 Contribution

of job control and other risk factors to social variations in coronary heart disease incidence. *The Lancet* **350**: 235–9
Maruyama M 1963 The second cybernetics: Deviation-amplifying mutual causal processes. *American Scientist* **51**: 164–79
McGinnis J M, Lee P R 1995 Healthy People 2000 at mid decade. *Journal of the American Medical Association* **273**: 1123–9
Michelson W H 1985 *From Sun to Sun: Daily Obligations and Community Structure in the Lives of Employed Women and their Families*. Rowman & Allenhald, Totowa, NJ
Moos R H 1979 Social ecological perspectives on health. In: Stone G C, Cohen F, Adler N E (eds.) *Health Psychology: A Handbook*. Jossey-Bass, San Francisco, CA, pp. 523–47
Newman O 1972 *Defensible Space: Crime Prevention Through Urban Design*. Macmillan, New York
O'Neil R, Greenberger E 1994 Patterns of commitment to work and parenting: Implications for role strain. *Journal of Marriage and the Family* **56**: pp. 101–12
Ottawa Charter for Health Promotion 1986 *Health Promotion* **1**: iii–iv
Park R, Burgess E, McKenzie R D (eds.) 1925 *The City*. University of Chicago Press, Chicago, IL
Pelletier K R 1996 A review and analysis of the health and cost-effective outcome studies of comprehensive health promotion and disease prevention programs at the worksite: 1993–1995 update. *American Journal of Health Promotion* **10**: 380–8
Robert S A 1998 Community-level socioeconomic status effects on adult health. *Journal of Health and Social Behavior* **39**: 18–37
Sanders-Phillips K 2000 Health promotion in ethnic minority families: The impact of exposure to violence. In: Schneider Jamner M, Stokols D (eds.) *Promoting Human Wellness: New Frontiers for Research, Practice, and Policy*. University of California Press, Berkeley, CA, pp. 294–324
Schneider Jamner M, Stokols D (eds.) 2000 *Promoting Human Wellness: New Frontiers for Research, Practice, and Policy*. University of California Press, Berkeley, CA
Stern P C, Young O R, Druckman D (eds.) 1992 *Global Environmental Change: Understanding the Human Dimensions*. National Academy Press, Washington, DC
Stokols D 1992 Establishing and maintaining healthy environments: Toward a social ecology of health promotion. *American Psychologist* **47**: 6–22
Stokols D 1996 Translating social ecological theory into guidelines for community health promotion. *American Journal of Health Promotion* **10**: 282–98
Stokols D 1999 Human development in the age of the internet: Conceptual and methodological horizons. In: Friedman S L, Wachs T D (eds.) *Measuring Environment Across the Lifespan: Emerging Methods and Concepts*. American Psychological Association, Washington, DC, 327–56
Stokols D 2000 Social ecology and behavioral medicine: Implications for training, practice, and policy. *Behavioral Medicine* **26**: 129–30
United States Public Health Service 1979 *Healthy People: The Surgeon General's Report on Health Promotion and Disease Prevention*. DHEW Publication No. (PHS) 79-55071. U.S. Government Printing Office, Washington, DC
Winett R A, King A C, Altman D G 1989 *Health Psychology and Public Health: An Integrative Approach*. Pergamon Press, New York

D. Stokols

Ecology, Cultural

The central interest of cultural ecology is the mutual relations of people and the environment (of places and resources) they occupy and use. Julian Steward coined the term in 1937, to distinguish it from biological, human, and social concepts of ecology. Steward defined it as:

> the study of the processes by which a society adapts to its environment. Its principal problem is to determine whether these adaptations institute internal social transformations or evolutionary change. It analyzes these adaptations, however, in conjunction with other processes of change. Its method requires examination of the interaction of societies with one another and with the natural environment (Steward 1968).

Cultural ecology, an inquiry dominated by geographers and anthropologists, has sought to make a place for environment in understanding and explaining the human condition (i.e., culture alone cannot explain culture or cultural change). It studies the significance of environment and the interactions humans have with it, with influences going in both directions: humankind's role in modifying the earth, and the environment's influences on human nature and culture. There is the undoubted fact that humans use the world around them and solve questions about how to make the material world meet human needs and wants. Adaptation of and to the material world is a great intellectual accomplishment of humankind.

Steward wrote the essay on cultural ecology for the 1968 edition of the *International Encyclopedia of the Social Sciences*. This article, after summarizing the main purposes and methods of cultural ecology in the early 1960s, traces the development of the field through the subsequent four decades. The purpose of the article is to illustrate the kinds of questions asked in cultural ecology and the way they are answered. Examples have been chosen for their illustrative power.

1. Major Themes in Cultural Ecology

The most important preoccupation in cultural ecologic research, particularly when done by anthropologists, has been with cultural adaptation. According to Steward, one studied (a) the connections between an environment and the technology people use to exploit it, (b) the patterns of behavior that result when particular technologies are used, and (c) the effects, if any, these have on other cultural patterns (Steward 1955). The resulting changes set in motion by cultural adjustments to environmental conditions constituted, in Steward's phrase, a process of multilinear cultural evolution (the idea being that there are significant, but

not universal, regularities in cultural evolution, that lead to parallel form, function, and sequence in quite disparate, unrelated cultures). Although complex, this approach is essentially a functionalist one, which argues that there are predictable relations between a given set of environment/technology relations and the evolution of other aspects of culture–social institutions and individual personality and behavior.

This theme of adaptive dynamics has been further developed by Gregory Knapp in *Andean Ecology* (1991). An associated theme, studied by Fredric Barth (1956), was economic specialization by different groups, with mutually beneficial reciprocal relationships. The resulting regions of specialization he termed econiches.

Works cited in this article which do not, because of space limitations, appear in the bibliography, can be found in one, or more, of the following four review essays in cultural ecology: Bassett and Zimmerer 2001, Butzer 1989, Grossman 1977, and Porter 1978. For reviews of cultural ecology by anthropologists, see Hardesty 1979, Moran 1990, Netting 1977, and Vayda and McCay 1975.

The second most important theme in cultural ecology can be called agricultural stability and change, an intersecting set of topics including population pressure and agricultural intensification (Boserup 1965, Netting 1968); carrying capacity (Allan 1949, Carniero 1956, Sahlins 1958); agricultural involution (Geertz 1963); environmental cognition and systems of indigenous knowledge (Porter 1965); systems analysis (Von Bertalanffy 1959); and energetics (Rappaport 1968). Two perspectives, synchronic, which focuses on contemporary conditions, and diachronic, which has a deep cultural historical, even archeological, basis, have been evident in much cultural ecological study (Butzer 1964, 1989, Denevan 1966). Agroecology, the study of field systems, crop management through habitat modification, and fertility management, has been a major interest of cultural ecologists, a good part of which has been archeological in nature—relict fields, terraces, irrigation systems, etc. (Denevan 1972, Doolittle 1988, Turner 1983, Dunning and Beach 1994).

Other themes and concepts that figured in cultural ecology by the mid-1960s include environmental risks and hazards (White 1961, Kates 1962); the behavioral environment (Kirk 1963) and environmental perception (Saarinen 1966); and medical ecology (May 1958, Hunter 1974, and reviewed by Pyle 1976, and Earickson et al. 1989).

2. Methods

The methods of the cultural ecologist are essentially those of the ethnographer—careful observation and recording of the ideas and material practices of the people whose culture is being studied, as well as participant observation and recording of life histories. The difference lies, perhaps, in the emphasis placed on the biophysical world the people use. Whereas the ethnographer pays great attention to kinship, law, language, origin myths, and religious beliefs and practices, the cultural ecologist devotes comparable attention to knowledge and beliefs regarding plants, soils, seasons, terrain, agriculture, livestock, and other aspects of the environment the people use. Another difference may lie in the degree to which cultural ecologists collect and carefully measure highly detailed data relating to resource use and the environment, and seek to understand the meaning of complex interrelationships among the data. An example is Eric Waddell's work in *The Mound Builders* (1972), in which relationships of sweet potato varieties, mound size, cold air drainage, size of soil aggregates after tillage, elevation of gardens, and frost hazards are considered.

Among cultural ecology's methods is an anthropological/ethnographic use of an approach associated with the linguist Pike (1966); namely, a focus on the internal, emic world of contextualized meanings of terms, in contrast to any universal, etic, or standardized manner of uttering a term (derived from phonemic vs. phonetic). The cultural ecologist also uses western science to the extent possible and when appropriate to the task at hand, whether it be carbon-14 dating methods, measurement of energy-water budget variables, or analysis of satellite imagery. The use of western science methods was strongly endorsed in the 1970s by such cultural ecologists as Karl Butzer (1982) and Harold Brookfield (1968). One further precept is that the cultural ecologist should be historical to the extent possible. A sensitivity to historical aspects and change through time helps ensure that analysis avoids freezing cultural groups into an ahistorical ethnographic present, an approach that permits them ethnography but denies them history.

Cultural ecology's use of the emic approach as a method and research framework frequently emerges through a common sense evaluation of how to do field work, and has been independently invented by field researchers any number of times. For example, the soils of the Pokot of western Kenya can be described using Western science—the US Soil Conservation Service's Comprehensive Classification System, but it would have no meaning for Pokot farmers, who have their own well-developed system of soil names with associated characteristics, typical locations, and uses. Since Pokot behavior is informed by their names for and understanding of soils, one learns much by using their classification (Porter and Sheppard 1998).

3. Scale and Location

The scale of cultural ecological studies reflects the intense nature of the research that produces it. Most cultural ecological study concerns small areas and

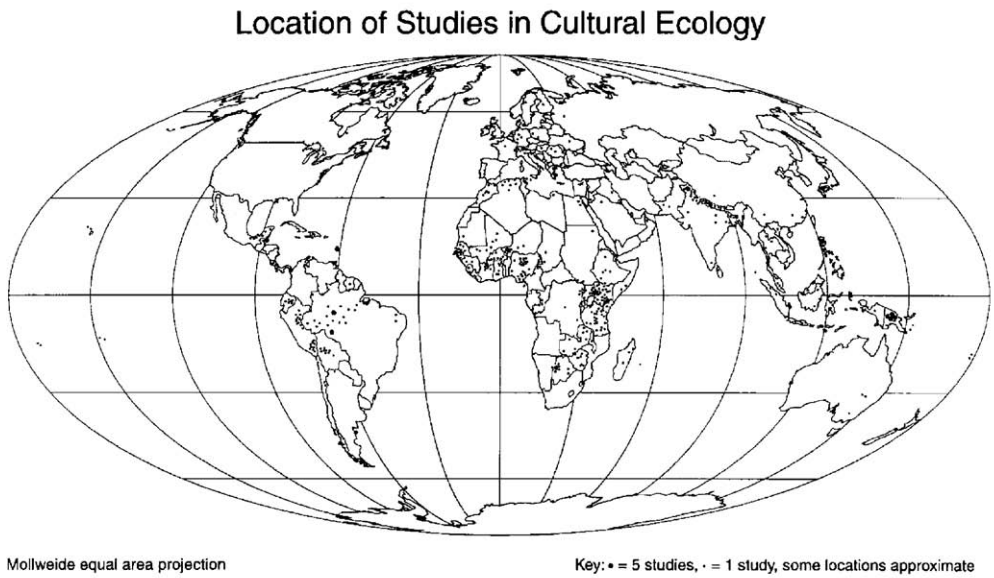

Figure 1
The bulk of the research shown in this figure has been done by these scholars: for *Papua New Guinea*—Harold Brookfield, Paula Brown, William Clarke, Lawrence Grossman, Paul Sillitoe, Andrew Vayda, and Eric Waddell; for *Africa*—William Adams, William Allan, Thomas Bassett, Simon Batterbury, Frank Bernard, David Brokensha, Karl Butzer, David Campbell, Judith Carney, Claudia Carr, Francis Conant, James Delehanty, Ton Dietz, Tom Downing, Neville andRada Dyson-Hudson, Robert Ford, John Galaty, Abe Goldman, Walter Goldschmidt, Leslie Gray, Philip Gulliver, John Hunter, Douglas Johnson, Robert Kates, Cindi Katz, C. Gregory Knight, Richard Lee, Peter Little, Gordon Matzke, Jacques May, Michael Mortimore, Robert Netting, Roderick Neumann, James Newman, Symmes Oliver, Philip Porter, Paul Richards, Bernard Riley, Dianne Rocheleau, Abdi Samatar, Richard Schroeder, Earl Scott, Thayer Scudder, Frederick Simoons, Derrick Thom, Mary Tiffen, Colin Trapnell, Matthew Turner, Donald Vermeer, Michael Watts, Gene Wilken, Edgar Winans, and Ben Wisner; for *Latin America and the Caribbean*—David Aagesen, Sheryl Beach, Anthony Bebbington, James Blaut, Karl Butzer, Robert Carniero, Oliver Coomes, William Denevan, William Doolittle, Nicholas Dunning, Arturo Escobar, Kent Flannery, Lawrence Grossman, Susanna Hecht, Anne Kirkby, Gregory Knapp, Diana Liverman, Kent Mathewson, Emilio Moran, Bernard Nietschmann, Ben Orlove, Eugene Parker, Campbell Pennington, Andrew Sluyter, Julian Steward, B. L. Turner, II, James and Connie Weil, Thomas Whitmore, Gene Wilken, and Karl Zimmerer, and for *South and Southeast Asia*—Nigel Allen, Jacques Barrau, Fredrik Barth, Piers Blaikie, Clifford Geertz, Harold Conklin, Roy Ellen, Charles Frake, Martin Lewis, Melinda Meade, John Metz, Haripryia Rangan, Paul Robbins, Marshall Sahlins, Stan Stevens, and Andrew Vayda

small groups of people—a village, a community, a valley. By judicious use of synedoche and caveat, the cultural ecologist may claim that the findings for the smaller group represent the larger culture. Indeed, following Brookfield's early lead (1962), cultural ecologists commonly work at multiple scales: the individual, the household, the community, the culture (and its territory) overall, as well as the culture in larger settings—national, international, and with reference to global political economy.

The 'third world' location of most cultural ecological research is illustrated in Fig. 1. It is based on a large bibliography on cultural ecology, collected over the years, plus searches, by title and subject, of several library databases (anthropology and geobase) using the linked key words cultural ecology. The search generated over 700 books and articles, either with the term cultural ecology in the title, or classified under the subject heading cultural ecology. The geographic locations of as many of these works as could be determined (over 550) were plotted on the map. The resulting pattern (an approximation) emphasizes the tropical and subtropical, colonial location of the bulk of the research. An exception is the work that has been done among native American Indian groups in the US and Canada. There are significant clusters in Papua New Guinea (where much early work originated), Africa, the Caribbean and Latin America (especially

Mayan, Andean, and Amazonian), and Himalayan south Asia.

The cultures that have been studied by cultural ecologists have tended to be preliterate, agrarian, and part of the former colonized world. The goal of the cultural ecologist has been to understand and convey the life-world and life-ways of indigenous cultures. It has sought to give a voice to the knowledges, beliefs, and values of indigenous peoples. Cultural ecology, therefore, constitutes, in the language of contemporary social sciences, a discourse. Its effects have been muted, since other discourses in the social sciences and in policy arenas are dominant. This may be changing, however, as witness the World Bank's recent focus on social capital and forms of reciprocity and community that tap local knowledge.

4. *Cultural Ecology 1960–2000*

Much happened in cultural ecological research in the final four decades of the twentieth century. The remaining discussion is divided into three sections: the evolution of cultural systems, the study of livelihood and environment, and recent critiques of cultural ecology.

4.1 *The Evolution of Cultural Systems*

People appraise the potentialities and constraints of their surroundings in different ways: (a) in the way they organize space (settlement patterns, field dispersal, use of different ecological zones), (b) in their use of local resources (irrigation, crop varieties, medicinal plants), (c) in the organization of internal relations (family, clan, age-set, bridewealth, neighbors, ritual and ceremony), and (d) in the way they connect with larger social and economic forces—local and national governments and the global economy (labor migrancy, remittances, production and sale of cash crops, contract farming, adoption of new technologies). Changes in circumstances in any of these domains can affect any of the other domains and occasion cultural change.

As an illustration of efforts to demonstrate how cultures evolve in response to changes in environmental/technological circumstances, consider the work of Walter Goldschmidt and co-workers in the Culture and Ecology in East Africa Project (1965). The central hypothesis of this project was that a major change in livelihood (say from livestock keeping to farming) would require changes in the nature and inter-relationships of social institutions and people, and at a deeper level, changes in the attitudes and personality characteristics of individuals. With a change in livelihood, social institutions and the repertoire of techniques must be refashioned or abandoned and replaced by new ones to meet changed circumstances.

The project centered on pastoralism, with farming used as its foil. Four cultural groups in east Africa (Sebei, Pokot, Wakamba, and Wahehe) were chosen because in each there was a group whose members were predominantly pastoralists and another whose members were predominantly farmers. It was hypothesized that the Sebei, for example, were 'originally' pastoralists, but over a period of generations, after some of their members took up farming on the higher potential parts of Mount Elgon in Uganda, growing bananas, maize, and coffee as a cash crop, their social institutions adapted to the requisites of horticulture. In the process they transformed a lightly peopled montane forest environment into one of dense human settlement.

First, the reader should envisage the dramatic contrast in settlement pattern between the farming and pastoral communities Goldschmidt studied. In the farming area the population density was 343 people/sq. km (888 people/sq. mile), a landscape of nearly continuous banana groves, which shaded coffee trees. Among the pastoralists, who grew some maize and finger millet, the population density was seven people/sq. km (19 people/sq. mile), with large open tracts of unpeopled grazing land, a sparse *Acacia* savanna.

Goldschmidt, in *Culture and Behavior of the Sebei* (1976), documented numerous ways in which institutions that were important for pastoralists, changed under a regime of cultivation. For example: the *pororyet*, a military organization based on age grades and designed for offensive action and rapid movement in open country, became among the farmers a resident militia of a fixed territory and made up of men of all ages; the *misisi*, a 'first fruits' thanksgiving harvest ritual, based on a carefully defined list of kin, was transformed into the *mukutanek* (sharing), a ritual beer party to which one invited neighbors, not kin. Its purpose came to be maintenance of good relations with neighbors. The *chomi ntarastit*, an oathing ceremony originally used to settle interclan disputes among the pastoralists, became a ceremony in which adult male residents of the farming community swore to abide by laws intended to preserve peace and public order; the *chichomnko lekok*, a ceremony among pastoralists performed between half-brothers, when they were toddlers, to ensure that there would be peace between them, was used by adults among the farmers to resolve tension and disputes over land and inheritance; age grades and male and female circumcision ceremonies became less formal and in certain instances were abandoned altogether in the shift from pastoral to farming livelihoods; the *tokapsoi* and the *toka che kichenkochi korket*, terms that refer, respectively, to a man's cattle and those that have been anointed and smeared for a wife, and over which she

has control (but which pass to her sons upon her death), were applied to farmed fields.

Edgerton, in *The Individual in Cultural Adaptation* (1971), explored the inner worlds of beliefs, attitudes, and personality of the Sebei and members of the other three cultures. His purpose was to see how pastoralists and farmers differed. He interviewed samples of about 30 adult men and 30 adult women in the pastoral and farming communities of the four cultures, administering a rich variety of test instruments. Although there were greater within-culture similarities than between cultures (i.e., 'culture' explained more than 'ecology'), there were 31 categories of difference between the people of the eight Culture and Ecology communities that were statistically significant. Some of these differences were attributable clearly to farming or herding. He discussed the various attributes along a continuum with herding at one pole and farming at the other. Five attributes were found close to the pastoral pole: affection, direct aggression, divination, independence, and self-control. Nearly as central in importance were adultery, sexuality, guilt-shame, depression, and respect for authority. In contrast, the attitude closest to the farming pole was disrespect for authority, but the following attributes were also characteristic of farmers: anxiety, conflict avoidance, emotional constraint, hatred, impulsive aggression, and indirect action. As one considers the requirements of a settled, sedentary life, and life in a community where one cannot easily escape the proximity of neighbors, in contrast to the demands of a life managing livestock, where one must be flexible in meeting the daily, unremitting needs of animals for water, graze, browse, salt, and safety, one can see ways in which the attributes emphasized in each lifeway make sense. In the broadest terms: 'Farmers must create institutions for the preservation of peaceable relationships, ... [P]astoralists must create ideals that bind together persons who need to preserve a sense of community in the face of the fact that they are forced by economic circumstances to operate as individualists' (Edgerton 1971).

The foregoing illustration of Steward's project to 'study the processes by which society adapts [or maladapts] to its environment,' could be expanded by the citation of many other scholars.

4.2 The Study of Livelihood and Environment

The core of this theme is the nexus of population and resources, from which many subsidiary questions flow—limits to growth, the carrying capacity of the earth, agricultural intensification, soil erosion and environmental degradation, destruction of forest lands, with associated species extinctions and loss of biodiversity, problems of the global commons (oceans, atmosphere, lakes, forest habitat, and coastlines). Key events in the study of livelihood systems were: (a) Harold Brookfield's 1964 article in *Economic Geography*, which argued the urgent need to include the interior dynamics of culture—that which lay in the beliefs, values, and attitudes of people, as well as their social institutions; (b) the appearance in 1965 of Ester Boserup's *The Conditions of Agricultural Growth* (to which we return); (c) Roy Rappaport's influential (1968) *Pigs for the Ancestors*, which illustrated systems theory, energetics, and homeostasis and feedback in ecological systems; (d) Robert Netting's (1968) classic work on the Kofyar, as well as his influential monograph, *Cultural Ecology* (1977); (e) two studies of disruptions in subsistence systems brought about by commercial activities: Bernard Nietschmann (1973), on the Miskito Indians of Nicaragua, and Lawrence Grossman (1984), on the highlands of Papua New Guinea; and (f) the publication of Piers Blaikie and Harold Brookfield's *Land Degradation and Society* (1987). Blaikie and Brookfield emphasized the importance of using a nested set of scales for research, in order to show how decisions and policies at national and international levels affecting land ownership, commodity pricing, development assistance, etc., had impacts on people at local levels. They showed how poor farmers were marginalized, both literally and figuratively, and were unable in their struggle to eke out a living (and despite their better instincts) to avoid degrading the environments they were forced to use.

Boserup's book challenged the classic argument of Thomas Malthus (that population can increase geometrically while agricultural productivity could only increase arithmetically). Her argument was that through slow sustained pressure of population on resources, people improved agricultural output. She focused on the means by which this was done: (a) the mobilization of labor, and (b) the tools appropriate to the vegetation managed by cultivators at the end of each fallow. She placed land on a continuum of use (forest fallow, bush fallow, short fallow, continuous cultivation) and examined a suite of expectations about the tools, the role of livestock, land ownership, and social and infractructural institutions and attitudes associated with different stages in land use intensification. Her work, supported by that of William Allan, whose *The African Husbandman* appeared the same year, transformed thinking in cultural ecology in many fruitful ways. Despite cogent critiques of her work, for neglecting capitalism, market relations, and the state, evidence of the impact and acceptance of her ideas is the now widespread use of the adjective Boserupian.

Environmental change induced by human action has also been an important theme in cultural ecology (Turner 1990). Some of it has speculated on population dynamics of past civilizations—Mayan, Incan, Egyptian (Denevan 1976, Butzer 1976). In contemporary terms it involves the inventorying of land use and land quality and subsequent tracking of changes in the geosphere and biosphere. These questions require a

'big-science' approach and involve the cooperation of different disciplines and international organizations. Among themes in this topic are land use change, climate change, pollution, and loss of plant diversity, farmland, and wilderness. Cultural ecologists contribute to some aspects of this inquiry.

4.3 Recent Critiques of Cultural Ecology

Cultural ecology as practiced in an earlier period ran four interrelated risks: (a) that of assuming equilibrium and balance in the society studied, (b) ethnographic presentism, i.e., failing to give adequate attention to history and change in the society, (c) use of an agreeable, internal, circular functionalism in explanation, and (d) failure to incorporate influences external to the group being studied. Given these potential flaws, scholars sometimes found cultural ecology inadequate and incomplete as a research paradigm and modified or added to the approach (Porter 1978). Although one might argue for an earlier beginning of the radical critique of cultural ecology (Brookfield and Kirkby 1974), one seminal event was the publication of Michael Watts' *Silent Violence: Food, Famine and Peasantry in Northern Nigeria* (1983). This book brought political economy, Marxian analysis of global capitalism, and the state more fully into the work of cultural ecologists. Watts in his writing and teaching has influenced a succeeding generation of cultural ecologists, many of whom prefer the term political ecology to describe their work. Political ecology has been a growth center of cultural ecology ever since, indeed, with efforts to go beyond political ecology to new, more effective approaches. An example is *Liberation Ecologies* edited by Peet and Watts (1996). The book argues for a single, all encompassing, poststructuralist, social science based on the discursive perspective and methods of critical theory as used in analysis in literature, the arts, and humanities. Both political ecology and cultural ecology are, they argue, inadequately theorized, and not sufficiently political. The various authors of *Liberation Ecologies* call for a reconceptualization of virtually everything: nature, environment, the state, class, gender, ethnicity, community, region, and identity. The central concern is with power relations and the contestations they generate at all scales, from the household (between wife and husband) to the state and the global economy. *Liberation Ecologies*, though full of stimulating ideas, does not, however, provide a unified alternative to political ecology, since its authors disagree on many questions.

Another critique of cultural ecology has been made by Karl Zimmerer (1994) in an article titled 'Human geography and the "new ecology": the prospect and promise of integration,' which proposes adoption of uncertainty principles and chaos theory in cultural ecological analysis, abandoning the underlying assumptions of homeostasis and ecological balance that characterized earlier work.

Among significant themes that have been developed are feminist political ecology (Carney 1993, Rocheleau et al. 1996, Schroeder 1999); the political ecology of protected areas (Neumann 1998, Stevens 1999, Nietschmann 1997); the political ecology of social movements (Bebbington 1996, Rangan 1996); and the theme of sustainability and global change (Chambers 1997, Turner 1990).

Although it has been critical of cultural ecology, political ecology should not be seen as an offshoot of cultural ecology. It is, rather, an enlarging of cultural ecology's domain of inquiry and the conscious incorporation of effects originating outside the ecological setting. Much early cultural ecological work was microscale and tended to study ensembles of things occurring at a place (site characteristics) as well as situational characteristics within the study area. Political ecology extended the scale of inquiry to include the global economy, intrusions of the state, markets, and other external influences.

Not all cultural ecologists accept the reconfiguration of cultural ecology as political ecology, and feel that with the addition of Marxian analysis and postmodern concerns over relations of power, knowledge, and resources, there has been a corresponding narrowing of inquiry and loss of the healthy eclecticism that has characterized cultural ecology, as well as moving it away from cooperation with the more positivist environmental sciences—ecology, paleoecology, and global change science (Turner 1997).

Cultural ecology remains a lively field of inquiry, asking questions that will not go away, and that can never be answered in fully satisfactory ways. Current approaches in cultural ecology, and its vigorous descendant, political ecology, reflect transformations in the social sciences themselves. At the same time there appears to be a diminution of the importance attached to the biophysical world and a western science understanding of it. In these postmodern times, when 'everything is socially constructed,' cultural ecology can be expected to continue to reflect changing modes of thought in the wider society.

See also: Cultural Geography; Ecology, Deep; Ecology, Political; Environmental Determinism; Environmentalism: Philosophical Aspects; Environmentalism: Preservation and Conservation; Global Environmental Change: Human Dimensions; Human–Environment Relationships; Land Degradation; Land Reclamation; Rural Geography

Bibliography

Bassett T J, Zimmerer K S 2001 Cultural ecology. In: Gaile G L, Willmott C J (eds.) *Geography in America at the Dawn of the 21st Century*. Oxford University Press, New York

Blaikie P M, Brookfield H C 1987 *Land Degradation and Society*. Methuen, London
Brookfield H C, Kirkby A 1974 *On Man Environment and Change*. International Geographical Union, Working Group on Man-Environment Theory. Work Notes No.1
Butzer K W 1989 Cultural ecology. In: Gaile G L, Willmott C J (eds.) *Geography in America*. Merrill, Columbus, OH
Chambers R 1997 *Whose Reality Counts? Putting the First Last*. Intermediate Technology Publications, London
Edgerton R B 1971 *The Individual in Cultural Adaptation*. University of California Press, Berkeley, CA
Goldschmidt W R 1976 *Culture and Behavior of the Sebei*. University of California Press, Berkeley, CA
Grossman L S 1977 Man-environment relationships in anthropology and geography. *Annals of the Association of American Geographers* **67**: 126–44
Hardesty D L 1979 *Ecological Anthropology*. Wiley, New York
Moran E J (ed.) 1990 *The Ecosystem Approach in Anthropology*. University of Michigan Press, Ann Arbor, MI
Netting R McC 1977 *Cultural Ecology*. Cummings Publishing Company, Menlo Park, CA
Peet R, Watts M (eds.) 1996 *Liberation Ecologies: Environment, Development, Social Movements*. Routledge, London
Pike K L 1966 Etic and emic standpoints for the description of behavior. In: Smith A G (ed.) *Communication and Culture*. Holt, Rinehart and Winston, New York
Porter P W 1978 Geography as human ecology: A decade of progress in a quarter century. *American Behavioral Science* **22**: 15–39
Porter P W, Sheppard E S 1998 *A World of Difference: Society, Nature, Development*. Guilford Press, New York
Steward J H 1955 *Theory of Culture Change: The Methodology of Multilinear Evolution*. University of Illinois Press, Urbana, IL
Steward J H 1968 Cultural ecology. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 4. pp. 337–44
Vayda A P, McCay B J 1975 New directions in ecology and ecological anthropology. *Annual Review of Anthropology* **4**: 293–306

P. W. Porter

Ecology, Deep

The term deep ecology was introduced by Norwegian philosopher and mountaineer Arne Naess (1912–) in 1972 at the Third World Future Research Conference in Bucharest. After a distinguished academic career, Naess chose early retirement to focus his expertise on addressing the global ecological crisis. Naess coined the terms deep ecology and shallow ecology to juxtapose what he regarded as two radically different approaches for problematizing (*Problematisieren*) and responding to the ecological crisis (Naess 1973). Deep ecology posits that along with humans' special capacities for reason and moral consciousness come equally special responsibilities, particularly in relation to the flourishing of nonhuman life and the ecological sustainability of the planet. In the years since introducing the term, Naess and other supporters of the deep ecology movement have written extensively, elaborating and popularizing the term in various directions. Considerable controversy and confusion has ensued, in part because deep ecology calls for a radical rethinking of our relationship to each other and nature, because provocative terminology and philosophical vagueness have been employed, and because deep ecology has unfolded over time, with substantial emendation. As a result, deep ecology has come to mean many things to many people. Regrettably, the term is now often used without discrimination to refer to three distinct entities: (a) the particular ecophilosophical approach advanced by Naess and other theoreticians of the deep ecology movement; (b) the international, grassroots, often activist oriented deep ecology movement; and (c) Naess' personal systematization of a philosophy of ecological harmony, Ecosophy T. After introducing deep ecology as a unique approach to ecophilosophy, the evolution of deep ecology will be reviewed and the remaining two entities will be briefly considered.

1. Deep Ecology as Ecophilosophy

The purpose of deep ecology as an ecophilosophical approach is to assist individuals in the process of weaving descriptive and prescriptive premises about the world, ecological science, and their ultimate beliefs into a cohesive framework for guiding decisions involving society and nature—ecologically inspired total views or ecosophies. This focus on praxis (responsibility and action) separates the deep ecology approach from more descriptive inquiries into environmental philosophy that focus on axiological questions, such as extending 'rights' to certain nonhumans or grading intrinsic value. The ontologically inspired deep ecology approach emphasizes eliminating the perception of fundamental people/environment and spiritual/physical cleavages. Its primary strategy for overcoming the ecological crisis is to help individuals avoid pseudorational thinking.

Naess argues that many regrettable environmental decisions are made in a state of 'philosophical stupor,' where narrow concerns are confused with, and then substituted for, more fundamental ones. In proposing the deep/shallow contrast, Naess applies his research on empirical semantics, philosophy of science, Spinoza, the inquiring skepticism of Sextus Empiricus, and Gandhi on nonviolent communication. His technical semantic distinction is directed at the individual's level of problematizing—the extent to which they can and do coherently and consistently trace their views, practices, and actions back to their ultimate beliefs or bedrock assumptions.

In relating this notion of persistently asking deeper questions to the ecological crisis, Naess broadens his concept of 'depth.' In the context of deep ecology as an ecophilosophical approach, depth refers to both the

general level of problematizing one employs in seeking out the underlying, coevolving causes of the ecological crisis and the extent of one's willingness to consider a broad array of social and policy responses, even if they necessitate changes that represent a radical departure from the status quo.

The 'shallow,' currently more influential, approach is identified with treating the symptoms of environmental degradation, such as pollution and resource depletion. Its central concern is the health and prosperity of people in the economically privileged countries. This reform-oriented approach is grounded not in ultimate premises that plumb the relationship between humans and nature, but in technological optimism, economic progress, and scientific management. A core premise is that all environmental problems are manageable—nature is a puzzle to be deciphered by human ingenuity and manipulated, albeit more efficiently, for human benefit. From this perspective, remedy for environmental problems is limited to economic, technological, and managerial reforms. This effort to palliate human impacts, rather than probe and address their underlying causes, favors a search for 'technical' solutions to what are more likely ethical, social, and political problems. By truncating the realm of problematizing, the shallow approach, perhaps inadvertently, prunes the set of conceivable social changes to a narrow incrementalism.

The 'deep' approach, as the other hand, while in no way discounting the exigency of addressing pollution and resource depletion, adopts a broader, long-term, more skeptical stance. Doubtful about technological optimism, critical of limitless economic growth, and decidedly against valuing nature in purely instrumental terms, it asks if the shallow approach's proposed solutions take into consideration the pervasiveness and severity of the problems they hope to rectify. Drawing on a wide diversity of philosophical or religious ultimate premises, which acknowledge that every living being has value in itself, the deep approach sees the flourishing of nature and culture as fundamentally intertwined. Nature is viewed as mentor, standard, and partner rather than vassal. A key premise is that environmental management is much more about managing the habits and desires of humans than managing nature. Remedy for environmental problems is sought by identifying and responding to the complex 'root' causes of the ecological crisis, dedicating special attention to protecting the wild and free from thoughtless human interference. Taking less for granted, the deep approach calls for the public questioning of every practice, assumption, and value that propels the ecological crisis.

By juxtaposing these two, almost caricatured, perspectives, Naess employs a technique of Gandhian nonviolent communication designed to confront core disagreements. The premise is that society's potential to overcome the ecological crisis rests on maneuvering discussion to its root causes. One of the primary root causes, Naess asserts, is the widespread disjunction between people's core beliefs and actions. People, in general, neither comprehend how their practices and everyday lifestyle choices harm the environment, nor recognize how these consequences may be in direct conflict with their core beliefs—this is the primary weakness of the shallow approach. A crucial, underlying hypothesis of the deep approach is that teasing out the presumed inconsistencies between an individual's actions and their fundamental beliefs will effect constructive change instead of generating more serious ancillary conflict.

Rather than simply calling for a new environmental ethic or a change in fundamental values, Naess' approach to ecophilosophy centers on transforming practice and policy by challenging individuals to develop more thoroughly reasoned, consistent, and ecologically inspired total views. Some will take issue with the core premise underlying this goal, namely that thoroughly reasoned and consistent positions do generally lead to improved policies and actions. There are, however, important indirect procedural benefits that can result from attempting to couple reason and values in decision making.

In any case, Naess intended his distinction between the shallow and deep approaches to environmentalism to be restricted to argumentation patterns and the diversity of policy and lifestyle changes that are given consideration. The distinction was never intended to shed light on the 'depth' of particular individuals or their values. Nevertheless, as an expert in semantics and communication theory, Naess cannot be exonerated for failing to anticipate the unfortunate derogatory connotations of his terminology.

2. *The Unfolding of Deep Ecology*

Naess' (1973) inaugural article, *The Shallow and the Deep, Long-Range Ecology Movement: A Summary*, gave both an appellation and a rudimentary philosophical framework to a movement that had been nascent for well over a century. Deep ecology as both a liberatory social movement and an ecophilosophical approach draws on at least five roots: (a) the nonanthropocentrism and reverence for wildness given voice by Henry David Thoreau, John Muir, and John Stuart Mill; (b) the ecological perspective gleaned from a scientific understanding of humans' role in creating environmental degradation, articulated by George Perkins Marsh, Aldo Leopold, Rachael Carson, and Paul Ehrlich; (c) the ecocentric vision and social criticism brought to life by the writing and activism of Paul Shepard, Gary Snyder, Edward Abbey, and Dave Brower; (d) the belief in the ultimate unity of all life, communicated through the nonviolent conflict resolution work of Gandhi; and (e) the call for individuals to be conscious of their value priorities and to counteract passivity and despair by actively making

them manifest in their daily life, explicated in the philosophical systematization of Spinoza.

Naess' brief initial article, along with establishing the deep/shallow contrast, outlined a seven-point survey, which was intended to summarize the shared norms and perspectives of the deep ecology movement. This survey introduced terms and slogans, such as 'rejection of the human-in-environment image in favor of the relational, total field image,' 'biospherical egalitarianism—in principle,' and 'complexity, not complication.' While Naess employed vagueness here intentionally to encourage widespread acceptance, these often enigmatic generalizations acted as a lightening rod for controversy.

Apart from the continuous emphasis on maintaining the 'diversity' of both human cultures and nature and the reincarnation of the 'relational, total field image' notion as Naess' subtle and rich concept of 'gestalt ontology,' these terms neither figure in Naess' more seasoned renditions of the significant tenets of the deep ecology movement nor his more mature descriptions of deep ecology as an ecophilosophical approach. Nevertheless, they do continue to resurface in many popular discussions and critiques of deep ecology. Most of the terms Naess introduces in the 1973 article, such as diversity, complexity, autonomy, decentralization, symbiosis, egalitarianism, and classlessness, do, however, figure prominently as key norms in his subsequent derivations of Ecosophy T (1977, 1989).

Naess' only book length discussion of deep ecology began with a short mimeograph, *Økologi og Filosfi* (*Ecology and Philosophy*), published in 1971. The fifth revision and expansion of this initial work resulted in the 1976 classic, *Økologi*, *Samfunn*, *og Livsstil* (Ecology, Community, and Lifestyle), which has been translated into Swedish, English (1989), Italian, Japanese, and Czech. Perhaps the earliest intimation that deep ecology should be separated into a social movement, an ecophilosophical approach, and Ecosophy T occurs in 1986, with Naess' article, 'The deep ecology movement: Some philosophical aspects.' This article offered an early presentation of the 'deep ecology platform,' which was prepared in collaboration with philosopher George Sessions in 1984.

In the late 1970s Naess' project was taken up and popularized by the California-based team of Sessions and sociologist Bill Devall. Sessions had earlier initiated the process of popularizing deep ecology with his *Ecophilosophy Newsletter*, published in six volumes, intermittently from 1976 to 1984. Devall and Sessions' collaboration culminated in the first English language, book length discussion of deep ecology (1985). Subsequently, the two deep ecology theorists worked independently. Devall published extensively on deep ecology practice (1988). Sessions continued to write on the historical foundations and theory of deep ecology, producing the most comprehensive deep ecology anthology to date (1995).

Two additional anthologies are dedicated to the exploration of deep ecology, Reed and Rothenberg (1993) focus on the Norwegian roots of deep ecology and Drengson and Inoue (1995) present an introduction to deep ecology theory and practice. Deep ecology essays and responses to critiques also figure prominently, along with ecofeminism and political ecology, in Zimmerman et al.'s leading environmental philosophy anthology (2001). In addition, Witoszek and Brennan (1999) produced one volume on deep ecology criticism, which includes dialogic responses by both Naess and his interlocutors. All of these collections, however, are necessarily highly selective. As the primary architect of deep ecology, Naess has more than 75 published articles and book chapters and more than 60 unpublished pieces to his credit. The most comprehensive selection of these articles appears in volume 10 of the 11 volume *Selected Works of Arne Naess* (2002) .

The most significant intellectual history of deep ecology and exhaustive effort to characterize what is 'distinct' about it was offered by Fox (1990). Subsequently, Glasser (1997) argued that Fox's distilling of deep ecology into 'Self-realization!' results in a vitiated version of deep ecology. McLaughlin (1993) takes up a discussion of deep ecology as a response to industrialism's role in creating and perpetuating the ecological crisis. The relationship between Naess' earlier philosophical work and deep ecology is explored by Glasser (1996), who also mines the policy implications of deep ecology. Cramer (1998) measures the degree of influence that deep ecology principles have had on American environmental politics.

While discussions of deep ecology appear in many popular and academic journals, only the *Trumpeter Journal of Ecosophy* has been devoted to advancing deep ecology theory and practice. *The Trumpeter* was under the editorship of Alan Drengson from 1983 to 1997. Subsequently, it has continued as an on-line journal, under the editorship of Bruce Morito.

3. *The Deep Ecology Movement*

The deep ecology movement is an informal, global affiliation of individuals who believe that overcoming the ecological crisis will require radical changes in the way humans relate to each other and nature. Its supporters are united not by their commitment to deep ecology as an academic ecophilosophy (which they may not recognize or embrace), but by their common commitment to ultimate premises that value nature for its own sake and their general rejection of anthropocentrism. What distinguishes the deep ecology movement from deep ecology as ecophilosophy is that it embodies praxis rather than merely philosophizes about it. The intuitions of deep ecology are not articulated as professional philosophy, but as art, poetry, ritual, conservation biology, grassroots acti-

vism, simple lifestyles, bioregionalism, and ecologically sustainable design, farming, fishing, and forestry.

The movement is perhaps best characterized as the group of individuals that do (or would) endorse the eight point Deep Ecology Platform. The 'Eight Points,' which summarize 10 years of thinking about deep ecology, were prepared to supersede the problematic seven point (1973) characterization. The first three points outline a high-level norm for protecting the planet's full diversity and richness of life forms—species, cultures, watersheds, landscapes, and ecosystems. Because the scale and character of human interference with the biosphere are currently excessive and thus incompatible with this norm and the thriving of human life and cultures (points 4 and 5), the next two principles outline a series of changes in practice and policy that can help bring them in line. The final point, without calling for specific actions or imposing particular priorities, urges those who embrace the prior seven points to help effect change.

The Eight Points are intended to express at a generic and abstract level the central views that supporters of the deep ecology movement hold in common. The Platform principles are not intended to characterize common views in particular situations. Since 1984, the Eight Points have been republished extensively along with detailed commentary and gained wide currency. Although they are frequently taken as the 'heart' of deep ecology, the eight points alone cannot effect the breadth and complexity of deep ecology as an ecophilosophical approach.

While the Platform can clearly serve as an 'ecological pillar' for Green politics, it has been criticized for not integrating the other three pillars. These concerns have been fueled by the sometimes inflammatory and ostensibly misanthropic statements that have been made by 'Earth Firsters' and others who have taken inspiration from deep ecology. While the Platform principles do not directly address social justice, grassroots democracy, and nonviolence, they are in no way inimical to them. Quite the contrary, the practical policy changes required by point six leave open immeasurable opportunities for collaboration with the peace and social justice movements. For instance, how can society characterize what constitutes a bundle of 'vital needs' for particular communities in particular situations (Point 3); how can society create new accounting systems that operationalize the distinction between 'quality of life' and 'standard of living' (Point 7); and how can society develop policies to support long-term population reduction through noncoercive and democratic means (Point 4).

4. *Ecosophy T*

Naess' personal application of the deep ecology approach results in a normative-derivational systematization akin to that of Spinoza or Aristotle. He refers to his own deep ecological total view as Ecosophy T. Naess' systematization is normative because it calls for a particular set of prescriptive premises. It is derivational by virtue of being a logical systematization, where premise-conclusion chains generate successively more specific norms (general ecological and ecopolitical principles and concrete decisions), which 'follow' from previously accepted premises and hypotheses.

Naess' survey of Ecosophy T is built on a single, highly aggregated and complex ultimate norm, 'Self-realization!' He likens it to 'the universal self,' 'the absolute,' and the Sanskrit ātman. Naess uses a capital 'S' to distinguish the norm from ego-realization and an exclamation mark to identify it as an imperative. 'Self-realization!' is an ecocentric-ontological analogue of the pareto rule in economics. The premise is that increased realization of any individual or species rests on advancing (or at least not hindering) the realization potential of all other species and individuals. While working from a single ultimate norm enables Naess to eliminate the potential for norm conflicts, it also embodies a certain artificiality that necessitates considerable word magic.

Naess draws his support for the Deep Ecology Platform from his Ecosophy T. Other supporters of the deep ecology movement, however, need not and should not draw directly on Naess' systematization. The Platform can be 'derived' from a broad variety of sometimes dissonant, ultimate norms, including those inspired by all of the world's major religions, ecofeminism, creation myths, Leopold's 'Land Ethic,' and E. O. Wilson's 'Biophilia Hypothesis' to name a few. Ultimate premises are essential building blocks for developing an ecosophy, but their breadth and abstractness insure that they only serve as general guides.

5. *Methodological Issues and Future Directions*

Deep ecology has been criticized for being too critical of pragmatic efforts to address acute environmental problems. It has also been accused of being inherently fascist, antihumanist, anti-feminist, and against having indigenous peoples share wilderness areas with nonhumans. None of these criticisms are borne out by a sophisticated reading of the literature on deep ecology as an ecophilosophical approach. This is not to suggest that deep ecology as an ecophilosophical approach has all the answers. In inspiring many people to reconsider their value priorities and more consistently relate them to their lifestyles and everyday actions, deep ecology opens more questions than it answers. Three key issues for future research loom on the horizon. First, how can society motivate value priorities for ecological sustainability when individuals do not accept the premise that nonhumans have intrinsic value? Second, does taking a deep ecological perspective over a shallow ecological perspective actually result in better policy—policies that result in long-

term protection of cultural and biological diversity, using democratic, noncoercive, and ethically unobjectionable, means. Third, how can deep ecology help to resolve conflicts in environmental decision making when they result not from pseudorational thinking, but real value conflicts.

See also: Environmental and Resource Management; Environmental Cognition, Perception, and Attitudes; Environmental Conservation: Ethical Concerns; Human–Environment Relationships; Nature, Concepts of: Environmental and Ecological Social Movements: Environmental Movements

Bibliography

Cramer P F 1998 *Deep Environmental Politics: The Role of Radical Environmentalism in Crafting American Environmental Policy*. Praeger, Westport, CT

Devall B 1988 *Simple in Means, Rich in Ends: Practicing Deep Ecology*. Gibbs Smith, Salt Lake City, UT

Devall B, Sessions G 1985 *Deep Ecology: Living as if Nature Mattered*. Gibbs Smith, Salt Lake City, UT

Drengson A, Inoue Y (eds.) 1995 *The Deep Ecology Movement: An Introductory Anthology*. North Atlantic Books, Berkeley, CA

Fox W 1990 *Toward a Transpersonal Ecology: Developing New Foundations for Environmentalism*. Shambhala, Boston

Glasser H 1996 Naess' deep ecology approach and environmental policy. *Inquiry* **39**: 157–87

Glasser H 1997 On Warwick Fox's assessment of deep ecology. *Environmental Ethics* **19**: 69–85

McLaughlin A 1993 *Regarding Nature: Industrialism and Deep Ecology*. State University of New York, Albany, NY

Naess A 1973 The shallow and the deep, long-range ecology movement. A summary. *Inquiry* **16**: 95–100

Naess A 1986 The deep ecology movement: Some philosophical aspects. *Philosophical Inquiry* **8**: 10–31

Naess A 1989 *Ecology, Community, and Lifestyle: Outline of an Ecosophy*. [Translated and revised by Rothenberg D]. Cambridge University Press, Cambridge, UK

Naess A 2002 Selected Papers: Deep Ecology, Vol. X of the selected works of Arne Naess

Reed P, Rothenberg D (eds.) 1993 *Wisdom and the Open Air: The Norwegian Roots of Deep Ecology*. University of Minnesota Press, Minneapolis, MN

Sessions G (ed.) 1995 *Deep Ecology for the Twenty-First Century*. Shambhala, Boston

Witoszek N, Brennan A (eds.) 1999 *Philosophical Dialogues: Arne Naess and the Progress of Ecophilosophy*. Rowman and Littlefield, Lanham, MD

Zimmerman M E, Callicott J B, Sessions G, Warren K J, Clark J (eds.) 2001 *Environmental Philosophy: From Animal Rights to Radical Ecology*, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ

H. Glasser

Ecology of Aging

The ecology of aging basically draws upon the idea that old age is a period of adult development profoundly influenced by the environment. Because age-related losses in vision, hearing, mobility, and cognitive capacity have a direct impact on the relationship between the organism and the environment, the older person should be particularly vulnerable to environmental demands. In other words, this research paradigm views contextual factors as being crucial determinants of everyday behavior and life satisfaction. According to Lawton and Nahemow (1973, p. 621), the ecology of aging is defined as 'a system of continual adaptation in which both the organism and the environment change over time in a non random manner.' Within this adaptational system, the ecology-of-aging paradigm emphasizes the role of the *physical and spatial environment* on behavior. Thus, housing research ('Aging in Place') may be regarded as most typical for the ecological approach to aging.

1. Development of the Field

Since aging has long been regarded as a process mostly determined by a biological program inherent in the organism, the explicit consideration of environmental variables which affect the course and outcome of human aging was an important step in the historical development of gerontology. This paradigm shift was greatly influenced by:

(a) The growing role of the social science perspective within gerontology which began in the late forties and set the stage for the emerging field of social gerontology.

(b) The prominent role of learning theories in psychology in the fifties and sixties which underscored the importance of the environment in all stages of human development.

(c) The Chicago school of urban sociology and its human ecology perspective (Burgess, McKenzie, and Park) in the 1920s and 1930s, the writings of Lewin in the thirties and forties, as well as the emergence of ecological psychology and of the social ecology view of human development (Barker, Bronfenbrenner, and Moos) in the sixties and seventies which, among other things, placed particular emphasis on the resource character of the physical and spatial environment and hence provided another set of roots for this discipline (Lawton 1982, Wahl 1992).

The conceptual and empirical productivity within the field grew steadily during the sixties and seventies. Gerontological research has been devoted to housing alternatives for the aged, relocation processes and outcomes, and aging in institutions ever since. By the eighties and nineties, the issue of dementia led to new

research challenges based on the assumption that environmental design becomes particularly crucial when the capacity for person-based adaptations in later life is exhausted.

2. Conceptual and Methodological Basics and Role Within Gerontology

2.1 Conceptual and Methodological Basics

The ecology of aging is particularly concerned with the description, explanation, and modification/optimization of the relationship between the aging person and his/her environment. Furthermore, by drawing from Lewin's basic insight that behavior must be seen as a function of the person and his/her environment, Lawton (1982) has suggested that research on the ecology of aging cannot proceed without acknowledging the 'ecological equation,' that is, the *interaction* between the aging individual and the environment as it relates to a variety of outcomes such as well-being, social behavior, or leisure activities. The term *transaction* has been suggested to underscore the notion that person and environment are hard to separate from each other and that the view of an ongoing process of mutual interchange may best reflect the everyday world (Altman and Rogoff 1987).

With these concepts in mind, research on the ecology of aging (as well as ecological research in general) was from the beginning concerned with examining the everyday life of individuals in their own day-to-day world. Hence, empirical methods, such as real-life observations in natural settings, the retrospective assessment of day-to-day behaviors through the use of diaries, or the reconstruction of internal representations of the environment have been preferred. Methods used to describe the environment, such as checklists for housing interiors or multidimensional evaluations of institutional settings, have also played an important role (Gitlin 1998).

Although ecological research has placed emphasis upon the physical environment, it is widely acknowledged that in reality, the physical and social environment are deeply interwoven. Some work on the ecology of aging has even put the major emphasis on the question of how social environments shape the elderly's behavior in specific socio-physical settings such as nursing homes (Baltes 1996).

2.2 Role of the Ecology of Aging Within Gerontology

Research on the ecology of aging strives to understand the environment-related tasks that typically confront the aging individual. These tasks include preserving autonomy in everyday life despite physical and mental impairments through the utilization of environmental resources within (including home modifications, see Gitlin 1998) and outside the home, initiating the relocation process if desired or necessary, and adapting to new living environments (such as a nursing home or other planned housing) after relocation. These classic research themes illustrate the fact that an interdisciplinary approach—covering disciplines such as psychology, sociology, architecture, geography, urban planning, and design—is essential for the understanding of the ecology of aging. Because the major research themes are so closely tied to the everyday world of the aging individual, empirical research on the ecology of aging also has strong practical implications and thus contributes significantly to the improvement of life quality in the later years.

3. Theoretical Approaches and Empirical Key Results

3.1 Overview of Theoretical Approaches

The ecology-of-aging paradigm encompasses a whole set of theoretical perspectives, all of which seek to understand how the socio-physical environment influences adaptational processes and outcomes in the everyday life of the elderly.

In the *Press–Competence Model*, the lowered competence of the older person, in conjunction with high 'environmental press,' is assumed to negatively impact on behavior in everyday life as is also expressed in the so-called *Environmental Docility Hypothesis* (Lawton 1982, Lawton and Nahemow 1973). For example, empirical research has revealed that the social interaction patterns of elders, particularly in planned housing, greatly depend on physical distances; long distances were found to effectively undermine social relations. The *Person–Environment Fit Model* has underscored the role of motivation within transactional processes (Carp 1987, Kahana 1982); basic as well as higher-order needs have been considered in their relation to the potential and limits of the environment. Empirical evidence supports the notion that a misfit between the resident's needs and the ability of the environment to fulfil these needs contributes to the explanation of lower life satisfaction. The *Person–Environment Stress Model* (Schooler 1982) argues that certain environmental conditions—such as lack of privacy or control in institutional environments—should be regarded as a potential threat for the aging process. Subsequently, intervention studies aimed to enhance the control of elders produced positive behavioral, emotional, and health outcomes (see Schulz and Heckhausen 1999, for a summary and theoretical integration of this research). Finally, the

Social Learning Model underscores the role of operant learning processes in circumscribed socio-physical settings, most notably care interactions in institutional settings (Baltes 1996). The empirical application of this model has produced profound insights into the effect of the social environment on dependent behavior in old age by identifying a robust interaction pattern characterized as the *Dependence–Support Script*. According to this script, the social environment of elders tends to strongly reinforce their dependent behaviors, while independent acts are mostly ignored.

3.2 Critique

Theoretical approaches within the ecology of aging have been critiqued for a number of reasons. For example, it has been argued that some of these models relegate a far too passive role to the elderly individual, who seems to be nothing more than a pawn of environmental forces. Lawton (1989) has addressed this critique by introducing the *Environmental Proactivity Hypothesis* which states that elders are also active and goal-directed shapers of their environments. It has also been argued that research on the ecology of aging has—in both theory and practice—been overly concerned with safety, orientation, and support (Parmelee and Lawton 1990). Meanwhile, the environment's potential for stimulating behavior and emotion, for creating strong cognitive–emotional bonds (*place attachment*; Rubinstein and Parmelee 1992) and thereby preserving one's personal continuity, has been neglected.

4. Future Considerations

Future research on the ecology of aging is faced with at least three challenges:

(a) *Theoretical challenges*. Most of the theoretical approaches in this field were developed in the sixties and seventies. One particularly fruitful theoretical avenue may be to strengthen the links between ecological theories and promising conceptual approaches within psychological gerontology. One prototypical example is provided by psychological control theories (see Schulz and Heckhausen 1999) which focus on person–environment relations and their alterations across the human life span.

(b) *Methodological challenges*. Although the use of naturalistic methods such as observations and *in vivo* data-collection strategies have proven themselves to be indispensable tools in research on the ecology of aging—as well as for gerontology in general—the methods used are rarely comparable. One of the most important methodological challenges is creating a single set of instruments to describe the physical and spatial environment of older people (Gitlin 1998).

(c) *Application and societal challenges*. Due to the rapidly growing number of older adults and particularly 'old-old' adults beyond the age of 85, one of the main challenges of today's and tomorrow's societies is providing care for the demented elderly. Research on the ecology of aging must be strongly committed to planning and evaluating supportive environments for these people. In addition, the growing impact of information and communication technology such as the Internet, the use of high-tech solutions in housing and in public community life, as well as the rapidly growing role of technology in geriatric care have given rise to a whole range of new environmental characteristics. Research on the ecology of aging will thus continue to encounter new challenges in the future.

See also: Aging and Health in Old Age; Caregiving in Old Age; Community Organization and the Life Course; Coping across the Lifespan; Human Behavioral Ecology; Human Ecology; Old Age and Centenarians

Bibliography

Altman I, Rogoff B 1987 World views in psychology: Trait, interactional, organismic, and transactional perspectives. In: Stokols D, Altman I (eds.) *Handbook of Environmental Psychology*. Wiley, New York, pp. 7–40

Baltes M M 1996 *The Many Faces of Dependency in Old Age*. Cambridge University Press, Cambridge, UK

Carp F M 1987 Environment and aging. In: Stokols D, Altman I (eds.) *Handbook of Environmental Psychology*. Wiley, New York, pp. 329–60

Gitlin L N 1998 Testing home modification interventions: Issues of theory, measurement, design, and implementation. *Annual Review of Gerontology and Geriatrics* **18**: 190–246

Kahana E 1982 A congruence model of person–environment interaction. In: Lawton M P, Windley P G, Byerts T O (eds.) *Aging and the Environment: Theoretical Approaches*. Springer, New York, pp. 97–121

Lawton M P 1982 Competence, environmental press, and the adaptation of older people. In: Lawton M P, Windley P G, Byerts T O (eds.) *Aging and the Environment: Theoretical Approaches*. Springer, New York, pp. 33–59

Lawton M P 1989 Environmental proactivity in older people. In: Bengtson V L, Schaie K W (eds.) *The Course of Later Life: Research and Reflections*. Springer, New York, pp. 15–23

Lawton M P, Nahemow L E 1973 Ecology and the aging process. In: Eisdorfer C, Lawton M P (eds.) *The Psychology of Adult Development and Aging*. American Psychological Association, Washington, DC, pp. 619–74

Parmelee P A, Lawton M P 1990 Design for special environments for the elderly. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*. Academic Press, San Diego, CA, pp. 464–87

Rubinstein R L, Parmelee P A 1992 Attachment to place and the representation of life course by the elderly. In: Altman I, Low S M (eds.) *Human Behavior and Environment, Vol. 12: Place Attachment*. Plenum, New York, pp. 139–63

Schooler K K Response of the elderly to environment: A stress-theoretical perspective. In: Lawton M P, Windley P G, Byerts T O (eds.) *Aging and the Environment: Theoretical Approaches*. Springer, New York, pp. 80–96
Schulz R, Heckhausen J 1999 Aging, culture and control: Setting a new research agenda. *Journal of Gerontology: Series B: Psychological Sciences* **54B**: P139–45
Wahl H W 1992 Ökologische Perspektiven in der Gerontopsychologie: Ein Blick in die vergangenen drei Jahrzehnte und in die Zukunft [Ecological perspectives in psychological gerontology]. *Psychologische Rundschau* **43**: 232–48

H.-W. Wahl

Ecology: Organizations

Inspired by the question, Why are there so many kinds of organizations? (Hannan and Freeman 1977), organizational ecology aims to explain how socio-political and economic conditions affect the relative abundance and diversity of organizations and to account for their changing composition over time. Ecological research typically begins with three observations: (a) diversity is a property of aggregates of organizations, (b) organizations' members often have difficulty devising and executing organizational changes fast enough to meet the demands of uncertain and changing environments, and (c) the community of organizations is unstable—organizations arise and disappear continually.

Consequently, contrasting adaptation approaches (see *Structural Contingency Theory*), which emphasize individual organizations' leaders cumulative strategic choices, ecological analyses formulate organizational change and variability at the population level, highlighting the creation of new and the demise of old organizations and populations.

Organizations, populations, and communities constitute the basic elements of ecological analysis. A set of organizations engaged in similar activities and with similar patterns of resource use constitutes a population. Organizational populations form as isolating processes, including technological incompatibilities, institutional actions such as government regulations, and environmental imprinting segregating one set of organizations from another. Populations develop relationships with other populations engaged in different activities that bind them into communities. Organizational communities are functionally integrated systems of interacting populations in which the outcomes for organizations in any one population are intertwined fundamentally with those in other interrelated populations. In broad terms, ecological theory and research focus on three themes: demographic processes, ecological processes, and environmental processes.

1. Demographic Processes

1.1 Age and Size Dependence

The relationship between organizational aging and failure is a central line of demographic inquiry. Until recently, the predominant view was the liability of newness (Stinchcombe 1965), which assumes that because they have to learn new social roles, create new organizational routines, and lack endorsements and exchange relationships, young organizations will have higher failure rates. A complementary viewpoint is that selection processes favor structurally inert organizations capable of demonstrating reliability and accountability, which requires high reproducibility. Since reproducibility, achieved through institutionalization and routinization, increases with age, older organizations should be less likely to fail (Hannan and Freeman 1984).

A related line of research examines how organizational size influences failure. Small organizations' propensity to fail results from several liabilities of smallness, including problems with recruiting and training employees, raising capital, and handling costs of regulatory compliance. Moreover, as organizations grow, increasingly they emphasize predictability, formalization, and control, increasing their reliability and accountability, and lowering their vulnerability to failure. The significance of 'smallness' stems in part from its relation to 'newness': since new organizations tend to be small, if small organizations have higher failure rates, empirical evidence of a liability of newness may actually reflect specification error. Although early studies consistently found a liability of newness, later studies controlling for time-varying size often do not, which has prompted formulation of alternative views on age dependence.

The liability of adolescence (Fichman and Levinthal 1991) predicts a $\cap$-shaped relationship between age and failure. New organizations begin with a stock of assets (e.g., goodwill, psychological commitment, financial resources) that buffers them during an initial 'honeymoon' period—even if early outcomes are unfavorable. The larger the initial stock, the longer the buffer. As initial endowments are depleted, however, organizations unable to meet ongoing resource needs because, for example, they are unable to establish effective routines or exchange relations, may fail.

Liabilities of newness and adolescence yield divergent accounts of age dependence for young organizations, but because processes underlying these models occur early in organizational lifetimes, they do not apply to older organizations. The liability of aging (Barron et al. 1994) identifies processes affecting older organizations and predicts that failure increases with aging. The liability of aging begins with another insight from Stinchcombe's (1965) essay: organizations are imprinted by their founding environment. As environments change, however, bounded rationality and

structural inertia make it difficult for individuals to keep their organizations aligned with environmental demands. Encountering a series of environmental changes may thus expose aging organizations to a risk of obsolescence. Aging may also bring about senescence: an accumulation of internal politics and precedent that impede action. Obsolescence and senescence pose separate risks. Senescence is a direct effect of aging; obsolescence a result of environmental change.

Among studies controlling for time-varying size, empirical estimates are mixed (Baum 1996). The divergent age dependence results might reflect nonproportionality (e.g., effect of age varies with size) masking 'true' age effects or alternatively variation in age dependence across populations. To resolve such questions, research must test alternative models' mediating variables directly rather than using age as a surrogate for all underlying constructs. Such an approach treats the alternatives as complementary rather than (falsely) as competing models.

1.2 Organizational Change

Most research on organizational change concentrates on content: a change to a more (less) advantageous configuration is considered adaptive (detrimental) (see *Organizational Change*). Complementing this focus, structural inertia theory (Hannan and Freeman 1984) emphasizes the change process and raises two questions.

1.2.1 How changeable are organizations? The structure in structural inertia theory refers only to core organizational features (goals, authority structures, core technology, and market strategy) related to mobilizing scarce resources for beginning an organization and the strategies used to maintain resource flows (Hannan and Freeman 1984). Peripheral features buffer an organization's core from uncertainty. Core features exhibit greater inertia than peripheral features, but the claim is not that they never change, only that 'the speed of reorganization is much lower than the rate at which environmental conditions change.' Nor does structural inertia theory remove individuals from responsibility or control over their organization's survival. Indeed, the challenge of overcoming structural inertia implies that the capacity of individuals to change organizations successfully is of great importance. Inertia also varies with age and size. Because older organizations have more formalized structures, standardized routines, institutionalized power structures, and dependencies, inertia should increase with age. Moreover, as organizations grow, they increasingly emphasize predictability, formalization, and control, becoming more inflexible. Large size, by buffering organizations from environmental demands, may also reduce the impetus for change.

Although not included in structural inertia theory, a complete understanding of change processes requires consideration of organizations' change histories (Amburgey et al. 1993). The more experienced an organization becomes with a particular change, the more likely it is to repeat it. If the change becomes causally linked with success in decision-makers' minds—rightly or wrongly—repetition is even more likely. The change process may thus itself become routinized, creating organizational momentum—the tendency to maintain direction and emphasis of prior actions in current behavior. To reconcile momentum with commonly observed periods of organizational inactivity, Amburgey et al. propose that local search processes (Cyert and March 1963) reinforce momentum immediately after a change occurs, but it subsequently declines.

Empirical studies yield mixed support for structural inertia theory; support for momentum appears stronger (Baum 1996). Overall, research indicates that, while inertia and momentum constrain them, organizations often initiate change in response to environmental shifts. Of course, inertia and momentum are not necessarily harmful: in addition to promoting performance reliability and accountability, in an uncertain environment they can keep organizations from responding too quickly and frequently to environmental change. Ultimately, whether inertia and momentum are adaptive depends on the hazardousness of change.

1.2.2 Is change beneficial? According to structural inertia theory, core change destroys or renders obsolete established routines, disrupts exchange relations, and undermines organizational legitimacy by impairing performance reliability and accountability. Thus, organizations may often fail precisely as a result of their attempts to survive—even attempts that might, ultimately, have proved beneficial. Because their structures, routines, and relationships are more institutionalized, older organizations are especially likely to experience disruption following core change. In contrast, large organizations can buffer themselves from disruption by, for example, maintaining both old and new routines during transition or overcoming deprivations and competitive challenges accompanying the change. If an organization survives the short-run hazard, its failure rate is predicted to decline as performance reliability, exchange relations, and legitimacy are re-established over time. Only if the rate of decline is faster than before the change, however, will the organization ultimately benefit from taking the short-term risk.

Empirical studies indicate that while core change is not necessarily disruptive in the short run, organizations do not necessarily improve their survival

chances in the long run either (Baum 1996). Two research design problems, however, appear likely to produce systematic underestimates of the hazardousness of change. First, the annual data typically used may not detect the deadliest core changes—those proving fatal within a year. Second, because organizations founded before a study period are not observed when they are youngest, smallest, and most likely to be subject to vulnerable change, their inclusion likely biases downward the risk of change.

2. Ecological Processes

2.1 Niche-width Dynamics

Niche-width theory (Hannan and Freeman 1977) focuses on two aspects of environmental variability to explain the differential survival of specialists, which possess few slack resources and concentrate on a narrow range of customers, and generalists, which attempt to appeal to the mass market and exhibit tolerance for more varied environments. Variability refers to the variance in environmental fluctuations about their mean over time. Grain refers to the patchiness of these variations with fine-grained variations frequent and coarse-grained infrequent. The key prediction is that in fine-grained environments, with large magnitude variations, specialists out-compete generalists regardless of environmental uncertainty. Specialists ride out the fluctuations; generalists are unable to respond quickly enough to operate efficiently. Thus, niche-width theory challenges the classic prediction that uncertain environments always favor generalists that spread their risk (see *Structural Contingency Theory*).

Table 1
Elaborations of the density dependence model

Model	Critiques[a]	Elaboration	Key references
Density delay	3	Incorporates imprinting effect of density at founding	Hannan and Carroll (1992)
Mass dependence	2,3	Weights each organization's competitive effect by its size	Barnett and Amburgey (1990)
Multiform models	2	Estimates intrapopulation microstructures and interpopulation (community) macrostructures	Barnett (1990)
Level of analysis	2,3	Estimates variation in density dependence across levels of geographic aggregation (city, state, region, nation)	Hannan et al. (1995)
Institutional embeddedness	1,4	Estimates population legitimation with relational density (endorsements by powerful actors and organizations)	Baum and Oliver (1992)
Nondensity measures	1	Estimates population legitimation with nondensity measures (certification contests, media content measures)	Rao (1994)
Localized competition	2,3,4	Weights competititive effects of organizations on each other by their proximity (size, price, location)	Baum and Mezias (1992), Baum and Haveman (1997)
Organizational niche overlap	1,2,4	Disaggregates population density for each organization based on resource (non)overlaps with other organizations	Baum and Singh (1994a, 1994b)
Location dependence	2,4	Incorporates information on spatial structure of organizational populations	Lomi (1995)
Time variation	1,2,3	Estimates variation in density dependence as a population ages	Baum (1995)
Dynamic selection	2,3,4	Incorporates variation in the strength competitive processes as a population ages	Baum (1995)
Competitive intensity	2,3,4	Weights each organization's competitive strength by its age	Barnett (1997)

a Critiques.
1 Proposed integration of ecological and institutional perspectives.
2 Assumption that each organization influences and is influenced by competition equally.
3 Inability to explain the concentration common to older populations.
4 Inattention to problems of unobserved heterogeneity.

In contrast to niche-width theory, which implies an optimal strategy for each population, Carroll (1985) proposes that, in environments characterized by economies of scale, competition among generalists to occupy the center of the market frees peripheral resources that are most likely to be used by specialists. He refers to the process generating this outcome as resource partitioning. His model predicts that in concentrated markets specialists can exploit more resources without engaging in direct competition with generalists, and so that while increasing market concentration increases the failure of generalists, it lowers the failure of specialists.

Although the specialist–generalist distinction is now common in ecological research, tests of niche-width and resource-partitioning theories' predictions are limited, and no studies explicitly contrast niche-width and resource-partitioning predictions. Recent studies frequently examine spatial and temporal environmental variation without reference to niche width, pointing to the need to elaborate niche-width models to encompass spatial as well as temporal environmental variation.

2.2 Population Dynamics and Density Dependence

Research on founding and failure in organizational ecology has paid considerable attention to endogenous population dynamics and density dependence. Prior population dynamics—patterns of founding and failure—shape current founding rates (Delacroix and Carroll 1983). Initially, prior foundings signal opportunity to entrepreneurs, encouraging founding. But as foundings increase, resource competition increases, discouraging founding. Prior failures have analogous effects. At first, failures release resources available to entrepreneurs. But further failures signal a hostile environment, discouraging founding. Resources freed by prior failures may also enhance the viability of established organizations, lowering failures (Carroll and Delacroix 1982). Density-dependent explanations for founding and failure are similar though not identical. Initial increases in the number of organizations can increase the legitimacy of a population, enhancing the capacity of its members to acquire resources. However, as a population continues to grow, competition with others for scarce common resources intensifies. Combined, the mutualistic and competitive effects imply a $\cap$-shaped relationship between density and founding and a $\cup$-shaped relationship between density and failure (Hannan and Carroll1992).

Studies of a wide variety of organizational populations provide empirical support for density-dependent founding and failure. By comparison, population dynamics findings are mixed and, when modeled together with population density, population dynamics effects are weaker and less robust. One explanation for the dominance of density dependence over population dynamics is the systematic character of density relative to the changes in density produced by founding and failure. A related explanation is the sensitivity of rate dependence estimates to outliers. These issues need to be examined more thoroughly before population dynamics is abandoned, which has been the trend in recent research.

Although support for density dependence is strong, it has received critical attention for its (a) proposed integration of ecological and institutional perspectives, (b) assumption that each organization influences and is influenced by competition equally, (c) inability to explain the concentration common to older populations, and (d) inattention to problems of unobserved heterogeneity. Several innovative elaborations, respecifications, and measures have been advanced recently to address questions raised by the initial formulation (summarized in Table 1). Although Hannan and Carroll (e.g., 1992, 1995) frequently rebuke such efforts, they nevertheless promise to improve substantive interpretation and realism of the original model.

3. Environmental Processes

Environmental processes are central to organizational ecology, shaping appropriate organizational forms and conditioning historical–structural relationships (e.g., the basis of competition).

3.1 Institutional Processes

Organizational environments are more than simply sources for inputs, information, and knowhow for outputs. Institutionalized rules and beliefs about organizations also figure prominently (DiMaggio and Powell 1983, Meyer and Rowan 1977). Institutional theory emphasizes that organizations must conform to these rules and requirements if they are to receive support and be perceived as legitimate. Density dependence theory views the relationship between ecological and institutional theory as complementary and synthesizes them within a single explanatory framework. However, institutional theory is increasingly conceived as contextual to ecological theory—the relationship between them is not only complementary, it is also hierarchical (Baum 1996).

From this point of view, the institutional environment constitutes the broader social context for ecological processes—the institutional environment may prescribe the criteria for judging whether an organization or entire population is worthy of continued survival. Ecological research on institutional processes

typically compares rates of founding, failure, change, or growth across organizational populations or time as the institutional arena of a particular population changes in terms of its political turbulence, government regulation, or institutional embeddedness (Baum 1996). Most past research has treated institutional change as exogenous. Studying how ecological dynamics of organizations shape institutional change can, however, enrich our understanding of institutional change and persistence.

3.2 Technological Processes

Technological innovation influences organizational populations profoundly by disrupting markets, changing the relative importance of resources, challenging organizational learning capabilities, and altering the basis of competition (see *Technology and Organization*). Supporting Schumpeter's characterization of technological innovation as a process of creative destruction, research supports the idea that technologies evolve over time through cycles of long periods of incremental change, which enhance and institutionalize an existing technology, punctuated by technological discontinuities in which new, radically superior technologies displace old, inferior ones, making possible order-of-magnitude or more improvements in organizational performance (Tushman and Anderson 1986).

The new technology can either be competence-enhancing, building on existing knowhow and reinforcing incumbents' positions, or competence-destroying, rendering existing knowhow obsolete and making it possible for newcomers to become technologically superior competitors. The technological ferment spawned by the discontinuity ends with the emergence of a dominant design, a single architecture that establishes dominance in a product class (Anderson and Tushman 1990), and technological advance returns to incremental improvements on the dominant technology. Although the universality of this technology cycle is debated, it has proved illuminating in a wide variety of industries.

Technological innovation creates opportunities for entrepreneurs to found new organizations and establish competitive positions as incumbents' sources of advantage decay. Technological innovation also creates uncertainty and risk for incumbents because its outcomes can be only imperfectly anticipated. An innovation's impact may not be known until it is too late for incumbents using older knowhow to compete successfully with new competitors; gambling too early on a given innovation may jeopardize an incumbent's survival if that technology turns out not to become dominant. Thus, underlying technologies and technological innovation may influence organizational populations' competitive dynamics and evolution profoundly. Ecological research relating technology cycles to population dynamics, although limited in scope, yields compelling support for this assertion (Baum 1996). Although past research typically treats technological change as exogenous, studying how ecological processes shape technological change can deepen our understanding of technology cycles by examining the dynamics of organizational support for new technologies.

See also: Darwinism: Social; Organizations, Sociology of; Population Ecology

Bibliography

Amburgey T L, Kelly D, Barnett W P 1993 Resetting the clock: The dynamics of organizational change and failure. *Administrative Science Quarterly* **38**: 51–73

Anderson P, Tushman M L 1990 Technological discontinuities and dominant designs: A cyclical model of technological change. *Administrative Science Quarterly* **35**: 604–33

Barnett W P 1990 The organizational ecology of a technological system. *Administrative Science Quarterly* **35**: 31–60

Barnett W P 1997 The dynamics of competitive intensity. *Administrative Science Quarterly* **42**: 128–60

Barnett W P, Amburgey T L 1990 Do larger organizations generate stronger competition. In: Singh J V (ed.) *Organizational Evolution: New Directions*. Sage, Newbury Park, CA, pp. 78–102

Barron D N, West E, Hannan M T 1994 A time to grow and a time to die: Growth and mortality of credit unions in New York City, 1914–1990. *American Journal of Sociology* **100**: 381–421

Baum J A C 1995 The changing basis of competition in organizational populations: Evidence from the Manhattan hotel industry, 1898–1990. *Social Forces* **74**: 177–205

Baum J A C 1996 Organizational ecology. In: Clegg S, Hardy C, Nord W (eds.) *Handbook of Organization Studies*. Sage, London, pp. 77–114

Baum J A C, Haveman H A 1997 Love thy neighbor? Differentiation and agglomeration in the Manhattan hotel industry. *Administrative Science Quarterly* **42**: 304–38

Baum J A C, Mezias S J 1992 Localized competition and organizational failure in the Manhattan hotel industry, 1898–1990. *Administrative Science Quarterly* **37**: 580–604

Baum J A C, Oliver C 1992 Institutional embeddedness and the dynamics of organizational populations. *American Sociological Review* **57**: 540–59

Baum J A C, Singh J V 1994a Organizational niche overlap and the dynamics of organizational mortality. *American Journal of Sociology* **100**: 346–80

Baum J A C, Singh J V 1994b Organizational niche overlap and the dynamics of organizational founding. *Organization Science* **5**: 483–501

Carroll G R 1985 Concentration and specialization: Dynamics of niche width in populations of organizations. *American Journal of Sociology* **90**: 1262–83

Carroll G R, Delacroix J 1982 Organizational mortality in the newspaper industries of Argentina and Ireland: An ecological approach. *Administrative Science Quarterly* **27**: 169–98

Cyert R M, March J G 1963 *A Behavioral Theory of the Firm*. Prentice Hall, Englewood Cliffs, NJ
Delacroix J, Carroll G R 1983 Organizational foundings: An ecological study of the newspaper industries of Argentina and Ireland. *Administrative Science Quarterly* **28**: 274–91
DiMaggio P J, Powell W W 1983 The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review* **48**: 147–60
Fichman M, Levinthal D A 1991 Honeymoons and the liability of adolescence: A new perspective on duration dependence in social and organizational relationships. *Academy of Management Review* **16**: 442–68
Hannan M T, Carroll G R 1992 *Dynamics of Organizational Populations: Density, Competition, and Legitimation*. Oxford University Press, New York
Hannan M T, Carroll G R 1995 Theory building and cheap talk about legitimation: Reply to Baum and Powell. *American Sociological Review* **60**: 539–44
Hannan M T, Carroll G R, Dundon E A, Torres J C 1995 Organizational evolution in a multinational context: Entries of automobile manufacturers in Belgium, Britain, France, Germany, and Italy. *American Sociological Review* **60**: 509–28
Hannan M T, Freeman J 1977 The population ecology of organizations. *American Journal of Sociology* **83**: 929–84
Hannan M T, Freeman J 1984 Structural inertia and organizational change. *American Sociological Review* **49**: 149–64
Lomi A 1995 The population ecology of organizational founding: Location dependence and unobserved heterogeneity. *Administrative Science Quarterly* **40**: 111–44
Meyer J W, Rowan B 1977 Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology* **83**: 340–63
Rao H 1994 The social construction of reputation: Certification contests, legitimation and the survival of organizations in the American automobile industry; 1895–1912. *Strategic Management Journal* **15** (Winter Special Issue): 29–44
Stinchcombe A L 1965 Social structure and organizations. In: March J G (ed.) *Handbook of Organizations*. Rand McNally, Chicago, pp. 153–93
Tushman M L, Anderson P 1986 Technological discontinuities and organizational environments. *Administrative Science Quarterly* **31**: 439–65

J. A. C. Baum

Ecology, Political

Political ecology had its origins in multidisciplinary efforts to integrate human and cultural ecology with political economy. The incorporation of ecology into political economy influenced political economists to take ecology more seriously, while political economy gave human and cultural ecologists a means to contextualize local level studies into broader political economic relations and into history. In the 1990s, some political ecologists have drawn on cultural studies and discursive approaches for insights into the ways that identity politics and the meanings ascribed to specific landscapes shape environmental politics at various scales—from the local through the global. This expansion has led to enhanced research by political ecologists on social movements, nongovernmental organizations (NGOs), gender relations, global through local environmental governance, and other areas. To date, however, political ecology is not a coherent theory oriented around a set of common assumptions, concepts, or research methodology. Rather, it is a dynamic and growing field that has become a forum for research, discussion, and debate among a diverse group of scholars concerned with the many ways in which power and politics affect and are affected by ecology and the biophysical environment.

1. *The Rise of 'Political Ecology'?*

Since the 1970s, political ecology has emerged as a multidisciplinary approach with which to analyze complex human–environmental interactions, especially those related to economic development and environmental destruction in the Third World (Peet and Watts 1996, Bryant and Bailey 1997). Although the field of political ecology has been greatly influenced by the 'regional political ecology' model set forth by geographers Piers Blaikie and Harold Brookfield (1987), political ecology is not as yet a comprehensive theory but a common ground where various disciplines intersect. Political ecology has been described in various ways: as 'a *framework* for analyzing human–environmental relations' (Grossman 1998); as a '*research orientation* that seeks to link macrolevel political economic processes with microlevel aspects of human ecology' (Dodds 1998); as a '*historical outgrowth* of the central questions asked by the social sciences about the relations between human society, viewed in its bio–cultural–political complexity, and significantly humanized nature' (Greenberg and Park 1994); and as a '*program or movement*... begun as a reaction to certain features of human ecology or ecological anthropology [as well as cultural ecology] as it was practiced in the 1960s and early 1970s' (Vayda and Walters 1999).

In part, political ecology emerged as part of a broader critique of cultural ecology, risk–hazard studies, and ecological anthropology that began in the 1970s and stemmed from the apolitical and microlevel analytical traditions of these fields. This critique was centered on the perceived neglect of the political dimensions of human–environmental interactions, the preoccupation with homeostatic processes, and the conceptualization of human communities as fairly autonomous and homogeneous (Painter and Durham 1995). The renowned anthropologist, Eric Wolf, was among the first to use the term 'political ecology' in his criticism of cultural ecological and ecological

anthropological approaches to the study of Alpine ecology (Wolf 1972). In his role as a discussant for an anthropological symposium on the 'Dynamics of Ownership in the Circum-Alpine Area,' Wolf emphasized the theoretical need to situate local ecological realities within the broader political economy.

In order to document trends in the historic rise of political ecology, the term 'political ecology' was used to search systematically a number of relevant digital databases of published articles and books, as well as theses and dissertations. These databases included the Social Science Index (scholarly articles published between 1974 and 1999), World Cat (books, articles, dissertations, reports published between 1000 and 1999), the University of California Digital Library (Melvyl–CAT, all books included in University of California holdings), and University of Michigan Dissertation Abstracts (theses and dissertations filed between 1861 and March 1999). Comparative analysis of the results of these searches revealed similar trends across databases. Between the early 1970s and the mid-1980s, a maximum of one article, book, dissertation, or report having the term 'political ecology' in the title of the work or as a specified key word and/or subject heading was found annually. From the mid-1980s through the early 1990s, the number of such works per database was in the range 5–10 annually (depending on the database). Between the mid- and late-1990s, this range increased to 15–25 per database.

In addition to showing a significant increase in the number of scholarly works defined as 'political ecology,' analysis of the search results also demonstrated a similar distribution of disciplinary interest among the authors of these works. Across the databases, geographers accounted for $\sim 33\%$ of the works, anthropologists 20%, political scientists 15%, sociologists 10%, and philosophers 4%. Other disciplines represented included history, education, agriculture, forestry, wildlife, economics, health sciences, public health, and urban/regional planning. Depending on the year, $\sim$ 65–85% of these studies were focused on the Third World. In addition to demonstrating the growing popularity of 'political ecology,' the search results also showed an expansion over time in the number of disciplines represented by the authors, suggesting that political ecology is being embraced by an increasing number of disciplines. Analysis of the search results also pointed out two major and separate threads within the works classified as political ecological. The first of these, and the one that is the focus of this review, emanated from, and was in part a reaction to, the earlier scholarly fields of cultural ecology and human ecology. The second major strand consisted of a loosely defined set of scholarship concerned more specifically with politics *per se*, including green movements and green politics, especially in Europe and at the international and global levels.

Although the increasing popularity of political ecology is indicated by the recent plethora of studies explicitly using the term, a growing number of essentially political ecological studies were conducted beginning in the 1970s. Although not specifically employing the term, these included work on various ecological crises, the role of the state in deforestation, the human and environmental repercussions of booms and busts in the production of export commodities, and the impact of colonial policies on human–environmental interactions (Little and Horowitz 1987, Grossman 1998). Especially prescient were the 'progressive contextualization methods' urged by Vayda (1983) and the 'cultural ecology of development' approach suggested by Grossman (1984).

Embraced by geographers, anthropologists, political scientists, sociologists, historians, and many others, political ecology is closely associated with the frequently cited work of geographers Piers Blaikie and Harold Brookfield (Blaikie 1985, Blaikie and Brookfield 1987) in which they developed their conceptual framework for 'regional political ecology.' Blaikie and Brookfield (1987, p. 17) provide this definition: 'Political ecology combines the concerns of ecology and a broadly defined political economy. Together this encompasses the constantly shifting dialectic between society and land-based resources, and also within classes and groups within society itself.'

Blaikie and Brookfield (1987, pp. 23–4) specify that 'regional political ecology' also is compatible with new directions in disaster and hazards research (e.g., Hewitt 1983) that linked hazardous events to processes of proletarianization and marginalization.

Although Blaikie and Brookfield are concerned primarily with land-based resources, other political ecologists have broadened the domain of political ecology to include other aspects of the environment such as marine resources, human biology, health and nutrition, and environmental governance and policy. In general, these varied political ecological analyses explain human–environmental interactions through their relations to political economic forces (Blaikie and Brookfield 1987). A vital concern is to understand local complexity in resource use within a larger, even global context. Central to Blaikie and Brookfield's framework is the resource user ('land manager') whose decisions must be evaluated in their 'historical, political, and economic context' (Blaikie and Brookfield 1987, p. 239). Also crucial to political ecological analysis is the relative power of, and collaboration among, relevant social actors that affect access to, and management of, natural resources. These stakeholders are then linked within and among levels of analysis through relations of power (Stonich 1993, Peet and Watts 1996). Many political ecologists follow the 'chain of explanation' suggested by Blaikie and Brookfield (1987, p. 27) and focus their research efforts, at least initially, at the local level, examining resource–use patterns in relation to households (as producers and/or consumers) and conflicts within and among households. It starts with the land managers

and their direct relations with the land (crop rotations, fuel-wood use, stocking densities, capital investments, and so on). Then the next link concerns their relations with each other, other land users, and groups in the wider society who affect them in any way, which in turn determines land management. The state and the world economy constitute the last links in the chain (Blaikie and Brookfield 1987, p. 27).

Political ecological studies, focused initially at the local level, often require the intricate analysis of human–environmental relationships that also was characteristic of earlier works in cultural ecology and ecological anthropology.

Political ecological analysis frequently integrates a number of essential components as outlined by Schmink and Wood (1987) in an early theoretical discussion. First, an assessment of the various contested ideologies that direct the use of resources and influence which stakeholders benefit from and which are disadvantaged by those ideologies. Also vital is a determination of the various international interests involved, including donor agencies, organizations, and private investors that promote particular patterns of natural resource use. Increasingly of concern is the function of the global economy in advancing specific patterns in access to and use of natural resources.

Crucial as well is the role of the state, particularly in terms of determining and implementing policies that favor the interests of certain social actors over those of others. Also central to the analysis is the relationship of class, ethnic, and gender structures to conflicts over access to productive resources, the interrelations among local resource users and groups of society who affect resource use, and diversity in the decisions of local resource managers (e.g., Stonich 1993, Rocheleau et al. 1996). Finally, political ecologists also tend to agree on the significant role of historical analysis in understanding the development of social and ecological relations and their subsequent links to environmental degradation (Blaikie 1985, Blaikie and Brookfield 1987, Guha 1990, Peluso 1993, Stonich 1993).

Efforts to characterize the evolving field of political ecology are difficult given the diversity, dynamism, and rapid growth of the field. One major focus of research using a political ecology approach has emphasized human impoverishment and environmental destruction (e.g., Johnston 1994) especially stemming from dominant development models (e.g., Stonich 1993, Painter and Durham 1995). These studies emphasized the relations among capitalist accumulation, state policy, surplus extraction, environmental destruction, and human impoverishment. Scholars often couched their arguments in opposition to prevailing popular explanations of environment degradation: Malthusian pressures, the irrationality of peasants, and/or technological ignorance (Painter and Durham 1995). A number of findings from these studies potentially can help understand deteriorating human and environmental conditions in other development contexts.

First, environmental destruction identified with the economic production systems of the poor usually is an outcome of their impoverishment, either absolutely or relative to other social classes. Impoverishment often is related to diminished access to land and other natural resources and to increased repression and violence at the hands of state authorities and more powerful individual and corporate interests engaged in land speculation. Moreover, environmental destruction associated with the decisions of the poor usually is intricately linked to existing environmental parameters, especially environmental marginality.

Second, as a consequence of their vulnerability and lack of power, smallholder producers often have received a disproportionate share of the blame for environmental decline. In contrast, political ecological research has demonstrated that a great deal of land and other natural resources has been degraded by the activities of more powerful private, public, and corporate interests. Large-scale enterprises that have acted destructively frequently have been granted land on concessionary terms by the state. This allows these more powerful stakeholders to treat land as a low-cost input, and makes it more economical to move elsewhere after the environment is degraded rather than try to conserve natural resources. Third, the same policies and practices that result in wealthy interests receiving land on favorable terms are responsible for the impoverishment of smallholders, because such policies institutionalize and exacerbate unequal access to resources. In sum, the political ecology approach applied to Third World development has shown that the crucial issue underlying environmental destruction and human poverty is blatant inequality in access to resources within a socially institutionalized context (Painter and Durham 1995). While other approaches, not labeled as political ecology, have found similar results, political ecological studies tend to emphasize these conclusions.

2. *Current Trends and Debates in Political Ecology*

Recently, the research agenda of political ecology has expanded to encompass a much wider and loosely delimited range of scholarship. Of special significance are efforts to understand the social construction of the environment and the increased integration of various cultural and discursive approaches into political ecological analysis (e.g., Escobar 1999). Reflecting the growing poststructural concern with discourse in the social sciences more generally, scholars such as Escobar emphasize that perceptions and discourses about environmental problems are social constructions that reflect specific backgrounds, values, and

positions of power rather than absolute truths. Diverse constructions and associated discourses frame contested points of view about what constitutes 'environmental problems' as well as possible/suggested solutions to these 'problems.' Coupled with a concern for political economy, these scholars are examining the plurality of perceptions and definitions of environmental and natural resource problems.

A second important trend is the 'serious treatment of *politics* in political ecology,' especially attempts to integrate political action (from passive resistance, to civil society, to political parties) into questions of resource access and control (Peet and Watts 1996, p. 10). Examples of this view of political ecology have emphasized various scales or levels of analysis from the domestic (household) to the global. For example, scholars such as Rocheleau, Thomas-Slayter, and Carney have focused their efforts at examining household level gender politics involving property rights, labor, access, and control of natural resources (Carney 1996, Rocheleau et al. 1996). In comparison, other scholars have focused on the more macrolevel and specific cases such as the World Bank and post-UNCED legal frameworks. These scholars have attempted to demonstrate the linkages between power relations, institutions, and environmental governance. Among the most important issues identified by studies that 'seriously' consider the politics of political ecology are the frequent occurrences of coercive patterns of environmental conservation. This was clearly demonstrated by Peluso (1993), who showed how state power and forestry institutions in Indonesia are contested by Indonesian peasants in the context of coercive conservation measures.

A related and expanding research interest for political ecologists is civil society—especially the proliferation of social and environmental movements (including the environmentalisms of the poor) and NGOs concerned with environmental governance at various levels—from the local, to the national, to the international and global. This trend, in addition to the previously discussed expansion of political ecology into the arena of politics, may represent a movement within political ecology to integrate the two separate strands of scholarship discussed earlier.

Political ecology has been criticized for overemphasizing the explanatory power of structural forces (i.e., the political economy) while downplaying the crucial role of local resistance, household survival strategies, inter- and intrahousehold conflicts over natural resources, as well as cultural beliefs (Stonich 1993). While several political ecologists have considered how local level forces interact with broader structural forces (e.g., Stonich 1993), increased attention to civil society may help restore the balance to what is essentially part of the broader structure–agency debate within the social sciences.

Finally, a nascent but exciting new direction aims to integrate political ecology into new ecological paradigms concerned with chaotic fluctuations, instability, and disequilibria. The shift from earlier systems models of ecology that stressed stability and resilience to the new ecology of chaos presents significant challenges and opportunities to the theory and practice of a political ecology concerned with the complex dynamics of local human–environmental relations (Peet and Watts 1996).

The prospect of integrating political ecology and the 'new' ecology also highlights a key issue in current debates over political ecology that centers on the place and priority of the environment itself in political ecological research and explanation (Grossman 1998). Critics have faulted political ecology for failing to give due consideration to the environment itself (e.g., Zimmerer 1996, Grossman 1998, Scoones 1999). These critics maintain that some studies labeled as political ecology focus almost *exclusively* on the effects of politics and political economic forces *on* the environment—treating the biophysical environment solely as modified by human actions. While most critics acknowledge the importance of considering political economic factors, they point out that it is also crucial to analyze the significance of the environmental variables themselves and how they interact with political economic and other forces to affect human–environmental relations. In their provocative critique of political ecology from this perspective, *Against Political Ecology*, Vayda and Walters (1999) maintain that 'It may not be an exaggeration to say that overreaction to the "ecology without politics" of three decades ago is resulting now in a "politics without ecology".'

To some extent, the perceived imbalance between the priorities given to political economic and ecological factors in political ecological analysis and explanation pointed out by recent critics is tied to the growing importance of poststructuralism/postmodernism and discourse theory that is apparent in many of the new directions in political ecology. With its emphasis on social construction and discursive formations, discussed previously in this review, the constructionist approach has added a significant critical dimension to the earlier work by political ecologists. Especially important have been constructivist examinations of the role of science in environmental debates and the role of discourse in framing environmental problems and conflicts (especially but not exclusively in the Third World). Despite these contributions, various critics have pointed out several limitations associated with the strong idealist position inherent in constructionist arguments (see, e.g., the commentaries following Escobar's article in *Current Anthropology* (Escobar 1999)). First, the concept 'cultural and/or social construction' of the environment can overestimate the power of human behavior and actions to create, transform, or otherwise control environmental forces while simultaneously underestimating the transformative power of the

environment. The profound repercussions of the 1997–98 El Niño Southern Oscillations (ENSO) and La Niña events and of Hurricane Mitch in 1998 are persuasive examples of the power of the environment as an active agent capable of provoking widespread and significant human response. Critics maintain that political ecology must balance the cultural/social construction of the environment with a meaningful and comprehensive analysis of the environmental construction of the social and the cultural. Furthermore, they suggest that an overemphasis on constructivist discourse analysis may diminish the concern for the material issues and practice that first provoked the emergence of political ecology. From the perspective of more materially oriented political ecologists, the importance of understanding discursive formations lies precisely in what that understanding reveals about the behaviors of the diverse actors involved in social and environmental problems and conflicts.

3. *Future Directions*

The field of political ecology developed initially to provide a political economic perspective on the causes of environmental problems and degradation. During the 1990s, research interests broadened considerably but the concern with the fundamental importance of political economy remained central to the field. While political ecology may represent an emerging research agenda, neither past developments nor recent trends are theoretically integrated and it remains to be seen where existing convergences and tensions may lead. Current trends and new directions in political ecological research reveal significant stress between material and idealist perspectives that are linked to perhaps irreconcilable differences between scientific and poststructural/postmodern positions. There have been some efforts, however, to bridge the gap between these two disparate and often divisive positions (e.g., Peet and Watts 1996, Rocheleau et al. 1996). The fact that political ecology has engendered much recent debate and discussion among proponents and opponents with vastly different points of view suggests a vibrant and maturing field. It is not clear, however, if political ecology will mature into a unified field with its own comprehensive theory or rather remain an intellectual meeting ground for scholars from diverse disciplines. There is considerable debate among practitioners about the appropriate future for political ecology. There is still much to be done. As yet, a core literature has not been established and the growing number of studies conducted within the rubric of political ecology does not share any unifying principle, concepts, or methodology. Political ecologists have yet to address adequately crucial issues of scale—particularly the discontinuity between local-level complexity and specificity and larger-scale generalization and representation. In addition, most research to date has been multidisciplinary rather than interdisciplinary, despite the fact that much political ecological research has demonstrated the need for interdisciplinary approaches to understand the complex dynamics of human–environmental relations and ameliorate problems (e.g., Stonich 1993). A key implication of this review is that further research is warranted to expand the theoretical framework of political ecology, to enhance its interdisciplinarity, and to test its applicability in particular contexts as well as at the global level.

See also: Ecological Imperialism; Ecology, Cultural; Feminist Political Ecology; Geography; Human–Environment Relationships; Place in Geography; Place, Power of; Political Geography

Bibliography

Blaikie P M 1985 *The Political Economy of Soil Erosion in Developing Countries*. Longman, London

Blaikie P M, Brookfield H 1987 *Land Degradation and Society*. Methuen, London

Bryant R L, Bailey S 1997 *Third World Political Ecology*. Routledge, London

Carney J 1996 Converting the wetlands, engendering the environment. In: Peet R, Watts M (eds.) *Liberation Ecologies: Environment, Development, Social Movements*. Routledge, London, pp. 165–87

Dodds D 1998 Lobster in the rain forest: the political ecology of Miskito wage labor and agricultural deforestation. *Journal of Political Ecology* **5**: 83–108

Escobar A 1999 After nature: steps to an antiessentialist political ecology. *Current Anthropology* **40**: 1–16

Greenberg J B, Park T K 1994 Political ecology. *Journal of Political Ecology* **1**: 1–12

Grossman L S 1984 *Peasants, Subsistence Ecology, and Development in the Highlands of Papua New Guinea*. Princeton University Press, Princeton, NJ

Grossman L S 1998 *The Political Ecology of Bananas: Contract Farming, Peasants, and Agrarian Change in the Eastern Caribbean*. University of North Carolina Press, Chapel Hill, NC

Guha R 1990 *The Unquiet Woods: Ecological Change and Peasant Resistance in the Himalaya*. University of California Press, Berkeley, CA

Guha R, Martinez-Alier J 1997 *Varieties of Environmentalism*. Earthscan, London

Hewitt K (ed.) 1983 *Interpretations of Calamity, from the viewpoint of Human Ecology*. Allen and Unwin, London

Johnston B R 1994 *Who Pays the Price? The Sociocultural Context of Environmental Crisis*. Island Press, Washington, DC

Little P D, Horowitz M M, Nyerges A E (eds.) 1987 *Lands at Risk in the Third World: Local-level Perspectives*. Westview Press, Boulder, CO

Painter M, Durham W H (eds.) 1995 *The Social Causes of Environmental Destruction in Latin America*. University of Michigan Press, Ann Arbor, MI

Peet R, Watts M (eds.) 1996 *Liberation Ecologies: Environment, Development, Social Movements*. Routledge, London
Peluso N L 1993 *Rich Forests, Poor People: Resource Control and Resistance in Java*. University of California Press, Berkeley, CA
Rocheleau D, Thomas-Slayter B, Wangari E 1996 *Feminist Political Ecology: Global Issues and Local Experiences*. Routledge, London
Schmink M, Wood C 1987 The 'political ecology' of Amazonia. In: Little P D, Horowitz M M, Nyerges A E (eds.) *Lands at Risk in the Third World: Local-Level Perspectives*. Westview Press, Boulder, CO, pp. 38–57
Scoones J 1999 New ecology and the social sciences: what prospects for a fruitful engagement? *Annual Review of Anthropology* **28**: 479–507
Stonich S C 1993 '*I Am Destroying the Land!' The Political Ecology of Poverty and Environmental Destruction in Honduras*. Westview Press, Boulder, CO
Vayda A P 1983 Progressive contextualization: methods for research in human ecology. *Human Ecology* **11**: 265–81
Vayda A P, Walters B B 1999 Against political ecology. *Human Ecology* **27**: 167–79
Wolf E 1972 Ownership and political ecology. *Anthropology Quarterly* **45**: 201–5
Zimmerer K 1996 Ecology as cornerstone and chimera in human geography. In: Earl C, Mathewson K, Kenzer M S (eds.) *Concepts in Human Geography*. Rowman and Littlefield, Manham, MD, pp. 161–88

S. C. Stonich

Econometric Software

This article will describe some of the characteristics of the most widely used econometric software, trace the evolution of this particular type of computer program, and explore the role of computer software in the development of applied econometrics. In order to discuss econometric software it is necessary, first, to define econometrics. We begin the discussion in Sect. 1 with a brief introduction to econometrics. Sect. 2 describes the distinctive features of the current vintage of econometric software and the progression of developments that has produced it. Sect. 3 discusses some of the common characteristics of software in econometrics. Finally, Sect. 4 will suggest some of the inter-relationships between econometric software and popular methods in econometrics.

1. *Econometrics*

Econometrics is a collection of methods and tools used to fit equations (economic models) to data. It involves both theory and measurement, and an overarching view of the process by which data come to be observed. The analysis begins with a theoretical explanation of some phenomenon, or a hypothesis about how some variables to be observed are related to each other. Measurement begins with the recording of some aspect of economic activity. Model building is the process by which these economic measurements are organized within the framework of one or more mathematical equations suggested by the economic theory. The principle behind this analytical framework is that of a 'data generating process' (DGP). There is an underlying statistical regularity at work that leads to the analyst's observation of their measurements. It is the characteristics of this underlying regularity, usually formalized as a probability distribution over random events, which are the objects of the econometric study.

In a familiar application of theory to data, observers collect pairs of observations on prices (P) and quantities (Q) exchanged in a market, such as the market for tobacco or fine wine, under the umbrella of a theory of 'demand.' The theory is embodied in a 'demand equation,' $Q = f(P \mid \alpha,\beta)$, which states that quantities traded are influenced by prices, and the nature of that response can be described by the relationship $f(\ldots)$ and the 'parameters,' α and β. The equation is made operational by computing statistics that provide information about the co-variation of prices and quantities transacted in the market. Thus, β might be an overall measure of how responsive quantities are to changes in prices, and the data are used to deduce information about the specific value of β.

In an econometric analysis, the information in a data set which may include thousands or millions of values is distilled to a few equations which are claimed to describe the DGP. The intricacy of modern econometric techniques makes essential elaborate software with capabilities that extend well beyond the ubiquitous spreadsheet programs such as Excel (4) and Lotus (10) (which would be the nearest cousins to the software considered here). (Numbers in brackets refer to the list of computer programs in the Appendix.) Econometric software is used most frequently for the manipulation of large amounts of measured data for the purpose of constructing the equations that are the model of the process which produced the data.

2. *The Evolution of Econometric Software*

The earliest software in econometrics was single purpose sets of Fortran language ('Formula Translation') code. The researcher who needed a program for a particular purpose would typically write it, or borrow it from someone else. One can find some of these programs in the appendices in some of the classic pieces in the literature, such as Nerlove and Press's (1973) development of the multinomial logit model, the first edition of Draper and Smith's (1981) exposition of regression analysis, Zellner's (1971) early

work on Bayesian estimation, and, more recently in Kalbfleisch and Prentice's (1980) work on survival modeling. A typical journal treatment is Krailo and Pike's (1984) exposition of a logit model for panel data, which contains a Fortran program listing for a standalone estimation program. Until the late 1960s, the journal *Econometrica* occasionally would publish short articles documenting computer programs developed by the authors.

2.1 Program Libraries

The process of building applications from scratch was made easier by 'libraries' of interchangeable parts for software development. The precursor to a very large group of programs currently distributed as the 'BMDP Library' (2) was a group of individual routines developed at UCLA in the 1960s under the heading of the 'BMD' (BioMedical) programs. These special purpose routines included subroutines for linear regression, Spearman rank correlation coefficients, cross tabulations, and so on. The BMDs initially existed as boxes of 80 column punched cards of Fortran code in the drawers of statistics labs. They maintained a common structure that allowed authors to write their own data setup programs to 'plug into' the package. Another programmer's toolkit that was circulated widely in the 1970s was the 'IBM SSP' (Scientific Subroutine Package) which was part of the documentation for IBM mainframe computers. This book contained Fortran and C listings of programs for such functions as inverting or decomposing a matrix, and computing a variance. ('C' does not stand for anything. It was simply an extension of 'B,' itself a derivative of 'BCPL' (the BC programming language) and a contemporary of APL (a programming language). All were languages originally written to work with the Unix operating system on DEC's original PDP-11 computers.

Most of the difficult computations being done by econometricians at this time involved complex matrix algebra, data manipulation, and random number generation. A library of well-written subroutines combined with the efforts of the analyst was quite sufficient for application development. This approach to programming has not vanished from the landscape. The *Numerical Recipes* books (C, Pascal, and Fortran) by Press et al. (1986), contain the modern counterpart to the expertly written library of subroutines contained in the SSP. The BMD libraries still exist as the commercial BMDP (2) routines, and the IMSL programs (6) and NAG (Numerical Analysis Group) (14) routines are still used by scientific programmers and in the development of major software packages. (IMSL, itself, grew largely out of a public domain library of matrix programs, EISPACK/-LINPACK. These provided the backbone for many early matrix implementations. They have since evolved to the (still) public domain package LAPACK (www.scd.ucar.edu/softlib/LAPACK.HTML and, see also NetLib at ORNL).

2.2 Integrated Packages

Early econometrics programs were single purpose Fortran programs that users passed amongst themselves and which grew over time as they were extended. The concept of a multipurpose, integrated econometrics package emerged in mid-1960s in the form of three particular programs, TSP (Time Series Processor), the Troll system, and B34S (which began as BiMed34; see Stokes 1997). TSP (23) and B34S (1) have remained active and are still under development. The early incarnations of these packages were primarily devoted to different forms of linear regression and correlation, but they have grown into vast assemblages of hundreds of thousands of lines of code (still mostly Fortran). Development of today's two largest commercial statistics and econometrics packages, SAS (17) and SPSS (20) also began at this time, though generally in social sciences other than economics and focused on techniques largely different from those of interest to econometricians. The demise of the distribution of custom written, single purpose software began in the 1970s with the emergence of TSP, SAS, SPSS, and so on.

2.3 Development Settings

With the exception of E-Views (3), which is a fairly recent arrival, and of the language platforms, MatLab (13), S-Plus (16), Gauss (5), Maple (11), and Mathematica (12), which are arguably successors to Fortran, most of the major packages in current use began development as part of a cottage industry in the 1960s and 1970s, usually by an individual or small team of innovators. This is a distinct difference from the contemporary, corporate route to major software development. The modern development centers, such as SAS, Inc. and SPSS, Inc. are creations of the 1980s and 1990s. Nonetheless, many of the programs listed in the Appendix to this survey continue to be developed by small numbers of statisticians and econometricians whose names would be recognized by practitioners (most likely in association with literature encountered as part of their graduate studies).

2.4 Programming Languages

Wither the primitive programming languages, Fortran and C? Fortran dominated the scientific computing landscape for nearly four decades until the early 1990s,

and remains the language of choice for many scientific applications, e.g., in physics and astronomy. The practitioner interested in writing a Fortran or C program may use one of the development kits for the PC (Windows) setting, such as Lahey (7) and Watcom (24). Time-shared systems, such as Sun Microsystems, also provide Fortran and C compilers and program libraries as part of the operating platforms. These vendors have extended the lives of Fortran and C as numerical language platforms, but this style of programming is a dying art. For software developers, Fortran is giving way to C, for example, in the full conversions of RATS (15) and SAS (17) and in the mixed language construction of LIMDEP (8). Individuals doing custom numerical programming increasingly rely on Gauss and its counterparts, MatLab, etc.

Recent developer's programming language implementations have gravitated toward the studio format. The studio platform consists of integrated tools for merging code in one or more languages along with numerous libraries and ancillary tools. The Microsoft Software Developers Kit (SDK) is a popular example. This approach to programming has become essential for the development of major programs. One of the very useful innovations is the seamless creation of mixed language (e.g., Fortran and C++) platforms that allow developers to exploit the best features of both these languages. The end result is that C, its successor, C++, and Fortran, continue to play a major role in the development of modern econometric software, but only in the laboratories of developers, rather than in the toolkits of practitioners.

2.5 The Internet

Finally, in what may be a surprise given the breathless surrounding hype, it appears that the Internet has had little or no impact on econometric software beyond facilitating the transfer of files, programs, and information. It is otherwise difficult to discern any ways in which the actual mechanics of econometric computing have been affected by the creation or expansion of the Internet.

3. The Characteristics of Modern Software

The software used by modern econometricians is essentially of two types. Most empirical research is done using one of the large, integrated packages that contain collections of programs for a variety of models. (See the 'Econometrics' list in the Appendix.) In these programs, complicated models are built with single, compact instructions. If a technique has gained acceptance in the applied literature, then it will appear as an 'option' in at least some of these packages. In contrast, when a technique or model proposed by an econometrician involves new methods or novel kinds of computations, then it is not likely to exist yet as a feature in an existing estimation platform. The computational tools that this analyst needs will typically be developed using one of the more primitive programming languages, such as Gauss (5), MatLab (13), S-Plus (16) or, at an even lower level, one of the native programming languages, Fortran or C++.

3.1 Packaged Expertise

The term canned routine has long been used for programs that embody the econometric expertise needed by users. Some have argued that the use of such software locks practitioners into a limiting set of models. While the argument may once have had some merit, the huge variety of techniques and the customizability provided by most large packages makes the criticism moot for current purposes. In addition, contemporary software is mature enough that developers have incorporated optimal algorithms that purists who mimic the textbook mathematical formulations might not use, or even be aware of. Least squares provides an excellent example. The analyst who writes his own program for linear least squares regression is likely to rely upon the textbook matrix formula, $b = (X'X)^{-1}X'y$. But, this is not the most accurate method of computing a least squares coefficient vector. Many large packages use much more accurate techniques, such as QR and singular value decompositions that avoid inversion and moments accumulation altogether. For the individual researcher to develop complicated low level tools such as this may well be excessively difficult.

3.2 User Written Programs

The distinction between these two types of software is somewhat blurry since the programming languages are often distributed with packages of routines that have been written by expert users. In addition, developers of code for Gauss (5) and MatLab (13) often distribute their programs on the Internet simply by posting them on their own websites, perhaps along with the research report of which the software development was a part. For example, all the major integrated econometrics packages contain an estimator for the logit model for binary choice, whereas the Gauss user must program their own. However, so many widely distributed Gauss libraries of routines contain users' versions of the logit program that for practical purposes, Gauss does contain a logit estimator. In the other direction, large integrated packages usually contain their own 'programming languages,' which allow the user to construct capabilities that are not already contained in the program. Thus, while a program such as SAS (17) may not contain a specific command for estimation of a certain kind of survival model, it does contain a programming language for matrix manipulation as well as a means for a user to access certain utilities such as a nonlinear optimization program. Thus, the SAS user can develop

a new estimator within his familiar SAS platform. TSP (123), SAS (17), SPSS (20), Stata (21), Shazam (18), EViews (3), and LIMDEP (8) to name a few, are all large integrated packages that also contain extensive matrix and data manipulation languages and access to nonlinear optimization programs which allow users to extend the modeling capabilities of the package. Thus, software developers and their users have blurred the distinction between preprogrammed procedures and programming languages.

3.3 Application Studios

The modern applied econometrician will usually maintain a toolkit of software consisting of one or more large packages, and, for less frequent usage, a low level development kit such as Gauss or MatLab, or a Fortran compiler, such as Lahey (7). Developers, conscious of this and sensing an opportunity, have begun to gravitate to a developer's studio format for econometric software. As noted, all of the large packages provide low to medium level programming capability. Current trends, such as EViews, LIMDEP, RATS, SAS, Stata, and TSP have brought the programming features closer to the front end of these large, high level packages and Gauss with its large, expanding, and widely distributed library is approaching the same format from the opposite direction.

3.4 Common Features

Most general, integrated econometrics programs are quite similar. The model building process dictates the design, and as in many settings, a standard has evolved naturally. The following functions will be found in all the integrated packages, and, over time, will be mimicked by the users of the language packages as they collect their own routines:

(a) Data input and transformation;

(b) Model estimation—the estimation of model parameters, usually of several types;

(c) Statistical inference—testing hypotheses and forming probabilistic statements about quantities of interest;

(d) Specification, testing, and iterative refinement of model specification;

(e) Manipulation of model results for forecasting and simulation;

(f) Graphics and display of results.

Programs differ in some features and there is some amount of specialization, but most or all of the preceding will be found common to all. (Reviews are regularly published, e.g., in the *Journal of Applied Econometrics*.) The second function listed is one of the major dimensions on which programs differ. For example, EViews and RATS are particularly oriented toward analysis of time series data, SAS and Stata are especially complete in their presentation of techniques used in biomedical research, and LIMDEP is best known for its selection of models for discrete dependent variables. (Markets being what they are, there is substantial and growing overlap of these features.)

3.5 Interfaces

The means of communication between the user and the software is the interface. In order to describe this, it is necessary to make a distinction between mainframe and 'PC' (desktop style) operation. A second distinction between 'batch' and 'interactive' mode use is also necessary. The two distinctions are useful for describing the way certain computer programs work, although both modes do appear in both settings. Econometric software design began on large time shared mainframe systems. The design of such systems necessitated a batch mode form of interface. In this case, the user submits a set of instructions to a computer (operator, real or virtual) and returns later to retrieve their desired results. Modern computing carried out on the desktop, in contrast, usually involves an individual operating their own system interactively in the fashion of a large calculator, submitting single instructions (perhaps in the form of mouse clicked selections from menus) and instantly receiving the results of their directive. Econometric analysis often requires a hybrid of these two modes. Many estimation problems require huge amounts of computation that may take hours or days to complete, or lengthy scripts of operations. As such, a batch mode of operation is often necessary, even on a desktop computing platform. Therefore, there is normally some means of giving the program a directive equivalent to 'carry out the following sequence of commands,' or 'fetch the set of commands contained in a certain file on disk and carry them out.'

There is a qualitative difference in the typical look and feel of a desktop computer compared to a time-shared mainframe. Modern desktop software is based on graphical user interfaces (GUIs). The amount of menuing varies from one program to another, but all modern econometrics programs are based on the GUI format. In contrast, there is relatively little GUI in the time-shared environment. The distinction is only superficial, however. The numerical functionality of most econometric software is easily replicated in both mainframe and microcomputer environments (e.g., all of the econometrics programs listed below except EViews). This is one of the strong distinctions between econometric software and, say, the programs used for graphics design or word processing. In the final analysis, the purpose of most econometric software is numerical analysis that can be done equally well in batch or interactive mode.

3.6 Estimation Procedures

To describe some of the remaining features, it is convenient to define some terms. Estimation is a

process of culling from a data set a set of statistics that summarize the process by which the data were created. Estimation results range from single numbers, such as the parameter β in our earlier demand equation example, to one to one transformations of the original data, such as a periodogram, nonparametric hazard function, or kernel density estimator. Econometric software will contain 'estimation routines.' The term derives from the earliest days in which Fortran 'subroutines' were written for these purposes. The estimation routine consists of a body of code written to be invoked with a simple 'command' and provided with the specifics of a user's problem. The command itself or the broader context of use of the software must define the kind of model to be estimated. Commands might be of the form 'Using the data on Price and Quantity, fit a linear regression model.'

For example, one popular program, LIMDEP, uses the following, fairly typical style:

Regression; Lhs = Quantity; Rhs = One,Price $

that states in words that the user wishes the program to compute the constant term (ONE) and slope in a linear regression of quantity on price, $Q = \alpha + \beta P + \varepsilon$. All integrated packages have counterparts to this kind of command, which is provided to the program by means of some kind of text processor. Some packages provide the same functionality by asking the user to select the estimation framework, then the variables in the equation from menus, instead. Many programs also allow users to choose either mode of entry. Each has its virtues—the former allows recording of steps and reuse of commands whereas the latter is generally more convenient.

A programming language such as Gauss would not contain such a broadly defined procedure. Rather, the Gauss processor would allow the user to define, using matrix expressions, the algebraic steps involved in doing the computation, including how to compute the coefficients for the model and how to compute certain statistics that the integrated processor would produce automatically. The user would then use the language's built-in functions for displaying the results of their computations. Thus it would seem that a great deal more work is required of the Gauss user than the LIMDEP user. There is, but two issues motivate the former. First, as noted, many such sets of computations already exist, in elaborate routines that programmers have written and disseminated, so that these procedures can be used, often with a single command. Second, many estimation techniques do not exist (yet) in large integrated packages, so this is the only way to proceed. (Since this market is fast moving, it would be inappropriate to mention specific techniques. By the time this article appears in print, the information will be out of date.)

The remaining functions involve manipulation of model results. Some of these are handled automatically by the software while others must be assembled by the user. At this juncture, the user of the integrated procedure is likely to use some of the programming features of the package, such as a matrix computation. For example, the literature often produces new statistics or ways to manipulate the results produced by familiar models such as the linear regression model. The SAS, Stata, or TSP user is likely to handle this situation by constructing a 'procedure' with 'adjustable parameters' that might appear as follows:

'Procedure: Regression;

Dependent variable is (Y), Independent variables are (x1, x2)

Compute residuals as variable (e).

Compute and report statistics using procedure …

End of Procedure'

Now, to invoke their procedure, the user might use a single directive such as:

'Use Regression Procedure (Quantity, One, Price).'

The generality of the procedure allows it to be used for different problems just by changing the 'call' statement. At this point, the only difference between Gauss and SAS is the detail that will have to be programmed in the first line of the procedure.

3.7 *Accuracy*

One might think that given the textbook formulas for the necessary computations, all programs would get the same right answer for any particular problem. But, in fact, this is far from correct. Even in the absence of programming errors, there are large differences in the algorithms and methods used in computer packages. For practical purposes, two approaches are used for validating econometric software: replication and benchmarking. The first of these is strictly pragmatic. For most techniques, there are no applications with known *correct* answers against which to benchmark a program. As such, it is assumed that if two or more of the major programs give the same answer for a problem, then they both have it right. If they differ, then the tricky issue arises as to which, if either is correct and, again, the issue would have to be resolved in ad hoc fashion. A contemporary study that pursues this line of inquiry is Silk (1996), whose review of estimators for simultaneous equations models presents some surprising differences (see, also, McCullough and Vinod 1999.) The same approach has been taken in a review of programs for the computation of GARCH models by McCullough and Renfro (2000).

True benchmarking is now available for many common calculations. The tradition in econometrics begins with Longley's (1967) true results to 10 significant digits for a specific problem in linear regression. The Longley standard was used for many years as the yardstick for linear regression programs; if a program could match the Longley benchmark to at least eight digits, then it could probably rise above suspicion for practical usage. Later analysis has showed that the Longley hurdle is not particularly high, and clearing it is not all that difficult, even for fairly mediocre software. The Statistical and Engineering Division of the National Institute of Standards and Technology (NIST) has published (www.nist.gov/itl/div898/strd) a suite of 57 benchmark problems in four classes: univariate statistics, analysis of variance, linear regression, and nonlinear regression. The NIST problems have been solved using very high precision methods, and correct answers have been published. In detailed reviews, McCullough (1998, 1999) found fairly wide differences in the performance of several econometrics packages. Since econometrics programs tend to be revised, refined, and updated very frequently, these differences are likely no longer to be evident by the time this article is in print.

4. *The Role of Software in the Development of Econometrics*

A convincing case can be made that several parts of modern econometrics awaited the development of software to gain currency.

In primitive incarnations, econometric software operated on a table of data with variables in columns numbered from left to right. To communicate a desired computation to a program, the user would specify columns of the table, as in 'Regress the variable in column 5 on a constant, and the variables in columns 6, 8, and 10.' TSP was among the first to allow names for variables to appear in a symbolic instruction to the program. The software went an important step further in that it allowed the components of equations to appear symbolically in words and familiar algebraic symbols. Software had begun to evolve in this direction, but TSP brought the language of equation specification up to the level of compilers such as Fortran. This major innovation changed the standard for econometric estimation programs.

The most important single development in practical econometrics in the period prior to 1970 was arguably the LSQ (Least Squares) procedure in TSP. The combination of symbolic equation specification with TSPs LSQ procedure allowed the birth of a very large literature on linear and nonlinear equation systems. The seemingly unrelated regressions model had been proposed a decade prior by Zellner (1962) and Theil (with Zellner, 1962), but the model did not appear in wide use until the appearance of LSQ. The procedure went an important step further in that it allowed the Zellner estimator to be extended to nonlinear systems of equations.

At roughly same time, Joreskog proposed a maximum likelihood approach to factor analysis and the estimation of general linear equation systems. The analytic results had been known for some time, but solvers for the likelihood equations for a general hierarchical linear equation system had not been widely distributed. Joreskog's (1970) early work included 'A Computer Program for Minimizing a Function of Several Variables,' (Educational Testing Services: Research Bulletin 70-14). In this was published the numerical engine of a package published two years later as 'LISREL' (Linear Structural Relationships—originally, ACOVS for Analysis of Covariance Structures). ETS 70-14 contained a listing of Fortran computer code for optimization, including a very clever and effective technique for the line search step in optimization. LISREL spawned a very large and still growing industry in structural equations analysis in the social sciences. The technique is a major part of the methodology; the program, itself, has become Scientific Software, Inc. (9).

Of somewhat more recent vintage, is Heckman's (1979) sample selection model. Building on his work on labor supply models, Heckman proposed a method for a regression model with a particular type of nonrandom sampling in a seminal paper in 1979. The computations involved are now fairly routine and incorporated in most econometrics packages, but in the mid-1980s, the technique became a standard tool in the microeconometricians kit as the LIMDEP program (8) became the *de facto* standard approach for implementing it.

The work of Sims (1977), Granger (1969), and Litterman (1986) in the mid-1970s on vector autoregressions (VARs) in econometric time series analysis produced a new body of theory to challenge the structural simultaneous equations models. The RATS (Regression Analysis of Time Series) (15) computer program, developed by these authors carried the methodology into the mainstream of applied econometrics. The development has also been ongoing, both of the program (Estima, Evanston, IL) and of the methodology as can be seen in any reference on time series, such as Hamilton (1994). (RATS has recently spawned CATS—Cointegration Analysis of Time Series.) Once again, the computations and the methods are not overly difficult, but the fact that they have been conveniently integrated into a general-purpose package has been an important impetus to the dissemination of the methodology in the applied literature. RATS was also an early contributor among several that significantly advanced practice by providing the user with access to an optimization program that required only a basic, again symbolic, specification of the problem. This freed the researcher from having to

	Software	Type	Website
1	B34S	Econometrics	www.uic.edu/~hhstokes/hhstokes/b34s.htm
2	BMDP	Library	www.meridian-marketing.com/BMDB/index.htm
3	E-Views	Econometrics	www.eviews.com
4	Excel	Spreadsheet/Data	(Microsoft Office 2000—commercial)
5	Gauss	Programming language	www.aptech.com
6	IMSL	Library	www.mathstat.com.au/vni/imsl/index.html
7	Lahey	Fortran development	www.lahey.com
8	LIMDEP	Econometrics	www.limdep.com
9	LISREL	Econometrics	www.ssicentral.com/lisrel
10	Lotus	Spreadsheet/Data	(Lotus Smart Suite—commercial)
11	Maple	Mathematical analysis	www.maplesoft.com
12	Mathematica	Mathematical analysis	www.wri.com
13	MatLab	Programming language	www.mathworks.com
14	NAG	Library	www.nag.com
15	RATS	Econometrics	www.estima.com
16	S Plus	Programming language	www.mathsoft.com/splus
17	SAS	Econometrics	www.sas.com
18	Shazam	Econometrics	shazam.econ.ubc.ca
19	Soritec	Econometrics	fisisoft.com/products
20	SPSS	Econometrics	www.spss.com
21	Stata	Econometrics	www.stata.com
22	Systat	Econometrics	www.spss.com/software/science/systat/
23	TSP	Econometrics	www.tspintl.com
24	Watcom	Fortran development	www.powersoft.com/products/languages/fort77.html

learn a large amount of programming. Contemporary programs, e.g., TSP, LIMDEP, and SAS, still RATS, Gauss, and so on, all offer this capability.

The earliest work in applied econometrics required the analyst either to use someone else's program or to know how to write their own Fortran code. The latter type were in a distinct minority, so most researchers used software written by others. Current practice and contemporary software make it possible for researchers to choose among a vast array of prepackaged techniques and, with the aforementioned MAXIMIZE capabilities and the matrix programming languages of Gauss, B34S, LIMDEP, RATS, and so on, to create another vast array of extensions with easily acquired programming skills. The result has been an explosion in the variety of techniques and model formulations found in the published research.

5. *Conclusion*

The development of software has been a facilitator, and a platform for the dissemination of results and continued research in econometric techniques and methodology. But there are cross-currents in contemporary software design. As of 1999, econometric software design and publication has left the cottage industry of the 1960s and 1970s far behind; it is usually (though not always) a corporate endeavor. At the same time, technique, development, and extension of econometric methodology remain the bailiwick of a scientific community which has, to some extent, preserved the distinction between the purely commercial nature of, for example, the word processor market and the romantic, if antiquated, free exchange of scientific software among individuals and small teams of researchers. However, market competition appears to be working smoothly in the right direction, and no case could reasonably be made that market forces have inhibited the development of econometric techniques or software in the way they have, e.g., in the markets for operating systems and Internet browsers. Technological advance in software and econometric technique appears to be rapid and uninhibited.

6. *Appendix: Some of the Most Commonly Used Econometric Software*

The above is a representative list of the most widely used econometric software. For additional information about the individual programs, the reader may refer to the websites listed. Some of these programs are listed as econometrics although econometrics is not the major functionality of the program neither is its

largest user base econometricians. These would include, for example, SPSS and Stata.

Bibliography

Draper N R, Smith H 1981, *Applied Regression Analysis*, 2nd edn. Wiley, New York

Granger C 1969 Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**: 424–38

Hamilton J D 1994 *Time Series Analysis*. Princeton University Press, Princeton, NJ

Heckman J J 1979 Sample selection bias as a specification error. *Econometrica* **47**: 153–61

Joreskog K, Gruvaeus G 1970 A computer program for minimizing a function of several variables. Educational Testing Services, Research Bulletin Number 70-14

Kalbfleisch J, Prentice R 1980 *The Statistical Analysis of Failure Time Data*. Wiley, New York

Krailo M, Pike M 1984 Conditional multivariate logistic analysis of stratified case-control studies. *Journal of the Royal Statistical Society*, *Series A-Statistics in Society* **84**: 95–103

Litterman R 1986 Forecasting with Bayesian vector autoregressions—five years of experience. *Journal of Business and Economics Statistics* **4**: 25–38

Longley J 1967 An appraisal of least squares programs from the point of the user. *J. American Statistical Association* **62**: 819–41

Maddala G S 1983 *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York

McCullough B D 1998 Assessing the reliability of statistical software. Federal Communications Commission, manuscript

McCullough B D 1999 Econometric software reliability: Reviews, LIMDEP, SHAZAN, and TSP, *Journal of Applied Economics* **14**(2): 191–202

McCullough B D, Renfro C 2000 Benchmarks and software standards: A case study of GARCH procedures. *Journal of Economics and Social Measure*, in the press

McCullough B D, Vinod H D 1999 The numerical reliability of econometric software. *Journal of Economic Literature* **37**: 633–65

Nerlove M, Press S 1973 Univariate and multivariate log–linear and logistic models. RAND -R1306-EDA/NIH, Santa Monica, CA

Press W, Flannery H, Teukolsky S, Vetterling W 1986 *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge

Silk J 1996 Systems estimation: A comparison of SAS, Shazam, and TSP. *Journal of Applied Economics* **11**: 434–50

Sims C 1977 Exogeneity and causal ordering in macroeconomic models, *New Methods in Business Cycle Research: Proceedings from a Conference*, Federal Reserve Bank of Minneapolis, MN, pp. 23–43

Stokes H H 1997 *Specifying and Diagnostically Testing Econometric Models*, 2nd edn. Quorum Books, Westport, CT

Zellner A 1962 An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of the American Statistical Association* **57**: 500–9

Zellner A 1971 *Introduction to Bayesian Inference in Econometrics*. Wiley, New York

Zellner A, Theil H 1962 Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. *Econometrica* **30**: 63–8

W. H. Greene

Econometrics, History of

Econometrics, the statistical metric for economics, can be regarded as one of the main innovations which turned twentieth century economics into an engineering, or tool-based science, in which each application of economic theory requires special shaping to circumstances, whether for scientific purposes or in the policy domain (see Morgan 2001). The development and use of statistical tools for economics emerged in the early part of the twentieth century and, by mid-century, problems had been defined, solutions approached and usable concepts developed, so that one could legitimately refer to a distinct body of knowledge embracing both theory and practice. After 1950 econometrics became a mature field and the dominant method of applied economics.

The particularities of the history of econometrics have often been linked to the nature of its tasks and aims (see De Marchi and Gilbert 1989, Heckman 2000, Morgan 1990, Qin 1993). Economists have understood their fundamental methodological problem to be: how to establish knowledge about economic phenomena and economic laws given that most economic activity cannot be subjected to controlled experiments. Econometricians have assumed that this means that descriptions, regularities, and relations of the economy must be extracted from 'passive' observations: statistical data that result from the natural course of economic activities, with the use of statistical theories and tools. Econometrics therefore suffers from all the usual problems of an inductive science. When measurement or testing of theories is at stake, severe correspondence problems arise because of the huge gaps between the assumptions of those economic theories and the conditions of data collection.

The history of the field also has its external dynamics (Desrosières 1998, Epstein 1987). To begin with, technological changes play an important role, evidenced in the development of the field of mathematical statistics and machines of computation. The provision of data, the availability of trained econometricians, and fashions in economic theory are all relevant to the narrative, as are the stimuli from twentieth-century economic events and associated demands for econometric expertise. Econometrics in the later twentieth century became a mainstay of economic policy analysis and tool for economic intervention, yet at the same time it grew into one of the most technically demanding and methodologically sophisticated areas of economics.

1. The Development Era

In the latter nineteenth century, a few economists began to use graphs and tables of statistical data to provide empirical support for their arguments, or

to describe behavioral characteristics of economic phenomena (Klein 1997). They were quick to take up certain statistical techniques developed for biometrics, namely correlation and regression. But these 'measures of association' were not entirely adequate for economists' questions (Aldrich 1995, Morgan 1997). Where biometricians had, for example, a sample of heights of fathers and their sons, whose relation could be assessed with correlation or regression coefficients under reasonable assumptions about the individuals in the sample, economists typically used time-series observations that provided only one datum for each time date. Not only were such 'samples' considered peculiar from the viewpoint of the correlation theory of that time—which assumed that the observations within the sample had been independently drawn—but economists believed that such time-series observations reflected a number of concurrent behavioral characteristics or causal relationships varying with the time element and unit chosen.

This problem of extracting causal relationships between variables from data which contained different time patterns dominated early econometric work but the questions attacked were driven by economic and policy concerns. Measurements, for example, of market demand and supply relations for agricultural goods were undertaken at the American Bureau of Agricultural Economics during the 1920s when farmers faced falling prices in international markets and followed earlier European breakthroughs on these questions. Similarly, the problems posed by 'business cycles' were widely recognized by commerce, politicians, and economists, for the crash of 1920–1921 was only outdone by the even more spectacular and longer-lived depression of 1929–1933. Business cycle research institutes, which had begun to emerge even before the end of World War I, sprang up everywhere in the 1920s from Massachusetts to Moscow, but their truly international quest for description and explanation of the business cycle showed that the task of establishing measures of—let alone explanations for—that phenomenon, its characteristics, and its time patterns was incredibly difficult.

A concerted movement to create a more 'scientific' economics, based on both statistical and mathematical tools in economics, resulted in the foundation of The Econometric Society (in 1932), and its house journal *Econometrica* (in 1933). The most critical innovation of that decade, often overlooked, was that of the 'econometric model' formulated as an intermediary device to bridge the gap between economic theory and economic data. Such models represented elements of the economic theory in mathematical form, but were also attuned to the peculiarities of the time patterns of data evidence. With the use of such models, various kinds of knotty correspondence problems at the intersection between economic theory and data began to be understood and solutions proposed. For example, the problem of unraveling several statistical relations holding within a variable set (the multicollinearity of statistically correlated variables) was understood as different from the problem of separating out different economic relations which involved some of the same overlapping variable set (the identifiability of several behavioral economic relations). At the same time, problems of measurement errors ('errors in variables'), or of omitted variables from the model ('errors in equations') were attacked with the development of statistical tools and methods. But these 1930s answers were proposed as one-off solutions to specific applied problems, not general solutions for generally recognized problems.

However, econometrics was not without its detractors and the move to codify and formalize these somewhat disparate concepts, problems, and emergent solutions seems to have been triggered by the Keynes-Tinbergen debate in the late 1930s. Jan Tinbergen, a young Dutch economist, was the most creative of econometricians of the 1930s in the development of modeling, in understanding conceptual problems, and in suggesting solutions. He was also responsible for building and using the first macroeconometric model to analyze policy options. In this work, he developed ideas by Ragnar Frisch, a Norwegian econometrician (which later made them joint winners of the first Nobel Prize in economics in 1969) who in turn had drawn on statistical work by English statistician Yule and Russian econometrician Slutsky.

Tinbergen's macroeconometric modeling work was heavily attacked by Keynes and this attack prompted the development in the 1940s of rigorous foundations for econometrics. Despite their difference of opinion, both Tinbergen and Keynes can be seen, in retrospect, as reshaping economic theory and policy in the aftermath of the Great Depression so that, in the 1950s, Lawrence Klein (Nobel prize 1980), could develop the post-war generation of macroeconometric models on Keynesian theories with more secure statistical foundations.

The key development in this foundation work was the 'probability approach', fashioned in the early 1940s by Trygve Haavelmo (Nobel Laureate in 1989 for this work), a Norwegian econometrician, who proposed to use probability theory to establish a bridge between economic theory and passive economic observations. His bridge was designed to carry the burden in both directions: to provide guidance on how to specify econometric models as statistical hypotheses so that not only could economic theory be logically related to economic data but that there would also be a path for inference back from the data to the theory.

2. *Establishing the Mature Discipline*

Haavelmo's probability blueprint for econometrics had indicated directions for both theoretical and applied work and has been interpreted as creating a

'revolution' in econometrics. Parts of Haavelmo's program were immediately elaborated at the Cowles Commission (later Foundation), an American hothouse of econometric research in the mid 1940s (see Hildreth 1986). They assumed that the driving relations of the economy were best represented as a system of simultaneous equations that operated at a level hidden beneath the time-series observations. Their research was primarily concerned with developing appropriate identification conditions and estimation (measurement) techniques for such econometric models. In the years after 1950 an established orthodoxy developed based on Cowles' work and enshrined in textbooks in the 1960s. But, despite an active program of research at the theory level, providing technical extensions to Cowles' groundwork, cracks in that orthodoxy began to appear. There were alternatives to the Cowles program.

Applied econometric research revealed serious practical and methodological limitations in the Cowles approach, particularly in regard to model specification, model choice, and associated testing procedures. Their assumption that econometric work is anchored on *a priori* theories meant that it was rarely possible to specify, and identify uniquely, structural models for those theories. This made theories virtually unfalsifiable and the fitting of models to data an arbitrary affair, as Henri Theil and T. C. Liu pointed out in different contexts in the 1950s. This was made transparent in the specification and sensitivity analysis using Bayesian methods developed by Ed Leamer in the 1970s (Qin 1996). The Cowles assumption of simultaneity and associated lack of attention to the characteristics of time-series data, were both open to challenges. The Swedish econometrician Herman Wold, who had trained in the Russian probability school and time-series analysis in the 1930s, attacked the assumption of simultaneity and developed alternative dynamic interdependent equation models in the 1950s (Morgan 1991). These developments found a generalized form in the 1970s in the American econometrician Christopher Sims' time-series versions of Wold's models called Vector Auto-Regressions (VARs) (Qin and Gilbert 2001). Whereas Wold had shown how the Cowles structural models were special cases of his causal chain models, VARs turn out to be the least arbitrary representation of the information in time-series data.

The time-series problems of economic data, and possible solutions for Cowles measurement techniques, were analyzed at the Department of Applied Economics at Cambridge in the 1940s under Richard Stone (Nobel Laureate in 1984). However, it was the 1970s experience of stagflation in many economies (high unemployment and inflation) that spurred a more radical refashioning of economic theory and econometric approaches. The dynamic elements of economic modeling, taken seriously in the 1930s, once again became the focal issue of econometric measurement. Descriptive statistical time-series models (i.e., models not relying on economic theory) underwent a certain renaissance in the wake of disappointing prediction records from many large-scale macroeconometric models during the 1970s upheavals. Meanwhile, economic theories were being re-formulated to provide a more elaborate account of uncertainty at the microeconomic level in the form of assumptions about 'rational expectations' and their relation to the macro-level. This formed an important stimulus for Sims' work, but while VARs can be understood as the most general form of dynamic models, they were not necessarily the most interpretable in terms of economic phenomena. An alternative dynamics, analogous in form to servo-mechanistic control engineering, and associated with David Hendry and the London School of Economics was the 'error correction model' which could be more readily understood in terms of individuals' reactions to changing economic events. These developments effectively reversed the Cowles approach for they start from a summary of the data and move towards the simplest possible interpretable model that is coherent with the data.

The emergence of an especial microeconometrics occurred concurrently with these changes in macroeconometrics. In the early 1950s, Stone had struggled with the immense task of providing measurement structures appropriate for the separate components of demand in a total consumer expenditure system: a system simultaneous at the micro-level. By the 1970s, various statistical models appropriate for explaining behavior at the individual level had been adapted from 1930s biometrics into econometrics by Dan McFadden and James Heckman (joint Nobel Prize in 2000). Their treatment of survey data enabled measurement models to be developed for the discrete choices made by individuals. Such discrete choices dominate economic decision making in consumption decisions, work decisions, family formation decisions, etc. These measurement models were quickly taken up for they matched the individual utility-maximizing framework of neoclassical economic theorizing, the main orthodoxy of post 1950s western economics. The gradual appearance of large-scale 'panel data sets' (a set of individuals measured repeatedly over time) encouraged both theoretical and applied work at this microeconomics level and created a wider territory for the expansion of econometrics. Time-series work in microeconometrics was not completely abandoned however, for in the 1990s a more data-instigated style of analysis, reminiscent of the 1930s business cycle work, was developed for the analysis of financial data.

Post 1950s data-based research has not been bound by any straightjacket methodology but spawned several hybrid approaches. Guy Orcutt, believing that macro-time series data were inadequate for inference about the economy, developed the method of microsimulation, an intermediary style of econometric

research involving modeling, analysis, and simulation. This approach, difficult to achieve when he began his work in the later 1950s, has become practicable with modern computing power, and is now almost a necessity in certain areas of policy analysis. In the 1980s, attempts to model business cycles combined simulation, computable general equilibrium analysis, and the VAR model another hybrid to produce replacing 'estimation' (using statistical data to measure relations) with 'calibration': the matching of certain statistical properties predicted by the models with those found in time-series data.

3. *The Place of Econometrics in Economics*

During the early years of The Econometric Society, 'econometrics' had been defined as the union of economics, mathematics and statistics. And, in developing model intermediaries based on mathematical economic theory in that period, econometricians had developed a distinctive path compared to those found in psychometrics and sociometrics in the same period. After 1950, as economic theorizing became generally expressed in mathematical terms, the term econometrics stabilized on its current meaning: the use of statistical reasoning and methods as means to establish data-based descriptions of economic phenomena and empirically based counterparts for, and tests of, economic theories.

The spread of econometrics as an established part of economics in the post 1950s period may be understood as due not only to the strengthening of its foundations in statistical theory but also to the continuing expansion of data ('official' statistics), its establishment in the core undergraduate teaching program, and the development of cheap desk-top computing. Together these meant that econometric work became a standard tool of policy work in governments and international agencies as well as becoming endemic, in various different forms, in the sub-fields of scientific economic research. Econometric theory has developed into a formidable body of specialist statistical theory and the increasing gap between the difficult technical and theoretical questions and the apparent ease of applications might indicate a field where applications came loose from theoretical work. This has been militated by another tool, the development of specific software packages for econometrics, in which theoretical and technical developments can be quickly translated into modeling, measurement, and testing regimes at the level of applications.

This widening use of econometrics did not denote a settled field. The community in the latter part of the twentieth century found itself immersed in a number of vibrant and important arguments about how to do econometrics, based upon applied econometric experience as much as on considerations of statistical theory and philosophy of science. Should models be theoretically based or descriptively accurate? What are the relevant criteria, economic and statistical, which can be used to test the adequacy of these different models? Should the modeling process start with simple models that grow more complex, or start general and aim to become simple? Should the probability framework be classical or Bayesian? What implications does this difference have for the types of measurements made and the possibilities of inference? How far can applied econometrics be formalized and codified apart from tacit knowledge? The terms of the continued debates around such themes point to a mature field, one intimately connected to both statistical and economic theory, but in which the foundational arguments raised in early work continue to reverberate to the most recent time.

See also: Disciplines, History of, in the Social Sciences; Econometric Software; Economic Growth: Measurement; Economic Panel Data; Economics, History of; Economics: Overview; Keynes, John Maynard (1883–1946); Probability: Interpretations; Quantification in the History of the Social Sciences; Simultaneous Equation Estimation: Overview; Statistical Identification and Estimability; Statistical Methods, History of: Post-1900; Statistical Methods, History of: Pre-1900; Statistical Systems: Economic; Time Series: Economic Forecasting

Bibliography

Aldrich J 1995 Correlations genuine and spurious in Pearson and Yule. *Statistical Science* **10**: 364–76

De Marchi N, Gilbert C 1989 *History and Methodology of Econometrics*. Oxford University Press, Oxford, UK

Desrosières A 1998 *The Politics of Large Numbers*. Harvard University Press, Cambridge, MA

Epstein R J 1987 *A History of Econometrics*. North-Holland, Amsterdam

Heckman J J 2000 Causal parameters and policy analysis in economics: A twentieth century retrospective. *Quarterly Journal of Economics* **115**: 45–97

Hildreth C 1986 *The Cowles Commission in Chicago, 1939–1955*. Springer-Verlag, Berlin

Klein J L 1997 *Statistical Visions in Time*. Cambridge University Press, New York

Morgan M S 1990 *The History of Econometric Ideas*. Cambridge University Press, Cambridge, UK

Morgan M S 1991 The stamping out of process analysis in econometrics. In: De Marchi N De, Blaug M (eds.) *Appraising Economic Theories*. Edward Elgar, Aldershot, Hants, UK, pp. 237–63, 270–2

Morgan M S 1997 Searching for causal relations in economic statistics: Reflections from history. In: McKim V, Turner S (eds.) *Causality in Crisis: Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. University of Notre Dame Press, Notre Dame, IN, pp. 47–80

Morgan M S 2001 Economics. In: Porter R (eds.) *The Cambridge History of Science, Vol. 7, Modern Social and Behavioral Sciences*. Cambridge University Press, Cambridge, UK
Qin D 1993 *The Formation of Econometrics: A Historical Perspective*. Clarendon Press, Oxford, UK
Qin D 1996 Bayesian econometrics: The first twenty years. *Econometric Theory* **12**: 5500–51
Qin D, Gilbert C L 2001 The error terms in the history of time-series econometrics. *Econometric Theory* **17**: 424–50

M. S. Morgan and D. Qin

Economic Anthropology

Economic anthropology studies how human societies provide the material goods and services that make life possible. In the course of material provisioning and during the realization of final consumption, people relate to each other in ways that convey power and meaning.

1. A Definition

The degree to which something is 'necessary' for life has long been debated and differences between one society and another have environmental, historical, and cultural reasons; but some wants must be inescapably satisfied, otherwise death ensues. Therefore, there is a physical limit to relativism regarding material means of livelihood. On the other hand, nonmaterial goods such as the goodwill of deceased ancestors might be conceived as essential for the reproduction of a society. Most nonmaterial needs, however, have some material expression, such as food sacrifices during ancestor worship or wealth exchange during mortuary ceremonies. The domain of economic anthropology covers the recurring interaction of individuals, within and between social groups and with the wider environment, with the object of providing material goods and services necessary for social reproduction.

Traditionally, economic processes have been divided into production, distribution and circulation, and consumption. These analytical categories respond to observable social interaction in all societies, although the categories themselves are a product of scholarly Western tradition. People, however, engage in social relations that can be described as 'economic' and which can be analyzed as participating simultaneously in the production, distribution, and consumption categories.

Economic anthropology originally focused on the *Economic Life of Primitive Peoples* (Herskovits 1960) where many of the elements present in the Western economy (such as money, a market system) were absent. Direct observation of noncapitalist societies through ethnographic fieldwork produced impressive and contextually rich information on economic activities worldwide. The way in which anthropologists reacted to the confrontation of this diversity and how they coped with it in theoretical terms, generated most debates within economic anthropology.

2. The Intellectual Debates

The major debates in economic anthropology have centered around (a) the universal applicability of Western generated categories of analysis, (b) the question of value, (c) the question of history and connectedness between polities, and (d) the weight of culture (meaning) in economic processes.

2.1 Western Categories of Analysis

The main debate took place between those anthropologists who considered that the postulates of marginalist economics, that is, the rational (optimizing) allocation of (scarce) resources between alternative uses, were of universal application (Firth 1970); and those anthropologists who, following Polanyi (1957), thought that the theory of rational action in regard to economic decision making was only valid in the context of Western market economy, and that a real definition of the economy should be meaningful in any society whatever its process of allocation. The first group was eventually defined as 'formalist' while the second was called 'substantivist,' following Polanyi's distinction. The most useful concept that emerged during the debate, although it was not itself subject to dispute in regard to 'primitive' economies, was that of the 'embeddedness' of economic activities in other social processes. A second wave of the debate in the 1970s included anthropologists working with formal decision-making models and Marxian anthropologists working with the concept of 'mode of production' and with issues of 'transition' to capitalist economies and 'articulation' of different modes of production (Godelier 1977).

2.2 The Question of Value

The question of value is a function of exchange, of the need to reach some equivalency through comparison. Some aspects of value have constituted constant sources of debate within anthropology. First, Marx's distinction between use value and exchange value, and how it could be applied to noncapitalist societies or to peasant societies where factors of production were not fully commodified. Second, the classical labor theory of value, and how to use it in societies where the labor market, when it existed, was only marginally active. Third, the marginal utility theory of value and how to make sense of it when every good could have multiple uses and be valued according to different measures.

Finally, the cultural theory of value where local meaning attached to objects, people and situations was the measure of value, posed the problem of cross-cultural comparability in a connected world.

2.3 History and Connectedness

Another issue that became increasingly central was the need to think historically about the transformation of social relations and the need to study the interconnection between different societies through time. These problems were especially salient to those anthropologists influenced by dependency and world system theories, by European Marxist anthropology, and by what came to be known as the political economy perspective in anthropology (Roseberry 1988). The aim was to understand local processes as both being formed by and acting upon the larger processes of historical transformation (Wolf 1982).

2.4 The Weight of Culture

Cultural aspects had been taken into account in economic anthropology mainly as a context where material activities occurred. More recently, anthropologists have been arguing about the centrality of meaning for the understanding of economic processes. Bourdieu (1980) has developed the concept of 'habitus' where past material social relations are transformed into a set of dispositions, a scheme of perceptions that generate practice. Gudeman's (1986) proposition, on the other hand, stems from his preoccupation with the methodological problems that Western models of knowledge pose to the understanding of other cultures. He proposes to study how each society culturally models the processes by which it secures a livelihood, and shows that some societies have multiple articulated models of economic activity. Finally other anthropologists are using the term 'moral economies' to refer to individual and collective moralities and cultural values that pervade economic social relations.

3. Concepts and Theory

The classical categories of production, distribution and consumption, will be followed in order to present briefly the main concepts and theoretical developments that have unfolded within economic anthropology.

3.1 Production

Production generally is understood to be the human transformation of matter, through work, into some useful, consumable good. Sometimes, as in foraging and hunting, human work uses very simple technology to select and get whatever food is available in the environment. Human–nature interaction has been studied from an ecological perspective stressing energy exchanges between different species, but emphasizing 'culture' as a basic factor of human adaptation to environmental constraints. The environment, however, is not an ahistorical given where human populations dwell. Rather, past relations between individuals, groups, communities and larger polities are inscribed in the environment. Political ecology tries to include this history in its perspective.

In most societies, the process of production requires the use of complex technology and the design of cooperative labor processes. Knowledge both of the skills necessary to use a particular technology and of the coordination of the entire process is a crucial asset of control that generally is distributed unevenly among those participating in production. The way in which a society organizes access to land, labor, energy, technology, and information expresses economic and political social relations. Moreover, in all societies these relations are embedded in other institutional domains such as kinship, religion, politics, and 'pure,' disembedded, economic relations have seldom been observed. Peasant studies show, for example, how kinship and marriage, experienced through discourses of respect and love, pattern inheritance systems but also social relations of production within the household farming unit.

Cultural meanings attached to concrete tasks and/or collective or personal identities such as gender, age, caste, or ethnic group also contribute to shape work processes. The concept of 'division of labor' was designed originally to explain the 'integration of the social body' through the necessary dependence between specialized domains of work (Durkheim 1933). An example of this meaning can be found when speaking of the 'sexual division of labor' in a society where male members are ascribed certain work responsibilities while female members are ascribed others. In this sense the concept is broad and general. More frequently, 'division of labor' is used to address particular production processes and the assigning of individuals or groups to positions within the process. This has given anthropologists the opportunity to analyze minute interactions taking place between those participating in a concrete labor process.

3.2 Distribution and Circulation

The concept of distribution refers to the allocation of goods between different individuals or groups, while the concept of circulation refers to the movement of goods. These processes mediate between the 'production' and 'consumption' moments of an economy. They also reproduce differentiated categories of people in regard to the access of resources in general which makes them central aspects of social reproduction.

Economic anthropology has developed a typology of forms of distribution that was proposed originally by Polanyi (1957). Distribution was for Polanyi the element that provided continuity and structure to economic processes. Through a comparative method, he concluded that three main forms of distribution were used to integrate the economy: reciprocity, redistribution, and exchange. This typology expressed institutional arrangements not so much the form of particular transactions.

The concept of reciprocity had an early start in Mauss' 1923–4 (1968) essay on the 'gift.' The essay was based mainly on available ethnographic descriptions of the Potlatch and in Malinowski's 1922 (1961) description of the Kula ceremonial exchange. For Mauss, the 'gift' was an entirely different form of transaction from market exchange. The sustained relationship between persons and things was essential to the 'gift' value. Conversely, market exchange was based on the total disjunction between autonomous objects and individual agents. Mauss describes the 'gift' as a movement related to three obligations of a social and moral character: to give, to receive, and to return. Further developed by Sahlins (1965) who related reciprocal transactions to the social distance between the persons involved, reciprocity has become a useful concept in economic anthropology. Weiner (1992) has described how, by entering or remaining out of circulation, objects could create and regenerate social relations. Her work opens the concept beyond the give-return movement originally associated with reciprocity.

The concept of redistribution as an institutionalized process refers to centralized polities that concentrate goods through tribute or taxation systems and reassign them later between groups, individuals, and specific domains. Ethnographic examples range from 'Big-man' systems to strongly centralized state polities.

In Polanyi's typology, exchange, as an institutionalized process, relates to societies that integrate the economy through the market system. However, multiple forms of transaction can be called 'exchange,' be it the routes of trade partners crossing Australia, African market places, or the elaborate systems of ceremonial exchange in Melanesia. Exchange raises two fundamental problems of transactions: first, comparison between the items exchanged and, second, nonsimultaneity of the agents' needs. Comparison is the central question of value. Things exchanged always are valued, but how this valuation proceeds is very different from one society to the next. Generally, for a transaction to take place there must be some measure of value that enables the agents to reach an equivalence acceptable to all. When some sort of standard item is used as a measure of value we may speak of money, although some other functions that generally are associated with money, such as serving as a medium of exchange, may be lacking. Often, the process of reaching equivalences of value takes place during exchange such as in bargaining practices, while in other cases a central authority sets up a fixed rate of exchange. Yet in other cases, as the model of the market system pretends, value is reached automatically through the free circulation of all commodities subject to the constraints of supply and demand. Classical economists tried to find a universal measure of value in work: the energy spent in producing the commodities exchanged was seen as their only common element.

Most societies have various spheres of circulation where different measures of value apply. In Bohannan's (1959) example of the Tiv's (Nigeria) multicentered economy, he describes how different goods circulate in distinct spheres, each of which is marked with different moral values relating to subsistence, prestige, and alliances, and conversion from one sphere to another, while possible, always is sanctioned morally. The idea that various measures of value might simultaneously be at work in a society has proved very fruitful. Increasingly, anthropologists are paying attention to the circulation of goods among different individuals, social groups, or polities along chains of transactions. This perspective integrates the variation of meaning attached to goods or processes together with its material causes and consequences (Appadurai 1986).

3.3 Consumption

Consumption can be defined simply as the use of a good or service. Use sometimes implies the destruction of the good and precludes further use; otherwise, a good can be used in multiple ways, successively or simultaneously by different persons. The allocation of food to different ends among household members traditionally has been the focus of attention. Gender and age biases that seriously affect the intake of nutrients of certain members of a society have been studied. Issues of power, of cultural and long-term institutional constraints have come to the forefront as ethnographic research has explored the actual processes of consumption (Mintz 1986).

Economic anthropology focuses on the social relations that emerge through consumption processes in the intimate space of the household or in other, more open, public spaces. Some anthropologists underline the meaningful aspect of consumption acts conceived as signs in an information system expressing social relations (Douglas and Isherwood 1980). Others emphasize the role of consumption in the present-day, globalized world where goods, people, and information circulate widely across boundaries, and where cultural identity seems to be increasingly the center of personal meaningfulness, collective solidarity and empowerment (Miller 1995).

Bourdieu (1979) has shown how relations of differentiation and domination are constructed through

complex processes of consumption where the provisioning and access to material goods is entangled with the production and reproduction of personal and collective identities. His emphasis is on the reproduction of an entire social system. Another perspective stresses the systems of provision and underlines the connectedness between production, distribution, and consumption. From the point of production to final consumption, specific social relations contribute both to the material constitution and availability of particular goods (e.g., regarding quality or distribution networks) and to the social construction of particular meanings attached to those goods (Fine and Leopold 1993).

Most societies have different possible modes of provisioning equivalent goods and each mode conveys a particular field of meanings, power relations, and material well-being. Frequent modes of provision are those making goods available through the market (commodities), through the State (social welfare), through the household (self-provisioning), or through the community (solidary networks). People may change their modes of provision following economic, political, or other social pressures, but goods also, during circulation, may participate in different modes of provision. Here, as with multiple spheres of circulation, transfers between modes of provision have both material consequences and meaning.

The focus on consumption in anthropology has renewed the interest in objects (material culture) and on how they incorporate, circulate, create, and transform social relations.

4. *New Trends*

Suggestive contributions have been made by anthropologists studying nature. They point to the methodological need to go beyond the duality nature/society in the study of human/environment interaction. Ethnographic research also underscores the blurring of the natural/artificial boundaries, as biotechnologies become capable of producing life while commoditization reaches nature in unprecedented ways (marketing organs, cells, genetic material). Groundbreaking concepts are bound to emerge from these new realities.

Focus on the informal sector of the economy has also contributed a major breakthrough by expanding the concept of 'work.' Research on informal relations (economic or otherwise) is bound to prove fruitful as the realities of flexible production, the weakening of the nation-state model, the informational revolution, and massive population movements point to the increasing value of personal networks in the structuring of contemporary societies.

Presently, economic anthropology is moving beyond the boundaries of a conceptual economic domain. While it remains useful to retain the material emphasis of economic processes, in practice, material relations should be studied together with their cultural expressions. On the other hand, ideas of worldwide connectedness—in systems of provisioning, relations of production, cultures of consumption, labor migration, etc.—need to be taken into account. The practice of economic anthropology seems bound toward the study of social reproduction as a whole (Narotzky 1997).

See also: Boserup, Ester (1910–99); Consumption, Sociology of; Development, Economics of; Development: Social-anthropological Aspects; Development: Socioeconomic Aspects; East Asian Studies: Gender; Economic Development and Women; Economic Geography; Economic Sociology; Exchange in Anthropology; Family and Consumer Sciences: United States; Family and Kinship, History of; Family as Institution; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Family Theory: Feminist–Economist Critique; Feminist Economics; Fertility: Institutional and Political Approaches; Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Southeast Asia; Gender and Technology; Gender, Economics of; Gender History; Household Production; Informal Sector; Labor Movements and Gender; Latin American Studies: Gender; Markets: Anthropological Aspects; Mode of Production; Money: Anthropological Aspects; Moral Economy and Economy of Affection; Near Middle East/North African Studies: Gender; Political Economy in Anthropology; Southeast Asian Studies: Gender

Bibliography

Appadurai A (ed.) 1986 *The Social Life of Things*. Cambridge University Press, Cambridge, UK

Bohannan P 1959 The impact of money on an African subsistence economy. *Journal of Economic History* **19**: 491–503

Bourdieu P 1979 *La Distinction*. Editions de Minuit, Paris

Bourdieu P 1980 *Le sens pratique*. Editions de Minuit, Paris

Douglas M, Isherwood B 1980 *The World of Goods*. Penguin, London

Durkheim E 1933 *On the Division of Labor in Society*. MacMillan, New York

Fine B, Leopold E 1993 *The World of Consumption*. Routledge, London

Firth R (ed.) 1970 *Themes in Economic Anthropology*. Tavistock Publications, London

Godelier M 1977 *Horizon, trajets marxistes en anthropologie*. F. Maspero, Paris

Gudeman S 1986 *Economics as Culture*. Routledge and Kegan Paul, London

Herskovits M 1960 *Economic Anthropology*. Knopf, New York
Malinowski B 1961 [1922] *Argonauts of the Western Pacific*. Dutton & Co, New York
Mauss M 1968 [1923–24] Essai sur le don. Forme et raison de l'échange dans les sociétés archaïques. In: Mauss M (ed.) *Sociologie et anthropologie*. Presses Universitaires de France, Paris
Miller D 1995 Consumption as the vanguard of history. A polemic by way of an introduction. In: Miller D (ed.) *Acknowledging Consumption*. Routledge, London
Mintz S W 1986 *Sweetness and Power*. Penguin, New York
Narotzky S 1997 *New Directions in Economic Anthropology*. Pluto Press, London
Polanyi K 1957 The economy as instituted process. In: Polanyi K, Arensberg C M, Pearson H W (eds.) *Trade and Market in the Early Empires*. Free Press, Glencoe, IL
Roseberry W 1988 Political economy. *Annual Review of Anthropology* **17**: 161–85
Sahlins M 1965 On the sociology of primitive exchange. In: Banton M (ed.) *The Relevance of Models for Social Anthropology*, ASA Monographs 1. Tavistock Publications, London
Weiner A B 1992 *Inalienable Possessions*. University of California Press, Berkeley, CA
Wolf E R 1982 *Europe and the People without History*. University of California Press, Berkeley, CA

S. Narotzky

Economic Development and Women

The consequences of economic development on women during the last half of the twentieth century have been tumultuous as traditional cultural protections and limitations prescribed for women are constantly being challenged and reconfigured. Around the world women have organized to influence development policies and to moderate the adverse impacts of rapid socio-economic change, an activity collectively referred to as women in international development (WID). This combination of activists, practitioners, and scholars has lobbied international agencies, national governments, and nongovernmental organizations so successfully that the consideration of women has become an integral part of development planning. The cumulative result of this political activity, given focus and visibility through the United Nations (UN) World Conferences for Women, has been the creation of a global women's movement, perhaps the most significant social movement of our times.

The first section documents the incorporation of women's issues into international development policy at the UN and at the national levels as a result of pressure by women's organizations, and reviews the four world conferences for women. The second section follows the evolution of programming for poor women in developing countries and underscores the significance of organizing women at the village or community levels. Significant debates about the goals and process of such programs are summarized. The final section shows how topical conferences organized by the UN provided a venue for inserting critical women's issues into official UN documents concerning human rights, population, and habitat.

1. *Affecting International Development Policy*

The UN was established in 1945 with its 50 members drawn largely from industrialized countries. By the mid-1960s, a majority of the UN members represented developing countries of Asia and Africa. Their concern with economic development propelled this new dimension into global politics. The first UN Development Decade 1960–70 emphasized infrastructure and industrial projects and was modeled on the successes of the Marshall Plan for redevelopment of a devastated Europe. Former colonial countries, which had often resisted granting independence, sought to maintain their trading advantage with former colonies by supplying funds and personnel to assist in modernization of these countries. Other industrialized countries as well as the former USSR and the Eastern Bloc offered bilateral assistance to selected countries. Respective UN agencies established programs in health, agriculture, employment, and education; the World Bank provided loans at low interest rates. Since some were much better positioned than others to benefit from these programs, the immediate result of these vast infusions of funds was increased inequality of income within the countries.

Concerned with this trend, the Economic and Social Council (ECOSOC) reacted to the UN Report on the World Situation by passing a resolution in 1963 tying social and economic issues together. Social issues clearly applied to women, since women delegates at the UN were often leaders of women's organizations in their own newly independent countries that were demanding a voice in planning at home, they supported a 1963 UN General Assembly resolution calling for women to be appointed to those bodies preparing national development plans (Jayawardene 1986, Snyder and Tadesse 1995) (see *Colonization and Colonialism, History of*; *National Liberation Movements*).

The debates about development priorities were reflected in the Second Development Decade 1970–80, with its emphasis on basic human needs. The refocus of development programming from macrodevelopment toward outreach to people was critical as it allowed advocates for women in development to document the distinct ways that economic development affected women as opposed to men. Poor women everywhere perform economic activities, from farming to microenterprise to household production; because income data excluded subsistence and exchange activities,

women's roles were invisible to planners who subsequently ignored or undercut women's livelihoods. Impetus for including women in development planning emerged, as do so many ideas whose time has come, from three distinct directions: from women in the UN, from women in Washington, DC, and from national women's organizations in developing countries.

1.1 Women Influencing the UN

The nineteenth-century women's movement focused on women's legal rights, particularly the right to vote (see *Women's Suffrage*). These issues, including education, continued to dominate the UN Commission on the Status of Women after it was set up in 1947 as the result of lobbying, especially by women from Latin America who had been active in a similar body of the Organization of American States. Despite a limited budget and staff, the Commission, working with international women's organizations participating at the UN, was successful in steering a series of conventions through the UN General Assembly on women's rights relating to citizenship, marriage, and inheritance. Planning for the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) began in 1965; before it was recognized as a treaty in 1980, additional articles were added to reflect documents from the women's conferences (see below). UN conventions must be ratified by member states before they take on the force of law in a particular country, but ratification does not always mean implementation. Nonetheless, the existence of CEDAW and other UN documents exerts strong pressure on governments to expand women's rights. Women's groups often present 'shadow' reports to the UN Committee that monitors implementation when they believe their governments are not complying with the Convention (Fraser 1999, Winslow 1995).

At the same time, pressure by developing country delegates was building to consider women's economic roles, in addition to legal and social rights that were still the primary focus of the Commission. A 1969 report of a meeting of experts reiterated the evidence of increased inequality between rich and poor under the capitalist approach to development, and since development ignored women's economic roles, the report stressed the importance of 'integrating women in development', a phrase widely adopted in subsequent documents. A second experts' meeting in 1972 was convened to explore the relationships between women and development and was co-sponsored by the Commission. Rapporteur of the meeting, and member of the UN Development Planning Committee, was Ester Boserup, whose landmark book *Woman's Role in Economic Development* had been published in 1970 (Boserup 1970). Of her role in that meeting, Boserup writes that the Secretariat of the Commission 'saw it as a means to get members of the Commission to change their focus from the generally unpopular subject of abstract women's rights to the popular one of economic development' (Boserup 1999, p. 49).

Boserup had found that even in countries with a large participation of women in agriculture, women were nonetheless classified as housewives or 'non-active' persons. To challenge these data categories, Boserup assembled case studies from Asian and African countries that provided hours worked by women and men in both paid and unpaid work, including agricultural work, household enterprises, and the fetching of fuel and water (Boserup 1970). Women's work varies by agricultural system: in Africa south of the Sahara women are the primary farmers while men assist in clearing the fields and protecting their homes. The plow changed agricultural techniques and required men as well as women to work in the fields but gave men control over new technologies, to women's disadvantage. Boserup believed that population increase drove the technological change which in turn altered women's roles. Male migration to cities is easier where women are farmers as in Africa; female migration predominated in Latin America where women had little agricultural role (Boserup 1999). Her work sparked a tremendous interest in producing further case studies to corroborate or challenge this evolutionary presentation.

1.2 Adding a Concern for Women to National Policy

The UN influences the world view on issues through debate, resolutions, and world conference, but it does not set national policy. The shift of goals in the Second Development Decade documents caused the US Congress to amend its Foreign Assistance Act of 1960 to include an emphasis on basic human needs. As a result of lobbying by the women's caucus (Women in Development or WID) of the Society for International Development (SID), the amended law of 1973 incorporated language of the UN 1969 resolution to integrate Women in Development programming (Tinker 1990). An office of Women in Development was established within the US Agency for International Development to carry out the Congressional mandate.

The United States was the major donor country at the time; its policies were widely observed in the West. Because the WID amendment required existing programs to consider women's roles when implementing them, it had an immediate impact on the development community and precipitated resolutions in UN implementing agencies and changes in assistance policies in many countries. These rapid changes occurred as a result of the efforts of women within the institutions concerned as well as activists in national and international women's organizations. (Women's activities in UN agencies and in NGOs are reported in Meyer and Prugl 1999.)

1.3 UN World Conferences for Women

The growing strength and visibility of the reinvigorated women's movement supported the CSW's request for a conference to mark International Women's Year in 1975. The UN concurred, proclaiming the themes for the year and conference as 'equality, development, and peace' reflecting the primary preoccupations of the three UN Blocs of East, West, and Third World. The official World Conference for Women was held in Mexico City in June 1975. The upsurge of support for further exploration of women's changing status led the UN to declare a Decade for Women 1976–85, and to hold both a Mid-decade Conference in Copenhagen in 1980 and one discussing 'Forward Looking Strategies' in Nairobi in 1985. In 1995 a Fourth UN World Conference for Women was convened in Beijing.

At each major UN world conference since 1972, nongovernmental organizations have organized a parallel event, an NGO Forum, that is open to the public. These meetings combine professional panels with the gaiety of a fair; they provide a way to disseminate new ideas and research to UN delegates as well as among the growing NGO community. Participants at the parallel NGO conferences far exceeded official delegates and provided most of the media coverage. The number of women and a few men attending NGO activities has risen steadily: 6,000 at Mexico City, 8,000 at Copenhagen, 14,000 at Nairobi, and 20,000–25,000 at Beijing; official delegates of states and UN agencies have increased to about 3,500 while the number of NGO representatives allowed to observe the official proceedings has grown to nearly 1,000. Nearly 30,000 participants and numerous media gave exceptional visibility to women's issues during the Beijing meeting.

Official UN conference documents are the subject of many rounds of debates in regional meetings and in preparatory committees, and the language is freighted with global debates of the day (Tinker and Jaquette 1987). Limited amendments are possible during the two-week conferences, but overtime activists learned how to insert phrases and paragraphs throughout the process so that the resultant documents reflected issues of concern raised by women's organizations and research centers. Although official UN documents lack the force of law, the moral power of conventions is considerable and provides activists in individual countries with a powerful tool to demand reform at home.

The psychological dimensions of this mobilization process should not be underestimated. The four world conferences and the many events that surrounded them provided the opportunity for women to recognize that women's agendas are legitimate, and united women across ideological and national boundaries. The networking that grew out of these lobbying efforts fed the global women's movement.

1.4 Development as a Woman's Issue

Of the three themes of the Second Decade, development was the newest and most salient to the Global South. Information on women's roles in development was scarce. A major seminar, convened by the American Association for the Advancement of Science, preceded the Mexico City conference and brought together over 100 women and men who had studied these issues. The official conference created the UN International Research and Training Institute for the Advancement of Women (INSTRAW) to facilitate further documentation, and the UN Statistical Office began to request sex-disaggregated data from all member countries. The International Labor Organization produced studies on women workers around the world (Beneria 1982). Scholars from the developing countries wrote reports for the Copenhagen conference. This outpouring of literature on women and development documented women's economic contributions to family and country, the existence of women-headed households, the impoverishment of women that followed monetization of the economy, the activities of women's organizations and NGOs in poverty alleviation programs, and the limitations placed on women by religion, culture, class, and law. Development policies have been reconceptualized as a result of these studies that make women's work—and the discrimination against them—visible.

2. WID Programming

The United States, as the major donor country, exerted strong political influence within the UN agencies as well as with the Office of Economic Cooperation and Development, a consortium of European and North American countries involved in development assistance. WID programming became an integral part of these agencies and of the many nongovernmental organizations which received funding from them. These implementing groups accepted the underlying liberal economic approach to development while criticizing policies that impacted negatively on the poor, especially women. In contrast, many academic women writing on WID issues identified themselves as marxist and emphasized the subordination and oppression of women resulting from patriarchy and capitalism (see *Feminist Theory*).

To convince skeptical development bureaucrats, practitioners advanced the 'efficiency' argument: including women as economic actors would enhance the possibility of meeting project goals. Since previous programs that dealt with women focused on women as mothers and wives, the WID office consciously avoided projects on maternal and child health or on population, to avoid the stereotype of dependent mother. Subsequently, women's health, family planning, and education programs were often combined

with income-generating projects (see *Reproductive Rights in Developing Nations* and *Reproductive Rights in Affluent Nations*).

At first the tendency was to add programs for women to existing projects with little understanding of the political or economic practices in the recipient countries. For example, providing information on subsistence crops to men in Africa was not effective since women are the primary farmers in areas south of the Sahara. Similarly, including women in village meetings to decide where a well might be placed ignored cultural prescriptions against women speaking in public.

To reach poor rural women and understand their needs, women's organizations, international NGOs, and development agencies began to focus their projects on women. The UN funneled grants to women's community groups through the Fund for Women, now UNIFEM, that was set up at the Mexico City Conference (Snyder 1995). Rural women farm, collect fuel and water, and process, prepare, and cook food. More efficient cookstoves, easier methods of grinding rice or grain, and closer wells or piped water reduced women's work hours. Groups of women were assembled to learn the new technologies; literacy was often part of the training so women could read simple instructions. Activists took the issues of women in agriculture, technology, and energy to UN World Conferences on these topics. Lobbying at these conferences broadened the base of support for WID beyond women's organizations.

Whatever the goal of the particular development project, other women's issues quickly surfaced; a characteristic of women's groups globally is that they address a wide range of concerns. For example, after the end of the Korean war, village women were brought together every month to receive birth-control pills. These women had seldom been out of the family compound once they were married. To open up their world by meeting other women and talking to the educated women from Seoul who convened the meetings expanded their horizons. These women extended their activities to reforestation, another government priority and therefore acceptable to husbands. Some groups used their income to visit shrines: for many women, riding a train was a novel experience.

As men migrated to cities in search of jobs, many rural women became temporary or permanent household heads. In contrast, in Latin America, women migrated while men stayed on the farms. Globally, about one-third of all households are supported by women; women with children are generally among the poorest households. Women-headed households became a new focus for programs directed toward reducing poverty.

The rapid monetization of the economy required married and unmarried women, urban and rural, to earn cash beyond their noncash economic activities. Income projects became the focus of many organizations. Early programs that emphasized crafts and sewing often produced poor-quality goods and seldom understood marketing. Even projects built on traditional skills seldom earned women significant profits.

Microcredit schemes have proved to be more widely successful. An early model is the Grameen Bank in Bangladesh. Landless women are organized into groups that use social pressure as collateral to ensure repayment. Members of the Bank must agree to Sixteen Decisions which encourage fiscal discipline, healthy lifestyles, and smaller families. The Grameen Bank trains the poor to reduce spending as well as to save and utilize credit. After learning to save even a few cents at each weekly meeting, women are allowed to borrow small amounts of money for entrepreneurial activities: raising a cow, selling street foods, buying a pedicab for a male relative, or even acquiring land to farm. Originally men were also borrowers, but their repayment rates were much lower; today the Grameen Bank members are about 98 percent women.

The Grameen Bank approach has been widely replicated, often without the specific social commitments required of Bank members. In Latin America, village banking is the predominant model. Women actually organize and run small banks. This model is more adapted to higher income countries. Microcredit programs have been widely introduced in the South (see Mehta et al. (1995) for a comprehensive list) and also in the United States.

Studies indicate that when women earn and control income it increases the health of their children, reduces domestic violence, and generally raises their bargaining power in the family (Tinker 1990, pp. 123–49, 150–61; see *Family Bargaining*). Increased self-confidence of the women contributes to their greater role in community and women's organizations. Other writers caution against the increased indebtedness that results from microcredit schemes. Still, credit at 16 percent per year is clearly an improvement over loans from moneylenders whose charges may be as high as 40 percent per month. Thus most families with loans increase their standard of living even if their new enterprises are not successful. Further, the poor understand the dangers of becoming beholden to moneylenders. A seven-country study of street food vendors found that the source of their capital was family and savings, not loans (Tinker 1997).

2.1 Debates About Program Efficacy

Development programs for women are multifaceted. Assisting women in traditional subsistence activities reduced the pressures on their spare time and allowed them to pursue income activities. When men migrated, women learned to make decisions and to write to their absent husbands. In cities, women often managed community organizations, earned money selling street

foods, rented out rooms even in their squatter houses, and participated in building their own homes, all with the help of development or advocacy organizations. The unintended impact of these projects was often more significant than the original goal; social engineering is not an exact science.

Debates continue about which programs actually alter existing gender relationships or challenge the systematic subordination of women (Sen and Grown 1987, Moser 1993, Mosse 1993, Kabeer 1994). Many who work within the system to increase access to resources for women and the poor are vocal critics of the dominant economic model. Others argue for programs that strategically undermine gender hierarchy and often oppose capitalist reformism. These perspectives may seem difficult to reconcile, but both groups challenge programs that adversely affect poor women, and both support the empowerment of women. In fact, most programs organize women as part of the process; once organized, women's world views expand and their sense of power increases.

A second debate concerns the positioning of women's programs within the bureaucracy. Many involved in development organizations are concerned that women's issues are still treated apart from the larger projects. Such separation results in low budgets and marginalization. These women advocate 'mainstreaming' women's programs and encourage women's organizations and NGOs to monitor project implementation to ensure women's participation (Jahan 1995). Others fear the submersion of women in mainstream programs and urge a balance with programs focused on women.

A third debate concerns the term gender, a useful analytical tool for scholars but confusing in the real world. Not only is this abstract term difficult to translate, it is almost always used as a euphemism for women. In some countries, relabeling women's programs as gender programs has diverted funding for training from women to men. Many activists believe that the use of the term gender undercuts the political importance of organizing women and creating the global women's movement, perhaps the most significant social movement of the twentieth century (Basu 1995).

3. *Contemporary Agenda for Women in International Development*

As globalization is having a greater impact on the livelihood of women and men in the South than development assistance, efforts to change policy have shifted toward international trade and monetary institutions. Structural adjustment measures imposed by the International Monetary Fund generally resulted in reduced social spending, affecting women as caretakers as well as workers. National indebtedness in the South has sparked a campaign for debt forgiveness, arguing that the gap between rich and poor cannot be closed while countries must spend a high percentage of their foreign earnings repaying debt. The World Trade Organization is challenged by a host of activists across the political spectrum. NGOs organized Women's Eyes on the World Bank.

Women's organizations are also challenging patriarchy by declaring that women's rights are human rights. Enunciated at the UN World Conference on Human Rights in Vienna in 1993, this concept challenges customary law that enshrines women's subordination and places domestic relations under civil law, and strengthens CEDAW.

Resolutions passed at the UN Conferences on Population, 1994, and Habitat, 1996, further strengthen women's rights. Enhancing women's education, eliminating violence against women, and allowing women to control their own fertility are now proclaimed as preferable ways to limit population growth (see *Gender and Reproductive Health*). Access to land and housing is recognized as a critical tool to alleviate poverty among poor women (see *Land Rights and Gender*).

Women are affecting trade policies and changing laws about women's rights by strengthening international networks and by increasing their participation in democratic institutions, particularly in those countries formerly communist countries in Europe (Funk and Meuller 1993) or in Asia (Tinker and Summerfield 1999) and in the postauthoritarian states of Latin America (Jaquette and Wolchik 1998). These dual roles in the civil society are supported by many development agencies who previously limited themselves to economic activities.

See also: Boserup, Ester (1910–99); East Asian Studies: Gender; Economic Globalization and Gender; Fertility Transition: Economic Explanations; Gender and Technology; Islam and Gender; Labor Markets, Labor Movements, and Gender in Developing Nations; Land Rights and Gender; Latin American Studies: Gender; Lone Mothers in Nations of the South; Nation-states, Nationalism, and Gender; Near Middle East/North African Studies: Gender; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender; Southeast Asian Studies: Gender

Bibliography

Basu A 1995 *The Challenge of Local Feminisms* (with McGrory E). Westview Press, Boulder, CO

Beneria L (ed.) 1982 *Women and Development*. Praeger, New York

Boserup E 1970 *Woman's Role in Economic Development*. Allen and Unwin, London (reprinted 1998. Earthscan, London)

Boserup E 1999 *My Professional Life and Publications 1929–1998*. Museum Tusculanum Press, Copenhagen, Denmark
Fraser A 1999 Becoming human: the origins and development of women's human rights. *Human Rights Quarterly* **21**(4): 853–906
Funk N, Mueller M 1993 *Gender Politics and Post-communism: Reflections from Eastern Europe and the Former Soviet Union*. Routledge, New York
Jahan R 1995 *The Elusive Agenda: Mainstreaming Women in Development*. Zed Press, London
Jaquette J S, Wolchik S L (eds.) 1998 *Women and Democracy: Latin America and Central and Eastern Europe*. Johns Hopkins University Press, Baltimore, MD
Jayawardene K 1986 *Feminism and Nationalism in the Third World*. Zed Press, London
Kabeer N 1994 *Reversed Realities: Gender Hierarchies in Development Thought*. Verso, London
Mehta R, Drost-Maasry A, Rahman R 1995 *Credit for Women: Why Is It So Important?* International Center for Research on Women, Washington, DC
Meyer M K, Prugl E (eds.) 1999 *Gender Politics in Global Governance*. Rowman and Littlefield, Lanham, MD
Moser C O N 1993 *Gender Planning and Development: Theory, Practice and Training*. Routledge, London
Mosse J 1993 *Half the World, Half a Chance*. Oxfam, Oxford, UK
Sen G, Grown C 1987 *Development, Crises, and Alternative Visions: Third World Women's Perspective, a Report from DAWN: Development Alternatives with Women for a New Era*. Monthly Review Press, New York
Snyder, M 1995 *Transforming Development: Women, Poverty and Politics*. Intermediate Technology, London
Snyder M, Tadesse M 1995 *African Women and Development: a History*. Zed Press, London
Tinker I (ed.) 1990 *Persistent Inequalities: Women and World Development*. Oxford University Press, New York
Tinker I 1997 *Street Foods: Urban Food and Employment in Developing Countries*. Oxford University Press, New York
Tinker I, Jaquette J 1987 UN Decade for Women: its impact and legacy. *World Development* **15**(3): 419–27
Tinker I, Summerfield G 1999 *Women's Rights to Housing and Land: China, Laos, China*. Lynne Rienner, Boulder, CO
Winslow A (ed.) 1995 *Women Politics, and the United Nations*. Greenwood Press, Westport, CT

I. Tinker

Economic Education

Economic education (or economics education) focuses on the scholarship of teaching economics. It encompasses the organizations advocating the need for economic literacy as well as those delivering the subject of economics. Scholars in the area address the content to be taught, methods of teaching, evaluation of those methods, and information of general interest to teachers of economics. Although economic education topics can address issues at any education, training, or schooling level, they are typically concerned with the secondary and tertiary levels. A comprehensive review of work in economic education can be found in Becker (1997) and Siegfried and Walstad (1998). The primary academic journal devoted to the topic is the *Journal of Economic Education*.

1. Organizations

In the United States, the end of World War II sparked the creation of the Joint (now National) Council on Economic Education. In the 1950s, the Committee on Economic Education was created by the American Economic Association; both work to advance the teaching of economics within the United States and throughout the rest of the world. Numerous other organizations (e.g., Junior Achievement) raise and spend money worldwide in the name of economic education, although many have a public relations mission that goes beyond the teaching of economics as a social science and a means of analysis. Most of the developed countries including the former Soviet Union now have programs aimed at increasing the economic literacy (or specific economic issue awareness) of their citizens.

Within the United States, the organizations that advocate teaching economics at precollege levels were successful in getting economics included in the Goals 2000 Educate America Act as one of 10 subjects in which every student should demonstrate competence. This 1994 United States Government Act also specifies the goal that every US adult citizen will be sufficiently literate in economics to understand and compete in a global economy. Although endorsed by the United States Department of Education, these national disciplinary content standards are not mandates from the federal government. Unlike Australia, for example, where level 12 curriculum exams (which include economics) are mandated for all students, the United States standards are intended to be a resource that local school districts adopt voluntarily.

Beyond policy statements and getting economics into precollege curricula, documentation of the actual learning outcomes of these organized efforts has been and will likely continue to be mixed. For instance, opinion surveys from North America, Europe, and Australia suggest that respondents can identify economic principles in some contexts (e.g., international trade) but not in others (e.g., minimum wage legislation) (Anderson 1999). Rapid depreciation of knowledge and skills gained in economics courses also suggests that more than a few courses in economics are required if individuals are to 'think like economists,' which at least at the upper-secondary and tertiary levels is the primary goal of courses in economics (Siegfried et al. 1991, p. 199).

2. Content

Samuel Brittan's seminal survey in Great Britain, 'Is There an Economic Consensus?' (Brittan 1973), provided evidence of the degree to which noneconomists

are thinking like economists by measuring the degree of agreement among the views of economists on economic issues and the views of a group of non-economists. His comparison groups were academic, government and business economists, members of parliament, and journalists. His questions were drawn from secondary and tertiary multiple-choice tests: the *Scottish Test of Economics Comprehensive*, prepared at the Esmee Fairbairn Economics Research Centre at Edinburgh's Heriot-Watt University, and the National Council on Economic Education's *Test of Understanding in College Economics* (TUCE).

In the 1970s, economic education researchers operated as if multiple-dimension learning outcomes could be measured validly and reliably by multiple-choice, fixed-response test questions. These standardized, advanced, secondary, and tertiary tests of economics typically include the textbook microeconomic concepts related to (a) scarcity and opportunity costs, (b) equity and efficiency, (c) cost, revenue, and profit maximization, (d) demand and supply analysis, (e) market structures, and (f) market failures and externalities. The macroeconomic concepts include (a) national income accounting, (b) real versus nominal values, (c) investment and saving, (d) money demand (liquidity) and money supply, (e) aggregate expenditure (demand) and supply, (f) stagnation and growth, and (g) monetary and fiscal policy. (The investment and saving, liquidity and money supply, and aggregate expenditure and supply framework found in tertiary economics textbooks is typically called the IS-LM-AS model.) Concepts from international trade and finance are treated on both micro and macroeconomic portions of these examinations (Saunders 1991).

Economic educators do not dispute the importance of the above macroeconomic concepts for those thinking of majoring in economics. Because so few students go on to become economists (Economists, doctored, 1998), however, the debate concerns whether the content in introductory or principles courses should be driven by the desires of those who teach more advanced courses. Unlike the mathmatically oriented economics required for more advanced study, the vocabulary of macroeconomics required for everyday life can be found in articles from the *Economist*, *Business Week*, the *Wall Street Journal*, and like publications. Kennedy (1992) provides the macroeconomic concepts that appear regularly in the news media but receive scant attention in textbooks and standardized tests. Kennedy's five key relationships that have met the test of time are (a) the discouraged–encouraged worker phenomenon and changes in reported unemployment, (b) inventories and production forecasting, (c) nominal versus real interest rates in a variety of guises, (d) inflation rate differences and exchange rate changes (purchasing power parity), and (e) the loose tendency for real, but not nominal, interest rates to be equal across countries (interest rate parity).

Once past these basic macroeconomic relationships, the debate among macroeconomists comes to the forefront. For example, Mankiw (1990, pp. 1645–6) believes that 'The IS-LM model, augmented by the Phillips curve, continues to provide the best way to interpret discussions of economic policy in the press and among policy makers.' David Romer states that even though countless teachers, students, and policymakers have found the IS-LM-AS model a powerful tool for understanding the economy, it must be updated to reflect the reality of monetary policy in different countries (Romer, 2000). Romer's reality, and also recent media headlines in the United States, place less emphasis on the money supply and more emphasis on interest rate rules.

Until macroeconomists settle the debate about the appropriateness of alternative models for macroeconomic analysis, students and their teachers of macroeconomics will be left to their own devices regarding the preferred method of analysis. Other than for definition and recall-type questions, economic educators can be expected to continue to have difficulty devising fixed-response, multiple-choice questions in this environment.

In microeconomics, Shapiro and Varian (1999) question the usefulness of concepts and nonrealistic textbook applications found on standardized multiple-choice tests. They argue that a brand new economics is not needed; instructors need only get past the concepts and ideas that no longer have much relevance in the information age. Becker (in press) summarizes those concepts and principles not covered or emphasized in principles and intermediate textbooks but which are emphasized by Shapiro and Varian (1999) in their analysis of the information industries: (a) bundling and complementarity, (b) experience goods and property rights, (c) signaling, screening, and selection, (d) expectations and risk, (e) switching costs and lock-ins, (f) cost- versus value-based pricing, (g) innovation- versus price-based competition, (h) competition within and between standards, (i) demand and supply dependence, and (j) network economies and externalities.

Competitive markets as described in textbooks adequately represent many markets (e.g., agricultural commodities, base metals, and similar primary goods), but they are not adequate for all items of interest to students. When quality depends on price (as in used-car markets, insurance, and labor markets), equilibrium may be characterized by inequality between quantities demanded and supplied; identification and a neat separation of demand and supply curves may not be appropriate. Textbook discussions of rising supply curves are problematic when marginal costs are approximately zero, as is the case for many goods at the beginning of the twenty-first century.

To get clear of the concept-laden approach to teaching economics, the national disciplinary content standards developed as part of the United States Department of Education Goals 2000 project attempt

to avoid the rhetoric employed in microeconomics (e.g., economies of scale, elasticity, diminishing marginal returns) and the controversies of macroeconomics. An attempt is made to make the language of economics accessible by 95 percent of high-school graduates (Siegfried and Meszaros 1998, pp. 143–4). In the process, however, the committee writing the standards may have made the standards so simple that they are of no practical use (Pennar 1997). Economic educators are struggling with the problems of changing the focus of examples to capture current situations better, and adapting analytical techniques to fit those examples in a meaningful way. In early 2000, however, there is no universal agreement on what knowledge, skills, and abilities are required to be economically literate.

3. *Methods of Teaching*

The Report of the Commission on Graduate Education in Economics calls for bringing 'real-world issues into the classroom' (Krueger et al. 1991). However, the Commission's assumed instructional mode is lecture. It concludes that the lack of emphasis on exposition skills is at least in part 'a judgment that the appropriate style of professional communication is something (graduate) students can figure out for themselves by watching their teachers' (p. 1049). The report of a committee of economics faculty members from prestigious liberal arts colleges in the United States (Kasper et al. 1991) emphasizes the need for graduate students to have breadth in content coverage to become good teachers of undergraduates, but it gave no attention to the need for breadth and training in teaching methods.

Until the late 1990s, economists paid little attention to teaching methods even though the American Economic Association's Committee on Economic Education had been active since the 1950s. Academic economists spoke of 'exposing' students to concepts and 'imparting' knowledge. For the transmittal of factual information and demonstrations of mathematical and graphical constructions, the lecture and blackboard were the simplest delivery modes and teachers of economics used little else (Becker and Watts 1996).

Some circumstantial evidence suggests that declining numbers of economic majors at the tertiary levels in the 1990s led academic economists to devote more time and energy to teaching. For example, Becker (2000) reports that at the 1999 Allied Social Science Association meetings in New York, 12 sessions focused on the teaching of economics, ranging from Nobel Laureate Paul Samuelson's principles textbook to faculty advisors dealing with student apprehensiveness about economics. At the 1998 ASSA meetings in Chicago, 14 sessions were devoted to teaching economics, including one headed by Nobel Laureate Ronald Coase on teaching business economics. As recently as 1996, the San Francisco meetings listed six sessions, and the January 1994 Boston meetings had four sessions on economic education, with similar small numbers in the 1980s.

Along with other educators, economic educators are now seeing the importance of small classes that require in-class interaction (Benzing and Christ 1997). The student: teacher ratio is the most important constraint in the teaching and learning process because two-way communication becomes increasingly difficult as class size increases (Rosen 1987). In small classes, each student has an opportunity to get to know and communicate with the teacher and fellow students. Emphasis can easily move from the teacher exposing and imparting knowledge to students doing economics.

Numerous activities have been developed by economic educators to get students actively involved in the classroom. Becker and Watts (1996) and Keenan and Maier (1995) describe many of these activities that can be used in the teaching of economics. For example, a 'think, pair, share' classroom activity is a cooperative learning method that might be initiated by having students think about the best way to describe the incidence and length of economic expansions in the United States since World War II. Next, they are required to 'pair up' with neighboring students and compare, discuss, and refine their answers. Finally, one member shares the team's responses with the class if asked to by the instructor. The think, pair, share activity requires only a few minutes, but provides a break in standard lectures by involving students and it allows instructors to assess quickly, and on a regular basis, what students are learning.

Students are not alike in their willingness to get involved in classroom group activities. Cooperative classroom learning experiences can require large amounts of time. As long as decisions outside the classroom are made on an individually competitive basis, students may choose to perform individually in the learning environment as well. Instructors in economics are now expected to employ alternative teaching strategies—strategies that can keep students actively involved, with both practice and feedback. Researchers in economic education have found that technology can complement traditional classroom activities (Agarwal and Day 1998). As demonstrated by the exponential growth in the number accessing the *Journal of Economic Education* World WideWeb site, and the creation of the 'Online' section of that journal, economists are increasing their use of the Internet in their teaching. At the tertiary level departments of economics are exploring ways to offer courses via the Internet. However, unless computers and other high-tech systems can replace the instructor and support personnel, their introduction always adds to the cost of instruction.

Replacing the instructor with technology that adapts to the needs of individual learners is a dream that no instructional developer has come close to achieving. Physical capital may be a poor substitute for human capital in economic education. Rosen (1987) observed that no major innovation in education has occurred since the printing press made inexpensive book publication possible. Higher education is a major contributor to technological advances elsewhere in the economy, yet practitioners in higher education appear to be unable to capitalize on this technology in teaching. The World Wide Web may lead to changes in educational practices, but it is too early to say much about the effectiveness and efficiency of Internet tools in teaching economics.

The speed with which economists embrace new technologies in teaching and participate in the scholarship of teaching depends on individual interests, monitoring methods, reward structures, and market forces. In a search for the best of these technologies, economic educators have learned that the conclusion of Siegfried and Fels (1979, p. 953) still holds: 'Different students learn economics in different ways. The best teaching strategy provides alternative learning methods.'

4. *Assessment of Teaching and Learning*

Assessment of teaching and learning in economic education is of two forms: formal quasi-experimental work aimed at publication and informal classroom review of students and teachers.

4.1 *Quasi-experiments*

Rendigs Fels provided the impetus for quantitative analysis of teaching techniques in economic education with his 1969 Presidential Address to the Southern Economic Association, 'Hard Research on a Soft Subject: Hypothesis Testing in Economic Education.' Until the 1990s, researchers typically specified a multiple-choice score, or the change in scores from the beginning to the end of the course, as the dependent variable produced by human capital inputs (preknowledge of economics or some other subject, grade points, previous courses of study), utilization measures (time spent by student or teacher in given activities), and technology, environment or mode of delivery (lectures, group work, computer use).

Of all the variations in outcome measures and models considered, Becker (1997) concluded that the only variables that consistently appear as significant and meaningful explanatory variables of the postcourse scores are the preaptitude measures, such as the precourse test scores. There also is some work that suggests the qualifications of the teacher, math skills, prior exposure to economics, gender, and type of teacher and student interaction are important in explaining success in tertiary economics courses. However, there are major problems with the input–output approach on which economic education research is based.

First, production functions are only one part of a student's decision-making framework. Observed outcomes are the result of the entire decision-making process, and each student's learning goals and process may be different. Second, data loss and the resulting prospects for sample selection bias in the standard learning model design are tremendous, with 20–40 percent of those taking the precourse tests not taking the postcourse tests. Third, from statistics we know that failure to reject the null hypothesis does not imply its acceptance. That an experimental program shows no statistical significance over the control does not imply that the experimental and control programs have equal effects. Finally, and possibly most important, education is a multiproduct output that may not be reflected in a single multiple-choice test score. In the 1990s, economic education researchers began to build models and employ estimation techniques aimed at overcoming these problems in testing for learning effects. These developments are discussed in Becker (1997), with new findings typically published in the *Journal of Economic Education*.

4.2 *Classroom Assessment of Students*

Within the United States, fixed-response, multiple-choice tests are a staple of assessment in economics, especially at the principles level. In most of the rest of the world, even at the principles level, free-response questions are the norm. An often-asked but unanswered question is whether the added cost of grading essays, reports, term papers, and the like is worth the information gained by students, instructors, and their institutions. In large enrollment principles classes, cost considerations require the use of multiple-choice tests, but this does not imply that they cannot be used creatively to get students involved and interacting with each other. For example, after a quiz has been machine scored, and returned to students, the instructor can have students interact with those sitting around them to determine the correct answers—a frenzy of activity ensues. After several minutes, the room is quieted and each student is allowed to resubmit a new answer sheet for partial credit.

Individual multiple-choice questions have also been used outside an examination setting to stimulate student interaction in class. For example, a single multiple-choice question can be projected in a traditional large lecture hall. Each student answers the question on a sheet distributed in the classroom. For a second question, each student marks 'A. Certain' to 'E. Doubtful' to indicate confidence. Students then discuss their answers with neighbors—the lecture hall is abuzz. After a few minutes, students continue the

process with question three (a repeat of question one) and final confidence given in question four (a repeat of question two). Student attendance and participation increase with the use of this activity, if students get some credit for attempting answers, as well as for the correct answer to number three, when responses are scanned and machine scored. The instructor also gets feedback on student confidence in what they are doing.

From Harvard seminars on college teaching, Light (1990, 1992) reported on the subjects students respect and from which they learn most. The crucial features are immediate and detailed feedback, with frequent points of assessment, and high demands and standards but with ample opportunity to revise and improve work as part of the grading process (i.e., learning from mistakes). Students also claim to learn from the reinforcement of their peers. Multiple-choice tests are crude instruments for assessing student learning but, as already illustrated, they can be used in educationally sound ways. Allen Kelley's TIPS program (Kelley 1973) was innovative in the early 1970s and its principal components (frequent multiple-choice testing in large classrooms with immediate detailed feedback) are easily implemented by anyone with access to machine scoring. It is especially attractive to anyone teaching in a small computer laboratory where assessment material can be delivered directly and uniquely to each student in the class, as I am doing in my teaching of business and economics statistics. In a large high-tech auditorium, where each student's response on a keypad can be monitored by the instructor, software programs permit rapid display of responses and their distribution, offering instructor and students immediate checks on learning.

Especially outside the United States, as already noted, there is a belief that no matter how they are delivered, multiple-choice questions are inferior to questions in which students construct their own responses. But within the United States, the perception is that they measure the same thing (Hansen 1998, p. 83). Walstad and Becker (1994, p. 196), for example, reported in their study of the Advanced Placement examination in economics (administered to selected twelfth graders in the USA) that 'the coefficients of correlation for the multiple-choice score and the essay score were relatively high (0.69 micro and 0.65 macro), … consistent with previous studies that found no difference or only slight differences, in what the two types of tests and questions measure.' Becker and Johnston (1999), however, point out that a correlation coefficient of 0.69 implies that only 48 percent of the variability in scores on one set of fixed-response (multiple-choice) or construct-response (essay) questions is explained by the other.

More importantly, once account is taken of the measurement error and simultaneous relationship that exists between scores on essay and multiple-choice components of an exam, the Becker and Johnston two-stage least-squares estimates demonstrate that they measure different dimensions of student knowledge. Hence it is important to use both forms of testing and not to abandon essay questions in favor of the multiple-choice form or vice versa. Second, females do not seem to be disadvantaged by multiple-choice questions, although they may have an advantage at essay writing, again suggesting that no one form of test should be employed. Economic educators such as Hansen (1998, p. 83) now suspect that tests with all essay response questions, or the exclusive use of multiple-choice response questions, must be avoided in the assessment of economics.

4.3 Assessment of Teachers

In economics departments at colleges and universities in the United States, there is an almost exclusive reliance on end-of-semester student evaluations of teaching (SET) as the measure of the instructional product (Becker and Watts 1999). That 'the primary purpose of the common end-of-course evaluation form is to provide comparative data for administrators … ' (Walstad and Saunders 1998, p. 339) is troubling. Typically SET scores explain less than 50 percent of the variability in other teaching outcomes (e.g., test scores, scores from trained classroom observers, postcourse reviews, etc.). Thus, when administrators use SET scores to compare individual instructors against an average, they are condemning half of their teachers based on a suspect instrument. Teachers (agents) can be expected to respond to SET and, even if inappropriate, adjust their behavior in accordance. For that reason, there are principal-agent problems associated with end-of-semester SET.

Unfortunately, the SET forms used in economics usually ask few questions that deal with what education specialists say is important: active student learning and group (or collaborative) learning. Furthermore, although academic economists call for the use of better applications and examples in teaching, these items are among the least-often-asked questions on SET. Although there is lip service about implementing new technology in teaching economics, questions about the use of technology are rare on SET forms. The teacher's overall effectiveness, communication skills, organization and planning, and knowledge of material are among the top four items on which student opinion is sought (Becker and Watts 1999). These items, however, fail to address what instructors can do to improve teaching. An instructor interested in improving the learning of his or her current students needs feedback before a semester is nearly over. Questions asked of students must elicit responses that suggest a desirable change in instructor behavior. They need not be of the standard SET type.

In addition to my participation quizzes, similar classroom checks on student understanding such as the 'minute paper' provide a proven framework to

assess what students are and are not learning during the course (Chizmar and Ostrosky 1998). The minute paper is typically assigned in the last minutes of each class when maintaining student attention is difficult. Each student is required to respond to two questions: (a) what was the most important thing you learned in class today? and (b) what question is unanswered? Although answers may or may not be graded, these written responses can be used to monitor both comprehension and attendance. It is only one of many vehicles for monitoring the process of education from the start of a course to its termination, which fits well with the industrial model of total quality management. On resident campuses with ample computer facilities and programming staff, electronic technology makes periodic assessment easy. Students need not complete assessment instruments in a classroom; they can complete questionnaires at their leisure sitting at terminals. Students can be required to complete the periodic questionnaires as part of a course requirement. If the unit processing the forms is independent of the instructor, individual student responses can be associated with the student or reported anonymously to the instructor.

Classroom observation, peer review of teaching materials, and subsequent student enrollments and student behavior in subsequent economics courses are some alternatives to SET, but these are not being used in the evaluation of teaching in economics (Becker and Watts 1999). For example, peer review of teaching is used by only 37 percent of the United States economics departments at research universities and by 52 percent of the baccalaureate institutions, as defined by the Carnegie Foundation for the Advancement of Teaching. Wherever peer reviews are employed, the Department Chair typically appoints the evaluators. Classroom observation is the most common form of peer review of teaching. Evaluators also examine course syllabi, instructional materials, and examinations. Few departments in the United States ask evaluators to consider student drop rates, although some give attention to grade distributions.

5. *Conclusion*

The scholarship of teaching economics is diverse. It encompasses the organizations advocating and delivering economic education. More importantly, it provides a body of knowledge and views regarding what is and should be taught, how it is and should be taught, its assessment and evaluation with regard to student learning and attitudes, and teacher performance in generating those changes in students.

See also: Economic History; Economics: Overview; Education: Phenomena, Concepts, and Theories; Educational Philosophy: Contemporary Views

Bibliography

Agarwal R, Day A E 1998 The impact of the Internet on economic education. *Journal of Economic Education* **29**: 99–110

Anderson M E 1999 *Neoclassical Economics and the Australian Community*. University of Melbourne–Melbourne Institute of Applied Economic and Social Research Faculty of Economics and Commerce, Melbourne, Australia

Becker W E 1997 Teaching economics to undergraduates. *Journal of Economic Literature* **35**: 1347–73

Becker W E 2000 Teaching economics in the 21st century. *Journal of Economic Perspectives* **14**: 109–19

Becker W E, Johnston C 1999 The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record* **75**: 348–57

Becker W E, Watts M 1996 Chalk and talk: A national survey on teaching undergraduate economics. *American Economic Review* **86**: 448–53

Becker W E, Watts M 1999 How departments of economics evaluate teaching. *American Economic Record* **89**: 344–9

Benzing C, Christ P 1997 A survey of teaching methods among economics faculty. *Journal of Economic Education* **28**: 182–8

Brittan S 1973 *Is There an Economic Consensus*? Macmillan, London

Chizmar J, Ostrosky A 1998 The one-minute paper: Some empirical findings. *Journal of Economic Education* **29**: 3–10

Economists, doctored 1998 *Economist* May 9: 57

Hansen W L 1998 Integrating the practice of writing into economics instruction. In: Becker W, Watts M (eds.) *Teaching Economics: Alternatives to Chalk and Talk*. Edward Elgar, Cheltenham, UK, pp. 79–118

Kasper H 1991 The education of economists. *Journal of Economic Literature* **29**: 1088–109

Keenan D, Maier M H 1995 *Economics Live: Learning Economics the Collaborative Way*. McGraw-Hill, New York

Kelley A C 1973 Individualizing education through the use of technology in higher education. *Journal of Economic Education* **4**: 77–89

Kennedy P 1992 Teaching tools: On journalists' use of macroeconomic concepts. *Economic Inquiry* **30**: 194–201

Krueger A O 1991 Report of the commission on graduate education in economics. *Journal of Economic Literature* **29**: 1035–53

Light R J 1990 *The Harvard Assessment Seminars*. First Report, Harvard University Graduate School of Education and Kennedy School of Government, Cambridge, MA

Light R J 1992 *The Harvard Assessment Seminars*. Second Report, Harvard University Graduate School of Education and Kennedy School of Government, Cambridge, MA

Mankiw N G 1990 A quick refresher course in macroeconomics. *Journal of Economic Literature* **28**: 1645–60

Pennar K 1997 Economics made too easy. *Business Week* January **20**: 32

Romer D 2000 Keynesian macroeconomics without the LM curve. *Journal of Economic Perspectives* **14**: 149–69

Rosen S 1987 Some economics of teaching. *Journal of Laboratory Economics* **5**: 561–75

Saunders P 1991 The third edition of the test of understanding in college economics. *Journal of Economic Education* **22**: 255–72

Shapiro C, Varian H R 1999 *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, Boston

Siegfried J, Fels R 1979 Research on teaching college economics: A survey. *Journal of Economic Literature* **17**: 923–69

Siegfried J J, Meszaros B T 1998 Voluntary economics content standards for America's schools: Rationale and development. *Journal of Economic Education* **29**: 139–49

Siegfried J J, Walstad W 1998 Research in teaching college economics. In: Walstad W, Saunders P (eds.) *Teaching Undergraduate Economics: A Handbook for Instructors*. Irwin McGraw-Hill, Boston, pp. 141–66

Siegfried J, Bartlett R L, Hansen W L 1991 The status and prospects of the economics major. *Journal of Economic Education* **22**: 197–224

Walstad W B, Becker W E 1994 Achievement differences on multiple choice and essay tests in economics. *American Economic Review* **84**: 193–6

Walstad W, Saunders P 1998 Using student and faculty evaluations of teaching to improve economics instruction. In: Walstad W, Saunders P (eds.) *Teaching Undergraduate Economics: A Handbook for Instructors*. Irwin/McGraw-Hill, Boston, pp. 337–55

W. E. Becker

Economic Geography

Economic geography is the study of the geography of economic activities. The field includes the study of where economic activities are located (their pattern or distribution), why they are found where they are (the processes by which they are located), and what their impact is on that place and other places. In addition, normative studies of optimal location (where activities should be located) also fall within economic geography.

1. *The Scope of Economic Geography*

Economic geography includes a wide range of topics that span production, consumption, and distribution activities. The subject matter of economic geography encompasses the traditional sectors of economic activity: the primary sector (agriculture, fishing, forestry, and mining), the secondary sector (manufacturing, including processing and assembly), and the tertiary sector, or services. Services, the most rapidly-growing part of most of the world's economies, are often disaggregated into quaternary (information-related) and quinary (decision-making) services. The interrelationship of these sectors and changes in them affect employment and how one makes a living. In addition to the sectors of activity, economic geography has seen as a central objective to understand the uneven development of regions, one of the foremost impacts of the location of economic activities.

Most human impacts on the earth's environment have been the result of economic activity of one type or another. These processes take place at various scales (local, regional, national, global) simultaneously (see *Scale in Geography*). The interconnected nature of human economic activity is seen in the interaction between subsistence farming and the market, which provides nonfarm work and income as well as goods that may compete with traditional commodities.

The primary sector remains integral to the field. The range of forces that affect agriculture include the physical environment, economic behavior and demand, institutional, social and cultural influences; in turn, agriculture affects the natural environment (Grigg 1995). Among the more prominent forces that affect agriculture are technological change, which has permitted massive increases in productivity, and the increasingly global links from growers to consumers. Globalized agribusiness illustrates the central role of large, transnational firms in the organization of this (and other) economic activities (Goodman and Watts 1997, Le Heron 1993).

Dependence on the primary sector has numerous disadvantages. Oversupply has weakened countries that are producers of many mineral commodities. Deforestation is a common result of high levels of demand for scarce types of wood. Energy and other natural resources are typically addressed separately (see *Energy, Geography of*; *Resource Geography*).

The secondary sector, manufacturing and processing industries, is the core of industrial geography, although the latter increasingly incorporates any or all of the sectors as firms have begun to operate across traditional sectoral boundaries (see *Industrial Geography*). Although economic geography formerly was largely focused on production, it has grown to include the services, finance, information, technology, consumption, and state regulation (Martin 1994).

Services, the tertiary sector, comprise intangible activities ranging from personal services such as cutting hair to research and legal advice. Services have expanded enormously in recent years as higher incomes permit more people to travel and eat in restaurants and to use computer software.

A conventional textbook organization of economic geography begins with population and resources—population because people are both workers and consumers in the economy, and resources because what is available locally will determine to some degree what must come from elsewhere via trade. Distribution activities (especially transportation), agriculture, cities, the international economy and variations in economic development are standard topics.

2. *A Brief Historical Look*

Economic geography has perhaps the widest scope of all of the systematic areas of geography. It originated to a large degree from the identification of resources that had commercial or economic value. Extraction and manufacturing, comprising commercial geography, formed the principal focus of research by economic geographers in the first quarter of the twentieth century. The intervening decades saw a great deal of

research related to transportation (see *Transportation Geography*) and cities, where most economic activity takes place (see *Urban Geography*), industry location and, increasingly after the Second World War, to retailing and (what would now be called) the geography of the service economy (see *Service Economy, Geography of*; *Retail Trade*; *Market Areas*). Most research remained quite local, however, concentrating on specific areas and individual industries. As economic activity became less dependent on direct inputs from agriculture, mining and other primary activities, economic geographers found less relevance from their colleagues in physical geography.

Theoretical concerns in economic geography for many years centered on examining patterns linked to theories and models, such as central place theory and empirical verification (Berry et al. 1988). During this period, economic geographers were among those involved in the founding of regional science as an interdisciplinary field (see *Regional Science*). Regional science also brought to economic geography an array of models for optimal location of industrial and service activities and of transportation and trade flows (see *Location Theory*; *Spatial Optimization Models*; *Spatial Interaction*; *Spatial Equity*). Many of the foci of optimization modeling concern flows related to economic phenomena, such as journeys to work and shopping trips. It has been difficult to utilize equilibrium approaches to explain some flows, such as of traded goods. Theories of trade, for example, do not encompass adequately the growth in intrafirm and intraindustry trade or the effect of government decisions and policies (see *International Trade: Geographic Aspects*).

3. *The Concerns of Economic Geography*

It has been necessary to cross several disciplinary boundaries, not only that with economics. It is common for geographers to rely on research in planning, sociology, engineering, and other fields in order to understand better the processes that form the patterns found on the landscape. Within the core of economic geography, several topics helped to set the direction of the future: the world economy and the capitalist system, firms and linkages among them, regional development and planning, and work and employment.

3.1 *Understanding the World Economy*

Paradoxically, understanding the economic geography of places and regions within the world economy has become more significant with the process of globalization. Globalization has taken place simultaneously with the rise of the service industries and their concentration in world cities (see *Globalization: Geographical Aspects*; *Cities: Capital, Global, and World*).

Regional development, the impact of change in the type and nature of economic activity, has been a significant focus of both theoretical and policy-related research since the 1970s, as massive spatial changes accompanied the end of the golden era of economic growth. Regional development attempts to capture the range of economic, political, and social influences on regional patterns of employment opportunity. The patterns and processes of change typically exhibit uneven development, which is central to the study of economic geography (Massey 1995). This includes economic restructuring, the term used for developed economies, development theory, referring to less developed countries, and economic transformation or transition, referring to the evolution from central planning to a market economy (see *Economic Restructuring: Geographic Aspects*; *Uneven Development, Geography of*; *Economic Transformation: From Central Planning to Market Economy*). As active policy initiatives fell in importance during the 1980s, research became more local in focus, highlighting the unique cultural and social circumstances of particular localities (see *Local Economic Development*; *Community Economic Development*).

Research into the nature of capitalism attempts to understand how the capitalist system works through analytical approaches. A major conclusion of analytical research on capitalism is that the 'golden path of accumulation' in a capitalist system is unstable, and that a focus is needed on disequilibrium paths, centrally those surrounding technical change (Webber and Rigby 1996). The slowdown of accumulation, or fall in profitability, that began in 1973 and resulted in falling productivity can be attributed to technical change, which enhances the ability of firms, industries and regions to avoid the persistently falling rate of profit. While the disequilibrium view finds many adherents, most research in this area on technical change is limited to process innovations, and not to new products, which many would credit with the bulk of capitalism's ability for Schumpeterian 'creative destruction.' The view of the economy as evolutionary and of firms as heterogeneous finds stimulus in several branches of economics outside the neoclassical mainstream, and especially on the topics of economic growth and technological change (Nelson 1998).

The political economy approach explicitly recognizes that 'capitalism is not just a phenomenon of economic geography. It is also at one and the same time a social, cultural, and political geography' (Peet and Thrift 1989). From this perspective and others, economic geography now incorporates social, cultural and political processes on a regular basis (Lee and Wills 1997).

The incorporation of cultural, social, political, and institutional influences within economic geography is fundamental to the topic of industrial districts and

networks, which has highlighted the social embeddedness of these influences. On this phenomenon, identified throughout the world, processes beyond the purely economic are seen as critical (Lee and Wills 1997, Martin and Sunley 1998). These include institutions, culture, and social ties and conventions, all of which affect dramatically economic geography, whether production or consumption.

To understand the capitalist world economy requires understanding the role of money and finance. Financial flows have become not only international but continual as markets in London, New York, and Tokyo serve as continental financial markets. The nature of global money is just beginning to be understood (see *Finance, Geography of*).

Money, knowledge, and other phenomena move throughout the world economy via networks, whether physical or virtual. Attempting to understand telecommunications has been especially important as those technologies have become pervasive within and among firms as well as elsewhere in everyday life. The transforming systems of computers and telecommunications are outcomes of technological change taking place at a large (global or regional) scale as new infrastructures enable new forms of production, including many producer services (see *Information Society, Geography of*).

Technological change is a source of resilience in the capitalist system, and repeatedly alters the nature of economic activity and its geography. This is the case particularly through the creation of new products, which broaden the range of possibilities for consumers and present substitutes for inputs. The topic of technological change also examines patterns of technological change, including the origins of new innovations, or places where technological goods and services are produced, whether called high-technology regions or technology districts (see *Technology Districts*).

3.2 Understanding Firms

Early studies focused on actual spatial patterns of industrial location, as well as optimal patterns, a theme that continues (see *Location Theory*). The object in geographers' study of firms is less the organizational or industrial form of these entities than their actual behavior. Understanding the mechanisms by which firms and industries operate and their impact on industrial location and regional development also is a central issue in *Industrial Geography*. Along these lines, of particular interest have been studies of vertical integration and disintegration, or what is inside a firm and what is obtained from other firms.

As companies became vertically integrated, absorbing necessary processes in-house, studies of large firms evolved into what was known as the 'geography of enterprise.' Globalization is an effect of the actions of large (multinational, transnational, global) companies, whether through trade, foreign direct investment (see *Foreign Investment: Direct*), or other arrangements (Cox 1997, Dicken 1998). This work has been complemented if not largely superseded by the view that large firms are less influential in their own right than merely taking advantage of divisions of labor, an interpretation that goes a long way toward explaining the geography of production (Massey 1995, Sayer and Walker 1992).

In addition to large firms, entrepreneurship and small firms are a central feature of both flexible accumulation and creative destruction, and therefore have received increasing attention as their significance in most economies has grown. This is most evident in the case of regions where new firms and new industries are common (see *Technology Districts*).

3.3 Connections Among Firms

A surge of flexibility appeared to signal the vertical disintegration or 'outsourcing' of production as firms increasingly 'bought' rather than 'made' needed physical and service inputs. Flexible production became the dominant paradigm in economic geography, yet it is but one of several manifestations of a dramatic change or restructuring of economic activity (Barnes 1996, Martin 1994). The shift toward services and the interdependence of goods and services stand out in this regard.

Vertical integration had not vanished, however, but was transformed through complex relationships with suppliers and subcontractors. The global economy operates through complex webs of linkages, both vertical and horizontal, that form production chains that take different form in each industry (Dicken 1998) (see *Globalization: Geographical Aspects*). Large firms remain central to many industrial districts and other centers of flexibility because of their ability to control global networks. Despite their power, large firms have had to make painful adjustments and adaptations of flexible forms and processes. Adjustments to structural conditions are not simple, as sunk costs and uncertainty make even the largest firms unable to make optimal choices (see *Industrial Geography*).

The study of the interdependence of firms takes two forms: aggregate input–output relations and more detailed firm-specific linkages. Input–output structures, while often based on surveys of firms, provide economy-wide, usually national, structures of interindustrial links, but also regional and interregional. Disaggregated studies of linkages, by contrast, focus on a single firm or a small number of firms, illuminating the geographical area within which a firm's economic activities have influence. This may range from very local to global, depending on the firm and its sector, and only detailed study identifies which.

3.4 Economic Development and Restructuring

Economic development, long considered a national-level process, has been viewed widely to have broken-down to more regional and local processes. It is less frequently national states and more frequently regions where the dynamics of economic change take place (Scott 1998, Storper 1997).

Some of this work has looked at the restructuring as a transition from the capitalist 'golden age' to something different (Webber and Rigby 1996). Economic restructuring within capitalism is a result of the several distinct processes: sectoral change, especially from manufacturing to services (deindustrialization and tertiarization), flexibilization, the increasing feminization of the labor force, the new technologies of the information economy, hyper-consumerism, globali-zation, and deregulation and privatization (Martin 1994, Townsend 1997).

Fundamental to analysis of restructuring is the idea that interfirm competition is more intense and less stable than previously, with new rivals emerging from both other industries and from newly founded firms. New firm formation, or entrepreneurship, is particularly prominent in new industries, aided by the instability fostered by technological change. New firm formation is perhaps most a product of social class and of industrial history that create the culture and social norms in places (Massey 1995).

Spatial manifestations of restructuring—changes in firms' locational decisions—are seen both inter-regionally as some regions rise and others decline, and intraregionally as some suburban nodes become cores of economic activity, especially of producer services. Innovative milieus represent an intriguing interplay between firms and territories, where those territories include not only other firms but also infrastructures, knowledge, and institutions and organizations (see *Economic Restructuring: Geographic Aspects*).

The optimal location of activities falls under the umbrella of location theory, and is closely related to the 'new' economic geography that has developed within the discipline of economics in recent years (see below). That research has developed models to determine optimal and/or equilibrium patterns of location (see *Location Theory*). It should be noted that location theory is suited poorly to incorporate the realities of industrial location, especially the maximization of the quality of inputs rather than the minimization of transport and input costs.

3.5 Work and Workers

Uneven development and the spatial division of labor created are among the most fundamental constructs in economic geography. The spatial division of labor puts at the forefront the fact that not only different numbers of jobs (or varying amounts of economic opportunity) are found in different places, but also that different kinds of jobs are found in different locations. This view of spatialized social relations results in a geography not of the distribution of jobs but of power relations, of dominance and subordination, of enablement and influence (Massey 1995). Class, gender, and the division of labor are 'intertwined structures' that affect virtually all aspects of work and the geographies of the labor market (Hanson and Pratt 1995, Peck 1996, Sayer and Walker 1992) (see *Spatial Labor Markets*).

The types of work made available in a place affect its potential for entrepreneurship and economic development. Some occupations are more amenable to new conditions and innovative activity, and a great deal appears to be cumulative. Existing theories, however, generally are unable to account for empirical patterns found.

3.6 The 'New' Economic Geography

A much more narrow definition of economic geography is that employed by economists practicing what is termed the 'new' economic geography (Krugman 1991, 1995, 1996, Ottaviano and Puga 1998). In this view, economic geography is 'the location of production in space,' the most striking feature of which is its agglomeration or concentration—'clear evidence of some kind of increasing returns' (Krugman 1991).

Models in this genre tend to focus on two regions (urban and rural or rich and poor) and two sectors, and aim toward or assume equilibrium or multiple equilibria, rather than the reality of disequilibrium. The highly restrictive models fail to approximate 'real places' or to incorporate spatial scale (Martin 1999). Echoes of this critique of mainstream economic growth theory lament the failure to include technological advance as an essentially disequilibrium process, a theory of the firm in which capabilities and differences across firms are central elements, and incorporation of a richer body of institutions (Nelson 1998). The difference between 'formal theory' and 'appreciative theory' is useful in the context of the debate on the new economic geography.

The 'new economic geography' models the centripetal forces that foster geographic concentration and agglomeration against the centrifugal forces such as immobile factors (including land, natural resources, and people) (Krugman 1998). Several 'modeling tricks' are used to arrive at an equilibrium solution, or multiple equilibria, to topics and situations studied by geographers several decades ago, such as central places and rank-size distributions (Berry et al. 1988). It is not surprising that 'geography turns out to be perhaps the most naturally 'nonlinear' area of economics' (Krugman 1998), but to geographers the critical point is the diversity of reasons why this so, not merely the mathematical approximation of it (Martin 1999).

Geographers are more inclined to focus on the empirical reality present in the world and the many geographies they represent (Lee and Wills 1997). Rather than linear, mechanistic metaphors, biological and evolutionary metaphors may capture more accurately the emergence of economic change. The broader conception draws not only from interdisciplinary links with other social sciences, but also from economics, where development in both institutional and evolutionary economics have meshed well with geographers' broader view of the world. In general, economists prefer to assume equilibrium, whereas geographers', restricted cognitive understanding of the world, and unequal locational advantage as the rule rather than the exception (Sheppard et al. 1998). The world is much more diverse than a single set of models can hope to capture (Martin 1999).

Spillovers and externalities (and interfirm linkages) are at the center of the debate. The presence of economies of scope and diverse locales turns the standard increasing returns model on its head, because it is based on an industrial organization that benefits more from intraindustry linkages than from scale of production. Likewise, knowledge spillovers across industries are ignored in the 'new' economic geography because they are not well understood or systematically measured and tracked (Krugman 1991), and they are seen as a consequence of market interactions involving economies of scale at the level of the individual firm (Krugman 1995). By contrast, untraded interdependencies, which are not able to be modeled, are central to much of the current thrust of research by geographers on learning in regions and localities (Pinch and Henry 1999, Storper 1997).

3.7 *Clusters, Agglomeration, and Territorialized Learning*

Among the most powerful processes for regional transformation and renewal is the process of regional learning. If organizations and regions can learn, it is because they are able to build and use relationships with other organizations in their environment. Low-tech as well as high-tech firms are able to draw upon spillovers of knowledge in this way (Maskell et al. 1998).

The presence of horizontal (interindustry) linkages and spillovers is expected especially within industrial districts and technology districts, now seen as special cases of clusters. Paradoxically, as globalization continues, regions—whether city-states such as Singapore, metropolitan areas, or cross-border territories—have become more significant as the loci of economic activity (Scott 1998, Storper 1997). Each region evolves a unique mix of products and technologies, skills and resources, conventions and customs, leading to a set of capabilities not found elsewhere. These geographically-specific capabilities—'people' in Krugman's list of centrifugal forces—can serve as enduring competitive advantages of agglomeration (Maskell et al. 1998). They also create the variety that is evident in the economic geography of our world.

4. *Conclusion*

The breadth of economic geography makes it virtually certain that research focuses on one or more of the field's topics. Economic geography has blossomed well beyond its original focus on the location of production to embrace the many other human forces—social, cultural, political, and institutional—that affect and are affected by economic activity.

See also: Economic Sociology; Globalization: Geographical Aspects; Regional Geography

Bibliography

Barnes T J 1996 *Logics of Dislocation: Models, Metaphors, and Meanings of Economic Space*. Guilford Press, New York

Berry B J L, Parr J B, Epstein B J, Ghosh A, Smith R H T 1988 *Market Centers and Retail Location: Theory and Applications*. Prentice-Hall, Englewood Cliffs, NJ

Cox K R (ed.) 1997*Spaces of Globalization*. Guilford Press, New York

Dicken P 1998 *Global Shift*, 3rd edn. Guilford Press, New York

Goodman D, Watts M J (eds.) 1997 *Globalising Food: Agrarian Questions and Global Restructuring*. Routledge, London

Grigg D B 1995 *An Introduction to Agricultural Geography*, 2nd edn. Routledge, London

Hanson S, Pratt G 1995 *Gender, Work, and Space*. Routledge, London

Krugman P 1991 *Geography and Trade*. MIT Press, Cambridge, MA

Krugman P 1995 *Development, Geography, and Economic Theory*. MIT Press, Cambridge, MA

Krugman P 1996 *The Self-Organizing Economy*. Blackwell, Oxford, UK

Krugman P 1998 What's new about the new economic geography? *Oxford Review of Economic Policy* **14**(2): 7–17

Lee R, Wills J (eds.) 1997 *Geographies of Economies*. Arnold, London

Le Heron R 1993 *Globalized Agriculture: Political Choice*. Pergamon, Oxford, UK

Martin R 1994 Economic theory and human geography. In: Gregory D, Martin R, Smith G (eds.) *Human Geography: Society, Space, and Social Science*. Macmillan, London and University of Minnesota Press, Minneapolis, pp. 21–53

Martin R 1999 The new 'geographical turn' in economics: Some critical reflections. *Cambridge Journal of Economics* **23**: 65–91

Martin R, Sunley P 1998 Slow convergence? The new endogenous growth theory and regional development. *Economic Geography* **74**: 201–27

Maskell P, Eskelinen H, Hannibalsson I, Malmberg A, Vatne E 1998 *Competitiveness, Localised Learning and Regional Development*. Routledge, London

Massey D 1995 *Spatial Divisions of Labor: Social Structures and the Geography of Production*, 2nd edn. Routledge, New York

Nelson R R 1998 The agenda for growth theory: A different point of view. *Cambridge Journal of Economics* **22**: 497–520

Ottaviano G I P, Puga D 1998 Agglomeration in the global economy: A survey of the 'new economic geography.' *The World Economy* **21**: 707–31
Peck J 1996 *Work-Place: The Social Regulation of Labor Markets*. Guilford Press, New York
Peet R, Thrift N 1989 Political economy and human geography. In: Peet R, Thrift N (eds.) *New Models in Geography*. Blackwell, Oxford, UK, pp. 3–29
Pinch S, Henry N 1999 Paul Krugman's geographical economics, industrial clustering and the British motor sport industry. *Regional Studies* **33**: 815–27
Sayer A, Walker R 1992 *The New Social Economy: Reworking the Division of Labor*. Blackwell, Oxford, UK
Scott A J 1998 *Regions and the World Economy*. Oxford University Press, Oxford, UK
Sheppard E, Plummer P S, Haining R P 1998 On economists and geographers: A response to Harwitz. *Journal of Regional Science* **38**: 675–76
Storper M 1997 *The Regional World*. Guilford Press, New York
Townsend A R 1997 *Making a Living in Europe: Human Geographies of Economic Change*. Routledge, London
Webber M J, Rigby D L 1996 *The Golden Age Illusion: Rethinking Postwar Capitalism*. Guilford Press, New York

E. J. Malecki

Economic Globalization and Gender

Economic globalization and gender is a topic that has grown in importance since the1970s. Although 'globalization' has occurred in specific periods for centuries, it was not until the early 1970s that attention was drawn to the importance of understanding the gendered effects of economic globalization and, in turn, their implications for sustainable human development.

1. *Definition and Context*

Economic globalization involves a wide variety of processes, opportunities, and problems related to the spread of economic activities among countries around the world. There have been many periods in which it occurred, most recently including the latter nineteenth century to WWI, the quarter century after WWII, and the late 1960s/early 1970s to the present.

The latest period has involved several major trends, as capitalism has spread throughout more of the world. First, there has been an increased reliance on markets (versus government involvement in the economy) by most nations (including industrialized countries, developing countries, and formerly socialist countries like China with over one-fifth of the world's population). Second, many developing countries have shifted to the more open export-oriented approach based on production for external trade from an import substitution development strategy (production of essential goods for the internal market). Third, multinational corporations (MNCs) in manufacturing, service, and finance sectors have moved into new tiers of countries and have established burgeoning networks of subcontractors in many areas. Fourth, since the late 1970s, economic globalization has also involved structural adjustment policies (SAPs), mandated by the International Monetary Fund (IMF) as a condition for granting countries loans. SAPs require governments to take many steps that further promote globalization. They also require austerity measures that fall heavily on the poor, particularly women. Fifth, there have been shifts in the power of key institutions internationally. On the one hand, the influence of many national governments has been eroded by the rising importance of institutions like the MNCs, the IMF, and World Bank (WB), and trade organizations such as the World Trade Organization (WTO). On the other hand, there has been an increase in nongovernmental organizations (NGOs) advocating for the rights of groups of citizens (see Pyle 1999 for more information regarding each of these five trends).

Economic globalization has involved an increase in the international movements of goods and services, capital (portfolio investments or foreign direct investment by MNCs), and labor as people migrate for employment. This phase of globalization has been particularly facilitated by changes in technology (telecommunications and information technology) and transportation (UNDP 1999). It is often characterized as 'neoliberalism' because it is putatively based on increased reliance on markets (and correspondingly less on governments), liberalization of trade and investment policies, and a growing openness among economies. However, economic globalization has occurred in a very uneven manner. Countries are integrated into the global economy to very different degrees (UNDP 1999). This has resulted in rising inequality and tension, which are increasingly considered the flip side of the increased reliance on market forces and the changes in the international power structure. In the 1970s, scholars began to explore the gendered impacts of changes in the global economy. Gender has both a biological dimension (categorization as male or female) and socially constructed components. The latter reflect a society's views regarding appropriate roles for men and women and are reinforced by economic, political, social, cultural, and religious institutions. Global organizations such as MNCs and the IMF often use these social constructions to their advantage; however, globalization can undermine them or cause them to be more firmly defended. This article outlines the importance of the evolving study of the gendered impacts of economic globalization, the current issues regarding research, theory, and methodology, and likely future directions.

2. History and Changing Emphases

Before the 1970s, women or gender issues were rarely mentioned in the development literature. That changed with the publication of Boserup's (1970) book *Woman's Role in Economic Development*. Most regard this as a turning point that spurred thinking about women in the development process and initiated what has become a very large literature. The conceptual approaches taken and the strategies for change they imply have, however, changed over the years. Scholars and activists associated with the Women in Development (WID) approach of the 1970s sought to make women's roles visible. Women's issues typically were tacked on to existing approaches or institutions, however, rather than fundamentally integrated. WID focused on individuals and on fostering change within existing structures. Seeing these limitations, others developed a focus on gender. This approach emphasized unequal power relations between men and women and the ways they are institutionalized in households, culture, schools and religious organizations, the state, and international organizations. (See Kabeer 1994 and Visvanathan et al. 1997 for more information on this history; see Visvanathan et al. 1997 for selected readings.)

In the 1990s, the international political economy changed considerably with the ending of the Cold War and the impetus this provided to all the dimensions of globalization listed above. The discourse shifted to examination of the gendered impacts of economic globalization, broadening the focus to include gendered effects in all countries—industrialized and formerly socialist as well as those the development literature has concentrated on—and adding important dimensions to the analysis. For example, since globalization has involved increasing movements of people as well as goods, services, and capital, many industrialized countries have immigrants from developing countries who have been profoundly affected by these forces. This conceptualization includes them.

In spite of these substantive changes, and although gender increasingly is recognized as a relevant perspective, there continued to be a lack of attention to gender in the 1990s by most economists and many sociologists specializing in the study of development or globalization. Feminists continued to break new ground, however, in exploring these issues. Feminist sociologists have long critiqued approaches to development such as modernization, dependency, and world systems theories. (See Ward 1993 for an example regarding the world system approach.) Feminist economists have asked new types of questions, transformed some traditional economic approaches to include gender, and posed alternative approaches to development (Beneria 1995, 1999, Sen and Grown 1987, World Development 1995). Many now recognize the critical importance of understanding the gendered impact of these dimensions of economic globalization—and their impact, in turn, on growth and sustainable human development. For example, it has been found that low investment in women's education, often a result of SAPs, can hinder a country's economic growth (World Bank 2001).

These conceptual changes and the activism they spawned since the 1970s have fostered the rise of NGOs, many of which were founded to alleviate the effects of globalization and to address women's needs. International institutions have been forced to incorporate gender issues into their thinking, planning, and programs. The United Nations designated 1975 the International Year for Women. It held World Conferences on Women in 1975, 1980, 1985, and 1995, bringing people together from around the world to assess the status of women and draft policy recommendations to address gender inequalities. Special divisions have been established within the UN and the WB; considerable information and data have been collected (United Nations 2000). The United Nations Development Programme (UNDP) has emphasized human development throughout the 1990s; its annual *Human Development Report* focused on gender and human security in 1995 and globalization with a human face in 1999 (UNDP 1995, 1999).

There have been critiques of the efforts of these institutions to mainstream gender issues. In response, these organizations have developed more inclusive approaches, engaged experts in the field of gender and development as employees or consultants, and established substantive websites to disseminate information and invite discussion. The UN sponsored Beijing Plus 5 in 2000 to assess progress for women since the 1995 Conference on Women in China. The WB published a new report that focuses on the importance of gender for development and examines the factors that affect gender equality (WB 2001). Particularly active in advocating for women is the United Nations Development Fund for Women (UNIFEM), which grew out of the First World Conference on Women. UNIFEM Director Heyzer drafted an Agenda for Women in the twenty-first century (Heyzer 1995). UNIFEM's new volume (UNIFEM 2000) is designed to help women hold powerful institutions accountable for gendered outcomes of globalization—including the UN, IMF and WB, MNCs and other businesses, and civil society organizations such as unions, community-based groups, and business or professional associations.

3. Current Issues

Because economic globalization and gender is a broad topic with multiple dimensions, this section can only highlight some of the key issues scholars and activists are addressing regarding research, theory, and methodology. Newer work is beginning to cross disciplinary boundaries, offering opportunities to examine the

gendered effects of globalization from perspectives that may include economics, sociology, political science, gender studies, or anthropology. Increasingly, women from the Third World are shaping the questions asked, the type of research conducted, the methodologies used, the policy recommendations suggested, and the activist endeavors undertaken (Heyzer 1995, Kabeer 1994, Sen and Grown 1987).

As economic activities become more global, some benefit, but the hardships of many others are increased. There are many aspects that can be considered in assessing the differing effects of globalization on men's and women's income earning opportunities and household economic responsibilities—including access to employment in formal or informal sector jobs, wages and earnings (relative to the amount required for a reasonable standard of living), the types of occupations, job conditions (health and safety, rights to organize, work-life balance), opportunities for skill acquisition and promotion, and the effects of all these on one's position in the household and community. The analysis is complicated, however, because research reveals that the impact also varies by class, race/ethnicity, age, and other attributes. Ironically, although technological advances are a foundation of this phase of globalization, and even though technology offers a possible means for women to access information, develop broader networks of support, and increase their standards of living, women do not have the necessary access and are not receiving equal benefits of technological advances.

The internationalization of basic economic activities—production, trade, and finance—produces effects that differ by gender. The globalization of trade can have divergent effects on women. For example, if women work in sectors producing goods for international trade, they may benefit from economic globalization—as long as they make a living wage and their employer (often a MNC or its subcontractor) remains in their area. Many female small business owners are forced out of business by the cheaper imports that trade liberalization brings. Gender differences in the impact of trade typically have not been considered as trade organizations such as the WTO formulate policies.

Much production (clothing, electronics, toys, shoes, and sporting goods) and provision of services (data entry, reservations, business services) have moved internationally, becoming part of the 'global assembly line.' Many MNCs prefer lower cost women workers who they believe are unlikely to resist adverse conditions. Women working in MNCs or their subcontracting networks often consider this employment a better option than otherwise available. However, working conditions can be oppressive, characterized by long hours, fast pace, few breaks, no opportunities for advancement, harassment, and unsafe and unhealthy workplaces. In addition, such employment is precarious. When labor costs rise as workers seek to improve their conditions, MNCs use some combination of automating and new technologies, suppressing worker demands, moving to other developing countries with lower labor costs, or establishing subcontracting networks where workers are lower cost and can be terminated immediately. These strategies typically have adverse effects on women.

Financial globalization involves not only foreign direct investment by MNCs but also portfolio investment (buying/selling of stocks) and international loans and aid. As conditions for receiving loans, the IMF and WB typically require countries to adopt SAPs. These involve opening their borders to foreign trade and cutting government employment and expenditures on health, education, housing, and food subsidies. The IMF argues these policies will increase revenues needed for loan repayment. The effects of SAPs fall heavily on women, however, who try to maintain their families' standards of living. They take on added household responsibilities (to compensate for decreased government services) and obtain extra income earning activities, often in the informal sector.

As a result of these economic aspects of globalization, many women are increasingly pushed into the informal economy (Moghadam 1998, Pyle 1999). Poverty levels rise. Large numbers of women, particularly from lower-income Asian countries, are forced to migrate internationally to find employment, often leaving families behind. Women in many countries must resort to earning a living as domestics or sex workers. Conditions in these occupations can be highly exploitative.

As a result of background studies in these areas, scholars are developing a broader theoretical understanding of the systemic linkages among the global expansion of capitalist production, trade, and finance and the increases of women in the informal sector, rising inequality and female poverty, and higher levels of female migration (Pyle 1999). Feminist perspectives are being used to question and transform orthodox models of human behavior and, in turn, their implications for understanding the gendered impacts of globalization (Beneria 1999).

Research on economic globalization and gender involves problems, however, that include: gaining acceptance for new concepts that challenge and replace traditional views, operationalizing concepts empirically, obtaining appropriate data, developing alternative policies, and fundamentally reconsidering the usefulness of current approaches and institutions (whether they can be transformed or new ones are needed). For example, the reconceptualization of women's 'work' to include unpaid work in the household is controversial; there are conflicts about how to measure and value it. In addition, to adequately examine the gendered effects of globalization, data that measure the desired concepts and is comparable cross-country must be available. However, many countries have little data available; in others, in-

formation may be incomplete or of questionable quality.

There are the dilemmas involved in developing alternatives to policies like SAPs. Such innovative policies would enable countries to obtain loans but the focus would be on how to foster sustainable human development and gender equality rather than on how to ensure loan repayment. There is also the key question of what the focus of strategies for change should be—whether on changing the individual's position or the group's (and, if the group, at what level—family, community, nation, region, international) and whether to increase women's capabilities or confront the social, economic, and political structures that disadvantage or discriminate against females. This is particularly important to think carefully about, given the context of the larger shift in the distribution of power (from national governments to MNCs and international organizations like the IMF).

4. *Future Directions*

In the future, there will be continued progress in the creation of concepts and theories that enhance our understanding of the gendered impacts of globalization and the importance of these effects for sustainable human development. This work will be based on new ways of looking at issues, a wider array of studies of the gendered effects of globalization, and analyses that examine the more nuanced and complex patterns that emerge from such studies. This research may also involve the generation of new methodologies. As in the past, these concepts and theories will be important for developing strategies for change.

To illustrate each of these points briefly, scholars and activists will continue to develop and extend new concepts and perspectives. For example, concepts such as 'caring labor' will be acknowledged, respected, and utilized more widely (UNDP 1999). There will be increased attention to the gender dimensions of the insecurity created by globalization. Beneria is examining women's security globally, focusing on women's different employment problems, paid/unpaid labor share, female-headed households, and economic restructuring. By fully understanding the many aspects of insecurity, better policies can be developed to address them.

Newer methodologies may be developed based on these innovative conceptualizations and perspectives; existing methods will be extended and made more gender sensitive. For example, work will continue on the construction of broader indicators that measure women's well-being, economic security, and progress in participating more fully in economic, political, and social life.

Each of the issues outlined in the last section will be more fully studied at many levels of analysis—from the household level to the international. Multilevel analyses reveal more of the two-way linkages between the micro- and macrolevels. Not only do macroeconomic changes have gendered impacts on people and households at the microlevel, but microlevel processes contribute to macroeconomic well-being. For example, studies have shown that when women have access to income (a microlevel event often shaped by macrolevel forces), their children are healthier and better educated and, in turn, contribute more to sustainable development (WB 2001).

Case studies that explore the gendered effects of globalization on people who differ by culture, race/ethnicity, class, age, sector, and region as well as sex will shed considerable light on the role of these factors in shaping the gendered effects of economic globalization. This will provide insight into what broader socioeconomic changes may be needed for more gender equality as globalization continues.

Along with these advances, there will be continued fundamental challenges to disciplines that study development and globalization and their theoretical underpinnings (Beneria 1999). Disciplinary approaches will be recast to integrate gender perspectives; new interdisciplinary paradigms or areas of study may be created.

Women will continue to resist adverse effects of globalization and seek to improve their positions. Creative strategies will be crafted to enhance women's capabilities, empower them, and increase their socioeconomic security. These approaches will include access to and creative use of new technologies. NGOs will extend their work in these areas—not only locally or within a country, but increasingly at national, regional, or international levels. For example, as the gendered effects of globalization are more widely recognized, people in similar jobs or situations worldwide (e.g., sweatshops) may form alliances in order to address the problems they face.

There will be efforts to increase women's voices at all levels of government and develop ways to strengthen the state, an institution that can still be important for garnering the benefits and combating the problems of worldwide economic changes although its power has been eroded by forces such as MNCs and the IMF's structural adjustment policies. There may be continued changes in the level of awareness of gendered effects of globalization by international institutions (WB and IMF, UN and its many divisions, and the ILO) and in their interest and ability to design programs and policies to alleviate the adverse impacts. It is also possible that entirely new forms of global governance will be created that are more gender aware and focus on human development (health, education, safety, and opportunities for employment and income earning) rather than on finance and trade (UNDP 1999).

See also: Boserup, Ester (1910–99); Economic Development and Women; Gender and Feminist Studies in

Economics; Gender and Place; Gender, Economics of; Gender Ideology: Cross-cultural Aspects; Islam and Gender; Labor Markets, Labor Movements, and Gender in Developing Nations; Land Rights and Gender; Lone Mothers in Nations of the South; Nation-states, Nationalism, and Gender; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender

Bibliography

Beneria L 1995 Toward a greater integration of gender in economics. *World Development* (special issue) **23**: 1839–50

Beneria L 1999 Globalization, gender and the Davos Man. *Feminist Economics* **5**(3): 61–83

Boserup E 1970 *Woman's Role in Economic Development*. St. Martin's Press, New York

Cagatay N, Elson D, Grown C (eds.) 1995 Gender adjustment and macroeconomics. *World Development* **23**(11)

Heyzer N 1995 A women's development agenda for the 21st century. In: Heyzer N (ed.) *Commitment to the World's Women*. UNIFEM, New York

Kabeer N 1994 *Reversed Realities: Gender Hierarchies in Development Thought*. Verso, London

Moghadam V 1998 Gender and the global economy. In: Lorber J, Ferree M M, Hess B (eds.) *Revisioning Gender*. Sage, Thousand Oaks, CA

Pyle J 1999 Third world women and global restructuring. In: Chafetz J (ed.) *Handbook of the Sociology of Gender*. Kluwer, New York

Sen G, Grown C 1987 *Development, Crises, and Alternative Visions*. Monthly Review Press, New York

UNIFEM 2000 *Progress of the World's Women 2000*. UNIFEM, New York

United Nations 2000 *The World's Women 2000: Trends and Statistics*. United Nations, New York

United Nations Development Programme 1995 *Human Development Report 1995: Gender and Human Security*. Oxford University Press, New York

United Nations Development Programme 1999 *Human Development Report 1999: Globalization with a Human Face*. Oxford University Press, New York

Visvanathan N, Duggan L, Nisonoff L, Wiegersma N (eds.) 1997 *The Women, Gender and Development Reader*. Zed, London

Ward K 1993 Reconceptualizing world system theory to include women. In: England P (ed.) *Theory on Gender/Feminism on Theory*. Aldine, New York

World Bank 2001 *Engendering Development*. World Bank, Washington, DC

J. L. Pyle

Economic Growth: Measurement

Economic growth is the sustained increase in welfare of an economy—nation, region, city—together with the ongoing changes in that economy's industrial structure; public health, literacy, and demography; and distribution of income. In the long run, as this economic transformation evolves, so do social, political, and cultural norms. Societies change profoundly and multidimensionally, as economic performance improves.

To measure economic growth is to quantify this increase in welfare and to endow with numerical precision these large-scale economic and social changes. Given the breadth of possibilities, it is impossible to undertake this measurement exercise without guidance—what can be pared away, what is essential—from *some* view on the causes of growth (see, e.g., *Economic Growth: Theory*).

This article sets down some key (measurement) facts concerning economic growth, and documents how they have evolved, if at all, over time. In doing this, the article attempts also to illustrate the historical interplay between two lines of research, measurement of and theories about economic growth, each influencing the other.

1. National Income

The panorama above of profound social and economic changes can be simplified dramatically by concentrating on just a single key economic variable, income per capita. (We will return in Sect. 8 below to issues of broader structural transformations).

Income per capita is the per head measure of the total value of all goods and services produced in an economy. Taking national income—measured by either gross national product (GNP) or gross domestic product (GDP), or its regional counterpart—and dividing it by population in the appropriate nation or region gives a convenient first measure on the state of economic well-being. Since total income is the same as total output, this measure might sometimes be usefully replaced by output per worker, or *labor productivity*, where the denominator is then the size of the labor force; or, even output per worker-hour, where the measure then takes into account the time spent working by the labor force.

In some detailed analyses, these alternatives can provide different useful insights into economic performance—different countries, at different times, have had their labor force markedly different from their population, or have had workers and firms make different choices on the length of their workday. However, for the kind of long-horizon, large-scale developments that are typically of interest in economic growth, these differences are inessential. Potentially more important is whether this one measure can suitably proxy for the wide spectrum of different variables of concern in economic growth.

Across countries, per capita income is positively correlated with a broad range of alternative indicators for economic performance—including life expectancy, (the negative of) infant mortality, and adult literacy

(Ray 1998). Thus, economists and many other social scientists have focused, primarily although not exclusively, on growth in per capita income as the preferred first measure of economic growth more generally.

One obvious and crucial shortcoming in this, however, is that income per capita, by definition, says nothing about income inequality. Thus, a potential tension between economic growth and income inequality arises, on top of an analytically-driven hypothesis that rising average incomes might be associated with a widening of the income distribution. We will return to this below in Sect. 7.

2. Early Work

Economic growth—significant, sustained increases in welfare—is a relatively recent occurrence (Maddison 1982, Romer 1986). While steady improvements in the lot of humanity over the last 50,000 years had been noted, most of those came tortuously slowly.

The Netherlands began the eighteenth century as the leading economy in the world. But economic growth here was neither sustained nor pronounced. Almost a century afterwards, labor productivity in the Netherlands was 6 percent *lower* than in 1700.

In 1785 the UK had risen to become the new world leading economy. However, despite immense strides in technological progress due to the Industrial Revolution, the UK's labor productivity grew by only 0.5 percent per year from 1785 through 1820. From then through the beginning of the twentieth century, the UK remained the world's economic powerhouse. But in this time UK labor productivity grew by only 1.4 percent per year. Only after the US overtook the UK following the end of the nineteenth century, did the world begin to see the kind of long-run, sustained growth rates that are now typical: US labor productivity grew at 2.3 percent from 1890 through 1979.

3. Kaldor's Stylized Facts

The concept of GNP was only developed in the early twentieth century, not for measuring economic growth, but for assessing the state of a national economy. Not long after, Kaldor (1961) observed that the ongoing growth in per capita national income came with several empirical regularities:

(a) Income per capita grows steadily over long stretches of time;

(b) Factor income shares—the fractions of total income accruing to the labor force and to physical capital—show no secular trend;

(c) The ratio of total national income to the aggregate stock of physical capital—the total value of machines, highways, bridges, buildings, physical infrastructure—remains roughly constant.

These observations, typically labeled *Kaldor's stylized facts*, constitute a first important growth measurement that economists sought to understand and to explain.

According to Kaldor's measurements, while economic growth occurs, its benefits divide up in a stable way between rewards to capital and labor—no one factor input, taken as an aggregate, benefited more than another from economic growth. Moreover, since capital's income share can be viewed as the stock of capital multiplied by capital's rate of profit, Kaldor's stylized facts also implied that the profit rate is, in the long run, constant.

4. Physical Capital Accumulation

The most successful of the different explanations for Kaldor's growth observations was given in Solow (1956). This immensely influential work produced a widespread conventional wisdom that was simultaneously a profoundly important policy implication. The line of work, however, then led to yet further measurements (Solow 1957) whose later reinterpretation would eventually overturn the original consensus.

Conventional wisdom following Solow (1956) had investment—the accumulation of physical capital—constituting the critical engine of growth. Solow (1956) assumed that total output obeyed a *production function* taking values that increased with the quantities of physical capital and labor employed in the economy. Because no well-behaved function can give at all times a perfect fit relating output, capital, and labor, Solow (1956) hypothesized further an unobservable residual to the production function relation, a residual that he interpreted as exogenously evolving *technical progress* or, more simply, technology. Together with some other reasonable assumptions on the behavior of savings (and therefore investment), labor force growth, and technology's impact on the effectiveness of labor and capital, the model generated outcomes successfully reproducing Kaldor's stylized facts.

The result envisioned had economic growth in two conceptually distinct phases. First, there is the long-run steady state, where the capital-output ratio is constant, and growth in per capita income occurs entirely from technical progress. Second, there is transition to the steady state: along this transition path, physical capital accumulates rapidly due to relatively high investment, driving up the economy's growth rate as the economy itself converges towards its long-run, steady-state growth path. If the second phase were that more relevant to the real world, as many supposed, then accumulating physical capital—building machines, factories, roads, and bridges—constituted the engine of economic growth. Policy-

makers from Joseph Stalin in the Soviet Union through Lee Kuan Yew in Singapore through leaders in developing country after developing country took this implicit advice to form a cornerstone of their development plans. Encouraging savings—voluntary or forced—became key to economic progress.

In Solow's work, technology bore two different meanings. One, because by assumption it influenced the effectiveness of factor inputs, technology was also labeled *total factor productivity* (TFP)—it improved the productivity of all the factor inputs in an economy. Two, because it emerged as a residual in the production function accounting exercise, it was also viewed, derisively, as a 'measure of our ignorance' (on the part of the modeller). Early research on growth measurement, therefore, took one of its goals to be the minimization of this residual, the idea being that the smaller TFP became, the more successfully one had explained economic growth.

Measurements undertaken by Solow (1957) showed that for the US over the first half of the twentieth century, almost 90 percent of economic growth was due to TFP, with only the remaining one-tenth accounted for by capital accumulation. Taken at face value, this surprising empirical finding sharply contradicted the conventional wisdom and policy advice many had taken to follow from Solow (1956). The hoped-for reconciliation was that by measuring accurately enough the factor inputs—accounting for improvement in quality of capital and labor—this Solow residual would be diminished.

By contrast, more modern work (Sect. 6 below) favors the first interpretation of TFP, that is, that it is correctly viewed as enhancing the productive potential of an economy. Thus, economies with large TFP are considered high-growth, high-performance economies. The research focus has shifted from trying to minimize measured TFP to trying to understand its evolution, typically in terms of its being driven by science and research and development (R&D).

Another important idea arose from this early work on measuring economic growth: convergence, or the tendency for economies not already on the steady state path to approach it. Stated more fully, the Solow (1956) model implies that capitalist economies where capital and labor are rewarded their *marginal products*—the contributions each make, on the margin, to final output—are economies that are stable around their steady states. If an economy is not already at steady state, it has a tendency to go there. Economies relatively poor have a tendency to converge to, or catch up with, those relatively rich.

5. *Rich and Poor*

Difficult as it initially was to compile systematic national income accounts for even the advanced countries, obtaining comprehensive and comparable data across countries was an order of magnitude more complex. This task has only recently been accomplished (Summers and Heston 1991) to a degree where researchers can now meaningfully compare incomes and therefore economic growth across countries.

Because different countries collect and report income data in their individual, respective currencies, cross-country comparisons can be made only after converting national incomes to a common basis. To achieve this, one possibility is to use official current exchange rates to convert national–currency figures into numbers denominated in some single currency. However, such exchange rates are highly volatile, and need not always accurately reflect the purchasing power, or the command over real resources, of the respective national currencies. Importantly, currency exchange rates cannot reflect the prices of nontradeable goods and services. To the extent that nontradeables are relatively lower-priced in poorer countries, conversion using official exchange rates *underestimates* the incomes of relatively poor countries, and thus exacerbates the true gap between rich and poor.

As just one example, consider the following. Even though between 1985 and 1990, Asia was the world's fastest growing region, its share of world GDP—calculated using official exchange rates—fell from 8 percent to 7 percent. The reason for this is that Asian currencies depreciated sharply against the US dollar in this 5-year period. In this instance using exchange-rate adjusted national incomes disguised (and indeed misled on) actual changes in economic performance.

Summers and Heston (1991), in collaboration with the United Nations, constructed rescaled national incomes for a broad cross-section of countries. They used a *purchasing power parity* (PPP) correction to overcome the measurement difficulty just described. Taking into account prices on close to 700 individual items—traded and nontraded—across different countries, and then adjusting for differential spending patterns, the Summers–Heston data provide the most reliable indicators to date, allowing comparison of economic growth across countries. These data now constitute the professional standard, and have been used in over 1,000 empirical studies on economic growth.

Allowing cross-country comparisons has opened up striking new questions in both the measurement and theory of economic growth. Using 1993 official exchange rates, Switzerland was the world's richest economy; Tanzania, the world's poorest. The ratio of per capita incomes between them approached an incredible 400. By contrast, when PPP-corrected and averaged across the last five years of the 1980s to smooth out short-run fluctuations, the ratio between richest and poorest economies (now the US and Chad, respectively) falls to a little over 40. This latter number accords better with less formal evidence and with casual observation on differing costs of living. However, even though PPP correction has reduced the

dramatic disparity between rich and poor across the world, the divide obviously remains considerable.

The Summers–Heston data reveal the following: between 1960 and 1989 world per capital income grew by 2.25 percent per year. Averaged over the five years 1960–64, per capita income at the bottom 10th-percentile point (in the cross-section of countries) was 0.22 times the world average, with the 90th-percentile point 2.70 times the world average (Quah 2001). The 90:10 ratio was, therefore, 12. Almost three decades later, averaged over 1985–89, per capita income at the bottom 10th-percentile point had fallen to 0.15 times the world average while the 90th-percentile point had risen to 3.08 times the world average. The 90:10 ratio thus nearly doubled over this period of high average growth worldwide. More detailed study of the growth dynamics in this post-Second World War cross-section of countries has revealed, not just a widening of income disparities across rich and poor, but also an emerging polarization, with the middle-income group of countries dwindling in number (Quah 1997a).

This finding, that in the post-War era, poor economies do not obviously catch up to the rich, has its interpretation disputed by research arguing that, provided one conditions on the appropriate indicator variables—investment and schooling, in particular—then convergence does occur. That kind of convergence, however, is no longer between rich and poor, but is convergence of each economy towards its own specific steady state growth path (Barro and Sala-i-Martin 1992, Levine and Renelt 1992). Hundreds of auxiliary variables have been suggested in these cross-country growth and convergence studies as appropriate alternative or additional conditioning variables (see, e.g., the compilation of studies in Durlauf and Quah 1999).

The evidence in Durlauf and Johnson (1995) shows that poor economies do indeed tend to converge to relatively low income levels, and rich ones to high, even after conditioning on savings and literacy rates. This study suggests, therefore, that as economic growth proceeds, then even with conditioning, the polarization into rich and poor described in Quah (1997a) remains empirically relevant.

6. *Ideas—Science and Technology*

Observing that country disparities are large and growing, that is, that convergence is not happening; and that so much of economic growth is due to TFP has motivated some researchers to investigate models of *endogenous technological change*. (One notable exception is Mankiw et al. (1992) who argued that taking into account *embodied human capital*—the quantity of schooling embodied in the labor force—suffices to explain both observations.)

In work on endogenous technology, TFP is no longer taken to be exogenously given. Instead, economic theory is used to shed light on the incentives and forces shaping TFP's evolution (Aghion and Howitt 1998, Grossman and Helpman 1991, Romer 1990). The problem of explaining cross-country disparities then becomes one of understanding the dynamics of cross-country technology dissemination (Coe and Helpman 1995, Helpman 1993). This work predicts that knowledge inputs—scientists and engineers, R&D, patents, knowledge spillovers, ideas—drive economic growth. The key empirical findings are twofold: First, significant spillovers occur across countries, so that worldwide knowledge accumulation drives worldwide economic growth. Second, country-specific knowledge stocks, resulting from that country's own R&D, do matter importantly, and help explain each country's pattern of economic growth.

The hypothesis of knowledge-driven economic growth, while suggested by the logical sequence of earlier measurements and theoretical developments, remains controversial. For one, many observers consider it implausible that the poorest economies—those in sub-Saharan Africa, say, in the latter part of the twentieth century—are poor because they do not do enough scientific research. This, however, is an overly-literal interpretation of knowledge inputs. The ability of people in different economies to absorb and economically exploit scientific developments or even just best-practice production techniques obviously differ. The question then becomes one of incentives to use ideas—why do the very poorest economies not use frontier technologies to raise the productivity of factor inputs? The economics of creating and disseminating ideas, or *intellectual property* more broadly, are subtle; so too, but much less-studied, the economics of using ideas.

However, even in the richest economies, empirical evidence is ambiguous on knowledge-driven economic growth. US scientists and engineers employed in R&D grew five-fold from 200,000 in 1950 to over 1 million in 1990; the growth rate of US GDP failed to increase by anything remotely comparable over this period (Jones 1995).

7. *Inequality*

As measuring economic growth has continued to focus on progressively fewer observable variables, an implicit tension has correspondingly sharpened: if welfare is proxied only by average income, then income inequality—or more generally, the distribution of income across people—might be unobserved worsening to where an average seriously misleads as an indicator for the economic welfare of the population overall.

The Kuznets (1955) curve describes a key empirical regularity between growth and inequality: Cross-sectionally across economies, inequality is low when per capita incomes are low, increases at moderately higher per capita income levels, and finally falls again at the very highest levels of per capita income. Put

differently, the relation between inequality and levels of development is an inverted U. This early finding suggested to many researchers the possibility that, indeed, economic growth might disadvantage the very poor, due to increases in inequality.

More recent empirical analyses (Li et al. 1998) have shown, however, that as economies grow, the change in inequality is small: most of the measured variation in inequality is that *across* countries. Over time, inequality can change but much, much more slowly than economic growth varies.

8. Structural Drift

The more recent focus on measuring economic growth using just changes in per capita income has de-emphasized the evolution in industrial structure of economies as growth proceeds. By contrast, research such as that in Kuznets (1966) consistently documented such basic changes. Quah (1997b) has noted that such structural drift continues, away from manufacturing, towards high value-added intangible services.

Over the nineteenth century, as economies grew, their industrial composition shifted away from an emphasis on rural agriculture, and towards manufacturing and urbanization. Economic activity became progressively more markets-based, so that traditional informal, nonmarket work at home was replaced by formal employee arrangements. Education became more institutionalized; foreign commerce increased; and natural resources were relied on progressively less.

9. Where Next?

Measuring economic growth is an intricate and subtle exercise. It has a scope that can be large and unmanageable. In practice, the history of measuring economic growth has been one of intimate interplay with economic theorizing, leading from improvements in the accuracy and relevance of growth data, to hypotheses on the critical sources of economic growth, to explaining deep changes in the structure of national economies.

Total factor productivity, previously a measurement residual to be minimized, has since been reinterpreted, and has emerged as the key driver for economic growth. Ideas—scientific and engineering knowledge, R&D, expertise—constitute the prime candidates for explaining TFP, but the supporting empirical evidence remains weak.

See also: Development, Economics of; Development: Social; Development: Social-anthropological Aspects; Economic History, Quantitative: United States; Modernization, Sociological Theories of; Time Series: General

Bibliography

Aghion P, Howitt P 1998 *Endogenous Growth Theory*. MIT Press, Cambridge, MA

Barro R J, Sala-i-Martin X 1992 Convergence. *Journal of Political Economy* **100**(2): 223–51

Barro R J, Sala-i-Martin X 1995 *Economic Growth*. McGraw-Hill, New York

Coe D T, Helpman E 1995 International R&D spillovers. *European Economic Review* **39**(5): 859–87

Durlauf S N, Johnson P A 1995 Multiple regimes and cross-country growth behavior. *Journal of Applied Econometrics* **10**(4): 365–84

Durlauf S N, Quah D 1999 The new empirics of economic growth. In: Taylor J B, Woodford M (eds.) *Handbook of Macroeconomics*, Vol. 1A. North Holland Elsevier Science, Amsterdam, Chap. 4, pp. 231–304

Grossman G M, Helpman E 1991 *Innovation and Growth in the Global Economy*. MIT Press, Cambridge, MA

Helpman E 1993 Innovation, imitation, and intellectual property rights. *Econometrica* **61**(6): 1247–80

Jones C I 1995 R&D-based models of economic growth. *Journal of Political Economy* **103**(3): 759–84

Jones C I 1998 *Introduction to Economic Growth*. Norton, New York

Kaldor N 1961 Capital accumulation and economic growth. In: Lutz F A, Hague D C (eds.) *The Theory of Capital: Proceedings of a Conference Held by the International Economic Association*. Macmillan, London, Chap. 10, pp. 177–222

Kuznets S 1955 Economic growth and income inequality. *American Economic Review* **45**(1): 1–28

Kuznets S 1966 *Modern Economic Growth: Rate, Structure, and Spread*. Yale University Press, New Haven, CT

Levine R, Renelt D 1992 A sensitivity analysis of cross-country growth regressions. *American Economic Review* **82**(4): 942–63

Li H, Squire L, Zou H 1998 Explaining international and intertemporal variations in income inequality. *Economic Journal* **108**: 1–18

Maddison A 1982 *Phases of Capitalist Development*. Oxford University Press, New York

Mankiw N, Romer D, Weil D N 1992 A contribution to the empirics of economic growth. *Quarterly Journal of Economics* **107**(2): 407–37

Quah D 1997a Empirics for growth and distribution: Polarization, stratification, and convergence clubs. *Journal of Economic Growth* **2**(1): 27–59

Quah D 1997b Increasingly weightless economies. *Bank of England Quarterly Bulletin* **37**(1): 49–56

Quah D 2001 Cross-country growth comparison: Theory to empirics. In: Dreze J (ed.) *Advances in Macroeconomic Theory, Vol. 1 of Proceedings of the Twelfth World Congress of the International Economic Association, Buenos Aires*. Macmillan, London, Chap. 16

Ray D 1998 *Development Economics*. Princeton University Press, Princeton, NJ

Romer P M 1986 Increasing returns and long-run growth. *Journal of Political Economy* **94**(5): 1002–37

Romer P M 1990 Endogenous technological change. *Journal of Political Economy* **98**(5, part 2): S71–S102
Solow R M 1956 A contribution to the theory of economic growth. *Quarterly Journal of Economics* **70**(1): 65–94
Solow R M 1957 Technical change and the aggregate production function. *Review of Economics and Statistics* **39**(3): 312–20
Summers R, Heston A 1991 The Penn World Table (Mark 5): An expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics* **106**(2): 327–68

D. Quah

Economic Growth: Theory

Growth theory was developed in the 1950s and 1960s, and some early growth models are collected in Stiglitz and Uzawa (1969). After a lull in the 1970s, the arrival of better data in the 1980s prompted a burst of activity during which economists refined their models so that they could better answer question like: Why did growth take off two hundred years ago and why in England and then the West, and why do Africa, Latin America, and most of Asia lag so far behind? And, what measures will improve things, especially in the poor countries? It is mostly the new crop of models that will be surveyed here.

1. *Growth*

What determines the rate at which our standard of living grows? Economic theory has settled on three broad categories:

(a) the progress of science and productive knowledge;
(b) the growth of individual skills; and
(c) incentives.

Incentives impinge on every economic decision, of course, including the decision to accumulate knowledge and skill. Still, (a) and (b) are listed separately because science and skills have evolved, in part, spontaneously or accidentally as people went about their business or pursued their curiosity.

1.1 *Science*

The world's wealth grows thanks mainly to the growth of knowledge. As with anything else, when adding to knowledge, there is strength in numbers. Even before the fifth century BC and the beginnings of science, early stone tools, farming, the domestication of animals, the wheel, and other inventions spread not by parallel and independent invention, but by imitation. If I was alone in the world, I would have to do my own inventing. But if there are two of us and you have invented the wheel already, I don't need to reinvent it, I can just copy your design. In copying you, moreover, I do not deprive you of your wheel—we both can make a wheel even though you alone came up with the idea. Thus, knowledge is a form of social capital that grows when anyone adds to it, and a bigger world will grow faster simply because more people will think up more new ideas. From the first stone tools to the steam-engine, electricity, and the computer, the mass of the people who derived some use from new ideas have had no part in coming up with them. So, if the overriding constraint on growth is scientific progress, a more populous world is one in which living standards will rise at a faster rate.

Arrow (1962) argued that we get new ideas when we use our old ideas, and that invention is incidental to normal production activity. He called this process 'learning by doing' and he assumed that the process operates in the machine-producing industry: The more machines we make, the better at it we get. The learning process operates at the level of the industry as a whole in that each producer learns from the experience of all other producers.

Because it relies exclusively on learning by doing, however, Arrow's growth mechanism contains only a small part of the engine that drives modern growth. After all, how many of our big inventions were made fortuitously at the workplace? Even when it is an accident—penicillin and the Post-it note, for example—a discovery usually occurs, instead, in the research lab. Moore's law, the observation that the speed of computers doubles every two years, is driven by research on the computer chip, and not by the mere fact that the world produces a lot of chips. Of course, if we made a greater number of chips, we would probably also be choosing to do more research to make them faster; the rate of production and the rate of growth of chip-quality would, in this sense, be positively related, but it would still be the research scientist and not the production worker whom we would have to thank for that quality growth.

Romer (1990) recognizes the importance of research. In his model, the more researchers we have, the faster we can invent, and the faster we grow. As in Arrow's model, then, here too a bigger economy grows faster. This positive effect that the scale of the world has on growth holds up well historically. Kremer (1993) finds that the rise in the world's population has, at least until around 1900, gone hand in hand with a rise in the rate of growth of income per head. The industrial revolution thus took place only when people came up with enough ideas—e.g., the steam engine—that then kicked off the industrial era.

1.2 *Skills*

Science cannot be all there is to growth and development. For one thing, large parts of the world

seem almost untouched by the many scientific inventions of the past century—many people are still illiterate and still live without electricity and running water. Moreover, the world is becoming more populous, and its cadre of scientists is growing by leaps and bounds, yet since the Second World War, income per head has grown more or less steadily at about 1.5 percent per year. This, too, Jones (1995) tells us, is not consistent with science-driven models which predict that a bigger world should grow faster as long as the well of ideas has not run dry. But there is not much evidence that the supply of new ideas is dwindling—on the contrary, computer-aided simulation has made research easier.

The constraining force on growth seems, instead, to now be the cost of adopting new ideas—the fact that to put new ideas into practice requires resources and skill. True, some technologies are so user friendly that their use demands no skill at all; to use a light bulb, for instance, all you have to do is screw it in. But you cannot do much with a computer if you have no education and no experience with computers. And, since the costs of technological adoption are likely to be proportional to the number of users, adoption is as costly in a big world as in a small one. Society inevitably spends much less on research and invention than its does on implementing what it has invented.

Even in the USA, research effort absorbs a mere twentieth of what is spent on educating and training the workforce. And in a developing country that imports all of its technology, the constraint on growth is purely the speed with which its citizens can enhance their skills and get the requisite machines. Lucas (1988) models such a world, a world in which countries share the same technology but not the skills of their workers. In his model, people augment their skills everywhere at the same rate, and if labor cannot move from one country to another, differences in skill-levels and, hence, in development persist indefinitely. Because education and work-experience differ so much among countries, his model explains differences of development among regions and countries. On the other hand, an Arrow–Romer type of model can explain the growth history of the world as a whole.

1.3 Incentives

We now come to incentives. Incentives to invent things, and to implement inventions, perhaps by acquiring the needed skills, or by investing in something else. It is impossible, for example, to explain European stagnation during the Dark Ages without mentioning the barriers that authority placed on intellectual progress. Baumol (1990) and Mokyr (1992) argue that, historically, material progress has depended less on the state of science and the supply of potential innovations that on whether people have had the incentive to implement them:

> The museum at Alexandria was the center of technological innovation in the Roman Empire. By the first century BC, that city knew of virtually every form of machine gearing that is used today, including a working steam engine. But these seemed to be used only to make what amounted to elaborate toys. The steam engine was used to open and close the doors of a temple (Baumol 1990 p. 910).

In other words, good ideas were plentiful two thousand years ago, but the environment for using them productively was unfriendly. The Roman Empire, Europe in the Dark Ages, China in the Middle Ages, and various nations today, have all been hostile to new ideas.

Hostility to technological change can show up in a variety of ways. Parente and Prescott (2000) show, for example, that established monopoly rights over an old technology can allow it to survive the threat of entry by firms wanting to use a better new one, and that this can have a big drag on development. The real challenge, however, is to understand how the policies are chosen. Krusell and Rios-Rull (1996) show that even a democracy may well choose to erect a barrier to technological change, because the majority of the voters may stand to lose from the adoption of a new technology as it would devalue their skills.

The incentive to save has also varied a lot historically and it varies now from one place to another because of policy differences. However, growth theory predicts only modest effects of frugality—a permanent rise in national saving has only a temporary effect on income growth. How long the stimulus lasts depends on how much physical capital matters in production. The evidence on whether frugality leads to even a temporary rise in growth is mixed (Barro and Sala-i-Martin 1995).

2. Inequality

Growth theory also aims to explain inequality. The world has gotten richer, but inequality in living standards is greater than ever. The poverty that most Africans, Asians, and Latin Americans now find themselves in isn't just relative—it isn't just that living standards in the poor world seem low because they are dwarfed by the great wealth in the USA, Western Europe, and Japan. On the contrary, living standards in the poor world are low by any historical standard. People there often face starvation and epidemic disease, and they do not live much better than their distant ancestors did. At the end of the twentieth century, US income per head is only six times what it was at the start of the century, yet it is 30 times as high incomes in Chad or Ethiopia. This means that people in Chad and Ethiopia are much worse off than Americans were a century ago. Why is the gap so wide today and why aren't the poor countries catching up?

2.1 The Causes of Inequality

Although one can always ask for a still prior cause, the following five are causes of inequality that seem, to the author, fundamental:

2.1.1 Geography. Some points on the globe are natural sites for economic activity—e.g., the estuary of a large river. Such advantages of location are compounded by economies of agglomeration such as the cost savings afforded by the sharing of cable lines, roads, railways, parks, parking lots, airports, etc. In cities, one can more easily share knowledge as well, and, as Lucas (2001) shows, this can explain why rents are so high in city centers. These economies of agglomeration draw economic activity to some points in geographical space and not others, and output per square foot of land varies a lot from one place to another.

2.1.2 Specialization. Economic efficiency dictates that people who start out as equals should sometimes end up different. One reason is specialization—we cannot all be doctors and university professors, someone must clean the streets, wait tables, etc. Lucas (1988, Sect. 5) explains how a natural advantage in specializing can backfire in the long run. He describes a world in which an initial comparative advantage in farming can cause a region to become a food producer; growth potential may, however, lie not in farming but in industrial goods, goods that people living in regions that do not have good farmland will turn to. Such people will eventually become expert manufacturers, whereas farmers will in the long run lag far behind because they specialized in a good with no growth potential.

2.1.3 External effects and market failure. Backwardness offers a society the option to 'free ride' on the efforts of the leaders. Eeckhout and Jovanovic (1998) argue that the option to imitate the ideas of others can lead to a stable situation in which some will prefer the role of a perpetual follower because this allows him to copy their inventions at a fraction of their original cost. Similarly, not everyone can always use cutting edge equipment. Jovanovic (1998) argues that firms in poor regions will use a technology and equipment that embodies it after the world's leaders are done with it and after it has become cheaper—an external benefit that the leader bestows on the follower indirectly, through prices. A country can be poor because its internal markets provide incorrect signals and intervention can then improve things. Murphy and co-workers (1989) model a situation where technological development is unprofitable for any single firm unless enough firms decide to go ahead and invest simultaneously—the 'big push' idea. Similar market failures occur in the models of Azariadis and Drazen (1990), and Galor and Zeira (1993).

2.1.4 Policies. The overriding cause of poverty in the world—and surely the one that dwarfs every other—is bad or corrupt government. The world's governments vary from democracies at the one end of the spectrum to totalitarian governments—right-wing, left-wing, and theocratic—at the other. As a result, people in different countries face vastly different policies, institutions, and laws and, hence, incentives. A policy that is bad for the average citizen may persist if it benefits or protects a minority that has somehow managed to gain power and influence. Corruption and bad policies repel foreign capital and cause domestic capital to flee. The burden falls on the average citizen who has to pay high taxes, tariffs, license fees, bribes, etc. It is hard to imagine, for example, how Burma can develop while its citizens are shut off from contact with the outside world.

3. How to Reduce World Inequality

The conflict between equality and efficiency, familiar in discussions of personal income differentials, is present at the world level as well. Some of the world's inequality is unavoidable and, indeed, simply signals that markets work. Nothing much can or should be done about inequality over geographical regions *per se*, or inequality that represents a reward for past sacrifices such as a high rate of national saving, investments in education, and so on.

An enlightened government can set taxes and subsidies so as to correct for market failures. Capitalism may seem to be at a loss in such circumstances, but perhaps not: in the late nineteenth century, a handful of bankers coordinated the actions of a big chunk of American business and the railroads in particular—with anticompetitive side effects to be sure—and seem to have engineered the big push that propelled the USA into industrial leadership. Later on, the US government started to regulate the banks and American business in general.

Whatever good governments may have done domestically, it is quite evident that on the international front government policies have done great harm. Governments have set policies that limit the flow of people, capital, and ideas. If so, the ultimate culprit for a country's poverty is the national boundary that keeps the citizens hostage to the policies of their rulers. Once this much is clear, several solutions to the problem of world inequality spring to mind at once.

3.1 Aid

One can help the poor world by taxing the rich countries, and, since the poor hold a majority at the

moment, a world democracy would, presumably, do just this. But under today's political system, we rely, instead, on voluntary contributions—aid—which, even when they occur, do not lead to improved government policies, and provide only a temporary fix.

Another solution is to bribe the offending politicians into give up power, or at least changing their policies. Such bribery has, in fact, been going on for decades, under the name of 'conditional lending'—a country gets a loan if its government agrees to adopt a pro-trade policy, etc. This is better than unconditional aid, but it, too, has not worked because the 'conditions' of loans cannot be enforced as funds are diverted from their intended uses and into private bank accounts.

3.2 Freer Migration of Labor

A better solution is to let people vote with their feet more freely than they can at present. As in Tiebout (1956) a country that adopts bad policies would find its constituents leaving for greener pastures. The lesson of the European Union is that when the boundaries come down, policies must harmonize and economic welfare converges. In spite of its size, the unified US labor market has far less income inequality than the world does, and immigrants manage to catch up to the US average after just a couple of generations. Right now, the rich countries limit immigration to a trickle so as to avoid a stampede that could prove disruptive, but immigration rates can probably be higher than they are today and still not endanger the fabric of life in the rich world. The rich countries are now unifying their markets (e.g., NAFTA and the EU), but markets in the poor world—the former Eastern Bloc, Africa, and Asia—are, for the time being, left out.

4. Where We Are Headed

If we can manage to avoid another global conflict, the political organization of our world will probably ease into a federal structure in which each nation will retain much of its sovereignty, but in which its government will set its policies in the face of free labor mobility. This Tiebout type of a world organization will produce far less inequality than we see today, as each government will have to adopt policies that make its jurisdiction an attractive place to work and live.

Without centralized intervention, however, the world's rate of growth would be less than optimal. Each local government would not, on its own accord, take enough trouble to save the ozone layer, the rainforests, and the endangered species, and, in particular, it would not do enough to promote activities that push the frontiers of knowledge out—research and development, education, etc.—or enough to protect intellectual property. A global market failure involving these activities will need correcting by policies set at the world-government level. A research subsidy will raise the world's growth rate, but it will also raise inequality because research will be done in the skill-abundant parts of the world, which also is where the highest-paid workers will live. The world federation that we are headed for will thus still face the perennial tension between efficiency and equity—it will still not be possible to have the largest pie and to distribute it equally. But the sort of inequality that we are now so used to will be a thing of the past.

See also: Business History; Development, Economics of; Development: Social; Development: Social-anthropological Aspects; Economic Growth: Measurement; Economic History; Science and Development; Technological Innovation

Bibliography

Arrow K 1962 The economic implications of learning by doing. *Review of Economic Studies* **29**(3): 155–73

Azariadis C, Drazen A 1990 Threshold externalities in economic development. *Quarterly Journal of Economics* **105**: 501–26

Barro R J, Sala-i-Martin X 1995 *Economic Growth*. McGraw-Hill, New York

Baumol W 1990 Entrepreneurship: Productive, unproductive, and destructive. *Journal of Political Economy* **98**(5): 893–921

Eeckhout J, Jovanovic B 1998 *Inequality*. New York University, New York

Galor O, Zeira J 1993 Income distribution and macroeconomics. *Review of Economics Studies* **60**: 35–52

Jones C I 1995 Time-series tests of endogenous growth-models. *Quarterly Journal of Economics* **110**(2): 495–525

Jovanovic B 1998 Vintage capital and inequality. *Review of Economic Dynamics* **1**(2): 497–530

Kremer M 1993 Population growth and technological change: One million B.C. to 1990. *Quarterly Journal of Economics* **108**: 681–716

Krusell P, Rios-Rull J-V 1996 Vested interests in a positive theory of stagnation of growth. *Review of Economic Studies* **63**: 301–29

Lucas R E 1988 On the mechanics of economic development. *Journal of Monetary Economics* **22**(1): 3–42

Lucas R E Jr 2000 Externalities and cities. *Review of Economic Dynamics* **4**(2): 245–74

Mokyr J 1992 Technological inertia in economic history. *Journal of Economic History* **52**(2): 325–38

Murphy K, Shleifer A M, Vishny R W 1989 Industrialization and the big push. *Journal of Political Economy* **97**(5): 1003–26

Nelson R, Phelps E 1966 Investment in humans, technological diffusion, and economic growth. *American Economic Review* **56**(1/2): 69–75

Parente S L, Prescott E C 2000 *Barriers to Riches*. MIT Press, Cambridge, MA

Romer P 1990 Endogenous technological change. *Journal of Political Economy* **98**(5): S71–102

Stiglitz J, Uzawa H 1969 *Readings in the Modern Theory of Economic Growth*. MIT Press, Cambridge, MA

Tiebout C 1956 A pure theory of local expenditures. *Journal of Political Economy* **64**(5): 416–24

B. Jovanovic

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Economic History

1. The Pioneer Period (1776–1914)

It is never easy to track the birthday of an intellectual discipline. Like many children of Clio, economic history was taking shape even before it was born. Politicians, literary men, philosophers, theologians, artists, and scientists have always used economic phenomena of the past to illustrate or elucidate their theories on private property, unequal distribution of wealth, interest and usury, the labor theory of value, and the just price. Merchants, lawyers, businessmen, and other 'practitioners' (as Paul Samuelson labeled these authors) in their writings frequently turned to the past, searching for similar experiences or comparative data that could be applied in their daily work. Later, the rapid growth of international commerce and the rise of powerful merchant states after the Middle Ages inspired the leading writers of the Mercantilist Era to develop some ideas on the history of trade, the balance of trade, monetary flows, exchange rates, and state intervention. Carlo M. Cipolla has quite rightly observed that if economics was preceded by 'protoeconomics' (the so-called political arithmetics of the seventeenth century), then the works of many mercantilist authors constitute a kind of 'protoeconomic history.'

Adam Smith (1723–1790) was the first modern scholar to use economic history as a tool of economic analysis. In his famous book *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776) he described and analyzed the working of Mercantilism in its historical perspective in order to criticize it and to propose a better, more efficient system, the free market system. Smith preferred inductive reasoning and therefore saw economic history as a kind of laboratory in which he could test his theoretical assertions. His work laid the foundations for classical economics. The most eminent representatives of this tradition, however, did not follow the path of the founding father and deprived the newborn theoretical economics of its historical dimension. David Ricardo (1772–1823), James Mill (1773–1836) and William Stanley Jevons (1835–1882) developed economics into an analytical instrument, based on mathematical abstraction and logical theory. Ricardo, whose works ironically would inspire so many economic historians, liked to present his ideas as abstract mechanisms without referring to historical events or even existing institutions. Most classical economists remained interested in institutional economics but concentrated their scholarly efforts more exclusively on constructing economic laws and concepts to explain the working of the market, leaving economic history aside. Notable exceptions, such as Alfred Marshall (1842–1924), used economic history rather in an eclectic way, presenting institutional structures and developments as clarifying illustrations of their theoretical analyses. As a result, by 1900 economists elaborated elegant and symmetric models of general equilibrium in which the historical presence of time was completely lacking.

From the 1840s onwards, a German Historical School emerged, reacting systematically against the deductive theoretical approach of the English classical economists by emphasizing that the inductive method, based upon economic history, should be the starting point of all theoretical reasoning. As the German economists were also interested, as were the English, in institutional economics, but, in contrast to the latter, wanted to add a specific historical dimension to it, they focused their research on describing and explaining the changes over time of institutions and systems. Inspired by their historical research they constructed theories of economic development in stages (the so-called 'Stufen' theories). Friedrich List (1789–1846) was a pioneer in this respect, subdividing the historical development of production into four subsequent stages of dominant economic activity: the stage of hunting and collecting, the agricultural stage, the agricultural–manufacturing stage, and, finally, the agricultural–manufacturing–trading stage. During the second half of the nineteenth century, his colleagues and followers, such as Bruno Hildebrand (1812–1878), Wilhelm Roscher (1817–1894), Karl Knies (1821–1898), Karl Bücher (1847–1930) and Gustav von Schmoller (1838–1917), elaborated the 'Stufen' idea still further. They introduced other 'economic criteria' as determining variables in the process of development, for example the way of exchanging goods (barter, money, and credit) or the distance between producer and consumer (closed domestic economy, city economy, national economy, and world economy).

The younger members of the German Historical School, active at the end of the century and at the beginning of the twentieth, remained faithful to the 'Stufen' hypothesis but widened the scope of their research. Werner Sombart (1863–1941), for example, added social and cultural variables to the economic ones, characterizing the stages more broadly as 'pre-capitalism,' 'early-capitalism,' 'high-capitalism,' 'late-capitalism,' and 'socialism.' Max Weber (1864–1920), being a sociologist as well as an economist, deepened his analysis by focusing on the links between Protestantism and the rise of capitalism in early modern Europe. Karl Marx (1818–1883), finally, combined the German Historical School's approach with the analytical methods of the classical economists when analyzing the transitions from a primitive society to economies based on slavery, later, on feudalism, which then evolved into capitalism, and from there would lead to socialism–communism.

Many young historians from neighboring countries, completing their studies in Germany, were fascinated by the themes of the German Historical School and integrated them into their own research. The Belgian historian Henri Pirenne (1862–1935), for example,

would analyze in depth the institutional development of industrial and commercial capitalism in Flanders and Brabant during the late medieval period. In Germany these scholars, who in their countries would soon hold university chairs themselves and have students of their own, became attracted by the study of economic phenomena in the past. They were also profoundly influenced by the new methodology that in the second half of the nineteenth century had found acceptance in the history departments of many German universities. This new methodology was the famous 'seminar' system that Leopold von Ranke had successfully introduced at the University of Berlin in 1833. Within these seminars, students had a direct contact with primary sources. Students were trained to solve problems of identification and interpretation by combining the techniques of a rapidly increasing variety of auxiliary sciences (paleography, chronology, numismatics, metrology, etc.) with the fine skills of internal and external source criticism. With the integration of this rigorous methodology, the economic historian started to distinguish himself from mainstream economists who showed less concern with the collection, use, and general validity of the data.

By studying economic–historical development in such a systematic way, the members and imitators of the German Historical School became, *de facto*, the founding fathers of economic history. Although the scholars contributed little to the further development of economic theory, their historical approach would remain influential during the whole twentieth century. Several economists of the twentieth century, in particu′ lar those studying modern economic development, such as Joseph Schumpeter (1883–1950), Arthur Lewis (1915–1990), and Walt W. Rostow (1916), were inspired by the ideas and methods of the School when elaborating their own theories.

During the second half of the nineteenth century, an English Historical School had also emerged, reacting in its turn against the too abstract approach of the classical economists. A first group of English scholars, under the influence of Positivism, became interested in historical statistics, proclaiming that serial history was a necessary tool for verifying or falsifying abstract economic theories and laws. At the middle of the century, Thomas Tooke (1774–1858) with his *History of Prices* had already paved the way, but it was James E. Thorold Rogers (1823–1890), publishing a large-scale history of agricultural prices in England between 1866 and 1902, who became the real pioneer of the group. A second group studied economic development in a historical perspective, similar to what the Germans had already done and were still doing. However, the English, as could be expected, were focusing in the first place on their own national experience. Arnold Toynbee (1852–1883), for example, studied the English Industrial Revolution and William J. Ashley (1860–1927), in the same vein, decided to undertake a survey of the whole English economic past. John H. Clapham (1873–1946), on the contrary, compared the development of France and Germany during the nineteenth century and also wrote a masterly survey of the economic history of Britain. Some American scholars can be considered belonging to the same group, because of their interest in economic institutions and their evolution over time: Thorstein Veblen (1857–1929), for example, focused on American capitalism during the second half of the nineteenth century.

The success of the German and the English Historical Schools was undoubtedly the most important but by no means the only factor that explained the promotion of economic history as an independent discipline within the historical sciences. Two other causes should be mentioned as well. First, the spread of Marxism, which became more and more influential among intellectuals and in academic circles, also favored the interest in economic history. The conviction that the production of goods and services determined political, social, cultural, and spiritual processes stimulated scholars to study themes such as the means of production, the relationships between social groups, landed property, the distribution of income and wealth, the evolution of wages, labor conditions, and other related topics. A second factor that certainly contributed to the development of the new discipline was the industrial revolution and the subsequent industrialization of great parts of the European continent in the second half of the nineteenth century. The new production methods, new techniques and, above all, the impressive economic growth in many countries created a favorable intellectual climate to study in a systematic manner the context and the conditions in which the new economic phenomena took shape.

Before World War I, economic history had developed into a healthy and respected discipline that confirmed its new status with an increasing number of students and with specific journals, university chairs and congresses. Of course, there were marked national differences. In Germany, hardly a surprise considering the importance of the Historical School in this country, a course on 'Wirtschaftsgeschichte' (economic history) was already announced at Heidelberg University in 1853, but never took place. The famous London School of Economics organized lectures on economic history right from the outset (1895), but the first university chair in the world would be installed in the United States. This happened at Harvard University where W. J. Ashley (1860–1927), a fellow of Oxford, accepted this position in 1892. The first chair for economic history in Britain was established at Manchester in 1910 and held by George Unwin (1870–1925). In the Netherlands, N. W. Posthumus (1880–1963) occupied the first chair of economic history, at Rotterdam, in 1913. The chair of economic history at the Sorbonne was created in 1927 and given to Henri Hauser (1866–1946). Meanwhile, economic history

already had its first journals. Again, Germany was the first country with a specific journal entirely devoted to economic and social history: this was the *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte* with its first issue in 1903. France followed with the *Revue d'histoire économique et sociale* in 1908. The famous *Annales d'histoire économique et sociale* that would strongly promote research in economic history only dates from 1929 and was therefore preceded by a.o. the *Economic History Review* (UK, 1927) and the *Journal of Economic and Business History* (USA, 1928). Except for the Italian *Rivista di Storia Economica* (1936) and the American *Journal of Economic History* (1941), all major journals date from the postwar period.

2. Between the Wars

In the inter-bellum period, economic history would expand further. However, it shifted its focus, towards economic problems arising from the war, such as war and postwar inflation, or towards problems becoming critical during the thirties, such as the chaos resulting from the world crisis. In the view of economic historians, the use of historical statistics could help substantially in deepening the analysis in both cases. An International Scientific Committee for Price History was created in 1931, on the initiative of Edwin F. Gay (Ashley's successor in 1902) and William H. Beveridge (1879–1963), to set in hand price studies in England, France, Germany, Austria, the Netherlands, the United States, Spain, and Poland. From 1934 onwards, several price studies concerning these countries were published. After World War II the Committee would not be active any more, but publications of historical price series still multiplied, not only in the countries mentioned above, but also in Belgium, Denmark, Russia, Italy, and many other countries.

In France, the construction of long-term price series aimed originally at verifying or falsifying quantitative monetary theory. François Simiand (1873–1935) was a pioneer in this respect, but he on extended his research to explore also the role of wages in the economy, basing it, once again, on historical statistics. Ernest Labrousse (1895–1989) took over Simiand's legacy, orienting it, however, towards the analysis of cyclical fluctuations in Europe during the *Ancien Régime*. His student, Jean Meuvret (1901–1971), integrated demography into the cyclical analysis, as did the German economic historian Wilhelm Abel (1904–85), in his famous study on agricultural fluctuations in central Europe since the late Middle Ages (*Agrarkrisen und Agrarkonjunktur*). Economists and economic historians, in particular in the Anglo-Saxon countries, were more interested in the study of the 'modern' business cycle. Scholars, such as Arthur F. Burns, W. C. Mitchell, Arthur D. Gayer, Joseph Schumpeter, Ernst Wagemann (with his Institut für Konjunkturforschung at Kiel), Walt W. Rostow and many others, published quantitative historical surveys of the British, American, German, and other business cycles during the nineteenth and twentieth centuries. The tradition would be continued after the World War II by R. C. O. Matthews, Philip A. Klein, Robert A. Gordon, and Erik Lundberg, among others.

Apart from the creation of the International Committee on Price History, two other important innovative initiatives were taken during the inter-bellum period, which, as was also the case with the history of prices, would continue to influence postwar economic historiography substantially. First to be mentioned is the ambitious project of publishing a multi-volume *Cambridge Economic History of Europe*, written by a large, international team of qualified scholars. The first volume appeared in 1941 and was followed by many others. The creation in 1929 of the French journal *Annales d'histoire économique et sociale* by Marc Bloch (1886–1944) and Lucien Febvre (1878–1956) was a second important event. The two founders had the ambition to encourage a new kind of social and economic history, focusing on daily material life, and to make this kind of history truly interdisciplinary by integrating the social sciences. In its first years, the *Annales* remained faithful to the quantitative-historical Simiand–Labrousse tradition, its themes of research being very diversified. Outside of France, the impact was still minimal. Gradually the journal moved more systematically in the direction of historical anthropology and global history in which the qualitative approach was gaining ground. When Fernand Braudel (1902–1985) took over control of the journal in the early 1950s, the Braudelian vision of a qualitative 'new global history' became the predominant idea of the *Annales* School. Its impact on European economic history now grew quickly, and with some time lag, the influence could also be felt in England, and in the United States as well.

3. The Heyday (1950–1980)

No doubt, the pre-war initiatives, mentioned above, were crucial for the postwar expansion of economic history to become one of the main disciplines of the economic and historical sciences, but several new initiatives taken after World War II by economists and economic historians alike were as decisive. During the immediate postwar decades, the Anglo-Saxons clearly dominated the scene of innovation. Keynesianism became the great fashion in economics, but it was soon complemented by theories of economic growth. The more dynamic approach of the growth theories stimulated the interest of the American economists, such as Moses Abramovitz and Colin Clark, in long-term economic-historical development. In 1950, during the Paris world conference of historical sciences, Clark incited the economic historians to join the economists in taking up research on long-term economic growth.

The appeal was heard. At the end of the 1950s, the American Economic History Association and the National Bureau of Economic Research were joining efforts in studying the evolution of income and wealth in the United States during the nineteenth century. During the first international economic history conference in Stockholm in 1960, themes on industrialization, human capital, and technology as factors of economic growth became predominant in the modern history sessions. Based on historical experience, concepts and theories on the 'take-off' (Walt W. Rostow), on the 'great spurt' (Alexander Gerschenkron), and on capital formation and capital utilization (Evsey Domar) were formulated.

Soon the scope of the discussions widened, leading, *inter alia*, the following new questions.

(a) Had planned economies a comparative advantage or disadvantage vis-à-vis the free-market economies in promoting economic growth?

(b) How to explain regional growth disparities in the West from a historical point of view?

(c) How to explain uneven economic development in the world?

(d) Finally, how to specify the links between growth and income?

The last question would be analyzed in depth by Nobel Prize winner, Simon Kuznets (1901–1985), who started to use reconstructed historical series of national accounts for that purpose, emphasizing that economic growth could only be studied scientifically if it was based on trustworthy macroeconomic statistics. Kuznets' appeal would generate a large-scale research movement of reconstructing national accounts. Phyllis Deane and W. A. Cole's successful book *British Economic Growth, 1688–1959*, first published in 1962, used such statistics, and encouraged the movement still further. Scholars in the United States, France, Germany, Belgium, the Netherlands, Spain, Switzerland, Sweden, Finland, and Portugal were launching projects to reconstruct historical national accounts of their own countries. In the United Kingdom, Charles Feinstein, Nicholas F. R. Crafts and others soon would start working on correcting and improving the estimates of Deane and Cole. Research on the links between growth and income, on the other hand, increased interest in the social aspects of growth and development. Marxist and non-Marxist economic historians in the following years would expand this field of research very much, emphasizing the importance of socio-institutional factors in the process of growth and development.

As far as the economic history of the *Ancien Régime* was concerned, the influence of Michael M. Postan (1899–1981) was prominent during the immediate postwar years. His long-term Malthusian demographical approach would become a model of interpretation for European economic development during the Middle Ages as well as during the early modern period. The Postan model would lead to the creation of the Cambridge Group for the History of Population and Social Structure under the directorship of Edward A. Wrigley, Roger S. Schofield and Peter Laslett. The French would join the English Malthusian demographical approach, although Michel Fleury, Louis Henry, Jean Meuvret and Pierre Goubert analyzed short- and medium-term demographical fluctuations in particular. For that purpose, Fleury and Henry would introduce the prosopographical method into historical demography and develop a method for family reconstitution based on the information in the parish registers. Emmanuel Le Roy Ladurie, on his part, would integrate, somewhat later, Postan's long-term model into his own research on French and European agricultural history in the early modern period.

During the same period the *Annales* School, now firmly dominated by Fernand Braudel, was extending its influence via the Ecole Pratique des Hautes Etudes de la Sorbonne, not only among the French economic historians, but also in the whole of Europe and soon beyond. From the late 1950s onwards until the early 1980s, the School would inspire a large part of economic historical research on:

(a) The late Middle Ages (for example Georges Duby and Jacques Le Goff).

(b) The early modern times (for example Pierre Jeannin, Jean-François Bergier, Pierre Deyon, Bernard H. Slicher van Bath, Jerzy Topolski, Pierre Chaunu, Felipe Ruiz Martin and Herman Van der Wee).

(c) The French Revolution period (for example François Furet and Denis Richet).

Originally, research was often focused on geographical studies of economic development, following the example of Braudel's pioneering study on the Mediterranean, but, in contrast to Braudel, most of these studies combined qualitative approach with quantitative analysis. Gradually, the qualitative 'global history' approach became predominant. Within this framework, under the influence of Marxist economic historians such as Immanuel Wallerstein, the socio-institutional aspects were getting increasingly more attention in the analysis of the agricultural, commercial, industrial, and financial development of Europe during the *Ancien Régime*.

While the methodological trend in the *Annales* School from the 1950s onwards was moving in the direction of a more qualitative approach to economic history, in the United States the trend during that period moved clearly in the opposite direction. In 1954, at Purdue University in Lafayette (Indiana), economists and economic historians already used primitive computers for the processing of statistical data. Three years later, during a conference of the American Economic History Association in 1957, Alfred H. Conrad and John R. Meyer launched the idea of a 'New Economic History,' aiming at integrating economic theory, quantitative methods, and his-

tory into one discipline. They illustrated the relevance of their idea by applying a microeconomic input–output method when analyzing the profitability of slavery in the American South during the nineteenth century. Nobel Prize winner Robert Fogel and Stanley L. Engerman later took up the theme. This led to a large-scale research project, its conclusions being published in *Time on the Cross*, 1974.

Some young economic historians received Conrad and Meijer's idea with enthusiasm. It gave birth to 'Cliometrics,' soon a well-established subdiscipline of American economic history. In the early 1960s, Robert P. Thomas had already applied the cost–benefit analysis method to explore the effects of the British Navigation Laws on the economy of the American colonies. More spectacular was the research undertaken by Robert Fogel and Albert Fishlow on the contribution of the railroads to American economic growth. To measure their impact, Fogel constructed a hypothetical world and used a counterfactual proposition: what would have happened to an economy if, contrary to the facts, one or more important features of that economy had never existed? Or applied to his research: if a railway system had not been introduced in the United States during the second half of the nineteenth century, could an alternative transport system, such as a system of existing and new canals, have generated as many social gains and dynamic effects as the railways did? The results of Fogel's cliometric test were fascinating: the difference would have been minimal since the 'social savings' would have represented less than 5 percent of national income in 1890. The cliometricians, now organized in an autonomous association, the Cliometric Society, and holding their own yearly meetings, began revisiting all major themes of American economic history, hoping that their methodology would lead to a complete reinterpretation of that history. Major themes studied by the cliometricians were:

(a) the economic reconstruction of the South after the Secession War;

(b) the impact of labor and migration on American growth at the end of the nineteenth century (cf. the model of Jeffrey Williamson, applying macroeconomic equilibrium theories);

(c) money markets and banking systems as factors of growth in the United States in the nineteenth century;

(d) the American Industrial Revolution;

(e) the Great Depression of the 1880s;

(f) the World Crisis of the 1930s; and

(g) the role of the residual factor in American productivity growth.

The methodological innovation, introduced by the American New Economic History, would spill over to the rest of the world, in particular to Canada, Australia, and some European countries, particularly the United Kingdom, Germany, France, and Sweden. Moreover, from 1970 onwards each world congress of the International Economic History Association would have a session on methodology, dominated from the beginning by the American cliometricians. However, the impact of the New Economic History outside the United States remained limited and certainly did not gain widespread recognition. F. Redlich, for example, described the counterfactual deduction in 1968 as 'fictitious history' and 'not history at all.' Was it because neoclassical theory was found to be less applicable to the European pre-industrial period than to the nineteenth-century American economy? Alternatively, were institutional factors, such as the separation between economists and economic historians who are normally active in the arts faculties, responsible for the emotional aversion towards the blend of history, economic theory, and econometrics? Even in the United States itself, dissidents soon emerged. Nobel Prize winner Douglass North, who had been applying the New Economic History methods in his studies on American growth, withdrew from cliometrics and reoriented his research to the institutional aspects of growth, using the historical experience to construct and verify his hypotheses. In his view, the protection of 'property rights' and the lowering of 'transactions costs' were the decisive factors in the history of economic growth. In the same vein as North's New Institutional History, but more innovating and with a much larger impact in and outside the United States, was Alfred D. Chandler Jr's founding of the 'New Business History.'

When creating the Center of Entrepreneurial Studies at Harvard in the late 1930s, Joseph Schumpeter and Arthur Cole had hoped to replace the hagiographical tradition of prewar and interwar business history by a more scientific approach. However, it was Chandler who, from the 1960s onwards, realized that ambition fully with his books *Strategy and Structure* (1962), *The Visible Hand* (1977), *Scale and Scope* (1990) and *Inventing the Electronic Century* (2001). Careful historical investigation, combined with in-depth institutional economic analysis, revealed the crucial role of business management and business organization in the successful rise of American capitalism. This was soon completed by studies on the rise and development of European and Japanese capitalism. Chandler's hypotheses and methodology became standard all over the world and are still so. Up to the present day, they are the main source of inspiration to the 'New Business History.' A last new economic history group, emerging after World War II, should be mentioned: the 'New Urban History,' integrating sociological, geographical, and economic theories and quantitative methods into the historical analysis of towns and urbanization. In 1962, an interdisciplinary Urban History Group was established, followed three years later by the 'Commission Internationale pour l'histoire des villes' directed by scholars such as Henri Aubin, Hektor Ammann, and Philippe Wolff. However, the Anglo-Saxon historians, once again,

were pioneers in the field, H. J. Dyos, Jan de Vries, Paul M. Hohenberg, Lynn Hollen Lees, Peter Clark and many others.

Postwar economic history was also influenced by the social effects and economic problems generated by the spectacular expansion of the Western economy during the postwar period. Indeed, research was much inspired by the increasing number of publications, such as the report of the Club of Rome (1972), revealing the problems caused by economic growth, or by movements protesting against the negative effects of growth in the West, such as the activist policies, formulated by the UNCTAD in the course of the 1960s and 1970s. Responding to these reactions, new themes emerged in the programs of the world conferences of the International Association of Economic History. These included, from the 1970s onwards, in particular: 'Natural Resources and Economic Development in History,' 'Oil in the World Economy,' 'Environment and Urbanization in History,' 'Ecological History,' 'The Economic History of Leisure and Recreation,' 'Typology of Colonial Economic Development,' 'The Economic History of the Third World,' 'Loans, Debts and Economic Development in the Nineteenth and Twentieth Centuries,' *inter alia*. The Cold War and its ideological conflicts would lead to a number of studies comparing the efficiency of the two economic systems in an historical perspective.

4. *Shifts and Changes (1980–2000)*

From the 1980s onwards, new shifts occurred in the postwar development of economic history. The first important shift originated from two new circumstances: economic history, as an autonomous discipline, crossed its traditional European and North American borders and acquired strong positions also in many Latin American and Asian countries, Australia and New Zealand and, to a lesser extent, the African continent. Moreover, globalization of the world economy, from the 1980s onwards, clearly accelerated its pace. The effect of both factors on economic history was twofold. Firstly, seen in relative terms, research was re-orienting itself increasingly in the direction of contemporary history, at the expense of medieval and early modern history. Secondly, a geographical re-orientation took place: research on non-European topics, particularly concerning Asian and Latin American countries and again in relative terms, expanded remarkably.

Another important shift was thematic. Economic development was increasingly considered a part of a much more complex reality, in which mental, cultural, and socio-institutional variables were as crucial as the economic ones for understanding and explaining it. This changing view on the place of economic history in the scientific approach to reality generated many new fields of research. Themes such as 'Ethnic Minority Groups and their Impact on Economic Development,' 'Guilds, Economy, and Society,' 'Women and the Practice of Credit,' 'Film: an Industry on the Crossroads of Economics, Politics, and Arts,' 'Consumption, Lifestyle, and the Standard of Living,' 'Gender and Labor Migration,' 'Courts as Economic Institutions,' etc., were very successful during the conferences.

A third important shift was the increasing interest of economic historians iicro-historical research: an interest in the experience of the individual ('*l'expérience vécue*'), an interest in the uniqueness of historical events ('*das Einmalige*'), an interest in the narrative and the descriptive at the expense of a more analytical history, and looking for recurrent patterns in historical development. Many economic historians preferred the meticulous approach of specific, atypical, idiosyncratic cases to broad but abstract theories or structural change analysis. Themes such as 'Cities at War,' 'The Economy of Private Households,' 'Education and Career of Business Managers,' 'Potatoes and Famines in mid-nineteenth Century Europe,' were significant for the rise of this kind of micro-history.

At the end of the twentieth century, economic history had expanded and diversified its research so much that it would be difficult to summarize its development in a nutshell. In a simplifying view, one could distinguish two main tendencies. A first one could be considered a modern version of the traditional socioeconomic historiography, presenting in a descriptive way the results of new research in the archives, particularly concerning aspects of daily life not yet studied much; for example, sexuality, criminality, discrimination against women, food, clothing, games, leisure, fashion, the environment, and domestic services. The second tendency, and by far the more important, aims at integrating elements of economic theory and methodology into historical analysis. This second tendency is split into two subgroups. A first group of scholars, now a minority within the discipline and mostly members of departments of economics in Europe and Northern America remains faithful to the American 'New Economic History.' However, by organizing themselves in a Cliometric Society (and partly also in the European Economic Association), they isolate themselves somewhat from the rest of the discipline. The second subgroup, the more numerous one, also uses economic theory and statistics. These scholars, however, do not formulate their hypotheses in the formal way as the cliometricians do. They prefer not to verify or falsify their hypotheses with econometric techniques as used by the cliometricians, and they include demographical, cultural, anthropological, mental, and institutional variables in their economic analysis.

See also: Anti-Semitism; Business History; Capitalism; Development, Economics of; Economic Anthropology; Economic Growth: Measurement; Econo-

mic History, Quantitative: United States; Economic Sociology; Economics, History of; Economics: Overview; Historical Demography; Historicism; History of Technology; Industrial Relations, History of; Industrialization, Typologies and History of; Marshall, Alfred (1842–1924); Quantification in History; Social History

Bibliography

Boyer R 1989 Economie et histoire: vers de nouvelles alliances. *Annales, Economies, Sociétés, Civilisations* **44**: 1397–426

Cipolla C M 1991 *Between History and Economics. An Introduction to Economic History*. Blackwell, Oxford, UK

Clark S (ed.) 1999 *The Annales School. Critical Assessments in History*. Routledge

Dumke R H 1986 Clio's Climacteric? Betrachtungen über Stand und Entwicklungstendenzen der Cliometrischen Wirtschaftsgeschichte. *Vierteljahrschrift fur Sozial- und Wirtschaftsgeshich* **73**: 457–87

Field A J (ed.) 1987 *The Future of Economic History*. Kluwer, Boston, MA

Fogel R 1970 Die neue Wirtschaftsgeschichte, Forschungsergebnisse und Methoden. *Kölner zur Sozial- und Wirtschaftsgeschichte* **8**. English version in *The Economic History Review, Second Series* **1966**

Grantham G 1997 The French cliometric revolution: A survey of cliometric contributions to French economic history. *European Review of Economic History* **1**: 353–405

Kadish A 1989 *Historians, Economists, and Economic History*. Routledge, London

Kindleberger C P 1997 *Economic Laws and Economic History*. Cambridge University Press, Cambridge, UK

McClelland P D 1975 *Causal Explanation and Model Building in History, Economics and the New Economic History*. Cornell University Press, Ithaca, NY

McCloskey D N 1991 Counterfactuals. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave. The Word of Economics*. Macmillan, London, pp. 149–54

Redlich F 1965 'New' and traditional approaches to economic history and their interdependence. *The Journal of Economic History* **25**: 480–95

Rouzet F, Lescent-Giles I 1998 French economic history in the past 20 years. *Neha-Bulletin* **12**: 75–101

Snooks G D 1993 *Historical Analysis in Economics*. Routledge, London

Subacchi P 1995 Meta-economic history: A survey of the Eleventh International Economic History Congress. *The Economic History Review, Second Series* **48**: 602–11

Temin P 1981 The future of the new economic history. *Journal of Interdisciplinary History* **12**: 179–97

Tilly C, Tilly L A, Tilly R 1991 European economic and social history in the 1990s. *The Journal of European Economic History* **20**: 645–71

Tilly C, Tilly R 1971 European economic history in the 1970s. *The Journal of Economic History* **31**: 184–98

Van der Wee H 1993 International economic–historical research in retrospect (1960–1990). In: Riis T (ed.) *A Special Brew … Essays in Honour of Kristof Glamann*. Odense University Press, Odense, Denmark, 91–105

Van der Wee H, Delbeke J 1983 Quantitative research in economic history after 1945. In: Fremdling R, O'Brien P K (eds.) *Productivity in the Economies of Europe*. Klett-Cotta, Bamberg, Stuttgart, Germany, pp. 11–29

E. Aerts and H. Van der Wee

Economic History, Quantitative: United States

Economic history is a hybrid discipline, concerned with recording and interpreting the economic events of the past. It should not be confused with the 'history of economic thought', the study of the intellectual history of economics as a discipline. Although the practice of compiling historical statistics goes back to the nineteenth century, if not earlier, the term 'quantitative economic history' refers to a self-conscious movement dating from the 1950s and 1960s, also known as the 'new economic history' or 'cliometrics', which set out to revolutionize the field by incorporating explicit economic analysis and quantitative methods into the study of past economic performance. In subsequent decades, practitioners have extended and redirected this agenda, by drawing upon diverse branches of economic theory to analyze the historical evolution of market and nonmarket institutions, technology, preferences, and other elements commonly taken as exogenous by mainstream economics. The resulting literature stresses the influence of historical experience on ensuing economic patterns, an intertemporal linkage known as 'path dependence.' This article traces the main tendencies within quantitative economic history since the mid-nineteenth century, though it is by no means a comprehensive survey.

1. The Origins of Economic History

Recognition of economic history as a distinct subject is traceable to the consolidation of British neoclassical economics at the turn of the twentieth century, under the towering influence of Alfred Marshall and his Cambridge successor, A. C. Pigou. The analytical core of the dominant school held that economies tended towards 'long-run equilibria' that were independent of initial conditions (although Marshall himself was by no means dogmatic about this doctrine in practice). Having lost out in the struggle for the disciplinary identity of economics itself, the 'historical school of economics' emerged at that time as a reaction against the neoclassical paradigm and the laissez-faire program with which it was associated. Drawing ideas and inspiration from the 'German historical school' of Gustav Schmoller and Friedrich List, early historical economists emphasized the virtues of inductive studies

as contrasted with the deductive or axiomatic approach. Methodological preferences were reinforced by their affinity with 'neomercantilist' policies, including promotion of national industries, social reforms, and the regulation of enterprise. The organization of economic history as an academic field in Britain was prompted as much by a search for refuge as by the emergence of a distinctive research agenda. William James Ashley, first president of the Economic History Society at its founding in 1926, lamented: 'the theoretical economists are ready to keep us quiet by giving us a little undisputed garden plot of our own; and we humble historians are so thankful for a little undisputed territory that we are inclined to leave the economists to their own devices' (quoted in Koot (1987, p. 7).

Similar tensions persisted in the United States, but the separation between economics and economic history was slower to develop. The inductive approach of the 'institutional' school associated with such figures as John R. Commons and Thorstein Veblen was dominant within labor economics and in other parts of the parent discipline, displaced only by the rise of Keynesian economics in the 1930s and 1940s. In the same era, 'progressive' historians such as Frederick Jackson Turner and Charles Beard put economic conflict and change at the center of their 'economic interpretation of history.' Although neither of these groups would be recognized today as economic historians, they did provide outlets for those dissatisfied with the deductive methodological approach. Thus, the (American) Economic History Association was founded only in 1940, and lacked the ideological identity and fervor of its British counterpart.

Between the 1930s and the 1950s, a few exceptional figures such as Carter Goodrich could have a significant impact on both economics and economic history. But the majority of EHA members were historians, whose works made little use of economics and had little influence on the thinking of economists. The same could be said for entrepreneurial history, which flourished for a time at the Harvard Center for Entrepreneurial History, under the leadership of Arthur H. Cole. In retrospect, many of the detailed descriptive studies from this era have lasting value, perhaps best reflected in monographic series entitled *The Economic History of the United States*, published between 1945 and 1962. But as economics became increasingly analytical and quantitative, economic history at the time appeared to be something of a stodgy backwater.

2. *The Rise of Quantitative Economic History*

The two decades after World War II saw a significant expansion of economic history within economics, much of it quantitative in character. This was mainly an American development, and largely (although by no means exclusively) focused on the economic history of the United States. The immediate stimulus was the new postwar interest in economic growth. Following the lead of Simon Kuznets, father of the national income accounts, a new generation of economic historians sought to extend the statistical record of the American economy back to its origins early in the nineteenth century, drawing upon the rich potential of the decennial census and making ingenious use of other diverse archival sources. The major results of these efforts were presented in landmark volumes published by the National Bureau of Economic Research (Parker 1960, Brady 1966). Estimates developed by Robert Gallman suggested that growth in the United States had reached modern rates by the middle of the nineteenth century, and should not therefore be seen as the consequence of political upheavals at the time of the Civil War (as hypothesized by Charles Beard). Adapting Gallman's estimates into a regional framework, Richard Easterlin showed that the slave economy of the South also shared in the growth acceleration, though per capita income in that region was some 25–35 percent below the national average.

A second and more controversial thrust was the deployment of tightly constructed deductive models to address specific hypotheses about this statistical record. The advent of this 'cliometric' approach is often dated from the publication in 1958 of 'The Economics of Slavery in the Ante-Bellum South' by Alfred H. Conrad and John R. Meyer (reprinted in Fogel and Engerman 1972). In this article, Conrad and Meyer set out to settle a perennial debate over the 'profitability of slavery,' by treating slaves as a form of capital and posing the question as an explicit rate-of-return-on-investment calculation. Their conclusion was that investments in slaves were just as profitable as alternatives such as railroad bonds. Other studies tackled broader questions about the origins of growth and the impact of government policies. Douglass North (1961) interpreted the acceleration after 1815 as a dynamic response to the potential of expanding markets, primarily through interregional specialization and trade. Fogel (1964) and Fishlow (1965) generated comprehensive estimates of the contribution of railroad construction to national economic growth, by posing explicit historical counterfactual circumstances within an analytical framework. Both the methods and the conclusions of these provocative works were the subjects of intense academic debate for many years after publication.

One strand of this discussion represented a reaction by 'old economic historians' against attempts to subsume historical complexity into what they regarded as hypothetical and unrealistic abstractions. A second strand, however, came from within the economics discipline, reflecting dissatisfaction with the limitations of deductive models as the working tools for structuring research in economic history. In 1972, two collaborative efforts summarized the major results of

the New Economic History (Davis et al. 1972, Fogel and Engerman 1972). The overall yield was impressive, an upgrading of the quantitative base of historical knowledge, and a richer appreciation of the presence of market forces in history. Signs of growing unrest were also evident, however, and in retrospect these volumes have the character of closing chapters in a methodological cycle that had largely run its course by that year. This interpretation may be elaborated by a review of two major research areas.

2.1 The Economic Analysis of Technological Change

Economic historians of all stripes had no difficulty agreeing that a central component of economic growth was technological change, broadly defined to include organizational change and the diffusion of existing techniques, in addition to the more famous 'inventions' that have long been standard fare in textbooks. But the dominant tradition in economic analysis has long been accustomed to regarding technological change as exogenous, not part of the domain of a discipline specialized in analysis of the *implications* of such change for the economy. One might reasonably hope therefore that quantitative economic history would be able to say more about the timing, pace, and direction of particular historical technologies, if not to uncover historical laws of motion for technological change in general.

In the context of American economic history, these questions were brought to the fore by a landmark book published in 1962 by the British economic historian H. J. Habakkuk, *American and British Technology in the Nineteenth Century*. Habakkuk hypothesized that American industry pioneered the development of capital-intensive methods as compared with Great Britain, in response to the scarcity of labor relative to land in the United States. This apparently straightforward thesis subsequently proved exceptionally difficult to capture analytically or to confirm empirically. An extended stream of contributions by Peter Temin, Robert Fogel, Nathan Rosenberg, Paul David, Alexander Field, John James, David Hounshell, and Kenneth Sokoloff, among others, succeeded in elaborating many characteristic features of the technologies emerging as an American national specialty in the nineteenth century. In comparison with the British, American techniques were not only capital intensive, they were also resource using, faster paced, skill saving (at least for traditional craft skills), and adapted for long production runs of standardized final products. As useful as this accumulated descriptive knowledge may be, the question remains: What sort of economic analysis would allow the quantitative economic historian to come to terms with the origins and nature of this list of characteristics?

The long-running debate over the 'labor scarcity paradox' brought out most clearly that so long as the analytical framework was defined by standard neoclassical analysis, the subject matter would be transmuted into the language of 'relative factor abundance' and substitution, terms derived from the assumption that a 'production function' is given by technical factors. Within this constraint, the economic historian was powerfully driven towards the exercise of rationalizing observed historical choices. The impulse was observed most clearly when the choice-of-technique question was transported to the other side of the Atlantic, and cliometricians assessed the popular indictment of British entrepreneurs for their reluctance to adopt American methods. Virtually without exception, the conclusion was that British choices were entirely appropriate responses to prevailing factor prices and market conditions. Although this verdict may be valid, the uniformity of results understandably gave rise to doubts about whether it constituted an authentic inference from historical evidence, or merely an assumption imposed on the evidence by the methodology.

This approach to technology proved unduly constraining for a number of reasons. As subsequent research has demonstrated, the 'factor endowments' taken to be fundamental determinants were themselves not really exogenous, being subject to influence not just by policy but by technology as it unfolded through time. For example, the large inflows of unskilled immigrants to the United States are most readily explained by the growth of industrial jobs adapted to that type of labor, a result in other words of an attribute of the new technology. Even the apparent American abundance of mineral resources, coal, iron ore, copper, petroleum, and many more, turned out on closer inspection to have been more the *result* of technological development and capital investment than its cause (David and Wright 1997). To a significant extent, minerals constituted a 'knowledge industry' even in nineteenth-century America. As such, its development was part of the broader evolution of national technology, or learning. Evidently, the development of an analytical *interpretation* of such phenomena required a new set of tools.

2.2 The Slavery Debate

No topic tested the limits of deductive economic history more fully than slavery. Conrad and Meyer (1958) set out to measure the 'profitability of slavery,' and thereby to test the thesis that slavery would have died out for economic reasons, in the absence of the Civil War. Among others, Yasuba pointed out that because slave imports were ended by legislation in 1807, a more appropriate measure of the private economic viability of slavery was the level of 'capitalized rent' embedded in the price of slaves 'high and rising on the eve of secession, according to Yasuba's estimates (1961, reprinted in Fogel and Engerman 1972). The 'profitability' finding is best understood as

an indication that slave prices were approximately correct as capitalizations of the expected stream of future returns, which may be taken as evidence that 'commercial' motivations were dominant in the operation of the slave system. The achievement of consensus around this broad proposition, however, redirected attention towards more difficult and challenging historical issues, including the political and ideological dimensions of slavery, and the longer-term consequences of slavery as a mode of social organization.

This was the setting for the appearance in 1974 of *Time on the Cross*: *The Economics of American Negro Slavery*, by Robert W. Fogel and Stanley L. Engerman. Fogel and Engerman (1974) went well beyond the previous consensus, presenting total-factor-productivity calculations implying that slave plantations were as much as 35 percent more efficient than free family farms. Not only was the slave system highly productive, but the book went on to assert that material and social conditions among slaves were far better than historians had recognized. Using such indicators as diet, clothing, housing, life expectancy, and family stability, the authors compared the lives of slaves favorably with those of immigrant workers in the northern states. It is hardly surprising that these incendiary claims generated an explosive reaction.

The full range of the ensuing debate cannot be recounted here. As with the response to Conrad and Meyer, one part of the reaction represented traditional historians objecting to the coldhearted nature of the calculations, and the apparently cavalier treatment of the feelings and behavior of the slaves themselves. Another component, however, came from within the profession. Detailed scrutiny of the statistical evidence ultimately led to significant modifications of the relatively benign picture of slavery, most notably in health conditions and family life. It would be a mistake to characterize the alternative views on these issues along simplified dimensions, such as the economist's inclination to presume that historical operators are rational and market oriented. For example, subsequent research suggests that slaves were bought and sold much more frequently than Fogel and Engerman believed, often separating not only husbands from wives, but even mothers from their children (Tadman 1989). On this count, *Time on the Cross* if anything actually understated the ruthlessly profit-driven character of American slavery.

Debates over empirical evidence are a normal feature of any social science. The dimension of the slavery debate with more enduring import for quantitative economic history concerned the *conceptual* underpinnings of the analysis of slavery, specifically the emphasis on calculations of efficiency in production. Because plantations produced multiple outputs with heterogeneous inputs, it emerged that the efficiency results were highly sensitive to relative values assigned to cash crops (such as cotton) vs. food crops (such as corn), and to factors of production (such as farm land of varying quality, or male vs. female slave labor). The debate may be followed in David et al. (1976), and a subsequent exchange in the *American Economic Review*. The deeper issue was whether it was possible, or perhaps whether it was meaningful and historically informative, to separate the calculation of productive efficiency as a technical matter, from the task of locating slavery as a system of property rights in historical time and space. From this perspective, property rights in slaves may be seen to have had a decisive influence on the course of economic development in the southern states, not only on human capital investment, but on geographic patterns of settlement, transportation infrastructure, and urbanization.

In several respects, the course of the slavery debate highlights the new directions in which quantitative economic history has moved since the 1970s: study of the evolution of institutions, incorporation of political, social and ideological dimensions into economic history, and identification of lasting effects of earlier historical phases on later events. These evolutionary tendencies are well illustrated by the careers of two of the founding figures of the New Economic History.

2.3 The Nobel Prizes

In October 1993, the Royal Swedish Academy of Sciences awarded the Nobel Prize in Economics to Robert William Fogel and Douglass North, noting that 'they were pioneers in the branch of economic history that has been called the "new economic history" or "cliometrics".' In essence, the two economic historians were cited for helping to launch the quantitative economic history movement. In what must be a rarity among Nobel laureates, at the time of the award both men had moved some distance from the approach for which they were honored.

North's early work in American economic history was both quantitative and analytical. He developed new statistical time series on the US balance of payments, and on the long-term decline in transportation costs. North (1961) drew upon concepts from theories of location and comparative advantage to interpret the quantitative record of regional performance. His first venture outside American economic history (North and Thomas 1973) was consistent with this focus on market incentives, emphasizing the role of limited government and secure property rights in explaining the comparative historical performance of nations. By 1981, North's range of explanatory material had broadened considerably, to include *ideology* (including norms and belief systems) as a primary element in comprehending the persistence of order in the world, essential for the effective functioning of markets. North (1990) cast the net even more widely, pointing to the critical economic role of

institutions, defined broadly to include all 'the humanly devised constraints that shape human interaction.' One part of North's personal evolution derived from an expansion of his research horizons, generating an awareness that the effectively functioning market society found in the United States is exceptional in a global context. In many ways, however, North's later work marked a return to themes and interests that had occupied economic historians for generations prior to the cliometric revolution.

Fogel's intellectual trajectory was rather different. Fogel (1964) and Fogel and Engerman (1974) built sweeping historical interpretations on the results of complex and ambitious quantitative procedures. After 15 years of further research on slavery, Fogel (1989) devoted more than half of his next book to an account of the rise of antislavery opinion in Europe and the United States, emphasizing the force of religious, political and ideological factors. Fogel wrote: 'I began [this research], like many other cliometricians, not because I was especially interested in the history of American slavery, but because an accident of scholarship made the economics of slavery a major testing ground for the application of cliometric methods. Once drawn into the subject, however, it was the substance of the issues that maintained my interest. Although my principal professional expertise was, and is, in the areas of economics and demography, I found myself led down a road that forced me to grapple with the work of colleagues in cultural, political, and religious history' (p. 13). To be sure, Fogel's later work on historical nutrition and physiology remains solidly quantitative in its essential scope and methods, but his interest in subjective and ideological aspects of history has persisted. It seems safe to say that by 1990, cliometrics had outgrown the historical naivete of its youth.

3. *New Directions in Quantitative Economic History*

In recent decades, quantitative economic history has become steadily more diverse and expansive in both its methods and its subject matter. Some of these tendencies reflect growing diversity in the composition of the cliometric intellectual community, which at the start of the twenty-first century includes many more women and scholars of differing nationality than was true in the 1960s and 1970s. Once almost exclusively an American phenomenon, cliometrics is now also pursued by small but thriving groups in Europe, Latin America, Canada, and elsewhere. Another force for diversification is the expanding range of topics now covered within economics itself, including ecology, law, family, gender, and race, all of which have strong historical dimensions. Since the full range of scholarship cannot be surveyed here, a brief account will be given of two general areas in which historical aspects of economic life seem especially compelling.

3.1 *Increasing Returns, Path Dependence, and Technology*

In 1984, William N. Parker organized a session at the annual meetings of the American Economic Association, on the role of economic history in graduate economics education, the proceedings of which were published in Parker (1986). Among the many spirited defenses of history by such economists as Kenneth Arrow, Robert Solow, Peter Temin, D. N. McCloskey, and others, one contribution proved to be particularly lasting as an illustration of path dependence: Paul A. David's account of the rise and persistence of the QWERTY typewriter keyboard. Although many evidently superior layouts were designed and marketed during the typewriter's post-Civil War diffusion, the standard became 'locked in' to QWERTY by the 1890s, illustrating the power of historical accident and 'network externalities' to shape the course of technological history. Drawing in part upon theoretical work by W. B. Arthur, David argued that because choices were interdependent 'since manufacturers, employers and typists all wanted to use the same keyboard that others were using' the system never shook off the circumstances of its origin, despite the suboptimality of the equilibrium choice.

The QWERTY story has emerged as a popular metaphor for the proposition that 'history matters' in economic life. The typewriter example and its meaning have been hotly controversial, and its appropriate scope and significance have by no means been settled. But its relevance for quantitative economic history has been to confirm the historical character of the economics of technology. As noted above, the study of technological change had never fit comfortably within the New Economic History. Nathan Rosenberg, a member of the first cliometric generation whose research on technological change diverged from the reigning paradigm, had long been engaged in *Exploring the Black Box* (1994), i.e., trying to understand technological change as it would be seen by technical people. In *The Lever of Riches*, Joel Mokyr (1990) elaborates the many context-specific factors that have served to encourage, retard, or channel technological progress. The broad conclusion of this research is not to deny the proposition that technology is responsive to incentives, quite the contrary, but to argue that these responses are sharply conditioned by the historical and institutional setting, including the cognitive structures through which knowledge is codified and transmitted.

3.2 *Institutions and Economic History*

In the heyday of the New Economic History, 'institutional' was used as a term of dismissal if not

contempt, to refer to old-fashioned descriptive work that lacked theoretical foundations. As the field has evolved, 'institutional' economic history has again become respectable, although in new and more analytical clothing. Indeed, many argue that the study of institutions is at the heart of the case for historical economics. Whereas the parent discipline specializes in the functioning of markets, the domain of the economic historian is the institutional foundations that allow markets to function as they do. These subject areas have become more pressing in the light of the varied experiences of formerly planned economies in their transition to market systems.

In its first phase, the so-called New Institutional Economics identified 'institutions' with the property rights, rules and regulations defined and enforced by the state, interpreting institutional change as responses to exogenous shifts in population, technology, or the scope of markets. Beginning with North and Thomas (1973), numerous historical studies examined the implications of these institutions for economic performance, broadly confirming the general proposition that securely enforced property rights are favorable for economic efficiency and growth. Deviations from this pattern are attributed commonly to the influence of special interest groups, with transactions costs standing as barriers to more efficient systems.

Dissatisfaction with this framework gradually began to build, with the realization that the simple cause-and-effect scenarios took as exogenous much of what was most historically interesting. An essential next step was to broaden the concept of institutions, to include not only explicit and state-enforced rules, but also informal and uncodified constraints on behavior. To say that institutions are 'constraints', on individual behavior, however, does not mean that they must be seen as having been 'imposed' on the society in question. By analogy with market prices, institutions may be viewed as exogenous to the individuals, yet endogenous to the society as a whole. When the underlying conception has been broadened in this way, a natural theoretical framework is provided by game theory, depicting institutions as self-enforcing equilibria arrived at through the interaction of historical players. Avner Greif (1994, 1997) has been a pioneer in the application of game theory to historical institutions, using these tools to analyze the bases for the rise of long-distance trade in late medieval Europe.

Greif argues that appropriate historical uses of game theory must be context specific. In many strategic situations there are multiple equilibria, implying that outcomes are not uniquely predetermined. Constructing a context-specific model that captures the essence of a situation requires microlevel historical study, to provide the specific details of the model, and to facilitate the task of confronting predictions with historical evidence. Thus, the process has a large inductive component, drawing on cultural, political, and social factors and also upon economic motivations. Although this specificity may seem to limit the potential scope for application, Greif (1994) was able to use these tools to extract sweeping implications from the contrasting cultural beliefs of the Maghribi traders (who used collective social instruments to discipline the behavior of their members) and the Genoese (from whose rule-based system the Western legal traditions descend). Because institutional evolution typically entails individual behaviors that are interdependent, such models often exhibit increasing returns and path dependence, similar to patterns observed in technological systems subject to network externalities.

In many ways, modern quantitative economic history has returned to the roots of economic history, in its renewed appreciation of institutions and culture. But there is a major difference. Whereas earlier versions of 'institutional economics' tended to be 'anti-theory', stressing the sheer weight of historical reality as contrasted with abstract rationalizations, the newer forms of institutional economic history are firmly lodged within the parent discipline, drawing on frontier developments in economic theory as well as quantitative methods. The ability to deploy more eclectic theoretical tools allows cliometricians to extend their reach into new subject areas, that previously were preserves of purely descriptive work. Examples include historical studies of financial markets and financial institutions, which now use theories of information and principle-agent relationships; and the history of labor markets and institutions, where the roles of norms, prejudices and family institutions are too prominent to be ignored.

An illustration of this progression is Claudia Goldin's book on the economic history of American women (Goldin 1990). The early chapters developed new quantitative evidence on relative wages and labor force participation rates of American women, interpreting two major upward surges (1820–1850 and 1890–1930) in terms of supply and demand forces, driven largely by technology. By the latter half of the book, however, Goldin found that 'institutional' aspects of the record were inescapable, from the 'marriage bars' of the 1920s and 1930s 'which forced women to resign from jobs upon marriage' to patterns of socialization and peer influence, detectable in behavioral shifts from cohort to cohort across the twentieth century. Significantly, these insights were not advanced as a rejection of quantitative economics, but as an enrichment of the analytical and quantitative tools within the core disciplines of economics and history.

Reflecting on her own evolution, Goldin wrote: 'I began this study more as an economist but have ended with a fuller appreciation of how the distant past affects the present, how norms and expectations impede change, how discrimination can survive even in highly competitive markets, and how slow genuine change can be.' Much of Goldin's personal account

could be applied to economic history as a whole, in its increasingly self-conscious identity, not as a branch of applied economics, but as a distinctive approach to the study of economic life in general.

See also: Economic Growth: Measurement; Historical Demography; Historical Geography; Quantification in History; Quantification in the History of the Social Sciences

Bibliography

Brady D S (ed.) 1966 *Output, Employment, and Productivity in the United States after 1800*. Studies in Income and Wealth, Vol. 30. National Bureau of Economic Research, New York

David P A, Gutman H G, Sutch R, Temin P, Wright G 1976 *Reckoning with Slavery: A Critical Study in the Economics of American Negro Slavery*. Oxford University Press, New York

David P A, Wright G 1997 Increasing returns and the genesis of American resource abundance. *Industrial and Corporate Change* **6**: 203–45

Davis L E, Easterlin R A, Parker W N, Brady D S, Fishlow A, Gallman R E, Lebergott S, Lipsey R E, North D C, Rosenberg N, Smolensky E, Temin P 1972 *American Economic Growth: An Economist's History of the United States*. Harper and Row, New York

Fishlow A 1965 *American Railroads and the Transformation of the Ante-Bellum Economy*. Harvard University Press, Cambridge, MA

Fogel R W 1964 *Railroads and American Economic Growth: Essays in Econometric History*. Johns Hopkins University Press, Baltimore, MD

Fogel R W, Engerman S E (eds.) 1972 *The Reinterpretation of American Economic History*. Harper and Row, New York

Fogel R W, Engerman S E 1974 *Time on the Cross: The Economics of American Negro Slavery*. Little, Brown, Boston

Fogel R W 1989 *Without Consent or Contract: The Rise and Fall of American Slavery*. W W Norton, New York

Goldin C D 1990 *Understanding the Gender Gap: An Economic History of American Women*. Oxford University Press, New York

Greif A 1994 Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy* **102**: 912–50

Greif A 1997 Microtheory and recent developments in the study of economic institutions through economic history. In: Kreps D M, Wallis K F (eds.) *Advances in Economic Theory, II*. Cambridge University Press, Cambridge, UK, pp. 79–113

Habakkak H J 1962 *American and British Technology in the Nineteenth Century*. Cambridge University Press, Cambridge, UK

Koot G M 1987 *English Historical Economics, 1870–1926: The Rise of Economic History and Neomercantilism*. Cambridge University Press, Cambridge, UK

Mokyr J M 1990 *The Lever of Riches: Technological Creativity and Economic Progress*. Oxford University Press, London

North D C 1961 *The Economic Growth of the United States, 1790–1860*. Prentice-Hall, Englewood Cliffs, NJ

North D C 1981 *Structure and Change in Economic History*. W W Norton, New York

North D C 1990 *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, New York

North D C, Thomas R P 1973 *The Rise of the Modern World: A New Economic History*. Cambridge University Press, Cambridge, UK

Parker W N 1960 *Trends in the American Economy in the Nineteenth Century*. Studies in Income and Wealth, Vol. 24. National Bureau of Economic Research, New York

Parker W N 1986 editor *Economic History and the Modern Economist*. Blackwell, New York

Rosenberg N 1994 *Inside the Black Box: Technology and Economics*. Cambridge University Press, Cambridge, UK

Tadman M 1989 *Speculators and Slaves: Masters, Traders and Slaves in the Old South*. University of Wisconsin Press, Madison, WI

G. Wright

Economic Panel Data

1. Introduction

A panel (or longitudinal or temporal cross-sectional) data set is one which follows a number of individuals over time, and thus provides multiple observations on each individual in the sample. Prominent examples are the University of Michigan's Panel Study of Income Dynamics, the National Longitudinal Surveys of Labor Market Experience, and National Longitudinal Survey of Youth in the US and the Social Economic Panel, the Expenditure Index Panel of Intromart, and the Labor Mobility Survey from the Organization of Strategic Labor Market Research in the Netherlands. These labor market data samples contain thousands of individuals and variables followed over a number of years. In addition, panel data are also common in marketing studies, biomedical sciences, and financial market analysis, etc.

The increasing popularity of panel studies is partly a consequence of more cost effective ways of developing and maintaining panels (e.g., Mátyás and Sevestre 1996). More importantly, panel data offers many more possibilities for exploring analytical and substantive issues than purely cross-sectional or time series data. However, new data sources also raise new issues. This entry reviews the major advantages and limitations of panel data in the context of specific econometric methodologies. Section 2 gives an overview of the major advantages of the informational content of panel data. Section 3 reviews typical specifications for the linear models. Section 4 discusses issues of nonlinear models. Section 5 considers the implication of sample attrition and sample selection. Conclusions are in Sect. 6.

2. Advantages of Panel Data

Panel data offer several advantages over a single cross-sectional or time series data. For instance, the use of panel data may:

2.1 Improve the Accuracy of Parameter Estimates

Most statistical estimators converge to the true parameters at the speed of the square root of the degrees of freedom. Models for panel data are based on many more equivalent observations than a cross-sectional or time series data, and hence can yield more accurate parameter estimates.

2.2 Lessen the Problem of Multicollinearity

Many economic variables tend to move collinearly over time. The shortage of degrees of freedom and severe multicollinearity problems found in time series data often frustrate investigators who wish to determine the individual influences of each explanatory variable. Panel data may lessen the degree of multicollinearity in the time dimension by appealing to inter-individual differences.

2.3 Allow the Specification of More Complicated Behavioral Hypotheses

A single time series is not useful for discriminating hypotheses which depend on different social–demographic factors. Cross-sectional data cannot be used to model dynamics. Cross-sectional data can show what proportion of the labor force is unemployed or describe the distribution of wage rates at a point in time. Aggregate time series data can provide indications of general trends and cyclical patterns in unemployment, wages, income, and so forth. But neither cross-sectional nor time series data can provide information regarding how many of those unemployed in one month can find employment in the next month. Neither can the data provide information on whether the tendency of welfare families to stay on welfare occurs simply because the same factors that cause them to be on welfare keep them there or whether, in addition, the experience of receiving welfare has some addictive effect that induces continuing welfare dependence. For instance, suppose that a cross-sectional sample of married women is found to have an average yearly labor force participation rate of 50 percent. At one extreme this might be interpreted as implying that each woman in a homogenous population has a 50 percent chance of being in the labor force in any given year while at the other extreme it might imply that 50 percent of the women in a heterogenous population always work and 50 percent never work. In the first case, each woman would be expected to spend half of her married life in the labor force, and half out of the labor force, and job turnover would be expected to be frequent, with an average job duration of two years. In the second case, there is no turnover, and current information about work status is a perfect predictor of future work status. Panel data, by providing sequential observations for a number of individuals, offer the possibility to distinguish inter-individual differences from intra-individual dynamics and to construct a proper recursive structure to study the effects of changes by comparing the differences before and after the changes (e.g., Heckman 1981).

2.4 Provide Possibilities for Reducing Estimation Bias

A standard assumption in statistical specification is that the random error term representing the effect of omitted variables is orthogonal to the included explanatory variables. Otherwise, the estimates are subject to omitted variable bias when these correlations are not explicitly allowed for. Panel data provide means to eliminate or reduce the omitted variable bias through various data transformations when the correlations between included explanatory variables and omitted variables follow certain specific patterns (e.g., Baltagi 1995), Chamberlain 1984, Hsiao (1986). For instance, suppose that

$$y_{it} = \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \underset{\sim}{\gamma}' \underset{\sim}{z}_i + u_{it}, \quad i = 1, \ldots, N, \; t = 1, \ldots, T \quad (1)$$

where the explanatory variables $\underset{\sim}{x}_{it}$, vary across individuals i, and over time t, and the explanatory variables $\underset{\sim}{z}_i$, vary across i, but stay constant over t, such as sex status, social-demographic background variables, etc. the error term u is assumed to be uncorrelated with $\underset{\sim}{x}$ or $\underset{\sim}{z}$. The least squares regression of y on x and z will yield consistent estimates of $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$. But if $\underset{\sim}{z}$ are correlated with $\underset{\sim}{x}$, yet are unobservable, the least squares regression of y on $\underset{\sim}{x}$ will yield bias estimates of $\underset{\sim}{\beta}$. However, in the pairwise difference,

$$y_{it} - y_{i,\,t-1} = \underset{\sim}{\beta}' (\underset{\sim}{x}_{it} - x_{i,\,t-1}) + (u_{it} - u_{i,\,t-1}) \quad (2)$$

the explanatory variables $\underset{\sim}{z}_i$ longer enter. Regressing $(y_{it} - y_{i,\,t-1})$ on $(\underset{\sim}{x}_{it} - x_{i,\,t-1})$ can yield consistent estimates of $\underset{\sim}{\beta}$.

MaCurdy's (1981) life-cycle labor supply of prime-age males under certainty is an example of this approach. Under certain simplifying assumptions, MaCurdy shows that a worker's labor supply function can be written as in Eqn. (1), where y is the logarithm of hours worked, $\underset{\sim}{x}$ is the logarithm of the real wage rate, and z is the logarithm of the worker's (unobserved) marginal utility of initial wealth, which, as a summary measure of worker's lifetime wages and property income, is assumed to stay constant through time but to vary across individuals. Given the economic problem, not only is $\underset{\sim}{x}_{it}$ correlated with z_i, but

every economic variable that could act as an instrument for $\underset{\sim}{x}_{it}$ (such as education) is also correlated with $\underset{\sim}{z}_i$. Another example is the 'conditional convergence' of the growth rate. The growth rate regression model typically includes investment ratio, initial income, and measures of policy outcomes like school enrolment and the black market exchange rate premium as regressors; however, an important component, the initial level of a country's technical efficiency z_{i0}, is not observed. Since a country that is less efficient is also more likely to have lower investment rate or school enrolment, one can easily imagine that z_{i0} is correlated with the included regressors. Neither model can be estimated with cross-sectional data, but both can be estimated using Eqn. (2) if panel data are available.

2.5 Reduce Computational Complexities

The use of panel data is often perceived as increasing computational complexities because the data contain both cross-sectional and time series dimensions. In many cases, however, panel data become a distinct advantage computationally. For instance, if one were to estimate a dynamic model using time series data with censored or missing observations in between, the computation can be horrendous because of the need to compute multiple integrals when the conditional variables are censored. To see this, consider the model

$$y^*_{it} = \gamma y^*_{i,\,t-1} + u_{it} \tag{3}$$

where u_{it} is independent normal with mean 0 and variance σ^2_u. Conditional on $y^*_{i,\,t-1}$, y^*_{it} is normally distributed with mean $\gamma y^*_{i,\,t-1}$ and variance σ_u. If instead of observing y^*_{it}, however, one observes the censored counterpart y_{it}, where $y_{it} = y^*_{it}$ if $y^*_{it} > 0$ and $y_{it} = 0$ if $y^*_{it} \leqslant 0$, then the conditional distribution of y_{it} given $y_{i,\,t-1} = 0$ becomes

$$\int_{-\infty}^{0} (2\pi\sigma_n^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma_u^2}(y_{it} - \gamma y^*_{i,\,t-1})^2\right\} f(y^*_{i,\,t-1})\,\mathrm{d}y^*_{i,\,t-1} \tag{4}$$

where $f(y^*_{i,\,t-1})$ denotes the marginal density of $(y^*_{i,\,t-1})$. When there are a number of censored observations over time, Eqn. (4) becomes relatively complicated to solve. With panel data, one can often adopt models that ignore observations that are censored and focus on those individuals for which successive observations are greater than zero to simplify the computation (e.g., Arellano et al. 1999).

2.6 Simplify Statistical Inference

In estimating a dynamic model using time series, the limiting distribution depends on whether the system is stable, integrated, or explosive. With panel data, as long as the cross-sectional dimension approaches infinity, the limiting distribution of the coefficients of lagged dependent variables is typically normal no matter what the time series properties of a process are. The limiting distribution remains the same whether the time series dimension is fixed (Anderson and Hsiao 1982) or approaches infinity (Levin and Lin 1993, Phillips and Moon 1999).

2.7 Obtain More Accurate Prediction of Individual Outcomes

When people's behavior is similar in nature, namely, they satisfy De Finnetti's (1964) exchangeability assumption, each individual's behavior can be better understood by observing the behavior of other similar individuals, hence can yield more accurate predictions of individual outcomes by pooling the data (e.g., Hsiao et al. 1989).

3. Linear Models

Panel data, by its nature, put emphasis on individual outcomes (see *Longitudinal Data*). Different individuals may be subject to different influences. In explaining individual behavior, one may extend the list of factors *ad infinitum*. It is neither feasible nor desirable to include all the factors affecting the outcome of all different individuals in a model specification. It is typical to leave out those factors that are believed to have insignificant impact or are peculiar to a given individual. However, when important factors peculiar to a given individual are left out, the individual outcomes conditional on the included explanatory variables may not be viewed as random draws from a common population, so standard statistical procedures may mislead. Therefore, one focus of panel data modeling is to determine whether heterogeneities exist among cross-sectional units over time that are not captured by the included explanatory variables, say $\underset{\sim}{x}$. If they are, one needs to control the impact of such unobserved heterogeneity in order to draw inference about the population characteristics.

If the unobserved heterogeneity among N cross-sectional units over T time periods can not be completely captured by $\underset{\sim}{x}$, there are two primary ways to model it. One is to incorporate it into the error term. The other is to let the coefficients vary across individuals and over time. For instance, if the heterogeneity among individual units is due to some individual specific and time invariant variables, $\underset{\sim}{z}_i$, that affect y_{it} linearly, then the scatter diagram between y and x conditional on $\underset{\sim}{z}$ may look like Fig. 1 in which the broken-line circles represent the point scatter for an individual over time, and the broken straightline represents the mean relationship for the ith individual

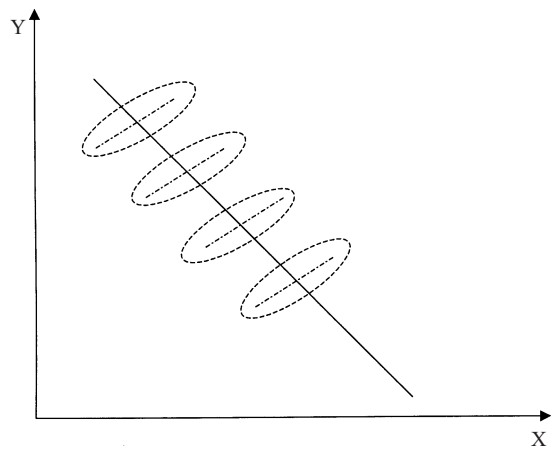

Figure 1
Scatter diagram of y and x conditional on z

conditional on $\underset{\sim}{z}_i$. The dark line indicates a simple pooled estimates of a model

$$y_{it} = \alpha + \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + v_{it} \tag{5}$$

where v represents the effects of omitted factors and the subscript, it, indicates observations of the ith cross-sectional unit at tth time period. It is obvious that the simple pooled estimates give a very misleading relationship between y and $\underset{\sim}{x}$. For data of the form of Fig. 1 it is necessary to specify a variable intercept model of the form

$$y_{it} = \alpha_i + \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + u_{it} \tag{6}$$

where α_i represents the effect of unobserved $\underset{\sim}{z}_i$.

A more complicated structure to capture the unobserved heterogeneity lets

$$y_{it} = \alpha_{it} + \underset{\sim}{\beta}'_{it} \underset{\sim}{x}_{it} + u_{it} \tag{7}$$

but this specification is not very useful for drawing inferences about the population because the coefficients are individual and time specific. Neither can it be estimated because there are more unknown parameters than degrees of freedom. In order to take advantage of panel data, a structure has to be introduced that allows the possibility to draw inference about the population characteristics while controlling for the impact of unobserved heterogeneity.

Suppose that the unobserved heterogeneity is due to some individual specific but time invariant variables, then an unrestricted linear model will have the form

$$y_{it} = \alpha_i + \underset{\sim}{\beta}'_i \underset{\sim}{x}_{it} + u_{it} \tag{8}$$

If T is large, Eqn. (8) can be estimated by Zellner's (1962) seemingly unrelated regression method; however, most panel data contains a large number of individuals observed over a short period of time. When T is fixed, increasing N increases the number of unknown parameters, α_i and $\underset{\sim}{\beta}_i$, $i = 1, \ldots, N$. This is the classical incidental parameter problem. The presence of the incidental parameters violates the usual regularity conditions for the consistency of the maximum likelihood estimator (MLE). Neither is Eqn. 3.4 very useful to draw inference about the population because each individual has a different behavioral pattern. Assuming Eqn. (8) is tantamount to assuming that the cross-sectional observations come from heterogeneous population. In this case there is no particularly appealing reason to pool the data.

To draw inference about the population, one must impose restrictions on Eqn. (7) or Eqn. (8). For instance, if the slope coefficients are the same, $\beta_i = \beta$, for all i as in Eqn. (6), the incidental parameter α_i can be eliminated by pairwise differencing of individual time series observations as in Eqn. 2 or by taking deviations from the individual time series mean:

$$(y_{it} - \bar{y}_i) = \underset{\sim}{\beta}(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i) + (u_{it} - \bar{u}_i) \tag{9}$$

where $\bar{y}_i = (1/T)\sum_{t=1}^{T} y_{it}$, $\bar{\underset{\sim}{x}}_i = (1/T)\sum_{t=1}^{T} \underset{\sim}{x}_{it}$ and $\bar{u}_i = (1/T)\sum_{t=1}^{T} u_{it}$. Applying the least squares method to Eqn. (9) yields the so-called covariance or within estimator of β. The covariance estimator of β is consistent when either N or T or both N and T tend to infinity (e.g., Hsiao 1986).

Alternatively, one can assume that $\underset{\sim}{\beta}_i$ and/or α_i in Eqn. (8) are randomly distributed with common mean $(\underset{\sim}{\beta}', \alpha)$ and covariance matrix Δ or assume that some of the coefficients are randomly distributed with common mean and constant covariance matrix, while others are fixed and different. Decomposing $\underset{\sim}{x}_{it}$ into $(\underset{\sim}{w}_{it}, \underset{\sim}{z}_{it})$, various formulations can be represented as placing different restrictions on the unconstrained model Eqn. (8):

$$\underset{\sim}{y} = W\underset{\sim}{\delta} + Z\underset{\sim}{\gamma} + \underset{\sim}{u} \tag{10}$$

where $\underset{\sim}{y} = (\underset{\sim}{y}'_1, \ldots, \underset{\sim}{y}'_N)$, $W = \mathrm{diag}(W_i)$, $Z = \mathrm{diag}(Z_i)$, $\underset{\sim}{y}_i = (\underset{\sim}{y}_{i1}, \ldots, \underset{\sim}{y}_{iT})'$, W_i and Z_i are $T \times k_1$ and $T \times k_2$ submatrices of T observed values of $(1, \underset{\sim}{x}'_{it})$, $\underset{\sim}{\delta} = (\underset{\sim}{\delta}'_1, \ldots, \underset{\sim}{\delta}'_N)'$ and $\underset{\sim}{\gamma} = (\underset{\sim}{\gamma}'_1, \ldots, \underset{\sim}{\gamma}'_N)$ are $Nk_1 \times 1$ and $Nk_2 \times 1$ vectors of $(\underset{\sim}{\alpha}'_i, \underset{\sim}{\beta}'_i)$, $i = 1, \ldots, N$. Suppose that the coefficients of $\underset{\sim}{w}_{it}$ are subject to stochastic constraints:

$$\underset{\sim}{\delta} = A_1 \bar{\underset{\sim}{\delta}} + \underset{\sim}{\varepsilon} \tag{11}$$

where A_1 is an $Nk_1 \times m$ matrix with known elements, $\bar{\underset{\sim}{\delta}}$ is an $m \times 1$ vector of constants, $\underset{\sim}{\varepsilon}$ is random with mean $\underset{\sim}{0}$, covariance matrix C and $E\underset{\sim}{\varepsilon}W' = \underset{\sim}{0}$, $E\underset{\sim}{\varepsilon}Z' = \underset{\sim}{0}$.

Suppose that the coefficients of z_{it} are subject to deterministic constraints:

$$\gamma = A_2\bar{\gamma} \tag{12}$$

where A_2 is an $Nk_2 \times q$ matrix with known elements, and $\bar{\gamma}$ is a $q \times 1$ vector of constants.

The model given by Eqns. (10)–(12) is of a mixed form: the coefficients of W_i are assumed to be random and the coefficients of Z_i are assumed fixed, (e.g., Hsiao and Tahmiscioglu 1997). When $W = 0$, $A_2 = \varrho_N \otimes I_{k_2}$, we have a common model for all NT observations, where ϱ_N denotes an $N \times 1$ vector of ones, I_p denotes a $p \times p$ identity matrix and $\otimes$ denotes Kronecker product. When $W = 0$ and $A_2 = I_{Nk_2}$, model Eqn. (8) results. when $W = 0$, $Z_i = (\varrho_T, X_i)$, $A_2 = (I_N \otimes \iota_{k_2} : \varrho_N \otimes I^*_{k_2-1})$, ι_{k_2} is a $k_2 \times 1$ vector of $(1, 0, \ldots, 0)'$, and $I^*_{k_2-1}$ is a $k_2 \times (k_2 - 1)$ matrix of $\binom{0'}{I_{k_2-1}}$, model Eqn. (6) (e.g., Mundlak 1978) is a special case. When $X_i = \varrho_T$, $A_1 = \varrho_N$, $\Delta = \sigma^2_\delta I_N$, the model is the error components model (e.g., Balestra and Nerlove 1966). Finally when $Z = 0$, $A_1 = \varrho_N \otimes I_{k_1}$, $C = I_N \otimes \Delta$, we have the random coefficients model (e.g., Swamy 1970).

Substituting Eqns. (11) and (12) into Eqn. (10) yields

$$y = WA_1\bar{\delta} + ZA_2\bar{\gamma} + v \tag{13}$$

where $v = W\varrho + u$, $Evv' = WCW' + \mathrm{Cov}(u)$. Model Eqn. (13) can be estimated by the generalized least squares (GLS) method. The individual coefficients δ can be estimated by a Bayes estimator which is a weighted average of the GLS estimator of $\bar{\delta}$ and the individual least squares estimator of δ (e.g., Hsiao et al. 1993).

The issue of whether to treat β_i and/or α_i or a subvector of them as fixed or random is an issue of whether conditional on x, one can treat the outcome variable as a random draw from a common population. If it is, then the individual and time subscript, it, can be viewed as purely a labeling device and *a priori*, $E(y_{it} \mid x) = E(y_{js} \mid x)$. This is what de Finnetti (1964) called the exchangeability criterion and it is more appropriate to treat β_i and α_i as random. If the distribution of β_i is not independent of x_{it} or conditional on x, individual outcomes are more appropriately viewed as generated from a heterogenous population, then the subscript, *it*, contains specific information about which heterogenous population the particular observation is generated from. It is therefore more appropriate to treat β_i and/or α_i as fixed parameters. Unfortunately, *a priori*, very little knowledge is available about which coefficients should be treated as random and different, and which coefficients should be treated as fixed and different. Hsiao and Sun (2000) have argued that the random versus fixed effects specification should be viewed as a model selection issue and suggest a Bayesian approach for model selection. Their limited Monte Carlo studies appear to favor the Schwarz (1978) information criterion to select the appropriate specifications. The Schwarz criterion works here because although from a Bayesian perspective there is no real difference between random effects and fixed effects since every parameter has a distribution, there is a difference between how the prior is formulated under the random and fixed effects formulation. Under the di Finnetti exchangeability criterion, β_i is distributed independently of x_i with a common mean $\bar{\beta}$ and constant variance C (e.g., Eqn. (11)). On the other hand, under the fixed effects formulation, each β_i is supposed to be independently distributed with a different mean and different variance because of the correlation between β_i and x_i or because each β_i is a random draw from heterogenous population.

4. Nonlinear Models

The estimation of a nonlinear panel data model involving individual specific effects is particularly difficult to handle. If the individual specific effects are treated as random, just as in the linear model, we no longer have the incidental parameters problem; however, the estimation of the structural parameters requires integrating out the individual specific effects which can be nontrivial. For instance, consider a binary choice model in which y_{it} takes the value of either 1 or 0. Suppose that the probability of observing $y_{it} = 1$, $P(y_{it} = 1 \mid x_{it})$, is given by $F(\beta' x_{it} + \alpha_i)$, where the individual specific effects α_i is randomly distributed with density function $f(\alpha_i)$. Even though conditional on α_i, y_{it} may be independently distributed over i and t, unconditionally, y_{it} is correlated over T. The log-likelihood function is of the form

$$\log L = \sum_{i=1}^{N} \log \int \prod_{t=1}^{T} F(\beta' x_{it} + \alpha)^{y_{it}} [1 - F(\beta' x_{it} + \alpha)]^{1-y_{it}} f(\alpha)\, \mathrm{d}\alpha \tag{14}$$

Equation (14) is a function of finite number of parameters and maximizing it under weak regularity conditions will yield consistent estimator of β (and the parameters characterizing the density function of α_i). However, maximizing (Eqn. (14)) is often computationally infeasible. For instance, if $F(\cdot)$ takes the form of an integrated standard normal distribution function, then a probit model results. Conditional on α_i it involves a univariate integration. Maximizing (Eqn. (14)) but without the conditioning involves the evaluation of T-dimensional integrals (e.g., Hsiao 1986).

On the other hand, suppose that the individual specific effects are fixed. First, in general there is no simple transformation of the data as in Eqns. (2) or (9) to eliminate the individual specific effects (incidental parameters). Second, the estimation of the individual specific parameters and the structural (or common) parameters are not independent of each other. When a panel contains a large number of individuals but only over a short time period, the error in the estimation of the individual specific coefficients propagates into the estimating of the structural parameters, and hence leads to inconsistency of the structural parameter estimation (e.g., Anderson and Hsiao 1982).

Many approaches have been suggested for estimating fixed effects nonlinear panel data models. One approach is by conditioning the likelihood function on the minimum sufficient statistics for the incidental parameters to eliminate the individual specific parameters (Andersen 1973, Chamberlain 1980). For instance, consider a binary logit model (e.g., see *China: Sociocultural Aspects*) in which the probability of $y_{it} = 1$ is given by

$$P(y_{it} = 1) = \frac{e^{\alpha_i + \beta' x_{it}}}{1 + e^{\alpha_i + \beta' x_{it}}} \tag{15}$$

For ease of exposition, assume that $T = 2$. It can be shown that when $x_{i1} = 0$ and $x_{i2} = 1$ the MLE of β converges to 2β as $N \to \infty$, which is not consistent (e.g., Hsiao 1986). However, one may condition the likelihood function on the minimal sufficient statistic for α_i . We note that those individuals i satisfying $y_{i1} + y_{i2} = 2$ or $y_{i1} + y_{i2} = 0$ provide no information about β since the former leads to $\hat{\alpha}_i = \infty$ and the latter leads to $\hat{\alpha}_i = -\infty$. Hence one can concentrate on those individuals with $y_{i1} + y_{i2} = 1$. For those individuals satisfying $y_{i1} + y_{i2} = 1$, let $d_i = 1$ if $y_{i1} = 0$ and $y_{i2} = 1$ and $d_i = 0$ if $y_{i1} = 1$ and $y_{i2} = 0$. Then

$$\text{Prob}(d_i = 1 \mid y_{i1} + y_{i2} = 1) = \frac{e^{\beta'(x_{i2} - x_{i1})}}{1 + e^{\beta'(x_{i2} - x_{i1})}} \tag{16}$$

Eqn. (16) no longer contains the individual effects α_i and is of the form of a standard logit model. Therefore, maximizing the conditional log-likelihood function on the subset of individuals with $y_{i1} + y_{i2} = 1$ is consistent and asymptotically normally distributed.

Alternatively, one can transform the data to eliminate the individual effects if the nonlinear model has a latent linear structure, and then apply semiparametric methods. For instance, consider a binary choice model with $y_{it} = 1$ if $y_{it}^* > 0$ and $y_{it} = 0$ if $y_{it}^* \leqslant 0$, where

$$y_{it}^* = \alpha_i + \beta' x_{it} + u_{it} \tag{17}$$

If u_{it} follows a logistic distribution, a logit model results; if u_{it} is normally distributed, a probit model results. Then

$$\begin{aligned} \beta' x_{it} > \beta' x_{i,t-1} &\Leftrightarrow E(y_{it} \mid x_{it}) > E(y_{i,t-1} \mid x_{i,t-1}) \\ \beta' x_{it} = \beta' x_{i,t-1} &\Leftrightarrow E(y_{it} \mid x_{it}) = E(y_{i,t-1} \mid x_{i,t-1}) \\ \beta' x_{it} < \beta' x_{i,t-1} &\Leftrightarrow E(y_{it} \mid x_{it}) < E(y_{i,t-1} \mid x_{i,t-1}) \end{aligned} \tag{18}$$

Rewriting (Eqn. (18)) into the equivalent first difference form, Manski (1975) proposes a maximum score estimator (MS) that maximizes the sample average function

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T} sgn[(x_{it} - x_{i,t-1})'\beta](y_{it} - y_{i,t-1}) \tag{19}$$

where $sgn[(x_{it} - x_{i,t-1})'\beta] = 1$ if $(x_{it} - x_{i,t-1})'\beta \geqslant 0$ and $sgn[(x_{it} - x_{i,t-1})'\beta] = -1$ if $(x_{it} - x_{i,t-1})\beta < 0$. The MS estimator is consistent but is not root-n consistent, where $n = N(T-1)$. Its rate of convergence is $n^{1/3}$ and $n^{1/3}(\hat{\beta} - \beta)$ converges to a random variable that maximizes a certain Gaussian process.

A third approach is to find an orthogonal reparameterization of the fixed effects for each individual, say α_i, to a new fixed effects, say g_i, which are independent of the structural parameters in the information matrix sense. The g_i are then integrated out of the likelihood with respect to an uninformative, uniform prior distribution which is independent of the prior distribution of the structural parameters. Lancaster (2001) uses a two period duration model to show that the marginal posterior density of the structural parameter possesses a mode which consistently estimates the true parameter.

While all these methods are quite ingenious, unfortunately, none can claim general applicability. For instance, the conditional maximum likelihood cannot work for the probit model because there are no minimal sufficient statistics for the incidental parameters that are independent of the structural parameters. The data transformation approach cannot work if the model does not have a latent linear structure. The orthogonal reparameterization approach works only for particular models. In short, the methodology for and consistency of nonlinear panel data estimators must be established case by case.

5. *Sample Attrition and Sample Selection*

Missing observations occur frequently in panel data. If individuals are missing randomly, most estimation methods for the balanced panel can be extended in a straightforward manner to the unbalanced panel (e.g., Hsiao 1986). For instance, suppose that

$$d_{it} y_{it} = d_{it}[\alpha_i + \beta' x_{it} + u_{it}] \tag{20}$$

where d_{it} is an observable scalar indicator variable which denotes whether information about (y_{it}, x_{it}') for

the ith individual at tth time period is available or not. The indicator variable d_{it} is assumed to depend on a q-dimensional variables, w_{it}, individual specific effects λ_i and an unobservable error term η_{it},

$$d_{it} = I(\lambda_i + \underset{\sim}{\delta}' \underset{\sim}{w}_{it} + \eta_{it} > 0) \tag{21}$$

where $I(\cdot)$ is the indicator function that takes the value of 1 if $\lambda_i + \underset{\sim}{\delta}' \underset{\sim}{w}_{it} + \eta_{it} > 0$ and 0 otherwise. In other words, the indicator variable d_{it} determines whether $(y_{it}, \underset{\sim}{x}_{it})$) in Eqn. (20) is observed or not.

Without sample selectivity, that is $d_{it} = 1$ for all i and t, Eqn. (20) is the standard variable intercept (or fixed effects) model for panel data discussed in Sect. 2. With sample selection and if η_{it} and u_{it} are correlated, $E(u_{it} \mid \underset{\sim}{x}_{it}, d_{it} = 1) \neq 0$. Let $\theta(\cdot)$ denote the conditional expectation of u_{it} conditional on $d_{it} = 1$ and $\underset{\sim}{x}_{it}$, then Eqn. (20) can be written as

$$y_{it} = \alpha_i + \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \theta(\lambda_i + \underset{\sim}{\delta}' \underset{\sim}{w}_{it}) + \varepsilon_{it} \tag{22}$$

where $E(\varepsilon_{it} \mid \underset{\sim}{x}_{it}, d_{it} = 1) = 0$. The form of the selection function is derived from the joint distribution of u and η. For instance, if u and η are bivariate normal, then we have the Heckman (1979) sample selection model with

$$\theta_{it} = \theta(\lambda_i + \underset{\sim}{\delta}' \underset{\sim}{w}_{it}) = \sigma_{w\eta} \frac{\phi(\lambda_i + \underset{\sim}{\delta}' \underset{\sim}{w}_{it})}{\Phi(\lambda_i + \underset{\sim}{\delta}' \underset{\sim}{w}_{it})} \tag{23}$$

where $\sigma_{w\eta}$ denotes the covariance between w and η, $\phi(\cdot)$ and $\Phi(\cdot)$ are integrated standard normal density and distribution, respectively and the variance of η is normalized to be 1. Therefore, in the presence of sample attrition or selection, regressing y_{it} on $\underset{\sim}{x}_{it}$ using only the observed information is invalidated by two problems—first, the presence of the unobserved effects α_i, and second, the 'selection bias' arising from the fact that $E(u_{it} \mid \underset{\sim}{x}_{it}, d_{it} = 1) = \theta(\lambda_i + \underset{\sim}{\delta} \underset{\sim}{w}_{it})$.

When individual effects are random and the joint distribution function of $(u, \eta, \gamma_i, \lambda_i)$ is known, both maximum likelihood and two or multi-step estimators can be derived (e.g., Heckman 1979). The resulting estimators are consistent and asymptotically normally distributed. The speed of convergence is proportional to the square root of the sample size. However, if the joint distribution of u and η is mis-specified, then even without the presence of α_i, both the maximum likelihood and Heckman (1979) two step estimators are inconsistent. This sensitivity of parameter estimate to the exact specification of the error distribution has motivated interest in semiparametric methods.

The individual specific effects in Eqn. (20) are easily eliminated by time differencing for those individuals that are observed for two time periods t and s, i.e., those individuals i satisfying $d_{it} = d_{is} = 1$. However, the sample selectivity factors are not eliminated by time differencing. Ahn and Powell (1993) note that if θ is a sufficiently 'smooth' function, and $\hat{\underset{\sim}{\delta}}$ is a consistent estimator of $\underset{\sim}{\delta}$, observations for which the difference $(\underset{\sim}{w}_{it} - \underset{\sim}{w}_{is})' \hat{\underset{\sim}{\delta}}$ is close to zero should have $\theta_{it} - \theta_{is} \simeq 0$. They propose a two-step procedure. In the first step, consistent semiparametric estimates of the coefficients of the 'selection' equation are obtained. The result is used to obtain estimates of the 'single index, $\underset{\sim}{w}_{it}\hat{\underset{\sim}{\delta}}$,' variables characterizing the selectivity bias in the index equation. The second step of the approach estimates the parameters of the equation of interest by a weighted instrumental variables regression of pairwise differences in dependent variables in the sample on the corresponding differences in explanatory variables: the weights put more emphasis on pairs with $\underset{\sim}{w}_{it}'\hat{\underset{\sim}{\delta}} \simeq \underset{\sim}{w}_{i,t-1}'\hat{\underset{\sim}{\delta}}$.

Kyriazidou (1997) generalizes this concept and proposes to estimate the fixed effects sample selection models in two steps: in the first step, estimate $\underset{\sim}{\delta}$ by either Chamberlain (1980) conditional maximum likelihood approach or the Manski (1975) maximum score method. In the second step, the estimated $\hat{\underset{\sim}{\delta}}$ is used to estimate γ, based on pairs of observations for which $d_{it} = d_{is} = 1$ and for which $(\underset{\sim}{w}_{it} - \underset{\sim}{w}_{is})' \hat{\underset{\sim}{\delta}}$ is 'close' to zero. This last requirement is operationalized by weighting each pair of observations with a weight that depends inversely on the magnitude of $(\underset{\sim}{w}_{it} - \underset{\sim}{w}_{is})' \hat{\underset{\sim}{\delta}}$, so that pairs with larger differences in the selection effects receive less weight in the estimation. The Kyriazidou (1997) estimator takes the form

$$\hat{\underset{\sim}{\beta}} = \left\{ \sum_{i=1}^{N} \sum_{1 \leqslant s < t \leqslant T_i} (\underset{\sim}{x}_{it} - \underset{\sim}{x}_{is})(\underset{\sim}{x}_{it} - \underset{\sim}{x}_{is})' K \left[\frac{(\underset{\sim}{w}_{it} - \underset{\sim}{w}_{is})' \hat{\underset{\sim}{\delta}}}{h_N} \right] d_{it} d_{is} \right\}^{-1} \left\{ \sum_{i=1}^{N} \sum_{1 \leqslant s < t < T_i} (\underset{\sim}{x}_{it} - \underset{\sim}{x}_{is})(y_{it} - y_{is}) K \left[\frac{(\underset{\sim}{w}_{it} - \underset{\sim}{w}_{is})' \hat{\underset{\sim}{\delta}}}{h_N} \right] d_{it} d_{is} \right\} \tag{24}$$

where K is a kernel density function which tends to zero as the magnitude of its argument increases and h_N is a positive constant that decreases to zero as $N \to \infty$. Under appropriate regularity conditions, Eqn. (24) is consistent but the rate of convergence is much slower than the standard square root of the sample size.

6. *Conclusions*

Panel data have opened up avenues of research that simply could not have been pursued otherwise, but their power depends on the extent and reliability of the

data as well as on the validity of the restrictions upon which the statistical methods have been built. Otherwise, panel data may provide a solution for one problem, but aggravate another. For instance, consider the income-schooling example given by Griliches (1979):

$$y = \alpha + \beta_1 S + \beta_2 A + u \tag{25}$$

where y is a measure of income, S is a measure of schooling, A is an unmeasured ability variable which is assumed to be positively related to S. If S and A are uncorrelated, the least squares regression of y on S will yield consistent estimates of β_1. If S and A are correlated, then the least squares estimator of β_1 will suffer from omitted variable bias. If A is a purely 'family' variable in the sense that siblings have exactly the same level of A, then estimation of β_1 from differences between the brothers' earnings and differences between the brothers' education is consistent. But, if ability, apart from having a family component also has an individual component related to the schooling variables, then the within family estimate of β_1 is again biased and the bias is not necessarily less than that of the least squares estimate.

This example demonstrates that the usefulness of panel data in providing particular answers to certain issues depends on the validity of the assumptions implicit in the model. Since there is no universally acceptable way of modeling unobserved heterogeneity in panel data, it would be more fruitful to explicitly recognize the limitations of the data and focus attention on developing models and estimation methods based on what is observable, avoiding imposing arbitrary assumptions on the model and data. For additional discussions on panel data methodology, see Baltagi (1995), Hsiao (1986), and Mátáyes and Sevestre (1996).

See also: Panel Surveys: Uses and Applications

Bibliography

Ahn H, Powell J L 1993 Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* **58**: 3–80

Andersen E B 1973 *Conditional Inference and Models for Measuring*. Mentalhygrejnisk Forlag, København

Anderson T W, Hsiao C 1982 Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* **18**: 47–82

Arellano M, Bover O, Labeaga J 1999 Autoregressive models with sample selectivity for panel data. In: Hsiao C, Lahiri K, Lee L F, Pesaran M H (eds.) *Analysis of Panels and Limited Dependent Variable Models*. Cambridge University Press, Cambridge, UK, pp. 23–48

Balestra P, Nerlove M 1966 Pooling cross-section and time series data in the estimation of a dynamic model: the demand for natural gas. *Econometrica* **34**: 585–612

Baltagi B H 1995 *Econometric Analysis of Panel Data*. Wiley, New York

Chamberlain G 1980 Analysis of covariance with qualitative data. *Review of Economic Studies* **47**: 225–38

Chamberlain G 1984 Panel data. In: Griliches Z, Intriligator M (eds.) *Handbook of Econometrics*, vol. II. North Holland, Amsterdam, pp. 1247–318

De Finnetti B 1964 Farsight: its logical laws, its subjective sources. In: Kyburg H E Jr, Smokler H E (eds.) *Studies in Subjective Probability*. Wiley, New York, pp. 93–158

Griliches Z 1979 Sibling models and data in economics, beginning of a survey. *Journal of Political Economy* **87**: supplement 2, S37–S64

Heckman J J 1979 Sample selection bias as a specification error. *Econometrica* **47**: 153–161

Heckman J J 1981 Heterogeneity and state dependence. In: Rosen S (ed.) *Studies in Labor Markets*. Chicago University Press, Chicago, pp. 91–139

Hsiao C 1986 *Analysis of panel data*. Econometric Society monographs No. 11. Cambridge University Press, New York

Hsiao C, Appelbee T W, Dineen C R 1993 A general framework for panel data models—with an application to Canadian customer dialled long distance telephone service. *Journal of Econometrics* **59**: 63–86

Hsiao C, Mountain D C, Tsui K Y, Luke Chan M W 1989 Modeling: Ontario regional electricity system demand using a mixed fixed and random coefficients approach. *Regional Science and Urban Economics* **19**: 567–87

Hsiao C, Sun B 2000 To pool or not to pool panel data. In: Krishnakumar J, Ronchetti E (eds.) *Panel Data Econometrics: Future Directions; Papers in Honor of Professor Pietro Balestra*. North Holland, Amsterdam, pp. 181–98

Hsiao C, Tahmiscioglu A K 1997 A panel analysis of liquidity constraints and firm investment. *Journal of American Statistical Association*. **92**: 455–65

Kyriazidou E 1997 Estimation of a panel data sample selection model. *Econometrica* **65**: 1335–64

Lancaster T 2001 Some econometrics of scarring. In: Hsiao C, Morimune K, Powell J (eds.) *Nonlinear Statistical Modeling: Proceedings of the 13th International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*. Cambridge University Press, New York, pp. 393–402

Levin A, Lin C 1993 Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties. Department of Economics, University of California, San Diego

MaCurdy T E 1981 An empirical model of labor supply in a life cycle setting. *Journal of Political Economy* **89**: 1059–85

Manski C F 1975 Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* **3**: 205–28

Mátyás L, Sevestre P (eds.) 1996 *The Econometrics of Panel Data—Handbook of Theory and Applications*, 2nd edn. Kluwer, Dordrecht, The Netherlands

Mundlak Y 1978 Pooling of time series and cross section data. *Econometrica* **46**: 69–85

Phillips P C B, Moon H R 1999 Linear regression limit theory for nonstationary Panel Data. *Econometrica* **67**: 1057–111

Schwarz G 1978 Estimating the dimension of a model. *Annals of Statistics* **6**: 461–4

Swamy P A V B 1970 Efficient inference in a random coefficient regression model. *Econometrica* **38**: 311–23

Zellner A 1962 An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of American Statistical Association* **57**: 348–68

C. Hsiao

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Economic Restructuring: Geographic Aspects

The term 'economic restructuring' typically is used in a very general sense to indicate changes in the constituent parts of an economy. In many respects it is a 'catch-all' phrase, incorporating factors such as the changing nature of industrial sectors, corporate organization, management strategies, occupational structures, and working practices. However, the term restructuring has also become associated with a series of geographical studies with distinctive perspectives on changing economies.

1. *Economic Restructuring: Common Underlying Themes of Geographic Studies*

Geographical studies of economic restructuring seek to explain geographical variations in economic development. The approach taken in most such studies differs from the neoclassical approach within economics, which attempts to discover universal principles or 'laws' of economic behavior that are relevant in all places at all times. In addition, the neoclassical approach places faith in the ability of market mechanisms within a capitalist economy to produce the most efficient allocation of resources in a way that is beneficial for all concerned. In marked contrast, the restructuring approach within geography sees economic behavior as highly dependent upon the social customs, cultural norms, institutional practices, and wider forces of capitalist development in operation at the time, and in the place, in which that behavior takes place. Furthermore, rather than a consensual outcome which is to the benefit of all concerned, the process of restructuring under capitalism typically is seen as a struggle between competing interest groups, leading to winners and losers. Geographical studies have therefore focused upon both the creative and destructive outcomes prompted by the continuous need for capitalist firms to generate profits (Storper and Walker 1989).

The restructuring approach has generated a vast body of literature undertaken by geographers and other social scientists interested in geographical issues, including regional economists and urban sociologists. These studies have been prompted by recognition of the fact that geography is a central element in the restructuring process. Not only does restructuring vary from one place to another but, in addition, the differences between areas are used by various agents—governments, corporations, entrepreneurs, trade unions, and workers—to affect the outcome of the restructuring process itself. Geography—sometimes referred to as place or space—is, therefore, not just a reflection of these processes, it is a constituent part of the restructuring process itself (see *Place in Geography*). A good example of the role of space is when a corporation moves employment from an area in which labor costs are high to a low labor cost region, or from an area in which workers are resistant to new working practices to one in which they are more compliant.

Restructuring studies focus upon the processes leading to restructuring; the mechanisms through which this restructuring is achieved; and the social consequences of these processes in terms of factors such as: income distribution, occupational structures, the prevalence of unemployment, and the incidence of two-career households. Nevertheless, it is possible to group restructuring studies by a number of empirical foci and theoretical approaches spatial divisions of labor; new industrial spaces; globalization; services; feminization; the public sector; and knowledge economies.

2. *Spatial Divisions of Labor*

The first geographical approach to focus explicitly upon economic restructuring was spatial divisions of labor (Massey and Meegan 1982, Massey 1995). This concept highlights the fact that throughout the first part of the twentieth century there was a shift in regional economic structures away from sectors towards functions. Within regional sectoral specialization, regions tended to specialize in all aspects of the production process of particular products, such as ships, cars, or textiles. Classic examples of regional sectoral specialization are the car manufacturing complex of Detroit or the once world-leading shipbuilding region of the Clyde in Scotland. However, following a shift towards regional functional specialization, in the second half of the twentieth century regions tended to specialize in particular stages in the production process of a wide range of products. The spatial divisions of labor approach therefore implies that there is a tendency for different types of work to be undertaken in different places. Thus, corporate headquarters tend to be concentrated in large cities, design centers may be concentrated in high-technology clusters, whilst the routine assembly of products may be based in regions with low labor costs.

Spatial divisions of labor are manifest at other scales in addition to the region. At the city-wide scale areas of manufacturing specialization are often termed 'industrial districts.' However, decentralization in the latter half of the twentieth century promoted functional specialization as new forms of work—and especially services—were introduced into suburban areas following the outmigration of people from inner cities. Spatial divisions of labor can also be seen at the international level, leading to the new international division of labor thesis. Hence, the advanced capitalist

regions of North America, Europe, and Japan may be seen as global centers of corporate control, finance, product design, and innovation, whilst routine assembly jobs are undertaken in developing countries. However, research soon made it clear that there were few, if any, clear-cut geographical patterns, and the shift from sectors to functions was a tendency or trend rather than a specific outcome. Hence, different industrial sectors produced complex spatial divisions of labor, some of which overlapped and some of which did not.

The spatial division of labor approach attributed this geographical restructuring to the continual need of capitalist firms to enhance their profitability. There is much debate about the processes that led to low productivity and declining profitability in many industries in the second half of the twentieth century. One important influence was growing international competition from low-cost locations outside the Western world. Other factors were a lack of investment in new machinery and boring repetitive work (often associated with labor unrest). Another factor prompting the desire for economic restructuring in the early part of the 1980s was the influence of corporate raiders who issued junk bonds to gain control of major corporations. They then enhanced share values by asset-stripping: putting new management into particular corporate divisions to enhance productivity through reductions in manpower and the introduction of new working practices, and then selling-off the revitalized units. Prosecutions for insider dealing led to a progressive decline in the influence of the raiders but gradually pensions funds and major financial institutions began to exert similar pressures on corporations to enhance productivity and profitability.

Geographical studies drew attention to processes that were particularly important in the late 1970s and early 1980s as companies began to respond to declining profitability at that time (Massey and Meegan 1982). First, there was rationalization, a euphemism for the reduction of industrial capacity through factory closure or massive workforce cutbacks. This process was especially prominent in the UK in the 1980s under the Thatcher governments. Thus the first nation to experience the Industrial Revolution became the first nation to experience a massive process of deindustrialization. Job losses and factory closures were especially pronounced in the older industrial areas of the UK, a pattern which was replicated in the US, Europe, and Australia. Rationalization frequently was associated with spatial relocation as routine production functions were transferred to areas with low labor costs in developing countries. Not only were shopfloor workers made redundant, intermediate level white-collar workers were also laid off as corporations stripped out their elaborate managerial and administrative hierarchies (a process sometimes termed 'delayering' or 'downsizing'). Second, there was the process of investment and technical change. This involved capital investment in new machinery and equipment in an effort to increase productivity and the quality of products. This frequently was also associated with losses of employment. Third, there was the process of intensification: increases in labor productivity through managerial and organizational change.

Many of these changes were made to increase flexibility. Rather like restructuring, this is a word with a wide range of meanings but in the context of economic restructuring it has taken on two further important associations. First, there is functional flexibility, the capacity of firms to extend the range of skills of their employees. This has often involved breaking down the barriers between different types of occupation. Second, there is numerical flexibility, the capacity of employers to alter the numbers of workers employed at any given time. This has involved a wide range of strategies including the increased use of temporary workers, part-time staff, casual employees, job-sharing, shift working, flexitime, annualized hours, and a wide range of innovative forms of job contract.

Two important factors underpin the need for flexibility. First, there is the ever increasing pace of technological change, both of products and the means by which they are produced. Increased use of computerization has tended to break down some of the divisions between mechanical-based and electronic-based occupations. The other factor is increasing market volatility. Markets for certain mass-produced goods have become saturated while niche markets for specialized goods change rapidly. Producers therefore need to be able to adjust quickly and efficiently to variations in demand.

As with other aspects of industrial restructuring, the capacity to introduce flexibility shows complex geographical patterns and wide variations among regions. For example, in an attempt to escape from labor hostility to flexible working practices in older industrialized areas with a tradition of trade unionism, many companies have established new factories in areas that are free of such traditions. The location of new US car plants by Nissan in Smyrna, Tennessee, and Toyota in Georgetown, Kentucky, illustrate this trend. Restructuring has greatly affected the economic fortunes of different places, favoring some and devastating others. This led to a series of studies—especially in the UK—that became known as the localities approach, focusing upon the collective impact of restructuring upon the life-chances of people in different places (Cooke 1989).

One of the most quoted and influential aspects of the spatial divisions of labor approach was Massey's use of the geological metaphor. The economic geography of an area was likened to various phases or 'rounds' of investment in factories, plant, machinery, and labor. Each round leaves an imprint in the form of a layer or strata of development. Some of these layers become eroded over the years by subsequent rounds of

the restructuring process but collectively they serve to shape the economic landscape (see also *Local Economic Development*;*Uneven Development, Geography of*).

2.1 Japanization

Concepts of flexibility were initially imported into the US and Europe from Japan in the 1980s. During their period of reconstruction after the second world war Japanese manufacturers became eminently successful at combining high productivity with high quality products. Initially they did this to overcome the disadvantages faced by Japanese companies compared with western producers but in so doing made a virtue of necessity. Shortages of capital meant that Japanese firms had smaller amounts of machinery than their Western counterparts and had to learn ways of using existing machines more flexibility (for example, by changing the dies of dedicated machines more rapidly so they could be used for more than one purpose).

Limited capital also meant that they were forced into making contracting arrangements with numerous small producers to deliver components just at the time they were needed in the final assembly plant. Not only did this cut down on expensive stocks of components (since these larger firms were initially short of capital) but it enabled closer relationships to develop between firms. A process of collective learning thus ensured the delivery of high quality components. This process, known as contracting-out, distancing, or outsourcing, is another important form of economic restructuring. These processes have been copied widely by Western manufacturers as part of the restructuring process. This is sometimes termed Japanization (Elger and Smith 1994) or the development of lean production (Harrison 1994).

Japanese manufacturing practices drew attention to a rather different structure to regional functional specialization. Because of the need for close links with suppliers, large Japanese firms began to perpetuate and enhance regional sectoral specialization. Indeed, there has emerged a second strand to the restructuring approach that argues that clustering of firms making the same products (known as regional agglomeration) was once again an important feature of the advanced economies.

3. New Industrial Spaces

The second approach in restructuring studies, focusing on new industrial spaces, has been prompted by the fact that some regions lead the world in the design and production of particular products (Castells and Hall 1994). The classic case is Silicon Valley in California that leads the world in semiconductors (Saxanien 1994). Other examples include racing car production in southern England, textiles in northern Italy, and machine tools from the Baden-Württemberg region of Germany. Rather than regional functional specialization, these examples suggest we are witnessing the resurgence of the industrial districts of an earlier age and there are various schools of thought that have attempted to explain their existence (see also *Technology Districts*).

Scott (1988) explained these clusters as the outcome of transaction costs. Transaction costs are the costs of transforming inputs into products and services. Companies are faced with three choices in minimizing transaction costs. They can undertake activities internally (in-house) through some form of *hierarchy*; subcontract activities to external suppliers and contractors through the *market*; or adopt a halfway house and engage in collaborative activities with other companies through networks. Scott argued that changing economies of scale and scope were affecting transaction costs and leading to firms clustering together geographically in collaborative networks.

Regulation theory has also been called upon to explain industrial agglomeration. This theory, originally developed by a group of French Marxian economists in the 1970s, attempts to relate developments in particular nations to broader trends in the evolution of capitalism (Amin 1994). Regulation theory holds that over time broad sets of conditions will emerge that enable production and consumption in capitalist societies to be brought together in a stable state, known as a regime of accumulation. For example, in the early part of the twentieth century, increases in productivity were brought about by increasing the efficiency with which inputs were used (known as the intensive regime of accumulation). This gave rise to the system of large factories based on internal economies of scale (see *Regulation Theory in Geography*).

Economies of scale are factors that cause the average cost of a commodity to fall as the amount of output of the commodity increases. Internal economies of scale are those factors that apply to individual firms (irrespective of the size of the industry to which they belong). These internal economies usually ensure that the optimal scale of output is large in order to recoup the high costs of research, development, and specialized capital equipment. It is these factors that led in the past to the growth of large factories. This system has been termed Fordism after the system of mass production pioneered by Henry Ford in Detroit for the production of cars.

External economies of scale, in contrast, apply when the industry to which the firm belongs to (rather than the firm itself) is large. External economies include the development of specialized ancillary services, pools of skilled labor, and specialist component suppliers. If the output of the latter is large, this can reduce costs, thus potentially benefiting the industry as a whole.

Economies of scope are factors that make it cheaper to produce a range of commodities (rather than to produce individual types of commodity in isolation) and again these economies of scope can be internal or external.

Regulation theorists argue that since the 1980s we have witnessed the emergence of a new regime of accumulation based on flexibility. Increasing market volatility and technological change have meant that internal economies of scale have been breaking down, leading to horizontal disintegration: the tendency for producers to become smaller and more numerous. In addition, there have been declining internal economies of scope, leading to vertical disintegration: the creation of specialist subcontractors (Scott 1988). Smaller firms have been able to take advantage of external economies of scale and scope, often by clustering together to reduce the costs of interacting with the numerous specialist suppliers upon which they are dependent.

The regulation theorists argued that whereas the regime of accumulation was a very broad abstract concept, the ways in which the problems of capitalist systems were resolved depended upon a mode of regulation that differs in some respects from nation to nation. This latter concept involves the norms, regulations, and institutional practices that govern issues such as trade and competitiveness policy, finance, and labor law. These issues will depend upon the particular historical development of the nation state concerned. Critics point out, however, that the transaction-cost approach is based on an economic logic which ignores these historical factors (Sayer 1989). This critique has led to an approach in restructuring studies that recognizes the social embeddedness of economic activities.

3.1 The Social Embeddedness of Economic Activity

The concept of social embeddedness recognizes that all economic activity depends upon the social context in which it takes place (Granovetter 1985). Economic systems are seen as the product of conventions applied by highly 'reflexive' (i.e., thinking and aware) human agents. All economic systems depend upon systems of meaning that are usually specific to particular places. In the context of regional development this has led to a focus upon the institutional structures that encourage or inhibit economic activity. Work has therefore focused upon the cultural factors that generate a strong sense of common purpose, trust, and skill development amongst geographically clustered firms. This has led to the concept of the learning economy (Lundvall and Johnson 1992) and the learning region (Morgan 1997). In addition, it has also led to a focus upon the local work cultures—such as that in Silicon Valley—that facilitate factors such as high rates of technological innovation (Saxanien 1994).

4. Globalization

Those who are sceptical about the importance of these 'new industrial spaces' have pointed to the overwhelming power of the forces of globalization, and this forms another major focus within studies of economic restructuring (Dicken 1998). Globalization is—like most aspects of restructuring—a highly controversial and much disputed concept. The term is usually associated with the growing importance of multinational (or transnational) companies that operate in more than one country. However, large corporations such as Ford have long manufactured outside of their parent nation, producing goods for consumption in distant markets. Globalization therefore refers to a more recent process in which the operations of transnationals, both in the spheres of production and distribution, are increasingly integrated on a global scale. Thus, products typically incorporate components made in diverse locations throughout the world. Furthermore, whilst the appearance of products such as cars may be tailored to meet the demands of particular local market conditions, this diversity is underpinned by many common elements (such as the floorpans and engines in the example of automobiles) to reduce production costs. Globalization is also associated with the global spread of Western values of materialism and consumerism (see also *Globalization: Geographical Aspects*).

Globalization has had a crucial effect upon economic restructuring. In particular it has led to the emergence of so-called 'world cities' (or 'global cities'); command centers of the new world economy which house the headquarters of the major corporations and the financial system (Sassen 1994) (see also *Cities: Capital, Global, and World*). Examples include New York, Los Angeles, London, Paris, and Tokyo. The occupational structures of global industries are such that these cities tend to be characterized by extremes of affluence and poverty. While few cities can claim true global status, there is a sense in which all urban centres are now global, for they are all affected by events outside their boundaries and they are all engaged in a fierce competition to attract mobile capital investment into their districts. Those who are sceptical of the new industrial spaces thesis argue that it is the elaborate structures of global corporate hierarchies that shape economies rather than industrial clusters in specific regions.

One approach to reconciling the conflict between the globalization and 'new industrial spaces' schools of thought is through the concept of neo-Marshallian nodes (Amin and Thrift 1992). This concept is named after Alfred Marshall, one of the first economists to draw attention to the positive advantages that firms can derive by locating closely together in a geographical cluster, such as common pools of knowledge and skill. This concept draws attention to the importance of geographical clusters of small firms but at

the same time recognizes that they are frequently dependent upon the activities of major corporations (see also *Industrial Geography*).

5. Services

Geographic studies of economic restructuring also focus upon the growth of services (Daniels 1993) (see also *Service Economy, Geography of*). Although services are notoriously difficult to define, a common theme is economic activity that does not involve the material transformation of a physical product. However, services are defined, there is little doubt that service activity is dominant in terms of total employment and is on the increase in Western societies. Growth has been especially pronounced in financial and business services, retailing, leisure, and entertainment industries. Growth in services has tended to concentrate outside the traditional manufacturing centres of the Western economies (for example, in the 'Sunbelt' of the US and in the South East of the UK). However, new telecommunications systems are enabling back-office activity (i.e., relatively routine service activity) to be decentralized away from areas of rapid service growth into older industrial areas.

The restructuring of services has had profound effects upon the social structure of economies. In particular, it has led to social polarization with increases in both the affluent and the less well off. Whereas the declining traditional heavy manufacturing industries had substantial proportions of blue-collar, relatively well-paid, middle-income jobs, services tend to be associated with both relatively high-paying and low-paying jobs. Closely connected with the growth of services has been the increasing numbers of women in the workforce. Whereas men were employed in most of the relatively well-paid jobs in manufacturing, many of the new relatively low-paid service jobs (but also some of the well-paid service jobs) are held by women.

6. Feminization

Although the spatial divisions of labor and the new industrial spaces approaches to economic restructuring recognized that women were often employed in routine assembly work, feminist perspectives have highlighted the neglect of gender issues in many of these earlier studies. Early regulation theory, for example, focused upon work dominated by men—most notably the assembly of cars or high technology research and design work—even though women frequently were engaged in routine 'Fordist' assembly work. Regulation theory tended to adopt the reserve army of labor theory to explain the position of women in the labor market (i.e., it assumed that they were a variable pool of workers taken on in times of high demand but rapidly laid off in times of recession). A number of flaws were soon apparent in this theory. First, it simply did not fit the facts; in the last quarter of the twentieth century women were an ever increasing proportion of the workforce rather than a reserve army. Second, previous studies frequently ignored the socially embedded character of women's employment. The issue of skill illustrates this point well. For example, rather than being an innate attribute, women's so-called 'nimble-fingers,' which supposedly make them suitable for low-paid routine assembly work in light industry, are now seen as socially ascribed. Subsequent work has shown how the combination of workplace, household, and community arrangements structure the job opportunities available to women (Hanson and Pratt 1995). Gender issues are therefore a crucial element in processes of economic restructuring, as are many other socially ascribed human characteristics related to factors such as family status, ethnicity, disability, and age (see also *Gender and Feminist Studies in Geography*).

7. Economic Restructuring and the Public Sector

Although initially applied to the private sector, concepts of restructuring have also been used to analyze the extensive changes in public sector organizations in recent years. In a public sector context the flexible firm model involves a core of public sector managers who draw up contracts for work to be subcontracted to various private sector organizations, a structure which is already utilized in many US and European local governments. As with the restructuring of private sector organizations, the geography of public sector restructuring shows complex overlapping patterns, but some generalizations are possible. In the UK, for example, the contracting-out of services tends to be greater in districts in which local management is in favor of change, where there are fewer logistical problems for private sector contractors, and where there are relatively compliant trade unions (Pinch 1997) (see also *Welfare State, Geography of*).

8. Knowledge Economies

The most recent strand of economic restructuring work by geographers has focused around knowledge economies (see also *Information Society, Geography of*). Various terms have been used to draw attention to the increasing importance of knowledge in contemporary societies including: knowledge-based capitalism (Florida 1995), the network society (Castells 1996) and reflexive accumulation (Lash and Urry

1994). This role of knowledge is reflected in the ever-increasing pace of technological change and the rapidity of changes in consumer fashions and styles. This means that rather than compete by providing goods and services at the least cost, many firms increasingly compete on the basis of factors such as quality, performance, style, and design. Rapid developments in the sphere of telecommunications—teleconferencing, faxes, email, the Internet, and various types of specialist computer networks—mean that knowledge can be disseminated much more quickly throughout corporate hierarchies or networks of collaborating firms. The term telematics is used to indicate services that link computer and digital media equipment over these telecommunications links. Central to these developments are new communications channels based around satellites and fiber optics that permit the exchange of greater and more complex amounts of information much more quickly over longer distances.

So far there has been relatively little restructuring research on the impact of telematics and the information economy (Graham and Marvin 1996). Certainly telematics facilitate globalization and ever increasing distances between the various elements in the production chain. In addition, they facilitate telecommuting: linking people into global information networks and enabling them to work in their own homes. While some claim new technologies will result in the dissolution of city life, paradoxically there has been a recent resurgence of interest in cities as sites of innovation and creativity. The reason for this is that creativity in knowledge-based industries is fostered by having dense networks of people in close physical contact in places where trust and reciprocity can facilitate learning. Furthermore, much of the expensive infrastructure for telematics is heavily concentrated in urban areas.

9. *Overview*

Rather than a single integrated approach, geographical studies of 'economic restructuring' represent a series of foci: on spatial divisions of labor, new industrial spaces, globalization, services, feminization, and knowledge economies. Nevertheless, these studies share a common desire to understand the many complexities of economic restructuring without recourse to simple 'laws' of economics that are assumed to work in all places at all times. In addition, they are all committed to unraveling the crucial role of geography as a constitutive element of restructuring processes. Although many generalizations have been possible, the geographical manifestation of economic restructuring shows many complex, ever changing, and cross-cutting patterns. This has led to a dynamic field that has become known as the 'new' regional geography (for example, Bryson et al. 1999) (see also *Economic Geography*). Although somewhat distinct from the approaches typically utilised by economists, it has attracted considerable interest from a wide range of scholars in related disciplines such as urban studies, planning, sociology and cultural theory.

See also: Globalization: Geographical Aspects; Location Theory; Urban System in Geography

Bibliography

Amin A (ed.) 1994 *Post-Fordism: A Reader*. Blackwell, Oxford, UK

Amin A, Thrift N 1992 Neo-Marshallian nodes in global networks. *International Journal of Urban and Regional Research* **16**: 571–87

Bryson J, Henry N, Keeble D, Martin R (eds.) 1999 *The Economic Geography Reader*. Wiley, Chichester, UK

Castells M 1996 *The Rise of the Network Society*. Blackwell, Oxford, UK, Vol. 1

Castells M, Hall P 1994 *Technopoles of the World: the Making of 21st Century Industrial Complexes*. Routledge, London

Cooke P 1989 *Localities*. U. Hyman, London

Daniels P W 1993 *Service Industries in the World Economy*. Blackwell, Oxford, UK

Dicken P 1998 *Global Shift: Transforming the World Economy*, 3rd edn. Paul Chapman Publishing, London

Elger T, Smith D 1994 *Global Japanization*. Routledge, London

Florida R 1995 Towards the learning region. *Futures* **27**: 527–36

Graham S, Marvin S 1996 *Telecommunications and the City: Electronic Space, Urban Places*. Routledge, London

Granovetter M 1985 Economic action and social structure: the problem of embeddedness. *American Journal of Sociology* **91**: 481–510

Hanson S, Pratt G 1995 *Gender, Work and Space*. Routledge, London

Harrison B 1994 *Lean and Mean: The Changing Landscape of Corporate Power in the Age of Flexibility*. Basic Books, New York

Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK

Lash S, Urry J 1994 *Economies of Signs and Space*. Sage, London

Lundvall B-A, Johnson B (eds.) 1992 *National Systems of Innovation*. Frances Pinter, London

Massey D 1995 *Spatial Divisions of Labour*, 2nd edn. Macmillan, London

Massey D, Meegan R 1982 *The Anatomy of Job Loss*. Methuen, London

Morgan K 1997 The learning region: institutions, innovation and regional renewal. *Regional Studies* **31**: 491–503

Pinch S 1997 *Worlds of Welfare: Understanding the Changing Geographies of Social Welfare Provision*. Routledge, London

Sassen S 1994 *Cities in a World Economy*. Pine Forge Press, London

Saxanien A 1994 *Regional Advantage; Culture and Competition in Silicon Valley and Route 128*. Harvard University Press, Cambridge, MA

Sayer A 1989 Post-Fordism in question. *International Journal of Urban and Regional Research* **13**: 666–96

Scott A 1988 *New Industrial Spaces: Flexible Production and Regional Economic Development in the USA and Western Europe*. Pion, London

Storper M 1995 The resurgence of regional economies, ten years later: the region as a nexus of untraded interdependencies. *European Urban and Regional Studies* **2**: 191–221

Storper M, Walker R 1989 *The Capitalist Imperative: Territory, Technology and Industrial Growth*. Blackwell, Oxford, UK

S. Pinch

Economic Sociology

In his 1936 presidential address to the American Sociological Society, Henry Pratt Fairchild lamented the estrangement of economics and sociology. 'The time is now ripe,' he declared, 'to correct this lesion, and to bring about such a rapprochement of the two sciences as shall serve to round out and solidify them both' (Fairchild 1937 p. 4). Economic institutions, noted Fairchild, ought 'to be incorporated in sociological theory on a par with the family, the state, or the school' (p. 7). In Fairchild's vision, economists would continue to study economic relations while sociologists studied the social organizations within which those economic relations took place.

As a recognized specialty, economic sociology only began to fulfill Fairchild's vision some four decades after his 1936 exhortation. True, economic activity has figured importantly in sociology from the time of Karl Marx, Max Weber, and Emile Durkheim (Swedberg 1998). As economists' studies of economic growth proliferated after World War II, furthermore, sociologists often joined them in examining the interplay of economic growth with changes in population, family structure, education, and other social phenomena. But those sociologists typically defined themselves as specialists in modernization, development, social mobility, or social change rather than as economic sociologists. Talcott Parsons and Neil Smelser's (1956) attempt to synthesize economic and sociological approaches attracted considerable attention, but failed to galvanize a new specialty.

Economic sociology as such began to gain coherence and strength in the 1980s. Arthur Stinchcombe, for example, published his *Economic Sociology* in 1983 and two years later Mark Granovetter (1985) published an influential article on embeddedness. Here are some other milestones: the *Encyclopedia of Social Sciences* published in 1937 contained no articles devoted to sociologists' treatments of explicitly economic processes; the *International Encyclopedia of the Social Sciences* (1968) included a general article on Economy and Society but none called economic sociology; while this encyclopedia singles out economic sociology as a specialized field.

In its recent form, economic sociology offers a sociological account of economic phenomena, more specifically, the production, consumption, distribution, and transfer of assets, goods, and services. In the process, it goes well beyond Fairchild's vision. It subverts the division of labor prescribed by Fairchild. The new agenda treats economic processes and behavioral assumptions—such as markets, rationality, or self-interest—as products of underlying social processes. As Harrison White (1988, p. 232) puts it, market activity is 'intensely social—as social as kinship networks or feudal armies' (for other general statements, see e.g., Smelser and Swedberg 1994a, Dobbin 1994, Barber 1995, Beckert 1996, Ingham 1996, Bourdieu 1997, Swedberg 1997, Fligstein 1998, Collins 1999).

Recent changes in the world economy have made the agenda of economic sociology both more urgent and more promising. Growth of international capital markets, expansion of multinational firms, pursuit of global commodity marketing, collapse and transformation of centrally planned economies, and recognition that capitalist economies depend on complex institutional infrastructures all challenge economic sociologists to locate economic transactions in their social settings.

1. *Approaches to Economic Activity*

Within recent economic sociology, we find three somewhat distinct approaches to economic phenomena. They vary with respect to economics in two regards: their proximity to standard economic explanations, and their proximity to conventional economic subject matters. We might call them *extension*, *context*, and *alternative*.

Extension theorists apply relatively standard economic models to social phenomena economists themselves have not treated extensively or effectively, for example, household behavior, sporting competition, religious recruitment, and compliance with states.

A *context* approach identifies features of social organization that work as facilitators or constraints on economic action. This position is intent on revamping economists' portrayals of individual and collective decision-making, for example, by specifying conditions other than short-term gain that influence decisions. Advocates of context often speak of the 'embeddedness' of economic phenomena in social processes, and often refer to interpersonal networks when they do so. As Michel Callon (1998, pp. 10–11) puts it: 'What benefits accrue from social network analysis? Answer: a simple explanation of the possibility that agents have of calculating, when caught up in situations of extreme uncertainty. They do not have to open up to their environment in order to exchange or get information, or to negotiate and coordinate their decisions so as to lay the foundations of a possible order. They are open and connected: it is from these connections that they derive their ability to calculate. *Homo clausus* of economic activity is re-

placed by *homo apertus* of social network analysis.' Given their orientation to the agenda of economics, followers of this approach have continued to focus on firms and different kinds of markets.

In the *alternative* perspective, sociologists propose competing accounts of economic transactions. Rather than expanding the economic approach or complementing it, one prominent view argues that in all areas of economic life people are creating, maintaining, symbolizing, and transforming meaningful social relations. As a result, the subject matter certainly includes firms and markets but also ranges over households, immigrant networks, informal economies, welfare transfers, or organ donations. Indeed, economic sociologists who pursue extension and context strategies often fail to recognize alternative analyses as part of their specialty.

By and large the *extension* approach pays little attention to culture, interpersonal ties, or categorical differences, such as class, gender, race, or nationality. *Context* theorists typically take just one of these—culture, ties, or categorical differences—and make them an explicit part of the explanation. *Alternative* perspectives give more explicit attention to combinations of culture, networks, and categories.

To detect differences among the three approaches in concrete areas of economic life, it is therefore convenient to ask repeatedly to what extent and how they employ culture, interpersonal ties, and categories as part of their explanation. Let us employ those rules of thumb in looking at four contrasting areas of economic life: firms, labor markets, households, and consumer markets. The first two sectors figure prominently in standard economic analyses, the second two, less so.

1.1 Firms

Economic sociologists study features of firms that managers and economists have thought to be ruled by rational efficiency and show how these corporate features depend on social categories, interpersonal ties, and culture. What, for instance, motivates workers in capitalist firms? The simplest economic answer is that employer and worker strike a bargain in the exchange of human capital for wages. Economists and economic sociologists have moved out of that efficiency-driven position by examining other organizational rationales for payment systems. Trond Petersen (1992), for instance, documents how variation in types (and not just levels) of payment within department stores establishes a hierarchy of jobs and returns from work.

As Aage Sorensen's (1994) useful synthesis of standard thinking about systems of compensation shows, the literature on payment systems divides between analysis from employers' perspectives and from workers' perspectives. From the employers' perspective, the literature looks at variation in payment systems as incentive structures, techniques for manipulating motivation. From the workers' perspective, a less abundant literature portrays payment systems as opportunity structures.

While such analyses provide thin organizational context for a well-known economic process, other sociologists offer competing accounts of organizational rationales. A case in point is Calvin Morrill's (1991) ethnography of executives employed by a large American corporation. Responding to structural transformations within the firm along with a rash of hostile takeovers these executives generated a new public style of competition and conflict resolution. The new system by no means simply matched material rewards to task performance.

Instead the system demanded adherence to a symbolically charged code of honor. Honorable executives or 'white hats' who followed the rules were rewarded not only by reputation and esteem but, Morrill found, expanded power and greater access to resources than their dishonorable 'black hat' colleagues enjoyed.

This sort of analysis takes the economic sociology of firms in a far different direction than wage: performance matching. Pure efficiency explanations of payment systems, as effective performance measures or reward structures, fail to account for the creation, changing acceptability, and diversity of arrangements for compensation.

1.2 Labor Markets

When it comes to labor markets, many economic sociologists have shifted radically away from models assuming socially unfettered rational exchanges between workers and employers. Instead, sociologists identify a vast set of structural, cultural, and categorical factors shaping the operation of labor markets. Consider the problem of how people find jobs. Traditional economic accounts of labor markets look at the intersection of two search procedures: workers searching for job opportunities, employers searching for workers. The individual on each side, the theory tells us, is trying to maximize self-interest by looking for the best possible bargain. Employers want to find the worker with the best qualifications, while workers are seeking the highest price for their labor.

Mark Granovetter's 1995 [1974] study seriously complicated this neat equation by introducing third parties. A job seeker's existing social networks, he showed, shaped the job search process in fundamental ways, by providing information and access to jobs. Granovetter coined the term 'strength of weak ties' to convey a key finding: connections to distant acquaintances were significantly more effective referral sources than close friends or family. More generally,

Granovetter points out that economic transactions are embedded in complex social relations which constrain them.

In a similar vein, other economic sociologists have examined the operation of migration networks, informal economy, ethnic niches, and internal labor markets (see e.g. Portes 1995). Such studies extend into the examination of culture (e.g. within the segments of labor markets) and of categorical difference (e.g. discrimination by gender, race, and ethnicity). In his detailed analyses of labor markets within corporations, for example, Ronald Burt (1998) has shown that, on the whole, men's breadth of connections to the firm's various power centers directly affects their relative success while involvement in self-contained cliques inhibits men's advancement. For women, however, involvement in a small circle that includes a powerful boss actually enhances their chances for success.

This work frequently moves away from contexts or extensions of standard economic work into clear alternatives. Economist Chris Tilly and sociologist Charles Tilly (1998, p. 19) state the program for a full-fledged alternative in the following terms: 'What vision, then, do we hold of work and labor markets? At the center of our analysis stands an image of a worker and an employer, each involved in extensive social networks that connect them not only with each other but with many others inside or outside the place of employment. The worker and employer are bargaining out the terms of work, using the material and organizational resources (including existing and available technologies) at their disposal. They draw on previously existing social relations—both work-generated and otherwise—in seeking their advantages. In the process, they generate new agreements, understandings, and social ties, which in turn constrain subsequent interaction. They interact within strong limits set by their shared culture and by the previous history of the productive organization.'

2. *Households*

Firms and labor markets occupy center stage in mainstream economic thought. Household economies, in contrast, have typically remained at the sidelines, part of a nonrational social world seemingly run by affect, not efficiency. The predominant model of what Amartya Sen (1983) calls the 'glued-together' family further discourages questions about the allocation, distribution, and expenditure of family resources among family members. Reacting to the pioneering work of Gary Becker (1981), feminist economists and their allies in economic sociology have challenged straightforward applications to households of firm-level models. Their work has integrated bargaining models, multiple utility schedules, and gender differences into the analysis of household economies. Another group of sociologists studying families and gender dynamics addresses parallel issues, but works in almost complete isolation from economic sociology.

One difficult problem has preoccupied both groups of researchers in sociology: causes and effects of household divisions of labor. Why do wage-earning women continue to do the bulk of household work? Why does men's household work respond so little to changes in household circumstances, and indeed, even tend to decline when their wives' income surpasses theirs? Why are household tasks often as segregated by gender as are labor markets? Why are women's earnings, even when substantial, commonly treated as discretionary income?

Answers to the housework puzzle have ranged from extensions of economistic models which see women's unpaid household labor as exchange for their husband's economic support to culturalist models where ideologies of gender shape material household exchanges. Julie Brines (1994) combines both perspectives. Brines argues that what she calls a dependency model accounts for women's domestic labor, but it fails to explain their husbands.' When it comes to men, Brines claims, adhering to masculine gender expectations leads dependent husbands to do less, not more, housework. Housework, from this perspective, acts as a symbolic marker of traditional gender relationships.

To get beyond this dichotomy will require us to find a model of interaction and bargaining within households. Marjorie DeVault's (1991) observations on food preparation as a bargaining form of social interaction among wives, husbands, and children points in that direction. Thus labor markets and household economies yield similar lessons. Without models of meaningful social interaction, economic sociology cannot get much farther in explaining such complex phenomena.

3. *Consumption*

If households have remained peripheral to economics and economic sociology, consumption splits down the middle. Some aspects of consumption have long been recognized and extensively documented by both groups of specialists. Other dimensions of consumption have become the province of specialists in family, class, gender, ethnicity, and community, in studies significantly removed both from economics and economic sociology.

Within self-defined economic sociology explicit studies of consumption have taken second place to the study of production. To be sure, for half a century sociologists and economists collaborated in surveys of consumer expenditures and behavior, a line of work that significantly influenced market research. Yet that nucleus excluded many aspects of consumption, such as family dynamics of consumption and social con-

nections among consumers. On their own, economic sociologists have related consumption to inequality and class position. The best documented cases tell of the acquisition of goods as cultural markers of class position, one aspect of which Thorstein Veblen (1953 [1899]) treated as 'conspicuous consumption' and another aspect of which Pierre Bourdieu (1984) identifies as the formation of 'cultural capital.'

The representation of class through consumption is only a special case of the relationship between social position and consumption patterns: other studies establish variation by ethnicity, gender, race, and locality. All such studies pose the classic economic sociology problem of embeddedness. Given variation in consumption from one social location to another, how do social relations actually shape the acts of consumption?

Paul DiMaggio and Hugh Louch (1998) provide one answer. They survey pre-existing noncommercial ties between buyers and sellers in consumer transactions involving the purchase of cars, homes, as well as legal and home repair services. DiMaggio and Louch find a remarkably high incidence of what they call within-network exchanges. A substantial number of such transactions take place not through impersonal markets but among kin, friends, or acquaintances. Noting that this pattern applies primarily to risky one-shot transactions involving high uncertainty about quality and performance, DiMaggio and Louch conclude that consumers will be more likely to rely on such noncommercial ties when they are unsure about the outcome. As in studies of firms, labor markets, and households, analyzing relations between economic transactions and social ties promises great returns for economic sociology.

Meanwhile, outside of economic sociology, in company with numerous anthropologists, historians, and social critics, sociologists have created a wide-ranging literature on cultures of consumption. Three questions dominate these studies: how does consumption vary with social position? How do changing patterns of consumption affect the quality of social life? and how does material culture create social meaning? This work has produced an enormous range of rich descriptive material and a great variety of conflicting interpretations. For example, students of consumption outside of self-defined economic sociologists have reproduced a great controversy that has troubled social commentators for centuries: do commodification and monetization of goods and services depersonalize and corrupt the social relations involved? Or, on the contrary, does their contribution to efficiency, productivity, and connections actually augment social well being? Economic sociologists and other students of consumption would greatly benefit from closer interaction between their separate spheres of investigation.

Firms, labor markets, households, and consumption do not of course exhaust the range of economic sociology. Scholars have investigated a wide variety of economies, from financial markets, gift economies, industries, international flows of goods and capital, to ethnic economies, education, religion, health, sexual economies, leisure, and the environment. They have paid attention as well to comparative and historical developments (see e.g. Smelser and Swedberg 1994b). In these various fields we again find examples of extension, context, and alternative models. For example in the study of technological change, some sociologists elaborate on existing economic choice models, some examine the immediate social context of innovation and its applications, while others treat technological innovation as a form of power struggle.

Although it is convenient to distinguish extension, context, and alternative, we should not exaggerate the rigidity of the boundaries between them. Indeed, in recent sociology, lines of investigation that began as attempts to provide contexts for economic analysis (for example, in studies of informal economies and gender inequality) have repeatedly developed into alternatives to standard economic analyses. Nor should we imagine that economists have stood still in these regards; institutional economists have been moving into intensive conversations with economic sociologists of all three tendencies.

In the immediate future it would be unwise to try to forward one perspective to the exclusion of the others. Each yields its own insights into the joint agenda of economics and sociology. Extension views build on existing economic theory, context accounts provide the challenge to improve on existing economic theory, correcting it when necessary, while alternative models replace existing economic theory with other principles of explanation. In the long run, the alternative perspective opens up the prospect of going beyond the embedding of economic activity. It examines ostensibly economic phenomena as the construction of meaning, formation of relationships, and organization of categories.

For understanding the complex economies of the twenty-first century, the analytical contribution of sociologists will be crucial.

See also: Consumer Culture; Consumption, Sociology of; Durkheim, Emile (1858–1917); Entrepreneurship; Family and Consumer Sciences: United States; Financial Institutions in Economic Development; Firm Behavior; Household Production; Labor Supply; Labor Unions; Labor, Division of; Market Structure and Performance; Marx, Karl (1818–89); Money, Sociology of; Networks: Social; Schumpeter, Joseph A (1883–1950); Theory: Sociological; Weber, Max (1864–1920)

Bibliography

Barber B 1995 All economies are embedded: the career of a concept, and beyond. *Sociological Research* **62**: 387–413

Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
Beckert J 1996 What is sociological about economic sociology? Uncertainty and the embeddedness of economic action. *Theory and Society* **25** (December): 803–40
Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA
Bourdieu P 1997 Le champ économique. *Actes de la recherche en sciences sociales* **119**: 48–66
Brines J 1994 Economic dependency, gender, and the division of labor at home. *American Journal of Sociology* **100** (November): 652–88
Burt R S 1998 The gender of social capital. *Rationality and Society* **10**: 5–46
Callon M 1998 Introduction: the embeddedness of markets in economics. In: Callon M (ed.) *The Laws of the Markets*. Blackwell, Oxford, UK, pp. 1–57
Collins R 1999 *The Multiple Fronts of Economic Sociology*. (March) Editorial, ECONSOC Econsoc.@listproc.hcf.jhu.edu.
DeVault M 1991 *Feeding the Family: The Sociological Organization of Caring as Gendered Work*. University of Chicago Press, Chicago, IL
DiMaggio P, Louch H 1998 Socially embedded consumer transactions: for what kinds of purchases do people most often use networks? *American Sociological Review* **63** (October): 619–37
Dobbin F 1994 Cultural Models of Organization: the social construction of rational organizing principles. In: Crane D (ed.) *The Sociology of Culture: Emerging Theoretical Perspective*. Blackwell, Oxford, UK, pp. 45–90
Fairchild H P 1937 Business as an institution. *American Sociological Review* **2** (February): 1–8
Fligstein N 1998 Agreements and disagreements in the new sociology of markets. Paper presented at a conference on Economic Sociology, University of Pennsylvania, December 12
Granovetter M 1985 Economic action and social structure: the problem of embeddedness. *American Journal of Sociology* **91**: 481–510
Granovetter M 1995 *Getting a Job: A Study of Contacts and Careers*, 2nd edn. University of Chicago Press, Chicago
Ingham G 1996 The new economic sociology. *Work, Employment & Society* **10**: 549–64
Morrill C 1991 Conflict management, honor, and organizational change. *American Journal of Sociology* **97** (November): 585–621
Parsons T, Smelser N J 1956 *Economy and Society: A Study in the Integration of Economic and Social Theory*. Free Press, Glencoe, IL
Petersen T 1992 Payment systems and the structure of inequality: conceptual issues and an analysis of salespersons in department stores. *American Journal of Sociology* **98**: 67–104
Portes A (ed.) 1995 *The Economic Sociology of Immigration: Essays on Networks, Ethnicity and Entreprepneurship*. Russell Sage Foundation, New York
Sen A 1983 Economics and the family. *Asian Development Review* **1**: 14–26
Smelser N J, Swedberg R 1994a The sociological perspective on the economy. In: Smelser N J, Swedberg (eds.) *The Handbook of Economic Sociology*. Russell Sage Foundation, New York, pp. 3–26
Smelser N J, Swedberg R (eds.) 1994b *The Handbook of Economic Sociology*. Russell Sage Foundation, New York
Sorensen A B 1994 Firms, wages, and incentives. In: Smelser N J, Swedberg (eds.) *The Handbook of Economic Sociology*. Russell Sage Foundation, New York, pp. 504–28
Stinchcombe A L 1983 *Economic Sociology*. Academic Press, New York
Swedberg R 1997 New economic sociology: what has been accomplished, what is ahead? *Acta Sociologica* **40**: 161–82
Swedberg R 1998 *Max Weber and the Idea of Economic Sociology*. Princeton University Press, Princeton, NJ
Tilly C, Tilly C 1998 *Work Under Capitalism*. Westview Press, Boulder, CO
Veblen T 1953 [1899] *The Theory of the Leisure Class*. Mentor, New York
White H C 1988 Varieties of markets. In: Wellman B, Berkowitz S D (eds.) *Social Structures: A Network Approach*. Cambridge University Press, Cambridge, UK

V. A. Zelizer

Economic Status, Inheritance of: Education, Class, and Genetics

1. Introduction

Economic status is transmitted from parents to offspring. The perpetuation across generations of a family's social class, or their position in the distribution of income, is generally thought to reflect the combined effects of the genetic and cultural transmission of traits, such as cognitive functioning, that contribute to economic success, as well as the inheritance of income-enhancing group memberships and property. The superior education enjoyed by the children of higher status families contributes to this process of economic inheritance. While recent research has illuminated important aspects of this account, the factors contributing to the extent of intergenerational transmission of status and the ways that genetic and cultural transmission and the inheritance of property and memberships contribute to this process remain obscure.

First, the extent of intergenerational economic status transmission is considerably greater than was thought to be the case a generation ago, with some intergenerational correlations of parent–offspring income exceeding Francis Galton's original estimate (two-thirds) for height (Galton 1889, p. 97). Second, the genetic inheritance of traits contributing to the cognitive skills measured on IQ and related tests explains very little of the intergenerational transmission of economic status, even if the heritability of IQ is quite high. Third, the combined genetic and cultural inheritance processes operating through superior wealth, cognitive levels, and educational attainments of those with well-off parents, while important,

do not fully explain the intergenerational transmission of economic status. The article will identify some as yet overlooked individual traits that enhance economic success in the members of both generations and are transmitted across generations.

2. The Intergenerational Transmission of Economic Status

Economic status may be measured in discrete categories—by membership in hierarchically ordered classes, for example—or continuously, by earnings (wages and salaries), income (earnings plus income from property and other sources), an occupational prestige index, or wealth.

Continuous measures of status—like height in Galton's example—allow a simple metric of persistence, the intergenerational correlation coefficient ρ, the square of which measures the fraction of the variance in this generation's measure of economic success that is statistically associated with the same measure in the previous generation. The persistence of economic status is measured using a first order Markov process

$$y_o = (1-\beta_y)\bar{y}+\beta_y y_p+\varepsilon_y \qquad (1)$$

We use subscripts 'o' and 'p' to refer to offspring and parental measures, respectively, so y_o is an individual's economic status, adjusted so that its mean, $\bar{y}$, is that of the parental generation, β_y is a constant, y_p is the individual's parental y, and ε_y is a disturbance uncorrelated with y_p. In this setup, inspired by Galton's treatment of height, and extended in Goldberger (1989), regression to the mean is measured by $1-\beta_y$. Little explanatory power is gained by using higher order Markov processes to take account of the effects of earlier generations (Warren and Hauser 1997, Behrman and Taubman 1989). The intergenerational correlation is

$$\rho_y = \beta_y \frac{\sigma_{y_p}}{\sigma_{y_o}}$$

where σ_y is the standard deviation of y. if y is the natural logarithm of wealth, income, or earnings, the standard deviation of y is a common measure of inequality. Thus, if inequality is unchanging across generations, so $\sigma_{y_p} = \sigma_{y_o}$, then $\rho_y = \beta_y$. In the empirical work reviewed below earnings, income, wealth, and other measures of economic success are measured by their natural logarithm unless otherwise noted. Thus, β_y is the percentage change in offspring's economic success associated with a one percent change in parents' economic success.

Early studies following the work of Blau and Duncan (1967) estimated intergenerational correlations for income or earning among men in the US to be in the neighborhood of 0.15, leading Becker and Tomes (1986) to conclude that:

> Aside from families victimized by discrimination, regression to the mean in earnings in the United States and other rich countries appears to be rapid … Almost all earnings advantages and disadvantages of ancestors are wiped out in three generations (S32).

But the appearance of such high levels of mobility was an artifact of two types of measurement error: errors in reporting income, particularly when individuals were asked to recall the income of their parents, and transitory components in current income uncorrelated with underlying permanent income (Bowles 1972, Bowles and Nelson 1974, Atkinson et al. 1983, Solon 1992, Zimmerman 1992). The high noise to signal ratio in both generations' incomes depressed the intergenerational correlation, and when corrected using a variety of methods and distinct data sources, the intergenerational correlations for economic status appeared quite substantial, as is indicated by the data in Table 1. Another useful survey of these data is Bjorklund and Jantti (1999), and the discussion in Mulligan (1997) and Solon (2000). Estimated intergenerational correlations generally rise with age, are greater for sons than daughters, and are greater when multiple years of income or earnings are averaged. Behrman and Taubman (1989) using the Michigan Panel Survey of Income Dynamics find that the estimated intergenerational correlation of parental income and offspring earnings is 0.58 when 18 years of earnings are used compared to 0.37 for a single year.

Unlike the entries in a transition matrix, μ_{ij}, which indicate the extent to which one's prospects are

Table 1
Intergenerational persistence of some economic characteristics, β_i

Economic characteristic	Number of Estimates	Range	Average
Years of schooling	8	0.14–0.45	0.29[a]
Log earnings or wages	16	0.11–0.59	0.34[a]
Log family income	10	0.14–0.65	0.43[a]
Log family wealth	9	0.27–0.76	0.50
Log family consumption	2	0.59–0.77	0.68

Source: Mulligan (1999).
a If recent studies of the US only are included these averages of 0.35, 0.33, and 0.38 respectively

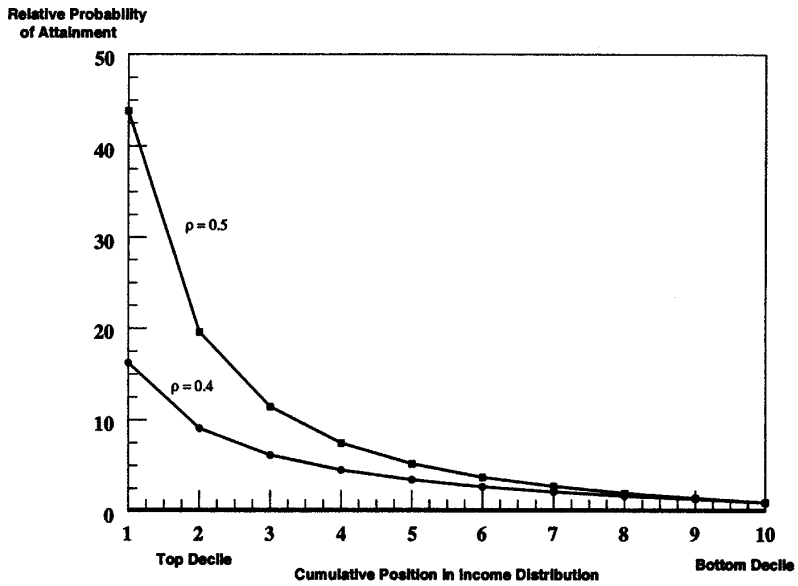

Figure 1
Intergenerational status transmission when $\rho = 0.4$ and $\rho = 0.5$. The vertical axis represents the probability of attaining the decile on the horizontal axis or higher deciles, for an individual from the top decile in parental background, relative to an individual from the bottom decile in parental background

conditioned by one's origins, the intergenerational correlation, ρ_y, is relatively difficult to interpret. However, the degree of conditioning implied can be illustrated by given values of ρ_y assuming that the underlying relationship is linear in y_o and y_p and that both are distributed normally. Figure 1 represents the transition probabilities implied by these assumptions and two different correlation coefficients, $\rho = 0.4$ and $\rho = 0.5$. The horizontal axis represents cumulative position in the income distribution, and the vertical axis represents the probability that an individual in the highest (first) decile of parental income attains at least this position divided by the probability that an individual in the lowest (10th) decile of parental income attains at least this position.

It can be seen, for instance, that an individual whose parents are in top decile is roughly 44 times as likely to attain a position in the top decile as an individual whose parents are in the bottom decile when $\rho = 0.5$, and 16 times as likely when $\rho = 0.4$. Similarly an individual whose parents are in the top decile is about 20 times as likely to attain a position in the top quintile as an individual whose parents are in the bottom decile when $\rho = 0.5$, and nine times as likely when $\rho = 0$. Studies allowing for nonlinear effects in Eqn. (1) suggest that our assumption of normality may lead these figures to understate the actual degree of persistence in the tails of the income distribution (Corak and Heisz 1999, Cooper et al. 1994).

The role of various forms of inheritance in the transmission process is now considered. Equation (1) is merely a summary of the results of a number of distinct processes having little in common except that they result in parent–offspring similarities for traits statistically associated with the degree of economic success in both generations. While candidates for the list of income generating traits with strong parent–offspring similarity are many, there are few for which both economic relevance and parent–offspring similarity have been empirically demonstrated. Among these are cognitive performance, the level of schooling, and ownership of wealth, each illustrating distinct transmission processes.

A common measure of economic success, income—that is, the sum of labor earnings (wages and salaries) and returns to assets—is treated as a phenotypic trait influenced by the individual's genotype g, environment e, and ownership of income earning assets. Genotypic and environmental influences jointly determine individual skills and other traits relevant to job performance, sometimes termed 'human capital.' Among the environmental influences are cultural transmission from parents, schools and other learning environments. Income, human capital, and assets are transformed to their natural logarithms, and then all variables normalized to have zero mean and unit variance, represented by y, l, and k, respectively. Here and below, $\tilde{y}$ is total income, while y is earnings from labor, represented as a return to human capital. So

$$l = h_y g + \beta_{le} e + \varepsilon_l \quad (2)$$

$$\tilde{y} = \omega l + \pi k + \varepsilon_{\tilde{y}} \quad (3)$$

which imply the reduced form

$$\tilde{y} = \omega(h_y g + \beta_{te} e) + \pi k + \varepsilon'_{\tilde{y}} \quad (4)$$

where $\varepsilon'_{\tilde{y}}$ is a disturbance uncorrelated with the independent variables, and the constants ω, h, β_{te} and π give the effect in standard deviation units of a standard deviation change in the relevant variable. Thus h_y^2 is the heritability of earnings, namely that portion of the variance of individual earnings that would be explained by the variance of g, were e and g uncorrelated.

Using Eqn. (4), the intergenerational correlation $\rho_{\tilde{y}}$ can be decomposed, yielding additive terms expressing the contribution of each of the above variables. Let r_{uv} be the simple correlation between variables u and v. The intergenerational correlation coefficient can be expressed as

$$\rho_{\tilde{y}} = \sum r_{j\tilde{y}_p} \beta_{\tilde{y}_o j} \quad (5)$$

where $r_{j\tilde{y}_p}$ is the simple correlation of the parental economic status measure ($\tilde{y}_p$) with some another variable j, and $\beta_{y_o j}$ is the normalized regression coefficient of the variable j in an equation like (7) below, predicting the current generation's economic status (y_o). Then if parent–offspring similarity in g, e, and k are the only sources of parent–offspring similarity in $\tilde{y}$ (i.e., $\varepsilon'_{\tilde{y}_p}$ is uncorrelated with $\varepsilon'_{\tilde{y}_o}$), the intergenerational correlation of incomes can be decomposed as follows

$$\rho_{\tilde{y}} \equiv r_{\tilde{y}_o \tilde{y}_p} = \omega(h_y r_{g_o \tilde{y}_p} + \beta_{te} r_{e_o \tilde{y}_p}) + \pi r_{k_o \tilde{y}_p} \quad (6)$$

expressing three fundamental mechanisms of intergenerational transmission of economic status: genetic, cultural and asset-based. What can be said about the relative importance of each? To answer this question we consider particular components of the genetic and environmental influences on economic success.

3. *The Role of Genetic Inheritance of Cognitive Skill*

Both the similarity of parents' and offspring's scores on cognitive tests and the statistical association of test scores and earnings are well documented. This suggests an appealing explanation of the process of intergenerational status transmission via inheritance of cognitive skills.

Correlations of IQ between parents and offspring are substantial, ranging from 0.42 to 0.72, the higher figure referring to average parental vs. average offspring IQ (Bouchard and McGue 1981). The contribution of cognitive functioning to earnings has been established using survey data to estimate the natural logarithm of earnings y as a function of a measure of parental economic and/or social status y_p, years (and perhaps other measures) of schooling s, and performance on a cognitive score c—often, in US data sets, the Armed Forces Qualification Test (AFQT), a cognitive test developed to predict vocational success—as well as an error term and other variables, such as work experience, race, and sex, which will not be covered here. This equation may be written in a normalized form as

$$y_o = \beta_{y_o y_p} y_p + \beta_{y_o s_o} s_o + \beta_{y_o c_o} c_o + \varepsilon_{y_o}. \quad (7)$$

Sixty-five estimates of $\beta_{y_o c_o}$ have been located, appearing in Fig. 2. The mean of these estimates 0.15, indicating that a standard deviation change in the cognitive score, holding constant the remaining variables, changes the natural logarithm of earnings by about one seventh of a standard deviation. By way of contrast the mean value of $\beta_{y_o s_o}$ is 0.22, suggesting a somewhat larger independent effect of schooling. We checked to see if these results were dependent on the weight of overrepresented authors, the type of cognitive test used, at what age the test was taken and other differences among the studies and found no significant effects (Bowles et al. 2001).

Does the fact that measures of cognitive functioning are robust predictors of individual earnings, along with the similarity of the cognitive scores of parents and offspring imply a major role for genetic inheritance of cognitive ability in the process of intergenerational economic status transmission? To answer this question one needs to understand the determination of schooling (s_o) and offspring cognitive scores (c_o) which appear as the independent variables in Eqn. (7). Suppose h_c^2 is the heritability of cognitive skill and that the relevant processes are described by

$$c_o = h_c g_o + \beta_{c_o y_p} y_p + \beta_{c_o c_p} c_p + \varepsilon_{c_o} \quad (8)$$

$$s_o = \beta_{s_o y_o} y_p + \beta_{s_o c_o} c_o + \varepsilon_{s_o} \quad (9)$$

where all variables are expressed in normalized form, ε_{c_o} and ε_{s_o} are error terms uncorrelated with the independent variable in their respective equations.

The cognitive test ideally would be taken by an adolescent just prior to the minimal school leaving age, thereby being a good predictor of adult cognitive functioning while still having an effect on the level of schooling attained. Figure 3 illustrates the causal model implied by Eqns. (7)–(9).

Equation (5) allows a decomposition of $\rho_{\tilde{y}}$ showing that the genetic inheritance of IQ contributes to the intergenerational status transmission process because it results in a correlation of parental income with offspring cognitive level $r_{c_o y_p}$, which in turn affects offspring income both directly ($\beta_{c_o y_o}$) and indirectly via its effect on the level of schooling attained ($\beta_{y_o s_o} \beta_{s_o c_o}$). Using Eqn. (5) to decompose $r_{c_o y_p}$ one has

$$r_{c_o y_p} = \beta_{c_o y_p} + r_{c_p y_p} \left(\beta_{c_o y_p} + \frac{h_c^2}{2} \right) \quad (10)$$

The genetically transmitted portion of this correlation is the term $r_{y_p c_p} h_c^2/2$, since the direct part r^g from g_p to

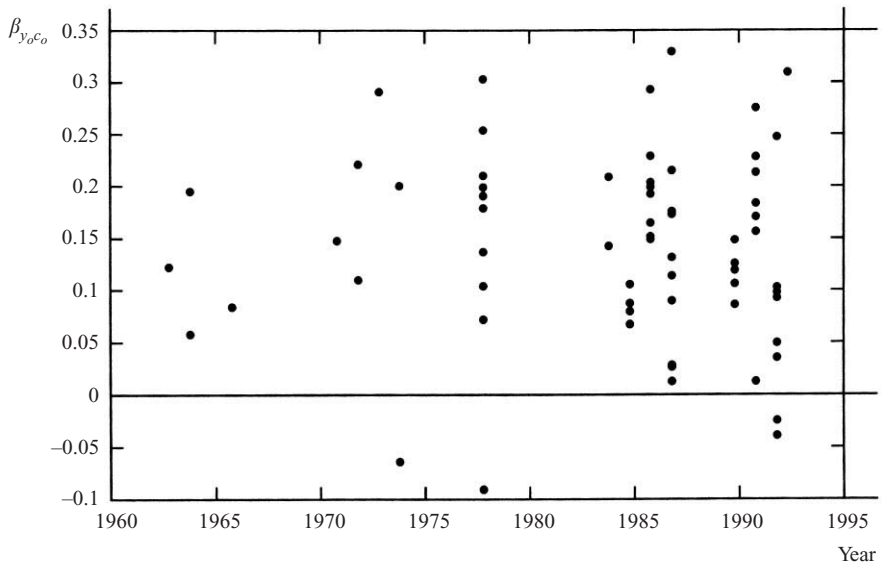

Figure 2
Normalized regression coefficient ($\beta_{y_o c_o}$) of cognitive score on the logarithm of income or earnings by year: 65 estimates from 24 studies (Bowles et al. 2001)

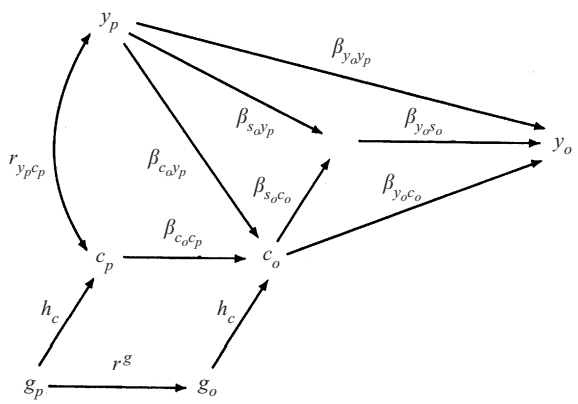

Figure 3
A causal model of intergenerational status transmission

g_o is just the genetic relatedness of parents and natural offspring, which is 1/2. Summing the direct and indirect effects one obtains the contribution of genetic inheritance of IQ to intergenerational transmission of earnings

$$\rho_c^g = r_{y_p c_p} r^g h_c^2 (\beta_{y_o s_o} \beta_{s_o c_o} + \beta_{y_o c_o}) \qquad (11)$$

Representative values have already been introduced for $\beta_{y_o s_o}$ (0.22) and $\beta_{y_o c_o}$ (0.15). Estimates (Jencks 1979) of $r_{y_p c_p}$ suggest a value of 0.35, and as an estimate of $\beta_{s_o c_o}$ the preferred estimate of Winship and Korenman (1999) is taken, which is 0.53. Therefore, $\rho_y^g = 0.047\, h_c^2$, indicating that even for $h_c^2 = 1$, the genetic inheritance of IQ accounts for a rather modest portion of the observed levels of intergenerational economic status transmission. The calculation is

$$\rho_y^g = 0.35 \frac{h_c^2}{2} (0.22(0.53) + 0.15) = 0.047 h_c^2$$

This of course assumes there is no assortative mating of parents. If parental IQ phenotypes have correlation m, and if the heredity coefficient h_c is the same for parents as for offspring, then $r^g = (1 + h^2 m)/2$. It follows that the maximum value ρ_y^g could have, which occurs when $h_c = 1$, is $0.047(1+m)$. The considerably lower values for h_c^2 suggested by recent research of Devlin et al. (1997), Otto et al. (1995), and Plomin (1999) are consistent with the conclusion that one twentieth or less of the observed intergenerational status transmission is due to genetic inheritance of IQ. Some of the data used in this exercise are not estimated very precisely (h_c^2 in particular) but the main point is robust with respect to reasonable alternative values.

Indeed it is thought that available estimates overstate the importance of general cognitive skill as a determinant of earnings for the simple reason that taking a test is more than a little like doing a job—the results measure performance, which is the joint effect of skill along with other contributors such as the disposition to follow instructions, persistence, work

ethic, and other traits likely to contribute independently to one's earnings. Eysenck (1994, p. 9) writes:

> Low problem solving in an IQ test is a measure of performance; personality may influence performance rather than abstract intellect, with measurable effects on the IQ. An IQ test lasts for up to 1 hour or more, and considerations of fatigue, vigilance, arousal, etc. may very well play a part.

Thus, some of the explanatory power of the cognitive measure in predicting earnings does not reflect cognitive *skill* but rather other individual attributes contributing to the successful performance of tasks. The importance of the performance rather than capability aspect of the test score as a determinant of earnings is suggested Bishop (1991), who found a measure of computational speed on time tested exams to be a more robust predictor of earnings than two alternative measures—a normalized sum of academic tests as well as a measure of technical competence.

4. *The Inheritance of Wealth and Educational Attainment*

The intergenerational inheritance of income-earning assets provides an alternative explanation that competes for simplicity with the inherited cognitive skills account. Table 1 shows that the intergenerational correlation for wealth status is quite substantial—within the range of parent–offspring similarity in cognitive scores—and that it exceeds the persistence of family income (to which it contributes) which in turn exceeds the persistence of earnings (which do not include income from property). While comparisons of imprecisely estimated parameters across differing data sets are notoriously unreliable, Mulligan (1997) uses a single data set and common methods for his estimates of persistence with respect to earnings, income, wealth, and consumption. His results confirm that consumption, wealth, and income regress to the mean more slowly than do earnings or wages.

These data are consistent with the inherited wealth account of economic status persistence, but for most individuals and families, income from property constitutes a negligible fraction of their total income. Only among the very well-to-do is property a major source of income. Correspondingly, very few individuals receive inheritances of significant magnitude. Mulligan (1997) estimates that estates passing on sufficient wealth to be subject to inheritance tax in the US constituted between two and four percent of deaths over the years 1960–1995.

Because wealth inheritance contributes to intergenerational income persistence, one would expect to find that persistence varies with the level of wealth of the individual's parents. For instance, in a rare study with sufficient sample size to derive good estimates of persistence for different positions in the distribution of income (Corak and Heisz 1999) the degree of income persistence for the very rich was markedly greater than for the rest of the distribution, with $\rho_{\tilde{y}} = 0.8$ for the top percentile.

However, the apparent simplicity of the inherited wealth account is misleading. The intergenerational persistence of wealth is not explained simply by bequests but reflects as well parent–offspring similarities in traits influencing wealth accumulation, such as orientation towards the future, sense of personal efficacy, work ethic, schooling attainment, and risk-taking. Some of these traits covary with the level of wealth: less well off people are more likely to be risk averse, to discount the future, and have a low sense of efficacy, for example (Bardhan et al. 2000, Fong 2000).

Like wealth levels, schooling attainments persist across generations although, as Table 1 indicates, less so. Nonetheless, father–offspring correlations for years of schooling reported by Behrman and Taubman (1989) are substantial (0.34, and would be closer to 0.4 were they corrected appropriately for errors in measurement). As in the case of wealth, the persistence of schooling attainments results from actions taken by parents and offspring, and these are influenced by beliefs and preferences that are themselves subject to intergenerational transmission. Some of the individual dispositions favoring high levels of school attainment are correlated with parental schooling levels and incomes, thereby fostering higher levels of educational attainment by the children of the well educated. This tendency is exacerbated where the children of the less well off experience schooling as a hostile or unpleasant environment, and where their families, limited incomes and inability to borrow make the students' potential labor services valuable to the family.

Because schooling attainment is persistent and has clear links to skills and perhaps other traits that are rewarded in labor markets, this human capital based account of intergenerational status transmission has strong *prima facie* plausibility, especially when deployed with the inherited cognitive skills account. Indeed, it was once commonly assumed that when adequate measures of schooling quality were developed the only effects of parental economic status on offspring earnings would operate through effects on cognitive functioning and schooling, with the direct effect of parental on offspring earnings vanishing, that is, $\beta_{y_o y_p} = 0$ in Eqn. (7).

But as the measurement of school quality has improved over the years, estimation of $\beta_{y_o y_p}$ have proven resilient. Note that $\beta_{y_o y_p}/\rho_y$ measures the importance of the direct effect of parental economic status relative to the total amount of intergenerational transmission. Mulligan (1997), controlling for a large number of measures of school quality as well as the AFQT and standard educational and demographic variables, finds that an estimate of parental income is an important and statistically significant predictor of

the natural logarithm of the hourly wage rate in 1990 and 1991 in the National Longitudinal Study of Youth. A little more than two-fifths of the gross statistical association between parental and offspring economic success (ρ_y) apparently operates independently of the influences of these conditioning variables. Specifically, the effect of an estimate of the logarithm of parental income on offspring's logarithm of income in a regression conditioned on measures of schooling quantity and quality, employment, and cognitive performance is between two-fifths and one-half of the estimated effect of parental income unconditioned on these variables. Mulligan's work repeats the finding in Atkinson et al. (1983) for a sample with direct measures of incomes of fathers and sons in the UK in which two-thirds or more of the substantial intergenerational transmission of income status is independent, in the sense just defined, of the covariation of parental income with a range of measures of son's schooling including the number of Ordinary and Advanced level exams taken, and the selectivity of the school. Finally, Bowles and Nelson (1974) found that between 33 and 60 percent of the covariation of parental economic status and respondent's income (not its logarithm) was not accounted for by the statistical association of parental status with childhood IQ of years of schooling.

It is possible, of course, that $\beta_{y_o y_p}$ measures the effects of wealth inheritance (none of the three studies includes measures of parental wealth or bequests) directly on earnings of indirectly via effects on unmeasured determinants of earnings, such as quality of schooling. But this seems unlikely given the very small number of individuals receiving any significant bequest in the samples in question. A more cautious interpretation is that it indicates an important gap in our understanding of how economic status is passed on from generation to generation: roughly half of the intergenerational transmission coefficient is unaccounted for. It is also true that typically one can statistically account for less than half of the variance of the earnings or income using the conventional variables described above, but this fact does not explain the limited success in accounting for the intergenerational correlation, as this measures only that part of the variation of earnings that can be explained statistically by parental economic status.

5. *The Relative Importance of Genetics, Culture, and Bequests*

The face that the earnings and schooling attainments of identical twins (more technically, monozygotic twins, which is abbreviated to 'mz' in equations) appear to be substantially more similar than fraternal twins (or dizygotic, which is abbreviated to 'dz' in equations) suggests that genetically inherited traits other than cognitive skills may account for some of the intergenerational correlation of earnings. Taubman (1976) uses a sample of monozygotic and dizygotic twins to measure the relative importance of genes and environment in income determination. Figure 4 presents data from the Twinsburg, US sample, the Australian Twin Register, and the Panel Study of Income Dynamics (US). Direct inferences about the degree of heritability of income from these data are likely to be misleading for the following reasons. First, the identical twins appear to have shared more similar environments than dizygotic twins and other siblings (Loehlin and Nichols 1976), so environmental differences contribute to the greater similarity of the monozygotic twins. Second, the twin data are not corrected for errors in measurement. An adequate treatment of this problem requires information on the degree of correlation of errors, which seems likely to be higher for the identical than the other twins. If this supposition is correct, error correction would increase the monozygotic correlations by less than the dizygotic correlations. Third, the identical twins are all same-sex pairs, while the dizygotic twins include mixed-sex pairs, and sex contributes to the within pair difference in economic success. However the difference in the earnings correlation between monozygotic and same-sex pairs of dizygotic twins in the NRC (Taubman 1976) sample and the Swedish Twin Registry (Isacsson 1999) is 0.24 (in both data sets). The brothers' data set addresses this latter issue, but includes, as social siblings, unrelated members of the same household, so the genotypic correlation among them is less than 0.5. All three problems imply that the naïve calculation of heritability from these data will constitute an overestimate.

It may be of interest, nonetheless to see what this estimate yields. Assuming the degree of similarity of the environment influence on earnings and schooling attainments is identical for the two sets of twins, one can calculate the heritability of these phenotypic traits

Twin and Sibling Earnings Similarity, r_y

Study	Monozygotic Correlation	Dizygotic Correlation	Brothers' Correlation
Twinsburg[a]	0.56	0.36	
Twinsburg[b]	0.63	0.37	
Australian Twins[c]	0.68	0.32	
U S Brothers[d]			0.45

Figure 4
Earnings and genetic similarity of monozygotic twins, dizygotic twins, and non-twin brothers.
Notes: (a) Ashenfelter and Kruger (1994), with earnings measure log wage; (b) Rouse (1999) and personal communication, with earnings measure log wage; (c) Miller et al. (1995), with earnings measure log occupational income; (d) Bjorklund and Jantti (1999) with earnings measure log annual earnings

as $h_i^2 = 2(r_i^{mz} - r_i^{dz})$, where r_i^{mz} and r_i^{dz} are the correlations of the monozygotic and dizygotic twins for trait i. Estimates of h_i and β_{ie_i} are not very sensitive to assumptions concerning the degree of assortative mating, measured by the correlation of parental phenotypes, m_i but depend critically on assumptions concerning the difference in the environmental correlations experienced by the dizygotic and monozygotic twins, $r_{e_i}^{mz}$ and $r_{e_i}^{dz}$, respectively. Relaxing these assumptions requires that h_i and β_{ie_i} be estimated simultaneously, using the equations below, where the subscript i refers to the trait in question

$$r_i^{mz} = \beta_{ie_i}^2 r_{e_i}^{mz} + h_i^2$$
$$r_i^{dz} = \beta_{ie_i}^2 r_{e_i}^{dz} + h_i^2(1 + h^2 m_i)/2$$

The greater the difference in the environmental correlations, the lower is the estimate of h_i^2 and the higher is the estimate of $\beta_{ie_i}^2$. Correspondingly, larger values of m_i increase the estimate of h_i^2 and decrease the estimate of $\beta_{ie_i}^2$. If the correlation for earnings-affecting environments for monozygotic and dizygotic twins are 0.8 and 0.7 respectively (rather than both being 0.8), and if $m_i = 0.25$, the estimates reported obtain. Twin correlations from Ashenfelter and Krueger (1994) for hourly wages and from Behrman and Taubman (1989) for years of schooling attained yield estimates of $h_i^2 =$ 0.4 both traits. Assuming that the correlation of parental genotypic and phenotypic income is h_y itself, the estimated contribution genetic inheritance to the intergenerational transmission of these measures of status is just $(r^g h_y^2)$ or one-fifth. While this estimate may overstate its influence, it is not doubted that genetic inheritance plays a role, perhaps a substantial role in the intergenerational transmission of economic status, even if (as has been seen) this process operates substantially independently of the cognitive skills measured on available tests.

Similar methods yields estimates for the environmental effects. If the correlation of parental income and the environmental influences on offspring income is approximated by the correlation of parental (permanent) income or other measures of socioeconomic status with offspring schooling attainment, the estimate of $r_{e_o y_p}$ is about 0.45. An estimate of β_{le}^2 is obtained from the Ashenfelter–Krueger twin data as $(r^{mz} - h^2)/r_e^{mz}$ where r_e^{mz}, the correlation of monozygotic twins' environments, is assumed to be 0.8, giving a value of 0.2. Thus, the environmental contribution to the intergenerational correlation of earnings $(\beta_{le} r_{e_o y_p})$ is the same magnitude as the genetic contribution, namely 0.2, giving an intergenerational correlation for earnings of 0.4.

Note that $\beta_{le} = \sqrt{0.2} = 0.45$ is an estimate of the effect on earnings, in standard deviation units, of a standard deviation change in the environmental influences on human capital and hence on earnings, and may be compared with β_{ys}, the mean estimate of which (Bowles et al. 2001) is just half of β_{le}, suggesting that

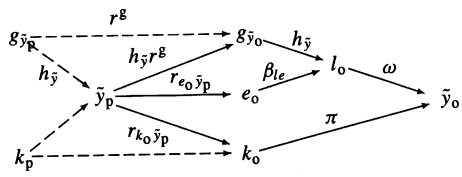

Figure 5
An expanded model of intergenerational transmission of income. The dashed lines represent the causal model but are not used in calculations

while educational attainment captures important aspects of the relevant environments it is far from inclusive.

To complete the analysis we turn to the intergenerational correlation for income, including return to property, and thus take account of the intergenerational transmission of assets. Figure 5 illustrates the causal model of the determination of income underlying Eqns. (2) and (4), with the paths from y_p through g_{y_o} and e_o to y_o depicting the transmission process for earnings addressed in the paragraphs immediately above. The contribution of wealth inheritance to the transmission process is simply the correlation of parental income and offspring wealth $(r_{y_p k_o})$ multiplied by the contribution of offspring wealth to offspring income, expressed as a normalized regression coefficient, π. An estimate of π is simply γ, the percentage change in income associated with a one percent change in assets suitably normalized. If the rate of return on one's assets is constant, γ is simply the share of income from assets as a fraction of total income, (e.g., if the rate of return on capital is 7 percent, and the average capital holding is three times the average income, γ is about 0.2). Expressing π as a normalized regression coefficient we have

$$\pi = \gamma \frac{\sigma_k}{\sigma_{\tilde{y}}}$$

The inequality of property holdings is considerably greater than the inequality of earnings or of income, and as a result γ is likely to differ among families, with values approaching zero for the asset poor, and approaching unity for the very wealthy. This implies that importance of wealth inheritance in the intergenerational persistence of economic status is likely to differ by wealth level. If the variance of the logarithm of assets is four times the variance of the logarithm of incomes (the variance of a logarithm is a common unit free measure of dispersion) we have $\pi = 0.2(2) = 0.4$. Finally if $r_{y_p k_o}$ is midway between the intergenerational correlations for income (0.43) and for wealth (0.50), or 0.46, the property inheritance contribution to the intergenerational income correlation becomes $0.2(2)(0.46) = 0.18$.

	Earnings	Income
environmental	$r_{e_o\tilde{y}_p}\beta le = 0.2$	$\omega r_{e_o\tilde{y}_p}\beta le = 0.14$
genetic	$h_y^2 r^g = 0.2$	$\omega h_y^2 r^g = 0.14$
wealth		$\pi r_{k_o\tilde{y}_p} = 0.18$
intergenerational correlation	= 0.4	= 0.46

Figure 6
Contribution of environmental, genetic, and wealth inheritance to intergenerational transmission

To calculate the contribution of g_o and e_o to the intergenerational correlation of income one simply multiplies the contributions estimated above by the path from earnings (or human capital) to income. One needs to assume that $h_y \cong h_{\tilde{y}}$ and $r_{e_o y_p} \cong r_{e_o \tilde{y}_p}$. This (following the above reasoning) is estimated as

$$\omega = (1-\gamma_o)\frac{\sigma_l}{\sigma_{\tilde{y}}}$$

If the variance of human capital is four-fifths as large as the variance of income, then $\omega = 0.72$. Of course this value would be close to unity for population groups with no source of income other than earnings (if $\gamma_o = 0$, then $\sigma_{\tilde{y}} = \sigma_l$), and could be much smaller among the wealthy.

The implied contributions to the intergenerational persistence of income based on the above assumptions are genetic = 0.14, environmental = 0.14, asset based = 0.18, giving an intergenerational correlation of 0.46 which, as one would expect, somewhat exceeds the analogous correlations for earnings alone. These results are illustrated in Fig. 6.

The substantial contribution of wealth inheritance to intergenerational transmission of economic status among the wealthy is unsurprising, and the mechanisms through which it works relatively transparent. The same cannot be said for the genetic and environmental contributions, for as has been seen, neither IQ nor schooling provides an adequate account of these influences. Thus, both genetic and environmental influences remain black boxes. We know they are important but we do not know why.

6. Conclusion

This article has shown that recent evidence points to a higher level of intergenerational transmission of economics position than previously thought. Moreover, although the level of intergenerational IQ inheritance is also considerable, the latter accounts for little of the former. Indeed, some combination of environmentally and genetically transmitted noncognitive personality traits probably account for most of the correlation between the economic positions of parents and children. Personality differences, like group membership, affect earnings and display parent–offspring similarity. They thus join the genetic and cultural inheritance of cognitive skills, property bequests, and other influences of wealth, and parent–offspring similarity in schooling attainments as aspects of the process of intergenerational transmission of economic status.

We do not know, of course, if a more inclusive view of the transmission process and of the traits being transmitted taking account of group effects, personality, and other aspects would substantially improve this empirical account of the intergenerational correlation. For many of the traits identified data are available that would permit an assessment of their importance in the transmission process.

However, for many of the personality variables, there is a lack good research on the transmission process. Heritability estimates for these variables are quite reliant on comparisons of monozygotic and dizygotic twins which may substantially over-estimate heritability, and consequently underestimate the role of vertical cultural transmission and other influences of siblings' shared (family) environments. It is known that schooling contributes to cognitive functioning independently of genetic inheritance but that most of the contribution of schooling to economic success is unrelated to the learning of cognitive skills in school (Bowles et al. 2001, Gintis 1971). It seems likely that schools contribute to economic success in part by fostering personality traits rewarded in labor markets, rather than simply by enhancing cognitive functioning, but this appears not to be the case for either the Rotter Scale or the social maladjustment measures studied by Osborne.

See also: Behavioral Genetics: Psychological Perspectives; Class: Social; Familial Studies: Genetic Inferences; Gender, Class, Race, and Ethnicity, Social Construction of; Social Inequality in History (Stratification and Classes); Social Stratification; Socioeconomic Status and Health

Bibliography

Ashenfelter O, Krueger A 1994 Estimates of the economic return to schooling from a new sample of twins. *American Economic Review* **84**(5): 1157–72

Ashenfelter O, Rouse C 2000 Schooling, intelligence, and income in America. In: Arrow K, Bowles S, Durlauf S (eds.) *Meritocracy and Economic Inequality*. Princeton University Press, Princeton, NJ, pp. 89–117

Atkinson A B, Maynard A K, Trinder C G 1983 *Parents and Children: Incomes in Two Generations*. Heinemann, London

Bardhan P, Bowles S, Gintis H 2000 Wealth inequality, credit constraints, and economic performance. In: Atkinson A, Bourguignon F (eds.) *Handbook of Income Distribution*. North-Holland, Dortrecht, The Netherlands

Becker G S, Tomes N 1986 Human capital and the rise and fall of families. *Journal of Labor Economics* **4**(3): S1–39

Behrman J R, Taubman P 1989 Is schooling 'mostly in the genes'? *Journal of Political Economy* **97**(6): 1425–46
Behrman J R, Taubman P 1998 The intergenerational correlation between children's adult earnings and their parent's income: Results from the Michigan panel survey of income dynamics. *Review of the Income and Wealth* **36**(2): 115–27
Bishop J 1991 Achievement, test scores and relative wages. In: Kosters M (ed.) *Workers and their Wages*. AEI Press, Washington, DC, pp. 146–86
Bjorklund A, Jantti M 1999 *Intergenerational Mobility of Socioeconomic Status in Comparative Perspective*. Stockholm University and Abo Akademi University
Bjorklund A, Eriksson T, Jantti M, Raaum O, Osterbacka E 1999 Brother correlations in earnings in Denmark, Finland, Norway and Sweden compared to the United States, unpublished
Blau P, Duncan O D 1967 *The American Occupational Structure*. Wiley, New York
Bouchard T J Jr., McGue M 1981 Familial studies of intelligence. *Science* **212**: 1055–9
Bowles S 1972 Schooling and inequality from generation to generation. *Journal of Political Economy* **80**(3): S219–51
Bowles S, Nelson V 1974 The 'inheritance of IQ' and the intergenerational reproduction of economic inequality. *Review of Economics and Statistics* **56**(1)
Bowles S, Gintis H, Osborne M 2001 The determinants of individual earnings: Skills, preferences, and schooling. *Journal of Economic Perspectives*, forthcoming
Conlisk J 1990 Monotone mobility matrices. *Journal of Mathematical Sociology* **15**(3–4): 173–91
Cooper S, Durlauf S, Johnson P 1994 On the evolution of economic status across generations. *American Statistical Association Papers and Proceedings* 50–8
Corak M, Heisz A 1999 The intergenerational earnings and income mobility of Canadian men: Evidence from the longitudinal income tax data. *Journal of Human Resources* **34**(3): 505–33
Dardoni V 1993 Measuring social mobility. *Journal of Economic Theory* **61**: 373–94
Devlin B, Daniels M, Roeder K 1997 The heritability of IQ. *Nature* **388**: 468–71
Eysenck H J 1994 Intelligence and introversion–extraversion. In: Sternberg R J, Ruzgis P (eds.) *Personality and Intelligence*. Cambridge University Press, Cambridge, UK, pp. 3–31
Fong C 2000 Essays on endogenous preferences and public generosity. Ph.D. thesis, University of Massachusetts
Galor O, Zeira J 1993 Income distribution and macroeconomics. *Review of Economic Studies*. **60**(1): 35–52
Galton F 1889 *Natural Inheritance*, Macmillan, London
Gintis H 1971 Education, technology, and the characteristics of worker productivity. *American Economic Review* **61**(2): 266–79
Goldberger A S 1989 Economic and mechanical models of intergenerational transmission. *American Economic Review* **79**(3): 504–13
Hauser R M, Warren J R, Huang M-H, Carter W Y 2000 Occupational status, education, and social mobility in the meritocracy. In: Arrow K, Bowles S, Durlauf S (eds.) *Meritocracy and Economic Inequality*. Princeton University Press, Princeton, NJ, pp. 179–229
Isacsson G 1999 Estimates of the return to schooling in Sweden from a large sample of twins. *Labour Economics* **6**: 471–89
Jencks C 1979 *Who Gets Ahead?* Basic Books, New York
Kanbur R, Stiglitz J E 1986 Intergenerational mobility and dynastic in equality. Princeton Econometric Research Program Memorandum.
Loehlin J, Nichols R 1976 *Heredity, Environment, and Personality*. University of Texas Press, Austin, TX
Miller P, Mulvey C, Martin N 1995 What do twin studies reveal about the economic returns to education: A comparison of Australian and U.S. findings. *American Economic Review* **85**(3): 586–99
Mulligan C 1997 *Parental Priorities and Economic Inequality*. University of Chicago Press, Chicago
Mulligan C 1999 Galton vs. the human capital approach to inheritance. *Journal of Political Economy* **107**: S184–224
Otto S P, Christiansen F B, Feldman M W 1995 Genetic and cultural inheritance of continuous traits. Working Paper No. 0064, Morrison Institute for Population and Resource Studies, Stanford University, CA
Plomin R 1999 Genetic and general cognitive ability. *Nature* **402**(2): c25–9
Rouse C E 1999 Further estimates of the economic return to schooling from a new sample of twins. *Economics of Education Review* **18**: 149–57
Solon G R 1992 Intergenerational income mobility in the United States. *American Economic Review* **82**(3): 393–408
Solon G R 2000 Intergenerational mobility in the labor market. In: Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*. North-Holland, Amsterdam
Sternberg R J, Wagner R K, Williams W M, Horvath J 1995 Testing common sense. *American Psychologist* **50**(11): 912–27
Taubman P 1976 The determinants of earnings: Genetic, family, and other environments; a study of white male twins. *American Economic Review* **66**(5): 858–70
Thorndike E L 1919 Intelligence and its uses. *Harper's Monthly Magazine* **140**: 227–35
Thorndike R, Stein S 1937 An evaluation of the attempts of measure social intelligence. *The Psychological Bulletin* **34**(5): 275–85
Warren J, Hauser R 1997 Social stratification across three generations: New evidence from the Wisconsin longitudinal study. *American Sociological Review* **62**: 561–72
Williams W M, Sternberg R J 1995 *Success Acts for Managers*. Harcourt Brace, Orlando, FL
Williamson J, Lindert P 1980 *American Inequality: A Macroeconomic History*. Academic Press, New York
Winship C, Korenman S 1999 Economic success and the evolution of schooling with mental ability. In: Mayer S, Peterson P (eds.) *Earning the Learning: How Schools Matter*. Brookings Institution, Washington, DC, pp. 49–78
Zimmerman D J 1992 Regression toward mediocrity in economic stature. *American Economic Review* **82**(3): 409–29

S. Bowles and H. Gintis

Economic Transformation: From Central Planning to Market Economy

The analysis of transformation from central planning into a market economy is one of the youngest fields of geographical research. While in the 1980s neither the overwhelming majority of theoreticians nor practical experts could conceive a collapse of socialism, nowadays library shelves are stuffed with literature dis-

cussing the transformation of Central and Eastern European countries. As a research field, transformation analysis was not merely new but also unexpected. This is why descriptive case studies predominate its ever increasing literature, while its theoretical foundations and terminological background are rather defective.

1. Transformation into the Market Economy: Concepts, Models, Myths

More than ten years after the collapse of the central planning economy, transformation analysis is still analysis is still characterized by controversial concepts and models, misconceptions, and myths. A short review of these delusions is inevitable.

1.1 The Concept of Transformation into a Market Economy

Transformation into a market economy involves a process in which the Central and Eastern European countries, having drifted to the periphery of the world economy after 1945, transform their economic systems from central planning into market economy, in step with the transition of highly industrialized countries.

Consequently, transformation into a market economy involves a threefold process:

a fundamental change of the economic and social system, that is, the elimination of the central planning system and the building up of the market economy;

an economic catch-up of former socialist countries; and

an adjustment to the highly industrialized countries' transition from a Fordist-type mass production into a 'post-Fordist' type flexible production.

1.2 Models of Transformation into a Market Economy

Highlighting the individual subprocesses and complexity of transformation into a market economy is of significance because at the moment unidimensional models prevail in the analysis of transformation into a market economy. Two such models have become widespread: the model of delayed development and the model of stages.

The model of delayed development assumes that, with a significant delay in time, subsequent to 1990 the countries of Central and Eastern Europe started following the same course of development as trodden by highly industrialized countries ever since the mid-1970s. Among others, Dostal and Hampl (1996), Enyedi (1996) and Gorzelak (1996) are all of the opinion that in the 1950s and 1960s basically similar models of extensive economic development became prevalent in highly industrialized countries and socialist countries. In the early 1970s, they think, the courses of these developmental models began to move away from one another. On the one hand, from this time on, highly industrialized countries left this course and the Fordist mass production characterizing the extensive period, and adopted an increasingly flexible production, thus shifting to the so-called intensive (or 'post-Fordist') period. On the other hand, socialist countries went on to maintain their earlier economic development model. What is more, they even intensified it to an 'over-extensive kind of development,' as Dostal put it.

Due to the discrepancy between the models of development, by the 1980s the gap between the Western and the Central and Eastern parts of Europe had widened and the process finally led to the collapse of the socialist system in 1989–90. In the wake of this political breakpoint, each single country in Central and Eastern Europe sank into a deep 'transformational recession' and they got out of it only in the middle of the 1990s. Dostal, Hampl, Enyedi and Gorzelak think that, from this economic turning-point, i.e., from the end of the transformational recession and the beginning of a consolidation of market economy, the Central and Eastern European countries have also been following the post-Fordist flexible production model.

At this point the critical remark needs to be made of the model of delayed development: namely that it completely neglects the fact that despite certain similar features in terms of extensive economic development, there are fundamental differences between a central planning system and a free market economy. Moreover, it must be stressed that delay in development processes is not tantamount to a mere time delay, but also involves economic, social, and regional distortion of the processes, for the shapes these processes take in the former socialist countries are rather different from their Western counterparts.

In contrast to the model of delayed development, the other unidimensional model, that of stages, does not focus on historical continuity, but on the momentum of breaking away from the previous development course. Fassmann (1997) says that the model of stages may be divided into three stages. The first one is the initial state of affairs of transformation into market economy, namely the period of central planning. The next stage is an intermediary stage: a gradual breakdown of the central planning system and a parallel building up of the market economy. The third stage of the model includes stabilization of the market economy. In Fassmann's view, the intermediary stage may be further divided into two substages. One of them is the period of transformational measures, in which the élite in control of transformation into the market economy makes the most important economic decisions (e.g., decisions concerning the liberalization of the economy, the privatization of state-owned property, the discontinuance of government subsidies, etc). The next stage

is a later period of transformational phenomena, a time when the results of economic political measures become manifest in practice (e.g., in restructuring foreign trade, the changes of ownership statuses, increased unemployment, and emerging enterprises).

All the above said notwithstanding, fault is also found with the stage model. Following the scheme of causality, it is seen that the model overemphasizes the role of its actors and attributes the regional processes of transformation into a market economy to a single factor: the activity of the participants of transformation. This is to say that on the one hand, the stage model neglects the—incredibly long-lasting, in some cases several centuries old—regional disparities, which had a significant impact on the regional processes of the transformation into the market economy after 1990. Another factor the stage model overemphasizes is the significance of the goals and ideals of the élite that steers the transformation into a market economy. Though these objectives undoubtedly influenced the participants' activities, it does not follow that the actual reality of transformation into a market economy would or should have any kind of palpable objectives or a final status whatsoever.

1.3 The Myths of Transformation into a Market Economy

The individual subprocesses and complexity of transformation into a market economy are worth stressing also, because certain myths still make their influence felt in research. In the first place there are three particularly powerful myths: the overvaluation of the regional policy pursued in the socialist era, the unique historical character of the transformation into a market economy, and the significance of global processes.

In the period between 1960 and 1990 the political leaders of the countries of Central and Eastern Europe declared an egalitarian regional policy, which claimed to reduce dramatically or even eliminate regional disparities. However, there was an enormous gap between the declared objectives and reality: the socialist regional policy based on a completely centralized redistribution of development resources—presumed to be purposeful—yielded fictitious results. As Meusburger and Klinger have pointed out, the socialist regional policy somewhat mitigated regional disparities between living standards and residential infrastructure but it was simply incapable of reducing the disparities in really tough local factors, in particular in the field of human resources.

Discernment is further impeded by the fact that in the countries of Central and Eastern Europe before 1990 regional statistics were frequently and deliberately manipulated, forged, and framed in categories incomparable to those used in highly industrialized countries. Socialist regional policy rather obscured and concealed the internal regional disparities of the countries of Central and Eastern Europe, instead of reducing or eliminating such disparities. An uncritical acceptance of the self-image offered by socialist regional policy and an overestimation of its performance imply the danger that the regional consequences of transformation into the market economy are studied from a false point of view, and that the effects of regional disparities brought along from the socialist era, which have a significant impact on transformation into market economy, are neglected.

The second myth related to the transformation into a market economy (namely, the overvaluation of the unique historical character of the transformation process) is an unintentional 'children's disease' of transformation research. It does not follow, from the mere fact that transformation from a central planning system into a market economy is a unique process of economic history, that this process cannot be described and interpreted within the framework of concepts and processes characterizing the economy of highly industrialized countries. The prime reason why this myth is to be considered dangerous is that it depicts the processes that arose after 1990 as if they were something completely new in the region.

Just to give an example: for ideological reasons, in the countries of Central and Eastern Europe there had been no unemployment in the socialist era. To be more precise, excess manpower was employed—from state subsidies—at factories. Actually, there already was significant hidden unemployment in the 1980s, but the official socialist policy declared it nonexistent. After 1990, hidden unemployment suddenly became manifest, as a result of emerging market conditions and opening the economy to the world market. In consequence, unemployment figures skyrocketed in each country of Central and Eastern Europe, something that was conceived as a new and unprecedented process, and as a rule, new phenomena were frequently interpreted as symptoms of the newly emerging regional disparities. Based on analyses of East Germany and Hungary, Cséfalvay (1997) pointed out that the regional pattern of unemployment was nothing new at all—rather it rested on the different processes forcing people, regions, and countries to the periphery—and these had existed already in the socialist era.

The third myth of transformation research includes the overvaluation of the role played by global processes in the regional processes of transformation into a market economy. This myth takes the planet Earth for a mere arena of general processes, and presumes that such processes should appear in every country and region sooner or later. Historical experience shows, however, that the Earth is not an arena for global processes, rather it is a multifaceted mosaic of regions with divergent economic strength and development opportunities. Plenty of the processes that emerged in Central and Eastern Europe in the course

of its transformation into the market economy and have been taken for global processes actually differ markedly from the processes characteristic of highly industrialized countries.

An extremely fast tertiarization took place, for instance, in all the countries of Central and Eastern Europe after 1990. While the rate of people employed in the service provision sector remained below 45 percent before 1990, the corresponding figure was between 55 percent and 60 percent in the mid-1990s. In economic and regional studies it has become a matter of doctrine to consider an increase in the rate of employment in the tertiary sector as one of the major indicators of postindustrial development, and to see a close connection between economic progress and the high proportion of people employed in the service-provision sector. However, in Central and Eastern Europe the sudden increase of the rate of people employed in the service provision sector was not an indication of postindustrial development at all; it was merely the outcome of a succession of winding ups and the closing down of nonprofitable socialist factories, a single, rapid, and dramatic deindustrialization. The example of Central and Eastern European countries is indicative of the fact that there may be very different explanations behind tertiarization—a process considered to be global. On the one hand, the assumption that a high rate of people employed in the tertiary sector involves a high capability of economic performance does not hold true for Central and Eastern Europe, for here GDP per capita is less then 50 percent of the average in the European Union. On the other hand, the underlying reasons for deindustrialization, which actually induce tertiarization, are markedly different. While in Western Europe these processes are triggered by a relocation of mass production, in Central and Eastern Europe the main reason is the disintegration of socialist industry. Finally, another significant difference is that in Western Europe the process of tertiarization took several decades, but in Central and Eastern Europe it required only a few years.

2. *The Consequences of Transformation into the Market Economy in Central and Eastern Europe*

All the subprocesses of transformation into the market economy—like the change in the economic system, catch-up, and the process of adjustment—resulted in peculiar regional consequences in the different countries of Central and Eastern Europe.

2.1 *The Regional Processes of Transformation into the Market Economy*

The change that took place in economic system, the disintegration of the centrally planned economy, and the evolution of the market economy—particularly the privatization of government-owned enterprises and the liberalization of trade—triggered a dramatic revaluation of the individual regions. Thus, subsequent to 1990, three regional processes of extreme importance took place simultaneously. In the centers of former socialist large-scale industry, privileged before 1990, the elimination of inefficient plants resulted in the evolution of crisis zones. Simultaneously, since mass dismissals affected primarily those who had low qualifications, the areas poor in human resources sank down to the level of a periphery. Finally, selective dismissals suddenly raised the value of large cities, for these have favorable human resource features and they are centrally located. With respect to their process in the course of time, the majority of these events of regional concern took place in the first half of the 1990s, thus most Central and Eastern European countries have left this stage behind.

Another subprocess of transformation into the market economy includes the efforts made for catching up with the economic achievements of the West. This endeavor took shape in two fundamental regional processes: in the strengthening of the nation-state concept at a high level of organization and in a loosening of the earlier closed borders of nation states. The concept of the strengthening of the nation state at a high organizational level and the renaissance of nation states in Central and Eastern Europe—and in particular in South Eastern Europe—are generally explained by a revival of ethnic conflicts. Even though ethnic conflicts are deeply rooted in the history of these countries, catch-up efforts are highly perceptible in the present renaissance of nation states. This is something to do with the fact that the different countries of the region had set their admission to the European Union as a primary goal. However, it soon turned out that 'nonproblematic' small states had much greater chances for an early admission than states under internal ethnic tensions or those loaded with seriously underdeveloped regions (i.e., the peaceful separation of the Czech Republic and Slovakia in 1993). Nevertheless, the renaissance of nation states is still a rather controversial phenomenon. This is because Central and Eastern European countries have achieved independence and sovereignty—after many decades or for the first time in history—exactly at the time they have to give up an increasing number of elements of sovereignty as a result of Europe's increasing integration and to the benefit of a supranational level of organization.

The other regional consequence of catch-up processes involves the loosening of earlier strictly-closed borders, which ran its course in a similarly controversial manner in Central and Eastern Europe. Though political motifs had significant roles in this process as well—just think of the demolition of the Berlin Wall and the Iron Curtain in 1989—economic incentives went deeper. With borders closed no Euro-

pean integration or economic catch-up may be realized. Only free crossing—the flow of resources, capital, goods, labor force, and information—may create chances for Central and Eastern European countries for a catch-up with the West. In Central and Eastern Europe, however, the majority of borderland areas have become mere passageways of the crossroads type except for the single area of the Austrian-Hungarian border region, which has turned into a rapidly growing economic zone. Capital, labor force, and information all find their ways through such crossroads as well. However, the flow of resources is not targeted at the borderlands themselves; they merely pass such regions and go on to farther located centres.

Finally, the third subprocess of transformation into the market economy includes adjustment to the transition of highly industrialized Western countries. This also triggered differentiated regional processes in Central and Eastern Europe. In the last two decades of the twentieth century the industrialized countries of the West saw a transition characterized—at least from the point of view of the adjustment Central and Eastern European countries—by two fundamental processes: the penetration of flexible production methods and the relocation of the labor force and cost-intensive mass production. Thus Central and Eastern European countries have got three alternatives from which to choose for adjustment. Fordist reindustrialization means the acceptance of relocated high-tech industrial mass production from highly-developed countries. Another way is given in the post-Fordist neoindustrialization or the promotion of naturalizing flexible production methods. The third alternative offers the possibility of developing post-Fordist central regions primarily by strengthening the transfer of capital and knowledge in the large cities. From among these three alternatives for Fordist reindustrialization, or the settlement of labor-intensive industrial production, is unambiguously predominant, while post-Fordist neoindustrialization and the evolution of post-Fordist regional centers are isolated phenomena in Central and Eastern Europe.

2.2. *The Regional Pattern of Transformation into a Market Economy in Central and Eastern Europe*

The best known and the most popular regional pattern of Central and Eastern European transformation into the market economy was created by Gorzelak (1996). In Gorzelak's view the losing and winning regions of transformation from centrally planned economies into market economies follow the traditional west vs. east developmental division of Central and Eastern Europe suit. The losing regions along the eastern borders of Poland, Slovakia, and Hungary constitute a single continuous underdeveloped zone he calls the 'eastern wall'. In contrast to these, the winning regions are located mainly in the western part of Central and Eastern Europe. If the winning cities are connected by a fictitious line, the form of a 'boomerang' is given, commencing from the port of Gdansk, proceeding towards the west through Poznan and Wroclaw to Prague and then, following the curve of the boomerang to pass Brno, Vienna, and Bratislava down to Budapest. However, this boomerang is not a consistent formation because the Budapest–Bratislava–Vienna triangle has a special role within it. This is the region where the speediest transition to a market economy was experienced and which had become growth regions already by the middle of the 1990s. The northern arm of the boomerang, however, is more of a fictitious supplementation of this developmental zone, while Warsaw and Lodz have been left out of its merely on account of their geographical distances.

Already its shape suggests that this 'boomerang' wishes to be a counterpart of the well-known western 'blue banana' (i.e., the primary economic zone of European urbanization that stretches from southeast England through the Ruhr area to the large agglomeration in South Germany and North Italy) and therefore assumes that such a developed zone evolved in Central and Eastern Europe after 1990. As a critical remark it must be mentioned, however, that the two zones significantly differ from each other in several respects. On the one hand, the highly industrialized zone of the 'blue banana' of Western Europe is not a recent development on our map, for the region at issue has always been the innovative driving force of Europe, whether the point is the spread of literacy in the early Middle Ages, Reformation in the seventeenth century, or industrialization at a later period. As against this, the cities of the 'Central European boomerang' have not invented but received many of these innovations from the 'blue banana', moreover, with a considerable delay in time.

Another significant difference is that the cities of the 'blue banana' are interlinked with myriad economic and social ties; the economic and social relations of the cities making the 'boomerang' are rather feeble, though historical development was similar to the western region. The reason is that the cities of Central and Eastern Europe have never focused on one another; on the contrary, each of them has always had a western orientation. Prague, for instance, was linked with Berlin and Nuremberg, while Budapest focused on Vienna and Munich.

3. *What Has the Future in Store?*

At the turn of the millennium the majority of Central and Eastern countries are over the hump of the most sensitive period of transformation into the market-economy, the period which required the

greatest social sacrifices. Private property has prevailed in the economy, the most important institutions of the market economy have been established, foreign working capital has flowed into the countries and the economy has stabilized. Transformation analysis has been more or less capable of following these changes, though a few questions remain unanswered. In the operational market segments, for instance, we have uncertain knowledge first and foremost of the housing and labor markets. There is the issue of a missing coherent 'grand theory' which would be able to give a consistent framework to the explanation and interpretation of the regional processes and consequences of transformation into the market economy.

However, in the beginning of the twenty-first century the above issues are not the only questions to deal with. Radical new issues will have to be faced by Central and Eastern European countries. One of the most serious challenges of the first few decades of this century will certainly include European integration, and more specifically their admission to the Union. This is a challenge that will certainly trigger whole trains of changes both in the western and in the eastern regions of Europe.

See also: Commodity Chains; Development Theory in Geography; Economic Geography; Economic Restructuring: Geographic Aspects; Ecotourism; Globalization: Geographical Aspects; Globalization: Political Aspects; International Business; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; International Trade: Geographic Aspects; Multinational Corporations; Political Geography; Regulation Theory in Geography; Scale in Geography; Soviet Studies: Economics; Soviet Studies: Geography; World Trade Organization

Bibliography

Cséfalvay Z 1997 *Aufholen durch regionale Differenzierung? Von der Plan- zur Marktwirtschaft: Ostdeutschland und Ungarn im Vergleich*. Erkundliches Wissen 122, Steiner, Stuttgart, Germany

Dostal P, Hampl M 1996 Transformation of East-Central Europe: General principles under differentiating conditions. In: Carter F W, Jordan P, Rey V (eds.) *Central Europe after the Fall of the Iron Curtains, Geopolitical Perspectives, Spatial Patterns and Trends*. Wiener Osteuropastudien 4, Lang, Frankfurt am Main, Germany, pp. 113–28

Enyedi G 1996 New regional processes in post-socialist central Europe. In: Carter F W, Jordan P, Rey V (eds.) *Central Europe after the Fall of the Iron Curtains, Geopolitical Perspectives, Spatial Patterns and Trends*. Wiener Osteuropastudien 4, Lang, Frankfurt am Main, Germany, pp. 129–36

Fassmann H (ed.) 1997 *Die Rückkehr der Regionen. Beiträge zur regionalen Transformation Osmitteleuropas*. Beiträge zur Stadt- und Regionalforschung 15, Österreichische Akademie der Wissenschaften, Wien, Austria

Fassmann H, Lichtenberger E (eds.) 1995 *Märkte in Bewegung: Metropolen und Regionen in Ostmitteleuropa*. Böhlau, Vienna, Austria

Gorzelak G 1996 *The Regional Dimension of Transformation in Central Europe. Regional Policy and Development 10*. Regional Studies Association, Kingsley, London

Hajdú Z (ed.) 1993 *Hungary: Society, State, Economy and Regional Structure in Transition*. Centre for Regional Studies, Pécs, Hungary

Meusburger P, Klinger Á (eds.) 1995 *Vom Plan zum Markt: Eine Untersuchung am Beispiel Ungarns*. Physica-Verlag, Heidelberg, Germany

Strubelt W (ed.) 1996 *Städte und Regionen—Räumliche Folgen des Transformationsprozesses*. Leske–Budrich, Opladen, Germany

Z. Cséfalvay

Economics and Ethics

The economy is a central component of society. To a large extent, it determines our wellbeing, our freedoms and opportunities, our social life, our health, and many other features of our lives. It also determines how equally or unequally these things are distributed across the people. Indeed, the economy is perhaps more important than anything else in determining how our lives go. Consequently, an essential task for ethics is to assess how well the economy is doing and how it might be improved.

Since the working of the economy is complex, and the science of economics is technical, philosophers rarely have the skills needed to progress far with this ethical task. It commonly falls to economists instead. The branch of economics that is devoted to the ethical assessment of the economy is known as welfare economics. It is a sort of applied ethics. Because of its special subject matter, it has its own special methods. But because it is applied ethics, it needs some foundations in ethical theory. Inevitably, welfare economists working back towards the foundations of their subject have encountered moral philosophers working forwards towards applications of theirs. Several particular topics lie within the domains of interest of both welfare economists and moral philosophers. Section 4 gives some examples.

When they look for foundations for their work, many welfare economists have simply adopted an existing ethical theory. They have formalized it—sometimes distorting it in the process—and applied it to their subject. One example comes from the theory of 'optimal tax,' which flourished in the 1970s. Optimal tax theorists tried to devise a scheme of taxation that would best achieve a particular objective. They took the objective from ethical theory. Some (e.g., Mirrlees 1971) adopted the utilitarian objective of maximizing the total of people's wellbeing, others (e.g., Phelps 1973) the 'maximin' objective of maximizing the wellbeing of the worst-off person. The latter group was inspired by the publication in 1972 of John Rawls' *A*

Theory of Justice (Rawls 1972) However, they distorted Rawls' view, because Rawls favored maximizing the worst-off group's access to 'primary goods,' rather than the worst-off person's wellbeing.

Other welfare economists have been less passive and more critical in their attitude to ethical theory. They have themselves contributed to the theory's development. They have found that the methods of economics are useful equipment for doing so. Like most of economics, welfare economics is a rather formal discipline, and it has evolved a range of formal tools to serve its purposes. These tools contribute to the ethical assessment of economies, but they have turned out to be useful for other ethical purposes too. Moreover, theoretical tools from other branches of economics have also turned out to be useful in ethics. Section 3 gives examples of formal methods developed in economics that can be valuable in moral philosophy.

In sum, the connections between economics and ethics are multifarious, and run in both directions. It is therefore not surprising that many important moral philosophers have also been important economists. Adam Smith, Karl Marx, John Stuart Mill, and Henry Sidgwick are examples from the past. As a general rule, economist-philosophers have been utilitarians. Economists in general, too, have generally shown utilitarian inclinations. Since the rise of 'classical' economics in the eighteenth century, economists have mostly been concerned with increasing the wellbeing of the people—a concern that is utilitarian in a broad sense.

Liberalism is another philosophy that is closely associated with economics. However, most liberal economists are liberal for instrumental reasons. Their liberalism derives from their utilitarianism; they believe that economic freedom is a good way to promote people's wellbeing. For example, in *An Inquiry into the Nature and Causes of the Wealth of Nations*, Adam Smith (1776) argued that extensive free trade deepens the global division of labor, which in turn leads to economic growth, and that is good for people. More recently, a similar conclusion has been drawn from some theorems in welfare economics. The theorems show that free economic competition leads to efficiency, defined in a special way, and efficiency is held to be generally beneficial. However, there are many occasions when economic freedom is undoubtedly not generally beneficial. For instance, monopolies may use their power to delay innovations, and companies freely pursuing their own profit may damage the environment. In these cases, economists are typically in favor of restricting freedom by regulation. This shows that their support for freedom is only instrumental.

Utilitarianism, understood broadly, remains the dominant moral philosophy of economics. But within economics, utilitarianism has been radically altered and even emasculated as a result of the progress of economic theory. So close is the association between economics and utilitarianism that these changes have strongly influenced the progress of utilitarian thinking as a whole. Sections 1 and 2 outline the development of utilitarianism within economics.

1. *Utility Theory*

Utilitarianism is an ethical theory, but its development within economics has largely been driven by the development of an associated psychological theory, which has come to be called 'utility theory.' Utility theory is the economist's core account of people's behavior. This section describes it, and Sect. 2 describes its influence on economists' moral theory. Utility theory says that each person chooses her actions so as to maximize her utility. But 'utility' is a term of art, which can be defined in various ways, and different definitions lead to very different sorts of utility theory. This has allowed the theory to undergo a major metamorphosis in the course of two centuries.

Utility theory's origin may plausibly be traced to an article by the mathematician Daniel Bernoulli, published in 1738 (Bernoulli 1738). Bernoulli thought that each person acts to maximize her *emolumentum*. (He wrote in Latin.) *Emolumentum* may be translated as 'benefit' or 'advantage,' so Bernoulli held a version of utility theory in which utility is benefit or advantage. Later, Jeremy Bentham (1789) made the more specific assumption that each person acts to maximize her pleasure. (He treated pain as negative pleasure.) So he held a version of utility theory in which utility is pleasure. (Bentham himself used the term 'utility' differently.) Bentham's view was nothing other than the ancient doctrine of psychological hedonism, but he gave it such detailed analysis that it may now be seen as an early form of the mathematical utility theory we have today.

Utility theory became important in economics only in the 1870s, when the so-called 'neoclassical' school rose to dominance. This was when economists first took a serious interest in the behavior of consumers, and utility theory was the theory of behavior they adopted. At the beginning of the neoclassical school, the economist William Stanley Jevons (1871) explicitly took his utility theory from Bentham, and went far beyond Bentham in his analysis.

Like Bentham, Jevons thought that each person acts to maximize her pleasure. But in defense of this view he said: 'Call any motive which attracts us to a certain course of conduct, pleasure … and it becomes impossible to deny that all actions are governed by pleasure.' Jevons evidently hoped to make his theory of behavior as nearly as possible true by definition. Later economists followed his lead. In the 1930s, in the course of a movement in economics known as 'ordinalism,' it took them to a radically different version of the theory. The doctrine of ordinalism was proclaimed by Lionel Robbins in 1932 (Robbin 1932). Ordinalist utility theory was formally set out in a pair of articles by John Hicks and R. G. D. Allen (1934). It was refined by later

work into what is best called 'axiomatic utility theory.' This is the version that is generally presented in textbooks of economics today.

Axiomatic utility theory is a theory of preferences. Given all the possible acts a person might do, it assumes the person to have preferences amongst them. It treats a preference as a disposition to choose: to prefer one act to another is to be disposed to choose the first over the second if faced with a choice between them. A person's preferences constitute a relation amongst possible acts, and the theory assumes that this relation satisfies a number of axioms. Provided it does so, a formal proof shows that a mathematical function called a 'utility function' can be defined to represent the preferences in a particular sense. The utility function assigns a number called a 'utility' to each of the acts, and it does so in such a way that one act is assigned a greater utility than another if and only if it is preferred to the other. This is the sense in which the function represents the preferences. The dispositional definition of preference ensures that, when faced with a choice between two acts, the person always chooses the one that is preferred. This one has the greater utility. So we may conclude that the person always acts to maximize her utility. In this way, utility theory is confirmed, but only because a person's utility is defined to make it true. Utility is defined as what the person maximizes.

Whereas Bernoulli's and Bentham's utility theories made substantive commitments about what a person maximizes, the axiomatic theory does not. It is not committed to hedonism, or even to an assumption that people always act self-interestedly. Since the commitments of earlier theories were not very plausible, this was a significant advance in economics.

In a sense, ordinalism made utility theory true by definition. However, it did retain some commitments. It is not always possible to define a utility function for a person in a way that makes her a utility maximizer. We can only be sure this is possible if the person's preferences satisfy the axioms of the theory. This is the theory's commitment: that preferences satisfy the axioms. In effect, the axioms ensure that the person can be construed as maximizing something, and utility is defined as what she maximizes.

The principal axiom is transitivity: that if one alternative is preferred to a second, and the second to a third, then the first is preferred to the third. If a person's preference relation is transitive, it constitutes an ordering of the options. The utility function represents this ordering. It puts the options in the same order as the preferences do. Any utility function that puts the options in this same order will represent the preferences equally well. Consequently, many different utility functions can represent the same preferences. The only significant feature of a utility function is its ordering of the options. A utility function defined this way is said to be 'ordinal.' Hence the name 'ordinalism.'

Axiomatic utility theory was later extended by John von Neumann and Oskar Morgenstern (1944) and by Leonard Savage (1954), independently of earlier work on the same lines by Frank Ramsey (1931). The extended theory recognizes that the outcome of an act is normally uncertain. It assumes that a person's preferences among acts are determined by the possible outcomes that may result from each act, and by the likelihood that they will result. It assumes that the preferences satisfy the same axioms as before, and also some more axioms that are concerned with the pattern of outcomes that may result from each act.

As in the simpler theory, a theorem shows that the preferences can be represented by a utility function. But because the extended theory makes more assumptions, it is able to establish stronger conclusions. In the extended theory, the utility function assigns utilities both to acts and to their possible outcomes. As before, one act is assigned a higher utility than another if and only if it is preferred. The special feature of the extended theory is that the utility of an act is the expected utility of its possible outcomes. (Expected utility is defined as the sum of the utilities of the possible outcomes, each multiplied by the outcome's probability of resulting.) Let us call this the 'expectation property' of the utility function. It means that, of two acts, the person always prefers the one that has the greater expected utility. She maximizes expected utility. Once again, this is only because of the definition of utility. Provided a person's preferences satisfy the axioms, the person can be construed as maximizing the expectation of something, and utility is defined as that thing.

The extended theory is known as 'axiomatic expected utility theory.' An important feature of it is that the utility function that represents preferences is not merely ordinal. Within the theory, not only does a utility function order acts in accordance with the preferences, but it also has the expectation property. This turns out to imply that differences—increases or decreases—in utility can be significantly compared with each other. Many different utility functions can still represent the same preferences. But if, according to one of these functions, the difference between the utilities of one pair of acts is greater than the difference between the utilities of another pair, then the difference is also greater according to any other function that represents the same preferences. A utility function in which differences of utility are comparable is said to be 'cardinal.'

2. *Utilitarianism and the Pareto Principle*

Utilitarianism in a broad sense is simply a concern to promote people's wellbeing. More precisely understood, it is the view that one ought to act so as to maximize the arithmetic total of people's wellbeing. Important consequences follow from the precise version that do not follow from the broader one. For

example, suppose we assume that the richer a person is, the less extra wellbeing will be brought her by adding an extra dollar to her wealth. This plausible assumption is known as 'diminishing marginal benefit of wealth.' An argument in favor of equality can be derived from it, together with precise utilitarianism. Diminishing marginal benefit suggests that rich people derive less benefit from extra money than do poor people. So, as extra money becomes available, giving it to poor people adds more to the total of wellbeing than giving it to rich people does. Precise utilitarianism therefore favors sending money towards the poor. This argument for equality is subject to important limitations, but it was nevertheless a significant contribution that economics made to utilitarian thinking. It was popular among economists before the 1930s, including Alfred Marshall and A. C. Pigou.

However, the arrival of ordinalism brought skepticism about the measurement of wellbeing. If we are even to make sense of precise utilitarianism, we must be able to make sense of an arithmetic total of wellbeing. Two conditions are necessary for that. First, each person's wellbeing must be an arithmetic quantity. To be more precise, it must be a cardinal quantity, which is to say that differences—increases or decreases—in a person's wellbeing must be significantly comparable with each other. Second, differences in one person's wellbeing must be significantly comparable with differences in any other person's: we must be able to say whether a particular increase or decrease in one person's wellbeing is greater or less than a particular increase or decrease in another's. Ordinalism made economists skeptical about both these conditions.

The source of their skepticism was epistemic. The ordinalists regarded a person's preferences as observable in principle. Since they defined utility to represent preferences, they regarded utility too as observable in principle. Because of this, they assigned ordinal utility a high epistemic status; Robbins called it a 'scientific' concept. But they were much more doubtful about other features of a person's psychology.

What about wellbeing? The ordinalists seem to have taken it for granted that a person prefers one thing to another if and only if it gives her more wellbeing. This view may be called 'preferencism.' On the face of it, preferencism is implausible. For example, presumably a person might altruistically prefer one thing to another, even if it gives her less wellbeing, because it is better for someone else. But the ordinalist espoused the doctrine of operationalism. According to operationalism, the meaning of a concept such as wellbeing is given by the operations we use in order to apply it. The ordinalists recognized no operation that would allow us to apply the concept of wellbeing except the operation of observing people's preferences. So they thought that the only meaning wellbeing can have is given by the preferencist formula. This means that utility measures wellbeing in an ordinal way: higher utility indicates higher wellbeing. The order of wellbeing therefore counts as a scientific concept, but the ordinalists acknowledged no other aspect of wellbeing. Wellbeing could not be counted as an arithmetic quantity, and nor could differences of wellbeing be compared between people. The preconditions of precise utilitarianism could therefore not be met.

Nevertheless, the ordinalists were utilitarian at heart. Consequently, they clung to the flimsy remnant of utilitarianism that survived their epistemic censure. They accepted the 'Pareto criterion,' named in honor of Vilfredo Pareto (1927). The Pareto criterion says that one state of affairs is better than another if everyone either prefers the first to the second or is indifferent between them, and if someone definitely prefers the first. This sufficient condition for one alternative to be better than another was the sole evaluative principle that the ordinalists were prepared to recognize. Few pairs of alternative economic arrangements will satisfy it, and so will few pairs of alternatives that are interesting elsewhere in ethics. So shrinking utilitarianism to the Pareto criterion is to cripple it. Ordinalism stultified welfare economics for decades. In particular, it denied that any judgment could be made about the value of equality.

A possible escape from the ordinalist straightjacket seemed to be offered by expected utility theory. Because expected utility theory yields cardinal utility functions, some authors thought it might allow wellbeing to be measured cardinally. But this thinking was confused about the relation between utility and wellbeing. Granted preferencism, utility measures wellbeing ordinally. But even granted preferencism and granted also that cardinal utility can be defined within expected utility theory, it does not follow that this cardinal utility measures wellbeing cardinally. Differences in utility are comparable, but it does not follow that they represent comparable differences in wellbeing. Furthermore, the second condition of precise utilitarianism remains unsatisfied. Nothing in expected utility theory makes one person's wellbeing comparable with another's.

Nevertheless, John Harsanyi in the 1950s produced no less than two new and important arguments that were intended to show that, indeed, the cardinal utilities defined by expected utility theory do have a place in ethical judgments.

His first argument (Harsanyi 1953) was this. Suppose we are comparing societies, asking which is the best. Harsanyi argued that this question could be answered by imagining a person who is allowed to choose among the different societies. She enters the one she chooses, but without knowing which position she will occupy within it. She has an equal chance of finding herself in the best position, or the worst, or anywhere else. Harsanyi claimed that the society she would choose under these circumstances is the best.

This is an appeal to the idea of a 'veil of ignorance,' which was later taken up by John Rawls.

From the subject's point of view, she is choosing between acts whose results are uncertain: the result of choosing a particular society will be that she ends up in an unknown position within that society. According to expected utility theory, for this person there will be a utility assigned to each position in the society. The utility for her of entering the society will be the expectation of all these individual utilities. Since she is assumed to have an equal chance of ending up in each position, this expectation is just the average utility of all the positions. So, of the alternatives available, the person will choose the one with the greatest average utility. If we assume every society has the same population, this is also the one with the greatest total utility. According to Harsanyi, then, the best society is the one with the greatest total utility. So this argument does indeed give a place in ethical judgments to utility, as expected utility theory defines it.

Does it also support utilitarianism? Precise utilitarianism implies that the best society is the one with the greatest total of wellbeing. So if utility measures wellbeing cardinally, Harsanyi's conclusion resembles precise utilitarianism. But Harsanyi's argument falls short of a complete argument for utilitarianism, because we still have no reason to think utility does measure wellbeing cardinally. Furthermore, the problem of comparing one person's wellbeing with another's remains.

Harsanyi's second argument (1955) was a formal theorem that is too complex to describe here. It is entirely distinct from the first argument, and it does not rely on a veil of ignorance. Still, it arrives at a similar conclusion, that the best society is the one with the greatest total of people's utility. For the same reasons as before, it does not amount to a complete defense of utilitarianism. Nevertheless, Harsanyi's two arguments cast an important new light on utilitarianism, and went a long way towards providing it with completely new foundations.

These foundations are based on preferences, and they depend on preferencism. Within moral philosophy as well as economics, it has become increasingly popular in recent years to base utilitarianism on preferences rather than, say, pleasure and pain. One reason is the prestige given to preference-based utilitarianism by Harsanyi's impressive work. His work issued, in turn, from the dramatic revision of utility theory caused by ordinalism. These are examples of the strong influence that economics has had on utilitarian thought.

3. *Tools for Ethics*

Economists have developed for their own purposes various formal tools that also turn out to be useful in ethics. Utility theory is one: apart from its contribution to utilitarian thought, it has other significant applications within ethics. For one thing, it constitutes a partial account of practical rationality; it specifies conditions that a person's preferences ought rationally to satisfy. An account of practical rationality is an essential background to ethical theory. For another thing, utility theory can easily be converted from a theory of preferences to a formal theory of the structure of value; its axioms can be made to apply to betterness instead of preferences (see Broome 1991).

Game theory is a second prominent example of a tool useful in ethics. The 'prisoners' dilemma' and related 'games' have led to some radical rethinking of the nature of rationality. Some authors (e.g., Gauthier 1986) have argued that they call for an understanding of individual rationality that somehow embodies an ethic of cooperation. Bargaining theory too—a branch of game theory—has proved useful at the foundations of contractualism (see Gauthier 1986, Binmore 1994, 1998).

A third example is social choice theory. Social choice theory is a branch of welfare economics that is principally concerned with 'aggregation.' It examines how the interests of individuals are aggregated together to determine the overall goodness of a society. One example of a principle of aggregation is the utilitarian principle that overall goodness is the arithmetic total of people's wellbeing.

Social choice theory originated in Kenneth Arrow's *Social Choice and Individual Values*, published in 1951 (Arrow 1951) (see Arrow's theorem). As an ordinalist, Arrow was concerned with aggregating preferences, specifically. However, his conclusion was negative; he demonstrated that people's preferences cannot generally be aggregated together in a plausible way to determine an overall ordering of the alternatives, ranking them from best to worst.

As a result, since Arrow's time, social choice theorists have gone beyond the bounds of ordinalism. They are no longer concerned with aggregating preferences only. A leader in this movement is Amartya Sen, whose *Collective Choice and Social Welfare* both codified social choice theory and set the scene for this new departure (Sen 1970). Ordinalism combined a utilitarian concern for people's wellbeing with skepticism about a quantitative notion of good. One can give up the skepticism, and many economists nowadays unashamedly take for granted a quantitative notion of a person's good or wellbeing. Or alternatively one can give up the utilitarianism and focus concern on something that is less prone to skepticism. Sen has lately been promoting people's capabilities—what they are enabled to do by the resources at their disposal—as the proper focus of concern.

4. *Shared Topics*

Economists and moral philosophers share an interest in several specific, practical subjects. But the co-

operation between the discipline is less than it should be. This section mentions two examples.

4.1 The Value of Life

For practical purposes, welfare economists are faced with the need to set a value on people's lives. For example, railways can be made safer by spending money on them. One benefit of doing so will be that fewer people are killed in train accidents. Economists need to decide how much money it is worth spending to achieve this benefit. So they need to determine how much good is done by saving a person's life or, equivalently, how much harm is done by a person's death.

Since ancient times, philosophers, too, have been interested in the value of life. Some modern discussion (e.g. Nagel 1979 and Feldman 1991) has been stimulated by an argument of Epicurus.' Epicurus tried to show that death is harmless to the person who dies, and the modern issue is still whether death is an evil and whether life has value. Some other modern philosophical writing takes it for granted that life is valuable, and asks more quantitative questions. For example, one issue that has been discussed is whether the value of a life is more affected by events that happen late in life or by those that happen early in life (e.g., Velleman 1991).

This is just the sort of question that should concern economists in their practical work, since it affects the comparative value of saving young people compared with saving old people. So this is a place where ethical theory could contribute to practical economics. Conversely, a more formal framework, of the sort that economic methods could supply, would improve the philosophical argument. But actually there has been virtually no contact between the economists and the philosophers working in this area. Economists have generally circumvented the need for theory by basing their valuations on the preferences people have about risks to their lives, unmoderated by theory. Perhaps because of this untheoretical approach, philosophers have reciprocally ignored the work of economists.

4.2 Equality

For a long time, welfare economists have been concerned with the value of economic equality; examples of important recent contributions are by Atkinson (1970) and Sen (1973). The same subject has also recently become important in moral philosophy; leading works are by Parfit (1995) and Temkin (1993). In this area, too, the contact between philosophers and economists has been limited, and both disciplines would benefit greatly from more.

The philosophers can fairly be said to have discovered an important distinction. Two very different moral concerns can have similar effects. One—generally called 'prioritarianism'—gives priority to improving the wellbeing of badly-off people, over improving the wellbeing of well-off people. Prioritarianism will generally lead one to steer resources towards worse-off people, which will generally have the effect of moving those people closer to equality with the better-off. So it will generally lead one to promote equality, but equality is not its direct object. A different concern is simply to reduce the distance between the better-off and the worse-off; this is sometimes called simply 'egalitarianism.'

The distinction is genuine, but because their effects are similar, it is difficult to specify the precise difference between prioritarianism and egalitarianism. But the formal methods of economics have the resources to do so. Indeed, the economic literature on inequality implicitly contains detailed analyses of both concerns. Despite that, the distinction has not been clearly recognized in economics; it is something that economics needs to learn from philosophy. Conversely, all the formal analysis that is already available within economics could help philosophers clarify the distinction they are making, and better understand its consequences. On the subject of equality, more cooperation would be very fruitful.

To summarize: economics and ethics have learnt a great deal from each other, but they could learn a great deal more.

See also: Ethics and Values; Market and Nonmarket Allocation; Practical Reasoning: Philosophical Aspects

Bibliography

Arrow K J 1951 *Social Choice and Individual Values*. Wiley, New York

Atkinson A B 1970 On the measurement of inequality. *Journal of Economic Theory*. **2**: 244–63

Bentham J 1789 *An Introduction to the Principles of Morals and Legislation*. Oxford University Press, Oxford UK

Bernoulli D 1738 Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 5 [trans. 1954 Exposition of a new theory on the measurement of risk. *Econometrica* **22**: 23–36

Binmore K 1994 *Game Theory and the Social Contract*. MIT Press, Cambridge, MA, Vol. I

Binmore K 1998 *Game Theory and the Social Contract*. MIT Press, Cambridge, MA, Vol. II

Broome J 1991 *Weighing Goods*. Basil Blackwell, Cambridge, MA

Epicurus 1926 Letter to Menoeceus. In: Bailey C (trans./ed.) *Epicurus: The Extant Remains*. Clarendon Press, Oxford, UK, pp. 83–93

Feldman F 1991 Some puzzles about the evil of death. *Philosophical Review* **100**: 205–27

Gauthier D 1986 *Morals by Agreement*. Oxford University Press, New York
Harsanyi J 1953 Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* **61**: 434–5
Harsanyi J 1955 Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* **63**: 309–21
Hicks J, Allen R G D 1934 A reconsideration of the theory of value. *Economica* **14**: 52–76 and 196–219
Hurley S L 1989 *Natural Reasons*. Oxford University Press, New York
Jevons W S 1871 *The Theory of Political Economy*. Macmillan, New York
Marshall A 1920 *Principles of Economics*, 8th edn. Macmillan, New York
Mirrlees J A 1971 An exploration in the theory of optimum income taxation. *Review of Economic Studies* **38**: 175–208
Nagel T 1979 Death. In: Nagel T (ed.) *Mortal Questions*. Cambridge University Press, Cambridge, UK, pp. 1–10
Pareto V 1927 *Manuel D'Economie Politique*, 2nd edn. V. Giard & E. Briere, Paris
Parfit D 1995 *Equality or Priority? The Lindley Lecture*. University of Kansas, KS
Phelps E S 1973 Taxation of wage income for economic justice. *Quarterly Journal of Economics* **87**: 331–54
Pigou A C 1932 *The Economics of Welfare*, 4th edn. Macmillan, London
Ramsey F P 1931 Truth and probability. In: Ramsey F P, Braithwaite R B (eds.) *Foundations of Mathematics and Other Logical Essays*. K. Paul, London
Rawls J 1972 *A Theory of Justice*. Clarendon Press, Oxford, UK
Robbins L 1932 *An Essay on the Nature & Significance of Economic Science*. Macmillan, London
Savage L J 1954 *The Foundations of Statistics*. Wiley, New York
Sen A K 1970 *Collective Choice and Social Welfare*. Holden Day, San Francisco
Sen A K 1973 *On Economic Inequality*. Clarendon Press, Oxford, UK
Sen A K 1987 *The Standard of Living*. Cambridge University Press, Cambridge, UK
Smith A 1776 *An Inquiry into the Nature and Causes of the Wealth of Nations*. Whitestone, Dublin, Ireland
Temkin L S 1993 *Inequality*. Oxford University Press, New York
Velleman D 1991 Wellbeing and time. *Pacific Philosophical Quarterly* **72**: 48–77
von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

J. Broome

Economics, History of

Economics emerged from the more venerable discipline of political economy during the latter half of the nineteenth century. The science of economics offers explanations of phenomena such as money, prices, and the production and exchange of goods. At its broadest sweep, it seeks the means to human happiness and prosperity.

1. Historical Background

While insights on economic phenomena can be found in the works of Plato (c. 427–347 BC), Aristotle (384–322 BC), and the medieval commentaries their work spawned, it was only in the early modern period, coincident with the Scientific Revolution, that a full-fledged discourse emerged. Nicolaus Copernicus (1473–1543), Jean Bodin (1530–96), Thomas Mun (1571–1641), William Petty (1623–87), and John Locke (1632–1704) are some of the more prominent figures of the sixteenth and seventeenth century who wrote on the subjects of money and trade. Copernicus and Bodin articulated the quantity theory of money, that a growth in the money stock results in a rise in prices, a phenomenon observed in Europe since the influx of gold and silver from overseas. Mun, a leading mercantilist and advocate of net exports as the key to England's prosperity, recognized that market forces transcend legal and institutional arrangements, and that genuine wealth comes from the growth of domestic production and the enrichment of the land. Petty was one of the first to devise quantitative measurements of economic phenomena, including per capita output. Locke refined the quantity theory of money, noting the velocity of money, and devised a labor theory of value. He also wedded the historic right to private property with the virtue of industry, and thus launched the liberal doctrine of political economy which subsequently played a profound role in the American constitution.

In the eighteenth century, the most distinguished treatises and essays on political economy were issued by Richard Cantillon (c. 1680–1734), David Hume (1711–76), François Quesnay (1694–1774), Ferdinando Galiani (1728–87), James Steuart (1712–80), and Adam Smith (1723–90), but there were hundreds of lesser known contributors to the subject. Hume and Quesnay were the most influential thinkers prior to Smith. Hume's *Treatise of Human Nature* (1739–40) argued that property claims were artificial and not natural, and thus had no deeper links to justice. Promise-keeping and contracts were privileged as the original steps in the formation of human society. In a series of *Political Essays* (1750), Hume made a number of brilliant insights on money, including his celebrated specie-flow mechanism whereby the gold supply and price level equilibrate via international trade. Hume also recognized that a sudden and unanticipated growth in the money stock can have real effects on output, at least in the interim between the influx of gold and the adjustment of prices. Hume also celebrated the advent of commerce as the means to bring peace and civility. Whereas Locke had highlighted the advent of money in prehistoric times as the critical moment when the accumulation of wealth became ethical, Hume pointed to the much more recent ubiquitous use of money, 'where no hand is empty of it,' as the more critical transition in human history.

Quesnay founded the first school of economics, known as Physiocracy, or the rule of nature. The Physiocrats or *les économistes* maintained that wealth only comes from the gifts of nature, from the fact that one plants one seed in the spring and obtains two in the fall. Wealth thus only truly grows in the agrarian sector, and so all policy measures should be taken to promote the cultivation of the land. All manufacturing activities were deemed sterile, merely transmuting materials but incapable of augmenting them. Quesnay devised one of the first models in economic theory, his celebrated *tableau économique*, which depicted the economy as a circular flow of money and goods between three sectors, the farmers, the artisans, and the landowners. The tableau adumbrated a nascent version of the multiplier via the 'zig-zag' transference of wealth from one sector to another over the course of a year. Physiocracy was widely influential, particularly on the work A. R. J. Turgot (1727–81) and Smith.

Adam Smith's *Wealth of Nations* (1776) canvassed the entire field of economics, including the theories of value and distribution, trade and development, public finance, even economic history and the history of economics. Smith rightly is admired for his concept of the 'invisible hand,' that highlighted the unintended and beneficial consequences of self-interested behavior. The idea that private vices might have public benefits had already been articulated by Bernard Mandeville (1670–1733), in his parody of human society, the *Fable of the Bees* (1714). Smith explored at length the problem of virtuous action in his *Theory of Moral Sentiments* (1759), which grounded human action in a natural capacity for sympathy and the desire for the approval of others. For all his emphasis on the pursuit of wealth, Smith saw it as a hollow dream, the futile pursuit of trinkets and baubles. But because we are deceived into thinking that wealth brings happiness, we work hard and thus keep the world going. Arguably no one has surpassed Smith in weaving such a penetrating and cynical view of human nature into such an expansive theory of material well-being.

Smith's ideas were extended and refined by, most notably, Thomas Robert Malthus (1766–1834), Jean-Baptiste Say (1767–1832), David Ricardo (1772–1823), and John Stuart Mill (1806–73), a tradition now known as the Classical school of political economy (see *Political Economy, History of*). The Classical economists emphasized a cost-of-production (labor) theory of value and the competing claims of the three economic classes (laborers, landowners and capitalists) over the annual product, most aptly depicted as the fall harvest of corn (wheat). Landowners and their servants were engaged in unproductive labor insofar as they did not produce tangible goods for consumption or investment. Malthus advanced a theory of population growth that reinforced the view that capital accumulation would reach a plateau in the not too distant future. Say popularized many of Smith's principles and also argued against the possibility of gluts in production.

Ricardo devised a novel theory of rent, as a return to the original and indestructible powers of the soil. Rent was a pure residual, and did not enter into the formation of prices. Ricardo highlighted the inverse tendencies of wages and profits, and recognized the importance of transfers of capital between different sectors of the economy. He also worked out the principle of comparative advantage in trade, predicated on the international immobility of capital. Mill extended the analysis by introducing nation-specific patterns of demand and rigidities in foreign exchange rates.

For the Classical economists, pricing came about after distribution, which was in turn determined by natural laws, for example, Malthusian laws governing the growth of population and the diminishing returns of the soil. Because of this emphasis on the natural order, the Classical economists tended to view their laws as inexorable, the result of forces deeper then any human action or process of deliberation. Even money was seen as having a motion of its own that transcended human agency. Scarcity, doom, and gloom were prevalent motifs to that set of writings, notwithstanding the remarkable growth rate of the European economy throughout the nineteenth century. Economists worried about the onset of the stationary state, whereby the profit rate fell and capital accumulation came to a halt; overpopulation resulted in a world of 'standing room only.' It is now apparent that they underestimated the potential for technological invention and innovation, both in the form of labor-saving and capital-saving procedures. Even if they did not comprehend the full ramifications of the industrial revolution of the late eighteenth century, there had already been many indicators in the agrarian sector that new techniques of plant breeding, crop rotation, and fertilization could have dramatic consequences for the accumulation of wealth. The reasons for their pessimism are still not fully understood.

2. *Neoclassical Economics*

Despite numerous appreciations by the Classical economists of the importance of utility or use-value in the determination of prices, it was only in the 1870s that economic theorists supplanted the labor theory of value with that of utility. Inspiration came from Jeremy Bentham (1748–1832), who believed that utility could be quantified and all actions reduced to a calculus of pleasure and pain. This was accompanied by a recognition that economic phenomena stemmed from the deliberations of the mind, and thus the locus of economic agency shifted from that of classes to individuals. Insofar as all labor created utility, the Classical distinction between productive and unproductive labor no longer obtained. Herein lay the germ of the more optimistic tenor that characterizes the

Neoclassical period. Wealth is nothing more than the maximization of utility; it is mental not material and thus knows no bounds. The Classical preoccupation with scarcity and the stationary state were thus swept aside.

The leading instigators of this so-called 'Marginal Revolution' were, in Britain, W. S. Jevons (1835–82), F. Y. Edgeworth (1845–1926) and Alfred Marshall (1842–1924) and, in Europe, Léon Walras (1834–1910) and Carl Menger (1840–1921). In the United States, the banner was carried by John Bates Clark (1847–1938) and Irving Fisher (1867–1947). The British group explicitly dropped the adjective 'political' from the name of the discipline, and sought significant ways to enhance its professional standing, through the formation of the British Economic Society and the *Economic Journal* in 1891, and by expanding and upgrading the study of economics in the university curriculum. Economics already had a much greater standing in France, due to its close links with the study of engineering and law, but there too efforts were taken by Charles Gide (1847–1932) and others to consolidate its standing as a full-fledged science. In 1878, every French university established a chair in economics and, in 1887, the *Revue d'économie politique* was founded. In Vienna, Menger attracted a number of students, giving rise to a full-fledged 'Austrian School' that survived until the Second World War. American economists garnered recognition with the founding of the American Economic Association in 1885, and the creation of three journals, the *Quarterly Journal of Economics* (1887), the *Journal of Political Economy* (1892) the *American Economic Review* (1891).

The Classical theory of Smith and David Ricardo had looked more to long-run 'natural' prices than to those that fluctuate daily in the market place. With the new emphasis on utility and demand, economists such as Jevons, Walras, and Menger maintained that every observed price was the manifestation of individual mental deliberations. There was no underlying more fundamental price to discern over the long run. The price of a good was determined at the margin, at the point where a given economic agent chose to forgo an additional unit of one specific good, and spend her money on other goods. Similar tradeoffs were construed for labor and leisure, for investments and savings, and thus the wage and profit rates were equally seen to be the effects of deliberations at the margin.

Jevons's grounding of the theory of prices in a concept of utility paved the way for the mathematization of economics. Market purchases were conceived as continuous functions of the utility that accrued to each additional infinitesimal unit of a good. Price ratios were posited to be commensurate to the last additional increments of utility that would be gained by additional purchases of two different goods. Insofar as each individual achieved this level of satisfaction, subject to budgetary constraints, their income would be perfectly distributed. Edgeworth found the means to incorporate Lagrangian techniques, thereby bringing added rigor to the concept of maximization. He also devised the indifference curve and contract curve, thus enabling a clearer representation of the links between individual deliberations and market maneuvers.

Marshall enriched the analysis of the demand curve, with rigorous definitions of elasticity and consumer surplus. He preferred to focus on the case of partial equilibrium, thus invoking numerous *ceteris paribus* conditions so as to isolate cases of comparative statics. Money income, for example, was held constant, as were the prices of all other commodities. Such simplifications, while unrealistic, were nonetheless insightful. To Marshall's credit, he struggled throughout his life with the trade-off between theory and empirical veracity, and left numerous accounts of these inquiries.

Walras went further in that he posited an economy in general equilibrium, whereby a silent auctioneer would establish those prices which would clear all markets simultaneously. By depicting the economy as a series of equations, Walras was able to demonstrate that there was an equal number of unknown variables. Hence, an algebraic solution could in principle be found. More specifically, given n productive services (land, labor, and capital) producing m finished goods, individuals with specific endowments, budget constraints, and a given set of prices, would achieve the point where supply equals demand across the entire economy. Walras made a number of simplifying assumptions, for example, that input coefficients for production were fixed, and that prices could be defined in terms of a single good, serving as the *numéraire*. However, he was unable to demonstrate rigorously that there was such a general equilibrium. Important refinements of this theory were undertaken by numerous others, notably Eugen Slutsky (1880–1948), Gustav Cassel (1866–1945), Abraham Wald (1902–50), John Hicks (1904–89), Oskar Lange (1904–65), and Paul Samuelson. In the late 1950s Kenneth Arrow and Gerard Debreu provided the capstone for the theory of general equilibrium.

The most salient feature of economic theory since the 1870s is its systematic use of mathematics, particularly the calculus. Much of the original momentum to transform literary political economy into a mathematical discipline was drawn from the wish to emulate physics. Even in the eighteenth century, Adam Smith among others found ways to inject mechanistic imagery into economic theory, the tendency of prices, for example, to gravitate to their natural price. John Stuart Mill, in his *System of Logic* (1843), developed the first extensive argument for viewing political economy as a science much like Newtonian mechanics. Both sciences were primarily axiomatic and deductive, and nonexperimental. Verification came after one had

derived the central principles from commonplace assumptions about the behavior of bodies, both human and physical. But political economy, Mill argued, could not make predictions with the kind of numerical precision of astronomy, and hence was deemed an inexact science.

What motivated the early marginalists to develop a mathematical theory was the realization that the phenomena themselves were quantitative, and that the mind deliberated in quantitative terms. They drew inspiration from new developments in logic and psychology, notably the work of George Boole (1815–64), Alexander Bain (1818–1903), and Theodore Fechner (1801–87). Strong analogies to the science of physics, notably mechanics and thermodynamics, also motivated the recasting of economics into mathematical form. Jevons, Walras, Edgeworth, Clark, Fisher, and Marshall each independently arrived at mathematical renditions of the theory of value and distribution. They found ways to apply geometry, algebra, probability theory, and the calculus. Economists have since adopted other mathematical techniques, such as topology, set theory, and have even linear programming trants to Wassily Leontie (1960–99).

In the period 1870 to 1930, many economists began to undertake the collection of large quantities of data and to estimate key economic variables or arrive at patterns such as business cycles. Efforts to measure and estimate key variables reach back to Petty, and statistical societies and investigations were commonplace by the mid-nineteenth century. But the systematic application of statistical methods and probability theory—what has come to be known as econometrics—only got underway by the Neoclassical period. Jevons, Henry Moore (1869–1958), and W. C. Mitchell (1874–1948) all attempted to discern business cycles by the regression of price data. Other important contributors to econometrics before the Second World War were Slutsky, Udny Yule (1871–1951), and Ragnar Frisch (1895–1973). Efforts to manage and plan the economy after the Great Crash of 1929 gave a significant boost to econometric research in Europe under the aegis of Jan Tinbergen (1903–94) and Trygve Haavelmo (1911–99) and in the United States in the form of the Cowles Commission. But while the Econometric Society was formed in 1930 and, in 1933, launched its own journal, *Econometrica*, the field only achieved full academic standing in the post-war period, and has obviously been much assisted with the advent of electronic computers (see *Econometrics, History of*).

3. *Applied Economics*

The Classical economists frequently engaged in public debates on such subjects as taxation, unionization, and foreign trade, but they nonetheless viewed their theoretical claims as in a separate sphere from their more practical ones. The art of political economy, the application of economic principles, was grounded in theoretical reflection, but also drew upon other bodies of knowledge, such as history, legal theory, and moral philosophy.

With the early Neoclassical economists, theory and policy became more closely linked, even though, paradoxically, the theory depicted an ideal world that was ahistorical and apolitical. The bridge principles which would link the world of Walrasian general equilibrium to the world as we know it have yet to be fully spelled out. Nevertheless, the tools of marginal utility were incorporated by Edgeworth and A. C. Pigou (1877–1959) quite readily into a branch called welfare economics, which addressed among other topics the age-old questions of distributive justice. In the first decades of the twentieth century, Vilfredo Pareto (1848–1923) devised some important tools for the ranking of alternative allocations of resources. A given allocation was deemed to be a Pareto optimum if no alternative arrangement could make at least one person better off without worsening the state of another. A Pareto improvement is a step toward such an optimum.

Social welfare functions took utility to be additive and measurable, even though the early Neoclassists had cautioned against such assumptions. There is, as yet, no measurable unit of utility, a 'util,' and economists generally have abandoned such a search, despite the implicit use of cardinal measures in some areas of analysis. Furthermore, although appeals to utility attempt to locate judgments of the 'good' in the hands of individuals, welfare economics necessarily incorporates some ethical judgments, both in determining what is socially desirable and in making policy recommendations in a world of scarcity. The general stance, however, is one of ethical neutrality. As Pigou decreed, economic welfare is synonymous to the good.

Another important field in applied economics is money and banking. Here, the school in Sweden, notably Cassel and Knut Wicksell (1851–1926) stands out. Irving Fisher was instrumental in formulating in algebraic terms what came to be known as the quantity equation, and also undertook steps to measure the average velocity of money, the rate at which a monetary instrument passes from one hand to another in the course of a year. But the person to transform most fully our understanding of money and reluctance to spend was John Maynard Keynes (1883–1946), who unpacked the underlying motives for the demand for money in terms of precautionary, transactionary, and speculatory needs. Keynes revived the seventeenth-century concern for hoarding money as a primary cause of economic slumps, whereby interest and profit rates are so low as to deter any investment. To escape the 'liquidity trap,' Keynes recommended government intervention in the form of fiscal and monetary stimuli. He placed considerable faith in fiscal policy as the

means to redistribute wealth, alleviate unemployment, and eventually liberate humankind from its passion for money. Keynes envisioned a world of developed economies run by decentralized institutions, whereby individuals would no longer be confined by economic necessity and could thereby transcend the cupidity that characterized the system of capitalism.

4. Heterodox Economics

Although the Neoclassical theory became the dominant position by the 1930s, there were many alternative streams of thought, some of which still survive to this day. In Germany, for example, such thinkers as Friedrich List (1789–1846), Gustav Schmoller (1838–1917), Werner Sombart (1863–1941), and Max Weber (1864–1920) were adamantly historical and sociological in their approach to the study of economic phenomena. They saw the Neoclassical commitment to rationality and mathematical methods as misguided. Others, such as Karl Kautsky (1854–1938) and Rosa Luxemburg (1871–1919), followed in the footsteps of Karl Marx (1818–83), who had emphasized strife between economic classes and upheld the labor theory of value. Marxist economics had its longest lineage in Eastern Europe, in the works of Vladimir Lenin (1870–1924), Nikolai Bukharin (1888–1938), Oskar Lange (1904–1965), and Michal Kalecki (1899–1970). Nikolai Kondratieff (1892–1931), while working under the Soviet system, made insights on the long waves of capitalism that were seen to be antithetical to Marxism.

In late-Victorian and Edwardian England, new schools of thought were founded that were either historical or socialist in orientation. Under the inspiration of Thorold Rogers (1823–90) and Arnold Toynbee (1852–83), economic history secured a firm toehold in the university curriculum, where subsequently it has flourished in the hands of William Ashley (1860–1927), G. D. H. Cole (1889–1959) and Phyllis Deane. The Fabians—George Bernard Shaw (1856–1950), Sidney Webb (1859–1947), and Beatrice Webb (1858–1943)—were instrumental in founding the London School of Economics (1895), which soon abandoned its socialist roots. Built up by such eminent economists as Edwin Cannan (1861–1935), Lionel Robbins (1898–1984), and Friedrich Hayek (1899–1992), the LSE became a leading institution for economic research and education. Sir John Hicks was the dominant figure at Oxford in the mid-twentieth century. At Cambridge University, two students of Keynes's, Piero Sraffa (1898–1983) and Joan Robinson (1903–83), cultivated neo-Ricardian theories which drew attention to capital rigidities and imperfect competition. The contributions of Dennis Robertson (1890–1963), Richard Kahn (1905–89), and Nicholas Kaldor (1908–86), however, insured that Cambridge to this day is primarily aligned with Neoclassical thinking.

Yet another group drew inspiration from Carl Menger (1840–1921), whose *Principles of Economics* appeared in 1871, the same year as Jevons's most important economic treatise. In one respect, these works bore a strong similarity in that they both treated prices in terms of the utility theory of value. But Menger was steeped in the tradition of Immanuel Kant (1724–1804) and G. F. Hegel (1770–1831), and was most sharply opposed to the German historical school of the mid-nineteenth century. He did not appreciate utilitarianism, nor the turn toward mathematical techniques. Together with his pupils, Friedrich von Weiser (1851–1926) and Eugen von Böhm-Bawerk (1851–1914), the Austrian school made strong commitments to methodological individualism, to the heroic efforts of the entrepreneur, and to the efficacy of market mechanisms. Introspection and *a priori* reasoning were the main methods of inquiry, insofar as economic phenomena were the products of a logic of choice. The Austrians also drew attention to the role of information, the scarcity of time, and the importance of defining economic needs and wants.

Both Joseph A. Schumpeter (1883–1950) and Hayek harked from Vienna, but moved considerably past the teachings of Menger. Both also emigrated to America before the Second World War and, in Hayek's case, settled in England, so that their respective doctrines came to evince an interesting blend of Austrian rationalism and Anglo-American empiricism. Schumpeter's *Capitalism, Socialism and Democracy* (1942) and Hayek's *Road to Serfdom* (1944) are two of the most popular nontechnical works of the twentieth century. Like Keynes, they attempted to predict the fate of capitalism as an economic system and speculate on the political measures required to sustain the status quo.

In America, under the inspiration of evolutionary biology and Social Darwinism, a school that came to be known as Institutionalism took shape with the work of, most notably, Thorstein Veblen (1857–1929), John R. Commons (1862–1945), and Wesley Clair Mitchell (1874–1948). They viewed economic phenomena in evolutionary terms; for example, human rationality was but one step from animal instinct. Human institutions were replete with contradiction; human striving was in many respects futile. Economic phenomena were treated in broader sociological terms, as part of the government and legal system. Although the Institutionalists failed to generate a large following, they were a viable force in the economics profession right through the 1940s and, more recently, have made a second coming in the form of New Institutionalism, as expounded by Douglass North most notably. Here, the salient concepts are transaction costs and institutions, which are defined as rule-setting bodies that serve to minimize such costs. From a different standpoint, Ronald Coase's theorem shows

that there are no externalities given zero transaction costs and well-defined property rights.

5. *Postwar Developments*

Since World War Two, Neoclassical economics, particularly in the United States, has become a formidable edifice. American economists garner the majority of the Nobel prizes in economics, and many non-American economists find it essential to spend some period of their career studying or teaching in the United States. Paul Samuelson's *Foundations of Economic Analysis* (1947) provided the first mathematical exposition of price dynamics; Milton Friedman restored the monetarist doctrines of the pre-Keynesians and, more recently, John Muth, Thomas Sargent, and Robert Lucas have inspired a train of thinking known as the New Classicism. The efforts of Frank Knight (1885–1972) and Jacob Viner (1892–1970) made the University of Chicago the most prominent center for the study and dissemination of economic theory. More Nobel Laureates in economics have had an affiliation with that university than with any other.

Undoubtedly the most original and influential subfield to develop since the war has been game theory, which had antecedents in previous work, such as that of Blaise Pascal (1623–1662) and Augustin Cournot (1801–77), but which truly came into its own with the publication of John von Neumann (1903–57) and Oscar Morgenstern's (1902–77) *Theory of Games and Economic Behavior* (1944). Game theory has underscored the importance of uncertainty, of suboptimal outcomes, and of strategizing. It has also drawn attention to the significance of information as an economic resource. Similar kinds of analysis can be found in decision theory and rational choice theory, such as Arrow's impossibility theorem.

Economists tend to specialize in one branch, some of which have longstanding roots, for example, labor economics or public finance. Other significant postwar developments are cliometrics (Robert Fogel), experimental economics (Vernon Smith), and household economics (Gary Becker). As with all of the aforementioned, the predominate aim is model building driven by mathematical tractability. Economists rarely seek laws of a general nature. In contrast to the economists of the Classical era, which might even be characterized as the glorious age of universal principles—the law of diminishing returns, of the falling rate of profit, and Say's law, Neoclassical economists prescribe more limited generalizations, such as the Phillips curve or the Heckscher–Ohlin conditions for trade. The search for general laws seems like a task long forgotten. Nevertheless, the Neoclassical theory is assumed to be universal in scope, to apply not only to fully developed nations but to all human societies, past and present. The model of rationality that lies at its conceptual foundations is purported to capture human agency at its most fundamental.

Notwithstanding all of the mathematical refinement that has transpired in the past century or so, there remains a vociferous minority who point to severe limitations of the Neoclassical paradigm and call for more public ownership or social controls. Herbert Simon and, more recently, Amartya Sen have called into question the model of human rationality to which most Neoclassicists subscribe. Other prominent critics since the 1950s are Joan Robinson, Gunnar Myrdal (1898–1987), Kenneth Boulding (1910–93), Robert Heilbroner, and John Kenneth Galbraith. They desist from singing the praises of capitalism, but do not view socialism as a viable alternative. Criticism is done with full knowledge and understanding of the mainstream theory, and for that reason these economists, and several more, have won large audiences for their writings. The Neoclassical theory has been remarkably robust in the face of such criticism, but its longevity is by no means guaranteed.

See also: Econometrics, History of; Game Theory; Hume, David (1711–76); Keynes, John Maynard (1883–1946); Locke, John (1632–1704); Mathematical Models in Philosophy of Science; Mill, John Stuart (1806–73); Political Economy, History of; Smith, Adam (1723–90)

Bibliography

Black R D C, Coats A W, Goodwin C D W (eds.) 1972 *The Marginal Revolution in Economics: Interpretation and Evaluation*. Duke University Press, Durham, NC

Caldwell B J (ed.) 1990 *Carl Menger and His Legacy in Economics*. Duke University Press, Durham, NC

Hausman D 1992 *The Inexact and Separate Science of Economics*. Cambridge University Press, Cambridge, UK

Hollander S 1987 *Classical Economics*. Blackwell, Oxford, UK

Ingrao B, Israel G 1987 *La Mano Invisible*. Guis. Laterza & Figli Spa, Roma-Bari 1990 *The Invisible Hand*. MIT Press, Cambridge, MA

Jaffé W 1983 *William Jaffé's Essays on Walras*. Cambridge University Press, Cambridge, UK

Maloney J 1985 *Marshall, Orthodoxy, and the Professionalisation of Economics*. Cambridge University Press, Cambridge, UK

Mirowski P 1989 *More Heat than Light: Economics as Social Physics, Physics as Nature's Economics*. Cambridge University Press, Cambridge, UK

Morgan M S 1990 *The History of Econometric Ideas*. Cambridge University Press, Cambridge, UK

Morgan M S, Rutherford M (eds.) 1998 *From Interwar Pluralism to Postwar Neoclassicism*. Duke University Press, Durham, NC

Schabas M 1990 *A World Ruled by Number: William Stanley Jevons and the Rise of Mathematical Economics*. Princeton University Press, Princeton, NJ

Spiegel H 1991 *The Growth of Economic Thought*, 3rd edn. Duke University Press, Durham, NC

Weintraub R 1985 *General Equilibrium Analysis*. Cambridge University Press, Cambridge, UK

Whitaker J K 1990 *Centenary Essays on Alfred Marshall.* Cambridge University Press, Cambridge, UK

M. Schabas

Economics: Overview

Economics is often defined as the study of the production, distribution, and utilization of scarce resources. Although this description is a useful characterization of much economic analysis it is in some ways unsatisfactory. To many economists this definition confines the scope of economic analysis so that some subjects studied by economists, such as the determinants of occupational safety (see *Safety, Economics of*) or the determinants of health care (see *Health Economics*), may not be included. The history of economic analysis (see *Economics, History of* and *Political Economy, History of*) is a fascinating demonstration of how the topics that economists take up for study have changed systematically over the last two centuries.

More generally, economics is the study of social relations that result from the wants that people have. This broad definition no doubt encompasses most, if not all, of economics. However, even this broad definition of economics omits some important activities that economists undertake as part of their profession. Economists have been major contributors, for example, to the development of specialized software useful to economists and other social and behavioral scientists (see *Econometric Software*) and to the development of mathematical tools of broad applicability in the social and behavioral sciences (see *Game Theory: Noncooperative Games*).

Most broadly, then, some would simply define economics as the studies that economists do.

1. *The Subject and Method of Economics*

Economics is divided either by subject matter or by the primary method used for study. The subject matter of economics is often divided into the areas of microeconomics, the study of the behavior of individual agents, and macroeconomics, the study of the behavior of aggregate units.

The vast majority of economists who work in academia, business, or government are working on problems that would be called microeconomics. However, the discussion of the aggregate economy, and public policies related to the aggregate economy, often pique the public interest for prolonged periods far more than do the issues of microeconomics. Aggregate unemployment and its movement is always of interest (see *Unemployment: Structural* or *Monetary Policy*, for example), but Antitrust Policy or health care costs (see *Health Economics*) are only on the public agenda occasionally. To show the magnitudes involved, we may compare the number of university graduates of business schools (who almost certainly have studied microeconomics) with the number of graduates from economics departments (who almost certainly have studied macroeconomics). In the USA in the academic year 1995–1996, for example, there were 13 times as many business school graduates (227,000) as there were economics graduates (17,000).

Macroeconomic analyses often focus on key indicators like the unemployment rate, gross national product, or price inflation (see *Macroeconomic Data*). It is increasingly common for these analyses to be based on the aggregation of microeconomic analyses, although this is not always the case.

Microeconomic analyses often focus on the details of a particular market or organizational structure. The breadth of subjects taken up using the tools of microeconomics is truly breathtaking, ranging from art, culture, and sports (see *Art and Culture, Economics of* and *Sports, Economics of*) to housing and agriculture (see *Housing Economics* and *Agriculture, Economics of*). As such, these analyses often abstract from other analyses that may be relevant in principle, but are thought to be irrelevant in practice.

Economics is often also distinguished by the methods that are used. Very generally, economic studies are often said to be either theoretical or empirical, and sometimes both. Although the techniques of analysis may vary dramatically among theoretical studies, ranging from the purely verbal to the highly mathematical, these studies have in common the attempt to provide a general framework of analysis that is independent of the direct observation of economic behavior.

Empirical studies in economics are less concerned about generality and more concerned about the direct observation of economic behavior. In principle, it is a commonly held goal of many economists to provide both a theoretical framework and the observational analyses that serve to test the implications of the assumptions on which the theory depends. However, this is not always possible or desirable.

The editors of the *Journal of Economic Literature* have provided a useful and exhaustive set of categories for the literature of economics. The categorization was done after an exhaustive set of studies of the nature of the vast literature of economics and with the assistance of a number of distinguished economists. Broadly speaking, there are 18 broad categories and a residual category that contains a group of special topics. The 18 categories are General Economics and Teaching; Methodology and History of Economic Thought; Mathematical and Quantitative Methods; Microeconomics; Macroeconomics and Monetary Economics; International Economics; Financial Economics;

Public Economics; Health, Education, and Welfare; Labor and Demographic Economics; Law and Economics; Industrial Organization; Business Administration, Business Economics, Marketing, and Accounting; Economic History; Economic Development, Technological Change, and Growth; Economic Systems; Agricultural and Natural Resource Economics; and Urban, Rural, and Regional Economics. As this list demonstrates, the categories of economics are not really mutually exclusive and it is often useful to categorize a particular economic study under more than one heading.

It should be apparent from this list of categories that modern economic analysis covers a very broad set of topics. Indeed, the list of topics included changes continuously.

2. *The Profession of Economics*

One of the main changes in economics since around 1900 is the increasing professionalization of the subject. At the beginning of the twenty-first century many people earn their living directly or indirectly from economic research. This may be seen in the phenomenal growth in the volume of published economic scholarship. The first economics journal, the *Zeitschrift für die gesamte Staatswissenschaft* (Tübingen), was founded in 1844. It was 22 years before another journal, the *Quarterly Journal of Economics* (Cambridge, MA), was created. By 1900 there were still only seven economics journals. By 1962 the number of economics journals had increased to 76, a growth of about one per year, but only 24 of these were published in the USA.

In 2001 there are approximately 610 economics journals, of which 300 were published in the USA. The growth in the number of journals in the last half of the twentieth century was at a rate 10 times greater than was the case in the first half of the twentieth century. These journals today publish about 2,600 issues and contain over 350,000 printed pages each year.

The growth and professionalization of economics has led to a change in the way that economists organize themselves. The oldest professional organization of economists, the American Economic Association, was founded in 1885 with fewer than 200 members. By 1962 the membership had grown to 11,000, most of the growth occurring between 1940 and 1962. By 2001 membership had reached 22,000, but it had changed very little during the previous two decades.

Other general-purpose organizations of economists have not fared so well. The Royal Economic Society was formed in 1890 and by 1959 had 5,000 members. In 2001, however, the Royal Economic Society only has 3,300 members. At the same time, however, there has been considerable growth in the number of specialist organizations of economists. Worldwide there are now 257 economics associations, of which 77 are headquartered in the USA. Many of these organizations have small memberships, but they often remain viable because they serve the strongly focused interests of a special group.

Bibliography

Economic Journal. Blackwell, Oxford, UK and Macmillan, New York, for the Royal Economic Society, Vols. 1–110, 1891–2001

Journal of Economic Literature. American Economic Association, Evanston, IL, Vols. 7–35, 1969–1997

Quarterly Journal of Economics. MIT Press for Harvard University, Cambridge, MA, Vols. 1–116, 1886–2001

Zeitschrift für die gesammte Staatswissenschaft. Verlag der H. Laupp'schen Buchhandlung, Tübingen, Germany, Vols. 1–141, 1844–1985

O. Ashenfelter

Economics, Philosophy of

Like many of the social sciences, economics grew out of philosophy, and the concerns of economists continue to intersect with those of philosophers. Philosophical reflections on (a) scientific method and social ontology; (b) the nature of rationality, self-interest, and preference; and (c) welfare, justice, equality, and freedom are of abiding significance to economists and other social scientists.

1. *Economic Methodology*

Philosophical reflection on economics is ancient, but the conception of 'the economy' as a distinct object of study dates back only to the eighteenth century. Aristotle addresses some problems of economics mainly as problems of household management. Scholastic philosophers addressed ethical questions concerning economic behavior, and they condemned 'usury'—that is, the taking of interest on money. With the increasing importance of trade and nation-states in the early modern period, 'mercantilist' philosophers and pamphleteers addressed questions concerning the balance of trade and the regulation of the currency. Only in the work of the physiocrats and especially of Adam Smith do scholars begin to think of the economy as an object of study with its own principles and laws.

Eighteenth century philosophers wrote in the shadow of Newton's towering accomplishments. David Hume is unabashed about his hopes to develop

a science of mind and society in the image of Newton's science of the solar system. To that end, he seeks out general laws of individual thought and action, out of which larger-scale orderly relations will arise, in just the way that patterns in planetary motion arise from the laws of motion and gravitation governing individual bodies. Thus Hume traces the rise in prices and the temporary increase in economic activity that follow an increase in currency to the perceptions and actions of individuals who first spend the additional currency.

It then lay to Adam Smith to extend such a method to a systematic *Inquiry into the Nature and Causes of the Wealth of Nations* and to make explicit an implication of previous causal inquiries such as those of Hume, the physiocrats, and of many of the mercantilists. That implication (which is arguably the foundation for all social science) is that the social, aggregative implications of individual choices are often *unintended*. In spending their additional gold imported from abroad, traders do not intend to increase the price level. But that is what they do nevertheless. In this way unintended regularities that arise from the intentional choices of individuals can rule those individuals just as surely as the regularities of nature, and, like nature, societies can be objects of scientific investigation.

1.1 Classical Economics and the Method a Priori

Although Adam Smith made some general comments about scientific method in his *History of Astronomy* and other essays, he wrote little explicitly about the method of economics. The first extended reflections on economic methodology had to wait until Nassau Senior (1836) and John Stuart Mill (1836). Their essays must be understood against the background of the prevailing economic theory. Like Smith's economics (to which it owed a great deal) and modern economics, the 'classical' economics of the middle decades of the nineteenth century traced economic regularities to the choices of individuals facing social and natural constraints. But, as compared to Smith, more reliance was placed on severely simplified models. In David Ricardo's *Principles of Political Economy*, a portrait is drawn in which wages above the subsistence level lead to an expanding labor force, which in turn requires more intensive agriculture or cultivation of inferior land. The extension of cultivation leads to lower profits and higher rents; and the whole tale of economic development leads to a gloomy stationary state in which profits are too low to command any net investment, wages return to subsistence levels, and only the landlords are affluent. From the time Ricardo's *Principles* was published (1819), the data available to classical economists was never in accordance with the trends the theory predicted. Yet the theory continued to hold sway for more than half a century, and the unfavorable data were explained away as due to various 'disturbing causes.'

It is not surprising then that Mill's account of the method of economics would emphasize the relative autonomy of theory. Mill distinguishes between two main kinds of inductive methods. The method *a posteriori* is a method of *direct* experience. It is only suitable for phenomena in which few causal factors are operating or in which experimental controls are possible. Mill's famous methods of induction are detailed specifications of the method *a posteriori*. In his method of difference, for example, one holds fixed every causal factor except one and checks to see whether the effect ceases to obtain when that one factor is removed.

Unfortunately, such direct inductive methods cannot be used to study phenomena in which many causal factors are in play. If, for example, one attempts to investigate whether tariffs enhance or impede prosperity by comparing the prosperity with and without tariffs, the results will be irregular and unreliable because other causes besides tariffs will also differ across societies and times. So one needs instead to employ the method *a priori*. Despite its name, this is an *inductive* method, but it is an indirect inductive method. One first determines the laws governing individual causal factors in domains in which direct inductive methods are applicable. Having then determined the laws of the individual causes, one investigates their combined consequences deductively. Finally, there is a role for 'verification' of the combined consequences, but owing to the causal complications, this testing has comparatively little weight. The testing of the conclusions serves only as a check on one's deductions and as an indicator of whether there are significant disturbing causes that one has not yet accounted for. Mill gives the example of the science of the tides. One determines the law of gravitation by studying planetary motion, in which gravity is the only significant causal factor. Then one develops the theory of tides deductively from that law and information concerning the positions and motions of the moon and sun. The implications of the theory will be inexact and sometimes badly mistaken, because many subsidiary causal factors influence tides. By testing the theory one can uncover mistakes in one's deductions and evidence concerning the role of the subsidiary factors. Because of the causal complexity, such testing does little or nothing to confirm or disconfirm the law of gravitation, which has already been established.

Because economic theory includes only the most important causes and necessarily ignores many minor causes, its claims, like claims concerning tides, are inexact. Its predictions will be imprecise, and sometimes dead wrong. But it is possible nevertheless to develop and confirm economic theory by first establishing in simpler domains the laws governing the major causal factors and then deducing their con-

sequences in different circumstances. For example, statistical data tell a mixed story about the relationship between minimum wages and unemployment; and there is no data at all about what the consequences for employment would be of an extremely high minimum wage. On the other hand, everyday experience teaches one that firms can choose among more or less labor-intensive processes and that a high minimum wage will make more labor-intensive processes more expensive. Since one also has good reason to believe that firms try to keep their costs down, one has good reason to believe that a high minimum wage will increase unemployment.

According to Mill, economics is not only inexact and committed to the method *a priori*. In addition, he maintains that it is a *separate science*. What distinguishes economics as a discipline is not only its concern for a certain domain of phenomena, but also its limitation to a particular set of causal factors, which predominate in this domain. With respect to this domain, one can (at a certain level of approximation) ignore the myriad causal factors that influence all social phenomena and which are the subject matter for sociology and develop economics separately. In defending a view of economics as in this way inexact and separate and a view of economists as following the method *a priori*, Mill was able to reconcile his empiricism and his commitment to Ricardo's economics.

Although Mill's views on economic methodology were challenged later in the nineteenth century by dissident economists who believed that the theory was too remote from the contingencies of policy and history, Mill's methodological views dominated the mainstream of economic theory for well over a century. Mill's vision survived the transformation of economics from classical to neoclassical and is clearly discernable in the most important methodological treatises concerning neoclassical economics, such as John Neville Keynes' *The Scope and Method of Political Economy* (1891) or Lionel Robbins' *An Essay on the Nature and Significance of Economic Science* (1935). Indeed, Hausman (1992) argues that current methodological practice closely resembles Mill's methodology.

1.2 Twentieth Century Philosophy of Science and Economic Method

Beginning in the 1930s, mainstream economists began to have a bad conscience about their traditional methodology, which some saw as insufficiently empiricist. Robbins is a transitional figure, because at the same time that he clung to Mill's methodology (1935, Chap. 4) and to the identification of the domain of economics with the predominance of particular causal factors (the allocation of scarce means that have alternative uses—1935, Chap. 1), he also offered his apparently empiricist critique of interpersonal utility comparisons as untestable value judgments (1935, Chap. 6). Terence Hutchison argued that the propositions of pure theory were so hedged with *ceteris paribus* conditions that they were not testable (1938). Paul Samuelson argued for a need to separate the 'operationally significant' economic wheat from the chaff, and in his theory of revealed preference he provided a model for how to do so (1947). Other economists cited survey data to argue that the theoretical propositions of economics were false. The confusing methodological situation stabilized in the 1950s with arguments by Fritz Machlup (1955) and especially Milton Friedman (1953) that economists need be concerned to fit only data concerning prices and quantities and that the 'realism of assumptions' is irrelevant.

Although confused, mistaken, and inconsistent with the practice of mainstream economists—including his own practice—Friedman's views have dominated the methodological conceptions of mainstream economists over the past two generations. His views are confused because they conflate many different kinds of 'assumptions' and 'realism.' Basic generalizations, propositions concerning initial conditions, and antecedents of conditional claims are all called 'assumptions.' Assumptions are called 'unrealistic' if they are false, incomplete, or not approximately true. Friedman's views are mistaken, because (as Mill already emphasized) market data are generated by too many causal factors to provide effective tests of economic theory. The eclecticism that Friedman's official methodological position counsels—'don't worry about what theory says, just ask whether it fits market data'—contradicts the firm and narrow commitment to a specific theory that characterizes mainstream economics, including Friedman's own work. Nevertheless, Friedman's views were warmly embraced because they freed mainstream economists to ignore criticisms of their theories and to abandon all empirical work apart from econometrics.

In the 1970s and 1980s other streams of contemporary philosophy of science began to influence the methodological reflections of economists. Karl Popper's view that scientific theories must be falsifiable, which had been defended by Hutchison, was taken up and emphasized by Mark Blaug (1992). According to Popper, the hallmark of science is to formulate theories so that they can be exposed to empirical testing and to reject theories that fail tests. Hutchison and Blaug object that mainstream economic theories are not harshly tested and that mainstream economists are unwilling to surrender them when they fail the few tests they are subjected to. In his extensive publications, Lawrence Boland (1982) has also emphasized the relevance of Popper's views to economic methodology, although Boland's interpretation of Popper is rather different than Blaug's or Hutchison's.

Imre Lakatos' methodology of scientific research programs (which represented a wedding of some of

Thomas Kuhn's insights to Popper's philosophy of science) was for a while widely discussed (Latsis 1976, de Marchi and Blaug 1991). Although still emphasizing the importance of empirical criticism, Lakatos insisted that theories should not be abandoned until superior alternatives are found, and his emphasis on heuristics struck a responsive chord.

A spate of prominent contemporary writers on methodology: Roger Backhouse, Bruce Caldwell, Neil de Marchi, D. Wade Hands, and E. Roy Weintraub have a more ambivalent relationship to Popper's and Lakatos' philosophy. Caldwell wrote a series of searching though charitable criticisms of Popperian views. Backhouse, de Marchi, Hands, and Weintraub were all at one time enthusiastic about Lakatos' views, though to differing degrees all have moved away from them now. Backhouse and de Marchi remain the closest. Hands has become a proponent of the application to economics of new work on the sociology of science, while Weintraub's work on the stabilization of concepts in economics has been more influenced by literary theorists

More recently, alternatives to philosophy of science influenced by work in literary theory and sociology have also developed a following among economists. In a series of stylistically brilliant works, Deirdre McCloskey has criticized the whole project of exploring a normative methodology for economics and has urged instead that economists attend to their rhetoric—that is, to their ways of persuading one another (1985). This work has been highly controversial because many of McCloskey's formulations apparently reduce questions about the correctness or incorrectness of economic claims to questions about what most economists accept. Such a view would imply that minority views are always mistaken. McCloskey denies such a radical reading of her position, but it is not clear how to avoid it without incorporating normative methodological commitments into the rhetoric of economics.

Philip Mirowski's work on methodology is heavily historical and less radical epistemologically than McCloskey's. He has explored the influence on economics of the formal analogy between utility theory and physics (1989). More recently, his work has shown increasing concern with sociological influences. For other work that tests the boundaries of conventional methodology, see Mäki et al. (1993).

Over the last 15 years, economic methodology has become a large field. It has its own *Journal of Economic Methodology* and occupies a large part of the journal, *Economics and Philosophy*. Dozens of monographs on economic methodology have been published. Graduate programs in the field have been established (including a Ph.D. program at Erasmus University in Rotterdam), and there are regular sessions on economic methodology at meetings of economists and philosophers. The field is very diverse, and each of its leading figures has a distinctive approach. Although one can roughly identify Popperian–Lakatosian and sociological–literary schools, there is no easy way to categorize contemporary methodology.

Several methodologists have been particularly interested in the role of explicitly causal notions in economics (see *Causes and Laws: Philosophical Aspects*; *Causation: Physical, Mental, and Social*). Nancy Cartwright (1989), who is also a distinguished philosopher of physics, has (unlike most philosophers writing on economics) paid a great deal of attention to econometrics. She has defended the importance of specifically causal considerations in science, and has argued that econometricians have made important contributions to a philosophical understanding of causation. Cartwright holds that economists should be understood as attempting to identify causal *capacities* and that the work of econometricians can be understood as contributing to this task.

Daniel Hausman (1992) defends a modernized variant of Mill's methodology that depicts economics as inexact and permits the plausibility of the basic principles of economics to count in the assessment of economic theories. But he is skeptical of the view that a small set of causal factors do indeed predominate in the domain of economics and defends a greater attention to the results of experiments, surveys, and field work. He has increasingly emphasized the role of causal generalizations in mainstream microeconomic theory.

Kevin Hoover (2000)—unlike most methodologists—has written on problems concerning macroeconomics. Although he has addressed technical issues in econometrics, his main interest has been in causality. Like Cartwright, Hoover has placed particular emphasis on what one can learn about causation by studying the work of economists.

Tony Lawson (1997) finds his inspiration in the 'transcendental realism' of Roy Bhaskar. Lawson sees himself not merely as enunciating a distinctive approach to methodology, but as laying the foundations for a new economics. Crucial to the approach and to his methodological views is a strongly realist ontology (see *Realisms and their Opponents: Philosophical Aspects*) which takes the objects of scientific investigation to be causal mechanisms and tendencies that lie as it were 'beneath' the irregularity of phenomena. This realist ontology leads him, like Cartwright, Mäki, and increasingly Hausman, to regard economic theories as identifying causal mechanisms and tendencies.

Uskali Mäki is harder to categorize. Like Lawson, he is concerned with realism, but whereas Lawson attempts to redirect economics and its methodology on the basis of a specific realist ontology, Mäki elucidates versions of realism to which economists have been implicitly committed. Like Cartwright, Hausman, and Hoover, Mäki emphasizes the importance of causal notions in economics. Like Hands and Mirowski, Mäki also applies insights from the

sociology of science to the understanding of economics.

There is a great deal of other work, too. A number of economists and philosophers have attempted to apply to economics the structuralist view of scientific theories developed by Patrick Suppes, Joseph Sneed, and Wolfgang Stegmüller. (See for example Hands 1985 and Balzer and Hamminga 1989.) Alexander Rosenberg, who is also a distinguished philosopher of biology, has pursued a number of different themes. His *Microeconomic Laws: A Philosophical Analysis* (1976) was one of the first modern philosophical treatments of the methodological problems of economics. In that book Rosenberg argues that economics fits reasonably well within a standard philosophical model of the natural sciences. But shortly afterwards Rosenberg changed his views radically, and he now defends the position that mainstream economics is best understood either as applied mathematics or as a normative social and political theory (1992).

Contemporary economic methodology thus moves in many directions. Although some of it is directed toward the conduct of economics, one regularly finds a sort of methodological schizophrenia whereby in theory economists cling to positivist or Popperian philosophical views that are radically at odds with their practice, which is roughly Millian. Other work in methodology is oriented toward the philosophy of science, and it appears that the study of economic methodology has made significant contributions to philosophical investigations of causation and explanation. It is controversial whether methodology can or should be a distinct and relatively self-contained field rather than addressing itself toward economists or toward philosophers.

2. *Methodology, Rationality, Preference, and Self-interest*

The study of rationality, preference, and self-interest is a second area of overlap between economics and philosophy. Mainstream economic theory is built around a variant of 'folk psychology.' According to folk psychology, human actions are the consequence of beliefs, desires, and circumstances that determine the feasibility and consequences of actions. The same elements figure in economic theories of choice. Sometimes economists suppose that agents have complete knowledge and thereby avoid having to refer specifically to beliefs, but in other contexts beliefs are explicitly modeled—most typically as subjective probability judgments. In place of desires, economists postulate that economic agents possess stable, complete, and transitive preferences. Given additional technical conditions, such preferences can be represented by a continuous 'utility function' such that $U_P(A) > U_P(B)$ if and only if an agent P prefers A to B. 'Utility' is simply a way of indicating how an alternative is ranked by an agent. It is not a substantive object of choice, and indeed it makes no sense to regard an agent as seeking or preferring utility. Mainstream economic models relate choice to preference by maintaining that, subject to the constraints, agents maximize utility—which means merely that among the feasible alternatives agents choose whatever they most prefer.

The previous paragraph presents the standard theory of choice as an empirical theory purporting to describe, predict, and explain how agents choose in fact. But the same theory also functions as a theory of rationality. Preferences are rational if they are complete and transitive (and perhaps satisfy other axioms, too). Degrees of belief are rational if they satisfy the axioms of the calculus of probability. Choices are rational if they are utility maximizing. Given this model of rationality, the empirical claim of the standard theory of rational choice can be restated simply as the claim that economic agents are in fact rational.

According to this theory of rationality, it is irrational to choose one feasible option over another when one prefers the second and according to this theory of choice people will in fact never choose one feasible option over another if they prefer the second. Since according to these theories choice is and ought (rationally) to be determined by the agent's preferences, it is easy mistakenly to infer that these theories maintain that agents are and ought to be self-interested. But whether a choice is self-interested depends on the *content* of the preferences that lead to it, not on whether the choice is determined by preferences. People who prefer to sacrifice their interests to others are not self-interested, and the fact that their choices result from these preferences does not make their choices self-interested. Unlike the standard theories of choice and rationality, the views that it is rational to pursue one's own interest and that people in fact do so are 'substantive' theories of choice and rationality. Unlike the standard 'formal' theory, which merely specifies a certain *structure* of choice and preference, self-interest theories specify an objective that agents do or ought rationally to have.

Theories of rationality are normative theories. They *prescribe* what one ought to do—not, to be sure, as a matter of morality, but as a matter of rationality or prudence. (Irrationality is foolish rather than evil.) Incorporating as it does a normative theory, economics is unlike any of the natural sciences. The reason why economics incorporates a theory of rationality is that human actions, unlike the actions of oak trees or potassium, can be criticized or justified as well as explained. This fact has important methodological consequences and indeed (as just argued) establishes one important distinction between the social and the natural sciences.

The first and perhaps most important methodological consequence is that explanations of individual

choices cite the *reasons* why the agent acted. Although a number of philosophers in the 1950s, who were influenced by Wittgenstein, argued that such explanations that cite an agent's reasons could not be causal explanations, nowadays most philosophers are persuaded by Donald Davidson's argument (1963) that satisfactory explanations which cite an agent's reasons must also be causal explanations. Agents may have reasons for performing an agent that are not in fact responsible for the action. Davidson argued that the distinction between the reasons that are 'effective' and those that are merely rationalizations is that the first unlike the second are causes of the action.

So the fact that the explanations economists offer of individual choices cite reasons for those choices does not imply that economists are not giving causal explanations, and it does not imply a strong anti-naturalist distinction between the structure of the natural and social sciences. But the fact that the causal explanations economists offer also cite reasons is nevertheless of great interest. It explains why economists must hold that agents are to some extent rational. (If they weren't then they couldn't have reasons for their actions.) Furthermore, it implies that the actions of economic agents are subject to rational *appraisal*. Were the agent's reasons *good* reasons? Was the action *justified*? Once one begins appraising choices, one is just one step away from ethical questions. Readers should be aware that the standard theories of rationality and choice are controversial. For more detail on the theories and these controversies, see *Intentionality and Rationality: An Analytic Perspective* and *Rational Choice Explanation: Philosophical Aspects*.

3. *Welfare, Justice, Equality, and Freedom*

Economists have an ambivalent attitude toward ethics. On the one hand, many have been concerned to insist that mainstream economics is a 'positive' science, whose conclusions are entirely independent of any moral commitments. On the other hand, economists freely give out normative policy advice. Some of this advice is purely technical, like the advice that a civil engineer might offer on where to locate a bridge, but much of it is not. Most mainstream economists are in fact committed to a specific view of ethics that emphasizes welfare, and they also adopt a distinctive theory of welfare. (When economists emphasize *efficiency* they are virtually always concerned with efficiency at promoting individual welfare.) These features of normative economics are not entailed by positive economics or the theory of rationality, but they are heavily influenced by them.

There are a variety of different ethical bases for assessing social arrangements. In addition to considering the welfare of the members, one can also ask whether the rights of the members are protected and whether procedures and the distribution are just, whether the members are treated equally, and what sorts of freedoms and opportunities people enjoy (see *Rights*; *Justice: Philosophical Aspects*; *Equality: Philosophical Aspects*; *Freedom: Political*). All of these considerations are important, but in economic policy pronouncements and normative theory (which indeed is called 'welfare economics'), typically only considerations of welfare enter.

The reason for this narrow normative focus is that welfare can be tied tightly to the standard models of rationality and choice. Suppose that individuals are rational and—in addition—that they are self-interested. As already pointed out, the standard theory of rationality does not imply self-interest, but self-interest is also commonly assumed in most economic models. It follows that individuals will prefer alternative A to B if and only if they believe that A serves their interests better than B. If individuals have perfect knowledge, which is again a common assumption in economic models, it follows that individuals will prefer A to B if and only if A in fact serves their interests better. Finally, if one identifies an individual's welfare with whatever is in the individual's interest, it follows that A is better for and individual than B if and only if the individual prefers A to B—or in other words that welfare is the satisfaction of preference. This view of welfare has the additional advantage that it prevents concerns about paternalism—to which many economists are opposed—from even arising, since by definition it could never be good for individuals to overrule their preferences.

Most normative economics is thus doubly narrow. Not only is it characterized by an almost exclusive preoccupation with welfare, but it is also committed to a particular view of welfare as the satisfaction of preferences. This double narrowness is doubly unfortunate. The limitation to considerations of welfare means that in policy discussions economists treat nonwelfarist ethical criteria as exogenous constraints to be left for someone else to worry about. The commitment to a preference satisfaction view of welfare has the drawback that this view of welfare is false. Because individuals are not always self-interested and because their beliefs are not always true, individuals do not always prefer what is good for them. There are a variety of good reasons to be hesitant about paternalism, but the view that individuals are always perfect judges of their own good is not one of them. There is some work, particularly by Amartya Sen (1992), that stretches these limits on normative economics, but such work is the exception to the welfarist rule (see also Hausman and McPherson 1996; *Economics and Ethics Welfare: Philosophical Aspects*).

4. *Conclusions*

Mainstream economists, whether engaged with positive theory, the theory of rationality, or with normative and policy investigations, are overwhelmingly com-

mitted to an exaggerated simplification of a plausible view of individual agents as possessing stable and consistent preference rankings, which, given the constraints, determine their choices. Market outcomes are the unintended consequences of these choices. Rationality is defined by this structure of choice. Welfare, like choice, is determined by these same preference rankings. Although there are philosophical issues concerning economics that have no particular connection to this core commitment, most work in the philosophy of economics has attempted to comprehend and evaluate the basic model of choice and its applications in positive economics, normative economics, and the theory of rationality.

See also: Methodological Individualism: Philosophical Aspects; Prisoner's Dilemma, One-shot and Iterated; Social Science, the Idea of

Bibliography

Backhouse R E 1997 *Truth and Progress in Economic Knowledge*. Edward Elgar, Lyme, NH

Balzer W, Hamminga B (eds.) 1989 *Philosophy of Economics*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Blaug M 1992 *The Methodology of Economics*, 2nd. edn. Cambridge University Press, Cambridge, UK

Boland L A 1982 *The Foundations of Economic Method*. Allen and Unwin, London

Caldwell B J 1982 *Beyond Positivism: Economic Methodology in the Twentieth Century*. Allen and Unwin, London

Cartwright N 1989 *Nature's Capacities and their Measurement*. Clarendon Press, Oxford, UK

Davidson D 1963 Actions, reasons and causes. *Journal of Philosophy* **60**: 684–700

de Marchi N, Blaug M (eds.) 1991 *Appraising Economic Theories: Studies in the Methodology of Research Programs*. E. Elgar, Aldershot, UK

FriedmanM 1953 The methodology of positive economics. In: *Essays in Positive Economics*. University of Chicago Press, Chicago, pp. 3–43

Hands D W 1985 The structuralist view of economic theories: the case of general equilibrium in particular. *Economics and Philosophy* **1**: 303–35

Hands D W 1992 *Testing, Rationality, and Progress: Essays on the Popperian Tradition in Economic Methodology*. Rowman and Littlefield, Lanham, MD

Hausman D M 1992 *The Inexact and Separate Science of Economics*. Cambridge University Press, Cambridge, UK

Hausman D M, McPherson M S 1996 *Economic Analysis and Moral Philosophy*. Cambridge University Press, Cambridge, UK

Hoover K 2000 *Causality in Macroeconomics*. Cambridge University Press, Cambridge, UK

Hutchison T W 1938 *The Significance and Basic Postulates of Economic Theory*. Repr. A. M. Kelley, New York, 1960

Latsis S J (ed.) 1976 *Method and Appraisal in Economics*. Cambridge University Press, Cambridge, UK

Lawson T 1997 *Economics and Reality*. Routledge, London

McCloskey D N 1985 *The Rhetoric of Economics*. University of Wisconsin Press, Madison, WI

Machlup F 1955 The problem of verification in economics. *Southern Economic Journal* **22**: 1–21

Mäki U, Gustafsson B, Knudsen C (eds.) 1993 *Rationality, Institutions and Economic Methodology*. Routledge, London

Mill J S 1836 On the definition of political economy and the method of investigation proper to it. Repr. *Collected Works of John Stuart Mill*, Vol. 4. University of Toronto Press, Toronto, ON, 1967

Mirowski P 1989 *More Heat Than Light*. Cambridge University Press, Cambridge, UK

Robbins L 1935 *An Essay on the Nature and Significance of Economic Science*, 2nd edn. New York University Press, New York

Rosenberg A 1992 *Economics–Mathematical Politics or Science of Diminishing Returns*. University of Chicago Press, Chicago

Samuelson P A 1947 Foundations of Economic Analysis. Harvard University Press, Cambridge, MA

Sen A 1992 *Inequality Reexamined*. Harvard University Press, Cambridge, MA

Senior N W 1836 *Outline of the Science of Political Economy*. Repr. A. M. Kelley, New York, 1965

Weintraub E R 1991 *Stabilizing Dynamics*. Cambridge University Press, Cambridge, UK

D. M. Hausman

Ecotourism

Ecotourism is criticized for being a vague and elusive notion, but, as with *Sustainable Development* (SD) (see *Sustainable Development*) this may prove its strength as a unifying concept which varied interests can identify with. However interpreted, ecotourism has commanded a degree of attention out of all proportion to its market significance because it appeals to all stakeholders. Destinations welcome the prospect of non-destructive natural resource utilization; increasingly discerning and differentiated tourists turn more and more to alternative, specialist, vacation experiences, and the tourism industry recognizes market potential.

1. Defining 'Ecotourism'

It is no coincidence that much of the discussion about ecotourism parallels the weighty debate surrounding the definition of SD. This is because the high ground claimed by ecotourism is that it constitutes a form of sustainable tourism. The most commonly used definition of ecotourism is that coined by the Ecotourism Society as 'responsible travel to natural areas which conserves the environment and improves the welfare of local people' (Lindberg and Hawkins 1993,

p.8). How this pivotal definition has evolved, and is evolving, mirrors the SD debate. The principles that ecotourism should embody fluctuate somewhere between a strong and weak green stance.

The first of these principles is that ecotourism should be 'green', recognizing a two-way link with nature conservation. This is incorporated in early definitions, such as that of Ceballos-Lascurain in 1987 (credited with first coining the term and giving it its preliminary definition) who suggested that ecotourism 'consists in traveling to relatively undisturbed or uncontaminated natural areas with the specific objective of studying, admiring, and enjoying the scenery and its wild plants and animals, as well as any existing cultural manifestations (both past and present) found in these areas' (Weaver 1998, p.15). Orams suggests that this perspective is essentially passive, requiring only that the activity does not result in negative impacts on the physical environment (Weaver 1998, p.16).

The second principle, incorporated in the Ecotourism Society definition, is that ecotourism should be responsible. This implies a more active position, involving behavioral/lifestyle changes by participants together with actions that contribute to environmental conservation and local livelihoods (Weaver 1998).

The third principle, essentially pragmatic, is that in order to be sustainable, ecotourism must recognize the interests of all stakeholders, including the tourism industry and the tourists themselves. Tourism is an economic activity which must be capable of making a profit if it is to survive, and delivery of a successful product depends on tourist satisfaction.

If ecotourism embodies these principles, symbiotic relationships between the varying interests should follow, with environmental protection resulting both from and in enhanced standards of living for local populations, continued profits for the tourism industry, sustained visitor attraction, and revenue for conservation.

2. *The Ecotourism Spectrum*

The reality of ecotourism in practice is, more often than not, divergent from its conceptualization in theory. Inevitably, the identity of ecotourism as one of the fastest growing tourism market segments leads to the term's widest use, and abuse, in the travel and tourism trade. Tour operators frequently relabel their products as a marketing ploy. Many travel experiences prefixed with 'eco' such as ecotour, and eco-(ad)ventures are often little more than examples of environmental opportunism (Wight 1994). Ecotourism, thus interpreted, may be ecologically based but not ecologically sound. Furthermore, it is unlikely to incorporate sociocultural integrity or, especially in the longer term, economic viability, as it compromises the very resources upon which it is based.

It is therefore evident that ecotourists are not an homogenous group. Several writers have identified a spectrum of expectations, behavior, and consequent impacts. Lindberg suggests a fourfold typology, ranging from handcore, through dedicated and mainstream, to casual nature tourists (Fennell 1999, p.57). Orams suggests that the continuum of ecotourism paradigms is bounded by polar extremes representing high human responsibility on the one hand and low human responsibility on the other (Weaver 1998, p.16). Paralleling the deep:shallow ecology debate (see *Ecology*, *Deep*) distinctions have also been made between deep, ecocentric, and shallow, human-centric, ecotourism.

Clearly ecotourists differ over many dimensions, but the typical profile is an average age of over 40 years, higher than average income and educational attainment, and a slight preponderance of males over females (Fennell 1999).

3. *The Conservation/Development Link*

It has become increasingly recognized that unless the needs of host populations are considered ecotourism will not be sustainable, as there will be no incentive for local communities to conserve. The creation of National Parks and wildlife reserves may satisfy the viewing needs of tourists, the profit maximizing aims of the tourism industry, and, if properly managed, the requirements of conservationists, but frequently the local population not only may not benefit but also may actively lose out. Within such protected areas, particularly in developing countries, they may be physically excluded from the resources upon which their livelihoods depend. This has led to the accusation of ecotourism as ecoimperialism (see *Ecological Imperialism*). The essential link between conservation and development is part of the much wider Parks People debate, but ecotourism is receiving increasing attention from policy makers and nongovernmental organizations as a theoretically nondestructive operationalization of the 'use it or lose it' philosophy, simultaneously making conservation pay and paying for conservation. It has also captured the interest of development agencies as a pro-poor strategy to enhance local livelihoods, although there is little evidence to date that tourism interventions have been driven by poverty elimination objectives.

4. *Ecotourism into the Millennium*

4.1 *Growth and Change*

The general consensus is that ecotourism is here to stay and set to grow. This is because it is a reflection of

much wider trends. Its increasing popularity may be attributable to being 'new,' alternative tourism, a reaction against the characteristics of mass consumption inherent in conventional mass tourism. Also, as the World Tourism Organization recently suggested, such travel may be regarded as a primary antidote to high-tech living. As well as constituting a lucrative niche, tour operators can respond to pressures of green-consumerism, increasing market credibility by offering a more responsible product. Inevitably, growth brings change and one of the most controversial aspects of change concerns scale of operation.

Many protagonists maintain that ecotourism must be small scale, low volume. Their view is that growth undermines sustainable ecotourism economies. However, the notion of small-scale, limited, specialist visitation, invariably highly priced, draws accusations of elitism; indeed Wheeller refers to ecotourism as 'egotourism' (Lew and Hall 1998, p.28). More fundamentally, David Western, the former head of the Kenya Wildlife Service, maintains that 'the principles applied to the mass market can do more for conservation—and alleviate more harm—than a small elitist market. Ecotourism, accepted in this way, is shifting from a *definition* of small-scale nature tourism to a set of *principles* applicable to any nature-related tourism … Surely what matters is not scale or motive, but impact' (Lindberg and Hawkins 1993, p.10). Coincidentally, many are also skeptical of the minimal economic impact, local, regional, or national, if ecotourism is restricted to small scale.

The question of scale however, raises issues of the carrying capacity of a locality (see *Human–Environment Relationship: Carrying Capacity*). Certain problems arise if the scope of ecotourism is broadened to embrace all nature tourism.

4.2 Divergent Theory and Practice

The potential benefits of ecotourism are widely documented, but despite various guidelines (Lindberg and Hawkins 1993, Lindberg et al. 1998), there is a considerable gap between theory and practice. This is partly attributable to the inevitable lag 'between the cutting edge of theoretical discussion and the application of theory to real-world situations' (Lew and Hall 1998, p.202). However, the disappointing performance of ecotourism to date in achieving several of its goals, such as increased local involvement, indicates the need to identify not only the nature of this gap but also its cause.

4.3 The Context

One cause is the failure to recognize the wider context within which ecotourism is cast. As Butler (1998, p.34) asserts 'tourism is part of the global system and cannot be tackled in isolation, spatially, economically or temporally.' It is vital that a move is made beyond a tourism-centric view, as it is 'inappropriate to discuss sustainable tourism any more than one might discuss any other single activity … we cannot hope to achieve sustainability in one sector alone, when each is linked to and dependent upon the others' (Butler 1998, p.28). These linkages exist not only between sectors but also between levels and interests, but there are discontinuities. What is sustainable for one interest in one place at one point in time in unlikely to be so for another. Williams and Shaw (1998) suggest place as an integrative concept, whereby prospects for tourism sustainability are mediated by contingencies of place. However, such contingencies are shaped by economic, sociocultural, political, ecological, institutional, and technical forces which are endogenous and exogenous as well as dynamic. Witness, for example, the inextricability of the global and the local as priority tourism projects in the Mekong subregion were postponed by the Asian Development Bank because of the Asian financial crisis.

Not only does ecotourism need to be set into the much wider context, but also it is imperative that it is set into specific context with regard to tourism. Milne (1998, p.47) suggests that 'in attempting to achieve more appropriate forms of tourism, it is also essential that we steer away from creating a dichotomy between 'alternative' and 'mass' tourism. Such a division serves little real purpose and diverts our attention away from the interlinked nature of all types of tourism development.'

The essential requirement of not examining ecotourism in isolation is also evident when considering the needs of foreign and domestic tourists. These two groups may construct divergent 'imaginative geographies' of the natural and cultural resources of the destination. This is particularly evident where cross-cultural differences are great. In China, domestic tourists favor artificial site enhancement compared with preservation in a pristine state, and have a higher tolerance of crowding and litter. These characteristics are clearly an anathema to foreign ecotourists.

This question of ethnocentricity may even create schisms between seemingly consensual viewpoints. Hinch (2001) surmises that, at first sight, the ecocentric world view of 'true' ecotourists should be congruent with that of indigenous cultures who tend to see themselves as being one with the land. However, he highlights potential discord should the ecotourists witness the hunting or trapping of species that may be an essential component of traditional livelihoods.

4.4 Geographical Perspectives

The contextual aspect is all important as ecotourism both shapes and is shaped by the social, cultural, geopolitical and geo-economic milieux within which it is set. Human geography and ecotourism thus share many common research foci. These include global–

local relationships, and the constraining and enabling conditions according to scale and circumstance (Hall and Page 1999). It is evident that human geography and ecotourism can each inform, and be informed by, the other. Ecotourism is not insignificant as a face and force of contemporary tourism.

See also: Ecological Economics; Environmental Challenges in Organizations; Environmental Risk and Hazards; Environmental Vulnerability; Human–Environment Relationships; Humanistic Geography; Tourism; Tourism, Geography of

Bibliography

Butler R 1998 Sustainable tourism—looking backwards in order to progress? In: Hall C M, Lew A A (eds.) *Sustainable Tourism: A Geographical Perspective*. Longman, Harlow, UK, Chap. 3, pp. 25–34

Fennell D A 1999 *Ecotourism: An Introduction*. Routledge, London

Hall C M, Page S J 1999 *The Geography of Tourism and Recreation*. Routledge, London

Hinch T D 2001 Indigenous territories. In: Weaver D B (ed.) *The Encyclopaedia of Ecotourism*. CABI, New York, Chap. 22, pp. 345–50

Lew A A, Hall C M 1998 The geography of sustainable tourism: Lessons and prospects. In: Hall C M, Lew A A (eds.) *Sustainable Tourism: A Geographical Perspective*. Longman, Harlow, UK, Chap 16, pp. 199–203

Lindberg K, Epler-Wood M, Engledrum D (eds.) 1998 *Ecotourism: A Guide for Planners and Managers, Vol. 2*. The Ecotourism Society, Bennington, VT

Lindberg K, Hawkins D E (eds.) 1993 *Ecotourism: A Guide for Planners and Managers, Vol. 1*. The Ecotourism Society, Bennington, VT

Milne S S 1998 Tourism and sustainable development: The global-local nexus. In: Hall C M, Lew A A (eds.) *Sustainable Tourism: A Geographical Perspective*. Longman, Harlow, UK, Chap. 4, pp. 35–48

Weaver D B 1998 *Ecotourism in the Less Developed World*. CABI, Wallingford, UK

Wight P 1994 Environmentally responsible marketing of tourism. In: Cater E, Lowman G (eds.) *Ecotourism: A Sustainable Option?* John Wiley, Chichester, UK, Chap. 3, pp. 39–55

Williams A, Shaw G 1998 Tourism and the environment: Sustainability and economic restructuring. In: Hall C M, Lew A A (eds.) *Sustainable Tourism: A Geographical Perspective*. Longman, Harlow, UK, Chap. 5, pp. 49–59

E. Cater

Ecumene

Ecumene (sometimes oecumene) is a rendition of the Greek word *oikoumene*, which the ancient Greeks applied to the inhabited (i.e., known to them) world. It conveys a sense of communion (as, notably, in religious ecumenism), a community of independent but interacting members, or a set of separate but interrelated societies and cultures. The notion is useful to anthropologists, geographers, historical sociologists, and world historians when they wish to indicate a large region that represents a sphere of persistent and effective interaction among a group of societies that have been influencing one another and have been shaped to some significant degree by a shared history. The concept of ecumene may thus serve to reflect the subdivision of the world into several separate spheres of culture-historical experience. More recently, with the growing interest in 'globalization,' the notion of a 'global ecumene' has emerged; the notion reflects the idea that the entire world is now becoming a single interaction sphere, and that processes (such as integration and constant cultural exchange) that used to characterize the more restricted ecumenes of the past now operate on a worldwide scale. This article will focus on the use of the term in anthropology, and to some extent in world history and historical sociology. The term has been most used by scholars interested in large-scale cultural dynamics over both time and space and who therefore tend to share in a common scholarly discourse.

1. Early Perspectives

Nineteenth century evolutionary anthropology (primarily Anglo-American) assumed that local societies were expressions of general cultural evolution; in effect, the ethnographic world was not mapped geographically but was viewed in terms of postulated temporal stages. By contrast, German ethnology (which was more firmly grounded in first-hand observation) emphasized the geographical distribution of cultural traits; it regarded cultural contacts and diffusion as the primary mechanisms of cultural dynamics. This fitted well with the commonplace experience that neighboring societies usually tend to be culturally similar, forming what German ethnology called 'cultural provinces'—a concept that stressed cultural continuities within regions.

Early twentieth-century American anthropology, whose worldview was derived from German anthropology by way of its founder Franz Boas, adopted this spatial perspective and saw the ethnographic world as being subdivided into 'culture areas' shaped by regional diffusion. This, however, was mainly true of the 'primitive' ethnographic world, which has usually been peripheral to the world centers of historically significant cultural developments. Even in the 'primitive' world, the diffusionist view raised some problems. The fact that North America, for example, had several culture areas rather than one, demanded some explanation of the cultural discontinuities between the culture areas (a problem that Ruth Benedict proposed

to solve by postulating distinctive regional patterns of culture). When one turned to regions of historically complex societies (such as Europe, the Mediterranean world, the Middle East, and South or East Asia) one found long-standing interactions in the midst of marked cultural heterogeneity.

The mounting dissatisfaction with the simplistic view of a world organized into a series of disparate culturally homogenous areas led Alfred Kroeber (1945) to draw attention to the existence on the world map of a central and culturally diverse area of 'civilization' ('a great web of cultural growth' stretching from Gibraltar to China). Within it, there were centers of consistent creativity, generating ideas and techniques that diffused through this area and sometimes to the peripheral areas beyond. Kroeber adopted the term oikoumene for it and his discussion may be seen as an early and prescient invitation to deal with the wider non-'primitive' world from an anthropological perspective. In this view, a culture area became a particular kind of ecumene—local and relatively, but not entirely, homogenous. As Kroeber (1936) had previously shown, such an area exhibited center-periphery relations through which diffusion throughout it took place.

However, Kroeber's 1945 invitation to expand the anthropological vision to worldwide cultural dynamics had scarcely any takers (for an unusual exception, see Hewes 1970). In the 1930s, anthropology began to abandon its regional ethnological perspectives in favor of a concentration on cultural dynamics. This turn took two directions. It either studied them within single societies or, when it dealt with contact, it studied it as it occurred between two neighboring societies. In the latter case, anthropology did develop a rich collection of ideas about cultural transmission, but the notion of the ecumene had no place in such a locally focused endeavor and it lay dormant in anthropology for some 30 years. In the meantime, the issues it raised came to the fore among world historians and historical sociologists (see Grew 1995).

2. *World History*

A 'universal history' had been a major project of the eighteenth century Enlightenment, and the idea was revived in the earlier twentieth century in the comparative analysis of civilizations by such historians as Spengler (1926–1928) and Toynbee (1934–1961). Though these comparative studies did not fashion a unified historical narrative on a world scale, they did bring to attention the existence of 'high' periods in the histories of non-Western societies. This undermined the Eurocentrism of world history of the time, in which the first ecumene of any importance seemed to have arisen only with the fifteenth century expansion of the West. Now, attention was being drawn to the dominance of preceding regional ecumenes, such as the Islamic world or China-centered eastern Asia, with the West being frequently peripheral to them. Some historians, notably McNeill 1963), began to frame world history in terms of a succession of dominant ecumenes, thereby wrenching it out of the complete grip of Western history.

The disappearance of the Western colonial empires after the Second World War reinforced the trend, for it raised questions about the very primacy of the West in world history, the reasons for it, its transience, and the lessons all this held for the immediate future. Here, post-colonial political and sociological concerns merged with historical ones, calling into being a historically oriented political science and political sociology. To be sure, the focus at first remained Eurocentric, as shown in one of the first formulations of these new perspectives—that by Wallerstein's (1974) 'world-system' analysis. Wallerstein took modern Western expansion to be a historically unique occurrence arising from specifically Western historical factors, such as the rise of capitalism. Nevertheless, the formulation was a first step toward an exceptionally fruitful debate and a series of analyses and reanalyses (e.g., Abu-Lughod 1989, Frank and Gills 1992, Smith 1991).

The debate brought to historical consciousness a variety of preceding smaller 'world-systems,' ecumenes that coexisted with one another in different patterns of integration, and that went back into prehistory. This new world history transcends the traditional national units and focuses on such interregional processes as long-distance trade, economic interdependence, interregional enterprise, large-scale imperial polities, and continent-wide cultural relations between centers and peripheries. It recognizes large areas of economic, political, and cultural interaction, their dynamics, and their periodic demise and reorganization into new ecumenes (Mazlich and Buultjens 1993). As is natural with any new paradigm, it has led to detailed examinations of local dynamics of economic, political, and cultural exchanges. This development represents, in some sense, a retreat from worldwide perspectives and it often provides correctives to the sweeping assertions of the grand picture. But many of these local studies take the world setting as a necessary context for their analyses and thus address the broader issues.

3. *Modern Anthropology*

Similar currents began to sweep anthropology in the later 1960s. The paradigm that had dominated anthropology since the 1930s and that dealt with the systemics of single cultures was lapsing. The demise of colonial rule made an ahistorical approach to ethnography increasingly unrealistic, and this called attention to the larger forces swirling around the anthropologist's supposedly isolated societies. With the introduction of

historicity into modern ethnography, it has become obvious that historicity is also necessary in dealing with the ethnographic past.

This new vision of former 'primitives' as being bounced about by the surrounding world was also reinforced by the popular journalistic discussions of 'globalization,' including large-scale international migrations, ethnic diasporas, world labor markets, economic integration, transnational economic and cultural flows, multiculturalism, cultural 'creolization,' the unifying role of the cyber net, and the divisive role of networks. These processes are sometimes seen as dismantling the separate old regional ecumenes that used to divide the world. What looms instead is a single community of interacting societies whose very borders appear increasingly to come into question—in brief, a single 'global ecumene' whose nature is being explored by anthropologists with the use of anthropological concepts (e.g., Foster 1991, Hannerz 1992). Thus, Kroeber's ancient oikemene has been transposed out of culture history and into the contemporary era.

In sum, the concept of the ecumene allows anthropologists to deal with interactions among disparate societies. Previously, the only concept that aggregated groups of societies was that of the culture area. But culture areas projected a homogenizing view of cultural interaction and it tied cultural diffusion to territorial continuities. The ecumene, by contrast, allows anthropologists a more realistic view of cultural dynamics—one in which very different societies interact and influence one another selectively and without ceasing to be different. As the mechanisms of such influence become better understood, it becomes clear that they can operate over discontinuous territory and thus encompass, in the modern age, the entire globe. At the same time, it should be noted that much of this new concern in anthropology has more to do with subject matter than theory, for it revives such old-standing anthropological interests as innovation, diffusion, culture contact, and selective acculturation.

See also: Borders, Anthropology of; Center–Periphery Relationships; Civilizations; Cosmopolitanism; Creolization: Sociocultural Aspects; Cultural Evolution: Theory and Models; Diaspora; Diffusion: Anthropological Aspects; Diffusion: Geographical Aspects; Global History: Universal and World; Globalization and World Culture; Globalization, Anthropology of; Human Rights, Anthropology of; Hybridity; Multiculturalism; Multiculturalism, Anthropology of; World Systems Theory

Bibliography

Abu-Lughod J 1989 *Before European Hegemony: The World-System, AD 1250–1350*. Oxford University Press, New York

Foster R J 1991 Making national cultures in the global ecumene. *Annual Review of Anthropology* **20**: 235–60

Frank A G, Gills B K 1992 The five thousand year world system: An interdisciplinary introduction. *Humboldt Journal of Social Relations* **18**: 1–79

Grew R 1995 Review of Costello P World historians and their goals: Twentieth century answers to modernism. *History and Theory* **34**: 371–95

Hannerz U 1992 The global ecumene as a network of networks. In: Kuper A (ed.) *Conceptualizing Society*. Routledge, London

Hewes G W 1970 Mapping the growth of the Old World ecumene. In: *VIIth International Congress of Anthropological and Ethnological Sciences VIII*. Nauka, Moscow, pp. 373–79

Kroeber A L 1936 Culture elements distributions: III. Area and climax. *University of California Publications in American Archaeology and Ethnology* **48**: 1–242

Kroeber A L 1945 The ancient oikoumene as an historic culture aggregate. *Journal of the Royal Anthropological Institute* **75**: 9–20

Mazlich B, Buultjens R (eds.) 1993 *Conceptualizing Global History*. Westview Press, Boulder, CO

McNeill W H 1963 *The Rise of the West: A History of the Human Community*. University of Chicago Press, Chicago

Smith A K 1991 *Creating a World Economy: Merchant Capital, Colonialism, and World Trade, 1400–1825*. Westview Press, Boulder, CO

Spengler O 1926–1928 *The Decline of the West*. 2 Vols. Knopf, New York

Toynbee A 1934–1961 *A Study of History*. 12 Vols. Oxford University Press, New York

Wallerstein I M 1974 *The Modern World-System*. 2 Vols. Academic Press, New York

I. Kopytoff

Edgeworth, Francis Ysidro (1845–1926)

It is only recently that Edgeworth has become fully recognized as one of the greatest economists of his time, as well as one of its greatest statisticians—a unique combination. In both disciplines his best work was so subtle and profound that it took many years for its meaning to be understood. Edgeworth is thus a striking example of one whose reputation as an original thinker has grown ever since his death, slowly at first and then quite rapidly. After a century, Edgeworth's thought is only beginning to find an echo. *The future will no doubt acknowledge the profundity of his vision more and more* (Allais 1978, p. 191, referring to *Mathematical Psychics* (1881); original emphasis, as in all quotations here).

1. Life

Francis Ysidro Edgeworth was born on February 8, 1845 at the family home Edgeworthstown in County Longford, Ireland, the youngest of five sons. His father Francis Beaufort Edgeworth (1809–1846) was one of 22 children of Richard Lovell Edgeworth

(1744–1817), an inventor and educationist whose favorite child was the celebrated novelist Maria Edgeworth (1767–1849). Father and daughter were much interested in political economy and numbered among their acquaintance Bentham Malthus, and Dugald Stewart; after her father's death Maria also developed a warm friendship with David Ricardo.

Edgeworth's father, Francis Beaufort Edgeworth, was a favorite of both his elderly father and his half-sister Maria. His mother, 2 years younger than her step-daughter Maria, was Frances Anne Beaufort (1769–1865), of Huguenot stock and elder sister of the admiral who devised the famous Beaufort marine wind scale. Naturally averse to mathematics, Francis left Trinity College, Cambridge without a degree. After a brief courtship at age 22 he married Rosa Florentina Eroles (1815–1864), who was 16, poor, with 'beautiful eyes,' and daughter of an exiled Spanish general. After some years in Italy Francis struggled unsuccessfully as owner of a private school in Eltham near London, and in 1840 returned home to Edgeworthstown to run the estate. He died aged 37 after a long illness, in the worst year of the Irish famine.

Despite his disappointing career Edgeworth's father was not without talent. His writings on poetry and philosophy sufficiently impressed John Stuart Mill to lead him to enquire, in 1840, whether Francis might be available to tutor 'the eldest son … of a person of very high rank,' adding mysteriously that 'there is, probably, no situation of the kind in England in respect to which more important consequences may depend on its being well filled' (Mineka 1963, p. 439). Moreover, Francis's friendship was highly valued by the poets Alfred Tennyson and Edward FitzGerald and by the great mathematical physicist Sir William Rowan Hamilton.

Edgeworth himself was actually baptized Ysidro Francis Edgeworth, a form he sometimes used even as an adult. 'Marshall, remembering his mixed parentage, used to say: "Francis is a charming fellow, but you must be careful with Ysidro."' (Keynes 1926, p. 152). He was educated by tutors at home, committing to permanent memory whole books of Homer, Virgil, Milton, and Pope. A notable linguist, he was well-read not only in classical Greek and Latin but also in French, German, Italian, and Spanish literature. In 1862 he entered Trinity College, Dublin to read classics and in 1867 transferred to Oxford, where in 1869 he gained a First in the honors examination in classics and philosophy.

He spent the next 22 years in London, living on the heights of Hampstead in two bare rooms that he continued as a *pied-à-terre* after he moved to Oxford in 1891. Supplementing his slim private means by occasional earnings as a part-time teacher in further education, he studied law systematically (specializing in commercial law) until he was called to the Bar at the Inner Temple in 1877; he never practiced. He also read widely in mathematics and physics, his interests perhaps influenced by those of his father's friend Hamilton. Certainly, by 1877 he had gained 'a confident and creative mastery of the calculus of variations' (Stigler 1986, p. 306).

In 1871 he became an 'original' (and lifelong) member of the new Savile Club in central London, whose friendly and distinguished atmosphere apparently suited him well, so much so that during the years 1878–1882 he was quite active in the club's affairs as either a Committee member or an Honorary Secretary or both. Indeed, it was precisely then and not (as is sometimes alleged) as a reclusive Oxford don that he wrote and published *Mathematical Psychics* (1881). Two years later he began the many articles on probability and statistics whose quality carried him upwards through a series of appointments at King's College, London, until in 1891 he was appointed Drummond Professor of Political Economy at Oxford, tenable at All Souls. He retired from this post in 1922 but the College remained his principal home until he died (from pneumonia) in Oxford on February 13, 1926. As the last surviving male Edgeworth in Europe he inherited the family estate from his brother Antonio in 1911 but did not live there.

Although Edgeworth never married it was not for want of trying. He was smitten by the handsome and imperious Beatrice Potter who found him 'gentle-natured … excessively polite and diffident … pathetic … pedantic'—and boring, even though she glimpsed 'a hidden fire burning within.' Insufficiently discouraged, in June 1889 he published a glowing review of her early work and pursued her to a Co-operative Congress in Ipswich where, in spite of his 'agonized expressions of romantic regard,' she remained adamantly unimpressed (MacKenzie and MacKenzie 1982, pp. 283–4, 286–7). Three years later she married Sidney Webb.

The words diffident, excessively polite, and pedantic recur often in friends' descriptions of Edgeworth's character, which perhaps accounts for Heilbroner's remark that: 'Neurotically shy, he tended to flee from the pleasures of human company' (1999, p. 173). However, just as frequently those same friends also described Edgeworth as affectionate, humorous, and kind; 'in my house … few visitors were so beloved as he' (Bonar 1926, p. 652). Known for his hospitality to foreign visitors, he had 'the widest personal acquaintance in the world with economists of all nations' (Keynes 1926, p. 151). The young Joseph Schumpeter, visiting Edgeworth at All Souls, was memorably welcomed to breakfast by his host with 'rock pheasant' (presumably, cold roast grouse) and champagne (Samuelson 1951, p. 53n).

Edgeworth was twice President of Section F of the British Association for the Advancement of Science (1889, 1922), President of the Royal Statistical Society in 1912, one of the original Fellows of the British Academy (1903), and an honorary Doctor of Civil Law from the University of Durham (1889). The most

important of his few honors was his sole editorship of the *Economic Journal* (*E. J.*), which lasted from 1890 until J. M. Keynes succeeded him in 1911, whereupon he became Chairman of the Editorial Board. Keynes, busy at Versailles in 1919, co-opted Edgeworth to serve as Joint Editor and he continued active in this capacity 'up to the last day of his life' (Keynes 1926, p. 140).

No complete bibliography of Edgeworth's prolific writings in economics and statistics has yet been published but an incomplete count yields four very short books, about 300 articles and shorter papers, and more than 200 book reviews. Perhaps reflecting his position as editor of the *E. J.*, after 1890 his major papers in economics tended to be surveys, learned and complex and delighting in brilliant paradoxes, rather than articles on original themes. Many of these, especially the surveys of the theories of distribution, index numbers, international trade, monopoly, railway rates, returns to scale, and taxation, he gathered together and reprinted in Edgeworth (1925). All his many papers in probability and statistics (except (1887) and a few others) were reprinted in 1996.

2. *Mathematical Psychics*

Although subtitled 'An Essay on the Application of Mathematics to the Moral Sciences,' *Mathematical Psychics* actually contains *three* analytically distinct essays, all rather jumbled together. The first is on methodology and comprises pp. 1–16 (undated page references here are all to (1881)); Appendix I—'On Unnumerical Mathematics'; and Appendix II. It develops two separate themes, the first being a visionary stress on the importance of maximum problems for economics, at both the individual and social level. It goes on to argue that in many respects problems in economics are analogous to those of classical mechanics, so that '*Mécanique Sociale* may one day take her place along with *Mécanique Célèste*, throned each upon the double-sided height of one maximum principle' (p. 12). The second theme is less grandiose and develops in detail the importance of 'unnumerical mathematics' for positive economics. The emphasis here on second-order conditions for a maximum sounds very modern: 'the condition that a required curve shall be, or shall not be, convex … this very relation of concavity, not a whit more indefinite in psychics than in physics … quarried from such data as the law of decreasing utility, of increasing fatigue, of diminished returns to capital and labour' (p. 92).

Edgeworth wrote most of his second essay, on the *Utilitarian Calculus*, when he was primarily a mathematically inclined philosopher. It consists of pp. 56–82 (which essentially reprints (1879)); Appendices III and IV; pp. 116–8 and 122–5 of Appendix VI; and pp. 126–34 of Appendix VII. Earlier, pp. 35–79 of his monograph *New and Old Methods of Ethics* (1877) had applied the calculus of variations and the psychophysics or Fechner and others to Sidgwick's utilitarianism. The resulting 'exact utilitarianism' sought 'the greatest quantity of happiness of sentients, exclusive of number and distribution,' where 'sentients' explicitly refer not just to mankind but to the whole animal creation (1877, p. 35). After a pioneering and detailed variational analysis, employing Lagrange multipliers for both finitely and infinitely many constraints as well as the idea of a continuum of rational agents, he concluded that for modern humankind '(T)here is no indication that … a law of distribution other than equality is to be wished' (1877, p. 78).

The substance of some comments by Alfred Barratt on this monograph were incorporated in Edgeworth's article in *Mind* (1879), which enhanced his earlier analysis by considering the possible negative utility from work as well as positive utility from consumption; but also diminished it disastrously, by treating different individual capacities for enjoyment not, as before, as essentially random variations but instead as systematic differences in the sentient's position in the 'order of evolution.' This drastically modified his previous conclusions concerning the optimality of equal distribution, since now '*the more capable of pleasure shall have more means*' (1879, p. 398, 1881, p. 64), and led him to some startling arguments in favor of inequality.

When he wrote those indefensible passages Edgeworth was not yet an economist. But two years later, as virtually an apprentice in economics to his friend, neighbor, and mentor William Stanley Jevons, he wrote his masterpiece, the *Economical Calculus*. This third essay, entirely on positive economics, consists of pp. 16–56; Appendix V; pp. 118–22 of Appendix VI; and pp. 134–48 of Appendix VII. Its fundamental theorem asserts that: 'Contract without competition is indeterminate … Contract with *perfect* competition is perfectly determinate … Contract with more or less perfect competition is less or more indeterminate (p. 20). The wholly normative analysis of the earlier *Utilitarian Calculus* depended critically on two doubtful assumptions, the cardinal measurability of individual utilities, and their interpersonal comparability, but this theorem needs neither of them. Indeed, it is richly ironical that Edgeworth, the most uninhibited of utilitarians, should have produced in this third essay the most far-reaching (though implicit) ordinalist analysis of any neoclassical economist.

Few works in economics have been so misunderstood as the *Economical Calculus*. Over the years commentators have discovered there not what it actually contains but only what their own preoccupations have predisposed them to find. For example, those with the *déformation professionnelle* sometimes induced by Walrasian general equilibrium theory have tended to confuse the aims of Edgeworthian recontract with those of Walrasian *tâtonnement*, while those who believe neoclassical economics to be overly tender towards capitalism have found in

Mathematical Psychics 'a kind of Panglossian Best of All Possible Worlds' (Heilbroner 1999, p. 173). This is travesty. Edgeworth believed that *imperfect* competition, with its few competitors, occurs frequently in economic life. Hence, by the theorem, indeterminate contract should occur frequently too, with its 'characteristic evil' (p. 29) of '*deadlock*, undecidable opposition of interests, unceasing strife and tumult' (the last phrase a translation of Edgeworth's Greek quotation from Demosthenes). This impasse would then lead to 'a general demand for a *principle of arbitration*' (p. 51), a demand which Edgeworth proposes to meet by proffering the following utilitarian rule: the arbitrator should choose that contract which maximizes the total utility of all the parties concerned. 'Thus the *economical* leads up to the *utilitarian calculus*' (p. 56). In the theorem, *articles of contract* are whatever is subject to bargain, such as commodities or property rights, and parties may *recontract* away from any existing contract. *The field of competition* consists of *all* those parties (e.g., sellers and buyers) willing and able to enter into contracts for the articles in question. In these general conditions, which include 'the political struggle for power' (p. 16) as well as commercial exchange, there may well be neither prices nor markets for the articles. Hence, the theory of supply and demand in perfect markets, as already worked out by Jevons, Marshall, and Walras, does not apply to the general case of Edgeworthian contract and it becomes necessary to devise radically new concepts of equilibrium, appropriate to this general setting.

Edgeworth's first such concept is a *settlement*, 'a contract which cannot be varied with the consent of all the parties to it,' and the second *final settlement*, 'a settlement that cannot be varied by recontract within the field of competition' (p. 19). A settlement can, therefore, withstand any variation from it proposed by any party to the original contract, while a final settlement can withstand any variation proposed by any interested agent, whether an original contractor or not. A settlement is, thus, precisely an allocation that yields *maximum d'ophélimité*, a concept first defined 16 years later by Pareto and now known universally as *Pareto optimality*, its Edgeworthian origins long forgotten. Note that while settlement is a concept of positive economics, Pareto optimality belongs to welfare economics.

Edgeworth explicitly provides for coalitions, whose members may recontract between themselves in order to improve their positions under prevailing contracts. As pointed out first by Shubik (1959), from this perspective the set of final settlements is precisely the core of the relevant cooperative game. (While working with John von Neumann in late 1940, Morgenstern 'showed him the *contract curve*' (Rellstab 1992, p. 83), but their epoch-making book of 1944 has no reference to Edgeworth.)

When there are only two agents they are the only possible contracting parties, so all settlements are also final settlements. Now for Edgeworth, competition necessarily involves at least three agents (see 1894, p. 378), so two agents represent '(α) Contract without competition' (p. 19). For this case Edgeworth invented and deployed general utility functions and indifference curves to show that the final settlements form a continuous contract curve containing infinitely many points. Contract is, thus, completely indeterminate. (His analysis here remains central to standard microeconomic theory, which uses the 'exchange box' invented in 1906 by Pareto to amend Edgeworth's original exposition; by a delicious irony, economists usually attribute Pareto's device to Edgeworth!)

At the opposite pole to bilateral exchange lies perfect competition, where indefinitely many traders exchange infinitely divisible articles of contract among themselves. Here Edgeworth shows each resulting market equilibrium allocation to be a settlement (the first proof of the *First Fundamental Theorem of Welfare Economics*), as well as a final settlement. He conjectures the true result (proved rigorously 91 years later by Debreu and Scarf) that, in the perfectly competitive replica economy on which his analysis is based, every final settlement is a market equilibrium allocation of some initial endowment. Since the latter allocations are usually unique or at least finite in number, this shows that perfect competition is perfectly determinate.

Edgeworth next displays extraordinary technical virtuosity in dealing with the intermediate case of imperfect competition, where he proves that for a replica economy with $2n$ agents (where $1 < n < \infty$), the set of final settlements always shrinks in size as n increases, so that contract becomes less and less indeterminate as the degree of competition increases. The key property needed for this striking result is the strict convexity of individual indifference curves, which Edgeworth actually proves by using assumptions that imply cardinal measurability of utility, but which he could just as well have assumed directly, without need of cardinality.

This completes the proof of the theorem (for details, see Newman 1994). As a coda, Edgeworth joyfully and correctly proves that his arbitration rule always produces a settlement. But this has two grave difficulties. First, the rule itself depends on cardinal measurability of utility. Second, an acceptable rule should produce not just a settlement but a final settlement. Otherwise, for some agents the utilitarian allocation might require their allocated utility levels to be below their initial levels, and so its implementation would require compulsory redistribution.

3. *Works in Statistics*

Soon after *Mathematical Psychics* appeared, and possibly influenced by Galton, Edgeworth became attracted intellectually by the 'law of error,' i.e., the

normal distribution and the central limit theorem. For the next 10 years he devoted his formidable powers and energy to probability, mathematical statistics, and applied statistics, publishing over 30 papers and one brief monograph (1887). Although subsequently he split his creative efforts more evenly between statistics and economics, in those later years he published a further 75 papers in statistics. R. A. Fisher rated Edgeworth of 'equal eminence' with Karl Pearson in statistics, while a distinguished historian of the subject has judged him to be 'the leading theorist of mathematical statistics of the latter half of the nineteenth century' (Stigler 1987, p. 98).

Edgeworth set himself the task of adapting classical probability theory, developed originally by Pierre Simon Laplace and others for the physical sciences, to the problems of statistical inference in the social sciences. For example, regarding the returns of the Bank of England, 'to estimate the probability that the differences in the averages for different weeks and months are not accidental … under what circumstances does a difference in figures correspond to a difference in fact' (quoted in Stigler 1986, p. 308). In 1883 an early result of these efforts was possibly the first derivation of Student's *t* distribution (reprinted as McCann 1996, II, No. 1), followed in 1885 by a pioneering version of the analysis variance for a two-way classification (McCann 1996, III, no. 2).

His methodological views on statistical inference in the social sciences were given in his centrally important paper *Methods of Statistics* of 1885 (McCann 1996, II, no. 3—this reprint has quite illegible diagrams). Generally, Edgeworth took an undoctrinaire Bayesian approach but was quite willing also to use frequentist methods rather than inverse probability, as when in 1905 he used significance tests to compare means (McCann 1996, I, no. 26). In (1887) and elsewhere he stressed the importance of maximizing (expected) utility rather than probability in devising appropriate statistical procedures: 'what have we to do with probability or frequency, except as a guide to our actions?' (1887, p. 53). But that monograph has no hint of an axiom system that would unify probability and utility, such as those developed much later by Ramsey, de Finetti, and Savage.

He bequeathed his name to statistical posterity with his development in 1905 of a new series expansion—subsequently the *Edgeworth series*—as a means of adding correction terms to the central limit theorem in order to deal with skew distributions (McCann 1996, I, no. 26). Less eponymous but more widely known is his term coefficients of correlation and his demonstration in 1892 of their role in expressing the parameters of the multivariate normal distribution (McCann 1996, III, no. 16 and II, no. 21). A less appreciated novelty, anticipating R. A. Fisher, was his proof in 1908–9 of the asymptotic efficiency of maximum likelihood estimators (McCann 1996, I, no. 12).

Finally, as so often with Edgeworth, a pioneering contribution in 1888 to inventory theory (McCann 1996, III, no. 4) was ignored by economists and statisticians alike until very recently, even though it 'filled a gaping hole in orthodox classical monetary theory' (Laidler 1991, p. 184). This delightfully written essay on banking demonstrated rigorously and quantitatively what had been known to bankers intuitively and qualitatively for centuries, that a bank's prudential reserve level rises proportionately, not with its liabilities, but with their square root.

Bibliography

Allais M 1978 Theories of general economic equilibrium and maximum efficiency. In: Schwödiauer G (ed.) *Equilibrium and Disequilibrium in Economic Theory*. Reidel, Dordrecht, The Netherlands

Bonar J 1926 Memories of F. Y. Edgeworth. *Economic Journal* **36**: 647–53

Edgeworth F Y 1877 *New and Old Methods of Ethics*. Parker, Oxford, UK

Edgeworth F Y 1879 The Hedonical Calculus. *Mind* **IV**: 394–409

Edgeworth F Y 1881 *Mathematical Psychics*. Kegan Paul, London

Edgeworth F Y 1887 *Metretike*. Temple, London

Edgeworth F Y 1894 Competition and regulation. In: Palgrave R H I (ed.) *Dictionary of Political Economy*, Vol. I. Macmillan, London

Edgeworth F Y 1925 *Papers Relating to Political Economy*, 3 Vols. Macmillan, London

Heilbroner R L 1999 *The Worldly Philosophers*, 7th edn. Simon and Schuster, New York

Keynes J M 1926 Francis Ysidro Edgeworth 1845–1926. *Economic Journal* **36**: 140–53

Laidler D 1991 *The Golden Age of the Quantity Theory*. Princeton University Press, Princeton, NJ

McCann C R Jr (ed.) 1996 *F. Y. Edgeworth: Writings in Probability, Statistics and Economics*, 3 Vols. Edward Elgar, Cheltenham, UK

MacKenzie N, Mackenzie J (eds.) 1982 *The Diary of Beatrice Webb: Volume One 1873–1892*. Virago, London

Mineka F E (ed.) 1963 *The Earlier Letters of John Stuart Mill 1812–1848*, 2 Vols. University of Toronto Press, Toronto, ON

Newman P 1994 Edgeworth's Economical Calculus. *Metroeconomica* **45**: 99–126

Rellstab U 1992 New insights into the collaboration between John von Neumann and Oskar Morgenstern on the 'Theory of Games and Economic Behavior'. In: Weintraub E R (ed.) *Towards a History of Game Theory*. Duke University Press, Durham, NC

Samuelson P A 1951 Schumpeter as a teacher and economic theorist. In: Harris S E (ed.) *Schumpeter: Social Scientist*. Harvard University Press, Cambridge, MA

Shubik M 1959 Edgeworth market games. In: Tucker A W, Luce R D (eds.) *Contributions to the Theory of Games IV*. Princeton University Press, Princeton, NJ

Stigler S M 1986 *The History of Statistics*. Harvard University Press, Cambridge, MA

Stigler S M 1987 Edgeworth as a statistician. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics*, 4 Vols. Macmillan, London

P. Newman

Education and Economic Growth

The manner and rate of *economic growth* among nations and its causes has been a topic for theory building, speculation, and investigation since the inception of the discipline of economics. Early economists such as Adam Smith and John Stuart Mill recognized that there must be a link between *education*, broadly defined, and economic growth. Explorations of the nature of the linkage did not seriously get under way until the early 1960s with the work of American economists Theodore Schultz (1961) and Gary Becker (1993), who developed a theory of human capital formation and then sought to analyze the rate of return of investment in education and training. The last half of the twentieth century has abundantly documented that linkages exist but that they are neither inevitable nor straightforward in their relationships and effects. Early naive theories of *human capital* that proposed strict linear relationships between formal education and economic growth at both the individual (micro) and societal (macro) levels have now been abandoned in the face of overwhelming evidence that only few and modest direct relationships can be obtained no matter what definition or method of analysis is used.

1. *Defining Growth and Education*

Reasons for this maze of complexity are not hard to find. First we have to consider that when we speak of 'economic growth' there are many different ways in which it can be conceptualized and measured. Changes in Gross National Product (GNP), Gross Domestic Product (GDP), import/export ratios, unemployment rates, and various indices of economic activity are just a few of the many ways, individually or in combination, by which economic growth can be operationally defined. The means by which rates of growth are calculated can also have a marked impact on whether 'growth' is discerned or not (e.g., year-to-year change, rolling averages, decade-to-decade change). Any growth also inevitably occurs within convoluted national, regional, and global contexts that includes substantial changes over time in factors such as population growth, immigration patterns, human capital, technological innovations, structural unemployment, evolving taxation and accounting policies, trade agreements and tariffs, 'hidden' or 'black market' economies, regulatory environments, and political leadership. The degree to which these changes are measured and factored into economic growth models to create 'value-added' or 'adjusted' rates varies considerably among economists and particular schools of thought within which they work.

We also need to consider that as the world economy changes over time, particular sectors of national economies may also alter significantly. Increasingly in many nations applications of knowledge and information drive economic indices of productivity rather than material goods. Changing production patterns that emphasize customization and just-in-time delivery have also significantly altered both our conceptions and means of economic growth. Large corporations within the highly interdependent global economy often have worldwide revenues that exceed the GNP of many nations, and the dispersion of corporate assets among many countries makes quantifying economic growth in any kind of reliable manner increasingly difficult.

Further complications arise when a particular model of economic growth (with its attendant assumptions and foci) is analytically and econometrically related to education. 'Education' is a broad concept that includes a variety of subcategories or sectors. The 'formal' system would include primary, secondary, college/university, and certification forms of education (e.g., secondary education equivalency examinations, specialized training programs, and examinations that lead to industry-recognized certification). The formal system can be further subdivided into public and private spheres. Economists and other researchers generally have scant information about the private sector. Virtually all models of economic growth include substantial data available from the public sphere and only modest amounts from the private sector. In some nations, this means that better than one-quarter of the formal education sector is ignored as 'missing' and frequently forgotten data. Formal education in many nations can also be divided into academic and vocational/technical in terms of its focus and once again differing results are seen when considering the impact of these different forms of education on the economy.

An 'informal' education sector embraces non-credit forms of education and peer-learning systems (operating without the knowledge of management) that occur on the job and within the culture generally (e.g., evening 'courses' offered to the local community hosted by a local secondary school but having no formal relationship to credentialing or degree-granting institutions). There also exists the arena of job-related training or education which can take the form of job-embedded education (internships, apprenticeships, union-led training, formal company programs)

and industry- or state-led training (e.g., vocational/technical education for adults) or retraining, often conducted under the guise of either economic competitiveness or ameliorating the effects of structural unemployment. Investment in these types of *informal education* schemes (over $3.8 billion between 1963 and 1986 in developing nations) are particularly difficult to optimize owing to the dramatic influence of economic policies within competitive markets. Success also requires a close match between labor market conditions and the kinds of training received.

Adult basic education and general efforts to teach *literacy* must also be considered as important sectors of education that generally operate in isolation from the world of primary and secondary education. The economic effects wrought by adult education are difficult to measure and vary considerably among nations owing to other factors. Adult basic education, for example, is fairly large even in highly industrialized nations such as the USA and may be a much larger sector than formal education in some lightly industrialized nations. Finally, we must also recognize at the individual level that there is much informal education that occurs as a person accumulates life experiences, reads, interacts with others, and thinks. The cumulative effects of these highly personal forms of education are unknown in their effects upon economic growth.

Models of economic growth that seek to document the impact of 'education' usually focus on one or at most two of the above subcategories and frequently employ different data sources. Not surprisingly, they sometimes come to dramatically different and frequently contradictory conclusions that arise due to the measures for economic growth adopted, the methodology employed, the scope of the inquiry, and the data sources consulted.

2. *The Continuum of Development*

Further complications arise when moving up and down the continuum of highly industrialized, information-intensive economies and minimally industrialized, information-modest economies (a growing North–South divide on the planet). Economic growth models generally focus on the capacity of a national economy and focus on measures such as global market share, employment in wage-related activities, income contributions (personal and corporate taxes) and redistribution schemes, and other factors that embody within them an implicit belief that these are the proper measures for 'growth' and 'progress' for all cultures, peoples, and nations. Movement 'up' the development continuum to greater industrialization and an intensification of the role of information within the economy are seen as valuable and essential for every nation. This view elevates economic considerations as defined principally by economists who live in particular types of economies into a global standard against which all nations are measured. It ignores noneconomic factors such as population growth (e.g., expansive growth is generally frowned upon by economists but may be viewed positively by noneconomists), political considerations (e.g., socialist regimes may view economic goals differently from the majority of economists who work within capitalist frameworks), sociocultural milieus (e.g., individual and group behaviors and class structures) and issues such as fertility, life expectancy, and child mortality.

Economists who have looked at some of these issues in relation to 'developing nations' routinely suspect that 'traditional growth accounting estimates are likely to underestimate the true contribution of education to economic growth and social welfare in general' (Woodhall 1994, p. 2646). Additional data sources not generally accounted for in approaches to data include the hidden economies revolving around the informal labor sector, black markets, and other exchanges of goods that are not quantified or quantifiable in many nations of the world. For example, it is estimated that only about half of the GNP in developing nations is comprised of the share of earnings from labor vs. a range of 60–70 percent in highly industrialized nations. Some economists have recommended that dualistic models be created to address adequately the many distinct differences between economies on the two extremes of the development spectrum. Others have suggested that really four distinct types of economies can now be found among the developing nations of the world: (a) newly industrialized nations in Asia, (b) relatively highly educated Latin American nations, (c) large continental economies such as India and China, and (d) marginal rural economies found in Asia, Africa, and Latin America.

A number of disparate findings have emerged when viewing the impact of education on earnings along the development continuum. Higher average levels of formal education are frequently associated with higher earnings in most highly *industrialized nations*, although notable exceptions are sometimes seen such as nations in Eastern Europe and Russia (Decker et al. 1997, Filmer and Pritchett 1999a, Baldi et al. 2000). Nations at the other end of the spectrum generally demonstrate negative associations with education at university level once educated labor is no longer scarce. This is probably due to the fact that the expansion of the educated *labor force* has far outstripped the supply of jobs requiring such high levels of formal education and this disequilibrium exacerbates civil unrest. Oversupply has led many economists to recommend that postsecondary education be scaled back in many developing nations and these resources be targeted to primary education interventions. Investments in the education of women (a key source of inequality in many nations) and in general education (as opposed to vocational education) yield returns equal to or better than their alternatives, although once again there are

notable exceptions based on differences in economic policies, social structures, and the distribution of education (López et al. 1998, Filmer and Pritchett 1999b).

3. Human Capital Theory

Economists classically view labor, land, and capital (human and otherwise) as the three principal inputs into production. Studies of the contribution of education to economic growth have been one line of investigation within the human capital dimension—often held in tension with inputs from the other two sectors in complex multidimensional models to tease out the residuals that can be 'explained' in a statistical sense by education as defined by the researcher. These 'growth accounting' models have generally shown a fairly modest effect of education on economic growth ranging from as low as 1 percent to highs of around 30 percent with a mean finding of around 10 percent since the 1970s. Perhaps not surprisingly, returns from primary-level education (basic literacy and numeracy) are generally found to be much higher than for other levels and the decline can be steep as one ascends the levels of formal education.

A working assumption that underlies virtually all human capital approaches going back to Nobel Prize winner Gary Becker (Becker et al. 1995) and others, is that people will invest at an individual level up to the point where they believe they are going to realize a full return on their investment. This is often driven by a belief that there are certain fundamental skills that one must acquire in order to compete or advance successfully in the labor market. (It is important to note that costs subsidized by the state or other entities rather than the individual, routinely result in most students choosing to go for a higher level of education or credential than they would elect if they had to bear the full cost themselves.) Schools serve as a powerful socializing influence both in reinforcing the value of education and in shaping the kinds of educational experiences and expectations that people hold. In many societies, schooling serves as a modernizing function whereas in others (e.g., strict Islamic societies), it may serve to restrict impulses frequently associated with modernization or open capitalist societies.

Formal education frequently acts as a filter or a screen that determines a threshold level that a job applicant must meet in order to be considered for an open position. One of the many reasons why there are such modest direct relationships between education and economic growth or even education and earning in a narrow sense is that market supply has a bearing on the level of education deemed minimally acceptable by an employer. In times of ample supply of applicants, employers often require a much higher level of education as a key screening tool. Adjusting education supply functions in the light of these changing demands in order to increase economic productivity is both conceptually and politically difficult and is generally discouraged. In part, this is because employers and policy makers recognize that education helps positively shape a variety of other aptitudes, dispositions, and motivation critical to work success (Woodhall 1994). It is important to note also that wages are affected by many other factors extrinsic to employees and employers, such as taxation, price controls, inflation, and state and local policies and regulations. For these reasons, the strong form of the screening hypothesis that argued that higher levels of education signals higher wages over one's entire career is no longer viewed as tenable. A weaker form of the *screening hypothesis* that views credentials and higher levels of formal education as a filter at the initial hiring stage with accompanying early wage differentials is generally accepted. Competing theories about the operation of labor markets have produced divergent explanations for why educational attainment and levels of lifetime earnings generally hold over time in many industrialized nations. It has been amply demonstrated that 'people with the same level of education do have different earnings depending on their race, gender, ethnicity, ability, and social background' (Cippollone 1994, p. 1655).

Human capital theory has been substantially modified from its early days when considering the impact of education upon the economy. Understanding of human capital has expanded exponentially beyond education *per se* to embrace issues such as age distributions within given nations, rurality vs. urbanity, and threshold factors to economic growth in some nations such as basic health, sanitation, housing, water supply, and nutrition. The modest claims for education's impact on economic growth advanced today are a reflection of both the complexities involved and what Theodore Shultz (1961) has called 'adjusting to economic disequilibrium.' The problems associated with the human capital approach have led to several competing theories about the relationship between education and economic development. One theory takes an institutional approach and argues that supply and demand in education is shaped by larger social behaviors within the society. It is these social conditions, rather than education *per se*, that is influential on economic growth. A second approach focuses on labor market segmentation and internal labor markets, which in turn influence both education and economic growth.

A comprehensive causative model for economic growth embracing all sectors of education has proven elusive. While the general premise of a link between education and economic growth appears sound, the complexities associated with explicating the exact nature of this premise challenge the best theories and economists.

See also: Adult Education and Training: Cognitive Aspects; Economic Growth: Measurement; Economic Growth: Theory; Education and Employment; Education and Income Distribution; Educational Institutions and Society; Educational Policy, Mechanisms for the Development of; Educational Productivity; Human Capital: Educational Aspects; Literacy Education; Vocational Education and Training

Bibliography

Adjibolosoo S B-S K 1999 *Rethinking Development Theory and Policy*. Praeger, New York

Baldi S, Khalaf G, Perie M, Sherman J, Snyder T D 2000 *International Education Indicators: A Time Series Perspective, 1985–1995*. National Center for Education Statistics, US Department of Education, Washington, DC

Becker G 1993 *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, 3rd edn. University of Chicago Press, Chicago

Becker G S, Schwartz P S, Febrero R (eds.) 1995 *The Essence of Becker*. Hoover Institution Press, Stanford, CA

Blaug A 1992 *The Economic Value of Education: Studies in the Economics of Education*. Edward Elgar, Lyme, NH

Boekema F, Morgar K, Bakkers S 2000 *Knowledge, Innovation and Economic Growth: The Theory and Practice of Learning Regions*. Edward Elgar, Lyme, NH

Carnoy M 1995 *International Encyclopedia of Economics of Education*, 2nd edn. Pergamon, New York

Cipollone P 1994 Education and earnings. In: Husén T, Postlethwaite T N (eds.) *The International Encyclopedia of Education*. Pergamon, New York Vol 3, pp. 1655–60

Decker P T, Rice J K, Moore M T, Rollefson M R 1997 *Education and the Economy: An Indicators Report*. US Department of Education, Washington, DC

Filmer D, Pritchett L 1999a The effect of household wealth on educational attainment: Evidence from 35 countries. *Population and Development Review* **25**: 85–120

Filmer D, Pritchett L 1999b Educational enrollment and attainment in India: Household wealth, gender, village and state effects. *Journal of Educational Planning and Administration* **13**: 135–63

Glewwe P 1999 *The Economics of School Quality Investments in Developing Countries: An Empirical Study of Ghana*. St. Martins Press, New York

López R, Thomas V, Wang Y 1998 *Addressing the Education Puzzle: The Distribution of Education and Economic Reform*. World Bank, Washington, DC

Mayer S E, Peterson P E 1999 *Earning and Learning: How Schools Matter*. Brookings Institution Press, Washington, DC

O'Connell P J 1999 *Adults in Training: An International Comparison of Continuing Education and Training*. Organization for Economic Cooperation and Development, Paris

Organization for Economic Cooperation and Development 2000 *Education at a Glance: OECD Indicators – Education and Skills*. Organization for Economic Cooperation and Development, Paris

Schultz T W 1961 Investment in human capital. *American Economic Review* **51**: 1–17

Woodhall M 1994 Human capital concepts. In: Husén T, Postlethwaite T N (eds.) *The International Encyclopedia of Education*. Pergamon, New York Vol 5, pp. 2643–7

D. W. Cheek

Education and Employment

Schooling is a social mechanism which, as a rule, dissociates the learner physically for a certain life period from the regular world of work and other life spheres. This is undertaken in order to prepare learners in a more rational manner for coping successfully with the diversity of work and other life tasks through explanations, rules, general reasoning strategies, etc. The more efficient the industrial society became in producing wealth, the more the education system and generated competencies contributing to the production of goods and services expanded. Regarding the world of work, education has:

(a) a qualifying (in the French and German connotation) function of fostering the cognitive and possibly affective and sensorimotor capabilities which might be useful to cope with job tasks;

(b) a status-distributive function: the level of 'educational attainment' determines to a certain extent the monetary resources and the social recognition which will be available to individuals in their subsequent life; for education is a powerful factor—and has increasingly become so over the nineteenth and twentieth century—in opening up or closing access to prestigious occupations and providing the means for professional achievement and thus to remuneration and socioeconomic status.

These basic functions are undisputed. However, it is generally assumed that education is bound to be imperfect in preparing for the world of work, partly because rational learning through dissociation from practice has its price in less direct preparation for occupational tasks and because education is expected to serve broader functions. Also, other factors contribute to professional success, e.g., sociobiographic background, genetically determined abilities, specific dynamics of credentials, processes of transfer from education to employment, and finally lifelong learning and personal development. Thus views vary as regards to the actual ways education is and ought to be shaped as well as the actual impacts learning has and ought to have on subsequent employment and work.

1. Imperfections and Limits of the Linkages of Education and Employment

Even if education was expected to be closely geared to the 'requirements' of the employment system and particularistic social factors were in play there are

obvious imperfections and uncertainties which make close linkages unlikely. Available research and past debates for instance pointed out the following issues:

(a) Occupational dynamics: The employment system is very dynamic in terms of changes of job tasks within given occupations, and most people have to expect occupational mobility in terms of employers or occupations. This challenges the view that one should prepare for a very specific bundle of job tasks.

(b) Indeterminate work tasks for highly qualified workforce: The higher the educational level required for a certain occupational area, the more difficult it is to identify the competencies required.

(c) Planning gap: There is an unavoidable time gap between the identification of new job assignments and the provision of respective competencies on the part of school leavers and graduates, because years are needed for the revision of curricula, and the actual 'production' of new graduates.

(d) Generalists' vs. specialists' paradigms: Views vary among experts regarding the extent to which education should be general or specialized according to areas of knowledge or occupations.

(e) Emphasis on vocational education vs. recurrent education: Finally, views differ similarly on the extent to which the growing role of lifelong learning might reduce the need for pre-career vocational or professional education and training, on the change of learning abilities over the life course, job requirements in different career stages, and the economic and social conditions for lifelong learning (cf. Schuller and Tuijnman 1999).

In all societies, controversies can be observed as regards the extent to which education is viewed as instrumental in securing the individual's and the society's economic success. There are also strong national differences with respect to four areas of values.

(a) An 'educational meritocracy,' i.e., a strong impact of vocational educational attainment on subsequent career, seems to be realized most and highly appreciated in Japan, whereas lifelong opportunities for counteracting those links are emphasized in the USA.

(b) A general or specific approach of education and training: Where a generalist's view prevails, for example in Anglo-Saxon countries and in Japan, specialized training tends to be viewed as 'narrow' both in terms of restraining professional flexibility and personality development. Where a specialist's view dominates, notably in France and to a certain extent in Germany, the acquisition of specific knowledge is considered as a process of exemplary in-depth learning which ensures substantial transfer to other areas of expertise, and is viewed as compatible with a broadly cultured personality.

(c) Concepts of employment and work-related identity: In France and Germany, employment and work are generally expected to be a major force for individual identity. This tradition has reinforced a high level of pride in skilled workers in Germany. In contrast, sense of affiliation to one's own employing organization tends to be viewed as a major source of identity and pride in Japan. In the Anglo-Saxon tradition, identity linked to work seems to be a phenomenon confined to high-level occupations, as the distinction underscores between (high-level) 'professions' and other 'vocational' areas.

(d) The role policies actually play and should play in shaping the relationships between education and employment: In the Soviet Union and affiliated countries, strong efforts were made from the 1950s through the 1980s to plan education according to the perceived demands of a planned economy. In the USA, faith in the self-regulatory forces is most widespread as far as educational preparation for employment and work in a market-driven economy is concerned. In European countries, different degrees of macroplanning and steering of education are considered essential in order to strike the balance between economic and other social and cultural rationales of education in a predominantly market-driven economy.

The different values are so pertinent for the actual links between education and employment in the respective societies that interpretations of research findings and discussions of their policies are most likely to reinforce national characteristics of education–employment relationships.

2. *Trends of Educational Attainment*

Expansion of education seems to have been a perennial trend since World War II. On average, about one third of the adult population in industrial societies had at least some upper secondary education in 1960; this proportion reached two-thirds in the late 1990s (OECD 2000).

From the late 1950s through the early 1970s, optimism prevailed that the economy requires a substantial increase of labor with advanced levels of education and that educational expansion is instrumental to economic growth. Comparisons focused on general and full-time schooling, thus neglecting vocational training. Accordingly, full-time enrollment rates of 17 year olds were 86 percent in the USA in 1970, about three-quarters in Canada and Japan, but only ranged in Western Europe from about 60 percent in Sweden to about 20 percent in Germany (OECD 1977). Average years of schooling, measured in 1975, ranged from about 11 years in the USA and the UK to less than six years in France, Spain, and Portugal.

The 1973 'oil crisis' was a turning point of the political mood, and concern about unemployment in general began to dominate over problems of employability of youth and potential mismatches between education and employment due to an over-

Table 1
Percentage of population from select countries having attained at least upper secondary education and university-level education, by age group 1998

	Upper Secondary Age Group				University-Level Age Group			
Country	25–34	35–44	45–54	55–64	25–34	35–44	45–54	55–64
USA	88	88	87	80	27	26	29	22
Norway	93	88	79	65	27	25	22	12
Japan	93	91	77	57	23	23	15	9
Germany	88	87	84	76	14	16	15	10
Czech Republic	92	88	84	74	7	5	5	3
UK	63	62	58	53	17	17	15	11
France	75	63	56	41	15	10	10	6
Philippines	56	53	35	35	26	26	18	18
Jordan	48	41	28	15	30	26	18	9
Spain	53	38	23	12	21	16	11	6
Argentina	36	29	21	15	11	10	7	4
Turkey	24	19	13	7	7	7	6	3
China	18	22	10	10	5	3	3	3
Zimbabwe	16	14	14	12	2	2	2	2
India	11	9	6	3	8	7	5	3

Source: OECD 2000, pp. 37–8.

supply of labor with higher or other advanced levels of education and training (OECD 1977). At that time, the OECD began to include part-time enrollment and apprenticeship training into their figures of upper secondary enrollment, thereby observing an average increase of the secondary education graduation quota from about 60 percent in industrial societies around 1970 to more than 80 percent in the late 1990s. Differences ranged from more than 90 percent in 1998 in New Zealand, Japan, and Germany to less than 60 percent in Portugal (OECD 2000).

Similarly, the proportion of the respective age group being awarded a university-level degree increased from less than 10 percent on average in industrial societies around 1970 to almost 20 percent at the end of the century. A comparison of the educational attainment (Table 1) shows that the differences between the age groups are small in the USA, where the major expansion of enrolment took place before the 1970s, and high in Spain, where educational expansion is most pronounced in recent decades.

3. *Income Differences by Education*

In all societies, educational attainment and income correlates. In European OECD member states, for example, those with upper secondary education earned on average in the 1990s about 1.3 times, those with nonuniversity tertiary education 1.5 times and those with university-level education about twice as much as persons without upper secondary education (see OECD 2000). Income differences according to educational attainment tend to be higher for women than for men, though they earn less than men. Differentials declined moderately on average over the last three to four decades of the twentieth century, though not consistently over time and across countries. Variations across countries cannot fully be explained by patterns of supply and demand; other social factors come into play which contribute to varying income inequality.

Calculations based on the human capital approach show that individual returns for higher education vary between countries from less than 5 percent to more than 20 percent. In some countries they surpass the interest rates of business capital whereas in others, they remain below that level.

4. *The 'Match' between Education and Employment*

The extent to which educational qualifications match the employment demands is a frequent topic of debate. One tries to establish—both for recent graduates and the total labor force—whether qualifications correspond to the occupational structure

(a) horizontally (links between subjects and occupational categories), or

(b) vertically (appropriateness of the level of education to the occupational status).

Altogether, concerns about a horizontal match are more pronounced in countries emphasizing specialization in education and employment than in countries considering education predominantly as a general preparation. They tend to be more pronounced as well regarding higher levels of education than lower ones: A substitution of a car mechanic by a baker seems to

be more acceptable than that of an engineer by a chemist. But more leeway for substitution exists than expected. For example, the fact that about half of the Germans having undergone an apprenticeship training are employed five years later in occupations not closely linked to the type of training does not tend to be considered as a major wastage, but rather as an indicator that in-depth specialized vocational training fosters substantial potentials for flexibility and transfer of skills.

More attention has been paid to problems linked to possible vertical mismatches. In the 1960s, concern was frequently voiced that countries with a low proportion of the population with advanced education might fall behind the others with respect to economic growth. From the 1970s through the 1990s, in contrast, the controversies on the agenda concerned whether a trend toward 'overeducation' or 'overqualification' could be observed. During the 1990s, a mix of concerns about overeducation in some areas and lack of competencies in other areas, e.g., qualifications required for new technologies, was prevalent in many countries.

'Overeducation' and in some cases 'undereducation' was analyzed with the help of various measures (cf. Hartog 2000, Paul et al. 2000):

(a) Educational and occupational statistics seem to suggest that the expansion of advanced levels of education clearly surpassed the demand of the employment system.

(b) Employers' expectations similarly tended to support the idea of substantial increase of 'overeducation' in the 1970s and 1980s, though they also point out lack of competencies in some domains.

Surveys of former students also show a higher frequency of education seemingly surpassing the level of job requirements than that of 'undereducation,' but they indicate that many graduates make use on the job of the higher level of competencies.

The progress of research notwithstanding, the issue of 'overeducation' seems to be perennial, because the tensions do not fade away between often steep occupational hierarchies and a relatively open access to the highest levels of education. However, concern tended to be voiced less dramatically in the 1990s than in the 1970s (Teichler 1999). This might be explained by widespread expections that a 'knowledge society' is likely to emerge. Also, research has shown that most persons seemingly overqualified acquired a position only slightly lower than they had strived for.

5. *Education, Unemployment, and 'Employability'*

Youth unemployment and various modes of precarious employment became a major concern in most industrialized societies since the mid-1970s and in most developing countries for long periods. The unemployment quota among youth is substantially higher than that of the total labor force in most countries. Also, in most industrial societies, the higher unemployment quotas are, the lower is the level of educational attainment. OECD calculations made in the late 1990s suggest that Europeans trained below upper secondary education have to expect four years of unemployment over their life course, those trained on upper secondary education somewhat more than two years, and those trained on tertiary education level somewhat more than one year.

This led to a growing attention to the education of persons not achieving advanced levels of education, called for example, 'the forgotten majority' or the 'forgotten half.' A discussion was triggered as to whether an expansionist education policy could strengthen the employability potentials of young people and whether certain types of education and training were more successful than others in preparing youth for the world of work.

Various models were suggested to classify modes of vocational education and training on secondary education level (cf. Tessaring 1998). During the late 1970s, the apprenticeship training model (Lutz 1994) became popular in international debates, because the youth unemployment rate in Germany and neighboring countries was not substantially higher than overall unemployment rates. During the 1980s, the pendulum of debate swung in favor of a general approach of vocational education combined with initial and further on-the-job training within the employment system, as customary in Japan (Dore and Sako 1989).

Efforts were also made to increase educational and training opportunities in countries previously characterized by a low enrollment rate at the upper secondary education level. Recently new mixes of education and work experience have emerged without any single model being regarded as clearly superior.

6. *Transition from Education to Employment*

In the 1990s, increased attention was paid to the processes of transition from education to employment. An OECD study defined transition as a period of up to 10 years when youth might be searching for a satisfying and well-paid job, orientation and efforts to supplement competencies might interact, and employment opportunities might differ from those for the older workforce (Werquin et al. 1997, OECD 1999).

Part of the respective research focuses on the job start as a major indicator of the relationships between education and employment. Other studies analyze the speed of transition and the frequency of unemployment and precarious employment over various years. A further type of analysis tries to examine whether the process of transition is facilitated by certain types of

prior education and training, e.g., a close 'coupling' of education and training to respective occupations, notably through the apprenticeship systems (cf. Shavit and Müller 1998), whereas others consider 'market' regulation, as observed notably in the USA, as most promising because it facilitates adjustments at any time in the career.

Various studies address factors other than educational achievement that might explain transition to employment and further career. 'Credentials,' for example, both certify and symbolically overemphasize educational achievements. Also, sociobiographic background comes into play not merely in terms of particularistic advantages in the transition processes, but also as in terms of social skills not taught through formal educational processes. Further, the terms of 'key qualifications' point out the importance of competencies in part fostered by education and training. Moreover, transition from education to employment also can be viewed as a moment under reduced meritocratic rule when young people can seek for chances through diligent search, smart tactics, and the demonstration of talents so far not rewarded in education. Finally, some analyses indicate moderate impacts of targeted measures of improving transition to employment, notably short training programs for 'at-risk youth' and occupational guidance and counseling as well as placement arrangements by public employment agencies, private recruitment, and placement firms, and by placement support on the part of educational institutions.

7. *Concluding Observations*

Education became increasingly an important determinant of employment and career in the nineteenth and twentieth centuries. By and large, long periods of vocational learning are rewarded by a higher status and more interesting and independent work. However, this is affected by labor market situations of demand and supply. Also, education becomes, in the process of expansion, more and more a prerequisite for career success, whereby other factors might gain weight with respect to the details of selection and allocation. Finally, the growing dynamics within the occupation system make the influence of vocational education less visible when life-long education grows.

International comparison has pointed out a striking variety of modes of education and training and of links between education and employment. The debates continue on the virtue of general vs. specific education, on hopes set in educational expansion vs. a closer match between education and the demands of the economy, on the advantages of a close coupling between training and work vs. open market regulations, etc. But international comparison also questions the beliefs widespread in the 1960s that one could identify the most successful, modern way of shaping the education system in a most promising way for serving economy, culture, and personality development. This did not necessarily lead to a dominance of relativistic approaches, but rather often stimulated the search for new mixes, e.g., in combining general education with work experience, opening up avenues for tertiary education while intensifying measures for educationally disadvantaged youth, improving education and concurrently measures that address the transition processes directly as well as 'employability' in other respects.

Even in a 'knowledge society,' one does not expect that a more or less pure educational meritocracy will emerge, because other factors might increasingly gain momentum. But education seems to remain the single most important determinator of career and other life chances.

See also: Adolescent Work and Unemployment; Education and Income Distribution; Educational Sociology

Bibliography

Carnoy M 1994 *Economics of Education: Research and Studies*. Pergamon Press, Oxford

Dore R P, Sako M 1989 *How the Japanese Learn to Work*. Routledge, London

Hartog J 2000 Human capital as an instrument of analysis for the economics of education. *European Journal of Education* **35**: 7–20

Lutz B 1994 The difficult recovery of 'professionalism.' In: *Apprenticeship: Which Way forward*? OECD, CEREQ, Paris, pp. 19–28

Nijhof W J, Streumer J N (eds.) *Key Qualifications in Work and Education*. Kluwver, Dordrecht, Germany

OECD 1977 *Entry of Young People into Working Life*. OECD, Paris

OECD (ed.) 1999 *Preparing Youth for the 21st Century*. OECD, Paris

OECD 2000 *Education at a Glance*. OECD, Paris

Paul J-J, Teichler U, van der Velden R (eds.) 2000 Higher education and graduate employment. *European Journal of Education* **35** (2) (special issue)

Psacharopoulos G (ed.) 1987 *Economics of Education: Research and Studies*, 1st edn. Pergamon, Oxford, UK

Shavit Y, Müller W (eds.) 1998 *From School to Work: A Comparative Study of Educational Qualifications and Occupational Destinations*. Clarendon Press, Oxford, UK

Schuller T, Tuijnman A (eds.) 1999 *Lifelong Learning: Policy and Research*. Portland Press, London

Teichler U 1999 Higher education policy and the world of work: Changing conditions and challenges. *Higher Education Policy* **12**: 285–312

Tessaring M 1998 *Training for a Changing Europe: A Report on the Current Vocational and Training Research in Europe*. CEDEFOP, Thessaloniki, Greece

Werquin R, Breen R, Plannas J (eds.) 1997 *Youth Transitions in Europe*: *Theories and Evidence*. CEREQ, Marseilles, France

U. Teichler

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Education and Gender: Historical Perspectives

Formal systems of education in the modern world have a two-way relationship with societal gender inequality: they reflect prevailing ideologies about gender roles and, under some circumstances, have ameliorated gender inequality in the larger society. This article traces major historical changes in gender inequality in education, primarily focusing on the United States from the period immediately prior to the creation of formal state-supported systems of schooling to the present. Changes in gender inequality in access to education have been dramatic: once largely denied formal schooling, in many parts of the world girls now have equal or better access to secondary and, sometimes even, higher education than boys. Former gender differences in curricular opportunities have been narrowed significantly, especially in the West, and gender bias in texts, social interaction, and school-based activities has been dramatically reduced as well. Gender inequalities and bias in education are institutionalized more deeply in parts of the world where women still are restricted largely to the private sphere.

1. Access

1.1 Education Before State-supported Systems

The story of gender and education in the West has to start in the years before formal state-supported systems of education were common. In the seventeenth- and eighteenth-century United States, for example, schooling was provided through a patchwork of arrangements and formal schooling opportunities were largely closed to girls. Some girls did receive instruction in town and private schools: girls who were 'smuggled in' after the boys had gone home for the day (Tyack and Hansot 1990, Sadker and Sadker 1994). Early 'dame schools'—schools privately subsidized by parents—did admit girls, and it was from the dame schools that public school systems later evolved. The exclusion of girls from early formal schooling in the United States reflected the prevailing ideology that women were destined for domesticity; public life was reserved for males.

1.2 Early State-supported Education

State-supported systems of education emerged in the United States by the first third of the nineteenth century (the last third of the century in the American South). One distinctive feature of American state-supported schooling, or 'common schooling,' as it was known, is that it was coeducational. Boys and girls were educated together in the same school buildings—often one-room schools in rural areas providing what we would today regard as an elementary education (Tyack and Hansot 1990). The early institutionalization of coeducation put limits on the forms that gender inequality in education could take and gave girls roughly the same access to schooling as boys. Coeducation was preferred in the American case

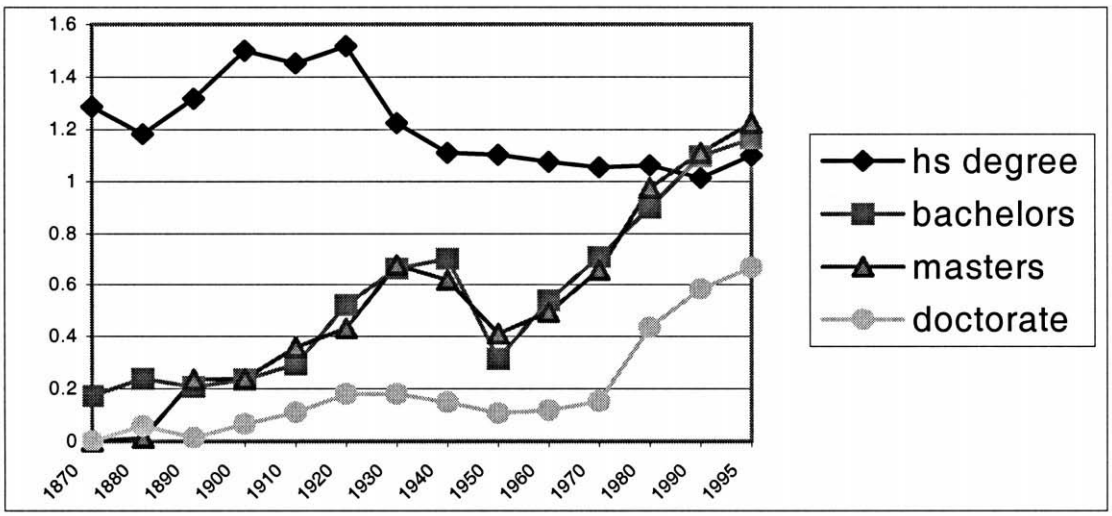

Figure 1
Female-to-male ratio of degrees awarded in the United States, 1870–1995 (sources: US Bureau of the Census 1975, 1998; US Department of Education 1998)

mostly because of the economies of scale it allowed rather than from an ideological commitment to educating boys and girls side-by-side. Single-sex school buildings were more common in Western Europe and Britain during this formative period of state-supported schooling in the nineteenth century, and the United States remained somewhat distinctive by allowing girls similar overall access to state-supported schools.

1.3 School Expansion, Access, and Attainment

Cities and townships started to establish secondary schools around the mid-nineteenth century, although secondary schools enrolled only a tiny fraction of eligible youth before the early twentieth century. From the outset of the establishment of secondary schools, girls actually outnumbered boys. The federal government started to keep statistics on the number of high school graduates in 1870; in that year there were 7,000 male graduates compared to 9,000 female graduates (US Bureau of the Census 1975). From the nineteenth century to the present, American women have maintained their advantage over men with respect to likelihood of high school graduation (see Fig. 1; a ratio greater than one indicates a female advantage).

Gender distinctions were embedded more institutionally in nineteenth-century higher education than in elementary and secondary education, perhaps because a large proportion of colleges and universities were private and higher education remained the province of the elite until after World War II. Nineteenth-century American women had many fewer opportunities to attend college than their male peers had; Oberlin became the first American college to admit women but it was one of the only university options American women had until after the Civil War (Sadker and Sadker 1994). The gender inequality in access to higher education was reflected in stark gender differences in degrees obtained: in 1870, male bachelor's recipients outnumbered female recipients by almost 6 to 1. The gap closed gradually; by 1920, the male advantage in bachelor's degrees conferred was slightly less than 2 to 1 and in 1982 women surpassed men in the number of bachelor's degrees conferred for the first time—an advantage that has continued to widen (see Fig.1). From 1890 on, trends in gender inequality in receipt of masters degrees basically mirrored the trends in baccalaureates. Inequalities at the doctoral level, however, took much longer to close to any significant degree and at this level the male advantage is still significant: in 1995, men were 1.5 times more likely to get doctoral degrees than women.

1.4 The United States in Comparative Perspective

The relatively early timing of rough gender parity, or even female advantage, in access to secondary and baccalaureate-level education in the United States is unusual, as is the relatively early expansion of secondary and higher education to non-elites, and there are some countries in which women still experience significant disadvantages in access to education. In Britain and other English-speaking democracies, most of Europe, and most of Latin America, gender parity at the secondary and baccalaureate levels was achieved far more recently than in the United States. Although gender inequality in access to education has declined over time from its once substantial levels in most Asian countries, including the rich countries of Japan and Korea, gender inequality remains greater there than in any European state. Finally, large differences in women's and men's access to education still exist in much of Africa and the Middle East, with Muslim countries exhibiting the most gender inequality in access (Brint 1998, Stromquist 1995). The overall pattern is that societies in which women have been most fully incorporated into the public sphere exhibit the greatest gender equality in access to education.

2. Gender, Curriculum, and Social Climate

Being in the same schoolhouse, or even the same room, with boys did not and still does not mean that girls received the same education as boys. Unlike access, in the case of curriculum and social climate the picture is not one of increasingly similar experiences for boys and girls; looking back at the last 200 years, gender inequality increased before it decreased.

2.1 The Nineteenth Century and 'Identical Coeducation'

By the mid-nineteenth century, when the common school was established firmly in American society, the available evidence suggests that boys and girls were by and large exposed to the same kind of curricular materials and were expected to gain similar things from schooling: basic literacy and numeracy, and appropriate political and moral values (Clifford 1982). By 1850, the relatively sizable male advantage in literacy that had existed at the time of the Revolution largely had been erased. Americans' reasons for supporting basic literacy and numeracy differed by gender: for boys for their future roles in the public sphere and for girls for their future roles as mothers of the republic (Tyack and Hansot 1990). One possible reason why a 'common' education generally was provided to boys and girls throughout the nineteenth century is that the ideology and practice of separate spheres was established so firmly in American society that educating boys and girls together for purposes of basic literacy and numeracy hardly challenged it. And the adult social roles depicted in their curricular materials, particularly the ubiquitous McGuffey

Readers, clearly reflected a world of separate spheres (Tyack and Hansot 1990). Further, it is likely that noninstructional activities, such as chores, were rigidly differentiated by gender.

2.2 The Progressive Era and Gender Differentiation

Throughout the late nineteenth century, attendance rates at the high-school level were low and the curriculum was almost entirely academic. High schools in some cities were single-sex, but as was the case for common schools the most frequent pattern was to have coeducational high schools. Available data suggest that girls and boys took quite similar courses (Latimer 1958). All that changed with the rapid expansion of the high school in the early twentieth century and its corresponding transformation from a primarily academic, or classical, curriculum to a vocational curriculum. Vocational education was proposed as a means to prepare students for adult life; in the case of gender, it meant preparing boys and girls for starkly different expected adult social roles. Progressive-era educational reformers increasingly called for differentiated education for boys and girls (Cremin 1961).

Instruction in the core academic subjects remained largely coeducational. But the core academic subjects occupied a decreasing portion of most students' time as the twentieth century unfolded. Gender-differentiated vocational education became much better established in public high schools. The new curriculum in home economics, for example, was targeted to girls. Industrial education, especially those courses providing 'manual training,' was almost entirely for boys. Girls were the ones who flocked to new courses in commercial education. And boys became more likely to take advanced elective academic courses, especially in math and science (Tyack and Hansot 1990, Clifford 1982). Further, physical education and the increasingly diverse opportunities for extracurricular activities, including sports, were differentiated by gender sharply. These transformations set the institutionalized gender parameters of American education until the 1970s, when the more extreme forms of gender differentiation lost their social legitimacy.

2.3 The 'Discovery' of Sexism in Schools

By the late 1960s and early 1970s, social science scholarship began to take a critical look at education and social inequality; this early work, however, focused primarily on class (see, e.g., Bowles and Gintis 1976). Gender bias in the formal knowledge conveyed by schools and in taken-for-granted social interactions in schools and classrooms was not 'discovered' until slightly later, when feminist scholars and activists turned their attention to classrooms, curriculum, and texts (for early examples, see Frazier and Sadker 1973, Stacey et al. 1974). The forms of gender bias and gendered social interactions identified in these works were not new social phenomena; what was new, however, was the definition of them as a social problem to be rectified. Note that the feminist critique of sexism in schools could not have been mounted successfully until a time, such as the 1970s, when a more general, society-wide ideology of equality between women and men was gaining legitimacy.

Early critics identified a number of ways in which American schools commonly exhibited bias against girls. As summarized by Tyack and Hansot (1990), the most important charges included: biased textbooks, favoritism of boys in classroom interactions, sex-typed counseling of students into courses and careers, a paucity of athletic opportunities for girls, the overrepresentation of males in high-level math and science courses, and male domination of educational administration. The cause of eliminating sex bias in schools got a boost in 1972 when the federal government passed Title IX of the 1972 Education Amendments, which prohibited sex discrimination in any educational program receiving federal funds. (In the United Kingdom, the 1975 Sex Discrimination Act similarly boosted the cause of gender equality in education (Arnot et al. 1999).) Title IX enforcement was slow and uneven, but the act itself, coupled with other antidiscrimination measures passed in the 1960s and 1970s, helped promote a social climate that made gender bias socially unacceptable and provided leverage for local activists and educators to implement change in schools. The major credit for the transformation in gender relations in schools, however, has to go to the rapid changes that took place in the family and the workplace.

It is difficult to evaluate the degree to which gender bias has been eliminated in the thousands of schools in a decentralized system such as that in the United States, but the most progress has probably been made with respect to biased textbooks, course and career counseling, and athletic opportunities. These forms of bias were institutionalized more formally and easier to change via centralized authority. A recent assessment of the schools' progress toward elimination of gender bias reports that girls are no longer underrepresented in some forms of science (e.g., biology) but remain underrepresented in courses in advanced math, computer science, physics, and calculus; the gender gap in sports participation narrowed in the 1990s but girls are still much less likely to participate in a team sport; and boys are more at risk of dropping out or repeating a grade than girls (American Association of University Women 1999).

Gender differences in vocational preference testing was eliminated long ago, and schools have no doubt made much progress toward counseling similar boys and girls toward similar future plans (Tyack and Hansot 1990). Although more difficult to gauge, recent

studies suggest that boys still get more attention from teachers and are more likely to be called on in class—forms of microinteraction that are not only difficult to change by policy decree but also are often unnoticed by those engaging in the practices (Sadker and Sadker 1994). In some important respects, however, the gender gap has been inverted. In the US and the UK, among other places, girls at the secondary level now outperform boys. In Britain, interestingly, this has caused not a public celebration over girls' success but a moral panic over boys' low achievement (Arnot et al. 1999).

3. Summary

Although in the West important gender differences in educationally-based opportunity structures remain, the institutionalized forms of gender bias that prevailed in the past have been greatly reduced. Nonetheless, those changes have not been sufficient to eliminate the significant gender differences in educational experiences or outcomes that are produced by daily microinteractions among students and between students and adults (Brint 1998). Further, some gender differentiation is the result of active choices made by boys and girls themselves that follow from gender socialization and, at the same time, are reinforced by the continuing sexism and gender differentiation common in other spheres of life, including the family and the workplace. Although at present significant forms of gender inequality and bias remain in schools, especially at the most selective levels, education is, at the same time, the most gender-neutral social institution in the contemporary Western world and in some important respects girls and women now outperform boys and men in selected Western countries. Educational gender inequality and bias remain more institutionalized in societies in which women are more restricted to the private sphere.

See also: Coeducation and Single-sex Schooling; Education (Higher) and Gender; Education (Primary and Secondary Schools) and Gender; Educational Philosophy: Contemporary Views; Educational Philosophy: Historical Perspectives; Gender and School Learning: Mathematics and Science; Gender and Technology; Gender-related Development; Women's Suffrage

Bibliography

American Association of University Women 1999 *Gender Gaps: Where Schools Still Fail Our Children*. Marlowe, New York

Arnot M, David M, Weiner G 1999 *Closing the Gender Gap*. Polity Press, Cambridge, UK

Bowles S, Gintis H 1976 *Schooling in Capitalist America*. Basic Books, New York

Brint S 1998 *Schools and Societies*. Pine Forge Press, Thousand Oaks, CA

Clifford G J 1982 Marry, stitch, die, or do worse: Educating women for work. In: Kantor H, Tyack D B (eds.) *Work, Youth, and Schooling: Historical Perspectives on Vocationalism in American Education*. Stanford University Press, Stanford, CA

Cremin L A 1961 *The Transformation of the School: Progressivism in American Education – 1876–1957*, 1st edn. Knopf, New York

Frazier N, Sadker M 1973 *Sexism in School and Society*. Harper & Row, New York

Latimer J F 1958 *What's Happened to Our High Schools?* Public Affairs Press, Washington, DC

Sadker M, Sadker D 1994 *Failing at Fairness: How America's Schools Cheat Girls*. C. Scribner's, New York

Stacey J, Bereaud S, Daniels J (eds.) 1974 *And Jill Came Tumbling After: Sexism in American Education*. Dell Publishing, New York

Stromquist N P 1995 Romancing the state: Gender and power in education. *Comparative Education Review* **39**: 423–54

Tyack D B, Hansot E 1990 *Learning Together: A History of Coeducation in American Schools*. Yale University Press, New Haven, CT

US Bureau of the Census 1975 *Historical Statistics of the United States – Colonial Times to 1970*, Bicentennial edn. US Department of Commerce, Washington, DC

US Bureau of the Census 1998 *Statistical Abstract of the United States*. US Department of Commerce, Washington, DC

US Department of Education 1998 *Digest of Educational Statistics*. Department of Commerce, Washington, DC

P. B. Walters

Education and Income Distribution

It is widely believed that income is differentiated by education, although the degree of the difference varies from country to country, from industry to industry, from occupation to occupation, and from individual to individual. Education contributes to raising an individual's productivity and thus to increasing his or her earnings capacity provided that workers' wages are paid according to their productivity. This is called the marginal productivity theory in economics.

We can point out five reasons why an individual productivity is raised by education:

(a) Individuals not only acquire skills through formal education but also through job training. The human capital theory is common in this approach where an individual skill in production process and thus earning capacity requires formal (i.e., schooling) and informal training.

(b) Why does formal schooling raise productivity? The answer is that it imparts fundamental skills and technologies such as reading, writing, mathematical calculation, physics, chemistry, engineering, etc.

(c) Educated people can use other resources skillfully such as machines and new technologies. Also, they can decide quickly whether these resources and information are useful in raising productivity or lowering cost.

(d) Education or schooling can socialize people into functioning effectively in society and economy. For example, schooling teaches people to be motivated, to be cooperative and patient, and to take roles in leadership, etc. This idea does not explain the increase in productivity through skill acquisition or knowledge of technique, but emphasizes mental and social aspects, which are likely to raise a group's productivity. In other words, it is useful in modern society where a team production is common, and concerned with an organizational explanation.

(e) The human capital theory emphasizes the importance of both formal schooling and training in the production process. The concept of trainability suggests that an individual who received formal schooling is ready to receive training because he or she can learn from training quickly, and adopt it efficiently in the production process. At the same time, it is likely that a company can provide educated employees with training at cheaper cost because of the above properties.

Education certainly contributes to higher productivity. However, empirical findings regarding points (a)–(e) are mixed. This is, in particular, true when we contrast between the developing countries and the developed countries. The difference in farmers' productivity in the developing countries can be verified by the degree of literacy rate and/or the enrolment rate in primary or secondary schools. However, nonagricultural jobs in the developed countries do not support the positive correlation between education and productivity. The reason is that more formal education raises the possibility that educated people may take jobs where productivity is higher. In other words, it is not easy to identify whether higher productivity in certain nonagricultural jobs can be explained by their higher educational attainment or that educated people take jobs where productivity is already higher.

There are similar problems in the relationship between education and income. Although it is true that education certainly raises an individual's earning capacity, its empirical verifications are very mixed. At the same time, it is possible that some reasons other than the fact that education raises earning capacity may be responsible for the positive correlation between education and income. The present entry reviews these two issues fairly extensively, namely empirical effects of education on income, and possible other causes to explain the positive correlation between education and income.

Before proceeding to these issues, it is necessary to make a distinction between income and earning. Income consists of earnings (labor incomes) and nonlabor incomes such as interests, rents, pensions, etc. 'Earnings' is a better terminology when we are concerned with the effect of education on income because earnings are direct rewards for working activities. However, since income is used throughout this entry note that here income normally implies (labor) earning.

The human capital income function is the most popular form of empirical studies, which was developed by Mincer (1974), and its form was applied to many countries. The fundamental form of the Mincer type of human capital function (i.e., log earnings of individual (i) is written as follows:

$$InY_i = a + bS_i + cT_i + dT_i^2 + u_i$$

where Y_i is annual income, S_i is years of formal education, T_i is the number of years an individual has worked since the end of formal schooling, and u_i is an error term. T represents potential labor market experience, which is supposed to indicate the influence of job training and of work experience. The reason why the squared form of T is introduced is that the influence of job training and of work experience has a quadratic form.

Empirical estimations based on this formula or its variations have been fairly successful in many countries. The implications and interpretations of this formula may be proposed in the following ways. First, if the estimated coefficient of b were positive with statistical significance, it would imply that the effect of education on income is positive. At the same time, the estimated magnitude of b would indicate the rate of return to education. Based on many empirical studies it is known that the estimated coefficient of b is positive and statistically significant, and thus that education certainly raises an individual earning capacity. The estimated rates of return to education are therefore different between countries, industries, and educational levels. Thus, it is risky to provide any conclusions about them. It is, nevertheless, possible to proclaim that they would be ranged between 0.05 and 0.25.

Second, the estimated R^2, which shows the degree of fit in statistical estimations based on this formula, indicates the degree to which the model is successful in estimating the variance of income. More concretely, it shows the degree to which education is able to explain the income inequality (i.e., variation of income). A large number of empirical studies suggest that about 10–40 percent of the variation in observed incomes can be explained by the model with a linear schooling term and a squared form in labor market experience.

Third, the relative importance of the schooling term and of the labor market experience term is different between countries, industries, occupations, etc. In other words, the influence of formal education is stronger than that of experience in one country, one industry, or one occupation in the determination of income, while the latter is stronger than the former in another country, another industry, or another occupation. It is nearly impossible to obtain a general feature about the relative importance of formal schooling versus labor market experience in the

determination of income. It is difficult to determine which variable is more important to raise an individual earning capacity between formal schooling versus job training (or learning by doing).

There have been several serious modifications or alternative ideas for the original Becker (1964) formulation of education and the Mincer-type human capital earnings function both theoretically and empirically. Some of them are explained here.

First, one modification was presented by Card (1995) who extended the idea of Becker. Becker (1964) originally considered a model in which an individual finishes his or her formal schooling before entering the labor market. In reality, an individual goes back to formal schooling either full-time or part-time after entering the labor market, and then goes back to work after completely their education. In other words, an individual moves back and forth between full-time or part-time schooling and part-time or full-time work.

Card (1995, 1999) developed a model, which incorporates the movements of an individual in and out of the labor force. Concretely, Card was concerned with individual heterogeneity in optimal schooling choice. There are two sources which affect an individual choice, namely (a) differences in the costs of (or taste for) schooling and (b) differences in the economic benefits of schooling.

In addition, there is an issue of a specification bias in the effect of education on income. More concretely, ability (innate ability) of an individual may affect the determination of income, and an exclusion of ability from an earnings function is likely to provide a biased estimation of the effect of education on earnings.

A more formal analysis gives the following result. Ignoring other variables, a simple earnings function can be written,

$$Y = \alpha + \beta S + \gamma A + u$$

where Y is income, S is education, and A is a measure of ability. When we ignore ability, we obtain a biased estimator of β as follows,

$$Eb_{Ys} = \beta + \gamma b_{AS} = \beta + \gamma \operatorname{cov}(AS)/\operatorname{var} A$$

where the return to education is estimated with a bias. Thus, it is necessary to include A when ability has an independent positive effect on earnings, and the relationship between the excluded ability and included schooling variable is positive.

The classical study by Griliches (1977) concluded that the ability bias caused by the excluded ability was minor. There is a consensus not only based on the study by Griliches' but also the subsequent studies that the exclusion of ability is not a serious problem. However, there remain several technical problems regarding the effect of ability on earnings. First, even if a popular variable such as IQ is included, there is a question about whether this indicates a proper measure of ability: a professional baseball player or a professional singer who receives a very high income does not necessarily do so because of a high IQ. Second, even if we assume that IQ is a relevant measure, it includes considerable measurement errors.

Another variable, which attracted many researchers in this field, is family background. It is observed that children's schooling achievements are highly correlated with the characteristics of their parents' status such as father's education, mother's education, parents' income, etc. This is called intergenerational transfer of family status, and some specialists such as Bowles and Gintus (1976) proposed that the effect of education on income must be understood. The positive correlation between education and income is superficial because the effect of family background on income is more intrinsic than that of education in the presence of the predetermined effect of family background on children's education.

A recursive type of an econometric model, which explains the causal relationships from family background to children's education, and from children's education to income, is capable of verifying this property. Griliches and Mason (1972) for the USA, Tachibanaki (1980) for France, Tachibanaki (1996) for Japan and Tachibanaki (1997) for the UK empirically support that the causal chain from family background to children's education and income is valid in the developed countries.

Econometricians and statisticians are enthusiastic about various estimation methods to identify the true effect of individual ability, family background, and measurement errors in income and educational attainment. They often use sibling and twin-data, offspring data, etc. In order to consolidate the analysis it is necessary to apply more sophisticated estimation methods to these data. Since their methods and techniques are beyond the scope of this entry, they are not described here (see Solon 1999, Card 1999).

See also: Educational Productivity; Human Capital: Educational Aspects; Income Distribution; Income Distribution: Demographic Aspects; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values

Bibliography

Becker G S 1964 *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. Columbia University Press, Columbia, NY

Bowles S, Gintus H 1976 *Schooling in Capitalist America*. Basic Books, New York

Card D 1995 Earnings, schooling and ability revisited. In: Polachek S (ed.) *Research in Labor Economics*. JAI Press, Stanford, CT, Vol. 14, pp. 23–48

Card D 1999 The causal effect of education on earnings. In: Ashenfelter O C, Card D (eds.) *Handbook of Labor Economics*. Elsevier Science, Amsterdam, Vol. 3A, pp. 1801–64

Griliches Z 1977 Estimating the returns to schooling: Some econometric problems. *Econometrica* **45**: 1–22

Griliches Z, Mason W 1972 Education, income and ability. *Journal of Political Economy* **82**: 74–103
Mincer J 1974 *Schooling, Experience and Earnings*. NBER Columbia University Press, New York
Solon G 1999 Intergenerational mobility in the labor market. In: Ashenfelter O C, Card D (eds.) *Handbook of Labor Economics*. Elsevier Science, Amsterdam, Vol. 3A, pp. 1761–800
Tachibanaki T 1980 Education, occupation and earnings: A recursive approach for France. *European Economic Review* **13**: 103–22
Tachibanaki T 1996 *Wage Determination and Distribution in Japan*. Clarendon Press, Oxford, UK
Tachibanaki T 1997 United Kingdom. In: Tachibanaki T (ed.) *Wage Differentials: An International Comparison*. Macmillan Press, London, pp. 210–38

T. Tachibanaki

Education and Learning: Lifespan Perspectives

For many people the likelihood of being able to retain the same job, or even the same profession, for the whole of one's working life is becoming increasingly remote. In these circumstances the idea of learning and education continuing to take place throughout life is a necessity, not simply a desirable goal.

There are and always will be good practical reasons for paying particular importance to learning and training during a young person's early years. The need to equip a person with the skills and capabilities that make it possible to have an independent life, and deal with the demands and responsibilities of a job sufficiently well to be employable, is most pressing in the early stages of life. It is understandable that once someone is capable of making a living and functioning as a mature adult, further learning and training may receive less priority. Even so, the chances are that at various times a man or woman will have to face new challenges and adapt to changing circumstances. For that reason it is highly desirable that individuals are encouraged to be aware that they may repeatedly be required to engage in learning and studying activities, throughout their whole lives.

1. *Do There Exist Fixed Periods for Learning?*

Because it is so important for adults to maintain their capacity to learn, it may seem odd that some people continue to regard learning and education as being activities that take place exclusively in the preadult stages of life. However, there are at least two reasons why that association continues to be made, if erroneously, even today. First, many people have been led to believe that a person's mental capacities are relatively fixed, and unchangeable. Second, it is widely assumed that there are certain critical periods for learning, and that these occur in early childhood. Both these assertions imply that learning and educational experiences that take place after a person has reached adulthood have at best limited effectiveness. Therefore, it is inferred, childhood is the proper time for learning.

In reality, neither the belief that mental abilities are fixed nor the idea that there are critical periods for learning is at all firmly founded on factual evidence. Although the view that every person possesses inherent mental abilities that cannot be changed is still widely shared (see, e.g., Herrnstein and Murray 1994), it is challenged by evidence from a variety of sources (Ceci 1996, Howe 1997). Proof that mental abilities can change, often very considerably, has been provided by several kinds of research findings. First, for instance, results obtained in adoption studies demonstrate that children adopted at birth into nurturing families tend to be more intelligent that their biological relatives (Schiff et al. 1982, Schiff and Lewontin 1986). That finding would not occur if it were true that people inherit a fixed intelligence. Second, the results of a number of intervention studies that were designed to provide compensatory educational experiences for young children reared in disadvantaged early environments, provide further evidence that substantial changes can occur in a young person's mental capacities (Ramey et al. 1984, Zigler and Muenchow 1992). That finding too would be an impossible one if it were really true that each person has fixed inherent mental abilities. Third, further evidence of the malleability of people's mental capacities has been obtained from studies in which average performance levels at intelligence tests is compared between different generations (Flynn 1987, Flynn 1991). Once again, the findings provide evidence of substantial rises, challenging the belief that a person's mental ability is fixed at an early age. Finally, findings obtained from a variety of studies assessing the influence of factors such as schooling provide additional proof that a person's mental capacities are indeed alterable rather than being fixed (Baltes and Reinert 1969, Cahan and Cohen 1989).

The notion of critical periods, which is the other phenomenon that seemed to provide support for the view that learning and training experiences are only fully effective when they take place early in life, has also been sharply criticized. In the case of humans, it does not stand up to scientific scrutiny (Bruer 1999, Gopnik et al. 1999). It is true that in certain nonhuman species there are restricted periods, at an early stage of life, during which certain forms of learning have to occur if they are ever to take place at all. In humans, however, that is broadly not the case, with a few minor exceptions. Generally speaking, the timing of learning is not so crucial.

In humans, the only circumstances in which delaying learning has disastrous consequences for a child are ones in which there has been massive deprivation. For example, language development can suffer permanently when a severely neglected child is deprived of any opportunities to experience language, and perceptual development may be disrupted permanently as a result of undetected pathological conditions that affect vision (Bruer 1999). But these very unusual and extreme critical periods in humans are largely nonexistent.

That is not to say that it does not matter at all when learning takes place. Some forms of learning are achieved more easily or effectively if they take place early in life. For instance, individuals who acquire a second language in childhood may learn to speak it without an accent, whereas people who do not commence learning the language until adulthood may never learn to speak without a detectable accent (Newport 1990). However, the fact that some forms of learning take place more readily or more efficiently earlier rather than later is no justification for believing in the existence of rigid critical periods.

The vast majority of skills and capabilities that typically are learned by children and young people can also be learned by adults: in healthy nonelderly individuals substantial age-related decreases in learning capacity are unusual. Consequently, there are few if any impassable barriers to learning across the whole life-span. However, adult learners may encounter various practical obstacles. For instance, the kinds of learning arrangements and learning contexts that are appropriate for children and young people may be less appropriate for older adults. In order to maximize adult learning opportunities it may be necessary to devise arrangements that are more flexible than the school-based schedules that have been designed for children tend to be.

2. *The Need for Flexibility*

One of the practical reasons why it is necessary to provide especially flexible arrangements for encouraging adult learning is that, compared with children, older people have more responsibilities, more demands upon their time, and more complicated lives. Although a man or woman with a job and a family may be just as capable of learning as a younger person, and equally keen to learn, in all likelihood it will be necessary for learning activities to be fitted in with all kinds of other activities. In children, it is realistic to assume that whole days and weeks may be almost exclusively devoted to learning, but in most adults that desirable state of affairs will only rarely occur. Partly for that reasons, the sheer number of hours that an adult can devote to learning will be restricted, in comparison with the amount of time that a younger person is likely to have available.

The fact that the sheer amount of time that an older learner can devote to studying may be restricted does not inevitably mean that the older person will make less progress. Even when, as at school or in a full-time college or university course, the whole day is nominally set aside for training and learning purposes, it does not always follow that young students will be engaged productively in learning activities throughout the day. For most young people, there is a limit to the extent that they can continue concentrating on learning tasks over lengthy periods of time (Csikszentmihalyi et al. 1993). Perhaps as a consequence of boredom or fatigue, or inability to concentrate, a school pupil may spend much of the time relatively ineffectively, going through the motions of studying but not giving learning tasks the sustained concentration that can be achieved by a more committed individual.

For these reasons, highly motivated adults who are especially committed to their studies and capable of sustaining their concentration on studying tasks may achieve more in a given amount of time than a younger person does. That is most likely to be the case if the training and studying arrangements are flexible enough to enable the adult to engage in study activities at the same time as meeting their other. A full-time course will be a viable option for some adults but not for others. Part-time courses in further education will be preferable for many adults.

Some older students will favor the especially flexible kinds of arrangements for learning that are offered by institutions such as Britain's Open University and also by comparable institutions in many other countries. Here the actual timing and scheduling of learning and studying sessions is largely decided by the student. For middle-aged and older students, there are also some 'Third Age' institutions, which typically combine an emphasis on flexibility with a lack of emphasis on factors such as competition and assessment. These can discourage older learners who may perceive them as irrelevant or aversive. For other learners, however, the extra flexibility on nontraditional courses is only achieved at a cost, because the individuals have to take more responsibility for structuring their time and also have to make decisions concerning how and when to engage in study activities. Broadly speaking, the more flexible the arrangements, the greater the demands that are made upon the person's maturity and independence as a learner.

3. *Problems Faced by Older Learners*

In comparison with school learners, older people will usually benefit from the major advantage of being able to decide on the particular course of study to be followed, rather than being expected to study at a particular subject simply because that is what the school requires. Broadly speaking, older students are

likely to be highly motivated to learn. On the other hand, there are certain problems that are more likely to be confronted by older than younger learners. The added responsibilities of an older person's day-to-day life may not only constrain the hours available for study activities but may also produce distractions at times that are nominally free for learning. Older people worry about their jobs and their relationships. Parents, in particular, are likely to be concerned about their children. Responsibilities like these may make it difficult for an adult learner, compared with a younger person, to concentrate fully on learning. Of course, children at school have anxieties, too, but the fact that older learners are likely to be responsible for other people as well as themselves may create distractions that are especially hard to ignore.

Another difficulty that is more prone to be experienced by older learners stems from the frustration that can accompany individuals' awareness of the slowness of their progress. For example, compared with a young child who is learning to play the piano and is perfectly content to keep practicing at the simple tunes that music teachers set for their child pupils, an adult beginner may have different expectations of what a learner ought to be able to achieve. An adult beginner at the piano may be anxious to learn to play the kind of music that they enjoy listening to. For older learners it may be frustrating to discover that it will be some years before they will reach a stage at which they can begin to play that music. They may be unhappy to perceive that in order to become able to do so it will be necessary to spend numerous hours painstakingly attempting to master easier musical compositions. These may provide few of the aesthetic satisfactions associated with more advanced music. For children, on the other hand, their actual rate of progress may be equally slow, but lacking either a knowledge of more advanced music or an adult expectation of the amount of progress that ought to be made, for the younger performer there may be no comparable frustrations or disappointments.

Similar frustrations may confront adult learners of a second language. Once again, an adult learner may actually be making just as much actual progress as a younger student. However, being more aware of the differences between one's goals and one's actual achievements, and being more conscious of how difficult it is to make progress, the adult is likely to experience more discontent than the child does.

Adult learners may be taken aback to find themselves experiencing other obstacles that less mature students encounter. For example, the activity of studying contains elements that make it a difficult one for many people to sustain, irrespective of age. Studying tends to be solitary, it requires concentration, and it takes effort. Many learners, young and old, find themselves inadequately equipped to meet the demands that are imposed upon a young learner by the necessity to function effectively in that combination of circumstances (Csikszentmihalyi et al. 1993) Put simply, a learner of any age may be unprepared for the sustained and sometimes lonely efforts that studying and learning demand.

Of course, to a considerable extent people are able to learn how to learn more effectively. In many countries 'how to study' courses, which attempt to meet the particular requirements of adults who are returning to education, are widely available. However, it takes time and effort for a learner to gain the kinds of habits that help make it relatively easy for a person to engage in continuous study sessions (Howe 1990). Those learners who do succeed in establishing firm study habits experience fewer anxieties than other learners, and are less likely to feel the desire for distractions, or for company, that is common in learners who have had relatively little experience at independent learning (Csikszentmihalyi 1993, Howe 1998).

Some learners of all ages are handicapped by experiencing a severe fear of failing (Covington 1998; Howe 1998). A person who fears failure may genuinely want to succeed, but be so anxious to avoid failing that they cannot make progress because of an unwillingness to engage in difficult learning activities that are not perceived as resulting in certain success.

Fear of failure is linked to a lack of confidence, and for a self-confident and experienced learner, the prospect of an occasional failure may be not at all threatening. After all, seen objectively, a failure to achieve something is nothing more than a sign that further progress is needed, or that more work must be done. However, a less confident student may perceive possible future failures very differently. Someone who has failed often in the past, and found the experience a humiliating one, and one that is a real blow to one's self-esteem, may easily start to feel that failure must be avoided at all costs (Covington and Omelich 1981). To such a person it may appear that the easiest way to avoid failing is to simply not try. After all, if one has not even tried to achieve something, one cannot really fail at it.

Unfortunately, of course, although not trying at all may be a good strategy for avoiding failure, it is certainly not a good way to promote learning. In order to become an effective student, someone who is suffering from fear of failure will need to learn to see possible failures in a different light—as mere signals that more work is needed rather than as proof of the individual's inadequacy. This may be a difficult process, and one that requires time and patience, and the support of sympathetic teachers.

The view of education as a life long process is increasingly being seen as a necessity rather than just a possibility. Undoubtedly, most adults are just as capable of learning as younger people are. There are differences between older and younger students, of course, and, as we have seen, there are both advantages and disadvantages of being a more mature learner. Because arrangements for learning need to be more

flexible when it is necessary for a student to deal with all the responsibilities of adult life, is not always easy for the adult learner to identify educational programs that are suited precisely to the individual's particular needs. However, in many cases the fact that older learners are likely to be especially strongly motivated to learn, and can make informed choices and decisions about the topics to be studied, provides ample compensation for any difficulties.

See also: Adult Education and Training: Cognitive Aspects; Aging and Education; Education in Old Age, Psychology of; Educational Learning Theory; Life Course: Sociological Aspects; Lifelong Learning and its Support with New Media: Cultural Concerns; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development; Plasticity in Human Behavior across the Lifespan

Bibliography

Baltes P, Reinert G 1969 Cohort effects in cognitive development in children as revealed by cross-sectional sequences. *Developmental Psychology* **1**: 169–77

Bruer J T 1999 *The Myth of the First Three Years*. Free Press, New York

Cahan S, Cohen N 1989 Age versus schooling effects on intelligence development. *Child Development* **60**: 1239–49

Ceci S J 1996 *On Intelligence—More or Less: A Bio-ecological Treatise on Intellectual Development*. Harvard University Press, Cambridge, MA

Covington M V 1998 *The Will to Learn*. Cambridge University Press, New York

Covington M V, Omelich C L 1981 As failures mount: affective and cognitive consequences of ability demotion in the classroom. *Journal of Educational Psychology* **73**: 796–908

Csikszentmihalyi M, Rathunde K, Whalen S 1993 *Talented Teenagers: The Roots of Success and Failure*. Cambridge University Press, New York

Flynn J R 1987 Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin* **101**: 271–91

Flynn J R 1991 *Asian Americans: Achievement Beyond IQ*. Erlbaum, Hillsdale, NJ

Gopnik A, Meltzoff A N, Kuhl P K 1999 *The Scientist in the Crib*. Morrow, New York

Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York

Howe M J A 1990 *The Origins of Exceptional Abilities*. Blackwell, Oxford, UK

Howe M J A 1997 *IQ in Question: The Truth about Intelligence*. Sage, London

Howe M J A 1998 *Principles of Abilities and Human Learning*. Psychology Press, Hove, UK

Newport E L 1990 Maturational constraints on language learning. *Cognitive Science* **14**: 11–28

Ramey C T, Yeates K O, Short E J 1984 The plasticity of intellectual development: insights from preventive intervention. *Child Development* **55**: 1913–25

Schiff M, Duyme M, Dumaret A, Tompkiewics S 1982 How much could we boost scholastic achievement and IQ scores? A direct answer from a French adoption study. *Cognition* **12**: 165–96

Schiff M, Lewontin R 1986 *Education and Class: The Irrelevance of IQ Genetic Studies*. Clarendon Press, Oxford, UK

Zigler E, Muenchow S 1992 *Head Start: The Inside Story of America's Most Successful Educational Experiment*. Basic Books, New York

M. J. A. Howe

Education: Anthropological Aspects

The study of education would appear *prima facie* to be central to the anthropological project, and courses that cover anthropological aspects of education are taught in a number of universities, particularly in the USA. Educational anthropology, however, has yet to create as secure a niche for itself as a separate academic discipline as, for example, medical or legal anthropology, which have become increasingly influential in the study of technologically advanced, as well as pretechnological, societies. One reason for the lack of a clearly defined anthropology of education has been that its areas of interest and methodological approaches have overlapped so much with those of other disciplines, in particular social psychology, sociology, and, of course, the growing field of educational studies. Indeed, some of the most influential work in the study of education has been produced either in collaborative projects between anthropologists and those from other disciplines or by the latter bringing anthropological insights to bear in their own work. The influence of anthropology—in terms of scope, theory and methodology—on the study of education, therefore, is rather more than it may at first appear.

1. Scope of Study

For anthropologists, the study of education is far from confined to the examination of formal educational institutions such as schools or to the experiences of those who have not yet been defined as adults. Education, in an anthropological sense, includes the whole range of learning experiences of individuals from conception to death. Indeed, to avoid the narrower connotations of the word 'education,' anthropologists have tended to use much broader terms such as 'socialization' and (less often) 'enculturation,' both of which emphasize that an individual's knowledge of what it means to be a social being is constantly refined throughout their life. The underlying assumption in the anthropological project is that every society or subsociety has an idea of what constitutes an 'educated person' (Levinson et al. 1996), but that this concept differs between societies and changes over time. One classic distinction in the study of education that many anthropologists have made is

between verbal and nonverbal forms of teaching (Lave and Wenger 1991, Singleton 1999). The former generally takes place in formal settings, such as schools, the latter in informal settings, such as apprenticeships. In verbal forms of teaching, the responsibility for ensuring correct understanding lies with the teacher, while in nonverbal forms it lies with the student who learns through close observation of the teacher.

In order to study such broad concepts as 'socialization' and 'enculturation,' anthropologists need to look at many features of the particular society they are examining, from kinship and religion to political and economic systems. When anthropologists study formal educational institutions—such as schools, companies, military, religious and sporting organizations—they also take a holistic approach. They examine how these institutions reflect wider social values and attempt to inculcate them into students through formal and informal means. For example, culturally specific ideas about children and childhood in a society impinge on the educational environment in that society and lead to very different types of classroom experiences. At the same time, classroom activity in any society is presented as a 'natural' process, even though it is clearly historically and culturally contingent.

These characteristics of the anthropology of education can be understood more easily if we look at how they have developed historically and at some of their underlying theoretical and methodological assumptions.

2. *Anthropology and Education: A Historical Overview*

A history of the development of the anthropological study of education might be divided into four main periods. Ethnographers at the end of the nineteenth and early twentieth centuries generally included, in their accounts of the life-cycle of the peoples they studied, some description of how children were cared for, educated, and made full members of society, often through complex initiation rites. In many ways, these accounts differed little from those provided by travelers, missionaries, and social commentators over the previous centuries who documented their observations of the treatment of the young in the societies they visited or in their own societies. Indeed, some comparative educationalists trace their subject back to Herodotus and Xenophon in Ancient Greece and Cicero in Ancient Rome (Trethewey 1976). Only occasionally would the earliest Victorian anthropological theorists, such as Edward Tylor, draw on examples of educational practices from around the world in order to support their theories about how societies evolved from 'savagery' to 'civilization,' but this was never done in a systematic manner.

The beginning of what we now know as the anthropology of education and socialization, therefore, should perhaps most usefully be linked to the first attempts to theorize systematically about the socialization methods of different societies. In this project, Arnold Van Gennep, Franz Boas, and Bronislaw Malinowski were perhaps particularly important.

Van Gennep's work on rites of passage introduced the idea that, while the rituals that mark the transition between life-cycle stages within any society are often both arbitrary and flexible, the underlying pattern of ritual performance often shows a remarkable similarity across societies. At the same time, Boas in the USA and Malinowski in the UK were among the first to challenge Victorian theories of social evolution which equated the poor performance of certain groups in formal Western educational institutions with cultural or biological inferiority. They did this in part by showing how socialization practices are designed to meet the particular 'needs' of a society at a given point in time. Their ideas of cultural relativity—that intelligence can only be judged in the context of a particular culture and not on a universal scale—helped form the basis of a counterargument to the use of universal tests of intelligence in a debate which continues to rage vigorously today.

In Britain, students of Malinowski such as Raymond Firth, Phyllis Kaberry and Audrey Richards continued into the 1930s to include in their ethnographies detailed descriptions of how children became full members of society. Followers of the alternative structural-functionalist tradition of A. R. Radcliffe-Brown, however, steered away from the study of education; they considered it to be too close to psychology and the concept of 'culture' which they eschewed in favor of the study of 'society.'

The second major phase of the development of the anthropology of education was the establishment of the Culture and Personality School in the 1930s and 1940s. This School, which enjoyed by far its greatest influence in the USA, primarily involved the application of psychoanalytic principles to ethnographic material. It believed that by studying child-rearing practices such as feeding, weaning, and toilet training, it was possible to understand the basis on which personality types were developed in different societies. Many of the leading US anthropologists of the period (Ruth Benedict, Cora DuBois, Abram Kardiner, and Ralph Linton) produced work within this paradigm. The School is probably best known, however, for the work of Margaret Mead; her argument that sex roles were determined culturally, as opposed to biologically, played an important part in the re-examination of cultural assumptions about men and women in many industrial societies.

The Culture and Personality School probably hit its height of influence in the wartime period with the publication of, so-called, National Character Studies, such as Benedict's classic account of Japanese society, *The Chrysanthemum and the Sword*. In the postwar

period, however, such approaches became increasingly criticized for exaggerating the homogeneity of personality types within any one society and the link between early child-rearing practices and later adult behavior, for ignoring the significance of relationships which exist between societies, and for reifying society rather than regarding it as a social construction. The result of sustained criticism was a dramatic decline in the influence of the Culture and Personality School. Some remnants of its approach could be seen, however, both in the way data on socialization practices were collected in the Human Relations Area Files, and in the more quantitative studies of childrearing carried out cross-culturally by Whiting and Child in the *Six Cultures Series* (1966). Such research, however, in its quest for similarities that held across cultures, provided little insight into the culturally and historically contingent nature of socialization and educational practices and it had little influence in the UK where the study of children and education virtually disappeared from the work of anthropologists for the next 20 years (see Richards 1970).

What might be seen as the third phase in the development of the anthropology of education began in the late 1960s. This phase was linked, to some extent, with the growth of civil rights movements in North America and Western Europe. In the USA, it focused in particular on what anthropologists generally termed 'cultural discontinuities' in the educational experience of minority and immigrant groups within the US educational system. Detailed ethnographic studies which examined how children, who did not come from the mainstream, white, middle-class society, experienced education in the USA were drawn on to develop programs of multicultural education and greater sensitivity to the educational needs of children from different ethnic backgrounds. As John Ogbu (1981) pointed out, however, these new programs could sometimes be at the expense of essentializing—freezing in time—and stereotyping the 'culture' of minority groups.

In Britain, similarly detailed ethnographic studies were being undertaken on British schools but with the focus on issues of class rather than ethnicity. Several of these (Hargreaves 1967, Lacey 1971) were undertaken by members of the social anthropology department at Manchester University whose head at the time, Max Gluckman, was keen to see anthropological methods used in the study of industrial societies. Probably the most influential work in the UK in this period, however, was Paul Willis' (1977) study of working-class culture in a British secondary school. Significantly, while Willis described himself as a sociologist, he drew heavily on the ethnographic methods used by social anthropologists.

The research that probably had the most lasting influence on the field of educational anthropology from this period, though, was carried out by the French scholar, Pierre Bourdieu who developed the concepts of 'cultural capital' (which the elite in society pass on to their offspring) and 'symbolic violence' (which schools impose on nonelite students so that they learn to silence themselves when in the company of those with greater social standing). While Bourdieu is also often described as a sociologist, the original insights for his later work can be found in the mainstream anthropological research that he carried out among the Kabyle peasants in Algeria (see Bourdieu 1977).

At the same time as ethnographers were examining the ways in which education systems reproduced class and ethnic inequalities, feminist anthropologists in many Western societies became interested in studying the mechanisms whereby gender patterns and inequalities were reproduced through the same system (Okely 1978). Overall, therefore, by the end of the 1970s the anthropology of education could be described as one of the most politically active areas of the discipline as well as one of the few areas where Western anthropologists were actively making studies of, and becoming involved in, policy-making in their own societies.

From the end of the 1970s, anthropologists of education became rather more outward looking, although they often remained politically committed. In part this involved more applied work being undertaken by anthropologists in developing countries. This work often identified closely with the ideas of the anticolonial movement and was sometimes highly critical of the manner in which Western educational systems had been 'imposed' on developing countries and tried to show how this had sometimes led to social disintegration rather than integration (Foley 1977). Perhaps more influential for policy in the West, however, were an increasing number of detailed ethnographic studies by anthropologists of the operation of the education systems of societies whose economies seemed to be rapidly developing at a time when those of the USA and Western Europe seemed, in the late 1970s and early 1980s, to be stagnating.

A large number of works on the anthropology of Japanese education, in particular, began to appear in the light of Japan's astonishing economic growth during the 1980s which many commentators—both inside and outside Japan—believed could best be explained by its education system. While some politicians in the West seemed to suggest that Japan might provide a model for the reform of their own education—particularly preschool and elementary—systems, most anthropologists pointed out the problems inherent in the piecemeal importation of elements of a foreign education system and the impossibility of adopting a system wholesale. Instead, they urged policy-makers to use descriptions of foreign education systems as a mirror for examining the strengths and weaknesses of their own systems (Rohlen 1983).

The fourth major phase in the history of the anthropology of education involved a change of focus

from studying those who run and organize education and other socialization systems to looking at the cognitive worlds of those being educated (Amit-Talai and Wulff 1995). While there had been calls within anthropology since the 1970s, influenced largely by the work of Edwin Ardener, for children, along with women, to be given their own voice in ethnographies (see Hardman 1973), this really only began to emerge two decades later. An impetus for this development may in part have been the United Nations Convention on the Rights of the Child. The Convention certainly sparked intense debate in many countries throughout the 1990s about what constituted a 'child' and what were 'children's rights'. Ideas about the 'child' and 'childhood' were shown to differ greatly between societies with consequent implications for issues of punishment and control in the formal educational institutions in those countries; some commentators indeed felt that the Convention represented an overly Anglo-Saxon view of the child that was being ethnocentrically imposed on other societies (Boyden 1990).

3. *Assumptions*

There are two overlapping sets of assumptions that need to be examined in order to explain the (often unconscious) different approaches that anthropologists have taken to the study of education. The first set relates to whether the anthropologist's primary interest is in the way education systems structure the behavior of individuals or in the way individuals construct education systems. This dichotomy is often described as one between a focus on 'structure' or 'agency.'

The second set of assumptions relate to the ideas—held both by the researcher and the society being researched, though these may not be the same—about how the process of socialization actually operates. For example, in the case of socializing children, is the child a passive object waiting to be molded by society, or an already formed individual which just needs an environment in which it can naturally develop, or a person intent on pursuing its own interests and on whom society has to impose its will (see Jahoda and Lewis 1988)? Sometimes these distinctions are discussed in terms of the child as *tabula rasa*, the Apollonian child (the child as naturally good), and the Dionysian child (the child as innately wilful).

Approaches which concentrate on how socialization systems—particularly formal education systems—help to construct, disseminate, and legitimate a commonly understood and accepted ritual and symbolic language and system of meaning within a society—thereby structuring the experience of individuals who are often presented as *tabula rasa*—are often described as Durkheimian or functionalist. Such analyses have been particularly influential in the study of large-scale modern societies where it is impossible for all the individuals to know each other personally. Formal education systems have played a major role in the formation of the modern nation state; they have often been the main means though which a common language has been disseminated and a national identity constructed in a population. Through their interaction with the economic and religious elements of the social system, they have helped to develop a national work ethic and inculcate national spiritual values. The study of education, therefore, has clearly been important in the work of anthropologists of ethnicity and nationalism such as Benedict Anderson and Ernest Gellner.

Marxist theories have also focused on the way in which education practices produce social beings but rather than seeing these as aids to social integration they have concentrated on how they maintain and reproduce social difference. Marxist and neoMarxist approaches to the study of education largely replaced functionalist and Durkheimian approaches in the 1970s with studies of how class and ethnic inequalities were reproduced through the education system in the USA and UK. These studies showed how an education system can be systematically designed—by means of 'linguistic codes' or a 'hidden curricula'—to provide differential educational experiences for different class, ethnic, and gender groups in a society. While the structural effects of class reproduction could be examined quantitatively, the actual mechanisms of the 'black box' (as it came to be called) of education could be discerned only through intensive qualitative research undertaken by ethnographers.

Social action theorists have focused on how the individual actors in the educational setting help to construct or resist its formal structures. Unlike Durkheimians and Marxists, they do not see individuals in an education system as passive but as active participants in a process of 'cultural production'. Where active participation supports the integrative or the social reproductive functions of education, the analyses of social action theorists overlap with those of Durkheimians and Marxists. As Willis (1977) showed in his account of working-class youth in the UK, for example, the subcultures of resistance formed by these youths to the middle-class curriculum often only served to recreate the class structure of the wider society. Social action theorists, however, generally suggest that there is a range of reactions to the learning process in any educational setting as well as a range of teaching styles. In describing these, they often draw, implicitly at least, on Max Weber's concept of 'ideal types.'

Understanding the differences between the positions outlined above is important in the discipline of anthropology as anthropologists are generally themselves the main research tools. The assumptions on which anthropologists base their research effect not only what they look at but also the types of questions they ask and the conclusions they are likely to draw.

4. Methodology

While there are a number of antecedents to the ethnographic approach to the study of society, it was largely patented and systematized by Malinowski in the 1920s and became closely associated with the discipline of social anthropology. The application of this approach may be the most significant contribution made by anthropology to the study of education. Almost all anthropologists of education have relied on some form of field research, ranging from total participant observation to informal interviews (Woods 1986). Through all of these techniques, the researcher tries to gain an understanding of the participants'—teachers,' pupils,' parents,' officials'—views on the educational process. In its ideal form, indeed, the researcher becomes through participant observation socialized in the society or institution being studied and thereby subjectively comes to empathize with and understand its values.

While field research has been shown to be a very effective means of understanding how socialization works in the broadest sense, it has perhaps particularly been suited to the study of education in 'bounded' institutions—institutions with clearly defined boundaries in terms of time and space—such as schools, churches, and other environments where children are taught basic social categories formally and informally. When children start formal schooling in Western societies, for example, they learn how concepts of time—work time vs. play time, school time vs. home time, holiday time vs. term time—and space—public vs. private, male vs. female, safe vs. dangerous—are divided in the extra-domestic environment. Such institutions, therefore, provide an excellent setting in which to investigate what are considered to be, in a particular society at a particular point in time, its core social values and how these values are socialized—through teaching the meaning of dominant symbols and key ritual performance—and legitimated through reference to ideas of 'tradition' and 'culture.' Field research also allows the researcher to gain a sense of the extent to which these values are accepted, rejected or ignored by those being taught them.

5. Contribution to the Study of Education

The discipline of anthropology has brought a number of important points to bear on the study of education—a relativistic perspective of socialization practices, a holistic approach, an ethnographic method that allows for the simultaneous study of structure and agency in the educational process—which have all become standard practice in the field even if they are not always associated directly in researchers' minds with anthropology. At the same time, the study of education has become increasingly common in anthropological accounts, but often under the headings of the study of ethnicity and nationalism or cognitive and psychological anthropology or the growing field of the anthropology of children rather than as a distinctive field of educational anthropology. It may be that the systematic anthropological study of education has been too threatening to anthropologists many of whom, of course, are themselves full-time educators.

See also: Anthropology; Anthropology and History; Cultural Diversity, Human Development, and Education; Education: Phenomena, Concepts, and Theories; Educational Sociology

Bibliography

Amit-Talai V, Wulff H (eds.) 1995 *Youth Cultures: A Cross-Cultural Perspective*. Routledge, London

Boyden J 1990 Childhood and the policy makers: A comparative perspective on globalisation of childhood. In: James A, Prout A (eds.) *Constructing and Reconstructing Childhood: Contemporary Issues in the Sociological Study of Childhood*. Polity Press, London, pp. 184–215

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK

Foley D 1977 Anthropological studies of schooling in developing countries: Some recent findings and trends. *Comparative Education Review* **21**: 311–28

Hardman C 1973 Can there be an anthropology of children? *Journal of the Anthropological Society of Oxford* **4**: 85–99

Hargreaves D H 1967 *Social Relations in a Secondary School*. Routledge and Kegan Paul, London

Jahoda G, Lewis I M (eds.) 1988 *Acquiring Culture: Cross-Cultural Studies in Child Development*. Croom Helm, London

Lacey C 1971 *Hightown Grammar: The School as a Social System*. Manchester University Press, Manchester, UK

Lave J, Wenger E 1991 *Situated Learning*. Cambridge University Press, Cambridge, UK

Levinson B A, Foley D E, Holland D C (eds.) 1996 *The Cultural Production of the Educated Person: Critical Ethnographies of Schooling and Local Practice*. SUNY, Albany, NY

Ogbu J 1981 School ethnography: A multilevel approach. *Anthropology and Education Quarterly* **12**: 3–29

Okely J 1978 Privileged, schooled and finished: Boarding education for girls. In: Ardener S (ed.) *Defining Females: The Nature of Women in Society*. Croom Helm, London, pp. 93–122

Richards A 1970 Socialization and contemporary British anthropology. In: Mayer P (ed.) *Socialization: The Approach from Social Anthropology*. Tavistock, London, pp. 1–32

Rohlen T P 1983 *Japan's High Schools*. University of California Press, Berkeley, CA

Singleton J C 1999 *Learning in Likely Places: Varieties of Apprenticeship in Japan*. Cambridge University Press, Cambridge, UK

Trethewey A R 1976 *Introducing Comparative Education*. Pergamon Press, Oxford, UK

Whiting B B, Child I L 1966 *Six Cultures Series*. Wiley, New York

Willis P 1977 *Learning to Labour*. Saxon House, Farnborough, UK

Woods P 1986 *Inside Schools: Ethnography in Educational Research*. Routledge and Kegan Paul, London

R. Goodman

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Education: Cultural and Religious Concepts

Because not everyone assigns the same meanings to education, culture, religion, and concepts of education, it is important to explain how those terms are intended throughout this article.

Education means intentional efforts to improve people's knowledge, skills, goals, or values. Culture means the beliefs, patterns of behavior, and physical objects shared in common by members of a group and passed from one generation to the next through education and daily experience. Groups that display particular cultures can be of various kinds—family, community, regional, national, ethnic, occupational, social-class, religious, and more. Religion, as a subcategory of culture, is typically an integrated system of beliefs about a supreme being or gods (theology); the origin and control of the universe (cosmology); the nature of existence (ontology); rules governing human relations (ethics, morals); knowledge and its proper sources (epistemology); and the goal of life (teleology).

Concepts of education are beliefs about what is worth learning and how people should acquire that learning. A group's concept of education can be inferred from the dominant educational agencies, objectives, curriculum content, methods of teaching, and techniques of evaluation found within that group. To demonstrate how different groups' concepts can result in variegated modes of education across cultures, the following discussion is designed to illustrate the proposition that all cultures pursue the same five basic educational goals, but cultures can vary in: (a) the specific learning objectives and curriculum content under those goals; and (b) their system for delivering instruction. The article closes with a brief interpretation of historical trends in cultural conceptions of education.

The examples cited in support of this proposition focus on three cultural categories—national, ethnic, and religious. However, it is apparent that those three categories are not mutually exclusive. For example, national, ethnic, and religious are essentially identical in such countries as Saudi Arabia (82 percent Saudi ethnic origin; 99 percent Muslim) and Portugal (98 percent Portuguese ethnic origin; 92 percent Catholic) (Calhoun 1999). However, in such multicultural countries such as Canada, India, Indonesia, and the USA, significant differences are found among national, ethnic, and religious cultures.

1. *Educational Goals*

By outright statement or by implication, all cultures direct education toward the same five general goals: (a) knowledge and support of cultural traditions; (b) citizenship; (c) moral conduct; (d) vocational preparation; and (e) physical and mental health. However, because of the sociohistorical differences among cultures, the specific learning objectives and curriculum content under each of these goals can vary from one culture to another.

1.1 *Knowledge and Support of Cultural Traditions*

The first goal is to teach each new generation the culture's dominant worldview and history, including the traditional language, social organization, explanations of phenomena (physical and social sciences), and arts. For example, to communicate within their group, Pakistanis learn Urdu, Tunisians learn Arabic, and Peruvians learn Spanish, Qechua, or Aymara. To understand social organization, Tongans learn the traditional chieftain structure of their society, North Koreans learn the country's version of communism, and South Africans study their republic's multiparty system and national assembly. In the realm of literature, India's youths study the epic poems *Ramayana* and *Mahabarata*, students in Taiwan read the *Analects of Confucius*, and those in Spain read Cervantes' *Don Quixote*.

1.2 *Citizenship*

A society's system of governance determines what people need to know about their own role as a citizen and about how their own role relates to the role of other people. Because societies' social systems often differ, citizenship education varies across cultures. The proper behavior of serfs in a feudal society differs from that of citizens in a socialist state or in a capitalistic democracy. The roles of leaders in the Roman Catholic hierarchy differ from the roles of leaders in Quaker, Presbyterian, and Buddhist cultures.

1.3 *Moral Conduct*

Moral education concerns proper ways to act toward other people and, in some cultures, proper ways to act toward supernatural forces (gods, ancestral spirits), nonhuman beings (animals of specified types), and physical surroundings (sacred forests, mountains, and waterways). All cultures teach such virtues as honesty, loyalty, respect, and courage. All disapprove of theft, deceit, and murder. Most usually prescribe additional moral rules not found in every other culture, e.g., obeying God's commandments is an essential moral rule in Judaic, Christian, and Islamic cultures but not in Chinese Taoism or international Communism (Thomas 1997).

Cultures often differ in how they identify the people to whom their moral rules should apply. For example, most cultural traditions teach, at least by implication, that moral rules should guide relationships with friends ('our kind of people' or 'members of our

group') but not with enemies ('those who threaten our welfare'). Therefore, one moral education objective can be that of identifying the social conditions (war, economic competition, and religious confrontations) under which the moral rules should and should not apply to selected groups. Furthermore, as social conditions change, the way moral principles affect personal-social relations often change as well. During World War II, the British and the Germans were enemies, so they were taught to hate, deceive, and slay each other. In contrast, a half century later the British and Germans, both members of the European Union, were taught to apply their common set of moral principles to each other.

1.4 Vocational Preparation

Societies vary markedly in the types and complexity of vocations for which each new generation is trained. Cultures that depend chiefly on hunting and gathering or on agriculture for obtaining food and clothing have far fewer types of vocations than do highly industrialized, occupationally specialized cultures. As a consequence, the detailed learning objectives and instructional content for occupational training can differ significantly across cultures. Furthermore, where vocational preparation is obtained can vary greatly. In dominantly nonmechanized agricultural societies, children learn their vocation through family apprenticeships—working beside their parents in the fields. In highly industrialized societies, most youths prepare for occupations in schools (learning communication, computation, social interaction, and technical skills) and in on-the-job programs away from their homes.

1.5 Physical and Mental Health

Variations from one group to another in the exact aims and content of health education can result from cultural differences in beliefs about: (a) what constitutes health, sickness, and deviant behavior; (b) what causes sickness and deviant behavior; and (c) how to treat sickness and deviance.

In their conceptions of sickness, cultures can disagree about the sorts of distressing conditions that call for healing measures.

Diviners, curers, oracles, shamans, and doctors the world over are not only consulted about bodily ills, but also about mental ills, social problems, and calamities of supernatural provenance which express jealousy, hatred, and suspicions emanating from conflicts over land, money, and inheritance, over marital and sexual disputes, and from political ambitions and rivalries. Further, the bodily ills are commonly taken to be mere epiphenomena: themselves material outcomes of immaterial forces and agencies that inflict punishment for social misdeeds (Worsley 1982 p. 327).

Contrasting cultural views of deviant behavior can be illustrated by a cluster of symptoms (impaired sociability, language, and communication which are accompanied by a constricted range of interests and activities) that are interpreted in typical Western industrialized cultures as indicative of autism, a regrettable malady that should be remedied. However, cultures that attribute autistic-like behavior to possession by spirits can elevate individuals who exhibit such behavior to a respected position reserved for enchanted members of society. Such is the case of individuals classed as *Nit-ku-bon* in West Africa, where an infant who would be diagnosed as autistic in Western cultures is said to be a 'marvelous child' (Ellenberger 1968).

2. Systems for Delivering Instruction

The term 'educational delivery system' refers to a society's manner of organizing the people and institutions that provide education for the populace. One way to identify differences between cultures in their delivery systems is by recognizing (a) the nature of systems' structures and (b) the manner in which components of systems acquire responsibility for particular educational goals.

2.1 Delivery Systems' Structures

A society's educational delivery system can be understood in terms of its institutions, personnel, and roles. An educational institution consists of a group of people (personnel) who together are pursuing the same educational objectives, with each person playing a particular role in the sense of assuming a specific part of the responsibility for achieving the objectives.

Types of institutions found in virtually all cultures are known by such general labels as family, school, church, youth clubs, and television programs. Subtypes that are committed to more specific objectives and roles are identified by more exact labels. Under the designation school there can be such subtypes as nursery school, secular junior high school, Islamic Koran school, junior college, Greek Orthodox seminary, military academy, Chinese language school, cooking school, etc.

In many institutions, roles are identified by titles that enable knowledgeable members of the culture to estimate the responsibilities and skills that those titles imply. Examples of titles and of national cultures in which such titles are used include *docent* (France), *imam* (Egypt), headmaster (England), *guru besar* (Indonesia), and superintendent of schools (USA).

Sometimes roles are traditional ones not overtly defined but learned informally by neophytes' copying the behavior of established members of the institution, with the neophytes then learning what is expected of them by their being praised for performing the role

Table 1
Learning opportunities in representative nations

Nation	Percent of children in school	Teacher/pupil ratio	Percent of population literate	Computers per 1,000 persons	Computers on worldwide network per million persons
China	99	24.3	81.5	2.1	1.8
Denmark	99	10.4	100	270	9,670
Germany	100	18.2	100	165	5,794
Morocco	72	28.3	43.7	1.7	8.6
Netherlands	99	15.7	100	200	11,110
Norway	99	12.5	100	273	19,289
Pakistan				1.2	0.1
Peru	91	27.4	88.7	5.9	35
Russia	100	12.5	98	18	140
Saudi Arabia	62	13.3	62.8	35	1.5
United Kingdom	100	21.2	100	186	7,513
United States	96	18.7	95.5	328	23,012

Source: Calhoun D R (ed.) 1999 *Encyclopaedia Britannica Book of the Year*, pp. 832–7, 880–5

properly and being admonished for deviating from accepted practice. Cultural differences in informally learned educational roles within the family can be illustrated with parents' behavior in Tahiti and Nepal. Levy (1996, p. 134) observed that children in Tahiti were allowed to go about their business of growing up without much verbal direction by adults; and when Tahitian parents did apply directive verbal interventions, their remarks were primarily negative: 'don't do this' or 'stop doing that.' In Nepal, parents offered much verbal guidance to their offspring as the children were learning to perform a new task, and this guidance was usually in positive, encouraging terms. Only after a child had achieved a measure of mastery would negative remarks appear, and then in a moral form, such as 'it's not right to do that,' rather than, 'don't do that.'

In other instances roles are formally described, as in the description of duties included in a contract that a headmaster signs when employed by a school system. Role definitions may also assume the form of the curriculum in a teacher-education program or of the skills and knowledge individuals must display in order to receive a license to teach in an accredited school.

In sum, cultures can differ in the types and subtypes of educational institutions they include and the roles that personnel are expected to perform.

2.2 Institutions' Assignments

The responsibility for achieving educational goals is usually divided among a society's educational institutions, with this division of labor often varying from one culture to another. Typical variations can be portrayed for the realm of moral education.

Whereas every society depends heavily on the family for teaching moral values, cultures often differ in the institutions outside the family that are authorized to promote morality. Some nations mandate a separate class in public and private schools for teaching religion, with much of the subject matter stressing moral behavior. Such is the practice in Finland, France, Germany, The Netherlands, Italy, Norway, Spain, Thailand, and much of Latin America. Other countries, such as Indonesia and Taiwan, require a class in moral education that is not religious in character but includes an emphasis on citizenship and civic responsibility. Public schools in the USA are proscribed from teaching religion, and they rarely maintain moral education classes. However, as in other countries, teachers in all subject matter areas in US schools are expected to foster moral behavior by the way they routinely maintain discipline and encourage amicable social relationships in their classes.

Variations among cultures in the responsibility for achieving educational goals can also result from variations in learners' access to such educational media as schools, books, television programs, and computer networks. Differences in access are principally the result of the availability of such media in a society. Although the latter half of the twentieth century witnessed a marked increase in the availability of schools in virtually all societies, by the twenty-first century the opportunities for formal education still differed significantly among nations. Even more dramatic than the variations in access to schools were the differences among schools in the preparation of teachers, teacher/pupil ratios, and the quality of facilities—reading materials, television, computers, and access to the worldwide computer network. The extent of five such differences in 12 countries near the close of the twentieth century is indicated in Table 1, which reports: (a) the percentage of nations' school-

age children enrolled in school; (b) teacher/pupil ratios; (c) the proportion of the general population over age 15 that was literate as the result of available educational facilities; (d) numbers of personal computers per 1000 persons; and (e) numbers of computers connected to the worldwide network.

3. Historical Trends

Over the nineteenth and twentieth centuries, interactions among cultures increased at an accelerating pace and affected societies' educational practices. The most striking trend has been observed in traditional societies replacing their indigenous educational institutions and teaching methods with ones deriving from Europe and North America. The original purveyors of the European-style school were Christian missionaries who accompanied Western military and commercial forces that colonized much of the non-Western world. Even after most former colonies gained political independence following World War II, the expansion of Western educational practices continued, expedited by foreign-aid projects for developing regions and by advanced communication technologies, such as television and international computer networks.

By the early twenty-first century, the Western-style school—with its constant flow of technological innovations—continued to be the dominant mode of formal education throughout the world. However, this intrusion of European and North American culture into indigenous cultures has not gone unchallenged. Education at the family level has continued to follow long-established traditions in many regions. In addition, conservative forces in various societies have sought to stem the encroachment of formal Western education on non-Western societies' inherited schooling practices. Such resistance has focused particularly on curriculum content. The most frequent curricular changes have been in the fields of history, literature, and the arts. In numerous societies, history books have been rewritten to interpret events from the vantage points of ethnic and religious groups that have gained significant political influence. Indigenous literature and art have been widely encouraged. Less common, and more controversial, have been efforts to teach science and mathematics that reflect traditional cultural worldviews. For example, in India during the 1990s, Hindu nationalists of the Bharatiya Janata Party sought to replace Western science in schools by introducing Vedic mathematics and the ancient science of *vastu shastra*. Party spokesmen charged that Western science was a source of imperialism and rationalism that conflicted with Hindu tradition (Gutek and Thomas 1998). In Australia, curriculum developers sought to accommodate Aboriginal political interests by issuing science-curriculum directives that accorded equal importance to: (a) Aboriginal modes of interpreting natural phenomena; and (b) scientific methods deriving from European/American tradition and practiced internationally (Maratos 1995).

What has been referred to as 'the crises in Islamic education' is the controversy in Moslem societies over the teaching of a Western scientific perspective which includes such precepts as: (a) conclusions about 'truth' should be based on empirical evidence; and (b) all conclusions are tentative and subject to revision on the basis of additional empirical evidence and logic. In Islamic tradition, the sources of true knowledge are the principal holy book, the *Qur'an*, and the sayings of the Prophet Mohammed recorded in the *Sunnah*. In any conflict between the revelations in the Holy Scriptures and 'scientific' theories (such as Darwin's theory of evolution), the 'scientific' theories are judged to be wrong. However, matters not treated in the holy books—such as computer technology or the geology of petroleum production—can be settled by empirical, scientific means. Moslem educators thus face the challenge of convincing learners that they should: (a) limit the application of scientific precepts to issues not dealt with in the holy scriptures; and (b) accept on faith the authority behind the doctrines that bear on matters treated in the scriptures. This same challenge is faced in other religious persuasions as well—Judaism, Christianity, Hinduism, Buddhism, Taoism, and more. It seems likely that the foregoing trends will continue in the future, with education systems across the world becoming increasingly alike in form and content. At the same time, ethnic and religious groups will probably continue their efforts to imbue curricula with content reflecting their own cultural traditions.

See also: Cultural Diversity, Human Development, and Education

Bibliography

Calhoun D R (ed.) 1999 *Britannica Book of the Year*. Encyclopaedia Britannica, Chicago

Ellenberger H F 1968 Impressions of psychiatric problems observed during a sojourn in Dakar. *Canadian Psychiatric Association Journal* **13**(6): 539–46

Gutek G L, Thomas R M 1998 Education. In: *Encyclopaedia Britannica Book of the Year*. Encyclopaedia Britannica, Chicago, p. 202

Levy R I 1996 Essential contrasts: Differences in parental ideas about learners and teaching in Tahiti and Nepal. In: Harkness S, Super C M (eds.) *Parents' Cultural Belief Systems*. Guilford, New York, pp. 123–42

Maratos J 1995 Ideology in science education: The Australian example. *International Review of Education* **41**(6): 357–69

Thomas R M (ed.) 1990 *International Comparative Education*, 1st edn. Pergamon, Oxford, UK

Thomas R M 1997 *Moral Development Theories—Secular and Religious*. Greenwood, Westport, CT

Worsley P 1982 Non-western medical systems. *Annual Review of Anthropology* **11**: 315–48

R. M. Thomas

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Education, Economics of

The economics of education is conceptually very broad. It could logically cover all aspects of the demand for schooling and of the impacts of schooling on subsequent outcomes. This discussion, however, concentrates on the more limited range of issues related to the organization, funding, and performance of schools.

Many facets of the broader field overlap those of labor economics, public finance, and growth theory, and they can best be put into those contexts—leaving this discussion to the unique facets of schools. The economics of education also borrows from other disciplines such as sociology and psychology, but the underlying behavior perspective remains unique to economics.

The economics of education is naturally linked to the study of human capital. Human capital refers to the skills and productive capacity embodied in individuals. While implicitly part of economics for several centuries (Kiker 1968), this idea has recently developed into a central concept in both theoretical and empirical analyses with the foundational work of T. W. Shultz (1961), Gary Becker (1993 [1964]), and Jacob Mincer (1970, 1974). Nonetheless, while abstractly dealing with skills and capabilities of individuals, human capital—to be both predictive and testable—must be defined in terms of more concrete measures. This requirement most often brings up schooling as a clear, measurable aspect of human capital that can proxy for major skill differences. While an expansive view could include all research topics that touch on schooling, it is useful to define the economics of education more narrowly in terms of unique aspects not covered in other subdisciplines: namely, the education sector itself. This discussion is also limited by available data and analyses. Due to a dearth of reliable outcome data from higher education as well as schools in other countries, the economics of education has concentrated its study largely on American K-12 schools. (The study of the economics of higher education, largely lacking student performance information, has directed most attention to issues of access and attendance and particularly of the influence of financial aids and costs on these (see for example McPherson and Schapiro 1990 or Kane 1994). Because these analyses have such a different perspective from those related to primary and secondary education, they are not included in this discussion.)

1. *Outcomes*

It is natural for economists to think in terms of a production model where certain factors and influences go in, and products of interest come out. This model, called the 'production function' or 'input–output' approach, is the model behind much of the analysis in the economics of education. The measurable inputs are things like school resources, teacher quality, and family attributes, and the outcome is student achievement. Let us focus on the outcome, first, and then move on to inputs, where more controversy exists.

Table 1
Percentage of population attaining upper secondary education or more, by country: 1992

	Age group			
Country	25–34	35–44	45–54	55–64
G-7				
Canada	81	78	66	49
France	67	57	47	29
Germany	89	87	81	69
Italy	42	35	21	12
United Kingdom	81	71	62	51
United States	87	88	83	73
Other				
Australia	57	56	51	42
Austria	79	71	65	50
Belgium	60	52	38	24
Czechoslovakia	87	79	68	51
Denmark	67	61	58	45
Finland	82	69	52	31
Ireland	56	44	35	25
Netherlands	68	61	52	42
New Zealand	60	58	55	49
Norway	88	83	75	61
Portugal	21	17	10	7
Spain	41	24	14	8
Sweden	83	76	65	48
Switzerland	87	84	78	70
Turkey	21	14	9	5

Source: NCES (1996).

Historically, the most frequently employed measure of schooling has been attainment, or simply years of schooling completed. The United States leads the world in investing in schooling. By the mid-1970s, three-quarters of US students completed high school, culminating a long national investment period that started with just 6 percent graduating from high school at the turn of the century. Many developed and developing countries, however, have recently mimicked this trend, as illustrated in Table 1, which displays the increasing completion of secondary schooling for more recent age cohorts. By tracing schooling across different age groups, the growth in attainment is readily seen. The value of school attainment as a rough measure of individual skill has been verified by a wide variety of studies of labor market outcomes. In the US, Mincer (1974) pioneered the approach that is now standard. Psacharopoulus (1985) has taken this analysis to the rest of the world. The estimation considers the relationship between earnings (viewed as a direct

measure of individual productivity) and schooling and labor market experience.

However the difficulty with this common measure of outcomes is that it assumes a year of schooling produces the same amount of student achievement over time and in every country. This measure simply counts the time spent in schools without judging what happens in schools—thus, it does not provide a complete or accurate picture of outcomes. Today, more attention is focused on quality outcomes, and the most common measure is tests for cognitive skills. To the extent that the productive value of schooling largely involves the ability to make decisions under uncertainty and to adapt to new ideas and technologies (cf. Nelson and Phelps 1966, Welch 1970), cognitive skills appear to be a valid measure of human capital. This conclusion is supported by recent studies that show a direct correlation between test scores and earnings in the labor market.

While early studies of wage determination indicated relatively modest impacts of variations in cognitive ability after holding constant quantity of schooling, more recent direct investigations of cognitive achievement find generally larger labor market returns to measured individual differences in cognitive achievement (e.g., Bishop 1989, 1991, O'Neill 1990, Grogger and Eide 1993, and Murnane et al. 1995). (A parallel line of research has employed school inputs to measure quality, but, while some controversy exists, this has not been as successful. Specifically, school input measures have not proved to be good predictors of wage or growth. The strategy faces the measurement problems discussed below and ignores the fact that families, peers, and individual ability also enter into the equation.) Similarly, society appears to gain in terms of productivity; Hanushek and Kimko (2000) demonstrate that quality differences in schools have a dramatic impact on productivity and national growth rates.

Attention to quality and cognitive skills also matches a growing policy interest. As countries have moved toward more universal access to schooling, they have turned to issues of quality of publicly provided schooling.

2. *Production*

Let us now look at the other side of the education equation—the inputs, or determinants, of student achievement. Because outcomes cannot be changed by fiat, much attention has been directed at inputs—particularly those perceived to be relevant for policy such as school resources or aspects of teachers.

(This is differentiated from the more common approach in educational research of 'process-outcome' studies where attention rests on the organization of the curriculum, the methods of presenting materials, the interactions of students, teachers and administrators, and the like. An entirely different approach–true experimentation–has been much less frequently applied, particularly when investigating the effects of expenditure differences. The best known is an experiment in class size reduction in the United States during the mid-1980s, Project STAR (Word et al. 1990).) 'Analysis of the role of school resources in determining achievement begins with the 'Coleman Report,' the US government's monumental study on educational opportunity released in 1966 (Coleman et al. 1966). The report was conducted in compliance with a mandate of the Civil Rights Act of 1964 to investigate the extent of inequality in the nation's schools. Although this was not the first such effort, it was much larger and much more influential than any previous (or subsequent) study.

The study's greatest contribution was directing attention to the distribution of student performance—the outputs as opposed to the inputs. Instead of addressing questions of inequality by simply listing an inventory of differences of schools and teachers by race and region, it highlighted the relationship between inputs and outputs of schools.

Unfortunately most of the attention the report generated focused on the report's conclusions rather than its innovative perspective. The controversial conclusion was that schools are not very important in determining student achievement; on the contrary, families and, to a lesser extent, peers are the primary determinants of performance variance. The findings immediately led to a large (but decentralized) research effort to compile additional evidence about input–output relationships in schools (There were also extensive analyses of the report's methodology and of the validity of its inferences. See, for example, Bowles and Levin (1968), Cain and Watts (1970), and Hanushek and Kain (1972).)

The underlying model that evolved as a result of this research is very straightforward. The output of the educational process, that is, the achievement of individual students, is directly related to a series of inputs. Some of these inputs—the characteristics of schools, teachers, curricula, and so forth—are directly controlled by policy makers. Other inputs—those of families and friends plus the innate endowments or learning capacities of the students—are generally not controlled. Further, while achievement may be measured at discrete points in time, the educational process is cumulative; inputs applied sometime in the past affect students' current levels of achievement.

The selected input measures have been fairly similar across studies. Family background is usually characterized by such sociodemographic characteristics as parental education, income, and family size. Peer inputs, when included, are typically totals of a student population's sociodemographic characteristics for a school or classroom. School inputs include teacher background (education level, experience, sex, race,

and so forth), school organization (class sizes, facilities, administrative expenditures, and so forth), and district or community factors (for example, average expenditure levels). Except for the original Coleman Report, most empirical work has relied on data constructed for other purposes, such as a school's standard administrative records. (As discussed elsewhere (Hanushek 1979, 1986), a variety of empirical problems enter into the estimation and the subsequent interpretation of results. The most significant general problems are the lack of measurement of innate abilities of individuals and the imprecise measurement of the history of educational inputs. Both the quality of the data and the estimation techniques are very important in interpreting specific findings, but, as discussed below, these findings have less impact on the aggregate findings illuminated here.) Based on this, statistical analysis (typically some form of regression analysis) is employed to infer what specifically determines achievement and what is the importance of the various inputs into student performance.

3. *The Importance of Measured School Inputs*

The state of knowledge about the impacts of resources is best summarized by reviewing available empirical studies. Most analyses of education production functions have directed their attention at a relatively small set of resource measures, and this makes it easy to summarize the results (Hanushek 1997). Table 2 provides a tabulation of estimates of the impact of school inputs. The 90 individual publications that appeared before 1995 and that form the basis for this analysis contain 377 separate production function estimates. The table indicates the sign of the estimated impact of a given resource on student performance; positive estimates imply that more resources are associated with higher student outcomes and the opposite for negative estimates. Additionally, information is provided about statistical significance, or our confidence that there is a real effect of the resources.

For classroom resources, only 9 percent of studies on teacher education and 14 percent of studies on teacher–pupil ratios found a positive and statistically significant relationship between these factors and student performance. (This summary concentrates on the results. Details of the underlying analyses and the selection of studies for this tabulation are found in Hanushek (1986, 1997). The first column provides a count of the total number of estimates addressing the impact of the given resource.) Moreover, these studies were offset by another set of studies that found a similarly negative correlation between those inputs and student achievement. Twenty-nine percent of the studies found a positive correlation between teacher experience and student performance; however, 71 percent still provided no support for increasing teacher experience (being either negative or statistically insignificant).

Studies on the effect of financial resources provide a similar picture. There is very weak support for the notion that simply providing higher teacher salaries or greater overall spending will lead to improved student performance. Per pupil expenditure has received the most attention, but only 27 percent of studies showed a positive and significant effect. In fact, 7 percent even suggested that adding resources would harm student achievement. It is also important to note that studies involving pupil spending have tended to be the lowest quality studies, and thus there is substantial reason to believe that even the 27 percent figure overstates the true effect of added expenditure.

Table 2
Percentage distribution of estimated effect of key resources on student performance, based on 377 studies available through 1995

Resources	Number of estimates	Statistically significant		Statistically insignificant
		Positive	Negative	
Real classroom resources				
Teacher: pupil ratio	277	14 percent	14 percent	72 percent
Teacher education	171	9	5	86
Teacher experience	207	29	5	66
Financial aggregates				
Teacher salary	119	20 percent	7 percent	25 percent
Expenditure per pupil	163	27	7	34
Other				
Facilities	91	9	5	86
Administration	75	12	5	83
Teacher test scores	41	37	10	53

Source: Hanushek (1997), revised.

4. Do Teachers and Schools Matter?

Because of the Coleman Report and subsequent studies discussed above, many have argued that schools do not matter, and that only families and peers affect performance. Unfortunately, these interpretations have confused measurability with true effects.

Extensive research since the mid-1960s has made it clear that teachers do indeed matter. The simple definition of teacher quality used here is an output measure based on student performance, instead of the more typical input measures based on characteristics of the teacher and school. High-quality teachers are ones who consistently obtain higher than expected gains in student performance, while low quality teachers are ones who consistently obtain lower than expected gains. When this approach has been used, large variations in performance have been uncovered.

In fact, the degree to which teacher differences affect performance is impressive. Looking at the range of quality for teachers within a single large urban district, teachers near the top of the quality distribution can get an entire year's worth of additional learning out of their students compared to those near the bottom (Hanushek 1992).

If teacher quality is central to the performance of schools, then research and policy quite naturally should focus on this. Unfortunately, it has been difficult to link quality—measured by impacts on student achievement—to labor markets. While considerable work has been done to assess the gross flows of individuals into teaching and how those are affected by economic conditions, this does not provide a reliable link to quality. Murnane et al. (1991), Hanushek and Pace (1995), and Ballou and Podgursky (1997) show how various characteristics of teachers including test scores and quality of college are related to labor market conditions and school choices in the United States. Similarly, Hanushek and Rivkin (1997) for the US and Dolton and van der Klaauw (1995) for the UK find that teacher supply adjusts to salaries and labor markets. Nonetheless, these leave open the essential concern about any systematic impacts on school performance. Accurate assessment of teacher labor markets requires reliable quality measures along with information about hiring, retention, and mobility of teachers, and such data have not been available.

5. Study Quality

The previous discussions do not distinguish among studies on the basis of any quality differences, so the results could be distorted by not adjusting the tabulations. While 'study quality' has a variety of subjective components, the available estimates can be separated by a few objective components of quality. First, while education is cumulative, frequently only current input measures are available, which results in analytical errors. Second, schools operate within a policy environment set almost always at higher levels of government. In the United States, state governments establish curricula, provide sources of funding, govern labor laws, determine rules for the certification and hiring of teachers, and the like. In other parts of the world, similar policy setting, frequently at the national level, affects the operations of schools. If these attributes are important—as much policy debate would suggest—they must be incorporated into any analysis of performance, but little progress has been made in identifying or measuring the relevant aspects of policy. Each of these problems potentially leads to biases in the estimated effects of educational inputs. The adequacy of dealing with these problems is a simple index of study quality.

The details of these quality issues and approaches for dealing with them are discussed in detail elsewhere (Hanushek 1997) and only summarized here. The first problem is ameliorated if one uses the 'value added' vs. 'level' form in estimation. That is, if the achievement relationship holds at different points in time, it is possible to concentrate on the growth in achievement and on exactly what happens educationally between those points when outcomes are measured. This approach largely eliminates prior inputs of schools and families, which will be incorporated in the initial achievement levels that are measured (Hanushek 1979). The latter problem of imprecise measurement of the policy environment can frequently be ameliorated by studying performance of schools operating within a consistent set of policies—e.g., within states in the US or similar decision making spheres elsewhere. Because all schools within a state operate within the same basic policy environment, comparisons of their performance are not strongly affected by unmeasured policies (Hanushek and Rivkin 1997). If the available studies are divided by whether or not they deal with these major quality issues, the prior conclusions about research usage are unchanged (Hanushek 1997).

An additional issue, which is particularly important for policy purposes, concerns whether this analytical approach accurately assesses the causal relationship between resources and performance. If, for example, school decision makers provide more resources to those they judge as most needy, higher resources could simply signal students known for having lower achievement. In such a case, the statistical results might reflect the impact of achievement on resources, instead of the other way around. One way to deal with this is to take into account explicitly the resource allocation process. When done in the case of class sizes, the evidence has been mixed (cf. Angrist and Lavy 1999, Hoxby 2000, Rivkin et al. 2000). An alternative involves the use of random assignment experimentation rather than statistical analysis to break the influence of sample selection and other possible omitted factors. The benefits of this approach are well demonstrated in other fields (e.g., medicine or agri-

culture). By randomizing who gets the treatment and comparing outcomes across treatment and control groups, it is less likely than is the case in standard statistical analyses that problems of model misspecification or of selection into treatment situations will bias the estimates of program effects. With one major exception, this approach nonetheless has not been applied to understand the impact of schools on student performance. The exception is Project STAR, an experimental reduction in class sizes that was conducted in the State of Tennessee in the mid-1980s (Word et al. 1990). To date, it has not had much impact on research or our state of knowledge. (While Project STAR has entered a number of policy debates, the results remain controversial (Krueger 1999, Hanushek 1999).)

6. *Investigations Outside the United States*

Researchers have conducted studies on student performance around the world, although with lower frequency. Because student performance data are much less readily available outside the US, most such studies have been special purpose studies that are frequently difficult to generalize. Nonetheless, the results appear qualitatively very similar.

In less developed countries, there is a slightly stronger correlation between resources and achievement, but the overall picture remains the same (Hanushek 1995). This similarity is surprising, because one might expect more impact of resources in schools that begin with very low levels of resources.

For developed countries, the evidence is even scarcer and more difficult to summarize. A review by Vignoles et al. (2000) points to a small number of studies outside the US and shows variation in results similar to that reported elsewhere. Nonetheless, the limited research makes any generalizations difficult.

7. *Efficiency*

Efficiency involves the relationship between inputs and outputs in a production process. The underlying notion is that production is efficient if given inputs produce the maximum output. The simplicity of this statement, however, obscures a variety of complexities that arise when the concept is actually applied.

Typically, economics does not devote much attention to the analysis of efficiency. If there are competitive markets, the behavior of producers and consumers tends to drive outcomes toward efficient production. Because of this basic theorem, it is common simply to assume efficiency.

In education, however, efficiency is never a given. First, education is generally publicly provided. Governmental organizations, which do not face the same incentives as private firms, cannot be expected to move toward efficient production. In particular, few school personnel are rewarded or punished based on the outcomes students obtain, making the incentives for performance minimal or nonexistent. Second, it is difficult to find information on school efficiency. With complete information, parents might be expected to pressure schools to use resources more efficiently—either through local political processes or through moving to districts that did better (Tiebout 1956). Because student performance is influenced by factors schools can't control—i.e., input from family and friends—simply observing student achievement does not accurately reflect school contribution. Thus, parental and voter pressures on schools are lessened by imperfect information.

The underlying political economy of schools that can preserve such apparently large inefficiencies in resource use deserves considerably more research.

8. *Competition and Incentives*

The lack of direct incentives within schools has led to investigations of circumstances where incentive forces may be stronger. While there have been a wide variety of proposed changes (e.g., Hanushek et al. 1994), most have not been implemented or analyzed sufficiently to draw any conclusions about their efficacy.

The classic argument for competition comes from Friedman (1962) with his arguments for vouchers. The well-known arguments suggest that separating the finance of schooling from the production of schooling by allowing students to choose the school they attend would improve individual satisfaction with the outcomes. Within the United States, this approach has been vigorously debated, but few examples of implementation have occurred.

The most celebrated application has been Milwaukee, Wisconsin, where vouchers have been available to a (constrained) number of poor children since 1990 (Witte 1999). A number of privately financed alternatives have also been offered (Howell et al. 2000). Most of the attention to these voucher programs has centered on the student outcomes of students in them compared to public schools. (A variety of controversies have developed in these analyses. Perhaps the primary analytical issue is dealing with a selection of students and an appropriate comparison group. See, for example, the discussions in Witte (1999), Rouse (1998), and Greene et al. 1998).) Because these programs have been very marginal to the education system, there has been little suggestion that the public schools have made any adjustments in response. Thus, these choice experiences have not provided information about how public schools might react to a larger, more institutionalized program.

Private schools offer one possibility for better understanding the effects of competition. They must

compete with the public schools in order to attract clients, and thus they are subject to stronger pressures to provide high levels of performance. The literature on Catholic school performance is summarized in Neal (1998) and Grogger and Neal (2000). The evidence has generally indicated that Catholic schools on average outperform public. This superiority seems clearest in urban settings, where disadvantaged students face fewer options than others. This evidence is, nonetheless, subject to some caveats. First, as recognized since some of the earliest work on the topic (Coleman et al. 1982), it is difficult to separate performance of the private schools from pure selection phenomena. Specifically, since private school students could have attended public schools but instead pay extra for private schooling, they are clearly different than the public school students with identical measured characteristics. A variety of alternative approaches have been taken to deal with the selection problem, and a rough summary of the results after those efforts is that there remains a small advantage from attending Catholic schools. (Grogger and Neal (2000) suggest, however, that there is no advantage to attending private elite schools—a surprising result given the extra cost usually involved in that. These results are possibly due to selection problems, but it not a simple relationship because most people would expect positive selection into these elite schools.) Second, this literature says little about the distribution of school quality within the Catholic sector. Within the public sector, it is widely believed that school quality varies considerably across schools, and one might suspect that the same holds true for private schools.

A related issue is the reactions of public schools to the private sector. Hoxby (1994) demonstrates that public schools in areas that have larger concentrations of Catholic schools perform better than those facing less private competition. This analysis provides the first consistent evidence suggesting that public schools react to outside competition.

Perhaps the most important potential element competition comes from other public school jurisdictions. Specifically, households can choose the specific jurisdiction and school district, à la Tiebout (1956), by their choice of residential location. While adjustment is costly, these choices permit individuals to seek high-quality schools if they wish. Residential location decisions are of course complicated, involving job locations, availability of various kinds of housing, school costs and quality, and availability of other governmental services. Given choice opportunities plus voting responses, this model suggests pressure on schools to use resources effectively. Otherwise one might expect housing values to be affected.

The simple choice model would suggest naturally that larger numbers of schools or school districts per student would offer more opportunities for residents and thus more competition across schools. This simple model motivates the empirical analyses of Borland and Howsen (1992) and its extension and refinement in Hoxby (2000) and in Hanushek and Rivkin (forthcoming). The simple inclusion of measures of concentration indicates that areas with less choice have poorer schools on average (Borland and Howsen 1992). Hoxby (2000) pursues alternative strategies to look at the causal impact of concentrations and finds a larger estimated impact of competition on the performance of schools. Similarly, Hanushek and Rivkin (forthcoming) find competitive impacts on performance of urban schools and that these are particularly important for low income students who have fewer mobility options.

9. *Finance of Schools*

Public finance economists have devoted considerable energy to the study of the finance of schools, largely from the viewpoint of traditional tax and expenditure policy. The educational policy component of this involves equity in the provision of education. These considerations are largely motivated by the many studies that show wide disparities in achievement by race and socioeconomic background (e.g., Coleman et al. 1966, Hanushek 2001). Nonetheless, the typical discussion of equity concentrates primarily on spending disparities across schools and districts.

Indeed, one of the most significant policy issues since the early 1970s in the United States has been the appropriate way for states to fund local schools. While most states have employed a compensatory aid formula to ameliorate some local governments' difficulty in raising taxes, these measures have only partially solved the problem. The issue became the subject of court action in the 1960s, with the case of *Serrano vs. Priest* in California. Suits in many other states followed. As a result, some significant narrowing of spending variations occurred (Murray et al. 1998).

However there has been surprisingly little analysis of the impacts of these suits. This is important to undertake, because the prior research discussed above does not show a clear link between increased funding and school quality or student performance. From the perspective of educational outcomes, then, the value of spending changes related to equalization court cases and the related legislative actions remains an open question (Hanushek 1991). To date, little research has focused on the achievement impacts of school finance policies. The one analysis of the impact of the Serrano case found no lessening of variation in student outcomes after spending was equalized across districts (Downes 1992). A broader, nationwide analysis of district spending on earnings outcomes also found no beneficial relationship between the two, except perhaps for black females (Hanushek and Somers 2001).

10. Some Conclusions and Implications

The existing research suggests inefficiency in the provision of schooling. It does not indicate that schools do not matter. Nor does it indicate that money and resources never impact achievement. The accumulated research simply says there is no clear, systematic relationship between resources and student outcomes. The persistence of this raises important questions about current education policy, as well as a real need for further research.

The main conclusion of this research is that policy decisions should not focus on school resources, because the impact of resources on student achievement is unknown at this time. The policy solution seems to be to establish incentives—rewards or consequences related to student outcomes—and then to permit local schools to develop their own game plans for meeting these goals. While a wide variety of incentives have been proposed—such as vouchers, merit pay, or contracting out—most have not been implemented or analyzed sufficiently to determine their efficacy.

A final major issue raised by these analyses is their implication for other kinds of studies (Hanushek 1996). Most studies involving human capital are interested in its effect on other aspects of behavior. For these studies schooling is largely tangential to other interests, and researchers are typically simply looking for an easy and readily available measure of school quality. The most common has been spending per pupil. However, the inefficient production of human capital introduces natural measurement problems. Direct spending is no longer a good measure of quality because it has no perceivable bearing on performance. Second, it is well known that families have considerable influence on student achievement, implying that school resources are only part of the equation. Both factors suggest that studies which measure student achievement just through resource investment have a high potential for distortion.

See also: Economic Sociology; Education: Skill Training; Educational Institutions and Society; Educational Institutions, History of; Educational Productivity; Educational Sociology; Human Capital: Educational Aspects; Mobility: Social; School as a Social System; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values

Bibliography

Angrist J D, Lavy V 1999 Using Maimondides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* **114**: 533–75

Ballou D, Podgursky M 1997 *Teacher Pay and Teacher Quality*. W. E. Upjohn Institute for Employment Research, Kalamazoo, MI

Becker G 1993 *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, 3rd edn. University of Chicago Press, Chicago

Bishop J H 1989 Is the test score decline responsible for the productivity growth decline? *American Economic Review* **79**: 178–97

Bishop J 1991 Achievement, test scores, and relative wages. In: Kosters M H (ed.) *Workers and their Wages*. The AEI Press, Washington, DC, pp. 146–86

Borland M V, Howsen R M 1992 Student academic achievement and the degree of market concentration in education. *Economics of Education Review* **11**: 31–9

Bowles S, Levin H M 1968 The determinants of scholastic achievement–an appraisal of some recent evidence. *Journal of Human Resources* **3**: 3–24

Cain G G, Watts H W 1970 Problems in making policy inferences from the Coleman Report. *American Sociological Review* **35**: 328–52

Coleman J S, Campbell E Q, Hobson C J, McPartland J, Mood A M, Weinfeld F D, York R L 1966 *Equality of Educational Opportunity*. US Department of Health, US Government Printing Office, Washington, DC

Coleman J S, Hoffer T, Kilgore S 1982 *High School Achievement: Public, Catholic and Private Schools Compared*. Basic Books, New York

Dolton P, van der Klaauw W 1995 Leaving teaching in the UK: A duration analysis. *The Economic Journal* **105**: 431–44

Downes T A 1992 Evaluating the impact of school finance reform on the provision of public education: The California case. *National Tax Journal* **45**: 405–19

Friedman M 1962 *Capitalism and Freedom*. University of Chicago, Chicago

Greene J P, Peterson P E, Du J 1998 School choice in Milwaukee: A randomized experiment. In: Peterson P E, Hassel B C (eds.) *Learning from School Choice*. Brookings Institution, Washington, DC, pp. 335–56

Grogger J T, Eide E 1995 Changes in college skills and the rise in the college wage premium. *Journal of Human Resources* **30**: 280–310

Grogger J T, Neal D 2000 Further evidence on the effects of Catholic secondary schooling. In: Gale W G, Rothenberg Pack J (eds.) *Brookings–Wharton Papers on Urban Affairs, 2000*. Brookings Institution Press, Washington, DC, pp. 151–93

Hanushek E A 1979 Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources* **14**: 351–88

Hanushek E A 1986 The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature* **24**: 1141–77

Hanushek E A 1991 When school finance 'reform' may not be good policy. *Harvard Journal on Legislation* **28**(Summer): 423–56

Hanushek E A 1992 The trade-off between child quantity and quality. *Journal of Political Economy* **100**: 84–117

Hanushek E A 1995 Interpreting recent research on schooling in developing countries. *World Bank Research Observer* **10**: 227–46

Hanushek E A 1996 Measuring investment in education. *Journal of Economic Perspectives* **10**(Fall): 9–30

Hanushek E A 1997 Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis* **19**: 141–64

Hanushek E A 1999 Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis* **21**: 143–63
Hanushek E A 2001 Black–white achievement differences and governmental interventions. *American Economic Review* **91**: 24–8
Hanushek E A, Kain J F 1972 On the value of 'equality of educational opportunity' as a guide to public policy. In: Mosteller F, Moynihan D P (eds.) *On Equality of Educational Opportunity*. Random House, New York, pp. 116–45
Hanushek E A, Kimko D D 2000 Schooling, labor force quality, and the growth of nations. *American Economic Review* **90**: 1184–208
Hanushek E A, Pace R R 1995 Who chooses to teach (and why)? *Economics of Education Review* **14**: 101–17
Hanushek E A, Rivkin G forthcoming Does public school competition affect teacher quality? In: Hoxby C M (ed.) *The Economics of School Choice*. University of Chicago Press, Chicago
Hanushek E A, Rivkin S G 1997 Understanding the twentieth-century growth in U.S. school spending. *Journal of Human Resources* **32**: 35–68
Hanushek E A, Somers J A 2001 Schooling, inequality, and the impact of government. In: Welch F (ed.) *The Causes and Consequences of Increasing Inequality*. University of Chicago Press, Chicago
Hanushek E A et al. 1994 *Making Schools Work: Improving Performance and Controlling Costs*. Brookings Institution, Washington, DC
Howell W G, Wolf P J, Peterson P E, Campbell D E 2000 Test-score effects of school vouchers in Dayton, Ohio, New York City, and Washington, DC. Evidence from randomized field trials. In: *Paper Prepared for the Annual Meetings of the American Political Science Association*, Washington, DC
Hoxby C M 1994 Do private schools provide competition for public schools? National Bureau of Economic Research, Working Paper no. 4978
Hoxby C M 2000 The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics* **115**: 1239–85
Kane T 1994 College entry by blacks since 1970: The role of college costs, family background, and the returns to education. *Journal of Political Economy* **102**: 878–911
Kiker B F 1968 *Human Capital: In Retrospect*. University of South Carolina, Columbia, SC
Krueger A B 1999 Experimental estimates of education production functions. *Quarterly Journal of Economics* **114**: 497–532
McPherson M S, Schapiro M O 1990 Keeping college affordable: Government's role in promoting educational opportunity. August
Mincer J 1970 The distribution of labor incomes: A survey with special reference to the human capital approach. *Journal of Economic Literature* **8**: 1–26
Mincer J 1974 *Schooling Experience and Earnings*. NBER, New York
Murnane R J, Singer J D, Willett J B, Kemple J J, Olsen R J 1991 *Who Will Teach?* Harvard University Press, Cambridge, MA
Murnane R J, Willett J B, Levy F 1995 The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics* **77**: 251–66
Murray S E, Evans W N, Schwab R M 1998 Education-finance reform and the distribution of education resources. *American Economic Review* **88**: 789–812
Neal D 1998 What have we learned about the benefits of private schooling? *FRBNY Economic Policy Review* **4**(March): 79–86
Nelson R R, Phelps E 1966 Investment in humans, technology diffusion and economic growth. *American Economic Review* **56**: 69–75
O'Neill J 1990 The role of human capital in earnings differences between black and white men. *Journal of Economic Perspectives* **4**(Fall): 25–46
Organization for Economic Co-operation and Development 1996 *Education at a Glance: OECD Indicators*. OECD, Paris
Psacharopoulos G 1985 Returns to education: A further international update and implications. *Journal of Human Resources* **20**: 583–604
Rivkin S G, Hanushek E A, Kain J F 2000 Teachers, schools, and academic achievement. National Bureau of Economic Research, Working Paper No. 6691 (revised)
Rouse C E 1998 Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics* **113**: 553–602
Schultz T W 1961 Investment in human capital. *American Economic Review* **51**: 1–17
Tiebout C M 1956 A pure theory of local expenditures. *Journal of Political Economy* **64**: 416–24
Vignoles A, Levacic R, Walker J, Machin S, Reynolds D 2000 The relationship between resource allocation and pupil attainment: A review. Research report No. 228, Department for Education and Employment, UK
Welch F 1970 Education in production. *Journal of Political Economy* **78**: 35–59
Witte J F Jr 2000 *The Market Approach to Education*. Princeton University Press, Princeton, NJ
Word E, Johnston J, Bain H P, DeWayne Fulton B, Boyd Zaharies J, Nannette Lintz M, Achilles C M, Folger J, Breda C 1990 *Student/Teacher Achievement Ratio (STAR), Tennessee's K-3 Class Size Study: Final Summary Report, 1985–1990*. Tennessee State Department of Education, Nashville, TN

E. A. Hanushek

Education for Gifted and Talented Children

The term 'gifted' is used to identify children who are superior in mental ability or very high academic achievers. Generally they are children who learn easily and rapidly, reason or use logic fluently, think well theoretically and abstractly, are advanced in verbal and/or mathematical skills, and retain or remember information exceptionally well. Their cognitive precocity is exhibited in their ability to function mentally at levels far beyond what is typical or normative for their chronological age.

Intelligence test scores are often used, along with other indicators of ability such as achievement test scores and teacher rating scales, to identify gifted children. While an IQ of 140 or above has often been the cutting level for classification as gifted, following the lead established by Terman (1925) in his study of

gifted children, a broad range of IQs from 120 upward have been used in real practice.

The terms 'talent' or 'talented' have often been used as synonymous with 'gifts' or 'gifted,' and the phrase 'gifted and talented' was, and still is, widely used with little effort to distinguish the two terms. For some practitioners 'talent' has been used to denote special ability in the arts while for others it refers both to academic talent (verbal talent, mathematical talent, etc.) and to artistic talent (music, drama, visual art, etc.). Leadership has also often been identified as a form of special talent.

Pioneering research on gifted children began with the work of Lewis Terman at Stanford University in 1920. Using the Stanford–Binet Intelligence Scale that he had developed and published in 1916, he went on to identify 1,000 children with IQs above 135 (most had IQs above 140) who were then on average 12 years of age, and to study their lives in great detail. The first major finding of the study, published in 1925, was that the children overall were not, as commonly thought, a group of social misfits characterized by eccentric behavior, but rather were generally quite normal except for their academic superiority and very good health. While the sample was quite biased in favor of white middle-class families, the findings were widely accepted in educational settings and came to focus educators' attention on their special need for more high-level and challenging instruction.

Understanding of intelligence and its role in giftedness was enhanced a great deal by the discovery of factors of intelligence and the fact that people have differentiated and unique patterns of cognitive capabilities—numerical, verbal, spatial, fluency—all of which are really components of general intelligence. Later, the same analytical tools of factor analysis were used to show that some aspects of intelligence are quite purely genetically determined (fluid) and others acquired through the interaction of genetic disposition with learning opportunities (crystallized). Thus, reasoning is more closely a fluid aspect of intelligence, whereas mathematical ability may be learned capacity that in turn manifests itself in intelligent behavior.

The componential approach to understanding intelligence and giftedness in a sense ran amok when psychologist Guilford (1950), using new methods of factor analysis, seemingly found hundreds of factors or facets of intelligence. Guilford proposed a 'structure of intellect' in which there were three dimensions: content or input on which thinking operates; operation or thinking processes and skills; and products or outcomes of the operations. The model was widely embraced, especially because of its inclusion of divergent thinking as an operation. However, it lost favor in the 1980s and 1990s as Guilford's factor analysis methods were questioned.

A major conception of human ability and giftedness appeared in Sternberg's new theory of human intelligence (1985). His componential theory of human intelligence is a cybernetic model that involves selective acquisition of knowledge, complex thinking or processing operations, and the creative production of new knowledge. A first major aspect of intelligence is a set of metacomponents, or cognitive operations, such as recognizing and clarifying problems, designing solution strategies, and judging the solutions. Another aspect is knowledge acquisition components, how we encode, retain, and use knowledge. Finally, there are performance components that are pieces of knowledge or skills needed to perform specific tasks. Sternberg's theory is now widely recognized in the world of gifted education, but no good test has emerged from Sternberg's theory and no specific identification of gifted children is carried out in the field of gifted education based on the theory.

1. *New Focus on Talent*

Publications by Gardner (1983), Bloom (1985), Csikszentmihalyi et al. (1993), and Gagné (1999) have presented new research and theoretical conceptions of specific human abilities, and paved the way for increasing attention among educators to the nature, development, and nurture of youth talents. After five decades of research on intelligence and components or factors of intelligence, Gardner proposed a new theoretical formulation of multiple intelligences that posited seven discrete general abilities: linguistic, logical-mathematical, spatial, musical, bodily-kinesthetic, interpersonal, and intrapersonal. Gardner acknowledged that the list may not be definitive; there may be other intelligences. Furthermore any or all of the seven intelligences might well be broken down into subcomponents. Gardner also acknowledged that calling the seven abilities 'intelligences' was somewhat arbitrary and that they might well be called 'talents' or 'abilities.' Whatever they are called, it is clear that Gardner's conception of multiple intelligences has had a major impact on school practice in general and gifted education in particular, but not so much in the area of identification of specific intelligences or talents in children as in the area of instruction directed toward development of gifted children's intelligences or talents.

Bloom carried out research focused on concert pianists and sculptors as representative of the arts, research mathematicians and research neurologists representing cognitive or intellectual fields, and swimmers and tennis players representing athletics. The subjects had all achieved world-class recognition before age 35. From the extensive interviews conducted with 120 subjects Bloom concluded that finding or identifying a child's talent strength or potential early was essential for the long-range development of the talent, that both parents and teachers play critical roles in supporting the child's talent development, and that motivational encouragement was essential at each stage of the child's talent development.

Confirmation of the need to identify specific youth talents as a prelude to nurturing talents came in other research of multipotentiality among the gifted. Specifically it was found that the idea that gifted children have potential for many different fields of achievement was a false conception. Instead, by the time they reach middle school and high school they have a limited range of talent potentials.

Recent research by Csikszentmihalyi et al. (1993) in which they studied talent development among teenagers with special talents in mathematics, science, music, art, and athletics, also concluded that it is essential for long-range talent development that youths' specific talents be identified as a prelude to the nurture of talents. They also concluded that there is a long-range process of differentiation of youths' specific talents, and integration of those talents with other personal characteristics to arrive finally at high-level expertise, creative achievement, or occupational fulfillment.

Gagné (1999) proposed that, out of several basic and original factors of giftedness, specific talents emerge as special forms of aptitude or ability as a result of experiences at home, in school, with peers, and in the broader community surrounding a child. The basic gifts or forms of giftedness are intellectual, creative, socioaffective, sensorimotor, and other unspecified, genetically determined abilities. Gagné concluded that there are six talent fields (academic, technology, arts, leisure activities, athletics, and communications-business) and six aptitude domains (intellectual, creative, socio-affective, personal, sensorimotor, and nonconventional). This author also surveyed students ages 9 to 17 and found 10 talent categories represented among them: academic, artistic, cognitive, creative, communication, athletic-physical, games, language, personal-social, and vocational-technical. As talent development relates to and/or is carried out in schools Feldhusen (1999) summarized these conceptions as representing talents in four domains: academic-intellectual, artistic, vocational-technical, and interpersonal-social.

In summary, there is an urgent need in the field of gifted education to move on beyond the simple dichotomous conception of youth as 'gifted' or 'ungifted' to a multi-aptitude or talents view of human abilities. Recent factor analyses of tests of human abilities (Carroll 1993) revealed once more, and more definitively than ever before, that there is a diversity of human abilities, but at the most macro level they can be categorized as general intelligence or g, fluid intelligence, crystallized intelligence, general memory and learning, broad visual perception-space relations-visualization, auditory perception, cognitive speed, and memory-retrieval ability. These basic human abilities underlie differentially the domains or fields of human talent that grow in youth with the nurturing experiences of homes, parents, peers, school, and the general community in which they live.

2. *Identification*

The field of gifted education places great stress on the identification process and makes pretense of finding the 'truly' or 'really' gifted child. Although there is little research verifying or validating identification processes, there is much discussion in the literature on finding gifted children, and also on finding those who are less easily found—such as children from some minority racial groups, physically handicapped children, and youth from culturally impoverished homes. The essential question to be answered, but which is rarely addressed, is the reliability of the process and/or the ratio of false positives and false negatives to a hypothetical total population of gifted children in a community. False positives are those who are incorrectly identified as gifted and who therefore fail to profit from, or do not need, special educational programs. False negatives, the far more serious problem, are those children whose abilities are at the level where need for the program is evident, but they are not identified as gifted and eligible for the program.

Children may be identified for a gifted program using an identification system based on Renzulli's three-ring conception of giftedness: ability, task commitment, and creativity. While children who are very high achievers are often declared not to be gifted, simply very hard workers, the Renzulli conception recognizes what is called 'task commitment' as a major component of giftedness. Task commitment may represent high-level and sustained interest in a field or topic of study as well as readiness to work long and hard at a task.

The other two factors of giftedness—abilities and creativity—constitute quite different areas for assessment. While valid and reliable measures of youth abilities are readily available, assessment of creativity in the identification process is much more tenuous. While the tests developed by Torrance (1981) were designed to measure creative ability, their validity and reliability for the assessment of creativity remain highly questionable.

The best combination of measures for the assessment of youth talents seems to be intelligence tests like the Wechsler Intelligence Scales, that yield scores for information, comprehension, arithmetic, similarities, vocabulary, digit span, picture completion, picture arrangement, block design, puzzle assembly, coding, and mazes as basic aptitudes; or the Differential Aptitude Tests, that yield scores for verbal reasoning, numerical reasoning, abstract reasoning, perceptual speed and accuracy, mechanical reasoning, space relations, spelling, and language usage.

Beyond the basic aptitudes or abilities, standardized school achievement tests offer valuable assessments of talented youths' precocity or advanced levels of learning, especially when they are administered off level; that is, when advanced forms of the tests are used in relation to a child's age or grade level. The major

standardized achievement tests available in the United States are the Iowa Test of Basic Skills, Stanford Achievement Tests, Metropolitan Achievement Tests, SRA Achievement Series, California Test of Basic Skills, Sequential Tests of Educational Progress, and the Scholastic Aptitude Test. All focus on student achievement in basic school subjects. However, the only test that has been used widely for off-level testing is the Scholastic Aptitude Test that has been widely used in the ubiquitous academic talent searches.

A third and valuable type of assessment is rating scales, most often completed by teachers who have observed children in the classroom over an extended period of time. Widely used are the Scales for Rating The Behavioral Characteristics of Superior Students by Renzulli et al. (1998), and the Purdue Academic Rating Scales by Feldhusen et al. (1997). The Renzulli scales assess learning, motivation, creativity, leadership, artistic ability, musical and dramatic talent, precision and expressiveness in communication, and planning. The Purdue Academic Rating Scales assess competence in mathematics, science, social studies, English, and foreign language learning. It is a misfortune of the field of gifted education that rating scales are often developed locally and used in the identification process without any attention to the validity and reliability of the scales.

A fourth area of assessment for identification is evaluation of student products portfolios, and/or performance ratings by skilled observers. This method of assessment is most appropriate for the several domains of the arts. Winner and Martino (1993) delineated well the early signs and development of giftedness in visual arts and music, and presented a comprehensive inventory of assessment procedures including tests, portfolios, and observation of performance. Clark and Zimmerman (1998) have also studied the nature of artistic talent intensively over several decades, and they summarized their work concluding that the development and assessment of artistic talent is a long-term process, not the one-time evaluation often used by schools to identify academically gifted children. Identification of talent in the arts usually involves not only portfolio evaluation but also day-by-day observational assessment by knowledgeable teachers of students' work in progress in the art classroom, as is also the case of music teachers and student musicians in the music classroom.

Identification of cognitively gifted and academically talented youth is most often carried out by combining or summarizing the collected data and judging if a child has reached a criterion level for giftedness or talent designation. Schools assume that children who are designated 'gifted' need the special educational service offered, but there is little evidence that they profit from it more than they would from remaining in the regular classroom. Research on boredom among gifted youth in school indicates that while many are indeed bored, so also are many youth who are not identified as gifted (Feldhusen and Kroll 1991). However, the same research indicates that the boredom of gifted youth is due mainly to the slow pace and low level of teaching in regular classrooms, while less able youth are simply not interested in much of typical classroom subject matter.

In summary, the identification process for all areas of giftedness and talent should best be viewed as a long-range process of helping children, parents, and teachers come to understand children's unique aptitudes and talents as a basis for appropriate educational planning and programming. There is probably little value in one-shot designation of children as gifted. A wide variety of tests, observations, and product assessments can be used over time to identify the nature and levels of emerging talent.

3. *Developing Gifts and Talents*

There is no definitive educational plan or program known to be uniquely appropriate for the development of gifted and talented youth. Generally it is assumed that, because of their precocity and capacity to learn faster than children of low or average ability, they need accelerated instruction, at a higher level commensurate with their achievement levels as revealed in the identification process, and at a faster pace. Children with IQs of 120 and higher are generally ready for instruction at advanced levels of the curriculum, and they learn new advanced material much more rapidly than children of low and average ability.

A wide variety of strategies are used to provide appropriate and intensive educational services to gifted and talented students, including:

(a) full-time classes in all subjects at the level of grades K–6;

(b) special classes in particular subjects (e.g., honors);

(c) enrollment part time at a college while in high school or taking college-level courses offered in the high school program (e.g., Advanced Placement);

(d) early admission to elementary school, middle school, high school, or college; and

(e) enrollment in a special school for gifted or talented students at any, several, or all grade levels.

The above options are offered throughout the United States and in some other countries. More common, however, are the following, far less intensive, services:

(a) pullout, resource room, or withdrawal programs in which gifted children work for 3–6 hours a week on creative enrichment, problem solving, and other supplementary activities not in the regular school curriculum;

(b) cluster grouped in an otherwise mixed ability classroom with a teacher trained to do accelerated and enriched learning activities with children in the gifted cluster; and

(c) individualized teaching in a regular mixed ability classroom guided by an individual learning plan for the gifted or talented student.

There is widespread assumption, also, that gifted and talented youth need the social experiences of interaction with peers of similar levels and types of giftedness and/or talent, and that they need higher level models and mentors in their own areas of talent. Peer models serve to motivate talented youth to higher level achievements and often serve as friends for gifted students who are otherwise socially rejected by less able peers.

Mentors and/or achievement models provide gifted and talented youth with blueprints or images of their own potential and ultimate achievements in a talent domain. Mentors are really key educational tools or resources in the education of gifted and talented youth. They teach, they motivate, and they provide emotional support to gifted and talented youth. Haeger and Feldhusen (1989) described a program of mentoring for talented youth at Purdue University, in which university faculty and graduate students served as mentors for special experiences on Saturday mornings. All students who were assigned to mentors were first required to take a Saturday morning course for nine weeks in the subject matter area in which they would have the mentoring experience. Haeger and Feldhusen found that prior background knowledge in the mentoring area was essential for successful mentoring experiences. Pleiss and Feldhusen (1995) carried out an intensive review of the theory and research literature on mentors, role models, and heroes in the lives of gifted and talented students and concluded that there is abundant support for their value in nurturing talent and social development of precocious youth.

Communities vary in their educational resources and attitudes toward gifted and talented youth. The great educational movement for inclusion or heterogeneous grouping of all youth for common educational experiences has spread to countries throughout the world. It has generated opposition to programs and services for gifted and talented youth in many schools. While inclusion may be workable and desirable for students of low and average ability, there is ample evidence that precocious youth who are also fast learners and able to pursue learning at deeper and more abstract levels are best educated in special classes with peers of similar ability and teachers who have been trained to work with the gifted and talented.

4. Summary

Gifted students are those who have superior general intellectual ability as revealed by intelligence and achievement tests and teacher ratings of their classroom performance. Talented students excel in one or more areas, such as science, mathematics, writing, arts, leadership, or technology, and are identified with tests, rating scales, portfolios, and performance ratings. Identification of gifted and talented students is a long-range and continuous process leading to better and better understanding of youths' talent potentials.

Gifts and talents are present in students at varying levels, and thus in planning educational services for them, their levels of ability must be linked to appropriate levels and types of service. There is a wide variety of services available to use in programs for talent development. Intensive and accelerated classes and programs, mentors, and schools focused on particular areas of talent such as mathematics, science, art, music, leadership, or technology serve best to meet the developmental needs of talented youth and to help them achieve their full potential.

See also: Gifted and Talented Individuals: Developmental and Educational Overview; Giftedness, Psychology of; Intellectual Functioning, Assessment of; Intelligence: Central Conceptions and Psychometric Models; Intelligence: Historical and Conceptual Perspectives

Bibliography

Bloom B S 1985 *Developing Talent in Young People*. Ballantine Books, New York

Carroll J B 1993 *Human Cognitive Abilities: A Survey of Factor-analytic Studies*. Cambridge University Press, New York

Clark G A, Zimmerman E D 1998 Nurturing the arts in programs for the gifted and talented. *Phi Delta Kappan* **79**(10): 747–51

Csikszentmihalyi M, Rathunde K, Whalen S 1993 *Talented Teenagers*. Cambridge University Press, New York

Feldhusen J F 1999 Talent identification and development in education: The basic tenets. In: Cline S, Hegeman K T (eds.) *Gifted Education in the Twenty-first Century*. Winslow Press, New York, pp. 89–100

Feldhusen J F, Kroll M D 1991 Boredom or challenge for the academically talented. *Gifted Education International* **7**(2): 80–81

Feldhusen J F, Hoover S M, Sayler M F 1997 *Identifying and Educating Gifted Students at the Secondary Level*. Royal Fireworks Press, Unionville, NY

Gagné F 1999 The multigifts of multitalented individuals. In: Cline S, Hegeman K T (eds.) *Gifted Education in the Twenty-first Century*. Winslow Press, New York, pp. 17–46

Gallagher J J 1997 Issues in the education of gifted students. In: Colangelo N, Davis G H (eds.) *Handbook of Gifted Education*. Allyn and Bacon, Boston, pp. 10–23

Gardner H 1983 *Frames of Mind, The Theory of Multiple Intelligences*. Basic Books, New York

Guilford J P 1950 Creativity. *American Psychologist* **5**: 444–54

Haeger W W, Feldhusen J F 1989 *Developing a Mentor Program*. DOK Publishers, East Aurora, NY

Pleiss M K, Feldhusen J F 1995 Mentors, role models, and heroes in the lives of gifted children. *Educational Psychologist* **30**(3): 159–69

Renzulli J S, Smith L H, White A J, Callahan C M, Hartman R K 1998 *Scales for Rating the Behavioral Characteristics of*

Superior Students. Creative Learning Press, Mansfield Center, CT
Sternberg R J 1985 *Beyond IQ: A Triarchic Theory of Human Intelligence*. Cambridge University Press, New York
Terman L M 1925 *Genetic Studies of Genius: Vol. 1. Mental and Physical Traits of a Thousand Gifted Children*. Stanford University Press, Stanford, CA
Torrance E P 1981 *Administration, scoring and norms-technical manual for: Thinking creatively in action and movement*. Scholastic Testing Service, Bensenville, IL
Winner E, Martino G 1993 Giftedness in the visual arts and music. In: Heller K A, Monks F J, Passow A H (eds.) *International Handbook of Research and Development of Giftedness and Talent*. Pergamon Press, New York, pp. 253–81

J. F. Feldhusen

Education (Higher) and Gender

In this article I sketch the relative position of men and women in higher education. I examine three distinct aspects of the educational system: access, college experiences, and post-collegiate outcomes. Women fare relatively well in the area of access, less so in terms of the college experience, and are particularly disadvantaged with respect to the outcomes of schooling.

Higher education is defined as formal enrollment and degree acquisition at the college and university level. While educational systems differ across countries, most distinguish between primary (elementary), secondary (high school) and tertiary levels of education (see UNESCO 1998 for details). As elementary and secondary education become increasingly universal, higher education becomes increasingly salient as the basis for educational stratification.

1. *Access to College in the United States*

In the United States women now represent the majority of students in the system of higher education. In 1995, 55.4 percent of students enrolled in college were women. During the 1990s, women high school graduates were five percentage points more likely to enroll in college than were men (64.7 versus 59.7 percent, respectively). Women's share of degrees climbed steadily during the 1970s and 1980s, a period when the fraction of college-age young adults enrolled in school increased slowly but steadily. By 1982, women surpassed men in the number of bachelor's degrees earned. Women have garnered more bachelor's degrees than their male counterparts ever since. By 1995, 54.6 percent of bachelor's degree recipients were women. Women earned 59.5 percent of two-year degrees, 52.7 percent of master's and professional degrees, and 39.3 percent of doctoral degrees (National Center for Educational Statistics 1997).

Women remain slightly overrepresented in lower-ranked schools, as indexed by higher acceptance rates, higher student/faculty ratios, lower standardized test scores, and lower fees (Jacobs 1999). The small remaining sex gap is due to two factors: (a) the relative scarcity of women in schools with large engineering programs (engineering programs tend to be above average in selectivity); and (b) the tendency of women to enroll in school part time, since lower-status institutions are more likely to accept part-time students. Selected reports on admissions as well as enrollment from leading institutions indicate that women are well represented among recent entering classes, except in schools that prominently feature engineering programs. There is some evidence that some liberal arts colleges are worried that they are enrolling too many women, and are starting to take steps to recruit men (New York Times 1998).

The parity women have achieved in college completion is a recent phenomenon, but the 1950s and 1960s represented a historically depressed level. Women represented 41.3 percent of college graduates in 1940, slipping to 23.9 in 1950 and remaining at an historically low 35.0 percent in 1960 (US Bureau of the Census 1975). Goldin (1997) estimates, based on retrospective reports from the 1940 Census, that women's college enrollment rates exceeded 90 percent of men's from the late 1890s until the mid 1920s, although the inclusion of normal schools arguably inflates Goldin's figures.

For the entire twentieth century in the United States, women have comprised a large proportion of students in primary and secondary schools. Women's rate of enrollment among 5–19 year olds has exceeded 90 percent of men's rate since as early as 1850, and 98 percent since 1890. Women have represented the majority of high school graduates since at least 1870: in 1920 over 60 percent of high school graduates were women (US Bureau of the Census 1975, 369–70, 379). The median years of schooling completed by women exceeded men for most of this century until the G. I. Bill after World War II enabled men to surpass women.

2. *International Comparisons*

Women in the United States have surpassed their counterparts in other countries in access to schooling at both the secondary and tertiary levels for more than a century. Today, the United States enrolls more college students per capita than virtually any other country, and women's share of college enrollments in the United States exceeds that in most other countries. (Data for selected countries are presented in Jacobs

1996. See also UNESCO 1998.) In most of the advanced industrial countries of Europe, women's share of enrollments is quite high. But even here, substantial variation persists, with women's share ranging from 40 percent of college students in Switzerland and 41 percent in Germany to 55 percent in France and 61 percent in Portugal. Women fared well in terms of schooling in the socialist countries of Eastern Europe (Kelly and Slaughter 1991), and socialist regimes in developing countries typically emphasized schooling for girls in their initial years in power. The post-socialist experience warrants close scrutiny, as women's status is eroding in many spheres in these countries (Biaecki and Heyns 1993).

Women's share of enrollment in Latin American colleges and universities is often quite high: Brazil, 53 percent; Argentina, 47 percent; Chile, 42 percent. Asian countries follow: in both China and India one-third of college students are women. African countries include many with the lowest share of female enrollments in the world. Within each of these regions, there is substantial variation in the share of women's enrollment.

Gender disparities are highest at the tertiary level, as young men are typically the first to pursue college. Gender disparities in expenditures are greater than those in enrollments, because college education is more expensive than elementary or secondary schooling.

3. *The Educational Process*

Women and men experience college differently, and face markedly different outcomes. Of the many respects in which the college experience differs by gender, here I consider fields of study and faculty representation.

Women and men pursue different fields of study in college. In the United States, 30 percent of women would have to change fields of study in order to be distributed in the same manner as men (Jacobs 1995a). The sex typing of fields of study is a worldwide phenomenon yet it varies between countries. For example, 51.6 percent of engineering students are women in Kuwait, compared with 3.3 percent in Switzerland and Japan (UNESCO 1998). One of the most striking contrasts is within the divided Germany: in the former East Germany, 32.4 percent of engineering graduates were women, compared with only 7.5 percent in West Germany. In Poland, 62.7 percent of mathematics and computer science degrees went to women, compared with 35.9 percent in the US and 21.0 percent in Egypt.

In the United States during the early 1960s, women were concentrated in an extremely limited range of fields. Education drew almost half of women undergraduates, and over 70 percent of women graduates were concentrated in just six fields: education, English, fine arts, nursing, history, and home economics. Now business is the leading field of study for women. In 1995, women garnered 52.3 percent of life science bachelor's degrees, 46.8 percent of mathematics degrees, 48.0 percent of business degrees, but only 17.2 percent of engineering degrees. Segregation across majors declined substantially from the mid-1960s through the mid-1980s, but has reached a plateau since 1985 (Jacobs 1995a).

Women have not always been segregated into separate fields from men. Founders of the most prominent women's colleges tried hard to maintain curricula that matched or exceeded men's in scope. During the early years of land-grant schools, no separate curriculum for women existed. A peculiarly feminine curriculum began to emerge with the development of home economics. This development reflected an enduring emphasis on female domesticity, but also was promoted in part by women academics, who were excluded from other fields and sought to create a field of expertise and set of job opportunities for which they would be uniquely suited. At the same time, this development contributed to the emergence of a distinctively feminine college experience for young women that served to limit the career prospects of many female graduates.

Men represent the great majority of college and university faculty worldwide (see UNESCO 1998, Jacobs 1996). In the United States, 33.4 percent of faculty were women in 1993. Women's representation declines with the prestige of the institution: 45.3 percent in public two-year schools, 33.9 percent in the public comprehensive schools, and 23.3 percent in public research universities. The number of women also declines with faculty rank. Women represent 42.1 percent of assistant professors, 30.0 percent of associate professors, and 17.0 percent of professors (National Center for Educational Statistics 1997). The US record actually looks quite favorable by comparison with the professoriat in Britain and France, which were 2.3 and 8.7 percent female, respectively, during the early 1980s. Graham (1978) notes that women's representation on the faculty of US colleges and universities actually declined between 1930 and 1970, before beginning a sustained advance during the 1970s and 1980s.

There are many reasons why women's entrance into faculty positions is so low. Until recently women were a small proportion of Ph.D. recipients; women are concentrated in a limited number of fields; women entered academia in large numbers during a period of retrenchment, and pursued fields that were facing sharp declines in enrollments (Slaughter 1993). Nonetheless, women's progress remains far slower than would be expected. Viewed optimistically, if a sizable fraction of women who are currently assistant professors are granted tenure, then the sex composition of the faculty will change dramatically from 2000–2020. Parity is unlikely for quite a long time because of the

number of fields in which women Ph.D.s remains severely underrepresented (Ransom 1990).

Much research has examined the position of women faculty members (see Chamberlain 1988, Jacobs 1996 for reviews). Studies have examined gender inequality with respect to hiring patterns, promotion rates, publication rates, mobility between institutions, job satisfaction, turnover, salaries, and the sense of personal and professional marginalization.

The notion of cumulative disadvantage seems to be a reasonable summary of the underrepresentation of women in faculty positions. In other words, women have been disadvantaged to some extent in every stage of the academic career process. This would account for women's underrepresentation in the higher echelons of university administration, in higher ranks, and in higher-status institutions. Graham (1978) suggests that the extreme exclusion of women from Ivy League institutions undermined the position of all women faculty, because, with the emergence of the research university as the pinnacle of the higher education system, these schools came to set the pattern for higher education as a whole.

A number of researchers see the position of women faculty as evidence of a 'chilly climate' for women throughout higher education (Sandler 1986). But the effects of faculty composition on students continue to be debated. Tidball (1980, 1986) finds that the proportion of female faculty is strongly associated with the number of women high achievers, even in coeducational schools. This finding is probably less vulnerable to the lack of institutional controls than are Tidball's findings regarding women's colleges, and is a result that bears further scrutiny with longitudinal data. Rothstein (1995) finds that women students with female advisors are more likely to continue their education after college. Evidence on student satisfaction with same-sex advisors (Erkut and Mokros 1984) and faculty (Ehrenberg et al. 1995) is mixed. These issues require more detailed investigation of particular environments on particular groups of women, such as math and science majors. Sadker and Sadker (1993) make the case for gender bias in the classroom, but this evidence is principally based on research in high schools.

4. *Educational Outcomes*

Much of the discussion on gender inequality in the labor market has been written in response to the writings of the human capital school of economics which holds that gender inequality stems from inadequate investments on the part of women (Jacobs 1995b). As we have seen, this is no longer the case in the United States. England (1992) has noted that working women have surpassed men in median years of schooling completed for much of the century; only during the period since the G. I. Bill have men surged past women. By the 1990s, working women once again caught up with men in average educational attainment (author's estimate, based on the March 1993 Current Population Survey).

Gender differences in earnings persist despite the parity in education attained by women. The median annual income of female college graduates is approximately 70 percent of that earned by their male counterparts (Jacobs 1996). Women earn less than men even with the same level of education. Indeed, the sex gap in earnings within educational levels is similar to that in the labor force as a whole.

The economic benefits of college have increased since the mid-1970s. The gender gap in earnings has narrowed at all educational levels, due in part to the decline in men's earnings (Bernhardt et al. 1995). Yet the gender gap in earnings among college graduates remains similar to that at other educational levels.

A significant portion of the gender gap in earnings can be attributed to gender differences in majors. Majors play a larger role in early career earnings, although they may influence later career earnings indirectly through occupational tracking. Formal schooling does not exhaust the range of possible sources of skill differences between men and women. The gender gap in wages may be due in part to gender differences in skills acquired in on-the-job training and informal experience. Space does not permit an in-depth exploration of all the sources of gender inequality in the labor market. The point here is that the gender gap in earnings in the United States does not stem from the fact that women spend too few years in formal schooling.

Studies find that higher education results in more support for egalitarian gender role attitudes on the part of women, particularly if the students took courses related to women's roles in society (Pascarella and Terenzini, 1991, p. 316). Other authors suggest that education increases women's support for feminism. Freeman, for example, (1975) maintains that education raises women's expectations and creates a sense of relative deprivation, leading them to support feminism.

As noted above, economists distinguish between the private returns to schooling—such as higher wages and higher household income—and the social or public returns, which may involve improvements to health, welfare, and society. Two principal nonmonetary effects of schooling that have been extensively researched are improved infant health and lower fertility. In developing countries, women with more education marry later, are more likely to use contraceptives, desire smaller families, have their first child later (but breastfeed for fewer months), and have a lower total fertility rate than do less educated women (Schultz 1993a, 1993b). Much of the research described by Schultz compares girls with primary education to those with no education in developing countries, but the depressing effect of education on

fertility is a consistent finding in affluent countries as well.

The 1994 United Nations International Conference on Population and Development (1995), held in Cairo, stressed the importance of empowering women in order to promote sustainable economic development and population control. Women's education is a central element in this agenda. Education may thus have significant implications for women's status with respect to gender relations throughout society, but these effects are historically contingent and depend on the character as well as the extent of women's education.

5. Conclusion

I have suggested that access, process, and outcomes are distinct aspects of higher education that need to be examined separately. The trends in these areas often do not coincide with one another, and consequently separate explanations of these facets of higher education are needed. For example, women remain a minority of faculty, and are disadvantaged in terms of rank and institutional prestige. Yet as students in the United States, women represent a majority of students at nearly all levels of higher education, and are not distinctly disadvantaged in terms of institutional position. Clearly, treating women's standing among the faculty and in the study body as one phenomenon will not do, since the extent of women's progress differs between these two statuses.

In my view, the principal challenge facing research on gender in education is to go beyond documenting specific gender effects to developing a more theoretically motivated account of the status of women in the educational system. This perspective would have to account for the relative status of women in each aspect of the educational system, as well as accounting for variation across time and space. The challenge is to situate gender inequality economically, historically, culturally, and politically. The substantial research in various fields on women in education should set the stage for the next generation of researchers to tackle some of the fundamental issues regarding gender and the educational system. In particular, the relationship between gender inequality in education and that in the rest of society is a fundamental question for future theory and research.

See also: Education and Gender: Historical Perspectives; Education (Primary and Secondary Schools) and Gender; Educational Philosophy: Contemporary Views; Educational Philosophy: Historical Perspectives; Family and Gender; Gender and School Learning: Mathematics and Science; Gender and Technology; Gender, Economics of; Gender-related Development; Statistical Systems: Education

Bibliography

Biaecki I, Heyns B 1993 Educational attainment, the status of women, and the private school movement. In: Moghadam V M (ed.) *Democratic Reform and the Position of Women in Transitional Economies*. Clarendon Press, Oxford, UK, pp. 111–34

Bernhardt A, Morris M, Handcock M S 1995 Women's gains or men's losses: A closer look at the shrinking gender gap in earnings. *American Journal of Sociology* **101**(2): 302–28

Chamberlain M K 1988 *Women in Academe: Progress and prospects*. Russell Sage Foundation, New York

Ehrenberg R G, Goldhaber D D, Brewer D J 1995 Do teachers' race, gender and ethnicity matter? Evidence from National Educational Longitudinal Study of 1988 (NELS). *Industrial and Labor Relations Review* **48**(3): 547–61

England P 1992 *Comparable Worth: Theories and evidence*. Aldine de Gruyter, New York

Erkut S, Mokros J R 1984 Professors as models and mentors for college students. *American Educational Research Journal* **21**(2): 399–417

Freeman J 1975 *The Politics of Women's Liberation*. McKay, New York

Goldin C 1997 Career and family: College women look to the past. In: Blau F D, Ehrenberg R G (eds.) *Gender and Family Issues in the Workplace*. Russell Sage Foundation, New York

Graham P A 1978 Expansion and exclusion: A history of women in American higher education. *Signs* **3**(4): 759–73

Jacobs J A 1995a Gender and academic specialties: Trends among college degree recipients during the 1980s. *Sociology of Education* **68**(2): 81–98

Jacobs J A (ed.) 1995b. Introduction. In: Jacobs J A (ed.) *Gender Inequality at Work*, Sage Publications, Thousand Oaks, CA, pp. 1–22

Jacobs J A 1996 Gender inequality and higher education. *Annual Review of Sociology* **22**: 153–85

Jacobs J A 1999 Gender and the stratification of colleges. *Journal of Higher Education* **70**(2): 161–87

Kelly G P, Slaughter S (eds.) 1991 *Women's Higher Education in Comparative Perspective*. Kluwer Academic Publishers, Dordecht. The Netherlands

National Center for Educational Statistics 1997 *Digest of Educational Statistics*. US Government Printing Office, Washington, DC

New York Times 1998 *US Colleges Begin to Ask, Where Have the Men Gone*. December 6, pp.1, 38

Pascarella E T, Terenzini P T 1991 *How College Affects Students*, 1st edn. Jossey-Bass, San Francisco

Ransom M R 1990 Gender segregation by field in higher education. *Research in Higher Education* **31**(5): 477–94

Rothstein D S 1995 Do female faculty influence female students' educational and labor market attainments? *Industrial and Labor Relations Review* **48**(3): 515–30

Sadker M, Sadker D M 1993 *Failing at Fairness: How America's schools cheat girls*. Scribner, New York

Sandler B R 1986 *Campus Climate Revisited: Chilly for women faculty, administrators, and graduate students*. Association of American Colleges, Washington, DC

Schultz T P 1993a Returns to women's education. In: King E M, Hill M A (eds.) *Women's Education in Developing Countries: Barriers, benefits and policies*. Johns Hopkins University Press, Baltimore, pp. 51–99

Schultz T P 1993b Investments in the schooling and health of women and men, quantities and returns. *Journal of Human Resources* **28**: 694–734
Slaughter S 1993 Retrenchment in the 1980s. The politics of prestige and gender. *Journal of Higher Education* **64**(3): 250–82
Tidball M E 1980 Women's colleges and women achievers revisited. *Signs* **5**(3): 504–17
Tidball M E 1986 Baccalaureate origins of recent natural science doctorates. *Journal of Higher Education* **57**(6): 606–20
United Nations Educational, Scientific and Cultural Organization (UNESCO) 1998 *Statistical Yearbook*. UNESCO, Paris
US Bureau of the Census 1975 *Historical Statistics of the United States: Colonial times to 1970*, bicentennial edn. Dept of Commerce, Bureau of Census, Washington, DC

J. A. Jacobs

Education, History of

'History of education' is an ambiguous term in at least two respects. First of all, the term has come to designate different levels of reality. On the one hand, it refers to education as a universal social practice, and to the manifold forms and successive transformations that this practice has experienced over time. On the other hand, it refers to a field of intellectual activity, namely a branch of historical study, devoted to describing, interpreting, and analyzing this practice. Obviously, both levels are not disconnected from each other. Rather, it was a crucial stage in the transformation of the former, namely the crystallization of education into fully institutionalized and increasingly inclusive 'national education systems,' which gave rise to the emergence, not only of educational theory in general, but also of the historical study of education in particular. This historical connection, however, gave rise to yet another ambiguity, an ambiguity inherent in the emergent field of study as such. From its very origins, this field has oscillated between different disciplinary orientations, namely Education and History, i.e., between systemic self-reflection committed to educational ideals and reform, and more detached research committed to the rules of historical method. Against this background, the article briefly describes the process of educational system formation and some of its key features (Sect. 1) in order then to set out the institutional conditions and disciplinary models that have contributed to shaping the history of education into either an adjunct to educational theory (in terms of, e.g., *Geschichte der Pädagogik* or *histoire de la pédagogie*) or into a specialized field of historical study (in terms of, e.g., *Historische Bildungsforschung* or *histoire 'historienne' de l'éducation*) (Sect. 2). The article then goes on to describe in greater detail the important renewal the field has experienced, especially from the 1960s onwards, as a consequence of the influx of concepts and methods stemming from innovations both in social and intellectual history and in the social sciences more generally (Sect. 3).

1. *The Rise of the Modern Educational System*

Although education has occurred, and continues to do so, in manifold forms and at all times—in families, through catechizing, at the workplace, in youth clubs and organizations, or in borstals—education in modern times has by and large become institutionalized in the form of large-scale public systems of education. These constitute the modern type of educational organization *par excellence* (Thomas et al. 1987). Processes of the formation and eventual consolidation of 'the modern educational system,' while starting towards the end of the eighteenth century, came to completion only towards the turn from the nineteenth to the twentieth century (Müller et al. 1987). Moreover, educational system formation was only possible as a component of more encompassing processes of social transformation, and of concomitant processes of state and nation-building, which lead to the specifically modern pattern of societal organization explained as 'functional differentiation' by recent social theory (Luhmann 1982). Drawing on related concepts, one may characterize educational system formation by key features such as *inclusion*, *nationalization*, the system's *internal articulation* as well as *systemic self-reflection*.

Inclusion refers to one of the central aspects of modern social structures, namely the access of a population as a whole to all of the functionally specific subsystems constituted at the level of society. More precisely, inclusion means the universalization, not of so-called 'service roles' such as those of a politician, a physician, or a teacher, but of the corresponding 'service-receiving roles' specific to each of society's subsystems, namely the roles of the electorate, of patients, or of pupils, respectively. Thus, the educational correlate to the inclusion process at the level of society at large was, in terms common to present-day debates on education and development, the 'universalization of primary education.' This process started in the Protestant countries of Central and Northern Europe as early as the seventeenth century; was largely achieved, in most of Western Europe, by the end of the nineteenth century; and came to an end, in some countries of Southern Europe, Latin America, and Asia, only in the middle of the twentieth century. And yet, it still constitutes an unmet challenge in large parts of the Southern Hemisphere. At the level of national legislation, important steps towards universalizing primary education often went hand in hand with measures undertaken to assure greater inclusion also in the political and legal systems. As illustrated by the historical experience, not only of England and

Italy, but in varying forms also of France and Prussia or to some extent Russia, such measures consisted in, e.g., the extension of the franchise or, more generally, steps designed to turn hitherto dependent subjects into national citizens. At the local level, however, national school laws were gradually implemented as an answer to the radically altered communication requirements that were attendant on social structures increasingly characterized by new patterns of differentiation, and ensuing processes of rationalization, mobilization, and industrialization. In this sense, the process of inclusion in the emerging educational system corresponded to the gearing of that system to its actual function, for this function consists precisely in laying the foundations of the capability for universal communication across all subsystems of society. In sociohistorical terms, thus, it was essentially the achievement of universal mass schooling that marked the decisive transition towards the constitution of institutionalized education as a modern 'system.'

Nationalization refers to another process that came to fruition from the eighteenth century onwards. To the extent that the traditional union of schools and Church(es) gave way to the differentiation between education and religion, which were more and more conceived of as distinct spheres of social action and meaning, it was the increasingly rationalized and secularized nation-state that was thought to serve as the natural support and guardian of the education of its citizens. This view was underscored by the fact that the eventual achievement and continuous running of an inclusive educational system presupposed organizational resources and required an administrative infrastructure which no corporate social actor other than the state was deemed able to provide. Alternatively, the state and state-supported agencies were ready to use compulsory mass schooling also for pushing through the realization of more general state-related goals, such as social mobilization and equalization, national integration, and collective identity-building, or political socialization and indoctrination. Thus, whether in Europe, in the Americas, or in Asia, the realization of an all-inclusive educational system by way of universalizing compulsory mass schooling went hand in hand with the increasing intervention of central state-agencies and the building of comprehensive administrative frameworks. Education Ministries were established for securing the financial resources and building up the administrative infrastructure necessary in order not only to guarantee generalized supply, but also to successfully run standardized networks of institutionalized schooling. Just like the realization of compulsory mass schooling, this process took more than a century to come to completion. It was a long way, indeed, from late-eighteenth-century programs of establishing a coherent state-run scheme of educational institutions modeled on Catholic Church organization (in the pioneering Italian principalities of Parma and Piedmont and, subsequently, in Imperial France) to the all-encompassing organization of institutionalized schooling as a 'public legal organism' (as this accomplishment was acclaimed in late-nineteenth-century social analysis).

The emergent system's *internal articulation* was the inevitable complement to inclusion. Since education systems in Europe were not created in a vacuum, their system formation processes started from quite an array of existing institutions which were stemming, on their part, from premodern establishments such as medieval Cathedral schools and Renaissance universities, church-run colleges and academies for the sons and daughters of the nobility, endowed grammar schools and *Gymnasien* established by enlightened princely rulers, as well as municipal petty schools and the nineteenth-century forms of drawing schools, science schools, higher-grade, technical and vocational schools. It was, then, a significant feature of the differentiation process that, during the initial stage of inclusion and due to such origins, the emergent educational systems were rather decentralized, uncoordinated, and open in their organization. The more the continuing process of inclusion advanced, however, the more acute the need for internal specification and rational articulation became apparent. Specification, on the one hand, implied the removal of ambiguities and unregulated competition between institutions that were to become clearly defined alternative tracks or types of schooling. The systems' articulation, on the other hand, raised the issue of connecting hitherto uncoordinated institutional components with the aim of constructing a rational 'sequence,' or 'ladder,' of successive levels of education. The outcomes of such processes of defining educational systems' institutional structures can then be conceived of as ways of coping with divergent expectations and requirements: with inclusion (by organizing universal access to primary schooling for an entire age-group) and selection (by consecutively channeling pupils into different tracks and levels), or with qualifying for universal communication across all fields of society and inculcating differential 'habitus' as appropriate to particular groups, trades, and areas of activity. As a consequence, the issue of defining an educational system's career sequence has always been closely associated with a nation's public policy in general, and with its social and economic policy in particular. This applied especially to the construction of sequence between the 'lower' or vernacular schools (e.g., the so-called 'English' or 'German' schools) and the 'learned' or Latin-based establishments, or between, to take up a graphic phrasing coined by Prost (1968), *l'école du peuple* and *l'école des notables*. Thus, from the late nineteenth century 'systematization' of the major Western European education systems; over the definition, during the first half of the twentieth century, of examination-based transition procedures from the primary to the secondary level of education;

up to the successive introduction, from the 1960s onwards, of comprehensive types of schooling in many, but not all, countries throughout the world, contrasting models and programs for educational sequencing have been at the very core of educational policy debate. At the same time, these models and programs have fostered again and again the production of both educational reflection—particularly reformative reflection—and educational research.

2. *History of Education in Connection with Systemic Self-reflection*

It is a significant feature of modernity that in subsystems of society that have crystallized in connection with the far-reaching transformation processes explained in terms of 'functional differentiation,' system-specific self-descriptions intensify to such an extent that they take the form of explicitly stated reflection theories. This common pattern of self-referential theories *of* a societal subsystem developed *within* that system holds true also for educational theorizing pursued within the emergent educational system (Luhmann 1982). The tremendous increase, from the late eighteenth century onwards, in the production of educational ideologies, educational theories, and educational knowledge more generally has to be seen in connection with the growing acceptance of programs and policies fueling the establishment of state-run 'national systems of education.' In particular, the conceptual model of systemic self-reflection intrinsic to the major subsystems of society makes intelligible the institutionalization of educational theory as a field of academic study in close temporal and ideological connection with the passing of the decisive laws on compulsory schooling in most of the Western European nations. The historically unprecedented scale of universal mass education, its inclusive character, and, especially in the Catholic countries of Southern Europe, its secularized organization not only presupposed an administrative infrastructure of a previously unknown magnitude, but also required the intensification of reflection in new and more sophisticated forms. Moreover, the transition to universal mass education also entailed the need for public legitimization of the new system, not least in the form of throwing into relief distinctive national traditions of education and schooling. Salient examples include the foundation of the *Musée Pédagogique* in France (1879) and of its counterpart in Spain (1882) which were meant to serve as depositories of their respective nation's educational memoirs; the virtual invention of national educational tradition in monumental works such as Gabriel Compayré's *Histoire critique des doctrines de l'éducation en France* (1879) and Ferdinand Buisson's *Dictionnaire de pédagogie* (1887–88); or the launching of the *Monumenta Germaniae Paedagogica* meant to collect 'the materials, scattered over archives and libraries, that are relevant to the history of German education and schooling' (1886–1938).

In addition, the necessities of systematic teacher training, and attendant considerations of social utility and responsibility, impinged heavily on the emergent forms of the historical interpretation of educational traditions. These necessities and considerations favored a genre of educational history that remained subordinated to the categories and problematiques of systematic—i.e., philosophico-reflective—theorizing, a genre in other words which gave preference to narratives of the great educators and philosophers of the past, and to interpretations of their respective doctrines realized in terms of normative hermeneutics. Interestingly enough, this genre has been common to most European countries. It can be traced from early British works over, for instance, Dutch and Scandinavian examples as well as the impressive range of, in particular, German and Italian histories of educational thought up to recent specimens in quite a number of countries (Heinemann 1979–85). This genre as well as a good many of the studies commonly assigned to the history of education have aimed at producing—and in fact continue to produce—not primarily historical research in the manner of historians, but, instead, justification and legitimization, the construction of enlightened norm consensuses and the reinterpretation and reactualization, with respect to issues of present-day educational reform, of the normative and theoretical substance contained in historical texts, models, or institutions. In other words, this tradition of an 'educational' history of education—of *Geschichte der Pädagogik* or *histoire de la pédagogie*—has not aimed at a detached historicization of lines of educational thought or of past experiences, but at certain forms of theoretical and normative 'traditionalization.'

Thus, seen from the vantagepoint of sociohistorical analysis, the history of education closely associated, in terms of institutional and disciplinary affiliation, with educational theory has from its origins shared the ambiguities proper to the latter field. Just as educational studies in general, history of education has oscillated between rigorous research and systemic self-reflection. Accordingly, the field has given room to a considerable variety of constructions of reality, the varying presuppositions of which ensue from their being related to distinct system logics and problem perspectives: to either issues of objective method and explanation as debated from the vantage point of social science research, or to issues of meliorism and reform immanent in the dynamics of modern education, and to attendant needs for interpretation, orientation, and professional knowledge. The respective proportions, however, in which these component modes of reasoning have given shape to historical studies on education arise from an intricate interplay of institutional conditions, intellectual traditions, and disciplinary models (Schriewer et al. 1993). Thus,

recent analyses of the state of the art have—on the grounds of reviews of substantive interests, research topics, and methodological approaches that are given preference to as well as of the field's disciplinary status, specialized journals, and associations—contrasted two major models. These are, on the one hand, the German model, which is well represented also in countries such as Switzerland, Austria, and Hungary as well as in Italy and Scandinavia, and, on the other hand, the institutional-*cum*-epistemological model prevalent in France and the United Kingdom, respectively (Compère 1995). In the former model, due to a long-standing tradition of educational-*cum*-philosophical theorizing as well as to the institutional and intellectual integration of the history of education into education as a discipline firmly established in academia, the system-reflection style of educational history described above has by and large prevailed up to the 1970s, and still has a strong impact up to the present. By contrast, the tradition of education conceived as a philosophico-hermeneutic field of theorizing gave way to educational studies based on the social and human sciences both in France, from the 1930s onwards, and in the United Kingdom, in the aftermath of World War II. Consequently, these developments also facilitated the alignment of historical studies on education with history and sociology proper, and the intellectual opening of such studies to wider sociological and anthropological frames of reference. Thus, in France and the United Kingdom, just as in the United States, history of education has turned into an academic discipline in its own right which is fairly independent of educational studies in general, based on its own specialized journals, and elaborated by scholars who are (also) trained as historians. This *histoire 'historienne' de l'éducation* has in turn encouraged initiatives leading to the emergence of a social history-related *Historische Bildungsforschung* also in countries traditionally influenced by the German model.

3. *History of Education as a Field of Social Science Study*

Fueled by debates on a 'new history of education' which arose in the United States at the beginning of the 1960s, conceptual models and methodological approaches heavily influenced by modern social history (see *Social History*), intellectual history (see *Intellectual History*), and the social sciences more generally have since then considerably altered the field. The publication, in 1960, of Philippe Ariès' *L'enfant et la Vie Familiale sous l'Ancien Régime* and Bernard Bailyn's *Education and the Forming of American Society* can be seen as a turning point. A stimulating and intellectually demanding period was about to begin. From the 1960s to the 1980s, the 'new history of education' brought a critical perspective into a discipline that had been dominated by parochialism and myths of progress. Ever since its repertoire has come to include the *longue durée*, a focus on structures and mentalities, a special attention to serial and quantitative methods, and the ambition of 'total history.' These changes, moreover, are reflected in corresponding shifts from narrative history to problem-centered history, from descriptive to explanatory approaches, and from conceptions of unilinear evolution to a vision of historical processes that emphasizes their often conflict-laden, asynchronous nature. While accomplishing such shifts, scholars like Michael Katz and David Tyack, Antoine Prost and Dominique Julia, and Lawrence Stone and Harold Silver have been deeply influenced by social history perspectives. They have turned the legitimizing narratives of the past upside down, providing new arguments about the emergence and development of educational systems, the role of schooling in social reproduction, the educational strategies of various social classes, and the action taken by the nation-state in establishing mass schooling. They accordingly shifted the attention from the evolution of educational ideas to the relationships between education and society.

While in the early 1980s a sense of accomplishment was shared by a majority of the researchers involved in rewriting the history of education, only a few years later, the resources of the 'new' history of education seemed to be exhausted. The accumulation of data, the considerable increase in well-conducted studies based on extended series of archival sources, and the growing specialization of historical inquiry nevertheless left the impression that educational historians had fallen short of addressing actually meaningful issues. In addition, efforts meant to raise new questions, related for instance to minority and gender issues, did not lead to an epistemological breakthrough, but rather to an even wider fragmentation of the field. Innovative solutions were to be found, then, in the adoption of theoretical frameworks and methodological approaches that focus not only on the organization of practices and discourses over time, but also on the elucidation of how these practices and discourses have constructed and reconstructed the individual lives of people and, thereby, social reality. In this sense, educational historians are facing the same challenges and dilemmas like social scientists more generally. To the extent that the difficulties of coping with the plurality of meanings henceforth present in the analysis of social phenomena became apparent, the 'new' complexity of historical work has nourished old suspicion as far as theory is concerned (Lagemann 2000). The transition from 'History' to 'histories' implies a reconfiguration of the field that invites educational historians to reconceptualize their work and to adopt a more comprehensive approach, shifting from a social history of culture to a cultural history of society (Chartier 1998). Still, there is no consensus regarding the changes that are occurring in historical

studies in education (Cohen and Depaepe 1996). Nevertheless, drawing on an often used conceptual division between educational actors, school practices, educational ideas, and educational systems, it is possible to identify major tendencies that have come to shape historical research in education (Nóvoa 1998).

3.1 The History of Educational Actors and the Rediscovery of Experience

For a long time, historians have described the history of educational institutions rather than that of educational processes or interactions. Consequently, the history of educational actors has been one of the most neglected areas of research. The rediscovery of the concept of 'experience' opened new possibilities, well illustrated by numerous texts on autobiographies, teachers' professional careers, or students' life stories. One of the reasons why feminist theories have been so powerful in educational history is that they help to reveal histories silenced by traditional narratives. From these points of view, modernity dispossessed the actors of the educational process from their subjectivity through the imposition of a structural logic and a populational reasoning. People were classified as categories (the teachers, the learners, the children at risk, etc.) and managed as populations. However, after a long tradition of conceptualizing the world in terms of structure and representation, time had come to see the world as 'experience.' This implied a new 'epistemology of the subject' which, by dissolving the 'natural' divisions between public and private, discourse and practice, or between text and context, prepared the way for a shift toward a self-conscious study of pupils and teachers, and for bringing the individual actor in education back into the historical portrayal. Moreover, new conceptions of the subject and of experience allow attention to be paid not only to living experience, but also to the way people elaborate it. Finally, the concept of 'experience' has to be seen not only at an individual level, but also in its collective dimension, which is characterized by a redefinition of identities and belongings through the participation of people in different communities of meaning.

3.2 The History of School Practices and a New Concept of Culture

The history of school practices is another field that has traditionally been neglected by educational historians. Explanations of this 'black box' phenomenon are directly related to the concept of 'culture.' Schools have always been regarded as places of culture: at first, in an idealistic vision of putatively universal values and knowledge to be transmitted to the young generations; afterwards, in a critical perspective of ideological inculcation and social reproduction. In both cases, educational historians failed to look at the internal production of a distinctive school culture that is not independent of social struggles and conflicts, but cannot be explained as being merely determined by the external world either. More recent conceptualizations, in contrast, have defined school culture as both an ensemble of norms (which organize the branches of knowledge and behaviors that must be brought into schooling) and an ensemble of practices (which allow the transmission of such knowledge and the integration of such behaviors). Moreover, it is possible to identify tendencies of historical research that successfully try to refocus historical inquiry onto the internal functioning of the school by providing thoroughgoing analyses of organizational structures, curriculum issues, school subjects, and day-to-day realities. Thus, one of the most interesting lines of work tries to understand the permanence of a 'grammar of schooling' that resists changes and reforms. Another line of research focuses on the history of the curriculum, viewed as the outcome of a complex interplay of competing values and traditions. Several authors are voicing the need for deconstructing a 'natural-history' conception of the curriculum, and therefore seek to problematize the curriculum as text and historicize its evolution over time. A third strand of recent research worth mentioning is concerned with the history of school subjects. Historians are asking salient questions regarding the formation of school subjects, the institutionalization of school knowledge, and the ways in which it helps to organize our understanding of the world. Finally, one should not fail to point out a research trend concerned with the 'silence' observed within the traditional history of education about meaning and culture of classrooms, or the everyday education practice (Grosvenor et al. 1999, Ruiz Berrio 2000).

3.3 From the History of Educational Ideas to a Sociohistory of Knowledge

In contrast to traditional histories of educational ideas, the 'new' cultural history of education is interested in the production, diffusion, and reception of educational discourses throughout time and space. In other words, it is interested in 'a historical imagination in the study of schooling that focuses on knowledge as a field of cultural practice and cultural production' (Popkewitz et al. 2001). Therefore, the research interest has shifted to the discursive practices that regulate schooling, especially in situations of conflict or rupture. In this sense, several authors are calling for a historiography, not of the general terms, but of the conceptual structures that determine edu-

cational discourses, and for analyses, not of prominent ideas, but of fields of historically defined discursive problematiques. Such an approach is also important for analyses of the roles played by educational experts in providing 'legitimate' or 'scientific' interpretations of school realities. It helps to understand the original institutionalization of educational studies, not as a narrative of intellectual progress, but as a form of 'governmentality.' The question is not how to identify different ideas, but how to apprehend the rise of expert systems of knowledge (Drewek et al. 1998). The same holds true for the analysis of the complex transformation procedures that change academic subjects (integrated into their distinctive conceptual spaces) into school subjects (integrated into the arena of schooling), or for the study of the production and diffusion of textbooks that are not simply 'delivery systems' of factual information, but the results of cultural constructions imposing particular ways of seeing and interpreting the world. In this sense, knowledge has increasingly come to be taken as a text that redefines subjectivity and identities, that elaborates rules and behaviors, and that configures meanings and beliefs.

3.4 Comparative-historical Research on Education

The publication of Fritz Ringer's *Education and Society in Modern Europe* (1979) heralded a virtual breakthrough of yet another paradigmatic approach to the field, namely the comparative-historical research on education. Although some forms of cross-national analysis or international account had not completely been missing in previous periods, the programmatic views expressed by Ringer laid claim to new and ambitious objectives: 'There is simply no other means of arriving at explanations, and not just descriptions, of change in education than the comparative approach.' Substantive research inspired by such views has remarkably increased from the 1970s onwards. It covers, for instance, quasi-experimental investigations of the relationships between education and economic growth; large-scale theory-based analyses of the varying processes of educational system formation; as well as cross-national analyses of the emergence and varying discipline formation patterns of education as an academic field of study in the context of European university systems and the emerging social sciences more generally. Comparative-historical research has also shed new light on the role that nation-specific social meanings, widely accepted interpretative schemes, or 'symbolic structures' deeply rooted in a particular civilization have played in shaping distinctive academic cultures and educational institutions (Archer 1979, Müller et al. 1987, Schriewer et al. 1993).

Not always, however, have attempts at providing convincing explanation through the combination of historical research and cross-national, or cross-civilizational, analysis been successful. Consequently, much debate has been devoted to issues of explanation (see *Comparative History*). While traditionally, and to some extent up to the present, mainstream models of explanation have followed lines of reasoning as defined in the orthodox philosophy of science, far-reaching shifts in both problem-awareness and conceptual models have taken place in the wake of recent theory developments (Schriewer 2000). Among these are, for instance, world-system models and analyses which, by drawing on Braudelian concepts, not only reintroduced a macrohistorical perspective into comparative research, but also tended to replace the latter by analyses of global diffusion processes and of the working of an emerging 'world polity' (Thomas et al. 1987). New forms of reconciling history and comparison are also suggested, however, by insights from historical sociology as well as by interdisciplinary developments discussed under headings such as 'self-organization,' 'morphogenesis,' or 'evolutionary complex systems research' (Gulbenkian Commission 1996). Thus, comparative-historical research on education informed by concepts and models of this kind seems to be particularly suited, not only to radically historicizing educational phenomena, but also to linking historical analysis up with insight relevant to social theory.

See also: Education and Economic Growth; Education: Anthropological Aspects; Education, Economics of; Education: Phenomena, Concepts, and Theories; Education (Primary and Secondary Schools) and Gender; Educational Institutions and Society; Educational Institutions, History of; Educational Philosophy: Historical Perspectives; Life Course in History; Nationalism, Historical Aspects of: The West

Bibliography

Archer M S 1979 *Social Origins of Educational Systems*. Sage, Beverly Hills, CA

Ariès P *L'Éducation et la Vie Familiale sous l'Ancien Régime*. Plon, Paris

Bailyn B 1960 *Education and the Forming of American Society—Needs and Opportunities for Study*. Vintage Books, New York

Buisson F (ed.) 1887–88 *Dictionnaire de Pédagogie et D'Instruction Primaire*. 2 parts in 5 Vols. Hachette, Paris

Chartier R 1998 *Au Bord de la Falaise: L'histoire Entre Certitudes et Inquiétude*. Albin Michel, Paris

Cohen S, Depaepe M (eds.) 1996 *History of Education in the Postmodern Era* (*Paedagogica Historica* Vol. 32 no. 2). C.S.H.P., Ghent, Belgium

Compayré G 1879 *Histoire Critique des Doctrines de L'Éducation en France depuis le Seizième Siècle*. Hachette, Paris [Reprint edition 1970. Slatkine, Geneva]

Compère M M 1995 *L'Histoire de l'Éducation en Europe*. Peter Lang, Bern, Switzerland

Drewek P, Lüth C, Aldrich R, Scholtz H, Schriewer J, Tenorth H E (eds.) 1998 *History of Educational Studies. Geschichte der Erziehungswissenschaft. Histoire des Sciences de l'éducation* (*Paedagogica Historica*, Supplementary Series, Vol. III). C.S.H.P., Ghent, Belgium

Grosvenor I, Lawn M, Rousmaniere K (eds.) 1999 *Silences and Images. The Social History of the Classroom*. Peter Lang, New York

Gulbenkian Commission 1996 *Open the Social Sciences. Report on the Gulbenkian Restructuring of the Social Sciences*. Stanford University Press, Stanford, CA

Heinemann M (ed.) 1979–85 *Die historische Pädagogik in Europa und den USA*. 2 Vols. Klett-Cotta, Stuttgart, Germany

Lagemann E C 2000 *An Elusive Science: The Troubling History of Education Research*. University of Chicago Press, Chicago

Luhmann N 1982 *The Differentiation of Society*. Translated by S. Holmes and C. Larmore. Columbia University Press, New York

Monumenta Germaniae Paedagogica. Schulordnungen, Schulbücher und pädagogische Miscellaneen aus den Landen deutscher Zunge, 62 Vols. 1886–1938. Launched by Kehrbach K, edited by Gesellschaft für deutsche Erziehungs- und Schulgeschichte. Weidmann, Berlin

Müller D K, Ringer F K, Simon B (eds.) 1987 *The Rise of the Modern Educational System. Structural Change and Social Reproduction 1870–1920*. Cambridge University Press, Cambridge, UK, and Editions de la Maison des Sciences de l'Homme, Paris

Nóvoa A 1998 *Histoire et Comparaison. Essais sur l'Éducation*. Educa, Lisbon, Portugal

Popkewitz T, Franklin B, Pereyra M (eds.) 2001 *Cultural History and Education: Critical Essays on Knowledge and Schooling*. Routledge, New York

Prost A 1968 *Histoire de l'Enseignement en France 1800–1967*. Armand Colin, Paris

Ringer F K 1979 *Education and Society in Modern Europe*. Indiana University Press, Bloomington, IN

Ruiz Berrio J (ed.) 2000 *La Cultura Escolar de Europa. Tendencias Históricas Emergentes*. Biblioteca Nueva, Madrid, Spain

Schriewer J (ed.) 2000 *Discourse Formation in Comparative Education*. Peter Lang, New York

Schriewer J, Keiner E, Charle C (eds.) 1993 *Sozialer Raum und Akademische Kulturen/A la recherche de l'espace universitaire européen*. Peter Lang, New York

Thomas G M, Meyer J W, Ramirez F O 1987 *Institutional Structure. Constituting State, Society, and the Individual*. Sage, Newbury Park, CA

J. Schriewer and A. Nóvoa

Education in Old Age, Psychology of

1. Introduction

From an educational perspective lifespan development can be described as a continuous and active process of coping with developmental tasks, i.e., demands, challenges, and chances that depend on people's environment and life situation in given phases of the aging process. Specific developmental tasks are conceptualized as a consequence of the interaction between biological maturity, normative conceptions of 'successful' aging or development in society and individual plans, aims, needs, and values. The significance of environment and life situation for lifespan development is twofold: first, coping successfully with demands of the environment and life situation increases the potential for coping with future demands and initiates further development; second, chances offered by environment and life situation contribute to the realization of specific developmental gains.

The here intended understanding of development as a process of personal growth (focusing on stable, not only temporary gains) is connected closely to a concept of education that implies two meanings: the process of being engaged in educational activities (person educating himself) and the result of this engagement (the educated person). Education does not only lead to the establishment of differentiated knowledge systems, but also facilitates a broader understanding of own experience and action; two developmental gains of high significance.

The insight that developmental gains are possible even in very old age is essential for lifespan developmental psychology. Human development is to be understood as a lifelong process characterized by multidimensionality and multidirectionality (Baltes 1987). Aging implies different processes of change in different dimensions of the person; in each dimension gains and losses can be observed simultaneously; changes in one dimension are a poor predictor of changes in other dimensions. Accumulated and organized experiences, knowledge systems, and strategies for coping effectively with familiar problems and tasks are important developmental gains (in the sense of psychological growth) in older adulthood; important developmental losses do occur in physiological and neurophysiological functions and processes, c.f. cognitive flexibility, coping with unfamiliar problems and tasks, and speed of information processing (which itself decrease performance of short-term memory) (for an overview see Kruse and Rudinger 1997).

From the perspective of this article, education is a precondition for reaching developmental gains as well as for the opportunity to compensate for developmental losses (e.g., via the process of selective optimization with compensation) and for individual abilities to maintain or re-establish a personal satifying perspective on life in old age.

The first part of this article focuses on the importance of lifespan developmental psychology for understanding education in old age. In the following parts the competence-maintaining and competence-enhancing functions of education are discussed in the context of three topics: life competencies of the old as a human capital for society, productive life in old age, and potentials for development in old age.

2. Life Span Developmental Psychology and Education

Four conclusions for the understanding of education in old age can be drawn from the results of lifespan developmental psychology.

(a) Accumulated and organized experiences, differentiated knowledge systems as well as effective strategies for coping with tasks and problems and for taking the chances offered by environmental conditions and life situation must be understood as a consequence of education processes in earlier phases of the lifespan, i.e., active coping with tasks, demands, challenges, and chances in younger years. In this context the considerable continuity of lifelong development becomes clear, since achievements in earlier years extent to later years as long as they are used sufficiently.

(b) The use and further improvement of experiences, knowledge, and strategies for effective action is essential for the realization of developmental gains in old age. This fact points to the importance of education in old age. Generally, there is a sufficient amount of plasticity, i.e., capacity to learn and change. As a consequence, education in old age, aiming to contribute to the use and further improvement of experiences, knowledge, and strategies for effective action, should be able to contribute to effective coping with developmental tasks and 'successful' aging. From our perspective, there is a demand for education in old age, since people cannot cope successfully with developmental tasks or realize the potentials offered by cultural innovation (e.g., in the technical field) in old age without engaging in a process of coping with self, life situation, and environment.

(c) Education in old age must not be analyzed only in terms of possible gains, depending on development in earlier phases of the lifespan. Possible developmental losses are of equal importance here. In this context, two aspects should be considered: first, gerontological research could show that cognitive training has an impact on the decline in the 'mechanics' of human intelligence, i.e., the basic processes underlying intellectual performance (Kliegl et al. 1989). Moreover, through cognitive training the compensation of specific impairments and losses can be encouraged. Intellectual performance and everyday competence can be improved by new learning and memory strategies that compensate at least in part for deficits in sensory and motor functions. The 'selective optimization with compensation,' i.e., the choice (selection) of well-developed functions, skills, and strategies and their differentiation (optimization) with the aim of compensating for losses in other functions, skills, and strategies forms an important basis of 'successful aging' (Baltes and Baltes 1990).

(d) In the context of possible developmental losses—especially in very old age—decreases in health status and social integration as well as limitations of life expectancy should be mentioned. However, empirical findings do show that even in very old age most people are able to cope effectively with developmental losses and to maintain or re-establish a personal satisfying perspective on their life, pointing to resilience in old age. It should be noted that resilience is established through personal and environmental resources (for an overview see Kruse and Wahl 1999, Staudinger and Fleeson 1996). Concerning environmental resources, gerontological studies show that the opportunity to participate in cultural and social activities that offer stimulation and support promotes effective coping with developmental losses. These findings can be interpreted as support for an educational perspective that points to the significance of education for coping with critical life events, e.g., via providing information about effective coping strategies or institutional services. In the case of definite losses in certain functions and skills (e.g., visual impairment as a result of macular degeneration), compensation (e.g., use of cassettes or touch-typing) as well as increased engagement in interests and activities where these functions and skills are not necessary can be interpreted in two ways: first, as a process of education, second, as encouraged by educational activities. Several intervention studies show the possible impact of educational opportunities on the effectiveness of individual coping processes (see the overview in Wahl et al. 2000).

From the here intended understanding of education and its effects, the close relationship between education and competence becomes clear. The latter is defined as the skills and abilities to maintain or re-establish an independent, task-oriented, and personally meaningful life in a stimulating, supportive (physical, social, and infrastructural) environment, encouraging self-responsible coping with problems, developmental tasks, and challenges (Kruse 1995).

In old age education has competence-maintaining and competence-enhancing functions, i.e., the rationale of providing educational programs and services for the old is to offer opportunities for using and further improving competencies. In the following, our discussion will concentrate on three topics that proved to be important for gerontological research as well as for societal perception and valuation of old age: (a) life competencies of the old as a human capital for society; (b) productive life in old age; (c) potentials for development in old age.

3. Life Competencies of the Old as a Human Capital for Society

The term life competencies refers to experiences, strategies, and knowledge systems that people have acquired in earlier phases of the lifespan (i.e., our understanding of education as a result of active coping

with the challenges and demands of the environment and life situation in a given period of life). Life competencies are built up in the context of effective coping and enable people to maintain or re-establish a personal satisfying perspective on their life when confronted with serious problems, tasks, and challenges in later years. Building up life competencies in earlier years is a basic requirement for successful development in advanced age, i.e., effective coping with demands of life in old age. Such demands include practical and psychological as well as interpersonal and ethical demands. Consequently, our understanding of life competencies is not limited to the experiences, strategies, and knowledge systems acquired in the context of occupational activities.

Life competencies are also reflected in ethical judgements, voluntary activities for other people, as well as in the willingness and readiness to take responsibility for oneself, for others, or for society. In this context it should be mentioned that empirical findings show that active coping with developmental tasks and the chances and limits of life can lead to the establishment of 'expert knowledge' or 'wisdom' with respect to questions of life (see Smith and Baltes 1990). Expert knowledge or wisdom is not limited to old age, it can be developed in earlier years as well. The only prerequisite for the development of expert knowledge or wisdom is the conscious and responsible preoccupation with a multitude of problems, tasks, and challenges in different periods of life and different contexts of development. Therefore empirical findings support a perspective on lifelong educational processes that accentuates the development of the person rather than the cultivation of specific knowledge and skills (Kruse 1997).

With the term human capital we refer to the significance of life competencies for society and culture, i.e., processes of initiating societal and cultural change and the extent to which societal and cultural change is determined by life competencies of the old. Societal and cultural development depends on the possibility to use individual life competencies. It is here necessary to establish the respective infrastructural conditions, e.g., opportunities for volunteers to qualify and to use given experiences, strategies, and knowledge systems effectively.

The contribution of specific educational services to the realization and use of life competencies of the old, i.e., the conversion of individual life competencies into human capital, is of interest for educational research and practice. In a later part of this article we revert to educational services in which (a) an intergenerational perspective is put forward (i.e., members of different generations are addressed) and (b) individual experiences, strategies, and knowledge systems are made explicit and used by stimulating a dialogue or discourse. At this point we are content with stating that, for example, stimulating an exchange between the young and the old by inviting older people to visit schools, to participate in specific lessons (e.g., on history), and to engage in the transmission of experiences and knowledge is appreciated equally by the young and the old. In establishing intergenerational contact which enables people to realize and use specific life competencies, the challenge for educational practice is twofold: first, experiences, strategies, and knowledge systems must be explored and integrated systematically; second, discussion or discourse must be initiated and guided with methodological competence.

4. *Productive Life in Old Age*

Even if productivity is commonly associated with the making of things (production) and material values, and—consequently—speaking of productive life in old age generally refers to societal productivity, i.e., enrichment, promotion, maintenance, and relief of social systems, the term productivity is neither restricted to participation in the labor market or voluntary activities nor confined to manual expression (Baltes and Montada 1996). Instead, an adequate definition of productivity must also consider intellectual, emotional, and motivational expressions of productivity. Using an extended definition of productivity, several possibilities of leading a productive life in old age can be distinguished: being interested in development, living conditions, and vital interests of younger people, the transmission of information to younger generations and the (self-)responsible reflection of experiences and knowledge systems of younger generations are examples of intellectual and emotional productivity in old age, since intergenerational discourses can initiate emotional and intellectual differentiation in older and younger participants. Moreover, by leading an independent and responsible life, even when confronted with serious problems or borderline situations, older people can give a good example of how to cope with problems and difficulties to younger people, i.e., build a productive context for others: an illustrative example of emotional and motivational aspects of productivity (see Kruse 1995, Baltes and Montada 1996).

From this context it becomes apparent that even the dialogue between young and old can be understood as an educational process in which young and old people can enrich each other by exchanging specific experiences and knowledge as well as general life competencies (see examples in Wahl et al. 2000). For our understanding of education in old age, two conclusions can be drawn: first, older people are able to fulfill educational functions for others by offering skills, experiences, and knowledge or by leading an independent and responsible life which can be regarded as exemplary for younger generations. However, it should be mentioned that for fulfilling educational functions for others it is essential that the transmission

of experience and knowledge is free from prejudiced beliefs which presuppose that own knowledge is superior to knowledge of other people and that—therefore—others have to be instructed to adopt their experiences and knowledge systems. Second, dialogue and discourse with younger people can initiate further development. In our view, intergenerational discourse is an important means to realize potentials for development. From the reports of services and institutions concerned with education in old age, it seems obvious that intergenerational encounters—either informal or in the context of formal educational settings—do have significant impact on developmental processes in the young and in the old and therefore should be promoted intensively.

5. *Potentials for Development in Old Age*

Cognitive training programs and programs for maintaining and improving everyday competence (i.e., skills and abilities for leading an independent and responsible life and for coping effectively with problems, tasks, and challenges of environment and current life situation) have proven to be effective in numerous empirical studies (for a review see Wahl and Tesch-Römer 1998). Findings do indicate an apparent cognitive and behavioral plasticity which can be explained by functional and morphological changes in neuronal networks. Functional plasticity refers to the increased resting potential of neurons (on the basis of which cellular excitation processes can be speeded up), morphological plasticity refers to the production of new synapses and enrichment of dendrite growth in continuous activation processes (Kruse and Rudinger 1997). However, if there is a lack of continuous stimulation in sensory, sensorimotor, and cognitive functions, the probability of losses in cognitive and everyday competence—especially in very old age—increases dramatically. This special risk is caused by increased vulnerability of physiological and neurophysiological systems in old age. Baltes (1984) refers to developmental contexts characterized by a lack of stimulation and training of functions as 'intelligence-decelerating' environments.

Losses resulting from insufficient stimulation and training can be traced back to a loosening of neuronal switch-off mechanisms, to reduced capacity of the synapses, and to a reduced number of dendrites. For this reason the statement can be made that a certain amount of stimulation and activation as well as a certain amount of training cognitive and everyday competencies is a necessary condition for maintaining an independent and responsible life in old age. In this context it should be mentioned that even an independent, responsible, and personally meaningful organization of daily activities alone can offer a sufficient amount of stimulation, activity, and training of cognitive and everyday competencies; this leads us back to the former statement that informal educational activities (such as continuous use) contribute to education in old age. Therefore intervention measures for maintaining and enhancing competencies in old age should not only be restricted to the improvement of individual skills and abilities but also consider the importance of environmental conditions for the relationship between competencies and actual performance.

See also: Adult Development, Psychology of; Adulthood: Developmental Tasks and Critical Life Events; Education and Learning: Lifespan Perspectives; Life Management, Developmental Psychology of; Lifelong Learning and its Support with New Media: Cultural Concerns; Lifespan Development, Theory of; Personality Development in Adulthood

Bibliography

Baltes P B 1984 Intelligenz im Alter (Intelligence in old age). *Spektrum der Wissenschaft* **5**: 46–60

Baltes P B 1987 Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* **23**: 611–26

Baltes P B, Baltes M M 1990 Psychological perspectives on sucessful aging: The model of selective optimization with compensation. In: Baltes P B, Baltes M M (eds.) *Sucessful Aging*. Cambridge University Press, New York

Baltes M M, Montada L (eds.) 1996 *Produktives Leben im Alter* (*Productive Life in Old Age*). Campus, Frankfurt, Germany

Kliegl R, Smith J, Baltes P B 1989 Testing-the-Limits and the study of adult age differences in cognitive plasticity of mnemonic skill. *Developmental Psychology* **25**: 247–56

Kruse A 1995 Entwicklungspotentialität im Alter. Eine lebenslauf- und situationsorientierte Sicht psychischer Entwicklung (Potentials for development in old age. A life-span and situation perspective on psychic development). In: Borscheid P (ed.) *Alter und Gesellschaft*. Wissenschaftliche Verlagsgesellschaft, Stuttgart

Kruse A 1997 Bildung und Bildungsmotivation im Erwachsenenalter (Education and motivation for education in adulthood). In: Weinert F E, Mandl H (eds.) *Psychologie der Erwachsenenbildung*. Hogrefe, Göttingen

Kruse A, Rudinger G 1997 Lernen und Leistung im Erwachsenenalter (Learning and performance in adulthood). In: Weinert F E, Mandl H (eds.) *Psychologie der Erwachsenenbildung*. Hogrefe, Göttingen

Kruse A, Wahl H-W 1999 Persönlichkeitsentwicklung im Alter (Personality development in old age). *Zeitschrift für Gerontologie* **32**: 279–93

Smith J, Baltes P B 1990 Wisdom-related knowledge: Age/cohort differences in response to life planning problems. *Developmental Psychology* **26**: 494–505

Staudinger U, Fleeson W 1996 Self and personality in old and very old age: A sample case of resilience? *Development and Psychopathology* **8**: 867–85

Wahl H-W, Kruse A, Rott C, Schmitt E 2000 On the application of geropsychological knowledge. General issues and results of a community-based evaluation study. In: Kantas A, Velli T,

Hantzi A (eds.) *Societally Significant Applications of Psychological Knowledge*. Ellinika Grammata, Athens
Wahl H-W, Tesch-Römer C 1998 Interventionsgerontologie im deutschsprachigen Raum: Eine sozial- und verhaltenswissenschaftliche Bestandsaufnahme (Intervention gerontology in German speaking countries: A review of social and behavioral research). *Zeitschrift für Gerontologie und Geriatrie* **31**: 76–88

A. Kruse and E. Schmitt

Education in Psychiatry

1. Introduction: Medical Education in General

Medical education worldwide is in a state of creative ferment, greater now than at any time since the early twentieth century. (The *Flexner Report* of 1910 had initiated groundbreaking reforms.) International consensus, reached by intensive global consultation conducted by the World Federation for Medical Education, is expressed in the *Edinburgh Declaration* (1988) and the Recommendations of the *World Summit on Medical Education* (World Federation for Medical Education 1994). These global mandates have been instruments for change worldwide, and the major impetus for reform.

Medical education currently is the driving force in many sectors of professional administration in medicine. To cite a few examples: postgraduate Deans administer salaries of doctors in postgraduate training; physicians in training are posted only to academically accredited hospitals; and many countries require continuing medical education credits for doctors to hold hospital appointments.

The contemporary managerial revolution in health care and the medicine-for-profit ideology have imposed an unanticipated educational limitation. Since ancient Greece the vocation of Medicine has been hallowed as a calling; to foster learning has always been esteemed. Now even born teachers can have their investment in teaching curtailed by their institution, if their educational zeal is judged to intrude on the potential for generating income through clinical work.

1.1 Reform and Stasis

Efforts to reform medical school curricula commonly fail because, as medical education research demonstrates repeatedly, the institution omits to change the committee structure (governance) of the medical school to be congruent with new educational goals. A curriculum left in the hands of departmental heads inevitably remains static. A Curriculum Committee, given autonomy and its own resources such as budgets, answerable to the Dean and not to departments, is essential. Departments are then still free to pursue their specific noneducational goals, such as research and clinical work, to which in any case they give greater priority (Abrahamson 1996).

Medical education has long derived crucial research impetus from the social sciences. Pioneering sociologists, most fortunately for medicine, elected to study medical education (Merton et al. 1957) when they started investigating the professions to determine why professionals do not exploit their clients more than is the case. The classic New York study at Cornell, *The Student-Physician*, followed by the investigation of Kansas medical school by Chicago sociologists, led to the conspicuous leadership which (Becker et al. 1961) medicine has provided in higher education generally. Sociologists, by studying the medical school as a social organization (Bloom 1988), have also explored obstruction to educational innovation. The resistances to curricular change in medical schools derive from apathy on the part of some medical teachers; the necessary critical mass of teaching staff, educationally informed, interested, and involved in the education of medical students, is not achieved. Medical teachers, in common with their counterparts in the rest of academia, commonly assume they are equipped to teach simply by virtue of their expertise in their research or specialist discipline: they are ignorant of the medical education literature, and do not know or respect the findings of medical education research (Miller 1980). They have very often not had personal instruction about how to teach.

Medical students learn best, they repeatedly convey when surveyed, when active student participation is encouraged; when the emphasis is on applied problem-solving (rather than learning of factual materials); when instruction is student-centered; and when teachers are humanistically oriented.

The shibboleth that everybody must learn the same syllabus has proved amazingly resilient. Emphasis on memorization of fact, resulting in information overload, has been the age-old obstacle to active learning. Teaching and learning methods are being revolutionized by electronic information technologies, problem-based learning, and simulated (standardized) patients. Scope is being given for critical thinking, and the development of clinical competence made a priority. Time has been freed for special interest options of students. Educational leadership is essential, often not provided by Deans deflected by administrative priorities; and students have to be valued partners as change agents at every level, participating actively in monitoring the quality of their education.

1.2 The Applicant Pool and Selection

There is now considerable attention being given to widening the range of entrants to medical school, to avoid discrimination according to gender (male and female entrants are now equal numerically), ethnicity, social class, etc. Applying to enter medical school,

given the great social desirability of the profession, is much documented and potential entrants are assiduously briefed and groomed (Iserson 1997).

Selection procedures have been investigated intensively. Entry criteria have centered on cognitive abilities unduly (e.g., school-leaving examination results), with nonintellectual attributes (e.g., motivation, empathy, humanism, drive, and determination) neglected. Once admitted, the student in a good medical school is responsibly respected and fostered (Merton et al. 1957, Coombs 1998).

2. *Basic Medical Education*

Medical education must be viewed as a continuum, with three constituent components: (a) basic medical education; (b) specialist training; and, (c) continuing education to maintain professional competence throughout professional life.

The medical school curriculum is the preparation for ensuing specialist training. The medical school phase results in a generic doctor, who can enter further training appropriate to a particular specialty. No doctor, in *developed countries*, relies any longer on his or her medical school education as the basis for his or her professional competence: all doctors, general practitioners as well, enter specialist training.

2.1 *Psychiatry in the Medical School Curriculum*

Psychiatry is a specialty in the field of medicine. All psychiatrists, in the first place, are medical doctors. Education in psychiatry starts, therefore, with entry to medical school, and basic medical education provides the start of the professionalization process.

In psychiatry specifically, the World Federation for Medical Education and the World Psychiatric Association in partnership have planned and instituted *The Core Curriculum in Psychiatry for Medical Students* (1999), the first ever global specification in any specialty, of the knowledge, skills, and attitudes which every doctor worldwide needs to learn about that specialty while at medical school. The *Core Curriculum in Psychiatry* specifies the minimum psychiatric instruction required by all medical students before proceeding to further training as specialists or as general practitioners.

Teachers of Psychiatry must concern themselves with the total curriculum of the medical school, and be involved with the governance of the medical school, serving on the educational committees that plan and implement the overall curriculum. Psychiatrists in medical schools must concern themselves with, and contribute to, all the canonical components of medical education in general: access (application to medical school, selection procedures, dropout, and wastage); curriculum; teaching methods; educational objectives; assessment and examinations; and evaluation and accreditation. Moreover, the departmental teaching of psychiatry must accord with the general institutional objectives of the medical school.

All branches of medicine, psychiatry of course also, are urgently confronted by the necessity to reorient the curriculum, to ensure not only curative medicine is taught, but also prevention of illness and promotion of health. Because of the great variation existing among countries and different cultures, any description of the program for psychiatry is of necessity schematic, but general principles can be set out which will apply in any setting.

Psychiatry must have a major place in the basic medical curriculum. (a) Understanding body–mind interaction, the unity of the person, is pertinent to the whole of medical practice, constantly advanced by research in the neurosciences. (b) Skills that are learned in psychiatry (e.g., the ability to form a good relationship with a patient, to assess the mental state, and to impart distressing information) are important for all doctors. (c) Psychiatric illness is common among patients seen by specialists in the other branches of medicine: among general medical outpatients, for example, about 15 percent given a medical diagnosis also have an associated psychiatric disorder; up to 40 percent of those given no specific diagnosis in fact have a psychiatric disorder. Psychiatric disorders are even more frequent among patients seen in general practice (Clare and Lader 1982). (d) Psychiatrists treat only a small portion of any nation's psychiatric morbidity; a great bulk of psychiatric disorder is managed, adequately or otherwise, by doctors whose psychiatric training often is confined to their exposure to psychiatry as medical students. All medical students require the psychiatry program, regardless of the branch of medicine they are destined to practice in the future.

2.2 *Broad Goals in Teaching Psychiatry*

When surveyed, teachers of psychiatry make clear the educational outcomes they consider the psychiatric program should achieve (Walton 1986).

(a) *Psychological-mindedness*. Communication skills and empathy with patients' emotional responses, and capacity for the interpersonal relationship between clinician and patient—the basis of the medical interview.

(b) *An objective approach to behavior*. Behavioral science teaching (notably psychology and sociology) provides students with scientific knowledge about behavior, augmenting their common-sense understanding with empirical knowledge.

(c) *Contact with psychiatric patients*. Relating clinically to patients, with a wide range of psychiatric illness and disorder, permits students to dispel their own anxieties about mental patients, and to overcome the prejudices and preconceptions customary in their society; they need therefore to develop skills in history-

taking, in examining the mental state, and in psychiatric interviewing of a patient repeatedly over a period of time.

(d) *Descriptive psychiatry*. Theoretical instruction about symptoms, signs, and syndromes, dealt with in the psychiatric textbooks, augments the clinical training when medical students are given proper access to patients.

(e) *Treatment methods*. A great proportion of all ill health being psychiatric, general practitioners and specialists in branches of medicine other than psychiatry inevitably deal with much the greater part of the mental health needs of the community. Knowledge about the main psychiatric treatment methods and skills is therefore essential for all future doctors.

2.3 Psychiatrists as Medical School Teachers

Psychiatrists have to overcome traditional disadvantages. Medical school teachers in general tend to view psychiatry as a backward branch of medicine. These nonpsychiatrist colleagues are further dismayed and confused by the doctrinal differences among psychiatrists (one a biologically oriented doctor favoring the medical model, the other a humanist concerned with psychodynamic processes, and practicing some or other form of psychotherapy). Medical school teachers in general are disconcerted when these contrasting ideologies come to be applied to the patients they refer to psychiatrists. It can sometimes seem relatively arbitrary how any patient will be diagnosed and treated—too much appears to depend on the outlook of the individual psychiatrist. Coterie positions are struck (see Coyle and Mollica of Harvard in *Current Opinion in Psychiatry* 1998, who have iconoclastically advocated that psychotherapy skills be 'deferred' to psychologists, social workers, counselors and clergy, whom they view as better suited for attending to the 'local taxonomy of human suffering').

Medical students acquire their professional values and outlook, in large measure, by modeling themselves on leading and respected teachers in the medical school. Investigations by sociologists, classically those of Merton(1957) at Cornell medical school in New York and Chuval (1980) in Israel have demonstrated that the status of psychiatry, of teachers of psychiatry, and of psychiatric patients in medical schools is not at the level of, say, medicine or surgery, with a seriously adverse effect on the teaching of psychiatry.

2.4 Medical Student Attitudes to Psychiatry

'Psychologically minded' medical students showed a greater interest in and a more positive attitude towards psychiatry than students scoring on tests as 'organically oriented.' Medical students' preference for psychiatry as a future career is stable, the same before as after psychiatric instruction. Certain learning experiences, particularly the clinical clerkship in psychiatry when in particular types of unit, can change students' attitudes towards psychiatry in a positive way. Only a very few percent of medical students choose psychiatry as a future career (see *Mentally Ill: Public Attitudes*).

2.5 The Science Base of Psychiatry

The sciences in the medical curriculum related to psychiatry include the biological sciences and the behavioral sciences. Science teaching in medicine has long been censured: a customary practice worldwide, instead of extending the 'basic' sciences throughout the curriculum, has been to compartmentalize the sciences in a so-called preclinical phase. Medical schools have imperatively to examine their entire educational program, in order to integrate science teaching properly with clinical instruction.

2.6 The Biological Sciences

Competing departmental interests fragment the sciences relevant to medicine into separate subjects (e.g., anatomy, physiology, pharmacology, etc.) at a time when integration of knowledge is the hallmark of advance (Eisenberg 1988). The *new biology* calls for an integrated view of whole structures, whether at molecular or macroscopic level. Medical schools have not responded sufficiently to the paradigm shift in science. Traditional basic science teaching in the medical schools stands charged with being outmoded, simplistic, and responsible for much of the information overload that mars medical curricula. Reform is a matter for the medical school as a whole, and is not to be left to the science disciplines in isolation from the clinical subjects.

2.7 Behavioral Science Teaching

Psychology and sociology in relation to medicine are viewed by departments of psychiatry as essential subjects in the curriculum. Despite this endorsement, behavioral sciences teaching remains problematic. Medical teachers in general are skeptical about the 'human sciences' and not sympathetic to such issues as Kant's fourth question: *Was ist der Mensch?*

Medical students and medical teachers alike are critical about the widespread failure in behavioral science teaching, given in the first or second year. Too

often a condensed general psychology and sociology course is provided, which inevitably is perceived as lacking medical relevance. Equally misguided is to include behavioral science as an optional subject in an otherwise compulsory curriculum. Experience has shown that both clinically experienced medical (psychiatrists) and nonmedical (psychologists, sociologists) teachers must be actively involved in the planning and delivery of behavioral science courses, clinical cases included throughout, to avoid implications of irrelevance or discordance with students' concerns as future doctors (see *Behavior Therapy: Psychiatric Aspects*).

2.7.1 Psychology. Students should acquire a basic overall knowledge of the findings, methodologies, and theories of psychology which are relevant to the practice of medicine; be aware of how patients' emotions, attitudes, values and experiences influence their response to illness and to its treatment; have a knowledge of learning processes and their relevance to medicine; possess knowledge about the main aspects of psychological development of humans from birth to old age; possess skills relevant to effective doctor–patient communication and particularly to interviewing; have knowledge of techniques of assessment used to test the reliability and validity of investigation procedures and therapeutic trials; and attain attitudes to development that will enable them as doctors to comprehend each patient as a complete person living in his or her own social environment.

2.7.2 Sociology. By the end of the course each student should understand: the various methods for defining and measuring health; know the importance for health and medical practice of social institutions such as the family, the community, the economy, and the law; understand the problems of equity and inequality in the provision and utilization of health services, related to age, gender, social class, and religion; grasp the changes in society and in the practice of medicine which have affected health and disease and the development of social policy; comprehend welfare provisions, the advantages and disadvantages of a country's health service, and problems of planning for change within a health service; know about the social (and sociological) factors that influence the process of becoming ill and the doctor–patient relationship, and the effect that ill health and hospitalization have on the lives of patients and their families; be able to discuss critically the role of preventive medicine and health education, and the role of self-help groups; the process of medical professionalization; and describe some of the research methods used to evaluate health and medical practice (see *Psychiatry and Sociology*).

2.7.3 Biostatistical sciences. The aims of courses in the probabilistic and information sciences are: to explore ways of making valid deductions from medical data; familiarizing students with the basic statistical terminology as found in the medical literature; to introduce the concepts necessary for designing as well as analyzing comparison and experimental studies in medicine; and methods of accessing and organizing medical knowledge; and computer literacy.

2.8 Clinical Psychiatry in the Curriculum

Instruction in clinical psychiatry must be integrated with the other main clinical subjects of the medical curriculum. The main teaching and learning methods in use are threefold: (a) lectures; (b) small group teaching; and (c) the clinical clerkship, a full-time attachment to a unit providing psychiatric care, in a general hospital psychiatric unit or in a psychiatric hospital, for six to eight weeks.

2.8.1 Objectives for clinical psychiatry in medical school. The knowledge, skills, and attitudes to be achieved by the end of study (educational outcomes) are general, deal with organic and with functional syndromes, and with treatment and management:

(a) *General*. The student should be able to: conduct a diagnostic interview including a mental state examination; relate patients' symptoms to their past experiences, personality, and social circumstances; give an account of his or her emotional responses to patients of different kinds and the way in which these reactions can influence judgment and hence the patient's management; understand patients' emotional responses to doctors and the way these can influence the presentation of illness; grasp the main principles of, and indications for, counseling and psychotherapeutic intervention.

(b) *Organic psychiatry*. The student should be able to: distinguish between organic and nonorganic psychiatric disturbance; recognize common organic mental disorders, including confusional states and dementia in different age groups; and management of organic syndromes in and out of hospital.

(c) *Functional syndromes*. The student should, for example, be able to describe: the management of depressive psychoses; schizophrenia and related psychoses; recognize common symptoms of neuroses; define and recognize the signs and symptoms of psychiatric illness; and the social services available for their management.

(d) *Treatment methods and agencies*. The student should be able to outline: the drugs commonly used; the principles of behavior modifications and their main clinical applications must be known; also important concepts in psychodynamic approaches, for example unconscious conflict and defense mechan-

isms; and the main agencies for the care and rehabilitation of the psychiatrically ill and those with learning disabilities in the community.

(e) *Examinations*. It is necessary to stress the importance of examinations, because examinations drive the curriculum. Students learn to pass examinations, whatever more high-minded educational emphasis teachers may seek to convey when specifying learning objectives. The assessment methods used should always be congruent with the objectives of the course. Only too often the methods of assessment are discordant. The professional examination needs to be composite: e.g., essay questions; short answer questions; an oral examination; continuous assessment during the clinical attachment, etc.

2.9 The Generic Medical Graduate

The product of basic medical education is a young doctor not yet equipped for independent practice on graduation. The goal of medical school is to produce a pluripotential doctor, who must then have postgraduate training to acquire the abilities needed for specialist (or generalist) practice. Prior to beginning such postgraduate training, the graduate is in the phase of transition from medical student to doctor.

The *junior hospital doctor* becomes a member of a consultant team in hospital. The notorious defects have been insufficient clinical supervision, insufficient feedback from consultants, excessive hours, and poor training standards with inadequate educational provision. Unless the general hospitals in which they work have psychiatry liaison services or psychiatric units, junior doctors receive no psychiatric instruction, and little incentive to consider training as a psychiatrist.

3. Specialist Training in Psychiatry

Training to become a specialist is the sector of medical education of particular significance to the specialty organization in each country (e.g., the American Medical Association, the American Psychiatric Association, etc.). General practice (family medicine), requires training that has a substantial psychiatry component, because of the epidemiological fact that the majority of psychiatric patients are seen not by specialist psychiatrists but by generalists.

Postgraduate training is in transition at present, not only nationally, but also worldwide. In part, this process of re-evaluation and reform results from the altering roles, responsibilities, and relationships of doctors, an upheaval so extensive and multifarious that the *World Summit on Medical Education* was designated: 'The Changing Medical Profession' (World Federation for Medical Education 1994).

Local context is crucial. Culture and history of each country are critically important in determining patterns of health care. Appropriately therefore each country continually reviews and reorients postgraduate medical education. Moreover, there are major differences between countries in their organizational frameworks for funding, delivering, and assessing postgraduate medical education. The professional nomenclature of psychiatric diseases and disorders varies: e.g., in North America the *DSM-IV* is used (American Psychiatric Association 2000) while in countries of Europe *ICD-10* is the diagnostic system (World Health Organization 1992).

3.1 Structure and Implementation of Specialist Training

Postgraduate training in almost all countries is the concern of the *Ministry of Health*, and takes place mainly in the hospitals of the country's health service. Universities are associated with health services in the postgraduate training programs, e.g., twenty psychiatric hospitals in The Netherlands have psychiatric training approval, eight of which are university hospitals. Psychiatry is one of the medical specialties now undergoing profound change (Caldicott 1996).

Specialty organizations customarily set standards and hold examinations in postgraduate education. Their particular role differs considerably among countries. In the United Kingdom, taken as an example, the Royal Colleges and their Higher Training Committees have major responsibilities. Each College (including the Royal College of Psychiatrists) through its Higher Training Committee produces its own assessment standards. The Colleges set formal competitive examinations to regulate entry into higher specialist training. During specialist training in the UK the Higher Training Committee vets the quality of the training posts and the progress of each trainee specialist, through inspection visits and regular formative assessment of the trainee.

3.2 Clinical Training

Postgraduate training in psychiatry involves a *sequence of posts*, providing a wide range of clinical experience and making up a rotational training scheme, usually in one geographical area and often in one approved hospital complex. The *rotational training scheme* includes a university department or hospital, and allows experience in the various *subspecialties* of psychiatry, including in addition to adult general psychiatry: child and adolescent psychiatry, mental retardation, geriatric psychiatry, psychotherapy, general hospital (liaison) psychiatry, and other special fields.

3.3 Theoretical

The *academic content* of postgraduate training in psychiatry should include an introduction to general

psychopathology, pharmacology, genetics, interviewing, diagnosis and classification, organic disorders, drug and alcohol dependence, schizophrenia, mood disorders, neurotic disorders, personality disorders, learning disability, psychiatric disorders of childhood and adolescence, old age psychiatry, forensic psychiatry, drug treatments, counseling and interpretative psychotherapies, behavioral and cognitive therapies, rehabilitation and community care.

3.4 Assessment Procedures

The *specialist examination* is taken at differing stages of training in different countries, and is also phased differently in the various specialties in each country. In order to sit the UK Royal College of Psychiatrists examination, trainees must have completed three years of psychiatric training and must be in a training post. After such general training, they may choose either to become general psychiatrists, or to enter additional training for a psychiatric subspecialty: e.g., in The Netherlands an additional two years in child psychiatry is required to obtain the Certificate of Completion of Specialist Training in Child Psychiatry.

In addition to the specialist examination, which as indicated above may be during the course of training, a number of countries have also introduced formal *summative* assessment at the end of specialist training. Equally important to trainees is feedback about performance, specific help about any weakness, and planning of future training.

3.5 Psychiatry Among the Medical Specialties

Countries differ in the separate specialties they identify. Psychiatry is not uniformly recognized, not even in all the countries in the European Union, regulated by law: the European Specialist Medical Qualifications Order, 1995. When psychiatry is one of the recognized specialties in a country, there are differences in training requirements including duration, e.g., it has been said that psychiatric training in the UK takes six to seven years, but four years in Portugal, Greece, or Belgium (with no formal examination in Belgium).

3.6 Subspecialties in Psychiatry

After general specialist training, young psychiatrists elect to undergo further training in subspecialties, and in professional life are designated according to the branch of the specialty in which they work. Adult psychiatry, child and adolescent psychiatry, geriatric psychiatry, forensic psychiatry, addiction psychiatry, etc. are differentiated, the range of categories varying to some extent in different countries.

3.7 National Differences in Standard Setting

In broad terms, five different patterns of responsibility for standard setting can be identified worldwide.

(a) National medical associations, as in Norway and Portugal, take the leading responsibility. In other words, the medical profession itself provides and dispenses the postgraduate training in the country. This is also the case in Germany, and in certain Central European countries.

(b) Designated professional bodies, distinct from the national medical association, may have major responsibilities, as in the United States where the specialty Board of Psychiatry, and in the United Kingdom where the Royal College of Psychiatry approve training posts as well as assess the competence of doctors in training.

(c) The government (Ministry of Health) together with the regional medical administration may take the main role. Sweden is an example of this arrangement.

(d) Universities may be the chief responsible bodies, as in Finland and southern Europe.

(e) Postgraduate Institutes in some countries are, or were, the chief postgraduate authority, as in Central and Eastern Europe.

3.8 New Developments

In Europe, within each country, there has been steady evolution of postgraduate training over the past 20 years, with formal definition, recognition, and support given to the educational programs. *International legislation* controls important aspects of medical education in the European Union, in the interests of free intercountry movement of doctors among the member states.

So that women, in particular, are not penalized, training arrangements for *doctors with domestic or other special responsibilities* have to be specified, or parents who have interrupted their training for family reasons are precluded from re-entering medical practice. European Union directives on postgraduate medical education now also specify that the minimum number of hours of work in part-time training (for general practice, at least 60 percent of the hours worked by those employed full-time). Similarly, doctors training part-time in hospital specialties will be required to extend their already lengthy training on a pro-rata basis in order to qualify for specialist registration.

Postgraduate medical training has secured the *confidence of the public* in the medical care provided by accredited specialists. To cite the UK again simply as an example, the statutory basis for the medical education infrastructure has been long regarded as a major strength of the national system, the General Medical Council, the body responsible for the register of qualified medical practitioners, held in high repute. However, malpractice and legal infringements have

recently undermined the long-standing confidence in self-regulation by the medical profession.

Doctors in postgraduate training have other professional needs which training programs must cater for. They require, for example, to learn the limits of their own competence, to identify and remedy deficits in their knowledge and skills, to learn when and how to seek the assistance of others, and to develop effective working relationships with professional colleagues and others. Structured training programs are therefore obligatory, which incorporate modern educational practice (Borman and O'Grady 1997).

4. *Continuing Medical Education (Professional Development)*

As a third great sector of medical education, maintenance of competence throughout professional life is now obligatory, and accepted as such by employing authorities, doctors themselves and the general public who require quality medical care and accountability by doctors for the clinical service they provide. Continuing medical education (CME), or professional development, consists of those educational activities undertaken by practicing doctors to maintain and update their clinical competence.

Medical journals and other medical scientific publications always were and continue to be paramount in disseminating new knowledge and promoting best clinical practice (Vysohlid and Walton 1990). Special continuing education courses were first introduced, in some middle European countries, for doctors practicing in state health services in 1878. As increasing numbers of countries introduced social insurance, furthermore, continuing improvement of their employees' professional competence also came to be expected. This direct intervention and involvement of the state, to a lesser or greater extent, in promoting new knowledge and experience for doctors in the public health service is the original model of CME.

As pertains to all specialists, so also it is the personal responsibility of every psychiatrist to remain competent. This ultimate obligation of individual physicians calls for proper means to assess constantly the quality of their work and effectiveness of their professional services, such as a system of performance review. Agreed standards are necessary, and computerized medical records greatly enhance *audit*, the process of examining one's clinical performance in relation to peers.

Being performance related, CME goes beyond information and knowledge, and focuses on the carrying out of actual clinical responsibility and services. The material included in CME, over and above lectures, case presentations, conferences, scientific meetings and conferring among clinical colleagues and their hospital firms or clinical teams, must be based on understanding of how adults learn: adults are oriented to tasks they undertake themselves, and to the personal interests they pursue. CME material must be relevant to the practitioner's own daily responsibilities. Old-style courses, in which doctors were brought together to sit through series of lectures haphazardly given by outside experts, are not congruent. Mere attendance at CME courses for which credit is given cannot be assumed to improve professional competence.

There is widespread agreement that the CME system in a country must be closely integrated with the health care system, and related to the health needs of the population. Some countries are already imposing sanctions on doctors who do not engage in approved CME activities. Self-evidently, differences in the CME system between countries are to be expected and, indeed, greatly varying patterns are compounded by national realities in the health care services, the varying professional bodies with authority for medical education and training, and the forms of statutory regulation of the medical profession.

Since the Second World War, opportunity was provided for CME in practically all North American and European countries. CME is now always adapted to identified and actual needs of the health problems of the population, and to the health services, particularly in relation to primary health care and ambulatory settings. Coordination between medical school programs and subsequent specialty training programs has powerfully consolidated the provision for continuing medical education.

The provision of clinical information made available by modern information technology is the basis now needed for CME. Distance learning can combat the relative unavailability of CME for doctors working in rural areas and distant from the main teaching center.

All doctors have now to accept personal responsibility for maintaining professional competence throughout their working lives through CME. In addition, regular monitoring and revalidation are increasingly advocated. This third, and longest, phase of medical education is the most important, and both basic medical education and postgraduate training must now be planned and implemented as preparatory for lifelong CME, to ensure professional competence of all medical doctors throughout professional life. The imperative charge, on all nations, is for provision of effective medical services by sufficient, responsible, ethical doctors.

See also: Differential Diagnosis in Psychiatry; Education: Phenomena, Concepts, and Theories; Ethics and Psychiatry; Ethics for Biomedical Research Involving Humans: International Codes; Medical Profession, The; Mental Health and Normality; Nosology in Psychiatry; Professionalization/Professions in History; Psychiatry, History of

Bibliography

Abrahamson S 1996 *Essays on Medical Education*. University Press of America, Lanham, MD

American Pychiatric Association 2000 *Diagnostic and Statistical Manual of Mental Disorders*. Fourth Edition, Washington, DC

Becker H S, Geer H B, Hughes E C 1961 *Boys in White: Student Culture in Medical School*. University of Chicago Press, Chicago

Bloom S W 1988 Structure and ideology in medical education: an analysis of resistance to change. *Journal of Health and Social Behavior* **29**: 294–306

Chuval J 1980 *Entering Medicine: The Dynamics of Transition*. Pergamon Press, Oxford, UK

Clare A W, Lader M (eds.) 1982 *Psychiatry in General Practice*. Academic Press, London

Coombs R H 1998 *Surviving Medical School*. Sage, New York

Coyle J T, Mollica R 1998 Psychiatry in medical education: a comment. *Current Opinion in Psychiatry* **11**: 493–5

Eisenberg L 1988 Science in medicine: Too much or too little or too limited in scope? *The American Journal of Medicine* **84**: 485–7

Flexner A 1910 *Medical Education in the United States and Canada: a Report to the Carnegie Foundation for the Advancement of Teaching*. Updyke, Boston

Iserson K V 1997 *Get into Medical School: A Guide for the Perplexed*. Galen Press, Tucson, AZ

Merton R K, Reader G, Kendall P 1957 *The Student Physician*. Columbia University Press, New York

Miller G 1980 *Educating Medical Teachers*. Harvard University Press, Cambridge, MA

Vysohlid J, Walton H J 1990 Development of continuing medical education in Europe: A review. *Medical Education* **24**: 406–12

Walton H J (ed.) 1986 *Education and Training in Psychiatry: A Case Study in the Continuity of Medical Education*. Oxford University Press, London

Walton H J 1994 Continuing medical education in Europe: A survey. *Medical Education* **28**: 333–42

World Federation for Medical Education 1988 The Edinburgh Declaration. *Lancet* **8608**: 464

World Federation for Medical Education 1994 Proceedings of the world summit on medical education, 1993. *Medical Education* **28**(1): 1–171

World Health Organization 1992 *International Classification of Diseases*, 10th edn. World Health Organization, Geneva, Switzerland

World Psychiatric Association and World Federation for Medical Education 1999 Core curriculum in psychiatry for medical students. *Medical Education* **33**: 204–11

H. J. Walton

Education: Phenomena, Concepts, and Theories

1. Classical and Christian Traditions

The Latin verb *educare* can be translated as 'to pull up,' and it is in this sense that it is understood to mean 'to educate': a person is pulled upwards. As both a concept and a play on words, 'education' entails a difference in height, and a process by which this difference is overcome. In Western literature, the concept of the 'teacher' and 'educationalist' has been associated with asymmetry ever since Homer's *Iliad*, and the direction of this asymmetric relationship cannot be reversed at any point in the educational process (Homer 1987, Book 9, p. 430ff.). 'Teachers' are appointed to train 'pupils,' who in turn are to learn what the teacher knows and represents. Education is subject to the person of the teacher, who is equipped with limited authority, the purpose of which is to make him- or herself superfluous upon successful completion of the educational process.

With regard to the terminology, a distinction is to be made between *educatio* and *eruditio*. The first area is that of moral education—in the wider sense a training of the body and the senses—whereas the second area is concerned with scholarly education and instruction. The two areas are often understood as opposites, as was the case in ancient Sparta, and later in the theory of Rousseau. The fundamental friction between the two has remained to this day, and educational discourses typically lend weight to one side or the other. This was also the case in Graeco-Roman times. A key question was how positive education should or must be. The question was answered by means of two consequential concepts, relating to rhetoric and public teaching on the one hand, and self-education on the other. Sophist teachers, who—as portrayed in the dialogs of Plato—publicly announced truths and virtues, were juxtaposed with Socratic dialogs, in which Socrates guided his untutored interlocutors to formulate answers to the problems he posed. Finding the solution to the problem required the 'pupil' to engage in independent thought with the maieutic (or Socratic) assistance of the 'teacher'; the teacher does not present a positive truth in a didactic manner.

The fundamental question of whether virtues can be taught is posed by Socrates in Plato's dialog *Meno*. The question is not answered, however, at least not in the expected sense. The first historical negation of the 'teacher' (of virtues; Plato 1999, 96a–c) is accompanied by a rejection of the teaching of virtues in favour of 'divine providence' (Plato 1999, 100a). This links education with Plato's theory of the soul (1999). Education is not simply rhetoric because—according to Socrates—truth cannot be taught in public. Knowledge is innate in the soul, rather than learned from experience. It is the teacher's task to encourage pupils to think for themselves, thus becoming aware of these timeless ideas. In other words, pupils need to be goaded into conscious reflection.

The relationship to psychology is of fundamental importance to the theory of education and the demands made of it. Without the concept of the soul, education cannot be considered to have lasting 'inner' effects. Yet Platonic theory rejects the notion that the soul can be easily affected by education. Education is

not merely the imitation of nature. Its two main areas, morality and knowledge, require pedagogical efforts going beyond imitation. The means of education, i.e., dialog, admonition, models, or guidance towards new insights, are not given simply on the basis of nature. At the same time, pedagogical efforts are limited. Education imparts knowledge and skills, but it cannot reach all of mankind. The Socratic wariness of 'teachers' who independently and publicly pronounced truths remained intact on into Christian times.

2. *Modern Aspects of Educational Theory*

With the Renaissance it became possible to forge a conceptual link between human dignity and education, i.e., to keep education free from doctrines of humiliation. In this sense, education (*eruditio*) became humanistic, but without the canon of classical teaching being lost. However, the adoption of Platonism in the Italian Renaissance, particularly in Florence (Hankins 1991), allowed the theory of 'inner' education—related to innate ideas, rather than doctrines or text books—to be strengthened. In the seventeenth and eighteenth centuries, new Platonic theories of education determined cultural life, i.e., literature, art, music, and key aspects of philosophy. Education was the aim and the expression of lasting ideas which, in turn, were intended to be understood in a manner which was independent of certain Christian doctrines. In England, in particular, the independence of education and the consequent fight against orthodoxy in the seventeenth century amounted to a political force in which new Platonism and the natural sciences were allies.

In this respect, a distinction is to be made regarding the education of children. Since the Reformation, education has concentrated on the minds and souls of children. This was strengthened by post-medieval catechisms which stipulated the form of learning, and emphasised the necessity of education beyond the parish and the church (Bast 1997). It was envisaged that the establishment of elementary school education in the eighteenth century would supplement education at home, without too much weight being attached to knowledge during lessons. The focus on Christian morals was of fundamental importance. These were supposed to become spiritualised, i.e., more closely related to the Augustinian 'inner teacher.' Moreover, the pietistic conception of the child resulted in the pedagogical dissemination of the Christian (particularly New Testament) view of the child as a paragon. Only adults whose inner being could once more become that of a child were close to Christian mercy. With pietism, education became a missionary power with the goal of bringing forth 'new people' and subsequently a 'new world' of belief.

In contrast to esthetic education, which determined the concept of 'the educated' and made it independent of teachers, the education of children has typically been seen from the perspective of educational methods. With the introduction of printing, instructional theories on the education of children could be disseminated for the first time. Thus, expectations were established, examples of which can be found in Christian pedagogy as early as the seventeenth century (e.g., Mitternacht 1666). Education was reduced to a methodological basis. In the eighteenth century, sensualistic psychology strengthened the trend for education to be understood as a target-oriented effect. The Platonic concept of 'innate ideas' became less influential, and was replaced by theories of learning. These stated that the inner world is established by processes of learning, and that learning can be guided by education. Hence, education, and subsequently morals, would make the person what he is, and what he becomes (D'Holbach 1770/1994, p. 346ff.).

The aspect of the inner being was further developed by Rousseau (1969), who combined earlier notions about the 'nature' of children (Mercier 1961) with Buffon's theory of human development. Basically, Rousseau proposed a solution to the problem of mercy that was widely discussed in French educational circles during the seventeenth and early-eighteenth centuries. His approach was that if the Augustinian theory of initial sin can be rejected, and if children can be assumed to develop on the basis of their natural goodness, education can be understood to be a new form of mercy. It cannot be taken as given; it occurs together with the development of nature, subject to the proviso that nature is protected against abuse by society. Rousseau therefore refers to *éducation negative*, in which virtues and reason are not imparted in a positive manner. The Socratic nature of *Emile* also becomes apparent in that the teacher (governor) is an omniscient authority. In contrast to the teacher in *Meno*, however, Rousseau acts as if he were simply allowing matters to take their course. The key concept for childhood—'*laisser meurir l'enfance dans les enfants*' (Rousseau 1969, IV/p. 324)—has Augustinian features, but the 'inner master' is not intended to, and should not, be a representation of the outer one.

Rousseau also embraces a certain version of the two-world theory: in childhood, nature is this life, and society the next world. The only difference is that the roles are played in reverse order, and do not follow the same sequence as in the case of Augustine. It is envisaged that a good education will engender a good society, conditional on the ability to keep the existing society out of education. At the same time, natural education is supposed to abstain from influencing the child's future (Rousseau 1969, IV, p. 301ff.). It would be 'barbaric' if it were geared towards the citizen (*le citoyen*) rather than the man (*l'homme*). One can speak of the education of *men* if nature is intact, and one can thus develop in accordance with one's own potential. In other words, Rousseau embraces the theory of

development. Education is not only concerned with the mind or learning, but with the whole of nature and its development. This presupposes that childhood is a world of its own (Rousseau 1969, IV, p. 319). This doctrine is far-reaching, and has constantly attracted new followers who understand learning as genetic and relate it to the autonomous child 'developing' from within (Oelkers 2000).

3. Educational Phenomena and Pedagogical Theories

There is a high degree of historical continuity in educational theories. Theories of childhood are just as dependent on the history of the concept as theories of education are. This holds even if they relate to societies which are in stark contrast to that of the eighteenth century, or to new media which neither Rousseau nor Locke could have foreseen. Hence, educational theories are not randomly constructed; they refer back to historical principles which are flexible enough to be applied to various situations and environments. At least, this is the case if very general statements or theories stipulating what can and cannot be deemed a valid educational phenomenon are discussed and accepted.

General theories of education are those which do not refer to certain environments, such as Plato's Athens, to certain doctrines, such as Augustine's Christianity, or to certain practices, such as catechism lessons. It is possible to generalise education by referring to doctrines of the soul, concepts of society, nature, progress or universal morals. It is almost always the case that theories mix these elements, frequently by shifting the relations between them and utilizing their immanent dualisms. The 'mind' of a child can be seen as the result of 'natural development'; such 'natural development' can oppose 'society'; 'morals' and 'nature' can be seen as either complementary or opposing forces; education can serve to impart society's morals or to oppose them, etc.

Educational theories only become specific when they do more than simply develop dualisms. They must at least differentiate between inner and outer, between process and product, between children and adults, between intention and effect and between objective and experience. These elements must be capable of being linked together, and cannot be understood as opposing forces if a positive theory of education is to be created. Such a theory is normative, rather than merely descriptive. It is geared towards practice and aimed at bringing about improvement. This demands the definition and justification of the pedagogical intent, as well as the prediction of outcomes. The demands placed on education can only be limited if the theory can differentiate between inner and outer. Objectives do not immediately constitute effects; educational processes are controlled by means of target-oriented rhetoric and experience, without the two ever being identical.

When the above is applied to the sphere of anthropology, the following statement can be made: children are not adults. Moreover, they do not simply become adults in accordance with the intentions of their educators—they also gather their own experience, independent of educational targets. In a literal sense, children are not the 'product' of education, primarily because each educational outcome can be changed by subsequent experiences. A pedagogical intent is necessary, but it is not the only causal factor. Education is a permanent and hazardous attempt to adapt to each new child. The subjects or themes of educational theory emerge in response to educational phenomena which may be of a moral, esthetic, or pragmatic nature. This is conditional on the identification of a certain need or deficit which requires pedagogical processing.

In contrast to the classical theories, modern educational theories are unlimited. They extend beyond the interpersonal relations of 'masters' and 'pupils,' and may relate to the 'world' or 'society'; they allow for a generalisation of the child as well as for abstract objectives such as 'emancipation'; they are no longer bound by religious theories. Accordingly, educational phenomena are general. In other words, they determine both the object and the responsibilities of education. The term 'education' may relate to literature and the arts, philosophy and all areas of knowledge, as well as all kinds of beliefs, but it may also relate to morality and mores, societal development or the general betterment of mankind. To this end, there are several competing theories allowing for very little standardisation.

In view of the interrelations between the theories, however, educational theories should not simply be equated with the names of their authors. The classification of theories according to names, or schools and directions associated with these names, is the most frequent lexical form of organisation. Definitions of this kind are not only obscure, they fail to produce real classifications, because the grouping principle barely allows any exclusions to be made. Ultimately, everything that can be related to the abstract relationship between humankind and the world can be described as 'humanistic,' 'dialogical,' 'progressive' or 'critical'—or indeed the opposite.

However, one can only speak in terms of 'educational theories' if primary criteria are used, allowing for differentiation and suggesting definitions. It is not sufficient merely to confirm the doctrines of certain pedagogical authors. However, part of the reflection on education is determined by this very desire for approval with respect to historical parameters perceived as archetypes. In this sense, educational theories should—and can—educate in their own right. They focus reflection on and perception of educational phenomena to the area of validity of doctrines.

When reflecting on the topic of 'education,' however, one is forced to consider the following unavoidable problems: observation of children reveals not only a 'natural development,' but a complex world of perceptions that must constantly link the inner and outer aspects without being able to rely on stages or supports. Accordingly, the concept of 'education' cannot contain the dualism of 'development' or 'impression.' Both concepts are helpful in the politicisation of education: 'development' is understood in the sense of siding with children, whereas 'impression' is understood as a means of strengthening educational authority. However, this alone does not constitute an educational theory. A distinction must be made between educational theory and the public rhetoric about education. A theory of education cannot merely confirm general aims by reference to preconceived doctrines, nor can it rely on 'education' being seen merely as a moral suggestion.

4. *Educational Theories*

Against this backdrop of theoretical tradition, at least five types of educational theories can be distinguished. This differentiation is made on the basis of the theories' underlying approach, which may be related to theory systems or general models of reflection as widely varying as philosophy, psychology, social theory, or anthropology. They can be grouped into idealistic, pragmatic, cultural, individualistic, and political approaches to educational theory, at the very least. These are not different aspects of one approach, but different ways of reflecting on the general theme of education. The methods used vary, and include analytical, empirical and historical approaches. The main topic of 'education' is outlined in terms of a normative core which, as far as is possible, is justified by hard facts. The theory itself is a construct which in some cases tends more towards *educatio*, and in other cases leans more towards *eruditio*.

Idealistic theories of education are based on moral aims or principles, which can be justified in an ideal or 'transcendental' manner. The two most influential approaches in the history of the theory are those of Plato (1999) and Kant. In the first approach, education serves to organize the state in accordance with the philosophical ideas of goodness and justice. The second approach is characterized by a form of education that centers on personal morality in accordance with the categorical imperative of action. Individual liberty is limited by law, and action is guided by strict virtues. The idea of the general good justifies the ideal state; practical reason justifies moral action. Consequently, educational theory has its normative core in the notions of goodness and morality, both of which are to be transmitted to children and citizens.

Pragmatic theories of education emphasise the paramount importance and hazardous nature of one's own actions. Since Darwin, these have been justified by an evolutionary-theoretical approach, in the sense that education is understood to be a continuous readjustment to changing worlds of experience. Against this background, Dewey (1985) sees education as 'a process of the continuous reconstruction of experience.' There is no morality that is formulated outside or independent of experience; the general public and democracy represent the regulative bodies for moral exchange. Similarly, Mead (1970) does not see education as an isolated chain of events between the inner and outer aspects. According to Mead, the 'organized self' (1970, p. 162) learns and gathers experience within the framework of its group, i.e., subject to the proviso of continual exchange processes (1970, p. 265). Education 'is' nothing else.

Cultural theories regard education as Bildung or initiation (Peters 1965), and are justified in a 'transcendental' manner. The process of initiation opens regulated ways of learning to culture and society. Those who intend to assess education must become involved with it. The educated man/woman is a normative ideal that develops within the process of education, i.e., does not exist independent of it. This is conditional on the historical existence of culture, i.e., the major fields of knowledge and ability into which the student is initiated. The individual personality develops in the process of this initiation. According to Bruner (1996), education itself can also be perceived as a 'culture' which, by means of specific institutions, ensures that education is provided for successive generations, with constant or increasing levels of performance. Schools and universities and, in a broader sense, the educational media are institutions for the continual transfer of knowledge and learning which can also be described in terms of general systems theory (Luhmann and Schorr 1988).

Individualistic theories of education, such as that advanced by Herbart (1808), focus on morals, understood to be a characteristic or potential of the personality. 'Morals' are based on practical ideas with the intention of determining moral education. Since Herbart, it has been possible to draw up a concept of morals on the basis of esthetic judgments, i.e., as an immediate and involuntary assessment of the excellence or impropriety of an occurrence or an action. This judgement is not simply subject to the law, it occurs spontaneously. The concept of spontaneity means that individualistic theories are in stark contrast to all theories which expect education to be a fixed inner result. Since Kierkegaard (1980), there have been different attempts to lessen the friction between the ethical and the esthetic, but the conflict between duty and affection in education has never really been resolved. The conflict recurs regularly: today's constructivist educational theories stress the spontaneous creation of one's own world, while moral theories refer to the 'inner' results of education, which are not supposed to change spontaneously.

At the beginning of the twenty-first century, political theories of education are primarily geared towards democracy. The totalitarian theories of the twentieth century have largely disappeared. It is no coincidence that a new reception of the eighteenth century civic society emerged after 1989, and that Dewey's *Democracy and Education* (1916) has become the focus of international pedagogical discussion. In contrast to Dewey, new approaches to a political theory of education include expanding its scope to involve the deliberative public (Gutmann 1999), and specifying its tasks and subjects to involve civic virtues (Callan 1997). The fundamental friction has survived: on the one hand, political democracy educates by means of its problems and procedures; on the other hand, education is intended to promote democracy. Citizens must be equipped with the capacity to participate actively in public life. Yet even the best school is not an embryonic society, anticipating on a small scale what later occurs on a much larger scale. The object of the theory is determined and non-determined at one and the same time.

5. *Educational Criticism*

Educational theories are simultaneously positive and critical. They draw attention to deficits which are to be made up or avoided in the future. This calls for a critical approach to present and past education. The underlying objective is for the future to be better than the past, and this would seem to be feasible if and inasmuch as education can be improved. Criticism of education *per se* is different. This sort of criticism rejects the claims that a theory of education is able to group, describe, and control an object or a range of subjects pertaining to 'education.' The normative alignment of most of the educational theories is seen as an illusion, their expectations of effectiveness are related to technological deficits in educational practice, and the identification of the 'child' as the object of education is subject to criticism on both the moral and psychological levels.

Education as an illusion was the subject of the psychoanalytical criticism of the 1920s in particular (e.g., Bernfeld 1925). According to this critical approach, the theory of education reflects ideals and thus expectations which can be understood as reflections of the subconscious, i.e., which refer back to the authors of the theory. Their illusory nature of this discourse means that it does not need to take note of educational realities. The critique of education advanced within the framework of general systems theory refers to the fact that the education system does not have access to any technology which might guarantee its objectives, it is referred to as a fragile form of communication which needs to be self-directing. Similarly, constructivist approaches to educational criticism underline the autonomy and self-organization of learning which cannot be incorporated by educational theories in which adults decide and act vicariously for their children. Furthermore, these kinds of theories tend to regard children, or learners in general, as the objects of education, i.e., to impose a certain course of action on them irrespective of their approval.

This was the very problem with Kant's ethics which, for good reason, stipulated that persons should not be treated as objects. However, Kant could not eliminate compulsion from education. The problem is insoluble as long as the school of thought is determined by dualisms such as compulsion or liberty. The general criticism of education uses the dualisms rather than avoiding them. For example, 'inner' is played off against 'outer,' 'self-organization' is confronted with 'outside determination,' and the discrepancies between the aims and the realities of education are exposed. Educational theory, in contrast, requires relations which allow for graduation, and which can be related to ongoing experiences in changing situations. The objects of the theory are those who act, i.e., those who can grant more or less liberty, use more or less compulsion, realize better or worse objectives, adapt precisely or less precisely to realities, relate to a child's personality in a more or less appropriate manner, etc.

It is necessary for a happy medium to be found in each case. Educational theory must be adjusted to a temporal and social succession of situations which are in constant flux. In this respect, educational theory represents a continual learning experience which needs to be able to correct its aims and intentions. Most of the educational theories mentioned above are not geared towards self-correction, primarily because their basic premises are not supposed—or allowed—to be modified. Their expectations are geared towards public rhetoric; education is supposed to achieve general targets, and is not understood as an experiment that may succeed or fail, but can only succeed if its approach can be corrected in response to experience. This experimental understanding of education has been embraced in the theory of Dewey (1985), in particular. However, it had only a very inconsistent effect on his own efforts to formulate an educational theory.

This is no coincidence. On the one hand, 'education' is a normative expectation which precludes modifications; on the other hand, it is a fragile practice which adjusts expectations, bringing them into line with possibilities. Educational theories should take account of this differentiation and process both levels. Expectations regarding education can be understood and rationalized in a manner which distances itself from public rhetoric (Siegel 1988); the practice of education can be formulated in terms of contractual theory (Gauthier 1990) or action theory; reflection on pedagogical matters can respond to uncertainties about the paradoxes of the educational field and pick out unavoidable dilemmas as its central theme. In this way, 'education' will no longer be an object, but a series of subjects and themes geared towards public

understanding, with no more than disciplined reflection at their disposal. In contrast to the popular reform pedagogy, educational theories are no longer perceived as guarantees for success. This also applies to the attempts to infer the nature of 'education' from general models of philosophy, psychology or other disciplines. In view of the high level of competitiveness between the general theories, their mere existence is no guarantee for their application in practice.

The normative type of pedagogical tradition still determines the field of reflection to a considerable extent. It suggests generalizations and tends to preclude practical or empirical differences. The concept of 'education' survives in public rhetoric, as do its suggestive powers and standards of argumentation. This explains why it can be so easily called into question and is constantly coming under fire. In this respect, education is nothing more than the belief in it. In fact, educational expectations are renewed irrespective of periodic or even chronic crises, meaning that better theories do not simply replace poorer ones. Progress does not occur in the form of a general increase in knowledge; the theories exist in a complex field of public expectations, tough competition, unrelenting demand, and high moral conformity.

The theoretical dynamics created between science, philosophy, and practice cannot be understood without this field of expectations. On the one hand, they are subject to very long-term historical assumptions; on the other hand, responses are expected on a short-term basis, because 'education' is a priority in the public perception, calling for immediate explanation and action. Because societal expectations are so high, educational theory has neither the time nor the distance to subject itself to self-assessment. It does not simply react to a scientific problem, it reacts to public demand. As such, it is and remains attractive, but at the same time creates its own particular difficulties.

See also: Cross-cultural Study of Education; Education: Anthropological Aspects; Educational Institutions and Society; Educational Philosophy: Contemporary Views; Educational Philosophy: Historical Perspectives

Bibliography

Augustine A 1993 *De Magistro* (*Le Maître*). Jolbert B [trans. and ed.]. Editions Klincksieck, Paris

Bast R J 1997 *Honor Your Fathers. Catechisms and the Emergence of a Patriarchal Ideology in Germany, 1400–1600*. Brill,

Bernfeld S 1925 *Sisyphos oder die Grenzen der Erziehung*. Internationaler Psychoanalytischer Verlag, Leipzig/Wien/Zürich

Bruner J 1996 *The Culture of Education*. Harvard University Press, Cambridge, MA

Callan E 1997 *Creating Citizens. Political Education and Liberal Democracy*. Clarendon Press, Oxford, UK

Clément d'Alexandrie 1960–1970 *Le pédagogue*. Livre I-III. Marrou H-I (ed.) Cerf, Paris

Dewey J 1911/1985 Contributions to *A Cyclopedia of Education* Volumes 1 and 2. *The Middle Works* 1899–1924, Vol. 6: *How We Think* and *Selected Essays 1910–1911*. In: Boydston J A (ed.) Southern Illinois University Press, Carbondale/-Edwardsville, IL, pp. 357–467

D'Holbach P-H Th. 1770/1994 *Système de la Nature ou des lois du monde physique et du monde moral*. 1994 edn. Diderot. T. I. (ed.). Georg Olms, Hildesheim/Zürich/New York

Gauthier D 1990 *Moral Dealing: Contract, Ethics, and Reason*. Cornell University Press, Ithaca, NY/London

Gutmann A 1987/1999 *Democratic Education*. Princeton University Press, Princeton, NJ

Hankins J 1991 *Plato in the Italian Renaissance*. 3rd edn. Leiden/New York/Köln (Columbia Studies in the Classical Tradition, ed. Harris W V, Vol. XVII)

Herbart J F 1808 *Allgemeine Pädagogik aus dem Zweck der Erziehung abgeleitet*. Johann Friedrich Röwer, Göttingen, Germany

Homer 1987 *Iliad*. (trans. Lattimore R). Chicago University Press, Chicago

Kierkegaard S 1843/1980 *Das Gleichgewicht zwischen dem Aesthetischen und dem Ethischen in der Herausarbeitung der Persönlichkeit. Entweder/Oder*. Zweiter Teil Band 2. (trans. Hirsch E). Gütersloher Verlaghaus Gerd Mohn, Gütersloh, Germany, pp. 165–356

Luhmann N, Schorr K E 1988 *Reflexionsprobleme im Erziehungssystem*. Suhrkamp, Frankfurt am Main, Germany

Marrou H-I 1982 *Histoire de l'éducation dans l'antiquité*. Tome I: Le monde grec. Tome II: Le monde romain. 7e éd. Editions du Seuil, Paris (English translation 1956)

Mead G H 1970 *Mind, Self and Society from the Standpoint of a Social Behaviorist*. Chicago University Press, Chicago

Mercier R 1961 *L'enfant dans la société du XVIIIe siècle (Avant l'Emile)*. Unpublished doctoral thesis, Faculté des Lettres, Université de Paris, Paris

Mitternacht S 1666 *Paedia, Das ist: Unvorgreiffliches und wolgemeintes Bedencken/von der Erziehung und Unterweisung der Kinder*… Johann Scheiben, Leipzig, Germany

Oelkers J 2000 Rousseau and the *Image* of 'Modern Education.' Unpublished manuscript, Pādagogisches Institut (Fachbereich Allgemeine Pädogogik), Universität Zürich, Zürich

Peters R S 1965 Education as initiation. *Philosophical Analysis and Education*. Routledge and Kegan Paul Humanities Press, London, pp. 87–111

Plato 1999 Gorgias-Ménon. *Oeuvres complètes*. Vol. III/2nd Part. Croiset A, Bodin L (eds.) 17th edn. Les Belles Lettres, Paris

Riché P 1995 *Education et culture dans l'Occident barbare VIe-VIIIe siècle*. 4th edn. Editions du Seuil, Paris

Rousseau J J 1969 *Oeuvres complètes*. Gagnebin B and Raymond M (eds.) T. IV: Emile. Education, Morale, Botanique. Editions Gallimard, Paris

Siegel H 1988 *Educating Reason. Rationality, Critical Thinking, and Education*. Routledge, New York

J. Oelkers

Education (Primary and Secondary Schools) and Gender

Since the 1970s, the number of empirical studies on differences in the academic performance and social

development of boys and girls in elementary and secondary school has escalated. In 1992, the American Association of University Women Educational Foundation (AAUW) estimated that over 1,000 articles and studies on girls and their education in kindergarten through 12th grade had been reviewed for their comprehensive report, *How Schools Shortchange Girls*. Gender studies of this type typically show how societal structures, primarily schools, distribute resources inequitably among boys and girls and what effect these distributions have on various outcomes such as achievement, self-esteem, educational attainment, and occupational choice. It now appears that the distribution of resources is changing, and assumptions about the treatment and consequences of certain actions are being challenged as the gender gap between boys and girls closes along several key dimensions.

Recent US data indicate that the achievement test score gap between girls and boys has been steadily declining during the 1980s and 1990s, and girls now perform as well as boys on most national tests.

In 1996, girls outperformed boys on tests in subjects such as reading and writing, and performed nearly equal to boys in mathematics (Campbell et al. 1998). In light of these findings, the question has been raised if schools are still 'shortchanging girls.' Kleinfeld (2000) argues that the idea that girls are victimized at school is wrong; girls receive higher grades in school, perform better than boys on standardized tests in many subjects, and enter and complete college more frequently than young men. Males are more likely to be placed in special classes for learning disabilities, are more difficult for teachers to handle, and are given less encouragement by their teachers to succeed.

As differences in the academic performance and educational attainment of boys and girls continue to decrease in magnitude, increasingly researchers are suggesting that girls and boys are a great deal more similar than they are different. This is not to say that there are no gender differences, but they appear when examining specific knowledge and skills in particular samples of females and males. Gender research has moved from assumptions of homogeneity within the sexes to a more detailed examination of intra-gender differences among girls and boys. Focusing on race, ethnicity, and social class has made recent research on gender more complex, and allowed for more fine grained distinctions of the academic and social challenges facing boys and girls as they progress through school.

1. *The Elementary School Years*

When they enter kindergarten, most young children have clear ideas about gender and strive for conformity with gender-stereotyped roles. Boys are committed more rigidly to sex roles than girls, consistently choosing and preferring sex-typed toys and activities. By age 10, there appears to be more flexibility in sex-segregated behaviors, although most children continue to prefer to spend time with same-sex peers (Thorne 1993).

Large-scale studies show that females are more likely to start their formal schooling at an earlier age than males and are less likely to fall behind academically and repeat a grade. Girls typically enter kindergarten with education-related skills such as small muscle development and language competence, abilities that are highly valued by their teachers. Mastery of these skills often eases the transition for girls from preschool into kindergarten. Boys, on the other hand, tend not to be as proficient in these capacities when they enter school and receive more attention than girls from their teachers, who give them specific and helpful feedback and challenge them with more complex and abstract questions. Teacher–student interaction studies in elementary school reveal that teachers tend to view boys' academic problems as motivational, but when girls exhibit similar problems they are seen as lacking ability (Sadker and Sadker 1994).

2. *Middle School*

The differential treatment by teachers of girls and boys continues through middle school where, it has been argued, girls' sense of competence is undermined and they become silent in the classroom (Brown and Gilligan 1992). Findings by Eccles et al. (1998) indicate that gender differences are more likely to be observed where gender stereotypes are endorsed by girls and boys. Girls who believe the stereotype that boys are more competent in a subject like math will be more likely to lack confidence and to avoid challenges in that subject.

Beginning in early adolescence, females consistently report greater dissatisfaction with their appearance and bodies than males. This phenomenon appears to have intensified over the last several decades. Perceptions of one's attractiveness are correlated with self-esteem for both males and females. Some claim, however, that females are more susceptible to the influences of the media, their teachers, parents, and peers, and have substantially lower self-esteem than males (Orenstein 1994). A more cautious view of these findings and other studies are discussed in two major meta-analyses conducted by Kling et al. (1999) and Feingold (1994), which show small but statistically significant gender differences in self-esteem favoring males for both cross-sectional and longitudinal studies.

In middle school, girls continue to demonstrate an academic advantage over boys, outperforming them with respect to grades in academic and nonacademic

subjects. However, there appear to be no significant differences in test performance between boys and girls (Willingham and Cole 1997). Females at this age are more likely than males to report talking to their parents about selecting courses, school activities, and their classes. Boys continue to skip school and get into trouble more often than girls. At home, however, there appear to be few differences in how parents monitor behavior. Males and females are likely to report similar levels of parental control regarding television viewing, reviewing homework, and limiting the time they can go out and be with friends (US Department of Education 1995).

3. Secondary School

Despite growing support among the general public and supplementary federal funding for the development of nonsexist curricular materials and programs in the 1970s, gender equity issues continued to plague American secondary schools well into the early 1980s. Although in high school girls consistently received higher grades than boys, boys tended to perform better than girls on standardized tests. In trying to discover what contributed to these results, researchers examined several different factors, including educational expectations, course-taking behaviors, gender bias of standardized tests, peer group interactions, and extracurricular participation. Results of their efforts encouraged the development of a number of different programs and activities to increase teacher and parent awareness of the problems facing girls in school (AAUW 1995). In the 1990s, it appears that these initiatives have been effective, and gender differences are slight or nonexistent in some domains.

4. Educational Expectations and Occupational Aspirations

Educational expectations and occupational aspirations, which have been shown to be highly sex-typed, have tended to be predictive of adult occupational attainment (Marini and Brinton 1984). Today, however, most adolescents expect to attend college after high school graduation, and these expectations do not vary by gender (Schneider and Stevenson 1999). Comparing large-scale longitudinal data from 1972 to 1992, Green et al. (1995a) show that in 1972 males were more likely than females to expect to complete college. However, by 1992, females and males were equally likely to expect to complete their education with at least a bachelor's degree. In 1992, for the first time, females were more likely than males to expect to earn a graduate degree. These results do not hold across all groups: Hispanic females are the least educationally ambitious and are least likely to expect to complete college or earn a graduate degree.

Males and females both aspire to professional occupations that require higher education degrees. However, gender differences continue to persist among certain occupations. Females are more likely than males to aspire to careers in education, health, psychology, and cosmetology, whereas males are more likely than females to aspire to careers in engineering, auto mechanics, and electronics. No gender differences are apparent for career aspirations in computer programming or business.

5. Course-taking Patterns

The courses students choose to take in high school have a significant impact on their lives after high school graduation. Students who take advanced courses in science and mathematics are more likely to attend selective colleges than those who do not. In the 1980s, girls lagged behind boys in the number of mathematics and science courses they enrolled in. In the 1990s there has been a change, with girls now taking nearly as many mathematics and science courses as boys (Wirt et al. 1999). However, a gap still exists with respect to the types of courses girls and boys take in these subjects. Females are less likely than males to take a fourth year of mathematics or science, which includes calculus and physics. Even females who earned high grades in previous mathematics and science classes often elect not to take these courses. Boys are more likely than girls to take physics as well as biology and chemistry. Girls outnumber boys in humanities and social studies; they are also more likely than boys to be enrolled in basic data entry and word processing courses.

Advanced placement (AP) and honors courses are geared to help students prepare for advanced placement tests for college entrance, where a scoring within a specified range allows a student to earn college credit for a specific subject. Consistent with their general course-taking patterns, females are more likely than males to take honors courses in foreign languages, English, and social studies; take about the same number of honors classes as boys in science, and slightly fewer in mathematics (Dwyer and Johnson 1997). In the 1990s, the number of girls taking AP exams increased more rapidly than the number of boys. By 1996, the number of males and females taking AP exams in the natural sciences and mathematics also increased, with the number of females increasing slightly more than males. Although more females were taking AP exams in the 1990s, the relative performance of girls and boys in most subjects was very similar (Willingham and Cole 1997).

6. Other Tests of Academic Performance

The small sex differences observed on the AP exams can also be found on other standardized tests. Recent results from the National Assessment of Educational Progress (NAEP), a nationally representative exam of student knowledge in selected subjects, show that the highest standardized test scores in reading and writing across all sampled grade levels were earned by girls. Boys earned higher scores on average in history, geography, mathematics, and science (Campbell et al. 1998). Males' test scores are about 10 percent more variable than females' which means that men are likely to outnumber women at the top and bottom of the scale, even though there are no sex differences in the mean scores (Hedges and Friedman 1993). Therefore, if the cutoff for admission to a highly selective program is scoring in the top 5 percent, fewer than half as many women as men may be admitted.

7. Experiences in School

Gender differences also characterize many aspects of adolescents' peer relationships. Adolescent girls' peer networks are often larger than those of adolescent boys, include more opposite-sex friends and fewer close friends. Nevertheless, girls' friendships are characterized by greater intimacy than are those of boys (Connolly and Konarski 1994). Females and their friends are more likely than males and their friends to have positive values and attitudes regarding their schooling experiences. While 43 percent of females say their friends think studying is very important, only 29 percent of males believe studying is very important to their close friends. Nearly 56 percent of females compared to 41 percent of males believe that their friends value receiving good grades (Green et al. 1995b).

8. Extracurricular Participation

There has been an increasing awareness that much of gender socialization occurs outside of the classroom in the extracurricular and informal activities of the school (Eder and Parker 1987). Data from the National Education Longitudinal Study of 1988–94, a nationally representative longitudinal sample of over 20,000 adolescents, show differences in gender participation. Girls were more likely to participate in artistic, religious, social, and service activities. White females were more likely to participate in most extracurricular activities than Hispanic, African–American, or Asian–American females. Male high school graduates were more likely than females to have taken at least one year of physical education, and outnumbered girls in team sports. Girls outnumbered boys in performing arts, school government, and literary activities, but boys were more likely than girls to occupy positions of leadership in some of these activities. As for activities outside of school, girls were more likely to take classes and be involved in activities with their parents than were boys.

9. Single-sex Schools

One of the responses to the negative implications of gender bias in schools has been a movement toward creating more single-sex schools (US Department of Education 1993). Research on the advantages of single-sex schools is mixed, particularly when taking into account selection bias issues. However, as more parents and educators become concerned with issues of harassment, differences in self-esteem, and the academic performance of females, the number of female single-sex schools may increase, even though there is not overwhelming evidence that girls fare better academically in these environments.

10. Leaving School before High School Graduation

Although dropout rates have been declining steadily, there are some important distinctions with respect to gender. Dropout rates are especially high among Hispanic girls. In 1995, 30 percent of Hispanic females aged 16–24 had dropped out of school and not yet passed a high school equivalency test. In contrast, dropout rates for white and African–American females declined, and those for white, African–American, and Hispanic males remained stable (US Department of Education 1995). One of the most frequent reasons teenage girls give for leaving school is being pregnant. Teenage pregnancy is linked with poor school performance, and is also distributed differentially among girls of different race and ethnicity, with Hispanic girls the most likely to become pregnant and leave school. Typically, across all racial and ethnic groups, males tend to take less responsibility than females for decisions about child bearing and childcare.

11. Transitioning from High School

Most young people will enter some type of postsecondary school following high school graduation. Today, females are slightly more likely than males to obtain some form of postsecondary education. In a recent study by AAUW (1999), among those high school students who entered the work force after

graduation, males were significantly more likely than females to describe their employment as a career rather than a job, to say they never seriously considered attending college, and to believe that they could get a decent job without a college education. Females in this group were more likely to feel that lack of money was a barrier to education and that better information about financial aid would have influenced their chances of attending college.

Even though there have been substantial gains in closing the gender gap in elementary through high school, there are still small differences that can have substantial implications. For example, low self-esteem may encourage a girl or boy to take less challenging courses, which can ultimately affect future plans after high school. Young people, regardless of gender, have talent and ability in particular areas when given the opportunity to succeed. The life paths that young people choose need to be constructed on the basis of their talents and efforts not their gender.

See also: Coeducation and Single-sex Schooling; Education and Gender: Historical Perspectives; Education (Higher) and Gender; Educational Philosophy: Contemporary Views; Educational Philosophy: Historical Perspectives; Families as Educational Settings; Family and Gender; Family and Schooling; Gender and School Learning: Mathematics and Science; Gender-related Development; Public Policy Schools; Sex-role Development and Education; Statistical Systems: Education

Bibliography

American Association of University Women Educational Foundation 1992 *How Schools Shortchange Girls*. American Association of University Women, Washington, DC

American Association of University Women Educational Foundation 1995 *Growing Smart: What's Working for Girls in School*. American Association of University Women, Washington, DC

American Association of University Women Educational Foundation 1999 *Gaining a Foothold: Women's Transitions Through Work and College*. American Association of University Women, Washington, DC

Brown L, Gilligan C 1992 *Meeting at the Crossroads: Women's Psychology and Girls' Development*. Harvard University Press, Cambridge, MA

Campbell J R et al. 1998 *NAEP 1996 Trends in Academic Progress*. National Center for Education Statistics, Washington, DC

Connolly J A, Konarski R 1994 Peer self-concept in adolescence: Analysis of factor structure and of associations with peer experience. *Journal of Research on Adolescence* **4**: 385–403

Dwyer C, Johnson L 1997 Grades, accomplishments, and correlates. In: Willingham W, Cole N (eds.) *Gender and Fair Assessment*. Erlbaum, Mahwah, NJ pp. 127–56

Eccles J, Wigfield A, Schiefele U 1998 Motivation to succeed. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology: Vol. 3. Social Emotional and Personality Development*, 5th edn. Wiley, New York

Eder D, Parker S 1987 The cultural production and reproduction of gender: The effect of extracurricular activities on peer-group culture. *Sociology of Education* **60**: 200–13

Feingold A 1994 Gender differences in personality: A meta-analysis. *Psychological Bulletin* **116**: 429–56

Green P et al. 1995a *Trends Among High School Seniors, 1972–92*. National Center for Education Statistics, Washington, DC

Green P et al. 1995b *A Profile of the American High School Senior in 1992*. National Center for Education Statistics, Washington, DC

Hedges L, Friedman L 1993 Gender differences in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research* **63**: 94–105

Kleinfeld J 2000 The myth that schools shortchange girls. In: Paul E L (ed.) *Taking Sides: Clashing Views on Controversial Issues in Sex and Gender*. Dushkin/McGraw-Hill, Guilford, CT pp. 202–11

Kling K, Hyde J, Showers C, Buswell B 1999 Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin* **125**: 470–500

Marini M, Brinton M 1984 Sex typing in occupational socialization. In: Reskin B (ed.) *Sex Segregation in the Workplace: Trends, Explanations, Remedies*. National Academy Press, Washington, DC pp. 192–232

Orenstein P 1994 *School Girls: Young Women, Self-Esteem, and the Confidence Gap*. Doubleday, New York

Sadker M, Sadker D 1994 *Failing at Fairness: How America's Schools Cheat Girls*. c.Scribner's Sons, New York

Schneider B, Stevenson D 1999 *The Ambitious Generation: America's Teenagers, Motivated but Directionless*. Yale University Press, New Haven, CT

Thorne B 1993 *Gender Play: Girls and Boys in School*. Rutgers University Press, New Brunswick, NJ

US Department of Education 1993 *Single-sex Schooling: Perspectives from Practice and Research*. Office of Educational Research and Improvement, Washington, DC

US Department of Education 1995 *Condition of Education 1995*. The National Center for Education Statistics, Washington, DC

Willingham W, Cole N 1997 *Gender and Fair Assessment*. Erlbaum, Mahwah, NJ

Wirt J et al. 1999 *The Condition of Education 1999*. The National Center for Education Statistics, Washington, DC

B. Schneider

Education: Skill Training

1. Introduction

More than 200 years ago, Adam Smith argued that the incentives to invest in workers' skills are similar to the incentives to invest in physical capital. Workers or their employers invest in skills if the benefits or

'returns' from this investment exceed its cost. These investments manifest themselves in increased worker productivity. When workers receive the returns from investments in skills their wages, salaries and fringe benefits rise. When employers receive the returns from these investments their profits rise.

This article describes how the concept of 'human capital' anticipated by Adam Smith applies to decisions to acquire more schooling or to invest in vocational skills on the job. The idea is that workers and their employers act as if they evaluate investments in skills using the same cost-benefit calculus that they would use when assessing the net benefits from any other investment. The familiar forces of supply and demand influence these decisions. As more individuals acquire schooling, the supply of better-educated persons increases. All other things being equal (i.e., holding demand constant), the number of better educated persons employed increases, as does national output, but wages or salaries for this class of workers decline. Similarly, as more persons become educated, the supply of less educated persons declines and this group of workers' wages tends to increase.

An important difference between investing in workers and investing in new plants and equipment is that individuals can move. An employer that provides generous training opportunities to its employees may find that it has provided generous training opportunities to other employers' workers. This problem creates an important disincentive for employers to train their employees. Accordingly, this article focuses on the incentives for workers to acquire vocational skills either while in school or on-the-job and then discusses the evidence on the incidence, duration, and costs of training. The last section of this article comments on circumstances where these incentives may be insufficient and where the public sector might productively influence private training decisions.

2. *Incentives to Invest in Education and Training*

Economic theory predicts that competitive markets provide workers and employers with incentives to invest in training (Becker 1980). Individuals acquire more schooling because the value of future benefits associated with that schooling exceeds its costs. For the individual who decides to go to school, the benefit from schooling is the difference between her earnings with the schooling minus what her earnings would have been without the additional schooling. Individuals realize these earnings gains from schooling during each year of their working lives. Because individuals reap these gains in the future, they should be appropriately discounted to account for the 'time value of money.'

Most analyses of the cost of schooling focus on the direct costs and the opportunity costs of schooling. The former costs consist of the value of resources used to teach students new skills. The value of these resources to society is the value of what they would produce if they were used in some other activity. Analysts often estimate this amount by the (unsubsidized) cost of tuition, supplies, and transportation associated with attending school.

Opportunity costs are another important cost of schooling, especially for adults. This cost arises because when a person is in school they may not be working, or if they are working, they are likely to be working in a job in which their productivity is relatively low or they are working fewer hours. Consequently, schooling is associated with lost earnings for the individual and lost output for the economy. Analysts usually measure this cost by a person's wage times the number of additional hours that she would work had she not gone to school. Because older workers are more productive, the opportunity cost of schooling is higher for them than it is for teenagers and young adults.

The decisions of individuals about whether to attend school depend on the relative magnitudes of these varying components of the benefits and costs of schooling. A simple model of this decision process is instructive for its empirical implications. Suppose the earnings gain due to schooling is constant throughout a worker's working life and is denoted by B_i. The subscript associated with the term B_i indicates that the benefits from schooling vary among individuals in the population. If an individual is contemplating attending and completing college, then we can think of the term B_i as being the college wage premium. Suppose further that the worker's career has N years remaining, future gains are discounted at a rate of r, the direct costs of schooling are K, and earnings during the schooling period would have been Y_0. Under these conditions the decision to attend school depends on whether the benefits exceed its costs as is summarized by the following expression:

$$B_i\,(1-(1/r)^N)/r > K + Y_0 \qquad (1)$$

This expression implies that people who attend school (a) anticipate higher gains than those who do not attend; (b) anticipate working longer after completing school; and (c) face relatively low direct and opportunity costs of schooling (Heckman et al. 1999).

The foregoing expression explains why more academically able persons attend school and acquire more years of schooling than do their less able counterparts. These persons acquire new skills more efficiently and as a result experience a larger increase in their productivity as a result of their education. Young people also should be more likely to attend school than older persons for two reasons. First, they have more years left in their working lives during which they can realize a return on their investment. Second, they have less labor market experience and can not

earn as much if they were to go to work. The expression given in (1) also implies that public subsidies of education should increase enrollments in schooling because they lower an individual's cost of schooling. Such subsidies should encourage persons to enroll in school who would not have otherwise attended school, because their anticipated gains from schooling, B_i, are relatively low, or they are older, or they face a relatively high opportunity cost of schooling.

The familiar forces of supply and demand for workers with differing levels of schooling influence the decisions of individuals to attend school. If schooling is a good investment (i.e., the inequality given by (a) holds), more persons attend school. Because of the increase in the supply of better-educated workers, the prevailing wages for these workers decline and therefore so does the gain from schooling, B_i. As the gain from schooling declines, enrollment rates also decline. In this way, market forces determine school enrollment.

We observe an example of this adjustment process in the United States during the 1970s, when young college educated workers experienced a marked decline in their relative earnings. These individuals were from a large, well-educated birth cohort that entered the workforce at the same time that labor force participation rates for women were increasing rapidly. The increase in supply pushed down their earnings. A similar pattern was observed in Korea and Japan during the 1980s, when the ratio of college-educated to high school educated workers' earnings declined (Kim and Topel 1995).

The decision to invest in skills on-the-job involves the same cost-benefit analysis as that given by expression (1) above. The difference between the decisions to invest in schooling and in on-the-job training is that employee and employer interests are not necessarily the same. When the skills that workers acquire on the job are easily transferred to other employers, the incentives of firms to invest in their employees' skills are greatly diminished. The employer can profit from its training investments only if the employee stays with the firm. In order for an employer to have sufficient incentive to invest in employee skills, it needs some device that ties the worker to the firm. One way that companies can foster long-term relationships with their employees is to offer them some form of deferred compensation.

Another solution to the problem of providing on-the-job training to employees is for the worker to pay for the training. In principle, an employee could sign a contract in which she agrees to compensate the employer for training costs if she were to leave prior to some specific date. But another way to achieve this outcome is for the employee to pay for the training by accepting a lower salary during her early years on the job. This outcome would occur if essentially two market transactions take place when a firm hires a worker. The first transaction is the familiar one in which an individual sells her time for a wage. The second transaction results because workers seek training opportunities from employers (Rosen 1972). They buy this training from employers by accepting lower pay. Accordingly, the wages and salaries that we observe workers receiving are the difference between the value of their time and initial skills that they bring to the workplace minus the cost of the training provided by the employer.

3. *Sources of Training in the Private Sector*

There is a great deal of information about the amount of schooling obtained by people in countries throughout the world. In the United States, the web pages for the US Departments of Education and of Labor include a wealth of information on school enrollment, educational attainment, and academic performance by different groups in the population. By contrast, much less is known about sources of training outside of school. Estimates of the amount of training acquired on-the-job usually are imprecise.

There are several reasons why we do not know enough from existing data sources about the incidence and intensity of vocational training received by the workforce. First, most data sources contain only unrefined measures of the incidence, the source, and duration of training as reported by the employee. These data sources do not include measures of the costs of training other than the employee's wage. Even less common are employer reports of training activity that are representative of the experiences of the workforce. Further, if more effort were made to survey employers about their training expenditures, there remains the problem that many firms do not keep separate accounts on the resources spent on employee training.

A second reason that we do not know as much about on-the-job training as we do about schooling is that such training is conceptually difficult to measure. If firms provide training in the context of a formal program, it is relatively easy to characterize the content of the program and to measure the number of hours spent in training and the dollars spent on the program. However, much private sector training is provided informally, especially in the United States (Lynch 1994). Employees acquire new skills by trial and error, by watching their peers and supervisors perform their jobs, or while they are mentored by other employees. Such training represents an investment, because the employer incurs the costs associated with a new employee's lower productivity. These costs include the wages employers pay employees to shadow other workers, the lost output associated with the time that other workers spend away from their jobs, and the lost output associated with employee errors.

In one of the few large surveys of employers, US employers reported that only 12 percent of the training

that they provided their most recent hires during their first three months on the job occurred in a formal setting. The majority of time in training occurred while the employee was on the job (Barron et al. 1989). Surveys of representative samples of the US labor force also underscore the importance of informal training in the workplace. Among individuals who responded to special supplements of the US Bureau of the Census' Current Population Surveys (CPS) and who said that they needed specific skills in order to qualify for their current (or last) job, 47 percent reported that they acquired this training informally either on another job with the same employer or from a previous employer. Further, among those reporting that they had acquired training while in their current jobs, 37 percent said that those skills had been acquired informally while on the job (Ashenfelter and LaLonde 1997).

4. *The Incidence of Skill Training in the Workplace*

In spite of existing data sources, do we know roughly how often and how much training workers acquire after they leave school? Vocational training is relatively common. In the US for example, the CPS indicates that a majority of workers are employed in jobs that require specialized training. Among these workers, 41 percent reported that they had received additional training while in their current job.

An important and consistent finding in the literature on the incidence of training is that there is a strong positive correlation between an individual's formal schooling and whether she reports having received on-the-job training (Lillard and Tan 1992; Lynch 1992; Altonji and Spletzer 1991). This relationship between worker skills and the propensity to receive training is not unique to the United States. Studies of worker training in Great Britain and Germany also reveal a positive correlation between employees' formal schooling and whether they have received formal on-the-job training (See Nickell 1982; Greenlaugh and Stewart 1987; Booth 1991; and Pischke 1996). These findings hold especially for formal company programs, and training received off the job but paid for by the employer. The relationship between schooling and the incidence of skill training appears to be less strong when the training is provided informally. But, this result may simply underscore the difficulty we have in identifying and measuring informal training.

The positive relationship between schooling and the receipt of training suggests that training is more productively provided to more skilled or academically able people. In the context of expression (1), research indicates on-the-job training is more productive when firms provide it to individuals with larger gains (B_i) from schooling. Studies show that even among persons who have attained similar levels of schooling, the better educated among them receive more training (Altonji and Spletzer 1991). One nationally representative US survey found that less than 12 percent of young economically disadvantaged persons reported receiving vocational training during the previous 18 months. By comparison, the percentage of young adults in the population at large that reported receiving training during the same period was approximately 25 percent (Lillard and Tan 1992).

The strong complementarity between workers' prior skills and on-the-job training also has implications for workforce policy. People with limited or poor quality schooling are likely to receive limited additional training once they enter the labor force. Governments can not expect that private firms will compensate for deficiencies in the public education system by providing training in basic skills to their employees. Instead, research suggests that employers would increase the incidence of skill training if the educational attainment and general skills of the workforce rose.

5. *Estimates of the Amount Invested in Training*

Although the available measures of training investments are not very precise, it is still possible to estimate the amount invested in training each year. The most direct approach is to add together the value of the time that workers spent being trained or training others plus the resource costs associated with providing the training (Mincer 1994). This approach requires detailed information on the employees' use of time. On occasion such information is collected from workers in the form of a diary, in much the same way that government statistical agencies collect information for consumer expenditure surveys. Given the costs of gathering this type of data it is not often used to estimate the cost of training on an economy wide basis.

A less expensive approach for estimating the annual cost of on-the-job training is to infer the amount of training received by employees from the rate of their earnings growth during their careers (Mincer 1962). This inference relies on an important empirical regularity. The wages and earnings of individuals rise with the amount of labor force experience that they have acquired, but rise at a decreasing rate. When individuals are young, their wages and salaries rise relatively rapidly as they acquire more work force experience. But as they move into middle age the rate of growth in their earnings slows and for some groups stops all together. This rise in earnings indicates that young workers become more productive as they get older.

If we assume that all of the increase in workers' earnings results from training received on-the-job during their careers, and if we assume that investments

in training earn a 'normal' rate of return, then we can infer the total investment in private training each year from the growth in workers' earnings. In keeping with these assumptions, several studies report that employees' wage growth is largest when they report that they are receiving training in their current positions. As they begin to report that they are receiving less training in their current jobs, their wage growth also slows (Brown 1989). Mincer's and Higuchi's (1993) study of US and Japanese manufacturing industries finds that workers in industries in which firms provide more training also have earnings that rise more rapidly with years of service. In another study of professional employees in a large US manufacturing firm, about 18 percent of the growth of employees' earnings is explained by participation in the company's formal on-the-job training programs (Bartel 1992).

If workers receive training that raises their general skills, the foregoing approach probably overstates private investments in skill training. There are other reasons that workers' earnings grow during their careers besides training. For example, workers change jobs frequently, especially early in their careers, in order to obtain better pay. This 'job matching' process leads to higher worker productivity as workers' skills become better matched with the tasks demanded by employers. But the resulting rise in worker productivity does not come about from investments in training, but rather from investments in 'job search.'

By contrast, the indirect approach to estimating annual training investments could understate the total investment in training if some of the skills that workers acquire are 'firm-specific.' These are skills that raise workers' productivity with their current employer, but not elsewhere. Because workers can not transfer these skills to another firm, it is more likely that the employer is willing to invest in training. Under these circumstances, the wage gains from training should be smaller than the productivity gains. This gap between workers' wage gains and their productivity gains indicates that the employer receives a return on its investment in training. Among the few studies that examine the relationship between training and productivity, at least two of them suggest that the effect on worker productivity of training received during the first few months on the job is twice as large as its effect on wages (Barron et al. 1989; Bishop 1994). This finding suggests that a significant component of vocational training may involve 'firm-specific' skills provided by the employer.

Studies of on-the-job and vocational training in the United States suggest that both the direct and indirect approaches yield similar estimates of annual investments in training. According to these studies, the annual cost of vocational training amounts to between $150 and $200 billion per year. By comparison, the annual costs of formal schooling in the US, including the forgone earnings while individuals are in school, have been estimated at around $500 billion dollars.

6. *The Impact of Training on Workers' Earnings*

Many studies find a significant positive correlation between the incidence and the amount of training received by employees and their wages. This correlation might indicate that on-the job training yields positive returns. But it also could simply indicate, as discussed above, that individuals who are more productive and highly paid also are more likely to receive training from their employers. Distinguishing between these two hypotheses is difficult. Employers do not provide training randomly to their employees nor do workers randomly acquire training on their own. However, studies that attempt to address this difficulty continue to find a positive relationship between measures of the incidence and duration of training and workers' wages. Research suggests that the gross returns to wages from a year's worth of vocational training is approximately 10 percent per year or more (Mincer 1994).

This evidence on the impact of skill training suggests that the returns from on-the-job training are comparable if not larger than the returns from formal schooling. One explanation for this finding is that despite analysts' best efforts, they still have not satisfactorily accounted for unobserved attributes that simultaneously influence wages (or productivity) and the propensity to receive training. There are several reasons, however, to expect that the returns from skill training to be larger than the returns from formal schooling. First, the returns from formal schooling may be lower than the returns from training because some schooling costs probably are better characterized as consumption instead of as an investment. Second, because workers acquire some specific skills and at the same time may leave the firm, the returns from training need to be larger than the returns from schooling in order to compensate both the worker and the employer for the risk associated with turnover.

Finally, as new technologies are introduced some skills may become obsolete. One study suggests that the value of workers' vocational skills may depreciate by as much as 10 percent per year (Lillard and Tan 1992). As a result, if workers' skills are likely to depreciate, the returns from vocational skills must be larger in order for there to be sufficient incentive to invest in training.

7. *Concluding Remarks*

Increased interest in the incentives to invest in vocational skills has coincided with several economic developments during the last two decades of the twentieth century. First, the rise in earnings inequality in the United States and Great Britain appears prompted in part by a decline in employers' demand for the services of relatively unskilled workers (Katz and Murphy 1992). The long-term prospects of many

of these individuals is poor because students in these countries appear to end their formal schooling less skilled and ready to learn than their counterparts in other industrialized nations (Berg 1994). Second, a large number of workers displaced from traditional sectors of industrialized economies have suffered long-term earnings losses (Jacobson et al. 1993). As these displacements spread to the services sectors, interest in the feasibility of retooling such workers has increased. Finally, the low unemployment rates during the late 1990s in North America have pressured employers to put greater emphasis on training their existing workers rather than to expect to be able to hire employees with the required skills.

Some commentators and researchers argue that these developments require attention from the public sector because employers and their employees lack adequate incentives to invest in skill training (Lynch 1994). To be sure, as this article has described, a well functioning competitive labor market provides incentives for workers to acquire and their employers to provide skill training by compensating workers with higher wages and firms with increased profits. Individuals face similar incentives when deciding whether to acquire more schooling. But there may be instances when these incentives are insufficient. Imperfect capital markets may restrict access to funds, which prevent individuals and firms from making otherwise productive investments in skills. These instances also arise when vocational skills can be easily transferred from one employer to another. The possibility that employers can poach other firms' employees diminishes incentives to invest in employee training. Likewise, imperfect or asymmetric information about workers' skills also may reduce incentives to invest in vocational training. Few studies address this possibility, but one that does suggests that skills are not fully 'portable' at least during the first year on the job (Bishop 1994). If it is difficult for workers to reveal their skills to a potential employer, employees may invest less in new skills than they would otherwise.

The merits of policies that require the public sector to play a role in private training decisions depend on whether companies and their employees underinvest in training. Such a view remains open for debate. As discussed above, there are several reasons why such a debate will be difficult to resolve. First, measures of vocational training tend to be unrefined and often not comparable among studies of firms in the same country, much less studies of different countries. Second, some components of training, especially informal on-the-job training, are conceptually difficult to measure. Evidence discussed here indicates that such informal training is an important source of training, especially in the United States. Finally, even if researchers concluded there has been too little investment in vocational training in industrialized economies, the conceptual difficulties associated with measuring different types of training indicate that it will be hard to enforce employer mandates to provide more training (Brown 1990).

See also: Adult Education and Training: Cognitive Aspects; Cost–Benefit Analysis; Education and Economic Growth; Education and Employment; Education, Economics of; Human Capital: Educational Aspects; Vocational Education and Training

Bibliography

Altonji J G, Spletzer J R 1991 Worker characteristics, job characteristics, and receipt of on-the-job training. *Industrial and Labor Relations Review* **45**: 58–79

Ashenfelter O, LaLonde R 1997 The Economics of Training. In: Lewin D, Mitchell D, Zaidi M (eds.) *Handbook of Human Resources*. JAI Press, Greenwich, CT

Barron J M, Black D A, Lowenstein M A 1989 Job matching and on-the-job training. *Journal of Labor Economics* **7**: 1–19

Bartel A P 1992 Training, wage growth and job performance: Evidence from a company database. Working Paper No. 4027, National Bureau of Economic Research, March

Becker G S 1980 Investment in human capital: Effects on earnings. In: *Human Capital*, 2nd edn. University of Chicago Press, Chicago, pp. 15–44

Berg P 1994 Strategic adjustments in training: A comparison with Germany. In: L Lynch (ed.) *Training and the Private Sector*. University of Chicago Press, Chicago, pp. 77–107

Bishop J 1994 The impact of previous training on productivity and wages. In: L Lynch (ed.) *Training and the Private Sector*. University of Chicago Press, Chicago, pp. 161–99

Booth A 1991 Job-related formal training: Who receives it and what is it worth? *Oxford Bulletin of Economics and Statistics* **53**(3): 281–94

Brown C 1990 Empirical evidence on private training. In: Ehrenberg R (ed.) *Research in Labor Economics*, Vol. 11. JAI Press, Greenwich, CT, pp. 97–113

Brown J N 1989 Why do wages increase with tenure? On-the-job training and life cycle wage growth observed within firms. *American Economic Review* **79**: 971–91

Greenlaugh C, Stewart M 1987 The effects and determinants of training. *Oxford Bulletin of Economics and Statistics* **49**: 171–90

Heckman J, LaLonde R, Smith J 1999 The Economics and Econometrics of Active Labor Market Programs. In: Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*, Vol. 3, Elsevier Science B.V., Amsterdam

Jacobson L, LaLonde R, Sullivan D 1993 *The Costs of Worker Dislocation*. W E Upjohn Institute for Employment Research, Kalamazoo, MI

Katz L, Murphy K M 1992 Changes in relative wages, 1963–1987: Supply and demand factor. *Quarterly Journal of Economics* **107**(1): 35–78

Kim D, Topel R 1995 Labor Markets and Economic Growth: Lessons from Korea's Industrialization, 1970–1990. In: Freeman R, Katz L (eds.) *Differences and Changes in Wage Structures*. University of Chicago Press, Chicago

Lillard L A, Tan H W 1992 Private sector training: Who gets it and what are its effects? In: Ehrenberg R (ed.) *Research in Labor Economics*, Vol. 13, JAI Press, Greenwich, CT, pp. 1–62

Lynch L M 1992 Private sector training and the earnings of young workers. *American Economic Review* **82**: 299–312

Lynch L M 1994 Payoff to alternative training strategies at work. In: R Freeman (ed.) *Working under Different Rules*. Russell Sage Foundation, New York
Mincer J 1962 On-the-job training: Costs, returns, and some implications. *Journal of Political Economy* **70**(Supplement): 50–79
Mincer J 1994 Investment in US education and training. Working Paper No. 4844, National Bureau of Economic Research, August. Cambridge, MA
Mincer J, Higuchi Y 1993 Wage structures and labor turnover in the United States and Japan. *Journal of Japanese and International Economics* **2**: 97–133
Nickel S 1982 The determinants of occupational success in Britian. *Review of Economic Studies* **59**: 43–53
Pishchke J 1996 Continuous training in Germany. Working Paper National Bureau of Economic Research Inc., Massachusetts Institute of Technology
Rosen S 1972 Learning and Experience In the Labor Market. *Journal of Human Resources* **7**(3): 326–42

R. LaLonde

Educational Assessment: Major Developments

The role of assessment and evaluation in education has been crucial, probably since the earliest approaches to formal education. However, change in this role has been dramatic in the last few decades, largely due to wider developments in society. The most dramatic change in our views of assessment is represented by the notion of assessment as a tool for learning. Whereas in the past, we have only seen assessment as a means to determine measures and thus certification, there is now a realization that the potential benefits of assessing are much wider and impinge on all stages of the learning process. In this article, we will outline some of the major developments in educational assessment, and we reflect on the future of education within powerful learning environments, where learning, instruction, and assessment are fully integrated.

1. *Definition of the Concept*

Much has been written about assessment, measurement, and evaluation in educational practice. However, there are different meanings and it is necessary to draw as clear a distinction as possible between these three different concepts. In this context, semantic distinctions embedded in the literature are considered. What measurement, assessment, and evaluation have in common is the operation of testing. Each commonly (but not always) makes use of tests, but none of the three operations is synonymous with testing. Moreover, the types of the instruments used for each of the three operations may be different (Fig. 1).

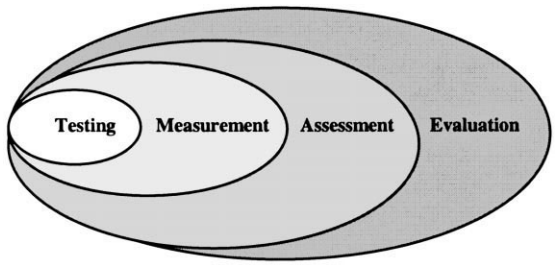

Figure 1
Relationships between the most commonly used different concepts in educational assessment

Worthen and Van Dusen (1991) define 'measurement' as the process of making observations about some attribute or characteristic of a person, a behavior, an event, or even an object, using clearly specified procedures to translate those observations into quantifiable or categorical form. Educational measurement is broader than educational testing. In addition to the fact that measurement frequently employs tests, it can also employ instruments, checklists, and other methods that are not considered as testing tools. However, measurement is rarely undertaken for its own sake. It is a basic research procedure that may be a component of both assessment and evaluation. Measurement itself places no value on what is measured or on the results of the measurement process.

In evaluation, quite the opposite is the case. The major attributes studied in evaluation are precisely chosen because they represent educational values. Objectives are educational values and they define what is sought to develop in learners as a result of exposing them to a set of educational experiences. Evaluation literature provides multiple perceptions of evaluation, differing in level of abstraction and in their descriptive or judgmental character. At the most general level, evaluation has been defined as a formal appraisal of the quality of educational phenomena (Popham 1988), as a determination of goal achievement, or as a measurement of merit (Scriven 1967). These definitions can be characterized as judgmental. More elaborated definitions are given, describing evaluation as 'data-collection of education materials for providing information for decision-makers.' These definitions have a more descriptive character. The definition of evaluation as 'an act of collecting systematic information regarding the nature and quality of educational objects' (Nevo 1995) combines description and judgment.

Closely related to the notion of measurement and evaluation is the term 'assessment.' According to Worthen and Van Dusen (1991), 'assessment' refers to

the process of transferring test scores—or results of alternative means of appraising learning outcomes—into statements about student performance. Such learner assessment is built on the bedrock of measurement's value neutrality, but its capstone is evaluation whenever the assessment data contain implicit judgments of the merit or worth of the performance of individuals or groups of students.

The difference between assessment and evaluation is also made clear by Scriven (1967). 'The goal of evaluation is that of providing answers to significant evaluation questions (e.g., assessing student performance), whereas its roles refer to various ways in which those answers are used (e.g., using these achievement data to determine the adequacy of the curriculum).' The distinction made by Nevo (1995) between 'primary' and 'secondary' functions of student evaluation is relevant here. Primary functions of assessment are those in which the student is the object of the evaluation and it is intended to provide information for the improvement of learning to certify individual accomplishment. Secondary functions of student assessment are those where the object of evaluation is not the student, but rather the teacher or the school or some other object, and the results of the evaluation are used for school accountability, or to control students or motivate them. Nevo uses the British synonym of 'student assessment,' for student evaluation. He also points at the difference between this British use of student assessment, and the recent distinctions between 'student testing' and 'student assessment,' referring to performance testing and authentic testing.

2. *Functions and Criteria of Student Assessment*

One of the main issues in student assessment is clarity about the type of instructional decision that is supported by the assessment procedure. Nitko (1989) identifies four types of instructional decisions that are supported by tests. These are discussed below.

2.1 *Placement Decisions*

This is deciding in which instructional sequence or at what level in the instructional sequence a student should begin in order to avoid unnecessary repetition of what is already known, and to allow the more rapid attainment of new goals. Placement or prior knowledge state tests are given before a student begins a unit of instruction. They may focus on the prerequisite knowledge and skills for proposed learning or on the outcome knowledge and skills to determine whether students have already attained the desired outcome of the instruction. Thus, instructors use prior knowledge state tests to place students in instructional modes by (a) determining the degree to which prerequisite entry behaviors or skills are present or absent, (b) determining entering mastery of course objectives, and also for (c) matching students to alternative instructional modes, based on student characteristics.

2.2 *Monitoring Decisions*

This is deciding (a) whether students are learning appropriately, and (b) whether the assigned learning activity is working effectively, or a different activity should be assigned. So-called progress assessments or assessments of growth are the instruments for the assessment of a student's progress toward the final learning objectives. Progress assessments provide feedback to students on how their learning is progressing and can lead to the (formative) decision that there is a need to remediate or relearn a certain instructional unit. Progress assessments, taken during instruction, are used by instructors to (a) choose or modify subsequent learning activities, (b) prescribe remediation of group or individual deficiencies, and (c) provide on-going feedback to the student for the purpose of directing advanced or remedial study.

2.3 *Attainment Decisions*

This involves deciding, at the end of a particular instructional segment, whether a student has attained the desired instructional goals. Final tests are used for certifying student learning and evaluating teacher effectiveness.

2.4 *Diagnostic Decisions*

This is deciding which learning outcomes a student has not acquired and the probable cause(s) of the failure to acquire them, in order to remediate or correct incomplete or erroneous prior learning. Diagnostic assessments are designed to provide specific information on individual learning deficiencies and misunderstandings. When prior knowledge state assessments, progress assessments, and final assessments are designed and interpreted properly they can also provide information of a diagnostic nature since students and teachers can use this information in order to regulate learning processes.

Another issue in student assessment relates to criteria setting. A distinction can be made between norm-referenced and criterion-referenced tests. Norm-referenced tests determine the level of students relative to some groups (e.g., class, school, age-group), whereas criterion-referenced tests determine the level of a student relative to some predetermined criterion set by the curriculum, the teacher, or some authority. Nevo (1995) mentions that while norm-referenced tests are suitable for selection procedures, criterion-referenced tests are the tests schools predominantly need.

3. Innovations in Assessment

3.1 Consequences of the Developments in Society

Economic and technological change, bringing with it significant changes in the requirements of the labor market, poses increasing demands on education and training. For many years, the main goal of education has been to make students knowledgeable within a certain domain. Building a basic knowledge store was the core issue (Dochy and McDowell 1997). Students taking up positions in modern organizations in the twenty-first century need to be able to analyze information, to improve their problem-solving skills and communication, and to reflect on their own role in the learning process. People increasingly have to be able to acquire knowledge independently and use this body of organized knowledge in order to solve unforeseen problems. As a consequence, education should contribute to the education of students as lifelong learners.

3.2 Paradigm Change: From Testing Towards Assessment

Many authors (Dochy and McDowell 1997) have pointed out the importance of instruction to promote students' abilities as thinkers, problem-solvers, and inquirers. Underlying this goal is the view that meaningful understanding is based on the active construction of knowledge and often involves shared learning. It is argued that a new form of education requires reconsideration about assessment (Dochy et al. 1999). Changing towards new forms of learning, with a status quo for evaluation, undermines the value of innovation. Students do not invest in learning that will not be honored. Assessment is the most determining factor in education for the learning behavior of students. Traditional didactic instruction and traditional assessment of achievement are not suited to modern educational demands. Such tests were generally designed to be administered following instruction, rather than to be integrated with learning. As a consequence, due to their static and product-oriented nature, these tests not only lack diagnostic power but also fail to provide relevant information to assist in adapting instruction appropriately to the needs of the learner. Furthermore, standard test theory characterizes performance in terms of the difficulty level of response choice items and primarily focuses on measuring the amount of declarative knowledge that students have acquired.

This view of performance is at odds with current theories of cognition. Achievement assessment must be an integral part of instruction, in that it should reflect, shape, and improve student learning. Assessment procedures should not only serve as a tool for crediting students with recognized certificates, but should also be used to monitor progress and, if needed, to direct students to remedial learning activities. The view that the evaluation of students' achievements is something which happens at the end of the process of learning is no longer widespread; assessment is now represented as a tool for learning (Dochy and McDowell 1997).

Birenbaum and Dochy (1996) describe this movement as a paradigm change between two cultures in the measurement of achievement and relate this to developments in the learning society. In the traditional so-called testing culture, instruction and testing are considered to be separate activities. Instruction is the responsibility of the teacher, whereas testing is the responsibility of the psychometric expert, who can use elaborate procedures for test development and sophisticated psychometric models for the analysis of test responses. The work of such experts within testing agencies was stimulated by the demand for objectivity and fairness in testing and a high level of standardization where important decisions were to be based on test scores. Stemming from such demands, common tools for assessment in education were easy-to-score items such as multiple-choice items and true/false items. The testing culture fits well with the traditional approach to education where teaching is seen as an act of depositing the content which students receive, memorize, and reproduce. The most efficient way of testing this process is by standardized tests delivered by external testing agencies.

The changing learning society has generated the so-called assessment culture as an alternative to the testing culture. The assessment culture strongly emphasizes the integration of instruction and assessment. Students play far more active roles in the evaluation of their achievement. The construction of tasks, the development of criteria for the evaluation of performance, and the scoring of the performance may be shared or negotiated among teachers and students. The assessment takes all kinds of forms such as observations, text- and curriculum-embedded questions and tests, interviews, performance assessments, writing samples, exhibitions, portfolio assessment, and project and product assessments. Several labels have been used to describe subsets of these alternatives, with the most common being 'direct assessment,' 'authentic assessment,' 'performance assessment,' and 'alternative assessment.'

3.3 New Methods of Assessment

Investigations of new approaches (e.g., Birenbaum and Dochy 1996) illustrate the development of more 'in context' and 'authentic' assessment. Nisbet (1993) defines the term authentic assessment as 'methods of assessment which influence teaching and learning positively in ways which contribute to realizing educational objectives, requiring realistic (or "authentic")

tasks to be performed and focusing on relevant content and skills, essentially similar to the tasks involved in the regular learning processes in the classroom' (p. 35).

Assessment of such 'authentic' tasks is highly individual and contextualised. The student gets feedback about the way he or she solved the task and about the quality of the result. Evaluation is given on the basis of different 'performance tasks,' performed and (reviewed) assessed at different times. The evaluation criteria have to be known in advance. When students know the criteria and know how to reach them, they will be more motivated and achieve better results. This form of evaluation gives a more complete and realistic picture of a student's ability (achievement). It not only evaluates the product, but also evaluates the process of learning. Students get feedback about their incorrect thinking strategies. Within the new forms of 'new assessment,' much attention is paid to authentic problem-solving, case-based exams, portfolios, and the use of peer and self-assessment. (Birenbaum and Dochy 1996).

In traditional education, the question 'who takes of the exam and who defines the criteria?' is seldom asked. Most of the time, it is the teacher. New forms of education do pose this question. Student themselves, other students, or the teacher and students together, are responsible for assessment. The type of student self-assessment referred to most frequently in the literature is a process which involves teacher-set criteria and where students themselves carry out the assessment and marking. Another form of student self-assessment is the case where a student assesses herself or himself, on the basis of criteria which she or he has selected, the assessment being either for the student's personal guidance or for communication to the teacher or others. According to Hall (1995) there are two critical factors for genuine self-assessment: the student not only carries out the assessment, but also selects the criteria on which the assessment is based. Similarly, peer assessment can indicate that fellow students both select the criteria and carry out the assessment. Any situation where the tutor and students share in the selection of criteria and/or the carrying-out of the assessment is more accurately termed co-assessment (Hall 1995). However, it is still frequently the case that teachers control the assessment process, sometimes assisted by professional bodies or assessment experts, whereas students' assessments and criteria are taken seriously but considered to be additional to the assessment undertaken by the teacher or professor rather than replacing it. Implementing forms of self-, peer, and co-assessment may decrease the time-investment professors would otherwise need to make in more frequent assessment (Dochy et al. 1999). In addition to that advantage, using these assessment forms assists the development of certain skills for the students, e.g., communication skills, self-evaluation skills, observation skills, and self-criticism.

4. *Assessing New Assessment Forms*

Judgments regarding the cognitive significance of an assessment begin with analyses of the cognitive requirements of tasks as well as the ways in which students try to solve them. Two criteria by which educational and psychological assessment measures are commonly evaluated are validity and reliability. One can say that based on these criteria the results above are not yet consistent and, depending upon the assessment form, there is a larger or smaller basis to state that the evaluation is acceptable.

It is however important to note that Birenbaum (1996) mentions that the meaning of validity and reliability has recently expanded. Dissatisfaction with the available criteria, which were originally developed to evaluate indirect measures of performance, is attributed to their insensitivity to the characteristics of a direct assessment of performance. The most important element of new assessment models is the reflection of the competencies required in real-life practice. The goal is to ensure that the success criteria of education or training processes are the same as those used in the practice setting. Hence, as notions of fitness of purpose change, and as assessment of more qualitative areas are developed, the concepts of validity and reliability encompassed within the instruments of assessment must also change accordingly. This means that we should widen our view and search for other and more appropriate criteria. It should not be surprising that a new learning society and consequently a new instructional approach and a new assessment culture cannot be evaluated solely on the basis of the pre-era criteria.

5. *Validity-related Issues*

Although performance assessment appears to be a valid form of assessment, in that it resembles meaningful learning tasks, this measure may be no more valid than scores derived from response choice items. Evidence is needed to ensure that assessment requires the high-level thought and reasoning processes that they were intended to evoke.

The authors of the 1985 Standards define test validity as 'a unitary concept, requiring multiple lines of evidence, to support the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores' (AERA, APA, NCME 1985, p. 9). All validity research should be guided by the principles of scientific inquiry reflected in construct validity.

Within the construct validity framework, almost any information gathered in the process of developing and using an assessment is relevant when it is evaluated against the theoretical rationale underlying the proposed interpretation and inferences made from test scores (Moss 1992). Thus, validation embraces all the experimental statistical and philosophical means

by which hypotheses are evaluated (Messick 1994). Validity conclusions, then, are best presented in the form of an evaluative argument (Cronbach 1989) that integrates evidence to justify the proposed interpretation against plausible alternative interpretations.

Kane's argument-based approach is in line with Cronbach's view on validity. According to Kane (1992), to validate a test-score interpretation is to support the plausibility of the corresponding interpretative argument with appropriate evidence (a) for the inferences and assumptions made in the proposed interpretative argument, and (b) for refuting potential counter arguments. The core issue is not that we must collect data to underpin validity, but that we should formulate transparent, coherent, and plausible arguments to underpin validity.

Authors like Kane and Cronbach use validity principles from interpretative research traditions, instead of psychometric traditions, to assist in evaluating less standardized assessment practices.

Other criteria suggested for measuring validity of new assessment forms are the transparency of the assessment procedure, the impact of assessment on education, directness, effectiveness, fairness, completeness of the domain description, practical value, and meaningfulness of the tasks for candidates, and authenticity of the tasks. According to Messick (1994), these validation criteria are, in a more sophisticated form, already part of the unifying concept of validity, which he expressed in 1989. He asserted that validity is an evaluative summary of both evidence for and the actual as well as potential consequences of score interpretation and use. The more traditional conception of validity as 'evidence for score interpretation and use' fails to take into account both evidence of the value implications of score interpretation and the social consequences of score use.

Messick's unifying concept of validity encompasses six distinguishable parts—content, substantive, structural, external, generalizability, and consequential aspects of construct validity—that conjointly function as general criteria for all educational and psychological assessment. The content aspect of validity means that the range and type of tasks used in assessment must be an appropriate reflection (content relevance, representativeness) of the construct-domain. Increasing achievement levels in assessment tasks should reflect increases in expertise of the construct-domain. The substantive aspect emphasizes the consistency between the processes required for solving the tasks in assessment, and the processes used by domain-experts in solving tasks (problems). Further, the internal structure of assessment— reflected in the criteria used in assessment tasks, the interrelations between these criteria, and the relative weight placed on scoring these criteria—should be consistent with the internal structure of the construct-domain. If the content aspect (relevance, representativeness of content, and performance standards) and the substantial aspect of validity is guaranteed, score interpretation based on one assessment task should be generalizable to other tasks assessing the same construct. The external aspect of validity refers to the extent that the assessment scores' relationship with other measures and non-assessment behaviors reflect the expected high, low, and interactive relations. The consequential aspect of validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use (Messick 1994).

In line with Messick's conceptualization of consequential validity, Frederiksen and Collins (1989) proposed that assessment has 'systematic validity' if it encourages behaviors on the part of teachers and students that promote the learning of valuable skills and knowledge, and allows for issues of transparency and openness, that is to access the criteria for evaluating performance. Encouraging deep approaches to learning is one aspect which can be explored in considering consequences. Another is the impact which assessment has on teaching. Dochy and McDowell (1997) argue that assessing high-order skills by authentic assessment will lead to the teaching of such high-order knowledge and skills.

With today's emphasis on high-stakes assessment, two threats to test validity are worth mentioning: construct under-representation and construct-irrelevance variation. In the case of 'construct-irrelevance variation,' the assessment is too broad, containing systematic variance that is irrelevant to the construct being measured. The threat of 'construct under-representation' means that the assessment is too narrow and fails to include important dimensions of the construct being measured (Dochy and Moerkerke 1997).

5.1 Special Points of Attention for New Assessment Forms

The above implies in our view that other criteria suggested for measuring validity of new assessment forms will need to be taken into account, i.e., the transparency of the assessment procedure, the impact of assessment on education, directness, effectiveness, fairness, completeness of the domain description, practical value and meaningfulness of the tasks for candidates, and authenticity of the tasks.

Authentic assessment tasks are more sensitive to construct under-representation and construct-irrelevance variation, because they are often loosely structured, so that it is not always clear to which construct-domain inferences are drawn. Birenbaum (1996) argues that it is important to specify accurately the domain and to design the assessment rubrics so they clearly cover the construct-domain. Messick (1994) advises the adoption of a construct-driven approach to the selection of relevant tasks and the development of scoring criteria and rubrics, because

it makes salient the issue of construct under-representation and construct-irrelevance variation.

Another difficulty with authentic tasks, with regards to validity, concerns rating authentic problems. The literature reveals that there is much variability between raters in scoring the quality of a solution. Construct under-representation in rating is manifested as an omission of assessment criteria or ideosyncratic weighting of criteria such that some aspects of performance do not receive sufficient attention. Construct-irrelevance variation can be introduced by the rater's application of extraneous, irrelevant, or idiosyncratic criteria. Suggestions for dealing with these problems in the literature include constructing guidelines, using multiple raters, and selecting and training raters. However, even within communities of problem-solvers, there is no common sense about the solution to an ill-defined problem. Since universally accepted (objective) criteria do not exist, argumentation can be an important means to judge the quality of solutions for an ill-structured problem. Therefore, product and process quality criteria need to be identified, defined, and communicated to candidates and raters (Dochy and Moerkerke 1997).

6. *Reliability-related Issues*

Reliability in classical tests is concerned with the degree in which the same results would be obtained on a different occasion, in a different context, or by a different assessor. Inter- and intra-rater agreement is used to monitor the technical soundness of performance assessment rating. However, when these conventional criteria are employed for new assessments (for example using authentic tasks), results tend to compare unfavorably to traditional assessment, because of a lack of standardization of these tasks.

The unique nature of new forms of assessment has affected the traditional conception of reliability, resulting in the expansion of its scope and a change in weights attached to its various components (Birenbaum 1996). In new assessment forms, it is not about achieving a normally distributed set of results. The most important question is to what extent the decision 'whether or not individuals are competent' is dependable. Differences between ratings sometimes represent more accurate and meaningful measurement than would absolute agreement. Measures of inter-rater reliability in authentic assessment, then, do not necessarily indicate whether raters are making sound judgment and do not provide bases for improving technical quality. Suen et al. (in press) argue that the focus on objectivity among raters as a desirable characteristic of an assessment procedure leads to a loss of relevant information. In high-stake decisions, procedures which include ways of weighing high-quality information from multiple perspectives may lead to a better decision than those in which information from a single perspective is taken into account. Measuring the reliability of new forms of assessment stresses the need for more evidence in a doubtful case, rather than to rely on making inferences from a fixed and predetermined set of data.

In line with these visions on reliability is Moss' idea (1992) about reliability. She asserts that a hermeneutic approach of 'integrative interpretations based on all relevant evidence' is more appropriate for new assessment because it includes the value and contextualized knowledge of the reader, than the psychometric approach that limits human judgment 'to single performances,' results of which are then aggregated and compared with performance standards.

7. *Conclusion: Future Research and Developments*

The assessment culture leads to a change of our instructional system from a system that transfers knowledge into students' heads to one that tries to develop students who are capable of learning how to learn. The current societal and technological context requires education to make such a change. The explicit objective is to interweave assessment and instruction in order to improve education. Although promising results have been reported, the first attempts to introduce new assessment procedures have not been entirely positive. Madaus and Kellaghan (1993) reported problems with the organization, time, and costs of assessment programs. A number of lessons can be learned from the early applications of new assessment programs.

Firstly, one should not 'throw the baby out with the bath water.' Objective tests are very useful for certain purposes, such as high-stake summative assessment of an individual's achievement, although they should not dominate an assessment program. Increasingly, measurement specialists recommend so-called balanced or pluralistic assessment programs, where multiple assessment formats are used (Birenbaum 1996). There are several motives for these pluralistic assessment programs: a single assessment format cannot serve several different purposes and decision-makers; and each assessment format has it own method variance, which interacts with people.

There is a need to establish a system of assessing the quality of new assessment and implement quality control. Various authors have recently proposed ways to extend the criteria, techniques, and methods used in traditional psychometrics (Kane 1992). Others, like Messick (1994), oppose the idea that there should be specific criteria, and claim that the concept of construct validity applies to all educational and psychological measurements, including performance assessment.

Further, there is a need for well-designed and well-evaluated heuristics, which would help teachers to implement and design high-quality new assessment procedures. Teachers need clear criteria for the development and use of assessment procedures, and easy-to-use techniques for quality improvement and quality control.

Surely, there is no one ideal assessment format for all cases. All assessment formats can have negative effects on teaching and learning. From this perspective, Frederiksen's criticism (1984) can be more broadly interpreted than simply as an exposure of the negative effects of multiple-choice items; it also shows that some aspects of any assessment may have negative as well as positive educational consequences. However, new assessment forms with their potential to develop and assess learning achievements which are congruent with contemporary views of the nature of knowledge and the needs of society for educated persons, is an avenue which we must continue to explore.

See also: Classroom Assessment; Competencies and Key Competencies: Educational Perspective; Content Validity; Educational Evaluation: Overview; Educational Research for Educational Practice; Generalizability Theory; Instructional Design; Instructional Psychology; Reliability: Measurement

Bibliography

American Educational Research Association, American Psychological Association, National Council on Measurement in Education 1985 *Standards for Educational and Psychological Testing*. Washington, DC

Birenbaum M 1996 Assessment 2000: Towards a pluralistic approach to assessment. In: Birenbaum M, Dochy F R J C (eds.) *Alternatives in Assessment of Achievement, Learning Processes and Prior Knowledge*. Kluwer Academic, Boston

Birenbaum M, Dochy F R J C (eds.) 1996 *Alternatives in Assessment of Achievement, Learning Processes and Prior Knowledge*. Kluwer Academic, Boston

Cronbach L J 1989 Construct validation after thirty years. In: Linn R L (ed.) *Intelligence: Measurement, Theory and Public Policy*. Invington, New York, pp. 147–71

Dochy F, McDowell L 1997 Assessment as a tool for learning. *Studies in Educational Evaluation* **23**(4): 279–98

Dochy F, Segers M, Sluijsmans D 1999 The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education* **24**(3): 331–50

Dochy F, Moerkerke G 1997 The present, the past and the future of achievement testing and performance assessment. *International Journal of Educational Research* **27**(5): 415–32

Frederiksen N 1984 The real test bias influences of testing on teaching and learning. *American Psychologist* **39**(3): 193–202

Frederiksen J R, Collins A 1989 A system approach to educational testing. *Educational Researcher* **18**(9): 27–32

Hall K 1995 *Co-assessment: Participation of Students with Staff in the Assessment Process. A report of Work in Progress*. Paper given at the 2nd European Electronic Conference on Assessment and Evaluation, EARLI-AE list European Academic and Research Network (EARN) (http://listserv.surfnet.nl/archives/earli-ae.html)

Kane M 1992 An argument-based approach to validity. *Psychological Bulletin* **112**: 527–35

Madaus G F, Kellaghan K 1993 British experience with authentic testing. *Phi Delta Kappan* **74**: 458–569

Messick S 1994 The interplay of evidence and consequences in the validation performance assessments. *Educational Researcher* **23**(2): 13–22

Moss P A 1992 Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research* **62**(3): 229–58

Nevo D 1995 *School-Based Evaluation. A Dialogue for School Improvement*. Pergamon, London

Nisbet J 1993 Introduction. In: *OECD Curriculum Reform: Assessment in Question*. Organization for Economic Cooperation and Development, Paris, pp. 25–38

Nitko A J 1989 *Designing Tests that are Integrated with Instruction Educational Measurement*. Macmillan, New York

Popham W J 1988 *Educational Evaluation*, 2nd. edn. Prentice Hall, Englewood Cliffs, NJ

Scriven M J 1967 The methodology of evaluation. In: Tyler R W, Gagné R M, Scriven M (eds.) *Perspective of Curriculum Evaluation*. AERA, Monograph Series Evaluation, No. 1. Rand McNally, Chicago

Suen H K, Logan C R, Neisworth J T, Bagnato S (in press) Parent-professional congruence: Is it necessary? *Journal of Early Intervention*

Worthen B R, Van Dusen L M 1991 Nature of evaluation. In: Husén T, Postlethwaite T N (eds.) *International Encyclopedia of Education*, 2nd edn. Pergamon Press, Oxford/New York, pp. 4345–56

F. Dochy

Educational Evaluation: Overview

In goal-directed human behavior, choosing between alternatives requires the assessment or the 'evaluation' respectively of the different possible behaviors. A rational choice can only be made when there is at least some factually based appraisal of each alternative's consequences. Thus an evaluation of former consequences of behavior ('experience') should always be the basis for rational decisions to keep the risks of unwanted consequences as low as possible. Trying to minimize the risks is also a question of ethics if other persons are affected by the consequences of an action as well. However, educational evaluation is realized in systems which are so complex that most evaluation attempts meet fundamental difficulties. Dealing with these problems requires not only scientific knowledge but also practical experience. This article can only give a general insight into some fundamental and methodological questions.

1. Fundamental Questions of Educational Evaluation

1.1 Definition of Evaluation

It is difficult to give a complete definition of an empirical construct. Rossi and Freeman (1993, p. 5), however, offer a formal definition: 'evaluation is the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of social intervention programs.' Put differently, evaluators 'use social research methodologies to judge and improve the ways in which human services policies and programs are conducted, from the earliest stages of defining and designing programs through their development and implementation.' (Rossi and Freeman 1993, p. 5) This is a definition close to science. The more one gets involved in the everyday business of evaluation, the less important become special methods of the social sciences. Often, simple questionnaires or interviews can be helpful. Generally speaking, evaluation is about the systematic appraisal of alternative options of behavior. Anything beyond that depends on the respective field of application and the emphasis chosen by evaluators working there.

1.2 Concepts of Evaluation

Due to the many forms of evaluation, different concepts of evaluation have been developed; for an overview see Scriven (1991).

The following forms of evaluation can be distinguished:

Microevaluation (the focus is on individual aspects of the evaluated program) vs. *macroevaluation* (global results are of interest);
Internal evaluation (those responsible for the scheme evaluate the program themselves) vs. *external evaluation* (development of the program and its evaluation are separate);
Summative evaluation (the review of the results takes place after a scheme/program has been completed) vs. *formative evaluation* (a measure and its effect are constantly checked in order to optimize it).

All these models (as well as others) can be used as general schemata to prepare evaluation studies. However, in each individual case it needs to be considered which aspects are especially relevant. 'Model' in this sense means aid for structuring and can offer a first orientation. In practice most evaluation studies (and most of them are focused on practical issues) cannot be completely described with just one model because of the complexities in real life. Therefore several dimensions and evaluation criteria are required.

1.3 Aspects and Problems of Practical Evaluation

Instead of an 'ideal' definition (which would be impossible) or an ideal classification scheme, the multifaceted aspects and problems of real evaluation attempts are outlined. The following questions are discussed: What is evaluated where and with what purpose? Which direct and indirect costs arise during the evaluation? With what are the results compared?

Exemplary answers (see Table 1) are given to these questions for the Third International Mathematics and Science Study (TIMSS), where the mathematical and scientific achievements of students from selected years in primary schools and secondary schools have been compared internationally.

1.3.1 What can be evaluated? As evaluation serves to optimize both actions and decisions, the consequences of both need to be compared with each other, i.e., evaluated. Concerning the alternatives in the educational field, the following are especially important:

Objectives: e.g., the consequences of different curriculums for a school subject or a vocational training (see *Curriculum Evaluation*);
Conduct of individuals: e.g., different didactical approaches in one seminar;
Techniques and procedures: e.g., comparison of different teaching methods;
Programs: e.g., the results of a new program to reduce unemployment in young people;
Systems: e.g., international comparison of school achievement (see *School Effectiveness Research*).

Different aspects are often intertwined. If, for example, techniques are to be evaluated then different individuals (e.g., teachers) use these techniques. This

Table 1
Basic evaluation questions for TIMSS

1. **What?:** International comparison of the efficiency of schools with regard to the achievements of students in mathematics and science
2. **Where?:** Schools; more or less randomly chosen schools from the 45 participating countries
3. **Objectives?:** International comparison of school systems and evaluation of the national school system with regard to the international results
4. **Cost/Benefit?:** No explicit statements, but large expenses are justified if the evaluation improves political decisions that affect many people
5. **Comparisons?:** Curriculums, school lessons, students' achievement

makes separating the effects problematic and may even raise the question of an interaction between the teachers' personalities and the teaching methods. It might get even more complicated when some of the aspects mentioned above vary systematically, e.g., when different educational programs have different objectives. Consequently, the results have to be interpreted keeping that in mind.

1.3.2 Where can evaluation take place? Generally speaking, all educational and educational-psychological measures can be evaluated. These evaluation studies can take place at very different locations. Each confronts the evaluator with very specific problems:

Laboratories: e.g., a model school can be introduced to test different forms of tuition. This makes the work for the evaluator easier as both teachers and parents are more motivated. Nevertheless, it is problematic how far the results from the laboratory situation can be transferred to the real life situation. This is a methodological problem as internal and external validity cannot be maximized at the same time (see Scriven 1991).
Educational Institutions: e.g., schools, universities. They are most often the place where educationally relevant evaluation takes place. Often the psychologists need to be able to communicate and cooperate well as they are dependent on the support of the other people involved.
Educational Systems: the evaluation of larger reform models or the comparison of whole educational systems like TIMSS (see Beaton et al. 1996 and Mullis et al. 1998 for a US account, Baumert et al. 1997 and Baumert et al. 2000 for a German account). Even more than with individual organizations one has to anticipate that the objective of the measures might encounter opposition from some of the people involved.

1.3.3 What objectives can an evaluation have? Other than in science, evaluators cannot choose the objectives of their evaluations in practice themselves. Evaluations in practice usually aim at improving existing 'systems' (see *School Improvement*). The evaluator has to consider the needs of those affected by the consequences of the evaluation. Additionally, the evaluator can only help the principals and those affected by the evaluation's consequences to clarify the objectives. The evaluator might draw attention to facts that have so far not been considered or point out contrasting interests of different subgroups. Nevertheless, evaluators need to remain neutral to all parties as well as all alternatives that are to be evaluated (for further information see e.g., American Psychological Association 1992).

1.3.4 What costs arise? Evaluations are expensive. Those evaluations that are carefully planned, carried out, and appraised especially require full commitment from all those involved. 'Costs' in addition to the cost of financing include:

Doubts concerning the present situation: The fact of an evaluation being carried out might be seen as criticism and depreciation of the present teaching efforts;
Unrest for all those involved: The examination of measures one has got accustomed to and the introduction of new measures can often disturb the implicitness that is important for the success of teaching;
Impairment of the personal situation: A control of the effectiveness among the employees of an institution can upset their self-esteem and reputation;
Additional workload for all those involved;
Delay in time: especially if a measure could be implemented immediately without the evaluation and without the risk of a negative outcome;
Possible damage for those affected by the implemented measure, if for example a new way of teaching languages does not prove to be as effective as the old method.

These costs and disadvantages of evaluation are only acceptable if they are outweighed by the benefit. Thus evaluators do not only need to be competent in their field but also in convincing others of their project. Still, it would be asking too much to assume that society needs to finance evaluations for every important new measure—even if the outcome of the cost–benefit analysis is positive.

1.3.5 Comparing the results. The basis for evaluations is always the comparison of different alternatives. If there are several independent alternatives then their consequences can be compared directly with each other. In this case a *summative* evaluation is possible.

If just one action alternative can be evaluated then the appraisal can have different criteria:

Expectation of the participants ('the students have been satisfied with my classes');
Former situation ('my intervention has been a success as the climate in the class has improved');
Personal objective ('I have reached XXX');
External objective ('the desired improvement in mathematics was achieved by the new teaching method');
Standard with group-reference ('with this learning method results were achieved that are better than the average of the last course').

In all this cases a functional appraisal is possible, especially with the objective of a *formative* evaluation. It lacks, however, the possibility of deciding that this option is better than another one (and which one). The alternative chosen might even have been the comparatively worst option under the circumstances. Without an alternative this cannot be decided, even if the outcome was positive (this class was better than last year's).

Which methodological approach is chosen (appraisal of one alternative or a comparison of groups as in experimental designs) depends on the objectives of a study. Causality can only be demonstrated when different groups are compared. If the cost–profit relationship is assessed it might be sufficient to focus on one alternative, even if the result is less meaningful.

2. Science's Contribution to Evaluation Studies

2.1 Reasons for Employing Scientifically Trained Evaluators

Evaluations of all kind are present in work situations, e.g., a school principal giving a judgment about a teacher. Evaluations like this are frequently done and many professionals believed (and sometimes still do) that the use of science for evaluations is superfluous. In contrast to this stands the rising figure of scientists and graduates actually working in this field. The reasons for this should lie primarily in the aspects of technical competence and objectivity.

Experts should have at least some knowledge in critical areas like basic research as well as in methodology (for planning the design and editing the results statistically). In addition to that they should have a higher ability of analyzing the often complicated connections found in most evaluation studies.

The image of objectivity plays an important role especially in critical evaluation studies, e.g., the evaluation of politically significant reform plans which only get financed if there is a massive dissension between the different political groups. Even though scientists are usually independent of political programs and ideas, they still have their own social convictions as well as preferences for certain theories. Consequently, politicians tend to assign those evaluators to the task who they assume to be on 'their' side. This may lead to the image of evaluators who can be bought. Different evaluation studies of the same project leading to different results can make the situation even more difficult (a famous example was the discussion about the effects of the comprehensive school, see Leschinsky 1999).

2.2 Planning and Designing an Evaluation Study

A significant advantage scientists have is their familiarity with the design of studies that allow a comparison of the relevant groups. For the planning and implementation of scientific evaluation studies, it is furthermore important that before the data is ascertained it is clear how it will be analyzed.

The design should try to exclude as many disruptive factors as possible. As studies differ so much there is nothing like an 'ideal' design. Other than in pure research it is not always helpful to stick to the original design throughout the whole evaluation. Whenever a program that is being evaluated shows serious shortcomings then it needs to be corrected in the interest of those concerned, even if it might go against scientific principles.

Another difficulty may arise with the sample. For being able to generalize the results, it needs to be representative. It is, however, also the psychologist's duty not to deceive participants about the purpose of the study (see e.g., American Psychological Association 1992). The truth might deter some people ('The new method is probably only ideal for the very weak children but we also need good pupils in the study') which then has serious statistical consequences.

2.3 Evaluation Reports

Mostly, evaluation results only present a basis for further discussion for those having to make a decision. Thus it is especially important how the decision-making groups learn about the results.

Usually, scientists are not used to informing laypersons (politicians, journalists, teachers, etc.): Only short summaries can be easily understood and scientific jargon (e.g., using the subjunctive and avoiding clear positions) is not accepted. The report has to be written in such a way that laypersons can comprehend the results and recommendations by the evaluators.

Even more problematic are oral reports, especially when controversial issues are concerned. The evaluator can easily be turned into the agent of one position. At the same time, it ought to be part of the professional ethics of an evaluator that although one attacks the scientific position of a colleague, personal integrity should remain untouched.

For guidelines on scientific evaluation reports see, for example, Fink (1995).

3. Present State of and Outlook for Evaluation Studies

In the last years there has been a renewed evaluation trend; this also holds true for the educational sector. In both schools and universities, new projects have been realized. In times of recession and tight budgets no one is prepared to pay for something of which the benefit is unknown. The public experts want to be informed about the efficiency of these costly institutions. In the universities, evaluations so far have been national. Some countries have an established ranking system as can be found in the USA (for an overview see House 1993) or the UK, others, like Germany, do not. For schools this has been the TIMSS study on the international level.

Realistically, evaluation studies in the educational sector cannot have the same degree of meaningfulness as pure research studies. However, this is not the intention of these studies either. Evaluation has served its purpose when it has provided discussions about a 'best' measure or the 'ideal' design with a rational basis. For this objective, there is no better alternative.

Luckily, the demand for evaluation studies is rising. The acceptance for evaluations is increasing in the whole educational sector. Nevertheless, they are limited by tight public budgets.

See also: Comparative Research in Education: IEA Studies; Curriculum Evaluation; Educational Assessment: Major Developments; Educational Policy, Mechanisms for the Development of; Program Evaluation; School Effectiveness Research; School Improvement; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values; Teaching: Trends in Research

Bibliography

American Psychological Association 1992 *Ethical Principles of Psychologists and Code of Conduct* **47**(12): 1597–611

Baumert J, Bos W, Lehmann R (eds.) 2000 *TIMSS/III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. Leske & Budrich, Opladen, Germany

Baumert J, Lehmann R, Lehrke M, Schmitz B, Clausen M, Hosenfeld I, Köller O, Neubrand J 1997 *TIMSS–Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: Deskriptive Befunde*. Leske & Budrich, Opladen, Germany

Beaton A E, Mullis I V, Martin M O, Gonzalez E J, Kelly D L, Smith T A 1996 *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study* (*TIMSS*). TIMSS International Study Center, Boston College, Chestnut Hill, MA

Fink A 1995 *Evaluation for Education and Psychology*. Sage, Thousand Oaks, CA

House E R 1993 *Professional Evaluation—Social Impact and Political Consequences*. Sage, Newbury Park, CA

Leschinsky A (ed.) 1999 *The Comprehensive School Experiment Revisited: Evidence from Western Europe*. Lang, Frankfurt am Main, Germany

Mullis I V, Martin M O, Beaton A E, Gonzalez E J, Kelly D L, Smith T A 1998 *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study* (*TIMSS*). TIMSS International Study Center, Boston College, Chestnut Hill, MA

Rossi P H, Freeman H E 1993 *Evaluation—A Systematic Approach*. Sage, Newbury Park, CA

Schmidt W H, McKnight C C, Cogan L S, Jakerth P M, Houang R T 1999 *Facing the Consequences: Using TIMSS for a Closer Look at US Mathematics and Science Education*. Kluwer Academic, Boston, MA

Scriven M 1991 *Evaluation Thesaurus*. Sage, Newbury Park, CA

H. Wottawa and D. Pult

Educational Innovation, Management of

Managing innovation in education (especially in schools) is a complex task for the principal, and also the staff, and often includes collaboration with outside partners. Innovations demand learning and are seldom neutral in the eyes of stakeholders, therefore they imply the understanding and handling of intertwined technical, political, and cultural perspectives.

1. Management

Management and leadership are not exactly the same thing, but many have noticed that for school leaders they tend to overlap. Leaders engage in mobilization, vision building, problem solving, and learning, whereas managers do the goal setting, facilitation, coordination, monitoring, and rewarding business. Both roles are clearly needed to facilitate innovation and change. There is a growing interest in that teachers now engage more in leadership roles among their peers.

2. Innovation

Innovations are programs, material, schemas, structures, etc., that promise to be new and better than the old ones. They denote the technical content of improvement. However, the generation, diffusion, and use of innovations are also political and entrepreneurial processes governed by their special rationalities (Fullan and Stiegelbauer 1991). Research had to learn by its own mistakes that there are technical, political, and cultural perspectives on innovations. Neglect of one or two of these still causes school reforms to fail (House and McQuillan 1998).

2.1 Technical Implementation of Innovations

Innovation as a technical concept originally belonged to the research, development and dissemination strategies of change (RDD) and a related approach advocating 'planned change,' both especially strong from the 1950s. Belief in expert-based, top-down change was founded on scientific positivism and models of rational decision making for implementation. It sometimes simplified the technical problem to ideas of 'teacher-proof' innovations. More sophisticated work within this tradition was synthesized in early work by Havelock (1969) in a system analysis-based view of the linkages and linking roles between providers and users of innovation. Out of this research about change, agent characteristics and training evolved. Increasing interest in the social psychology of organizations culminated in the organization devel-

opment movement of the 1950s–1970s. RDD strategies also incorporated T-groups and other group dynamic devices in order to change the climate of organizations, making staff more receptive to innovations. Teams or departments can still be sent to 'temporary systems' (Miles 1964), resorts where the regular system's norms and hierarchies cannot easily dominate and thereby allow the training of more creative relations. The problem of how learning from temporary systems can survive in the regular system always remains.

2.2 Political and Cultural Competition

The negotiation of interests and the legitimacy of authority are the focus of the political perspective. Strategies used by individuals or groups to serve their interests within an organization are called micropolitics. A school is fairly often informally run by a dominant coalition of teachers and the principal may be lucky to belong to it. Micropolitics is almost inevitably a part of management as a principal strives to achieve control of and commitment from the staff. When it is done with a hidden agenda, including, e.g., cooptation, and control of information flows, it is even more micropolitical. Micropolitic tensions often imply a gender aspect.

The cultural perspective rests on an image of community where values and shared meanings hopefully merge into cultural integrity. Contrary to issues within the political perspective, cultural issues are hardly negotiable, as their meaning may have deep roots in ideology or group identity. This is one of the reasons why control of the recruitment of organization members greatly facilitates common action.

The political definitions of outcomes from innovations are stressed by Cuban (1998), who isolates five important criteria by which the success or failure of them are judged in the USA. Three are favored by the policy elite—effectiveness, fidelity, and popularity, and two by practitioners—adaptiveness and longevity. The effectiveness and fidelity criteria are connected to a bureaucratic concept of top-down authority guarding control and technical rationality, which is preferable for elites.

There are professional and popular movements working from a bottom-up perspective with agendas of their own, reflecting differing views about, e.g., learning and teaching. Movements engage in contests over professional culture. Often they are recruited as partners to elite initiatives.

Simola (1998) named a certain blend of utopian and rationalistic ideas 'wishful rationalism.' It gives educational reformers 'self-created problems,' a most important component for group identity to build around. It can be added that wishful rationalism probably has one of its most fruitful seeding grounds in the literature of management and leadership.

3. Solutions and Problems of Mixed Character

Teacher efficacy illustrates nicely the need to use several perspectives in searching for solutions to problems of innovation and change. Teacher work is basically characterized by technical uncertainty, which makes high self-efficacy hard to achieve. Rosenholtz (1989) found that receiving positive feedback on performance and collaborating with other teachers were among the most important sources of teacher efficacy. This is in line with general theory on self-efficacy, which predicts that mastery experiences is the most important determinant, but that also social persuasion (pep talk and general feedback), and vicarious experiences (e.g., modeling by peers) are important. It is suggested that the implementation process should contain early and strong support because of its vital importance for teachers' feelings of practice mastery and which also increases user commitment.

Networking between teachers from different schools (but with the same interest in subjects, children of a certain age, etc.) is seen as a powerful instrument for professional development and change. It frees participants from limitations in institutional roles and geographic isolation, and builds a kind of temporary system. Networks are organizationally flexible. They support 'just-in-time learning,' learning that is tied to the actual work teachers do. Networks stimulate teachers to take leadership roles among peers. Within them, more egalitarian and sustained interaction between researchers and practitioners may be built (Lieberman and Grolnick 1996).

There have, however, been warnings against networking as just fostering a discussion culture. Some think that the tendency to talk together instead of working and learning together is already too strong within the teaching profession.

Great efforts are made to disseminate complex concepts of constructivistic instruction in science and mathematics to teachers using a 'systemic reform' design. Measures at several levels will conjoin: teacher education, curriculae, testing, accountability systems, information, support, etc. Strong coalition building at regional and local levels is needed to achieve this and still the result is highly dependent on the capacity of single schools. High-capacity schools have cohesive staffs developing 'collegial learning communities' and using professional networks to scan their environment for innovations and models. Low-capacity schools lack a critical mass of teachers willing to or used to work like this (Knapp 1997).

A 'learning community' in the sense just mentioned is about teachers learning together at the same school, and another term for that might be a 'learning organization.' However, Leithwood and Seashore Louis (1996, Chap. 13) point to different conceptions inherent in those terms. Organizational learning implies deep change of norms and repertoires, a process

which can be highly stressful for participants and which is by no means linear or harmonious. They find that this transformational change must be balanced by evolutionary changes. 'Breakthroughs' have to be complemented by 'small wins.' The small adaptions of everyday life help to shape continuity and stability in relations and routines, and thus the possibility to build a professional community. So schools need both, and the challenge is to get them into a positive dynamic. There is some evidence that community building preferably foregoes deep organizational learning. Leadership that is simultaneously authoritative and supportive is strongly associated with professional community. So is trust among staff and between staff and leaders, and it also fosters risk taking, which is good for learning.

Louis and Miles (1990) found that successful urban schools struggling to change in hard environments used evolutionary planning; i.e., iterative processes of learning through experience. This must be clearly distinguished from the strategy of 'muddling through,' which is to take advantage of any hopeful events that might occur under severe constraints. Evolutionary planning demanded strong coordination by cross-role groups or steering groups which must be legitimate in taking control over the process. Legitimacy came both from the principal and from power sharing. Vision building was important, but contrary to many assumptions it did not depend especially much either on formal data collection or on the leaders' charisma. It was rather an interpretative act making the staff realize a common task and a common fate. 'Deep coping' involved measures such as rolling planning, substantial restaffing, attempts at increasing school control over the environment, etc. Actions aimed at building system capacity, such as coordinators and task forces, were more effective than actions aimed at personal capacity building.

However, for ambitious change within schools, another, competing concept is also recommended, i.e., 'strategic planning,' which is a revitalized planned change approach. In the version of Kaufman and Herman (1991), it is a bold adventure in which the school interacts with its constituencies in making missions for the future and deriving long-term goals and strategies from them. Involving stakeholders means that the approach uses its political in addition to its technical potentialities. The whole idea of strategic planning has, however, been criticized for being contraproductive in using too formal and technocratic planning methods.

4. *Models from Industry and Markets*

The 1980s–90s saw a great inflow of innovative ways of organizing schools coming from the business sector. This marks a change of control over schooling as both professional and political judgements are circumscribed by administrative and economic models. Total quality management (TQM) is such a broad concept, geared to market mechanisms. It has one line of thought focusing on the tight steering of processes ('doing it right from the beginning'), and another focusing on the capacity to control the environment. The latter may mean that schools are recommended to influence their input by putting up their own preschools and parent training programs, and by controlling publishers of school books. It should also engage in consumer research, as 'in enterprises that depend on public support for their existence, perception is truth' (Bradley 1993, p. 4).

See also: Curriculum as a Field of Educational Study; Educational Institutions and Society; Educational Institutions, History of; Educational Leadership; Educational Planning Worldwide; Educational Policy, Mechanisms for the Development of; Educational Research and School Reform; Innovation: Organizational; Schools, Micropolitics of

Bibliography

Bradley L H 1993 *Total Quality Management for Schools*. Technomic, Lancaster, PA

Cuban L 1998 How schools change reforms: redefining reform success and failure. *Teacher's College Record* **99**: 453–77

Fullan M, Stiegelbauer S 1991 *The New Meaning of Educational Change*. Teacher's College Press, New York

Havelock R 1969 *Planning for Innovation Through the Dissemination and Utilization of Knowledge*. CRUSK, Institute for Social Research, University of Michigan, Ann Arbor, MI

House E R, McQuillan P J 1998 Three perspectives on school reform. In: Hargreaves A et al. (eds.) *International Handbook of Educational Change*. Kluwer, Dordrecht, The Netherlands, pp. 198–213

Kaufman R, Herman J 1991 *Strategic Planning in Education: Rethinking, Restructuring, Revitalizing*. Technomic, Lancaster, PA

Knapp M S 1997 Between systemic reforms and the mathematics and science classroom: the dynamics of innovation, implementation, and professional learning. *Review of Educational Research* **67**: 227–66

Leithwood K, Seashore Louis K (eds.) 1996 *Organizational Learning in Schools*. Swets and Zeitlinger, Lisse, The Netherlands

Lieberman A, Grolnick G 1996 Networks and reform in American education. *Teacher's College Record* **98**: 7–45

Louis K S, Miles M B 1990 *Improving the Urban High School*. Teacher's College Press, New York

Miles M B 1964 On temporary systems. In: Miles M B (ed.) *Innovations in Education*. Teacher's College Press, New York

Rosenholtz S J 1989 *Teachers' Workplace. The Social Organizations of Schools*. Teacher's College Press, New York

Simola H 1998 Firmly bolted into the air: wishful rationalism as a discursive basis for educational reforms. *Teacher's College Record* **99**: 731–57

R. I. Lander

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Educational Institutions and Society

The analysis of relationships between educational institutions and society presupposes at least minimal conceptions of institutions and societies as social systems

1. Educational Institutions

1.1 School Systems as Institutions

There is widespread agreement in sociological theory that institutions are social creations that serve to solve basic problems of society—e.g., to produce essential goods (economic institutions), to regulate sexual needs and biological reproduction (marriage and family), to resolve conflict (legal and political institutions), etc. Educational systems can also be regarded as large-scale problem-solving institutions. Confronting human genetic plasticity and the ability to learn to meet the complex needs of social life reveals the necessity of cultural reproduction via educational institutions. Educational systems form the highly institutionalized and organized part of the total educational service of a society for the new generation.

Given this background, the question arises as to which societal problems an educational system responds in a highly complex society and what it contributes as an institution for society at large. The answer is that above all educational systems solve cultural reproduction problems, i.e., problems of reproducing the culture and social structure of a society in the light of the biological exchange of its members. Each new generation, for instance, must learn to read and write if the society is not to decline into a more primitive state. A certain technology can only be maintained if we also have future technicians and engineers who, for instance, can build and repair television sets, cars, and computers.

To solve these problems of reproduction, school systems have been created during long historic processes of intentional 'institution building.'

1.2 School Systems as Places for Socially Controlled and Organized Learning

Content and ways of learning are structured and arranged in educational institutions on behalf of outside authorities. They express the power of social groups interested in specific content and processes of education.The socially desired learning processes are enforced in school systems by legal procedures and sanctions, by providing learning material and personal resources, and by systematically developing teaching know-how via elaborated training procedures.

1.3 School Systems as Socialization Agency

The third and central element of our concept of schools characterizes them as socialization agencies. Any process is labeled socialization if it at the same time constitutes psychological structures and reproduces cultural and social structures. School systems socialize children and adolescents via training qualifications and teaching values, without which the individual would be unable to function in society. At the same time, social reproduction and societal functioning is perpetuated. School systems are thus subsidiary arrangements; they emerge when mere participation in the social life of the family and kin group no longer suffices to teach all that is required for the functioning of society.

It is evident that reproduction is the main function of school systems, but school systems are also instruments of social change, proposed by social groups in charge of educational government.

2. The Structural-functional Approach: Society and Institutionalized Education

In the 1960s, two sociological theories became prominent as frameworks for the formulation of the relationships between educational institutions and society: first the structural functionalism of Parsons, and secondly the framework of English social anthropologists (Halsey).

Here for the first time the basic idea is formulated rigorously:

> (a) All societies maintain themselves with the help of a culture, i.e., with a complex set of attitudes and skills which are not contained in the genetic constitution of an individual but must be learned. This social heritage must be transmitted by social organization. Education has this function of cultural transmission in all societies.
> (b) Individuals' personalities must be formed so that they fit into a culture. Everywhere, education has the function of forming social personalities. As culture is transmitted by influencing personality, education contributes to integration of society whereby society can be seen as a mechanism which enables humanity to adapt to its environment, to survive, and to reproduce. (Halsey 1973 p. 385)

In his famous essay 'The school class as a social system,' Parsons (1959) explicated two functions of modern educational systems, namely socialization and selection.With socialization Parsons conceptualizes the process of passing skills on to each new generation, thus allowing them to fulfill tasks in an occupation and as a citizen. The function of selection leads to an allocation of students to occupational positions matching their abilities and interests. The structure of the school class is ideally suitable for fulfilling these tasks by modeling the processes of qualification and allocation. Dreeben (1968) continued these lines of

reasoning by contrasting the social structure of the family and the school and by explicating the corresponding norms and values. Thereby the family represents a particularistic institution where each member occupies a unique status. In contrast, the school is seen as a universalistic social system where students are treated as equals and difference is attributed by universalistic norms of achievement.

An influential analysis of the relationship between school system and society originates from the French language area. The sociologists Bourdieu and Passeron (1977) became the first since the important work of Durkheim (1956), to provide internationally recognized contributions clarifying relationships between the educational system and social class (Bourdieu and Passeron 1977). Using the example of the higher educational system in France, they postulated the split between the official ideology of achievement based on upward mobility and the actual rules for social reproduction. In reality, students with the correct cultural 'habitus,' belonging to the family culture in the bourgeoisie and the upper class, have greater opportunities. These strata reproduce their privileged positions via the school system. The ideology of achievement as the basis for social mobility serves to veil the real process. Thus the school system does not pass on a society's cultural heritage. Instead the culture of the ruling class makes primary use of it to maintain its privileged position.

3. *The Dual Function of Educational Systems in Modern Societies*

Beyond a more affirmative or critical view on the social functions of educational systems in modern societies, the schooling process can be analyzed from two perspectives: one with the view on reproducing and changing society, the other on constituting individual qualifications and orientations.

To see the social functions of 'human processing' via schools is indispensable for an adequate understanding of schooling. In a historical perspective, this becomes especially evident. From an individual perspective, school represents the most important cultural context for learning and the development of competencies and values. At the beginning of the twenty-first century, the relationships between educational institutions and society can be characterized as follows:

(a) Reproduction of cultural systems, which are often characterized as knowledge and skills, is institutionalized within the school system. This involves mastering basic symbolic systems ranging from language and writing to acquiring specific professional qualifications. Accomplished by instruction, this function is referred to hereafter as the *qualifying function*.

(b) The second function refers directly to the society's social structure. Social structure is defined as the distribution of positions in the occupational and social area. By way of school-based learning of qualifications, verified by a detailed system of examinations, the young generation is allocated to stratified occupational positions (*allocation and selection*).

(c) School systems are instruments of social integration. Reproduction of norms, values, and belief systems serve to reinforce prevailing power relationships (see Bernstein 1996, for contemporary theory on the relationship between knowledge and control). This accounts for the central political function of the socialization process in educational institutions (*integration and legitimization*).

In Fig. 1, the three reproduction functions of educational systems in modern societies are summarized. They appear as contributions to a defined

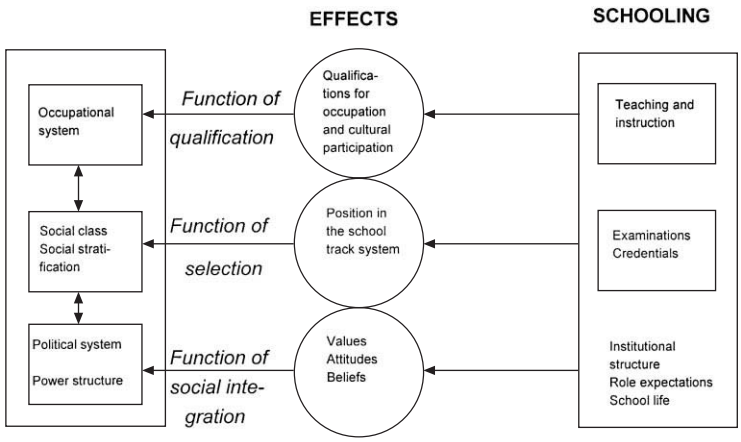

Figure 1
Social functions of schooling

social system which are necessary for its survival. To those involved, functions appear as tasks and goals. Therefore, reproduction can be converted into social change via educational institutions.

The school's societal references are shown on the left in Fig. 1. Systemic problems may arise which can only be solved adequately by educational institutions. This may be the case with regard to necessary qualification patterns for the occupational system, the legitimate structure of social mobility and social reproduction, and the legitimacy of the societal power structure or the political system.

The educational mechanisms through which societal problems are solved are shown on the right in Fig. 1. Instructional processes, examinations, and the institutional structure are differently involved in the process of fulfilling societal functions.

4. *The Historic and Action-theoretical Approach: The Evolution of Relationships between the Educational System and Society*

From the background of the functional analysis above, the economic, social, and political significance of educational systems is clearly visible. At the same time it becomes evident that the functional relationships between educational institutions and society are at work only in modern societies. Because educational systems do not follow a path of social determinism, their historic evolution can be traced back to observable processes of their creation. They are the result of human agency and social interaction and thus distinguish themselves fundamentally from natural evolution, as do all other areas of social reality (Archer 1995).

Educational systems in the European culture have emerged in the context of religious movements and interests. Until the eighteenth century it was the church, particularly, which provided personnel for schools, maintained buildings, selected curricula, set the definition for instruction and access to education, and shaped opportunities after graduation. In comparison to the multifunctional relationships between educational institutions and societies in modern times, this represented a case of mono-integration. The primary goal of educational institutions consisted in the cultural reproduction of religious knowledge and thereby maintained a monopoly in the Christian way of understanding human existence. As a book religion, Christianity depended on the training of religious experts. During the Reformation, providing direct access to the Bible became an essential part of the Christian way toward salvation. Therefore everybody should be able to read. This belief became an important impetus for teaching reading and writing to everybody. The practice of teaching was conceived along the lines of preaching and roles of teachers ran parallel to those of priests. Thus they determined the definition of knowledge in terms of pedagogical practice and the types and quality of educational output which seemed necessary.

Naturally, the typology in Table 1 fails to represent the total variety of historic phenomena covering educational institutions. There existed important national variations (see Archer 1979 for France, England, Russia, and Denmark) and particularly variations by religious faith. During the sixteenth century the Reformation was crucial for the emergence of alphabetization and literacy. It was also an important basis for the rise of national educational systems at the end of the eighteenth century.

In order to understand the emergence of national educational systems in European countries after the French Revolution, three historic predecessors have to be kept in mind: (a) it could build on advanced literacy

Table 1
Typology of relations between educational institutions and society (mono – integration vs. multiple integration)

	Mono-integration	Multiple integration
Functions	Maintenance of religion, of a monopoly in interpreting the meaning of human existence	Relation to political systems Relation to economic systems Relation to the social structure
Resources	Provided by the church	Provided by the state
Personnel	Integrated role of priest and teacher	Professional staff with defined autonomy
Content	Religious content selected and controlled by the church	Pluralistic content, as a result of democratic bargaining
Control	Ownership	Democratic and bureaucratic control
Hierarchy	Subordination and domination	Bargaining power of corporate actors
Access	Elective access	Universal access by ability and interest
Promotion rules	Conformity to religious convictions Suitability for expert positions	Achievement in defined subject matters
Careers	Positions in the hierarchy of the church	All divisions of the labor market: Dependent on qualifications

(for Switzerland, see Messerli 1999), (b) there existed scholarly communities at universities since the middle ages, communicating in one common language (Latin), (c) private schools served the need for commercial and industrial qualifications in the cities. Accordingly, there had been scattered possibilities of education in the cities since antiquity. The universities represent a complex history in which the conflict between reason and faith arose early as a heritage of antiquity. Development of scientific thinking has been an important motivation for establishing public education since the Renaissance and the rise of humanism.

The rise of modern education became an essential part of the emergence of state structures from the nineteenth century onward. This process turned out to be a historic struggle for power in which identifiable actors and interest groups proceeded within the framework of historic and structural conditions. Their activities during a process of structural elaboration led to educational institutions, which only gradually assumed the characteristics of a coherent system in the course of the nineteenth and twentieth centuries. Speaking of an educational system means that pathways through the school system were systematically linked to each other, that a professionally trained staff emerged, and that coherent curricula were developed. During this process, the 'meritocratic' principle was established as the only legitimate way of promotion.

The three main social functions described above developed along this historic route. At the individual level they implied that long-term learning became a universal instrument of individual life planning. A parallel process could be observed at the social level. It implied the increasing inclusion of all social classes in the educational system. During the nineteenth century the struggle for social participation in education took place at the primary school level, in the twentieth century at the secondary and tertiary level. Thus discussion about the role of education reflects the political discourse of society. Yet it is embedded within the political rules of the society and the real power structure, which becomes visible in historic and comparative perspectives (see Heidenheimer 1997).

5. *Empirical Analysis of the Relationship between the Educational System and Society*

Historic analysis of the relationship between the educational system and society is not limited to the topic of how the educational system evolved parallel to and as part of the developing welfare state during the last two centuries.

5.1 *Educational Expansion*

Beyond system development, the factual educational participation at various educational levels can be described. Doing this, it becomes apparent that all parts of the population had become literate in most European nations toward the end of the nineteenth century. Expansion of secondary education and particularly tertiary schooling occurred cyclically, yet it was only during the 1950s that it began in large numbers. The sharp increase in educational participation from the 1950s to the 1970s began to level off from the 1980s.

Opinions vary greatly in the discussion of causes for this educational expansion. Doubt exists primarily about the thesis that it resulted from conscious planning and pressure by political decision makers. Actually it followed a complex interrelationship of supply and demand, of system-building processes on the supply side and accumulating consumer demand. Pressure from parents and students emerged in the twentieth century through perspectives on qualification-related life planning. These focused on the perception of opportunity structures influenced by economic developments in an era of rapid technological change and economic globalization. On the political level, a highly productive educational system turned out to be a key factor for successful global competition.

5.2 *The Educational System and Social Structure*

The nineteenth and twentieth centuries reflect the struggle over participation in education by privileged and disadvantaged social groups. Efforts were exerted to achieve a classless society by education, by implementing the principle of meritocracy, and by forcing the nobility to take examinations. Educational expansion brought about egalitarian effects and strengthened the role of the equal rights of all citizens (Müller 1998). Moreover, parallel to expansion the twentieth century saw an observable increase in educational opportunities for the farming sector as well as for workers and employees. Yet relative inequality of opportunity (Fend 1982) still existed in all industrialized nations at the end of the twentieth century.

The greatest progress in educational participation during the second half of the twentieth century could be observed among girls and women. They profited most from educational expansion (Blossfeld and Shavit 1993, Dahrendorf 1965, Müller and Karle 1993, Schimpl-Neimanns and Lüttinger 1993).

6. *Practical Questions*

Linking a theory of educational institutions to historical processes of their creation opens up the perspective that humans are responsible for shaping educational institutions for the demands of future developments. Questions of quality in the educational systems and their effectiveness in the light of limited resources have top priority in the political agenda. At

the opening of the twenty-first century one central issue deals with the problem of how to shape the qualifying function of educational institutions within the framework of international economic competition. Answers to questions about equal opportunity and the principle of meritocracy will shape the acceptance of education in the coming years. Questions concerning democratic control of knowledge and values will be of great significance for the peaceful coexistence of people in a global civil society.

See also: Education and Employment; Educational Institutions, History of; Educational Policy: Comparative Perspective; Educational Policy, Mechanisms for the Development of; Educational Sociology; Educational Systems: Europe; Institutions

Bibliography

Apple M W 1999 *Power, Meaning, and Identity*. Peter Lang, New York

Archer M S 1979 *Social Origins of Educational Systems*. Sage, London

Archer M S 1995 *Realist Social Theory: The Morphogetic Approach*. Cambridge University Press, Cambridge

Bernstein B B 1996 *Pedagogy, Symbolic Control and Identity. Theory, Research, Critique*. Taylor and Francis, London

Blossfeld H-P, Shavit Y 1993 Dauerhafte Ungleichheiten. Zur Veränderung des Einflusses der sozialen Herkunft auf die Bildungschancen in dreizehn industrialisierten Ländern. *Zeitschrift für Pädagogik* **39**(1): 25–52

Bourdieu P, Passeron J-C 1977 *Reproduction in Education, Society, and Culture*. Sage, London

Dahrendorf R 1965 *Bildung ist Bürgerrecht. Plädoyer für eine aktive Bildungspolitik*. Nannen-Verlag, Hamburg, Germany

Dreeben R 1968 *On What is Learned in School*. Addison-Wesley, Reading, MA

Durkheim E 1956 *Sociology and Education*. Free Press, Glencoe, IL

Fend H 1982 *Gesamtschule im Vergleich*. Beltz, Weinheim, Germany

Halsey A H 1973 The sociology of education. In: Smelser N J (ed.) *Sociology. An Introduction*. New York, Wiley, pp. 381–434

Heidenheimer A J 1997 *Disparate Ladders. Why School and University Policies Differ in Germany, Japan and Switzerland*. Transaction Publishers, New Brunswick, NJ

Messerli A 1999 Lesen und Schreiben 1700 bis 1900 Unpublished postdoctoral dissertation, University of Zurich, Switzerland

Müller W 1998 Erwartete und unerwartete Folgen der Bildungsexpansion. In: Friedrichs J, Lepsius M R, Mayer K U (eds.) *Die Diagnosefähigkeit der Soziologie*. Westdeutscher Verlag, Opladen, Germany, pp. 81–112

Müller W, Karle W 1993 Social selection in educational systems in Europe. *European Sociological Review* **9**(1): 1–23

Parsons T 1959 The school class as a social system: Some of its functions in the American society. *Harvard Educational Review* **29**(4): 297–318

Schimpl-Neimanns B, Lüttinger P 1993 Die Entwicklung bildungsspezifischer Ungleichheit: Bildungsforschung mit Daten der amtlichen Statistik. *Zuma Nachrichten* **32**: 76–115

H. Fend

Educational Institutions, History of

1. Origins

Any attempt to specify a starting point for the history of educational institutions is necessarily arbitrary. This is due to the uneven and contingent distribution of early historical sources across time and space, as well as to the fuzziness of the term 'educational institutions.' Broadly speaking, the term refers to a set of rules that govern the aims and activities of the adult generation in socializing the young generation into more or less age-graded roles. Knowledge about how to profit from the extended period of human infancy in order to further physical and moral/cognitive development must have been part of the everyday wisdom of the older generation long before precepts concerning education appeared in religious texts. And even after the introduction of educational institutions in a more narrow sense of the term (schools), these rule-sets governing everyday socializing/educational practice remained of prime importance since these formal institutions were highly specialized and, for centuries to come, reached only a small minority of a population.

The boundaries between religious or general cultural practice and educational practice and the respective institutions in these fields are not clearly marked. Thus, for ancient Jewish education the basic set of texts is identical with the literature guiding religious practice: the Pentateuch, the oldest part of the Old Testament, i.e., the books Moses 1–5. Of course, those parts of the text referring directly to education are clearly recognizable as such but, obviously, their interpretation (e.g., in the Talmud) depends on the religious context and vice versa (Schwenck 1996, pp. 153–77).

For example, we are most familiar with an educational practice that is not primarily channeled through educational institutions in the narrow sense of the term but rather builds on participation in cultural, social, and political activities in a highly dynamic society, namely that prevailing in ancient Greece during the classical period (for the following, see Marrou 1948, Schwenck 1996). This does not hold, of course, for Sparta, where the male child left the family at the age of 7 years to join a state-controlled educational system for the next 12 years bent on producing fearless and obedient warriors.

Athens at that time had not developed a public system of schools; military training of 2 years' duration was introduced only in the late fourth century BC. Education of the young (boys) was the responsibility of the family, which availed itself of the services of private teachers. Boys of fathers having full citizenship had the right, at age 10 years, to be present at most cultural events. Although the works of the poets and historians were available in writing and many of the (male) citizens must have been literate to be able to

fulfill their public duties, Greek culture of the classical period was to a large degree an oral culture. Festivals, the theater, and competitions in presentations of dance, music, poetry, drama, and athletics provided ample opportunity to familiarize the young generation with the rich heritage of Greek civilization. This experience was understood to be more than a means to achieve esthetic refinement; it exposed the young to reflections on human destiny and justice, concepts of virtue, models of conduct and of good governance. It was preparation for citizenship and moral education. Physical education and music education, too, were considered fundamental for the formation of the free citizen. Especially the former was seen as part of the preparation for military service, and the communes were responsible for providing and controlling localities for athletic activities. The growing prestige of the Panhellenic games may, however, have served to weaken the nexus with military training.

Many elements of the (nonformal) curriculum of the classical period reappeared in the curriculum of the public schools provided by the larger communes during the Hellenistic era, normally dated from 338 BC when Philipp II of Macedonia defeated Athens and their confederates in the battle of Chaironeia. In both periods the curriculum is strictly oriented towards general as opposed to vocational education. The adult role emulated in the schools is not that of the skilled professional but that of the citizen potentially responsible for different general tasks in and for his community and in need of general moral as well as cognitive competences.

Educational institutions of a more restricted vocational orientation had already been introduced long before the apex of Greek civilization in the Sumerian city states of the third millenium BC. These were highly differentiated societies facing complex administrative challenges and had important trading relations. To document and control their diverse activities they needed systems of book-keeping for the different administrative fields and functions. Considering the state of development of the sign system for writing, mastering this task required systematic learning. Consequently, writing schools for trainees of the administration were established (Feldbusch 1988, pp. 1473ff). Comparable institutions specializing in the education of scribes, administrators, and probably also artisans had already existed in ancient Egypt (Craig 1981, Adick 1992, pp. 164ff).

The Hellenistic system of postelementary education which prospered not only in the Greek communes but also in many regions bordering on the Aegean and Mediterranean Seas, and which functioned in the eastern part of the empire until the fall of Byzantium in 1453, is important for further developments during the Middle Ages in Europe in several respects: it provides the context in which early Christian concepts of education are developed in close contact with classical pre-Christian educational theory. The curriculum of the Enkyklios Paideia/Septem Artes liberales remains a point of reference in the debate on the contents and objectives of postelementary education, even gaining in importance with the growth of first-hand knowledge of the classical texts in central and western Europe.

2. *The Middle Ages*

The cathedral schools first founded towards the end of the eighth century and spreading throughout the Carolingian Renaissance provide evidence of this type of long-term curricular continuity (Verger 1986). Although their primary purpose was the education of the clergy via an in-depth study of the Holy Scripture, they also introduced their students to a version of the trivium grammar, rhetoric and—around the end of the tenth century—dialectic. Thus, these schools offered a relatively broad general education, including the study of classical nonreligious texts, as a basis for the professional preparation for the clergy. Some of these schools developed into universities in the first half of the thirteenth century, the largest and, with respect to the quality of its liberal arts and theological faculty, perhaps the most important among them the University of Paris, which attracted large numbers of students from England, the German-speaking regions of Europe and Italy. Among its famous teachers were Albertus Magnus (1240–48), Bonaventura (1248–57) and Thomas of Aquino (1252–58 and 1268–72) (Verger 1986, p. 421). Their names, like Duns Scotus, Roger Bacon and Occam, also point to the importance of the Christian orders, foremost among them Dominicans and Franciscans, as centers of the rich intellectual life and scholarly learning of the twelfth and thirteenth centuries. The attractiveness of the University of Paris was comparable to that of law schools in northern Italy (e.g., Bologna and Pavia), which served as models for the rest of Europe. The hierarchical relation between professional schools of theology, law (Orléans) or medicine (Montpelier) at the top and the ancillary and preparatory functions of the liberal arts faculty, typical of the university up to the end of the eighteenth century, is already discernible in some of the French cathedral schools of the eleventh and twelfth centuries. This may have helped to preserve the idea that higher education should be more than just a rather restricted professional training. When, by the late Middle Ages, much of the knowledge acquired in classical antiquity and preserved in the centers of learning of the eastern part of the Roman Empire had been recovered, this offered the opportunity to widen the scope of scholarly learning to approximate the full range of the classical liberal arts curriculum by including parts of the quadrivium: arithmetic, geometry, music, and astronomy. To be sure, this was already the vision of Alcuin—teacher and close counselor of Charlemagne, learned author of textbooks on the seven liberal arts and very active in

the propagation of cathedral schools in the early ninth century. In most of the French cathedral schools, however, teaching of the liberal arts seems to have been restricted to the trivium.

When the first German universities were founded in the second half of the fourteenth century (beginning with Prague in 1348), they followed the organizational pattern developed in the older European universities: the liberal arts faculty was by far the largest but ranked lowest. The faculties of theology and law enjoyed the highest prestige and the best remuneration of their teachers. In most of these universities, however, the arts faculties covered the full range of the arts, including natural science on the basis of Aristotle's writings (Paulsen 1919, Vol. I, Rüegg 1993/1996, Vol. 1).

The larger and more successful European universities were international in the composition of their student bodies and teaching faculties, which, of course, was rendered possible by the use of Latin as the lingua franca of higher learning. Students were attracted by famous teachers and these, too, traveled much in search of better conditions for their work. Thus, knowledge spread fairly quickly, but was accessible only to that small minority who mastered the medium of communication. Schools that could have transmitted the necessary skills were small and widely dispersed. Demand for this type of instruction was low for a host of reasons. Until the end of the eighteenth century there was no clear demarcation between the respective functions and objectives of schools and universities.

By the end of the fourteenth century, most European towns of some importance had one or more schools. They were attached either to a parish, a convent, or a chapter, or directly sponsored and controlled by the communes. The latter form, with the supervision of the town's council, seems to have prevailed in Germany. But here, too, the relation between school and church remained a tight one. In many cases the minister of the church was himself the teacher or he may have acted as rector of the school, responsible for the selection of a teacher who, in turn, chose his own teaching assistants. The pupils were taught in two or more groups formed on the basis of age and learning progress. The main objective of the town school was to teach the pupils to read and to write Latin. It seems remarkable that these schools took in not only the children of well-to-do families but sometimes also a sizable proportion of poor children living on charity (Paulsen 1919, Vol. I, 18ff).

3. *The Humanistic Movement and the Reformation*

Around the turn of the sixteenth century, universities and schools were shaken by two interdependent events: the humanistic movement, originating in the Italian Renaissance, and the Reformation (Rüegg 1993/1996, Vol. 2). The humanistic movement was a polemic revolt against the culture and learning of the Middle Ages, which it condemned as a dark, narrow, uninspired, mechanistic, unnatural and bookish educational tyranny; and it was a movement for a fresh reading of the poets and philosophers of classical antiquity in their original language—Greek, Latin, Hebrew—a re-reading freed from the pressure to prove this literature's compatibility with Christian values and dogma. It was, last but not least, a plea for a study of (human) nature, unfettered by the authority of classical literature (for a contrasting evaluation, compare Paulsen 1919, Vol. I, Chaps. 3–5 and Garin 1964, pp. 25ff).

The movement's immediate effect on the schools seems to have been rather limited, depending on the conditions encountered locally: the prestige of the classical languages was raised, new textbooks written or recommended by devotees of the movement were introduced, some schools hired a poet to teach the basics of Latin poetry and eloquentia. In a small number of schools, Greek was introduced as a new subject. The long-term effects, on the other hand, were important to both schools and universities: The curriculum of the liberal arts faculty was considerably extended to include Greek language and literature which was, like Latin literature, to be represented in a fairly wide selection of poetry philosophical, historical, mathematical, and other scientific texts. Most arts faculties also offered courses in Hebrew. This increase in service functions opened teaching positions for humanists and doubtless raised the prestige of the arts faculties. In the larger schools created in the course of the sixteenth century by combining several smaller Latin schools, too, Greek was offered as a new subject; in some of them even an introductory course of Hebrew was taught. Several of these larger academic schools in fact covered areas of knowledge that were traditionally part of the curriculum of higher education

As for the effects of the Reformation on the development of educational systems, only two aspects will be mentioned here: it supported the relatively fast spreading of literacy in the mainly Protestant areas of northern Europe (the concept of universal priesthood combined with the accessibility of the sources of faith translated into the vernacular to every literate person). Sweden seems to have been most successful in this respect. Already in the seventeenth century, admission to full church membership was tied to the ability to read. Prussia was the first state to initiate steps towards compulsory elementary education early in the eighteenth century. Other Protestant states followed suit early in the nineteenth century. Fairly reliable quantitative data on the development of attendance at primary schools and on the spreading of literacy are available only for the decades from about 1830 onwards. In both respects, they show a considerable

advantage for the Protestant territories (Flora 1975, Flora et al. 1983).

In central Europe, in particular, the Reformation changed the relation between the university and the polity. The universities became part of a territorial state with a distinctly denominational character that assumed control over the reorganization of the religious system and of the educational institutions still closely connected with it. This process entailed a loss of autonomy on the part of the universities, a loss which increased with the growth of the central state administration This is clearly visible in the Protestant territories, where the ruling dukes and princes were actively engaged in the process of reorganizing the universities and of creating a new type of academic school (Landesschule, Fürstenschule), financed directly by the state and recruiting its students more or less statewide. The control of the elementary schools, financed by the communes with few additional resources provided by the state, remained mostly in the hands of the church. Towns continued to be responsible for the establishment and financing, mostly via fees, of Latin schools or different types of intermediate schools without foreign language instruction for children of well-to-do families (Leschinsky and Roeder 1976). To remit (partly) fees for motivated pupils from poor families seems to have been widely practiced.

In England, the impact of the Reformation on school development may have been even stronger than in central Europe: 'modernization of the educational system dates from the Reformation. It was during this period—from the early mid-sixteenth to the mid-seventeenth century—that grammar schools were established in most market towns, with separate school buildings being erected and teaching being developed as a profession, and a systematic approach to the structuring of the content of education … took root' (Simon 1987, p. 93).

Despite the changes observable in the schools and universities during the sixteenth century, they still had more in common with their precursors of the late Middle Ages than with the modern system of education which slowly evolved at the end of the eighteenth century and which developed with increasing rates of change and growth from the middle of the nineteenth century onwards. The main function of the university still is the transmission of a cultural tradition legitimized by traditions. Where such traditions were not established as in the case of Protestant theology, this led to bitter and fruitless quarrels between adherents of dissenting factions of Protestantism, who had their strongholds in different universities.

There seems to have been little room for creative minds in the fields of philosophy, mathematics, and natural science. Descartes, Spinoza, Newton, and Leibnitz stayed out of its bounds. Hopes for the progress of knowledge and for innovations with the potential of improving the lot of mankind concentrated not on the university but on the Academy, an association of men of proven ability, especially in mathematics, science, and technology (Stichweh 1984, pp. 67ff). That the academies established in Germany did not pose a threat to the position of the universities is mainly due to the fact that the latter, from the start of the eighteenth century, had embarked on the road to an ambitious reform.

4. *Reforming the University*

The University of Halle, founded in 1694, acted as forerunner and model in this development. It succeeded in attracting a number of outstanding teachers, among them August Hermann Francke, one of the heads of the Pietistic movement and an original pedagogical thinker, Christian Thomasius, a jurist influenced by Grotius and Pufendorf, who also taught and published in the German language, Christian Wolf, mathematician and an eminent advocate of rationalistic philosophy, who also published textbooks in German, and J. S. Semler, a theologist (1752–91) who was among the first to introduce the historical–critical method into the study of the Holy Scripture. The curriculum of the faculty of philosophy was extended to include history, geography, mathematics, and the natural sciences; the law school offered lectures on, among others politics and natural law. Halle clearly pointed the way to a modern university that was no longer just a teaching institution but had begun to accept systematic research as a major function and service to society.

The second model university of the eighteenth century was that at Göttingen, founded in 1734. Its focus was on the modern disciplines: political sciences, history, law, mathematics, and natural sciences; however, it was equally important for the development of the study of the civilizations and literature of classical antiquity. Lectures were given regularly in German. There existed something like a common understanding that it was an obligation of the university to extend knowledge in the respective disciplines by doing original research. Like A. H. Francke in Halle, both Gesner and Heyne, the dominant figures in the field of classical studies in Göttingen, were influential in improving the preparation of teachers and the educational practice in schools.

Whereas in France the universities and colleges of the Ancien Régime were suppressed by imperial order in 1793 and, in their stead, a series of professional schools were organized (Verger 1986), universities in Germany consolidated their dominant position in higher education and succeeded in safeguarding a sphere of autonomous decision making in the organization of their teaching and research. The pattern of the further development throughout the nineteenth century and beyond was already clearly discernible in

the conceptual framework elaborated by Fichte, Schleiermacher and W. v. Humboldt for the foundation of the Berlin university in 1810: the traditional ranking of the faculties was to be abolished; a central role was ascribed to philosophy, in the sense that a philosophical spirit ought to pervade work in all fields of study. The mission of the university was defined as science in the service of truth, i.e., doing research to extend the boundaries of knowledge. This mission explicitly included the students. They were expected, fairly early in the course of their studies, to choose topics of interest where they could do original work on a restricted scale, the results of which were to be presented and discussed in the seminar. The seminar provided a new setting for teaching and learning, adapted to the new mission of the university: a regular meeting of a professor with his more advanced students for the purpose of discussing original research presented either by the professor or by a student.

A growing specialization of both teachers and students and the formation of subdisciplines was one important effect of these developments, regarded as ambivalent by those concerned with the preparation of teachers for academic secondary schools, that profession to which most students of the philosophical faculty aspired (Paulsen 1919, Vol. II, pp. 259–78). For a comparative perspective on the 'transformation of higher learning … expansion, diversification, social opening, and professionalization in England, Germany, Russia and the United States' see Jarausch (1982).

5. *Schooling in the Eighteenth and Nineteenth Centuries*

The first decades of the nineteenth century were not only a period of change for the universities. They also saw the evolution of the dominant type of academic secondary schools (Gymnasium), the curriculum of which centered on the classical languages Latin and Greek and, to a lesser degree, on mathematics and German, without completely neglecting the study of history, geography, and science. Relations between secondary schools and universities were clearly articulated with the introduction of the final examination (Abitur) and the examination of teachers qualifying for employment (Müller and Zymek 1987). Parallel developments are observable in France with the evolution of the Lycée and the introduction of the Baccalauréat in 1820.

With respect to the development of elementary education, wide regional variations in the actual enrollment of that age group for which compulsory education was mandated are observable in the first half of the nineteenth century. The highest proportions of enrollment around 1820 was reached in Saxony, totaling more than 80 percent. An average for the whole of Prussia in 1816 was (probably over-) estimated at 60 percent. Three decades later, the least developed province of Prussia topped this mark with 70 percent of the age group under compulsory schooling actually enrolled. Obviously, there must have been a rather steady growth of the provision of elementary education over the period (Leschinsky and Roeder 1976, pp. 137ff). The expansion of the provisions for teacher training seems even more remarkable: 17 new seminars for the education of elementary school teachers were established during the period, raising the total number to 38. Their yearly output of qualified teachers seems to have been sufficient to meet the growing demand.

In England, compulsory attendance was not introduced until the final decades of the nineteenth century. But there, too, a rapid increase in literacy throughout the century is to be observed, much of it due to the willingness of working-class parents to pay for the education of their children in one of the many local private schools, which they often preferred to the cheaper schools established by the churches or by associations crusading for the education of the lower classes, as Thomas W. Laqueur contends (Stone 1976, pp. 192ff).

6. *The Emergence of Educational Systems*

At the start of the twentieth century, coherent systems of education emerged in the more developed European nations. The elementary schools no longer stood apart from the rest of the system, but developed their function as feeder institutions for the secondary schools. Within the secondary system or between secondary and primary schools, new types of schools were established, varying in curricular focus prestige and entitlements conferred upon them. Well into the twentieth century, the classical academic secondary school with its clear focus on Latin and Greek languages and literature—lycées, collèges in France, grammar schools and, of course, the small number of highly prestigious boarding schools, the so-called 'Public Schools,' in England, and the Gymnasium in Germany—succeeded in defending its top position. Below this type ranked secondary schools that did not offer Greek or neither Latin nor Greek but, instead, taught modern languages and reserved more hours of instruction for the sciences: enseignement spécial in France, higher-grade schools in England, Realschulen, Oberrealschulen, Realgymnasien in Germany. Many of these schools emerged owing to the activities of local pressure groups convinced that the classical curriculum did not meet the educational needs of their children. The ambitious blueprints for their modernized curricula, integrating elements of vocational and of general education, however, had already been proposed from the second half of the seventeenth

century onwards (Leschinsky and Roeder 1976, pp. 168–208). Some of these types of secondary schools eventually achieved parity with the classical academic secondary schools—if not in prestige, at least in entitlements to matriculation (Prost 1969, Müller et al. 1987, Müller and Zymek 1987). Central and regional bureaucracies, responsible for the development and functional control of the system, were expanded early in the nineteenth century in France and Germany. This happened considerably later in England, where an effective substitute had been developed in the form of the Royal Commissions 'with the brief of examining and making recommendations for the reform of all levels of education' (Müller et al. 1987, p. 88). Curricular frameworks and standards of examinations for different types of schools and teacher training were specified by the central administration, traditionally leaving most leeway for curricular decisions on the level of the individual school in England. The inclusion of females had progressed considerably throughout the nineteenth century, although parity was not achieved until the second half of the twentieth century.

7. *A Global Model of Schooling*

In reviewing the long-term evolution of educational institutions in the more developed countries of western, northern and central Europe and in the USA, one gains the impression that, despite national peculiarities and the weight of regional traditions, the systems emerging at the turn of the twentieth century are basically similar in structure and functioning. They seem to represent what has more recently been termed a 'global model of schooling.' The thesis that we are witnessing the 'formation of a world educational system' (Adick 1995) implies that the dynamics fueling the development of systems of mass education presupposes the existence of the modern nation state and the concept of citizenship but that it is to an important degree transnational. This thesis has received substantial support in empirical research conducted by John W. Meyer, one of its principle proponents, and his collaborators (Meyer et al. 1992). Thus, an analysis of constitutions enacted since about 1870 shows that practically all of them include statements concerning the individual right of access to education and the authority and obligation of the state to provide and regulate the educational institutions (Boli-Bennet in Meyer and Hannan 1979). The Prussian constitution of 1794 (Allgemeines Landrecht) can, at least in some respects, be cited as an important early precursor of these later developments. A study of primary school curricula enacted between 1920 and 1986 also reveals a high degree of communality with respect to, e.g., the subjects taught, hours of instruction per subject, and basic changes in the content of subjects across a large sample of countries (Meyer et al. 1992). Further support comes from case studies of educational developments in colonial and postcolonial West Africa (Adick 1992, 1995) that document autochthonous initiatives in the course of the nineteenth and early twentieth centuries to establish a school system for the native population. The problem of whether the curriculum should cultivate native African traditions or emulate the best international models was, and still is, openly debated (see Reichmuth (1995) for the influence of Christian and Muslim communities on education in Nigeria and Wiegelmann (1999) for the development of schooling in Francophone Senegal, with special attention to the problem of the language of instruction, afflicting many of these countries).

More recently, the evolution of a global model of schooling has been actively promoted by international agencies: OECD's 'Reviews of National Policies for Education' are based upon a continuously refined set of common criteria and standards used in the evaluation of national educational systems. The European Union, too, may induce a further rapprochement between educational systems of member states; mutual recognition of qualifications certified in schools and universities can be viewed as a step in this direction. Large international quantitative studies of school achievement presuppose communalities in the curricula of primary and secondary schools, their success in constructing achievement tests, developed in international groups of researchers and valid in countries otherwise different, offers proof that such communalities exist in fact; and decisions of national educational administrations taken in reaction to test results may again push the trend towards globalization.

To be sure, the global model still leaves room for variation between countries, e.g., with respect to ensuring elite education for minorities of high achievers, in the degree of comprehensiveness of higher secondary education, in the organization of different types and levels of vocational and higher education, and in the importance assigned to specific subjects like, for example, religion.

A final remark addresses the problem of large gaps in the global model: both the quality and the inclusiveness of educational provisions are far below accepted international standards even on the level of primary education in a number of developing countries, despite considerable progress achieved in the decades between 1960 and 1980. The report prepared for the 1990 'World Conference on Education for All' presents a long list of deficiencies in, e.g., equity of access, completion of primary cycle preparation of teachers, quality of instruction, availability and quality of instructional materials, and managerial competences of personnel at the central, regional, and school level (Lockheed and Verspoor 1990). The recommendations, elaborated in the report, for action to overcome these deficiencies may have been effective in some countries. In others, the situation may well have deteriorated further in the course of the 1990s. This, at least, is the impression gained from a survey of the

conditions of primary schools conducted in 14 of the least-developed countries in 1995 (Postlethwaite 1998). In significant parts of our world primary education for all still constitutes a vision rather than a reality.

See also: Cultural Diversity, Human Development, and Education; Educational Institutions and Society; Educational Sociology; Social Mobility, History of; Universities, in the History of the Social Sciences

Bibliography

Adick C 1992 *Die Universalisierung der modernen Schule. Eine theoretische Skizze zur Erklärung der weltweiten Verbreitung der modernen Schule in den letzten 200 Jahren mit Fallstudien aus Westafrika*. Ferdinand Schöning. Paderborn, Germany

Adick C 1995 Formation of a world educational system. In: Roeder P M, Richter I, Füssel J-P (eds.) *Pluralism and Education. Current World Trends in Policy, Law, and Administration*. Institute of Governmental Studies Press, University of California, Berkeley, CA

Craig J E 1981 The expansion of education. *Review of Research in Education* **9**: 151–210

Feldbusch E 1988 Entstehung der geschriebenen Sprache. In: Ammon U, Dittmar N, Mattheier K J (eds.) *Sociolinguistics. An International Handbook of the Science of Language and Society*. Walter de Gruyter, Berlin, Vol. II, pp. 1469–79

Flora P 1975 *Indikatoren der Modernisierung. Ein historisches Datenhandbuch*. Westdeutscher Verlag, Opladen, Germany

Flora P et al. 1983 *State, Economy, and Society in Western Europe. Vol. I. The Growth of Mass Democracies and Welfare States*. Campus Verlag, Frankfurt, Germany, Macmillan Press. London, St. James Press, Chicago

Garin E 1964 *Geschichte und Dokumente der abendländischen Pädagogik I. Mittelalter*. Rowohlt Taschenbuchverlag, Reinbek, Germany

Jarausch K H (ed.) 1982 *The Transformation of Higher Learning 1860–1930*. Klett-Cotta, Stuttgart, Germany

Leschinsky A, Roeder P M 1976 *Schule im historischen Prozess. Zum Wechselverhältnis von institutioneller Erziehung und gesellschaftlicher Entwicklung*. E. Klett Verlag, Stuttgart, Germany

Lockheed M E, Verspoor A M 1990 Improving primary education in developing countries. A review of policy options. Draft prepared for World Conference on Education for All, Bangkok, 1990

Marrou H I 1948 *Histoire de l'Éducation dans l'Antiquité*. Editions du Seuil, Paris

Meyer J W et al. 1992 *School Knowledge for the Masses: World Models and National Primary Curricular Categories in the Twentieth Century*. Falmer Press, Washington, DC

Meyer J W, Hannan M T 1979 *National Development and the World System, Educational, Economic, and Political Change, 1950–1970*. University of Chicago Press, Chicago

Müller D K, Ringer F, Simon B 1987 *The Rise of the Modern Educational System: Structural Change and Social Reproduction 1870–1920*. Cambridge University Press, Cambridge, UK, Editions de la Maison des Sciences de l'Homme, Paris

Müller D K, Zymek B 1987 *Datenhandbuch zur deutschen Bildungsgeschichte. Bd. II, 1. Teil. Sozialgeschichte und Statistik des Schulsystems in den Staaten des Deutschen Reiches 1800–1945*. Vandenhoeck und Ruprecht, Göttingen, Germany

Paulsen F 1919 *Geschichte des gelehrten Unterrichts auf den deutschen Schulen und Universitäten vom Ausgang des Mittelalters bis zur Gegenwart*. 2 Vols. Veitand, Leipzig, Germany

Postlethwaite N 1998 The conditions of primary schools in least-developed countries. *International Review of Education* **44**: 289–317

Prost A 1969 *Histoire de l'Enseignement en France, 1800–1967*. Colin, Paris

Reichmuth S 1995 Christian and Muslim communities and their influence on education in Nigeria. In: Roeder P. M, Richter I, Füssel H-P (eds.) *Pluralism and Education. Current World Trends in Policy, Law, and Administration*. Institute of Governmental Studies Press, University of California, Berkeley, CA

Roeder P M, Richter I, Füssel H-P (eds.) 1995 *Pluralism and Education. Current World Trends in Policy, Law, and Administration*. Institute of Governmental Studies Press. University of California, Berkeley, CA

Rüegg W (ed.) 1993/1996 *Geschichte der Universität in Europa*. 4 Vols. C. H. Beck, Munich, Germany

Schwenk B 1996 *Geschichte der Bildung und Erziehung von der Antike bis zum Mittelalter*. Deutscher Studien Verlag, Weinheim, Germany

Stichweh R 1984 *Zur Entstehung des modernen Systems wissenschaftlicher Disziplinen. Physik in Deutschland 1740–1890*. Suhrkamp Verlag, Frankfurt/Main, Germany

Stone L (ed.) 1976 *Schooling and Society. Studies in the History of Education*. Johns Hopkins University Press, Baltimore

Verger J (ed.) 1986 *Histoire des Universités en France*. Bibliothèque Historique Privat, Toulouse, France

Wiegelmann U 1999 *Alphabetisierung und Grundbildung in Senegal. Ein empirischer Vergleich zwischen modernen und traditionellen Bildungsgängen und Schulen*. IKO-Verlag für interkulturelle Kommunikation, Frankfurt/Main, Germany

P. M. Roeder

Educational Leadership

The quality of education provided in schools and colleges is commonly thought to be influenced by the quality of the leadership directing these organizations. But the meaning of leadership, how it differs from everyday management and administration, and how it is manifested in educational organizations are matters of considerable dispute. Differing views of the forms and effects of leadership in education are central to this article, as is the question of what, if anything, is distinctive about 'educational' leadership. Issues of power, ethics, democracy, and the consent and empowerment of organizational participants are central to the exercise of leadership. They are particularly salient in educational organizations, because of their formative and symbolic role in shaping a society's future. Such issues lead to disagreements about how

educational leadership should be conceived and researched, and ensure that it will remain a highly contested concept.

1. Generic Leadership Versus Educational Leadership

Scientific research on school administration and educational leadership began in the late 1950s and early 1960s, when researchers in the 'theory movement' began utilizing the techniques of the social sciences to examine these topics (Willower and Forsyth 1999). Although researchers recognized that schools and colleges were totally different from business and industrial organizations, little attention was given to differences among types of organizations in the early research. Instead, leadership and administration were treated as generic organizational issues for many years, with few questions asked about what might be distinctive about these functions or processes in educational settings. To the extent that a generic approach is considered sufficient, one only needs to look at the generic literature on leadership and administration in organizations. The main argument in this article, however, is that generic approaches, although valuable, are insufficient for comprehending effective leadership in educational organizations.

Many definitions of leadership exist, but most acknowledge that it involves the ability to influence others—often defined as the ability to get them to do what they otherwise would not do—coupled with the ability to get people to accept a mission or vision of the goals or purposes a group, a team, or organization should follow (Hoy and Miskel 1996, Leithwood and Duke 1999). To distinguish between leadership and management, some definitions emphasize that leadership involves bringing about change or innovation, while management involves maintaining the smooth operation of the existing organization. Research on leadership has evolved from early studies that focused on the traits or abilities of effective leaders, to recognition of the importance of the situation or context that leaders face. Thus, a 'contingency' approach asks what traits in what situations are important for effective leadership (Hoy and Miskel 1996).

1.1 The Importance of Context

The organizational structure of typical schools and colleges involves a context in which teachers are solo practitioners doing discretionary, professional work. This workplace impedes direct supervision and fosters the autonomy of employees (Lortie 1975). This context thus reduces the effectiveness of a directive, 'command and control' style of leadership and seems conducive to a more collaborative approach. This raises the question of how leaders can best improve the effectiveness of schools and teachers. The early research on educational leadership emphasized strong, directive leadership by school administrators, but more recent research suggests that a more participatory and collective approach to leadership is needed, to obtain the consent of employees and to build a sense of 'professional community' in schools (Newmann and Wehlage 1995). Some of the contextual aspects of schools, such as the tension between their bureaucratic and professional features, create what Ogawa et al. (1999) call the 'enduring dilemmas of school organization.'

How much teamwork and democracy are really needed in schools seems to be both a cultural and philosophical question. In the English-speaking democracies, cultural values press for a democratic approach, even if it is seldom fully achieved. In other societies, for example many Asian nations noted for high student achievement, schools are usually run in a more traditionalistic, 'top-down' manner. The Asian nations show that, in a cultural context supportive of student effort, traditional values, and the importance of learning, traditionally structured schools, utilizing top-down management and teaching, can be very successful. In more pluralistic Western societies—less supportive of effort, traditional values, and learning—effective schooling is more difficult to achieve. Firestone and Louis (1999) provide an interesting discussion and set of references on these themes.

To the extent that schools, wherever they are located, can be instructionally effective through some combination of high academic standards, strong teachers, and a milieu supportive of learning, they may not need to undertake school improvement efforts. But many, if not most schools lack these qualities and need systematic improvement efforts. Research has shown repeatedly that teachers in these more typical schools often lack the motivation and/or the skills or capacity to successfully implement new programs or teaching approaches (Fullan 1991). The need to enlist their involvement and effort in school improvement ventures creates a demand for educational leadership, i.e., leadership primarily focused on facilitating and improving the core function of educational organizations: teaching, and learning. The need for school improvement also calls for leaders adept at motivating and managing the change process in organizations.

1.2 Models of Leadership

In a comprehensive review of the research literature on school leadership, Leithwood and Duke (1999) delineate six models of leadership as predominating in the articles published since 1988 in four leading English-language educational administration journals. Of the three most frequently mentioned models,

'instructional leadership' led with 13 mentions, followed by 'transformational leadership' (11 mentions), and 'contingent/leadership styles' (nine mentions). The other three leadership models uncovered were 'moral leadership,' 'managerial leadership,' and 'participative leadership.' Instructional leadership is close in meaning to the broader concept of 'educational leadership' explicated here. Transformational leadership involves leadership that inspires and heightens the commitments and capacities of organization members toward the attainment of organizational goals. In contrast, 'transactional leadership' involves the exchange of valued things between leaders and followers in a more extrinsic relationship that lacks the quality of a mutual pursuit of higher purposes.

Leithwood and Duke (1999, p. 54) state that the contingent/leadership styles approach focuses on 'how leaders respond to the unique organizational circumstances or problems they face as a consequence, for example, of the nature and preferences of coworkers, conditions of work, and tasks to be undertaken.' Moral leadership (see Hodgkinson 1991) focuses on 'the values and ethics of the leader, so authority and influence are to be derived from defensible conceptions of what is right or good' (Leithwood and Duke 1999, p. 50). Managerial leadership 'focuses on the functions, tasks, or behaviors of the leaders' (p. 53) in the belief that competent performance of these functions will facilitate work within the organization. Participative leadership emphasizes the group's decision-making processes and collective involvement. As argued earlier in this article, it represents the kind of leadership increasingly in demand in democratic societies.

1.3 Democratic or Shared Leadership

Fearing that leaders tend to abuse their power—in education as elsewhere—critics call for more democratic, shared or collective governance arrangements (Chapman et al. 1995). Advocates of educational leadership, while recognizing the possibility of negative leadership, emphasize the importance of leadership for organizational improvement and extol the value of leaders capable of communicating an inspiring vision of a more dynamic and effective organization. However, whether educational leadership should be viewed, normatively or empirically, as the property of one or a few individuals or as a collective organizational property is an increasingly important question, both in research and practice.

Since the 1970s, scholars embracing critical theory and post-positivist orientations have mounted a sustained critique on mainstream approaches to social science, organizational theory, and leadership (Heck and Hallinger 1999). As Leithwood and Duke (1999, p. 63) note, three features of the theory movement in educational administration have come under attack: 'a positivist orientation toward knowledge, a structural–functionalist conception of organizations, and a control-oriented, managerial view of leadership.' Arguing that knowledge is actually 'socially constructed,' that hierarchical relationships in organizations reinforce unjust inequalities in society, and that a control orientation toward leadership and administration impedes reform and empowerment, critics such as Foster (1989) contend that educational leadership must focus on social justice and emancipation.

Along with concerns for democracy and justice, pressures for more shared or participative leadership also increasingly flow from demands for higher performance and better student achievement. The case for 'high-involvement' management for high-performance schools is well stated by Darling-Hammond (1996, pp. 146–7).

Efforts to invent nonbureaucratic twenty-first-century organizations tend to use incentives and structures that motivate through collaboration rather than coercion; build strong relationships rather than rely solely on rules for governing behavior; encourage quality by structuring work around whole products or services rather than disconnected piecework; and create information-rich environments that support widespread learning and self-assessment among workers rather than rely primarily on hierarchical design and supervision of work routines. Such organizations aim to stimulate greater thoughtfulness and creativity rather than focusing largely on enforcing compliance with predetermined procedures. Their success, then, depends on the creation of new opportunities for teacher and school learning, new modes of accountability, and new kinds of incentives for continual improvement and problem solving.

Research on school improvement efforts and, particularly, on the professional development of teachers to facilitate these efforts, shows how the theory and practice of educational leadership can move toward approaches that are more participatory and democratic. Smylie and Hart's (1999) perceptive synthesis of research on school leadership and teacher learning and change illustrates this well. Building on work on human and social capital development, they show how teacher collaboration in improvement initiatives and teachers' sense of membership in a professional community can be fostered through more participative and distributed forms of leadership. Smylie and Hart (1999, p. 428) contend that, rather than a hierarchical, bureaucratic style of leadership, 'from social and human capital development perspectives, attention is drawn toward relational and interactional aspects of school leadership. These perspectives call for a broader conception of school leadership, one that shifts from a single person, role-oriented point of view to a view of leadership as an organizational property shared among administrators, teachers, and perhaps others' (p. 428). This view 'sees leadership as participative and inclusive, as infused throughout the social system of

the school in ways that foster collective responsibility, mutual trust and obligations, and joint accountability' (p. 436).

Smylie and Hart emphasize that school principals still have vital roles to play in this shared approach to leadership, for example, in creating opportunities for interaction, fostering social bonds, and establishing support systems. Further, they stress that efforts to move to this participative model always encounter resistance because they 'challenge deeply entrenched patterns of isolation and autonomy that define teachers' work' (p. 430). Leadership and improvement strategies for pulling teachers out of these deeply engrained norms and habits are the subject of extensive research and writing (Sykes 1999), some of which cautions against manipulating teachers against their will.

2. School Culture and Change

Firestone and Louis' (1999) review of research on 'schools as cultures' emphasizes that the shared culture of isolation and autonomy among teachers in many schools frequently inhibits change and improvement efforts. Nevertheless, they argue that, within some limits, leaders can begin a process to change a school's culture toward one more conducive to participation, professionalism, and improvement. This can be done, in part, through the effective use of 'symbolic leadership,' in which the leader uses symbolic language and ceremonies, highlights key values and shared meanings, and reinterprets events to renew or redirect the organization's culture. A related strand of research, in the broad field of school improvement and organizational change literature (Louis et al. 1999), involves the growing interest in leadership approaches to foster 'organizational learning' in schools (Leithwood and Louis 1998). The well-known 'continuous improvement' and 'total quality management' approaches represent views of how organizational learning can be institutionalized in organizations.

Differences in national cultures associated with such core values as 'equality versus inequality, collectivism versus individualism, masculinity versus femininity, and uncertainty avoidance versus tolerance for ambiguity' are among the factors that affect leadership and the culture of schools (Firestone and Louis 1999, p. 302). At the same time, the increasing forces of globalization produce pressures and trends that tend to reduce national differences and 'internationalize' or standardize the character of education systems and educational administration (Chapman et al. 1999). For example, growing international pressures to honor diversity and end discrimination based on race, religion, gender, etc., are changing the culture of schools worldwide and opening the doors of school leadership to people from all backgrounds. Thus, effective educational leadership is likely increasingly to require more inclusive and collaborative approaches that foster collective responsibility and commitment, thereby promoting professional development and school improvement.

See also: Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Leadership: Political; Leadership, Psychology of; School Administration as a Field of Inquiry; School as a Social System; School Management; Schools, Micropolitics of

Bibliography

Chapman J, Froumin I, Aspin D (eds.) 1995 *Creating and Managing the Democratic School*. Falmer, London

Chapman J, Sackney L E, Aspin D 1999 Internationalization in educational administration: policy and practice, theory and research. In: Murphy J, Louis K S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco

Darling-Hammond L 1996 Restructuring schools for high performance. In: Fuhrman S H, O'Day J A (eds.) *Rewards and Reform: Creating Educational Incentives that Work*. Jossey-Bass, San Francisco

Firestone W A, Louis K S 1999 Schools as cultures. In: Murphy J, Louis K S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco

Foster W 1989 Toward a critical practice of leadership. In: Smyth J (ed.) *Critical Perspectives on Educational Leadership*. Falmer Press, London

Fullan M 1991 *The New Meaning of Educational Change*. Teachers College Press, New York

Heck R H, Hallinger P 1999 Next generation methods for the study of leadership and school improvement. In: Murphy J, Louis K S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco

Hodgkinson C 1991 *Educational Leadership: the Moral Act*. SUNY Press, Albany, NY

Hoy W K, Miskel C G 1996 *Educational Administration: Theory, Research, and Practice*, 5th edn. McGraw-Hill, New York

Leithwood K, Duke D L 1999 A century's quest to understand school leadership. In: Murphy J, Louis K S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco

Leithwood K, Louis K S 1998 *Organizational Learning in Schools*. Swets and Zeitlinger, Lisse, The Netherlands

Lortie D C 1975 *Schoolteacher: a Sociological Study*. University of Chicago Press, Chicago

Louis K S, Toole J, Hargreaves A 1999 Rethinking school improvement. In: Murphy J, Louis K S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco

Newmann F N, Wehlage G G 1995 *Successful School Restructuring*. Center on Organization and Restructuring of Schools, School of Education, University of Wisconsin–Madison. Madison, WI

Ogawa R T, Crowson R L, Goldring E 1999 Enduring dilemmas of school organization. In: Murphy J, and Louis K. S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco, CA

Smylie M A, Hart A W 1999 School leadership for teacher learning and change: A human and social capital development

perspective. In: Murphy J, Louis K S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco
Sykes G 1999 The 'new professionalism' in education: an appraisal. In: Murphy J, Louis K S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco, CA
Willower D J, Forsyth P B 1999 A brief history of scholarship in educational administration. In: Murphy J, Louis K S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco, CA

W. L. Boyd

Educational Learning Theory

1. Introduction

There are distinct traditions in education that derive from differing perspectives on the phenomena of the domain. We organize our discussion around three general perspectives that have developed in educational research. We recognize that other organizing principles could be chosen, and that many of our colleagues would characterize the field in different terms. Our version groups together many research contributions that could be distinguished in important ways. We have arrived at this grouping in our own effort to understand broad trends in educational research.

The perspectives correspond to three general views of knowing and learning, which following Case (1992) we refer to as *empiricist*, *rationalist*, and *pragmatist–sociohistoric*. (Case used the simpler term 'sociohistoric.') Empiricism, typified by Locke and Thorndike, emphasizes consistency of knowledge with experience. Rationalism, typified by Descartes and Piaget, emphasizes conceptual coherence and formal criteria of truth. Pragmatism, typified by Dewey and Mead, and sociohistoricism, typified by Vygotsky, emphasize that knowledge is constructed in practical activities of people as they interact with each other and their material environments. Current manifestations of these three perspectives are the *behaviorist* perspective, the *cognitive* perspective, and the *situative* perspective.

All three of these perspectives have contributed, and continue to contribute, important insights to fundamental scientific knowledge and understanding of cognition and learning, and have influenced educational practices significantly. While each perspective is valuable, they frame theoretical and practical issues in distinctive and complementary ways, somewhat in the way that physics, chemistry, and biology frame issues surrounding genetic replication in different ways. This essay is drawn from an extended discussion of the three perspectives in Greeno et al. (1996).

2. Overview of the Three Theoretical Perspectives

There are three thematic issues that characterize the perspectives:

(a) The nature of knowing.
(b) The nature of learning and transfer.
(c) The nature of motivation and engagement.

In the behaviorist/empiricist view, knowing is an organized accumulation of associations and components of skills. *Learning* is the process in which association and skills are acquired, and *transfer* occurs to the extent that behaviors learned in one situation are utilized in another situation. Motivation is a state of the learner that favors formation of new associations and skills, primarily involving incentives for attending to relevant aspects of the situation and for responding appropriately. There are three traditions that we consider contributed to this view. *Associationism*, which goes back to Locke and David Hume, viewed knowing as the associations between ideas, and learning as building new associations. *Behaviorism* took the position that knowing could be characterized only in terms of observable connections between stimuli and responses, and learning in terms of forming and strengthening or weakening and extinguishing those connections through reinforcement or non-reinforcement. *Connectionism* (or neural networks) treats knowledge as the pattern of connections between neuron-like elements, and learning as the strengthening or weakening of those connections.

The cognitive/rationalist perspective on knowledge emphasizes understanding of concepts and theories in different subject matter domains and general cognitive abilities, such as reasoning, planning, solving problems, and comprehending language. There are three traditions of research that we consider to be branches of the rationalist perspective. The oldest of these is *gestalt psychology*, which emphasized the structural nature of knowledge and the importance of insight in learning. A second tradition, *constructivism*, was originally developed by Piaget and is focused on characterizing the cognitive growth of children, especially their growth in conceptual understanding. The third tradition, *symbolic information processing*, was developed in cognitive science by Chomsky, Simon, Newell, and others, and is focused on characterizing processes of language understanding, reasoning, and problem solving. All three traditions emphasize the importance of organized patterns in cognitive activities. The constructivist and information-processing traditions also focus on procedures and operations for representing and reasoning about information. Learning is understood as a constructive process of conceptual growth, often involving reorganization of concepts in the learner's understanding, and growth in general cognitive abilities, such as problem solving strategies and metacognitive processes. Discussions of motivation often emphasize that much learning occurs without the need for extrinsic incentives, as in the case

of learning one's first language, and focus on ways to foster the intrinsic interest of learners in ideas and concepts.

The situative/pragmatist–sociohistoric perspective views knowledge as distributed among people and their environments, including the objects, artifacts, tools, books, and the communities of which they are a part. Analyses of activity focus on processes of interaction of individuals with other people and with physical and technological systems. Several research traditions have contributed to the situative perspective. The best established of these is *ethnography*, including the study of cultural practices and patterns of social interactions, as well as discourse analysis and conversation analysis in activity theory, sociolinguistics, anthropology, and sociology. Another research tradition is *ecological psychology*, which studies behavior as physical interaction in which animals, including people, participate in physical and technological systems. A third research tradition is *situation theory*, in logic and philosophy, which analyzes meaning and action as relational systems and is developing a reformulation of logic to support these relational analyses. Knowing in this perspective is both an attribute of groups that carry out cooperative activities and an attribute of individuals who participate in the groups. Learning by a group or individual involves becoming attuned to constraints and affordances of the material and social systems with which they interact. Discussions of motivation in this perspective often emphasize engagement of individuals with the functions and goals of the community, including interpersonal commitments and ways in which individuals' identities are enhanced or diminished by their participation.

3. *Acquiring and Applying Associations: The Behaviorist/Empiricist View*

When people's knowledge is viewed as having associations between ideas or stimuli and responses, learning is the formation, strengthening, and adjustment of those associations. The research on basic associative processes of learning has important implications for teaching and learning. One is the importance of individual students having opportunities to give responses of the kind that they are to learn, and of feedback that is contingent on the individual student's responses.

Researchers in behavioral conditioning also found that effective learning usually requires significant preparation, or *shaping*, in which the learner becomes oriented to the general conditions where learning will occur. In conditioning experiments with animals, shaping involves a period in which the trainer attends carefully to the animal's activity in the learning environment, first providing reinforcement for being near the goal (e.g., a disk that a pigeon can peck), then for orienting toward the goal, then for touching it, and finally only for pecking it; the response that is desired. This kind of instruction by approximation has clear parallels in school learning, where skilled teachers attend to students' progress and provide encouragement for students' efforts as they achieve better approximations to the behaviors they need to succeed.

Analysis of complex tasks into learning hierarchies (Gagné 1968) has been used in designing instruction sequences and computer-based systems for learning routine skills. The hypothesis that smaller units of behavior need to be mastered as prerequisites for more complex units provides a basis for arranging sequences of instruction in which students are able to succeed by learning in small steps. This *decomposition hypothesis* is currently being questioned by many in the cognitive community (e.g., Resnick and Resnick 1991), based on a concern that instruction limited to components can result in mechanical knowledge without sufficient development of the conceptual basis for procedures that are learned.

In the view that knowing is having associations, learning in a new situation depends on how many, and which kinds of, associations needed in the new situation have already been acquired in the previous situation. The idea of *transfer* in conditioning involves *gradients* of similarity along stimulus dimensions, so that a response learned as an association to one stimulus generalizes more strongly to other stimuli that are similar to it in all respects, and less strongly to stimuli that differ from it in one or more dimensions. Thorndike (1903) expressed this as a theory of transfer based on common elements.

4. *Acquiring and Using Conceptual and Cognitive Structures: The Cognitive/Rationalist View*

Research on children's *learning* has been a significant topic in developmental psychology. Piaget's influence on educational practice has been considerable, especially informing efforts to organize science learning in a way that involves students' discovery of principles and concepts. Current research is developing accounts of the rich intuitive conceptual understanding by children, which undergoes significant change as they grow older. The research emphasizes that children's learning must be viewed as transforming significant understanding that they already have, rather than as simple acquisitions written on blank slates.

Information-processing models characterize learning in terms of procedures that represent task information, set goals, and transform symbolic expressions to satisfy the main problem goal. As an example, Anderson (1983) simulates three kinds of learning processes: proceduralization, tuning, and automization. Researchers have also investigated reasoning and

understanding that depends on mental representations, called *mental models*, that provide a kind of simulation of events, rather than descriptions of events (Gentner and Stevens 1983). In reasoning with a mental simulation, a model represents properties of the system, and operating on the model changes some of those properties in ways that correspond to changes in properties of the system.

An important theme in the cognitive view of learning is the concept of *metacognition*, the capacity to reflect upon one's own thinking, and thereby to monitor and manage it. These strategies have been studied under many labels, all pointing to the importance of self-conscious management of one's own learning and thinking processes. This theme was introduced by developmental psychologists (e.g., Brown 1978, Flavell and Wellman 1977) who noted that a reflective, self-monitoring capacity discriminated developmentally advanced children from their less advanced peers.

In the cognitive perspective, *transfer* is assumed to depend on acquiring an abstract mental representation that is invariant across situations. A large body of research has found that students often fail to transfer their learning. Brown (1989) pointed out that in research about children's ability to transfer, the deck is stacked in favor of finding that transfer does not occur. Children are asked to solve a problem, then a new problem is presented, and the experimenter observes whether the new problem is solved in a way that uses the initial solution. Usually, experimenters do their best to hide the relation between the two problems when the second problem is presented, so that if children do transfer, we can be sure they did so spontaneously. More importantly, for Brown, the potential generality of the initial solution is not made clear. When Brown and Kane (1988) taught solutions of problems and asked children to explain why the solutions were examples of general themes, thus calling attention to their potential generality, the children in their experiments transferred much more successfully.

There is an important theoretical and educational principle in these results about transfer. The manner in which solutions of problems are presented can make a major difference for the generality of what is learned. If students understand the solution as an example of a general method, and if they understand the general features of the learning situation that are relevant to use of the method, the abilities they learn are more likely to apply generally.

5. *Becoming Attuned to Constraints and Affordances through Participation: The Situative/Pragmatist–Sociohistoric View*

When *knowing* is viewed as practices of communities and of the abilities of individuals to participate in those practices, then *learning* is the strengthening of those practices and participatory abilities. Systems in which individuals learn to participate in social practices are very common, and include apprenticeship and other forms of being initiated into the practices of a group. Lave and Wenger (1991) reviewed several studies of learning involving apprenticeship and concluded that a crucial factor in the success of such a system is that learners must be afforded legitimate peripheral participation, which involves access to the practices that they are expected to learn and genuine participation in the activities and concerns of the group. Lave and Wenger characterized learning of practices as processes of participation in which beginners are relatively peripheral in the activities of a community, and as they become more experienced and adept, their participation becomes more central. A crucial issue in the nature of learning is whether, and in what ways, the peripheral participation of beginners is legitimate. For an environment of apprenticeship to be a productive environment of learning, learners need to have opportunities to observe and practice activities in order to progress toward more central participation.

The view that learning occurs through participation is at the root of the practices of apprenticeship, where apprentices are guided and supervised by masters. In successful apprenticeship learning, masters teach by showing apprentices how to do a task (modeling), and then helping them as they try to do it on their own (coaching and fading). Lave and Wenger (1991) emphasized how an apprentice's identity derives from becoming part of the community of workers. The motive for becoming a more central participant in a community of practice can provide a powerful motivation for learning. Of course, what is learned in apprenticeship may not generalize easily to other contexts. Collins et al. (1989) attempted to characterize how the modeling, coaching, and fading paradigm of apprenticeship might be applied to learning the cognitive subjects of school, in an approach they called 'cognitive apprenticeship.'

A major goal of educational reform is to have students participate more actively in learning communities, including participation in formulating and evaluating questions and problems, and constructing and evaluating hypotheses, evidence, arguments, and conclusions (Brown and Campione 1996). Abilities for participating in these activities have to be learned, and the research literature on that kind of learning is sparse. Several projects have focused on creating classroom practices of discussion and inquiry, and the investigators in those projects have discussed some aspects of the process of establishing norms and expectations by the students that support productive collaborative learning (Cohen 1986, Lampert 1990, Slavin 1983).

In the view of learning as coming to participate in a community of practice, *transfer* becomes a problematic issue. The question is whether transfer applies to new practices within the community (e.g., for school

communities this might mean working new problems or accomplishing new kinds of tasks) or to practices outside the community (e.g., for school these might be work environments). Many of the resources and supports that occur within a community of practice do not carry over to a different community, and so the problem of transfer becomes one of marshalling the resources needed to be successful in a new environment. This requires sophisticated social and information-processing skills: the kinds of skills that businesses think they will need in the future.

For transfer to be possible, there must be some constraints and/or affordances that are invariant under the transformations that change the learning situation into the transfer situation. For transfer to occur, learners must become attuned to those invariants in their initial learning. One of the ways to be attuned is to have an abstract representation that can apply in the new situation, but this is only one possible way for attunement to occur, and may not be the typical way for many learned activities to generalize (Greeno et al. 1993).

See also: Behaviorism, History of; Education: Phenomena, Concepts, and Theories; Instructional Psychology; Learning Theories and Educational Paradigms; Lifelong Learning and its Support with New Media: Cultural Concerns; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development; Motivation, Learning, and Instruction.

Bibliography

Anderson J R 1983 *The Architecture of Cognition*. Harvard University Press, Cambridge, MA

Brown A L 1978 Knowing when, where, and how to remember: A problem of metacognition. In: Glaser R (ed.) *Advances in Instructional Psychology*. Lawrence Erlbaum, Hillsdale, NJ, Vol. 1, pp. 77–166

Brown A L 1989 Analogical learning and transfer: What develops? In: Vosniadou S, Ortony A (eds.) *Similarity and Analogical Reasoning*. Cambridge University Press, Cambridge, UK, pp. 369–412

Brown A L, Campione J 1996 Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In: Schauble L, Glaser R (eds.) *Innovations in Learning: New Environments for Education*. Lawrence Erlbaum, Mahwah, NJ, pp. 289–325

Brown A L, Kane M J 1988 Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology* **20**: 493–523

Case R 1992 Neo-Piagetian theories of cognitive development. In: Sternberg R J, Berg C A (eds.) *Intellectual Development*. Cambridge University Press, New York, pp. 161–96

Cohen E G 1986 *Designing Groupwork*. Teachers College Press, New York

Collins A, Brown J S, Newman S E 1989 Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In: Resnick L B (ed.) *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*. Lawrence Erlbaum, Hillsdale, NJ, pp. 453–94

Flavell J, Wellman H M 1977 Metamemory. In: Kail R V Jr, Hagen J W (eds.) *Perspectives on the Development of Memory and Cognition*. Lawrence Erlbaum, Hillsdale, NJ

Gagné R M 1968 Learning hierarchies. *Educational Psychologist* **6**: 1–9

Gentner D, Stevens A L (eds.) 1983 *Mental Models*. Lawrence Erlbaum, Hillsdale, NJ

Greeno J G, Collins A, Resnick L B 1996 Cognition and learning. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Macmillan, New York, pp. 15–46

Greeno J G, Smith D R, Moore J L 1993 Transfer of situated learning. In: Detterman D K, Sternberg R J (eds.) *Transfer on Trial: Intelligence, Cognition, and Instruction*. Ablex, Norwood, NJ, pp. 99–167

Lampert M 1990 When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal* **17**: 29–64

Lave J, Wenger E 1991 *Situated Learning*: *Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK

Resnick L B, Resnick D P 1991 Assessing the thinking curriculum: New tools for education reform. In: Gifford B R, O'Connor M C (eds.) *Changing Assessment: Alternative Views of Aptitude, Achievement, and Instruction*. Kluwer, Boston, pp. 37–75

Slavin R E 1983 *Cooperative Learning*. Longman, New York

Thorndike E L 1903 *Educational Psychology*. Lemke and Buechner, New York

A. Collins, J. G. Greeno, and L. B. Resnick

Educational Media

Since Thorndike (1912) recommended pictures as a labor saving device for teachers, educational media have been the focus of research in many disciplines. Media such as computers and television are perceived widely as important because they provide essential communication links allowing the exchange of educational and other information between diverse and geographically separated groups and individuals.

Depending on the discipline conducting investigations, media have been defined from points of view that include journalistic, communication, philosophical, historical, economic, sociological, and psychological perspectives. The result is a mountain of studies exploring the influence of media on outcomes such as knowledge acquisition, political opinion, social learning, access to information, values, attitudes, business development, perceptions of reality, economic development, socialization processes, and learning in settings such as communities, schools, homes, and work. Most of this research has progressed without the unifying aid of theory development. The central focus of most studies is a concern with the impact of media on the learning and motivation of individuals,

groups, and societies. The emphasis in this review will be on educational uses of media from the perspective of social-psychological research.

1. Three Social-psychological Definitions of Educational Media

European, North American, and Asian social-psychological researchers tend to define media in one of three ways. First is the 'media as vehicles' definition where two or more information storage and delivery devices are examined for their capacity to provide different learning benefits and efficient access by learners to knowledge and instruction. The goal in many of these studies is to find the media or mix of media that provide the greatest learning benefit for different learners and curricula. Examples of media that fit this definition are books, cinema, television, radio, newspapers, and computers when used for teaching and learning. Two types of studies are often conducted. Comparison studies examine the relative learning and cost benefits of two or more alternative media (see for example, summaries and meta-analyses of these studies by the Kuliks and their co-workers (Kulik and Kulik 1991) to determine which medium or mix of media contributed more learning benefit. In most instances, the results of meta-analytic studies indicated that computer-based instruction resulted in learning benefits of approximately one-half a standard deviation over traditional media with a variety of curricula and learners. Other studies in this area question the relative efficiencies of different educational uses of media (for example, Levin 1983) and their contribution to instructional design (efficiency is often defined as the cost of different media for learner access to instruction in time, money, mental, or physical effort), time required to reach a preset learning criterion, and ease of use. The results of Levin's studies suggest that media can be very cost-efficient for educational purposes when the numbers of students to be served are quite large and geographically separated.

1.1 Media as Vehicles Plus Teaching Methods

A second approach can be found in the integrated 'media as vehicles plus teaching method' studies. Here media are defined not only as vehicles but also by the instructional methods or strategies they permit to teach specific curriculum content to specific learners. Most North American and European researchers with this view believe that it is not possible or beneficial to separate the delivery and instructional content functions of media (see Kozma 1994). A familiar 'media plus method' example is the use of the 'interactivity' between media and learner possible in computer-based instructional programs. This definition encompasses the capacity of some media to permit many real-time, interactive instructional 'transactions' (for example, a learner might respond to computer provided questions followed by corrective or elaborative feedback from the computer about the learner's responses). Typical research in the 'media and method' approach examines the learning benefits of developed instructional media programs such as Sesame Street or the Jasper Woodbury science program (Kozma 1994). An offshoot of this approach is exemplified by 'media attributes' studies. Researchers in this area study the capabilities of different media to provide a model for some of the mental processes required for learning. For example, Salomon (1979/1994) described studies that examined the capacity of cinema and television to 'zoom' from a wide field into a more narrow or specific view of a complex picture. After a zooming treatment where learners who had initial difficulty noticing parts of paintings significantly increased not only the number of details they reported but also increased the number of alternative views about the purpose of the artist. Salomon reasoned that his subjects had learned not only to notice more aspects and interpretations of paintings, but also had learned the perceptual and mental process of 'zooming' from this attribute or symbolic mode of real-time visual media.

1.2 Media as Learning Environments

A number of European, Asian, and some North American researchers have tended to take a 'media as learning context' approach to educational media inquiry. In this definition, media are viewed as contexts where education occurs, much like research on schools or homes as settings for educational experiences. Examples of this approach were described in an edited series by De Corte et al. (1992). The learning context focus investigates the ways that 'learning with media' in an educational context (as contrasted with 'learning from media' in the media as vehicles and methods approach) results in transformed learning environments and more durable learning-to-learn strategies for students. In these studies, the units of analysis are both the learner and the educational system. When new electronic media are added to educational settings '*Most everything has changed—curriculum*, *activity*, *engagement in learning*, *instruction*, *social interaction*, *teacher's roles*, (and) *modes of student evaluation* …that far transcended the mere introduction of computers' (Salomon 1992, p. 259, emphasis in original). The primary concern of learning context studies is to determine ways that educational media can modify and improve instructional settings. A more recent example of this work can be found in various studies reported by Kiesler and Kraut (see a summary of alternative views in Kiesler and Kraut 1999) on the

negative social and psychological consequences of increased Internet use in distance, media-based education, and communication.

2. Disputes about Methodological Issues

Educational media research has suffered from methodological disputes at least since the 1960s when Lumsdaine (1963) cautioned researchers about problems in the design of educational media studies. He claimed that the design of studies up to that date had confounded the effects of medium with the separate learning effects of the content and instructional strategies contained in media presentations. Salomon and Clark (1977) summarized the methodological concerns of Lumsdaine and others by distinguishing between research with media and research on media. Research with media employed newer media to study the effects of new (vs. old) teaching methods and curriculum reforms. An example would be the use of television to study the learning benefits of altering the length of educational presentations. Learning benefits found in these studies were sometimes mistakenly attributed to the medium. Research on media examined the unique contributions of various media to learning of different curriculum elements by different groups of learners.

2.1 Recent Methodological Disputes

Clark (1994) has argued that all studies that find learning or motivational benefits from media are flawed methodologically. He suggested that all learning benefits attributed to media are actually due to the uncontrolled and confounded effects of different instructional methods and information content in different treatments. He also claimed that all learning benefits attributed to any one medium or media attribute would also be found to be available from other media or attributes of media and therefore the benefits of media are economic (some media cost less than others to deliver the same instruction with similar learning benefits). He summarized his argument by suggesting that media do not influence learning any more than the truck that delivers food to a market influences the nutrition of its customers. Kozma (1994) and others have disagreed with Clark and other methodological critics of media studies. He argues for the medium plus method approach and suggests that Clark's 'separation of media from method creates an unnecessary and undesirable schism between the two…(because)…a medium's capabilities enable methods and the methods that take advantage of these capabilities of media are going to influence learning—methods must be confounded with media' (Kozma 1994, p. 16). A number of other media plus method and media as context researchers have agreed with Kozma (see a number of papers on this topic in Clark 1994).

3. Future Directions for Educational Media Research

A variety of new directions seem to develop constantly in educational media research. Most of these directions are focused on new electronic media or novel computer software that is adapted to educational uses. Saettler's (1968) history of educational media inquiry suggested that this has been a pattern for the past century. Each new medium attracts a group of new researchers to tend to ignore the lessons of past media studies and ask redundant questions. However, productive new areas are developing.

At least three novel or rediscovered research areas are developing as this article is being written. Vigorous research is being carried out in cognitive load theory, in economic studies of media, and in national surveys of educational media use in schools and universities.

3.1 Cognitive Load Theory and Educational Media Design

One group of theory-oriented researchers, stimulated by studies in Australia on cognitive load by Sweller et al. (1998), are examining ways to design educational media presentations so that cognitive overload is avoided. Sweller and his co-workers have developed a method of measuring the amount of cognitive load imposed and mental effort required by an instructional presentation. They have also identified a number of learning problems caused by common educational media design strategies and ways to overcome these problems. One example is the split attention effect. When learner attention is split between different sensory modes or media, Sweller hypothesizes that working memory is often overloaded. Studies by Mousavi et al. (1995) and Mayer (1997) have claimed that presenting novel and difficult concepts to learners in integrated auditory and visual symbolic modes overcomes the split attention effect and results in more learning than information presented in either mode alone for high visual ability and low prior knowledge learners. The theoretical explanation advanced for these findings is that integrated visual and auditory information is more efficiently processed in working memory. The added benefit of integrated auditory explanation and visual depiction of concepts and processes is presumably caused by auditory and visual sensory memory 'buffers' that specialize in storing information that exceeds the limits of working mem-

ory. This activity is hypothesized to extend the representational duration of information to be learned so that elaboration processes have more time to operate and transfer information to long term memory. It is likely that this area will receive much more attention from media researchers since it fits nicely with the current interest in 'multi-media instruction' (see a review and summary by Hedberg et al. 1993). This approach is also one of the features of a new theory of instructional design for teaching complex knowledge.

3.2 New Research on the Economics and Efficiencies of Educational Media

Recently, some of the media-as-vehicle researchers have attempted to encourage studies that examine the 'cognitive efficiency' of instructional presentations that use newer electronic media (see Cobb 1997) at a variety of levels (national, regional, small group, individual) and outcomes (for example, cost per unit of achievement or speed of learning a set of concepts). Cobb suggests that some media and symbolic modes lead to quicker and/or less demanding learning and performance outcomes than other media or symbolic modes for some people. So we have a question with at least two interactive independent variables: the media or representational mode used for presenting an instructional method (for instance, an example presented in either pictorial or verbal modes) and the individual or group differences that would predispose learners to process the method easier and/or faster during learning (for instance, high visual but low verbal ability learners will likely learn faster from pictures than from narrative descriptions of examples). The cost of learning is, after all, one of the most important issues for those concerned with the application of research to solving practical problems.

Cobb illustrates research in cognitive efficiencies by inventing a situation where we want to teach someone to recognize the song of a specific species of bird. He describes different media/modes of presenting bird songs to learners including audio recordings and musical notation. He asks 'How many hours are needed to learn a bird song with a recording vs. with sheet music?'(Cobb 1997, p. 26). He suggests that we show learners the musical notation for the song or play an audio recording of the song. If enough time is allowed for learners and if their motivation is adequate to support their persistence in difficult treatments such as the musical notation of the bird song, then all treatments should eventually produce learning. Yet few media specialists would consider teaching bird songs with musical notation to a majority of learners with adequate hearing and auditory discrimination ability. Is it possible however, that the two modes have different efficiency characteristics for different learners? While the example requires a stretch of the imagination, consider learners who have musical training and auditory discrimination problems. It seems plausible to assume that this small subset of learners might be able to recognize bird songs more efficiently with musical notation than with audio recordings. The work of Levin and co-workers (for example, Levin et al. 1987) has been useful for conceptualizing the various analysis protocols for cost-benefit and cost-effectiveness assessment of media for learning and performance. Levin's (1983) 'replacement method' of determining costs is a very conservative estimate of the economic gains from using media in a school or home setting. Additional work needs to be done in this area to facilitate the measurement of efficiency benefits during learning (for example, speed of learning, lowering the mental effort required to learn).

3.3 National Surveys of Educational Media Use in Schools

A number of governments currently are interested in surveying educational media use in schools and universities. The reason for this interest is, in part, a belief that preparing students for future work and social responsibilities requires skill and knowledge about media. In addition, there is considerable national and international interest in using the World Wide Web and computer-based teaching. Examples of educational surveys in the US have been described by Becker (1999) who has reported statistics on computer use in American schools since 1993 to support national policy decisions on educational media use and research. Among the findings of these surveys is that Internet connections in North American schools have increased from 35 percent in 1993 to over 90 percent in 1999. Becker and his co-workers are curious about the relationship between Internet use and educational philosophy. They have suggested that teachers who have a more constructivist approach (for example, a belief that students construct their own knowledge) use the Internet more in the their teaching than teachers with a traditional instructivist philosophy (for example, that students must be 'taught' knowledge). These studies reflect the educational media as context approach to research.

See also: Communication: Electronic Networks and Publications; Computer-assisted Instruction; Internet: Psychological Perspectives; Lifelong Learning and its Support with New Media: Cultural Concerns; Media and History: Cultural Concerns; Media, Uses of; Technology-supported Learning Environments

Bibliography

Becker H 1999 *Internet Use by Teachers: Conditions of Professional Use and Teacher Directed Student Use, in Teaching Learning and Computing: 1998 National Survey*. Center for Information

Technology and Organizations, University of California, Irvine. Internet access:http://www.crito.uci.edu/TLC/findings/Internet Use/startpage.htm
Clark R E 1994 Media will never influence learning. *Educational Technology Research and Development* **42**(2): 21–30
Cobb T 1997 Cognitive efficiency: Toward a revised theory of Media. *Educational Technology Research and Development* **45**(4): 21–35
De Corte E M C, Mandl L H, Verschaffel L 1992 *Computer-based Learning Environments and Problem Solving*. Springer-Verlag, New York
Hedberg J G, Harper B, Brown C 1993 Reducing cognitive load in multimedia navigation. *Australian Journal of Educational Technology* **9**(2): 157–81. Access on Internet—http://cleo.murdoch.edu.au/gen/aset/ajet/ajet9/su93p157.html
Kiesler S, Kraut R 1999 Internet use and ties that bind. *American Psychologist* **54**(9): 783–4
Kozma R 1994 Will media influence learning: Reframing the debate. *Educational Technology Research and Development* **42**(3): 1–19
Kulik C, Kulik J 1991 Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior* **7**: 15–94
Levin H 1983 *Cost-effectiveness: A Primer*. Sage, Pasadena, CA
Levin H, Glass G, Meister G R 1987 Cost-effectiveness of computer assisted instruction. *Evaluation Review* **11**(1): 50–72
Lumsdaine A 1963 Instruments and media of instruction. In: Gagne E (ed.) *Handbook of Research on Teaching*. Rand McNally, Chicago
Mayer R 1997 Multimedia learning: Are we asking the right questions? *Educational Psychologist* **32**(1): 1–19
Mousavi S, Lowe R, Sweller J 1995 Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology* **87**: 319–34
Saettler P A 1968 *A History of Instructional Technology*. McGraw Hill, New York
Salomon G 1979/1994 *Interaction of Media, Cognition and Learning*. Jossey Bass, San Francisco
Salomon G 1992 Effects with and of computers and the study of computer-based learning environments. In: De Corte E M C, Mandl L H, Verschaffel L (eds.) *Computer-based Learning Environments and Problem Solving*. Springer-Verlag, New York
Salomon G, Clark R E 1977 Re-examining the methodology of research on media and technology in education. *Review of Educational Research* **47**(1): 99–120
Sweller J, Van Merrienboer J, Paas F 1998 Cognitive architecture and instructional design. *Educational Psychology Review* **10**(3): 251–96
Thorndike E L 1912 *Education*. Macmillan, New York

R. E. Clark

Educational Philosophy: Contemporary Views

The issues that are discussed nowadays in the philosophy of education, and the particular ways in which they are dealt with, must be understood as part of the history of the subject. Contrary to the post-Second World War English-speaking scene, where philosophy of education concerned itself primarily, though not exclusively, with analysis and accounts of schooling, the European counterpart occupied itself mainly with problems situated within the broader field of child-rearing including discussions of the concept formation (*Bildung*). Furthermore it is important to realize that continental philosophy of education placed itself firmly within the legacy of the Enlightenment tradition. From this philosophical stance, education was the means to becoming human, i.e., rational: becoming free from one's inclinations and passions, putting oneself under the guidance of reason.

Construing education as a means to realizing one's true nature was not interpreted in a narrow means/end fashion, however, in spite of some claims to that effect. Nor should it be seen as highly individualistic and alienating, for being rational was a state of affairs which was universal and which, being the same for everyone, precluded false consciousness and alienation. The educational implications of this Enlightenment tradition were spelled out mainly by German philosophers of education starting from Kant and Herbart during the eighteenth and nineteenth centuries. Through their work the project of Enlightenment became the project of education. However, a radical pluralism has swept the world and this has been the cause of a crisis in education. But it should be noted that the crisis of education, and the parallel crisis of philosophy of education, are part of a wider crisis of rationality itself. The questions whether reason, and reason alone, can decide what should be done, and if, moreover, rational thinking is even possible at all, are at the heart of the matter. For example, phenomenology, existentialism, neo-Marxism, the Frankfurter Schule (that critiqued the use of a rationality which subsumes the particular under the general because of a will to dominate) and others, have criticized the overambitious Enlightenment project of rationality. Though the critique of reason by reason belongs itself to the tradition of the Enlightenment, the assumption that our enlightened rationality is adequate is, however, called in question—particularly from Nietzsche onwards (cf. Peters et al. 2000, Blake et al. 2000).

Not surprisingly the traditional approach or framework of education has been criticized by twentieth-century philosopher educators. Their starting point in education is from the child-centered movement, or reform-pedagogy. For them child-rearing could no longer be characterized by activities pursued by adults in order to bring children to adulthood. From this position the educator (the parent and/or the teacher) is, first of all, the adviser to the child, and the facilitator of what she or he really wants. It is the child, it is argued, who is from the very beginning responsible for the learning process. Parents should respect their children's values. All of this is of course not unproblematic. It was doubted whether it is possible that

an individual can discover within him or herself what he or she really wants. The closed *cogito* rightly criticized is, so it seems, introduced again here uncritically. Values do not arise privately and individually but have to be thought of—from the beginning—as bestowed upon people by significant others. Furthermore, how could parents and teachers behave as if the values they embrace have the same significance for them as the values their children accept? Finally, it is not clear how parents could avoid initiating their children in the values they live by—must not parents inevitably influence their children in very important ways. In the Anglo-Saxon context, traditionally philosophy had assumed a foundational role for education. The works of the great educators and their employment in educational institutions reflects this belief. Yet the discourse and practices related to this belief were haphazard, discontinuous, and unsystematic. The field was therefore ripe after the Second World War for a coherent and systematic rationalization of its beliefs and practices.

The earlier works by Hardie (1942) and O'Connor (1957), written from a positivistic philosophical position, were not to serve as paradigms for the emerging professionalism of philosophy of education. It was instead Israel Scheffler (1960) and Richard Stanley Peters (1966) who would become the leading authors of the analytic philosophy of education. Their work had its roots in the early work of Moore, Russell, Wittgenstein, Tractatus, Austin, Ryle, and Ayer—clearly not an exhaustive list. Scheffler's *The Language of Education* (1960) was to stimulate considerable discussion on the concept of teaching, Peters' analysis of the concept of education (*Ethics and Education* 1966) was to provide an example of conceptual analysis which was to serve as a paradigm. In practice analytic philosophers of education claimed that they were attempting to clarify the criteria used in the use or application of concepts, by clarifying the rules or conditions under which concepts were used or applied. Borrowing the notion of language as rule governed activity from the work of the later Wittgenstein, though adapted improperly because ultimately they were searching for foundations, and for necessary and sufficient conditions, they adopted the research program of analysis and clarification. First under attack from sociologists of education and then from the general vocational thrust of education in Thatcher's Britain, philosophy of education lost much of its power and influence.

Within North America, while not abandoning the gains made by the analytic approach, not only methodologically but also professionally, attempts were made to broaden the field of philosophy of education, but whereas the *Proceedings of the Philosophy of Education Society* (USA) and *Educational Theory* have a decidedly different feel about them compared to the *British Journal of Philosophy of Education*, there is no one clear paradigm which has emerged in either to replace the analytic approach (see *Educational Theory* 1991, **41**(3) for a summary of the state of the art in North America). If philosophy of education was at its height in 1981 with the publication of the Yearbooks' special issue on philosophy of education, and if the division of chapters in the yearbook—into ethics, epistemology, metaphysics, etc.—revealed that philosophy of education had gone 'professional' by becoming, as editor Jonas Soltis seemed to see it, a collection of subfields of subfields philosophy, then it had also lost touch with educators. By seeking legitimacy in the ideas of philosophy—for example in Rawls, Marx, the phenomenologists and the Frankfurt school—they were no longer communicating with educators. On the contrary it was writers like Illich, Freire, and Bowles and Gintis, who were being discussed in education and not Scheffler, or Soltis, or... More recently it is Apple, Giroux, and McLaren whose works are being widely published and discussed. However not all of this contested 'ground' has been relinquished. For example, philosophers like Harvey Siegel writing on critical thinking, Denis Phillips on objectivity, especially in relation to educational research, and Robin Barrow on curriculum matters (among other things), do have a voice in this conversation on education. But it is clear that the general interest nowadays has shifted to thinkers such as Rorty, Derrida, Lyotard, and Foucault. A strong and sustained critique of foundationalism has emerged in recent years in philosophy, and thus philosophy of education is slowly catching up with some of these developments.

At the different centers in Europe in which philosophy of education is practiced, a varied landscape emerges. Some philosophers of education continue to follow the traditional paths as if nothing has really changed. Within the traditional approach, representatives of several paradigms can be found: phenomenology; existentialism; transcendental (Kantian) pedagogy; critical rationality; *geisteswissenschaftlich*-hermeneutic; and critical-emancipatory. However, in general, the scene is dominated mainly by those who have put these positions in question (see for instance *Zeitschrift für Pädagogik*, 1990, **36**(1)). A theme present in a large number of discussions is the relationship between theory and praxis (see, e.g., Hermann 1987 who reconciles both by the use of common sense). Another approach is to be found in the system-theory of Luhmann and Schorr (1982). Here the basic assumption is that theory cannot formulate any rules which make clear how a practitioner ought to act. Instead educational praxis has to be characterized by the self-sufficiency of the system. There is also the insistence on *Allgemeinbildung* (see Pleines 1987), which could be translated as general formation. The theme of the general is also approached in a different way in Kohlberg's developmental psychology (cf. Oser 1986). The task of the educator comes down to establishing the right conditions in

which the structures can articulate themselves and can change. An important stance which dates back to Horkheimer and Adorno is the critical-emancipatory. There reason signifies the ability to free oneself from a prereflective bond with nature and to be differentiated from it. The capacity of distancing contains the possibility of transforming nature into an object to be dominated. Authors such as Habermas, who is critical of the early starting point of the conscious subject, start from language. There the focus is on the emancipation of the individual in an intersubjectivity of unconstrained agreement with others. The interest in the latter, i.e., in critical theory, is of course not only present in the European context but is there worldwide. So is the interest for Wittgensteinian and postmodern philosophy of education generally. But before turning to these general strands it is important to point out that there are not only schools of thought which can be identified within philosophy of education, but there are also a number of topics which are given particular attention to in the debate nowadays. Among these are: liberalism and communitarianism, citizenship and nationalism, education and the market, multiculturalism, information technology and it goes without saying that pragmatism and feminism (including the ethics of care and its relation to virtue ethics) generate a considerable interest as well (cf. Blake et al. 2001).

From a Wittgensteinian position education can be conceived as a dynamic initiation into a form of life (cf. Smeyers and Marshall 1995). From this position parents are seen as the first educators and from this responsibility, the responsibility of the state concerning schooling can be derived. Parents offer the child the truths by which they live: what moves them, what appeals to them, what supports the idea of human being hoping that she or he will participate. This is an initiation into what is self-evident, or the initiation into the form of life, as Wittgenstein puts it. Insofar as meaning is involved, it will be clear that attention will be given to the particular aspect of a situation. This cannot be separated from the quest for justification. What falls outside one's actions and one's language is not intelligible and thus cannot be justified. Only through educational practice will it become clear what is there for us and what is valuable for us. For Wittgensteinian epistemology and ethics, there is no either/or. It is always language-and-the-world, always the individual-and-the-others. Wittgensteinian philosophy of education will try to make explicit the grammar of the educational language and its presuppositions. It will not offer new facts, nor hypotheses that can be tested. There would be no attempt to offer general justifications through ordinary language, nor to lay out necessary and/or sufficient conditions for the use of certain concepts. What it would offer is a discussion of the nature of an educational question. Here it is fully accepted that there is not one interpretation, but only an interpretation. Besides its task in indicating the incorrectness of certain arguments and the impossibility of accepting certain presuppositions, there is then the explication of reasons that carry one's engagement. Evidently, here philosophy of education cannot remain indifferent to developments within the educational context, but must take a stance (cf. Smeyers and Marshall 1995).

Postmodern philosophy of education on the other hand focuses on another aspect of the present *Zeitgeist*. A profound objection to modernity has always been that the modern technical genius for finding effective means to ends has too much diverted attention from serious consideration of our chosen or implicit ends themselves, whether ethical, economic or, educational. Modernity is instrumentalist. Can objection be heard no less from traditionalists than from the Left or any other quarter. Under the postmodern condition, as Lyotard has described it, the obsession with efficiency and effectiveness that he criticized, and memorably labeled as performativity, has finally parted company altogether from controversial, political questions of what we should be trying to achieve. All kinds of business and activity are measured and ranked against each other, with ever less concern for the rationale for doing so. Thus performativity obscures differences, requiring everything to be commensurable with everything else, so that things can be ranked on the same scale and everyone can be accountable against the same standards. This in turn entails the disvaluing, and perhaps the eradication, of what cannot be ranked. Schools in the United Kingdom, to give one example, are now ranked in gigantic league tables on the basis of the tested achievements of their pupils. Yet while their examination scores are neurotically compared, their social and moral ethos goes by the wayside. Schools and pupils are now required to speak their truth on pain of punishment. This they are to do in the methodical and routine terms of examinations and tests and moreover in the format acceptable to management.

Despite this drive to optimize the system's performance under centralized control, there has also been a countervailing tendency toward dispersal and differentiation. Such variety would, supposedly, empower consumers by giving them a wider range of choice in the marketplace. In itself this apparent contradiction is easily related to historic trends in modernity: to secularization and bureaucratization in the sphere of culture, and rationalization and liberalization in the sphere of society. What is new are the terms in which these processes have come to be understood and celebrated. Power has moved elsewhere, to consumers rather than producers, or at any rate to those (advertising agencies, marketing consultants and facilitators of all kinds) who have the expertise to manipulate consumer choice and sell the educational product just like any other. It is to the consumers that educators must now make confession. Thus confession supplants profession, method replaces thoughtfulness, and pres-

entation skills and image management come before scholarly authority. The family, too, suffers bewilderment in its attempts to educate the young. Like teachers, parents find themselves the objects of suspicion and distrust. An ethos of care sits uneasily with an emphasis on children's rights and, above all, on the individual child's right to choose. The family is scrutinized for its outcomes: how are your chances of good examination results, or your earning potential, affected by being brought up by a single parent? Children in this position may be compensated by being offered, in that most revealing phrase, quality time by the parent who has left, as if this improves the child's opportunity of turning into a quality product (by the standards of quality control). The courses of relevant experts in parenting skills will help obscure the fact that bringing up children is not really regarded as valuable in a society dominated by economic rationality. But the demands of performativity cause more than bewilderment: as Lyotard tells us, they cause terror. We have to be operational or disappear: to fit in with bureaucratic systems, to do the things that can be entered on computer databases, or pay the penalty. In this context one turns to a number of writers and theorists who show a way out of the intellectual paralysis of postmodernity. These writers—chiefly Jacques Derrida, Jacques Lacan, Jean-François Lyotard, and Michel Foucault—have mostly repudiated the title of postmodernists, yet they are clearly postmodernist thinkers in the sense that they help us think about the postmodern condition in which we find ourselves. They have set the stage for the contemporary debate at the time of the turning of the century. As well in the Wittgensteinian philosophy of education as in its postmodern counterpart, the link with education itself has thus been restored (cf. Blake et al. 1998).

See also: Critical Theory: Frankfurt School; Educational Philosophy: Historical Perspectives; Wittgenstein, Ludwig (1889–1951)

Bibliography

Blake N, Smeyers P, Smith R, Standish P 1998 *Thinking Again: Education After Postmodernism*. Bergin & Garvey, New York

Blake N, Smeyers P, Smith R, Standish P 2000 *Education in an Age of Nihilism*. Falmer Press, London

Blake N, Smeyers P, Smith R, Standish P (eds.) 2001 *The Blackwell Guide to the Philosophy of Education*. Blackwell, Oxford, UK

Habermas J 1981 *Theorie des kommunikativen Handelns*. Suhrkamp, Frankfurt am Main, Germany

Hardie C D 1942 *Truth and Fallacy in Educational Theory*. Teachers College Press, Columbia University, New York

Hermann U 1987 *Kontinuität und Traditionsbrücke in der Pädagogik*. Comenius, Ànster

Luhmann N, Schorr K E (eds.) 1982 *Zwischen Technologie und Selbstreferenz: Fragen an die Pädagogik*. Suhrkamp, Frankfurt am Main, Germany

Lyotard J.-F 1984 *The Postmodern Condition. A Report on Knowledge* [Bennington G, Massumi B trans.]. Manchester University Press, Manchester, UK

O'Connor D J 1957 *An Introduction to the Philosophy of Education*. Routledge and Kegan Paul, London

Oser F 1986 *Transformation und Entwicklung: Grundlagen der Moralerziehung*. Suhrkamp, Frankfurt am Main, Germany

Peters M, Marshall J, Smeyers P (eds.) 2000 *Nietzsche's Legacy for Education: Past and Present Values*. Bergin & Garvey, New York

Peters R S 1966 *Ethics and Education*. Allen & Unwin, London

Pleines J E (ed.) 1987 *Das Problem des Allgemeinen in der Bildungstheorie*. Königshausen und Neumann, Würzburg, Germany

Scheffler I 1960 *The Language of Education*. Thomas, Springfield, IL

Smeyers P, Marshall J (eds.) 1995 *Philosophy and Education: Accepting Wittgenstein's Challenge*. Kluwer, Dordrecht, the Netherlands

Wittgenstein L 1953 *Philosophical Investigations/Philosophische Untersuchungen* [Anscombe GEM trans.]. Basil Blackwell, Oxford, UK

P. Smeyers

Educational Philosophy: Historical Perspectives

Philosophical reflection on the nature of education and related issues is about as old as philosophy itself; Plato's *Meno* and *Republic* are not only foundational works of Western philosophy and political theory, but are also the first great classics in philosophy of education. And just as philosophy itself has been an extremely diverse field for the past two and a half millennia, and much of its terrain has been—and still is—bitterly contested, so it is with philosophy of education. To make the field even more complex to describe, during the twentieth century most of the significant intellectual and social trends, and not just the strictly philosophical ones, had an impact upon the work of philosophers of education. Analytical and ordinary language philosophy, and logical positivism, have had adherents in philosophy of education, but so have Marxism, critical theory, phenomenology, hermeneutics, postmodernism, feminist epistemology and social theory, and neoliberalism; furthermore, philosophers of education have often been embroiled in the disputes that rage over such controversial issues as multiculturalism, children's rights, national testing of students and the loss of standards, national curriculum, teacher training and accreditation, and the interface between school and work. There also have been important regional differences in the way in which the field has been conceptualized and institutionalized. This article sketches, necessarily in broad strokes, the development of this lively and diverse domain.

1. Classic Philosophers Writing on Education

Philosophy of education as a recognizable branch of philosophy, or area of specialization within it, is a product of the twentieth century; most commentators would agree that its institutionalized development was relatively insignificant until just before World War II. However, many of the writings of philosophers of the past dealt with educationally relevant issues, although these educational implications were most often left implicit rather than made the object of explicit discussion; direct discussion of education was only sporadic. The nature of knowledge, the nature of human cognition, the rational bases of belief and the causes of unbelief (or of false belief), the key aspects of critical thinking, the formal and informal logical fallacies that plague reasoning in everyday life, the nature of morality and the good life, the importance of personal autonomy, the nature of a just society, the obligations borne by those who aspire to full citizenship in a society, the education of the citizen, the logical features of inquiry in the human and social sciences, are examples of matters that have been grist to the philosophical mill, and which have great relevance to the field of education.

Indeed, it can be argued more strongly that philosophy always serves a vital educational purpose, for, as Amelie Oksenberg Rorty wrote in the introduction to a collection of essays she edited about philosophers of the past who did educationally relevant work: 'Even "pure" philosophy—metaphysics and logic—is implicitly pedagogical. It is meant to correct the myopia of the past and the immediate' (Rorty 1998, p. 1). On this account, then, almost any philosopher's work is also a work in philosophy of education, so it is no surprise that Rorty's volume has discussions of the educational philosophies of—or, more accurately, the educational relevance of the work of—a diverse group including Plato, Aristotle, Augustine, Aquinas, Maimonedes, Rene Descartes, Thomas Hobbes, Spinoza, John Locke, Rousseau and other early Romantics, Kant, Hegel, Mill, the Logical Positivists, and Dewey. Notable omissions are the nineteenth- and twentieth-century figures J. F. Herbart, Herbert Spencer, Bertrand Russell, Wittgenstein, and the two philosophers whose writings most explicitly addressed philosophical issues in education in the second half of the twentieth century—Richard S. Peters and Israel Scheffler. Chambliss, in an essay-length review of the history of philosophy of education, discusses a set of authors that overlaps with the group in the Rorty volume (Chambliss 1996); for a work that attempts to integrate historical and contemporary writing in the field, see Noddings (1995).

1.1 Philosophy vs. 'Cultured Reflection'

An important issue arises here, one that has not always been recognized by writers on the history of philosophy of education: When philosophers of the past—even noteworthy ones such as those listed above—have turned explicitly to address issues concerning education, it has not always been the case that their writings have been philosophical on even a charitable interpretation of this contested term. Some philosophers, to be sure, have developed educational positions that are closely related to, or which are fairly tightly derived from, their philosophical positions; but many others, when they have turned explicitly to education, have engaged in what has been called 'cultured reflection' upon education rather than in closely reasoned philosophical argumentation. The British analytical philosopher D. J. O'Connor noted as far back as 1957 that too often expressions like 'philosophy of education' or 'philosophical reflection on the bases of education' were 'no more than vague though high-sounding titles for miscellaneous talk about the aims and methods of teaching,' and he suggested that these usages 'could well be dropped in the interest of clarity' (cited in Phillips 1985, p. 3863). His suggestion was viewed, in some quarters, as imperious.

A book by John Locke can serve as one example of this nonphilosophical genre; his *Some Thoughts Concerning Education* (1693/1947), which has been reprinted many times and has been cited widely in histories of educational thought, deals with the education of the young gentleman who is likely to inherit a country estate, and it is filled with advice on such matters as hardening the youngster by having him wear leaky shoes in wet weather. There is little (if anything) here that derives from Locke's foundational work in empiricist epistemology or political philosophy; rather, his often interesting educational ideas seem to stem from his wide range of experience as a man of the world. On the other hand, his classic philosophical writings—which are not widely cited in histories of education, and which do not explicitly discuss the field at all—have been a major influence on educational thought; for example through their impact on associationist psychology, the child study movement of the late nineteenth century, and liberal social theory.

In the twentieth century there were many so-called 'philosophers of education' whose work was in this genre of 'cultured reflection' on education rather than being philosophical in any recognizable sense. As the twentieth century progressed there were growing numbers of nonphilosophical writers with strong educational or ideological commitments who self-identified as philosophers of education, and who thus helped to establish something of a gulf between philosophy of education and the parent discipline. (In this context it is interesting to note that Amelie Rorty had no contributor to her volume who is identifiable as a member of the philosophy of education community.) The works of these individuals are nonphilosophical in the sense that they do not contribute to the deeper

development or defense of any philosophical theses, and they would not be useful in courses on the traditional epistemological, metaphysical, logical, or axiological problems of philosophy offered in departments of philosophy.

1.2 Examples of Directly Relevant Philosophical Work

On the other hand, Plato's educational thought was a much more integral part of his philosophical and political thought. In his *Republic*, for example, after developing the metaphysical theory of the underlying realities, the 'forms,' he detailed the educational process that would lead the future philosopher–rulers to gain knowledge of them; their education would culminate in them beholding 'the form of the Good,' in the light of which they would be ably to govern wisely. Plato also discussed the education of the subordinate classes (the auxiliaries and the common citizens) that would allow them to take their allotted places in society—a society marked by complete justice, with everyone serving in the role for which nature had most fitted them. Plato's *Meno* is also regarded as a classic work in philosophy of education, and it has stimulated discussion down to the present; Plato depicts Socrates teaching a slave boy a variant of the Pythagorean theorem, not by direct instruction but by a process of 'Socratic questioning' in which the relevant information is drawn from him—illustrating Plato's theory (closely related to his epistemology) that learning is a process of remembering (for one cannot inquire about something that one does not already know).

Jean-Jacques Rousseau's novel *Emile* (1762/1955) is another work that has a niche in the history of philosophy, and an even more central place in the historical development of philosophy of education. A rather undisciplined thinker whose work is often marked by unresolved contradictions, Rousseau has nevertheless had a major impact on Western thought, in part through his influence on thinkers at the time of the French Revolution; his work on the 'social contract' is still discussed by political philosophers. His novel about the young man, Emile, which traces his life from birth until parenthood, created something of a furor upon its publication, and it was publicly burned in France. The novel—a founding work of modern progressive education—illustrates the philosophical view that all that is natural is good, and all that is 'man-made' (including the products of the civilization of Rousseau's time) is bad; Emile is taken to live in the country where a tutor arranges for him to have a 'natural' education. Emile acquires morality by suffering the natural consequences of his acts; his religious views are 'naturalistic' and are not scripturally based, and are acquired from the 'Savoyard Vicar' whom he meets while on a walk in the countryside. Emile is given very little by way of formal instruction in any topic that he has not developed a natural desire to pursue. (One of the major inconsistencies is apparent here, for—as many commentators have noted—the boy's tutor actually does not leave matters to natural happenstance but is extremely directive, albeit in a subtle way; furthermore it is notable that Rousseau does not provide a coherent account of the concept of 'natural,' which is so central to his line of argument (see Hardie 1962, Rorty 1998). The novel is notorious, too, for its treatment of Sophie, the young woman whom Emile eventually marries; in contrast to the rather progressive education of the latter, Sophie is raised to be 'pleasing' to her future mate. Late-twentieth century feminist writers and others have produced scathing analyses of this aspect of Rousseau's work (e.g., Rorty 1998).

There can be little doubt that, since the days of Plato, the major philosopher to see an intimate relationship between philosophy and education was the American pragmatist John Dewey (1859–1952). In the course of his long and productive philosophical life Dewey published more than 40 books and 800 journal articles; and yet for many years he regarded his *Democracy and Education* (1966; first printed in 1916) as the work that also best summarized his general philosophical position. Indeed, in this book—probably the most famous modern work in philosophy of education—Dewey claimed that philosophy *was* 'the general theory of education'; education, he said, 'is the laboratory in which philosophic distinctions become concrete and are tested' (Dewey 1966, pp. 328, 329). Dewey's views on the intimate relation between activity, and learning and thinking, stemmed from the evolutionary epistemology that he developed after reading the psychological writings of William James, and were directly translated into his recommendations for educational practice. Dewey's views on the classroom as a small community, and on the social functions of education, were part and parcel of his social philosophy that was an interesting amalgam of neoliberalism and communitarian ideas.

2. The Rise of Philosophy of Education as a Specialization

There are a number of widely recognized subdisciplines within the field of philosophy: epistemology, metaphysics, moral philosophy, social and political philosophy, philosophy of science, philosophy of language, logic, philosophy of mathematics, philosophy of law, history of philosophy, and so on. During the early decades of the twentieth century philosophy of education started to emerge as another subfield, although it never won the widespread acceptance accorded to these other specializations within the philosophical community. Indeed, the overwhelming majority of philosophers of education around the

world are members of schools or departments of education or pedagogy rather than being in departments of philosophy; a very small number hold joint appointments.

Despite their relative physical and intellectual isolation from their colleagues in academic philosophy, philosophers of education have interests that span the subdisciplines mentioned above. For example, there are some who pursue epistemological issues as they relate to education—such as the epistemic status of the bodies of subject matter taught in schools, or the epistemological issues that arise in educational research; some—taking a political theory, or neo-Marxist, or even a postmodernist, perspective—are concerned with issues such as what groups in society have the power to decide the structural features of the educational system, or what is taught (or not taught) in schools; others are interested in the philosophical issues surrounding citizenship, and the educational rights of individuals (including parents and children) versus sociocultural groups, in democratic societies—and the related issue of the nature of equality of opportunity; some are concerned with the philosophical issues involved in the field of moral education and moral development; others are concerned with issues that fall into the category of philosophy of psychology, such as the nature of human abilities, the relation between intelligence and creativity, and so forth.

The philosophers of education who pursue such issues are quite interdisciplinary, and follow the relevant literature in educational and psychological research, pedagogy, educational theory and policy, as well as philosophy; they tend to publish in a wide variety of journals, and not just in the philosophical ones. Other philosophers of education, reflecting the fact that the majority of those in the field are actively involved in the preparation of teachers, apply philosophical ideas drawn from a wide variety of sources (plus ideas drawn from other disciplines) to issues of classroom pedagogy and curriculum; there is no clear counterpart to their area of expertise to be found in traditional departments of philosophy. They tend to publish only in educational journals, and their work is unknown to the vast majority of their peers in such departments.

It is difficult to pinpoint the precise date of the emergence of philosophy of education as a field in its own right; the origins seem to lie in the second half of the nineteenth century when many 'normal schools' devoted to the training of teachers were established, and university departments of education or pedagogy were emerging around the world. Influential public intellectuals such as Herbert Spencer, T. H. Huxley, and Cardinal J. H. Newman in England were occasionally writing on education in a philosophical vein—again blurring the boundary between philosophy and philosophy of education, and illustrating the fact that one does not have to be a professional philosopher to sometimes write philosophically interesting material.

The expression 'philosophy of education' first seems to have been used in a systematic way in an encyclopedia produced in 1911–13; but the John Dewey Society was not established until 1935. The Philosophy of Education Society was founded in North America in 1941, and similar societies in the United Kingdom, and Australia and New Zealand, followed some 25 years later (Kaminsky 1996). Philosophical writing about education was not, of course, confined to the English-speaking world, but in Europe the intellectual landscape traditionally was carved up quite differently, leading to a different conception of philosophical reflection on education.

On the Continent there seems to have been a longstanding and stronger appreciation by those scholars working in the field of education that philosophical reflection was a vital part of the 'pedagogical sciences,' and that it was too important to be left to a small number of academic specialists. Thus there has been a lot of philosophical activity, but few have been labeled as doing 'philosophy of education' *per se*—rather, many have been engaged in discussing 'pedagogical theory' or 'pedagogical science.' As the German educationist Wolfgang Brezinka has put it, in the English-speaking world 'educology (or pedagogics) as an autonomous scientific discipline does not exist' (Brezinka 1992, p. 3). He goes on to show that in German-speaking countries, pedagogics as a mixed normative–descriptive discipline is understood to be a theory suitable for 'supplying guidelines for praxis.' Pedagogics should simultaneously 'attain a comprehension of reality and a determination of what should be.' It cannot 'limit itself to studying what is,' but is 'at least in part also a normative science which … develops guiding ideals and measures existing reality against its own claims.' It 'includes normative decisions' and should combine 'the establishment of facts' with the 'critical evaluation of these facts in the service of an obligatory norm' (Brezinka 1992, p. 5).

The Belgian scholar Paul Smeyers also sees marked differences between the way philosophy of education has been pursued in Europe and in the English-speaking world. Especially since World War II, Anglo-Saxon philosophy of education has been primarily—though not exclusively—a philosophy of schooling. Continental philosophy of education, however, concerned itself mainly with problems within the broad field of child-rearing. It is also important to realize that the Continental philosophy of education has to be situated within the legacy of the Enlightenment tradition. In this philosophical stance, education was regarded as the 'means' by which individuals became rational members of the human race (Smeyers 1994, p. 4456).

But it should not be assumed that all European writers adopt the same stance towards philosophy of education; the intranational differences are almost as great as the cross-national ones. For example, in Holland there are several philosophers of education

whose work strikes a resonant note with that of their colleagues in the English-speaking world, and who are recognizably within the analytic tradition that up until recently was dominant in the latter (and indeed they often publish in English, for example van Haaften et al. 1997); and in the USA there are some who work in a manner close to that of many of their colleagues on the Continent.

3. *Some Features of the Modern History of Philosophy of Education*

As indicated above, the field of philosophy of education is extremely diverse, with interesting cross-national differences, and with much variation even within the local scenes in North America, the UK, Australasia, The Netherlands and other nations on the Continent. This makes it impossible to generalize about the development of the field as a whole, but a few historical highlights from the twentieth century can be mentioned as significant.

3.1 *The 'isms' Approach*

For several decades around the middle of the twentieth century it was common to find philosophers of education suggesting that educators should first select a philosophical position that they found appealing, and then 'deduce' the implications for educational practice that followed from this. The 'isms' recommended as possibilities were such positions as idealism, realism, Thomism, existentialism, and pragmatism (see Broudy, in Lucas 1969). This approach was reflected in textbooks, journal articles, conference proceedings, and there was even a series of instructional films; the approach could be found in Europe as well as in the English-speaking world, and philosophers of all stripes seem to have accepted it as a valid approach.

It was some time before it was realized that this approach rested upon a serious philosophical mistake, and even late in the century an occasional textbook was released bearing its stamp. There is not a simple one-to-one correspondence between a person's deep philosophical commitments (such as to idealism or realism) and his or her everyday actions, or actions in the classroom. Idealists and realists, for example, are not offering different empirical theses about the world (e.g., that matter does, or does not, exist), but are giving different philosophical analyses of the concepts of 'matter' and 'existence.' Ludwig Wittgenstein captured the core issue well in his *Zettel* (Wittgenstein 1970, p. 413):

> One man is a convinced realist, another a convinced idealist and teaches his children accordingly … But the idealist will teach his children the word 'chair' after all, for of course he wants them to do this or that, e.g., to fetch a chair. Then where is the difference between what the idealist-educated children say and the realist ones? Won't the difference only be one of battle cry?

Those who advocated the 'isms' approach to philosophy of education were also making use of an over-simple notion of implication; a variety of premises are required to link an abstract position such as realism to the realm of schooling; clearly it is a non-sequitur to argue, as one adherent of the 'isms' approach did in 1957, that 'reality is independent of man's knowledge of it,' therefore 'man must learn impersonal science, and history; and the school plays the role of transmitter of knowledge' (see Phillips 1971, pp. 11–12).

3.2 *The Prominence of Analytic Philosophy of Education*

The 'isms' approach started to decline as the influence of analytic philosophy spread from 'pure' philosophy into philosophy of education during the late 1950s and 1960s. Philosophical analysis—considered broadly as involving careful attention to rooting out vagueness or ambiguity in concepts, to making the use of language more precise, and to scrutinizing trains of argument in order to expose hidden assumptions and logical imprecision—can be traced back at least to the work of Plato. The *Republic* is a book-length attempt to throw light on the notion of justice, and Plato depicts the main character in the book, Socrates, showing up the defects in the arguments and assumptions of those whom he engages in dialog. The tradition never died, but in the work of Plato and subsequent philosophers analysis was mixed together with metaphysical and speculative philosophical activity. During the twentieth century there were notable accomplishments that led to the rise of analytic philosophy as a separate movement in philosophy (for example, the work of Gottlob Frege, Russell, the early Wittgenstein, and the members of the Vienna Circle)—a movement that was for some time quite hostile towards metaphysics. For several decades analysis became the dominant type of philosophical activity in the English-speaking philosophical world, although there certainly were analytic centers in Europe (e.g., the logical positivists centered in Vienna).

Analytic philosophy of education (APE) was chiefly influenced by the so-called 'ordinary language philosophy' of the later Wittgenstein, and Gilbert Ryle's 'philosophical behaviorism' (which also drew from Wittgenstein) was another important influence—an extract from Ryle's *The Concept of Mind* (1949) was reprinted in one of the early compendiums in analytic philosophy of education edited by Israel Scheffler (Scheffler 1958).

As suggested by the titles of many books, the analytic philosophers of education focused much of their attention on the language of education; their work on the concept of 'education' itself can serve as a

typical example of their work. Hirst and Peters (1970), writing separately and together, developed an analysis according to which individuals have been educated insofar as they have acquired a broad cognitive perspective which changes them for the better; educating individuals is equivalent to initiating them into a new and valuable form of life about which they will come to care.

They wrote that if 'this analysis is correct, therefore, teachers who enter the profession because they are concerned about education, would be striving to initiate others into a form of life, which they regard as desirable, in which knowledge and understanding play an important part' (Hirst and Peters 1970, p. 20). Their analysis placed a great deal of emphasis on the normal usage of words; it would be extremely odd, they suggested, if a person claimed to be educated and yet virtually always preferred to play bingo than pursue good literature or study science and the like. Their work, as a result, was later attacked as reflecting a class-based set of attitudes, and it was argued that some classes in British society do not have the same linguistic intuitions as Peters and Hirst; one rebuttal was even titled 'In Defense of Bingo.' (Robin Barrow, in a reassessment of their work, defends Hirst and Peters against the general charge outlined above; see Barrow 1994. See also the debate between Peters, Woods and Dray in Peters 1973.)

Other topics that were widely discussed by analytic philosophers of education in the English-speaking world in the 1960s and 1970s were the (conceptual) differences between teaching and indoctrinating, the conceptual relationship between teaching and learning, the concept of autonomy, the concept of development, the logical structure of the disciplines taught in schools, and the conceptual bases of progressive education.

The criticism concerning reliance on ordinary usage, however, coupled with the related but philosophically more interesting attack on Wittgenstein's ordinary language philosophy by Gellner in his *Words and Things* (1959), and an increasing air of sterility associated with APE, all took their toll. The fall of analytic philosophy of education was also accelerated by the spread of neo-Marxism and critical theory, and then the rise of postmodernism and feminist philosophy, all of which found fertile soil in philosophy of education (Burbules 1994, Kohli 1995). In a relatively short period the center of interest in philosophy of education shifted from concepts and language usage, to the pressing social and political issues (and the associated social injustices) that play out in educational contexts. Many contemporary philosophers of education are socially more activist than the analytic philosophers of education, who are seen (rightly or wrongly) as having been educationally and socially rather conservative. Analytic philosophy of education, then, has only a shadow of its former influence; but, controversially, some would argue that on balance the analytic movement was a positive force and its example is sorely missed in the present climate where debates are marked by a certain lack of rigor (see Barrow 1994; and the debates in Kohli 1995).

3.3 What is 'Philosophy of Education'?

The debates over ordinary language philosophy, sketched above, inspired much discussion about the nature of philosophy of education itself, although philosophers and philosophers of education have long been debating the nature of their discipline. In 1969 Christopher Lucas edited a collection of essays with the revealing title *What is Philosophy of Education?* and debate has continued down to the present—as recently as 1998 the Newsletter of the Philosophy of Education Society of Great Britain published a number of letters from members on the topic 'where should philosophy of education go?'; one of the chief issues was how close the relationship should be between philosophy of education and the general field of philosophy itself.

A related development has been the growing number of encyclopedia articles devoted to philosophy of education. To cite just several of these: Phillips wrote a 16,000 word essay for an international encyclopedia in 1985; the second edition of this work in 1994 contained more than 30 entries covering the discipline and its history, and in 1996 Chambliss edited a one-volume encyclopedia covering the field with more than 200 entries (Chambliss 1996).

See also: Dewey, John (1859–1952); Education: Phenomena, Concepts, and Theories; Educational Philosophy: Contemporary Views; Pestalozzi, Johann Heinrich (1746–1827); Pragmatism: Philosophical Aspects; Rousseau, Jean-Jacques (1712–78)

Bibliography

Barrow R 1994 Philosophy of education: Analytic tradition. In: Husen T, Postlethwaite T N (eds.) *International Encyclopedia of Education*, 2nd edn. Pergamon Press, Oxford,pp. 4442–7

Brezinka W 1992 *Philosophy of Educational Knowledge*. Kluwer, Dordrecht, The Netherlands

Burbules N 1994 Marxism and educational thought. In: Husen T, Postlethwaite T N (eds.) *International Encyclopedia of Education*, 2nd edn. Pergamon, Oxford, pp. 3617–22

Chambliss J (ed.) 1996 *Philosophy of Education: An Encyclopedia*. Garland, New York

Dewey J 1966 *Democracy and Education*. Macmillan, New York

Gellner E 1959 *Words and Things*. Victor Gollancz, London

Hardie C D 1962 *Truth and Fallacy in Educational Theory*. Teachers College Press, New York

Hirst P, Peters R S 1970 *The Logic of Education*. Routledge, London

Kaminsky J 1996 Philosophy of education, professional organizations in. In: Chambliss J (ed.) *Philosophy of Education: An Encyclopedia*. Garland, New York, pp. 475–79

Kohli W (ed.) 1995 *Critical Conversations in Philosophy of Education*. Routledge, New York
Locke J 1693/1947 Some thoughts concerning education. Reprinted in: Penniman H (ed.) *John Locke: On Politics and Education*. Van Nostrand, New York
Lucas C J (ed.) 1969 *What is Philosophy of Education?* Macmillan, London
Noddings N 1995 *Philosophy of Education*. Westview, Boulder, CO
Peters R S (ed.) 1973 *The Philosophy of Education*. Oxford University Press, London
Phillips D C 1971 *Theories, Values and Education*. Melbourne University Press, Melbourne, Australia
Phillips D C 1985 Philosophy of Education. In: Husen T, Postlethwaite T N (eds.) *International Encyclopedia of Education*, 1st edn. Pergamon, Oxford, pp. 3859–77
Rorty A (ed.) 1998 *Philosophers on Education: New Historical Perspectives*. Routledge, New York
Rousseau J J 1762/1955 *Emile*. Dent, London (Trans. B. Foxley)
Ryle G 1949 *The Concept of Mind*. Hutchinson, London
Scheffler I (ed.) 1958 *Philosophy and Education: Modern Readings*. Allyn and Bacon, Boston
Smeyers P 1994 Philosophy of education: Western European perspectives. In: Husen T, Postlethwaite T N (eds.) *International Encyclopedia of Education*, 2nd edn. Pergamon, Oxford, pp. 4456–61
Van Haaften W, Korthals M, Wren T (eds.) 1997 *Philosophy of Development*. Kluwer, Dordrecht, The Netherlands
Wittgenstein L 1970 *Zettel*. University of California Press, Berkeley, CA

D. C. Phillips

Educational Planning Worldwide

This article includes a tentative definition of what educational planning consists of, a brief history of the planning of education systems during the last decades of the twentieth century, an overview of the main evolutions in educational planning in those years, a presentation of the main features of educational planning today, the identification of new areas of concern for the planning community, and suggestions concerning some foreseeable future perspectives in the field of educational planning.

1. *Tentative Definition of Educational Planning*

Planning is the development and statement of clear goals, and the design of efficient and effective strategies to achieve these goals (Davis 1994). It is a process through which scarce resources should be allocated as efficiently as possible, in order to better respond to specific needs, in this case countries' educational needs. These are still huge in a number of countries, in particular developing ones.

Traditionally, educational planners are responsible for the building of new schools and institutions, the repair of existing ones, the forecasting of new enrollments, the allocation and (re)deployment of teachers, the projection of costs and budget estimates, and the planning of didactic materials. In each of these areas, they make thorough diagnoses, identify key problems, assess different policy options and strategies, present them to policy makers for decision, translate them into a number of programs and projects, and then monitor and assess the implementation of the latter. But these tasks have evolved considerably over the past years, and have become more and more complex, due to important evolutions in the general environment in which planners work. These will be dealt with below.

Among the various tasks of educational planners today one could mention (Caillods 1997): (a) participating in discussions about the political, economic, and cultural aspirations and goals of the concerned society that should be properly reflected in the objectives set for the education system; (b) developing, on this basis, an acceptable set of goals and objectives; (c) assessing the present state and past development of school systems, institutions, programs, and projects; (d) assisting, on this basis, in the design of policies, programs, organizational structures, and financial allocations for the next 10 to 20 years; (e) devising mechanisms for resource allocation according to specific criteria in line with policies; (f) identifying appropriate partners; (g) setting up the proper incentives and regulations so that the different actors identified take the most appropriate decisions; (h) monitoring the implementation of the designed policy; (i) ensuring a transparency of the educational markets; and (j) taking appropriate compensatory measures if needed (Ross and Levacic 1999).

As a consequence, for some people, educational planning is a technical exercise, based on the collection of a number of quantitative data, the creation of information systems, and the best use of research results. For others, it is a far more empirical process, involving policy making and implementation (Farrell 1994).

2. *Some Historical Facts*

The principles and methods of educational planning were first broadly developed in the former USSR and after World War II, in socialist states, mainly in Central Europe but also in other parts of the world. The ideological control that characterized those societies led to highly centralized, command-oriented education systems, with considerable attention paid to the control of the curriculum (Farrel 1994).

However, even if industrialized market-economy nations have always shown a strong desire not to be ideologically associated with the concept of planning, many have also participated to a considerable extent in the formulation of policies to regulate and control education. Plans, in that case, tended to be more indicative than directive, more decentralized, and

more incentive-oriented. Centralized educational planning has also played an important role among developing nations, as from the 1960s. The UNESCO International Institute for Educational Planning, created in 1963, has been specifically designated to help these countries develop their educational plans, essentially through the training of central planners. In the 1970s, major donor agencies insisted on educational planning as a precondition for external aid.

As a result of strong government commitment associated with strong social demand for education, in the 1960s and 1970s, developing countries made impressive efforts and progress in education. But in the mid-1970s, and especially in the 1980s, it became clear that, in most cases, planning had not produced the expected results. First, despite the extensive development of education systems, many children were still out of school, and many others had not acquired the basic competencies that they required. Moreover, declining quality, imbalances between supply and demand for educated manpower, and continuing or widening gaps between rural and urban areas could be observed; the absolute numbers of illiterate adults had increased; and a rise in costs and expenditures was noted. Second, it appeared that planners were often incapable of adapting rapidly to changing environments. Reforms and policy changes rarely complied with the calendar set by the plan. Moreover, centralized planning often failed to introduce a reform at local and grassroots levels successfully, largely because of the lack of involvement and empowerment of those directly affected by the process (teachers in particular). Finally, the progressive multiplication of levels of intervention and of actors involved tended to limit the value of any planning exercise conducted by a single actor, even one as predominant as the central government.

All these problems were exacerbated by the financial crisis that hit most nations in the mid-1970s, following the 'oil shock.' The collapse of the Central and Eastern European socialist states, interpreted by many as the failure of centralized and imperative planning, added to the questioning of the merit of central planning (Hallak 1990).

3. *Recent Evolutions*

The lack of confidence in all central planning that resulted from the economic recession—and above all from the questioning of the role of the central government—has nevertheless not led to the disappearance of national plans and programs. Quite the contrary, it is now widely recognized that:

(a) The traditional approaches used by educational planners (projection of a cohort, diagnosis, feasibility studies, etc.) are useful, as long as their range of implementation is limited to sectoral or subsectoral reforms, and to programs or projects affecting one particular level or category of education.

(b) Planning appears to be indispensable at a time when all countries have to face rapid changes, in a general context of globalization resulting from the worldwide integration of economic and financial sectors, and closely linked to the explosion of new information and communication technologies (Hallak 1998).

(c) The spreading of democracy in various regions of the world and the resulting decentralization process has called for more rather than for less planning at local, regional and institutional levels, and has highlighted the need for a new role for central authorities in promoting system-wide efficiency and equity (McGinn 1994).

(d) The challenges to be addressed by educational authorities are still of great importance: in 2000, more than 113 million children had no access to primary education; 875 million adults were illiterate; gender discrimination continued to permeate education systems; youths and adults were denied access to the skills and knowledge necessary for gainful employment and full participation in their societies, etc. (World Education Forum 2000).

A worldwide consensus thus seems to be emerging on the need for participatory planning that can take various forms according to the countries concerned and that should be applied not only at national, but also at regional, local, institutional, schools and program/project level (Govinda 1999).

4. *Main Features of Educational Planning Today*

Observing the worldwide trends in educational planning, one can note that there are still a number of countries where 'traditional planning' is still currently in use and that in a few countries, long-term plans—for 10 years' duration—are being formulated to guide policies and reforms.

However, as a result of the various evolutions described above, educational planning has undergone major changes over the past years, which can be summarized as follows.

4.1 *From Centralized to Decentralized and Participatory Planning*

The call for less central public, and, consequently, for more local public and private education supply, has led to the development of decentralized and participatory approaches in the field of educational planning. Hybrid systems of administration and management have been established, characterized by the devolution of responsibilities to those who are ultimately responsible for and affected by the results of planning, namely schools and various educational institutions (for routine management), local authorities (for experimentation and development), and the private sector (in the case, in particular, of

vocational training programs (Atchoarena and McArdle 1999) and of nonformal educational projects). As a counterbalance, the need to assume new responsibilities at the central level has emerged, in particular as regards the design of a globally coherent framework and the setting up of coordination mechanisms, thereby ensuring that the educational goals established are achieved (Hallak and Poisson 1997).

4.2 From Quantitative to Qualitative Planning

A major issue for educational planners lies in systems' inefficiency, and as a result increased attention has been paid to quality control and regulation. There is now a general feeling that matters of curriculum, pedagogy, educational materials, teacher training, assessment of educational outcomes, and accreditation of standards require much more attention than they are receiving, in particular on the part of planners and experts (IBE 2000). Particular attention should be paid to the four pillars defined by the Delors Commission; namely, 'learning to know,' 'learning to do,' 'learning to live together,' and 'learning to be' (Delors et al. 1996, pp. 85–97). At the same time, there is a call for greater teacher accountability, increased emphasis on outcome rather than input, and more efficient control/regulation systems.

4.3 From Rigid Systems to School-based and Learner-centered Approaches

It is now widely recognized that planning processes should be tailored to the specific characteristics of individual schools, classrooms, and even pupils in order to accommodate the diversity of environments in which learning takes place in a single country. In addition, greater emphasis is needed on the effective implementation of measures taken at the central level. In that respect, school-based and learner-centered approaches are being developed that have important implications for the management of education systems at various levels. The new roles to be played by teachers (more as tutors than as speakers) must be carefully taken into consideration and encouraged, possibly through incentives.

4.4 From Standardized Planning Supply-driven to Planning Though Incentives More Demand-driven

The involvement of new actors in the field of educational management, together with the greater focus on outcomes rather than inputs have led to reflection on ways of stimulating the involvement of all the concerned stakeholders. Incentives have appeared, in this framework, as useful tools to influence the attitudes and involvement of teachers, of communities resistant to education, of pupils, of private actors, of employers, etc. ('incentives' means here 'any reward or sanction, which has, as its effect, the modification of behavior'). These incentives can be simple financial rewards to a particular target group; legal arrangements encouraging or discouraging some initiatives and activities inside or outside the public sector; or any other nonfinancial rewards (such as awards, investitures, medals, affiliation with prestigious associations, etc.). The use of incentives is seen as particularly important in the case of teachers, as a means of stimulating their involvement in their work and their participation in teams, in the adaptation of new contents, and the implementation of new pedagogical methods, as well as in individual experimentation.

Among all the evolutions described, undoubtedly the use of incentives appears to be one of the most important, as it constitutes a way of promoting the desired behavior of key actors of the educational process, and, even more importantly, a means of reconciling regulation by the market with educational development monitoring and planning by the central government.

5. Areas of Concern for the Planning Community

As a consequence of the various evolutions detailed above, several areas of particular concern for educational planners can be distinguished.

5.1 The Need for 'Informed Environments'

Incentive-based planning requires very comprehensive, up-to-date, and operational information systems. Education Management Information Systems (EMIS), more and more associated with Geographical Information Systems (GIS)—which combine statistical and spatial analysis—should be modified in this sense. They should thus go on providing data used to secure or to allocate resources, as well as information about the rewards and penalties available to decision makers to encourage local authorities, institutions, and individuals to perform their tasks. They should also provide information about forms of alternative education and their costs and probable effects. The decentralization of data use, linked with the decentralization of authority and of incentives, constitutes a key issue to be addressed (Kemmerer and Windham 1997).

5.2 The Shift from Planning Toward Management

The design of adequate information models, the establishment of flexible budgets, the handling of qualitative aspects of education, the development of individual decision making through the use of incentives, the creation of proper certification and assessment mechanisms, etc. all imply a shift from a planning approach toward a management approach. This trend calls for a diversification within the plan-

ning profession and the involvement of new categories of experts (Caillods 1991). Additional disciplines, such as law, geography, etc., should be taken into account in strategic planning in order to enrich the planning methods. Training in planning should also target new categories, such as local administrators (in particular at city level), members of nongovernmental organizations, entrepreneurs, journalists, etc.

5.3 Renewed Tools for Educational Financing

The decentralization of the management of education systems necessitates cost-sharing schemes between both central and local authorities (which may mean drastic savings in salary costs for central administration), and public and private authorities. Fiscal reforms, cost-recovery mechanisms, etc. may also be required. In addition, grants-in-aid may be needed, in particular to support disadvantaged youth (among them, the disabled). Finally, learner-centered approaches could lead to an educational voucher scheme, consisting of allocating a capital of time, in other words, 'a study-time entitlement,' that each individual should be allowed to manage freely.

5.4 New Opportunities and Mechanisms for Policy Dialog

There is a crucial need for proper mechanisms of consultation and policy dialog in order to make possible the active participation of the main partners in the educational development; namely, the education profession, teachers' unions, families, communities, the private sector, the mass media, and various segments of the public sector (authorities responsible for health, for instance). This dialog should be organized in such a way that it is independent of party politics and the government of the day. Otherwise, if there is a frequent change of government policy, nothing long-term and solid will be able to take place. This may require certain social and political conditions in the country concerned.

5.5 The Fundamental Roles Played by Research and Experimentation

The provided adaptation of educational planning to today's needs should be based on the results provided by appropriate research (forecasting and scenario models, microeconomic studies, determinants of school achievement, methods of public-sector management, mechanisms of assessment and certification, etc.).

6. Future Perspectives

The World Education Forum, organized in Dakar in April 2000, provided the opportunity to assess the achievements, lessons, and failures of education systems since 1990. It made use of the results of the Education For All 2000 Assessment, which included national assessments of the progress achieved since Jomtien in 183 countries. Its *Framework for Action* (World Education Forum 2000) provided broad orientations for future activities of all actors in the field of education for the ensuing years, in particular planners.

Among the main goals laid down in the *Framework* are: (a) expanding and improving comprehensive early childhood care and education, especially for the most vulnerable and disadvantaged children; (b) ensuring that by 2015 all children, particularly girls, children in difficult circumstances, and those belonging to ethnic minorities, have access to and complete free and compulsory primary education of good quality; (c) ensuring that the learning needs of all young people and adults are met through equitable access to appropriate learning and life-skills programs; (d) achieving a 50 percent improvement in levels of adult literacy by 2015, especially for women, and equitable access to basic and continuing education for all adults; (e) eliminating gender disparities in primary and secondary education by 2005, and achieving gender equality in education by 2015, with a focus on ensuring girls' full and equal access to and achievement in basic education of good quality; and (f) improving all aspects of the quality of education and ensuring the excellence of all, so that recognized and measurable learning outcomes are achieved by all, especially in literacy, numeracy, and essential life skills.

To reach all of the above-mentioned goals, participating stakeholders suggested that national plans for action in education be set up or developed. Their preparation should imply a highly participatory process taking place at country level, within a well-integrated sector framework clearly linked to poverty elimination and development strategies. The national action plans of Education for All may well constitute a new starting point for planning worldwide.

See also: Cultural Diversity, Human Development, and Education; Education: Phenomena, Concepts, and Theories; Educational Innovation, Management of; Educational Institutions and Society; Educational Policy: Comparative Perspective; Educational Policy, Mechanisms for the Development of; Educational Research and School Reform; Educational Systems: Asia; Educational Systems: Australia; Educational Systems: Europe; Educational Systems: North America; Progressive Education Internationally; Soviet Studies: Education

Bibliography

Atchoarena A, McArdle T 1999 *Reforming Training Governance: The Jamaican Experience*. UNESCO, Paris

Caillods F 1991 *Educational Planning for the Year 2000*. IIEP, Paris
Caillods F 1997 The new educational environment: Planning for participation. In: Windham D, Kemmerer F (eds.) *Incentives Analysis and Individual Decision-making in the Planning of Education*. UNESCO, Paris
Davis R G 1994 Educational planning: Models and methods. In: Postlethwaite T N, Husen T (eds.) *The International Encyclopaedia of Education*, 2nd edn. Pergamon, Oxford, UK
Delors J, Al Mufti I, Amagi I, Carneiro R, Chung F, Geremek B, Gorham W, Kornhauser A, Manley M, Padrón Quero M, Savané M-A, Singh K, Stavenhagen R, Won Suhr M, Nanzhao Z 1996 *Learning: The Treasure Within*. UNESCO, Paris
Farrel J P 1994 Planning education: Overview. In: Postlethwaite T N, Husen T (eds.) *The International Encyclopaedia of Education*, 2nd edn. Pergamon, Oxford, UK
Govinda R 1999 *Reaching the Unreached Through Participatory Planning: School Mapping in Lok Jumbish, India. School Mapping and Local-level Planning*. UNESCO, Paris
Hallak J 1990 *Investing in the Future: Setting Educational Priorities in the Developing World*. Pergamon, Oxford, UK
Hallak J 1998 *Education and Globalisation*. UNESCO, Paris
Hallak J, Poisson M 1997 *Public Authorities: Renewing Legitimacy*. UNESCO, Paris
International Bureau of Education 2000 *Globalization and Living Together: The Challenges for Educational Content in Asia*. IBE, Geneva
Kemmerer F N Windham D M 1997 Linking information to incentives. In: *Incentives Analysis and Individual Decision-making in the Planning of Education*. UNESCO, Paris
McGinn N F 1994 Politics of Educational Planning. In: Postlethwaite T N, Husen T (eds.) *The International Encyclopaedia of Education*, 2nd edn. Pergamon, Oxford, UK
Ross K, Levacic R (eds.) 1999 *Needs-based Resource Allocation in Education via Formula Funding of Schools*. UNESCO, Paris
World Education Forum 2000 *The Dakar Framework for Action. Education For All: Meeting our Collective Commitments*. UNESCO, Paris

J. Hallak

Educational Policy: Comparative Perspective

Who and what are the key actors and institutions in education policy and how do they vary among countries? A difficult question is made more complicated by the fact that many languages lack a distinct equivalent to the English term 'policy,' whereas French and German have several terms to cover the concept of 'education.' But although politically developed countries vary in the degree to which parents, teachers, and private agencies share in influencing education policies, there are two sets of institution that are ubiquitous in this arena. These are the tiers in governmental authority and the levels of educational provision, ranging from preschool to university level education.

1. Centralized and Decentralized Systems

Whether the relationships between governmental tiers and the hierarchies of educational institutions were in the foreground of academic attention varied among the earlier generation of scholars of comparative education. But one can discern some conceptual continuity between the classification of education along a centralized–decentralized continuum systems by writers like Isaac Leon Kandel, and the analytic schemas utilized by influential comparative education analysts of the 1970s.

For Margaret Archer, political centralization was the most important variable determining paths of educational system development, which she defined as 'a nation-wide and differentiated collection of institutions devoted to formal education, whose overall control and supervision is at least partly governmental and whose component parts and processes are related to one another' (Archer 1979, p.54).

In her historical analysis, Archer compared two centralized unitary European systems, those of France and Russia, with the decentralized unitary systems of the UK and Denmark. The process through which education became anchored in the public sphere in centralized political systems like those of France reflected different policy priorities from those in decentralized areas. As monarchs and their bureaucracies developed earlier controls over national and local administrations they sought to use these to lay down uniform rules regarding school establishment. Where states were able to go beyond regulation to direct operation of schools they aimed initially for unification, and subsequently for systematization of school systems.

In terms of the tiers of governmental organization, some less centralized political systems were classified as federal ones, especially where strong state autonomy is grounded in greater longevity than the federation, as in the case of Germany and especially Switzerland. During the Napoleonic period, the French and their Swiss sympathizers attempted to impose a centralized republican regime, but widespread resistance forced a return to a confederal system. Reactions against this led, after 1848, to the founding of some higher education institutions by the Swiss federal government, a development which states rights advocates kept from occurring in the USA. For pre-1933 Germany, the relationship between governmental tiers and education was influenced strongly by the varying role of Prussia, which came to set standards for the other states of the 'German Reich.'

An opportunity for augmenting the traditional centralization–decentralization schema is provided by the fact that most technical and lay descriptions distinguish between three tiers of government—local, regional, and national, as well as three levels of education—primary, secondary, and tertiary. Much can be gleaned by seeking to distinguish among

interactions and control patterns of different intensity within the matrix constituted by the two sets of institutions.

Other than the 'top-down' and 'bottom-up' of classifications, the analytical schema based on morphological symmetry distinguishes differing degrees of structural affinity mainly by assessing the strength of horizontal linkages. The more symmetrical the financial and administrative interaction between analogous levels, such as between secondary education and regional government, the higher the degree of *isomorphism*. The less symmetrical the correspondence of financial and administrative interaction between the analogous structures, as when local governments are shut out from running or financing primary schools, the greater the tendency to be *anisomorphic* (or unisomorphic). Relational sets that are neither very symmetrical or asymmetrical can be conceptualized as being *heteromorphic*.

2. Isomorphic Relations and Policy Emphasis

At first glance we would expect greater similarity between the German and Swiss systems, since in these federal systems the role of Swiss 'cantons' and German states generate more autonomous policy-making power than we can expect to find in the unitary and centralized Japan. However, this expectation is not borne out upon closer analysis. For, 'it is overly ambitious to attempt universalistic expectations, across time and place, about how latent political motives drive decentralization policies' (Lauglo 1996, p. 22).

The Swiss pattern differs most significantly from the German one with regard to the maintenance of the communities' responsibility for operating the primary schools. Germany is one of numerous European systems that operate without local school boards or significant community influence on education policy. By contrast the Swiss communities have successfully defended these powers and retained the dominant role in financing primary schools, as Table 1 indicates. They have sustained their own identities by resisting the trend toward merger of small local governments, which in the course of the 1970s led to the disappearance through amalgamation of some two-thirds of German communities.

The role of the local Japanese school boards is not nearly as strong since their members are appointed and not elected, and since their power to affect hiring and retention decisions regarding teachers and principals is much more limited than in the Swiss case. Isomorphism is strengthened through enhanced regional responsibility for most secondary schools. This is lodged with cantons in Switzerland and mainly with the prefectures in Japan, but the autonomy of the latter with regard to areas like curriculum and personnel is much more limited, hence the heteromorphic classification.

Morphic affinities are less strongly manifested with regard to higher education, since in both these countries the national government agencies operate only one sector of the higher education institutions. Nevertheless, the fact that the Japanese Education Ministry, *Monbusho*, is in charge of the 96 national universities, provides some parallel to the Swiss case of national responsibility for the two technical institutes. The proportion of all higher education students in these institutions, 21 percent and 18 percent respectively, is very similar, whereas the German federal government operates no civilian higher education institutions.

2.1 The Anisomorphic German Case

The German case is distinguished from the other two insofar as the bulk of both operating and financing responsibility for all three levels of education is predominantly centered in one level of government, that of the state. Thus teacher appointments for both primary and secondary schools are centered in the state ministries to a degree not paralleled, especially for primary schools, in Japan or Switzerland. Since the appointment and financing responsibility for higher education are also lodged at this level, Germany constitutes a case that departs most strongly from the isomorphic pattern, since the inequality among the governmental tiers in education policy is so accentuated by the concentration of power in the states. The following attempts to explain how derivatives of party and bureaucratic influence and power positions can help to make the rationale for this placement more apparent.

One cross-national case of contrasting structural relationships can be found in the continued existence of small schools in which teachers have to instruct pupils from several grades in the same classroom. Such schools were gradually phased out in Germany in the course of communal amalgamation in the 1970s, whereas in typical Swiss mountain communities, quite a number of them were still operating in the 1990s.

In the anisomorphic German setting, the dominant state governments can impose strong uniform standards within their boundaries. But they can also pursue policies that differ somewhat from those pursued by the federal government, as well as by other states. Party majorities contrasting with the one in power nationally pursue different upper secondary school policies in both Japanese prefectures and German states. But the contrasting pattern was sustained more strongly in Germany, partly because the financing and credentialing powers of the national tier were lower than in the Japanese case (Heidenheimer 1997).

With regard to financing sources for public education, Switzerland is more like the USA and less like Germany. In the first two countries the states and

Table 1
Isomorphism and public educational funding, 1993

		Sources of Public Funds %		
		Local	Regional	National
Germany	Primary & Secondary	19	80	1
(Anisomorphic)	Tertiary	1	83	16
Japan	Primary & Secondary	0	75	25
(Heteromorphic)	Tertiary	0	14	86
Sweden	Primary & Secondary			
(In transition)	Tertiary	1	3	96
USA	Primary & Secondary	44	48	8
(Quasi-isomorphic)	Tertiary	6	58	36
Switzerland	Primary & Secondary	44	52	4
(Isomorphic)	Tertiary	1	55	44

Source: OECD 1995, pp. 117–18

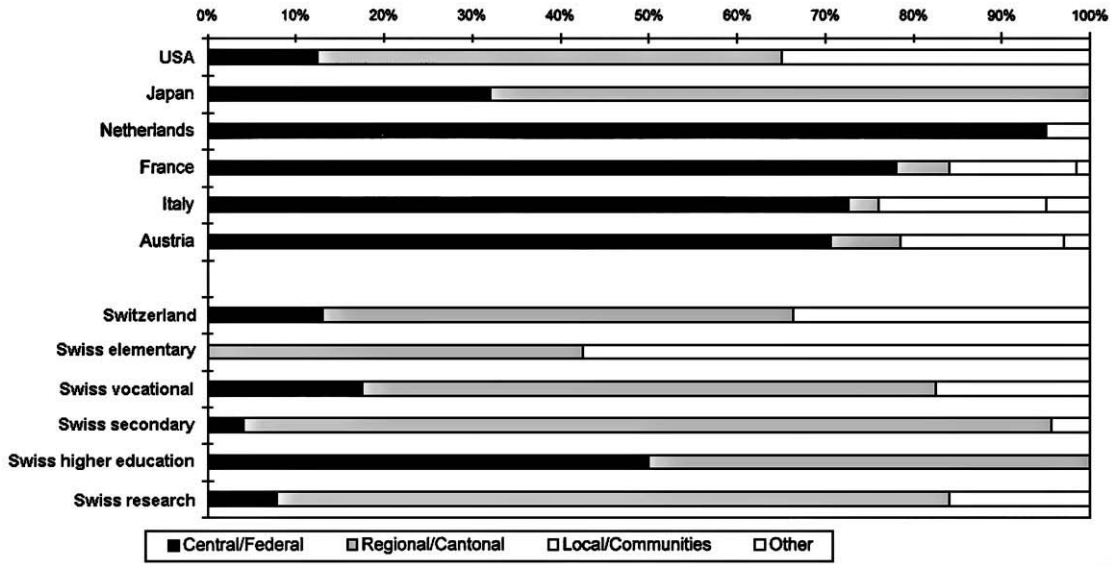

Figure 1
Public education expenditures by level of government, 1988: Swiss isomorphism in contrast

cantons bear around half of the costs, the communities about a third, and the federal government only between 10 and 15 percent. Both are thus examples of bottom-heavy financing backing up relatively bottom-heavy decision making. But the Swiss pattern approaches the isomorphic model; as Fig. 1 shows, communities bear the majority of costs for primary schools, the cantons bear most of the financing of secondary and vocational education, and the federal government bears half the cost of higher education and almost all the burden of research financing. The US distribution is very similar, but there the national government does not operate any tertiary institutions. Hence we classify it as quasi-isomorphic.

What is implied by regarding Germany, a federal country, as exemplifying an anisomorphic model, whereas the highly centralized Japanese unitary system is classified in the intermediate, or heteromorphic, category. Does this essentially lead to the hypothesis that Mombusho, the powerful national ministry, is on balance, faced by a greater series of constraints, than are the collective powers of the state education ministers, and their conference of German school ministers?

Whereas the weaker German and Swiss national education agencies have to share jurisdiction for the vocational dual sector with economic ministries and business interests, the Japanese ministry has not had to interact as directly with a bureaucratic rival. In contrast to other policy sectors that are of concern to several ministries and perhaps also foreign governments, the Japanese education sector was long unique for remaining predominantly subject to the influence of only one ministry and one party. Whereas a greater range of bureaucratic and political interests were at play in Japanese policy areas like agriculture and environment, education policy was uniquely 'locked into the status quo' because issues in this sector were 'purely domestic concerns under the jurisdiction of a single ministry' (Schoppa 1993, p. 260).

On the other hand some expert observers assert that the dominant stereotype of Japanese education conveys a false picture about the uniformity of the postwar education experience. It focuses attention on one level—that of academic secondary education—of a multitiered institution whose other levels can be understood to be organized along often different principles (Kelly 1993, p. 207).

In the longer run Mombusho proved to be more successful in overcoming opposition from the teachers' union, and its allies on the level of secondary schooling, than in implementing its objectives for the tertiary level. This can be attributed to the greater strength of top-down bureaucratic influence, from the ministry to the prefectures, than the potential of bottom-up opposition party influence. The attempts of Socialist prefectural party leaders to oppose policies like those leading to prestige stratification of high schools were undercut by the way that prefectural bureaucrats followed the Mombusho line. Even though the Japanese prefectures have considerable potential education policy autonomy, they have seldom made significant use of it. It is telling that the Japanese Prefectural Superintendents Association has been characterized as an example of an 'incorporated interest' (Schoppa 1991, p. 146).

The reason why Germany and Switzerland constitute (an)isomorphic countermodels becomes clearer when one considers how decision-making powers reflect the way educational finances are shared by the three levels of government. Take the influence of Swiss communities over primary schools: They are much more powerful because of the way the selection of school personnel and other key powers are connected to a strong communal financing role, which is twice as large as in Germany. Also Swiss direct democracy has served to guarantee communal viability by providing veto protection against unwelcomed amalgamation from above.

While the similarity of the US and Swiss profiles in Table 1 is striking, there are two reasons for classifying the USA as only a quasi-isomorphic system. One is that like Germany, but unlike Switzerland, the US federal government does not administer any significant set of civilian higher education institutions in the way that the Swiss operate the Swiss technical institutes. The second is that the USA has a much larger set of private higher and secondary institutions, which has an indirect effect on the intensity of relations between government tiers and educational levels. Thus it contributes to promoting greater competition among US than Swiss universities. This would also hold for the contract between German and more mixed public/private Japanese universities.

By applying an 'aquatic' lens, one can discern the configurations through which swarms of adolescents are directed through gradually shifting channels in river beds whose capacity for carrying traffic loads undergo subsurface. Focusing this aquatic lens on the Swiss flow, one can see how size and political topography cause many streams there to be both smaller and slower. Mainly because the injunctions of referenda democracy safeguarded the powers of small and medium-sized communities and cantons, sandbars and dams kept the streams from changing or intermingling as much. Referenda democracy caused Swiss national interest groups to be superior to political parties for coordinating localized channel maintenance efforts (Heidenheimer 1997, p. 133).

3. *Tiers of Decision Making*

Further insights become apparent when examining how countries compare in the allocation of decision-making powers for lower secondary education as between four levels of governance, those of the school, local school boards, and regional and national governments. Given its propensity toward 'bottom-up' and direct democracy, we might expect Swiss schools to enjoy more autonomy than those elsewhere. But in fact it was found in a 1993 OECD level of decision-making study that the Swiss came in last among 14 countries, four places behind Germany, in the degree to which lower secondary teachers and parents at the school level could share in decision making (OECD 1993, p.136). This finding makes the surprising lack of Swiss school principals somewhat more plausible, since it bears out that local and cantonal school boards preempt schools' decision-making powers even more than in other countries with occasionally intrusive local school boards, like the USA.

The essential element that has kept the Swiss system bottom-heavy, and maintained the decision-making powers of the communes, is the ability of citizens to challenge both legislative and executive decisions through referendums. The decision-making powers of the parliaments and parties are limited because parliaments can only produce proposals for decisions which must expect to be submitted to plebiscitary finalization, creating an additional decision level,

which increases greatly decision-making costs. The fact that referenda proposals from all three levels of government are simultaneously submitted to voters about four times a year contributes crucially to keeping the Swiss education system close to the isomorphic model. Otherwise the municipalities' autonomy and influence might have declined somewhat in the way they did in both Germany and Japan.

Another indicator of the great state and cantonal level predominance of German and Swiss education can be deduced from the way in which the terms of office of top political and bureaucratic officeholders do or do not parallel each other. The average tenure of German state education ministers, five years, is about equal to the period of office of their top administrators, the state secretaries. When a Zürich education minister left office in 1995 he shared a platform with six rectors of Zürich University who had succeeded each other during his 24 years in office. Several Swiss cantonal education directors served terms during which as many as a dozen Japanese education ministers came and went. In Japan the average tenure of national education ministers is only about one or two years, or much shorter than the terms of office of their senior civil servants, or of many university presidents.

4. *Isomorphism and Equality Goals: Switzerland and Sweden*

An advantage of centralized educational systems has often been held to be that the greater uniformity of resources and curriculums which they can offer tend to be associated with reducing inequalities of achievement, especially those based on social class. Indeed comparative studies have shown centralized systems like those of Japan and Sweden as demonstrating decreasing dependence of educational achievement on social origins different from most other countries. (Shavit and Blossfeld 1993, p.112) Comparative research has shown that 'real and fairly substantial' equalization of education opportunity took place in Sweden in recent decades. The lower level of education inequality compared with Germany is attributed partly to the greater school continuation encouraged by comprehensive schools which were uniformly adopted in Sweden from the 1960s. These were only partially implemented in some West German states so that school continuation continued to be ' heavily dependent on grades and teachers assigned evaluations' (Eriksson and Jonsson 1996, pp. 91, 199–204).

After having imposed numerous education reforms under a strong national board of education, Sweden in the 1970s started to move gradually toward the Swiss modal structure. Both national and regional education boards were abolished, and much of their power was transferred to local government, which in turn empowered their schools. Teachers who had been employees of the state and counties came to be hired by municipalities. These reforms were pushed through by the Social Democrats, who had previously been so committed to supporting equality goals through centralism, over the opposition of the bourgeois parties and the teacher's union. One of the new goals was to allow individual schools to profile themselves as to content and ways of working. This led to uncertainty, complexity, and often changing divisions of labor.

What effect this tendency had on equality goals in education has come to be strongly contested. Some have continued to argue that 'proposals for decentralization of education acts against democracy' since 'transferring authority to local communities or individuals acts to reduce the achievements of educational outcomes that favor the whole society.' But some Swedes argue that they are observing a redefinition of policy norms, with 'a new type of egalitarianism' emerging as part of a more 'rational' long-term agenda. In education this trend moves away from the standardized system which had sought to facilitate the 'less privileged' to enter the more prestigious school tracks. Now an upgrading of vocational programs is held to exemplify a greater recognition of diversity so as to stress 'equally valued' service rather than 'equal provision' (Gustafsson and Lidström 1996, p. 156).

The Swedish move toward isomorphism illustrates how curricular contents and decision making at the level of the school can be linked dynamically to changes in the relationship between governmental tiers and educational levels. Gustafsson and Lindström claim that these changes have brought about 'a new balance' between the tiers of government, as well as different educational and other ingredients in the 'emerging Swedish welfare definitions,' which takes Sweden's new position in the European context into account.

5. *Educational Polices and National Choices: Some Illustrative Extrapolations*

5.1 *Japanese Uniqueness*

What makes Japan appear so unique as an 'educational society par excellence' may not be due to a uniform tendency expressed in how steep all Japanese educational ladders seem to those who climb them. Rather it may be concluded that the Japanese heteromorphic configuration had helped produce somewhat contrasting sets of ladders at the various levels of education. At the lower secondary school level the prevalent type seems to be that of the boarding ladder, designed for facilitating entry into new realms of knowledge and cooperative learning. At the upper secondary level the students are trained to perform

more on hook ladders, that become attached to the entrance ports of various groups of universities. Finally at the university or tertiary level it is the companion ladder, ensconced within a secure and protected setting, that constitutes the predominant model.

The synecdochal streak in the heteromorphic Japanese case is also anchored to the economics of the tertiary sector. Relatively low public expenditures there had been made feasible by leaving the bulk of tertiary expansion to be handled by the private sector. Thus private universities came to enroll the majority of students, in contrast to the minority enrolled in private US institutions.

5.2 American Test Rankings

The question as to whether, at the start of the new millennium, the USA's strategic position as the 'sole remaining superpower' was matched by rankings in education would probably elicit very divergent responses. Typical ones from those focused on the university sector would probably be 'yes' or 'maybe', but there would be little chance of getting such a reply from those focused on the secondary school sector. The story behind this situation, as in Germany, involves crossed wires between the federal government and the largest member state.

Until the 1970s public education in California had stood out on both quality and quantity. But this changed drastically, largely as a result of a 1978 referendum capping tax revenues for education. From this followed a decline of quality thresholds of the primary and secondary levels, which in turn fed into the poor performance of US high school students on the international achievement tests which gained visibility in the 1980s. To meet demands for leadership on this front, President George Bush and the state governors led by Arkansas governor Bill Clinton met at a highly publicized conference in 1989 and adopted the goal of getting US rankings in these subjects to the international peak by the end of the century. A decade later 49 states had indeed adopted some sort of standards which schools and school districts were to be evaluated on, but the goal of moving US achievements significantly upwards in international rankings had made scarcely any progress.

5.3 Small Country Choices

The advantages and disadvantages of strong isomorphic attachments are strikingly illustrated by the way in which education changes have occurred in relation to the potential entry into the European Union of two rich small countries like Sweden and Switzerland. For the Swedes an adaptation of their dominant equality goals became accepted as part of the cost of entry into Europe. For the Swiss an unwillingness to tamper with the sacrosanct powers of referenda democracy and local self-government have gone hand-in-hand with the failure of campaigns to make European Union membership attractive for the Swiss voters in the way that they had in Sweden and Austria (Kobach 1997).

5.4 German Cultural Status

In the European context Germany's anisomorphic inheritance has engendered what might be called a costly case of synecdoche, i.e., where the part is bigger than the whole. Germany's largest state, Northrhine Westphalia, not only spends more on education than does the federal government. But due to the Social Democrats remaining in continuous control, its ministers also have a dominant voice in German education, especially higher education, with its great dependence on federal funding. This has permitted them to veto proposals to introduce student fees as a significant source of funding, in the way that this had been pushed through under Tony Blair in the UK. The continued financial plight of German universities contributed to a decline in their attraction for foreign students. In turn this and similar factors contributed to a declining cultural weight of the enlarged post-unification Germany—as manifested, for instance, in the failure of attempts to upgrade German as an official language within the European Union.

6. Conclusion

The fruits of this analysis can perhaps serve as a basis toward efforts to seek to identify how governmental hierarchies affect the character of the ladders that link the various educational levels. The article has tried to show how isomorphic or nonisomorphic system configurations help to determine whether and to what extent the presumed 'imperatives' of centralization or decentralization are actually borne out, and sustained over time. Given a certain configuration, institutional outcomes may come out differently from what a path deterministic model would anticipate. In this manner it can be understood why a more decentralized Swiss political system has led to a larger national government role in operating higher education than in Germany; or why the more centralized Japanese higher education system encompasses a larger private sector than is found in these and other European countries.

See also: Comparative Research in Education: IEA Studies; Cross-cultural Study of Education; Education: Phenomena, Concepts, and Theories; Edu-

cational Institutions and Society; Educational Institutions, History of; Educational Planning Worldwide; Educational Systems: Europe; Educational Systems: North America

Bibliography

Archer M S 1979 *Social Origins of Educational Systems*. Sage, London

Bardetscher H (ed.) 1997 *Die schweizerische Konferenz der kantonalen Erziehungsdirektoren 1897 bis 1997*. Verlag Haupt, Bern, Germany

Chapman J D, Boyd W, Lancer R, Reynolds D (eds.) 1996 *The Reconstruction of Education: Quality, Equality and Control*. Cassell, London

Clark B R (ed.) 1985 *The School and the University*. University of California Press, Berkeley, CA

Eckstein M, Noah H J 1993 *Secondary School Examinations: International Perspectives on Policies and Practice*. Yale University Press, New Haven, CT

Erikson R Jonsson J O (eds.) 1996 *Can Education Be Equalized: The Swedish Case in Comparative Perspective*. Westview, Boulder, CO

Gustafsson G, Lidström A 1996 Redefining the concept of educational equality through decentralization. In: Chapman J D et al. (eds.) *The Reconstruction of Education:Quality, Equality and Control*. Cassell, London, pp. 149–59

Hega G 1999 *Consensus Democracy: Swiss Education Policy Between Federalism and Subsidiarity*. Peter Lang, New York

Heidenheimer A J 1977 Achieving equality through educational expansion. Problems in the Swedish experience. *Comparative Political Studies* **10**: 413–32

Heidenheimer A J 1986 Politics, policy and policey as concepts in English and Continental languages: An attempt to explain divergences. *Review of Politics* **1**: 3–29

Heidenheimer A J 1997 *Disparate Ladders: Why School and University Policies differ in Germany, Japan and Switzerland*. Transaction, New Brunswick, NJ

Husén T 1979 *The School In Question: A Comparative Study of the School and its Future in Western Societies*. Oxford University Press, New York

Husén T Kogan M (eds.) 1984 *Educational Research and Policy: How Do They Relate?* Pergamon Press, New York

Kandel I L 1933 *Comparative Education*. Houghton Mifflin, New York

Kelly W 1993 Finding a place in metropolitan Japan. In: Gordon A (eds.) *Postwar Japan in History*. University of California Press, Berkeley, CA, pp. 189–238

Kobach K W 1997 Spur thy neighbor: Direct democracy and Swiss isolationism. *West European Politics* **3**: 185–211

Lauglo J 1996 Forms of decentralization and their implications for education. In: Chapman J D et al. (eds.) *The Reconstruction of Education:Quality, Equality and Control*. Cassell, London, pp. 17–46

Müller W, Karle W 1993 Social selection in education systems in Europe. *European Sociological Review* **9**: 1–23

Neidhart L 1987 Regierungs- und Verwaltungssystem der Schweiz und Bundesrepublik Deutschland. In: Windhoff-Heritier A (ed.) *Verwaltung und ihre Umwelt: Festschrift für Thomas Ellwein*. Westdeutscher, Germany, pp. 170–93

Organization for Economic Cooperation and Development (OECD) 1995 *Education at a Glance: OECD Indicators*. OECD, Paris

Page E 1991 *Localism and Centralism in Europe*. Oxford University Press, Oxford, UK

Schriewer J 1999 Coping with complexity in comparative methodology. In: Alexander R, Broadfoot P, Phillips D (eds.) *Learning from Comparing: New Directions in Comparative Education Research, Vol. 1: Contexts, Classrooms and Outcomes*. Symposium Books, Oxford, UK, pp. 3

Schoppa L 1993 *Education Reform in Japan*. Routledge, London

Shavit Y Blossfeld H P (eds.) 1993 *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*. Westview, Boulder, CO

Tarlton C D 1965 Symmetry and asymmetry as elements of federalism. *Journal of Politics* **27**(4): 861–74

Watson K, Modgil S, Modgil C (eds.) 1997 *Power and Responsibility in Education*. Cassell, London

Weiler H N 1989 Education and power: The politics of education decentralization in comparative perspective. *Education Policy* **3**(1): 31–43

Windolf P 1997 *Expansion and Structural Change: Higher Education in Germany, the United States, and Japan, 1870–1990*. Westview, Boulder, CO

A. J. Heidenheimer

Educational Policy, Mechanisms for the Development of

Discussions of education policy mechanisms have typically focused on the broad array of institutions that influence and enact policy aimed at changing or regulating patterns of practice and governance in schools, classrooms, and supporting organizations. However, as education policy has gained greater prominence on national agendas and the institutions that produce it have grown more diverse, the notion of 'policy mechanism' has expanded to include the governmental and nongovernmental institutions that formulate policy, as well as the ideas, interests, and strategies that shape the options policy makers have at their disposal. This article examines both the institutions that develop education policy, and the sources that give content to those policies.

1. Institutions as Arenas for Policy Development and Enactment

Since policy analysis emerged as a field of inquiry within educational studies more than 30 years ago, the emphasis has been on explaining why particular policies are adopted, what policy makers expect them to accomplish, and the extent to which they are

implemented and actually change teaching and learning. These studies are primarily intended to improve policy by suggesting how their conclusions might inform future initiatives and strengthen the connections between policy and educational practice. Within this paradigm, institutional characteristics are assumed to be critical factors in explaining the enactment and implementation of specific policies.

The overarching conclusions from such studies are twofold: (a) political institutions define the framework in which policy making occurs and thus shape its outcomes; and (b) different institutional arrangements advantage some societal interests and policy alternatives at the expense of others (Kelman 1987). These effects occur through the structure of rules and norms that determines who may participate in the policy making process, who is granted formal authority to choose among proposals, and what procedures must be followed before those with formal authority can make a decision. In education policy, institutional arrangements are particularly critical to explaining policy outcomes, because even in the most centralized systems, political power is exercised and policy decisions are made in multiple venues or arenas. As Mazzoni (1991) notes, these arenas are never neutral in their allocation of access and power: 'They legitimate a set of participants, establish the institutional and social context—including the governing "rules of the game"—mediate the potency of resources and strategies, and encourage some means (and discourage other means) of reaching agreements (Mazzoni 1991, p. 116).

1.1 Interactions Among Governmental Institutions

Much policy research in education, while not always explicit about this focus, examines the interaction in arenas that share authority but operate under different rules and norms. These analyses take a number of forms, but one approach has been to use case studies that compare variation in policy outcomes across national or subnational jurisdictions, and to explain observed differences by focusing on the rules that define the respective roles of different policy-making institutions. To some extent, generalizations can be made about the relationship between the institutions that produce particular policies and the nature of those policies. For example, a common basis of comparison is the difference between policy making in parliamentary and presidential systems. Another relates to the respective domains of civil servants, judges, and elected officials. Because of their professional expertise, education bureaucrats are best equipped to deal with policies requiring high levels of technical knowledge, while judges specialize in interpreting existing law in case-by-case applications, but both groups will typically be less effective in addressing highly politicized or controversial policy issues whose resolution requires the legitimacy of the mandate held by elected politicians. Nevertheless, such generalizations have to be qualified, depending on how a country's or a region's historical development and political culture have shaped its political and educational systems.

A study of judicial policy development is illustrative of this interaction between history and institutional function. In their comparative analysis of education policy making by courts in Germany, the UK, Israel, and the USA, Jung and Kirp (1984) conclude that the judiciary has assumed a more central position in the USA because of constitutional provisions that make judges interpreters of the constitution and enforcers of its guarantees. In contrast, in the other three countries, the courts occupy a less central policy-making position because their role has been circumscribed constitutionally, and because of a greater emphasis on accountability through legislative representation of majoritarian interests and through professional norms guiding decision making by education bureaucrats. These historical institutional differences help to explain why policies involving questions of individual and group rights that are decided in political or administrative arenas in many countries are often transformed into judicial issues in the USA.

Institutional effects on policy outcomes can also be observed in comparing policy-making institutions within a single country. For example, in his study of the politics of school finance in the USA, Wong (1999) finds that the national government's focus on social redistribution to address inequities based on race and social class is not reinforced by state and local allocative practices. Not only are governmental levels often operating at cross-purposes in their fiscal decisions, but the institutional rules at each level favor different constituencies and policy goals. Even in France, where education governance is highly centralized, officials recognize that there are multiple arenas in which policies can be decided, and that different arenas have different balances of power and provide incentives for different types of coalitions. Therefore, a policy outcome may depend in large measure on the arena in which the issue is decided. Across the 30 French cases that Baumgartner (1989) examined, participants ranged from small groups of experts to shared decision making by political appointees and civil servants to massive public debates and street demonstrations, with the intensity of conflict surrounding an issue a significant predictor of whether it was decided by a broad range of political participants or a narrow group of experts.

1.2 The Decentralization Debate

Probably the most common distinction among models of education policy making is the extent to which decisions governing finance, personnel, and curricu-

lum are made centrally or by decentralized institutions such as local boards of education or even councils at individual schools. The optimal balance between centralized and decentralized decision making has been debated on both normative and empirical grounds in industrialized and developing countries alike (Lauglo 1995, Weiler 1990). Over the last 20 years of the twentieth century, there was a move toward greater decentralization of education governance. The arguments in favor of such a strategy have been economic (enhancing cost effectiveness through more efficient deployment of resources), political (ensuring wider representation of legitimate interests), and educational (customizing the learning process to local conditions). Counterarguments focus on the need for the central government to enhance equity in the face of regional resource disparities; represent the interests of the broader polity in socializing youth; and ensure some level of curricular standardization, especially where population mobility is considerable. Despite the liveliness of the debate, research has generated few firm conclusions about the relationship between the degree of decentralization in institutional arrangements governing policy decisions and educational outcomes such as student achievement (Cummings and Riddell 1994, Elmore 1993). However, it is reasonable to conclude that decentralization of some decisions will continue to have political utility as a way to diffuse conflicts over educational goals, even as central states continue to assert greater control over their educational systems (Weiler 1990).

1.3 The Role of Nongovernmental Institutions

The design of policy-making institutions and research on those institutions have long taken into account nongovernmental entities, such as political parties and interest groups, in the formation of public policy. Both pluralist and corporatist models of decision making emphasize the role of voluntary interest groups, with the former focusing on their competition for power and influence, and the latter structuring that pluralism by allowing major interest groups to act in partnership with the state in policy formation. These two models have been particularly effective in explaining the role of large, well-established interests, such as teacher unions, in a variety of countries.

However, even in the context of corporatist states and especially in pluralist systems such as that of the USA, less formal institutions also function as key actors in the policy process. As researchers have gained a more nuanced understanding of the policy process, they have focused on the informal institutions that shape policy. The terms used to describe these informal groupings include 'policy networks,' 'issue networks,' and 'policy communities.' They refer to groups of actors who share an interest in a policy domain such as education, and who are linked by direct and indirect contacts with each other. Members of a policy network may come from within government (e.g., ministry officials, legislative staff) or from outside government, including interest group representatives, academics, and analysts in universities and policy research institutes, as well as journalists. A pivotal player in an issue network and in explaining the emergence of particular problems and proposals on the policy agenda is the 'policy entrepreneur'—someone either inside or outside government who is willing to invest their time, energy, and reputation promoting a particular policy concept in the hope of a future return (Kingdon 1995).

The actions of policy networks and policy entrepreneurs are particularly significant in explaining 'policy borrowing,' or the diffusion of innovations from one jurisdiction to another and across national borders—for example, the move to decentralize education governance and give parents greater choice in their children's schooling. Over the last two decades of the twentieth century, the volume and rate of policy diffusion within education significantly increased through a variety of mechanisms, including international organizations such as the World Bank and the Organization for Economic Cooperation and Development (OECD), associations of elected officials, international networks of politicians sharing similar political ideologies, and policy research institutes or 'think tanks' with national and international audiences. Most research on this phenomenon has been based on comparative case studies, although attempts have been made to calculate diffusion rates and identify major explanatory variables through quantitative analyses (for examples of each type of study, see Mintrom and Vergari 1998, Halpin and Troyna 1995). Although much remains to be learned in this area, research has found that policy makers are likely to 'borrow' policy innovations from other jurisdictions based on both the reliability and trustworthiness of the innovators, and the degree of similarity between the ideology and characteristics of the innovators's education system and those of the 'borrowers.'

2. The Sources of Policy

The focus on informal mechanisms acknowledges that policy development does not begin with the enactment process, but also includes the more diffuse agenda-setting process during which adverse educational conditions are identified as policy problems; new or preexisting proposals are framed as solutions to those problems; and political officials are convinced (often by policy entrepreneurs) that they have sufficient incentive to act on the issue (Kingdon 1995). During agenda setting and enactment, participants in the

policy-making process shape the content of a policy by defining its objectives and selecting the instruments or strategies used to achieve those objectives. The goals and strategies chosen depend to a large extent on how a policy problem is defined, with the entire process of problem definition and policy choice dependent on the competing motivations of participants.

2.1 Interests and Ideas as Policy Sources

Understanding the motivations of those who advance and enact policy proposals can shed additional light on the process by which policy is developed, the sources that give policies their substantive content, and, more generally, on the dynamics of the larger political system. However, over the last two decades of the twentieth century, the nature of those motivations has been the subject of a major normative and empirical debate within economics, political science, and policy analysis.

On the one side is the rational choice model, borrowed from microeconomics. It assumes that political actors know what their self-interest is, and pursue it in the political arena through bargaining with other actors who are also pursuing their self-interest and whose support is needed to build a coalition representing enough votes to enact a given policy. In this model, policy is the result of a series of negotiations or exchanges among self-interested parties.

On the other side are those who argue that accounts of policy making based solely on self-interest are incomplete, and that other goals or motivations are needed to explain policy sources and outcomes. This category of motivations is typically referred to as 'ideas,' and includes both the notion of a public interest that transcends self-interest and, in the case of education, theories about how children ought to be educated and who ought to be responsible for that education. Those who see ideas as a source of policy typically view policy making as more than just the result of whatever bargains can be struck. Rather, it is also characterized by deliberation among participants whose preferences are not fixed. During the deliberative process, participants consider various policy options on their merits. As a result, participants come to understand better their own interests and those of others, and policy decisions are more than just the result of 'splitting the difference' among competing interests (for a discussion of these two models applied to education policy, see the chapters in McDonnell et al. 2000).

Much of the debate over the validity of interests versus ideas in explaining policy outcomes has been waged deductively and on normative grounds, rather than on the basis of empirical research. However, over the past decade, researchers have realized that neither model provides a full explanation. Much of public policy is motivated by self- or group interest, whether it be politicians' desire to be reelected or interest groups's efforts to enhance their members's material advantages. Nevertheless, as Kingdon (1993) argues, self-interest provides an incomplete explanation for policy outcomes. Not only are ideas necessary to persuade others to accept even the most self-interested policy options, but comprehensive and farreaching policies are more likely than narrow, specialized ones to be motivated by ideology and by concepts that transcend self-interest (Kelman 1987). In fact, ideas and interests are so intertwined that it is often difficult to measure their independent effects (Kingdon 1993). One can acknowledge, for example, that teacher unions seek to improve their members's salaries and working conditions, but also that they represent professional values that may lead them to espouse policy options that may benefit the educational enterprise generally, but not necessarily the self-interest of all their members (e.g., consider peer evaluations leading to the dismissal of unqualified teachers).

2.2 Policy Instruments

Whether goals, such as access to advanced courses or higher student achievement, are the result of self-interest or ideas, the policies that embody them must also incorporate strategies for translating those goals into actions that lead to changes in school and classroom practices. To make systematic comparisons across individual policies and to understand how policy design influences implementation and ultimate effects, analysts have focused on the concept of policy instruments or tools. These are generic mechanisms that translate substantive policy goals into the actions needed to effect change, and they can be thought of as the basic building blocks of any policy. Although researchers differ in how they conceptualize categories of policy instruments, and in the total number of distinct instruments they identify, they agree that different types of instrument embody different assumptions about how a policy problem is defined, the causal process through which government action can change the behavior of policy targets, and what the expected effects of that action will be (McDonnell and Elmore 1987).

Two classes of policy instruments are included in all categorizations and are used frequently in education policy: mandates impose rules on targets and are intended to produce compliance (e.g., school-leaving policies, nondiscrimination requirements), and inducements provide money and other tangible payoffs on the condition that targets perform specified actions (e.g., grant-in-aid programs). However, a variety of other instruments are used in education policy, including: capacity-building strategies to promote

longer-term investments in enhancing human capital; hortatory policies that motivate action by providing information or appealing to policy targets's values (e.g., published reports on school quality); and system-changing instruments that transfer authority from one institutional sector to another (e.g., privatization schemes).

3. Future Research

Despite a growing recognition that much can be learned from examining policy mechanisms comparatively, such research is still limited. However, the current worldwide emphasis on implementing policies aimed at comprehensive school reform has given a new impetus to comparative education policy research (Fowler 1995). The most productive direction for that research is likely to be studies based on models that conceptualize policy making as the interplay of interests and ideas, constrained by the institutional rules and norms of the different arenas that share authority over education.

See also: Education, Economics of; Educational Innovation, Management of; Educational Institutions and Society; Educational Planning Worldwide; Educational Policy: Comparative Perspective; Policy Knowledge: Universities

Bibliography

Baumgartner F R 1989 *Conflict and Rhetoric in French Policy-making*. University of Pittsburgh Press, Pittsburgh, PA

Cummings W K, Riddell A 1994 Alternative policies for the finance, control, and delivery of basic education. *International Journal of Educational Research* **21**: 751–76

Elmore R F 1993 School decentralization: Who gains? Who loses? In: Hannaway J, Carnoy M (eds.) *Decentralization and School Improvement*. Jossey-Bass, San Francisco, pp. 33–54

Fowler F C 1995 The international arena: The global village. In: Scribner J D, Layton D H (eds.) *The Study of Educational Politics*. Falmer Press, Bristol, PA, pp. 89–102

Halpin D, Troyna B 1995 The politics of education policy borrowing. *Computers & Education* **31**: 303–10

Jung D J, Kirp D L 1984 Law as an instrument of educational policy-making. *American Journal of Comparative Law* **32**: 625–78

Kelman S 1987 *Making Public Policy*. Basic Books, New York

Kingdon J W 1993 Politicians, self-interest, and ideas. In: Marcus G E, Hanson R (eds.) *Reconsidering the Democratic Public*. Penn State University Press, University Park, PA, pp. 73–89

Kingdon J W 1995 *Agendas, Alternatives, and Public Policies*, 2nd edn. HarperCollins, New York

Lauglo J 1995 Forms of decentralisation and their implications for education. *Computers & Education* **31**: 5–29

Mazzoni T L 1991 Analyzing state school policymaking: An arena model. *Educational Evaluation and Policy Analysis* **13**: 115–38

McDonnell L M, Elmore R F 1987 Getting the job done: Alternative policy instruments. *Educational Evaluation and Policy Analysis* **9**: 133–52

McDonnell L M, Timpane P M, Benjamin R (eds.) 2000 *Rediscovering the Democratic Purposes of Education*. University Press of Kansas, Lawrence, KS

Mintrom M, Vergari S 1998 Policy networks and innovation diffusion: The case of state education reforms. *Journal of Politics* **60**: 126–48

Weiler H N 1990 Comparative perspectives on educational decentralization: An exercise in contradiction? *Educational Evaluation and Policy Analysis* **12**: 433–48

Wong K K 1999 *Funding Public Schools: Politics and Policies*. University Press of Kansas, Lawrence, KS

L. M. McDonnell

Educational Productivity

Educational productivity is the extent to which learning is increased while minimizing costs. Costs may be considered broadly as including not only total monetary costs but also the difficulty of implementing new programs and procedures, possible staff retraining, educational incentives, student time required, and similar considerations, which may lack explicitly estimated monetary costs. Learning can also be considered broadly as including such things as knowledge, understanding, critical thinking, creativity, skills, attitudes, and career success and other long-term adult outcomes. Despite such broad concerns, most empirical research on educational productivity has focused on achievement on standardized tests of knowledge and cognitive abilities related to specific skills and subjects such as reading, mathematics, and science in primary and secondary education. Although the monetary benefits, costs, and their relations are difficult to estimate, much psychological research identifies on educational practices and conditions that increase learning without necessarily increasing costs.

1. Economic Perspectives

Interest in educational productivity accelerated in the last few decades as policymakers, scholars, and educators realized more fully, the contribution of human, social, and cultural capital to economic growth and individual and national wellbeing—an insight that may be traced to Adam Smith's 1776 Wealth of Nations and to Aristotle and other thinkers in Western

and non-Western civilizations. Indeed, parents and educators have long sought the best means of raising and educating children. Today, moreover, civic leaders and citizens more clearly recognize that in addition to physical assets such as farms, industrial plants, machinery, and inventions, the abilities of individuals and their capacity to work together partly determine their wellbeing and future prospects. In addition, it is in the interests of nations and individuals to have productive or efficient education that conserves public funds, tuition, and students' time, which can be used for other purposes, including attaining more advanced education, starting careers and pursuing the other aspects of their lives.

The late sociologist James S. Coleman precipitated much interest in educational productivity by suggesting, on the basis of one of the largest student surveys ever undertaken, that learning does not depend on the amount of money spent on schools. This article sets forth the factors on which learning does depend, as revealed by psychological research of the last few decades of the twentieth century.

Productivity research in most industries lies in the province of economists who identify and measure activities that maximize benefits and minimize costs. Yet, it has proven difficult to estimate the monetary benefits and costs of education. Economists can estimate wealth and income, but the values of other qualities of life, such as citizenship, creativity, and capacity for teamwork, are difficult to evaluate. Policymakers and educators often have little grasp of the total and component costs of schooling and their bearing on the production of learning. They seldom estimate, for example, the monetary value of parent and student time contributed to the leaning process. Even the often-observed relation of the number of years of education and adult income cannot be unequivocally attributed to schooling, since both may partly be attributable to children's abilities and their parents' socioeconomic status and child-rearing practices.

Leading economists, moreover, have compiled a great deal of research on the effects of total and component expenditures on education (see Hanushek 1994). Their analyses show that beyond minimal levels, higher expenditures on teachers' years of experiences and advanced university degrees, physical facilities, smaller class sizes, and other items for which costs can be estimated have little or no influence on how much students learn.

For several reasons, educational productivity research and concerns have been concentrated in the USA. International achievement surveys in mathematics, science, reading, and other subjects show that the longer US students are in school, the more they fall behind those in other countries. Near the end of secondary school, they score last among the economically advanced Asian and European countries. Because they fall behind while in primary and secondary schools, the problem appears to be more attributable to schooling than to families and society. Even so, the per-student costs (adjusted for purchasing power) of schooling are higher in the USA than most other economically advanced countries. Finally, unlike most industries in the past few decades, the productivity of US schooling has declined: Inflation-adjusted costs have risen sharply, while students have made similar or smaller gains in achievement. Such analyses of productivity indicators across time and among nations have provoked much discussion among scholars, legislators, and citizens.

2. *Research on Psychological Productivity Factors*

Since economists have shown little relation between costs and quality, it would seem useful to focus on educational technology, that is, instructional programs, methods, and conditions. Here progress is apparent. Psychologists and educational researchers have measured the chief psychological variables in the education process and discovered their causal relations. Given such information, policymakers and educators can better analyze schooling to make it more productive.

The nine factors, shown in Table 1, have strong, consistent influences on how much students learn within and outside school. Potent, consistent, and widely applicable to many education settings, these nine psychological factors fall into the three groups—student aptitude, instruction, and environment. From two research traditions, considerable research corroborates the influence of these nine factors on learning. 'Meta-analyses' (statistical syntheses of multiple studies) shows that some methods of instruction produce far greater learning than others. From meta-analyses, Walberg and Lai (1999) compiled sizes of effects of 275 educational methods and conditions on the amount students learn. These effects are based on thousands of comparisons of experimental and control groups. Along with cost, feasibility, philosophy, and other considerations, educators can use the 'effects' sizes as criteria in designing and evaluating teacher education and school programs.

A second basis for the factors draws upon the social science tradition of 'secondary analysis,' in which large local, national, or international surveys are statistically analyzed to detect the size of simultaneous effects of multiple causes on outcomes. Analyses of several dozen surveys allowed productivity analysts to estimate the relative effects of the factors in Table 1, including those that are difficult to manipulate merely for the sake of research.

These two research traditions strengthen confidence in the productive efficacy of the factors and complement one another. More often employed by psy-

Table 1
Nine educational productivity factors

Student aptitude
1 Ability or preferably prior achievement as measured by the usual achievement tests
2 Development as indexed by chronological age or stage of maturation
3 Motivation or self-concept as indicated by personality tests or the student's willingness to persevere intensively on learning tasks
Instruction
4 The amount of time students engage in learning
5 The quality of the instructional experience including method (psychological) and curricular (content) aspects
Psychological environments
6 The 'curriculum of the home'
7 The morale or student perception of classroom social group
8 The peer group outside school
9 Minimal leisure-time television viewing (unless educational)

chologists, experiments employing random assignment to treatment and control groups allow strong inferences about causality but, because of costs and other constraints, they are often conducted with small groups of students in homogeneous conditions. Meta-analysis allows inferences from a broad base of many studies and can yield estimates of possible differential efficacy for educational conditions and types of students. Secondary analysts employ statistical rather than experimental control, which may be less telling of causality but often allows wider generality because samples are drawn randomly from nations or even groups of nations. The most confident conclusions come from methods and conditions that yield similar results for many kinds of students in widely varying settings, employing both research methods.

An additional basis for the productivity factors is the biographical study of the highest achievers, secondary school students qualified to compete in the international Mathematical Olympiad from China, Japan, Taiwan, and the USA, and eminent men and women from a variety of fields. Since research showed that the nine factors each promote achievement, it was hypothesized that the very highest levels of accomplishment require concentrating and optimizing all nine factors to attain peak performance in mathematical problem-solving ability and in preparation for outstanding careers. For top mathematics performance, data collected on biographical questionnaires supported this hypothesis.

To test the hypothesis with respect to adult eminence, a sample of pre-twentieth-century eminent men was drawn, based on the length of their biographies in American and European encyclopedias. A sample of twentieth-century eminent women was also identified by their outstanding contributions to such fields as art, music, science, writing, politics, and religion. Ratings made by biographers and ratings made by content analysts from biographies confirmed the hypothesis. Many eminent people, while not necessarily highly successful in school, persevered through difficulties and concentrated on their chosen fields, even in the first 13 years of life. Parents, teachers, and others identified their special talents, encouraged them, and provided opportunities for them to develop to their fullest. The general corroboration of the determining effects of the productivity factors was particularly important in these samples, because nearly all other research focused on academic achievement in school, rather than adult accomplishments.

3. *Specific Effects of Productivity Factors*

The economic idea of scarcity can promote understanding of how the factors affect learning. The first two sets of five factors in Table 1 may be termed essential, since without minimal levels of each, students are unlikely to learn. For example, if they are completely unmotivated, or given no time to learn, or taught in a language completely foreign to them, they learn little. These essential factors, moreover, appear to substitute, compensate, or trade-off one another, possibly at diminishing rates of return. Immense quantities of time, for example, may be required for a moderate amount of learning, if motivation, ability, or instructional quality is minimal. Thus, no single essential factor overwhelms the others; all appear important. In addition, concentration in one field makes high accomplishment in another more difficult; the best chess player is unlikely to be best at physics or drawing.

Although the environment factors in Table 1 are consistent statistically- or experimentally-controlled influences on learning, they may directly supplement as well as indirectly influence the essential classroom factors. In either case, the powerful influences of out-of-school factors, especially the home environment must be considered. For example, if students spend 12 years of 180 6-hour days in elementary and secondary school, they will have only spent about 13 percent of the waking, potentially educative time during the first

18 years of life in school. If more of the 87 percent of the student's waking time nominally under the control of parents, that is spent outside school, were to be spent exposed to academically-stimulating media and conditions in the homes and peer groups, then students might learn much more.

The strong influences of ability, motivation, and age are largely reflected in achievement tests in the academic subjects. Achievement at any point in time also reflects the previous home environment, exposure to mass media, and amount and quality of instruction. Current achievement predicts subsequent achievement, and the older the student, the better the prediction. High achievers in secondary school science, for example, can be expected to achieve well in science, engineering, medicine, and other university courses and programs with a heavy science content. Although 'late bloomers' can be identified, they are rare, especially in the more cognitively demanding fields; hence, the importance of a good early start. Another consequence of the cumulative nature of achievement is the need to consider 'value added' indicators of school productivity. Schools, for example, with relatively poorly prepared students, who nonetheless make good gains, should be given credit even though their end results may only be average.

Quantity of instruction or engaged learning time, along with the other factors, can be viewed as a scarce ingredient in learning. Quality of instruction can be abstractly understood as providing encouragement, information, questioning, and other stimuli to ensure the fruitfulness of engaged time. Diagnosis and tutoring can help ensure that instruction is suitable to the individual student. Inspired teaching can enhance motivation to keep students persevering. Quality of instruction, then, may be considered as an efficient enhancer of learning time. Some methods are more effective than others, though not necessarily widely employed. Expensive reductions in class sizes, for example, have led to highly inconsistent and debatable achievement results. Some methods of teaching, however, such as direct instruction, have excellent research records and do not necessarily cost extra money, since many teachers can learn these individually, and others during routine staff development. Still other methods are traditional and should always be included in teachers' repertoires.

The four psychological environments defined in Table 1 can enlarge and enhance learning time. Good classroom morale may reflect match of the lesson to student aptitude, the stimulating properties of the academic group, or, in general, the degree to which students concentrate on learning, rather than diverting their energies, because of unconstructive social climates. Peer groups outside school and stimulating home environments provided by parents can help by enlarging learning time and enhancing its efficiency; students can both learn in these environments and become more adaptable to formal schooling.

Because students spend so much time outside school, their parents play decisive roles in how well they do inside school. In addition to encouraging and supervising homework, parents can improve academic conditions at home. What might be called 'the alterable curriculum of the home' is much more predictive of academic learning than is family socioeconomic status. This curriculum refers to informed parent–child conversations about school and everyday events; encouragement and discussion of leisure reading; monitoring, discussion, and guidance of television viewing and peer activities; deferral of immediate gratification to accomplish long-term goals; expressions of affection and interest in the child's academic and other progress as a person; and perhaps, among such careful efforts, caprice.

Turning to the last and only negative factor, leisure-time media exposure including television (although many television programs can be educational) and radio, can displace homework, reading, and other educationally stimulating activities. They may also dull the student's keenness for academic work. Some of the average of 28 hours a week spent viewing television by US secondary school students might usefully be added to the mere four or five weekly hours of reported homework (Walberg 1984). Ten of those hours a week for several years could be sufficient to learn tennis, a foreign language or a specialized science.

Many experimental studies and analyses of survey data show the pervasive and fairly uniform effect of the productivity factors in the usual school subjects and on students of different ages, countries, and cultural groups. One recent study, for example, concerned the possible differential effect of the factors on mathematics achievement for African-American and other, mostly white students. The question was asked if some factors have greater or lesser effects on the African-American than on the other students. The longitudinal analysis of mathematics scores of a national sample of more than 10,000 students showed equal effects of the factors on African-American and other students. Thus, for example, more study time, better instruction, and a more stimulating home environment had equal effects on how much African-American and other students learned, and how favorably their attitudes toward subject matter changed.

4. *Current and Future Initiatives*

Even though educational productivity research is extensive, policymakers and practitioners seem to have made only partial use of it. Part of the problem could be an information gap. For this reason, the International Academy of Education, an honorary scholarly group headquartered in Brussels, Belgium, began a series of pamphlets on 'best practices.' In Geneva, Switzerland, the International Bureau of Education, a

part of UNESCO, distributes about 4,000 copies of the series to education officials throughout the world, who are free to reproduce the booklets. The booklets are also available on the Internet pages of the two organizations, and educators may download them for reproduction.

The barriers to implementing well established achievement-raising practices, however, may not be completely attributable to a lack of information but to failures in (a) setting achievement as schools' top priority, (b) choosing the most appropriate procedures to attain specified achievement goals, (c) specifying system-wide procedural plans and milestones, and (d) monitoring progress, particularly with respect to procedural and achievement goals. Because school systems have been slow to accomplish these things, policymakers are making radical changes to provide incentives for putting effective methods in place. These include closing failing schools, financially supporting public school students whose families want to send them to private schools, and giving authority for school governance to mayors and parents, rather than educators. At the same time, public and private education providers are experimenting with curriculum and instruction that can be provided through the Internet, and evaluating other technologies that may provide more convenient and productive education. Although it is difficult to say where all this will lead, it does appear certain that attention to and research on the problem of educational productivity, will continue to expand.

See also: Academic Achievement: Cultural and Social Influences; Cost–Benefit Analysis; Creativity Training; Education, Economics of; School Achievement: Cognitive and Motivational Determinants; School Effectiveness Research; School Improvement

Bibliography

Ahlgren A, Walberg H J 1973 Changing attitudes towards science among adolescents. *Nature* **245**: 187–90

Campbell J R, Wagner H, Walberg H J in press Academic competitions and programs to challenge the exceptionally talented. In: Heller K A, Monks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of Research and Development and Giftedness and Talented*. Pergamon, Oxford, UK

Hanushek E A 1994 *Making Schools Work: Improving Performance and Controlling Costs*. The Brookings Institution, Washington, DC

Walberg H J 1983 Scientific literacy and economic productivity in international perspective. *Daedalus* **112**(2): 1–28

Walberg H J 1984 Improving the productivity of America's schools. *Educational Leadership* **41**: 19–27

Walberg H J, Lai J-S 1999 Meta-analysis as a basis of educational policy. In: Cizek G (ed.) *Handbook of Educational Policy*. Academic Press, San Diego, CA, pp. 419–53

Walberg H J, Zeiser S in press Productivity, accomplishment, and eminence. In: Colangelo N, Davis G (eds.) *Handbook of Gifted Education*. Allyn and Bacon, Boston, MA

H. J. Walberg

Educational Research and School Reform

Educational research is a growing industry. Its aim is to make a contribution to knowledge. How this knowledge can be used, and in some cases if it can be used, to reform education forms the focus of this article.

In 1899, William James in his *Talks to Teachers on Psychology* emphasized that education was 'an art' and not a science and could therefore not deduce schemes and methods of teaching for direct application from psychology: 'An intermediary inventive mind must make the application by using its originality' (p. 8). Nearly 100 years, later De Landsheere (1994 p. 1871) 'Like medicine, education is an art. That is why advances in research do not directly produce a science of education, in the positivistic meaning of the term, but yield increasingly powerful foundations for practice and decision-making.' Masters (1999), on the other hand, while making comparisons with medical research, suggests that in education research syntheses, sharper-focused research questions, and better communications between researchers and research utilizers, would do much to improve the link between research and school reform.

Husén (1994) suggested that the relationship between research and practice in policy-making had, for a long time, been conceived in a rather simplistic way. 'Policy makers wanted research that primarily addressed their pressing problems within the framework of their perceptions of the world of education.' They also wanted 'recipes' for their immediate problems (Husén 1989).

In 1960, there seemed to be general agreement that it would take at least one generation, about 15 years, between the results being produced by any of the then current research projects and their implementation at the school or classroom level. Since the 1960s, educational research has become a large enterprise with governments supporting it financially. Increasingly it is hoped that research can give pointers to the elements of reform that can either help to raise cognitive achievement, affective development and skills or cut the costs of education without damage to the desired outcomes.

This article deals with some of the different aspects of educational research, the contexts in which policy-making and research take place, the disjunction between researchers and policy-makers, and models of

research utilization, and also presents some examples of the establishment of links between research results and action or reform.

1. Different Aspects of Educational Research

Before proceeding with the link between research and reform, it would be desirable to deal with the various aspects of research.

Most policy-makers want, as mentioned above, clear results from research for the problems they have. These results should preferably be accompanied by a set of suggestions for reform or action. The research from which the results emanate should be 'sound.'

Given that it is often extremely difficult, if not impossible, to conduct good experiments in education from which 'cause and effect' can be safely identified, researchers often have to depend on natural variation using survey samples. However, this depends on natural variation being there for the phenomenon under investigation. Class size is an example of where there is some variation in school systems but it is often within fairly narrow limits. Within a country, the average class size may be 25 children per class at a certain grade level. But the range is only 20–30. Yet, in other countries, the average class size may well be over 40 or even over 60. And in these other countries the range will be small. As another example, a country may wish to examine the differences between private and public schools, but if there are no private schools, then there is no natural variation. Furthermore, there is always the criticism that cross-sectional and longitudinal studies are 'correlational' and assumptions must be made about 'cause and effect.'

The research must be 'sound.' Yet it can be argued that there are many studies that have technical deficiencies such that the results cannot be trusted. These deficiencies can include poor conceptualization, poor measurement, poor sampling and incorrect estimates of sampling error, inappropriate analyses, and false interpretations and conclusions. Some could argue that more unsound than sound research is produced in the field of education. The International Academy of Education produced a summary of 'The requirements of a good study' for sample surveys in international educational achievement studies (International Academy of Education 1999, Appendix 2). This attempts to describe in lay terms the kinds of points in research publications that policy-makers should look for when trying to decide whether or not the research was well conducted and trustworthy. It would be useful if summaries such as this one were to be produced for different types of research.

It is important that readers of research reports can identify if the research conducted was unsound and cannot be trusted. Many senior policy-makers do not have this ability and must rely on those skilled in these matters in their own ministries of education. Unfortunately, not all ministries have such people. The dangers of implementing policy based on poor research are obvious.

It is also the case that policy-makers need results that have, if possible, been replicated and that are generalizable to a grade level or general level (e.g., junior high school or elementary levels of education) in an educational system, and where the effects are large. Research that is only applicable to a few specific schools (case-study approach) is not of interest to them. Furthermore, the research should ideally be applicable to several key subjects in school and not just to one subject. Ministries allocate resources to schools and they want the resources to have wide application in each school and not just to one subject. Unfortunately, having results on the effect of variables on several subjects at the same time is costly and time consuming.

Keeves (1994) has distinguished between research studies associated with the generation of change and those that serve to maintain and consolidate existing conditions. It is the second type of research that he calls 'legitimatory research.' Thus the motivation for commissioning research must also be taken into account when viewing the links between research and change.

2. Contexts for Policy-making and Research

2.1 For Policy-making

As already stated, policy-makers are interested in their problems within their current frame of reference and understanding of education. They perceive of what researchers know as 'fundamental' research to be of no, or only peripheral, interest to them. What the policy-makers perceive to be of interest can change with a change of government. For example, the matters of bussing, educational vouchers, and private schools had different priorities under the Carter and Reagan administrations in the USA. Equality of educational opportunity, on the other hand, has had a high priority under various administrations in various countries. In the second project of the Southern Africa Consortium for Monitoring Educational Quality (SACMEQ), chief policy-makers in 14 southern African countries were asked to rate issues that were of interest to them in primary schooling. One of the issues was that of private tuition outside school. This was given very low priority despite the fact that the first study of SACMEQ had shown the practice to be very widespread, for the most part to affect achievement even after home background had been accounted for. It was later given a higher priority after the researchers had interacted more with the policy-makers.

Policy-makers have allegiances that influence both what they regard as relevant, innocuous or even 'dangerous' research and also their willingness to take account of research findings. Stories abound of ministries or government departments either ignoring research results or delaying the publication of government-sponsored research because the results did not accord with the current political interests. Ways of delaying the publication of research results that governments do not like have been hilariously resumed in 'Yes Minister' (Lynn and Jay 1982). Not only can research be delayed or ignored but also only partial results can be selected to help them argue a case. Keeves (1994) has dealt with many of the problems that arise with those in authority when conducting legitimatory research.

Governments, as a rule, last only a few years. The research they sponsor should be completed within that time period and preferably in less than two years. There are few research projects that can be completed in this time period when it is considered that the instruments have to be developed and tried out before the major data collection can begin. The findings should be made available in time for the decisions that have to be taken and which will go ahead with or without the necessary knowledge base. Operational decision-making cannot wait for the results of specifically sponsored research results.

It has sometimes occurred that governments have initiated a reform, often without much evidence of its worthwhileness, and then asked for research to be undertaken to assess its effect. Where the reform has been total, it is too late to compare the effects of the reform with any other form of education. Even if the reform has been in some and not all schools, the problem has arisen that the government wants quick results and yet it takes some time for an innovation to 'settle down.' It is not good, or even dangerous, to judge its effects in the first year of operation.

2.2 For Researchers

The researchers often tend to have different backgrounds and value systems from those of the policy-makers. Researchers have usually undertaken their work in research institutes or university settings. Researchers have been trained to think widely and to question all things. Their allegiance tends to be more to fundamental or conclusion-oriented research. They pay much more attention to how the quality of their research is received by their research peers than by government agencies.

In most cases, researchers are not involved in the planning and implementation of reform. They rarely have any interest in how to plan the costs involved in any reform. They can rarely phrase their results in the language of planning and implementing reform. Indeed, once they have published their research report they tend to lose interest in what happens after that.

Status in the research system depends on the reputation the individual researcher gains from his or her research work. Status in the administrative world tends to be based on seniority and length of service and not on the quality of research work.

3. Disjunction Between Researchers and Policy Makers

From the above, it can be seen that the 'ethos' pervading in the contexts in which researchers and policy-makers live is very different. In nearly all cases, the researcher has to write a proposal for the research project that he or she must submit for funding. Very often the researcher views the proposal from the point of view of its likely contribution to knowledge whereas the policy-maker examined it from the point of view of the relevance to the ministry's agenda. Education is value based and has certain underlying ideologies. It is a short step, therefore, to the political arena. It can often happen the governments view social science research with suspicion because when they want the status quo to be preserved social science research that wants to go to the root of the matter in the sense of critical inquiry can be seen as subversive radicalism. On the other hand, some social scientists—because of the relationship between education and certain political and social philosophies—have begun to act as evangelists and this has damaged their credibility.

In the 1960s, governments began to support social and behavioral sciences research on a hitherto unprecedented scale. This was particularly true in countries such as the USA, UK, and Sweden (Aaron 1978). It was not long before there were discrepancies between expectations on the part of the policy-makers and the actual research performance. There was a decreasing credibility of research work on the part of the policy-makers that began in the early 1970s.

4. Models of Research Utilization

Weiss (1979) distinguished among seven different models of research utilization in the social sciences (see also Husén 1994):

(a) *The R&D model*. This is a 'linear' process from basic research to applied research and development to application. Weiss pointed out that the applicability of this model was limited in the social sciences because knowledge in the field does not lend itself easily to 'conversion into replicable technologies, either material or social' (Weiss 1979, p. 427).

(b) *The problem-solving model*. Here the results from a specific project are expected to be used directly in a decision-making situation. The process consists of: identification of missing knowledge → acquisition of research information either by conducting a new study or by reviewing the existing body of knowledge → interpretation of the results given the policy

options available → decision about which policy to pursue. This model has sometimes been known as the 'philosopher-king' approach where researchers are supposed to provide the information from which policy-makers derive guidelines for action. The problem-solving model often tacitly assumes about goals but social scientists do not agree among themselves about the goals of certain actions. Nor is there necessarily agreement among all policy-makers about the goals.

(c) *The interactive model.* This model assumes an ongoing dialogue between researchers and policy-makes. This is usually a disorderly set of interconnections and back-and-forthness.

(d) *The political model.* Research findings are used as ammunition to defend a particular point of view. It is often the case that policy-makers have already made up their mind about taking a particular course of action and should there be research results that legitimize this, then they will use it.

(e) *The tactical model.* This is a negative approach in that it is a way of deferring any decision. Thus, a controversial issue can be buried or postponed by policy-makers calling for more research or further analyses.

(f) *The enlightenment model.* According to Weiss (1979, p. 428), this is the model where social science research most frequently enters the policy arena. Research can 'enlighten' policy-makers because as research results become available they sensitize informed public opinion about ways of thinking of educational problems and come to shape the way in which people think about social issues. This is sometimes known as the percolation effect. People are helped to redefine problems through research so that they begin to think of the problems in a different way.

(g) *The research-oriented model.* Weiss (1979) refers to this as the 'research-as-part-of-the-intellectual-enterprise-of-society' model. Social science research together with other intellectual inputs, such as philosophy, history, journalism, and so on, contribute to widening the horizon for the debate ion certain issues and to reformulating the problems. This is somewhat akin to the enlightenment model outlined above.

Shavelson (1988, pp. 4–5), when dealing with the utility of social science research, sought to reframe the issue of 'utility' by suggesting 'that the contributions lie not so much in immediate and specific applications but rather in constructing, challenging, or changing the way policy-makers and practitioners think about problems.' He suggested that one cannot expect educational research to lead to practices that make society happy, wise, and well-educated in the same way that the natural sciences lead to a technology that makes society wealthy. The assumption that 'educational research should have direct and immediate application to policy or practice rests on many unrealistic assumptions'. Among these are relevance to a particular issue, provision of clear and unambiguous results, research being known and understood by policy-makers, and findings implying other choices than those contemplated by policy-makers.

5. *Some Examples*

It is relatively easy to identify some research that has had an enlightenment effect. The work of Piaget in identifying 'stages of development' began to have an effect on work in curriculum development by the end of the 1950s. Bloom's 'Taxonomy of Educational Objectives' and 'Model of School Learning' had a marked effect on many later research projects, the results of which entered the market place and influenced educational thinking. Coleman's 'Equality of Educational Opportunity' produced the concepts of 'school climate' and later 'social capital,' both of which entered educational thinking. The 'Plowden Report' in England produced the concepts of parental attitudes and 'educational priority zones,' both of which entered the general educational thinking. Carroll's 'model of school learning' and later his 'human cognitive abilities' research both influenced how educators thought of school learning. However, in all of these cases, the process took a long time. In some special cases in Sweden in the late 1950s and early 1960s research studies led directly to some action especially in the revision of the curriculum. However, the direct effect examples are rare.

In the first two decades or even more of the work of the International Association for the Evaluation of Educational Achievement (IEA) (Postlethwaite 1994), a massive amount of effort went into attempting to identify the major determinants of educational achievement in each of the participating countries. It was to the chagrin of the research workers that very little attention was given to the results. It was rather the 'horse race' national mean achievement levels that gained the attention of the press and the policy-makers. In the 1970s the then Minister of Education talked of the results being like an electric shock because they showed that the gap between north and south Italy was much larger than had been expected. The Third International Mathematics and Science Study (TIMSS) of IEA (see, for example, Beaton et al. 1996) in the 1990s only presented aggregate estimates for the test scores and selected questionnaire variables in their first set of publications and did not present analyses identifying the major determinants of either between country differences, or between-school or between-student differences within countries. Nevertheless, it was these aggregated estimates that attracted enormous media publicity and comment. In some cases the comment was appropriate such as about the high achievement scores in Japan but the low attitudinal scores towards the learning of the subject. In other cases, several of the journalists' comments and even those of some ministry officials proffering policy

Table 1
Selected policy actions for the Mauritius report

Policy suggestion	Relevant department(s)	Time	Cost
Consultations with staff, community, and experts			
1. The Ministry should ask the Staffing Branch to meet with Teacher Union representatives to discuss the potential 'conflict of interest' that arises from the current practice of allowing the teachers of around 80 percent of Grade 6 pupils to receive high financial rewards for teaching their own pupils as private clients outside of school hours	Staffing Branch and Teacher Union	Short	Low
Reviews of existing planning and policy procedures			
2. The Inspectorate should be asked to review and, if necessary, establish benchmark standards for the educational environment that are deemed to be 'reasonable for the proper functioning of primary schools'	Inspectorate	Short	Low
Data collection for planning purposes			
3. The Supply Branch of the Ministry should investigate the reasons for the relatively high percentage of Grade 6 pupils in Rodrigues who reported that they lacked basic learning materials	Supply Branch	Medium	Low
Education policy research projects			
4. The Planning Unit should undertake an investigation into the practice of grade-repeating in Mauritius in order to determine whether the extra year of schooling being received by the 'average Grade 6 pupil' can be justified on either educational or economic grounds	Planning Unit	Medium	Low
5. The Ministry should plan to undertake a follow-up survey of the same target population employed during SACMEQ's initial project in order to examine changes in important educational indicators over time	Planning Unit	Medium	Moderate

suggestions assumed that the relationships between certain variables between countries would be the same as those within countries (Ross 1997). The question must be asked as to why the press and the ministries seem to be more interested in the mean score differences rather than in the determinants of such differences. Is it possible that an interface of persons skilled in translating research findings into policy suggestions is needed?

Several of these kinds of projects did attempt to communicate to schools that participated in the research and would send each school a school report. This typically consisted of giving the test scores for each child in the sample in the school, the school's mean scores on the test as a whole and on subscores, the mean scores of similar schools and the mean scores of the nation. It was assumed that these kinds of results would enable school principals to identify the weak parts of achievement in his/her school and then be able to take remedial action. It would seem that some school principals did take the feedback seriously, but many did not do so.

By the beginning of the 1990s, some progress was being made in the attempt to have educational planners and administrators work more closely with educational researchers in order to have more direct links between research and policy-making. One international study in the 1990s had the strategy of working out in conjunction with ministries of education the research questions and then working out the actions to be taken as a result of the findings. This latter step was undertaken *before* the research reports were finalized for publication. This study was that of the Southern Africa Consortium for Monitoring Educational Quality (SACMEQ). The last chapter of each national publication listed the suggested policy actions categorized by cost and time needed for implementation. An example is given in Table 1 of

selected suggested policy actions for the Mauritius report (Kulpoo 1998, pp. 86–92). There were 38 policy suggestions in all.

It should be noted that the units undertaking the research study were units within the Ministry of Education and not special research units either within the ministry or in higher education institutes. The actual suggestions were checked, and in one or two cases modified, by the senior ministry personnel before the book was published. To what extent the suggestions were implemented is not yet known. On the other hand, there was a similar study in Zimbabwe with similar policy suggestions in 1991 and a report documenting the implementations of the suggestions has been published (Ross 1995). Despite the fact that some of the suggestions were implemented, there was no improvement in Grade 6 reading achievement by 1997 (Shumba 1998). It is unclear whether the implementation was not well conducted or whether only some of the easier but not crucial suggestions were implemented. This will also apply to the other SACMEQ reports.

In the USA, the United States Department of Education (1986) published a booklet under the title *What Works: Research About Teaching and Learning*. This booklet reported the results of research syntheses. It was aimed at all involved in education and presented in one or two pages the results of research and the practical implications of those results. This was followed by a series of similar booklets on different aspects of education. These booklets were made available to all schools in the USA. This was an attempt to communicate directly with school principals and teachers and was said to be very popular.

From the mid-1980s onwards there has been a movement initiated by Ministries of Education known as the 'Indicators Movement.' The Organization for Economic Co-operation and Development (OECD) has coordinated this work and produced a series of annual reports called 'Education at a Glance' (see the most recent one at the time of writing, OECD 1998). Data are collected and aggregated and reported at the national level. In some cases these results are of interest. There is the danger that the correlation of two variables/indicators at the national level is not the same as at the within country between school or between student levels (cf., Ross 1997). Where policy is made on the basis of the national level correlations and the within-country correlations are different, this can lead to misleading policy. It is important to examine the relationships at different levels before arriving at suggested action for policy-making. The OECD has begun to overcome this problem by having initiated a 'Program for International Student Assessment' (PISA), where every four years students aged 15 years (i.e., just before the end of compulsory schooling in most countries) will be tested in mathematics, science, and reading as key subjects and variation in achievement will be analyzed at student, school, and country levels. Data were also collected on student and school background variables (indicators) and will be related to variation in student outcomes.

There has been a host of research studies on such topics as learning, programmed instruction, leadership (authoritarian, democratic, and laissez-faire), grouping of students and achievement outcomes, learning methods, motivation (expectations of success and failure), grading practices, creativity and teaching methods, language (and foreign language) learning, the regularity and sequencing of in-service teacher training programs, teaching styles, and so on. Many suggestions have emanated from such research but in some cases there were contradictory results from different studies or the suggestions, even from sound research, were ignored. In some countries, there has been a tendency to initiate 'bottom-up' reform from examples from single schools. These have been case-study type research and have suffered from two defects: they could not be generalized in the sense of knowing how the innovation would work in many schools, and in many cases they were successful because of the charisma of a particular individual. When attempts have been made to replicate the innovation on a larger scale, there have been problems.

6. *Conclusions*

A great deal of money goes into educational research and yet the results are rarely used directly. Indirectly and over time research results 'percolate' down (one of the Weiss models of research utilization) to the district and school levels and then demands are made on the policy-makers to 'do something.' Even when policy-makers take the results seriously, there remains doubt about how well the implementation of policy suggestions was carried out.

If research is in some cases to have a more direct link to school reform it would seem as though several steps should be taken:

(a) Researchers should talk to the policy-makers before the research begins and ensure that the design and proposed analyses will answer the policy-makers' concerns.

(b) Projects should be conducted in a sound technical way. This requires that the funders have a cadre of persons that can advise them about the technical soundness of the proposed research. This is a big step because there are not many ministries even in highly developed countries that possess such persons.

(c) Projects should, wherever possible, be replicated.

(d) Researchers should, after consultation with the policy-makers, make a list of suggested actions for reform.

(e) Ministries of education should train and employ a cadre of policy researchers who are able to translate the research results into suggested actions that constitute the crucial actions and not only those easy to

conduct. These persons may not necessarily be within the ministry but could be in university departments.

(f) Policy suggestions should be acted upon and well implemented.

(g) Studies should be conducted to assess the effect of the implementations.

Until these kinds of steps are taken, research will continue to have its indirect enlightenment effect.

Finally, it should be remembered that there are certain reforms that do not depend on research. When the great debate raged in the 1960s in Europe about comprehensive schooling versus selective schooling it was the then Minister of Education in England and Wales who wrote, when asked why England has not preceded 'going comprehensive' with research such as that conducted in Sweden, 'It implied that research can tell you what your objectives should be. But it can't. Our belief in comprehensive reorganization was a product of fundamental value judgements about equity and equal opportunity and social division as well as about education. Research can help you to achieve your objectives, and I did in fact set going a large research project, against strong opposition from all kinds of people, to assess and monitor the process of going comprehensive. But research cannot tell you whether you should go comprehensive or not—that's a basic value judgement' (Husén and Kogan 1984).

See also: Comparative Research in Education: IEA Studies; Educational Evaluation: Overview; Educational Innovation, Management of; Educational Institutions and Society; Educational Planning Worldwide; Educational Policy, Mechanisms for the Development of; Educational Research for Educational Practice; School Improvement

Bibliography

Aaron J H 1978 *Politics and the Professors: The Great Society in Perspective*. Brookings Institution, Washington, DC

Beaton A E, Mullis I V S, Martin M O, Gonzales E J, Kelly D, Smith T A 1996 *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. TIMSS International Study Center, Boston College, Chestnut Hill, MA

De Landsheere G 1994 Educational research, history of. In: Husén T, Postlethwaite T N (eds.) *International Encyclopedia of Educational Research*. Pergamon Press, Oxford, UK, Vol. 4, pp. 2053–9

Husén T 1989 Educational research at the crossroads? An exercise in self-criticism. *Prospects* **19**(3): 351–68

Husén T 1994 Educational research and policy-making. In: Husén T, Postlethwaite T N (eds.) *International Encyclopedia of Educational Research*. Pergamon Press, Oxford, UK, Vol. 3, pp. 1793–7

Husén T, Kogan M (eds.) 1984 *Educational Research and Policy: How Do They Relate?* Pergamon Press, Oxford, UK

International Academy of Education 1999 *The Benefits and Limitation of International Studies of Educational Achievement*. IIEP, Paris

James W 1899 *Talks to Teachers on Psychology: and to Students on Some of the Life's Ideals*. Longmans Green, London

Keeves J P 1994 Legitimatory research. In: Husén T, Postlethwaite T N (eds.) *International Encyclopedia of Educational Research*. Pergamon Press, Oxford, UK, Vol. 6, pp. 3368–73

Kulpoo D 1998 *The Quality of Education: Some Policy Suggestions Based on a Survey of Schools: Mauritius*. SACMEQ Policy Research: Report No. 1. Ministry of Education and Human Resource Development, Mauritius and International Institute for Educational Planning, Paris

Lynn J, Jay A (eds.) 1982 *Yes Minister: The Diaries of a Cabinet Minister by the Rt. Hon. James Hacker MP*. Vol. 2. British Broadcasting Corporation, London

Masters G N 1999 Valuing educational research. *ACER Newsletter Supplement*, No. 94. Australian Council for Educational Research, Melbourne, Australia

OECD 1998 *Education at a Glance: OECD Indicators*. Organisation for an Economic Co-operation and Development, Paris

Postlethwaite T N 1994 Educational achievement: comparative studies. In: Husén T, Postlethwaite T N (eds.) *International Encyclopedia of Educational Research*. Pergamon Press, Oxford, UK, Vol. 3, pp. 1762–69

Ross K N 1995 From educational research to educational policy: an example from Zimbabwe. *International Journal for Educational Research* **23**: 301–403

Ross K N 1997 Research and policy: a complex mix. *IIEP Newsletter* **15**(1): 1–4

Shavelson R 1988 Contributions of educational research to policy and practice: constructing, challenging, changing cognition. *Educational Researcher* **17**: 4–11, 22

Shumba S 1998 Was there a change in reading-literacy levels for Grade 6 pupils in Zimbabwe during the period 1991 to 1995? In: Machingaidze T, Pfukani P, Shumba S (eds.) *The Quality of Primary Education: Some Policy Suggestions Based on a Survey of Schools*. SACMEQ Policy Research Report No. 3. IIEP (UNESCO), Paris and Ministry of Education and Culture, Zimbabwe

United States Department of Education 1986 *What Works: Research About Teaching and Learning*. US Department of Education, Washington, DC

Weiss C 1979 The many meanings of research utilization. *Public Administration Review* **39**: 426–31

T. N. Postlethwaite

Educational Research for Educational Practice

After 100 years of systematic research in the fields of education and educational psychology, there is still no agreement about whether, how, or under what conditions research can improve educational practice. Although research and educational practice each have changed substantially since the beginning of the

twentieth century, the question of how science can actually contribute to the solution of real educational problems continues to be controversial.

Because there are no general rules or strategies for translating research into practice, the first part of the present article focuses briefly on some of the basic problems concerning the relation between theory and practice in education. Next, six different approaches that attempt in various ways to make research findings relevant to education, particularly in schools, are discussed. These approaches concern: (a) using theoretical knowledge to improve technology for teaching and learning; (b) facilitating teacher expertise as a practical application of research on teaching; (c) using studies on classroom learning to provide the scientific basis for ability grouping, mastery learning, and adaptive teaching; (d) using research on cognitive development and learning as a source of scientific information about how to train student's learning competencies and self-instructional skills; (e) using educational research to design, implement, and evaluate new models of schooling and instruction; and (f) using research findings as a source of background knowledge for practitioners. Suggestions for how research can be made more relevant to educators and for ways in which educational practice can be more research-based are proposed on the basis of these different approaches.

1. The Gap between Research and Educational Practice

Most research in education and educational psychology is conducted with the explicit or implicit goal of directly or indirectly improving educational practice. As Anderson and Burns (1985) stated, 'the primary purpose of classroom research is to help educators to improve the conditions of learning and the quality of learning of increasingly large numbers of students' (p. ix). The kinds of necessary research, the required theoretical insights, and the ways that scientific findings should be translated into practice to attain this goal are issues that have not as yet been resolved.

When empirical and experimental research on educational phenomena began, there was a widespread vision, conviction, and expectation that it would generate a scientifically informed basis for educational practice. For example, Rein (1893) stated that 'there is only one way in teaching that corresponds to nature; to follow carefully the laws of the human mind and to arrange everything according to these laws. The attainment of adequate instructional procedures follows from knowledge of and insight into these laws' (p. 107).

When these ambitious hopes did not seem to be realized, there was substantial disappointment among both scientists and practitioners. It was not clear how teachers could use such very general and vague recommendations as those, for example, based on a synopsis of classical learning theories by Thorpe and Schmuller (1954). They concluded that instruction was especially effective when learners were motivated, when task demands matched learners' aptitudes, when learners received sufficient opportunities to relate elements of the learning task to the learning goal, when learners could use external criteria to judge their progress, and when the learning process occurred under conditions that facilitated adaptation to the total situation. Although these conclusions are probably valid, it is not surprising that such nonspecific psychological statements and such obvious educational recommendations led the educational community to the cynical conclusion that learning theorists were perhaps the only ones who could derive personal benefit from learning theory.

Dissatisfaction was not limited to problems in finding practical applications for learning theory. The results from research on teaching also seemed unproductive in providing educators with practical guidance. This can be seen, for example, in Bloom's (1966) resigned summary of the state of research: 'large class, small class, TV instruction, audiovisual methods, lecture, discussion, demonstration, team teaching, programmed instruction, authoritarian and non-authoritarian instructional procedures, etc. all appear to be equally effective methods in helping the student learn more information or simple skills' (p. 217).

The limited applicability of educational science for educational practice has resulted in several metatheoretical debates that have contributed only a little to improving the research situation or to solving practical problems. Some examples of the debated issues are:

(a) Is the gap between scientific and practical work the result of a production deficit in research, a reception deficit in the practitioners, or deficits in translating theoretical knowledge into practical suggestions? This debate has led at least to the insight that scientists and practitioners have different ways of perceiving the world, prefer to define their problems in different ways, and thus speak different 'languages.'

(b) How relevant is basic research, as opposed to applied research, for education? This debate is especially unproductive, both theoretically and practically. On the one hand, it is difficult to differentiate pure or basic research from applied research; on the other hand, it has become clear that findings from both research prototypes have specific advantages and disadvantages with respect to their practical applications.

(c) Is it more appropriate to use quantitative or qualitative methods in educational research? This debate has almost become a religious war for many, even though the theory of science convincingly teaches that both approaches are necessary and complementary components of any system of research in the social sciences.

2. Theory and Research

These metatheoretical debates and the many concrete attempts to derive practical applications from research findings have led to a variety of different proposals for solving the research–practice problem. Two positions are especially characteristic of these proposals: The first is that theory should take precedence; the second is that empirical work should take precedence.

The first characteristic position, that theory should be preeminent, can be captured by a quotation from Mook (1983), who, in agreement with many other scientists, recalled the truth of the classical adage that nothing is more useful than a good theory:

> Ultimately, what makes research findings of interest is that they help us in understanding everyday life. That understanding, however, comes from theory or analysis of the mechanisms; it is not a matter of generalizing the findings themselves … The validity of those generalizations is tested by their success and predictions and has nothing to do with the naturalness, representativeness or even non-creativity of the investigations on which they rest. (p. 386)

Examples of the ways in which theoretical models have aided in the solution of practical problems include the use of attribution theory to provide a better understanding of students' intuitive explanations for their success and failure; the use of findings from the expert–novice research paradigm to estimate the relative importance of general abilities and domain-specific knowledge in the solution of demanding tasks; and the use of theories of prejudice in the analysis of social conflicts.

The second characteristic position, focused on empirical research, is invoked in instances where there are competing theories, where context conditions determine psychological processes, or where descriptive or causal models must be transposed into prescriptive statements. Sequential research strategies are proposed to solve such problems. One example is the description of 'six steps on the road from pure learning research to technological research and development in the classroom' suggested by Hilgard (1964, pp. 405–11): step 1—research on learning with no regard for its educational relevance; step 2—pure research on learning with content that is similar to school subject matter; step 3—research on learning that is relevant to education because the subjects are students, the material learned is school subject matter, and/or the learning conditions are similar to classroom situations; step 4—research on learning and instruction in special laboratory classrooms; step 5—use of experimentally tested learning and teaching strategies in normal classrooms; step 6—developmental work related to advocacy and adoption, for a wider use of the research-based educational strategies.

These and similar sequential research strategies are followed in many research and development centers around the world. The differences among the programs are, however, too large to allow conclusions about the success of such a strategy. In principle, the possibilities and limits of this approach also depend on the basic relation between educational research and educational practice. Good educational practice is not simply applied science, nor does it consist in the correct use of findings from applied research. Science can at best only provide an important and useful, but always limited, theoretical basis for the planning, analysis, control, support, reflection, and better understanding of educational practice.

The limits of scientific prediction and explanation of educational outcomes are expressed in the conclusions reached by Haertel et al. (1983) based on a comparison of different psychological models of educational performance. They concluded that:

> Classroom learning is the multiplicative, diminishing-returns function of four essential factors—student ability and motivation and quality and quantity of instruction—and possibly four supplementary or supportive factors—the social–psychological environment of the classroom, education-stimulating conditions in the home and peer group, and exposure to mass media. Each of the essential factors appears to be necessary but insufficient by itself for classroom learning; that is, all four of these factors appear required at least at minimum levels for classroom learning to take place. It also appears that the essential factors may substitute, compensate, or trade-off for one another in diminishing rates of return. (p. 75f)

3. Models for Bridging the Gap between Research and Educational Practice

There is no general rule that specifies correct, expedient, or appropriate ways to use scientific findings to solve the practical tasks of education. More importantly, there can in principle be no such rule. There are, however, various approaches available that show how research can be used in different ways and what strategies can be used to translate research into practice.

3.1 From Learning Theories to Teaching Technologies

Educational technology consists of systems of rules, tools, and activities designed to bring students to the point where they can achieve specific learning goals systematically, effectively, and economically. Even more, Glaser and Cooley (1973) perceive technology 'as a vehicle for making available to schools what psychologists have learned about learning' (p. 855). Such standardized technologies should, of course, be derived from explicit and tested scientific theory. Although not a necessary component of educational technology, technical devices are frequently used with the purpose of minimizing the negative consequences of fluctuation in attention to errors or mistakes, whether by students or teachers.

The first researcher who, from theoretical grounds, wanted to replace normal instruction in the classroom with almost revolutionary mechanical teaching techniques was Skinner (1954). In the early 1950s Skinner argued that the behaviorist models of learning and conditioning, and the psychological laws specifying how to shape behavior through contingent reinforcement of desired reactions, were the scientific prerequisites for successful and effective instruction. It was his belief that these principles could not be used in classroom teaching because teachers were not able to present the material to be learned in sufficiently elementary components and were not able to provide the requisite number of reinforcements in the necessary contiguity to learners' behavior, verbal or nonverbal. Thus, he recommended programmed instruction and the use of teaching machines.

Beyond Skinner's initial attempts, the development of educational technologies made dramatic progress in the 1970s, 1980s, and 1990s. Computer-assisted instruction, integrated learning systems, intelligent tutoring systems, instructional media, and intelligent learning environments are some of the key concepts referring to the development, study, and use of modern technical aids for improving educational practice.

In particular, modern computer technology offers a variety of possibilities for the technical design of controlled, but flexible, adaptive, and intelligently guided instructional systems and learning processes. The prerequisites for these systems are general learning theories and domain-specific models of knowledge acquisition that allow the following:

(a) analysis of the learning goal in terms of competent performance (specification of the learning task);

(b) description of the initial state of the learner (specification of individual differences in relevant skills and knowledge);

(c) determination of the methods that effectively lead to knowledge acquisition (specification of instructional methods);

(d) assessment of the effects of these instructional methods (specification and measurement of learning progress).

Although there is no disagreement that educational technologies provide an up-to-date tool for education as well as an excellent opportunity for employing research results in a practical way, the relative failure of educational computing in general, and in mathematics in particular, must be acknowledged. The high expectations supported by research that rose in the early 1980s, with respect to the potential of the computer as a lever for the innovation and improvement of schooling, have not been redeemed. Elsewhere De Corte (1994) has argued that a major reason for this relative failure of computers in education is that the machine has been mainly introduced as an add-on to an existing, largely unchanged classroom setting. In addition, due to a lack of good communication, teachers usually had only low expectations about computer support for their teaching. Also, there was in most cases no question at all of an orientation toward modifying teachers' conceptions about educational goals and their beliefs about learning.

3.2 Research on Teaching and the Acquisition of Teacher Expertise

A large part of educational research is directed toward investigating educational productivity, teacher effectiveness, and the relation between instructional quality and learning outcomes. Although the empirical findings are very impressive, they are also quite diverse and are not stable across different studies. In a synthesis of the effects of instructional variables on learning, the following rank ordering of noncontent-related factors was found (with decreasing effect strength): reinforcement, accelerated learning programs, cues and feedback, cooperation programs, personalized instruction, adaptive instruction, tutoring, higher-order questions.

If one considers that the effects of single instructional variables can be masked or mediated by combined, compensatory, and interactive effects from other teaching, student and context variables, it becomes apparent that training single teaching skills cannot be a practical implication of these research results.

It seems considerably more effective to provide a combined training program for teachers that uses the results from research on teaching. The paradigm of the expert teacher provides an interesting theoretical perspective for the practical application of such research findings. Prompted by work in cognitive psychology, this paradigm investigates how experienced and successful teachers differ from inexperienced and/or unsuccessful teachers in their knowledge about education and in their teaching activities. The research results from this paradigm have already been applied in the design of teacher training.

This paradigm allows the integration of two research traditions that have been portrayed as alternative approaches (Stolurow 1965): 'Model the master teacher or master the teaching model?'. In the expert teacher paradigm, it is possible to directly and systematically relate research on teaching and learning in the classroom to teacher training. Such training should not be directed toward producing a uniform type of 'expert teacher', but should explicitly assume that there are interindividual differences in the teacher competencies to be trained.

3.3 Studies of Classroom Learning

In contrast to the subject pools used in experimental investigations of learning, the classroom is best described as a group of students with very different

learning prerequisites who are to be instructed at the same time in the same place. Massive interindividual differences in cognitive abilities, knowledge, skills, motivation, interests, attitudes, and study habits affect both students' learning behavior and their learning achievements. The relevant research concerning these factors is very extensive. It has led to a large number of psychological theories about how personality states and traits are relevant to learning, and has contributed to the development of measurement models and measurement instruments for diagnosing individual differences in learning prerequisites and learning outcomes.

Carroll (1963) provided an important contribution to the practical application of this psychological research with his 'model of school learning.' Described simply, Carroll proceeded from the assumption that, given equivalent learning time, students with different aptitudes would diverge in their learning performance; that is, some students would not attain the required performance goal. To avoid this outcome, each learner must be allowed the learning time he or she needs to attain a specific learning goal.

This model has had a strong influence on subsequent research; in many respects it has been further specified, broadened, and extended, and it has also been used for a variety of different practical applications. For example, the following concepts and procedures can be traced to Carroll's original model: time on task (the time during which a student is actually involved in a learning task); the personalized system of instruction; individually prescribed instruction; mastery learning; and, largely independent of Carroll's model of school learning, the aptitude–treatment interaction (ATI) research program. All these approaches have been concerned with the question of whether and to what extent undesired effects of individual student aptitude differences on learning outcomes can be reduced by variations in instruction.

Research on the dependence of school achievement on aptitude differences and instructional conditions has led to important practical applications. Three of these are grouping students for learning, mastery learning, and adaptive teaching.

The educational benefits of grouping students for learning are controversial and seem to depend on the effects of a variety of factors. Furthermore, these benefits are both student-specific and criterion-specific. A successful practical application of ATI research has not been made.

Mastery learning is the attempt to ensure that as many students in a class as possible (90–95 percent) meet a required learning criterion (90–95 percent correct items in a criterion test). This goal is achieved by a limited increase in learning time (about 20 percent) and optimal instructional use of the learning time. Although what one can actually expect from mastery learning in normal school instruction is somewhat controversial, there is no doubt that this is a fruitful model that allows productive research and practical applications.

Adaptive teaching is based on many ideas from Carroll's model of school learning and ATI research. Adaptive teaching involves both learner adaptation to the environmental conditions presented and teacher adaptation of the environmental conditions to the learners' present state, to effect changes in that state. The teacher's goal, in other words, is to make environmental variations 'nurtured' rather than merely 'natural' for each learner concerned (Corno and Snow 1986, p. 607). 'Adaptive teaching' is thus an omnibus term covering a large number of different standardized and informal procedures for adapting teaching in variable, flexible, and dynamic ways to students' individual differences, so that optimal learning conditions can be provided for each student.

3.4 Fostering Students' Learning Competencies and Self-instructional Skills

In the above characterization of adaptive teaching it was noted that the optimization of individual learning requires not only that instruction be compatible with the cognitive and motivational demands of the student, but also that students' learning competencies and learning motivation be explicitly fostered. There is considerable research in cognitive psychology, developmental psychology, and educational psychology that addresses how to promote competencies and motivation. At the forefront are investigations on transfer of training, learning to think, the promotion of metacognitive skills, and the development of self-instructional skills.

Much of the theoretically oriented research also includes suggestions for applying the findings. Relevant support programs have been tried for elementary, secondary, and university students. Although these programs have demonstrated that intellectual skills and thinking performance cannot be arbitrarily trained, they have also convincingly shown that the training of metacognitive competencies and self-instructional skills can result in lasting improvements. When such programs have realistic goals, are provided over an extensive time frame, and if possible are directed toward the acquisition of domain specific knowledge, all students can profit.

3.5 The Use of Research Findings for New Models of Schooling

Research can be used not only to improve existing educational institutions, school organizations and instructional conditions, but also to plan new models for schooling. Such an approach to change is needed when it is unclear whether an educational goal can be reached by gradual and piecemeal changes in current

activities or whether fundamental reform is more necessary. Both theoretical discussion and practical experiences are available for using research findings in this way. Such change requires considerations of the factors that must be considered in planning a new educational system, the strategies that should be employed, the problems that must be addressed, and the implementation tasks that must be performed.

3.6 Research as a Source of Background Knowledge for Practitioners

The practical application of research consists in more than the instrumental use of research findings. In addition, science and research have an educational function; that is, they provide individuals with knowledge about themselves and the world, and allow them to act rationally.

Because people think, decide, and act primarily on the basis of their available knowledge (whether it is correct or incorrect, complete or incomplete, objective or biased), it is important and necessary for those involved in the educational process (politicians, administrators, principals, teachers, parents, students, and researchers) to be able to change or replace their personal beliefs, intuitive knowledge, prejudices, and suppositions with reliable and valid information. Such a replacement is not only a prerequisite for reflective, responsible, and effective individual action, but also a condition for rational communication and discourse among educational practitioners. Although it is not possible to measure the effectiveness of this function of research, it nonetheless seems plausible and important to translate as many research results as possible into the language of the practitioner.

A potentially very promising strategy to bridging the theory–practice gap is a combination of the last two strategies, namely using research for the design, implementation, and evaluation of new models of education and schooling, on the one hand, and research as a source of background knowledge for teachers on the other. This strategy consists in the creation and evaluation in real classrooms of complex instructional interventions that embody our present (hypothetical) understanding of effective learning processes and powerful learning environments. In order to have a reasonable chance of success, such attempts at totally changing the classroom environment and culture should be undertaken in partnership between researchers and knowledgeable practitioners.

Such partnership is essential not only to promote good two-way communication, but also in view of modifying and reshaping teachers' beliefs about education, learning, and teaching (see De Corte 2000). The positive effects of this partnership idea have, for instance, been shown by Huberman (1990). In a qualitative study of the linkages between researchers and practitioners, he has shown that contacts between both parties during a research project can result not only in better applications of the outcomes of a single study, but also in the establishment of more enduring forms of collaboration.

A specific approach which is in line with this holistic strategy is represented by scholars who advocate the use of so-called design experiments, and who aim at the development of a design science of education. According to Collins (1992, p. 15) 'a design science of education must determine how different designs of learning environments contribute to learning, cooperation, motivation, etc.' As a result a design theory would emerge that can guide the implementation of educational innovations by identifying the variables influencing their success or failure. This intervention approach has a twofold goal: it intends to advance theory building, while at the same time contributing to the optimization of classroom practices.

In this respect, Brown (1994) argued that theory building is crucial for conceptual understanding as well as for practical dissemination. In fact, Brown's project 'Fostering communities of learners' is one of the most representative examples of the design experiment approach. In this project the classroom is fundamentally redesigned using innovative components such as reciprocal teaching, the 'jigsaw method' of cooperative learning, and the creation of a new classroom culture. Reciprocal teaching establishes a learning environment in which a small group of students is provided with guided practice in acquiring four cognitive strategies with a view to improving their skill in reading comprehension. These strategies are: asking questions, summarizing, clarifying comprehension difficulties, and making predictions about the future content of a text.

A major feature of reciprocal teaching is that it takes the form of a dialog, in which the teacher and the pupils take turns in leading the discussion; gradually the pupils' share and responsibility in the dialogs increases. The 'jigsaw method' implies that small research groups of pupils each study a subtopic of a broader theme. Afterwards they regroup into learning groups in which each pupil is the expert in one subtopic, so that by combining their knowledge they all learn about the whole topic. A new classroom culture and climate is created based on four qualities, namely individual responsibility, mutual respect, establishment of a community of discourse, and a restricted number of participant structures (such as reciprocal teaching).

In the perspective of the further dissemination of this new kind of learning environments it is necessary to keep in mind that they should be realizable in existing classrooms. In this respect the notion of a partnership between researchers and practitioners is also crucial in the light of the necessity of the research–practice reciprocity. Whereas practitioners can help in translating theory into practice, and thus in making classroom teaching more research-based, their

partner role can also contribute to make research more practice-driven.

Finally, the orientation of this intervention approach to theory building raises a major challenge for the researchers, namely the development of an appropriate methodology for designing experiments in complex classroom settings in such a way that sound and theoretically valid conclusions can be drawn from a wide range of quantitative and qualitative empirical data.

4. Conclusions

While the effectiveness of these six different approaches to the problem of bridging the theory–practice gap has not been systematically studied, it is plausible that each of them has not only some merits but also some weaknesses. Moreover, much will depend on how each approach is used in translating research results into practice. In this respect, it can be argued that to be successful in making psychological theory and research applicable to education, one should develop a strategy that combines the following basic characteristics:

(a) good communication with practitioners, which means that the relevant outcomes are translated in such a way that they are palatable, accessible, and usable for the teachers;

(b) an orientation toward a fundamental change of teachers' belief systems about the goals of education and about good teaching and effective learning (in line with the view described above);

(c) a holistic (as opposed to a partial) approach to the teaching–learning environment, i.e., all relevant aspects of the learning environment should be addressed.

The relation between research and practice is complex and difficult. As a consequence, there is no simple, uniform solution to the problem of translating scientific results into practical suggestions. Nonetheless, some general conclusions can be drawn.

First, many different types of research are applicable to educational practice. The value of a scientific study for educational practice is not in any way directly related to the extent to which it mirrors typical features of an applied educational setting.

Second, research results can be used in educational practice in different ways. Six typical examples have been noted: the development of educational technologies, teacher training, the optimization of instruction, fostering students' learning competencies, the design of new models of schooling, and the use of scientific information as background knowledge for practitioners.

Third, these different strategies for applying research results are not mutually exclusive. Rather, it is desirable to combine several variants to improve the practical situation.

Finally, exploring the processes by which theoretical knowledge can be transformed into practical action must itself be an important domain of educational research.

See also: Comparative Research in Education: IEA Studies; Educational Innovation, Management of; Educational Policy, Mechanisms for the Development of; Educational Research and School Reform; Science and Development; Self-regulated Learning; Teacher Expertise

Bibliography

Anderson L W, Burns R B 1985 *Research in Classrooms*. Pergamon Press, Oxford

Bloom B S 1966 Twenty-five years of educational research. *American Educational Research Journal* **3**: 211–21

Brown A L 1994 The advancement of learning. *Educational Researcher* **23**(8): 4–12

Carroll J B 1963 A model of school learning. *Teacher College Record* **64**: 723–33

Collins A 1992 Toward a design science of education. In: Scanlon E, O'Shea T (eds.) *New Directions in Educational Technology*, *NATO-ASI Series F: Computers and Systems Sciences*. Springer-Verlag, Berlin, Vol. 96, pp. 15–22

Corno L, Snow R E 1986 Adapting teaching to individual differences among learners. In: Wittrock M C (ed.) *Handbook of Research on Teaching*. Macmillan, New York, Vol. 3, pp. 605–29

De Corte E 1994 Toward the integration of computers in powerful learning environments. In: Vosniadou S, De Corte E, Mandl H (eds.) *Technology-based Learning Environments: Psychological and Educational Foundations (NATO ASI Series F: Computer and Systems Sciences*. Springer-Verlag, Berlin, Vol. 137, pp. 19–25

De Corte E 2000 Marrying theory building and the improvement of school practice: A permanent challenge for instructional psychology. *Learning and Instruction* **10**: 249–66

Glaser R, Cooley W 1973 Instrumentation for teaching and instructional management. In: Gage N L, Travers R M W (eds.) *Second Handbook of Research in Teaching*. Rand McNally, Chicago, pp. 832–57

Haertel G D, Walberg H J, Weinstein T 1983 Psychological models of educational performance: A theoretical synthesis of constructs. *Review of Educational Research* **53**: 75–91

Hilgard E R 1964 A perspective on the relationship between learning theory and educational practice. In: Hilgard E R (ed.) *Theories of Learning and Instruction. The Sixty-third Yearbook of the National Society for the Study of Education*. The National Society for the Study of Education, Chicago, Part 1, pp. 402–15

Huberman M 1990 Linkage between researchers and practitioners: A qualitative study. *American Educational Research Association* **27**: 363–91

Mook D A 1983 In defense of external invalidity. *American Psychologist* **38**: 379–87

Rein V 1893 *Basic Principles of Education*. Göschen, Stuttgart, Germany

Skinner B F 1954 The science of learning and the art of teaching. *Harvard Educational Review* **24**: 86–97

Stolurow L M 1965 Model the master teacher or master the teaching model. In: Krumbholtz J D (ed.) *Learning and the Educational Process*. Rand McNally, Chicago, pp. 223–47
Thorpe L P, Schmuller A M 1954 *Contemporary Theories of Learning*. Ronald Press, New York

F. E. Weinert and E. De Corte

Educational (School) Counseling: United States

At the outset it is important to note that this article is focused on counseling in elementary, middle, and high schools. The obvious omissions from this list include postsecondary institutions such as junior colleges, community colleges, vocational–technical schools, four-year colleges, and universities. These institutions were not included in this discussion because their historical antecedents and current status are quite different from those of counseling in other educational institutions and the space allocated to this piece does not allow for a discussion of counseling services in both groups of institutions.

1. *Historical Perspectives*

It seems likely that the delivery of guidance and counseling services in educational settings began in the USA near the end of the nineteenth century in San Francisco in the California School of Mechanical Arts. In 1895 George Merrill began offering services that included occupational exploration, career counseling, job placement, and follow-up services. Jesse B. Davis also pioneered in offering guidance services in high schools in Detroit and Grand Rapids, Michigan. Although his program also included vocational guidance, it also focused on character development and developing appropriate behavior (Miller 1968).

During the first two decades of the twentieth century, other programs were established in Cincinnati, New York City, and Seattle. However, one of the most important developments that emerged from this period actually arose from the work of Frank Parsons in the Vocational Guidance Bureau and his posthumous publication *Choosing A Vocation* (Parsons 1909). Parsons was a social reformer who was concerned about a number of groups including immigrants and school dropouts. He recognized the importance of making wise career choices and set several programs in motion to improve their decision-making processes. In his book he outlined a three-step model for this process: (a) develop a clear understanding of one's self, (b) acquire a comprehensive knowledge of occupations, and (c) using what Parsons called 'true logic,' choose an occupation that is suited to one's abilities, aptitudes, and interests. Parsons' simple model of occupational choice making, underpinned by the development of numerous psychometric instruments in the 1920s and 1930s was to underpin the vocational counseling and guidance process in the USA and elsewhere until the 1950s. In the 1950s, Super (1951), Holland (1997), and others published more sophisticated models of occupational choice and career development that would change the way counselors think about the process. In reality, much of what passes for career counseling in schools today still follows Parsons' model.

Most of the early developments described thus far have been centered in high schools. Although many people advocated the need for guidance services for children, activity in this area was limited to a few school districts such as the Baltimore city schools. In 1964, funds for elementary school guidance programs and the training of elementary school counselors became available because of amendments to the National Defense Education Act and since that time elementary school counseling programs have increased rapidly in the USA.

1.1 *Directive Approach*

To this point the general term guidance has been used more often than counseling to describe the services provided to students in educational institutions. Today, many school counselors in the USA are still called guidance counselors because of the early emphasis on guidance services. Guidance services include testing and appraisal; providing educational, vocational, and personal information; classroom guidance; consultation with teachers and parents; educational and occupational placement; and individual and group counseling.

Although the term counseling appears in connection with many of the descriptions of services provided in the late nineteenth and early twentieth-century services, the process of counseling received little attention in the professional literature until 1939 when E. G. Williamson published *How to Counsel Students*. Williamson (1939) set forth a counseling model that could to some degree be linked to Parsons' (1909) tripartite model, although it had six steps. These steps were (a) collect data about the student and his or her environment using tests, inventories, observations, and other data collection devices; (b) collate and synthesize the data that has been collected; (c) diagnose the student's problem; (d) predict what will occur to the student based on the various choices available to the student; (e) provide counseling; and (f) follow up to determine how the student has fared given the course of action taken, and recounsel as necessary. Williamson's approach, which relies heavily

on appraisal, became known as the directive approach to counseling because of some of the counseling strategies recommended. For example, Williamson suggested that counselors identify the most appropriate course of action for students, and if the student rejected the suggestion, send them out to consider their options before proceeding.

1.2 Nondirective Approach

In the 1940s and 1950s Carl Rogers (1942, 1951) provided counselors with a very different approach to counseling students that soon became known as the nondirective approach. Rogers' philosophy and his counseling techniques were simple. He believed that people have within them an innate ability to come to terms with and solve their own problems. The counselor's role, according to Rogers, is to create a climate in which individuals feel free to explore their innermost self-perceptions and feelings. In Rogers' counseling framework, appraisal of the individual was largely eliminated, as was diagnosis. Not surprisingly, Rogers' optimistic view of human beings and their ability to cope with their problems was widely embraced.

Historical hindsight suggests that the enthusiastic acceptance of Rogers' ideas was not the wisest course of action that could be taken by counselors working in educational settings. Most school counselors are faced with serving large numbers of students and therefore must provide counseling sessions that are relatively short in duration and few in number. The client-centered model, although it was touted as a briefer model of counseling than the psychoanalytic approaches that held sway in the 1940s and 1950s, turned out to be an approach that required several sessions of fairly long duration. Moreover, counselors found that children needed more structure than the client-centered model provided and as a result did not respond well in counseling. Although many elementary school counselors incorporate aspects of the client-centered model into their work, other briefer, more action-oriented approaches to counseling appear to be more popular (Gibson et al. 1993), a fact that reflects the status of client-centered counseling with other school counselors working at other levels as well.

2. Current Status and the Future of School Counseling

The National Defense Act of 1958 and the increased emphasis on counseling that began with the acceptance of Carl Rogers' ideas in the late 1950s and early 1960s marked the beginning of a rapid movement toward professionalization by school and other types of counselors. The result was the development of accreditation standards by the Council for the Accreditation of Counseling and Related Educational Programs (CACREP), higher standards for licensure to practice in school settings by every state, and the publication of a variety of position papers regarding the role and function of school counselors by the American School Counselors Association. Today school counselors utilize dozens of models of counseling to assist students with the wide range of problems they bring to them. School counseling, which began in the USA, has been adopted to varying degrees by all English-speaking countries, most Western European countries regardless of language spoken, and some African and Asian nations. Interestingly, the career development needs of students in all countries, including the USA, will continue to shape the services offered to students both in the USA and in countries such as Japan, the People's Republic of China, and Taiwan. However, a number of other forces continue to influence school counseling including concerns about the family, efforts aimed at substance use and violence prevention, and concerted efforts aimed at educational reform.

2.1 Counseling

Baker (2000) indicated that the need for counseling in contemporary services is substantial, and presented a list of concerns that are addressed routinely in individual and group counseling sessions. The problems on Baker's list were aggressive behavior, underachievement, grief, suicidal ideation, eating disorders, substance abuse, stress, problem pregnancy, career planning, and coping with dysfunctional families. Earlier Conyne (1994) provided data that support to some degree Baker's observations. He reported that 25 percent of high school students in the USA drop out of school, suicide is the second leading cause of death among the 11–24 age group, and that AIDS has become one of the leading causes of death among 15–24 year olds. Other data, such as the report that one in five students report at least one bout of depression before age 18 (Lewinsohn et al. 1993), also lend credibility to Baker's list of problems.

It should be noted at this point that most school counselors assume a very limited role when counseling students who have severe mental health problems because they lack the time required and may not be trained to deal with severe mental health problems. However, there is support from both quantitative and qualitative research for both individual and group counseling that school counselors perform according to the conclusions of Whiston and Sexton (1998). They reviewed the research literature regarding school counseling interventions and concluded that students appear to benefit from social skills training, activities related to coping with families, and career planning activities.

Whiston and Sexton (1998) also reviewed the research literature focusing on what they called guidance curriculum activities and drew different conclusions. Generally speaking, research has not been supportive of the efficacy of classroom guidance activities, although there were some positive results. Classroom guidance activities devoted to the development of self-esteem were generally not effective according to Whiston and Sexton. However, they concluded that research regarding the use of classroom guidance to improve academic achievement was somewhat effective. At the conclusion of their review, Whiston and Sexton made two types of recommendations. One recommendation was that more and better research regarding school counseling interventions is needed. The second recommendation was that school counselors should strive to improve the quality of their interventions.

2.2 Classroom Guidance

Given the rather limited support for classroom guidance, should school counselors abandon it as a strategy and use their time to provide additional counseling activities? The idea of eliminating large group interventions seems untenable since school counselors are faced with providing services to 350 to 1,000 students and the use of a large group delivery mechanism is an efficient means of reaching large numbers of students. Whiston and Sexton's (1998) recommendation was on target. There is a need for the improvement of the delivery of classroom guidance activities.

Another line of research, that dealing with resiliency, provides an even more compelling rationale for the improvement of classroom guidance, counseling interventions, and other strategies used by school counselors. Resiliency is the innate ability of children and adolescents to cope with numerous environmental difficulties and emerge with positive mental health (Lewis 2000). Although resiliency is innate, it seems to be fostered by a number of protective factors that school counselors can impact. Lewis (2000) reports that resilient individuals have a positive attitude toward school and develop good reading skills by age 10, have good reasoning and communication skills, believe that they can influence their own lives, set career success as a high priority, have special interests and hobbies, and have coping strategies when facing adversity. Factors that seem to promote resiliency include friendship-making skills, positive mentoring at school, positive affectional ties with alternative care givers, a caring environment, high expectations, and opportunities for participation. Clearly, the fostering of resiliency cannot be developed through a single intervention. However, social skills such as friendship-making skills, decision-making skills, coping with factors such as stress, study skills, and many other variables that positively affect resiliency, can be influenced if a variety of strategies are employed.

To this point only counseling and classroom guidance have been discussed. School counselors use a number of other strategies such as consultation to impact some of the other factors discussed.

2.3 Consultation

Consultation is a triadic process engaged in by a school counselor (consultant) and a caregiver (consultee) for the benefit of a student (client). The objectives of consultation are twofold. The most obvious of these is to improve the educational or psychological functioning of the student. The second goal of consultation is to improve the functioning of the consultee, who is typically a teacher or a parent. Although there is limited research on consultation with parents, the research that is available is generally supportive of this activity as is the rather extensive number of studies focusing on teacher consultation (Brown et al. 1998). Using consultation and parent education (Muro and Kottman 1995), school counselors help parents develop more positive parenting strategies and establish more positive, supportive family environments. Consultation with teachers typically focuses on academic achievement, disruptive classroom behavior, or a combination of the two. Depending upon the theoretical orientation of the counselor, interventions may range from individual behavioral contracts to entire classroom interventions employing democratic decision-making strategies.

2.4 Peer Intervention

Peer interventions have also been used successfully by school counselors, both as a means of preventing the development of problems as well as interventions to help students who have developed academic and psychological problems. For example, peer counseling programs have been employed successfully to improve communication skills, enhance decision making, teach relaxation strategies, and increase assertiveness (Baker 2000). Peer mediation programs have also been used successfully to reduce physical and verbal aggression in a large number of schools and peer tutoring programs have been instrumental in improving educational achievement among students at various grade levels (Brown 1999). Students have also acted as cocounselors in successful smoking cessation programs for children and adolescents as well as in other roles involving the prevention and elimination of addictions. The roles that school counselors play in peer counseling and tutoring programs include recruit-

ing peer counselors, mediators, and tutors; training them for their roles; supervising their activities; and evaluating the outcomes of their work.

2.5 Other Counseling Activities

Counselors at all levels utilize the interventions discussed to this point. However, educational planning and placement, career planning and placement, and scheduling students' classes are activities that are very much a part of high school counselors' roles, engaged in to a limited degree by middle and junior school counselors, and almost totally absent from the roles of elementary school counselors. Assessment for educational and career planning, including the interpretation of tests and inventories to students and parents, follows this same pattern, with elementary school counselors being only marginally involved in these types of activities and counselors at higher levels becoming progressively more involved. In the USA, school counselors have been caught up in the assessment of educational progress that characterizes the educational reform movement in many schools. The result is that elementary school counselors often coordinate the end-of-grade tests, while middle and high school counselors coordinate the end-of-course tests that are used to assess achievement in individual subject areas. High school counselors may also be involved in administering college admissions tests as well as general academic competency tests required for graduation by some schools.

3. Conclusions

If there is one role that is routinely ignored by large numbers of school counselors it is the evaluation of their work. Baker (2000) indicated that the increasing demand for accountability requires that the outcomes of school counseling programs be evaluated, but few counselors seem to be heeding his warning. The result is that a few schools are requiring that counselors conduct follow-up studies to determine opinions about the effectiveness of their programs. Even fewer schools require that specific interventions such as group counseling be evaluated periodically. If Baker is correct about the need to demonstrate the effectiveness of school counseling programs, conducting evaluation studies will become a more important part of the counselor's role in the future.

Finally, it is difficult to predict whether conducting evaluation studies will be an important activity within school counseling programs in the future because of the demands placed on their time by ever-changing educational institutions. It is even more difficult to predict the overall future of school counseling. The number of school counselors has increased dramatically since the 1960s and there is increasing interest in providing counseling services in schools throughout the world. However, counseling services come at a cost. As school officials set priorities designed to improve the educational experiences of students, school counselors will, as Baker has suggested, have to demonstrate that they are worthy of funding. If this is to occur attention must be paid to improving the quality of some of the interventions used, such as classroom guidance and presenting evaluative data that support the services provided.

See also: Adolescent Vulnerability and Psychological Interventions; Antisocial Behavior in Childhood and Adolescence; Counseling Psychology; High School Dropout; Mental Health Programs: Children and Adolescents; Rogers, Carl Ransom (1902–87)

Bibliography

Baker S B 2000 *School Counseling for the Twenty-first Century*, 2nd edn. Merrill, Upper Saddle River, NJ

Brown D 1999 *Proven Strategies for Improving Learning and Achievement*. Clearinghouse on Counseling of Student Services, Greensboro, NC

Brown D, Pryzwansky W B, Schulte A 1998 *Psychological Consultation: Introduction to Theory and Practice*, 4th edn. Allyn & Bacon, Boston

Conyne R 1994 Preventive counseling. *Counseling and Human Development* **27**: 1–10

Gibson R L, Mitchell M H, Basile S K 1993 *Counseling in the Elementary School: A Comprehensive Approach*, 2nd edn. Allyn & Bacon, Boston

Holland J L 1997 *Making Vocational Choices*, 3rd edn. Personality Assessment Resources, Odessa, FL

Lewis R E 2000 Resiliency: Pathways to protective factors and possibilities for self-righting narratives. In: Capuzzi D, Gross D R (eds.) *Youth at Risk*, 3rd edn. American Counseling Association, Alexandria, VA, pp. 41–75

Miller F W 1968 *Guidance Principles and Services*. Merrill, Columbus, OH

Muro J J, Kottman T 1995 *Guidance and Counseling in the Elementary and Middle Schools: A Practical Approach*. Brown & Benchmark, Madison, WI

Parsons F 1909 *Choosing a Vocation*. Houghton-Mifflin, Boston

Rogers C R 1942 *Counseling and Psychotherapy*. Houghton Mifflin, Boston

Rogers C R 1951 *Client-centered Therapy*. Houghton Mifflin, Boston

Super D E 1951 Vocational adjustment: Implementing a self-concept. *Occupations* **30**: 88–92

Whiston S C, Sexton T L 1998 A review of school counseling outcome research: Implications for practice. *Journal of Counseling and Development* **76**: 412–26

Williamson E G 1939 *How to Counsel Students*. McGraw-Hill, New York

D. Brown

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Educational Sociology

Educational sociology is a field of sociology that directs attention to the social forces that affect all aspects of education in a given society, or across many societies. As such, it complements the study of education from other scientific perspectives.

The term 'educational sociology' is thought to have been first used by the philosopher John Dewey at the end of the nineteenth century, and university courses began to appear with that title in the United States already in the early twentieth century.

However, in the late nineteenth and early twentieth centuries, the precursors of modern sociology had recognized the importance of the social context of education. This was especially true of Emile Durkheim, and to a lesser extent Max Weber, Karl Marx, and others who in one way or another included education in their analyses of society. Thus, the sociological study of education can trace its origins to the writings of the early European founders of sociology itself. Since its early beginning, educational sociology has been the site of major theoretical and methodological developments. It has also produced a very large body of research knowledge about education in different societies and cultures. Noblit and Pink conclude in their discussion of the field that it '…is even more varied and vibrant than other subfields of the (sociology) discipline' (Noblit and Pink 1995, p. 8).

The purpose of this article is to provide a brief description of the origins, the characteristics of, and the major issues current in educational sociology.

1. *The Sociological Perspective in the Study of Education*

From its early beginnings, the sociological study of education was viewed with optimism. In one of the first widely used textbooks in the United States, *The Sociology of Teaching*, Waller commented 'I am convinced that Educational Sociology is a fruitful and challenging idea, and I have tried to put the evidence for this belief into my book' (Waller 1932, p. viii). A sociological approach to the study of education is one that focuses on social factors rather than psychological, biological, or physical factors in the study of education. The social factors of interest to educational sociologists include those that are related to social structures as well as social processes. Thus, for example, the study of schools as social organizations and bureaucracies is a well-established field in educational sociology. This organizational context also provides the framework for studies of processes at the individual level, such as the acquisition of attitudes, values, identity, and academic achievement.

Among the many other topics included in educational sociology are the family and its relationship to the socialization of the young, and its relationship to schooling. Teachers and students also receive attention, and interests range from the study of the recruitment and training of teachers to the study of the teaching profession and of teachers in schools.

Educational sociology also includes macrolevel studies of educational systems, for example, of educational expansion, of education in different ideological systems, and of the relationship between class structures and what takes place within educational systems. The study of education and social change includes the link between education and social and economic development, the relationship between technology and education, and the factors related to the formation of educational policy and planning.

2. *Educational Sociology or Sociology of Education?*

One of the most uncomfortable divisions within the sociological study of education has been the extent to which the field is seen as a 'science,' represented by the dispassionate investigation of educational phenomena, or as the development of normative policies and plans to build particular kinds of educational systems.

In the 1960s this basic division was characterized as a tension between the empirical and normative 'modes' in the study of education. Even today there are many who see the sociological study of education primarily as a change agent in education, either through the formulation of policy, or through the effects of research itself. In this respect, the discipline is seen as a normative agent which is concerned with specifying what education or schools 'ought' to be or to do in carrying out the educative function. This point of view dominated the early development of the field, at least in the United States. This approach was sometimes specifically labeled as 'educational sociology.'

However, there were those who viewed the sociological study of education as a science. The label 'sociology of education' has often been used to designate this perspective, and this term has its roots in the sociological tradition that attempted to emulate the natural sciences and create a 'science of society.'

The significance of the names educational sociology and sociology of education became an issue in the United States with the decision to change the name of the *Journal of Educational Sociology*, which was founded in 1927. When the journal came under the sponsorship of the American Sociological Association in 1963, it was decided that its name be changed to *Sociology of Education*.

Today sociologists and educational researchers tend to use the two terms interchangeably. The designation of the field to some extent depends on disciplinary location, such as in departments of sociology, where it

is largely known as sociology of education or in departments of education where it is often called educational sociology. Furthermore, textbooks and courses sometimes are given other labels such as 'social foundations of education,' 'social context of education,' and 'education and society.'

Language and culture also may affect the way that the field is labeled. For example, in German the sociological study of education is related to the three words *Soziologie* (sociology), *Bildung* (education), and *Erziehung* ('to bring up'), and the field is frequently referred to as *Bildungsoziologie* or *Erziehungsoziologie* (Chisholm 1996).

In this article, the terms educational sociology and the sociology of education will be used interchangeably without implying that there are subtle meanings that differentiate between the two terms.

3. *Some Classical Theoretical Contributions to Educational Sociology*

Virtually all social thinkers, since the time of Plato and Aristotle, have considered the education of the young as part of their theories of society. From a social point of view, the process of education has always been linked with other social processes, either as a source of social stability or a source of social conflict. However, the discipline of sociology itself did not emerge until the early nineteenth century with the writings of August Comte (1798–1857). Many of these early sociologists included education in their studies of society.

The formal sociological study of education did not emerge until Emile Durkheim (1858–1917) began to teach, and write about, education and educational institutions to students at the Sorbonne where he was made Professor of Pedagogy and Sociology in 1913. Durkheim quite rightly can be called the father of educational sociology. As a teacher of teachers at the Sorbonne in Paris, Durkheim considered education to be '...an eminently social matter' and therefore highly suitable for sociological analysis. In his introductory lecture upon taking up the Chair of Pedagogy in 1902, Durkheim began with the following words: 'As a sociologist, it is above all as a sociologist that I shall speak to you of education (1956, p. 114). Later in the lecture he added: '…if there was ever a time and a country in which the sociological point of view was indicated, in a particularly urgent fashion, for pedagogues, it is certainly our country and our time' (1956, p. 133).

Durkheim's contribution to the sociological study of education lay primarily in his focus on the relationship between education and society. He believed that in every society there is a notion of the 'ideal person' and that every society had some form of education system to produce the ideal. This notion of the ideal person varies from society to society, and even in the same society over time. But, according to Durkheim, the education system adjusts to meet this need. Because of the importance of education for the survival of society, Durkheim regarded all education as moral education. For him, morality was the set of duties and obligations that influence the behaviors of individuals. He saw modern morality as resting on reason rather than religion, and he constructed a theory of educational practice that included what he considered the three basic elements of morality, namely discipline, attachment to social groups, and individual autonomy. Unlike his major works in sociology, Durkheim's works in education are collections of his course lectures and include letters and book reviews.

The importance of the social aspects of education did not escape other early sociologists. For example, Max Weber (1864–1920) included education in his analysis of social stratification, and educational credentials formed an important part of his understanding of social class, social status, and the formation and maintenance of elites in society. Although Karl Marx (1818–83) never engaged in a systematic study of education, it was nevertheless an important social institution in his grand theory about class exploitation and class conflict. Unlike many other fields in sociology, the sociological study of education can trace its origins to the writings of the early European founders of the discipline itself.

4. *Some Modern Theoretical Contributions to Educational Sociology*

In the 1970s educational sociology began to move into the mainstream of sociological debate in both theory and research methodology. At that time the sociological study of education was criticized as being too conservative and constrained by the priorities and preoccupations of educators rather than those of sociologists. At the same time, sociologists were criticized as being too preoccupied with analyses of classroom interaction on the one hand and the relationship between education and status attainment on the other.

In the late 1960s and early 1970s the microlevel study of educational processes was developing in the United Kingdom. During this period Bernstein (1971) conducted research on language codes and learning, and pointed to the hitherto ignored symbols and meanings in the classroom situation as a possible locus of inequality in educational processes. Bernstein's research gradually led him to focus on the ways knowledge is valued, classified, and controlled in various kinds of pedagogic processes.

In the minds of some, this 'new' sociology of education, which included antipositivist sociological

movements such as symbolic interactionism, phenomenology, ethnomethodology, Marxist sociology, and the sociology of knowledge, constituted a paradigm shift in the sociology of education comparable to a scientific revolution in the Kuhnian sense. Perhaps the most important aspect of the 'new sociology of education' is that it directed attention to the study of classroom interaction, classroom language and the curriculum. Knowledge itself became the main focus of study, namely how knowledge was defined, controlled, and transmitted to students. During the 1970s and 1980s the 'new sociology of education' stood in opposition to the 'old' and was manifested particularly in resistance theory and social and cultural reproduction theories. Although it is clear that developments in educational sociology have moved into other 'newer' directions (Noblit and Pink 1995), the influence of the 'new sociology of education' continues in much of current theoretical debate and research.

A third new direction in educational sociology emerged in the United States. Sociologists began to use causal models to investigate the multiple determinants of educational outcomes, and subsequently, the effect of educational attainments on occupational and career attainments. Although the use of highly quantifiable data and statistical techniques was not new in the sociological study of education on either side of the Atlantic, the development of status attainment models added social-psychological variables to the educational and occupational attainment process. The use of causal (path) models has dominated aspects of educational research from the 1970s to the present day in many countries.

A fourth and more recent development in educational sociology is that of critical theory, particularly with respect to the study of education as a form of oppression by dominant groups, or education as means of liberation from oppression. Critical theory emerged out of the Frankfurt School in Germany in 1923. During the ensuing years, including a period of dispersion during the Second World War, members of the Frankfurt School developed a range of theoretical approaches which were a reaction to the scientism of positivist sociology, and an attempt to develop a general theoretical approach which combined theory and practice. These writers included Horkheimer, Adorno, Fromm, Marcuse, and Habermas. Although the focal points and precise details of the writers vary, the general perspective and its implications for education are coherent and relevant. Critical theory is based on the assumption that the power of capitalism has come to dominate all aspects of social life. Concepts such as 'the totally administered society,' 'one dimensional man,' and 'communicative competence' typify the portrayal of social life in modern capitalism. The purpose of critical theory is not simply to explain and analyze, but also to emancipate and transform society in the process (the unity of theory and practice).

By the late 1970s educational sociology had moved into the mainstream of sociological debate generally. These debates became so important that some writers suggested that educational sociology had become a battleground for theories and policies about inequality and social reform. By the mid-1980s educational sociology had become very much a mainstream area in the discipline of sociology generally.

5. *Educational Sociology in Various Countries*

Early surveys of research in educational sociology, for example in the Federal Republic of Germany, France, and Britain, suggest that the functionalist perspective was dominant during the 1950s and 1960s. However, during this period it was in France that Marxist approaches were developed.

There is now growing documentation about the current development of educational sociology in Western countries. Surveys of research in educational sociology, for example, in the United States (Dreeben 1994), the former Federal Republic of Germany (Chisholm 1996), the Netherlands (Wesselingh 1996), and South Africa (Muller 1996), to name a few, present a picture of a rapidly growing discipline with shifting and multiple paradigms.

In many countries educational sociology was late in becoming established. In India, for example, the earliest identifiable studies in educational sociology were conducted in 1953 and 1967 (Chitnis 1982). In 1974 there were only 89 educational studies that could be classified and reviewed as sociological, but by 1980 the number had more than doubled.

Also in Spain, like many other countries outside the United States and Britain, educational sociology was slow in getting established. Early studies occurred in the 1960s and were typified by an emphasis on human capital theory, particularly in relation to economic development in Spain (Subirats 1990). Spain is an example of a country where sociology was not favored politically. However, after 1970 educational sociology began to develop in more radical ways. Much influenced by French sociology, educational sociology since the mid-1970s has been dominated by the neo-Marxist perspective, with research being conducted on schooling and social reproduction, into the effectiveness (or noneffectiveness) of schools, the link between school and work, and particularly on the role of universities in social reform.

6. *Themes and Developments in Educational Sociology*

Although educational sociology has branched into a number of identifiable subfields such as the sociology of the curriculum and the sociology of teaching, there

are a number of themes that have dominated both theory and research across all specialties. In this section a number of these themes are described in terms of their relevance for educational knowledge and research.

6.1 Macro/Micro Perspectives and Structure and Agency

Since the 1970s, one of the major issues in sociology generally has been the dualism between macro and micro approaches to the study of human behavior. Macro approaches typically focus upon aspects of social structures as explanatory variables, and thus tend to adopt deterministic and constraining theories about human behavior, with little or no acknowledgment of individual autonomy or freedom. Typical variables for the macro-oriented sociologists are those of social class, gender, ethnicity, capitalism, and organizational hierarchies. Conversely, micro-oriented sociologists tend to focus upon individual actors, regarding them as autonomous actors with few or no constraints by the outside social and structural environment. Typical variables for the microsociologists are the subjective meanings of human interaction, resistance, action, and voluntarism.

In the study of education this has meant an almost irreconcilable gulf. Those adopting a macroperspective have tended to study the statistical relationships between variables, largely measuring structural characteristics such as socioeconomic status, social class, gender, ethnicity, and the like. Those taking a microapproach have focused on interaction patterns, individual attitudes, values, or beliefs, or the subjective meanings that individuals construct about educational realities. The impasse between the 'two sociologies' has prompted some sociologists to observe that the only solution is to continue to develop both macro- and microsociology.

Related to the macro/micro controversy, and perhaps an obstacle to solving it, has been the question of structure and agency. The conceptualization of these two dimensions of social action also represents a dualism in the sociological study of education. This dualism focuses around explanations of social action: large-scale structural (and therefore deterministic) forces or individual voluntarist forces. Thus, social action is seen as existing on two distinct levels. Examples of the former in the sociology of education are structuralist neo-Marxists, who explain social reproduction in schools in terms of the contradictions of the capitalist economic system. Regarding the agency frame of reference, those who adopt an interpretive perspective tend to see the source of action at the individual or the microlevel, but with a reference to the structural context.

6.2 Critical Theory, Critical Pedagogy, and Postmodernism

The sociological study of education has been influenced by the intellectual movements of the late twentieth century, such as postmodernism, deconstructionism, and critical pedagogy. These intellectual movements approach the study of education from a critical, oppositional, and sometimes neo-Marxist perspective.

Critical Pedagogy, derived from the critical theory perspective, advocates that the classroom should become the arena for '…intellectual resistance towards the ways in which our roles, our lives, and our subjectivities are defined and constituted' (Shapiro 1995, p. 191). Ultimately those who espouse critical pedagogy in classrooms seek to establish a truly democratic culture through a form of education that is empowering and liberating in terms of the dominant ideology.

Modernism refers to the industrial revolution, the artistic and literary movements of the twentieth century, and the beliefs and values of the Enlightenment, such as the idea of progress, and belief in reason and principles such as equality, liberty, and justice. In contrast, postmodernism rejects linear views of progress, relies more on a dialectical analysis of the opposing forces in society, and stresses a democratic, emancipatory and antitotalitarian theory and practice, particularly with respect to education.

Postmodernism, as a social theory and intellectual movement, differs from critical pedagogy and critical theory in that the notion of an underlying reality is denied. The notion of a postmodernist break or rupture with modernity has occurred through the penetration of the media, advertising and television. The surface signs (language) take on their own meaning. It has been argued that postmodernist social theory makes the notion of intellectual resistance and liberation meaningless (Shapiro 1995), and thus with little educational relevance.

As theoretical movements, it remains to be seen as to how these two will influence the social understanding of educational processes and educational policy and practice in the twenty-first century.

6.3 Equality and Excellence

In most Western societies, the concept of equality has been central to much of the planning for and expectations from educational systems. Generally speaking, structured social inequality is defined as a social problem, and education is usually seen as a way of resolving it. Furthermore most modern democracies espouse the principles of meritocracy in educational opportunity and social rewards. In this context, education plays a double role, for not only is it

expected to provide a major mechanism for bringing about greater equality, but it also serves as a mechanism for self-betterment, and therefore for social mobility.

Clearly then, since education is linked with equality and inequality, it is essential to clarify what is meant by the term. Two seemingly contradictory dimensions underlie notions of equality with respect to education. The first of these includes the notions of fairness and equity, and generally means that 'random inequalities of nature' should be acknowledged and taken into account in the educative process. Thus, naturally bright children should be treated in ways which will allow that brightness to be developed to its fullest. Not to do so would be unequal and inefficient.

However, there is a second dimension, and that is that education should help to rectify (mitigate) the 'random inequities of nature.' In other words, insofar as is possible, every person should have an equal opportunity to excel in ways that are not constrained by natural differences. Sociological research in education has addressed the origins of inequality in education, the consequences of inequality, and policies that may make educational access and outcomes more equitable. Much of the research into the importance of family background and schooling in educational and occupational attainments has the question of inequality underlying it. Ultimately, the extent to which education either creates or reinforces these inequalities means that it operates in an inequitable manner. Thus, the subfields of adult education, the curriculum, learning, special education, and teaching are all concerned, in one form or another, with equality and efficiency in the educational system.

6.4 *Gender and Education*

The process of differentiation by sex in education occurs early in schooling and continues throughout the education process. In most countries, the attainments and achievement of males and females at different education levels, and in different subject areas, differs considerably. In a study of science achievement in ten countries (Australia, England, Finland, Hungary, Italy, The Netherlands, Sweden, Thailand, and the United States), between 1970 and 1984, Keeves and Kotte (1992) found that the social roles of women changed in all countries with respect to participation in the labor force and declining fertility rates. They also found that the participation of women in secondary and tertiary levels had increased, and in some countries, the participation rates for women were higher than that for males.

The results of their study showed that in all 10 countries boys outperformed girls in science, even when other variables were taken into account, such as home background, aptitude of student, values and motivation to do science, and attitudes toward science. The authors concluded, however, that their model left much unanswered in explaining why it is that boys do better in science than girls: '…it is not firmly established exactly how the effects of being male or female operate to influence achievement outcomes in science' (Keeves and Kotte 1992, p. 163). Although it can be argued that sex differentiation has its origins in the family socialization processes during infancy and childhood, the school remains a key institution that accounts for some unequal educational outcomes. The school both constructs and reconstructs gender relations such that the subordinate status of females is continually confirmed. The fact that females are staying in school longer in many countries does not mean that gender differences are being eradicated. Questions of gender in education apply to teachers as well as students. Sex inequalities in student academic and career outcomes have often been related to the influence, subtle or otherwise, of the sex of teachers. More importantly the sex inequalities among teachers regarding their work experience and career opportunities have become a topic of importance for understanding the influence of gender in education generally.

Continued research into the processes of maintaining sex divisions in schooling will remain high on the agenda of educational sociology. Furthermore, social policies to eradicate gender inequalities in schooling can only become effective when they are based on sound research evidence, which until now remains inadequate.

6.5 *Cultural Diversity and Multicultural Education*

The movement of peoples across and between national boundaries continues to pose important research issues for educational sociology. The schooling of minority groups within countries has been a long-standing item on the sociology research agenda, particularly in countries such as the United States, Britain, Canada, and Australia. In these and similar countries the emergence of heterogeneous populations through voluntary and nonvoluntary migration has called into question again whether education serves as an assimilationist mechanism or a protector of unique cultural systems and identities. In reality the debate over the notion of a core curriculum and a multicultural curriculum has yet to be resolved, particularly with respect to learning among members of race and ethnic minorities (Ogbu 1992).

Considerable research has been conducted on the factors that contribute to the academic achievement of minority groups in various countries. However, these studies have often assumed homogeneity within minority groups themselves, and have also assumed that members of minority groups share similar values

toward education and its outcomes. However, research findings make it clear that this is not the case, and that for indigenous groups and other racial and ethnic minorities, 'differentness' and 'otherness,' both between and within groups, are major factors in explaining educational processes and outcomes.

Ogbu (1992) has argued that both countries which advocate a strong core curriculum (as in Germany, Japan, South Korea, and Taiwan), and those which support a more pluralistic multicultural curriculum, overlook the wider societal and dominant cultural impact on the learning processes of minority students.

Educational sociologists will focus more on the processes related to the education of refugees, to the consequences of new international economic and social alliances, such as the European Community, and improved knowledge about the determinants of academic achievement across and within minority groups of all types.

6.6 *Dropouts and Burnouts: Alienation in Schools*

School retention rates have steadily increased since the 1960s in almost all countries. Nevertheless, there will remain considerable concern over those students in the twenty-first century who do not complete secondary school. There are numerous reasons for this concern. Rumberger (1987) gives five: (a) the short-term increase in dropouts since 1968, at least in the United States, (b) the numbers of minority group students with high dropout rates are increasing, (c) as academic standards increase, dropout rates will also increase, (d) the standard of educational credentials for virtually all jobs can be expected to increase, and (e) completion rates will increasingly be used as an indicator of performance for secondary schools.

Most of the research on school dropouts has regarded the phenomenon as pathological, and as representing a loss to the individual as well as to society. Lee and Burkham (1992) group the causes of dropping out into three categories: (a) the personal and psychological characteristics of dropouts, (b) the academic and psychological behaviors of dropouts, and (c) schools as contributors to dropping out. Thus, they argue that dropouts (or 'at risk' students) tend to come from low status, minority backgrounds, from single parent or step-parent households, are alienated, make lower grades, have higher rates of truancy, and do not attend schools (such as Catholic schools) which have lower dropout rates (1992, pp. 422–3).

LeCompte and Dworkin (1991) draw similar conclusions, but direct attention to the equivalent phenomenon among teachers, namely burnout. They argue that teachers also experience alienation and often leave the teaching profession because of burnout. However they observe that many teachers who experience burnout and lose their commitment do not leave the profession, a process they call teacher entrapment. This phenomenon poses two problems for the quality of teaching in schools: (a) new programs are difficult to implement with unenthusiastic teachers, and (b) staffing becomes a problem when teachers who should leave, do not. The authors suggest that school reforms fail because of inappropriate measures of school success or failure, of mismatched reforms, of quick-fix reforms, or reforms that leave the alienating structure unchanged. In short, the problem of alienation in schools, both from individual and structural causes, results in problems that are only beginning to be researched. Educational sociologists in most countries will continue to research these processes as long as there are school dropouts and burned-out teachers.

7. *Educational Sociology and Educational Policy and Planning*

Educational planning generally is understood to be the identification, development, and implementation of strategies designed to attain, efficiently and effectively, the educational needs and goals of students and society. The practice of educational planning is not a new phenomenon and, according to some, can be traced to the writings of Plato, to Renaissance scholars, and more recently to post-Second World War experts.

What characterizes current activities is that planners have tended to rely upon nonsociological assumptions in formulating their planning models. Demographic projections, manpower models, rate-of-return approaches, and school mapping all approach the task of educational planning as though the social aspects of educational behavior were irrelevant. Each of these techniques has at its base an implicit theory that assumes that the needs of society can be clearly specified and measured in economic terms, and that the social behavior of individuals in response to these needs is unproblematic. It is for this reason that writers argue that educational planning needs a social theory of society that includes structure and process, and both macro- and microdimensions.

We have learned from research and from all theoretical perspectives that the actors in the planning process, the politicians, bureaucrats, and planners are part of the very system they want to change. The functionalists can explain how it is that these persons act largely to preserve the status quo, even though the planning programs may appear very radical. The Marxists can explain in whose interests planning decisions are made, while the interactionists can explain at the microlevel how and why the actors in the planning process come to define planning objectives in particular ways, and how competing views are 'defined away.'

Finally, sociological research has helped us understand the importance of consultation with all interested parties in the educational planning process. It has been argued that the compensatory education and desegregation programs in the United States were flawed by lack of adequate information about objectives and methods. Resistance to any social reform can indicate conflicting interests which could have been avoided had adequate information been obtained beforehand. It is only by understanding these social forces that means can be taken to ensure that the objectives and strategies of educational planning are appropriately, efficiently and effectively formulated and implemented.

8. *Educational Sociology in the Twenty-first Century*

The rapid development of educational sociology during the late nineteenth and twentieth centuries has resulted in considerable accumulated knowledge about educational systems and how they work in different cultural and organizational settings. The discipline has become increasingly relevant for policy making, as well as for theoretical development and methodological application. It is also a discipline which is increasingly comparative, as knowledge about educational systems across national and cultural boundaries adds to both our knowledge of consistencies in educational processes, as well as the idiosyncratic properties of education in different structural and cultural settings.

Because it is problem-driven, the future of educational sociology will continue to reflect the preoccupations of countries in the performance of schools and the needs of society. In this respect, Durkheim was correct in arguing that education was imminently social. It will continue to be so, and will continue to attract the vigorous attention of sociologists and educational researchers who take a sociological perspective.

See also: Durkheim, Emile (1858–1917); Education and Economic Growth; Education and Employment; Education and Gender: Historical Perspectives; Education, Economics of; Education (Higher) and Gender; Educational Institutions and Society; High School Dropout; Human Capital: Educational Aspects; Social Inequality and Schooling

Bibliography

Bernstein B B 1971 *Class, Codes and Control. Volume 1: Theoretical Studies Towards a Sociology of Language*. Routledge and Kegan Paul, London

Chisholm L 1996 A singular history? The development of German perspectives on the social analysis of education. *British Journal of Sociology of Education* **17**(2): 197–211

Chitnis S 1982 Sociology of education in India: Emerging trends and needed research. In: Nayar P K B (ed.) *Sociology in India: Retrospect and Prospect*. B R Publishing, Delhi, pp. 155–83

Durkheim E 1922/1956 *Education and Sociology*. Free Press, Glencoe, IL

Dreeben R 1994 The sociology of education: Its development in the United States. *Research in Sociology of Education and Socialization* **10**: 7–52

Keeves J P, Kotte D 1992 Disparities between the sexes in science education: 1970–1984. In: Keeves J P (ed.) *The IEA Study of Science III: Changes in Science Education and Achievement: 1970 to 1984*. Pergamon, Oxford, UK, pp. 141–64

LeCompte M D, Dworkin A G 1991 *Giving Up on School: Student Dropouts and Teacher Burnouts*. Corwin Press, Newbury Park, CA

Lee V E, Burkam D T 1992 Transferring high schools: An alternative to dropping out? *American Journal of Education* **100**(4): 420–53

Muller J 1996 Dreams of wholeness and loss: Critical sociology of education in South Africa. *British Journal of Sociology of Education* **17**(2): 177–95

Noblit G W, Pink W T 1995 Mapping the alternative paths of the sociology of education. In: Pink W T, Noblit G W (eds.) *Continuity and Contradiction: The Futures of the Sociology of Education*. Hampton Press, Cresskill, NJ, pp. 1–29

Ogbu J U 1992 Understanding cultural diversity and learning. *Educational Researcher* **21**(8): 5–14

Rumberger R W 1987 High school dropouts: A review of issues and evidence. *Review of Educational Research* **57**(2): 101–21

Shapiro S 1995 The end of radical hope? Postmodernism and the challenge to critical pedagogy. In: McLaren P (ed.) *Postmodernism, Postcolonialism and Pedagogy*. James Nicholas Press, Albert Park, Australia, pp. 187–204

Subirats M 1990 Sociology of education in Spain. In: Giner S, Moreno L (eds.) *Sociology in Spain*. Consejo Superior de Investigaciones Cientificas, Madrid, pp. 169–73

Waller W 1932 *The Sociology of Teaching*. Russell and Russell, New York

Wesselingh A 1996 The Dutch sociology of education: Its origins, significance and future. *British Journal of Sociology of Education* **17**(2): 213–26

L. J. Saha

Educational Systems: Asia

Asia is a vast continent with a large population. It accommodates a vast number of countries with diverse cultures and religions, at different states of economic development and with different political systems. Education, being a human activity, therefore also demonstrates a spectrum of varieties in Asia. Nonetheless, there are certain detectable trends in education

Table 1
Countries in Asia

East Asia	China, Hong Kong (China), Taiwan (China), Macau (China), Japan, Korea (Republic of 'South Korea'), Korea (People's Republic of 'North Korea'), Mongolia
South East Asia	Brunei Darussalam, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand, Vietnam
South Asia	Bangladesh, India, Maldives, Nepal, Pakistan, Sri Lanka

This table follows largely UNESCO classification, with the exception of Iran

development in the region, given strong trends of globalization and increased interactions between nations and cultures.

This article refers to subregions which are often known as East Asia, South East Asia and South Asia (see Table 1). It does not include three other major groups of nations: systems that belong to the former Soviet Union, systems in the Middle East, and the Pacific Islands.

In most of these nations/regions, there have been dramatic developments in basic education in the last two decades of the twentieth century, approaching universal basic education (which is defined differently in different systems) with varied degrees of accomplishment amidst various difficulties. There were also deliberate developments of vocational education in the past two decades, but it faces varied degrees of challenges given the change in the economy and the labor market. Higher education is relatively underdeveloped except in some of the most developed economies.

1. *The Structures*

The education systems in Asia are very much influenced by the West. The earlier systems of education, even in noncolonial China, and Japan, were modeled after the old German system and later the US system. The $6+3+3$ system of primary, junior secondary and senior secondary schools has become the common pattern in Mainland China, Japan, Taiwan, and South Korea in East Asia.

With the exception of Japan and Thailand, most of the nations and regions in Asia experienced colonial rule, and that has significance in the structure of the education systems. Quite a few systems in Asia adopt a traditional British model that took shape during colonial times. Hong Kong, Britain's last colony in Asia until 1997, preserves a $6+3+2+2$ system, with a three-stage secondary schooling: two years junior secondary, two-year senior secondary, leading to O-Level (Ordinary Level) examinations, and an additional 2 years of pre-university A-level (Advanced Level) study. Singapore maintains a similar system, except that there are staggered years of secondary schooling according to students' abilities. The British model is still the norm in past British colonies such as Pakistan, Sri Lanka, Malaysia, Bangladesh, Maldives, and Nepal. India, which was the largest British colony, maintains the same pattern, except that it has an integrated eight years' basic education instead of nine.

Meanwhile, there are still traces of French influence in Vietnamese, Lao, and Cambodian education; these regions belonged to French Indochina. Likewise, Indonesia used to be a Dutch colony, and maintains a system that is more or less in line with the Dutch system. The Philippines were earlier colonized by Spain, and later mandated by the USA. They follow a very American system. Macao, a former Portuguese colony, has a blend of schools following the Portuguese and Chinese systems, respectively. In Mongolia and North Korea, the former Soviet model was followed.

2. *The Cultures*

Asia hosts a number of strong cultures. The strongest among them are the East Asian culture, the Indian culture, and the Moslem culture.

The East Asian culture, often also called the Confucian culture, prevails in communities that use chopsticks: the Chinese communities (the Chinese Mainland, Hong Kong, Taiwan, and Macao), Japan, the two Koreas, Vietnam, and a large part of Singapore. Although there are significant variations between, and even within, these societies, they are often seen as societies of collectivism and conformity. They are cultures in which education inherits the strong examination orientation in ancient China, where formal education was the only path for upward social mobility. Societies in East Asia still place high values on formal education. These societies therefore often enjoy a high attendance rate, and are champions in international comparisons of educational achievements, particularly in the realms of science and mathematics. However, the examination orientation has also rendered the systems rather weak in the relevance of the curriculum and flexibility in learning. Among others, the East Asian education systems also place serious emphasis on the social or moral aspects of students' development.

The Indian culture is an oversimplified umbrella term that covers different traditions derived from Hinduism. Against common beliefs, the South Asian

subcontinent societies encompass a large variety of religions, including Buddhism, Christianity, and Islam. However, it is safe to say that Hinduism underpins a large part of the education philosophy in South Asia. The education systems in South Asia often carry a strong religious orientation, and emphasize the relations between humans and nature. Because of the British colonial past, many of the South Asian systems of education also bear a strong framework borrowed from the traditional British system.

There is a strong tradition of Islamic culture in some of the South East Asian nations: Malaysia, Indonesia, and Brunei Darussalam. There are also sizeable Islamic communities in China, India, and Singapore. The Moslems in Asia seem to have developed a particular paradigm about education, which is able to integrate the emphasis on human–God relations in traditional Islamic education, with the learning of science and modern technology within the education system.

The cultural factor plays an important part in the shaping of education systems in Asian nations. Overall, there is an emphasis on the social (human–human) and religious (human–God) dimensions in the educational philosophy, and separation of such from knowledge and skills. Hence, in most of the education systems in Asia, there have been comprehensive and deep-rooted traditions of education about morality, attitudes, and values.

The strong cultural traditions in Asia have influenced not only education within the respective nations/regions, but have also affected the education of the Diaspora communities. The Chinese and Indian immigrants in other countries, for example, regardless of their countries of origin, are known for the outstanding academic performance of their children.

3. *The Economy and Education*

The continent also comprises nations in rather polar economic conditions. On the one extreme, there is Japan, which is among the strongest economies in the world. On the other extreme, there are the South Asian countries such as Nepal, Bangladesh, and Pakistan who are among the least-developed economies. In between, there are the newly emerged economic powers, the so-called little dragons: Singapore, Hong Kong, Taiwan, South Korea. There are the two giant economies: China and India; both are still in the 'developing' category, but are growing rapidly. Then there are the other South East Asian nations, once called 'advanced developing economies,' Malaysia, Thailand, the Philippines, and to some extent, Indonesia. Other nations have just recovered from the inefficiency of planned economies: Vietnam, Laos, Cambodia, Mongolia, and more recently, North Korea.

The economic strength of these nations affects their education systems in at least two dimensions. The economy constrains the resources available, and hence affects the supply of education. The economy also carries with it manpower structure and manpower requirements that affect the demand for education among the population. However, a special feature of education systems in Asia is that strong cultural elements often override economic conditions and enable many Asian nations to develop their education systems beyond that which the economy could seemingly warrant. Again, typified by China and India, followed by Sri Lanka and Mongolia, many Asian systems demonstrate a very strong literacy rate and enrolment ratio, when compared with systems in other parts of the world.

Most of the economies in Asia have seen positive growth in the last two decades of the twentieth century, despite intermittent interruptions by political instability or financial crises. One distinct move among such economic growth, common to the Asian economies with almost no exception, is the opening of the local market to foreign investments. This has sped up the process of globalization among Asian societies. Such a trend has tremendous implications for education, both in terms of human resources demand and of the philosophies in the development of education.

4. *Enrollment in Basic Education*

It was not a surprise that critical international endeavor on *Education for All* was launched in Asia (in Jiomtian, Thailand, 1990). Asia hosts the largest student population in the world. Among others, China and India alone host a total of 250 million primary school students, which is more than one-third of the world's students in primary education.

Population is always an issue in Asia. Population poses two types of problems for the education system: scale and growth. The sheer scale of the population in some of the Asian countries has made access to education an almost formidable problem. In addition, the problem is amplified by the rapid growth of the population in some nations.

There are very small systems of education in Asia, such as that in Brunei with around 40,000 students, and the Maldives, which has around 50,000, but they are not typical of Asian systems.

China and India host the two largest systems of education in the world, with primary school populations of around 140 million and 110 million, respectively. In both countries, the policy emphasis in the last two decades of the twentieth century has been on expansion of access to basic education. Both countries have launched major expansions of basic education since the mid-1980s (1985 in China, and 1986 in India). Both have demonstrated spectacular

improvements in access and school attendance; however, both countries are still struggling with access problems because of the size and growth of the population.

In China, there have been rigorous measures of birth control, which have effectively slowed down the growth in the population. There has also been rather successful mobilization of community resources in the building of schools. Hence universal attendance is achieved in most parts of the country. The national enrolment ratio for primary education, age 6–12, has been over 97 percent. However, there are pockets of population where either the economy is underdeveloped or the minority culture does not accept formal schooling as it is. There are both supply and demand problems of even basic education within these pockets. It has to be borne in mind that with the vast population in China, even a small percentage would mean millions of children out of school.

In India, there has been an unprecedented increase in the scale of basic education. There has also been great success in mobilizing nongovernmental organization of education, particularly in the nonformal sector. However, the increase in access is often offset by the rapid growth in the population. In the past two decades, India has demonstrated spectacular expansion in the number of children in schools, yet there is a decrease in the enrollment ratio because of the even faster growth in the population.

Both China and India face challenges of school drop-out. There are broadly two causes of school drop-out: (a) inability of the families to bear the private costs of schooling, and (b) income-earning economic activities that distract children in their teens from schools. In China, this is more serious in rural areas of the newly prospering economies, where there is usually a drastic increase in the private costs for formal schooling, and where there are plenty of other opportunities in economic activities for the young. This is particularly serious in the final years of the nine-year compulsory schooling. In India, school drop-out is also more serious in the final years of the eight-year basic education, and occurs mainly in rural villages where schooling does not promise foreseeable employment opportunities in the modern sector.

The problems facing China and India are not limited to those two giant countries. Indonesia, for example, hosts a student population of more than 26 million. Although Indonesia does not have a vast area compared with China and India, its over 600,000 islands have made universalization of basic education an even more formidable task. Other countries such as Pakistan and Bangladesh share the problem of rapid population growth.

Population and access are less of a problem in the more developed economies. Japan, South Korea, Taiwan, Hong Kong, and Singapore have seen universal or near-universal basic education since the 1970s. The demographic situations in these societies are fairly stable; the economy poses demand for more education, and both the public and private costs are largely met with little difficulty.

Low female enrollment is a serious issue in many South Asian systems of education. It is also an issue in Japan and Korea at tertiary education level, very much because of the social expectations for women. In most other societies in Asia, there is a general increase in the percentage of female students in higher education. In some systems, such as Hong Kong and Taiwan, there are more female university students than male.

5. *Schools and the Curriculum*

Schools are a relatively new concept in many Asian countries. Before the school model was borrowed from the West, education took place as private tutoring, or within the temples. Still, in many Asian countries, particularly those where the population is strongly Islamic or Buddhist, education or learning is a concept that goes far beyond schools, and schools are seen only as training geared for the modern sector.

Unlike in the West, where the early schools were largely associated with the church, schools in Asia were started deliberately separate from the traditional religions. In most of the Asian systems, strong government intervention in the school system was only a post-World War II phenomenon. In the Asian context, there was the Karachi Declaration of 1962, which urged all countries in the region to achieve universal basic education within 20 years. In the decades that followed, universal education was the policy goal in most Asian countries, although the target goal of universal basic education still remains a remote vision in many countries.

The concept of a school varies from society to society. Physically, schools in Asia may vary from modern concrete buildings, which are not very different from any urban school in other parts of the world, to makeshift bare shelters, with the blackboard as the only symbol of a classroom. In the mid-1980s, there was a movement in China to supply each student with a desk and a chair. At exactly the same time, India launched the 'Operation Blackboard' in order to equip each class with a good blackboard.

There have been success stories of mobilization of local communities in building and running schools. This is now common practice in China, India, Pakistan, and Bangladesh, and is well regarded among the developing world.

Asian classes are relatively large by world standards. It is not unusual to see classes of over 50 or 60 students, at both primary and secondary levels. Such classes may be seen in relatively underdeveloped rural areas in South Asia, but are also common in developed urban cities in Korea, Japan, and the Chinese com-

munities. In many cases, the large class is seen as a necessary community for collective learning, rather than a matter of economy.

Most of the Asian systems of education maintain a traditional curriculum. Few systems in the region have seen fundamental changes in the curriculum, and the modes of teaching or learning. Typically, schools teach the national language and mathematics as the core subjects, plus a foreign language, which since the 1980s has been predominantly English in almost all countries.

Among others, most of the systems in South Asia place a high value on the English language, and the societies use English as their language for wider communication among different language groups within the countries.

The Asian schools, East Asia and South Asia in particular, have a strong tradition in mathematics. This has made the East Asian systems of education envied by many outside the region, who regard mathematics as the most reliable evidence for student performance. The high level in mathematics learning is sometimes regarded as contributing to the success of the East Asians and South Asians in the American Silicon Valley, or the recent championship of India in software industries.

However, the Asian systems of education are also known for the frequency and intensity of examinations, both inside and external to schools, and the pressure therefore brought upon students. Private tutorials are common in many Asian countries. In East Asian societies, private tutorials are formalized in after-school tutorial schools. Typified by the *juku* in Japan, a large percentage of students attend tutorial schools as a normal part of their lives. In Korea, the tutorial classes start even earlier than schools in the morning. The examinations are often seen by educators as distorting the aims and relevance of learning. The reduction of adverse examination pressures has been the major theme in recent educational reforms in China, Singapore, Taiwan, and Hong Kong.

Nonetheless, the examinations play a crucial role as quality controller, and have contributed to the extraordinarily small variance among students' achievements. This has often lent itself to supporting movements for national assessments in some Western countries.

In most of the strong cultures in Asia, there was also a strong oral tradition in imparting knowledge. This has been translated into reading, reciting, memorizing, and copying, which are a large part of school activities. Such a tradition is a common point of reform in China, Singapore, Hong Kong, and Taiwan. Meanwhile, there are also recent discoveries, e.g., 'rote-learning' could be a particular way of learning, and practicing is seen as an alternative path to understanding.

There are also strong elements of 'moral' education in Asian schools. They are implemented either through religious or ethics education, in schools that are sponsored by religious bodies. Even in secular schools, it is regarded as essential to have elements of education in moral values, attitudes, and conscience, and these are often more important than the cognitive development of students. Such moral education is fulfilled through formal lessons (such as in China, Singapore, and Taiwan), internal publicity and communications (such as campaigns, assemblies, or media within the schools), and extracurricular activities. In many Asian systems of education, the schools pay particular attention to a large variety of extracurricular activities as a way of training of student leadership or social awareness.

6. *Challenges of the Knowledge Society*

The coming of the knowledge society has posed new challenges to the Asian systems of education. Most of the Asian economies have seen a rapid growth in the service sector that demands educated human resources. In many major cities in Asia, the 'white-collars' have exceeded the 'blue-collars' and there is little room for uneducated 'cheap labor.'

In these circumstances, the traditional high respect for education in Asian cultures has become a blessing, and that has been witnessed through the strong community support for basic education in many Asian countries in the last few decades of the twentieth century. Most Asian societies have no phobias about modern technology, and that attitude assists the development of learning technologies and lifelong learning.

The systems have inherited strong cultural traditions that place high importance on the development of the person. That is another strength, given the emerging emphasis of 'social competency' or 'character education,' which is increasingly seen as essential to a society where acquisition of knowledge is no longer seen as the major function of education.

However, the strong tendency to conformity in the systems, which has led to rigidity in the curriculum and uniformity in expectations, might pose a challenge to the Asian systems. There is a perceived need to facilitate and enhance diversity and creativity among the younger generation, and this is on the agenda of educational reform in many systems of long tradition in Asia.

See also: China: Sociocultural Aspects; East Asian Studies: Culture; East Asian Studies: Society; Educational Systems: Australia; Educational Systems: Europe; Educational Systems: North America; Japan: Sociocultural Aspects; Korea: Sociocultural Aspects; Scientific Academies in Asia; South Asia: Sociocultural Aspects; South Asian Studies: Culture;

Southeast Asia: Sociocultural Aspects; Southeast Asian Studies: Society

Bibliography

Unesco 2000 *World Data on Education*, Unesco

K.-M. Cheng

Educational Systems: Australia

As the individual governments of the six Australian states and the two territories have the major responsibility for education there is no one national system. Despite this, the provisions in each jurisdiction are similar. Collaboration on matters of policy take place in a council of ministers of education and the roles of the federal government have increased in recent years. Education is provided as part of preschool education, school education, and tertiary education (consisting of vocational education and training as well as higher education). Major elements of the education systems in Australia are represented in Fig. 1.

1. *National Context*

In 1997, Australia had a population of 18.5 million in an area of 7.7 million km^2. Although the overall population density is low, it is a highly urbanized society; the two largest cities contain 40 percent of the population. Outside the cities, the country is sparsely populated; 30 percent of primary schools have fewer than 100 students and 30 percent of secondary schools have fewer than 500 students. Commentators have noted that centralized administrative structures emerged to promote uniformity of educational provision for a dispersed population. Although the Australian population is mainly of European background, recent immigration has produced greater ethnic and cultural diversity. One-fifth of the population was born overseas. English remains the language of most activities in education. About 2.5 percent of Australian school students are of Aboriginal descent, some of whom still live in isolated communities.

The Government is a parliamentary democracy. After the federation of former colonies in 1901, states retained responsibility for education. Two territories have education powers similar to the states. Political responsibility, federally and in each state, resides with a minister who is accountable to the relevant parliament. State governments provide about 60 percent of all public expenditure on education (90 percent of the expenditure on school education). The federal government provides 40 percent (including most of the expenditure on universities). For the financial year 1996 to 1997, total public and private expenditure on education was A$29,300 million, which represented 5.7 percent of GDP. Government outlays amounted to A$24,500 million. Over half of these outlays (A$13,900 million) were directed to primary and secondary schooling (ABS 1998a).

The World Bank classifies Australia as a high-income country. In 1996 to 1997, per capita GDP was just over A$24,000. Literacy among adults is nearly universal and 43 percent hold a post-school qualification. Just over 12 percent have university degrees and 11 percent have a vocational qualification. Amid general affluence, unemployment among young people has remained high in recent years. In November 1996, some 11 percent of 15–19-year-olds were unemployed and a number of programs are concerned with

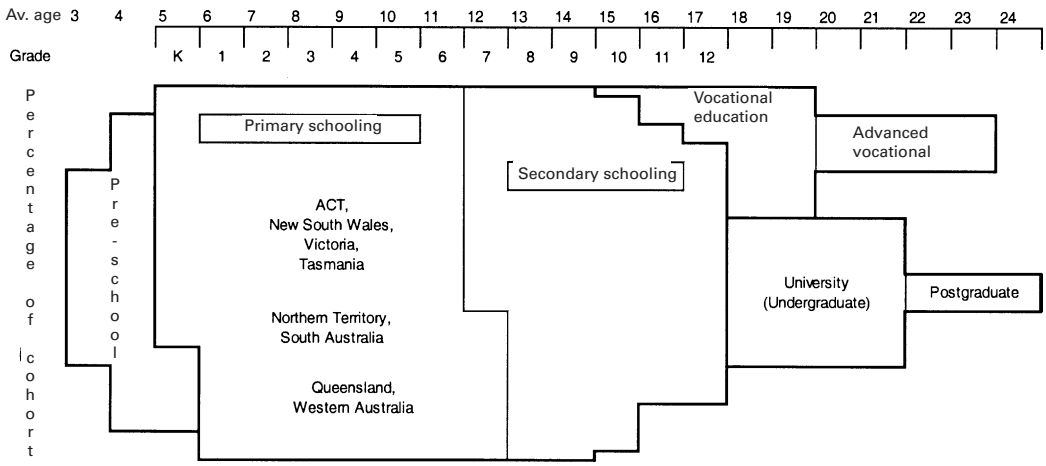

Figure 1
Representation of the Australian Education System

transitions between education and work (OECD 2000, p. 128). Shifts in patterns of employment towards service industries have also resulted in changes to education and training programs.

2. Preschool Education

Although the states and territories provide preschool education there is no national policy on its provision. The age at which children attend preschool varies across jurisdictions and differences exist in the organization of preschools. A national survey reported in 1996, that 188,200 children younger than five years attend preschool. Children of three and four years old make up 92 percent of children at preschool. A majority of 4-year-olds attend preschool on a part-time basis (ABS 1996).

3. Primary and Secondary Schooling

Education is compulsory from six to 15 years (16 in Tasmania). Most children start primary school at five years of age and continue for either six or seven years (depending on the state), and students complete primary school at the age of 11 or 12 years. Students in Australian primary schools usually have one teacher for most subjects, and are promoted each year. During the 1990s, there emerged a national focus on the development of literacy especially in the early years. This focus is embodied in a *National Literacy and Numeracy Plan* agreed to by federal, state, and territory education ministers (DEETYA 1998).

Secondary education is provided for either five or six years depending upon the length of primary education in the state. The first one or two years of secondary school typically consist of a general program followed by all students. In subsequent years, a basic core of subjects is supplemented with optional subjects available to students. Students in secondary schools generally have a different teacher for each separate subject area. In the final two years of secondary schools students have more scope to specialize and a range of elective studies is provided from which students choose five or six. Almost all (95 percent) final-year students study English and 85 percent study a mathematics subject. Subjects from the sciences and studies of society and environment are taken by 70 percent of students. Just fewer than half the students include a technology subject, 30 percent take an arts subject, and 10 percent study a foreign language (Ainley et al. 1994).

One of the most marked changes during the 1980s was an increase in the percentage of students who remained to complete secondary school. The percentage of commencing secondary students remaining to the final year of school rose from 35 percent in 1980 to 77 percent in 1993. It has since declined a little to 72 percent in 1997. Since 1976, the school completion rate has been higher for girls than boys. In 1997, 78 percent of girls, compared to 66 percent of boys, remained to the final year of school (DETYA 1999).

For both primary and secondary schools the length of the school year is approximately 200 days; beginning at the end of January and ending in the third week of December and divided into four terms of nine to 11 weeks. The summer vacation is five or six weeks. Students attend school five days per week. In primary schools, there are five hours of tuition per day and in secondary schools there are 5.5 hours tuition per day.

3.1 Government Schools

State education departments recruit and appoint the teachers in government schools, supply buildings, equipment, and materials, and provide limited discretionary funding for use by schools. In most states some responsibility for administration, staffing, and curriculum has been devolved to regional education offices and schools (which has become more extensive in recent years).

In 1997, the average sizes of primary and secondary schools were approximately 230 and 780 respectively (MCEETYA 1997). During the 1990s there was a planned reduction in the number of government schools so as to increase the average school size. Most government secondary schools are comprehensive and coeducational. There are a small number of selective-entry secondary schools in some states. In Tasmania and the Australian Capital Territory, the final two years of government schooling is in separate senior secondary colleges.

In 1997, the ratio of students to teachers was 17.9 in primary schools and 12.7 in secondary schools. In the 1996 to 1997 financial year, per capita expenditure was A$4,686 in government primary schools and A$6,447 in government secondary schools (MCEETYA 1997). In the largest system, in 1993 the average class size in primary schools was 27.4 and the average size of English classes in secondary schools was 22.9 (MCEETYA 1994).

Taxation revenues provide almost all the financial resources for the operation of government schools. Although parents are not officially required to pay fees for students to attend government schools, many schools seek voluntary contributions from parents and raise funds from other local sources.

3.2 Non-government Schools

Non-government schools enrolled 30 percent of students (26 percent of primary and 34 percent of

secondary school students) in 1997, a proportion that has gradually risen since 1970. Most private schools have some religious affiliation, most commonly with the Catholic Church (66 percent of private school students are in Catholic schools). Other private schools vary from long-established prestigious schools to relatively new lower-cost schools. A range of funding sources including government grants supports private schools. In 1997, 44 percent of private school income was derived from fees or donations, 18 percent from state government grants and 38 percent from federal government grants. Government grants comprised 72 percent of the income of Catholic schools and 34 percent of the income of independent schools. In 1997, per capita expenditure was estimated as A$3,996 and A$6,510 in non-government primary and secondary schools respectively (MCEETYA 1997).

3.3 Special Education

Special education programs are provided for students with physical, intellectual, and emotional disabilities. A national study in 1992 estimated that about 2 percent of students in Australian schools had some form of disability. At that time, about 30 percent of students with disabilities were enrolled in special schools, 25 percent were enrolled in special classes or units attached to other schools, and 45 percent were enrolled in regular school classes. Over the past 20 years, policies have been directed towards supporting education for a greater proportion of students with disabilities in regular school classes (deLemos 1994).

3.4 Curriculum

Although there is no national curriculum, governments collaborated in the early 1990s to develop curriculum statements and profiles in eight key learning areas: English, mathematics, science, technology, studies of society and the environment, the arts, health and physical education, and languages other than English. The statements provide an account of the aims and content of each area of learning as a framework for curriculum in primary and secondary schooling. The profiles provide an outline of a sequence of students' development of knowledge, understanding and skills and a framework for assessing and reporting student progress and achievements. Although the national statements were not adopted, states and territories use statements and profiles in some form for curriculum development, incorporating variations that reflect local policies and priorities. In 1990, the state and federal education ministers established the Curriculum Corporation, a semi-autonomous body with a charter to develop curriculum materials on a commercial basis.

Within states, central authorities specify curriculum guidelines and schools have considerable autonomy in deciding curriculum detail, textbooks, and teaching methodology at primary and junior secondary level. At the senior secondary level, the curriculum is specified in detail by a state authority responsible for examining and certifying student achievement.

3.5 Assessment and Monitoring of Student Achievement

All states operate statewide assessment and monitoring programs for primary schools. There is population testing of students in government schools in English and mathematics at Grade 3 and Grade 5. The purposes are to report information about student achievement, based on a common test to parents, teachers, and schools and to monitor the overall performance of the education system. Most of the assessment instruments used in these programs of testing incorporate a range of items including multiple choice and open-ended response items, and make use of modern measurement techniques that have enabled forms of criteria-referenced reporting to be implemented. Formal statewide assessment programs are currently less common in the lower secondary grades than in primary school. Evidence from the *National School English Literacy Survey* (Masters and Forster 1997) conducted in 1996, indicates the extent of differences in early literacy achievement; the top one-fifth of students in Year 3 performed at a level close to the average for Year 5.

3.6 Teachers

A little over 207,000 (full-time equivalent) teachers were employed in schools in 1997: approximately 104,000 in primary and 103,000 in secondary schools. Overall 65 percent of teachers are female (77 percent in primary and 53 percent in secondary schools) (ABS 1998b). Teaching is an occupation with relatively high prestige, being the third top of 16 categories on the occupational prestige scale. Salaries are determined at state level and there are differences between states. The annual starting salary is approximately A$30,000 and for a teacher with 10 years' experience is about A$45,000. Teacher training occurs in universities, but states determine acceptable qualifications. A total of 4 years of higher education is the normal length of initial training for secondary teachers; typically, a 3-year degree in a major discipline and a 1-year diploma of education. Primary teachers normally complete a 3-year diploma of teaching, but many later improve their qualifications with specialist diplomas or a 4th year, resulting in a degree.

3.7 Information and Communication Technologies in Schools

From the mid-1990s, efforts have been made to incorporate information and communication tech-

nologies in school programs. The intention is for students to become proficient at using information technologies, graduate as skilled users of the technology, and use these tools to support learning across the curriculum. Schools vary in the extent to which students have access to these technologies and high costs are an impediment. A national survey in 1998, estimated that 64 percent of schools operated with a ratio of between 6 and 15 students per computer; the average figure being approximately 12 (Meredyth et al. 1999).

4. *Vocational Education and Training*

Vocational education and training (VET) is an important component of education in Australia. A variety of courses are provided through institutes of technical and further education (TAFE) and other organizations. Most (over 95 percent) vocational education and training is provided in colleges (or institutes) of technical and further education (TAFE). Courses include a range of vocational training from entry-level employment preparation to trades (e.g., through apprenticeships) through advanced vocational courses to para-professional and professional courses. In addition, many recreation and leisure programs are provided.

4.1 *Students*

In 1997, approximately 121,000 TAFE students graduated with a qualification from a vocational course of at least 200 hours, or one semester, duration (NCVER 1998). Some 20 percent of young people participate in a TAFE program by the age of 19 and a further 15 percent (mostly males) undertake apprenticeship training through TAFE (Long et al. 1999). Overall, there were 1.4 million enrolments in VET programs in 1997. Participation is characterized by part-time attendance and a wide age range (people aged 15–24 years comprise 38 percent of the clients). Entry to many courses is possible after Year 10, but in practice nearly half of the entrants to vocational courses have completed Year 12. In the early 1980s the corresponding figure was one-fifth. This change has arisen from a combination of labor market conditions and increased school retention rates.

4.2 *Apprenticeship*

Apprenticeship is an important component of VET. Over 4 years, an apprentice works for an employer (or group of employers) and attends a training institution (traditionally a TAFE institute) part-time (typically a total of 800 hours). Recent changes have been a response to perceived limitations (inflexibility, limited occupations, old technology, and lack of access for women) and declining numbers. In 1985, traineeships were introduced to provide a shorter and more flexible approach to entry-level training. Traineeships, typically involved a one-year program, with an employer incorporating on-the-job and off-the-job training, mostly in office-based and retail industries. More recently, apprenticeships and traineeships have been integrated as part of a more unified entry-level training system.

4.3 *Organization*

Responsibility for administration of VET lies with the state and territory governments. In 1997, recurrent and operating funding of the TAFE system was about $3,800 million, with the states and territories providing 56 percent and the federal government 25 percent, with the remaining 19 percent coming from fees and other sources. A national vocational education and training system was established in 1992. A ministerial council determines national policy and priorities, strategic directions, funding arrangements, and planning processes for vocational education and training, on the advice of the Australian National Training Authority. State training agencies manage the delivery of VET in a manner consistent with the national strategic plan.

5. *Higher Education*

Prior to 1990, higher education comprised universities and colleges of advanced education (which placed more emphasis on teaching and less on research). In 1990, formal distinctions between the two types of institution were removed and since then amalgamations stimulated by federal government policy have reduced the number of institutions from around 70 to 36. In addition to the 36 public institutions there are two private universities. Most universities operate under state legislation and are autonomous bodies responsible for their own governance, allocation of funds, staffing, and academic courses.

5.1 *Students*

Students commencing higher education courses are required to complete secondary school, or otherwise demonstrate a capacity to complete a course. There is keen demand for higher education places at most institutions. In 1997, there were nearly 660,000 students enrolled in universities with slightly fewer than 500,000 enrolled in undergraduate courses. The basic undergraduate course is a bachelor degree of three or

four years' duration, and these enroll 75 percent of all students. The fields of study with the largest numbers of students are: the arts, humanities, and social sciences (23 percent); business, administration, and economics (23 percent), and science (15 percent) (DEETYA 1997). Approximately half of the students who complete secondary school proceed to university by the age of 19, this being approximately 38 percent of each cohort. In 1997, a little more than half (54 percent) of university students were female, a figure that has risen steadily over the 1980s and 1990s. Despite the expansion of higher education during the 1980s, university students are still drawn disproportionately from those with managerial and professional home backgrounds (Long et al. 1999). In 1997, approximately 60 percent of students were enrolled in full-time study, 27 percent in part-time study and the remainder in external studies.

5.2 *Organization and Funding*

The federal government funds universities. In 1996, the operating revenue of universities amounted to just over A$8,000 million, 57 percent of which came from federal government grants. In addition to government funding, institutions receive payments from students who are required to contribute to the cost of their education through the *Higher Education Contribution Scheme* (a deferred, income-contingent repayment program that operates through the taxation system), and from full fee-paying students.

6. *Education at the Turn of the Millennium*

At the beginning of the new millennium (2001), there are a number of issues that will challenge the education systems of Australia. The balance, and funding, of public and private schooling, establishing a new mix of central authority and local responsibility in public school systems, and developing new forms of quality assurance seem likely to continue as important issues in the next few years. So also will questions of securing greater levels of engagement among students in the middle years. Providing resources for an appropriate mix in programs of education and training beyond school has emerged as an issue as the demand for education continues to rise. In senior secondary school and in higher education there will most likely be new initiatives to facilitate the pathways between education and work. In addition the ongoing issues of providing programs for a dispersed, and an increasingly diverse, population will continue.

See also: Curriculum as a Field of Educational Study; Educational Institutions and Society; Educational Planning Worldwide; Educational Policy: Comparative Perspective; Educational Systems: Asia; Educational Systems: Europe; Educational Systems: North America

Bibliography

Ainley J, Robinson L, Harvey-Beavis A, Elsworth G, Fleming M 1994 *Subject Choice in Years 11 and 12*. Australian Council for Educational Research, Melbourne, Vic

Australian Bureau of Statistics (ABS) 1996 *Child Care, Australia (4402.0)*. ABS, Canberra, NSW

Australian Bureau of Statistics (ABS) 1998a *Expenditure on Education, Australia (5510.0)*. ABS, Canberra, NSW

Australian Bureau of Statistics (ABS) 1998b *Schools Australia 1997 (4221.0)*. ABS, Canberra, NSW

DeLemos M 1994 *Schooling for Students with Disabilities*. Australian Government Publishing Service, Canberra, NSW

Department of Employment, Education, Training and Youth Affairs (DEETYA) 1997 *Selected Higher Education Student Statistics, 1997*. DEETYA, Canberra, NSW

Department of Employment, Education, Training and Youth Affairs (DEETYA) 1998 *Literacy for All: The Challenge for Australian Schools*. Australian Government Publishing Service, Canberra, NSW

Department of Education, Training and Youth Affairs (DETYA) 1999 *Education Participation Rates, Australia—1997*. DETYA, Canberra, NSW

Long M, Carpenter P, Hayden M 1999 *Participation in Education*. Australian Council for Educational Research, Melbourne, Vic

Masters G, Forster M 1997 *Mapping Literacy Achievement: Results of the 1996 National School English Literacy Survey*. J S McMillan, Canberra, NSW

Meredyth D, Russell N, Blackwood L, Thomas J, Wise P 1999 *Real Time: Computers, Change and Schooling*. J S McMillan, Canberra, NSW

Ministerial Council on Education, Employment Training and Youth Affairs (MCEETYA) 1994 *National Report on Schooling in Australia: 1993. Statistical Annex*. Curriculum Corporation, Melbourne, Vic

Ministerial Council on Education, Employment Training and Youth Affairs (MCEETYA) 1997 *National Report on Schooling in Australia: 1997*. Curriculum Corporation, Melbourne, Vic

National Centre for Vocational Educational Research (NCVER) 1998 *Australian VET: TAFE Graduate Destination Survey 1997—National Report*. NCVER, Adelaide, SA

Organisation for Economic Co-operation and Development (OECD) *From Initial Education to Working Life: Making Transitions Work*. OECD, Paris

J. Ainley

Educational Systems: Europe

National coherence and the feeling of national identity depends mainly on a common frame of reference of language and values, and this depends upon a common

language and mainly common systems of formal schooling, at least at the basic level. The modern national state, represented by the industrialized countries in Europe, was shaped by the mandatory primary school which was legislated in many European countries and became an instrument of socializing children into the respective national cultures, the mother tongues, and the predominant values. It also contributed to establishing a national language over and above the various dialects or local languages. When, in 1870, the French parliament decided to make *école primaire* mandatory all over the country, French was spoken by only half the population within the geographical area today referred to as France. In peripheral areas German, Italian, Spanish, and Breton were spoken.

1. Similarities

The school systems in most European countries have a rather similar historical background. Primary school has a relatively short history. It was established along with mandatory school attendance over a few decades in the nineteenth century, in most cases by legislation in national parliaments. What is now referred to as secondary school has its roots in the Middle Ages with its monastery and cathedral school. These took care of an elite and were preparatory for university studies. The German *Gymnasium*, the British grammar school, and the French *lycée* were typical of this type of school. Well into the twentieth century, preparation for entry into the secondary schools was taken care of privately.

The parallel between the first grades of secondary school and the last ones in primary school were brought into serious question after 1945. Behind this was the demand for greater equality of opportunity, but also the social and economic changes in a society where formal education increasingly became crucial for individual career and social status.

In the late 1930s steps began to be taken in reforming British education. The Labour governments in the 1940s to 1960 were strongly in favor of a comprehensive school system which would replace the tripartite system of 1944. Part of the breaking up of the old structure in France was the presidential decree of 1959 which introduced a *cycle d'orientation* (middle school) of two years after five years in *école primaire*. In Germany a similar, two-year *Orientierungsstufe* comes after four years in *Grundschule*. The rationale behind such an arrangement has been to get ample opportunity and time to find out which students are able to proceed in the academic and vocational programs respectively.

Reform movements after 1945 in several countries aimed at making secondary school available and comprehensive for all young people. The 1959 French decree by President de Gaulle made the transfer from primary to secondary more flexible by the *cycle d'orientation*. The same occured in many places in the Federal Republic of Germany where the more radical reformers were demanding a *Gesamtschule* (comprehensive school) which would cover the entire period of mandatory schooling. The German system, however, was more institutionalized and rigid than the French, and parallel to the new system in Germany, transfer could occur from the fourth grade in the primary school to the first grade in the nine-year *Gymnasium*. In France an intermediary type of school at the lower secondary level was introduced, called *collège d'enseignement secondaire*, which broadened the opportunities for the majority of students. In Italy, for a long time after 1945, the *scuola media*, the lower secondary school, played an important role in the debate about school reform. The debate primarily referred to northern Italy; the southern part of the country was lagging behind in this respect.

It is not possible here to describe each national system of education in Europe. Therefore, the article is limited to the systems in France, Germany, the UK, and Russia, all of which have influenced the other countries by serving as models. Each of these systems is represented by an outline of their structure. Common to all of them is the division into three main stages: primary, secondary, and higher education. They also all have in common a mandatory primary stage, whereas for a long time the secondary institutions educated a social and intellectual elite in separate, adademically oriented schools, such as *Gymnasium* in Germany, *lycées* in France, and grammar schools in the UK, all preparing for higher education.

2. Differences

The various national systems differ in several respects (see OECD 1998, European Commission 1995, 1991, Husén et al. 1992, Eurydice European Unit 1991). Some countries have a 'comprehensive' school that caters for all children of primary school age and covers the first 8–10 years in primary school which then runs parallel to the secondary for the next few years. The Nordic countries today have a comprehensive system which gradually has been gaining ground in France, the UK, and some of the *Länder* (states) in Germany. In these countries there are also private schools more-or-less independent of the state. Most of the countries in Eastern Europe introduced postwar comprehensive systems modeled after the one in the USSR.

Transfer from primary to secondary school occurs, as mentioned above, in various countries and at different age levels. In Germany transfer has in most *Länder* taken place at the age of 10 after four years of primary school. In the UK it takes place at the age of

11. In the Netherlands, most children start school at the age of four but it becomes mandatory at five, and primary school lasts until the age of 12. In Italy, mandatory school is from 6 to 14; children have to take an examination called *licenza elemenare* after which they can proceed to the next stage, *scuola media*, which is also comprehensive.

Secondary school is thus organized very differently in the various countries. What in Sweden is referred to as *gymnasium* consists in Germany of grades 11–13 (in some *Länder* grades 11 and 12). It is an academically oriented school running parallel to schools with vocational programs.

3. School Structures in Four Countries

3.1 Germany

In the late 1940s Western Germany became a federation of states which had full sovereignty in matters of education. After four years in the *Grundschule*, transfer to a two-year *Förderstufe* can take place. This stage can be part of a *Gymnasium* (nine-year), *Realschule* (six-year) or a *Hauptschule*.

A striking feature of the German *Gymnasium* has traditionally been the multitude of subjects in the final grades. The purpose is to give a broad general education which can be a heavy task for the students. In this respect the German system differs radically from the British with its specialization on three or four subjects during the final two grades of secondary school.

The number of students reaching *Abitur* (examinations necessary for university entrance) was for a long time rather low, for instance less than 10 percent of the age group in the 1960s, but during the last few decades it has increased to 25–30 percent.

In the DDR (East Germany) before unification there was a 10-year polytechnically oriented school where during the last few grades students could work and learn in business enterprises. The *Gesamtschule*, the basic and/or comprehesive school was the model taken over from the USSR. It was also introduced in the Federal Republic (West Germany) in some states, such as Hesse, in the 1960s.

3.2 France

The school system in France had for a long time two kinds of parallelism. There were public schools running parallel with religious (Catholic) schools. Furthermore, there were municipal schools parallel with *lycées* run by the state. The structure at the beginning of the twenty-first century is as follows. After *école maternelle*, the preschool which traditionally covers a larger portion of the age groups than in most other European countries, comes the five-year *école primaire*. Then follows secondary school, *collège* or *lycée* of which the first two grades, along with the corresponding grades in the *école primaire*, make up a *cycle d'orientation* (middle or transition school), after which a decision is taken as to whether or not the student is capable of transferring to an academically oriented *lycée* or will have to finish schooling in a *collège d'enseignement géneral*, which is the final stage of madatory schooling. In 1959 the latter was increased to nine years but in 1975 a *classe terminal* was added. Thus the 10-year school covers the age range 6 to 16.

Grade repeating (*redoublement*) has traditionally been a dominant feature of the French system. As late as 1980 about one third of the students in *école primaire* consisted of grade repeaters. This means that in the upper grades students on the average are one year older than 'normal' for the grade.

3.3 The United Kingdom

During World War II, in 1944, the British parliament passed an Education Act which provided access to secondary school for all young people in the age range 11 to 16. After five years in primary school they could transfer to secondary school with mandatory attendance up to the age of 15. Since schooling began at five, transfer to secondary school took place at 11. Then students had to go through a selection procedure, the '11 plus' examination, which more or less decided the allocation of students to three types of secondary school: grammar schools, technical schools, or secondary modern schools, which all had five-year programs. About one fourth qualified for grammar school, the rest were roughly equally distributed among the other two types. Schools reformers, particularly in the Labour party, were proponents for a unitary school, a comprehensive school for all students. Comprehensivization occurred in the UK after the early 1960s and the 11 plus examination was abolished in most of the UK. But the students are as a rule differentiated according to ability within the schools.

At the end of the nineteenth century the great majority of schools in the UK were divided into two stages: primary from 5 to 11 years of age and secondary from 11 to 16 (or for those taking a preparatory program for university entrance, to 18). The distinction now in some parts of the UK is between three stages of schooling: first school from 5 to 7 or 8, middle school from 8 to 12 (or 9 to 13), and upper school from 12 or 13 to 16 or 18. Some 5 percent of children of primary school go to private schools. At the secondary stage the percentage is 8–10.

After five years in secondary school the student can take General Certificate of Secondary Education

(GCSE) external examinations. Those who intend to go on to higher education can, after two more years of specialization in three or four subjects, take the examinations for the General Certificate of Education advanced level (A-levels).

The school reform of 1988 means that for the first time a national curriculum has been introduced, along with national assessments of student competence at several stages.

3.4 Russia

The disintegration of the USSR meant that several of the republics in the Union tended to move away from the model which was dominant, namely the Russian one. The literacy of the population within an essentially agrarian economy had until the 1930s been a central task, which meant that adult education played a dominant role in the educational system. When universal schooling for children was legislated the big cities were given priority. Thus Moscow and Leningrad early provided 9–10 year schooling for the majority of children, whereas in rural areas for a long time five-year primary school was common. A comprehensive eight-year basic or primary school has traditionally been common in Russia, at least in the urban regions. A secondary two-to-four-year school with various vocational programs followed. Higher education has also had a clear-cut vocational bias which reflects the 'polytechnical' orientation that has characterized the communist educational philosophy, a kind of symbiosis between theory and practice. Whereas schooling in many districts was easily available for all, selection to higher education has been rather tough. For a long time adult education played a prominent role in Soviet education with evening courses, distance teaching, etc.

The system was, and still seems to be, strongly centralized. Curricula were conceived and issued from Moscow. The Academy of Pedagogical Sciences thereby played an important role in preparing a common curriculum. The reforms that followed upon the disintegration of the USSR implied in certain respects a decentralization.

4. Private School Systems

In most European countries there has traditionally been a private sector both at the primary and secondary level. This sector is in the UK referred to as 'independent' and is financed by fees and private funds, but in some instances students from less affluent backgrounds can be excused fees. In the 1990s between 6 and 9 percent of the age groups at primary school level went to independent ('preparatory') schools and 12–15 percent at the secondary level. Many of these schools in the UK are boarding schools; the secondary schools are referred to as 'public schools', the most famous of which are Eton, Rugby, and Harrow. Elite universities, such as Oxford and Cambridge, for a long time recruited roughly half of their students from the independent schools.

In France a 'cultural struggle' went on for more than a century between the state and the Catholic Church about what basically had to do with religious instruction. The Catholic schools, which in around 1900 recruited about 20 percent of the age group at the primary stage, had until then no public financial support, but gained one in 1905. Recurrent attempts have been made, as late as in the mid-1980s, to incorporate the Catholic schools into the state system, but the attempts have failed. In other Catholic countries, not least in Spain, the church has a firm hand on the school system and the curriculum, whereas in the Protestant countries the state usually rules the system. In the Netherlands there are, in practice, three systems; one Catholic, one Protestant, and one secular, each with its own central administration. The curriculum is, however, on the whole common. In Belgium there are one Flemish, one French-oriented, and one German-oriented system, each with its own Minister of Education.

Denmark has a long tradition of independent schools, which goes back to the reformer Grundtvig. The pedagogical cornerstones are student self-reliance and free approaches to learning.

In 1988 a school reform in Great Britain was passed which reduced the influence of the local education authorities (LEAs) and made the Ministry of Education more powerful. Thus, a majority of parents in a local school could decide to 'opt out' from the LEA and had the right to get financial support directly from the Department of Education in London.

There have in recent years been tendencies in several countries to make the boundaries between public and private schools more fluent with regard to the role of both central and local authorities.

5. Higher Education

Three national systems of higher education have been influential outside the boundaries of the respective countries: Germany, France, and the UK. The German system has influenced particularly the countries in northern and central Europe, the French has been a model in some East European countries, whereas the British has been emulated in former colonies.

The idea of the university, which Wilhelm von Humboldt launched at the beginning of the nineteenth century and implementend in Berlin, had a strong influence both on the European continent and in northern Europe. The basic idea was the unity of

teaching and research which, among other things, was realized by introducing departments at the undergraduate level and special institutes for research at the graduate level. For instance, the model established by Wilhelm Wundt in 1879 by setting up an institute at the university in Leipzig had an important impact on the development of experimental psychology on both sides of the Atlantic. Great distinction was made at the time between basic and applied research and between theory and practice. In German philosophy of education at higher level the word *Bildung* was of central importance. The study of classical languages and history was regarded as contributing a good deal to general education, although in Germany natural sciences were also appreciated. The faculties of natural sciences, as well as those of medicine, played a more important role at German universities than at UK ones. Teaching was provided mainly by professors, the 'chair holders,' one in each discipline.

French higher education on the whole got its present structure during the era of Napoleon, with the establishment of the *grandes écoles*, elite institutions for training various professions, such as civil engineers, senior civil servants in the ministries, and teachers in the upper grades in the *lycées*. The teaching put high intellectual demands on stringency and analysis. The professors were expected to be in contact with what happened in their respective disciplines but not to devote too much of their time to research, which was conducted mainly outside the universities in *Centres Nationeaux de Recherche Scientifique* which were placed directly under the Ministry of Education. As an outcome of the student revolution of 1968 the university system and its administration was thoroughly reformed by the *lois d'orientation*. For instance, the University of Paris, which until then had been entirely concentrated at the Sorbonne, was divided into a dozen new units. The new law created almost one thousand *unités d'enseignement et recherche*, something in between faculties and departments for various disciplines. More influence was given to the students with regard to instruction and the governance of the bigger units. The French system, not least the university system, is traditionally rather centralized and therefore rather rigid and difficult to reform.

The UK had, until 1992, 47 universities, which had the right to grant examination certificates including the doctoral degree. Parallel to the universities in this dual system were about 40 polytechnics with two-or three-year, vocationally oriented programs. The polytechnics have constantly struggled for university status, among other things by acquiring research resources. In 1992 they were incorporated into the university system, which meant that the UK now had 88 institutions which could call themselves universities. This also means that there are great differences in quality and status. Elite institutions like Oxford and Cambridge, which are very selective and account for only three percent of the total number of university students, are often regarded as 'typical' of the UK system.

In 1988 a reform of the system made regular evaluations a condition for funding. British universities have traditionally enjoyed a high degree of autonomy vis-à-vis the central government. Thus, until 1988, the University Grants Commission, which consisted entirely of representatives of the universities, received a lump sum from the Government which was then distributed by the Commission among the universities. From 1988 there have been Higher Education Funding Councils for England, Wales, and Scotland, one for the universities and one for what were then polytechnics and college. In these bodies the majority of the members still have experience from the academic world. But there are more bodies ordering services in terms of training and research than distributors of available funds. Formal purchasing contracts are negotiated which makes funding dependent on the number of students in the respective disciplines, and on the 'productivity' demonstrated.

For a long time the elite universities in Britain could take credit for bringing about 90 percent of those who entered studies to an academic degree. This is taken after three or four years of study, which is shorter than in most countries on the continent where the basic degree can take five to eight years of study. Since UK university students begin their study earlier than on the continent, the basic degree is usually taken at the age of 21 or 22. The specialization of university studies is facilitated by the fact that the two final years in secondary school studying for A levels is specialized to three or four subjects.

See also: Comparative Research in Education: IEA Studies; Cross-cultural Study of Education; Education: Phenomena, Concepts, and Theories; Educational Institutions and Society; Educational Policy: Comparative Perspective; Educational Systems: Asia; Educational Systems: Australia; Educational Systems: North America

Bibliography

European Commission 1995 *Key Data on Education on the European Union*. Task Force for Education, Training and Youth, Brussels

European Commission, Directorate-General XXII 1991 *A Guide to Higher Education Systems and Qualifications in the EU and EEA Countries*. Office of Official Publications of the European Communities, Luxembourg

Eurydice 1999 *Organisation of Higher Education Structures in Europe*. Office of Official Publications of the European Countries, Luxembourg

Eurydice European Unit 1991 *The Structures of the Education and the Initial Training Systems in the Member States of the European Community*. Eurydice European Unit and CEDEFOP, Brussels

Husén T, Tuijnman A, Halls W D 1992 *Schooling in Modern European Society: A Report of the Academia Europaea*. Pergamon Press, Oxford
OECD 1998 *Education at a glance*: *OECD Indicators*. OECD, Paris

T. Husén

Educational Systems: North America

North America is home to two national systems of public, compulsory education that differ in size, degree of centralization and organizational structure, but have been shaped by similar historical and cultural forces. In this article, common cultural dynamics are discussed and an overview of other sub-systems of education, such as military schools and religious schools, is provided.

1. The Effect of Religion and Language on Early Forms of Schooling

Early in the seventeenth century, Father Superior Paul Le Jeune sat down with 'a small Indian boy on one side and a small Negro on the other' and attempted to instruct the children in the Catholic faith. While brief and ineffective, 'as neither of the three understood the language of the others,' Le Jeune's efforts epitomize the tenor and focus of schooling in North America during the seventeenth and eighteenth centuries (Parkman 1983, p. 413).

In both the French and English colonies, schools were used to maintain existing social hierarchies and assimilate non-European populations by disseminating religious views, inculcating moral virtues (European), and instilling literacy. Early colleges, (e.g., the Jesuit College in Quebec, or Harvard College) provided instruction in religious dogma and the Christian sacral languages. Literacy was tied to religious indoctrination, and the Bible was probably the most common school 'text' (see Kett 1977 or Sutton 1988). As Cremin noted: 'during the first phases of the Virginia settlement, for example, education was in the hands of the ministers, who used it both as an instrument for promoting discipline and order among the colonists, and as a device for winning the loyalty of the natives, (Cremin 1970, p. 9).

The impact of schooling was limited though, due to the fact that during the seventeenth, eighteenth, and early nineteenth centuries most of the population received little if any formal education. There were early attempts at vocational (or domestic) education—for example, the Ursuline nuns of Montreal received 'a thousand francs to buy wool, and a thousand more for teaching girls to knit.' (Parkman, 1983, 1357), but most 'education' consisted of skills learned in the family or apprenticeships. Even slaves participated in these forms to a wide extent: Equiano notes that he was trained as a ship's pilot as a slave (Equiano 1995, p. ix). Education in the colonies was unsystematic, polymorphous, and above all, practical.

2. Emerging Systems: Education to Transform Society

In the USA, networks of public schools developed rapidly in the early 1800s. Early architects of the public education system, such as Thomas Jefferson, believed that a secular, public system of schools would serve as the foundation of democracy in the new republic. Benjamin Rush, lobbied (albeit unsuccessfully) to create a state-funded system of education in the US State of Pennsylvania. Horace Mann worked tirelessly in the cause of the 'common school' in Massachusetts and worked to establish 'normal schools' to train teachers—a group whose educational accomplishments were often astonishingly meager in the 1800s.

Driving these movements for educational reform was a widespread belief in the transformative power of education and a desire to build a system of education that would change society—a cultural tendency that has been piquantly labeled as a propensity to 'tinker toward utopia' (Tyack and Cuban 1995). Rush was one of the first to articulate these ideas in a systematic way. He noted: 'Our schools of learning … will render the mass of people more homogenous, and thereby fit them more easily for uniform and peaceable government.' Enumerating 'a few advantages of learning upon mankind,' he wrote 'A free government can only exist in an equal diffusion of literature … and where learning is confined to a "few" people, we always find monarchy, aristocracy, and slavery' (Rush 1787, pp. 1–2, 14).

Yet, during this same period, many state legislatures began to systematically prohibit literacy and other forms of schooling to both slave and free African populations by denying them access to education. Nearly 150 years after thinkers like Thomas Jefferson proposed universal schooling as the foundation of a democratic society, racially segregated schools were common in the USA. At the same time that profound educational experiments, such as Lowell's incorporation of educational programs for his mill workers were occurring, African-Americans were denied the opportunity to learn to read.

It was not until the late 1800s when educational progressives like the philosopher John Dewey gained widespread political ascendancy and began to 'tinker' with public schools in order to transform the social order. From the late 1800s through to the end of the World War II, scholars and politicians from a

wide range of backgrounds worked to extend and systematize public schooling. Yet these reformers typically assumed, as did Rush, that schooling would homogenize new waves of immigrants and make them 'good citizens' of Canada or the USA.

3. *Conflict of Ideals: Maintaining Society or Transforming it through Schooling?*

The historical development of the North American school systems was affected by conflicting ideals about what schools should do for society. One end of the continuum is the view that school should be a vehicle for instilling morals, proper behavior, and transmitting a stable social order. At the other is the view that schools should allow students the widest range of self-determination and exploration and eliminate old prejudices thereby paving the way for a new, and better, social order.

3.1 *Colonization, Native Peoples and Schooling*

Early in the development of both the US and Canadian systems of education, special schools were created for the children of Native or First Peoples that stripped them of their culture and language and attempted to replace them with English, European moral sensibilities, and 'Christian' virtue. These sub-systems of education, including Bureau of Indian Affairs schools in the USA and church-run boarding schools in Canada, continued to function in this manner well into the twentieth century. At the start of the twenty-first century, some groups in Canada are still pursuing legal action against such schools for the abusive treatment of children. From the 1960s on, there has been considerable educational reform among Native communities including the development of a system of tribal colleges in the USA.

3.2 *From Segregation to 'A Nation at Risk': The Federal Role in US Public Schools*

In the former Confederate States of the USA, freed slaves made significant educational advances during Reconstruction until 1873, when the political imposition of Jim Crow laws and subsequent legalized segregation (flowing from the Supreme Court decision of Plessy vs. Ferguson, 1896) reversed the educational advancement of African-Americans and held it in abeyance until the 1950s. Beginning with the early Civil Rights movement in the USA, renewed political battles occurred over the role of the nation's schools. The US Supreme Court decision of Brown vs. the Board of Education (1954) overturned the legal support for school segregation, fomenting widespread social unrest and unprecedented school reform. In one of the few instances where federal troops have been used to countermand the orders of state leaders, national guardsmen opened a high school in Birmingham, Alabama and allowed a handful of black teenagers to enter. The US federal role in public elementary and secondary schools has continued to grow, and has increasingly become the focus of intense political conflict. Events such as the launch of the Russian satellite Sputnik and the publication of *A Nation at Risk* set off tremendous public debate about how schools were failing children and what could be done about it. While many have interpreted school reform since 1980 as an 'attack on the public schools,' attempts to create charter schools, provide vouchers for parents to send children to private schools, or pass laws making home schooling easier, are evidence of a continued belief that education is the best tool for creating a better society.

3.3 *Achieving Equity: Conflict over Language, Religion, and the Curriculum*

A concern with educational equity has dominated educational concerns since the 1960s but the focus has shifted from assuring equal access to ethnic or racial minorities to one of who ultimately controls what values the school will transmit. These battles—over language, the teaching of religion (e.g., offering Creation Science), and general course content—have been far more vociferous in the USA than in Canada. In the US constitution, there is no clear enumeration of educational rights comparable to that found in the Canadian Charter of Rights and Freedoms which specifically details the educational rights of language minority parents. This leaves open the possibility of political referendums, like the California act ending bilingual education, to substantially change the focus of schooling.

4. *The Systems Today*

Despite the progress made in establishing comprehensive, egalitarian systems of public education through the secondary level, the public school systems in the USA and Canada have not yet been able to fully rectify persistent problems of social inequality and conflicting cultural values. Yet the history of tension and educational experimentation has resulted in highly decentralized systems of free, public education, often emulated by other industrialized nations. In the Canadian system, under the British North America Act of 1867, provinces have wide discretion in setting educational policy, and there is no national ministry of

education. In the USA, individual states have broad powers to set educational standards but locally elected school boards exert considerable control on the day-to-day operation of the schools. Federal control over schools in the USA is indirect, typically using laws that require states to meet certain standards in schools in order to obtain federal funds for education to ensure compliance.

In both nations education is compulsory from about ages 6 or 7 to about age 16. In Canada, students make the transition from elementary to secondary schooling between sixth and eighth grade (with a slightly different system in Quebec). In the USA, despite the highly successful junior high school movement of the early twentieth century, there is no standard form of early secondary school organization. Completion rates for schooling are similar for the two nations: roughly 85 percent of 25- to 34-year-olds in 1996 had completed upper-secondary school (OECD 1998, p. 39). However, this statistic belies the fact that many US students do not complete high school but later complete general educational development (GED) tests. This statistic also fails to document the fact that minority students have statistically significant lower mean achievement scores, higher dropout rates, and lower rates of college enrollment than the majority students do. Detailed statistical reports on enrollment and other educational indicators are published yearly by the National Center for Education Statistics in the USA and by Statistics Canada.

The cultural emphasis on public school as the means of promoting social change or maintaining a peaceful and orderly society has often led scholars to ignore other nation-wide systems of education. In any catholic assessment of systems of education in North America, the role of tertiary education, military schools, and religious schools must be addressed.

4.1 The Tertiary Sector: Dominance in the World System

At the university level, in the USA, and to some extent in Canada, the cultural oppositions that affect elementary and secondary education have largely disappeared. If we compare two institutions that have had a profound impact on the development of college education in general—Harvard University and the University of Virginia—we find two distinct historical roles. For much of its history, Harvard educated the New England social elite and the Protestant clergy. UVA, while also transmitting social privilege, had a secular orientation that was replicated in the land-grant colleges set up around the US. Although land-grant colleges were sometimes modeled on elite private colleges (e.g. the University of Georgia and Yale University), these institutions focused on improving the conditions of the state populace and offered courses in pedestrian subjects such as agricultural or domestic science. Enrollment grew rapidly; most states now boast a network of high quality public colleges or universities accessible to a large segment of the populace.

Over time the differences between highly ranked private and public universities has declined, giving rise to the 'research university' that combines undergraduate training with advanced research in virtually every discipline currently studied. Many US and Canadian universities have eclipsed the European institutions on which they were originally modeled as international centers of knowledge production, and the USA is widely recognized as the world leader in the production of Ph.D. recipients.

4.2 Military Schools and Education

Over half of high school graduates go on to some form of college in the USA and Canada, but a large percentage of the population still uses the military as a means for educational advancement. For children of working-class families, military training and military service offer low-cost educational opportunities. Not only does the military provide basic instruction to its enlisted men and women, in both nations it also runs an independent system of elementary and secondary schools for the children of military personnel.

The US military also has a substantial tertiary presence. It operates highly regarded military colleges such as the Army Academy at West Point and funds students in public and private colleges through the officer training (i.e., the Reserve Officer Training Corps). Less visible, but not less important to tertiary education, the US Department of Defense funds extensive research projects, providing substantial external sources of revenues for many colleges.

4.3 Religious Schools and Churches

Churches and religious schools form another diverse system outside of public schooling. 'As in colonial times, the churches continued to serve as centers of formal and informal education.' (Cremin 1970, p. 383). The role of churches as centers for disseminating education and promoting social change was dramatically seen in the USA, when in the 1960s, Sunday sermons melded into political calls for ending segregation. At present, US and Canadian churches are host to a wide array of meetings that have an explicitly educational component. Large suburban churches typically offer a range of weekly classes and meetings ranging from independent piano lessons to church-sanctioned marriage classes. These activities are distinct from, and constitute a considerable educational impact over and beyond, the activities of private (US) or state-related (Canada) religious schools. Religious schools (primarily, but not exclusively Christian) are found throughout North America, ranging from local

kindergarten programs up to world-class universities like Notre Dame. Both nations experienced considerable political conflict over the role government should play in supporting these schools. In the USA political battles and court challenges (such as the conflict in the 1840s regarding Catholic schools) have generally been resolved to reinforce the separation of church and state, although groups like the Amish have mounted successful court challenges that have limited the state's ability to enforce compulsory schooling laws. In Canada many school boards were organized along denominational lines, and there is considerable variation in the relationship between provincial educational authorities and church-supported schools.

5. *Experimentation and Innovation: The North American Educational Dynamic*

Mr. Fredrick Douglas, a former slave, penned one of the most compelling statements of North American beliefs in the transformative power of education. Douglas wrote: 'Mr. Auld [his owner] … at once forbade Mrs. Auld to instruct me further, telling her, … that it was unlawful, as well as unsafe, to teach a slave to read. The very decided manner with which he spoke, … served to convince me that he was deeply sensible of the truths he was uttering … and the argument which he so warmly urged, against my learning to read, only served to inspire me with a desire and determination to learn' (Huggins 1980, pp. 42–3).

North Americans share a culture that places extreme emphasis on the power of education to transform both the individual and society and typically view schools as important institutions of social change or control. However, there is considerable cultural conflict over how schools are to be used: many citizens hold the view that a good government should mandate a system of free and comprehensive public education that is uniform for all, while others view any government intrusion into educational matters with extreme aversion.

North Americans also share a cultural tendency to engage each new educational innovation as a chance to move one step forward toward social utopia. This facet of North American culture may appear to others in the world as an obsession with educational reform, yet, this cultural dynamic has served to create a system of public education that is widely emulated. In the 1980s and 1990s, the performance of US and Canadian students on international tests of academic achievement caused many to declare the failure of North American schools and praise the educational culture of East Asia. Few saw the irony of the situation: educational reformers heavily influenced the Japanese school system. Educational policy makers in East Asia continue to look to North America as a center of educational reform and research.

Educational reform initiatives originating in North America will continue to influence educational systems around the globe. Looking toward the future, the most significant changes in systems of education will be wrought by the historical confluence of new electronic communication networks and a resurgence of political efforts to decentralize the control of schools. This can be seen in both the home schooling and charter school movements.

Widespread access to immense databases and electronic communication now allows parents to provide children with access to a world-class curriculum within the home, thereby weakening old claims that 'home learning' is greatly inferior to 'schooling.' While the majority of home-educated children belong to fundamentalist, Christian groups, a wide range of parents, anywhere from 300,000 to over half a million around 1995, now opt to school their children at home (US Department of Education 2000).

Charter schools, in the USA, have also increased in the wake of the educational restructuring movement of the 1980s and 1990s. These schools are publicly funded and are subject to governmental regulation but operate under fewer restrictions than ordinary public schools. There is great variation in state laws regarding the establishment of these schools, and only a small percentage of the US school-age population attend charter schools. But again, the ability of a small number of adults to access an immense amount of literature, music, and art via computer-based networks has made the establishment of such schools quite viable.

The demand for these, and other alternatives to public education, will continue to arise from the important cultural role that education has in North American society. The early part of the twentieth century saw the dramatic expansion of an organized public school system, then viewed as the best way to move society forward. The early part of the twenty-first century will no doubt see dramatic decentralization as families push for more autonomy over their child's education and individuals of all ages continue to resist any attempt to control what they can, or cannot, learn.

See also: American Studies: Education; Cross-cultural Study of Education; Education: Anthropological Aspects; Educational Institutions and Society; Educational Policy: Comparative Perspective; Educational Systems: Asia; Educational Systems: Australia; Educational Systems: Europe; Special Education in the United States: Legal History.

Bibliography

Cremin L 1970 *American Education: The Colonial Experience 1607–1783*, 1st edn. Harper and Row, New York

Equiano O 1995 *The Interesting Narrative*. Penguin Books, New York
Huggins N 1980 *Slave and Citizen: The Life of Frederick Douglas*. Little, Brown and Company, Boston
Kett J 1977 *Rites of Passage Adolescence in America, 1790 to the Present*. Basic Books, New York
National Commission on Excellence in Education 1983 *A Nation at Risk*. Washington, DC
Organisation for Economic Co-operation and Development (OECD) 1998 *Education at a Glance, OECD Indicators*. OECD, Paris
Parkman F 1983 *France and England in North America*. Library of America, New York, Vol. 1
Rush B 1786 *A Plan for the Establishment of Public Schools and Diffusion of Knowledge in Pennsylvania*. Thomas Dobson, PA
Sutton J 1988 *Stubborn Children*. University of California Press, Berkeley
Tyack D, Cuban L 1995 *Tinkering Toward Utopia: A Century of Public School Reform*. Harvard University Press, Cambridge, MA
US Department of Education. National Center for Education Statistics. 2000 *Issues Related to Estimating the Home-schooled Population in the United States With National Household Survey Data, NCES 2000–311*. Washington, DC

G. K. LeTendre

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Eff

Efficacy: Political

Political efficacy refers to citizens' beliefs that they can affect a political system. People with a high sense of political efficacy believe that they can, through their own individual or collective efforts, influence a political system that is regularly responsive to the actions of ordinary citizens. It is primarily a cognitive concept, a set of beliefs about one's citizenship role in relationship to governmental institutions. It is related to, but distinct from, affective or behavioral measures of political engagement.

Efficacy represents a subjective assessment of the political system. It differs from objective measures of a citizen's ability to affect a political system, such as political knowledge, or subjective measures of political engagement, such as political interest or civic obligation. One may feel politically efficacious while neither feeling motivated to participate in politics, nor judged by others to be capable of doing so.

Although political efficacy is generally measured at the individual level by social scientists, in practice it may also represent a propensity to draw upon collective social resources to effect political change. Political efficacy is significant not only because efficacious citizens have a greater propensity to engage in political action, but also because the presence of high levels of political efficacy within a population is important in creating a reservoir of citizen support for the political system.

1. *Intellectual Context*

Rooted in a social psychological theory of 'ego strength' (Lane 1959), Campbell et al. (1954) introduced political efficacy as one of several independent components of political engagement that predicted voter turnout but not voter preference. This basic role as a predictor of political participation has not changed much, although significant elaborations have been made in our understanding of that role.

Several years after its introduction as a predictor of political participation, scholars began to look at political efficacy as an indicator of existing and aspiring democratic systems. Almond and Verba (1963), attempting to understand the process by which aspiring democracies might achieve stability, conducted a comparative investigation of how political culture sustains existing liberal democracies. In order to connect the 'civic culture' concept, with its focus on political efficacy, to non-Western democracies, the authors included Mexico as one of five countries studied. Additional research by Indian scholars, with an awareness of the relative youth of Indian democracy, focused on the egalitarian distribution of political efficacy. They compared levels of political efficacy between different castes and between Hindus and Muslims. In particular, this research assessed whether lower castes had developed any sense of political efficacy since their inclusion in the Indian political system (Vir 1989).

Two separate lines of research have also used the concept of political efficacy as a yardstick to assess the health of mature liberal democracies. One is the differing levels of political engagement and participation by women and minorities, caused in part by differing levels of political efficacy (Abramson 1983). The second line emerged with the availability of repeated measures of political efficacy in the American National Election Studies (ANES). Converse (1972) showed decreasing levels of political trust and external (see below) political efficacy in the US, trends which continue to the present day. Easton's (1965) concepts of diffuse and specific support for a political system were applied to research on declining levels of trust and efficacy. According to Easton, democracies could continue to function well in the face of public dissatisfaction with current 'outputs' as long as a sufficient reserve of diffuse support (support for the structure of the political system) existed. However, Easton predicted that persistent public dissatisfaction and lack of specific support could erode diffuse support over the long term. This added urgency to findings of declining trust and efficacy.

2. *Conceptual Distinctions*

The distinction between internal and external political efficacy represents one of the most significant conceptual changes in understandings of political efficacy since the 1970s. While both types of efficacy refer to beliefs about the relationship of citizen political participation to the political system, they identify different explanations for the presence or absence of political efficacy. Internal political efficacy represents beliefs about the impact a person may have on the political process as a result of their own skills and confidence (or people just like themself). External political efficacy represents the impact a person believes they may have on the political process as a

consequence of political institutions' responsiveness. This distinction was implicit in earlier definitions of political efficacy, which addressed both components with the understanding that the two were strongly related (Lane 1959): internal feelings of efficacy arose in part because the political system was responsive, and at the same time, people perceived a political system to be responsive because they believed their actions could influence it. Indeed, subsequent studies of internal and external efficacy always found a significant and positive relationship between the two. But it was the performance over time of specific items measuring political efficacy in the ANES that suggested that the two components could behave quite differently (Converse 1972).

Balch (1974) first demonstrated empirically that the four ANES political efficacy questions behaved in ways congruent with a distinction between internal and external efficacy. In particular, he found that internal, but not external, political efficacy items had significant positive correlations with political interest, knowledge, and activity. External, but not internal, political efficacy items correlated positively and significantly with political trust. In general, these patterns have held up in subsequent research, to the extent that correlations between political efficacy and political engagement, activity, and trust serve as checks on the external validity of proposed internal and external political efficacy measures (Craig et al. 1990). Although this distinction has gained widespread acceptance, scholars have too often treated political efficacy as a single construct.

The performance of both internal and external efficacy, both in the aggregate and on the individual level, over time points to similarities and differences in the nature of the two concepts. Internal political efficacy has largely remained stable since the late 1960s while external political efficacy has declined steadily. Both aspects of political efficacy are relatively stable at the individual level. Internal political efficacy is strongly related to changes in an individual's socioeconomic background and life stage (Abramson 1983). By contrast, external political efficacy appears to be sensitive to changes in the political environment, but is less responsive to changes in individuals' characteristics. While the presence or absence of specific elected officials may affect external political efficacy, individuals' reactions appear to be heavily related to their partisanship. External efficacy rises for those whose favored party is in power, with the presidency representing this incumbency effect for the US (Craig 1993, Miller and Listhaug 1990). Such effects may be quite specific, even distinguishing between party control and local and federal levels (Stewart et al. 1992).

Some scholars distinguish between political trust as a judgment about governmental performance regardless of citizen participation, while external political efficacy is a judgment of the responsiveness of government to citizens' expressed preferences. However, as Balch (1974) noted earlier, the correlation between political trust and external political efficacy is often higher than the association among political efficacy items. This persistent pattern has led some scholars to question the distinction between external political efficacy and political trust. As a result, external political efficacy and political trust have at times been treated as broad indicators of political system support. In part this is due to their similar pattern of decline over time among Western democracies (Craig 1993, Miller and Listhaug 1990).

3. *Nature of Political Efficacy*

Prewitt's (1968) assessment of the concept of 'political efficacy' focused on the personal antecedents as well as the consequences of political efficacy. Education, socioeconomic standing, and ethnic majority status were positively related to political efficacy. Higher levels of political efficacy in turn predicted higher levels of a range of political behaviors, voting being the central concern. Political efficacy was also thought to play a key role in general citizen support for a democratic system, although this hypothesis was largely untested. Subsequent research provides confirmation for some of these findings and broadens others. Education and socioeconomic standing continue to predict levels of political efficacy, while the disadvantage accrued to ethnic minority status appears to be dependent on local political context (Abramson 1983, Vir 1989).

Thus, the earliest studies of political efficacy examined personal antecedents and went on to argue that political efficacy affects the political system. One major trend in research on political efficacy has been to complete the circle of causation by examining the structural antecedents of political efficacy. Pateman's work (1970) occupies a central place in this tradition. She drew on the work of liberal democratic theorists as well as findings by Almond and Verba (1963) to argue that all forms of participation, broadly conceived to include nonpolitical settings, could educate citizens to take part in self-government. Political efficacy was one measure of such educative effects and although her work suggests a specific focus on internal political efficacy, much research following her thesis has focused on external political efficacy.

Of special concern to Pateman were the effects of workplace conditions. Workplaces that emphasized democratic decision making would teach necessary skills for political participation and be intrinsically valuable, contributing to citizens' self-development. But Pateman's general thesis—that nonpolitical and political participation has educative effects in a democracy—has been tested mainly in areas besides work. Soss (1999), for example, successfully argued that how public policy is implemented has significant conse-

quences for socializing political efficacy. Clients generalize from the responsiveness of administrators in public benefit programs, and from opportunities to shape the implementation of policy to responsiveness of government in general, producing an impact on external political efficacy that contributes to voting turnout.

Finkel (1985, 1987) tested Pateman's underlying assertion that political participation has an educative effect on citizens. Using panel studies in the US and West Germany, he found that although levels of internal political efficacy affect conventional participation, the reverse is not true. By contrast, external political efficacy is affected more directly in a reciprocal manner by political participation.

Related research examined the effects of the larger political system on political efficacy. Scholars have identified persistent gender and race differences in internal political efficacy. Blacks, Hispanics, and women consistently report weaker feelings of internal and external political efficacy than whites and men. Abramson's (1983) explanation of this difference with regard to schoolchildren has been mirrored in later studies of adults. Abramson's basic argument, based on previous studies including his own with children, is that black children are influenced indirectly by the presence or absence of blacks in positions of political power. Differences among minorities with respect to political efficacy reflect actual low levels of political influence.

While some scholars emphasized the structural antecedents of political efficacy, others examined the causes and consequences of declining external political efficacy and trust. Gamson (1968) proposed that unconventional behavior would result among citizens whose sense of personal political efficacy was strong but trust in the political system was low. Subsequent evidence for this interaction effect between trust and political efficacy has been rather weak, however (Craig 1993; Sigelman and Feldman 1983). Scholars seeking to test Gamson's thesis have disagreed on a number of issues concerning its operationalization. Greater support for Gamson's hypothesis appears to lie in considering the context of unconventional behavior, especially the presence or absence of viable challengers to a political regime.

Other scholars have investigated the causes of declining levels of political support, including external political efficacy and political trust. Such work has been influenced by Easton's (1965) analysis of the implications of specific and diffuse political support for a political system. There is disagreement over whether political trust measures specific support while external political efficacy measures diffuse support for a political system, or whether both are indicators of specific support. This research found multiple causes of the decline of political support but has assigned at least part of the blame to a general perception of governmental unresponsiveness. Craig (1993), who focuses on the importance of procedural values to the American public, argues that the gap between a growing sense of 'democratic entitlement' among the public and the failure of political elites to respond accordingly is the main cause of declining political support. Miller and Listhaug (1990), in a comparative study of Sweden, Norway, and the US, argue that a perceived lack of representation by the party system has led to an erosion of political support. Declining specific support for particular institutions and parties is translated into a lack of diffuse support for the party system in general.

4. *Measurement of Political Efficacy*

The American National Election Study (ANES) has provided the most frequently used measures of political efficacy. The main advantage of these measures is their continuity. Three of the five original items used in 1952 have been asked on each ANES study, including two items that are taken as measures of external efficacy, and one that is taken as a measure of internal efficacy. Aside from this, a single potential measure of external political efficacy, asked on the General Social Survey since 1973, is the only other measure available over time.

Although many scholars have adopted the ANES measures, a number of different measures of political efficacy exist, often combining political trust or social alienation with political efficacy. The differentiation of efficacy by level of government (local, state, federal), and even specific governmental institutions appears to be the most useful recent innovation (Reef and Knoke 1999).

If the continuity of the ANES measures is their main advantage, the imprecision of these measures constitutes their main drawback. Craig et al. (1990) analyzed new efficacy and trust items alongside regular ANES efficacy and trust items in the 1987 pilot survey. Their results cast doubt on the separation of external political efficacy and trust, at least when both address incumbents rather than political procedures. While they validated one ANES item used consistently to measure internal political efficacy, their suggestions for additional internal political efficacy measures were not consistently adopted. In addition, a potential hazard to the use of ANES measures over time is a change in the response set. Beginning in 1988, respondents were also offered a 'neither agree nor disagree' option. Scholars who study these items suggest that the change has not substantively affected the continuity of these measures.

Political efficacy has proven to be exceptionally amenable to combining multiple investigative techniques. Such techniques include more qualitative methods such as in-depth interviews, open-ended questions in structured interviews, and even observational research, as well as quasi-experimental work

(Almond and Verba 1963, Craig 1993, Soss 1999). This research generally involves using qualitative methods to generate hypotheses or illustrate trends in secondary data, and using quantitative analysis of secondary data to confirm hypotheses.

5. Future Directions

Recent trends in research on political efficacy focus on the relationship between specific political institutions and political efficacy as contrasted with political systems as a whole. Future research will likely rely on a reciprocal framework, examining the consequences of efficacy for support of specific institutions and the 'educative effects' of such institutions. This may include a focus on the implementation of public policy, along the lines of Soss's (1999) work or specific governmental institutions, particularly local and state. Earlier work along these lines consistently found effects for external, but not internal, political efficacy.

The effect of political culture on political efficacy has been under-examined. The few comparative studies of political efficacy have consistently supported the basic findings about political efficacy, such as the effects of education. But such studies have also consistently found that political history plays a significant role, either in mediating the impact of events or of individual characteristics on political efficacy, or as an independent contributor to existing levels of political efficacy (Almond and Verba 1963; Finkel 1987; Miller and Listhaug 1990).

Future research is likely to compound fragmentation of a concept that has undergone at least one significant modification since the 1970s. At the same time, research that points to specific effects by government institutions or political culture may provide insight into the more general nature of political efficacy. In addition, recent focus on trust as a fundamental element of social capital suggests that it would be wise to incorporate political efficacy into the analysis. The fact that both trust and efficacy have declined significantly indicates that attempts to rebuild reservoirs of social capital ought to attend to both trust and efficacy.

See also: Action, Collective; Action, Theories of Social; Citizen Participation; Civic Culture; Democracy; Democratic Theory; Ego Development in Adulthood; Ego Psychology and Psychoanalysis; Participation: Political; Political Culture; Political Protest and Civil Disobedience; Social Movements, Geography of; Social Movements, History of: General

Bibliography

Abramson P R 1983 *Political Attitudes in America: Formation and Change*. Freeman, San Francisco

Almond G A, Verba S 1963 *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton University Press, Princeton, NJ

Balch G I 1974 Multiple indicators in survey research: The concept sense of political efficacy. *Political Methodology* **1**: 1–43

Campbell A, Gurin G, Miller W E 1954 *The Voter Decides*. Row, Peterson, Evanston, IL

Converse P E 1972 Change in the American electorate. In: Campbell A, Converse P E (eds.) *The Human Meaning of Social Change*. Sage, New York

Craig S C 1993 *The Malevolent Leaders: Popular Discontent in America*. Westview Press, Boulder, CO

Craig S C, Niemi R G, Silver G E 1990 Political efficacy and trust: A report on the NES pilot study items. *Political Behavior* **12**: 289–314

Easton D 1965 *A Systems Analysis of Political Life*. Wiley, New York

Finkel S E 1985 Reciprocal effects of participation and political efficacy: A panel analysis. *American Journal of Political Science* **29**: 891–913

Finkel S E 1987 The effects of participation and political efficacy and political support: Evidence from a West German panel. *Journal of Politics* **49**: 441–64

Gamson W A 1968 *Power and Discontent*. Dorsey Press, Homewood, IL

Lane R E 1959 *Political Life: Why People Get Involved in Politics*. Free Press, Glencoe, IL

Miller A H, Listhaug O 1990 Political parties and confidence in government: A comparison of Norway, Sweden, and the United States. *British Journal of Political Science* **20**: 357–86

Pateman C 1970 *Participation and Democratic Theory*. Cambridge University Press, Cambridge, UK

Prewitt K 1968 Political efficacy. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York

Reef M J, Knoke D 1999 Political alienation and efficacy. In: Robinson J P, Shaver P R, Wrightsman L S (eds.) *Measures of Political Attitudes*. Academic Press, San Diego

Sigelman L, Feldman S 1983 Efficacy, mistrust, and political mobilization: A cross-national analysis. *Comparative Political Studies* **16**: 118–43

Soss J 1999 Lessons of welfare: policy design, political learning, and political action. *American Political Science Review* **93**: 363–80

Stewart M C, Kornberg A, Clarke H D, Acock A 1992 Arenas and attitudes: A note on political efficacy in a federal system. *Journal of Politics* **54**: 179–96

Vir D 1989 *Political Efficacy in Urban India: A Sociological Exploration*, 1st edn. Classical, New Delhi, India

J. L. Sullivan and E. Riedel

Egalitarianism: Political

Equality is one of the foundational principles of modern democratic states. Every citizen—man or woman, black or white, rich or poor—must be treated equally in certain respects. They must, for example, have equal rights and equal opportunities. Only racists and a few other such extremists would deny this. Yet it remains intensely controversial how far this equality

of treatment must go. Must citizens enjoy economic and social equality as well as political equality in the sense of equal civil and political rights, for instance? Here conservatives, liberals, socialists, and feminists part company. This article looks briefly at the origins of egalitarianism, and then at the different forms of egalitarianism advocated in contemporary democracies.

1. *The Origins of Egalitarianism*

As just noted, in modern democracies equality in one or other of its versions has gained widespread acceptance as a political principle. For most of human history, however, this was not so. The idea that all human beings are, at some level, one another's equals can be found in the major religions and in a variety of cultures. But this foundational equality did not translate into a demand that human beings should associate politically on an equal footing. On the contrary, although a society of equals was sometimes presented as a Golden Age, or a utopian vision of the future, the current human condition was thought to require hierarchical relations in which men exercised authority over women, masters over slaves, lords over vassals and so forth. Accordingly if we search through the texts of classical or medieval political thought, we do not find equality held up as a principle that should inform existing practice (see Lakoff 1964)

This began to change in the early modern period, with the emergence of the centralized state and the accompanying gradual transition to a market economy. The state was founded on the assumption that everyone living within its territorial boundaries was to be equally subject to its authority and its laws. This idea was expressed with particular clarity by contractarian thinkers such as Hobbes, Grotius, and Locke, who argued that the state emerged from a covenant between men whose natural condition was one of equality: each had equal rights by nature, and they would therefore consent only to a state whose laws afforded them equal protection. Meanwhile the gradual shift away from feudal economic relations, in which people were linked together in a vertical hierarchy, toward a market economy, in which contracts and exchanges took place between individuals who were at least formally equal to one another, fostered the notion that each person had an equal right to freedom, and an equal right to the fruits of his labor.

By the end of the eighteenth century these principles had become widely accepted in Western societies, and found expression in the two great political documents of that era—the American Declaration of Independence and the French Declaration of the Rights of Man and of the Citizen—both of which boldly inscribed the principle of equality at the heart of political life. In practice, however, equality was still very far from being realized: large sections of the population (most notably women and the working class) were still denied the vote, many Americans found it possible to reconcile the equal rights of mankind with the practice of slavery, and economic opportunities still depended very largely on inherited social position. The story of the last two centuries can be seen from one angle as the steady (though still far from complete) overcoming of these inequalities in the name of equality itself.

2. *Political Equality*

The fierce debates in the nineteenth and early twentieth centuries over the extension of full political rights first to the working class and then to women have long since ended, and yet the issue of political equality is still far from settled. One problem has to do with representation. Even though every citizen in contemporary democracies has an equal right to vote, those elected are still primarily male, middle class, and drawn from majority ethnic groups. Does this matter? The answer depends on whether political representation means the representation of interests, and if so whether the elected members are able adequately to represent the interests of those who belong to a different section of the population. Critics argue that political equality requires the direct presence, in elected bodies, of people drawn from different groups in society, who are then able to articulate the perspectives and interests of those groups in an undistorted way (see Phillips 1995).

A second problem has to do with majorities and minorities—a problem, it should be noted, that may arise even within fully representative political institutions. At first sight it might seem as though political equality entails the majority principle for making decisions, since this is the only method of deciding issues that gives each vote an equal weight. But consider a case in which a relatively apathetic majority defeats a minority whose members care intensely about the issue that is being decided; or consider a case in which a minority group finds itself repeatedly outvoted in a series of decisions that affect its interests. Reflection on these cases suggests that political equality entails more than just formally equal voting rights. It requires something like an equal chance to influence the outcome of the decision procedure. But it is hard to design institutions that can achieve political equality in this fuller sense (see Beitz 1989).

3. *Equality of Opportunity*

Alongside political equality, equality of opportunity—the principle that each person should have an equal chance to gain jobs, offices, and other positions of advantage—has become a staple of liberal-de-

mocratic thinking. Yet here, too, we find ongoing disagreement about how the principle should be understood. In its original formulation, it meant simply that there should be no legal barriers preventing those with certain characteristics (class or religious affiliation, for example) from attending schools and universities, entering careers or holding public office. But this left in place the practical obstacles, chiefly lack of resources, that effectively closed off such options to many citizens. Many people therefore now take equality of opportunity to mean what John Rawls has called fair equality of opportunity, the principle that 'those who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the social system' (Rawls 1971, p. 73). Taken literally, this calls for the elimination of all those disparate influences—family background, school choice, and so forth—that today help to determine a person's likely social destination. Most liberal thinking opts for the middle ground: the state should try to mitigate disadvantage by, for example, curtailing the right to inherit, and insuring that children from all backgrounds have access to good schools, but it should not interfere directly with the transmission of cultural and other advantages through the family (see Fishkin 1983 for an analysis).

4. *Equality of Outcome*

The most radical versions of egalitarianism hold that equality of opportunity is not enough, since even in its stronger Rawlsian guise it still permits a person's life prospects to be determined by their talent and ability. Instead it should be the goal of public policy to make every citizen as materially well off as every other, except insofar as people can be held personally responsible for having less (thus equality of outcome, as it is usually understood, does not require that people who gamble away their share of resources should subsequently be compensated).

Several questions may be asked about equality of outcome. The most basic has to do with the respect or respects in which people's condition is to be made equal—what is sometimes referred to as the currency of equality. The natural currency might seem to be that of welfare: how much happiness or well-being each person enjoys. This suggestion faces acute difficulties of measurement—how can we decide when one person is happier or has more welfare than another?—but beyond that it may be criticized on grounds of principle. Someone's level of welfare depends in part on how far they are able to realize their aims and ambitions, but should they not be held responsible for having these aims and ambitions in the first place? Should equality involve giving more material resources to people whose welfare is reduced because they cannot fulfil expensive ambitions such as consuming fine wines or collecting great works of art?

Some contemporary liberals, most notably Ronald Dworkin, have therefore advocated equality of resources as a better alternative to equality of welfare (Dworkin 2000). People are equal, on this view, when they have access to resources of equal value, which they can then choose to use in whatever way their preferences dictate. Dworkin proposes the device of an auction as a way of determining when external resources are equally distributed. He has greater difficulty when he comes to deal with internal resources, the talents and abilities that determine what people can do with the external resources they are given. How can equality be achieved here without forcing highly talented individuals to use their talents in ways that they may strongly dislike—the so-called 'slavery of the talented' (Roemer 1985)?

A further problem for equality of resources is that it appears not to accommodate people whose special needs mean that they derive less welfare than others from a given bundle of resources almost regardless of their preferences about how to live. In recognition of this problem, a third interpretation of equality of outcome has been proposed by Amartya Sen: basic capability equality, according to which people should as far as possible be made equally capable of achieving a set of functionings, such as being well nourished, moving about, etc. (Sen 1992). Sen claims that this conception focuses, properly, on what resources can do for people, as opposed to the bare fact of having resources, while it avoids the difficulties that arise when equality is defined in terms of happiness or preference satisfaction. It does, however, face the problem of comparing different capability sets. If I can do A, B and C and you can do D, E and F, how do we decide whether we are equal or unequal to each other, without surreptitiously referring to a criterion such as welfare?

5. *Social Equality*

Equality of outcome faces serious problems beyond that of finding a suitable currency in which to express it. The most rehearsed of these is perhaps the incentives problem: if everyone is to be made equally well off at the end of the day, what incentive is there for people to use their talents in productive ways? A different line of criticism challenges the view that equality of outcome is something we should care about ethically. Instead, it is argued, we should be concerned about sufficiency—about insuring that everyone reaches a standard of living sufficient to satisfy their essential needs and interests, or alternatively about giving priority to the worst off—about insuring that public policy attaches special weight to improving the material situation of those who have least. (See Raz 1986,

Chap. 9, Frankfurt 1987, Parfit 1998 for arguments of these kinds.)

Another view maintains that inequalities of outcome matter only when they create social inequalities. On this view, what matters is that each member of a given society should feel himself or herself to be the equal of every other, that there should be no inequalities of status such that A regards himself as socially inferior to B in an across-the-board sense. One way to achieve this is to prevent inequalities from accumulating in such a way that when B enjoys more goods of one kind than A, he also enjoys more goods of other kinds. Michael Walzer has applied the term 'complex equality' to a society where different people are ahead along different dimensions of advantage (some enjoy greater wealth, others more power, yet others are more highly educated, and so forth) and there are blocking mechanisms that prevent people from using their advantaged position on one dimension to gain unjust advantages on others (Walzer 1983). One merit of this way of understanding equality is that it allows justice and equality to be reconciled, even in cases where justice itself mandates inequality, such as cases in which some people deserve more than others. Such justified inequalities need not threaten social equality provided Walzer's conditions are met. (See Miller 1999, Chap. 11 for a fuller discussion.)

See also: Democracy; Democracy, History of; Democracy: Normative Theory; Egalitarianism, Sociology of; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Social Stratification

Bibliography

Beitz C R 1989 *Political Equality: An Essay in Democratic Theory*. Princeton University Press, Princeton, NJ

Dworkin R 2000 *Sovereign Virtue: The Theory and Practice of Equality*. Harvard University Press, Cambridge, MA

Fishkin J S 1983 *Justice, Equal Opportunity and the Family*. Yale University Press, New Haven, CT

Frankfurt H G 1987 Equality as a moral idea. *Ethics* **98**: 21–43

Lakoff S A 1964 *Equality in Political Philosophy*. Harvard University Press, Cambridge, MA

Miller D 1999 *Principles of Social Justice*. Harvard University Press, Cambridge, MA

Parfit D 1998 Equality or priority? In: Mason A (ed.) *Ideals of Equality*. Blackwell, Oxford, UK, pp. 1–20

Phillips A 1995 *The Politics of Presence*. Clarendon, Oxford, UK

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Raz J 1986 *The Morality of Freedom*. Clarendon, Oxford, UK

Roemer J 1985 Equality of talent. *Economics and Philosophy* **1**: 151–86

Sen 1992 *Inequality Reexamined*. Clarendon, Oxford, UK

Walzer M 1983 *Spheres of Justice: A Defense of Pluralism and Equality*. Martin Robertson, Oxford, UK

D. Miller

Egalitarianism, Sociology of

Egalitarianism is an ideal that motivates many normative political visions and many social movements. It is a relatively modern concern because it is only in relatively advanced nations that states have the capacity to redistribute systematically on a grand scale. Indeed, they have had the capacity to tax incomes effectively only since the early nineteenth century. Most contemporary discussions of equality are reformist and addressed to reducing inequalities, but many earlier movements were radical and addressed to remaking society. There are two classes of obstacles to achieving egalitarianism: causal problems and value-theory problems. The chief causal problems are that, in achieving equality, we typically must affect other things, such as productivity (Rawls 1971, Okun 1975, Bowles and Gintis 1998). We cannot achieve equality while holding all else equal. The chief value-theory problems are suggested by the questions: equality of what? and; how do we measure it?

The philosophical literature on egalitarianism is deeply equivocal on the questions of what we should equalize and how we could measure it. Social scientists most commonly suppose that the answers are simple; we should equalize incomes, measured in monetary terms only. The most conclusive arguments are about obstacles to achieving equality or even to reducing inequality very much. These causal arguments are about the tradeoffs that we must make if we are to have complete or even greater equality. I will first address these tradeoffs and then turn to measurement issues. Finally, I will briefly comment on international inequality.

1. Tradeoffs

The chief appeal of the market organization of an economy is that it enhances production—but it does so by introducing massive inequalities of benefit from social organization. Most of the concern with tradeoffs in achieving greater equality is with the effects on incentives for production. Having strict equality of incomes is equivalent to turning national production into one vast problem in which the collective national income is produced by collective action. That income would then be divided equally among all citizens (or all families or producers) independently of the scale of their personal contributions to the total product. Hence, individuals would have no incentive to contribute much to the collective good beyond altruism or public-spiritedness. Even if we merely wish to reduce inequality of income, we must tax away part of the higher incomes to transfer money to those with the lowest incomes. As we do that, the highest earners may increasingly prefer to substitute leisure for income, because their effective hourly wage is reduced so that leisure, which is untaxed, costs less than other things

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

their income could buy for them. Consequently there will be lower overall production and there will be less to tax and transfer, so that greater equality will entail lower average income. If government transfer is inefficient, not least because it requires labor that is otherwise not productive, the reduction of average income will be exacerbated. There are other detrimental incentive effects. Personal investment in human capital to be productive will not be rewarded if incomes are equal, and it may therefore be substantially reduced. Also, the incentives to entrepreneurs to take large risks will be eliminated.

For many collective achievements, a hierarchical organization seems to be necessary. Although hierarchy without inequality of wealth or income is conceivable, high status in a sharp hierarchy is likely to be an unequal reward in its own right. Very generally, we require extensive division of labor if our society is to be prosperous, diverse and interesting, and such division is itself likely to be hierarchical and unequal in its rewards of status and other nonmonetary emoluments. Émile Durkheim (1933) argued that the equality that fits with relatively primitive conditions of mechanical solidarity gives way to the division of labor and the organic solidarity that it entails. The emoluments from hierarchy and many other things cannot be equalized (Hirsch 1976). Finally, inequality of wealth may actually be enormously beneficial to technological progress in many areas, because the relatively wealthy will buy new technologies while their prices are far too high for a mass market. Their early entry into the market often then drives down the costs of production so that the new technologies become much more widely available, as has been true for computers in the 1980s and 1990s. The green revolution of developing nations was made possible by the prior investment in wealthy nations in agricultural innovations that, as they became cheaper, became attractive to much poorer nations. For example, chemical fertilizers, highly productive seeds, and light tractors have all become cheap. Similarly, after wealthy nations had eliminated smallpox infection in their own populations, they worked successfully to eliminate it everywhere, in part for their own benefit. They now attempt to do the same for polio and there are promising prospects for malaria, cholera, and other pervasive diseases. F. A. Hayek (1960, p. 40) supposes that 'if all had to wait for better things until they could be provided for all, that day would in many instances never come.'

Finally, the evidence of twentieth-century experiments in national programs of reducing inequality suggest a grim tradeoff. Because incentives can be extraordinarily powerful in stimulating unequal efforts, any substantial degree of equality in a whole society is likely to require strong government control over distribution and wages and perhaps even over production. If government is empowered to intervene in such ways to make major redistribution, it is very likely also to be empowered to do other, generally harmful things. Stalin and Mao sat in capricious, brutal judgment of artists and their work. Many national leaders, including Nehru and many revolutionary idealists, have supposed that they should design and run the economy even though none of them has been a competent economist. Indeed, even the seemingly best economists historically had theories that, if put into policy, would have been destructive. As David Hume, Adam Smith, and James Madison insisted, weak government has many appeals.

2. *Measurement*

Conceptually, perhaps the first instinct is that we should equalize welfare or well-being as far as possible. In the days of cardinal utility theory, comparing welfares was a simple matter in principle, although impossible in practice. In the twentieth century, economists largely rejected cardinal welfare theory and adopted an ordinal theory in which it is not even possible to make interpersonal comparisons. If we can make at least ordinal comparisons, then we can compare whole states to each other, so that we might be able to say that India is worse off than China, or vice versa, or that one very general policy is more egalitarian than another. Such ordinal evaluation can have seemingly anomalous implications. For example, an inegalitarian Bangladesh can be compared to Iceland if its 250,000 best off citizens are better off than the citizens of Iceland, while a Bangladesh with the same resources overall, but more equally distributed, might not be comparable one way or the other to Iceland. Hence, the inegalitarian Bangladesh might be found to have greater welfare than Iceland, while the egalitarian Bangladesh might be found to be, on a Rawlsian or other assessment, more just than the inegalitarian Bangladesh. The egalitarian Bangladesh might, however, not be comparable to Iceland either on a Rawlsian or an ordinal utilitarian assessment. The first of these pairs permits a welfarist comparison, the second permits a Rawlsian justice comparison, and the third permits neither comparison. Although such implications may sound strange on first hearing, they are likely to be commonplace in an ordinal world. If such implications sound very strange, we may find it difficult to redirect our understanding to think in ordinal terms. But that is what we must do if we wish to make consequentialist sense of our real world.

Because welfare is hard to measure, many analysts focus on other things that might be equalized. Perhaps because they seem like a reasonable proxy for welfare, resources are the most common alternative focus. Among the earliest visions of an egalitarian society was that of the seventeenth-century Diggers in England, who wanted equal plots of land for each family to farm (Winstanley 1973). They thought that even the wealthy should be reduced to simple farming,

because such a life of hard work would make them better Christians. Hence, theirs was a program of equalization both up for the poor and down for the wealthy. (Many political movements and regimes from the Diggers to Pol Pot and the Taliban have sought to achieve equality by reducing the welfare of the best off rather than by raising the welfare of the worst off.) The central focus of much of the debate on equality has long been private ownership of property, especially of the means of production. For Karl Marx, such ownership meant that nonproductive owners gained much of the profits of production. The route to greater equality was public ownership of such property so that all income would be in the form of wages from labor.

3. *International Inequality*

In the twentieth century, several major societies attempted to achieve greater equality in the midst of other societies that were highly inegalitarian or to overcome their own poverty by closing themselves off from the world economy and centrally directing their own development. Such autarky can have substantial costs. The Soviet recourse to autarky was partially imposed from outside, but Nehru's relatively autarkic policies in India were plausibly motivated by desires to keep India out of the commercial mold of the West and to maintain control over the prospects for greater equality. Unfortunately, one nation cannot be as creative and innovative as the entire rest of the world, and impoverished Indians were made relatively worse off internationally during the 40 years of Nehru's legacy. When a nation blocks inequalities that other nations allow, many citizens have strong incentive to emigrate. Some states have blocked such emigration with harsh controls on movement, so that again egalitarianism forces a grim tradeoff.

In much of the contemporary public discussion of the poverty of developing nations, the focus is often on achieving greater equality through straight redistribution from the wealthy nations, as though the issue were the static problem of current distribution of what is already available. This is almost entirely wrong-headed. The crux of the problem of grievous inequality between very poor communities and prosperous communities is dynamic; it is to enable people in impoverished communities to become engaged in productive work that will raise their level of welfare. Transfers that are aimed at providing work opportunities make sense, as do transfers that are aimed at momentary emergencies such as drought and starvation in the Sahel. Transfers that merely equalize static distributions cannot be practically justified in a world in which the problem is inherently dynamic.

What impoverished nations need is investment in production facilities, such micro-lending institutions as the *Grameen Bank*, human capital, and good legal systems. The chief differences between the advanced and impoverished economies are not natural resources but human capital and the social capital of relevant institutions. The truth of this claim is now revealed to an extreme degree by the wealth-generating information industries whose inputs are almost exclusively human and social capital (and, of course, intellectual property). The actual contribution of natural resources to production in the advanced economies is very slight (apart from energy resources, which are not found in many of the most advanced economies). Any move to address inequality that ignores this issue of production is hardly of any interest. We must consider institutions for improving the prospects for internal economic production. Indeed, apart from emergency cases, this should be the first and primary focus for reducing international poverty and gross inequality.

See also: Altruism and Self-interest; Collectivism: Cultural Concerns; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Hayek, Friedrich A von (1899–1996); Hierarchies and Markets; Income Distribution; Public Goods: International; Utilitarian Social Thought, History of; Utopias: Social

Bibliography

Bowles S, Gintis H 1998 *Recasting Egalitarianism: New Rules for Communities, States, and Markets*. Verso, London

Durkheim É 1933 *The Division of Labor in Society*. Macmillan, New York

Hardin R 1988 *Morality Within the Limits of Reason*. University of Chicago Press, Chicago

Hayek F A 1960 *The Constitution of Liberty*. University of Chicago Press, Chicago

Hirsch F 1976 *Social Limits to Growth*. Harvard University Press, Cambridge, MA

Okun A 1975 *Equality and Efficiency: The Big Tradeoff*. Brookings, Washington

Rawls J 1971 A *Theory of Justice*. Harvard University Press, Cambridge, MA

Roemer J E, Arneson R J *Equal Shares: Making Market Socialism Work*. Verso, London

Winstanley G 1973 *The Law of Freedom in a Platform or, True Magistracy Restored*. Schocken Books, New York

R. Hardin

Ego Development in Adulthood

Jane Loevinger's theory of ego development (Loevinger 1976, Hy and Loevinger 1996) has provided an approach to personality, emotion, and values that is distinct from approaches that are concerned with the assessment of stable dimensions or traits.

Table 1
Characteristics of levels of ego development

Level	Code[a]	Impulse control	Interpersonal mode	Conscious preoccupations	Example
Impulsive	E2 (I-2)	Impulsive	Egocentric, dependent	Bodily feelings	… shut up
Self-protective	E3 (Delta)	Opportunistic	Manipulative, wary	'Trouble,' control	… date as many guys as she likes
Conformist	E4 (I-3)	Respect for rules	Cooperative, loyal	Appearances, behavior	… ask her mother's and father's advice
Self-aware	E5 (I-3/4)	Exceptions allowable	Helpful, self-aware	Feelings, problems, adjustment	… make her own decisions about any job she wants to have
Conscientious	E6 (I-4)	Self-evaluated standards, self-critical	Intense, responsible	Motives, traits, achievements	… do what she wants as long as it is within reason and it does not hurt anyone including herself
Individualistic	E7 (I-4/5)	Tolerant	Mutual	Individuality, development, roles	… do anything men can do but still being a feminine person in the process
Autonomous	E8 (I-5)	Coping with conflict	Interdependent	Self-fulfillment, psychological causation	… realize her potential, regardless of …the role restrictions society may try to impose
Integrated	E9 (I-6)		Cherishing individuality	Identity	[very rare, probably less than 1% of responses; no example given]

Source: revised after Hy and Loevinger (1996, Table 1.1, p. 4).
[a]Notation in parentheses is original notation; 'E' notation is notation revised by Hy and Loevinger (1996).

Instead, ego level theory focuses on how individuals organize meaning and information relevant to the self or ego. Loevinger suggests that such organization of meaning changes as a function of the conceptual complexity of individuals. She distinguishes several ego levels, which are derived from developmental theory. However, these levels not only serve to assess time-related or age-related aspects of personality organization, but also are important in understanding individual differences.

1. *Definition of Ego Levels*

Loevinger's theory borrows from both cognitive-developmental theory and from psychodynamic approaches (Loevinger 1976). From cognitive-developmental theory, it borrows and develops the notion that more mature levels of ego functioning emphasize aspects of complex and abstract thinking—and thus are oriented towards objectivity and impersonality. From psychodynamic theory, it introduces the assumption that abstract thinking is not merely impersonal, but that others and reality in general are approached in terms of the relation or meanings they have for the individual—with his or her needs, anxieties, and goals, with the impulses and methods of controlling them, personal preoccupations and ambitions, interpersonal attitudes and social values (Blasi 1998). All of these aspects are described in terms of a series of less to more complex levels of ego development.

Ego levels are assessed through the sentence completion test (Hy and Loevinger 1996). This test consists of 36 sentence stems such as 'Raising a family … ,' 'Being with other people … ,' or 'A girl has a right to … .' Individuals are asked to complete these stems, and responses are coded along several dimensions such as (a) impulse control and character development; (b) interpersonal style, such as independence, manipulativeness, need to belong, autonomy; (c) conscious preoccupations; and (c) cognitive style, such as tolerance of ambiguity and tendency towards objectivity. Table 1 presents coding criteria and sample responses for the item, 'A girl has a right to … .' At lower levels, responses are dictated by impulsivity, egocentrism, and bodily feelings, while at the higher ends they are characterized by acceptance of conflict, the realization of interdependence and a cherishing of individuality, and a focus on self realization and psychological causes.

2. *Ego Levels and Age/Development*

Since ego levels are derived from developmental theory, an obvious question to ask is how they relate to age or development. Research indicates that from childhood to adulthood, age is positively related to ego level. However, throughout adulthood, the relation between age and ego level may vary; depending on the circumstances, ego level may even decline. This is to be expected if one accepts that age and development are not necessarily identical. Thus individuals may grow older but not necessarily more mature or complex. Hence, one would expect patterns of change to vary with the conditions in which individuals age.

In support of the notion that age and development can diverge in adulthood, Loevinger et al. (1985) suggest that the pressure towards conformity in adulthood can actually lead to a lowering of ego level. In general, one would expect that ego level is enhanced by experiences that challenge one's adaptive strategies. If individuals are in stable niches, no ego level growth may be observed. As Block (1982, p. 289) stated, 'The passage of time does not mean development.'

Helson's work (e.g., Helson and Roberts 1994) has paid particular attention to the contexts that may encourage or discourage growth in ego level in her long-time study of women's development. While in general ego level is strongly predicted by such variables as intelligence and education, women's life path (life stimulation and difficult experiences) additionally affected ego level. For example, with respect to life stimulation, women who had remained in traditional family roles or who had never made a commitment to family or career roles were the lowest in ego level, whereas women who had careers had the highest levels. Women high in ego level also were more likely to tell difficult stories about issues of abandonment or searching for an independent identity.

3. *Ego Levels, Values, and Beliefs*

Loevinger's theory maintains that ego levels assess a particular conceptual style by which individuals integrate self-relevant experience. Research has provided ample evidence that this notion is valid. Individuals of higher ego level have been found to be more open-minded than lower ego level individuals, and to hold less conventional beliefs and values (McCrae and Costa 1980). They are also more likely to question fundamental religious beliefs and values, and to be open to different religions. Overall, they are more tolerant, less authoritarian, and less compliant. They are also more likely to engage in complex thinking about new events and situations (Helson and Roberts 1994), and are less traditional in their endorsement of gender roles (McAdams 1985). They are more likely to display patterns of unusual creativity (Helson and Wink 1987).

Individuals of higher ego levels are also more differentiated from others than are those of lower ego levels. For example, high ego level women were found to have more appreciation of others as autonomous and independent beings (Helson and Wink 1987). In a related vein, research by Labouvie-Vief and coworkers (e.g., Labouvie-Vief et al. 1995) also reported

that at lower ego levels, individuals describe their parents primarily in terms of the immediate emotional bond and the global quality of the relationship (such as 'My mom is the greatest mom in the world, I love her!') or in terms of their parental roles (e.g., 'he is a good provider'). In contrast, individuals of higher levels of ego development are better able to view parents as persons separate from the immediate parent–child bond.

Ego levels also imply that individuals differ in the way they modulate their impulses. This is shown in research (Labouvie-Vief et al. 1989) in which individuals described different emotional states: happy, sad, anger, and fear. At lower ego levels, individuals described their emotions statically in terms of concrete actions, bodily symptoms and their concerns to modulate them along conventional and role-defined ideals. However, more mature individuals explored the tension between conventional standards of emotion regulation and what they feel is a deeper self which was defined in terms of contrasts and transformations over time and across contexts.

4. Ego Level and Adjustment

Loevinger claimed that ego level and adjustment are completely independent dimensions. Indeed, when adjustment is measured in terms of subjective well-being (Helson and Wink 1987, McCrae and Costa 1980, Vaillant and McCullough 1987), ego level appears to be independent of adjustment, at least where adult samples are concerned (Noam 1998). For a wide range of adults, ego level does not appear to predict such indices as life satisfaction, emotional adjustment, positive affect, or well being. Why should ego level not be predictive of adjustment? Helson and Wink (1987) suggested that two forms of maturity characterized adults. The first, 'ego level,' is primarily related to growth in cognitive complexity, tolerance of ambiguity, objectivity, and intrapsychic differentiation; this form is related to measures of individuality and creative expressiveness, including such variables as recognition for work, interest in self-understanding, and philosophical and religious interests. The second indicator, 'competence,' was defined as the ability to align with social norms and to regulate behavior by reducing friction and obtaining rewards in social life. This indicator was related to such variables as emotional security and lack of defensiveness and it was negatively correlated with use of drugs and medication and positively with confidence and lively engagement. Similarly, Vaillant and McCullough (1987) found that ego level appeared to be part of a cluster completely distinct from a psychosocial maturity cluster.

Labouvie-Vief (1999) explains such evidence in her theory of self and emotion. She proposes that ego level and adjustment result from two independent core emotion regulation strategies. One of those, affect optimization, is a strategy aimed at maximizing positive emotions and damping negative ones; the second is a strategy of affect amplification, or affect complexity, in which individuals are open to negative affect. This strategy is strongly related to ego level and conceptual complexity. Individuals adopting this strategy process and explore information with the general aim of achieving a more balanced, objective view of self and reality.

5. Conclusion

Ego level represents an important approach to understanding individual differences and development in adulthood. Ego levels are meaningfully related to how individuals organize their experience in terms of their life stories, their values, and their emotions. They emphasize the ability to be open, to engage in complex thinking, to be tolerant and oriented towards objectivity. Although ego levels mature with development in early life, in adulthood their further development appears to be dependent on the context; indeed, constraints of institutions may even hamper further ego development. In addition to adjustment criteria that primarily emphasize individual well-being and positive affect, ego levels provide a broader view of adjustment. They add to positive affect an emphasis on openness, tolerance, and objectivity as additional criteria for mature adaptation in the life course.

Despite the demonstrated importance of ego levels, many approaches to personality and adaptation have not yet made the dimension of complexity that underlies ego level an integral part of their assessment. To integrate ego level and other cognitively based approaches with general theories and assessment approaches to adaptation will constitute an important domain of future research.

See also: Ego Psychology and Psychoanalysis; Personality and Conceptions of the Self; Psychology of the Real Self: Psychoanalytic Perspectives; Self-knowledge: Philosophical Aspects; Self-regulation in Adulthood

Bibliography

Blasi A 1998 Loevinger's theory of ego development and its relationship to the cognitive-developmental approach. In: Westenberg P M, Blasi A, Cohn L D (eds.) *Personality Development: Theoretical, Empirical, and Clinical Investigations of Loevinger's Conception of Ego Development*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 13–26

Block J 1982 Assimilation, accommodation, and the dynamics of personality development. *Child Development* **53**: 281–95

Helson R, Roberts B W 1994 Ego development and personality change in adulthood. *Journal of Personality and Social Psychology* **66**: 911–20

Helson R, Wink P 1987 Two conceptions of maturity examined in the findings of a longitudinal study. *Journal of Personality and Social Psychology* **53**: 531–41

Hy L X, Loevinger J 1996 *Measuring Ego Development*, 2nd edn. Erlbaum, Mahwah, NJ
Labouvie-Vief G 1999 Emotions in later life. In: Bengtson V (ed.) *Theories of Adult Development and Aging*. Springer, New York
Labouvie-Vief G, DeVoe M, Bulka D 1989 Speaking about feelings: conceptions of emotion across the life span. *Psychology and Aging* **4**: 425–37
Labouvie-Vief G, Diehl M, Chiodo L M, Coyle N 1995 Representations of self and parents across the life span. *Journal of Adult Development* **2**: 207–22
Loevinger J 1976 *Ego Development: Conceptions and Theories*, 1st edn. Jossey-Bass, San Francisco
Loevinger J 1993 Measurement of personality: true or false. *Psychological Inquiry* **4**: 1–16
Loevinger J, Cohn L D, Redmore C D, Bonneville L, Streich D, Sargent M 1985 Ego development in college. *Journal of Personality and Social Psychology* **48**: 947–62
McAdams D P 1985 *Power, Intimacy, and the Life Story: Personological Inquiries into Identity*. Dorsey Press, New York
McCrae R R, Costa P T 1980 Openness to experiences and ego level in Loevinger's sentence completion test: dispositional contributions to developmental models of personality. *Journal of Personality and Social Psychology* **39**: 1179–90
Noam G 1998 Solving the ego development–mental health riddle. In: Westenberg P M, Blasi A, Cohn L D (eds.) *Personality Development: Theoretical, Empirical, and Clinical Investigations of Loevinger's Conception of Ego Development*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 271–95
Vaillant G E, McCullough L 1987 The Washington University sentence completion test compared with other measures of adult ego development. *American Journal of Psychiatry* **144**: 1189–94

G. Labouvie-Vief

Ego Psychology and Psychoanalysis

Ego psychology represents a significant historical as well as current perspective within psychoanalysis. Most broadly, ego psychology extends classical psychoanalytic drive theory, combining biological and psychological views of individual development with frameworks referring to complex sociocultural dimensions. In this way, ego psychology moves psychoanalytic theory and practice from study of unconscious events and psychopathology to explorations of adaptive processes and conscious experience within a matrix of interpersonal, familial, and broader sociocultural forces (see, for example, Erikson 1950, Hartmann 1936, Hauser and Safyer 1995). In contemporary theoretical and clinical discussions, the construct of the ego refers to cognitive processes (e.g., self-reflectiveness); perceptual processes (e.g., self-observing); defenses (e.g., suppression); and executive functions (e.g., planning). Moreover, these components can operate at several levels of consciousness—rather than conscious *or* unconscious—within the same individual (Vaillant 1998).

1. Early Formulations

Freud's definition of the ego evolved through three phases (Rappaport 1959). The ego was initially defined with the least precision, the term being used interchangeably with one's own person or self. Despite this ambiguity, one aspect of ego processes was already being highlighted—the defensive functions of the ego. In this first definition then, the ego was conceptualized as focused on preventing painful memories from entering awareness. In a second definition, Freud distinguished central aspects of ego functioning: secondary process (e.g., conscious ideas), reality principle, and repression. The third definition began with Freud's introduction of the structural model, constructing the ego as a coherent system, which—together with the id (drives, nonrational ideas) and superego (conscience) systems—is a component of the tripartite personality organization model. In this most comprehensive formulation, ego processes refer to many connections among perceptions, ideals, moral principles, and conflicts generated within and between id, superego, and ego structures. In this third model of the ego, the groundwork was laid for our current view of the ego representing the individual as an active agent with his or her own independent interests, rather than passively responding to other psychological forces (e.g., conflicts, moral precepts, responses to others) (Rappaport 1959).

Subsequent theoretical, clinical, and empirical contributions have led to clarification and extension of a number of constructs first suggested in these earlier models of ego functioning. Their additions to ego psychology have led to an increasingly refined appreciation of significant relations between ego psychology perspectives and other domains of knowledge, ranging from the social sciences to clinical theory and practice.

2. Contemporary Perspectives on Ego Psychology

Between the late 1930s and the early 1980s many important new additions to ego psychology appeared in psychoanalytic literature. Freud was among the contributors. For although he continued to focus on drives and unconscious mental processes, together with changes in clinical technique, he clearly remained invested in developing ego theory during his later years (Schafer 1995, p. 224, Freud 1933, 1939, 1940). Psychoanalyst critics of ego psychology inaccurately characterize the perspective as minimizing the significance of drives, the pleasure principle, unconscious fantasy, unconscious conflict, danger situations, or transference. 'Instead, one aims to develop further the context for each of these dynamic concepts. The [ego psychology] theory is designed to encompass the total mind and not just its damaged aspects and its primitive, id, or primary-process aspects' (Schafer

1995, p. 224). A major assumption within ego psychology is that through theorizing about mature, rational, and stable modes of function it becomes possible to more precisely conceptualize the 'non-primitive' and the so-called 'normal.'

The last decades of the twentieth century witnessed further extensions of the fine-grained formulations of ego functions begun by Hartmann and colleagues (Hartmann 1936, Hartmann et al. 1964). For example, understanding development and change in such processes as remembering, language learning, communicating, and experiencing insight have been discussed by a wide number of writers with respect to the phenomena themselves and their relation to problems of organization within aspects of the personality, particularly in terms of ego, id, and superego structures. However, in current psychoanalytic theory it has been difficult to trace ego psychology links to current major interests in such broad and important domains as object relations, intersubjectivity, and constructivism. Much of ego psychology deals with the individual as he or she is embedded in the broad biological and social field. But the theory has not consistently conceptualized reciprocal influences between individual development/functioning and the surrounding social fields. Schafer (1995) argues for the relevance of ego psychology for reconciling the theoretical tensions between constructionist and objectivist stances within psychoanalytic thinking; tensions manifest in discussions of psychic reality, communications, and social conditions. 'In psychoanalysis, it is specifically ego psychology that tries to give an account of the development of this complex mixture of constructivism, conventionality, and systematic thought, and what is required to maintain and extend it. Those who reject Freud's and Hartmann's ego psychology … are simply ignoring this area of complete conceptualization and explanation. They claim too much for intersubjectivity alone' (Schafer 1995, p. 230).

During the evolution of ego psychology and the pioneering work of Hartmann and his colleagues (Hartmann et al. 1964), it was already clear that there was not a monolithic theory of psychoanalysis. This pluralistic perspective is more obvious now than ever before. Yet with one exception, we find no new theoretical contributions to the ego psychology literature. The single exception is Erik Erikson.

3. *Extensions of Ego Psychology: Erik Erikson*

Erikson is best known for his writings on ego identity (e.g., Erikson 1959). However, Erikson's impact reaches even further through his creative work at the interface of psychoanalytic theory and the social sciences including history, economics, anthropology, sociology, and all of developmental psychology. Keenly focused on the intertwining of individual development with sociocultural forces, Erikson has been termed by some as the father of psychohistory (Smelser 1998). Consistent with the ego psychology perspective, Erikson considers individual development and functioning within family, community, and cultural contexts. Recognizing biological forces with respect to psychosexual development, Erikson goes further, conceptualizing historical, sociological, political, and economic influences through theoretical analyses and intensive analyses of major figures (for example, Ghandi (Erikson 1969) and Luther (Erikson 1958)). Central in Erikson's work are connections between ego development and society, the intertwining of generations with respect to ego identity development, the 'meshing of individual neurosis and social history' (Smelser 1998), and the necessity for interdisciplinary efforts in understanding human development: 'only psychoanalysis and social science together can eventually chart the life-cycle interwoven throughout with the history of the community' (Erikson 1959, p. 18i; Smelser 1998).

In light of the interdisciplinary nature of Erikson's work considering the individual within his or her social matrix—including intersections of specific life histories and historical eras—it is not surprising that his contributions are not readily incorporated into earlier ego psychology conceptualizations. And for the very same reasons his work has neither been drawn into more mainstream psychoanalytic theories nor into discussions of clinical practice (Wallerstein 1997). On the other hand, social scientists and other scholars applaud and utilize his creative insights in psychoanalysis and neighboring disciplines (Smelser 1998, Cavell 1998, Friedman 1999). Perhaps the most salient explanation for the marginalization of Erikson from the traditional ego psychology literature is the challenge of aligning his complex formulations (drawn from many conceptual frameworks) into the primarily 'individual mind' categories of ego psychology, with precise theoretical definitions.

One further strand of Erikson's work is his conceptualization of the human life-cycle along the lines of developing individual's capacities to master increasingly complex social relationships and maintain continuity with his or her own history and community. Vaillant (e.g., 1993), a major clinical scholar, has called attention to this core theme in Erikson, beginning with his early adolescent (identity) writings and continuing to later considerations of adult development (Erikson 1982).

4. *Ego Psychology in Contemporary Psychoanalytic Practice*

Translating the refined theoretical distinctions of ego psychology into the practice of psychoanalysis is a daunting challenge. Since the early 1980s a group of writers have risen to this important challenge, by

delineating an ego psychology approach to the psychoanalytic technique (Gray 1994, Busch 1997, Kris 1982). Examining implications of ego psychology concepts (e.g., defenses, executive ego functions) and clinical interactions, these contributions attend to the content and style of patient communications, encouraging 'close monitoring,' the clinician's listening to the patient's associations as 'constructions of the ego filtered through a lens measuring the degree of risk' (Busch 1997, p. 407). While not minimizing the importance of unconscious fantasies, drives, and motivations, this clinical approach informed by ego psychology represents 'an attempt to read the text as written by the patient, rather than a reading of the meaning hidden behind the text' (Busch 1997, p. 410).

The idea of risk leads to consideration of defenses, long considered a central construct within ego psychology, first most comprehensively conceptualized with respect to the many forms of ego defenses by Anna Freud (1936). Recent writing on defense processes—including diverse mechanisms as anticipation, suppression, displacement, and denial—draws attention to conscious and unconscious aspects of these functions and their vital importance in adaptation in contrast to psychopathology and dysfunction (Vaillant 1993). Gray's (1986) view of clinical interpretation focuses on the patient's current abilities to grasp new self knowledge, allowing novel constructions into their consciousness: 'The interpretive task is to estimate sensitively the patient's ability to comprehend in order to make a formulation that is not too superficial, yet does not stimulate more reactive defenses' (Gray 1986, p. 253). In other words, this attentiveness to the patient's defenses and other adaptive capacities must be maintained if the psychoanalyst is to involve 'those healthy aspects of the patient's ego that seek to fill out understanding and memory, while searching for pleasure in the form of more effective ego functioning' (Busch 1997, p. 421, Klein 1976). It is of interest that ego defenses, of all hypothesized ego processes, have been the focus of the most sustained empirical efforts at investigating ego psychology constructs.

5. *Empirical Research: Ego Processes and Ego Development*

In addition to the gap from ego psychology to clinical practice, a second important gap is from ego psychology to systematic empirical research. Operationalizing ego psychology constructs and propositions, and designing systematic studies, has posed daunting challenges for investigators. Nonetheless, several programs of research have focused on specific questions and hypotheses regarding ego functions and ego development. Arguably, the ego psychology domain is too vast to be included in any single program of empirical studies. Selected constructs relevant to ego psychology—in particular defenses and ego development—are targets of previous and continuing successful research efforts.

5.1 *Ego Defenses*

The study of ego defense mechanisms has been problematic to clinicians as well as to empirical researchers. 'Investigators who have attempted systematic research on defense mechanisms have often found themselves without a common language for describing them and without replicable methods for assessing them' (Vaillant 1998, p. 1147, Hauser 1986). The multifold problems of classification, conceptual definition, operationalization, and adequate assessment design have plagued dedicated clinical researchers for many years. Since 1971, Vaillant has persistently provided conceptualizations, assessment strategies, and enlightening narrative and quantitative findings regarding the diverse array of defense processes (see, for example, Vaillant 1994). Based on a theoretically-derived hierarchy of defense processes, Vaillant constructed operational definitions and scoring approaches for examining interview material gathered through clinical interviews. His approach has been applied in numerous studies of college men (Vaillant 1997), and inner city men (Vaillant 2000). In addition to dealing with questions about conceptualization and assessment of defenses, much of the interest in Vaillant and colleagues' work has to do with the 'transmutation of less adaptive defenses into more adaptive defenses' (Vaillant 2000, p. 98). Ego development, in particular as manifest through transformation of defenses, is a major theme throughout Vaillant's many contributions (Vaillant 1993, 2000).

5.2 *Ego Development*

Ego defense processes and ego development are intertwined in numerous ways, including the observation that maturation of defenses reflects one important aspect of ego development. The many complex facets of relations between defenses and ego development have not yet been fully articulated (Hauser 1986, 1993). One illustration of the interplay between defenses and ego development is illustrated through two extant models of ego development. One view considers the ego as representing the collection of ego mechanisms, while the other model emphasizes the centrality of a single ego process, comprising integrative activities (e.g., of thoughts and feelings, of present and past ideas) (Hauser 1979).

Bellak et al. (1973) assume the 'collection' model. From this perspective, ego development refers to an unfolding of multiple functions, including adaptive strengths, cognitive processes, defenses, and perceptions of others (Hauser and Daffner 1980). The

second model of ego development focuses on integrative or synthetic processes and the individual's overall frame of reference. Elaborated by Loevinger and colleagues, this view is linked with a widely-used sentence completion assessment technique (Loevinger 1976, Hy and Loevinger 1996). For over 25 years empirical studies have incorporated this theoretical approach and psychometrically rigorous assessment instrument to study ego development in children, adolescents, and adults. Moreover, since the late 1970s these studies have encompassed psychopathology variables together with cross-cultural dimensions (see, for example, Cohn 1991, Westenberg et al. 1998, Hauser 1976, 1993, Hauser et al. 1991, Loevinger 1979). Growing knowledge of connections among individual behavior, individual experience, and ego development has added invaluably to diagnostic and longitudinal studies of psychopathology and development arrests in varied contexts (Hauser 1999, Hauser and Safyer 1995, Hauser et al. 1991). An ongoing longitudinal investigation of adolescent and adult development over the past 20 years focuses on ego development, attachment, and family contexts (Hauser et al. 1991, 1998).

In Loevinger's model of ego development, it is assumed that each individual holds a customary orientation to the self and world, and that there is a continuum (ego development) along which these frames of reference can be arrayed. 'In general, ego development is marked by a more differentiated perception of one's self, of the social world, and the relations of one's feelings and thoughts to those of others' (Candee 1974, p. 621). Thus, in contrast to Bellak's model, this assessment of ego development does not distinguish specific ego processes (e.g., defense mechanisms) from one another, since they are considered aspects of one integrated structure of the developing ego (Loevinger 1979b). Interesting empirical questions then arise—for instance, how are defense mechanisms and other ego processes are linked with ego development (e.g., Hauser et al. 1998).

6. *Future Directions*

In terms of ego psychology theory, there are suggestions of new integrations of this theory with contemporary psychoanalytic studies of focuses on intersubjectivity, object relations development, and dialogue. Related to this is the theoretical challenge to find connections of the highly creative work of Erik Erikson to other psychoanalytic theory and practice. A second direction in the recent effort to apply ego psychology observations and theory to clinical practice with respect to the nature of clinical observations as well as techniques relevant to facilitating the expansion of insight and development. Arguably, these advances in moving from ego psychology at a theoretical level to ego psychology within clinical theory and practice are important ones that will likely continue in the next decades. Finally, several research programs are inspired by ego psychology constructs. Longitudinal studies of ego development now include psychopathology and normal development. Systematic studies of ego defenses and related processes are embedded within these longitudinal investigations. Additional operational definitions of ego processes will likely be developed in the future. It is vital and completely consistent with the work of such ego psychologists as Heinz Hartmann and Erik Erikson, that the perspectives and constructs of ego psychology become even more explicitly integrated with the insights, theories, and empirical approaches of neighboring disciplines (Holzman and Aronson 1992). These disciplines include those investigating aspects of stress, social context dimensions (e.g., family, school, neighborhood), coping, and individual development.

See also: Defense Mechanisms; Ego Development in Adulthood; Erikson, Erik Homburger (1902–94); Freud, Sigmund (1856–1939); Identity in Childhood and Adolescence; Person and Self: Philosophical Aspects; Personal Identity: Philosophical Aspects; Personality and Conceptions of the Self; Psychoanalysis, History of; Psychoanalysis in Sociology; Psychology of the Real Self: Psychoanalytic Perspectives; Self: History of the Concept; Self: Philosophical Aspects

Bibliography

Bellak L, Hurvich M, Gediman H K 1973 *Ego Functions in Schizophrenics, Neurotics, and Normals: A Systematic Study of the Conceptual, Diagnostic, and Therapeutic Aspect*. Wiley, New York

Busch F 1993 'In the neighborhood': Aspects of a good interpretation and its relationship to a 'developmental lag' in ego psychology. *Journal of the American Psychoanalytic Association* **41**: 151–78

Busch F 1997 Understanding the patient's use of the method of free association: An ego psychological approach. *Journal of the American Psychoanalytic Association* **45**: 407–23

Candee D 1994 Ego developmental aspects of new left ideology. *Journal of Personality and Social Psychology* **30**: 620–30

Cavell M 1998 Erik Erikson and the temporal mind. In: Wallerstein R S, Goldberger L (eds.) 1998 *Ideas and Identities: The Life and Work of Erik Erikson*. International Universities Press, Madison, CT, pp. 33–47

Cohn L 1991 Sex differences in the course of personality development: A meta-analysis. *Psychological Bulletin* **109**: 252–66

Erikson E H 1958 *Young Man Luther: A Study in Psychoanalysis and History*. Norton, New York

Erikson E H 1959 Identity and the life-cycle: Selected papers. In: *Psychological Issues* (monograph I). International Universities Press, New York

Erikson E H 1969 *Gandhi's Truth: On the Origin of Militant Nonviolence*. Norton, New York
Erikson E H 1982 *The Life-cycle Completed: A Review*. Norton, New York
Freud A 1936 *The Ego and the Mechanisms of Defense*. International Universities Press, New York
Freud S 1933 New introductory lectures on psychoanalysis. In: *The Standard Edition of the Complete Psychoanalytical Works of Sigmund Freud*. Hogarth Press, London, Vol. 22
Freud S 1939 Analysis terminable and interminable. In *The Standard Edition of the Complete Psychoanalytical Works of Sigmund Freud*. Hogarth Press, London, Vol. 23
Freud S 1940 An outline of psychoanalysis. In: *The Standard Edition of the Complete Psychoanalytical Works of Sigmund Freud*. Hogarth Press, London, Vol. 23
Friedman L J 1999 *Identity's Architect: A Biography of Erik H Erikson*. Scribner, New York
Gray P 1986 On helping analysands observe intrapsychic activity. In: Richards A D, Willick S P (eds.) *Psychoanalysis, The Science of Mental Conflict: Essays in Honor of Charles Vrenner*. Analytic Press, Hillsdale, NJ, pp. 245–62
Gray P 1994 *The Ego and Analysis of Defense*. Aronson, Northvale, NJ
Hartmann H 1936 *Ego Psychology and the Problem of Adaptation*. International Universities Press, New York
Hartmann H, Kris E, Lowenstein R 1964 *Papers on Psychoanalysis*. International Universities Press, New York
Hauser S T 1976 Loevinger's model and measure of ego development: A critical review. *Psychological Bulletin* **83**: 928–55
Hauser S T 1993 Loevinger's model and measure of ego development: A critical review II. *Psychological Inquiry* **4**: 23–9
Hauser S T 1999 Understanding resilient outcomes: Adolescent lives across time and generations. *Journal of Research on Adolescence* **9**(1): 1–24
Hauser S T, Daffner K 1980 Ego functions and development: Empirical research and clinical relevance. *McLean Hospital Journal* **5**: 87–109
Hauser S T, Powers S, Noam G 1991 *Adolescents and Their Families: Paths of Ego Development*. Free Press, New York
Klein G 1976 *Psychoanalytic Theory: An Exploration of Essentials*. International Universities Press, New York
Kris A 1982 *Free Association: Method and Process*. Yale University Press, New Haven, CT
Rappaport D 1959 An historical survey of psychoanalytical ego psychology. *Introduction to Psychological Issues* **I**: 5–17
Schafer R 1995 In the wake of Heinz Hartmann. *International Journal of Psychoanalysis* **76**: 223–35
Smelser N J 1998 Erik Erikson as social scientist. In: Wallerstein R S, Goldberger L (eds.) *Ideas and Identities: The Life and Work of Erik H Erikson*. International Universities Press, Madison, CT, pp. 49–66
Vaillant G E 1993 *The Wisdom of the Ego*. Harvard University Press, Cambridge, MA
Vaillant G E 1994 *The Ego and the Analysis of Defense*. Aronson, Northvale, NJ
Wallerstein R S 1997 Erikson's concept of ego identity reconsidered. *Journal of the American Psychoanalytic Association* **46**: 229–47
Wallerstein R S, Goldberger L (eds.) 1998 *Ideas and Identities: The Life and Work of Erik H Erikson*. International Universities Press, Madison, CT
Westenberg P M, Blasi A, Cohn L D 1998 *Personality Development: Theoretical, Empirical, and Clinical Investigations of Loevinger's Conception of Ego Development*. Erlbaum, Mahwah, NJ

S. Hauser

Elaboration in Memory

1. Introduction

Elaboration is a conscious, intentional process that associates to-be-remembered information with other information in memory. Examples of elaborations include retrieving a semantic associate of a to-be-remembered item (e.g., 'cat' for 'dog'), retrieving the taxonomic category of a to-be-remembered item (e.g., 'fruit' for 'pear'), and identifying the relationship between two to-be-remembered items (e.g., recognizing that the to-be-remembered items 'sour' and 'grapes' may constitute an idiom). The construct of elaboration has been used in cognitive psychology to explain why a variety of instructional and materials manipulations produce superior memory performance. In addition, differences in the spontaneity or efficacy of using elaborations have been proposed as explanations of differences in memory performance across populations (e.g., young subjects vs. elderly subjects). At a general level, the construct of elaboration is consistent with cognitive psychology's tripartite emphasis on intentional processing, mental contents as mediating variables, and information transformation. The succeeding sections describe the intellectual antecedents of the construct of elaboration in modern cognitive psychology, the specific genesis of the contemporary use of elaboration in the levels-of-processing literature, the theoretical mechanisms that have been proposed to explain the beneficial effect of elaboration on memory performance, and three important empirical applications of the construct. These applications involve the explanation of population differences in memory performance, differences in the memorability of various types of materials, and differences in performance on various memory tests.

2. Historical Antecedents of the Construct of Elaboration

The modern concept of elaboration emerged from a series of information-processing models emphasizing the roles of filtering and rehearsal of information as critical determinants of memory performance (Broadbent 1958, Atkinson and Shiffrin 1968, Baddeley and Hitch 1974). Filter models of attention (Broadbent 1958) emphasize that attention plays an important role in sensory processing by giving priority to information arriving through certain sensory moda-

lities or channels while filtering information arriving through other modalities or channels. These models introduced the ideas that earlier stages of information processing were required for later processing and that only a limited subset of information received this initial processing. These notions of filtering were adopted in stage models of memory (Atkinson and Shiffrin 1968). Stage models assume three distinct types of memory stores (a sensory store, a short-term memory store, and a long-term memory store) and that information flows sequentially from one store to the next. Information entering the sensory store was assumed to be filtered by attentional processes so that only a limited set of this information entered the short-term memory store. Upon entrance to the short-term store, an intentional rehearsal process, generally referred to as maintenance rehearsal, was assumed to maintain information, mediating its transfer to the long-term store. Transfer to the long-term memory store was assumed to be a necessary condition for later memory performance. Thus, stage models emphasize that conscious, intentional processing is a critical influence on memory (see Baddeley and Hitch (1974) for more complex descriptions of intentional processes that might operate in short-term or working memory).

The emphasis of stage models on maintenance rehearsal led to a series of investigations of the properties of maintenance rehearsal. In the current context, the most critical studies were those which demonstrated that, contrary to the assumptions of the stage model, maintenance rehearsal did not necessarily enhance memory (Craik and Watkins 1973, Woodward et al. 1973). For example, Craik and Watkins (1973) demonstrated that increasing the duration of maintenance rehearsal by manipulating characteristics of a secondary task did not enhance memory. This counterintuitive finding led to a distinction between type 1 (maintenance) and type 2 (elaborative) rehearsal, with only the latter type of rehearsal assumed to affect retrieval from long-term memory. Thus, by the early 1970s, filter models of attention and stage models of memory had emphasized the role of conscious, intentional rehearsal processes in memory and ensuing empirical investigations had demonstrated that maintenance rehearsal did not necessarily enhance memory.

3. *Specific Genesis of the Modern Construct of Elaboration in the Levels-of-processing Literature*

Filter models of attention and stage models of memory focused attention on the process of memory storage. In the early 1970s, the levels-of-processing approach presented an alternative characterization of memory storage (Craik and Lockhart 1972). In contrast to stage models, the levels-of-processing approach assumed that to-be-remembered information was processed in three sequential stages of *perceptual analysis*. Presented materials were assumed to be subjected to an orthographic analysis, a phonological analysis, and then a semantic analysis. As in stage models, the character and nature of this processing were assumed to influence memory. Specifically, it was assumed that earlier stages of analysis (i.e., orthographic and perceptual analyses) led to transient memory traces, whereas semantic analyses led to longer-lasting memory traces. Thus, the levels-of-processing framework predicts that semantic processing should lead to better memory than phonological or orthographic processing. Consistent with this prediction, numerous empirical studies demonstrated that semantic orienting questions (e.g., 'is a pear a fruit?') during study lead to better memory than phonological or orthographic orienting questions (e.g., 'does pear rhyme with fair?') during study. Despite this support for the levels-of-processing approach, the congruity effect (Craik and Tulving 1975) presented a major challenge to this approach. The congruity effect is the finding that memory is better when responses to orienting questions are positive (i.e., 'yes') rather than negative (i.e., 'no'), independent of the level of processing of the orienting question. For example, the word 'pear' is better remembered if the orienting question is 'Is a pear a fruit?' than if the orienting question is 'Is a pear a weapon?' The congruity effect presents a problem for the levels-of-processing approach because it demonstrates substantial variability in memory performance with a fixed level of processing.

To explain the congruity effect, Craik and Tulving (1975) argued that elaboration—the association of to-be-remembered information with other information—was critical to memory performance. Thus, the congruity effect was assumed to occur because 'positive' orienting questions created episodic associations between the to-be-remembered item and information in the orienting questions (e.g., 'pear' and 'fruit') whereas 'negative' orienting questions did not form such associations (e.g., 'pear' and 'weapon'). Further, Craik and Tulving (1975) proposed that elaboration was a fundamental determinant of memory performance, hypothesizing that it was the basis of levels-of-processing effects. Specifically, they argued that the elaborations produced by semantic orienting questions were more likely to enhance memory than the elaborations produced by phonological or orthographic orienting questions.

4. *Theoretical Mechanisms of the Effect of Elaboration on Memory Retrieval*

Two classical approaches to memory retrieval provide explanations of elaboration's effects on memory. These mechanisms are compatible and one can conceptualize effects of elaboration as arising from their joint influence. The first mechanism arises from associationist models of memory (e.g., Raaijmakers

and Shiffrin 1981). In these models, to-be-remembered information is associated with episodic or list representations and to other information presented or retrieved during a study episode. During a memory test, episodic and other representations can be retrieved and used as cues for to-be-remembered information. In this perspective, elaboration enhances memory because it creates additional cues for to-be-remembered information. A simple example demonstrates the benefit of additional cues. The presentation of the word 'dog' in the study list creates an association between this item and the representation of the study list. If a subject then elaborates the presentation of the word 'dog' by recognizing that 'dog' is associated with 'cat,' this elaboration creates a second cue (i.e., 'cat') for 'dog.' In addition to providing an account of the effects of elaboration, a focus on multiple cues also provides an account of other memory effects such as those of encoding variability and imagery. Encoding variability effects occur when information studied from two perspectives is remembered better than information studied twice from a single perspective (Reder et al. 1974). For example, studying the word 'bank' in relation to its meanings as 'the side of a river' and 'a financial institution' produces better memory than studying 'bank' twice in the context of either of the individual meanings. Effects of imagery occur when words that have concrete referents (e.g., 'dog') are remembered better than words with abstract referents (e.g., 'justice') (Paivio 1971). The effects of encoding variability and imagery can be ascribed to the availability of additional retrieval cues. In the case of encoding variability, the second meaning of a word provides an additional cue, while in the case of concrete words, a visual/imaginal representation provides an additional cue.

The second mechanism focuses on the idea that elaboration creates distinctive associations (Craik and Jacoby 1979). This approach assumes that cues are more effective when they have unique associations to to-be-remembered information. Thus, elaboration can enhance retrieval by creating distinctive or unique cues. For example, if 'xylophone' is part of a to-be-remembered list, the elaboration 'a musical instrument beginning with x' should enhance performance more than the elaboration 'a word containing many vowels.'

This focus on distinctiveness has been used to explain the standard levels-of-processing effect (i.e., semantic > phonological > orthographic) by assuming that semantic cues are more distinctive than phonological or orthographic cues. Further, investigators working in this tradition have demonstrated that nonsemantic information can be well remembered if it is distinctive. For example, Hunt and Elliott (1980) have demonstrated that orthographic distinctiveness can enhance recall and recognition.

The preceding account of the effects of elaboration emphasizes the cue-dependent nature of memory. The availability and distinctiveness of cues are critical determinants of memory performance. This focus on cue dependence helps explain how powerful mnemonic devices such as hierarchical organization (Bower 1970) produce exceptional memory performance. In hierarchical organization, each cue is associated with a relatively small number of to-be-remembered items which, in turn, serve as cues for other items. Thus, the initial retrieval of a single cue at the top of a hierarchical organization can permit access to a large amount of information. Consistent with this view, hierarchical organization has been implicated in the exceptional performance of many mnemonists (e.g., Chase and Simon 1973).

5. *Applications of the Elaboration Construct*

5.1 *Interactions with Populations*

Difficulties in using elaborations have been proposed as explanations of differences in memory performance across populations. Most notably, difficulties in using elaborations have been proposed to explain why both young children and older adults have poor memories. Consistent with this account, studies of young children have demonstrated that young children are much less likely to create spontaneously elaborations than older children. For example, Barnes et al. (1996) demonstrated that six-year olds were less likely to construct elaborations about a text than 15-year olds, even when the two groups of subjects were equated on prior knowledge of the text. Similar arguments have been applied to explain the memory deficits of older adults. In addition, it has been proposed that the elaborations produced by older adults are more idiosyncratic than those produced by younger adults. These idiosyncratic elaborations are unlikely to be effective cues when they are retrieved during a later test. Consistent with this view, Zabrucky and Moore (1995) demonstrated that elaborations of text generated by older adults were less likely to be related to the text (i.e., more idiosyncratic) than the elaborations generated by younger adults. Further, the number of idiosyncratic elaborations generated was negatively correlated with recall and accounted for much of the variance in recall performance across older and younger subjects.

Additional evidence that deficits in elaboration produce the memory problems demonstrated by young children and older adults is found in studies manipulating elaborative processing. These studies demonstrate that manipulations designed to increase elaboration minimize the memory deficits demonstrated by younger children and older adults. For example, Cherry et al. (1993) demonstrated that asking older adults to create elaborations that focused on explaining a text minimized age differences in recall, even though such differences occurred in a control

condition in which participants were not directed to construct elaborations.

The elaboration construct has also been applied to the explanation of memory difficulties in a wide variety of patient populations (i.e., learning disabled, brain injured, and mentally retarded persons). As with younger children and older adults, it is assumed that these patients do not spontaneously or effectively use elaborations. In this context, numerous studies have demonstrated that teaching elaboration techniques to these patients enhances their learning and memory. For example, Scruggs et al. (1994) demonstrated that fourth and fifth graders with mild learning disabilities demonstrated a substantial benefit in text recall when encouraged to construct elaborations.

5.2 Interactions with Materials and Prior Knowledge

A large number of studies have investigated the interactions of elaborative processing with the nature of the to-be-learned materials and the participant's prior knowledge. These investigations help specify the conditions under which elaborations enhance memory. The semantic relatedness of to-be-learned materials is one of the most important determinants of memory with strongly related materials generally being remembered better than weakly related materials. Further, the effects of elaboration interact with those of semantic relatedness, so that elaborative processing produces a larger benefit for weakly related than strongly related materials, often equating overall memory performance on these materials. For example, Horiba (1996) examined recall of pairs of sentences that were either strongly or weakly related when participants were given elaboration or memorization instructions. Horiba found that strongly related sentences were remembered better than weakly related sentences under memorization instructions, but that elaboration produced a greater benefit for the weakly related sentences, equating performance on strongly and weakly related sentences. Explanations of such finding emphasize that elaborations occur spontaneously for strongly related materials, but not for weakly related materials. Consequently, elaborative instructions produce a greater benefit on weakly related materials.

This focus on the beneficial effects of elaborative processing for weakly related materials can be used to explain circumstances in which weakly related materials produce better memory than strongly related materials. For example, Erdfelder and Bredenkamp (1998) hypothesized that atypical events presented in text are better remembered than typical events because subjects must extensively elaborate the atypical events to produce a coherent mental representation of the text.

One assumption underlying the preceding account of the beneficial effects of elaboration on weakly related materials is that participants can retrieve the semantic information necessary to create the appropriate elaborations. This assumption suggests that participants' degree of prior knowledge should interact with the effects of elaborative processing. Consistent with this suggestion, numerous studies have demonstrated that elaborations produce greater memory benefits when participants have more extensive prior knowledge. Willoughby et al. (1993) demonstrated this finding by examining memory for facts learned about well-known and unfamiliar animals. They compared memory when subjects were asked to explain why a fact about an animal was true to memory in a repetition control condition. The elaboration condition produced greater memory relative to the control condition for familiar animals, but not for unfamiliar animals. The absence of an elaboration effect for unfamiliar animals is consistent with the view that prior knowledge is critical to the beneficial effects of elaborative processing.

These interactions of elaborative processing with the structure of materials and participants' prior knowledge play an important role in memory for, and comprehension of, text (Kintsch and Van Dijk 1978). While an extensive discussion of the text comprehension literature is beyond the scope of this article, there are two important parallels between the literature on text comprehension and memory that should be noted. First, consistent with the views derived from standard memory paradigms, elaborations of a text are assumed to have important effects on the memory representation of the text. Second, the issue of the materials and knowledge states that permit elaborations has been a topic of extensive study. One example that represents some of the issues arising in such investigations is Lea's research (Lea 1995) on inference generation. Lea demonstrated that subjects will use propositional logic (e.g., given p or q and not p, they will infer q) to generate elaborations about a text, even when such elaborations are not necessary to maintain a coherent text representation.

5.3 Interactions with Memory Tests

As discussed above, numerous studies have demonstrated that elaborating materials during study enhances performance in episodic memory tests. A large number of recent studies have examined the effects of elaboration on implicit memory tests. Implicit memory (Schacter 1987) demonstrates an influence of a prior episode that is independent of the participant's conscious awareness of that episode. Implicit memory tests generally attempt to show that performance on a task is enhanced by a prior episode even though participants are not necessarily consciously aware of the effect of the prior episode on their performance. Examples of such tests include perceptual identification and fragment completion. In perceptual identification, words are presented very briefly (e.g., for

33 ms) and participants are asked to identify them. Some of the items presented in the perceptual identification test are items that were presented in an earlier study episode and the test attempts to see if identification is better for these old items than for new items. Differences in perceptual identification between old and new items are generally referred to as 'priming.' The fragment completion test operates analogously to the perceptual identification test, except that, instead of rapid presentation, letters are deleted from the to-be-identified item.

Graf and Mandler (1984) hypothesized that elaborative processes affect explicit memory tests such as recognition and recall, but not implicit memory tests such as perceptual identification and fragment completion. Their rationale was that implicit memory tests are primarily sensitive to integrative processing of perceptual information. Assuming that elaborative processing occurs following item perception, one would not expect variations in elaborative processing to influence the integration of perceptual information.

The literature investigating these hypotheses has generated a complex pattern of results, with some manipulations of elaborative processing (e.g., levels-of-processing, Jacoby and Dallas 1981) producing no detectable effects on implicit memory and other manipulations of elaborative processing (e.g., generation, Masson and MacLeod 1992) demonstrating robust effects on implicit memory tests. Further, even in the cases where effects were not detectable in individual experiments, meta-analyses have demonstrated effects. For example, Jacoby and Dallas (1981) demonstrated that levels-of-processing produced no effect on perceptual identification. However, Brown and Mitchell (1994) demonstrated effects of levels-of-processing on perceptual identification in a meta-analysis. A final complication in this literature is that results demonstrating effects of elaborative processing on implicit memory tests may arise because explicit memory processes contaminate performance in the implicit memory test.

While the preceding summary demonstrates a literature in ferment, and one that has far from definitive conclusions, two hypotheses may help rationalize this complex pattern of results. First, it may be the case that the small effects demonstrated in meta-analyses arise from contamination by explicit memory processes. If this is the case, results such as those of Jacoby and Dallas (1981) support Graf and Mandler's original hypothesis that elaborative processing does not necessarily influence performance on implicit memory tests. Second, there may be some heterogeneity in the effects of different types of elaborative processing on implicit memory tests.

Thus, while the levels-of-processing manipulation used in the experiment of Jacoby and Dallas (1981) may not affect performance on implicit memory tests, the generation manipulation used by Masson and MacLeod (1992) may do so. Establishing such heterogeneity in types of elaborative processes would be consistent with cognitive psychology's tradition of differentiating types of encoding processes (see introductory sections).

6. *Conclusion*

The modern construct of elaboration arose from cognitive psychology's emphasis on intentional processing and information transformation—an emphasis reflected in filter models of attention and stage models of memory. Its specific genesis was in attempts to explain the congruity effect demonstrated in the levels-of-processing literature. Elaboration of the relation between an orienting question and to-be-recalled information was assumed to enhance performance, producing the congruity effect. Theoretical accounts of effects of elaboration have emphasized that the availability and distinctiveness of retrieval cues are critical to memory performance. The construct of elaboration has been widely applied with differences in the ability to create or use elaborations being proposed as explanations of differences in memory due to the effects of age, materials, participant's knowledge, and type of memory test.

See also: Conscious and Unconscious Processes in Cognition; Memory and Aging, Cognitive Psychology of; Memory Development in Children; Memory for Meaning and Surface Memory; Memory: Genetic Approaches; Memory Problems and Rehabilitation; Memory Retrieval; Unconscious: History of the Concept

Bibliography

Atkinson R C, Shiffrin R M 1968 Human memory: A proposed system and its control processes. In: Spence K W, Spence J T (eds.) *The Psychology of Learning and Motivation: Advances in Theory and Research, 2*. Academic Press, New York

Baddeley A D, Hitch G 1974 Working memory. In: Bower G H (ed.) *The Psychology of Learning and Motivation, 8*. Academic Press, New York

Barnes M A, Dennis M, Haefele-Kalvaitis J 1996 The effects of knowledge availability and knowledge accessibility on coherence and elaborative inferencing in children from six to fifteen years of age. *Journal of Experimental Child Psychology* **61**: 216–41

Bower G H 1970 Organizational factors in memory. *Cognitive Psychology* **1**: 18–46

Broadbent D E 1958 *Perception and Communication*. Pergamon, Oxford, UK

Brown A, Mitchell D 1994 A reevaluation of semantic vs. nonsemantic processing in implicit memory. *Memory and Cognition* **22**: 533–41

Chase W G, Simon H A 1973 The mind's eye in chess. In: Chase W G (ed.) *Visual Information Processing*. Academic Press, New York

Cherry K E, Park D C, Frieske D A, Rowley R L 1993 The effect of verbal elaborations on memory in young and older adults. *Memory and Cognition* **21**: 725–38
Craik F I M, Jacoby L L 1979 Elaboration and distinctiveness in episodic memory. In: Nilsson L (ed.) *Perspectives on Memory Research: Essays in Honor of Uppsala University's 500th Anniversary*. Erlbaum, Hillsdale, NJ
Craik F I M, Lockhart R S 1972 Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* **11**: 671–84
Craik F I M, Tulving E 1975 Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology* **104**: 268–94
Craik F I M, Watkins M 1973 The role of rehearsal in short term memory. *Journal of Verbal Learning and Verbal Behavior* **12**: 599–607
Erdfelder E, Bredenkamp J 1998 Recognition of script-typical versus script-atypical information: Effects of cognitive elaboration. *Memory and Cognition* **26**: 922–38
Graf P, Mandler G 1984 Activation makes words more accessible, but not necessarily more retrievable. *Journal of Verbal Learning and Verbal Behavior* **23**: 553–68
Horiba Y 1996 The role of elaborations in L2 text memory: The effects of encoding task on recall of casually related sentences. *Modern Language Journal* **80**: 151–64
Hunt R R, Elliott J M 1980 The role of nonsemantic information in memory: Orthographic distinctiveness effects upon retention. *Journal of Experimental Psychology: General* **109**: 49–74
Jacoby L L, Dallas M 1981 On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General* **110**: 306–40
Kintsch W, van Dijk T A 1978 Toward a model of text comprehension and production. *Psychological Review* **85**: 363–94
Lea R B 1995 On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning Memory and Cognition* **21**: 1469–82
Masson M E J, MacLeod C M 1992 Reenacting the route to interpretation: Enhanced preceptual identification without prior perception. *Journal of Experimental Psychology: General* **121**: 145–76
Paivio A 1971 *Imagery and Verbal Processes*. Holt, New York
Raaijmakers J G W, Shiffrin R M 1981 Search of associative memory. *Psychological Review* **88**: 93–134
Reder L M, Anderson J R, Bjork R A 1974 A semantic interpretation of encoding specificity. *Journal of Experimental Psychology* **102**: 648–56
Schacter D L 1987 Implicit memory: History and current status. *Journal of Experimental Psychology: Learning Memory and Cognition* **13**: 501–18
Scruggs T E, Mastropieri M A, Sullivan G S 1994 Promoting relational thinking: Elaborative interrogation for students with mild disabilities. *Exceptional Children* **60**: 450–57
Willoughby T, Waller G T, Wood E, MacKinnon G E 1993 The effect of prior knowledge on an immediate and delayed associative learning task following elaborative interrogation. *Contemporary Educational Psychology* **18**: 36–46
Woodward A, Bjork R, Jongeward R 1973 Recall and recognition as a function of primary rehearsal. *Journal of Verbal Learning and Verbal Behavior* **12**: 608–17
Zabrucky K, Moore D 1995 Elaborations in adults' text recall: Relations to working memory and text recall. *Experimental Aging Research* **21**: 143–58

E. Hirshman

Electoral Geography

Electoral geography provides a spatial perspective to the study of the organization, conduct, and outcome of elections, all of which are inherently geographical. It focuses on the nature of, and interactions among, four maps. There are two main input maps—the geography of votes for a party or candidate (itself in part a function of the geography of the social groups from which each draws its main support) is set within a geography of constituencies: their interactions produce the first output map—the geography of representation—that can be converted into a second output map, a geography of power.

This study of spatial aspects of elections is a small subfield within geography. No synthesizing text has appeared since 1979 (Taylor and Johnston 1979), although the subject matter studied has expanded substantially since (Johnston et al. 1998). Its main concerns are with: spatial variations in the support for parties and candidates, and in the campaigning for that support; the processes by which territorially-defined electoral constituencies are defined within different electoral systems; the geographies of representation which result from the pattern of votes across constituencies; and the geography of political power—of 'who gets what, where' because of politically- and electorally-influenced decisions on the location of public goods and services.

1. *The Geography of Votes*

The classic works on the geography of votes include several by André Siegfried, who pioneered *géographie électorale*. He drew cartographic correlations between patterns of party support and the physical, social, and economic landscapes, sophisticatedly linking environment to voting through agrarian systems, land ownership, and rural social stratification and tradition, demonstrating that social factors underpinning party support have environmental foundations. Support for parties in the Ardèche during the Third Republic (1871–1940) reflected variations not only in geology and relief, but also in agricultural activities and the distribution of industry: parties obtained their support in different sociospatial milieux (Siegfried 1949; see also Dogan 1967 and the volumes edited by Lacoste 1986, follow the Siegfried tradition).

Siegfried showed substantial continuity in the Ardèche's geography of votes throughout the Third Republic, as did another seminal work (Key 1955). Consistent geographies of support reflect stable electoral cleavages, with parties sustained within the same milieux over sequences of elections. Those cleavages are broken occasionally, either temporarily, producing a deviating election from the normal sequence, or permanently, whereby one or more elections is

critical in the realignment from one cleavage to another. Changed cleavage structures are reflected in changing geographies of support.

The concept of electoral cleavage was broadened in Lipset and Rokkan's (1967) identification of those characterizing most European countries. Two—between subject and dominant cultures (the core-periphery cleavage) and between churches and secular interest groups (the church vs. state cleavage)—followed the national revolution from the seventeenth century onwards, whereas those between the primary and the secondary economy (the urban–rural or town–country cleavage) and employers and workers (the class cleavage)—were products of the industrial revolution. The class cleavage occurs in nearly all European countries but the others' relative strength reflects local history and culture; understanding a country's geography of votes involves mapping its cleavage structures and exploring how parties mobilise support in different social milieux and places.

Key's theory underpinned studies of American electoral mosaics, notably Archer and Taylor's (1981) detailed analysis of the changing geography of support for American Presidential candidates, which they related to the geography of the country's political economy. Shelley et al. (1996) linked maps of support for presidential candidates and for parties at Congressional elections to the geographies of various interest groups and of policy outcomes (such as support for various policies in the House of Representatives from different parts of the country).

1.1 Local Variations Within Cleavage Structures

The cleavage model suggests that the geography of votes at an election can be 'read-off' from the geography of the cleavage structure(s): if the class cleavage dominates, for example, knowledge of each area's class structure should successfully predict spatial variations in each party's support. This was widely assumed in Great Britain: Butler and Stokes' (1969, 1974) classic study identified considerable uniformity across the country in both voting at individual elections (though with some variations, such as differences in working-class voting between resort towns and mining areas) and interelection shifts in support between parties—the so-called 'uniform swing.'

These assumptions were challenged by geographical work showing that the cleavage and uniform swing models could not account for changing geographies of support. People in the same class position voted differently in different parts of Great Britain (Johnston 1985) and interparty, interelection swings also varied spatially (Johnston 1983). Parties performed even better than the cleavage model suggested where their class base was strong; the geography of support was spatially more polarised than the geography of cleavages. (These findings were contested by some who believe that all spatial variations can be accounted for by the characteristics of individual voters: McAllister and Studlar 1992, Johnston and Pattie 1998).

One major reason for this polarization has been identified in geographical studies. As class dealignment advanced after 1960 and cleavages became less deep-seated in British culture, short-term economic factors increasingly influenced voting decisions. American work on 'economic voting' (Lewis-Beck 1988) shows people influenced by their perceptions of the economic situation: they reward governments which deliver prosperity, and punish those which do not. Such retrospective voting may be either egocentric (people who are prospering individually vote to re-elect governments) or sociotropic (people vote to reward governments that deliver national prosperity, but punish those blamed for economic difficulties). Alternatively, economic voting decisions may be prospective: voters support the party considered most likely to deliver future prosperity (national or personal).

Studies of the 1983 and 1987 British general elections suggested that economic voting accounted for the growing spatial polarization in support for the two main parties—the so-called north–south divide. Similar types of people in different areas perceived a different national economic situation and voted accordingly (Johnston et al. 1988)—the incumbent government performed better in the prospering 'south' than in the depressed 'north.' Pattie and Johnston (1995) introduced a third component to the retrospective voting model: as well as evaluating the government's record on the production of national and personal prosperity, voters also judged its economic performance against the state of their own region's economy.

1.2 Local Influences on Voting Decisions

Most voting decisions result from interactions, many indirect, between political parties and electors, which include local interactions that influence election outcomes. Key's (1949) book on elections where party cues were irrelevant provided a major stimulus; the Democratic party dominated politics in the southern US for nearly a century, and voters' main choices were among contenders for its nomination at primary contests. Many candidates performed best in their home areas, creating a 'friends and neighbors effect.' This was formalized by Reynolds (1969) and extended by Cox (1969): voting decisions are influenced by spatially-biased information flows—what Cox called the acquaintance-circle process—which generate not only support for local candidates but also more general 'neighborhood effects' whereby an area's majority opinion is accentuated if those initially favoring minority views are won over through a process of 'conversion by conversation.'

Results consistent with the neighborhood effect have been produced by studies using ecological data,

but their generating processes can only be inferred (Miller 1977, Eagles 1995a, 1995b)—hence Dunleavy's (1979) critique. The few process studies that have been undertaken (notably Huckfeldt and Sprague 1995) provide strong evidence that 'those who talk together vote together' (however, see Curtice 1995 and Pattie and Johnston 1999).

Study of neighborhood, or local contextual, effects has been extended by geographers interested in the specificity of particular places, especially their local cultures. Cox (1969) called this the forced-field process: a local political culture is established by parties and/or other interest groups (such as trades unions) and forms the basis for mobilizing the local electorate around either particular views on general issues or local issues. The resulting geography of votes may show some areas deviating significantly from wider trends, as in Cox's (1970) work on Wales. The approach was extended by Agnew's (1987) case studies from Scotland, Italy, and the US on mobilization within local milieux. He codified the processes involved through six claims (Agnew 1996):

(a) Class and community affiliations take on different meanings according to the nature of individual places within the spatial division of labor;

(b) Communication systems limit interaction across space in some situations, leading to separate networks within which politically relevant information flows and is interpreted;

(c) There are continuing tensions between local areas and their encompassing national states, which foster spatial variations in political mobilization;

(d) All cleavages—class, gender, race, age, etc.—have local as well as national histories within which the structuration of political movements occurs;

(e) Many political parties claim local roots and policy directions in their electoral manifestos, which can result in spatially-varying support; and

(f) The microgeography of everyday settings—home, work, leisure, etc.—can stimulate local distinctiveness that is reflected in voting patterns.

1.3 Local Campaigns and Election Outcomes

Mobilizing support in communities is particularly active during the campaign period preceding an election. Although national and regional media are used widely, local campaigns are crucial for disseminating information about candidates and canvassing support and turnout. This requires expenditure, and the more that candidates spend in their constituencies the better their performance (Jacobson 1980, Johnston 1987, Pattie et al. 1995); spending is a good surrogate for other aspects of local campaigning, demonstrating that the more intensive a party's campaign in a place, the better its relative performance (Denver and Hands 1997)—including campaigns for tactical votes to unseat an unpopular incumbent (Johnston et al. 1988).

2. The Geography of Constituency Definition

Almost all electoral systems allocate (at least some) legislative seats by territorially-defined constituencies. Definition of constituency boundaries is crucial: victory or defeat can depend upon a constituency's voters' characteristics.

Interest in constituency definition developed during the US 1960s 'redistricting (or reapportionment) revolution' (Baker 1966). Two 'abuses' were practiced widely where parties controlled redistricting: malapportionment, whereby constituencies with widely differing populations were created—parties maximized their returns (the ratio of seats to votes) by having small constituencies where they were strong but large ones in their opponents' heartlands; and gerrymandering, whereby parties delimited constituencies, whatever their size, so that their supporters were in a majority in as many as possible (Morrill 1973, 1981). Malapportionment was outlawed by the Supreme Court in the 1960s and strict equality criteria were set. Gerrymandering was considered much harder to tackle, because of difficulties proving intent to discriminate against those minorities protected under the Fourteenth Amendment, used in most of the relevant cases (Grofman 1990); the *Voting Rights Act 1965* and its successors provided a basis for challenging maps which apparently discriminated against black voters, although the Courts have been inconsistent in their rulings (Grofman 1998). Redistricting is largely nonpartisan in its conduct elsewhere, but parties in the UK influence the details to their own electoral benefit (Rossiter et al. 1999).

Defining constituencies, using smaller areas as building blocks, exemplifies the modifiable area unit problem well-known in spatial statistics. Gudgin and Taylor (1979) developed a sophisticated theory of the process which they linked to the production of electoral bias—both intentional, through explicit malapportionment and gerrymandering strategies, and unintentional, through nonpartisan procedures with outcomes consistent with those strategies: other geographical factors are important bias generators, such as district size effects (Taagepera and Shugart 1989). Bias identification has been extended using Brookes's (1960) index, subdivided into six components, all spatial (Johnston et al. 1999, Rossiter et al. 1999).

3. The Geography of Representation and the Geography of Power

Relatively little has been done on the second pair of maps: electoral geographers have concentrated on the inputs to electoral systems rather than their outputs.

Studies of the geography of representation focus on the translation of votes into seats, especially under the first-past-the-post electoral system. A party's representation is spatially more polarized than its support

among the electorate in many cases, as in the UK during the 1980s (Johnston et al. 1988): such 'geographical divisiveness' was presented as a major rationale for reforming the country's electoral system (Jenkins 1998).

Spatial polarization of representation may be reflected in the geography of public policy, with governing parties favoring areas they draw legislators from. Shelley et al. (1996) illustrated this for the US where the allocation of some public goods involves pork-barreling: individual Representatives and Senators use their positions on Congressional Committees to win benefits for their constituencies (Johnston 1980, Stein and Bickers 1995).

4. Conclusions

Elections are based on geographies. Electoral geographers stress the importance of the creation and interaction of the geography of votes and the geography of constituencies, for which they have developed novel analytical procedures (such as: Archer and Taylor's (1981) adaptation of factor analysis for identifying electoral continuity and critical elections; Rossiter and Johnston's (1981) computer simulation of the UK redistricting procedure; Johnston and Hay's (1982) adaptation of entropy-maximizing methods to estimate detailed voting patterns at various spatial scales; Jones et al's. (1992) multilevel modeling of voting patterns at scales from the individual upwards; and Johnston et al's. (1999) extension of Brooke's algebra for estimating biases in election outcomes). They have shown that the map of votes cannot be 'read off' from a map of cleavages but must take account of other influences on voting decisions, many of which (such as the friends and neighbors effect, the neighborhood effect, the pattern of local culture, and the geography of campaign mobilization) involve processes that are differentiated geographically. These interact with the geography of constituencies to produce electoral outcomes that frequently are biased towards one party, and which in turn produce geographies of representation within which geographies of policy implementation are inscribed.

See also: Electoral Systems; Participation: Political; Party Systems; Political Representation; Polling; Voting: Class; Voting: Compulsory; Voting: Issue; Voting, Sociology of; Voting: Tactical

Bibliography

Agnew J A 1987 *Place and Politics: The Geographical Mediation of State and Society*. Allen & Unwin, Boston

Agnew J A 1996 Mapping politics: how context counts in electoral geography. *Political Geography* **15**: 129–46

Archer J C, Shelley F M 1986 *American Electoral Mosaics*. Association of American Geographers, Washington, DC

Archer J C, Taylor P J 1981 *Section and Party: A Political Geography of American Presidential Elections from Andrew Jackson to Ronald Reagan*. Research Studies Press, Chichester, UK and New York

Baker G E 1966 *The Reapportionment Revolution*. Random House, New York

Brookes R H 1960 The analysis of distorted representation in two-party, single-member systems. *Political Science* **12**: 158–67

Butler D E, Stokes D E 1969 *Political Change in Britain: Forces Shaping Electoral Choice*. Macmillan, London

Butler D E, Stokes D E 1974 *Political Change in Britain: The Evolution of Electoral Choice*, 2nd edn. Macmillan, London

Cox K R 1969 The voting decision in a spatial context. *Progress in Geography* **1**: 81–117

Cox K R 1970 Geography, social contexts and voting behavior in Wales, 1861–1951. In: Allardt E, Rokkan S E (eds.) *Mass Politics*. Free Press, New York

Curtice J 1995 Is talking across the garden fence of political import? In: Eagles M (ed.) *Spatial and Contextual Models in Political Research*. Taylor & Francis, London

Denver D T, Hands G 1997 *Modern Constituency: Electioneering: Local Campaigning the 1992 General Election*. Cass, London

Dogan M 1967 Political cleavage and social stratification in France and Italy. In: Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments*. Free Press, New York

Dunleavy P 1979 The urban basis of political alignment: social class, domestic property ownership and state intervention in consumption processes. *British Journal of Political Science* **9**: 409–43

Eagles M (ed.) 1995a *Spatial and Contextual Models in Political Research*. Taylor & Francis, London

Eagles M (ed.) 1995b *Spatial and Contextual Models of Political Behaviour*. *Political Geography* **14**(6/7)

Grofman B (ed.) 1990 *Political Gerrymandering and the Courts*. Agathon Press, New York

Grofman B (ed.) 1998 *Race and Redistricting in the 1990s*. Agathon Press, New York

Gudgin G, Taylor P J 1979 *Seats, Votes and the Spatial Organization of Elections*. Pion, London

Huckfeldt R, Sprague J 1995 *Citizens, Politics and Social Communication: Information and Influence in an Election Campaign*. Cambridge University Press, Cambridge, UK

Jacobson G C 1980 *Money in Congressional Elections*. Yale University Press, New Haven, CT

Jenkins R 1998 *The Report of the Independent Commission on the Voting System*. HMSO, London

Johnston R J 1980 *The Geography of Federal Spending in the United States*. Research Studies Press, Chichester, UK

Johnston R J 1983 Spatial continuity and individual variability. *Electoral Studies* **2**: 53–68

Johnston R J 1985 *The Geography of English Politics: The 1983 General Election*. Croom Helm, London

Johnston R J 1987 *Money and Votes: Constituency Campaign Spending and Election Results*. Croom Helm, London

Johnston R J, Hay A M 1982 On the parameters of uniform swing in single-member constituency electoral systems. *Environment and Planning A* **14**: 61–74

Johnston R J, Pattie C J 1998 Composition and context: region and voting in Britain revisited during Labor's 1990s' revival. *Geoforum* **29**: 309–29

Johnston R J, Pattie C J, Allsopp J G 1988 *A Nation Dividing? The Electoral Map of Great Britain 1979–1987*. Longman, London
Johnston R J, Pattie C J, MacAllister I, Rossiter D J, Dorling D F L, Tunstall H 1998 Spatial variations in voter choice: modelling tactical voting at the 1997 general election in Great Britain. *Geographical & Environmental Modelling* **1**: 153–79
Johnston R J, Rossiter D J, Pattie C J 1999 Integrating and decomposing the sources of partisan bias: Brookes' method and the impact of redistricting in Great Britain. *Electoral Studies* **18**: 367–78
Johnston R J, Shelley F M, Taylor P J (eds.) 1990 *Developments in Electoral Geography*. Routledge, London
Jones K, Johnston R J, Pattie C J 1992 People, places and regions: exploring the use of multi-level modelling in the analysis of electoral data. *British Journal of Political Science* **22**: 343–80
Key V O Jr 1949 *Southern Politics in State and Nation*. Knopf, New York
Key V O Jr 1955 A theory of critical elections. *Journal of Politics* **17**: 3–18
Lacoste Y (ed.) 1986 *Géopolitiques des Régions Françaises*. Fayard, Paris, 3 Vols.
Lewis-Beck M S 1988 *Economics and Elections: The Major Western Democracies*. University of Michigan Press, Ann Arbor, MI
Lipset S M, Rokkan S E 1967 Cleavage structures, party systems, and voter alignments: an introduction. In: Lipset S M, Rokkan S E (eds.) *Party Systems and Voter Alignments*. The Free Press, New York
McAllister I, Studlar D T 1992 Region and voting in Britain 1979–87: territorial polarization or artifact? *American Journal of Political Science* **36**: 168–99
Miller W L 1977 *Electoral Dynamics*. Macmillan, London
Morrill R L 1973 Ideal and reality in reapportionment. *Annals of the Association of American Geographers* **63**: 463–77
Morrill R L 1981 *Political Redistricting and Geographic Theory*. Association of American Geographers, Washington, DC
Pattie C J, Johnston R J 1995 'It's not like that round here': region, economic evaluations and voting at the 1992 British general election. *European Journal of Political Research* **28**: 1–32
Pattie C J, Johnston R J 1999 Context, conversation and conviction: social networks and voting at the 1992 British general election. *Political Studies* **47**: 887–9
Pattie C J, Johnston R J, Fieldhouse E A 1995 Winning the local vote: the effectiveness of constituency campaign spending in Great Britain, 1983–1992. *American Political Science Review* **89**: 969–86
Reynolds D 1969 A spatial model for analyzing electoral behavior. *Acta Sociologica* **12**: 122–30
Rossiter D J, Johnston R J 1981 Program GROUP: the identification of all possible solutions to a constituency delimitation problem. *Environment and Planning A* **13**: 231–38
Rossiter D J, Johnston R J, Pattie C J 1999 *The Boundary Commissions: Redrawing the UK's Map of Parliamentary Constituencies*. Manchester University Press, Manchester, UK
Rossiter D J, Johnston R J, Pattie C J, Dorling D F L, MacAllister I, Tunstall H 1999 Changing biases in the operation of the UK's electoral system, 1950–1997. *British Journal of Politics and International Relations* **1**:
Shelley F M, Archer J C, Davidson F M, Brunn S D 1996 *Political Geography of the United States*. Guilford Press, New York
Siegfried A 1949 *Géographie Électorale de l'Ardèche sous la IIIe République*. Colin, Paris
Stein R M, Bickers K N 1995 *Perpetuating the Pork Barrel: Policy Subsystems in American Democracy*. Cambridge University Press, Cambridge, UK
Taagepera R, Shugart M S 1989 *Seats and Votes: The Effects and Determinants of Electoral Systems*. Yale University Press, New Haven, CT
Taylor P J, Johnston R J 1979 *Geography of Elections*. Penguin Books, London

R. Johnston

Electoral Systems

An electoral system is a set of rules governing an election. The study of electoral systems has tended to be restricted in a number of respects. First, most research focuses on elections to legislative assemblies and to political executives, even though, of course, many nongovernmental agencies also hold elections as the means of filling various offices. Second, many issues of central concern in the study of electoral systems—such as paradoxes emanating from the application of particular voting rules—are usually studied within the 'theory of voting' branch of rational choice analysis; in some respects, therefore the study of electoral system can be understood as more of an 'applied' social science, while the 'theory of voting' is 'pure' theory.

1. *Elements of Electoral Systems*

Following Rae (1971), many scholars distinguish between three main elements of an electoral system: ballot structure, constituency structure, and the electoral formula.

1.1 *Ballot Structure*

There are several different ways in which this can vary, including:

(a) The number of votes each voter is permitted. In some systems the voter casts as many votes as there are positions to be filled (e.g., plurality systems with multimember constituencies—as employed in many state legislatures in the USA before the 1960s); in others, voters may cast only one vote irrespective of how many positions are being filled (e.g., the Single Non-Transferable Vote system used in Japan before 1994).

(b) The kind of information the voter can reveal in voting. Some systems permit two different kinds of vote to be cast; for example, in the German Additional Member System (AMS) the voter votes for a party candidate in a constituency, as well as for a party for

the party list component of the ballot. In many versions of party list systems the voter can reveal both which party they prefer, and which of that party's candidates they prefer. Again, by combining two separate elections, a system may make it possible for voters to reveal extra information about their preferences. For example, in Israel since 1996, voters reveal not merely which party they prefer, but also which party they wish to form the government: the Israeli voter can vote both for a party in the election for the legislature, and for a particular candidate to be prime Minister following that election. Furthermore, in some systems, voters can reveal only their first preference (whether for a party or a candidate), while in others they can provide a rank ordering of the different parties (or candidates); the plurality system used in British parliamentary elections is an example of the former, while the Single Transferable Vote (STV) employed in the Republic of Ireland is an example of the latter. Finally, some electoral systems limit the expression of voter preference to those candidates or parties listed on the ballot, while others allow the 'writing in' of names not on the ballot, or allow the voter to register opposition to all those listed on the ballot by voting for 'none of the listed candidates.'

(c) Whether the ballot is open or secret. In the twentieth century it has become the norm for national elections to be contested under a secret ballot, but until the late nineteenth century balloting was a public act. In the USA, parties in the nineteenth century printed their own ballot papers, and in several states the parties used different colors of paper for the ballot so they could tell who had voted for them. Whether the ballot should be secret or not was a major controversy in nineteenth-century political debates, and some of those who advocated an extensive franchise (e.g., John Stuart Mill) still favored an open ballot.

1.2 Constituency Structure

There are three main ways in which electoral systems vary in this regard.

(a) The election can be conducted on an 'at large' basis, or an election can be held in each of several constituencies. Generally, though with some exceptions, 'at large' elections are held only in smaller territories and when a proportional mechanism is used to allocate seats in an assembly (e.g., The Netherlands). Most national and subnational assembly elections are divided into constituencies.

(b) Whether the constituencies are organized on a territorial or some other basis. Although there have been debates about the desirability of a functional, rather than a territorial, basis for representation, most electoral systems use the latter. The odd exceptions include the 'university seats' of the British Parliament that were abolished before the 1950 election. Any territorial subdivision of a polity involves the drawing of boundaries for the constituencies, and how the boundaries are drawn may affect the overall election result. This issue becomes most significant in systems using nonproportional means for allocating seats, but it is still an issue in those using proportional mechanisms.

(c) Whether the constituencies are single-member or multimember. Systems that employ proportional means for allocating seats in an assembly have to use multimember systems, as does the SNTV; other systems do not. Generally, polities using the plurality system for allocating seats have moved towards the exclusive use of single-member constituencies during the twentieth century.

1.3 Electoral Formula

This is the way in which votes are connected to the office(s) being contested—in an assembly it is those rules governing how votes are translated into seats. There are three main kinds of formulae:

(a) *Majoritarian*. Under these systems, the winning candidate or party must obtain more than 50 percent of the vote. The two most commonly used means for achieving this are the Alternative Vote System (AV), and the double Ballot System (DB). The AV, which is used in Australia, involves voters showing their second, and subsequent, preferences on the ballot form; when no candidate obtains more than 50 percent of the vote, the lowest-placed candidate is eliminated and his/her voters have their second-choice votes counted. This process continues until one candidate attains more than 50 percent of the vote. The DB requires a second election to be held in all constituencies in which no candidate obtains more than 50 percent of the vote; this second contest includes only the two leading candidates from the first election.

(b) *Plurality*. Here the candidate(s) with the largest number of votes wins. In relation simply to its electoral formula, the SNTV system is merely a plurality system; however, its very different ballot structure—voters having only one vote each in multimember constituencies—mean that it is a radically different system in its operation.

(c) *Proportionality*. This formula attempts to provide a 'fit' between the share of the votes that the candidates or parties obtain, and the share of the seats the parties get in the assembly. There are two main devices for achieving this. One is to use a formula, of which there are a number of variants including d'Hondt and St. Lagüe, that distributes seats to parties on the basis of their share of the vote. The other method is to permit voters to rank-order all candidates, and to redistribute the 'wasted votes' of those candidates who have more votes than they need to get elected; that, broadly speaking, is the basis of STV. Disproportionalities in supposedly proportional systems can occur for three main reasons:

(i) the ways in which electoral boundaries are drawn, if the system is divided into constituencies, thereby giving an advantage to particular parties;

(ii) relatively small district magnitude—that is, a small number of seats being contested in each constituency; the fewer the seats in each constituency, the more likely it is that, in aggregate the allocation of seats between parties will diverge from the distribution of total votes. (STV has to be employed in conjunction with relatively small district magnitude; it becomes unwieldy with large district magnitude because of the complexities involved in expressing preferences between a large number of candidates.); and

(iii) in party-list systems, through the use of thresholds which require that parties failing to secure a minimum share of the vote will fail to win any seats at all. A threshold of between 2 and 3 percent of the total vote usually reduces the number of parties represented in a legislature quite dramatically, while a threshold of 5 percent (as used in Germany) restricts seriously the number of parties represented.

2. *Classification of Electoral Systems*

There are a very large number of possible electoral systems, including many hybrids of the most common types. (An example of such a hybrid is the system used for the French National Assembly, which is a variant of the DB; a second ballot is used, but it is not confined to the two leading candidates from the first ballot. Instead, any candidate with more than 12.5 percent of the vote is permitted, but not required, to take part in the second ballot.) Given this large variety, considerable attention has been given to classifying the main types of system. It is widely agreed that the main distinction is between 'proportional' systems and other types, but there is some disagreement on two matters.

The first is whether the principal division is a bifurcation between 'proportional' and 'non-proportional' systems, or whether a more complex main division is appropriate. The second disagreement concerns how to treat a variety of electoral systems that are sometimes grouped together under the heading 'semiproportional'; these include the SNTV and the Cumulative Vote (CV). (The latter is a system used in Illinois for many years that, in a three-member constituency would give three votes to each voter but would allow the voter to give all three votes to one candidates, give two votes to one candidate and one to another, or to give a single vote to each of three candidates.) Sartori (1994), for example, argues in favour of treating the SNTV as a form of proportional representation (PR), but most scholars do not do so.

The most common classification is to distinguish between those systems that have mechanisms for generating proportional results that are internal to the systems themselves, and which do not depend for their proportionality on the behavior or strategies of voters or candidates, from those that do not.

2.1 *Proportional Electoral Systems*

Within the class of proportional systems there are two main variants—the *party-list* and the *STV* systems, and within the party-list subclass there are four main sub-subclasses:

2.1.1 *Party-list*

(a) *Pure party-list*. In these systems the order of the party's candidates on the list, and hence their likelihood of being elected, is determined exclusively by the party. The voter can vote only for a party and not for a candidate. (Such systems are relatively uncommon, but this system is used in Israel and was used in France in 1945–6, and in a slightly less pure form during the rest of the fourth Republic.)

(b) *Voter-choice from party-list*. These systems allow the voter to express a choice between candidates on the party-list, though great variations in the procedures are evident. In some systems it is actually difficult for voters to propel a popular candidate from a low position on the list to a higher one; in other systems it is much easier to do so. These types of system are the most common form of the party-list, having been used in many European countries.

(c) *Mixed systems*. The AMS is the principal variant of this type. A proportion of the seats (sometimes one half) are contested in single-member constituencies under plurality voting. The remaining seats are contested using a party-list, and the allocation of seats from the party-list is used so as to correct for any disproportionalities resulting from the elections in the constituencies. This type of system, which is used in Germany and New Zealand, produces election results that are rather similar to those that would occur under other party list systems.

(d) *Panachage*. Used in Switzerland, this system permits voter choice between candidates of more than one party. Some experts argue that it is not really a form of party-list system because of this rule, although its operation more closely resembles such systems than any other types.

2.1.2 *STV*

Unlike party-list systems, STV achieves proportionality in the distribution of seats in a legislature through a mechanism that allocates seats to particular candidates. A party acquires seats through given candidates of theirs receiving sufficient votes to win seats, rather than the parties being allocated votes through a formula that distributes seats directly to the parties.

2.2 *Nonproportional Electoral Systems*

Among the nonproportional systems, there are three main subclasses:

(a) *Plurality*. The first-past-the-post system has been used in many Anglo-Saxon democracies, and it is still the system used in Canada, and most of the USA, as well as for elections to the UK Parliament.

(b) *Majoritarian*. The most common variants of systems requiring the winning candidate (or party) to win more than 50 percent of the vote are the AV and DB. Outside national elections, there are some majoritarian systems that require the winning candidate to receive more than 50 percent of the vote—one example being the Democratic National Convention which, until the 1930s, required that the nominee for the presidency receive two-thirds of the delegates' votes. Since such systems can produce many rounds of balloting and still fail to produce a result, they are wholly unsuitable for elections at which electors cannot interact personally with each other in an effort to achieve a definite result.

(c) *Semi-proportional*. This is a subclass containing several rather different kinds of system, including the CV, SNTV and the Limited Vote (LV). The last is rather similar to the SNTV in that there are multi-member constituencies, but, whereas in SNTV voters have only one vote, under LV they have more than one vote, but fewer votes than there are seats being contested. It is employed for some seats in the Spanish Senate, and was used in a few urban constituencies in Britain between 1867 and 1883. Semiproportional systems acquire their (rather misleading) name from the fact that they tend to generate greater representation for smaller parties than does the plurality system. However, the objection to classifying them as a form of proportional system is not that they tend to produce disproportional results—though in fact they do tend to do so—but rather that how proportional they are depends on the particular behaviour of voters and parties. In some situations they may not be proportional at all. By contrast, proportional systems are designed to generate a degree of proportionality.

3. *Consequences of Electoral Systems*

Public and academic interest in electoral systems tends to be much greater when the franchise is being extended (as in Britain in the 1860s), or new legislative assemblies are introduced (Britain in the 1990s) when democratic regimes have collapsed or are unstable (Europe from the 1920 to the 1950s), and when regimes are democratizing (Eastern Europe in the 1990s), or redemocratizing (Latin America in the 1980s and 1990s). One of the main focal points of attention is the likely consequences that one form of electoral system will have compared with others.

In the 1950s, and in light of the experience of collapsed democracies in the interwar years and with the potentially unstable regimes in postwar France and Italy, Duverger (1959) propounded an argument that has come to be known as Duverger's Law: 'the simple-majority single ballot system favours the two-party system'. He also argued that PR tended to be associated with multipartism. His argument was often linked to the widely held view in Britain and the USA at that time that PR was responsible for unstable governments with multiparty systems, of the kind found in the fourth Republic, whereas plurality systems were a recipe for stable two-party democratic government.

Since Duverger's Law was first propounded it has had its critics, and at the time of writing many political scientists would point to its limitations. It does not apply when there are strong regional cleavages, for example; they also reject the popular view that PR was causally responsible for unstable government. That view does not fit well with the evidence: most European countries since 1945 have used some form of PR, and most have had multiparty systems and coalition governments, yet most have not had unstable regimes. Although there are some exceptions, notably Sartori (1994), many political scientists now argue that the electoral system has only a limited effect on a regime's party system, and hence on its potential for instability. Increases in the number of parties in a system, for instance, tend not to follow the introduction of PR, but rather tend to precede it; moreover, the timing of the introduction of PR is significant in relation to whether party fragmentation will occur. An early switch to PR following democratization does tend to produce fragmentation—because the larger parties have had limited opportunities to develop loyal electorates. A switch much later does not lead to fragmentation, because the established parties now have loyal electorates who are not likely to defect to new or small parties.

Nevertheless, the precise form of electoral system used can have an impact on election results: changing the rules does change the political game. For example, the switch from plurality voting to AV in British Columbia in the early 1950s probably did prevent the socialist CCF party from gaining power in that decade. Similarly, the switch to PR from DB for the 1986 Assembly elections in France helped to minimize the Socialists' losses and to establish the National Front as a credible party that could win seats in the Assembly. Electoral systems cannot transform the social bases of electoral politics, but they can make it more likely that some parties will win office, or share power in government, and make it less likely that others will do so. For that reason, politicians are tempted sometimes to change them. For the most part, though, they are not changed frequently—partly because those in office are the people who usually benefit from the working of the existing system, and partly because the full consequences will be unknown, and hence might actually run counter to the interests of those seeking change.

See also: Democratic Theory; Minimum Winning Coalition, in Politics; Parliamentary Government;

Party Systems; Pluralism; Political Representation; Primary Elections; Proportional Representation; Voting: Class; Voting: Compulsory; Voting: Retrospective/Prospective; Voting, Sociology of; Voting: Tactical

Bibliography

Blais A, Massicotte L 1996 Electoral systems. In: LeDuc L, Niemi G, Norris P (eds.) *Comparing Democracies*. Sage, Thousand Oaks, CA, pp. 49–82

Duverger M 1959 *Political Parties*, 2nd edn. Methuen, London

Lijphart A 1994 *Electoral Systems and Party Systems*. Oxford University Press, Oxford, UK

Rae D W 1971 *The Political Consequences of Electoral Laws*, 2nd ed. Yale University Press, New Haven, CT

Sartori G 1994 *Comparative Constitutional Engineering*. Macmillan, Basingstoke, UK

A. Ware

Electrical Stimulation of the Brain

Direct electrical stimulation of brain tissue, in both animals and humans, has proven over more than a century of experimentation to be a truly powerful tool for revealing functional relations between brain and behavior as well as connections within the brain. Various sorts of current have been delivered through electrodes that are either applied acutely to localized brain regions or chronically implanted deep within the brain or on its surface. This article will describe experimental studies that have used animal models in laboratory settings, and clinical studies in which human patients have received electrical stimulation of the brain (ESB) in order to observe its effects, including especially possible therapeutic effects.

1. *Brain Stimulation and the Elicitation of Motor Responses*

In the late nineteenth century, pioneering experiments mapped motor functions of mammalian cerebral cortex using ESB (e.g., Fritsch and Hitzig 1870). These investigators, using dogs, were able to elicit discrete motor responses of specific muscles and muscle groups by applying brief electrical stimuli directly upon the exposed cortical surface. These studies laid the empirical foundation for our current understanding of topographic motor maps on the brain's surface; later work filled in the outline provided by these early investigations. Central concepts began to emerge that eventually shaped current thinking about brain function and guided much later research. These include the notion of functional localization within brain structures, an idea originally set forth in the pseudoscientific ideas of phrenology. With the contributions of Fritsch and Hitzig (and others), however, a valid basis was afforded for specific types of localization. Not only was the frontal region of mammalian cortex recognized as a major source of neural output controlling responses of the striate musculature, but this region was observed to contain a map which was topographic in nature; that is, adjacent muscle groups are represented in adjacent groups of neurons. Furthermore the principles of both proportional and contralateral representation became apparent. Muscle groups capable of relatively finely calibrated, skilled movements (e.g., digital and some facial muscles) are represented by very much larger populations of neurons than muscles capable only of grosser movements, so that there is a topological distortion within the topographical map (proportional representation); also, cortical motor neurons in each of the brain's hemispheres exert their influence over muscles on the opposite side of the body (contralateral representation).

2. *Brain Stimulation and Perceptual Prostheses*

Analogous concepts have been developed for the cortical representation of somatosensory responses originating at the skin: topographical maps exist on each cortical hemisphere representing the contralateral side of the body's surface, with topological distortions similar to those of motor cortex, but corresponding instead to differences in sensory acuity, with the digits and some facial areas over-represented in terms of quantity of neural tissue involved. The somatosensory maps have been localized to the central and posterior parts of the cortex (in humans, the parietal lobe). Auditory and visual maps have also been studied in detail; the two halves of the visual field (left and right), for instance, are mapped onto the contralateral cortical hemispheres in the most posterior part of the cortex (in humans, the occipital lobe). Although multiple maps exist, the principle of proportional representation is clearly evident in the primary map, where the areas of greatest acuity (at and near the fixation point) are greatly over-represented compared to peripheral parts of the visual field.

Since it is widely accepted that all sensory responses are due to some unique pattern of electrical activity in the corresponding sensory cortex, it has been evident for some time that, in theory at least, an appropriate pattern of ESB within any sensory region of cortex might be arranged to mimic any specific normal, exteroceptive pattern of sensory stimuli in the corresponding modality. The possibility thus exists of developing prostheses for artificial vision or hearing,

for example, using brain stimulation to evoke perceptual responses in persons lacking functional peripheral receptors or even sensory nerves.

Considerable exploratory work has been done on these problems, although practical prostheses based on ESB are not at the moment close to realization. A promising approach developed recently involves the implantation of relatively large (30–100) arrays of microelectrodes arranged to deliver ESB to the auditory or visual parts of cortex. For instance, cats trained to lever-press to an external auditory stimulus will respond similarly to very modest levels of ESB delivered over an array of microelectrodes implanted within auditory cortex (Rousche and Normann 1999); furthermore, patients with profound or total bilateral hearing loss may receive acoustic information through stimulation of electrodes implanted in the auditory regions of the brainstem (Laszig and Aschendorff 1999). By the same token, visual responses have been demonstrated in a blind patient receiving intracortical ESB over an array of electrodes implanted in the occipital lobe (Schmidt et al. 1996).

3. *Brain Stimulation and Associative Learning*

Since the 1980s a major neuronal network has been delineated, responsible for much of what memory theorists call 'procedural learning' (e.g., Thompson and Kim 1996). The elements of this network specifically responsible for plasticity and associative learning are located within the cerebellum, and include the mossy and climbing fiber systems afferent to (providing information for) intracerebellar structures; the cerebellar cortex, especially its Purkinje cells and their inputs and output; and the deep cerebellar nuclei, especially the anterior interpositus nucleus. A useful model that has frequently been used to study this network is the classical conditioning of eyeblink and other discrete motor responses in the rabbit: a neutral or conditioned stimulus, the CS, which does not initially elicit a response, is paired repeatedly with an eyeblink eliciting stimulus (the unconditioned stimulus or US), usually a brief airpuff delivered to the cornea. After 100–200 trials the CS acquires the property of eliciting the eyeblink (associative learning or acquisition). If the rabbit is subsequently exposed to repeated CS presentations in the absence of the US, the conditioned response weakens and eventually disappears, in a process referred to as extinction. An appealing feature of this preparation involves the similarities seen here between human and animal learning. The form of the response, the curves of both learning and extinction, and certain more detailed aspects such the timing of the response within each trial and the changes in timing that are seen as conditioning proceeds, are all quite similar when comparing rabbits and humans. It is generally accepted that this particular form of conditioning, which is simple and readily adapted for well-controlled laboratory studies, shares many features with human learning of motor skills, including both simple and complex motor learning and performance.

Validation of the role of these cerebellar network elements and circuits in associative learning has involved the application of numerous neurobiological approaches in both animals and humans. The use of ESB to selectively activate the specific cerebellar neuronal networks involved in procedural learning has been particularly fruitful (e.g., Thompson et al. 2000). The earliest use of brain stimulation to demonstrate the direct involvement of the cerebellum in learning was reported by Brogden and Gantt, who showed that a variety of motor responses, including eyelid closure, forelimb and hindlimb flexion, head turn, and pinna movement elicited by electrical stimulation of cerebellar white matter (chiefly in lobule HVI) were readily conditioned to auditory and visual CSs, using conventional stimulus pairing procedures of classical (Pavlovian) conditioning (Brogden and Gantt 1937, 1942). Recent experiments have been aimed at identifying and characterizing more closely the relevant pathways, using electrical microstimulation delivered over chronically implanted semi-microelectrodes in the awake, behaving rabbit during conditioning. For instance, ESB of the inferior olive, which elicits specific motor responses, substitutes effectively for the unconditioned stimulus when paired with a CS in classical conditioning, thereby lending support to the role of olivary efferent pathways (climbing fibers) in providing information about the unconditioned stimulus. By the same token, studies of the conditioned stimulus pathway have established that ESB delivered to either the pontine nuclei or the middle cerebellar peduncle, activating the mossy fiber projections to cerebellum, serves the same function as the tone that has long been used in conventional eyeblink classical conditioning. Steinmetz et al. (1989) combined these effects in a single preparation, and showed that conditioning is reliably obtained using pontine ESB as the CS and olivary ESB as the US, dispensing altogether with exteroceptive stimulation.

In subsequent experiments associative learning has been shown in a preparation in which both CS and US are delivered directly to, and confined within, the cerebellum. Four chronic stimulating electrodes are implanted in rabbit cerebellum, one bipolar pair in cerebellar cortex, providing an electrical CS activating cortical parallel fibers and thence Purkinje cells, and another bipolar pair in underlying white matter of lobule HVI, delivering an electrical US that elicits unconditioned responses. Paired CS–US presentations lead reliably to the development of conditioned responses; after conditioning is established, repeated presentations of the CS alone lead to extinction as in conventional conditioning. Also, increased local excitability is observed in cerebellar cortex following conditioning; that is, the threshold for eliciting

responses to ESB delivered over the CS electrode is significantly decreased after conditioned responding is well established. Pseudorandom unpaired presentation of CS and US does not produce any CRs, indicating that true associative learning, and not sensitization, is responsible for these results (Shinkman et al. 1996).

Furthermore, associative learning that is localized within the cerebellar cortex has been demonstrated in exploratory studies, discussed in Thompson et al. 2000). These investigators delivered ESB over a single bipolar electrode implanted in the Purkinje cell layer of cerebellar cortex and used this stimulation both as the CS (very low-level stimulation that does not elicit any response) and as the US (higher-level stimulation through the same electrode that reliably elicits a discrete unconditioned motor response). When paired CS–US trials are followed by CS-alone trials, in a paradigm formally identical to a standard classical conditioning protocol, both acquisition and extinction of conditioned responses is observed, and following conditioning there is increased local excitability in cerebellar cortex (the threshold for response elicitation decreases). The effects of highly localized blocking of neuronal activity have also been studied in this preparation, using intracerebellar injections of muscimol, a substance that reversibly inactivates the firing of neurons in the immediate vicinity of the injection. The expression of previously established conditioned responses is blocked by injection of muscimol into the region surrounding the electrodes, but occurs normally in the absence of muscimol. Because muscimol affects only neuronal cell bodies, and does not influence fibers of passage, it can safely be concluded that its presence has no remote effects other than those attributable to the lack of output from the region of inactivation. These results suggest that the observed plasticity involved in classical conditioning depends upon changes in the interaction between neuronal elements in cerebellar cortex, especially parallel fibers and Purkinje cells, and in the deep cerebellar nuclei.

4. *Brain Stimulation as Reward: 'Pleasure Centers' in the Brain*

In 1953 James Olds, a postdoctoral fellow at McGill University in Montreal, was studying behaviors elicited in rats by the application of ESB. He was using a relatively new technique in which stimulating electrodes are chronically implanted under surgical anesthesia so that stimuli can be delivered at specified locations in the brain while the subject is awake, relatively unencumbered, and engaging in ongoing behaviors. Although the focus of these experiments was on behaviors directly elicited by the stimulus, Olds had the remarkable insight, based upon some of his subjects' behaviors, that the stimulus was in some cases apparently rewarding (or, more technically, reinforcing) whatever behavior immediately preceded the application of the stimulus. That is, in B. F. Skinner's terminology, the stimulus was serving a 'reinforcing' rather than an 'eliciting' function. Olds quickly undertook a series of experiments to verify this intuition, and ascertained that in specific subcortical regions (e.g., lateral hypothalamus, septal area), ESB can exert powerful reinforcing effects (Olds and Milner 1954).

In terms of potency, the reinforcing effects of ESB delivered using appropriately placed electrodes far outweigh those of conventional (i.e., biological or adaptive) reinforcers. To establish that this was so, Olds unearthed a long-forgotten apparatus called the Columbia Obstruction Box, which had been used in the 1930s at Columbia University in an ill-conceived attempt to 'compare motives.' Rats were placed on a platform separated by an electrified grid from a goal. In the original experiments, the amount of footshock the subject would tolerate to reach the goal was compared for deprived rats responding for, e.g., water, food, or sex (the flaw here, of course, is that deprivation levels for different goals do not fall on the same metric). Nevertheless, levels of footshock absorbed by implanted rats to reach rewarding ESB far exceeded those for all other reinforcers, including the previously top-ranked one—access for a mother rat to her newborn pups.

An enormous literature on various aspects of this extraordinary finding emerged over subsequent decades. A particularly intriguing line of experimentation has linked rewarding ESB with the reinforcing properties of addictive substances such as cocaine and opiates. It has now been established that the rewarding effects of ESB may be augmented or potentiated in the presence of such substances. Indeed, it appears likely that addictive substances and rewarding ESB may both exert their reinforcing effects upon a variety of behaviors by activating the same neural circuits involved in the brain's response to conventional or natural reinforcers (e.g., Wise 1996).

5. *Brain Stimulation in Human Patients*

Electrical stimulation of the brain through chronically implanted electrodes, with the aim of alleviating various symptoms in human patients, has been used infrequently but to considerable effect since the 1950s. Early studies (e.g., Heath and Mickle 1960, Mark and Ervin 1970) focused on very severe affective disorders. In at least some patients, it was shown that symptoms including violent or inappropriately aggressive behavior could be triggered by stimulation of certain limbic structures, notably the amygdala, and abolished or prevented by ablation (removal) of these structures or by stimulation of regions (e.g., the septal area) whose activity is mutually antagonistic to amygdaloid excitation.

Recent studies have concerned a variety of other disorders, including conditions as diverse as Parkinson's disease, multiple sclerosis, intractable epilepsy, and chronic pain. It is now well established, for instance, that stimulation of certain parts of the human motor nervous system, collectively designated as the extrapyramidal system, can lead to dramatic improvement in the motor disorders experienced by patients suffering from Parkinson's disease. In this case the stimulation serves to functionally block the abnormal activity of these brain regions; although similar effects can be achieved by ablation, the use of ESB is clinically preferable (e.g., Montgomery 1999), because it is reversible, readily calibrated and adjusted, and may avoid such side effects as cognitive impairment.

While studies of the clinical efficacy of ESB in reducing or abolishing chronic pain have yielded conflicting results, the most common finding has been an impressive reduction of symptoms and a corresponding increase in patients' quality of life (e.g., Young et al. 1985). This particular application, like many others, has its origins in the animal experimental literature. It is interesting to note that some of the most effective electrode placements in human pain patients (e.g., brainstem periaqueductal and periventricular gray matter) correspond to regions where stimulation had been shown much earlier to be both rewarding and analgesia-producing in rats.

Newer stereotaxic techniques allow for guided, precise electrode placement; at the same time, advances in electrode technology permit the application of highly localized, parametrically controlled stimulation. It seems reasonable to expect that clinical applications for ESB will continue to expand, and that directions for such expansion will continue to be informed by basic experimentation on the effects of ESB in laboratory animals.

See also: Brain Stimulation Reward; Cerebellum: Associative Learning; Electroencephalography: Basic Principles and Applications; Electroencephalography: Clinical Applications; Eyelid Classical Conditioning; Topographic Maps in the Brain

Bibliography

Brogden W J, Gantt W H 1937 Cerebellar conditioned reflexes. *American Journal of Physiology* **119**: 277–8

Brogden W J, Gantt W H 1942 Intraneural conditioning. *Archives of Neurology and Psychiatry* **48**: 437–55

Fritsch G, Hitzig E 1870 Ueber die elektrische Erregbarkeit des Grosshirns. *Archiv fur Anatomie, Physiologie und wissenschaftliche Medicin* no. 428, pp. 300–32

Heath R G, Mickle W A 1960 Evaluation of seven years' experience with depth electrode studies in human patients. In: Ramey E R, O'Doherty D S (eds.) *Electrical Studies on the Unanesthetized Brain*. Hoeber, New York, pp. 214–42

Laszig R, Aschendorff A 1999 Cochlear implants and electrical brainstem stimulation in sensorineural hearing loss. *Current Opinion in Neurology* **12**: 41–4

Mark V H, Ervin F R 1970 *Violence and the Brain*. Harper & Row, New York

Montgomery E B 1999 Deep brain stimulation reduces symptoms of Parkinson disease. *Cleveland Clinic Journal of Medicine* **66**: 9–11

Olds J, Milner P 1954 Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology* **47**: 419–27

Rousche P J, Normann R A 1999 Chronic intracortical microstimulation (ICMS) of cat sensory cortex using the Utah Intracortical Electrode Array. *IEEE Transactions on Rehabilitation Engineering* **7**: 56–68

Schmidt E M, Bak M J, Hambrecht F T, Kufta C V, O'Rourke D K, Vallabhanath P 1996 Feasibility of a visual prosthesis for the blind based on intracortical microstimulation of the visual cortex. *Brain* **119**: 507–22

Shinkman P G, Swain R A, Thompson R F 1996 Classical conditioning with electrical stimulation of cerebellum as both conditioned and unconditioned stimulus. *Behavioral Neuroscience* **110**: 914–21

Steinmetz J E, Lavond D G, Thompson R F 1989 Classical conditioning in rabbits using pontine nucleus stimulation as a conditioned stimulus and inferior olive stimulation as an unconditioned stimulus. *Synapse* **3**: 225–33

Thompson R F, Kim J J 1996 Memory systems in the brain and localization of a memory. *Proceedings of the National Academy of Sciences USA* **93**: 13438–44

Thompson R F, Swain R A, Clark R E, Shinkman P G 2000 Intracerebellar conditioning – Brogden and Gantt revisited. *Behavioural Brain Research* **110**: 3–11

Wise R A 1996 Addictive drugs and brain stimulation reward. *Annual Review of Neuroscience* **19**: 319–40

Young R F, Kroening R, Fulton W, Feldman R A, Chambi I 1985 Electrical stimulation of the brain in treatment of chronic pain: Experience over 5 years. *Journal of Neurosurgery* **62**: 389–96

P. G. Shinkman

Electroencephalography: Basic Principles and Applications

1. Introduction

Electrodes placed upon the scalp record the integrated post-synaptic potentials generated by inhibitory as well as excitatory activity in the subjacent neuronal population. If this activity were distributed randomly in polarity and in time, the net surface potential would be zero. Correlative studies between the scalp and the pial surface have resulted in the estimate that about $6\,cm^2$ of cortex must be synchronously active in order for a net voltage deflection to be detected on the scalp. Thus, EEG potentials arise from highly nonrandom synchronized activity in very large neuronal assem-

blies. This coordination of discharge is mediated by cortical-cortical, cortico-thalamic and thalamo-cortical processes which, in view of the high degree of reciprocity in intra-cortical connections, may reflect some degree of reverberatory activity. Such synchronization of oscillatory discharges may involve electrotonic as well as trans-synaptic interactions.

Zero-phase synchronization of high frequency oscillations between anterior and posterior brain regions takes place shortly after presentation of sensory stimuli. Energetic contemporary research suggests that synchronization of high-frequency oscillatory activity among distant brain regions may play an important role in 'binding'—the integration of features of a complex stimuli, extracted by spatially separated 'feature detectors,' into a perceptual whole.

These aspects of brain organization and information processing are accessible from quantitative analyses of EEG recordings from the scalp. Usually, such recordings are obtained from an array of electrodes placed in accordance with the internationally standardized 10/20 Electrode Placement System and affixed to the scalp with paste. Data can be recorded from different numbers of electrode positions varying with the purpose of the study, but 19 channel recordings most commonly are used. Numerous economical and reliable EEG recording instruments are commercially available. While EEG machines initially recorded ink traces upon moving paper, present technology utilizes desktop computers for digital collection and storage of EEG data, so-called digital EEG or DEEG. Analysis usually is performed by computer after visual selection of artifact-free data.

These analyses usually utilize the Fast Fourier Transform (FFT) to quantify the amount of power at selected frequency intervals in the EEG from every recording channel to compute the power spectrum. Normative QEEG databases have been constructed for the power spectrum and the covariance matrix of spontaneous brain electrical activity. QEEG methods provide sensitive and precise measures for study of normal as well as abnormal brain functions.

2. *Neurophysiological Basis of the EEG*

An anatomically complex homeostatic system regulates the EEG power spectrum. Brainstem, thalamic, and cortical processes involving large neuronal populations mediate this regulation, utilizing all neurotransmitters. Some of the neuroanatomical interactions and neurochemical transmitters comprising this control system have been clarified by recent neurophysiological and neuropharmacological research (Llinas 1988, Lopes Da Silva 1996, McCormick 1992, Steriade et al. 1990). On the basis of such studies, it is believed that pacemaker neurons distributed throughout the thalamus normally oscillate synchronously. Efferent thalamo-cortical projections globally distributed across the cortex receiving cortico-thalamic feedback produce the rhythmic electrical activity known as the alpha rhythm (7.5–12.5 Hz), which dominates the EEG of an alert healthy person at rest and can be seen readily as the most prominent peak in the power spectrum. Nucleus reticularis can hyperpolarize the cell membranes of thalamic neurons by gamma-amino-butyric acid (GABA) release, slowing the dominant alpha rhythm into the lower theta range (3.5–7.5 Hz) and diminishing sensory throughput to the cortex. Slow delta activity (1.5–3.5 Hz) is believed to originate in oscillator neurons in deep cortical layers and in the thalamus, normally inhibited by input from the ascending reticular activating system in the midbrain. Activity in the beta band (12.5–20 Hz) is believed to reflect cortico-cortical and thalamo-cortical transactions related to specific information processing. Activity in the 20–60 Hz band is referred to as gamma, and recent studies have included so-called 'higher frequency' activity as high as 600 Hz. These very fast oscillations appear to arise from cortico-cortico transactions and geniculo-cortical projections.

Evidence is reviewed below to support the belief that a complex homeostatic system regulates the stability and predictability of the normal EEG and may be genetically specified. A circuit summarizing the hypothetical neuroanatomy and neurochemistry of the homeostatic system postulated to achieve regulation of the EEG power spectrum has been published (Hughes and John 1999), in accordance with the insights provided by the research cited above. Activation of the mesencephalic reticular formation (MRF) causes inhibition of nucleus reticularis by cholinergic and serotonergic mediation which releases the thalamic cells from inhibition by nucleus reticularis. The dominant activity of the EEG power spectrum becomes more rapid, with return of alpha activity and the higher frequency beta activity (12–20 Hz), and the flow of information through the thalamus to the cortex is facilitated. The cortex can activate n. reticularis directly by glutamatergic pathways to suppress the arrival of information to the cortical level and, by dopaminergic projections, the striatum can inhibit the MRF. Such inhibition of the MRF enables inhibition of thalamic neurons to occur and blocks the flow of sensory information through the thalamus to the cortex.

This model suggests that deficiencies or excesses of any of the neurotransmitters should produce marked departure from the homeostatically regulated normative EEG spectrum. Such neurotransmitter perturbations are widely believed to make decisive contributions to much psychiatric pathophysiology. Abundant documentation exists in the literature that demonstrates the sensitivity of QEEG to psychiatric disorders (for a review, see Hughes and John 1999).

Since this regulatory neurophysiological system is common to all humans, the EEG power spectrum can

reasonably be expected to be stable and characteristic for all healthy individuals, with high specificity reflecting their common genetic heritage. These measures can also be expected to be sensitive to many modulations of the resting brain state as well as to brain dysfunctions present in some developmental, neurological and psychiatric disorders. The data supports both of these expectations. A thorough review of the abundant such evidence is beyond the scope of this article; only illustrative examples will be cited.

3. Principles of QEEG

3.1 Normative Databases

The QEEG has been found to display remarkable stability and reproducibility among healthy, normally functioning individuals of any age, ethnic or cultural background (John and Prichep 1993). After proper editing to remove artifacts and transformations to achieve normal distributions and appropriate age-regression, the values of extracted QEEG measures referred to normative databases possess high predictability (Fein et al. 1983, Oken and Chiappa 1988). Test–retest replicability, both short- and long-term, has been shown to be extremely high (Kondacs and Szabo 1999).

3.2 Data Acquisition and Analysis

In QEEG, multichannel recordings (usually 19 electrodes at standardized positions) of eyes closed resting EEG are edited visually. An artifact-free sample, usually 1–2 min long, is analyzed using the FFT. The power at each frequency of the EEG is averaged across the entire sample, providing a Power Spectrum with very high test–retest replicability.

The bandwidth of interest has long been considered to extend from about 1–20 Hz. traditionally separated into four wide frequency bands, typically defined as Delta (1.5–3.5 Hz), Theta (3.5–7.5 Hz), Alpha (7.5–12.5 Hz), and Beta (12.5–20 Hz). More recently, the region of interest has been extended to include Gamma or 'high frequency activity,' with bands up to 100 Hz and beyond. Intracerebral as well as scalp recordings have related oscillatory activity at these high frequencies to cognition and information processing (Curio 2000).

4. Stability and Specificity of the EEG Power Spectrum—The Basis of QEEG

Confirmed reference databases provide reliable 'population norms,' thereby minimizing the need to collect control data. Comparison of individual QEEG results to such norms enables sensitive detection of developmental, neurological or psychiatric dysfunctions. Further, QEEG stability facilitates identification of transient brain activity related to temporary changes of brain state due to cognitive activity or other influences, by comparing sequential QEEG samples from the same person to each other or to a preceding 'self-norm.' QEEG studies using statistical evaluation of computer extracted features relative to population or self reference norms are often referred to as 'neurometric' studies (John et al. 1977, 1988). A major advantage of QEEG vs. all other contemporary brain imaging technologies is the unique availability of confirmed, ethnic fair, normative data that reduces or eliminates the need for separate control studies. The high specificity of QEEG is beyond the confidence level achieved by many routinely used clinical tests, such as mammograms, cervical screenings, or CT brain scans (Swets 1988). Further, the demonstrated high test–retest replicability provides potential simplification in repeated measure experimental designs.

Using analog methods, the eyes closed, resting EEG power spectra of large samples of healthy normally functioning individuals were found to show systematic changes, initially from ages 17–64 (Matousek et al. 1967), and then from ages 1–21 (Matousek and Petersen 1973). Using general purpose digital computers, not only were the systematic changes with age soon confirmed, but no significant differences were found between the age-regressed EEGs of normally-functioning Swedish children and white or black US children (John et al. 1980).

Normative data now exist from 1 to 95 years of age, for each electrode in the standardized International 10/20 System, for absolute power, relative power, mean frequency, coherence, and symmetry, as well as for covariance matrices which quantify inter-relationships among brain regions (John et al. 1988, John and Prichep 1993). These multivariate composite measures are unique to QEEG. Since cognitive processes as well as psychiatric disorders rarely engage focally localized brain activity, QEEG measures are, therefore, well suited to studies in these areas.

4.1 Independence from Cultural and Ethnic Influences

Soon after the initial demonstration of cross-cultural equivalence of normative data, normally functioning black children in Barbados were found to display the same values of the EEG power spectrum as the US and Swedish groups (Ahn et al. 1980). The independence of the normative QEEG descriptors from cultural and ethnic factors enables objective assessment of functional as well as anatomical brain integrity in persons of any age, origin, or background, repeatedly confirmed in studies from Barbados, China, Cuba, Germany, Holland, Japan, Korea, Mexico, Netherlands,

Sweden, US, and Venezuela (Ahn et al. 1980, Alvarez et al. 1987, Gasser et al. 1982, Harmony et al. 1987, Jonkman et al. 1985, Yingling et al. 1986, Diaz de Leon et al. 1988, Duffy et al. 1993, 1995, Matsuura et al. 1993, Oken and Chiappa 1988, Ritchlin et al. 1992, Valdes-Sosa et al. 1992, Veldhuizen et al. 1993).

5. *QEEG Sensitivity to Changes in Brain State*

The temporal resolution of state changes detectable by QEEG is far superior to that of any other modality except the magnetoencephalogram, or MEG. The very high short-term replicability provides excellent sensitivity to transient changes in brain state. Such changes can reflect allocation of brain resources to various cognitive, motor or perceptual tasks, or functional alterations imposed by some internal or external influences.

5.1 Detection of Brain Regions Engaged by Mental Activity

QEEG has been used extensively to detect brain regions engaged during performance of various activities. Studies performed shortly after discovery of the EEG (Berger 1937) found that during mental activity, dominant alpha waves were replaced by low voltage fast beta activity, termed 'desynchronization.' This phenomenon has been exploited in numerous studies of 'event-related desynchronization,' ERD, or event-related synchronization, ERS (Petsche 1990, Pfurtscheller 1992). Very narrow band QEEG analyses compare the power spectrum in a reference state to the power spectrum immediately after presentation of a sensory stimulus, imposition of a cognitive task, or onset of a change in the environment (e.g., hearing a musical composition). Differential ERD or ERS in various brain regions enables topographic maps to display the loci of significant state alterations correlated with the experimental procedures. The detected topography has, in some cases, been confirmed by other functional imaging techniques such as regional cerebral blood flow (rCBF). Other QEEG studies of cognitive processes have focused upon changes during cognitive tasks relative to the profile before initiating the task (Fernandez et al. 1995, Harmony et al. 1996).

One variation on this theme uses the strategy of labeling several attributes of a stimulus by presenting each aspect at a distinctive repetition or 'tracer' rate. By very narrow band analysis of the power spectrum of every recording derivation, the modulation of power at the tracer frequency in different brain regions provides a functional image of the particular brain regions engaged in the sequential steps required to process the stimulus dimensions labeled by each frequency (John et al. 1997). This technique recently has been adapted to MEG studies of binocular rivalry (Tononi and Edelman 1998).

Another productive variation has been to examine the changes in phase angles of oscillatory rhythms between brain regions, rather than changes in local power, as cognitive tasks are performed (Gevins et al. 1981). The logic of this approach is based upon the expectation that transmission from one cortical region to another will be reflected in the phase angle of some critical frequency between the interacting regions. Topographic maps based upon cortico-cortical delays among brain regions allow inferential construction of the sequence of steps in the information processing system used in a given task.

5.2 Adaptive Segmentation of Brain Fields

Evidence has been provided that the topographic fluctuation of brain voltage fields in real time is quantized into intervals during which the field is stable in one topography followed abruptly by a period of stability with a different topography, etc. These brief stable periods have been called microstates (Lehmann 1971). Computer classification methods have been applied to the field contours manifested during a prolonged observation and four prototypic field patterns have been identified (Pascual-Marqui et al. 1995). Evidence has been published that different modes of ideation characterize these different microstates (Strik and Lehmann 1993).

Normative data have been computed from large groups of healthy individuals from age 6 to 90 and show that the same four basic microstate modes exist across this representative age span, with average duration of about 80 milliseconds (Koenig et al. in press). This duration is approximately the same as that of a 'perceptual frame' (Efron 1967), which is the period during which sequential stimuli will be perceived as simultaneous. This similarity suggests that microstates may reflect the parsing of time into discontinuous brief samples which characterize subjective perception. Evidence has been provided that microstate duration is altered from this stable value in abnormal conditions such as ADHD, Alzheimer's disease, and schizophrenia (Lehmann 1994).

Using the mathematically very different approach of spatial principal component analysis (SPCA) based upon the covariance matrices across the 10/20 scalp array of electrodes, it has been found that over 90 percent of the variance of the spontaneous EEG and event related potentials can be accounted for by 4–5 SPCs in normally functioning individuals (Duffy et al. 1995, Easton and John in press). It is noteworthy that the topography of these SPCs closely corresponds to that of the prototypic microstates. These stable patterns of spatial covariance in the scalp EEG may reflect

the organization of the major functional neuroanatomical systems which represent global resources allocated as needed by the brain.

5.3 QEEG Assessment of the Depth of Anesthesia

Studying several hundred actual surgical procedures, a set of invariant reversible QEEG changes during loss and return of consciousness were found which were independent of the particular anesthetic agents used (John et al. 2001). These invariant changes were translated into multiple stepwise discriminant functions and embedded in a real-time monitor of the depth of anesthesia, called the Patient State Analyzer, or PSA (Physiometrix, Inc.). The PSA has received approval from the Food and Drug Administration (FDA) for routine use as an intraoperative monitor of the depth of anesthesia.

The successful development of a quantitative index for the level of consciousness gives evidence that QEEG is a technology with the potential of making powerful contributions to basic and applied psychological research into the relationships between brain and behavior.

6. Three-dimensional QEEG Source Localization Images

In the 1990s, mathematicians, physicists, and neurophysiologists devised numerous approaches to the problem of localizing the sources of scalp voltages within the depths of the brain. Some of these methods are now used to pinpoint neurosurgical targets in patients with intractable seizures. Perhaps the most promising approach to the inverse solution of distributed generator sources is low-resolution electromagnetic tomography (LORETA) (Pascual-Marqui et al. 1994). This method has been applied to the localization of intra-cranial epileptic foci and space-occupying lesions. Locations identified by LORETA correspond well to the actual location of the abnormalities as confirmed by subdural electrodes or during neurosurgery.

Topographic mapping of QEEG data has been extended to QEEG tomography, or QEEG-T, using variable resolution electromagnetic tomography (VARETA) (Valdes-Sosa 1997). A standard probabilistic MRI atlas in Talairach space, constructed from more than 300 normal MRIs (Evans et al. 1993), was evaluated to constrain the hypothetical generators of the distributed inverse solution to voxels composed of gray matter. For 3500 voxels, age regressed voxel-normative data were calculated across the age range 5 to 97 years over spectra ranging from 0.39 to 19 Hz in 0.39 Hz intervals. Statistical probability tomographic images (SPTI) were then obtained by Z-transformation of each voxel of the raw VARETA image at each frequency, obtained from any individual subject, relative to the corresponding voxel-normative data. Spatial statistical methods evaluate deviations from the norms of anatomical regions of interest. Results are displayed on slices from the probabilistic MRI atlas with each voxel color-coded for statistical significance. A complete description of the VARETA method and the underlying mathematical basis has very recently been published (Bosch-Bayard et al. 2001).

The accuracy of this QEEG-T method has received excellent confirmation in MRI and neurosurgical studies of space-occupying lesions (Fernandez-Bouzas et al. 1999) and is routinely in use in our laboratories to study brain dysfunctions in psychiatric and neurological patients. It has been used recently to localize invariant QEEG correlates of the loss and return of consciousness during anesthesia (John et al. 2001) QEEG-T is an economical brain imaging technology which provides very high resolution of brain electrical activity in space and time, yielding statistically interpretable images. These features make the method an attractive competitor to functional brain imaging alternatives such as PET, fMRI and MEG, which are not only more expensive (all), lack normative data (all), require substantial time for image acquisition and therefore lack temporal resolution (PET, fMRI), or impose a radiation burden which limits repeated testing for safety reasons.

7. QEEG Sensitivity to Brain Dysfunctions in Psychopathology

A voluminous literature attests to the utility of the conventional EEG in neurological and psychiatric disorders, but test-retest reliability between as well as within clinical EEG practitioners is poor. QEEG supplements the EEGer's trained eye with literally many hundreds of reliable quantitative measures which can be computed automatically and can generate precise statistical comparisons of each of these numerous measures from an individual relative to databases of age-matched normal subjects or of patients having different diagnoses. A broad consensus exists across hundreds of EEG and QEEG studies on the high incidence and particular types of abnormalities found in different psychiatric disorders.

Replicated findings of QEEG abnormalities have been reported in cerebrovascular disease, degenerative dementias, major affective disorders, mood disorders, obsessive-compulsive disorder, schizophrenia, learning disabilities, attention deficit disorder with and without hyperactivity, and mild head injury. Discriminant functions have been developed and independently replicated which accurately classify patients into DSM categories, using QEEG profiles. Using cluster analysis based upon QEEG measures, in several DSM categories subtypes have been identified

within samples of patients homogeneous by diagnosis. Subtype membership correlated well with response to treatment or evolution of illness.

Space precludes detailed descriptions here of the abnormal QEEG features distinctive for each of these disorders (see *Recovery of Function, Neural Basis of*). More complete details have been published elsewhere (Hughes and John 1999). The evidence overwhelmingly shows that QEEG assessment provides a valuable adjunct to clinical diagnosis and treatment selection in behavioral, developmental, neurological, and psychiatric disorders.

See also: Electrical Stimulation of the Brain; Electroencephalography: Clinical Applications; Epilepsy; Functional Brain Imaging; Magnetoencephalography; MRI (Magnetic Resonance Imaging) in Psychiatry; Recovery of Function, Neural Basis of; Schizophrenia; Sleep States and Somatomotor Activity

Bibliography

Ahn H, Prichep L S, John E R, Baird H, Trepetin M, Kaye H 1980 Developmental equations reflect brain dysfunction. *Science* **210**: 1259–62

Alvarez A, Valdes P, Pascual R 1987 EEG development equations confirmed for Cuban school children. *Electroencephalography and Clinical Neurophysiology* **67**(4): 330–2

Berger H A 1937 Electroencephalogram of man. *Archives of Psychiar Nervenkr* **106**: 577–84

Bosch-Bayard P, Valdes-Sosa P, Virues-Alba E, Aubert-Vazquez E, John E R, Riera-Diaz J, Trujillo-Barreto N 2001 3D EEG Tomography: Statistical parametric mapping of EEG source spectra. *Clinical Electroencephalography* **32**(2): 47–61

Curio G 2000 Ain't no rhythm fast enough: EEG bands beyond beta. *Journal of Clinical Neurophysiology* **17**(4): 339–40

Diaz de Leon A E, Harmony T, Marosi E, Becker J, Alvarez A 1988 Effect of different factors on EEG spectral parameters. *International Journal of Neuroscience* **43**: 123–31

Duffy F H, McAnulty G B, Albert M S 1993 The pattern of age-related differences in electrophysiological activity of healthy males and females. *Neurobiology Aging* **14**: 73–84

Duffy F H, Jones K J, McAnulty G B, Albert M S 1995 Spectral coherence in normal adults: unrestricted principal components analysis: relation of factors to age, gender, and neuropsychologic data. *Clinical Electroencephalography* **26**: 30–46

Easton P, John E R in press Spatial principal component analysis (SPCA) of QEEG. *Clinical Electroencephalography*

Efron E 1967 The duration of the present. *Annals New York Academy of Sciences* **138**: 713–29

Evans A C, Collins D L, Mills S R, Brown E D, Kelly R L, Peters T M 1993 3D statistical neuroanatomical models from 305 MRI volumes. *Proceedings of IEEE-Nuclear Science Symposium and Medical Imaging Conference* **95**: 1813–17

Fein G, Galin D, Johnstone J, Yingling C D, Marcus M, Kiersch M E 1983 EEG power spectrum in normal and dyslexic children. 1. Reliability during passive conditions. *Electroencephalography and Clinical Neurophysiology* **55**: 399–405

Fernandez T, Harmony T, Rodriguez M, Bernal J, Silva J, Reyes A, Marosi E 1995 EEG activation patterns during the performance of tasks involving different components of mental calculation. *Electroencephalography and Clinical Neurophysiology* **94**: 175–82

Fernandez-Bouzas A, Harmony T, Bosch J A E, Fernandez T, Valdes P, Silva J, Marosi E, Martinez-Lopez M C G 1999 Sources of abnormal EEG activity in the presence of brain lesions. *Clinical Electroencephalography* **30**(2): 46–52

Gasser T, Bacher P, Mocks J 1982 Transformation towards the normal distribution of broadband spectral parameters of the EEG. *Electroencephalography and Clinical Neurophysiology* **53**: 119–24

Gevins A S, Doyle J C, Cutillo B A, Schaffer R E, Tannehill R S 1981 Electrical potentials in human brain during cognition. *Science* **213**: 918–22

Harmony T, Alvarez A, Pascual-Marqui R 1987 EEG maturation of children with different economic and psychosocial characteristics. *International Journal of Neuroscience* **31**: 103–13

Harmony T, Fernández T, Silva J, Bernal J, Diaz-Comas L, Reyes A, Marosi E, Rodriquez M 1996 EEG delta activity: an indicator of attention to internal processing during performance of mental tasks. *International Journal of Psychophysiology* **24**: 161–71

Hughes J R, John E R 1999 Conventional and quantitative electroencephalography in psychiatry. *Journal of Neuropsychiatry and Clinical Neuroscience* **11**(2): 190–208

John E R, Ahn H, Prichep L S, Trepetin M, Brown D, Kaye H 1980 Developmental equations for the electroencephalogram. *Science* **210**: 1255–8

John E R, Easton P, Isenhart R 1997 Consciousness and cognition may be mediated by multiple independent coherent ensembles. *Consciousness and Cognition* **6**(1): 3–39

John E R, Karmel B Z, Corning W C, Easton P, Brown D, Ahn H, John M, Harmony T, Prichep L, Toro A, Gerson I, Bartlett, Thatcher R, Kaye H, Valdes P, Schwartz E 1977 Neurometrics: numerical taxonomy identifies different profiles of brain functions within groups of behaviorally similar people. *Science* **196**: 1393–1410

John E R, Prichep L S 1993 Principles of neurometrics and neurometric analysis of EEG and evoked potentials. In: Niedermeyer E, Lopes Da Silva F (eds.) *EEG: Basic Principles, Clinical Applications and Related Fields*. Williams and Wilkins, Balimore, MD, pp. 989–1003

John E R, Prichep LS, Friedman J, Easton P 1988 Neurometrics: Computer assisted differential diagnosis of brain dysfunctions. *Science* **293**: 162–9

John E R, Prichep L S, Valdes-Sosa P, Bosch J, Aubert E, Gugino L D, Kox W, Tom M, di Michele F 2001 Invariant reversible QEEG effects of anesthetics. *Consciousness and Cognition* **10**: 165–83

Jonkman E J, Poortvliet D C J, Veering M M, DeWeerd A W, John E R 1985 The use of neurometrics in the study of patients with cerebral ischemia. *Electroencephalography and Clinical Neurophysiology* **61**: 333–41

Koenig T, Prichep L S, Valdes-Sosa P, Braeker E, Lehmann D, Isenhart R, John E R in press Developmental norms for brain microstates. *Nature Neurosci*

Kondacs A, Szabo M 1999 Long-term intra-individual variability of the background EEG in normals. *Clinical Neurophysiology* **110**: 1708–16

Lehmann D 1971 Multichannel topography of human alpha EEG fields. *Electroencephalography and Clinical Neurophysiology* **31**: 439–49

Lehmann D 1994 Conscious mind state and electric brain states: the atoms of thought. *International Journal of Psychophysiology electrophysiological* **18**: 122
Llinas R R 1988 The intrinsic properties of mammalian neurons: Insights into central nervous system function. *Science* **242**: 1654–64
Lopes Da Silva F H 1996 The generation of electric and magnetic signals the brain by local networks. *Comprehensive Human Physiology* : 509–31
Matousek M, Petersen I 1973 Norms for the EEG. In: Kellaway P, Petersen I (eds.) *Automation of Clinical Electroencephalography*. Raven Press, New York, pp. 75–102
Matousek M, Volavka J, Roubicek J, Roth Z 1967 EEG frequency analysis related to age in adults. *Electroencephalography and Clinical Neurophysiology* **23**: 162–7
Matsuura M, Okubo Y, Toru M, Kojima T, He Y, Hou Y, Shen Y, Lee C 1993 A cross-national EEG study of children with emotional and behavioral problems: A WHO collaborative study in the Western Pacific region. *Biological Psychiatry* **34**: 59–65
McCormick D A 1992 Neurotransmitter actions in the thalamus and cerebral cortex and their role in neuromodulation of thalamocortical activity. *Progress in Neurobiology* **39**: 337–88
Oken B S, Chiappa K H 1988 Short-term variability in EEG frequency analysis. *Electroencephalography and Clinical Neurophysiology* **69**: 191–8
Pascual-Marqui R D, Michel C M, Lehmann D 1994a Low resolution electromagnetic tomography: A new method for localizing electrical activity in the brain. *International Journal of Psychophysiology* **18**: 49–65
Pascual-Marqui R D, Michel C M, Lehmann D, Molnar M 1995 Segmentation of brain electrical activity into microstates: model estimation and validation. *IEEE Trans Bioned Eng* **42**: 658–65
Petsche H 1990 The EEG and thinking. *Electroencephalography-EMG* **21**(4): 207–18
Pfurtscheller G 1992 Event-related synchronization (ERS): An electrophysiological correlate of cortical areas at rest. *Electroencephalography and Clinical Neurophysiology* **83**: 62–9
Ritchlin C T, Chabot R J, Alper K, Buyon J, Belmont H M, Roubey R, Abramson S B 1992 Quantitative electroencephalography: A new approach to the diagnosis of cerebral dysfunction in systemic lupus erythematosus. *Arthiritis and Rheumatism* **35**(11): 1330–42
Steriade M, Gloor P, Llinas R R, Lopes Da Silva F, Mesulam M M 1990 Basic mechanisms of cerebral rhythmic activities. *Electroencephalography and Clinical Neurophysiology* **76**: 481–508
Strik W K, Lehmann D 1993 Data determines window size and space oriented segmentation of spontaneous EEG map series. *Electroencephalography and Clinical Neurophysiology* **87**: 169–74
Swets J A 1988 Measuring the accuracy of diagnositc systems. *Science* **240**: 1285–93
Tononi G, Edelman G M 1998 Consciousness and the integration of information in the brain. *Advances in Neurology* **77**: 245–80
Valdes-Sosa P 1997 Quantitative electroencephalographic tomography. *Electroencephalography and Clinical Neurophysiology* **103**(1): 28 (abstract)
Valdes-Sosa P, Bosch J, Grave R, Hernandez J, Riera J, Pascual R, Biscay R 1992 Frequency domain models of the EEG. *Brain Topography* **4**: 309–19
Veldhuizen R J, Jonkman E J, Poortvliet J D C 1993 Sex differences in age regression parameters of healthy adults-normative data and practical implications. *Electroencephalography and Clinical Neurophysiology* **86**(6): 377–84
Yingling C D, Galin D, Fein G, Peltzman D, Davenport L 1986 Neurometrics does not detect 'pure' dyslexics. *Electroencephalography and Clinical Neurophysiology* **63**: 426–30

E. R. John

Electroencephalography: Clinical Applications

1. Introduction

The electrical activity of the brain, as manifested in the power spectrum of the electroencephalogram (EEG), is precisely regulated by complex homeostatic systems in the brain. Numerous quantitative electroencephalographic (QEEG) descriptors of spontaneous brain electrical activity can be readily extracted by simple computing algorithms. Measurements made from any individual can be objectively evaluated using statistical parameters which can be obtained from standardized, age-regressed, normative databases, and are: (a) highly stable and replicable within the individual, with few false positive findings in healthy, normally functioning persons; (b) independent of ethnic influences or cultural background of the individual; and (c) highly sensitive to many forms of developmental dysfunction, neurological disorders, or psychopathology. An overview of current understanding of the neurophysiological basis of the EEG, fundamentals of QEEG analysis, and of three-dimensional tomographic as well as topographic QEEG imaging, and a brief summary of major areas of basic research in which such methods are applicable, has been provided in a companion article in this encyclopedia (see *Magnetoencephalography*).

Invaluable assessments of the structural and functional integrity of the brain are accessible from quantitative analyses of the scalp EEG. QEEG methods provide sensitive and precise measures for study of normal as well as abnormal brain functions. Contemporary research suggests that synchronization of oscillatory activity among distant brain regions, as well as localized coherent activity, may play an important role in the processing of afferent information, particularly in 'binding' the diverse features of a complex stimuli into a perceptual whole, and in the storage of memories. Such processes are reflected in quantitative measures of the EEG power spectrum, symmetry, and coherence (synchronization) within and between the cerebral hemispheres. In most studies to date, results have been visualized as topographic maps. Validated methods for three-dimensional

QEEG tomography have very recently become available and provide statistical parametric 'virtual functional magnetic resonance imaging (MRI)' images, upon which the statistical significance of deviations relative to reference data can be encoded by colors assigned to each voxel. The importance of assessing the possible contribution of such objective neurophysiological and neuroanatomical measures to better understanding of the abnormal behaviors exhibited in psychopathology cannot be overemphasized.

QEEG methods have been used in numerous studies of patients with psychiatric disorders. This article will be restricted to QEEG applications in psychiatric patients. Due to the limitation of space, only illustrative articles can be cited here. A more comprehensive survey has recently been published (Hughes and John 1999), to which this article is indebted.

2. *QEEG Sensitivity to Brain Dysfunctions in Psychopathology*

A voluminous literature, using subjective visual pattern recognition, attests to the utility of the conventional EEG in neurological and psychiatric disorders. While the astute electroencephalographer is capable of a high level of recognition of pathophysiological patterns, test–retest reliability between as well as within clinical EEG practitioners is poor. Algorithms for computer pattern recognition, the computer's unsurpassed capacity to measure and calculate, together with the availability of normative databases across the human life span, have added an important additional dimension to electroencephalography. QEEG objectively supplements the EEGer's trained eye with literally many hundreds of reliable quantitative measures which can be computed automatically from a brief recording (John and Prichep 1993). Precise statistical comparisons of each of these numerous measures from an individual can be generated relative to a 'self norm' reference state or to mean values from databases of age-matched normal subjects or of patients with different diagnoses. A broad consensus exists on the high proportion of abnormalities found in drug-free patients with different psychiatric disorders and their distinctive nature. A brief overview of these findings follows.

3. *Cerebrovascular Disease and Multi-infarct Dementia*

In cerebrovascular disease, several QEEG parameters are highly correlated with regional blood flow or metabolism. Sensitivity and specificity are high for detection of ischemia-related diffuse or focal impairment (Nuwer 1994). These studies show sensitivity generally greater than 80 percent, false positive rates below 5–10 percent, and correlations of $r > 0.7$ between EEG and regional cerebral blood flow (rCBF) as measured using $Xenon^{133}$ or SPECT. EEG slowing (especially increased theta activity) is highly correlated with decreased rCBF (Passero et al. 1992) or diminished regional cerebral glucose metabolism.

QEEG can be quite abnormal even when the CT is still normal, such as in the first 1–3 days after a stroke or when the degree of ischemia is mild enough to cause dysfunction without infarction.

Differentiation between multi-infarct dementia (MID) and Alzheimer's dementia (AD) has been reported in several studies evaluating asymmetries (Clark et al. 1991) and coherence (Leuchter et al. 1992). Focal asymmetries and hypocoherences are more prevalent in MID.

4. *Dementias*

QEEG studies in dementia patients report increased delta and/or theta power, decreased mean frequency, decreased beta power, and decreased occipital dominant frequency. Many of these studies regard increased slow activity prior to reduction of alpha power as the earliest electrophysiological indicator appearing in Alzheimer's disease (Prichep et al. 1994, 1995). The amount of theta activity shows the best correlation with cognitive deterioration and with clinical outcome in longitudinal follow-up (Prichep et al. 1995). Increased delta appears to be a correlate of severe advanced dementia, subsequent to an increase of theta.

Multiple studies report accurate discrimination of AD patients from elderly patients with late onset depression and from normals using EEG or QEEG measures of slow activity (Robinson et al. 1994). Many of the distinctive features of AD, such as diffuse increased delta and/or theta power, decreased beta power and mean frequency, are absent in depression and are focal in multi-infarct dementia, enabling these disorders to be differentiated from AD.

5. *Learning Disorders and Attention Deficit (ADD/ADHD)*

Specific developmental learning disorders (SDLD) affect 4–6 percent of all school age children. Attention deficit disorders with or without hyperactivity (ADHD or ADD) have a prevalence of 6–9 percent in school age children. While both ADD/ADHD (see below) and SDLD are believed to be distinct neuropsychiatric entities, there is considerable comorbidity between the two disorders. Precise and accurate determination of the presence of ADD/ADHD vs. SDLD can be of critical importance in avoiding the potentially devastating impact of these disorders on children and their families. QEEG can contribute usefully to this distinction as well as in separating children with social or motivational factors underlying school problems from those with organic dysfunction.

The conventional EEG has been reported to be abnormal in 30–60 percent of children with ADD or with specific developmental learning disorder. Abnormalities reported in that literature often included diffuse slowing and decreased alpha activity. In QEEG studies, a high incidence of excess theta or decreased alpha and/or beta activity has been reported in SDLD children (Ahn et al. 1980, Lubar 1991) with theta or alpha excess often seen in children with ADD or ADHD. The types of QEEG abnormality have been shown to be related to academic performance in SDLD children (Harmony et al. 1990) and to cognitive, emotional, and behavioral problems found in a cross-national study of children (Matsuura et al. 1993).

A large percentage of children with attention deficit problems (over 90 percent) show QEEG signs of cortical dysfunction, with the majority displaying frontal theta or alpha excess, hypercoherence, theta or alpha excess, and a high incidence of abnormal interhemispheric asymmetry. Using QEEG measures, it has been possible to discriminate replicably between ADD/ADHD vs. normal children, with a specificity of 88 percent and a sensitivity of 94 percent and between ADD/SDLD children with a sensitivity of 97 percent and a specificity of 84.2 percent (Chabot et al. 1996).

6. *Mood Disorders*

The incidence of abnormal conventional EEG findings in mood disorders appears to be substantial, ranging from 20 to 40 percent (Duffy et al. 1994). The incidence of EEG abnormalities appears to be higher in (a) manic than depressed patients, (b) female than male bipolar patients, and (c) nonfamilial cases with late-age onset. Whether an 'abnormal' EEG is a necessary correlate of a clinically effective series of ECT treatment is controversial. This suggestion, analogous to assertions about the prognosis of treatments with clozapine in schizophrenia, will require further study.

Numerous QEEG studies have found increased alpha and/or theta power in a high percentage of depressed patients (Alper 1995, John et al. 1988a, 1998b). It is noteworthy that antidepressants reduce alpha activity, in contrast to the increased alpha caused by neuroleptics, suggesting that appropriate medication achieves normalization of deviant QEEG features in successful pharmacotherapy of psychiatric patients (Galderisi et al. 1994, Saletu et al. 1994, Schellenberg et al. 1994).

Interhemispheric asymmetry, especially in anterior regions, has been reported repeatedly, as has decreased coherence. In bipolar illness, in contrast to unipolar depression, alpha activity is often reduced (Knott et al. 1999, Clementz et al. 1994) and beta activity increased. This difference may serve to separate unipolar from bipolar patients presenting in a state of depression, without prior history of mania (Prichep et al. 1990).

Current treatment of bipolar disorder often includes polypharmacy, with the use of the anticonvulsant medications, carbamazepine, and sodium valproate. The successful use of these agents suggests overlap between convulsive disorders and bipolar illness.

7. *Anxiety, Panic, Obsessive Compulsive, and Eating Disorders*

The available studies suggest a high incidence of QEEG abnormalities in patients with anxiety disorders, panic disorders, and obsessive compulsive disorder (OCD). Diminished alpha activity has been found in anxiety disorder and increased theta activity has been reported in OCD. Two subtypes of OCD patients have been described. One, with increased alpha relative power, responded positively (82 percent) to serotonergic antidepressants, while the second, with increased theta relative power, failed to improve (80 percent) (Prichep et al. 1993). Epileptiform activity can occasionally be found in patients with tics (or stuttering), in addition to nonspecific diffuse slow activity. In patients with panic disorder, paroxysmal activity was four times more common than in depressed patients. Temporal lobe abnormalities, in particular, have been emphasized in QEEG studies in this type of patient.

In anorexia nervosa, abnormal background activity in the EEG can be seen in nearly 60 percent of patients, possibly related to the effect of starvation on cerebral metabolism. Paroxysmal abnormalities are seen in about 12 percent of these patients. In intractable binge eating, 'soft' neurological and EEG signs can appear. Patients with eating disorders frequently have a history of physical or sexual abuse as children, so the increase in EEG abnormalities in this group may be related to their abuse history. Alternatively, dietary and nutritional deficiencies may contribute to altered brain function. Both anticonvulsant and antidepressant drugs have been helpful in some of these patients (Grebb et al. 1984).

8. *Mild Head Injury or Postconcussion Syndrome*

Patients with complaints of cognitive memory or attentional deficit after mild head injury without loss of consciousness frequently present for neuropsychological evaluation for worker's compensation and disability benefits. Objective evidence of brain dysfunction in such cases is critical in the endeavor to separate the truly dysfunctional patient from the malingerer. QEEG evidence can play a critical role in such cases. While the absence of abnormal brain electrical activity cannot definitively exclude the possibility of brain dysfunction, the presence of abnormalities, especially those most frequently associated with unequivocal traumatic brain insult, must be considered supportive of such claims.

There is a high incidence of diffuse axonal injury in about 50 percent of the 50,000 patients/year with head injury who recover, and who characteristically do not display structural lesions on CT scan or even in MRI scans early after injury. Among those who recover after moderate head injury, 90 percent have cognitive or neuropsychological deficits (Rimel et al. 1982). Among such patients, studies involving many hundreds of cases have reported normal neurologic examinations but abnormal QEEG (Jerrett and Corsak 1988).

Numerous EEG and QEEG studies of severe head injury (Glasgow Coma Scale 4–8) and moderate injury (GCS 9–12), using samples of 50–200 patients, have agreed that increased theta and decreased alpha power and/or decreased coherence and asymmetry often characterize such patients. Changes in these measures provide the best predictors of long-term outcome (Thatcher et al. 1989).

There is a consensus that the most characteristic QEEG abnormalities persisting after mild or moderate head injury are similar in type but lesser in degree to those found after severe head injury, namely increased focal or diffuse power in the theta band, decreased alpha, low coherence, and increased asymmetry. Similar abnormalities have been reported in boxers (Ross et al. 1983) correlated with the numbers of bouts or knockouts and in professional soccer players who were 'headers' (Tysvaer et al. 1989).

The consistency of these observations, across a wide range of different degrees and types of head trauma, lend additional credibility to the multiple reports of discriminant functions based on QEEG variables which successfully separated normal individuals from patients with a history of mild to moderate head injury, years after apparent clinical recovery (Thatcher et al. 1989).

9. *Schizophrenia*

Evaluation of EEG and QEEG literature on schizophrenia is complicated by the evident heterogeneity of the illness, and the diversity of medication histories and dosage levels at the time of examination. In spite of these potential sources of difference among findings, considerable agreement nonetheless appears. Numerous EEG and QEEG studies of background activity have been performed on carefully evaluated groups of schizophrenic patients. Reviews of numerous qualitative studies indicate an incidence of abnormal conventional EEG findings in 20–60 percent of schizophrenic patients (Small 1993). With few exceptions, a substantial amount of agreement emerges from this body of literature. Alpha power is consistently reported as deficient (John et al. 1988a, 1998b), as well as altered alpha mean frequency or diminished alpha responsiveness. Numerous studies have reported increased beta activity in schizophrenia (Morihisa et al. 1983, Karson et al. 1987). It was pointed out earlier that neuroleptics typically increase alpha power (Galderisi et al. 1994, Saletu et al. 1994, Schellenberg et al. 1994) and reduce beta power (Niedermeyer 1987, Herrmann and Schaerer 1986), suggesting possible normalization of deviant features by medication.

Increased delta and/or theta activity has also been reported in a large number of studies. Such increased slow activity can apparently result from neuroleptic treatment, although there are reports of increased delta in patients off medication for several weeks (Morihisa et al. 1983) and reduction of delta or theta when medication is resumed (Galderisi et al. 1992, Saletu et al. 1994).

Among the various QEEG parameters, one which serves to discriminate schizophrenic patients from controls is the presence of increased amounts of delta activity in the left anterior temporal area. A decrease in fast alpha (10–12/sec) on the frontal areas has been called 'hypofrontality' (Gattaz et al. 1992).

An interesting observation with therapeutic implications is that there is positive correlation between the degree of QEEG abnormality and the degree of clinical improvement (Omori et al. 1992). This observation has raised the question as to whether this 'degradation' of the EEG is a necessary condition for a clinical improvement with clozapine. Clinical relationships include correlations between (a) negative symptoms and delta waves in the temporal areas (Gattaz et al. 1992); (b) blink rates, alpha power, and smoking (Klein et al. 1993); and (c) a more favorable prognosis and any EEG abnormality (Omori et al. 1992).

10. *Heterogeneity Within the Schizophrenic Population*

Results from a small number of studies are inconsistent with the consensus of a QEEG profile with increased delta or theta, decreased alpha and increased beta in schizophrenia. For example, increased slow activity has not been found by some workers, and increased alpha and decreased beta have occasionally been reported. It should also be noted that not all of the studies reporting the increased slow activity, decreased alpha, increased beta profile found the indicated deviation in all of these frequency bands.

This inconsistency might arise from the coexistence of several subtypes with different QEEG profiles within the population of schizophrenic patients, so that observations might vary depending on the fortuitous mixture of subtypes within the samples collected for a particular study. This heterogeneity has been recently documented in a study of a large sample of medicated, nonmedicated, and never-medicated schizophrenics, using cluster analysis based on QEEG variables. Five subtypes were detected, with QEEG profiles characterized by delta plus theta excess, theta excess, theta plus alpha excess with beta deficit, theta

excess and alpha excess with beta excess (John et al. 1994a, 1994b). Never-medicated patients were found in three of these subtypes. Schizophrenics with QEEG profiles corresponding to some of the groups identified by this cluster analysis have been reported to display differential responses to treatment with haloperidol (Czobor and Volavka 1991) or risperidone (Czobor and Volavka 1993). Additional evidence of heterogeneity in the schizophrenic population has been provided in QEEG studies by other groups (Roemer et al. 1991).

Findings of asymmetry in schizophrenia have been inconsistent. However, when the electrode array covered both anterior and posterior regions, power was highest over the right hemisphere in anterior but over the left hemisphere in posterior regions (Guenther and Breitling 1985, Roemer et al. 1991). This conclusion was supported by the cluster analysis just cited, in which this asymmetry pattern was found in every frequency band for all five subtypes (John et al. 1994a, 1994b).

Increased interhemispheric coherence in anterior regions has been consistently found (John et al. 1994a, 1994b, Nagase et al. 1992). Taken in conjunction with the decreased frontal coherence in depressive illness cited below, useful QEEG separation of schizophrenics from bipolar depressed patients should be attainable using measures of coherence.

11. Substance Abuse

Replicated reports have appeared of an increased beta (relative power) in alcohol dependence (John et al. 1988a, 1998b, Bauer and Hesselbrock 1993), and in children of alcoholic fathers or men at risk for alcoholism (Volavka et al. 1985). Increased alpha power, especially in anterior regions, has been reported in withdrawal, as well as after acute exposure to cannabis (Struve et al. 1994). Increased alpha, as well as decreased delta and theta, have been reported in crack cocaine users in withdrawal (Prichep et al. 1996).

12. Conclusion

The citations above document the wide applicability of QEEG as an adjunct to diagnosis, prognosis, treatment selection, and outcome evaluation in behavioral problems, developmental disorders in children, mild head injury, mood disorders, primary degenerative dementia, and schizophrenia. Familiarity with these data and with QEEG methods should enable psychologists to make major contributions to improve the management of patients in all of these categories. The EEG and Clinical Neuroscience Society (ECNS) offers training programs and board certification examinations in QEEG for psychologists, as well as for neurologists and psychiatrists. Psychologists interested in acquiring this professional qualification should contact ECNS.

See also: Anxiety and Anxiety Disorders; Attention-deficit/Hyperactivity Disorder (ADHD); Attention-deficit/Hyperactivity Disorder, Neural Basis of; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Functional Brain Imaging; MRI (Magnetic Resonance Imaging) in Psychiatry

Bibliography

Ahn H, Prichep L S, John E R, Baird H, Trepetin M, Kaye H 1980 Developmental equations reflect brain dysfunction. *Science* **210**: 1259–62

Alper K 1995 Quantitative EEG and evoked potentials in adult psychiatry. In: Pankseep J (ed.) *Advances in Biological Psychiatry*. JAI Press, Greenwich, CT, Vol. 1, pp. 65–112

Bauer L O, Hesselbrock V M 1993 EEG autonomic and subjective correlates of alcoholism. *Journal of Studies Alcohol* **54**: 577–89

Chabot R J, Merkin H, Wood L M, Davenport T L, Serfontein G 1996 Sensitivity and specificity of QEEG in children with attention deficit or specific developmental learning disorders. *Clinical EEG* **27**(1): 26–34

Clark C M, Ammann R W, Martin W W R, Ty P, Hayden M R 1991 The FDG\PET methodology for early detection of disease onset: A statistical model. *Journal of Cerebral Blood Flow and Metabolism* **11**: A96–102

Clementz B A, Sponheim S R, Iacono W G, Beiserm 1994 Resting EEG in first-episode schizophrenia patients, bipolar patients, and their first-degree relatives. *Psychophysiology* **31**: 486–94

Czobor P, Volavka J 1991 Pretreatment EEG predicts short-term response to haloperidol treatment. *Biological Psychiatry* **30**: 927–42

Czobor P, Volavka J 1993 Quantitative electroencephalogram examination of effects of risperidone in schizophrenic patients. *Journal of Clinical Psychopharmacology* **13**(5): 332–42

Duffy F H, Hughes J R, Miranda F, Bernad P, Cook P 1994 Status of quantitative EEG (qEEG) in clinical practice 1994. *Clinical EEG* **25**: R6–R22

Galderisi S, Maj M, Mucci A, Bucci P, Kemali D 1994 QEEG alpha 1 changes after a single dose of high-potency neuroleptics as a predictor of short-term response to treatment in schizophrenic patients. *Biological Psychiatry* **35**: 367–74

Galderisi S, Mucci A, Mignone M L, Maj M, Kemali D 1992 QEEG mapping in drug-free schizophrenics. Differences from healthy subjects and changes induced by haloperidol treatment. *Schizophrenia Research* **6**: 15–24

Gattaz W F, Mayer S, Ziegler P, Platz M, Gasser T 1992 Hypofrontality on topographic EEG in schizophrenia: Correlations with neuropsychological and psychopathological parameters. *European Archives of Psychiatry and Clinical Neuroscience* **241**: 328–32

Grebb J A, Yingling C D, Reus V I 1984 Electrophysiologic abnormalities in patients with eating disorders. *Comprehensive Psychiatry* **25**: 216–24

Guenther W, Breitling D 1985 Predominant sensorimotor area left hemisphere dysfunction in schizophrenia measured by brain electrical activity mapping. *Biological Psychiatry* **20**: 515–32

Harmony T, Hinojosa G, Marosi E, Becker J, Rodriguez M, Reyes A, Rocha C 1990 Correlation between EEG spectral parameters and an educational evaluation. *International Journal of Neuroscience* **54**: 147–55

Herrmann W M, Schaerer E 1986 Pharmaco-EEG: computer EEG analysis to describe the projection of drug effects on a functional cerebral level in humans. In: Lopes Da Silva F H, Storm van Leeuwen W, Remond A (eds.) *Clinical Applications of Computer Analysis of EEG and other Neurophysical Signals*. Elsevier Science, Amsterdam, pp. 385–445

Hughes J R, John E R 1999 Conventional and quantitative electroencephalography in psychiatry. *Journal of Neuropsychiatry and Clinical Neuroscience* **11**(2): 190–208

Jerrett S A, Corsak J 1988 Clinical utility of topographic EEG brain mapping. *Clinical EEG* **19**(3): 134–43

John E R, Easton P, Prichep L S, Fridman J 1994a Standardized Varimax descriptors of event related potentials: II. Evaluation on psychiatric patients. *Psychiatry Research* **55**: 13–40

John E R, Prichep L S 1993 Principles of neurometrics and neurometric analysis of EEG and evoked potentials. In: Niedermeyer E, Lopes Da Silva F (eds.) *EEG: Basic Principles, Clinical Applications and Related Fields*. Williams and Williams, Baltimore, MD, pp. 989–1003

John E R, Prichep L S, Alper K R, Mas F, Cancro R, Easton P, Sverdlov L 1994b Quantitative electrophysiological characteristics and subtyping of schizophrenia. *Biological Psychiatry* **36**: 801–26

John E R, Prichep L S, Fridman J, Easton P 1988a Neurometrics: computer assisted differential diagnosis of brain dysfunctions. *Science* **239**: 162–9

John E R, Prichep L S, Fridman J, Essig-Peppard T 1988b Neurometric classification of patients with different psychiatric disorders. In: Samson-Dollfus D (ed.) *Statistics and Topography in Quantitative EEG*. Elsevier, Paris, pp. 88–95

Karson C N, Coppola R, Morihisa J M, Weinberger D R 1987 Computed electroencephalographic activity mapping in schizophrenia. *Archives of General Psychiatry* **44**: 514–17

Klein C, Andreson B, Thom E 1993 Blinking, Alpha-brain waves and smoking in schizophrenia. *Acta Psychiatria Scandinavia* **87**: 172–8

Knott V, Bosman M, Mahoney C, Ilivitsky V, Quirt K 1999 Transdermal nicotine: Single dose effects on mood, EEG, performance and event-related potentials. *Pharmacology and Biochemical Behavior* **63**(2): 253–61

Leuchter A F, Newton T F, Cook I A, Walter D O, Rosenberg-Thompson S, Lachenbruch P A 1992 Changes in brain functional connectivity in Alzheimer-type multi-infarct dementia. *Brain* **115**: 1543–61

Lubar J F 1991 Discourse on the development of EEG diagnostics and biofeedback for attention-deficit/hyperactivity disorders. *Biofeedback and Self-Regulation* **16**: 201–25

Matsuura M, Okubo Y, Toru M, Kojima T, He Y, Hou Y, Shen Y, Lee C 1993 A cross-national EEG study of children with emotional behavioral problems: A WHO collaborative study in the Pacific region. *Biological Psychiatry* **34**: 59–65

Morihisa J M, Duffy F H, Wyatt R J 1983 Brain electrical activity mapping (BEAM) in schizophrenic patients. *Archives of General Psychiatry* **40**: 719–28

Nagase Y, Okubo Y, Matsuura M, Kojima T, Toru M 1992 EEG coherence in unmedicated schizophrenic patients: Topographical study of predominantly newer medicated cases. *Biological Psychiatry* **32**: 1028–34

Niedermeyer E 1987 EEG and clinical neurophysiology. In: Niedermeyer E, Lopes Da Silva F (eds.) *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*. Urban & Schwarzenberg, Baltimore, MD, pp. 97–117

Nuwer M R 1994 Electroencephalograms and evoked potentials: Monitoring cerebral function in the neurological intensive care unit. *Neurosurgery Clinics* **5**: 647–59

Omori M, Koshino Y, Murata T, Murata I, Horie T, Isaki K 1992 Quantitative EEG of elderly schizophrenic patients. *Japanese Journal of Psychiatry & Neurology* **46**: 681–92

Passero S, Rocchi R, Vatti G, Burgalassi L, Battistini N 1992 Quantitative EEG mapping, regional cerebral blood flow neuropsychological function in Alzheimer's disease. *Dementia* **6**(3): 148–56

Prichep L S, Alper K R, Kowalik S C, John E R, Merkin H A, Tom M, Rosenthal M S 1996 Quantitative electroencephalographic characteristics of crack cocaine dependence. *Biological Psychiatry* **40**(10): 986–93

Prichep L S, John E R, Essig-Peppard T, Alper K R 1990 Neurometric subtyping of depressive disorders. In: Cazzulli C L (ed.) *Plasticity and Morphology of the Central Nervous System*. Kluwer, Dordrecht, The Netherlands

Prichep L S, John E R, Ferris H S, Reisberg B, Almas M, Alper K R, Cancro R 1994 Quantitative EEG correlates of cognitive deterioration in the elderly. *Neurobiology Aging* **15**(1): 85–90

Prichep L S, John E R, Reisberg B, Ferris H S, Almas M 1995 Quantitative EEG (QEEG) prediction of cognitive deterioration in normal subjectively impaired elderly: A longitudinal study. *Psychiatry Research* accepted with revision

Prichep L S, Mas F, Hollander E, Liebowitz M, John E R, Almas M, DeCaria C M, Levine R H 1993 Quantitative electroencephalographic subtyping of obsessive compulsive disorder. *Psychiatry Research* **50**(1): 25–32

Rimel R W, Giordani B, Barth J T, Jane J A 1982 Moderate head injury: Completing the clinical spectrum of brain trauma. *Neurosurgery* **11**: 344–51

Robinson D J, Merskey H, Blume W T, Fry R, Williamson P C, Hachkinski V C 1994 Electroencephalography as an aid in the exclusion of Alzheimers disease. *Archives of Neurology* **51**: 280–4

Roemer R A, Shagass C, Dubin W, Jaffe R, Katz R 1991 Relationship between pretreatment EEG coherence measures and subsequent response to ECT. *Neuropsychobiology* **24**: 121–4

Ross R J, Cole M, Thompson J S, Kim K H 1983 Boxers—computed tomography, EEG, and neurological education. *Journal of the American Medical Association* **249**: 211–13

Saletu B, Kufferle B, Grunberger J, Foldes P, Topitz A, Anderer P 1994 Clinical, EEG mapping and psychometric studies in negative schizophrenia: Comparative trials with amisulpride and fluphenazine. *Neuropsychobiology* **29**: 125–35

Schellenberg R, Milch W, Schwarz A, Schober F, Dimpfel W 1994 Quantitative EEG and BPRS data following haldol-decanoate administration in schizophrenics. *International Clinical Psychopharmacology* **9**: 17–24

Small J G 1993 Psychiatric disorders and EEG. In: Niedermeyer E, Lopes Da Silva F (eds.) *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Williams & Wilkins, Baltimore, MD, pp. 581–96

Struve F A, Straumanis J J, Patrick G 1994 Persistent topographic quantitative EEG sequelae of chronic marijuana use: A replication study and initial discriminant function analysis. *Clinical EEG* **25**(2): 63–75

Thatcher R W, Walker R A, Gerson I, Geisler F H 1989 EEG discriminant analyses of mild head trauma. *EEG Clinical Neurophysiology* **73**: 94–106

Tysvaer A T, Storli O V, Bachen N I 1989 Soccer injuries to the brain. A neurologic and electroencephalographic study of former players. *Acta Neurologica Scandinavia* **80**: 151–6

Volavka J, Pollock V, Gabrielli J, Mednick S A 1985 The EEG in persons at risk for alcoholism. *Recent Developments in Alcoholism* **3**: 21–36

E. R. John

Electronic Democracy

Electronic democracy refers to the use of information technology (IT) to expedite or transform the idea and practice of democracy. The capacity of computers to store vast quantities of data, and the relative ease with which such data can be accessed, combined with many forms of communication based on computers, satellites, telephone lines, and fiber optic cables, which enable information to be transmitted almost instantaneously from place to place, has created an infrastructure which can be applied to the operation of democracy. The capacity to store and access data allows citizens to gain knowledge about what is done in their name, just as politicians can find out more about those they claim to represent. Equally, political parties and interest groups can make similar use of these data stores, as well as using the communication networks to solicit support and funds for their causes. Finally, forms of electronic communication make possible systems of political participation, whereby citizens can vote online or engage in interactive communication with those who represent them or act on their behalf.

These political uses of IT (and there are many others besides those outlined above) do not secure universal acclaim. Just as there are those who see in electronic democracy a panacea for many contemporary problems—political apathy, alienation, democratic deficits, and so on—there are others who view it as creating new problems, or exacerbating old ones, by increasing the potential for surveillance and manipulation, for increasing inequalities of power rather than diminishing them. The emergence of the Internet has greatly increased the urgency and intensity of this debate, but it did not create it. The debate about electronic democracy, and the possibilities for its realization, predate the information superhighway.

1. *The Origins of Electronic Democracy*

Electronic democracy is neither the direct product of technological advance nor of abstract thought, but rather of the engagement of the two. Although the discussion of electronic democracy came into focus in the 1980s, it can be argued that it was a feature of every change in the technology of communication, from the printing press to the World Wide Web. Each new system of communication raises questions about the distribution and content of knowledge, about who has the right to know and say what. Knowledge is a key resource in politics. Dictatorships tend to restrict and control it, just as democracies are supposed to make it freely available. So it is that with each new technology of communication new questions are raised about how these technologies should be organized and used. Should the press or broadcasting be privately owned? Who should have access to the newsprint and the airwaves? To this extent, the debate about electronic democracy is just another moment in the long history of the politics of communications.

Nonetheless, we can trace the particular concern with electronic democracy, the use of electronic means to enable citizens to participate in politics, to the late 1970s and early 1980s, when reports emerged of experiments such as 'televoting' in Hawaii (Becker and Staton 1981) and the QUBE system in Columbus, Ohio (Neustadt 1985). What both of these early initiatives represented was the use of television and telephone technology to encourage citizens to take a greater interest in politics, to reduce the costs of participation, and increase levels of knowledge. Both created the possibility of local referenda in which citizens were given information on a particular issue and then encouraged to register their views by electronic means. These initiatives were not confined to North America, and other experiments emerged in Europe, most notably the Netherlands (for a survey, see McLean 1989). Roughly parallel with these developments were changes in political campaigning, especially in the United States. Politicians began to use computer-based databases, themselves the product of interest group and consumer behavior, to organize systems of direct mailing so that particular constituencies could be targeted with customized political propaganda. By the late 1980s, it was possible to identify a large variety of examples of the use of IT in politics, and to talk of the emergence of 'teledemocracy' (Arterton 1987) and of 'the electronic commonwealth' (Abramson et al. 1988). These various uses of IT were often pragmatic and instrumental responses to particular political problems: to increase participation, to reach key voters, to raise money. What gave them a wider significance was the ways in which a number of political theorists incorporated these experiments into a case for a new notion of democracy—or perhaps more accurately to revive the fortunes of an old one. A key figure in this was Benjamin Barber whose book *Strong Democracy* (1984) used examples of electronic participation to advocate an alternative to the hegemonic dominance of liberal democracy.

There have always been positive arguments for a liberal democratic regime based on a respect for

individual preferences, limited popular participation, and the principle of representation. Just as there have been good reasons for not advocating the direct system of democracy employed in Ancient Greece (and reincarnated in the writings of Rousseau and Marx). But while there are principles which help decide between these alternatives, practical considerations have also been implicated. The utilitarian case against liberal democracy, for instance, drew attention to the impossibility of assembling large numbers of people for political discussion, while other, later theorists have drawn attention to the difficulty of collating and comprehending the information needed for decision-making in complex modern societies. To this extent, the arguments against direct democracy have been practical ones.

What Barber argued was that, with the advent of new technologies of communication, these practical objections no longer hold the force they once did. They can be overcome by enabling people to participate electronically, recreating the experience of mass participation without the practical requirement of gathering them together. For Barber, given the ability to overcome the practical problems, the case for direct participatory democracy can be made, and can be seen to trump the rival claims of liberal democracy. Such thoughts have received yet greater impetus with the emergence of the Internet.

2. *Cyberdemocracy*

Set against the speed, flexibility, and apparently ever-proliferating possibilities of the web, the early experiments in televoting seem extraordinarily crude. Now every party and almost all politicians boast their own website, from which they solicit funds and set out their stall. More importantly, there has been an explosion of websites dedicated to advocating or realizing systems of direct participation. These extend from policy discussion forums to systems of direct participation. The accompanying rhetoric is now one in which politicians talk of the 'electronic town meeting' or of the 'virtual parliament,' and where academics produce volumes on 'cyberdemocracy' (Tsagarousianou et al. 1998).

This notion of a 'cyberdemocracy' refers to a number of different developments in democratic practice. The most obvious is the opportunity that the web provides for citizens to inform themselves. They can call up websites that provide reports and data on central and local government; they can access the arguments on any number of debates and issues. It is a small step from here to the opportunity to register a vote, simply by clicking on an option just as one makes a purchase from Amazon.com, or by entering a chat room to voice an opinion. From the other side, parties now tailor their websites to elicit support and to persuade the casual surfer. Mainstream interest groups do the same, while their more activist, radical colleagues go further. Campaigns against genetically-modified foods and motorway building, together with such movements as Reclaim the Streets or those opposing the World Trade Organisation, have used the web to coordinate their activities. Meanwhile their opponents in government have increasingly used the web to re-establish their fraying legitimacy, by appearing to engage directly with their citizens and by bypassing the traditional means of communication offered by the mass media.

'Cyberdemocracy', in encompassing all these different practical forms, also displays a similar theoretical eclecticism. For some, the opportunities offered by the web allow for the emergence of new democratic forms, ones that, as with Barber's strong democracy, allow for the rediscovery of the direct tradition. The emphasis, though, is less upon participation as *control*, as taking hold of the levers of power in the name of the people, as *deliberation*: the use of the net to constitute a public sphere within which people deliberate upon the collective good. The web is seen to enable a new deliberative democracy, which sets itself against a liberal democracy concerned, according to such critics, only with realizing pre-established individual preferences. The web allows individual citizens to revise their preferences in the light of this knowledge and understanding of other people's preferences. But cyberdemocracy is not claimed just by those who oppose liberal democracy. Liberal democrats too see in the web the potential for enhancing the practices and principles that enhance their version of democracy. The opportunity of unmediated contact between citizens and parties or interest groups, the potential for online accountability, provide means for enhancing the efficiency and efficacy of liberal democracy.

To the extent that the web produces new forms of access and participation, it does represent an important moment in the constant transformation of politics. But to the extent that traditional political divides replicate themselves in the use of the Internet, electronic democracy also forms part of an established continuity. The debate about the implications of electronic democracy similarly straddles the familiar and the novel.

3. *The Debate about Electronic Democracy*

The argument about electronic democracy runs along several different dimensions. Three are singled out here. The first is that of access. Critics of electronic democracy argue that its credibility depends on some principle of universal access. Where traditional forms of political participation (visiting the polling station to vote) allow for near universal access, computer-mediated participation is exclusive rather than inclusive. The barriers to entry include limits on access

to computer technology—not everyone in the industrialized West has access to a telephone let alone a computer. Furthermore, the networks which form the web are not evenly distributed; as Manuel Castells (1996) has illustrated, the political economy of systems of communication marginalizes the less-developed world. Finally, even where technical access is possible, there are political, social, and cultural factors which intervene to thwart the capacity of some people—women, for instance—to make full use of the technology (Wajcman 1991).

Other critics of electronic democracy (Lyon 1994) focus on the technology's capacity to centralize power rather than devolve it. The network of computer databases can become the infrastructure for a system of surveillance which allows central state authorities to monitor the activities of its citizens. Such fears have become linked to discussion of the desire for, and capacity of, the state to regulate the web. Through such devices as ECHELON it has become possible for states to monitor citizens' use of email and the web. In such circumstances, the democratic capacities of the web are replaced by dictatorial ones. For skeptics of electronic democracy (e.g., Sussman 1997), the democratic prospect held out by the web is illusory.

The last dimension of the critique of electronic democracy focuses on the underlying theory, the ideas which animate the interpretation of IT as opening up a new or improved form of democracy. Insofar as electronic democracy appeals to any one version of democracy—whether direct or liberal—it relies upon the principles which sustain that particular version. Those unimpressed by such claims will be equally unimpressed by the claims made for electronic democracy. As Iain McLean (1989) points out, the division between competing notions of democracy does not derive from a lack of conceptual clarification or from technical inadequacies in existing forms of participation, but are fundamental to the very idea itself. The thought that the deliberative democracy promised by the application of electronic democracy is simply the 'best' fails to recognize the case to be made against the deliberative form (Weale 1999). In summary, there is an important debate to be had about electronic democracy, but it will not be resolved by the application of technology; it will be just another phase in the endless debate about democracy itself.

4. *The Future*

Whatever the future brings, and it will establish undreamt of forms of participation and transformative moments in the conduct of politics, it will not bring an end to the arguments around principle and ideology that fuel our notion of democracy.

See also: Age Structure; Communication: Electronic Networks and Publications; Computers and Society; Democracy; Participation: Political; Political Communication; Television: General

Bibliography

Abramson J B, Arterton F C, Orren G R 1988 *The Electronic Commonwealth: The Impact of New Media Technologies on Democratic Politics*. Basic Books, New York

Arterton C 1987 *Teledemocracy*. Sage, London

Barber B 1984 *Strong Democracy: Participatory Politics for a New Age*. University of California Press, San Francisco

Becker T, Staton C D 1981 Hawaii televote. *Political Science* **33**(1): 52–6

Castells M 1996 *The Rise of the Network Society*. Blackwell, Oxford, UK

Lyon D 1994 *The Electronic Eye*. Polity Press, Cambridge, UK

McLean I 1989 *Democracy and the New Technology*. Polity Press, Cambridge, UK

Neustadt R 1985 Electronic politics. In: Forrester T (ed.) *The Information Technology Revolution*. Blackwell, Oxford, UK, pp. 561–8

Sussman G 1997 *Communication, Technology, and Politics in the Information Age*. Sage, London

Tsagarousianou R, Tambini D, Bryan C 1998 *Cyberdemocracy: Technology, Cities and Civic Networks*. Routledge, London

Wajcman J 1991 *Feminism Confronts Technology*. Polity Press, Cambridge, UK

Weale A 1999 *Democracy*. Macmillan, London

J. Street

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Electronic Economy: Governance Issues

Stories about the digital economy, the E-conomy, have moved from specialty journals to the business pages and from there to the front pages of mainstream newspapers. The issues have evolved from the narrowly technical discussions of how best to configure a data network and the narrowly legal issues of patents and copyrights into significant policy battles over taxes, privacy, and intellectual property that have significance for the character of the marketplace and the polity. At the same time, the discussion has evolved from contrasting futurist presentations, a new Jeffersonian era of frontiersmen freed from government constraint set against a neo-Orwellian technology enabled 'big brother,' into a more sophisticated and pragmatic debate about issues such as surveillance and control, private and public (Bar et al. 2000). The policy issues raised in the Internet economy will not be easily fixed by simple technical tinkering, adding new digitally sensitive wording to old rules. Rather, they portend fundamental struggles about the nature of the polity and the economy. The resolutions will not be a matter of policy management, but of political choices about basic values.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

At a minimum the collection of networks, computers, and information tools that emerged in the last decade of the twentieth century represented a new leading economic sector. Since the first industrial revolution there has been a sequence of such leading sectors, each bringing its 'explosion of invention and innovation that revolutionizes productivity in a narrow slice of the economy' (Cohen et al. 2000). But is there more? A case can be made that the emergence of an E-conomy represents the creation of a new set of general tools that is fundamentally transforming *each sector of the economy*. *This perspective sees a shift in* the very terms and dynamics of market competition. Stephen Cohen, Brad DeLong, and John Zysman have suggested that the new technologies 'do not just amplify productivity in one sector but give all economic sectors new "tools." They are tools for thought that amplify brainpower in the way the technologies of the Industrial Revolution amplified muscle power' (Cohen et al. 2000).

Two processes drive this revolution. Certainly *one process* is technological advance. The conventional measure of this advance throughout the 1980s and the first half of the 1990s was advance in semiconductor technology with the successive iterations of the memory chip or microprocessor reflecting the amount of information that could be stored or processed (Moore 1965). Now, particularly since 1997–98, it is tracked by the expanding network bandwidth that measures the amount of information that can be transmitted. The two technological advances together create a vector of information capacity.

The *second process* is one of social innovations and reorganizations. Together those innovations have constructed the networked business system, which includes digitally enabled reductions in transaction costs, but more importantly involves quite fundamental innovations. The Silicon Valley system, which rests on repackaging and redirecting people and technology and backing them with risk capital, provided new answers to real problems of industrial innovation and production (Saxenian 1994, Cohen and Fields 1999).

No revolution happens without substantial dislocation and uncertainty over the meaning of measures and indicators. Critically, though, the structural transformation of the economy we are watching needs to be separated from any debates about the valuations of stock prices (Weber 1997). The debates around these points are maturing toward more substantive questions, such as how does the digitalization of the economy change, if at all, the major causal forces of the macroeconomy.

For now, the Internet economy, the intersection of the technological advances of processing and bandwidth with the business process innovations, rests principally on business applications and business reorganization. But as broadband networks capable of delivering streaming media and other new applications reach the home, consumer applications that are intimately entangled with social and community evolutions will become the drivers that induce the ongoing sequence of technical innovation. Will this be viewed as another mass media such as television and radio in the twentieth century, as another communications instrument such as telephone and telegraph in the nineteenth century, or as with the book in the fifteenth century which became the base of the Renaissance? (Eisenstein 1980).

The dynamic, all encompassing nature of this transformation creates pitfalls for any brief discussion or any research strategy. First, traditional analyses or research strategies focus on particular threads, themes, and issues. They do so with good reason. Understanding relationships, specifying causal links, is difficult enough when dealing with limited, specific and fairly conventional arguments and issues. Weaving a broad pattern of relationships all too often produces not great insight but superficiality. Yet the more profound the transformation, the more difficult such a traditional strategy becomes. As context itself changes, what were once parameters become variables. Research problems cannot be so easily isolated nor causal relationships cleanly specified.

The issues, policy and political, present themselves in four packages. Each package is required to grapple with the whole; each constitutes in itself a set of social science puzzles. The four packages are: (a) networks and tools; (b) e-business and e-society; (c) the productivity dilemma resolved; and (d) governance and politics.

1. Networks and Tools—the Infrastructure of the E-conomy

The paraphernalia of this new Internet economy rests on advanced data networks and is facilitated by new software tools. The transformation of network communications underpins and expresses the evolving e-commerce and e-society story. The new Internet-based economy grew extraordinarily quickly in comparison with earlier communications and transportation technologies. It took telephone technology 38 years to reach a penetration level of 30 percent of US households. Television was faster, reaching 30 percent of households after about 17 years. The Internet, however, surpasses any previous communications technology in diffusion speed by accomplishing the same penetration in less than seven years (United States Internet Council 1999).

One of the main reasons for this rapid diffusion is that the new technological system, the Internet/web system for linking and addressing computers, could be initially deployed over the older, existing infrastructure for voice communication, telephones. This technologically-driven transformation interacted with and

fueled a move away from the older voice infrastructure, plain old telephones (POTS) over a switched analog, then later switched digital network, in a public utility, or regulated monopoly, model, and toward competing public providers, private data networks, and 'networks of networks' linked by routing data.

The deployment of the new infrastructures to date has demonstrated three features that are critical to understanding this transformation. *First*, the American network revolution has been user driven, that is, the products and architectures have been defined by large sophisticated users rather than by network providers. American data networks were defined, configured, and often controlled, even owned, by the users (Bar and Borrus 1994). The dominant American provider, AT&T, did not simply define the network rules to which customers adapted. It was competitive experimentation among user-defined networks that unlocked an avalanche of innovation. A primary difference between the European networks such as Minitel and the Internet is just that—the Internet is user controlled and developed. The Internet system that permitted and facilitated the new e-business strategies is now in turn being shaped by the emergence of e-business—its latest user constituency.

Second, the infrastructure of the E-conomy, the innovations and the evolving networks themselves have been developed and deployed by private companies in competitive markets. There has been intense and growing competition both among network providers—AT&T, Sprint, Deutsche Telekom, France Telecom, and the like—as well as among the equipment suppliers to these networks, companies such as Lucent, Nortel Networks, Alcatel, and Ericcsson, and newly established firms from Silicon Valley, such as Cisco and Ascend, and newly important firms from other regions, such as Nokia of Finland. Because this competition rages both at the network levels, equipment level, and the software to operate the equipment, the battleground is very complicated.

That private competition was driven by two public policy initiatives: (a) the push to deregulate telecommunications infrastructure, including the divestiture of AT&T, and (b) a series of investments and policy decisions by DARPA, NSF, and other research agencies that nurtured packet-switching technology into an open infrastructure that could be freely exploited and commercialized in many different ways (Naughton 1999, Hafner and Lyon 1996, Rosenzweig 1998).

Third, the network revolution may have been launched in the United States. However, the next generation of advanced networking, particularly wireless, will include network configurations and technologies that are being developed outside the United States. Certainly, American companies such as Qualcomm and Motorola are major players in this wireless world, but advanced wireless networking, for example, is presently defined in European markets. The wireless revolution has been a story of Europe's conscious and effective use of standard setting as a development tool (Bach 2000).

In sum, the network revolution was generated by user-defined needs, private provision, and intense and increasingly international competition. In addition to accelerating innovation, the legacy of user-control has been enshrined in the culture and to a large extent in the technology, and is now becoming an issue of policy debate (Bar et al. 2000). Legislation and indeed the culture that surrounds the Internet reinforce each other and make the social structure around the Internet quite robust.

2. *E-business and E-society*

Our society may ultimately be transformed by digital tools for thought and the data networks that connect the now omnipresent computers. It is important to remember that we are still in the earliest days of technological innovation, with only the first glimmers of the commercial and social implications to come. Although there is much speculation on the subject, the question of how new communities will be shaped by the new technological possibilities is, inherently, unknowable—there are simply too many degrees of freedom (Smith and Kollock 1999, Turkle 1995). The telephone was proposed to be a means of listening to symphonies, before switched signals made it a tool for social and business communication. We are still in a phase of rapid innovation where unforeseen technological change may undermine even the most far-seeing predictions about the social implications of communication.

What is clear is that the transformation, at least for the moment, is being driven by business applications. Business to business (B2B) and business to consumer (B2C) projects have induced the development of a new data communications infrastructure. If significant changes in the broad dynamics of the E-conomy emerge in the first half of the decade, it appears that they will begin with the transformation of business. This transformation goes far beyond the introduction of computers into the workplace: as Erik Brynjolssen and Lauren Hitt point out, for every dollar spent on computers or other information technology equipment, businesses spend about 10 dollars to reorganize their information, production, and social systems in order to use the new technology most efficiently (Brynjolssen and Hitt 2001). The real story is thus not about technology as such, but in the way technology enables and induces reorganization.

3. *The Productivity Dilemma Resolved*

It is this complex reorganization that impacts levels of productivity. And it is productivity, not technology

per se, that ultimately determines what people are paid, the types of skills and training required, and hence who can work and what schools must be like. Firms are the unit that will adjust to the new productivity dynamics and terms of competition and it is consequently on the level of the firm that we will first see the transformative capacity of the new digital technologies. The story of business transformation is, like the story of the network infrastructure on which it rests, central to understanding the community, policy, and governance issues. The phenomenon of e-commerce, or more broadly e-business, is unfolding extraordinarily rapidly. The conventional wisdom about its course and consequences shifts with equal rapidity. Because agreed upon 'facts' and the terms of discussion change so quickly, it is essential, but difficult, to establish a stable framework for observing the emerging transformation of business strategy and practice.

How will e-commerce and digital networks change the organization of economic activity and market competition? We are, as noted, in an era of experimentation with the ultimate outcomes emerging both from a series of marketplace struggles that will change how sellers and buyers interact and policy struggles that will develop around questions of regulating and taxing those processes. A series of rapidly changing codewords point at these rival experiments, for example, disintermediation, the elimination of actors who control information or goods in a value chain (a *filiere*, if you will), has given way to the notion of reintermediation, the introduction of new kinds of players (Shapiro and Varian 1999).

The flux that one sees in the language represents not just fad and fashion but real, fundamental uncertainty. The rapid churn of venture capital funding is not a surprise—it represents a series of bets in a setting where efficiency and competitiveness are hard or impossible to predict because the marketplaces in which those characteristics will emerge are still in flux. Theories of industry organization and competition policy already are shifting their emphasis and focus from issues of scale to scope, from the economies of production to the economics of networks and standards (Shapiro and Varian 1999). But these categories can only take us so far. The marketplaces themselves are being altered fundamentally as decentralized auctions (such as eBay) and reverse auctions with buyers setting the terms (such as priceline.com), change the ways prices are set and goods are sold. Significant parts of the markets for information products (from finance to media) have moved entirely online—and the limits to that move have not yet been seen (Negroponte 1995).

The transition can be monitored by looking not at one but all four of the following processes: (a) competing business strategies, and the tools on which those strategies rest; (b) evolving market structures, boundaries of the firm that contribute to market structure, and the parameters that shape the types of firms that will win; (c) changes in how the marketplace works as e-tools change the ways that buyers and sellers interact; and (d) the creation of cyber-market places in sectors such as media and finance and perhaps others where the entire industry moves its goods and interchanges onto networks.

Studying this transformation of markets by the tools of E-commerce reopens a series of traditional social science problems. Let us consider two. Put simply, will there be one E-conomy or several? Will the new economics of communication and information finally sweep away national boundaries creating the long awaited but never quite arriving convergence of national systems under an electronic manifestation of globalization? This question recapitulates older debates over whether there is one capitalism or several (Hilferding 1923, Gerschenkron 1962, Shonfield 1965, Katzenstein 1978, Zysman 1983, Hall 1986, Soskice 1999). A single E-conomy view builds on the common understanding of the world as composed of technological leaders and followers. No matter which country or region leads the world market, the powerful communications instruments of the digital technology will produce a single global market and similar business models under the governance of convergent, presumably neo-liberal regulatory regimes. Other forms of regulation and protectionism will either be circumvented, swept aside, or else doom the offending jurisdiction to technological backwardness.

The 'techno-determinist' view is partially right in suggesting that new capabilities erode or at least alter inherited market and regulatory structures (Castells 1997). However, this initial story is much too simple. A second view is that globalization becomes a matter of playing out on a large stage nationally rooted innovations. Historical perspective is important here. The first chapter in the e-globalization story was Japanese production innovation (Womack et al. 1991). A second chapter consisted of American technological revival built around innovation in computing, networks, and contract manufacturing to neutralize the Japanese advantage in production (Borrus and Zysman 1997). A third chapter, the Internet story, was launched by and rooted in American technology and deregulation.

Leading-edge users of the Internet and e-commerce are beginning to emerge in Europe and Asia as well as the United States. As new e-commerce technologies, usage patterns, business models, and legal frameworks develop outside the US, they will challenge the early dominance of American policy and market leadership. Regionally- and nationally-based lead users, with regionally specific market needs and operating in distinct policy environments, will drive distinctive lines of technology development in which local firms may have competitive advantages.

The distinction between these two perspectives is ultimately an empirical one. Convergence can be

measured over time, if not precisely, and divergent paths of development will be visible in familiar forms: distinct technologies and applications, differentiated market structures and business models, different structures of comparative advantage, and ultimately distinct policy and legal frameworks.

The second problem is that the manner in which technology is embedded by law and practice into society shapes the trajectory of technical evolution. This is dramatically clear in the case of open source software. Instead of retaining private ownership of the source code (the list of instructions in a programming language that comprise the basic recipe for a piece of software), open source software projects such as Linux, Apache, and others distribute source code freely, permit modification by users, and allow redistribution of modified versions without royalties or licensing fees to the author. Open source inverts the standard logic of intellectual property rights. It is built collaboratively by a geographically dispersed community who work in a highly parallel, relatively unstructured way and without direct monetary compensation. Because open source software is nonrival, nonexcludable, and subject to collective creation, it looks like an 'impossible public good' from standard perspectives (Smith and Kollock 1999). But in fact it is successful both technically (producing software that is as good or better, technically, than proprietary software) and in a social sense (sustaining a community of developers whose commitment to the collective project is extraordinary). The political economy of open source is ultimately a story about both economics and culture—but an economic and cultural logic that have each been modified, in specific ways, by the Internet (Weber 2000). The question of how the model scales and whether it can be extended to other areas of knowledge production generates dramatic arguments about business models, which are ultimately arguments about the structure of economic production in a digitizing economy and society.

4. *Governing, or at Least Steering, the E-conomy*

All market economies rest on rules (Polanyi 1944). Just as there were never separate rules for a fax economy or a telephone economy, there will not be separate rules for an Internet economy—rather, the decisions we make about the E-conomy, rules both for the networks of information technologies that are defining new market relationships and for the new business system that has helped generate it, will be critically important for the 'traditional' economy. We are beyond the point where simple tinkering will suffice. We are also beyond the point where we can sustain the illusion that the Internet can exist apart from and independent of the rest of the economy and society. The cyber world is intertwined with, not independent of, the 'traditional' world. Libertarian fantasies about the Internet economy have been left behind. Regulation in this new era ultimately is a story about how political decisions interface with technical possibilities. Lawrence Lessig's recent work differentiates between 'West Coast Code,' the code that defines how the networks and the programs work in a technical sense, and 'East Coast Code,' the laws and regulations drawn up by policy-makers (Lessig 1999, Rheingold 1993, Reid 1995). The key debates will ultimately be about what regulation can and should do to modify the technical possibilities that shape what kind of network world we have, what kind of code-constructed realities and businesses we develop, and hence what kind of economy we build.

The process of translating values for a new era, or creating new rules that encompass new values, is intertwined with qualitative changes in business and social life that reopen established policy bargains. The choice of rules will often have major outcomes in terms of private gain. It is hard to imagine a higher set of stakes as a new governance 'system' is simultaneously being created at home, domestically, and as a bargain amongst several national systems. Framing the initial policy debates is difficult enough. Perhaps because it was carried out mainly among technologists, the early debate was punctuated by libertarian conceptions that government and perhaps even traditional forms of political contestation somehow could be excluded from something loosely called cyberspace (Barlow 1996). There is now broad recognition that technical choices about the marketplace touch the character of our communities and polities in fundamental ways (Lessig 1999). And that, conversely, the choices we seek to make about community in an epoch of data networks in turn shape the marketplaces (Post 1995). The resulting political debate is not simply one of advantage and interest, but rather it concerns fundamental values and basic choices about markets, community, and democracy. These kinds of debates recast political alliances and reshape the character of the polity as actors redefine their interests in the New Economy (Litan and Swire 1998).

There is a set of profound issues already on the agenda today: competition policy, taxation, intellectual property rights, and privacy. They open debates about what kind of E-conomy, about what rules for Internet facilitated business reorganization, about what sorts of virtual communities with what rights of speech and anonymity, about what types of network arrangements, about which architectures of code are called for. For example, basic issues of competition policy are now being fought out over the fate of Microsoft (US vs Microsoft Corporation: Conclusions of Law and Final Order 2000, Bork 2000). More broadly, how the network and other key elements of infrastructure (such as operating systems, as in the case of Microsoft) are regulated will shape who can participate in the new E-conomy, and on what terms. It will powerfully influence national patterns of e-

commerce, both the speed of the transformation and the models which are adopted, which is why decisions about Microsoft are ultimately much more important than simply deciding the future of that one company.

A second issue set touches on more traditional questions of governance and the state, such as taxation. Though the numbers are still small, there is little doubt that a few years hence quite significant portions of commerce will take place over electronic networks (Shop.org/Boston Consulting Group 1999). The question of whether and how to tax the Internet will become ever more a debate about how and where to tax overall, rather than a sideshow. The intensity and urgency of the new debate will depend directly on the influence that the shifting locus of transaction has on government revenue streams and their capacity to deliver services to their community.

But the most basic issues are about information—who owns it, and what to do with it. At a minimum, these are questions about intellectual property, privacy, and speech. Here, new technologies will stretch familiar debates out into new territories of political values and the political theories that lie behind those values. Privacy is one dramatic example. Extraordinary amounts can be known about us by monitoring our activity in a computer-based economy. The bank or credit card company knows what we buy and may monitor spending patterns for risks of default, such as possible divorce suggested by unexpected patterns of flower or jewelry purchase. The grocery store knows what you eat. An Intelligent Transportation System that manages traffic flows may know where you are, where you go, and when. How may the data be used? Can it be combined? Or sold to an insurance company? Or provided to the IRS?

A new relationship between privacy vis-à-vis the government, and privacy vis-à-vis commercial interests is emerging. Is the law in the United States—originally crafted to protect the citizen against the state and state actions such as wiretap—appropriate in an era in which privately controlled data may provide a precise picture of our life, far more detailed than anything the government can know? What rights should government have to access this privately collected data? One proposed policy approach would provide property rights in data about ourselves that we as individuals can sell or withhold. But that property right, it appears, may not include genetic data about us obtained by physicians or surgeons in medical practice. A second approach would propose that privacy is a function of the community's needs, that there must be limits on the personal data that can be gathered or dispensed for the sake of security of the community. The questions of privacy blur quickly into questions of security of communication and transaction, and in turn to a balance between privacy and community safety and protection. The complexity of the values involved in these kinds of decisions ensures that the debates will be tortuous.

What issues such as privacy, intellectual property, free speech, consumer protection, or taxation have in common is that the debate about the kind of communities we would prefer becomes entangled with the way we will run our new electronic marketplaces. Consider the question of security and encryption. Once again the balance between personal privacy and national security as well as police purposes has to be reset. Similarly the question of jurisdiction, which political entity is responsible, for matters such as taxation and consumer protection have to be rethought. Consider the balance between free speech and the protection of minors against inappropriate content. Should the solutions be technical—filtering devices that block access to what parents would reject? Or should we instead seek solutions that require public rules? Some issues simply will be forced onto the political agenda as network-based transactions multiply. Whatever your views on taxation on the net, simply speculate how the debate would work if a majority of commercial transactions were on the net, but untaxed. Public services from roads to schools would not end; but the structure of taxation would be forced to change in a profound manner.

Complicating each of these debates tremendously is that they cannot be settled definitively in one country. If privacy rules are different in Europe and the United States, how do companies from AOL to IBM operate? (Litan and Swire 1998). While particular issues are thought out and fought out, care must be taken to assure that diverse national solutions are sufficiently reconciled to assure the operations of the global information system. An era in which national monopoly telephone systems could be imagined as closed almost watertight systems provisioned by a national cartel of suppliers and connected at the borders or halfway across an ocean has given way to the era of interconnected data systems, global service providers, and global equipment companies. Connection at the borders of separate systems will be difficult; more likely rules may be harmonized, that is made identical, or made interoperable. But a solution will have to be found.

The evolution of the E-conomy is a remarkable opportunity for the further development of empirical and theoretical social science. This is not a separate research domain for a few observers interested in technological evolution and the politics of technological change. Information technology is penetrating every aspect of business, the economy and society as a whole. Just as expressions such as 'e-commerce' and 'Internet economy' will be meaningless in light of the large share of commercial transactions and economic activity carried out over digital networks, differentiating between research on 'Political Economy' and 'Political E-conomy' is likely to become pointless. Understanding the way technological, organizational, and political-legal changes interact in shaping a future characterized by ubiquitous digital networks connect-

ing countless nodes globally will be essential for our understanding of political economy, whether domestic, comparative, or international, and for our understanding of society as a whole.

Keeping track of these changes will be difficult, to say the very least. In the short run, bargains will most likely be struck by existing coalitions or new coalitions composed of familiar interest groups in an existing institutional environment. But in the long run, the political sociology of communities of all kinds will change dramatically. Individuals will come to define their interests differently, new interest groups will emerge, new coalitions will rise, and new bargains will be struck and embodied in entirely new or fundamentally altered institutions. Indeed, the changes underway have the potential to radically change the very configuration and orchestration of social and political life. And we are only starting to study the very beginning of this transformation.

Bibliography

Bach D 2000 International cooperation and the logic of networks: Europe and the global system for mobile communications (GSM). *Berkeley Roundtable on the International Economy* (BRIE), Working Paper 139, Berkeley, CA

Bar F, Borrus M 1994 The future of networking in the US. *E-conomy Project*, Working Paper 11, Berkeley, CA

Bar F, Cohen S, Cowhey P, DeLong B, Kleeman M, Zysman J 2000 Access and innovation policy for the third-generation internet. *Telecommunications Policy* **24**(6/7) July/August: 489–578

Barlow J P 1996 A cyberspace independence declaration. http://www.eff.org/pub/Misc/Publications/JohnPerryBarlow/ barlow0296.declaration, downloaded May 11, 2001

Bork R 2000 The case against Microsoft. http://www.procompetition.org/research/bork.html, downloaded May 11, 2001

Borrus M, Zysman J 1997 Globalization with borders: the rise of wintelism as the future of global competition. *Industry and Innovation* **4**(2) December: 141–66

Brynjolssen H, Hitt L 2001 Beyond Computation. *Journal of Economic Perspectives*. **14**(4): 23–48

Castells M 1997 *The Power of Identity: The Information Age: Economy, Society, and Culture*. Blackwell, Oxford, UK, Vol. 2

Cohen S, DeLong B, Zysman J 2000 Tools for thought: What is new and important about the 'e-conomy'? *Berkeley Roundtable on the International Economy* (BRIE), Working Paper 138, Berkeley, CA

Cohen S, Fields G 1999 Social capital and capital gains, or virtual bowling in Silicon Valley. *California Management Review* **41**(2): 108–30

Eisenstein E 1980 *The Printing Press as an Agent of Change*. Cambridge University Press, Cambridge, MA

Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Harvard University Press, Cambridge, MA

Hafner K, Lyon M 1996 *Where Wizards Stay Up Late: The Origins of the Internet*. Simon and Schuster, New York

Hall P 1986 *Governing the Economy*. Polity, Cambridge, MA

Hilferding R 1923 *Das Finanzkapital: eine Studie ueber die juengste Entwicklung des Kapitalismus*. Verlag der Wiener Volksbuchhandlung, Vienna, Austria

Katzenstein P (ed.) 1978 *Between Power and Plenty: Foreign Economic Policies of Advanced Industrial States*. University of Wisconsin Press, Madison, WI

Litan R, Swire P 1998 *None of Your Business: World Data Flows, Electronic Commerce, and The European Privacy Directive*. Brookings Institution Press, Washington, DC

Lessig L 1999 *Code and Other Laws of Cyberspace*. Basic Books, New York

Moore G E 1965 Cramming more components onto integrated circuits. *Electronic* **38**(8)

Naughton J 1999 *A Brief History of the Future: the Origins and Destiny of the Internet*. Weidenfeld and Nicolson, London

Negroponte N 1995 *Being Digital*. Alfred A. Knopf, New York

Polanyi K 1944 *The Great Transformation*. Beacon Press, Boston

Post R 1995 *Constitutional Domains: Democracy, Community, Management*. Harvard University Press, Cambridge, MA

Reid E 1995 Virtual worlds: culture and imagination. In: Jones S (ed.) *CyberSociety: Computer-mediated Communication and Community*. Sage, Thousand Oaks, CA

Rheingold H 1993 *The Virtual Community: Homesteading on the Electronic Frontier*. Addison-Wesley, Reading, PA

Rosenzweig R 1998 Wizards, bureaucrats, warriors, and hackers: writing the history of the internet. *American Historical Review* **December**: 1530–52

Saxenian A 1994 *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press, Cambridge, MA

Shapiro C, Varian H 1999 *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, Boston, MA

Shonfield A 1965 *Modern Capitalism: The Changing Balance of Public and Private Power*. Oxford University Press, New York and London

Shop.org/Boston Consulting Group 1999 *The State of Online Retailing 3.0*. http://www.shop.org, downloaded May 11, 2001

Smith M, Kollock P (eds.) 1999 *Communities in Cyberspace*. Routledge, New York

Soskice D 1999 Divergent production regimes: coordinated and uncoordinated market economies in the 1980s and 1990s. In: Kitschelt H, Lange P, Marks G, Stephens J D (eds.) *Continuity and Change in Contemporary Capitalism*. Cambridge University Press, Cambridge, MA

Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon and Schuster, New York

United States Internet Council 1999 *State of the Internet*. http://www.usic.org, downloaded May 11, 2001

US vs. Microsoft Corporation: Conclusions of Law and Final Order April 3, 2000. http://www.usdoj.gov/atr/cases/ms_index.htm, downloaded May 11, 2001

Weber S 1997 The end of the business cycle? *Foreign Affairs* **July/August**: 65–82

Weber S 2000 The political economy of open source. *Berkeley Roundtable on the International Economy* (BRIE), Working Paper 140, Berkeley, CA

Womack J, Jones D, Roos D 1991 *The Machine that Changed the World*. Harper Perennial, NY

Zysman J 1983 *Governments, Markets and Growth*. Cornell University Press, Ithaca, NY

J. Zysman and S. Weber

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Eliade, Mircea (1907–86)

Mircea Eliade was the best-known contemporary theorist of religion from the discipline of religious studies. He was one of a handful of scholars within the discipline who influenced scholars outside it. Unlike specialists in a single religion, he was influential as a theorist of religion as a whole. While he wrote specialized tracts on Australian religions, Indian religions, alchemy, yoga, shamanism, and the religion of his native Romania, he was most celebrated as one of the last comparativists and, even more, as the latest defender of the defining tradition of religious studies. Eliade sought to demonstrate the universality of religion by finding it in even seemingly secular domains. He also wrote fiction, and in Romania was a popular novelist and short story writer.

1. *Life*

Mircea Eliade was born in Bucharest on March 9, 1907. He began writing for Romanian weeklies as a teenager. In 1925, after completing the *lycée*, he entered the University of Bucharest and received an MA for his thesis on Italian philosophy from Ficino to Bruno. His mentor was the philosopher Nae Ionescu. While at university, he continued writing regularly—short stories, scholarly articles, reviews, and translations. He also founded a journal.

In 1928 Eliade went to India to study Sanskrit and Indian philosophy with Surendranath Dasgupta of the University of Calcutta. He aspired to become sufficiently trained linguistically to secure a professorship in comparative religion at the University of Bucharest—an especially ambitious goal since no chair in the field existed at the University at the time. Daring to hope that one would be created for him, he immediately immersed himself in Sanskrit grammar. He also discovered the world-affirming variety of yoga known as Tantrism. This outlook came to express Eliade's view of religion universally: that the sacred is to be found within the profane rather than outside it.

Impressed by Eliade's diligence and enthusiasm, Dasgupta soon took him into his household and treated him almost as an adopted son. When Dasgupta discovered that Eliade was having an affair with his 16-year-old daughter, Maitreyi, under his own roof, he threw Eliade out of his house, and the two parted ways forever. Undeterred, Eliade proceeded to spend six months in the ashram of Rishiskesh, learning the practice, not merely the philosophy, of yoga. Even during his time in India, which constituted his fieldwork, he recognized that his heart lay in comparative religion rather than in any single religion. While in India, he continued writing, turning out newspaper columns and magazine articles about his travels, scholarly articles on yoga, and several novels. After three years in India, he returned to Romania, pressured by his father to undertake his already deferred year of military service. He intended to return to India but never did.

Back home, Eliade worked on his doctoral thesis on yoga, which he had begun in India, and simultaneously published articles, wrote novels, worked for journals, gave radio talks, and organized public lectures—all on a host of topics. He came to be considered the leading cultural figure of his generation of young Romanians. In 1932 he did his year of military service, largely doing translations. In 1933 he received his doctorate. In the same year he published the novel *Maitreyi*, about his love affair with Dasgupta's daughter. The book won a prize and became a best seller. It was the Romanian literary sensation of 1933, and Eliade became famous.

At the end of 1933 Eliade began teaching at the University of Bucharest. He had hoped to be offered a lectureship but had to settle for serving as an assistant to Ionescu. He held the post for five years, lecturing on various topics in both religion and philosophy. Ionescu, whom he continued to idolize, was the intellectual mentor of the Iron Guard, the Romanian counterpart to the Nazis, and wrote diatribes against Jews. Proclaiming that Jews deserved to suffer because they had rejected Jesus as their savior, he demanded that Romania rid itself of all Jews by any possible means. In his autobiography Eliade self-servingly emphasizes his own break with Ionescu over the status of Jews, but recent research has uncovered scores of viciously anti-Semitic newspaper articles written by Eliade himself (see Berger 1989; see, in reply, Rennie 1996, Chap. 13). Eliade associated Jews with everything foreign to true Romanian culture. In stock anti-Semitic fashion, Jews were blamed for communism, materialism, democracy, and modernity. Eliade's official biographer, M L Ricketts, uncritically accepts Eliade's own presentation of his position (see Ricketts 1988, II pp. 903–12). According to both, the fervently Romanian nationalism that Eliade espoused was spiritual rather than political. The Government nevertheless arrested and detained Eliade in 1938 because of his zealous support for the Iron Guard. Eliade was subsequently transferred to a sanitarium.

In 1940 Eliade was allowed to leave the country to serve in the Roman Legation in London, and in 1941 he was transferred to the Legation in Lisbon, where he remained until the end of World War II. He returned only once to Romania. Had he returned after the war, he would have faced arrest as a collaborator. He left for Paris in 1945, where he spent 10 years writing the scholarly tomes that made him internationally famous. In 1946 he served as a visiting professor at the *Ecole des Hautes Etudes* of the Sorbonne. In 1956 he was a guest lecturer at the Divinity School of the University of Chicago, and the next year he became a permanent Professor of the History of Religions. In 1962 he was named the Sewell S. Avery Distinguished

Service Professor. He co-founded the journal *History of Religions*. He became the best-known scholar of religious studies in the world and the leading advocate of the field. He retired in 1983, at the age of 76, and was honored with the establishment of an endowed chair in his name. He died in Chicago on 22 April 1986. Shortly afterwards, there appeared the culminating opus of his career: the 16-volume *Encyclopedia of Religion*, edited by him.

2. *Theory of Religion*

The tradition perpetuated by Eliade goes back at least to the nineteenth-century comparativists Friedrich Max Müller (1823–1900) and C. P. Tiele (1830–1902) and was continued by such figures as P. D. Chantepie de las Saussaye, Nathan Söderblom, Gerardus van der Leeuw, W. Brede Kristensen, and Raffaele Pettazzoni. This tradition, alternatively called comparative religion, phenomenology of religion, the history of religions, and, most simply, religious studies, maintains that religion is religion and little, if anything, else. Accordingly, religion is not the subject of anthropology, sociology, psychology, or economics. It is the subject of religious studies alone. The focus of religious studies is less on the truth of religion than on its origin and function. Religion, according to this tradition, arises for the distinctly religious purpose of making contact with God.

The tradition is, however, ambiguous about its aims. Sometimes it claims to be merely *presenting* the adherent's point of view, which purportedly deems religion a route to God. Other times the tradition claims to be *endorsing* the adherent's point of view. Now religion really does serve to make contact with God. The tradition thereby ceases to offer a mere description of religion and now offers an explanation, or a theory, of religion. But then the issue of the truth of religion proves difficult to sidestep. For if religion succeeds in making contact with God, God must exist.

Eliade's own position is unclear. He himself wavers between presenting and endorsing the believer's point of view. Typically, he declares modestly that 'The ultimate aim of the historian of religions is to understand, and to make understandable to others, religious man's behavior and mental universe' (Eliade 1959a, p. 162). In contrast to the theologian, from whom he repeatedly distinguishes himself, the historian seeks only to explicate, not to advocate, the adherent's point of view: 'The procedure of the historian of religions is just as different from that of the theologian. All theology implies a systematic reflection on the content of religious experience, aiming at a deeper and clearer understanding of the relationships between God–Creator and man–creature. But the historian of religions uses an empirical method of approach. He is concerned with religio-historical facts which he seeks to understand and to make intelligible to others' (Eliade 1959a, p. 88).

Yet in actuality Eliade ventures far beyond the mere presentation of the believer's point of view. He does not hesitate to discuss dead religions, whose adherents are scarcely available for consultation. His chief source of information for even living religions is not informants but their texts and practices. Moreover, he is prepared to grant that the meaning of all religions is partly, even largely, unconscious. Of religious symbols, for example, he states that 'we do not have the right to conclude that the message of the symbols is confined to the meanings of which a certain number of individuals are fully conscious, even when we learn from a rigorous investigation of these individuals what they think of such-and-such a symbol belonging to their own tradition. Depth psychology has taught us that the symbol delivers its message and fulfills its function even when its meaning escapes awareness' (Eliade 1959b, pp. 106–7). Eliade goes as far as to assert that even most professedly non-religious moderns are in fact religious: 'The majority of the "irreligious" still behave religiously, even though they are not aware of the fact.... [T]he modern man who feels and claims that he is non-religious still retains a large stock of camouflaged myths and degenerated rituals' (Eliade 1959a, pp. 204–5). Eliade outright bars any exceptions: '[E]ven the most avowedly non-religious man still, in his deeper being, shares in a religiously oriented behavior' (Eliade 1959a, p. 211). In imputing to avowed atheists a religiosity of which they are wholly unconscious and which contradicts their own conscious one, he is doing more than describing the actor's point of view. His defense of his view of religion as irreducibly religious cannot therefore be that it is so for actors themselves. It must be that religion is itself irreducibly religious. He must be offering a theory, not just a description, of religion.

The aspects of religion on which Eliade focuses are myths, rituals, and symbols. All are means of encountering God. Religion for Eliade arises not from an encounter with God but from a *yearning* for an encounter. Religion provides both places and times for the encounter. It demarcates as 'sacred' those places, or spaces, where God has previously appeared. Adherents assume that God, who is elusive, will more likely appear where God has already been than in a new place. Even if adherents believe that God is everywhere, they still go to church to seek out God. Similarly, religion sets off as sacred those times, always long past, when God most often appeared. Because adherents believe that God was closest at hand during or just after creation, they use myths about creation not merely to tell them when God was nigh but, more, to carry them back to that time: 'It is easy to understand why the memory of that marvelous time haunted religious man, why he periodically sought to return to it. *In illo tempore* the gods had displayed their greatest powers. The cosmogony is the supreme divine manifestation, the paradigmatic act of strength, superabundance, and creativity. Religious man ... seeks to

reside at the very source of primordial reality, when the world was *in statu nascendi*' (Eliade 1959a, p. 80). Myth works like a magic carpet.

In the tradition of religious studies, Eliade maintains that the quest for God, or the sacred, is not a means to an end but the end itself. Adherents seek the sacred because they seek the sacred, not because they seek something else through it. Religion can have anthropological, sociological, psychological, and economic benefits, but those benefits are considered to be only the consequences of a quest for the sacred and not the purpose of the quest. Furthermore, Eliade tolerates no reciprocal causality, the way, for example, Max Weber does. Religion can have non-religious effects, but not vice versa. Other scholars in the field of religious studies do allow for multiple causes as well as for multiple effects of religion, but they still insist that the main cause is the religious one. For the uncompromising Eliade, the sole cause of religion is religious. Somehow religion cannot arise for both religious and non-religious reasons.

Because Eliade pits the religious cause against all other proposed ones, he pits the field of religious studies against the social sciences, which he continually faults for accounting for religion non-religiously, or 'reductively': '…a religious phenomenon will only be recognized as such if it is grasped at its own level, that is to say, if it is studied *as* something religious. To try to grasp the essence of such a phenomenon by means of physiology, psychology, sociology, economics, linguistics, or any other study is false; it misses the one unique and irreducible element in it—the element of the sacred' (Eliade 1958a, p. xiii). Eliade takes for granted that, as disciplines, the social sciences stand committed to a non-religious account of religion.

Only once, in all his writings, does Eliade qualify his condemnation of the social sciences as resolutely reductive in their approach to religion. In fact, the qualification comes in an otherwise unbending indictment of anthropologists for spurning the adherent's point of view, which he assumes to be religious: 'Indeed, as it is well known, the main interest of the western scholars has been the study of material cultures and the analysis of family structure, social organization, tribal law, and so forth. These are problems, one may say, important and even urgent for western scholarship, but of secondary importance for the understanding of the *meaning of a particular culture*, as it was understood and assumed *by its own members* (broader, sympathetic and intelligent studies like those of Marcel Griaulle, E. G. [sic] Evans-Pritchard or R. G. Lienhardt, and a few other cultural anthropologists are rather exceptional' (Eliade 1967, pp. 502–3 [italics in original]).

Eliade's characterization of anthropology was outdated even when he wrote the essay, and it has become ever more outdated ever since. At best, his characterization, which borders on caricature, applies to the British brand of social anthropology influenced by A. R. Radcliffe-Brown. But the interest of British social anthropologists in social structure has rarely been at the expense of an interest in religion, as the cases of Evans-Pritchard and Lienhardt, as well as that of Radcliffe-Brown himself, conspicuously attest.

More important than the accuracy of Eliade's depiction of anthropology and of other social sciences is his argument. For him, the social sciences wrongly disregard the believer's point of view. But in fact they do not. Far from ignoring, let us even suppose, the religious nature of religion for adherents, social scientists begin with it. The actor's point of view constitutes the data to be accounted for. If social scientists ignored 'the meaning of a particular culture, as it was understood and assumed by its own members', they would have nothing left to account for. Why religion has meaning for adherents is what social scientists want to know.

Yet if social scientists necessarily turn to adherents for the data to be accounted for, they do not necessarily turn to adherents for an account of the data. Believers are akin to patients; social scientists, akin to doctors. It is not that religion is a disease to be cured but that religion is better understood by doctors than by patients. Patients may provide their own diagnoses of their ailments, but the final diagnosis rests with the doctor. That a patient rather than the doctor has the disease does not make the patient an expert on it. A patient's own diagnosis may turn out to be correct, but not because it is the patient's.

Eliade's indictment of anthropologists for spurning the actor's point of view is even more tenuous because as noted, he himself spurns it. If he is entitled to find a religious function in the professedly non-religious activities of avowed non-believers, why are social scientists not entitled to find non-religious functions in the professedly religious activities of avowed believers? In truth, both are entitled to venture beyond the actor's point of view.

Eliade's additional argument for accounting for religion religiously is that an account of any phenomenon must match the phenomenon being accounted for. Just as an account of a work of literature must be literary, so an account of a religious belief or practice must be religious. Eliade is right to insist on symmetry between an account and the subject of the account. But he is wrong to assume that the phenomenon being accounted for can be characterized and thereby accounted for in only one way. A 'literary' work such as a novel can alternatively be characterized as a work of history, sociology, or autobiography, and can then be accounted for accordingly. A Freudian account of religion characterizes a believer's relationship to God as a child's relationship to the child's father, and then attempts to account for that relationship psychoanalytically.

For Eliade to reply that God is more than a human being would miss the point. Freud never claims to be

able to account for all aspects of God, only for father-like ones. He never claims that his characterization of the relationship between believer and God as filial is exhaustive. As a whole, social scientific accounts of religion claim at most to be offering necessary, not sufficient, accounts of religion. Freud recognizes the gap between a father and a God, and he is prepared to turn to religious studies to close it. He claims only to be demonstrating how father-like God is. Eliade, however, would bar Freud from offering an account, whether necessary or sufficient, of any aspects of God, and would do so on the question-begging grounds that religion is religion and not psychology.

See also: Anthropology, History of; Myth in Religion; Religion: Definition and Explanation; Religion, Sociology of; Religiosity, Sociology of; Ritual; Ritual and Symbolism, Archaeology of; Symbolism in Anthropology

Bibliography

Allen D 1998 *Myth and Religion in Mircea Eliade*. Garland/Taylor and Francis, New York

Berger A 1989 Fascism and religion in Romania. *Annals of Scholarship* **4**: 455–65

Eliade M 1954 *The Myth of the Eternal Return*. Trans. Trask W R, Pantheon Books, New York (also published 1955 Routledge and Kegan Paul, London; reprinted as 1959 *Cosmos and History*, Harper Torchbooks, New York; reprinted as 1971 *The Myth of the Eternal Return*, Princeton University Press, Princeton, NJ; 1991 Princeton University Press Mythos Series)

Eliade M 1958a *Patterns in Comparative Religion*. Trans Sheed R, Sheed and Ward, New York (reprinted 1963 Meridian Books, Cleveland; reprinted 1974 New American Library, New York)

Eliade M 1958b *Yoga: Immortality and Freedom*. Trans. Trask W R, Pantheon Books, New York; Routledge and Kegan Paul, London (2nd edn. 1969 Princeton University Press, Princeton, NJ; reprinted 1970 Princeton University Press, Princeton, NJ; 1990 Princeton University Press Mythos Series)

Eliade M 1959a *The Sacred and the Profane: The Nature of Religion*. Trans. Trask W R, Harcourt, Brace, New York (reprinted 1961 Harper Torchbooks, New York; reprinted 1968 Harvest Books, New York)

Eliade M 1959b Methodological remarks on the study of religious symbolism. In: Eliade M, Kitagawa J M (eds.) *The History of Religions: Essays in Methodology*. University of Chicago Press, Chicago

Eliade M 1963 *Myth and Reality*. Trans. Trask W R, Harper, New York (also published 1964 Allen and Unwin, London; reprinted Harper Torchbooks, New York, 1968; reprinted 1975 Harper Colophon, New York)

Eliade M 1964 *Shamanism: Archaic Techniques of Ecstasy*. Trans. Trask W R, Pantheon Books, New York; Routledge and Kegan Paul, London (reprinted 1970 Princeton University Press, Princeton, NJ)

Eliade M 1967 On understanding primitive religions. In: Müller G, Zeller W (eds.) *Glaube/Geist/Geschichte: Festschrift für Ernst Benz*. Brill, Leiden, The Netherlands

Eliade M 1969 *The Quest: History and Meaning in Religion*. University of Chicago Press, Chicago (reprinted 1975 University of Chicago Press Phoenix Books, Chicago)

Eliade M 1972 *Zalmoxis: The Vanishing God: Lomactive Studies in the Religious and Folklore of Dacia and East Europe*. Trans. Trask W R, University of Chicago Press, Chicago (reprinted 1978 University of Chicago Press Phoenix Books, Chicago)

Eliade M 1978–85 *A History of Religious Ideas*, 3 Vols. Trans. Trask W R (Vols. 1 and 2) Hiltebeitel A, Apostolos-Cappadona D (Vol. 3), University of Chicago Press, Chicago (also published 1979–85 Collins, London)

Eliade M 1987 (ed.) *Encyclopedia of Religion*, 16 Vols. Macmillan, New York

Eliade M 1981 *Autobiography, I: 1907–1937, Journey East, Journey West*. Trans. Ricketts M L, Harper and Row, San Francisco (reprinted 1988 University of Chicago Press, Chicago)

Eliade M 1988 *Autobiography, II: 1937–1960, Exile's Odyssey*. Trans. Ricketts M L, University of Chicago Press, Chicago

Idinopoulos T A, Yonan E A (eds.) 1994 *Religion and Reductionism: Essays on Eliade, Segal and the Challenge of the Social Sciences for the Study of Religion*. Brill, Leiden, The Netherlands

Leach E R 1966 Sermons by a man on a ladder. *New York Review of Books* **7** October 20: 28–31

Rennie B S 1996 *Reconstructing Eliade: Making Sense of Religion*. State University of New York Press, Albany, NY

Ricketts M L 1988 *Mircea Eliade: The Romanian Roots, 1907–1945*, 2 Vols. East European Monographs, Boulder, CO

R. A. Segal

Elias, Norbert (1897–1990)

Born on June 22, 1897 to Jewish parents in Breslau (now Wrocław, Poland), then part of Germany, Elias, after completing his secondary education, first set out to study medicine, but soon changed to philosophy. Only after having completed his dissertation in that discipline in 1925 did he turn to sociology, pursuing his studies at Heidelberg, then one of the leading universities in Germany. He was accepted by Alfred Weber, Max Weber's brother, as a candidate for the *Habilitation* (conferring qualification as a university lecturer or professor), but when his colleague and friend Karl Mannheim obtained a chair at the University of Frankfurt, Elias followed him in 1930 as his assistant. His thesis for the *Habilitation* was completed by the beginning of 1933, but with the Nazis having taken over power in Germany, Elias had to flee, leaving Germany first for France, and then, in 1935, for London, on a grant from a philanthropic organization. Between 1935 and 1938, he completed his now famous two-volume *Über den Prozeß der Zivilisation (The Civilizing Process)*, which, as a book in the

German language published on the eve of WW II, did not receive much attention before republication in 1969. Elias remained in England where he obtained a post as lecturer in sociology at the University of Leicester but not before 1954. After retirement in 1962, he spent 2 years as professor of sociology in Ghana. Later on, he moved first to Amsterdam and then to Bielefeld (Germany), and was also invited to a number of European universities as visiting professor during the 1970s. In the mid-1980s he moved back to Amsterdam where he died on August 1, 1990 (on his biography, see Elias 1990 and Mennell 1989). Only during the 1970s, after a few sporadic articles, he had begun to publish extensively, first in English and later on, during his Amsterdam and Bielefeld years, drawing much on older manuscripts, in German. Some of the manuscripts were published only after his death.

There are few sociologists for whom knowledge about their biography is as central in understanding both their fate as a sociologist and their influence on the discipline as in the case of Elias. His first major book, and by far the most influential one, was not recognized as a masterpiece before his retirement as a university teacher. During the 1950s and 1960s, Elias seemingly was cut off, and cut himself off from his former German contacts, such as Theodor W. Adorno. Neither he nor those very few sociologists who acknowledged the outstanding achievement of *Über den Prozeß der Zivilisation* (e.g., Johan Goudsblom or Dieter Claessens) were successful in campaigning for this book. It only became famous in the early 1970s, when copies of the 1969 edition began to circulate among students and when a best-selling German paperback edition appeared in 1976. Translations into about 15 languages testify to the wide interest the book still attracts. Elias later complained about the insufficient recognition of his other works he published, among other subjects, on the sociology of sciences (Elias 1974), on time (Elias 1988), on relationships between the established and the outsiders (Elias and Scotson 1965), on sports (Elias 1970a) and even on topics such as Mozart (Elias 1991a) or *The Loneliness of the Dying* (Elias 1987a). Yet, without any doubt *Über den Prozeß der Zivilisation* is by far his most important, and probably also his only truly original contribution to sociology.

This is not to say that this work, like his others, does not draw on many concepts and notions developed by fellow sociologists. The idea of conceiving of society in terms of processes, instead of states, the absence of which from contemporary sociology was a reiterating theme in Elias's writings from the 1960s onwards, was firmly rooted not only in those American and English sociologists, such as M. Ginsburg and W. Osgood, whom Elias began to study in London in 1935 (van Krieken 1998, pp. 26–28), but also in important strands of German sociology, not least two thinkers whose work he was surely well acquainted with, Karl Mannheim and Alfred Weber. From the latter, he probably borrowed the notion of the civilizing process (Weber 1920), but giving it an entirely different content, as Weber by that term referred to the continual accumulation of knowledge, conceived of as a transition from 'naivité' to reflection and to mastering of nature (a theme that is not entirely absent from Elias's late writings). Other important influences that are easily recognized were Sigmund Freud and the Dutch historian Johan Huizinga, whose *Herftij de middeleeuwe* (Huizinga 1924) was Elias's dominant source on the late Middle Ages. But the blend Elias formed from these, and other ideas (such as traces of Darwinism or even Malthusianism), together with his own research, was decidedly unique.

The main achievement lies in the rigor with which Elias portrays the development of certain forms of conduct or manners, and ties this development to large-scale processes of state-formation and modernization on the one hand, and the concomitant effects on the psychic level, above all the display of emotions, on the other. The first volume of *Über den Prozeß der Zivilisation*, drawing on a variety of etiquette manuals, lucidly describes how norms of conduct we now think as self-evident, such as our eating habits, or norms concerning bodily functions including nakedness and sexuality, gradually developed from the Middle Ages onwards (with the selection of this 'starting point' not meant to imply that the process had not begun before that time). The Middle Ages are described as a time when man could answer needs and desires more spontaneously, when emotions were displayed with much less restraint. The gradual replacement of fingers with forks and knives at the table, the diffusion of the idea that spitting or blowing one's nose onto the ground or into the fingers are to be considered embarrassments to others, or the refinement and suppression of violence—all these are examples of the changes of our outward, overt behavior in the direction of increasing 'civilization.' These changes are accompanied by a gradual shift of the psychic apparatus which is characterized by an increasing influence of the super-ego (especially feelings of shame and embarrassment), and also of the ego, over the human drives, or the id, and by a restriction of affects and impulses, a 'damping' of the amplitude of the swings of emotion.

These developments were, in the second volume of *Über den Prozeß der Zivilisation*, analyzed as embedded in large-scale historical processes. The rise of the European empires, especially French absolutism, and the concomitant development of monopolies of violence on the state level restrained aggressive behavior in everyday conduct. The French court was also the site where the nobility developed those manners that were continually refined in the courtiers' competition for the king's favor (a process described in more detail in Elias 1969). These manners also served as elements of distinction of the entire nobility in contrast to the *bourgeoisie*—but to no avail, as the

latter, and later on also other social strata, began to imitate and to take over the nobility's patterns of behavior. These interdependencies are part of the overall process of modernization, especially of social differentiation, which requires of modern man a degree of reflection, foresight, and calculation unknown to medieval man. The distinctive overall characteristic of this process is the increasing internalization of social constraints, or the gradual replacement of 'outer' constraints by self-constraint—a self-constraint we are only sometimes conscious of, as it is routinized and habitualized, or incorporated, in our everyday conduct.

The main thrust of the book thus resonates well with a theme that has become central to the social sciences and the humanities in the 1970s: the investigation of domination, or constraint, not (only) on the (macro) level of relationships between social groups, but rather as it makes itself felt on the (micro) level of emotions and of individual conduct. It is a topic that can be found in the writings of Herbert Marcuse, or of Michel Foucault of the mid-1970s, who both echoed, albeit with very different emphasis (and seemingly without relating to Elias), the 'discontents of civilization' Sigmund Freud had already articulated in the late 1920s. Elias's account was perhaps received with even more sympathy as it nonetheless retained an optimistic tone. On the one hand, it was sensitive to the 'costs of civilization' that were felt by many at that time. On the other, Elias regarded the process of civilization as possibly leading to a balance between individual needs and the requirements of social life and in the long run to a more peaceful world society.

The uniqueness of *Über den Prozeß der Zivilisation* has two (related) dimensions: In terms of content, it lies in its blend of sociology with historical and psychological insight. In terms of method, it lies in its linking of grand theory with careful scrutiny of detail. Many sociologists would readily acknowledge that sociology has to look at historical processes; however, few, if any, have actually delivered a picture that so vividly describes the changes in people's manners and habits that have occurred from the Middle Ages onwards. Elias indeed gave a detailed description of how 'medieval man' (women are indeed largely absent from his account) gradually became 'modern man'; that is, where others have rendered snapshots, Elias has produced a movie. We can indeed understand that medieval man, as depicted by Elias (i.e., more impulsive, aggressive, and yielding to emotions than modern man) is alive today, but in a more suppressed, constrained way.

This, together with its often lucid language, largely free of sociological jargon, has rendered *Über den Prozeß der Zivilisation* popular among a wider audience of non-specialists. It has also been received well by historians such as Robert Chartier or Robert Muchembled, and also in psychology and other humanities. This is not meant to imply that Elias has triggered new research agendas in these disciplines; rather, his work was considered as congenial and as providing an overarching framework for research that had been devoted to similar topics, but without his broad, all-encompassing approach. At the same time, it is often concluded that by assuming a long-term trend towards 'more civilized' behavior, Elias possibly has drawn too far-reaching conclusions. While accusations of eurocentrism in his work easily can be refuted, the observation that other cultures have developed comparable standards of civilization (in Elias's sense) remains a challenge to the theory. It is also questioned whether medieval man indeed was so different. Certainly some of what Elias says on the display of emotions is valid; but the Middle Ages also had highly developed cultural rules for constraining and guiding behavior. A related problem is that Elias in the final analysis was not able to account for the cruelties of the twentieth century, not least those of Nazism. Elias certainly was aware of the ambivalences of civilization, but they were recognized by him mostly on the individual level, to the neglect of the dangers inherent in such seemingly civilized (and civilizing) forces such as modern technics or the state monopoly of violence. It must be added that during his late life, he became more sensitive to these issues, but still arguing fervently in favour of the possibility (not the unavoidability) of worldwide peace.

Elias's account exhibits serious lacunae also on the more factual level (van Krieken 1989). Among the deficiencies of *Über den Prozeß der Zivilisation* is the absence of an account of the processes through which the manners of conduct, described by Elias only with reference to the etiquette manuals, were effectively brought about by (private, and later on public) educational practices. Even more importantly, some presumably decisive conditions of modern civilization (in Elias's understanding of that term) are entirely lacking. Among these, most conspicuous are, first, the disciplinary mechanisms that were invented or became more widespread during the eighteenth and nineteenth century, especially the school and the factory, an observation that also points to the almost complete absence of the working class from Elias's work. Second, going further back in history, we also find a neglect of religion, certainly one of the most important sources of (European) civilization in the form of monastic discipline and more generally the religious prescriptions (and proscriptions), not least of protestantism.

Elias's later theoretical work, claiming to develop a general 'theory of human beings' (*Menschenwissenschaft*), exerted little influence on sociology in general. The earlier focus on social processes in these later writings was complemented by more emphasis on the 'figurational' nature of society, a term coined by Elias to point out that man and society could and should be understood as relational entities (see especially Elias 1970b, 1987b). With that concept, he claimed to

overcome the alleged dichotomy of individual and society that he saw prevailing in most other sociologists' thought—an idea borne not so much from in-depth acquaintance with sociological theorizing but rather from his desire, obvious in most of his late works, to portray himself as a unique innovator of sociological theory. Figurations were analyzed mainly in terms of power—or more exactly of power differentials—of the dynamics resulting from these, and finally of the outcomes which, as he often argued, cannot be understood as a realization of the purposes of social agents but rather as unintended consequences. In other words, his sociology was mainly a structural-relational one, with emphasis on the dynamics of social relations (in contrast to much of the 1930s to 1960s structuralism). The price was a neglect of individual motivation as well as of broader cultural systems of meanings, such as values or norms. The absence of such aspects from Elias's sociology is perhaps a consequence of his reluctance, acquired very early in his grappling with philosophy, of conceding ideas an existence of their own. Only in his latest writings did he start to acknowledge the influence of systems of meaning and language (Elias 1988, 1991b). With respect to some sociologists' claim to apply or to extend Elias's figurational sociology, it must be borne in mind that he never developed anything approaching a systematic account. Therefore, references to this theory must be viewed as partly gratuitous, as the idea of a 'relational' sociology certainly is not Elias's original contribution.

In contrast, fruitful inspirations continue to come from his theory of the civilizing process. Research in this spirit is accomplished above all in those places where he spent the last two decades of his life, that is, The Netherlands, especially Amsterdam, and also Germany, whereas in the USA, in spite of a few admirers such as Lewis A. Coser, his work has not met with wide attention. There have been fine extensions of the theory to specific phenomena such as eating practices, the history of marriage, or the domestification of fire, demonstrating how Elias's ideas are especially applicable to aspects of everyday cultural practices. But there are also attempts at rendering his approach useful to macro-analyses, such as in de Swaan's (1988) work on the development of modern welfare states. There is still much debate concerning the idea of civilization as an irreversible process and its relationship to modernity. While some researchers have explored the possibilities, and the conditions, of decivilizing processes, others have investigated processes of informalization, that is, the development of less restricted, destandardized norms of behavior, a topic on which Elias (1992, pp. 33–60) commented himself in one of his last works. Overall, it seems to be difficult to use Elias's ideas on the civilizing process as a starting point for really new developments; rather, much research focuses on applications of his insights to as yet unexplored phenomena, or on refining the theory in order to make it capable to account for phenomena that seem to contradict it at first sight.

See also: Civilization, Concept and History of; Civilizational Analysis, History of; Civilizations; Emotion: History of the Concept; Modernity; Modernity: History of the Concept; Modernization, Sociological Theories of

Bibliography

de Swaan A 1988 *In Care of the State. Health Care, Education and Welfare in Europe and the USA in the Modern Era*. Polity Press, Cambridge, UK

Elias N 1969 *Die höfische Gesellschaft*. Luchterhand, Neuwied (1983 *The Court Society*. Pantheon, New York)

Elias N 1970a The genesis of sport as a sociological problem. In: Dunning E (ed.) *The Sociology of Sport*. Cass, London

Elias N 1970b *Was ist Soziologie?* Juventa-Verlag, Munich (1978 *What is Sociology?* Hutchinson, London)

Elias N 1974 The sciences: Towards a theory. In: Whitley R (ed.) *Social Processes of Scientific Development*. Routledge & K. Paul, London

Elias N 1939/1976 *Über den Prozeß der Zivilisation*. Suhrkamp, Frankfurt-am-Main (1978/1982 *The Civilizing Process*. Blackwell, Oxford, UK)

Elias N 1987a *Über die Einsamkeit der Sterbenden in unseren Tagen*. Suhrkamp, Frankfurt-am-Main (1985 *The Loneliness of the Dying*. Blackwell, Oxford, UK)

Elias N 1987b *Die Gesellschaft der Individuen*. Suhrkamp, Frankfurt-am-Main (1991 *The Society of Individuals*. Blackwell, Oxford, UK)

Elias N 1988 *Über die Zeit*. Suhrkamp, Frankfurt-am-Main (1992 *Time: An Essay*. Blackwell, Oxford, UK)

Elias N 1990 *Über sich selbst*. Suhrkamp, Frankfurt-am-Main (1994 *Reflections on a Life*. Polity, Cambridge, UK)

Elias N 1991a *Mozart. Zur Soziologie eines Genies*. Suhrkamp, Frankfurt-am-Main (1993 *Mozart. Portrait of a Genius*. Polity, Cambridge, UK)

Elias N 1991b *The Symbol Theory*. Sage, London

Elias N 1992 *Studien über die Deutschen*. Suhrkamp, Frankfurt-am-Main (1996 *The Germans*. Blackwell, Oxford, UK)

Elias N, Scotson J L 1965 *The Established and the Outsiders*. Cass, London

Huizinga J 1924 *Herftij de middeleeuwe*. (1972 *The Waning of the Middle Ages*. Penguin, Harmondsworth, UK)

Mennell S 1989 *Norbert Elias: Civilization and the Human Self-image*. Blackwell, Oxford (*Norbert Elias: An Introduction*; reprint of this version 1998 at University College Dublin Press, Dublin)

van Krieken R 1989 Violence, self-discipline and modernity. Beyond the Civilizing Process? *Sociological Review* **37**: 193–218

van Krieken R 1998 *Norbert Elias*. Routledge, New York, London

Weber A 1920 Prinzipielles zur Kultursoziologie (Gesellschaftsprozeß, Zivilisationsprozeß, und Kulturbewegung). *Archiv für Sozialwissenschaft und Sozialpolitik* **47**: 1–49

W. Ludwig-Mayerhofer

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Elicitation of Probabilities and Probability Distributions

Bayesian statistical and decision analyses require several inputs aside from observed data. Among these inputs are *prior distributions* and *utility functions* (Barnett 1982), which typically reflect the opinion(s) of an individual or group. *Elicitation* is the process of encoding the subjectively held beliefs of an individual (called the *elicitee*) or group into prior distributions or utility functions. Elicitation is thus a complex (and imprecise) mapping from human knowledge processing to a quantitative metric. As such, the study of elicitation methods has two primary divisions—first, the psychological aspects of how humans are able to articulate knowledge, and second, the mathematical aspects of how such articulated knowledge is captured for use in a Bayesian analysis.

1. Subjective Opinion within a Bayesian Framework

Subjective opinion is actually employed in several parts of any statistical analysis, Bayesian or frequentist (Lad 1996) (see *Decision Theory: Bayesian* and *Decision Theory: Classical*). The Bayesian model of decision making and inference is that prior beliefs about a particular attribute or state of nature are updated through data, and then used together with utilities to decide on a course of action. Formally, θ represents a (possibly multiattributed) state of nature such that $\theta \in \Theta$; $\mathbf{x}$ represents observed data; and

$f_\Theta(\theta)$	Prior
$f_X(\mathbf{x} \mid \theta)$	Likelihood
$f_\Theta(\theta \mid \mathbf{x}) = \dfrac{f_X(\mathbf{x} \mid \theta) f_\Theta(\theta)}{\int_\Theta f_X(\mathbf{x} \mid \theta) f_\Theta(\theta)\, d\theta}$	Posterior
$L(\theta, \mathbf{a}) = -U(\theta, \mathbf{a})$	Loss Function ($= -$Utility)
$D(\mathbf{x}) = \min_a \int_\Theta L(\theta, \mathbf{a}) f_\Theta(\theta \mid \mathbf{x})\, d\theta$	Optimal decision minimizes posterior expected loss

Under a personal and subjective view of Bayesian inference sometimes called subjective expected utility (Savage 1954), a decision maker will choose the course of action, a, that, based on prior beliefs updated by collected data, will have the least adversarial consequences when compared with other possible decisions.

Elicitation is concerned with mapping subjectively held opinions into prior probabilities, $f(\theta)$, and (subjective) preferences and value judgments into utilities (or exchangeably, loss functions) $L(\theta \mid a)$. These prior distributions and utilities represent an amalgamation of information available in addition to the data collected for a particular problem. As such, the most valuable elicited opinions are those of experts in the area of the problem, who may (or may not) have any training in statistical methods. Meyer and Booker (2000) and Chaloner (1996) outline reasons why subjective expert opinions are valuable in Bayesian analysis. These include the *a priori* information that can be quantified in the investigation of new, rare, complex, or poorly understood phenomena; when collecting large amounts of data is not feasible; when collected data alone is insufficient evidence for decision making; in designing experiments (see *Experimental Design: Overview*); in forecasting future events; and incorporating historical data and experiences that may not be well documented.

Spetzler and Staël von Holstein (1975) identify three stages of an elicitation process: the *deterministic phase*, which corresponds to choosing a functional form for a statistical model, the *probabilistic phase* where the subjective assessments are made, and the *informational phase* where verification checks are carried out. The primary focus of elicitation methods is on the middle phase (probabilistic). It is assumed generally, at least in statistical practice, that the first phase is fixed by the likelihood chosen for the data; the third phase is addressed more in the research on cognitive aspects of elicitation than the statistical aspects of encoding subjective beliefs and preferences.

2. Cognitive Aspects of Elicitation

In reporting subjectively held beliefs and preferences, there are several psychological heuristics that can lead to misrepresentation (see *Cognitive Psychology: Overview*). For more detailed discussion on these, early work on the subject is found in Kahneman et al. (1982), Kyberg and Smokler (1980), Hogarth (1987); updated coverage is detailed in Poulton (1986) and in Wright and Ayton (1994). The most common problems in eliciting subjective opinions come from:

(a) *Overconfidence*. Probability assessors tend to underestimate variability and the tails of the distribution.

(b) *Hindsight bias*. Most assessors believe they would have predicted correctly the outcome of an event; thus only the outcomes that actually occurred are viewed as having nonzero probability of occurrence.

(c) *Representativeness*. Expert judgments can be based on the synthesis of previously observed data. If that data came from small samples, it may not be representative. Other terms often used in conjunction with this heuristic are *base-rate neglect*, *small-sample fallacy*, and *misperception of randomness*. Another well-known aspect of representativeness is the *conjunction fallacy*, where higher probability is given to a well-known event that is a subset of an event to which lower probability is assigned.

(d) *Availability*. The probability of an event is judged by the frequency with which an event can be recalled in memory. The classic example of this is in the elicitation of beliefs about likely causes of death; botulism, which typically gets a great deal of press attention, is usually overestimated as a cause of death, whereas diabetes, which does not generate a great deal of media attention, is underestimated as a cause of death.

(e) *Adjustment and anchoring*. When an initial assessment is made, elicitees often make subsequent assessments by adjusting from the initial anchor, rather than using their expert knowledge.

To overcome possible biases introduced in the elicitation of probabilities and utilities by these heuristics, Kadane and Wolfson (1998) summarize several principles for elicitation:

(a) Expert opinion is the most worthwhile to elicit.

(b) Experts should be asked to assess only observable quantities, conditioning only on covariates (which are also observable) or other observable quantities.

(c) Experts should not be asked to estimate moments of a distribution (except possibly the first moment); they should be asked to assess quantiles or probabilities of the predictive distribution.

(d) Frequent feedback should be given to the expert during the elicitation process.

(e) Experts should be asked to give assessments both unconditionally and conditionally on hypothetical observed data.

3. *Eliciting Prior Distributions*

Eliciting prior distributions entails quantifying the subjective beliefs of an individual into a statistical distribution (see *Distributions, Statistical: Approximations*). Kadane and Wolfson (1998) discuss the distinction between *structural* and *predictive* methods of doing this. In structural methods, an expert is asked to directly assess parameters or moments of the prior distribution; in predictive methods, elicitees are asked questions about aspects of the prior predictive distribution; the predictive distribution is

$$f_Y(\mathbf{y}\,|\,\mathbf{x}) = \int_\Theta f_Y(\mathbf{y}\,|\,\theta) f_\Theta(\theta\,|\,\mathbf{x})\,\mathrm{d}\theta$$

and when no data has been observed (i.e., $\mathbf{x} = \{\varnothing\}$), this is known as the prior predictive distribution. In most complex problems, the predictive method is preferable, since it is more suited to the prescription for good elicitation methods outlined in Sect. 2. The functional form of the prior distribution is usually selected to be conjugate to the likelihood chosen to model the data.

The elicitation is generally executed by asking assessors to identify quantile-probability pairs (q, p) in some sense, where $p = P_X(X < q)$. Questions can be asked by specifying either a point or interval value for p or q, and responses can be given directly, as values, probabilities, or odds, or indirectly where preferences are stated between two alternatives.

To illustrate, consider various ways in which a prior distribution for the probability parameter in a binomial model might be elicited. The beta distribution is the usual choice for modeling the prior distribution, and functionally, the model is

$f_P(p\,\|\,\alpha, \beta)$	$\sim Beta(a, \beta)$	Prior
$f_X(x\,\|\,p, n)$	$\sim Binomial(n, p)$	Likelihood
$f_P(p\,\|\,\alpha, \beta, x, n)$	$\sim Beta(\alpha + x, \beta + n - x)$	Posterior

In this instance, x is the number of successes observed in n trials; from the way in which the prior parameters are updated by the posterior distribution, $\alpha + \beta$ can be viewed as a 'prior sample size' or measure of prior strength, and $\alpha/(\alpha + \beta)$ as a point estimate of the mean prior probability of a success. So one way to structurally elicit a prior would be to ask an elicitee to specify a prior sample size and probability of success, and this would uniquely define the parameters of the prior distribution. This, however, does not provide any feedback to the elicitee, and will almost surely be overconfident because of the way in which the variability of the distribution is never elicited directly. It also does not address the issue of the skewness of the distribution. Chaloner and Duncan (1983) propose a predictive method for eliciting this same prior distribution, where the elicitee is asked to specify the modal number of successes for given prior samples sizes, and then is asked several questions about the relative probability of observing numbers of successes greater than or less than the specified mode. The elicitee is also shown (graphically) the implied predictive distributions for future observations specified by their current answers, and then walked through a bisection algorithm to adjust the spread of the distribution.

While the predictive method just described is more difficult to implement, it is also more appropriate in terms of having the elicitee answer questions that will effectively allow them to communicate both their knowledge and their uncertainty within the constraints of the model. To illustrate, suppose that in the structural method, an elicitee specified a prior sample size of 15 and a prior mean of 1/3. In the predictive method, the same elicitee specified, for a sample size of 15, a predictive mode of five successes; the odds of observing four rather than five successes was specified as 0.8, and the odds of observing six rather than five success was specified as 0.7. The two priors, in the structural case, a beta (5,10) distribution, and in the predictive case, a beta (9.67,18.74), are shown in Fig. 1. This illustrates that way in which the predictive

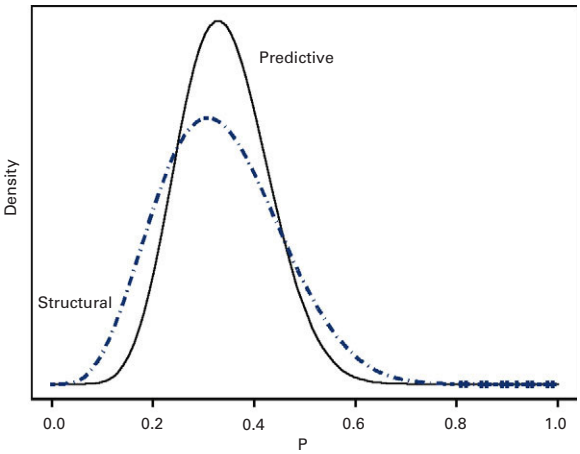

Figure 1
Structural and predictive elicited distributions for the beta-binomial model

method allows for better representation of the actual state of knowledge the elicitee has.

Existing protocols for eliciting prior distributions are described in some detail in Kadane and Wolfson (1998). Elicitation protocols are described as being either *general* or *application-specific*. General methods are those that are meant for specific classes of models, such as the beta-binomial, the normal linear regression model, the dirichlet-multinomial, and others. Application-specific methods are ones tailored to the problem under study; several examples of these are described both in Kadane and Wolfson (1998) and French and Smith (1997). Application-specific methods are used typically for complex situations, where there may be no conjugate or standard prior that is suitable, or where the elicitation process must be *decomposed* in order to provide sensible assessments.

4. *Eliciting Utilities*

The ultimate goal of eliciting utilities is to make decisions that minimize expected loss (see *Decision Theory: Bayesian*), under the constraints of uncertainty specified by the prior distribution and likelihood function. Determining expected loss means constructing loss functions that consist of outcomes based on a quantity of interest, $\mathscr{Q}$, and the loss associated with each outcome. The loss can be fixed or random. In general, there is a set $\mathscr{A}$ of possible actions that can be taken, and for each element of $\mathscr{A}$, there are losses that may or may not depend on $\mathscr{Q}$. The set $\mathscr{A}$ must be exhaustive; in other words, all possible actions must be enumerated, and each action must be exclusive of every other action.

The first stage of the elicitation process should be to choose the action set $\mathscr{A}$ and the quantity of interest $\mathscr{Q}$. The set of actions should be well defined and must contain at least two elements. These actions are usually dependent on some quantity of interest. For example, if the action set consists of choosing which family planning program to implement, the quantity of interest might be some quantification of the state of infant mortality. In choosing the quantity of interest, it is important to make sure that $\mathscr{Q}$ is a quantity that incorporates the uncertainty inherent in the problem under study. Generally, $\mathscr{Q}$ will be a random quantity defined by a statistical model, and as such, it can have several dimensions.

Once the action set $\mathscr{A}$ and the quantity of interest $\mathscr{Q}$ have been determined, the second stage of the elicitation is to construct loss functions based on them. This is the most difficult stage of the elicitation process. If several stakeholders are involved in the decision-making process, they may construct these separately or as a group; regardless, those assessing the loss functions must give careful thought to the consequences resulting from each action in $\mathscr{A}$ for any value of $\mathscr{Q}$. For useful overviews of loss function elicitation strategies, see works by Clemen (1996), Zeleny (1982), Edwards (1992), and Saaty (1980). In general, it is necessary to specify the form the loss functions should take; this is sometimes specified in advance, but it is advisable to consider this as part of the elicitation process.

There are several 'standard' loss functions (see *Decision Theory: Classical* for a discussion of them). One of the best-known examples of a loss function is the 0–1 loss function generally applied to statistical hypothesis testing. In this case, the set of actions is given by $\mathscr{A} = \{a_1$: reject the null hypothesis, a_2: do not reject the null hypothesis$\}$ (Berger 1985). An extension of this is the 0–k loss function, which is typically written (letting $\mathscr{Q} = \theta$)

$$L(\theta, a_i) = \begin{cases} 0 & \theta \in \Theta_i \\ k_i & \theta \notin \Theta_i \end{cases}, \quad i = 1, 2$$

There are many possible variants on this loss function that could be considered, and identifying the 'form' of the loss function is the more controversial of the tasks that must be performed in eliciting utilities. Given the conceptual framework of a 0–k loss function, consider the next elicitation task—identifying the values of k_1, k_2. Because k_1, k_2 can represent rather intangible quantities, it is often advantageous to specify their relative rather than absolute values. Consider that the 'optimal' decision in this example is to choose a_1 when

$$\frac{k_1}{k_2} > \frac{P(\theta \in \Theta_2 \mid \mathbf{x})}{P(\theta \in \Theta_1 \mid \mathbf{x})} = \frac{P(\theta \in \Theta_2 \mid \mathbf{x})}{1 - P(\theta \in \Theta_2 \mid \mathbf{x})}$$

Then the remaining elicitation task needed for decision-making is to elicit the ratio k_1/k_2, which

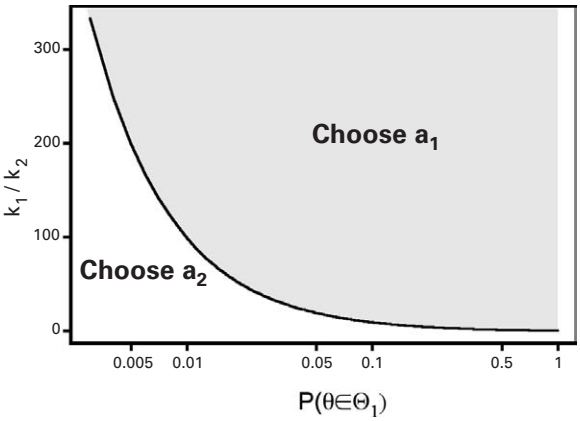

Figure 2
A decision maker should choose a_1 or a_2 based on the values of $P(\theta \in \Theta_1)$ and k_1/k_2

simplifies the elicitation task from eliciting k_1, k_2 directly or independently. The ratio on the right-hand side above can be viewed as the ratio of the posterior probabilities that the parameter lies outside the 'rejection region' ($\theta \in \Theta_2$) to the posterior probability that the parameter lies within the region (or $\theta \in \Theta_1$). From this, a general relationship can be drawn between the functional form of the model and the two quantities for which preferences need to be elicited; this is illustrated in Fig. 2.

Exploiting the relationships between the functional form of the model and loss function and the quantities about which preferences and value judgments must be elicited is a useful strategy. Many examples of the application of evaluating utilities and preferences are found in journals in the fields of risk analysis, accounting, and economics.

5. *Other Issues*

It is sometimes advisable or necessary to elicit the opinion of more than one individual. Genest and Zidek (1986) review methods for combining the opinions of several experts; Cooke (1991), Meyer and Booker (2000), and Morgan and Henrion (1990) provide guidelines for structuring elicitation tasks with groups of experts. French and Smith (1997) have several examples of analyses where elicitation of more than one expert is used. In general, there are two approaches used in group elicitation—the collective opinion of the experts is elicited in the whole or the opinions of several experts can be combined in some fashion.

Another issue that is discussed in the literature (see Kadane and Wolfson 1998 for further discussion) is that of validating and comparing elicitation methods. There are several criteria that might be considered; the most important of these is *coherence*, that the elicited opinions obey the laws of probability (see *Probability: Formal*). In general, this should be considered a minimal requirement for any elicitation method. Other considerations are the *reliability* of the elicitation methods in producing stable reflections of the elicitee's beliefs and preferences; also, the elicitation method can be validated by an examination of whether it obeys *proper scoring rules* (i.e., does not systematically encourage the misrepresentation of beliefs and preferences). Some practitioners of elicitation, particularly in the field of meteorology, advocate the *calibration* of expert opinion, whereby elicited opinions are adjusted to reflect the elicitee's ability to correctly predict events. Subjective Bayesians typically object to calibration because subjectively held beliefs and preferences are not meant to be infallible. A final consideration given in the validation and comparison of elicitation methods should be given to the modeling of errors-in-elicitation. Some cognitive aspects of this are described in Poulton (1996), but very little statistical literature is available on the subject.

See also: Bayesian Statistics; Multivariate Analysis: Overview; Spatial Autocorrelation; Spatial Sampling

Bibliography

Chaloner K M 1996 The elicitation of prior distributions. In: Berry D A, Stangl D K (eds.) *Case Studies in Bayesian Biostatistics*. Marcel-Dekker, New York, pp. 141–56

Chaloner K M, Duncan G T 1983 Assessment of a beta prior distribution: PM elicitation. *The Statistician* **32**: 174–80

Clemen 1996 *Making Hard Decisions*, 2nd edn. Duxbury Press, Belmont, CA

Cooke R M 1991 *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, New York

Edwards W (ed.) 1992 *Utility Theories: Measurements and Applications*. Kluwer Academic Publishers, Boston

French S, Smith J Q (eds.) 1997 *The Practice of Bayesian Analysis*. Arnold, London

Genest C, Zidek J V 1986 Combining probability distributions: A critique and an annotated bibliography. *Statistical Science* **1**: 114–48

Hogarth R M 1987 *Judgment and Choice*. Wiley, New York

Kadane J B, Wolfson L J 1998 Experiences in elicitation (with discussion). *The Statistician* **47**: 3–19

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK

Kyberg H E Jr., Smokler H E (eds.) 1980 *Studies in Subjective Probability*, 2nd edn. Krieger, New York

Lad F 1996 *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*. Wiley, New York

Meyer M A, Booker J M 2000 *Eliciting and Analyzing Expert Judgement: A Practical Guide*. Society for Industrial and Applied Mathematics and American Statistical Association, Philadelphia, PA

Morgan M G, Henrion M 1990 *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, New York

Poulton E C 1994 *Behavioral Decision Theory: A New Approach*. Cambridge University Press, Cambridge, UK

Saaty T L 1980 *The Analytic Hierarchy Process*. McGraw-Hill, New York
Savage L J 1954 *The Foundations of Statistics*. Wiley, New York
Spetzler C S, Staël von Holstein C A S 1975 Probability encoding in decision analysis. *Management Science* **22**: 340–58
von Winterfeldt D, Edwards W 1986 *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, UK
Wright G, Ayton P (eds.) 1994 *Subjective Probability*. Wiley, New York
Zeleny M 1982 *Multiple Criteria Decision Making*. McGraw-Hill, New York

L. J. Wolfson

Elites, Anthropology of

The term elite, formerly an expression referring to a theological elect, came into circulation with its current meaning in the nineteenth century as mass society took shape in the West. Elites have been under-theorized and understudied in the social sciences, and thus the term has a certain vagueness: both 'traditional' and 'modern' societies are said to have elites, and the privileges of elites may be grounded in genealogy, land ownership, control of capital, bureaucratic position, educational attainment, or religious status, for example. The first to theorize elites were the Italians Vilfredo Pareto and Gaetano Mosca, each working independently in the late nineteenth and early twentieth centuries to develop a theory of hierarchy in modern society that was less abstract and less tied to the dynamics of capitalism than Marx's theory of class.

Anthropologists have not, until recently at least, evinced great interest in writing about elites, and there is hardly a field of elite studies in anthropology. The first generation of anthropologists, attempting to develop holistic descriptions of entire societies, was not interested in detaching elites from broader social wholes in order to focus on them and, in any case, many had been drawn to marginal societies in low-rent neighborhoods of the world system partly to escape their own privileged backgrounds at home. As more recent generations of anthropologists have developed an interest in colonial and post-colonial society, urban anthropology and repatriated anthropology, elites have still received surprisingly little ethnographic attention. There are many reasons for this, including practical obstacles, the historical conventions of the discipline itself, and the temperamental preferences of anthropologists. Elites are busy, important people who have the power to elude or obstruct anthropologists, and whose lives are usually unsuited to the anthropologist's classic (and time-intensive) technique of investigation—participant observation (Hertz and Imber 1995). (While it is easy to imagine anthropologists helping peasants with the harvest, spending hours with poor womenfolk while they cook and clean, or idling with urban streetcorner society, it is harder to imagine them sitting unobtrusively at board meetings, loitering in the hallways of the upper floors of the Pentagon, or being repeatedly invited to dinner with the Rockefellers.)

Anthropologists have also continued to neglect the study of elites for reasons of convention: the traditional division of labor in the social sciences has given anthropologists dominion over the geographically and socially marginal, and anthropology has historically developed as a discipline that explains to the center the logic of lifeways on the periphery. This convention retains a certain force even in the era of repatriated and postcolonial anthropology. There is also, finally, for many anthropologists a temperamental obstacle to fieldwork with elites: the traditional anthropological methodology of participant observation has historically demanded of anthropologists a reflexive sympathy for the people they have studied—largely poor and marginal people. This, combined with the liberal politics of most anthropologists, makes fieldwork with elites psychologically unappealing and conflictual.

Insofar as first and second wave anthropologists wrote about elites, they often studied them not so much as topics in their own right but as components of larger social wholes or as products of social processes that were themselves the principal object of analysis. Examples include Dumont's (1970) work on caste, Barth's (1959) monograph on patron–client relations among the Swat Pathans of Pakistan, Kuper's (1961) study of an African aristocracy, and Sahlins' (1958) work on chiefs and followers in Polynesia. Marcus (1983) observes that such work on elites tended to cluster in particular ethnographic zones where kinship—the core subject matter of early anthropology—was particularly important in understanding the dynamics of elite status: South Asia, marked by its caste system; the Pacific Islands, home to hereditary chiefs and 'big men'; Africa, peopled by hereditary chiefs and kings; and the oligarchies of Latin America. Work by first wave anthropologists emphasised the importance in sustaining elite status of kinship relations, marriage strategies, land inheritance, patron–client relations, and ritual—what we might call the preindustrial sinews of elite power.

As colonization, industrialization, and missionization transformed societies around the world, new sinews of elite power emerged. Accumulation of capital (not just land) became increasingly important, as did access to education, geographical mobility, and positions in the expanding bureaucracies of new states. In many colonial societies the power of traditional, usually hereditary, elites was challenged by the rise of new elites who owed their standing to Western education, positions in emergent state bureaucracies, and trading opportunities provided by the colonial en-

counter. The process of decolonization usually mobilized, or became ensnared in, rivalry between rising and falling elites but, over time, some sort of complex hybridization of old and new elites was often effected through, for example, intermarriage, the conversion of capital into land, and the Western education of chiefly offspring. (This process broadly re-enacted a similar development that took place in the industrializing societies of Western Europe where the ancient aristocracies and rising bourgeoisies learned over time to accommodate to one another through marriage, education, and financial alliance.) Processes of elite formation and mutation under and after colonialism are discussed by Ferguson (1990), Geertz (1963), Leach and Mukherjee (1970), Luhrmann (1996), and Vincent (1971).

As societies become more complex, bureaucratized, and marketized, the bases of elite power shift and new kinds of specialized elites (from physicists to investment bankers) emerge. One of the distinguishing features of complex industrial societies is that they have a more pluralistic constellation of elites than do their predecessor societies. Moreover, if the dominance of elites in 'traditional' societies was often expressed in direct, face-to-face contexts, elite dominance in industrial society increasingly is mediated by bureaucracies, investment markets, and educational systems, distancing elites from those below them. And if the privileges of elites in traditional societies were often brazenly hereditary, reproduction of elites in industrial societies is, in theory at least, less regulated by kinship. Thus, Marcus and Dobkin (1992), who studied traditional elites on the Pacific island of Tonga and dynastic families in the US, argue that, while kinship appears to define the transmission and reproduction of elite status for the Gettys just as much as it does for the Tongan aristocracy, the analogy is superficial and deceptive: for families such as the Gettys kin relations of inheritance constitute a shell that conceals the corporate and bureaucratic character of the contemporary dynastic family dependent, as it is, on lawyers and money managers for its functioning and reproduction.

Cohen (1981) points out that in industrial, bureaucratic societies recruitment to elites is generally advertised as meritocratic, yet principles of heredity often find a way to infiltrate and suborn processes of promotion ostensibly based on merit as members of elites transmit their privileges to their children. For example, private schools become a means of converting the wealth of one elite generation into the superior educational attainment of its children, allowing the circumvention of the meritocratic prohibition on the simple inheritance of elite position. Moreover, as both Cohen (1981) and Bourdieu (1984) argue, elites come to value taste and manners as signs of distinction when recruiting a new generation of leaders, and accomplishment in this realm is asymmetrically transmitted to the children of existing elites as they are subjected to esoteric training in manners and taste in informal social contexts within the family, in private clubs, in peer groups, and at private schools. Secret societies, such as the freemasons and the fraternities at elite universities in the US, can also play a role in short-circuiting the processes of meritocracy.

Anthropologists engaged in fieldwork in Western societies have tended, until recently at least, to ignore Western elites. Thus, for example, anthropologists in the 1950s and 1960s kept out of the lively, at times bitter, debate in sociology and among public intellectuals provoked by the publication of Mills's (1956) argument that, far from being a meritocratic society governed by a plurality of separate elites, the US was ruled by a single 'power elite.' According to Mills this elite class was integrated loosely by the backstage functioning of private schools, elite universities, private clubs, and the marriage market and had become adept at disguising its privileges and its degree of integration from the mystified masses. Instead of joining this debate on elites in US society, American anthropologists engaged in fieldwork at home at this time were busy writing about 'the culture of poverty' or, in a logical sequel to earlier fieldwork abroad, studying processes of acculturation and economic struggle in Western enclaves of ethnic migrants from Third World countries. To this day, repatriated anthropology produces far more studies of the marginal and powerless—welfare mothers, drug dealers, immigrant street traders, and so on—than of elites.

While it makes sense in terms of the historical conventions of the discipline that anthropologists would continue to seek their subjects on the periphery, albeit a periphery redefined, the anthropological tradition for studying down is problematic. It helps to ghettoize anthropology since those who prefer to study the marginal can more easily be written off as marginal themselves. It leaves a lacuna in anthropology's traditional boast to be the science that studies all of human culture, everywhere. Above all, while it is important for anthropologists to make the point that the poor and the marginal have cultures that have value and are worth studying in their own right, there is danger in an asymmetric ethnographic gaze that imposes the risks of anthropological surveillance disproportionately on the least powerful members of society. If one of the crutches of elite power is invisibility, anthropologists have an important potential contribution to make to democracy and public discourse, as well as to anthropological scholarship, in adapting and redeploying their traditional methodology for the study of elites.

A milestone in the recognition of these problems came, in a historical context where anthropologists were questioning their complicity with processes of domination from Vietnam abroad to the ghettoes at home, with the publication of Nader's (1972) article calling on anthropologists to 'study up.' Nader drew attention to the problems of imposing a special

visibility on the poor and powerless and appealed to anthropologists to shine an ethnographic spotlight on the functioning of power in the US. Her article is often misrepresented as a call for anthropologists to engage in bounded studies of elite cultures, where in fact she called for studies of 'vertical slices' of social life that would enable the analysis of hierarchical relationships between elites and others and dramatize the essential nature of those relationships for the grounding of elite status and power. This was a way of reworking the traditional holism of anthropology. However, if the first generation of anthropologists had subordinated the study of elites to the analysis of societies as social wholes, Nader's is a call for holism that puts the privileges and power of elites front and center.

Marcus (1983) and Marcus and Dobkin (1992) have also called for more anthropologists to study elites in the West and Marcus, as editor of the *Late Editions* series of edited volumes, has provided a forum for younger anthropologists to experiment in a variety of voices with article-length portraits of, among others, elite subjects. Marcus's own approach to studying elites differs in emphasis from Nader's. He is more interested in cross-cultural comparisons between the functioning of Western and non-Western elites, and he has called for a move away from structural analyses of elite power in society toward fine-grained ethnographic analysis of elite subcultures as, more or less, self-contained entities. His own work on dynastic families explores, for example, the way such families cultivate taste and display wealth, intergenerational relationships in wealthy families, and the social production of cynicism, eccentricity, and piety among the super-rich.

Recent years have seen a number of innovative studies of Western elites. Rose (1989), in a highly literary response to Nader's call to study 'vertical slices' of American life, juxtaposes vignettes from the lives of poor African-Americans in Philadelphia with scenes from the lives of business elites and the country gentry in Pennsylvania. Weatherford (1981) has written, in a somewhat popular vein, about the culture of the US Congress. Wedel (2001) has written a book that mixes investigative journalism and anthropology in its portrait of IMF and World Bank officials assigned to reform the economies of post-Soviet Eastern Europe. Pedelty (1995) has written a vivid account of the journalistic elite that shaped Western understandings of wartorn Central America in the 1980s, exploring the ways in which these journalists' articles are shaped in response to the interests of elites both in Central America and the US. Luhrmann (2000), author of a previous (1996) book on a post-colonialist elite in India, has written a fine study of the socialization of American psychiatrists—an increasingly important category of expert in the production of American culture and identity. Harding (2000) has written an ethnographic portrait of Jerry Falwell, bringing anthropological categories to bear on the analysis of a key player in the US religious elite, and Wulff (1999) has written about the culture of international ballet performers.

Some of the most rigorous and innovative ethnographic work on Western elites has appeared in the new field of science studies. This may be in part because scientists are supposed to espouse a culture of openness, thus making them easier to study from a methodological point of view (especially if they are clustered conveniently in laboratories). Also, scientists are what we might call marginal elites, being esoteric enough that they more comfortably fit the anthropological template of the exotic than do, say, lawyers and investment bankers. Relevant examples in the anthropology of science include Latour and Woolgar's (1979) study of the making of scientific facts in a biology laboratory; Traweek's (1988) investigation of particle physicists in Japan and the US; Gusterson's (1996) examination of the culture of American nuclear weapons scientists in California; Helmreich's (1998) portrait of artificial life specialists at the Santa Fe Institute; and Rabinow's (1996) ethnography of Cetus, a biotechnology corporation, which explores the conjoining of venture capital and scientific entrepreneurship in the remaking of the world. These ethnographies all broadly examine the socialization of scientists; the role gender, race, and class play in their cultural worldviews; and the construction of scientists' special authority as experts.

This new wave of studies of Western elites does leave some obvious, and rather large, gaps. We still await authoritative ethnographies of most government bureaucracies, of military elites, and of cultural elites from university professors to television producers. (The new field of cultural studies, which is emerging as a powerful interlocutor for many anthropologists, has, strongly influenced by the Birmingham School, shown more interest in popular culture, especially its reception, than in documenting the culture of the producers of culture.) Above all, we lack studies of business and corporate elites, these subjects having largely been surrendered to business schools in an unspoken interdisciplinary division of labor.

Recent years in anthropology have seen some experimentation with new modes of research and writing. We can expect the maturation of elite studies in anthropology to accentuate this trend. Given that elites are often geographically dispersed, busy, and difficult to access, we can expect modifications to the traditional ethnographic method as more anthropologists move to multi-sited projects and to supplement participant observation with archival work, interaction in cyberspace, and condensed formal interviews with informants met only once—a strategy I have elsewhere referred to as one of 'polymorphous engagement.' It is also surely no coincidence that many ethnographers of elites have abandoned the objectifying conventions of classic ethnographic writing for more humanistic, journalistic, novelistic,

or ironic styles of writing, since in elites we have informants who are also sophisticated and powerful readers of our work and will demand to engage with our theories and conclusions, often in public fora. Learning how to respond gracefully and honestly, but without losing anthropology's critical edge, is our challenge for the future.

See also: Civilizations; Community Power Structure; Conflict: Anthropological Aspects; Development: Social-anthropological Aspects; Elites: Sociological Aspects; Hegemony: Anthropological Aspects; Inequality: Comparative Aspects; Leadership: Political; Pareto, Vilfredo (1848–1923); Power in Society; Social Stratification

Bibliography

Barth F 1959 *Political Leadership among Swat Pathans*. Athlone, London

Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste*, trans. Nice R. Harvard University Press, Cambridge, MA

Cohen A 1981 *The Politics of Elite Culture: Explorations in the Dramaturgy of Power in a Modern African Society*. University of California Press, Berkeley, CA

Dumont L 1970 *Homo Hierarchicus: An Essay on the Caste System*, trans. Sainsbury M. University of Chicago Press, Chicago

Ferguson J 1990 *The Anti-Politics Machine: Development, Depoliticization, and Bureaucratic Power in Lesotho*. Cambridge University Press, New York

Geertz C (ed.) 1963 *Old Societies and New States: The Quest for Modernity in Asia and Africa*. Free Press of Glencoe, New York, Collier-Macmillan, London

Gusterson H 1996 *Nuclear Rites: A Weapons Laboratory at the End of the Cold War*. University of California Press, Berkeley, CA

Harding S F 2000 *The Book of Jerry Falwell: Fundamentalist Language and Politics*. Princeton University Press, Princeton, NJ

Helmreich S 1998 *Silicon Second Nature: Culturing Artificial Life in a Digital World*. University of California Press, Berkeley, CA

Hertz R, Imber J B (eds.) 1995 *Studying Elites Using Qualitative Methods*. Sage, Beverly Hills, CA

Kuper H 1961 (1947) *An African Aristocracy: Rank Among the Swazi*. Oxford University Press, London

Latour B, Woolgar S 1979 *Laboratory Life: The Social Construction of Scientific Facts*. Sage, Beverly Hills, CA

Leach E, Mukherjee S N (eds.) 1970 *Elites in South Asia*. Cambridge University Press, Cambridge, UK

Luhrmann T M 1996 *The Good Parsi: The Fate of a Colonial Elite in a Postcolonial Society*. Harvard University Press, Cambridge, MA

Luhrmann T M 2000 *Of Two Minds: The Growing Disorder in American Psychiatry*. Knopf, New York

Marcus G E (ed.) 1983 *Elites: Ethnographic Issues*. University of New Mexico Press, Albuquerque, NM

Marcus G E, Dobkin P 1992 *Lives in Trust: The Fortunes of Dynastic Families in Late Twentieth Century America*. Westview Press, Boulder, CO

Mills C W 1956 *The Power Elite*. Oxford University Press, New York

Nader L 1972 Up the anthropologist: Perspectives gained from studying up. In: Hymes D (ed.) *Reinventing Anthropology*. Vintage Books, New York

Pedelty M 1995 *War Stories: The Culture of Foreign Correspondents*. Routledge, New York

Rabinow P 1996 *Making PCR: A Story of Biotechnology*. University of Chicago Press, Chicago

Rose D 1989 *Patterns of American Culture: Ethnography and Estrangement*. University of Pennsylvania Press, Philadelphia, PA

Sahlins M D 1958 *Social Stratification in Polynesia*. University of Washington Press, Seattle, WA

Traweek S 1988 *Beamtimes and Lifetimes: The World of High Energy Physics*. Harvard University Press, Cambridge, MA

Vincent J 1971 *African Elite: The Big Man of a Small Town*. Columbia University Press, New York

Weatherford J M 1981 *Tribes on the Hill*. Rawson, Wade, New York

Wedel J 2001 *Collision and Collusion: The Strange Case of Western Aid to Eastern Europe, 1989–1998*. St. Martin's, New York

Wulff H 1999 *Ballet across Borders: Career and Culture in the World of Dancers*. Berg, New York and Oxford, UK

H. Gusterson

Elites: Sociological Aspects

As commonly defined, elites are small groups of people who exert substantial power and influence over the public and over political outcomes. This power is based on the possession and control of various resources, including economic ones (capital and means of production), control of organizations, political support, symbolic means (knowledge, information, and the manipulation of symbols), and personal resources (such as charisma, or ambition). Some elite persons—those whose power is based on political support or charisma—are also referred to as leaders.

On this basis, the major elites in contemporary societies include political leaders (who in democracies are usually divided between the government and the opposition), the heads of the state bureaucracy, the military and the judiciary, the captains of business and industry, key media people (top journalists, editors, owners and managers of large-scale media organizations), trade union leaders, the intellectual–cultural–academic elite (influential writers, philosophers, publicists, and the like), religious or church leaders and leaders of civil society, of major interest groups, nongovernmental organizations and social movements (the latter are frequently referred to as non-established, counter-elites).

Vilfredo Pareto (1935), who introduced the term into the social sciences, defined elites as those who are most capable in any area of activity. But, as most commonly defined in the social sciences today, elites are not necessarily the best and the brightest; they are

simply those who, for whatever reason, have been able to gain the resources that lend them substantial power and influence. Occasionally the elite is equated with the upper class. However, by the above definition, elites are those who wield power, irrespective of their class background and socioeconomic status. In general, elites tend to stem from privileged class background, but not necessarily from the upper class. The top business elite belongs to the upper (capitalist) class. But other elites (e.g., trade union leaders, prominent media figures, church leaders) do not necessarily belong to that class, yet they still hold power, and therefore, elite positions.

Some elites command more power than others, and elites may thus be looked upon as a pyramid: at the top there are those who hold power at the macro, or national level—such as political leaders, and the top military brass. At the lowest level there are the elites of small units, such as heads of universities or local organizations. And in-between there are those who hold less power at the national level or hold power only in certain sectors or localites, such as the leaders of states or provinces (in a federal system) and city mayors, though the boundaries between the various levels are not sharply defined. The school of thought in the social sciences that focuses on the role of elites in society—or elite theory—has been concerned chiefly with top level or national elites.

1. *Classical Elite Theory*

Historically, elite theory developed as an alternative to class theory. While class theory, fathered by Karl Marx, stems from the middle of the nineteenth century, elite theory emerged only at the turn and the beginning of the twentieth century. Once it appeared, however, it did so with triple strength, under the aegis of the Italian classical social theorists Pareto (1935), Gaetano Mosca (1939), and Roberto Michels (1915).

The central idea of these founding fathers is, that all societies and organizations—including democratic ones—are divided into the few who rule and the many who are ruled. The ruling minority—dubbed by Pareto the governing elite or class, by Mosca the political or ruling class, and by Michels the oligarchical leadership—concentrates power in its own hands, uses it in its own interest, and then finds a formula to justify its rule. And the majority that is ruled, the masses, is largely powerless and has but little impact on policy outcomes.

To this Mosca adds that, according to democratic theory, majorities rule minorities. In fact, even in a democracy, the domination of an organized minority over an unorganized majority is inevitable. According to Michels, the very principle of organization spells oligarchy. And Pareto adds that although in principle governing elites are made up of those who excel, this is not necessarily so in practice. As long as a governing elite is composed of people with qualities of excellence and is open to absorb the most talented from the nonelite—while shedding the more degenerate of its own members—this ensures a circulation of elites, that keeps society in a state of equilibrium and gradual change. When this circulation is impeded, the equilibrium is upset and society becomes prone to violent upheavals.

According to critics, the notion that there are those who govern and those who are governed is self-evident. To transcend this truism, it would have to be shown that the governing elite has internal unity, which exacerbates its power. This, in fact, is the thesis of the, now classical, theorist C. Wright Mills. Mills (1956) sees even democratic (especially US) society as dominated by the top executive political, military, and corporate economic personnel, that together form a concerted power elite. At the bottom there are the masses, the powerless and apathetic receiving end of the power-elite's manipulation and rule.

Another critique leveled at these classical elite theorists was, that they overemphasized the similarity of elite rule in all types of regimes. Thus they had little to say on what still makes democracy (with all its iniquities) a distinctive regime. It fell to the lot of other theorists—known as liberal, democratic elite theorists—to fill this void. These included the German Max Weber who wrote at the turn and the beginning of the century, and surprislingly also Mosca in some later developments of his thought. In the post-war era they also included Joseph Schumpeter (1954), Raymond Aron (1968) and Suzanne Keller (1963), whose writings may now also be considered classical.

Rejecting the idea that democracy spells the rule of the people, and clarifying that it inevitably spells elite rule, these theorists still stress its distinctiveness. This lies not only in restrained elite competition and free elections that ensure the responsiveness of elites to the demands of the public, but also in the structure of elites. In a democracy, this is characterized by the plurality, autonomy and differentiation of elites, which enables them to countervail each other's power. These checks and balances increase the liberty of the public and the power-holders' responsiveness to it.

In opposition to this democratic elite theory, it has been argued that the virtues it attributes to democracy are more fiction than reality. Hence those who believe in them are overly complacent with respect to democracy, naively believing that 'the system that we ought to have is the very one that we do in fact have' (Pateman 1970, p. 16).

2. *Contemporary Elite Theories*

Such critiques have been one of the main causes of the fact that, since the 1960s, elite theory has been overshadowed by Marxist and neo-Marxist class theory. Yet elite analyses have not disappeared, and, indeed in quantity and variety have exceeded the classical writings on this topic.

Some of these new analyses—such as those by William Domhoff (1983), and Thomas Dye and Harmon Zeigler (1996)—may be seen as descendants of classical elite theory, in offering a skeptical view of democracy, albeit with a new twist. These observers of the US scene see the power of elites as exacerbated by their interlocking and consensus. They view popular elections as little more than rituals which lend the ill-informed masses sparce influence over policy. Indeed, it is the elites that preserve democracy, which works best when the masses are only minimally involved in politics. Intriguing as these ideas are, they lay themselves open to charges of excessive cynicism.

Such a charge is, in fact, implied by writers, such as Nelson Polsby (1985) and Giovanni Sartori (1987). These are representatives of the pluralist school of thought and they argue that democracies are distinguished by a plurality of elites, that occasionally are in coalition, and, occasionally, in opposition to each other. Such elites develop more links with the public and greater responsivenss to its wishes, than elites ever did before. Once again, swinging the pendulum to the other side, Robert Dahl (1985), who previously was a prominent pluralist, later took a different view in which, although in a polyarchy (i.e., democracy) there is a plurality of elites, the business elite is more powerful than the rest, and can make the government serve its interests.

Is it the case, then, that elites in Western democracies are plural, divided and in collision, or are they unified and in collusion with each other? An original scheme on this has been presented by John Higley and Michael Burton (1989) and some of their colleagues. According to Higley and colleagues elite configurations vary and have an impact on the structure of regimes. Disunified elites generate unstable regimes and ideologically unified elites give rise to stable autocratic regimes. By contrast, consensually unified elites—divided ideologically yet unified by their agreement on the rules of the political game—produce stable democracies.

Critics view this scheme as containing tautological elements (democracy prevails when elites adhere to its rules); but its proponents' response is that consensual unity historically always precedes the emergence of democracy. Accordingly, the major political turning points in modern history have been transformations of elites into consensually united ones. In any case the scheme is most useful for having produced a large number of studies of the emergence and development of democratic regimes.

Such developments, particularly transitions to democracy in various parts of the world, have also been the focus of other elite analyses. These concur that such transitions are usually crafted by elites. They take place where the previous elite oligarchy included some restricted competition between elites (Diamond and Linz 1989), through the forcible imposition of democracy by elites, or through compromises between contending elites (Karl and Schmitter 1991, Huntington 1991). At times they produce quasidemocracies, in which the elected elites' power is considerably greater than it is in representative democracies (O'Donnell 1994).

One widespread accusation, leveled at both classical and new elite theories, is that they are not only elite, but elitist theories, and that they applaud elite rule. A response to this charge has been made by a new version of democratic elite theory (Etzioni-Halevy 1993). This holds that the analysis of elites is crucial not as an endorsement, but as a scrutiny, of their power, without which it would be more pernicious. According to some, elites are the enemies of democracy. In fact, since elites are inevitable, all depends on their inter-relations. What makes democracy distinctive is the attempt to set them up as watchdogs for each other, through their relative autonomy, particularly that of the nongovernmental elites from the government.

This, indeed, is the central, implicit, metaprinciple of liberal democracy. It is nowhere so stated, but without it, the explicit principles of democracy, free competitive elections, human rights and civil liberties, would fall by the wayside. Without an opposition that is free from repression by government and not dependent on it for its livelihood, there could be no free elections. Without a judiciary that is exempt from government interference and cannot be dismissed by it, human rights could not be ensured. And without a media elite that cannot be gagged, freedom of expression, free elections, or human rights and civil liberties, could not be sustained.

Some of the main threats to democracy spring from the subversion of this metapricinciple as, even in Western countries, elites develop a close connection with each other. Thus, politicians frequently live in a cosy symbiosis with business magnates. A symbiosis has also developed between state elites and leaders of civil society, social movements and voluntary organizations.

The relative independence of such bodies has long been one of the hallmarks of democracy. But, recently, environmental movements have been institutionalized in parties, governments, and governmental organizations. And voluntary organizations let governments offload some of their responsiblities to them in return for funding. Thus, the boundaries between the state and civil society have been blurred and the latter can no longer serve as a check to government power. This puts the onus of preserving the watchdog metaprinciple of democracy more heavily on other elites, such as the media, which are not always well equipped to carry the burden.

3. *Recent Developments in Elite Analysis*

Elite and democratic elite theory, justifiably, has also been censured for paying insufficient attention to the

public, and to economic, class-based inequalities within it, particularly in Western (capitalist) democracies. Such inequalities and exploitation had long been the main theme of Marx-inspired class theory. However, with the breakdown of communism in Eastern and Central Europe, class theory has suffered a severe setback.

This opened an opportunity for elite analysis to incorporate socioeconomic inequality into its own domain. It did so by evincing a greater interest than before, in linkages between elites and the public, and in the impact of such linkages on both democracy and inequality. Such linkages involve a reciprocal relationship, whereby elites organize and promote the interests of certain constituencies, classes or groups within the public, which in turn support these elites. Linkages may be strong and clear, or weak and blurred. They may be direct, or through mediating bodies. Sustainable democracy relies on linkages that combine strength with extensive mediation (Pakulski 2000).

Much also depends on which groups within the public the elites link up with. Linkages between elites and the disadvantaged classes and groups within the public, are crucial for a stable democracy. For the disadvantaged—encompassing millions of people—cannot promote their own interests, unless they generate or adopt elites to represent them. And, unless the interests of all, including the disadvantaged, are represented in the political arena, there can be no democracy. The coupling of elites and the disadvantaged also makes elites struggle for a more egalitarian distribution of resources, and thus is essential in promoting progress towards greater equality.

This is, in fact, what happened in the West at the end of the nineteenth and the beginning of the twentieth centuries, as labor movement, union, and party-leaders linked-up with the working class and, to some extent, promoted their interests. At that time workers' wages rose, inequalities decreased, the working class gained the suffrage, and democracy was born.

Recently, in postindustrial society, the linkages between elites and the disadvantaged—those in the underprivileged labor market, the unemployed, the poor, the excluded—has become especially problematic. Together with insecurity of employment, this creates a destabilizing potential for democracy (Higley and Burton 1999).

At the same time, several elites have uncoupled from the disadvantaged. Labor-party leaders have turned their parties into 'catch-all' parties; leaders of social movements are more concerned with quality of life and identity issues, than with the redistribution of resources; and unions have been weakened and can no longer staunchly support the interests of the workers. This partial vacuum in elite–public linkages exacerbates inequality and insecurity. This, in turn, boosts fundamentalism and spurs the birth of populist and doubtfully democratic movements (Etzioni-Halevy 1997).

Another inequality is that between men and women in power positions. Until recently women were virtually excluded from such positions, even in advanced democracies, and this fact was largely ignored in elite analysis. But in recent decades women's presence in such posts has increased—even though they have not reached parity with men—and so has the attention paid to this issue in elite analysis. This analysis now shows that, to access elites, women must often offset the 'handicap' of gender, by even more privileged social background, and even greater human capital, than those of their male counterparts. Once in decision-making posts, women tend to have less hierarchical orientations than men; thus their increased representation in them constitutes a crucial factor in decreasing inequalities in society at large (Vianello and Moore 2000).

New times create new elites, and the new elite analysis also focuses on these. The transition from communism in Eastern and Central Europe has been led by elites and involved novel elite configurations. Yet a marked continuity of elite personnel has been widely noted, so that the pressing question is whether changes in elite values and behaviours have accompanied the transition.

Meanwhile, in the West the information society has generated new conglomerations of power in the high-tech, computer, and Internet centered fields; and the elites at their head have few visible links with the public. Further, globalization has given rise to transnational organizations—such as the International Monetary Fund and the European Union—and these have sprouted a second layer of elites beyond those of the state. These elites have the ability to shape people's life chances around the globe, yet they maintain but a minimal accountability to the public, to the detriment of democracy in global or transnational organizations. New transnational grassroots movements whose counter-elites would balance globalization from above with globalization from below, could vitalize transnational democracy.

See also: Aristocracy/Nobility/Gentry, History of; Class: Social; Community Power Structure; Elites, Anthropology of; Leadership: Political; Leadership, Psychology of; Pareto, Vilfredo (1848–1923); Pluralism; Political Elites: Recruitment and Careers; Power in Society; Power: Political; Social Inequality in History (Stratification and Classes); Social Stratification

Bibliography

Aron R 1968 *Progress and Disillusion*. Pall Mall, London

Dahl R A 1985 *A Preface to Economic Democracy*. University of California Press, Berkeley, CA

Diamond L, Linz J 1989 Democracy in developing countries. In: Diamond J, Linz J, Lipset S M (eds.) *Latin America*. Lynne Rienner, Boulder, CO, pp. 1–52

Domhoff G W 1983 *Who Rules America Now?* Prentice-Hall, Englewood Cliffs, NJ
Dye T, Zeigler H 1996 *The Irony of Democracy*, 10th edn. Wadsworth, Belmont, CA
Etzioni-Halevy E 1993 *The Elite Connection*. Polity Press, Cambridge, UK
Etzioni-Halevy E 1997 Elites and the working class. In: Etzioni-Halevy E (ed.) *Classes and Elites in Democracy and Democratization*. Garland Publishing, New York, pp. 310–26
Higley J, Burton M G 1989 The elite variable in democratic transitions and breakdowns. *American Sociological Review* **54**: 17–32
Higley J, Burton M G 1999 Elites, mass publics, and democratic prospects in post-industrial societies. *International Review of Sociology* **9**: 221–37
Huntington S 1991 *The Third Wave*. University of Oklahoma Press, Norman, OK
Karl T L, Schmitter P C 1991 Modes of transition in Latin America, Southern and Eastern Europe. *International Social Sciences Journal* **43**: 269–84
Keller S 1963 *Beyond the Ruling Class*. Random House, New York
Michels R 1915 *Political Parties*. Hearst's International Library, New York
Mills C W 1956 *The Power Elite*. Oxford University Press, London
Mosca G 1939 *The Ruling Class*. Kahn H D (trans.), McGraw-Hill, New York
O'Donnell G 1994 Delegative democracy. *Journal of Democracy* **5**: 55–69
Pakulski J 2000 Responsive or Responsible Elites? Unpublished
Pareto V 1935 *The Mind and Society*. Livingston A (trans.), Harcourt Brace Jovanovitch, New York
Pateman C 1970 *Particpation and Democratic Theory*. Cambridge Univesity Press, Cambridge, UK
Polsby N W 1985 Prospects for pluralism. *Society* **22**: 30–34
Sartori G 1987 *The Theory of Democracy Revisited*. Chatham House, Chatham, NJ
Schumpeter J A 1954 *Capitalism, Socialism and Democracy*. Routledge, London
Vianello M, Moore G (eds.) 2000 *Gendering Elites*. Macmillan Press, Basingstoke, UK

E. Etzioni-Halevy

Emergent Properties

Emergence and emergent properties are concepts that are used across many scientific disciplines. They are the key elements of a theoretical and methodological tradition—most prominently represented in the 1920s—claiming that the whole is more than the sum of its parts. This amounts to saying that the properties of a whole (systemic qualities) cannot be deduced by summing or averaging the properties of its components (nondeducibility thesis). Vice versa, the appearance of genuinely novel properties characterizing compounds or systems as wholes (i.e., emergent properties) cannot be solely predicted from knowing the properties of the constituents and their relations. Emergence theories are most elaborate in philosophical thought. However, an influential strand of twentieth-century social theory (Emile Durkheim, Talcott Parsons, Niklas Luhmann) borrowed some elements of this philosophical tradition and adopted the idea that the social (i.e., social organization at any level) should be regarded as an emergent property. Arguing against the powerful approach of reductionism, this sociological tradition asserts that systemic properties of the social world cannot be deduced from properties of the constituents (i.e., individuals) and their respective interactions. They need to be accounted for at the system level. In the latter part of the twentieth century, the scientific discourse about emergence has regained prominence in many scientific disciplines—not to speak of a veritable renaissance. More recent approaches to the understanding of emergent properties include, among others, the theory of deterministic chaos, complex-system theory, and theories of self-organization and self-referentiality. Interdisciplinary in orientation, they focus on the complex dynamics between interactions and relations at different organizational levels (multilevel analysis) and stress the coexistence of system concepts and constituent concepts in explaining composition outcomes. After a brief history of the concept in philosophical thought, the major theoretical discourses about emergence in the context of social theory will be discussed. Next, there is a brief discussion of the recent interdisciplinary approaches that incorporate the idea of emergence. Finally, methodological issues related to the measurement and the inference of emergent properties are discussed.

1. *The Philosophical Tradition of Emergence*

The concepts of emergence and emergent properties have a long philosophical tradition (see Stephan (1999) and Beckermann et al. (1992) for extensive reviews). Unanimously, the origins of the idea of emergence are attributed to John Stuart Mill (1974). Considered to be the classic source, Mill distinguished between two types of joint effect produced by the interplay of various causes (composition of causes). In the first type, the joint effect of several component causes is identical with the sum of the separate effects. With respect to the second type, the joint effect is heterogeneous to the separate causes. The properties of a system or a compound generated by the second type of joint effect were later to be labeled *emergent*. Offering an alternative to mechanism and vitalism, the 1920s may be regarded as the heyday of the philosophical tradition that grew out of Mill's fundamental insight.

It was in this decade that those later to become *classic* authors—Samuel Alexander, Conwy Lloyd Morgan, Charles Broad, and Roy Sellars—published their influential works. According to Stephan (1999), 'weak' versions of emergentism may be distinguished from 'strong' ones. The former loosen some of the

assumptions deemed to be essential for defining emergent properties in the latter versions. Both strands of theory assert that emergent characteristics of wholes are not the sum of the properties of their parts (nonadditivity). They cannot be understood by simply summing (or averaging) the behaviors of the constituents; rather, they must be described in terms of their own. The novelty thesis claims that in evolution genuinely novel (systemic) properties emerge. Examples are the advent of mind or reflective thought. In more abstract terms, this amounts to saying that emergent system properties are different in type from the properties of the system's constituent components. The two strands differ, however, with regard to the core assumptions of emergence theory—deducibility and predictability of systemic properties. Strong versions maintain that emergent properties are not deducible or predictable from the properties of the constituents and their relations (nonpredictability and nondeducibility). In this view, emergence may be characterized as the antithesis of microreductionism (Auyang 1998). By contrast, in weak variants of emergence theory, all systemic properties that distinguish a system from the qualities characterizing its elements must be regarded as candidates for representing emergent properties. The late 1960s and early 1970s saw a renaissance of ideas of emergence. Centering around the fundamental issue of the mind-brain relationship—the psycho-physical problem—Bunge (1977) and Popper and Eccles (1977) offered theories strongly drawing on the idea of emergent properties.

2. *Sociological Conceptualizations of Social Reality*

From its inception, sociological thought has been characterized by two dramatically different conceptualizations of social reality. In particular, these differences pertain to the fundamental understanding of the relationship between the individual and social organization. Methodologically, this problem may be couched in terms of the relation between the micro- and macro-levels—in other words, the micro–macro link. One tradition, most commonly referred to as the individualistic–reductionist position, asserts that the properties of the social at any level may ultimately be explained by the interplay between the properties of individuals (e.g., motives, preferences), who are the constituents of society. Having strong roots in the Scottish moral philosophers (David Hume, Adam Smith, Adam Ferguson), the development of twentieth-century social theory has been characterized by various attempts to explain social phenomena by the rules governing the behavior of individuals. Basically, this approach asserts that social organization results from individual actions. Fundamentally opposed to this view is the other sociological tradition, often classified as the collectivist position. Its basic assumption is that the social world constitutes a reality of its own. Historically, this position is closely intertwined with the demands of sociology as a young scientific discipline to claim its territory and to define its own true subject matter vis-à-vis already well-established disciplines. By asserting that society is a reality of its own, nonreducible to the rules governing individual behavior, sociology distinguishes itself especially from economics and psychology. Searching for the particular subject matter of sociology, proponents of this position asked about the novel qualities and the specificity of the social vis-à-vis other types of reality, such as the mental realm, the biological or physical realm. It is with this latter tradition in sociology that the ideas of emergence and emergent properties have been most commonly associated. For this reason, the historical development of the latter tradition is discussed below.

3. *The Sociological Tradition of Emergence: Society as an Emergent Property*

In late-nineteenth and early-twentieth-century sociology, the ideas of emergent properties are reflected in the conceptualization of society as representing a reality of its own endowed with separate characteristics nonreducible to the properties of its constituents (i.e., individuals). The sociological tradition that incorporates this assumption strongly draws on the thesis of emergence as nondeducibility. In the early days of sociology, Emile Durkheim (1938) most prominently advocated this approach to the understanding of social phenomena. In Durkheim's view, society is not solely the sum of individuals; rather, it is the *association* of individuals. Hence, it is the *nature of this association* that constitutes the decisive factors of social life and determines the proper characteristics of social phenomena ('faits sociaux'). This theoretical stance was epitomized in the famous methodological rules introduced by Durkheim (1938) according to which the social realm needs to be explained by social phenomena.

In modern sociological thought, developed in the second half of the twentieth century, general systems theory has advanced to one of the most influential currents, implicitly or explicitly advocating ideas of emergence (Parsons 1937, 1951, Luhmann 1995). With the Durkheimian tradition, systems theory shares the fundamental methodological conviction of the nondeducibility of the social world. Social phenomena at different levels of social organization (i.e., society, organizations, social groups) show novel properties that cannot be explained by reducing them to the properties of and interactions between their constituents (i.e., behavioral properties of individuals). Hence, a genuine *social* theory is needed to account for social reality as an emergent property.

This line of thought, advocated by Durkheim in late-nineteenth-century sociology, was continued in the mid-twentieth century by Talcott Parsons, albeit with some major modifications (Parsons 1951). For Parsons, the old problem raised by Hobbes about the origin of social order and social integration constitutes the fundamental sociological issue to which social theory needs to provide a satisfactory answer. The core of Parsons' answer to this fundamental problem is his thesis on the independence of institutionalized values and norms vis-à-vis the reality of individuals. The common system of ultimate values, constituting a reality *sui generis*, helps to establish order at various levels of social life and to prevent a free-for-all. In essence, Parsons declares common-value integration as the emergent property of social action systems. Parsons derives this conclusion from his systems-theoretical approach to the relationship between individual actors and social organization, whereby the starting point is the assumption that all reality realms constitute systems (i.e., the biophysical, personality, cultural, and social systems). Despite fundamental differences between various types of systems, they all share the system character as a common feature. By drawing a sharp distinction between the social system (the social reality) and the personality system (the human being), Parsons maintains that each system needs to solve its own particular functional problem (i.e., system stability), for which each system relies on self-regulating mechanisms. In the case of the social system, social order and social integration are the functional problems, and processes of socialization and social control are the self-regulating mechanisms. Despite their own system logics, the two systems interact with each other in such a way that each system constitutes the environment for the other. The interactions between systems and their environment include the mutual exchange of resources, necessary for each system to keep up its functional capacity (i.e., system maintenance). Within this systems-theoretical framework, the social system and the personality system mutually presuppose each other; they are of equivalent status. This implies that there is no ontological priority of one system over the other. Hence, the properties of the social world cannot be deduced from the rules governing the personality system. The society is thus an emergent property.

Within the tradition of systems theory, Niklas Luhmann (1995) has provided the latest elaborate attempt to conceive of social reality as a separate, nonreducible system, reproducing itself solely on the basis of its own elements. Sharing some common ground with Parsons' theoretical approach, Luhmann designs a systems theory that makes use of fundamental ideas developed in molecular biology. In particular, Luhmann applies the concept of autopoiesis advocated by Maturana to social reality. Arguing at a very high level of abstraction, social systems are autopoietic systems in as much as they are self-reproduced by their own components, which are defined as acts of communication. This amounts to saying that social systems operate on the basis of communicative acts engendering subsequent acts of communication. Hence, the social system is an independent, self-reproductive whole. Like Parsons' theory, Luhmann's approach attributes great significance to the strict distinction between the social system and the psychological system. These two separate systems reciprocally constitute the environment of the other, for any system needs appropriate environmental relations for its self-reproduction. Given the system–environment relationship, the autopoiesis of the social system is not conceivable without the autopoiesis of the psychological system and vice versa. In other words, the two autopoietic systems enable each other to operate. From this it follows that the two systems are equivalent (and not hierarchically dependent), which precludes the reductionist assumption that systemic properties of one system can be deduced from those of any other system. Consequently, social systems are emergent properties.

4. *The Renaissance of Ideas of Emergence in Late-Twentieth-Century Theoretical Thought*

In the latter part of the twentieth century, ideas of emergence have again attracted considerable attention in various scientific fields. Several new approaches, such as synergetics (Haken 1978) and other versions of self-organization theory (Krohn and Küppers 1992, Krohn et al. 1990) or chaos theory (Newman 1996), address interdisciplinary research questions at a relatively high level of abstraction that allude to problems raised in classical emergence theory. Examples of these research problems are the investigation of the spontaneous and self-confined emergence of new structures and systemic properties or the examination of predictability in chaotic dynamic processes. Interestingly enough, this body of research seldom refers to the tradition of emergence theory and the notion of emergence itself remains rather vague. This suggests that, for these theoretical approaches, the concept of emergent properties is used less as an analytical research tool and more as a guiding research idea.

This is most apparent in the booming field of theories of self-organization. Krohn and Küppers (1992), for example, provide a rather vague concept of emergent properties. They define emergence as the sudden appearance of novel properties that can be explained neither by the properties themselves nor by the relations of the elements involved, but rather by specific self-organizing process dynamics. To complicate matters, the notion of self-organization is also a shifting concept. Depending upon the strand of system theory, the meaning attributed to self-organization turns out to be completely different. Searching for the least common denominator, self-organization

may be understood as interactions between the components of a system and its organization that lead to modifications on both sides (Hejl 1990, p. 115). It follows that the dynamics of self-organization engender new structures endowed with emergent systemic properties. In theories of self-organization then, self-referential organization of systems and emergent properties are causally related. Within this theoretical framework, social phenomena, such as social networks and social movements, easily lend themselves to the conceptualization of emergent properties generated by self-organizing process dynamics.

The issues addressed in theories of self-organization discussed above may be generalized to include the problem of how to conceive of the relationship between microinteractions and macrostructures—in short, the micro–macro link. With respect to social theory, the great bulk of previous work adheres to either a bottom-up method (microreductionism) or a top-down method (systems theory). Both methods, taken by themselves, have their limits. The reductionist tradition in social theory must at least acknowledge systemic properties as marginal conditions of individual behavior. Vice versa, the collectivist tradition is often short of explaining many systemic properties. Against this background, the need for a circular movement of theory building when conceptualizing the micro-macro link (i.e., interlevel explanations) is evident. Hence, approaches that attempt to couple systems concepts with constituent concepts are in demand. A theoretical framework of this type 'simultaneously accounts for the behaviors of systems on one organizational level, the behaviors of their constituents on a lower level, and the relations between them' (Auyang 1998, p. 5). In this respect, the approach of synthetic microanalysis proposed by Auyang (1998) may be regarded as one example.

5. *Emergent Properties: Methodological Issues*

Methodologically, the question arises of how to assess emergent properties. The key to answering this question lies in the exploration of the link between the micro and macro levels of social phenomena. From this perspective, the claim that the whole is more than the sum of its parts amounts, first, to furnishing proof of the *specific assumptions* implied in either the 'strong' or 'weak' versions of emergence theory (see Sect. 1). According to the 'strong' version, properties at the higher levels of aggregation (i.e., the macro level) qualify as emergent, if they *cannot* be deduced or predicted from the properties of units at lower levels (i.e., the micro level). In the 'weak' version, they qualify, as long as they can be expressed as complex functions of the lower level characteristics. Second, the assessment of emergent properties must be based on evidence regarding the *impact* of system properties (i.e., contextual properties) on lower-level units and/or relationships that operate independently of the effects of any lower-level characteristics. Couched in more technical terms, the second requirement implies that the attributes of units operating at the micro level are a function of context (i.e., the macro level). The two tasks identified here to assess emergent properties address a measurement issue on the one hand and an issue about modeling techniques on the other. Each issue will be addressed in turn.

5.1 *The Measurement Issue*

'Strong' and 'weak' versions of emergence theory agree that higher-level properties that can be expressed as simple functions of lower-level attributes do not qualify as emergent. Examples include contextual characteristics defined as the sum, as a proportion or as the mean of lower-level properties (e.g., percentage of community residents above the age of 65 years; average income of regions). The two strands of emergence theory differ, however, with regard to the theses of deducibility and predictability. The 'strong' version claims that emergent properties are *global* attributes that are not expressible as functions of lower-level properties. This implies that system properties need to be measured at the system level as opposed to the level of system elements. One example of such an approach would be the description of political systems characterized by the degree of corporatism, whereby the measurement of corporatism would be based on expert judgement.

The measurement issue involved in 'weak' versions of emergence theory includes the application of complex rules of aggregation. Expressed as complex functions of lower-level attributes, system properties can be deduced or predicted from lower-level units. The income distribution characteristic of regional contexts—expressed by a complex function of individual incomes—serves as an example. A methodological perspective that is well suited to identifying emergent properties in line with 'weak' versions of emergence theory is *network analysis*, most prominently advocated by scholars such as Ronald S. Burt, Mark Granovetter and Barry Wellmann. This technique truly addresses the fundamental idea of the whole being more than the sum of its parts by focusing on *relational* characteristics and by introducing complex rules of aggregation. The social closure of elites, for example, cannot be inferred from the attributes of individual members of the elite. It is a property inferred from the *relational* attributes of the network members. Hence, the relational order of network elements (i.e., the structure) is a system property that truly qualifies as an emergent property. The second basic assumption of network analysis is the embeddedness of lower-level units in higher-level systems (i.e., networks) (see Granovetter 1985). This implies that lower-level outcomes (e.g., preferences, tastes, action orientations,

performance at the individual level) are mediated by higher-level system properties. This is to say that individual and collective actors respond to their social context. Couched in more technical terms, the notion of embeddedness addresses the issue of the additive and interactive effects of context on lower-level outcomes. This remark leads up to the second issue raised above, namely the modeling techniques required to assess emergent relationships.

5.2 The Impact Issue

To assess the effects of system properties measured at various levels of a hierarchical structure on lower-level outcomes, multilevel models (or contextual models) are the method of choice. A variety of models are used to specify the link between the macro and micro levels of social phenomena. While the names of the models and their statistical properties do differ (e.g., hierarchical linear model, random coefficients model, hierarchical mixed linear model), the basic idea is 'to analyze data that consist of multiple macro units (contexts) and multiple micro units within each macro unit' (DiPrete and Forristal 1994, p. 334). Within these models, two types of effect can be distinguished: main or additive effects and interactive effects. Main effects of higher-level properties on lower-level outcomes (e.g., behavior) provide evidence that context matters over and above the impact of any lower-level attributes. This indicates that the relationships identified through the effect analysis cannot be reduced to individual differences. Examples of substantive research along these lines include the contextual effects of schools and classes on student achievement or the impact of macro contexts, such as organizational structure and labor market segmentation, on career outcomes. The second type of effect involves interaction effects of properties measured at various levels of a hierarchical structure. The existence of cross-level interaction effects provides evidence of how higher-level properties (i.e., the macro context) affect the impact of lower-level attributes (i.e., those at the micro level) on the outcome of interest. Variation in the strength of the effect of micro-level properties on the outcome when system properties are included indicates the presence of relationships that can be qualified as emergent. The reason is that interaction effects involving cross-level components show that they are more than the sum of their parts. Substantive research employing this complex modeling strategy includes, for example, the ways in which country characteristics (e.g., economic development; policy programs) affect the influence of maternal education on the reproductive behavior of women.

See also: Durkheim, Emile (1858–1917); Evolution: Self-organization Theory; Law, Autopoiesis in; Luhmann, Niklas (1927–98); Mill, John Stuart (1806–73); Neural Circuits: Models of Emergent Functions; Parsons, Talcott (1902–79); Statistical Analysis: Multilevel Methods

Bibliography

Auyang S Y 1998 *Foundations of Complex-system Theories in Economics, Evolutionary Biology, and Statistical Physics*. Cambridge University Press, Cambridge, UK

Beckermann A, Flohr H, Kim J (eds.) 1992 *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Walter de Gruyter, Berlin

Bunge M 1977 Emergence and the mind. *Neuroscience* **2**: 501–9

DiPrete T A, Forristal J D 1994 Multilevel models: methods and substance. *Annual Review of Sociology* **20**: 331–57

Durkheim E 1938 [1895] *The Rules of Sociological Method*. Free Press, New York [*Les Règles de la Méthode Sociologique*. Presses Universitaires de France, Paris]

Granovetter M 1985 Economic action and social structure: the problem of embeddedness. *American Journal of Sociology* **91**: 481–510

Haken H 1978 *Synergetics*, 2nd edn. Springer, Heidelberg, Germany

Hejl P M 1990 Self-regulation in social systems. In: Krohn W, Küppers G, Nowotny H (eds.) 1990 *Self-organization: Portrait of a Scientific Revolution. Sociology of the Sciences*, Vol. XIV. Kluwer, Dordrecht, The Netherlands pp. 269–92

Krohn W, Küppers G (eds.) 1992 *Emergenz: die Entstehung von Ordnung, Organisation und Bedeutung*. Suhrkamp, Frankfurt am Main, Germany

Krohn W, Küppers G, Nowotny H 1990 *Self-organization: Portrait of a Scientific Revolution. Sociology of the Sciences*, Vol. XIV. Kluwer, Dordrecht, The Netherlands

Luhmann N 1995 [1984] *Social Systems*. Stanford University Press, Stanford, CA [*Soziale Systeme*: *Grundriss einer allgemeinen Theorie*. Suhrkamp, Frankfurt am Main]

Mill J S 1974 [1843] *A System of Logic: Ratiocinative and Inductive. Collected Works*, Vols. VII and VIII. University of Toronto Press, Toronto, ON

Newman D V 1996 Emergence and strange attractors. *Philosophy of Science* **63**: 245–61

Parsons T 1937 *The Structure of Social Action*. McGraw-Hill, New York

Parsons T 1951 *The Social System*. Free Press, New York

Popper K R, Eccles J C 1977 *The Self and Its Brain*. Springer, Heidelberg, Germany

Stephan A 1999 *Emergenz: von der Unvorhersagbarkeit zur Selbstorganisation*. Dresden University Press, Dresden, Germany

M. Buchmann

Emigration: Consequences for Social Structure in the Sending Population

International migration is the movement of people across national boundaries. Emigration is international migration from the perspective of the migrants' country of origin. Emigrant flows are the

movements of emigrants with common origins or destinations, and emigrant stocks refer to the number of emigrants present in a destination. In many uses of the term, emigration involves a change in residence that is long term, although there is no single accepted definition of the length of time that a person has to be out of the country of origin in order to be considered an emigrant. Tourist or business related trips are not counted as emigration, whereas movements that involve a permanent change in place of residence are. Between these two extremes lies considerable ambiguity in deciding when an out-of-country movement becomes emigration. The United Nations defines a long-term emigrant as 'a person who moves to a country other than that of his or her usual residence for a period of a least a year (12 months), so that the country of destination effectively becomes his or her new country of residence' (United Nations 1998, p. 10).

Part of the ambiguity in defining emigration arises from the presence of circular international labor migration. The development and expansion of international communication and transportation networks has stimulated and facilitated the movement of labor across national borders by increasing global access to information on living standards and employment opportunities in high-income countries, and by lowering the time and monetary costs of international travel. Labor recruitment activities targeted at workers in low-income countries and state-sponsored guest worker programs have played an important role in generating and sustaining the international flow of labor. The same communication and transportation networks that facilitate emigration to high-income countries also facilitate return migration to emigrant places of origin. Recent approaches to emigration recognize the importance of return flows of information, financial resources, and emigrants. This article addresses the consequences in sending populations of both emigrant outflows and the return flows that emigrants generate. The consequences are divided into three broad groupings: demographic, economic, and social.

1. Measuring Emigration

Measuring emigration is difficult because it is a dynamic, duration-dependent process—a migrant is not necessarily an emigrant at the time of departure, but rather becomes an emigrant after a period of time in the country of destination. Although it is possible to measure intended length of stay at the time of departure, migrant intentions often change and not all international moves are legally sanctioned and thus documented at the point of departure or arrival. Many governments do not register departures or collect information on international migration flows. Lack of comparability across countries in the migration statistics reported makes it difficult to derive precise estimates of emigrant stocks and flows.

A number of approaches have been used to derive rough estimates of emigration using census and survey data. When emigrants from a country go to a small number of foreign destinations, information on place of birth from censuses conducted in the countries of destination can be used to estimate emigrant stocks and flows. Zlotnik (1992) used recent census data from countries in the Americas to estimate the number of emigrants from Latin American countries living in other countries in the region. Similarly, Hatton and Williamson (1998) used census data in major migrant receiving countries to estimate gross emigration rates in the second half of the nineteenth and early twentieth centuries for selected European countries. A second approach to estimating the magnitude of emigrant flows makes use of the demographer's balancing equation:

$$E = P_1 - (P_0 + B - D + I)$$

The number of emigrants (E) is derived as the difference between the enumerated population at the end of the period (P_1), and the expected population produced by natural increase ($B - D$) and immigrant additions (I) to the initial population (P_0). Baines (1985) used a variant of this approach to estimate emigration from England and Wales in the late nineteenth century. A third approach to measuring emigration makes use of questions in household surveys about the international migration of current household members and nonresident relatives. Woodrow-Lafield (1996) used this approach to derive estimates of emigration from the USA, and Bean et al. (1998) used it to derive estimates of emigration from Mexico to the USA. The first two census-based approaches capture long-term emigration, but miss many short-term emigrants who are not present in the destination at the time of the census. The survey-based approach misses long-term emigrants who no longer have relatives in their community of origin, but is more effective at collecting information on short-term emigration and return flows.

2. Emigrant Flows

Migration scholars identify three periods of major emigration in the modern world (Massey et al. 1998). The first period, from 1500 to 1800, encompassed the flow of European emigrants into areas of colonization in the Americas, Africa, Asia, and Oceania, as well as the forced emigration of African slaves into the Americas. The precise number of European emigrants during this period is not known, but the number of African slaves was close to 10 million. The mass migration of Europeans to the Americas in the nineteenth and early twentieth centuries marks the

second major period of emigration. From 1800 to 1925, more than 48 million people emigrated from Europe in response to economic dislocations created by economic development at home and the availability of agricultural land and industrial employment overseas (Massey et al. 1998). Major emigrant-sending countries were the British Isles, Germany, the Scandinavian countries, Italy, Austria-Hungary, Poland, Russia, Spain, and Portugal. The third period of major emigration began in the 1960s and continues up to the present. Emigration in this period is marked by a major shift in origins. Most of the world's emigrants now come from Africa, Asia, and Latin America. Many of the former emigrant-sending countries in Europe have become important emigrant destinations.

Most emigrant flows are accompanied by significant counterflows of returning emigrants. The percentage of emigrants who eventually return to their country of origin varies substantially by period, and by country of origin, and destination. During the early periods of European overseas emigration return flows were relatively small because of the high cost of transportation. However, as the cost of intercontinental transportation came down, return emigration rose. At the end of the nineteenth and start of the twentieth centuries the number of immigrants who left the USA to return to their countries of origin was equivalent to 30 percent of new immigration. While very few Russian and Irish immigrants returned, nearly 50 percent of Italian and Spanish immigrants to the USA returned home (Hatton and Williamson 1998). In the current period, the number of return emigrants in many national contexts exceeds the number of emigrants who remain in the country of destination. Emigrants are drawn back to their countries of origin by strong social ties at home, limited integration into the host societies—including legal restrictions on settlement put in place by the destination government—and the high purchasing power of emigrant earnings in low-income countries of origin (Appleyard 1998a).

3. *Demographic Consequences of Emigration in the Sending Population*

Permanent emigration represents a loss of population in the place of origin. Whether this loss translates into a long-run reduction in population size depends on the age and sex composition of emigrant flows, and whether birth and survival rates rise in response to the loss of population. Malthusian population theory suggests that a reduction in the population pressure on existing resources through emigration could trigger a rise in birth and survival rates in the sending population. A demographic response of this type would cancel out or reduce the impact of emigrant loss on long-run population size. The development context within which most contemporary emigration originates makes a rise in birth rates an unlikely response to emigration. The same forces of economic development and modernization that stimulate emigration also operate to reduce fertility in the sending population. In general, countries that have experienced high emigration show lasting effects on future population size. For example, during the peak years of mass emigration, 1864–1914, emigration losses reduced potential population growth by 40 percent in Norway and by 25 percent in Sweden (Lowell 1987, p. 5). In Mexico, cumulative emigration to the United States has reduced the expected population of major emigrant-sending states by almost 50 percent (Verduzco and Unger 1998, p. 427).

The long-run effects of emigration on population size stem from the fact that emigration entails not only the loss of the emigrant but the loss of the emigrant's descendants as well. The emigration of women who are ready to begin childbearing has a larger impact on long-run population size than the emigration of women who have completed childbearing. When emigration is selective of men, total fertility may also get below what it would have been because a greater proportion of women never marry. Ireland is an example of a country that experienced high rates of non-marriage among women due to the emigration of men.

The loss of emigrants and their descendants is associated with long-term or permanent emigration. Short-term emigration has separate but potentially significant effects on future population size. When emigration is temporary and sex-selective, it disrupts family formation and the pace of childbearing by delaying marriage and separating couples. Van de Walle (1975) provides evidence from the Swiss canton of Ticino of how the seasonal migration of men in the late nineteenth and early twentieth centuries stemmed population growth and led to low fertility by increasing the proportion of women who never married and by separating husbands and wives during their childbearing years.

4. *Economic Consequences of Emigration in the Sending Population*

The economic consequences of emigration for the sending population stem primarily from the outflow of labor and the return flow of emigrant earnings (remittances). When emigration is voluntary and motivated by economic reasons, emigrants are overwhelmingly drawn from the working-age population. For the sending country emigration reduces the supply of labor, which in turn leads to lower unemployment and higher wages for those workers who remain. Hatton and Williamson (1998) find evidence that emigration from Ireland in the nineteenth century had

precisely this effect—the departure of labor improved the economic conditions of non-emigrants faster than would have been the case if emigration had not occurred. Similarly, contemporary emigration in South Asia, much of which is directed toward oil-producing countries in the Middle East, has pushed up wages for skills that have become more scarce due to the temporary absences of workers (Appleyard 1998b). In addition to raising wages, emigration may induce the adoption of more advanced labor-saving technologies, and thereby speed up economic development (Simon 1989).

The impact of emigration on the aggregate supply of labor depends in part on the labor force participation of emigrants prior to departure and on the response of households to the loss of an emigrant worker's labor. For example, in Sri Lanka close to two-thirds of emigrant workers are female household workers, many of whom were not in the labor force before emigrating. The effect of this emigration on unemployment in Sri Lanka is therefore difficult to measure because many of these women would not be in the domestic labor force if they had not emigrated (Gunatilleke 1998). Households may also respond to the emigration of a household worker by increasing the economic activity of remaining household members. This intrahousehold substitution of work diminishes the impact of emigrant departures on the labor market. For example, in countries where female labor force participation is low and emigration is selective of men, the substitution of women's labor for men's would alleviate any possible labor shortages created by emigration and would thus leave wage rates unchanged.

The emigration of highly skilled workers represents a loss for the sending country in terms of public investments made in their education and in terms of the economic growth that the emigrants would have generated if they had remained. Known as the brain drain, the emigration of highly educated and skilled workers is a cause of concern in emigrant-sending countries. The loss of potential new employment due to emigration is expected to be greatest when the skills possessed by emigrants and non-emigrants are complements of one another, that is, demand for one set of skills is dependent upon demand for the other. Not all emigration, however, represents a loss of skills for the sending country. Many emigrants return, some with new skills which lead to increased productivity at home. Of course, not all skills learned abroad are transferable to origin labor markets because of national differences in the level of development and in the demand for distinct skills. Many jobs performed by emigrants in high-income countries also provide little or no opportunity for acquiring new skills.

Each year emigrant workers send home, or return with, a portion of their earnings. The transfer of emigrant earnings (remittances) has important economic consequences in the sending population. At the macro-level, remittances from emigrant workers have been a 'life saver' for the economies of some countries. In Bangladesh and Pakistan, both major emigrant labor sending countries, remittances constituted about 70 percent of export earnings during the 1980s (Shah 1998, pp. 24–5). During the 1990s remittances amounted to 50 percent of exports in the Dominican Republic, and in El Salvador remittances were nearly as great as the volume of exports (Lowell and de la Garza 2000, p. 6). In addition to being an important source of foreign exchange, remittances are a substitute for scarce credit, and remittances spent on goods and services generate new employment. In an analysis of the effect of emigration and remittances on the labor market in El Salvador, Funkhouser (1992) found that the strongest effect of emigration occurs through remittances and not through the loss of emigrant workers. Money earned abroad and spent at home creates investment and employment opportunities in the provision of goods and services that remittances are used to purchase. The labor market impact of remittances will be greatest in the local areas where remittances are spent. Market towns in emigrant-sending regions, for example, may benefit more from remittances in terms of employment growth than rural emigrant-sending communities.

At the household level remittances significantly raise consumption levels and the overall standard of living. The most common use of remittances after covering day-to-day living expenses is investments in housing. Aside from meeting a basic need, putting remittances into residential real estate is a popular strategy for storing the value of saved emigrant earnings when other forms of investment or savings are not viable options. Residential real estate can be used to generate rental income, and it can be converted back to cash by being sold. The heavy use of remittances for household consumption as opposed to investment in economic production has lead to pessimistic assessments by some analysts of the capacity of remittances to stimulate economic development. One common argument is that labor emigration to high-income countries and the return flow of emigrants and remittances create a condition of dependence on emigrant earnings in sending communities. Studies conducted in emigrant-sending communities from diverse settings reveal substantial variation in the relative use of remittances for consumption and investment purposes, however. In some communities remittances have played a key role in agricultural modernization, in others remittances have spurred the development of cottage industries which produce goods for domestic and export markets. The use of remittances for productive investments appears to be influenced by local access to productive resources. Greater numbers of return emigrants use remittances for investments in agriculture and entrepreneurial activities in communities where the expected returns on productive investments are positive.

5. Social Consequences of Emigration in the Sending Population

Many of the social consequences of emigration in sending populations are related, first, to the disruption of family life produced by temporary out-of-country absences, and, second, the diffusion of new ideas and behaviors brought back by visiting and returning emigrants. The emigration of selected household members leaves households divided across national borders. One of the consequences of the disruption of family life is a change in definitions of family roles and role relations. For example, in countries where male household members emigrate and female members stay behind, married women often become *de facto* household heads during their husband's absence, and take on many of the roles otherwise reserved for men. The experience of greater autonomy in economic decision making can produce long-run changes in household gender roles. The prolonged absence of a spouse or parent also places families under stress, resulting in increased rates of family disintegration and divorce and a weakening of family control over youths. On the other hand the periodic absence of household heads, whether male or female, does not always have negative repercussions for families. In many low-income countries temporary labor emigration is a standard part of life, extending back several generations. Under these circumstances, family systems evolve and adapt coping mechanisms to meet the basic needs of emigrant and non-emigrant members (Thomas-Hope 1998). Kinship networks are often an important source of support for households with emigrant members.

During their visits to destination countries emigrants are exposed to new ideas, behaviors, and lifestyles that they gradually and selectively adopt. These new ideas, experiences, and behaviors are transmitted back to family and friends in the community of origin through correspondence, telephone conversations, and most importantly through visits and return migration. Levitt (1998) calls the ideas and norms that emigrants bring home with them, social remittances. 'They include norms for interpersonal behavior, notions of intrafamily responsibility, standards of age and gender appropriateness, principles of neighborliness and community participation, and aspirations for social mobility' as well as 'expectations about organizational performance such as how the church, state, or the courts should function' (Levitt 1998, p. 933). The circular flow of community members back and forth between the sending and destination communities produces continuous, sustained exposure to destination norms and behaviors. Emigrants and non-emigrants in places of origin filter, transform, and adapt these norms and behaviors to meet their own needs and circumstances.

See also: International Migration by Ethnic Chinese; International Migration by Ethnic Germans; International Migration by Ethnic Russians; International Migration by Ethnic South and Southeast Asians; Migration, Economics of; Migration into the Americas and between American Countries; Migration out of Europe; Migration: Sociological Aspects

Bibliography

Appleyard R (ed.) 1998a *Emigration Dynamics in Developing Countries, Vol. I: Sub-Saharan Africa*. Ashgate, Aldershot, UK

Appleyard R (ed.) 1998b *Emigration Dynamics in Developing Countries, Vol. II: South Asia*. Ashgate, Aldershot, UK

Baines D 1985 *Migration in a Mature Economy: Emigration and Internal Migration in England and Wales, 1861–1900*. Cambridge University Press, Cambridge, UK

Bean F D, Corona R, Tuirán R, Woodrow-Lafield K A 1998 The quantification of migration between Mexico and the United States. In: *Migration between Mexico and the United States: Binational Study*. Mexican Ministry of Foreign Affairs and US Commission on Immigration Reform, Austin, TX, Vol. 1, pp. 1–89

Funkhouser E 1992 Mass emigration, remittances, and economic adjustment: The case of El Salvador in the 1980s. In: Borjas G J, Freeman R B (eds.) *Immigration and the Work Force: Economic Consequences for the United States and Source Areas*. University of Chicago Press, Chicago, pp. 135–75

Gunatilleke G 1998 Macroeconomic implications of international migration from Sri Lanka. In: Appleyard R (ed.) *Emigration Dynamics in Developing Countries, Vol. II: South Asia*. Ashgate, Aldershot, UK, pp. 113–46

Hatton T J, Williamson J G 1998 *The Age of Mass Migration: Causes and Economic Impact*. Oxford University Press, New York

Levitt P 1998 Social remittances: Migration driven, local-level forms of cultural diffusion. *International Migration Review* **32**(4): 926–48

Lowell B L 1987 *Scandinavian Exodus: Demography and Social Development of 19th Century Rural Communities*. Westview Press, Boulder, CO

Lowell B L, de la Garza R 2000 *The Developmental Role of Remittances in U.S. Latino Communities and in Latin American Countries*. Inter-American Dialogue and Tomás Rivera Policy Institute, Claremont, CA and Washington, DC

Massey D S et al. 1998 *Worlds in Motion: Understanding International Migration at the End of the Millennium*. Clarendon, Oxford, UK

Shah N M 1998 Emigration dynamics in South Asia: An overview. In: Appleyard R (ed.) *Emigration Dynamics in Developing Countries, Vol. II: South Asia*. Ashgate, Aldershot, UK, pp. 17–29

Simon J 1989 *The Economic Consequences of Immigration*. Blackwell, Oxford, UK

Thomas-Hope E 1998 Emigration dynamics in the Anglophone Caribbean. In: Appleyard R (ed.) *Emigration Dynamics in Developing Countries, Vol. III: Mexico, Central America and the Caribbean*. Ashgate, Aldershot, UK, pp. 232–84

United Nations 1998 *Recommendations on Statistics of International Migration*. Statistical Papers, no. 58, rev. 1. United Nations, New York

Van de Walle F 1975 Migration and fertility in Ticino. *Population Studies* **29**(3): 447–62

Verduzco G, Unger K 1998 Impacts of migration in Mexico. In: *Migration between Mexico and the United States: Binational Study*. Mexican Ministry of Foreign Affairs and US Commission on Immigration Reform, Austin, TX, Vol. 1, pp. 395–435

Woodrow-Lafield K A 1996 Emigration from the USA: Multiplicity survey evidence. *Population Research and Review* **15**: 171–99

Zlotnik H 1992 Empirical identification of international migration systems. In: Kritz M M, Lim L L, Zlotnik H (eds.) *International Migration Systems: A Global Approach*. Clarendon, Oxford, UK, pp. 19–40

D. Lindstrom

Emotion and Expression

Emotion can be expressed in the face, voice, posture, and gesture. Since Darwin, most researchers have concentrated on facial expressions. The scientific study of these expressions has been an important site for disputes over the existence of a universal human nature and the extent to which human culture is grounded in biology.

1. *What is at Stake in Discussions of Facial Expression?*

The nature and origins of facial expressions of emotion may not seem at first glance to be a matter of momentous concern for the social sciences and humanities. However, the emotions are widely believed to be a critical feature of moral agency, and are even more widely believed to be a critical part of esthetic response. If all healthy people display, recognize, and respond to the same limited number of emotional expressions, this suggests a similar uniformity in emotion itself. In that case, moral and esthetic judgments may have universal validity. Conversely, if humans expressions of emotion are as diverse as the concepts embodied in different languages and if humans can only understand the expressive repertoire of their own cultural group, this would seem to support cultural relativism about emotion and thus about ethics and esthetics.

A related debate concerns the extent to which emotional expressions can be given evolutionary explanations. If so, then the emotions they express are presumably also part of the evolutionary inheritance of humanity. The idea that emotions are the products of evolution has been used to argue for a range of highly controversial claims. One argument runs from the premise that emotions have evolved to the conclusion that emotions are universal. Another argument runs from the premise that emotions have evolved to the conclusion that emotions are biologically determined in the sense that that they cannot easily be altered by changing environmental factors. Because of the critical role of emotions in human social interactions, biological determinism about emotions supports the view that certain forms of social organization are biologically inevitable.

2. *The Origins of Research on Facial Expression*

The scientific study of facial expression began with the nineteenth century anatomical investigations of Bell (1844) and Duchenne (1867/1990). The most influential work of the nineteenth century, however, was undoubtedly Darwin's *The Expression of the Emotions in Man and Animals* (Darwin 1872). The book started life as a chapter of Darwin's *Descent of Man* (1871) but grew so large that it became a separate work. In *Descent*... Darwin aimed to show evolutionary continuity between human ethical behavior and animal behavior and also, via his theory of sexual selection, between the human esthetic sense and the esthetic sense of animals. In *Expression*... he aimed to show evolutionary continuity between the human and animal emotions by demonstrating continuity between the facial expressions of humans and animals. Darwin explicitly targeted Bell's claim that the muscles of the human face were created by God to express human emotions. This led him to argue that many movements that now express emotion were vestiges of previous ways of life. 'With mankind some expressions, such as the bristling of the hair under the influence of extreme terror, or the uncovering of the teeth under that of furious rage, can hardly be understood, except in the belief that man once existed in a much lower and animal-like condition' (Darwin 1872, p. 12). Some authors have denied that Darwin ascribed any current adaptive functions to emotional expressions (e.g., Fridlund 1994), but, to take just one example, Darwin argued that humans, unlike chimpanzees, open their mouths when terrified because, unlike chimpanzees, humans must breathe through their mouths when fleeing.

Darwin's analysis of the contributions of the various facial muscles to emotional expressions drew heavily on Duchenne and Bell, but he made substantial original contributions, particularly with respect to homologies between human and animal expressions. His conclusions conform closely to currently accepted views of the muscular basis of facial expression. Darwin also pioneered the use of photographs, including some images made by Duchenne, to test the ability of observers to infer emotional state from isolated facial expressions. Certain photographs elicited very general agreement from Darwin's subjects and he took these to show the 'true' expressions of the associated emotions. Photographs relating to facial expression are shown in Fig. 1.

Regrettably, Darwin only used photographs in England and relied on a less satisfactory technique to

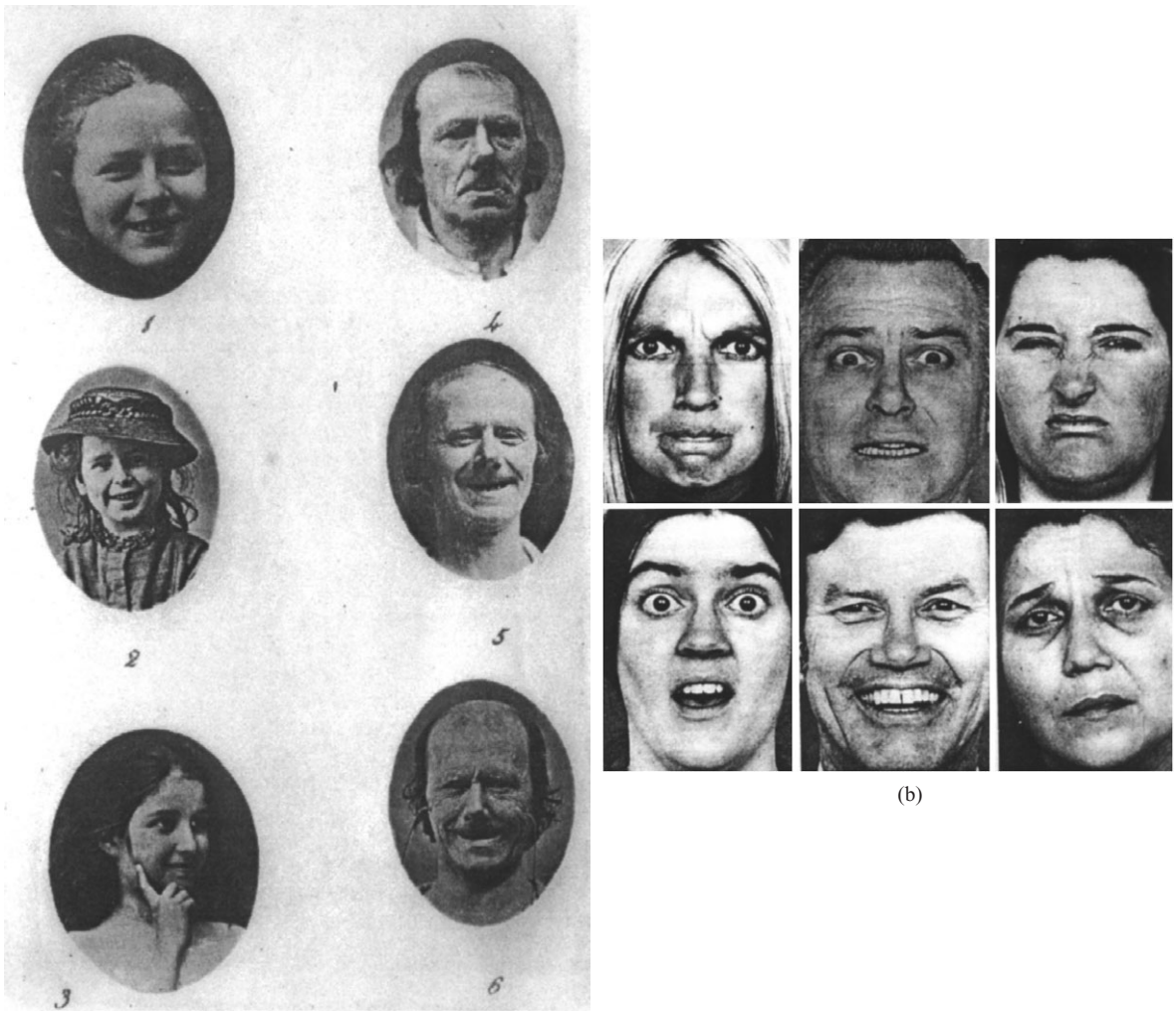

(a)

(b)

Figure 1
(a) Photographs used by Darwin to test the recognition of emotion from facial expression. 1–3, Degrees of mild amusement; 4, impassivity; 5, true smile; 6, false smile. Note the electrodes used by Duchenne to stimulate the muscles. (b) Some photographs used in Ekman and his collaborators' research on facial expression. From top left: anger, fear, disgust, surprise, happiness, and sadness. (Source: Ekman and Friesen 1975. © Paul Ekman. Reproduced with permission)

study the expression of emotion in other cultures. He dispatched questionnaires to missionaries, traders and others, asking them about the facial movements used to express emotion by the people amongst whom they lived. Unfortunately, Darwin's questionnaire suggested which facial components he expected to represent which emotions and thus made it very likely that his observers would read the desired results into their observations. Darwin's failure to use photographic evidence in his cross-cultural research is puzzling when considered against the background of the extensive contemporary interest in ethnographic photography.

3. *The Affect Program Theory of Facial Expression*

Darwin's work on emotional expression was neglected for the first half of the twentieth century. Anthropolo-

gists such as Weston La Barre (1947) and Ray Birdwhistell (1963) argued that culturally specific emotional states were signaled in a culturally specific code of facial expressions and gestures acquired by the individual during their upbringing. This culturalist tradition was displaced in the late 1960s by a powerful revival in the Darwinian approach. Many of Darwin's specific findings about facial expression were confirmed and extended, most notably by the psychologist Paul Ekman and his collaborators (Ekman 1972). One famous experiment used subjects from the Fore language group in New Guinea with a minimum of prior contact with Westerners and their cultural products. These subjects were given three photographs, each showing a face, and told a story which was designed to involve only one emotion. They were asked to pick the photograph showing the person in the story. This method has the advantage that no translation of the names of emotions is needed. The subjects were highly successful in picking the photograph intended to express the emotion appropriate for each story. The New Guinean subjects were also asked to act out the facial behavior of the people described in the stories. Videotapes of their responses were shown to US college students. The students were generally accurate in their judgements of the emotion intended by the New Guineans. The six facial expressions that figured in these experiments were labeled surprise, joy, sadness, fear, anger, and disgust. These, and a seventh previously confounded with disgust and now labeled contempt, have become the basis of Ekman's influential revival of the Darwinian approach to expression.

Ekman conceptualizes facial expressions as key components of affect programs, complex patterns of facial and other actions controlled by the central nervous system and triggered in response to some specific stimulus (Ekman 1984). This conceptualization derives, ultimately, from the idea of a fixed action pattern as it figures in the classical ethology of Konrad Lorenz and Niko Tinbergen. These founders of the modern study of animal behavior rediscovered Darwin's work on the emotions and interpreted it as prefiguring many of their own ideas (Lorenz 1965). Emotions, they argued, were innate patterns of behavior that had evolved by natural selection. Darwin's emphasis on the vestigial nature of many human facial expressions was interpreted using the ethological concept of ritualization (Tinbergen 1952). Patterns of behavior that once served a practical function, such as the opening of the mouth and retching exhibited in disgust, have been retained and modified by natural selection because of their value as signals. Hence the primary biological function of facial expressions in humans is as a signal of the person's emotional state. The distinctive experimental paradigm of classical ethology is the deprivation experiment, which aims to show that an animal produces a pattern of behavior even when it is deprived of experiences that would seem necessary for the behavior to be learnt. The ethologist Irenaus Eibl-Eibesfeldt (1973) examined the facial movements of children born deaf and blind and showed that these children used the same patterns of facial action as other children to display their emotions. He concluded that these behaviors were part of humanity's evolutionary heritage.

Ekman and his collaborators have handled cultural differences in the expression of emotion with the concept of a display rule, a concept exemplified in another of their experiments. Twenty-five American and 25 Japanese college students were shown neutral and stress-inducing films while alone in a room. The repertoire of facial behaviors shown during the stress phase by the two sets of subjects was very similar. However, when an experimenter was introduced into the room and allowed to ask questions about the subject's emotions as the stress film was shown again, the facial behavior of the Japanese diverged radically from that of the Americans. Video analysis showed the momentary occurrence of negative emotional expressions and their replacement with polite smiles. An important feature of the display rule conceptualization of cultural differences is that the evolved expressions remain intact but interact with culturally specific behaviors to determine the observable pattern of facial action.

4. *Universalism vs. Relativism*

The research leading to the affect program theory and subsequent discussion have helped to clarify some of the issues about the universality of emotion. The affect programs have the same output across cultures, but they do not have the same input. There are some universal elicitors of affect programs in childhood, such as very loud noises, which elicit fear. There are also systematic biases in the conditioning of affect program responses that could lead to a convergence in the eliciting conditions for adult responses. The general picture, however, is that affect programs come to be associated with whatever stimuli locally fulfill a broad functional role, so that the fear affect program comes to be associated with whatever locally constitutes a threat, the disgust response with whatever locally appears noxious or unclean, and so forth. Distinguishing between the universality of outputs and that of inputs makes it clear that at least some of dispute between universalists and cultural relativists about emotion is merely semantic. Universalists regard two cultures as manifesting the same emotion when they respond in the same way to different stimuli. Relativists insist that having the same emotion means responding in the same way to the same things.

Further clarification results from distinguishing the question of whether emotions are pancultural (found in all cultures) from the question of whether emotions are universal (found in all healthy individuals). The types of evidence normally gathered by universalists

are designed to show that emotions are pancultural and have little bearing on the question of universality. Emotions might be pancultural but still be like blood type or eye color, with several different types of individual in each population. Models of the evolution of social emotions typically predict that competing types will be maintained in the same population through competition. It is surprising that the issue of whether emotions have evolved is still so strongly linked to the issue of whether there is a single, universal, human emotional nature.

Finally, the debate over universality could be clarified by abandoning the last vestiges of the traditional dichotomy between learnt and innate behaviors. Some critics of the affect program theory have argued that a biological perspective on emotion is inappropriate merely because the emergence and maintenance of emotional responses depends upon environmental factors. Conversely, evidence that emotions are pancultural and thus likely to be the products of evolution is still thought to imply that these emotions are genetically determined and resistant to modification by environmental changes. These inferences ignore the facts that the environment plays a rich and constructive role in the development of even the most stereotypically biological traits, such as bodily morphology or sexual behavior. Evolved emotions, like the rest of evolved psychology, will likely make use of many reliable features of the environment of the developing child in order to construct and maintain themselves. They will be open to cultural and individual variation as a result of changes in these features, as well as through genetic variation. Griffiths (1997) has suggested that questions of universality can often be usefully reframed in terms of the biological concept of *homology*: are two emotions in different individuals or cultures homologous, that is, are they modified forms of a response in a common ancestor of those individuals or cultures?

5. *Contemporary Alternatives to the Affect Program Theory*

Cultural relativism about emotions, now known as social constructionism, has undergone something of a revival in recent years (Harré 1986). The older relativist tradition rested its case on anthropological fieldwork. The modern revival relies less on empirical data and more on conceptual argument. Many social constructionists start from the widely accepted idea that emotions involve a cognitive evaluation of the stimulus. They argue that emotions will inherit the cultural differences in the way emotion stimuli are conceived. If two cultures think differently about danger, then, since fear involves an evaluation of a stimulus as dangerous, fear in these two cultures will be a different emotion. This argument does not impress universalists, many of whom will allow that the two emotions are interestingly different but maintain that their similarities are more than sufficient to make them the same thing, namely fear. As well as purely semantic disputes of this kind, relativists and universalists focus on different parts of the domain of emotion. The affect program states seem to require minimal cognitive evaluation of the stimulus via pathways that the neuroscientist Joseph le Doux has christened the *low road* to emotion (Le Doux 1996). Social constructionists often refuse to regard these isolated physiological responses as emotions, reserving that term for the broader cognitive state of a person involved in a social situation in which they might be described as, for example, angry or jealous. It is thus unclear whether the debate between contemporary social constructionists and their universalist opponents has a substantial factual dimension, or whether it reflects only a preference on the one side for tractable, reductive explanations and a desire on the other side to forestall the scientific neglect of the social, transactional aspects of human emotion.

The main contemporary alternative to the affect program theory is the paralanguage theory of emotional expressions, a view whose leading exponents are Fridlund (Fridlund 1994) and James A. Russell (Russell and Fernendez-Dols 1997). Paralanguage theorists argue that the affect program theory exaggerates the automated nature of emotional expressions and neglects their voluntary, performative aspect. They deny that facial expressions are involuntary expressions of discrete emotional states. Paralanguage theorists have accumulated considerable evidence of audience effects on the production of emotional expressions. They take this to show that expressions are produced for the benefit of the audience rather than to express the feelings of the person making the display. They question the affect program view of emotions themselves, as well as of emotional expression, arguing that emotional states vary continuously along several axes, rather than falling into a few discrete types. Paralanguage theorists argue that the experimental results of Ekman and his collaborators are largely artifactual. They are particularly critical of the use of forced choices from a list of emotion terms dictated by the experimenter. This procedure forces a range of responses to a range of facial signals into a few boxes defined by the experimenter. The critics argue that this cannot provide evidence for the existence of a limited number of stereotyped facial expressions, since this assumption is built into the experimental procedure.

Although paralanguage theorists criticize the experiments usually taken to establish the universality of emotion, some of them accept that emotional expressions have evolved and would therefore presumably expect the flexible signaling system they postulate to be at least substantially pancultural. Fridlund argues that an evolutionary perspective actually favors the

paralanguage view, since involuntary signals of true emotional state would be subverted by the evolution of dissimulation and deceit. While this is an important perspective on the evolution of emotional expressions, it does not constitute the decisive argument that Fridlund seems to suppose. Veridical signals do evolve in nature, often by making use of so-called *hard to fake* signals. The cost of being unable to suppress a signal of emotional state may be balanced by the advantage of being believed. A purely theoretical argument based on the evolutionary dynamics of signaling systems seems more likely to support the view that 'examples of emotional behavior lie along a continuum from expression to negotiation' (Hinde 1985, p. 989) than a purely expressive or purely manipulative picture of emotional expression.

6. *Conclusion*

Facial expressions of emotion are an important site for the debate between universalism and relativism because both sides believe that here they can offer substantial evidence for their positions. An examination of this evidence need not serve only to support one or other side of the debate. The actual details of emotional expression, its evolution, and its development in the individual can be used to question the validity of the dichotomies between biology and culture, fixity and malleability, universality, and idiosyncraticity which are assumed by both sides.

See also: Adulthood: Emotional Development; Culture and Emotion; Emotion: History of the Concept; Emotion in Cognition; Emotion, Neural Basis of; Emotions and Health; Emotions, Evolution of; Emotions, History of

Bibliography

Bell C 1844 *Anatomy and Philosophy of Expression as Connected with the Fine Arts*, 3rd edn. Murray, London

Birdwhistell R L 1963 The kinesic level in investigations of the emotions. In: Knapp P H (ed.) *Expression of the Emotions in Man*. International Universities Press, New York, pp. 123–39

Darwin C 1871 *The Descent of Man*. Murray, London

Darwin C 1872 *The Expressions of Emotions in Man and Animals*, 1st edn. Philosophical Library, New York

Duchenne G B 1867/1990 *The Mechanism of Human Facial Expression*. Cambridge University Press, Cambridge, UK

Eibl-Eibesfeldt I 1973 Expressive behaviour of the deaf and blind born. In: von Cranach M, Vine I (eds.) *Social Communication and Movement*. Academic Press, London, pp. 163–94

Ekman P 1972 *Emotions in the Human Face*. Pergamon, New York

Ekman P 1984 Expressions and the nature of emotions. In: Scherer K R, Ekman P (eds.) *Approaches to Emotion*. Erlbaum, Hillsdale, NJ, pp. 319–43

Ekman P, Friesen W V 1975 *Unmasking the Face: A Guide to Recognising Emotions from Facial Expressions*. Prentice Hall, Englewood Cliffs, NJ

Fridlund A J 1994 *Human Facial Expression: An Evolutionary View*. Academic Press, San Diego, CA

Griffiths P E 1997 *What Emotions Really Are: The Problem of Psychological Categories*. University of Chicago Press, Chicago

Harré R 1986 *The Social Construction of the Emotions*. Blackwell, Oxford

Hinde R A 1985 Was 'Expression of the Emotions' a misleading phrase? *Animal Behaviour* **33**: 985–92

La Barre W 1947 The cultural basis of emotions and gestures. *Journal of Personality* **16**: 49–68

Le Doux J 1996 *The Emotional Brain*. Simon and Schuster, New York

Lorenz K 1965 *Preface to The Expression of the Emotions in Man and Animals by Charles Darwin*. University of Chicago Press, Chicago

Russell J A, Fernandez-Dols J M (eds.) 1997 *The Psychology of Facial Expression*. Cambridge University Press, Cambridge, UK

Tinbergen N 1952 Derived activities: their causation, biological significance, origin and emancipation during evolution. *Quarterly Review of Biology* **27**: 1–32

P. E. Griffiths

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Emotion: History of the Concept

This article deals with the history of the concept of emotion, not with the validity or acceptability of particular interpretations or theories of emotion. It will, of course, be inevitable that the discussion will deal with theories of emotion, that is, where the concept is lodged. The emphasis is, in part, on the historical situation and development of these theories, which were situated in the sociocultural temper of the times. The story starts with a survey of ancient Greek thought and its successors, mostly in the context of insightful premature speculations that, in the fashion of the times, mix quasiscientific thought with practical, in this case ethical, considerations. We then jump to the Renaissance and René Descartes, with the not unexpected claim that the new sciences were replacing all that went before. This is followed by the scientism of the nineteenth century leading up to the last great philosopher-psychologist, William James, who was able to combine humanistic insights with scientific rigor. James and Wilhelm Wundt lead us into the modern period and the need for analysis and synthesis.

Early intellectual history was preoccupied with rationality, and when emotions were discussed they tended to be identified more with the magic and mysteries of human life. The Orphic mysteries dominated much of thinking; submission to the supernatural involved too many emotional attitudes to make it possible to speak clearly about feelings and emotions. In addition discussions of the feelings and emotions—when they did occur—were confounded with considerations of ethics and aesthetics. Emotions were

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

discussed in the context of how to control and how properly to use them. They were part of a philosophy of life, and codes of behavior used the emotions as their foundation. This was especially true of the pre-Socratic thinkers but persisted long afterwards, after the Aristotelian interruption.

With notable exceptions, and Aristotle stands out among them, these discussions of the emotions from pre-Socratic to modern, i.e., Cartesian, days centered on their distracting influences on the intellect, on ways of controlling them (see, for example, Empedocles and Heraclitus). Later on, the emphasis in Plato and the early Greeks is on the intellect, how we know, and what we can know. But the theme of the distracting emotions was not forgotten and even medieval scholastics such as St. Thomas Aquinas were concerned with the way emotions interfere with pure thought and how one might control them.

In the *Timaeus* Plato had postulated three different faculties of the mind—thought, spirit, and appetite. This compartmentalization, and the previous positioning of the emotions as distinct from the mental, intellective faculties, was discarded by Aristotle. He unified the mind and its effects and influences. Mind cannot be conceived without the body and—importantly—the converse is asserted as well. This biological context allows Aristotle to consider feelings as natural phenomena. But Aristotle does not allow for simple, pure, affective processes. Rather he requires a cognitive element of a percept or idea, as well as an affective component of pleasantness/unpleasantness, and a conative effort or activity (cf. Hammond 1902).

The middle ages were in the hands of the scholastics and speculations about the emotions relegated them often to mere expressions of animal spirits and to a position very distinct from the spirit and the intellect. With the advent of Descartes and the modern age all that changed. In 1649 in his *Les passions de l'âme* (*The passions of the soul*), Descartes discarded the work of the ancients as 'defective' and emphasized the modernist credo to start afresh. Whereas his predecessors frequently had speculated on the number and character of the emotions that constituted the emotional life, Descartes was the proper seventeenth-century scientist and postulated six primary passions, with all the rest constructed of those six: wonder, love, hate, desire, joy, and sadness. This approach to passions has stayed with us unabated for three and a half centuries, though Descartes's love, desire, and wonder have fallen by the wayside to be substituted by other, more contemporary states. And the centuries have failed to solve the question of 'how many?'—the answer ranges between three and 18 (Ortony and Turner 1990).

The next major change in the conception of emotion came later in the seventeenth century with Spinoza's detailed analyses of the various emotions based on the insistence that they must be taken as natural and lawful phenomena rather than as a bothersome intrusion into mental life. In the next century Kant followed that trend and gave feelings the status of a special class of psychical processes—a distinction that held well into the twentieth century. Much of the nineteenth century was devoted to the detailed classification of the emotions—a refinement of the enumeration approach. Dominant among the classifiers was Wundt, whose system went from simple to complex feelings and then to true emotions. The composite feelings are often temporary; when permanent they represent moods. Combining composite feelings into a unitary whole results in the emotions. This simplified version does not do justice to a system where an emotion such as grief may be analyzed in terms of a half dozen or more types and tokens of feelings. The late nineteenth century occasioned several equally complex and in the end ineffective classificatory schemes. All this was brought to an end by a theory that arose in three different countries and dominated the psychology of emotion for decades to come—the James–Lange–Sergi theory.

The theory is generally attributed to William James, though the name of the Danish physiologist Carl Georg Lange is frequently included. James published his theory in 1884 in the journal *Mind*, followed by some elaboration in his *Principles of Psychology* in 1890 and an important defense in 1894. Lange published in a monograph in 1885 in Danish and two years later in German, and the Italian Guiseppe Sergi presented his ideas in a book on pleasure and pain in 1894, followed by a more specific treatment in 1896. There is no indication that any of the three had much knowledge of the others when they presented their ideas. In brief, James claimed that bodily reactions (including both autonomic and skeletal ones) were produced by some inciting fact and that the perception/feeling of these changes is the emotion. Lange similarly noted that the emotions were nothing more that the sensations produced by the vasomotor system, and Sergi noted that visceral/autonomic functions are stimulated directly and produce the emotions. And those three were not alone; there were similar suggestions from different parts of the globe. What was it about the end of the nineteenth century that engendered this particular point of view in such disparate places? Probably the commitment to an explicit materialism that pervaded scientific thinking. In psychology such an approach demanded implicitly that mental states should—whenever possible—be derived from observable physical events. The road went straight from physiology to experience, and it was not until the turn of the century that psychological theory languages were developed *sui generis*. In the process the definition of James's exciting facts that generated bodily reactions was neglected, though current thought has caught up with the question and primarily assigns these to cognitive evaluations. (For a discussion of these developments and the psychology of emotion during the subsequent century see Mandler 1979.)

With the advent of the commanding Jamesian position, the philosophy of emotion became an adjunct rather than the *avant garde* of emotion theory. For the next century philosophers would list, from different vantage points, specific factors as essential to emotion, such as: certain feelings, positive and negative cognitions, bodily sensations, facial and bodily expressions, and action tendencies. Increasingly psychological evidence was used to buttress one or another position. In the flavor of the century and in keeping with the complex common-sense meaning of emotion a Wittgensteinian position has often been invoked, which gives emotions the characteristic of family resemblances with one or more of the above factors relevant from case to case.

The Jamesian position demanded that for each different emotional experience there should be an appropriately different bodily/visceral precedent. However, a century of studies has failed to show any evidence that different emotional states are preceded by different bodily states. That fact was central in the extensive and crucial critique of the James–Lange position by Cannon (1927). Even though James did not restrict 'bodily events' to the viscera, Cannon noted that visceral autonomic changes were not differentiated or fast enough to account for emotional experiences, nor can visceral changes alone produce emotional experiences. It is this last observation that starts the bridge to the current scene.

In 1924, the Spanish physician Marañon had shown that productions of sympathetic nervous system arousal (by the injection of adrenalin) produced either descriptions of bodily symptoms or 'as if' reports, experimental subjects claiming they felt as if they experienced some emotions. Stanley Schachter picked up the thread imaginatively in 1962 when he (with Jerome Singer) showed that adrenalin injections together with the experimental induction of cognitive states related to specific emotions produced true emotional states or experiences. Essentially the Schachter experiments established what has since been called the 'constructivist' approach to emotional states. The states are the result of potentially independent cognitive and visceral/arousing events experienced as subjectively unified conscious emotions. The notion that cognitive factors played an important part in emotional life had been considered since Aristotle, but now became close to established gospel. The cognitive opening also made possible a new and wide variety of analyses of emotional phenomena and during the last quarter of the twentieth century many different theories of emotion emerged. (For a review of the state of affairs before the proliferation of theories, see Arnold 1970.) More important perhaps was the realization that these evaluative cognitions may well differ from one culture to another (see, for example, Lutz and Abu-Lughod 1990). A field that had been relatively quiescent during the earlier decades of the century erupted with new theories, several new journals, and new organizations. Emotion became—in some of these approaches—all-embracing, with some mental states devoid of any visceral involvement becoming candidates for emotional status. Given the large number of possible cognitive states, enumeration of emotional states became popular once again, though James, for example, had already argued that any such attempt reflected subjective sociocultural prejudices. As the number of theories multiplied, current fashions dictated neurophysiological projections as well as evolutionary speculations.

When the core of the emotion concept moved from ethics into the realm of science there arose an increasing insistence on the importance of visceral, i.e., sympathetic nervous system, activity as part of any reasonable definition of this common language concept. The nineteenth century showed that such activity can be dissociated from the other, cognitive, component of emotional experience. We are left with a question raised specifically in recent years: what—apart from an interminable list—in principle produced the visceral arousal. Questions about a principled source of all emotions had been raised in the nineteenth century by John Dewey in the US and Frederic Paulhan in France who noted that emotions are preceded by conflict, for example, among action tendencies. Current speculations about the source of visceral arousal have come from both psychological and physiological sources. Psychologically it has been proposed that any event or situation that is inconsistent with the individual's expectations (such as to continue some action or perceive an expected sequence) produced SNS arousal. What is proposed is a difference detector linked to the sympathetic nervous system (Mandler 1984, 1992). Physiologically, the research program has been in response to the question 'how is the initial state of bodily arousal evoked?' The most promising proposal is that 'the amygdala detects and organizes responses to natural dangers (like predators) and learns about novel threats and the stimuli that predict their occurrence' (LeDoux 1998).

Neurophysiology has been a widely favored discipline during the last decade of the twentieth century, and recent proposals have mirrored that approach to a psychology of emotion. At the same time there has been an increasing awareness of cultural determinants of different emotional states, and the concurrent realization that so-called cognitive determinants are reflections of sociocultural forces. Is it sheer *fin de siècle* optimism that we can look toward a new synthesis of a neuropsychology of arousal and a cultural analysis of cognitive values?

See also: Culture and Emotion; Emotion and Expression; Emotion, Neural Basis of; Emotions, Evolution of; Emotions, History of; Emotions, Psychological Structure of; Emotions, Sociology of; Intentionality and Rationality: A Continental-European Perspec-

tive; Intentionality and Rationality: An Analytic Perspective

Bibliography

Arnold M B (ed.) 1970 *Feelings and Emotions: The Loyola Symposium*. Academic Press, New York

Cannon W B 1927 The James–Lange theory of emotions: A critical examination and an alternative theory. *American Journal of Psychology* **39**: 106–24

Hammond W A (ed.) 1902 *Aristotle's Psychology: A Treatise on the Principles of Life (De anima and Parva naturalia)* [trans. Hammond W A, Sonnenschein S]. London

LeDoux J 1998 Fear and the brain: Where have we been, and where are we going? *Biological Psychiatry* **44**: 1229–38

Lutz C A, Abu-Lughod L 1990 *Language and the Politics of Emotion*. Cambridge University Press, Cambridge, UK

Mandler G 1979 Emotion. In: Hearst E (ed.) *The First Century of Experimental Psychology*. L Erlbaum Associates, Hillsdale, NJ

Mandler G 1984 *Mind and Body: Psychology of Emotion and Stress*. WW Norton, New York

Mandler G 1992 Emotions, evolution, and aggression: Myths and conjectures. In: Strongman K T (ed.) *International Review of Studies on Emotion*. Wiley, Chichester, UK

Ortony A, Turner T J 1990 What's basic about basic emotions? *Psychology Review* **97**: 315–31

G. Mandler

Emotion in Cognition

A view held by influential philosophers in the eighteenth century, and inherited by psychologists in the nineteenth century, was that mental life could be separated into three mental faculties: conation (will and action), cognition (thinking and knowing), and affect (emotion). In the twentieth century this distinction was reflected in two movements that each gave priority to one of these faculties: behaviorism to conation, cognitive psychology to cognition. In both affect was neglected. Recently it has become clear that neither conation nor cognition can be understood without understanding the influences of affect.

In this article the following is assumed: emotion (affect) and cognition can be distinguished for purposes of analysis, though in practice they are seldom separate. Emotion is often used as a superordinate category that includes both specific emotions such as happiness and anger as well as moods such as anxiety. Specific emotions are typically elicited by events, and usually have known objects: for instance one becomes angry at a particular person for a known reason. Moods are longer lasting and may be objectless (free floating) but their underlying mechanisms are the same as those of specific emotions. Cognition, which includes both conscious and unconscious processes, is the representation and use of knowledge in language, thinking, remembering, attending, perceiving, and the like. To understand emotion in cognition we can ask, therefore, how a specific emotion such as anger affects reasoning, how the fear-based mood of anxiety affects attention, and so forth.

1. *The Idea that Emotions Interfere with Rational Cognition*

Emotion has often been regarded as antipathetic to cognition as in the clichéd adage 'the heart versus the head,' as if emotions were storms that blow the ship of selfhood from its rationally calculated course.

In the Western intellectual tradition, suspicion of emotions was perhaps most clearly articulated by Darwin (1872), who argued that emotional expressions are actions that occur whether or not they are any use. He included the bristling of hairs when angry or frightened, and the curling of a lip on one side of the mouth in contempt which he thought was a vestige of baring the teeth in snarling preparation to bite. Had he been writing today, he would have pointed to the smiles and gestures people make when they speak on the telephone. For Darwin, emotional expressions were those actions that derive from our evolution as beasts, or from our individual development from infancy. They are leftovers from earlier times, which cannot be inhibited by taking thought: behavioral equivalents of the appendix in our digestive system which in distant ancestors had a function but which, in adult humans, has a function no longer. The implication is that such expressions, and by implication the emotions themselves, have a regrettable influence on adult cognition. In a notebook Darwin wrote: 'Our descent, then, is the origin of our evil passions!!—The Devil under form of baboon is our grandfather.' Although he would never have published this sentiment, the argument in his 1872 book is consonant with it. The atavistic quality Darwin ascribed to emotional expressions gave biological credence to a distrust of emotions as enemies of reason that had been voiced by Plato, then nurtured by the Stoics and many subsequent thinkers.

2. *Functional Conceptions of Emotion*

There is, however, another line of argument. It is that emotions are the essential guides of cognition. In Western thought, the line derives from Aristotle, who argued that emotions are evaluations: beliefs about what we value. This is the view that has prevailed in modern cognitive psychology. So an emotion is a psychological process that mediates between goals (valued concerns) and events. It is typically elicited by an appraisal of an event in the world that affects an

important goal or concern (see e.g., Oatley and Jenkins 1996).

Aristotle's main treatment of emotions is in his book *Rhetoric*, which is not so much about persuading people by sly means, as about reaching useful conclusions socially when no exact demonstration of the correct course of action can be made. If there were an exact demonstration—an unequivocal argument on some matter—says Aristotle, we should follow it. But without such a demonstration, there must be discussion, as occurs for instance in a law court or in politics. Now emotions become essential. They are aroused in people who listen to the case on this side and that. 'The emotions,' said Aristotle, 'are all those feelings that so change [people] as to affect their judgements, and are also attended by pain or pleasure. Such are anger, pity, fear, and the like, with their opposites.'

Most problems of practical human life have no fully rational solutions. Therefore different people need to represent different points of view, including their emotional implications. This occurred recently in political and media discussions of the Yugoslavian conflict: on one side the tragic fate of Kosovan refugees, on the other the terrifying effects of bombing on civilians. In such discussions, conclusions are reached that are not ideal, but are better than those reached without considering emotional implications or without thinking in a dialogical way.

Since the advent of cognitive science in the late 1950s, psychological conceptions of thinking have been dominated by the idea of search across a space of possibilities. In chess, for instance, one searches ahead across moves and their continuations. In reasoning, one searches across manipulations of a problem to reach a solution. But, as quickly became clear, most search spaces for real problems quickly become indefinitely large, so they elude simple search processes. Research attention was then drawn to heuristics that help to prune search trees.

Following especially from a paper of Simon (1967), emotions came to be seen as heuristics of this kind. Simon's argument was that although, at that time, computer processes ran without interference until they terminated, in any realistic example of cognition every possibility cannot be foreseen, and so unexpected events occur. A system is therefore needed to interrupt ongoing processes, so that a more urgent process can take control. Such a system corresponds with what we think of as emotion: the process of managing cognition in environments which, like the ones we live in, are never fully predictable.

2.1 Three Classes of Problem for Human Rationality

Simon stressed the unpredictability that derives from limitations of resources, but really the problem is more extensive. Outcomes of actions are not fully predictable for three classes of reasons. First, there is indeed the issue of resources: human mental models are always incomplete and often incorrect; time, movement, strength, etc. are also always limited. Hence the unexpected often occurs. Second, goals are often multiple and sometimes incommensurable, as is made clear every time the career aspirations of a politician are threatened by media exposure of questionable sexual goals. Moreover, even where goals are commensurable, as for instance when mediated by money, there is no principled way of deciding whether to maximize gains or minimize losses. Third, the human biological arrangement is to offset limitations of resources (the first point) by distributing cognition and action among several agents. This has allowed a huge net gain in the possibilities of agency as compared with what could be accomplished by the individual, as may be seen in the products of human culture. But the arrangement introduces a new domain of unpredictability because our knowledge of beliefs and plans of other agents with whom we interact is also limited. In general, then, although we should strive for rationality, fully rational thought and action in most of life is impossible.

A solution to these three classes of problems has derived from evolution: the system of human emotions as heuristics that manage priorities of action in environments that are never certain, when there are multiple goals, and when we interact with other people.

2.2 Definition of Emotion

A working definition of emotion is due to Frijda (1986): an emotion occurs when an event is appraised as relevant to a goal. It is a form of readiness approximately appropriate to the event that elicited it. It gives priority to some goals rather than others, and an urge towards action of a certain kind. It is often accompanied by distinctive types of thought, expression, and physiological response, and it can interrupt, or compete with, alternative courses of action. An example: imagine you are reading quietly at home when the phone rings with news of an accident to a close friend. You experience a hollow feeling in the stomach, and in anxiety you put all thoughts of reading aside. Mental processing is switched to the urgent concern for your friend, and you make plans about what you might do to help.

2.3 Emotions and Optimization of Cognitive Processing

Darwin's book on *Expression* can be read as a treatise on optimization of behavioral mechanisms that were each adapted to a particular environment at a par-

ticular phase of evolution or development. Behavioral mechanisms are acquired by accretion, and in general these mechanisms cannot be deleted. Any mechanism, then, may or may not be useful in future environments, and may or may not be adaptable to new purposes. Darwin saw that the adaptations of evolutionary change have occurred in a piecemeal fashion, over extended periods. His principal concern was to argue against the idea that humans are perfect creations, so he emphasized aspects of behavior that in adult human life are imperfectly adapted. A century later, the argument for evolution had been won. Simon accepted it, saw the optimization problem more generally, and recognized the functional place of emotions within it.

Emotions, on this view, are not antipathetic to cognition; they are its managers within a system of only-partly-adapted cognitive-behavioral mechanisms that operate in a world where perfect solutions are rare. Emotions are not especially enemies of rationality. Rather, human rationality is like an archipelago in a huge and changing sea. We can rationally choose to take the shortest route to some destination, rather than setting off in the opposite direction. As self-interested economic agents we can rationally choose the lowest price for a product. As friends, we can rationally do what we have promised, knowing that someone depends on us. But the sea of ignorance of the exact effects of our actions is large. Emotions are heuristics that bridge across the straits between the islands of what we know rationally how to do.

3. *Experimental Studies of Effects of Emotions on Cognition*

Since the 1970s, experimental studies of effects of emotions have been made in increasing numbers. A pioneer was Isen who induced mild happiness in people by testing a skill and telling them they had done well, or by giving them small gifts, or by having them watch comedy films. She found several types of effect. As compared with people in neutral moods, those made happy (a) were more cooperative for instance in making larger donations to charity and in helping strangers, (b) did better in some intellectual tasks such as creative problem solving, and (c) recalled happy rather than unhappy memories. One of Isen's most striking studies (Isen et al. 1987) was to have subjects watch a funny or a neutral film. She then asked them individually to solve a problem: to use only a box of tacks, a book of matches, and a candle, to fix the candle to a corkboard wall and light it. Less than 20 percent of subjects who watched the neutral film solved the task, but in one experiment 75 percent and in another experiment 58 percent of subjects made happy by the funny film solved the problem within 10 minutes. The tacks did not work to pin the candle directly to the wall. The solution was to empty the box of tacks, pin the box to the wall, and use it to hold the candle. Happiness, therefore, made cognition more flexible. In terms of optimization, happiness minimizes issues of safety (which are maximized by fear and anxiety), and it improves the search for unusual associations.

At about the same time Bower (1981) was experimenting with inducing emotions by hypnosis. In one experiment, he asked subjects to recall incidents from their childhood, moving haphazardly around their memory for 10 minutes and briefly describing each incident they recalled. Next day he asked them to classify the recalled incidents as pleasant, unpleasant, or neutral. On the following day, he induced either a happy or sad mood, and asked the subjects to recall again the incidents they had recalled previously and later categorized. He found that subjects who were put into a happy mood recalled substantially more incidents they had classed as pleasant compared with unpleasant. Subjects in a sad mood remembered slightly more of the unpleasant than the pleasant incidents.

Explanations for this type of phenomenon have shifted. Bower at first thought that the explanation was of state dependency: incidents of any arbitrary kind that are memorized in one mood state are more available for recall when the same mood state recurs. Results of many studies now imply an explanation called mood congruence: autobiographical incidents with a particular emotional meaning to the subject are more available to recall when that same emotion recurs.

Effects of the mood of anxiety have also been studied. Whereas sadness has large effects on memory, anxiety has its largest effects on attention. A typical effect is shown in the so-called Stroop task in which words are printed in different colored inks. In Stroop's original task, people were asked to look at words such as 'red,' 'yellow,' and 'blue,' and to name the color of the ink in which each was printed. People were significantly slowed wherever a word, e.g., 'red,' was different from the color in which it is printed, e.g., in yellow ink. The explanation is that attention required for the task of naming the ink colors is captured by the meanings of the words. In the emotional Stroop task, the words are not of colors but they are emotionally significant: words like 'cancer,' 'robbery,' and so forth. Where such words have special emotional significance for subjects, attention is captured, and such subjects are slowed in saying the ink color of the words. For instance, Foa et al. (1991) found that people who had been victims of rape were slowed in saying the ink colors of words related to rape compared with ink colors of words that were not so related. In general, methods such as this have indicated that people who are anxious attend selectively to issues of their anxious concern.

Many other laboratory findings and principles have been established. Emotion has also been shown to affect perception, categorization, and the represen-

tation of knowledge. Perhaps most widely it has been shown to affect tasks that involve reasoning and social judgements (Forgas 2000). Principles include the finding that happy emotions tend to make for cognition that uses short cuts or schematic solutions, whereas negative emotions tend to make for more careful, step-by-step, processing (Schwarz and Bless 1991). In terms of maximization and minimization, in environments that are experienced as safe and friendly, cognitive processes are switched to maximizing gains according to current goals, to cooperative plans, or to assimilating new information to current knowledge structures. Negative emotions signal that something has gone wrong in progress toward a goal. Processing is switched so that schemata can accommodate, and careful attention is given to events of the environment, and to consistency of internal models.

It is clear that the relations among emotions and cognition need to be understood at several different levels of processing (Teasdale 1999). It is also clear that interactions between emotion and cognition can be approached using experimental methods in brain sciences as well as in psychology. For instance, it has been found that people with lesions of the prefrontal region have experimentally demonstrable defects in emotional processing; they also show impairments in social planning as it affects friendships, employment, and the making of social arrangements (Damasio 1994). Although this association between emotion and social planning is consistent with much cognitive research on emotions, neurological findings such as this have been influential in enabling acceptance of emotions as guides of cognition rather than disturbers of rationality.

4. *Research in Psychiatric Epidemiology*

Though experimental studies in this area are robust, many of them neglect people's most pressing emotional concerns. This domain has, however, been explored in studies of people's cognitive-affective responses to emotionally important events and problems in their lives. Perhaps most significant in this area has been the finding by Brown and Harris (1978) of how clinical depression is typically caused. When people become depressed, they experience a range of cognitive changes, and typically come to regard life as hopeless and unsustainable, so that they cannot cope with daily tasks, and contemplate giving up life altogether. The immediate causes of such breakdowns are severely adverse events such as bereavement from loved spouses, children, or parents, or severe protracted difficulties such as marital conflict, or unemployment and consequent lack of income.

Comparably, people become clinically anxious in response to severe threats to their way of life and to their systems of previous belief. Although this is true for many kinds of clinical anxiety states, most striking in this regard is the syndrome of post-traumatic stress disorder, in which people who suffer war or civil disaster, have their lives and value systems upturned. Such people's cognitive systems switch chronically to states of hypervigilance, accompanied by vivid memory flashbacks to the disaster that occurred.

Cross-connections between experimental studies and research in psychiatry have been made (see, e.g., Power and Dalgleish 1997). For instance, depressed people have been found to suffer biases of memory towards recalling incidents of loss and defeat, and people who are clinically anxious show biases of attention that can be demonstrated on tasks such as the emotional Stroop test (see, e.g., Mathews 1993).

5. *Conclusion*

For a long time cognition was studied as if it involved merely the impartial representation and use of knowledge. But emotions are, as Aristotle argued, evaluations that mediate between concerns and the events that impinge on knowledge. They are guides of cognition, selecting among beliefs, arranging priorities among goals, directing attention, biasing access to remembered material, and providing heuristics that influence reasoning, judgement, and planning.

See also: Culture and Emotion; Emotion, Neural Basis of; Emotions, Evolution of; Emotions, Psychological Structure of; Human Cognition, Evolution of; Mood-dependent Memory; Social Cognition and Affect, Psychology of

Bibliography

Aristotle c. 330 BC *Rhetoric and Poetics*, [Trans. Roberts WR]. Random House, New York

Bower G H 1981 Mood and memory. *American Psychologist* **36**: 129–48

Brown G W, Harris T 1978 *Social Origins of Depression: A Study of Psychiatric Disorder in Women*. Tavistock, London

Damasio A R 1994 *Descartes' Error*. G P Putnam, New York

Darwin C 1872 *The Expression of the Emotions in Man and Animals*. Murray, London

Foa E B, Feske U, Murdoch T B, Kozak M J, McCarthy P R 1991 Processing of threat related information in rape victims. *Journal of Abnormal Psychology* **100**: 156–62

Forgas J P (ed.) 2000 *Feeling and Thinking. The Role of Affect in Social Cognition*. Cambridge University Press, New York

Frijda N H 1986 *The Emotions*. Cambridge University Press, Cambridge

Isen A M, Daubman K A, Nowicki G P 1987 Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology* **52**: 1122–31

Mathews A 1993 Biases in emotional processing. *The Psychologist* **6**: 493–9

Oatley K, Jenkins J M 1996 *Understanding Emotions*. Blackwell, Cambridge, MA

Power M, Dalgleish T 1997 *Cognition and Emotion: From Order to Disorder*. Psychology Press, East Sussex, UK

Schwarz N, Bless H 1991 Happy and mindless, but sad and smart? The impact of affective states on analytic reasoning. In: Forgas J P (ed.) *Emotion and Social Judgment*. Pergamon, Oxford, pp. 55–71

Simon H A 1967 Motivational and emotional controls of cognition. *Psychological Review* **74**: 29–39

Teasdale J D 1999 Multi-level theories of cognition–emotion relations. In: Dalgleish T, Power M J (eds.) *Handbook of Cognition and Emotion*. Wiley, Chichester, UK, pp. 665–81

K. Oatley

Emotion, Neural Basis of

In order to provide a fundamental basis for understanding the neural basis of emotion, it is useful to have a clear view of the events that give rise to emotions and the functions of emotion, for these help to define the input systems in the brain involved in producing emotions, and the output systems in the brain that are involved in expressing emotion. With this basis, developed fully by Rolls (1999, 2000a) the neural systems that underlie emotions are then considered.

1. *The Events that Give Rise to Emotions*

Emotions can usefully be defined as states produced by rewards and punishers. Rewards are stimuli for which an animal (including humans) will work, and punishers are stimuli that an animal will work to escape from or avoid. Rewards and punishers are reinforcers in that they alter the probability of behavior. Some stimuli are unlearned or primary reinforcers (e.g., the taste of food if the animal is hungry, or pain). Other stimuli may become reinforcing by learning, because of their association with such primary reinforcers, thereby becoming 'secondary reinforcers.' For example, fear is an emotional state which might be produced by a sound (the conditioned stimulus) that has previously been associated with a painful stimulus (the primary reinforcer.)

The different emotions can be described and classified according to whether the reinforcer is positive or negative, and by the reinforcement contingency. An outline of such a classification scheme is shown in Fig. 1.

The operation of the following factors enables a very wide range of different emotions to described:

(a) The reinforcement contingency (see Fig. 1).

(b) The intensity of the reinforcer (see Fig. 1).

(c) Any environmental stimulus might have a number of different reinforcement associations. For example, a stimulus might be associated both with the presentation of a reward and of a punishment, allowing states such as conflict and guilt to arise.

(d) Emotions elicited by stimuli associated with different primary reinforcers will be different.

(e) Emotions elicited by different secondary reinforcing stimuli will be different from each other (even if the primary reinforcer is similar).

(f) The emotion elicited can depend on whether an active or passive behavioral response is possible. For example, if an active behavioral response can occur to the omission of a previously rewarded stimulus, then anger might be produced, but if only passive behavior is possible, then sadness, depression, or grief might occur.

By combining these six factors, it is possible to account for a very wide range of emotions. It is also worth noting that

(a) emotions can be produced just as much by the recall of reinforcing events as by external reinforcing stimuli;

(b) cognitive processing (whether conscious or not) is important in many emotions, for very complex cognitive processing may be required to determine whether environmental events are reinforcing or not;

(c) emotions normally consist of cognitive processing which determines the reinforcing valence of the stimulus, and an elicited mood change if the valence is positive or negative; and

(d) and that stability of mood implies that absolute levels of reinforcement must be represented over moderately long time spans by the firing of mood-related neurons, a difficult operation which may contribute to 'spontaneous' mood swings, depression which occurs without a clear external cause, and the multiplicity of hormonal and transmitter systems which seem to be involved in the control of mood.

2. *The Functions of Emotion*

The functions of emotion can be summarized as follows.

(a) The *elicitation of autonomic responses* (e.g., a change in heart rate) *and endocrine responses* (e.g., the release of adrenaline). These prepare the body for action.

(b) *Flexibility of behavioral responses to reinforcing stimuli*. Emotional (and motivational) states allow a simple interface between sensory inputs and motor outputs. The new understanding is that some genes specify which stimuli are rewarding and punishing, and these stimuli are the goals for flexible actions. For example, genes specify that strong somatosensory stimuli produce a state of pain which animals (including humans) are built to perform actions to avoid or escape. The state produced by the painful stimulus is an emotional state which has the adaptive function of identifying and maintaining the goal for an action. This is an efficient design, for the genes need only specify the goals for actions, and not the actions themselves, so that the actions can be selected flexibly to meet the goals. It would be very inflexible and

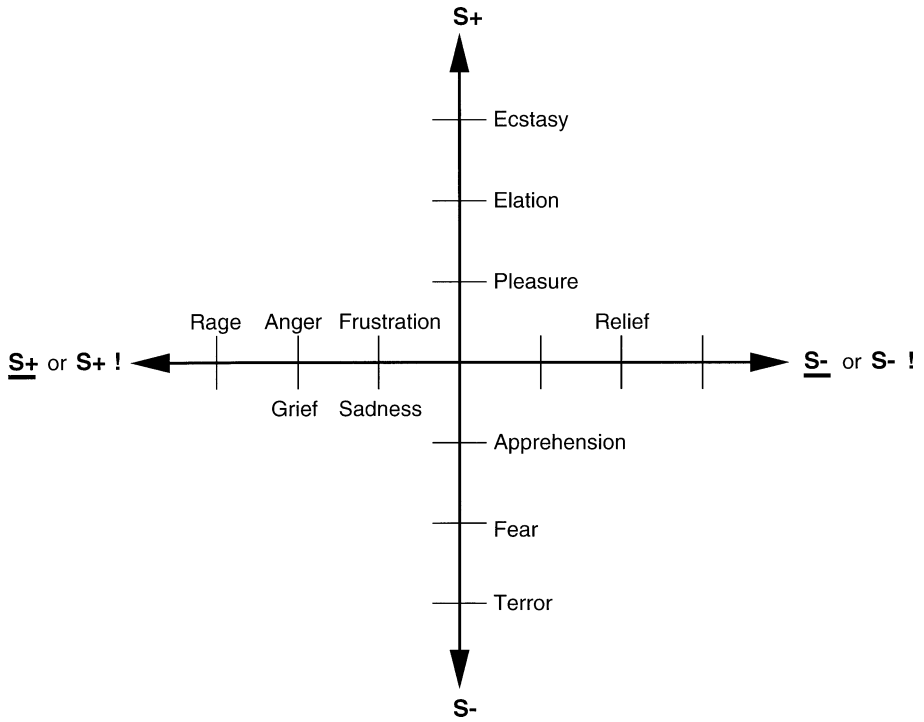

Figure 1
Some of the emotions associated with different reinforcement contingencies are indicated. Intensity increases away from the centre of the diagram, on a continuous scale. The classification scheme created by the different reinforcement contingencies consists of (a) the presentation of a positive reinforcer (S+), (b) the presentation of a negative reinforcer (S−), (c) the omission of a positive reinforcer (S+) or the termination of a positive reinforcer (S+!), and (d) the omission of a negative reinforcer (S−) or the termination of a negative reinforcer (S−!).

inefficient for the genes to specify actions themselves. In this way, emotions provide an efficient interface between sensory systems and actions. The hypothesis is that our emotions correspond to the goal-specifying states arising from stimuli that genes have found it adaptive and possible to specify. Rolls (1999) provides in Table 10.1 the start of a list of some of the primary reinforcers that genes appear to specify, and thus an idea of how a wide range of different emotions have evolved because of their adaptive implications.

(c) Emotion is *motivating*. For example, fear learned by stimulus-reinforcement association formation provides the motivation for actions performed to avoid noxious stimuli.

(d) *Communication*. For example, monkeys may communicate their emotional state to others, by making an open-mouth threat to indicate the extent to which they are willing to compete for resources, and this may influence the behavior of other animals. This aspect of emotion was emphasized by Darwin (1872).

(e) *Social bonding*. Examples are the emotions associated with the attachment of the parents to their young, and the attachment of the young to their parents.

(f) The current mood state can affect the *cognitive evaluation of events or memories* (see Oatley and Jenkins 1996), and this may have the function of facilitating continuity in the interpretation of the reinforcing value of events in the environment.

(g) Emotion may facilitate the *storage of memories*. One way in which this occurs is that episodic memory (i.e., one's memory of particular episodes) is facilitated by emotional states. This may be advantageous in that storing many details of the prevailing situation when a strong reinforcer is delivered may be useful in generating appropriate behavior in situations with some similarities in the future. A second way is that the current emotional state may be stored with episodic memories, providing a mechanism for the current emotional state to affect which memories are recalled. A third way in which emotion may affect the storage of memories is by guiding the cerebral cortex in the representations of the world which are set up. For example, in the visual system, it may be useful to build

perceptual representations or analyzers which are different from each other if they are associated with different reinforcers, and it may be less likely to build them if they have no association with reinforcement.

(h) Another function of emotion is that by enduring for minutes or longer after a reinforcing stimulus has occurred, it may help to produce *persistent and continuing motivation and direction of behavior*, to help achieve a goal or goals.

(i) Emotion may trigger the *recall of memories* stored in neocortical representations. Amygdala back-projections to the cortex could perform this for emotion in a way analogous to that in which the hippocampus could implement the retrieval in the neocortex of recent (episodic) memories (Rolls and Treves 1998).

(j) Rewards and punishers, and the emotional states they produce, provide a common currency for the behavior selection process between competing alternative actions.

3. Brain Mechanisms of Emotion

3.1 Overview

Animals are built with neural systems that enable them to evaluate which environmental stimuli, whether learned or not, are rewarding and punishing, that is will produce emotions and will be worked for or avoided. Sensory stimuli are normally processed through several stages of cortical processing to produce a sensory representation of the object before emotional valence is decoded. For example, in the taste system, taste is analyzed in primates to provide a representation of what the taste is in the primary taste cortex, and this representation is independent of the reward value of the taste in that it is not affected by hunger. In the secondary taste cortex, in the orbitofrontal region (see Fig. 2), the reward value of the taste is represented, in that neurons only respond to the taste if the primate is hungry. In another example, in the visual system representations of objects which are view-, position- and size-invariant are produced in the inferior temporal visual cortex after many stages of cortical processing, and these representations are independent of the emotional valence of the object.

Then in structures such as the orbitofrontal cortex and amygdala, which receive from the inferior temporal visual cortex, associations are learned between the objects and the primary reinforcers associated with them, by the process of stimulus-reinforcement association learning. In the orbitofrontal cortex and amygdala emotional states are thus represented.

Consistent with this, electrical stimulation of the orbitofrontal cortex and amygdala is rewarding, and damage to these structures affects emotional behavior by affecting stimulus-reinforcement association learning. These brain regions influence the selection of behavioral actions through brain systems such as the ventral striatum and other parts of the basal ganglia.

3.2 The Amygdala

The amygdala receives information about primary reinforcers (such as taste and touch), and also information about visual and auditory stimuli from higher cortical areas (such as the inferior temporal cortex) that can be associated by learning with primary reinforcers (Fig. 2). Bilateral removal of the amygdala in monkeys produces tameness, a lack of emotional responsiveness, the excessive examination of objects, often with the mouth, and the eating of previously rejected items such as meat (the Kluver-Bucy syndrome). In analyses of the bases of these behavioral changes, it has been observed that there are deficits in learning to associate stimuli with primary reinforcement, including both punishments and rewards. The association learning deficit is present when the associations must be learned from a previously neutral stimulus (e.g., the sight of an object) to a primary reinforcing stimulus (such as the taste of food).

Further evidence linking the amygdala to reinforcement mechanisms is that monkeys will work in order to obtain electrical stimulation of the amygdala, that single neurons in the amygdala are activated by brain-stimulation reward of a number of different sites, and that some amygdala neurons respond mainly to rewarding stimuli, and others to punishing stimuli. The association learning in the amygdala may be implemented by associatively modifiable synapses from visual and auditory neurons on to neurons receiving inputs from taste, olfactory or somatosensory primary reinforcers (LeDoux 1996). Consistent with this, Davis (2000) has found that at least one type of associative learning in the amygdala can be blocked by local application to the amygdala of a NMDA receptor blocker, which blocks long-term potentiation (LTP), which is a model of the synaptic changes that underlie learning (see Rolls and Treves 1998). Consistent with the hypothesis that the learned incentive (conditioned reinforcing) effects of previously neutral stimuli paired with rewards are mediated by the amygdala acting through the ventral striatum, amphetamine injections into the ventral striatum enhanced the effects of a conditioned reinforcing stimulus only if the amygdala was intact (see also Everitt et al. 2000).

An interesting group of neurons in the amygdala (in, for example, the basal accessory nucleus) responds primarily to faces. They are probably part of a system which has evolved for the rapid and reliable identification of individuals from their faces, and of facial expressions, because of the importance of this in primate social behavior. Consistent with this, activation of the human amygdala can be produced in neuroimaging studies by some face expressions, and

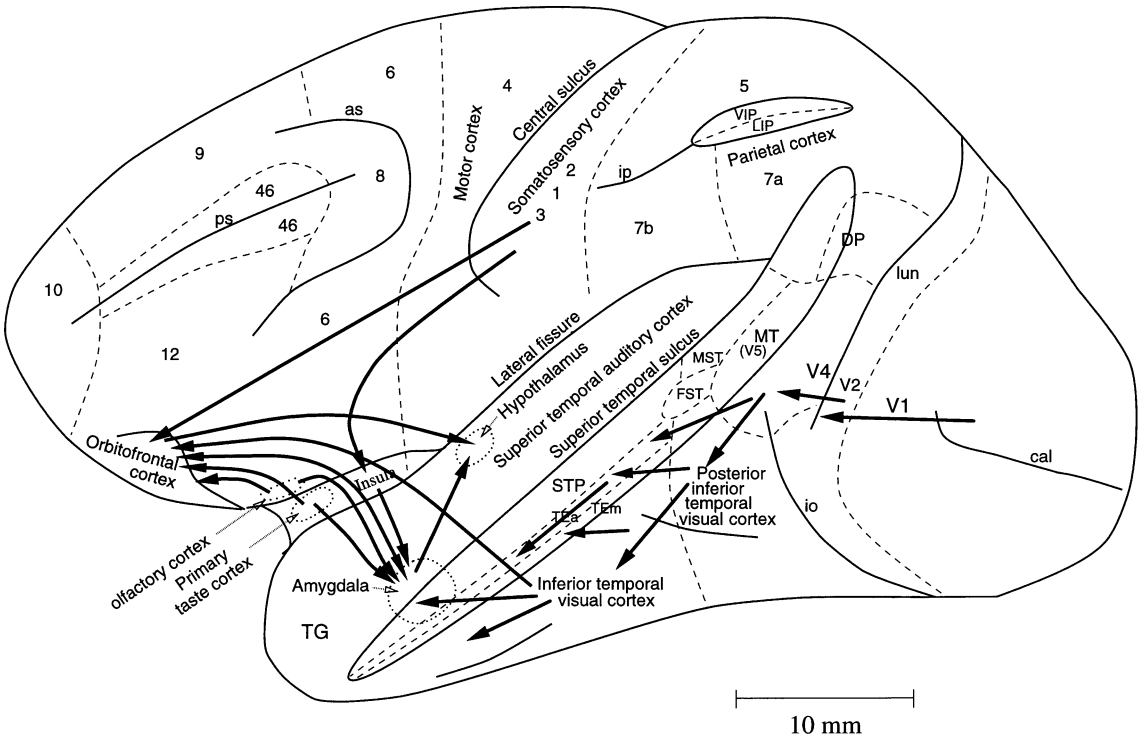

Figure 2
Some of the pathways involved in emotion described in the text are shown on this lateral view of the brain of the macaque monkey. Connections from the primary taste and olfactory cortices to the orbitofrontal cortex and amygdala are shown. Connections are also shown in the 'ventral visual system' from V1 to V2, V4, the inferior temporal visual cortex, etc., with some connections reaching the amygdala and orbitofrontal cortex. In addition, connections from the somatosensory cortical areas 1, 2 and 3 that reach the orbitofrontal cortex directly and via the insular cortex, and that reach the amygdala via the insular cortex, are shown. as, arcuate sulcus; cal, calcarine sulcus; cs, central sulcus; lf, lateral (or Sylvian) fissure; lun, lunate sulcus; ps, principal sulcus; io, inferior occipital sulcus; ip, intraparietal sulcus (which has been opened to reveal some of the areas it contains); sts, superior temporal sulcus (which has been opened to reveal some of the areas it contains). AIT, anterior inferior temporal cortex; FST, visual motion processing area; LIP, lateral intraparietal area; MST, visual motion processing area; MT, visual motion processing area (also called V5); PIT, posterior inferior temporal cortex; STP, superior temporal plane; TA, architectonic area including auditory association cortex; TE, architectonic area including high order visual association cortex, and some of its subareas TEa and TEm; TG, architectonic area in the temporal pole; V1—V4, visual areas 1—4; VIP, ventral intraparietal area; TEO, architectonic area including posterior visual association cortex. The numerals refer to architectonic areas, and have the following approximate functional equivalence: 1,2,3, somatosensory cortex (posterior to the central sulcus); 4, motor cortex; 5, superior parietal lobule; 7a, inferior parietal lobule, visual part; 7b, inferior parietal lobule, somatosensory part; 6, lateral premotor cortex; 8, frontal eye field; 12, part of orbitofrontal cortex; 46, dorsolateral prefrontal cortex

lesions of the human amygdala may cause difficulty in the identification of some face expressions.

3.3 The Orbitofrontal Cortex

The orbitofrontal cortex receives inputs from the inferior temporal visual cortex and superior temporal auditory cortex; from the primary taste cortex and the primary olfactory (pyriform) cortex (see Fig. 2); from the amygdala, and from the midbrain dopamine neurons. Damage to the caudal orbitofrontal cortex in the monkey produces emotional changes. These include decreased aggression to humans and to stimuli such as a snake and a doll, and a reduced tendency to reject foods such as meat. These changes may be related to a failure to react normally to and learn from nonreward in a number of different situations. This failure is evident as a tendency to respond when responses are inappropriate, for example no longer rewarded.

For example, monkeys with orbitofrontal damage are impaired on Go/NoGo task performance (in which they should make a response to one stimulus to obtain a reward, and should not make a response to another stimulus in order to avoid a punishment), in that they Go on the NoGo trials. They are also impaired in an object-reversal task in that they respond to the object which was formerly rewarded with food. They are also impaired in extinction in that they continue to respond to an object which is no longer rewarded. Further, the visual discrimination learning deficit shown by monkeys with orbitofrontal cortex damage, may be due to the tendency of these monkeys not to withhold responses to nonrewarded stimuli.

The primate orbitofrontal cortex contains neurons which respond to the reward value of taste (a primary reinforcer), in that they only respond to the taste of food when hunger is present (which is when food is rewarding). It also contains neurons which learn to respond to visual stimuli associated with a primary reward such as taste, and which reverse their responses to another visual stimulus in one trial when the rewards and punishers available from those visual stimuli reverse. Further, these visual responses reflect reward, in that feeding the monkey to satiety reduces the responses of these neurons to zero. Moreover, in part of this orbitofrontal region, some neurons combine taste and olfactory inputs, in that they are bimodal, and are in 40 percent of cases affected by olfactory-to-taste association learning and by feeding the monkey to satiety, which reduces the reward value. In addition, some neurons in the primate orbitofrontal cortex respond to the sight of faces. These neurons are likely to be involved in learning which emotional responses are currently appropriate to particular individuals, and in making appropriate emotional responses given the face expression.

It has been found that another class of neurons in the orbitofrontal cortex of the monkey responds in certain nonreward situations. For example, some neurons responded in extinction, immediately after a lick had been made to a visual stimulus which had previously been associated with fruit juice reward, and other neurons responded in a reversal task, immediately after the monkey had responded to the previously-rewarded visual stimulus, but had obtained punishment rather than reward. Another class of orbitofrontal neuron responded to particular visual stimuli only if they were associated with reward, and these neurons showed one trial stimulus-reinforcement association reversal. Another class of neuron conveyed information about whether a reward had been given, responding, for example, to the taste of sucrose, or for other neurons of saline.

These types of information may be represented in the responses of orbitofrontal neurons because they are part of a mechanism which evaluates whether a reward is expected, and generates a mismatch (evident as a firing of the nonreward neurons) if reward is not obtained when it is expected (see Rolls 1999, 2000a, 2000b). These neuronal responses provide further evidence that the orbitofrontal cortex is involved in emotional responses, particularly when these involve correcting previously learned reinforcement contingencies, in situations which include those usually described as involving frustration.

It is of interest and potential clinical importance that a number of the symptoms of frontal lobe damage in humans appear to be related to this type of function; altering behavior when stimulus-reinforcement associations alter, as described next. Thus, humans with frontal lobe damage can show impairments in a number of tasks in which an alteration of behavioral strategy is required in response to a change in environmental reinforcement contingencies (Rolls et al. 1994; Damasio 1994). Some of the personality changes that can follow frontal lobe damage may be related to a similar type of dysfunction. For example, the euphoria, irresponsibility, lack of affect, and lack of concern for the present or future which can follow frontal lobe damage may also be related to a dysfunction in altering behavior appropriately in response to a change in reinforcement contingencies. At one time, following a report by Moniz (1936), prefrontal lobotomies or leucotomies (cutting white matter) were performed in humans to attempt to alleviate a variety of problems, and although irrational anxiety or emotional outbursts were sometimes controlled, intellectual deficits and other side effects were often apparent (see Valenstein 1974). Thus these operations have been essentially discontinued.

To investigate the possible significance of face-related inputs to orbitofrontal visual neurons described above, the responses to faces that were made by patients with orbitofrontal damage produced by pathology or trauma were tested. Impairments in the identification of facial and vocal emotional expression were demonstrated in a group of patients with ventral frontal lobe damage who had socially inappropriate behavior (Hornak et al. 1996). The expression identification impairments could occur independently of perceptual impairments in facial recognition, voice discrimination, or environmental sound recognition. Thus the orbitofrontal cortex in humans appears to be important not only in the rapid relearning of stimulus-reinforcement associations, but also in representing some of the stimuli, such as face expression, which provide reinforcing information. Consistent with this, neuroimaging studies in humans are showing representations which reflect the pleasantness of the taste and smell of food, the pleasantness of touch, and also of quite abstract rewards and punishers such as winning or losing money (O'Doherty et al. 2001.)

The behavioral selection system must deal with many competing rewards, goals, and priorities. This selection process must be capable of responding to many different types of reward decoded in different brain systems that have evolved at different times,

even including the use, in humans, of a language system to enable long-term plans to be made. These many different brain systems, some involving the implicit (unconscious) evaluation of rewards, and others explicit, verbal, conscious, evaluation of rewards and planned long-term goals, must all enter into the selector of behavior. Although poorly understood, the issue of emotional feelings are part of the much larger problem of consciousness, and may involve the capacity to have thoughts about thoughts, that is higher order thoughts (see Rolls 1999, 2000a).

See also: Adulthood: Emotional Development; Amygdala (Amygdaloid Complex); Emotion and Expression; Emotion in Cognition; Emotions, Evolution of; Emotions, Psychological Structure of; Emotions, Sociology of; Infancy and Childhood: Emotional Development; Intrinsic Motivation, Psychology of

Bibliography

Darwin C 1872 *The Expression of the Emotions in Man and Animals* 3rd edn. University of Chicago Press, Chicago

Damasio A R 1994 *Descartes' Error*. Putnam

Davis M 2000 The role of the amygdala in conditioned and unconditioned fear and anxiety. In: Aggleton J P (ed.) *The Amygdala: Second Edition. A Functional Analysis*. Oxford University Press, Oxford, UK, Chap. 6, pp. 213–87

Everitt B J, Cardinal R N, Hall J, Parkinson J A, Robbins T W 2000 Differential involvement of amygdala subsystems in appetitive conditioning and drug addiction. In: Aggleton J P (ed.) *The Amygdala: Second Edition. A Functional Analysis*. Oxford University Press, Oxford, UK, Chap. 10, pp. 353–90

Hornak J, Rolls E T, Wade D 1996 Face and voice expression identification in patients with emotional and behavioral changes following ventral frontal lobe damage. *Neuropsychologia* **34**: 247–61

LeDoux J E 1996 *The Emotional Brain*. Simon and Schuster, New York

Moniz E 1936 *Tentatives opératoires dans le traitement de certaines psychoses*. Masson, Paris

Oatley K, Jenkins J M 1996 *Understanding Emotions*. Blackwell, Oxford, UK

O'Doherty J, Kringelbach M L, Rolls E T, Hornak J, Andrews C 2001 Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience* **4**: 95–102

Rolls E T 1999 *The Brain and Emotion*. Oxford University Press, Oxford, UK

Rolls E T 2000a Précis of *The Brain and Emotion*. *Behavioral and Brain Sciences* **23**: 177–233

Rolls E T 2000b Neurophysiology and functions of the primate amygdala, and the neural basis of emotion. In: Aggleton J P (ed.) *The Amygdala: Second Edition. A Functional Analysis*. Oxford University Press, Oxford, UK, Chap. 13, pp. 447–78

Rolls E T, Hornak J, Wade D, McGrath J 1994 Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery and Psychiatry* **57**: 1518–24

Rolls E T, Treves A 1998 *Neural Networks and Brain Function*. Oxford University Press, Oxford, UK

Valenstein E S 1974 *Brain Control. A Critical Examination of Brain Stimulation and Psychosurgery*. Wiley, New York

Young A W, Hellawell D J, Van de Wal C, Johnson M 1996 Facial expression processing after amygdalotomy. *Neuropsychologia* **34**: 31–9

E. T. Rolls

Emotional Inhibition and Health

1. Introduction

Emotional inhibition constitutes part of a maladaptive interaction between an individual and their social environment. Emotional inhibition may be classified into genetic, repressive, suppressive, and deceptive inhibition and the extreme form emotional implosion (Traue 1998). Overt emotional inhibition generally is characterized by reduced expressiveness, unemotional language, and shyness, all of which is related to dysfunctional bodily reactions and may be adaptive in a short-term social stress situation. In the long run, emotional inhibition is likely to have a harmful effect on the individual, which may occur along any of three pathways: neurobiological, social-behavioral, and cognitive. Most psychotherapeutic techniques are directed at emotional behavior and experience and focus largely on changes in intra- and interindividual emotional regulation and the construction of meaning from emotional experience.

2. Emotion and Inhibition

Emotions may be considered transactions between individuals and their social environment. They give personal meaning to external and internal stimuli and communicate meaning from the individual to others. Emotions consist of interpretations of intero- and exteroceptive stimuli, intentions, physiological patterns of arousal, and motor behavior including overt emotional expressiveness. The interaction of these different components in the individual and the social and physical environment are mediated by the central nervous system (CNS). From a regulation point of view, emotional expressiveness has two important functions: first, it serves a communicative function by facilitating the regulation of person–environment transactions and, second, the feedback function of behavioral expressions controls the intra-individual regulation of emotion. This means that an experience

may be influenced indirectly by actively responding towards a negative emotional stimulus in the environment, in order to attenuate it, or directly through self-regulation. Thus, expressive behavior can serve simultaneously as a component of emotional processes and as a coping response.

Both psychoanalysis and neurophysiology advanced to the concept of inhibition in different contexts in the twentieth century. For many years, an inverse relationship between expressive behavior and autonomic responsivity has been documented, such that the inhibition of overt emotional expressiveness can lead to an autonomic over-reaction. This has been considered a significant factor in the etiology and maintenance of psychosomatic disorders. A number of early researchers in the first two decades of the twentieth century reported measurements of high physiological activity in subjects suppressing emotional expression. These studies led to the concept of internalization and externalization, where two behavioral coping styles for dealing with psychic tension were discerned: either behaviorally, outwardly directed, or physiologically, within the individual. According to this concept, the term 'internalizer' describes a person exhibiting a low level of overt expressiveness under stress yet a high level of physiological excitation, whereas an 'externalizer' is characterized by high expressiveness and a low level of physiological expressiveness in social situations. Different health models explain on internalizing and externalizing in terms of stress and coping or in terms of inhibition as physiological work. Over time, the work of inhibition acts as a low-level cumulative stressor. As with all cumulative stressors, sustained inhibition is linked to increases in stress-related diseases, and various other disorders such as cardiovascular and skin disorders, asthma, cancer, and also pain (Traue and Pennebaker 1993).

3. *The Model of Inhibition and Health*

A model explaining how emotional stress, under a given social situation, can trigger or modulate health disorders is described below (see Fig. 1). Health disorders and illness behavior are conceived as different, but related, processes, and distinct mechanisms may contribute dependently and independently to different aspects of a given disorder and its behavioral consequences (Traue 1998).

On a phenomenological level, emotional stress can occur on a severity dimension ranging from daily stressors, through more traumatic life events, to more chronic or severe psychotrauma. A common underlying factor in such situations is that each has to be coped with by the individual. Emotional stress can be considered to be processed by way of an '*inhibition-implosion*'-dimension (to implode means to collapse or cause to collapse inwards in a violent manner as a result of external pressure), modulated by dispositional factors (innate, personality, and socialization).

Of particular importance are innate and socialization factors. Possible individual differences in hypothalamic mechanisms and opioidergic pathways corresponding with increased vulnerability to stress have been suggested. Emotional processing relating to the inhibition-implosion idea has been discussed in relation to several topics such as control, suppression, Type C personality, repression, alexithymia, or ambivalence. Each of these concepts covers different aspects of overt emotional expressiveness from a personality or coping perspective. Inhibition is the most general term for the incomplete processing of emotional stress when these stressors induce bodily changes (physiological, endocrinological, or immunological) and the cognitive, emotional, and behavioral processing are dysfunctional such that subjective experience and spontaneous expression of emotions and action tendencies are separately or simultaneously attenuated and intra- and interpersonal regulation is disturbed. This process may be the result of innate or acquired behavior.

3.1 *Classification of Emotional Inhibition*

A classification of involved mechanisms differentiates between genetic, repressive, suppressive, and deceptive inhibition.

Genetic inhibition describes the genetically determined basis of behavioral inhibition. Studies working with young children have classified those children who were least able to initiate interaction in a social situation with other children and adults as behaviorally inhibited. Most children's degree of behavioral inhibition has been shown to be stable over a period of five years. In those children classified as inhibited, increased levels of arousal, norepinephrine, and salivary cortisol have been found.

Repressive inhibition refers to emotional processing with attenuated subjective experience of emotional arousal. Emotionally expressive responses in repressive inhibition could be based solely on the cognitive interpretation of the situation and are organized nonspontaneously. The individual, being unable to feel their own arousal, processes insufficient response-information for an appropriate response, which in turn decreases the need to express emotions or cope with an emotional stressor. In addition, the cognitive interpretation of the situation without the emotional component may be wrong or misleading. Prolonged bodily arousal and impaired coping could result. Repressive coping style is the best known model for repressive inhibition.

Suppressive inhibition is best circumscribed by suppression of emotional arousal. The emotional arousal is

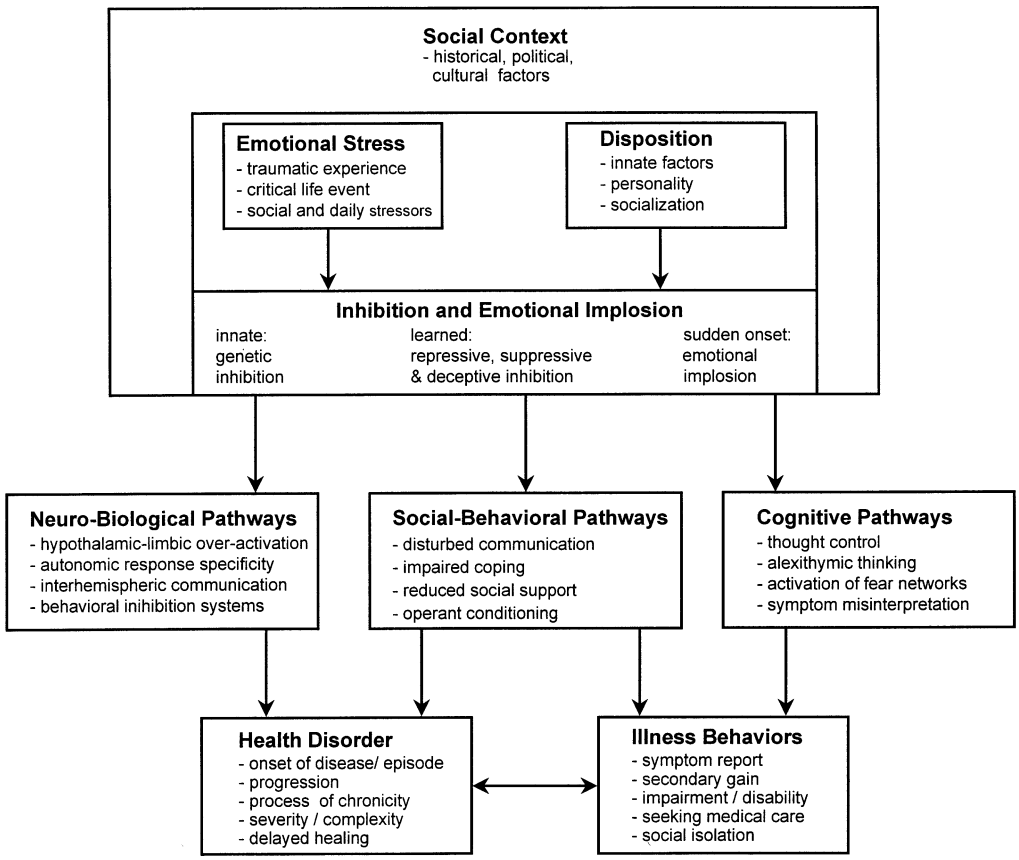

Figure 1
Pathway model of emotional inhibition with neurobiological, social-behavioral, and cognitive pathways from stress/disposition to emotional processing and subsequent health disorders or illness behaviors (adapted from Traue 1998)

recognized by the individual, but spontaneous expressive and cognitive behaviors are suppressed involuntarily or habitually. Suppressive inhibition of emotions results from interactions between innate factors and socialization. For example, if individuals show increased responses under stressful encounters, they are prone to socialization conditions of punishment and negative reinforcement, initiating a learning history with decreases in spontaneous expressiveness and increases in bodily arousal.

Deceptive inhibition describes the phenomenon whereby individuals under emotional stress, aware of their bodily reactions and their urge for expressiveness, can suppress this need voluntarily or try to present a false response to a receiver. Whether an individual is pokerfaced or displays a false emotional response, such inhibition requires psychological energy, reducing the individual's coping capacity and providing additional stress.

In reviewing the psychophysiological and psychosomatic data, it might appear that emotional inhibition is always harmful. However, it should be noted that although the correlations between bodily processes and the above four forms of inhibition support such a notion, in certain circumstances inhibition could be beneficial for the individual and his/her relationship with the social environment. Inhibition becomes toxic when it is related, first, to physiological, endocrinological hyperarousal or immunological dysfunction, second to long-standing disregulation of emotions within the individual on a cognitive and behavioral basis, and third, if inhibition disturbs the individual's social relations.

Inhibition represents a risk factor for health under normal stressors. If the severity of the stressors is dramatically high, the mental and physical health consequences are inevitable. In traumatic stress situations like rape, criminal bodily attacks, or torture the victim may well lose control over strong emotional responses. The emotional responses of horror, panic, and loss of control could literally cause a violent breakdown in the mental and bodily systems. Such an

emotional implosion (Traue 1998) is observable in the symptom pattern of post-traumatic stress disorder (PTSD): cognitive and behavioral avoidance of trauma stimuli, numbing of general and emotional responsiveness (emotional anesthesia), detachment from other people, persistent symptoms of arousal like disturbed sleep, exaggerated startle response and somatic complaints (see *Post-traumatic Stress Disorder*). In addition, persons with PTSD suffer an increased risk of social phobia, major depressive and somatizing disorder (*DSM-IV*, American Psychiatric Association 1994).

The psychological, physical, and social symptoms in PTSD are a form of emotional processing involving an extreme form of inhibition. While inhibition generally develops over a long time span through interaction between innate and socialization factors, implosion can occur in a very short time as a result of a single and extreme man–made stressful event.

4. *Pathways from Emotional Inhibition to Health Disorders and Illness Behaviors*

Emotional stress modulated through innate, personality, and socialization factors can trigger, maintain, or worsen health disorders and related illness behavior through neurobiological, social-behavioral, and cognitive pathways. With respect to illness behaviors, the pathways include biases in symptom reporting, secondary gain by presented symptoms, subjective feelings of being impaired, pressure to seek medical help and social isolation.

4.1 *Neurobiological Pathways*

Emotional inhibition is strongly neurobiologically based. The behavioral inhibition system and the behavioral activation system have been discussed as possible neurobiological structures. Empirical evidence from between-subject studies shows that inhibited, repressed, or suppressed emotional expressiveness is linked to greater autonomic arousal, both under conditions of emotion induction and voluntary deception. There is rich empirical evidence for neurobiological correlates of inhibition in respiratory, cardiovascular, muscular, digestive, and immune functions. Immune functioning is of particular interest because it is the immune system that may be relevant in all sorts of infectious, allergic, and neoplastic illness processes. Inhibited style of processing upsetting events can compromise immune functions, resulting in higher serum antibody titers, decreased monocytes counts, and poorer natural killer cell activities. Other fields of research relevant to the neurobiological pathways of inhibition include hypothalamic-limbic overactivation, prolonged activation of physiological response specificity, hemispheric brain lateralization of emotion processing and faulty interhemispheric communication, the neuroregulation of action, and the behavioral activation and behavioral inhibition systems.

4.2 *Social-behavioral Pathways*

Several social-behavioral pathways connect inhibition-implosion to health disorders and illness behaviors. First of all, neurobiologically innate factors (shyness, behavioral inhibition, hypersensitivity, introversion) are superimposed by classical and operant conditioning in the socialization of an individual. Since individuals with these characteristics in early childhood are conditioned more easily, the process of socialization is of greater importance to them than it is to others. Under critical developmental conditions, the gap between emotional expressiveness and physiological hyperactivity may increase. A lack or deficit in emotional expressiveness will hinder interpersonal communication. It is thought that deficits in interpersonal communication disturb the development of emotional competence that is important for sharing experiences, maintaining psychological and physical contact, and adapting to the social environment (Rimé 1995). These are deficiencies in the healthy coping competencies that have been termed emotional intelligence. Other consequences of inhibited emotional expressiveness are disturbed social relations resulting in social isolation and a disrupted social support network.

Normally, persons respond to emotion-evoking stimuli with emotional expression, and such reactive expression is created by facial muscle activity and movements with reafferent neuronal signals in the CNS, which contribute to the individual's emotional experience. Subjective emotional experience does not, however, depend exclusively on this nervous input as claimed in the facial feedback hypotheses, but feedback does contribute positively to sensitivity towards the physiological aspects of emotion. If this sensitivity is disrupted, an individual will not perceive adequately increased muscle tension or other autonomic nervous system (ANS) reactions caused by stress, and consequently will not initiate healthy relaxing behavior. It has been demonstrated that the hypothesis of deficient perception of muscle tension holds for myogenic pain. These patients were significantly less able to judge the extent of their muscle strain than were controls.

It is conceivable that, under unfavorable circumstances, expressive behavior for mainly negative emotions like anger and aggressiveness is punished socially and thus justifiably avoided. The suppression of expressive behavior can be achieved by an additional increase in muscle activity. Such avoidance behavior or inhibition is very adaptable in the short

term, and it helps to modify a socially stressful situation. This reduction of emotional expressiveness is conditioned by the learning mechanism of negative reinforcement (the avoidance of punishment).

4.3 Cognitive Pathways

Memories and thoughts of emotional stressors are generally unpleasant. Increasing severity of the stressful encounter makes imagining the event painful or even unbearable in the case of traumatic experience. Most individuals attempt to suppress the thoughts surrounding traumatic events. As soon as the inhibition work begins, the urge to distract oneself and the mental energy put into this process fuels the images. Consequently, triggered intrusions and unwanted thoughts make life more stressful than before. In addition, thought control interferes with natural ways of coping such as sharing the experience with important others and thinking through the event. Therefore, inhibition may be dangerous because it hampers the individual who has suffered a critical life event in resolving the stressful experience cognitively and behaviorally.

As part of the inhibition process, individuals may tend to exclude the emotional content of the stressful encounter from their language representation of the event, called alexithymia, or low level thinking style. Although this may help to avoid negative emotionality in the short term, it hinders complete processing and integration of the stressful experience. The lack of integration into the self-concept makes an individual prone to the activation of fear networks. Finally, impaired or unfinished cognitive processing can make an individual prone to misinterpreting bodily symptoms. Instead of understanding bodily reactions as part of emotional responses, the individual conceptualizes the bodily reactions as symptom patterns and seeks medical help for illnesses. Badly advised medical treatment procedures result in iatrogenic diseases trapping the individual in a vicious cycle. Cognitive appraisal of a situation depends partly on facial feedback as a source of emotional information. When the expressive components of emotional reactions are systematically repressed by inhibition, the individual 'unlearns' an accurate assessment of stressful circumstances. This learning mechanism occurs because the estimated load of a stress situation is dependent not only on external features of the situation, but also on the subjective experience of stress-conditioned reactions. When, however, an inhibited person takes bodily reactions into account in the evaluation of a situation, their judgment will be impaired when the original physiological components of mainly negative emotions are interpreted as 'symptoms.'

In clinical studies, psychosomatic patients suffering, for example, from headaches have been found to report significantly lower stress levels than a control group, but showed nearly twice as much neck muscle tension as the controls. Although the arousal and muscle tension data indicated higher levels of stress, patients were unable or unwilling to report those stressors. It would appear that patients tend to interpret their stressors in terms of bodily symptoms, rather than underlying levels of stress.

5. Rituals and Therapeutic Interventions

Most societies seem to contain at least some implicit knowledge that emotional inhibition has negative health implications. The conflict resulting from the need for emotional regulation on the one hand and the need for disclosure, sharing, and catharsis on the other hand leads to a variety of cultural phenomena to solve this conflict. These include older universal cultural rituals (such as rituals of, grief, or lament), or religious acts such as confessions. The Wailing Wall in Jerusalem, where people have been going for centuries to deliver a written prayer, is possibly an example of an ancient disclosure phenomenon. Contemporary Western societies have introduced psychotherapy for enhancing emotional expressiveness. Here, talking or writing about emotions is encouraged, as is the acting out of emotions in role-plays. Assertiveness training aims at effective expression of emotion and catharsis-based techniques like confrontation are modern remedies for anxiety, PTSD, and the like. All of these techniques seem to have in common that they are directed at the construction of meaning from emotional experience. Culture, which may be construed as a subset of possible meanings, is therefore a very important mediating factor in this process. New approaches to cross-cultural psychology have proposed intercultural differences in emotional inhibition based on dimensions like masculinity–femininity, collectivism–individualism, uncertainty avoidance, and power distance.

See also: Coping across the Lifespan; Emotions and Health; Emotions, Psychological Structure of; Stress and Coping Theories; Stress and Health Research; Stress: Psychological Perspectives

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders DSM-IV*, 4th edn. American Psychiatric Association, Washington, DC

Rimé B 1995 Mental rumination, social sharing, and the recovery from emotional exposure. In: Pennebaker J W (ed.) *Emotion, Disclosure, and Health*. American Psychological Association, Washington, DC, pp. 271–92

Traue H C 1998 *Emotion und Gesundheit. Die psychobiologische Regulation durch Hemmungen (Emotion and Health. Psyc-*

hobiological Regulation by Inhibition). Spektrum, Heidelberg, Germany
Traue H C, Pennebaker J W (eds.) 1993 *Emotion, Inhibition and Health*. Hogrefe & Huber, Seattle, WA

H. C. Traue

Emotions and Health

Emotions have long been believed to play a significant role in health. However, definitive evidence on the nature of this role and the mechanisms involved is still being developed. On the one hand, patients often have emotional reactions to the diagnosis of serious or life-threatening illness, responses that can influence emotional as well as physical well being. On the other hand, emotions and their expression or inhibition have been linked with the development of various disease conditions. This article explores what is currently known about these links between emotions and health and the implications for both theory and intervention.

1. Background

Following a period dominated by behaviorism and stimulus-response approaches to behavior, there has of late been an explosion of interest in the emotions, their nature, antecedents, and consequences. Beginning in the 1980s, this interest has been brought about by a growing acceptance of a cognitive mediational approach to human behavior which goes beyond a simple stimulus-response model of human behavior to seeing behavior as being strongly mediated by internal psychological processes of which, of course, emotions are a part (Lazarus 1993).

At the most basic level, emotions are the affective or feeling states that people experience in response to external events, internal cognitions, or a combination of both. Beyond this, however, there are numerous controversies about the nature of human emotions. For example, there is debate over whether emotions are best described in terms of categories or dimensions, how many emotions can be delineated, and the ways in which emotions differ across cultures. These controversies, however, are beyond the scope of this article that specifically is concerned with the interrelationship of emotions and physical health. For the purposes of this article we will be concerned with emotions as feeling states which can be differentiated into identifiable emotions, are embedded in a social context, may be only experienced internally or also expressed outwardly, and are associated with identifiable patterns of physiological activity.

Work on emotions and health bears a close affinity to work on the relationship of stress to health (see *Stress and Health Research*) as well as work on personality and health (see *Personality and Health*). The relationship of emotion to stress has been delineated very persuasively by Lazarus (1993) who notes that the term stress has often been used to refer globally to negative emotional responses that people have to situations involving threat or harm and argues that stress should be understood as a subset of the emotions. This view has the advantage of bringing together two closely related but generally separate research literatures and allows for a more differentiated understanding of stress and its effects on health. The relationship of the work on emotions and health to that on personality and health derives from the fact that at least some of the personality traits found to be related to health are associated with emotion (Friedman 1990).

We begin our discussion with a consideration of the emotional impact of illness and then move on to a discussion of emerging evidence for the role of emotion in the development of disease.

2. Emotional Responses to Illness

Diagnosis with a chronic or life-threatening disease has been described as a crisis for patients which produces strong emotional responses. Patients diagnosed with serious conditions frequently experience intense feelings of fear, anxiety, grief, anger, and other emotions. These emotional responses often lessen with time as the person comes to grips with the condition but can lead to serious emotional maladjustment. Key determinants of quality of emotional adjustment appear to be age, the severity of the condition, coping strategies, and social support. Older patients appear to have less difficulty adjusting to disability than younger persons, perhaps because of an expectation of greater disability as one gets older whereas more severe disease is often associated with greater emotional upset. Further, coping strategies appear to be critical determinant of emotional adjustment. For example, positive adaptation to breast cancer appears to be facilitated by finding a sense of meaning in the experience of cancer as well as developing a sense of mastery and control over the disease and using downward comparison to help one feel good about how one is doing in the face of the disease (Taylor 1983). Finally, having supportive social relationships is critical to positive adaptation (see *Social Support and Recovery from Disease and Medical Procedures*).

Emotional responses to serious illness can also have important implications for the person's physical well being. For example, 10-year survival rates for breast cancer patients who responded to their diagnosis with feelings of helplessness, denial, or stoic acceptance

have been found to be substantially lower than those for patients who responded to their diagnosis with what has been termed a 'fighting spirit,' which included optimism in the face of their disease and a determination to do everything in their power to overcome the disease (Pettingale 1984).

3. Emotions and the Development of Disease

Emotions have also been linked to the development of a variety of disease conditions. In what follows, evidence is examined concerning emotions that are associated with a higher risk of disease as well as others which appear to be protective against ill health (see *Health Risk Appraisal and Optimistic Bias*).

3.1 Emotions as Health Risk

Not surprisingly, emotions associated with increased risk of disease are negative ones. Evidence has accumulated linking various negative emotions to the experience of ill health as well as the development of various disease conditions (Friedman 1990). Further, there is a tendency for the experience of different negative emotions to be correlated with each other and as such it makes sense to refer to a general tendency towards the experience of negative emotions. This tendency is known as negative affectivity and is closely related to the personality trait of neuroticism. A number of studies have found negative affectivity to be a common ingredient running through various health-related measures of stress, emotion, coping, and illness experience. As such it has been argued that findings relating such measures as life stress, pessimism, loneliness, low life satisfaction, and other measures of psychological distress to poor health can be explained, at least in part, as being a reflection of this general tendency to experience negative emotions and interpret experience in negative terms (Clark and Watson 1991).

Other evidence indicates that it is not just the experience of negative emotions that is associated with poorer health but more specifically the lack of expression of these emotions. Since the 1980s research evidence has been accumulating on the negative health effects of emotional inhibition. In particular, people with repressive coping styles as well as those actively inhibiting the expression of emotion have been found to have generally poorer health and to show negative psychophysiological as well as immunological changes as a result of suppressing the expression of negative feelings (Pennebaker 1995).

Although these general variables of negative affectivity and emotional inhibition are associated with poor health generally, there is also good reason to believe that a full understanding of the relationship of emotion to health requires examination of the ways in which specific emotions relate to health status and to specific disease conditions. In particular, the relationships of different emotions to specific health conditions are somewhat variable, suggesting at least a certain amount of specificity in the role of emotions in health.

Most likely the strongest relationship with health outcomes has been obtained for anger and its cognitive and dispositional counterpart, hostility. Even though anger and hostility can be conceptually differentiated they are discussed together because they are clearly closely related, measures of these two constructs generally share a common core of variance, and they are frequently discussed together in the research literature. Numerous studies have found significant associations between anger/hostility and deaths from coronary heart disease as well as mortality from all causes, an association that remains significant when other risk factors are statistically controlled. In the case of coronary heart disease this association has been described as being at least as large as those reported for more traditional risk factors for CHD such as cigarette smoking, high blood pressure, and high serum cholesterol (Miller et al. 1996). Anger and hostility also appear to be related to other disease conditions such as cancer and arthritis but the strength of these relationships appears to be somewhat weaker than that found with CHD. Interestingly, while it appears to be the expression of anger that is associated with CHD, the suppression of anger is associated with cancer (Friedman 1990).

Depression is a second emotional state with significant health implications. The most obvious association between depression and health is with suicide (see *Suicide*). However, depression has been found to have an association with a number of other medical conditions as well. On the one hand, there is evidence that depression is associated with an increased likelihood of conditions such as heart disease and cancer as well as with overall mortality (Wulsin et al. 1999). On the other hand, depression has been found to be associated with poorer prognosis for patients diagnosed with conditions such as AIDS, cancer, and heart disease. For example, work by Frasure-Smith et al. (1995) has found depression to be a significant predictor of mortality among cardiac patients whereas similar findings have been obtained for cancer patients (Pettingale 1984) and patients with AIDS (Patterson et al. 1996).

Relatedly, grief has been found to have significant health implications. Several studies have found increased illness as well as mortality among individuals who have been recently bereaved. This increased mortality has been found to be greatest in the first six months following bereavement and to derive from a number of causes including accidents, violence, heart disease, and cancer (Wulsin et al. 1999).

Finally, anxiety and fear have been found to have significant health implications. For one thing, anxiety can be converted into physical complaints through

somatization. For example, studies of patients' visits to physicians have found that a large portion of such visits are often the result of physical complaints with an emotional basis, often anxiety (Barsky and Klerman 1983). In addition, anxiety has been found to be associated with conditions as diverse as CHD, asthma, ulcers, arthritis, and headaches (Friedman 1990; Friedman and Booth-Kewley 1987) and severe fright has been linked to sudden death from heart attacks (Hafen et al. 1996).

3.2 Emotions as Protector of Health

The flip side of the coin is that positive emotions may be protective of health. Evidence for this is not as strong as for the effects of negative emotions but links are being established. As with negative emotions, there is a tendency for dispositional measures of positive emotion to correlate with each other, a construct termed positive affectivity. Positive affectivity is largely independent of negative affectivity and, whereas negative affectivity is closely associated with neuroticism, positive affectivity is closely related to extroversion. Current evidence indicates that on its own, positive affectivity is generally unrelated to health complaints. However, conceptually, positive affectivity is closely related to optimism and low levels of positive affectivity appear to be an important ingredient in depression. As such there has been speculation that it is the positive affectivity components of these emotional states that are responsible for their relationships to physical health (Clark and Watson 1991).

Perhaps the strongest evidence for positive emotions as a protector of health comes from work on optimism. Optimism refers to a person's tendency to look on the positive side of events and to believe that events will work out for the best. As such optimists can be expected to experience positive emotions more often than more pessimistic individuals. And, in fact, as noted, optimism is associated closely with positive affectivity. Research on optimism indicates that optimists tend to live longer, healthier lives than more pessimistic individuals and also to be less susceptible to viral infections, such as the common cold. In addition, optimistic cancer patients tend to live longer than more pessimistic ones and optimism is associated with greater immune system competence as indicated by the ratio of T4 to T8 cells (Peterson and Bossio 1991).

4. Mechanisms Linking Emotions and Health

As noted above there is accumulating evidence that emotions are significantly related to health and indications that emotional states may play a significant role in the development of disease as well as in the healing process. How can these relationships be explained? What are the mechanisms linking emotions and health? The answers to these questions are still emerging but a number of plausible hypotheses have been proposed. These hypotheses relate to both indirect and direct effects that emotions may have on disease processes.

4.1 Indirect Links Between Emotions and Health: The Health Behavior Connection

One possibility is that emotions influence health indirectly through their effects on behavior, particularly health behaviors. Numerous studies have noted that stress, which often includes emotional upset, is associated with poorer health behaviors. For example, higher levels of stress have been associated with higher levels of smoking as well as increased alcohol and drug usage and poorer eating habits. Specific emotions also appear to be associated with health behaviors. For example, there is evidence that higher levels of hostility are associated with greater consumption of cigarettes and alcohol as well as less physical exercise and poorer self-care (Smith 1992). It is well establised that these health behaviors are associated with poorer health outcomes (see *Health Behaviors*) and thus it seems highly plausible that emotions may well affect health outcomes in this manner.

4.2 Direct Links Between Emotions and Health: The Physiological Connection

There also appear to be direct physiological links between emotions and health. In addition to their experiential aspects emotions are very much physiological events. There has been much controversy over the years about the precise relationship of emotions to physiology, whether there are specific patterns of physiological changes that correspond to specific emotions, and whether the phenomenological experience of emotion precedes or is the result of such physiological changes, but there is no doubt that emotions are associated with a variety of physiological changes. These changes occur in a number of different bodily systems including the nervous, cardiovascular, respiratory, endocrine, and muscular systems and vary considerably in intensity. A substantial number of studies have attempted to chart these changes and relate them to specific emotions. Results from such studies have been mixed in that physiological changes found for specific emotions tend to vary considerably between studies and with the methods used for evoking emotion as well as measuring physiological changes. In many respects, this variation is not surprising given the complexity of physiological events and the metho-

dological hurdles in establishing their relationship to particular emotions (Cacioppo et al. 1993). However, there is no doubt that emotions are associated with substantial physiological changes and it seems highly plausible that these changes may mediate the relationship between emotions and health.

Although there is considerable controversy about the precise relationship between emotions and physiology, progress has been made in charting likely physiological pathways between emotions and specific diseases. Some of the most detailed evidence to date comes from work on the relationship of anger and hostility to cardiovascular disease. A number of studies have found that individuals high in anger and hostility tend to show heightened physiological, and particularly cardiovascular, responses to various stressors. Although results have not been entirely consistent there is evidence that individuals who are high in hostility, and who are thus more prone to the experience of anger, show larger stress-related increases in blood pressure, heart rate, and stress-related hormones than do individuals lower in hostility. This heightened physiological reactivity appears particularly likely when the person is faced with social stressors such as harassment or other social challenge and is considered by some to be at least a marker if not a causal agent for the future development of cardiovascular disease (Smith 1992). One possible mechanism is that increased blood pressure and catecholamine reactivity as a result of repeated episodes of anger may be related to increased likelihood of injury to the interior walls of blood vessels which in turns accelerates the process of atherosclerosis (Manuck 1994).

In addition to heightened reactivity in the face of stressors, it is also quite possible that high hostile individuals experience more situations evoking anger, either because they gravitate toward such situations and/or because their actions tend to evoke hostile responses from others. For example, there is evidence that high levels of hostility are associated with generally high levels of interpersonal conflict and low levels of social support. In addition, high hostility has been associated with the experience of more frequent and severe daily hassles as well as major negative life events. This tends to suggest a synergism between poor interpersonal relationships and high physiological reactivity which would further increase the high hostile individual's vulnerability to disease (Smith 1992).

Heightened physiological reactivity is one link between emotions and health but it is certainly not the only one. Another important link concerns the relationship of emotion to immune functioning. In this respect there is a growing literature showing that the tendency to experience certain emotions can affect immuno-competence, as can transient moods. First, indices of immuno-competence have been found to be lower among individuals with depressive symptomotology as well as among recently bereaved individuals strongly suggesting that dysphoric moods may lead to the suppression of immune functioning (O'Leary 1990). One possible explanation for this is that withdrawal responses and the generally more passive coping responses associated with helplessness and depression are associated with the activation of the hypothalamic-pituitary-cortical axis which leads to increases in cortisol output which in turn have the effect of suppressing immune functioning (Antoni 1987). Such processes have important implications for a variety of immune related diseases including cancer, AIDS, and autoimmune diseases. This may well be one of the reasons why individuals with a so-called 'Type C' personality, which includes the traits of being unassertive, patient, and unexpressive of negative emotions (see *Personality and Health*), all of which are related to passive coping, are more likely to develop cancer (Friedman 1990). Second, humor and positive emotions have been found to be related to increases in immuno-competence among other physiological changes. For example, several studies have found that laughter is associated with increases in concentrations of lymphocytes as well as increased natural killer cell activity and overall immune responsiveness, and a reduction in the secretion of cortisol. Similarly, studies on optimism have found that optimists have stronger immune functioning than do people with a more pessimistic attitude towards life (Hafen et al. 1996). Finally, there is growing evidence that the suppression and expression of emotion have both physiological and immunological effects. Inhibition of emotion has been shown to lead to changes in electrodermal as well as cardiovascular activity. Also, in addition to evidence cited above that suppression of negative emotions is related to reduced immune functioning, there is evidence from a growing number of studies that the expression of negative feelings as a form of catharsis can result in improved immune functioning. For example, several studies have now demonstrated that having people write or talk about traumatic experiences they have not disclosed to other people previously results in an increase in immune responsiveness. This occurs when they focus on the emotions involved but not when they focus only on the facts of what happened (Pennebaker 1995).

4.3 Common Underlying Factors Between Emotions and Health: Constitutional Vulnerability

A third possible link between emotions and health is that there may be common underlying factors for both emotionality and the occurrence of disease. For example, there may be genetic factors that influence, on the one hand, the emotions a person experiences and how those emotions are expressed and, on the other hand, the occurrence of ill health in general as well as specific disease conditions. As yet no such genetic links have been established but this does not

mean that they do not exist. With the rapid progress currently being made in the genetics of disease as well as work on the genetics of personality such links may well be established in the future.

5. Reducing Health Risk

The emerging pattern of linkages between emotions and health raises the question of what can be done to reduce the health risk from negative emotions and boost the benefits from positive ones. Three possibilities are considered here but these are best seen as only representative of interventions that may be effective.

First, psychotherapeutic techniques can be used to help people deal more effectively with negative emotions as well as boost positive emotions. Numerous techniques have been developed, and there is reason to believe that at least some of these may have positive benefits for physical health as well as mental health. For example, psychotherapy interventions have been shown to be effective in boosting immune responsiveness and improving physical health in both cancer patients and noncancer populations. Psychotherapy with cancer patients has been shown to reduce recurrence and increase five-year survival rates which is consistent with data showing that psychotherapy with a variety of medically ill patients can lead to improved health outcomes (Fawzy and Fawzy 1994). Of course, psychotherapy does more than simply reduce negative emotions and boost positive ones but such data, in conjunction with the evidence on the role of emotions in health, suggest the potential of psychotherapy for reducing the negative effects of emotions on health.

Closely related to these effects of psychotherapy in general is work showing the psychological and physical benefits of writing or talking about traumatic events. Research by Pennebaker and others has produced substantial evidence for the positive psychological and physical benefits of writing as well as talking about traumatic events. Overall, when people, ranging from university students to holocaust survivors, are induced to write or talk about traumatic experiences positive psychological benefits are found and there are reductions in harmful arousal as well as improvements in immune functioning and subsequent reductions in the number of times the person seeks medical attention. Such changes are particularly evident when the person writes or talks about traumas that had not been previously disclosed to others and when he or she writes about the emotions involved and not just the facts of the situation (Pennebaker 1995). This strongly argues for the health benefits of disclosure of negative experiences to others and suggests an apparently powerful way of reducing the harmful effects of emotion on health.

Finally, evidence on the role of anger and hostility in heart disease strongly argues for interventions directed towards reducing these potentially harmful emotions. A number of strategies have been suggested and there is evidence that teaching people to be aware of their feelings and then utilizing a series of strategies for dealing with difficult life situations can have positive effects both on the person's relationships and psychological adjustment and possibly on their physical health as well (Williams and Williams 1997). Definitive evidence on the effects of these interventions is not yet available but the results to date are promising.

6. Implications and Future Directions

Findings concerning the relationships of emotion to health are a prime example of the way in which psychological and social factors are closely intertwined with biological processes in determining one's health status. Disease is clearly a biological phenomenon but it is one that is strongly influenced by psychosocial factors. Also, phenomena such as emotions that are generally viewed in the realm of the psychological have a definite basis in biological processes. As such one cannot speak of psychological processes without consideration of the biological and vice versa.

Although the evidence relating emotions to health has been growing rapidly there are still many questions that remain to be answered. One major question concerns the degree to which specific emotions are related to specific disease conditions. There has long been speculation about such relationships but as yet there has been no clear resolution of this question. Whereas work on such topics as negative affectivity tends to suggest a general relationship between negative emotions and ill health, other work tends to suggest at least a degree of specificity. This is a question that needs further work and one which will benefit from a clearer specification of how one can best differentiate between different emotions both psychologically and physiologically.

Relatedly, there are still many questions about the mechanisms linking emotions and health. Evidence for the roles of physiological reactivity and immunocompetence in this linkage is growing but much remains to be learned. Exactly how is it that different emotions produce these physiological changes? What other biological processes operate in linking emotion to health? How are genetic factors involved in this relationship? These are fascinating questions that still remain to be addressed.

Finally, much remains to be done with respect to capitalizing on what has been learned in this area for improving health. Current interventions have been described but these would appear to only scratch the surface of what might be possible with respect to

harnessing psychological methods for improving physical health.

See also: Depression, Hopelessness, Optimism, and Health; Emotional Inhibition and Health; Explanatory Style and Health; Health: Self-regulation; Personality and Health; Stress and Health Research

Bibliography

Antoni M H 1987 Neuroendocrine influences in psychoimmunology and neoplasia: A review. *Psychology and Health* **1**: 3–24

Barsky A J, Klerman G L 1983 Overview: Hypochondriasis, bodily complaints, and somatic styles. *American Journal of Psychiatry* **140**: 273–83

Cacioppo J T, Klein D J, Berntson G G, Hatfield E 1993 The psychophysiology of emotion. In: Lewis M, Haviland J M (eds.) *Handbook of Emotions*. Guilford Press, New York

Clark L A, Watson D 1991 General affective dispositions in physical and psychological health. In: Synder C R, Forsyth D R (eds.) *Handbook of Social and Clinical Psychology: The Health Perspective*. Pergamon Press, New York

Fawzy F I, Fawzy N W 1994 Psychoeducational interventions and health outcomes. In: Glaser R, Kiecolt-Glaser J K (eds.) *Handbook of Human Stress and Immunity*. Academic Press, San Diego, CA

Frasure-Smith N, Lesperance F, Talajic M 1995 Depression and 18-month prognosis after myocardial infarction. *Circulation* **91**: 999–1005

Friedman H S (ed.) 1990 *Personality and Disease*. Wiley, New York

Friedman H S, Booth-Kewley S 1987 The 'disease prone personality': A meta-analytic view of the construct. *American Psychologist* **42**: 539–55

Hafen B Q, Karren K J, Frandsen K J, Smith N L 1996 *Mind, Body, Health*. Allyn and Bacon, Boston

Lazarus R S 1993 From psychological stress to the emotions: A history of changing outlooks. *Annual Review of Psychology* **44**: 1–21

Manuck S B 1994 Cardiovascular reactivity in cardiovascular disease: 'Once more unto the breach'. *International Journal of Behavioral Medicine* **1**: 4–31

Miller T Q, Smith T W, Turner C W, Guijarro M L, Hallet A J 1996 A meta-analytic review of research on hostility and physical health. *Psychological Bulletin* **119**: 322–48

O'Leary A 1990 Stress, emotion, and human immune function. *Psychological Bulletin* **108**: 363–82

Patterson T L, Shaw W S, Semple S J, Cherner M, McCutchan J A, Atkinson J H, Grant I, Nannis E 1996 HIV Neurobehavioral Research Center (HNRC) group Relationship of psychosocial factors to HIV disease progression. *Annals of Behavioral Medicine* **18**: 30–9

Pennebaker J W (ed.) 1995 *Emotion, Disclosure, and Health*. American Psychological Association, Washington, DC

Peterson C, Bossio L M 1991 *Health and Optimism*. Free Press, New York

Pettingale K W 1984 Coping and cancer prognosis. *Journal of Psychosomatic Research* **28**: 363–64

Smith T W 1992 Hostility and health: Current status of a psychosomatic hypothesis. A review. *Health Psychology* **11**: 139–50

Taylor S E 1983 Adjustment to threatening events: A theory of cognitive adaptation. *American Psychologist* **38**: 1161–73

Williams V, Williams R B 1997 *Life Skills*. Times Books, New York

Wulsin L R, Vaillant G E, Wells V E 1999 A systematic review of the mortality of depression. *Psychosomatic Medicine* **61**: 6–17

G. D. Bishop

Emotions, Children's Understanding of

1. Theoretical Foundations

A *social constructivist* theoretical position is taken in this article on children's understanding of emotion, which means that children construct their understanding of emotional experience in the context of cultural beliefs, relationship transactions, and an affordant environment. The child's cognitive developmental level may be thought of as an organismic constraint on how and with what degree of complexity children construct their understanding of emotion. A *functionalist* perspective is also evident here in that how we come to understand our emotional experience is rooted in the goals that we have vis à vis our encounters with our environment (Campos et al. 1994). One of the more prominent aspects of western culture's views of how emotion 'works' is the emphasis on cognitive mediation of emotion or appraisal (e.g., Lazarus 1991). This western folk belief about emotions as stemming from one's thoughts and images derives from an emphasis on emotion as an internal, individual experience as opposed to a number of other cultures' emphasis on emotion as an interactive experience (Kitayama and Markus 1994). What do we know about children's acquisition of this belief, namely that emotion results from one's own cognition (as opposed to deriving from one's kinship system, from ancestral spirits, or from bodily illness)? Research undertaken by P. L. Harris (Harris 1989) suggests that in the preschool years western children demonstrate in their understanding of others' experience that the perspective or belief held by the other is significantly involved in determining how the other feels. As they move through middle childhood, children begin to invoke mental factors as well for how to change emotional experience. For example, Harris and Lipian (1989) interviewed children in boarding schools to find out how the children thought they could alleviate negative feelings such as home-sickness and sadness. Older children (age 10) were much more likely to cite such strategies as distraction, focusing on something positive, reinterpreting the situation, and so forth. Younger children (age six) tended to restrict their strategies to suggestions about how to change the situation rather than looking inward to seeing whether and how one might alter one's internal experience. This greater reliance by the older children on internal,

mental states as part of the 'emotional package' impacts on several critical features of emotion, among them which cues are used to identify emotions reliably, how emotional-expressive behavior is regulated, and importantly, how to cope with aversive emotions.

Goals play an important role in children's appraisals of what gives rise to emotions and feelings (e.g., Levine 1995, Wellman and Banerjee 1991). Liking or wanting something or the opposite, not liking or not wanting something, are prominent anchors in how preschool children determine how they or others are likely to feel. Such a goal orientation allows the young child to develop expectations for emotional experience and beliefs about 'the right way' to feel. Socializing agents, whether they are parents, peers, or the media, influence what is liked or not liked, and thus affect children's beliefs about how to feel when they do or do not get what they want. Harris (1995) has also focused on the role that children's cognitive skills have on their awareness of emotions relative to contexts. He makes an important distinction in children's understanding of intentional targets of emotions vs. causes of emotions. He contends that children primarily anchor their emotions in appraisal of the object, person, or event at which the emotion is directed (e.g., mad *at* you, scared *of* spiders), as opposed to the causal or precipitating event (e.g., a squabble between siblings leads to anger, finding a big spider in one's bed results in fear). The significance of this idea of intentional targets of emotions as a contextual anchor for children's understanding of their subjective emotional experience is important theoretically, for it points toward the transactional encounter between person and environment; that is, children construe their emotions as directed at something. Functionalist views of emotion as described earlier emphasize this goal-oriented nature of emotion as opposed to other theoretical positions that emphasize the biological causes of emotions.

2. Emotion Understanding and Skills of Emotional Competence

A useful way to examine children's understanding of emotion is to think of their acquiring a set of skills that facilitate their being able to pursue their goals efficaciously. My use of the term *emotional competence* is such a construct, and Table 1 lists those skills relevant both to emotional competence and emotion understanding (Saarni 1999). These skills are learned in conjunction with children's cognitive development and the social contexts in which they are engaged. I shall address children's understanding of emotion relative to each of these six skills, and in some cases refer the reader to other articles in the encyclopedia that are likely to provide more detailed coverage.

Table 1
Skills relevant to emotional competence and understanding

1. Awareness of one's emotional state, including the possibility that one is experiencing multiple emotions, and at even more mature levels, awareness that one might also not be consciously aware of one's feelings due to unconscious dynamics or selective inattention
2. Skill in discerning others' emotions, based on situational and expressive cues that have some degree of cultural consensus as to their emotional meaning
3. Skill in using the vocabulary of emotion and expression terms commonly available in one's subculture and at more mature levels to acquire cultural scripts that link emotion with social roles
4. Capacity for empathic and sympathetic involvement in others' emotional experiences
5. Skill in realizing that inner emotional state need not correspond to outer expression, both in oneself and in others, and at more mature levels the ability to understand that one's emotional-expressive behavior may impact on another and to take this into account in one's self-presentation strategies
6. Capacity for adaptive coping with aversive or distressing emotions by using self-regulatory strategies that ameliorate the intensity or temporal duration of such emotional states (e.g., 'stress hardiness')

2.1 Understanding One's Own Emotional Experience

Given that infants' emotional experience and the self-conscious emotions (e.g., shame, pride) are addressed in *Infancy and Childhood: Emotional Development* and *Self-conscious Emotions, Psychology of*, respectively, I shall focus on how older children come to understand that they can have multiple feelings directed at the same target or situation. Young children's experience of ambivalence has been found to be a sequential process with different appraisals attached to the polarized emotional responses. Research with older school-age children suggests that they have developed the ability to consider at the same time one set of ideas that is at odds with another set of ideas, enabling them to integrate opposite valence emotions (happy and sad) about different targets that co-occur in a situation (e.g., 'I'm glad I get to live with my dad, but I'm sad about not being able to live with mom too'). In early adolescence simultaneously opposite valence emotions

about the same target can be integrated (e.g., 'I love my dad, even though I'm mad at him right now for not following through on his promise'). In late childhood children also understand that emotions interact with one another; that is, a positive emotion following a negative one can ameliorate the negative feelings one had and vice versa. Similarly, older children report that memories of previous feelings can influence their experience of current emotions. It is not known whether younger children might experience simultaneously two (or more) emotions but can only cognitively construct an explanation about the experience that focuses on one emotion at a time. When children experience an intense emotion, they must also cope with their feelings and the eliciting situation. During this coping process, they may well begin to cognitively construct a more complex system of appraisals about the emotion-eliciting situation or relationship. Finally, younger children (preschoolers to about age seven) scan a situation for its emotional meaning and finding one, stop scanning. Older children do not exhaust the search for further possible appraisals that might yield multiple emotional responses.

2.2 Understanding Others' Emotional Experience

By late preschool, children can decode the culturally common meanings of prototypical emotional facial expressions, and they have learned as well common situational elicitors of emotions. During this age period they make steady progress in realizing that others have minds, intentions, beliefs, or what has otherwise been referred to as 'inner states.' At around 8–10 years of age they understand the need to take into account unique information about the other that might qualify or make comprehensible a nonstereotypic emotional response or a response that differs from how oneself would feel in the same situation. The attributions children apply to their expectations for which emotions are likely can be characterized by three dimensions: *stability*, which refers to the extent an event will recur; *locus*, which refers to whether the causes of emotion lie within the person (internal locus and thus stable) or in a context (external locus and thus variable). Another feature is the perceived *controllability* of the eliciting event. Different combinations of these attributions have been found to characterize different features of children's understanding of others' as well as their own emotional experience.

2.3 Understanding Emotional Lexicon and Scripts

Children's ability to represent their emotional experience through words, imagery, and symbolism of varied sorts allows them to communicate their emotional experiences to others across time and space, and these representations of emotional experience facilitate further cognitive elaboration and comparison with others' representations about emotional experience. Some of the developments in awareness of our own multiple feelings or in understanding others' emotions described earlier could not be undertaken without access to a language or representational system for symbolically encoding and communicating emotional experience. Research on early mother–child feeling-state conversations indicates that access to an adult who is interested in one's feelings may be pivotal to children having opportunities to talk about emotions and acquire a lexicon of emotion vocabulary. Mothers tended to use conversations about feelings as a functional way to guide or explain something to their children, whereas the children were more likely to use feeling-descriptive words simply to comment on their own reaction or observation of another (Dunn et al. 1991). Thus, they were learning to communicate their own self-awareness of feeling states to their mothers, who in turn were likely to communicate *meaningfulness* to their children by using guiding, persuading, clarifying, or otherwise interpretive feeling-related language. Other research has also documented that if families were characterized as high in frequency of anger and distress expression, the children were less likely to be engaged in discourse about feelings. But if the families were low in frequency of negative emotional expression, then when a negative emotional event did occur for the child, there was a greater likelihood of an emotion-related conversation to ensue between child and parent (Dunn and Brown 1994). Acquisition of emotion-descriptive concepts continues throughout childhood and into adolescence, but little research has examined these older age groups. One way that emotion language develops further in the school-age child and adolescent is evidenced in their greater ability to add variety, subtlety, nuance, and complexity to their use of emotion-descriptive words with others.

2.4 Understanding Empathy and Sympathy

With empathy children affectively share in another's emotional experience; with sympathy they understand another's emotional experience but are more detached from it. Some individuals experience vicariously another's emotion so strongly that it becomes aversive; they tend to focus on alleviating their own personal distress and generally are not likely to show prosocial behavior toward the distressed person. Preschoolers are capable of responding empathically to others' distress, and what continues to develop in their understanding is what to do if they are to respond prosocially and appropriately. As they learn more

subtle cues of emotion, their range for responding empathically or sympathetically is similarly extended. *Prosocial Behavior and Empathy: Developmental Processes* considers this topic in much greater detail.

2.5 Understanding of Emotion Dissemblance and Management of Expressive Behavior

Research has demonstrated that by 10–11 years, children understand that the expression of emotion, whether genuine or dissembled, is a regulated act. At this age they are able to make reference to the degree of affiliation with an interactant, status differences, and controllability of both emotion or circumstances as contextual features that affect the genuine or dissembled display of emotion. However, across all ages, the most common reason cited for when genuine feelings would be expressed was if they were experienced as very intense (and thus less controllable). When children are asked to explain why feelings are not genuinely expressed, four broad categories of motivation are apparent in their responses: (a) to avoid anticipated negative outcomes, (b) to protect one's self-esteem or to cope more effectively with how one feels, (c) to avoid hurting someone else's feelings, and (d) to abide by norms and conventions (e.g., rules of etiquette). By school entry children generally understand that how one looks on one's face is not necessarily how one feels on the inside. They can often provide justifications for how appearances can conceal reality, in this case, the genuine emotion felt by an individual. Many of these justifications included describing the intent to conceal one's feelings and to mislead another to believe something other than what was really being emotionally experienced. Children younger than six can readily adopt pretend facial expressions, but they are not likely to be able to articulate the embedded relationships involved in deliberate emotional dissemblance. These embedded relationships refer to how the self wants another to perceive an apparent self, not the real self. In other words, by age six children readily grasp that emotional dissemblance has as its basic function the creation of a false impression on others.

2.6 Understanding of Emotion Regulation and Coping Strategies

Two general patterns have emerged with regard to age: as children mature, they generate more coping alternatives to stressful situations, and they become more able to make use of cognitively-oriented coping strategies for situations in which they have no control. Embedded in both of these age-related patterns is greater cognitive complexity that is associated with becoming older: (a) the ability to appraise accurately when a situation is simply not under one's control; (b) the ability to shift intentionally one's thoughts to something else less aversive; (c) the ability to use symbolic thought in ways that transforms the meaningfulness of a stressful encounter or situation (reframing); and (d) very importantly, the ability to consider a stressful situation from a number of different angles and thus consider different problem solutions relative to these different perspectives. As to when these general developmental patterns in coping are clearly in place, it is likely that by 10 years old most children will have a fairly well-developed coping repertoire that includes the emotion-focused strategies of cognitive distraction and transformation. This assumes, of course, that the family lives of these children have been adequately supportive and that harsh traumas have not preoccupied the children's emotional resources such that inflexibility or dissociative processes have taken hold in the child's self- and emotion-regulation. In conclusion, children's understanding of emotion develops in tandem with their exposure to social situations in which emotions are elicited and with their cognitive development. Enhanced emotion understanding facilitates children's social competence and provides a buffer against pathology, whereas deficits in emotion understanding may function either as precursors to subsequent maladaptive functioning or as outcomes of pathology.

See also: Conflict and Socioemotional Development; Coping across the Lifespan; Culture and Emotion; Emotions, Psychological Structure of; Friendship: Development in Childhood and Adolescence; Infancy and Childhood: Emotional Development; Infant and Child Development, Theories of; Personality Development in Childhood; Play and Development in Children; Prosocial Behavior and Empathy: Developmental Processes; Self-development in Childhood; Self-regulation in Childhood; Social Cognition in Childhood; Social Competence: Childhood and Adolescence; Social Learning, Cognition, and Personality Development; Socialization in Infancy and Childhood; Temperament and Human Development; Values, Development of

Bibliography

Campos J J, Mumme D L, Kermoian R, Campos R G 1994 A functionalist perspective on the nature of emotion. *The Development of Emotion Regulation. Monographs of the Society for Research in Child Development* **59**: 284–303

Dunn J, Brown J 1994 Affect expression in the family, children's understanding of emotions, and their interactions with others. *Merrill-Palmer Quarterly* **40**: 120–37

Dunn J, Brown J, Beardsall L 1991 Family talk about feeling states and children's later understanding of others' emotions. *Developmental Psychology* **27**: 448–55

Harris P, Lipian M S 1989 Understanding emotion and experiencing emotion. In: Saarni C, Harris P L (eds.) *Children's Understanding of Emotion*. Cambridge University Press, Cambridge, UK, pp. 241–58
Harris P L 1989 *Children and Emotion: The Development of Psychological Understanding*. Basil Blackwell, Oxford, UK
Harris P L 1995 Children's awareness and lack of awareness of mind and emotion. In: Cicchetti D, Toth S L (eds.) *Emotion, Cognition and Representation*. University of Rochester Press, Rochester, NY, Vol. 6, pp. 35–7
Kitayama S, Markus H R (eds.) 1994 *Emotion and Culture*, 1st edn. American Psychological Association, Washington, DC
Lazarus R S 1991 *Emotion and Adaptation*. Oxford University Press, New York
Levine L 1995 Young children's understanding of the causes of anger and stress. *Child Development* **66**: 697–709
Saarni C 1999 *The Development of Emotional Competence*. Guilford Press, New York
Wellman H M, Banerjee M 1991 Mind and emotion: Children's understanding of the emotional consequences of beliefs and desires. *British Journal of Developmental Psychology* **9**: 191–214

C. Saarni

Emotions, Evolution of

The study of the evolution of emotion dates back to Charles Darwin. There are at least three different approaches to the subject in the current literature. The widely accepted 'basic emotions' view developed in the 1960s out of human ethology and the psychological and physiological study of emotion. The evolutionary psychology movement and the paralanguage or 'behavioral ecology view' draw in two very different ways on ideas that originated in behavioral ecology and sociobiology during the 1960s and 1970s to suggest amendments to this orthodox view.

1. *The Darwinian Legacy*

Darwin did not see his *Expression of the Emotions in Man and Animals* (Darwin 1872) as a theory of the evolution of emotion itself, because he defined emotions as subjective feelings. Following Herbert Spencer, Darwin distinguished the emotions as those feelings caused by external states, in contrast to those caused by bodily states, such as hunger and pain. Later evolutionary theorists have preferred behavioral and physiological definitions of emotion, with the result that Darwin's theory of the expression of emotions has been interpreted as a theory of the emotions themselves.

Darwin proposed three principles of the evolution of expression (see also *Emotion and Expression*). The most important of these was the 'principle of serviceable associated habits,' which is a straightforward application of Darwin's theory of instincts to the case of emotion. Darwin believed that instinctive behaviors derive from habits acquired by psychological reinforcement. The consistent acquisition of the same habit for many generations causes it to become a hereditary, or instinctive, behavior by the inheritance of acquired characteristics, in which Darwin was a firm believer. Most of the distinctive behaviors associated with particular emotions, such as the erection of the hair in fear, reflect long-since vanished lifestyles in which those behaviors were rewarded and reinforced in each generation until they were finally incorporated into the hereditary material as instincts. Darwin supplements this principle with two others, the 'principle of antithesis' and the 'principle of direct action.' His antithesis principle postulates an intrinsic tendency for opposite states of feeling to produce opposite behaviors. Darwin remarks of a submissive dog that:

> Not one of the movements, so clearly expressive of affection, are of the least direct service to the animal. They are explicable, as far as I can see, solely from their being in complete opposition or antithesis to the attitude expressive of anger (Darwin 1872, p. 51).

Darwin explains the behaviors left over after the application of these two principles as the results of the 'direct action' of the nervous system. Excess nerve energy built up in an emotional episode is released in behaviors such as sweating and trembling for no other reason than that it must go somewhere and that these channels are physiologically available for its release.

2. *The Ethological Tradition*

The founders of ethology saw themselves as the direct heirs of Darwin's work on mental evolution (Lorenz 1965). Their account of the evolution of emotion retains Darwin's principles, but reinterprets them to fit the theory of evolution as it emerged in the 1930s in the 'modern synthesis' of Darwinism and Mendelian genetics. The principle of serviceable associated habits is transformed into the ethological concepts of 'ritualisation' and 'derived activity' (Tinbergen 1952). Derived activities are behaviors that originally evolved for one purpose but were later selected for another purpose. Ritualized behaviors are derived activities that originally evolved to fulfill some practical function but which were later selected to function as signals. Thus, although piloerection in fear and rage does not make a human being appear larger to an opponent, it does communicate his or her emotional state. Derived activities require a special pattern of evolutionary explanation. They cannot be understood

purely in terms of the function they currently perform and the selection pressures that currently maintain them in the population. This is particularly obvious in the case of signals. Piloerection is not intrinsically better as a signal of fear than smiling or laughing. This particular behavior was selected as a signal only because it was already associated with certain emotional states in the distant past. It was associated with those states, not because it was a signal, but because it made the animal appear larger.

The concept of ritualization allowed ethology to reconstruct Darwin's principle of serviceable associated habits whilst avoiding his commitment to the inheritance of acquired characteristics. Darwin's descriptions of the psychological rewards that led to the reinforcement of emotional behaviors are equally plausible as descriptions of the original selective advantage of those behaviors. Darwin's other two principles are equally open to reinterpretation. The principle of antithesis is explained by the selective value of unambiguous signals. It is as important for a dog to signal that it wants to avoid conflict as for it to signal aggression. Hence there can be selection of behaviors merely because they look different from the behaviors that signal aggression. The principle of direct action was transformed into the ethological concept of a displacement activity. Early ethologists shared Darwin's view that instinctive motivations cause a build up of mental energy that must be released in some behavior or other. An example commonly given is that of an angry cat that is unwilling to attack and begins to wash itself. Niko Tinbergen remarks: 'I think it is probable that displacements do serve a function as outlets, through a safety valve, of dangerous surplus impulses' (Tinbergen 1952, p. 23). This wholesale reinterpretation of Darwin's three principles works so smoothly and allows the retention of so much of the detail of Darwin's work that the early ethologists seem almost unaware of the differences between Darwin's theory and their own (Lorenz 1965).

3. *The Affect Program Theory of Basic Emotions*

Darwin's detailed accounts of human facial expressions were confirmed and extended during the 1960s and early 1970s and many of his suggested homologies between human and primate facial expressions were confirmed. In addition, it became generally accepted that some emotional expressions, the so-called 'basic emotions,' are found in all cultures and have an evolutionary history at least as long as the history of the human species (Ekman 1973, see also Emotion and Expression). The basic emotions are usually referred to as fear, anger, disgust, contempt, joy, sadness, and surprise, although in normal usage each of these words has a much wider meaning. During the same period, the hydraulic metaphors that the ethologists had used to describe psychological processes were replaced by a computational model of emotion. Each basic emotion corresponds to an 'affect program' stored somewhere in the brain. When activated, this program coordinates a complex of actions that includes facial expression, autonomic nervous system changes, expressive vocal changes, and muscular–skeletal responses such as flinching or orienting. The concept of an affect program inherits many of the features of the earlier ethological concept of a fixed action pattern. Both concepts suggest that certain apparently complex behaviors are, in reality, atomic units of behavior that unfold in the same, stereotyped sequence whenever they are triggered. The affect programs are controlled by an 'automatic appraisal mechanism.' This is a specialized neural system that applies its own distinctive rules for stimulus evaluation to a limited set of data derived from the earliest stages of the processing of perceptual information. Considered together, the appraisal mechanisms and affect programs form a cognitive module in the sense favored by more recent evolutionary psychologists (Barkow et al. 1992).

The affect program theory retains two distinctive features of the earlier, ethological model of emotions. First, emotional behaviors are the inevitable and involuntary accompaniment to an underlying emotional state and as such are reliable signals of that state. Second, deliberate attempts to disguise emotions are subject to 'leakage' from the operation of the involuntary emotional response. Voluntary attempts to suppress emotional behavior can only operate by simultaneously using the muscles involved in the expression for some other purpose. They cannot interfere with the actual operation of the emotion module (see *Emotion and Expression*).

The affect program theory was accompanied by a distinctive theory of the evolution of the emotion system. The system was interpreted as an ancient form of cognition that had originally operated on its own and had later been supplemented by higher cognitive functions. This view was supported by MacLean's theory of the 'triune brain,' according to which the emotions are located in the 'paleomammalian' portions of the brain while higher cognitive functions are realized in newer, 'neomammalian' structures (MacLean 1952). The survival of these ancient forms of behavior control in primates was explained by their value as fail-safe responses ensuring that vital behaviors are performed whenever necessary even if that means they are performed too often. This view of the emotion system as a collection of primitive but reliable fail-safe mechanisms remains influential (Panksepp 1998).

4. *Sociobiology and the Emotions*

Sociobiology brought a new perspective to bear on the evolution of emotion in the 1970s and 1980s. It also moved the focus of investigation from the basic

emotions to the moral and quasi-moral emotions involved in human social interaction. Emotions such as trust, loyalty, guilt, and shame play an obvious role in mediating the competitive social interactions that were the focus of most research in human sociobiology. Numerous sociobiologists made brief comments to the effect that moral emotions must have evolved as psychological mechanisms to implement evolutionary stable strategies of social interaction (Weinrich 1980). Robert A. Frank suggested that the moral emotions evolved as solutions to 'commitment problems' (Frank 1988). A commitment problem arises when the winning strategy in an evolutionary interaction involves making a binding but conditional commitment to do something that would be against one's own interests if the condition were ever met. If such a commitment is to be credible, some special mechanism is needed which would cause the organism to act against its own interests. Frank suggests that emotions such as rage and vengefulness evolved to allow organisms to engage in credible deterrence, threatening self-destructive aggression to deter a more powerful aggressor. Conversely, emotions such as love and guilt evolved to allow organisms to engage in reciprocal altruism in situations where no retaliation is possible if one partner fails to reciprocate.

Sociobiologists criticized ethology for its lack of a theoretical framework with which to predict how humans would behave, accusing it of being little more than descriptive natural history (Barash 1979, Barkow 1979). In contrast, sociobiology seemed to make strong predictions that clashed with some aspects of the affect program theory of basic emotions. From a sociobiological perspective it makes no sense for organisms to possess involuntary expressive behavior. The application of evolutionary game theory to emotional behavior predicts that this behavior will be designed to manipulate the expectations of other organisms rather than to 'express' emotional states transparently. This theoretical argument has been used by some to reject the affect program view of basic emotions. Alan Fridlund has argued that emotional behavior should be regarded as a paralanguage of social signals whose production depends at least as much on an organism's social context as on its emotional state (Fridlund 1994). Fridlund and other paralanguage theorists have documented audience effects on the production of the basic emotions and have argued that this is inconsistent with the affect program theory. They have also tried to show that the classic facial expressions of emotion are not strongly correlated with underlying emotional state and that people rely more on context than on behavior in attributing emotions to others (Russell and Fernández-Dols 1997). It is unclear, however, that this data contradicts the predictions of the affect program theory. Signaling behavior exhibits audience effects in many organisms, such as domestic chickens, in which that behavior is presumably part of a highly stereotyped behavior sequence controlled by a relatively simple cognitive mechanism.

Mark Hauser has suggested that Fridlund's arguments bear on questions about the biological function of emotional behavior whilst the affect program model is concerned with the mechanisms that produce that behavior (Hauser 1996, pp. 495–6). In some places, however, Fridlund does seem to be discussing the nature of the underlying emotional processes and not merely their biological function. He argues that since emotions are signals and are produced because of the impact they will have on other organisms, a particular emotion defined behaviorally need not correspond to any particular motivational state (emotion) of the organism producing it.

5. *Evolutionary Psychology and the Emotions*

The evolutionary psychology movement of the late 1980s and 1990s has been less inclined than sociobiology to question the accepted view of basic emotions. The evolutionary psychologists believe that the mind as a whole is made up of modules that operate independently of one another. Like the emotion system, these modules operate on specific kinds of data using algorithms that differ from those used by other modules. Hence, evolutionary psychology endorses the affect program theory of basic emotions, but wants to go further, both by adding to the complexity of the known affect programs and by finding modular mechanisms underlying other emotional behaviors. John Tooby and Leda Cosmides suggest a number of cognitive and behavioral systems that they would expect to receive output from the affect programs, in addition to those currently known to receive such output (Tooby and Cosmides 1990). They also suggest that the affect programs will be elicited by simple, perceptual stimuli that have played a recurrent role in human evolutionary history. Current empirical research suggests, however, that few simple perceptual stimuli elicit affect programs in the absence of any individual experience of those stimuli. Most evolutionary influences on sensitivity to simple perceptual stimuli seem to take the form of learning preparedness: the relative ease with which one association is acquired rather than another. Genuine pan-cultural antecedents of emotion seem to be relatively abstract states of affairs having to do with the significance of a situation for the organism and requiring local knowledge to translate them into perceptible features of the stimulus situation (Ekman and Davidson 1994, pp. 144–77).

The leading evolutionary psychologist, David Buss, has argued for the existence of a module for sexual jealousy, one of the additional modules predicted by Tooby and Cosmides. Buss argues that sexual jealousy has simple perceptual elicitors such as unusual scents, changed sexual behavior, excessive eye contact, and

violation of rules governing personal space. 'These signals alert us to the possibility of infidelity, since they have been statistically linked with relationship loss over the long course of human evolutionary history' (Buss 2000, p. 45). The jealousy module uses special-purpose algorithms and, like the basic emotions, it functions as a fail-safe mechanism: 'failing to detect an infidelity was more costly than mistakenly accusing an innocent partner of betrayal. Evolution, as a consequence, forged a hypersensitive defense system, designed to sound the alarm not just when an infidelity has been discovered, but also when the circumstances make it slightly more likely' (Buss 2000, p. 224). When triggered, the module produces various forms of violence against female sexual partners, including, under certain environmental conditions, murder. A noticeable contrast between these recent theories and more traditional accounts of the evolution of emotion is the absence of the idea that emotions represent a more primitive form of behavioral control that can be contrasted to rational, planned action. The emotions are seen as just another cognitive module reflecting details of the environment of evolutionary adaptedness.

6. Conclusion

Current theories of the evolution of emotion all stem from the intellectual lineage stretching back to Darwin himself. There are a number of disagreements between current researchers, however. These include the number of separate, evolved emotions, the nature of the stimuli that produce evolved emotions, and a cluster of issues about the evolutionary function of emotional behaviors and their relationship to underlying emotional states, issues which are often discussed in terms of whether emotional behaviors should be primarily regarded as expressions or as signals.

See also: Culture and Emotion; Emotion: History of the Concept; Emotion, Neural Basis of; Emotions, History of; Emotions, Sociology of; Infancy and Childhood: Emotional Development; Facial Expressions

Bibliography

Barash D P 1979 Human ethology and human sociobiology. *Behavioral and Brain Sciences* **2**(1): 26–7

Barkow J H 1979 Human ethology: Empirical wealth, theoretical dearth. *Behavioral and Brain Sciences* **2**(1): 27

Barkow J H, Cosmides L, Tooby J (eds.) 1992 *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, Oxford, UK

Buss D M 2000 *The Dangerous Passion*. Simon and Schuster, New York

Darwin C 1872 *The Expressions of Emotions in Man & Animals*. Philosophical Library, New York

Ekman P (ed.) 1973 *Darwin & Facial Expression: A Century of Research in Review*. Academic Press, New York

Ekman P, Davidson R J (eds.) 1994 *The Nature of Emotion: Fundamental Questions*. Oxford University Press, New York

Frank R H 1988 *Passions within Reason: The Strategic Role of the Emotions*. Norton, New York

Fridlund A 1994 *Human Facial Expression: An Evolutionary View*. Academic Press, San Diego, CA

Hauser M D 1996 *The Evolution of Communication*. MIT Press, Cambridge, MA

Lorenz K 1965 Preface In: Darwin C *The Expression of the Emotions in Man and Animals*. University of Chicago, Chicago

MacLean P D 1952 Some psychiatric implications of physiological studies on frontotemporal portions of the limbic system (visceral brain). *Electroencephalography & Clinical Neurophysiology* **4**: 407–18

Panksepp J 1998 *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, Oxford, UK

Russell J A, Fernández-Dols J M 1997 *The Psychology of Facial Expression*. Cambridge University Press, Cambridge, UK

Tinbergen N 1952 Derived activities: their causation, biological significance, origin and emancipation during evolution. *Quarterly Review of Biology* **27**(1): 1–32

Tooby J, Cosmides L 1990 The past explains the present: emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology* **11**: 375–424

Weinrich J D 1980 Towards a sociobiological theory of emotions. In: Plutchik R, Kellerman H (eds.) *Emotion: Theory, Research & Experience, Vol 1: Theories of Emotion*, Academic Press, NY, pp. 113–140

P. E. Griffiths

Emotions, History of

1. Using the Past for New Understanding

This field of study builds on two assumptions. First, historians of emotion believe that emotion, the response to stimuli that combines physiological reactions and cognitive evaluations, plays a significant role in historical action. In this they differ from some historians who have argued, for example in interpreting rioting, that purely rational calculations underlie behaviors even when there are admitted emotional overtones such as anger or fear. Second, historians of emotion believe, and demonstrate, that emotions are subject to change, according to shifts in historical context. In this they differ from some researchers who argue that emotions are human constants, based on inherent biological and psychological factors, varying only by personality. Historians of emotion join anthropologists and many sociologists in seeing emotional standards and reactions as group variables, at

least in part. Whereas anthropologists mainly study different emotional configurations across contemporary cultures, historians of emotion deal with changes over time, venturing a periodization of key developments. Historians of emotion also contribute to emotions-study a concern for the social results of emotional standards, as opposed to purely personal qualities.

2. *Major Themes*

Historical research has become a significant contributor to the exploration of emotion and the role of emotion in larger social interactions. Although there are a number of different approaches to emotions history, and several different rationales, most historians agree in the focus on 'changes' in emotional standards and experience. They seek to chart what the changes were, granting that explicit standards, or what is often called emotional culture or emotionology, are easier to delineate than actual individual or collective experience. However, they demonstrate that standards themselves, have important results both in emotional life and in social and institutional behavior. Here, indeed, is one of the chief methodological pillars of the field. Historians also try to explain what the 'causes' of change were, which helps connect emotions to larger social developments. They are also eager to show what the 'results' were, not just in individual perceptions of emotions and capacities for display, but in other institutions and interactions.

3. *How the Field Became Explicit*

History cannot be written without dealing with emotional issues and expressions. Even before the rise of psychohistory—the application of psychological theories, particularly Freudian psychoanalytic concepts, to historical data—biographers often referred to emotional qualities of their subjects. Classic attempts to convey the larger spirit of a historical period, like Johan Huizinga's characterization of the waning Middle Ages, often included emotional dynamics, as did descriptions of reactions to wars or crowd behavior. With all this, however, the principal focus was not on emotion *per se*, but on emotion as a component of some other phenomenon.

The history of emotions received a considerable boost in the 1930s. The rise of social history in France, seeking to understand the experience of ordinary people in the past, involved exploring topics not previously part of the historical repertoire. Innovators like Lucien Febvre (1973) specifically called for research on emotions as part of the larger social history effort. At about the same time, German sociologist Norbert Elias sketched his theory of a 'civilizing process,' (Elias 1978, 1982) arguing that standards of self-control became more rigorous in Europe's upper classes in the seventeenth and eighteenth centuries. His primary focus involved physical behavior, such as eating habits, but his attention to more refined manners also involved discipline over emotion. Neither Febvre nor Elias generated an immediate response; indeed, Elias' theories are used currently more widely than ever before.

The development of psychohistory in the 1970s certainly called new attention to the historical exploration of emotion. Historians of childhood like Philip Greven (1977) and David Hunt (1970) thus explored the emotional impact of characteristic styles of childrearing in the seventeenth and eighteenth centuries, using psychoanalytic theory; Greven focused on the consequences of new styles of discipline associated with religious movements in colonial America. Most psychohistorians, however, concentrated primarily on biography, as in the classic study of Luther by Erik Erikson (1962). Still, historians influenced by Freudianism have continued to contribute, particularly when they have turned to larger cultural standards and individual experience around these standards. Generational emotional results of the experience of World War I have thus figured in some explanations of the rise of German Nazism (including the impact of father absence).

However, most explicit attention to emotions history resulted from a slightly different set of developments, in social history and related social sciences, as a framework established in the 1970s translated into explicit research from the mid-1980s onward. Social historians, now dominant in many areas of historical research, continued to widen their exploration of how ordinary people functioned in the past. A major extension of these inquiries, launched in France but spreading more widely, argued for inquiry into deeply-held values and assumptions—*mentalités*—concerning subjects such as death, community, or the forces of nature. Not surprisingly, attitudes about emotion quickly figured into this mix; sweeping studies of fear were one key result. English and American historians, for their part, broadened the investigation of family patterns into qualitative relationships among family members. In this vein they explored shifts in the definition and valuation of love, both marital and parental, and also changes in the ways anger was viewed in family life. Mentalities and family historians alike painted a distinctive picture of emotional standards in early modern—sixteenth to seventeenth century—European and American societies, compared to more modern norms. They also paid explicit attention to processes of change, particularly by the eighteenth century.

Two related strands of social history inquiry thus produced a new understanding of the extent to which past culture, and institutions like the peasant village or the family, could not be grasped without assessing the emotional dynamics involved. The same inquiry

underlined the extent to which emotional dynamics varied over time—with the historical period—and were therefore subject to explainable change.

Concurrently, various social scientists exploring emotion, began to probe the historical record as well, in this case with primary, though not exclusive, attention to more recent developments. American sociologists and social psychologists began to look at new emotional standards at work, as service sector jobs expanded in the twentieth century. They also investigated changing definitions of family-related emotion. While most anthropologists concentrated on painting rich, but fairly static pictures, of emotional standards in various societies, some began to build in attention to change. Thus Robert Levy, in his classic study of the Tahitians (1973), dealt with the impact of Western colonialism on Tahitian emotional culture, while Catherine Lutz (1988) in a variety of studies looked at shifts in emotion over time, while relying heavily on contemporary field work descriptions for her basic categories. A number of European sociologists, some strongly influenced by Elias' theories, began to examine changes in 'emotion management,' particularly in the twentieth century but in some cases over a longer sweep of time, focusing on ways emotional impulses were to be controlled. A Dutch group, in particular, worked on the emotional standards that accompanied more informal manners and relationships in the twentieth century. By the 1980s, explicit work in the history of emotion was well underway (see Wouters 1995).

4. *Gains and Gaps in Historical Research on Emotion*

Once launched, emotions history inevitably gained additional vantagepoints and emphases. The list of emotions historically explored now includes envy, disgust, joy (tentatively), grief, shame, and guilt, as well as the initial roster.

The takeoff of emotions history has also widened in terms of place and time—crucial indications of a project's significance in this empirically specific discipline. Thus several studies deal with changes and continuities in emotional standards and expressions in Chinese history. Historians addressing medieval Europe have generated an impressive literature, as in an important set of studies on anger and projects on wonder and empathy (Rosenwein 1998). Classicists are also dealing with love, in the context of further explorations of family history. The growing range still begs for expansion and connection. Changes within the Middle Ages need more explicit linkage to the patterns noted in the early modern centuries, while the Renaissance, as a possible period in emotions history, has not clearly surfaced at all. Geographical gaps are at least as obvious as the challenges in chronology and linkage. Specialists dealing with a number of key areas—Eastern Europe, Africa, Latin America, additional parts of Asia—have yet to generate focused histories of emotion, as opposed to more haphazard overlaps and connections with other topics like family life. Relatedly, key theories about emotional change—such as the civilizing process idea—have yet to be seriously tested beyond the Western context, so despite some speculation there is no consensus about possible applicability. Nevertheless, it is fair to say that the process of moving emotions history toward a full historical canvas is underway.

5. *Two Turning Points*

The most elaborate work, however, continues to emphasize developments in the USA and Western Europe. Here, in addition to broadening the varieties of emotions considered and the contexts in which they operate, historians and kindred social scientists have particularly emphasized two points of particularly revealing transformation. An important set of changes redefined emotional culture between the late seventeenth and early nineteenth centuries, with timing somewhat varying with the emotional configuration involved. A subsequent turning point intervened in the second quarter of the twentieth century, possibly a bit earlier in the USA than in Western Europe. Within this periodization framework a host of questions remain disputed or unanswered, but the framework itself has proved widely useful in explaining significant recastings of emotional guidelines and experience and in beginning to amass evidence about why emotions change and with what larger results.

In the early modern emotionological transformation, two factors propelled change. First, growing commercialization helped place new value on the family, as communities became somewhat more competitive. Family ties were alternatives to older, now frayed communitarian bonds. Same-sex friendships, by the same token, declined. Along with this, Protestantism, attacking the idea of special credit for celibacy, greatly improved the religious status of marriage (and, in destroying convents, made marriage more essential for women). By the seventeenth century Protestant writers urged emotional sharing in marriage. From these twin bases, actual emphasis on love as a key ingredient in forming and evaluating marriage measurably expanded by the mid-eighteenth century. A concomitant, noteworthy by this point, was a more systematic attack on anger in the family, as contradictory to love and also to the greater equality in human relationships being promoted by Enlightenment ideology. Jealousy was also redefined and more systematically reproached, particularly as the idea of a jealous defense of honor came under attack as uncivi-

lized. Growing emphasis on parent–child love, combined with further erosion of community controls, helped generate the growing reliance on guilt over shame in childrearing (Stearns and Stearns 1988). Techniques such as isolation increasingly supplanted both shaming and physical punishments, convincing children that love was temporarily interrupted and could be restored only by sincere repentance—often establishing guilt as a durable emotional enforcer of character standards. Grief, though more fully amplified in the nineteenth century, also followed from a greater emphasis on love, along with new beliefs that deaths should be brought under greater control. The apparatus available to commemorate death changed in order to accommodate growing grief needs (Rosenblatt 1983).

Emotional change also developed outside the family. The acceptability of sadness and melancholy declined, with people (initially, particularly men) urged to present a cheerful demeanor. More broadly, emotions were increasingly seen as operating under individual control, rather than sweeping through the body as an out-growth of physical humors. Correspondingly, responsibility for personal emotion management increased. Words like temper began to suggest a type of anger that should be controlled, while outright neologisms, such as tantrums, indicated a childish level of rage that could disqualify an adult from serious consideration (Stearns and Stearns 1986). The rise of new standards of disgust followed from cultural redefinitions of emanations from the body and from growing social inequality, with disgust used to mark and justify boundary lines between respectable and unrespectable people and situations (Corbin 1986).

Historians are still working out some of the implications of these changes, in altering the context of family life but also spilling over into wider public arenas. The transformation of criminal justice, shifting from shaming to imprisonment, clearly reflected the new commitment to guilt, for example, along with growing disgust at crowd reactions to public punishments. Ramifications of emotional change could reach into a wide spectrum.

Redirections in emotional culture between 1900 and 1950 were also extensive. Here, the dominant motifs are greater control over intensity plus growing informality in social relationships. Some historians also add an increasing democratization of emotional standards, with fewer systematic social divisions between respectable groups and others. Suspicion of intensity showed in several arenas. Anger, previously approved for men as a motivation for business and professional life (though dangerous in the home), was now more systematically reproved, with new techniques designed to defuse it. Grief was increasingly disallowed. Even more revealing attacks on intensity turned against too much stress on guilt, which might damage children's self-confidence. Maternal love, praised without major qualification in the nineteenth century, now fell victim to extensive criticism because of its overpowering qualities; mothers should keep their affection under tighter bounds. Romantic love, which had evolved toward increasingly idealistic statements in the nineteenth century, seemed excessive as well by contemporary standards; many prescriptive writers urged more limited, realistic emotional connections, while also freeing sexuality from some of the emotional commitment promoted in nineteenth-century advice literature (Seidman 1991, Stearns 1994).

The new concern about intensity was no accident, and it had some of the same range of effects noted in the earlier Western transformation. Causes centered on changes in economic structure, and therefore emotional functions, along with the rise of a new, psychologically-informed expertise. Growing consumerism urged emotional investment in acquisition and goods, and this diverted from some of the emotional investments that figured powerfully in the eighteenth and nineteenth centuries. Corporate and service employment required higher levels of emotional self-manipulation, in order to turn on a friendly smile and avoid bureaucratic confrontations. 'Impersonal, yet friendly' became a literal watchword in business training programs, with considerable emotional implications. At the same time, popularizing experts now referred to their psychological, rather than moral credentials, creating a picture of a more fragile mental structure than had previously prevailed. A recent study (Pfister and Schnog 1997) further probes the particular American fascination with psychological explanations and remedies, as a framework for contemporary emotional life.

The consequences of change involved, as with earlier redirections, both personal and social features. Americans became more nervous about emotional display, convinced that intensity reflected immaturity on their part; hence a recent study emphasizes a proclivity to check with others about onsets of jealousy, in order to provide a peer-mirror to guard against lack of restraint. Other studies (Salovey 1994) show the consequences of rigorous emotional control on the job, in distorting responses even in private relationships. Legal changes, including the rise of no-fault divorce, incorporated the growing belief that moderns should be able to keep anger and grief in check, in the most trying circumstances. Even changes in protest rates—the marked decline of labor strikes after 1958—derived in part from successful efforts to persuade workers that anger was childish, a reflection of personal flaw rather than a valid emotional response to objective problems. The reach of this new change in basic emotional style, in sum, was considerable.

Finally, amid growing constraints on certain emotional expressions, a recreational culture arose that provided some symbolic contrasts. Intense spectatorship, for example at sporting events, played off against

normal emotion management, and athletes and actors were allowed displays of anger, grief, or vivid camaraderie that were not usually countenanced—providing, in other words, what has been described as a compensatory surrogate culture.

Exploration of the two great transformations has involved the work of a substantial number of historians and historical sociologists. Inevitably, even in a new field, some disagreements have surfaced. Specific emphases and attributions of cause differ in the approaches to the early modern transformation. French work divides between explorations of distinctively Catholic features (as in the work on fear) and a sense that French emotional culture moved in harmony with other parts of Europe (as in the work on familial affections) (see Delumeau 1978). Divisions, though not outright clashes, also inform the work on the twentieth-century turning point. European sociologists have worked more on the process of 'informalization' than have their American counterparts, arguing that contemporary emotional relations work because people can assume successful incorporation of the rules of restraint (Wouters 1992). American historians accept this conclusion, in part, but additionally point to important new regulations and tensions. No coherent schools have emerged as yet, but there are fruitful debates.

Both of the great changes in emotional standards raise a host of questions, headed by the problem of social class. The rise of new standards of love in the eighteenth century involved a variety of social groups; proletarians, free from the constraints of property arrangements, may have moved early to emphasize a type of love more fully (Gillis 1985). But some of the standards particularly attached to the rising middle class. Emphasis on heightened emotional sensitivities between parents and children, for example, while no monopoly of a single class, moved into a bourgeois lexicon; and many observers criticized workers for failure to display the finer parental feelings. Changes in the twentieth century did have a democratic air, in that nineteenth-century attempts to downplay certain groups as impossibly crude noticeably declined. Romantic love was a case in point. American jurists in the late nineteenth century had been willing to allow lack of love as a valid reason for divorce for middle-class people, under the heading of mental cruelty, but they would not do the same for the working class because such emotional refinement was impossible at this level. In the twentieth century, however, the now-modified romantic love ideal was proclaimed for all. Still, issues of how particular emotional standards affected social classes are only now beginning to receive needed attention. One study, for example, shows how new hostility to the use of fear in childrearing, an American middle-class staple by 1900, captured Catholic working-class culture only in the 1950s. In other instances, middle-class prescriptions were not simply more slowly assimilated by other groups, but were positively defied, as in some of the more emotionally exuberant facets of twentieth-century African American culture (Stearns and Lewis 1988).

6. *Current Issues and Future Directions*

The success of historical work to date considerably adds to available information on varieties of emotional formulations and the ways values and emotions interact. It thus amplifies the constructionist approach to research on emotion, arguing that a significant part of emotional codes and responses vary with context—by place, but also by time period. The exploration of change and causation builds a further ingredient into constructionism, often missing from anthropological studies, in showing how one emotions variant may transform into another, responding to wider factors in economic, political, and cultural structure.

The historical agenda also raises problems, even aside from the range of areas and periods, and types of emotions, still open to this kind of analytical inquiry. The relationship between widely-prescribed cultural standards and other aspects of emotional experience—aspects that usually leave less explicit past records—is an obvious issue. Standards count in themselves: they affect social rules and practices, based on expected emotional behavior; they color evaluations of others and of self; and they unquestionably shape actual experience. Thus a culture that discourages romantic love—a common pattern in premodern peasant societies—almost certainly has less of it. But there are disjunctures as well, and historians must be sensitive to these. Among other things, emotional experience may vary somewhat less than standards do. Historians who initially assumed that the absence of explicit encouragement (indeed, the considerable explicit hostility) to marital love, in prescriptive materials before the seventeenth century, meant that there was in fact little, were assuming too much, and more recent researchers have improved on the evidence available in generating more nuanced statements of contrast and change.

A related issue, now happily beginning to gain constructive attention, involves subcultures. As noted previously, most emotions history to date has focused on mainstream standards issuing from dominant social groups. Various social classes, religious contingents, even age categories form emotional cultures of their own. Research in American history, particularly, is beginning to reveal patterns of emotionology in various religious and racial groups, and in the working class. These patterns sometimes link to mainstream trends, if with a certain chronological lag, but they sometimes deliberately run counter—as in the enthusiastic religious sects that arose around 1900 as an

antidote to growing mainstream hostility to emotional intensity.

Comparison also deserves more attention. The extensive focus on large transformations invite fine-tuning. Most research on both the early modern and the twentieth-century turning points has been either strictly national (just the USA or just France) or has assumed large transatlantic, Western trends. Yet we know that in the year 2000, in key particulars, regional patterns vary—the French get angry when jealous, the Dutch sad, while Americans scurry to get other people's opinions. We need explicit historical comparisons to determine when, and above all why, distinctions of this sort emerged.

History can also establish middle-level generalizations about emotional change, based on the accumulation of cases. We can begin to use partial models of the factors most likely to cause change, and about the length of time normally required between the introduction of widely-shared new standards, and actual changes in social arrangements and behaviors (including childrearing). It took about a century, for example, for warnings against scaring children with bogeymen stories to gain sufficient acceptance that the prescriptive emphasis could be cut back (after the 1920s). Thirty years elapsed between new strictures against anger in the workplace, and incorporation of the same standards in parenting advice; and almost 40 years between the new culture's inception and clear implementation in laws and protest levels. These are straws in the wind, toward more systematic statements about timing.

In what is still a new field of explicit research, despite promising results to date, dealing with empirical and theoretical gaps inevitably forms a challenging agenda for the future. Identifying subcultures has barely begun, for example in dealing with racial differences over emotional goals in the USA. Geographical omissions are troubling in themselves, and they limit opportunities for comparison. Even for the Atlantic world, where the literature is richest, comparative opportunities have barely been taken up. With all this, two additional challenges remain: further integration with history, and further integration with the other disciplines that examine emotion.

The number of historians interested in emotion has expanded steadily. But findings about emotional change are not usually incorporated into mainstream historical study. Family historians come closest to recognizing the importance of this topic, as an element for example, along with birthrates and the roles of family members, in establishing periodization. But the day when historians of protest or politics consider the role of shifts in emotional standards has not yet arrived. The connection is possible: at a time when the emotional reactions of American politicians are carefully scripted in light of the dominant emotionology, there is obvious invitation to explore mutual interactions between culture and political reactions in the twentieth century (Kofler 1997). The same holds for studies of sports or labor relations. While the record to date is tentative at best, there is reason to assume that fuller integration will occur, as historians become more comfortable with the idea that emotional standards shift and that shifts have important consequences.

The even greater challenge for the future is interdisciplinary. Historians of emotion have found ready audience among anthropologists, sociologists, and some social psychologists (particularly, of course, those theoretically wedded to a discursive, cultural-constructionist approach). But actual interdisciplinary projects, that would deal with change and impact in relation to current emotional manifestations, have yet to emerge. Mainstream psychology focuses primarily on invariant aspects of emotional display, such as facial expressions, and ignore issues of cultural variance and change (Harré and Stearns 1995). We lack the collaboration necessary to use historical analysis to help test the boundary lines between inherent emotional components and the role of cultural standards, yet precisely this collaboration constitutes a vital next step. Projects that would identify recent change and subject the emotional results to standard psychological scrutiny are possible—for example, the tentative signs, since the 1970s, of partial modification of the contemporary hostility to grief. But constructing the interdisciplinary apparatus that can handle such issues is no easy task, given the growing split between cultural analysis and scientistic psychology. The desirable agenda is obvious; its realization, distinctly problematic.

Even as things stand, the history of emotions is contributing not only to an understanding of emotion, but also to a wider assessment of emotion's impact on behaviors and social institutions. Too often, emotions research focuses on individual experience. Even in the short time explicit emotions topics have commanded their attention, historians have advanced the exploration of why particular emotional articulations matter—how they affect the experience of family life and friendship, how they connect to regimens at work, even—in some extensions currently being developed—how they can help shape new political styles. Historians will continue to probe emotional characteristics in the past, and particularly cases of significant change; but they will also be eager to fold emotion actively into our larger understandings of how societies function. Momentum continues to build in this active new disciplinary partner in emotions research.

See also: Biography: Historical; Body, History of; Collective Memory, Psychology of; Cultural History; Cultural Psychology; Elias, Norbert (1897–1990); Emotion: History of the Concept; Emotions, Evolution of; Emotions, Sociology of; Family and Kinship, History of; Love and Intimacy, Psychology of;

Psychiatry, History of; Psychoanalysis, History of; Psychohistory

Bibliography

Corbin A 1986 *The Foul and the Fragrant: Odor and the French Imagination*. Harvard University Press, Cambridge, MA

Delumeau J 1978 *La Peur en Occident, XIVe-XVIIIe Siècles: Une Cité Assiégée*. Fayard, Paris

Elias N 1978 *The Civilizing Process: the History of Manners*. Blackwell, Oxford, UK

Elias N 1982 *The Civilizing Process: State Formation and Civilization*. Blackwell, Oxford, UK

Elvin M 1989 Tales of the Shen and Xien: Body-personal and heart-mind in China during the last 150 years. *Zone* **4**: 266–349

Erikson E H 1962 *Young Man Luther: a Study in Psychoanalysis and History*. Norton, New York

Febvre L 1973 *A New Kind of History*. Harper and Row, New York (Original work published 1933)

Gay P 1984–1998 *The Bourgeois Experience: Victoria to Freud* (2 Vols.). Oxford University Press, New York

Gillis J R 1985 *For Better, for Worse: British Marriages, 1600 to the Present*. Oxford University Press, New York

Greven P 1977 *The Protestant Temperament: Patterns of Child Rearing, Religious Experience and the Self in Early America*, 1st edn. Knopf, New York

Harré R, Stearns P N (eds.) 1995 *Discursive Psychology in Practice*. Sage, London

Hunt D 1970 *Parents and Children in History: The Psychology of Family Life in Early Modern France*. Basic, New York

Kofler A (ed.) 1997 Emotion and culture. *Innovation: The European Journal of Social Sciences* **10**: Special Issue

Lantz H R 1982 Romantic Love in the Pre-Modern Period: A Social Commentary. *Journal of Social History* **15**: 349–70

Levy R I 1973 *Tahitians: Mind and Experience in the Society Islands*. University of Chicago Press, Chicago

Lutz C 1988 *Unnatural Emotions: Everday Sentiments on a Micronesian Atoll and Their Challenge to Western Theory*. University of Chicago Press, Chicago

Lystra K 1989 *Searching the Heart: Women, Men and Romantic Love in Nineteenth-Century America*. Oxford University Press, New York

Pfister J, Schnog N (eds.) 1997 *Inventing the Psychological: Toward a Cultural History of Emotional Life in America*. Yale University Press, New Haven

Rosenblatt P C 1983 *Bitter, Bitter Tears: Nineteenth Century Diarists and Twentieth-Century Grief Theories*. University of Minnesota Press, Minneapolis

Rosenwein B H 1998 *Anger's Past: The Social Uses of an Emotion in the Middle Ages*. Cornell University Press, New York

Lutz C 1988 *Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Challenge to Western Theory*. University of Chicago Press, Chicago, IL

Salovey P 1994 *The Intelligent Emotions*. Brenzel Publishing Company, Florence, KY

Seidman S 1991 *Romantic Longings: Love in America, 1830–1980*. Routledge, New York

Stearns C, Stearns P 1986 *Anger: The Struggle for Emotional Control in America's History*. University of Chicago Press, Chicago

Stearns C, Stearns P 1988 *Emotion and Social Change: Toward a New Psychohistory*. Holmes and Meier, New York

Stearns P 1994 *American Cool: Constructing a Twentieth-Century Emotional Style*. New York University Press, New York

Stearns P, Lewis J 1988 *Emotional History of the United States*. New York University Press, New York

Wouters C 1992 On status competition and emotion management. *Journal of Social History* **29**: 699–718

P. N. Stearns

Emotions, Psychological Structure of

As is the case with many concepts in the social sciences, the term 'emotion,' used widely in everyday language, constitutes a hypothetical construct, that is, a conceptual and operational definition of an underlying phenomenon that constitutes the object of theory and research. While laypersons use 'emotion' interchangeably with terms such as affect, feeling, sentiment, or mood, psychologists define the construct as a process of changes in different components rather than a homogeneous state. Furthermore, it is assumed that the differentiation of the emotions (e.g., into fear, anger, or joy) is based on specific configurations of changes in the components.

1. Components

Starting with Greek philosophy, there has been an uninterrupted tradition of postulating at least three major reaction components of emotion: physiological arousal, motor expression, and subjective feeling. These three components are often referred to as the 'emotional response triad.'

1.1 Physiological Arousal

The widespread notion that emotions upset the body's equilibrium state suggests that the presence of arousal or activation is one of the major preconditions for using the label 'emotion.' References to physiological changes such as temperature sensations, respiratory and cardiovascular accelerations and decelerations, trembling and muscle spasms, as well as feelings of constriction in internal organs are frequently part of emotion descriptions. The important role of neurophysiological changes in emotional episodes is due to the emotion-eliciting event disturbing ongoing homeostatic regulation and smooth behavioral coordination, and the preparation of appropriate adaptive responses (e.g., producing the necessary energy for appropriate actions such as fight or flight).

1.2 Motor Expression

The changes in facial and vocal expression, as well as in gestures and posture, during emotion episodes are also considered to be central components of the

emotional response. According to Darwin ([1872] 1998) these expressions are rudiments of formerly adaptive behaviors (e.g., clenching one's teeth as a rudiment of a biting response). Whatever their origin, most scholars have explained emotional expressions in terms of their communicative functions, broadcasting the reaction of an individual and its corresponding behavior intentions to other members of the group.

1.3 Subjective Feeling

The fact that individuals can verbally report a multitude of qualitatively different feelings, as encoded in a rich emotion vocabulary, constitutes the most important facet of emotional episodes for many scholars in the field of emotion. These internal sensations, often considered as necessarily conscious experiences, are often called 'qualia,' irreducible qualities of feeling that are unique to the specific emotional experience of a particular individual. More recently, it has been suggested that the feeling component of emotion can be conceptualized as a reflection of all changes in components during an emotion episode, that is, the results of event appraisal, motivational change, and proprioceptive feedback from motor expression and physiological reactions. It is important to differentially define the concepts in this fashion as the tendency to use emotion (the process as a whole) and feeling (one of its components) as synonyms results in confusion (Scherer 2000a).

1.4 Behavior Preparation

It has been argued that the classic response triad needs to be complemented by another component—behavior preparation or action tendencies (Frijda 1986, Scherer 1984). While both the physiological and expressive components imply the preparation of adaptive action in a general, often unspecific way, the notion of specific action tendencies being part of the response syndrome suggests a more explicitly motivational function. Emotions change ongoing goal-directed behavior and produce action tendencies that are specifically adapted to dealing with the environmental contingency that has elicited the emotional response.

1.5 Cognitive Processes

In the wake of the cognitive revolution in the social and behavioral sciences, there has been increasing recognition that emotions may have very strong effects on perceptual and cognitive processes such as attention, thinking, memory, problem solving, judgment, decision making, and the like (see contributions in Dalgleish and Power 1999). Thus, the emotional response syndrome also includes a cognitive component. It can be further argued that the emotion construct should not be restricted to the response components but should also include the specific characteristics of the processes that elicit and differentiate the emotions. Thus, proponents of appraisal theories of emotion suggest that emotional reactions are determined by the subjective evaluation of events with respect to their significance for the well-being and goal attainment of individuals (see contributions in Scherer et al. 2001). The recursiveness of the appraisal–response sequence (e.g., the evaluation of an event as dangerous may produce fear which in turn may affect the ensuing evaluation of subsequent events) strengthens the suggestion to include cognitive appraisal processes as a component of the emotion construct.

While the notion that emotions are syndromes of component changes is widely accepted, there is debate as to the relative importance of these components, particularly as to which ones should be considered as necessary and sufficient conditions for the occurrence of emotion. The components are expected to interact, that is, to influence each other. Thus, one issue of research concerns the possibility that motor expression may produce specific feelings or at least increase or decrease the intensity of subjective experience (as postulated by proprioceptive feedback theories; see Scherer 2000a for further detail).

2. Process

While it is customary to talk of 'emotional states,' emotion researchers generally emphasize that emotion is a process in time. Generally, research has shown that different emotions have different time patterns including differential duration and varying onset or offset patterns. What is seen as part of an emotion episode partly depends on the specific definitions proposed by a researcher. In what follows, a number of sequential phases or stages of typical emotion episodes are described. Emotions can be evoked by a number of different factors, including external events or behaviors, internal neurophysiological changes, or memory recall. All of these potential origins will be subsumed under the term 'stimulus event' (triggering an emotion episode). Since external events seem to be the most frequent elicitors, the description below is biased toward such types of stimulus events (e.g., a sudden noise that might represent an explosion in the vicinity).

Salient stimulus events are immediately evaluated by an automatic evaluation process on lower, subcortical levels of the central nervous system (CNS), focusing attention on the stimulus. This preliminary evaluation may result in orientation and/or defense

responses as well as in feelings of pain or pleasure. Stimulus events thus marked for requiring further, more controlled or effortful processing are subject to being appraised by evaluation processes on higher levels of the CNS, in particular the cortical association areas and memory (Dalgleish and Power 1999, Leventhal and Scherer 1987). In this phase, the causation of the event (i.e., the objects and individuals likely to be responsible), the process involved, the consequences to be expected, the implications for the well-being and the immediate goals of the individual, and the ability to modify the consequences or to adjust to them, are evaluated (a process described in detail by appraisal theories of emotion; see contributions in Scherer et al. 2001).

The appraisal process generally produces motivation towards an adaptive response to the event. Depending on the importance of the consequences to be expected, this urge may generate a primordial goal for the specific moment in time (e.g., survival or self-assertion), superseding all prior needs, desires, or goals. Once the consequences of the event and the individual's response potential are evaluated in at least a preliminary fashion, the current goal/plan hierarchy needs to be restructured. This stage can be called 'goal/need priority setting.' Next, action alternatives are examined with respect to the cost involved and the likelihood of success. The result of this selection produces more specific action tendencies or behavior preparation (Frijda 1986), enabling the organism to rapidly engage in a certain class of behaviors (e.g., mobilizing the necessary energy for fight or flight). In this phase, the internal monitoring of the event appraisal and the physiological and expressive reactions (as reflected in subjective feeling, see above) play a central role in determining the need for emotion control or regulation, inhibiting execution of the activated action tendencies. The next phase consists of behavior execution, that is, the triggering of motor commands that produce concrete overt behavior, in terms of instrumental motor behavior and verbal or nonverbal communicative acts. In many cases the individual will, during or after the emotion episode, probably talk about the emotional experiences with others (emotional communication).

So far, the description has focused on the part of the emotion process consisting of an individual's reactions to an event. In many cases this reaction is embedded in a social context (particularly when an interaction partner has provoked the reaction) and is only the beginning of an interactive process in which each reaction of one of the protagonists is a stimulus event that is likely to modify the emotion process. Examples are the escalation of anger or the soothing of sadness through the empathic response of another. This process of interpersonal emotion regulation or escalation has been well described in the context of marital interactions (Gottman 1993). In social interaction emotions, and particularly emotional expression, are often 'managed' in the pursuit of strategic goals such as self-presentation, intimidation, adulation, and so forth (Goffman 1959, Manstead 1991).

3. *Differentiation*

So far, the structure of emotion has been defined in terms of the underlying theoretical construct, in particular with respect to components and process. However, the problem of structure also arises when the differentiation of the emotions is addressed: How many emotions are there and what are they? This question is answered very differently depending on the theoretical stance adopted (see contributions in Ekman and Davidson 1994). Below, a number of major theoretical positions and their views on this issue are briefly described (see Scherer 2000b for greater detail).

3.1 *Dimensional Models*

The first explicit dimensional model was formulated by Wundt (1905) who used both introspective and experimental methods to study subjective feeling experiences. He proposed that feelings varied by their positions on three independent dimensions: pleasantness–unpleasantness, rest–activation, and relaxation–attention. This model strongly influenced early emotion psychology. It has an intuitive appeal since it has been shown that humans have a tendency to perceive all kinds of meaning in terms of valence (i.e., positive vs. negative) and activation (active–passive; see Osgood et al. 1975). Following this tradition, many dimensional theorists argue that the differentiation is sufficiently well described by one or both of these central dimensions. Dimensional models have been particularly successful in social psychological and neurophysiological approaches to emotion (see Scherer 2000b).

3.2 *Discrete Emotion Models*

Theorists in this tradition postulate a limited set of basic emotions such as fear, anger, joy, sadness, and disgust, characterized by a number of fundamental adaptive responses to prototypical events (e.g., fear/flight in response to an attack by a powerful enemy, disgust/avoidance in the case of spoiled food). Thus, the structure of emotion differentiation is given by prototypical, phylogenetically evolved adaptation patterns. Some models suggest that the differentiation and the number of fundamental emotions is determined by evolved neural circuits (circuit models; e.g., Panksepp 1998) or by phylogenetically continuous classes of motivation such as aggression or nurturance (motivational models, e.g., Plutchik 1980). Based on

Darwin ([1872] 1998), Tomkins (1984) suggested the existence of a small number of discrete, basic emotions in the form of neuromotor programs. Tomkins' suggestion has been extended by Ekman and Izard, who worked towards obtaining empirical evidence for theoretically derived consequences of discrete emotion theory, such as early ontogenetic onset of the discrete emotion patterns, discrete patterning of prototypical facial expressions, and universality of these patterns (Ekman 1992, Izard 1991). Theorists in this tradition have suggested that these basic emotions can be mixed or blended, thus explaining the large variety of emotion-descriptive verbal labels in many languages.

3.3 Meaning-oriented Models

Frequently, the semantic fields of terms in the emotion vocabularies of different languages are used to describe the structure of emotion, assuming that there is a mapping between semantic and psychological structures and categories. For example, lexical models try to define basic emotions as prototypical linguistic categories (e.g., a prototypical 'anger' category covering terms such as irritation, rage, being cross, etc.; e.g., Shaver et al. 1987) and to define each category by listing, as is the case in a dictionary, the semantic prerequisites and implications for the use of the word (e.g., in the case of indignation, 'being denied what one thinks one has a claim on'). Close to this view are social constructivist models claiming that the emotions are constituted or constructed by socioculturally determined behavior and value patterns, varying greatly over cultures. For example, it is suggested that certain emotions exist only in specific cultures or that a particular emotion has disappeared because of historic changes in cultural values and lifestyles. These models all suggest that the structure of emotion differentiation cannot be reduced to basic psychobiological patterns (i.e., a limited number of adaptive behaviors) but needs to be modeled on the basis of sociocultural meaning, often reflected in the semantic structure of the emotion vocabulary (see Scherer 2000b for more detail).

3.4 Adaptational Models

Theorists focusing on the way in which evolution may have equipped organisms to respond automatically to significant events have developed what can be called 'biological preparedness' models. For example, Öhman (1988) has developed a model that suggests that humans have unconscious detection mechanisms for stimuli that are potentially harmful (such as snakes and spiders). The automatic detection of such threat stimuli serves to efficiently prepare emergency reactions on the level of the autonomic nervous system (ANS). LeDoux's (1996) work shows that the neural architecture of the fear system (at least in rats) is optimally suited for the production of such rapidly and automatically elicited but relatively nondifferentiated responses.

3.5 Componential Models

Componential emotion theorists postulate that emotions are elicited by a cognitive (but not necessarily conscious or controlled) appraisal of antecedent situations and events and that the patterning of the reactions in the different response domains (physiology, expression, action tendencies, feeling) is produced by the outcome of this evaluation process. According to this view, the structure of emotion is the result of the dimensions or criteria used in the appraisal process to determine the meaning of the stimulus event to the individual and of the response patterns (facial and vocal expression, physiological symptoms) generated by the appraisal results (see the chapters authored or co-authored by Ellsworth, Frijda, Lazarus, Roseman, Scherer, and Smith in Scherer et al. 2001). Lazarus (1991) sees emotion structure as determined by appraisals on a small number of dimensions called 'relational themes' (e.g., 'loss of a cherished person or thing' giving rise to sadness). In contrast, the component process model proposed by Scherer (1984; see also chapter in Scherer et al. 2001) suggests that one can distinguish as many different emotional processes as there are differential patterns of appraisal results. However, it is suggested that there are modal emotion outcomes corresponding to universal organism–environment interactions (e.g., 'blocked goals') which tend to be labeled by short emotion terms (e.g., anger) in most languages. The other appraisal theorists adopt intermediate positions.

This review shows that the current psychological models of emotion vary quite strongly in their postulates regarding the structure of emotion with respect to focus on specific components, presumptions about the nature of the process, and the explanations offered for the differentiation of individual emotions. A two-dimensional coordinate system can be used to plot these structural differences among psychological theories of emotion. One dimension consists of the different components of the emotion syndrome, as described above: cognitive processes, peripheral physiological responses, motivational changes, motor expression, and subjective feeling. The second dimension consists of the different stages of the emotion process, as outlined above. Table 1 shows where in this space the models reviewed above can be located.

Table 1 illustrates the origin of some of the differences in assumptions about the psychological structure of emotion: The more a model is concerned with the output end of the emotion process (e.g., the selection of adaptive responses), the more it sees structure as being determined by the type of action tendency (e.g., a flight tendency characterizing the fear response). In contrast, the more a model focuses on

Table 1
Comparison of major emotion theories with respect to their focus on components and phases in the emotion process

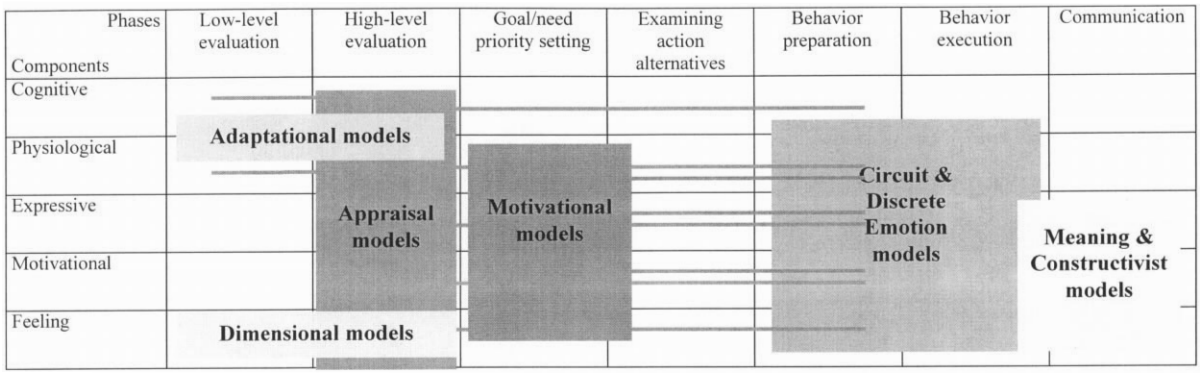

The solid area, representing a particular type of model, is drawn across the intersections of components and phases that theorists in the respective tradition focus on most. The horizontal lines projecting from the solid areas indicate that a particular model includes in its theorizing at least some of the components/phases crossed by these lines.

the appraisal phases, that is, the input end of the emotion process, the less constrained is the presumed structure of emotion differentiation, allowing for a multitude of potential outcomes of the appraisal process. Obviously, it also matters which component of emotion a specific model focuses on. Models oriented toward the output end of the process tend to focus on the motivational (action tendencies) and motor expression components whereas those focusing on the input side highlight the cognitive component as being involved in structural differentiation. The table also allows one to determine which theories are committed to a psychobiological view (the models focusing on the physiological and expressive components, i.e., adaptational, motivational, circuit, and discrete emotion models) in contrast to those that lean more towards the phenomenological or sociocultural aspects of emotion (the models focusing on the subjective feeling component, i.e., dimensional, meaning, and constructivist models). Appraisal models tend to span the complete range of components. In spite of these differences, the claims of the different models are quite compatible once one allows for the fact that they focus on different components of emotion and different phases of the process.

4. Outlook

The explosive growth of emotion research in the last quarter of the twentieth century has produced both conceptual and methodological tools that promise to revolutionize the study of emotion. Largely confined to philosophical speculation based on verbal reports of subjective experience for much of its history, and hampered by terminological confusion, the study of emotion is increasingly based on a consensus concerning fundamental conceptual distinctions and employs methods that complement the traditional reliance on verbal report. In particular, the objective assessment of facial and vocal expression as well as of neurophysiological reactions allows a multicomponent approach to measurement that is likely to do greater justice to the complex structure of emotion episodes. Another promising development is the rapprochement between psychological and neuroscience approaches to emotion, with insights on the neural architecture of the emotion system, obtained through brain imagery, providing information on process characteristics. Our understanding of emotion processes has also been advanced by copious work on the development of emotion (see *Infancy and Childhood: Emotional Development*) and on affective disturbance.

Progress has been somewhat less marked with respect to the social functions of emotion: While there is much research on the influence of mood on social cognition, there have been only very few efforts to study empirically the role of emotion in social interaction. Phenomena like emotional contagion, affective influence, anger escalation, or emotional climate in groups constitute promising areas for future research on the social context of emotion processes.

See also: Culture and Emotion; Emotion and Expression; Emotion: History of the Concept; Emotion, Neural Basis of; Emotions, Evolution of; Emotions, History of; Self-conscious Emotions, Psychology of

Bibliography

Dalgleish T, Power M 1999 *Handbook of Cognition and Emotion*. Wiley, Chichester, UK

Darwin C [1872] 1998 *The Expression of Emotions in Man and Animals*, 3rd edn. HarperCollins, London

Ekman P 1992 An argument for basic emotions. *Cognition and Emotion* **6**: 169–200

Ekman P, Davidson R J 1994 *The Nature of Emotion Fundamental Questions*. Oxford University Press, New York

Frijda N H 1986 *The Emotions*. Cambridge University Press, Cambridge, UK

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY
Gottman J 1993 Studying emotion in social interaction. In: Lewis M, Haviland J M (eds.) *Handbook of Emotions*. Guilford Press, New York, pp. 475–88
Izard C E 1991 *The Psychology of Emotions*. Plenum Press, New York
Lazarus R S 1991 *Emotion and Adaptation*. Oxford University Press, New York
LeDoux J 1996 *The Emotional Brain The Mysterious Underpinnings of Emotional Life*. Simon and Schuster, New York
Leventhal H, Scherer K R 1987 The relationship of emotion to cognition. A functional approach to a semantic controversy. *Cognition and Emotion* **1**: 3–28
Manstead A S R 1991 Emotion in social life. *Cognition and Emotion* **5**: 353–62
Öhman A 1988 Preattentive processes in the generation of emotions. In: Hamilton V, Bower G H, Frijda N H (eds.) *Cognitive Perspectives on Emotion and Motivation*. Kluwer, Dordrecht, The Netherlands, pp. 127–44
Osgood C E, May W H, Miron M S 1975 *Cross-cultural Universals of Affective Meaning*. University of Illinois Press, Urbana, IL
Panksepp J 1998 *Affective Neuroscience The Foundations of Human and Animal Emotions*. Oxford University Press, New York
Plutchik R 1980 *Emotion: A Psychobioevolutionary Synthesis*. Harper and Row, New York
Scherer K R 1984 On the nature and function of emotion A component process approach. In: Scherer K R, Ekman P (eds.) *Approaches to Emotion*. Erlbaum, Hillsdale, NJ, pp. 293–317
Scherer K R 2000a Emotion. In: Hewstone M, Stroebe W (eds.) *Introduction to Social Psychology. A European Perspective*, 3rd edn. Blackwell, Oxford, UK, pp. 151–91
Scherer K R 2000b Psychological models of emotion. In: Borod J (ed.) *The Neuropsychology of Emotion*. Oxford University Press, New York, pp. 137–62
Scherer K R, Schorr A, Johnstone I T 2001 *Appraisal Processes in Emotion: Theory, Research, Application*. Oxford University Press, New York
Shaver P, Schwartz J, Kirson D, O'Connor C 1987 Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology* **52**: 1061–86
Tomkins S S 1984 Affect theory. In: Scherer K R, Ekman P (eds.) *Approaches to Emotion*. Erlbaum, Hillsdale, NJ, pp. 163–95
Wundt W 1905 *Grundzüge der physiologischen Psychologie*, 5th edn. [*Fundamentals of physiological psychology*]. Engelmann, Leipzig, Germany

K. R. Scherer

Emotions, Sociology of

1. Introduction

In this article, a brief social history of the Western bias against emotions will clear the air for their current study. Works defining the beginnings of the 'sociology of emotions' are described; substantive topics addressed by the field are organized under a cybernetic macro–micro framework that emphasizes its contribution to sociology in general.

The sociology of emotions is premised on several recent observations that contradict popular prejudices. First, some constant emotional posture is necessary for mental functioning. Second, rather than depicting the asocial, idiosyncratic part of human existence that escapes through the seams of social organization, emotion is socially patterned and is a vital part of the maintenance of social organization. Third, rather than shaping the mere cultural display of biologically fixed emotions, in most cases, social processes are involved in their very constitution.

Definitions of emotion typically include organized mental states comprising cognitive appraisals, impulses to action, and to varying degrees, feelings. Hochschild (1983) likens emotions to senses that signal what is personally relevant about surrounding events. Scheff has cautioned that the attempt to define emotion is a rock on which many a ship has split. One difficulty is that emotion labels are often ideological. Contemporary American salaries for corporate executive officers averaging 400 times more than the lowest organizational worker can be viewed in terms of 'purely objective rewards for effectiveness,' self-interest, or greed.

2. The Recent Shift in the Western Evaluation of Emotion

Until the end of the twentieth century, Western society paid scant attention to emotions except to warn against them as a source of bias. The last quarter of the century however, saw a significant reversal from the devaluation of emotion to recognition of its crucial importance. Emotion, once seen as the enemy of rationality is now viewed as the foundation for our engagement in the world. Studies of brain trauma now identify certain emotional processes as necessary for rational choice-making and social adaptation. Artificial intelligence research reveals the importance of emotional predispositions in selecting what is relevant to a particular line of thought out of an unmanageable array of objective possibilities. Philosophers contend that 'emotion is what one sees the world in terms of.' Thus, emotion sets the very agenda for thought.

3. The Discursive Origins of the Word Emotion

The practice of carving affective experience into many separate subtypes and grouping them together as the same kind of psychic phenomena called 'emotion,' is relatively recent, and suspect to many scholars. The word did not appear in the Oxford Dictionary until 1580, after terms like 'self' and 'consciousness' lost

their negative connotation and individual experience became a subject worth attention. Before that, rage, love, and pity, etc. were not seen in terms of the broader category, 'emotions.' As the ascribed identities in the dark ages gave way to identities based more on achieved status, subjective experience became more relevant. Personal identity depended more on interpersonal skills and self-reflective behavioral control. People became more aware of, and interested in, their emotions.

Since antiquity, the more narrow term, 'passion,' stemming from 'pathos' and associated with suffering, referred to psychic events that overcame individuals. The broader term 'emotion' in France came from the Latin 'movement' or 'migration,' and meant that which 'moved' the person psychically. While this movement could now be for good or bad, the term still carried the connotation of persons torn between their hearts and minds (Sarbin 1986). This placed cognition as the exclusive link to reasoning with emotion as its threat. Thus, the implication of the term 'passion' still lurks in the broader word 'emotions.' Positive emotions such as awe, appreciation, gratitude, and determination are typically glossed over. Baily (1983) suggests that this is because emotion is used strategically as a discursive practice to account for, and excuse, untoward action. Nonetheless, sociological analysis depends on giving attention to a broad range of both dangerous and life-affirming emotions. All humans are emotional in the broad sense. Bureaucratic unease with emotion, is itself, emotion.

4. *Scientific Devaluations of Emotion in the Nineteenth Century and Lessons Learned*

William James (1902) and Charles Darwin (1955) set the scientific stage for the twentieth century depreciation of emotion. Darwin saw emotion as a vestigial remnant of the primordial bodily preparation for action. Its evolutionary survival lay only in its expression as cues to others; emotion had no current value to the individual. To James, emotion was primarily internal; it was essentially the body responding to the body, and not important in relating persons to the world. What lasted from James was the debated idea that an internal, defining substrate exists for major emotions.

Ironically the very 'subjectivization' of postmodern mentality that made the broad classification 'emotion' make sense and seem important has also led to distortions insofar as they are seen as self-contained, totally internal processes, independent of ongoing human action and particular social contexts. Emotions do not exist separated from action in an inner depository of the self, like cookies in a jar waiting to be discovered by inner reflection (Baumeister 1986).

With all the diverse emotional lexicons in the world, it is an illusion to insist that there is a distinctive process or state existing somewhere in the person that coincides with each emotion word. However emotion is viewed, its relation to cognition must be handled carefully. To the extent that the two can be separated, they are nonetheless highly interdependent; changes in one system produce changes in the other. Though they can be experienced in serious tension with each other, recent findings, alluded to above, have demonstrated the necessity of certain emotions for adaptive behaviors. Emotion and cognition are now seen, not in terms of their oppositional contrasts, but as parts of a single, though multivalent, system in human mentality.

5. *Origins of the Sociology of Emotions: 1975–80*

Particular emotions were, indeed, important to the classical sociologists (Marx and alienation; Weber and disenchantment, etc.). Their treatment, however, was not concerned with the light they shed on the broader social-emotional process in general (Smith-Lovin 1995). It was not until 1975 that American sociologists, for example, called for a study of emotion as a social process in its own right. In that year Arlie Hochschild (1975) argued that a purely cognitive view of human life was not sufficient for sentient humans. Her insistence that emotion was socially patterned was partly explained by her concept of feeling rules, i.e. culturally and historically variable norms defining what should be felt in various situations. Often these rules served ideological functions. While Foucault was stressing hegenomic power in forms of knowledge, Hochschild was stressing the hegenomic control of the heart. 'Emotion work' described the individual processes that maintained these norms. This occurs when actors feel a pinch between what they actually feel and what they feel they should feel. Techniques used to change emotions include surface acting to mask or enhance the emotion and deep acting wherein the actual emotion is changed. Recent studies also focus on how other persons encourage or cut off the actor's expression of emotion and in some cases, feeling itself.

Scheff's (1979) model of the emotional person allowed for the possibility that culture could violate inherent emotional needs. He stressed how culture provided, or failed to provide, the cultural rituals needed to purge the physiological build-up of distressful emotions. Such rituals evoked the distressful experience so it could be relived emotionally but simultaneously offer a new cognitively distanced perspective resulting in cathartic, 'aesthetic balance.'

In the following decade (Scheff 1990) refined his proposal that shame and pride were the most important human emotions, forming indicants of the

individual's social bond. This bond was the *sine qua non* for human adaptation. The current overemphasis on impersonal forms of social organization puts this bond in jeopardy and increases incidents of shame. Through a series of escalating shame–anger spirals, rage and violence emerges. Most conducive to violence is unacknowledged shame, that is, shame not identified as such by the actor. Observable shame markers are identified that manifest themselves independently of the actor's knowledge and allow for testable hypotheses.

Kemper's (1978) Power and Status Theory argued that 'social power,' the ability to force others' compliance against their will, and 'social status,' conceived as favors freely given, were critical dimensions determining emotional states. A person's 'structural' place in hierarchies of excesses or deficits along these dimensions automatically produced certain distressful emotions depending on whether one attributed the cause to one's self or others. At a minimum his deterministic scheme drew attention to the fact that a great number of emotions had to do with one's place on the social map.

Shott (1979) presented a theory that stressed the importance of emotions in voluntaristic social control. Prior research had stressed the importance of role-taking as a process that fused self-control and social control. Role-taking is the capacity to respond to one's own on-coming behavior as would others, and to use their anticipated response as a guide to further conduct. It had been understood as essentially cognitive. But cognition was not sufficient to ensure normative social control. Role-taking could as soon be used to manipulate others without the self-reflexive emotions of guilt, shame and embarrassment or the other-directed empathetic emotions of sympathy. The personal discomfort these emotions produced gave actors a personal stake in the avoidance of behaviors producing them. In these cases harming others is harming one's self.

In her focus on chronically failed emotion-work, Thoits (1990) coined the term 'emotional deviance.' The label of mental illness is often the result of reactions by others or one's self to the inability to feel appropriately. Sources of discrepant feelings are multiple roles, role transitions and unrealistic demands. Thoits expanded Hochschild's treatment of emotion-work strategies identifying points at which they could easily fail. She also drew attention to the emotional dimensions of coping behaviors and social support, the latter seen as emotion work on others. Support may not alleviate pain but it can help victims cope with it.

Assuming the close relationship between emotion and cognition, fundamentally diverse life forms and thus fundamentally diverse ways of thinking will produce important emotional differences. This is the basis for Social Constructionism, articulated by Rom Harré (1986). Like Baily, he sees the use of the term 'emotion' as a social strategy—a discursive practice used for identity maintenance. He does not deny bodily concomitants, or the perceived reality of the experience, but to him the physiology involved is not a 'natural sign' of a biological inner state. Instead, emotional experience is a function of diverse local languages and moral orders.

Denzin (1984) argued against the study of emotions as an abstract class, separated from the lived experience of situated individuals. Emotion was to be studied by narrative and as reflected in cultural reflections such as film.

6. *The Cybernetic Program for the Sociology of Emotions*

In 1985, the sociology of emotions was formally recognized as a subsection in the American Sociological Association. Several scholars independently suggested a cybernetic program for emotion studies. This balanced the cultural and structural shaping of emotions with equal attention to how individual emotional experience worked inadvertently to maintain, confirm and change the social arrangements that initially shaped them. This view is clearly illustrated by incest taboos shaping feelings of sexual attraction and disgust. Aggregate individual experiences of these sexual aversions confirm and maintain the kinship arrangements and feeling rules that shaped them in the first place. Without these feelings, the structures of kinship would collapse. The advantage of such an approach is that it allows for a detailing of the individual interactional processes that form the foundation of emergent structures. Without this, our theory of social structure becomes dangerously reified (Scheff 1990). Regardless of the emergent quality of structures, the detailed link between concrete people and institutions must be traced. Collins' (1984) conception of emotional energy emanating from 'interactional ritual chains' argues for a similar articulation between micro- and macro-levels.

7. *Cultural and The Social Structural Shaping of Emotion*

Emotions presented as culturally distinct include the Japanese amae, an accepted adult feeling of dependence and a desire to be cuddled that would make most Westerners uncomfortable. Legit is an emotion of the Philippine Llongots that combines achievement of group membership with force and creation. They feel legit when they 'take heads.' There would be little debate that certain cultures encourage and discourage certain emotional experiences and that these lead to important behavioral dispositions. Through history one can trace the emotional tenors of particular epochs

like the anxious prudishness of the Victorian bourgeoisie and the shift to the ambivalence of the more narcissistic discontent of today. Shifts in the incidences of swooning imply the degree to which culture socializes the body. Other examples of the cultural shaping of emotion include the literature on emotion and gender. Important here is the devaluation of emotion and the belief that women are inherently more emotional than men. There is no reason to think that men are not as in touch with their own perceived manly feelings as are women. Hochschild (1983) brought attention to 'status shields,' i.e. shared, interpretive dispositions that protect those of privilege against (a) the displaced feelings of others, (b) the negative interpretations of one's own emotional expressions, and (c) the conviction of others that one's emotions are unwarranted and not important. The undermining of women's anger by interpreting them as 'crabby' because of their menstrual period or otherwise 'hysterical,' is an obvious example of a structured lack of status shields. The aggregate of such interpretations serve male dominance.

Emotions are shaped by the ways whole societies are organized, as well as the organizations within them. As an example of the first type of shaping, Hochschild (1989) demonstrated how macro-level changes in women's entry into the workforce effected micro-level gender imbalances in feelings of gratitude between couples in the household. Because of traditional gender roles, men's relatively small contributions to housework evoked more gratitude than the larger contributions from their working wives. This resulted in a 'second shift' for women and interpersonal tension. The study of gendered inequalities in how much it takes to evoke gratitude has important implications for power relations in general.

In spite of the devaluation of emotion in many businesses and bureaucracies, they organize emotional processes unwittingly or consciously. Feelings can impede organizational efficiency but they are also necessary for it. Any workplace must deal with loyalty and disenchantment, stirrings about status and power, anxieties about keeping control and being controlled, pride, defensiveness, fear of change, and the comforts of the familiar. Emotional labor, according to Hochschild (1983), is the commercial exploitation of the principle of emotion work. The airline stewards she studied were paid to 'be nice' regardless of insults and paradoxically asked to be genuine. Loyalty can supercede error; receptionists and credit collectors are paid to make persons feel certain ways. Many studies deal with divisions of emotional labor in organizations and the simultaneous relations between formal and informal emotion rules. Hospitals must deal with protecting their staff from the anguish of patients and the different emotional roles played by practitioners and nurses. The emotion work necessary from female public defenders of rapists has also proved of interest to sociologists.

8. *Social Interaction and Emotions Supporting Structures*

Individual emotional processes acting as concrete support for structural arrangements most obviously include emotion work and the reflexive role-taking emotions. Below, other such processes are identified.

What persons care about determines emotion. Persons care about their identities. Identities are connected with social structure because they are embedded in social roles. Emotions will determine what roles are chosen as central to identities and resulting role performance will be affected by emotions (e.g., confidence) as well as producing them (e.g., pride). The emotional aspects of identity work are, then, simultaneously commitments to structural role performances. The work by Heiss and Smith-Lovin on affect control as well as that by Stryker and Serpe on identity theory describes this 'maintenance' process (Smith-Lovin 1995).

A crucial process in maintaining the Western belief that each emotion is an underlying physiological state is the 'documentary method.' This process underwrites the reality of socially constructed 'facts' by a circular logic. An unobservable belief (like the fixed existence of a certain emotion in the person) is presumed, and overt events are selectively used as evidence of its existence or explained away if they refute it. This process is especially evident in the interpretation of gendered behavior.

See also: Alienation, Sociology of; Charisma: Social Aspects of; Culture and Emotion; Emotion: History of the Concept; Emotions, History of; Goffman, Erving (1921–82); Marx, Karl (1818–89); Personality Structure; Psychoanalysis, History of; Psychoanalysis in Sociology; Psychological Anthropology; Shame and the Social Bond; Socialization, Sociology of; Symbolic Interaction: Methodology; Weber, Max (1864–1920)

Bibliography

Baily F G 1983 *The Tactical Uses of Passion*. Cornell University, Ithaca, NY

Baumeister R F 1986 *Identity*. Oxford University Press, New York

Clark C 1997 *Misery and Company: Sympathy in Everyday Life*. University of Chicago Press, Chicago

Collins R 1984 The role of emotion in social structure. In: Scherer K R, Ekman P (eds.) *Approaches to Emotion*. Lawerence Erlbaum Associates, Hillsdale, NJ

Darwin C R 1955 *The Expression of the Emotions in Man and Animals*. Philosophical Library, London

Denzin N K 1984 *On Understanding Emotion*. Jossey-Bass, San Francisco

Elias N 1978 *The History of Manners: The Civilizing Process: Volume I*. Pantheon Books, New York

Ellis C, Flaherty M G 1992 *Investigating Subjectivity: Research on Lived Experience*. Sage, Newbury Park, CA

Ellis C, Flaherty M (eds.) 1994/95 *Social Perspectives on Emotion*. JAI Press, Greenwich, CT, Vol. III

Erikson R, Cuthbertson-Johnson B (eds.) 1998 *Social Perspectives on Emotion*. JAI Press, Greenwich, CT, Vol. IV
Fineman S 1993 *Emotions in Organizations*. Sage, London
Franks D D, Gecas V (eds.) 1992 *Social Perspectives on Emotion*. JAI Press, Greenwich, CT, Vol. I
Franks D D, Smith T S (eds.) 1999 *Social Perspectives on Emotion*. JAI Press, Greenwich, CT, Vol. V
Gordon S L 1981 The sociology of sentiments and emotion. In: Rosenberg M, Turner R H (eds.) *Social Psychology: Sociological Perspectives*. Basic Books, New York
Griffiths P 1997 *What Emotions Really Are: The Problem of Psychological Categories*. University of Chicago Press, Chicago, Chap. 6
Harré R 1986 *The Social Construction of Emotion*. Blackwell, New York
Hochschild A R 1975 The sociology of feeling and emotions: selected possibilities. In: Millman M, Kanter R M (eds.) *Another Voice*. Anchor/Doubleday Press, Garden City, NY
Hochschild A R 1983 *The Managed Heart: Commercialization of Human Feeling*. University of California Press, Berkeley, CA
Hochschild A R 1989 The economy of gratitude. In: Franks D D, McCarthy E D (eds.) *The Sociology of Emotion: Original Essays and Research Papers*. JAI Press, Greenwich, CT
James W 1902 *The Principles of Psychology*. Henry Holt and Co., New York, Vol. II
Kemper T D 1978 *A Social Interactional Theory of Emotions*. John Wiley, New York
Lepenies W 1992 *Melancholy and Society*. Harvard University Press, London
Lutz C 1988 *Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Challenge to Western Theory*. University of Chicago Press, Chicago, IL
Sarbin T R 1986 Emotion and act. In: Harré R (ed.) *The Social Constuction of Emotion*. Blackwell, New York
Scheff T J 1979 *Catharsis in Healing, Ritual and Drama*. University of California Press, Berkeley, CA
Scheff T J 1990 *Microsociology: Discourse, Emotion, and Social Structure*. University of Chicago Press, Chicago, IL
Shott S 1979 Emotion and social life: a symbolic interactionist analysis. *American Journal of Sociology* **84**: 1317–34
Smith-Lovin L 1995 The sociology of affect and emotion. In: Cook K S, Fine G A, House J S (eds.) *Sociological Perspectives on Social Psychology*. Allyn and Bacon, Boston
Thoits P A 1989 The sociology of emotion. *Annual Review of Sociology*. **15**: 317–42
Thoits P A 1990 Emotional deviance: research agendas. In: Kemper T D (ed.) *Research Agendas in the Sociology of Emotion*. State University of New York Press, Albany, NY
Weigert A J 1991 *Mixed Emotions: Certain Steps Toward Understanding Ambivalence*. State University of New York Press, Albany, NY
Wentworth W M, Ryan J (eds.) 1994 *Social Perspectives on Emotion*. JAI Press, Greenwich, CT, Vol. II

D. D. Franks

Empiricism, History of

There are almost as many empiricisms as there are empiricists, but what these views or approaches have in common is an emphasis on the importance of experience to the formation of concepts and to the acquisition of knowledge. The foil to empiricism is rationalism, which emphasizes instead the importance to thought and knowledge of material that is in some sense independent of experience. The range of empiricist positions is vast, from the shocking view that all we can think or know about are our sensations to the mundane claims that experience plays some role in the formation of some of our concepts and in the justification of some of our beliefs. Empiricism of some form may seem both obviously correct and obviously the correct philosophy for science on the grounds that it is clear that one can only find out about the world by observing it; but this innocuous looking thought has been disputed, and in any case many forms of empiricism go far beyond it in their claims. It is also unclear whether empiricism ultimately supports or undermines claims to scientific knowledge.

1. History

Empiricism is a hardy perennial in the history of philosophy. It was extensively discussed in antiquity, often with considerable hostility. Plato disdained it, partly on the grounds that observation can only provide information about a realm of appearances less important and indeed less real than a more abstract realm that we may be able to grasp through a form of thinking that does not depend on experience. Mathematics was a model of knowledge for many ancient thinkers, and this is perhaps the area where empiricism is least plausible, since mathematical proofs do not seem to depend on observation. Nevertheless, empiricism did have defenders in the ancient world, with Epicurus often cited as its first clear exponent. He held not only that observation is the only source of knowledge about the world but that, insofar as they are taken as sources of information about the appearances of things, the senses never lie.

Empiricism also had its advocates in the medieval period. St. Thomas Aquinas wrote that there is nothing in the intellect which was not previously in the senses, though by this he meant to endorse only concept empiricism and not a strong form of knowledge empiricism. Concepts can only be formed with the help of sensory images, but some knowledge—for example, some knowledge about God—is available by means of logical argument independent of particular observations. Empiricism is perhaps most strongly associated with three philosophers of the seventeenth and eighteenth centuries: John Locke, George Berkeley, and David Hume. These British empiricists developed versions of both concept and knowledge empiricism. In some respects Berkeley (1710, 1713) can be seen as attempting to purify Locke's (1700) view of its nonempiricist elements and Hume (1739, 1777) as attempting the same for Berkeley. Both Locke and Hume appealed to Newton's achievements in physics for inspiration: Locke presented himself as

Newton's underlaborer, 'clearing ground a little, and removing some of the rubbish, that lies in the way to knowledge.' Part of that ground-clearing operation consisted of a vigorous rejection of the claim that we are born with innate knowledge, a characteristic point of dispute between empiricists and rationalists (cf Mackie 1976). Hume characterized his own project more ambitiously, as an attempt to do for the human sciences ('moral philosophy') what Newton had done for the physical sciences ('natural philosophy'), by providing a theory that explained the creation and behavior of ideas. Berkeley was less influenced by the model of natural philosophy, aiming instead to combat skepticism and to support theism. The result was an exceptionally lucid statement and defense of idealism, the view that the only things that exist are minds and their contents. Locke had not taken this step, hoping to hold on to a material world even if our knowledge of its nature must be severely limited, but Berkeley argued that only idealism was compatible with the principles of empiricism. Hume was an idealist too, but made less of a fuss than Berkeley about this aspect of their philosophies, in part because Hume wished to avoid metaphysics, even of a negative sort.

In the nineteenth century, John Stuart Mill carried forward the tradition of British Empiricism from Hume, including his idealism, though in Mill's (1872) hands the position is normally referred to as 'phenomenalism.' Mill held that physical objects are 'permanent possibilities of sensation': to say that there is a tree in the quad at a certain time is to say either that someone is having an appropriate arboreal experience at that time or that they would have had one if they had been in the quad. Berkeley famously preferred to account for the existence of the tree when no human is observing it by appealing to God's constant observation, but in one or two passages he clearly presages Mill's counterfactual alternative, which appeals to possible human observations rather than to actual divine observations.

Mill went beyond his British precursors, however, in his treatment of mathematics, the most obvious candidate for nonempirical knowledge. Most empiricists have adopted the strategy of admitting that pure mathematics is nonempirical but denying that it provides worldly knowledge, claiming instead that it tells us only about the logical consequences of our definitions. Mill, however, maintained that mathematics, however pure, is empirically based, an inductive inference from pervasive regularities in our experience, such as those involving the manipulation of objects.

Twentieth-century empiricism remained inspired by science. Logical positivists and logical empiricists saw empiricism both as a way to understand scientific claims and as a way to make philosophy itself more scientific and less metaphysical (cf. Ayer 1936). They especially pushed the conceptual side of empiricism, for example, in the verifiability theory of meaning, whose motto was that 'the meaning of an expression is its method of verification.' On this view, the experiences that would confirm a claim also tell us what it means, and a claim which no course of observation could either confirm or refute is either a mere definition or metaphysical nonsense. Empiricism in this century has also been promoted by scientists themselves, in movements such as operationalism in the physical sciences and behaviorism in psychology. Another influential empiricist strain in this period has been Pragmatism, especially as developed by the American triumvirate of William James, Charles Saunders Pierce, and John Dewey, and more recently in a radical form that rejects definitions and meanings by Willard Van Orman Quine (1951).

2. *Concept Empiricism*

Empiricism about concepts is the view that our concepts derive in some sense from our experiences. Hume offered what has come to be a canonical version of this position, arguing that all our ideas are copies of preceding impressions, meaning roughly that every concept is a faint image of the experiences that cause it. Thus, to take a favorable example, our idea of the color blue is supposed to a faint image of the particular blue experiences we have enjoyed. For Hume, the copy principle was not just a fundamental fact about our cognitive economy but also a powerful philosophical tool. Constructively, the copy principle is supposed to allow us to analyze the content of an idea or concept, since by tracing ideas back to their source impressions we determine what that content really is. Destructively, the copy principle is supposed to expose empty talk or thought, since if a putative idea has no possible source impression, it must be devoid of content. The copy principle obviously requires some general demarcation between impressions and ideas. Hume draws it subjectively, in terms of the relative 'force and vivacity' of perceptions, recognizing the vagueness of this distinction, but trusting that we will all understand the difference between the vividness of what one sees as compared with faintness of memory and imagination.

Hume's most famous applications of his copy principle were destructive, part of his program to expose the emptiness of metaphysics. Thus, although he allowed that we do have an idea of the 'necessary connection' that is supposed to glue causes to their effects, the source impression reveals this idea to be purely subjective. For that impression is merely the feeling of expectation of a familiar effect when we see a familiar cause, something that is obviously only in our minds. We cannot conceive one thing making another thing happen: all we can conceive is a pattern of objects. Indeed although Hume often did write as if we could conceive of external objects, he seems to have held that the copy principle rules out conception of these too.

Hume's antimetaphysical fervor was taken up by the logical positivists of the first half of the twentieth century. They took statements rather than ideas as the unit of analysis, and held that a statement is meaningful only if there is some specifiable experience that would verify it. They made an exception of mathematics and logic, which were allowed to count as meaningful even though not empirical, but only because they consisted of definitions and their consequences rather than substantive claims. Since philosophical truths have often been supposed to consist precisely of nonempirical truths about the world, it is not surprising that the verification principle was supposed to sweep away great swaths of traditional philosophical doctrine, not on the grounds that such claims are false, but that they are literally meaningless, just noise.

As with Hume, the logical positivist program also had a positive side: for those statements that did manage to be meaningful, their meaning comprised the experiences that would confirm them. From this it follows that all claims are claims about experiences. Some tempered this extreme view by speaking of observable states of affairs rather than experiences, thus making the data public. In the hands of some behaviorists in the philosophy of psychology, this brought about a kind of reversal of the positivists' original, experiential doctrine, since now rather than analyzing all claims as claims about experiences, experiences and other mental states are analyzed in physical terms. On this view, the real meaning of the claim that someone is happy is not that the person is enjoying a certain kind of inner life, but that they are disposed to behave in characteristically happy ways. Hostility to this position is the source of the story of the behaviorist who, upon meeting another behaviorist in the street, proclaims 'You're fine, how am I?'

Concept empiricism has been extensively criticized in both its negative and positive aspects. The demarcation that concept empiricism attempts to draw between genuine concept or meaningful statement on the one hand and nonsense on the other depends on making reasonably precise the notion of a copy of an impression or the conditions under which evidence confirms a hypothesis. This has proven remarkably difficult so that, for example, familiar accounts of confirmation have been shown to have the consequence that everything is evidence for anything else, a result that would make all statements meaningful and is anyway obviously false. The other conspicuous weakness of concept empiricism concerns the specific meanings it attributes to statements that it rightly classes as meaningful. The attempt to reduce the meaning of a statement to the evidence that would support it typically both attributes meaning the statement does not have and leaves out meaning it does have. This comes out clearly in the case of scientific hypotheses that traffic in unobservable entities and processes. A hypothesis about the behavior of subatomic particles does not include as part of its *meaning* the diverse and indirect kinds of empirical evidence in its favor, while that meaning does include claims about unobservable particles.

3. *Knowledge Empiricism*

Knowledge empiricism is the view that all knowledge of the world is based on experience. Put negatively, this is the denial of *a priori* knowledge of the world. Empiricism here opposes rationalism, according to which the fundamental source of knowledge of the world is reason, not experience.

Critics of knowledge empiricism generate putative examples of *a priori* knowledge. Empiricists deal with these cases in several ways, claiming either that the statement in question cannot be known *a priori*, because it is unknowable altogether or knowable but only through experience, or that that the statement does not really provide knowledge *of the world*. Thus, when René Descartes, the great seventeenth century rationalist, claimed to know *a priori* of his own existence, Hume could respond that the only knowledge one has of oneself is of one's experiences, which are known through themselves. Mathematical knowledge is another popular area for rationalists' examples of the *a priori* although, as we have seen, Mill claimed that even this knowledge is ultimately empirical, resting on observations of the result of adding together and dividing apart physical objects. The more common empiricist treatment of mathematical knowledge, however, is to admit it does not depend on specific experiences but to deny that it is about the world. Both Hume and the logical empiricists admit that definitions or analytic truths—such as that a bachelor is an unmarried man—can be known to be true without experience, but maintain that these statements have no factual content. They then analyze mathematical axioms as definitions and treat everything that is derived from them as analytic truths which can be known without experience but tell us nothing about the world.

4. *Empiricism and Scientific Knowledge*

Alongside the question of whether there can be any knowledge independent of experience is the question of how far empirical knowledge may extend beyond the experience upon which it is based. Since Hume's seminal discussion of the problem of induction, it has become difficult to see how knowledge can extend beyond actual experience at all. It is neither surprising nor worrying that any nondemonstrative inference from experience might in principle go wrong; but Hume gave a remarkably resilient argument that we can have no reason whatever for extrapolating in one direction rather than in another, even when dealing

with everyday claims about the immediate future or the unobserved past and present. The nub of the problem is that any justification for the future success of our methods of prediction or extrapolation seems itself to rely on such a method, leaving us arguing in what is apparently a vicious circle, not unlike trusting someone solely on the grounds that he or she claims to be trustworthy. Thus empiricism has been a friend to the skeptic: if Hume is right, we have no more reason to trust the most sophisticated scientific predictions than tea leaves or entrails.

Even empiricists who, ignoring the Humean problem of induction, help themselves to inferences from past to future observations have had the greatest difficulty in showing how scientific knowledge is possible, since scientific theories appear to traffic in claims about entities and processes that are unobservable, not just unobserved. As we have seen, this is denied by the concept empiricists, for whom no claims about the unobservable can even be formulated, but few recent empiricists have been willing to go down this road. Instead, they have had to face the problem of the underdetermination of theory by data, developed with special force in the work of Pierre Duhem (1954) and Willard Van Orman Quine (1975).

The problem is that no matter how much evidence there is in favor of a particular scientific theory, there exist in principle many other theories, incompatible with the first yet compatible with all the evidence. This creates an epistemic stalemate that seems to block warranted inference to any claim that goes beyond what is observable.

Empiricists have attempted to deal with the underdetermination problem in a number of ways. Karl Popper (1959) adopted a radical solution that was designed to solve the Humean problem of induction as well. It was to deny that scientists need to indulge in nondemonstrative inferences at all. Instead, they are to attempt to refute their theories and then replace them by better ones, where this process relies only on a deductive inference from the falsity of a prediction to the falsity of the theory that entailed the prediction. This is empiricism in the sense that the rejection of a theory is determined by observational data. Popper is one of the few philosophers of science to have attracted the respect of a significant number of practicing scientists, probably because scientists appreciate being reminded of the importance of considering what sort of results would tend to undermine a hypothesis, but Popper's view has been severely criticized by philosophers. Falsifying an interesting and successful scientific theory is not nearly as straightforward as Popper would have scientists pretend. Moreover, even if Popper's methodology could be put into practice, it would still be a form of skepticism, since according to him we never have any reason whatsoever to think that any scientific theory or prediction is even approximately correct, nor can we have any reason to believe any claim about the future.

Empiricism continues to play an important role in the study of knowledge within science studies. One example is the 'strong program' in the sociology of knowledge, which rests on a kind of psychological empiricism. As David Bloor (1991) puts the matter, the strong program assumes that perception and thought are distinct but asymmetrically related: perception influences thought more than thought influences perception. Within the philosophy of science, a recent empiricist approach to scientific knowledge and the underdetermination problem is Bas Van Fraassen's constructive empiricism. Van Fraassen (1980) introduces a technical sense of 'acceptance,' where to accept a scientific theory is not to believe that it is true, but only that it is empirically adequate, where a theory is empirically adequate just in case everything it says about what is observable is true.

Constructive empiricism is the view that the aim of science is to discover theories that are empirically adequate, not theories that are true. The content of the theory that goes beyond its empirical claims is to be regarded as a model which may serve its function of generating true predictions without itself being a true representation of an invisible realm. Constructive empiricism thus appears to solve the underdetermination problem, since although the theories underdetermined by the data will be incompatible with each other, their empirical consequences, which is all the constructive empiricist would have us believe, may be compatible.

Constructive empiricism can thus be seen as a form of moderate skepticism, but it is still too skeptical for those who think that knowledge or at least reasonable belief about unobservable entities and processes is possible. Constructive empiricism also faces the difficulty of providing a nonarbitrary distinction between what is observable and what is not, a tricky problem faced in one form or other by almost all forms of empiricism.

See also: Darwin, Charles Robert (1809–82); Experimentation in Psychology, History of; Galton, Sir Francis (1822–1911); Helmholtz, Hermann Ludwig Ferdinand von (1821–94); Hume, David (1711–76); Locke, John (1632–1704); Logical Positivism and Logical Empiricism; Mill, John Stuart (1806–73); Pavlov, Ivan Petrovich (1849–1936); Positivism, History of; Positivism: Sociological; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Ayer A J 1936 *Language, Truth and Logic*. Victor Gollancz, London

Bennett J 1971 *Locke, Berkeley, Hume: Central Themes*. Oxford University Press, Oxford, UK

Berkeley G 1710/1975 A treatise concerning the principles of human knowledge. In: Berkeley G (ed.) *Philospohical Works*. Dent, London

Berkeley G 1713/1975 Three dialogues between Hylas and Philonous. In: Berkeley G (ed.) *Philosophical Works*. Dent, London
Bloor D 1991 *Knowledge and Social Imagery*, 2nd edn. University of Chicago Press, Chicago
Duhem P 1954 *The Aim and Structure of Physical Theory*. Princeton University Press, Princeton, NJ
Fraassen B V 1980 *The Scientific Image*. Oxford University Press, Oxford, UK
Hume D 1739/1975 In: Selby-Bigg L A, Nidditch P H (eds.) *A Treatise of Human Nature*. Oxford University Press, Oxford, UK
Hume D 1777/1975 In: Selby-Bigg L A, Nidditch P H (eds.) *An Enquiry Concerning Human Understanding*. Oxford University Press, Oxford, UK
Locke J 1700/1975 In: Nidditch P H (ed.) *An Essay Concerning Human Understanding*. Oxford University Press, Oxford, UK
Mackie J L 1976 *Problems from Locke*. Oxford University Press, Oxford, UK
Mill J S 1872 *An Examination of William Hamilton's Philosophy*. Longmans, London
Popper K 1959 *The Logic of Scientific Discovery*. Hutchinson & Co, London
Quine W V O 1951 Two dogmas of empiricism. *Philosophical Review* **60**: 20–43
Quine W V O 1975 On empirically equivalent systems of the world. *Erkenntnis* **9**: 313–28

P. Lipton

Employers' Associations

Employers' associations are formal organizations which specialize in the aggregation, definition, processing, and promotion in the political arena of the collective interests and goals of a distinct social class defined by its dominant position in the division of labor and its power to invest in market economies. Although the term 'employers' associations' applies specifically to the role played in labor relations, while the term 'trade associations' refers to the role played in industrial policy, the former is often used more generally to identify any kind of business interest associations. It is in this broader sense that it is used here.

Employers' associations are representative voluntary organizations—i.e., they are led by elected leaders and managed by professional managers—with a permanent staff for the administration of associational activities and the delivery of associational services, a budgetary process for the acquisition of funds and the accountability of their use, and a set of specific criteria defining the rights and duties of members. They can be analyzed in terms of their organizational properties, i.e., the domain of representation, the internal structure, the resources, and the tasks performed, and in terms of their degree of organizational complexity and relative autonomy. As formal representative organizations, they differ from other forms of organizations working for business, like clubs, law firms, and public relations agencies, as well as from more informal types of business association.

Business is a powerful pressure group in contemporary society, and employers' associations are just one of the four major ways by which businessmen can exert power and influence, both individually and collectively, and both in the market and in the political arena. At the level of the individual firm, entrepreneurs and managers play a crucial role through investment, production, and employment decisions in the market—which shape the economic and political environment of government policies—and through lobbying and other forms of pressure politics. Collectively, businessmen can exert power in the market through agreements for controlling prices, market sharing, the allocation of raw materials, and in the political arena through collective political action by its own specific representative organizations. In fact, in spite of social cleavages and cultural differences among entrepreneurs, and in spite of the different roles they play in diverse organizational settings (large and small businesses, private and public firms, local firms, and multinational corporations), the business class often acts today as an organized collective actor through its own specialized representative organizations. Why does this happens? Or, in other words, why should capital owners and top executives, who possess the discretionary power to invest, develop a need for collective interest representation?

Historically, there have been three main reasons by which employers may be brought to politicize their interests and to form organizations. The first is that their power to invest can be challenged by workers' unions. In most historical instances, business efforts to organize and coordinate actions came as a response to prior attempts by workers to pursue their interests by collective mobilization. Employers' associations have concerned themselves with assuring a stable and reliable labor force, and with the setting of shared rules of the game in labor disputes in order to institutionalize class conflict. Trade unions have become their most important institutional partners.

A second reason for entrepreneurs to organize have been the attempts to avoid cut-throat competition on the market, to prevent—or at least limit—the access of foreign competitors to the domestic market, to form a common front vis-à-vis the suppliers of basic resources. In fact, while free competition is a basic feature of capitalist market economy, competition, when pushed too far, can generate crises and contradictions which threaten the viability of the system as a whole. Formal coordination rather than that spontaneously created by the market may be necessary through associational activity and through government action.

The last remark leads to the third main reason for employers to organize, i.e., the systematic effort to influence government economic policies: resistance to decisions which either limit the freedom to invest, or

aim at redistributing wealth, or tend to increase the state direct intervention in key productive sectors. Democracies are consensus-bound, and competing political parties try to win the electoral support of social groups larger than business through fiscal, welfare, and other interventionist policies aimed at economic redistribution. Such policies are feared by employers because, directly or indirectly, they may threaten their discretionary power to invest, and consequently, they stir a political reaction which often takes the form of business interest organizations.

These three reasons have combined in various ways in different historical contexts to foster business associability. However, since, as will be argued later, greater power in the market place goes together with greater difficulties in producing collective benefits, collective political action for business can be defined as a second-best choice (Martinelli 1991). Whereas, in fact, labor can exert power only through collective organization, capital can exercise power also if not collectively organized. Capital can to some extent release itself from its dependence on labor through technical change, but labor cannot release itself from its dependence on capital for employment. Since, while it involves considerable risks and uncertainties, collective action does not grant capital owners similar competitive advantages as economic action in the market place, business interests become politicized whenever the power to invest is no longer sufficient for them to realize their economic interests. Associate action is therefore a second-best solution.

1. *The Status of the Research on Employers' Associations*

Research on business and employers' associations is much less developed than research on the working class and trade unions, in spite of early movements to the contrary. As early as 1907, at the first meeting of the German Sociological Association, Max Weber argued for assigning a very high priority to the systematic study of patterns of associations in general and to the associative activities of business in particular, which he considered a significant new form of collective action. Descriptions of the political role of specific firms and individual entrepreneurs, and references to the working of business associations as part of general pressure groups studies, do exist; but specific studies on the strategies, organizational structures, resources, activities, and attitudes of associations devoted to the promotion and protection of employer and trade interests have been rare (Ehrmann 1957, Linz and De Miguel 1966).

Only in the last two decades of the twentieth century has the study of business interests associations been developed through the collective research work of a group of political scientists and economic sociologists (Schmitter and Streeck 1981, Lehmbruch 1984, Offe and Wiesenthal 1985, Windmuller and Gladstone 1985, Grant 1987, Sisson 1987, Coleman 1988, Martinelli 1991, Sadowski and Jacobi 1991). Research focused on business associability both in national systems and in industrial sectors was mostly framed in the paradigms of neocorporatism and organized pluralism.

A set of related research questions have been studied. The question of why should capital owners and top executives, who possess the discretionary power to invest, develop a need for collective interest representation, has already been discussed; three further sets of questions have been addressed by recent comparative empirical research. The first concerns the distinctive character of business associability, i.e., the ways in which entrepreneurs and managers form an interest group different from others, and what is specific about it. The second focuses on the logic of membership, i.e., what organizational arrangements are required, given the basically individualistic and competitive structure of business interests, for entrepreneurs and managers to be able to associate with each other and to form collective interest organizations. The third set of research questions focuses on the logic of influence, i.e., what is the role of employers and business associations in policy making, both in policy advocacy and in policy participation. In all these questions the main dependent variables are the organizational dynamics, and the strategies that business interest associations develop, both in their relations with their members, and with outside groups and institutions, for their collective action to be successful.

1.1 *The Distinctiveness of Employers' Associations*

In contemporary poliarchies business is a privileged interest; it differs from other interests because it can exert power directly through individual pressures or by collective organized action of business interest associations; and, indirectly, because of the crucial role entrepreneurs and managers exercise over production, investment, and employment decisions. The specific nature of entrepreneurs' collective action can be better appreciated by contrasting it with that of workers. As Offe and Wiesenthal (1985) argue, unions and business associations are different, because their respective collective actions are based upon two different logics. Unions' collective action is more complex, since the definition of collective interests of individual workers results from their interaction in the union; this leads to an ongoing contradiction between bureaucracy and internal democracy, aggregation of individual interests, and formation of a collective identity. Business associations have an easier task: they represent a pure form of individualistic rationality, because each member can perceive individually what is his interest, and the association's task is only the coordination of individual behaviors into the most

effective strategy. Besides, labor has no alternative modes of collective action, whereas business has.

This view has been criticized, because it underestimates the problems of organizing business (Schmitter and Streeck 1981). Business and employers' associations cannot be seen only as tools of market rationalization. Besides problems similar to those of unions, they have to confront further specific problems rooted in the diversity of economic sectors. Collective action for entrepreneurs is not easier; on the contrary, it is more difficult than for workers for three major reasons. First, business associations must organize actors who normally compete with each other and have less incentive to cooperate than workers. Second, entrepreneurial culture considers competition a core, legitimate value, whereas for workers it is often seen as a violation of the value of class and group solidarity; while an entrepreneur who pushes his competitors out of business is recognized by other entrepreneurs as an example of efficiency and success, a worker outbidding a fellow worker to get his job runs the risk of being ostracized for disloyalty and breach of solidarity. Third, business associations have to organize a range of interests wider and more complex than unions; their members are both buyers and sellers of factors of production, goods and services.

For these reasons, the problem of producing collective political benefits, as well as selective benefits in the form of specialized services for members, should be even more pronounced and more visible for employers than for workers.

1.2 Relations with Members and the Logic of Membership

As any representative organization, employers' associations must, first of all, be able to recruit members and win their allegiance and support. The task is not an easy one, first, because collective political action for business is second best, a choice to be taken only when the discretionary power to invest is no longer sufficient to realize its economic interests.

Second, this task is not easy because Olson's theory of the logic of collective action applies to employers' associations even more cogently than to other large interest groups. Why should a rational entrepreneur, belonging to a large group (where the link between the contribution of the single participant and the outcome of the action is very weak, as it is in an industrial labor dispute), bear the cost of the action—as in the case of lost production because of a strike—when he can benefit from it without participating in it? This is the reason why employers and business interest associations develop a wide array of activities and services, such as legal and fiscal assistance, constant information about laws and regulations, professional education and specialized training, and so on, which are used as selective benefits aimed at fostering continuing membership.

Third, business and employers' associations in the relations with their members face another fundamental problem, i.e., the management of diversity. Some potential conflicts between their members are due to competition, while others arise from relations of mutual exchange. In fact, the more homogeneous a business association is with regard to its membership, the stronger is likely to be the competition among its members in the market. On the other hand, the more heterogeneous a business association, the greater the diversity of interests involving rates of exchange between different functional areas or sectors of the economy (Schmitter et al. 1980).

The management of diversity can be achieved through organizational development, defined in terms of the capability to organize complexity and to obtain relative autonomy from the state. Comparative research has shown that employers and business associations are the more effective the more developed and encompassing they are in scope and purpose; the more external effects and interdependences they are able to internalize; the more specialized and coordinated they are internally; the greater is their control over their environment; the more safely their supply of strategic resources is institutionalized; and the greater is their autonomous capacity to act and to pursue long-term strategies regardless of short-term environmental constraints and fluctuations. The logic of membership requires an articulated organizational structure that is capable of adapting to fragmented and segmented interests, and of fostering internal social cohesion and awareness of common interests among its members. The management of the relations between the association and its members depends on the specific demands and perceptions, on the number and dimensions of firms belonging to a given economic sector or region, the location of firms, and the degree of competition of the relative markets, and the degree of interdependency among firms; all variables which can foster tensions between individual and collective interests, and which can therefore affect the success of business representation.

1.3 The Role of Employers, Associations in Policy Marking

Pluralist and corporatist approaches have both studied the role of employers' associations in policy making. The basic difference lies in the fact that the former approach tends to consider organized business as one political actor among others (although often more influential than others) and consequently to confine its role to policy advocacy in pressure politics, whereas for the latter business organizations are active both in policy making and in policy implementation as a private interest lobby, and as effective intermediators between their members and the state.

In policy advocacy, employers' associations compete with other actors—such as unions, consumer asso-

ciations, environmental groups—in the attempt to influence the agenda of public policies, the formation of their content, and the way in which they are implemented once agreed in parliament and by government. Employers' associations try to capture distributional benefits through exchanges with the state, providing information and pledges of political support. Basic requisites for successful advocacy are: the constant processing of specialized technical information about policy making, the skilful evaluation of the economic and structural effects of policy proposals on members; the ability to mobilize political support at crucial stages in the progress of a policy proposal (which in turn requires the maintainance of internal cohesion—not an easy task for employers' associations given the great diversity of interests within them); the ability to cultivate public opinion, and to win public support, by presenting the specific purposes of organized business as compatible with broader national goals (such as economic growth, more effective competitiveness of the national economy in the global market, and employment opportunities); and the ability to establish relations with members of legislatures and top bureaucratic officials and to become parts of policy communities and 'iron triangles.'

According to other scholars, employers' associations can assume a second basic role in the political arena: the active involvement in formulating and/or implementing public policies. In order to play this more complex role, employers' associations must be able to coordinate the complex range of information and activities that they are asked to assume by their members, by other organizations like trade unions, and by the state. They must be independent of both their members and the state, and in some cases they must assume responsibility for controlling their members' compliance with agreed policy and for sanctioning their behavior.

Organizational development—which is an important variable in the analysis of the logic of membership—is also relevant in the analysis of the logic of influence. In fact, organizational development is also linked to policy achievement: associations that possess a developed organizational structure, sufficient autonomy from, and authority over, their members and the state, and are legitimated by the state, can play an effective role in the political arena. In turn, demonstrating effectiveness in securing the implementation of agreements leads to the delegation of greater authority to an association, thus increasing its importance with respect to both its members and the state.

In some countries, employers' and business associations have actually acted as institutions of social regulation, alongside markets, public bureaucracies, and communities. Whilst the guiding principle of the market is dispersed competition, that of the community is spontaneous solidarity, and that of the public bureaucracy is hierarchical control, the guiding principle of employers' and other organized interests is organizational concertation (Streeck and Schmitter 1985).

The case of the economies of continental Europe provides an interesting case. Here, the renewed focus on the central role of the market and of large multinational corporations in today's global economy does not explain completely the complexity of institutional responses to the problems of economic governance and industrial adjustment. Business interest associations, together with large firms, international markets, and government agencies have been part of a balanced mix of coordinating institutions and regulatory mechamisms (Chiesi and Martinelli 1989). This regulation concerns both the external environment and relations with other collective actors, and the internal environment in which associations play a role of self-regulation and mediation among various business groups.

2. *Conclusion*

The logic of membership and the logic of influence remain the key issue in this study of employers associations. Emerging lines of research focus on the newly emerging supranational and transnational forms of interest intermediation, and the evolution of pressure politics in contemporary polyarchies. The weakenings of the nation state, and the increasing interdependence of productive and marketing processes, in fact induce employers' associations to take action beyond the national boundaries. The focus is on the adaptability of national interest associations, on search processes, on structural and behavioral modifications, on organizational learning, design, and engineering, in the context of globalization and of the formation of supranational political entities like the European Union.

See also: Action, Collective; Alliances: Political; Competitive Strategies: Organizational; Corporatism; Industrial Relations and Collective Bargaining; International Business; Multinational Corporations; Policy Process: Business Participation; Trade Unions, Economic Behavior of; Voluntary Organizations

Bibliography

Chiesi A, Martinelli A 1989 The representation of business interests as a mechanism of social regulation. In: Lange P, Regini M (eds.) *State, Market and Social Regulation*. Cambridge University Press, New York

Coleman W D 1988 *Business and Politics. A Study of Collective Action*. McGill–Queen's University Press, Montreal

Crouch C, Traxler F 1995 *Organized Industrial Interests in Europe: What Future?* Aldershot, London

Ehrmann H W 1957 *Organised Business in France*. Greenwood Press, Westport, CT

Farago P, Kriesi H P 1986 *Wirtschaftsverbande in der Schweiz*. Rugger, Gruesch
Grant W 1987 *Business and Politics in Britain*. Macmillan, Basingstoke, UK
Hollingsworth J R, Boyer R (eds.) 1997 *Contemporary Capitalism. The Embeddedness of Institutions*. Cambridge University Press, New York
Lehmbruch G 1984 Concertation and the structure of corporatist networks. In: Goldthorpe J (ed.) *Order and Conflict in Contemporary Capitalism*. Oxford University Press, Oxford
Lindblom C E 1977 *Politics and Markets*. Basic, New York
Linz J, De Miguel A 1966 *Los empresarios ante et poder publico: el liderazgo y los grupos de intereses en el empresariado espanol*. Madrid
Marin B (ed.) 1989 *The Politics of Private Business and Public Enterprise*. De Gruyter, Berlin
Martinelli A (ed.) 1991 *International Markets and Global Firms. A Comparative Study of Organized Business in the Chemical Industry*. Sage, London
Mintz B, Schwartz M 1985 *The Power Structure of American Business*. Chicago University Press, Chicago
Offe C, Wiesenthal H 1985 Two logics of collective action: Theoretical notes on social class and organizational form. In: Offe C (ed.) *Disorganized Capitalism*. Polity, Cambridge, UK
Perez Diaz V 1985 Los empresarios y la clase politica. *Papeles de Economia Espanola* **22**: 2–37
Sadovski D, Jacobi O (eds.) 1991 *Employers' Associations in Europe: Policy and Organization*. Nomos Verlagsgesellschaft, Baden Baden, Germany
Schmitter P, Streeck W 1981 *The Organization of Business Interests: A Research Design to Study the Associative Action of Business in the Advanced Industrial Societies of Western Europe*. International Institute of Management, Berlin
Sisson K 1987 *The Management of Collective Bargaining. An International Comparison*. Basil Blackwell, Oxford
Streeck W, Schmitter P C 1985 Community, market, state—and associations? In: Streeck W, Schmitter P C (eds.) *Private Interest Government: Beyond Markets and State*. Sage, London
Useem M 1984 *The Inner Circle: Large Corporations and the Rise of Business Political Activity in the US and UK*. Oxford University Press, London
Windmuller J P, Gladstone A (eds.) 1984 *Employers Associations and Industrial Relations*. Clarendon Press, Oxford

A. Martinelli

Employment and Labor, Regulation of

Employment regulation consists of the legal rules and institutions that constitute and structure the employment relationship. Employment regulations determine the nature of the labor contract, the obligations of each party, and the legal means by which each party may press for advantageous terms. It includes not only affirmative interventions by the state into the employment relationship through statutes, administrative regulations, executive orders, or other acts of positive law, but also the background contracting rules with which a buyer and a seller of labor power, that is, an employer and an employee, form an employment relationship.

The content of employment regulation determines the parties' bargaining power vis-á-vis each other (Hale 1923). For example, legal rules determine whether workers are free to join together into unions or whether such actions constitute unlawful conspiracies. Legal rules also determine whether and under what circumstances workers can lawfully strike and picket, and whether they can use secondary pressure or appeals to the general public to support their position. In these and many other respects, the legal rules that govern the employment relationship have powerful distributional consequences.

The content of employment regulation is a product of the each side's relative power in the political process. Employers and unions operate as pressure groups, lobbying directly or indirectly galvanizing public opinion to shape the substantive law of employment in ways that will give them advantage. Legislatures and judges respond to the competing demands of employers and employees when they decide cases and enact employment-related statutes (Kennedy 1997).

In most Western nations, there are two types of positive employment regulations—regulations that directly determine specific terms of the individual wage bargain and regulations that legally empower labor and management to engage in self-regulation (Kahn-Freund 1983). Most northern European countries have detailed legal codes that specify many terms of the employment relationship such as mandated hours of work, safety conditions, and childcare leave policies. Most also have specialized labor courts. In Germany and France, terms negotiated between unions and employers can be transformed into generally applicable legal standards for an entire industry or region (Weiss 1987). In contrast, Great Britain relies primarily on the common law to regulate the employment relationship, and leaves many important issues concerning employment outside the law altogether (Wedderburn 1986).

Countries differ as to the content as well as to the form of their employment regulation. For example, some countries treat the individual employment contract as imposing affirmative duties on both the employer and the employee. For this reason, in most parts of Europe and Latin America, an employer violates the employment contract by firing an employee without good cause. In the United States, on the other hand, the employment contract is at-will, which means that employees who are terminated unfairly have no legal recourse.

Countries also differ as to what kinds of rights, privileges, and protections they give to organized labor. Some countries give very broad protection to labor organizing, picketing, and strikes, and some do not.

There are also significant differences between countries as to the role they give to organized labor

in corporate decision-making. Germany, France, and the Netherlands have statutes that mandate that employees be represented in their firms by elected works councils. Germany and Sweden mandate that employee representatives serve on corporate boards of directors in certain types of firms (Summers 1980). In Great Britain and the United States there are no statutorily-mandated participation rights.

This article traces the development of employment regulation in the United States. While the specific evolution of employment law in the US is unique, parallels can be found in many countries.

1. Early Employment Regulation in the United States

In the United States, employment was regulated primarily by the common law until the 1930s. Prior to the nineteenth century, there were very few workers' as we understand the term today. America's workforce was largely made up of skilled masters, journeymen, apprentices, indentured servants, and slaves and others in status relationships. The employment relationship in the United States was regulated under the common law of the master–servant relationship.

The master–servant law was hierarchical in structure and harsh in application. Servants were required to give masters not only labor power, but also obedience. The law condoned the use of corporal punishment against insubordinate or lazy workers. In addition, most states had statutes making it a crime for a worker to quit his or her job. Other conditions for employment—such as the hours of labor, the length of rest hours, the grounds for dismissal, and so forth—were set forth in statutes, leaving little to individual variation or private bargaining (Steinfeld 1991).

In the mid-nineteenth century, there arose for the first time a significant number of people known as 'free laborers.' Some were immigrants and former farm laborers who migrated to cities to become factory workers. Some were former slaves, indentured servants, and others in the status of bound labor who were converted into free labor by the passage of the 13th amendment and antipeonage statutes after the Civil War. And some were journeymen and apprentices who became free as craft guilds declined (Steinfeld 1991, Tomlins 1993).

Free labor was not as free as its name suggested. A number of common law doctrines made the worker dependent upon his employer. For example, under the 'entire contract' doctrine, if a worker left the job for any reason prior to its term, the employer was not obligated to pay for work already performed. The entire contract doctrine gave employers an incentive to treat employees harshly, with the hope that they would quit and forego their right to payment for work already performed. In addition, it was unlawful for one employer to entice workmen from employment at another establishment. Thus, worker mobility was curtailed (Orren 1991).

Toward the end of the nineteenth century, a new common law doctrine was applied to govern employment relationships that had no fixed term. Beginning in the 1880s, courts began to hold that a contract of employment of indefinite duration is terminable by either party at any time for any reason or no reason at all (see *Payne* v. *Western and A. H. RR Co.*, 81 Tenn. 507 (1884)). Under this 'at-will rule,' an employer was permitted to fire a worker at will, and a worker could likewise quit at will and nonetheless be paid for work performed. Thus at its inception, the at-will rule was a liberating reform for American workers. It was the beginning of the modern law of individual employment (Orren 1991)

While the law of individual employment was undergoing a fundamental transformation in the late nineteenth century, the law of collective employee action was not. At common law, unions were treated as dangerous and unlawful combinations. Throughout the nineteenth century, strikers and pickets were indicted for criminal conspiracy. Organized labor and progressive reformers organized political opposition to the use of conspiracy doctrine against labor. As a result of such pressure, by 1900 some state courts changed their approach toward collective labor action, and held that peaceful collective action by employees was lawful so long as the union's objective was legitimate (see *Plante* v. *Woods*, 176 Mass. 492 (1900)).

2. Early Labor Legislation and the Doctrine of Substantive Due Process

Before 1900, most labor law was state law rather than federal law. Some labor organizations successfully pressured state legislatures to enact legislation favorable to workers, such as legislation exempting trade unions from the conspiracy laws, guaranteeing workers payment of wages, restricting the hours of work, and providing health and safety protections for minors and women. There was no federal labor legislation in these areas. The Supreme Court adhered to a narrow interpretation of the Commerce Clause that said that Congress had no power to regulate private employment except on the railroads. In addition, in the early twentieth century, Supreme Court adopted a doctrine, known as 'substantive due process,' that said that legislative protection for labor rights and employment standards was a violation of freedom of contract and therefore unconstitutional under the Due Process clauses in the 14th and 5th Amendment. The doctrine was articulated in the cases of *Lochner* v. *New York*, 198 U.S. 45 (1905), striking down a state law establishing maximum hours of work for bakery

employees, and *U.S.* v. *Adair*, 208 U.S. 161 (1908), invalidating a federal law prohibiting railroad employers from discriminating against employees because of union membership.

Ironically, for the first quarter of the twentieth century, the only federal legislation that had generalized application to American workers was not labor legislation but anti-monopoly legislation intended to break up the trusts and to promote business competition. The Sherman Antitrust Act, enacted in 1890, declared illegal 'every contract, combinations in the form of a trust or otherwise, or conspiracy, in restraint of trade or commerce among the Several states' (15 U.S.C. X 1 *et seq*). At the time of its passage, proponents believed it was a business regulation that would help restore free competition to the product markets (Sanders 1999). However, in 1908, the Supreme Court held that the Sherman Act applied to labor unions, and imposed severe sanctions on union members for conducting a product boycott in support of organizing efforts (*Loewe* v. *Lawlor*, 208 U.S. 274 (1908)). The case was a bitter lesson for the labor movement, effectively ending the use of product boycotts in union organizing (Ernst 1995).

After 1908, the American Federation of Labor (AFL) began a legislative campaign to obtain a labor union exemption from the Sherman Antitrust Act. It joined forces with the Democratic Party in the congressional and presidential elections, and was rewarded for its efforts by the enactment of the Clayton Antitrust Act of 1914 (15 U.S.C. X 12 *et seq*). On the surface, the Clayton Act created an immunity for labor from the operation of the Sherman Antitrust Act. But in 1921, in *Duplex Printing Press* v. *Deering*, 254 U.S. 443 (1921), the Supreme Court gave the Clayton Act a narrow interpretation and held that it did not protect unions that were engaged in secondary activity from antitrust liability.

Despite the legal setbacks for unions in the 1920s, the railroad unions succeeded in obtaining legislation that gave them rights to organize and bargain collectively. After many decades of tumultuous labor relations and several unsuccessful regulatory attempts, the railroad unions and the railroad operators together drafted the Railway Labor Act of 1926. The resulting statute gave railroad workers a right to organize, protected them against dismissal for union activity, and imposed on unions and railroads a duty to bargain and to maintain the status quo while negotiations were in progress.

3. *The 1930s: The New Deal and the Origin of Contemporary Employment Regulation*

The 1930s was the decade of the Great Depression. With the election of Franklin D. Roosevelt and a Democratic Congress in 1932, the government embarked on an ambitious program of economic and social reconstruction called the New Deal that included affirmative steps to assist American workers.

The New Deal Congress passed three major statutes to address the problems of American workers. The first, enacted in 1932, was the Norris–LaGuardia Act, 29 U.S.C. XX 101 – 15, which declared that it was the public policy of the United States to support unionism and collective bargaining. The Act took away the power of federal courts to issue injunctions in labor disputes. Then in 1935, Congress enacted the National Labor Relations Act (NLRA), 29 U.S.C. X 150 *et seq*, which gave workers a legally enforceable right to organize, strike, engage in collective bargaining, and take other 'concerted action for mutual aid and protection.' The NLRA also established an administrative agency, the National Labor Relations Board (NLRB), to enforce those rights. And in 1938, Congress enacted the Fair Labor Standards Act, which established a minimum wage and set maximum hours of employment (29 U.S.C. XX 201 B 19).

The three New Deal labor statutes dramatically restructured the relationship between private employment and the state in the United States. All three were explicitly redistributive in their goals, and they all embodied the premise that it is legitimate, proper, and constitutional for Congress to legislate about the substantive terms of the wage bargain. Thus, they reflected a radical transformation in Congress' view of the capacity of government to intervene in the private economy (Irons 1982). Yet, the three statutes also retained a commitment to private bargaining as the core mechanism for establishing terms of employment.

There were also two important Supreme Court cases in the 1930s. In 1937, in *West Coast Hotel* v. *Parrish*, 300 U.S. 379 (1937), the Supreme Court abandoned the doctrine of substantive due process and instead upheld a state law setting maximum hours of work for women and children. That same year, the Supreme Court decided *Jones & Laughlin v. NLRB*, 301 U.S. 1 (1937), upholding the constitutionality of the NLRA and at the same time adopting an expansive definition of the Commerce Clause. These two decisions laid the groundwork for Congressional regulation of the private economy. They provided a constitutional foundation for the two-faceted labor law regime of the postwar era—legal support for collective bargaining and government mandated minimum terms of employment. As a result of the New Deal statutes, combined with rank and file determination and union organizing efforts, union membership increased from under 3 million in 1933 to more than 12 million by the early 1940s.

4. *World War II and the Postwar Period*

During World War II, the US government turned to private industry to help operate a wartime economy.

President Roosevelt also appealed to organized labor to help ensure labor peace. The unions agreed and gave a no-strike pledge in exchange for significant union representation on the War Labor Board and other government agencies involved with the war effort. After the war, the years of frozen prices and pent-up consumer demand fueled a major inflation. Unions, whose members' wages had been frozen for years, engaged in an unprecedented strike wave in an attempt to catch up with rising prices. Employers saw this is a propitious time to attempt to amend the NLRA and weaken its protections for unions. Employer associations waged a vigorous campaign and, in 1947, secured passage of the Taft–Hartley amendments. The amendments curtailed union power by prohibiting secondary boycotts, banning closed shops, and restricting union-only hiring halls.

In the postwar decades, unions continued to grow in numbers and strength. In 1953, the two competing federations merged to form the AFL–CIO. Throughout the 1950s and 1960s, unions successfully negotiated wage increases, cost of living clauses, health and pension benefit programs, and just cause dismissal protection for millions of American workers.

One area in which the labor movement made great strides was in the establishment of in-plant arbitration systems. Labor arbitration is a system of dispute resolution that uses a third-party neutral decision-maker, selected by the parties, rather than a judge or hearing officer to decide disputes that arise under existing contracts. During World War II, grievance and arbitration clauses spread to many industries with the active approval, and at times insistence, of the War Labor Board (Atleson 1998). After the war, arbitration clauses became widespread in the collective bargaining agreements (Stone and Van 1981).

Collective bargaining provided the primary form of employment regulation for American workers in the postwar period. However, the gains and protections obtained through collective bargaining only applied to unionized workers. In the 1960s and 1970s, Congress enacted employment laws that had general applicability. In 1964, the civil rights movement achieved the enactment of the Civil Rights Act which contained Title VII, 42 U.S.C. X 2000e – 17, a broad prohibition on discrimination in employment. Also in that decade, the labor movement succeeded in securing the enactment of the Occupational Safety and Health Act of 1970, 29 U.S.C. 651–687 and the Employee Retirement Income Security Act of 1974, 29 U.S.C. 1001–1461.

5. *The Decline of Unions and the Rise of Individual Employment Rights*

The New Deal system of employment regulation accomplished its redistributive goals throughout the postwar period. In that period, labor's share of national income rose and the labor movement held considerable political power. However, by 1979, that began to change.

The decade of the 1980s was a watershed for American workers. There had been a slow decline in union density from the mid-1950s onward, but in the 1980s union density plummeted as a result of increased employer opposition and changes in the interpretation of the NLRA that made it more difficult for unions to operate effectively (Stone and Van 1988). After 1979, for the first time since World War II, real wages began to decline. Average hourly and weekly wages declined by 11 percent between 1973 and 1990. At the same time, a wave of corporate takeovers and mergers produced massive employee dislocation. The three interrelated factors of union decline, wage decline, and job loss produced soaring rates of profit for corporations and hardship for American workers (Bluestone and Bluestone 1992, Kuttner 1997).

In response to the massive layoffs of the 1980s, state law-makers began to provide protections for the job security of individual employees. State legislatures enacted measures to provide protection for whistle-blowers, protection for privacy on the job, and other job-related rights. Also, some state courts began to carve out exceptions to the at-will employment rule and provide remedies for workers who were wrongfully dismissed. The expansion of individual employment rights in the 1980s shifted employment regulation from the federal government to the states. The shift corresponded with the demise of the union movement and the end of the redistributive impact of the New Deal system of employment regulation (Stone and Van 1992).

6. *The Prospects for Transnational Labor Regulation*

In the 1990s, unions further declined and income distribution became ever more unequal. At the same time, trade liberalization policies and the development of trading blocs introduced a powerful threat to domestic employment regulation—the threat of global regulatory competition. Businesses faced with competitive pressures from global trade or presented with lucrative opportunities for direct foreign investment to pressure national governments to weaken their labor standards and defer enforcement of the regulations they have. The threat of capital movement generates a downward spiral of deregulatory reforms. As a result of this regulatory competition, firms move around the globe like checkers, jumping from place to place to avoid onerous employment laws.

At present, the most pressing issue of employment regulation lies at the transnational level. Labor organizations and policy-makers in many countries have made proposals for a regime of transnational labor

regulation that could impose minimum global labor standards and thus limit regulatory competition. However, to date, none of them has been adopted. Instead, trade experts insist that labor considerations should not be included in trade agreements (Bhagwati 1994). Thus, the world trading system is evolving without any effective mechanism to protect labor standards, creating an unregulated global labor market in which companies can easily avoid domestic labor unions and national labor regulation.

In the emerging global labor market, employment regulation will not disappear, but it will be transposed. Just as in the early twentieth century, the doctrine of substantive due process served as an ideological barrier to state intervention in the labor contract, so too the present-day commitment to a form of free trade stripped of social concerns is an ideological barrier to the development of global labor rights. In this environment, employment regulation will occur through invisible mechanisms, subsumed in the emerging notions of property and contract, and will have the effect of exacerbating problems of inequality and social instability. Today's free trade orthodoxy threatens to prevent transnational actors from addressing social concerns, raising living standards, or redressing inequalities. An alternative regime of transnational labor regulation will be necessary if the emerging global economy is to promote social justice and protect the dignity of workers worldwide (Stone and Van 1999).

See also: Databases, Core: Sociology; Discrimination; Discrimination, Economics of; Statistical Systems: Labor; Labor History; Labor Movements, History of; Labor Unions; Regulatory Agencies; Regulation: Working Conditions; Workplace Environmental Psychology; Work, History of; Work, Sociology of; Workplace Safety and Health

Bibliography

Atleson J 1998 *Labor and the Wartime State*. University of Illinois Press

Bluestone B, Bluestone I 1992 *Negotiating the Future: A Labor Perspective on American Business*. Basic Books, New York

Bhagwati J 1994 Fair Trade, reciprocity, and harmonization: The new challenge to the theory and policy of free trade. In: Deardorff and Stern (eds.) *Analytical and Negotiating Issues in the Global Trading System*. University of Michigan Press

Ernst D R 1995 *Lawyers Against Labor*. University of Illinois Press

Hale R 1923 Coercion and distribution in a supposedly non-coercive state. *Political Science Quarterly* **38**: 470

Irons P H 1982 *The New Deal Lawyers*. Princeton University Press, Princeton, NJ

Kahn-Freund O 1983 *Labor and the Law*, 3rd edn. Stevens and Sons

Kennedy D 1997 *A Critique of Adjudication*. Harvard University Press

Kuttner R 1997 *Everything for Sale: The Virtues and Limits of Markets*. Knopf, New York

Orren K 1991 *Belated Feudalism: Labor, the Law and Liberal Development in the United States*. Cambridge University Press

Sanders E 1999 *Roots of Reform*. University of Chicago Press, Chicago

Steinfeld R 1991 *The Invention of Free Labor: The Employment Relation in English and American Law and Culture*. University of North Carolina Press, pp. 1350–1870

Stone K, Van W 1981 The postwar paradigm in American labor law. *Yale Law Review* **91**: 1509

Stone K, Van W 1988 Labor and the corporate structure: changing conceptions and emerging possibilities. *University of Chicago Law Review* **55**: 73

Stone K, Van W 1992 The legacy of industrial pluralism. *Chi. Law Review* **59**: 575

Stone K, Van W 1999 To the Yukon and beyond: Local laborers in a global labor market. *Journal of Small and Emerging Business Law* **3**: 93

Summers C 1980 Worker participation in the U.S. and West Germany. *American Journal of Comparative Law* **28**: 367–93

Tomlins C L 1993 *Law, Labor, and Ideology in the Early American Republic*. Cambridge University Press, Cambridge, UK

Wedderburn W 1986 *The Worker and the Law*. Pelican Books, London

Weiss M 1987 *Labor Law and Industrial Relations in the Federal Republic of Germany*. Kluwer, Boston

K. V. W. Stone

Encoding Specificity in Memory

The notion of encoding specificity was proposed, and has ever since been vigorously championed, by Endel Tulving. It is integral to his concept of an *episodic memory* system, which pertains to memories for personal experiences as opposed to mere knowledge (Tulving 1972). Tulving conceptualizes each episodic memory as embodied in a physical entity, referred to as a *memory trace* or *engram*, that serves to bridge the temporal gap between the episode and its recollection. According to the encoding specificity principle, the likelihood of a given cue eliciting recollection of the episode depends on how the episode was encoded into the trace.

Nowadays, it is easy to dismiss this notion as obvious, but at the time of its introduction around 1970 it was more commonly dismissed as obviously wrong. Key to this change in fortunes has been an increasing appreciation of the role of encoding, and in particular idiosyncratic encoding, in the memory process.

1. Idiosyncratic Encoding

Tulving and his associates were quick to credit Richard Semon with a biological rendition of the encoding specificity principle, and indeed of their entire conception of the episodic memory system (Schacter et al. 1978). But prior to this acknowledgment Semon was virtually unknown to psychology. By the time his views were translated into English (Semon 1921), mainstream American experimental psychology was taking an objectivist turn, and before long encoding was thought to add nothing to what could be objectively discerned from the stimulus conditions.

There were, of course, other approaches to psychology, some of which underscored the subjective nature of encoding. For example, gestalt psychologists used ambiguous figures to demonstrate that one and the same stimulus could result in distinctly different perceptions (see *Perceptual Organization*), and moreover they conceptualized memory for such perceptions as embodied in memory traces. Nevertheless, it was not until the adoption of the information-processing metaphor as the dominant model of cognitive psychology in the second half of the twentieth century (see *Cognitive Psychology: History*) that the concept of encoding assumed its current theoretical prominence.

Among the many demonstrations of the role of encoding in memory that came with the dawning of the information-processing era was Conrad's (1964) finding that an error in the written reproduction of a short visual sequence of letters tended to be acoustically similar to the correct letter. Thus, the letter S was 65 times more likely to be recalled as an F than as a B, despite not looking more like an F than a B. The implication is that performance in this task is mediated by inner speech, even though the letters are presented and reported silently. Such encoding is perhaps virtually universal under these conditions, but under other conditions encoding can be willful and idiosyncratic. The cocktail party phenomenon, for example, illustrates an ability to attend selectively to just one of multiple speakers. On the other hand, that such willful control is imperfect is shown by a special sensitivity to our own name spoken in what would appear to have been an unattended conversation. In this case, encoding is unwilled, yet clearly idiosyncratic.

Throughout the 1960s Tulving, among others, emphasized the hitherto neglected *recall* phase of *remembering*. Recall of an episode at any given time, he argued, depended on what was being encoded at that time, as well as on how the episode had been encoded into its memory trace (see *Memory Retrieval*). More particularly, he came to promulgate the encoding specificity principle—the dictum that recall of the episode depends on the relation between current encoding and the information available in the episode trace. Thus, the encoding specificity principle is best illustrated by experiments in which control is brought to bear on the encoding conditions prevailing at both the study and test stages of a memory procedure. A few examples warrant review.

2. Item-specific Context Effects

In a seminal demonstration of encoding specificity, Tulving and Osler (1968) presented a list of 'target' words (e.g., STOMACH, HEALTH, EAGLE), each being shown alone for some subjects (STOMACH), in the presence of a weakly associated 'cue' word for others (*empty* – STOMACH), and in the presence of two weakly associated cue words for yet others (*empty, hurts* – STOMACH). They then tested for recall of the target words. Of the subjects in each of the three study conditions, some were given free recall instructions (i.e., no specific cues), others were given one of the two cue words of each target word (e.g., *empty* – ?), and yet others were given both of the cue words (e.g., *empty, hurts* – ?). Subjects who saw cues at neither study nor test recalled, on average, 44 percent of the target words. The various groups of subjects for whom the cue, or at least one of the two cues, presented at study was also presented at test recalled an average of 60–62 percent of the target words, depending on the group. Subjects for whom no cue word presented during study appeared at test, or for whom the target words had been presented alone at study and tested with one or both of their cues, recalled an average of 29–36 percent of the target words, again depending on group. The authors drew the then-provocative conclusion that, 'specific retrieval cues facilitate recall if and only if the information about them and about their relation to the (to-be-remembered) words is stored at the same time as the information about the membership of the (to-be-remembered) words in a given list' (Tulving and Osler 1968, p. 599).

With the distinction between episodic memory and general knowledge yet to be drawn, it would have been natural to attribute the ineffectiveness of the extralist cues (i.e., words not presented at study but presented at test) to the weakness of their associations to the target words, and the effectiveness of the intralist cues to the temporary boosting of these associations in the study phase of the procedure. Evidence contrary to this account was provided by Thomson and Tulving (1970). In one of their experiments, target words were studied either alone (e.g., FLOWER) or in the context of a weakly associated cue (*fruit* – FLOWER), and tested under conditions of free recall (i.e., in the absence of explicit cues), or in the presence of the weakly associated cue (*fruit* – ?) or of a strongly associated extralist cue (*bloom* – ?), with a different group of subjects serving under each of the six combinations of study and test conditions. When no context words had been present at study, the proportion of targets recalled was, relative to conditions of free recall (.49), increased by the strong cues (.68) but not by the weak cues (.43). By contrast, when

weakly associated context words had been present at study, the proportion of targets recalled was, relative to conditions of free recall (.30), increased by the weak cues (.82) but not by the strong cues (.23). This pattern of results is not readily reconciled with what Thomson and Tulving called the hypothesis of associative continuity, whereby no distinction is made between episodic memory and general knowledge. It is, on the other hand, in keeping with a principle of encoding specificity operating within the realm of a distinct episodic memory system.

The critical finding in each of these two studies is the study–test interaction, which indicates that the relative effectiveness of different test conditions depends on the conditions under which the items were studied. Such interactions constitute the evidence for encoding specificity, and more generally for the premise of a distinct episodic memory system. In one of the many subsequent illustrations of study–test interactions (Fisher and Craik 1977, see also Morris et al. 1977), target words were presented in the context of either semantically or phonemically related words (*sleet* – HAIL or *pail* – HAIL). Recall was then measured in response to extralist cues that also were either semantically related or phonemically related to the target words (*snow* or *bale*). The semantically related cues were more effective than the phonemically related cues for words studied in the context of semantically associated words (.36 versus .22), but not for words studied in the context of phonemically associated words (.16 versus .18). This interaction is at variance with one particularly influential idea of the time, namely Craik and Lockhart's (1972) notion that memory for an item is determined by the depth to which it was processed, with semantic encoding being assumed to reflect greater depth than phonemic encoding (see *Memory: Levels of Processing*). Fisher and Craik suggested supplementing this idea with that of encoding specificity.

That specificity of encoding can occur entirely within the semantic realm is implied by the finding that a recall cue for an unambiguous target word may be more or less effective depending on its consonance with the context of the target word at study. Thus, *something heavy* is a more effective cue than *something with a nice sound* for the target word PIANO presented in the study sentence, 'The man lifted the piano,' but not for the same target word presented in the study sentence, 'The man tuned the piano' (Barclay et al. 1974).

Perhaps the most influential support for the encoding specificity principle has been the phenomenon of recognition failure of recallable words (see *Recognition Memory, Psychology of*). In his quest for a truly general theory of episodic memory, Tulving came to relinquish the widely held assumption of a qualitative difference between the 'copy' cue in a *recognition* test (i.e., the target word itself presented as a test item) and other kinds of memory cue, and suggested instead that the copy cue was, like any other cue, fallible, and so could prove ineffective even though another cue would be effective (Tulving 1983). To evaluate this suggestion, Tulving and Thomson (1973) presented three consecutive lists of 24 weakly associated word pairs (e.g., *ground* COLD, *head* LIGHT, *bath* NEED), and following each of the first two lists they re-presented in a different order the left-hand words as recall cues for the target right-hand words. The subjects expected to be tested the same way for the third list, and indeed they were, but not until they had engaged in two other tasks. First, they free associated to each of a sequence of stimulus words (e.g., *want*, *dark*, *hot*). Unknown to the subjects, each of these stimulus words had been selected as likely to yield one of the target words from the just-studied list (namely, NEED, LIGHT, and COLD, respectively). The subjects then looked over their free associates and circled any they recognized as a target from the study list. Finally, they were given the recall test they had expected, with the left-hand words from the third study list presented as cues for their respective paired words. A substantial proportion of the words produced in the free association task went unrecognized and yet were produced in the ensuing recall test. Across three experiments, the probability of a target free associate being recognized as a target was only about half that of its being recalled in response to the weak associate with which it had been paired. Subsequent research has shown this phenomenon of recognition failure of recallable words to be remarkably robust. Tulving and his associates have interpreted it in terms of the encoding specificity principle: the targets were encoded into memory traces in such a way that the traces tended to be more responsive to the context words with which the target words had been studied than to copies of the targets themselves.

This interpretation implies that level of performance on a recognition test may well depend on the congruence between the cognitive environments at study and test. Just such a relation was demonstrated at the very time the encoding specificity principle was being developed (Light and Carter-Sobell 1970). Subjects studied sentences that included a capitalized two-word phrase (e.g., 'The STRAWBERRY JAM tasted great') in anticipation of a memory test for these phrases. In fact, they were given a recognition test in which the test items were word pairs and the task was to decide whether the second word of each pair (JAM) had been included in the study list. Recognition of a target word was much more probable if its context was the same as at study (STRAWBERRY JAM) than if it was different (TRAFFIC JAM). Across various subconditions and experiments, probability of recognition, corrected for guessing, ranged between two and three times greater in the same context condition than in the different context condition.

Of course, *jam* is a homograph, as were the other target words in this study. According to the encoding

specificity principle, however, context may affect recognition of any word, and indeed context effects in the recognition of words that are not transparently homographic have been demonstrated. Thus, Tulving and Thomson (1971) found that a target word (e.g., WATER) studied alone was more likely to be picked out in a recognition test when presented alone (0.70) than when presented in the context of either a weakly (WHISKEY) or strongly (LAKE) associated word (0.49 and 0.59), and that a target word studied in the context of a weakly or strongly related context word (WHISKEY–WATER or LAKE–WATER) was more likely to be picked out at test when presented in the same context (0.85 and 0.88) than when presented alone (0.77 and 0.70) or in a different context (0.58 and 0.66).

Context effects in recognition memory have been shown to extend to nonverbal stimuli, such as faces. In one study (Watkins et al. 1976), subjects were shown photographs of unfamiliar faces, each paired with a descriptive phrase (e.g., *keeps tropical fish*, *drives an Italian sports car*, *works in cancer research*). The phrases and faces were then shown in a recognition test, with half paired as before and half re-paired. The subjects' task was simply to identify the faces regardless of context. Probability of identification was greater for same-context faces than for re-paired faces (0.84 versus 0.71).

In each of these examples, as with most other studies cited as support of the encoding specificity principle, encoding was controlled by manipulation of the physical environment. According to the thinking behind the encoding specificity principle, however, the effects of the physical environment are mediated by the internal, or cognitive, environment. Consistent with this premise, it should be possible to show, for example, context effects in recognition memory by having the subjects impose contexts in a purely mental way, and such effects have indeed been demonstrated. For example, if instructions in the study phase of a recognition memory experiment call for the mental filling in of irregular shapes with specified colors, then subsequent identification of the shapes, regardless of color, tends to be more accurate if they are shown filled in with the imagined color than with another color (Watkins and Schiano 1982).

3. General Context Effects

Memory may depend on the reinstatement of, not only item-specific contexts, but also general contexts shared by many or all of the target items (Davies and Thomson 1988). For example, there have long been claims of a beneficial effect of using the same room for study and test (Abernathy 1940, Smith et al. 1978; but see Saufley et al. (1985) for an example of a failure to replicate this effect). Similarly, an item encountered in the presence of a particular odor is subsequently more likely to be remembered in the presence of the same odor than in the presence of a different odor (Cann and Ross 1989, Angleton and Waskett 1999). In the same vein, recall has been found to be more successful when study and test are both on land or both underwater than following a study-to-test switch in these environments (Godden and Baddeley 1975). Also, memory tends to be more effective when drug or mood state at test matches that at study (see *Mood-dependent Memory*). And, like item-specific context effects, general context effects can be psychogenic. For example, adopting the perspective of a burglar or of a home buyer while recalling a story about an old house has been shown to differentially affect recall of such items as 'rare coin collection' and 'leaky roof' (Anderson and Pichert 1978) (see also *Schemas, Frames, and Scripts in Cognitive Psychology*).

4. Criticisms

Not all aspects of Tulving's account of encoding specificity are widely accepted. For example, its logic notwithstanding, there is little appeal to Tulving's (1983) argument that the interaction of study and test conditions precludes characterizing one episode as more memorable than another. We do not deny that Mother Teresa was a better person than Adolf Hitler just because in some ways (e.g., as an orator, artist, or military strategist) she probably was not, so why should we deny that some memories (e.g., my memory for the birth of my first child) are stronger than others (e.g., my memory for putting on my socks that same day)? More importantly, the very idea of episodes being encoded into discrete traces within a distinct episodic memory system is problematic. For one thing, it is an instantiation of Tulving's notion of availability. This notion has never been adequately operationalized (Watkins 1978), and consequently the distinction between encoding specificity and such process oriented notions as transfer-appropriate processing (Morris et al. 1977) is shaky at best. Also, the relation among traces has not been worked out. If, for example, memory for an item from a study list is represented in a trace, then how is memory for the list represented?

It is important to be clear that such problems refer to a particular conception of encoding specificity, and not to the idea itself. Indeed, in one guise or another, encoding specificity has become essential to contemporary interpretations of episodic memory.

See also: Elaboration in Memory; False Memories, Psychology of; Implicit Memory, Cognitive Psychology of; Memory: Levels of Processing; Memory: Organization and Recall; Memory Retrieval; Recognition Memory, Psychology of

Bibliography

Abernathy E M 1940 The effect of changed environmental conditions upon the results of college examinations. *Journal of Psychology* **10**: 293–301

Anderson R C, Pichert J W 1978 Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior* **17**: 1–12

Angleton J P, Waskett L 1999 The ability of odours to serve as state-dependent cues for real world mnemonics: can Viking smells aid the recall of Viking experiences? *British Journal of Psychology* **90**: 1–7

Barclay J R, Bransford J D, Franks J J, McCarrell N S, Nitsch K 1974 Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior* **13**: 471–81

Cann A, Ross D A 1989 Olfactory stimuli as context cues in human memory. *American Journal of Psychology* **102**: 91–102

Conrad R 1964 Acoustic confusions in immediate memory. *British Journal of Psychology* **55**: 75–84

Craik F I M, Lockhart R S 1972 Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* **11**: 671–84

Davies G M, Thomson D M (eds.) 1988 *Memory in Context: Context in Memory*. Wiley, Chichester, UK

Fisher R P, Craik F I M 1977 Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory* **3**: 701–11

Godden D, Baddeley A D 1975 Context-dependent memory in two natural experiments: On land and under water. *British Journal of Psychology* **66**: 325–31

Light L L, Carter-Sobell L 1970 Effects of changed semantic contexts on recognition memory. *Journal of Verbal Learning and Verbal Behavior* **9**: 1–11

Morris C D, Bransford J D, Franks J J 1977 Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior* **16**: 519–33

Saufley W H, Otaka S R, Bavaresco J L 1985 Context effects: Classroom tests and context independence. *Memory and Cognition* **13**: 522–8

Schacter D L, Eich J E, Tulving E 1978 Richard Semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior* **17**: 721–43

Semon R 1921 *The Mneme*. Allen and Unwin, London

Smith S M, Glenberg A M, Bjork R A 1978 Environmental context and human memory. *Memory and Cognition* **6**: 342–53

Thomson D M, Tulving E 1970 Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology* **86**: 255–62

Tulving E 1972 Episodic and semantic memory. In: Tulving E, Donaldson W (eds.) *Organization of Memory*. Academic Press, New York

Tulving E 1983 *Elements of Episodic Memory*. Oxford University Press, New York

Tulving E, Osler S 1968 Effectiveness of retrieval cues in memory for words. *Journal of Experimental Psychology* **77**: 593–601

Tulving E, Thomson D M 1971 Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology* **87**: 116–24

Tulving E, Thomson D M 1973 Encoding specificity and retrieval processes in episodic memory. *Psychological Review* **80**: 352–73

Watkins M J 1978 Theoretical issues. In: Gruneberg M M, Morris P E (eds.) *Aspects of Memory*. Methuen, London

Watkins M J, Ho E, Tulving E 1976 Context effects in recognition memory for faces. *Journal of Verbal Learning and Verbal Behavior* **15**: 505–17

Watkins M J, Schiano D J 1982 Chromatic imaging: An effect of mental colouring on recognition memory. *Canadian Journal of Psychology* **36**: 291–99

M. Watkins

Encyclopedias, Handbooks, and Dictionaries

According to one informed estimate, if a copy of every important reference work known in Western history were set side by side in chronological order, some three kilometers of shelf space would be required for their display. Ever since Plato and Aristotle first called for the assembling of what they named 'encyclopedia,' a challenge quickly accepted by one of Plato's nephews, Speusippus (d. 339 BC), intellectuals have tried to summarize and codify what they knew, either for general literate audiences or for specialized groups of colleagues. Of these three kilometers of systematized learning, those volumes that treat the social sciences would appear in only the last few hundred meters of material. This reflects both the fact that social science was not part of the trivium/quadrivium arrangement of classical and renaissance learning, and that official academic status was accorded these fields only during the twentieth century in most research universities.

The history of encyclopedias, handbooks, and dictionaries (designations which until at least the eighteenth century were used interchangeably) is distinguished not only by sheer quantity of printed material, but also by the quality of thinkers and writers who created and contributed to them. Pliny the Elder's *Natural History*, for instance, served as a model of organizing such material for 1700 years after its appearance in 77 AD, despite an odd mixing of fact and fancy throughout its 37 books. Francis Bacon devised in 1620 a huge encyclopedia of 130 sections which was never completed, but his mode of organizing knowledge was still consulted by Diderot and later compilers 130 years later. In 1817 Samuel Taylor Coleridge also put forth an ambitious reorganizing of all human knowledge, which he interestingly titled *Encyclopædia Metropolitana*.

The urge, then, to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West. Even in 551 AD when the love of knowledge and external conditions for its preservation reached an apparent nadir, Cassiodorus saw to it that the labors of the monastery he founded at Vivarium be dedicated principally to the preservation of worthwhile information, producing the two-volume *Institutes of Divine*

and Secular Literature. Vincent of Beauvois has long been identified as the finest encyclopedist of the Middle Ages, producing his *Speculum Majus* (*The Greater Mirror*) in 1244, made up of 10,000 chapters in 80 books, an achievement which held the field for sheer scope until the eighteenth century.

Alphabetical as opposed to categorically arranged encyclopedias began to appear at the end of the first millenium, although most have followed the latter rather than the former method of organization. The first grouping of all knowledge within a format that is recognizably modern, along alphabetical lines, was the *Grand Dictionnaire Historique* produced by Louis Moreri in 1674. The audience for such a work was already developed, for his venture went through six editions in the following 17 years, followed by still more editions in English.

Were Moreri and his contemporary compilers alive today, they would probably be astonished at the level of continuing interest in reference sources. Informed academic reference librarians and professional bibliographers now claim that we are living in a golden era of reference materials, especially if one uses sheer quantity of output as the vital measure of the gold standard. One director of acquisitions wrote of 'so many choices, so little money' regarding new reference books (Wilkinson 1997), and the editor of *The Book Review* of *The Library Journal* analyzed 'reference for a new age' with particular attention to the mesh of electronic and printed materials suddenly dominating publishers' catalogs (Hoffert 1996). A third expert notes that flashy graphics and other user-friendly alterations have attracted many first-time users to reference materials who before would have shunned their perceived complexity (Mantell 1997). Even a cursory look at the current availability of dictionaries, handbooks, or encyclopedias in the social sciences indicates that in some fields, for which before 1990 virtually no reference books (let alone CD-ROMs) were available, now have at their disposal a half-dozen fresh entries in the genre. And even at considerable cost, they are being bought by research libraries almost as a matter of course, just as the electronic forms have begun to encroach on printed versions. The latter may continue to exist into the twenty-first century, but will probably be relegated to the status of secondary resources, subordinate to the newer formats. Or they might persist as specialized bodies of information for ever smaller audiences. Overall, it is fair to say that the reference materials 'business' is thriving as never before, especially when measured by library acquisition budgets aimed in its direction, and the ultimate effect this outpouring may have on the way social science is conceived and used remains difficult to predict.

That said, however, no-one seems to be arguing that the latest entries into the reference book market are supplanting in quality of information or breadth of coverage some of the old standards. *The Encyclopedia of the Social Sciences* from the early 1930s will never be rendered entirely obsolete simply because so many of its contributors have themselves become the subjects of articles in later works. What was true in 1910 of the Eleventh Edition of the *Encyclopaedia Britannica*, still the nonpareil standard of reference works in English, is similarly the case for a range of earlier works. When authors close to the origin of a given field of research accepted invitations by reference publishers to write about their own scholarly innovations (e.g., Werner Sombart on 'Capitalism' in the *ESS*, Benedetto Croce's article on 'Aesthetics' or Edmund Husserl's on 'Phenomenology,' both in the 14th edition of the *Britannica*), later authors have been obliged to accept such statements as first-order expert testaments, beginning their reformulations from there. Surely some of this is going on today, for instance, in newer fields like 'cognitive neuroscience,' 'psycholinguistics,' 'psychobiology,' or 'economic psychology,' and a century hence these may well be accepted as foundational statements. But, for the time being, it is hard to distinguish the ephemeral from the more or less permanent when evaluating the raft of reference works that appeared during the last decade or so of the twentieth century.

It was not always so difficult to name the leading works to which scholars or novices could turn when searching for basic information regarding the social sciences, especially if one strays beyond the confines of the English language. There was a time, of course, when dictionaries, handbooks, and encyclopedias (which in Greek means 'general education') could have been viewed as a special genre of scholarly display, with Samuel Johnson's *Dictionary* (1755), Diderot's *Encyclopedia* (1751), or Voltaire's *Philosophical Dictionary* (1764) initially setting the tone for many modern versions that followed. As noted above, for some centuries prior to the Enlightenment, the words 'dictionary' and 'encyclopedia' were used synonymously, lacking precise meaning, e.g., when Rabelais around 1534 introduced the word into French literature, in *Pantagruel*: 'he has opened for me the true well and abyss of encylopedic knowledge' (*The Complete Works of François Rabelais*, University of California Press, 1991, p. 201). Though Rabelais meant simply 'knowledge' when using the term, within only 25 years a German, Paul Scalich, had published his *Encyclopaedia*, the scope of which conformed to today's usage. Since then, of course, untold thousands of such works have appeared in every major language.

But, unlike scholarly articles and books, rolling off the presses with merciless regularity, substantial reference works have until recently been rarely enough published that they eluded the standard procedure of critical appraisal brought to bear on most nonreference works. And they are hardly the innocent, bland summaries that the naive reader might take them to be. Each new entry into this crowded market must either claim to bring together for the first time a body

of knowledge which to that point had not been assembled by quite the same rules as those which constitute the 'new' rubric, or it must offer some other, less lofty advantage, such as lower price or ease of use. All of this is currently going in the world of social science publishing, and at an unprecedented rate, mostly owing, one suspects, to the huge composing advantages supplied by personal computers.

The positive result of the new technology lies in the superabundance of suddenly available reference resources, some of which will surely define for the first time entire realms of intellectual inquiry. A less sanguine prospect, however, is the burden this torrent of innovative materials places on those who must evaluate or use them. As the range of choices widens, the provision of useable criticism becomes ever more costly, and probably less reliable.

This was not always the case, at least regarding the social sciences. A fairly limited selection of well-known reference tools were at the disposal of researchers in the late nineteenth and early twentieth centuries, which gave this growing general zone of study a tidy coherence that today is all but gone. And whereas such an orderly presentation of knowledge, then common to a given discipline or subdiscipline, might today seem either unnecessary or epistemologically indefensible, for the founders of the social sciences (*Geisteswissenschaften*) over one hundred years ago, the task of proving to more established disciplines that intellectual coherence indeed existed within their upstart fields became a truly vital effort at self-preservation.

Within what we could now broadly call sociology, for example, systematizations appeared fairly early. This is interesting for a number of reasons, not least of which is the fact that sociology as a discipline was defined and taught (particularly in the USA) some time before its sister disciplines within the social sciences. It often played the role of *avant garde* as this new way of carving up the academic world began rearranging disciplinary boundaries. While many of today's practitioners are familiar, at least with the titles, of the *Encyclopedia of the Social Sciences* and *The International Encyclopedia of the Social Sciences*, or Robert E. L. Faris's *Handbook of Modern Sociology* (1964) or Neil Smelser's *Handbook of Sociology* (1988), most are not aware of this genre's substantial tradition, especially as produced in Europe. It is now commonly supposed that Faris's was the first handbook of American sociology to be published, since it was widely distributed at the time and was fortunate in appearing just as sociology became a 'growth industry' on campuses worldwide. But even some reviewers at the time knew better, e.g., Joseph Gittler's opening line in his review: 'With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one' (Gittler 1965, p. 335). But it must have seemed like much less of a 'milestone in the field of sociology' as the publisher had characterized the *Handbook* through its mailed advertisements to ASA members if one considers the substance of the reviews it received. The book was probably taken as a handy cribbing device for graduate students and young instructors, glanced at perfunctorily, then parked on the shelf with similar volumes which Rand McNally had become good at marketing.

As Gittler was hinting, the history of sophisticated handbooks that sought to introduce audiences to new trends of analysis is in fact long and distinguished, though it is quite true that only a few of them were in English. Carl von Rotteck and Carl Welcker edited the *Staats-Lexikon oder Encyclopädie der Staatswissenschaften* (1834–43), the oldest the author has found, adding up to fifteen volumes at 800 pages each. (Even though the social sciences hardly existed at that time, formally speaking, the subject matter in the *Encyclopädie*, as well as the other sources named below, included much of what we would now recognize as such.) In 1852 and 1853, Charles Coquelin oversaw production of the *Dictionnaire de L'économie Politique* in two volumes (1858 pages). Such works 'came dear,' as Dickens would have put it. The copy the author examined bore an inscription from September 1853, noting that the price paid in Piccadilly was £3.80, a very substantial amount of money at the time. Male manual workers in England then earned on average 9s.6d per week (or about £2 per month), with child laborers making perhaps 2s. per week, and domestic servants settling for much less (as per Caird's survey of 1851, reported in Burnett 1984, p. 29). Considering that the price of bread was around 1s. or more per loaf, or 11% of a week's salary, these early reference works would cost in today's terms the equivalent of two months gross earnings, or many thousands of dollars each, that is, if one were to apply every pence earned to the cost of the books! Meanwhile in Italy, Gerolamo Boccardo brought out *Dizionario Della Economnia Politica e Del Commercio: Cosi Teorico Come Practico* (1857) in four volumes and 2800 pages, an equally lavish item.

The predecessor to Weber's famous *Grundriss der Sozialökonomik* was Gustav Schoenberg's *Handbuch der Politischen Oekonomie* (1882–85) in three volumes. Surely the most important of these early handbooks was edited by Conrad et al., *Handwörterbuch der Staatswissenschaften* (1895), an enormous work in seven volumes and about 6000 pages. Supplements were added in 1895 and 1897 of 2000 pages, and later editions (1898 and 1909–11) included two versions of Weber's famous *Agrarverhältnisse im Altertum* (first translated in 1976 as *The Agrarian Sociology of Ancient Civilizations*), the initial version 27 pages long, the second 136 pages. Ludwig Elster, part of Conrad's team, published a shorter version of the monumental work in only two volumes (1898), merely 2100 pages in length. The French were not idle, with Léon Say and Joseph Chailley's *Nouveau Dictionnaire d'Economie*

Politique (1891/92), amounting to two fat volumes of 2800 pages, and a supplement in 1900 of 271 more. Finally, the famous Robert Harry Inglis Palgrave edited *Dictionary of Political Economy* (1894, 1901, 1910), which in its last version came to three volumes and 2500 pages. This was soon followed in 1897 by William D. P. Bliss' *Encyclopedia of Social Reform*, one volume of 1439 pages (1908). What all this shows is that the notion of social science compendia has been around quite a while, and that most of the nineteenth-century European models were enormous publishing achievements. With the new century, ambitions and costs both shrank considerably.

Although in the space allotted here it is impossible to offer even a succinct history of social science reference books, some sense of how they evolved over the last hundred years or so might be conveyed simply by means of mentioning some titles which seemed to define a given discipline at a particular moment in its development. I will begin with my own field, whose history is more familiar to me than the related disciplines.

1. *Sociology*

Despite its senior status as a university discipline of independent identity relative to most of the other social sciences, Sociology has not generated as many references works as one might therefore expect. Thomas Nixon Carver's *Sociology and Social Progress: A Handbook for Students of Sociology* (Ginn, 1906; 810 pp.) was remarkable for its earliness. Yet the first real encyclopedic effort in English came only in 1992 under Edward Borgatta's editorship (Macmillan, 4 vols.), a full century after academic departments bearing the discipline's name appeared in several midwestern universities. (Two previous efforts, one edited by the noted scholar Michael Mann, did not live up to their titles as 'encyclopedias' proper.) Other useful compendia include Smelser and Swedberg's *The Handbook of Economic Sociology* (1994) and Frank Magill's *International Encyclopedia of Sociology* (1995, 2 Vols.).

The first handbook expressly designed for sociology was Alfred Vierkandt's *Handwörterbuch der Soziologie* (1931), a mere 690 pages long, followed in 1956 by Werner Ziegenfuss' *Handbuch der Soziologie* at double the length. And in 1962 Rene König edited the famous *Handbuch der Empirischen Sozialforschung* in two volumes of 2000 pages. In the USA one could begin with *An Introduction to Sociology* edited by Jerome Davis and H. E. Barnes (1927, 1931), a 900-page collaboration by seven authors. *Trends in American Sociology* followed in 1929, edited by George Lundberg, Read Bain, and Nels Anderson, with substantial chapters on the history of sociology, theory, social psychology, culture, rural sociology, urban sociology, educational sociology, social work, applied sociology, and methods. In 1937 Prentice-Hall sold *Man and Society* for $3.75, a 787-page work edited by Emerson Schmidt that tried to introduce students to all the social sciences simultaneously, covering everything from human values to income distribution to social psychology (Herbert Blumer's memorable chapter). The Gurvitch–Moore volume, *Twentieth Century Sociology*, published just following World War II (1945), seemed to have widespread influence, despite its execrable typos and layout. Its most distinctive feature, in addition to chapters on all the subfields of sociology at the time, was a second section of 250 pages that dealt with sociology in France, the USA, Britain, Germany, Latin America, Italy, Spain, Russia, Poland, Czechoslovakia, Romania, and Yugoslavia. This should have inspired later handbooks, and may account for the volume's popularity, since a distinctly ecumenical flavor swept the academy worldwide following WWII. Its authors Burgess, Parsons, H. P. Becker, Sorokin, MacIver, Znaniecki, Gurvitch, Roscoe Pound, Merton, Wach, Levi-Strauss, Faris, Albert Salomon, Bastide, and so on comprise a list of distinguished scholars that has not been matched since in a collaborative volume. The first recognizably 'modern' compendium was Joseph Gittler's *Review of Sociology* (1957), with chapters by recognized authors on the topics that soon became the staples of introductory textbooks. (Georges Gurvitch edited *Traité de Sociologie* in 1960, two volumes and 1000 pages, which seemed to have little impact in the USA.) And finally, Parsons himself had a go with his *American Sociology* (1968), with chapters by many scholars now in the twilight of their careers but who then were the bedrock of the mainstream, as it were.

One could say many things along these lines about the Smelser *Handbook*, which, given its size and unique position within sociology at present, deserves more than passing attention. This is especially so regarding the 'theories' that are put forth in the opening chapters as constituting 'the' new streams. For instance, neither Walter Wallace nor Jeffrey Alexander, the chapters' authors, wishes to consider postmodernism as a fact of existential and scholarly life, and therefore each throws himself into a strange time warp. Peterson had characterized Faris' introduction to his earlier compendium as 'blindly benign' (Peterson 1965, p. 254), which seems perfectly correct, even when rereading Faris today. One could say something similar about the theoretical terrain that is gingerly walked over in opening chapters of the Smelser *Handbook*. A range of old beliefs, now diminished by criticism from philosophers and literary critics—some of whom, incidentally, number among our most influential 'social' theorists—are trotted out as if nothing substantial had occurred to Western culture since 1964. It is as if the 'self' about which Parsons wrote with such conviction is still out there doing what selves have 'always' done. The same could be said about the precritical notion of 'science' and 'rationality' that figures in both chapters,

but especially in Alexander's, as if definitions of rational action and societal rationality are not now in themselves enormously contentious topics. Such limitations begin to indicate the intellectual boundaries of compendia of this sort, even those written by top scholars, for the adventurousness which might pervade a monograph composed by the same authors is diminished when they turn their hands to 'general' statements. Since reference sources, by common consent, are supposed to offer a summary of 'established knowledge' in a specific field or subfield, their authors typically draw back from the perimeters of discovery and innovation. Therefore, such works almost always lag behind the leading edges of thinking and offer their users a snapshot of what was confidently held to be true a dozen or more years before they are published. And with the ever-increasing rate of change in global intellectual culture, this poses a problem, particularly for printed reference sources, which may be met only via constantly updated electronic versions.

Comparison of Smelser's *Handbook*, at the time of writing the most recent entry into the genre, with earlier versions must remain fairly rough. Its predecessors employed other sorts of authors, aimed at different audiences (those not yet effectively lobotomized by videocretinization), and seem to have been composed in an atmosphere of genuine wonder and hopeful discovery regarding what social scientific study might yield for intellectuals and for the social order that supported them. This kind of childlike hope decomposed after the Gurvitch–Moore volume in 1945 and was replaced by a somewhat defensive, stuffy scientism in later compendia that reflected sociology's struggle to survive McCarthyism on one hand, and literary criticism on the other. But such books, while high on scientific rhetoric, seemed correspondingly low on political self-examination and savvy, while also revealing a foreshortened understanding of the philosophical problems necessarily part of sociological study. Still, it is uplifting to read these older books because their authors were usually straightforward and unpretentious. They tended to avoid enormous claims for the discipline, and offered up its fruits with a direct and unembattled character that nowadays is harder to find. It is probably no accident that after years of sociology bashing at the highest levels of government, both in the USA and in Europe, the field should have put out a summary package for broad consumption in hope of regaining lost ground, both within the cloister and beyond its high walls. We shall see if it succeeds as part of the long process of recertifying the discipline for a public which has been told repeatedly that nothing good can come of thinking about social reality in the way 'we' do.

Dictionaries for sociology began with Constantine Panunzio's pamphlet, *A Student's Dictionary of Sociological Terms* (California, 1937), continuing with Henry Fairchild's *Dictionary of Sociology* (1944). Only in the late 1960s were a number of competing volumes added to this small shelf of sociological reference books, including those of Geoffrey Mitchell (1968, 1979), George Theodorson and Achilles Theodorson (1969), and Thomas Hoult (1969). With Raymond Boudon and François Bourricaud's *Critical Dictionary of Sociology* (1989), celebrated scholars from Europe entered the field for the first time, followed in the 1990s with new dictionaries by David Jary (1991), Nicholas Abercrombie (1994), Gordon Marshall (1994, 1998), and Allan Johnson (1995), the former three being well-known British scholars. American sociologists have not of late contributed substantially to this genre.

2. *Anthropology*

Of all the social sciences, anthropology has been to date the least well served in terms of general reference books. This seems odd given the nineteenth century's penchant for vast compilations of quasianthropological data, e.g., Frazer's *Golden Bough* or Herbert Spencer's *Descriptive Sociology* (in fact comparative anthropology) in 12 volumes. The earliest dictionary of sociology I have found has now become a standard, Charles Winick's *Dictionary of Anthropology* (1956), with several later editions. *Anthropology A to Z* (1963), edited by Gerhard Heberer, Carleton Coon, et al., was based on an earlier German compilation of several hands. A recent trend among publishers to name reference books after their firm is typified in *The Macmillan Dictionary of Anthropology* (1986) edited by Seymour-Smith. The implication of this general practice—that a source's value to the user is more or less guaranteed by virtue of a firm's prestige—is not necessarily realized in practice. The most recent entry is *Barfield's Dictionary of Anthropology* (1997).

The discipline came into its own so far as reference sources go when Honigmann et al. produced the 1300-page *Handbook of Social and Cultural Anthropology* (1973) at a period when the social sciences reached a zenith of appeal among college undergraduates in the USA. Two misnamed volumes are Hunter and Whitten's *Encyclopedia of Anthropology* (1976) and Barnard and Spencer's *Encyclopedia of Social and Cultural Anthropology* (1996), both books of modest size. An interesting twist appeared with Bock's *Handbook of Psychological Anthropology* (1994), an instance, ever more common, of a reference book seemingly anticipating an intellectual or disciplinary trend before it gains acceptance at the institutional level. Such a rubric might well surprise both psychologists and anthropologists, relatively few of whom typically relate their two fields in any intimate way.

3. *Cognitive Sciences and Related Fields*

As soon as it became apparent in the 1970s to policymakers, governmental budget specialists, and

academics that improvements in psychopathic or sociopathic conditions could be more cheaply and quickly realized through the use of drugs and related somatic measures—as opposed to 'talking cures' or other socially based interventions—those fields that stress the linkage between individual physiology and the social order experienced a sudden renaissance. What seemed a neo-Darwinian rage for simplification overtook large zones of inquiry within the social and behavioral sciences. This has become so much the case during the 1990s that physiological reductionism often seems the order of the day, nearly extinguishing the behavioral component in explanations for an ever broadening range of ailments, from alcoholism to schizophrenia to general criminality. Given this intellectual environment, it is hardly surprising that a range of new reference books appeared which positioned themselves along a continuum of biobehavioral explanation of social action, from crude sociobiological reductionism to subtle neurological analysis of brain function and behavior. Much of this new wave of work stirs up controversy because of its real or imagined political consequences, most famously exemplified by the raucous reception given Herrnstein and Murray's *The Bell Curve* (1994), both in the ordinary as well as the academic press.

Among many such books published in the 1980s and 1990s, several might be mentioned for their innovative qualities. Michael Gazzaniga edited *Handbook of Cognitive Neuroscience* (1984), complemented by Dunlop and Fetzer's *Glossary of Cognitive Science* (1993), brought to even grander fruition in Wilson and Keil's *The MIT Encyclopedia of the Cognitive Sciences* (1999). Not unrelated are Wyer and Srull's *Handbook of Social Cognition* (1984, 1994) and Van Hasselt and Hersen's *Handbook of Social Development* (1992). More traditional, and connected with a field the roots of which go back at least to Saussure's posthumous masterpiece, *Course in General Linguistics* (1916), is Florian Coulmas's *The Handbook of Sociolinguistics* (1997), and what amounts conceptually to its complementary volume, Morton Gernsbacher's *Handbook of Psycholinguistics* (1994). These interestingly hybrid fields will undoubtedly grow in the twenty-first century, and it seems now that a biosocial focus will stimulate at least as much theorizing and research as did the more purely sociopolitical analyses during the preceding half-century.

4. Economics

Unlike some of the newcomers mentioned above, economics (or 'political-economy' as it was first known) has a long history relative to other social sciences, dating at least to 1776 and the publication of Smith's *Wealth of Nations*. Reference sources therefore abound. Before economics became an academic area, many guides to financial matters existed, of course, for example *The Hand-book of Trade and Commerce, or, A concise dictionary of the terms and principles of trade, commerce, manufactures, commercial and common law, etc, with tables of money, weights, and measures*, published in London in 1840. We also know from sources like Marx's *Theories of Surplus Value*, his three-volume anthology of annotated excerpts from the history of economic theory, that the literature on economic matters was by the 1850s quite well developed. Reference works, however, did not apparently surface until somewhat later. Palgrave's *Dictionary of Political Economy* (1894) is perhaps the most famous, and John Joseph Lalor's *Cyclopedia of Political Science, Political Economy, and the Political History of the United States: By the Best American and European Writers* (1881–1884) is another early example. So was Frank Bower's *A Dictionary of Economic Terms for the Use of Newspaper Readers and Students* (1905), kept in print until 1951 under subsequent editors.

More obviously pertinent to the discipline as it now exists was Herwith Price's *Economic Dictionary* (2 volumes, 1926–29), published in Berlin, as well as John Todd's *The Mechanism of Exchange: A Handbook of Economics* (3rd edn., 1927) (which became *The Science of Prices* in a 4th edition [1935] before reverting to its orginal title in its 5th). Frank Bower's brief *Dictionary of Economic Terms* reached its 10th edition in 1936, while Horton's *Dictionary of Modern Economics* appeared in 1948. Beginning in the early 1950s, the number and variety of dictionaries and handbooks concerning economics exploded, keeping pace with undergraduate enrollments at large public universities in the USA and elsewhere. Sloan and Zurcher's went through five editions between 1951 and 1970, while J. L. Hanson's reached its 6th edition in 1986, having been in print for 21 years. The latest dictionaries of economics include those by John Black (Oxford, 1997) and Bannock et al. (Wiley, 1998). On a larger scale, encyclopedias of economics began to appear relatively late, with the *Guilford Encyclopdia of Economics* in 1973/74, the *McGraw-Hill Encyclopedia of Economics* in 1982 and 1994 (2nd edn.), and *The Fortune Encyclopedia of Economics* in 1993. O'Hara has reverted in his choice of title to an earlier time, perhaps for political as much as intellectual reasons, with *Encyclopedia of Political Economy* in two volumes (1999). Another new entry is K. Liou's *Handbook of Economic Development* (1998), which broadens the discipline's terrain by including the global dimension.

5. Political Science

One of the earliest general reference works in this area is *Political Dictionary, forming a work of universal reference, both constititutional and legal, and embracing the terms of civil administration, of political*

economy and social relations, and of all the more important statistical departments of finance and commerce (London, 1845–46) in two volumes. It is instructive that those fields which in the mid-nineteenth century seemed closely enough linked, empirically and practically, to be treated within a single work, have since the mid-twentieth century been torn asunder into mutually incompatible and incomprehensible camps of specialists, who have little to say to one another except at ceremonial occasions. This situation is not necessarily an advance in social knowledge. The 14-volume *American Annual Cyclopaedia and Register of Important Events … : Embracing political, civil, military, and social affairs* (New York, 1862–1875) does not take in political science *per se*, though it would surely have uses to the historian of the discipline. Already mentioned, John Lalor's 1882 *Cyclopedia of Political Science […]* is among the first to use the term in the modern sense. Another amalgamated volume that typifies an era unafraid to mix disciplines was William Bliss' *Encyclopedia of Social Reform: Including political economy, political science, sociology and statistics* (1897), some 1439 pages in length.

Dictionaries and lexicons of political science are as numerous as encyclopedias or handbooks are scarce. William Connolly's *Terms of Political Discourse* (1974) and Maurice Cranston's *Glossary of Political Terms* (1966) were produced by scholars who subsequently became well-known for other work. Joseph Dunner's *Dictionary of Political Science* (1964), Jack Plano's *Political Science Dictionary* (1973), and Roger Scruton's *A Dictionary of Political Thought* (1982) are several among a score of similar works, which, as in neighboring fields, have been increasing in number, length, and complexity. Works of larger scope, in terms of length as well as conceptual breadth, include David Miller's *The Blackwell Encyclopaedia of Political Thought* (1987, 1991), Hawkesworth and Kogan's *Encyclopedia of Government and Politics* (1992, 2 vols.), and Goodin and Klingemann's *A New Handbook of Political Science* (1996).

6. *Psychology*

Just as the number of undergraduate majors in psychology typically overwhelms those of all the other social sciences, with ratios often running five or six to one or higher, so with reference books attached to this enormous field. The most distinguished early survey was surely William James's *Principles of Psychology* (1890, 2 Vols.), in print continuously ever since, in full and abridged versions. Clark Murray's *Handbook of Psychology* (1885, 1888), James Sully's *Teacher's Handbook of Psychology* (1886, 1909, 5th edn.), and the *Handbook of Psychology* (1889–91, 2 Vols.) by William James's equally famous contemporary, James Mark Baldwin, were each substantial compilations, and helped give psychology a head start toward producing a unified view of the social world vis-à-vis its sister disciplines. More eclectic viewpoints were expressed in Geoffrey Rhodes's *The Mind at Work: A Handbook of Applied Psychology* (1914) (with help from the noted sociologist, L.L. Bernard) and in Leander Keyser's *A Handbook of Christian Psychology* (1928, 2nd edn.). By 1931, however, the field had become thoroughly professionalized, as exemplified by Anderson, Buhler, and A. Freud, et al., *A Handbook of Child Psychology* (1933, 2nd edn.), and its companion volumes in the Clark University series, H. Banister et al., *A Handbook of General Experimental Psychology* (1934, 1125 pp.) as well as Allee, Allport, et al., *A Handbook of Social Psychology* (1935, 1995 pp.). Somewhat less innovative were W. B. Pillsbury's *Handbook of General Psychology* (1942), Kimball Young's *Handbook of Social Psychology* (1946), or Lindner and Selger's *Handbook of Correctional Psychology* (1947). Specialization within the discipline had already reached such a pitch that in 1950 Norman Munn offered *Handbook of Psychological Research on the Rat* (first published in 1933 as *An Introduction to Animal Psychology*).

The 1950s were a boom period for psychology reference works of the larger scale, reflecting the discipline's ever-widening influence in Western societies, with at least six more 'handbooks' entering the market. In more recent times innovation has once again become commonplace, with books like M. Gazzaniga's *Handbook of Psychobiology* (1975), Raaig, Veldhoven, et al., *Handbook of Economic Psychology* (1988), and Crawford and Krebs's *Sociobiology and Psychology* (1987), which later became the *Handbook of Evolutionary Psychology* (1998). The last volume intimates a strong connection currently being forged between researchers with refreshed enthusiasm for neo-evolutionary doctrines on the one hand, and sociobiological explanations for behavior on the other. This general tendency—in some ways a repudiation of earlier causal models for social action and psychological states—has been given substantial impetus by the Human Genome Project and related studies that attempt to link physiology with behavior in an unmediated fashion.

Encyclopedias have also been a more regular feature of psychology's disciplinary life than in other fields. Early efforts included McKean's *The Encyclopedia of Psychology: A series of graphic descriptions of human characteristics …* (Brussels, 1928) and Philip Harriman et al., *Encyclopedia of Psychology* (1946). The 1970s exploded with new versions varying substantially in size and scope. Robert Goldenson's *Encyclopedia of Human Behavior: Psychology, Psychiatry, and Mental Health* (1970, 2 vols.) brought together supplementary fields, just preceding the herculean effort by Benjamin Wolman et al. (300 editors and 1500 authors), *International Encyclopedia of Psychiatry, Psychology, Psychoanalysis and Neur-*

ology which took up 12 initial volumes (1977) plus a supplement (1983) and surely set a new standard for comprehensiveness of these allied research areas. Wolman also edited a one-volume *Encyclopedia of Psychiatry, Psychology, and Psychoanalysis* in 1996. Another half-dozen encyclopedias of psychology proper appeared during the late 1970s and 1980s, that of Raymond Corsini's running to four volumes (Wiley, 1984); Frank Magill's reference book production team brought out the *Survey of Social Science: Psychology Series* (1993) in six volumes, grandly renamed the *International Encyclopedia of Psychology* when reissued and reset in 1996. (Magill and his specialty editors have produced like volumes for all the social sciences in his *Survey* series.)

Ever since James Mark Baldwin published *Dictionary of Philosophy and Psychology* (1901–05, 3 Vols.; new edn., 1928), there has never been a shortage of dictionaries for the discipline. The author has identified 35 in English alone, some of them of continuing interest simply for the peculiarity of their design, at least by today's standards, e.g., William Taylor's *Taylor's Dictionary of Bio-Psychology, Genetics, and Various Branches of Applied Psychology* (1925, 148 pp.), or Vergilius Ferm's *A Dictionary of Pastoral Psychology* (1955). Of the more standard fare—a score or more appear in most standard bibliographies—many were compiled by the same scholars who oversaw production of the encyclopedias named above, with the most lavish entries all appearing during the 1990s.

Bibliography

Bliss W D P (ed.) 1897 *Encyclopedia of Social Reform*. Funk & Wagnalls, New York

Boccardo G (ed.) 1857 *Dizionario Della Econòmia Politica e Del Commercio*. Franco e Figli, Torino

Burnett J 1984 *Useful Toil: Autobiographies of Working People from the 1820s to the 1920s*. Penguin, New York

Chinoy E 1966 Review of Faris' Handbook of Modern Sociology. *American Sociological Review* **31**: 270–1

Conrad et al. 1895 *Handwörterbuch der Staatswissenschaften*. Gustav Fischer, Jena

Coquelin C (ed.) 1852/53 *Dictionaire de l'Economie Politique*. Guillaumin, Paris

Cupp W L 1959 Review of Merton et al. *Sociology Today*. *Social Forces* **38**: 77–8

Davis J, Barnes H E (eds.) 1927 *An Introduction to Sociology*. Heath, Boston

Elster L (ed.) 1898 *Worterbuch der VoIkswirtschaft*. Fischer, Jena

Faris R E L (ed.) 1964 *Handbook of Modern Sociology*. Rand McNally, Chicago

Gittler J (ed.) 1957 *Review of Sociology: Analysis of a Decade*. Wiley, New York

Gittler J 1965 Review of Rubert Faris' *Handbook of Modern Sociology*. *American Journal of Sociology* **71**: 335–6

Gurvitch G 1960 *Traite de Sociologie*. Presses Universitaires de France, Paris

Gurvitch G, Moore W E (eds.) 1945 *Twentieth Century Sociology*. Philosophical Library, New York

Hoffert B 1996 Reinventing reference for a new age. *Reference 97*: Supplement to *Library Journal* (November 15): S4–7

House F 1959 Review of Merton *et al.*, *Sociology Today*. *American Journal of Sociology* **45**: 211

König R (ed.) 1962 *Handbuch der Empirischen Sozialforschung*. Enke, Stuttgart, Germany

Lundberg G, Bain R, Anderson N (eds.) 1929 *Trends in American Sociology*. Harper and Brothers, New York

Mantell S 1997 Looking it up is looking up. *Publishers Weekly* **244**(September 29): 51–3

Merton R K et al. (eds.) 1959 *Sociology Today*. Basic Books, New York

Moore W E 1959 The whole state of sociology. *American Sociological Review* **24**: 715–8

Palgrave R H L (ed.) 1894 *Dictionary of Political Economy*. Macmillan, London

Parsons T (ed.) 1968 *American Sociology*. Basic Books, New York

Peterson R 1965 Review of Faris' *Handbook of Modern Sociology*. *Social Forces* **44**: 253–4

Say L, Chailley J (eds.) 1891/92 *Nouveau Dictionnaire d'Economie Politique*. Alcan, Paris

Schmidt E (ed.) 1937 *Man and Society*. Prentice-Hall, New York; Laupp, Tübingen, Germany

Vierkandt A (ed.) 1931 *Handwörterbuch der Soziologie*. Enke, Stuttgart, Germany

von Rotteck C, Welcker C (eds.) 1834–1843 *Staats-Lexikon oder Encyclopedie der Staatswissenschaften*. Hammerich, Ultona

Wilkinson F E 1997 Reference materials: so many choices, so little money. Librarians and publishers speak out! *Against the Grain* **9**(4 September): 1, 18, 20, 47, 167

Ziegenfuss W (ed.) 1956 *Handbuch der Soziologie*. Enke, Stuttgart, Germany

A. Sica

Endocrinology and Psychiatry

The term '*psychiatrie endocrinienne*' ('psychiatric endocrinology') was first coined by Laignel-Lavastine in 1908 at a psychiatric congress in Dijon. Since then hormones always have been viewed as magic bullets for the treatment of a variety of psychiatric conditions and even Sigmund Freud predicted that hormone treatment might one day replace psychoanalysis. The first systematic research in this field was conducted by the Swiss psychiatrist Manfred Bleuler, who attempted to establish a connection between psychopathology and the secretory activity of endocrine glands. Influenced by his mentor, the neurosurgeon Harvey Cushing from Boston, he first directed his interest toward the effect of pituitary-adrenocortical hormones that are excessively secreted in patients with an autonomous adenoma at the pituitary gland. This adenoma produces corticotropin in high amounts unrestrained by the high plasma cortisol concentration that would normally suppress corticotropin release

through negative feedback action. Bleuler defined the 'endocrine psychosyndrome' and his descriptions of the behavioral changes associated with endocrine dysfunction are still valid today.

This article illustrates how Bleuler's concept has been translated into hypothesis-driven research and how that has ultimately led to a novel pharmacology that holds the promise of better antidepressant treatments. In addition, some examples are given showing how peripherally monitored changes in hormonal secretion can be viewed as the readout of disturbed central neurobiology. Finally, the current status of several selected hormonal treatments is examined.

1. *Stress Hormones*

The majority of patients with Cushing's syndrome, whether this is due to high plasma corticosteroid levels caused by ectopic secretion or high-dosed corticosteroid treatments, develop mood disorders, some with psychotic features. Only in very rare instances do manic or schizophrenic syndromes prevail. Apart from these disturbances in mood and behavior, cognitive impairments are also observed. Cortisol was suspected to account for neuronal loss and damage because the hippocampus, a brain area critical for cognitive function and richly endowed with receptors to which cortisol binds (corticosteroid receptors), was found to be reduced in size in patients with elevated stress hormone secretion. Based on some simplistic concepts, the central actions of cortisol have been described as detrimental to neuronal viability. This impairment, however, is only true when extremely high corticosteroid levels are experimentally induced (Almeida et al. 2000). Under such conditions, which probably mimic the clinical situation only very rarely, glucose uptake by neurons could be reduced, thereby making neurons vulnerable to other insults such as hypoxia or increased neuronal firing rates (e.g., during seizures). Such neuronal damage does not occur in situations of stress-elicited hypercortisolism. On the contrary, corticosteroids are essential to coordinate the organism's ability to cope with stress and promote the interpretation and storage of novel information, as well as facilitating the extinction of behavior that is not relevant under the immediate highly demanding situation. Significantly, under certain circumstances corticosteroids can act as neuroprotectants rather than as neurotoxicants.

Much of the current confusion related to the excitement about the phrase 'stress is bad for your brain' stems from the different functions that the two corticosteroid receptor types, the glucocorticoid (GR) and the mineralocorticoid receptor (MR) may exert. As studies by Ron de Kloet and colleagues (1999) in Leiden, the Netherlands, have revealed, MRs play a role in behavioral reactivity during novel situations, whereas GRs are involved in the consolidation of learned information. Moreover, the neuroprotectant effect of naturally occurring corticosteroids is due to their binding at MRs, whereas synthetic corticosteroids, which selectively bind at GRs, render neurons susceptible to cell death. This does not negate the fact that continuous exposure to stress or acute serious trauma may have long-term effects, increasing an individual's risk of developing stress-related diseases such as depression, anxiety, sleep disturbances, or post-traumatic stress disorder.

Table 1
Common changes in stress hormone regulation of patients with depression

- The 24-hour secretory pattern of corticotropin and cortisol shows elevated values
- The urinary-free cortisol content is elevated over a 24-hour period
- Corticotropin-releasing hormone (CRH) is elevated in the cerebrospinal fluid
- Binding of CRH is reduced in the frontal cortex
- CRH is overexpressed in the paraventricular hypothalamus
- ACTH response to CRH is decreased
- ACTH and cortisol response is increased following CRH in dexamethasone-pretreated patients (combined DEX/CRH test)

One possible explanation is the long-term effect induced by stress hormones on gene activity. It is of note that GRs and MRs, once activated by corticosteroids, act as powerful transcription factors that bind at specific DNA sequences to activate or repress transcription of a manifold of genes. Hormone-activated MRs and GRs can also bind to other transcription factors and through this interaction indirectly modify the transcription rate of many genes that are not primarily hormone regulated. Stress-elicited hormone secretion and its consequences on gene activity represents a perfect model for the study of the influence of environmental psychosocial factors on gene expression in the context of the pathophysiology of depression.

While it is well-established that the risk of developing depression depends on a complex multigenetic predisposition, it is an equally well-founded fact that depression can be precipitated by life events. On a functional level, such life events are associated with an increased demand to maintain homeostasis. This condition is associated with increases in stress hormone levels, which exert activity at the level of gene expression by transforming GRs and MRs into transcription factors. Under these demanding conditions subtle changes in the genetic blueprint, e.g., single nucleotide polymorphisms, may no longer be compensated, resulting in the emergence of a clinical phenotype. Given that stress hormones account for such effects by transducing environmental challenges on the cellular and genomic level, it would seem worthwhile

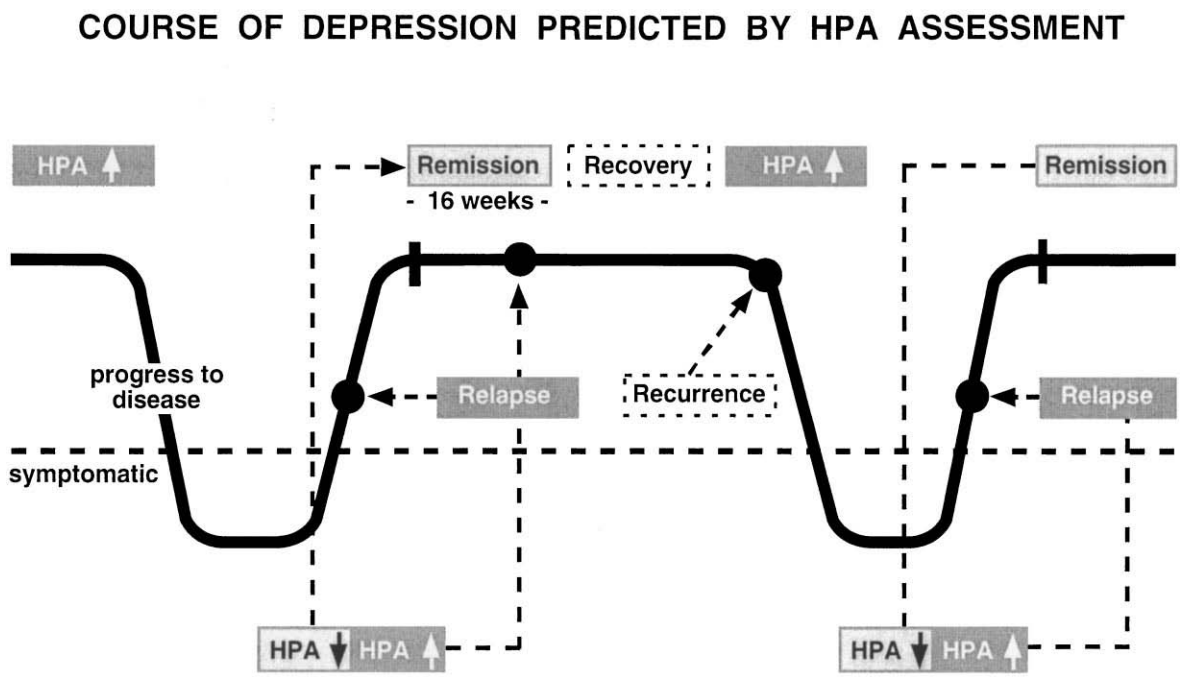

Figure 1
Stress hormones and depressive symptoms

to consider the reuptake options available for intervention. To better understand the rationale for today's pharmacological strategies, the relevant facts about stress hormones are listed in Table 1. It is important to note that elevations of stress hormones or aberrant responses to neuroendocrine probes are not a byproduct of emotional stress associated with depression. As illustrated in Fig. 1, stress hormone abnormalities precede the occurrence of depressive psychopathology, and conversely, a gradual normalization of stress hormone regulation is followed by an improvement in depressive symptoms. This time-course pattern rejects the possibility that abnormal test results are secondary to depression. A more straightforward interpretation links regulation of the stress hormone system to the pathophysiology of depression (Holsboer 1995).

The key regulator of peripheral stress hormones is corticotropin-releasing hormone (CRH), a neuropeptide synthesized mainly in the nucleus paraventricularis of the hypothalamus, an assembly of brain nuclei that serves as a relay between brain function and peripheral autonomous and endocrine regulation (Holsboer 2000). From the hypothalamus, CRH is released into small blood vessels that allow transportation of CRH to the pituitary where it evokes the synthesis and release of corticotropin. This hormone enters the blood circulation and once it arrives at the adrenal cortex, cortisol is secreted, which regulates CRH and corticotropin release through a well-balanced feedback loop. Key elements in this circuit are corticosteroid receptors that when hormone-activated, serve as a break on genes coding for CRH and corticotropin. The corticosteroid receptor hypothesis of depression submits that the signaling through these receptor molecules is impaired, resulting in a number of changes of gene activity, including that coding for CRH. This neuropeptide is not only involved in organizing the stress hormone system—which is traditionally called the 'hypothalamic-pituitary-adrenocortical (HPA) axis, although it involves many other brain areas including those believed to be involved in the neuropathology of depression—but also coordinates a number of behavioral signs and symptoms that aid for adaptation to stress.

In transgenic mice continuous overexpression of CRH, as documented to occur in depression (see Table 1), leads to increased anxiety-like behavior and other symptoms reminiscent of depression (decreased appetite, sleep disturbance, loss of libido, etc.) (Steckler et al. 1999). Similar findings in rats to which CRH was administered also suggest that CRH precipitates behavioral changes compatible with anxiety and the symptoms of depression. Consequently, research has focused on the possibility of reducing these negative effects of CRH by blocking its receptors. After the responsible receptor had been identified as the CRH-R1 subtype, drugs were developed that are capable of

antagonizing the actions of CRH specifically at these receptors. Indeed, in animal studies it has been shown that anxiety-like behavior under stress conditions or under conditions of withdrawal from drug abuse disappears when CRH-R1 antagonists are administered. Moreover, stress-induced sleep disturbances disappeared in rats when the administration of the stressor is combined with the CRH-R1 antagonist. On a neurochemical level it has been shown that CRH overproduction in the brain is associated with blunted serotonin responses to stress, whereas genetic invalidation of CRH-R1 receptors results in increased serotonin release. This neurotransmitter is strongly implicated in the pathophysiology of depression, and most currently prescribed antidepressants such as fluoxetine, paroxetine, and citalopram work on the principle of the enhancement of serotonin release. This favors the hypothesis that blockers of CRH type 1 receptors may also work as antidepressants through the enhancement of serotonergic neurotransmission and the correction of excessive central stress hormone drive. These considerations have led to the first clinical trial of a CRH-R1 antagonist in depression and its results have confirmed that this hypothesis-driven research has led to a novel pharmacotherapeutic concept (Holsboer 1999).

2. *Sex Hormones*

The differences between men and women extend to neuroanatomy, with male brains having a higher volume of cerebrospinal fluid and grey matter. Women's brains are denser in grey matter, consisting of neuronal cell tissue and connection-making dendrites. It is believed that the proportional difference between grey and white matter accounts for the different performances in neuropsychological tests where men demonstrate better spatial skills, but women are more verbally proficient. Moreover, estrogens, women's key sex hormone, secreted from gonadal glands, may explain some of the functional gender differences. These can be studied in postmenopausal women, whose endogenous estrogen production is reduced. Studies using a brain imaging technique have shown that estrogen administration to postmenopausal women can activate a brain region responsible for storing phonologic information, which may explain why women score higher in language tests.

Changes in estrogen secretion have been related to many psychiatric syndromes ranging from schizophrenia to depression and Alzheimer's disease. While the lifetime risk for schizophrenia is equal for men and women, it has been found that the age at which affected men have their first psychotic symptoms is three to four years earlier than in women who, after the menopause, have a second peak of disease onset. The later onset of schizophrenia in women has led to the hypothesis that estrogens protract the expression of schizophrenic symptoms (Häfner et al. 1993). One possible mechanism involved is an estrogen-induced decrease in dopaminergic neurotransmission. This attractive hypothesis still awaits corroboration from therapeutic tests using estrogens in combination with neuroleptics. Further studies will also have to address the contradiction of estrogen-elicited increases in the density of serotonin type 2A receptors in the forebrain and the role of these receptors in the pathogenesis, as well as in the treatment of psychoses. Numerous studies have shown that serotonin type 2A receptor blockade is a fundamental pharmacological requirement for the new drugs termed atypical antipsychotics, such as clozapine or olanzapine. On the other hand, serotonin type 2A receptor density has been found to be diminished in patients with schizophrenia, which would be consonant with decreased estrogen secretion with age.

Estrogens are widely prescribed, mainly by gynecologists and practitioners to alleviate mood swings associated with the premenstrual syndrome (Schmidt et al. 1998). Comparatively little systematic research has been conducted, although most published trials have reported beneficial effects. Whether estrogens are also effective in major depression is controversial, and careful inspection of the available data casts doubt on the possibility that the documented effects of estrogens upon serotonergic (increase of serotonergic 2A receptors in the forebrain), noradrenergic, dopaminergic, and cholinergic systems may translate into a psychotropic action comparable to established antidepressant drugs. The situation is perhaps different in older women where endogenous estrogen production is reduced, or in women where massive changes in estrogen secretion have occurred. Postpartal depression is a typical example of this condition and in these cases estrogen treatment has been reported to be beneficial. Perhaps the most meaningful application of estrogens in women with depression is its use as an adjunct to standard antidepressant treatment. For example, estrogen replacement treatment in combination with fluoxetine was found to be superior to antidepressant monotherapy in women who were 60 years or older. Another study where postmenopausal women with severe major depression were treated with imipramine, and where to those who did not respond either to estradiol or placebo was added as adjunct, the investigators failed to observe an improvement by the hormone. In the light of this, mild dysthymic and depressive symptoms may respond to estrogen supplementation either directly through central neuropharmacological actions or indirectly through the beneficial effects of estrogen on peripheral bodily functions such as libido, hot flushes, skin crawls, rheumatic pains, etc. (Santoro et al. 1999).

Perhaps the most compelling influence of estrogen on brain function is related to memory, cognition, and the apparent protective effect of estrogens with respect to the onset of Alzheimer's dementia. Multiple epi-

demiological studies have suggested that estrogen replacement therapy may protect against, delay the onset of, or slow the cognitive decline associated with Alzheimer's disease. In these studies, women who took estrogen replacement once they reached menopause had a significantly decreased likelihood of developing Alzheimer's disease. A cautionary note, however, seems warranted as this correlation does not necessarily reflect a causal mechanism. In favor of a specific role of estrogens as neuroprotectants are the results of a number of experiments investigating the interaction of these hormones with β-amyloid, which is thought to be causal in the development of cerebral plaques. These are in combination with fibrillary tangles, a characteristic neuropathologic feature of Alzheimer's disease. Estradial was found to protect against β-amyloid-induced cell degeneration and to reduce the production of this β-amyloid. The protective effect does not seem to involve estrogen receptor activation but rather is related to the antioxidative actions of these steroid hormones (Behl and Holsboer 1999). This is a clinically pertinent observation because the decreased risk of developing Alzheimer's disease might be outweighed by the increased risk of, for example, breast cancer (Grodstein et al. 1997). It was shown that the antioxidant effect of estrogens is independent of the hormonal effects via estrogen receptors. This opens up the possibility of developing 'designer estrogens' to provide long-term protection against the common causes of morbidity and mortality such as osteoporosis, coronary heart disease, and probably against cognitive decline associated with Alzheimer's disease.

Male sex hormones have only received scant attention in psychiatry despite their widespread use as anabolic drugs. The best-established effects of male sex hormones, especially of testosterone, are related to their effect on sexual function. Sexual desire, intense sexual feelings, and sexual activity are reduced in males with pathologically decreased testosterone levels, and are restored with testosterone treatment. In young women without sexual dysfunction one single dose of testosterone can induce markedly increased sexual arousal irrespective of sex hormone secretory activity. However, men who are not hypogonadal, i.e., have normal plasma testosterone levels, but a decreased libido and erectile function, do not respond to testosterone administration. In other words, decreased sexual arousal and erectile impotence are not responsive to testosterone in men with normal sex hormone status. That male sex hormones play a role in aggression has been inferred from studies in prisons, in which violent criminals were compared with nonviolent inmates. Many, but not all, of these studies agree that violent criminals have higher testosterone concentrations in both plasma and cerebrospinal fluid.

Perhaps the more important question relates to the psychiatric risk of testosterone administration in dosages used by young men to improve athletic performance or personal appearance ('body builders'). A recent examination of the effects of testosterone on mood and aggressive behavior found notable adverse psychiatric effects in a subgroup of 16 percent of the study population, which was found to exhibit manic symptoms. From animal experiments a large number of interactions of testosterone with neurotransmitters and neuropeptides have emerged. These findings, however, give no hint as to why supraphysiological dosages of testosterone might induce manic symptoms in a fraction of otherwise healthy males. There is no documented evidence that patients with mania have elevated testosterone despite their increased activity levels, including sexual activity. Likewise, no data exist showing that testosterone is decreased in depression and that men with plasma testosterone levels within the normal range benefit from testosterone treatment.

What still needs to be investigated is whether male sex hormones are beneficial as neuroprotectants and, specifically, if they are beneficial in reducing the formation of β-amyloid, the protein related to Alzheimer's disease. It is of note that the brain contains an enzyme that converts testosterone into estrogen.

3. *Neuroactive Steroids—Dehydroepiandrosterone*

A special situation has developed with regard to dehydroepiandrosterone (DHEA), which has received disproportional media coverage and has been promoted as a 'fountain of youth.' This steroid was first described by Adolf Butenandt, who pioneered steroid biochemistry and who later received the Nobel prize in recognition of his work. DHEA is often thought to be a 'weak androgen,' which is incorrect as it is produced only to a small extent by gonadal glands and stems mainly from the adrenal cortex. The error possibly derives from the fact that testosterone and DHEA decrease with increasing age, whereas cortisol secretion increases. The increasing cortisol:DHEA ratio with age, however, reflects the fact that the part of the adrenal cortex from which cortisol is secreted (zona fasciculata) remains unchanged, while the part of the adrenal cortex from where DHEA is produced (zona reticularis) gradually undergoes apoptotic change with the consequence that the very high level of DHEA, which is present mainly as sulfate conjugate (DHEAS), declines (Arlt et al. 1999). Particularly low DHEAS levels have been recorded in patients under chronic stress or with serious health problems ranging from arthritis to HIV infection.

DHEAS is produced not only in the adrenal cortex, but also in the brain. Indeed, plasma DHEA concentrations in rats have been found to be lower than those in the brain and this is not due to the accumulation of peripherally produced steroid. In contrast, astrocytes and neurons are capable of synthesizing DHEA from pregnenolone, a common steroid pre-

cursor molecule and a derivative of cholesterol. Because nerve cells have the capacity to produce DHEA, this steroid is called a neurosteroid (Rupprecht and Holsboer 1999). The main effect of DHEA in the brain is its antagonistic effect on gamma-aminobutyric acid A ($GABA_A$) receptors. Based on this pharmacological property, several studies in animals and humans were initiated that observed the memory-enhancing effects of DHEA, compatible with the actions of other $GABA_A$ antagonists. The finding of a DHEA-induced increase in rapid eye movement sleep, which has been implicated in memory storage, is in accord with this notion. In the USA DHEA has become an 'over-the-counter' remedy for all ailments associated with aging, but this occurs in the absence of a solid database that would allow the use of DHEA in clinical treatment.

Another recent approach is an initiative in France where 280 men and women between 60 and 79 years received 50 mg DHEA or placebo daily for one year (Baulieu et al. 2000). The overall clinical impression obtained from the data of these healthy individuals was that this kind of replacement therapy has beneficial effects on bone and skin status as well as on several behavioral characteristics including sexual function. More refined analysis of this study, when available, will perhaps help us to identify the role of DHEA in preventive medicine. Until such data are made accessible to the public, DHEA will be considered by health care professionals as a 'lifestyle' drug. The only exception that relates to psychiatric conditions are studies where DHEA was administered to patients with depression where DHEA levels are low. Preliminary results from these studies noted an improvement in ratings of depression and aspects of memory function, supporting the notion that decreased DHEA levels are amenable to pharmacotherapy in selected cases.

4. *Thyroid Hormones*

There is a well-documented relationship between thyroid disease and psychiatric symptomatology. In patients with hypothyroidism, depressive symptoms are very frequent and disappear when thyroxine (T_4) or triiodothyronine (T_3) is given as substitution. These two hormones are secreted by the thyroid gland and act at the interior pituitary to inhibit thyroid-stimulating hormone (TSH), and at the hypothalamus to inhibit the synthesis and secretion of thyrotropin-releasing hormone (TRH).

At baseline, patients with depression are euthymic, i.e., they have normal T_3, T_4 and TSH levels. However, when the pituitary is stimulated with TRH, about 25 percent of the patients show blunted TSH responses, which correlates with the CRH-elicited corticotropin increases. The blunted TSH response is explained by chronic hypersecretion of TRH from the hypothalamus producing desensitization of anterior pituitary TRH receptors and ultimately reducing responsiveness to exogenous TRH. These laboratory findings do not suggest that hypothyroidism should be treated with thyroid hormones, which, nevertheless, have been used in various ways to treat depressive illness, mainly in combination with other antidepressants. In the treatment of depression, T_3 mainly is used because this hormone is the active steroid hormone in the brain, locally converted from T_4 by a brain-specific enzyme called deiodinase that removes one iodine atom from the thyroxine hormone molecule.

The finding that brain concentrations of T_3 are as high as those of T_4, whereas serum T_4 levels are 50 times higher than those of T_3 is a strong argument for a key role of T_3 in the brain. T_3 formation is increased by antidepressants that increase the aforementioned deiodinase enzyme activity. T_3 enters neurons readily and binds to nuclear receptors modulating the activity of many genes directly or through protein-protein interactions. As of yet it is entirely unclear which of T_3's neurochemical effects are involved in its beneficial effects on brain function. It is of note, however, that T_3 affects the serotonergic system as, for example, T_3 administration to rats increases basal serotonin levels in the forebrain. Similarly, T_3 has been found to induce subsensitivity of presynaptic serotonin 1A autoreceptors, proposed as one of the major mechanisms responsible for the antidepressant effect of selective serotinin reuptake inhibitors such as fluoxetine. In line with the special role of T_3 in brain function is the finding that psychopathological and neuropsychological symptoms of hypothyroidism respond favorably, if the patients receive a balanced combination of T_3 and T_4 instead of the usual T_4 monotherapy.

T_3 has been used in doses of 25 μg/day in combination with antidepressants and most reports claimed that adjunctive thyroid hormone treatment may hasten the onset of response (Newman et al. 2000). While most of these studies reported intriguing findings, they are preliminary and await confirmation in studies using current standard methods for efficacy validation. Another interesting approach using T_3 is the augmentation of the antidepressant effect in patients who have failed to respond to their current antidepressant therapy. Nonresponse, partial response, or failure to achieve stable remission frequently haunt practicing clinicians. Many studies exist reporting that T_3 can amplify the antidepressant effect thereby shifting nonresponders or partial responders into responders. A meta-analysis found that all T_3 augmentation studies could confirm that T_3 was more effective than placebo when combined with antidepressants. However, when more rigid standards were applied, e.g., randomized control design, etc., the significant difference between T_3 and placebo as potential augmentors of antidepressant action disappeared. Comparing all available augmentation strategies, it seems clear that T_3 studies have provided

the most convincing database and that this justifies a critical evaluation of T_3 dosage and duration of treatment in combination with new antidepressants.

5. Summary and Outlook

This article has taken a few examples of how neuroendocrinology has shaped psychiatric research, although hormone treatments have not yet entered clinical routine. Estrogen replacement therapy is well-established, but it is doubtful that it works as an antidepressant and its clinical usefulness in preventing Alzheimer's disease is not yet validated. Whether similar effects can be expected from testosterone replacement in men is an open question. DHEA treatment has received much attention, but its promise as a rejuvenating hormone still has to stand the test of time. Although not unanimously accepted, the most robust database exists for T_3 augmentation studies, which need to be corroborated by applying current validation standards for clinical trials. The most exciting development to date is the novel opportunity to interfere with drugs that block the anxiogenic and depression-like effects of CRH. Perhaps this new pharmacology can put hormone treatment back on the map in psychopharmacology.

Where will the future lead us? Classically, a neuroendocrine factor has been defined based on phenotypic changes. Entering the era of functional genomics, we will depart from studying one gene at a time and will rather apply a global approach, which will result in a dramatic expansion of our knowledge of gene sequences, expression profiles, and protein functions. From this rapidly progressing genomic revolution we can expect the discovery of many new genes in the endocrine systems. This will pave the way to the identification of new drug targets and the analysis of the genetic susceptibility of individuals to develop diseases where endocrine dysfunction, e.g., impaired stress hormone regulation, plays a role. Finally, if single nucleotide polymorphisms are mapped for gene areas of interest, customized therapies for the prevention of singularly detrimental effects of hormone dysregulation will become available. These new opportunities will establish a new platform where endocrinology works in partnership with psychiatry at a behavioral, functional, and genomic level.

See also: Biopsychology and Health; Depression; Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of; Hypothalamus; Mental Illness, Etiology of; Peptides and Psychiatry; Psychoneuroimmunology; Schizophrenia: Neuroscience Perspective; Sex Hormones and their Brain Receptors; Stress, Neural Basis of

Bibliography

Almeida O F, Condé G L, Crochemore C, Demeneix B A, Fischer D, Hassan A H S, Meyer M, Holsboer F, Michaelidis T M 2000 Subtle shifts in the ratio between pro- and antiapoptotic molecules after activation of corticosteroid receptors decide neuronal fate. *FASEB Journal* **14**: 779–90

Arlt W, Haas J, Callies F, Reincke M, Hübler D, Oettel M, Ernst M, Schulte H M, Allolio B 1999 Biotransformation of oral dehydroepiandrosterone in elderly men: Significant increase in circulating estrogens. *Journal of Clinical Endocrinology and Metabolism* **84**: 2170–6

Baulieu E-E, Thomas G, Legrain S, Lahlou N, Roger M, Debuire B, Faucounau V, Girard L, Hervy M P, Latour F, Leaud M C, Mokrane A, Pitti-Ferrandi H, Trivalle C, de Lacharrière O, Nouveau S, Rakoto-Arison B, Souberbielle J-C, Raison J, Le Bouc Y, Raynaud A, Girerd X, Forette F 2000 Dehydroepiandrosterone (DHEA), DHEA sulfate, and aging: Contribution of the DHEAge Study to a sociobiomedical issue. *Proceedings of the National Academy of Science of the United States of America* **97**: 4279–84

Behl C, Holsboer F 1999 The female sex hormone oestrogen as neuroprotectant. *Trends in Pharmacological Sciences* **20**: 441–4

de Kloet E R, Oitzl M S, Joëls M 1999 Stress and cognition: Are corticosteroids good or bad guys? *Trends in Neurosciences* **22**: 422–6

Grodstein F, Stampfer M J, Colditz G A, Willett W C, Manson J E, Joffe M, Rosner B, Fuchs C, Hankinson S E, Hunter D J, Hennekens C H, Speizer F E 1997 Postmenopausal hormone therapy and mortality. *New England Journal of Medicine* **336**: 1769–75

Häfner H, Riecher-Rössler A, an der Heiden W, Mauer K, Fätkenheuer B, Löffler W 1993 Generating and testing a causal explanation of the gender difference in age at first onset of schizophrenia. *Psychological Medicine* **23**: 925–40

Holsboer F 1995 Neuroendocrinology of mood disorders. In: Bloom F E, Kupfer D (eds.) *Psychopharmacology: The Fourth Generation of Progress*. Raven Press, New York, pp. 957–69

Holsboer F 1999 The rationale for corticotropin-releasing hormone receptor (CRH-R) antagonists to treat depression and anxiety. *Journal of Psychiatric Research* **33**: 181–214

Holsboer F 2000 The corticosteroid receptor hypothesis of depression. *Neuropsychopharmacology* **23**: 477–501

Newman M E, Agid O, Gur E, Lerer B 2000 Pharmacological mechanisms of T3 augmentation of antidepressant action. *International Journal of Neuropsychopharmacology* **3**: 187–91

Rupprecht R, Holsboer F 1999 Neuroactive steroids: Mechanisms of action and neuropsychopharmacological perspectives. *Trends in Neurosciences* **22**: 410–16

Santoro N F, Col N F, Eckman M H, Wong J B, Pauker S G, Cauley J A, Zmuda J, Crawford S, Johannes C B, Rossouw J E, Merz N B 1999 Therapeutic controversy. Hormone replacement therapy—where are we going? *Journal of Clinical Endocrinology and Metabolism* **84**: 1798–1812

Schmidt P J, Nieman L K, Danaceau M A, Adams L F, Rubinow D R 1998 Differential behavioral effects of gonadal steroids in women with and in those without premenstrual syndrome. *New England Journal of Medicine* **338**: 209–16

Steckler T, Holsboer F, Reul J M H M 1999 Glucocorticoids and depression. *Best Practice & Research Clinical Endocrinology and Metabolism* **13**: 597–614

F. Holsboer

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Endogeneity: Methodology

The concept of endogeneity arises when a distinction is made between a dependent variable, to be explained, and independent variables, used as explanations. A variable is endogenous if it is the dependent variable. A variable is exogenous if it is used to explain variations in the endogenous variable. There are various forms of endogeneity, such as in recursive versus interdependent systems. The concept plays an important role in statistical analysis of social science data, sometimes but not necessarily called causal analysis.

1. *Endogenous vs. Exogenous Variables*

In several fields in the social sciences a distinction is made between endogenous and exogenous variables. A simple example may clarify. In studies of social inequality researchers try to explain differences in the level of education that people achieve. Education can be measured, for example, as years of schooling or as less than high school, a high school degree, and so forth. The goal of social science analysis will then be to figure out how the level of education achieved varies with other characteristics, such as a person's social background, sex, race, and more. Social background can be measured in terms of parents' socioeconomic status or income. An early example of this kind of analysis in sociology is Blau and Duncan (1967).

In this example, level of education is the endogenous variable. The variables used to explain variations in the level of education are called exogenous. More generally, the variables that show differences we wish to explain are called endogenous, while the variables used to explain the differences are called exogenous. Often this goes along with a causal imagery. Exogenous variables are thought of as causes, endogenous as their effects. But there is no necessary connection; one may use the terms without implying causality (Sobel 1995). A simpler terminology for exogenous and endogenous is independent and dependent variables.

This distinction between exogenous and endogenous variables is used almost universally in quantitative social science. It arises in statistical data analysis, as in large parts of sociology, political science, and psychology. It arises also in mathematical analysis of theoretical ideas, as in large parts of economics and some parts of political science. But the distinction clearly is of relevance also in more qualitative research, such as in fieldwork and historical research, though less often used explicitly in those contexts. The distinction is simply relevant whenever there is some phenomenon that the researcher is trying to explain, which then is the endogenous variable, and there are some explanatory factors that are used to explain the phenomenon, which then are the exogenous variables. For example, in Skocpol's (1979) study of social revolutions, two exogenous variables are the state and agrarian structures in a society, while the endogenous is whether a revolution occurs or not.

2. *Recursive Systems: Causal Chains*

The example above can be elaborated. Consider what happens once people have achieved a particular level or type of education. They then often enter the labor market hoping to find a job. But the kinds of jobs people can expect to get depend strongly on their education. Psychologists are not hired to build bridges and engineers are rarely employed to provide mental-health services to children in despair. So in a next step, in explaining the kinds of jobs people find, education becomes an important independent or exogenous variable. It was endogenous, but is now an exogenous variable for explaining the type of job, which then is the endogenous variable. But the job may in turn become an important factor, thus exogenous, in explaining the salary received, which then is the endogenous variable.

And so it may continue. A variable, which once was endogenous, may be exogenous in another setting. There is a technical term for such reasoning, not important in itself, but sometimes encountered (from Bentzel and Wold 1946; see Hausman 1998). The variables entering into the reasoning are said to constitute a 'recursive system': parents' socioeconomic status influences education, education influences the type of job, the type of job influences earnings and career opportunities, and so forth. In causal imagery, they make up a causal chain. And it may also be that parents' socioeconomic status not only influences education, but also the job their offspring gets, for example through the network ties parents may have to employers or others who make hiring decisions. So an exogenous variable may be exogenous not only to one of the endogenous variables but to several of them.

For a set of variables to constitute a recursive system, there can be no feedback effects between them. Education may influence the type of job. But the type of job can have no influence on education itself. That such is the case is almost always tautologically true, provided the problem is carefully framed. The level of education is determined chronologically prior to the type of job, so there can be no feedback effect. This does not rule out that expectations and desires about future jobs will influence the education sought. But expectations about the future are not the same as the future. Enough of us have desired to be glamorous rock stars, outstanding violinists, famous war generals, or influential academics. Some may even have sought educational tracks correspondingly. A few succeed, most end up other places than aimed for. But whatever education and training we acquire, it usually has some impact on where we end up, though not always enabling us to get where we aspired to.

3. Statistical vs. Mathematical Analysis

Often the analysis of the relationships between variables is statistical. Researchers then typically use multivariate statistical methods. One tries to assess the joint influence of, say, sex and race on earnings. The goal is to see how the average level of earnings varies with sex, race, and so forth, not to explain the outcome for each individual. Similarly, if the endogenous variable is the type of job, say blue-collar, white-collar, professional, etc., then the goal of the analysis is to figure out how the percentages falling into each of these groups vary with education, sex, race, etc. Statistical social science analysis is eminently Holmesian, taking its cue from Sherlock Holmes. He argued that 'while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician' (Doyle 1986, p. 175). And though few quantitative social scientists would accept the statement that aggregate relationships are a mathematical certainty or that we can successfully predict aggregate behaviors, the gist of Holmes' comment is correct and agreed upon: We investigate what on average goes on.

If the analysis is purely mathematical, where one tries to explore theoretical ideas by means of mathematical tools, then the writer traces out the logical implications of the relationships between the variables. For instance, from the knowledge of how producers and consumers act and interact in modern economies, economists will specify several mathematical equations that try to capture elements of these actions and interactions. A goal can be to investigate, using the mathematical equations as an aid for reasoning, what the impact of an increase in the interest rate will be for consumption, production, and finally the national product. For example, increasing the interest rate may lead to a decline in home sales, leading to a decline in building of new homes, with correspondingly increased unemployment in the home construction sector, and so forth.

4. Interdependent Systems: Mutual Dependencies

In some settings two or more variables may mutually influence each other, sometimes referred to as being interdependently related, constituting an interdependent system. For example, the amount of schooling a person obtains may depend on her family behavior: whether she is married or not, and the presence or absence of children. Those who marry early and get children may decide not to pursue lengthy education. But vice versa, a person's family behavior may depend on her educational behavior, whether she is in school or not (e.g., Marini 1984). Being in school may delay family formation and getting children. In this example, family status is endogenous to schooling and schooling is similarly endogenous to family status. They mutually influence each other.

In the social sciences many attempts have been made to assess the strengths and even possible directions of such mutual influences. These vary greatly in how successful they are.

A number of approaches have attempted to use cross-sectional data, with information about variables at one point in time, to determine the strength and direction of causation between them. One may obtain information on the amount of schooling and family status of say people 30 years old. One then tries, using advanced statistical techniques, to disentangle how the two mutually depend on each other. Such attempts suffer from the drawback that they do not take into account the sequencing of and the temporal relationships between the variables. They can hence not be considered particularly successful (see Alexander and Reilly 1981).

An example may help. A woman may get pregnant at age 18, after finishing high school. This may lead to a shorter subsequent education. But the shorter subsequent education did not influence the event of giving birth at age 18. The expectation or intent of not obtaining additional education may have influenced whether to go through with a pregnancy at age 18, but as noted above, an expectation about the future is not the same as the future. The shorter subsequent education may, however, shorten the time before the person has a second child, which in turn may delay further education. At age 30, we may observe only the family status at that age, the number of children, and the education obtained so far. But unless one also collected data on the sequence of events as they unfolded over the life cycle, one cannot determine how childbirth influenced education and vice versa how education delayed or led to additional childbirths. Without a temporal relationship between values on two or more processes, mutual dependencies are hard if not impossible to assess (Petersen 1995).

It is therefore widely recognized that such interdependencies are best studied by means of dynamic models using over-time data, where the sequence and types of transitions are studied as they occur over the life cycle. This was already pointed out by Lazarsfeld (1948).

In the context of the present example, with over-time data on several individuals, one can investigate how family status at age 18 influences schooling status at age 19, and vice versa how schooling status age 18 influences family status at age 19. Suppose that family status has no impact on schooling status, so that marital status does not influence whether one goes to school or not in the next year. But suppose that schooling status has an effect on family status, so that it does matter whether you are in school or not for whether you get married or not in the next year. One would then say that schooling is an exogenous variable

relative to family status. This is the same as saying that family status is endogenous to schooling status. If they both impact each other, then both are endogenous to each other. Having figured out what happens at age 19, one can continue with analyzing the mutual influences of family and schooling status at age 19 on the same variables at age 20, and so forth. One can sequentially unravel how the relationships between the variables evolve over time. Bentzel and Hansen (1954) argued that mutual dependencies between two or more variables should be studied in such a manner.

There is an extensive statistical literature that addresses how to investigate which variables mutually influence each other, requiring access to over-time data as described above (Engel et al. 1983, Bergstrom 1984, Petersen 1995). Probably the best way to advance knowledge about which variables are endogenous and which are not is to collect careful data on sequences of event or variables (e.g., Tuma and Hannan 1984). No statistical technique can compensate for deficient data.

But no matter the kind of approach used or even favored, quantitative or more qualitative, the distinction between endogenous and exogenous variables is essential. Usually, there is some phenomenon in which we try to explain variations: the endogenous variable. Usually, there are some factors used to explain those variations: the exogenous variables.

See also: Causal Inference and Statistical Fallacies; Causation (Theories and Models): Conceptions in the Social Sciences; Control Variable in Research; Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression); Factor Analysis and Latent Structure: Overview; Latent Structure and Casual Variables; Mediating Variable; Multivariate Analysis: Overview; Structural Equation Modeling

Bibliography

Alexander K L, Reilly T W 1981 Estimating the effects of marriage timing on educational attainment: Some procedural issues and substantive clarifications. *American Journal of Sociology* **87**(1): 143–56

Bentzel R, Hansen B 1954 On recursiveness and interdependency in economic models. *Review of Economic Studies* **22**(3): 153–68

Bentzel R, Wold H 1946 Statistical demand analysis from the viewpoint of simultaneous equations. *Scandinavian Actuarial Journal* **29**: 95–114

Bergstrom A R 1984 Continuous time stochastic models and issues of aggregation over time. In: Griliches Z, Intriligator M D (eds.) *Handbook of Econometrics*. North-Holland, Amsterdam, Vol. II, pp. 1145–1212

Blau P M, Duncan O D 1967 *The American Occupational Structure*. Wiley, New York

Doyle A C 1986 *Sherlock Holmes: The Complete Novels and Stories*. Vols 1 and 2. Bantam Books, New York

Engle R F, Hendry D F, Richard J-F 1983 Exogeneity. *Econometrica* **51**(2): 277–304

Hausman D M 1998 *Causal Asymmetries*. Cambridge University Press, New York

Lazarsfeld P F 1948 The use of panels in social research. *Proceedings of the American Philosophical Society* **92**(5): 406–10

Marini M M 1984 Women's educational attainment and the timing of entry into parenthood. *American Sociological Review* **49**(4): 491–511

Petersen T 1995 Models for interdependent event history data: specification and estimation. *Sociological Methodology* **25**: 317–75 (Edited by Peter V. Marsden.) Oxford, UK: Basil Blackwell, for the American Sociological Association

Skocpol T 1979 *States and Revolutions*. Cambridge University Press, New York

Sobel M E 1995 Causal inference in the social and behavioral sciences. In: Arminger G, Clogg C C, Sobel M E (eds.) *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Plenum Press, New York, Chap. 1, pp. 1–38

Tuma N B, Hannan M T 1984 *Social Dynamics. Models and Methods*. Academic Press, Orlando, FL

T. Petersen

Energy Economics

Energy economics is the field that studies human utilization of energy resources and energy commodities, and the consequences of that utilization. In physical science terminology, 'energy' is the capacity for doing work, for example, lifting, accelerating, or heating material. In economic terminology, 'energy' includes all energy commodities and energy resources, commodities or resources that embody significant amounts of physical energy and thus offer the ability to perform work. Energy commodities—for example, gasoline, diesel fuel, natural gas, propane, coal, or electricity—can be used to provide energy services for human activities, such as lighting, space heating, water heating, cooking, motive power, or electronic activity. Energy resources—for example, crude oil, natural gas, coal, biomass, hydro, uranium, wind, sunlight, or geothermal deposits—can be harvested to produce energy commodities.

Energy economics studies forces that lead economic agents—firms, individuals, governments—to supply energy resources, to convert those resources into other useful energy forms, to transport them to the users, to use them, and to dispose of the residuals. It studies roles of alternative market and regulatory structures on these activities, economic distributional impacts, and environmental consequences. It studies economically efficient provision and use of energy commodi-

Table 1
Physical properties of common energy resources and commodities

Resource/commodity	Energy form	Physical form	Timescale
Crude oil	Chemical	Liquid	Depletable resource
Refined petroleum products	Chemical	Liquid	Storable commodity
Natural gas	Chemical	Gas	Depletable resource
Processed natural gas	Chemical	Gas	Storable commodity
Coal	Chemical	Solid	Depletable resource; storable commodity
Trees/biomass	Chemical	Solid	Renewable, storable resource
Battery	Chemical	Solid	Storable commodity
Electricity	Electrical	Moving electrons	Nonstorable commodity
Wind	Mechanical	Moving gas	Renewable, nonstorable resource
Hydro	Mechanical	Moving liquid	Renewable, storable resource
Geothermal	Thermal	Solid or liquid	Renewable or depletable resource
Uranium	Nuclear	Solid	Depletable resource
Solar radiation	Radiation	Pure energy	Renewable: nonstorable resource

ties, and resources and factors that lead away from economic efficiency.

1. *Properties of Energy Resources and Energy Commodities*

Other than all embodying significant amounts of physical energy, energy resources or commodities vary greatly. They may embody chemical energy (e.g., oil, natural gas, coal, biomass), mechanical energy (e.g., wind, falling water), thermal energy (geothermal deposits), radiation (sunlight, infrared radiation), electrical energy (electricity), or the potential to create energy through nuclear reactions (uranium, plutonium.) They have differing physical forms. Crude oil, most refined petroleum products, and water are liquids. Water includes available energy only through its motion. Coal, most biomass, and uranium are solids. Natural gas and wind are gases, with wind including available energy based only on its movement. Geothermal energy is available through hot liquids (normally water) or solids (subterranean rock formations). Solar radiation is a pure form of energy. Electricity consists of electrons moving under an electrical potential.

Resources can be viewed as renewable or depletable. Some renewable resources can be stored; others are not storable. These issues will be discussed more fully in a subsequent section. Table 1 summarizes the energy form, physical form, renewable/depletable distinction for several common energy resources and commodities.

2. *Energy Conversion Processes*

A fundamental property of energy is expressed by the first law of thermodynamics: energy can be neither created nor destroyed (except through nuclear reactions transforming matter to energy.) Energy can be converted between forms, and human use of energy typically involves such conversions for human ends.

Energy conversion processes are basic to human experience. Fire, providing heat and light, is a process by which chemical energy stored in the fuel (say, wood) is converted to thermal energy and radiant energy. Chemical energy stored in wood is the result of photosynthesis, whereby plants convert energy in sunlight to chemical energy, stored in the plant material. Carbohydrates in food are converted within the human body to thermal energy and mechanical energy, providing body warmth and movement.

The industrial revolution was characterized by a change from the use of hand tools, using human mechanical energy, to machine and power tools. Machine tools allowed the conversion of energy in falling water to mechanical energy (water wheels), or the conversion of chemical energy in wood or coal to mechanical energy (steam engines) for industrial processes.

Humans now routinely harness complex sequences of energy conversion processes to provide desired services. Crude oil is separated into refined products such as gasoline, diesel oil, jet fuel, and heavy distillates, that embody chemical energy. Gasoline or diesel oil are burned explosively in internal combustion engines, converting chemical energy into thermal energy. Heated gases push engine pistons, converting thermal energy into mechanical energy. Some is lost as heated gases or as radiant energy. The mechanical energy moves the automobile and, in the process, is converted to thermal energy through friction within the automobile or between the automobile and the road or air. Some mechanical energy is converted to electrical energy by a generator in the automobile, to power electrical equipment. Some electrical energy is converted into chemical energy in the automobile battery. To start the car, chemical energy in the battery is converted to electrical energy which is then converted to mechanical energy to turn the engine.

Similarly, coal combustion converts chemical energy into thermal energy to create steam, which powers a turbine in an electric generating plant, converting the thermal energy into mechanical energy, and then into electrical energy. Electricity can power a motor (converting to mechanical energy), heat a room (to thermal energy), or light a bulb (to thermal energy, then to light). This energy is later converted to thermal energy, which ultimately is radiated into space.

Energy economics recognizes the fundamental physical realities that (a) no energy is created or destroyed, but that energy can be converted among its various forms, and (b) energy comes from the physical environment and ultimately is released back into it. Thus, energy economics is the study of human activities using energy resources from naturally available forms, through often complex conversion processes,

to forms providing energy services. Several issues of the demand for energy will be examined next.

3. Demand for Energy as a Derived Demand

Demand for energy is derived from wishes to use energy to obtain desired services. It is not derived from preferences for the energy commodity itself. Energy demand depends primarily on demand for desired services, availability and properties of energy conversion technologies, and costs of energy and technologies used for conversion.

For example, consumers use gasoline to fuel an automobile or other motorized vehicle, converting gasoline to mechanical energy for motive power. The amount of gasoline used is proportional to the miles the auto is driven, and inversely proportional to the efficiency by which gasoline is converted to useful mechanical energy, measured as miles per gallon (mpg) of gasoline used by the automobile. Demand for gasoline is thus derived from choices about distances vehicles are driven, and their energy conversion efficiencies.

Similarly, electricity is purchased by consumers only to perform functions using electricity. Typical electricity uses include lighting, refrigeration, space heating, air conditioning, clothes washing, drying, dish washing, water heating, and operating electronic equipment such as computers or televisions. Electrical energy is converted to mechanical energy (motors in refrigerators, air-conditioning units, vacuum cleaners), thermal energy (space heating, clothes dryers, water heating), or radiation (lighting, television, computer monitors.) Electricity demand is derived from demand for the underlying services—comfortable space, refrigeration, cleaning, entertainment, information processing.

In each case, efficiency of energy conversion equipment also determines energy demand. Typically, energy conversion equipment is long-lived—automobiles, air-conditioning units, refrigerators, televisions, computer systems, furnaces. Consumers or firms can usually choose among alternatives with various conversion efficiencies; such choices influence energy demand significantly. To the extent that consumers and firms purchase these units with an understanding of their conversion efficiencies, expectations of future energy prices can influence choices of particular equipment. For example, high natural gas prices can motivate consumers to invest in home insulation.

In general, increased energy prices reduce demand by reducing use of energy services and motivating selection of higher conversion efficiency equipment. For example, gasoline prices influence demand through vehicle miles and fuel efficiency of vehicles. Vehicle miles are influenced by cost per mile of driving, including per-mile gasoline costs, equal to the ratio pg/mpg (where pg is the gasoline price), and other costs. Increased gasoline prices lead consumers to purchase more fuel-efficient cars. Both factors imply that increased gasoline prices reduce gasoline demand, with the vehicle miles adjusting relatively quickly, and vehicular fuel efficiency adjusting slowly as vehicles enter the fleet. Except for firms selling energy resources or energy commodities, similar issues are important for industrial and commercial use of energy.

4. Demand Substitution among Energy Commodities

Some energy services can be provided by several different energy commodities. Homes can be heated using electricity, natural gas, oil, or wood, since each can be converted to thermal energy. Cooking can use electricity, natural gas, propane, wood, or charcoal. Thus, energy commodities typically are economic substitutes for one another: The demand for a particular energy commodity is an increasing function of prices of other energy commodities.

This substitutability of energy is made possible by, and is also limited by, the available set of energy conversion technologies. Typically, one conversion technology can be used only for a particular energy commodity. For home heating, a natural gas furnace cannot use oil, electricity, or wood. Because conversion equipment typically is very long lived, substitution among energy commodities occurs only slowly, and then only when new equipment is purchased. Short-run substitution can usually occur only if several energy conversion technologies are available simultaneously for use by particular consumers, for example, homes that have a central natural gas heating system plus portable electric space heating units. Thus, various energy commodities can usually be viewed as imperfect substitutes for one another, with much greater substitutability in the long run than in the short run.

The degree of substitutability can be altered sharply by the development of new conversion alternatives. For example, historically, automobiles were fueled only by gasoline or diesel fuel, but technologies currently being developed would allow them to be powered by electricity, natural gas, propane, hydrogen, or other energy commodities. Once such conversion technologies are commercialized successfully, gasoline and other energy commodities will become highly substitutable in transportation.

5. Is Energy an Essential Good?

In economics, an essential good is one for which the demand remains positive no matter how high its price

becomes. At the theoretical limit, for prices unboundedly high, consumers would allocate all of their income to purchases of the essential good.

Energy is often described as an essential good because human activity would be impossible absent use of energy: living requires food embodying chemical energy. Although energy is essential to humans, neither particular energy commodities nor any commercial energy commodities are essential goods. Particular energy commodities are not essential because consumers can convert one form of energy into another. Even the aggregate of all commercial energy cannot be viewed as an essential good. Experience from low-energy research facilities shows that an extremely energy efficient home needs relatively little energy. Solar energy could generate electricity or heat water. Travel could be limited to walking or riding bicycles. Solar-generated electricity or wood fires could be used for cooking. For high enough prices of purchased energy, demand for purchased energy by consumers could be reduced to zero. Thus, commercial energy is not an essential good.

6. *Optimality of Consumer Choice*

Debate is ongoing about the extent to which consumers understand conversion efficiencies of alternative technologies and act on this understanding, and the extent to which manufacturers of conversion equipment respond to consumer preferences. Labeling requirements have been one policy response to concerns that consumers otherwise would have insufficient information to enable them to choose among energy conversion equipment. New cars have stickers with estimates of mpg under different driving cycles. Refrigerator labels estimate annual electricity use and cost. The concern has led to the adoption of energy efficiency standards for household equipment, such as refrigerators, under the belief that labeling alone is not sufficient to motivate optimal consumer choices. Similarly, the imposition of corporate average fuel efficiency standards (CAFE) for automobiles is a legislative response to concern that, on their own, automobile manufacturers will make automobiles less fuel efficient than optimal. These behavioral issues and the appropriate policy responses have not been resolved fully in either the existing literature or the policy community.

7. *Depletable, Storable Renewable, Nonstorable Renewable Resources*

Based on the speed of natural processes, one can classify primary energy resources as depletable or renewable. Renewable resources can be further subdivided into storable or nonstorable resources. Renewable resources are self-renewing within a timescale important for economic decision making. Storable renewable resources typically exist as a stock which can be used or can be stored. Biomass, hydro power, and some geothermal, fall in this category. The amount used at one time influences the amount available in subsequent times. Nonstorable renewable resources—wind, solar radiation, run-of-the-river hydro resources—can be used or not, but the quantity used at a given time has no direct influence on the quantity available subsequently. Most energy commodities are storable (refined petroleum products, processed natural gas, coal, batteries), but electricity is not storable as electricity. Depletable resources are those whose renewal speeds are so slow that it is appropriate to view them as made available once and only once by nature. Crude oil, natural gas, coal, and uranium all fall in this category.

Table 2
US energy supply (domestic plus imports)

Sources of energy consumed in the US, 1999 (quadrillion BTU)			
Depletable resources		Renewable resources	
Coal	21.8	Hydroelectric power	3.4
Natural gas	22.1	Wood and waste	3.5
Petroleum: domestically produced	15.2	Geothermal	0.3
Petroleum: imported	22.5	Wind	0.04
Nuclear	7.7	Solar	0.08
Total depletable resources	89.3	Total renewable resources	7.4

Initially, all human energy use depended on renewable resources, in particular biomass resources used for food, heat, or light. In the United States, renewable energy—human, animal, water, wood, and wind power—dominated energy supply through the middle of the nineteenth century. Only during the second half of the nineteenth century did a depletable resource, coal, surpass renewable resource use. Crude oil and natural gas started supplying large quantities of energy only in the 1920s.

Now the dominant use of energy (other than as food) in developed nations is based on depletable resources, particularly fossil fuels. Table 2 shows that of the total sources of energy consumed in the United States in 1999, 92 percent was from a depletable resource and only 8 percent was from a renewable resource, of which almost all was hydroelectric and biomass (wood and waste).

But depletable resource use cannot dominate for ever. Once particular deposits have been used, they cannot be reused. Therefore, a future transition from depletable resources, particularly from fossil fuels, is inevitable. However, which renewable energy sources will dominate future consumption is not clear. And there is great uncertainty about the timing of a shift to renewable energy resources. Related is the unresolved question of future energy adequacy: will the renew-

able sources of energy be adequate to satisfy demands for energy, once the fossil fuel supplies move close to ultimate depletion? These issues will be examined further in the next section.

8. *Depletable Resource Economics and the Transition to Renewable Resources*

The study of depletable resource economics began with articles by Lewis Grey (1914) and Harold Hotelling (1931), which examined economically intertemporal optimal extraction from a perfectly known stock of the resource, with perfectly predictable future prices of the extracted commodity. Many, but not all, subsequent articles maintained the perfect knowledge assumptions.

The essential result is that under optimal extraction paths the resource owner recognizes (explicitly or implicitly) an opportunity cost, or rent, in addition to the marginal extraction costs. All information about the role of future prices and costs would be embodied in this opportunity cost. The competitive firm would extract at a rate such that the marginal extraction cost plus opportunity cost would equal the selling price for the extracted commodity. Price would thus exceed marginal cost, even if the firm were operating perfectly competitively. This opportunity cost would evolve smoothly over time. As the resource neared depletion, the opportunity cost and the marginal extraction cost even at very low extraction rates would together have increased until they equaled the commodity price, at which time extraction would cease.

In depletable resource theory, market prices would increase gradually to the cost of producing substitutes, reaching that cost only as the depletable resource were nearing depletion. Substitutes would be produced only in small quantities until near the time of depletion. Market forces would, automatically and optimally, guide commodity prices upward so that when the depletable resources were nearing depletion, commodity prices would have risen to a point at which the demands could be fully satisfied by the substitutes.

In reality, the economic cycle for depletable resources is far more complex and more prone to error and surprise. The cycle typically begins with innovations that allow the resource to be utilized. Technologies improve over time, partially guided by economic forces and public policy decisions, but often in somewhat unpredictable ways. Generally, the magnitude and location of the resource base remains unknown, and exploration is required to identify resource deposits. But exploration is costly. Therefore, typically it is optimal for companies to explore only until they find sufficient resources to satisfy their expectations of near-term extraction. These discovered resources—referred to as 'reserves'—typically are only a fraction of the resource base and do not provide reliable estimates of the overall size of the resource stock. Firms may extract optimally from the proven reserves but cannot know with any certainty the quantity or extraction costs of the undiscovered resources. Opportunity cost would depend on the future prices of the extracted commodity, but future commodity prices will themselves depend on future demand and supply, which in turn depend on uncertain future discoveries, and which are therefore very unpredictable.

With this additional complexity and uncertainty, although the opportunity cost concept remains important, it seems appropriate to focus more attention on responses of markets to random changes in technologies, reserves, and other market information. And it seems appropriate to abandon the notion that markets will automatically and optimally guide the system to a smooth transition to renewable resources.

One central result does remain, however. If there is only a limited stock of a resource—undiscovered plus discovered—then there will be only a limited number of years during which the resource can be extracted. Here, there is an important commonality between energy economics and ecological economics (see *Ecological Economics*). The typical pattern includes an initial period in which the resource is not used, before technology for extracting and/or using the resource is developed. Extraction rates rise over time, perhaps rapidly, as the technology develops and demand increases. Commodity prices would fall with falling extraction and finding costs. However, at some time, rising costs due to depletion of the resource start overtaking the decreasing costs due to technology advances. The extraction rate declines until ultimately all of the economical resource stocks are depleted. At that point, the consumers of the depletable resource must substitute some other means of satisfying energy service demand.

If markets work well, the renewable resources will then be available in sufficient quantities and at reasonable costs. A transition to renewable energy resources will have been accomplished successfully, and there will always be an adequate energy supply to satisfy all demands at the prevailing market price.

However, given the complexity and the uncertainty, it is not obvious that the transition will work as automatically, as smoothly, or as optimally as is suggested above. It may be that fossil fuels saved for the future will ultimately never be needed because substitute forms of energy become available at a lower cost and at an earlier time than expected. Or it may be that fossil fuels are depleted rapidly, but that costs of renewable resources remain well above expected levels, or that quantities of renewable resources are more limited than expected. In either case, in retrospect everyone might wish to have made very different public policy decisions.

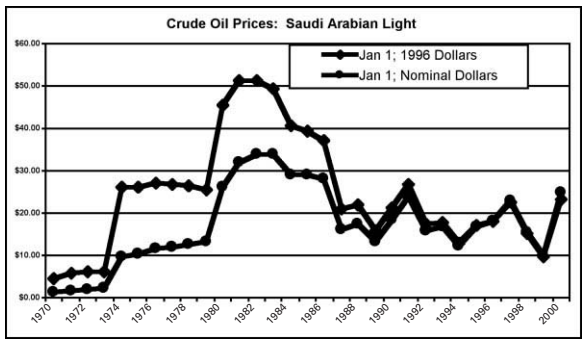

Figure 1
World crude oil prices: nominal dollars and real dollars (1996), beginning of year prices

9. *World Oil Prices*

In 1973, energy markets were disrupted when war broke out between Israel and neighboring nations. Oil supply was reduced by some member nations of the Organization of Petroleum Exporting Countries (OPEC). At that time there was relatively little excess worldwide oil extraction capacity. Consumers and producers expected further disruptions. Therefore, the supply reductions led to disproportionately sharp increases in the world price of oil, increases that remained well after the war. Another price increase occurred in 1979. The price of Saudi Arabian light jumped from $2.10 per barrel (/bbl) at the beginning of year (boy) 1973 to $9.60/bbl at boy 1974. The boy 1979 price of $13.34/bbl was followed by a boy 1980 price of $26.00 per barrel. Not until 1986 did prices drop to levels consistent with the current levels. Prices are well above the pre-1973 levels, even when adjusted for inflation, and remain very volatile. Figure 1 shows these boy annual prices in nominal dollars and in 1996 dollars, adjusted using the US implicit GNP price deflator.

Many explanations have been offered for the persistence of the world price increases, including assertions that they simply reflected a realization that oil resources were being depleted, that they reflected only random movements, and that they reflect independent choices by various nations attempting to satisfy developmental goals. The most obvious explanation, that they reflect a collusive exercise of market power by OPEC members, is the best supported by economic data.

These sudden jumps in oil prices led directly to a wide range of economic impacts, and indirectly to impacts through governmental programs responding to the perceived crisis. Worldwide recessions followed both price jumps, partly caused by direct disruptions to industry, partly by reductions in real income of oil-importing countries, and partly by tight monetary policies imposed to decrease oil-market-induced inflation. Price and allocation controls were placed on refined petroleum products in the USA, resulting in widespread shortages, manifested by long gasoline lines. And prices of other energy resources and commodities increased significantly in response to the oil price increases.

The oil price increases and ensuing economic problems led to profound changes in energy policy and energy markets worldwide. For example, in the USA, President Nixon declared 'Project Independence' to reduce sharply US dependence on imported oil. Oil exploration and development increased extraction capacity throughout the non-OPEC nations, particularly in the North Sea (Norway and the UK), China, and Mexico. Energy 'conservation' programs were enacted to reduce energy demand, particularly for oil. For example, corporate average fuel efficiency (CAFE) standards were imposed on the US automobile industry. Public- and private-sector investments in renewable energy R&D increased sharply, with emphasis on solar, wind, geothermal, and hydro power. New national and international organizations were created, for example, the International Energy Agency (IEA) in Paris, and the US Department of Energy. Programs were initiated to reduce economic impacts of oil disruptions: the US government purchased large quantities of crude oil for storage underground; oil-importing nations agreed to IAE-monitored minimum levels of oil storage. Although some changes have been reversed—for example, investment in energy R&D has declined since 1986—oil policies, particularly for oil-importing nations, are still shaped by these profound changes in the world oil price.

10. *Energy Conversion Industries*

In addition, activities are associated with commercial conversion of energy from one form to another, particularly to electricity, from hydro power, coal, natural gas, oil, nuclear fission, wood and waste products, geothermal, wind, or solar radiation. Energy conversion industries, for economic success, must be able to sell their product at a price higher than the cost of energy commodities used as inputs plus per unit capital and operating costs of the facilities. Energy conversion is never perfectly efficient, and some input energy is lost into the environment. Therefore, the price per Btu of electricity must be substantially greater than the price of energy commodities used to generate electricity.

Technological advance can be very important. New technologies are becoming available that increase the conversion efficiency from natural gas or coal to electricity, and which can be expected to have lower operating and capital costs. Such technological

advances can be expected to bring prices of these energy commodities closer together over time.

In addition to these technological changes, there are important ongoing changes in economic structure of the electricity production and distribution industry, throughout the world. In many countries, state-owned industries generate, transmit, and distribute electricity. In others, private electricity suppliers are subject to special economic regulation. The reason for governmental ownership or control seems to stem from two factors. Electricity is fundamental to economic activity, and many people have not trusted private industry. Second, production, transmission, and distribution of electricity have shown significant increasing returns to scale, and the industry has been viewed as a natural monopoly. Fearful that an unregulated monopoly would exercise market power and overprice electricity, most nations have chosen to own or to control tightly the industry.

Recently, however, smaller geographically distributed electricity generating plants that could reasonably compete with one another, have become attractive economically. Thus, the possibility for competition in electricity generation has been recognized. In addition, it is now realized that an electricity utility sells two classes of product: electricity delivery services (wires), and electricity. Although these two classes of product traditionally were bundled together into a price per kilowatt hour of electricity, in principle, the two could be unbundled and sold by separate companies. Electricity delivery service is characterized by increasing returns to scale, but electricity itself is not. Therefore, the possibility is open for a competitive market structure to sell electricity to consumers, separately from the electricity delivery services.

In some localities this movement toward privatization and deregulation seems to be very successful; in others, it has not been. At the time of writing, there is an intense and heated debate about whether deregulation of the electricity industry is appropriate, and if so, what is the appropriate form of deregulation.

11. *Environmental Consequences of Energy Use*

Much important environmental damage stems from the production, conversion, and consumption of energy. Costs of this environmental damage generally are not incorporated into prices for energy commodities and resources; this omission leads to overuse of energy. Concern about this issue is common to energy economics, environmental economics (see *Environmental Economics*), and ecological economics (see *Ecological Economics*), with energy economics and environmental economics literature attempting to assign monetary valuation of the impacts, and ecological economics rejecting the idea that a monetary value could be placed on environmental impacts.

Environmental impacts currently receiving most attention are associated with the release of greenhouse gases into the atmosphere, primarily carbon dioxide, from combustion of fossil fuels. The three primary fossil fuels—coal, petroleum, and natural gas—each include carbon. During combustion, carbon combines with oxygen to produce carbon dioxide, the primary greenhouse gas. Carbon dioxide accumulates in the atmosphere and is expected to result in significant detrimental impacts on the world's climate, including global warming, rises in the ocean levels, increased intensity of tropical storms, and losses in biodiversity. Fossil fuels account for 98 percent of the US carbon dioxide net releases into the atmosphere, and 82 percent of the releases of greenhouse gases, measured on a carbon equivalent basis.

Energy use leads to additional environmental damages. Coal combustion, particularly high sulfur coal combustion, emits oxides of sulfur, which, through atmospheric chemical reactions, result in acid rain. Automobile gasoline combustion releases oxides of nitrogen and volatile organic compounds, which, in the presence of sunlight, result in smog. Electricity generating facilities often use large quantities of water for cooling and release the heated water into lakes or oceans, leading to local impacts on the ecosystem. Extraction of oil or mining of coal can lead to subsidence of the land overlying the extracted deposits.

Pervasive environmental impacts of energy use, absent governmental intervention, imply that significant costs of energy use are not included in the price energy users face. These so-called externalities (see *Environmental Economics*) lead to overuse of energy and provide strong motivation for interventions designed to reduce energy use.

See also: Energy, Geography of; Energy Policy: International; Environment and Development; Environmental and Resource Management; Environmental Economics; Environmental Policy; Environmentalism: Preservation and Conservation; Resource Geography; Resource Institutions; Sustainability Transition: Human–Environment Relationship; Sustainable Development

Bibliography

Adelman M A 1993 *The Economics of Petroleum Supply: Papers by M A Adelman 1962–1993*. MIT Press, Cambridge, MA

Barnett H, Morse C 1963 *Scarcity and Growth: The Economics of Natural Resource Availability*. Johns Hopkins University Press, Baltimore, MD

Berndt E R, Wood D O 1975 Technology, prices, and the derived demand for energy. *Review of Economics and Statistics* **57**: 259–68

Bohi D R 1981 *Analyzing Demand Behavior: A Study of Energy Elasticities*. Johns Hopkins University Press, Baltimore, MD
Bohi D R, Toman M A 1996 *The Economics of Energy Security*. Kluwer, Boston, MA
Cato D M, Rodekohr M, Sweeney J 1976 The capital stock adjustment process and the demand for gasoline: A market share approach. In: Askin B, Kraft J (eds.) *Econometric Dimensions of Energy Supply and Demand*. D C Heath, Lexington, KY
Cropper M L, Oates W E 1992 Environmental economics: A survey. *Journal of Economic Literature* **30**: 675–740
Dasgupta P S, Heal G 1979 *Economic Theory and Exhaustible Resources*. Cambridge University Press, Cambridge, UK
Denny M, Fuss M, Waverman L 1981 The substitution possibilities for energy: Evidence from US and Canadian manufacturing industries. In: Berndt E R, Field B (eds.) *Modeling and Measuring Natural Resource Substitution*. MIT Press, Cambridge, MA
Energy Information Administration 2001 *Annual Energy Review 2000*. US Department of Energy, Washington, DC
Energy Modeling Forum 1982 *World Oil*. The Energy Modeling Forum, Stanford University, Stanford, CA
Federal Energy Administration 1974 *Project Independence Report*. US Federal Energy Administration, Washington, DC
Gately D 1984 A ten-year retrospective on OPEC and the world oil market. *Journal of Economic Literature* **22**(3): 1100–14
Grey L C 1914 Rent under the assumption of exhaustibility. *Quarterly Journal of Economics* **28**: 466–89
Griffin J M, Steele H B 1986 *Energy Economics and Policy*, 2nd edn. Academic Press, New York
Heal G 1976 The relationship between price and extraction cost for a resource with a backstop technology. *The Bell Journal of Economics* **7**(2): 371–8
Hickman B G, Huntington H, Sweeney J (eds.) 1987 *Macroeconomic Impacts of Energy Shocks*. North-Holland, Amsterdam
Hotelling H 1931 The economics of exhaustible resources. *Journal of Political Economy* **39**: 137–75
Hudson E A, Jorgenson D W 1974 US energy policy and economic growth: 1975–2000. *The Bell Journal of Economics* **5**: 463–514
Kneese A V, Sweeney J L 1993 (eds.) *Handbook of Natural Resource and Energy Economics*. Elsevier Science Publishers, Amsterdam, Vol. III
Nordhaus W D 1994 *Managing the Global Commons: The Economics of Climate Change*. MIT Press, Cambridge, MA
Slade M E 1982 Trends in natural resource commodity prices: An analysis of the time domain. *Journal of Environmental Economics and Management* **9**: 122–37
Solow R M 1974 The economics of resources or the resources of economics. Papers and Proceedings. Richard T Ely Lecture. *American Economic Review*
Sweeney J L 1977 Economics of depletable resources: Market forces and intertemporal bias. *Review of Economic Studies* February: 125–41
Sweeney J L 1984 The response of energy demand to higher prices: What have we learned? *American Economic Review* **74**(2): 31–7
Taylor L D 1975 The demand for electricity: A survey. *The Bell Journal of Economics* **6**: 74–110
Weyant J 1993 Costs of reducing global carbon emissions. *Journal of Economic Perspectives* **7**(1): 27–46

J. L. Sweeney

Energy, Geography of

The geography of energy is concerned with the interrelationship between space and energy as the most important factor in human activity. The production, distribution, and consumption of energy determine all of the basic functions of our existence: work and production, housing, education, transport, communication, and even recreation. Accordingly, an energy chain (Chapman 1989, see also Brücher 1997, Scheer 1999) connects production and consumption, with various stages interacting with space, the environment, and society (Fig. 1).

1. *From Energy Chain to Energy System*

The use of all kinds of energy (e.g., heat, electricity) and energy resources (e.g., coal, oil) follows such an energy chain, giving it an individual profile. Take, for instance, the example of coal in the old mining areas (the numbers refer to Fig. 1). Underground mining manifests itself at the surface with various installations for the extraction (2) preparation (4), and processing (5) of the mined resource with each installation, connected by transport lines (3). After (possible) storage (6), coal is burnt to produce heat for the consumer (8) or transformed (7) into secondary energy (e.g., coke, electricity). All mining activities harm the environment in various ways through noise, pollution, surface subsidence, and heaps of mine-tailings. At relatively few locations, the high energy content of this fuel and the great labor intensity required to procure it has resulted in extensive conurbations (e.g., Ruhr (Germany), Midlands (UK), Pittsburgh (USA)). Mining is also an important political object, whether as a basis of energy supply and industry, or as a problem area characterized by economic decline. In comparison to the energy chains of other resources, like uranium, it is evident that there are not only substantial differences between them, but also between their specific impacts on space and on political decision making.

The generation of electricity from water power only, as in Norway, or the exclusive supply of energy by hydrocarbons, as in Saudi Arabia, are rare exceptions. Nearly every country has to use different resources and to combine their respective energy chains. Such an energy mix is based on availability, price level, socioeconomic structures, supply security, environmental protection, and so on. Furthermore, there is a need to adapt to the strategies of the nongovernment energy utilities and find an appropriate degree of acceptance by the electorate (e.g., 'green' anti-nukes in Germany). Given these and other considerations, every government shapes its own individual energy policy. The less energy a country owns (or estimates it owns), the more influential the impact of policy on the economy of energy will be; a good example for this is

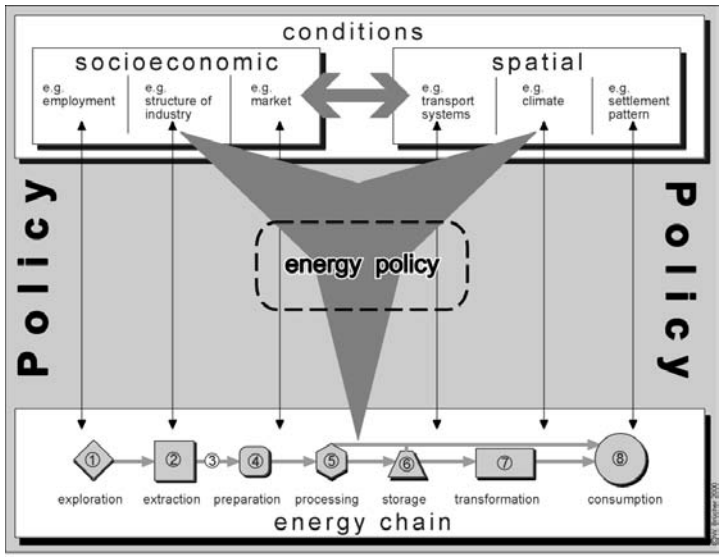

Figure 1
Interactions between policy, socioeconomic and spatial conditions and the energy chain

the decision of France to produce a maximum of nuclear electricity.

In any given spatial-temporal situation the ecological, and technical conditions of energy supply, together with the energy policy and the utility structures, form a complex system with its specific impact on space. Referring to Debeir et al. (1989), (1991) and Chapman (1989) we call it an energy system. One can find examples based on slavery, on the steam engine, or on the power grid. In the wake of technical and economic progress, and with mounting requirements for energy and its changing forms of application, existing systems are being replaced by new ones. Such a transition from an old to a new system is always combined with a fundamental change in the interrelationship between the energy chain and space.

2. *Changing Energy Systems: From 'Energy from Space' to 'Energy for Space'*

Before the Industrial Revolution, energy was obtained exclusively from biomass (e.g., wood, food muscle power), from water and wind, and thus exclusively from the surrounding space. Was this not, ultimately, the decisive incentive for the historical conquests of territories? As the energy content of these resources is very low, there has always been a need to concentrate it: through irrigation and fertilizing, the grain harvest and thereby the human workforce can be increased. Similarly, a ton of charcoal holds twice the energy content of a ton of wood. Let us have a look at the energy system of the Romans. In their capital, up to half of the population was comprised of slaves, the 'engines' of antiquity. They were 'powered' by bioenergy derived from cereals supplied from the whole empire via a converging network of roads and shipping routes. Finally, the system destroyed itself through the excessive consumption of energy for transport over excessively long distances. This is the inherent limit of any energy system depending on space (e.g., charcoal, water mills, working animals; see Debeir et al. 1989, 1991).

A new system, the exact opposite of the former, permitted the breakthrough of the Industrial Revolution. Since then, energy has been gained mostly from fossil fuels, each with an increasing level of energy content, from coal to uranium. The resources are exploited at a restricted number of locations. From there, highly concentrated energy can be distributed ubiquitously by modes of transport fed by the same fuel. Finally, with electricity and the internal combustion engine it is possible to transform one form of energy into another and this at any point on the earth, as with a diesel generator in an isolated Amazon village. In the past, the target was to concentrate a low energy harvest from a limited space. Now, the main purpose is to maximize energy efficiency, which means overcoming with minimum loss the distance between production and consumption.

3. *The Subjects and Profile of a Geography of Energy*

Of course, the geography of energy subjects are characterized by great variety, for example: the energy

chain of the resources (e.g., uranium: from pitchblende mining to nuclear power station); the original energy systems of different countries (e.g., the interrelationship between resources, huge distances, and import-export in Canada, see Chapman 1989); the impacts of energy policy on space (e.g., supply in eastern Germany after reunification), or specific problems in the Third World (e.g., firewood in the Sahel). Of topical interest to the current situation is the competition between nuclear power, fossil fuels, and renewable energies, with regard even to global change.

The geography of energy is often considered as a part of industrial geography (see *Industrial Geography*). Thus, Chardonnet's work *Géographie Industrielle* (1962) includes a volume on *Les sources d'énergie*. Indeed, the production and transformation of energy comprise industrial forms of processing; a power station is as much an industrial plant as a steel mill. In terms of location factors, energy production is similar to industrial production. Nonetheless, even in some publications entitled 'Geography of energy,' such as George (1950), a specific profile is missing. As in industrial geography, these publications focus too much on the location problem, leading an individual energy resource like coal to be treated as if it were a branch of industry (see George 1950, Sevette 1976). In contrast to these earlier approaches, more modern concepts have also been developed: Fernie (1980), Curran (1981), and Chapman (1989), for instance, insist on the dominant influence of policy, and Chapman shows a new methodical approach oriented by the energy chain. Energy is also being increasingly judged as an important sector in geography (see Calzonetti and Solomon 1985). Nevertheless, the differences between the energy and industry sectors in structure, procedures, and interaction with spatial and socioeconomic conditions have not yet been sufficiently elucidated. There are at least five fundamental differences:

(a) Let us begin with the energy chain. In industrial processing too, there is a sequence of stages between the extraction of raw materials and the production of finished products. Industrial processing, however, lacks the imperative cohesion between the beginning and the end: one can produce an unlimited variety of goods out of steel, even in so-called 'footloose industries' (industries without special links to location). In contrast to this, moreover, all 'energy resources' (except for the raw materials for the chemical industry, like oil) are entirely transformed and conducted to an energetic end use, passing by stages on exactly calculated locations. Furthermore, unlike industrial products, consumed energy cannot be recycled.

(b) Three types of energy are line-bounded: electricity, district heat, and gas (except liquefied natural gas). Due to the varying demand during the day, the week and the year, but also according to changing economic trends, the load curve fluctuates continuously. Therefore, line-boundedness requires large, even supranational organizations for transport and distribution (e.g., the European 'Union pour la Coordination du Transport de l'Electricité') which are not needed by industry.

(c) Whereas industry produces innumerable goods, energy is generated and used in only very few forms. But as it is basic to all human activities, there is an extraordinary demand for energy resources. They are extracted in large amounts, transformed in huge facilities (e.g., power stations, refineries) and shipped by bulk transport systems over great distances (e.g., tanker, pipelines, power grid). Only very big utilities, if not state companies, can afford that.

(d) Enterprises of such dimensions are both the very sought-after political instruments of the governments and competing poles of political power themselves. While everyone feels concerned by energy policy (e.g., taxes, petrol price, and environmental protection), there is no comparable interest in an 'industrial policy.' In principle, inter-relations between policy and the energy sector are much tighter than with industry, as political power is always established on the control of energy. Political influence is omnipresent, on and between the different levels: the tensions regarding the deregulation of the European electricity market concern local and regional utilities as well as state companies (e.g., ENEL in Italy) and even the European Union.

(e) Only parts of the energy sector belong to the secondary sector: for instance, the preparation, processing, and transformation into electricity, coke or petrol. Simple mining, however, belongs to the primary sector, whereas transport, distribution, and supply are parts of the tertiary sector. Such a vertical structure, from the primary to the tertiary sector, is valid for each energy resource, from oil exploration to the petrol station, as well as for the big companies themselves.

It is true that these characteristics give an individual profile to the geography of energy. Nevertheless, the latter does not figure among the traditional fields of economic geography, in the manner of agricultural or transport geography, which are oriented by one of the three economic sectors. The geography of energy should rather be conceived of as an interdisciplinary area without sharp limits, focusing on the individuality of the energy sector. This individuality is determined by the fact that it embraces all three economic sectors, remains steadily under strong political influence, and interacts with space along the energy chain.

4. *Renewable Energies—To a New Energy System 'from Space'?*

The interrelationship between energy and environment is one of the central topics in current geographic research. Undoubtedly, the energy chains' cumulated emissions are by far the most serious cause of

worldwide ecological damage. Simultaneously, there is rising worry about the dwindling supply of resources. Against this background, we are witnessing a struggle concerning the forms of energy use (e.g., nuclear power), the markets (e.g., deregulation) and last but not least, the available reserves (e.g., recent wars in the Near East). Against this background, too, the need to increase the use of renewable energies is urgent. That has already caused conflicts between the established forces of fossil fuel economy and the pioneers of a new sustainable energy system (see Scheer 1999). Conflicts are also taking place on a second level. On the one side, the traditionally centralized, very lucrative supply systems based on sophisticated energy chains are easy to control by large enterprises or governments. On the other side, renewable forms of energy can be produced and used at the same location, as for instance through wind or solar photovoltaics on a remote farm. As there is no need for an energy chain, there is neither a basis for control or for much profit, wind and sunshine being free of charge and taxes.

It is too early to outline future scenarios. Possibilities for radical change are in the offing, however, not only in the energy economy itself but also in business structures and policy, with conflicts from the local to the global level. For the established companies, decentralized production and consumption of renewable energies would not only mean ruinous losses, but also a step out of the energy chain, and hence out of influence. Furthermore, the state would lose an instrument of political control. Thus it is not surprising that the traditional system is already adapting to a new one. The large companies, the oil multinationals at the head, are integrating renewable energies into their activities. They want to remain on the top technological level and to prevent a shortage of reserves. In addition, they are trying to conserve their business structures, their networks and, in this way, their control over the system. This is manifested too, in the resistance of large producers to decentralized power generation, as in the example of the resistance in Germany to providing an adequately high level of reembursement for feeding surplus energy from wind generated electricity into the power grid. There are obvious tendencies, too, to recentralize those dispersed microstructures, for instance, by assembling numerous wind turbines in Denmark in 'windparks'. This is also evident in gigantic plans to use nonpolluting hydrogen in densely populated industrialized countries (see Schaefer et al. 1997): The electricity needed for producing hydrogen by electrolysis should be generated by a renewable energy, too—why not by water power in Canada? From there, hydrogen could be shipped to specialized filling stations in European conurbations—via a new type of energy chain.

The transition from one energy system to the next is always slow and fluid; one is adopted without the former being completely abandoned. Preindustrial ways of using energy from biomass or muscles continue to be practiced, although almost exclusively in the Third World. Today we are again witnessing the signs of such a transition to a new energy system: a fundamental change from the rapid and definite exploitation of cumulated, but exhaustive reserves to the sustainable use of resources that are continuously 'recharged,' directly and indirectly, with solar energy. In a certain way, it is a return to a system of harvesting energy from space, as the cereals were for Roman slaves. It will be generated in small quantities at numerous dispersed locations, although with different methods and different degrees of impact on space than before the Industrial Revolution. The availability of resources as well as environmental constraints and policy will decide in the future to what extent and over which period of time fossil fuels can or must be replaced by renewable energies, and centralized generation be replaced by decentralized generation, respectively.

Bibliography

Brücher W 1997 Mehr Energie! Plädoyer für ein vernachlässigtes Objekt der Geographie. [More energy! A plea for a neglected subject of geography]. *Geographische Rundschau* **49**: 330–5

Calzonetti F J, Solomon B D (eds.) 1985 *Geographical Dimensions of Energy*. Reidel/Kluwer, Dordrecht, The Netherlands/Boston/Lancaster, UK

Chapman J D 1989 *Geography and Energy. Commercial Energy Systems and National Policies*. Longman/Wiley, New York

Chardonnet J 1962 *Géographie Industrielle. I: Les Sources d'Énergie*. Sirey, Paris

Curran D W 1981 *La Nouvelle Donne Énergétique*. Masson, Paris

Debeir J-C, Deléage J-P, Hémery D 1986 *Les Servitudes de la Puissance: Une Histoire de l'Énergie*. Flammarion, Paris

Debeir J-C, Deléage J-P, Hémery D 1989 *Prometheus auf der Titanic: Geschichte der Energiesysteme*. Campus, Frankfurt, Germany, New York and Maison des Sciences de l'homme, Paris

Debeir J-C, Deléage J-P, Hémery D 1991 *In the Servitude of Power: Energy and Civilization through the Ages*. Zed Books, London

Fernie J 1980 *A Geography of Energy in the United Kingdom*. Longman, London/New York

George P 1950 *Géographie de l'Énergie*. Génin, Paris

Manners G 1964 *The Geography of Energy*. Hutchinson, London

Schäfer K, Soyez D, Dörrenbächer, H P 1997 Kanadischer Wasserstoff für Deutschland. [Canadian hydrogen for Germany.] *Geographische Rundschau* **49**: 368–73

Scheer H 1999 *Solare Weltwirtschaft. [Global Solar Economy.]* Kunstmann, München, Germany

Sevette P 1976 *Géographie et Économie Comparée de l'Énergie*. Institut Economique et Juridique de l'Energie, Université de Grenoble, Grenoble, France

Ter Brugge R 1984 Spatial structure in relation to energy production and consumption. *Tijdschrift voor Economische en Sociale Geografie* **75**: 214–22

W. Brücher

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Energy Policy: International

Energy policy is a special sector of policy dealing with energy problems. Energy (i.e., the capacity to realize work) is an essential ingredient of the activity of human beings. The minimum average energy requirement for an adult human to stay alive is about 1,000 kilocalories per day. For an adult engaged in normal activities, it is about 2,000 kcal a day. To give an idea of the amount of energy involved in human activity, 2,000 kilocalories is the energy contained in a cup of petroleum; this is approximately the amount of energy the human body extracts from 1 kilogram of food which one eats every day.

The stages of development of man from primitive man (one million years ago) to today's technological man can be closely correlated with energy consumption as indicated in Fig. 1, which shows daily consumption of energy per capita for six stages in human development.

Primitive man (East Africa about 1,000,000 years ago), without the use of fire, had only the energy of the food he ate (2,000 kcal a day).

Hunting man (European about 100,000 years ago) had more food and also burned wood for heat and cooking.

Primitive agricultural man (Fertile Crescent in 5000 BC) was growing crops and used animal energy.

Advanced agricultural man (Northwestern Europe in AD 1400) had coal for heating, water power, wind power and animal transport.

Industrial man (in England in 1875) had the steam engine.

Technological man (in the US in 1970) consumed 230,000 kcal/day.

Starting with the very low energy consumption of 2,000 kcal per day, which characterized primitive man, energy consumption grew, in one million years, to almost 250,000 kcal per day. This enormous growth of per capita energy consumption was only made possible by:

(a) increased use of coal as a source of heat and power in the nineteenth century;

(b) use of internal combustion engines that led to the massive use of petroleum and its derivatives; and

(c) electricity generated initially in hydroelectric sites and afterwards in thermoelectric plants.

The world average annual per capita energy consumption is equivalent to 1,500 kg of petroleum equivalent; with a population close to 6 billion people, this corresponds to almost 9 billion tones of oil equivalent, valued at approximately 1 trillion dollars.

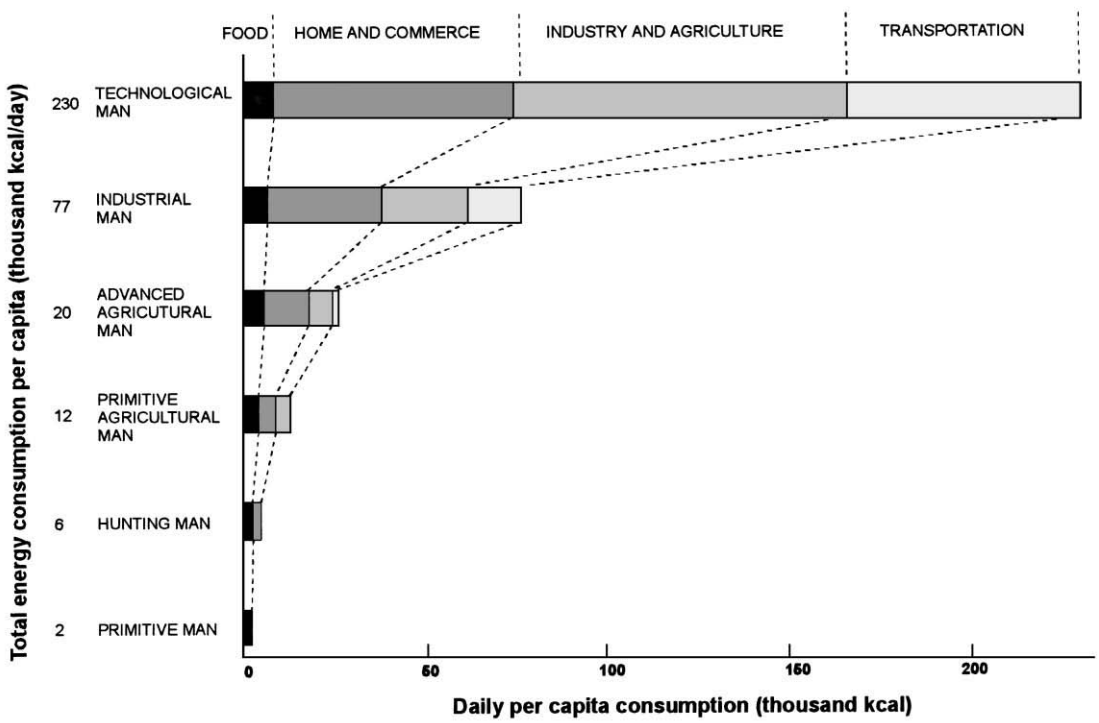

Figure 1
Stages of development and energy consumption

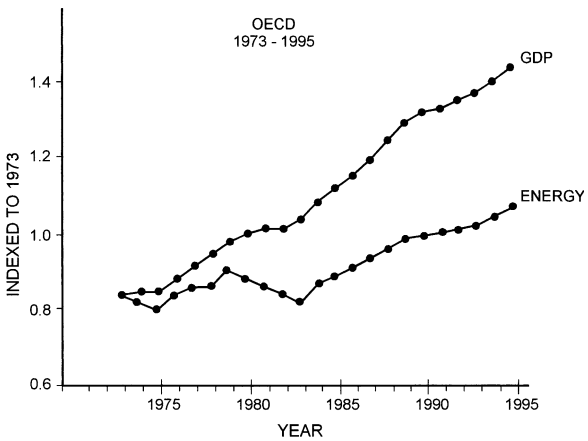

Figure 2
Evolution of GDP and energy in the OECD

To guarantee the supply and consumption of such huge amounts of energy is the subject of energy policy. In addition to that, there is a strong relationship between energy consumption and environmental degradation.

The energy–environment connection has been the object of many studies and it is sometimes possible to establish a cause and effect relationship between energy use and environmental damage. In 400 BC, for example, Plato mourned the lost forests, described by Homer centuries before, that had covered the barren hills of Greece. In this particular case, it was the use of wood, mainly for shipbuilding and forges to produce weapon that led to the destruction of the ancient Greek forests. A more recent example is the land degradation and desertification taking place in some areas of Africa because of the non-sustainable use of trees as fuel wood.

To harmonize energy consumption and environmental concerns is another facet of energy policies. Fossil fuel energy resources, such as coal and gas, are distributed unevenly across continents and countries. As a consequence, energy trade is an extremely important activity that involves not only the setting of prices, but of policies that regulate such trade. If we take the case of oil, more than 30 percent of the world's supply comes from the Middle East, which led to the formation of a cartel of countries that were the main petroleum exporters (OPEC). This cartel was able to raise the price of oil by a factor of 4 between 1973 and 1979, with very serious consequences for the world economy.

As a result of this crisis, several countries adopted national energy policies to protect themselves from undesirable dependence on Middle East oil imports; France, for example, dramatically expanded its nuclear program for the generation of electricity to replace coal and oil as fuels. The United States under President Nixon launched programs to reduce Middle East oil imports and President Carter introduced the expensive synfuels program to produce high-grade fuels from shale and coal, which were abandoned

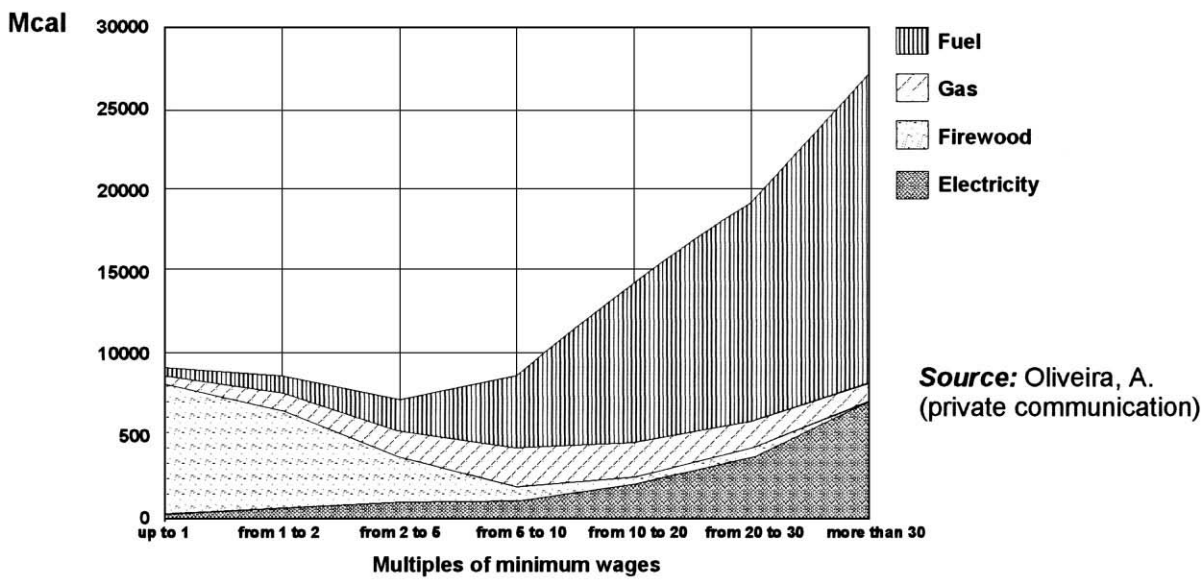

Figure 3
Average energy demand by income segment (Brazil)

when the OPEC cartel lost its strong hold on oil supplies.

In the United States and Western European countries, another important energy policy launched was 'energy conservation,' which is a shorthand description for a series of technical fixes leading to a more rational use of energy. In the past, conventional wisdom led us to think that the growth of national income of a given country or region occurs through an increase of population, number of houses, automobiles, and domestic appliances with a consequent proportional increase in energy consumption. Such a perception proved to be wrong.

A number of 'rationalization' measures adopted by OECD countries after 1973 were very successful in 'decoupling' energy consumption and GDP (Gross Domestic Product) taken usually as a measure of economic growth (see Fig. 2).

Such uncoupling, at the aggregate national level, does not address the question of how different groups in society consume energy and what type of energy, which can be quite different as income grows. Figure 3 shows average energy demand by income segment in Brazil.

To address such problems, one needs special energy policies, such as subsidizing energy for the poor. Particularly conspicuous in the last few years are energy policies to address environmental concerns; such problems can be local, regional, or global.

1. Policies that Promote Sustainable Energy Futures

Energy policies to reduce local and regional environmental impacts have been in effect for a long time. Local pollution has to do with local governments, since it deals with clean air, fresh supplies of clean water, the removal and disposal of solid wastes and liquid effluents, street cleaning, etc. However in many developing countries, a large fraction of the population still lives among the rubble and residues that it produces, due to the lack of resources to remove waste and to build sewers and engineering works needed for the supply of water.

Regional pollution is caused mainly by automobiles, energy production, and heavy industry, which are inherent to more prosperous societies. Large cities and adjoining areas, such as Los Angeles, Mexico City, and São Paulo, have long been 'suffocating' under the pollution caused by the emissions and smog resulting from the burning of fossil fuels. Sometimes the amount of pollution produced is large enough to cause regional and even trans-border problems, such as 'acid rain,' some of which originates in the USA but is responsible for the destruction of life in the Canadian lakes. The same has happened to lakes in Scandinavia, due to industrial activities on the south of the Baltic Sea. Regional pollution has to be dealt with at the state or national level and eventually among a number of countries.

Global environmental problems and concerns are more recent, the most conspicious of them being the destruction of the stratospheric ozone layer by CFCs and the 'greenhouse effect.' These problems result from changes in the composition of the atmosphere and have little to do with national borders. The causes of such global problems are gases originating anywhere in the world and the problem can only be tackled at the international level.

The ozone layer problem was succesfully resolved by a concerted effort of a few countries that led to the Montreal Protocol. Subsequent decisions are effectively leading to the phaseout of CFCs from the face of the Earth. Other substances have been developed to replace them. The outstanding problem is the 'greenhouse effect' that is proving to be much harder to solve since the most important greenhouse gas is CO_2 (carbon dioxide), the unavoidable result of the combustion of fossil fuels that is responsible for 75 percent of all energy consumed in the world today.

The basic policy in this case is to reduce CO_2 production through improved efficiency in energy use and to search for proper replacements such as renewable energy sources. Appreciable success has been achieved in industrialized countries to reduce CO_2 emission mainly through rationalization in the use of energy. CO_2 emissions in those countries have levelled off since 1980 or are growing very slowly since then. In developing countries, however, emissions are growing at a fast rate.

The main policy instrument to solve this problem is the Framework Convention on Climate Change (FCCC) signed in 1992 during the Rio Summit and the Kyoto Protocol (1997) that adopted a binding resolution according to which industrialized countries, as a whole, would reduce their CO_2 emissions by 5.2 percent below the 1990 level in the period 2008–12. The United States accepted a reduction of 7 percent to their 1990 level. Since emissions have been growing by approximately 1 percent since 1990 in this country the reductions required would be much higher. Some estimates indicate that the reduction required will be approximately 25 percent.

In order to make it easier for industrialized countries to reduce their emission the Kyoto Protocol in addition to setting caps on emission for the industrialized countries has introduced features that could induce industrialized and developing countries to cooperate in reducing CO_2 emissions. They are the 'flexibility mechanisms' by which emission reductions in developing countries (or economies in transition, i.e., the former Soviet Union and countries of East Europe) can be counted as reductions in industrialized countries. In this category are 'Emissions Trading,' 'Joint Implementation,' and 'The Clean Development Mechanism.'

Emissions trading (ET) is the more general, market-based instrument. It enables achievement of limits at least cost by taking advantage of marginal cost differentials in emissions abatement among countries. Since greenhouse gases are a uniformly mixing global pollutant, both the damages from emissions and the benefit from emission reductions are independent of their origin. In order to minimize the cost of given global emission reduction, abatement should take place where the costs are lowest. (Emissions trading is defined in Article 16 bis and referred to in Article 3.10 and 3.11 of the Protocol.) It is restricted to trading among the Annex I countries (industrialized countries).

Joint Implementation (JI) limits emissions trading to developed countries that have accepted emission reduction commitments under the Kyoto Protocol, as listed in its Annex I. Countries that are certified to be below their national limits are allowed to sell greenhouse gas offsets to other Annex I countries who are above their own limits. The latter's motivation is the opportunity to meet their commitments at lower cost than through purely domestic actions (defined in Article 6, and referred to in Article 3.10 and 3.11 of the Protocol).

The Clean Development Mechanism (CDM) is an instrument for joint action on emissions reduction (and therefore trading) between the developed countries listed in Annex I of the Kyoto Protocol and developing countries, which are not listed in Annex I (covered in Article 12 and referred to in Article 3.12 of the Protocol). Article 12 of the Kyoto Protocol was the result of a political compromise. It conciliated two sharply opposing views that almost wrecked the Kyoto Conference in 1997. For some countries, an essential element for the Kyoto Protocol was emission trading based on adoption of mandatory reduction commitments by all countries. This feature was contained in a United States Senate Resolution, but was unacceptable to the developing countries. The creation of the CDM emerged as the solution.

In Annex I countries, there will be competition among the several mechanisms defined in the Kyoto Protocol to meet emission reduction requirements. The portion of resources allotted to each of them will prove to be an important internal problem to be resolved. The Kyoto Protocol left undefined what part of the emission reduction commitments could be met outside national borders through use of the flexibility mechanisms. A small percentage would encourage domestic action in Annex I countries, reducing the financial flow to developing countries. A larger one would discourage domestic actions and eventually tend to lead to an inordinate increase of emissions in Annex I countries. In all likelihood, the share for the CDM compared with other mechanisms will be determined by the market.

The Kyoto Protocol has not yet been ratified by the key industrial countries, but—if and when ratified—it might prove to be as effective as the Montreal Protocol in solving one of the most difficult global environmental problems of our time.

See also: Energy Economics; Energy, Geography of; Environmental Policy; Industrial Ecology; Industrialization

J. Goldemberg

Engineering Psychology

As its name implies, engineering psychology is a branch of psychology that is concerned with design. In particular, it seeks to ensure that human capabilities, limitations, and tendencies are considered in the design of things people use—from ordinary consumer products to sophisticated technologies—with the aim of maximizing performance and safety. The colloquial term 'user-friendliness' captures the essence of this goal, and in pursuing it, engineering psychologists draw upon and contribute to many domains of psychological knowledge.

1. Interdisciplinary and Disciplinary Relationships

Throughout its history, which dates back to the Second World War era, engineering psychology has struggled to establish a clear identity. Partly this is a result of its interdisciplinary perspective and partly it stems from a confusing array of labels with which it has been associated. To appreciate the nature and current status of the specialty, therefore, it is necessary to sort out these relationships.

The core concept that engineering psychology shares with specialists in other disciplines—including some in engineering, computer and cognitive sciences, biomechanics, medical sciences, architecture, and gerontology—is that of the human–machine system: the idea that human and machine function as partners in a collective enterprise rather than as independent operators. Whereas design professionals typically focus on the efficiency and reliability of the machine component, and personnel specialists, on the proficiency of the human component (see *Personnel Selection, Psychology of*), the system perspective emphasizes the importance of their interaction. A machine that is difficult to use will result in poor system performance no matter how well it functions as a machine.

For example, airplane accidents officially attributed to 'pilot error' were at one time quite commonly found to involve misreading the cockpit altimeter. Since this instrument provided highly accurate altitude infor-

mation and rarely malfunctioned, it was beyond reproach from an engineering standpoint. From a human–machine systems standpoint, however, it was seriously deficient in that altitude readings were easily confused. Hence if a pilot under stress mistook 1,000 for 10,000 feet, it was as much the fault of the display design as it was of the human operator. Fortunately, this problem was diagnosed by engineering psychologists and corrected through human-oriented design (Sanders and McCormick 1987). The moral: good design requires joint consideration of human and machine characteristics.

The approximately 11,000 professionals who subscribe to this philosophy are known as 'human factors specialists' or 'ergonomists,' and despite a variety of disciplinary backgrounds, their shared interest has evolved into a discipline of its own (Van Cott and Huey 1992). In the USA, it is most commonly referred to as 'human factors'; in the rest of the world, as 'ergonomics.' Since nearly half of these professionals were trained as research psychologists, engineering psychology represents a substantial component of this interdisciplinary field. Within the field of psychology, however, where the number of professionals is well in excess of 150,000 and growing, it constitutes but a tiny specialty area—one that few other psychologists could readily identify. Moreover, larger, well-established specialties such as cognitive, experimental, educational, environmental, and industrial/organizational, (I/O) psychology have also developed an interest in design applications. For this reason, it has been suggested that the engineering psychology specialty is in transition and may eventually be absorbed into the others (Howell 1993). Its relationships with cognitive, experimental, and I/O specialties are particularly noteworthy as we shall now see.

2. *Background and Changing Emphases*

Although it had many precursors, engineering psychology emerged as a recognized specialty following the Second World War. During the war, teams of research psychologists and engineers had been assembled to investigate—and correct—problems such as the excessive number of accidents in certain aircraft, the difficulty of maneuvering large ships, and the failure of radar operators to detect critical signals. Most of the psychologists recruited for this purpose were academic scientists whose peacetime work had involved studies of basic human functions such as sensation, perception, motivation, and learning. Within the field they were known as experimental psychologists because they relied chiefly on the experimental methods of science in their investigations.

The success of these wartime 'human engineering' efforts, as they were called, had four noteworthy consequences. First, it convinced the engineers and psychologists who had been thrown together to rectify specific design problems that the collaboration had much broader potential. The engineers discovered that psychological knowledge is useful for design; the psychologists discovered new applications and directions for their research.

Second, it prompted transition of the psychology–engineering collaboration into peacetime applications. Military human-engineering units continued their research and development (R&D) activities, and extended their influence by writing user-oriented requirements into contracts for new systems, and by supporting civilian researchers. Consequently, some large military contractors (e.g., aircraft manufacturers) and civilian industries (e.g., The Bell Telephone system) saw fit to form human-engineering groups of their own. These groups comprised mainly of psychologists who reviewed plans and prototypes, defined usability requirements, and conducted research as a basis for setting those requirements.

The third consequence of the wartime human-engineering effort was the effect it had on the subsequent work of the participating psychologists. When they returned to their respective laboratories, many of them continued lines of research stimulated by their wartime activities, often under military sponsorship. For instance, Cornell University psychologist James J. Gibson developed a whole new theory of perception based on his wartime experience, one that revolutionized thinking on how we construct a three-dimensional world out of a two-dimensional retinal image (Gibson 1966).

The final consequence was the institutionalization of the amalgam of psychology and design with the establishment of the Society of Engineering Psychologists and the interdisciplinary Human Factors Society, both in 1958. Specialized journals (e.g., *Human Factors, Ergonomics*), graduate training programs, textbooks, courses, research laboratories, and consulting firms followed.

Then as now, the content of engineering psychology was dictated largely by the nature of the prevailing technologies and the accompanying task requirements for human operators (Howell 1993). At mid-century, humans exercised manual control over most systems, hence the focus was on the design of information displays (mostly dials), manual controls, and their interaction. The most relevant psychological knowledge was that involving sensation, perception, motor functions, reaction time, and perceptual-motor skill.

Today's technologies present operators with the daunting combination of more information, less direct control, and greater responsibility. Often the human's task is that of monitoring complex, powerful, highly automated machines and acting only when something goes wrong or there is reason to alter the program—a role commonly referred to as 'supervisory control' (Sheridan 1992). Consequently the psychological emphasis has shifted to more 'mental' topics such as attention, cognition, mental workload, information-

processing capacity, decision making, and problem solving.

Throughout its history, engineering psychology has maintained an active involvement in both science and application. When the literature provides a plausible scientific basis for a design decision, that knowledge is incorporated; when it does not, it prompts new lines of research. Some of this research seeks specific answers to narrowly defined questions, such as whether to use menus or icons in a particular software program (see also *Human–Computer Interaction*). Some, however, addresses much broader, more fundamental issues—such as the nature and limits of human attention—and in so doing expands the body of psychological knowledge.

In this regard, it is interesting to note that the work of engineering psychologists played an important role in a comprehensive 'paradigm shift' that overtook psychology, starting in the 1960s. At a time when the dominant psychological philosophy (behaviorism) held that mental events are unsuitable subjects for scientific investigation because they are not open to public inspection, British engineering psychologist Donald E. Broadbent and his colleagues were gathering data on radar operators that would culminate in a general model of how humans process information (Broadbent 1958).

Broadbent's model, particularly his theory of attention, was thus instrumental in the 'revolution' that rekindled psychologists' interest in cognition (i.e., mental processes). So, too, were contributions such as the Tanner–Swets theory of signal detectability (Swets et al. 1961), which established that the seemingly mechanistic task of sensing things includes an important cognitive element (decision making); *Hick's Law* (Hick 1952), which described human reaction time in terms of the mental operations required to translate a stimulus into a response; and a number of others. Because engineering psychologists dealt with applied problems such as detecting radar signals, interpreting messages obscured by noise, following moving targets, and remembering strings of letters and numbers, they found it necessary—and useful—to explore the mental functions that underlay these skills, thereby challenging the tenets of behaviorism.

Whereas engineering psychology's origins and current orientation are closely linked to experimentation (hence the experimental and cognitive branches of the field), it also has ties with another specialty—I/O psychology—that evolved from a very different tradition: psychometrics, or the measurement of individual differences. The experimental approach seeks to understand how the typical human functions under a particular set of circumstances—in the present context, in order to design for the modal user. The individual-differences approach seeks to develop measures (e.g., tests, questionnaires) that enable one to understand and measure how people differ—in the context of I/O psychology, in order to select and place them in jobs that best fit their knowledge, skills, and abilities (see also *Personnel Selection, Psychology of*; *Job Analysis and Work Roles, Psychology of*). Simply put, the former tradition tries to fit the machine (or work setting) to the human; the latter, the human to the machine (or work setting).

Over the years, the two traditions have drawn closer together philosophically and methodologically in recognition of the fact that both derive from the same basic model (complex goal-oriented systems) and seek the same end (achieving a good fit between the system components). I/O psychologists now include experimentation among the tools they use in organizational research, and engineering psychologists recognize that measuring individual differences is important for design because one size does not fit all. Increased ethnic and gender diversity among today's pilots, for example, requires rethinking a number of the cockpit design features that were geared to the fairly homogenous male pilot population of yesteryear. And our aging general population requires considering age-related differences in the design of consumer products, home-care medical devices, highways, and working conditions.

What is perhaps surprising is that despite these converging trends and the occasional reference of one to the other in modern texts, engineering and I/O psychology remain far apart in their institutional existence (Ilgen and Howell in press). There is very little overlap between their respective membership in professional organizations, journal content, graduate training programs, research efforts, or job titles.

The one notable exception is the content area of training in which both specialties have a long-standing interest. Engineering psychologists are particularly interested in the design and evaluation of devices (e.g., simulators, computer-based instructional systems) used for developing skills, whereas I/O psychologists are concerned with a much wider array of training issues (e.g., from simple job training to team building and even management or organizational development—see also *Simulation and Training in Work Settings*). It is worth noting, however, that whatever the focus, training constitutes a third strategic leg on the system performance stool—a complement to the design and selection strategies favored by engineering and I/O psychologists respectively.

3. *Engineering Psychology Today: Examples*

Whether or not engineering psychology survives as an independent specialty, the philosophy that it represents has clearly established itself within psychology and, by whatever name, design-oriented research will continue to grow. Thus it is important to grasp the essential nature of such research, and that is best done through examples. Two illustrations are presented below: the first, an example of fundamental research that was prompted by a design issue; the second, an

example of the application of existing psychological knowledge to a design problem. These examples also illustrate methodologies typical of engineering psychology.

3.1 Mental Models

Many technologically advanced systems have failed to perform as effectively as they might because the logic built into the machine component either is not transparent to the user, or is inconsistent with his or her conception of it, or both. That is, the 'artificial intelligence' and the 'human intelligence' involved are not congruent.

For example, computer-based systems designed to educate students in a particular subject matter (so-called 'intelligent tutoring devices') have often failed because they could not adequately diagnose the thinking behind the learner's answers and hence could not provide feedback in a useful form. And highly automated 'process-control' systems, such as those built into nuclear power plants, have encountered problems when something went awry because the humans responsible for diagnosing the fault were not sufficiently familiar with how the machine executed its control function. This was one of the many factors that contributed to the highly publicized accident at the Three Mile Island plant in 1979 (Rubinstein and Mason 1979).

Overcoming this limitation in the design of intelligent tutors, process-control automation, and other such systems cannot be achieved without an understanding of how people conceptualize the underlying physical or logical processes. Realization of this fact has led to a great deal of research on the mental representation of such structures, commonly referred to as 'mental models,' some of it aimed at specific applications and some at teasing out general cognitive principles.

A key requirement for either endeavor, of course, is devising ways to access private mental content—to describe these mental models and validate them. And that, in turn, has produced a number of new 'knowledge elicitation' and 'cognitive task analysis' techniques that are helping us gain a deeper understanding of human cognition (Cooke 1992). Some, known as 'process tracing' methods, involve collecting verbal protocols in a systematic way from individuals while they are performing a task, and then analyzing those protocols to infer the underlying thought processes. Others, such as the 'Pathfinder' technique and a collection of so-called 'policy-capturing' methods, deduce these processes using sophisticated statistical analysis applied to a lot of relatively simple judgments or choices required of the operator.

One technique, known as SAGAT, was designed explicitly to capture the operator's mental picture (or situation awareness') at various stages in the performance of highly complex tasks such as air-to-air combat or command decision making. This approach involves introducing systematic 'knowledge probes' at various points in the conduct of a simulated exercise, thereby piecing together the evolution of the operator's mental model (Endsley 1995).

Taken together, therefore, the design problems posed by the growing complexity of today's systems, plus the frequent disconnects between human and machine modes of thinking, has given rise to new insights into how the mind works as well as new methods for extracting that information.

3.2 The TADMUS Project

In 1988 a civilian airliner was mistakenly shot down by the USS Vincennes, a cruiser equipped with the latest intelligence technology, and the incident prompted international concern. It also prompted a Congressional investigation and a commitment by the US Navy to diagnose and rectify whatever went wrong. For many years, the Navy's Office of Naval Research (ONR) had been supporting fundamental research on human decision making—if fact, had been its leading sponsor. Therefore, ONR accepted the challenge of applying this vast body of knowledge to the general situation involved in the Vincennes incident, and the Tactical Decision Making Under Stress (TADMUS) project was born.

The aim of this project was to examine everything that was known about how humans make decisions; how their performance is affected by the complexities, uncertainties, and urgency (i.e., the stresses) typical of military 'command and control' operations; and how to ameliorate the effects of stress.

Needless to say, this ambitious goal called for—and received—a substantial investment in both time and money: more that $10 million was committed over an eight-year period, and the results were impressive. Drawing on basic cognitive literatures, hypotheses were developed and tested in a high-fidelity simulation of the shipboard Combat Information Center (CIC), and promising innovations were evaluated by operational personnel in research conducted at training sites, and even in actual shipboard demonstrations (Cannon-Bowers and Salas 1998).

At the end of the 1990s, TADMUS-based concepts on how to train teams to perform this vital function and sustain their performance, especially in the face of stress, were being incorporated into training curricula, and proven design improvements were being considered in planning the next generation of CIC interface equipment. The interface innovations include better (user-oriented) display of the massive amount of information available to decision-makers, and machine 'decision aids' to help them process this information rapidly and accurately. The literature on 'mental models' played an important role in both the training and design innovations.

In sum, the TADMUS project, mounted to address an urgent problem, has brought a wide variety of psychological knowledge to bear on that problem and, as a result, is virtually certain to result in improved CIC performance in the future. Moreover in the process, new knowledge was gained about team training, display for 'situation awareness,' decision aiding, and a number of other generic issues.

Viewed together, these two examples—mental models research and the TADMUS project—should provide ample material from which to construct a valid mental model of the specialty of engineering psychology as it exists today. Both its enduring features (the goal of improving human–machine systems, the bidirectional linking of design issues with psychological research, the multidisciplinary perspective, the driving role of technology) and its current focuses (cognitive functions; complex, heavily automated systems) are vividly illustrated. Additional sources of general information on engineering psychology can be found in two widely used texts (Wickens 1992, Sanders and McCormick 1987), the *Handbook of Human Factors* (Salvendy 1987), a chapter in the *Handbook of I/O Psychology* (Howell 1991), and chapters devoted to the topic that have appeared periodically in the *Annual Review of Psychology* (see Howell 1993).

See also: Human–Computer Interaction; Job Analysis and Work Roles, Psychology of; Personnel Selection, Psychology of; Simulation and Training in Work Settings

Bibliography

Broadbent D E 1958 *Perception and Communications*. Pergamon Press, London

Cannon-Bowers J A, Salas E (eds.) 1998 *Making Decisions under Stress: Implications for Individual and Team Training*. American Psychological Association, Washington, DC

Cooke N J 1992 Eliciting semantic relations for empirically-derived networks. *International Journal of Man–Machine Studies* **37**: 721–50

Endsley M R 1995 Toward a theory of situational awareness. *Human Factors* **37**: 32–64

Gibson J J 1966 *The Senses Considered as Perceptual Systems*. Houghton-Mifflin, Boston, MA

Hick W E 1952 On the rate of gain of information. *Quarterly Journal of Experimental Psychology* **4**: 11–26

Howell W C 1991 Human factors in the workplace. In: Dunnette M D, Hough L M (eds.) *Handbook of Industrial and Organizational Psychology*. Consulting Psychologists Press, Palo Alto, CA, Vol. 2, pp. 209–69

Howell W C 1993 Engineering psychology in a changing world. *Annual Review of Psychology* **44**: 231–63

Ilgen D R, Howell W C in press The National Research Council Committee on Human Factors (or human factors: It's more than you think). *The I/O Psychologist* **36**

Rubinstein T, Mason A F 1979 The accident that shouldn't have happened: An analysis of Three Mile Island. *IEEE Spectrum* November: 33–57

Salvendy G 1997 *Handbook of Human Factors*, 2nd edn. Wiley, New York

Sanders M S, McCormick E J 1987 *Human Factors in Engineering and Design* 6th edn. McGraw-Hill, New York

Sheridan T B 1992 *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, Cambridge, MA

Swets J A, Tanner W P, Birdsall T G 1961 Decision processes in perception. *Psychological Review* **68**: 301–40

Van Cott H P, Huey B M (eds.) 1992 *Human Factors Specialists' Education and Utilization*. National Academy Press, Washington, DC

Wickens C D 1992 *Engineering Psychology and Human Performance*, 2nd edn. Harper Collins, New York

W. C. Howell

English Revolution, The

The term 'English Revolution' refers to the profound changes experienced by the English political system from 1640 to 1660. Generations of historians have labeled this as one of the decisive political events of modern European history; at the same time it connected the social, political and religious changes of the sixteenth century with those occurring after 1660, up to the 'Glorious Revolution' of 1688–9.

1. Outline of the Debate on the English Revolution

Contemporaries such as the Earl of Clarendon, as a rule, did not use the term revolution (see instead he wrote *The History of the Rebellion and Civil Wars in England* (Macray W D (ed.) 6 vols. 1880, repr. London 1958). Although a generation later the word was added to the political vocabulary, it was applied to the events of 1688–9, that is, the overthrow of the Stuart monarchy, the succession of William and Mary and the new constitutional settlement. It was a foreigner, Francois Guizot, who, after the French revolution, was the first to refer to the events of the years 1640 to 1660 as the 'English Revolution' in his *Histoire de la république d'Angleterre et de Cromwell* (2 vols. 1854). At this time, liberal British historians such as Hallam and Macaulay were still describing those years as an untidy interruption in the process of the evolution of parliament and liberty, which had reached its first climax with the 1689 Bill of Rights. Then S. R. Gardiner developed his concept of the Puritan revolution (*The First Two Stuarts and the Puritan Revolution*, London 1876). His magisterial *History of England from the Accession of James I to the Outbreak of the Civil War 1603–1642* (10 vols., repr. New York 1965), which he later continued in his *History of the Great Civil War 1642–1649* (3 vols. London, 1886–91) and his *History of the Commonwealth and Protectorate 1649–1656* (3 vols. London 1894–1901), placed these

years firmly into the context of the heroic struggle of the English House of Commons against arbitrary monarchy. This Whig interpretation of early modern English history dominated for nearly a century, despite the vigorous attacks of Marxist historians such as Christopher Hill, who understands the English revolution as the violent replacement of a feudal society by a new capitalist order. This enhances the importance of the events which Hill regards as 'comparable in many respects with the French Revolution of 1789' (*The Century of Revolution 1603–1714*, Edinburgh 1961). Thus in Whig and Marxist accounts alike the revolution in general and the parliamentary cause in particular were progressive and dynamic, and thereby contributed to the patriotic myth of the distinctiveness and exceptionalism of English political tradition.

Such comprehensive views of a wide-ranging English revolution have been challenged fundamentally by recent scholarship. This new revisionism disputes the notion of an inevitable, long-term historical process of unfolding progress, culminating in the revolutions of the seventeenth century. Detailed research on events rather than trends, often concentrating on the regions rather than the capital, not only pointed to chronological fissures and contingencies, but also deconstructed the concept of a progressive parliament as the main historical agent. Every postrevisionist assessment of the English revolution must take care not just to repeat old paradigms, to avoid anachronistic conceptualization and to pass judgment on the basis of varied and often contradictory evidence.

2. *Origins and Causes*

Up to Stone's magisterial summary of *The Causes of the English Revolution* (1986) these have been placed into the context of the major economic and social changes of sixteenth-century England: the long-term inflation of the so-called price revolution, the secularization of the massive properties of the church in the wake of the Reformation, and new developments in trade and industry. All this contributed to the destabilization of the traditional social fabric, that is, to a temporary decline of the aristocracy and to a rise of the gentry wherever its members started to exploit their estates by the rules of an incipient capitalism. But no direct links between economic and social change on the one hand and the English revolution on the other have yet been established. Thus the latter cannot be labeled an early bourgeois revolution.

2.1 *The Long-term Crisis of the Constitution*

The political philosopher James Harrington (1611–77) attributed the outbreak of the civil war of 1642 to 'the dissolution of government' and this opinion has stood the test of time, remaining valid in the light of modern research. Since the Middle Ages the English monarchy has been based on the consent and co-operation of the political nation, that is those who, like the aristocracy, had an interest in participating in their country's politics, and possessed the means to do so. In this relationship parliament had served as the point of contact, giving its assent to the legislation of the crown and voting taxes to support its policies. This vital constitutional role had been confirmed and even strengthened during the Reformation, which had been effected by the King in Parliament. But towards the end of the sixteenth century this relationship began to disintegrate. The political elite was no longer willing to shoulder the rising financial and bureaucratic burdens of effective royal government, which represented the emergence of the early modern European state on the road towards absolutism. The structural weakness of the English monarchy became evident, when in the early years of the reign of Charles I, the more localist-minded members of parliament refused to meet the rising costs of war and thereby prevented the king from acting as a major player in the field of European politics in the age of the Thirty Years War. The crown, trying to bypass parliament's assent to taxation, mounted a constitutional offensive, which provoked a crisis of the constitution. Whereas the king and his advisers stressed the unlimited powers of the monarch, who ultimately was answerable only to God, the Commons accused them of trying to abolish the ancient English constitution and claimed to be the last bulwark of liberty. The conflict reached its climax when parliament forced Charles I to assent to the *Petition of Right* (outlawing nonparliamentary taxes, arbitrary imprisonment, and the billeting of troops) in 1628. Then, in 1629, the king dissolved parliament and declared that he would not call it again, but establish a government of personal rule.

2.2 *Conflict over Religion*

Historians always have agreed that religion was the cardinal issue of the English revolution, and John Morrill has recently described the English civil war as the 'last great war of religion.' With the final triumph of the Protestant Reformation, anti-Catholicism became the distinctive mark of English political culture. In the age of counter-Reformation Popery, especially in alliance with monarchical absolutism, was regarded as a constant threat to the ancient liberty of an English nation still haunted by the memory of the Armada. In the field of religion, transcending the limited sphere of the political nation, a wider public could be addressed and even mobilized. Contesting Popery promoted activism at all levels of society.

In this context the politics of religion implied momentous consequences, and contributed to the aggravation of tensions even more than the constitutional offensive of the crown. The king, as head of

the English church, gave the impression of undermining the Calvinist foundations of the Anglican faith when he obviously favored a novel Arminian stance in theology and worship. At the same time, under the protection of his Catholic queen, Catholicism gained influence at the royal court. Soon many contemporaries believed that their king was, if not the author, then at least the instrument, of a comprehensive Popish plot to subject England to papal rule again.

2.3 The Multiple Kingdom

Historians in search of the origins of the English revolution have only recently stressed the British dimension of the conflict. Since the accession of James VI of Scotland to the English throne in 1603, a regal union of both kingdoms had existed, and since 1541 the English monarch had also held the title of King of Ireland, reflecting the fact that the island had been an English colony since the Middle Ages. In addition to their different social and economic structures, the three kingdoms were separated by contrasting religious establishments, with the result that the repercussions of the English crown's religious policy were enhanced. When Charles I tried to adapt a radical Calvinist Scottish church to the doctrines and organization of the English Anglican Church, he provoked the armed rebellion of his Scottish subjects in 1639. This was to be an important precondition for the formation of effective English opposition to Charles I. In the face of English pressure to introduce the Reformation, the Irish had remained staunch Roman Catholics. In 1641, in view of the conflict between the king and the English parliament, they seized the opportunity to take up arms against their oppressors. This contributed decisively to the outbreak of the English civil war, which might also be called the war of the three kingdoms.

2.4 Charles I—the Personal Factor

The English revolution was not an accidental event because it took place in a political and cultural system, which could not stand crisis and strain. Nevertheless, its outbreak and course were also determined by certain elements of contingency, in the first place, the personality of king Charles I. Revisionist historians in particular have exposed the numerous political blunders of his constitutional and ecclesiastical policies, which finally led to the breakdown of the monarchy. This is, thus, presented more as a contingent event in itself than as the result of structural conditions and long-term developments. Yet his personality and his aims and actions were neither erratic nor contradictory. Right up to his end on the scaffold, they were linked to one single great purpose: the establishment and preservation of a monarchy, in which the king's powers would be limited neither by law nor by consent, because the ruler is answerable only to God. He always stood for absolutism and thus the quarrels he provoked raised fundamental issues, which form the subject matter of revolutions.

3. From Resistance and Rebellion to Revolution and Restoration

3.1 Stages and Turning Points

In 1639 the Scots rose against the ecclesiastical policy of Charles I, forcing the king to abandon his personal rule and to recall parliament in order to finance first the war and later the costs of military defeat. The unpopularity of his policies meant that he could command no following, either in parliament or elsewhere, and therefore had to give his assent to those statutes and measures which the opposition in parliament regarded as the necessary safeguards against royal absolutism and popery. They abolished the instruments of arbitrary rule and unparliamentary taxation, secured the position of parliament, and had the king's chief adviser, the Earl of Strafford, executed. This policy of enforced reforms changed from a defense of the traditional political order into a violation of the ancient constitution as soon as the leaders of parliament had to take precautions against a potential royal backlash and proposed to abolish the king's right to appoint ministers and control the army. Revolutionary aims like these soon alienated conservative members of the political nation, provided the king with a party, and finally opened the way for the escalation of the constitutional conflict between crown and parliament into civil war between Cavaliers and Roundheads (1642–5).

But even the military victory of parliament, which commanded superior resources, did not put an end to the conflict with the monarch. In the eyes of the majority of the nation there was no alternative to monarchy, and so even after defeat Charles was never willing to assent to, or even compromise on, the reduction of his royal prerogative. He counted on divisions among his adversaries and the second civil war in 1648 seemed to vindicate him, when risings of the royalists who were now supported by the Scots had to be put down. The confusion and turmoil of those years gave rise to and accelerated a process of political radicalization. The victorious army with its leading general, Oliver Cromwell, became the center of power, purged the House of Commons of those who still favored negotiation, and staged the trial and execution of the king (January 30, 1649). Together with the abolition of the House of Lords and the proclamation of March 17, 1649, in which parliament declared 'that the office of a king in this nation … is unnecessary, burdensome, and dangerous to the liberty, safety, and public interest of the people …' this marked the

climax of a political revolution in which the British monarchy was replaced by the republic of the Commonwealth.

At the same time new rifts opened between the more conservative 'Rump Parliament,' as it was called after the purge and the army which, in the field of religion, pressed for a 'godly reformation.' In 1653, Cromwell dissolved parliament by force and the Protectorate was established. Though parliaments still formed part of the new constitution laid down in the 'Instrument of Government', power rested with Oliver Cromwell, who became head of state as lord protector and his council. Thus, an only thinly veiled military dictatorship led the way back to the monarchy. After Cromwell's death in 1658 the revolutionary process came full circle when, with the overwhelming assent of the political nation, the Stuart monarchy was restored in 1660.

3.2 Forces and Factors

A common denominator of revolutions is the active role of the people, the masses applying 'violence from below.' Especially during the first stage of the English revolution, the populace of London repeatedly exerted pressure on the House of Lords and the king, whose assent was needed to sanction decisive reforms. They acted in accord with the leaders of the opposition in the House of Commons, who were able to organize mass demonstrations and petitions through a network of Puritan congregations. And when in December 1641 parliament published its program for radical political and ecclesiastical reform in the so-called 'Great Remonstrance,' this revolutionary move testified to the new revolutionary alliance between the representatives of the nation and a wider public. More than 20,000 sermons, pamphlets, and newsletters were printed between 1642 and 1649, opening up a public sphere for a general debate on religion and politics. New radical forces emerged, such as the popular democratic movement of the Levellers. This movement confronted parliament with a wide range of grievances, and, in its *Agreements of the People*, presented its drafts for a new written constitution, which was to guarantee every freeborn Englishman's birthright. And when the victorious army became a major factor in politics, this was also affected from below. Common soldiers elected agitators as their representatives to the council of the officers and declared that they were not a 'mere mercenary army,' but had the right of political participation.

In the end, the revolutionary regime was based on the minority in Cromwell's army that took an active interest in politics. This points to the extent to which, in consequence of progressive radicalization, the popular basis of support for the revolution was shrinking. But at the same time it was becoming more forceful, especially when linked to radical Puritanism, which, through the doctrine of predestination, could provide its members with the self-confidence that it takes to want to turn the world upside down.

4. Achievements

Because they refuse to take a long-term perspective and sometimes even try to separate the period of resistance and the civil war from the republican experience, many revisionists do not concede that the English Revolution in the long run secured the place of parliament within the framework of the constitution. But after a second attempt to establish absolutist rule in England had been averted by the resistance of the political nation in the Glorious Revolution of 1688–9, the introduction of constitutional monarchy, which had been proclaimed as the policy goal of the parliament of 1642, was finally achieved. Though it did not effect substantial social change, the English Revolution had unfolded the complete repertoire of political change: not only the right of resistance but also the idea of the sovereign people sitting in judgment upon the king, and the theory of individual rights and social contract as the basis for a written constitution. Thus, it both marked a decisive step in the process of political modernization (see *Modernization and Modernity in History*), and provided a case study for future revolutions, as the American struggle for independence was to show (see *American Revolution, The*).

See also: Collective Memory, Anthropology of; Revolutions, History of; Political History (History of Politics)

Bibliography

Hill C 1973 *The World Turned Upside Down. Radical Ideas during the English Revolution*. Temple Smith, London

Hirst D 1999 *England in Conflict, 1603–1660. Kingdom, Community, Commonwealth*. Edward Arnold, London

Hughes A 1991 (ed.) *The Causes of the English Civil War*. Macmillan, Basingstoke and London

Morrill J 1993 *The Nature of the English Revolution*. Longman, London

Pocock J G A (ed.) 1980 *Three British Revolutions: 1641, 1688, 1776*. Princeton University Press, Princeton, NJ

Russell C 1990 *The Causes of the English Civil War*. Clarendon Press, Oxford, UK

Stone L 1986 *The Causes of the English Revolution*, 2nd edn. Routledge, London

Underdown D 1985 *Revel, Riot, and Rebellion. Popular Politics and Culture in England, 1603–1660*. Clarendon Press, Oxford, UK

P. Wende

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Enlightenment

1. Definition and Meaning

Enlightenment is an epochal concept of modernity. It can be compared with others which preceded it such as Renaissance and Baroque, or followed like Romanticism, the latter mostly as its antithesis. The word *Enlightenment* is molded exactly on its German original, *die Aufklärung*, whose first conceptual definition is due to Immanuel Kant; in 1784 he had tried to explain the meaning of the word to the readers of a Berlin journal. The first point to be borne in mind therefore is that the term was condensed with theorical precision in a German area and that *Enlightenment* (second half of the nineteenth century) or *Illuminismo* (beginning of the twentieth century) were created as translations of the German term. But the concept of *Aufklärung* has had several metaphoric equivalents in European culture since the eighteenth century. The best known are linked to the term *light*, which in its turn is connected with *reason*: therefore Enlightenment was conceived from the very beginning as the *Age of Enlightenment*, emancipating reason, the time of man's coming-of-age. This is what Kant meant. Reason is a critical tool allowing man to get rid of prejudice, to change the world, dominating nature through science. Enlightenment is conceived as a culture of conscious change, that is of *reforms*, the secularization of a term which was originally in the singular and expressed a religious concept of modernity.

It does not mechanically coincide with rationalism, but it is one of its essential and critical moments. In fact a rationalistic conception of the world precedes Enlightenment and really belongs to the seventeenth century, as philosophers like Descartes and Spinoza could show. With them reason is a system, whereas with Enlightenment it becomes method. Reason becoming method is an essential point of enlightened epistemology, which not only uses Locke's empiricism, but also Newton's celebrated motto: 'Hypotheses non fingo.'

2. Times, Spaces and Routes

As for all complex epochal categories it is not easy to imprison it in a definite period of time or in space. The fact that the concept originated in a German area does not necessarily mean that the center of Enlightenment is linked to that environment. From a historiographic point of view at least three great European cultures claimed to have given origin and identity to Enlightenment: France, England, and the German world. As far as France was concerned, the capital of Enlightenment (light) was Paris, in the name of the *philosophes* and of the *Encyclopédie*, which synthesized its essential knowledge. London could not only boast Locke's and Newton's contribution, or the precocity of the *freethinkers*, and later the great lesson of Scottish intellectuals, but also its own political system, already based on the separation of powers, and above all on the rapid evolution of public opinion, favored by a free press. The Germans in their turn not only claimed that they had first created the conceptual condensation of the term, but also having given (through Leibniz, Wolff and Kant) Enlightenment a philosophy. Actually these claims implicitly conceal all the ideological processes of the following age, which perceived cultural identities on a national basis, in antithesis with Enlightenment, thus denying a fundamental presupposition, cosmopolitanism.

If one wants to understand the Age of Enlightenment today one has to put aside any simplification identifying center and peripheries, but accept its polycentric nature, which should be considered as a European and worldwide circulation of ideas, to which Kant attributed the choice of universal emancipation. This does not mean that one should ignore the diversities which induced some scholars to imagine the existence of various kinds of Enlightenment. A polycentric interpretation probably helps better than others to save the unity and the diversities of Enlightenment at the same time.

It is never easy to define the period of time enclosing a concept of epoch. If one compares the cultural events of France, England, and Germany with those of other areas from the Mediterranean sea to the north of Europe, one can take the 1680s, indicated by Paul Hazard as the beginning of the 'crisis of European conscience,' as a starting point, provided one does not fix dates or individual events. They are the years of the revocation of the Edict of Nantes, of the *querelle* between Ancients and Moderns, of the appearances in European culture of *Etrangers Symboles* ('foreigners symbols'), like the Chinese Mandarin, archetype of virtuous atheists. In England it is the period of intellectual maturity with Locke and Newton, but also of freethinkers like Anthony Collins and John Toland. Thomasius and Leibniz are the protagonists of the *Frühe Aufklärung*. In Holland refugees like Pierre Bayle and Jean Leclerc elaborate the tools of a great critical culture. Only slightly later the adventure of Giambattista Vico, Pietro Giannone, Alberto Radicati, di Passerano, and Ludovico Muratori has its beginning in Italy. As for the Iberian peninsula, the *décalage* is greater.

The 1730s and 1740s can be considered years of transition, with the important exception of Voltaire's and Montesquieu's France, who inaugurate the Age of Enlightenment. In England the radical culture of freethinking dies out, whereas the Augustan season asserts itself through Samuel Johnson. In Italy Radicati and Giannone succumb to the success of the erudites and of the enlightened catholics (from Scipione Maffei to Muratori). The latter's Spanish equivalent is Father Fejio.

The 1750s and 1760s are the season of Enlightenment at the height of its glory. The circulation of ideas links together all the European centers. Not only Paris, London, Berlin, or the cities of the great cultural mediation like Amsterdam and Geneva, but also Madrid, Lisbon, Milan, Naples, St Petersburgh, and Stockolm become involved in a close dialog. These are the days of the *Encyclopédie*, of the *Contrat Social*, of Cesare Beccaria's *Dei Delitti e Delle Pene*. Learned periodicals become journals of opinion. The first great reforms are introduced in the fields of economy, legislation, politics, and morals. Pombal's Portugal opens a confrontation with the Jesuits in which all the Bourbon states are going to be involved and will have an echo in the great European enlightened culture.

The network of Freemasonry develops the political language of a civil society, deeply entwining itself with all the other tools of European intellectual sociality. Reforming projects are most fruitfully nourished by Utopian ideals. Enlightened absolutism makes its tentative appearance in Austria, Russia, Prussia, Tuscany, Denmark, and Sweden. If sovereigns are drawing near to the philosopher's projects, there is also a fruitful exchange between enlightened culture and government's experience of a new generation of officials.

The 1770s and 1780s, when little by little the generation of the great representatives of Enlightenment (Voltaire, Rousseau, Diderot, D'Alembert, d'Holbach) disappears, tackle an Atlantic event like the American Revolution, which gave Europe first-hand ideological material on questions not yet approached, such as the self-determination of peoples, parliamentary representation, and federative connection within a republican model: liberty, democracy, and civil ethics as a shared and written constitutional project. Europe's boundaries widen with Russia entering the Mediterranean, and the inevitable modifications of the eastern balance. Historians like Edward Gibbon and law reformers like Gactano Filangieri take note of it. The models of reason change. Freemasonry itself, which Lessing thought the ideal tool to diffuse Enlightenment, lives a deep restlessness and the clash between operative rationality and taste for the occult and for mysticism. Raynal, who had enjoyed Diderot's collaboration, accuses the European colonial models, denouncing both their violence against other civilizations as well as slavery and the inclination towards despotism; he goes so far as to assert that the torch of freedom would be carried to Europe by a black Spartacus.

It is not easy to find an *ad quem* term. The beginning of the French Revolution, a starting point of an epochal crisis, is plausible beyond a logic privileging the French course. Some frontier areas are soon involved in the revolutionary turmoil. The process seems to take longer in Germany, England, Spain, Portugal, and Italy, not to mention in Eastern Europe. The French Revolution and war put a stop to the politics of enlightened absolutism. The crisis of enlightened culture and its transformation into something different (from Jacobinism to counter-revolutionary resistance) last through most of the 1790s. As for England—which experiences the first unsettling industrial revolution—it carries out and uncompromising war not only against the French Revolution, but against the Napoleonic Empire as well; and yet this does not modify its constitutional evolution, nor does it alter substantially the role of public opinion. This is true also with regard to the United States, which will later appear to an observer like Alexis de Tocqueville (*De la démocratie en Amérique*, 1835) as the most mature and complete project of the Enlightenment, from which the Europe of *Restauration* would have to learn freedom and democracy anew.

3. *Operative Tools and Fields*

An analysis of the categories of space and time connected with the concept of the Enlightenment implicitly shows its operative tools (reason, experience, criticism, scientific models, reforms, utopian tensions, theories of progress, and of public and private happiness). Each of these terms has its own internal dynamics, but also tends to adapt to highly differentiated realities. Enlightenment today does not appear so much like a clear unitarian philosophy, according to its reconstruction by Ernest Cassirer (1932), as like a process of general emancipation which moves hypothesis, possibilities, and strategies which might appear different and antithetic at first sight. This is clear from a brief analysis of its main fields of influence: religion, ethics, politics, economy, history, social transformation, and cultural change.

As regards religion, in the long term secularization had its roots in the first period of modernity (Humanism, Renaissance, Reformation, libertinism). Criticism of revealed religions is an important element of English freethinking in the early eighteenth century, with repercussions not only in Holland, but also in France and Italy. Materialism, pantheism (the term was coined by Toland in 1704), and the negation of the immortality of the soul involve both the protestant and the catholic world, as the religious adventures, destined to be defeated, of Jean Meslier, Giannone and Radicati di Passerano could show. The progressive reduction of magical, miraculous, and prophetic forecast practices are generalized as the refusal of superstition. Fruits of the application of critical models linked with philosophy and science, they create various intersections between faith and reason, from reasonable Christianity to enlightened catholicism, to natural theology, to deism. Tolerance, which is not only the cohabitation of different religious confessions on the same territory, but also the possibility (enunciated by Pierre Bayle) of individual conscience to err, without any church arrogating to itself the power of

constraint, is a good ground for testing important conquests. In the wake of a great juridical tradition Voltaire will take this battle into the very core of Enlightenment, elaborating for European public opinion the theoretical premises for the choices of Prussian, Austrian, and Tuscan enlightened absolutism.

D'Holbach's *coterie*, which represents the most radical experience of the French Enlightenment from a religious point of view (with two phases: the first an antichristian offensive; the second the elaboration of a social model where ethics and politics were separated from religion) is only one of the components, though the most aggressive and sensational, of the French and European Enlightenment. It is all too easy to underline its elitist nature, limited to a small group speculating on the complicity of European publishers interested in creating a flourishing market of clandestine books. One should not underestimate this aspect, after a scholar like Robert Darnton explained the relation between publishing and sedition, between forbidden books and crisis of political, ethical and religious social structures.

The most lasting result of the Enlightenment in this field was the widespread consciousness not only among laymen but among priests as well, including enlightened catholics, that religious choice mainly belonged to the private sphere. In this sense the most mature enlightened project had accepted and adapted for a larger civil society the rights of conscience proposed by Bayle at the time of the 'crisis of European conscience.'

4. *Emancipation, Reform and Utopia*

In its most essential and creative moments, each innovative choice of the Enlightenment was preceded by an ethical reflection which cannot be mechanically reduced to an individual reckoning of pleasures and sorrows. The utilitarian moral itself had imagined with Bernard de Mandeville that private vice could become public virtue; but in the middle of the eighteenth century it was perceived more as a tension to ensure the maximum happiness to the largest possible number of men. From this point of view the analysis of Montesquieu's works, of the 'economists,' of Pietro Verri, Beccaria, and Gaetano Filangieri are important because of the strong relation they establish with a new ethical project. The Scottish Enlightenment also starts from morals to discover economy and sociology. In fact one of the concepts most charged with future, that of 'civil religion' theorized by Rousseau in his *Contrat social*, was based on a profound thought of natural law (above all through Jean Barbeyrac) which had tried to set the field of man's rights and duties free from the traditional links of Patristics and Scholasticism.

Rousseau and Mably had transformed this inheritance and founded a new system of rights and duties of the citizen. The state was taking upon itself new tasks in the fields of instruction, cultural control and assistance, thus widening its public dimension and secular action; as a consequence religious choice would tend to become a private question, involving the intimate and inviolable sphere. The most significant lesson of the Enlightenment was not so much a lay morality opposed to the religious one, as the possible dialog on the field of citizenship between different kinds of ethic: a difficult project, continuously interrupted, which modernity has to tackle all the time.

Also from the point of view of political models the Enlightenment does not show a unified front, but a series of possible choices, from enlightened absolutism to the return of a republican culture, to the bureaucratic centralization of the 'well-ordered police state,' to constitutionalism, to debates on representation, to projects of direct democracy. Each of these terms in fact can be used to define very different experiences, unless there is the intervention of a contextual historical event. There is a sharp distinction, for example, between the forced modernization of Peter the Great, which is an example of reformism to the point that it attracted Voltaire's attention, and the strategies of 'enlightened' princes like Frederick II, Joseph II, Peter Leopold, and even Catherine the Great.

In the same way, with regard to republican thought, one should not confuse the patrician models of the republics which survived the success of absolute power (Holland, Venice, and Genoa), or the mannered republicanism one learnt by reading the classics, with the new republican tensions which arose as a consequence of the claims for freedom of the Poles, the Corsicans, the Genevans and above all the American colonies. But enlightened absolutism itself attracts or repels enlightened thinkers because of two equal possibilities. The first is the use of centralization in order to carry our reforming politics, where the accent was not so much on strength, however necessary to win resistance, as on the involvement in the project of new fields: law, politics, economy, religion, culture, together with the elimination of particularisms and unjust social relations. This implied the other necessity, that of transforming the model into constitutional forms. The process of emancipation, but also of adaptation to the complexities of the real world one wants to change, Enlightenment moves between the two poles analyzed by Franco Venturi (1971): utopia and reform, closely entwined. These two tensions vivify each other. But the course of the Enlightenment revolves round another two essential terms: freedom and democracy. The first term above all was the main object of Montesquieu's *L'esprit des lois* (1748). It is not possible here to go beyond a schematization recognizing virtue as the essential value of republics, as distinguished from honor, the real ethical and social cement of monarchies. Montesquieu does not choose among the different political forms. He only discards Asian

despotism and is very diffident against absolutism, which he wants to temper not only by separating the three powers, or through the binding strength of the laws, but also through the concrete presence of intermediate bodies, including the nobility.

The other term of reference for the political theory of Enlightenment is Rousseau. He opposes to eighteenth-century cosmopolitism the love for one's own country, which requires the practice of direct democracy. Rousseau's democracy is right for a small republican country. It does not admit representatives. Between majority and minority only the former represents the general will, because sovereignty derived from the social pact cannot be divided. The Genevan's challenge was even more complex: he questioned a developmental model of civilization based on the progress of art and science, but also on the basis of social inequality. Democracy was the courage to modify ownership, justice, and instruction. A large part of later political thought is linked to a comparison between Montesquieu and Rousseau, to establish a connection between freedom and democracy, the aspiration to absorb the Genevan's Utopian drive and transform it into a spirit of reform. An exemplary text is *Dei delitti e delle pene*, which questions the right to punish in an unjust society because it is unequal.

Various other experiences belong to Enlightened politics: cameralism, an administrative model mainly used by Prussian and Austrian officials; the physiocratic theory of legal despotism, in which a monarchic power guarantees certain natural laws, among which is ownership. From the above considerations it appears that the question of enlightened politics is narrowly linked with the question of law, which becomes 'science of legislation.' The weaker field was feminine political and cultural identity. For this question one would have to wait for Mary Wollstonecraft's adventure and her work on women's rights.

Economy becomes a science during the eighteenth century, evolving from late-mercantilist models. The physiocratic project established the essential points of the productive cycle based on agriculture and on free corn trade. If agriculture was the only wealth-producing sector, while transformation and circulation were the task of industry and commerce, taxing had to become progressive with respect to land, whose ownership was one of the fundamental natural rights. Adam Smith extended liberalism to sectors such as industry and commerce, which were not considered subsidiary, but highly productive. In this way the first strong perception of an industrial revolution under way in England became a theory. During the last decades of the century there was a return to protectionism.

What happens with economy could well be verified with other sciences, not only social and natural, but human as well. The Scottish school creates the premise for sociology, by studying social stratification or the concept of civil society. Economy itself, by defining a stages theory based on production forms, laid down the basis of anthropology. Consolidated disciplines like physics were renovated, as well as chemistry, geology, agronomy, and geography itself. Linneus and Buffon had reclassified nature, thus rendering it an open book. Natural history had given a common language to natural and biological sciences.

As for history, one can only register a change in some fundamental paradigms: the loss of the superiority of sacred history, the transformation of ecclesiastical history into history of religions, the emergence of the history of civilization within new models of universal history, the flourishing of civil history as institutional history, the first appearance of the history of science.

5. *Circulation and Practices: the Growth in Literacy*

If one examines this process from the point of view of cultural history and of general transformation of intellectual models, European and now world-wide circulation of ideas can be measured through parameters which are apparently external to intellectual history, but not unimportant. I am referring first of all to the international book market, but also to the creation of an incredibly large number of tools, thanks to which the consultation of books made possible, from public and private libraries to bibliographies, catalogs, abstracts, reviews. Book and newspapers have an almost universal circulation. In Europe and in America there has been an increase in literacy. The world of potential readers has expanded not only thanks to the advent of the middle classes but because the distance between male and female instruction tends to diminish in urban spaces. Somebody called this the revolution of middle class readers. Even reading habits change; there is a greater tendency to undirected reading, not linked to learned professions but to the consumption of leisure time. The demands of the new readers also bring about a change in genres. The novel has an incredible development, offering as it does to all classes means for education, identification and escapism. Among the forbidden books of the second half of the eighteenth century, the *philosophique* and *pornographique* genres often follow the same semiclandestine circuits, thus contributing to the modification of the cultural and ethical general models of society. Music and theatre as well are often fields for ideological formation and dispute. Journalism follows in the tracks of this expansion of the book market, and from the objectivity of abstracts, which are made for a public of learned people, moves to the subjectivity of journals, the concision of reports. Journals of opinion are essential for the formation of this new

international public sphere which discovers languages, interlocuters, and spaces. To this sphere belong the salons, public, and private spaces at the same time. Intellectual sociality takes a number of forms, from the old and new academies, to clubs, to cafés, down to the taverns where conversation and gossip are a way for dilating news. The Enlightenment–Academies relation is supported by a rich bibliography, starting with Reinhard Koselleck's (1959) work which laid down the basis for a more through analysis of the links between Freemasonry and Enlightenment. The former is now considered the privileged center for the formation of the political language on which the assemblary model of the French Revolution will be created.

Many centers of intellectual communication reverse the prevailing male image one received when analyzing the concept from the point of view of intellectual history. Women are present in the famous French salons, where they have a personal control over an ephemeral though socially important genre like conversation. They start being accepted in some lodges and academies, their role thus changing from that of consumers of books and newspapers (also of a literature for women, produced by men) to that of writers. Firstly, they occupy the space of women's literature, then they take and active part in the journalistic debate. The inclusion of women in the public sphere does not alter the fact that their role is still subordinate. Women's writing often tends to self-identification, but it remains unfinished, secret. And this corresponds to a social model which is prevalent in male Enlightenment. Rousseau, who had put the question of public education of the citizens, had kept women separate; Filangieri himself had followed suit. This explains the novelty of Wollstonecraft's denunciation. In fact the public sphere created by High Enlightenment seems to be threatened by new classes joining the world of writing and communication, forming an international Grub street and living on collaborations and translations. The creation of a more aggressive political language, and of the use of reason both as prophecy and announcement of the crisis, breaks the balance of the previous generations which had reached the highest levels of intellectual communication in the academies, in the teaching and in the publishing world. In this phase women seem to live less apart. The fact remains that on this question Enlightenment as emancipation was a partial project delegated to the future, because it had not fully accepted the other half of its universe.

6. *Building an Epocal Concept*

Finally it should be said that the French Revolution, whilst celebrating the Enlightenment as its parent, strongly contributed to its interruption first of all because what had survived as *idéologie* had to adapt to the heavy Napoleonic tutelage and stabilization, and secondly because the behavior and the propaganda of Napoleonic France were experienced by the European cultures as a violence which forced them to find anew their national, religious, and identifying roots. This explains why part of Romanticism set itself up as an antithetic culture to the Enlightenment which was considered responsible for the political conditions in Europe. Friedrich Meinecke had spotted the opposition between cosmopolitism and the national state (1908). While new problems required new kinds of culture, certain fundamental values, after being driven underground during the *Restauration*, after the Napoleonic age, sprang up again in all the moments of opening to freedom and democracy which are scattered in the history of Europe. In French history the ideological elaboration of the Third Republic renewed the strong link between Revolution and Enlightenment again, the latter being dragged by the former.

The twentieth century, however, really needed to reflect on Enlightenment again, having a critical period troubled by nationalistic, fascist, and totalitarian cultures which threatened liberty, democracy, and equality. These threats were answered by Cassirer's (1932), Becker's (1932), Hazard's (1935, 1946) and many other's meek projects. From these roots there originated an important historiography (Venturi (1971), Gay (1967), Koselleck (1959), Pocock, Darnton, Jacob (1946, 1980), Roche and many others) which to this day has an echo in the International Society for Eighteenth Century Studies and in the dense network of National Societies. Margaret Candee Jacob has written about a 'Radical Enlightment,' which was masonic, pantheist, and republican. John Pocock (1999), more recently, has located in the English world a 'Conservative Enlightenment,' while Jonathan Israel (2001) widens the chronological terms of a European 'Radical Enlightenment' and 'its making of modernity,' starting from 1650.

There are also many publications on the limitations, deficiencies and responsibilities of Enlightenment in the evils of modernity, to begin with T. W. Adorno and M. Horkheimer (1947). But in the light of the dramatic problems of our times many evoke the necessity of a further Enlightenment, of a future of Enlightenment which will resume the interrupted project. One can well imagine that it will inevitably be a new one, because the past can give inspiration, but not necessarily return. It is true, however, that we must find our own answers to emancipation, freedom and Utopia for a world otherwise risks being an icy globality without universalism.

See also: American Revolution, The; Democracy, History of; Democracy: Normative Theory; French Revolution, The; Locke, John (1632–1704); Montesquieu, Charles, the Second Baron of (1689–1755); Neyman, Jerzy (1894–1981); Political Economy,

History of; Rationalism; Rationality in Society; Rousseau, Jean-Jacques (1712–78)

Bibliography

Adorno T W, Hoorkheimer M 1947 *Die Dialektik der Aufklärung*. Fischer, Frankfurt am Main, Germany [Engl. transl. 1972 *Dialectic of Enlightenment*. Allen, New York]

Becker C L 1932 *The Heavenly City of Eighteenth Century Philosophers*. Yale, New Haven, CT, USA

Benitez M 1996 *La face cachée des Lumières. Recherches sur les manuscrits philosophiques clandestins de l'âge classique*. Universitas, Paris; Voltaire Foundation, Oxford

Cassirer E 1932 *Die Philosophie der Aufklärüng*. Mohr, Tübingen, Germany

Crocker L G 1959 *An Age of Crisis. Man and World in Eighteenth Century French Thought*. Johns Hopkins University Press, Baltimore, MA, USA

Delon M (ed.) 1997 *Dictionnaire européen des Lumières*. PUF, Paris

Diaz F 1986 *Dal Movimento dei Lumi al Movimento dei Popoli*. Il Mulino, Bologna

Ferrone V, Roche D (eds.) 1997 *L'Illuminismo, Dizionario Storico*. Laterza, Roma

Gay P 1967–1969 *The Enlightenment. An Interpretation*. Knopf, New York

Giarrizzo G 1995 *Massoneria e Illuminismo nell'Europa del Settecento*. Marsilio, Venezia

Guerci L 1986 *L'Europa del Settecento. Permanenze Mutamenti*. Utet, Turin

Hazard P 1935 *La crise de la conscience européenne*. Boivin, Paris

Hazard P 1946 *La pensée européenne de Montesquieu à Lessing*. Boivin, Paris

Israel J I 2001 *Radical Enlightenment: Philosophy and the Making of Modernity 1650–1750*. Oxford University Press, Oxford, UK

Jacob M C 1980 *The Radical Enlightenment*. Unwin, London

Jacob M C 1994 *Living Enlightenment: Freemasonry and Politics in Eighteenth Century Europe*. Oxford University Press, New York

Koselleck R 1959 *Kritik und Krise Zur Pathogenese der bürgerlische Welt*. Alber, Freiburg-München

Mortier R 1969 *Clartés et ombres au siècle des Lumières*. Droz, Geneva

Outram D 1995 *The Enlightenment*. Cambridge University Press, Cambridge, UK

Pocock J G G A 1999 *The Enlightenments of Edward Gibbon*. Cambridge University Press, Cambridge, 2 vols.

Porter R, Teich M (eds.) 1979 *The Enlightenment in National Context*. Cambridge University Press, Cambridge, UK

Ricuperati G 1994 Le categorie di periodizzazione e il Setteceneto. Per un introduzione storigrapca. *Studi Settecenteschi* **14**: 9–106

Ricuperati G. (ed.) 2000 *La reinvenzione dei Lumi. Percorsi storiografici del Novecento*. Firenze, Olschki

Scheneiders W (ed.) 1965 *Lexikon der Aufklärung. Deutschland und Europa*. Beck, München

Schmidt J (ed.) 1996 *What is Enlightenment? Eighteenth Century Answers and Twentieth Century Questions*. University of California Press, Berkeley, CA

Tortarolo E 1999 *L'illuminismo, Ragioni e Dubbi della Modernitò*. Carocci, Rome

Venturi F 1971 *Utopia and Reform in the Enlightenment*. Cambridge University Press, Cambridge

Venturi F 1969–1990 *Settecento riformatore*, 5 vol (in 7 tomes). [1989 English transl. of the III, *The End of the Old Regime in Europe, 1768–1776: The First Crisis*. Princeton University Press, Princeton; 1991 English transl. of vol IV (in two tomes) *The End of the Old Regime in Europe, 1776–1789*. Princeton University Press, Princeton, 2 vols.]

G. Ricuperati

Entertainment

Entertainment is play, but play put on display. In the modern, Western or Westernized world plays and games have become commodities for sale in that special branch of economic life called show business. There is no limit to the media, the art forms, or the genres that might be marketed by the agents of show business, but within television and music the label 'light entertainment' marks the quintessence of entertainment: ease, lightness, cultural expressions with no other aim than pleasing as long as the program or the piece of music lasts. Entertainment, in the modern sense of the word, may thus be defined as the commercialization of the pursuit of happiness.

Late in the reign of Queen Elizabeth I, Sir John Harington wrote a *Treatise On Playe* (Harington 1968), where he defines play as 'a spending of the tyme eyther in speeche or action, whose only end ys a delyght of the mynd or the speryt.' In Elizabethan culture play developed into commercial entertainments, and Harington's definition corresponds with most modern definitions of play and entertainment. His distinction between entertaining words and entertaining actions also seems to correspond with the ancient Roman distinction between *jocus* (cf. joke) and *ludus*, but anachronism and ethnocentrism may lead to a distorted understanding if the modern concept of entertainment is projected onto cultures distant in time or space. Thus, during the Empire the Roman plays assumed explicit political functions. An operative definition of entertainment must therefore include a discussion of the interaction between history, structure, and functions in the emergence and the enjoyment of entertainment.

1. Play and Entertainment

'Only when playing is man fully man,' Friedrich Schiller declared in his manifesto *Über Ästhetische Erziehung*, and his words are echoed in Johan Huizinga's (1937) influential book *Homo Ludens*, where it is claimed that all cultural institutions are derived from play. Huizinga's sketch of the history of the concept of play, however, shows that the meaning varies from culture to culture. Thus, in ancient Greek

the verb indicating 'play' literally meant 'doing what children do' (which raises a parallel problem of historical and cultural differences in the definition of childhood). In spite of the risk of anachronism and ethnocentrism, Huizinga presents the following general definition of play: it is an activity that is freely chosen, taking place in a space and time separated from the ordinary world, operating by fixed rules or by a general 'rule' of fictiveness, and leading to results, which are uncertain and unproductive. If freedom is fundamental to play, this would exclude the medieval carnivals since they were linked to a social and religious calendar, and it is indeed very likely that Huizinga's definition of play itself is tied to modern, liberal society. If this is so, the process of secularization is what makes possible the transfer of forms of religious ritual to forms of play and entertainment, and this again implies that what appears to be the same activity actually is perceived very differently according to cultural norms.

One of Huizinga's critics, Jean Henriot (1989), has suggested that a theory of play should consider the development of the idea of play rather than the practices of play. At the individual level the child learns to consider certain activities or certain attitudes towards these activities as play, and at the social level people learn to distinguish between for instance education, work, politics, war, and play.

Ideas such as play, work, etc. govern our perception of society, and play may even be interpreted as one of those ideas in which society reflects on itself. Henriot advocates a historical definition of play, but he also notes that it is always characterized by detachment and mental lightness, and here he comes close to Georg Simmel's sociological perspective. Simmel suggests that all social processes may be considered as content and as form, and in play the forms are articulated for their own sake, while the content is suspended: in an entertaining conversation the aim is not to convince the other, the pleasure of verbal exchange is the only aim; and in a quiz-show the aim is not to get an answer to a problem, the playful reproduction of social forms known from the church (catechization) or the school (examination) is the only aim. The word 'entertainment' actually refers to play as a source of 'sociability' (Simmel). It is derived from Latin: *inter* means 'among' or 'in between' and *tenere* means 'to hold,' and therefore the meaning of entertainment is 'to hold something mutually' or 'to keep something up.' Entertainment is a social relationship and is concerned with social relations. Play has its origins in the fact that people cannot be alone, as La Bruyère stated in his *Caractères* [1688], and one of the pioneers of modern entertainment, the composer Jacques Offenbach emphasized the importance of mutual support by declaring that his dream was 'to found a share holding company of mutual insurance for the prevention of boredom' (Kracauer 1976, p. 101).

Undoubtedly there are great differences between cultures as to the degree to which a culture can or may distinguish between the contents and the forms of its everyday practices, but Simmel's perspective offers the possibility of a mediation between a historical and a formal understanding of play and entertainment; it indicates in what sense Huizinga was right when he saw a connection between play and the other cultural institutions; and it explains why play or play-like activities have been so influential in the description and interpretation of cultures distant in time as well as space from Mikhail Bakhtin and Emmanuell Le Roy Ladurie on the Renaissance carnival to Clifford Geertz on Balinese cock-fighting.

2. *Bread and Circuses*

In Western culture spectacles, circuses, mimes, etc. go back to antiquity. The idea of the ludic, i.e., the playful, is derived from the Roman concept of *ludus* (chariot races); and the idea of the aleatory, i.e., the hazardous, is derived from the Roman word for dice. In the course of Roman history it is possible to follow a restructuring of religious rituals—such as gladiatorial fights in connection with funeral rites—into cultural forms much closer to the modern idea of entertainment. Even though the giant spectacles of the Roman Empire functioned as part of a system that should solidify and glorify the status of those in power, they have become immensely powerful models for modern entertainment and modern critique of entertainment. Juvenal's verdict on a social policy that replaced the active political commitment of its citizens by the passive reception of 'bread and circuses' has inspired a 'mythology of negative classicism', i.e., comparisons of the development of modern society with the social decay of classical Rome (see Brantlinger 1983, p. 18). Just as the Romans staged enormous sea battles, mythological or historical, so English imperialism was supported by show business in gigantic reproductions of the imperial wars in the colonies, for instance the wars in South Africa staged as Savage South Africa. Colosseum has become a popular name for modern establishments within show business, and Las Vegas of course boasts a Caesar's Palace.

3. *Carnival*

In the Middle Ages, Roman festivals such as the Saturnalia and Lupercalia metamorphosed into the still-existing tradition of carnival. Entertainment is often characterized as recreation, and re-creation is what characterizes the carnival. The rural and pagan carnival would recreate the natural world, the Christian carnival would recreate the religious community, and the urban carnival would recreate the power and splendour of the city—or different groups and corporations within the city. During the carnival the

world is turned upside down, and in this period of creative chaos the sense of community is reanimated.

In contrast to commercial entertainment however, the participants in the carnival were not free to choose their leisure, and therefore their performances cannot be called pure entertainment. From the perspective of the church or the city council who presided over the event, the aim of the carnival was to make the different groups of society accept hierarchy. The function of the blasphemous words and gestures was to frighten people away from changing society or at least to restrict their demands for social and spiritual change to the carnival season.

That the authorities could use the carnival is the only plausible explanation of their pivotal role in staging the processions and plays of the carnival. But as long as the carnival lasted the situation was potentially open, the functions and the future of society were put at stake, and the safety valve might turn into an explosive political risk. Mikhail Bakhtin (1968) emphasizes the emancipatory aspects of the carnival in his influential book *Rabelais and his World*, while Emmanuelle Le Roy Ladurie in his *Le Carnaval De Romans* shows the confrontation between disciplining forces and rebellious forces within the carnival.

4. *Courtly Diversions*

In Europe in the sixteenth, seventeenth, and eighteenth centuries the ways of popular culture and elite culture separated. Politically, as power became centralized the powerful made a spectacle of themselves through fireworks, pageants, and allegorical plays, while reducing their subjects to spectators. Economically, the Reformation and the rise of the middle classes, who knew that time was money, was accompanied by a war against idle leisure.

To the aristocracy as well as the middle classes, the experience of time, the abstract and empty time of the clock, became a new social experience, but while the aristocracy actively cultivated a culture of pastimes and diversions, the puritan middle classes rejected all pastimes that would divert their attention from work. While the aristocracy in an oblique way became the pioneers of modern entertainment, the puritans just as obliquely shaped the pattern of modern cultural critique.

The culture of *divertissements* at the French court developed as the process of centralization and the politics of absolutism changed the 'nobility of the sword' into courtiers. Leisure time at court became a psychological as well as a political problem. To aristocrats used to fighting, hunting, etc., life at court was inactive and boring, and they had to do something to pass away their time. Diversions would compensate for the lack of activity and power and at the same time the entertaining discussions, dances, etc. would create a social bond between them. The intricate equestrian displays known as *carrousels* epitomized this domestication of a warrior class through entertainment. A trace of this historical development may still be seen in the predilection for a lavish, baroque style in merry-go-rounds and other fairground architecture.

5. *Commerce and Buying Fun*

In the eighteenth century the word entertainment (or in Germany: *Unterhaltung*) had a breakthrough, but this was in reference to private or domestic entertainment. The expansion of the book market contributed to changing the private home into a sphere of entertainment, and in so far as stories of love or adventure were written for the market, the principle of publicity entered the private sphere. The court was replaced by the market as the provider of entertainment, and these entertainments were open to everyone with sufficient money. In 1725 Daniel Defoe noted that writing had become 'a very considerable Branch of the English Commerce' (Watt 1957, p. 55), and an increasingly popular genre such as 'the penny dreadful' indicates the importance of the market. In the Scandinavian languages the word commerce became synonymous with having fun.

However, at the beginning of the nineteenth century Joseph Strutt wrote *The Sports And Pastimes of the People of England* (1801) in which it is obvious that the idea of public entertainment for a general public was only beginning to get a foothold. In his table of contents the pastimes are divided according to time, place, and social group: 'Rural exercises practised by persons of rank,' 'Rural exercises generally practised,' 'Pastimes usually exercised in towns and cities, or places adjoining to them,' 'Domestic amusements of various kinds; and pastimes appropriated to particular seasons.'

In contrast to premodern entertainments the break with hierarchy was a general feature of the modern culture of entertainment. When Philip Astley (Speaight 1980) founded the first circus at Westminster Bridge in 1770—gathering itinerant showmen, jugglers, clowns, and acrobats under the same roof, with riding masters as a tribute to aristocratic culture—he addressed his public as 'the Nobility, Gentry and others.' But he, as well as other entrepreneurs within show business, soon learned that it would pay off to put the 'others' in the foreground. If the new entertainment industry used amusements of the old aristocratic culture, they would inevitably democratize the diversions of the past. Søren Kierkegaard made 'leveling' a key concept in his analysis of modern, urban culture; Georg Simmel wrote of 'leveling' in his analysis of 'The Metropolis and Mental Life' from 1903; and in 1913 Georges d'Avenel published a book

on the leveling of amusements in modern times: *Le nivellement des jouissances.*

6. *The Mass Entertainment Industry of the Nineteenth Century*

According to Asa Briggs (1960) five conditions stand out for the development of a mass entertainment industry in the nineteenth century. First, a large and concentrated urban population. Second, sufficient money to buy regular, cheap entertainment. Third, sufficient leisure time to make a commercial exploitation possible. Fourth, urban transportation systems to take the public to the places and palaces of entertainment. Fifth, the application of technology to entertainment. The money, the leisure time, the transportation, and the technology making mass reproduction possible are self-evident preconditions for the rise of a mass entertainment industry. The process of urbanization, however, needs some further comment.

In the nineteenth century the concentration of the urban population was preceded by an uprooting of people—from the countryside, from tradition, etc.—and the isolation of the individual in the middle of the urban crowd called forth a need for the mental support of entertainment. Just as the popular culture of fairy-tales, carnival, etc. had given form and meaning to the culture of the preindustrial society, the mass entertainment industry shaped a sense of community between urban dwellers, who could no longer rely on time-honored traditions or pre-established social relations. Boredom may be termed the cultural disease of the nineteenth century. Similar to the boredom experienced at the courts of the seventeenth and eighteenth centuries, this was the psychological reflection of a loss of cultural meaning, a loss of a common cause. A recurrent theme in nineteenth century literature was the characterization of boredom (*ennui*, *Langeweile*) as a modern invention, a modern monster, the constitutional evil of the century (see Sagnes 1969). In the same year that the cosmopolitan Georg Carstensen opened the Tivoli Gardens in Copenhagen, Søren Kierkegaard published his *Either-Or* with an ironical chapter on boredom as the source of all creation including the Creation, and as a remedy to this boredom he suggested public entertainments on a grand scale.

The grandest scale was represented by the city itself and the crowd, and urban life became a leading motive in all kinds of modern entertainment: panoramas such as the London Colosseum showing a panorama of London as seen from the cupola of Saint Paul's Cathedral, crime and detective stories, melodramas, music (from Offenbach's 'La vie parisienne' to the role songs of the music halls). Many of the new establishments of entertainment were even constructed as small cities, for instance panoramas with their spectacular 360 degree representations of cityscapes; world exhibitions; amusement parks such as the Tivoli Gardens; and zoological gardens with gates, walls, streets, squares, and animals living in cages shaped like buildings from their native countries.

City life was entertaining in itself, as witnessed by cafés, promenades, show windows, etc., and the appeal to visual pleasures was noticed by most observers. Charles Lamb wrote of the spectacle as the essence of the city; Théophile Gautier declared that the era of purely visual spectacles had arrived; and in the 1880s a new and abundant use of the terms 'show' and 'show business' (see Nasaw 1993) marked that visual turn of entertainment which has continued with film, television, and even music videos in the twentieth century.

7. *A Map of Modern Entertainment*

A map of modern entertainment may take as its starting point the minimizing expression: 'only entertainment.' This expression indicates two things: (a) that entertainment has become a specific sphere in society, and (b) that entertainment has no importance. As a specific sphere, as 'show business,' entertainment differs from play although it is based on forms of play. And as a sphere of activities allegedly of no importance, entertainment differs from art although the absence of intentions may be seen as an extreme version of the esthetic attitude defined by Kant as 'disinterested pleasure.' However, the reduction of entertainment to 'only entertainment' may be interpreted as an effort to deny or exorcise the real importance of entertainment in modern society: as a substitute for religion, as a purveyor of a new mythology, as a part of politics, and as a key sector in the economy.

As mentioned above, boredom seems to have been the mental reflection of the processes of modernization, which changed the status of religion, politics, economy, and private life. Uprooted from traditional social ties, modern man was thrown back on himself. Entertainment, however would offer a reconstruction of social ties based on the elementary experiences of the body and the gaze. Circus, melodrama, waxworks, the whole showbusiness, literally show that the body and the gaze are the stuff entertainment is made of. On one hand the concept of diversions indicates that entertainment literally should divert attention from this boredom. On the other hand entertainment became an institution or a cultural practice which should shape the relationship between the individual and modern life, the mass and modernity.

Play is the basis of entertainment and similarly to play, the forms of entertainment may be divided into four categories, each articulating a specific relation between the individual and society—or the mass and power. Following Roger Caillois (1967) the main forms of entertainment may be interpreted as four different ways of handling the oppositions between

identification and alienation and activity and passivity (or action and passion).

(a) Activity and identification are leading principles in chariot races, chivalrous tournaments, and sports as well as in the shooting parlors of the amusement parks, and in the narratives of action and adventure whether told in novels, films, or TV shows. Just as the hero does, the audience identifying with the hero seeks the recognition of their identity in the hero's victory. The characteristic emotional experience of action, adventure, and competition is suspense, which gives the reader or spectator a vicarious feeling of taking part in the action.

(b) Passivity and identification are leading principles in all games of chance, in horror stories, etc. The heroes are more often than not victims, and if they end up as winners it is not by their own merit. Identification therefore does not consist in recognition of their name but in identification of the number which makes them winners. The characteristic emotional experience of this type of entertainment is the thrill, which gives the reader or spectator a vicarious feeling of being victimized or saved by alien forces.

(c) Activity and alienation are leading principles in all kinds of theater, masquerades, variety shows, literary parodies, etc. The heroes here take on the heroic or villainous or just different persona of another person, and if they gain a name, it is through their talent for making themselves another person. The whole gamut of feelings is of course invited by this type of entertainment.

(d) Passivity and alienation are leading principles in slapstick comedy, in many kinds of music and dance, in swings, merry-go-rounds, roller-coasters, etc. The heroes here hand over themselves and their destiny to impersonal forces, and in contrast to the gambler do not look for any profit but the pleasure of vertigo. Even though emotional distance from the capers and clowns of all sorts is one of the enjoyments of comedy, the readers or spectators who give themselves over to laughter are also looking for the pleasure of losing control.

Even though play often involves a gathering of onlookers, it is not, as in public entertainment, based on a division between actors and spectators. This division of course complicates any mapping of emotional investments and gains. In action and adventure the investments of reader or spectator parallel the hero, whereas in comedy the investments of the reader or the spectator are characterized by their contrast to the troubles of the 'hero' (i.e., the fool or the clown). And even though a horse race confronts parties in a competition where they can only rely on their own skill in their endeavor to be named winners, the spectators may very well make it an occasion for gambling. The map thus needs further elaboration, but already this provisional map shows the possibility of describing and comparing forms of entertainment across the different media. What they have in common are typical, and sometimes stereotypical, ways of staging self-experience and social bonds.

8. *Entertainment Now*

After 1900 show business has developed exponentially with the new media, but now the system of genres and types seems well established. Sampling, parody, and similar kinds of recycling indicate a growing familiarity with the world of entertainment. Also a new entertainment form such as computer games is characterized by using the formulaic worlds of adventure, crime, gothic tales etc. In spite of computer cafés, this medium also confirms the trend toward privatization and even individualization of entertainment: the radio made it possible to enjoy a symphony on the sofa, the television set made the sitting room into a cinema, the transistor radio and the home video made it possible for individuals to chose when they would enjoy a specific program, the Walkman made the individual into a closed electronic circuit. In contrast to this process of extreme individualization a number of new entertainment forms offer the pleasure of being a member of a crowd: rock festivals, sport events, cult movies. Entertainment no longer forms a distinctive sphere, or it is at least often discussed as an aspect of education, politics, work, or even war (wargames). If entertainment is a dramatization and an interpretation of modern life, this invasion of the other spheres of social life may reveal a growing need for understanding how society—in Simmel's words—is possible.

See also: Adolescents: Leisure-time Activities; Globalization and World Culture; Leisure and Cultural Consumption; Leisure, Psychology of; Leisure, Sociology of; Media, Uses of; Play, Anthropology of; Popular Culture; Sport and Exercise Psychology: Overview; Sport, Anthropology of; Sport, Sociology of; Television: Genres; Television: History; Television: Industry

Bibliography

Bakhtin M 1968 *Rabelais and His World*. MIT Press, Cambridge, MA

Brantlinger P 1983 *Bread and Circuses. Theories of Mass Culture as Social Decay*. Cornell University Press, Ithaca, NY

Briggs A 1960 *Mass Entertainment: The Origins of a Modern Industry*. Griffin Press, Adelaide, Australia

Burke P 1978 *Popular Culture in Early Modern Europe*. Temple Smith, London

Caillois R 1967 *Les jeux et les hommes*. Gallimard, Paris

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Harington J 1968 A treatise on playe. In: Harington J (ed.) *Nugae Antiquae*. Georg Olms, Hildesheim, Germany, Vol. 2, pp. 144–88

Harris N 1973 *Humbug. The Art of P. T. Barnum*. Little, Brown, Boston
Heers J 1971 *Fêtes, jeux et joutes dans les sociétés d'Occident à la fin du Moyen Age*. Librairie J. Vrin, Paris
Henriot J 1989 *Sous couleur de jouer. La métaphore ludique*. José Corti, Paris
Huizinga J 1970 *Homo Ludens: A Study of the Play Element in Culture*. Temple Smith, London
Kracauer J 1976 *Jacques Offenbach und das Paris seiner Zeit*. Suhrkamp, Frankfurt am Main, Germany
Le Roy Ladurie E 1979 *Le Carnaval de Romans. De la Chandeleur au Mercredi des Cendres 1579–1580*. Gallimard, Paris
McLean A F Jr 1965 *American Vaudeville as Ritual*. University of Kentucky Press, Lexington, KY
Mehl J 000 *Les jeux au royaume de France du XIIIe au début du XVIe siècle*. Fayard, Paris
Nasaw D 1993 *Going Out. The Rise and Fall of Public Amusements*. Basic Books, New York
Rearick C 1985 *Pleasures of the Belle Epoque. Entertainment and Festivity in Turn-of-the-century France*. Yale University Press, New Haven, CT
Sagnes G 1969 *L'ennui dans la littérature française de Flaubert à Laforgue (1848–1884)*. A. Colin, Paris
Schiller F 1967 *On the Aesthetic Education of Men* [trans. Wilkinson E M, Willoughby L A] Clarendon Press, Oxford
Speaight G 1980 *A History of the Circus*. Tantivy, London, UK
Veyne P 1976 *Le pain et le cirque. Sociologie historique d'un pluralisme politique*. Seuil, Paris
Watt I 1957 *The Rise of the Novel. Studies in Defoe, Richardson and Fielding*. University of California Press, Berkeley, CA

M. Zerlang

Entrepreneurship

The entrepreneur is the head of the firm and coordinates the factors of production; introduces new methods, processes, and products and creates opportunities for economic growth; bears the risks connected with his or her activities; and, enjoys power and high status in capitalist market societies. Entrepreneurs predate capitalism and do exist in socialist and other alternative economic systems, but are especially associated with that economic system more than any other.

The term entrepreneur is only one of the concepts used to designate leadership in business and overlaps with others, such as capitalist, employer, investor, owner, manager, producer, chief executive, bourgeois, investor, work-giver, profit-maker, boss, businessman, and businesswoman. Each of these labels stresses one aspect of leadership in business institutions and of the relations with a variety of stakeholders, but the concept 'entrepreneur' seems the most encompassing and rigorously defined.

The very origin of the word 'entrepreneur' must be traced to the realm of war rather than of economic competition, although in a war for gain rather than for power or honor. The first references to the word appeared in sixteenth-century France, where it defined the fortune captain who hired mercenary soldiers to serve princes, leagues, or towns for pay. It was only later, in the eighteenth century that the concept applied to 'peaceful' economic actors who either undertook contracts for public works, or introduced new agricultural techniques in their land, or risked their own capital in industry.

The first theoretical accounts of the entrepreneurial function appeared in the second half of the eighteenth century with Cantillon and Turgot, who stressed the willingness to accept the risk and uncertainty inherent in economic activity as the key distinctive feature, a concept which was later developed by Knight in the twentieth century. Adam Smith and mainstream economics did not emphasize the specific role of the entrepreneur—with the notable exceptions of Jean Baptiste Say—who defined the entrepreneur as the person who combines the factors of production for profit, putting himself at the center of several relations among capitalists, workers, technicians, merchants, and consumers—and John Stuart Mill—who separated the entrepreneurial function (the payment for which is profit) from that of providing capital (the payment for which is interest). It was, however, Joseph Schumpeter, who above all, conceptualized entrepreneurship as innovativeness, thus giving it a dynamic quality lacking in earlier formulations.

Contemporary economics assumes that entrepreneurial services are highly elastic and that failures in entrepreneurship are due to maladjustments to external market conditions and to lack of economic incentives. Only a few economists analyze more deeply the concept of the entrepreneur, conceiving him or her as a middleperson between markets, or a gap-filler alert to new opportunities. The study of entrepreneurship is thus more developed within other disciplines: sociology, business history, psychology, and anthropology.

1. *The Classical Theories*

The classical interpretations of entrepreneurship by Marx, Weber, and Schumpeter are framed into broad, ambitious studies of economic dynamics and cultural change which, although sometimes not convincing and in need of empirical verification and substantial revision, generally avoid the pitfalls of comparative static models and often ask the right kind of questions.

Marx's contribution to the study of entrepreneurship does not lie so much in *Das Kapital*'s (1867) abstract model of the capitalist mode of production, as in the analysis of the capitalist labor market and factory system, and in the historical sketch on 'primitive accumulation,' where he traces the processes of disenfranchising of labor from feudal and corporate ties and the transformation of land, merchant, and money capital into industrial capital; as well as in the

accounts of the events of 1848 and 1870 in France, where the different strategies of the various components of the bourgeoisie, such as industrial and financial capital, are analyzed. On the whole, in spite of its methodological errors—first of all the reifying of its economic model into more or less universal laws—Marx's analysis of the rise and the internal dynamics of capitalism played an important role in future works by Weber, Sombart, Schumpeter, and many others (both Marxists like Dobb, Sweezy, the 'dependencia' theorists, and non-Marxists like Polanyi and Barrington Moore), by identifying a set of basic structural conditions of a capitalist economy, where entrepreneurship can develop.

Weber's study of entrepreneurship is couched in his analysis of the interplay between religious ethos and modern rationalization in the rise of capitalism. The entrepreneur is clearly distinguished from his historical predecessors in traditional economies by virtue of his rational and systematic pursuit of economic gain, reliance on calculation measured in relation to this economic criterion, the extension of trust through credit, and the subordination of consumption in the interest of accumulation. These are the elements of the rational economic actor's 'instrumental rationality' (*Zweckrationalitat*), by which he or she establishes a systematic relationship between preferred goals and the most suitable means to achieve them. In his ponderous essays on the economic ethics of great world religions published in the *Archiv für Sozialwissenschaft und Sozialpolitik* and in his major work *Wirtschaft und Gesellschaft* on the relations between economy and society, Weber argued that the advent of ascetic Protestantism provided an especially fruitful breeding ground for the mentality of economic rationality, since through a complex process material success is a sign of ascetic realization.

Weber's analysis has been variously and sometimes convincingly criticized; but his most significant and lasting contribution lies, as Brigitte Berger (1991) puts it, in 'his ability to show how the expansion of "instrumental rationality" characteristic of the modern entrepreneurial phenomenon also impelled—slowly and incrementally, through the efforts of individuals and groups in their everyday activities and practices—the formation of distinctly modern institutions in all spheres of life, the public as well as the private.' The degree to which the forces of rationalization responsible for dislodging individuals from their embeddedness in nature, religion, and tradition continue to shape contemporary and future developments is the study of some current researchers, particularly those influenced by Kellner.

The classical interpretations of Marx and Weber are relevant, but Schumpeter is the theorist of entrepreneurship 'par excellence,' who provides the most thorough analysis of the entrepreneurial function. The analysis of entrepreneurship is at the center of Schumpeter's theoretical system. The entrepreneurial function is the key variable in his theory of development. He defined it as innovation, the introduction of a new combination of the factors of production (land and labor) that, when combined with credit, breaks into the static equilibrium of the circular flow of economic life and raises it to a new level. The entrepreneur changes the conditions of supply, combines existing resources in new ways, and thereby sets up a new production function. Entrepreneurship does not imply the requisite of property, is not based on the assumption of risk, and does not require belonging to a business organization. Schumpeter stressed the revolutionary character of the entrepreneur (and sometimes portrayed him with the same unilateral admiration as Marx showed for the revolutionary proletariat). But the conception of entrepreneurship as 'the function of innovation' opens the door to the critique that he exaggerates the importance of a function that seems potentially open to every person in business. To counteract this critique, Schumpeter drew on a range of sociological and psychological insights to demonstrate the entrepreneur's exceptional role and qualities.

Entrepreneurship, he argues, calls for a specific type of personality and conduct, which differs from the economic man. The entrepreneur takes advantage of rationally based components of his or her environment, such as money, science, and individual freedom, and he or she orients his or her conduct to rational values, but he or she is not the average product of bourgeois culture, which defines rationality from the narrower viewpoint of calculating one's short-term advantage. The entrepreneur acts on the basis of an autonomous drive to conquest and struggle, to achieve and create for its own sake, and also to establish a family dynasty. He or she is a bold leader, capable of thinking the new and grasping the essential, willing to act quickly, to understand by intuition, and to forgo the psychological resistances and social criticisms that always arise when new and innovative behavior is regarded as deviant and dangerous. This sets him off from the routine manager. While having some elements in common with religious and military leaders of the past, the entrepreneur is, however, the leader of a rational and antiheroic culture, and as a result does not excite the charismatic feelings and collective enthusiasm of those who make or defend whole civilizations.

Entrepreneurship is a specific historical phenomenon, which rests on the premise of a separate economic sphere differentiated from others. In previous epochs, the entrepreneurial function was fused with others in the actions of religious, political, and social leaders. In any historical society there is leadership, defined as the capacity to conceive and lead the making of innovations. What changes in the different historical contexts is the privileged sphere where leadership is applied, the one that is related to the core function for the survival and development of that

given society. Entrepreneurship is the specific historical form that leadership assumes in capitalism, its distinctive (and even essential) feature, closely linked to the bourgeois class. The bourgeoisie is the leading class, because bourgeois families have performed the innovating and leadership role in the economy and because they acquire, consolidate, and transfer prestige, power, and wealth to future generations.

The fate of entrepreneurship is closely linked to those of the bourgeoisie and capitalism. The conception of entrepreneurial innovation as the key element of capitalism implies that the weakening of the role of the innovative entrepreneur is seen as a basic factor—along with the fading away of bourgeois institutions—of the crisis of capitalism. Schumpeter predicted the progressive decay of the entrepreneurial function by virtue of the routinization of innovation in large oligopolies, which would render the entrepreneurial function superfluous and undermine the basis for continued bourgeois dominance. It is the very success of the capitalist firm that undermines the system.

Schumpeter's identification of the fate of the nineteenth-century entrepreneur with that of capitalism is largely responsible for his faulty prediction of the collapse of this system. In reality, capitalism has proved capable of fundamental transformations through that process of 'creative destruction' that Schumpeter had clearly perceived, but underestimated. Several different brands of capitalism exist, which proved compatible with the existence of very large firms and with state intervention and control of the economy.

In spite of his limitations, Schumpeter is the most influential scholar of entrepreneurship. Although often in a rather sketchy way, he asked the most relevant questions and provided important theoretical insights.

2. *The Multidisciplinary Approach to the Study of Entrepreneurship*

In recent decades entrepreneurship has been studied by various disciplines and through different approaches that were influenced by the classical studies, and particularly by Schumpeter's theory. For most economists—with notable exceptions from Schumpeter to Kirzner—the question of entrepreneurship is not problematic; entrepreneurial activities will emerge more or less spontaneously, as an instance of rational profit maximization, whenever economic conditions are favorable, such as availability of capital, labor, and technology, factor mobility, and access to markets. Other social scientists, mostly sociologists, social-psychologists, business historians, and anthropologists, tend to see entrepreneurship as a more problematic phenomenon, deeply embedded in societies and cultures; they focus on the influence of, and the mutual interplay among noneconomic factors, such as cultural norms and beliefs, class relations and collective action, state intervention and control, organizational structures, bounded solidarity and trust, deviant behavior and marginality status, and motivations for achievement. They also study how different historical and geographical settings may call for markedly dissimilar forms of entrepreneurship. Economic historians have pointed out that the personality type and behavior characteristics of last-century business leaders bears little resemblance to their counterparts of today. Comparative sociologists have shown strong differences in entrepreneurial roles and styles across countries and with regard to the firm's size and to the stages in the firm's life cycle.

Different disciplinary paradigms provide a first basis for classifying major approaches to the studies of entrepreneurship and management. Underlying these different approaches are the two fundamental dimensions of: (a) system variables versus actor variables (or, in different terms, macroanalytical models versus microanalytical models), and (b) structural versus cultural variables.

Metaphorically we can say that for entrepreneurship to emerge it is necessary that the 'seed' should find the appropriate 'ground.' Certain scholars focus on the seed, i.e., either on specific psychological traits of entrepreneurial personalities or on their social characteristics. Other scholars focus on the breeding ground, which is analyzed either in its structural factors (types of markets, factors of production, class and ethnic relations, state planning, etc.) or cultural factors (business ethics, social approval of economic activity, etc.). Others, still, emphasize the specific relation between the actor and the situation.

3. *The Entrepreneurial Personality*

Schumpeter argued that entrepreneurship calls for a specific type of personality and conduct, which differs from the simple rational conduct of the economic person. Psychological and sociopsychological studies focus on the distinctive traits of entrepreneurial behavior and personality. McClelland's comparative study tries to answer the question of why some societies are more likely to produce entrepreneurs than others. His major conclusion is that childhood experiences create in certain individuals a particular psychological factor, what he calls 'the need for achievement,' which is responsible for economic growth and decline. The inculcation of the achievement motive is due to child-rearing practices that stress standard for excellence, self-reliance training, maternal warmth, and low father dominance. In Western industrialized nations children internalize in family socialization attitudes and symbols which favor a higher need for achievement than do children raised elsewhere, and many of the high achievers have become entre-

preneurs. In underdeveloped countries McClelland finds fewer people with a high need for achievement; besides, of this smaller number, most go into fields other than business and industry. This interpretation of the motivational structure of entrepreneurship has rather overt normative implications; governments interested in economic growth should produce more high achievers by breaking traditional patterns and emphasizing other-directedness and should motivate the high achievers to enter business and industry. McClelland's need for achievement theory raises serious methodological problems and risks to be tautological; yet, because it is easily testable, it has fostered a large number of empirical researches.

In general, psychological approaches are questionable since they either tend to make economic activity too much a function of personality, underplaying the role of contextual influences, or to establish too simple a correlation between micro (personality) variables and macro (economic and sociological) variables. Gerschenkron (1962) criticizes this type of theory, arguing that the normal variation in child-rearing practices and in IQ will produce enough potential entrepreneurs in every society; the critical questions are the institutional factors that are available for entrepreneurial action.

A more sophisticated theory of the relation between socialization and entrepreneurial personality is that of Hagen (1962), who tries to answer both the questions of who the entrepreneurs are and why they come more often from certain social groups. Hagen's model combines a psychoanalytical interpretation of the entrepreneurial personality with an analysis of his or her condition as a member of a group that is deviant from the cultural norms of the larger society. According to Hagen, entrepreneurs tend to come from groups that suffered from withdrawal of status respect (i.e., groups who perceive that their purposes and values in life are not respected by other groups in society whom they respect and whose esteem they value). The loss of status by the social group leads to the breakdown of the family authority and is the trigger mechanism for changes in personality formation. However, the outcomes can be very different. Status withdrawal may only lead to sentiments of anger and anxiety and to retreatism. Or, in situations where the mother rejects the husband-father and shows an attitude of protective nurturance to her child, she may foster individualism and the development of a creative and self-reliant personality. Several generations are required to overcome social blockage, but then entrepreneurs can emerge.

The weakest point of Hagen's theory is that his attempts at verification are limited to a few historical cases. Yet, at least for the part based on status withdrawal, his model can rely on empirical evidence of disadvantaged minorities in complex national societies that contributed disproportionately to entrepreneurial activity and economic development (like the Jews and the Protestants in Europe, the Samurais in Japan, the Parsees in India). Hagen's model is in some way the *trait-d'union* between the psychological and the sociological approach; he considers macro-sociological variables, such as status withdrawal and social marginality, but only insofar as they influence personality development.

4. *The Social Context of Entrepreneurship: Marginal or Privileged Status?*

The context where entrepreneurship emerges has been studied with regard to: either the social position of the entrepreneurs (marginal/deviant or privileged), or the cultural values and attitudes that can favor it, or the structure of risks and opportunities. Sociological conceptions of entrepreneurship in terms of deviance and marginality status have a longstanding tradition. In *Der Moderne Kapitalismus* (1916–27, 1987), Sombart remarks that the creativity and the ability to break traditional values and patterns which characterize the capitalist entrepreneur can be found in all peoples, social groups, and religions, but that they are more frequent among the members of certain minorities, such as the heretics, the strangers, and the Jews. These groups are not completely accepted in the societies in which they live; they can therefore avoid more easily than others the traditional values and norms that regulated economic behavior in premodern Europe. Because of their minority status—both tolerated and oppressed—they tend to develop specific skills in the commercial and financial activities they are allowed to practice; and because of their acute sense of diversity, they maintain a strong group solidarity, which favors trust and therefore credit among the members of the group.

Building on both Sombart and Schumpeter, Hoselitz (1963) argues that entrepreneurs are deviant because of their ambiguous social position. Marginal groups such as the Jews and Greeks in medieval Europe, the Chinese in Southeast Asia, and Indians in East Africa, are peculiarly suited to make creative adjustments in situations of change and to develop genuine innovations in social behavior, because they act in a hostile social milieu, where prevailing attitudes are against innovation; they are excluded from political power so they concentrate on business; and being left outside the dominant value system, they are subjected to lesser sanctions for their deviant behavior.

More recent works on ethnic communities (Ward and Jenkins 1984, Waldinger et al. 1985) and women (Goffee and Scase 1985) follow a similar 'positional' approach. Structural factors within the larger society, such as racism, sexism, and credentialism, through processes of exclusionary closure create 'outsiders'; but some of these people form 'feeder groups' from which new entrepreneurs emerge, as in the case of

South African Indians who, embedded in organic networks and communities, were remarkably successful in circumventing the massive legal and political constraints of apartheid.

In their study (1992) of how Dominican and other immigrants 'made it' in the American economy Portes and Min Zho emphasize bounded solidarity and enforceable trust as sources of social capital, which do not stem from shared value orientations, but from the position of the ethnic minorities in the wider social structure. Citizens of China, Korea, or Cuba do not display any exceptional solidarity or *bienfeasance* in economic transactions when they are in their native countries. Such benefits stem from their being the members of an identifiable social minority in the host country. Bounded solidarity is created among immigrant customers, workers, and investors because they are treated as foreigners, and have a heightened awareness of the symbols of common nationhood. Enforceable trust is based on the ostracism of violators, cutting them off from sources of credit and opportunity in the ethnic economy.

The social marginality approach can be challenged on two grounds: on structural grounds, by those who argue that dominant classes in society can produce entrepreneurs more than marginal groups can, because of their access to economic, political, and social resources; and on cultural grounds, by those who stress core values and social approval as requisites of entrepreneurship. Social marginality and ethnic solidarity certainly play a role in the formation of entrepreneurship. But historical and comparative research on developed countries—both early industrializing countries (like Great Britain and France) and 'latecomers' (like Germany, Italy, and Japan)—show that most new entrepreneurs come from the already 'privileged' status groups of preindustrial societies, such as merchants, landowners, and wealthy artisans, who possess both the material and intellectual resources for economic achievement. Dobb's thesis of the revolutionary role played by yeomen and independent artisans (1946), Pirenne's analysis of the role of merchants in the formation of the urban bourgeoisie (1914), and Brentano's view of the acquisitive aristocracy as a protocapitalist class (1916), are all instances of the importance of social groups that occupied well-established positions in the 'traditional' societies and which paid a key modernizing role. Although in different ways, studies of modernization and of the world economy by Barrington Moore (1966), Bendix (1978), Wallerstein (1979), and others, show the importance of socially established classes in entrepreneurial formation and consolidation.

Even in a more open society as the United States, the contribution of immigrants and lower classes to the formation of entrepreneurship has been exaggerated. Scholars as different as Mills (1945) and the business historians of the Harvard Center for Entrepreneurial History (Miller 1952) found a rather stable recruitment pattern: most of the business elite in the period of the great American industrialization (1870–1910) came from landowning or entrepreneurial families, whereas lower classes contributed between 10 and 20 percent of the membership. A similar critique of the popular origins of entrepreneurs in industrialized countries comes also from major studies in social mobility (Lipset and Bendix 1959).

More recent studies of the relationship between entrepreneurship and social stratification focus on the social and cultural specificities of women entrepreneurs (Goffee and Scase 1985); and stress the importance of the middle classes as a breeding ground for small entrepreneurs. Martin and Norman (1970) argue that small entrepreneurs are particularly suited to contemporary society, because of facilitating conditions both on the supply side of entrepreneurial talents, and on the demand side of services requiring new market niches. And Bechhofer and Elliot (1981), see them as the true custodians of the traditional values upon which the capitalist social order was built (passionate individualism and independence, the moral evaluation of work, laissez faire competition, and the belief that inequality is the result of differential distribution of talent and effort).

5. *Entrepreneurship and Culture*

Studies stressing cultural context variables are indebted both to Weber and Schumpeter. The classical exposé is Weber's comparative analysis of religious ethics and economic action in the rise of capitalism. Schumpeter inspired the historical research of the Harvard Center for Entrepreneurial History. Neo-Weberian research focuses on the degree to which the forces of rationalization responsible for dislodging individuals from their embeddedness in nature, religion, and tradition continue to shape economic growth and social modernization. Following Weber and Parsons (and in a similar vein to Marshall Berman's perceptive study of modernity, 1982), scholars like Peter Berger, Brigitte Berger, and Kellner (Berger et al. 1973), stress the typical cognitive style that distinguishes modern consciousness, such as instrumental rationality and the propensity to combine and recombine activities for the achievement of calculated ends. Brigitte Berger's cultural approach to entrepreneurship argues that economic growth develops from the 'bottom up,' not from 'top down': ordinary individuals, competing with one another for achieving a variety of goals (including economic profit and self-advancement), create in their everyday activities, practices, habits, and ideas the basis for other distinctly modern institutions to emerge that may mediate between them and distant, large-scale structures of society.

Evidence for this cultural approach is found in various empirical researches, including: Martin's

(1990) analysis of the role of Protestant sects in generating a dynamic process among segments of the urban poor in contemporary Latin American cities that fosters entrepreneurial activities; Redding's study of the relation between basic aspects of Chinese culture—such as Confucian ethic and family attitudes—and the entrepreneurial behavior among overseas Chinese (1990); Landa's thesis of the entrepreneurial success of ethnically homogeneous middlemen groups in Africa and in Southeast Asia (1991).

These types of study can be challenged on various grounds. First, these studies often make some confusion between the hegemonic culture and marginal groups' subcultures: sometimes it is the contrast with the core cultural values which is presented as a cause of entrepreneurial success; some other times it is the convergence with the core value of instrumental rationality. Second, the concept of culture is often stretched so far as to include social interaction in general and all sorts of social networks, without paying attention to structural variables like patterns of solidarity, class, status and power relations, legal norms, and state arrangements and policies, which fundamentally affect social interaction. Third, cultural explanations can be accused of having little predictive power since they are invoked only after a particular group has demonstrated its economic prowess; and of being ultimately tautological, since if a certain minority is successful, it must be because it originally had, or later acquired, the right values.

Another approach that is framed in the cultural paradigm is the social attitudes (or cultural legitimation) model, developed by the members of the Harvard Center for Entrepreneurial History, such as Landes (1951), Sawyer (1952), and Cochran (1949) and by sociologists such as Lipset (1967).Comparing the United States and Latin America, both Cochran and Lipset explain differences in economic development in terms of the degree of legitimation of entrepreneurship. Cultural norms, role expectations, and social sanctions can either favor or hinder innovation. Comparing France and the United States, both Landes and Sawyer maintain that the different pace and degree of economic development in the two countries were due to their different historical heritage; while in France the feudal heritage had left a consistent residual of social attitudes hostile to entrepreneurship that limited the recruitment of entrepreneurs, in the United States the absence of a feudal past had allowed the growth of a sociocultural context which was especially receptive for innovation and entrepreneurship.

In a well-known debate, Gerschenkron (1962) refutes Landes and Sawyer's thesis by arguing that the error of giving too much importance to social attitudes lies in assuming the existence of a homogeneous and generalized value system in society. As counterevidence Gerschenkron brings the cases of eighteenth-century France's *fermiers généraux* and of nineteenth-century Russia's emancipated serfs, who became entrepreneurs in spite of an unfavorable cultural environment. Gerschenkron's critique is framed in his theory of the different paths to economic development; according to him, different countries develop through a different mix of what he calls the 'institutional agents' of development, such as private entrepreneurs, merchant banks, and governments.

6. *The Entrepreneur in the Structure of Risks and Opportunities*

Both nonmainstream economics and the situational approach to the study of entrepreneurship focus on markets as structures of risks and opportunities. Mainstream accounts of economics—like those by Knights and by post World War II development economists—assume economic actors to have knowledge of all the opportunities open to them and to behave in a fixed, rational, maximizing way, once appropriate incentives are introduced. But, when these assumptions are relaxed, and market segmentation, ignorance, impeded factor mobility, and pervasive administrative controls are taken into account, the 'extraordinary' role of the entrepreneur becomes apparent, as does the need to analyze more carefully the factors that can favor his or her formation. The most interesting contemporary interpretation of this kind is Kirzner's approach.

Under the influence of Schumpeter, Mises, and Hayek, Kirzner (1973) points out that human action is meaningful only in relation to the purposes, plans, and expectations of the actor; it is partly guided by maximizing criteria, but alertness, creativity, and judgment—the typical features of the entrepreneur—also influence what we do. For the market process to emerge, an additional element is required that is not comprehensible within the narrow conceptual limits of economizing behavior, that is entrepreneurial action. This action is favored by an appropriate system of economic and social incentives, such as profit, fame, prestige, power. Entrepreneurial competition is a discovery procedure of profit opportunities, and the competitive system is a fundamentally social process, dependent on the free interplay of individuals.

Pushing Kirzner's conception further, Lavoie (1991) argues that the theory of entrepreneurship could more fully account for economic change if it were built on the hermeneutical theory of language and culture. According to him, any individual interpretive process is embedded in culture and meaningful only in relation to culture. Lavoie adheres partially to methodological individualism, only insofar as it assumes that the social whole has no purpose, but is a complex resulting from the choices of the participating individuals. But he rejects its other assumption, shared by most economists, that the rational action of the isolated individual serves as the foundation of the analysis of

markets. Individual minds are interdependent parts of an integral process of cultural dynamics; they could not exist in isolated, cultureless, brains devoid of language. With Lavoie, we come full circle, back to culture.

By the mid-1960s, scholars from diverse disciplines contributed to what is known as the situational (or social development) approach. Instead of looking at the institutional conditions of entrepreneurship, they asked themselves what actually an individual does when being an entrepreneur. Some, like Glade, and Greenfield and Strickon, give primacy to social action over structural context; others, like Gibb and Ritchie, stress structural context over social action.

Glade (1967) starts criticizing McClelland and Hagen's models as instances of comparative statics, describing social behavior in two contrasting types of system, one underdeveloped and the other developed, which are both in a steady state; what they lack is a theory of change which explains the transition from one state to the other. Glade maintains that actors make choices and decisions within social settings that are opportunity structures changing over time. The integral features of any given situation are both an 'objective' structure of economic opportunity and a subjective capacity of participants to perceive and act upon such opportunities. Entrepreneurs are individuals who can recognize the new opportunities and take advantage of them, without losing them to others.

Glade's situational analysis to the study of entrepreneurship has been criticized and elaborated upon by Greenfield and Strickon (1986). They propose using Darwinian biology as a metaphor for the study of change; as Darwin rejected typological, essentialist biology, they reject fixed types of entrepreneurship—the analogue of the immutable species—and recognize existing diversity of behavior within specific human populations (or communities), which at its extremes encompasses innovation and novelty. These diverse behaviors interact with their environments, to produce outcomes that are evaluated both by the actor and others. Those innovations judged more advantageous in terms of the standard prevailing within the group may be selected, learned, and imitated, with the result being the establishment of a statistical pattern. Society, culture, religion, politics, and economics are not seen as entities with a reality of their own, but as statistical patterns abstracted from the variable behaviors of the members of specific communities.

An attempt to combine the situational approach with that of reference groups is the social development model exemplified by Gibb and Ritchie (1982). Entrepreneurs change continuously in terms of the situations they encounter and the social groups to which they relate. It is their interactions with specific social contexts and reference groups which produce distinctive ambitions and behavior. The wide range of influences and interactions makes it impossible to define a single entrepreneurial model, but it does not prevent the development of typologies of entrepreneurs, like the one they develop, where four types ('improvisers,' 'revisionists,' 'superceders,' and 'reverters') are identified as being at the center of different sets of influences.

7. *An Integrated Approach to the Study of Entrepreneurship*

The conclusion we can draw from the state of the art of the research on entrepreneurship is that the most interesting studies are often located at the borders between disciplines, such as those by economists who reject simple rational models and recognize the influence of social interaction and culture, or by sociologists and anthropologists who reject oversocialized conceptions of man and take into account the strategies of individual actors. The most convincing interpretations are those away from the extremes of the actor-system continuum, which dismiss oversimplified views of the interaction between the actor and the systemic context.

A recent example of this integrated approach, close to my own view, is that expressed by Aldrich and Waldinger in their essay on ethnic entrepreneurship (1990). It focuses on such diverse variables as structure of markets, access to ownership, state policies, group characteristics, predisposing factors, and resource mobilization. As for types of economic environments that might support new immigrant entrepreneurs, they identify core urban markets increasingly abandoned by large food retailers, markets where economies of scale are low, markets affected by instability or uncertainty, markets for exotic goods; as for access to ownership, the relevant conditions are the level of interethnic competition for jobs and businesses, patterns of residential segregation and succession, and state immigration and labor market policies; as for group characteristics, predisposing factors, and resource mobilization, they focus on selective migration trends, settlement patterns, culture and aspiration levels, ethnic social networks, and organizing capacities. Although Aldrich and Waldinger's preference is for structural and institutional variables (like market niches and group resource mobilization) over cultural variables (like group cultural heritage), they are willing to take into account all covariables which can reasonably affect entrepreneurial formation in ethnic communities.

What is needed in future research on entrepreneurship is a multidisciplinary comparative approach, capable of integrating the analysis of the context (market, social structure, culture) with a theory of the actor (both individual or collective) with his or her motives, values, attitudes, cognitive processes, and perceived interests.

See also: Competitive Strategies: Organizational; Leadership in Organizations, Psychology of; Leadership, Psychology of; Organizational Decision Making; Schumpeter, Joseph A (1883–1950); Weber, Max (1864–1920)

Bibliography

Aldrich H E, Waldinger R 1990 Ethnicity and entrepreneurship. *Annual Review of Sociology* **16**: 111–35

Bechhofer F, Elliott B (eds.) 1981 *The Petite Bourgeoisie: Comparative Studies of an Uneasy Stratum*. MacMillan, London

Berger P L, Berger B, Kellner H 1973 *The Homeless Mind: Modernization and Consciousness*. Random House, New York

Bottomore T, Brym R J (eds.) 1989 *The Capitalist Class. An International Study*. Harvester Wheatsheaf, New York

Cardoso F H 1972 *Empresario industrial e desenvolvimento econômico no Brasil*. Difusão Européia do Livro, São Paulo, Brazil

Chandler A D Jr 1962 *Strategy and Structure: Chapters in the History of Industrial Enterprise*. MIT Press, Cambridge, MA

Cochran T C 1949 Role and sanction in American entrepreneurial history. In: *Change and the Entrepreneur, Postulates and the Patterns for Entrepreneurial History*. Harvard University Press, Cambridge, MA

Cole A H 1959 *Business Enterprise in its Social Setting*. Harvard University Press, Cambridge, MA

Curran J, Stanworth J 1983 The sociology of petit capitalism: A trend report. *Sociology* **20**: 265–79

Dobb M H 1946 *Studies in the Development of Capitalism*. Routledge & Kegan Paul, London

Gerschenkron A 1962 *Economic Backwardness in Historical Perspective, a Book of Essays*. Harvard University Press, Cambridge, MA

Gibb A, Ritchie J 1982 Understanding the process of starting small businesses. *European Small Business Journal* **1**: 26–45

Glade W P 1967 Approaches to a theory of entrepreneurial formation. *Explorations in Entrepreneurial History, Series 2* **4**: 245–59

Goffee R, Scase R 1985 *Women in Charge: The Experiences of Female Entrepreneurs*. Allen & Unwin, London

Greenfield S M, Strickon A (eds.) 1986 *Entrepreneurship and Social Change*. University Press of America, Lanham

Hagen E E 1962 *On the Theory of Social Change: How Economic Growth Begins*. Dorsey Press, Homewood, IL

Hoselitz B F 1963 Entrepreneurship and traditional elites. *Explorations in Entrepreneurial History, Series 2* **1**: 36–49

Kirzner I M 1973 *Competition and Entrepreneurship*. University of Chicago Press, Chicago

Lavoie D 1991 The discovery and interpretation of profit opportunities: Culture and the Kiznerian entrepreneur. In: Berger B (ed.)

Lipset S M 1967 Values, education, and entrepreneurship. In: Lipset S M, Solari A (eds.) *Elites in Latin America*. Oxford University Press, London

Martinelli A, Smelser N J (eds.) 1990 *Economy and Society: Overviews in Economic Sociology*. Sage, London

Martinelli A 1994 Entrepreneurship and management. In: Smelser N J, Swedberg R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ

Marx K 1867 *Das Kapital*. Trans. 1906 *Capital*. The Modern Library, New York

McClelland D C 1961 *The Achieving Society*. D. Van Nostrand Co, Princeton, NJ

Miller W (ed.) 1952 Men in business. *Essays in the History of Entrepreneurship*. Harvard University Press, Cambridge, MA

Scase R, Goffee R 1982 *The Entrepreneurial Middle Class*. Croom Helm, London

Schumpeter J A 1934 *The Theory of Economic Development, an Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*. Harvard University Press, Cambridge, MA

Schumpeter J 1942 *Capitalism, Socialism and Democracy*. Harper and Row, New York

Sombart W 1916–27 1987 *Der Moderne Kapitalismus*. Munich, Germany

Weber M 1978 *Economy and Society: An Outline of Interpretative Sociology*. University of California Press, Berkeley, CA

A. Martinelli

Entrepreneurship, Psychology of

Small and medium sized enterprises are important for the economy because they are the major agents of economic growth and employment (about 99 percent of the European companies are small or medium sized and they provide 66 percent of the working places, European Council for Small Business newsletter, 1997). They add jobs faster than bigger companies in the developed and underdeveloped world. Small-scale firms are highly adaptable and able to act quickly and innovatively.

Psychological approaches to entrepreneurship have experienced a revitalization recently because of the future importance of small-scale entrepreneurs and because the entrepreneur is at the boundary line of individual work psychology (personality, work activities, etc.), organizational psychology (founders of the organization have an enormous influence on it), and market psychology (economic activities in the market). Many organizational issues, for example, the influence of human resource practice, can also be studied in entrepreneurs. Essentially, all aspects of psychology are implicated when studying entrepreneurs (Rauch and Frese 2000).

Relevant literature in this area is distributed in many outlets and can be found in such diverse journals as the *Journal of Applied Psychology*, *Academy of Management Journal and Review*, *Administrative Science Quarterly*, *Journal of Small Business Management*, *Journal of Business Venturing*, *Entrepreneurship Theory and Practice*, *Journal of Management*, *Small Business Economics*, *World Development*, *Strategic Management Journal*, *Organization Studies*, and there are many articles in conference procedures such as *Frontiers of Entrepreneurship Research*, and *International Council of Small Business Conference Proceedings*.

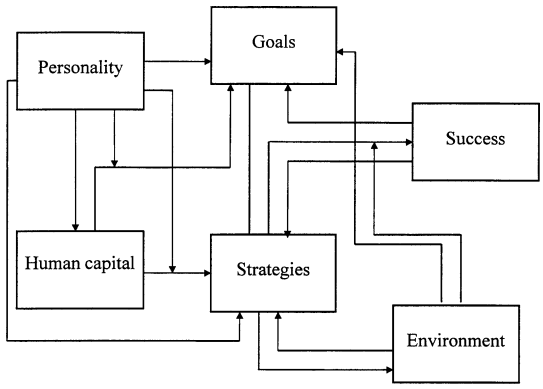

Figure 1
The Giessen–Amsterdam Model of small business owners' success

1. *A General Psychological Model of Entrepreneurial Success*

Figure 1 presents a general psychological model of entrepreneurial success. It helps us to organize this contribution, and also has some controversial implications. For example, it does not hypothesize any direct links from personality, human capital, or environment to success because we assume that there is no success without actions. Actions are mainly determined by goals and strategies. Therefore, according to this model, psychological strategies of actions are the bottleneck through which all of entrepreneurial success is accomplished or not accomplished. The model can also be used to understand the different levels of analysis: one can differentiate the organizational level and the individual level of the firm owner with regard to each issue in Fig. 1. The level of analysis issue has a slightly different meaning in the area of entrepreneurship because company size determines which level is the adequate one. In large companies, the right level of analysis is the organizational level, in small firms, the firm owner is typically the source of action of this firm. When there are only four or five employees in a firm, the owner usually has a much stronger impact on company policy, company culture, and the company's actions than in larger firms and an individual level of analysis can be used profitably for these firms.

2. *Definition Issues: Entrepreneurs, Business Owners, and Other Concepts*

There is no agreed-upon definition of entrepreneurship. Moreover, founders and owner/managers are a highly heterogeneous group that defies a common definition. Probably the best strategy is to use a behavioral definition because it does not make any further assumptions of success, growth, or failure (e.g., Hisrich 1990): entrepreneurship is the study of founders/managers of organizations. However, one should also be open to the fact that there is a growing interest in entrepreneurship within large organizations.

3. *Characteristics of the Entrepreneur*

It is necessary to separate the emergence and the success of entrepreneurs. There may be different processes by which a person decides to become an entrepreneur and by which a person achieves entrepreneurial success (Utsch et al. 1999). Personality characteristics may be more important for the decision to become a founder than for success.

3.1 *Personality and Emergence of Entrepreneurship*

McClelland's (McClelland and Winter 1971) early work suggested that need for achievement should be higher in people who start a business. This is indeed the case as a quantitative review shows (Rauch and Frese 2000). A similar result appears for locus of control (Rotter 1966). Business owners have a slightly higher internal locus of control than other populations (Rauch and Frese 2000). Other studies have found a high degree of innovativeness, competitive aggressiveness, and autonomy (Utsch et al. 1999), Protestant work ethic beliefs (Bonnett and Furnham 1991), or risk taking (Begley and Boyd 1987). The literature about the emergence of entrepreneurship highlights that entrepreneurs are different from managers and other groups.

More recently, researchers developed more sophisticated personality concepts that match the personality variables with the behavioral requirements of an entrepreneur, for example, the Entrepreneurial Attitude Scale (EAO), which consists of achievement, self-esteem, personal control, and innovation (Robinson et al. 1991) or task motivation theory (Miner et al. 1994).

3.2 *Personality and Success*

The most frequently studied personality characteristics were need for achievement, internal locus of control, and risk-taking. A quantitative review showed a weighted uncorrected mean correlation of 0.13 between need for achievement and success. It is important to note that there is a reduced variance in these samples because of the fact that emergence is also related to the achievement motive (Rauch and Frese 2000). Studies showed that achievement motive could be enhanced and that this leads to a higher success in business (e.g., McClelland and Winter 1971). A similar relationship with success also appeared for locus of control (Rauch and Frese 2000). In contrast, high risk-taking is not or even negatively

associated with business success (Rauch and Frese 2000).

3.3 Personality Reconsidered

Thus, there are differences between entrepreneurs and managers, and correlations between personality and success, but these correlations are not high. It is, therefore, understandable that criticisms of a personality approach have appeared.

However, both approaches—the personality proponents and its critiques—have overlooked the significant advances that have been made in personality research during the last 20 years that need to be made useful for entrepreneurship research. The most important issue is certainly that specific behaviors (such as starting up a business or using a certain approach to the market) works only through mediating processes (cf. the Giessen–Amsterdam model in Fig. 1). For example, planning mediated the relationship between achievement orientations and success.

Second, the personality variable has to be specific enough to predict specific (entrepreneurial) behavior. For example, Miner et al.'s (1994) task motivation theory explained 15–24 percent of variance in growth measures.

Third, interaction models suggest that one looks at which personality trait helps in which environment. Thus, one would have to look at interactions of personality with environmental conditions.

Finally, no one personality trait will ever have a strong relationship with success because success is determined by many factors.

4. Human Capital

Human capital theory is concerned with knowledge and experiences of small-scale business owners. The general assumption is that the human capital of the founder improves small firms' chances of survival (Bruederl et al. 1992). Human capital acts as a resource. However, human capital theory studies usually assume that experiences are translated into knowledge and skills. This assumption is problematic, however, because length of experience is not necessarily a good predictor of expertise (Sonnentag 1995). Therefore, it is not surprising that human capital factors, such as length of managerial or industry experiences or education, are not strong predictors of success, although in large-scale studies they usually are significant (Bruederl et al. 1992, Rauch and Frese 2000).

5. Goals

One can distinguish between goals related to the start-up of an enterprise and goals related to the existing enterprise. Goals or motives for becoming self-employed can be categorized into push and pull factors. Push factors imply that a current situation (e.g., the job or unemployment) is unsatisfying, pull factors are related to desires for being independent and doing what one likes to do. While there are differences among entrepreneurs, there is little evidence that goals related to developing a business are related to success (Frese 2000).

6. Strategies (Content, Process, Entrepreneurial Orientations)

From a psychological perspective strategies are directly related to goal-oriented actions. It is useful to distinguish between three dimensions of business strategy: content, process, and orientation. All three strategy dimensions can in principle be crossed with another.

First, strategic content is concerned with the type of business decisions vis-à-vis the customers, suppliers, employees, products, production factors, marketing, capital, competitors. Studies in this area are often done by economists although psychological issues are important as well, for example, how to convince banks to give credit, active strategies on the market, for example in developing a niche product, and so forth.

Second, the strategic process is concerned with formulation and implementation of strategic decisions (Olson and Bokor 1995). One issue is planning, which is related to success (Schwenk and Shrader 1993). Frese and his co-workers (Frese 2000) have recently presented a theoretical typology of psychological strategies that are differentiated along the dimensions of proactivity and planning. Planning strategy implies that a top-down planning process is used that is also highly proactive. Critical point planning implies that an important issue is planned but other issues are not. It is proactive and involves a smaller amount of planning. Opportunistic strategy implies very little planning but a high degree of proactivity: one looks out for opportunities and takes them without any detailed planning beforehand. Finally, the reactive strategy implies that one is neither proactive nor planning: one simply reacts to the demands of the situation. The latter is negatively correlated with success in various countries (four African countries), while all others are positively correlated with success (Frese 2000); however, the relationship between planning and success depends on the situation to a certain extent. A Dutch longitudinal study has also shown that there is a reciprocal determinism from a reactive strategy to failure and from failure to reactive strategy (Van Gelderen and Frese 1998).

Third, orientation implies an attitude towards one's strategy: why a strategy is played out. Lumpkin and Dess (1996) conceptualized entrepreneurial orientation to consist of five dimensions: innovation, pro-

activeness, risk-taking, autonomy, and competitive aggressiveness. Covin and Slevin (1986) showed among others that entrepreneurial orientation was highly related with company performance ($r = 0.39$, $p < 0.01$). The relationship between entrepreneurial orientation and success may also be contingent on environmental and organizational factors.

7. *Environmental Conditions*

Each enterprise is nested in a specific environment. The task environment can be divided into three bipolar dimensions: complexity, dynamism, and munificence. Complexity describes the homogeneity versus heterogeneity of an environment. In a complex environment it is more difficult to get and to consider all the necessary information than in an easy environment. Dynamism describes the variability and unpredictability of the environment. Munificence falls into two subconcepts: ease of getting customers and ease of getting capital.

Sharfman and Dean (1991) showed that munificence had no significant relationship with performance, but complexity and dynamism were positively related to success. Thus, an unfavorable environment—the dynamic environment—has positive consequences. According to Swaminathan (1996), organizations founded in adverse environments have a higher initial mortality rate. But beyond a certain age, the surviving organizations had a lower mortality rate than those founded in a more friendly environment.

8. *Other Issues of Psychological Entrepreneurship Research*

There are other psychological issues that have not been studied as much as the ones discussed above but that are potentially interesting (see Rauch and Frese 2000). Among them are the effects of networks, information search activities to get feedback, and organizational life cycle models. Moreover, there are leadership issues and one can study whether visionary leadership, communication, delegating, and performance facilitation are related to success. Social psychological and cognitive factors are most likely related to entrepreneurial outcomes, such as values and culture, attributional theory, and problem-solving styles. Other issues of this type are concerned with learning, minorities, human resource management, learning and training, feedback processing, transition from business founder to manager, financing, organizational culture, and others. One fascinating topic is the issue of making psychological entrepreneurship research useful for developing countries.

9. *Conclusion*

Psychological approaches to entrepreneurship are fascinating both for entrepreneurship and psychology. Entrepreneurship can profit from this interface between business and psychology because psychological variables are clearly related to entrepreneurial entry and success. Psychological variables (most notably action-related concepts) function as mediators in the process that leads to success (e.g., strategies). For psychology, entrepreneurship is interesting because it combines the following features.

(a) The level of analysis question is related to the dynamic of enterprise growth; in the beginning, a small scale enterprise is best described by looking at the owner. However, in somewhat more mature enterprises, the level of analysis has to change because more delegation, management, and implementation are necessary.

(b) Some interesting organizational hypotheses can better be studied with small-scale entrepreneurs than with large organizations. A good example is the study of contingency theories. Small-scale enterprises are more coherent than larger ones and, therefore, contingency models can be tested better.

(c) Even large organization attempt to mimic small enterprises, and stress intrapreneurship, innovation, and personal initiative. There is no doubt that future workplaces will stress innovation and personal initiative more strongly and we need to know how small-scale entrepreneurs act.

(d) Interdisciplinary cross-fertilization takes place in this area.

We have reported a number of different models in this review; they are often presented to be contradictory. For example, some people have pitted personality approaches against human capital approaches. As Fig. 1 shows, we assume that they coexist and can influence each other (e.g., IQ has an influence on the development of skills and knowledge). An integration of various approaches to make real headway towards understanding a societally important phenomenon—entrepreneurship—is called for and should produce challenging research.

See also: Business History; Entrepreneurship; Human Capital: Educational Aspects; Human Resource Management, Psychology of; Identity in Childhood and Adolescence; Innovation: Organizational; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Personality and Risk Taking

Bibliography

Begley T M, Boyd D B 1987 Psychological characteristics associated with performance in entrepreneurial firms and small businesses. *Journal of Business Venturing* **2**: 79–93

Bonnett C, Furnham A 1991 Who wants to be an entrepreneur? A study of adolescents interested in a Young Enterprise scheme. *Journal of Economic Psychology* **12**: 465–78
Bruederl J, Preisendoerfer P, Ziegler R 1992 Survival chances of newly founded business organizations. *American Sociological Review* **57**: 227–42
Covin J G, Slevin D P 1986 The development and testing of an organizational-level entrepreneurship scale. In: Ronstadt R, Hornaday J A, Peterson R, Vesper K H (eds.) *Frontiers of Entrepreneurship Research*. Babson College, Wellesley, MA, pp. 628–39
Frese M (ed.) 2000 *Success and Failure of Microbusiness Owners in Africa: A Psychological Approach*. Greenwood, Westport, CT
Hisrich R D 1990 Entrepreneurship/Intrapreneurship. *American Psychologist* **45**(2): 209–22
Lumpkin G T, Dess G D 1996 Clarifying the entrepreneurial orientation construct and linking it to performance. *Academy of Management Review* **21**(1): 135–72
McClelland D C, Winter D G 1971 *Motivating Economic Achievement*. Free Press, New York
Miner J B, Smith N R, Bracker J S 1994 Role of entrepreneurial task motivation in the growth of technologically innovative firms. Interpretations from follow-up data. *Journal of Applied Psychology* **79**(4): 627–30
Olson P D, Bokor D W 1995 Strategy process–content interaction: Effects on growth performance in small, start up firms. *Journal of Small Business Management* **27**(1): 34–44
Rauch A, Frese M 2000 Psychological approaches to entrepreneurial success: A general model and an overview of findings. In: Cooper C L, Robertson I T (eds.) *International Review of Industrial and Organizational Psychology, 2000*. Wiley, Chichester, UK, pp. 101–41
Robinson P B, Stimpson D V, Huefner J C, Hunt H K 1991 An attitude approach to the prediction of entrepreneurship. *Entrepreneurship Theory and Practice* **2**: 13–31
Rotter J B 1966 Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs* **609**(80): 1
Schwenk C R, Shrader C B 1993 Effects of formal strategic planning on financial performance in small firms: A meta-analysis. *Entrepreneurship Theory and Practice* **17**: 48–53
Sharfman M P, Dean J W 1991 Conceptualizing and measuring the organizational environment: A multidimensional approach. *Journal of Management* **17**(4): 681–700
Sonnentag S 1995 Excellent software professionals: Experience, work activities, and perceptions by peers. *Behaviour & Information Technology* **14**: 289–99
Swaminathan A 1996 Environmental conditions at founding and organizational mortality: A trial-by-fire model. *Academy of Management Journal* **39**(5): 1350–77
Utsch A, Rauch A, Rothfuss R, Frese M 1999 Who becomes a small scale entrepreneur in a post-socialist environment: On the differences between entrepreneurs and managers in East Germany. *Journal of Small Business Management* **37**(3): 31–41
Van Gelderen M, Frese M 1998 Strategy process as a characteristic of small scale business owners: Relationships with success in a longitudinal study. In: Reynolds P D, Bygrave W D, Carter N M, Manigart S, Mason C M, Meyer G D, Shaver K G (eds.) *Frontiers of Entrepreneurship Research*. Babson College, Babson Park, MS, pp. 234–48

M. Frese and A. Rauch

Environment and Anthropology

Anthropologists have been concerned with interaction between human populations and their environments since the middle of the nineteenth century. Early treatments were largely in terms of environmental determinism, replaced increasingly from around 1900 by 'possibilist' approaches which specified only a constraining role on society for environmental factors. A major paradigm throughout the twentieth century has been that of *ecological anthropology*, informed by conceptions of ecology which have been to varying degrees Darwinian, emphasizing the inter-relationship of social, cultural, biotic, and physical variables within a system, and centrally concerned with the concept of adaptation. This essentially positivist program has been augmented, and in part superseded, since the 1980s by one which uses interpretative, cognitive, and sociological approaches to understand peoples' perceptions of nature, environmental problems, and collective responses to them. These methodologies are not easily brought within the framework of ecological anthropology, and in recent years the term *environmental anthropology* has appeared as a more inclusive description. Approaches to human ecology which are concerned only with the biological characteristics of populations are not dealt with in this article.

1. *From Sociological Possibilism to Cultural Ecology*

During the first four decades of the twentieth century, anthropological theory was dominated by the ideas of Emile Durkheim (1858–1917) and Franz Boas (1858–1942), for whom environmental factors were for the most part secondary, and who accorded them a permissive rather than a determinant role. Nevertheless, the period saw the appearance of a number of classic studies, including Boas's own study and that of Mauss and Beuchat on Eskimo patterns of settlement.

The first explicit use of the concept of ecology in anthropology is found in the work of Julian Steward (1902–72) during the 1930s. For Steward, the concept of cultural adaptation is paramount, and the key adaptive strategies of a particular culture are located in a core of social institutions and technical arrangements directly focused on food-getting. The recognition of distinctive subsistence strategies provided the basis for the delineation of cultural types, which Steward maintained evolved multilineally, rather than in the unilinear fashion advocated by many nineteenth-century thinkers. In a modified form, Steward's ideas have continued to be influential, through the work of scholars such as Robert Netting. However, his theory of cultural ecology (see *Ecology, Cultural*) depends on a definition of adaptation, and a division between organic and superorganic levels of

explanation, and between a core of key adaptive traits and a neutral periphery, which recent writers have rejected.

2. *The Influence of Bioenergetics and Systems Ecology*

Advances within biology linked to the notion of ecosystem, the empirical measurement of energy flow, and the employment of the language of cybernetics and systems theory, led during the 1960s to a new formulation of ecological problems in anthropology. The special character assigned by Steward to the superorganic was passed over in favor of a view of human behavior in many respects functionally equivalent to that of other animals. Descriptions of ecological interactions began to involve computations of carrying capacity (see *Human–Environment Relationship: Carrying Capacity*), and estimates of energy transfer and efficiency (Ellen 1982). Adaptation became a matter of webs of mutual causality. In particular, there developed an interest in the way in which cultural institutions might serve to regulate certain systems of which human populations are part. All of these trends are demonstrated in the seminal work of Rappaport (1968), undertaken among the New Guinea Maring.

The technical and conceptual shortcomings of such approaches have been criticized in retrospect: inadequate samples and indices of what constituted success and adaptation and a theoretical naivete wedded to a prevailing sociological cybernetics which underestimated the difficulties of defining system boundaries (and thus of demonstrating feedback loops), and which neglected the role of historical contingency. This contributed to a skepticism concerning the more extreme proposition that certain kinds of small-scale society have built-in mechanisms for maintaining environmental balance through homeostasis. Although, other things being equal, it does seem that small populations with limited equipment and resources have benign effects on the environment (temptingly but erroneously interpreted as 'harmony'), for some in the environmental movement the inspiration offered by such cultures has led to a degree of idealization of the extent to which the values of particular nonindustrial and animistic peoples equate with environmental practices.

This debate has been particularly apparent with respect to the native peoples of North America. However, debunking the existence of 'ecological Edens' is not to downplay either the extent of the interconnections between apparently esoteric aspects of culture and the pragmatic well-being of subsistence systems, or the manifestly sophisticated knowledgeability of many nonindustrial peoples. Two studies which capture the former in a powerful way are Lansing's (1991) analysis of the linkage between the Balinese temple hierarchy and irrigation decision-making, and Richards' (1985) demonstration of the long-term rational behavior of Mende rice farmers in Sierra Leone. Both vindicate traditional local knowledge as against that of successive external bureaucracies, and thereby provide an antidote to top-down development programs. This general trend is evident in the growing literature on indigenous environmental knowledge (e.g., Warren et al. 1995) (see *Indigenous Knowledge and Technology*).

Some human systems *do* maintain a degree of integrity and are relatively well defined, while it is both practically and intuitively a relatively simple matter to isolate 'systems' which make sense of the data and which may be used in the investigation of a wide range of problems. Recognizing the openness of human systems is not inconsistent with employing some notion of closure to determine what universe of variables to examine. The arbitrary and analytic character of most boundaries, together with the reality of exogenous intrusions of varying types, magnitudes, and origins, does not inevitably mean accepting a notion of receding causation or infinite regress whereby the critical determinants always seem to lie outside the local system being examined—a view which comes dangerously close to providing no explanation at all. Vayda (1983) has sought to circumvent the problem by advocating a 'progressive contextualization' focusing on particular locally-significant 'people–environment interactions' and placing them in increasingly wider and denser contexts without making prior assumptions as to what the context might be. However, all explanations presume some kind of context and boundary (Moran 1990).

3. *Neo-Darwinist Approaches*

One response to the perceived deficiencies of systems approaches has been inspired by economic formalism and neo-Darwinism, both characterized by methodological individualism. This trend is evident in studies emphasizing how people actually cope with environmental hazards (Vayda and McCay 1975) (see *Environmental Risk and Hazards*), and in optimal foraging theory (Winterhalder and Smith 1981) and evolutionary ecology (Borgerhoff 1991). In this latter, adaptation and change are seen as providing an indefinite number of alternative strategies for enhancing reproductive fitness and survival depending on particular environmental circumstances. Such approaches have been criticized for their simple-mindedness at the level of ethnographic analysis, for using 'social' to mean no more than observable interactions between conspecifics, and for their failure to understand the cultural significance of subsistence, or to consider that the organization of consciousness might be relevant. However, more recently it has been acknowledged that the role of genes is more complex:

enabling the generation and sustenance of diverse kinds of social relationship and permitting (but not determining) the effective factors which contribute to their formation. At the same time, the properties of human social interaction themselves are seen to condition the adaptive value of strategies adopted; individuals thus adapting to the constraints of the social environment in which they develop, and the outcome of consumption influencing the value of alternative patterns of production.

4. *The Ecological and the Social as Mutually Constitutive but Separate Systems*

A second response to the difficulties of 'ecological functionalism' has been to explicitly disaggregate post-Stewardian general ecology, to speak instead of humans being simultaneously involved in nonconflateable ecological and social relations, and to play down possible causal connections between the two. In this model—most clearly stated in the work of Ingold (1986b)—ecological relations are those represented by material flows between individual organisms, and abstract and culture-bearing social relations are those established between persons through consciousness. Here the social and the ecological are co-extensive; both intra- and interspecific interactions simultaneously have social and ecological aspects. For example, predation is an ecological relationship of extraction with consequences for the flow of energy and materials; hunting—which may refer to exactly the same activities—on the other hand is constituted through social relations of appropriation and exchange. Natural selection operates at the level of ecological rather than social relations, within an effective environment of conscious internal representations and plans for action. In this model, consciousness lies beyond (rather than within) the field of ecological relations, social institutions and relationships never being incorporated within adaptationist or other Darwinist explanations.

One difficulty with this approach is a metaphysic which prevents the effective engagement of the social and ecological and which sidesteps the issue of adaptation. How one might influence the other, and how we might measure and compare different entities or sets of relations remains the major problematic of ecological anthropology, and one rarely, and then inadequately, addressed. There is plenty of empirical evidence to suggest that the social and ecological have measurable effects one upon the other. By emphasizing the continuity of the social landscape, such an approach fails to acknowledge the need in any one analysis to draw boundaries around datasets and thus create 'systems' of interacting parts which provide insights into existential reality. Such a view places renewed emphasis on the essentially social determination of modes of subsistence, acting, at most, as a filter for the expression of an underlying ecological dynamic.

5. *Historical Ecology*

A third recent approach, both in biology and anthropology, has been to reject synchronic ecology in favor of analyses which are historical in orientation, emphasizing contingency and process at the expense of the systemic and functional. At the highest level of abstraction, there has been expanded emphasis on the evolution of social and ecological systems (Ingold 1980), and on the co-evolution of genes, species, and culture (Durham 1991), focusing on positive feedback. Work in historical ecology has been concerned with, for example, historical demography (Netting 1981), the environmental causes and consequences of systemic collapses such as that of the classic Maya (e.g. Atran 1993), and the way particular human groups first make and then manage their environment (Crumley 1994). Some of the most dramatic studies have been with respect to tropical forests (Sponsel et al. 1996), long thought of as a byword for the most pristine of global ecosystems, and consequently the most vulnerable to human impact. Yet other studies have focused on how particular well-known functionalist theories of environment–social organization linkage might better be explained through particular histories (e.g. Kelly 1985). Such developments in anthropology have proceeded very much in parallel with the new emerging environmental history.

6. *Perceiving and Conceptualizing the Environment*

Another major area of growth since the late 1970s has concerned how people conceptualize the natural world, what they know of its objective properties, how this relates to the decisions they make about environmental information, and which of this might be used in pragmatic development contexts (see *Nature, Concepts of: Environmental and Ecological*, *Environmental Cognition, Perception, and Attitudes*). Research in this area has grown exponentially over the last decade, and the shape of the problematic altered so much that it is impossible to do justice to the subject here. There has been a spectacular growth in ethnobotany in particular (see *Ethnobiology*), in studies of indigenous knowledge with a view to countering top-down development models (e.g., Richards 1985) and making development policy more effective, and studies in the cognitive organization and cultural representation of natural kinds, and the concept of nature more generally (Descola and Palsson 1996). Research on the environmental knowledge of local peoples has shown how in looking at, say, the relations between humans and domesticates, cultural selection is important, strengthening the case for reconceptualizing

human evolution as the systemic co-evolution of both interdependent species, and of objects and their cultural images.

The emphasis on cognition, on how different cultures represent and 'construct' their worlds, the questioning of the objective notion of system, and developments in modern biology, have led to the emergence of a critical questioning of what is meant by the terms 'nature' and 'environment' (e.g., Ellen and Fukui 1996). Increasing knowledge of the extent to which even hunter-gatherer populations can modify their surroundings and the growing literature in historical ecology has led to further erosion of the nature–culture divide. A persuasive anthropological critique of the organism–environment distinction is to be found in Ingold's borrowings (e.g. 1986a) from the ecological psychology of Gibson (1979): the concept of environmental niche as a set of affordances, and the idea that organisms fit the world to themselves rather than being 'constructed' by them.

7. *The Environment as a Political Agenda*

The development of anthropological and social science theories of human ecology over the last 30 years has taken place against the backdrop of an emerging political agenda, one which has intruded into ecological theorizing, as in discussions on 'indigenous environmental knowledge' (Milton 1996) (see *Ecology, Political*). Under the guidance of figures such as Garrett Hardin and Kenneth Boulding, economic thinking has been placed in a more biospheric context, and the 'growth model' rejected both in relation to advanced industrial and developing societies. Practical concern for environmental degradation, the profligate use of finite resources, the calculated advantages of 'alternative' technologies, and anxieties about biodiversity conservation have spawned theories of sustainable development (Oldfield and Alcorn 1991, Redclift 1987). It has led to the interrogation of definitions of 'sustainability,' asking at what level systems are sustainable (local, regional, say); whether economic, ecological, cultural, and social sustainability are compatible; whether long-term ecological sustainability is to be preferred to shorter term economic sustainability; whether we want to sustain some aspects of a system at the expense of others (say, biodiversity over sustainable food production); whether we wish to sustain systems which derive their dynamic from human interventions rather than change an environment through conservation, and so on (see *Sustainable Development*). Some writing in this vein is distinctively Utopian; some is concerned with practical matters of implementing specific controls; some seeks to modify the existing world-system though retaining capitalist relations of production; other writing seeks a rapprochement between Marxism and environmentalism; and some seeks to interrogate the very notion of 'environmentalism' (e.g. Milton 1996). A specific but well-articulated debate in this subfield concerns what Hardin had called 'the tragedy of the commons,' the ecological consequences of economic development on the management of collectively owned resources (McCay and Acheson 1987). Specific substantive issues which have contributed to our understanding here include that relating to pastoralism and overgrazing (Homewood and Rodgers 1987) and the extent to which foraging is a sufficient lifestyle to survive in tropical rainforests (Headland 1987).

8. *Merging Paradigms*

Understanding the way in which human ecological relations interact with social consciousness is increasingly seen to require the simultaneous application of two apparently antithetical approaches: one which emphasizes the systemic character of interactions looked at spatially and in the short-term, and one which emphasizes linear change in the long-term. The first is about ecological and social context, the second concerns biological and cultural history (including, but not exclusively, the working out of the key mechanisms of natural selection and cultural adaptation). Human action in the environment is, therefore, inevitably the product of a subtle combining of genetic and cultural process. While these are historically the product of previous action, they go on to contribute to the context in which further action takes place, and neither culture nor environment, neither genes nor social relations, are independent variables.

It should be noted that approaches which confine the environment to relativistic internal conceptions, methodologically undermine the notion of environment as an objective backdrop or set of relations, the basis of ecological analysis hitherto. It is necessary, therefore, to recognize that ecological and social relations have both an existence externally as part of the context, and internally as part of a system of ideas and meaning. There is, if you like, a kind of 'triple duality' (pace Giddens): a theoretical notion of context and process, an empirical recognition of genes and social consciousness, and a methodological distinction between objective environment and its various cultural conceptions. All of these interact recursively to reproduce and to change the systems of which they are part.

Moreover, material relationships between humans (those we describe as economic) are predicated by a notion of value, and value which is not reducible to matter is based on a cultural division of tasks, on exchange between persons rather than individual biological organisms, and which can affect material transactions, including the flows of energy, nutrients, objects, and information between individuals of the same species. Through the increasing dependence on value, culture, and consciousness in human production and consumption, the economic becomes differentiated from the ecological, and as—through

exchange—economic relations become more separated from other status relations in local social formations, so the economy as a separate phenomenal form emerges.

The relationship between *Homo sapiens* and the environment cannot be reduced to essences or imperatives—not environmentalism, biologism, sociologism, nor ecologism. Ecology, as human ecology, cannot ever be an autonomous discipline, except in the limited sense of an understanding of the dynamics of human populations in their ecological context. Ecology is rather a problematic—a discursive practice—which brings into focus the articulation and integration of several fields of knowledge, skills, and disciplines.

See also: Human–Environment Relationships; Nature–Society in Geography

Bibliography

Atran S 1993 Itza Maya tropical agroforestry. *Current Anthropology* **34**: 633–700

Borgerhoff M M 1991 Human behavioural ecology. In: Krebs J R, Davies N B (eds.) *Behavioural Ecology: an Evolutionary Approach*, 3rd edn. Blackwell, Oxford, UK

Crumley C L (ed.) 1994 *Historical Ecology: Cultural Knowledge and Changing Landscapes*, 1st edn. School of American Research Press, Santa Fe, NM

Descola P, Palsson G (eds.) 1996 *Nature and Society: Anthropological Perspectives*. Routledge, London

Durham W H 1991 *Coevolution: Genes, Culture, and Human Diversity*. Stanford University Press, Stanford, CA

Ellen R 1982 *Environment, Subsistence and System: The Ecology of Small-Scale Social Formations*. Cambridge University Press, Cambridge, UK

Ellen R, Fukui K (eds.) 1996 *Redefining Nature: Ecology, Culture and Domestication*. Berg, Oxford, UK

Gibson J J 1979 *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA

Headland T N 1987 The wild yam question: How well could independent hunter-gatherers live in a tropical rainforest ecosystem? *Human Ecology* **15**: 463–91

Homewood K M, Rodgers W A 1987 Pastoralism, conservation and the overgrazing controversy. In: Anderson D, Grove R (eds.) *Conservation in Africa: People, Policies and Practice*. Cambridge University Press, Cambridge, UK

Ingold T 1986a *The Appropriation of Nature: Essays on Human Ecology and Social Relations*, 1st edn. University of Iowa Press, Iowa City

Ingold T 1986b *Evolution and Social Life*. Cambridge University Press, Cambridge, UK

Kelly R C 1985 *The Nuer conquest: The Structure and Development of an Expansionist System*. The University of Michigan Press, Ann Arbor, MI

Lansing J S 1991 *Priests and Programmers: Technologies of Power in the Engineered Landscape of Bali*. Princeton University Press, Princeton, NJ

McCay B J, Acheson J M (eds.) 1987 *The Question of the Commons: The Culture and Ecology of Communal Resources*. University of Arizona Press, Tucson, AZ

Milton K 1996 *Environmentalism and Cultural Theory*. Routledge, London

Moran E F (ed.) 1990 *The Ecosystem Approach in Anthropology: From Concept to Practice*. University of Michigan Press, Ann Arbor, MI

Netting R McC 1981 *Balancing on an Alp: Ecological Change and Continuity in a Swiss Mountain Community*. Cambridge University Press, Cambridge, UK

Oldfield M L, Alcorn J B (eds.) 1991 *Biodiversity: Culture, Conservation and Ecodevelopment*. Westview Press, Boulder, CO

Rappaport R A 1968 *Pigs for the Ancestors: Ritual in the Ecology of a New Guinea People*. Yale University Press, New Haven, CT

Redclift M 1987 *Sustainable Development: Exploring the Contradictions*. Routledge, London

Richards P 1985 *Indigenous Agricultural Revolution: Ecology and Food Production in West Africa*. Allen and Unwin, London

Sponsel L E, Headland T N, Bailey R C (eds.) 1996 *Tropical Deforestation: The Human Dimension*. Columbia University Press, New York

Vayda A P 1983 Progressive contextualism: Methods for research in human ecology. *Human Ecology* **11**: 265–81

Vayda A P, McCay B J 1975 New directions in ecology and ecological anthropology. *Annual Review of Anthropology* **4**: 293–306

Warren D M, Slikkerveer J, Brokensha D (eds.) 1995 *The Cultural Dimension of Development: Indigenous Knowledge Systems*. Intermediate Technology Publications, London

Winterhalder B, Smith E A (eds.) 1981 *Hunter-Gatherer Foraging Strategies: Ethnographic and Archaeological Analyses*. University of Chicago Press, Chicago

R. Ellen

Environment and Common Property Institutions

The twentieth century witnessed many efforts to increase the sustainability of natural resource systems through reforms of property-rights systems. Many of these reforms have not achieved their intended outcomes. Instead, some have generated counterproductive results. Since many reforms of property-rights systems are based on a conventional theory of how property-rights systems affect the incentives of those using natural resource systems, the multiple failures need to be viewed as evidence for an insufficient theoretical understanding of the relationship between diverse forms of property and the incentives and behavior of resource users. The currently accepted theory is based on idealized models of private property and government property. Researchers frequently equate common property with the absence of any property rights.

1. Conceptual Confusion

The debate about the relative merits of government, private, and common property has been clouded by

three confusions. Different meanings are assigned to terms without clarifying how multiple aspects relate to one another. The sources of confusion relate to the differences between (a) common-property and open-access regimes, (b) common-pool resources and common-property regimes, and (c) a resource system and the flow of resource units.

1.1 The Confusion between Common-property and Open-access Regimes

Open-access regimes (*res nullius*)—including the classic cases of the open seas and the atmosphere—have long been considered in legal doctrine as involving no limits on who is authorized to use a resource. If anyone can use a resource, no one has an incentive to conserve their use or to invest in improvements. If such a resource generates highly valued products, then one can expect that the lack of rules regarding authorized use will lead to misuse and overconsumption. Some local grazing areas, inshore fisheries, and forests are effectively open-access resources, but there are many fewer than presumed in contemporary policy literature.

Some open-access regimes lack effective rules defining property rights by default because the resource is not contained within a nation-state or no entity has laid claim successfully to legitimate ownership. Other open-access regimes are the consequence of conscious public policies to guarantee the access of all citizens to the use of a resource within a political jurisdiction. The concept of *jus publicum* applies to their formal status but effectively these resources are open access. A third type of open-access regime results from the ineffective exclusion of nonowners by the entity assigned formal rights of ownership. In many developing countries, the earlier confusion between open-access and common-property regimes has paradoxically led to an increase in the number and extent of local resources that are effectively open access. Common-property regimes controlling access and harvesting from local streams, forests, grazing areas, and inshore fisheries have evolved over long periods of time in all parts of the world, but have rarely been given formal status in the legal codes of newly independent countries.

As concern for the protection of natural resources mounted during the 1960s, many developing countries nationalized all land and water resources that had not yet been recorded as private property. The evolved institutional arrangements that local users had devised to limit entry and use lost their legal standing as a result of the national government claiming ownership. The national governments, however, lacked monetary resources and personnel to monitor the use of these resources effectively. Thus, resources that had been under a *de facto* common-property regime enforced by local users were converted to a *de jure* government-property regime, but reverted to a *de facto* open-access regime. When resources that were previously controlled by local participants have been nationalized, state control has frequently proved to be less effective and efficient than control by those directly affected, if not disastrous in its consequences.

1.2 The Confusion between a Resource System and a Property Regime

To add to the confusion, the term 'common-property resource' frequently is used to describe a type of economic good that is better referred to as a 'common-pool resource.' Traditional examples of common-pool resources include fisheries, water systems, and grazing lands. The Internet is an example of a common-pool resource enabled by modern technology. All common-pool resources share two attributes of importance for economic activities: (a) it is costly to exclude individuals from using the good either through physical barriers or legal instruments and (b) the benefits consumed by one individual subtract from the benefits available to others (Ostrom et al. 1994). Recognizing a class of goods that shares these two attributes enables scholars to identify the core theoretical problems facing all individuals or groups who wish to utilize such resources for an extended period of time. Using 'property' in a term that refers to a type of good reinforces the impression that goods sharing these attributes tend everywhere to share the same property regime, a notion that is empirically incorrect.

Common-pool resources share with public goods the difficulty of developing physical or institutional means of excluding beneficiaries. Unless means are devised to keep nonauthorized appropriators (those who harvest from the resource) from benefiting, the strong temptation to free ride on the efforts of others will lead to a suboptimal investment in improving the resource, monitoring use, and sanctioning rule-breaking behavior. Second, the products or resource units from common-pool resources share with private goods the attribute that one person's consumption subtracts from the quantity available to others. Thus, common-pool resources are subject to problems of congestion, overuse, and potential destruction unless harvesting or use limits are devised and enforced. In addition to *sharing* these two attributes, particular common-pool resources *differ* on many other attributes that affect their economic usefulness including their size, shape, and productivity and the value, timing, and regularity of the resource units produced.

Common-pool resources may be owned by national, regional, or local governments, by communal groups, or by private individuals or corporations, or they may be used as open-access resources by whomever can gain access. Each of the broad types of property regimes has different sets of advantages and disadvantages, but at times may rely upon similar operational rules regarding access and use of a resource. Examples exist of both successful and unsuccessful efforts to govern and manage common-pool resources by gov-

ernments, communal groups, cooperatives, voluntary associations, and private individuals or firms (Bromley et al. 1992). No automatic association of common-pool resources exists with common-property regimes or with any other particular type of property regime.

1.3 The Confusion between the Resource and the Flow of Resource Units

Common-pool resources are composed of resource systems and a flow of resource units or benefits from these systems. The resource system (or alternatively, the stock or the facility) is what generates a flow of resource units or benefits over time. As mentioned above, examples of typical common-pool resource *systems* include lakes, rivers, irrigation systems, groundwater basins, forests, fishery stocks, and grazing areas. Common-pool resource systems may also be *facilities* that are constructed for joint use, such as mainframe computers and the Internet. The resource units or benefits from a common-pool resource include water, timber, medicinal plants, fish, fodder, central processing units, and connection time. Devising property regimes that effectively allow sustainable use of a common-pool resource requires some rules that limit access to the system and other rules that limit the amount, timing, and technology used to withdraw diverse resource units from the resource system. Both types of rules are essential if a resource and its flow are to be managed sustainably.

1.4 Property as Bundles of Rights

A property right is an enforceable authority to undertake particular actions in a specific domain (Commons 1957). Property rights define actions that individuals can take in relation to other individuals regarding some 'thing'. If one individual has a right, someone else has a commensurate duty to observe that right. Schlager and Ostrom (1992) identify five property rights that are most relevant for the use of common-pool resources, including access, withdrawal, management, exclusion, and alienation. These are defined as:

(a) Access: the right to enter a defined physical area and enjoy nonsubtractive benefits (e.g., hike, canoe, sit in the sun);

(b) Withdrawal: the right to obtain resource units or products of a resource system (e.g., catch fish, divert water);

(c) Management: the right to regulate internal use patterns and transform the resource by making improvements;

(d) Exclusion: the right to determine who will have an access right, and how that right may be transferred; and

(e) Alienation: the right to sell or lease management and exclusion rights (Schlager and Ostrom 1992).

Private property is frequently defined as being the equivalent of alienation. Property-rights systems that do not contain the right of alienation are considered by many policy analysts to be ill-defined. Further, they are presumed to lead to inefficiency since property-rights holders cannot trade their interest in an improved resource system for other resources, nor can someone who has a more efficient use of a resource system purchase that system in whole or in part (Demsetz 1967). Consequently, it is assumed that property-rights systems that include the right to alienation will be transferred to their highest valued use.

Instead of focusing on one right, it is more useful to define five classes of property-rights holders as shown in Table 1. In this view, individuals or collectivities may hold well-defined property rights that include or do not include all five of the rights defined above. This approach separates the question of whether a particular right is well defined from the question of the effect of having a particular set of rights. 'Authorized entrants' include most recreational users of national parks who purchase an operational right to enter and enjoy the natural beauty of the park, but do not have a right to harvest forest products. Those who have both entry and withdrawal use-right units are 'authorized users.' The presence or absence of constraints upon the timing, technology used, purpose of use, and quantity of resource units harvested are determined by operational rules devised by those holding the collective-choice rights (or authority) of management and exclusion. The operational rights of entry and use may be finely divided into quite specific 'tenure niches' (Bruce 1999) that vary by season, by use, by technology, and by space. Tenure niches may overlap when one set of users owns the right to harvest fruits from trees, another set of users owns the right to the timber in these trees, and the trees may be located on land owned by still others. Operational rules may allow authorized users to transfer access and withdrawal rights either temporarily through a rental agreement, or permanently when these rights are assigned or sold to others. 'Claimants' possess the operational rights of access and withdrawal plus a collective-choice right of managing a resource that includes decisions concerning the construction and maintenance of facilities and the authority to devise limits on withdrawal rights.

2. Can Proprietors as well as Owners Manage Common-pool Resources?

'Proprietors' hold the same rights as claimants with the addition of the right to determine who may access and harvest from a resource. Most of the property systems that are called 'common-property' regimes involve participants who are proprietors and have four of the above rights, but do not possess the right to

Table 1
Bundles of rights associated with positions

	Owner	Proprietor	Claimant	Authorized user	Authorized entrant
Access	X	X	X	X	X
Withdrawal	X	X	X	X	
Management	X	X	X		
Exclusion	X	X			
Alienation	X				

Source: Ostrom and Schlager (1996, p. 133)

sell their management and exclusion rights even though they most frequently have the right to bequeath it to members of their family (see Berkes 1989, Bromley et al. 1992, McCay and Acheson 1987).

Empirical studies have found that some proprietors have sufficient rights to make decisions that promote long-term investment and harvesting from a resource. Place and Hazell (1993) conducted surveys in Ghana, Kenya, and Rwanda to ascertain if indigenous land-right systems were a constraint on agricultural productivity. They found that land users having the rights of a proprietor as contrasted to an owner in these settings made as efficient investment decisions and were as productive. Other studies conducted in Africa (Migot-Adholla et al. 1991, Bruce and Migot-Adholla 1994) also found little difference in productivity, investment levels, or access to credit as well as in land tenure regimes and intensity of cultivation (Turner et al. 1993). In densely settled regions, however, proprietorship over agricultural land may not be sufficient (Feder et al. 1988, Feder and Feeny 1991). In a series of studies of inshore fisheries, self-organized irrigation systems, forest user groups, and groundwater institutions, proprietors tended to develop strict boundary rules to exclude noncontributors; established authority rules to allocate withdrawal rights; devised methods for monitoring conformance; and used graduated sanctions against those who did not conform to these rules (Agrawal 1994, Blomquist 1992, Schlager 1994, Lam 1998). 'Owners' possess the right of alienation—the right to transfer a good in any way the owner wishes that does not harm the physical attributes or uses of other owners—in addition to the bundle of rights held by a proprietor. An individual, a private corporation, a government, or a communal group may possess full ownership rights to any kind of good including a common-pool resource. The rights of owners, however, are never absolute. Even private owners have responsibilities not to generate particular kinds of harms for others. In conventional theory it is always presumed that owners will manage the resources they own efficiently. This does not mean, however, that owners are more likely than proprietors to manage resources in a sustainable manner. Sustainability requires restricting the use of the flow to the 'sustainable yield,' which may or may not be the efficient level (Clark 1980). Thus, the private owner of a common-pool resource may efficiently 'use up' a resource when the value of the flow is high and the future is discounted at conventional levels.

The world of property rights is, thus, far more complex than simply government, private, and common property. These terms better reflect the status and organization of the holder of a particular right than the bundle of property rights held. All of the above rights can be held by single individuals or by collectivities. Some communal fishing systems grant their members all five of the above rights, including the right of alienation (Miller 1989). Members in these communal fishing systems have full ownership rights. Similarly, farmer-managed irrigation systems in Nepal, the Phillippines, and Spain have established transferable shares to the systems. Access, withdrawal, voting, and maintenance responsibilities are allocated by the amount of shares owned (Martin and Yoder 1983, Siy 1982, Maass and Anderson 1986). On the other hand, some proposals to 'privatize' inshore fisheries through the devise of an Individual Transferable Quota (ITQ) allocate transferable use rights to authorized fishers but do not allocate rights related to the management of the fisheries, the determination of who is a participant, nor the transfer of management and exclusion rights. Thus, proposals to establish ITQ systems, which frequently are referred to as forms of 'privatization', do not involve full ownership.

Further, all three types of owners may fail to manage common-pool resources in a sustainable manner. The possibility that individuals who jointly own a resource system and govern and manage that resource in a sustainable manner has only been seriously recognized the 1980s. Prior to that there were many individual studies that documented individual 'success' cases, but these were treated as relatively unique cases and not representing any more general phenomena. While the number of successful instances of self-organized resource systems now appears to be very large, it is also the case that some communities do not organize themselves to develop indigenous property rights systems. When the resources they jointly use are valuable, the lack of self-organization can, indeed, lead to tragic overuse, if not the destruction of the resource system. Thus, the question of why some groups organize and

others do not organize is key to an understanding of common-property institutions and the environment.

3. *On the Origin of Self-governed Common-pool Resources*

Recent research, thus, challenges the generalizability of the conventional theory that predicts that no one will organize to ensure the long-term sustainability of a jointly used natural resource system. While the conventional theory is generally successful in predicting outcomes in settings where appropriators are alienated from one another or cannot communicate effectively, it does not provide an explanation for settings where appropriators are able to create and sustain agreements to avoid serious problems of overappropriation. Nor does it predict well when government ownership will perform appropriately or how privatization will improve outcomes. A fully articulated, reformulated theory encompassing the conventional theory as a special case does not yet exist. On the other hand, scholars familiar with the results of field research substantially agree on a set of variables that enhance the likelihood of appropriators organizing themselves to avoid the social losses associated with open-access, common-pool resources (McKean 2000, Wade 1994, Ostrom 1990, Baland and Platteau 1996). Considerable consensus exists that the following attributes of resources and of appropriators are conducive to an increased likelihood that self-governing associations will form.

3.1 *Attributes of the Resource*

3.1.1 Feasible improvement. Resource conditions are not at a point of deterioration such that it is useless to organize or so underutilized that little advantage results from organizing.

3.1.2 Indicators. Reliable and valid indicators of the condition of the resource system are frequently available at a relatively low cost.

3.1.3 Predictability. The flow of resource units is relatively predictable.

3.1.4 Spatial extent. The resource system is sufficiently small, given the transportation and communication technology in use, that appropriators can develop accurate knowledge of external boundaries and internal microenvironments.

3.2 *Attributes of the Appropriators*

3.2.1 Salience. Appropriators are dependent on the resource system for a major portion of their livelihood or other important activity.

3.2.2 Common understanding. Appropriators have a shared image of how the resource system operates and how their actions affect each other and the resource system.

3.2.3 Low discount rate. Appropriators use a sufficiently low discount rate in relation to future benefits to be achieved from the resource.

3.2.4 Trust and reciprocity. Appropriators trust one another to keep promises and relate to one another with reciprocity.

3.2.5 Autonomy. Appropriators are able to determine access and harvesting rules without external authorities countermanding them.

3.2.6 Prior organizational experience and local leadership. Appropriators have learned at least minimal skills of organization and leadership through participation in other local associations or learning about ways that neighboring groups have organized.

4. *The Importance of Larger Political Regimes*

It is important to stress that many of these variables are affected strongly by the type of larger political regime in which appropriators are embedded. Larger regimes can facilitate local self-organization by providing accurate information about natural resource systems, providing arenas in which participants can engage in discovery and conflict-resolution processes, and providing mechanisms to back up local monitoring and sanctioning efforts. Perceived benefits of organizing are greater when users have accurate information about the threats facing a resource. The costs of monitoring and sanctioning those who do not conform to rules devised by users are very high, when the authority to make and enforce these rules is not recognized. Thus, the probability of participants adopting more effective rules in macro regimes that facilitate their efforts over time is higher than in regimes that ignore resource problems entirely or, at the other extreme, presume that all decisions about governance and management need to be made by central authorities. If local authorities are not formally

recognized by larger regimes, it is difficult for users to establish an enforceable set of rules. On the other hand, if rules are imposed by outsiders without consulting local participants in their design, local users may engage in a game of 'cops and robbers' with outside authorities.

5. Conclusion

The conventional theory of common-pool resources, which presumed that external authorities were needed to impose new rules on those appropriators trapped into producing excessive externalities on themselves and others, is now considered a special case of a more general theoretical structure. For appropriators to reformulate the institutions they face, they have to expect that the benefits they receive from an institutional change will exceed the immediate and long-term expected costs. When appropriators cannot communicate and have no way of gaining trust through their own efforts or with the help of the macroinstitutional system within which they are embedded, the prediction of the earlier theory is likely to be empirically supported. Ocean fisheries, the stratosphere, and other global commons come closest to the appropriate empirical referents (Ostrom et al. 1999). If appropriators can engage in face-to-face bargaining and have autonomy to change their rules, they may well attempt to organize themselves. Whether they organize depends on attributes of the resource system and the appropriators themselves that affect the benefits to be achieved and the costs of achieving them. Whether their self-governed enterprise succeeds over the long-term depends on whether the institutions they design are consistent with design principles underlying robust, long-living, self-governed systems. The theory of common-pool resources has progressed substantially during the second half of the twentieth century. There are, however, many challenging puzzles still to be solved.

See also: Environment and Development; Environment Regulation: Legal Aspects; Environmental and Resource Management; Environmental Economics; Property Rights; Resource Geography; Sustainable Development

Bibliography

Agrawal A 1994 Rules, rule making, and rule breaking: examining the fit between rule systems and resource use. In: Ostrom E, Gardner R, Walker J M (eds.) *Rules, Games, and Common-Pool Resources*. University of Michigan Press, Ann Arbor, MI

Baland J M, Platteau J P 1996 *Halting Degradation of Natural Resources. Is There a Role for Rural Communities?* Clarendon Press, Oxford, UK

Berkes F (ed.) 1989 *Common Property Resources; Ecology and Community-based Sustainable Development*. Belhaven Press, London

Blomquist W 1992 *Dividing the Waters: Governing Groundwater in Southern California*. ICS Press, San Francisco

Bromley D W, Feeny D, McKean M, Peters P, Gilles J, Oakerson R, Runge C F, Thomson J (eds.) 1992 *Making the Commons Work: Theory, Practice, and Policy*. ICS Press, San Francisco

Bruce J 1999 *Legal Bases for the Management of Forest Resources as Common Property*. Community Forestry Note 14. Food and Agriculture Organization of the United Nations, Rome

Bruce J W, Migot-Adholla S E (eds.) 1994 *Searching for Land Tenure Security in Africa*. Kendall/Hunt, Dubuque, IA

Clark C W 1980 Restricted access to common-property fishery resources: A game theoretic analysis. In: Lin P T (ed.) *Dynamic Optimization and Mathematical Economics*. Plenum Press, New York

Commons J R 1957 *Legal Foundations of Capitalism*. University of Wisconsin Press, Madison, WI

Demsetz H 1967 Toward a theory of property rights. *American Economic Review* **57**: 347–59

Feder G, Feeny D 1991 Land tenure and property rights: Theory and implications for development policy. *World Bank Economic Review* **5**(1): 135–53

Feder G, Onchan T, Chalamwong Y, Hangladoran C 1988 *Land Policies and Farm Productivity in Thailand*. Johns Hopkins University Press, Baltimore, MD

Lam W F 1998 *Governing Irrigation Systems in Nepal: Institutions, Infrastructure, and Collective Action*. ICS Press, Oakland, CA

Maass A, Anderson R L 1986 *... and the Desert Shall Rejoice: Conflict, Growth, and Justice in Arid Environments*. Krieger, Malabar, FL

Martin E, Yoder R 1983 Review of farmer-managed irrigation in Nepal. In: Martin G, Yoder R (eds.) *Water Management in Nepal: Proceedings of the Seminar on Water Management Issues, July 31–August 2*. Ministry of Agriculture, Agricultural Projects Services Centre, and the Agricultural Development Council, Kathmandu, Nepal

McCay B J, Acheson J M (eds.) 1987 *The Question of the Commons: The Culture and Ecology of Communal Resources*. University of Arizona Press, Tucson, AZ

McKean M A 2000 Common property: what is it, what is it good for, and what makes it work? In: Gibson C C, McKean M P, Ostrom E (eds.) *People and Forests: Communities, Institutions, and Governance*. MIT Press, Cambridge, MA

Migot-Adholla S, Hazell P, Blarel B, Place F 1991 Indigenous land rights systems in sub-Saharan Africa: A constraint on productivity? *World Bank Economic Review* **5**(1): 155–75

Miller D 1989 The evolution of Mexico's spiny lobster fishery. In: Berkes F (ed.) *Common Property Resources: Ecology and Community-Based Sustainable Development*. Belhaven, London

Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, New York

Ostrom E, Burger J, Field C B, Norgaard R B, Policansky D 1999 Sustainability – Revisiting the commons: local lessons, global challenges. *Science* **284**(5412) (April 9): 278–82

Ostrom E, Gardner R, Walker J M 1994 *Rules, Games, and Common-Pool Resources*. University of Michigan Press, Ann Arbor, MI

Ostrom E, Schlager E 1996 The formation of property rights. In: Hanna S, Folke C, Mäler K-G (eds.) *Rights to Nature*. Island Press, Washington, DC
Place F, Hazell P 1993 Productivity effects of indigenous land tenure systems in sub-Saharan Africa. *American Journal of Agricultural Economics* **75**: 10–19
Schlager E 1994 Fishers' institutional responses to common-pool resource dilemmas. In: Ostrom E, Gardner R, Walker J M (eds.) *Rules, Games, and Common-Pool Resources*. University of Michigan Press, Ann Arbor, MI
Schlager E, Ostrom E 1992 Property rights regimes and natural resources: A conceptual Analysis. *Land Economics* **68**(3): 249–62
Siy R Jr. Y 1982 *Community Resource Management: Lessons from the Zanjera*. University of the Philippines Press, Quezon City, Philippines
Turner II B L, Hyden G, Kates R (eds.) 1993 *Population Growth and Agricultural Change in Africa*. University Press of Florida, Gainesville, FL
Wade R 1994 *Village Republics: Economic Conditions for Collective Action in South India*. ICS Press, San Francisco

E. Ostrom

Environment and Development

1. Environment and Development

Since the 1950s people have become increasingly aware of problems like marine oil-spills, wildlife extinction, and pollution. However, before the early-1970s few, except ecologists and other environmental scientists, were familiar with the terms 'environment' or 'development.' Today it is generally accepted that for development to be successful it must be environmentally sensitive (Barrow 1995).

2. The Term 'Environment'

'Environment' may be defined as the total external conditions and influences affecting the life and fortunes of organisms, effectively 'the parameters of life.' It is shaped by living and nonliving elements that often interact in complex ways, making it a challenge to understand and predict future situations. The study of these interactions, 'ecology,' was founded as an academic subject in 1869 by Ernst Häckel, became established by the 1930s, and underwent an expansion in the 1970s. Ecology has been described as 'scientific natural history' and, along with 'environment,' became familiar to the general public in the early 1970s. The media often uses the term ecology loosely as a label for environmental concern (for example, an 'Ecology Party was launched in the UK in the early 1970s).

Humans often adapt well to the environment they live in, and may be able to modify it to better suit their needs. Sometimes they are uncomfortable, fail to improve or even sustain conditions, or cause environmental degradation and then have to move away or suffer. A given environment may offer more opportunities than others but, ultimately, there are limits to human activity. If humans can satisfy their needs on a sustained basis without damaging the environment they may be said to have achieved 'sustainable development.' There are often many workable development strategies that could be pursued, some 'fit-in' with nature and at the other extreme others boldly change conditions. Wise development should also assess and prepare for environmental threats (see *Environmental Surprise*; *Environmental Risk and Hazards*). Agriculture is essentially an attempt to modify natural environments by inputs of energy (human labor or fuel-based mechanization), technology, and knowledge to sustain higher outputs of food or other crops.

People often talk of 'their local environment' or an 'alpine' or a 'desert environment,' but there is essentially only one global environment or biosphere. The biosphere can be subdivided into a vast number of more or less interacting 'ecosystems' (each a recognizably distinct assemblage of organisms adapted to particular environmental conditions). The ecosystem is the basic functional unit of ecology (see Miller 1990).

3. The Term 'Development'

Attempts to define 'development' reflect the current values of those making the definition, it is therefore difficult to agree one definition that is universally accepted and precise. Some see it as a learning process, others argue it is 'positive social transformation.' Another definition that is widely accepted in the West is: *an ambiguous term for a multidimensional process involving material, social and organizational change, accelerated economic growth, the reduction of absolute poverty and inequality*. There is some overlap with *modernization*, so that a group of people could develop yet remain unmodernized, or vice versa. Modernization is generally accepted to be a change toward a 'modern' economic, social and political system developed in Western Europe and North America between the seventeenth century and present. Many in the arts, sciences, philosophy, and social argue that there have been signs of a paradigm shift since the mid-1960s, away from modern to *postmodern* development, involving less support for a mechanistic, reductionist and compartmentalized approach to the world's problems, in favor of a more *holistic*, multidisciplinary stance. (A researcher adopting a holistic approach seeks to study the whole, which may be more than the sum of the parts, so exploration of linkages and interactions is as important as studying component parts.) Whether the approach to development is

modern or postmodern in character, it will meet difficulties if it does not make optimum use of environmental opportunities, heed physical and socioeconomic limits, and prepare for possible hazards.

4. Attitudes Toward Environment and Development

Peoples' perception of development affects their exploitation of the environment. It is frequently argued that a faulty attitude toward development leads to environmental misuse and socioeconomic problems for humans (Crump 1991) (see *Environmental Cognition, Perception, and Attitudes*; *Nature, Concepts of: Environmental and Ecological*).

4.1 Attitudes Toward Environment and Development from Prehistoric Times to the Nineteenth Century

Before humans developed a sedentary lifestyle and their numbers increased beyond a few persons per square kilometer, environmental damage was largely temporary and limited. People moved on before pollution or erosion became too bad and, with the probable exception of some game animals, seldom overexploited resources (see *Human–Environment Relationship: Prehistoric*). Although it has always been important to recognize natural threats and opportunities, the need to be concerned about environment and development has mainly grown as people have settled and multiplied, and as industry and technology have caused new impacts. By roughly 3000 BC around the Mediterranean the struggle to survive had led to forest clearance, soil erosion, and loss of fauna through hunting enough for several ancient chroniclers to record environment and development problems. One Roman geographer even recognized a link between the expansion of free trade and exploitative agriculture damaging the land.

Peasants, with few alternatives to fall back upon for survival, have often evolved sustainable livelihood strategies and sound rules of stewardship to ensure common resources were used without being degraded. Other social groups: the very poor, refugees, and the greedy have often exploited and damaged the environment in their struggle to survive or to profit. The poor may be forced to sacrifice sustainable livelihood and future opportunities to survive in the short term unless given aid; those driven by greed or commerce cause havoc through environmentally insensitive ethics (see *Environmental Conservation*: *Ethical Concerns*).

In medieval Europe the prevailing view was that the universe reflected Divine Perfection, consequently there was little attempt to think about development or tolerance for those who did. After the Renaissance some explored the possibility of shaping secular human development; for example, Thomas Moore published *Utopia* in AD 1516. By the mid-sixteenth century Europe's merchant class had grown in influence, possibly assisted by the spread of Calvinistic ethics (Weber 1958). There was a growing, relatively liberal, capitalist economy and an increasing interest in science and rational study. From the mid-eighteenth century this commerce and scientific reason, together with an Industrial Revolution and urban migration had caused huge changes of attitude, a shift to what has been termed the 'modern' outlook.

This modern, western civilization is often accused of particular insensitivity toward the environment. However, there are environment and development problems in virtually all societies: in atheistic, Christian, Islamic, Buddhist and Hindu countries (the latter, some claim, are more 'in-tune' with nature), regardless of whether the economy is socialist or free-enterprise capitalist. The intensity and type of environment and development problems may vary with the impact of commerce, technology and population growth, but they are found everywhere and pose a global challenge.

Beyond the spreading cities, degraded farmlands and forests of Europe and eastern North America several islands began to suffer damage to biota and environment by the eighteenth century. For example, by the 1760s legislation was being enacted to try to protect forests on Tobago, Mauritius, and St. Helena, and many other islands were in need of such controls. Even oceanic fisheries, seal and whale stocks were coming under pressure by the early-nineteenth century (see *Human–Environment Relationship: Carrying Capacity*). Nevertheless, few people could yet see the world's frontiers closing, if there were problems they were sure there was plenty of new land and resources to move on to. There were also naturalists in Europe and the Americas who raised public interest in the natural environment (an early example being Gilbert White's 1798 study of nature; reputedly the fourth-most published English language book) (for writings on the development of environmental attitudes see Wall 1994).

4.2 Attitudes Toward Environment and Development from the Nineteenth Century to the 1940s

By the nineteenth century the dominant attitude adopted toward development in the Western nations was that, whenever possible, the environment should be exploited, and unspoiled areas were wasted or even threatening and should be 'tamed.' As the nineteenth century progressed the idea that humans could ultimately conquer environmental problems became established. Achievements between 1830 and 1930 seemed to confirm such beliefs, as engineering and science made huge progress. But as pollution and overcrowding of industrial cities increased (see Ponting 1991) and as scientists learnt more, it became clear that controlling the environment would not be as

easy as had once been hoped, and that there was a growing need for better stewardship and development ethics.

In Europe, Russia and North America concern about growing cities, pollution and the clearance of forest and wilderness prompted a diverse group of intellectuals, romantics and 'proto-anarchists' to question the way development was progressing (these included Thoreau 1854 and Kropotkin 1899). These intellectuals shared a concern about the moral and ethical issues associated with development and environmental degradation. Some saw the need to value nature for the inspiration and solace it could give which might help improve the human character. While more utilitarian environmentalists argued for a return to closer proximity to nature, and for development to establish decentralized, cooperative communities that would offer an escape from what was being inflicted by large scale industrialization and increasing state control.

The idea of protecting the environment against development existed here and there long before the nineteenth century, around the world people have protected certain plants or animals, forests or areas for practical or religious reasons, establishing sacred groves, hunting parks, holy trees, and taboos which help ensure rational and sustainable use of common resources (see Berkes 1989). By the early- to mid-nineteenth century a growing number of naturalists, hunters, and foresters could see that what had once seemed limitless lands and resources were being degraded so they set up societies to promote conservation, and lobbied for laws to protect wildlife, forest and soil. National parks and reserves began to appear in the USA, Africa, India and other countries, and a number of forestry services were established. In the USA *preservationists* like John Muir worked to establish unspoilt wilderness reserves, and *conservationists* like Gifford Pinchot argued for areas where environmental protection could be combined with compatible land use or resource exploitation (see *Environmentalism: Preservation and Conservation*). On both sides of the Atlantic a number of nongovernmental organizations (NGOs) were established between 1830 and 1930 to promote conservation and wildlife protection (like the Sierra Club in California—founded in 1892); as well as continuing to promote environmental concern some of these bodies played an important role after the Second World War in establishing powerful new environmental NGOs like the International Union for the Conservation of Nature and Natural Resources (IUCN) and Friends of the Earth.

Publication of Darwin's *The Origin of Species by Means of Natural Selection* in 1859 helped established concepts like 'natural selection' and 'survival of the fittest'. These ideas and scientific advances in the nineteenth century led to some rethinking of long accepted ideas about humankind and nature. From the mid-nineteenth century the development of science, particularly concepts of the evolution of species, encouraged some to argue that the human environment–development relationship was strongly influenced or even largely controlled by physical conditions like climate. Race was also often seen as another determinant of development. These *environmental determinists* have been divided into the 'scientific' (e.g., Ellsworth Huntington 1915), who tried to objectively explore how nature might affect human fortunes, and those who promoted 'crude and subjective' arguments. The latter especially were cited to justify racial and political extremism, notably that which helped support fascism in Germany. Similar problems were caused by supporters of *social Darwinism*, who argued for development to be shaped by competition, self-interest and eugenics, rather than mutual aid, co-operation and tolerance of the less developed (Pepper 1984). Although, much discredited by the 1940s, it is probable that these concepts helped discourage social scientists from exploring environment and development issues until well into the 1970s.

Serious drought and soil degradation in the US Midwest 'Dustbowl,' especially between 1932 and 1938, displaced many farmers and darkened the air as far as Chicago and Washington DC. Novelist John Steinbeck drew on the misery caused by the disaster for his novel *The Grapes of Wrath* (1939), as did the folksinger Woody Guthrie; both were at first seen as subversives criticizing the American lifestyle before they provoked public concern (see *Desertification*). The US Government had begun to pass 'New Deal' measures to control the land degradation by the mid-1930s, notably the foundation of the US Soil Erosion Service (in 1933) and its successor, the US Soil Conservation Service (in 1935). Just as the public in North America and Europe had started to see signs that there were environmental limits to development the Second World War diverted attention and resources.

4.3 Attitudes Toward Environment and Development After the 1940s

During the immediate post-War period attention was mainly focused on economic and industrial reconstruction and improving food production. The British, struggling to raise vegetable oil production, invested heavily in the Tanganyika Groundnut Scheme in the late 1940s and early 1950s, largely by failing to consider basic environmental parameters the development efforts failed and wasted large sums of money. This was one of the first environment and development disasters, soon followed by oil-spills, nuclear waste escapes, and other problems. The Cold War diverted attention and funds from environmental studies and development programs, and continued to do so until the 1980s; but by triggering antinuclear weapons

protest movements, it also helped to establish environmental politics in Europe. A few writers explored environment and development issues in the 1940s and 1950s (e.g., the writings of Teilhard de Chardin and Aldo Leopold helped stimulate environmental interest in the 1960s and 1970s), and some NGOs and public figures began to bring conservation and pollution issues to the attention of the public in North America and Europe. In 1949 the UN held one of the first post-War meetings, the Conservation Conference at Lake Success, New York State. In 1948 the IUCN, an NGO body which has acted as a very influential network supporting conservation and promoting interest in environment and development issues was founded, now known as the Union for World Conservation, it continues to be highly important (for a history of the IUCN and conservation, see Holdgate 1999).

By the early 1960s attitudes towards environment and development had begun to change in Western nations, this is probably attributable to peoples' growing affluence and leisure time, development of the media, improved knowledge of the structure and function of nature (in part thanks to the exchange of information and research during the International Geophysical Year of 1957–1958 and the International Biological Programme of 1964–1965), and well-publicized pollution accidents and other development-related environmental problems. It has also been suggested that the US Civil Rights Movement and 'hippie' culture encouraged people to ask questions about development and to value the environment more. Another significant development was that NGOs and groups of concerned individuals began to fight environmental issues in law courts. Interest in environmental matters 'took-off' in California, spreading to the rest of North America and Europe, reaching a peak between about 1966 and 1974. During this period what is variously described as 'environmentalism,' an 'environmentalist movement,' an 'environmental movement,' or an 'ecology movement' spread (the first is probably the best term to adopt). A number of popular books helped provoke public interest in environmentalism, one of the most seminal and earliest being *Silent Spring* (Carson 1962), which voiced concern about DDT pollution.

Today *environmentalism* is generally seen as a moral code or set of mediating values to guide human conduct. Environmentalism is often confused with *ecologism* which is more a political ideology requiring its adherents to seek ethical and moral changes and to move away from giving priority to the welfare of humankind in any evaluation of development options. Environmentalism generally allows followers to put humans before nature and to seek managerial solutions rather than make far-reaching changes of development ethics and morals. Environmentalism encompasses a huge diversity of beliefs and attitudes, some pragmatic and willing to draw-upon science, others more romantic, and some keen to avoid modern science. In general mid-1960s to mid-1970s environmentalism was dogmatic, rich in advocacy rather than workable proposals, and focused on the need for conservation, and the threats of over-population, pollution, and careless technical innovation (Farvar and Milton 1972), it was also largely unpoliticized and restricted mainly to middle-classes and intellectuals. Many environmentalists, especially the *neo-Malthusians* (e.g., Ehrlich 1972), warned of a coming 'crisis.' In the UK a popular environmentalist journal *The Ecologist* was founded in 1970 and public interest was stimulated by authors like Goldsmith et al. (1972) and Ward and Dubos (1972). The feelings of 'crisis' were fanned when the USA and USSR space programs released pictures of the world from space, prompting awareness that war, pollution, and other development problems took place within a finite and vulnerable system—a realization captured at the time by the expression 'Spaceship Earth.' At about this time the Club of Rome published *The Limits to Growth* (Meadows et al. 1972), which argued that in a finite world the established patterns of development were akin to cancer. Amidst bitter debate this helped initiate a rethinking of development ethics and the nature of economics (see *Cassandra/Cornucopian Debate*).

In 1969 James Lovelock proposed the *Gaia hypothesis*, there have been several variants of this, but they all essentially argue that life on Earth has not simply adapted to physical conditions it is part of a complex inter-relationship. The hypothesis is disputed, but if accepted, the implication for development is that humans must carefully 'fit in' and ensure they do not upset any of the delicate mechanisms that maintain a stable environment, failure to do could result in catastrophic ecological change (see *Gaia*).

Warnings that the Earth faces an environment and development crisis have increasingly been issued since the 1960s. The cause is identified as one or a combination of the following: population growth; cavalier use of nature; misapplication of technology; faulty development ethics. Those concerned with environment and development face a dilemma: problems may be expensive, socially damaging, or impossible to solve once they happen, it therefore makes sense to try to monitor crucial thresholds and act to prevent a crisis (a turning-point after which action is less effective or hopeless); yet there are risks in warning of threats and acting before there is irrefutable proof unless a so-called 'win–win' approach can be found which gives benefit whatever the outcome (see *Precautionary Principle*). There are risks that accepting there is a crisis will divert attention from actually solving problems and will promote undesirable 'crisis management,' rather than careful problem avoidance. Recognition of a crisis may be a mistaken over-generalization and lead to failure to uncover causes.

In 1973 the UN Environmental Programme (UNEP) was established to promote environmental care worldwide. Based in Nairobi, rather than Paris or

Washington or some other developed country capital, to help ensure developing countries are not alienated, the UNEP often struggles to gain adequate funding. North America and Europe started to pass legislation in the late 1960s to develop more pro-active planning approaches which aim to reduce or avoid unwanted impacts. One of the most important examples of such legislation was the 1969 US National Environmental Protection Act (NEPA) (passed in 1970), among its achievements was the promotion of *environmental impact assessment* (EIA) procedures, and within ten years it had been copied in many countries around the world.

Nevertheless, in spite of continuing environmentally-damaging accidents, conflict over natural resources (e.g. the UK–Iceland 'Cod Wars' between 1952 and 1976), growing politicization of environmentalism, and the 1972 UN Conference on the Human Environment (held at Stockholm), popular interest waned after 1974. Ironically, the 1973–74 OPEC oil crisis led to fears of energy scarcity which diverted public attention and led to allocation funding to pay for more expensive fuel imports so there was less for environmental care. In many poor countries debts began to mount after 1973 prompting increased exploitation of natural resources to earn foreign exchange which accelerated environmental damage also, as funds became scarce, poverty grew and the poor often resorted to environmentally-damaging livelihood strategies. So, as environmental problems grew there was less money to spend on tackling them, added to this there were political changes in the USA which by the late 1970s saw less government enthusiasm for expenditure on environmental protection.

4.4 Attitudes Toward Environment and Development After the 1970s

At the time of what was effectively the world's first international ecology 'summit,' the 1972 UN Conference on the Human Environment at Stockholm, environmental matters were poorly reported, environmentalism was hardly politicized (outside Germany and a few other countries), few states had ministries or departments for the environment, and there was disinterest, distaste or even fear of calls for limits to growth and to protect nature. A number of representatives, especially those from developing countries, voiced the views that environmental care was a luxury their nations could not afford; that it was a ploy used by richer countries to delay further development and subjugate or exploit them (see *Ecological Imperialism*); or, at the very least, was hypocrisy coming from states which had damaged their own environments to grow rich. When the 1992 UN Conference on Environment and Development met at Rio de Janeiro virtually all countries had ministries of environment; most delegates accepted environmental concern was a vital, integral part of development and not a luxury; the prevailing view was that rich and poor countries were interdependent parts of one global environment; and that practical and possibly painful measures must be taken to solve problems and promote more development in poor countries. In twenty years there had been a 'sea change.'

Certain publications helped achieve this attitude shift, one was *Our Common Future*, the 'Brandt Report' (Independent Commission on International Development Issues 1980), which spread the idea of developed and developing country interdependence eroding the view 'aid is charity' (i.e., the rich would benefit from helping the poor develop and by protecting the world environment). Other influential publications included: the *World Conservation Strategy* (1980) and the 'Brundtland Report' (World Commission on Environment and Development 1987). These publications rekindled public interest in environment and, to a greater extent than before, in development issues; they also spread the concept of *sustainable development*. Sustainable development was more palatable than zero growth called for by supporters of the limits to growth argument; it links aid, environment and development in a concept which, although difficult to give a precise and universal definition for, argues for stewardship to minimize environmental and social damage, efforts to ensure future options are not reduced and, if need be, aid to help achieve these aims (see *Sustainable Development*) (Reid 1995).

During the early 1980s the expression *green* was increasingly applied to environmental issues, largely as a friendly media 'shorthand.' The meaning is vague, but in the main stands for environmentally-friendly and environmentally-sensitive, so 'greening' means 'improving environmental care,' and a 'green,' someone concerned for the environment. 'Greening' has marked the spread of environmental concern to a wider spectrum of Western (later other) societies, than in the 1970s. Interest has spread beyond the middle-classes to the public as a whole, and business, the legal profession, technology, the media, consumer protec tion bodies, marketing and advertising, economics and social studies, theology and philosophy have 'gone green.' This is especially important with respect to economics and commerce, given the crucial role these have in development. Economics has been seeking to develop environmentally-sensitive approaches, an early green economics publication being *Blueprint for a Green Economy* (Pearce et al. 1989). By the late 1980s many large bodies, including the World Bank and most aid agencies, had established environmental departments and green guidelines.

Many interpret this greening as indicating a new social movement: the *green movement*, certainly since the mid-1970s a new 'green' relationship has developed between the social sciences, natural sciences, policy making, and development implementation. Some see

the green movement as a 'cultural attack' on the ills of modern society and economics (Adams 1990, p. 71), and it has been argued it is a manifestation of a postmodern and postindustrial cultural shift. The green movement is very diverse, but most would accept it embraces environmental sensitivity, social responsibility, grass-roots democracy, and nonviolence.

Although many greens scorn politics, or support whatever party they feel will give the best protection for the environment, an aspect of greening has been the spread of environmental politics. Green politics first originated in New Zealand and West Germany with roots in anti-nuclear weapons protest, feminism, and the US Civil Rights Movement, early romantic environmentalists (like Henry Thoreau), and more recent writers on *social ecology*, environmental ethics and *deep ecology* (such as Arne Naess) (see *Ecology, Deep*). Starting in Hamburg during the early 1970s, green politics established a newspaper—*Die Grünen*—and candidates were elected to German State legislatures, and a few years later, to the national Parliament. Early German activists include Petra Kelly, Herbert Gruhl, and Rudolph Bahro. Only a scattering of greens have been elected to governments in Europe and other countries (virtually none in the USA), and if anything support has waned wince the mid-1980s, however they triggered a greening (somewhat cynical) of established political parties. At the close of the twentieth century there can be scarcely a single country in the world without green NGOs (increasingly networking globally), sympathetic newspaper, radio and television editors, politicians interested in cultivating an environmentally-sensitive image, and growing commercial interest in environment.

It has become increasingly apparent that various 'development pressures' are damaging, or have destroyed, established livelihood strategies, leading to environmental damage and human misery. For example, the spread of sedentary agriculture has disrupted seminomadic herders; economic policies may trigger shifting cultivation leading to land degradation; population growth or land development policies may promote settlement of unsuitable and vulnerable soils. As traditional ways breakdown new ones must be found which improve and sustain human welfare without environmental damage. Development therefore faces more challenges than ever before.

Governments, business, and NGOs increasingly adopt *environmental management* measures to try to ensure the *polluter pays* (i.e., whoever is responsible for environmental damage meets the cost of remedy) and minimize unwanted impacts, using precautionary techniques like environmental and social impact assessment, environmental accounting, and eco-auditing. Foreign aid is nowadays almost always conditional on the application of such measures, and is increasingly focused on reducing or avoiding environment and development problems. Since the late 1980s concern about *transboundary environmental problems* has increased, these include acid deposition; stratospheric ozone depletion; chemical pollution; loss of biodiversity; and global warming. There has been progress in understanding the structure and function of the environment and in techniques for measuring and analyzing data—for example, satellite remote sensing and geographical information systems (GIS) (for handling constantly updated data). At the Rio 'Earth Summit' in 1992, participating nations signed a Declaration on Environment and Development; Conventions on Biodiversity (protection) and Climate Change (control), and endorsed an environment and development action program: *Agenda 21* (Keating 1993). Nations have started to negotiate and ratify solutions for environment and development problems, but often back-off when it comes to actually spending money or foregoing possible profits (see *International and Transboundary Accords, Environmental*). Governments and individuals are becoming aware that consumption and technological advances can mean environmental damage and move toward better monitoring and enforcement of controls for environmental problems.

See also: Environment and Anthropology; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Economics; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Psychology: Overview; Green Parties; Greening of Technology and Ecotechnology; Human–Environment Relationship: Carrying Capacity; Planning Issues and Sustainable Development; Precautionary Principle; Sustainable Development

Bibliography

Adams W M 1990 *Green Development: Environment and Sustainability in the Third World*. Routledge, London

Barrow C J 1995 *Developing the Environment: Problems and Management*. Longman, Scientific & Technical, Essex, UK

Berkes F (ed.) 1989 *Common Property Resources: Ecology and Community-based Sustainable Development*. Belhoven Press, London

Carson R 1962 *Silent Spring*. Houghton Mifflin, Boston

Crump A 1991 *Dictionary of Environment and Development: People, Places, Ideas and Organizations*. Earthscan, London

Ehrlich P R 1972 *The Population Bomb*. Simon and Schuster, New York

Farvar M T, Milton J P (eds.) 1972 *The Careless Technology*. Garden City Press, New York

Goldsmith E, Allan R, Davol M, Lawrence S 1972 *Blueprint for Survival*. Penguin, Harmondsworth, UK

Holdgate M 1999 *The Green Web: A Union for World Conservation*. Earthscan, London

Huntington E 1915 *Civilization and Climate*. Yale University Press, New Haven, CT

Independent Commission on International Development Issues 1980 *North–South: A Programme for Survival*. Pan, London

International Union for Conservation of Nature and Natural Resources, UNEP, WWF 1980 *World Conservation Strategy: Living Resources for Sustainable Development*. IUCN, Gland, Switzerland
Keating M 1993 *The Earth Summit's Agenda for Change: A Plain Language Version of 'Agenda 21'*. Earthscan, London
Kropotkin P 1899 *Fields Factories and Workshops* (original in Russian; English edn. 1974 ed. by C. Ward, Hutchinson, London)
Meadows D H, Meadows D L, Randers J, Behrens III W W 1972 *The Limits to Growth (a Report for the Club of Rome's Project on the Predicament of Mankind)*. Universe Books, New York
Miller G T Jr 1990 *Living in the Environment: An Introduction to Environmental Science*, 6th edn. Wadsworth, Belmont, CA
Pearce D, Markandya A, Bastier E 1989 *Blueprint for a Green Economy*. Earthscan, London
Pepper D 1984 *The Roots of Modern Environmentalism*. Croom Helm, London
Ponting C 1991 *A Green History of the World*. Sinclair-Stevenson, London
Reid D 1995 *Sustainable Development: An Introductory Guide*. Earthscan, London
Thoreau H D 1854 *Walden; Or, Life in the Woods*. Ticknor and Fields, Boston (later published as *Walden*. Penguin, New York, 1983)
Wall D 1994 *Green History: A Reader in Environmental Literature, Philosophy and Politics*. Routledge, London
Ward B, Dubos R 1972 *Only One Earth: The Care and Maintenance of a Small Planet*. Penguin, Harmondsworth, UK
Weber M 1958 *The Protestant Ethic and the Spirit of Capitalism* (English translation by T. Parsons—originally published in 1904 and 1905). Charles Scribner, New York
White G 1977 [1798] *A Natural History of Selborne*. Penguin, Harmondsworth, UK
World Commission on Environment and Development 1987 *Our Common Future* (the Brundtland Report). Oxford University Press, Oxford, UK

C. J. Barrow

Environment Regulation: Legal Aspects

1. Definition of ER

ER is the guiding by law of human activities with a view to preventing them from damaging the natural environment.

'Regulation' means the setting of specific rules for specific purposes mostly involving administrative agencies in the implementation of the rule. Their form can be laws or sublegal rules. This distinguishes regulation from more fundamental constitutional norms, but also from basic rules of civil law. With the opening up of civil law for more specific political purposes such as consumer and environmental protection, and, more generally, with the growing understanding of any law as an instrument of organizing society, civil law instruments can also be regarded as regulation. Regulation is distinct from decision-making in individual cases, in that the former consists of rules addressing many cases and takes the form of an ordinance or otherwise, while the latter typically results in administrative acts and contracts. In a broad sense, however, regulation is understood to mean the whole setting of state intervention into the market by means of the introduction of goal-oriented law, the setting up of related administrative agencies, and the application of the law in individual cases.

In many states a pattern of ER has evolved consisting of (a) laws protecting specific media, such as flora and fauna, the atmosphere, water, and the soil against various kinds of dangerous activities (media-related laws); (b) laws concentrating on specific dangerous activities and directing them to respect the various media mentioned above, such as the marketing and use of toxic substances, the disposal of waste, the application of risky technologies (such as genetic engineering, nuclear power), etc. (activity-related laws); and (c) laws on instruments applicable for various activities and media, such as land-use planning, environmental impact assessment, access to environmental information, environmental liability, etc. (horizontal or instrument-related laws). Laws organizing the utilization of natural resources, such as laws on land-use planning, minerals exploitation, fishery, agriculture, and forestry, can also be regarded as ER to the extent that they are concerned with environmental issues. A recent trend in many countries is to assemble ER in comprehensive codes. France, the Netherlands, and Sweden have successfully completed such an exercise.

2. Legal Perceptions of the Environment

The type of ER a society chooses depends on how it perceives the environment (Winter 2000). The relationship between nature and the law varies both with the historical development of societies and with the current geographic situation in any given society. In subsistent economies nature is regarded as a powerful 'fate' or agent to which people must defer. The law (e.g., guild rules) impedes individual initiative and thus indirectly prevents overexploitation of the environment. In the industrializing world nature is seen as a resource which can be exploited in order to raise living standards. The law is preoccupied mainly with releasing and assisting individualistic economic energy and inventiveness. In a way it is also environmental law, but in the sense not of protecting but of utilizing nature. In wealthy societies nature is thought of as the environment of the individual, which must be protected in order to secure healthy physical conditions. ER is mainly a child of this kind of society. In today's globalized world a new common perception has not yet emerged. The developing countries still picture

nature as an exploitable resource whilst the developed countries, carrying the environment concept to the global level, would like to have the third world's nature preserved as their carbon sink, gene bank, etc.— in other words, as *their* environment. Ongoing international discourses tend to foster the vision of nature as a global *biosphere*. The interacting physical and biological processes of this biosphere have developed mechanisms of negative loops which can absorb many disturbing events. Humanity is, however, about to instigate more and more positive loops which may endanger the livability of the earth. Therefore, today's ER has, by content as well as level, increasingly become globalized itself.

3. *Difficulties and Responses*

One major difficulty ER must cope with is *uncertainty*. Often it is scientifically unresolved whether suspected environmental damage does or will exist and, if so, whether it was or will be caused by a certain human undertaking. For instance, for a long time it was unclear if acid rain damage to forests was in fact caused by sulfur dioxide emissions. Today it is uncertain whether the global warming caused by 'greenhouse gases' will persist or be reversed by counter-effects, such as cooling effects brought about by the vegetation growth promoted by increased carbon dioxide emissions. In former times when ER was a matter of police power (i.e., restraint, governmental control of society) a determined and immediate *danger* of damage was required for ER to come into effect. Often, the damage did materialize and ER could only look for repair and compensation.

Today in many countries ER follows the *precaution principle*, which requires that preventive safeguards shall be applied even in situations where there is a *risk*, but no certainty, of damage. Also, a primary concession for a risky undertaking may more easily be modified or revoked if better knowledge reveals formerly unknown risks.

Another difficulty to cope with is that ER often involves economic loss, e.g., when an emission standard requires the installation of costly technology. From a long-term perspective environmental benefit will often prove to be economically rewarding (new markets for integrated technology and products, avoidance of disasters, etc.). But individual actors facing regulation normally have to incur adaptation costs. Legal cultures differ as to how they take such costs into consideration. The Anglo-American legal culture is more prepared to balance ecological gain against economic costs. ER is then inappropriate, if the costs outweigh the benefits (cost–benefit analysis, CBA). By contrast, continental European culture somewhat one-sidedly looks at the ecological risks, thereby revealing a basic skepticism against unrestricted relativization of environmental goods. Best available techniques (BAT) are demanded notwithstanding the cost, except if they are obviously extreme. Both of these approaches meet with difficulties: CBA, by monetarizing intangible ecological goods, may distort a fair balance, whereas BAT, by suppressing cost considerations, may cause them to reenter on the political and implementation level where costs tend to be dramatized. The solution is probably to be found somewhere in between over- and underrationalization: depart from BAT, require to quantify excessive cost–benefit ratios, but allow this only to the extent that the related goods have a market value.

A third difficulty is the frequent divergence between scientific and everyday life experience. As demonstrated by the controversy over genetic engineering, the scientific community and the general public often mistrust each other's perception of risks. This tension influences ER on all levels including legislation, sublegal standard-setting, and decision-making in individual cases. Legal cultures differ as to whether they are more scientifically or politically oriented in this respect. For instance, sublegal rule-making for environmental standard-setting in the USA is wide open for public participation, whilst the EC tends to rely on committees of experts (the so-called 'comitology'). More systematic approaches suggest an iterative procedure: first an open scoping of the valuation framework and necessary research; second, the elaboration of scientifically grounded draft standards; third, exposure of the drafts to public scrutiny; fourth, possibly a loop back to the scientific body; fifth, the taking of the decision by the responsible body (SRU 1996). Individual licensing of dangerous installations, on the other hand, is open to public participation both in the USA and the EC member states. In relation to large industrial or infrastructure projects, public inquiries have often been conducted in a hostile atmosphere far from the ideal of rational deliberation some scholars propose. In North America, mediation techniques have been used frequently, and successfully, to resolve such hostilities. By contrast, in Europe mediation is much less frequent and has seldom been successful. This difference is not easy to explain. Perhaps in Europe decision-makers are less used to open discussion, and perhaps the public is also less inclined to strike deals on environmental issues.

4. *Regulatory and Economic Instruments*

ER has frequently employed what are known as 'regulatory' or 'command and control' instruments, i.e., precise standards (environmental quality goals, emission limits, product quality standards, etc.), licencing requirements, monitoring obligations both of the industry and the administration, powers to order adjustment measures, etc. Economists have criticized this kind of instrument as inefficient. Instead they propose so-called 'economic' instruments, es-

pecially pollution charges and tradable pollution rights. They say that regulatory instruments disregard the different pollution abatement costs of the parties concerned, and thus require a great deal of money to be spent on costly technology with marginal protection effect. Economic tools are more efficient as they allow the entrepreneur to decide whether to pay and pollute or to invest and not to pay. A closer look at the enforcement level in the command and control framework, however, reveals that administrative agencies proceed with some flexibility when requiring abatement investments. Also, strict and equally applicable standards provide the enforcement agency with a most valuable bargaining position (Gawel and Lübbe-Wolff 2000, p. 65). Another criticism is that regulatory tools need more administrative personnel than other instruments. Again, a closer look shows that economic instruments must also be monitored closely in order to work effectively. Inversely, economic instruments do not work when it comes to preventing serious accidents or abating 'hot spots' of pollution. The parties concerned adopt a trading approach, in which they expect something in return for any concessions. This atmosphere may cause all concerned to lose sight of the material problem. This is why many countries have been reluctant to move from command and control to economic measures when it comes to ER.

However, under specifiable conditions economic instruments may indeed be more effective, or at least more efficient, than command and control. For instance, the Californian Acid Rain Program fixed the maximally tolerated emission quantity of sulfur dioxide, allocated pollution rights, and allowed intra-firm and, later, inter-firm trading of these rights. The scheme was deemed successful (Sorrell and Skea 1999, p. 43). Whether it was more effective than a regulatory option would have been is impossible to say, but it was probably more efficient. In more abstract terms emissions trading may work when the target can be agreed upon, the polluting substances are singular, and the number of polluters limited. As to polluter charges, the German 'ecotax' on energy consumption can serve as an example. The tax must be borne by every consumer, but it is collected by the energy utilities. The aim of reducing energy consumption could hardly be reached by command and control measures. This shows in more abstract terms that the charge does work if an unlimited number of persons need to be addressed, and if there is no 'hot spot' situation.

5. *Regulatory Measures and Self-control*

Besides reacting to the heteronomy innate in ER, the actor may also protect the environment by *autonomous self-control*. Autonomy can be founded on the moral beliefs of individual actors or organizations, as well as on social norms developed out of social interaction or association. Professional organizations and standardization bodies have learned to consider environmental effects when designing product or production standards. Many large companies have introduced self-binding rules (e.g., the Responsible Care declarations of the European chemical industry) which are intended to guide their personnel and at the same time inform the public. Many have also established environmental management apparatuses and schemes. In addition, the market mechanism has shown some capacity to make an autonomous push towards environmental protection. For instance, companies may develop environmentally adapted products in anticipation of future consumer preferences. The line between the market and the political sphere is blurred when consumers participate in boycott action, such as that taken against the Shell oil company in the Brent Spar case, where Shell had planned to dump an old oil platform in the North Sea but was prevented from doing so by public action.

These various phenomena of autonomy of the individual, company, association, or market mechanism are explained in various ways. One strand of theory suggests some innate force of self-organization or autopoiesis of social systems (Teubner 1996). Another points to the fact that self-regulation which is effective and not mere paperwork is normally incited, structured, and supervised by heteronomous ER. Indeed, the internal environmental management schemes in place in many companies were instigated (but not imposed) and framed by voluntary auditing laws such as the Auditing Regulation of the EC. Voluntary commitments incurred by a company or group, or association of companies, are often triggered and guided by the threat that otherwise the government might issue a binding regulation or order. Sometimes they are framed as a contract providing for the monitoring of effects and for sanctions in case of noncompliance (for a legal framework for voluntary covenants see the draft German Environmental Code (articles 35–39) Ministry for the Environment 1998). Therefore, in a practical perspective, self-regulation as an instrumental solution must always be tuned with a binding framework of ER.

6. *Legal Protection*

Regulation normally constitutes a two-tiered relationship between the addressees and the government. In rule of law systems judicial review of this relationship is provided. Depending on the system, legal recourse may be granted against regulation in the form of individual administrative acts, as well as sublegal acts and even laws. In ER there are addressees who by implication have the right of appeal, and the situation with this kind of regulation is further complicated because it is characterized by a three-tiered relationship involving third parties (neighbors, etc.). These also ask for legal protection against environmental risk. Civil law remedies directed at the polluter, such as

injunctive reliefs and tort or strict liability, come alongside public law remedies directed at the administrative organs, such as the action for quashing an administrative act (e.g., a construction authorization for a dangerous installation), or for issuing an administrative act (e.g., an order to prohibit the marketing of a toxic substance).

These remedies are, however, confronted with the problem typical of ER that often the environmental damage reaches beyond the individual sphere of the affected third party (consider the destruction of a biotope on propertied land implicating the extinction of a species), or that the third party is only indirectly or morally affected (contamination by pesticides of groundwater which may be used for drinking water purposes, destruction of a forest resort frequented by birdwatchers) (Prieur 1996, p. 842). Legal systems differ in the extent they open their legal remedies up to those 'trans-individual' interests. This may be done by allowing other agents than the individual to defend the related collective interest, and by extending the scope of legal arguments to collective concerns. In the realm of liability Brazilian law, for instance, allows for an action called 'civil-public' which can be filed by the public prosecutor as well as by environmental associations for restoration of or compensation for ecological damage (Leme Machado 1998, p. 290). As to third party appeal against the administrative authorization of dangerous activities, French and English law, for instance, only require the plaintiff to show to be affected in a factual interest (Winter 1996, p. 129), whereas according to the narrower German concept the plaintiff must show that a subjective right or at least an interest individually protected by law was violated (Kloepfer 1998, p. 510). US law takes a middle position by requiring that (a) the plaintiff's factual interest is affected and (b) this interest lies within the zone of interests protected or regulated by the relevant law (Findley and Farber 1999, p. 78). The overall European trend points in the same direction.

Legal systems differ also with regard to the density of judicial review of administrative action. Some reduce court review to errors in law (USA, UK, France) whilst others extend it to errors in fact (Germany). Case law has, however, more and more blurred this line. The overall practice has become to reduce or extend court checking, depending on whether the previous administrative proceeding was court-like or not, if the matter involved complex technical, prognostic or evaluative questions or not, etc. (Findley and Farber 1999, p. 120, Kloepfer 1998, p. 532).

7. *Transnationalization of ER*

ER has been transnationalized step by step. There are three main ways in which this has happened. The first may be called the 'market trajectory.' It is put in force by national regulation of product quality (e.g., threshold values for exhaust from cars or for pesticide residues in food). This constitutes an import restriction for foreign products and at the same time a competitive disadvantage for the internal export industry. This again creates an incentive for both the ER and the foreign state to strive for the harmonization of product standards. As a first step a regime (i.e., rules and institutions) is created which checks whether the distortion of trade implied in the national ER can be justified by the public interest in the envisaged environmental protection measure (so-called 'negative integration'). An example for this is Articles 28–30, Treaty of the European Community (ECT) in combination with judicial review under Article 226 ECT of national ER by the European Court of Justice (ECJ). As a second step the regime issues product standards of its own ('positive integration'). The competence basis for this in the EC is Articles 95, 175 ECT (Krämer 2000).

The second development pattern may be termed a 'relocation trajectory.' It sets off with national regulation of production processes (e.g., emission standards for installations). Such regulation will already have been undermined, in its implementation or even in its design phase, because the industry concerned, envisaging cost differentials, threatens to move to low-standard countries. This leads the ER state to push for an international regime. As the avoidance of environmental investment is not regarded as 'commercial dumping,' those regimes normally practice 'positive integration.' Depending on the sophistication of the regime the joint criteria are framed only in broad language or in the form of precise standards. In the EC the pertinent competence basis is Article 175, ECT.

The third pattern, which may be termed a 'commons trajectory,' starts with situations where adequate ER is absent. This leads to an overexploitation of environmental resources (e.g., destruction of tropical forests), or overutilization of the environment as a medium for residues (e.g., destruction of the ozone layer, climate change). Not only the inactive home country but also foreign countries suffer from such environmental damage because it transcends state borders. This triggers their interest in creating international regimes for the protection of transnational environmental 'goods.' Some of these goods (the high seas outside sovereign rights, the ozone layer, the climate) are legally understood as international common heritage, others (tropical forests, genetic resources) as sovereign rights of the host states but of common concern.

As to the organizational capacity of the international environmental regimes (Young 1994, p. 140), the basic pattern is the founding agreement, a secretariat and the conference of the contracting parties (COP), whose conclusions are not directly binding but need to be ratified. Sometimes technical committees are set up

which accelerate the production of strategies, standards, and implementation measures. More advanced regimes have introduced simplified procedures of adopting binding conclusions (e.g., the opting out procedure). In the long run the international regimes, because of their diversity and number (there are about 250 international environmental protection agreements—some of which form a joint regime), need to be restructured in order to enlarge membership, remove overlapping competences, and improve decision-making and implementation effectiveness. This is even more urgent in order to build worldwide 'positive integration' capacity alongside the worldwide 'negative integration' established by international regimes such as the WTO (Esty 1994).

See also: Environment and Development; Environmental and Resource Management; Environmental Cognition, Perception, and Attitudes; Environmental Economics; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; Environmentalism, Politics of; International and Transboundary Accords, Environmental; Law and Risk; Nature, Concepts of: Environmental and Ecological; Regulation and Administration; Regulation Theory in Geography; Risk, Sociology and Politics of

Bibliography

Bell S 1997 *Environmental Law*, 4th edn. Blackstone, London

Esty D C 1994 *Greening the GATT*. Institute for International Economics, Washington, DC

Findley R W, Farber D A 1999 *Cases and Materials on Environmental Law*, 5th edn. West Group, St. Paul, MN

Gawel E, Lübbe-Wolff G 2000 *Effizientes Umweltordnungsrecht*. Nomos, Baden-Baden, Germany

Kloepfer M 1998 *Umweltrecht*, 2nd edn. Beck, Munich

Krämer L 2000 EC *Environmental Law*, 4th edn. Sweet & Maxwell, London

Leme Machado P A 1998 *Direito ambiental brasileiro*, 7th edn. Malheiros Editores, Sao Paulo, Brazil

Ministry for the Environment (Germany) 1998 Environmental Code. Draft. Duncker & Humblot, Berlin

Prieur M 1996 *Droit de l'environnement*, 3rd edn. Dalloz, Paris

Sands P H *Principles of International Environmental Law*. Manchester University Press, Manchester, UK

SRU (Rat von Sachverständigen für Umweltfragen) 1996 *Umweltgutachten 1996*. Metzler-Poeschel, Stuttgart, Germany

Sorrell S, Skea J (eds.) 1999 *Pollution for Sale*. E. Elgar, Cheltenham, UK

Teubner G (ed.) 1996 *Global Law Without a State*. Dartmouth, Aldershot, UK

Winter G (ed.) 1996 *European Environmental Law: A Comparative Perspective*. Dartmouth, Aldershot, UK

Winter G 2000 *Umwelt–Ressource–Biosphäre. Ansichten von Natur im Recht*. Manholt, Bremen, Germany

Young O R 1994 *International Governance*. Cornell University Press, Ithaca, NY

G. Winter

Environmental Adaptation and Adjustments

The nature of the relationship between people and their physical and built environments is constantly changing. The societal choices regarding natural resource use, alterations in the physical world, or responses to environmental threats vary widely. The timing and scale of these choices affect people and places differently, are time dependent (affect current and future generations), and leave a very distinct geographical imprint. Within the social science literature, these responses are categorized as human environmental adjustments and adaptation.

1. Adjustments to Environmental Threats

Humans continually make decisions in response to the physical environment. These decisions are labeled *adjustments*. Some adjustments are purposeful, but many others are incidental and undertaken unwittingly. The term adjustment is used to describe intentional actions taken by individuals or society to cope with risk or the uncertainty of environmental threats. The concept was initially used in the context of coping with extreme natural events (e.g., floods, earthquakes), but has since been expanded to include a range of nature–society interactions. Understanding human adjustments to environmental extremes has been a fruitful research path in the social sciences since the mid-twentieth century (Cutter 1994, Mileti 1999).

It is the interaction of environmental conditions and human systems that gives rise to environmental hazards. The range and choice of adjustments to hazards are influenced by elements in both the human and physical systems. Some adjustments are too costly to implement, while others may not be successful and ultimately lead to greater costs in the future. While benefits may be derived from some adjustments, they also come at a cost, or are not shared equally among all potential beneficiaries. While there could be hundreds or thousands of potential individual or collective adjustments to environmental threats, normally they are classified into three major categories: (a) modifying the causes of environmental threats, (b) reducing human vulnerability, and (c) distributing losses.

1.1 Modifying the Causes of Environmental Threats

There are many examples of successful and less successful attempts at modifying the cause of environmental threats. For example, in the 1970s, considerable research was undertaken to explore the

range of modifications that could reduce the physical impact of natural hazards (White and Haas 1975). Cloud seeding to reduce the severity of severe storms, underground explosions to relieve fault stress, and damming of rivers to prevent catastrophic flooding are notable examples. As the range of environmental hazards broadened in later decades, so did examples of modifying the causes of these new threats: pollution reduction from its source (point and nonpoint); governmental restrictions in the manufacturing and use of known cancer-causing chemicals (DDT); and federally mandated reductions in chemicals that cause human cognitive dysfunction (lead) or endocrine disruption (dioxins, furans, and PCBs) (Colborn et al. 1996). Within the global change literature, the term *abatement* has been used to describe the range of actions designed to reduce the source or cause of that environmental threat, but the meanings are similar. Greenhouse gas reductions (through fuel switching or conservation) are examples of modifying the cause of the global warming hazard.

1.2 Reducing Human Vulnerability

Broadly, vulnerability is the potential for loss, yet within the environmental hazards literature there are conflicting characterizations of the term as well as methodologies for measurement (Cutter 1996). Vulnerability variously refers to personal or individual potential for loss, societal vulnerability to hazard events and losses, or the resilience of society to environmental threats (Dow 1992). Biophysical vulnerability refers to elements in the physical world that contribute to the loss potential, such as climate perturbations (increased drought, sea-level rise), human modification of ecosystems (deforestation, desertification), or location (relocation) of industrial activity (such as toxic facilities, coal mining, or other extractive industries). Social vulnerability examines the characteristics of various social groups (wealth, gender, race/ethnicity, age) that either constrain or enhance human ability to respond adequately to or cope with environmental threats. Both social and biophysical vulnerability have explicit spatial outcomes that can be mapped and thus compared from local to global scales.

Vulnerability is increasingly used in the hazards and global change communities to describe the differential impact of environmental changes on people and the places where they live (Blaikie et al. 1994, Bohle et al. 1994, Hewitt 1997, Comfort et al. 1999, Oliver-Smith and Hoffman 1999). Vulnerability science is still in its infancy and has not progressed far enough for the development of a universal set of specific indicators as of yet.

Purposeful, short-term adjustments designed to modify vulnerability are often referred to as *mitigation*. Mitigation includes land-use planning (Burby 1998), the reduction of losses through building codes, or improved warning systems to reduce human losses. More often than not, mitigation actions are prompted by a specific hazard event. However, it is the incidental and incremental adjustments taken prior to a hazard event that often attain greater importance in reducing overall losses. The elevation of housing structures so that the ground floor is above base flood levels (flood hazard), the construction of 'safe rooms' in houses to reduce the human losses from tornadoes, fish advisories to reduce human consumption of fish caught in polluted water to decrease bioaccumulation of toxins in human fatty tissue, and carbon sequestration as a means for mitigating greenhouse gas emissions are examples of mitigation actions to reduce human vulnerability.

1.3 Sharing and Distributing Losses

Insurance, relief, and recovery operations are all examples of individual or collective loss sharing. When a disaster strikes, governmental agencies and nongovernmental organizations are mobilized to provide assistance in the form of temporary shelter, food and water, and power. The costs of such operations are normally not borne by the individual disaster victims, but rather by society at large. While the impacts are local and may outstrip the ability of a community to respond, the economic costs are more widely shared. Relief organizations (Red Cross and Red Crescent Societies), faith communities, and governmental agencies (US Federal Emergency Management Agency, UN Disaster Relief Organization) all assist in recovery and the rebuilding of devastated communities.

Insurance is another example of loss sharing, but not everyone has equal access to it. In the United States, the public looks toward insurance as the mechanism for loss compensation (Kunreuther and Roth 1998). Insurance as an adjustment will become less viable in the future as the costs of disasters increase and the realization that neither the insurance nor the reinsurance industry have sufficient resources to pay for a worst-case event. Other examples of loss sharing are financial incentives and disincentives (taxation) which encourage or discourage risky adjustments.

2. Adaptation

Adaptation describes longer-term activities that take into account the threat of environmental extremes or changes. Adaptations often take place unknowingly, simply reflecting the modification of environmental conditions (Turner et al. 1990) or the changing occupancy of threatened regions (Kasperson et al.

1995). Adaptation is rooted in social and cultural systems. Biological adjustments are also considered adaptive such as the assimilative capacity of biogeochemical cycles to absorb pollutants without immediate systems degradation. As with other aspects of culture, these adaptations, while appropriate in some circumstances, are not always the best solution to responses to environmental threats. Often, adaptation to one hazard or environmental threat prompts greater vulnerability to others.

3. Factors Influencing the Choice of Adjustments

Under ideal circumstances, individuals and society would have a full set of choices that they could select from in adjusting or adapting to environmental threats. How people choose adjustments is not always clear-cut, simple, rational, or even economic (Burton et al. 1993). There are a number of factors that constrain the set of adjustments and influence the choices that are made and these are described below.

3.1 Local Control and Local Focus

Local control and focus facilitate the adoption of adjustments as communities and individuals see a direct benefit from their hazard reduction efforts. When a homeowner decides to take action to reduce the risk of storm damage to the home, or when a city council votes to remove flood-prone land from development, both are engaging in purposeful adjustments. For mitigation efforts to become successful, individuals (and communities) must become aware of the hazards that affect them and realize they have a personal stake in avoiding, minimizing, and mitigating the costs and impacts of hazards on the local community.

Mitigation has been encouraged at the local level, because hazard control strategies implemented at the national level have been only partially successful. Flood mitigation policy in the United States provides a useful illustration. The structural era of flood control—federal construction of channels, dams, and levees—led in some instances to an increase in vulnerability from more extreme events (Platt 1996, 1999). These structures created a false sense of security and an assumption that the flood threat had been eliminated, resulting in increased human activity in floodplains, thereby increasing vulnerability to flood hazards. Given the relative lack of success in actually controlling floods (modifying the cause), it has become apparent that a less costly and more appropriate approach to flood mitigation should focus on land-use patterns within floodplains.

In the United States, for example, the National Flood Insurance Program (NFIP) requires appropriate floodplain development by communities that want to participate in this national insurance program. While spreading the loss, this program does have its drawbacks and critics (Platt 1999). The NFIP has the potential to encourage development within the floodplain precisely because low-cost insurance is provided by the federal government, thus reducing the incentive for homeowners to mitigate on their own such as floodproofing their home or moving away from the floodplain altogether.

3.2 Perception of Risks and Hazards

The perception of hazards and risks play an important role in understanding how and why individuals choose specific adjustments (Burton et al. 1993). Depending on the perception of the threat, those at risk will either seek additional information, seek actions to protect themselves, or continue with their normal activities. Hazards not considered threatening, whether due to lack of information, lack of scientific certainty, or some other factor, result in little or no preparation. Factors that influence the perception of hazards include hazard frequency and past experience, cultural/social attributes (age, wealth status, gender), and the credibility of hazard information. The perception of the environmental threat also colors the choice of adjustments people are willing to undertake to curb damaging effects. For example, the public is aware of the causes and potential impacts of global warming, although the linkages between individual human activities and global impacts are not completely understood. Because individuals do not 'see' their own contributions in greenhouse gas emissions or feel any personal responsibility for the hazard, they are unlikely to adopt any short-term adjustments such as purchasing more fuel-efficient cars or driving less.

3.3 Experience

As noted above, a belief must exist that lives and property are susceptible to harm from a particular hazard before individuals or organizations will undertake mitigation efforts (Palm 1990). Adjustments based on experience are subject to the material capabilities and personality constraints discussed next. It has been shown that individuals with little or no experience with a hazard are less likely to accept information, recognize the seriousness of the threat, or take appropriate action to protect themselves from harm. For some hazards there is little direct human experience. In these instances, the choice of adjustments will be based primarily on the perception of the threat and belief in the scientific certainty about potential impacts. In the case of global warming, the perception of the threat is low (requiring very few adjustments), whereas in the case of nuclear power

plants the public perception of risk is fairly high, resulting in very strict siting, regulatory, and safety controls.

3.4 Socioeconomic Characteristics and Personality

Socioeconomic characteristics such as gender, wealth, education, age, race, and ethnicity are not only correlated with risk and hazard perception, but also with adjustment adoption. For example, women have been shown to perceive risks differently, to be more risk-averse than men, to experience risk and hazards more often, to take a longer time to recover after disasters, and to undertake substantial preventative or mitigative actions in response to environmental hazards (Enarson and Morrow 1998).

Wealth also constrains the adoption of adjustments. Insurance, while technically feasible, is not an option for families with limited incomes. As a general rule, the poor are more susceptible to hazards owing to lack of resources, poor housing quality, and an inability to recover quickly. A host of equity literature is also supportive of race and ethnicity as contributing factors in creating vulnerability. Race and ethnicity shape the adoption of adjustments through differences in hazard perception and experiences, an understanding of and belief in the science behind the hazard information, and attitudes toward the dissemination of information.

Finally, personality influences adjustments since they are further narrowed by one's belief in their ability to control events. There is evidence that suggests that 'locus of control' influences the choice of adjustments. For example, if one feels powerless in the face of natural events (a fatalistic belief), then that individual is more likely to remain in place despite hurricane evacuation orders. Conversely, individuals who feel that they can control events around them, may take other courses of actions in response to an environmental threat such as floodproofing a home, evacuating out of harm's way, or reducing energy consumption to produce fewer greenhouse gas emissions.

4. Adaptation Versus Mitigation

Adapting to environmental threats means undertaking long-term changes in the ways we use the environment and interact with it without trying to prevent the landscape transformations from initially taking place (Kempton et al. 1995). Adaptations (sea walls and dikes, permanent relocation of populations, permanent closure of facilities) do little to reduce the source of the threat. Rather, the actions taken are in direct response to reducing human vulnerability.

Mitigation refers to the short-term actions that relieve losses (insurance, housing rehabilitation), but it is also synonymous with preventative actions that guard against future hazards and their associated losses. The implementation of either adaptation or mitigation strategies depends on the nature of the environmental threat, and will require public education and information, technology development and transfer, and economic, financial, legal, and institutional mechanisms. Further, each strategy will be distinguished by its capability for implementation at the local, regional/state, national, or international levels.

The drought hazard provides a good example of the difference between adaptive and mitigative responses. Societal dependence on the water budget for a particular area can lead to disaster when these supplies fail. While the drought 'trigger' may be found in the physical world, drought impacts should be viewed as resulting from the variability in the supply of and demand for water.

Since most droughts are local or regional occurrences, the initial policy responses tend to develop from local and/or regional concerns. In the United States, the federal government is involved in most aspects of water resource planning, but most drought policy is enacted by individual states and there is no consistency among them. In other developed countries, drought initiatives generally come from the central government, but normally in the form of aid to get the local constituents to change behavior. In Australia, for example, the national government encourages agricultural producers to adopt self-reliant procedures in managing climate variability. The government provides education in risk management, sustainable agriculture, and drought preparedness and only provides direct assistance under exceptional drought circumstances. These actions tend to be more mitigative and encourage water consumers to modify their behavior and practices to reduce the wasting of water.

In developing countries, drought response tends to be more adaptive. Since the GNP of many of these nations is much lower than those of developed countries, there is less money available for the government to spend on drought mitigation. The adjustment burden, therefore, falls on individual families who adapt by planting crops in seasonal floodplains, using drought-resistant varieties, or growing quick-ripening crops in preference to species that promise a higher average yield. These societies have managed to adapt successfully by being rich in resourcefulness despite their relative lack of financial wealth.

Geographic scale is crucial in understanding the distribution of hazards, their impact, and human responses to their reduction. While initiatives to curb losses may come from national or even global authorities, greater success is realized through the efforts at the local level. The ability to undertake significant mitigation or adaptation actions in response to environmental threats will depend on the

range of adjustment alternatives, the perception of the viability of them, their cost, and their relative impact in reducing the source of the environmental threat. These factors will vary over time and space and across social groups, economic sectors, and political jurisdictions.

See also: Climate Impacts; Consumption, Environmentally Significant; Environment Regulation: Legal Aspects; Environmental and Resource Management; Environmental Change and State Response; Environmental Planning; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; Environmental Vulnerability; Environmentalism, Politics of; Environmentalism: Preservation and Conservation; Global Environmental Change: Human Dimensions; Precautionary Principle; Sustainable Development

Bibliography

Blaikie P, Cannon T, Davis I, Wisner B 1994 *At Risk: Natural Hazards, People's Vulnerability, and Disasters*. Routledge, New York

Bohle H G, Downing T E, Watts M J 1994 Climate change and social vulnerability: The sociology and geography of food insecurity. *Global Environmental Change* **4**: 37–48

Burby R (ed.) 1998 *Cooperating with Nature: Confronting Natural Hazards with Land-use Planning for Sustainable Communities*. Joseph Henry, Washington, DC

Burton I R, Kates R W, White G F 1993 *The Environment as Hazard*, 2nd edn. Guilford Press, New York

Colborn T, Dumanoski D, Myers J P 1996 *Our Stolen Future*. Dutton, New York

Comfort L, Wisner B, Cutter S, Pulwarty R, Hewitt K, Oliver-Smith A, Wiener J, Fordham M, Peacock W, Krimgold F 1999 Reframing disaster policy: The global evolution of vulnerable communities. *Environmental Hazards* **1**: 39–44

Cutter S L (ed.) 1994 *Environmental Risks and Hazards*. Prentice-Hall, Englewood Cliffs, NJ

Cutter S L 1996 Vulnerability to environmental hazards. *Progress in Human Geography* **20**: 529–39

Dow K 1992 Exploring differences in our common future(s): The meaning of vulnerability to global change. *Geoforum* **23**: 417–36

Enarson E, Morrow B H (eds.) 1998 *The Gendered Terrain of Disasters: Through Women's Eyes*. Praeger, Westport, CT

Hewitt K 1997 *Regions of Risk: A Geographical Introduction to Disasters*. Longman, Harlow, UK

Kasperson J X, Kasperson R E, Turner B L II (eds.) 1995 *Regions at Risk: Comparisons of Threatened Environments*. United Nations University Press, Tokyo

Kempton W, Boster J S, Hartley J A 1995 *Environmental Values in American Culture*. MIT Press, Cambridge, MA

Kunreuther H, Roth R J Sr (eds.) 1998 *Paying the Price: The Status and Role of Insurance Against Natural Disasters in the United States*. Joseph Henry, Washington, DC

Mileti D S 1999 *Disasters by Design: A Reassessment of Natural Hazards in the United States*. Joseph Henry, Washington, DC

Oliver-Smith A, Hoffman S M (eds.) 1999 *The Angry Earth: Disaster in Anthropological Perspective*. Routledge, New York

Palm R 1990 *Natural Hazards: An Integrative Framework for Research and Planning*. Johns Hopkins University Press, Baltimore

Platt R H 1996 *Land Use and Society: Geography, Law, and Public Policy*. Island Press, Washington, DC

Platt R H 1999 *Disasters and Democracy: The Politics of Extreme Natural Disasters*. Island Press, Washington, DC

Turner B L II, Clark W C, Kates R W, Richards J F, Matthews J T, Meyer W B (eds.) 1990 *The Earth as Transformed by Human Action*. Cambridge University Press, Cambridge, UK

White G F, Haas E 1975 *Assessment of Research on Natural Hazards*. MIT Press, Cambridge, MA

S. L. Cutter and J. T. Mitchell

Environmental and Resource Management

1. Setting the Scene

The worldwide decay in environmental quality conditions and the gradual depletion of natural resources, sometimes referred to as the 'new scarcity,' has been a dominant theme for research and public policy during the latter part of the twentieth century. The global interest in environmental matters is partly caused by the increased pressures that a mounting population and increased production exert on the earth's natural resource base. In addition, as personal incomes rise and leisure time becomes more freely available in the developed world, concern with more immediate human needs has been accompanied by an interest in preservation and conservation for the sake of future generations as well as other regions of our world. We observe an increasing interest in 'quality'—next to 'quantity'—as an important constituent of individual and collective welfare.

It seems likely that the issue of environmental degradation will stay with us until far in the next millennium. Already Plato in his Kritias was complaining about human activity which had turned the landscape of Attica into a skeleton and a wasteland. We also know about environmental regulations in medieval European cities which aimed to control the use of coal burning or the noise annoyance caused by horse-drawn carriages. And the early stage of the Industrial Revolution demonstrated an abundance of urban environmental quality degradation. But in our era, the size and the intensity of resource use and of related environmental decay have taken such massive forms that the stability and sustainability of many ecosystems—both locally and globally—is threatened. This has stimulated the launching of the concept of sustainable development in research and policy making at the end of the 1980s, as well as the international

research on biodiversity loss and climate change risks during the 1990s.

Much of the early interest of the last century centered on public health in dense agglomerations. As sociomedical knowledge on how diseases spread developed, urban authorities sought to improve the local environment by measures like sewage control and clean water supply to reduce the diffusion of germs and infection. This trend, much later, led to policies embodied in various initiatives of clear air legislation to reduce local atmospheric pollution that causes smog and other harmful effects. Wealthy societies, and the better off within poorer societies, with the time and resources to expand, became concerned with the built environment and with shaping nature in ways that they found aesthetically pleasing. Over the centuries this has led to specific patterns of landscapes in the countryside and the provision of parks and gardens in urban areas. This has created a general concern on quality of life, in addition to economic dependence on resources and environment.

The resource and environment base of our earth is a complex system that comprises an enormous diversity of forms consisting of biotic and nonbiotic components. Essentially, the resource and environment base is made up of three main interrelated classes: biotic resources emerging from soil or water (such as vegetation and animal populations); abiotic resources with productive or consumptive values (such as minerals and energy); and environmental components needed for our well-being (such as clean water or fresh air).

Traditionally, resources are subdivided into renewables (which may be replenished with or without human intervention) and nonrenewables (which cannot be reproduced within a meaningful time horizon). Both types of resources can be depleted.

The main problem in environment and resource management is the interwoven nature of the 'new scarcity' question. Resource use leads to pollution, and abatement of pollution requires additional resource use. Clearly, technological innovation (e.g., recycling) may help to alleviate some of these problems, but with a rising population, and rapidly rising consumption and mobility levels it will be extremely difficult to achieve a sustainable development over a time scale involving many human generations. Furthermore, resource and environment issues may manifest themselves at local or regional scales, but they are part of a globally interwoven ecosystem. Consequently, the 'new scarcity' has clear spatial and temporal horizons which extend far beyond the current level of thinking and acting.

Against this background, the concept of sustainable development has become 'en vogue.' Although this notion has already a longer period of existence, it became fashionable after the publication of 'Our Common Future' by the World Commission on Environment and Development (WCED) (also known as the 'Brundtland Commission'). The Commission called attention for the need to consider our planet as an integrated social, economic, ecological and political system which needs collective initiatives and actions in order to ensure continuity under changing conditions. The report suggested that economic growth and environmental protection can go together, which was subsequently tested by Duchin et al. (1994) using a multisector–multiregion model for the world. Their conclusion was negative.

2. *Back to the Roots*

The scientific analysis of the physical conditions of our earth has already a history of many centuries, as is witnessed by the body of knowledge collected in such disciplines as biology, physical geography, archaeology and chemistry. The social sciences were latecomers in studying issues of environmental quality and nature. Of course, there are early examples of social science research on environmental and resource issues, for instance, by nineteenth-century scientists like Malthus and Marx. But it lasted until the 1960s and 1970s when the urgency of environmental degradation and resource depletion had become so widely recognized that various social science disciplines started to develop a pathway for environmental research. This does not mean that in some disciplines attention had never been given to resource and environment issues. On the contrary, in economics we observe a long-standing interest in resource scarcity, in relation to agricultural land use and mineral use. The Physiocrat thinking in economics even advocated that nature—in particular, land and water—were the real sources of economic wealth. But this early interest in resource use was mainly instigated by the (positive) productive contribution of physical resources to economic activity. Only a few economists in the first part of the twentieth century recognized the 'unpriced scarcity' nature of the environment and of many resources (notably Pigou). Marshall introduced the concept of 'externalities,' which meant that the social value of various goods and services, including nature and the environment, is not or is insufficiently reflected in market prices, so that a socially undesirable use of these goods and services will follow. This may be the result of myopic behavior of selfish people, the public good nature of various resources, the lack of control by property rights or proper regulations.

The real interest in environmental issues started essentially in the 1970s when the growing recognition of a worldwide environmental decay and of severe resource depletion (in combination with a population explosion) received an unexpected but welcome support in the oil crisis. This sudden event was complemented by the First Report to the Club of Rome ('The Limits to Growth'), which was based on a scenario analysis with a systems dynamics model of the World.

Although the scientific contents of the latter study left much to be desired—as a result of conservative information on resource availability and insufficient incorporation of negative behavioral and technological feedback mechanisms—it created a shock effect among social scientists. This gave rise to an intensive debate between growth optimists and pessimists (Daly and Townsend 1993).

This also marked the beginning of the social science interest in environmental sciences. Economists constructed abstract models of economic growth and resource use (Dasgupta and Heal 1979) and developed a theory of environmental policy for correcting environmental externalities (Baumol and Oates 1988). Economists and psychologists began to investigate how people value environmental change, environmental policy and ecosystem management, using stated preference and revealed preference methods (Hanley and Spash 1993). Demographers started to investigate the relationship between resource scarcity, population growth and migration. Decision theorists tried to develop new tools for policy-making that were more tuned to the often qualitative and unpriced nature of environmental goods (such as multicriteria and multi-objective decision tools). Many disciplines worked together in integrated modeling and assessment, addressing both ecosystem and global scales (climate). Finally, statisticians became involved in the formidable task of developing new statistical data that would map out and monitor energy use, environmental deterioration and the like. So, in some 25 years time there has been an explosive interest among social scientist in environment and resource issues.

It should be noted that there are many ways for a simultaneous analytical treatment of economics and environment. Since the 1960s, a great many attempts has been made to link economics to the ecology (Costanza et al. 1997). An important contribution began simply with a reflection on the principles of the materials balance for resources (extracted or collected, transformed, consumed and emitted) and on the need to take account of an economic viewpoint of such processes (Ayres et al. 1999). Several attempts have also been made to build economic and social accounting systems that could incorporate the measurement of economic welfare and performance together with the measurement of environmental indicators and performances. The integration of economics with ecology has also been approached from the viewpoint of land-use—where economic and ecological processes have the most disruptive effects—and of urban environments. In addition, the interaction between economics and ecology has next been dealt with for situations with global risks and uncertainties.

In the past decade, one can observe an avalanche of publications on environmental issues, not only in popular journals, but also in academic publication channels (journals, books). Most of these publications demonstrate clearly the multidisciplinary nature of environmental sciences, which caused researchers from various disciplines to cooperate, with more or less success. In particular, the field of 'ecological economics' (and its journal with the same title) has shown the relevance of multidisciplinary thinking in environmental and resource economics.

3. *A Short Account of Core Topics*

In this section, we will offer a concise overview of major topics in the area of management and economics of environment and resources. These issues are covered in detail in van den Bergh (1999). A very readable introduction is Kahn (1998), while a thorough treatment is provided by Perman et al. (1999).

A major focal point with already a long tradition is—as mentioned above—natural resource management, for both renewable and nonrenewable resources (Clark 1990; Dasgupta and Heal 1979). Questions analyzed in this context concern optimal extraction of non-renewable resources under competitive conditions (such as the well-known Hotelling cake-eating choice dilemma). In the literature on these issues, much attention is given to the (as yet uncertain) options offered by technological innovations and to the statistical uncertainty in estimating the resource stock (so far all estimates appeared to be rather conservative). This relates to the problem that the correct way of measuring resource scarcity is subject to debate, and has given rise to a range of scarcity indicators. From a management perspective, much attention has also been devoted to the implications of limited competition (e.g., monopolistic and oligopolistic market forms). Also renewable resources have gained much interest, in particular in the agricultural, fishery and forestry sector. Important issues here are uncertainty, re-investment rules and multiple-use conditions. Finally, in recent years we have observed an intensified interest in two particularly important resources with a wide societal and global coverage, viz. water and energy. These two resources are important concerns in both the developed and the developing world, not only because of their immediate far reaching implications of emerging scarcity but also because of the changing markets of these goods towards more deregulation and privatization.

A recurrent theme in environmental and resource economics is the economic valuation and evaluation of environmental goods, such as clean air or water. The main problem here is the nonpriced nature of many environmental goods, so that the environment runs the risk to be forgotten in any cost–benefit analysis for plan or project evaluation. The question is then whether it is possible to restore a missing market, e.g., by introducing shadow prices via taxation schemes or an artificial market, or to edict regulations to ensure at least a minimum respect for environmental goods,

e.g., in the form of standards. Although from a market theory perspective shadow prices (including taxes and subsidies) are often regarded as superior, in practice many forms of standard stetting have emerged, as they are more unambiguous and better testable. In the literature a wide variety of valuation approaches has been proposed. These can be categorized as travel-cost methods (based on revealed preference approaches using generalized travel costs), contingent valuation methods (based on state preferences, i.e., experimental surveys or questionnaires), and hedonic pricing methods (based on revealed preferences related to property values). A prominent question in valuation theory is the reliability of the various value estimates. Various considerations are relevant in this respect: the strategic contents of responses to interviews or survey questionnaires; the distinction between use and nonuse values; the choice between 'willingness to pay' and 'willingness to accept compensation' measures; and scale and double-counting issues (Hanley and Spash 1993, Turner et al. 1999). Recently, benefits transfer and meta-analysis have been used to avoid having to undertake costly new valuation studies. For environmental policy purposes various adjusted approaches have also been developed, such as social cost–benefit analysis and multicriteria evaluation methods. These methods have been widely applied. Nevertheless, their application faces various caveats, in particular problems associated with choosing a correct social rate of discount, incorporating spatial or temporal equity issues, and addressing problems where only qualitative or fuzzy information is available.

Another topic that makes up an indispensable component of the study of the environment and natural resources is the geographical dimension. Many environmental problems may seem to be local or regional issues, but have a worldwide impact. This is caused by two factors, viz. international (or interregional) trade and the spatial diffusion of pollution (or the spatial dimension of externalities in general). For example, the high efficiency of the meat industry in Western Europe is co-determined by the importance of tapioca from Thailand, which leads to widespread environmental decay in the latter country. Phenomena like ozone-layer destruction, the rise in ocean levels and desertification are other illustrations of the geographically interwoven nature of the environmental problem. As a result, cross-boundary and global environmental problems have provoked much scientific debate, not only on the causes but also on the policy responses (e.g., international environmental agreements, trade regulations, etc.). Such issues play also a recurrent role in negotiations on trade liberalization in the context of the World Trade Organization (WTO).

In a more theoretical setting, much attention is also given to equilibrium-distorting effects of environmental externalities in international trade. Issues investigated there are the effects of imperfect competition, industrial specialization and interlinkages; the consequences of unilateral environmental regulations; trade in pollutants (e.g., toxic materials); and the position of Third World countries. Cooperative and non-cooperative game strategies are studied in the context of global warming, climate change, acid rain, stratospheric ozone and global biodiversity. And finally, from a geographical perspective, much attention has been paid to point versus nonpoint source pollution problems, land use, urban environmental issues, locational issues and transportation. In retrospect, the spatial dimension of resource and environment issues has been addressed on various occasions and from different perspectives, but there is certainly still a broad research field in the geography of the environment.

Next, we will address the problem of environmental and resource policy. The frequent absence of markets for environmental goods, i.e., market failures, leads to a distortion in the price system and hence to an inefficient allocation, which causes environmental decay, resource depletion and the acceptance of high environmental risks. There has been a formidable range of publications on environment and resource policy, in particular on the choice of policy instruments; on different taxation regimes, technical and product standards; on subsidies, tradable permits, deposit-refund systems, stimulating self-regulation via covenants; and on joint implementation. In economic theory, efficiency has been the dominant policy selection criterion. However, in practice instrument choices will also be influenced by other criteria, such as effectiveness (relating to uncertainty, monitoring and control), equity (social and political feasibility), sustainability (ecosystem stability and resource availability) and international feasibility (agreements, laws and policy coordination). Various topics have been discussed in relation to the choice of instruments of environmental policy: the impact of imperfect markets; the encouragement of technological innovation; research and development; environmental tax reforms; and the role of transaction costs. Certain proposals for environmental tax revisions, or 'ecotaxes,' in popular jargon, have been argued by some to create a double dividend, that is, a beneficial outcome for both environment and employment. In general, however, economists are very skeptical about the existence of 'win–win' (or 'no-regret' or 'low-fruit picking') policies. This is partly related to the inevitable efficiency–equity trade-offs in policy-making. Finally, in the past few years we have witnessed a large flow of scientific contributions on international agreements and negotiations regarding transboundary or global environmental and resource issues. Harmonization of international regulations appears to be still a field fraught with many conflicts and divergent policy strategies.

As a last topic, the scientific apparatus developed to come to grips with the complexity of the new scarcity

will be mentioned. Standard monodisciplinary approaches cannot always be applied in a straightforward manner, as always multiple dimensions and disciplines have to be taken into consideration. Examples range from lessons obtained from thermodynamic theory to marketing principles being applied to eco-labeling. Mass balance calculations are needed to undertake life cycle analysis, and analogies of ecosystem models may be used in the context of 'industrial metabolism.' It is therefore no surprise that a uniform analysis framework has not emerged. On the contrary, we observe a great diversity of methods and models. Some important classes are: input–output and materials balance models; optimal control and dynamic programming models; partial and general equilibrium models; policy evaluation methods such as cost–benefit and multi-criteria analysis; game theoretic models; resource extraction models; natural resource accounting methods; evolutionary and other dynamic simulation approaches; meta-analysis; and experimental methods.

The previous overview has clearly highlighted the need for proper communication in the environmental sciences. Progress—in terms of new insights and impacts on policy—will only be made if environment and resource issues are jointly analyzed within common themes in research.

4. *Towards a Sustainable Environment*

Doomsday prophets have spent much time building an image of the world where environmental deterioration would herald apocalyptic perspectives. It is surprising to note that in many parts of the world we have seen a recovery of threatened ecosystems and environmental quality conditions as a result of active policy intervention, technological progress and behavioral change. It seems as though economic growth is not necessarily at odds with environmental sustainability.

Environmental sustainability is often divided into weak and strong sustainability. 'Strong' refers to the idea that every component of the environment and nature has to be maintained, with 'Deep ecology' as its most extreme proponent. 'Weak' refers to a change that makes some environmental components worse off, provided that the overall net balance is still positive. The distinction relates to ethical positions as well as views on the possibility of substitution (as opposed to complementarity) between nature or environmental goods and services on the one hand and socio-economic artifacts on the other hand. The question of weak and strong environmental sustainability is also co-determined by the degree of disaggregation of the environmental system under consideration, by the geographic scale at which environmental problems are studied and by the time horizon.

The notion of sustainable development is related to the old debate on growth versus environmental protection (Daly and Townsend 1993). In the context of endogenous growth theory it seems plausible that economic progress may be accompanied by environmental quality improvement, provided a sufficient part of public and private expenditures is allocated for environmental protection measures, for environmental investments and for environmental research and development. This has inspired a debate on sustainable growth. Clearly, such a discussion can only lead to conclusive answers if it is supported by empirical facts. This discussion on the 'decoupling' or 'delinking' hypothesis has laid the foundation for the so-called 'environmental' or 'green Kuznets' curve, where the underlying assumption is that after a stage of environmental decay as a result of economic growth, a new stage may set in during which economic growth supports a more environmental benign development. This growth optimism, however, is thus far not supported by sufficiently reliable statistical data, although it is clear that there are many examples of environmental improvement in a growing economy. But the green Kuznets curve does certainly not hold for all environmental indicators. A related important issue is the attempts to correct the conventional measure of welfare, GDP per capita, for environmental decay. Although this would not lead to a perfect indicator of welfare or progress, it could render a much better indicator for public and private decision-making, notably by central banks, international financial institutions and financial markets.

Just as the academic interest in environmental problems has evolved, so has the institutional setting in which they are addressed. There has been a movement away from micro-policy making in the hands of city councils, regional public bodies and the like, to macro-policy making, involving global agencies such as the United Nations and World Bank. Indeed, one of the pressing issues of contemporary environmental debate is the appropriate jurisdiction for different levels of government.

It is thus clear that the concept of sustainable development has been the source of much creative environmental research in the past years. But effective implications for environment and resource management still have fully to emerge.

5. *Concluding Remarks*

Environmental analysis extends across virtually all academic disciplines. Indeed, many disciplines such as economics and engineering have specific subareas of interest specifically devoted to environmental issues. The multidisciplinary and transdisciplinary nature of environmental research makes it difficult to keep track of the key literature.

In a few decades time, an overwhelming amount of research effort and research findings on environment

and resource issues can be observed. Environmental science has become a rapidly evolving field with a strong social science component. Environmental and resource economics has generated many insights about environmental policy design, natural resource management, development and growth policies, international trade and environment, transboundary pollution issues, and international agreements and policy coordination.

A rapidly evolving field is environmental business management (sometimes referred to as 'environmental management' or 'business and the environment'). This brings together scientific insights and business experiences. Many firms, especially larger ones, have accepted the importance of environmentally sustainable action and strategy, for various reasons. It has not only led to many in-house adjustments (waste treatment, environmental accounting, internal organization), but also to strategic decisions about 'green marketing,' eco-labeling, investments and involvement of various stakeholders. So far, however, a close link with the theoretical and empirical literature in traditional environmental and resource economics is lacking. A fundamental and longstanding question that their synergy could resolve is how firms make decisions about environmental management in complex and international situations, which are dominated by multiple stakeholders, rapidly changing markets and new information and communication technologies. This could provide information on what type of environmental regulation is most effective and efficient in the long run.

In conclusion, resource and environmental economics has created a breadth and depth of scientific insights into the new scarcity. Both theoretical and empirical knowledge has been generated in a surprisingly rapid pace. A multidisciplinary orientation has offered a great strength to this relatively young discipline in attacking complex and topical policy issues.

See also: Ecological Economics; Ecotourism; Environment Regulation: Legal Aspects; Environmental Change and State Response; Environmental Health and Safety: Social Aspects; Environmental Justice; Environmental Planning; Environmental Policy; Environmental Security; Environmental Vulnerability; Environmentalism, Politics of; Human–Environment Relationship: Indicators; International Justice: Philosophical Aspects; Regulatory Agencies; Resource Institutions; Sustainability Transition: Human–Environment Relationship; Sustainable Development; Water Pollution: Terrestrial and Groundwater

Bibliography

Ayres R U, Button K J, Nijkamp P (eds.) 1999 *Global Aspects of the Environment*. Elgar, Cheltenham, UK, Vols. 1 and 2

Baumol W J, Oates W E 1988 *The Theory of Environmental Policy*, 2nd edn. Cambridge University Press, Cambridge, UK

Clark C W 1990 *Mathematical Bioeconomics: The Optimal Management of Renewable Resources*, 2nd edn. Wiley, New York

Costanza R, Perrings C, Cleveland C (eds.) 1997 *The Development of Ecological Economics*. Elgar, Cheltenham, UK

Daly H E, Townsend K N (eds.) 1993 *Valuing the Earth: Economic, Ecology and Ethics*. MIT Press, Cambridge, MA

Dasgupta P S, Heal G M 1979 *Economic Theory and Exhaustible Resources*. Cambridge University Press, Cambridge, UK

Duchin F, Lange G M, Thonstad K, Idenburg A 1994 *The Future of the Environment: Ecological Economics and Technical Change*. Oxford University Press, Oxford, UK

Hanley N, Spash C L 1993 *Cost–Benefit Analysis and the Environment*. Elgar, Cheltenham, UK

Kahn J R 1998 *The Economic Approach to Environmental and Natural Resources*, 2nd edn. The Dryden Press, Fort Worth, TX

Perman R Y, Ma Y, McGilvray J 1999 *Natural Resource & Environmental Economics*. Longman, London

Turner R K, Button K, Nijkamp P (eds.) 1999 *Ecosystems and Nature: Economics, Science and Policy*. Elgar, Cheltenham, UK

van den Bergh J C J M (ed.) 1999 *Handbook of Environmental and Resource Economics*. Elgar, Cheltenham, UK

P. Nijkamp and J. C. J. M. van den Bergh

Environmental Archaeology

1. Description

As a result of historical and cultural differences in the development of the field, what is called environmental archaeology in English, Spanish, and Italian is known in other languages as the equivalent of ecological archaeology or archaeology of the milieu. Because of the underlying differences in perspective and the absence of a coherent theoretical, methodological, or technical framework, it is a 'field of studies' rather than a 'discipline.' Its focus could be defined as: 'the study of the past of our terrestrial milieu, its relationship to humans, and the latter's relationship to their environment.'

Recent reference works omit the concept (Mignon 1993) or define it as 'a broad range of specialized studies in archaeology that pertain to prehistoric human-environment interactions' (Michaels 1996). These studies apply methods from the natural and life sciences to the study of past environments and the ways people exploited and changed them. Depending on the application or the discipline of origin, they are commonly divided into the following.

Geoarchaeology includes geomorphology, archaeopedology, and archaeometry (the application of geophysical prospection methods to archaeology) as well as (sometimes) isotopic dating and characteriz-

ation methods, and the rapidly expanding field of soil micromorphology of ancient sites and landscapes. Among the *archaeobotanical* methods one finds palynology, macroscopic and microscopic analyses of charcoal, fruits, or seeds, and the study of fossil plant crystals (phytoliths). The latter is growing in importance as it does not depend on the preservation of organic materials.

The study of ancient animal remains is the preserve of *archaeozoology* (or zooarchaeology, see *Zooarchaeology*), which identifies the animals, determines their age and sex, reconstructs their population dynamics, etc. to better understand their significance to prehistoric people. In this domain, isotope analyses have recently added a dimension.

Finally, bioarchaeology deals with humankind: human growth and growth deficiencies, activity levels, nutrition, parasites, illnesses, and much more. In the latter three of these fields, DNA studies promise to open new avenues.

In the meantime, the essential unity of environmental archaeology (as opposed to the disciplinary diversity of the approaches to it) is gradually being recognized (Reitz et al. 1996). The tardiness of that development and the difficulties it encounters are rooted in its intellectual origins.

2. *The Roots and Early Development of Environmental Archaeology*

Since the fourteenth century the natural sciences have flourished by focusing on issues concerning closed systems, in which independent, irreversible changes have played a negligible role. They successfully encroached on many areas which for centuries had been considered outside their remit, creating an opposition to the historical disciplines (Evernden 1992). By the mid-nineteenth century, they had fundamentally changed our worldview to the point that, in everyday life, 'nature' was increasingly preferred over 'tradition' or 'the past' to explain self-evident observations (Girard 1990). As the then dominant schools of thought dealing with the long term, 'natural history' and *Kulturgeschichte* illustrate the fundamental dilemma of environmental archaeology.

'Natural history' is the forerunner of evolutionary biology; in both one first explains the observed functioning of physical or biological systems in terms of rigorously derived laws or general principles, and then attempts to reconstruct their possible antecedent states (Jonas 1982). In the absence of an accepted explanatory theory, early nineteenth-century natural historians either assumed that God created things as they are, or avoided the problem altogether by focusing on systematics rather than origins. J.-B. de Lamarck (1744–1829) is an exception, accepting intragenerational changes but not their inheritance (1802, 1809). For the *Kulturgeschichte* school, past and present are part of an irreversible, complex process, and writing history involves simplifying it through a continued back and forth between the observed past and the present in which one interprets. The nonrecurrent event is considered an essential aspect of human behavior, and the historian is concerned with the impacts of chance as well as of the role of individuals. A fundamental refusal of the positivist position dominant in the natural sciences underpins these ideas. W. Dilthey (1833–1911) summarized it as follows: 'We explain nature, we understand humanity' (Dilthey 1988).

Beginning in the 1850s, new approaches emerge. Charles Darwin (1809–82) explains (1859) long-term 'natural' change by arguing that the first living beings could have been much simpler than the present ones. Distinguishing ontogenetic from phylogenetic evolution allows him to explain the past and the present of all species by invoking 'variation' and 'natural selection,' mechanisms operating at the metalevel of the long-term existence of species, rather than at that of the individual and/or the single generation. Moreover, the theory of evolution breaks the iron grip of reversibility and/or replicability on explanation, and facilitates the reintroduction of historical explanation in the realm of nature. At roughly the same time, sociologists such as E. Durkheim (1893, 1895), building on the work of G. Vico (1668–1744), C.-L. de Montesquieu (1689–1755), M.-J. de Condorcet (1743–94), and A. Comte (1798–1857), argued for the importance of the study of the social and functional dynamics of societies, while anthropologists such as E. B. Tylor (1871, 1878) and L. H. Morgan (1877) argued for a universal pattern of development of all societies. Their 'social evolution' introduced 'naturalist' elements into the study of 'culture,' breaking with the *Kulturgeschichte* school.

Emerging in this climate, between the humanities and the natural sciences, prehistoric archaeology and environmental studies constituted an ambiguous middle ground because they dealt with phenomena which change qualitatively and irreversibly during the period of observation, and with the interface between people and their natural environment. Neither fit the mechanistic approach of the natural sciences because these exclude the study of qualitative change, nor did they fit the historical approach, which focuses on the human, nonrecurrent, aspects of behavior. Like 'natural history,' prehistoric archaeology reconstructs a past that moves from the simple to the complex, and in order to write it, simplifies the complex present. It involves description of the transformation of processes (in which neither chance nor idiosyncrasies play a part). This contrasts with classical, medieval, and later archaeology in Europe, as well as the archaeology of the great civilizations, which have evolved in the historical realm. To this day, these two archaeologies operate very differently.

In 1866 E. Haeckel defined his 'new science of ecology,' as 'the science of the relationships of the organism with its environment, including all conditions of existence in the widest sense' (Haeckel 1866). He does not include humankind in it—in effect, he defines 'environment' in much the same way as 'nature' was defined a few centuries earlier, as 'non-organism.' Darwin on the other hand reconfigured Lamarck's concept of 'environment' as 'the conditions of natural selection,' and included mankind in his 'web of life.' Thus, the distant past and the environment both became objectifiable and separable, fit to be studied by rigorous, 'scientific' methods within a 'naturalist' concert of disciplines, but dealing with humanity as well as the nonhuman natural world. It is thus no surprise that environmental archaeology enters the arena of the archaeological disciplines as the handmaiden of prehistory.

The study of the natural environment of prehistoric man developed first in those parts of Europe in which traditional 'classical' or 'literary' archaeology was the least dominant: Scandinavia, the Low Countries, Northern Germany, Switzerland, and the UK. The first known formal association between prehistoric archaeology and the biological sciences is the so-called 'Danish Royal Commission for Research on Shell Middens' instituted in 1848, which included the leading people of the day, such as J. J. A. Worsaae, J. Steenstrup, and J. S. Forchhammer. This is part of a blossoming interest in past environments in the nineteenth century, in which geology played an important part, as exemplified by G. de Geer's (1858–1943) work on varve chronology in Sweden, and E. J. Lennart von Post's work on prehistoric palynology (1884–1951). In Switzerland, A. Morlot followed in these footsteps when the first lakeshore dwellings were excavated in the 1850s and 1860s (1861).

In The Netherlands, at about the same time, paleogeographers seem to have been convinced that landscapes reflect processes in which natural and human dynamics both play an important part. In 1920, a plant physiologist (A. E. van Giffen) was appointed to a chair in archaeology at the newly founded Biologisch-Archaeologisch Instituut of the University of Groningen, with the aim of doing integrated research on the origins and history of the Dutch coastal system. He notably stimulated archaeozoological and later palynological research, and instituted a university degree (at MA and Ph.D. level) in environmental archaeology just after World War II.

In the area of archaeological pollen analysis, the next steps were set by German scholars working in the moors in the northern part of their country (Bertsch, Reinerth). Their efforts eventually led to the foundation, around 1960, of the influential Institut für Küstenforschung in Wilhelmshafen (north Germany) by W. Haarnagel. But already before the war UK (H. Godwin 1933) and Scandinavian (Joh. Iversen) scientists had enlarged the temporal and spatial scope of palynological research.

In the UK, partly following in the footsteps of V. G. Childe (1928, 1929), J. G. D. Clark (1939, 1940), C. Vita Finzi (1969), E. Higgs (1972), and others gave the field an enormous impetus in the 1950s through to the 1970s under the name economic archaeology—emphasizing geology and archaeozoology. From the late 1960s, ideas of the Cambridge school of 'new geography' were introduced in UK (environmental) archaeology, initially by D. L. Clarke (1968).

In the USA the first attempts at studying environmental matters were again related to shell-middens (E. W. Gifford 1916). J. Steward (1937, 1955) and L. White (1949, 1975) are early but very influential protagonists of the idea that the environment greatly influences, if not determines, the core aspects of any culture, and proposed the combined, comparative study of culture and the environment. W. Wedel (1941) and E. Haury (1950) provided early stimuli through their fieldwork in other parts (the Great Plains and the Southwest, respectively). The expeditions organized by Braidwood in the 1950s and 1960s (Braidwood and Willey 1962) to various early prehistoric sites in the Near East constituted an important step because they brought specialists in all the necessary disciplines together in the field, disseminated the environmental approach to near eastern archaeologists all over the world, and provided training to 'stars' of the next generation (e.g., R. McC. Adams, K. V. Flannery). K. Butzer's work is an important turning point (1965, 1971) because of its breadth and interdisciplinary perspective, integrating not only the environmental disciplines but also geography, for example (1982).

3. *Oppositions Between Nature and Culture, Past and Present are at the Core of the Discipline*

As a result of the context in which it was born, and notwithstanding the exponential increase in the number of its practitioners during the last 30 years of the twentieth century, and the increasingly direct confrontation of the modern world with environmental problems, environmental archaeology has had difficulty in defining its identity by resolving the dichotomies which underpin it: 'nature–culture' and 'past–present.'

Until well into the 1970s, most environmental archaeologists viewed humans as reacting to (adapting and exploiting) the conditions set by the dynamics of nature. Then, partially under the impact of the 'green' movement, awareness grew of human impact on the environment, particularly in the area of climate, vegetation, and animal life. As the focus in environmental studies in general shifted to sustainability, a new perspective emerged. People were henceforth also

seen as proactive, albeit generally in a destructive sense. The difference can be illustrated by looking at the distinction, in French, between *le milieu*, that is, that which human groups are in the midst of, and *l'environnement*, that which surrounds human groups. Although in both cases, the focus is on what surrounds human beings, in the former the perspective is human. Nature's cohesion, its unknown aspects, its strangeness, and force are amplified, whilst the confusion and the handicaps of humanity are accentuated; change is attributed to nature, and people have no choice but to adapt to it.

In the latter case, the perspective is nonhuman: the cohesion and strength of nature are diminished, while the same properties are accentuated in humanity. The known aspects of nature seem to be more important, and it thus seems more controllable and loses its dangerous appearance. Humanity, in this perspective, is viewed as the source of change: people create their environment.

More recently, recognition is growing that both these points of view are essentially anthropocentric. Humans should be seen as 'just another unique species' (Foley 1987), and one cannot separate social dynamics from natural ones, even though different approaches are used to investigate them. McGlade (1995) thus argues: 'there is no natural (sub-)system, there is no social (sub-)system, there are only socio-natural interactions.' These interactions are driven by different kinds of dynamics within and between the two realms. Inanimate dynamics concern only energy and matter, whilst among plants and animals the role of information is limited. Within the life span of individual organisms, it is confined to stimulus-response and behavioral optimization, while over longer periods genetic transmission may induce structural changes. Only among human beings does the exchange of information regularly result in structural changes within an individual's lifetime. And although human impact on the environment is as immediate as that of other species, the human reaction to environmental change requires that a society becomes aware of environmental problems before it can respond to them (Luhman 1985). That requires a conjunction of both environmental and social dynamics, and interaction between these dynamics takes the form of resonance, rather than direct cause and effect. But bridging the gap also involves different ways of linking past and present. Ecologists have for some time been aware of this (e.g., Martinez Alier 1990), but had not really developed that angle until some adopted the complex systems approach, which offers a very tempting perspective to anyone striving to integrate the natural, life, and social sciences with the humanities:

(a) It conceptualizes problems in both natural and social dynamics in a language which is independent from any of the disciplines.

(b) It studies interaction between phenomena at different spatiotemporal scales by viewing macroscopic, apparently stable phenomena as the result of unstable interactions between microscopic entities.

(c) It reformulates natural dynamics from an irreversible temporal perspective by dealing with both process and event within the same approach, and placing each in the context of the other, thus relativizing them and facilitating description of historical processes in 'natural science' terms and vice versa

(d) Its dynamic descriptions of apparent structures introduce the notion that any structure operates in a field of tension and can be viewed as a flow, thus bridging the gap between pattern and process, and linking structures dynamically to their environment.

(e) By highlighting the metastability of many apparently stable patterns and structures, and the nonlinear character of many processes, it deals with chance and necessity, continuity and change, tradition and innovation, from one and the same perspective, thus moving away from our traditional bias in favor of stability, and our focus on investigating change.

(f) It invites us to rethink the relationship between cause and effect in all disciplines, because the effect of a change in a state variable depends to a very great extent on the state of the entire system at that moment, so that similar causes can have different results, and different causes similar results.

4. *The Challenge*

Surmounting the 'nature–culture' and 'past–present' oppositions seems to be coming within reach (and why not, as many other cultures have never created the oppositions in the first place, cf. Descola 1994?). But achieving it requires that environmental archaeology gains its own place under the sun in the 'concert of disciplines' by addressing questions which are central to the long-term relationship of people with their environment, and thus reflect the interests of ecologists, historians, geographers, and others as well. For example:

(a) Many purely ecological issues which demand a very long time frame to deal with them effectively, but for which historical or instrumental data are only available for at most a hundred years or so. Paleoclimatology operates at a global scale, for example, but the effects of climate differ between regions, and need to be investigated with techniques in which environmental archaeologists excel. Moreover, there are such issues as the long-term evolution of biodiversity (Odgaard 1999), biological successions, species interactions (Lawton 1999), etc. In such areas, ecology's origins as a natural science have often limited observation to short-term dynamics, and have confined explanation to a search for universals without consideration of individual events. And what about the long-term evolution of scalar interactions?

(b) Issues concerning human–environmental coevolution (Thompson 1999). Observing the behavior of human populations in a particular area over 10,000 years, for example, highlights the growing interdependency between the species involved and its effects on the resilience of that system (van der Leeuw 1998). Such studies therefore have an original contribution to make to present-day questions about 'sustainability.' Similarly, a long-term perspective on the relationship between changes in the natural dynamics and in human resource exploitation, technology, and spatial organization helps us to understand different landscapes (Naveh 1998). That in turn enables us to look at a much wider range of possible socionatural states of resilience than we would otherwise.

(c) At the human end of the spectrum, we will need to stand back and develop better overall ideas on such topics as (i) the dynamics responsible for long-term human evolution; (ii) the genesis of human societies; (iii) the way human biological and social evolution have slowly transformed our environment and in many cases made it dependent on its interaction with society; and (iv) the dynamics by which humans have over time, dealt with their need to subsist in that environment. Here archaeology has made considerable advances, but the coevolutionary aspects which are the particular domain of environmental archaeology have not kept pace.

(d) To create a viable future, we need to find new ways to articulate the different kinds of socionatural components in a landscape dynamic. As Di Castri (1997) has argued, the way the world is changing demands of ecologists that they move toward studying systems in which human beings dominate. This is exactly where environmental archaeologists are stronger than most 'pure' ecologists. Playing a role in such a reorientation of ecology, combining strands in ecology such as landscape ecology and restoration ecology with new movements in archaeology such as historical ecology, seems a very worthwhile challenge.

See also: Bioarchaeology; Environmental Adaptation and Adjustments; Environmental History; Environmental Sciences; Environmentalism: Preservation and Conservation; Evolutionary Approaches in Archaeology; Food Production, Origins of; Zooarchaeology

Bibliography

Butzer K W 1971 *Environment and Archaeology: An Ecological Approach to Prehistory*. Aldine, Chicago

Butzer K W 1982 *Archaeology as Human Ecology: Method and Theory for a Contextual Approach*. Cambridge University Press, Cambridge, UK

Darwin C 1859 *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Murray, London

Descola P 1994 *In the Society of Nature: A Native Ecology in Amazonia*. Cambridge University Press, Cambridge, UK

Di Castri F 1997 Editorial: Landscape ecology in a changing globalised environment. *Landscape Ecology* **12**: 3–5

Dilthey W 1988 *Introduction to the Human Sciences: An Attempt to Lay a Foundation for the Study of Society and History. [Einleitung in die Geisteswissenschaften]*. Wayne State University, Detroit, MI [1883 Trans. R J Betanzos]

Evernden N 1992 *The Social Creation of Nature*. Johns Hopkins Press, Baltimore, MD

Foley R A 1987 *Just Another Unique Species*. Cambridge University Press, Cambridge, UK

Girard R 1990 Innovation and repetition. *SubStance* **62/63**: 7–20

Haeckel E 1866 *Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-wissenschaft, mechanisch begründet durch die von Charles Darwin reformierte Descendenz-Theorie*. Reemer, Berlin

Jonas H 1982 *The Phenomenon of Life. Toward a Philosophical Biology*. University of Chicago Press, Chicago

Lamarck J-B de 1809 *Philosophie zoologique*. Chez Dentu et L'Auteur, Paris

Lawton J H 1999 Are there general laws in ecology? *Oikos* **84**: 177–92

Luhman N 1985 *Ecological Communication*. Polity Press, London

Martinez Alier J 1990 *Ecological Economics. Energy, Environment and Society*. Blackwell, Oxford, UK

McGlade J 1995 Archaeology and the ecodynamics of human modified landscapes. *Antiquity* **69**: 113–32

Michaels G 1996 Environmental archaeology. In: Fagan B (ed.) *The Oxford Companion to Archaeology*. Oxford University Press, Oxford, UK, p. 205

Mignon M R 1993 *Dictionary of Concepts in Archaeology*. Greenwood, Press, Westport, CO

Morgan L H 1877 *Ancient History, or Research in the Lines of Human Progress from Savagery though Barbarism to Civilization*. Leacock, Gloucester, MA

Naveh Z 1998 Ecological and cultural landscale restoration and the cultural evolution towards a post-industrial symbiosis between human society and nature. *Restoration Ecology* **6**: 135–43

Odgaard 1999 B V Fossil pollen as a record of past biodiversity. *Journal of Biogeography* **26**: 7–17

Reitz E, Newsom L A, Scudder S J 1996 Issues in environmental archaeology. In: Reitz E, Newsom L A, Scudder S J (eds.) *Case Studies in Environmental Arcchaeology*. Plenum, New York, pp. 3–16

Thompson J N 1999 The raw material for co-evolution. *Oikos* **84**: 5–16

Tylor E B 1871 *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art and Custom*. Murray, London

Tylor E B 1878 *Researches into the Early History of Mankind and the Development of Civilisation*, 2nd edn. Estes & Lauriat, Boston

Van der Leeuw S E (ed.) 1998 *The ARCHAEOMEDES Project: Understanding the Natural and Anthropogenic Causes of Land Degradation and Desertification in the Mediterranean Basin*. Publications of the European Union, Luxemburg

S. E. van der Leeuw

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Environmental Challenges in Organizations

Environmental challenges in organizational and management studies initially referred to the restrictions that the external environment imposes onto the survival of organizations. In the last decades of the twentieth century it has come to designate the survival of both organizations and the natural (or ecological) environment. The transformation occurred simultaneously in the practice of corporate environmental management, and in the studies concerning organization and environment. This article presents both the empirical responses that have been developed by formal organizations in facing environmental challenges as well as studies on environmental management and ecological sustainability.

1. The 'Environment' in Organizational and Management Studies

The environment in organizational and management studies has historically been treated as an organizational contingency (see *Structural Contingency Theory*). The environment was synonymous with the organizational environment, something abstract and a-temporal, having little to do with the natural environment (Shrivastava 1994). The main concern of theories of organizations in the 1960s was the restriction that the external environment imposes on the survival of organizations, rather than the environmental burdens that organizations cause to the natural environment.

The pervasive assumption in both research and management practice that organizations had a capability, and indeed should, control their environments was questioned in the work of the population ecology perspective (see *Ecology: Organizations*). The title suggests that the organizational ecology (as it is also known) is an ecological theory of organizations. In fact, it represents the transplantation of concepts and calculus techniques from the discipline of ecology to the world of formal organizations. The application of biological models to the analysis of the 'ecology of organizations' has nothing to do with the natural environment. The use of the concepts of 'environment' and 'ecology' as metaphors to express the 'organizational environment' or 'the ecology of organizations' was not an intentional strategy to deny the importance of environmental issues in organization studies. The all-encompassing meaning of the terms themselves is what could mislead one to think the opposite. Nonetheless, what makes these perspectives vulnerable to criticism is the dominant epistemological position about research in organization studies that they tend to assume. The positivistic rigidity that dominated organizational studies until the late 1970s resulted in a scope of research that did not allow much space for its own reflexivity.

The emergence of new metaphors and paradigms for organizational analysis in the late 1970s undermined the historic hegemony of functionalist perspectives (see *Organizations, Metaphors and Paradigms in*). These new theoretical approaches questioned some taken for granted assumptions about research and management, transforming organization studies into a 'contested terrain' (Clegg et al. 1996). The acknowledgment that the external environment affects organizational action in the same way that organizations impact on the natural environment began to appear in early 1990s, giving rise to the development of research and theories on organization and environment.

Implicit in the studies about environmental issues in organizations is the assumption that a theoretical imperfection allows companies to externalize their environmental and social costs on to society. Market prices do not always reflect total costs of products (Hawken 1993). There is a discontinuity between empirical outcomes and theoretical representations of organizational action. Environmental and social costs that occur during the life cycle of products—the extraction of raw materials, manufacturing, use, and final disposal or recycling—are not considered in most companies' accounting systems (Grey et al. 1993). Increasingly demanding regulatory measures, public demand for environmental protection and compensation, and consumer demand for cleaner processes and products, constitute some of the current environmental challenges requiring organizations to eliminate or internalize these costs. Firms responded to these challenges by developing novel environmental strategies and management practices, described in the next section.

2. Organizational Responses to Environmental Challenges

Until the 1970s it was common for companies to avoid compliance with environmental regulations. From 1970 to 1985 the general attitude of companies was to go only as far as to comply with centralized regulatory measures that commanded them to internalize environmental costs—normally achieved by installing pollution-control equipment. Typically, firms resisted the command and control type of regulation because it did not generate incentives for them to assume higher levels of environmental responsibility (Fisher and Schot 1993).

From 1985 to the mid-1990s, dialogue between legislators and industry resulted in business accepting some responsibility for solving environmental problems, at the same time that new regulatory mechanisms were being proposed as alternatives to the command and control approach. Industry associations, in partnership with governments and other

interest groups, responded to the consumer and public demand for environmental protection, and shareholder rejection of environmental risks by instituting voluntary business codes of environmental practice. The standardization of environmental management systems such as the European Eco-Management and Audit Scheme (EMAS) and the International Organization for Standardization (ISO) 14000 environmental series constituted another type of voluntary program developed in the 1990s. Eco-labeling programs have also been put into practice, aiming to inform consumers whether manufacturers' products meet certain criteria for environmental impacts during the various stages of a product's life.

Pollution prevention was the basic principle guiding these initiatives, which focuses on minimizing or eliminating waste before it is created by production processes. As a whole, voluntary schemes have not generated significant improvements in corporate environmental practices, and regulatory approaches based on extended producer responsibility are expected to emerge in the early decades of the twenty-first century. Corporations will be required to develop product stewardship strategies to minimize pollution not only during manufacturing but also to reduce environmental costs associated with the entire life cycle of their products (Hart 1997).

Economic incentives for the incorporation of ecological issues into business practices became more evident in the environmental management literature in the second half of the 1990s. Increasing costs of pollution control, the development of new technologies for pollution prevention, and cost-savings from waste reduction, emerged as driving forces for proactive corporate environmental management (also known as beyond compliance environmental practices). Ecological requirements started to be seen as an integral part of quality management principles and potential sources of competitive advantage. Porter and van der Linde (1995) asserted that environmentally friendly regulation could act as an incentive for companies to innovate and use resources more efficiently. The result would be a gain in both corporate profits and environmental improvements—a double dividend or win–win scenario. Support for this approach, however, has not been unanimous. Empirical evidence questions the extent to which one can generalize the results of specific case studies to the entirety of commerce (Howes et al. 1997).

Finding hidden opportunities for organizations to profit from environmental strategies became an appealing theme for both business and academic communities. Nonetheless, the double dividend debate inspires a more crucial issue in the study of organization and environment: the role of business in promoting sustainable development. Are corporate environmental management practices conducive to ecological sustainability? Can incremental improvements lead modern industrial societies to sustainable patterns of production and consumption? Are radical changes imperative to ecologically sustainable societies?

3. *Research on Environmental Management and Ecological Sustainability*

Research on organizational responses to environmental challenges as well as the identification of necessary changes for the creation of ecological sustainable organizations and societies frame the focus of current use of the 'environment' in organizational and management studies. Typically, the literature adopts one or other of radical or reformist approaches (Egri and Pinfield 1996). Radical studies normally have their foundations in areas such as deep ecology, social ecology, and eco-feminism, which consider biospecies egalitarianism a prerequisite for economic advancement in harmony with nature. In this view, radical changes in the values guiding organizational actions are required. Conversely, reformist approaches stress the necessity of working within the rules that currently guide managerial action. The studies assume that organizations—and more specifically, business enterprises—are both the responsible agents, and the ones most likely to be sufficiently powerful to promote environmental reform and lead modern societies towards *Natural Capitalism*—a term proposed by Hawken et al. (1999) to simplify the arguments favoring organizations to value ecosystems services by internalizing environmental costs.

The lack of evidence that organizations will change their behavior out of altruistic motives leads reformists to assume that incrementalism is the guiding principle of organizational change. The gradual incorporation of ecological values into corporate strategies and practices would happen over time, they argue. Incrementalism is the basic principle of technical approaches from the natural sciences to environmental management, and most corporate responses to environmental challenges, described in the previous section, are exemplars of this perspective.

By the end of the 1990s the number of publications in organization and environment remained small, compared with traditional areas of research, but the natural environment gradually gained ground in the study of organizations. Environmental issues entered research through traditional areas of organizational studies such as institutional theory and resource-based theory. Together with the (technical) area of environmental management, the majority of these studies tended to be reformist in nature, urging scholars to attend to the difference between practices of environmental management and research, education, and management for ecological sustainability (Roome 1998).

Environmental management practices in corporations are expected to reduce the impact caused by

industrial and commercial activity, but whether these practices are conducive to sustainable development remains uncertain. In practical terms, the business community tended to equal the concept of 'sustainable development' with 'eco-efficiency'—an extension of environmental management techniques in corporations. Theoretically, from the reformist perspective, there is still a notable lack of understanding whether incremental innovations in systems of production and consumption are conducive to ecological sustainability. Conversely, from the radical-perspective, if a biocentric world is to be achieved, such environmentalists have yet to develop a theory to make this transition feasible.

4. *Future Directions for Theory and Research in Organization and Environment*

In the context of the social sciences, particularly in organizational and management studies, environmental challenges increasingly demand consideration of interdependencies between natural and artificial (social, economic, technological) systems. In the emerging area of organization and environment the importance of corporations as the principal agencies responsible for environmental disruptions is acknowledged. Radical theorists demand a complete revision of the values guiding business: only a cultural revolution can lead to ecologically sustainable societies, they suggest. Reformists assume that ecological modernization of systems of production and consumption is the only viable alternative in the quest for sustainable development (Spaargaren and Mol 1992).

There is already widespread acceptance that 'sustainability' cannot be confined to the realms of focal organizations but requires co-ordination of economic, social, and environmental prerequisites at the industry and societal levels. Human rights issues, such as equity in trade and empowerment of minority groups, are seen as integral elements of ecologically sustainable organizational strategies and practices (Welford 1997). Economic globalization makes boundaries between social, economic, and environmental systems more diffuse and the development of interdisciplinary studies and cross-level analysis increasingly necessary.

The expertise accumulated in organizational and management studies, used in conjunction with disciplines such as engineering, economics, and sociology, can be expected to generate knowledge about current systems of production and consumption as well as alternative models for their reorganization. The mechanisms for the integration of ecological principles into business strategies and practices that foster or inhibit organizational learning and change need further research (Roome 1998). Understanding structural and behavioral prerequisites that might increase the pace of environmental innovations and inter-firm collaboration is also crucial for the creation of more efficient industrial systems, and consequent dematerialization of the economy. In sum, research about organizational strategies and practices that can lead to ecologically sustainable systems of production and consumption, constitute the main environmental challenge for organizational and management studies in the early decades of the new millennium.

See also: Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; Environmental Stress and Health; Environmentalism, Politics of; Human–Environment Relationships; Industrial Ecology; Industrial Metabolism

Bibliography

Clegg S, Hardy C, Noord W (eds.) 1996 *Handbook of Organization Studies*. Sage, London

Egri C, Pinfield L 1996 Organizations and the biosphere: Ecologies and environment. In: Clegg S, Hardy C, Noord W (eds.) *Handbook of Organizations*. Sage, London

Fischer K, Schot J (eds.) 1993 *Environmental Strategies for Industry: International Perspectives on Research Needs and Policy Implications*. Island Press, Washington, DC

Grey R, Bebbington J, Walters D 1993 *Accounting for the Environment*. Paul Chapman and the Association of Chartered Certified Accountants, London

Hart S 1997 Beyond greening: Strategies for a sustainable world. *Harvard Business Review* **75**: 67–76

Hawken P 1993 *The Ecology of Commerce*. Harper Collins, London

Hawken P, Lovins A, Lovins H 1999 *Natural Capitalism: Creating The Next Industrial Revolution*. Earthscan, London

Howes R, Skea J, Whelan B 1997 *Clean and Competitive? Motivating Environmental Performance in Industry*. Earthscan, London

Porter M, van der Linde C 1995 Toward a new conception of the environment–competitiveness relationship. *Journal of Economic Perspectives* **9**(4): 97–118

Roome N (ed.) 1998 *Sustainable Strategies for Industry*. Island Press, Washington, DC

Shrivastava P 1994 Castrated environment: Greening organization studies. *Organization Studies* **15**(5): 705–26

Spaargaren G, Mol A 1992 Sociology, environment and modernity: Ecological modernization as a theory of social change. *Society and Natural Resources* **5**: 323–44

Welford R 1997 *Hijacking Environmentalism: Corporate Responses to Sustainable Development*. Earthscan, London

R. J. Orssatto

Environmental Change and State Response

Environmental change is one of a myriad of pressures or demands made upon state resources and attention. Many of the responses to environmental stress that

occur involve uncoordinated human responses greatly affected by markets. Accordingly, producers and consumers respond to changes in prices, relative incomes, and external constraints. But frequently market 'signals' do not reflect social values, as in the case of intergenerational equity, for example, or the deleterious effects of environmental degradation are not internalized in market prices and remain as 'externalities.' As a result, states often choose to intervene with collective actions aimed at managing environmental change and reducing the associated adverse social and economic effects.

Accompanying environmental changes will be events and consequences that are potential 'signals' of threat to society. Such signals occur with or without societal recognition. If no recognition occurs, there is no (or limited) social learning or stimulation of action by the state, and the lack of response may accelerate the human driving forces of change as a means of positive feedback. If the degradation is recognized, whether by individual farmers, the media, or governmental officials, and elicits recognition and concern, then the response mechanisms of the state are set into action. Once the state receives and recognizes that an environmental change poses a threat, it typically moves to interpret the threat, to assess its social meaning and significance, and to locate it on the policy agenda (see *Environmental Policy*). Responses occur at many different levels, so scale may become an important factor in state response, particularly 'scaling-up' and 'scaling-down' problems. These refer to the extent to which patterns and causes of relationships, such as how institutions affect human behavior, observed at one scale can be expected to operate at higher or lower scales (Ostrom 1990, Young 1999). Such issues are only now beginning to receive concerted attention in the social sciences.

All societies, including states, engage in processes by which those at risk, the mass media, societal groups, institutions, and the state itself, assess whether a particular change involves a risk worthy of attention and how serious the risk is (see *Environmental Risk and Hazards*). This 'social processing of risk' includes information flow; the communication channels and social networks through which individuals, groups, and institutions interact; the formation of public perceptions and concerns; and the degree to which social capital and trust exist. Modern environmental groups play an important role in state response to environmental risk by targeting particular environmental threats for action, in calling such concerns to media attention, and by shaping public values and attitudes. Meanwhile, a global shift in public values has led some analysts (e.g., Inglehart 1990) to highlight the emergence of 'green ideologies' related to the growth of 'postmodern' values and lifestyle changes with economic development, and the effects of such shifts on heightened public concerns over environmental issues.

This societal processing function has been described in a 'social amplification of risk' framework (Fig. 1), in which social structures and processes 'amplify' or 'attenuate' signals to society concerning the seriousness and the manageability of the environmental risks (Kasperson et al. 1988). Since most of society learns about the flow of environmental change and risk events through information systems rather than through direct personal experience, risk communicators, and especially the mass media, are major agents (or what are termed 'social stations') of risk amplification and attenuation. Particularly important in shaping group and individual views of risk are the extent and content of media coverage, and particularly the way the environmental problem is framed. Information about environmental change flows through multiple communication networks—the mass media represented by television and newsprint, the more specialized media of particular professions and interests (including, increasingly, the Internet or the information superhighway), and, finally, the more informal personal networks of friends and neighbors. The mass media cover risks selectively, emphasizing those that are rare and dramatic while downplaying, or attenuating, more commonplace but often more serious risks, such as soil erosion or climate change. Environmental risks and risk events, thus, compete for scarce space in the media's coverage, and the outcome of this competition is a major determinant of whether a risk will be socially amplified or attenuated.

In postindustrial states, large organizations—multinational corporations, business associations, and government agencies—mostly set the contexts and terms of society's debate about environmental problems. The President's Commission on the Accident at Three Mile Island (1979) in the USA for example, concluded that the 'mind set' that permeated the institutions charged with managing nuclear safety represented the primary problem in ensuring the safety of the nuclear technology used to produce electricity. Breakdowns in internal organizational communications can also contribute to attenuation of risk, as occurred in the space shuttle *Challenger* accident, when the risk concerns of technical experts failed to reach top decision makers within the US National Aeronautics and Space Administration. Large corporations develop markedly different kinds of organizational cultures that influence their ability to assess the environmental effects of their activities and products. The behavior and interactions of institutions and organizations are major sites of risk amplification and attenuation and require detailed attention in gauging how different countries respond to similar environmental challenges.

Apart from the array of societal processes involved in the social amplification of risk, states also use a variety of means to inform themselves of potential or emerging environmental hazards (see *Environmental Risk and Hazards*). *Monitoring* implies an active and

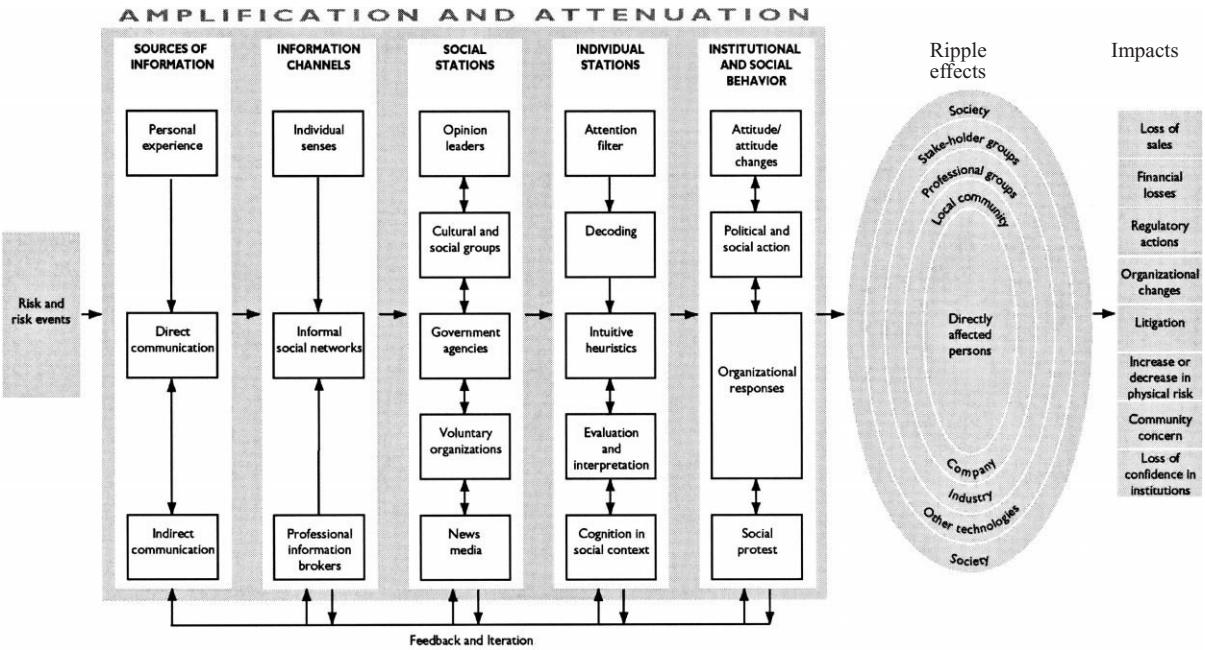

Figure 1
The social amplification and attenuation of risk

often systematic search by institutions or individuals for information or signals of environmental threats. *Alerting* refers to the warning function associated with the identification and assessment of risk. Both functions play key roles in identifying social problems and initiating societal responses. The perception of an environmental hazard begins with the first sense of danger but changes as increased information develops. The identification of risks is tied closely to social perceptions and occurs in a variety of arenas within the state—scientific research, the mass media, environmental groups, and the dissemination of industrial products. Global environmental risks, it should be noted, pose new challenges to traditional monitoring and alerting systems (see *Global Environmental Change: Human Dimensions*). High scientific uncertainty and political stakes are often present in such issues, and past experience is often unreliable as a guide to future risks at the global scale. And so societal alerting and assessment systems assume a larger role and burden in threat anticipation.

State alerting systems for environmental threats range over a wide diversity of institutions, roles, and contexts. Knowledge of environmental change does not always arise from purposeful monitoring of the environment but often occurs by serendipity or by chance. In many cases, the motivation for monitoring is aimed at the realization of some other societal goal, such as health promotion, economic growth and development, or policy effects. Additionally, particular institutions or social groups may experience a change that suggests a departure or discontinuity not previously experienced. Mountain dwellers of Nepal, for example, learn of environmental change through a wide array of experience—increased mudslides, declining crop yields, and growing scarcity of fuelwood, whereas the Nepali government learns about it through accounts of local crises trickling up from remote areas or through reports of failures of development or aid programs (Jodha 1995).

With the initial assessment that an environmental problem exists, it may become a policy issue (see *Environmental Policy*). How the issue is conceived and socially constructed is critical; political ideology and cultural values are major factors influencing this social construction. The power structure of society directly influences this process, of course, and particular values or ideologies may be dominant in the 'framing' of the issue. Authorities of the state, for example, have constructed deforestation in Borneo as a problem of 'shifting cultivation,' whereas deforestation and associated fires in Amazonia have been seen as the inevitable price of either frontier development or hamburgers for US markets (Kasperson et al. 1995). If the environmental problem is constructed as a concern only to those most directly involved with the natural resource utilization process (the manager closest to the environmental source), adjustments may be largely cosmetic or of the 'hidden hand' variety. If, however, a problem achieves sufficient salience as to require

concerted collective action, and comes together with other political confluences needed for policy action, interventions to ameliorate the risk may be undertaken. Whether policy actions or interventions occur will depend upon whether 'policy windows' exist, that is, where opportunities for pushing pet proposals or a favored conception of the problem are available (Kingdon 1995). These interventions may be particularly likely at 'critical junctures' where three different streams—problems, policies, and politics—which are ordinarily largely independent of one another come together (see *Environmental Policy*).

Various management strategies are available to state responses, including such options as bearing the risk, preventing or reducing the risk, initiating precautionary actions, spreading or reallocating the risk, fostering greater adaptiveness and resilience (see *Environmental Adaptation and Adjustments*), or ameliorating the consequences after they have registered their initial impacts. Among these risk-response strategies, *avoidance* refers to action to evade the problem or exposure to it (as in relocation from threatened areas); *consequence mitigation* refers to ameliorating the impacts (as through disaster aid or food relief); *bearing the risk* means deciding to absorb the stress and impacts caused by the environmental degradation; *risk spreading* refers to strategies, such as insurance, that reallocate the risk over larger populations without reducing the risk itself; and *adaptation* involves strategies by which society changes itself in order to reduce further or future impacts of the risk. Each of these approaches typically makes use of mechanisms appropriate to particular political cultures, such as taxes, regulations, marketable allowances, liability or legal resources, technology-based approaches, social and economic incentives, and dissemination of information (Rayner and Malone 1998). Structural changes—those responses that seek to address directly the basic nature of the driving forces—may also occur but such intervention is more fundamental and politically costly. Technology development, redistribution of land, wealth, and power, and fundamental reorientation of social and economic priorities are examples of such state responses. Still other far-reaching relevant state responses include enhanced status for women, family planning, debt cancellation, land reform, increased support for the traditional rural sector, and inclusion of environmental costs in national accounting systems.

Substantial differences exist among countries in the nature of responses to environmental change (see *Environmental Policy*). Vogel (1990) argues that environmental concerns in Japan have focused on public health, attention in Germany has centered more on the protection of nature, and the USA and Britain have occupied intermediate positions between Japan and Germany. Variations in national political cultures, the role of nongovernmental organizations, and the power of business in the policy process all contribute to these differences. Different countries also reveal substantial variations in regulatory styles, with the USA displaying an adversarial regulatory style, with demanding and specific quantitative standards, whereas Britain and many other European countries employ more flexible standards and greater discretion in enforcement (Stern et al. 1992).

Global environmental risks may pose distinctive challenges for the state management systems, institutions, and approaches that deal with them. The effectiveness of institutions, it has been argued (Young 1999), depends upon the match between the characteristics of the institutions and the nature of the environmental systems with which they interact. Institutional failures may arise because of mismatches between environmental processes and institutional designs and capabilities. Environmental management often strives for creating relatively high scientific certainty and anticipation, effective 'rational' choice among options, and a 'command-and-control approach to management. Increased attention is being given to 'adaptive management' approaches that may be more robust, resilient, and effective over time with many types of transboundary and global risks (Lee 1993; see also *Environmental Adaptation and Adjustments*). Such a management style proceeds from the recognition that environmental risks often cannot be comprehensively assessed, that large uncertainties typically characterize major elements of the risk problem, and that surprises are certain to occur. Another recent response that has gained favor, particularly in Europe, has been the growing invocation of the precautionary principle as an approach to managing uncertain environmental risks.

Bibliography

Inglehart R 1990 *Culture Shift in Advanced Industrial Society*. Princeton University Press, Princeton, NJ

Jodha N S 1995 The Nepal middle mountains. In: Kasperson J X, Kasperson R E, Turner B L II (eds.) *Regions at Risk: Comparisons of Threatened Environments*. United Nations University Press, Tokyo, pp. 140–85

Kasperson J X, Kasperson R E, Turner B L II (eds.) 1995 *Regions at Risk: Comparisons of Threatened Environments*. United Nations University Press, Tokyo

Kasperson R E, Renn O, Slovic P, Brown H, Emel J, Goble R L, Kasperson J X, Ratick S J 1988 The social amplification of risk: A conceptual framework. *Risk Analysis* **8**: 177–87

Kingdon J W 1995 *Agendas, Alternatives, and Public Policies*, 2nd edn. HarperCollins, New York

Lee K 1993 *Campass and Gyroscope: Integrating Science and Politics for the Environment*. Island Press, Washington, DC

Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK

President's Commission on the Accident at Three Mile Island 1979 *The Need for Change: The Legacy of TMI*. Government Printing Office, Washington, DC

Rayner S, Malone E (eds.) 1998 *Human Choice and Climate Change, Vol. 4: What Have We Learned?* Battelle Press, Columbus, OH

Stern P C, Young O R, Druckman D (eds.) 1992 *Global Environmental Change: Understanding the Human Dimensions*. National Academy Press, Washington, DC

Vogel D 1990 *National Styles of Regulation: Environmental Policy in Great Britain and the United States*. Cornell University Press, Ithaca, NY

Young O R 1999 *Institutional Dimensions of Global Environmental Change: IDGEC Science Plan*. IHDP Report No. 9. International Human Dimensions Programme on Global Environmental Change (IHDP), Bonn, Germany

R. E. Kasperson

Environmental Cognition, Perceptions, and Attitudes

Individuals' primary interaction with the natural environment, as with other aspects of their experience, is through their senses. They form feelings and impressions that both shape and are shaped by their basic cognitive structures and values. At deeper levels of processing, they form more explicit beliefs and attitudes, even theories and ideologies. These cognitions, perceptions, and attitudes affect both individuals' own well being and that of their environment. People feel better in a healthy natural environment and feel better knowing that their natural environment is healthy. The health of that environment depends on the choices that they make, as consumers, workers, citizens, officials, tourists, and industrialists (Garner and Stern 1996, Kaplan and Kaplan 1989).

At the present time, environmental protection enjoys broad support. It is expressed in public opinion polls (e.g., Dunlap et al. 2000), in the popularity of nature programming on television, and in the growth of 'socially responsible' mutual funds, which control 13 percent of the US market and typically screen firms for their environmental behavior. However, the fate of the environment depends on the actual support (or opposition or indifference) emerging in specific decisions. These choices are enormously varied, in their options, consequences, and decision-makers—not to mention their national, social, cultural, institutional, and economic settings.

The diversity and complexity of these decisions poses a challenge to researchers. One response is performing case studies, fully respecting the particularities of each situation. Environmental studies scholars have produced many such accounts (Slovic and Dixon 1992). Their observations shape and constrain the theories of environmental scientists, seeking general, testable theories about environmental behavior. These scientists have typically adopted one of the two classic compromises for producing general accounts of heterogeneous phenomena. (a) *Holistic* approaches look for recurrent behavioral archetypes, groups of people who respond in predictable, but different ways (e.g., Cotgrove 1979, Douglas and Wildavsky 1982). For example, one might posit 'eco-freaks' who resolutely oppose any intrusion into pristine areas or 'technocrats' who automatically dismiss such concerns. (b) *Reductionist* approaches assume that all individuals respond to common factors, but interpret and weight them differently—sometimes leading to different actions (Stern 1992). In them, eco-freaks and technocrats (and everyone else) would attend to similar outcomes and processes. However, they might view them very differently, in terms of their attitudes and beliefs. Both kinds of account coexist in dual-process theories of attitude and cognition. In them, holistic first impressions create an orientation toward a stimulus (e.g., an environ mental cause or insult). A parallel process involves analytical thinking, elaborating that impression. The result might be interpreting it in terms of more detailed (holistic) archetypes or (reductionist) cross-cutting concerns. Such analysis can overturn first impressions. However, it can also be constrained by them. Many common cognitive processes encourage the perseveration of beliefs. For example, people tend to seek confirmatory, rather than contradictory evidence; to resolve ambiguity in favor of current beliefs; to exaggerate the definitiveness of small samples of evidence (Gilovich et al. in press). How hard people work to understand an issue should depend on how important it seems, what returns they expect from their efforts to learn, and how great is their general need for cognition. Those factors, in turn, reflect and shape the opportunities that they have to affect their environment.

1. *Holistic Accounts*

A general account of environmental behavior is only useful if researchers can unambiguously translate complex real-world situations into its necessarily abstract terms. For holistic theories, that translation process may be holistic itself, attempting to identify specific cases with theoretical archetypes. Holistic interpretation allows astute observers to use the full set of available data. Moreover, if those observers write well, the coherence of their accounts should help readers to retain and integrate the information that is presented. Archetypes are a natural way to organize evidence, even if that sometimes leads to neglecting discordant details.

Unfortunately, holistic interpretation also allows motivated observers to exploit the richness of the evidence, in order to fashion desired accounts. For

example, they might exploit the privileged position of hindsight, so as to craft interpretations that affirm their theories. Or, they might strategically identify other people with positive or negative archetypes, in order to advance a political agenda. The very act of categorizing others is an exercise of power. It can reduce those individuals to stereotypes and present the observer as possessing superior insight (and perhaps even morals). It can feed prejudices and benefit from the natural tendency to create simplistic views of other people.

These possibilities are not lost on the partisans in environmental conflicts, who value the opportunity to demean their opponents and flatter their allies. As a result, holistic characterizations are often offered and evaluated by individuals without scientific credentials. Conversely, scientists can have a political following because of the conclusions they reach, independent of the quality of their work. Pundits can drape themselves in the trappings of academia, seeking its rights, but not responsibilities, in order to achieve political ends. The narrative character of holistic accounts means that both professional and amateur ones have the same outward form. Commentators may have insights to offer. However, they are in a different business, without the scholarly constraints of sharing sources and reflecting previous research. Thus, readers need to check the scientific pedigree of environmental narratives. What training did the authors have? What peer-review screens did their work pass?

With ongoing environmental issues, more systematic measurement of holistic archetypes might be possible. If the archetypes are clearly defined, it should be possible to operationalize them in an individual difference measure (e.g., Dake 1992). If these measures pass conventional psychometric criteria (e.g., reliability, construct validity), then they allow objective tests of their theory (e.g., '"corporate greens" have the deepest commitment to consumption-related issues'; 'environmental pessimists have the greatest scientific knowledge and institutional distrust'). Of course, this strategy also requires similarly systematic characterization of the other elements of the prediction: what the environmental issues are, how those individuals perceive them, what their observable behavioral responses are, what meaning they attach to those actions, and so on.

If archetypes cannot be measured authoritatively, then individuals may be characterized by more observable categories, such as gender, race, income, or nationality. The price paid for the precision of such characterizations is losing the theoretical force of a well-crafted archetype. For example, surveys in the US have found particularly high environmental concern among African–Americans (Vaughan 1993). Speculatively, that concern could reflect greater direct exposure to environmental problems (so that they have more reasons for concern), greater awareness of their immediate surroundings, greater physiological sensitivity (e.g., because of multiple exposures or weakened immune systems), greater suspicion of societal institutions, fewer resources for self-protection, and so on. An analogous array of hypotheses follows from Slovic's (1999) finding of similar attitudes toward the environment across most US populations. The one exception was white males who, on average, were more favorable toward technology and less concerned about the environment. Characterizing the subset of white males responsible for this overall group difference is a starting point for a theory of archetypes, sorting out the experiences or circumstances creating these distinctive views.

1.1 Reductionist Approaches

Comparable measurement issues confront reductionist approaches, looking at how people respond to a common set of concerns, as expressed in different environmental issues. In these models, individuals evaluate an action in terms of its expected effects on the probability of occurrence for each of a set of consequences. Those consequences might be experienced personally (e.g., changes in taxes, wilderness opportunities, social standing), socially (e.g., changes in health, justice, property rights), or environmentally (e.g., changes in biodiversity, the introduction of novel life forms). An action is attractive to the extent that it increases the chances of good things happening and decreases the chances of bad ones. These models belong to the general class of rational-actor models. They preserve the basic logic of orderly decision-making found in neoclassical economics, but vary in how strongly they endorse its normative and optimality assumptions.

The theoretical challenge for researchers is identifying a concise set of consequences, with important impacts in multiple situations. If that set exists, then general statements are possible, regarding 'what matters to people, when it comes to the environment.' Although an appealing aspiration, such sweeping summaries are limited by the actual relevance of the issues. For example, social norms may loom very large for curbside recycling, where all one's neighbors can see what one does. Those norms may have little effect on roadside littering when no one is looking—unless the norms have been internalized as a personal value or the action (not littering) has become a matter of habit.

Thus, differences in the importance of issues can reflect (earnestly considered) differences in situations, and not just (hypocritical) situational ethics. As a result, some of the most successful attempts to predict behavior from beliefs and attitudes have focused on specific domains, such as home energy conservation (Stern 1992). These studies find that conservation increases when people see greater benefits and fewer barriers. These patterns can be sought in everyday

circumstances or in the context of interventions intended to encourage such behaviors (e.g., recycling or political campaigns). Good communication should improve citizens' understanding of the issues, making their actions more consistent with their values (and more predictable).

Like holistic approaches, these models face the challenge of assessing beliefs and values, for the people whose behavior they hope to predict. Even a valid model can fail empirical tests, if its terms are measured poorly. One form of protection for reductionist models is provided by their additive structure: they sum the impacts of favorable and unfavorable factors, allowing good and bad expectations to cancel one another. Such models are relatively insensitive to errors in measurement and specification (Dawes 1988). As a result, they will have some predictive validity as long as they include rough approximations of the relevant factors—or factors correlated with them. Unfortunately, such robustness inhibits theoretical development. It is hard to interpret the weights estimated for the factors in the model or to know whether they are actual drivers of behavior or just surrogates. An important exception to this rule arises when one consideration overrides all others, as something that must or must not be done. Such a decision model is noncompensatory, in the sense that there is no way to overcome the stigma associated with an action—or to diminish the moral high ground that it confers (Flynn et al. 2001).

2. *Perceived Risks and Benefits of Environmental Changes*

Assessing beliefs and values is a central activity for cognitive and social psychology. Assessing them for environmental issues has posed significant challenges to the state of those sciences. Environmental changes can involve novel, unintuitive processes, surrounded by uncertainty and couched in unfamiliar terminology. They may extend over time, individuals, and species, in ways that are hard to grasp, much less evaluate. They pose choices as esoteric and fateful as those arising from modern medicine (e.g., advanced directives regarding life support, genetic screening, rationing scarce resources among competing needs). The environment suffers directly when people make poor choices. It suffers indirectly when investigators mismeasure what people want regarding the environment—hence send inaccurate signals to policy makers and citizens. These practical concerns have encouraged environmental social scientists to break new ground in these traditional assessment activities.

The difficulties that face investigators, in posing these questions, echo the difficulties that face individuals, when life poses them. As a result, the research agenda has combined descriptive and prescriptive goals, trying to understand current behavior in order to facilitate better informed behavior in the future. Those improvements are sought both in beliefs and values. People need to understand how the environment works and how changes can affect their lives (health, emotions, children, etc.). They need to understand themselves well enough to know what values to assign to these impacts.

As a result of these applied concerns, studies of beliefs often have an evaluative framing. They ask, not just what do people know, but whether people know enough to make effective decisions. When people face specific decisions, the adequacy of their knowledge can be evaluated relatively precisely—especially compared to general tests of environmental knowledge, which generally have no analytical foundation for selecting items. A decision-making perspective allows specifying what people need to know, and not just what it would be nice to know.

In the context of values, the evaluative framing asks whether people have succeeded in constructing stable opinions about specific environmental choices, consistent their relevant basic values. It recognizes that, even when a topic is familiar (e.g., declining fish populations, invasive species, suburban sprawl, vector-borne disease), the specific choice may not be. As a result, individuals cannot just 'read off' a response, from a universal utility or value function. In effect, they may not know what they want. In such cases, researchers need to decide whether to take people as they are, with fragmentary positions, or help them to derive more reflective ones. The former strategy risks a sin of omission if respondents are swept along by the incomplete set of issues evoked by a particular question. The latter strategy risks a sin of commission, if interacting with respondents distorts, rather than deepens, their understanding.

These questions of competence can be asked regarding both citizens and technical specialists. The answers have implications for the roles that these individuals should play in environmental decision making. When citizens lack the desired competence, a case can be made for rule by experts (or for intensive education). When even the experts don't know very much, a case can be made for moving slowly (unless that allows even more poorly understood processes to go on unchecked). The current debate over genetically modified organisms revolves on whether anyone understands them well enough to make sound choices. Those who doubt that even the experts know enough to proceed often invoke the 'precautionary principle.' At times, it is just a (holistic) slogan of distrust. At times, it is a request (or demand) for incorporating deeper forms of uncertainty into cost, risk, and benefit analyses. These concerns pose significant challenges to the measurement and characterization of expert knowledge (Funtowicz and Ravetz 1990, Kammen and Hassendzahl 1999).

The emerging science of belief and value assessment draws on two converging research traditions. One

leads from psychophysics to attitude research. The second leads from decision theory to behavioral decision theory.

2.1 Psychophysical Measurement

Emerging from the natural sciences, many early psychologists focused on what came to be called psychophysics, determining the psychological equivalent of physical stimuli. The conceptualization underlying these studies holds that a combination of physiological and psychological mechanisms translates those stimuli into an internal state of arousal. If asked, individuals can report on that state, either verbally (by word or number) or by some other action (e.g., pressing on a handgrip, adjusting a rheostat to show the light intensity equivalent to the loudness of a tone).

The most straightforward environmental applications of the psychophysical approach ask people how they feel about a specific environment (e.g., how natural, attractive, rich…is it?). They might be asked to use direct ratings (Daniel and Vining 1983) or to make similarity judgments, which are then subjected to multi-dimensional scaling (Gärling 1976). Alternatively, they might be asked to sort pictures that vary in the roles played by natural and built objects (Cantor 1977). Using pictures frees respondents from having to encode verbally what they see and like. As a result, it may capture attitudes that people have difficulty expressing, because they lack the words or hesitate to use them (e.g., because of social unacceptability or uncertainty about terminology). One set of properties, identified in such studies, is coherence, complexity, legibility, and mystery (Kaplan and Kaplan 1989).

Psychophysical studies create a significant interpretative burden for investigators, who must decode the meaning for respondents of both stimulus and response (Fischhoff and Furby 1988). Just what features did respondents notice? What did they overlook? What did respondents believe about the services that the focal environment provides and the peril it faces? Did they detect hints (intended or inadvertent), regarding what to say? Would they respond similarly in less hypothetical circumstances? It takes a suite of converging studies, with suitable manipulation checks, to establish what people think they are evaluating and how deep their understanding is. Such interpretative issues have been at the center of the continuing controversy regarding contingent valuation, a method advanced by some resource economists for monetizing goods not traded in any conventional marketplace (e.g., the survival of an endangered species, the improvement of atmospheric visibility) (Mitchell and Carson 1989). In the attitude measurement tradition, interviewers ask respondents how much they are willing to pay in order to prevent an adverse environmental change or to accept as compensation for it. If valid, the responses to these questions could be used in cost-benefit analyses. In that role, they would represent environmental changes that would otherwise be neglected in policy making dominated by economic analyses. However, these are intellectually demanding questions. By definition, they require explicit, quantitative evaluations of complex, novel stimuli with many potentially relevant details.

Attitudes have both strengths and weaknesses as anchors for evaluating environmental goods. The ease with which they are evoked means that individuals typically have some feelings to draw upon, whatever question they face. On the other hand, such initial feelings can be hard to overcome when more analytical responses are required (e.g., willingness to pay for a specific good) (Kahneman et al. 1999). Strong feelings can blur the details of serious environmental questions. Debates over the validity of contingent valuation have focused on whether respondents are properly sensitive to the amount (or 'scope') of the environmental change being evaluated (Arrow et al. 1993). The death of migratory birds saddens many people. Envisioning many dead birds creates more sadness than the death of few. The challenge is to translate degrees of sadness into appropriate monetary equivalents. Doing so may require intensive, interactive processes, which depart from the nonreactive norm of attitude research.

Even if such elicitation processes are successful, questions can be asked about the appropriateness of their goal. At the extreme, monetization assumes that everything of value has an economic equivalent. A more modest aspiration is to assign economic values wherever that is possible, but without assuming that they are everything. However, one might still worry that the focus on economics 'anaesthetizes moral feeling' (Tribe 1972). Furthermore, it would take a very disciplined process not to let monetized effects swamp diverse other ones.

Despite the residual ambiguity of their results, these studies place important constraints on studies using other procedures. Methodologically, features that demonstrate a robust influence in attitude studies should be explicitly represented in the stimuli used with other methods. Otherwise, respondents need to guess about the features that are missing, based on those that are present. Substantively, the basic preferences reflected in attitude studies should be observed in studies using in other methods. If not, then doubts must be raised about the robustness of the preferences (or the methods).

In the attitude research tradition, evaluation focuses on a single option. Formally speaking, those evaluations could reflect the alternatives as well, in terms of the opportunity costs of forgoing those other options. However, in cognitive terms, the nature and consequences of those alternative actions will inevitably be less clear than those of the focal one. With many options, the alternatives may not even be fully

enumerated. However, even when there are but two complementary options (e.g., develop/don't develop), thinking about one need not call the other to mind. As a result, each option might evoke different concerns. Asking about it alone does not allow individuals to express their full suite of relevant values.

2.2 Decision Theoretic Measurement

The early days of the modern environmental movement presented many technology managers with opposition that they could not, or would not, understand. One natural response was that their opponents were misinformed. The initial research response was to determine how well people understood the magnitude of different risks, conceptualized as annual fatality rates. The research found (a) a strong correlation between lay and statistical estimates; (b) systematic deviations from the best-fit curve, partially attributable to the relative *availability* (in memory) of examples of the different causes of death; (c) considerable ordinal (and even ratio–scale) consistency in judgments across response mode (e.g., fatalities vs. relative death rates); and (d) considerable inconsistency in absolute estimates, across response modes. Thus, respondents seemed to have a fairly robust intuitive feeling for relative risk levels, which emerged however these (unusual) questions were asked (Lichtenstein et al. 1978).

However, rather different judgments emerged when respondents judged 'risk,' rather 'average year fatalities.' Moreover, the risk judgments were better predictors of their attitudes toward technologies (Slovic et al. 1979). Many studies have sought the determinants of subjective definitions of 'risk.' One early candidate was 'catastrophic potential,' in the sense that people see greater risk in technologies capable of claiming many lives at a single time. Anecdotal support for this hypothesis can be found in the attention drawn to plane crashes and public concern over nuclear power (despite citizens' recognition that, in an average year, the death rate from nuclear power is negligible). Subsequent research suggested, however, that catastrophic potential was not as significant as the uncertainty associated with technologies that can produce large accidents (Slovic et al. 1984). Avoiding uncertain technologies reflects risk aversion, rather than an attitude toward how deaths are distributed over time. These are different ethical stances, with different policy implications. Concern over catastrophic potential means providing incentives for locating technologies remotely and disincentives for creating technologies that could, conceivably, affect large numbers of people. Concern over uncertainty means providing incentives for research that sharpens risk estimates and disincentives for highly innovative technologies (which can't be known that well).

The search for defining attributes of 'risk' can be traced to Starr's seminal (1969) claim that, for a given level of benefit, the public tolerates higher levels of risk when they are incurred voluntarily (e.g., skiing), rather than imposed involuntarily. As evidence, he presented aggregate estimates of annual deaths and economic benefits from eight sources. In his plot, he sketched in two parallel 'acceptable risk' lines, an order of magnitude apart, for the voluntary and involuntary risks. This sort of *revealed preference* analysis assumes that society manages to achieve acceptable risk-benefit tradeoffs from technologies. If so, then that tradeoff should be seen in lay perceptions of risks and benefits (which drive political processes).

A study testing this hypothesis asked citizens to judge 30 technologies in terms of current risks and benefits, acceptable risks (given the benefits), voluntariness, and eight other features (e.g., catastrophic potential, dread, controllability, known to science, known to the public) (Fischhoff et al. 1978). It found:

(a) discrepancies between lay and statistical estimates of risk and benefit;

(b) a weak correlation between estimates of current risks and benefits (so society was not seen as exacting greater benefit from riskier technologies);

(c) no increase in the risk-benefit correlation, when mean judgments of the risk attributes were partialed out (indicating that respondents did not see society as having set a double standard for, say, voluntary and involuntary technologies);

(d) a belief that the acceptable level of risk for most (but not all) technologies was lower than respondents' current estimate (rejecting the hypothesis of effective societal risk management);

(e) a significant correlation between judgments of current benefits and acceptable risk levels (indicating a willingness to have more risk with greater benefit); and

(f) a large increase in predicting acceptable risk levels from perceived benefits, when many of the risk attributes were partialed out (suggesting the desire for a double standard, accepting greater risk when, say, it is incurred voluntarily).

3. Prescriptive Applications of Cognitive Research

3.1 Setting Public Policy

If robust, such results provide guidance to public policies affecting the environment. At the simplest level, they show a willingness to accept risk-benefit tradeoffs (possibly surprising some critics of lay rationality). They also show attributes of risk that might be considered when designing technologies or regulatory mechanisms. Different definitions of risk can lead to different social priorities. That is true even for different ways of measuring a single attribute. For example, mortality risk can be summarized as prob-

ability of premature death or as lost days of life expectancy. The latter measure puts a special premium on deaths of young people (where many days are lost with each life). With the former measure, a death is a death (Crouch and Wilson 1981, Fischhoff et al. 1981). Either could consider the distribution of risk across the population (e.g., among those who do and do not benefit from a technology), or leave equity considerations to others. Technical specialists often follow professional conventions when computing their estimates. Nonetheless, those definitions still incorporate ethical considerations. Whether citizens (or their representatives) are consulted is a reflection on accepted political processes.

In cognitive and administrative terms, dealing with a large set of attributes is unwieldy (and, at some point, intractable). As a result, many studies have looked for redundancy among risk attributes. Using various data-reduction techniques, typically they have found that much of the variance in ratings of the attributes can be accounted for by two factors. One seems to reflect how well risks were known, by science and by those exposed to it. The second seems to reflect how much risk is dreaded, inequitable, involuntary, and personally threatening.

If such issues are on citizens' minds, democratic societies need ways to incorporate them in public policies, lest trust be eroded. Institutionalizing analytical procedures that incorporate public values (e.g., definitions of risk) is one possibility. Processes that directly involve citizens have increasingly been advocated, with environmental issues both driving and reflecting societal pressure for more consultative government and industry (e.g., Canadian Standards Association 1997, UK Royal Commission on Environmental Pollution 1998). The US Environmental Protection Agency (1993) fostered some 50 state and regional efforts to set priorities among risks. These panels' ability to structure their work as they saw fit, and to receive clear explanation of technical matters, seemed critical to their success (Davies 1996). Nonetheless, it was unclear what features any particular panel had considered, nor what weight it gave to each feature. Without such transparency, the credibility of its conclusions could lack credibility, for individuals who were not party to the deliberations. Furthermore, skeptics might wonder whether participants really understood the environmental risks that they were judging.

3.2 Ensuring Citizen Competence

Thus, cognitive approaches to preference elicitation draw upon two of the major streams in psychology, as it has developed over the past century-plus. The psychophysical stream envisions values as emerging from direct perceptual processes; these visceral responses, then, require intellectual effort, in order to translate them into terms that address specific policy questions. The decision theory stream envisions specific values as being derived intellectually from basic values; these, then, must be connected with individuals' life experiences, lest they be but transient artifacts of the elicitation procedure.

Both streams recognize the following steps that individuals must traverse, if they are to formulate meaningful positions—with either researchers or policy makers as their audience:

(a) Encode the task—in order to understand the choices being offered.

(b) Access relevant personal values—in order to structure the evaluation process.

(c) Integrate the relevant values, as interpreted in the specific context—in order to construct a preference.

(d) Translate that preference into terms recognizable by the investigator or policy maker.

The first step to ensuring such shared understanding is creating a full task specification, characterizing all features that matter to investigators or respondents (e.g., Fischhoff and Furby 1988). How hard investigators have to work to convey these task definitions depends on how far they share a common frame of reference with respondents. For example, as the vote approaches on a widely debated referendum, many citizens know what commitment it implies and can respond reliably to a poll eliciting their values. When that occurs, many details 'go without saying.' Omitting them both saves time for less obvious details and respects conversational norms.

Communicating the remaining details has been approached from both a bottom-up and a top-down perspective. The former involves creating a supply curve for facts, focusing presentations on the facts that respondents most need to know, but currently do not. If successful, it allows using recipients' attention most wisely, as well as evaluating the adequacy of the understanding that they achieve. Sometimes, only a few things really matter. Sometimes, many do. Top-down communication takes advantage of people's well-known ability to absorb information more quickly, when they can organize it into *chunks*, clusters of facts cognitively processed as units. Mnemonists have mastered integrating diverse facts into highly flexible templates. Ordinary people tend to organize information spontaneously into less coherent, domain-specific *mental models* (Johnson-Laird 1983). Invoking one takes advantage of existing beliefs, allowing some task features to go without saying, while making others easier to absorb. However, mental models can also encourage unintended inferences, while hindering the processing of unexpected features. Like other metaphors, they require some interpretation, and may evoke varying images. Thus, here as elsewhere, empirical research is needed to design the best communications, then to evaluate their success (Fischhoff 1999).

4. Conclusion

Environmental decisions are enormously diverse. Individuals can have a hard time grasping the issues involved and deciding how they feel about them. These difficulties are part of the task of researchers, hoping to characterize or facilitate environmental decision making. Driven by the importance of these problems, environmental social scientists have cast a wide net, for potentially relevant methods and theories—extending and integrating them along the way. They make it possible to go beyond speculation and anecdote, regarding what people believe and want about specific environmental issues. Most people report caring about the environment. Its fate and theirs depend on how well they understand what is happening with it and convey their desires for its future.

See also: Ecological Inference; Environmental Psychology: Overview; Environmental Risk and Hazards; Environmental Surprise; Psychophysics; Risk, Sociological Study of; Spatial Thinking in the Social Sciences, History of

Bibliography

Arrow K, Solow R, Portney P, Leamer E, Radner R, Schuman H 1993 Report of the NOAA panel on contingent valuation. *Federal Register* **58**: 4601–14

Canadian Standards Association 1997 *Risk Management (CSA-850)*. Canadian Standards Association, Ottawa

Cantor D V 1977 *The Psychology of Place*. Architectural Press, London

Cotgrove S F 1979 *Catastrophe or Cornucopia*. Wiley, New York

Crouch E A C, Wilson R 1982 *Risk/Benefit Analysis*. Ballinger, Cambridge, MA

Dake K 1992 Myths of nature—culture and the social construction of risk. *Journal of Social Issues* **48**(4): 21–37

Daniel T C, Vining J 1983 Methodological issues in the assessment of landscape quality. In: Altman I, Wohlwill J F (eds.) *Human Behavior and Environment*. Plenum Press, New York, Vol. 16, pp. 39–84

Davies J C (ed.) 1996 *Comparing Environmental Risks*. Resources for the Future, Washington, DC

Dawes R M 1988 *Rational Choice in an Uncertain World*. Harcourt Brace Jovanovich, San Diego, CA

Douglas M, Wildavsky A 1982 *Risk and Culture*. University of California Press, Berkeley, CA

Dunlap R, VanLiere K D, Mentig A G, Emmet Jones R 2000 Measuring endorsement of the new ecological paradigm. *Journal of Social Issues*. **56**(3): 425–42

Fischhoff B 1999 Why (cancer) risk communication can be hard. *Journal of the National Cancer Institute Monographs* **25**: 7–13

Fischhoff B, Furby L 1988 Measuring values: A conceptual framework for interpreting transactions. *Journal of Risk and Uncertainty* **1**: 147–84

Fischhoff B, Lichtenstein S, Slovic P, Derby S L, Keeney R L 1981 *Acceptable Risk*. Cambridge University Press, New York

Fischhoff B, Slovic P, Lichtenstein S, Read S, Combs B 1978 How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences* **8**: 127–52

Flynn J, Kunreuther H, Slovic P (eds.) 2001 *Risk, Media and Stigma*. Earthscan, London

Funtowicz S O, Ravetz J 1990 *Uncertainty and Quality in Science for Policy*. Kluwer, London

Gärling T 1976 Structural analysis of environmental perception and cognition—multidimensional scaling approach. *Environment and Behavior* **8**: 385–415

Garner G T, Stern P C 1996 *Environmental Problems and Human Behavior*. Allyn and Bacon, Boston

Gilovich T, Griffin D, Kahneman D (eds.) in press *The Psychology of Judgment: Heuristics and Biases*. Cambridge University Press, New York

Johnson-Laird P N 1983 *Mental Models*. Cambridge University Press, New York

Kahneman D, Ritov I, Schkade D 1999 Economic preferences or attitude expression? *Journal of Risk and Uncertainty* **19**: 203–35

Kammen D, Hassendzahl D 1999 *Should We Risk It?* Princeton University Press, Princeton, NJ

Kaplan R, Kaplan S 1989 *The Experience of Nature: A Psychological Perspective*. Cambridge University Press, Cambridge, UK

Lichtenstein S, Slovic P, Fischhoff B, Layman M, Combs B 1978 Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory* **4**: 551–78

Mitchell R C, Carson R T 1989 *Using Surveys to Value Public Goods*. Resources for the Future, Washington, DC

Royal Commission on Environmental Protection 1998 *Setting Environmental Standards*. Stationery Office Ltd., London

Slovic P 1999 Trust, emotion, sex, politics, and science. *Risk Analysis* **19**: 698–701

Slovic S, Dixon T 1992 *Being in the World: An Environmental Reader for Writers*. Macmillan, New York

Slovic P, Fischhoff B, Lichtenstein S 1979 Rating the risks. *Environment* **21**(4): 14–20, 36–9

Slovic P, Lichtenstein S, Fischhoff B 1984 Modeling the societal impact of fatal accidents. *Management Science* **30**: 464–74

Starr C 1969 Societal benefit versus technological risk. *Science* **165**: 1232–8

Stern P C 1992 Psychological dimensions of global environmental change. *Annual Review of Psychology* **43**: 269–302

Tribe L H 1972 Policy science: Analysis or ideology? *Philosophy and Public Affairs* **2**: 66–110

US Environmental Protection Agency 1993 *A Guidebook to Comparing Risks and Setting Environmental Priorities*. US Environmental Protection Agency, Washington, DC

Vaughan E 1993 Individual and cultural differences in adaptation to environmental risks. *American Psychologist* **48**: 673–80

B. Fischhoff

Environmental Conservation: Ethical Concerns

1. Major Debates on Ethical Systems

Should our environmental responsibility emerge from the concern for consequences of our action or from the

criteria that we chose to take those actions? There are several frameworks that have been used such as: Libertarianism, Contractualism, Relativism, Deontology, and Teleology, for analyzing ethical dilemmas.

Bentham (1748–1832) developed a simple universal principle for deciding what is good, calling it the principle of utility. According to this theory, the doer of any good act tries to maximize the amount of happiness to the doer and to others affected by it. Critics have argued that hedonic calculus could not be the basis of estimating moral worth of an action, arguing

> that because some pleasures are inherently better than others (getting an education, say, is better than getting drunk), they have inherent worth that makes them desirable, it throws in question that the very possibility of hedonic calculus follows; it would mean that pleasure and pain are no longer considered the basic unit in terms of which the worth of all acts and goals are to be measured (Goldberg, 1995, p. 117).

The boundary of feelings of those sentient beings whose pain affects us becomes part of our moral boundary.

What are the different ways in which ethical concerns are manifested when describing the environmental problems or prospects? One of the persistent issues, which emerges in the dialogue on ethics, is the distinction among facts and values. Goldberg states: 'facts concern the way things actually are, the evaluations are the judgments about things ideally as they ought to be' (1995, p.5). Amartya Sen, in a very significant contribution (1981), observed that what we choose to describe involves value judgments and how we choose to describe it also involved a value judgment. The descriptive 'is' becomes the normative 'ought'.

2. *Determinants of Domain of Responsibility*

2.1 *Boundary of Pain*

Does the degree of pain suffered by other species determine the boundary of our responsibility? Some have argued that predation is a rule in nature and thus why should an anthropomorphic view of nature be decried so much. But survival of the fittest is also a feature of a natural system. Will that be acceptable as a basis of human social evolution? Darwin is believed to have remarked that humans were not necessarily on the top of the ecological chain as some higher form than others. Each organism may have adapted to its niche and thus may have advantages over other species in that niche: no single species is higher or lower than another. Our responsibility for another species may emerge from the pain we cause to it. The other view, as we shall see later, is that not every species may experience pain. Jeremy Bentham (1748–1832) held that

> any being capable of suffering should have his or her experiences taken into account by utilitarian calculations. However, according to right theorists, any version of utilitarianism, no matter how carefully conceived it is, fails to provide sufficient protection for innocent life, human or non-human (Pluhar, 1998, p. 165).

Pluhar asks,

> Do animals have a *prima facie* right to life or a *prima facie* right not to be tortured? One might hold that some animals have no serious right but do have a right not to be tortured or one might hold that that they have a *prima facie* right to humane treatment that could be overridden by the need to preserve allegedly more morally significant lives. (Pluhar, 1998, p. 162).

But then this argument raises another dilemma, which is about hierarchy of moral responsibility. René Descartes (1596–1650) proposed 'beings that are not rational are incapable of suffering.' According to him, non-human animals are merely organic machines without consciousness, unlike humans, who allegedly are amalgams of material bodies and immaterial minds (souls). Though he claims that he does not deny that non-human animals are capable of sensation, he is denying that they can suffer.

2.2 *Responsibility for Conservation Arising out of Greater Human Purpose*

Dower asks,

> Should we care for the environment because other life forms in nature, or nature itself, have a value, moral status independent of our interests, or because it is in our own collective interests to protect it? (1998, p. 769)

The scope of moral responsibility needs to be defined and different cultures experiencing tremendous loss of environment are defining it implicitly by making false compromises between development and environment. It appears as if one could have the former without the latter (unless we genuinely believe that Western societies are indeed more developed than the rest, as they certainly offer more consumer choices to their citizens). Human purpose could be to prosper, without impoverishing other human and non-human sentient beings, but it could also be defined by various cultures in negation of certain rights of others. Can ethics be determined by voting on it? We do not think so. A great degree of consumerist culture survives in most democratic European societies and yet its sustainability is questionable. The democratic way of arriving at the human purpose does not make it more legitimate and moral.

Human purpose can be defined in minimalist terms. One cannot solve all environmental problems but one can certainly solve some. Should one not try to solve a few because others remain? At the same time, our concern extends not just to those we know, see, or recognize. Our responsibility is to society and the biosphere at large, hence the international agreements such as the Convention on Biological Diversity (1992).

2.2.1 Agricultural ethics. The responsibility towards long-term sustainability of land, biodiversity, and well-being of animals, has been contrasted with a desire to intensify agriculture by crossing the natural barriers among species by using biotechnology or other technologies. For some, biotechnological tools can lead to better environment, if these help in reducing or eliminating chemical pesticides. However, these tools can do the opposite by unleashing environmental risks. The biotechnological revolution involving incorporation of genes from one species into another has raised a great many ethical issues. The invocation of the precautionary principle itself has become contentious. If in doubt, this principle advises that we err on the side of conservation. Some have interpreted this to mean that no research need be done to explore biotechnological alternatives, even if these were to have the potential for solving some nutritional and food needs. Others argue that the issue of hunger is not one of production, but of distribution. Ethical dilemmas arise when we enable the hungry to obtain sufficient food (through the public distribution system) through environmentally destructive land use practices (such as cultivation on marginal lands). Whether biotechnology can help solve this problem is as much an issue of ethics as one of policy and institutions. For instance, it can reduce pressure to bring more land under cultivation by increasing productivity of existing land. Ethical dilemmas also emerge when:

(a) intellectual property rights are granted over life forms (such as the Harvard onco-mouse) or other organisms violating the sanctity of life, as well as granting monopoly to those who did not and can not create life (except for modifying it into the laboratory),

(b) risks are taken in releasing in the environment genes which, in their natural conditions, did not have the possibility of diffusing among species (e.g., through transgenics tolerant to herbicides),

(c) animals are treated with hormones or other reagents which increase their productivity, yet affect their well-being or shorten their life cycle, and

(d) solutions are not developed to grow crops in problem soils, such as alkaline or salt-affected soils, through biotechnological means.

2.2.2 Asymmetry of rights and responsibilities. Risks are involved not just when things are changed but also when they are left unchanged. It is this position justifying inertia that populist arguments have made their mainstay. While exploring biodiversity or associated knowledge systems, ethics of extraction assumes asymmetrical rights and responsibilities (Gupta, 1994, 1995, 1999). We never acknowledge the creativity and innovations as well as traditional knowledge systems of local communities and individuals conserving resources in our writings (the whole discipline of ethnobiology has been a testimony to this). We do not share benefits with them fairly, never share what we have learned from them and others in their language. The Honey Bee network set up in 1989 to document and disseminate grassroots creativity and green innovations is an exception (http://www.sristi.org/honeybee.html). We complain when the same communities and individuals are at times forced to follow environmentally unfriendly actions. Studies have shown that the regions of high biodiversity in the tropics are also inhabited by the poorest people, who have the lowest educational levels (despite exceptions), who have high emigration of males and consequently a high proportion of households headed or managed by women. Yet discourse on environmental ethics has seldom reflected on these systematic relationships.

2.3 Ecocentrism/Biocentrism/Deep Ecological Ethics

The rights of not just the animate but also the inanimate, of not just the human but equally other species, and of not just the born but also the unborn are articulated in a combination of biocentric, ecocentric, and deep ecological ethical perspectives. Ecocentric theorists include Kenneth Goodpaster, Lawrence Johnson, Holmes Rolston, Baird Callicott, Arnold Leopold, John Rodman, etc., who essentially argue for moral rights of all beings and ecosystems. A lake has a personality if endowed with values, in the same way as mountains and species. Munshi (1952), a famous writer, in his lecture entitled 'A Gospel of Dirty Hand' tried to link the soil with soul. He did not see any way we could understand the relationship between nature and human beings if we did not see the linkage between nutrient cycle, hydrological cycle, and local institutions. Of course, he contrasted the ethic in which nature was held supreme, with the local tribes who, overawed by nature, remained what some may call 'primitive.' Other tribes or social groups, which overpowered nature, vanished into oblivion because they crossed the limits of nature. He argued for moderation. Some ecologists argue, however, not just for moderation but for a hands-off policy. The

> recognition of moral status of non-sentient living things can thus be depicted as the next step in the history of moral development (Rawles, 1998, p. 276).

Pedersen (1993) distinguishes two strands of modernity drawing from Giddens' work (1990) which may have a bearing on the evolution of ecocentric ethics. He refers to separation of space from place and of time from ecological space. Such a conception of ecological space does not require cultural association with a locality. Callicott (1994) hopes that such an international environmental ethic can evolve which can be compatible with local vernacular cultural traditions linked to space and time in a very different manner. The evolution of global time and of global space may also lead to the emergence of a global social group without a particular living space or habitat to qualify for specific environmental value.

Biegart (1999) recalls the contrast of perspectives between Native Americans and European settlers about water (surface flow). He gave the example of the late Philip Deere of Oklahoma, medicine man of the Muskogee Nation, who termed rivers and streams as the veins of the world. Clogging them, one could say would clog not just the life in them but the life of humans as well. The sacredness of water in all such cultures indicates that by polluting waters we are also polluting the spirits that sustain these waters.

2.3.1 Accountability towards perfect strangers and other non-persons. Human needs cannot always take priority over the needs of nature and other living beings. How do we determine our accountability towards the future generation, which is unknown to us? The future generation is made up of 'perfect strangers'—i.e., who are not known and are unknowable. We do not hear the voice of the unborn. What should be the responsibility of the present generation to discern the needs and preferences of such sentient and non-sentient beings with whom we are unable to communicate? There needs to be some response using contemporary as well as traditional value systems (Gupta 1991).

The cultures of the world have evolved means of generating and monitoring responsibility towards other living beings. But this responsibility need not emanate only from human value systems. Goodpaster observes,

> to be worthy of moral respect, a unified system need not be composed of cells and body tissue: it might be composed of humans and non-human animals, plants and bacteria. (Rawles, 1998, p. 279).

Johnson argues that 'various beings other than individual organisms can meaningfully be said to have interests, and that these interests are morally significant.' The beings in question include species and ecosystems. Brennan critiques this position and suggests that the claim of Goodpaster

> rests upon a naïve and scientifically outmoded view of ecosystems and species, neither of which have the characteristics that Johnson has attributed to them.

Brennan disagrees with the claim that

> ecosystems have interests, because he takes this claim to presuppose a view of ecosystems as goal-directed that the scientific community has largely rejected. (Brennan, in Rawles 1998, p. 279).

2.4 Socio-psychological Roots of Ethical Consciousness: The Internal Values

Hill (1978) brings in a personal psychological aspect while identifying roots of ecological ethics. He adds,

> my psychological argument is that truly ethical behavior originates wholly from the healthy, unhurt, undistressed parts of individuals; and unethical behavior originates from the hurt part. If we want people to behave ethically, then we must provide environments that are supportive of the healthy part of individuals (Hill 1992, p. 11).

Stone (1987) argues that monist (one best way to resolve ethical dilemmas) arguments have to be tempered by moral pluralism. The latter implies that one looks at the ethical basis of not just the action choice, but also the motivations of the actors and the institutional context of both the actors and the actions. There could be other planes as well. Can we not use a universalistic ethical principle in one part of our life space and use pluralistic values in other parts? Innovations from the Honey Bee data base on grassroots innovations managed by the Society for Research and Initiatives for Sustainable Technologies and Institutions (SRISTI) seem to indicate segmentation of life space.

The institutional behavior is the one where internal commands replace external demands. One does a thing not because one is being supervised but because that is the right thing to do. Environmental ethics is at a crossroads. We are looking for new indicators that will generate internal responsibility for caring for nature across different cultural contexts and worldviews. But what constitutes nature and what determines whether responsibility for its care globally, regionally, and locally will invoke equally strong internal commands: these questions remain open. Legal and public policy instruments are evolving and seem to indicate increasing concern for environmental care. But calls for such concern in the current geopolitical context from Western countries immediately invites criticism from developing countries. They see this call from the West as the sign of new emerging environmental protectionism. They argue that Western societies accumulated wealth by destroying their environment, which gives rise to a

precipitous argument: that therefore developing societies also need to have the right to accumulate wealth by destroying their own environment. Moral discourse will have to take a center stage in each of these polarized polities.

2.4.1 Environmental perception and cognition. Dunlap et al. (1993) have provided some of the most striking evidence against the notion that concern for the environment stems from 'postmaterialist values' (Muller-Rommel 1989). In a worldwide survey of citizen concern for the health of the planet, a high level of environmental concern was found in developing as well as developed countries. Such a concern is a necessary condition, though is insufficient for taking effective action. Stern et al. (1995) looked into the factors that may influence these environmental concerns of the citizens. One of the important findings of the authors is that activation of personal stable values is possible through organized efforts which try to influence the values in the direction of conservation ethic or otherwise. A study done by Gupta looked at the profile of the green consumers (Gupta et al. 1997). It revealed four types of consumers: (a) active mobilisers, (b) populist mobilisers, (c) passive practitioners, and finally, (d) those who were indolent and indifferent. The fact that there was not much difference in the proportions for each category indicated the possibility that social values could gravitate to either side depending on the nature of effect, available information, and action of organized interest, would have on internal values.

Austin and Schill (1991) and Boyce (1995) have argued for taking into account the need for environmental justice while looking at the issues of environmental care. SRISTI (1993) has argued for the need to combine ethics, equity, excellence, efficiency, environment, and education.

Gandhi provided a thoughtful summary of environmental ethics when he said that the world had enough for everyone's need but not enough for everyone's greed. The concepts of *aparigrah* (not accumulating more than one needs), *ahimsa* (non-violence) and frugality developed by him provided a practical guide to responsible living.

In the emerging modern consciousness in which a human being is no more responsible for his or her individual moral space but for the whole world, Amato (1982) argues that claims of guilt and gratitude will ultimately make us humble and bring us into harmony with our collective destiny. The exploration of environmental ethics is thus a journey into an abyss of responsibility for self and society, for present and future generations, and for the human and the non-human sentient and non-sentient beings, things, places, and also for time. We need to extract a slice of our time from the womb of ever-forgiving nature, which is learning to forget forgiving. That is the real tragedy and also the challenge.

See also: Environmental Cognition, Perception, and Attitudes; Environmental Justice; Environmentalism: Philosophical Aspects; Environmental Planning; Environmental Policy; Environmentalism, Politics of; Natural Concepts, Psychology of

Bibliography

Amato II J A 1982 *Guilt and Prejudice: A Study of The Origins of Contemporary Conscience*. Greenwood Press, Connecticut, p. 218

Austin R, Schill M 1991 Black, brown, poor and poisoned: Minority grassroots environmentalism and the quest for eco-justice. *Kansas Journal of Law and Public Policy* **1**(1): 69–82

Biegert C 1999 Sacred waters of aquatic and marine biodiversity. In: Posey D (ed.) *Cultural and Spiritual Values of Biodiversity, Intermediate Technology Publications*. UNEP, Nairobi, Kenya, pp. 402–3

Boyce J K 1995 Equity and the environment, social justice today as a prerequisite for sustainability for the future. *Alternatives* **21**(1): 12–7

Brennan A 1992 Moral pluralism and the environment. *Environmental Values* **1**: 15–33

Callicott B J 1980 Animal liberation: A triangular affair. *Environmental Ethics* **2**: 319–21

Callicott B J 1984 Non-anthropocentric value theory and environmental ethics. *American Philosophy Quarterly* **21**(1984): 299–309

Callicott B J 1994 Towards a global environmental ethic. In: Tucker M, Grim J (eds.) *Worldviews and Ecology: Religion, Philosophy, and the Environment*. Orbis, Maryknoll (Excerpts taken from Posey D (ed.) *Cultural and Spiritual Values of Biodiversity*. Intermediate Technology Publications, UK and UNEP, Nairobi, Kenya, p. 730)

Chambers P 1999 Marine and aquatic biodiversity. In: Posey D (ed.) *Cultural and Spiritual Values of Biodiversity*. United Nations Environment Programme, Nairobi, Kenya, Chap. 10, pp. 397–433

De Young R 1999 Environmental psychology. In: Alexander D E, Fairbridge R W (eds.) *Encyclopedia of Environmental Science*. Kluwer Academic Publishers, Hingham, MA

Dower N 1998 Developmental ethics. In: Chadwick R (ed.) *Encyclopedia of Applied Ethics*. Academic Press, USA, pp. 767–79

Dunlap R E, Gallup G H, Gallup A M 1993 Of global concern: Results of the health of the planet survey. *Environment* **35**(November): 6–15, 33–39

Giddens A 1990 *The Consequences of Modernity*. Stanford University Press, Stanford, CA

Gillon R 1998 Bioethics, overview. In: Chadwick R (ed.) *Encyclopedia of Applied Ethics*. Academic Press, pp. 305–17

Goldberg D T 1995 *Ethical Theory and Social Issues: Historical Texts and Contemporary Readings*. Holt, Rinehart and Winston, p. 527

Goodpaster K E 1978 On being morally considerable. *Journal of Philosophy* **75**: 303–25

Gupta A K 1991 Sustainability through biodiversity: Designing crucible of culture, creativity and conscience. Presented at the International Conference on Biodiversity and Conservation held at Danish Parliament, Copenhagen, November 8. (IIM WP No. 1005, 1992, Ahmudabad)

Gupta A K 1991 Building upon people's ecological knowledge : Framework for studying culturally embedded CPR institutions. Presented at the 2nd Annual Conference of the International Association for the Study of Common Property, Winnipeg, Manitoba, 26–29 September (IIM WP No. 1004, 1992 Ahnudabad)
Gupta A K 1994 Ethical issues in prospecting biodiversity. Background paper for evolution of Pew Conservation Scholar Program Guidelines for ethical issues in exploring biodiversity. IIM WP No. 1205, Ahmedabad
Gupta A K 1995 Ethical dilemmas in conservation of biodiversity: Towards developing globally acceptable ethical guidelines. *Eubios Journal of Asian and International Bioethics* **5**: 40–6
Gupta A K 1999 Ethical dilemmas in the conservation of biodiversity. In: Shantharam S, Montgomery J F (eds.) *Biotechnology, Biosafety, and Biodiversity: Scientific and Ethical Issues for Sustainable Development*. Science Publishers, NH, pp. 93–128
Gupta A K, Shukla S, Darshit S, Pawan M, Muralikrishna S 1997 Consumer response to green market opportunities. Paper presented at the National Workshop on Organic Food Products, August 2, ISAM, Ahmedabad
Hill S 1978 Ecology, ethics and feelings: The re-evaluation of existing value and the search for absolute values. *Proceedings of the Seventh International Conference on Unity of the Science*, Nov 24–6, Boston. The Cultural Foundation Press, New York, pp. 593–607
Hill S 1992 Ethics, sustainability and healing. *Proceedings of the Alberta Round Table on Environment and Economy and the Alberta Environmental Network*, EAP Publication, pp. 1–17
Johnson L 1991 *A Morally Deep World*. Cambridge University Press, Cambridge, UK
Muller-Rommel F 1989 *New Politics in Western Europe: The Rise and Success of Green Parties and Alternative Lists*. Westview Press, Boulder, CO
Munshi K M 1952 *A Gospel of Dirty Hand*, Ministry of Broadcasting and Information (Government of India), Publication Division, New Delhi, pp. 21–184
Pedersen P 1993 Nature, religion and identity: A contribution to the anthropology of knowledge about nature. *Paper presented to the Nordic Research Seminar on Anthropology and Nature*, Faroe Islands, October 28–30, 1993
Pluhar E 1998 Animal rights. In: Chadwick R (ed.) *Encyclopedia of Applied Ethics*, Academic Press, Vol. 1, pp. 161–72
Rawles K 1998 Biocentrism. In: Chadwick R (ed.) *Encyclopedia of Applied Ethics*. Academic Press, USA, Vol. 1, pp. 275–83
Rolston III H 1988 *Environmental Ethics: Duties to and Values in The Natural World*. Temple University Press, Philadelphia, PA, p. 391
Sen A 1981 Description as a choice. Oxford Economic paper, p. 353–369
Solomon R 1996 *A Handbook for Ethics*. Harcourt Brace, New York, p. 146
Society for Research and Initiatives for Sustainable Technologies and Institutions (SRISTI) 1993 Creativity, innovation, entrepreneurship and networking at grassroots level: Conserving biodiversity through documentation, experimentation and value addition in local innovations for development and diffusion of sustainable technologies and institutions. Mimeo, Ahmedabad
Stern P C, Dietz T, Kalof L, Guagnano G A 1995 Values, beliefs, and proenvironmental action: Attitude formation emergent attitude objects. *Journal of Applied Social Psychology* **25**: 18, 1611–36
Stone C D 1987 *Earth and Other Ethics*: *The Case for Moral Pluralism*. Harper and Row, New York, p. 280

A. K. Gupta and R. Sinha

Environmental Determinism

Environmental (or geographical) determinism views the natural environment as the basic factor controlling human achievement, an environment incorporating location and the geophysical and biophysical features native to the earth, including climate, structure, minerals, soil, flora, and fauna—all that is intrinsically 'earthy' rather than formed or shaped by human action. Some blur this issue by incorporating human-modified features, others reduce the rigidities of 'determinism' to 'influences,' strong or slight. Some trace environmental influences deep within the human psyche. Environments vary with the objects 'environed.' 'Environmentalism' can also refer to the search for environmental relationships and efforts to salvage the ecological system. Contextual awareness is therefore needed to pluck one theme from a varied setting, interpreting 'environment' and 'determinism,' separately and together.

Frequently dismissed as passé, environmental determinism nevertheless touches sensitive and perhaps unresolved issues—the remote and continuing impact of nature on humans, human and racial genesis, ethnic and national origins, conditions that favor or frustrate economic and intellectual achievement, the global constraints and opportunities that define all hopes of present or future accomplishment—perhaps the whole panorama of human studies, now extending into biogenetic and biosociological realms.

1. *Prehistory from Hippocrates to Montesquieu*

Frequently entangled with astrological or 'cosmic determinism,' environmental determinism permeated Chinese, Indian, Babylonian, and Greek philosophy before the dawn of Western Civilization. By the fourth century BC, Hippocrates had combined the 'elements' of earth, air, fire, and water with geographic location to promote economic and communal health, evaluating swampy hollows and windy uplands accordingly. Greeks generally perceived their Mediterranean shores as a 'golden mean' remote from northern winters and southern heat, a zone favoring moderation, intellectual pursuits, and civil discourse. Open steppes were fine for savage Scythians, and sun-burned Africans and Asians were deemed to lack the spirit essential for

liberty and war. As Glacken (1967) remarked, 'the familiar assertions of modern times … are found in a cruder form in antiquity: warm climates, produce passionate natures; cold, bodily strength and endurance; temperate climates intellectual superiority; … and a fertile soil produces soft people, a barren one makes them brave.'

Greek thought cast long shadows across Medieval Europe, with variant determinisms facing the unforeseen. P. Sorokin found few who failed to mention 'geographic influences,' and Renaissance sage Jean Bodin (1530?–1596 AD) scanned 2000 years of earth and sky for principles to guide the French monarchy through religious and ethnic strife. Two centuries of Enlightenment later, Montesquieu (1689–1755) might ignore the stars, but still insisted that 'climate is the first, the most powerful, of all empires.' Such epigrams evoked challenge from Church and philosopher alike. Theologians could scarcely accept 'revealed religion' as a matter of climate, Voltaire found bad weather no excuse for bad government; and skeptic Hume could not correlate national characteristics with atmosphere. In rejoinder, Montesquieu affirmed he had never thought that blind fatality could generate intelligent beings, that physical and moral causes were always intertwined, thereby provoking later determinist-possibilist debate, but his 'determinism' lingered beyond the Encyclopedic era as the particularizing sciences, geography included, took shape.

2. *Cosmos, Teleology, and Environment*

Montesquieu's 'environmentalistic' instances were echoed by Immanuel Kant (1724–1804) as he lectured on the roles of space and time in organizing geographical and historical phenomena, and shortly thereafter Alexander von Humboldt (1769–1859) and Karl Ritter (1779–1859), founding fathers of modern geography, generated similar interests, although not necessarily in deterministic fashion. Humboldt, seeking law in the natural Cosmos, repeatedly warned that presuppositions regarding environmental impact on human character involved 'wanton trespass' beyond the limits of knowledge, pointedly observing that New World grasslands lacked the pastoral nomadism of the Old World, even as American desert skies generated no parallels to Arabian astronomy (Tatham, in Taylor 1957). Yet environmental awareness informed his views of esthetic life, and (however secular) he sought always to manifest 'the perpetual influence of physical nature on the moral system and on the very fate of mankind' (Anuchin 1977).

Such thoughts were central for Ritter, confident (with J. Herder) that Divine teleology joined earth and humans as body and soul, relating historical progression to geographical form. The continents, framed by landform and climate, presented natural regions which fostered cultural education and differentiation. Localizing human–historical data in 'Erdkunde,' Ritter thought that once the 'general laws' of nature and humans became clear, humans could glance 'backwards and forwards, to determine from the whole of a nation's surroundings, what the course of its development is to be and to indicate, in advance of history, what ways it must take to retain the welfare which Providence has appointed for every nation whose direction is right and whose conformity to law is constant' (Dickinson and Howarth 1933). Much already seemed apparent, for civilization, dawning in the Orient, reached its contemporary zenith over the fretted coastlines and narrow European seas, and could brighten to an American sunset. Such qualified determinism, stripped of its teleology, could harden into the 'scientific' determinism that was to follow.

3. *The Anthropogeographical Tradition*

Deterministic thought flowed strongly through the late-1900s, with Darwin, Marx, and Freud becoming household names, and sociologist Comte and historian Buckle articulating 'social laws.' It would have been surprising had some not invoked 'geographical environment' as determinative. As memories of Ritter and Humboldt faded, and Departments took shape in Germany, influxes from the natural sciences reinforced physical geography, while human geography paused uncertainly. Darwinism seemed a welcome barrier against academic Balkanization, bonding all life forms to an ancient earth, rather than presuming the adaptation of earth to humans (Freeman 1962). 'Survival of the fittest' was linked with race and culture, natural law plausibly extended through 'Social Darwinism' to society, and 'Evolution' accounted as excluding teleology. Re-evaluation of such interpretations now seems essential, given neo-Lamarckian thoughts, and P. Kropotkin's cooperative evolution. Yet a 'superficial gilding of Darwinian rhetoric' entered geographical discourse (Livingstone 1992). Certainly biological Darwinism was initially appropriated by Friedrich Ratzel (1844–1904) as he sought application of natural-scientific law, with his trilogy of space, location, and social dynamics situated in environmental settings. Critics found his first volume on 'Anthropogeographie' presenting 'man as essentially passive, [responding] with Euclidian precision [to] the laws of the physical environment which determine human activities, distributions, and organization, in both space and time' (Dickinson and Howarth 1933).

But subsequent readings have softened early criticism, for Ratzel followed analyses of the historical impact of geographical features with attention to cultural migration and growth, measuring and mapping the 'co-variants of human distributions' (Dickinson 1969). In 'Politische Geographie,' evolving societies and states filled their respective

lebensraum, expanding toward natural or ethnically shaped boundaries—an idea conceptually fertile and generally devoid of the nationalistic provocation some discern among later geopoliticians.

Ratzel's American disciple, Ellen Churchill Semple (1863–1932), excised organic metaphor and extended illustrative examples in 'translating' anthropogeography for the 'Anglo-Saxon world.' 'Influences of Geographic Environment' were traced through the history of America, the World, and the Classical Mediterranean. Particular attention was given to mountain passes, rivers, sea routes, cultural migrations, contact zones, and the subsequent emergence of political entities. To critics her examples seemed tendentious, furthering a dubious mono-causal approach to historical and cultural complexities. Darwinian concepts obtruded in statements that 'Man is a product of the earth's surface,' molding every thought and act, strengthening body and wits together. Yet, she warned, 'geographic determinant' was shunned and 'geographic control' mentioned only with caution (Semple 1911).

4. *Physiography and Climatic Change*

Attempts to find 'anthropogeographical laws' variously appropriate for inhabitants of mountains, plains, coastlands, peninsulas, and scattered islands aroused skepticism, but were compatible with the views of W. M. Davis (1850–1934), founding father among American geographers. Discarding Ritter's and Guyot's teleology for evolutionary 'cause-and-consequence,' he postulated that it was only the relationship of earth to inhabitants that gave geography its distinctive and unifying character, a theme encapsulated as 'physiography' and 'ontography,' all life-forms included—a formulation which 'set the pattern for the pursuit of man–land relationships on a deterministic basis' for the first decades of the twentieth century (Dickinson 1976).

Other American universities, pioneered by Chicago, followed this Davisian concept (Blouet 1981), even as Ellsworth Huntington (1876–1947), moving from physiography to ontography, became intrigued with the potential linkages of climate and civilization, a theme pursued from the dune-fringed lakeshores of Central Asia to tree-rings in the American Southwest, from Yucatan to Palestine, and beyond. Typed as a climatic arch-determinist by the 'Pulse of Asia,' he revised his ideas with criticism and experience, replacing progressive desiccation with climatic oscillation, and searching for climatic optima within the cyclonic Westerlies, where sequences of fine and stormy weather seemed to stimulate. Field observations revealed abandoned villages, evidence of bygone agriculture, alluvial deposition, and glacial retreat. Combing historical records and voluminous correspondence (selective fact gathering, some said), he sought to match, measure, and correlate historical and geographical patterns. Using maps, graphics, and sometimes vivid prose, he personified geography (and climatic determinism) to the American public. Not unaware of challenge, he accounted climate a largely overlooked factor in the selective triad, a process blending natural environment, eugenics, and cultural impetus—dying, however, with his magnum opus unfinished (Martin 1973).

5. *Stop-and-go Determinism*

Huntington exchanged similar views with Griffith Taylor (1880–1963), British born and Commonwealth oriented, who, migrating to Australia, moved through geology and climate to geography, seeing environment as suturing his physical and human concerns. As weather-person and physiographer, he accompanied Scott's last Polar expedition, and thereafter Antarctica's icy wastes and Australia's desert Outback were engrained on his environmental consciousness. He found his own assessment of the economic margins curtailing development and population growth at odds with those of 'boosters' advocating 'Australia Unlimited.' Somewhat pugnaciously, he relinquished his Geography Department at Sydney for the hopefully more congenial University of Chicago, whence he moved to found the Department at Toronto, incidentally close to the margins of the vast Canadian Shield, where settlement was restricted by geology and climate.

Taylor's outspoken 'geocratic' determinism had its limitations. Conceding Western Europe and Eastern America to the Possibilism advocated by 'We-ocratic' microgeographers, he scanned the earth as a whole, affirming that, facing expanses of barren rock, frigid cold, and harsh aridity, habitation was reduced toward the vanishing point (Taylor 1957). Correlating the racial distributions of anthropology with 'evolution on the map,' he viewed geography's task as environmental analysis. As a 'stop-and-go-determinist,' he sought 'Nature's plan,' believing that 'humans are like the traffic controller' in a city, accelerating, slowing, or stopping the flow of traffic, but ever alert to the economic cost of disregarding Nature's directions, including those of 'geopacifics,' focused on the benefits of international cooperation.

6. *Russian Environment and Marxist Thought*

Beyond the bounds of the West extended vast steppes, forests, and waterways, and Tsarist historians 'attributed many peculiarities in the historical development of Russia to the direct influence of the natural environment,' K. M. Ber affirming 'the fate of peoples is determined in advance and seemingly

inevitably by the nature of the locality they occupy' (Anuchin 1977). Such theses breed their antitheses, especially when urgent priority is given to industrialization. Wrestling with the determinism of dialectical materialism, geographers such as N. N. Baranskiy sought to link productivity with environment, but were overruled when the Communist Party accepted Stalin's 1938 edict, precluding consideration of 'the geographical environment' as 'a determining influence,' and affirming modes of material production as crucial (Matley 1966). With 'geographical deviationists' silenced and environmental determinism shunned as 'Western,' Soviet geography separated into 'physical' and 'economic–cultural' branches, reflecting 'incompatible' natural and social laws.

Following Stalin's death in 1953, some geographers ventured public challenge. V. A. Anuchin, finding 'inhuman' physical geography and 'unnatural' economic geography objectionable, urged a synthesis expressive of dialectical–materialistic 'determinism' rather than the 'indeterminism' of those who, denying causation, separated humans from Nature. Invoking Marxist phraseology, he presented the human race, emergent through labor and tool-making, as qualitatively distinct, if thoroughly 'natural.' Not all geographical exegesis was considered relevant to humans, for earth's 'landscape envelope' included untouched components (such as sea-floors and Himalayan peaks) remote from nature–human relationships, and excluded from 'the effective environment.' There, the spread of human population and its products reshaped 'the natural environment' into 'the geographic environment,' complexly uniting inorganic, organic and societal components into one humanized world, roughly equivalent to Western concepts of 'cultural landscapes.' Within this ultimate and seamless 'geographic environment,' natural elements continued, so Western-style environmental determinists—although likely to mistake conditions for causes, and be mechanically direct rather than dialectical when mediating natural laws into human settings—could thus appear valid, given occasional and accidental circumstance. Consequently, 'certain occurrences and phenomena in the geography of population and economy … may be determined, admittedly in an indirect but yet decisive manner, by the geographic environment and even [by] its natural complex alone.'

Anuchin's appeal for a merger of nature and humanity into one 'geographic environment' seemingly failed, but his impact appeared in 1963 when the Communist Party, prompted by geographers including N. N. Baranskiy and Y. G. Saushkin, rescinded Stalinist definition of geographic environment as 'a purely natural category,' even as dusty Virgin Lands, shrinking Aral Sea, and planning complications reinforced contentions that a unified geography could indeed assist in rehabilitating nature and humanity in Soviet 'territorial complexes.'

7. *'Possibilities' in Western Geography*

Given Soviet insistence that social laws reflected dialectical materialism, such changes were not viewed as concessions to Western-style 'possibilism,' and Russian geographers were becoming aware that few Western colleagues appreciated description as 'environmental determinists.' After 1900, European opinion had moved as 'determinism' and 'organism' became suspect, and society, not environment, was scanned for essential causes of social change. (Martin and James 1993). Alerted by sociological critiques of Ratzel, French geographers followed Vidal de la Blache and Brunhes into 'possibilism.' Environment was conceded its essential role, but *genres de vie* (or lifestyles) were followed into 'regional personality.' British geographers, also regionally inclined, were perhaps less precise, casually invoking Darwinian 'response to environment' (Freeman 1962), and indulging in 'frankly boring' debates about determinism (Livingstone 1992).

In the seminal 1920–40 era, pragmatic American geography was redefined by European—especially Germanic—influences. Anthropologist Franz Boas promoted anti-environmentalism, W. M. Davis moved from physiographic environmentalism toward regionalism (Blouet 1981), and others, H. H. Barrows in particular, urged geographers to move to undeterministic 'human ecology.' More decisively, C. O. Sauer, alluding to O. Schluter and S. Passarge, challenged all definitions presupposing environmental relations as parasitic and biased, moved tangible 'landscape' to the fore, and left environmental ties to be proved. Agreeing with A. Hettner, R. Hartshorne defined 'the nature of geography' in 'chorological' or 'areal differentiation' terms, thus deflecting attention to regionalism rather than environmentalism—a 'deviation' from geography's *raison d'être*.

With environmentalism excluded from definitions, some sensed a 'death blow to the doctrine of environmental determinism' (Dickinson 1976), but subsequent statements suggest otherwise. Questioning views that the valid search for environmental relationships necessarily presupposes 'determinism,' some note shifts to influences—'probabilistic,' 'possibilistic,' or simply 'permissive' (Lewthwaite 1966). Many, broadening their concept of geography to encompass historically formulated 'traditions,' followed 'the man–land tradition,' with mutual interaction sometimes progressing from land to humans deterministically. Others, following 'spatial traditions,' abstract geometrical 'deterministic systems' from the overall environment. Some have traced analogies, with cultural variations, between Western environmentalism and Oriental geomancy. Others have seen 'eurocentricity' at the core, and some, inspired by Marxism, view environmentalism as a class-derived phenomenon.

8. Environment Remains

Doubtless many a component—eurocentricity, imperialism, nationalism, racism, class, and other interests—have been involved in origin and expression. However, if ideas be judged on their own merits, not pronounced 'guilty by association' but evaluated for potential insights, then 'environmental determinism' cannot be dismissed out of hand. Buttimer (1993) affirms that 'the most demanding challenge for the geographer today is how to incorporate nature and the biophysical environment in descriptions of context while avoiding the resuscitation of environmental determinism.' Of course, with mind as well as matter involved in all human thought and action, a robotic response to environment is automatically excluded, but nevertheless 'physical causation' is an unavoidable fact, evident when an earthquake or hurricane strikes. In such situations 'geographical controls' are real enough, and at the abstract conceptual level (presumably in conjunction with other and alternative concepts) environmental determinism has its analytical value.

Whatever our views, the environment is always with us: driven out the door, it comes back through the window. 'Environmentalism' covers many things, not least recurrent crusades to avert the spoliation of natural or cultural treasures (Manners and Mikesell 1974). However, environment surrounds and touches all things, whereas 'environmental determinism' suggests ultimate limits to human aspirations. Willy-nilly, 'environment' seems likely to evoke abiding interest.

See also: Cassandra/Cornucopian Debate; Cultural Geography; Darwinism; Darwinism: Social; Determinism: Social and Economic; Environmental Psychology: Overview; Marxist Geography

Bibliography

Anuchin V A 1977 *Theoretical Problems of Geography*. Ohio State University Press, Columbus, OH

Blaut J M 1999 Environmentalism and eurocentrism: A review essay. *The Geographical Review* **89**: 391–408

Blouet B W 1981 *The Origins of Academic Geography in the United States*. Archon Books, Hamden, CT

Buttimer A 1993 *Geography and the Human Spirit*. Johns Hopkins University Press, Baltimore, MD

Chappell J E Jr 1981 Environmental causation. In: Harvey M E, Holly B P (eds.) *Themes in Geographic Thought*. St. Martin's Press, New York

Dickinson R E 1969 *The Makers of Modern Geography*. Routledge and Paul, London

Dickinson R E 1976 *Regional Concept: The Anglo-American Leaders*. Routledge and Paul, London

Dickinson R E, Howarth O J R 1933 *The Making of Geography*. Clarendon Press, Oxford

Freeman T W 1962 *A Hundred Years of Geography*. Aldine, Chicago

Glacken C J 1967 *Traces on the Rhodian Shore*. University of California Press, Berkeley, CA

Lewthwaite G R 1966 Environmentalism and determinism: A search for clarification. *Annals of the American Association of Geographers* **56**(1): 1–23

Livingstone D N 1992 *The Geographical Tradition*. Blackwell, Oxford

Manners I R, Mikesell M W (eds.) 1974 *Perspectives on Environment*. Association of American Geographer, Washington, DC, Publication No. 13

Martin G J 1973 *Ellsworth Huntington: His Life and Thought*. Archon Books, Hamden, CT

Martin G J, James P E 1993 *All Possible Worlds. A History of Geographical Ideas*, 3rd edn. Wiley and Sons, New York

Matley I M 1966 The Marxist approach to the geographical environment. *Annals of the American Association of Geographers* **56**(1): 97–111

Pattison W D 1964 The four traditions of geography. *Journal of Geography* **63**: 211–6

Peet R 1985 The social origins of environmental determinism. *Annals of the American Association of Geographers* **75**(3): 309–33

Semple E C 1911 *Influences of Geographic Environment on the Basis of Ratzel's System of Anthropo-geography*. H Holt, New York

Speth W W 1978 The anthropogeographic theory of Franz Boas. *Anthropos* **73**: 1–31

Speth W W 1999 *How it Came to Be: Carl O. Sauer, Franz Boas and the Meanings of Anthropogeography*. Ephemera Press, Eltensburg, WA

Taylor G (ed.) 1957 *Geography in the Twentieth Century*, 3rd edn. Philosophical Library, New York, Methuen, London

Yoon H-K 1982 Environmental determinism and geomancy: Two cultures, two concepts. *GeoJournal* **6**: 77–80

G. R. Lewthwaite

Environmental Economics

It may seem strange to someone unfamiliar with economics to assume people's behavior can be described independently from the environmental systems that sustain their very existence. Nonetheless, this is exactly what conventional economic analysis does. Production, exchange, and consumption decisions are usually modeled in ways completely separate from the natural environment. Over the last 30 years of the twentieth century there has been a growing realization among ecologists, and to a lesser degree in the economics profession as a whole, that this approach can lead to serious mistakes. Environmental economists study the effects that environmental systems have on the positive predictions and normative recommendations of economic models (see Solow 1991).

Early in the history of economic modeling, Hotelling (1931) and more recently Dasgupta and

Heal (1979) as well as numerous other economists recognized the importance of intertemporal issues for economic choices involving renewable and nonrenewable natural resources. These types of economic analyses are treated separately in the *Encyclopedia*.

Environmental economics assumes there are things people care about, either directly because they enhance individual well-being or indirectly because they influence the production of other commodities, that are: (a) scarce; (b) available outside markets; and (c) influenced by the activities of others. These objects of environmental choice (OECs) have diverse characteristics and result from a variety of activities. Six topic areas organize this overview: the allocation problems they pose, the design of policy, measures for the amounts of environmental amenities and their economic values, OECs as risks, the distributional implications, and interactions between growth and environmental resources.

1. Problems—Public Goods, Externalities, and Congestion

The definitions for public goods and externalities are closely linked to the conditions required to sustain private property rights. As a result, the allocation problems they create are sometimes reduced to descriptions of causes and remedies for market failures.

Pure public goods are nonrival and nonexcludable in consumption (Cornes and Sandler 1996). Each person's consumption of a public good does not reduce the amount available to others. Moreover access to these goods cannot be limited. While it is hard to provide examples satisfying these ideal conditions, it is easy to cite examples of goods and services with attributes close to a pure public good's features. Local air quality, measured in terms of the concentration of airborne pollutants, varies with location but within large areas is approximately the same for all individuals living there. The global concentrations of greenhouse gases (e.g., CO_2, methane, etc.) can be described as examples of pure public 'bads' because emissions of them at any location have equivalent effects on their atmospheric stocks. However, the consequences of those accumulations for climate change and local weather conditions will depend on where people live. Perfect adherence to the ideal definition of a public good is not essential to consider the implications of 'publicness' for the efficiency of different resource allocation mechanisms.

Externalities describe a class of interactions between two or more economic agents. One group's (or individual's) consumption, production, or exchange activities have unintended effects on the production or consumption of others. These external effects resemble public goods (or bads) in that they can have the same nonrival and nonexcludable attributes.

Congestion describes a phenomenon whose externality effects reduce the nonrival attributes of a public good. For example, over some levels of recreational use of a beach or a wilderness area (and the levels will be different in each case) further users can be supported without extra cost or diminution in the satisfaction each person realizes from his or her use. However, at some point, this simultaneous use diminishes each person's experience. The phenomena linking those losses to the added users of a resource at the same time is described as congestion.

While all three of these concepts are relevant outside environmental applications, they have found special meaning when environmental resources provide the services treated as OECs or as the assets whose services provide the conduits for the negative or positive interactions between economic agents.

These concepts are important for at least two reasons. First, the incentives created by private markets do not assure that resources will be allocated to their highest valued uses when one or more of these problems is involved. Second, the information available from markets cannot be assured to signal and evaluate the performance of economic institutions. Much of environmental economics can be interpreted as responses to these issues.

2. Policies—Context, Design, and Performance

Economic analysis seeks to abstract from the 'details' of a problem, focusing instead on the incentive properties of each decision context. In the case of environmental problems these details can matter to the insights offered by economic analysis. For example, economic descriptions of production and consumption activities must be consistent with materials and energy balances (Kneese and Bower 1968). When they are not, it is easy to overlook the residuals and the problems that must be faced in disposing of them. This insight is important because most externalities arise from using the services of an environmental asset (i.e., the ambient atmosphere or a free flowing river) for residual disposal. The form and severity of environmental externalities depend on both the types of residuals involved and the functions of the ecosystem receiving them. As a rule, the ecosystems are open access resources whose services in some uses are rival applications. They may well also be subject to depletion over time.

Economic prescriptions for these externalities (as well as for congestion) tend to focus on creating the 'right' incentives. Private choices should reflect their full or social incremental cost (i.e., the private marginal cost of producing any commodity plus the marginal external costs due to any damages associated with a commodity's production, consumption, and ultimate disposal). When the marginal damages associated with the production or consumption of a good are

measured per unit of each residual emitted, it is sometimes described as the price that should be charged for that emission. Pigouvian taxes or effluent charges set equal to these marginal damages are one type of policy response to externalities.

Under simplified conditions, efficient resource allocation in the presence of environmental externalities (or congestion) requires that, at the margin, agents face the full cost (or benefit) of their actions including the damages caused directly or indirectly from each choice. Those experiencing damages are assumed to take defensive actions and costs computed recognizing these choices, but the injured parties do not need to be compensated (Baumol and Oates 1988).

With full information, policy can take a variety of forms. These can be charges per equivalent unit of impact (taking account of the ecosystem) to the agents causing externalities; permits to emit pollution that can be traded; and even subsidies to abate emissions. Each of these policy measures can be made to appear equivalent in terms of the incentives for incremental choices. Each approach requires different implementation rules and a different set of adjustments to reflect their effects at the extensive margin of choice (e.g. the entry and exit decisions of firms). Charges vs. subsidies will affect incentives to enter an industry (Carlton and Loury 1980). Tradable emission permits must be denominated in units reflecting the effects of emissions on the services associated with ambient environmental quality that defines the OECs. They must also consider both the initial distributions of permits among firms, the rules for trading, and their ability to be saved (if temporal patterns of emissions are important).

Unfortunately, in the setting where actual decisions must take place, information is scarce and uncertainty abounds. If uncertainty affects the marginal control costs, then it can affect the preferred policy instrument. Comparing charges vs. marketable permits (denominated in units of the OEC) Weitzman (1974) has shown the choice depends on the relative steepness of the marginal benefit and cost schedules. Tradable quotas are favored if the marginal damage function is steeper than marginal abatement cost and effluent charges for the reverse. His results depend on the approximate linearity of these schedules. Moreover, the conclusion is a relative statement not an absolute judgment. Composite strategies using price and quantity instruments together can be shown to be better in the sense that they enhance overall welfare (Roberts and Spence 1976).

This summary implies that with perfect information either instrument can lead to an efficient resource allocation. Charging for emissions or distributing permits to pollute seems to convey 'rights' to different economic agents to use ecosystems' services. The conclusion that they can be equivalent may seem counterintuitive. While surprising it is one reflection of the general implications of the Coase (1960) theorem. That is, in a world of perfect information, no transaction costs, and no income effects (from assigning rights to ecosystem services) the theorem demonstrates the assignment of rights will not affect the efficient outcome. It is often used in arguments that externalities can be resolved without public sector intervention by assigning property rights to ecosystem services. In most cases the practical relevance of this conclusion is an empirical question. Nonetheless, the economic logic underlying the theorem also explains why the available policy instruments can be equivalent.

Until recently, most environmental policy was driven by the broad goals of the existing legislation. In the USA the rulemaking process, implementing this legislation, has favored command-and-control (CAC) regulations. Specific technologies were prescribed to meet objectives such as using the best available control technology to establish the lowest emission rate. These types of engineering standards specified, at the unit process level, emission rates for the same types of pollutants that sometimes were quite different for separate equipment within the same plant. The result was very different incremental costs for controlling the same pollutants. Abatement costs were also increasing over time as more stringent controls were imposed. Areas found to be in nonattainment with environmental standards faced restrictions on further economic activity. The result was a policy environment seeking some ways of increasing the flexibility in how firms could respond. Some regions in the USA had started the process of 'regulatory reform' and introduced some movement toward the use of economic incentives when the 1990 Clean Air Act dramatically changed the US regulatory climate. It offered the first large-scale emission permit trading system as a way of realizing more stringent controls for sulfur dioxide. Ellerman et al. (2000) estimate that emission trading yielded cost savings by 33 to 67 percent of the most reasonable CAC alternative. Permit trading schemes are being considered or used in small regional markets for an array of environmental resources—including air and water pollutants, wetlands, habitat for endangered species, and carbon dioxide (see Tietenberg 1990).

Most economic choices about market goods and services cannot be analyzed realistically without recognition of their environmental consequences. This requirement is symmetrical. Environmental policy analysis cannot ignore the market and regulatory context for the products produced along with the environmental externalities. Two sets of insights have been especially important to reaching this conclusion. The earliest, associated with Buchanan (1969), argued that a monopolistic product market may, by restricting output, serve environmental goals. In this context, imposing effluent charges or restricting emissions with tradable quotas could actually lead to welfare losses.

The logic underlying this result is more general. Judgments about the preferred incentives to realize an

efficient resource allocation in an economy with externalities and multiple distortions are extremely difficult. An emerging issue indicating how such distortions can affect the design of environmental policy began with the promise of 'double dividends' from effluent charges on CO_2 emissions (Bovenberg and de Mooij 1994). Subsequent research rediscovered the issues associated with designing optimal product taxes in the presence of multiple taxes—an important theme of the optimal taxation literature (Auerbach 1985). The double dividend argument relies on using the revenues from environmental fees (or authorized permits) to reduce the marginal rates of one or more large, pre-existing taxes. It is suggested that the result of the tax substitution may be zero or negative gross cost. Subsequent research has extended arguments about the potential importance of pre-existing tax distortion to evaluations of the effects of new environmental regulations, suggesting in the case of SO_2, costs over 70 percent higher than they would be in the absence of the taxes (Goulder et al. 1997). Nonetheless it is important to recognize how specific this conclusion is to the circumstances described in each model. Changing the assumptions about how private goods relate to the externality in consumers' preferences will affect the size of the excess burden attributed to existing taxes and therefore the gains from reducing them. What is less likely to change is the attention the argument implies should be given to the political economy of how and why environmental legislation takes specific forms.

3. *Amounts: The Role of Nonmarket Valuation and Benefit-cost Analysis*

Nearly all analysis of policy instruments assumes we know how much environmental quality to provide. Answering this question involves more than deciding the amount of pollution permitted. It requires consideration of a wide range of natural assets whose amenity services are important to people. Everything from endangered species and their habitat to land and water resources that provide essential capacity for a wide range of outdoor recreation must be considered. Some choices must be made more carefully than others must because some allocations of a natural environment will irreversibly alter its ability to provide amenity services (Krutilla 1967, Krutilla and Fisher 1975).

The extent of 'publicness' of the OECs compounds the difficulties in making these choices. Free-riding incentives arising from nonexclusion assure that private action alone cannot be relied upon to realize the efficient level of these types of resources (Cornes and Sandler 1996). Recovering information equivalent to what is provided when market exchanges correctly signal scarcity requires economic 'detective work.'

Two strategies have been developed since Hicks (1943) first defined rigorously the monetary representations for the tradeoffs defined by economic choices (Freeman 1993). The first relies on the logic of revealed preference and uses choices of private goods to recover the tradeoff information. For example, it recognizes that a consumer's decision to purchase (or rent) a home in an area free of air pollution reflects, in part, a payment for that ambient quality in comparison to the next best alternative (Palmquist 2000). With a wide array of alternative conditions, the price gradients arising from different air quality conditions, with other housing attributes held constant, would reveal the incremental willingness to pay for improving air quality (Freeman 1993). Similar logic implies one could use the trips taken to recreation sites at locations with different travel costs to reveal the demand for the recreation these sites provide (Freeman 1993). Given a set of recreation sites, varying in quality conditions, consumers' visitation decisions offer the potential to measure the incremental value of those quality dimensions. Decisions to avert or mitigate the negative consequences of pollution offer the same potential to reveal underlying tradeoffs (Freeman 1993).

The logic underlying each strategy relies on reconstructing the unobserved elements required to define the economic tradeoff revealed with each choice. In the case of nonmarketed goods this process implies that a connection is also defined between the decision and the OEC of interest. For the tradeoff to be informative about an individual's preferences for the nonmarketed good these connections must convey a degree of exclusively (i.e., economic value of the amenity requires consumption of the linked private good). They must also specify how enhancements in an individual's ability to pay would change them. One way this control over the substitution and income effects of choice is implemented comes from the joint imposition of restrictions on the properties assumed in describing an individual's utility function. Weak complementarity (Mäler 1974) and the Willig (1978) condition are among the most commonly used. Weak complementarity maintains that the environmental amenity is of value when it is used with the private good. Without some positive consumption of that good and improvement in the amenity has no value to each person. The Willig condition adds to this restriction a requirement that the incremental value of increases in the amenity, per unit of the private good consumed, do not change with income.

Stated preference approaches to recovering economic tradeoffs offer the second nonmarket valuation strategy. They rely on presenting people believable (but hypothetical) choices with proposed financial consequences and interpreting their stated decisions as bounding economic tradeoffs (Mitchell and Carson 1989). An important motivation for proposing choices arises from the inability to link all the ways amenities enhance people's well being to observable choices of

private goods. Such unlinked reasons are sometimes referred to as nonuse or existence values (Krutilla 1967). Stated preference methods are controversial because they require that survey responses be treated as economically consequential (Diamond and Hausman 1994, Hanemann 1994).

The primary use for nonmarket valuation estimates has been in benefit-cost analyses of projects involving environmental resources and regulations intended to improve environmental quality. Benefit-cost analyses (BCA) assume there are reasons why an existing resource allocation is not efficient. The analysis seeks to evaluate whether an aggregate choice improves the efficiency (Arrow et al. 1996). Four aspects of applying BCA are especially important. First, the net benefits from an aggregate choice are summarized as a present value and are influenced by the time horizon and discount rate used. Both are potentially influential and controversial judgments (see Portney and Weyant 1999 for a summary of the issues in making a choice). Second, large changes in relation to the overall economy have general equilibrium effects that may require evaluating the effect at the household level and aggregating (i.e., computing net benefits at the household level rather than aggregating benefits and costs separately, see Hazilla and Kopp 1990). Third, BCA analyses will always face incomplete information about the changes being evaluated from a baseline to a new outcome. When an irreversible choice is being made and there is the possibility for resolving some of the uncertainty over time, the analysis should include the conditional value of any new information that would be consequential to the decision (Arrow and Fisher 1974). Finally, monetization of all factors creating benefits or costs is not essential for BCA to be informative. Observed choices (at the individual level) define bounds for the tradeoffs implied by a decision. This same logic can be used to delineate bounds for the missing elements in a BCA. That is, we define what the missing elements in a measured value for net benefits would need to be to support a specific choice (Krutilla and Fisher 1975).

4. *Risk: Perceptions and Values*

One of the most important classes of OECs involves environmental sources for changes in the probabilities of undesirable outcomes. For example, air pollutants have been associated with premature death and increased episodes of serious illness. These decisions are highlighted for two reasons. First, the framework for evaluating private and public choice must consider attitudes toward risk and the opportunities available to adjust to it. Within a framework that evaluates decisions made before the uncertain events are resolved, there has been considerable discussion of how these risks should affect monetary measures of policy outcomes and discounting. There seems to be broad consensus that attitudes toward risk and opportunities for private and public adjustment should be built into the monetary measures of *ex ante* net benefits and not to the discount rates (Portney and Weyant 1999). Second, studies of laypersons' risk perceptions and behavior in both experimental and survey contexts suggest marked discrepancies between subjective beliefs and experts' recommendations. Equally important this behavior contrasts with the predictions of an expected utility model for individual choice. These differences are especially important for small probability and long latency outcomes (Slovic 1987). Such findings raise questions with what constitutes an informed choice and potential limits to consumer sovereignty (Machina 1990). They also highlight the role of information and risk communication as policy instruments.

5. *Distribution*

The term distribution in most economic analyses of policies would lead to a discussion of equity effects, interpreted as evaluating the consequences of actions for different income groups. For environmental policy, distributional considerations are inherently more complex. At least four perspectives have been considered. First, the use of an aggregate net benefits criterion for evaluating policy itself is a distributional judgment. It implies departures from a strict efficiency standard and a willingness to admit policies where there exists the potential for improvements. Positive aggregate net benefits means those gaining could compensate those losing. It does not require the compensation to take place.

The second distributional perspective follows from the identification of long lived, potentially irreversible externalities—with greenhouse gases and nuclear waste being two widely recognized examples. In these cases the distribution of costs and benefits extends across generations and the equity issues are often framed in ethical terms. One important implication of this line of reasoning is that policies should treat all generations alike. As a result, this line of reasoning would seem to imply a zero discount rate should be used. However, there is an equally strong intuition from a different perspective suggesting it is not morally acceptable. It implies a decision criterion (i.e., the zero discount rate) that would demand excessively high savings rates of any one generation. This countervailing influence is important because Koopmans' (1960) classic analysis of choice rules over infinite dimensional consumption streams concluded that these rules must display impatience or some amount of positive discounting. Otherwise, it would require the first generation to sacrifice everything. With preference separability over time, Arrow concluded irreversibility does not change conclusions in favor of intertemporal discounting (see Portney and Weyant 1999). Of course

to the extent preferences or production display important substitutions or complementarities with irreplaceable environmental resources, then the guidance will depend on the details of those linkages (Portney and Weyant 1999).

Economic and environmental feedback effects can have significant distributional consequences within short time horizons. Several cases illustrate how this third perspective can arise. Deciding between command and control vs. incentive based policies can seem clear-cut, considering the total costs of meeting a uniform standard (Tietenberg 1990). However, the environmental transfer system, linking emissions to ambient concentrations, can imply the cost savings are realized by allowing the cleanest areas to become more dirty while those exceeding the standard become cleaner. These properties can impact political acceptability as well as different income and demographic groups.

Much less is known about the economic feedback loops connecting environmental policy in one sector to its impacts on others. The economic models used to compute general equilibrium responses usually describe activities as if they all took place at the same 'point-in-space' (Hazilla and Kopp 1990). Recent adaptations to these models (Espinosa and Smith 1995) and to household equilibrium location models (Sieg et al. 2000) confirm the importance of relaxing these assumptions. Indeed, the results from the equilibrium location model suggests that spatially differentiated, but consistent, air quality improvements can induce housing price adjustments (within a closed system) that imply some households would actually be worse off after an improvement in their air quality. This result arises because the housing price increases induced by some households' relocations can offset the air quality gains for those households with a small taste for environmental public goods.

The last distributional perspective, paradoxically, receives the greatest attention in agencies' design of policies (in the US some legislation specifically identifies concern for the impacts of regulations on small businesses) and often the least in academic discussions. It concerns the impact of regulations on industries and different sized firms within a sector. Empirical evidence confirms that environmental regulations can have effects on the timing, location, and scale of plant investments. Considering only air quality regulations in the US, they find that activity shifts to areas that meet ambient standards, not subject to stringent area-wide regulations. Less activity takes place in areas that are not in attainment with national ambient air quality standards. When new plant investments do take place they seem to involve larger up-front investments in comparison to what takes in the attainment areas. In one of the sectors, these authors studied (industrial organic chemicals) their results suggest plants grew to smaller sizes with regulation than in the preregulation era (Morgenstern et al. 1998).

6. *Growth and the Environment*

One of the most durable questions posed in considering interactions between the economy and the environment concerns whether the environmental system poses some 'ultimate limit' on the scale of economic activity (see Barnett and Morse 1963). Sometimes expressed in terms of raw materials (or natural resources) and other times in terms of ability to provide a sustainable food supply and waste absorptive services to a growing world population, the issue most recently has been expressed in terms of whether growth in income has a self-regulating property. That is, as household income grows, the taste for environmental amenities increases. Providing these services requires reducing pollution, protecting natural environments that provide habitat for endangered species, and enhancing absorptive capacity of ecosystems for some types of pollution.

At one scale, this issue has been posed using what is termed an 'environmental Kuznets curve' (Grossman and Krueger 1995). Pollution is hypothesized to have an inverted U-shape with household income. It increases until income growth creates demand for amenities that ultimately reduces pollution and leads to policies that restore environmental quality. To date, the questions raised in this line of research have been based on empirical analyses, relying on the contemporaneous performance of economies varying in pollution and household income, as if they could be treated as descriptions of a single economy moving through a dynamic process.

This new formulation of the old question is potentially important. It casts the issue in its most relevant form. Economists are concerned about the services provided by environmental resources because market institutions cannot be assured to signal the scarcity of these resources. Understanding both the static distortions to patterns of use and the dynamic nonmarket incentives to bring about changes in the utilization of these natural assets is important for designing new institutions that recognize the role of environmental resources in assuring that it is possible to continue to enhance economic well being.

See also: Cassandra/Cornucopian Debate; Economic Geography; Economic Growth: Theory; Economics: Overview; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Planning; Public Goods: International; Social Movements: Environmental Movements

Bibliography

Arrow K J, Fisher A C 1974 Environmental preservation, uncertainty, and irreversibility. *Quarterly Journal of Economics* **88**: 312–9

Arrow K J, Cropper M L, Eads G C, Hahn R W, Lave L B, Noll R G, Portney P R, Russell M, Schmalensee R, Smith V K, Stavins R 1996 Is there a role for benefit cost analysis in environmental, health and safety regulation? *Science* **272**: 221–2
Auerbach A J 1985 The theory of excess burden and optimal taxation. In: Auerbach A J, Feldstein M S (eds.) *Handbook of Public Economics*. North-Holland, Amsterdam, Vol. 1
Barnett H J, Morse C 1963 *Scarcity and Growth*. Johns Hopkins University Press, Baltimore, MD
Baumol W F, Oates W E 1988 *The Theory of Environmental Policy*, 2nd edn. Cambridge University Press, Cambridge, UK
Bovenberg A L, de Mooij R A 1994 Environmental levies and distortionary taxation. *American Economic Review* **84**(4): 1085–9
Buchanan J M 1969 External diseconomics, corrective taxes and market structure. *American Economic Review* **59**: 174–7
Carlton D W, Loury G C 1980 The limitations of pigouvian taxes as a long run remedy for externalities. *Quarterly Journal of Economics* **95**: 559–66
Coase R H 1960 The problem of social cost. *Journal of Law and Economics* **3**: 1–44
Cornes R, Sandler T 1996 *The Theory of Externalities, Public Goods, and Club Goods*, 2nd edn. Cambridge University Press, Cambridge, UK
Dasgupta P S, Heal G M 1979 *Economic Theory and Exhaustible Resources*. Cambridge University Press, Cambridge, UK
Diamond P A, Hausman J A 1994 Contingent valuation, is some number better than no number. *Journal of Economic Perspectives* **8**: 45–64
Ellerman A D, Joskow P L, Schmalensee R, Montero J P, Bailey E M 2000 *Markets for Clean Air: The U.S. Acid Rain Program*. Cambridge University Press, Cambridge, UK
Espinosa J A, Smith V K 1995 Measuring the environmental consequences of trade policy: A non-market CGE analysis. *American Journal of Agricultural Economics* **77**(August): 772–7
Freeman A M III 1993 *The Measurement of Environmental and Resource Values: Theory and Methods*. Resources for the Future, Washington, DC
Goulder L H, Parry I W H, Butraw D 1997 Revenue raising versus other approaches to environmental protection: The criticasignificance of pre-existing tax distortions. *Rand Journal of Economics* **28**: 708–31
Grossman G M, Krueger A M 1995 Economic growth and the environment. *Quarterly Journal of Economics* **110**: 353–77
Hanemann W M 1994 Valuing the environment through contingent valuation. *Journal of Economic Perspectives* **8**(4): 19–43
Hazilla M, Kopp R J 1990 Social cost of environmental quality regulations: A general equilibrium analysis. *Journal of Political Economy* **98**: 853–73
Hicks J R 1943 The four consumer surpluses. *Review of Economic Studies* **11**(1): 31–41
Hotelling H 1931 The economics of exhaustible resources. *Journal of Political Economy* **39**: 137–75
Kneese A V, Bower B T 1968 *Managing Water Quality: Economics, Technology and Institutions*. Johns Hopkins University Press for Resources for the Future, Baltimore, MD
Koopmans T C 1960 Stationary ordinal utility and impatience. *Econometrics* **28**: 287–309
Krutilla J V 1967 Conservation reconsidered. *American Economic Review* **57**: 777–86
Krutilla J V, Fisher A C 1975 *The Economics of Natural Environments: Studies in the Valuation of Commodity and Amenity Resources*. Johns Hopkins University Press for Resources for the Future, Baltimore, MD
Machina M J 1990 Choice under uncertainty: problems solved and unsolved. In: Hammond P B, Coppock R (eds.) *Valuing Health Risks, Costs and Benefits for Environmental Decision Making*. National Academy Press, Washington, DC
Mäler K-G 1974 *Environmental Economics: A Theoretical Inquiry*. John Hopkins University Press for Resources for the Future, Baltimore, MD
Mitchell R C, Carson R T 1989 *Using Surveys to Value Public Goods, The Contingent Valuation Method*. Resources for the Future, Washington, DC
Morgenstern R D, Pizer W A, Shih J S 1998 *The Costs of Environmental Protection*. Paper No. 98-36, Resources for the Future, Washington, DC (May)
Palmquist R B 2000 Property value model. In: Mäler K-G, Vincent J (eds.) *Handbook of Environmental Economics*. North Holland, Amsterdam
Portney P R, Weyant J P (eds.) 1999 *Discounting and Intergenerational Equity*. Resources for the Future, Washington, DC
Roberts M J, Spence M 1976 Effluent charges and licences under uncertainty. *Journal of Public Economics* **5**: 193–208
Sieg H, Smith V K, Banzhaf H S, Walsh R 2000 Estimating the General Equilibrium benefits of Large Policy Changes: The Clean Air Act Revisited. NBER Working Paper 7744 presented to Econometric Society, Boston, Massachusetts (June)
Slovic P 1987 Perception of Risk. *Science* **236**: 280–5
Solow R M 1991 Sustainability: An Economist's Perspective. 18th J. Seward Johnson Lecture, Marine Policy Center, Woods Hole Oceanographic Institution (June 14)
Tietenberg T H 1990 Economic instruments for environmental regulation. *Oxford Review of Political Economy* **6**(1): 17–33
Weitzman M L 1974 Prices versus quantities. *Review of Economic Studies* **41**: 477–91
Willig R D 1978 Incremental consumer's surplus and hedonic price adjustment. *Journal of Economic Theory* **17**(2): 227–53

V. K. Smith

Environmental Health and Safety: Social Aspects

Environmental health, especially related to toxic substances, is a highly politicized area. As the use of chemicals in industry, agriculture, and the household has become central to the whole economy, many problems have been detected. The physical and mental health impacts of toxic substances have generated significant social movement activity, have led to major government regulation, and have demonstrated the capacities of laypeople to make important advances in scientific knowledge.

1. The Scope and Significance of Environmental Factors on Health

Much disease is caused by substances and conditions in people's surrounding environment, including chemicals and air particles in factories, pesticides in agriculture, toxic wastes in residential neighborhoods, and radiation in the atmosphere. The 'environment' is so broad a term that we could virtually subsume all disease processes under the umbrella of unhealthy living and working conditions. But when experts speak of social aspects of environmental health, they generally focus on the health effects caused by toxic substances in people's immediate surroundings. These chemical-related, air-pollution-related, and radiation-related diseases and symptoms are the phenomena that have generated much conflict, policy making, legislation, public awareness, media attention, and social movement activity.

These social outcomes were prefigured by prior developments in occupational health. Many harmful effects of chemicals and substances are first found in the work setting, and then generalized to the community level. There was much corporate, governmental, and scientific opposition to detecting and remediating occupational causes of disease, setting a tone for responses to environmental conditions. Even a clearly work-related condition such as black lung disease was the source of great contention, with coal miners pitted against corporations and the federal government (Smith 1981). Mesothelioma (a form of lung cancer) and asbestosis (a pulmonary disease) are now generally understood to be caused by asbestos, although there was great conflict until quite recently (Brodeur 1985).

Rachel Carson's *Silent Spring* (1962) sparked the first mass public attention to environmental health effects. Carson demonstrated how pesticides, especially DDT (dichlorodiphenyl-trichloroethane), caused morbidity and mortality in animals and humans. Her work led to significant pesticide regulation in the USA, the National Environmental Protection Act, the Environmental Protection Agency (EPA), and the new wave of the environmental movement.

The 1978 Love Canal situation was the first occasion on which human health was central in an environmental crisis, prompting the creation of the Superfund Program by the US EPA. Residents were warned by state health officials that toxic chemicals permeated their neighborhood in Niagara Falls, New York, where a chemical company had donated a dump site to the city to build a school. The revelation meshed with residents's awareness of a history of experiencing health effects. Residents and their scientist allies discovered elevated rates of miscarriage, birth defects, cancer, and chromosome damage (Levine 1982). The residents organizing to get action by state and federal governments led many other communities around the country to act similarly.

2. Types of Environmental Health Effect

2.1 Lay Discovery of Environmental Health Effects

Much of the knowledge about environmental health effects has come from actually or potentially affected people. Some situations involved immediate health effects, such as the dioxin-laced waste oil in Times Beach, Missouri (USA). Others involved episodes without immediate effects, but where effects were expected to follow, as at the Three Mile Island (TMI) nuclear plant. Sometimes there was the discovery of prior contamination episodes, such as in the case of radiation releases from the Hanford, Washington (USA) nuclear facility. Sometimes there are unexplained clusters of disease, of which people are seeking the cause, but do not yet have a specific emission to trace it to, as with Gulf War illnesses.

Environmental justice activism expanded these concerns as Blacks, Latinos, and Native Americans in the USA discovered excess lead poisoning and disproportionate placement of landfills, incinerators, and toxic waste sites in poor and minority areas (Bullard 1993, Bryant and Mohai 1992). Women's growing involvement in toxic waste activist groups has heightened the importance of gender issues (Blocker and Eckberg 1989).

Laypeople's success in identifying environmentally induced diseases and their likely causes is striking. A valuable example is the lay discovery of the health effects of the Hanford nuclear reservation. Citizens began with initial questioning, followed by extensive self-education through reading, talking with neighbors, querying officials, and analyzing primary sources. Ultimately, they compelled the release of massive numbers of documents, documented many accidental and planned radioactivity releases, identified thyroid disease among 'downwinders,' and were instrumental in US Energy Secretary Hazel O'Leary's 'openness' policy that brought to light many releases and accidents, as well as many unethical radioactivity experiments on captive populations (Kaplan 1997).

2.2 Physical Health Effects

People are exposed through many routes: water supply, contamination of foodstuffs and animal feed, chemical factory and transportation leakages and explosions, trash incineration, contamination from toxic waste sites, and exposure to household chemicals, fertilizers, and pesticides (see *Environmental Stress and Health*).

Toxic substances have been implicated in diseases of all organ systems. The physical health effects of toxins are not a separate category of symptoms and diseases, but are best understood as forms of poisoning. The

disputes over environmental health effects are therefore usually not about the symptoms *per se*, but about their etiology. Although a few toxins have 'signature' diseases (e.g., asbestosis and mesothelioma), most do not. Hence, neither sufferers nor clinicians are likely to attribute individual cases to a toxic exposure. There has also been intense debate on the presence of abnormalities that are not specific diseases, such as chromosome damage, eye blink reflexes, and reproductive abnormalities (Brown and Mikkelsen 1997, Colborne et al. 1996).

2.3 Mental Health Effects

Toxic contamination episodes often leave a residue of fear and anxiety, and case studies show that depression is common. In contrast with natural disaster situations, people have to fight to gain public recognition of the pollution in human-caused disasters. Litigation contributes to the ongoing depression by forcing people to relive their illness experience through the seemingly endless process of depositions, testimony, and consultations with independent experts. During the year after the TMI accident, mothers experienced new episodes of affective disorder at a rate three times higher than a nonexposed control group in another location (Bromet et al. 1982).

Post-traumatic stress disorder (PTSD) has been found in many toxic waste sites. People retained a 'death imprint' of memories of the disaster, and felt a strong 'death anxiety.' Many experienced 'death guilt' concerning what they thought they should have done to prevent a loved one's death (Lifton and Olson 1976).

Bruce Dohrenwend (1981) compared demoralization (the emotional state of the nonmentally ill general population) of TMI-area residents with the clientele of a community mental health center. With a higher score indicating more disorder, the overall mean score of the mental health center clients was 28.3, for TMI-area women 30, and for TMI-area men 25. Of those interviewed directly after the accident 26 percent showed severe demoralization. Even two years later, TMI residents had more symptoms, poorer task performance, and higher levels of serum catecholamines (a biochemical marker of stress) than did controls who lived near an undamaged nuclear facility (Baum et al. 1983).

People in contaminated communities often reassess the environment, seeing it as a malevolent force. Even the home may become threatening. Residents lose faith that the world is just, and may even think that they deserve their fate (Edelstein 1988). Victims lose trust in government and in the ordinary operations of the economy and polity, since the established institutions do not help them. This failure precipitates further emotional distress.

2.4 Community-wide effects

Contamination episodes harm communities, disrupting ordinary social relations as well as damaging health. Victims suffer loss of community bonds (Erikson 1976), have conflict with local government officials (Levine 1982), develop anxiety about their surroundings (Edelstein 1988), and may argue with other residents over what to do (Couch and Kroll-Smith 1992). Communities facing contamination episodes are often in the midst of 'chronic technological disasters' (Couch and Kroll-Smith 1985), long-term, unfolding problems that often haveno visible starting point, but which eat at the social fabric. These 'invisible environmental contaminants' are felt by residents as 'occult,' mysterious factors (Vyner 1988).

2.5 Social Responsibility

Disease sufferers and their allies target the corporations they hold responsible for contamination episodes. Residents have often expected local, state, and federal government support for their claims. Government failure to support victims's claims has led to increased mistrust of government. Residents in many contaminated communities have become radicalized in response to the combination of corporate pollution and government inaction (Szasz 1994). Toxic contamination episodes at a local level can have dramatic national effects. The dioxin release at Seveso in Italy led to major conflict in parliament over many issues, from remediation to abortion access for exposed women. In Japan, PCB (polychlorinated biphenyl) contamination of cooking oil resulted in national-level wrangling, as political parties took up the issue (Reich 1991). In the case of the Chernobyl meltdown in the USSR, there were many international repercussions as radiation passed over borders. Citizen response to toxic contamination has been a major element of popular social movements in the USA, at a time when social movements have otherwise been in decline. This has been less the case in other countries, perhaps because many of them already have powerful Green parties and other similar movements.

3. Issues and Conflicts in the Definition and Measurement of Environmental Health

3.1 Standards of Proof

Environmental health effects pose difficult research issues, such as inadequate history of the site, lack of clarity of the route of contaminants, determining appropriate sampling locations, small numbers of

cases, getting appropriate control groups, migration in and out of exposed areas, lack of knowledge about the characteristics and effects of certain chemicals, and unknown or varying latency periods for toxins. Epidemiology studies usually have such small numbers of exposed persons that observed rates must often be 20 times that of expected rates in order to attain statistical significance (Neutra 1990). New approaches argue that small effect sizes should not be viewed as obstacles, since even low relative risks are very powerful if large numbers of people are affected. Similarly, consistency across many studies should be acceptable as evidence of etiological linkages, even if not all of those studies meet statistical significance (National Research Council 1991). Ozonoff and Boden (1987) distinguish statistical significance from public health significance, since an increased disease rate may be of great public health significance even if statistical probabilities are not met. They believe that epidemiology should mirror clinical medicine more than laboratory science, by erring on the safe side of false positives. Although it is not always easy to evaluate the effects of removal of a hazard, researchers have produced such findings. In the well-known childhood leukemia cluster in Woburn, Massachusetts (USA), no new cases of leukemia developed after contaminated wells were shut, and birth weights returned to normal at the Love Canal and the Lipari, New Jersey (USA) landfill sites after exposure declined (National Research Council 1991).

Despite the many shortcomings in environmental epidemiology data, reviews of the research show that most controlled studies find associations between hazardous waste sites and human health (National Research Council 1991).

3.2 Other Obstacles to Knowledge on Environmental Health

In medical training, environmental medicine is relegated to a minor position (Institute of Medicine 1988). In one survey of all Massachusetts physicians, those whose practices are located in communities with Superfund sites were no more knowledgeable about environmental health hazards than other physicians (Brown and Kelley 1996). Medical centers and research institutions are not necessarily interested, since environmental health has low prestige and little funding, and can put medical facilities in conflict with local centers of power.

Corporations have fought against recognition of environmental health effects because of the financial cost involved in settlements, fines, production restructuring, and alternative forms of disposal. Government agencies have been reluctant to affirm most relationships between contaminants and disease. One reason is the fear that business will be harmed. Another is that governmental failures to regulate chemicals and other toxic substances are a source of public anger, and governments do not want to be criticized for their shortcomings.

3.3 Lay Involvement

Environmental health has become a major public concern due to dramatic environmental catastrophes such as the contamination at Love Canal, the dioxin exposure at Times Beach the dioxin explosion at Seveso, and the Union Carbide explosion of methyl isocyanate in Bhopal, India. Citizens were often the first to point out problems, and even when they were not, they typically took the lead in studying the situation. Lay-initiated health studies have been central to the development of environmental epidemiology.

However, many epidemiologists and government officials have criticized lay involvement, assuming that it must be biased. An example can be seen in the Woburn toxic waste crisis, where a major health survey was conducted by an alliance of Harvard biostatisticians and Woburn residents. They examined childhood leukemia, pregnancy outcomes, and childhood disorders from 5,010 interviews, covering 57 percent of residences with telephones. The researchers trained 235 volunteers to conduct the survey, taking precautions to avoid bias. Childhood leukemia was found to be significantly associated with exposure to water from two municipal wells. Use of that well water had been associated with perinatal deaths since 1970, eye/ear anomalies, and CNS/chromosomal/oral cleft anomalies. Water exposure was associated with kidney/urinary tract and lung/respiratory disease. This study would not have been possible without community involvement, yet government agencies and the American Cancer Society charged that the study was biased (Brown and Mikkelsen 1997).

4. New Methods and Measures Needed for Improved Knowledge and Action

Much environmental epidemiology is carried out in crisis situations, using very small samples. More routine monitoring is necessary, such as through regular national health surveys and toxic disease registries. Scholars need to decide what size of geographical area is most useful and must make clear connections between community contexts, environmental variables, and health outcomes. They need to decide what are the best health outcomes, both in terms of data access and in terms of plausible connections to environmental quality. Better environmental measures are needed to avoid the problem of existing data sources that give a wide array of units

of measurement for such items as toxic releases, type of releasing facilities, water quality, and air quality. Extracting data from the US EPA is cumbersome, requiring tedious Freedom of Information Act requests and lengthy waits. In the USA it is also important to link cross-state data, since researchers are unable to match exposure in one state with morbidity and mortality in another. Cancer incidence and mortality is not sufficient. Other less serious illness must also be studied.

Scholars need to come up with creative models that study health outcomes in relation to a variety of inputs, including political structure, public participation, access to health services, economic factors, population density, racial/ethnic proportions, environmental quality, and access to the natural environment.

See also: Crowding and Other Environmental Stressors; Environmental Stress and Health; Occupational Health and Safety, Regulation of; Workplace Safety and Health

Bibliography

Baum A, Fleming R, Singer J E 1983 Coping with victimization by technological disaster. *Journal of Social Issues* **39**: 117–38

Blocker J, Eckberg D L 1989 Environmental issues as women's issues: General concerns and local hazards. *Social Science Quarterly* **70**: 586–93

Brodeur P 1985 *Outrageous Misconduct: The Asbestos Industry on Trial*. Pantheon, New York

Bromet E J, Parkinson D K, Schulberg H C, Dunn L O, Gondek P C 1982 Mental health of residents near the Three Mile Island reactor: A comparative study of selected groups. *Journal of Preventive Psychiatry* **1**: 225–76

Brown P, Kelley J 1996 Physicians' knowledge of and actions concerning environmental health hazards: Analysis of a survey of Massachusetts physicians. *Industrial and Environmental Crisis Quarterly* **9**: 512–42

Brown P, Mikkelsen E J 1997 *No Safe Place: Toxic Waste, Leukemia, and Community Action*. University of California Press, Berkeley, CA

Bryant B, Mohai P (eds.) 1992 *Race and the Incidence of Environmental Hazards*. Westview Press, Boulder, CO

Bullard R (ed.) 1993 *Confronting Environmental Racism: Voices from the Grassroots*. South End Press, Boston

Carson R 1962 *Silent Spring*. Houghton-Mifflin, New York

Colborne T, Dumanoski D, Meyer J 1996 *Our Stolen Future: Are We Threatening Our Fertility, Intelligence, and Survival? A Scientific Detective Story*. Dutton, New York

Couch S R, Kroll-Smith J S 1985 The chronic technical disaster: Towards a social scientific perspective. *Social Science Quarterly* **66**: 564–75

Couch S R, Kroll-Smith J S (eds.) 1992 *Communities at Risk: Collective Responses to Technological Hazards*. Peter Lang, New York

Dohrenwend B 1981 Stress in the community: A report to the President's commission on the accident at Three Mile Island. *Annals of the New York Academy of Science* **365**: 159–74

Edelstein M 1988 *Contaminated Communities: The Social and Psychological Impacts of Residential Toxic Exposure*. Westview Press, Boulder, CO

Erikson K 1976 *Everything in its Path: The Destruction of Community in the Buffalo Creek Flood*. Simon and Schuster, New York

Institute of Medicine 1988 *Role of the Primary Care Physician in Occupational and Environmental Medicine*. National Academy Press, Washington, DC

Kaplan L 1997 The Hanford Education Action League: An informed citizenry and radiation health effects. *International Journal of Contemporary Sociology* **34**: 255–66

Levine A 1982 *Love Canal: Science, Politics, and People*. Heath, Lexington, MA

Lifton R, Olson E 1976 The human meaning of total disaster: The Buffalo Creek experience. *Psychiatry* **39**: 118

National Research Council 1991 *Environmental Epidemiology, Volume 1, Public Health and Hazardous Wastes*. National Academy Press, Washington, DC

Neutra R R 1990 Counterpoint from a cluster buster. *American Journal of Epidemiology* **132**: 18

Ozonoff D, Boden L I 1987 Truth and consequences: Health Agency responses to environmental health problems. *Science Technology and Human Values* **12**: 70–7

Reich M 1991 *Toxic Politics: Responding to Chemical Disasters*. Cornell University Press, Ithaca, NY

Smith B E 1981 Black lung: The social production of disease. *International Journal of Health Services* **11**: 343–59

Szasz A 1994 *Ecopopulism: Toxic Waste and the Movement for Environmental Justice*. University of Minnesota Press, Minneapolis, MN

Vyner H M 1988 *Invisible Trauma: The Psychosocial Effects of the Invisible Environmental Contaminants*. Lexington Books, Lexington, MA

P. Brown

Environmental History

1. The Term Environmental History

Environmental History deals with the history of human impacts on nature and the interactions between humans and nature. It asks how nature influences humans, how humans intervene in nature and how nature and humans interact. To be able to understand these processes, it also investigates changes in nature not caused by human action. The terms nature and environment are largely seen as being synonymous. In traditional, preindustrial societies generally they describe the natural environment, comprising elements such as the landscape, rivers, climate and weather, animals, plants, etc. These natural factors are also of importance for industrial societies. At the same time, their environments are shaped more and more by human action citing, for example, the environment of modern towns, in which today more than half of the

world's population lives. In addition, nature itself increasingly is shaped by human action. In Europe, there are hardly any and world-wide relatively few regions where humans did not interfere, so that in today's world we can speak of a man-made or even an artificial nature.

To understand such diverse developments the term environment as well as environmental history has to be defined very broadly. In practice, however, mostly it signifies the natural environment, being aware that in modern societies a pristine nature hardly exists any more. But this terminological ambiguity is acceptable since environmental history above all is interested in the interactions between humans and nature and the transitions between these two. In doing this kind of research, the contribution of other disciplines such as geography, geology, biology, and many others is important, and it is sometimes argued that environmental history as a matter of principle should adopt an interdisciplinary approach. In practice, however, interdisciplinary approaches have proved to be demanding so that most of the contributions to environmental history still come within from single disciplines, not just from historians, but also from geographers, biologists, social scientists, and others (Worster 1988).

2. *The Emergence of Environmental History*

The relationship between the natural environment and humans has been a topic in Western and other civilizations from ancient times (Glacken 1967). Considering the great importance of nature for preindustrial societies, it was understandably seen as a powerful agent, bringing about abundance, beauty, and health as well as need, disasters, and suffering. In this context, deterministic arguments developed, which were especially prominent in the nineteenth and early twentieth centuries explaining, for example, differences in societies, behavior, or racial characteristics as caused by the natural environment. Also prominent were traditions which placed humans into a wider cosmos of nature and propagated a harmony between these two. It is, however, difficult to tell whether these were idealistic concepts of a mainly religious or philosophical nature or to which degree they actually shaped the daily practice.

These long established approaches have not disappeared, but towards the end of the nineteenth century in industrial societies the role of nature became less important, especially in the academic debate. It was not until very recently that a new understanding developed, this time stressing the importance of human action in shaping the environment. An understanding of these influences had developed in the nineteenth century, but it was mainly after the 1950s that this approach became prominent and helped to establish the new field of environmental history as we know it today. At that time, it had become all too obvious how far reaching and often disastrous the effects of human intervention in the environment could be.

In the 1950s and 1960s environmental history was taken up as a subject particularly in the United States. It spread to Europe in the 1970s and 1980s and has expanded world-wide since the 1990s. Among the more interesting recent publications are studies on the environmental history of Asia (Elvin and Ts'ui-jung 1998, Grove et al. 1998), Africa (Beinhart and Coates 1995), and Latin America (Martínez and Jácome 1999). Now (at least) two journals in English exist which are specifically devoted to this topic: *Environmental History*, with an emphasis on the US, and *Environment and History*, whose outlook is more global, paying special attention to the non-Western world. In addition, more and more conferences are taking place; the number of publications is increasing rapidly, and environmental history is entering the curriculum at school and university level. This applies above all to the US where environmental history is most firmly established and where the American Society for Environmental History was founded in 1974 (Miller and Rothman 1997). Furthermore, the European Society for Environmental History is well established, while others are being formed. The field, therefore, is marked by rapid expansion, and the best way to get information about it is to consult the H-environment web page (www2.h-net.msu.edu/~environ/).

Given the debates of impending ecological crises, early environmental history developed a strong interest in its causes, which were often seen as a peculiar feature of the Western world. Among the explanations put forward were 'Christian roots of our current ecological crisis'; the rise of modern 'mechanistic' science; the repudiation of traditional customs and of a softer, female understanding of nature; and, of course, modern industrialization. Coming with these explanations was the notion that historically a balance between humans and nature had existed which was then destroyed by one or several of these factors (Merchant 1989, Cronon 1995).

For the existence respectively destruction of this balance different periods are named. A radical position blames the first appearance of humans; others emphasize the transition to agricultural societies, while mostly the development of the modern world, especially industrialization is held responsible, since it did not just destroy the existing balance and small-scale structures, but also initiated developments which brought about today's environmental problems: a rapid increase in population; the move to and the rise of big cities; the intensification of agriculture; the use of fossil fuels; a massive consumption of resources; the production of new, artificial materials; and new forms, and disastrously high levels of pollution. Numerous studies were undertaken showing in great detail how

human interference had altered the landscape; how the discharge and waste of industry, household, and cities had destroyed many forms of life; had endangered human health and had slowly but seemingly unavoidably contributed to the increasing ecological crisis (Turner and Clark 1990, Miller and Rothman 1997).

The results of these developments were felt globally and even began to affect the biosphere, putting at risk not just humans but all forms of existence. Against this background, conventional, that is, anthropocentric approaches were criticized and it was demanded they be replaced by biocentric notions which did more justice to animals and nature. Most radical amongst them is the Gaia-hypothesis. It sees the earth and the biosphere as a living organism which needs to be protected, but can also defend itself against human interference. Global warming is interpreted in this way. This thesis, however, was and still is a minority view. Much more widely shared is the argument that ecology has become so important that it has established a new scientific (and political) paradigm.

3. *Recent Developments and Debates*

Among the many developments in the recent debate within environmental history the following merit special attention. The concept of ideal, constant ecological balances is today seen an oversimplification. Rather, it is a feature of such balances that they are not stable and that none of them is more 'natural' than the other. Second, it has been shown that nature in Europe, North-America, China, and many other parts of the world has been altered by humans for several centuries, if not millennia, while the controversy of whether preindustrial societies existed in harmony or in conflict with their environment is often shaped by presentism. In this context, it is debated whether the concepts of nature and environment are also socially constructed. And finally, there are numerous indications that modern science, modern institutions, and modern form of productions are not solely harmful for the environment, but can also contribute to a better, more complex understanding of it.

3.1 *The Concept of a Stable Past*

For a time, environmentalists subscribed to the concept that natural processes culminated in stable balances and that these balances represented and ideal state. For some this ideal had existed some time in the past while others saw it as an objective to be realized in the future. Today, it is commonly accepted that the concept of ideal balances and of an ideal nature is misleading (Cronon 1995). Ecological balances are inherently instable, and there are no clear-cut ecological criteria except for human preferences to favor one or other. We may prefer the current climate to the ice-age, but ecologically speaking the ice-age may have been more stable. As a result we cannot avoid anthropocentric perspectives, since no ideal-state independent of our knowledge and preferences exists, and also since we cannot look at nature through the eyes of an animal, a plant, or the biosphere. Consequently, there have to be debates and value judgments about such preferences. Scientific knowledge or historical findings on their own cannot provide answers, but they offer valuable information.

In these debates, it has to be taken into consideration that humans have altered their environment for centuries. Even early societies such as hunters and gatherers effected great changes, for example, by using fire to clear trees or to hunt animals. Also, preindustrial societies could exploit their resources to such a degree that these ran out, thereby causing grave crises. A much debated example is the case of wood forests. For many societies, wood was the most important resource and was needed not only as fuel but also to build houses and ships, to produce tools, and numerous objects for everyday use. In many parts of Europe in the late eighteenth century, there were dramatic warnings that a shortage of wood was imminent. We now know that many of these warnings were grossly exaggerated, but behind the rhetoric, elements of a real crisis existed (Radkau 2000).

The probably best known example for far-reaching human interference in nature can be found in China, where it has taken place for almost 3,000 years (Elvin and Ts'ui-jung 1998). Other examples are the disappearance of forests in the Mediterranean (McNeill 1992) or in Scotland, where today a great deal of opposition exists to reforestation schemes in the barren Scottish Highlands—preferring the present state of nature. Similarly, the first nature reserve in Germany, the Lüneburg heath is the result of massive interference through deforestation and grazing by large numbers of sheep. Today, it requires constant management to keep it as it is.

3.2 *Ecological Imperialism and Traditional Structures*

The global dimension of ecological developments has been emphasized by the concept of ecological imperialism, most prominently put forward by Crosby (1986). He argues that Western imperialism succeeded not so much because of superior technology or arms, but because of the deadly viruses and bacteria the white conquerors brought with them. He describes an ecological clash of cultures which left the indigenous population largely helpless and brought about the death of millions of them. In the meantime, numerous other studies on the environmental consequences of the European expansion have appeared on, for ex-

ample, the consequences of British forest policy in India, the adaptation of European agrarian techniques, the environmental effects of the plantation-system, or of monocultures, etc. (Griffiths and Robben 1997, Grove et al. 1998).

Before these studies were published, the traditional methods of agriculture or pasture in Africa or Asia were often described as being harmful for the environment. The indigenous population, it was argued, had too little, if any knowledge and tended to overexploit the resources, for example, through overgrazing by keeping too many cattle or by destroying the remaining wild animals. As a consequence, nature reserves were established and local people often forced to settle elsewhere; new, Western technologies in agriculture were introduced, and the traditional methods widely discredited. It is interesting to note that similar arguments had been put forward in the eighteenth and nineteenth centuries in Europe and North America. Here, too, traditional methods and customs were first attacked as being inefficient and harmful to nature, and later almost completely abolished (Hughes 2000, Radkau 2000).

In the 1980s and 1990s these assumptions about traditional methods were increasingly challenged. At that time, many of the new, large-scale projects had failed; monocultures and the vast agricultural agglomerates had caused enormous ecological problems; and despite newly introduced plants and fruits crop failure as well as famines had continued to occur. In this context, environmental historians, too, (re)discovered the advantages of traditional methods and argued that these had been particularly well adapted to the differing local conditions; that they had not been based upon ignorance, but on knowledge accumulated over generations, and that they had been better suited to their environment than many of the modern techniques, which furthermore often went together with processes of centralization, thereby further alienating farmers from their roots. The new, centralized projects were heavily criticized and it was argued that they should be replaced by traditional, local practices. These were seen as having been part of a balance with the environment, destroyed by the white colonialists. Given the widespread disappointment with Western approaches, this point of view gained prominence, as it emphasized the negative, destructive consequences of modern science, bureaucratic centralization, the capitalist profit-motive, and market-oriented production. Similar arguments had been developed by environmental historians working on earlier periods of European and North American history. But in the former colonies these arguments had an important political edge and in addition corresponded to developments in the debate on imperialism, which also emphasized its destructive effects (Grove et al. 1998, Crosby 1986).

These debates have yielded important results. They have shown that traditional methods—far from damaging the environment—often were particularly well suited to the local and regional situation; or that small units often provide more scope to protect nature and the environment. At the same time, there was a tendency to portray too simplistic a picture. The mingling of species is not necessarily negative. Rather, it is inevitable, and in most cases nature and culture benefit from new biological combinations (Diamond 1997; Griffiths and Robben 1997). In addition, traditional methods, too, interfered with nature and were not always as environmentally-friendly as is often assumed. The same can be said about traditional institutions, for which the commons are a much debated example.

3.3 The Debate on the Commons

Until recently, this debate was dominated by the notion of the 'tragedy of the commons,' according to which the communal ownership of resources almost unavoidably leads to their exhaustion. For each individual it makes sense to use as much of the communal resources as possible. Even if they use less, there still is a tendency to exhaust the resources since everybody else would continue to maximize their own advantage. Therefore, so the argument goes, no inbuilt mechanism to preserve the commons exists, especially not in time of crisis. A similar argument was put forward in the eighteenth and nineteenth centuries, when agrarian reformers and (mainly big) landowners tried to abolish the still existing commons. In the recent environmental debate this notion has gained renewed prominence. The depletion of fish stocks in the oceans or the pollution of water, soil, and air are seen as proof that communal, that is, free goods tend to be exploited. As a result, property rights are advocated, arguing that private owners of goods have an inherent interest to guarantee their existence over a longer period of time.

Recent research by environmental and other historians has led to a reappraisal of the commons. Their findings show that different mechanisms of control and regulation to curtail individual excesses did in fact exist. Many commons survived for decades, others for centuries, indicating that efficient mechanisms to keep a stable balance must have existed. In addition, the establishment of private property rights could and did lead to rapid exhaustion if the new owners were interested primarily in a quick and high return and/or adopted methods such as monocultures not suited to the local conditions (Radkau 2000).

Not all commons existed for long periods of time. There were also cases of resources being exploited and the land becoming exhausted. Such cases seem to have increased in Europe towards the end of the eighteenth and especially during the nineteenth centuries, mainly due to the rapid increase in population. There are many indications that the traditional methods did not

offer enough scope to deal with this increase. This is not surprising, since historically speaking such a rapid increase is an exceptional phenomenon and exhaustions of resources did not usually occur. Preindustrial societies had a much less sophisticated and efficient technology, a lower level of science, or a less developed infrastructure, which limited the growth of the population, the economy and the exploitation of resources. As a result, interference with the environment was less widespread and far less damaging than after industrialization (Merchant 1989, Hughes 2000).

Considering today's environmental problems and the enormous population pressure the debate on the merits of commons respectively private property has lost some of its importance. There are historical examples both of exhaustion and of prudent care, so that an ideal-type juxtaposition of these two concepts is not helpful. The same can be said for the relationship between modern and traditional forms of the economy or of dealing with nature. Here, too, a juxtapositon is problematic not the least since historically the boundaries between the different spheres and approaches have been anything but clear cut. Practitioners of traditional methods were often influenced by new findings while the development of modern science was shaped by traditional assumptions and knowledge. This is also true for the colonies, even though here the confrontation between the two worlds was particularly intense. Very often, however, the existing structures, for example, in agriculture remained intact. And it has even been argued that the science of ecology originated in the colonial context where the indigenous plants and animals posed a challenge to colonial officials inducing them to have a closer look at the exotic environment in their postings (Grove 1995).

These and other examples show that since the 1980s more complex approaches and arguments have come to the forefront in environmental history. Here, as in other branches of historical studies, the identification of culprits and victims or the juxtaposition of traditional and modern has proved not to be very productive, not the least since the representatives of traditional methods largely were portrayed as victims, not as agents and historical subjects. The insistence on the complexity of historical processes, is, of course, not specific to environmental history, but here it has an added dimension. Given the enormous diversity of environments which in turn are shaped from place to place by very different natural factors, it is very difficult to come to general conclusions or to transfer findings from one place to another.

3.4 *Deforestation and* 'Waldsterben'

The advance of more complex arguments has also contributed to widely accepted assumptions being questioned, for example, the deforestation paradigm. It is a widely-held assumption that increased forest removal in mountainous areas enhances the frequency and severity of floods in the adjacent lowlands. Recurrent flooding in Switzerland in the nineteenth century established this paradigm in the scientific community as well as in public perception. In Germany, the flooding of the Rhine and the Oder in recent years are seen as further proof of this assumption. And in the Ganges delta the inhabitants blamed the clearing of forests in the Himalaya area for the recurrent floods in recent years. However, new research carried out in the floodplains of Bangladesh and in Switzerland has cast doubts upon these widely-held views. In Bangladesh, the deforestation paradigm disagrees with empirical facts, and Switzerland similar results exist, which show that the floods there corresponded to significantly higher rainfall.

In the case of the *Waldsterben*, too, recent findings have fundamentally challenged previous knowledge. In the 1980s the fear of dying forests was one of the major topics in the environmental debate, above all in Germany. It was agreed generally that as a result of air pollution, caused by industry, households, and traffic, most—some even argued all—of the forests would disappear. Now recent research has come to the conclusion that there probably never was such a thing as widespread *Waldsterben*. On the contrary, over the last decades trees in most parts of Europe have—apart from some very small areas—grown faster than before. It seems, that the thesis of the dying forest was based on a misleading research method (Brüggemeier 1998).

These and other findings show, that in environmental history, too, the concepts and terms used are socially constructed. Nature may be a powerful agent, guided by its own laws and often being beyond human influence, but it cannot speak for itself, so that to talk about nature or the environment, constructions are inevitable. At the same time, nature itself exists, of course, independent of such constructions so that environmental history can make a unique contribution to the current methodological debate on construction and reality.

3.5 *Recent Developments in Industrial Societies*

These examples should not give a wrong impression. There can be no doubt that we still face very serious environmental problems. But it seems that our knowledge of the historical processes and causes leading to them is vaguer than has long been realized, that these processes were often more complex, and that some of our assumptions have to be questioned, including the developments within industrial societies. Here, the period after the Second World War seems to be of special importance. On the one hand we have—beginning with the USA—the rise of the consumer societies, which use a greatly increased amount of

resources, above all energy, and produce markedly more emissions and waste (Pfister 1995). There are many indicators that the current environmental problems originated in this period and that their negative effects for the environment are still accumulating. On the other hand, it was also in this period that the environmental movement and environmental policies established themselves. It is still too early to judge their achievements but it can at least be debated whether these have been greater than is commonly perceived (Brüggemeier 1998, Simmons 1993).

When dealing with industrial societies a problem facing all environmental history becomes even more evident: how to judge human interference in the environment. In industrial societies, this interference is, of course, especially widespread and has destroyed much of the pre-existing nature. But will it do to see it only as having been destructive, or would it be more appropriate to emphasize the aspect of change? Are the terms man-made or artificial nature merely descriptive or do they stand for new concepts and new criteria? And what could these criteria be?

Undoubtedly, a lot of research still has to be done in environmental history. We know how some of today's major ecological problems developed, but the knowledge very often is incomplete and above all differs greatly from country to country. The same can be said about the environmental movements; changes in technology, law, and politics; the attitudes of political parties and institutions, of different groups in society; of the changing perception of environmental problems or of the many facetted attempts—and failures—to do something about them. If environmental history takes up these issues it will have to pay particular attention to international aspects on the level of causes as well as policy, taking account, for example, of the different international conferences and treaties of recent years McCormick 1989, Bailey 1995).

4. *Sustainability and Prospects*

For these topics and for environmental history in general the concept of sustainability could prove to be useful. It is true that this concept is very vague and that a great number of different, often conflicting definitions exist. However, many of them share three aspects which are interesting for environmental history: societies are sustainable, if they do not consume more renewable resources than grow every year; if they do not produce more emissions than the environment can absorb; and if of other resources they only consume as much as can be replaced by alternatives.

It is obvious that modern industrial societies do not fulfill any of these criteria. In fact, they are so far off the mark that one can doubt whether it makes sense to apply them. The same cannot be said of preindustrial societies. Almost by definition they depended practically completely on organic, that is, renewable resources; they produced very little emissions, and it could be argued—according to the three criteria mentioned above—that they were sustainable. At the same time, politically, economically, and socially, they were often in a precarious situation, given the recurrent crises such as famines or plagues. The enormous importance of environmental factors for these societies has been stressed in recent publications. These take up the long established debate on the influence of nature on human societies, but avoid its deterministic shortcomings and have again opened up a very important area of research (Diamond 1997).

Are ecological sustainability and social stability difficult to reconcile? This is an oversimplification, but the concept of sustainability could lead to new insights for environmental (and other) historians. At the same time environmental history could make important contributions to the current debate about the sustainability of industrial societies, which—in theory—is not limited to ecological concerns but also integrates political, economic, and social ones.

The development of environmental history over the last decades has not established a new paradigm nor has it closed the gap between science and the humanities. But, environmental history has raised important new questions; it has (re)introduced the topic of nature and the environment into historical studies, not only through specialist research, but especially by adding new perspectives to established areas and problems; and while it has not closed the gap to science, it has built many bridges. In short, environmental history has grown up. It will continue to flourish and make important contributions.

See also: Climate, History of; Deforestation–Forestation; Economic History; Environment and Development; Environmental Archaeology; Environmental Challenges in Organizations; Environmental Policy; Environmental Vulnerability; Environmentalism: Philosophical Aspects; Environmentalism: Preservation and Conservation; Global Environmental Change: Human Dimensions; Human Ecology; Human–Environment Relationships; Land Degradation

Bibliography

Bailey R (ed.) 1995 *The True State of the Planet*. Free Press, New York

Beinhart W, Coates P 1995 *Environment and History. The Taming of Nature in the USA and South Africa*. Routledge, London

Brüggemeier F J 1998 *Tschernobyl, 26 April 1986. Die ökologische Herausforderung*. Deutscher Taschenbuch Verlag, München, Germany

Cronon W 1995 (ed.) *Uncommon Ground: Toward Reinventing Nature*, 1st ed. Norton, New York

Crosby A W 1986 *Ecological Imperialism. The Biological Expansion of Europe, 900–1900*. Cambridge University Press, Cambridge, UK
Diamond J 1997 *Guns, Germs, and Steel. The Fates of Human Societies*, 1st ed. Norton, New York
Elvin M, Ts'ui-jung L (eds.) 1998 *Sediments of Time: Environment and Society in Chinese History*. Cambridge University Press, Cambridge, UK
Glacken C J 1967 *Traces on the Rhodian Shore. Nature and Culture in Western Thought from Ancient Times to the End of the Eighteenth Century*. University of California Press, Berkely, CA
Griffiths T, Robben L 1997 *Ecology and Empire: Environmental History of Settler Societies*. University of Washington Press, Seattle, WA
Grove R H 1995 *Green Imperialism. Colonial Expansion, Tropical Island Edens and the Origins of Environmentalism, 1600–1860*. Cambridge University Press, Cambridge, UK
Grove R, Damodaran V, Sangwan S (eds.) 1998 *Nature and the Orient. The Environmental History of South and South-East Asia*. Oxford University Press, Delhi, India
Hughes J D 2000 *The Face of the Earth: Environment and World History*. Sharpe, Armonk, NY
Martínez B G, Jácome A G 1999 *Estudios sobre historia y ambiente en América. I: Argentina, Bolivia, México, Paraguay*, 1st edn. Instituto Panamericano de Geografia e Historia, Mexico
McCormick J 1989 *The Global Environmental Movement: Reclaiming Paradise*. Belhaven Press, London
McNeill J R 1992 *The Mountains of the Mediterranean World: An Environmental History*. Cambridge University Press, Cambridge, UK
Merchant C 1989 *Ecological Revolutions. Nature, Gender, and Science in New England*. University of North Carolina Press, Chapel Hill, NL
Miller C, Rothman H 1997 *Out of the Woods. Essays in Environmental History*. University of Pittsburgh Press, Pittsburgh, PN
Pfister C (ed.) 1995 *Das 1950er Syndrom. Der Weg in die Konsumgesellschaft*. Paul Haupt, Bern, Stuttgart, Wien
Radkau J 2000 *Natur und Macht. Eine Weltgeschichte der Umwelt*. Beck, München, Germany
Simmons I G 1993 *Environmental History. A Concise Introduction*. Blackwell, Oxford, UK
Turner B L II, Clark W C (eds.) 1990 *The Earth as Transformed by Human Action*. Cambridge University Press, Cambridge, UK
Worster D (ed.) 1988 *The Ends of the Earth. Perspectives on Modern Environmental History*. Cambridge Univeristy Press, Cambridge, UK

F-J. Brüggemeier

Environmental Justice

Despite significant improvements in environmental protection over the last quarter of the twentieth century, millions of Americans continue to live in unsafe and unhealthy physical environments. Hardly a day passes without the media discovering some community or neighborhood fighting a landfill, incinerator, chemical plant, or some other polluting industry. This was not always the case. In the 1970s, the concept of environmental justice had not registered on the radar screens of environmental, civil rights, or social justice groups. Nevertheless, it should not be forgotten that Dr. Martin Luther King, Jr. went to Memphis in 1968 on an environmental and economic justice mission for the striking black garbage workers. The strikers were demanding equal pay and better working condition. Of course, Dr. King was assassinated before he could complete his mission.

1. *Birth of a New Movement*

The national environmental justice movement emerged around the issues of fairness, social equity, and environmental protection. This new movement embraces the principle that all communities are entitled to equal protection and enforcement of environmental, health, employment, housing, transportation, and civil rights laws and regulations that impact quality of life. The environmental justice framework rests on developing tools and strategies to eliminate unfair, unjust, and inequitable conditions and decisions. The framework also attempts to uncover the underlying assumptions that may contribute to and produce differential exposure and unequal protection. It brings to the surface the *ethical* and *political* questions of 'who gets what, when, why, and how much' (Bullard 1996). The national environmental justice movement/framework has its origins in the United States of America.

1.1 *Definition of Environmental Justice*

Environmental justice is defined as the fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies. Fair treatment means that no group of people, including racial, ethnic, or socioeconomic groups should bear a disproportionate share of the negative environmental consequences resulting from industrial, municipal, and commercial operations or the execution of federal, state, local, and tribal programs and policies (US Environmental Protection Agency (EPA) 1998).

1.2 *Early Struggles*

Another landmark garbage dispute took place in Houston, Texas, a decade after the Memphis garbage strike. African–American homeowners in Houston began a bitter fight to keep the Whispering Pines sanitary landfill out of their suburban middle-income neighborhood. The residents formed the Northeast Community Action Group or NECAG. In 1979, NECAG and their attorney, Linda McKeever Bullard,

Figure 1
Summit delegates hold a rally on the steps of the nation's capital, Washington, DC, 1991 (Photo by R D Bullard)

filed a class action lawsuit to block the garbage facility from being built. The 1979 lawsuit, *Bean vs. Southwestern Waste Management, Inc.*, was the first of its kind to challenge the siting of a waste facility under civil rights law.

The landmark Houston case occurred three years before the environmental justice movement was catapulted into the national limelight in the rural and mostly African–American Warren County, North Carolina. The environmental justice movement has come a long ways since its humble beginning in Warren County where a hazardous-waste landfill ignited protests and over 500 arrests. This marked the first time Americans had been arrested protesting the siting of a waste facility.

The Warren County protests provided the impetus for US General Accounting Office to conduct an independent investigation. That study revealed that three out of four of the off-site, commercial hazardous waste landfills in Region 4 (which comprises eight states in the South) were located in predominantly African–American communities, although African–Americans made up only 20 percent of the region's population (US General Accounting Office 1983). Nearly 15 years later, the state of North Carolina committed over $25 million to cleanup and detoxify the controversial Warren County landfill.

The protests also led the Commission for Racial Justice (1987) to produce *Toxic Waste and Race*, the first national study to correlate waste facility sites with demographic characteristics. Race was found to be the most potent variable in predicting where these facilities were located—more powerful than poverty, land values, and home ownership or occupancy status.

1.3 The People of Color Summit

The 1991 First National People of Color Environmental Leadership Summit was probably the most important single event in the young movement's history. Held in Washington, DC, the 4-day Summit was attended by over 650 grassroots and national leaders from around the world. Delegates came from all 50 states including Alaska and Hawaii, Puerto Rico, Chile, Mexico, and as far away as the Marshall Islands (Fig. 1).

The Summit broadened the environmental justice movement beyond its antitoxics focus and demonstrated that it is possible to build a multiracial

grassroots movement around environmental and economic justice. On September 27, 1991, Summit delegates adopted 17 'Principles of Environmental Justice.' By June 1992, Spanish and Portuguese translations of the Principles were being used and circulated by NGOs and environmental justice groups at the Earth Summit and Global Forum in Rio de Janeiro.

2. Government Response

The federal EPA manages, regulates, and distributes risks. During its 30-year history, the EPA has not always recognized that many government and industry practices (whether intended or unintended) have adverse impact on poor people and people of color. In 1992, after meeting with community leaders, academicians, and civil rights leaders, the EPA (under the Bush Administration) admitted there was a problem, and established the Office of Environmental Equity. The name was changed to the Office of Environmental Justice under the Clinton Administration.

The agency produced *Environmental Equity: Reducing Risks for All Communities*, *a* comprehensive report that examines environmental justice concerns in low-income and people of color communities (US Environmental Protection Agency 1992). The EPA also established a 25-member National Environmental Justice Advisory Council or NEJAC under the Federal Advisory Committee Act. The NEJAC is comprised of stakeholders representing grassroots community groups, environmental groups, nongovernmental organizations, state, local, and tribal governments, academia, and industry.

2.1 Environmental Justice Health Symposium

In February, 1994, seven federal agencies, including the Agency for Toxic Substances and Disease Registry (ATSDR), National Institute for Environmental Health Sciences (NIEHS), Environmental Protection Agency (EPA), National Institute of Occupational Safety and Health (NIOSH), National Institutes of Health (NIH), Department of Energy (DOE), and Centers for Disease Control and Prevention (CDC) sponsored a national health symposium, 'Health and Research Needs to Ensure Environmental Justice.' The February conference was attended by more than

Figure 2
President Clinton signs Executive Order 12898, Washington, DC, 1994 (Photo by White House Press Office)

1,000 scientists, researchers, health professionals, educators, government officials, and community stakeholders.

Environmental justice leaders offered their own environmental protection framework. This framework adopts a public health model of prevention (i.e., elimination of the threat before harm occurs) as the preferred strategy, shifts the burden of proof to polluters/dischargers who do harm, who discriminate, or who do not give equal protection to people of color, low-income persons, and other 'protected' classes, allows disparate impact and statistical weight or an 'effect' test, as opposed to 'intent,' to infer discrimination, and redresses disproportionate impact through 'targeted' action and resources (Fig. 2).

2.2 Executive Order 12898

In response to growing public concern and mounting scientific evidence, President Clinton on February 11, 1994 issued Executive Order 12898, 'Federal Actions to Address Environmental Justice in Minority Populations and Low-Income Populations.' The Order attempts to address environmental injustice within existing federal laws and regulations. It reinforces the Civil Rights Act of 1964, Title VI, which prohibits discriminatory practices in programs receiving federal funds. The Order also focuses the spotlight back on the 1969 National Environmental Policy Act (NEPA), a law to ensure for all Americans a safe, healthful, productive, and aesthetically and culturally pleasing environment (Council on Environmental Quality 1997).

3. Communities under Siege

A growing body of evidence reveals that people of color and low-income persons have borne greater environmental and health risks than the society at large in their neighborhoods, workplace, and playgrounds (Bryant 1995, Bullard 1999a, 1999b, 1999c, 2000, Stretesky and Hogan 1998). A recent study from the Institute of Medicine (1999) concluded that government, public health officials, and the medical and scientific communities need to place a higher value on the problems and concerns of environmental justice communities. The IOM study confirmed what most impacted communities have known for decades. People of color and low-income communities are (a) exposed to higher levels of pollution than the rest of the nation, and (b) experience certain diseases in greater numbers than the more affluent, white communities (Institute of Medicine 1999).

3.1 Unequal Protection

The *National Law Journal* uncovered glaring inequities in the way the federal EPA enforces its giant Superfund laws. White communities see faster action, better results, and stiffer penalties than communities where blacks, Hispanics, and other people of color live (Lavelle and Coyle 1992). Clearly, communities that are located on the 'wrong side of the tracks' are at special risk from exposure to environmental hazards. Race has been found to be independent of class in the distribution of air pollution, contaminated fish consumption, location of municipal landfills and incinerators, abandoned toxic waste dumps, polluting industries, and lead poisoning in children (Commission for Racial Justice 1987, Agency for Toxic Substances and Disease Registry 1988, Pirkle et al. 1994).

3.2 Childhood Lead Poisoning

Lead poisoning is the number one environmental health threat to children of color. Figures reported in the July 1994 *Journal of the American Medical Association* on the Third National Health and Nutrition Examination Survey (NHANES III) revealed that 1.7 million children (8.9 percent of children aged 1 to 5) are lead poisoned, defined as blood lead levels equal to or above 10 micrograms/deciliter. The NHANES III data found African–American children to be lead poisoned at more than twice the rate of white children at every income level (Pirkle et al. 1994). Over 28.4 percent of all low-income African–American children were lead poisoned compared to 9.8 percent of low-income white children. During 1976–1991, the decrease in blood lead levels for African–American and Mexican–American children lagged far behind that of white children.

In California, a coalition of environmental, social justice, and civil libertarian groups joined forces to challenge the way the state carried out its lead screening of poor children. The Natural Resources Defense Council, the National Association for the Advancement of Colored People Legal Defense and Education Fund (NAACP LDF), the American Civil Liberties Union, and the Legal Aid Society of Alameda County, California won an out-of-court settlement worth $15 million to $20 million for a blood lead-testing program. The lawsuit, *Matthews vs. Coye*, involved the failure of the state of California to conduct federally mandated testing for lead of some 557,000 poor children who receive Medicaid. This historic agreement triggered similar lawsuits in several other states that failed to live up to the federal mandates.

4. The Path of Least Resistance

The southern US has become a 'sacrifice zone' for the rest of the nation's toxic waste. Lax enforcement of environmental regulations has left the region's air, water, and land the most industry-befouled in the US. Louisiana typifies this pattern. Nearly three-fourths of Louisiana's population, more than 3 million people,

Figure 3
St. James Citizens for Jobs and the Environment and allies hold a press conference in opposition to Shintech's proposed polyvinyl chloride plant, Convent, LA., 1997 (Photo by Environmental Justice Resource Center)

get their drinking water from underground aquifers. Dozens of aquifers are threatened by contamination from polluting industries. The Lower Mississippi River Industrial Corridor has over 125 companies that manufacture a range of products including fertilizers, gasoline, paints, and plastics. Environmentalists and local residents have dubbed this corridor 'Cancer Alley' (Motavalli 1998, Bullard 2000).

4.1 Corporate Welfare

Louisiana is also a leader in doling out corporate welfare to polluters. Industries routinely pollute the air, ground, and drinking water while being subsidized by tax breaks from the state. In the 1990s, the state wiped off the books $3.1 billion in property taxes to polluting companies. The state's top five worse polluters received $111 million dollars over the past decade (Barlett and Steele 1998) (Fig. 3).

Subsidizing polluters is not only bad business, but it does not make environmental sense. A growing body of evidence reveals that 'states with lower pollution levels and better environmental policies generally have more jobs, better socioeconomic conditions and are more attractive to new businesses' (Templet 1995). Louisiana and other states could actually improve their general welfare by enacting and enforcing regulations to protect the environment.

4.2 Dumping on Native Americans

It should not be a surprise to anyone to discover that Native Americans have to contend with some of the worst pollution in the US (Tomsho 1990, Taliman 1992). Native American nations are prime targets for waste trading. More than three dozen Indian reservations were targeted for landfills, incinerators, and other waste facilities in the early 1990s. Native American groups and their allies defeated most of these waste proposals.

Racism operates in energy production (mining of uranium) and disposal of wastes on Indian lands (Churchill and LaDuke 1983). The practice has left many sovereign Indian nations without an economic

infrastructure to address poverty, unemployment, inadequate education and health care, and a host of other social problems. In 1999, residents of the Eastern Navajo reservation filed suit with the Nuclear Regulatory Commission to block a permit for uranium mining in Church Rock and Crown Point, New Mexico.

4.3 Transboundary Waste Trade

Hazardous waste generation and international movement of hazardous waste pose some important health, environmental, legal, and ethical dilemmas. Shipping hazardous wastes from rich communities to poor communities is a global problem. In the real world, all people, communities, and nations are not created equal. Unequal interests and power arrangements have allowed poisons of the rich to be offered as short-term remedies for poverty of the poor. This scenario plays out domestically (as in the US where low-income and people of color communities are disproportionately impacted by waste facilities and 'dirty' industries) and internationally (where hazardous wastes move from OECD states to non-OECD states).

The conditions surrounding the more than 2,000 maquiladoras, assembly plants operated by American, Japanese, and other foreign countries, located along the 2,000-mile US–Mexico border may further exacerbate the waste trade. The industrial plants use cheap Mexican labor to assemble imported components and raw material and then ship finished products back to the US. Nearly a half million Mexican workers are employed in the maquiladoras.

All along the Lower Rio Grande River Valley maquiladoras dump their toxic wastes into the river, from which 95 percent of the region's residents get their drinking water (Hernandez 1993). In the border cities of Brownsville, Texas and Matamoras, Mexico, the rate of anencephaly—babies born without brains—is four times the national average. Affected families have filed lawsuits against 88 of the area's 100 maquiladoras for exposing the community to xylene, a cleaning solvent that can cause brain hemorrhages, and lung and kidney damage.

The Mexican environmental regulatory agency is understaffed and not equipped adequately to address the problem (Barry and Simms 1994). Only time will tell if the North American Free Trade Agreement (NAFTA) will 'fix' or exacerbate the economic, public health, and the environmental problems along the US–Mexico border.

5. Conclusion

The environmental justice movement emerged in response to environmental and social inequities, threats to public health, unequal protection, differential enforcement, and disparate treatment received by the poor and people of color. The movement redefined environmental protection as a basic right. It also emphasized pollution prevention, waste minimization, and cleaner production techniques as strategies to achieve environmental justice for all Americans without regard to race, color, national origin, or income.

The poisoning of African–Americans in Louisiana's 'Cancer Alley,' Native Americans on reservations, and Mexicans in the border towns all have their roots in a system characterized by economic exploitation, racial oppression, and devaluation of human life and the natural environment. Unequal political power arrangements have allowed poisons of the rich to be offered as short term economic remedies for poverty and unemployment. Having industrial facilities in one's community does not automatically translate into jobs for nearby residents. Similarly, tax breaks and corporate welfare programs have produced few new jobs by polluting firms.

The environmental justice movement has established clear goals of eliminating unequal protection. This call for environmental and economic justice does not stop at the US borders but extends to communities and nations that are threatened by the export of hazardous wastes, toxic products, and 'dirty' industries. Environmental justice leaders are demanding that no community or nation, rich or poor, urban or suburban, black or white, should be allowed to become a 'sacrifice zone' or dumping grounds. They are also pressing governments to live up to their mandate of protecting public health and the environment.

See also: Environment and Development; Environmental and Resource Management; Environmental Change and State Response; Environmental Health and Safety: Social Aspects; Environmental Risk and Hazards; Environmental Stress and Health; Environmental Vulnerability; Environmentalism: Philosophical Aspects; Environmentalism, Politics of; Neighborhood Revitalization and Community Development; Public Health; Social Movements: Environmental Movements; Sustainable Development

Bibliography

Agency for Toxic Substances and Disease Registry 1988 *The Nature and Extent of Lead Poisoning in Children in the United States: A Report to Congress*. US Department of Health and Human Services, Atlanta, GA

Barlett D L, Steele J B 1998 Paying a price for polluters. *Time* **(November 23)**: 72–80

Barry T, Simms B 1994 *The Challenge of Cross Border Environmentalism: The US–Mexico Case*, 1st edn. Resource Center Press, Albuquerque, NM

Bryant B 1995 *Environmental Justice: Issues, Policies, and Solutions*. Island Press, Washington, DC

Bullard R D (ed.) 1996 *Unequal Protection: Environmental Justice and Communities of Color*. Sierra Club Books, San Francisco

Bullard R D 1999a Dismantling environmentalism in the USA. *Local Environment* **4**: 5–19
Bullard R D 1999b Building just, safe, and healthy communities. *Tulane Environmental Journal* **12**: 373–404
Bullard R D 1999c Leveling the playing field through en vironmental justice. *Vermont Law Review* **23(Spring)**: 454–78
Bullard R D 2000 *Dumping in Dixie: Race, Class and Environmental Quality*, 3rd edn. Westview Press, Boulder, CO
Churchill W, LaDuke W 1983 Native American: The political economy of radioactive colonialism. *Insurgent Sociologist* **13(1)**: 51–63
Commission for Racial Justice 1987 *Toxic Wastes and Race in the United States*. Public Data Access, New York
Council on Environmental Quality 1997 *Environmental Justice: Guidance under the National Environmental Policy Act*. CEQ,Washington, DC
Hernandez B J 1993 Dirty Growth *The New Internationalist* **August**
Institute of Medicine 1999 *Toward Environmental Justice: Research, Education, and Health Policy Needs*. National Academy Press, Washington, DC
Lavelle M, Coyle M 1992 Unequal protection. *The National Law Journal* **September 21**: 1–2
Motavalli J 1998 Toxic targets: Polluters that dump on communities of color are finally being brought to justice. *E Magazine* **9 (July/August)**: 28–41
Pirkle J L, Brody D J, Gunter E W, Kramer R A, Paschal D C, Glegal K M, Matte T D 1994 The decline in blood lead levels in the United States: The National Health and Nutrition Examination Survey (NHANES III). *Journal of the American Medical Association* **272**: 284–291
Stretesky P, Hogan M J 1998 Environmental justice: An analysis of superfund sites in Florida. *Social Problems* **45(May)**: 268–87
Taliman V 1992 Stuck holding the nation's nuclear waste. *Race, Poverty & Environment Newsletter* **(Fall)**: 6–9
Templet P H 1995 The positive relationship between jobs, environment and the economy: An empirical analysis and review. *Spectrum* **(Spring)**: 37–49
Tomsho R 1990 Dumping grounds: Indian tribes contend with some of the worst of America's pollution. *The Wall Street Journal* **(November 29)**
US General Accounting Office 1983 *Siting of Hazardous Waste Landfills and Their Correlation with Racial and Economic Status of Surrounding Communities*. Government Printing Office, Washington, DC
US Environmental Protection Agency 1992 *Environmental Equity: Reducing Risk for All Communities*. EPA, Washington, DC
US Environmental Protection Agency 1998 *Guidance for Incorporating Environmental Justice in EPA's NEPA Compliance Analysis*. EPA, Washington, DC

R. Bullard

Environmental Planning

For millennia, humans have realized the importance of conserving important natural resources and minimizing the damage that their activities cause to the natural environment in which they live. They have followed two broad strategies to protect environmental resources and reduce or eliminate these impacts. The first is to shape land use patterns so as to protect people against natural hazards such as floods, to prevent inappropriate consumption of natural resources, and to segregate or locate appropriately activities that pollute the environment. The second is to control emissions or discharges directly.

Environmental planning embraces those activities of government agencies and private organizations whose primary purpose is to manage the impacts of the human use of land on the natural environment, by taking thought for the future and choosing courses of action that promote appropriate goals. In parallel with environmental planning, environmental pollution control, usually in the form of media-specific laws such as the United States' Clean Air and Clean Water Acts, seeks to control discharges, emissions, and environmental pollution in general by a variety of means, the most important of which will be summarized below. While obviously important, cultural, social, psychological, and economic considerations are not the focus of environmental planning.

Because of the scope and complexity of the activities it addresses and their far-reaching and variable impacts, the field of environmental planning defies easy definition. Furthermore, conceptions of the relationships between humanity and the environment change over time, so that the character, goals, methods, and, indeed, the ethical foundations of environmental planning, are continuously evolving. At the beginning of the twenty-first century environmental planners carry out such varied responsibilities as analysis of the natural environment as a component for comprehensive plans; the preparation of environmental impact statements for governmental programs; the drafting of plans for recreational areas, parks, and other types of open space; and the development policies for the protection of prime farmland, wetlands, flood plains, and other natural resource areas.

1. *Brief History of Environmental Planning*

Although the term 'environmental planning' is of relatively recent origin, humankind has undertaken the activities it now embraces for millennia. At least five thousand years ago, the earliest urban civilizations—the 'hydraulic' societies of Sumer, Egypt, and the Indus Valley—undertook major programs to harness natural forces. Large-scale irrigation projects provided the water needed to supply the broad agricultural foundation on which the urban civilizations were built. Certainly, the Romans engaged in a substantial amount of what we would today include in environmental planning as they built extensive public works, such as the dams, reservoirs, aqueducts, and sewerage systems that served their far-flung cities.

In Islamic cities from Morocco to Iran, abattoirs, tanneries, and textile dyeing operations were relegated to the outer, downstream, edges of the medina (the old city), far from the Grand Mosque, the central souk (market), and the principal residential areas. Across the Islamic world after the seventh century, gardens were associated with paradise, and many palaces contained elaborate gardens that gave their owners respite from the harshness of the climate. Superb examples such as the fourteenth century Alhambra palace of the Nasrid caliph, Yusuf I, in Grenada, Spain, (Hoag 1977), are creations that fall within the broad definition of environmental planning.

In eleventh- and twelth-century Europe, after a long hiatus following the collapse of the Roman Empire in the fifth century, Italian cities adopted municipal codes that regulated both the disposal of wastes (forbidding, for example, their discharge within the town) and the number and location of pollution-generating activities (Zupko and Laures 1996). A growing interest in nature found expression in the Renaissance gardens of Florence, Rome, and Tivoli (McHarg 1992). Here, in the sixteenth century, artists such as Bramante, Raphael, and Palladio imposed a simple geometry on the forms of nature. A century later, landscape gardeners such as André Le Nôtre developed formal Baroque gardens at Versailles and Vaux-le-Vicomte, that 'testified to the divinity of man and his supremacy over a base and subject nature' (McHarg 1992). A strongly divergent tradition developed in England in the eighteenth century. Landscape gardeners such as Lancelot 'Capability' Brown and William Kent created the famed English landscape for their aristocratic clients, a setting for their manor houses that appeared natural but was as much an artifice of the human mind as the formal gardens of Paris and Rome (McHarg 1992). Nature and its forms became something to be respected and idealized.

In the nineteenth century, interest in public health mounted, especially in the link between sewage discharges from rapidly growing cities, pollution of their water supplies, and outbreaks of typhoid and dysentery. Many major cities undertook the construction of public water supply and sewerage systems. In 1872, responding to a growing interest in the conservation of critically important natural areas, during the administration of President Grant, the United States created Yellowstone National Park, the first in what was to become an impressive system of parks. Frederick Law Olmsted designed a number of parks, the best known of which is Central Park in New York City.

The American city planning movement, which began in the early twentieth century, drew inspiration from the public health, housing, and City Beautiful movements, and from early German efforts to zone land according to its most appropriate land use. But it was concerned more with the management of new development than it was with the moderation of human impacts on the natural environment and the conservation of natural areas. Environmental concerns took a back seat during World War I, the Roaring Twenties, the Depression, World War II, and the reconstruction and normalization years immediately thereafter. After World War II, Aldo Leopold called for a new kind of ethic that called on humankind to serve as the steward of nature, not its sovereign (Leopold 1949), and in 1962 Rachel Carson's *Silent Spring* fired the American imagination with her chronicle of the devastation that country's profligate waste disposal, pesticide use, and weak environmental policies were wreaking on its environment (Carson 1962). Ian McHarg, winner of the prestigious 1999 Japan Prize for City Planning, argued eloquently and persuasively that plans for the development and redevelopment of rural and urban areas must be based on a comprehensive inventory of ecological data that presents a complete picture of the geological, hydrological, pedological, limnological, physiographic, climatological, and biological systems in the area that is the subject of the plan (McHarg 1992).

It was not until the 1960s and the early 1970s that a newly felt concern for the environment, and a deepening awareness of the damage that humankind was inflicting on it, resonated with something deep in people's souls. This found expression in the United States and across the world, in the enactment of a substantial body of environmental protection laws. More than in any preceding era, public and private decision-makers have been called upon to take into account the environmental impacts of the various policies and infrastructural systems they propose.

2. *The Intellectual Context of Environmental Planning*

At the beginning of the twenty-first century, environmental planning lies at the juncture of two major fields of policy formulation and implementation: comprehensive growth management and environmental pollution control.

Comprehensive growth management seeks to advance a broad spectrum of goals and objectives. It addresses the location, density, use, height and bulk, and phasing or timing of urban development and redevelopment, and the provision of various infrastructural systems. At the same time, it seeks to protect critically important natural resources, such as farmland, flood plains, marshes, and other areas of critical ecological concern. Since 1970, nation after nation and governmental jurisdiction after governmental jurisdiction have come to recognize that any effort to manage urban and regional development and redevelopment must be grounded in an understanding of the natural environment within which they are occurring. Comprehensive planning takes into account the underlying geology of the area, the impact of proposed development on ground water, rivers, lakes, wetlands,

and coastal waters, the impact of manufacturing and transportation activities on air quality, the effects of urbanization on the natural biota, the impacts of suburban development on the supply of prime agricultural land, and the effects of solid and liquid waste disposal on the natural environment. Environmental planning is a key component of comprehensive growth management.

Environmental pollution control policies are broadly concerned with air, water, and earth pollution, and such global problems as loss of biodiversity, global warming, and depletion of the stratospheric ozone level. Decades of contemporary efforts to protect the environment and reduce environmental pollution have confirmed that humanity and the natural environment are interdependent. Changes in the natural environment can cycle back and cause serious harm to mankind. Transportation, land development, and land use affect environmental quality, whether it be by covering over wetlands, changing the hydrological characteristics of watersheds so as to increase flooding and aggravate drought conditions, converting farm land to residential developments and shopping centers, or creating automobile-dominated transportation systems that generate a major proportion of dangerous environmental pollutants. Environmental pollution controls often include the management of land development and the design of transportation systems so as to reduce pollution. Environmental planning, the management of land development, and the design of transportation systems are key components of environmental pollution control.

3. *Emphasis in Current Theory and Research*

In the field of environmental planning, there are a number of areas of research that are attracting interest from academics and practitioners.

3.1 *Research on the Scientific Basis of Environmental Planning*

Considerable work is being done in the natural sciences and ecology to strengthen the scientific underpinnings of an effective environmental planning effort. Some subjects are fairly specific, such as determining a river's flood plains for differing flood recurrence intervals, or developing more cost-effective ways to clean polluted industrial sites. Others are much broader, such as determining whether the global warming that the world is currently experiencing is the cause of the rise in atmospheric carbon dioxide concentrations, or vice versa, and determining the nature and probability of the effect of various hazardous materials on the human body.

3.2 *Implementing Environmental Planning*

Environmental planning embraces more than the various studies that environmental planners perform in the process of preparing a comprehensive plan. It includes a careful assessment and selection of the most appropriate strategies of implementing the various goals of environmental planning and protection. Among the most important, with examples from the American experience, are:

3.2.1 Strategies based on disclosure. Environment assessment and impact statements such as those mandated by the United States' National Environmental Policy Act of 1969 (NEPA), and numerous state counterparts, with respect to major governmental programs have a significant effect on the environment. This enables affected governmental agencies and citizen groups to find out about proposed actions and to assert their influence to modify their outcome.

3.2.2 Strategies based on increased intergovernmental coordination, often accompanied by greater public participation in the decision-making process. For instance, the American Intergovernmental Coordination Act and NEPA call for collaboration among the various governmental agencies with responsibilities in a particular region.

3.2.3 Command and control strategies. These approaches have long been the workhorse of environmental planning and protection. A government establishes a regulation that restricts private activity, such as a flood plain zoning ordinance or a statute seeking to reduce air pollution by setting permissible emission levels. This regulation sets a standard with which the private actor must comply, and then applies sanctions to those who violate the standard. Command and control strategies may take a number of forms, from simply specifying certain environmental standards, to controlling market access, to prohibiting certain kinds of actions, to requiring that firms internalize formerly externalized environmental costs.

3.2.4 Strategies based on market-based incentives. Many American jurisdictions have experimented with techniques that seek to influence private actors' behavior by providing economic incentives or disincentives that induce citizens to change their behavior simply because it is more profitable. 'Transferable

development rights' (TDR) systems allow landowners whose land has been severely regulated to sell to other landowners the right to build at somewhat higher density than they would be permitted to without the TDRs (see Sect. 3.4). Also, the 1990 amendments to the United States' Clean Air Act established a system for trading emissions allowance for oxides of sulfur (SO_x) and nitrogen (NO_x) that permits electric utilities to meet statutory requirements either by reducing their emissions by installing more effective air pollution control devices or by buying emission rights from other producers, whose emissions are well below their assigned level (see Sect. 3.4). Finally, many nations have created tax incentives that seek to increase after-tax income of individuals and companies that engage in conduct that is either environmentally protective or environmentally enhancing.

3.3 Sustainability

The 1989 Report of the World Commission on Environment and Development (The Brundtland Commission) proposed a summary definition of sustainable development: 'development that meets the needs of the present without compromising the ability of future generations to meet their own needs.' This oft-quoted exemplary statement articulates the basic principle that we should live off the income of our natural capital and not consume non-renewable resources in a profligate manner. As such, it focused heavily on the natural environment and called for a policy of wise stewardship of the world's natural resources and a reduction in the amount of pollution that humankind was discharging into the environment. The concept of sustainability has since been expanded to include profound social, political, and equitable considerations, such as (Newman and Kenworthy 1999; see also, Beatley and Manning 1997):

(a) 'The elimination of poverty, especially in the Third World, is necessary not just on human grounds, but also for environmental reasons.' Third World economic and social development, with the attendant reduction in rates of population increase and a reduction in international strife, is a necessary prerequisite for achieving a just and balanced steady world state.

(b) 'The First World must reduce its consumption of resources and the production of wastes.' First World citizens consume many times as many natural resources and discharge much greater amounts of pollutants into the environment than do their Third World counterparts. If Third World standards of living are raised without a corresponding lowering of the material consumption of the First World, the effects on the environment will be devastating.

(c) 'Global cooperation on environmental issues is no longer a soft option.' We have already recorded potentially irreversible changes in the environment, such as the loss of a number of species and a decrease in biodiversity, and the reduction in the stratospheric ozone layer. One respected scientist has predicted that, unless we drastically change our policies toward tropical rain forests, as many as one-quarter of the world's species will have become extinct by the year 2020 (Wilson 1992). Humanity must make difficult decisions or risk devastating systemic changes in the environment that could well lead to the extinction of the human race.

(d) 'Change toward sustainability can occur only with community-based approaches that take local cultures seriously.' Many of the activities that have adverse impacts on the environment are managed locally, so that local environmental planning strategies comprise a critical and fundamental element of a policy that moves toward sustainability.

The growing recognition of the central importance of achieving a policy of sustainable development so that human activities do not irretrievably disrupt basic ecological systems is reflected in the creation of governmental institutions such as the United States' President's Commission on Sustainable Development and in numerous United Nations and international conferences, such as the Rio Declaration, coming out of the United Nations Conference on Environment and Development in Rio de Janeiro in 1992.

More and more, environmental planning will be premised on principles of sustainable development at all scales. For instance, the automobile-based urban sprawl that characterized late twentieth-century urban development in many countries is profligate in that it consumes much more prime farmland and areas of critical environmental concern than a more compact, mass-transit-oriented pattern of development would. It requires the use of resources to construct new infrastructure systems, at the periphery of our metropolitan areas, while under-utilizing the installed capacity that already exists in our inner cities. Here environmental plans are critical components of comprehensive growth management systems for designing and creating new urban communities.

3.4. Market-based Incentives: Economic Dimensions of Environmental Planning

Another area that is the subject of considerable research and experimentation is the realignment of economic incentives to encourage actions that conserve the environment and discourage actions that damage it. At the national and international scale, analysts such as Daly and Cobb (1989) and Repetto and McGrath (1989) advocate the revision of the system of national accounts that is used to determine gross domestic product. Late twentieth-century practices do not create national capital asset accounts that

include natural resources, such as forests, minerals, and prime agricultural land. Nor do they depreciate these assets when timber is harvested, minerals extracted, or rich soil lost to erosion. Instead, the wood, coal, iron ore, and other raw materials are registered as income when they are sold, with no corresponding entry for the draw down in the capital asset. This is in stark contrast to buildings and other man-made capital assets, which are depreciated each year because their cost is an important input into production. Other proposals include the imposition of 'green taxes' on activities that pollute the atmosphere. The idea is to shift the incidence of major taxes from 'goods,' such as income and real property, to 'bads,' such as environmental pollution and energy derived from nonrenewable natural resources (Hawken 1993). Along the same lines is the successful American program for reducing emission of oxides of sulfur and nitrogen by large electric power generating plants by setting up a market on which emissions allowances for these pollutants may be traded. In an effort to introduce more decisional flexibility than has been possible under traditional 'command and control' methods of regulation air emissions, the US Environmental Protection Agency assigned each of some 111 utilities in the Midwest and East regions of the country a maximum emission level of tons of SO_x and NO_x per million BTUs of energy produced. If a company was able to reduce its emissions below the prescribed level, it could sell the balance of its permissible emissions to a company that was producing more than its allotment, permitting the latter to exceed that level by the amount of the emissions purchased. The program introduces decisional flexibility because plant managers are enabled to calculate the most efficient way to meet tightening environmental standards. This program has been widely credited with making it possible for electric utilities to reduce their emissions of these two pollutants by more than 50 percent in the ten years since the program was set up. The 1997 United Nations Framework Convention on Climate Change at Kyoto produced a Protocol that proposed to apply these ideas to emissions of carbon dioxide and other so-called 'greenhouse gases,' in an effort to address the build-up of these gases in the atmosphere.

In the field of environmental planning, TDR programs, which exhibit certain similarities to the trading of emission allowances of SO_x and NO_x, have been used in a number of jurisdictions to protect prime farmland and areas of special environmental significance, such as flood plains, beaches, and marshes. The local government imposes zoning limitations on key land such as farmland or flood plains that reduce the density at which they may be developed and then assigns the owner a certain number of development rights, usually calculated on a per acre basis. The owner may sell these rights to a developer who owns property in an area where the government wishes to encourage more dense development. The developer will be able to develop at a higher density than would have been possible, had he not purchased the TDRs.

The general objective of these market-based incentives is to shift from primary reliance on the command and control system of managing urban growth and protecting environmental resource lands to an approach that uses economic incentives and disincentives to encourage environmentally beneficial activities, and discourage activities that despoil the natural ecosystem.

3.5 Environmental Justice

Another dimension of environmental planning that has long been of concern but has become more central in the last decade or so—and also adumbrated in the *Brundtland Report*—is that of equity: Are the environmental benefits and burdens of particular projects distributed evenly throughout the various segments of society? Or do the benefits tend to accrue to the more prosperous and powerful, and the burdens to the poor and those citizens who are members of disadvantaged minority groups? At the local level, this may involve the decision to locate a trash-to-steam plant near a poor African–American community. At the international level, it may concern the construction of a major waste tire disposal facility or of a 'ship-breaking' operation in a developing country that is in desperate need of international exchange.

Considerations of environmental justice lead to determining the range of the detrimental impacts from a proposed project, and then defining the various communities that exist within that range. If all or most of the people within the range are poor or members of minority groups, the actor should seek alternatives.

See also: Environment and Development; Environmental and Resource Management; Environmental Change and State Response; Environmental Justice; Environmental Policy; Environmental Policy: Protection and Regulation; Environmentalism: Preservation and Conservation; Planning Ethics; Planning Issues and Sustainable Development; Republican Party

Bibliography

Beatley T, Manning K 1997 *The Ecology of Place: Planning for Environment Economy, and Community*. Island Press, Washington, DC

Carson R 1962 *Silent Spring*. Fawcett, Greenwich, CT

Daly H, Cobb J 1989 *For the Common Good: Redirecting the Economy Toward Community, the Environment, and a Sustainable Future*. Beacon Press, Boston

Hawken P 1993 *The Ecology of Commerce*. Harper Business, New York

Hoag J 1977 *Islamic Architecture*. Rizzoli, New York
Leopold A 1949 *A Sand County Almanac*. Oxford University Press, New York
Marsh W M 1998 *Landscape Planning: Environmental Applications*. Wiley, New York
McHarg I 1992 *Design with Nature*. Wiley, New York
Newman P Kenworthy J 1999 *Sustainability and Cities: Automobile Dependence*. Island Press, Washington, DC
Plater Z, Abrams R, Goldfarb W, Graham R 1999 *Environmental Law*. Michie and Co., Charlottesville, VA
Repetto R, McGrath W 1989 *Wasting Assets: Natural Resources in the National Income Accounts*. World Resources Institute, Washington, DC
Wilson E 1992 *The Diversity of Life*. Norton, New York
World Commission on Environment and Development 1989 *Our Common Future*. Oxford University Press, Oxford
Zupko R, Laures R 1996 *Straws in the Wind: Medieval Urban Environmental Law*. Westview Press, New York

J. C. Keene

Environmental Policy

1. *History and General Characteristics of Environmental Policy*

1.1 *History*

For as long as people have been altering their environment, they also have been polluting and depleting their natural resources. A striking historical illustration of the links between population growth, economic needs, and environmental change concerns the immense degradation of forests since the twelfth century in Europe to meet the timber supply for an increasing population. Only in the eighteenth century, in response to the devastating effects of the overexploitation of forests, was the concept of sustainable development formulated, with a view to the future need for a regular supply of timber and the preservation of the other functions of the forest.

In the twentieth century, international environmental policy evolved over several periods, reflecting the state of the environment, changes in political consciousness of the environment, the development of scientific knowledge, the application of technologies, and the changing structure of the international legal order. From the 1930s through the 1960s, a science-based literature gradually evolved which laid the foundation for political action in the field of environmental protection. In 1962, Rachel Carson published *Silent Spring*, an attack on the indiscriminate use of pesticides and a call for a change in the way humankind viewed the natural world. The book attracted immediate attention and contributed to a heightened public awareness of an endangered environment. Carson's research focused on the indiscriminate spraying of pesticides like DDT. As a result of her writing, DDT was banned as an insecticide in the USA and in most countries of the world. Today, the book is rightly considered a milestone in the history of environmentalism which has influenced modern thinking far beyond DDT. In 1970, a new awareness of the place of humankind in nature, of the sensitivity of nature, and of human responsibility, culminated in what may be considered the first systematic modern legislative and institutional response to the growing environmental challenge, being the National Environmental Policy Act (NEPA) by the US Congress and the establishment of the Environmental Protection Agency (EPA). Key initiatives of environmental policy date from the late 1960s and early 1970s when public demonstrations and riots were on the agenda in industrialized countries and when environmental pollution had become a major public health concern, initially because of the new phenomenon of acid rain in North Europe. These years of public unrest and distrust witnessed a slow erosion in 'business as usual,' and the rising call for 'the conservation of nature' soon led to the chaotic birth of the first green parties in Europe and, at the global level, led to the UN Conference on the Human Environment (so called Stockholm Conference) of 1972—the beginning of the active period of international environmental policy with new insights into the relationship between human society and the earth's natural capital. The Stockholm Conference set the scene for novel international activities at the global and regional level and also influenced legal and institutional development. In 1973, the United Nations Environmental Programme (UNEP) was created as the central environmental focal point in the United Nations, and other international organizations slowly followed suit and presented environmental protection concepts of their own, as did for example the Food and Agriculture Organization of the United States (FAO), World Health Organization (WHO), World Meteorological Organization (WMO), United Nations Educational, Scientific and Cultural Organization (UNESCO), and the World Bank.

1.2 *Characteristics of Modern Environmental Policy*

Knowledge of the impact of humankind on the natural world has grown rapidly since the emergence of the term 'environment' in the early 1960s. The emission and impact of sources of pollution became more understood, and new sources of pollution were discovered slowly. Ecology as a science was now regarded seriously by most scientists and the public. During the 1970s, the volume and influence of ecological studies increased, and new technologies to investigate the environment were developed, such as radar to observe the earth's surface. In the 1980s, purely scientific matters began to acquire a political dimension. Regional and global environmental threats became part

of the international community agenda as scientific evidence identified the potential consequences of climate change, ozone depletion, and loss of biodiversity. Local issues, like the forest destruction in the tropics or desertification, were recognized to have transboundary or even global consequences for mankind.

While the substance of environmental policy formulation is largely dominated by science, political essence is often characterized by the delay between cause and effect.

Environmental issues, like the ongoing dramatic forests losses, the increasing emission of greenhouse gases, and the loss of biological diversity call for *long-term strategies* in the field of environmental policy. The dimension of time is relevant for both international and national environmental policy. The experiences of the last few decades of the twentieth century demonstrate that democratic governments are better equipped to deal with long-term issues than authoritarian or totalitarian regimes. But political democracies as well as market economies may themselves be guided by parameters and values that make only limited allowance for long-term aspects for the sake of short-term concern for voting results or shareholder value. In practice, party leaders need to determine exactly what balance should be struck between quick successes and effective long-term policies. Environmental consequences of economic or social policies do not necessarily affect those persons or generations who enjoy its immediate benefits, and the momentary intangibility has given rise to the call for intergenerational equity and sustainable development. Political action to address these concerns has, moreover, to deal with the fact that science is not always clear on cause and effect of environmental degradation.

Originally conceived by German forest managers in the eighteenth century, the notion of *sustainable development* was highlighted in the final report by the World Commission on Environment and Development ('Brundtland Report' 1987) as 'development that meets the needs of the present without compromising the ability of future generations to meet their own needs.' The concept of sustainable development evades precise definition; in essence, it highlights two primary policy concerns of environmental policy: the integrated consideration of environmental, economic, and social issues and the expectation that policymakers will strike a fair balance between the interests of present and future generations (intergenerational equity). Having made numerous appearances in international treaties and other instruments, sustainable development is today firmly accepted as the key concept of international environmental and development policy. A single operational-oriented definition of the concept is impossible in view of the diversity of social, economic, and cultural conditions prevailing in each individual state. On the international level, the key challenge of environmental policy today lies in the ongoing search for a workable and effective consensus on the meaning of sustainable development which reflects the realities of both industrialized and developing states.

2. *Areas of Environmental Policy*

In terms of their origin and their effect, environmental problems may be divided into local and regional, international, and global issues. Such distinctions neglect that all sources of pollution and degradation originate form one local point, but they are useful in considering and formulating appropriate policy responses on local, regional, and global levels.

The thematic issues and areas of environmental policy are diverse and faceted.

(a) Global climate change is one of the greatest environmental challenges. For the most part, the increase of the concentration of greenhouse gases in the troposphere and the artificial warming of the earth's atmosphere seems to be due to human activities, including particularly fossil fuel use, land use change, and agriculture. The overwhelming weight of scientific authority predicts an alarming increase in temperature of about 1.4 °C to 5.8 °C by 2100 relative to 1990 and a rise of the average sea level of between 9 cm and 88 cm by the year 2100. Such changes would entail grave consequences for economies, the environment, matters of health, and food production.

(b) The accelerating decline of many of the world's forests represents another great problem facing the global community. The increasing human needs of a growing population in the developing countries, agricultural expansion, overharvesting of industrial wood and fuelwood, unsustainable commercial logging, forest fires, air pollution, overgrazing, pests, and diseases cause the ongoing massive forest destruction today, mainly in the tropics.

(c) The extinction and ecosystem destruction goes along with a relatively unknown loss of biodiversity. It is recognized that between 50 percent and 90 percent of all species currently in existence dwell in tropical forests. The global loss of species and natural habitat is proceeding at an alarmingly fast rate, with unknown but potentially disastrous future consequences.

(d) Another major issue concerns soil degradation which has affected about one-third of the world's agricultural soils and which often leads to desertification. Deserts account for about 7 percent of the earth's total land surface; the livelihoods of over 1 billion people in more than 100 countries are jeopardized by desertification, as farming and grazing land becomes less productive.

(e) Local and regional problems of the pollution of air and water have in part been addressed with success in industrialized states in the past decades, but continue to grow in most parts of the developing

world, in particular in the still growing number of megacities.

(f) The main pressure on land use is caused by an expanding economy and the demand for food by a growing population. Global food demand is expected to increase by 50 percent by 2015 and by more than 110 percent by 2050, with further pressure on the remaining nondomesticated areas worldwide.

(g) High population growth, the ongoing industrialization, and improving standards of living contribute to a higher demand for fossil energy. In the developed countries, energy use per capita is already extremely high and continues to increase slowly, and the much smaller developing nations' share of commercial energy consumption is expected to increase to nearly 40 percent by 2010. Energy consumption in the transportation sector is also expected to grow in both the developed and the developing countries as higher incomes will lead to a rising number of private cars in use. Energy consumption leads both to local and global problems of air pollution.

(h) Not only the earth's land surface, but also the oceans are threatened by ecological damage. The main environmental concerns to international waters stem from water quality degradation and overexploitation of marine resources. Land-based human activities cause 70–80 percent of the marine pollution and constitute a serious problem especially in coastal zones which provide 90 percent of the world's fishing catch and where 75 percent of the world's population live.

(i) The disposal and transport of industrial and domestic hazardous waste is another main problem on the national and international environmental agenda; many industrial products and chemical wastes are either nondegradable or difficult to dispose of.

(j) The development of modern tourism is characterized by a notable diversification in the location of new tourism destinations, particularly in Asia, North Africa, Latin America, and the Caribbean, with the result not only of economic growth, but also of an increasing environmental degradation in these regions.

3. *Instruments of Environmental Policy*

The aim of environmental policy interventions through different kinds of instruments is to correct failures or to control specific behavior in achieving society's environmental goals.

3.1 *Constitutional Provisions*

Constitutional provisions to protect the environment are an instrument of national environmental law, like Art. 20a of the Basic Law of the Federal Republic of Germany, introduced in 1994. Art. 20a provides that 'pursuant to its responsibility for this and future generations, the State shall protect the natural basis of human existence by legislation within the framework of the Constitution, and by the exercise of its executive and judiciary powers.' This rule establishes the protection of the environment as one of the goals of the State, although is does not accord any rights directly to the citizen.

3.2 *Command-and-control Approaches*

Two major kinds of tools are available for environmental policy: on the one side the more reactive command-and-control approaches (CAC) and on the other side preventive market-based instruments that create economic incentives not only for industries but also for private persons to reduce harm to the environment. Governments have traditionally employed CAC regulations in the 1970s and 1980s. This approach was subsequently reconsidered when economic studies pointed out that CAC regulations are often cost-ineffective, inflexible, and of limited efficacy. In particular, the private sector lacks an incentive to invest in developing a new and better control technology or strategy when the government dictates flat CAC rules.

3.3 *Market-based Instruments*

Limited inefficiency of CAC-regulations and the need to stimulate polluter's self-interest in the search for environmentally useful innovations have resulted in the development of alternative instruments achieving a better integration of economic and environmental decision making on a political and/or legal basis. The so-called market-based policies allow private industry and markets to remain relatively free in choosing the most efficient means of achieving environmental goals, and market incentives provide for increased flexibility and cooperation between government and industry; nevertheless, market-based-instruments require at least government supervision. Various types of market-based policies exist today.

Emissions trading was introduced in 1968 in the USA, and the attitude and perception towards tradable allowances among policymakers have since evolved; this concept allows a polluter to benefit from the results of a more effective actor in the same branch by way of paying for his additional effort. The interest in emissions trading especially for controlling greenhouse gases has grown during the 1990s, greatly influenced by the successful establishing of a tradable permit system to control sulfur dioxide and ozone-depleting substances in the USA. A special trading system for greenhouse gases is now envisaged in Art. 17 of the Kyoto Protocol (1997) to the UN Convention on Climate Change. The basic idea of emissions trading is simple. Each participant is subject to an allowed level of emission, and he is allowed to exchange parts of his emission allowances with another entity. Economically, emitters with low abatement costs will reduce emissions and sell their excess allowance, whereas polluters with high abatement

costs will prefer to buy them. The allowable emission total can be reduced over successive time periods. The Kyoto Protocol allows for two other types of flexible market-based innovations, namely Joint Implementation among industrialized countries and projects in developing states governed by the the so-called Clean Development Mechanism.

Another approach to promoting pollution prevention relies on *voluntary self-commitments* by industry, such as the voluntary self-commitment of European car manufacturers to achieve reductions in fuel consumption. Such voluntary programs are faster to implement than regulations and may produce faster reductions. The quantitative assessment of voluntary programs is still difficult partly due to the fact that these programs are still new.

Market incentives can also come from *positive subsidies* for environmental purposes, in forms of grants, soft loans, tax allowances, or special environmental taxes (so-called 'eco-taxes') levied on products that are harmful to the environment when they are used in the production process, consumed, or disposed of. At the same time, subsidies granted for other than environmental reasons often result in the degradation of the environment. Below-market financing costs for some electricity infrastructure, tax preferences for oil and mineral extraction, agricultural commodity price programs or tax preferences for private vehicle transport will in effect harm the environment if they support measures which encourage the emission of environmentally harmful pollutants, the generation of waste, and the excessive use of natural resources. In consequence a debate to reduce or to eliminate such subsidies has begun, so far without major effect on state practice. The financial consequence of environmental policies for the public budget rightly often dominate public debates about environmental reform, and the reduction of subsidies provides an important response to such budgetary concerns.

The *environmental impact assessment* may today be called a classical instrument. It provides for a systematic check procedure for relevant public or private projects, e.g., for the construction of power stations, airports, motorways, but also for some agricultural or infrastructural projects. Such assessments are intended to provide the governmental authorities and the private entrepreneurs with better environmental information about the consequences of a project in the planning stages.

3.4 Neoliberalism and Environmental Policy

The paradigm of economic neoliberalism raises hopes for more prosperity, but it also raises questions concerning the protection of the environment. The call for a lean state allows for and fosters a liberal economy which has dominated national and international debates in the 1990s and has been reinforced by the globalization of the economy. Inasmuch as free markets by themselves do not price negative environmental effects, economic neoliberalism will have to be combined with governmental or intergovernmental structures equipped and authorized to gather sufficient information and to exercise sufficient authority to protect the environment as a public good. This holds true, in principle, for the national and for the international level.

4. Principles of International Environmental Relations

In addition to substantive legislation describing in concrete terms the rights and obligations of the subjects of national or international environmental law, the international implications of environmental issues are reflected in a large number of principles, rules, or purely political goals which are still under development.

4.1 Sovereignty Over Natural Resources

Environmental issues do not respect national boundaries. Thus, environmental issues pose new challenges for state sovereignty, which is traditionally defined as affirming the independence of states in deciding about their own affairs, including their choices to define policies on the natural environment within their territory. Therefore, sovereignty over natural resources implies the right of a state to protect, but also to pollute or even to destroy the environment on its own territory. However, this right ends where environmental degradation is no longer restricted to the national territory and activities within one state's jurisdiction adversely affects conditions in another state. The obligation to protect the rights of other states within their own territory is generally accepted as a maxim of international law and frequently quoted in international environmental protection conventions and declarations, for instance as principle 21 of the Stockholm Declaration.

With regard to contemporary environmental problems with no single identifiable source like ozone depletion, global warming, polluted waters, eroding soils, loss of forests, and biodiversity, states have to assume responsibility not only for the ecological well-being of their own territory, but also of the earth's ecosystem. Thus, sector-specific regimes are developed among states which recognize that it is in the national interest to address jointly issues of global and regional environmental management.

4.2 Common but Differentiated Responsibility

Without disregarding the fact that efficiency in environmental protection can be achieved only by joint effort and obligation, the principle of common but

differentiated responsibility reflects the fact that industrialized nations and developing countries differ both in their responsibilities and capabilities. Many problems now existing in the field of environmental protection have arisen solely as a consequence of industrial production of the developed countries. Because developing countries are at an earlier state in the development process, they now insist on pursuing their own economic and social development, which includes polluting the environment in a manner formerly practiced by those states which today count among the industrialized nations. Thus, integration of developing states into global environmental regimes remains one of the key issues in designing international environmental policies.

In view of a potential conflict between economic necessity and ecological responsibility, a number of provisions in international treaties concerning environmental issues take into account the special situation and particular requirements of developing countries; in concrete, the relevant rules typically address issues of technology transfer and funding. For the time being, the principle of common but differentiated responsibility is a purely political concept, reflecting practical conditions and options without being legally binding.

4.3 Scientific Uncertainty and the Precautionary Principle

Decisions on international environmental issues have to be taken in the face of scientific uncertainty: new scientific information can either confirm the assessment of risks or determine a less serious risk than estimated. The precautionary principle requires that a nation shall take measures to prevent environmental degradation even while there is a lack of full scientific knowledge. While originating in the context of the marine environment, the precautionary principle is now being applied to nonmarine issues such as the ozone depletion and climate change. One of the first texts to mention the precautionary principle was the preamble to the Convention for the Protection of the Ozone Layer; in the Rio Declaration it is embodied in principle 15. Although the precautionary principle appears in a number of international declarations, its basis and scope remain controversial. The growing practical acceptance of the concept was reflected in the successful negotiation of the Biosafety Protocol to the Biodiversity Convention in 2000.

4.4 Preventive Principle

The principle of preventive action states that it is much more effective to prevent pollution than to remedy its effects or to assign liability for environmental damage. Therefore, states are obligated to take measures designed to protect the environment early enough to keep environmental harm from occurring. The preventive approach has been endorsed *inter alia* by the 1972 Stockholm Declaration, the 1982 World Charter for Nature, and in a large number of international treaties. Principle 11 of the 1992 Rio Declaration calls on states to enact 'effective environmental legislation.' Not only incorporated into specific environmental treaties, but also into treaties of more general application, such as the European Community Treaty (ECT), the preventive principle involves the obligation constantly to take into account environmental concerns in all area of governmental activity and control. Instruments like early-warning systems, risk assessments and stronger monitoring provisions reflect the anticipatory approach. Despite differences in national legal systems, it will be of elementary importance to ensure that these principles are implemented in every state in a manner sufficiently efficient to benefit the environment.

5. International Environmental Law

5.1 Global Issues

International environmental issues have increasingly burdened national governments and the factual degradation of the environment poses serious problems for the maintenance of national sovereignty. International policies to respond to such degradation reflect this factual situation, and in the long run the deliberate acceptance of the restriction of national sovereignty by way of acceptance of international agreements may be more advantageous from the vantage point of national freedom than insistence on national legal sovereignty and the absence of binding agreements. This paradox of environmental policy, however, remains to be universally understood and accepted.

A period of international environmental law began with bilateral fisheries treaties in the nineteenth century to halt overexploitation; yet practical effects were narrowed by emphasis only on border-related environmental problems. Early attempts to develop international environmental rules focused on the conservation of rivers and seas and the conservation of wildlife; the first whaling convention was adopted in 1931. The second phase in the development of international environmental policy was marked by the creation of the UN and the specialized agencies in 1945. Subsequently, international organizations on the regional and global level slowly began to address environmental issues, and the new range of environmental concerns included a focus on the causes of pollution resulting from certain ultrahazardous activities. However, recognition of the relationship between economic development and environmental

protection hardly existed. Nevertheless, growing policy concern over environmental protection prompted not only new actions on the local and national level, but also led to international efforts. The first culminating point was the Stockholm Conference in 1972. The goal set by governments at this conference was to 'protect and enhance the quality of the human environment for present and future generations' (Principle 1 of the Stockholm Declaration on the Human Environment). In the mid-1980s, a renewed internationalization began to take place: the increasing interdependence among states called for new foundations for international environmental cooperation. New regional and global conventions concerning environmental issues were adopted, and the first treaty to protect the ozone layer was concluded in 1985.

The next fundamental breakthrough in the international negotiations of universal environmental issues was embodied and articulated by the United Nations Conference on Environment and Development in Rio de Janeiro in 1992 (UNCED), attended by 176 states and several thousand nongovernmental organizations. UNCED was concerned with a new paradigm of environmental politics, namely the new search for the balance between environmental protection and economic development. The conference adopted two legally binding treaties, the Convention on Biological Diversity and the Framework Convention on Climate Change, and three nonbinding instruments, the Rio Declaration, the Principles on Forests, and the AGENDA 21. UNCED has reaffirmed the need to integrate the concept of sustainable development and its environmental dimensions into economic policies. The World Bank, in particular, has since insisted more than other international organizations on the need for a holistic response to economic social and environmental issues and has started to work on so-called comprehensive development frameworks with its client states. In principle, a consensus has also emerged on the importance of technology cooperation as a strategy of sustainable development; while developing states have called for the transfer of machinery and technical equipment, developed states prefer to discuss the transfer of knowledge of science and technology and support for the institutional capacity to address environmental issues.

In the follow-up to UNCED, various institutional innovations have been introduced, like the Commission of Sustainable Development (CSD) within the United Nations. Also, funds have been set up for the purpose of transferring financial resources and technology for special sectors, e.g. the Multilateral Ozone Fund. The Global Environment Facility (GEF), initiated in 1990, and substantially reformed in 1994, is responsible for the financing of projects in the field of global environmental protection, particularly for measures in four focal areas (protection of the ozone layer, climate, international waters, and biological diversity). The GEF stands out today as an institutional recognition for the requirement of environmental cooperation between North and South, with a novel voting structure which prevents North or South dominating. While in the GEF industrialized states have recognized in a binding way that effective global environmental policy presupposes support of the developing states by the developed world, developing states in turn have accepted that efforts within their economies and societies are required in the effort to address the global issues. It is true that the funds for the GEF are currently limited, less than US$1 billion per year.

In 1997, the nineteenth UN General Assembly Special Session (UNGASS) took place to review progress in the field of sustainable development as set out in the Rio Declaration and Agenda 21 ('Earth Summit +5'). Five years after Rio, some positive results had been achieved, reflected in the increasing willingness of states to accept international treaties to protect the world's climate, the ozone layer, biodiversity, and to combat the spread of deserts and soil degradation. Influential segments of civil society have been formed to address the long list of issues on the environmental agenda. Nevertheless, significant environmental problems remain in all regions of the world, and *communiqués* formulated by professional speech-writers cannot do away with the fact that the state of the global environment continues to deteriorate. Generally speaking, international environmental policy making is today seriously hindered by the absence of an effective body with a voice and authority comparable to the influence of other international organizations such as the World Trade Organization (WTO) or the International Monetary Fund (IMF). This institutional deficiency will become more and more noticeable as national governments will remain inclined to focus on their national agendas and on short-term issues. Thus, institutional improvement will remain on the agenda on the international level as a prerequisite for more effective policy making.

5.2 Regional Issues

The environmental rules of regional organizations must be considered in the context of their overall legal and political structure. Thus, the environmental law and policy of the EC is highly influenced by the Treaty of Amsterdam, adopted on October 2, 1997, which has introduced important changes to the environmental provisions of the EC treaty (ECT). Sustainable development has now been formally introduced as one of the goals of the European community (Article 2, ECT). Furthermore, the principle of integration, requiring the integration of environmental policy demands into other areas of community policy, is now embodied in a single provision in the very first part of the ECT (Article 6, ECT).

As the USA, Mexico, and Canada moved to liberalize regional trade through negotiations of the North American Free Trade Agreement (NAFTA), environmental concern played a significant role in the discussions. Despite many environmental provisions in the Agreement, i.e., environmental, health, and safety standards as well as sanitary and phytosanitary measures, the provisions were seen by the US environmental community as inadequate in response to the environmental risks posed by the liberalized trade among the three nations. In consequence of this criticism, the parties have negotiated a Side Agreement, the North American Agreement on Environmental Cooperation (NAAEC), which went into force on January 1, 1994 together with NAFTA.

With an increasing importance of regional economic associations, it will be important to formulate environmental policies which are not in practice subordinated and evaded in the application of the rules on economic progress.

6. *Prospects*

In the developing world, about 2 billion people live on less than US$2 per day, and it is estimated that only half of them have access to clean and safe water. Development experts point out that these numbers may grow rapidly in the coming decades, and it has been rightly said, from an environmental viewpoint, that poverty acts as the most 'toxic substance.'

At the same time, the developed world relies on patterns of production and consumption which are often wasteful and which are not sustainable in the long run if adopted by people in the developing states. In these facts lies the basic cause and primary challenge of growing environmental problems. Ultimately, environmental policies will succeed or fail to the extent that they effectively address the underlying causes of these facts. If formulated against the forces of the market and without incentives for more environmentally benign technologies, such policies would fail. Also, the support of such policies by strong sectors of civil society is necessary. In the end, however, governments and international organizations are needed to formulate, accept, and implement effective environmental policy.

See also: Climate Policy: International; Deforestation–Forestation; Development: Sustainable Agriculture; Ecotourism; Environmental and Resource Management; Environmental Planning; Environmentalism, Politics of; Environmentalism: Preservation and Conservation; Land Degradation; Population Policy: International; Precautionary Principle; Sustainable Development; Tourism, Geography of

Bibliography

Baker R (ed.) 1997 *Environmental Law and Policy in the European Union and the United States*. Praeger Publishers, Westport, CT

Bowman M, Redgwell C 1996 *International Law and the Conservation of Biological Diversity*. Kluwer Law International, London

Brown Weiss E (eds.) 1998 *International Environmental Law and Policy*. Aspen Law & Business, Gaithersburg, MD

Campiglio L (ed.) 1994 *The Environment After Rio—International Law and Economics*. Kluwer Academic Publishers Group, London

Dolzer R, Thesing J 2000 *Protecting Our Environment—German Perspectives on a Global Challenge*. Konrad-Adenauer-Stiftung e.V., Sankt Augustin

Food and Agriculture Organization (FAO) 1995 *State of the World's Fisheries and Aquaculture*. FAO, Rome

Grubb M 1999 *The Kyoto-Protocol—A Guide and Assessment*. Royal Institute of International Affairs, London

International Panel on Climate Change (IPCC) 1996a, *Climate Change 1995, The Science of Climate Change—Summary for Policymakers and Technical Summary of the Working Group I Report*. Cambridge University Press, Cambridge, UK

Litfin K (ed.) 1998 *The Greening of Sovereignty in World Politics*. MIT Press, Cambridge, MA

Livernash R 1998 *Valuing the Global Environment—Actions & Investments for a 21st Century*. Global Environment Facility (GEF), Washington, DC

Sand P 1999 *Transnational Environmental Law—Lessons in Global Change*. Kluwer Law International, London

Sands Ph 1995 *Principles of International Environmental Law, Vol. 1: Framework, Standards, and Implementation*. Manchester University Press, Manchester, UK

Tarasofsky R (ed.) 1999 *Assessing the International Forest Regime*. IUCN Environmental Policy and Law Paper No. 37, Gland, Switzerland

Tolba M 1998 *Global Environmental Diplomacy*. MIT Press, Cambridge, MA

United Nations Environment Programme (UNEP) 1997 *Global Environment Outlook*. Oxford University Press, Oxford, UK

World Resources Institute, UNEP, UNDP, World Bank 1998 *World Resources 1998–99—A Guide to the Global Environment*. Oxford University Press, Oxford, UK

R. Dolzer

Environmental Policy: Protection and Regulation

Environmental policy is a commonly used term that eludes a simple, all encompassing definition. Essentially it is concerned with how best to govern the relationship between humans and the natural environment for the benefit of both. This broad definition has the merit of successfully capturing the unbounded character of the environment as a field of policy, touching as it does upon virtually *every* aspect

of social and economic life. However, it lacks specificity: most if not all policy areas are concerned with changing the world 'out there.' One way to proceed towards a more precise definition is to break the term down into its component parts—environment and policy. Albert Einstein famously defined the *environment* as 'everything that isn't me.' Environmentalists would argue that this fundamentally depersonalises 'the environment' which is inherently a subjective, spiritual experience specific to each and every one of us. However, they would agree that his notion of 'everythingness' is useful insofar as it confirms powerfully the complex and essentially unbounded nature of environmental policy, underlining the need for policy responses which are integrated with the main sectors of society. On this line of reasoning, which is now beginning to attract much greater political purchase, *every* policy area has environmental implications and should be conceived of as 'environmental.' In other words, the broad and holistic definition is actually more useful than it first appears. All this may seem like a fairly minor definitional problem but environmentalists argue that the 'everythingness' of the environment is why they find it so difficult to get politicians to take the matter seriously. The very interconnectedness of environmental problems render them unamenable to simple, soundbite politics and *ad hoc* policy interventions. Consequently, they tend to be marginalized.

Environmental policy commonly is defined in terms of the problems it addresses. So if we look at the traditional content of environmental policies, most are associated with controlling pollution, limiting the loss of natural habitats, managing waste such as sewage and litter, and reducing the impact of natural hazards such as flooding (McCormick 1991). These policies normally are put into effect using regulatory tools although recently there has been a pronounced trend towards using nonregulatory tools such as subsidies, eco-taxes, voluntary agreements, and information campaigns. However, environmental policy in the broader sense described above touches upon many more issues than this functional list of issues implies. Indirectly it is related to food and energy production, the transport of goods and services, human health and safety, consumer choices, the survival of biological species, and national and international security in the broadest sense. In short, environmental policy addresses nothing less than the complex, reciprocal links between modern society, the economy, and the environment. In that sense it is probably unique among policy areas in its breadth and complexity (Ingram and Mann 1983). Achieving a harmonious balance between societal, economic, and environmental systems is the main aim of sustainable development, the main organizing concept of contemporary environmental policy.

If 'the environment' is amorphous then the term 'policy' is equally slippery. A dictionary definition is 'a course of action or principle adopted or proposed by a government.' However, some policies comprise a whole series of decisions, layered on top of one another. Policy-making is recursive: policies are put into place to address problems then immediately retuned to fit new political demands. Moreover, policies may be adopted formally by national governments but the detailed aspects of implementation are often left to lower levels of government to determine. Therefore, when a local pollution control officer negotiates with the owner of a polluting factory over the details of site license or emissions permit, they are arguably making policy. Finally, policies also serve a symbolic purpose. Some are only ever statements of intent which governments have no real intention of enacting. Policy, therefore, should be thought of in terms of courses of action and inaction. To summarize, policy is as much an ongoing *process* of interaction as a discrete product or output of the political system, written down in legal texts.

1. *The Characteristics of Environmental Policy*

We can therefore summarize environmental policy as an evolving societal process of addressing the quality of the natural world: hardly a crisp and clear-cut definition! Perhaps the best way to illuminate the concept is to compare the environment with other policy areas such as health, the economy, and crime. According to Weale and Williams (1992), the objective characteristics of a policy area have an important bearing upon the form of politics and policies that develop around it.

1.1 *Public 'Goods' and 'Bads'*

Environmental quality is a classic example of what economists term a nonexcludable or 'public' good. By that they mean that the benefits of environmental quality accrue to most if not all citizens (or, to put it another way, the costs of degradation are borne by society rather than just the polluter). This seemingly self-evident observation has a number of important implications for the scope of environmental policy. First, large parts of the environment—the atmosphere and large areas of the ocean—are common property and nobody can be excluded from using them. This means that while benefits of resource use tend to be concentrated spatially, the costs or 'externalities' typically are dispersed across myriad individuals in the form of pollution (or economic 'bads'). Think of a smoking factory chimney pouring black soot on to the houses surrounding it, for example. In this example the polluter has an economic incentive to carry on consuming clean air in the knowledge that the bads will be born by the whole neighborhood. Second, while the financial cost of improving the environment often falls on one small section of society (e.g., the

owner of the factory), the benefits normally are spread widely among myriad of users. This is partly why polluters, who have a strong incentive to unite and fight to protect their right to pollute, have an immediate advantage over those who suffer the disbenefit. The latter are often too widely dispersed or individually suffer too little to mobilize into a coherent group. Third, those fighting for the 'public' interest or the welfare of natural entities do not necessarily act automatically to achieve their individual or group self-interests—an observation first made by the political scientist Mancur Olson. He argued that individuals have an incentive to 'free ride' by consuming environmental quality in the hope the cost will be born by some other party or society as a whole. For example, polluters of a stream will continue to pollute in the hope that a sufficient number of other users desist from polluting it. However, the public good aspect of environmental problems also conditions attempts made to abate environmental damage. Then, recreational users of our hypothetical river may well calculate that it is easier to free-ride (i.e., wait for someone else to tackle the polluter and share in the benefits of clean water) than do something about it themselves. One of the reasons why motorists continue to drive is that they cannot be excluded easily from the benefits (i.e., less air pollution and congestion) which are generated when some of their number give up their vehicles and turn to public transport.

1.2 The Need for Regulation

Because environmental damage has its origins in otherwise socially legitimate activities like energy and food production, the role of the state is often to police the consumption of public goods by limiting the level of damage that one section of society can impose upon others. Environmental policy is, therefore, inherently *regulatory* in nature, although regulation inevitably has distributive and redistributive consequences (Lowi 1972). Lowi argued that whereas distributive policies (e.g., welfare programs) produce close, dependent relationships between the state and interest groups, regulatory policies produce more conflictual forms of politics. However, not all interest groups enjoy the same level of influence or political power. Some routinely have more of 'a say' than others in determining the content of policy by 'capturing' the regulatory process. The precise manner and extent to which the state is captured by sectional groups is disputed hotly among political scientists (Dunleavy and O'Leary 1987). It is abundantly clear, however, that business groups do enjoy a structural advantage over other groups—one that exacerbates the collective action problems suffered by environmental groups. According to Lindblom (1977), the basic fact that *all* governments, whether Left or Right, need a healthy economy to provide jobs and welfare support, forces them to adopt measures that are in the interests of business without business having to take any observable action. Scientists too have their own independent policy objectives, and can be thought of as an interest group. In some situations 'epistemic communities' (Haas 1992) of scientific experts have had a decisive influence on the course of policy by steering decision-makers to their preferred solutions.

1.3 Complexity and Uncertainty

Once human development exceeds the planetary capacity for self-repair, environmental assets can collapse and disappear, never to reappear. Admittedly, some environmental assets and functions are repairable or substitutable (individual trees can be replanted; water can be transported), but many others (a delicate wetland habitat, for example) are not. Irreversibility complicates environmental policy making, differentiating it from other areas such as tax or social affairs where problems can be more easily revisited. Irreversibility implies that piecemeal policy responses should be substituted for ones that foresee problems and address them systematically. In practice, the environment often responds slowly and subtly to human interventions, and 'local' problems are sometimes only the observable tip of much more complicated and spatially dispersed set of processes. For instance, the pressure to pollute a local river with pesticides and fertilisers may stem from a national policy to intensify agricultural production in order to compete in world markets. The 'international' agricultural policy and the 'local' pollution policy are interconnected and should be developed as such. However, in practice it is often extremely difficult to trace the connections between and within particular spatial levels and sectors, leading to compartmentalized policy making. This raises the question of how decisions should be made at one level when the entire sequence of consequences for other levels or sectors cannot be predicted. Should decision-makers err on the side of caution or proceed in the hope that unfavorable consequences will not arise? This is one of the reasons why the precautionary principle—the idea that policy makers should go 'beyond science' to address potential problems when the threat of harm is significant and potentially irreversible—has become such an important issue of policy debate in areas such as genetically modified food and endocrine-disrupting chemicals (Jordan and O'Riordan 1999). But, whenever the state of the environment cannot be determined precisely before or after a decision is made, the scene is set for those with vested interests to twist scientific ignorance to suit their own cause. Political systems find it difficult to cope with genuine uncertainty in policy making. This is why difficult environmental

policy decisions routinely are handed to committees of experts to deliberate upon, delaying still further the adoption of protective measures.

1.4 Time and Space

We have noted that environmental quality is a classic example of a public good, but the costs and benefits of policy do not necessarily accrue to the same citizens. Consequently issues of equity and justice are central to most, if not all, environmental decisions. More often than not, those who suffer most from environmental damage are the poorest who lack the means to protect themselves, while those who benefit have a vested interest in maintaining the status quo. The precise spatial and temporal distribution of costs and benefits produces alliances of self-interest which battle to determine the scope and content of environmental policy. The construction of a bypass road around a congested village, for example, will greatly improve the welfare of its residents but at the cost of disruption to those living in the path of the alternative route. Balancing these different needs becomes a central question of *politics* (or decision-making) which is pervasive in all forms of environmental policy making. Another problem facing policy makers is how to account for the well being interests of future generations whose needs can only be guessed at. According to the most widely adopted interpretation of sustainable development, policy makers should address the needs of current generations of humans without compromising the ability of future generations to meet theirs. Clearly, this raises very thorny questions of justice and morality, to say nothing of due democratic process.

1.5 Political Borders

Unfortunately for policy makers, environmental problems show scant regard for political borders. Essentially, there are three types of international problem. The first arises when the legitimate activities of one state have international implications (e.g., acid rain and global warming). Then there are global processes such as trade and international debt that have environmental implications. Finally, there is the problem of how to preserve resources such as rainforests and regional seas that reside in one part of the world but are valued by all. Olson's theory of collective action illuminates the difficulties in achieving cooperation between otherwise sovereign states, especially when action is required from those who can legitimately say they are not to blame. Regional arrangements, such as those pioneered by the European Union, provide the best evidence of successful international diplomacy.

However, many of the disputes about who is responsible for causing and remedying global environmental problems such as global warming change hinge on a North vs. South axis, though not exclusively so. In these circumstances, environmental policy relies upon regimes—international conventions and protocols—to develop the necessary amount of trust between participants. In some circumstances institutional arrangements are put in place to deliver 'selective incentives' (Olson 1965) (i.e., technical and financial assistance) to bring reluctant countries on board whose participation is considered vital to the success of the regime (Jordan 1994).

1.6 Environmental Policy Integration

While the environment functions as an integrated whole, the bureaucracies set up to manage it typically are fragmented into different sections in order to avoid administrative overlap and inefficiency (Weale et al. 1991). The mismatch between institutions and environmental problems has been an enduring problem for policy makers since time immemorial. The environment is probably unique in the extent to which it cuts across traditional sectors of government. It is why environmental policy making often produces furious conflicts betwn environmental ministries and the older and often more powerful parts of government representing finance, agriculture, transport, and so on. Thus, one of the great unmet challenges of the 1990s has been to find mechanisms to *integrate* environmental concerns into all areas of decision-making (Jordan 2000, Weale and Williams 1992). According to the UN ('Brundtland') report that did so much in the late 1980s to set the terms of the international debate about sustainable development:

> Those responsible for managing natural resources and protecting the environment are institutionally separated from those responsible for managing the economy. The real world of interlocked economic and ecological systems will not change; the policies and institutions concerned must (World Commission on Environment and Development 1987).

In the 'real world' of bureaucratic politics described by Brundtland, each department defends its own position while resisting a line that, while potentially beneficial to society as a whole, clashes with its 'own' interests. So, even if the environment department wishes to push a greener line, it is likely to come up against fierce opposition from departments representing the interests of farmers, industrialists, or road builders.

1.7 Nonhuman Entities

Environmental policy is probably the only domain of policy that has as its main goal the protection of nonhuman entities. Just about everybody supports some form of environmental protection. But the crucial question is to what extent should moral

significance be ascribed to plants and animals, especially if it involves self-sacrifice and higher economic costs? There are essentially two ethical positions in this debate: technocentric and ecocentric. The clash between these value positions, which have remained remarkably resilient in spite of the widespread greening of party politics, policy making, and government throughout since the 1970s, is clearly identifiable in most contemporary environmental disputes. Those who adopt a *technocentric* view regard nature as a source of resources to be exploited, whereas *ecocentrists* feel that the nonhuman world has interests and moral significance independent of humans valuation—what economists term 'existence value.'

While the technocentric view is hierarchical, manipulative, and managerial, the ecocentric view, by contrast, embraces community scale, natural rhythms, and a morality based on ecological principles. Business tends to adopt a more technocentric line, which conditions the general direction of policy via the structural commitments outlined above. The 'green' case put by environmentalists is much more ecocentric. If anything, 30 years of environmental debate has led to a convergence of interests towards the center of this continuum of worldviews. To a greater or lesser extent environmentalism has made all industrialized societies much greener. This is reflected in and nourished by the discourse surrounding the term sustainable development which presents environment and development as two sides of the same coin, one being the necessary precondition of the other. Sustainability as defined by the Brundtland report is politically palatable precisely because it carries the comforting message that the world *can* grow its way out of ecological and geopolitical crisis.

Despite this convergence, contemporary environmental conflicts are still driven by the clash between worldviews. We see this in the current debate about genetically modified organisms (GMOs) and foods. The large chemical companies are technocentric in their outlook, stressing the benefits of manipulating natural processes for human benefit. Their faith in the power of science to achieve a transition to modified crops in a predictable and controlled manner sets them apart from environmental groups who stress the fundamental uncertainties involved—uncertainties which, if not respected, may result in disastrous and irreparable damage to natural biodiversity. At the root of the unfolding debate about GMOs is the question of whether humans have the moral *right* to 'play God' with nature. The technocentric has few moral qualms about this, whereas the committed ecocentric does. We miss a great deal if we try interpret environmental policy making solely in terms of the exercise of political power, the clash of institutions, and the development of the law. Beliefs, morals, and ethics make up the 'glue' that cements individuals together into interest groups, informing their political positions and policy preferences.

2. *The Main Elements of Environmental Policy*

2.1 *A Brief History*

As policy areas go, environmental policy is relatively youthful, the main planks dating back only as far as the early 1970s. This is not to say that there were no environmental provisions before that time, only that they were relatively fragmented, aimed mainly at securing human health, and developed in a reactive and *ad hoc* manner. Environmental policy gained its impetus from the great upsurge of public interest in environmental matters that occurred in the late 1960s. Landmark books such as Rachel Carson's *Silent Spring* and *Blueprint For Survival* by the *Ecologist* magazine suggested a global crisis was imminent—a message which found a keen and receptive audience in the USA and Western Europe. It is often said that the first evocative images transmitted from space of Earth floating precariously in the inky blackness of deep space, catalyzed and intensified public concern. They emphasized much more clearly than words the finiteness of the planet and did much to encourage a *global* perspective. It is also undeniably true that the unprecedented magnitude of human induced environmental change throughout the twentieth century had also produced an objective reduction in environmental quality. The European Union's (EU) environmental division, DG XI, the UK Department of the Environment (DoE), and the United Nations Environment Program (UNEP) all date from this period, as do the Norwegian and Danish environment ministries (Hanf and Jansen 1998). The policies adopted were mainly regulatory in nature, specifying process and emission standards that had to be met. In those countries with administrative systems that were less well developed, this early period saw the development of basic standards for the main environmental media—air and water—and for waste and noise pollution.

In the 1980s and 1990s, the focus shifted from dealing with the symptoms of environmental damage—pollution, waste, habitat loss—to addressing the underlying causes. The trend towards redesigning institutions and organisations to tackle problems at root is often labeled 'ecological modernisation' (Weale 1992). This involves moving from setting end of pipe standards to tackle the social processes that generate the emissions in the place. The shift towards more diffuse and less interventionist policy tools in part reflects political pressure from business for more flexible and economically efficient alternatives to regulation. The 'new' tools in the armoury of environmental policy now include:

(a) *Ecolabels*: these signal the environmental impact of products and processes to purchasers and investors. Eco-auditing and lifecycle assessments help to rate the ecofriendliness of products and processes by measuring their overall environmental impact.

(b) *Ecotax reform*: working with the grain of markets, taxes are designed to penalize economic 'bads' and, through hypothecation mechanisms, reward 'goods' such as labour. For example, in 1997 the UK adopted a tax on the waste deposited in landfills, the revenues from which are diverted to environmental projects and employee taxes.

(c) *Environmental assessment*: the aim here is to implement the principle of environmental policy integration. The US National Environmental Policy Act (1969) formally required an environmental assessment of individual projects such as a new road. The EU is developing plans to introduce similar provisions much earlier in the decision-making process, via the assessment of overarching policies and programs (strategic environmental assessment).

(d) *Integrated product policies*: using various devices to 'green' the production of goods from their design and production, through to final disposal. A good example is the EU's packaging waste directive that aims to reduce the overall quantity of waste material.

(e) *Strict liability regimes*: by making developers legally responsible for remedying the damage they cause, such regimes are seen as a means of encouraging a more consistently preventative approach to environmental issues. The European Commission has been trying to develop such a regime since 1993.

2.2 National Policy Styles

When we study the political response to environmentalism, we see that states and supranational organizations have followed slightly different paths reflecting their history and political mandate. By and large, the current allocation of responsibilities respects the principle of subsidiarity, which holds that problems should be addressed at the lowest appropriate administrative level. Thus, global problems are addressed at the global level (i.e., the United Nations), regional problems at the regional level (e.g., the EU), and the rest at the national or subnational level. Environmental pressure groups and scientists have learned to work at all levels, removing environmental policy from the domain of purely state-guided action and placing it in the context of a multilevel system of governance.

In spite of the internationalization of environmental policy making, individual states continue to pursue different approaches, reflecting their individual geographical circumstances, political and constitutional make-up, cultural orientation, and economic conditions (Skou Andersen and Liefferink 1997, Jänicke and Weidner 1997). The interaction of these circumstances gives each country a distinct national 'policy style' (Richardson 1982) which pre-exists its response to new political issues. The concept of national policy style would seem to contradict Lowi's basic thesis that the functional characteristics of a problem determine the main aspect of policy. Policy analysts continue to debate which is more important. It is clear, though, that Britain and Germany, two of the largest and most important actors in the EU, have a long history of approaching problems in very different ways (Weale 1992). The British style is highly consensual and reactive, and decisions are made by bargaining between actors in the context of informal rules. Since the 1980s Germany has adopted a more anticipatory approach based on explicit standards and programs which are designed to 'force' the development of clean technologies. Different again is the American 'style.' It is much more open, conflictual, and rule-bound, reflecting the adversarial nature of the US political system. The importance of the free market in American society partly explains why the USA has gone furthest in experimenting with economic instruments, such as tradable permits (Golub 1998).

Significantly, these national differences also explain why countries perceive the same environmental problem in very different ways. Significantly, the environment 'out there' and the policies adopted to protect it are:

> themselves … products of the political process. Countries will vary among themselves and over time as to what is considered to be an 'environmental problem' as well as with respect to what should be done, and by whom, about it (Hanf and Jansen 1998).

2.3 The Policy Process

The policy process normally is seen as having a series of sequential parts or stages. These are (a) problem emergence, (b) agenda setting, (c) consideration of policy options, (d) decision making, (e) implementation, and (f) evaluation. According to this 'textbook' view of policy, the first task facing environmental groups is to get a particular problem on the agenda for discussion and, if possible, consideration by policy makers. Policy makers then select the best course of action based on specialist advice, make the policy, then hand it on to administrators to be implemented. This stage-based view emphasises that policy is a process involving many different parts of the Government. It is also simple and intuitively appealing. But in practice, policy issues are interconnected, policy makers fumble around for solutions in the context of great uncertainty and many internal and external constraints. Often what was done in the past has a determining effect on how new issues are processed.

There is a multitude of different models and theories which try to explain how the machinery of government makes policy (Ham and Hill 1993, Parsons 1995, John 1998), each presenting a slightly different account of how the constituent elements interrelate. *Pluralists* believe that political power is widely, although by no means evenly, spread throughout society. Although

there are very powerful groups in particular policy sectors, no single group is successful continuously or capable of distorting the entire policy process to its advantage. In policy terms, pluralists assume that the agenda-setting process is open and competitive, with the government acting as an honest broker. Once adopted, though, policies must still be steered through the reefs and shoals of the implementation process. Because of the competitiveness of the policy process policy outcomes tend to be unpredictable.

Neopluralists on the other hand argue that business is in a 'privileged' position compared to other groups. Instruments like the mass media help to structure environmental politics by removing 'grand majority' issues concerning the fundamentals of the political order from the agenda, leaving citizens to debate an endless range of 'secondary' concerns (Lindblom 1977). Whereas pluralists assume that grievances are brought fully into the open, neopluralists argue that they are organized out of politics by institutional rules and routines. Policy is made and implemented within fairly small and stable groups of actors (or networks) clustered around particular government departments. According to neopluralists, policy outcomes generally reflect business preferences rather than those of environmental groups, who find it difficult to gain access to decision makers or conclude that the chance of bringing about change is so slim as to be not worth fighting for.

Finally, structuralists (e.g., *Marxists*) believe that the state is under powerful structural pressure to nurture economic growth regardless of the environmental implications (Benton 1996, Pepper 1996). On this view, the welfare state and most environmental controls are nothing more than a sham, put in place to pacify critics and keep the conflict between economic classes to manageable levels. For Marxists, environmental problems are rooted in the unequal distribution of resources within society. There can be no lasting solution to such problems until the social system is structured more equally. Although longer-term outcomes inevitably reflect the needs of business, there may, however, be limited, short-run concessions to environmental demands.

2.4 What Drives Environmental Policy?

Having defined the policy process and the main tools used, the obvious question is what actually drives it forward? Where do the inputs, or political issues, come from which create the agenda for environmental policy making? The answer partly is conditioned by which of the three models described above is used. These differences notwithstanding, ultimately the source of all green ideas is environmentalism, a broad social movement which emerged in the late 1960s reflecting a growing public demand for a better environment (Doyle and McEachern 1998). Environmentalism, the roots of which stretch far back into the past (Pepper 1996), is reflected in a long-term shift in public opinion towards 'quality of life' issues such as peace and tranquillity, and an increasingly active and diverse range of 'green' pressure groups. Environmental concern has exhibited a broadly cyclical pattern over time, with particularly pronounced peaks in the late 1960s and late 1980s. Closer scrutiny reveals that these short-term pulses coincided with periods of economic growth and social introspection.

The political scientist, Downs (1972), has produced a simple model that captures these dynamics. He argues that environmental issues evolve slowly in a culture of scientific enquiry and suppression. This is usually a fairly secretive phase during which gloomy prognoses are carefully explored away from the glare of media concern. The second phase in his 'issue attention cycle' was one of media attention and anxiety triggered by crises. Since the 1960s, the environmental movement has benefited enormously from a series of symbolic environmental crises from Amoco Cadiz, Seveso, Three Mile Island, to Chernobyl and GMOs.

The third phase is the reactive process in which political institutions rush to cope and adapt. The early 1970s and the period 1989–92 were both marked by important institutional innovations. The respective focal points were the international conferences at Stockholm in 1972 and Rio in 1992 which spawned a suite of international regulatory protocols. In turn these triggered a higher order of public and private awareness, much greater financial investment, and corporate planning. During the fourth phase the costs of protection sink in, and demands grow for regulation to be rolled back to allow enterprise and economic growth. So-called contrarians begin to question the accuracy of the more gloomy environmental claims (North 1995). However, downgrading the importance of environmental protection succeeds only in sowing the seeds of the next crisis, and so the cycle begins again. Significantly, however, political concern has not settled back to the original level, but has oscillated upwards around a steadily rising curve. The radical environmentalist critique has been blunted greatly as governments have adopted tougher and tougher environmental policies. But ratchet-like, the laws and public institutions put in place to protect the environment during each upsurge, have helped to fix environmental values in place, providing a platform for the next burst of policy making.

3. Conclusion

Environmental policy has reached a critical stage in its relatively short life. In the 1970s governments across the world responded to environmentalism by establishing agencies and laws to address specific issues such as oil pollution, acid rain and ozone depletion, as and

when they arose. The primary aim of these early innovations was to safeguard human health rather than protect the environment for its own intrinsic value. Policy makers regarded environmental problems as being essentially technical in nature and politically uncontentious; policies were overwhelmingly reactive rather than proactive, and sectoral rather than integrated across policy domains.

But the discourse surrounding environmental issues has begun steadily to metamorphose into one centered on sustainable development. Significantly, sustainable development is not simply development as conventionally understood, 'greened' by bolting on cleaner technologies and reducing waste, but a *process* of human development which simultaneously satisfies social, economic, and environmental goals, within and between different generations across the world. The emerging orthodoxy is that environmental thinking must now be integrated permanently into areas of policy that have traditionally been regarded as 'nonenvironmental' such as agriculture, transport, and energy. One of the most important lessons learnt since the 1970s is that the fundamental driving forces of pollution and habitat loss reside in these sectors. Unless and until these forces are addressed at their roots, environmental protection will always be a reactive, piecemeal, and *ad hoc* activity. Politically speaking, environmental policy integration represents a new and potentially difficult stage in the continuing development of environmental policy, although the idea behind it has been canvassed by environmentalists for decades (Dobson 1990). However, the political stakes are high because by greening other policy areas environmental policy is in effect progressively undermining its own, hard-won independent existence. The irony is that if and when the environment is reflected fully in every aspect of policy making there will be no need for an identifiable 'environmental' policy.

See also: Environmental Determinism; Environmental Planning; Environmental Policy; Environmental Vulnerability; Environmentalism, Politics of; Environmentalism: Preservation and Conservation; Human–Environment Relationships; Land Degradation; Land Reclamation; Population Ecology; Population Pressure, Resources, and the Environment: Industrialized World; Sustainable Development

Bibliography

Benton T (ed.) 1996 *The Greening of Marxism*. Guilford Press, London

Dobson A 1990 *Green Political Thought*. Harper Collins, London

Doyle T, McEachern D 1998 *Environment and Politics*. Routledge, London

Downs A 1972 Up and down with ecology: The issue attention cycle. *Public Interest* **28**: 38–50

Dunleavy P, O'Leary B 1987 *Theories of the State*. Macmillan, Basingstoke, UK

Golub J 1998 *New Instruments for Environmental Policy in the EU*. Routledge, London

Haas P 1992 Epistemic communities and international policy coordination—introduction. *International Organization* **46**(1): 1–35

Ham C, Hill M 1993 *The Policy Process in the Modern Capitalist State*. Harvester Wheatsheaf, New York

Hanf K, Jansen A-I (eds.) 1998 *Governance and Environment in Western Europe*. Longman, New York

Ingram H, Mann D 1983 Environmental protection policy. In: Nagel S (ed.) *Encyclopaedia of Public Policy*. Dekker Publishers, New York

Jänicke M, Weidner H (eds.) 1997 *National Environmental Policies*. Springer, Berlin

John P 1998 *Analysing Public Policy*. Pinter, London

Jordan A J 1994 Financing the UNCED agenda: The controversy over additionality. *Environment* **36** (3): 16–20, 26–34

Jordan A J 2000 Environmental policy. In: Dunleavy P, Gamble A, Holliday I, Peele G (eds.) *Developments in British Politics 6*. Macmillan, Basingstoke, UK

Jordan A, O'Riordan T 1999 The precautionary principle in contemporary environmental policy and politics. In: Tickner J, Raffersberger C (eds.) *Putting the Precautionary Principle into Practice*. Island Press, New York

Lindblom C E 1977 *Politics and Markets*. Basic Books, New York

Lowi T 1972 Four systems of policy, politics and choice. *Public Administration Review* **32**: 298–310

McCormick J 1991 *British Politics and the Environment*. Earthscan, London

North R 1995 *Life on A Modern Planet*. Manchester University Press, Manchester, UK

Olson M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA

Pepper D 1996 *Eco-socialism*. Routledge, London

Parsons W 1995 *Public Policy*. Edward Elgar, Aldershot, UK

Richardson J (ed.) 1982 *Policy Styles in Europe*. Allen and Unwin, London

Skou Andersen M, Liefferink D (eds.) 1997 *European Environmental Policy: The Pioneers*. Manchester University Press, Manchester, UK

Weale A 1992 *The New Politics of Pollution Control*. Manchester University Press, Manchester, UK

Weale A, O'Riordan T, Kramme E 1991 *Controlling Pollution in the Round*. Anglo-German Foundation, London

Weale A, Willliams A 1992 Between economy and ecology: The single market and the integration of environmental policy. *Environmental Politics* **1**(4): 45–64

World Commission on Environment and Development 1987 *Our Common Future (the 'Brundtland Report')*. Oxford University Press, Oxford, UK

A. Jordan

Environmental Psychology: Overview

Environmental psychology is the study of the impact of the physical environment on people and the impact of people on the physical environment. This broad definition will be qualified in several respects

to distinguish environmental psychology from psychology in general and from other subfields with similar aims, in disciplines such as human geography, physiology, and sociology. Environmental psychology is also known as an area of applied psychology, although as in many other such areas a substantial proportion of the research is devoted to theoretical and methodological development. Today environmental psychology provides a theoretical and methodological foundation for environmental planning, design, and management.

The first section will briefly outline how environmental psychology developed historically. A problem conceptualization follows which is used to define focuses of the environmental psychology research. Finally, some of the major research focuses are described.

1. Historical Development

Transformation of the earth's surfaces by human beings has a long history. A primary aim is to create protection against threats to survival. In the 1960s the scientific study of the psychological effects of the humanmade environment began, largely motivated by a concern for the deterioration of human environments (e.g., Proshansky et al. 1970). A sensible early approach was to compare psychological effects of the humanmade environment with those of the natural environment in which the human species evolved (see *Natural Environmental Psychology*).

Research strategies in environmental psychology were first outlined by Craik (1968), based on the dominant strategy in personality assessment. Measurement issues were naturally at the forefront in the early research in environmental psychology. In particular, it was necessary for progress that methods were developed to assess physical settings outside the psychological laboratory as well as to assess how people react to such settings.

Important substantive research problems were originally defined by design professionals whose training made them sensitive to many subtle effects of the designed environment on human behavior. They used this knowledge in designing environments. However, in the face of criticism they became concerned about the scientific basis of their work. Social scientists observed a decline in the quality of urban life. To physiological psychologists this opened new meaningful research problems of how health and well-being are related to the physical environment. Cognitive psychologists concerned about the ecological validity of their research discovered in environmental psychology the possibility for investigating the acquisition, representation, and use of everyday knowledge (of the physical environment). Social psychologists saw in environmental psychology a way out of the laboratory to do meaningful research in the real world. Since the new focus on the physical environment did not stop them from analyzing the social environment, they offered influential conceptualizations in which the physical environment is presupposed to interact with the social environment in shaping human behavior.

The history of environmental psychology as well as its current status is documented in a series of comprehensive reviews published in *Annual Review of Psychology*, see, for example, Sundstrom et al. (1996). In 1987 the *Handbook of Environmental Psychology* was published under the editorship of Daniel Stokols and Irwin Altman. Following this benchmark publication several edited series of books have appeared with reviews of research in different subfields (e.g., Zube et al. 1987–97). Current research is primarily disseminated in three journals with slightly different emphases: *Journal of Environmental Psychology*, edited by David Canter, represents the psychological perspective; *Environment and Behavior*, edited by Robert Bechtel, covers the broader interdisciplinary field of environment–behavior research; and *Journal of Architectural and Planning Research*, edited by Andrew Seidel, is primarily an outlet for architectural and environmental design research. Several textbooks have also been written. The most comprehensive of these (Gifford 1997) cites approximately 5,000 studies.

2. Problem Conceptualization

A research problem in environmental psychology has three facets: the environment, the person or group, and the activity the person or group is engaged in. Any of these facets may be analyzed at different depths, ranging from what may be termed a molar to a molecular level.

So, for instance, the environment may be conceived of as consisting of different molar types (e.g., home, workplace). They are partly defined by their spatial extension, partly by their function. Any such type of environment may furthermore be analyzed at the level of stimuli (e.g., noise, temperature) impinging on people. These two levels are the primary concern to environmental psychology. At a still lower level the physical characteristics of stimuli may be analyzed. This level is sometimes the focus when changes of the other levels are the goal, for instance, in constructing new environments.

Sociodemographic characteristics of people occupying environments are seldom a primary concern to environmental psychology whereas psychological characteristics are. Such characteristics include more stable ones referred to as abilities, attitudes, and personality traits. More transient characteristics such as knowledge, skill, and motivation are also of interest to environmental psychology.

Activities are clearly important as people's means for achieving goals. They are sometimes analyzed at

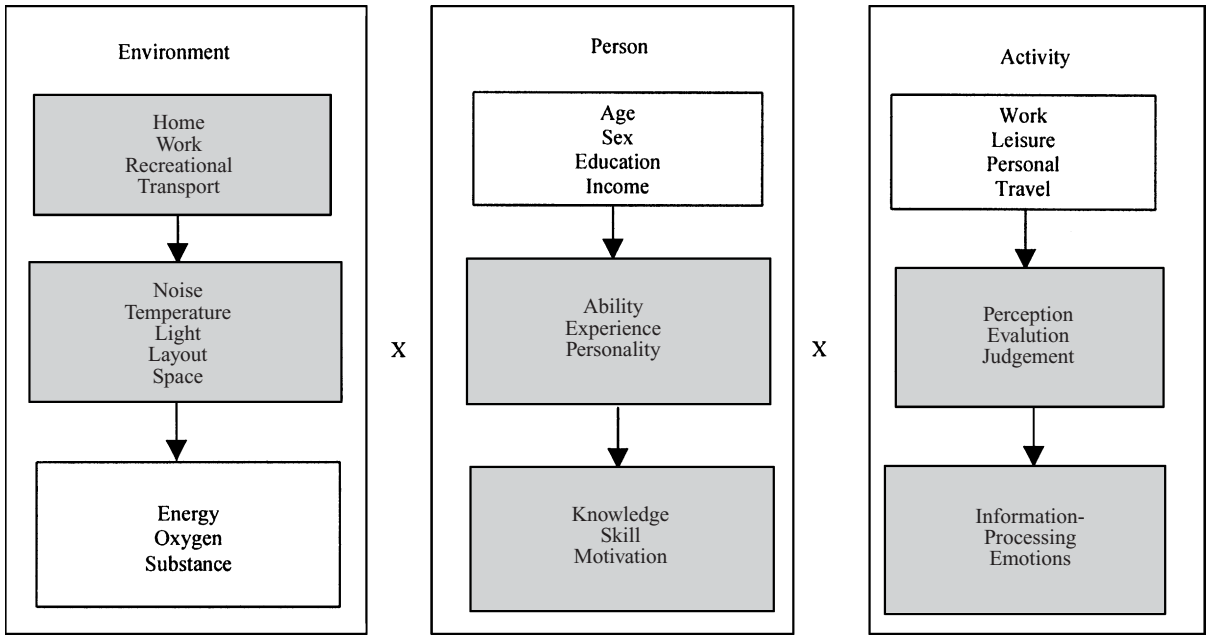

Figure 1
Facets of research problems in environmental psychology

the level of general function (e.g., work, leisure). However valuable this is, the main focus of environmental psychology is psychologically deeper. Here invariant components of all activities (e.g., perception, evaluation) are emphasized. Sometimes the focus is even on subcomponents of these component processes such as information processing stages or component emotional reactions.

At the psychologically most shallow level a study might focus, for example, on how sociodemographic factors are related to choice of leisure activities in the home. An in-depth study might instead focus on how noise levels affect information processing and mood in the workplace. Environmental psychology differs from psychology in general by its primary focus on molar environments (the filled boxes in Fig. 1), frequently also allowing longer exposures to the environment than is usual in most psychological research. Its focus on intraperson psychological processes distinguishes environmental psychology from other subfields in the disciplines of human geography, physiology, and sociology which study the effects of molar environments. It also differs from these subfields in going beyond sociodemographic factors with the aim of disentangling the ways ability, attitude, personality, knowledge, skill, and motivation modify how individuals react to the molar environment and how differences in intraperson psychological processes mediate such moderator effects.

3. *Research Focuses*

In this section three general areas of past and current environmental psychology research are described, highlighting in each a few major research focuses.

3.1 *Choice of the Environment*

Individuals either choose or are forced to choose where to be. This choice of environment may have profound future consequences for themselves and others. One research focus in environmental psychology is the choice itself. What characteristics of an environment make a person with certain characteristics choose this environment? What characterizes the choice process? Is it deliberate, habitual, or forced?

Research on environmental choice emphasizes different time scales. Some environmental psychology research targets residential choice and migration with possibly lifelong consequences (Aragonés et al. 2001; see also *Residential Environmental Psychology*). This is also a traditional area of research in social sciences such as economics, human geography, and sociology. A more frequent research focus with a shorter time scale is the study of people's daily choices of activity areas in the environment where they live, such as shopping locations, recreational facilities, and others. The demand for travel that such choices generate is being studied (see *Transportation, Psychology of*). In

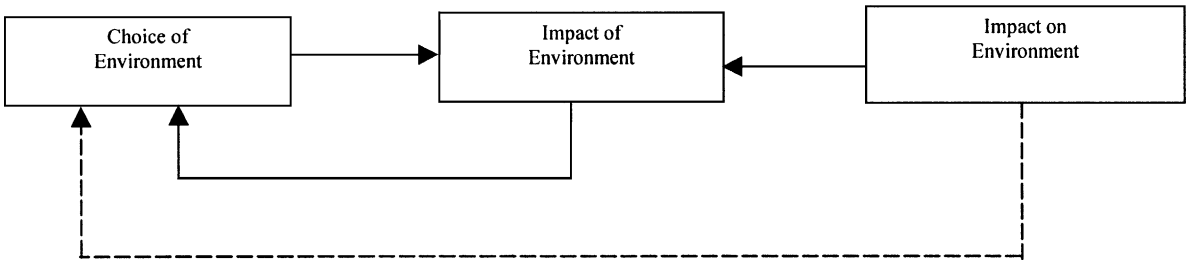

Figure 2
General areas of research in environmental psychology

addition, field and laboratory studies investigate acquisition, representation, and use of both spatial and nonspatial knowledge of environments. This research on cognitive maps has expanded in different directions. As evidenced by a recent major publication (Golledge 1999), it now attracts research interests also from cognitive and computer scientists, human geographers, neuroscientists, and zoologists. Another marker is the interdisciplinary journal *Spatial Cognition and Computation* under the editorship of Patrick Olivier and Stephen Hirtle. Primarily from the perspective that people are cost minimizers, Gärling and Golledge (2000) review and analyze research on how location choices depend on the degree and quality of spatial and nonspatial knowledge of the environment. A general conclusion is that this knowledge is always incomplete, is hierarchically organized, and therefore differentially accessible in the choice process. A complementary conceptualization of how knowledge and preferences may be integrated in environmental choices is offered by Canter (1991) who sees action goals as primarily driving preferences.

3.2 Impacts of the Environment

A large number of psychological and other effects of physical environments on human beings are assumed or expected and have been empirically demonstrated. Some of these effects are relatively direct, whereas others are modified by person characteristics and mediated by intraperson psychological processes. The social environment is another mediator of effects of the physical environment.

In large part because of its potential health outcomes, a dominating research focus in environmental psychology concerns the detrimental effects of environmental stressors such as noise, crowding, and extreme temperatures (Evans 1982) as well as threats of natural disasters, technological catastrophes, or accidents (Baum et al. 1983). Such effects are extensively documented in self-report, physiological, and performance measures. A complementary research focus which more recently has attracted attention is the study of beneficial effects of environments, for instance, the restorative (Kaplan and Kaplan 1989) and stress-reducing potentials of the natural environment (Ulrich 1983). The recurrent finding that sensory deprivation may have positive effects (Suedfeld 1980) bears theoretically on this current research with important practical implications.

Another research focus is the study of groups of people acting and interacting in physical settings (see *Community Environmental Psychology*). The impact of the physical environment is sometimes mediated by the social environment (e.g., Wicker 1979), for instance, when scarcity of space leads to excessive social interaction with adverse consequences. In other cases the physical environment facilitates social contact with positive effects on, for instance, residential satisfaction.

3.3 Impacts on the Environment

Since the early 1990s the scope of environmental psychology has broadened to include the study of methods to promote proenvironmental behaviors so that the detrimental impacts of people on the physical environment may be significantly reduced (Stern 1992). Extending a conceptual framework proposed by Russell and Ward (1982), a possible integration of this new research focus in environmental psychology with the traditional ones may be achieved as shown in Fig. 2. Forced or unforced choice of the environment has psychological consequences for the person. At the same time these choices have impacts on the environment, frequently negative when they are aggregated across many people. Sometimes there are regulatory feedback loops, such as when, for instance, crowding in an environment reaches intolerable levels. In other cases, however, the harmful effects only materialize in the future. Feed-forward loops (the broken arrow) should be provided so that people take these future consequences into account when making their choices. Yet, changing behavioral antecedents and consequences has not in general been a successful method. Current research therefore focuses on different sets of determinants of proenvironmental behaviors such as habits, attitudes, moral concerns,

and value orientations. This is likely to lead eventually to both a deeper understanding of behavioral change mechanisms and more effective methods of changing behavior.

4. Forecasting the Future

There are few reasons to believe that the current environmental problems of pollution and resource overuse will soon be solved. It is more likely that demand will increase on environmental psychology to contribute to solutions. Acceptance of new sustainable technologies such as zero-emission vehicles is one salient research focus in the future. One may also see more collaboration with economists, political scientists, and sociologists in developing effective methods for promoting proenvironmental behaviors (e.g., Foddy et al. 1999).

The development of telecommunication technology may have profound influences on how the physical environment will be organized in the future. Transformations of homes into workplaces are already taking place. Electronic shopping is increasing in frequency. To forecast the effects on people is a challenging future task.

The design of intelligent buildings with many automatically performed functions poses new research tasks in the immediate future. Although these new features are basically of benefit to users, people may still react negatively because of loss of perceived control. Also, automatic control may not always work—so users must know how to take over.

See also: Community Environmental Psychology; Environmental Determinism; Natural Environmental Psychology; Residential Environmental Psychology; Transportation, Psychology of; Wayfinding and Navigation Behavior

Bibliography

Aragonés J I, Francescato G, Gärling T (eds.) 2001 *Residential Environments: Choice, Satisfaction, and Behavior*. Greenwood Press, Westport, CT

Baum A, Fleming R, Davidson L M 1983 Natural disaster and technological catastrophe. *Environment and Behavior* **15**: 333–54

Canter D 1991 Understanding, assessing, and acting in places: Is an integrative framework possible? In: Gärling T, Evans G W (eds.) *Environment, Cognition, and Action*. Oxford University Press, New York, pp. 191–209

Craik K H 1968 The comprehension of the physical environment. *American Institute of Planners* **34**: 29–37

Evans G W (ed.) 1982 *Environmental Stress*. Cambridge University Press, New York

Foddy M, Smithson M, Hogg M, Schneider S (eds.) 1999 *Resolving Social Dilemmas*. Psychology Press, Philadelphia, PA

Gärling T, Golledge R G 2000 Cognitive mapping in spatial decision making. In: Kichin R, Freundschuh S (eds.) *Cognitive Mapping*. Routledge, London, pp. 44–65

Gifford R 1997 *Environmental Psychology*, 2nd edn. Allyn and Bacon, Boston

Golledge R G (ed.) 1999 *Wayfinding Behavior*. Johns Hopkins University Press, Baltimore, MD

Kaplan R, Kaplan S 1989 *The Experience of Nature*. Cambridge University Press, New York

Proshansky H M, Ittelson W H, Rivlin L G (eds.) 1970 *Environmental Psychology*. Holt, Rinehart, and Winston, New York

Russell J A, Ward L M 1982 Environmental psychology. *Annual Review of Psychology* **33**: 651–88

Stern P C 1992 Psychological dimensions of global environmental change. *Annual Review of Psychology* **43**: 269–302

Stokols D, Altman I (eds.) 1987 *Handbook of Environmental Psychology*. Wiley, New York, Vols. 1–2

Suedfeld P 1980 *Restricted Environmental Stimulation*. Wiley, New York

Sundstrom E, Bell P A, Busby P L, Asmus C 1996 Environmental psychology 1989–1994. *Annual Review of Psychology* **47**: 485–512

Ulrich R S 1983 Aesthetic and affective response to natural environment. In: Altman I, Wohlwill J F (eds.) *Human Behavior and Environment*. Plenum, New York, Vol. 6, pp. 85–125

Wicker A W 1979 *An Introduction to Ecological Psychology*. Brooks/Cole, Monterey, CA

Zube E H, Moore G T, Marans R W 1987–97 *Advances in Environment, Behavior, and Design*. Plenum, New York, Vols. 1–4

T. Gärling

Environmental Risk and Hazards

Environmental risks and hazards are a major focus of research, general concern, and public policy; and they provide an important arena for interdisciplinary collaboration between social scientists, natural scientists, and engineering. The terms encompass a wide range of definitions and practices, with considerable evolution of their meaning and use over time. This article reviews the definitions and intellectual history of each term, and considers several important developments and debates in current theory, methods, and research from the perspective of the social sciences. One of the distinctive characteristics of social science research on environmental risks and hazards is the strong influence on policy making and environmental management.

The overall cost of natural hazards, expressed either in monetary terms or through property loss or damage, is extremely high, estimated at more than $40 billion per year worldwide. Most of this is accounted for by floods (40 percent), tropical cyclones (20 percent), and drought (15 percent). Roughly two-thirds of the total cost stems from loss of life and property, and the rest is spent on attempts to reduce risk. On average, more

than 250,000 people die from hazards in a typical year, most of them in developing countries in Latin American, Africa, and Asia.

1. Definitions

Everyday use generally defines risk as the chance of loss or injury, and hazard as a source of danger. The research literature uses the terms in more complex and sometimes contradictory ways, although precise interpretations are provided in some regulatory frameworks.

Environmental hazards are defined as extreme events or substances in the Earth and its ecological system that may cause adverse effects to humans and things they value (NRC 1996). The idea of hazard includes both the event and its consequences. Environmental hazards include geophysical and meteorological phenomena such as earthquakes, droughts, and hurricanes, often called 'natural' hazards, as well as pollution problems and other 'technological' hazards. Most scholars argue that an event does not become a hazard until humans are exposed to it, and thus that hazards result from the interaction of humans and extreme events.

Environmental risks are often defined as the product of a hazard and the likelihood of its occurring, using a simple formula that defines a risk as the product of the probability of an event and its severity measured in terms of the population exposed, and the nature of the consequences. For example, the risks of a nuclear accident have been expressed in terms of the chance (e.g., one in one million) per exposure unit (e.g., a person exposed to a certain level of radiation) getting cancer (consequences).

2. Research Traditions

Although people have long been concerned with the documentation of disasters and other environmental dangers, formal social science approaches to natural hazards are often associated with a group of scholars based at the University of Chicago and led by geographer Gilbert White. White set out to understand the processes by which humans become exposed to natural hazards, such as floods, and the role that human adjustments may make in reducing hazard losses. Prior research had focused on engineering approaches to reducing hazard losses, such as constructing levees and dams, but White's work showed how such structures may actually increase losses by encouraging humans to occupy flood plains, and channeling water to create greater damage downstream (White 1958). A framework from 'human ecology' was employed to examine interactions between humans and extreme events in terms of human behavior and perceptions. White's students were able to demonstrate how views and adjustments to hazards such as drought (Saarinen 1966) and flood (Kates 1962) were influenced by the socioeconomic and personality characteristics of individuals, and by their knowledge and experience of previous events. The major categories of adjustment included modifying the environment (e.g., cloud seeding during drought), modifying human behavior (planting different seeds), or redistributing losses (crop insurance). The concept of adjustment and response was important because these human actions could reduce the physical and social impact of an extreme event.

This framework drew on behavioral theories of decisions based on bounded rationality, where individuals made decisions based on limited knowledge within the constraints of a social system. The research results influenced education and warning programs and legislation, such as hazard zoning, that attempted to reduce occupancy of hazardous locations, and improve understanding and response to hazard warnings. For example, the US National Flood Insurance Program attempts to regulate land use in flood plains as a way of reducing social and economic damage from flooding. After White moved to the University of Colorado he started the Natural Hazards Research and Applications Information Center which coordinated several national and international assessments of various hazards, and met regularly with stakeholders to discuss research results (White 1974).

Although geography provided the main disciplinary home for hazards research, sociology and anthropology were also very active (see *Disasters, Sociology of*). Sociologists, led by Enrique Quarentelli, developed a different approach to understanding natural hazards that used theories of collective behavior and social organization to understand response to disasters (Quarantelli 1978). Anthropology contributed ideas about culture and development, including research that showed the significance of indigenous knowledge and cultural traditions in attitudes and responses to hazards and risks in several regions of the world (Douglas and Wildavsky 1982).

In response to devastating droughts in Africa and earthquakes in Latin America as well as the influence of Marxist thinking in the social sciences in the 1970s, hazard losses began to be blamed less on individual decisions and more on larger scale political and economic structures (O'Keefe et al. 1977, Hewitt 1984). This critical, structural, or political economy perspective focused on processes that forced people to live in hazardous environments, limited their access to information and resources, and impoverished people to levels where they could not adjust or recover from extreme events. For example, studies of earthquake impact and recovery in Nicaragua and Guatemala showed that working class and poor households were living in fragile housing, often on steep slopes, as a result of poverty and unequal land tenure systems, and were unable to gain access to disaster relief that was

directed to the more powerful members of society (Cuny 1983).

An enduring concept to emerge from this work was that of 'vulnerability,' defined simply as the potential for loss in different places or among different social groups. Broader definitions encompass the characteristics of a person or group in terms of their capacity to anticipate, cope with, resist, and recover from the impacts of natural hazard (Blaikie et al. 1994). Vulnerability is conditioned by both biophysical and social conditions, including topography, poverty, access to information and insurance, gender, and ethnicity, and involves both external structural factors and individual capacity to cope with extreme events (see *Environmental Vulnerability*). It can be seen as the opposite of security, and as influenced by underlying dynamics of power relationships in society and access to resource entitlements (Watts and Bohle 1993).

Vulnerability studies often involved the detailed analysis and mapping of the distribution of sensitivity and resilience to hazards across geographic space, and between social sectors. Thus, they provided a deeper understanding of the distributive effects of hazards and responses to them. Women, children, and the elderly were found to be more vulnerable than the general population to many hazards.

For example, vulnerability to drought in Africa was associated with differential access to resources or entitlements to land, food, or assistance at a variety of scales ranging from the household to the global food system refracted through markets, environmental stresses, and multilateral trade and aid agreements. Influenced by Sen's analysis of famine and entitlements (Sen 1981) vulnerability emerges from the interaction between drought and individuals' abilities to produce food themselves, participate in the reciprocal or market exchange of labor and food, or gain effective access to relief and development assistance. The more frequently vulnerable groups include rural smallholders on marginal land, pastoralists, rural wage laborers, the urban poor, refugees, and displaced people, as well as women, children, the infirm, and the elderly (Downing 1996). Insights from such analyses have been used to develop vulnerability maps and famine early warning systems to assist in mitigation and relief efforts.

Other important concepts include that of 'hazardousness of place,' whereby the totality of hazards in a particular geographic location is analyzed (Hewitt and Burton 1971), and of hazards as part of everyday life in contrast to isolating them as departures from normal or accidents (Hewitt 1984). Both concepts point to the importance of viewing hazards within the larger context of social, political, and economic structures at scales ranging from the household to the community to the state. A growing interest in the hazards of the world's major urban areas has shown how the combination of population growth and migration, poor infrastructure, economic growth, inadequate institutions, and poverty is creating vulnerability to disasters (Mitchell 1999). Half the world's largest cities are located on earthquake belts or tropical cyclone tracks. The 1995 earthquake in Kobe, Japan, caused more than 6,000 fatalities and over $120 billion in economic loss. The Izmit, Turkey, earthquake in 1999 caused 20,000 fatalities and an estimated $20 billion in economic loss.

3. *Technological Hazards and Environmental Risk Assessment*

With the growth of the environmental movement and awareness about pollution problems, some researchers developed an interest in technological hazards, stimulated partly by the need for studies that would support the decisions of US regulatory agencies such as the Environmental Protection Agency or the Nuclear Regulatory Commission, and by a series of accidents and serious pollution incidents. Regulation of carcinogenic pollutants required careful assessments of the risks of environmental chemicals for human health. A 1974 study of nuclear reactor safety provided one of the first major examples of probabilistic environmental risk assessment, and prompted a vigorous debate about the appropriate methods for measuring and regulating pollution and accident risks (USAEC 1974).

Initial research on environmental risks focused on techniques for quantitative risk assessment that attempted to estimate the expected incidence of illness or death in human populations exposed to chemicals and other substances. The two major approaches were statistical or actuarial analyses based on past experiences, and laboratory experiments using animals or cell cultures. For example, risk assessment was used to estimate the expected rate of illness or death in a human population exposed to a hazardous chemical based on the number of experimental animals affected by various doses of the chemical as measured in laboratory experiments. Limitations of these environmental risk assessments included inadequate data or controls on past empirical occurrences, problems in extrapolating from high short-term experimental doses to longer-term exposure, and from animals to humans, and the likelihood of real-world multiple and interacting stresses. Risk estimates were often used to compare risks in order to allocate resources for risk management using techniques that compared a particular risk (such as coal-fired electricity generation) to risks of alternative technologies (e.g., hydroelectricity), everyday life (e.g., driving), or to the economic costs and benefits of the technology (e.g., health costs and low-cost electricity). Quantitative risk analysis has at its core the belief that decisions about technology can be made according to a monetary calculus in which costs, determined by quantifying loss of life and productivity, are balanced against total 'social

benefits.' Risk–benefit analysis became controversial because it involved calculations that included placing often inconsistent values on human lives, injuries, or ecosystem damage. Scientific rankings of risk also differed with the views of the public, as shown in the keynote article by Starr (1969) that showed the public to be 1000 times more willing, on average, to accept voluntary risks (e.g., driving) than those imposed upon them (e.g., pollution), and that the acceptability of risk was acceptability of a risk relative to the third power of the perceived benefits.

Scholars from the environmental hazards tradition such as Kates and Kasperson raised questions about public perception of risk, together with an important body of work by psychologists such as Slovic and Fischhoff (Fischhoff et al. 1981, Kates and Clark 1977, Slovic 1987). Research demonstrated that people rank risks not only on scientific studies of the probability of harm, but also on how well the process is understood, its visibility and association with cancer, the degree of catastrophe, how equitably the danger is distributed, how well individuals can control their exposure, and whether risk is voluntary or imposed. These studies helped to explain why the public was so concerned about the risks of nuclear power or toxins in the environment in comparison to more frequent or serious risks from driving or smoking. Kasperson, one of the pioneers of research into technological hazard, proposed the concept of the 'social amplification of risk' and which suggested that the actions of the media, government, and nongovernmental organizations, as well as disputes among scientists, can significantly increase or decrease public risk concerns (Kasperson et al. 1988).

As a result of the relation to decision making, the processes of risk assessment have become a continuing and frequent preoccupation of advisory councils and committees in many countries and jurisdictions, such as the US National Research Council and the European Community.

4. *Emerging Issues*

Another set of social science research activities focused on the distribution of risks among social groups and across landscapes. This work showed the unequal distribution of the benefits and risks from technologies, and demonstrated the disproportionate exposure of poor and minority populations to environmental risks and hazards. This 'environmental justice' literature has continued to influence research on environmental risks and hazards (see *Environmental Justice*).

Mapping the distribution of risks and vulnerability was facilitated by the development of computer-aided geographic information systems (GIS), which allowed for detailed spatial information on both physical and social variables to be integrated into databases and maps. These allowed the identification of hazard zones, and enabled a more rapid response to natural and technological disasters (Dangermond 1991). For example, GIS have been used to map earthquake hazards, droughts, and exposure to pollutants, and to facilitate more rapid response to emergencies by highlighting the spatial patterns of vulnerability and transport routes, as well as the location of key technical and human response and relief resources.

Social scientists also raised important issues about power and communication in risk assessment, analyzing the process by which certain groups were legitimized as experts, and the roles of scientists, government authorities, and the media in defining and managing risks (Beck 1992, Jasanoff 1993, Lash et al. 1996). The social constructionist perspective drew on discourse analysis to examine events such as the Bhopal chemical accident in India, and responses to the risks of bioengineering and climate change in terms of the framing of the debate, the use of language and imagery, and the use of science by different interest groups.

Growing demands for public participation in decisions about environmental risk fostered research on the role of nongovernmental organizations, and on models for involving stakeholders in both risk estimation techniques and management decisions. For example, research into the design and effectiveness of deliberative processes for formal and informal participation in decision making about risks provides a critical examination of different models of hearings and other forms of public input into decisions about location of hazardous facilities or development of pollution regulations.

Awareness of the potential of new global-scale environmental changes to alter the severity and pattern of environmental risks led some scholars to situate their research in the context of global climate and land use changes. Projections remain uncertain about how global warming might alter the frequency and intensity of meteorological hazards, or change the incidence of pollution at the local level. Nevertheless, researchers employed scenarios to estimate changes in hurricane risks associated with sea-level rise, or drought and epidemics associated with climate changes and their results contributed to international assessments of global change, such as those of the Intergovernmental Panel on Climate Change (IPCC 1991). The concept of vulnerability was another important contribution of risk and hazard research to global change studies, because it suggested the importance of socioeconomic context and conditions in determining the impacts of climate change or the causes of deforestation, and highlighted issues of distribution of impacts and the ability to adjust to changes (Liverman 1994).

Hazards and risks are created, responded to, and studied in a highly international context, whether it be the role of trade or climate change in transferring risks

from one region to another, the complex patterns of governmental and nongovernmental flows of disaster relief and development assistance, or the collaborations of scientists from different countries in research. International organizations such as the UN Disaster Relief Organization, the International Council of Scientific Unions, and various multinational environmental and charitable groups such as Oxfam, coordinate research and response to environmental risks and hazards. For example, the 1990s were declared the International Decade for Natural Disaster Reduction (IDNDR), with the aim of reducing hazard losses worldwide through improved prediction and reduction of vulnerability.

See also: Accidents, Normal; Adaptation, Fitness, and Evolution; Air Pollution; Climate Impacts; Desertification; Disasters, Sociology of; Environmental Justice; Environmental Planning; Environmental Security; Environmental Vulnerability; Organizational Errors: Catastrophic; Plagues and Diseases in History; Precautionary Principle; Water Pollution: Marine; Water Pollution: Terrestrial and Groundwater

Bibliography

Beck U 1992 *Risk Society: Towards a New Modernity*. Sage, London

Blaikie P M, Cannon T, Davis I, Wisner B 1994 *At Risk: Natural Hazards, People's Vulnerability, and Disasters*. Routledge, London

Cuny F C 1983 *Disasters and Development*. Oxford University Press, New York

Cutter S L 1993 *Living with Risk: The Geography of Technological Hazards*. Routledge/Chapman and Hall, New York

Dangermond J 1991 Application of GIS to the International Decade for Natural Hazards Reduction. *International Conference on Seismic Zonation, Proceedings* **4**: 445–68

Douglas M, Wildavsky A 1982 *Risk and Culture: An Essay On The Selection of Technical and Environmental Dangers*. University of California Press, Berkeley, CA

Downing T E (ed.) 1996 *Climate Change and World Food Security*. Springer-Verlag, Berlin

Fischhoff B, Lichtenstein S, Slovic P, Derby S L, Keeney R L 1981 *Acceptable Risk*. Cambridge University Press, Cambridge, UK

Hewitt K (ed.) 1984 *Interpretations of Calamity*. Allen & Unwin, Boston, MA

Hewitt K, Burton I 1971 *The Hazardousness of a Place: A Regional Ecology of Damaging Events*. University of Toronto Research Publication No. 6, Toronto, Canada

Intergovernmental Panel on Climate Change 1991 *Climate Change: The IPCC Response Strategies*. Island Press, Washington, DC

Jasanoff S 1993 Bridging the two cultures of risk analysis. *Risk Analysis* **13**(2): 123–9

Kasperson R E, Renn O M, Slovic P, Brown H S, Emel J, Goble R, Kasperson J X, Ratlick S 1988 The social amplification of risk: a conceptual framework. *Risk Analysis* **8**(2): 177–87

Kates R W 1962 Hazard and choice perception in flood plain management. Research paper no. 78, Dept. of Geography, University of Chicago, Chicago

Kates R W, Clark W C 1977 Managing technological hazard: research needs and opportunities: A report of three workshops. Institute of Behavioral Science, University of Colorado, Boulder, CO

Lash S, Szerszinski B, Wynne B (eds.) 1996 *Risk, Environment and Modernity: Towards a New Ecology*. Sage, London

Liverman D M 1994 Vulnerability to global environmental change. In: Cutter S (ed.) *Environmental Risks and Hazards*. Prentice-Hall, New York, pp. 326–42

Mileti D (ed.) 1999 *Disasters by Design: A Reassessment of Natural Hazards in the United States*. Joseph Henry Press, Washington, DC

Mitchell J K (ed.) 1999 *Crucibles of Hazard: Megacities and Disasters in Transition*. United Nations University Press, Tokyo, Japan

NRC (National Academy of Sciences, National Research Council) 1996 *Understanding Risk*: *Informing Decisions in a Democratic Society*. National Academy Press, Washington, DC

O'Keefe P, Westgate K, Wisner B 1977 Taking the naturalness out of natural disaster. *Nature* **260**: 566–7

Quarantelli E L 1978 *Disasters: Theory and Research*. Sage, Beverly Hills, CA

Saarinen T F 1966 Perception of the drought hazard on the Great Plains. Research paper no. 106, Dept. of Geography, University of Chicago, Chicago

Sen A 1981 *Poverty and Famines: An Essay on Entitlements and Famines*. Oxford University Press, Oxford, UK

Slovic P 1987 Perception of risk. *Science* **236**(4799): 280–5

Starr C 1969 Social benefit versus technological risk. *Science* **165**: 1232–8

USAEC (US Atomic Energy Commission) 1974 *Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants*. National Technical Information Service, Springfield, VA

Watts M J, Bohle H G 1993 The space of vulnerability: the causal structure of hunger and famine. *Progress in Human Geography* **17**(1): 43–67

White G F 1958 Changes in urban occupance of flood plains in the United States. Research paper no. 57, Dept. of Geography, University of Chicago, Chicago

White G F (ed.) 1974 *Natural Hazards: Local, National, Global*. Oxford University Press, New York

D. Liverman

Environmental Sciences

The term 'environmental sciences' groups a number of disciplines which, in academia, are not usually found together. Ecology is a biological science, meteorology a physical science, and geology and physical geography also have their separate departments. Urban and regional planning is a social science. Yet all these sciences have important commonalties.

First, they have shared historical roots. All entered the modern phase of their development around 1800. The term 'Humboldtian sciences' has been given to sciences which seek, as all of these fields do, to integrate

large numbers of facts and measurements in a holistic image of their subject matter. There is a strong link with Romanticism, which, in the sciences, aimed at conceiving large wholes and finding their underlying patterns of organization. Yet, during a century or more, the various Humboldtian sciences increasingly developed within their own disciplinary boundaries.

However, from 1960 onward, the environmental sciences have again moved closer toward each other. Societal pressure played a large part therein, and the very word 'environmental' developed its momentum in the public realm before it entered scientific discourse to the extent that it has now (Jamison et al. 1990). Ecology was first science to receive its status as 'the' environmental science, not long after the publication of Rachel Carson's *Silent Spring* in 1962 and more visibly after Earth Day in 1970. Governmental science policy in many western countries has been willing to recognize ecology's centrality.

Meteorology much later than ecology rose to the status of an encompassing environmental science. It acquired this status around 1985. The 'greenhouse effect' occasioned primarily by human societies' burning of fossil fuel, and the resulting 'global warming,' have inspired governmental policies on a number of levels. These policies in turn have exercised profound impacts on priorities in science policy. Many other sciences have oriented their agenda to issues related to global warming, and are now also identified as 'environmental sciences.'

1. Ecology

1.1 Alexander von Humboldt

Alexander von Humboldt (1769–1857) had received a university education in geology. An Inspector of Mines, he was embarking on a career in the service of the Prussian government, when the inheritance of his family fortune enabled him to give up his position. By then, he had been in contact with the naturalists on board of Captain James Cook's scientific expeditions in the Pacific. In 1798, Humboldt began his own journey to Latin America. Upon his return in 1804, he became the most famous man in Europe after Napoleon. Travel and naturalistic research in the tropics were given an enormous boost through the accounts of his voyage. Among Humboldt's many admirers may be counted the Dutch King and government, who sent a scientific expedition on a journey through Java in 1820, a first in Dutch colonial history.

1.2 Humboldtian Ecology

An important conceptual innovation of Alexander von Humboldt's was the transformation of 'landscape' into a subject matter for science. Previously, the word landscape referred to a particular genre in painting, and Humboldt made good use of the painterly gaze, which confers unity on what it sees. Humboldt thus grasped the landscape in effect as an individual being. Every region of the globe had its own visual characteristics and beauty. Foremost in the appearance of a landscape was its 'vegetation,' another new concept, since it grouped plants according to the place where they lived rather than following their taxonomical order. Humboldt discovered that very different species not only occurred together, but that there were social relations among them, and he called these groupings of plants 'associations' (Nicolson 1987).

The first modern textbook of ecology by the Danish botanist Eugenius Warming appeared in 1895, with the English translation appearing in 1909. Warming retained Humboldt's original precepts on plants in social associations. He worked them out in a precise manner, exposing the diversity of degrees and styles of interdependence, which exist in nature. Josias Braun-Blanquet in Montpellier pursued this line of thought, and he and others in the so-called Zürich–Montpellier school produced a comprehensive system of Europe's plant associations. This vegetation science would play a major role in the ecological underpinning of nature conservation practices in many countries of continental Europe (Nicolson 1989). Workers in the Zürich–Montpellier school could identify which rare species belong to a certain association—hence they knew when their nonoccurrence should cause concern. In the UK, a vegetation ecology of a different style has been equally useful for nature conservation, and its practitioners were consulted by the government to designate nature reserves since the late 1940s (Bowler 1992).

1.3 Darwin and Evolutionary Ecology

Charles Darwin's famous *Origin of Species* of 1859 is as much an ecological book as it is about evolutionary theory. Darwin owed much to Humboldt's holistic ecology, as Darwin saw the various species of animals and plants forming an 'entangled' whole. But whereas Humboldt stressed harmony in nature, Darwin regarded nature much more as a creative force. Nature continually disturbs settled equilibrium through the struggle of species scrambling for a place (Worster 1985). Evolutionary ecologists stress interactions between species (populations). They do not regard the climate as the deterministic factor in the make-up of animal communities, as ecosystem ecologists tend to do.

Population dynamics has strong ties with applied ecology: economic entomology (the study of pest insects in agriculture), and fishery ecology. Already by 1900, ecologists were working on biological pest control, that is, the method of holding one insect species in check by another. The publication of *Silent Spring* gave this research an important political boost (Perkins 1982, Palladino 1993).

Several approaches in population dynamics, in particular those of Robert May and C. S. Holling, have enlarged its theoretical basis in diverse ways. Both developed new mathematical techniques, which tackled the occurrence of more or less dramatic changes in ecological systems. Though interesting leads exist with regard to nature conservation, notably in the work of Daniel Botkin, evolutionary ecology has been so far less influential in this domain than systems ecology.

1.4 Ecosystems and Systems Ecology

Systems ecology is the mathematical, computer-implemented form of ecosystem ecology. The latter assumes that large entities such as landscapes are functionally integrated wholes and it attempts to uncover their structure and functioning.

The word 'ecosystem' was coined in 1935 by Arthur Tansley, a British ecologist. Tansley wanted to place the older association or community oriented ecology on a mechanistic footing. This idea remained largely programmatic until the 1950s, when Eugene Odum pulled together some theoretical approaches and physical methodologies. Odum's synthesis launched ecosystem ecology. It incorporated an emphasis on 'productivity,' which was added to the analysis of food web relations first pioneered by Charles Elton in the UK. Odum made use of newly available physical methodologies, most importantly the use of radioactive tracers. His approach was based on a recent discovery that several species accumulate radioactive isotopes, as they get picked up in minute quantities from the environment. Radioactive tracers permitted the elucidation of food webs both qualitatively (who eats what) and quantitatively (Kwa 1993).

The next step was to turn these quantified systems of food web relations into a more formal model that could be implemented on a computer. Through computer simulation of ecosystem models Odum's fellow ecosystem ecologists hoped to find the key to the rational exploitation and control of natural resources. They hoped to use the models to help them establish the conditions under which ecosystems preserve their stability and to identify the natural mechanisms that enable them to do so.

The first model of a complete ecosystem was built in 1959, an electric model, a material analog of an ecosystem. The electrical current was the equivalent of the carbon cycle; resistors were the equivalents of various categories of groups of plants and animals, etc. The first whole ecosystem model, implemented on a digital computer, followed in 1965. These models were then gradually expanded in order to achieve closer similarity with real ecosystems. Their mathematical complexity in terms of the number of variables reached a temporary zenith in 1973. It was hoped that the models would mimic the behavior of real ecosystems at the scale of a good part of a continent. This expectation may now seem naive, but we may note that during the period 1959–1973 mainframe computers gradually became accessible to larger segments of the academic community. Many expected to use this expanded computing capacity to find the solutions to hitherto unsolvable complex problems. The prime example here is the 'Limits to Growth' model developed by Dennis Meadows for the Club of Rome in 1972 (Ashley 1983).

One important assumption built into the models was that ecosystems would be steered by climatic factors such as precipitation, temperature, and hours of sunshine. The models thus represented ecosystems as automatically steered machines, which would return to equilibrium after not too big a disturbance. However, if the functioning of the ecosystem was threatened, the ecosystem manager should make sure that equilibrium was safely reached again.

Around 1967, the US Congress became convinced that systems ecologists would be able to develop an early warning system, so that natural catastrophes could be avoided. The environmental movement also placed much trust in ecology as the science that could save humanity from its own excesses. Everywhere in the western world, measures to promote environmental hygiene were framed in a systems ecological vocabulary. As early as 1971, ecologists began to feel the pressure of popularity and increasingly voiced disclaimers about their ability 'to save the planet.'

But by then professional ecologists were no longer needed for this purpose. Systems ecology had placed an emphasis on the circulation of matter and energy in the ecosystem, and other sciences could now push ahead with charting the various cycles of chemical substances to frame problems of pollution abatement and energy conservation. Another case in point is the carbon cycle, which at one point was understood as an ecological concept and which now has become a subject matter of meteorology.

2. Meteorology

Humboldt had invented the isotherms, lines on a map, which connect points with equal mean temperature. Together with isobars and a number of other graphic representations, such as of storm tracks, isotherms formed the basis of weather forecasting well into the twentieth century. Progress in meteorology was based on an ever-expanding system of atmospheric measurements. But the skillful combination of those involved many rules of thumb.

2.1 Vilhelm Bjerknes and the Bergen School of Meteorology

Modern meteorology was founded by Vilhelm Bjerknes and his 'Bergen school' in Norway. He visualized the atmosphere as consisting of three-

dimensional air masses moving around the globe, different in temperature, density, and humidity. An important step was achieved by Bjerknes' son Jacob, who developed a model of the extra-tropical cyclone and how it moves. Bjerknes was able to exploit an existing military observational network for weather forecasting in western Norway, as it enabled him to follow the movement of the cyclones. It brought him considerable popular support in the area. The major breakthrough occurred in 1919, when the concept of the 'polar front' was born. It would take several decades before the world's meteorologists were persuaded to 'see' a battle line, separating cold air masses from the north and warm air masses from the south. Bjerknes thought that this line stretched along the entire Atlantic, or indeed around the entire globe. Along the polar front the cyclones, or low-pressure areas, were formed, which would hit the coasts of Europe with rain and storm. Bjerknes put weather forecasting on a new footing. The analysis of polar fronts for the purpose of forecasting relied on large networks for atmospheric data gathering. Bjerknes pushed for it internationally, thereby also promoting the Bergen school ideas (Friedman 1989).

2.2 The Adoption of the Bergen School Ideas in the US

The Bergen school ideas received a decisive boost when Carl-Gustav Rossby (b. 1898), a Swede who had worked with Bjerknes in Bergen, came to work at the US Weather Bureau during 1926–1927. He encountered stiff resistance to the Bergen school ideas, but he was able to recruit some followers.

In the late 1930s Rossby, by then at MIT, developed a method of calculating the large-scale motion of the atmosphere, based on the movement of long waves (later to be called Rossby waves). He put the westward movement of the extra-tropical cyclones more rigorously into the context of the general circulation of the Earth's atmosphere, thus explaining semiquantitatively how the cyclones moved. On the basis of Rossby's work, precision in forecasting methods was greatly enhanced in succeeding years.

2.3 Numerical Prediction of the Weather on the Computer

John von Neumann, a mathematician, is one of the main architects of the modern computer. During World War II he had been working on the atom bomb at Los Alamos. After the war he initiated a project to build an 'all-purpose, automatic, high speed electronic computing machine.' Von Neumann also formulated the problems that computers were to solve among them the modeling of the atmosphere. The reason why von Neumann chose meteorology may have been the military importance of weather forecasts. Jule Charney became von Neumann's collaborator in 1948. In early April 1950, the first successful numerical weather prediction was completed on the basis of Charney's simple two-dimensional model of the atmosphere. Routine numerical weather prediction began by 1955.

The early models by Charney were intellectually derived from the Bergen school. However, in the mid-1960s these models were replaced in daily forecasting by models based on the so-called primitive equations which were basic physical equations describing the atmosphere. As a result, the original Bergen 'fronts' moved to the background in numerical meteorology.

2.4 The Attempt to Modify the Weather and the Climate

Between 1946 and 1973, there was a brief flourishing of the field of weather and climate modification. In 1946, Vincent Schaeffer and Irving Langmuir, working for General Electric in the US, discovered by chance that small quantities of frozen carbon dioxide could induce a cloud to release its contents. The many experiments that followed had a simple experimental format: clouds were sprayed with a reagent (most often silver iodide), usually from an airplane (Byers 1974). By 1965, just enough progress had been achieved in the eyes of the larger scientific community to endorse the idea of weather modification. Also, mergers took place between experimental cloud seeding and theoretical modeling of clouds and the atmosphere. One of these projects had been initiated at the National Hurricane Research Laboratory in Miami, which in 1963 had seeded two hurricanes. Results were evaluated as 'encouraging if not conclusive.' Around the same time, the US and the UK began to worry about Soviet plans to divert Siberian rivers away from their flow into the Northern Ice Sea. The West's need for its own expertise on this matter provided an important impetus for the development of General Circulation Models of the whole Earth's atmosphere.

Ten years later, funding for weather and climate modification was considerably reduced, despite the meteorologists continuing best efforts. But opposition had grown. There was increasing pressure by environmental activists against cloud seeding. Also, projected figures of jet exhausts of supersonic aircraft induced worries about anthropogenic damage to the ozone layer. In general, the long prevalent optimism about human interference in the environment was giving way to fears of irreparable damage.

3. Global Warming and Climate Change

Fear of unintended catastrophe is at the heart of the current concern about global warming. Never before has science revealed a risk, which is at once of such tremendous proportions and is surrounded by so many uncertainties (Jasanoff and Wynne 1998).

The General Circulation Models of the Earth's atmosphere provide the primary evidence for the existence of a tendency toward global warming, later reconceptualized as climate change, and also that it is primarily caused by the human use of fossil fuels. The evidence is assessed by the Intergovernmental Panel on Climate Change (IPCC), an international body of scientific experts appointed by national governments to provide policy-relevant advice. The IPCC's aim to seek broad consensus around scientific matters is contested by some scientists. To others, it is justified because the stakes are so high.

General Circulation Models are now becoming linked to models of the oceans, and the influence on the atmosphere by the vegetation masses of the earth is under investigation. If research on the causes of global warming exceed the disciplinary limits of meteorology, research on the possible effects of global climate change has drawn in a still larger cross-section of the sciences. Physical geography, regional planning, ecology, and environmental economics assess the impact of global warming on land-use and the landscape. A new science of integrated assessment has sprung up, which researches ways to combine all the available knowledge into so-called 'Integrated Assessment Models.' Models of this kind were used to advise the European Commission in proposing targets for energy-use reduction at the international negotiations in Kyoto, 1998.

Models appear to be the instrument of choice to master the complexity of the environment. Their range is wide, with scopes extending from the General Circulation Models to changes in the Delta of the Rhine as the expected result of increased rainfall caused by global warming. Over the past decades, the mathematical sophistication of models has been greatly improved, as have been the computers on which they are run. Yet, expectations with regard to precision in forecasting and the extent of possible human control over the environment have diminished in important respects as awareness of complexity and uncertainty has grown. Models of the environment have become highly ambiguous things, and it is easiest to say what they are not: truthful pictures of the world as it is and how it will evolve (Shackley 1997). Scientists usually take models as tools for further research, while policy makers need them to base their decisions on. Contested as they may be, models increasingly function as mediators between the environmental sciences and society.

See also: Environment and Development; Environmental Economics; Environmental Planning; Environmental Policy; Environmental Risk and Hazards; Environmentalism, Politics of; Environmentalism: Preservation and Conservation; Planning Issues and Sustainable Development; Planning Theory (North American); Research and Development in Organizations

Bibliography

Ashley R K 1983 The eye of power: The politics of world modeling. *International Organization* **37**: 495–535

Bowler P 1992 *The Fontana History of the Environmental Sciences*. Fontana Press, London

Byers H B 1974 History of weather modification. In: Hess W N (ed.) *Weather and Climate Modification*. John Wiley & Sons, New York, pp. 3–44

Friedman R M 1989 *Appropriating the Weather*. Cornell University Press, Ithaca, NY

Jamison A, Eyerman R, Cramer J 1990 *The Making of the New Environmental Consciousness*. Edinburgh, UK

Jasanoff S, Wynne B 1998 Science and decisionmaking. In: Rayner S, Malone E (eds.) *Human Choice and Climate Change: The Societal Framework*. Battelle Press, Columbus, OH, Vol. 1, Chap. 1, pp. 1–87

Kwa C 1993 Radiation ecology, systems ecology and the management of the environment. In: Shortland M (ed.) *Science and Nature*. British Society for the History of Science (BSHS Monograph 8), Oxford, UK, pp. 213–49

Nebeker F 1995 *Calculating the Weather*. Academic Press, New York

Nicolson M 1987 Alexander von Humboldt, Humboldtian science and the origins of the study of vegetation. *History of Science* **25**: 167–94

Nicolson M 1989 National styles, divergent classifications. *Knowledge and Society* **8**: 139–86

Palladino P 1993 On environmentalism: The origins of debates over policy for pest-control research in America, 1960–1975. In: Shortland M (ed.) *Science and Nature*. British Society for the History of Science, Oxford, UK, pp. 181–212

Perkins J H 1982 *Insects, Experts and the Insecticide Crisis*. New York

Shackley S 1997 Trust in models? The mediating and transformative role of computer models in environmental discourse. In: Redclift M, Woodgate G (eds.) *The International Handbook of Environmental Sociology*. Edward Elgar, Cheltenham, UK, pp. 237–60

Worster D 1985 *Nature's Economy: A History of Ecological Ideas*. Cambridge University Press, Cambridge, UK

C. L. Kwa

Environmental Security

Environmental security is concerned with the linkages between environment and security. There is no one definition as the topic is a fast-evolving one, but it is generally accepted to be a broadening of the traditional concept of national security to include environmental factors. Sometimes 'environmental' refers to the origin of a direct threat to security; for example, natural disasters or water scarcity, which can threaten the stability of a state or region. In other cases, the environment is seen as an indirect cause of violent conflict, as one of the factors (together with demographic, economic, political and social causes) in social impoverishment and inequity. Although a case can be made that political instability and conflict are them-

selves causes of environmental problems, most of the current work on environmental security examines the causal relationship in the other direction: from environmental causes to their effects on security, at national, regional and global levels. Thus, the key context is to look at the effects of environmental factors on security and not the other way round.

1. Evolution of the Concept

The concept of environmental security is diffuse because it has its origins in several different evolutionary strands and research traditions. One origin is in an expansion of the concept of military security. Another is the development of a welfare concept known as 'human security'. A third is an extension of the work in natural hazards, environmental risk assessment and the notion of ecological and geopolitical 'vulnerabilities', which has led to the concept of critical zones for environmental security. The report of the World Commission on Environment and Development (WCED 1987), which is best known for elaborating the concept of 'sustainable development,' highlighted environmental stress as a potential cause of military conflict. The 1986 nuclear accident at Chernobyl was a clear message to many people that human health and ecosystem integrity were interdependent. In 1987, President Gorbachev of the USSR proposed that 'ecological security' should be a top priority for international relations. The concept had entered the mainstream of academic and political discourse.

1.1 Military Security

The concept of national security to include nonmilitary aspects has evolved over the past few decades, as it has been accepted that issues such as ethnic and religious differences and environmental degradation affect national stability and the likelihood of conflict. Ullman (1983) argued for an expanded definition of security in which any action or sequence of events was a threat to national security if it had major and relatively sudden impacts on quality of life or significantly narrowed the range of choices available to a government or its people. Clearly, most environmental issues would not be security issues under this definition as they occur relatively slowly and do not, by themselves, significantly restrict people's ability to respond to them. However, some regional issues, such as water in the Middle East and the drying of the Aral Sea in central Asia (which has lost 70 percent of its volume since 1960), are significant enough to be considered a threat to national security, and an underlying cause in regional conflict. The assessment of threat is highly contextual and it is difficult to establish any direct cause and effect relationships.

The North Atlantic Treaty, under which the NATO alliance operates, was one of the first to recognize that security is not only based on military strength but also depends on economic and social security. This dimension has been increasingly important within NATO since the end of the Cold War, when environmental aspects of security were also integrated into NATO strategy (Lietzmann and Vest 1999).

1.2 Human Security

Human security was first defined as threats to physical security; in particular, freedom from fear and freedom from want. The Universal Declaration of Human Rights adopted by the United Nations in 1948 states that 'everyone has the right to life, liberty and the security of person.' By 1994, the concept of human security had evolved to encompass seven aspects of security: economic security, food security, health security, environmental security, personal security, community security and political security (UNDP 1994). At the same time, a distinction was made between human security and human development. 'Human development is a broader concept, defined as a process of widening the range of people's choices. Human security means that people can exercise these choices safely and freely' (UNDP 1994, p. 23).

Environmental security is highlighted as an individual component of human security and as a contributing factor to other aspects of human security, such as food, employment and health, because these depend on people's access to natural resources, like land, water and forests. Very often, these environmental security issues are played out at a regional level between national sharing borders and environmental resources.

1.3 Critical Zones and Vulnerabilities

A related concept is that of environmental criticality, when human activity has so radically altered the natural environment that it no longer can support human populations or their economic livelihoods, and it is beyond restoration by normal ecosystem processes or human efforts at remediation (Kasperson et al. 1996). This same idea lies behind the 'red zone' maps of Russian geographers and the concept of 'hot spots' of biodiversity at risk (Myers 1993).

Whether an area becomes a critical zone depends on the rate and extent of environmental degradation, but it appears that some areas are more vulnerable to reaching a critical state than are others. This has led to an interest in why certain combinations of geographic

area and human society seem to be more vulnerable than others to environmental risk, and what the thresholds and possible indicators of vulnerability and adaptation might be (Lonergan 1997).

2. Global Change and Environmental Security

More recently, the relationship between global environmental change and environmental security has been examined. Global change is manifested in many environmental effects. These stem from the warming and other changes in the atmosphere brought about through the increase of levels of carbon dioxide and other gases, and of contaminants produced through human economic activity. These environmental effects extend to changes in ocean currents, precipitation patterns on land and sea, a rising sea level, and stratospheric ozone depletion. In turn, global change impacts many aspects of ecosystem functioning and human economic activity such as loss of biodiversity, global biogeochemical cycles, agricultural and forestry production. Some of these changes are gradual and long term, while other changes relate to the frequency and severity of peak events, like hurricanes, droughts and other natural hazards. It is argued that both types of changes can significantly affect people's ability to cope with environmental risks and thus are likely to affect environmental security and conflict.

Global change has also led to a series of global environmental negotiations and agreements, as well as bargaining about trading rights to pollute. While these efforts at environmental diplomacy might seem to reduce the potential for environment insecurity and conflict, it is clear that in the short run, there are potential winners and losers among nation states as well as possible 'free-riders.' Thus the global environmental negotiations can exacerbate the potential for environmental disputes. These are revealed as ideological differences between industrialized and nonindustrialized (or developing) countries, whose interests are usually clearly different, but more importantly, they tend to emphasize regional disparities and potential for conflict.

Formal international scientific programs which seek to bridge the gap between science and policy (or research and action) are beginning to examine risk, vulnerability and security. For example, the Global Environmental Change and Human Security Project (GECHS) of the International Human Dimensions Programme on Global Environmental Change (IHDP) is studying the relationship between global change and environmental security (GECHS 1999). It is addressing some key questions: what types of environmental change threaten security? Which regions and groups of people are the most insecure and why? Can we predict future insecurities? What strategies are available to governments and communities to cope with environmental insecurity caused by global change? Another international science program, LUCC (Land-Use and Land-Cover Change) is focusing on critical regions and vulnerable places, especially what happens to the resiliency of people, agricultural production systems and biodiversity in the face of global changes like sea-level rise and increased frequency of extreme storm events (LUCC 1999)

3. Critical Zones

It has been argued that some environments, such as rainforest and tundra, are more sensitive to human exploitation and that some forms of human development, such as rapidly growing populations or 'frontier' economies (where exploitation has room to move and can therefore be more destructive) exacerbate environmental vulnerabilities. Amazonia is a tropical rainforest which has been damaged by frontier development and rapidly increasing in-migration.

Kasperson et al. (1996) examined nine regions of the world which were considered potentially critical zones according to three criteria: (a) the environmental changes must threaten the basic structure and function of the ecosystem; (b) the changes must endanger human well-being or sustained human use of the area; and (c) the threat must be great enough to overwhelm the capacity of the local people or government to deal with the degradation. They concluded that only the Aral Sea is presently a critical zone because the lake is essentially dead and surrounded by evaporated salts and highly contaminated soils. However, five of the areas studied are highly endangered and are in a pre-critical condition. These are: Eastern Sundaland (which includes the tropical forests of Borneo and the eastern Malay Peninsula); the semi-arid Ukambani region of southeastern Kenya; the deforested and highly eroded Nepal middle mountains; the arid plains in western Texas and New Mexico; and the Basin of Mexico in which Mexico City stands.

A different approach to critical zones is that being developed in the GECHS (Global Environmental Change and Human Security) project which is to identify indicators of environmental degradation and human insecurity to develop early warning of 'hot spots' for potential conflict (GECHS 1999). Data quality and reliability are major problems in this approach but some initial maps have been prepared based on an Index of Human Insecurity. The indicators include water scarcity, food import dependency ratio, energy imports as a percentage of consumption, access to safe water, expenditures on defense vs. health and education, indicator of human freedoms, urban population growth, child mortality, maternal mortality, income per capita, degree of democratization, and fertility rates (Lonergan et al. 1999). The results

show that the areas with the highest insecurity are most of Africa, the Middle East, and south and southeast Asia. Within these areas are so-called pivotal states which are balanced precariously between security and insecurity: between development and political and social chaos; and where what happens in that country can have far-reaching, even global effects. Algeria, Brazil, Egypt, Mexico, India, Indonesia, Pakistan, and South Africa have been identified as pivotal states.

4. *Environmental Security and Conflict*

A number of major research projects have focussed on the nature of the linkages between environmental security and conflict including the Peace and Conflicts Studies Programme at the University of Toronto (Homer-Dixon 1994, Homer-Dixon and Percival 1996); the Environment and Conflicts Project (ENCOP) in Switzerland (Bachler and Spillman 1996) and the International Peace Research Institute in Oslo (Gleditsch 1998, Hauge and Ellingsen 1998).

The relationship between environmental stress and conflict is a complex one and may not necessarily result in violent conflict; it can also lead to negotiation and peaceful resolution. Environmental stress almost always interacts with other social, economic and political factors before leading to conflict, and any resulting conflict can lead to further environmental stress. Usually environmental stress produces human insecurity problems like food insecurity, population movements and the breakdown of social and political institutions, and it is these secondary effects which can contribute directly to conflict. Thus, environmental stress is seen as sometimes a source of conflict, a catalyst or a trigger for conflict, depending on the circumstances. The social, economic and political context has always to be considered when assessing the potential effects of environmental insecurity.

The University of Toronto study identified three conditions of resource scarcity which could lead to conflict: (a) decreased quality and quantity of renewable resources (supply induced scarcity); (b) increased population growth or per capita consumption (demand-induced scarcity); and (c) unequal access to resources (structural scarcity). These types of scarcity can act singly but often combine to produce two common phenomena: resource capture and ecological marginalization. Resource capture occurs when a decrease in resources quantity or quality coincides with population increase, and the most powerful groups in society ensure that the resources are redistributed in their favor, leading to severe environmental scarcity for the poorer and weaker. Ecological marginalization occurs when population increase and inequitable access to natural resources combine to cause population migration, usually to more marginal and less inhabited environments like steeper mountain slopes and more-arid areas. The resulting higher population densities and lack of local environmental knowledge can lead to severe environmental damage by the migrants (Homer-Dixon 1994).

The ENCOP (Environment and Conflicts) Project sponsored by the Swiss Government and the Swiss Peace Foundation considers all conflicts to be social and political events even when environmentally induced. It found that the main contextual factors for environmental insecurity are: perceptions, especially when one group sees their interests as being threatened; economic vulnerability and resource dependency; institutional and technological capacity; cultural and ethno-political factors, especially where there are differences in wealth and power; the existence of democratic institutions and political stability; participation in decision making; and what mechanisms of conflict resolution are available. Based on a series of case studies, the ENCOP project found that environmentally induced conflicts were of three main types: (a) between different groups within a country; (b) through processes such as population displacement which leads to conflicts crossing national borders; and (c) when internationally shared resources like rivers, regional environments or global commons become degraded and lead to inter-state conflict.

5. *Environmental Migrants*

Environmental degradation has been cited as one of four root causes of refugee flows, together with political instability economic tensions and ethnic conflict (UNHCR 1993). The number of environmental refugees is estimated to be 10 million out of the 17 million official refugees and more than half of these are believed to be in sub-Saharan Africa (Westing 1992). But the numbers may be far greater and appear to be rapidly increasing. The greatest effects of climate change may be to displace millions of people through shoreline erosion, coastal flooding and agricultural failures.

Although the World Commission on Environment and Development identified environmentally induced population movements as a recent phenomenon (WCED 1987), there is much evidence that people have had to move from land affected by natural disasters, warfare and mismanagement throughout—and before—history. Despite the size of the problem, there has been relatively little empirical research on the role of environmental change in population movements. Some of this movement is across national borders, but it is estimated that for every individual who moves across an international border, there are two or three persons who have to move within their country. These are internally displaced persons (Myers 1995).

There are a number of difficulties in understanding the role of the environment in causing people to move.

Migration research has long shown that the decision to migrate is a combination of 'push–pull' factors, as people weigh the advantages and disadvantages of where they are presently located compared to where they plan to move. It is difficult to isolate the role of environment in this decision unless there is a specific environmental stress such as an earthquake or flood from which people are fleeing and after which they are likely to return. Other environmental 'push' factors include slower environmental degradation such as deforestation; soil deterioration through erosion, salinization or waterlogging; and desertification and falling water tables, which make agriculture less viable. Development projects such as dams and irrigation schemes have often involved enforced resettlement. In India, it is estimated that over 20 million people have been displaced by development projects since the 1960s (Fornos 1992).

6. *Critical Perspectives*

Although work on environmental security has gained momentum in the 1990s in both academic and policy circles, there are strong critics from both the security and environmental sides of the debate. Security critics point out that broadening the definition of security to include aspects like environment, food security, and economics, renders it useless as an analytical tool as it means different things to different people. Furthermore, it encourages the military to undertake 'environmental' missions which takes time and resources away from their real task of military preparedness. Environmental critics claim that the research in environmental security is methodologically weak and cannot find causal links between environment and security. They also see that security institutions are looking for new missions ('greening of security') to maintain their funding after the end of the Cold War. Developing country scientists see the concept of environmental security as a northern industrialized paradigm. There is also concern that the concept of environmental security encourages a military 'us versus them' approach to problem solving, where complex relationships between environmental change and human development require less simplistic and more collaborative solutions (Lipschutz 1992, Gleditsch 1997, GECHS 1999). Despite these criticisms, environmental security is rapidly developing as a focus at the nexus of research and policy.

See also: Environment and Development; Food Security; Land Reclamation; Retail Trade

Bibliography

Bachler G, Spillman K R 1996 *Environmental Degradation as a Cause of War*. Ruegger Verlag, Zurich, Vols. 2–3

Fornos W 1992 *Desperate Departures: The Flight of Environmental Refugees*. The Population Institute, Washington DC

GECHS 1999 *Global Environmental Change and Human Security (GECHS) Science Plan*. International Human Dimensions Programme on Global Environmental Change, Bonn, Germany

Gleditsch N P 1998 Armed conflict and the environment: A critique of the literature. *Journal of Peace Research* **35**: 381–400

Hauge W, Ellingsen T 1998 Beyond environmental scarcity: causal pathways to conflict. *Journal of Peace Research* **35**: 299–317

Homer-Dixon T F 1994 Enviromental scarcities and violent conflict. Evidence from cases. *International Security* **19**: 5–40

Homer-Dixon T F, Percival V 1996 *Environmental Scarcity and Violent Conflict: Briefing Book*. AAAS and University of Toronto, Toronto, Ontario, Canada

Kasperson J X, Kasperson R E, Turner B L 1996 Exploring environmental criticality. *Environment* **38**: 4

Lietzmann K M, Vest G D 1999 Environment and security in an international context. *Environmental Change and Security Project Report 5*. Woodrow Wilson Center, Washington, DC

Lipschutz R D (ed.) 1995 *On security*. Columbia University Press, New York

Lonergan S 1997 The role of environmental degradation in population displacement. *Global Environmental Change and Human Security Project Research Report 1*

Lonergan S C, Gustavson K, Harrower M 1999 Mapping Human Insecurity. In: Lonergan S C (ed.) *Environmental Change, Adaptation and Security*. Kluwer Academic, Dordrecht, The Netherlands

LUCC 1999 *Land-Use and Land-Cover Change Implementation Strategy*. International Geosphere-Biosphere Program, Stockholm

Myers N 1993 *Ultimate Security: The Environmental Basis of Political Stability*. W. W. Norton, New York

Myers N 1995 *Environmental Exodus: An Emergent Crisis in the Global Arena*. Climate Institute, Washington DC

Ullman R H 1983 Redefining security. *International Security* **8**: 129–53

United Nations High Commission for Refugees 1993 *The State of the World's Refugees: The Challenge of Protection*. Penguin Books, New York

UNDP 1994 *Human Development Report 1994*. Oxford University Press, New York

Westing A H 1992 Environmental refugees: a growing category of displaced persons. *Environmental Conservation* **19**: 201–7

World Commission on Environment and Development (WCED) 1987 *Our Common Future* Oxford University Press, Oxford, UK

A. V. Whyte

Environmental Stress and Health

1. *Introduction*

This article examines how the concept of stress has been used to specify environmental characteristics that may lead to disturbances of intended behavior, psy-

chological and physiological discomfort, and health aspects of residents living in the vicinity of noisy or toxical sources. The physical environments considered here have enduring characteristics and are not directly under the control of the residents. This fact constitutes one of the main characteristics of psychological stress: people may consider the environmental load to be harmful, but they usually cannot stop the emissions of the source directly. In addition, they often learn from authorities that the emissions are harmless, and yet they have to rely on authorities in order to reduce the environmental load. The focus of this article is on dose-response relationships between environmental characteristics and psychological reactions of the residents. The observed reaction-variance between residents is due partly to measurement errors both on the physical and the response side, and due partly to factors intervening between the physical characteristics and the responses. Nevertheless, in some cases there are clear dose-response relationships that can be used in order to predict the average main effects at given physical loads.

2. *Definitions of Environmental Stress*

Although there are many different concepts of stress in the respective fields of medicine, psychology, and sociology, it is generally understood that stress is aversive in some sense. Psychologists often favor the stress concept proposed by Lazarus (1966), and Lazarus and Launier (1978), which defines stress as a transaction between stimulus and response, i.e., the process that reflects the individual appraisal of a stimulus by persons who see an imbalance between environmental demands and their own response capabilities to cope adequately with the demands. It should be noted that this concept relies heavily on individual self-reports, but sometimes even the individuals concerned do not notice the dangers that a certain situation provides for health, e.g., the potential hearing loss implied by loud music.

2.1 *Characteristics of Environmental Stressors*

Environmental stressors are usually considered to fall into one of four distinct classes: cataclysmic events, stressful life events, daily hassles, and ambient stressors (Evans and Cohen 1987). Cataclysmic events comprise sudden catastrophes that affect many individuals at the same time. For instance, floods, major storms, earthquakes, volcanic eruptions, nuclear power plant accidents, chemical plant accidents, and the discovery of toxic waste dumps belong to this category. Cataclysmic events are seldom predictable, neither with respect to their beginning, nor with respect to their course, but they are usually expected to be rather short.

Stressful life events are major incidents in the course of life that require major individual adaptive responses. Such events include major changes in work, or residential environment, e.g., beginning a new job, moving to a new residential area, major construction work in the present residential area, or a major perceptible change in the operating conditions of a nearby stressor. The event as such is usually short, but the behavioral consequences may be long, or permanent.

Daily hassles are repeatedly occurring aversive events of ordinary life, such as arguments with colleagues, crowded classrooms, and traffic congestions on the daily route to work. Although they typically are more or less predictable, the individual has little means to avoid such hassles, and the duration is rather short.

Ambient stressors is a term proposed by Campbell (1983), denoting more continuous and intractable background characteristics of the physical environment. They often go unnoticed, like the continuous hum of the air conditioning, the permanent dust in an industrial area, and the faint hiss of the central heating system. Most people believe that they can adapt to ambient stressors, and they consider the costs of coping with such stressors to be higher than simply enduring them.

In addition, each type of environmental stressor can be categorized along several dimensions, of which we mention only three.

(a) The major dimension is the degree of controllability over the environmental stressor. Control may be either direct (e.g., switching off the neighbor's noisy lawn mover), indirect (e.g., closing the windows in order to reduce the noise penetration), social (e.g., calling authorities to stop the noise), or cognitive (e.g., knowing that the noise will stop in about an hour). It has often turned out that stress effects decrease with increased control, but sometimes control has secondary stress effects (e.g., closing the windows will increase the temperature inside the house).

(b) Related to cognitive control is the predictability of the stressor. High predictability of the occurrence and time course of an aversive agent may help people to cope with it. For instance, residents in the vicinity of city rail systems usually know exactly the time schedule of the trains and seldom complain about noise and vibration, but they feel irritated when a train is missing. Alternatively, knowing beforehand that a new airport will open in the neighborhood does cause greater stress effects than knowing that the old airport will close in the neighborhood (Hatfield et al. 1998).

(c) Stressors may be more or less perceptually salient or identifiable. Some environmental agents are of rather low intensity and may either be not detected at all (e.g., low radioactive radiation), others are only detected when intensity changes (e.g., the faint current of air conditioners). If an environmental agent is not perceptually salient, stress reactions may occur be-

cause people get informed about potential hazards, and the type and degree of stress reaction depends very much on the content of information and the credibility of its source.

2.2 Variables Intervening Between Stressors and Their Effects

Although it is true that varying degrees of environmental load cause varying degrees of stress reactions in the average, it is also well known that the same degree of environmental load does not cause the same degree of reaction in every person or in every situation. It depends on situational, personal, and social variables whether an environmental agent becomes a severe problem for a person or not. For instance, a faint burning smell or a sudden noise may cause greater alarm reactions at night than the same events occurring during daytime—this is an example of situational moderation. Individual differences in susceptibility are well known with respect to noise and smoke, i.e., persons who judge themselves to be rather susceptible to noise generally show greater noise reactions than persons who judge themselves less susceptible. A similar relation holds for smoke, and recent evidence (Winneke and Neuf 1992) fosters the belief that environmental susceptibility is a general individual trait, comprising susceptibilities for different environmental agents.

While individual differences are thought to be the result of the individual development of a person, social intervening factors are thought to be the result of social developments, shared by larger groups of the society. For instance, a plant that emits fumes and noise may cause less stress effects when a group evaluates the source positively in some other respect (e.g., as providing jobs). Of course, there is no clear-cut distinction between personal and social factors, because the individual development usually takes place within a certain society, and there is also a considerable individual variation with respect to sharing the views and evaluations of society in general. For an extensive discussion of moderating variables for noise, see Guski (1999).

3. Effects of Environmental Stressors

3.1 Effects of Noise

Noise is defined generally as unwanted sound, i.e., a sound that is already evaluated by somebody or some group as being aversive. Since the degree of aversiveness usually increases with the level of sound, the scientific study of noise effects is restricted to the middle and upper levels of environmental sound. Sound levels are usually given in decibels (e.g., dB),

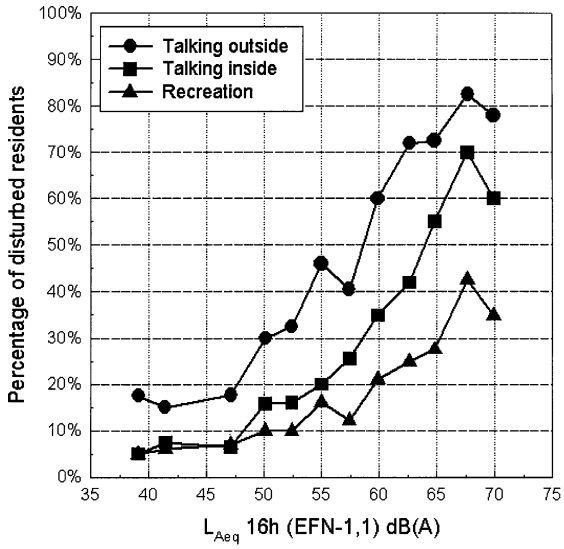

Figure 1
Percentage of residents disturbed by aircraft noise at Oslo Airport (after Gjestland et al. 1990)

but care should be given to the exact meaning of the term used for describing the sound level. For instance, a temporary maximum value of $L_{max} = 65$ dB is easily reached by every speaker at 1m distance—it may be a nuisance, but will cause no harm—while the sound energy of $L_{Aeq} = 65$ dB calculated for 16 hours of daytime is considered to be a risk for health. The main source of environmental noise-effect studies is transportation noise (road, rail, and air), because it affects almost everybody in the industrialized world. Lambert and Vallet (1994) estimated 20 percent of citizens of the European Community to be exposed to transportation noise with daytime levels of $L_{Aeq} = 65$ dB or higher.

The studies considered in this article are interdisciplinary field studies, using residents in the vicinity of traffic routes as subjects. The noise level is usually given as daytime and/or night-time sound energy, calculated from data about traffic volumes, together with a few supporting measurements at the outside of the subjects' lodgings. The questionnaires usually cover noise-effect questions (e.g., with respect to annoyance, disturbance of communication, disturbance of recreation, and sleep quality) in a closed response format, general questions about living conditions, demographic characteristics, and potential intervening variables, such as the individual noise sensitivity, health status, coping capacity, and trust/distrust in authorities of the sound source. Daytime environmental traffic noise has large statistical effects on the degree of annoyance, and the degree of reported disturbances of daily activities. Nighttime environmental traffic noise has somewhat smaller statistical

effects on the degree of general annoyance and reported sleep quality. A very typical result is the linear relationship between physical parameters of sound and the mean response of residents, at least as far as raw scores are used. Sometimes, the responses are given as percent 'highly annoyed' (percent HA), selecting only those residents who chose the upper 27 percent of the original scale. Because of the selection process, the dose–response curve is nonlinear in that case. A typical dose–response relationship is given in Fig. 1, showing the nonlinear relationship between the level of daytime aircraft noise in the residential area, and the average percentage of residents who say that they are disturbed while engaged in different activities during the day.

In comparing the effects of different sources of transportation noise, it is generally found that the mean general annoyance response is somewhat higher for aircraft noise than for road traffic and railroad noise at the same overall noise levels, if the noise levels are converted to day/night levels (L_{dn}), where night levels are weighted 10 dB stronger than day levels (Miedema and Vos 1998). The same authors (1999) conclude from a comparison of original data from several studies that personal attitudes to noise (e.g., fear or other negative evaluations of the source, and noise sensitivity) are very effective variables intervening between the physical impact and the annoyance response to noise.

With respect to somatic effects of noise, data are less clear, but there are several studies showing the prevalence of ischemic heart diseases to increase somewhat in middle-aged males with high levels of daytime road traffic noise (Babisch et al. 1999). Under nighttime transportation noise a fragmentation of the sleep structure is to be observed and can be classified as stress (Griefahn 1986, Griefahn and Muzet 1978). The effects can be observed by means of individual EEG recordings during the night, as well as by means of increased fallout of the stress hormones adrenaline, noradrenaline, and cortisol. Maschke et al. (1998) claim that an increased secretion of adrenaline and noradrenaline belongs to a defense situation, an increased secretion of cortisol marks a defeat situation. The results of an experimental longitudinal field study at Hamburg Airport show no uniform cortisol excretion reaction to persistent aircraft noise, but about half of 16 subjects display an increased cortisol excretion after a phase of adaptation during the disturbed night, and the increased cortisol levels are classified as exceeding the normal range.

3.2 Living on Polluted Soil

Residential areas are sometimes built in the vicinity of former industrial areas, and this can cause severe problems for health, if the soil is polluted. One of the best known examples is Love Canal (see Bell et al. 1995), where toxic agents had been disposed of over years into a creek, and several years after the closure of the chemical plant, residential buildings were erected in the area. When toxic evaporations were detected, health authorities alarmed the residents and offered financial help for those who wanted to move. The monetary compensations were judged to be insufficient by many residents, and a long-lasting public debate reflected the severe helplessness of many residents (Levine 1982).

Another example is a residential area in Dortmund (Germany), where toxic evaporations from a former chemical plant were detected and caused severe psychosomatic complaints (Matthies et al. 2000). The debate about possible health hazards, in combination with insensitive communication policy of the health authorities, led to ongoing psychosomatic effects even after the beginning of a complete soil exchange in the whole residential area. The amount and degree of complaints was not related to any chemical variable measured, but differed significantly from a control group in an unpolluted area.

3.3 Living With Nuclear Power Plants

Although the nuclear industry spends billions of dollars for public relations in the hope of changing public attitudes to nuclear power plants and nuclear waste, the general attitudes in Western societies seem rather skeptic. The skeptic attitudes are caused by the public experience of technological catastrophes like Three Mile Island (TMI) (1979) and Chernobyl (1986). The psychological effects of technological catastrophes on residents differ from those of natural hazards in many ways, e.g., they are more long lasting, and residents lose faith in public communications of authorities with respect to safety and control of health risks. Even seven years after the catastrophe at TMI, residents living near the damaged reactor reported more chronic stress symptoms and sleep disturbances than did control subjects (Davidson et al. 1987).

The catastrophe at Chernobyl caused severe psychological changes in the European population, especially with respect to the evaluation of nuclear power, the probability of a nuclear accident, and the health risks for children (Drottz-Sjöberg and Sjöberg 1990, Hüppe and Janke 1994). Many social scientists believe that the inconsistent and reluctant communication policy of public authorities has contributed to the stress symptoms of residents who were only slightly exposed to direct nuclear radiation.

4. Conclusions

Environmental characteristics of residential areas (such as noise, polluted soil, and nuclear power plants) may lead to disturbances of intended behavior, psy-

chological and physiological discomfort, and health aspects of residents. The relation between physical characteristics of the source and the aversive reactions can be described as a stress situation, because the reactions covary with the degree of emission, and residents usually cannot stop the emissions of the source directly. In addition, situational, personal, and social variables moderate the dose-response relationship.

See also: Crowding and Other Environmental Stressors; Environmental Challenges in Organizations; Environmental Risk and Hazards; Job Stress, Coping with; Post-traumatic Stress Disorder; Stress and Coping Theories; Stress and Health Research; Stress at Work

Bibliography

Babisch W, Ising H, Gallacher J E J, Sweetnam P M, Elwood P C 1999 Traffic noise and cardiovascular risk: The Caerphilly and Speedwell studies, third phase: 10-year follow up. *Archives of Environmental Health* **54**: 210–6

Bell P A, Greene T C, Fisher J D, Baum A 1996 *Environmental Psychology*, 4th edn. Harcourt Brace, Fort Worth, TX

Campbell J M 1983 Ambient stressors. *Environment* & *Behavior* **15**: 355–80

Davidson L M, Fleming R, Baum A 1987 Chronic stress, catecholamines, and sleep disturbance at Three Mile Island. *Journal of Human Stress* **13**: 75–83

Drottz-Sjöberg B-M, Sjöberg L 1990 Risk perception and worries after the Chernobyl accident. *Journal of Environmental Psychology* **10**: 135–49

Evans G W, Cohen S 1987 Environmental stress. In: Stokols D, Altman I (eds.) *Handbook of Environmental Psychology*. Wiley, New York , Vol. 1 pp. 571–610

Gjestland T, Liasjö K, Granöien I, Fields J M 1990 *Response to Noise Around Oslo Airport Fornebu*. Trondheim: Elab-Runit Sintef Gruppen. Acoustics Research Center. Report STF40 A90189

Griefahn B 1986 Noise pollution during the night—a possible risk factor for health. In: Sáenz A L, Stephens R W B (eds.) *Noise Pollution. Effects and Control. Scope 24*. Wiley, Chichester, UK pp. 265–82

Griefahn B, Muzet A 1978 Noise induced sleep disturbances and their effects on health. *Journal of Sound and Vibration* **59**: 99–106

Guski R 1999 Personal and social variables as co-determinants of noise annoyance. *Noise* & *Health* **3**: 45–56

Hatfield J, Job R F S, Carter N L, Peploe R, Taylor R, Morell S 1998 Attitude-mediated reaction to noise influences physiological responses: Evidence supporting causality. *Inter-Noise* '98, Paper No. 440

Hüppe M, Janke W 1994 The nuclear plant accident in Chernobyl experienced by men and women of different ages: Empirical study in the years 1986–1991. *Anxiety, Stress, and Coping* **7**: 339–55

Lambert J, Vallet M 1994 *Study Related to the Preparation of a Communication on a Future EC Noise Policy*. Bruxelles: LEN Report No. 9420

Lazarus R S 1966 *Psychological Stress and the Coping Process*. McGraw-Hill, New York

Lazarus R S, Launier R 1978 Stress-related transactions between person and environment. In: Pervin L, Lewis M (eds.) *Perspectives in Interactional Psychology*. Plenum, New York pp. 1–67

Levine A G 1982 *Love Canal: Science, Politics, and People*. Lexington Books, Lexington, MA

Maschke C, Harder J, Hecht K, Balzer H U 1998 Nocturnal aircraft noise and adaptation. *Noise Effects '98, 7th International Congress on Noise as a Public Health Problem* **2**: 433–8

Matthies E, Höger R, Guski R 2000 Living on polluted soil. Determinants of stress symptoms. *Environment* & *Behavior* **32**: 270–86

Miedema H M E, Vos H 1998 Exposure-response relationships for transportation noise. *Journal of the Acoustical Society of America* **104**: 3432–45

Miedema H M E, Vos H 1999 Demographic and attitudinal factors that modify annoyance from transportation noise. *Journal of the Acoustical Society of America* **105**: 3336–44

Winneke G, Neuf M 1992 Psychological response to sensory stimulation by environmental stressors: Trait or state? *Applied Psycholology: An International Review* **41**: 257–67

R. Guski

Environmental Surprise

Decision makers at all scales (individuals, firms, and local, national, and international governmental organizations) are concerned about reducing their vulnerability to (or the likelihood of) unexpected events or 'surprises.' After briefly and selectively reviewing the literature on uncertainty and surprise, a definition of 'surprise' is introduced that does not include the strict requirement that it apply to a wholly unexpected outcome, but rather recognizes that many events are often anticipated by some, even if not most observers. Thus, an 'imaginable surprise' is defined as an event or process that departs from the expectations of some definable community, yet is a concept related to, but distinct from risk and uncertainty. Therefore, what gets labeled as 'surprise' depends on the extent to which what happens departs from community expectations and on the salience of the problem (for elaboration see Schneider et al. 1998). Impediments to overcoming ignorance range from the need for more 'normal science' to phenomenological impediments (e.g., inherent unpredictability in some chaotic systems) to epistemological ignorance (e.g., ideological blocks to reducing ignorance). The substantive focus will concentrate on the theme of global change. Examples of 'imaginable surprises' in the global change context are presented. Improving the anticipation of surprises is an interdisciplinary enterprise

that should offer a skeptical welcoming of outlier ideas and methods, tempered by broadbased assessments that attempt to estimate the subjective probabilities of all imaginable outcomes (e.g., Schneider 2001).

Strictly speaking, a surprise cannot be anticipated; by definition it is an unexpected event. Potential climate change, and more broadly global environmental change, is replete with the truly unexpected because of the enormous complexities of processes and interrelationships involved and our insufficient understanding of them (such as coupling ocean, atmosphere, and terrestrial systems). However risk, hazard, and related research demonstrate repeatedly that the event, process, or outcome registered as a surprise by the community in question was frequently known or forecast by others or the same event was knowable within the competing frameworks of understanding (Darmstadter and Toman 1993).

Much of the current work on surprise has grown out of an extensive body of research on uncertainty. Yet, although widely acknowledged and studied, uncertainty remains a difficult concept to define or codify. Different conceptualizations and approaches to uncertainty abound in the literature, cross numerous fields of study and touch a wide range of problem types. Two basic options are invariably followed in the face of uncertainty. The first is to reduce the uncertainties through data collection, research, modeling, simulation techniques, and so forth. Following this option, the objective is to overcome uncertainty, to make the unknown known. But the daunting nature of uncertainties surrounding global environmental change (what some have labeled 'deep uncertainty'), as well as the need to make decisions before the 'normal' science option can provide resolution, force a second option—that of managing or integrating uncertainty directly into the decision-making or policy-making process. Before uncertainty can be so integrated, however, the nature and extent of the uncertainty must be clarified.

The fields of mathematics, statistics, and more recently physics, provide the 'science of uncertainty' with many powerful means and techniques to conceptualize, quantify, and manage uncertainty, ranging from the frequency distributions of probability theory, to the possibility and belief statements of Bayesian statistics. Addressing other aspects of uncertainty, fuzzy set logic offers an alternative to classical set theory for situations where the definitions of set membership are vague, ambiguous, or non-exclusive. More recently, researchers have proposed chaos theory and complexification theory to focus on expecting the unexpected in models and theory (e.g., Casti 1994).

The practical application of many of these techniques was originally pioneered by researchers in decision analysis (see Raiffa 1968). In the fields of economics and decision theory, researchers continue to study rational decision making under uncertainty and how to assess the value of collecting additional information (e.g., Clemen 1991). Methods for modeling risk attitudes, leading to the terms risk-prone and risk-averse, attempt to capture how different people faced with making a decision react to the uncertainty surrounding the expected outcomes. Uncertainty and its related context, surprise, are treated largely as the realization that events, currently unknown, will occur affecting the final outcome of a process or decision.

This acknowledgment of uncertainty has found a prominent place in many other fields of study, each one speaking its own language of uncertainty. Research on uncertainty cross-cuts a number of different disciplines. For example, researchers making risk assessments and setting safety standards find it most useful to distinguish between risk (the probability of a certain negative effect resulting from a hazard occurrence, given the specified level of exposure), variability (inter-individual differences in vulnerability and susceptibility), and uncertainty (model parameter variability and any unexplained residual). In work related to hazards and risk, sociologists, anthropologists, psychologists, and geographers have made important contributions to the discussions on risk perception, risk communication, and the social amplification of risk (Kahneman et al. 1982, Kasperson et al. 1988, see Gigerenzer 1996 for a criticism). Similarly, work on visualizing or graphically conveying uncertainty also crosses a diverse set of disciplines including psychology, computer science, and Geographic Information Systems (e.g., MacEachren 1992).

Wynne (1992) emphasizes that the modeling of environmental risk systems requires examination of not only the scientific evidence and competing interpretations, but also investigation of the nature, assumptions, and inherent limitations of the scientific knowledge behind the data and the model. He specifies four types of uncertainty—risk, uncertainty, ignorance, and indeterminacy—each overlaying dimensions of uncertainty. *Risk* refers to a situation when the system behavior is well known and the chances of different outcomes can be quantified by probability distributions. If, however, the important system parameters are known but not the associated probabilities, then in Wynne's definitions, *uncertainty* exists. *Ignorance* is that which is not known (or even that we are aware that we do not know it) and, for Wynne, is endemic because scientific knowledge must set the bounds of uncertainty in order to function. *Indeterminacy* captures the unbounded complexity of causal chains and open networks. *Uncertainty*, in part, stems not only from an incomplete understanding of determinate relationships, but also from the interaction of these relationships with contingent and unpredictable actors and processes. However, the extent to which situations are truly 'indeterminate', as opposed to simply containing a very broad distribution of subjective probability estimates, is not a straightforward classification, for often very ill understood phenomena can still be bounded to some extent by

existing knowledge, and thus not truly indeterminate. In the areas of environmental policy and resource management, policy makers struggle with the need to make decisions utilizing vague and ambiguous concepts (such as sustainability), with sparse and imprecise information, in decisions that have far-reaching, and often irreversible, impacts on both environment and society. Not surprisingly, efforts to incorporate uncertainty into the decision-making process quickly move to the forefront with the advent of decision-making paradigms, such as the precautionary principle, adaptive environmental management, the preventative paradigm or stewardship. Ravetz (1986) takes the concept of 'usable knowledge in the context of incomplete science' one step further by introducing the idea of usable ignorance. To Ravetz, acknowledging the 'ignorance factor' means becoming aware of the limits of our knowledge. Ravetz argues that ignorance cannot be overcome with any amount of sophisticated calculations. Rather, coping with ignorance demands a better articulation of the policy process and a greater awareness of how that process operates. He recognizes that one can only replace ignorance by gaining more knowledge, but stresses that by 'being aware of our ignorance we do not encounter disastrous pitfalls in our supposedly secure knowledge or supposedly effective technique' (p. 429).

The emphasis on managing uncertainty rather than mastering it can be traced to work on resilience in ecology (e.g., Holling 1986). Whereas resistance implies an ability to withstand change or impact within some measure of performance, resilience captures the ability to give with the forcing function, without disrupting the overall health of the system. In this framework, adaptation is an ecological mechanism whose aim is not to overcome or control environmental uncertainty, but to live with, and in some case, thrive upon, it.

To recap, risk typically is defined as the condition in which the event, process, or outcome, and the probability that each will occur, is known. In reality, of course, complete or perfect knowledge of complex systems, which would permit the credible calculation of objective or frequentist probabilities rarely exists. Likewise, the full range of potential outcomes is usually not known. Thus, risk almost always is accompanied by varying degrees of uncertainty. Uncertainty is usually defined as the condition in which the event, process or outcome is known (factually or hypothetically), but the probabilities that it will occur are not known, or are highly subjective estimates (see, e.g., Moss and Schneider 2000).

Strictly speaking, surprise is the condition in which the event, process or outcome is not known or expected. In this 'strict' meaning, the attribution of surprise shifts toward the event, process or outcome itself. We may expect surprises to occur, but we are surprised by the specific event, process, or outcome involved. This meaning, as noted, begs the issue of anticipation, because the very act of anticipation implies some level of knowledge or foresight. However, it may be possible to identify 'imaginable conditions for surprise' where the conditions that might induce surprises are known even though the actual surprise events are not, for example, rapid forcing of non-linear systems (e.g., Moss and Schneider 2000).

Because of the impracticality of the strict definition of surprise for policy making, various studies advocate the use of another meaning for surprise, one in which the attribution of 'surprise' shifts more toward the expectations of the observer. Holling (1986, p. 294) recognized this meaning of surprise as a condition in which perceived reality departs qualitatively from expectations. It is this more interpretive or relational meaning of surprise—which has been labeled 'imaginable surprise'—that portends to be most useful for global change studies (Schneider et al. 1998).

Almost every event may constitute an imaginable surprise to someone. But since global change phenomena and their environmental and societal impacts are a community-scale set of issues, little can be gained for our purposes by focusing on whether someone, somewhere, may or may not have once predicted or hinted at some surprise event. More fruitful is the recognition that groups, communities, and cultures may share expectations such that a particular event is likely to qualify as a surprise for most within them. In these cases, what gets labeled as a surprise depends upon the extent to which reality departs from community expectations, and on the salience of the problems imposed.

Imaginable surprise applies to communities of experts, policy makers, managers, and educators who share common ranges of expectation that are generated by group dynamics, leaders, and signal processors, including the dominant educational and research paradigms (Kasperson et al. 1988). For these communities, shared expectations follow from dominant interpretations among the expert community (e.g., global warming is likely), from their fit with broader policy agendas (e.g., environmentally benign economic development is possible), and from vested interest, conscious or unconscious, of an agency or group to maintain a particular view (e.g., global population growth is environmentally damaging, or, alternatively, good for the economy). Since policy making often reflects a blend of public and interest group perceptions of reality, the imaginable surprise formulation is much more relevant to global change policy issues than a strict definition of surprise as an unimaginable outcome.

Since natural and social global change science remains in a range of developmental stages, the unknowns are sufficiently large to warrant attention to divergent themes about similar processes and outcomes. To facilitate this range of research, (a) measures should be taken to ensure a more open

discourse and evaluation of alternatives, such as by a more open airing and professional evaluation as opposed to uncritical, 'equal time' and equal credibility often afforded to polarized viewpoints in the popular media of less dominant or unconventional views, including those by advocacy science and scientist; and (b) by reducing the redundancy of research focused on the dominant views and theses while still preserving a diversity of approaches within dominant paradigms, that is, create research 'overlap-without-cloning.'

The assumptions associated with the standard paradigm of global climate change impact assessment, for example, although recognizing the wide range of uncertainty, are essentially surprise free. One approach is to postulate low, or uncertain, probability cases in which little climate change, on the one hand, or catastrophic surprises, on the other hand, might occur and multiply the lower probability times the much larger potential costs or benefits. Analysts, however, customarily use a few standard general circulation model CO_2-doubling scenarios to 'bracket the uncertainty' rather than to postulate extremely serious or relatively negligible climatic change outcomes. A strategic approach, that is, one that considers a wide range of probabilities and outcomes, may be more appropriate for global climate change impact assessments given the high plausibility of surprises, even if we have but limited capacity to anticipate specific details right now (Moss and Schneider 2000).

An assumption in cost-benefit calculations within the standard assessment paradigm is that 'nature' is either constant or irrelevant. For example, ecological services such as pest control or waste recycling are assumed as constants, or of no economic value in most assessment calculations. Yet should climatic change occur in the middle to upper range of that typically projected, it is highly likely that communities of species will be disassembled, and the probability of significant alterations to existing patterns of pests and weeds seem virtually certain. Some argue that pests, should their patterns be altered, can simply be controlled by pesticides and herbicides. The side effects of many such controls are well-known. What is not considered in the standard paradigm is the consideration of a 'surprise' scenario such as a change in public consciousness regarding the value of nature that would reject pesticide or herbicide application as a 'tech-fix' response to global changes.

Finally, global change portends alterations to the basic processes that govern the state of the biosphere. Global change research, therefore, might do well to anticipate these alterations, an effort that will require more than the study of extant processes and conditions alone. Various modes of analysis and approaches appropriate for such explorations, but typically underutilized in the research community, should be encouraged. Among these are: (a) backcasting scenarios from posited future states and/or reconstructing past scenarios in alternative ways to identify events or processes that might happen (recognizing, of course, that diffusion processes usually are not reversible and diffusion-dominated systems cannot be meaningfully backcast); (b) increasing attention to and support for the study of 'outlier' outcomes, searching for the reasons they appear deviant and the lessons that might be drawn from them (e.g., Hassol et al. 1997); and (c) exploring the 'resilience' paradigm (e.g., precautionary principle) alongside the 'optimization' paradigm (e.g., aggregated cost-benefit analyses) to inform policy making and diagnose alternative outcomes and risk management strategies. Other means of improving the anticipation of surprise in global change science would emerge from convening additional expert groups and asking them for more exhaustive assessments of the issues. Balanced assessments will likely lead to recommendations that 'research as usual' be tempered with more alternative or even unusual research.

In summary, global change science and policy making will have to deal with uncertainty and surprise for the foreseeable future. Thus, more systematic analysis of surprise issues and more formal and consistent methods of incorporation of uncertainty into global change assessments will become increasingly necessary. Improvements in dealing with scientific surprise in climate change in particular and global change in general, therefore, require the research and funding communities to seek a better balance among traditional and experimental research alternatives (also Kates and Clark 1996, p. 31). This aim, in turn, requires strategies that will facilitate this balance, including the difficult problem of assessing 'quality' in an interdisciplinary context.

See also: Environmental Adaptation and Adjustments; Environmental Challenges in Organizations; Environmental Risk and Hazards; Environmental Vulnerability; Risk, Sociology and Politics of

Bibliography

Casti J L 1994 *Complexification: Explaining a Paradoxical World through the Science of Surprise*. Harper Collins, New York

Clemen R T 1991 *Making Hard Decisions: An Introduction to Decision Analysis*. PWS-Kent, Boston

Darmstadter J, Toman M A (eds.) 1993 Nonlinearities and surprises in climate change: An introduction and overview. In: *Assessing Surprise and Nonlinearities in Greenhouse Warming. Proceedings of an Interdisciplinary Workshop*. Resources for the Future, Washington, DC, pp. 1–10

Gigerenzer G 1996 On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky. *Psychological Review* **103**: 592–6

Hassol S J J, Katzenberger J (eds.) 1997 *Elements of Change 1996*. Aspen Global Change Institute, Aspen, CO, p. 268

Holling C S 1986 The resilience of terrestrial ecosystems: Local surprise and global change. In: Clark W C, Munn R E (eds.) *Sustainable Development of the Biosphere*. Cambridge Uni-

versity Press for the International Institute for Applied Systems Analysis. Cambridge, UK, pp. 292–317
Kahneman D P, Slovic P A, Tversky A 1982 *Judgment Under Uncertainty*. Cambridge University Press, New York
Kasperson R E O, Renn O, Slovic P H, Brown H, Emel J, Goble R, Kasperson J X, Ratick S J 1988 The social amplification of risk: A conceptual framework. *Risk Analysis* **8**: 177–87
Kates R W, Clark W C 1996 Environmental surprise: Expecting the unexpected? *Environment* **38**(6–11): 28–34
MacEachren A 1992 Visualizing uncertain information. *Cartographic Perspectives* **13**: 10–19
Moss R H, Schneider S H 2000 Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting. In: Pachauri R, Taniguchi T, Tanaka K (eds.) *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*. Intergovernmental Panel on Climate Change, Geneva, pp. 33–51. Available from the Global Industrial and Social Progress Research Institute at http://www.gispri.or.jp
Raiffa H 1968 *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Addison-Wesley, Reading, MA
Ravetz J R 1986 Usable knowledge, usable ignorance: Incomplete science with policy implications. In: Clark W C, Munn R E (eds.) *Sustainable Development of the Biosphere*. Cambridge University Press, New York, pp. 415–32
Schneider S H 2001 What is 'dangerous' climate change? *Nature* **41**: 17–19
Schneider S H, Turner B L, Morehouse Garriga H 1998 Imaginable surprise in global change science. *Journal of Risk Research* **1**(2): 165–85
Wynne B 1992 Uncertainty and environmental learning: Reconceiving science and policy in the preventive paradigm. *Global Environmental Change* **XX**: 111–27

S. H. Schneider

Environmental Vulnerability

Global changes are accelerating at an ever increasing pace with varying impacts. While some societies are becoming more globalized and equipped with modern technology, others live in poverty, they struggle to survive amidst the challenges of fluctuating daily provisions, inadequate shelter, and poor provision of essential services including health and water. Both societal groups may confront environmentally vulnerable situations. People living in industrialized cities may, for example, have to encounter technological hazards (e.g., toxic chemical accidents) and those living in less-developed locations may be at risk of poorly constructed homes and inadequate basic services provision such as water (White 1974, Kates et al. 1985).

A host of practitioners (nongovernment organizations, relief workers, disaster managers), scientists, academics, and policy makers are endeavoring to understand the processes that shape the welfare and daily conditions of people and how their lives are influenced by environmental change. One way of trying to understand what enables some to better manage changing events is through a careful probing and examination of 'environmental vulnerability.' In this article environmental vulnerability will be examined particularly from the perspective of human vulnerability. Environmental vulnerability will be defined followed by examples of how the concept has changed and been used over time.

1. *Defining Environmental Vulnerability*

There have been several attempts to define and capture what is meant by environmental vulnerability ranging from geophysical, engineering, and technological aspects of environmental vulnerability through to the social, political, economic, and political dimensions. In the 1970s and 1980s, the importance of integrating these dimensions as a means to understand how such factors shape people's lives and their exposure and vulnerability to changing environments was recognized. This thinking influenced the generation of theories and concepts that were subsequently adopted in the lexicons of development, disaster management, and in global change science. One key concept which reflects this holistic view is environmental vulnerability.

Environmental vulnerability is the product of a combination of circumstances related to a host of issues which include living with risks and hazards (sudden events such as floods, earthquakes, and slow-onset events such as droughts, e.g., see *Environmental Risk and Hazards*). Other factors such as health, poverty, and gender also exacerbate existing risks/hazards, thereby heightening vulnerability and weakening positive responses to such phenomena. Environmental vulnerability is thus described with reference to the physical and social environment or the milieu in which a group may find themselves (Fig. 1).

Having briefly described what is meant by environmental vulnerability, the focus now shifts to examine how it is used in a variety of contexts. In some fields, environmental vulnerability is associated with geophysical aspects of risk and 'hazard science.' In many cases, however, particularly those characterized by poverty, certain risks, hazards, and potential disasters are not only linked to natural events but also related to slow, continuous processes that are advanced by environmental and socioeconomic vulnerability. In these regions, 'traditional' hazard/risk analyses are often compounded by other mitigating factors such as lack of access to resources, lack of accessible information, reduced response to shocks because of poverty or other factors. Recognizing this complexity, a strong emphasis on the human dimensions of vulnerability has emerged which describes both *internal dimensions* (inability or ability to cope and respond) and *external dimensions* (such as ex-

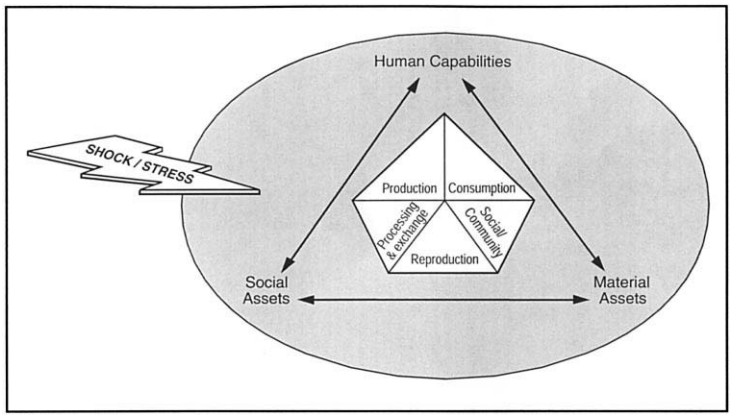

Figure 1
Configuration of vulnerability: an interaction between a household's capacity, flows of social and human assets, factors making up a 'living' in the face of a sudden/creeping shock or stress (adapted from Swift 1989, Chambers and Conway 1992, Mullins 1999)

posure to risk, shocks, and stress) (e.g., Chambers 1989, Swift 1989, Liverman 1990, Chambers and Conway 1992, Blaikie et al. 1994, Cutter 1996, Downing et al. 1996, Hewitt 1997). From this perspective, 'vulnerability' is not an absolute term but is a relational or social term. Vulnerable groups or regions are therefore those most exposed to a disturbance but who also possess a limited array of 'response mechanisms' or coping capabilities to respond to the event, crisis, or disturbance. Vulnerability also includes a 'sense of powerlessness' and has been described as a 'measure' of defenselessness and insecurity in the face of a crisis. Poor refugees, for example, dislocated from their families and family networks, lacking adequate shelter and forced to inhabit environmentally risky areas, would be most vulnerable to a sudden shock event such as an outbreak of cholera, a flood, or an earthquake.

These examples illustrate the complex nature of vulnerability and they also show how difficult it is to untangle the causes of vulnerability from a complex set of social and physical factors. They also illustrate the importance of rigorous analysis in unbundling the complex interplay between those components which shape and contribute to vulnerability (e.g., economic restructuring measures, violence and intimidation, health issues, loss of livelihoods, environmental degradation, and rapid technological development). For instance, several vulnerability assessments in poor regions examine a household's assets, 'bundles of goods,' capacities and abilities to respond to events, plus how these change or evolve over time.

Assessments of environmental vulnerability have also expanded to include the investigation of elements of both time and space. A group may, for example, become increasingly more vulnerable to an event as their coping responses become eroded over time. In several countries, HIV/AIDS, conflict, and violence are insidious, everyday factors that heighten vulnerability during times of shock or stress such as drought. Inter- and intraregional variations in environmental vulnerability also exist (see Kasperson et al. 1995). Some have therefore called for a multilayered approach when examining environmental vulnerability which involves various measurements to produce assessments of environmental vulnerability.

2. *Measuring Vulnerability in an Environmental Context*

Different organizations, scientists, and others have experimented and used an array of approaches and techniques to try and measure or target 'vulnerable groups.' Some of these include identifying indicators of vulnerability including vegetation change, flood and drought risk, food access and availability, and/or malnutrition. Methods and frameworks from which one can identify vulnerability have often been the product of a borrowing and a 'stitching' together of related approaches attempting to identify 'vulnerable environments' or 'regions of risk.' By using a variety of methods, maps of vulnerability highlighting those areas and groups that may be at social, physical, and or economic risk have been compiled (see Davies 1996, Kasperson et al. 1995). The dynamic nature of vulnerability, however, dictates that measurements of vulnerability have to be flexible and often include qualitative as well as quantitative techniques. The sustainable livelihoods approach has been used extensively in several parts of the world as a means to identify areas and groups that are vulnerable. In this approach the environment is seen to include both physical and social dimensions.

2.1 Sustainable Livelihoods Approach—Accessing a Measure of Vulnerability

A complex array of capabilities, resilience, and vulnerabilities are viewed collectively in the concept of sustainable livelihoods (see Chambers and Conway 1992, Carney 1998). This approach has become a very popular way of trying to gain a sense of the vulnerability and resilience of a community and its environment, particularly among those working in third world rural and urban contexts. The focus in such an approach is to try and determine how people live and make a living and from there try and capture those factors that may be undermining peoples' ability to produce a livelihood. In this approach people and their livelihoods are viewed together with their stores and resources (tangible assets e.g., land, water, trees and livestock, farm equipment, tools, etc.) and their claims and resources (intangible assets e.g., local networks such as relatives) within a specified environment.

In such ways livelihood analysis examines the resilience of an array of household strategies in the face of environmental change. This approach has been used in several country and household studies by, for example, organizations such as Care International, USAID, Oxfam, Save the Children Fund (UK), and others. A central element of such investigations has been to examine the interplay between the environment (e.g., such factors as land use management) and households at risk and explores how environmental factors complement or detract from flows of goods, services, and assets into a household, making such households more vulnerable or resilient to change. Environmental degradation, associated with deforestation, desertification, poor water, and land use management may, for example, prevent a household from being able to farm and or derive some benefit from the environment. A degraded environment may also heighten the risk of landslides, flooding, and other disasters.

3. Themes in Environmental Vulnerability

The last few paragraphs have been used to illustrate what is meant by the concept environmental vulnerability and how it can be used. This discussion is now extended to illustrate how environmental vulnerability has been used in various fields. Assessments of environmental vulnerability have collectively framed wider dialogue and discussion and provided a focal point in the fields of food security, climate and global change, environmental resource management, and disaster management. The focus over time for those using environmental vulnerability to inform food security has evolved to include a more holistic understanding of vulnerability. In this field, the locus of attention has moved beyond the examination of the numbers of people affected, their daily food caloric requirements, and tons of food available. Today, such efforts converge on the 'real people' including attempts to capture the daily suite of activities that people use to live and often to survive together with those factors that may undermine or encroach on these capacities. Vulnerability and chronic food insecurity, therefore, could potentially be driven by policy failures (e.g., war, economic strategies); resource poverty/bad luck, disaster (e.g., environmental hazards, asset distribution); as well as 'population transition' (e.g., demographic growth) (see Watts and Bohle 1993, Ribot et al. 1996).

As groups are compelled to exert greater pressures on resources in order to adjust, manage and, in some cases, survive, they may negatively shape and impact on their physical environment. Recognizing this relationship, development practitioners, scientists, and others have begun to adapt the concept to improve our understanding of these interactions. Sustainable resource utilization is being understood through the lens of environmental vulnerability (see *Environmental and Resource Management*). Some are using various interpretations of vulnerability to understand and unravel how institutions and issues of access and entitlement to resources shape and configure various environments (e.g., Leach et al. 1997). Access to natural resources such as trees, plants, grasses, which are exchanged for currencies and enable people to make a living, may be controlled by institutions at various levels (e.g., governments, state policies, relations within households). The strength of this institutional control may preclude individuals or groups gaining access to the resources they use and which in turn may heighten their vulnerability during times of need, shock, stress, or in changing environmental conditions (e.g., see *Global Environmental Change: Human Dimensions*).

Other practitioners including nongovernmental organizations (NGOs), governments, humanitarian, and relief agencies have also begun articulating their understanding of vulnerability as an integral component of disaster risk. Here vulnerability is used to inform mitigatory and risk-reduction measures to offset disasters. Organizations including The International Federation of the Red Cross and Red Crescent Societies, Catholic Relief Services, Save the Children Fund, and several disaster management and risk reduction networks, including those established in South America (e.g., LA RED), in southern Africa (PeriPeri), and in South Asia (e.g., Duryog Nivaran), are asking 'Why are some people and groups more vulnerable to disasters than others?' (e.g., Ariyabandu 1999, Holloway 1999). This focus on vulnerability and risk reduction to disasters has been underscored at the international level as one of the tenets underpinning the United Nations General Assembly Resolution (44/236) adopted in 1989 when the decade 1990–2000 was declared the International Decade for Natural Disaster Reduction (IDNDR).

The growing interest in vulnerability in recent years can be largely attributable to the variety of entry points that 'environmental vulnerability' gives those in fields influenced by the social and biophysical context of environmental change. When a sensitive understanding of vulnerability has been obtained then, it is argued, this knowledge can be used to inform a variety of interventions including development policy which incorporates disaster-related strategies. Several notions of vulnerability are therefore being crafted together into discourses on poverty (in both rural and urban settings), food security, disasters, rural and urban livelihoods, and assessments of local and global environments that are at risk. Despite this broad range of interest, environmental vulnerability remains a complex issue, demanding complicated methods to identify environmentally vulnerable groups or regions.

4. *Current and Future Potential Usage of Environmental Vulnerability*

Several examples of how environmental vulnerability can be used to inform planning, science, and development work can be found in a number of countries including those in Africa, South America, and South Asia. For example, in Mozambique, government, NGOs, and various other organizations (e.g., United Nations World Food Program) have used a vulnerability analysis and mapping (VAM) approach to demarcate and identify those areas and groups that are vulnerable. By using geographical information systems and other mapping techniques, areas that are prone to flood risk, rainfall shortage, suffer from poor vegetation and where food availability and food access are problems have been identified. This information is not only useful to define vulnerable areas but can also be used to inform and help target efforts by policy makers and other organizations that may be involved in development work.

In South Asia, disaster and risk-reduction networks such as Duryog Nivaran have been using environmental vulnerability as a concept and the basis for their disaster mitigation work in the region (Ariyabandu 1999). This network has illustrated that environmental vulnerability and mitigation is linked not only to understanding technological factors but also to increased sensitivity of how grass-root communities have used their experience to increase their resilience to risky environmental situations. Close examination of how gender plays a role in either heightening or reducing vulnerability has, for example, highlighted the role that women play in preparing and mitigating disasters such as droughts and cyclones. Through the sharing of their knowledge and experience women have enabled each other to face the challenges of both daily living as well as being strengthened for times of shock associated with extreme weather events.

5. *Environmental Vulnerability and the Future*

Much has been done, and still remains to be done, to advance and adapt the use of environmental vulnerability assessments to other contexts. Identification of critical thresholds of environmental vulnerability; improved indicators of environmental vulnerability in developed, technological, and developing worlds; and the identification of ways to capture issues of scale and space are current and future areas of interest. How these feed into and are in turn impacted by global change are central priority research areas of several local and international initiatives.

Vulnerability reduction, some have argued, has not been a priority in development and environmental policy and has often been a response to discrete events, linked only to disaster management. A critical area, therefore, is to better integrate environmental vulnerability and vulnerability assessments into development and environmental planning. These efforts will also hopefully assist in mitigating the consequences associated with possible changes in global environment which may accompany a warmer, more crowded, poorer, disaster-prone world.

Notwithstanding the difficulties of trying to understand, profile, and articulate notions of environmental vulnerability, it remains useful for a range of users including hazard/risk scientists, policymakers, donors, relief agents, and development practitioners. The complexity embedded in environmental vulnerability is perhaps its greatest strength because it opens ways for greater interaction, discussion, and discourse in an age of greater globalization and information exchange and one which may bring with it a variety of 'millennium bugs,' 'surprises,' and 'new vulnerabilities.'

See also: Environmental Challenges in Organizations; Environmental Surprise

Bibliography

Ariyabandu M M 1999 *Defeating Disasters: Ideas for Action*. A Duryog Nivaran publication funded by the Conflict and Humanitarian Affairs Department of the Department for International Development, UK (CHAD-DFID)

Blaikie P, Cannon T, Davis I, Wisner B 1994 *At Risk: Natural Hazards, People's Vulnerability and Disasters*. Routledge, London

Carney D (ed.) 1998 *Sustainable Livelihoods: What Contribution Can We Make?* Department for International Development, London

Chambers, R 1989 Editorial introduction: Vulnerability, coping and policy. *IDS Bulletin* **20**(2): 1–7

Chambers R, Conway G R 1992 *Sustainable Rural Livelihoods: Practical Concepts for the 21st Century*. Institute of Development Studies, Discussion Paper, 296, Institute of Development Studies IDS, University of Sussex, Brighton, UK

Cutter S L 1996 Vulnerability to environmental hazards. *Progress in Human Geography* **20**(4): 529–39

Davies S 1996 *Adaptable Livelihoods: Coping with Food Insecurity in the Malian Sahel*. Macmillan
Downing T, Watts M, Bohle H 1996 Climate change and food insecurity: Toward a sociology and geography of vulnerability. In: Downing T E (ed.) *Climate Change and World Food Security*. NATO ASI Series, Global Environmental Change, pp. 183–206
Hewitt K 1997 *Regions of Risk a Geographical Introduction to Disasters*. Longman
Holloway A (ed.) 1999 *Risk, Sustainable Development and Disasters: Southern Perspectives*. University of Cape Town, Cape Town, South Africa (a PeriPeri publication)
Kasperson J, Kasperson R, Turner B 1995 *Regions at Risk. Comparisons of Threatened Environments*. The United Nations University Press, Tokyo
Kates R W, Hohenemser C, Kasperson J X (eds.) 1985 *Perilous Progress: Managing the Hazards of Technology*. Westview Press, Boulder, CO
LA RED 1993 *The Network for Social Studies on Disaster Prevention: Research Agenda and Constitution*. ITDG, Lima, Peru
Leach M, Mearns R, Scoones I 1997 *Environmental Entitlements: A Framework for Understanding the Institutional Dynamics of Environmental Change*. Discussion Paper 359, Institute of Development Studies IDS, University of Sussex, Brighton, UK
Liverman D M 1990 Vulnerability to global environmental change. In: Kasperson R, Dow K, Golding D, Kasperson J (eds.) *Understanding Global Environmental Change: The Contributions of Risk Analysis and Management*. Clark University, Worcester, MA, pp. 27–44
Mullins D 1999 Livelihoods and vulnerability analysis: From basic concepts to fieldwork. Paper presented at a conference on *Risk, Sustainable Development and Disasters: Perspectives from the South*, a PeriPeri Institute initiative, Harare, Zimbabwe, 22 February
Ribot J C, Magalhaes A R, Panagides S S 1996 *Climate Variability, Climate Change and Social Vulnerability in the Semi-arid Tropics*. Cambridge University Press, Cambridge, UK
Swift J 1989 Why are rural people vulnerable to famine. *IDS Bulletin* **20**(2): 8–15
Watts M J, Bohle H G 1993 The space of vulnerability: The causal structure of hunger and famine. *Progress in Human Geography* **17**(1): 43–67
White G 1974 *Natural Hazards: Local, National and Global*. Oxford University Press, New York

C. Vogel

Environmentalism: Philosophical Aspects

Environmentalism is an intellectual, moral, and political movement that arose in response to global environmental crises. Although it had its precursors in forestry preservation and national parks campaigns of the twentieth century, the contemporary movement is distinguished by a perception that there is a global, complex, interrelated web of environmental problems that pose a threat to the health, well-being, and perhaps the very existence of humanity. Even local problems—the pollution of a lake, the disappearance of a species—are perceived by environmentalists to be symptomatic of a general malaise.

Developments in the 1950s and 1960s—concerns about population growth, nuclear war and nuclear technology, the appearance of Carson's book on the effect of pesticides (1965), and the development of the science of ecology—helped to create the environmentalist outlook (McCormick 1989, pp. 47–68). The Club of Rome's influential study *Limits to Growth* (Meadows 1972) used computer models to argue that exponential growth in population, use of resources, pollution, and industrial growth, as well as interaction between these factors are leading the world rapidly toward environmental catastrophe. Although the assumptions made by the Club of Rome have been subjected to extensive criticism, environmentalists have found no reason to doubt that serious problems exist, are becoming more severe, and cannot be solved without far-reaching moral and political changes.

Environmentalism can be understood as a theoretical position (a view about the nature and causes of environmental crises), as a moral position that poses the question of what our relation to nature ought to be, and as a social and political movement aiming to bring about a society capable of living within environmental limits. These aspects of environmentalism are interdependent but will be dealt with separately in the following sections.

1. *Explanations of Environmental Crises*

Imagine a commons, says Hardin, where a community of herdsmen graze their cattle. As time goes by, the demands they make on the pasture increase, and they are aware that it will be destroyed in the not so distant future unless they impose limitations on their activities. Nevertheless, each herdsman reasons that it is in his best interests to continue to add more cattle to the commons. Since the cost of adding an extra cow will be borne by the community as a whole, he will gain more than he loses from the addition to his herd. If, for the sake of future benefits, he puts limits on the size of his herd, then he is likely to lose future as well as present gains, since the other herdsmen cannot be trusted to do the same (Hardin 1972, p. 117). 'Ruin is the destination toward which all men rush each pursuing his own best interests in a society that believes in the freedom of the commons' (Hardin 1968, p. 1244).

Hardin applies this lesson primarily to the problem of population growth, but it can easily, and more plausibly, be adapted to account for other kinds of environmental problems—pollution, overuse of resources, and destruction of environments and species. Factory owners who spill wastes into the air or the water are exploiting the freedom of the commons.

Fisherfolk whose activities wipe out the fish populations on which they depend and farmers whose irrigation practices cause rivers to become saline can be understood to be operating according to the logic Hardin describes. In fact, the dilemma he identifies, or something like it, can be used to account for destruction of environments that are not commons (contrary to what he supposes). Farmers who recognize that a farming practice will eventually ruin their own land may continue to use it because this is the only way in which they can remain competitive and avoid losing out in the short term.

Hardin regards the 'tragedy of the commons' as the consequence of human nature coming up against environmental limitations. Others, noting that some societies have managed to live within these limitations, think that the underlying causes of environmental destruction are cultural and historical. White (1967) blamed Christianity, particularly puritanical Protestantism, for encouraging a disdain for nature. Leiss (1972) claimed that domination of nature is the rationale behind Western science and technology. Other environmentalists prefer to think that the problem is due to certain kinds of technological development (Commoner 1971) or a particular scientific worldview (Matthews 1991). Marxists blame capitalism for the exploitation of nature, although they have had to acknowledge that socialism is not necessarily a remedy. Anarchists regard environmental destruction as one of the consequences of political centralization and social hierarchy (Bookchin 1980), and ecofeminists have argued that domination of nature and domination of women have the same social-cultural roots (Plumwood 1993).

The interest in finding the cause of environment crisis has waned. Most contemporary environmentalists suspect that the problems they are concerned with have many causes, and that it is a mistake to try to reduce this complexity to a single explanation or to assume that all have the same cause. Although Hardin's 'tragedy of the commons' analysis does not provide an explanation of all environmental problems or a complete explanation of any, it has the virtue of explaining how environmental destruction can persist when almost everyone acknowledges the need to do something about it. Any proposal for social and political change must come to grips with this problem.

2. *Environmentalist Ethics*

Environmentalism was motivated from the beginning by a concern about human survival and well-being. This is a moral as well as a prudential worry. Those living now—at least those who are older—may escape the worst, but environmental damage is left as a legacy for their successors. A concern for future people and the future of humanity—so obvious as a motivation to most people that it hardly needs to be stated—continues to be the moral reason for the protests and actions of most environmentalists. However, environmental problems raise issues that require theoretical reflection on moral and social values. They pose questions about how human beings ought to live and what they should regard as conducive to their well-being. They create a need to think about requirements of justice in respect to both relations between existing people and our relations to future generations. Even more fundamentally, they bring into question attitudes toward nature, and give rise to reflections about how nature should be valued.

2.1 *Environmental Humanism*

The moral point of view that most environmentalists have adopted could be called 'humanism.' Humanism demands that every human individual be treated with equal respect. An action or policy can be judged morally right only if the interests of all affected individuals have been equally considered and taken into account—and these individuals, environmentalists insist, have to include not only everyone in the world but members of future generations.

Humanism takes many forms and humanist environmentalists often disagree with each other in their moral theories and the conclusions they reach about what should be done. Proponents of 'lifeboat ethics'—those who think that the wealthy should refuse aid to poor countries that do not make efforts to become environmentally sustainable—are consequentialist humanists who think that the greatest good, taking into account the interests of present and future individuals, is best served by this policy (Hardin 1974). Marxists who respond to environmental crises by favoring a democratic socialism that eliminates waste and exploitation are humanists who think that respect for individuals requires the promotion of egalitarian and cooperative social relations. Environmental humanists with a liberal concern for social justice emphasize the need for an equitable distribution of resources and risks among existing and future individuals.

The issue of what we should do for future generations is a central concern of environmental ethics which requires not only that we think about what we ought to preserve for our descendants. It also brings into question common ideas about the scope of our duties to future people. It is commonly accepted that we have duties of justice only to our immediate successors. However, environmentalists point out that some of our activities—for example, our use of nuclear technology—leave behind waste products that can harm people thousands of years from now.

2.2 *Ecocentric Ethics*

The main debate in environmental ethics is about how nature should be valued. Humanism—indeed the

whole Western ethical tradition—has been persistently challenged by those who object to its anthropocentrism—its assumption that only human beings and their interests are ethically considerable. Ecocentered environmental ethics—the view that value in nature cannot be reduced to what promotes human well-being—has always been a minority position among environmentalists. Nevertheless, ecocentered environmentalists have had considerable influence in the environmental movement and an even larger impact on ethical theory. More radically than the animal liberation movement, ecocentrism has challenged traditional ways of thinking about value and morality.

Leopold, commonly thought to be one of the first to put forward an ecocentered position, notes that Odysseus, returned home from his travels, had no moral qualms about killing the slave girls who had consorted with his enemies. 'The girls were property. The disposal of property was then, as now, a matter of expediency, not of right and wrong' (Leopold 1949, p. 217). We no longer think that any human being should be treated as the property of another, he continues. But our relation to the land, soil, plants and other creatures that are part of our biotic community remains an economic one. He calls for a land ethic that 'enlarges the boundaries of the community to include soils, waters, plants, animals, or collectively: the land' (Leopold 1949, p. 224).

Leopold claims that a land ethic is the natural result of a process of ethical evolution. Some philosophers make a case for ecocentrism by attempting to show that the scope of humanitarian ethics can and should be widened to include as objects of moral concern all living organisms—and perhaps other things in nature. Most morally sensitive people, says Taylor, are already convinced that many of the qualities that make human beings morally considerable are possessed by sentient animals. They too have interests and feelings, and can be hurt or benefited by what we do. These similarities are a good reason for saying that animals as individuals are deserving of moral respect. A plant or a nonsentient animal is like a sentient creature in that it has a good of its own—the full development of its biological powers. If we acknowledge that this is so, and acknowledge human dependence on the complex web of life, we will be predisposed to recognize each organism as having inherent worth—a value that compels our moral respect (Taylor 1986, pp. 59–98). Johnson argues that species, wilderness areas, and other natural systems have interests in their own right and thus are valuable for the same reasons that human beings and other sentient animals are valuable (Johnson 1991, pp. 97–183).

Taylor and Johnson differ on whether parts—individual organisms—or wholes—species and environmental systems—have inherent worth. However, the difference between individualism and holism in environmental ethics is less important than it seems. The fact that human beings and sentient creatures are conscious, and thus have a point of view, gives us a natural means of distinguishing individuals from each other and individuals from the wholes of which they are part. Where consciousness is not present, individuation and the distinction between part and whole are not likely to be ethically crucial. In fact, many ecocentered environmentalists are prepared to assign inherent value to both individual organisms and the systems that contain them. A more serious disagreement among ecocentered environmentalists exists between those who value creatures or environments because they have qualities similar to human qualities, and those who think that this reason for valuing nature is fundamentally human-centered and should be rejected.

Those who reject analogical arguments for valuing nature have to find another basis for their ecocentric ethics. Some find this in a worldview that sees the cosmos or the Earth as a self (Matthews 1991, pp. 142–7). Others think that a sympathetic engagement with sciences such as ecology will lead us to an ecocentric position (Callicott 1989, p. 125), or that we will be pushed in that direction by reflection on our intuitions. Richard and Val Routley (Routley and Routley 1980, pp. 121–3) ask us to imagine that we are the last people on the planet. Would we regard it as morally permissible to destroy its ecosystems if we did not have to worry about the well-being of future generations? They expect us to answer 'no,' and as a result to recognize that we are already inclined to value nature for its own sake. We are then required to identify the characteristics that give natural systems value—they suggest diversity, naturalness, integrity, stability, and harmony—and to give these valuable systems their due when we engage in moral reasoning.

Its critics regard ecocentered environmentalism as implausible or even potentially dangerous. Why should we extend the scope of our moral concern beyond creatures who are capable of suffering? How can we accept an ethics that might prescribe policies contrary to the human good? Moreover, ecocentric ethics faces conceptual problems. If it asks us to value anything that can be perceived to have a good of its own or every system that is balanced, ordered, and integrated, then how can we avoid a requirement to value for their own sake machines, bodily organs, or social systems? Why should such an ethics have a special regard for creatures now living and systems now existing rather than equally (or more) complex, integrated environmental systems that have existed or might exist?

However, even if ecocentered ethics turns out to be implausible or unacceptable, its objective of persuading us that nature should be valued for what it is—and not simply as a resource—is reasonable and can be achieved within the confines of a human-centered ethics. If, for example, things in nature are valuable because of their aesthetic qualities then we have reason to make an effort to appreciate and preserve them. In

doing so we are valuing them for what they are. Wilderness, landscapes, species, the land itself can also be appreciated as heritage or simply loved and cherished because these environments or creatures are part of our world. Regarding ourselves as part of a community that includes our natural environment (as Leopold recommended) can enhance our sense of belonging to a place and thus enrich our lives. Such valuing retains its human center—but it may give us sufficient reason for thinking that the last people should not destroy their world.

2.3 Deep Ecology

There is another approach to valuing the environment that is sometimes regarded as an ecocentric ethics. Instead of recommending that we widen the scope of our ethical concern, deep ecology asks us to broaden and deepen our sense of self. Influenced by Buddhism and Gandhi's philosophy, Arne Naess (1973) argues that self-realization requires an identification with others. Having this identification means that self-love becomes love for others, and individuals seek not only their own ends but also the good of those with whom they identify. Naess believes that recognition of the equality of living things makes it natural and desirable for the self to identify with all creatures. By doing so we realize ourselves in a more complete and satisfactory way.

Deep ecology has been criticized, particularly by ecofeminists, for trying to master the world by incorporating it into the self (Cheney 1987). They insist that beings in nature have an independent existence and should be appreciated for what they are. It is true that some of the feats of identity that deep ecologists claim to perform are pretentious if not silly ('I am the black bear'). But there is probably no great distance between deep ecologists and feminist critics. It seems undeniable that people can learn to care about natural creatures and environments to the point where they feel diminished if these things are destroyed, and that this kind of caring can be a motivation for environmental preservation.

3. Environmentalist Politics

Environmentalists believe that by promoting an environmental ethics they are making it more likely that environmental problems will be solved. However, few deny that there is a need for an environmental politics: not merely campaigns to save the environment or to educate citizens but an environmental political theory and a long-term strategy for achieving an environmentally sound society. There are two major questions for environmentalist political theory. The first is what it means for a society to live within environmental limits. The second is how this goal is related to important political values such as liberty, democracy, justice, equality, solidarity, or liberation of women.

3.1 From 'No Growth' to 'Sustainability'

The Club of Rome report and other warnings of imminent environmental disaster encouraged environmentalists to believe that an environmentally sound society had to be a 'no-growth society,' and that achieving it had to become a priority everywhere in the world. The goal of 'sustainable development' popularized by the World Commission on Environment and Development (1987) allows that development and growth can occur within the carrying capacity of the planet. Sustainable development is supposed to leave room for development within poorer countries and encourage those in wealthier countries to look for more environmentally sound ways of maintaining their living standards.

There is a lot of room for disagreement, not only about what the environment can sustain but also about the relation between environmental preservation and sustainability. Sustainable development does not necessarily mean that wilderness areas or threatened species will be preserved, or that the values that ecocentered environmentalists find in nature will be respected. 'Sustainable development' also raises normative questions about how much environmental risk it is legitimate to impose upon present and future people for the sake of development.

3.2 Environmentalism and Political Values

There are three basic ways in which environmentalists have answered the question of how environmental objectives should affect or be affected by political values. They can be called 'pessimist,' 'compatibilist,' and 'utopian.' Pessimists think that political values such as freedom and democracy have to be abandoned or severely limited, at least temporarily, for the sake of avoiding environmental disaster. Saving ourselves and the environment has to take priority over all other objectives. Compatibilists think that necessary environmental reforms are compatible with the preservation or achievement of political values—although some restrictions on the realization of these values have to be accepted. Utopians think that in the right sort of society people would not only live in harmony with nature but enjoy more freedom, democracy, justice, equality, and communal solidarity than we do at present.

3.2.1 The pessimist approach. Ophuls regarded environmental crises as calling for a new social contract. We need a Leviathan, a Hobbesian state that is not impeded from taking the steps necessary for

protecting the environment (Ophuls 1977, pp. 222–48). Heilbroner concluded from an assessment of human behavior and existing social systems that democracy cannot be expected to survive the transformation to an environmentally sound society (Heilbroner 1974, p. 90).

Ophuls and Heilbroner feared the imminent collapse of environmental systems. Those who do not regard the danger as so great or so immediate are not inclined to accept their assessment of our social possibilities. Ophul's advocacy of a Hobbesian solution to the political problems posed by environmental crisis has been widely condemned as unnecessarily authoritarian. It is also open to the criticism that an authoritarian solution is naive. There is no guarantee that a 'green dictatorship,' once installed, would be willing or able to save the environment.

3.2.2 The compatibilist approach. Liberal compatibilists argue that an environmentally sound society can be achieved within the framework of a free, just, and democratic society, although doing so is likely to require adjustments and restrictions. Most compatibilists agree that more restrictions will have to be placed on individual freedom. Polluting activities will have to be curbed, use of resources restricted, and, in some countries at least, population growth checked. The debate centers on what kinds of restriction would be effective and morally acceptable. Let us assume, for example, that restrictions on family size are an effective way of limiting population growth. The question remains whether the measures required constitute an unacceptable abuse of freedom. Would a scheme of incentives or disincentives be more acceptable from a moral point of view—even though they are likely to affect the poor more than the rich?

Many compatibilists also accept the necessity of restrictions on the scope of democratic decision making. Recognizing environmental rights of individuals, it has been suggested, would define the limits within which democratic decision making should operate. Christopher Stone (1974) argues that natural objects, such as rivers or wilderness areas, would be better protected if they were accorded legal rights.

Liberals who aim for an environmentally sound society cannot ignore questions of justice. Although environmental problems are global, they affect some people more than others. The burden of solutions to environmental problems can also be unfairly distributed. A pollution control measure that has the effect of raising consumer prices will disadvantage those who are in the least well-off sections of the population unless the costs are offset by other factors. A scheme of incentives or disincentives for limiting family size or the use of private cars and other environmentally harmful activities is likely to be experienced by the poor as a form of coercion. Revising or adapting theories of distributive justice to deal with environmental problems and devising measures for achieving an environmentally sound society compatible with justice is one of the biggest challenges for environmentalist political theory.

3.2.3 The utopian approach. One of the first responses to predictions of environmental disaster was *Blueprint for Survival* (Goldsmith 1972) which urged not only the reduction of waste and the stabilization of population but the creation of a decentralized social system. It recommended the relocation of the population of existing states into self-sufficient communities of about 500 people who would use small-scale, low-impact technology to provide for their own needs. The communalist vision of an environmentally sound society has been popular particularly among left-wing environmentalists who rejected both the capitalist state and centralized socialist bureaucracies, and turned to anarchist ideas about social organization and community. Many Green parties, for example *Die Grünen* of Germany, regard the achievement of decentralized, egalitarian, and democratic political and economic institutions as their ultimate political goal. The US bioregionalist movement advocates self-sufficient, 'human-scale' communities dedicated to preserving and enhancing the diversity of native ecosystems. Many ecofeminists think that only in a nonhierarchical society where qualities associated with the feminine are valued will nature and women be truly liberated.

The reasoning that motivates these visions of environmental utopia is that people who relate to each other in a face-to-face nonhierarchical community, exercise control over the conditions of their lives through participatory democracy, and have a close relation to their environment will live in harmony with each other and nature without the imposition of coercive political rule. They will not experience restrictions imposed on their activities for the sake of the environment as a sacrifice of their self-interest, because they value and identify with their community and its caring relation to the environment.

Some critics think that ecoanarchism depends on an implausible view of human nature. Be that as it may, traditional societies and small communities have often been nondemocratic, restrictive, hierarchical, and patriarchal, and they have not always lived in peace with their neighbors. Ecoanarchists tend to assume, without sufficient discussion, that decentralization will produce the best results as far as communal life is concerned rather than one of the worst. Ecoanarchists also fail to provide sufficient detail about how communities will coordinate their affairs in order to deal with environmental and other intercommunity problems. The tragedy of the commons, even if avoided at a local level, could reappear as a problem concerning the relation between communities.

Ecosocialists claim with some justification that Marx's criticisms of capitalism can be used to account for the existence of many environmental problems. Driven by the necessity for making profit, owners and managers of industry cannot afford to be too concerned about causing environmental damage. A socialist society, say ecosocialists, would be able to establish social relations on a more cooperative, democratic basis. To avoid the pitfalls, moral and environmental, of bureaucratic socialism, ecosocialists borrow heavily from ecoanarchists. They too want decentralization, close relations with nature, and participatory democracy. However, they are less prepared than ecoanarchists to do away with the state and its ability to coordinate economic activity and ensure that wealth is fairly distributed among individuals and communities (Pepper 1993, pp. 226–7).

Although it is not well known or popular among environmentalists, libertarians have also presented a utopian vision of a society that protects the environment and upholds individual rights of life, liberty, and property (the only values that libertarians think a state should protect). Right to property includes the right to enjoy it without having to suffer from the polluting activities of others. A state that rigorously protects individual rights would not allow anyone to pollute without permission of those whose lives and properties are affected (Machan 1984).

3.3 Environmentalism and Global Politics

A common failing of all approaches to political theory of the environment, whether pessimist, compatibilist, or utopian, is that they concentrate mostly on the problem of transforming national societies into environmentally sound societies. There has been comparatively little reflection on practical and normative issues in relation to global society: for example, how environmental policies should affect or be affected by ideas about global justice; how utopian visions of the good life in an environmentally sound society can be translated into a blueprint for the world as a whole; whether the creation of more effective and powerful cosmopolitan institutions should be an environmentalist objective; whether global democracy, redistribution of the world's resources, and/or the recognition and protection of environmental rights as universal rights are feasible and desirable objectives; whether individuals should come to regard themselves as 'planetary citizens' and what this would mean. Finding answers to these questions is going to become more urgent as the twenty-first century progresses.

See also: Ecology, Cultural; Ecology, Deep; Ecology, Political; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Conservation: Ethical Concerns; Environmental Justice; Environmental Policy; Environmental Policy: Protection and Regulation; Environmentalism, Politics of; Environmentalism: Preservation and Conservation; Global Environmental Change: Human Dimensions; Nature, Concepts of: Environmental and Ecological

Bibliography

Bookchin M 1980 *Toward an Ecological Society*. Black Rose, Montreal, Canada

Callicott J B 1989 *In Defense of the Land Ethic: Essays on Environmental Philosophy*. State University of New York Press, New York

Carson R 1965 *Silent Spring*. Penguin, Harmondsworth, UK

Cheney J 1987 Ecofeminism and deep ecology. *Environmental Ethics* **9**: 115–45

Commoner B 1971 *The Closing Circle: Nature, Man and Technology*. Alfred A. Knopf, New York

Goldsmith E 1972 *A Blueprint for Survival*. Penguin, Harmondsworth, UK

Hardin G 1968 Tragedy of the commons. *Science* **162**: 1243–8

Hardin G 1972 *Exploring New Ethics for Survival: Voyage of the Spaceship Beagle*. Viking, New York

Hardin G 1974 Living on a lifeboat. *Bioscience* **24**: 561–8

Heilbroner R L 1974 *An Inquiry into the Human Prospect*. Norton, New York

Johnson L E 1991 *A Morally Deep World: An Essay on Moral Significance and Environmental Ethics*. Cambridge University Press, Cambridge, UK

Leiss W 1972 *The Domination of Nature*. George Braziller, New York

Leopold A 1949 *A Sand County Almanac*. Oxford University Press, Oxford, UK

Machan T R 1984 Pollution and political theory. In: Regan T (ed.) *Earthbound: New Introductory Essays in Environmental Ethics*. Random House, New York, pp. 74–105

Mannison D S, McRobbie M, Routley R (eds.) 1980 *Environmental Ethics*. Research School of Social Sciences, Australian National University, Canberra, Australia

Matthews F 1991 *Ecological Self*. Routledge, London

McCormick J 1989 *Reclaiming Paradise: The Global Environmental Movement*. Indiana University Press, Bloomington, IN

Meadows D L 1972 *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*. Pan Books, London

Naess A 1973 *The Shallow and the Deep, Long-range Ecology Movement: A Summary*. Inquiry **16**: 95–100

Ophuls W 1977 *Ecology and the Politics of Scarcity*. W. H. Freeman, San Francisco

Pepper D 1993 *Eco-socialism: From Deep Ecology to Social Justice*. Routledge, New York

Plumwood V 1993 *Feminism and the Mastery of Nature*. Routledge, London

Regan T 1984 *Earthbound: New Introductory Essays in Environmental Ethics*. Random House, New York

Routley R, Routley V 1980 Human chauvinism and environmental ethics. In: Mannison D S, McRobbie M, Routley R (eds.) *Environmental Ethics*. Research School of Social Sciences, Australian National University, Canberra, Australia, pp. 96–189

Stone C D 1974 *Should Trees Have Standing? Toward Legal Rights for Natural Objects*. William Kaufmann, Los Altos, CA
Taylor P W 1986 *Respect for Nature: A Theory of Environmental Ethics*. Princeton University Press, Princeton, NJ
White L 1967 The historical roots of our ecological crisis. *Science* **155**: 1203–7
World Commission on Environment and Development 1987 *Our Common Future*. Oxford University Press, Oxford, UK

J. Thompson

Environmentalism, Politics of

Environmentalism is one of several 'new social movements' challenging institutionalized politics. That movement strives to protect the natural environment, both for its own sake and ours, from unsustainable exploitation and degradation through pollution, overpopulation, maldistribution, technological excess, and resource extraction. Politically, greens adopt a decentralized and participatory style of protest politics which is often orthogonal to the workings of formal political institutions. Internationally, various environmental protection regimes have been created by intergovernmental agreements and energized by nongovernmental organization activism.

1. Environmentalist Concerns

Although their concerns sometimes extend to the 'built' or 'social' environment, self-styled environmentalists generally seek to protect the 'natural' environment. Natural environments are rarely pure, of course, and few places are literally untouched by human hands. Heather dominates the Scottish moors because of past agricultural practices, for example, and eucalyptus the Australian bush because of Aboriginal 'firestick' land management. Many sites and species owe their social salience to the resonance they have with human history and heritage, symbolisms, and meanings. Thus, it is apt to say (more anthropocentrically than some 'deep ecologists' would prefer) that political concern is indeed with 'our' environment (Goodin 1992, Chap. 2).

Some of the interests which environmentalists have in view are similarly anthropocentric. Overpopulation (Ehrlich 1968), maldistribution (Brandt 1980), and resource depletion threaten to starve us: Malthus' (1803) old prophecy has acquired new urgency through modern computer modeling (Meadows et al. 1972, Council on Environmental Quality and the Department of State. 1980). Pollution threatens to poison us and technology to overwhelm us (Commoner 1971). Greenhouse gases threaten global warming, melting the polar ice caps, and CFCs threaten the ozone layer that protects us from the sun's ultraviolet rays (McKibben 1989). To nuclear power greens say, 'No thanks!', in part because they see it as unnecessary (like 'cutting butter with a chain saw'), in part because of risks of radioactive leaks from operating plants (such as at Chernobyl) or waste storage facilities, and in part because of fears of diversion of nuclear materials to bomb-building (Goodin 1982, Chap. 10, Lovins et al. 1980).

Anthropocentric concerns merge with biocentric ones, insofar as what is bad for humans is bad for other species and ecosystems as well. One of the cornerstones of the modern environmental movement —Rachael Carson's *Silent Spring* (1962)—took the effect of DDT on birds as a harbinger of the dangers of chemical pollution to human health. The interconnectedness of all life is an ecological theme echoed repeatedly in the titles of major international reports, such as the report of the World Commission on Environment and Development, *Our Common Future* (1987), and the manifesto of the 1972 Stockholm Conference, *Only One Earth: The Care and Maintenance of a Small Planet* (Ward and Dubois 1972).

Environmentalists are also concerned with the well-being of animals, species or ecologies for their *own* sake, rather than because of the fate they share with humans. Environmentalists campaign to 'Save the Whales' from extinction, not because they want to eat them, but because they think that those noble and intelligent creatures have as much right to live as people do. Environmentalist concern sometime extends beyond the animal kingdom to inanimate creation. Biodiversity is just as valuable among plants as animals. Natural landscapes, wild rivers, and majestic canyons have a value which is not reducible just to the value of the birds and fish that live in them.

Environmentalist concern also extends widely in both time and space. Environmentalists are concerned with the legacies that current activities might be leaving for distant future generations, and with the impact that they might have on distant locales which none of them personally ever expect to visit. The wrongness of the harms we are inflicting on those distant peoples and places, environmentalists say, is in no way mitigated by the sheer fact of their distance, either in time or space. Neither is it mitigated by the inevitable uncertainty attaching to action at such a distance. Whereas economic calculations ordinarily discount uncertain effects, greens urge a 'precautionary principle' which in the face of uncertainty assumes the worst and strives to protect against it (Goodin 1982, Chap. 9, Weale 1992, UN 1992, Principle 15).

Many environmentalists are thus keen advocates of 'environmental justice,' internationally as well as domestically. The maldistribution of resources is seen as a key cause of environmental degradation, and remedying the one as a key step in remedying the other (UN 1992). Various schemes for 'sustainable devel-

opment' and 'ecological modernization' suggest ways to 'develop without damage' (Dryzek 1997, Chaps. 7–9, Dobson 1998). Ecofeminists, however, suppose that man's dominance of nature is of a cloth with men's dominance of women, and problems of environmental degradation cannot be resolved without resolving problems of gender subordination (Plumwood 1993).

2. *Movement Politics*

Politically, the key environmental slogan is: 'Think globally, act locally.' Although the scope of their concerns is global, many environmentalists suppose that the most effective way of pursuing them is through decentralized structures and local action (Goldsmith et al. 1972). For greens, 'small' is indeed 'beautiful' (Schumacher 1973).

Much of environmentalist action takes the form of protest politics, ranging from street marches and publicity stunts to eco-sabotage. An example of the latter is the radical US group, 'Earth First,' which advocates 'monkey-wrenching' modern technology, including everything from cutting down power lines to driving spikes into trees in old-growth forests to provide them with a means of 'natural self-defence' against anyone who tries cutting into them with a chainsaw (Dryzek 1997, Chap. 9, Goodin 1992, pp. 133–5). Less radical groups such as Greenpeace engage in direct but nonviolent action, such as stationing their ship *Rainbow Warrior* squarely within the exclusion zone surrounding the French nuclear test site in the South Pacific.

Those techniques of direct action and nonviolent mass protest are relatively standard across the other 'new social movements' with which environmentalists often make common cause, among them the peace movement, the feminist, gay, and lesbian rights movements, and various other human rights movements. Most of those share the same broad concern (and many of the same specific concerns) with the socioeconomic institutions of contemporary capitalism and the acquisitive, materialist values underlying them. Many organize themselves as loose networks rather than formal membership organizations. Many pursue their political objectives substantially through nonconventional means, outside the ordinary channels of established parties and legislative assemblies (Offe 1985).

Movement politics do sometimes take an institutional form, when environmentalists organize themselves into explicitly 'green parties' and interest groups (see *Green Parties*). The most famous among them, *Die Grünen* in Germany, entered the *Bundestag* in 1983 and joined the governing coalition led by Social Democrat Gerhard Schröder in 1998. Green parties were characterized, initially at least, by the same egalitarian, nonhierarchical ethos as the 'new social movements' from which they sprang (Douglas and Wildavsky 1982, Spretnak and Carpa 1986, Dobson 1990, Chap. 4, Goodin 1992, Chap. 4). They have decentralized organizational structures, characterized by grassroots democracy and rotating leadership. For some greens (such as the German *Fundi* faction), that political style is valued in itself, even if it costs them any real political power; but for others (the *Realos*, who won the German factional fight), the whole point of a political party is to seize power to do good, and internal party structures ought be aimed toward those ends (Goodin 1992, pp. 109–11, Dryzek 1997, pp. 173–5).

3. *Environmental Policy*

Many accomplishments of the environmentalist movement have come at the local or national rather than international level, and through the efforts of mainstream political parties rather than greens themselves. Examples range from the 1970 National Environmental Policy Act (NEPA) and the Environmental Protection Agency (EPA) in the USA to the 1989 Dutch National Environmental Policy Plan. The latter sort of regulatory regime is praised as more flexible and consultative and, because of that, more capable of altering pollution-generating practices. The US EPA is criticized for its inflexible 'command-and-control' approach and a fixation with 'end of the pipe' solutions, stopping pollutants from being released into the environment at the very end of the process rather than preventing them from being created in the first place (Dryzek 1997, Chap. 8, Weale 1992). Economists urge the replacement of regulatory policies with economic incentives, tradeable pollution permits, and the like (Pearce et al. 1989). It turns out, though, that even relatively toothless reporting requirements, such as the Environmental Impact Statements required by the US NEPA, can be surprisingly effective at 'making bureaucracies think' about the environmental consequences of their proposed actions (Taylor 1974).

Many environmental problems transcend national boundaries, however. Coordinated international action is then required, but that need not necessarily be universal in form. Early in these debates, George Kennan (1970) pointed out that most threats to the global environment come from the 'rich men's club,' and coordinated action among a majority of rich countries would go a very long way toward reducing those threats, whether or not others decline to join. Some of the most promising international agreements —notably, the 1987 Montreal Protocol restricting emission of ozone-destroying chlorofluorocarbons (CFCs)—are structured in precisely this way.

Institutionally, the 1972 Stockholm Convention led to the creation of a formal United National Environmental Program, based in Nairobi. Much of the impetus for international environmental protection has, however, come through the loosely coordinated actions of nongovernmental organizations (NGOs). Some of those are coalitions of environmental activist organizations, while others represent coalitions of

scientists in an 'epistemic community' driving the development of policies such as those embodied in the 1987 Montreal Protocol (Haas et al. 1993).

See also: Environment Regulation: Legal Aspects; Environmental and Resource Management; Environmental Change and State Response; Environmental Determinism; Environmental Economics; Environmental Health and Safety: Social Aspects; Environmental Justice; Environmental Policy; Environmental Policy: Protection and Regulation; Environmentalism: Philosophical Aspects; Environmentalism: Preservation and Conservation; Green Parties; Resource Geography; Social Movements: Environmental Movements

Bibliography

Brandt W 1980 *North–South*. Pan, London

Carson R 1962 *Silent Spring*. Houghton Mifflin, Boston

Commoner B 1971 *The Closing Circle*, 1st edn. Knopf, New York

Council on Environmental Quality and the Department of State 1980 *The Global 2000 Report to the President of the US*. Pergamon Press, New York

Dobson A 1990 *Green Political Thought*. Unwin Hyman, London

Dobson A 1998 *Justice and the Environment*. Oxford University Press, Oxford, UK

Douglas M, Wildavsky A 1982 *Risk and Culture*. University of California Press, Berkeley, CA

Dryzek J S 1997 *The Politics of the Earth*. Oxford University Press, Oxford, UK

Ehrlich P R 1968 *The Population Bomb*. Ballantine, New York

Goldsmith E, Allen R, Allaby M, Davoll J, Lawrence S 1972 A blueprint for survival: towards the stable society. *The Ecologist* **2**(1): 8–22

Goodin R E 1982 *Political Theory and Public Policy*. University of Chicago Press, Chicago

Goodin R E 1992 *Green Political Theory*. Polity, Cambridge, UK

Haas P M, Keohane R O, Levy M A (eds.) 1993 *Institutions for the Earth: Sources of Effective International Environmental Protection*. MIT Press, Cambridge, MA

Kennan G F 1970 To prevent a world wasteland: a proposal. *Foreign Affairs* **48**(3): 401–13

Lovins A B, Hunter Lovins L, Ross L 1980 Nuclear power and nuclear bombs. *Foreign Affairs* **58**: 1137–77

Malthus T R 1803 *Essay on the Principle of Population*, 5th edn. Johnson, London [Winch D (ed.) 1992 Cambridge University Press, Cambridge, UK]

McKibben B 1989 *The End of Nature*, 1st edn., Random House, New York

Meadows D H, Meadows D L, Randers J, Behrens III W W 1972 *The Limits to Growth*. Universe, New York

Offe C 1985 New social movements: challenging the boundaries of institutional politics. *Social Research* **52**: 817–68

Pearce D, Markandya A, Barbier E B 1989 *Blueprint for a Green Economy*. Earthscan, London

Plumwood V 1993 *Feminism and the Mastery of Nature*. Routledge, New York

Schell J 1982 *The Fate of the Earth*, 1st edn., Knopf, New York

Schumacher E F 1973 *Small is Beautiful*. Blond and Briggs, London

Spretnak C, Carpa F 1986 *Green Politics*, rev. edn., Bear, Santa Fe, NM

Taylor S 1984 *Making Bureaucracies Think: The Environmental Impact Statement Strategy of Administrative Reform*. Stanford University Press, Stanford, CA

United Nations (UN) 1992 Rio declaration on environment and development. Annex 1 of *Agenda 21*, Report of the United Nations Conference on Environment and Development. A/Conf.151.26 (Vol.I). United Nations, New York

Ward B, Dubois R 1972 *Only One Earth: The Care and Maintenance of a Small Planet*. Penguin, Harmondsworth, UK

Weale A 1992 *The New Politics of Pollution*. Manchester University Press, Manchester, UK

World Commission on Environment and Development 1987 *Our Common Future*. Oxford University Press, Oxford, UK

R. E. Goodin

Environmentalism: Preservation and Conservation

Environmentalism represents one of the most influential and enduring social movements of our time. Bursting onto the American scene in the late 1960s, the modern environmental movement has not only taken root in the USA but has spread across the globe. The environmental movement is widely considered to be one of the most successful social movements of the late twentieth and early twenty-first centuries, especially in terms of gaining widespread societal acceptance of its goals (Dunlap 2000), and it is likely that it will continue to play an important role in regional, national, and global politics. This article traces the growth and diversification of modern environmentalism in the USA, briefly reviews the growth of environmentalism worldwide, and then examines the major stages of contemporary environmental activism—conservationism, environmentalism, and ecologism.

1. The Emergence and Growth of the US Environmental Movement

The organizational and ideological roots of contemporary environmentalism are commonly traced to the progressive conservation movement that emerged in the late nineteenth century in reaction to reckless

exploitation of the United States' natural resources. The progressive conservation movement is typically divided into two broad camps: utilitarian conservationists and romantic preservationists. While conservationists were interested primarily in the wise use and management of natural resources, preservationists were interested primarily in protecting natural areas from the encroachment of industrialization (Hays 1959, Nash 1982). Although the conservation movement faded from attention with World War I, it left a legacy of organizations (e.g., the Sierra Club and National Audubon Society) and government agencies (e.g., the National Park Service and the Forest Service). Additionally, the movement's concerns were continued throughout the early to mid-twentieth century by projects such as the Tennessee Valley Authority and the Civilian Conservation Corps, and by environmental luminaries such as Bob Marshall and Aldo Leopold.

More recent explorations of the historical origins of environmentalism have begun to highlight additional, yet often neglected, precursors to the modern environmental movement (Taylor 1998). From the mid-1800s to the mid-1900s, middle- and working-class whites developed an environmental agenda that focused on the creation of urban parks and the protection of public and industrial worker health and safety. This urban agenda, was, however, not strongly identifiable as a cohesive social movement. Unlike the conservation movement, it lacked a strong infrastructure and membership base, had few ties to powerful business and government interests, and lacked the strong symbolism and charismatic leaders that were the hallmarks of the conservation movement. Similarly, despite the existence of activism by and for people of color during this time, such activities have been eclipsed, not only by traditional historical accounts of environmentalism, but also by the public's greater acknowledgement of conservationists and preservationists, both currently and historically.

The shift from its historical precursors to the modern environmental movement is typically marked by two events: the publication of Rachel Carson's *Silent Spring* in 1962 and the April 22, 1970 Earth Day celebration that drew millions of participants. The modern environmental movement emerged during an era of widespread political activism and reform, and quickly achieved high levels of support from the public, activists, and even elites—all of whom found environmental issues to be relatively appealing and consensual compared to the civil rights and anti-Vietnam protests.

The growth of scientific evidence on environmental degradation, eloquently documented in *Silent Spring*, coupled with media-enhanced disasters like the Santa Barbara oil spill of 1969, generated widespread concern. Post-World War II affluence enabled larger numbers of people to spend leisure time in the outdoors, heightening their commitment to preserving areas of natural beauty. Affluence, combined with increased urbanization and education, also stimulated changes in social values, lessening concern with materialism and generating interest in the quality of life, including environmental quality (Hays 1987).

At the same time, traditional conservation/preservation organizations, such as the Sierra Club and the Wilderness Society, were aggressively battling threats to natural areas, and broadening their agendas to incorporate a variety of issues, including those espoused by the earlier generation of urban/industrial activists (e.g., pollution and public health). Legal changes enabled organizations to fight battles on several fronts, from congressional offices to courtrooms, where legal standing was becoming easier to achieve. Furthermore, new organizations, such as the Environmental Defense Fund and the Natural Resources Defense Council, began to emerge, typically aided by foundation funding (Mitchell et al. 1992).

While the environmental movement is composed of numerous local and regional organizations, it is the large environmental organizations, such as those previously mentioned, that form the most visible face of contemporary environmentalism in the USA. Large environmental organizations have combined memberships numbering in the millions; they solicit and control multimillion-dollar budgets; and they employ large professional staffs of lobbyists, lawyers, and scientists. A core group of these organizations form what is known as the national 'environmental lobby,' a relatively influential, bureaucratic, and professionalized group of organizations headquartered in Washington, DC (Brulle 2000).

Despite predictions of movement demise and evidence of a decline in public support after the initial Earth Day, the movement remained strong throughout the 1970s, maintaining relatively high public support and continuing to gain in organizational membership. Ironically, the Reagan administration generated renewed support for environmentalism in the 1980s, as its anti-environmental agenda stimulated increased public support for environmental organizations (Dunlap 1995). The twentieth anniversary of Earth Day in 1990 witnessed the extensive involvement of Hollywood, politicians, and corporations. Although much of this reflected superficial commitment, the event was successful in stimulating major membership growth spurts in many environmental organizations (Mitchell et al. 1992). While growth in these organizations subsided somewhat in the 1990s, the thirtieth anniversary of Earth Day in 2000 yet again revealed the staying power of environmentalism. Environmental issues remain important to the public, the environment has become a significant factor in many political campaigns, and recycling programs and 'ecologically correct' products have become increasingly popular.

The environmental movement, like any social movement, is far from monolithic. Indeed, much of the

success garnered by the 'mainstream' organizations (i.e., the environmental lobby) has engendered criticism that such organizations have become ossified, overly bureaucratic, and co-opted by the political and business interests of the status quo. Such criticism has fueled the development of a 'radical fringe,' a group of organizations and activists that advocate direct action (e.g., spiking trees, destroying bulldozers) and espouse 'deep ecology,' a philosophy grounded in self-identification with nature (Devall 1992). With the larger organizations ostensibly focusing more on their 'bottom line' (memberships and budgets) and safely maintaining their ties to powerful interests, radical groups argue that the mainstream environmental movement has become too willing to compromise in its efforts to protect the environment.

While radical groups argue that the mainstream movement has lost sight of nature, other environmentalists believe that the mainstream has lost sight of people. Grassroots groups fighting locally unwanted land uses (LULUs), such as toxic waste dumps or hazardous waste incinerators, have argued that the large organizations ignore their concerns in favor of more fashionable national and global issues (e.g., rainforest destruction or loss of biodiversity). Environmental justice activists have argued that the mainstream organizations have ignored issues of social justice (Bullard 1993). Mainstream organizations are not only dominated by middle- and upper-class whites, but such organizations, it is argued, have ignored environmental issues faced by members of the working class and people of color. Because these grassroots and environmental justice groups focus heavily on public health issues that have a great impact on families and communities, women have played a more significant leadership role in them than has been true of the mainstream environmental movement.

As can be expected with any successful social movement, the environmental movement has generated opposition. Opposition to environmentalism comes from an incredibly diverse set of groups and activists, including well-organized industry groups, conservative think tanks, grassroots groups, and militant anti-environmentalists (Switzer 1997). Although industry groups have existed since the movement's beginnings, it is the grassroots groups that have recently become more numerous as well as more organized. Whereas grassroots opposition was typically confined to the rural west (the Sagebrush Rebellion being the key antecedent in the 1970s and early 1980s), it has increasingly become a nationwide phenomenon. While some grassroots anti-environmental activism is funded by industry, much of it nonetheless stems from genuine local concern over the loss of control over management and use of land and other natural resources. The constituents of these grassroots groups include farmers, ranchers, miners, motorized recreationists, militia supporters, landowners—virtually anyone who feels negatively affected by governmental attempts to protect natural resources. Despite the growing activism of opposition groups, however, it remains to be seen whether they can succeed at significantly thwarting the publicly accepted environmental agenda in the USA.

2. *Environmentalism Around the Globe*

Environmentalism has clearly become a significant force in the USA. However, the USA is not unique in this regard, nor has environmentalism been confined to the 'richer' countries of the world. Not only is public support for environmental protection found around the globe (Dunlap and Mertig 1995), but activism on behalf of environmental protection has emerged in many nations, both industrialized and nonindustrialized.

Like the USA, other industrial nations have relatively well-organized and successful national-level environmental movements. The countries of Western Europe, in particular, have not only developed their own professionalized and institutionalized environmental organizations (with a corresponding growth in radical groups critical of the large organizations), but many have also established relatively successful Green parties that attempt to advance social as well as ecological concerns (Dalton 1994, Rootes 1997). The European Union's emphasis on environmental protection is probably one sign that the environmental movement(s) within Europe have been quite successful in promoting their concerns.

In many countries, activism on behalf of environmental issues has a long history, even though it has typically not been labeled as environmentalism until relatively recently. This is perhaps most true of the poorer countries of the world, where colonization and imperialism have historically threatened not only the natural resource base but the very survival and livelihood of local peoples. For many people in poorer nations, such historical threats have not disappeared. Thus, it is no surprise to see activists in countries as varied as Malaysia, India, Thailand, Brazil, Nigeria, and Kenya fighting against international corporations to protect their forests and rivers, to promote sustainable development, and to safeguard biodiversity from the onslaught of transnational seed companies and other forces of globalization. Paralleling the movement for environmental justice in the USA, these activists are motivated by more than a concern for environmental protection; central among their concerns are those of social justice, the protection of livelihood, and the survival and well-being of family and community (Guha and Martinez-Alier 1997). This 'environmentalism of the poor' also counts women among its strongest leaders, for they are the ones who are most likely to see the devastating impact that resource shortages and environmental degradation can have on families.

Around the world, as in the USA, environmental activists and groups have also faced increasingly organized and sometimes violent opposition. From the bombing of a Greenpeace ship in New Zealand and the use of Strategic Lawsuits Against Public Participation (SLAPP) against anti-road activists in Britain, to the detention of anti-dam activists in Malaysia and the violent—occasionally murderous—attacks upon activists in Mexico, Costa Rica, Nigeria, and Brazil, there have been numerous instances of a growing global 'backlash' against environmentalism (Rowell 1996). While these events are certainly sobering for those involved in the environmental movement, most observers recognize that backlash reflects fear of a social movement's success. As environmentalism gains strength around the world, it will undoubtedly generate growing opposition.

3. *The Trajectory of Environmental Activism in the USA*

Despite the historical presence of other forms of environmental activism, especially those focused on early urban and industrial concerns, the progressive conservation movement is widely recognized as being the principal ancestor of modern environmentalism. Although modern environmentalism continues to encompass the concerns and tactics of the progressive conservation movement, it goes beyond them in substantial ways. Because the environmental movement is so qualitatively different from the progressive conservation movement, including both its preservationist and utilitarian branches, some scholars treat them as two separate movements, others as distinct stages of one broad movement. The predominant characteristics of what we identify as the three main stages of environmentalism in the USA—'conservationism,' 'environmentalism,' and the recently emerged 'ecologism'—are juxtaposed below. In this discussion, the three stages are treated as parts of one broad and continuously evolving movement aimed at protecting environmental quality.

Rather than divorcing itself from its predecessors, each stage in this broad trajectory incorporates the concerns and tactics it inherited from earlier stages into its own, expanded agenda. For instance, ecologism does not ignore issues of resource management, the hallmark of conservationism, nor does it reject tactics such as political lobbying. In fact, the evolution of the movement has been marked by a gradual broadening of issues and growth in the diversity of groups dealing with such issues (Brulle 2000). While diversity exists in most mature social movements, and has been present in conservationism/environmentalism since the early split between Gifford Pinchot's utilitarian conservationists and John Muir's preservationists, the modern movement encompasses an extraordinary diversity of organizations, goals, ideologies, and tactics. The scheme used here attempts to display the predominant characteristics of each stage, but it should be emphasized that substantial diversity exists at each stage, especially the third.

In Table 1, major characteristics of each stage of environmental activism are noted, starting with the approximate beginnings of each one. Conservationism dates to the late nineteenth century, when the progressive conservation movement emerged; environmentalism includes the bulk of the modern environmental movement that arose in the 1960s; and ecologism, it is argued, emerged in the last two decades of the twentieth century—partly in response to perceived weaknesses within mainstream environmentalism.

Social movement scholars distinguish movements (here, stages of one long-term movement) by their interrelated set of goals, ideologies, and world views. Movement goals are the primary focus of action for activists; ideology is a broad statement of what the movement considers to be an 'ideal' state of affairs; and world view is the lens through which activists perceive the world, in this case the natural world that they are attempting to protect. Of special relevance is the distinction between an anthropocentric world view, where humans are the center of concern, and an ecocentric (or biocentric) world view, where ecosystems, other species, and all life on Earth are deemed important.

Despite the efforts of preservationists such as John Muir, conservationism was predominantly anthropocentric because it emphasized the wise use of resources for human benefit. Similarly, environmentalism has also been largely anthropocentric in its emphasis on environmental quality as crucial for a high quality of *human* life. Concerns for human health, outdoor recreational opportunities, and an esthetically pleasing natural world were all motivated by an underlying concern with the welfare of humans—even though the natural environment often derived benefit from these concerns.

The recent emergence of ecologism, however, has broadened earlier concerns by incorporating an ecocentric world view stemming largely from two noteworthy branches of the modern movement: deep ecology and radical environmentalism (Devall 1992). Adherents of ecologism believe that nature has a right to exist in and of itself, apart from human desires. Although this position existed in the conservation movement via John Muir and his followers, and was given new impetus by Aldo Leopold's 'land ethic' in the mid-1900s, only recently has it become a potent voice in the evolving movement. Advocates of ecologism hold as their primary goal the maintenance of ecological sustainability, for the entire earth and all of its inhabitants (Thiele 1999).

Social movements are further distinguished by the types of problems or issues that they address (their 'grievances'). The issues addressed by environmental

Table 1
Three stages of environmental activism

	Conservationism	Environmentalism	Ecologism
Approximate beginnings	Late nineteenth century	Middle twentieth century	Late twentieth century
Primary goals	Conservation of natural resources	Protection of environmental quality	Maintenance of ecological sustainability
Dominant ideology	Natural resources should be used efficiently for the good of all society	Environmental quality should be protected for a high quality of life	Ecosystems should be protected for the benefit of all species
Worldview	Anthropocentric	Anthropocentric	Ecocentric
Nature of issues	Geographically bounded (typically rural), specific, unambiguous	Geographically dispersed (often urban); delayed, subtle and indirect effects; potentially harmful to human health	Extremely diverse; systemic and synergistic effects; potentially irreversible and harmful to all life on Earth
Example	Over-logging of a specific forest area	Urban air and water pollution	Global environmental change
Cost of solution	Relatively small, localized	Often substantial	Potentially infinite
Tactics	Lobbying	Lobbying, litigation, citizen participation	Lobbying, litigation, electoral action, direct action, lifestyle change
Opposition/ culprits	Natural resource industries; local economic interests (loggers, hunters)	Corporations, economic growth, modern lifestyles	Status quo; excess human production, consumption, and population

activists have expanded over time, as environmental problems and society's awareness of them have grown. Not only has each stage seen new issues incorporated into a growing environmental agenda, but the issues themselves have been perceived quite differently over time. As modern technology and lifestyles have become more complex, the environmental problems they create have likewise increased in scale and complexity. Costs of solving environmental problems have also grown, reflecting a rapid change in the magnitude of the problems and our perceptions of them.

Conservationism typically dealt with specific and unambiguous issues, such as the protection of particular forestlands from logging. Compare this to air and water pollution, a typical concern of environmentalism. Pollution can diffuse over a broad area, its sources are often ambiguous, and its effects may be delayed and difficult to identify. In short, pollution is more subtle than logging, and controlling or stopping it is typically more difficult and costly.

Ecologism incorporates an even wider range of issues, as well as a significantly different perception of these issues. Especially notable is a greater emphasis on both micro and macro concerns. At the micro level, grassroots groups around the country are mobilizing NIMBY ('Not-In-My-Backyard') protests against garbage incinerators and hazardous waste sites. Minorities in particular are mobilizing against 'environmental racism,' the practice of locating environmentally noxious facilities in their communities. At the macro level, ecologism is concerned with issues of international import, such as global warming, ozone destruction, and the loss of rainforests. In addition, the growing radical wing of the movement, mentioned earlier, promotes an ecocentric world view and an emphasis on global ecological sustainability. In fact, radical groups such as Earth First!, as well as many of the grassroots groups, have developed explicitly in reaction to what they refer to as the 'reform' environmentalism and 'shallow' ecology embodied by the mainstream national environmental organizations.

These relatively new wings of the movement, while having some antecedents, reflect a qualitatively different approach to environmental issues. In fact, the major difference between ecologism and its forerunners is not so much in the types of issues it looks at, but rather in *how* it looks at them. Ecologism focuses on a bigger picture than either environmentalism or conservationism. Older issues such as concern for human health, and even the welfare of future generations, are important, but ecologism is marked by concern for overall ecological sustainability and recognition of the inherent interdependence of all life systems (Thiele 1999). Even NIMBYites are becoming concerned with the sustainability of current production systems, as witnessed by the growing NIABY ('Not-In-Anyone's-Backyard') attitude. Ecologism thus broadens environmental concern in various ways, fighting on a greater number of fronts—from local to global—and viewing other species and ecosystems as

having rights to exist independent of human interests. Concern for equity between species parallels a growth in concern for equity within the human race and across generations, as advocates of sustainable development attempt to rectify social injustice and poverty as well as environmental devastation throughout the world.

Ecologism entails an expanded critique of the status quo, based on a systemic and large-scale view of human impact on the natural environment. The pivotal concerns of ecologism are typically those that involve long-term, irreversible, synergistic, and often unpredictable consequences of human actions. Global warming and loss of biodiversity, for instance, are not only long-term and irreversible results of human actions, but they stem from an incredibly complex interplay of factors. While environmentalism has often addressed these issues, it has looked at them in a piecemeal fashion. Ecologism, on the other hand, views them in the context of the larger ecological–evolutionary global system. Advocates of ecologism talk about the end of nature, mass extinction, and the halt of evolution—unless human practices are altered, and soon. The purported causes and consequences, as well as the costs needed to remedy them, are therefore truly colossal.

A key aspect of social movements is the type of tactics they utilize. Just as the issues have broadened, so have the tactics employed by environmental activists. Tactically, conservationists relied heavily and successfully on lobbying government officials (epitomized by Pinchot's influence with Teddy Roosevelt). The solution to resource problems, they felt, came with governmental and scientific management of resources. In addition to traditional lobbying, environmentalists added litigation, research, and citizen participation to their repertoire, through the development of research- and legal-oriented groups as well as via letter-writing campaigns and mass protests (Mitchell et al. 1992).

Ecologism builds upon this tactical legacy—grassroots activists march and petition government officials, organizations engage in litigation and lobbying—but it increasingly employs more aggressive tactics such as consumer boycotts and various forms of 'direct action.' The growing radical wing of the movement is especially likely to engage in sit-ins, 'monkey-wrenching' of equipment, and other forms of 'ecotage' (from pouring quick rice in the radiator of a bulldozer to ramming a drift-net ship on the high seas) that are disavowed by mainstream environmentalists intent on 'working within the system' in order to reform it.

Recent years have also seen a rapid growth in other forms of activism, such as electoral action. More and more organizations are becoming active in political campaigns, not only publicizing candidates' records but publicly supporting selected candidates. In addition, advocates of ecologism, like their predecessors, promote lifestyle change, ranging from recycling and purchasing 'green' products, to reduced consumption and dietary changes. The tactics have clearly broadened and it is likely that they will continue to do so.

Finally, social movements can be distinguished by their opposition and the source of their grievances. Usually these are related, for those who benefit from environmentally harmful practices are most likely to oppose attempts to halt those practices. As our understanding of ecological problems has progressed, so has recognition of their embeddedness in the status quo. No longer are just a few 'robber barons' to blame for our problems; rather, as Pogo said: 'We have met the enemy, and he is us.'

Conservationism laid the blame for resource depletion at the feet of a relatively small group of people, and thereby stimulated limited opposition. Environmentalism, in contrast, blamed entire industries, modern lifestyles, and 'growthmania' in general, and in the process engendered broader opposition. Ecologism issues a critique that leaves few unscathed, for our entire species and the status quo (at least within industrialized nations) are to blame, albeit some aspects more so than others. Simply reforming current practices via legislation is therefore unlikely to suffice. As a simple example, rather than favoring installation of scrubbers on factory smokestacks to reduce pollutants, advocates of ecologism argue for alternative production techniques or giving up the product completely. Because the economic and social costs of halting human-induced environmental change could prove enormous, and leave little of modern life untouched, ecologism has the potential to stimulate enormous opposition.

4. *The Future*

The recent emergence of ecologism reflects the persistent evolution of societal concern over the impact of humans on the Earth. Based on the staying power of environmentalism over the past few decades, it is likely that the current wave of ecologism will continue to be a vital force. A major reason for this optimism is the great diversity of ecologism. Just as biodiversity is vital to the stability of ecosystems, tactical and ideological diversity may be among ecologism's greatest strengths. Still, given that ecologism has arisen out of frustration over the failure of environmentalism to halt ecological degradation, and given the immensity of the tasks (and opposition) it faces, the future for this latest phase of human efforts to protect the environment is uncertain. Ultimately, the true test of a social movement is its success in achieving its goals, not simply in maintaining its own survival.

See also: Ecological Imperialism; Ecology and Health; Ecology, Deep; Ecology: Organizations; Environmental Conservation: Ethical Concerns; Environmental

Justice; Farmland Preservation; Green Parties; Locational Conflict (NIMBY); Nature, Concepts of: Environmental and Ecological; Political Protest and Civil Disobedience; Social Movements: Environmental Movements; Social Protests, History of; Sustainability Transition: Human–Environment Relationship

Bibliography

Brulle R J 2000 *Agency, Democracy and the Environment: The US Environmental Movement from the Perspective of Critical Theory*. MIT Press, Cambridge, MA

Bullard R D (ed.) 1993 *Confronting Environmental Racism: Voices from the Grassroots*. South End Press, Boston

Dalton R J 1994 *The Green Rainbow: Environmental Groups in Western Europe*. Yale University Press, New Haven, CT

Devall B 1992 Deep ecology and radical environmentalism. In: Dunlap R E, Mertig A G (eds.) *American Environmentalism: The US Environmental Movement, 1970–1990*. Taylor and Francis, Philadelphia, PA, pp. 51–62

Dunlap R E 1995 Public opinion and environmental policy. In: Lester J P (ed.) *Environmental Politics and Policy*, 2nd edn. Duke University Press, Durham, NC, pp. 63–113

Dunlap R E 2000 The environmental movement at 30. *The Polling Report* **16**(1): 6–8

Dunlap R E, Mertig A G 1995 Global concern for the environment: Is affluence a prerequisite? *Journal of Social Issues* **51**(4): 121–137

Guha R, Martinez-Alier J 1997 *Varieties of Environmentalism: Essays North and South*. Earthscan, London

Hays S P 1959 *Conservation and the Gospel of Efficiency: The Progressive Conservation Movement, 1890–1920*. Harvard University Press, Cambridge, MA

Hays S P 1987 *Beauty, Health and Permanence: Environmental Politics in the United States, 1955–1985*. Cambridge University Press, Cambridge, UK

Mitchell R C, Mertig A G, Dunlap R E 1992 Twenty years of environmental mobilization: Trends among national environmental organizations. In: Dunlap R E, Mertig A G (eds.) *American Environmentalism: The US Environmental Movement, 1970–1990*. Taylor and Francis, Philadelphia, pp. 11–26

Nash R 1982 *Wilderness and the American Mind*, 3rd edn. Yale University Press, New Haven, CT

Rootes C A 1997 Environmental movements and Green parties in Western and Eastern Europe. In: Redclift M, Woodgate G (eds.) *The International Handbook of Environmental Sociology*. Edward Elgar, Cheltenham, UK, pp. 319–48

Rowell A 1996 *Green Backlash: Global Subversion of the Environmental Movement*. Routledge, London

Switzer J V 1997 *Green Backlash: The History and Politics of Environmental Opposition in the US*. Lynne Reinner, Boulder, CO

Taylor D E 1998 The urban environment: The intersection of white middle-class and white working-class environmentalism (1820–1950s). *Advances in Human Ecology* **7**: 207–92

Thiele L P 1999 *Environmentalism for a New Millennium: The Challenge of Coevolution*. Oxford University Press, New York

A. G. Mertig and R. E. Dunlap

Environments for Education

It is agreed generally that students' learning will be enhanced when family and school environments are supportive of education and in harmony with each other. In this article studies are examined that have investigated relationships between family and school educational environments and measures of students' school-related outcomes. Studies are organized into four general categories that are labeled as social-arithmetic environmental research, family and school structural characteristics, environmental press of families and schools, and interpretive analyses.

1. Social-arithmetic Environmental Research

One of the most significant sociological traditions in studies of education relates to attempts to describe empirically the social and educational conditions of society. Typically, the method that has been labeled as political arithmetic is used to explore associations between social and educational inequalities with the intention of possible educational reform by political means. A recent variant of political arithmetic has been developed in which statistical approaches such as structural equation modeling, logistic regression, and multilevel analyses are used to examine individual, social, and school influences on students' eventual educational and occupational attainments. Much of the research has lost, however, the political emphasis of earlier investigations. It might be more appropriate to classify the research as reflecting social rather than political arithmetic.

Social-arithmetic studies have often used adaptations of what has become known as the 'Wisconsin' model of status attainment. Typically, the research has shown that relationships between family background and young adults' status attainment are mediated by perceptions of parental and peer influences and by adolescents' aspirations. In addition, children's ability is related to successive measures of academic performance, independent of the effects of family social status.

A particularly interesting set of social-arithmetic studies has examined the longitudinal influences of children's early experiences on the educational outcomes of a group of US children in Baltimore. The investigations have shown that the quality of interactions with parents and teachers in early elementary school years is associated with children's reading and mathematics achievement 4 to 9 years later. In an attempt to disentangle the impact of family and school educational environments on outcomes, the Baltimore research has examined the differences between winter and summer learning. Winter learning refers to children's outcomes while school is in session and it reflects the influences of both home and school. In

contrast, summer learning that occurs during the long vacation is assumed to be related primarily to family influences. The research shows that for mathematics achievement there are consistent losses for poorer children during the summer vacation. It is suggested that schools are most beneficial for those children who need them most. The longitudinal investigations conclude that while families guide children's development throughout schooling, the quality of experiences in early school grades have a profound and continuing impact on children's eventual educational careers (Alexander et al. 1997).

The development of multilevel statistical models that recognize that students are members of classes, classes are nested within schools, and that schools are part of wider social structures such as neighborhoods, states, and countries, has allowed social-arithmetic research to examine the effectiveness of schools. A meta-analysis, using data from many countries, has shown that classroom conditions have a greater association with outcomes than do school organizational characteristics. In addition, the analysis indicates that across countries, students' family environment has a much stronger impact on academic outcomes than do the malleable features of schooling (Scheerens and Bosker 1997).

School effectiveness research in Britain has tended to show that it is possible to discriminate significantly between schools only when extremely successful and extremely unsuccessful schools are compared. From the British experience it is claimed that because of the limitations associated with present statistical modeling techniques, caution should be expressed when output league tables are used by governments and others to rank schools, and that parents need to be informed of the limitations of the research that generated the league tables (Goldstein and Spiegelhalter 1996). Similarly, an Australian study that combined multilevel analyses and structural equation modeling in a two-stage process concluded that because of modeling limitations, claims about educational effectiveness are at best tentative and at most spurious. It is suggested that future school effectiveness research should go into classrooms and examine those factors which optimize the quality of teacher–student interactions (Rowe and Hill 1998).

Although there have been substantial advances in the sophistication of statistical modeling in social-arithmetic research, there are concerns that such developments have not been matched by similar advances in the quality of the data that are being collected and in the elegance of theorizing. A summary of social-arithmetic research concludes that two melancholy propositions have been supported with increasing certainty:

(a) That class inequality is stubbornly resistant to social change when properly conceived in terms of relative rather than absolute improvements and equalization of life chances, and

(b) That educational reform, including reform of secondary schooling... has similarly failed to modify relative class chances though there may have been significant improvements in some countries with respect to the relative chances of women and some ethnic groups (Halsey 1994, p. 443).

2. *Family and School Structural Characteristics*

Family and school educational environments may be considered to consist of sets of structural constraints that provide opportunities for many children while limiting the choices of other children. In this section some of the structural characteristics that define children's educational environments are examined.

2.1 *Sibling Structure Variables*

Sibling variables such as sibship size (the number of children in families) and the birth order of children within families have often been used as indicators of children's family educational environments. Typically, these family structural characteristics are shown to have modest inverse associations with children's cognitive performance, especially verbal measures of achievement. A number of conceptual orientations have been presented to explain these relationships including the resource dilution hypothesis and the confluence model.

The resource dilution hypothesis indicates that sibling variables are related inversely to the cultural and material resources that parents are able to provide for their children. It is then proposed that such variations in the amount of family educational resources provided for different siblings are related to differences in children's educational outcomes. The confluence model proposes that children's cognitive development is related to the intellectual quality of family environments where quality is determined by influences such as sibship size, the age spacing among siblings, and whether individuals are only or last-born children (Zajone 1976). In the model, family intellectual environment is related to the average intellectual ability of all family members, including the child who is being studied. A family's intellectual environment is never stable, with changes being produced by alterations in family structure. The confluence model has generated vigorous controversy. It has been suggested, for example, that it may be a theory which is attempting to explain a social phenomenon that does not really exist (Retherford and Sewell 1991).

Much recent sibling structure research has examined whether the associations between sibling variables and educational outcomes are causal or attributable to other correlated factors such as socioeconomic status (Downey et al. 1999). In an analysis using change

models, for example, it has been shown that sibship size effects on cognitive performance disappear when family educational environments are taken into account (Guo and Van Wey 1999).

2.2 Single-parent Families

A number of theoretical orientations have been proposed to explain why the educational outcomes of children in single-parent households appear to be less favorable than the outcomes of children in two-parent families. The no-impact hypothesis suggests that the association between changing family structures and school outcomes can be attributed to a combination of other social contexts such as socioeconomic status, family ethnicity/race, and changing family economic circumstances. In the family socialization orientation it is suggested that the absence of a parent decreases the total amount of parental involvement in children's education that is related to poorer school outcomes. An instability, change, and stress perspective indicates that if children have to cope with the stress and instability that may accompany divorce, separation, or the death of a parent, then their school performance is likely to be affected negatively.

Although the findings from investigations that have examined the educational outcomes of children from single-parent families remain equivocal, the research is beginning to address a number of limitations of earlier studies. Investigations, for example, are beginning to (a) discriminate among different types of single-parent families such as divorced, separated, widowed, or never married, (b) include various social status measures as moderating variables, (c) examine the influence of changing family income as a background and a mediating variable, (d) separate the socioeconomic and ethnicity/race of families, and (e) differentiate between the effects on children's outcomes of families that are headed by fathers or mothers.

2.3 Ability Grouping in Schools

Perhaps the school structural characteristic that has been examined most widely has been ability grouping and whether tracking or streaming is related to students' academic outcomes. Some investigations have supported the divergence hypothesis of grouping which states that relative to what would have occurred if students had not been placed in ability groups, those students in high-ability groups gain in performance while those in low-ability groups lose in performance. However, findings from the research on ability grouping remain inconclusive. It is suggested that one reason for the inconsistent results might relate to how instructional resources are distributed within schools. If decisions about the allocation of teaching resources are associated with student placement, then the impact of ability grouping on outcomes might be primarily a function of the distribution of the quality of instruction. In addition, research has indicated that better educated parents can influence the tracks into which their children are placed, often independent of children's actual performance. The equivocal nature of the research suggests the need for further investigations of how ability groups are formed and how instructional resources are assigned to those groups; how families influence the assignment of children to ability groups; what the effects of ability grouping are on educational productivity and inequality; and what the effects of the alternatives to ability grouping are on learning outcomes.

3. Environmental Press of Families and Schools

In the development of a theory of personality, Murray (1938) suggests that to understand individuals' behavior it is necessary to define environments by the kinds of benefits and harms that they provide. The directional tendency implied in Murray's framework is designated as the press of the environment. An alpha press is defined as the benefit or harm that actually exists, as far as research can assess it, while a beta press is an individual's perception of the environment. In the following section the press of family and school educational environments are considered.

3.1 The Press of Family Environments

It was not until Bloom (1964) and a number of his students examined the environmental correlates of children's affective and cognitive outcomes that a 'school' of research emerged to assess the alpha press of families. Bloom defined environments as the conditions, forces, and external stimuli that surround, engulf, and impinge on individuals. These environmental characteristics may be physical and social as well as intellectual. It is proposed that such press variables form subenvironments with each subenvironment being related to the development of a particular human characteristic. A number of alterable alpha press variables have been identified as being consistently associated with school outcomes. These variables include family work habits, parents' academic guidance and support, language environment in the home, family stimulation to explore and discuss ideas and events, and parents' aspirations for their children (Marjoribanks 1994).

While alpha press studies have typically examined between-family environmental factors, there is a need to explore with greater sensitivity possible within-family influences. These latter family factors are often not shared by siblings such as the differential aspirations that parents have for their children or the experiences a child has with a chronic illness. The critical question becomes, why are children in the same

family often so different from one another? To answer this question perhaps interpretive analyses of families need to be undertaken in conjunction with alpha press investigations.

3.2 The Press of School Environments

There is now a large body of research that has defined classroom environments as the shared perceptions of students and sometimes of teachers. The schedules that have been constructed to assess the beta press of classroom and school environments are often referred to as high-inference measures. In contrast, low-inference techniques measure specific explicit phenomena such as the number of questions asked by students in a certain section of a lesson. Typically, perceptual measures are related to moderate amounts of variance in learning outcomes, often beyond the variance that can be attributed to student characteristics such as pretest performance or general ability. The practical implication of the research is that student outcomes might be improved by establishing classrooms that match those educational environments which have been shown to be associated with students' learning.

A limitation of most classroom learning environment instruments is that they measure an individual student's perceptions of a whole class, as distinct from students' perceptions of their own roles in the classroom. There is a distinction between the private beta press or idiosyncratic perception that each person has of an environment and the consensual beta press that the members of a group share about that environment (Fraser 1998). It is likely that future classroom and school environment research will be enhanced if personal as well as group forms of beta press schedules are adopted.

Many educators would argue that school learning environments need to be examined with a greater sensitivity than can be generated from perception measures. Researchers using such perceptual scales generally suggest themselves that the measures provide one portrayal of educational environments that is likely to be enriched by adopting other measurement measures. In the following section interpretive approaches to the investigation of educational environments are examined.

4. Interpretive Analyses

Increasingly in research related to family and school educational environments, concepts, and methodologies are being adopted from a number of theoretical orientations such as social phenomenology, ethnomethodology, symbolic interactionism, and critical discourse analysis. In educational environments, research of these interpretive perspectives has emphasized the need to examine the processes used by members of families and schools to define and manage their everyday lives. Investigations typically use variations of ethnographic methods to obtain accounts of why parents, teachers, and children perform certain acts and what social meanings they give to the actions of themselves and of others.

In an intensive study of parents' involvement with their third grade children, for example, Lareau and Horvat (1999) examined how inequality is perpetuated in school settings. They were concerned that previous research has not investigated adequately how family, cultural, and social resources are converted into educational advantages. From indepth interviews with the parents of 24 children, with staff in the children's school, and with members of the wider community, the study concluded that while a child's racial and social class are associated with social reproduction they do not determine it. In addition, the study indicates that those parents who have the capacity to activate intellectual capital are likely to have a substantial impact on their children's eventual status attainment. In an even more intensive ethnographic study, Goldenberg (1989) investigated the relationships among parental involvement, ability group placement, and the reading achievement of three first-graders. The findings indicated that teachers' perceptions of children's behaviors were related to the placement of children into academic groups. As a result, parents were able to affect group placement by influencing their child's classroom behavior.

Such interpretive analyses provide valuable insights into the relationships among family and school educational environments and children's school outcomes.

5. Future Research Directions

The complexity of the relationships among family and school educational environments and students' school outcomes indicates the difficult task that confronts parents and teachers when they attempt to design and implement programs to improve educational outcomes. Parents and teachers may create what they consider to be supportive and harmonious educational environments but children's perceptions and the meanings they impute to those environments may be influenced greatly by the social, cultural, and economic constraints that surround them. Investigations are required that integrate the most recent advances in the research orientations that have been considered in this article to examine how the cultural, economic, and social capital of families and schools are associated with the educational outcomes of boys and girls from different social status and ethnic/race groups. Unless such inclusive studies are undertaken it is unlikely that there will be significant advances in understanding how families and schools interact to influence students' learning.

See also: Educational Assessment: Major Developments; Environments for Learning; Families as Edu-

cational Settings; Family and Schooling; Learning and Instruction: Social-cognitive Perspectives; Parents and Teachers as Partners in Education; Socialization and Education: Theoretical Perspectives

Bibliography

Alexander K I, Entwisle D R, Horsey C S 1997 From first grade forward: Early foundations of high school dropout. *Sociology of Education* **70**: 87–107

Bloom B S 1964 *Stability and Change in Human Characteristics*. Wiley, New York

Downey D B, Powell B, Steelman I C, Pribesh S 1999 Much ado about siblings: Change models, sibship size, and intellectual development. *American Sociological Review* **64**: 193–98

Fraser B J 1998 Classroom environment instruments: Development, validity and applications. *Learning Environments Research* **1**: 7–33

Goldenberg C N 1989 Parents' effects on academic grouping for reading: Three case studies. *American Educational Research Journal* **26**: 329–52

Goldstein J L, Spiegelhalter D J 1996 League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society* (*A*) **159**: 385–409

Guo G, Van Wey I K 1999 Sibship size and intellectual development: Is the relationship causal? *American Sociological Review* **64**: 169–87

Halsey A H 1994 Sociology as political arithmetic. *British Journal of Sociology* **45**: 427–44

Lareau A, Horvat E M 1999 Moments of social inclusion and exclusion: Race, class, and cultural capital in family-school relationships. *Sociology of Education* **72**: 37–53

Marjoribanks K 1994 Families, schools and children's learning. *International Journal of Educational Research* **21**: 439–555

Murray H 1938 *Explorations in Personality*. Oxford University Press, Oxford, UK

Retherford R D, Sewell W H 1991 Birth order and intelligence: Further tests of the confluence model. *American Sociological Review* **56**: 141–58

Rowe K J, Hill P W 1998 Modeling educational effectiveness in classrooms: The use of multi-level structural equations to model students' progress. *Educational Research and Evaluation* **4**: 307–47

Scheerens J, Bosker R J 1997 *The Foundations of Educational Effectiveness*, 1st edn. Pergamon, Oxford, UK

Zajone R B 1976 Birth configuration and intelligence. *Science* **192**: 227–9

K. Marjoribanks

Environments for Learning

The term 'learning environment' expresses that learning is dependent on various environmental factors, which are created to various degrees by external factors. A learning environment is made up of an arrangement of teaching strategies and methods, learning materials, and media. The learning environment represents the current temporal, spatial, and social learning situation and also includes the relevant cultural context. The basis for concrete measures to create learning environments provides a fundamental concept for teaching and learning. In the following, two central (purist) concepts are compared with each other and are then confronted with a pragmatic concept. In this manner, a framework for orientation is provided with regard to the design and assessment of learning environments.

1. The Cognitivist-based Concept: The Design of Object-oriented Learning Environments

1.1 Basic Assumptions

In a cognitivist position on teaching and learning, all efforts are devoted to instruction and to the question, how teaching and learning are planned, organized, and guided. In this way, the learners understand and internalize the contents of knowledge according to their system and succeed in learning based on predefined teaching and learning goals. In the design of learning environments, one concentrates on the object of teaching and learning, which as a rule is interpreted as a 'pre-fabricated knowledge system.' The teaching/learning process is therefore viewed as a type of 'knowledge transfer.' The teacher's function is to present and explain the contents of knowledge and also to guide the learner and to ensure that advances in learning are made. The learner remains in more of a passive position (Lowyck 1991). Under these conditions learning is to a large degree a receptive process. In object-oriented learning environments, great emphasis is placed on the assessment of how successful learning is. Norm-oriented and/or learning goal-oriented tests evaluate to which learning results the instructional methods have led. Instruction and evaluation are, in this case, viewed as two separate units.

1.2 Problems of the Cognitivist-based Teaching and Learning Concept

There is still a lack of empirical findings that justify a rational structuring of the learning and teaching processes, as it is postulated in the cognitively oriented teaching and learning models and in the discipline of ID. Especially lacking is the evidence that the effects of single instructional measures are replicable. Theoretically problematic is the instruction's analytical process of separating entireties into elementary parts that are transferred separately. This reductionist procedure ignores the fact that understanding is dependent on the

entire knowledge structure and not on its isolated parts. Contestable also is the instructional methods assumption that the effect of different methods can be exactly predicted (Duffy and Jonassen 1991). Most apparent, however, are the practical problems of the cognitivist view with its uneven role separation between teachers and learners, which results in many cases in receptive behavior from the learner. The resulting shortage of self-initiative and self-responsibility for the learning process and its success increases the possibility that the learners find themselves in a passive position and feel demotivated, or at best extrinsically motivated (Deci and Ryan 1993). In addition, systematized and logically prepared knowledge has little to do with the complex and unstructured demands of everyday situations (Resnick 1987). In this way cognitive-oriented instruction often produces 'tacit knowledge,' which has been gained but which cannot be used in real situations.

2. *The Constructive-oriented View: Design of Situational Learning Environments*

2.1 *Basic Assumptions*

Constructivism has been discussed intensively again in educational psychology since the 1980s, and specifically constructivism as a scientific and cognition theory and also as a paradigm in sociology, cognition science, and psychology ('new' constructivism) (see Gerstenmaier and Mandl 1995). In questions of teaching and learning the 'new' constructivism becomes especially relevant: in contrast to radical constructivism, the new constructivism is more concerned with the thinking and acting subject. Central to this is the assumption that knowledge is not a copy of reality, but rather a human construction (Knuth and Cunningham 1993). While teaching and learning methods in the tradition of the cognitive view hold the instruction at the center of their efforts, the views of the constructivists distinguish themselves by placing the learner and the learning process at the center. The question of how knowledge is constructed and the connection between knowledge and action is more important than how knowledge is transferred (Gerstenmaier and Mandl 1995). The learner finds him/herself in an active position, while the teacher has the task of providing problem situations and the tools to solve them and, if necessary, is available to react to the needs of the learner. In the constructivist position, learning is viewed as an active constructing process, which is situated in a definite context. Evaluation is an integral part of the constructivist view, in which processes are more important than results; however, there are no predefined evaluation methods, which are theoretically derived and empirically 'secured.'

2.2 *The Situated Cognition Movement*

From a constructivist perspective, the learning environment must provide the learner with situations in which their own constructive achievements, social interactions, and participation processes are possible, which is why these are referred to as 'situated learning environments.' It is the goal of situated learning environments that the learner understands the new contents and is able to use the gained knowledge and skills in a flexible manner. In addition, the learner should be able to develop problem-solving skills and other cognitive strategies. Above all, it is due to the situated cognition movement that the ideas of context and social participation in authentic situations have found widespread use in the learning process. The term situated cognition, in a similar fashion to the term constructivism, is not clearly defined and there are several theoretical variations. Common to both is the view that thinking, acting, and learning take place situally and that knowledge is on the one hand constructed by the perceiving subject and on the other hand is shared amongst a community (Resnick 1991). Concepts within the situated cognition movement stem from various influences, which range from cognitive anthropology (Lave 1991) to ecological psychology (Greeno 1989) to socio-cognitive schools of thought. Since the end of the 1980s various approaches have been developed in instructional psychology which are related to the Situated Cognition movement, for example approaches like Anchored Instruction (CTGV 1997), Cognitive Flexibility (Jacobson and Spiro 1992) or Cognitive Apprenticeship (Collins et al. 1989). The commonality exists in the fact that the learning processes are embedded within solving authentic problems and that learning is not restricted to the aspect of knowledge acquisition, but rather is viewed under the perspective of enculturation.

2.3 *Problems in Constructivist-oriented Teaching and Learning Concepts*

The constructivist view runs into difficulties when the question is raised regarding the empirical underpinning of the postulated effects of situated learning environments. Although the number of studies has recently increased, the findings are far from sufficient. In addition, several findings of transfer research contradict the constructivist postulate that knowledge is bound to context (Anderson et al. 1996). There is the problem that the promotion of learning processes through constructivist principles has led in subsequent knowledge tests to poorer results. The meaning of these findings is first qualified, when these results are compared to the results of later applied tests of the learner, which show a positive influence of

situated learning. Problems of the theoretical kind arise when a more radical constructivist perspective is taken, which states that there is no objective reality those teachers can pass along to their students. Many supporters of situated learning environments, however, take a less extreme view and assume that, aside from the individual construction capacities of the learner, it is possible to stimulate learning processes externally (Lowyck 1991). Even in the framework of this moderate constructivist position, teachers are given such large degrees of freedom that they have problems in making the right choice. It must also be mentioned that central concepts such as activity and situation are conceptually vague and used in a heterogeneous manner (Renkl et al. 1999). The supporters of the constructivist view must also accept the practical criticism that their approach for both teachers and learners may be very time-consuming and may result in a poor cost-benefit relationship (Anderson et al. 1996). The lack of direction and support for the learner in situational learning environments can also lead to undesirable results; one example could be that the demands placed on the students are too high. This is especially the case with learners who have disadvantageous learning conditions, so that the danger of a 'scissor effect' arises: as a rule, more competitive (motivated) learners profit from situated learning environments than less competitive learners. The goal of situated learning, namely the enculturation and acquisition of expertise, cannot be sought after in many contexts. For example where orientation and a lower achievement level are sufficient, proven cognitivist ways are practicable and from time to time more economical (Winn 1996).

3. *The Attempt at a Pragmatic View: Structure of Problem-oriented Learning Environments*

3.1 *Basic Assumptions*

The use of theoretical teaching and learning models in practice is difficult in many respects, regardless of whether they are models for creating object-oriented or situated learning environments. The practical problems of both the cognitivist and the constructivist position have made it clear that theoretical one-sidedness is problematic when applied to practice. The lack of clear empirical finds from which action-guiding rules for or against a definite form of teaching and learning makes the situation even more confusing. Against this background the attempt is increasingly being made to take a middle position between the cognitivist and the constructivist views and to find a balance between the primacy of instruction in object-oriented learning environments and the primacy of construction in situational learning environments. To a large degree, consensus has been reached supporting a relative constructivist conception of learning as an active, self-guided, constructive, situative, and social process (Reinmann-Rothmeier and Mandl 1997). Knowledge-based constructivism is also currently under discussion, in which learning is based on the personal construction of meaning from the perspective of the learner, but also presupposes sufficient knowledge, whose acquisition is barely possible without instructional guidance and support (Resnick and Williams Hall 1998).

3.2 *Problem-oriented Learning Environments*

A moderate constructivist view of learning does not in any way exclude guidance and support from teachers. It is neither possible nor sensible to solely rely on the constructive ability of learners. One cannot constantly transfer prefabricated knowledge systems to learners according to fixed rules. In various attempts to integrate object-oriented and situated learning environments and thereby to utilize the characteristics of both approaches, problem-oriented learning environments are often postulated. With a desideratum of characteristics, which especially the Anchored Instruction view, the Cognitive-Flexibility theory, and the Cognitive Apprenticeship view have in common, some guidelines from problem-oriented structuring of learning environments can be formulated, which also can be combined with instructional measures of cognitivist character. (e.g., CTGV 1997, Reinmann-Rothmeier and Mandl 1997):

(a) *Learning in a situated manner with authentic problems*: The starting point of learning processes should be authentic problems—problems which due to their relationship to reality and their relevance motivate the learner to develop new knowledge or new skills. Problem-oriented learning environments are therefore designed in such a manner that they make possible and encourage dealings with realistic problems and authentic situations.

(b) *Learning in multiple contexts*: To ensure that newly gained knowledge or concepts do not remain fixed to a certain situation, the same contents should be learned in different contexts. Problem-oriented learning environments are therefore designed in such a manner that the learned material can also be applied to and utilized in the analysis of other problems.

(c) *Learning using multiple perspectives*: In learning it should be considered that single contents or problems could be examined from different perspectives or aspects. Problem-oriented learning environments are therefore designed in such a manner that knowledge and concepts can be learned and applied using multiple perspectives.

(d) *Learning in a social context*: Learning must not be done as an individual process, but rather group learning and group tasks should be used in as many

learning phases as possible in the framework of situated problem analysis. Problem-oriented learning environments are therefore designed in such a manner that they make it possible to have cooperative learning and problem solving in groups.

(e) *Learning with instructional support*: Learning without any instructional support is ineffective and causes excessive demands on the learners. Teachers cannot limit themselves to giving opportunities to learn; they must also guide learners whenever necessary and especially give specific guidance for problem solving. Problem-oriented learning environments are therefore structured in a manner that provides both diverse possibilities for learning in complex situations and the knowledge necessary to work on problem solving.

It is also important to explicitly consider motivational factors that are generally lacking in the structuring of cogvitivist learning environments. Motivational factors are implicitly used in situational learning environments but conceptually receive rather little attention (Stark et al. 1998).

3.3 Empirical Arguments for Problem-oriented Learning Environments

Considerations of the compatibility of instruction and construction show a certain relation to the so-called 'adaptive teaching,' which strives towards an optimal fit between the internal need for the support of the learner and the external offer of support. The classic ATI (Aptitude Treatment Interaction) research (Cronbach and Snow 1977) shows an interaction between individual learning conditions and the teaching methods employed. It showed in numerous cases that learners with less advantageous learning conditions profited from a highly structured learning situation, but learners in advantageous learning conditions did not. Teacher-centered forms of instruction in the sense of the cognitivist view are more effective in disadvantageous learning conditions (i.e., little previous knowledge, lower intelligence, high level of anxiety). Situated learning environments, which place high demands on the capability of the learner to construct knowledge, appear to be more effective for learners with advantageous learning conditions. The strongest predictor for subsequent learning results has, in recent studies, proven to be the domain-specific preknowledge. The call for the combination of instructional principles from cognitivist and constructivist teaching and learning approaches has also gained empirical support from other fields. In initial business school education, for example, the combination of multiple perspectives (in the form of multiple examples and instructional support (in the form of requests for self-explanation)) has proven effective. A further example is case-oriented computer learning programs in medicine, which support learning in authentic problem situations with directions and expert commentaries as support (Gräsel 1997). In addition, there have been further developments in the Anchored Instruction approach, in which narrative anchors have been enhanced with teacher-supported learning cycles (CTGV 1997). Positive results have also been shown in attempts to implement problem-oriented teaching and learning forms within conventionally structured (mass-) universities and in this way to combine them in a logical way with new, constructivist-inspired models of teaching and learning.

See also: Educational Learning Theory; Environments for Education; Instructional Design; Learning Theories and Educational Paradigms; Situated Learning: Out of School and in the Classroom; Teaching and Learning in the Classroom

Bibliography

Anderson J R, Reder L M, Simon H A 1996 Situated learning and education. *Educational Researcher* **25**(4): 5–11

Collins A, Brown J S, Newman S E 1989 Cognitive apprenticeship: Teaching the crafts of reading, writing and mathematics. In: Resnick L B (ed.) *Knowing, Learning and Instruction. Essays in the Honour of Robert Glaser*. Erlbaum, Hillsdale, NJ, pp. 453–494

Cronbach L J, Snow R E 1977 *Aptitudes and Instructional Methods*. Irwington, New York

CTGV (Cognition & Technology Group at Vanderbilt) 1997 *The Jasper Project: Lessons in Curriculum, Instruction, Assessment, and Professional Development*. Erlbaum, Mahwah, NJ

Deci E L, Ryan R M 1993 Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik* **39**: 223–38

Duffy T M, Jonassen D H 1991 Constructivism: New implications for instructional technology? *Educational Technology* **31**(5): 7–12

Gerstenmaier J, Mandl H 1995 Wissenserwerb unter konstruktivistischer Perspektive. *Zeitschrift für Pädagogik* **41**: 867–88

Gräsel C 1997 *Problemorientiertes Lernen. Strategieanwendung und Gestaltungsmöglichkeiten*. Hogrefe, Göttingen, Germany

Greeno J G F 1989 Situations, mental models and generative knowledge. In: Klahr D, Kotovsky K (eds.) *Complex Information Processing: The Impact of Herbert A. Simon*, Erlbaum, Hillsdale, NJ, pp. 285–318

Jacobson M J, Spiro R J 1992 Hypertext learning environments and cognitive flexibility: Characteristics promoting the transfer of complex knowledge. In: Birnbaum L (ed.) *The International Conference on the Learning Sciences. Proceedings of the 1991 Conference*, Association for the Advancement of Computing in Education, Charlottesville, VA, pp. 240–248

Knuth R A, Cunningham D J 1993 Tools for constructivism. In: Duffy T M, Lowyk J, Jonassen D H (eds.) *Designing Environments for Constructive Learning, NATO ASI Series*. Springer, Berlin, pp. 163–188

Lave J 1991 Situated learning in communities of practice. In: Resnick L B, Levine J M, Teasley S D (eds.) *Perspectives on Socially-shared Cognition*. American Psychological Association, Washington, DC, pp. 63–83

Lowyck J 1991 The field of instructional design. In: Lowick J, DePotter P, Elen J (eds.) *Instructional Design: Implementation Issue, Proceedings of the IBM/VU Leuven Conference*. La Hulpe, Belgium, pp. 1–10

Reinmann-Rothmeier G, Mandl H 1997 Lehren im Erwachsenenalter. Auffassungen vom Lehren und Lernen, Prinzipien und Methoden. In: Weinert F E, Mandl H (eds.) *Enzyklopädie der Psychologie D/I/4. Psychologie der Erwachsenenbildung*. Hogrefe, Göttingen, Germany, pp. 355–403

Renkl A, Gruber H, Mandl H 1999 Situated learning in instructional settings: From euphoria to feasibility. In: Bliss J, Säljö R, Light P (eds.) *Learning Sites*. Pergamon, Amsterdam, pp. 101–9

Resnick L B 1987 Learning in school and out. *Educational Researcher* **16**(9): 13–20

Resnick L B 1991 Shared cognition: Thinking as social practice. In: Resnick L B, Levione J M, Teasley S D (eds.) *Perspectives on Socially Shared Cognition*. American Psychological Association, Washington, DC, pp. 1–20

Resnick L B, Williams Hall M 1998 Learning organizations for sustainable education reform. *Daedalus* **127**(4): 89–118

Stark R, Gruber H, Mandl H 1998 Motivationale und kognitive Passungsprobleme beim komplexen situierten Lernen. *Psychologie in Erziehung und Unterricht* **44**: 202–15

Winn W 1996 Instructional design and situated learning: Paradox or partnership. In: McLellan H (ed.) *Situated Learning Perspectives*. Educational Technology Publications, Englewood Cliffs, NJ, pp. 57–66

H. Mandl and G. Reinmann-Rothmeier

Epidemiology, Social

Epidemiology is the basic science of public health: it focuses on the distribution of disease in the population and on the factors that explain that distribution. Social epidemiology is a subfield of epidemiology that concentrates on social factors as explanatory variables. There is increasing interest in the importance of social factors both for understanding the determinants of disease in the population and for the development of programs to prevent disease and promote health.

1. *Rationale for a Social Epidemiologic Approach to Disease Prevention*

1.1 *Magnitude of the Disease Problem*

Consideration of social factors in efforts to prevent disease is suggested by the sheer magnitude of many diseases (including coronary heart disease, cancer of various sites, arthritis, mental illness, diabetes, and stroke). While a one-to-one approach to diseases of great prevalence clearly is of value to patients, families, and friends, it does little to alter the distribution of disease in the population because new people develop disease even as sick people are treated and cured. Thus, an individual approach exhausts substantial medical care resources but does little to address those social and environmental factors that have initiated the problem. In this circumstance, an environmental approach to prevention is clearly more efficient.

1.2 *The Patterning of Disease Rates*

The continual swell of new 'at-risk' people in the population leads to the second reason for considering social factors in disease prevention research and programs: groups often have a characteristic pattern of disease over time even though individuals come and go from these groups. This pattern indicates that there may be something about groups that either promotes or discourages disease among individuals in those groups. One example of patterning of disease rates is provided by the case of coronary heart disease (CHD). CHD rates are far higher in the US than in Japan.Yet when men from Japan immigrate to the US, their CHD rates are five times higher compared to those in their country of origin. This increase in rate is observed even after account is taken of such CHD risk factors as blood pressure, cigarette smoking, diet, and serum cholesterol levels. It seems reasonable to suggest that something about the environment of the US causes people who settle here to slowly adopt the endemic rates of disease in the US (Marmot and Syme 1976). This example of patterned consistency of disease rates across geography, social, and cultural groups emphasizes the potential importance of environmental factors in the study of disease etiology. In consequence, a considerable amount of research has been done in hopes of identifying social and environmental factors so that interventions might be developed to prevent or control disease.

1.3 *Difficulties in Changing Behavior*

A third reason for the consideration of social factors in disease prevention is that the prevention of many diseases requires that people change their behavior. To prevent these diseases, people are increasingly asked to begin to do things they have not done previously, to stop doing things they have been doing for years, and to do more of some things and less of others. This behavioral approach to disease prevention contrasts with programs that attempt to do something 'to' people. While some diseases and conditions can be treated best by injection, surgery, or other nonbehavioral manipulations, most diseases of concern today cannot. Diseases such as cancer, diseases of the cardiovascular system, cirrhosis, and some respiratory diseases to a greater or lesser degree are associated with particular behaviors and hopefully can be prevented or treated by behavior change (Ferguson 1998,

Canada Department of Health and National Welfare 1974).

It is one thing to identify a risk behavior and another thing to insure that people will actually change that behavior. While there certainly are examples of successful programs to change behavior, the evidence suggests that behavior change is a very difficult and complex challenge (Syme 1986). While many different factors need to be considered in helping people change their behavior, one factor rarely considered is that of the environment. It is difficult to expect that people will change their behavior easily when many forces in the social, cultural, and physical environment conspire against such change (Leventhal and Cleary 1980). If successful behavior-modification programs are to be developed to prevent diseases, more attention will need to be given not only to the behavior and risk profiles of individuals, but also to the social environment within which people live.

2. *Social Factors and the Incidence of Disease*

2.1 *Personal Dispositions*

Since about 1970, research evidence has accumulated regarding the role of many social factors in disease etiology. While most of this work has been done in reference to coronary heart disease, other diseases also have been studied. Early studies emphasized the role of mobility and Type A behavior pattern as risk factors for disease, but subsequent research did not uniformly confirm these findings. A new generation of research has emerged from this early work that focuses on particular components of the behavior pattern such as antagonism, resentment, time urgency, anger, need for control, self-involvement, self-esteem, and hardiness (Syme 1992). The most impressive and consistent findings from this body of research have been on the concept of hostility which has been shown to be related to coronary heart disease in a wide variety of studies (Miller et al. 1996). Other research noted below has a limited sense of control to health, with other dispositions such as optimism and negative affect receiving increased attention.

2.2 *Social Support and Social Connection*

Historically, one of the earliest areas of interest in the social patterning of disease has been the beneficial effect of marital status on health. While this research area has yielded interesting results, it also has branched out to new research areas on social support and social connection.

It has been known for many years that people who are not married—whether single, separated, widowed, or divorced—have higher mortality rates than married people. Not only is having a spouse beneficial to health, but loss of spouse to death is particularly detrimental. Common to marital status research are the themes of social support and social connection. In an influential article Cassel (1976) pointed out that lack of social connection, such as disruption of marriage or moving to a new social setting, was associated with a wide range of disease outcomes. According to Cassel's hypothesis, the essential element in these situations was lack of social connection or support. In the absence of social connections a person becomes generally susceptible to illness, the specific type of which was determined by that person's particular risk situation. The concept of social support subsequently was studied in a large community sample in Alameda County, California (Berkman and Syme 1979), in Tecumseh, Michigan (House et al. 1982), and in eastern Finland (Kaplan et al. 1988). In these studies, a strong association between networks and mortality has been observed for all causes of death for men, but only sometimes for women. These findings persisted even after account was taken of a broad range of other risk factors such as age, obesity, smoking, physical activity, health practices, education, and family history.

Several reviews (Broadhead et al. 1983, House et al. 1988, Berkman and Glass 2000) have critically assessed this literature and have concluded that something of importance is going on but that the precise elements of this 'something' are not yet clear. One possible explanation for the inconsistent pattern of findings may be that a few simple questions about relationships (e.g., about marriage, clubs, and number of friends) may be enough to separate those with ties from those without ties in large, urban areas like Alameda County, but that these questions are not precise enough in smaller, more rural communities, or for women. Further work will be necessary to better define the importance of various dimensions of social networks and social support for health and wellbeing.

2.3 *Socioeconomic Status*

Socioeconomic status has the effect of grouping individuals and of nesting them within geographic, social, and occupational environments. Obviously, not all environments are equal. They are more or less desirable places to live or work for a variety of reasons: more or fewer public and private resources; better or poorer housing and infrastructure; higher or lower levels of social capital; and more or less evenly distributed wealth. These environmental differences are patterned by socioeconomic status. Yet, as noted above, socioeconomic status is often treated as if it was a characteristic of individuals. Poverty, richness, and social status are qualities of environments as well as of people.

Evidence that area socioeconomic status has an independent relationship with mortality, over and above individuals, has emerged in the 1990s. Haan and

colleagues showed that residents in a federally designated poverty area experienced elevated age, race, and sex-adjusted mortality over a follow-up period compared to residents of a nonpoverty area (Haan et al. 1987). In this study, the heightened risk of death persisted after adjustment for a wide range of demographic, behavioral, social, psychological, and health characteristics. These researchers concluded that the social environment contributed to the association between low socioeconomic status and excess mortality, independent of individual factors. Recently, other analyses have supported these results on a nationwide scale (Anderson et al. 1997).

Further support for this observation comes from England, where Blaxter (1990) investigated a wide variety of health outcomes in manual and nonmanual workers living in different types of areas. The interaction between social class of area of residence and individual social class characteristics was also reported by Yen and Kaplan (Yen and Kaplan 1996), and Shouls and colleagues (Shouls et al. 1996). Both studies found that poverty area of residence, together with individual characteristics, affect illness and adoption of risk behaviors.

Another approach to the study of area quality, one that focuses on economic patterns, involves attention to equality of income distribution, a concept which has no equivalent measure at the individual level. In recent years, evidence has accumulated to suggest that within populations of developed countries, it is the relative socioeconomic position as well as the absolute level of income that is associated with health (Wilkinson 1992). Wilkinson has shown that, in a sample of industrialized countries, life expectancy increases as the distribution of income in the country becomes more egalitarian. Data from the Luxembourg Income Study revealed a correlation of 0.86 between life expectancy and the proportion of the total net income received by the least well-off 70 percent of the population. In addition, when changes in income distribution and life expectancy were compared over time, increases in equality of income distribution were associated with greater increases in life expectancy.

Two other recent studies have explored the association between income distribution and mortality and morbidity within the United States. Although these studies used different measures of income distribution, they report similar results: as the equality of the income distribution increases, the age-adjusted rates of all-cause mortality decreases. In one of these studies (Kennedy et al. 1996), similar relationships are seen between state income equality and mortality from heart disease, malignant neoplasms, cerebrovascular disease, and homicide. In the second study (Kaplan et al. 1996), strong, significant correlations are observed between greater income equality and lower rates of all-cause and age-specific mortality, other health indicators (including rates of low birth weight, violent crimes, disability, per capita expenditures on protection, sedentary behavior, and current smoking), other social indicators (proportions of people who are unemployed, proportion receiving Aid to Families with Dependant Children, and proportion without health insurance). There is considerable debate as to the interpretation of these findings regarding income inequality. Some argue that income inequality at the population level simply represents individual differences within that population and that these individual characteristics cause disease (Gravelle 1998). Others suggest that income inequality affects health because it causes individuals with low income to experience relative deprivation with resultant feelings of distrust, shame, and hostility (Kawachi et al. 1997). A third view is that income equality is a reflection of political and economic forces in the community that generate inequality and that impact the distribution of such resources as schooling, medical care, social welfare, and working conditions (Lynch et al. 2000) which, in turn, affects health and wellbeing.

Social class differences are important determinants of health. Further research to better understand the meaning of social class gradients and of income inequality is important to help in the development of appropriate and effective interventions to improve health.

One of the most persistent disease patterns observed in public health research is that people in the lowest socioeconomic groups have the highest rates of morbidity and mortality. In a comprehensive review of 30 studies on this topic, Antonovsky noted the consistency of this finding dating from the twelfth century (Antonovsky 1967). Further, this differential has been observed throughout the world, regardless of whether the dominant diseases of death and disability were attributed to infectious or noninfectious causes and regardless of the specific methods used to assess socioeconomic status (Syme and Berkman 1976, Feinstein 1993).

In a massive nationwide survey of mortality in the United States, Kitagawa and Hauser (Kitagawa and Hauser 1973) found that mortality rates varied dramatically among socioeconomic groups for both men and women, whether socioeconomic status was studie in relation to education, income, or occupation: the lower the socioeconomic level, the higher the death rate. In addition, Kitagawa and Hauser found that those in lower socioeconomic groups had higher death rates for every cause of death except, among women, cancer of the breast and motor vehicle accidents. Higher rates of morbidity also have been observed among those in lower socioeconomic groups. These higher morbidity rates include virtually every disease as well as mental illnesses and conditions such as schizophrenia, depression, unhappiness, worry, anxiety, and hopelessness (Syme and Berkman 1976, Haan et al. 1989).

One problem with this observation is to identify those components of socioeconomic status that are

potentially amenable to intervention. Research done by Marmot and his colleagues (Marmot et al. 1978) on British civil servants provides an opportunity to identify some of these components. These investigators have demonstrated that British civil servants in the highest grade (administrators) had the lowest rate of coronary heart disease while those in the lowest grade (mainly unskilled manual workers) had rates four times as high. After account had been taken of such coronary heart disease risk factors as serum cholesterol, cigarette smoking, blood pressure, physical activity, glucose intolerance, and social support, the differences in rate between those at the top and bottom of the graded hierarchy was reduced to three times. However, about 60 percent of the difference in coronary heart disease rates among civil service grades remained unexplained after this adjustment. More interesting is the fact that workers in professional and executive jobs (grade 2) and in clerical jobs (grade 3) have coronary heart disease rates 2 and 3.2 times as high as administrators.

Although the relation of socioeconomic position to health is not linear in the general population, since the largest differences are usually seen at the broad lower end of the distribution, the finding of Marmot and colleagues poses a challenge. While it is reasonably simple to come up with possible explanations for the fact that those at the bottom have higher rates than those at the top, these explanations do not account for the fact that those close to the top have higher rates of disease than those at the top. Factors such as inadequate medical care, unemployment, low income, racial factors, poor nutrition, poor housing, and poor education may account for higher rates of disease among those in grade 5 but they seem unlikely to explain why professionals and executives in the British civil service have rates twice as high as administrators.

These socioeconomic differences in disease are not unique to British civil servants. They have been observed in a wide variety of populations in many different countries, and are not confined to a single age group or disease entity, having been observed for many body systems—including the digestive, genitourinary, respiratory, circulatory, nervous, blood, and endocrine systems (Haan et al. 1989, Marmot et al. 1978, Susser et al. 1985)—and diseases (e.g., most malignancies, congenital anomalies, infectious and parasitic diseases, accidents, poisoning and violence, perinatal mortality, diabetes, and musculoskeletal impairments) as well as for early stages of and progression to carotid atherosclerosis (Lynch et al. 1995, Lynch et al. 1997).

A number of approaches have been taken to explain the link between socioeconomic status and disease. One possible explanation is the concept of 'control of destiny.' It could be postulated, for example, that the lower down one is in the socioeconomic status hierarchy, the less control one has over the factors that affect life and living circumstance. This hypothesis is very general, and it does not specify whether control involves money, power, information, prestige, experience, or something else. Over the years, many social scientists have studied different concepts related to the idea of 'control of destiny': mastery, self-efficacy, locus of control, learned helplessness, controllability, predictability, desire for control, sense of control, powerlessness, hardiness, competence, and so on (Syme 1989).

Research on job stress provides an example of the feasibility of intervening on 'control.' For many years, researchers have explored the relationship between 'job stress' and disease. It was only after Karasek (Karasek et al. 1981), Theorell (Theorell et al. 1984), and their associates added the idea of job latitude and discretion to that of job stress that such a link was observed. These investigators have shown in several studies that occupational stressors have consequences for health especially when workers do not have sufficient latitude and discretion for coping with these stressors. When workers have little control over work pace and methods, higher rates of catacholamines are seen as well as higher rates of mental strain, coronary heart disease, and other health problems. The implications for prevention are perhaps clearer when one is able to focus on such concepts as worker discretion, latitude, and involvement rather than broader concepts such as socioeconomic status. This focus also directs attention to the work environment and not simply to the individual worker.

2.4 *Race and Ethnicity*

Another pervasive and persistent patterned regularity in the study of health and disease is seen between race and ethnic groups. For example, young and middle-aged Black Americans have disproportionately high morbidity and mortality rates (Lillie-Blanton et al. 1996). McCord and Freeman (1990) have estimated that Black men living in Harlem in 1980 had less chance of surviving to the age of 65 than men in Bangladesh, and recent data from a nationwide study by Geronimus (1992) show that this situation was even worse in 1990. Geronimus and colleagues indicate that about half of the higher mortality rate in Blacks is due to poverty. Other possible explanations for the differences in mortality rate and life expectancy might include discrimination, lifestyle factors, use of medical care, and biologic differences. Exploration of the relative importance of these factors is complicated by the fact that 'race' is such an imprecise term. Biologically, race is an indicator of the distribution of gene frequencies across populations, but, as Susser et al. (1985) point out, this distribution of genes is deeply conflated with geographic, social class, and cultural distinctions. Many physical anthropologists prefer to dispense with the term altogether because it is so vague.

Nevertheless, race clearly has social meaning. Race has impact on life choices and opportunities, on education and employment. Race influences the social environmental conditions in which people live. And race influences how people interact with each other and this may have health consequences. In an innovative and pioneering study, Krieger and Sidney (1996) studied the consequences of racial discrimination on blood pressure in 5,115 young adults in four US cities: Birmingham, Chicago, Minneapolis, and Oakland. Eighty percent of Black men and women reported experiences of racial discrimination in seven basic life situations. Racial discrimination and internalized responses to the discrimination were associated with elevations of blood pressure. Working-class Black adults who reported no experiences of discrimination, yet said they typically accept unfair treatment as a fact of life, had systolic blood pressure 7 mm Hg higher than white adults, and had higher blood pressure than Black adults who typically challenged and reported discrimination.

Racial categories are clearly associated with differences in life expectancy, disability, cirrhosis, infant mortality, birthweight, arthritis, diabetes, hypertension, anemia, and virtually every cause of death among adults. These are important and complex issues that deserve a high priority in our research agenda. Recent research on race, especially that focusing on Black–White differences, is providing important new findings (Geiger 1996, Fang et al. 1996, Laveist 1993). It is clear that this issue deserves serious and continued attention directed to a wider range of racial and ethnic groups.

3. Conclusion

The importance of social factors in the etiology of many diseases is becoming increasingly clear. The evidence for some factors remains tentative or unclear. Nevertheless, it is impressive that an increasingly large body of consistent findings is being generated in spite of these major methodologic problems. It may be premature to use the data emerging from this research in public health programs, but it is clear that they will be needed soon. The reason for this is that most serious diseases today are importantly influenced by the social environment. For these reasons, interventions in the social environment clearly will be necessary, and continued research on social factors therefore must become an important priority in both public health planning and program development.

Bibliography

Anderson R T, Sorlie P, Backlund E, Johnson N, Kaplan G A 1997 Mortality effects of community socioeconomic status. *Epidemiology* **8**(1): 42–7

Antonovsky A 1967 Social class, life expectancy and overall mortality. *Milbank Memorial Fund Quarterly* **45**: 31–73

Berkman L F, Glass T 2000 Social integration, social networks, social support, and health. In: Berkman L F, Kawachi I (eds.) *Social Epidemiology*. Oxford University Press, New York

Berkman L F, Syme S L 1979 Social networks, host resistance, and mortality: A nine year follow-up study of Alameda County residents. *American Journal of Epidemiology* **109**: 186–204

Blaxter M 1990 *Health and Lifestyles*. Routledge, London

Broadhead W E, Kaplan B H, James S A, Wagner E H, Schoenbach V J, Grimson R, Heyden S, Tibblin G, Gehlbach S H 1983 The epidemiologic evidence for a relationship between social support and health. *American Journal of Epidemiology* **117**: 521–37

Canada Department of National Health and Welfare 1974 *A New Perspective on the Health of Canadians: A Working Document*. Ottawa, Canada

Cassel J 1976 The contribution of the social environment to host resistance. *American Journal of Epidemiology* **104**: 107–23

Cohen S, Syme S L (eds.) 1985 *Social Support and Health*. Academic Press, New York, Orlando, FL

Dekker E 1975 Youth culture and influences on the smoking behavior of young people. In: *Smoking and Health. Smoking and Health. Proceedings of the Third World Conference. Department of Health, Education, and Welfare, Public Health Service, National Institutes of Health*. Washington, DC, pp. 381–92

Fang J, Madhavan S, Alderman M H 1996 The association between birthplace and mortality from cardiovascular causes among Black and White residents of New York City. *New England Journal of Medicine* **335**: 1545–51

Feinstein J S 1993 The relationship between socioeconomic status and health: A review of the literature. *Milbank Quarterly* **71**: 279–322

Ferguson K S 1998 Health behavior. In: Wallace R B, Doebbling B N, Last J M (eds.) *Public Health and Preventive Medicine*. Appleton and Lang, Stamford, CT

Geiger H J 1996 Race and health care—An American dilemma? Editorial. *New England Journal of Medicine* **335**: 815–16

Geronimus A T 1992 The weathering hypothesis and the health of Africa-American women and infants: Evidence and speculations. *Ethnicity and Disease* **2**: 207–21

Gravelle H 1998 How much of the relation between population mortality and unequal distribution of income is a statistical artifact? *British Medical Journal* **316**: 382–5

Haan M N, Kaplan G A, Camacho T 1987 Poverty and health: Prospective evidence from the Alameda County Study. *American Journal of Epidemiology* **125**: 989–98

Haan M N, Kaplan G A, Syme S L 1989 Socioeconomic status and health: Old observations and new thoughts. In: Bunker J P, Gomby D F, Kehrer B H (eds.) *Pathways to Health: The Role of Social Factors*. H. J. Kaiser Family Foundation, Menlo Park, CA, pp. 76–135

House J S, Landis K R, Umberson D 1988 Social relationships and health. *Science* **241**: 540–5

House J S, Robbins C, Metzner H L 1982 The association of social relationships and activities with mortality: Prospective evidence from the Tecumseh Community Health Study. *American Journal of Epidemiology* **116**: 123–40

Kaplan G A, Pamuk E R, Lynch J W, Cohen R D, Balfour J L 1996 Inequality in income and mortality in the United States: Analysis of mortality and potential pathways. *British Medical Journal* **312**: 999–1003

Kaplan G A, Salonen J T, Cohen R D, Brand R J, Syme S L, Puska P 1988 Social connections and mortality from all causes and from cardiovascular disease: Prospective evidence from Eastern Finland. *American Journal of Epidemiology* **128**: 370–80

Karasek R, Baker D, Marxer F, Ahlbom A, Theorell T 1981 Job decision latitude, job demands, and cardiovascular disease: A prospective study of Swedish men. *American Journal of Public Health* **71**: 694–705

Kawachi I, Kennedy B P, Lohner K, Prothrow-Stith D 1997 Social capital, income inequality and mortality. *American Journal of Public Health* **87**: 1491–8

Kennedy B P, Kawachi I, Prothrow-Stith D 1996 Income distribution and mortality: Cross sectional ecological study of the Robin Hood index in the United States (erratum: 5/96). *British Medical Journal* **312**(7037): 1004–7

Kitagawa E M, Hauser P M 1973 *Differential Morality in the United States*. Harvard University Press, Cambridge, MA

Krieger N, Sidney S 1996 Racial discrimination and blood pressure: The CARDIA study of young black and white adults. *American Journal of Public Health* **86**: 1370–8

Laveist T A 1993 Segregation, poverty, and empowerment: Health consequences for African Americans. *Milbank Quarterly* **71**: 41–64

Leventhal H, Cleary P D 1980 The smoking problem: A review of the research and theory in behavioral risk modification. *Psychological Bulletin* **88**: 370–405

Lillie-Blanton M, Parsons P E, Gayle H, Dievler A 1996 Racial differences in health: Not just Black and White, but shades of gray. *Annual Review of Public Health* **17**: 411–48

Lynch J, Kaplan G A, Salonen R, Cohen R D, Salonen J T 1995 Socioeconomic status and carotid atherosclerosis. *Circulation* **92**: 1786–92

Lynch J, Kaplan G A, Salonen R, Salonen J T 1997 Socioeconomic status and progression of carotid atherosclerosis: Prospective evidence from the Kuopio Ischemic Heart Disease Risk Factor Study. *Arteriosclerosis, Thrombosis, and Vascular Biology* **17**(3): 513–9

Lynch J Smith G D Kaplan G A House J S 2000 Income inequality and mortality: Importance to health of individual income, psychosocial environment, or material considerations. *British Medical Journal* **320**: 1200–1204

Marmot M G, Rose G, Shipley M, Hamilton P J S 1978 Employment grade and coronary heart disease in British civil servants. *Journal of Epidemiology and Comminity Health* **32**: 244–9

Marmot M G, Syme S L 1976 Acculturation and coronary heart disease in Japanese Americans. *American Journal of Epidemiology* **104**: 225–47

McCord C, Freeman H P 1990 Excess mortality in Harlem. *New England Journal of Medicine* **322**: 173–7

Miller T Q, Smith T W, Turner C W, Guijarro M L, Hallet A J 1996 A meta-analytic review of research on hostility and physical health. *Psychological Bulletin* **119**: 322–48

Shouls S, Congdon P, Curtis S 1996 Modeling inequality in reported long term illness in the UK: Combining individual and area characteristics. *Journal of Epidemiology and Comminity Health* **50**: 366–76

Susser M W, Watson W, Hopper K 1985 *Sociology in Medicine*. Oxford University Press, New York

Syme S L 1986 Strategies for health promotion. *Preventitive Medicine* **15**: 492–507

Syme S L 1989 Control and Health: A personal perspective. In: Steptoe A, Appels A (eds.) *Stress, Personal Control and Health*. Wiley, New York

Syme S L 1992 Social determinants of disease. In: Last J M, Wallace R B (eds.) *Public Health and Preventive Medicine*. Appleton and Lange, Stamford, CT

Syme S L, Alcalay R 1982 Control of cigarette smoking from a social perspective. *Annual Review of Public Health* **3**: 179–99

Syme S L, Berkman L F 1976 Social class, susceptibility and sickness. *American Journal of Epidemiology* **104**: 1–8

Theorell T, Alfredsson L, Knox S, Perski A, Svensson J, Waller D 1984 On the interplay between socioeconomic factors, personality and work environment in the pathogenesis of cardiovascular disease. *Scandinavian Journal of Work Environment and Health* **10**: 373–80

Wilkinson R G 1992 For debate—Income distribution and life expectancy. *British Medical Journal* **304**: 165–8

Yen I, Kaplan G A 1996 *Poverty area residence and prospective change in risk behavior: Evidence for the Alameda County Study*. Presented at Society for Epidemiologic Research Meeting, Boston, MA

Ziegler R G, Hoover R N, Pike M C, Hildesheim A, Nomura A M, West D W, Wu-Williams A H, Kolonel L N, Horn-Ross P L, Rosenthal J F, Hyer M B 1993 Migration patterns and breast cancer risk in Asian-American women. *Journal of the National Cancer Institute* **85**(22): 1819–27

S. L. Syme

Epigenetic Inheritance

'Epigenetic inheritance' is a term used to describe the transmission from one generation to the next of structural and functional variations that do not depend on genetic differences. In its broadest sense, the term includes the transmission of nongenetic variations in behavior, but its use is usually restricted to the inheritance in cell lineages of variations that are not based on changes in the genetic information carried in DNA. Such heritable epigenetic changes are part of normal development but there is increasing evidence that they can also be transmitted from one generation of organisms to the next. Epigenetic inheritance therefore has practical implications for medicine and agriculture, and theoretical implications for ideas about heredity and evolution.

1. *Definitions and Historical Background*

'Epigenetics,' 'epigenetic,' and 'epigenetic inheritance' are relatively new terms, derived from the much older term 'epigenesis.' According to the ancient and influential theory of epigenesis, development involves qualitative changes in organization, with the parts of the adult animal forming gradually and sequentially from an ovum that is amorphous and unorganized. In the rival theory of preformation, which was popular in the late seventeenth and eighteenth centuries, development was assumed to entail the growth or

unfolding of parts that already exist in miniature in the egg (Pinto-Correia 1997). Since the beginning of the nineteenth century, it has been accepted generally that development is through epigenesis, although a subtle form of preformation was introduced towards the end of the nineteenth century, when it was recognized that although developmental changes are qualitative, they are controlled by preformed determinants whose transmission and effects ensure hereditary continuity. For several decades after the rediscovery of Mendel's laws in 1900, the new science of genetics was concerned with the way these determinants, the genes, are transmitted from one generation to the next; their role in development received little attention until the second half of the twentieth century.

The introduction of the term 'epigenetics' by the British embryologist Conrad Waddington in 1947 (see Waddington 1975), reflected a growing interest in the relationship between genes and development, an interest which increased as the molecular nature of the gene was unraveled. Waddington coined the term epigenetics for the study of the causal mechanisms of development; the adjective 'epigenetic' is applied to the processes through which the genetic information in a fertilized egg is expressed and used in the production of the adult form, the visible phenotype. As the study of genetics and development progressed, it became clear that different cell types usually have identical genotypes, but express different sets of genes. (One important exception proved to be the cells of the immune system, where epigenetic changes involve alterations in DNA sequences.) Finding out which genes are expressed and what controls when and where they are expressed became the key to understanding the functional and structural specialization of cells and tissues during development. It also became clear that once a cell is committed to a particular form or function, that state is often transmitted to daughter cells, even though the stimulus that originally induced its development is no longer present. Thus, once a cell is committed to becoming a liver cell, it breeds true and by division gives rise to a lineage of liver cells; similarly, skin cells will give rise to skin cells, and kidney cells to kidney cells. This stable transmission of the alternative phenotypes of cells with the same genotype became known as 'cellular heredity,' 'cell heredity,' or 'epigenetic inheritance.'

Interest in the transmission of cellular states began to grow during the mid-1970s, as the molecular mechanisms and systems that underlie the control of gene expression started to be unraveled (Holliday 1990). Gradually, it was recognized that epigenetic information is transmitted not only between generations of cells in the cell lineages within an individual, but also between generations of individuals. The concept of epigenetic inheritance was therefore extended and now has both a narrow and a broad meaning. The narrow meaning limits it to cell memory in lineages within an organism, whereas the broad meaning includes the transmission of epigenetic information between any reproducing entities—cells, individuals, and even groups of individuals. This article focuses first on epigenetic inheritance in cell lines, and then discusses the consequences of the transmission of epigenetic variations between generations of individuals.

2. *Mechanisms of Transmission*

The systems that underlie epigenetic inheritance and enable the phenotypic expression of the genetic information in a cell to be transmitted to the next generation have been called epigenetic inheritance systems (EISs). Three broad types have been characterized (Jablonka and Lamb 1995): steady-state systems, structural systems, and chromatin-marking systems.

2.1 *Steady-state Systems*

These systems lead to patterns of gene activity persisting in cell lineages through the operation of regulatory feedback loops. The simplest example is that of a gene that controls its own transcription through positive feedback. When the activity of such a gene is induced by an external stimulus, some of the gene's product bind to its own control region. This binding leads to the continued activity of the gene. In this way, gene activity that initially depended on the presence of an external inducer is maintained by self-regulation—by the continuous induction of the gene by its own product. If the concentration of the gene-product is high enough and cell division is more or less equal, the gene's active state will be inherited, because the regulatory gene-product will be transmitted to daughter cells as an automatic consequence of cell division. This type of self-regulation means that cells with the same DNA sequence can have different but stable patterns of cell-heritable activities, which reflect their developmental history. Many examples of such self-perpetuating regulatory loops have been found in a variety of animals and plants. In most cases the systems are more complicated than the one described, with several genes and gene products involved in the feedback loop, but the principle is the same.

2.2 *Structural Inheritance Systems*

Cells contain many complex three-dimensional structures whose assembly seems to depend on the presence of existing similar structures. In ways that are not understood, old structures are used as templates for the production of the new structures that are required for cell growth and multiplication. The best-known examples of this type of inheritance come from the

study of single-celled organisms, such as the ciliated protozoan *Paramecium*. Genetically identical paramecia can show heritable differences in the patterns of ciliary rows on their cell surface, and experimentally altered patterns can be transmitted to daughter cells for hundreds of cell generations (Grimes and Aufderheide 1991). Although this type of inheritance is particularly conspicuous and stable in ciliated protozoa, structural inheritance is probably universal, since all cells rely on pre-existing structures when they produce new ones. The self-templating properties of some three-dimensional protein complexes is believed to underlie the transmission of diseases such as mad cow disease (BSE—bovine spongiform encephalopathy), where aberrant prion proteins seem to induce function-affecting changes in the conformation of normal prion proteins (Prusiner 1998).

2.3 Chromatin-marking Systems

Chromatin-marking systems are currently the best-understood EISs. Chromatin is the complex of DNA, proteins, and other macromolecules that make up chromosomes, and chromatin marks are the binding proteins or chemical groups that influence the activity of the DNA to which they are attached (Jablonka and Lamb 1995). The epigenetic information contained in chromatin marks is carried from one cell generation to the next with DNA. A good example of an inherited chromatin mark is a gene's methylation pattern. In many plants and animals, some of the cytosines in DNA are methylated. Although cytosine (C) is one of the four bases whose sequence encodes genetic information, the presence of a methyl group (C^m) does not affect its coding properties. However, cytosine methylation within or around a gene often determines whether or not the gene is active, and how easy it is to activate: usually low levels of methylation are associated with potential activity, while high levels are associated with inactivity.

Changes in methylation patterns occur during growth and development, so the same DNA sequence can have several different methylation patterns, each reflecting a different functional state, and each inherited stably through many cell divisions. The mechanism of transmission in this case is well understood, and depends on the way a cytosine in one strand of the DNA double helix is always paired with a guanine base (G) in the other. Methylation occurs symmetrically on the two strands, usually in CG doublets in which C^mG is paired with GC^m. Following DNA replication, in which the two parental DNA strands separate and are used as templates for building complementary daughter strands, the previously methylated sites remain methylated on the parental strand, but are unmethylated on the newly synthesized daughter strand. However, special enzymes recognize the asymmetrical, half-methylated sites, and preferentially methylate the cytosines of the new strand (Holliday 1990, Jablonka and Lamb 1995). Thus, because of the mirror symmetry of CG sites and the semiconservative nature of DNA replication, the specific pattern of methylation of a region of DNA is reproduced at every cell division.

Methylation patterns are not the only type of heritable chromatin marks. Chromatin marks involving DNA-associated proteins that affect the maintenance of gene activity can also be transmitted in cell lineages. The way these protein marks are replicated is beginning to be understood, and plausible models for the reproduction of nucleoprotein complexes which ensure that daughter cells inherit the state of activity of parental cells have been proposed (Lyko and Paro 1999).

3. Genetic and Epigenetic Inheritance Compared

The properties of the genetic and epigenetic inheritance systems overlap, but certain features are more characteristic of one than the other. Most stem from the fact that epigenetic systems are concurrently inheritance systems and response systems, whereas, with a few exceptions, information transmitted by the genetic system is not altered in response to external factors. The response function of the epigenetic system means that, unlike genetic variations, the origin of heritable epigenetic variations is not random—they are often induced by specific developmental or environmental changes. Frequently induced epigenetic variations are adaptive, although they need not be so. Whereas changes in the DNA sequence of a gene are rare, and reversion to the original sequence is even rarer, the response function of EISs means that, under some circumstances, the rate of change between epigenetic variants is very high and there may be coordinated changes in the epigenetic marks of several genes. However, although some inherited epigenetic changes differ from typical genetic changes in being nonrandom, frequent, reversible and adaptive, others resemble genetic mutations in that they are rare, random, nonadaptive, and unlikely to revert. No single criterion or set of criteria can be used to distinguish between the properties of the genetic and epigenetic systems.

4. Transgenerational Epigenetic Inheritance

Studies of the mechanisms of epigenetic inheritance were initially focused on cells in the cell lineages of multicellular organisms, but subsequently it was found that single-celled organisms such as bacteria and yeasts also transmit epigenetic variations to their descendants (Grandjean et al. 1998, Klar 1998). It was also realized that multicellular organisms sometimes transmit epigenetic variations between generations. Transgenerational transmission was recognized in two ways: either as genomic imprinting, in which a gene's

expression is influenced by the sex of the parent that transmitted it, or as the long-term inheritance of acquired variations over several generations.

4.1 Genomic Imprinting

During the formation of sperm and eggs, chromatin is modified. The way that it is modified in the male and female germ lines differs, so, at the beginning of embryonic development, the genes transmitted by the two parents are in different epigenetic states. If these differences persist during development and have phenotypic effects, the genes are said to be 'imprinted.' Imprints may involve differences in methylation patterns or in other types of chromatin marks. The effect of genomic imprinting is that the expression of some genes depends on the sex of the parent from which they were inherited. Normally imprints are erased and replaced in the germ line, so when a gene that was inherited from one sex is transmitted to the next generation through the opposite sex, the imprints are reversed. Imprinting seems to be widespread in both plants and animals, but currently it is being studied most intensively in mammals. It is known to be necessary for normal development and to affect the severity of expression of some genetic diseases, such as Prader-Willi and Angelman syndromes (Reik and Surani 1997).

4.2 Long-term Transmission

Most cases of the long-term transmission of epigenetic variations have been found in plants, which often have vegetative as well as sexual reproduction, but it has also been found in a variety of animals (Jablonka and Lamb 1995, 1998). Remarkably, the first morphological flower-mutant ever to be characterized, a peloric form of toadflax described by Linnaeus over 250 years ago, has turned out to be an epimutation, not a mutation; the difference between the normal and peloric forms is due to a difference in the gene's state of methylation, not in its DNA sequence (Cubas et al. 1999). Similarly, a well-studied case of variation in the coat color of the mouse has been shown to be the result of heritable epigenetic modifications, rather than of genetic or environmental effects (Morgan et al. 1999). Findings like these suggest that the long-term transmission of epigenetic variations may be far more common than was at one time thought likely.

5. Implications

Recognizing the ubiquity and importance of epigenetic inheritance has major implications for medicine and agriculture, as well as for the way that we think about heredity and evolution. Since the practical implications were the first to be generally recognized, they will be considered first.

5.1 Practical Implications

In recent years it has become clear that the variations transmitted through EISs can affect health and disease. First, many cancers involve cell-heritable epigenetic modifications, such as alterations in methylation patterns and other aspects of chromatin organization (Zingg and Jones 1997). Second, some hereditary diseases are caused by mutated imprinted genes, which show peculiar but characteristic patterns of inheritance (Reik and Surani 1997). In some cases the inherited conditions may not be caused by mutations, but by epimutations. Third, prionic diseases such as BSE and kuru seem to be the consequence of transmissible infectious protein complexes and the mechanism of reproduction and reconstruction of the infectious agents seems to involve structural, three-dimensional templating (Prusiner 1998). Fourth, some environmentally induced diseases, such as those caused by starvation or by taking thalidomide, seem to have long-term transgenerational effects (Holliday 1998). Epidemiological research programs and medical practice will have to accommodate information like this, and develop ways of recognizing, avoiding, and curing the diseases caused by heritable epigenetic changes.

In agriculture, the importance of epigenetic inheritance is already acknowledged widely, because it has caused many problems in genetic engineering aimed at crop improvement. Commonly, newly inserted foreign genes are heritably silenced through extensive DNA methylation, so ways of circumventing this problem have had to be developed. On the positive side, since some epigenetic variations can be induced by environmental changes, it may be possible to develop agricultural practices that exploit these inducing effects and thus develop epigenetically 'engineered' improved crops.

5.2 A Broader Concept of Inheritance and Evolution

Including EISs and other non-DNA inheritance systems (such as the transmission of information through behavior or language) produces a much richer and more general concept of heredity. Heredity does not depend solely on genes, and not only genotypes are inherited. Even at the strictly cellular level, the production and transmission of inherited variations involves several different mechanisms. With some of these, such as those underlying structural inheritance and steady-state systems, there is no real equivalent of the phenotype–genotype distinction, because the re-

construction of the phenotype is an integral part of the transmission mechanism.

With all EISs, the inheritance of acquired (induced) variation is common. Consequently, there is a 'Lamarckian' component in evolution, with the environment being both an inducer and a selector of epigenetic variations. The gradual broadening of the concept of heredity, and recognition that natural selection acts on several different types of heritable variation, which began in the last decades of the twentieth century, has challenged the widely accepted gene-centered neo-Darwinian version of Darwinism. A more comprehensive and more powerful Darwinian evolutionary theory is now being constructed, one that incorporates multiple and interacting inheritance systems.

See also: Behavioral Genetics: Psychological Perspectives; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Familial Studies: Genetic Inferences; Genetics and Development; Genotype and Phenotype; Lifespan Development: Evolutionary Perspectives; Lifespan Development, Theory of

Bibliography

Cubas P, Vincent C, Coen E 1999 An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401**: 157–61

Grandjean V, Hauck Y, Beloin C, Le Hégarat F, Hirschbein L 1998 Chromosomal inactivation of *Bacillus subtilis* exfusants: A prokaryotic model of epigenetic regulation. *Biological Chemistry* **379**: 553–7

Grimes G W, Aufderheide K J 1991 *Cellular Aspects of Pattern Formation: The Problem of Assembly*. *Monographs in Developmental Biology*. Karger, Basle, Vol. 22

Holliday R 1990 Mechanisms for the control of gene activity during development. *Biological Reviews of the Cambridge Philosophical Society* **65**: 431–71

Holliday R 1998 The possibility of epigenetic transmission of defects induced by teratogens. *Mutation Research* **422**: 203–5

Jablonka E, Lamb M J 1995 *Epigenetic Inheritance and Evolution: The Lamarckian Dimension*. Oxford University Press, Oxford, UK

Jablonka E, Lamb M J 1998 Epigenetic inheritance in evolution. *Journal of Evolutionary Biology* **11**: 159–83

Klar A J S 1998 Propagating epigenetic states through meiosis: Where Mendel's gene is more than a DNA moiety. *Trends in Genetics* **14**: 299–301

Lyko F, Paro R 1999 Chromosomal elements conferring epigenetic inheritance. *BioEssays* **21**: 824–32

Morgan H D, Sutherland H G E, Martin D I K, Whitelaw E 1999 Epigenetic inheritance at the agouti locus in the mouse. *Nature Genetics* **23**: 314–18

Pinto-Correia C 1997 *The Ovary of Eve: Egg and Sperm and Preformation*. University of Chicago Press, Chicago

Prusiner S B 1998 Prions. *Proceedings of the National Academy of Sciences of the USA* **95**: 13363–83

Reik W, Surani A (eds.) 1997 *Genomic Imprinting*. Oxford University Press, Oxford, UK

Waddington C H 1975 *The Evolution of an Evolutionist*. Edinburgh University Press, Edinburgh, UK

Zingg J-M, Jones P A 1997 Genetic and epigenetic aspects of DNA methylation on genome expression, evolution, mutation and carcinogenesis. *Carcinogenesis* **18**: 869–82

E. Jablonka and M. J. Lamb

Epilepsy

Epilepsy is a common disorder of the central nervous system (CNS) characterized by recurrent seizures. Seizures come in a variety of forms—some primarily motor and some affecting consciousness; some restricted to a local brain region and some involving the entire brain; some so dramatic as to be frightening and some so subtle as to be easily missed by the casual observer (Engel 1989, Engel and Pedley 1998). Current estimates are that approximately 1 percent of the population has a seizure disorder. Epilepsy respects no national, racial, or ethnic boundaries. Thirty to fifty percent of epilepsies in the U.S. (and other medically 'sophisticated' countries) are inadequately controlled by medication; that percentage is much higher in underdeveloped nations. Better understanding of the bases of seizure disorders is critical if we are to develop more effective treatments—and cures. This article provides a brief overview of our current insights into the cellular and molecular causes of epilepsy, and identifies the challenges for epilepsy research.

1. Definition of Terms

Although the term 'epilepsy' is usually used in the singular, 'epilepsy' (like 'cancer') refers to a diverse set of abnormalities, with different causes and different clinical expressions. The epilepsies are defined as chronic seizure conditions, and categorized by clinicians on the basis of the characteristic behavior and associated pattern of brain electrical discharge (the electroencephalogram, or EEG). This categorization scheme (Commission on Classification and Terminology of the International League Against Epilepsy 1981, 1989) has resulted in a useful—although complex—system for diagnosis and treatment, but generally ignores the underlying bases (molecular and cellular) for the seizures. A seizure can be viewed as a paroxysmal and synchronous discharge of a large population of neurons. Occurrence of an isolated, single seizure does not necessarily reflect brain pathology. Indeed, seizures can occur in the normal brain, and reflect normal responses to stress or injury.

Repetitive seizures, however, signal some chronic pathological condition of the CNS. Understanding that abnormality has been particularly difficult because the symptom—the seizure—is (for most epilepsies) episodic and unpredictable. Further, during the time between seizures (which may be months, or even years), the brain functions in an apparently normal fashion. Indeed, it is the unpredictability of seizures—as well as the loss of voluntary controls during these 'ictal' (i.e., seizure) episodes—that constitute major contributors to the morbidity associated with epilepsy.

2. *Categorization of Epilepsy Syndromes*

With advances in our understanding of underlying mechanisms of seizures, and of the processes by which a chronic seizure state is acquired (i.e., 'epileptogenesis'), it is now possible to categorize many epileptic disorders on the basis of the neurobiological cause rather than simply on the basis of the seizure behavior and EEG pattern. A general classification scheme, that covers most epilepsies, includes the following (overlapping) three categories: genetic disorders; developmental disorders; and acquired epilepsies (Fig. 1).

2.1 *Genetic Disorders*

A number of epilepsies have now been traced to single gene mutations. Many of the implicated genes code for ion channels (e.g., for sodium, potassium, or calcium) or for neurotransmitter receptor proteins (e.g., for acetylcholine or gamma-aminobutyric acid) (Berkovic and Scheffer 1999). Abnormalities in these proteins make at least some intuitive 'sense' in trying to understand seizure discharge, for much is known about how ion flux and neurotransmitters affect brain excitability. Other 'epilepsy genes' have been associated with structural brain abnormalities, such as seen in lissencephaly or tuberous sclerosis (Spreafico et al. 1999); abnormalities in these gene products result in developmental disorders generated through molecular signaling aberrations that we are just beginning to understand. Some of the implicated genes have functions yet to be elucidated. While single gene epilepsies are relatively rare, it is clear than some of the more common epilepsies (for example, childhood absence epilepsy—CAE, or juvenile myoclonic epilepsy—JME) are inherited in a more complex fashion, suggesting that multiple genes play roles in these disorders. Genetics also appears to play a significant role in determining seizure susceptibility, even in the normal population. Thus, major challenges for the field include characterizing and identifying the genes involved in multigenic seizure disorders, and gaining a better understanding of which genes are important in determining seizure predisposition. In addition, it is critical to understand the cellular pathways through which these genic abnormalities lead to neurophysiologic dysfunction. Identifying 'epilepsy genes' is not enough to solve the mystery of aberrant function. It is, however, a critical start—especially useful in developing new drugs (or other treatments) with specific molecular targets.

2.2 *Developmental Disorders*

The process of brain development is marvelously complex. Developmental neuroscience research has provided a rich background regarding the genetic and molecular signals involved in constructing a normal mammalian brain (Sanes et al. 2000). Abnormalities in these processes—in cell proliferation, migration, differentiation, and establishment of connections—may all lead to neurological disorders, including epilepsy (Schwartzkroin et al. 1995). Given the potential for developmental mistakes, it is actually quite remarkable that the incidence of neurological dysfunction is as low as it is. It should come as little surprise that seizure occurrence is higher among children (estimated at approximately 4 percent) than in the adult population—partly as a result of abnormalities in brain development (Spreafico et al. 1999), and partly because the properties of the immature brain render it particularly seizure-prone (Nehlig et al. 1999). The epilepsy incidence in children with other developmental disorders—cerebral palsy, mental retardation, autism, fragile X syndrome, etc.—is perhaps 10 times that in otherwise normal children, undoubtedly reflecting a general underlying brain pathology (Dodson and Pellock 2000). Since seizure incidence is so high among the young, since epilepsies may reflect abnormalities of genetics and/or brain development, and since the brain is so vulnerable to perinatal trauma (e.g., hypoxia, head trauma, infection), many epileptologists consider epilepsy to be primarily a set of disorders of childhood (Schwartzkroin and Moshe 2000). Why is it that these early insults are so likely to induce an epileptic state? Why do developmental brain abnormalities, such as seen in tuberous sclerosis or various forms of lissencephaly, correlate so highly with early seizure syndromes? We are just beginning to get a glimpse of the answers to these questions—and the general theme seems to revolve around the highly 'plastic' nature of the immature CNS. While in the adult, brain trauma often leads to cell death and loss of function, the immature brain is more resilient. Immature neurons appear to be less sensitive to injurious stimuli. The immature brain is capable of producing new neurons to a much greater extent than is the adult CNS. Immature nerve cells more readily make new connections when normal organization of the system is disrupted. While this 'plasticity'

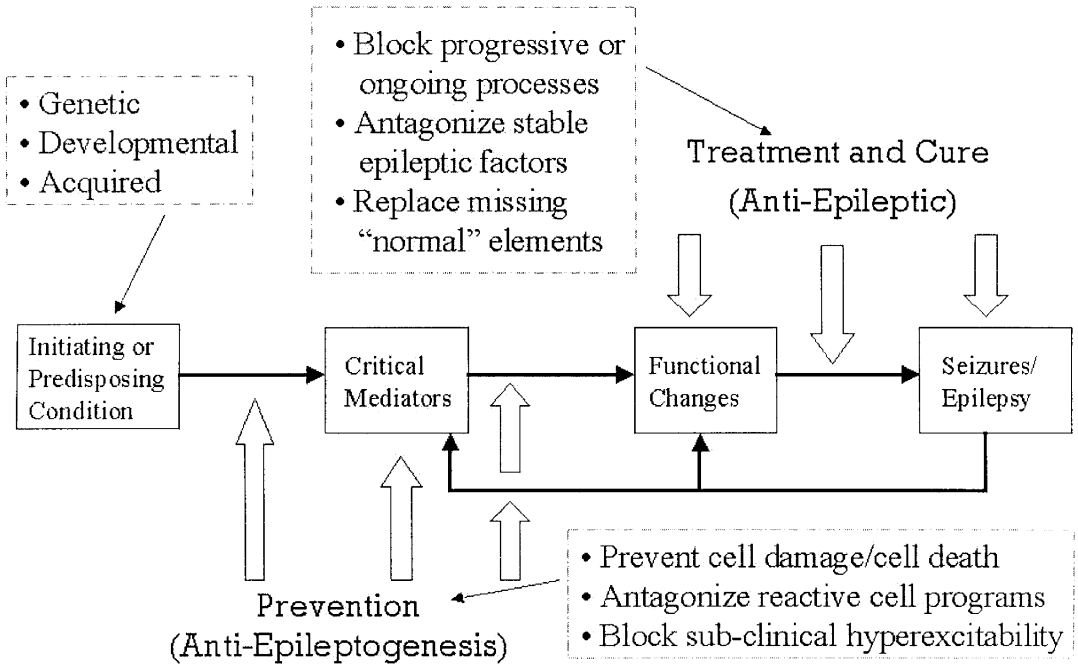

Figure 1
Diagrammatic representation of conditions that underlie epileptic disorders; potential points of intervention (open arrows) in the epileptogenic process (solid arrows); and strategies for prevention, treatment and/or cure

preserves a significant amount of behavioral function, it may come at a cost—for the new circuits established by injured neural elements may not be entirely normal. Marginal injury and inappropriate connections appear to be two major contributors to the development of abnormal electrical activity—i.e., seizures.

2.3 Acquired Epilepsies

Many epilepsies can be associated with a particular traumatic stimulus (e.g., head injury) or medical condition (e.g., a tumor or an infection), and thus are designated 'acquired.' The implication of this designation is that the brain is (was) otherwise normal, and that the identified insult is the sole cause for the epilepsy. This assumption allows us to examine the variables associated with the epileptic condition, and thus to better understand what mechanisms are responsible for development of the epileptic state. As suggested above, many of our current hypotheses revolve around the new connections made by the injured brain, and the loss of critical elements that control the electrical excitability of the system. However, while it may well be the case that a given traumatic stimulus is the critical 'trigger' for epileptogenesis, it remains something of a mystery why such injury in one individual will lead to an epileptic state but the same injury to another individual will not. This situation has led to the concept of 'predisposition' with respect to epileptogenesis (Fig. 1). That is, there may be subtle factors—gene-based differences in seizure susceptibility, subclinical injury, or developmental abnormalities—that provide a background upon which a brain insult plays out. A subtle, preexisting injury may make it more likely that an individual with head trauma develops epilepsy at a later time. This concept also implicates another important possibility that is often discussed, especially within the context of pediatric epilepsies—that epilepsy is a 'progressive' disorder (Heinemann et al. 1996, Engel et al. 2000), which develops as a result of active neuronal processes. Thus, an initial subtle and benign insult may 'kindle' into a more significant and clinically troubling condition over months, or even years.

3. Basic Neuronal Mechanisms Underlying Seizure Discharge

In general, it is thought that seizures arise as a result of an imbalance between the processes responsible for excitation and inhibition in the brain (Fig. 2). This

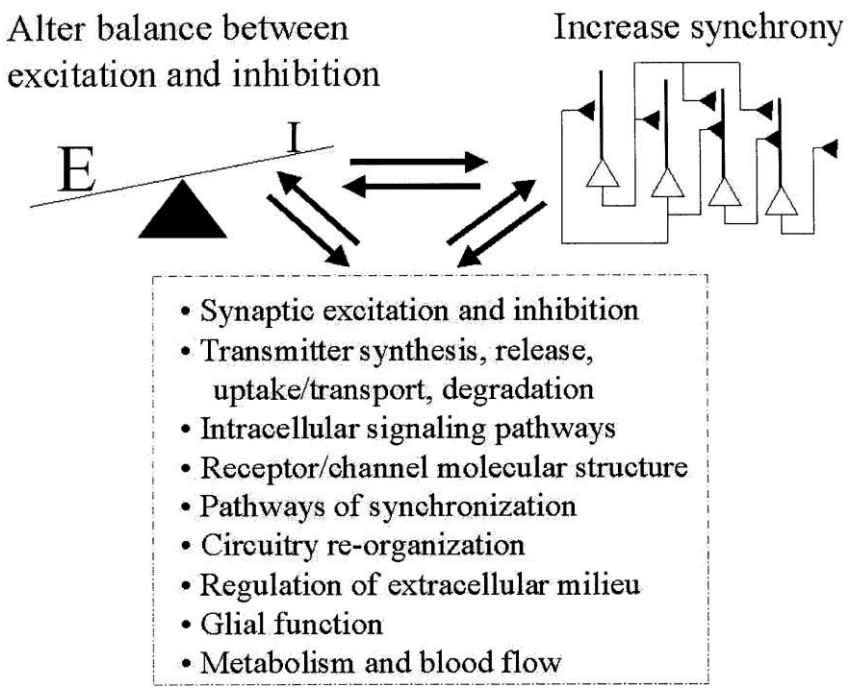

Figure 2
Basic mechanisms underlying seizure discharge and epileptogenesis—contributors to altered balance between excitation and inhibition, and increased neuronal synchrony

imbalance leads not only to a state of 'hyperexcitability' but also to a condition of 'hypersynchrony'; that is, the neurons discharge at a higher rate than normally, and discharge in a characteristically synchronous pattern that overrides the normal patterns of neuronal activity. There are a host of processes to consider in trying to understand these changes (Fig. 2) (Delgado-Escueta et al. 1996). Most research has focused on the mechanism through which nerve cells generate electrical activity (i.e., the ionic balance and ion-selective channels that are the basis for neuronal signaling) and the neurochemical communication systems through which neurons excite or inhibit each other. These are the basic neurophysiological 'building-blocks' of brain function, critical to any consideration of neural activity, in the normal brain as well as in the brain exhibiting a neurological disorder. Indeed, investigations to understand the basic mechanisms of the epilepsies have contributed greatly to our understanding of normal brain function, and to dysfunction in other neurological pathologies.

The excitation–inhibition balance in normal brain is delicately regulated—perhaps to allow for the maximum degree of 'plasticity'. While excitatory discharge is a sign by which we recognize neuronal electrical activity, it is the *inhibition* that sculpts the patterns of interaction that make up normal information processing. Thus, much of epilepsy research has focused on key inhibitory mechanisms, and many theories regarding underlying epileptic processes involve changes in the quantity—or quality—of inhibitory mechanisms. A key to much of this analysis is the critical role played by the predominant inhibitory neurochemical in the forebrain, gamma-aminobutyric acid (GABA). Reduction of GABA synthesis, loss or reorganization of neurons that use GABA for their neurotransmission, and/or alterations in the receptors sensitive to this molecule have all been proposed as mechanisms of epilepsy. Also noteworthy is the fact that many of the antiepileptic drugs currently in use exert their actions, at least in part, through GABA receptors. It is clear that even small changes in this inhibitory system can have serious consequences. For example, blocking only 15–20 percent of GABA receptors can, in experimental preparations, produce seizures. Thus, in elucidating underlying seizure mechanisms, we are faced not only with identifying the salient processes, but also understanding the significance of mechanistic disruption in *quantitative* terms.

Basic epilepsy research since the 1960s has focused on the cellular and molecular mechanisms responsible for *seizure activity*. We have still to approach the bigger question of how a *chronic seizure state* is established. Research into the basic mechanisms of epileptogenesis is now a priority in many laboratories, in large measure because investigators see an opportunity to 'cure'

epilepsy before it starts (Fig. 1). An important goal is to identify/diagnose 'at risk' patients, and treat them so as to disrupt the underlying processes of epilepsy development. Much of this research is driven by the idea that epilepsy development 'recapitulates'—at least in some important processes—early developmental processes. Thus, there is much interest in the signaling molecules that guide the development of neuronal connections and of mature electrophysiological properties. Indeed, it is the plasticity of these processes, even in the adult brain, that have become primary targets in research and in treatment development.

4. *Epilepsy and Brain Plasticity*

What are the 'plastic' processes that give rise to epileptic conditions? As suggested above, many investigators believe that they are the same processes that we have identified as the basis for higher brain functions, such as learning and memory. We can define 'plasticity' as the ability of the brain to respond (structurally and/or functionally) to new stimuli, with adaptive behavior. Such plasticity is one of the primary functions of the highly developed cortical regions of the mammalian brain, where we believe that learning takes place—and where (perhaps not coincidentally) most seizures are generated. Investigators have elucidated many processes that are involved in such complex neural function, including: *long-term potentiation* (and depression) of synaptic interactions; *sprouting* of new connections; changes in the number and *subunit composition* of receptors and channels; *proliferation* and/or *death* of neural and glial elements; and (of course) the up- (or down-) *regulation of the genes* that code for protein products involved in these changes (Fig. 2) (Engel et al. 2000). These are all normal brain mechanisms. They are all implicated in genesis of epileptic conditions. The epileptic state is characterized by a loss of regulation or control of these plasticities. A little plasticity is critical for higher brain function, but too much plasticity leads to pathology.

See also: Brain Development, Ontogenetic Neurobiology of; Electroencephalography: Basic Principles and Applications; Electroencephalography: Clinical Applications; Kindling, Neural Basis of; Neural Plasticity

Bibliography

Berkovic S F, Scheffer I E 1999 Genetics of the epilepsies. *Current Opinions in Neurology* **12**: 177–82

Commission on Classification and Terminology of the International League Against Epilepsy 1981 Proposal for revised clinical and electroencephalographic classification of epileptic seizures. *Epilepsia* **22**: 489–501

Commission on Classification and Terminology of the International League Against Epilepsy 1989 Proposal for revised classification of epilepsies and epileptic syndromes. *Epilepsia* **30**: 389–99

Delgado-Escueta A V, Wilson W A, Olsen R W, Porter R J (eds.) 1986 *Jasper's Basic Mechanisms of the Epilepsies*, 3rd edn. Lippincott Williams & Wilkins, Philadelphia

Dodson W E, Pellock J M (eds.) 2000 *Pediatric Epilepsy, Diagnosis and Therapy*, 2nd edn. Demos, New York

Engel J Jr 1989 *Seizures and Epilepsy*. F.A. Davis, Philadelphia

Engel J Jr, Pedley T A (eds.) 1998 *Epilepsy: A Comprehensive Textbook*. Lippincott-Raven, Philadelphia

Engel J Jr, Schwartzkroin P A, Moshe S L, Lowenstein D H (eds.) 2000 *Brain Plasticity and Epilepsy*. Academic Press, San Diego

Heinemann U, Engel J Jr, Avanzini A, Meldrum B S, Mouritzen-Dam A, Wasterlain C (eds.) 1996 *Progressive Nature of Epileptogenesis*. Elsevier, Amsterdam

Nehlig A, Motte J, Moshe S L, Plouin P (eds.) 1999 *Childhood Epilepsies and Brain Development*. John Libbey, London

Sanes D H, Reh T A, Harris W A 2000 *Development of the Nervous System*. Academic Press, New York

Schwartzkroin P A, Moshe S L (eds.) 2000 Pediatric epilepsy. *Mental Retardation and Developmental Disabilities Research Reviews* **6**(4)

Schwartzkroin PA, Moshe S L, Noebels J L, Swann J W (eds.) 1995 *Brain Development and Epilepsy*. Oxford University Press, New York

Spreafico R, Avanzini G, Andermann F (eds.) 1999 *Abnormal Cortical Development and Epilepsy*. John Libbey, London

P. A. Schwartzkroin

Episodic and Autobiographical Memory: Psychological and Neural Aspects

Episodic memory is the type of memory that makes possible the recollection of personally experienced events. During an act of retrieval from episodic memory, an individual is said to be remembering the past. The term is closely related, but not identical, to autobiographical memory, which refers to memory for and about a person's own life.

The relation between episodic and autobiographical memory can be illustrated through the case study of a man, known as K.C., who became profoundly amnesic in 1981 following a serious closed head injury (Tulving et al.1991). As a result of the injury, K.C. suffered a complete loss of episodic memory, as he cannot recollect a single thing that has ever happened to him. Despite extensive questioning, he cannot remember even things that should be highly memorable, such as the accidental death of a brother to whom he was close. It follows that K.C. has a

drastically reduced ability to recount his autobiography. Yet he can report autobiographical facts, such as his date of birth, the names of many of the schools he attended, the make and color of a car he once owned, and the fact that he has spent many summers at his parents' cottage. K.C. is unable to *remember* any aspects of his life, although he *knows* many important facts about his past. In healthy individuals, autobiographical memory encompasses both what can be recollected episodically, and self-relevant knowledge that can be expressed through semantic memory.

1. A Psychological History

Psychological writing on the topic of episodic memory began with William James in his classic *Principles of Psychology* (James 1890). To James, the concept of *memory* was equivalent to what is now called episodic memory. The defining characteristic of memory was the conscious recollection that accompanied the retrieval of personally experienced events. 'Memory requires more than the mere dating of a fact in the past. It must be dated in *my* past… I must think that I directly experienced its occurrence.' (p. 612). Thus James distinguished between acts of episodic memory and other manifestations of learning and memory (semantic memory, classical and emotional conditioning), which do not involve conscious recollection of the personal past. Later philosophers and psychologists emphasized the role of such personal memories in defining an individual's personal identity and autobiography (Shoemaker 1959).

Although James provided a strong foundation, there was very little research directed at memory for personally experienced events for the next several decades. Psychological research became dominated by the idea that observable behavior, and not conscious experience, was the appropriate subject matter for the field. The study of learned responses was not compatible with the study of the remembering of experiences.

Contemporary investigations began when Tulving (1972) proposed the distinction between episodic and semantic memory. This distinction can be characterized in two different ways, and it is important for these meanings to be differentiated and the meaning used in a particular context to be made clear. One distinction has to do with a particular kind of memory task or memory performance, while the other concerns different neurocognitive memory systems in the human brain. Episodic memory was originally defined as the variety of memory that receives and stores information about temporally dated episodes or single events, and the temporal–spatial relations among them, while semantic memory comprised organized knowledge about the world, including information about words, objects, and events. The episodic/semantic distinction introduced the possibility that the remembering of single, personal events should be studied separately from factual knowledge. For many years, the study of memory for single events was classified as episodic memory research and it proceeded largely independently of research on semantic memory. Even today, the prototypical laboratory list-learning task in which participants are exposed to study material and then tested for what they have learned is often classified as an episodic memory task.

Over the intervening years, it became clear that the 1972 definition of episodic memory was no longer satisfactory. Performance of a so-called episodic memory task, such as free recall, can depend not only upon the type of conscious recollection that was described by William James, but also on general knowledge about single events from the past. Episodic memory now refers to a neurocognitive system that renders possible the conscious recollection of events as they were previously experienced. The defining characteristic of episodic memory is its dependence on a special kind of conscious awareness called *autonoetic awareness*. An individual with autonoetic (or 'self-knowing') awareness is capable of roaming at will in subjective time, by recollecting aspects of past experiences, or imagining future experiences through anticipations and daydreams. Semantic memory is defined as a neurocognitive system that makes possible the acquisition, retention, and use of factual information. The retrieval of semantic information is accompanied by *noetic awareness*, which is experienced when one thinks about something one knows. Noetic awareness does not involve any feeling of reliving the personal past; it represents knowledge about some aspect of the world.

While episodic remembering is necessarily autobiographical, a person's autobiography can be expressed through systems of memory that are not episodic. The final criterion is the individual's phenomenal feeling, not the contents, or veridicality, or age, or confidence of what is retrieved. Therefore, it is possible for individuals to recall autobiographical aspects of their lives, even single events in the recent past, via noetic awareness only.

2. Neural Bases

Research in cognitive neuroscience has made significant progress in identifying the brain structures that support episodic and autobiographical memory. Most of the relevant knowledge about the relation between neuroanatomy and memory has been gleaned from the study of patients with circumscribed brain injuries. Especially important has been the identification of patients with selective losses of episodic or autobiographical memory, with a relative sparing of other varieties of memory, such as semantic memory. The

integrity of two brain regions, the medial temporal lobes and the prefrontal cortex, are critical for the normal operations of episodic memory. The former area is necessary for the establishment of episodic and autobiographical memories, and also participates in their retrieval for a limited time following encoding. The latter area underlies the ability to become autonoetically aware of the personal past and future.

2.1 Medial Temporal Cortex and Episodic Memory

It is now well established that several structures in the medial temporal cortex have a critical role in both the encoding and retrieval of episodic memory. This region, often referred to as the *medial temporal lobe system*, comprises a number of anatomically connected and functionally interrelated structures, including the hippocampi, entorhinal, perirhinal, and parahippocampal cortices. Midline diencephalic structures, such as the medial thalamus and the fornix, are also included in this system.

The importance of the medial temporal cortex in episodic memory was established in one of the classic case studies in cognitive neuroscience (Scoville and Milner 1957). In an experimental surgery, a patient known as H.M. underwent bilateral hippocampal excision with the goal of curing his severe epileptic seizures. While the surgery was effective in the prevention of seizures, it had the unintentional side-effect of rendering the patient amnesic for all postsurgical events. To this day, H.M. cannot remember anything that has happened to him since the surgery, a condition known as *anterograde amnesia*. Over the intervening years, many other patients with damage limited to the medial temporal lobe system have reinforced the relation between this brain region and anterograde amnesia for episodic and autobiographical memories.

The medial temporal lobes receive highly preprocessed inputs from multiple, diverse cortical and subcortical brain regions. This makes the hippocampal system well-placed anatomically to bind or conjoin the inputs, and combine them into a single complex event or episode that could subsequently be stored and remembered (Cohen and Eichenbaum 1993). When the hippocampi and related structures are damaged, affected patients are unable to form a 'memory trace' from the various aspects of an ongoing event; it follows that the event cannot be subsequently remembered.

The hippocampal system participates in the retrieval of episodic memories for a limited time after encoding. Patients with anterograde amnesia typically are unable to remember events that occurred just before they became amnesic. The extent of this *retrograde amnesia* varies between patients yet there is typically a temporal gradient in which memory for recent episodes is affected more than that for remote episodes. As time passes after encoding, episodic and autobiographical memories gradually become reorganized, or consolidated. Over a period of several weeks or months, recollection becomes mediated by distributed areas in the neocortex.

Structures in the medial temporal lobes are critical for the encoding and retrieval of both episodic and semantic memory. The majority of patients with amnesic syndromes have an equivalent or roughly equivalent impairment in their ability to autonoetically recollect their past (episodic memory) as well as noetically know about the past (semantic memory). Yet some recent reports have described a group of amnesic patients with profound losses of episodic memory, with little or no deficit in semantic memory (Vargha-Khadem et al.1997). Two distinctive features of the pathology were that: (a) it was confined solely to the hippocampus itself, while sparing the surrounding cortex, and (b) it was sustained very early in life, before the patients had the opportunity to build a base of autobiographical memories. While the patients are unable to remember any of their ongoing experiences, they have made satisfactory progress through school, and assembled a large body of semantic knowledge. The patients have some autobiographical knowledge, as they *know* many facts about their lives, although they cannot *remember* anything that has happened to them. An increasing body of knowledge suggests that episodic memory processes depend upon the hippocampus proper, while semantic memory is mediated largely by the perirhinal cortex in the temporal lobes (Aggleton and Brown 1999), although a definitive resolution of this proposal remains elusive.

2.2 Prefrontal Cortex and Autonoetic Awareness

The operations of episodic memory are dependent upon a sophisticated variety of conscious awareness, one that arguably does not exist in nonhuman animals. Autonoetic awareness refers to the ability to consider one's subjective experiences in the past, present, and future. It is the human prefrontal cortex, defined as those aspects of the frontal lobes anterior to motor and premotor areas, which subserves the capacity for autonoetic awareness.

An extensive literature in clinical neuropsychology indicates that patients who have suffered injuries restricted to the prefrontal cortex often suffer from impoverished awareness of their own mental life. Though patients are fully awake, alert, and have full command of intelligence and language capabilities, they often appear either unwilling or unable to engage in self-reflection or introspection (Luria 1973). Reflection upon the personal past defines episodic memory. Multiple lines of research converge to show that the integrity of the prefrontal cortex is necessary for successful episodic recollection of past experiences. Although patients with large prefrontal lesions are not amnesic— they can learn from experiences and update their knowledge—recalled information does not have the subjective, personal quality that comprises epi-

sodic recollection. Similarly, as patients report about autobiographical experiences, they often discuss their own lives as if they are casual observers (Wheeler et al.1997). The prefrontal cortex confers humans with the ability to recollect prior experiences in a personal, subjective way. It follows that, when patients with prefrontal lesions attempt to recount their autobiography, it is the episodic component which is compromised, much more than the semantic component.

Because of its reliance upon autonoetic awareness, episodic memory is related to other cognitive skills and behaviors that are not typically considered to be memory capabilities. Autonoetic awareness allows for self-reflection extending into both the past and future. Patients with large prefrontal lesions typically lack ambition, foresight, and initiative, and this brain area has been considered as important for 'memories of the future' (Ingvar 1985). As such, episodic memory shows a close family resemblance to some sophisticated mental achievements, like introspection, anticipation, and daydreaming, which require individuals to reflect upon personal experiences, past or present, that are not part of the immediate perceptual environment.

See also: Declarative Memory, Neural Basis of; Learning and Memory, Neural Basis of; Memory: Autobiographical; Memory, Consolidation of; Working Memory, Neural Basis of

Bibliography

Aggleton J P, Brown M W 1999 Episodic memory, amnesia and the hippocampal–anterior thalamic axis. *Behavioral and Brain Sciences* **22**: 425

Cohen N J, Eichenbaum H E 1993 *Memory, Amnesia, and the Hippocampal System*. MIT Press, Cambridge, MA

Ingvar D H 1985 'Memory of the future': An essay on the temporal organization of conscious awareness. *Human Neurobiology* **4**: 127–36

James W 1890 *Principles of Psychology*. Holt, New York

Luria A R 1973 *The Working Brain*. Basic, New York

Scoville W B, Milner B 1957 Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry* **20**: 11–12

Shoemaker S S 1959 *Journal of Philosophy* **56**: 868–82

Tulving E 1972 In: Tulving E, Donaldson W (eds.) *Organization of Memory*. Plenum, New York, pp. 381–403

Tulving E, Hayman C A G, Macdonald C 1991 Long-lasting perceptual priming and semantic learning in amnesia—a case experiment. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **17**: 595–617

Vargha-Khadem F, Gadian D G, Watkins K E, Connelly A, Van Paesschen W, Mishkin M 1997 Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**: 376–80

Wheeler M A, Stuss D T, Tulving E 1997 Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin* **121**: 331–54

M. A. Wheeler

Equality and Inequality: Legal Aspects

Law's role in the pursuit of equality remains a focal point of attention and controversy in both theory and practice. That inequality persists in many significant domains—employment, housing, education, wealth—is uncontroversial. However, the explanations and solutions for this continuing inequality have become more complex and elusive. Legal approaches developed to address first generation types of inequality—deliberate exclusionary treatment based on race, gender, or other identity classifications—are inadequate to address more complex, interactive, and structural forms of inequality. As the experiences and causes of inequality become more subtle and multifaceted, consensus over the role of law in addressing inequality has also fractured.

Many of the recurring disagreements in case law, commentary, and politics proceed from the clash between theories of equality. These theories in turn reflect differences in the assumptions about the causes of inequality, the appropriate balance between inequality and competing values, and the proper role of courts in addressing social problems. This essay lays out four conceptual and legal approaches to equality, focusing on the employment context as a case study. The employment context in the United States has become a proving ground for other areas of discrimination law and practice. It features a wide range of conceptual approaches to inequality. As such, it provides a useful vehicle for mapping the terrain of argumentation about legal approaches to inequality. (See *Discrimination*; *Affirmative Action: Comparative Policies and Controversies*; *Affirmative Action: Empirical Work on its Effectiveness*; *Injustice: Legal Aspects*.)

1. Formal Equality of Treatment

The dominant conception of equality in American antidiscrimination law is embodied in the concept of 'colorblindness.' The law pursues equality by eliminating race, gender, or other suspect classifications as a legitimate basis of making decisions. It focuses on the state of mind of the decision maker, and targets intentional reliance on race or gender as the violation of the equality norm. It is also process oriented, in the sense that it is the taking account of race or gender in the treatment of employees that constitutes both the harm and the wrongful conduct.

This paradigm mirrors the social and political context that led to the adoption of civil rights legislation. In the period preceding the enactment of Title VII, deliberate exclusion of women and people of color from private employment was lawful and even, for some, morally acceptable. Many companies had employment policies that segregated or excluded workers based on race and gender (Abrams 1989, Fiss

1976, Crenshaw 1988). Supervisors and co-workers displayed open hostility or bias to workers based on their sex or race, and fired, refused to hire, or segregated workers based on those attitudes.

The formal equality model, and its implementing principle of antidiscrimination, is closely linked to these factual and moral conditions. Within that paradigm, discrimination occurs when an actor with formal power over subordinates in the workplace hierarchy deliberately uses that power against a particular subordinate because of attitudes, motivations, or desires relating to sex or race.

This model defines the wrongful conduct as intentional, unequal treatment based on race, gender, or some other class-based characteristic that is viewed as arbitrary. For inequality to count as legally relevant under this theory, it must be demonstrably caused by this form of deliberate, class-based exclusion. The premise of this model is that these class characteristics are irrelevant to the underlying decision, and that taking class membership into account is thus arbitrary and unfair (Fiss 1971, Fallon 1984). This premise embraces the idea that employers may, and usually do, properly differentiate among employees based on qualifications relevant to effective job performance. By taking race or gender into account, employers unjustifiably depart from merit-based decision making. This violates basic notions of fairness, because people deserve to be treated 'on the merits,' based on an individualized assessment of their capacities (Fallon 1984, Gutmann 1996). Exclusion based upon characteristics that are beyond the control of the affected individual are viewed as particularly morally problematic (Fiss, Brest 1976). Decisions based on class membership thus offend dignitary values, in addition to undermining instrumental values of efficiency.

The goal of the formal equality model is 'colorblindness': employment decisions should be made without regard to employees' race or gender. Race, gender, and other categories of status operate as fixed, universal, and static categories based on membership in a set, unchanging group identified by predetermined, and often physical, characteristics such as skin color. The assumption is that group status is, by definition, irrelevant to decisions about capacity to do the job, that considering class status is by nature arbitrary, denigrating, and unfair, that decision makers can make decisions without noticing or taking account of racial and gender characteristics, and that equality can be achieved by a rule that prohibits employers from considering race or gender in making employment decisions.

This theory of equality provides remedies to individuals who have been subjected to a process tainted by racial or gender considerations. It is decidedly ahistorical and individualistic in its focus, and compensatory in its explicit remedial approach. Many who embrace a formal equality approach believe current racial inequality results from factors unrelated to racial discrimination. They question the legitimacy and efficacy of legal norms to address this form of inequality. (Browne 1993, Rabkin 1998, Sowell 1984). Perhaps the most vehement expression of this view in the US can be found in Justice Scalia's concurring opinion invalidating voluntary affirmative action in government contracting based upon its inconsistency with the principle of color blindness: 'To pursue the concept of racial entitlement—even for the most admirable and benign of purposes—is to reinforce and preserve for future mischief the way of thinking that produced race slavery, race privilege, and race hatred. In the eyes of government, we are just one race here. It is American' (*Adarand Constructors vs. Pena* 115 S. Ct. 2097 [1995]). (See *Affirmative Action: Comparative Policies and Controversies*; *Affirmative Action: Empirical Work on its Effectiveness*)

1.1 Critiques of Formal Equality

The dominance of formal equality in American antidiscrimination law has produced a wide array of criticism. One set of critiques questions the colorblindness prohibition on grounds of autonomy and freedom of association, a view expressed by Richard Epstein (1992). Other criticism proceeds from the view that colorblindness is an inadequate response to continuing inequality. Most fundamentally, critics point out that a formal equality theory leaves in tact a legacy of state-produced inequality and subordination. It ignores the role that group status continues to play in shaping the life expectancies, opportunities, and experiences of group members. It locks in the continued effects of past formal exclusion, imposing the consequences of that continued exclusion on the excluded groups themselves. By defining equality solely in terms of the conduct of the perpetrator, a formal equality approach normalizes conditions of marginalization and exclusion, and the group dimensions of these material conditions. It does not begin to address the perspective of those who experience exclusion, marginalization, and unequal enjoyment of resources and status (Freeman 1990, Crenshaw 1988, Gotanda 1991). At the same time, 'It conceals the distributive consequences of group-salient practices in a semantic code that defines instrumental rationality as race-neutral' (Siegel 2000). A moral theory that might be suitable for a just society, i.e., one that does not systematically disadvantage people based on their race, loses its moral premise for a society that is unjust in those ways (Gutmann, 1996, Sunstein 1994).

On a conceptual level, David Strauss has pointed out the logical inconsistencies of colorblindness as an antidiscrimination principle. To protect a designated group against treatment targeting that group, the state must pay attention to and indeed specially protect the interest of that group. Thus, colorblindness is contra-

dictory. Strauss also points out that the metaphor of colorblindness really doesn't fit its use in the non-discrimination context. Colorblindness in the literal sense prevents an individual from seeing color; colorblindness as an equality model prevents an individual from acting on what he or she sees. This approach presumes that individuals can see race but keep it from influencing their perceptions, conduct, and decisions (Strauss 1986).

Recent work in the social sciences challenges the assumption (which underlies the formal equality model) that conscious decision making adequately describes the current dynamics of exclusion and unequal treatment. Bias is produced not only as a result of motivation but also as a product of racial or gender schemata, multiracial group interactions, and patterns of conflict resolution or avoidance (Krieger 1995, Lawrence 1987, Sturm 1998). Bias or exclusion often results from interactive patterns of non-inclusion, as much as from exclusion motivated by bias (Donnelon and Kolb 1994, Wilkins and Gulati 1996). Members of nondominant groups may well be treated differently based on their race, notwithstanding the decision maker's lack of intention to do so.

Historians have also challenged the theory of formal equality for its ahistoricism. Colorblindness ignores historically constituted patterns of group inequalities that are 'socially pervasive (articulated across social domains), and socially persistent (articulated over time)' (Siegel 2000). Examining race and gender inequality from a historical standpoint reveals relations of group inequality embedded in the social organization of work, reproduction, and sexuality. 'In short, when considered from a historical standpoint, discrimination has no transcontextual or fixed form' (Siegel 2000).

Robert Post applies this structural account of the dynamics of bias to the law itself. Using what he calls a sociological account of the field, he contends that law does not and cannot cause employers to make judgments about qualifications in a way that is totally race and sex blind. Legal actors are themselves engaged in a social practice of shaping the norms and conventions by which sex and race are given social meaning. 'Antidiscrimination laws always begins and ends in history, which means that it must participate in the very practices that its seeks to alter and regulate' (Post 2000). To pretend otherwise is to allow courts to rely on unarticulated assumptions and premises that legitimate some racial or gender meanings and challenge others.

2. *Functional Equality of Treatment*

A modified approach to equal treatment attempts to take account of the more structural, interactive, cognitive, and unconscious character of bias in the second generation of discrimination. This theory provides a framework that continues to account for inequality as a result of unintentional as well as intentional biased treatment. Functional equality focuses on whether differences in treatment based on group membership produce unequal outcomes, regardless of whether a policy or specific intention explains that unequal treatment. It broadens the focus of analysis from formal and deliberate decisions to the social context and practices that produce decisions. Differences in treatment between majority and minority group members that are arbitrary, in the sense that they are not justified by criteria of economic rationality, violate this principle of equality. These differences in treatment are morally problematic because they are unfair. Fair opportunity requires that 'equal consideration should be given to all qualified candidates so that candidates are chosen on the basis of their qualifications, where qualifications are set that are relevant to the legitimate social purposes of the position in question' (Gutmann 1996). If differences in treatment are attributed to stereotyping, cognitive bias, or differential access to the networks that structure advancement, they violate this fairness principle, even if the decision maker did not consciously intend that result. The harm is not only the stigma or loss of dignity that results from substituting group characteristics for individual assessment. It is also the difference in access that results from patterns of exclusion that emerge from social interaction or cognitive bias (Krieger 1995, Sturm 1998).

This form of inequality is often difficult to identify at the level of individual decisions. For this reason, functional equality of treatment approaches often focus on patterns of unequal treatment affecting groups of similarly situated individuals. Unexplained disparities in treatment of similarly situated members of majority and minority groups are problematic under this approach (*Teamsters vs. United States*, 431 US 324 [1977]). Fault, in the sense of conscious intent to exclude based on irrelevant and suspect classifications, is not necessarily a criterion for discrimination. Thus, functional equality theories depart from morality-based conceptions of discrimination and move to a focus on remediation.

2.1 *Critiques of Functional Equality*

The move away from fault-based notions of discrimination and formal notions of equality involves the courts in more subtle, complex, and interactive forms of bias. Critics of this move have questioned whether this is an appropriate role for the law in general and for courts in particular. If cognitive bias is widespread, indeed, ubiquitous, is it appropriate to rely on the law to change embedded patterns of interaction (Krieger 1995)? Others have questioned

the viability of a model of discrimination that, by its nature, can only be detected by looking at patterns involving groups over time (Browne 1993, Rabkin 1998). Do courts, and employers for that matter, have the capacity to prevent and redress bias at the level of cognition and group interaction? However, theories of proof and remediation have developed that operate effectively at the group and structural level. These theories shade into a third conceptual approach to inequality: equal achievement.

3. *Equal Achievement or Equal Outcome*

This concept of equality focuses on differences in the material conditions of members of particular groups. It is rooted in social reality, and looks at the distribution of jobs, income, and status of members of groups that have previously been subjected to systematic exclusion and subordination (Fiss 1976, Fallon 1984, Freeman 1990). The focus of this theory is the group, rather than the individual who has been excluded as a result of group status. Groups continue to matter because they define the opportunity structures affecting their members. For these theorists, group status must be directly addressed because it is the level at which inequality operates. For some, groups also have independent normative value, reflected in the patterns of cultural and political meaning that emerge among individuals who identify as a group (Gotanda 1991). These theorists justify the focus on groups not only in instrumental terms, but also as an independent value.

The empirical premise of equal outcome or equal achievement is that current inequality continues to be the result of systematic and legally sanctioned subordination. Legal intervention is justified to remedy persistent and arbitrary differences in resources, social status, and opportunity based on group membership. The current system for allocating opportunities continues to reflect biased criteria that systematically disadvantage non-dominant groups. Thus, group-based remediation compensates for departures from merit-based decision making (Fiss 1971, Harris and Narayan 1994). The goal of this approach to equality is outcome oriented: disparities in group status should be reduced or eliminated.

The normative justifications for the equal achievement theory are largely compensatory and remedial. Past group-based subordination justifies current group-based remediation. This theory rests on the premise that the group continues to have social meaning. The group continues to experience harm that is a consequence of the history of group-based exclusion, and members of the group experience the current manifestations of that harm. For this reason, matching harm to individual injury is thus both unnecessary and counterproductive (Crenshaw 1988). Moreover, group-based remediation is necessary to overcome barriers to participation that are perpetuated through individual, neutral processes that do not compensate for the effects of past discrimination (Siegel 2000, Abrams 1989).

Other justifications for the achievement theory of discrimination are more pragmatic: bias is difficult to prove and process rules are easy to evade. Outcome measures serve as an adequate proxy for bias that continues to affect access to fair process and opportunity. If a process is fair and race or gender neutral, it will produce a result that more or less mirrors the composition of the pool. Outcome measures that strive for a workforce that mirrors the racial or gender composition of the pool simply accomplish directly what may be more difficult to achieve through reforming the decision-making process. These measures also provide an easy-to-implement remedy that responds to concerns about judicial capacity to detect and correct patterns of bias (Fiss 1971, Gutmann 1996).

3.1 *Critiques of Outcome Theories of Equality*

Adherents of formal equality criticize outcome-based theories as inconsistent with the principle of race neutrality. Some critics take the view that any approach that treats race or gender as a salient category, even for remediation, reinforces the continued significance of those categories and is thus discriminatory. They also question the appropriateness of remedial strategies that focus on groups. This skepticism proceeds in part from an increasing willingness to question the empirical assumption that current inequality is the result of past legal subordination. Since the 1970s, some scholars have developed alternative explanations for persistent inequality that emphasize characteristics of the racial or gender group, rather than patterns of exclusion or marginalization (Glazer and Moynihan 1975, Sowell 1984). Once the causal link between current economic or social inequality and legal exclusion or subordination is questioned, that skepticism undercuts the perceived legitimacy of group-based remedies. Indeed, some critics of outcome theories of equality question whether racial or gender status continues to have any meaning at the level of the group. Some conservative critics thus suggest that group-based remediation is just another form of racial or gender-based decision-making, and is thus illegitimate (Belz 1991, Abram 1986). Finally, they treat outcome-based equality theories as inconsistent with principles of merit-based decision-making and thus contrary to both efficiency and dignitary values (Browne 1993).

Outcome theories have also been criticized by progressives as an inadequate framework to define equality policy because they rest so heavily on pragmatic considerations and lack an adequate normative foundation. Outcome theorists typically view their strategies as contingent and temporary. They do not generally defend group-based redistribution as an end

in and of itself, but rather as a means to an end. Nor do they reveal what it was about current practice that is objectionable, other than the result. The absence of a clearly articulated normative justification invites challenges to the legitimacy of any efforts to address racial or gender bias. Outcome approaches, if used as the exclusive theory justifying equality initiatives, may thus stunt the refinement, rearticulation, and reformation of the normative meaning of racial and gender participation and exclusion. They may also reinforce the perception that equality and efficiency are competing goals (Sturm 2001, Sturm and Guinier 1996).

4. *Equal Access*

Equal access theories focus on inequalities that result from unfair criteria and structures of interaction. They target decision-making that is formally fair but functionally biased in favor of members of a particular group. The central question is whether the structure, process, or criteria of decision-making arbitrarily favors members of one group over another for reasons unrelated to the articulated goals underlying the decision (Sturm 2001).

The most widely articulated conception of equal access theory accepts the moral legitimacy of differentiating among applicants based on meaningful differences in their qualifications to carry out the functions of a position. Reasonable disagreements exist, however, concerning the social purposes relevant to a particular job, the relative importance of multiple purposes, and the best way to assess which applicants meet those qualifications. 'Those members of society who currently have the power to set qualifications may tend to value those qualifications that favor people similar to themselves' (Gutmann 2000). Moreover, these power differentials shape the structures for assessing individuals' qualifications and opportunities for advancement. These standards and structures are no more likely to produce valued social outcomes than others that do not exclude members of non-dominant groups. As such, these exclusionary standards and practices are unfair, and thus discriminatory (Strauss 1998).

The classic case of *Griggs vs. Duke Power Co.* (401 US 424 [1971]) lays out this theory through a fable: 'Congress has now provided that tests or criteria for employment or promotion may not provide equality of opportunity merely in the sense of the fabled offer of the milk to the stork and the fox.' If the goal of the employer is to find out who can drink milk, it is arbitrary and unfair to make that determination by evaluating all animals' capacity to drink milk from a saucer. Discrimination thus occurs when an employer relies upon conditions, criteria, or practices that favor members of a particular group, if that advantage is arbitrary in the sense that it does not advance the articulated goals of the employer.

This theory defines discrimination as conduct that deprives individuals or groups of meaningful access to employment opportunities through the use of standards, criteria, or structures that fail to minimize the expression of bias, or that arbitrarily advantage one group over the other (Friedman 1986). It builds on the practical insight that discrimination and inequality often result from arbitrary, unstructured, and unaccountable decision-making. It uses outcome measures not to define the harm, but to signal the likelihood that an employment process is biased, and thus to trigger scrutiny of the adequacy and fairness of the criteria and their application. The beneficiaries of this theory are both individuals, who are denied the opportunity to compete based on unbiased selection processes, and groups, which are systematically disadvantaged in their access to jobs.

Equal access theory has thus emerged as an integrative approach to discrimination that accepts the inevitability, if not the legitimacy, of merit-based distinctions but introduces structural and group notions of fairness to the question of how to define and implement merit-based decision making.

4.1 *Critiques of Equal Access Theory*

Equal access theory has been called both over- and under-inclusive. Its potential for radically restructuring employment decision-making prompts challenges based on the intrusiveness of such a theory into the autonomy of employers. This criticism often proceeds from the view that courts should not engage in the process of institutional restructuring and act illegitimately if they move beyond the role of dispute resolution and compensation for discrete individual harms. The under-inclusiveness critique challenges the acceptance of merit or market criteria as a morally defensible basis for distinguishing between the opportunities of potential employees. The weaker version of this critique, preferred by outcome theorists, argues that the continued pervasiveness of a caste system so infects the distribution of opportunities as to delegitimate a system of merit that produces racially identifiable results (Freeman 1990, Strauss 1998). The stronger version of this critique challenges the moral legitimacy of differentiating access to social goods based on any arbitrarily distributed characteristics, including talents and motivation. 'There is no more reason to permit the distribution of income and wealth to be settled by the distribution of natural assets than by historical and social fortune … . From a moral standpoint the two seem equally arbitrary' (Rawls 1974).

The differences in approach to equality summarized in this essay recur in and help to explain deep fractures in employment discrimination doctrine, legislative policy making, and public discourse about equality. Current debates often pose formal equality of treat-

ment and equal outcome as the only two available conceptions of equality, and force a choice between them. The increasing complexity of the dynamics of bias, along with the decreasing political and judicial viability of outcome theories of equality, have increased the attention to and importance of access theories of equality. These theories are, however, the least normatively and empirically developed. Emerging scholarship in the area of employment discrimination has begun to fill this crucial gap. It remains to be seen whether the law will respond dynamically to these new understandings of the dynamics and meaning of workplace inequality.

See also: Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Civil Liberties and Human Rights; Discrimination; Discrimination, Economics of; Discrimination: Racial; Equality of Opportunity; Equality: Philosophical Aspects; Gender, Class, Race, and Ethnicity, Social Construction of; Human Rights in Intercultural Discourse: Cultural Concerns; Injustice: Legal Aspects; Justice and its Many Faces: Cultural Concerns; Justice, Access to: Legal Representation of the Poor; Prejudice in Society; Race and the Law; Racism, History of

Bibliography

Abram M 1986 Affirmative action: Fair shakers and social engineers. *Harvard Law Review* **99**: 1312–25

Abrams K 1989 Gender discrimination and the transformation of workplace norms. *Vanderbilt Law Review* **42**: 1184–1248

Aleinikoff T A 1990 Case for race consciousness. *Columbia Law Review* **91**: 1060–1125

Andrews M 1999 *The Political Economy of Hope and Fear: Capitalism and the Black Condition in America*.

Barthlett K T, Harris A P 1998 *Gender and Law: Theory, Doctrine, Commentary*, 2nd edn.

Belz H 1991 *Equality Transformed: A Quarter-Century of Affirmative Action*

Brest P 1976 In defence of the antidiscrimination principle. *Harvard Law Review* **90**: 1–54

Browne K 1993 Statistical proof of discrimination: Beyond 'Damn Lies'. *Washington Law Review* **68**: 477–558

Crenshaw K 1988 Race, reform, and retrenchment: Transformation and legitimation in antidiscrimination law. *Harvard Law Review* **101**: 1331–87

Edsall T B, Edsall M D 1992 *Chain Reaction: The Impact of Race, Rights, and Taxes on American Politics*

Epstein R 1992 *The Case Against Employment Discrimination Law*

Fallon R 1984 Firefighters v. Stotts: Conflicting models of racial justice. *Supreme Court Review* **1**: 1–25

Fiss O 1971 A theory of fair employment laws. *University of Chicago Law Review* **38**: 235–314

Fiss O 1976 Groups and the equal protection clause. *Philosophy & Public Affairs* **5**: 107–77

Freeman A 1990 Antidiscrimination law: The view from 1989. *Tulane Law Review*. **64**: 1407

Friedman J 1986 Redefining equality, discrimination, and affirmative action: The access principle. *Texas Law Review* **65**: 41–99

Glazer N, Moynihan P (eds.) 1975 *Ethnicity: Theory and Experience*. Harvard University Press, Cambridge, MA

Gotanda N 1991 A critique of 'Our constitution is color-blind'. *Stanford Law Review* **44**: 1–68

Gutmann A 1996 Responding to racial injustic. In Appiah A, Gutmann A, *Color Conscious: The Political Morality of Race*

Krieger L H 1995 The content of our categories: A cognitive bias approach to discrimination and equal employment opportunity. *Stanford Law Review* **47**: 1161–1248

Lawrence C 1987 The id, the ego, and equal protection: Reckoning with unconscious racism. *Stanford Law Review* **39**: 317–88

Oliver M L, Shapiro T 1995 *Black Wealth, White Wealth: A New Perspective on Racial Inequality*

Post R 2000 Prejudicial appearances: The logic of American antidiscrimination law. *California Law Review* **88**: 1, 17

Rabkin J 1998 Racial divisions and judicial obstructions. In: Devins N, Douglas D M (eds.) *Redefining Equality*

Rawls J 1974 *A Theory of Justice*, pp. 74–5

Siegel R 2000 Discrimination in the eyes of the law: How 'color blindness' discourse disrupts and rationalizes social stratification. *California Law Review* **88**: 77, 82, 101

Sowell T 1984 *Civil Rights: Rhetoric or Reality?*

Strauss D 1986 The myth of colorblindness. *Supreme Court Review* **99**

Strauss D 1998 The illusory distinction between equality of opportunity and equality of result. In: Devins N, Douglas D M (eds.) *Redefining Equality*

Sturm S 1998 Race, gender, and the law in the twenty-first century workplace: Some preliminary observations. *University of Pennsylvania Journal of Labor and Employment Law* **2**: 639

Sturm S 2001 Second generation employment discrimination: A structural approach. *Columbia Law Review* **101**: 458–568

Sturm S, Guinier L 1996 The future of affirmative action: Reclaiming the innovative ideal. *California Law Review* **84**: 953–1036

Sunstein C 1994 The anticaste principle. *Michigan Law Review* **92**: 2410

Wilkins D, Gulati G M 1996 Why are there so few Black lawyers in corporate law firms? An institutional analysis. *California Law Review* **493**: 493–614

Williams J 1987 *Eyes on the Prize: America's Civil Rights Years, 1954–1965*. Penguin, New York

S. Sturm

Equality of Opportunity

Equality of opportunity is a concept that has been used frequently in the sociology of education, and in particular in discussions of social class differences in education, although the key ideas are applicable in many other fields. In the sociology of education the focus has been on whether there are equal opportuni-

ties for children with the same measured ability, or from different social class backgrounds in gaining access to higher or more privileged institutions such as grammar schools or universities. The concept could be applied equally well, however, to the unequal opportunities faced by boys and girls, or by members of different ethnic minorities, and it could be applied to questions of access to other institutions, such as the professions or the legislature. There are two main versions of the concept: a narrower sense of *equal treatment* in the competition for access for people with the same relevant abilities, and a broader sense of *equal rates of access* for people with different ascribed characteristics.

The narrower version of the concept has strong normative connotations. Equality of opportunity has frequently been taken to be a normative ideal in education and in liberal-democratic societies generally. A motion submitted to the British Trade Union Congress in 1896, for example, urged that 'our education system should be completely remodeled on such a basis as to secure the democratic principle of equality of opportunity' (quoted in Silver 1973, p. 11). Equality of opportunity has often been contrasted with other possible conceptions of social justice, such as equality of outcome, which would typically be associated with a socialist rather than a liberal-democratic social order.

As a normative ideal, equality of opportunity can be seen as a special case of the notion of fairness—that people with the same relevant traits or competencies should be treated in the same way. In the particular context of education, equality of opportunity frequently has been taken to mean that people with equal academic aptitude or ability should be given equal access to advantaged sectors of education. This was the dominant interpretation in the first half of the twentieth century, exemplified by the work of Gray and Moshinsky (1938) in Britain. Similarly in America, Lloyd Warner and his colleagues wrote 'we might speak of equality of educational opportunity if all children and young people exceeding a given level of intellectual ability were enabled to attend schools and colleges up to some specified level. This is the only practicable kind of equality of educational opportunity' (Warner et al. 1946).

Historically, the term was used frequently in the context of access to grammar-school education in Britain: the fact that affluent parents were able to secure fee-paying places at the prestigious grammar schools, places that were denied to equally able children of less affluent parents, was taken to be manifestly inequality of opportunity. More generally, one can think of any discriminatory procedure for regulating access to advantaged institutions as constituting inequality of opportunity.

In this respect, inequality of opportunity also becomes an explanatory concept as well as a normative ideal. Thus, fee-paying becomes one possible explanation for the observed class differences in access to grammar schools. The inability to afford fees is a possible mechanism that might explain why children of equal ability had unequal rates of access. (In practice, however, the abolition of fee-paying after World War II did not lead to a reduction in class inequalities, suggesting that fee-paying may not have had the explanatory role assumed by pre-war scholars; see Halsey et al. 1980.)

This explanatory sense of equality of opportunity implies a voluntaristic model of action where individuals are confronted with opportunities that they may choose whether or not to exercise. Inequality of educational opportunity implies that there are barriers, such as fee-paying or discriminatory selection procedures, which prevent some appropriately qualified individuals who wish to gain access from actually doing so. This interpretation lends itself to a rational choice account of educational decision-making, as in the work of Boudon (1974), and Erikson and Jonsson (1996).

However, in the post-war sociological literature, inequality of opportunity was used frequently as a synonym for, rather than an explanation of, class differences in access to privileged institutions. Thus, for example, the term 'inequality of opportunity' has been applied to any class difference in educational attainment regardless of the explanatory mechanism that might have generated that inequality. This approach has thus moved away from the earlier conception of equal opportunities for those with equal abilities, which would be consistent with some observed class differences in access, to a broader definition that ignores the unequal class distribution of measured ability.

The broadening of the concept in this way probably reflected a growing sociological distrust of the concept of IQ or measured ability. The broader interpretation could be justified on the grounds that the class distributions of ability and of preferences for education are themselves socially conditioned. This can therefore be thought of as a sociologically deterministic model.

More generally, the broader conception was congruent with the distinction between ascribed and achieved characteristics that became a focus of a great deal of post-war sociological research and theory. Discussions of inequality of opportunity could thus be integrated with mainstream sociological enquiry. Social class origin is an ascribed characteristic, and inequality of educational opportunity could be redefined as the persistence of ascribed characteristics in influencing educational outcomes. While much broader than the earlier conceptions, this interpretation was still distinct from that of equality of outcome in that it was consistent with inequalities between individuals.

Whether this broader conception serves much useful purpose, however, is unclear. Once the focus on

unequal treatment is removed, both the normative and the explanatory roles of the concept can be questioned. Moreover, since there is already an established distinction between ascribed and achieved characteristics, it is doubtful whether much is to be gained by broadening the concept so that it merely duplicates the notion of ascription. It may be better to preserve the original emphasis on equal treatment and to retain the concept as an explanatory rather than a purely descriptive one.

See also: Class: Social; Educational Institutions and Society; Equality and Inequality: Legal Aspects; Income Distribution; Inequality; Inequality: Comparative Aspects; Sex Differences in Pay; Sex Segregation at Work; Social Inequality in History (Stratification and Classes); Wealth Distribution

Bibliography

Boudon R 1974 *Education, Opportunity and Social Inequality: Changing Prospects in Western Societies*. Wiley, New York

Erikson R, Jonsson J O 1996 *Can Education be Equalized? The Swedish Case in Comparative Perspective*. Westview, Boulder, CO

Gray J L, Moshinsky P 1938 Ability and opportunity in English education. In: Hogben L (ed.) *Political Arithmetic: A Symposium of Population Studies*. Allen & Unwin, London

Halsey A H, Heath A F, Ridge J M 1980 *Origins and Destinations: Family, Class, and Education in Modern Britain*. Clarendon Press, Oxford, UK

Silver H (ed.) 1973 *Equal Opportunity in Education: A Reader in Social Class and Educational Opportunity*. Methuen, London

Warner W L, Havighurst R J, Loeb M B 1946 *Who Shall be Educated? The Challenge of Unequal Opportunities*. Kegan Paul, Trench and Trubner, London

A. F. Heath

Equality: Philosophical Aspects

The ideal of equality is that everyone shall have the same. Different versions of the ideal require that people have the same rights, or status, or opportunity, or condition, or whatever it is deemed important to equalize. Since being equal in one respect can force inequality in other respects, different conceptions of equality are rivals. Currently ideals of equality are many and various, and also controversial.

1. Lockean Equality: Equality as Nondomination

Locke (1690) affirms that all men are 'naturally in' a '*state of perfect freedom* to order their actions, and dispose of their possessions and persons, as they think fit.' Locke adds that this natural state is 'also one of *equality*, wherein all the power and jurisdiction is reciprocal, no one having more than another.' Being of the same 'species and rank,' men should be 'equal one amongst another without subordination or subjection.' Call this equality as nondomination. According to this Lockean ideal, all adult humans have equal natural rights to live as they choose so long as they do not wrongfully harm others. This implies that no person ought to find herself placed without her consent in a relationship in which she is subject to the will of another and forced to obey the other's commands. Much ink has been spilled puzzling over what might be meant by stating as Locke does that men are 'naturally in' a state of equal freedom. The puzzles dissolve if one construes the natural right to equal freedom as a moral right that each person possesses independently of institutional arrangements or customary beliefs in the society the person inhabits. The institutional arrangements in a particular society may include slavery, and customary beliefs may presume that slavery is perfectly justified, but nonetheless one may hold that each person has a moral right to equal freedom which implies that others are duty bound not to enslave her.

The Lockean claim of human equality conflicts with views that people may legitimately be born to a state of subjection to others as occurs in a feudal or caste society. This claim also seems to conflict with the view that at least if one is born and raised in a tolerably just society one is legitimately subject to the authority of its government in the absence of any consent one has given to this just authority. The claim also seems to conflict with the view that there are forms of subjection that are not morally legitimate and that should be viewed as violating the Lockean right of equality properly construed even if they secure the voluntary consent of all parties to the relationship. This is to claim that some right of equal freedom is inalienable (nontransferable) and nonwaivable (the right-bearer cannot release those bound by the duty to comply).

2. Democratic Equality

The Lockean idea that all adult persons have equal rights to liberty and property opposes the idea that by birth some persons legitimately acquire authority and power over others. But as stated the Lockean ideal does not require equality of political status, and is compatible with (for example) an hereditary monarchy, so long as the king respects everyone's Lockean rights. A further and independent requirement of equality is equality of democratic citizenship—that the coercive political order of government should be subject to popular control, with each qualified adult having an equal right to stand for political office and to vote in free and fair elections. Rousseau (1968) argues for democratic equality on the ground that only

in an ideal form of democracy can individuals submit to political authority while being autonomous in the sense of obeying only laws they give themselves and avoiding subjection to the will of another person.

Democratic equality can be interpreted more expansively. In this broad sense, democratic equality prevails in a society to the degree that each adult person has equal power to influence the outcome of political processes, or alternatively, to the degree that all adult persons with equal personal ability to influence others' opinions have equal power to influence political outcomes.

3. *Equality of Opportunity*

In a caste society, accidents of birth and social connections strongly influenced by birth tend to determine who enjoys positions of special privilege and above-average advantages. Desirable niches in the hierarchical social structure may be reserved for men rather than women, for those of noble lineage rather than commoners, for those who are members of a favored religion, clan, tribe, or skin color grouping.

3.1 *Formal Equality of Opportunity*

The ideal of formal equality of opportunity or careers open to talents challenges the legitimacy of caste privilege. Applied to places in educational institutions, formal equality of opportunity requires that such places be open to any applicants, and that applications be judged on their merits, with criteria of merit being selected so they are predictive of applicants' ability and ambition to succeed at the course of study to be undertaken. (It is consistent with formal equality of opportunity that schools charge tuition and fees and that among applicants selected on the basis of merit, only those sponsored by someone willing and able to pay tuition and fees are actually admitted to school.) The formal equality of opportunity ideal can be adapted to apply to competitions for desired positions in business enterprises and state bureaucracies and to bank loans made available to would-be entrepreneurs.

Formal equality of opportunity constrains the freedom of property owners to do whatever they might choose with property they legitimately own. At least, if formal equality of opportunity is enforced by coercive legal rules or social norms backed by informal sanctions, the freedom of property owners is to that extent limited. The ideal of formal equality of opportunity is compatible with any degree of inequality in the extent to which individuals have the opportunity to develop their native talents into skills that would render them qualified in meritocratic competitions. Hence, it could happen that in a society regulated by equality of opportunity all good jobs go to the children of the wealthy and well educated.

3.2 *Fair Equality of Opportunity*

A public school system funded by general taxes exhibits a commitment that goes beyond formal equality of opportunity in the direction of assuring to all individuals at least some minimally adequate opportunity to become trained and to acquire skills that will render them qualified for meritocratic competitive posts. The ideal implicit in the idea of education funded by the public, pressed to its limit, becomes substantive or fair equality of opportunity. This norm is satisfied when all have equal opportunity to develop their native talents into the skills needed for meritocratic success. If better-off parents provide special education and socialization for their children so that they have special opportunities to become qualified, these advantages are somehow exactly offset, in the society of fair equality of opportunity.

Rawls (1999) espouses fair equality of opportunity and specifies that in the society that satisfies this norm any two individuals drawn at random from the population who have the same ambition and the same native talent will have the same prospects of competitive success for posts that confer advantages. Ambition here is a desire to attain some class of advantageous positions coupled with the willingness to work hard to become qualified for meritocratic competition. Having the same native talent as another is having the same disposition to develop skill given an identical environment. This latter idea is inherently fuzzy, as one sees by noting that one person might develop greater skill than another under educational regimen A and lesser skill than the other under educational regimen B. The notion of ambition embedded in the idea of fair equality might also provoke unease, as when it is noted that a society in which all girls develop the ambition to be homemakers and all boys develop the ambition to have satisfying and lucrative employment could satisfy fair equality even though all meritocratic competitions are won by men. But the general intent of the Rawlsian norm is clear. One might regard the norm of fair equality of opportunity as the ideal of a classless society, in which being born of privileged parents does not affect in the slightest an individual's prospects for acquiring a privileged position in adult life except via the mechanism of inheritance of genes and via parental and other social effects on the development of ambition. Evidently the norm is intended to be at best very roughly approximated, because precisely measuring the degree to which a society deviates from fair equality as defined here is not possible even in principle.

4. *Equality of Condition*

In a democratic society with fair equality of opportunity prevailing, some people could end up leading grim, unenviable lives while some are extra-

ordinarily prosperous, and some of the prosperous may lead wonderful, enviable lives. This could occur whether the economy consists of state-owned or privately-owned firms. The discussion to follow focuses on a competitive market economy with private ownership.

Some would hold that equality of democratic citizenship along with equality of opportunity is equality enough, and that inequalities in life prospects that emerge in a market economy over time are not morally troublesome. Others disagree. Seminal works exploring this issue include Rawls (1999) and Nozick (1974).

4.1 Pure and Impure Egalitarianism; Equality vs. Priority

Call someone who claims that it would be morally desirable if the material condition of people were to become more nearly equal an egalitarian. This view takes many forms. One important distinction is between those who regard material equality as contributing to an ideal that is morally desirable but morally optional and those who hold on one or another ground that closer approximation to material equality is morally required. This distinction cross-cuts another, between those who believe that it is morally appropriate for the state to enact coercive social policy aimed at equality of condition and those who deny that such use of state power can be morally permissible. One might be an egalitarian either because one believes (a) that equality is instrumental to attaining some further moral goal or partially constitutive of some further moral goal or because one believes (b) that equality of some sort is *per se* morally desirable. The impure egalitarian has a reason of type (a) for affirming equality, the pure egalitarian a reason of type (b).

One should also take note of a position that may be confused with egalitarianism but is in fact distinct from it. The egalitarian holds that it is better if all have the same level of benefit. The prioritarian on the other hand holds that it is morally better to bring about a gain in benefit for someone, the lower that person's lifetime benefit level would be absent the particular gain in question. In short, the prioritarian gives priority to helping the worse off. Everyone's having the same, or closer to the same, is not a goal for the prioritarian (see Parfit 1995, Temkin 1993).

An egalitarian advocate, whether pure or impure, is better regarded as proposing a component of fundamental moral principle, not an immediate guideline for laws and public policies. Public policies should be set in terms that are operationalizable, feasible to administer. A society would presumably do better by its fundamental values if it does not attempt to enact them as public policies but instead chooses policies that seek to achieve feasible goals that are reasonable proxies for the fundamental values.

Equality is in one respect a clear notion: given some way of measuring people's condition, their condition is equal when everyone's is the same according to this measure. But when people's condition is unequal, the notion of equality itself does not provide a measure that ranks different possible distributions and specifies which are more or less equal than others and by what extent. Philosophers discussing equality in recent years have not paid much attention to this issue (but see Temkin 1993, and Sen 1992).

4.2 Equality among Whom?

The articulation of an ideal of equality requires a specification of the persons or entities whose condition should be rendered equal. One aspect of this issue is whether the relevant units to be equalized are individuals or social groups. Some might hold that equality of condition should hold across ethnic, supposed racial, or religious groups, or across groups living in different regions of a society. The discussion to follow simply assumes that individual human persons are the relevant units for an ideal of equality of condition.

Equality of condition might be deemed desirable to sustain either (a) among the inhabitants of each political society taken one at a time or (b) among all human persons alive anywhere. Those who favor equality within each political society but not across political societies are sometimes labeled 'patriotic egalitarians' and those who favor equality across political societies are labeled 'cosmopolitan egalitarians.' A further dimension of the 'equality among whom?' specification is time. An egalitarian might give priority to attaining a closer approximation to equality among persons alive now or among persons who shall live now and in the future. (Policies that affect population growth and economic growth will affect the degree of inequality obtaining among persons living at different times.)

Another aspect of 'Equality among whom?' is whether equality is supposed to obtain in the condition of persons at each moment of time, or in the lifetime overall condition of persons, or perhaps among age groups at each moment of time. Notice that inequality between young and old is consistent with equality among all persons over the life course.

4.3 Equality of What?

Under this heading theorists debate how to measure people's condition for purposes of determining who qualifies as better off and who worse off when policies to enhance the extent to which equality is achieved are being considered or applied. (But it is evident that this issue of the basis of interpersonal comparison must be addressed by a wide class of theories of distributive

justice, not just those that affirm equality of condition. Anyone who espouses the notion that government should redistribute from better off to worse off persons needs some metric for deciding who is better off and worse off, and the metric had better make sense or else the ethic of redistribution will be unpalatable.)

Perhaps the most obvious candidate measure is equality of money or, more broadly, equality of material resources. On the simplifying assumption that material resources can be divided into homogeneous lots, one implements this ideal by dividing all such resources into N identical lots, one for each of the N persons in the group whose members are to be made equal.

Equality of material resources might be thought problematic on at least two grounds. The first difficulty is that this ideal leaves intact and unremediated the inequalities in the talents and useful personal traits possessed by individuals. If equality of material resources obtains between two persons, otherwise similar except that one is born legless and armless, the handicapped individual must devote a large portion of her resource allotment to compensate for the lack of abilities that normal legs and arms provide. One's resources for achieving one's ends include personal resources as well as material resources. Excluding the former may seem an arbitrary limitation.

A second difficulty is that equality across individuals' holdings of material resources over their entire lives might be thought to leave too little role for personal responsibility in the proper determination of each person's holdings. One might seek an interpretation of the ideal of equality of condition that does not require society continuously to offset in the name of equality the foreseeable impact of the risky and conservative, wise and foolish, careless and prudent choices made by individuals as they pursue their individual life plans.

This concern about responsibility might regard it as instrumentally valuable (holding people responsible for their choices is a useful means to attain other goals) or as intrinsically morally desirable (it is better in itself if benefits go to the responsible and deserving rather than to the irresponsible and undeserving).

Responding to these difficulties, Dworkin (2000) has proposed that people should be made equal in the level of uncourted disadvantage they suffer in their lives. The idea is that people should be made equal in their unchosen circumstances so far as is feasible, but they should not be compensated for differences among them that stem from their different ambitions and preferences and their choices of action to achieve their individual aims in life. Accordingly, equality of resources is expanded to encompass personal resources and adjusted to integrate a notion of individual responsibility for how one's life goes.

The family of views that interprets the ideal of equality of condition as equality of resources of some sort is united in rejecting the family of doctrines that prizes equality of utility, or well-being, or welfare across persons. This welfarist family of views supposes that when one person's condition is assessed as better or worse than another's for purposes of distributive justice the relevant measure is how well the person's life is going, the goodness of the life for the person who is living it. The goodness of a life in this sense is different from the sense in which the life of a person who sacrifices herself for others might well be deemed noble or admirable or good. Such a life is good (admirable), but not by virtue of that fact good for the person whose life it is. Good in the welfarist sense is what a person aims at when she is being rationally prudent.

Resourcist theories of distributive justice take it to be an important principle that society should not seek to make people equal in their personal well-being. Two considerations support this principle. One is that reasonable people will persist in disagreeing about what is ultimately valuable or worthwhile in human life. What people owe to each other by way of equal justice cannot depend on agreement on some conception of human good, because no such agreement can be forthcoming. Hence individual well-being cannot serve as a standard for determining whether someone is better off or worse off. The second consideration is that there should be a division of responsibility between society and the individual, such that if each individual is allotted equal resources according to the appropriate conception, then what individuals do with their equal resources is their business, not society's.

The ideal of equality of welfare varies in content as it is filled in with different conceptions of welfare or well-being. On subjective conceptions, whether something is good for a person and to what degree depends on her attitudes and opinions regarding it. A subjective conception might identify individual welfare with desire satisfaction or life aim fulfillment or enjoyment. On objective conceptions, whether something is good for a person and to what degree is fixed independently of her subjective attitudes and opinions regarding it. One might identify individual welfare with entries on an objective list, such that the greater the extent to which the person gains items on the list, rated by their relative importance, the better her life is going. To give a welfarist conception the best run for its money, the most plausible conception of welfare must be used to fill in its content.

The welfarist responds to the claim that reasonable people cannot agree about what renders a life intrinsically valuable for the one who is living it partly by insisting that the claim greatly underestimates the degree of consensus on the nature of the human good that does exist among reasonable people and partly by allowing that to the extent that there is genuinely reasonable disagreement about the human good, to that extent there is only limited commensurability in this domain.

The responsibility objection against the ideal of equality of welfare has prompted the reply that the welfarist egalitarianism can accommodate a sensible norm of personal responsibility (Cohen 1989). But when one tries to work out the details it becomes clear that integrating an appropriate norm of personal responsibility into any ideal of equality of condition is so far an unsolved problem.

Consider this revision of equality of welfare: Society is responsible for bringing about an arrangement that induces equality of opportunity for welfare among individuals. Roughly, the level of opportunity for welfare provided by one's circumstances is the level of welfare one would reach if one conducted oneself as prudently as can reasonably be expected over the course of one's life. A prudent individual acts efficiently to advance his or her own welfare within the limits of legal and moral constraints. One determines whether a person has acted as prudently as can reasonably be expected by relativizing the requirement of prudence to that individual by taking account of the person's unchosen deficits in choice-making and choice-executing talent.

Equality of resources could be revised to become equal opportunity for resources, an ideal that is satisfied among a group of persons just in case all would have the same resources (or alternatively, the same expected level of resources) if all behaved as prudently as can be expected. As stated, both equal opportunity views may seem too unforgiving in the way they incorporate personal responsibility into an egalitarian ideal. If one fails to meet the specified standard of prudent conduct by even a tiny amount, there is no constraint on the extent to which one's fortunes may plummet without triggering egalitarian concern.

Another approach to interpreting distributive equality supposes that what people owe to each other by way of providing all opportunities for a choice-worthy life is a set of freedoms (Sen 1992, Nussbaum 1999). Freedom might seem just another way of referring to opportunity, but identifying these two approaches would be a mistake. For one thing, since an individual's opportunity for welfare provided by her circumstances is the welfare level she would reach if she behaved as prudently as could reasonably be expected, with less freedom the individual who has difficulty being prudent could have more opportunity for welfare.

The freedom metric for distributive justice is calibrated in terms of effective not merely formal freedom. The distinction is that one is formally free to do *X* if no law forbids the doing of *X* and no one wrongfully interferes with one's attempts to do *X*. One is effectively free to do *X* provided that if one chooses to do *X*, and genuinely tries to do *X*, one will do *X*.

The ideal of equality of effective freedom is indeterminate pending some specification of significant or valuable freedoms, the ones that should be equally secured for persons. To fix the content of the ideal, it is necessary to identify valuable modes of action or ways of being such as being well nourished, having adequate shelter, having relationships of friendship and love, sustaining close family ties, attaining significant work achievement, experiencing enjoyment in one's activities and avoiding useless pain, participating in the political life of one's community, gaining understanding of the world, and so on. Call these various doings and beings 'functionings,' and identify equality of freedom with the effective freedom or capability to function in these valuable ways.

Since individuals differ in their traits, they will require different packages of education and resources in order to get (close to) equal effective freedom to function. For the effective freedom approach, what is morally important is not the level of resources that one has access to *per se* but what one is enabled to do and be by means of these resources.

The equality of freedom approach limits the responsibility of society to provision of real freedom for individuals. The level of actual functionings reached is a 'don't care' from this standpoint. For example, society provides conditions in which each can exercise religious liberty, but whether and how religious freedom is exercised and the resultant spiritual life that people experience is a matter of culture, not justice as equality.

Equality of effective freedom or capability to function as stated does not adopt a stance on personal responsibility. Suppose that an individual has an initial effective freedom to function in some significant way, but squanders the resources that assured the freedom, and when equivalent resources that are supplied again, squanders them again. At some point in this cycle one might suppose that equality of effective freedom should not be sustained further.

However, how to integrate a sensible account of personal responsibility into an egalitarian distributive ethic is far from settled. No account asserted in current discussions has attracted a consensus of support.

4.4 Doubts about Whether Equality of Condition Is Morally Desirable

The question arises, why should everyone have the same? What makes equality of condition morally desirable? Some theorists suspect the answers are that nothing renders equality of condition *per se* desirable.

One seed of doubt about the intrinsic value of equality is planted by noting that if one can achieve equality or a closer approximation to it by worsening the condition of those who are better off without improving anyone's condition, from the standpoint of equality one must hold that this leveling down is desirable. This implication of equality is expressed in the slogan, 'Equality itself is as well pleased by graveyards as by vineyards.' But a change from the status quo that worsens someone's condition without

making anyone's condition better in any respect cannot be a change for the better in any respect, some think. This thought rejects the idea that equality of condition is per se desirable.

Asserting the intrinsic moral significance of equality of condition, theorists sometimes note that the wealthy in a prosperous world control enormous amounts of resources while the impoverished of the world lead grim lives in horrible conditions. But a plausible rejoinder (Frankfurt 1987) is that it is evidently not the gap between rich and poor *per se* that is morally troublesome but rather the sheer badness of the lives that the poorest people on the earth lead. For if the gap between rich and poor were *per se* significantly objectionable then the same-sized gap between super-rich and merely rich should be equally objectionable. But if the latter gap does not seem especially morally problematic, then the relative position of rich and poor cannot be regarded as *per se* especially morally problematic either.

5. *A Puzzle about Basic Equality Status*

An egalitarian holds that people should be made equal in some aspect of their condition on some moral ground. What justifies equality among persons on any conception? It is often supposed that nonhuman animals, even if their interests should be considered in determinations of action and policy, do not count for as much as humans. This discounting of animals strikes many as plausible, but what underlies it? The usual suggestion is that normal adult humans have cognitive capacities that render them intrinsically better than animals, so their interests and welfare should count for more. The puzzle then is that normal humans differ significantly in their possession of the cognitive capacities that are supposed to render all humans fundamentally equal in their basic moral status. If an ordinary gorilla has a lesser moral status than an ordinary human because the latter is smarter, why does not the ordinary human have lesser moral status than another human who is much smarter? It might be urged that what makes all persons equal and superior to other animals is that all normal humans have cognitive capacities at some threshold. Having cognitive capacities above the threshold does not confer superior moral status. But why and how this is so is not yet clear.

See also: Democracy; Egalitarianism: Political; Egalitarianism, Sociology of; Equality of Opportunity; Human Rights, History of; Law and Democracy; Locke, John (1632–1704)

Bibliography

Anderson E A 1999 What is the point of equality? *Ethics* **109**: 287–337

Arneson R 1997 Equality and equal opportunity for welfare. In: Pojman L, Westmoreland R (eds.) *Equality: Selected Readings*. Oxford University Press, Oxford, UK, pp. 229–42

Cohen G A 1989 On the currency of egalitarian justice. *Ethics* **99**: 906–44

Cohen G A 1992 Incentives, inequality, and community. In: Peterson G (ed.) *The Tanner Lectures on Human Values*. University of Utah Press, Salt Lake City, UT, Vol. 13, pp. 261–329

Dworkin R 2000 *Sovereign Virtue: The Theory and Practice of Equality*. Harvard University Press, Cambridge, MA

Frankfurt H 1987 Equality as a moral ideal. *Ethics* **98**: 21–43

Kagan S 1999 Equality and desert. In: Pojman L, McLeod O (eds.) *What Do We Deserve*? Oxford University Press, Oxford, UK, pp. 298–314

Kolm S C 1972 *Justice et equite*. Centre National de la Recherche Scientifique, Paris

Locke J 1690 *Second Treatise of Government*. Hackett Publishing, Indianapolis, IN

Mill J S 1977 *Considerations on Representative Government*. In: Robson J J (ed.) *Collected Works*. University of Toronto Press, Toronto, ON, Vol. 19, pp. 371–577

Nagel T 1991 *Equality and Partiality*. Oxford University Press, Oxford, UK

Nozick R 1974 *Anarchy, State, and Utopia*. Basic Books, New York

Nussbaum M 1999 Women and cultural universals. In: Nussbaum M (ed.) *Sex and Social Justice*. Oxford University Press, Oxford, UK, pp. 29–54

Parfit D 1995 *Equality or Priority*? Department of Philosophy, University of Kansas, KS

Rakowski E 1991 *Equal Justice*. Clarendon Press, Oxford, UK

Rawls J 1999 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Raz J 1986 *The Morality of Freedom*. Oxford University Press, Oxford, UK

Roemer J E 1996 *Theories of Distributive Justice*. Harvard University Press, Cambridge, MA

Rousseau J J 1968 *The Social Contract* (trans. Cranston M). Penguin, Harmondsworth, UK

Rousseau J J 1988 *The Basic Political Writings* (trans. Cress D A). Hackett Publishing, Indianapolis, IN

Scanlon T M 2000 The diversity of objections to equality. In: Clayton M, Williams A (eds.) *The Ideal of Equality*. Macmillan, Basingstoke, UK, pp. 11–59

Sen A 1992 *Inequality Reexamined*. Harvard University Press, Cambridge, MA

Tawney R H 1952 *Equality*, 4th edn. Allen and Unwin, London

Temkin L 1993 *Inequality*. Oxford University Press, Oxford, UK

Walzer M 1983 *Spheres of Justice: A Defense of Pluralism and Equality*. Basic Books, New York

R. J. Arneson

Equity

Equity is usually defined as fairness, impartiality; justness. Consequently, Equity in Law is said to exist when justice administered between parties appears as based upon natural reason or ethical judgement—whether or not it is in conformity with written (State) law. Indeed, equity can be considered either as a

concept immanent to the law or as a set of principles outside the law, the function of which is to allow the judge to give soft interpretations in concrete cases in order to reach not too unmerciful solutions. Hence, equity is working at large in legal systems where the judge plays an important role, like common law or other systems where state law has a reduced role, whilst the official use of it is very insignificant in other legal systems, such as positivist continental laws. Equity in Law can be examined under two main headings. Equity in Law has always been considered (a) *ethically* as an opportunity to refer to equity against positive (State) Law, and (b) *politically* as the way to open law to necessities borne from economic developments.

1. Equity as a Common Expedient

In this sense, equity is positioned between law and ethics. In the Western legal tradition, equity was even culturally perceived as the only foundation of law. Equity and justice go hand in hand in Western legal philosophy. But the Roman root *aequus* was translated by two words which acquired two very distinct meanings through the ages: 'equity' on the one hand, and 'equality' on the other. The distinction led to opposing political theories on law and justice.

Roman Law defined legal work as skillfulness in settling what is good and fair: *Ars boni et aequi* (Celsius and Ulp, *Dig*. I, 1, 1, pr.). Among traditional legal principles, we find also the rule according to which equity must be expected in everything, but above all in Law (Paul, *Dig*., 50, 17, 90). As Pretorian law, a great part of everyday Roman legal regulation was progressively built upon equity in order to complete positive law. The Roman distinction between *ius honorarium* and *ius civile* has been seen by some authors as a forerunner of what would later exist in England as a distinction between common law and equity: a 'body of rules existing alongside the original civil law, founded on distinct principles, and claiming incidentally to supersede civil law in virtue of a superior sanctity inherent in these principles' (Maine 1861, p. 25).

Equity is used to replace legal rules drawn from statute or common law when these rules are considered too inadequate or unyielding for the expectation of justice in a specific case. Hence, equity was always considered as a common expedient against legal (state) justice where this latter appeared too distant from ethics, moral or customs. This route to justice was historically employed to such a degree that the gambit was embodied in many legal traditions, re-emerging from time to time according to need.

In opposition to legal unfairness, equity is regarded as either a refuge (equity prompts legal rules to law givers) or a recourse (equity is the alternative source of a redress or remedy to those provided by established law).

From ancient times, equity was considered to be an efficient means to establish justice in law. Without equity, law would be unjust (*Summum ius*, *summa injuria*). Within legal anthropology, Henri Sumner Maine showed in his work *Ancient Law* (1861) that equity is considered an instrument to bring harmony to progressive societies. According to this author, equity is one of the three agencies by which primitive law has been adapted to the requirements of a developing society and an improving morality.

Equity was also considered in Western legal tradition as a complement or a corrective to positive law. There are a number of aphorisms on this topic, such as *Hæc æquitas suggerit*, *etsi deficiamur jure* or *Æquitas nihil aliud est quam jus quam lex scripto prætermisit*, i.e., equity must be followed where there is no legal norm to reach a solution. But when the man in the street talks of 'equity,' he is thinking of ideal justice which may even be contrary to the law. Here lies that faith in the fairness of a claim to some natural law existing above positive laws or statutes. Equity focuses on the imperfections of law and can even be preferred to a legal norm where it provides a useful corrective to legal institutional justice by improving an equitable outcome: *Placuit in omnibus rebus præcipuam esse justitiæ æquitatisque quam stricti juris rationem*. Equity is better than a solution drawn from a strict interpretation of law. Such a theory, elaborated from Aristotle onwards, runs through the western legal tradition through Roman Law, subsequent Canon Law (Grossi 1998) and right through continental legal regimes before the French revolution.

That law can be brought into harmony with society through equity is currently demonstrated by legal sociologists, whose research illustrates specific circumstances in which societies make appeals to alternative procedures where equity plays an important role. This is particularly clear when it is not ingrained within a culture to instinctively call on the law for help as the main way to resolve conflicts, such as in traditional Chinese or Japanese cultures. Moreover, equity is often grasped as an opportunity for a workable alternative to state law, where it is not envisaged by a written legal norm. In the last decades of the twentieth century, for instance, equity was cited in alternative movements by both Italian and Brazilian judges (*Alternative Use of Law* and *Alternative Justice*, in Arnaud 1993).

2. Equity as a Means of Adapting Legal Regulation to Political Prerequisites for the Development of an Economy

All through the building of the occidental legal tradition, whilst laws were considered a means to protect communities and insure their reproduction following a specific pattern, equity also appeared as

the way to open up law to necessities born from the development of trade. Here, equity appears (a) as an earlier technique to create a society for merchants, and (b) in the last decades of the twentieth century, as a means to implement a global legal order.

2.1 Equity as a Foundation of Trading Societies

At the end of the Middle Ages, the opposing pressures of the wish of kings to concentrate absolute powers in their hands, the opening up of everyday life to trade and merchants, as well as the emergence of a *bourgeoisie*, led the various national laws to adapt themselves to the development of landed and propertied interests. Western laws gradually expanded into opposing systems: the common law and equity tradition on the one hand and continental legal positivism on the other.

After centuries of legal rationalization, equity was definitively taken up from the discourse of ethics and located as one of the foundations of law, appealing to natural law and reason, even as against positive law and statutes (Hugo Grotius, Samuel Pufendorf, Christian Wolff and their disciples in Northern Europe; Jean Domat in France).

As to the continental romano-canonic tradition, equity was quickly made subservient to state law. Despite the influence of Domat's work on French jurisprudence, equity was forbidden in France by the *Ordonnance civile* of 1667. The French Revolution of 1789 was partially directed against the discretionary power of judges. A project to reintroduce equity as a tool in judicial interpretation in the French Civil Code was drastically cut down by Bonaparte. Finally, the judge was considered merely to be the mouthpiece of state law. It is true that we can find counterexamples in nineteenth-century France, such as judicial decisions by the 'Bon Juge Magnaud' who used equity to avoid enforcing the law against the poor. But this use of equity was a rarity and almost all such decisions were reversed by the court of appeal.

We can also identify some slight remains of equity in Civil Law, even in the French Civil Code (articles 565, 1135, 1854). The strongest examples are, however, drawn from the Swiss Civil Code, articles 1 and 4: 'The law must be applied in all cases which come within the letter or the spirit of any of its provisions. Where no provision applies, the judge shall decide according to the existing customary law, and, failing which, according to the rules which he would lay down if he had himself to act as legislator. Herein he must be guided by approved legal doctrine and case-law … Where the law expressly leaves a point to the discretion of the judge, or directs him to take circumstances into consideration, or to consider whether a ground alleged is material, he must base his decision on principles of justice and equity' (see Lloyd and Freeman, p. 1231).

At the same time, in the common law countries, legal developments went in the opposite direction. There, indeed, equity refers to that part of the law of England which is derived not from the custom of the realm nor the enactments of Parliament, but from the decisions of the old Court of Chancery, consolidated in 1873, along with the other superior courts in the supreme court of judicature. Such equity pertains to a right recognized by a court of equity through a specific procedure. Here, we are faced with procedural equity and not 'natural justice.' Jurisdiction in equity is one of the three clearly marked divisions into which the jurisdiction of the court falls, together with a miscellaneous jurisdiction and a common law jurisdiction. The equitable jurisdiction of the court originated in the power which the chancellor assumed in the Middle Ages to modify a strict and formal legal system on moral grounds.

There are a number of principles used as guidelines when it is necessary to make appeal to this notion of equity; equity acts *in personam*; equity will not suffer a wrong to be without a remedy; equity follows the law; equity looks to the intent rather than the form; equity imputes an intent to fulfil an obligation; equitable remedies are discretionary; delay defeats equities; he who comes into equity must come with clean hands; equity takes into account the balance of convenience; where there are equal equities, the law prevails and the first in time prevails; equity, like nature, does nothing in vain; equity never wants a trustee; equity aids the vigilant; equality is equity; equity will not assist a volunteer; equity will not permit a statute to be a cloak for fraud.

In England, because the remedies which the common law courts could give were initially limited by the forms of writs available, men turned to the Chancellor for equity for assistance. The two systems came to be administered by wholly independent tribunals whose decisions were frequently in conflict. The early jurisdiction of the Court of Chancery was in its nature extremely vague. From the beginning of the seventeenth century, it is possible to discern certain general principles in the exercise of the court's jurisdiction. The early seventeenth century was the principal turning point in the history of the court, for the common law lawyers were seeking to prohibit the Chancery from interfering with common law rights. King James I resolved the conflict between the two jurisdictions in favor of Chancery. From this time onward, although the conflict did not die down until the end of that century, the court gradually developed jurisdiction co-ordinate with that of the common law courts and on settled principles. In the eighteenth and nineteenth centuries, it came to follow precedent as strictly as they.

The principal distinction between the two systems was that, whereas the rules of common law were supposed to have been established since time immemorial, those of equity could in almost every case

be traced to the Chancellor who invented them. In the nineteenth century the jurisdiction of Chancery, now clearly defined, was frequently classified according to its relationship to the common law as being exclusive, concurrent or auxiliary. By the middle of the twentieth century, the Chancery division consisted of the Lord Chancellor and judges divided into groups which were respectively assigned matters within the peculiar jurisdiction of the division.

In the USA, equity was developed with relative consistency during the nineteenth century. The expansion of the equity system was subject to statutory grants bestowed to each state, which constitutes a specific feature of American legal history. During the Civil War, the equitable jurisdiction was generally attributed to courts of double jurisdiction. Since 1848, many states excluded the distinction between law and equity in favor of one 'civil action.' At last, there was a complete fusion of law and equity by the adoption, in 1938, of the federal rules of civil procedure, promulgated by the US Supreme Court under authority of an Act of Congress.

2.2 Equity as a Referent in the Framework of Contemporary Political Theory

Recently, the use of equity was proposed by neoliberal legal philosophers as a way to put into common practice some fundamental legal principles. Hence, equity has become a referent in the framework of globalization, re-emerging in many new areas of legal work, e.g., private, regional and international judicial assemblies, and directly in relation to global governance. The importance of the recourse to equity in contemporary neoliberal legal and political philosophy is due to (a) the predominance of the Anglo-American legal tradition in international legal relations, (b) the special focus given to the above-mentioned principle according to which 'equality is equity,' and finally (c) the claim for a minimal state.

After much controversy surrounding the Aristotelian idea of 'corrective justice,' equity was reintroduced as the legal principle of equal treatment of people in the allocation of legal benefits and obligations. John Rawls designates as *'equality of fair opportunity'* the equity of opportunities which cannot be reduced to a merely hypothetical right (Rawls 1971, pp. 14, 46). Here arises a paradox: the more the State intervenes in establishing equality of opportunities, the more it stands apart from the ideal of contractualism which characterizes the minimal state. Consequently, there could be no coherent basis for imposing common global lawful and unique interpretations (for example, of human rights). This means that it would not be possible to universally expect the same behavior through all institutions, whether national, local, or international.

Popper's claim for the importance of tradition, a tradition which offers the moral structure of a society, is grounded in the idea that herein lies the social meaning of justice and equity. On the one hand, tradition reveals the degree of moral feeling which each society attains; on the other, should there be conflict between opposing interests, tradition is the moral framework which allows us to find fair or equitable compromises (Popper 1963, Chap. 17, paragraph 3). Institutions alone give a society its legal structure appropriate to this moral background.

According to Friedrich Hayek, society is developing upon a set of individual interests pursued, although not always fully recognized, in the framework of the 'open society' or 'great society'—the global society. A free society is a pluralist society without a common hierarchy of peculiar achievements. The market supposes a discipline, a system of rules of behavior. The market is a game able to create wealth. Indeed, it gives each player adequate information in order to allow him/her to answer to needs of which he/she is not able to be directly aware. Thanks to this game, opportunities increase for each player since nobody is treated unequally. And there is no inequality when the result of the game is different for different people: on the contrary, the market mechanism itself produces impersonal flags, signs which show the players when and how they have to modify their actions to adapt to events of which they are only indirectly aware.

States from time to time wish to introduce social justice through legal regulation. But equity cannot be considered as a means of establishing social justice. Social justice, Hayek writes (Hayek 1973), destroys the spontaneous feature of the market, proscribes the possibility of a great society, and is incompatible with liberty. In other words, equality of treatment before the law—even in the name of charity—leads to arbitrary outcomes. There are requests and revolts against the abstraction of the rules which must be observed within the framework of the Great Society. Indeed, there is always a predilection for concrete human circumstances. But such feedback would only reveal a lack of intellectual and moral maturity faced with the necessities of an impersonal global order of humankind.

Globalization supposes an interdependence within all humankind as we move towards the realization of an open society in which people must accept risk. Consequently, globalization produces inequalities. Peace will result only from the observance of rules of fair behavior. Equity will be realized provided that everybody is treated equally according to these abstract rules of fair behavior. Equality of opportunities cannot be implemented beyond equal treatment before the law.

The State must not involve itself in a process of social justice with the purpose of giving to a group of

citizens what belongs to others. Success or failure must depend only upon each person's action and cannot flow from the knowledge and power held by public authorities. Authoritarian intervention generates disorder and cannot be equitable. To speak of concrete common goals is specific to small closed societies in which people were protected by community at the price of their freedom. In contrast, a great society of free people can exist only if everybody submits to 'equal' abstract rules.

In this context, it is easy to understand that the state must play a minimal role. The 'entitlement theory' (Nozick 1974) based upon a discussion of Hart's 'principle of fairness' (Hart 1961) challenges the claims for State action aimed at implementing equality among people.

Where the State is striking space, the judge becomes a masterpiece in the framework of the legal structure. The judge enjoys extended powers in such a society. He has to correct permanently the disturbances within an order which is spontaneous, not artificial, and which tells people what they have to do. Such a judge must decide in 'equity' in the sense that he has to decide if people's behavior is appropriate to what can normally be expected in comparison with the common conduct of the members of the same group.

The law of the judges is a translation of what people have learned from their actions. Judges' function is to maintain a permanent and efficient order in people's behavior. The judge contributes to improving a given order, an order which must not be upset by changes within relationships between citizens.

Such contemporary theoretical views on equity are particularly significant. Globalization is promoting neoliberal views on law, justice and equity. Indeed, equity has already become a common referent for regional and international Courts of Justice. This leads in turn to many ambiguities, for instance in the implementation of the European Union's *Equal Treatment Directive* against sex discrimination. This last rule is based upon a general feeling of equity … but supposes some affirmative action through Community (state) intervention (see the *European Law Journal* 1998, and Posner 1993, pp. 404–5). The same ambiguity can be found in the report by the Commission on Global Governance entitled *Our Global Neighborhood* (Commission on Global Governance 1995). In this report, written at the request of the United Nations, we read that 'Justice and equity are essential human values. Respect for them is indispensable for peace and progress … Where large numbers of citizens are treated unfairly or denied their due … discontent is inevitable and conflict likely.' However, 'a concern for equity is not tantamount to an insistence on equality.' Therefore, equity 'does call for deliberate efforts to reduce gross inequalities, to deal with factors that cause or perpetuate them, and to promote a fairer sharing of resources' (Commission on Global Governance 1995, p. 51). This is good as political principle, but what practical legal rule can flow from such abstract, inconsistent, and equivocal discourse?

Nevertheless, and despite of the lack of general agreement on its relationship with equality, equity appears as a newly contemporary and significant notion which promises great possibilities for redesigning law-making, the administration of justice, and the process of judicial decision-making.

See also: Civil Law; Common Law; Equality and Inequality: Legal Aspects; Equality: Philosophical Aspects; Globalization: Legal Aspects; Hayek, Friedrich A von (1899–1996); Justice and Law; Justice: Philosophical Aspects; Lex Mercatoria; Natural Law

Bibliography

Aequitas et Bona fides 1955 *Festgabe Simonius*. Helbing & Lichtenhahn, Basle, Switzerland

Arnaud A-J (ed.) 1993 *Dictionnaire encyclopédique de Théorie et de Sociologie du droit*. 2nd edn. Libr. générale de droit et de jurisprudence (LGDJ), Paris

Arnaud A-J 1998 De l'équité des marchands à l'équité du marché. In: *Entre modernité et mondialisation: Cinq leçons d'histoire de la philosophie du droit et de l'État*. LGDJ, Paris

Bonse E 1963 *Summum jus summa injuria. Individualgerechtigkeit und der Schutz allgemeiner Werte in Rechtsleben*. Tübingen, Germany

Bürgi W 1939 *Ursprung und Bedeutung der Begriffe 'Treu und Glauben' und 'Billigkeit' im schweizerischen Zivitrecht*. Stämpfli, Berne, Switzerland

Castán Tobeñas J 1954 *La Formulación judicial del derecho–Jurisprudencia y arbitrio de equidad*. 2nd edn. Instituto Editorial Reus, Madrid

Commission on Global Governance 1995 *The Global Neighbourhood*. The Report of the Commission on Global Governance, Oxford University Press, Oxford, UK

De Page H 1931 *A propos du gouvernement des juges, L'équité en face du droit*. Sirey-Bruylant, Paris

Esser J 1963 Wandlungen von Billigkeit und Billigkeitsrechtsprechung im modernen Privatrecht. In: *Summum ius summum injuria. Mitgliedern der Tübingen Juristenfakultät*. Mohr, Tübingen, Germany, pp. 22–40

European Law Journal 1998 Special issue 4/2 on Sex Equality.

Grossi P 1998 '*Aequitas canonica*' *Quaderni Fiorentini per la storia del pensiero giuridico moderno*. Giuffrè edn., Milan, Vol. 27

Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford

Hayek F A 1973 *Law, Legislation and Liberty*. Routledge & Kegan Paul, London

Lord Lloyd of Hampstead, Freeman M D A 1985 *Lloyd's Introduction to Jurisprudence*. 5th edn. Stevens & Sons, London

Maine H J S 1861 *Ancient Law: Its connection with the Early History of Society and its Relation to Modern Ideas*. J. Murray, London

Manai D 1985 *Le juge entre la loi et l'équité: Essai sur le pouvoir d'appréciation du juge en droit suisse*. Payot, Lausanne, France

Marini C M 1959 *Il Giudizio di equità nel processo civile*. Premesse teoriche, Padova, Italy

Mommsen Th (ed.) 1954 *Corpus Juris Civilis, Digestum*. Weidemann, Berlin, Vol. 1
Newman R A (ed.) 1973 *Equity in the World's Legal Systems: A Comparative Study*. Establissements Emile Bruylant, Brussels, Belgium
Nozick R 1974 *Anarchy, State and Utopia*. Basic Books, New York
Popper K R 1963 *Conjectures and Refutations*. Routledge & Kegan Paul, London
Posner R A 1993 *The Problems of Jurisprudence*. Harvard University Press, Cambridge, MA (1st edn. 1990)
Rabello A M (ed.) 1997 *Aequitas and Equity: Equity in Civil Law and Mixed Jurisdictions*. Harry and Michael Sacher Institute for Legislative Research and Comparative Law, The Hebrew University of Jerusalem, Israel
Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA
Ridder H 1950 Aequitas und equity. *Archiv für Rechts- und Sozialphilosophie* **39**: 181–200
Zaki M S 1990 Équité. *Archives de Philosophie du Droit* 87–118

A.-J. Arnaud

Ergonomics, Cognitive Psychology of

The term 'ergonomics' and the ideas behind it were introduced by the Polish scientist Jastrzebowski as early as 1857. However, only at the beginning of the twentieth century did growing concern about the health of the workforce develop which gave rise to the foundation of dedicated research institutes in a number of countries. Although at that time ergonomic problems were addressed, the rise of ergonomics as an organized field had to wait until World War II and thereafter. It was only in the 1950s that the Ergonomics Research Society in the UK and the Human Factors Society in the USA were founded.

Several subdisciplines of psychology address issues related to human work. To some degree ergonomics can be seen as a blend of engineering and of experimental psychology. It is not only the application of knowledge from experimental psychology to engineering, the fertilization is bidirectional. Among the influences which triggered and shaped the rise of cognitive psychology were developments in engineering such as information theory, control theory, and computer science. A core characteristic of these developments was the abstract nature of major concepts; they were functional and not linked to physical implementations of the functions. Hence they were also suited for the analysis of human performance. Among those who established the close relation between ergonomics and cognitive psychology which developed as human work became more and more 'mind-work' rather than 'body-work' were Paul Fitts in the USA and Donald Broadbent in the UK. Both of them contributed to the key issues of what is sometimes called 'cognitive ergonomics' or 'psychological ergonomics.' Three of these key issues will be briefly addressed.

1. Human–Machine Systems

Humans interact with machines to achieve certain goals at work. In some way human–machine systems are basic units of ergonomic analysis. They can differ vastly in terms of complexity, depending on the kind of machine involved. The term 'machine' in this context is to be understood broadly, ranging from simple tools such as hammers and screwdrivers to vehicles and computers and finally to largely automated technical systems such as can be found in power plants and in the chemical industry. Issues that arise from the study of complex human–machine systems are addressed in a number of other places in this encyclopedia (see *Expert Systems in Cognitive Science*; *Human–Computer Interaction*; *User Interface Design*).

1.1 Manual Control

A basic characteristic of a machine is its transformation of human movements into machine output, e.g., of movements of a computer mouse into movements of the cursor on the screen, or of rotations of steering wheels into changes in the direction of vehicles. In spite of the tremendous differences in terms of complexity, there are commonalities among many human–machine systems. Formally skilled operation of a machine requires the solution of a control problem, that is, a solution of the problem to find an appropriate input to achieve a desired output.

There are different types of solutions to a control problem. Two basically different ways are a closed-loop solution and an open-loop solution, which can be combined. In both cases the transformation implemented by the machine is (approximately) inverted and applied to the desired output. With closed-loop control the approximate inversion is an emergent characteristic of the closed-loop structure, whereas open-loop control requires an internal model of the transformation by the machine.

Manual-control tasks are typically studied by means of tracking setups which have the basic structure of a closed-loop system. The development of an internal model mostly implies the learning of an intermodal relation, specifically of the relation between visually perceived output of the machine and proprioceptively sensed self-produced movements. When such a relation follows a power law (which includes proportionality), it is easy to learn, and internal models of the transformation can be fairly precise. This becomes

apparent when the machine has to be operated in an open-loop fashion, with the feedback loop being interrupted. However, when the transformation involves other nonlinear relations, an inversion (e.g., move the mouse to the right for a cursor movement to the left), or dynamic components (such as an integration in the case of driving a vehicle), internal models tend to become less precise and the dependence on closed-loop control more pronounced.

Perhaps the most extensively studied variation of visuo-motor relations is called stimulus–response compatibility. For example, when a cursor moves to the right upon a mouse movement to the right, this would be a spatially compatible relation, and a directional reversal would produce an incompatible relation. More generally, the term 'compatibility' refers to the naturalness of a visuo-motor transformation, which is often fairly obvious to a user. Usually there are strong differences between compatible and incompatible relations when visual stimuli and responses share a common dimension (such as left–right). Whether or not a certain relation is compatible depends not only on the physical situation, but also on the mental representation of it. For example, when one has to respond to small and large numbers between 1 and 12 by pressing a right or left key, the assignment of small numbers to the left key and large numbers to the right key results in shorter reaction times (compatible relation) than the opposite assignment (incompatible relation); however, the normally incompatible relation becomes compatible when one imagines the numbers being arranged on the face of a clock.

1.2 Monitoring and Supervisory Control

Modern technology allows machines to become more and more complex, and manual-control tasks can mostly be automated. Modern airplanes are popular examples of this. There is a certain irony in such automation which should not be overlooked. On the one hand automation serves to avoid human errors and to reduce mental load, but on the other hand there are emergency situations in which the system changes from automatic to manual control. The irony is that a human operator who is deprived of the daily experience of controlling the machine in normal situations is required to master it in the difficult emergency cases. In the normal situation, the task of the operator in highly advanced transport and production systems becomes that of a supervisor. The task of the supervisor is no longer continuous control of the system output, but rather the monitoring of the system state, the discovery and diagnosis of abnormal states, and the taking of appropriate actions when abnormal states are discovered.

Monitoring can be a boring task: irregular system states are rare events, so there is little that happens. The classical problem which gave rise to the systematic study of maintained attention to displays which only rarely indicate critical events is that of the radar controller. In the course of a watch there is typically a fairly rapid decline of the probability of detecting a critical event, the so-called vigilance decrement. The same phenomenon can become a problem in industrial inspection tasks. In particular, the really rare events carry a risk of being overlooked. One of the factors which are likely to contribute to the vigilance decrement is temporary inactivation of the appropriate task set.

Monitoring can also be a task which places too high a burden on a human operator. There can be hundreds of state variables which have to be monitored, so the decision as to which state variable to inspect next can be difficult. In their inspection patterns operators adjust to the characteristics of the state variables. For example, state variables which change rapidly tend to be inspected more frequently than state variables which change only slowly. In addition, interdependences between state variables guide the inspection pattern. However, in the case of a failure the normal characteristics of the state variables may no longer be present, so that inspection patterns can be misguided.

The detection of abnormal system states can take minutes or hours, depending on a number of conditions. Among them is the rate of change of system states, which can be slow. Again, as does the pattern of inspection behavior, fault detection relies heavily on the operator's internal model of the system. Unfortunately the concept of an 'internal model' is somewhat elusive when it comes to a precise characterization. One of the reasons is that the internal model on which monitoring and supervisory control is based is not necessarily identical with the reportable knowledge of an operator. Both incidental observations and formal studies suggest that at least part of the knowledge on which the operators' monitoring and control actions are based is implicit in nature.

2. Workload and Multiple-task Performance

One of the classical and continuing problems is the assessment of workload. Its importance results from the notion that overload or underload can result in errors, accidents, and/or health risks. There are three types of measurements that are employed for workload assessments, physiological measurements (such as heart rate and heart-rate variability), subjective ratings, and behavioral data.

The assessment of workload by means of behavioral data is closely linked to the study of multiple-task performance. Everyday observations suggest that the mind can be more or less dedicated to the performance of a certain task. To bring this observation into the realm of scientific treatment, the concept of capacity

or of resources has been introduced: the more capacity is dedicated to the performance of a certain task, the higher is the performance in this task, and the less capacity is available for the performance of other concurrent tasks. From this general idea, it follows that there should be a tradeoff between performance levels in two concurrent tasks when performers vary their respective emphasis and, theoretically, the proportions of total capacity allocated to each of them. Performance tradeoffs indeed exist. For practical purposes, these simple concepts suggest that performance in a subsidiary task could be used to assess mental workload. However, so far all attempts to find a subsidiary task that could be used as a kind of yardstick in the assessment of mental load have failed.

The limits of multiple-task performance are important not only in the context of workload assessment, but also in the context of work design. For example, in supervisory control situations can arise which require different actions concurrently. Sometimes concurrent actions can be physically impossible, so that the problem becomes one of scheduling tasks according to priorities. Even when concurrent activities are physically possible, there is not only interference because of limited mental capacity, but also interference of a more specific kind, related, for example, to cross-talk within the perceptual or motor systems. A basic principle is that interference becomes more pronounced as the perceptual information for different tasks, the effectors involved, and the nature of the information (e.g., verbal vs. pictorial) become more similar.

3. *Skill Acquisition*

Human performance can be improved by practice, and the time needed for a certain action (such as the classical example of rolling a cigar) declines. Practice curves can be described by a power function for a great variety of skills; this fact is referred to as the 'power law of practice.' However, skilled performance has a number of additional characteristics. It tends to be less variable, and it tends to be associated with a reduced workload and less interference with the performance of concurrent tasks. The latter aspect of skill acquisition is called automatization. However, the reduction of dual-task interference is not a universal characteristic of skill acquisition, but there can also be certain secondary tasks for which dual-task interference increases.

In many cases the skills required to operate complex human–machine systems are acquired by means of special training devices, so that there are degrees of freedom in designing the training procedure. The design of training procedures can be a fairly complex process. Often it involves the identification of task elements which are practiced separately, special instructions, and special feedback. The training procedure shapes the internal model of the machine. Thus, a critical factor for the success of training procedures is the match of the practice situation and the final 'test' with the real system, in particular the match with respect to functional characteristics and with respect to the information available. Somewhat ironically, too much can be done to support the learner during training in terms of guidance and feedback; performance can become dependent on such additional guidance and feedback so that it will deteriorate as soon as it is no longer available.

See also: Artificial Intelligence in Cognitive Science; Automation: Organizational Studies; Dual Task Performance; Job Analysis and Work Roles, Psychology of; Job Design, Psychology of; Organizational Behavior, Psychology of; Simulation and Training in Work Settings; User Interface Design; Workplace Environmental Psychology

Bibliography

Damos D L (ed.) 1991 *Multiple-task Performance*. Taylor and Francis, London

Matthews G, Davies D R, Westerman S, Stammers R B 2000 *Human Performance. Cognition, Stress, and Individual Differences*. Psychology Press, Hove, UK

Proctor R W, Dutta A 1995 *Skill Acquisition and Human Performance*. Sage, Thousand Oaks, CA

Proctor R W, van Zandt T 1994 *Human Factors in Simple and Complex Systems*. Allyn and Bacon, Boston

Sanders A F 1998 *Elements of Human Performance: Reaction Processes and Attention in Human Skill*. Erlbaum, Mahwah, NJ

Wickens C D 1992 *Engineering Psychology and Human Performance*, 2nd edn. Harper Collins, New York

H. Heuer

Erikson, Erik Homburger (1902–94)

Erik Erikson was born on June 21,1902 in Frankfurt, Germany to a Danish Jewish mother, Karla Salomonsen. He was eventually given his stepfather's name (Homburger) and later christened himself, Erikson. He grew up in Germany, completed his gymnasium in Karlsruhe and in 1921 enrolled in art school. However, he soon left this and after a *Wanderschaft* period, he secured his first professional position in the late 1920s as a teacher, together with Peter Blos, in Vienna's Heitzing School founded by Eva Rosenfeld, Dorothy Burlingham, and Anna Freud. The latter became

Erikson's analyst and supervisor in play therapy with children. Erikson married a Canadian dancer, Joan Serson, in 1930. In 1933, now the father of two sons, he left his incompleted studies at the University of Vienna, was voted a full member of the Vienna Psychoanalytic Society and emigrated to the United States with his young family. He established a private practice in Boston and was made a member of the Boston Psychoanalytic Society despite that body's parochial policy of limiting membership to physicians. During the next 5 years, Erikson worked at the Judge Baker Center in Boston, the Murray Clinic at Harvard, secured a position at Yale with the Gesell clinic and Institute of Human Relations, and began an important professional relationship with the anthropologist, Margaret Mead.

In 1938, Erikson, whose surname until this time had been his stepfather's (Homburger), changed his name to Erikson, became a US citizen and moved to California to take a position at Macfarlane's Institute of Child Welfare in Berkeley and established a substantial private practice in San Francisco in 1942–3. While working with returning veterans at the Mount Zion Rehabilitation Clinic, he began a friendship with the Jungian analyst Joseph Wheelwright, who broadened Erikson's Freudian perspective. In 1949 he became a professor at Berkeley, in spite of lacking the usual academic credentials, but soon left after refusing to sign the McCarthy era loyalty oath—a 'Here I stand' gesture that may have prefigured his forthcoming book on Martin Luther. In 1950 he presented his and Joan's outline of the human life cycle to a White House Conference and saw his most influential book, *Childhood and Society*, published. He returned to the eastern US in 1951 to join the Austen Riggs Center in Stockbridge, Massachusetts as clinician and author. There he established important professional and personal relationships with ego psychoanalyst Robert Knight and David Rapaport and he completed his second major book, *Young Man Luther*, (1958). He published the monograph *Identity and the Life Cycle* in *Psychological Issues* (1959) expanding both on the life cycle outline and on ego identity, a concept, together with 'identity crisis' that was to become associated with his name in the mind of the public.

With help from David Riesman and Robert White, Erikson moved from Stockbridge to Boston and Harvard in 1960 and became a professor, but outside any particular academic department. In the midst of, and to some extent with reference to, the student unrest of the 1960s, Erikson began work on his second psychobiography, *Gandhi's Truth* (1969) and for which he received both the Pulitzer Prize and the National Book Award. In addition, *Insight and Responsibility* (1964), *Identity: Youth and Crisis* (1968), and a revision of *Childhood and Society* (1963/1985) were also completed during this period. These books reflected a growing shift in Erikson's thinking from the development of the individual, to the social context of individuals' development, to the moral and ethical implications of different forms of social organization for humankind. In this transition, the work of his friend, Paul Tillich, was influential. Clearly, Erikson had begun to move quite a distance from his orthodox psychoanalytic initiation by Anna Freud.

By the beginning of the 1970s Erikson had become one of the most respected and honored thinkers within the psychoanalytic developmental area, described by one reviewer as 'the most influential of living psychoanalysts.' He delivered two sets of honorary lectures: the Godkin at the John F. Kennedy School for Government at Harvard in 1972, and the Jefferson for the National Endowment for the Humanities in Washington, DC, in 1973. These eventuated in two more books, *Toys and Reasons* (1977) and *Dimensions of a New Identity* (1974). *Life History and The Historical Moment* (1975) occasioned significant criticism from both feminists and New Left sources who found Erikson still too politically and psychologically conservative. Erikson's most important final writings were on the last life cycle stage of *Integrity: Reflections on Dr. Borg's Life Cycle*, a commentary on Bergman's film, *Wild Strawberries*, *The Life Cycle Completed: A Review* (1982), and *Vital Involvement in Old Age* (1986) written with Joan Erikson and Helen Kivnick. Erikson continued to do clinical work until 1990. He died on May 12, 1994, in Harwich, Massachusetts.

Erikson lived through the world-shaping events of the twentieth century: World War I, the Great Depression, World War II, and the United States' coming-of-age in Vietnam. The national and ideological strands that the Eriksons gathered up and wove into his psychosocial developmental theory were numerous: Danish, German, North American, Jewish, Protestant, classical psychoanalytic, ego psychoanalytic, and even Jungian. His literary/theoretical excursions into a form of biography that he labeled 'psychohistory' further widened his intellectual scope to include Luther in Reformation Germany, Gandhi in pre-independence India, Hitler in pre-W.W.II Germany, Gorky in Bolshevik Russia, and Jesus Christ in biblical Palestine. Although only one of a number of psychoanalytic thinkers who left Europe for America in the 1930s, only Erikson integrated and transformed these disruptive experiences into a clinically, theoretically, and empirically useful outline of the stages of normal human development.

Erikson took a 'configurational' approach to children's play therapy, wherein he considered the spatial arrangement of items chosen by the child to be as important as the items themselves or the stories linking them. He carried this configurational approach over into his psychosocial developmental chart that portrays the human life cycle as consisting of eight central developmental crises, each having its more or less specific age-related time of ascendance. The first three stages: Trust vs. Mistrust (infancy), Autonomy

vs. Shame, Doubt (early childhood), and Initiative vs. Guilt (play age) parallel the Freudian psychosexual stages: oral, anal, and phallic. However, in contrast to Freudian libido theory, these stages of ego growth refer not just to erogenous body zones, but to modes of behavior and senses of oneself and of the world that are shaped by the child-rearing practices of particular parents embedded within their particular culture. The description of life cycle stages after childhood constitutes Erikson's unique contribution to psychoanalytic developmental theory. In these he outlines a sequence of ego development extending beyond childhood through the adult years: Industry vs. Inferiority (school age), Identity vs. Identity Confusion (adolescence), Intimacy vs. Isolation (young adulthood), Generativity vs. Stagnation (middle adulthood), Integrity vs. Despair (old age). Associated with each of these stages are social institutional 'rituals' (e.g., infancy: 'motherhood') as well as eventual 'virtues' (e.g., infancy: 'Hope').

The emergence of each of the eight stages in this sequence is governed by an epigenetic principle comprising a built-in developmental progression of individual physical and psychological change occurring with a social context of age-related 'average expectable' demands and provisions. Because Erikson located personality development firmly within the social context, he is to be credited with initiating psychosocial developmental theory. Optimal development in childhood consists of a balance between the opposite poles described above, tilted in favor of the more positive alternative. Beginning in adolescence, resolution of a developmental crisis incorporates both poles into a synthesis. For example, Identity development includes an element of Diffusion or Confusion; Generativity formation incorporates some Stagnation; and there is no Integrity without some accompanying Despair. The primarily positive resolution of each stage is considered to be crucial for the resolution of the next stage. However, consistent with Erikson's configurational approach, there is an aspect to the design of his developmental chart that allows for remediation of inadequately resolved issues or precocious resolution of 'future' developmental issues. Each stage is assumed to arise during every chronological period, although the nature of that stage is shaped by the dominant issue of the period. For example, a Basic Trust vs. Mistrust crisis re-occurring at Intimacy will take a form configured by the adult nature of relationships typical of early adulthood. Because they arise again at each succeeding chronological period, previously unresolved age-specific issues can be remediated through the provision of 'better-than-average expectable' conditions (e.g., counseling, psychotherapy, or enriched milieux).

In the broadest view, Erikson proposes that there are crucial periods ('crises') of development characterizing the human life cycle, and societies have developed institutions that, more or less, furnish a matrix within which these developmental crises are resolved. Ideally, societal rewards and demands mesh with individual needs and abilities at each of these stages. Whereas Freud furnished a view of personality from the 'inside out,' looking to the individual's unconscious in order to understand cultural products, Erikson, while accepting the essential Freudian dynamic principles, added a perspective from the 'outside in,' beginning at the level of the whole individual and describing a particular culture's impact on human growth. Also, in contrast to Freud, Erikson saw ego growth as continuous throughout the life cycle—from infancy to old age. Erikson has been criticized from a number of quarters. Orthodox psychoanalytic theorists (e.g., Jones, Eissler) saw Erikson's emphasis on social factors and the ego in the development of personality as paying insufficient attention to the causative role of early childhood experiences and the primacy of unconscious factors in determining later adult personality. In short, they found him insufficiently 'Freudian.' Early feminist critics (e.g., Weisstein, Millett) took Erikson to task for his assumption of a unique developmental trajectory for women based upon their biological differences from men—in other words, too 'Freudian.' Later feminists (e.g., Gilligan) criticized him for paying insufficient attention to women's differences from men in terms of their capacities for care and relationship. Object relations theorists (e.g., Fairbairn, Winnicott) and self theorists (Kohut) felt that Erikson was somewhat superficial in underemphasizing the importance of early mother–child interactions and discussing too much the importance of social factors in personality development. By contrast, some historians and sociologists (e.g., Kushner, Weinstein, and Platt) have found Erikson's theory too little concerned with underlying causative social factors and questioned the universality of his eight-stage model. Social critics from the left (e.g., Fromm, Kovel) thought Erikson too conservative in his apparent acceptance of the social status quo. And, finally, social scientists (e.g., J. McV. Hunt) found Erikson's writing too imprecise to allow for empirical research. Erikson, himself, was critical of attempts to validate his theory with experimental methods, preferring more clinical and social approaches.

Erikson's primary contribution, emerging from the smoke of all the controversy noted above, has been a comprehensive overview of the human life cycle that incorporates somatic, psychological, and social elements. The influence of his psychosocial approach has been pervasive throughout ego psychoanalytic theory, developmental psychological research, and educational counseling, although there is no 'Eriksonian' school of psychotherapy. Almost every psychology textbook on personality theory and developmental psychology of children, adolescents, and adults has a section devoted to Erikson. Although his descriptions of early developmental stages usually are covered, it is the identity issue at adolescence with which he is most

closely associated, followed by his descriptions of adult life cycle stages. Even political theory, especially the 'politics of identity,' has been influenced by his approach. Erikson has amended Freudian theory to make it more understandable, 'experience-near,' growth-focused, researchable, and socially useful. Although criticized by classical Freudians, Erikson's extension of psychoanalytic theory into the realm of normal, and optimal, human development, may have saved that theory from its excessive interiority and obscurantism. In addition, because of its essentially interpersonal nature, Erikson's theory has provided a valuable link between drive-based classical and ego psychoanalytic theory and relationship-based object relations theory. From a social–political perspective, the psychosocial developmental outline proposed by Erikson is sufficiently clear and putatively universal that it can provide a guide for social programs that focus on establishing optimal conditions for human development across the life cycle. This becomes increasingly practicable in view of the research (see Kroger 2000) that is accumulating supporting the validity of measures of the Eriksonian stages of Industry, Identity, Intimacy, Generativity, and Integrity.

See also: Adulthood: Developmental Tasks and Critical Life Events; Allport, Gordon W (1897–1967); Developmental Psychology; Human Development, Successful: Psychological Conceptions; Identity in Childhood and Adolescence; James, William (1842–1910); Jung, Carl Gustav (1875–1961); Life Course in History; Life Course: Sociological Aspects; Life Management, Developmental Psychology of; Lifespan Development, Theory of; Love and Intimacy, Psychology of; Midlife Psychological Development; Parenthood and Adult Psychological Developments; Personality Development in Adulthood; Personality Psychology; Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis, History of; Psychoanalysis in Sociology; Psychoanalysis: Overview; Psychohistory; Psychology: Historical and Cultural Perspectives

Bibliography

Coles R 1970 *Erik H. Erikson: The Growth of his Work*, 1st edn. Atlantic Monthly Press, Boston

Erikson E H 1958 *Young Man Luther: A Study in Psychoanalysis and History*, 1st edn. Norton, New York

Erikson E H 1959 Identity and the life cycle: selected papers. *Psychological Issues* **1** International Universities Press

Erikson E H 1963/1985 *Childhood and Society*, 2nd edn. Norton, New York

Erikson E H 1964 *Insight and Responsibility: Lectures on the Ethical Implications of Psychoanalytic Insight*, 1st edn. Norton, New York

Erikson E H 1968 *Identity, Youth and Crisis*, 1st edn. Norton, New York

Erikson E H 1969 *Gandhi's Truth: On the Origins of Militant Nonviolence*, 1st edn. Norton, New York

Erikson E H 1974 *Dimensions of a New Identity: The 1973 Jefferson Lectures in the Humanities*, 1st edn. Norton, New York

Erikson E H 1975 *Life History and The Historical Moment*, 1st edn. Norton, New York

Erikson E H 1976 Reflections on Dr Borg's life cycle. *Daedalus* **105**: 1–28

Erikson E H 1977 *Toys And Reasons: Stages in the Ritualization of Experience*, 1st edn. Norton, New York

Erikson E H 1982 *The Life Cycle Completed: A Review*. Norton, New York

Erikson E H, Erikson J M, Kivnick H Q 1986 *Vital Involvement in Old Age*. Norton, New York

Friedman L J 1999 *Identity's Architect: A Biography of Erik H. Erikson*. Scribner, New York

Kroger J 2000 *Identity Development: Adolescence through Adulthood*. Sage, Thousand Oaks, CA

Schlein S (ed.) 1987 *A Way of Looking at Things: Selected Papers from 1930–1980*, 1st edn. Norton, New York

J. E. Marcia

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Errors in Statistical Assumptions, Effects of

Each statistical procedure rests upon a collection of assumptions for its validity. For example, many statistical procedures assume that a data set consists of a collection of independent observations. A large number of these same procedures also assume that the observations all have the same distribution. Each of these assumptions, as well as any others that a procedure assumes, may or may not be reasonable in a given application. If one or more of the assumptions justifying a procedure fail to hold, the statistical properties of the procedure may be different than they would be if the assumptions hold. This article discusses ways to assess the sensitivity of a statistical procedure to violations of its assumptions.

1. Introduction

The development of almost every statistical procedure is accompanied by a list of properties possessed by the procedure under some specified statistical assumptions. For example, the t-test is a uniformly most powerful unbiased level α test if one assumes that the data $X_1,\ldots, X_n$ comprise a sample of independent random variables with common normal distribution. Of course, one can perform a t-test with data $X_1,\ldots, X_n$ regardless of whether one makes some or all of the

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

assumptions. One cannot expect the associated properties to hold, however, if one does not believe the assumptions. If one has reason to believe (or fear) that one or more of the assumptions used to justify a statistical procedure fail to hold, then one should be concerned with how sensitive the properties are to the assumptions. For example, suppose that the observations $X_1,\ldots, X_n$ have some distribution other than normal. How much does that affect the probability of type I error when I use the level α t-test? How much does it affect the power of the test at important alternatives? This entry will not attempt to answer all of the questions of this sort. Rather, this entry describes general ways to assess the sensitivity of statistical procedures to specific violations of their assumptions. An excellent discussion of the role of assumptions in statistical models was initiated by Box (1980).

Related to the question of sensitivity is the question of diagnosis. That is, can one tell whether or not the data suggest that certain assumptions are indeed violated? There are several statistical methods that have been developed to diagnose violations of assumptions of the linear regression model. Some of these are discussed in *Linear Hypothesis: Regression (Graphics)*.

For a given statistical procedure, one first needs to be aware of what assumptions justify the procedure. Then one needs to decide which aspects of the procedure and which assumptions the sensitivity of which one wishes to assess. One needs to decide which violations of those assumptions one is willing to contemplate and finally how one wishes to measure the sensitivity. Each of these issues is examined in a detailed study of one statistical procedure in Sect. 2.

Theoretical results concerning sensitivity can be difficult to derive, even for the most elementary types of violations. Hotelling (1961) and Efron (1969) analyze the effects of some specific types of violations on the distribution of the t statistic. Box (1953) discusses the effects of violations of assumptions on tests concerning variances.

2. *The t-test*

Since the t-test is such a popular statistical procedure, let us first consider violations of its assumptions. The assumptions for the validity of the t-test are that the data comprise a sample of independent random variables, all with the same normal distribution. Specifically, there are three key assumptions:

(a) the observations are independent;

(b) the observations all have the same distribution; and

(c) each observation has a normal distribution.

There are many ways in which these assumptions can be violated. For example, the observations might be dependent. (See *Time Series: General* for discussion of one type of model of dependence between observations.) The observations might have different distributions. Alternatively, some or all the observations might have distributions other than normal. Finally, it is possible that more than one assumption is violated. For example, each observation could have its own distribution, none of which is normal, and they could be dependent.

The most crucial properties of a test procedure are its size and power. These can be measured from the distribution of the test statistic given a set of assumptions about the data. For example, consider a test of the null hypothesis that the mean μ of a sample of random variables is 0. If one were interested in measuring the sensitivity to the assumption that the data are normal, and if one contemplates the possibility that a linear function $a(X-b)$ of each observation X has a Cauchy distribution, then one must reinterpret μ to be something other than the mean (e.g., the median, b). The power of the test can then be written as a function of $a\mu$.

2.1 *Theoretical Results*

Clearly, it is not reasonable to expect to find theoretical results that handle every possible departure from the assumptions. For some violations, it is possible to determine analytically how sensitive a procedure will be. For example, suppose that the observations $X_1,\ldots, X_n$ have normal distribution with mean μ and common variance σ^2, but that they have common correlation ρ. And suppose that one wishes to perform a t-test of the null hypothesis that the mean of the distribution is μ_0. The t statistic has the form

$$T = \frac{\bar{X}-\mu_0}{\sqrt{S^2/n}} \tag{1}$$

where $\bar{X}$ is the sample average and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{2}$$

Under the usual assumptions for the t-test, if $\mu = \mu_0$, then T has a t distribution with $n-1$ degrees of freedom. In the case of common correlation, it is possible to show that the random variable $U = T/\sqrt{1+n\rho/(1-\rho)}$ has t distribution with $n-1$ degrees of freedom. If ρ is greater than 0, the distribution of T will be more spread out than the nominal t distribution if $\mu = \mu_0$. This will have the effect of making the size of the t-test larger than its nominal value. For example, if $\rho = 0.05$ and $n = 10$ and one uses a test with nominal size 0.05; the actual size is 0.10. If the null hypothesis is false and $\mu \neq \mu_0$, then U has noncentral t distribution with $n-1$ degrees of freedom and noncentrality

parameter $\mu\sqrt{n/[1+(n-1)\rho]}$. The effect of this on the power of the test is not so obvious, but it is not as dramatic as the effect on the size.

For other violations, one can attempt to reason qualitatively what the effect will be. For example, suppose that the observations $X_1,\ldots, X_n$ come from a distribution with heavier tails than a normal distribution (i.e., observations far from the median occur with high probability). And suppose that one wishes to perform a t-test of the null hypothesis that the median of the distribution is μ_0. Some theoretical results concerning this case have been derived by Benjamini (1983). A heavy-tailed distribution will produce a value $\bar{X}$ with more variability than a normal distribution will. That is, we would expect to see large values of $|\bar{X}-\mu_0|$ with higher probability from a heavy-tailed distribution than from a normal distribution. But the effect of heavy tails on $\sqrt{S^2/n}$ should be even greater than the effect on $\bar{X}$ because the squaring gives the large deviations $X_i-\bar{X}$ much more prominence. Hence, one expects $|T|$ to be smaller if the data have a distribution with heavy tails than if the data have normal distribution. Not all violations lend themselves to qualitative analysis of this sort, however. And, as Benjamini (1983) points out, an intuitive analysis like this might only hold in special cases, as will be noted later in this article.

When one wishes to focus on a specific collection of possible violations, then one can perform simulations to assess both qualitative and quantitative sensitivity. Simulation has advantages and disadvantages when compared to theoretical analysis. A simulation can be designed to assess the sensitivity of any statistical procedure to any specific collection of violations of its assumptions. However, each simulation analysis assesses only the specific violations for which it was designed and does not tell us how sensitive the procedure is to other, albeit similar, violations. Of course, simulations are generally quick and easy to perform, so one could perform many of them to cover a large number of specific violations. The remainder of this section illustrates the sorts of simulations that one might perform in order to assess the sensitivity of various aspects of t-test to some of the more easily understood violations of assumptions.

2.2 *Violations of Normality*

Suppose that one thinks that the observations are normal with mean μ and variance σ^2, and that one wishes to test the null hypothesis that $\mu = 0$. Suppose that most interest lies in a test with level 0.05 and with the power of the test when $\mu = \sigma$. However, suppose that one is willing to contemplate the possibility that the data come from a distribution that is not normal. If that distribution does not have a mean or does not have a variance, then one will need to reinterpret μ and/or σ. For distributions without a mean, let μ stand for the median. For all distributions, let σ stand for 0.741 times the interquartile range (which happens to be the standard deviation when the data have a normal distribution). Now, one can simulate samples of size n from every distribution that one wishes to consider and compute the probability that $|T| > c$, where c is the 0.975 quantile of the t distribution with $n-1$ degrees of freedom. For distributions with mean (or median) 0, this probability will be the size of the test. For distributions with mean $\mu = \sigma$, this probability will be the power of interest. Table 1 reports the results of several simulations. Each row of the table represents two million simulated samples of size $n = 10$ or $n = 20$ from a distribution that is not normal. The simulated data from the t distribution have three degrees of freedom. The simulated data from the gamma distribution have parameter 2. The σ in the last column is 0.741 times the interquartile range. The 'Normal' rows give the theoretical values under the normal distribution assumption. For each sample the statistic T was computed. The table reports the estimated probability that $|T| > c$ for c equal to the 0.975 quantile of the t distribution with $n-1$ degrees of freedom.

Table 1
Simulated sizes and powers of the t-test under violations of normality

Distribution	Sample size (n)	Prob. of $\|T\| > c$	
		$\mu = 0$	$\mu = \sigma$
Laplace	10	0.042	0.593
	20	0.046	0.862
Cauchy	10	0.019	0.295
	20	0.020	0.325
t (3 degrees of freedom)	10	0.040	0.619
	20	0.043	0.848
Uniform	10	0.054	0.971
	20	0.052	1.0
Exponential	10	0.099	0.786
	20	0.081	0.997
Gamma (2)	10	0.077	0.838
	20	0.066	0.999
Normal	10	0.05	0.803
	20	0.05	0.989

A quick look at the rows of Table 1 corresponding to heavy-tailed distributions (Laplace, Cauchy, t) indicates what we argued earlier. That is, the probability of $|T| > c$ is smaller for these heavy-tailed distributions than for the normal. However, as is common with simulations, one must be careful not to generalize too far. If one had considered level 0.20 tests instead of level 0.05 tests, the results would have been slightly different. For example, with the Laplace distribution, the estimated level is slightly higher than 0.2 with samples of both sizes 10 and 20. This is consistent with the theoretical results of Benjamini (1983).

Table 2
Simulated sizes and powers of the t-test under violations of common distribution

Ratio of standard deviations	Sample size	Proportion with larger standard deviation	Prob. of $\|T\| > c$	
			$\mu = 0$	$\mu = \sigma$
2	10	0.1	0.047	0.809
	10	0.5	0.046	0.804
	20	0.1	0.048	0.984
	20	0.5	0.048	0.985
5	10	0.1	0.030	0.831
	10	0.5	0.037	0.800
	20	0.1	0.033	0.959
	20	0.5	0.046	0.982
10	10	0.1	0.017	0.829
	10	0.5	0.033	0.799
	20	0.1	0.015	0.948
	20	0.5	0.045	0.981
1	10	0	0.05	0.803
	20	0	0.05	0.989

2.3 Violations of Common Distribution

For the assumption that the observations have the same distribution, there are many possible violations. The data might split into two (or a few) groups, each of which has a common distribution. Alternatively, each observation could have its own distribution. Or there might be just one (or a few) observation(s) with distribution different from the others. All of these are possible even if one believes that each of the individual distributions has a normal distribution.

Table 2 gives the results of several simulations. As before, each simulation is based on two million simulated samples of size $n = 10$ or $n = 20$. Each sample consists of n normal random variables with a common mean μ but with either 10 percent or 50 percent having a larger standard deviation than the rest. The larger standard deviation is either 2, 5 or 10 times as large as the smaller one. The value of σ in the last column is

$$\sigma = \sqrt{(1-p)\sigma_1^2 + p\sigma_2^2} \tag{3}$$

where σ_1 is the smaller standard deviation, σ_2 is the larger standard deviation, and p is the proportion of the sample with the larger standard deviation. The 'Ratio = 1' rows give the theoretical values under the assumption that the data have a common normal distribution. The table reports the estimated probability that $|T| > c$ for c equal to the 0.975 quantile of the t distribution with $n-1$ degrees of freedom. Note how the size of the test can be significantly lowered by the inclusions of 10 percent of the observations with a standard deviation 5 or 10 times as large as the rest. The power of the test at $\mu = \sigma$ is not much affected by this type of violation. However, the value of σ/σ_1 increases as the ratio of the two standard deviations increases. If one had calculated the power at a fixed alternative, there would have been a dramatic decrease in power as the ratio of standard deviations increased and as the proportion with the larger standard deviation increased.

2.4 Violations of Independence

Violations of the independence assumption are a bit more difficult to think about. There are simple violations such as equal correlation that is not 0. Indeed, this case is the one that lent itself to an analytic solution earlier in this entry. Another type of violation of independence is time-series dependence. This means that the amount of dependence between observations depends on the order in which they were observed. Many data sets are collected without recording when the observations were made. This would make it impossible to perform a traditional time series analysis. However, we can still assess the effect of time series dependence on the distribution of a statistic like T, even if one did not record the observation times. For example, suppose that observations arrive in the order $X_1, \ldots, X_n$, regardless of whether or not one knows this to be true. One might consider the possibility that $X_i = \rho X_{i-1} + \varepsilon_i$ for $i = 2, \ldots, n$. Here $\rho \in [-1,1]$ and $\varepsilon_2, \ldots, \varepsilon_n$ are assumed to be mutually independent with variance $(1-\rho^2)$ Var (X_1). This type of dependence is called autoregressive of order 1. (For more detail on

Table 3
Simulated sizes and powers of the t-test under violations of independence

Sample size (n)	Correlation (ρ)	Prob. of $\lvert T\rvert > c$	
		$\mu = 0$	$\mu = \sigma$
10	0	0.05	0.803
	0.1	0.072	0.792
	0.3	0.138	0.773
	0.5	0.245	0.759
	0.7	0.412	0.761
	0.9	0.672	0.824
20	0	0.05	0.989
	0.1	0.074	0.982
	0.3	0.145	0.962
	0.5	0.256	0.928
	0.7	0.422	0.883
	0.9	0.688	0.864

autoregressive dependence, see *Time Series: ARIMA Methods*.)

Table 3 presents the results of several simulations. Each simulated sample has autoregressive dependence of order 1 with a different correlation. Each number in the table is based on 2,000,000 simulated samples. The rows with '$\rho = 0$' give the theoretical values under the assumption that the data are independent. The table reports the estimated probability that $|T| > c$ for c equal to the 0.975 quantile of the t distribution with $n-1$ degrees of freedom. Even correlation as small as 0.1 has a noticeable effect on the size of the test.

3. *Linear Regression*

The assumptions which justify the standard inferences in linear regression are a bit more complicated than those which justify the t-test. In turn, the types of violations of these assumptions are also more complicated. To be specific, consider simple linear regression. The data comprise a set of pairs $(Y_1, x_1),\ldots,(Y_n, x_n)$, and the numbers $x_1,\ldots,x_n$ are considered fixed, known constants. Alternatively, the remainder of the probabilistic statements are to be understood conditional on these values. The assumptions are that

(a) the random variables $Y_1,\ldots,Y_n$ are independent,
(b) every Y_i has a normal distribution,
(c) Var (Y_i) is the same for all i, and
(d) there exist numbers β_0 and β_1 such that $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$.

This section will briefly discuss violations of these assumptions, but not in as much detail as it did for the one-sample t-test. Failures of independence can come in the same forms that they did for the one-sample t-test. However, there are additional varieties available as well. For example, one might consider the possibility that there is correlation between Y_i and Y_j that depends on how far apart x_i and x_j are. Each Y_i can have a distribution that depends on x_i through more than just its mean and/or variance. Indeed, there are standard methods in regression analysis for diagnosing and dealing with cases in which it appears that Var (Y_i) depends on x_i. The dependence of $\mathrm{E}(Y_i)$ and x_i can be very general. It might seem reasonable in some problems to assume that $E(Y_i)$ is a monotone function of x_i, but not always. For more discussion of regression methodology, including diagnosing violations of assumptions, see *Linear Hypothesis: Regression (Basics)* and *Linear Hypothesis: Regression (Graphics)*.

4. *Sensitivity of Bayesian Procedures*

Bayesian procedures (see *Bayesian Statistics*), being statistical procedures, are also based upon probabilistic assumptions, and are thereby sensitive to violations of their assumptions. Typically, a Bayesian procedure is based on the same types of assumptions as any other procedure plus additional assumptions about the distributions of any parameters (such as μ and σ^2 in the t-test) that might appear in the statistical model. These latter assumptions are generally part of what is called the prior distribution of the parameters. Sensitivity of Bayesian procedures to violations of all of these assumptions can be assessed in a manner similar to what was done earlier in this article. However, those who use Bayesian procedures to represent subjective beliefs conditional on the observed data (see *Probability: Interpretations*) would be more interested in assessing sensitivity in a different manner.

For example, suppose that one wants an interval estimate of a parameter μ, and one models data $X_1,\ldots,X_n$ as independent normal random variables with mean μ and variance σ^2. One would also specify a prior distribution for (μ, σ^2). After observing the data values $X_1 = x_1,\ldots,X_n = x_n$, one would then compute the appropriate interval estimate $[a,b]$. Then one could ask what interval would have been computed if one had modeled the data as having a different distribution. For example, suppose $(X_i - \mu)/(0.6745\sigma)$ are modeled as independent with Cauchy distribution. (The factor of 0.6745 makes the interquartile range of X_i the same as it would have been if the data were normal, namely 1.349σ.) Under this Cauchy assumption, a different interval $[a',b']$ would be computed using the same observed data values $x_1,\ldots,x_n$. The change from the interval $[a,b]$ to the interval $[a',b']$ measures the sensitivity of the process of forming an interval estimate to this change of the assumption about the distribution of the data. This type of sensitivity of the process of forming an inference differs somewhat from the type of sensitivity of a particular procedure that was described earlier. Sensitivity of process is not a uniquely Bayesian concern,

however, and it can also be measured for changes to assumptions in non-Bayesian inference.

There have been some successful attempts to take considerations of sensitivity to assumptions into account explicitly in the formulation of Bayesian models. Model uncertainty (see Draper 1995) and Bayesian model averaging (Hoeting et al. 1999) attempt to develop methodology for expressing directly in the modeling process the belief that more than one out of several competing statistical models might be appropriate for the data, and it may not be clear which one is best. For example, one might not be able to decide whether to model data as normal random variables or as Cauchy random variables. The methods of model uncertainty and model averaging provide ways to pursue both models (as well as others that might seem plausible) without having to pretend as if we really believed firmly in precisely one of them.

Finally, there have been some methods developed to deal with specific common violations of assumptions. The assumption that all observations in a sample have the same distribution has received attention particularly in order to be able to accommodate the occasional outlying observation. One can model data as coming from a mixture of two or more distributions. One of the distributions is the main component on which most interest is focussed. The others represent the types of data that might arise when something goes wrong with the data collection procedure. Each observation can be modeled as coming from one of the component distributions with a probability associated with each component. Indeed, using powerful simulation methods (see *Markov Chain Monte Carlo Methods*) one can even compute, for each observation, the conditional probability that it arose from each of the components given the observed data.

5. *Conclusion*

This article has discussed the need for carefully acknowledging the probabilistic assumptions made to justify a statistical procedure and some methods for assessing the sensitivity of the procedure to those assumptions. If one believes that there are violations of the assumptions that might reasonably arise in practice, and if the procedure is overly sensitive to violations of those assumptions, one might wish to select an alternative procedure.

There are two popular approaches to choosing alternatives to standard statistical procedures when one fears that assumptions are likely to be violated. One is to use robust procedures, and the other is to use nonparametric procedures. Robust procedures are designed to be less sensitive to violations of specific assumptions without sacrificing too much of the good performance of standard procedures when the standard assumptions hold. Nonparametric procedures are chosen so that their properties can be verified under fewer assumptions. For more details on these types of procedures see *Robustness in Statistics* and *Nonparametric Statistics: The Field*.

See also: Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Robustness in Statistics; Statistics: The Field; Time Series: ARIMA Methods; Time Series: General

Bibliography

Benjamini Y 1983 Is the *t* test really conservative when the parent distribution is long-tailed? *Journal of the American Statistical Association* **78**: 645–54

Box G E P 1953 Non-normality and tests on variances. *Biometrika* **40**: 318–35

Box G E P 1980 Sampling and Bayes' inference in scientific modeling and robustness (with discussion). *Journal of the Royal Statistical Society*, (*Series A*) **143**: 383–430

Draper D 1995 Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society*, (*Series B*) **57**: 45–97

Efron B 1969 Student's *t*-test under symmetry conditions. *Journal of the American Statistical Association* **64**: 1278–302

Hoeting J A, Madigan D, Raftery A E, Volinsky C 1999 Bayesian model averaging: A tutorial (with discussion). *Statistical Science* **14**: 382–417

Hotelling H 1961 The behavior of some standard statistical tests under nonstandard conditions. In: Neyman J (ed.) *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA, Vol.1, pp. 319–59

M. J. Schervish

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Estimation: Point and Interval

Introduction

When sampling is from a population described by a density or mass function $f(x \mid \theta)$, knowledge of θ yields knowledge of the entire population. Hence, it is natural to seek a method of finding a good estimator of the point θ, that is, a good point estimator. However, a point estimator alone is not enough for a complete inference, as a measure of uncertainty is also needed. For that, we use a set estimator in which the inference is the statement that $\theta \in C$ where $C \subset \Theta$ and $C = C(\mathbf{x})$ is a set determined by the value of the data $\mathbf{X} = \mathbf{x}$ observed. If θ is real-valued, then we usually prefer the set estimate C to be an interval. Our uncertainty is quantified by the size of the interval and its probability of covering the parameter.

1. *Point Estimators*

In many cases, there will be an obvious or natural candidate for a point estimator of a particular parameter. For example, the sample mean is a natural candidate for a point estimator of the population

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

mean. However, when we leave a simple case like this, intuition may desert us so it is useful to have some techniques that will at least give us some reasonable candidates for consideration. Those that have stood the test of time include:

1.1 The Method of Moments

The method of moments (MOM) is, perhaps, the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. One of the strengths of MOM estimators is that they are usually simple to use and almost always yield some sort of estimate. In many cases, unfortunately, this method yields estimators that may be improved upon.

Let $X_1,\ldots, X_n$ be a sample from a population with density or mass function $f(x|\theta_1,\ldots, \theta_k)$. MOM estimators are found by equating the first k sample moments to the corresponding k population moments. That is, we define the sample moments by $m_j = \Sigma_{i=1}^n X_i^j/n$ and the population moments by $\mu_j(\theta_1,\ldots, \theta_k) = \mathrm{E}X^j$ for $j = 1,\ldots, k$. We then set $m_j = \mu_j(\theta_1,\ldots, \theta_k)$ and solve for $\theta_1,\ldots, \theta_k$. This solution is the MOM estimator of $\theta_1,\ldots, \theta_k$.

1.2 Maximum Likelihood Estimators

For a sample $X_1,\ldots, X_n$ from $f(x\,|\,\theta_1,\ldots, \theta_k)$, the likelihood function is defined by

$$\begin{aligned} L(\theta\,|\,\mathbf{x}) &= L(\theta_1,\ldots, \theta_k\,|\,x_1,\ldots, x_n) \\ &= \Pi_{i=1}^n f(x_i\,|\,\theta_1,\ldots, \theta_k) \end{aligned} \tag{1}$$

The values of θ_i that maximize this function are those parameter values for which the observed sample is most likely, and are called the maximum likelihood estimators (MLE). If the likelihood function is differentiable (in θ_i), the MLEs can often be found by solving

$$\frac{\partial}{\partial\theta_i}\log L(\theta\,|\,\mathbf{x}) = 0, \quad i = 1,\ldots, k \tag{2}$$

where the vector with coordinates $\frac{\partial}{\partial\theta_i}\log L(\theta\,|\,\mathbf{x})$ is called the score function (see Schervish 1995, Sect. 2.3).

Example If $X_1,\ldots, X_n$ are i.i.d. Bernoulli (p), the likelihood function is

$$L(p\,|\,\mathbf{x}) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \tag{3}$$

and differentiating log $L(p\,|\,\mathbf{x})$ and setting the result equal to zero gives the MLE $\hat{p} = \Sigma_i x_i/n$. This is also the method of moments estimator.

If we instead have samples $X_1,\ldots, X_n$ from a binomial (k, p) population where p is known and k is unknown, the likelihood function is

$$L(k\,|\,\mathbf{x}, p) = \prod_{i=1}^n \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i} \tag{4}$$

and the MLE must be found by the numerical maximization. The method of moments will give the closed form solution.

$$\hat{k} = \frac{\bar{x}^2}{\bar{x}-(1/n)\Sigma_i(x_i-\bar{x})^2} \tag{5}$$

which can take on negative values. This illustrates a shortcoming of the method of moments, one not shared by the MLE. Another, perhaps more serious shortcoming of the MOM estimator is that it may not be based on a sufficient statistic (see *Statistical Sufficiency*), which means it could be inefficient in not using all of the available information in a sample. In contrast, both MLEs and Bayes estimators (see *Bayesian Statistics*) are based on sufficient statistics.

1.3 Bayes Estimators

In the Bayesian paradigm a random sample $X_1,\ldots, X_n$ is drawn from a population indexed by θ, where θ is considered to be a quantity whose variation can be described by a probability distribution (called the *prior distribution*). A sample is then taken from a population indexed by θ and the prior distribution is updated with this sample information. The updated prior is called the *posterior distribution*.

If we denote the prior distribution, by $\pi(\theta)$, and the sampling distribution by $f(\mathbf{x}\,|\,\theta)$, then the posterior distribution, the conditional distribution of θ given the sample, $\mathbf{x}$, is

$$\pi(\theta\,|\,\mathbf{x}) = f(\mathbf{x}\,|\,\theta)\pi(\theta)/m(\mathbf{x}) \tag{6}$$

where $m(\mathbf{x}) = \int f(\mathbf{x}\,|\,\theta)\pi(\theta)\,\mathrm{d}\theta$ is the marginal distribution of $\mathbf{x}$.

Example Let $X_1,\ldots, X_n$ be i.i.d. Bernoulli (p). Then $Y = \Sigma_i X_i$ is binomial (n, p). If p has a Beta (α, β) prior distribution, that is,

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{7}$$

the posterior distribution of p given y, is

$$\begin{aligned} f(p\,|\,y) &= \frac{f(y, p)}{f(y)} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} \end{aligned} \tag{8}$$

which is a beta distribution with parameters $y+\alpha$ and $n-y+\beta$. The posterior mean, a Bayes estimator of p, is

$$\hat{p}_{\mathrm{B}} = \frac{y+\alpha}{\alpha+\beta+n} \tag{9}$$

2. Evaluating Point Estimators

There are many methods of deriving point estimators (robust methods, least squares, estimating equations, invariance) but the three in Sect. 1 are among the most popular. No matter what method is used to derive a point estimator, it is important to evaluate the estimator using some performance criterion.

One way of evaluating the performance of a point estimator W of a parameter θ is through its mean squared error (MSE), defined by $\mathrm{E}_\theta(W-\theta)^2$.

MSE measures the average squared difference between the estimator W and the parameter θ. Although any increasing function of the absolute distance $|W-\theta|$ would serve, there is a nice factorization

$$\begin{aligned}\mathrm{E}_\theta(W-\theta)^2 &= \mathrm{Var}_\theta W + (\mathrm{E}_\theta W-\theta)^2 \\ &= \mathrm{Var}_\theta W + (\mathrm{Bias}_\theta W)^2\end{aligned} \tag{10}$$

where we define the bias of a point estimator as $\mathrm{Bias}_\theta W = \mathrm{E}_\theta W-\theta$. An estimator whose bias is identically (in θ) equal to zero is called *unbiased*.

For an unbiased estimator we have $\mathrm{E}_\theta(W-\theta)^2 = \mathrm{Var}_\theta W$, and so, if an estimator is unbiased, its MSE is equal to its variance. If $X_1,\ldots, X_n$ are i.i.d. from a population with mean μ and variance σ^2, the sample mean is an unbiased estimator since $\mathrm{E}\bar{X} = \mu$, and has MSE

$$\mathrm{E}(\bar{X}-\mu)^2 = \mathrm{Var}\bar{X} = \frac{\sigma^2}{n} \tag{11}$$

Controlling bias does not guarantee that MSE is minimized. In particular, it is sometimes the case that a trade-off occurs between variance and bias. For example, in sampling from a normal population with variance σ^2, the usual unbiased estimator of the variance $S^2 = \frac{1}{n-1}\Sigma_{i=1}^n(X_i-\bar{X})^2$ has MSE $= 2\sigma^4/(n-1)$. An alternative estimator for σ^2 is the maximum likelihood estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})^2 = \frac{n-1}{n}S^2$. This is a biased estimator of σ^2 with MSE

$$\begin{aligned}\mathrm{E}(\hat{\sigma}^2-\sigma^2)^2 &= \left(\frac{2n-1}{n^2}\right)\sigma^4 < \left(\frac{2}{n-1}\right)\sigma^4 \\ &= \mathrm{E}(S^2-\sigma^2)^2\end{aligned} \tag{12}$$

showing that $\hat{\sigma}^2$ has smaller MSE than S^2. Thus, by trading off variance for bias, the MSE is improved.

Measuring performance by the squared difference between the estimator and a parameter is a special case of a function called a *loss function*. The study of the performance, and the optimality, of estimators evaluated through loss functions is a branch of decision theory. In addition to MSE, based on squared error loss, another popular loss function is absolute error loss, $L(\theta, W) = |W-\theta|$. Both of these loss functions increase as the distance between θ and W increases, with minimum value $L(\theta, \theta) = 0$. That is, the loss is minimum if the estimator is correct.

In a decision theoretic analysis, the worth of an estimator is quantified in its risk function, that is, for an estimator W of θ, the risk function is $R(\theta, W) = \mathrm{E}_\theta L(\theta, W)$, so the risk function is the average loss. If the loss is squared error, the risk function is the MSE.

Using squared error loss, the risk function (MSE) of the binomial Bayes estimator of p is

$$\begin{aligned}\mathrm{E}_p(\hat{p}_{\mathrm{B}}-p)^2 &= \mathrm{Var}_p\hat{p}_{\mathrm{B}} + (\mathrm{Bias}_p\hat{p}_B)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n}-p\right)^2\end{aligned} \tag{13}$$

In the absence of good prior information about p, we might try to choose α and β to make the risk function of $\hat{p}_{\mathrm{B}}$ constant (called an *equalizer* rule). The solution is to choose $\alpha = \beta = \sqrt{\mathrm{n}/4}$, yielding

$$\mathrm{E}(\hat{p}_{\mathrm{B}}-p)^2 = \frac{n}{4(n+\sqrt{\mathrm{n}})^2}$$

We can also use a Bayesian approach to the problem of loss function optimality, where we would use the prior distribution to compute an average risk $\int_\Theta R(\theta, W)\pi(\theta)\mathrm{d}\theta$, known as the *Bayes risk*. We then find the estimator that yields the smallest value of the Bayes risk. Such an estimator is called the Bayes rule with respect to a prior π.

To find the Bayes decision rule for a given prior π we write the Bayes risk as

$$\begin{aligned}&\int_\Theta R(\theta, W)\,\pi(\theta)\,\mathrm{d}\theta \\ &\quad = \int_\chi\left[\int_\Theta L(\theta, W(\mathbf{x}))\,\pi(\theta\,|\,\mathbf{x})\,\mathrm{d}\theta\right]m(\mathbf{x})\,\mathrm{d}\mathbf{x}\end{aligned} \tag{14}$$

where the quantity in square brackets is the expected value of the loss function with respect to the posterior distribution, called the *posterior expected loss*. It is a function only of $\mathbf{x}$, and not a function of θ. Thus, for each $\mathbf{x}$, if we choose the estimate $W(\mathbf{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk.

For squared error loss, the posterior expected loss is minimized by the mean of the posterior distribution. For absolute error loss, the posterior expected loss is minimized by the median of the posterior distribution. If we have a sample $X_1,\ldots, X_n$ from a normal distribution with mean θ and variance σ^2, and the prior is $n(\mu, \tau^2)$, the posterior mean is

$$E(\theta \mid \bar{x}) = \frac{\tau^2}{\tau^2+(\sigma^2/n)}\bar{x}+\frac{\sigma^2/n}{\tau^2+(\sigma^2/n)}\mu \qquad (15)$$

Since the posterior distribution is normal, it is symmetric and the posterior mean is the Bayes rule for both squared error and absolute error loss. The posterior mean in our binomial/beta example, $\hat{p}_B = \frac{y+\alpha}{\alpha+\beta+n}$, is the Bayes estimator against squared error loss.

We can loosely group evaluation criteria into large sample or asymptotic methods, and small sample methods. Our calculations using MSE and risk functions illustrate small sample methods. In large samples, MLEs typically perform very well, being asymptotically normal and efficient, that is, attaining the smallest possible variance. Other types of estimators that are derived in a similar manner (for example, M-estimators—see *Robustness in Statistics*) also share good asymptotic properties. For a detailed discussion see Casella and Berger (2001), Lehmann (1998), Stuart et al. (1999) or Lehmann and Casella (1998, Chap. 6).

3 Interval Estimation

Reporting a point estimator of a parameter θ only provides part of the story. The story becomes more complete if an assessment of the error of estimation is also reported. Informally, this can be accomplished by giving an estimated standard error of the estimator and, more formally, this becomes the reporting of an interval estimate. If $\mathbf{X} = \mathbf{x}$ is observed, an *interval estimate* of a parameter θ is a pair of functions, $L(\mathbf{x})$ and $U(\mathbf{x})$ for which the inference $\theta \in [L(\mathbf{x}), U(\mathbf{x})]$ is made. The *coverage probability* of the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is the probability that $[L(\mathbf{X}), U(\mathbf{X})]$ covers the true parameter, θ, and is denoted by $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$.

By definition, the coverage probability depends on the unknown θ, so it cannot be reported. What is typically reported is the confidence coefficient, the infimum of the coverage probabilities, $\inf_\theta P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$.

If $X_1,\ldots, X_n$ are i.i.d. with mean μ and variance σ^2, a common interval estimator for μ is

$$\mu \in \bar{x} \pm 2\frac{s}{\sqrt{n}} \qquad (16)$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation. The validity of this interval can be justified from the Central Limit Theorem, because

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \to n(0, 1) \qquad (17)$$

the standard normal distribution. We then see that the coverage probability (and confidence coefficient) of Eqn. 16 is approximately 95 percent.

The above interval is a large sample interval since its justification is based on an asymptotic argument. There are many methods for constructing interval estimators that are valid in small samples, including these.

3.1 Inverting a Test Statistic

There is a correspondence between acceptance regions of tests (see *Hypothesis Testing in Statistics*) and confidence sets, summarized in the following theorem.

Theorem *For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of H_0: $\theta = \theta_0$. For each $\mathbf{x} \in \chi$,* define a set $C(\mathbf{x})$ *in the parameter space by*

$$C(\mathbf{x}) = \{\theta_0 \colon \mathbf{x} \in A(\theta_0)\} \qquad (18)$$

Then the random set $C(\mathbf{X})$ is a $1-\alpha$ confidence set. Conversely, let $C(\mathbf{X})$ be a $1-\alpha$ confidence set. For any $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{x \colon \theta_0 \in C(\mathbf{x})\} \qquad (19)$$

Then $A(\theta_0)$ is the acceptance region of a level α test of H_0: $\theta = \theta_0$.

Example If $X_1,\ldots, X_n$ are i.i.d. $n(\mu, \sigma^2)$, with σ^2 known, the test of H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$ will accept the null hypothesis at level α if

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu_0 \leqslant \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (20)$$

The interval of μ values, $[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{\alpha/2}\sigma\sqrt{n}]$, for which the null hypothesis will be accepted at level α, is a $1-\alpha$ confidence interval for μ.

3.2 Pivotal Inference

Perhaps one of the most elegant methods of constructing set estimators is the use of pivotal quantities (Barnard 1949). A random variable $Q(\mathbf{X}, \theta) = Q(X_1,\ldots, X_n, \theta)$, is a *pivotal quantity* (or pivot) if the distribution of $Q(\mathbf{X}, \theta)$ is independent of all parameters. If we find a set C such that $P(Q(\mathbf{X}, \theta) \in C) =$

$1-\alpha$, then the set $\{\theta: Q(\mathbf{X}, \theta) \in C\}$ has coverage probability $1-\alpha$.

In location and scale cases, once we calculate the sample mean $\bar{X}$ and the sample standard deviation S, we can construct the following pivots:

Form of pdf	Type of pdf	Pivotal quantity
$f(x-\mu)$	location	$\bar{X}-\mu$
$\frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$	scale	$\frac{\bar{X}}{\sigma}$
$\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$	location-scale	$\frac{\bar{X}-\mu}{S}$

(21)

In general, differences are pivotal for location problems, while ratios (or products) are pivotal for scale problems. See also *Fiducial and Structural Statistical Inference*.

Example Suppose that $X_1, \ldots, X_n$ are i.i.d. exponential (λ). Then $T = \sum_i X_i$ is a sufficient statistic for λ and $T \sim$ gamma (n, λ). In the gamma pdf, t and λ appear together as t/λ and, in fact the gamma (n, λ) pdf $(\Gamma(n)\lambda^n)^{-1} t^{n-1} e^{-t/\lambda}$ is a scale family. Thus, if $Q(T, \lambda) = 2T/\lambda$, then

$$Q(T, \lambda) \sim \text{gamma}(n, \lambda(2/\lambda)) = \text{gamma}\,(n, 2)$$

which does not depend on λ. The quantity $Q(T, \lambda) = 2T/\lambda$ is a pivot with a gamma $(n, 2)$, or χ^2_{2n}, distribution, and a $1-\alpha$ pivotal interval is

$$\frac{2T}{\chi^2_{2n,\,\alpha/2}} \leqslant \lambda \leqslant \frac{2T}{\chi^2_{2n,\,1-\alpha/2}}$$

where $P(\chi^2_{2n} > \chi^2_{2n,\,a}) = a$.

3.3 Bayesian Intervals

If $\pi(\theta \mid \mathbf{x})$ is the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$, then for any set $A \subset \Theta$ the posterior probability of A is

$$P(\theta \in A \mid \mathbf{x}) = \int_A \pi(\theta \mid \mathbf{x})\, \mathrm{d}\theta$$

If $A = A(\mathbf{x})$ is chosen so that this posterior probability is $1-\alpha$, then A is called a $1-\alpha$ *credible set* for θ. If $\pi(\theta \mid \mathbf{x})$ corresponds to a discrete distribution, we replace integrals with sums in the above expressions.

The interpretation of the Bayes interval estimator is different from the classical intervals. In the classical approach, to assert 95 percent coverage is to assert that in 95 percent of repeated experiments, the realized intervals will cover the true parameter. In the Bayesian approach, a 95 percent coverage means that the probability is 95 percent that the parameter is in the realized interval. In the classical approach the randomness comes from the repetition of experiments, while in the Bayesian approach the randomness comes from uncertainty about the value of the parameter (summarized in the prior distribution).

Example Let $X_1, \ldots, X_n$ be i.i.d. Poisson (λ) and assume that λ has a gamma prior, $\lambda \sim$ gamma (a, b), where a is an integer. The posterior distribution of λ is

$$\pi(\lambda \mid \textstyle\sum_i X_i = \sum_i x_i) = \text{gamma}(a + \sum_i x_i, [n + (1/b)]^{-1}) \tag{22}$$

Thus the posterior distribution of $2[n+(1/b)]\lambda$ is $\chi^2_{2(a+\Sigma x)}$, and a $1-\alpha$ Bayes credible interval for λ is

$$\left\{\lambda: \frac{\chi^2_{2(a+\Sigma x),\,1-\alpha/2}}{2[n+(1/b)]} \leqslant \lambda \leqslant \frac{\chi^2_{2(a+\Sigma x),\,\alpha/2}}{2[n+(1/b)]}\right\} \tag{23}$$

We can also form a Bayes credible set by taking the highest posterior density (HPD) region of the parameter space, by choosing c so that

$$1-\alpha = \int_{\{\lambda: \pi(\lambda \mid \Sigma x) \geqslant c\}} \pi(\lambda \mid \Sigma_i x_i)\, \mathrm{d}\lambda \tag{24}$$

Such a construction is optimal in the sense of giving the shortest interval for a given $1-\alpha$ (although if the posterior is multimodal the set may not be an interval).

4. Other Intervals

We have presented two-sided parametric intervals that are constructed to cover a parameter. Other types of intervals include (*a*) one-sided intervals, (*b*) distribution-free intervals, (*c*) prediction intervals, (*d*) tolerance intervals.

One-sided intervals are those in which only one endpoint is estimated, such as $\theta \in [L(\mathbf{X}), \infty)$. Distribution-free intervals are intervals whose probability guarantee holds with little (or no) assumption on the underlying distribution. The other two interval definitions, together with the usual confidence interval, provide us with a hierarchy of inferences, each more stringent than the previous.

If $X_1, \ldots, X_n$ are i.i.d. from a population with cdf $F(x \mid \theta)$, and $C(\mathbf{x}) = [L(\mathbf{x}), U(\mathbf{x})]$ is an interval, for a specified value $1-\alpha$ it is a:

(a) *confidence interval* if, for all θ, $P_\theta[L(\mathbf{X}) \leqslant \theta \leqslant U(\mathbf{X})] \geqslant 1-\alpha$;

(b) *prediction interval* if, for all θ, $P_\theta[L(\mathbf{X}) \leqslant X_{n+1} \leqslant U(\mathbf{X})] \geqslant 1-\alpha$;

(c) *tolerance interval* if, for all θ and for a specified value p, $P_\theta([F(U(\mathbf{X})\,|\,\theta) - F(L(\mathbf{X})\,|\,\theta) \geqslant p] \geqslant 1-\alpha$.

So a confidence interval covers a mean, a prediction interval covers a new random variable, and a tolerance interval covers a proportion of the population. Thus, each gives a different inference, with the appropriate one being dictated by the problem at hand. Vardeman (1992) argues that these 'other intervals' provide an inference that is different from that of a confidence interval and are as important as confidence intervals.

5. *Conclusions*

Point estimation is one of the cornerstones of statistical analysis, and the basic element on which many inferences are based. Inferences using point estimators gain statistical validity when they are accompanied by an interval estimate, providing an assessment of the uncertainty. We have mainly discussed parametric point and interval estimation, where we assume that the underlying model is correct. Such an assumption can be questioned, and considerations of nonparametric or robust alternatives can address this (see *Robustness in Statistics*). For more on these subjects see, for example, Hettmansberger and McKean (1998) or Staudte and Sheather (1990). Full treatments of parametric point and interval estimation can be found in Casella and Berger (2001), Stuart et al. (1999), or Schervish (1995).

Bibliography

Barnard G A 1949 Statistical inference (with discussion). *Journal of the Royal Statistical Society, Series B* **11**: 115–39

Casella G, Berger R L 2001 *Statistical Inference*, 2nd edn. Wadsworth/Brooks Cole, Pacific Grove, CA

Hettmansperger T P, McKean J W 1998 *Robust Nonparametric Statistical Methods*. Wiley, New York

Lehmann E L 1998 *Introduction to Large-Sample Theory*. Springer-Verlag, New York

Lehmann E L, Casella G 1998 *Theory of Point Estimation*, 2nd edn. Springer-Verlag, New York

Schervish M J 1995 *Theory of Statistics*. Springer-Verlag, New York

Staudte R G, Sheather S J 1990 *Robust Estimation and Testing*. John Wiley, New York

Stuart A, Ord J K, Arnold S 1999 *Advanced Theory of Statistics, Classical Inference and the Linear Model,* 6th edn. Arnold, Oxford University Press, London, Vol. 2A

Vardeman S B 1992 What about other intervals. *The American Statistician* **46**: 193–7

G. Casella and R. L. Berger

Ethical Behavior, Evolution of

The distinction between good or bad and what we ought or ought not do constitutes the subject matter of ethics. Early students of the evolution of ethical behavior (EEB) and some sociobiologists attempted to direct the study of EEB into the domain of prescriptive ethics. Twenty-first century sociobiologists are not concerned with the nature of vice, virtue, or the rules of moral behavior, but with the question of the biological origin of ethical behavior. To disregard the distinction between the questions of ethics and the questions of the evolutionary causes and the bases of human ethical behavior is to misunderstand the discipline of EEB, which is not a branch of philosophy or of ethics, but a subdiscipline of sociobiology. No prescriptive code can be derived from the theory of evolution; therefore, the previously used term 'evolutionary ethics' is a misnomer that must be dropped from sociobiological usage. An excellent philosophical examination (Flew 1967) and a superior historical analysis (Richards 1987) on EEB are available.

The ideas on EEB will be exemplified by a review of arguments, first, of the originators of EEB; second, of ecologists, ethologists, geneticists, linguists, neurophysiologists, and cognitive scientists; and third, of sociobiologists and their supporters and opponents.

1. *The Originators of the Idea*

Charles Darwin, in *On the Origin of Species* (1859), established the homology between human and nonhuman primate morphology, but not between human and nonhuman behavior. In his revised edition of *The Descent of Man* (1885), he recognized that the moral sense or conscience differentiates humans from the other animals, and that 'The imperious word *ought* seems merely to imply the consciousness of the existence of a rule of conduct, however it may have originated.' He believed that the understanding of the EEB would be enlarged by studies of nonhuman behavior, and that nonhuman animals possessing parental and filial instinct would exhibit a kind of rudimentary intellectual and moral sense. He supported this by arguing that, first, social instincts lead animals to social groups and to perform 'social services'; second, as the mental faculties develop, images and feelings endure and are subject to recall; third, development of language facilitates communication; and last, acquisition of habits helps to guide the conduct of individuals within the community.

Humans, to Darwin, are social animals who have lost some early instincts but retain the protohuman instinctive love and sympathy for others. Some instincts are subject to group selection; the altruistic social behaviors of early humans were not for the good

of species or individuals, but for the welfare of the group. Humans ought to obey the instincts acquired and prescribed by the rule of conduct within society. These Darwinian views have been accepted, except that a small number of biologists finds the theory of group selection (that is, a natural selection operating on more than one member of a lineage group, in which the selection is beneficial for the group rather than for the individual) inadequate to explain altruism (behavior that increases the fitness of the other at cost to self).

There are three Darwinian contributions to EEB. First, the acceptance of morphology as a product of evolution repudiated creationism and thus, the divine source of morality, and at the same time implied that human behavior is also evolutionarily derived. Second, the rejection by Darwin of what John Dewey called 'the holy arc of permanence' meant that all human behavior, including ethics, evolved, is evolving, and will evolve. Third, EEB is concerned with the origin of ethical behavior, and not with the prescriptions of ethics.

Spencer asserted that ethical behavior began in nonhuman social animals, was developed by—and conforms to—the law of evolution, and reached its full expansion in cultural evolution. He argued that cooperation is beneficial to social units, is a cause of physiological and sociological divisions of labor, and it profits all members in a society, but only when social bargains are respected. Cooperation arose at the dawn of life when altruism was dependent on egoism, and egoism was dependent on altruism. Offspring are preserved and species maintained by altruistic acts. To Spencer, evolution was thus a progressive moral force operating within human society and leading to the ethical improvement of humankind.

Spencer replaced the Darwinian natural selection with the 'survival of the fittest,' and developed his concepts in economic terms that led to Social Darwinism. It is now realized that the nurturing of offspring in human and nonhuman families can only be explained in terms of the Spencerian concept of cooperation, which thus appears to be evolutionary. In social groups there must be selective pressure to encourage altruism and to discourage selfishness. By definition social life is altruistic. Spencer's ideas on cooperation were more fully developed by Kropotkin (1902).

T. H. Huxley (1894, reprinted in Nitecki and Nitecki 1993) supposed that protohuman morphology, human society, intellect, and morality are parts of nature, but that human behavior—unlike behavior in other animals—is not innate; therefore, there is a chasm between humans and other animals. Human character, by which he meant the moral and intellectual essence, is passed from generation to generation. He accepted the pessimistic view of human life as a continuous and uneven struggle against the cosmos, and assumed that competition is the exclusive element of the living world. Perhaps Huxley overstressed competition and selfishness as a natural instinct, and underrated cooperation and altruism as evolutionary factors. His panoramic view of life replaced the old body/soul dichotomy with the nature/culture dichotomy.

Although Huxley objected to Social Darwinism as expressed in the Spencerian 'survival of the fittest,' he nevertheless accepted the seventeenth-century perspective of natural selection as a 'struggle for existence.' It was this eternal struggle between humans and cosmos that colored his view of nature, and blinded him to the role of cooperation in evolution.

2. *The Unintended Developers of the Idea*

The conclusions of early psychological behaviorists were that behavior, with the exception of physiological actions, is purely a cultural phenomenon. Learning was also cultural. However, ecologists (Allee et al. 1949) and ethologists (Tinbergen 1951) established that behavior and learning are processes of the central nervous system, and are products of genetics *and* of social interactions. Thus behavior may be both innate and acquired. Animal social interactions are occasionally in a one-to-one relationship (cause and effect), but transmission from the stimulus to the response is mostly through a complex of neurological channels. Behavior is so complex that it does not allow itself to be controlled by a single gene.

The environment *and* the chemical releasers of animals control and influence behavior by acting upon endocrine glands and upon sense organs. If the cooperative human and nonhuman groups are units of selection, then survival is of *greater* importance to the group as a whole than to the individual.

Ecologists documented that mutualism occurs not only between organelles within cells and among individual organisms, but also in groups and even in kingdoms of organisms: herbivores cooperate with symbiotic algae, bacteria fix nitrogen within leguminous plants, and fungi with algae form lichens. Ecologists postulated that instincts are inborn nervous mechanisms that may be activated, and that instinct contributes to the fitness of both individuals and groups. Animals are social only when acting within the group of the same species.

In the reverse of anthropomorphism, ethologists applied the concepts derived from the studies of nonhuman behavior to humans, and established that both human and nonhuman physiology and behavior are constantly modified, changed, and influenced by the presence of other individuals. Without the presence of another individual, the terms 'good' or 'bad' behavior have no meaning.

Ethologists stressed the role of genes in human behavior, the role of information encoded in the genome, and also the role of development in forming the behavioral patterns. They argued that human nature is neither inherently good or bad, but is both, and therefore 'natural' cannot be equated with 'good' or 'bad.'

The geneticist Muller (1959) reasoned that altruistic acts foster the multiplication of altruistic genes, and that cooperation (or mutual aid) is a genetic factor of great value for the worldwide social organization. He agreed with Darwin that the protohuman group was the primary unit of selection in which altruism developed, but for altruism to become fixed in humans the group must remain isolated. Not only our morphology and our DNA, but also our behavior evolved.

Linguistic studies (Pinker 1997) proved that language capacity is already present in newly born humans. Ecologists had already pointed out that human behavior, including language, requires at least two individuals. Linguists added that this presence is necessary for language to develop. Therefore, the capacity for language is genetic, but it must be socially primed to be activated. Since human infants can do simple additions before they can talk, linguists reasoned that even arithmetic has an evolutionary basis. However, to many researchers the structural features of human language differentiate it from nonhuman communications. Others believed that the difference is of degree only.

Neurophysiological and cognitive studies recognized that all mental activities, including consciousness, are products of the physiological functions of the material brain. Mental activities are programmed by instinct and are controlled in recognized locations in the brain. Consciousness is innate in humans. These investigations shifted the argument from dualism to materialism. It is self-evident that the very long human infancy requires cooperation of parents and society. Infants recognize and learn altruism and love of mother, and how to reciprocate.

It follows, from the ecological to cognitive researches, that nonhuman and much human behavior have evolutionary bases. Biologists now agree that, at least in some nonhuman animals, complex behavior has more than simple genetic components, and that humans are *both* biological and cultural beings. These studies established the positive correlation between brain structure and mental activities, questioned the brain/mind dichotomy, and laid the foundations for the sociobiological analysis of EEB.

3. *Sociobiology*

Sociobiology is the systematic study of the biological basis of *all* social behavior in *all* organisms, including humans (Wilson 1975). Ecologists and neurophysiologists accepted the nonhuman and some human behavior and ability to learn as innate; sociobiologists concluded that altruism is also innate. Sociobiological emphasis on the analysis of altruism reinstated the almost dormant EEB. Originally sociobiologists alleged that, in human choices of ethical behavior, heredity plays a greater role than environment. They hypothesized that since men and women are sexually different, their behavior is also different, and the economic and political ranking of humans is natural.

Since sociobiologists contend that a straight line from nonhuman to human behavior exists, they presumed that the fusion of sociobiology with neurophysiology would explain the rules and properties of ethical behavior. Some sociobiologists proposed that ethics should be biologized and that the range of study of nonhuman behavior should be expanded into the domain of psychology. They assumed that ethical and moral concepts are guided by natural selection and by social forces, and that as long as some philosophers fail to accept the evolutionary bases of social behavior they will continue to reject EEB. Furthermore, the emerging sociobiology threatened the three popular existentialist dogmas: the uniqueness of the individual, the search for the meaning and purpose of human life, and belief in the freedom of choice of action. These assertions raised instant responses.

3.1 *Voices of Dissent*

Anthropologists (Sahlins 1976) immediately objected to the application of sociobiology to human behavior and morality. They claimed that culture cannot be reduced to biology because cultural environment controls the phenotype and behavior, and that human ethical behavior is fundamentally different from nonhuman. In this view sociobiology was mired in a Eurocentric patriarchal worldview.

Leftist and feminist critics (Caplan 1978) saw negative political consequences in genetic determinism and surmised that all social human behavior is exclusively environmental and without biological bases. They accused Wilson of Social Darwinism and sexism, that is, of the puritanical and moralistic defense of the old cultural interpretation of ethics, and of human behavior. They linked sociobiology to Nazism and racism. Those critics claimed not only that the science of sociobiology was wrong, but also that it was morally wrong, and their attack on sociobiology became, and continues to be, an attack on EEB.

The sociobiologists' injection of anthropomorphic terms for descriptions of nonhuman behavior was criticized for confusing homology with analogy and for introducing difficulties in analyzing behavior. Homology, when applied to morphology and behavior, requires common ancestry. Altruistic behavior is a function of the social group, but tendencies of organisms to form social aggregates do not appear homologous. Terms may imply homology, but, if altruism in nonhuman and human animals is not homologous, then the term altruism will incorrectly suggest that ethical principles have biological roots. However, terms usually have more than one meaning that is

decipherable only from the context of the sentence. If anthropomorphic terms are restricted to human actions and new terms are coined for nonhuman behavior, then the concept of EEB can be more clearly defined.

3.2 Soft Answers

The Christian tradition of the Fall, caused by original sin, held that humans are born evil, and Huxley claimed that cosmos is manifestly amoral. The struggle for existence, an intrinsically 'bad' property of all creation, did not need explication, but cooperation and altruism did, and all moral actions were judged in relation to the benefit of the group, however defined. Men sacrificing their lives in war were worshiped as heroes, and those dying for faith were beatified as martyrs. Cooperation and altruism have always been socially rewarded, while selfishness was condemned. It is this emphasis on virtue that the study of sociobiology introduced.

Thus altruism, the contrariant of the cold-hearted selfish power of the cosmos, demanded explanation. Because altruism was difficult to explain by natural selection, which acts only on individuals, Darwin proposed group selection. However, the term 'group selection' is too broadly used and includes population, community, and even species selection. For group selection to operate, all individuals must be related, altruistic, the group must be isolated and the gene flow restricted. Group living may be beneficial when it provides protection from predators (group size decreases predation), but it may also be harmful (it spreads diseases). Therefore, a surrogate model was proposed, that is, kin selection, a form of group selection, which was defined as the natural selection of genes in which the donor benefits its kin and thus favors the survival and reproduction of kin who possess the same genes. Although this process reduces the fitness of the donor, it increases the reproductive fitness of one or more kin and increases the frequency of genes of kin in the future generations.

Three other theories have been offered. First, inclusive fitness (Hamilton 1964) is not only the fitness of an individual and its offspring but also the fitness of *all* kin who share the same genetic constitution. This will insure the reproductive success of an individual, its offspring, and all relatives. Inclusive fitness and kin selection explain the genetic origin of cooperation and altruism.

Second, reciprocal altruism (Trivers 1971) is the genetic behavior causing cooperative behavior in expectation of return of altruistic acts. Reciprocity and most mutual aid are most common among relatives. It was assumed that human disoperation, aggression, reciprocity, kin-selected altruism, and general cooperation are not culturally controlled but are subject to natural selection. Thus, sociobiologists introduced a tit-for-tat concept to explain altruism, and they assumed that tit-for-tat is selected because there must be selection for altruistic acts.

The third theory argued that the unit of selection is a gene, not an individual, and the individual is only a carrier for the gene (Dawkins 1976). Selected genes are selfish and do not act for the survival of species, but only for the survival of the genes themselves. The value element of 'good' and 'bad' was thus replaced with altruism and selfishness as value-free properties. Since sociobiologists strongly adhere to the concept of natural selection, it follows that a struggle for existence, or even an 'arms race' should occur among genes. However, the role of genes in controlling human behavior and the contribution of human embryology to the ontology of behavior are still poorly understood.

Although genes responsible for such human behavior as altruism, homosexuality, and criminality have been claimed by sociobiologists to exist, it is doubtful whether they have been identified. It is also very difficult to differentiate what is genetically or environmentally fixed and what activates behavior, and it is too early to claim a detailed knowledge of the great complexity of behavior or of genes.

3.3 Voices of Assent

Zoology assumes that humans are animals; therefore, after Darwin, it became a proper approach to study the biological aspects of moral behavior, and to accept the comparisons of human with nonhuman social behavior. But, this produces nontrivial consequences. For if humans are animals, the definition 'animal' must include the properties that characterize humans. Thus, human ethical behavior becomes 'natural' and subject to natural selection. Many students of EEB recognize some dependence between evolution and ethics, and accept the role of natural selection in evolution, but are aware of the restrictions placed by natural selection on EEB. Natural selection is only a guiding force, and not the sole source of behavioral (and morphological) changes, and EEB in social groups is only one evolutionary component. The origin of ethical behavior in protohumans and its subsequent development are a subject matter of sociobiology, but the solution of prescriptive ethical problems, of what we ought or ought not do, is a concern of philosophy. While the study of evolution may explain EEB, it does not explain ethical values or goals.

Lumsden and Wilson (1983) emphasized that neither culture nor evolution is autonomous: 'culture is created and shaped by biological processes while the biological processes are simultaneously altered in response to cultural change.' Human ethics works because cultural forces alter the innate bases of

behavior. While genes may control the basic instincts, the behavioral responses to stimuli may be automatic; therefore, the response may not be always driven by natural selection.

The history of the idea of EEB circles around various theories of human nature, the interpretation of which is central to the entire gamut of dichotomies from nature/nurture to brain/mind. The evolutionary psychologist Cyrulnik (1993) called these dichotomies pseudoconcepts, because 'human behavior … is 100 percent innate and 100 percent acquired. Or, what amounts to the same thing, nothing is "innate" and nothing is "acquired." … the acquired can only be acquired by means of the innate, which in turn is always shaped by the acquired. Nature is nurtured!'

What EEB is asking is not prescription, but how we came to think about prescription. Although sociobiologists accept deep evolutionary bases of morality, few of them are comfortable with the concept of 'evolutionary ethics.' The use of this term has begun to disappear, not because all the questions on EEB have been answered, but because the term 'evolutionary ethics' unnecessarily confuses the ethical meanings of 'good' and 'ought' with the biological origin of altruistic behavior.

See also: Altruism and Prosocial Behavior, Sociology of; Cultural Evolution: Theory and Models; Ethics and Values; Evolutionary Theory and Education; Natural Selection; Psychological Development: Ethological and Evolutionary Approaches; Value Pluralism; Values, Anthropology of; Values, Development of; Values: Psychological Perspectives; Values, Sociology of; Virtue Ethics

Bibliography

Allee W C, Emerson A E, Park O, Park T, Schmidt K 1949 *Principles of Animal Ecology*. Saunders, Philadelphia

Caplan A 1978 (ed.) *The Sociobiology Debate*. Harper, New York

Cyrulnik B 1993 *The Dawn of Meaning*. McGraw-Hill, New York

Darwin, C 1859 *On the Origin of Species*. Murray, London

Darwin C 1885 *The Descent of Man*. Murray, London

Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford, UK

Farber P L 1994 *The Temptations of Evolutionary Ethics*. University of California Press, Berkeley, CA

Flew A G N 1967 *Evolutionary Ethics*. Macmillan, London

Hamilton W D 1964 The genetical evolution of social behavior. *Journal of Theoretical Biology* **7**: 1–16, 17–51

Kropotkin P 1902 *Mutual Aid: A Factor of Evolution*. Heinemann, London

Lumsden C J, Wilson E O 1983 *Promethean Fire*. Harvard University Press, Cambridge, MA

Muller, H J 1959 The guidance of human evolution. *Perspectives in Biology and Medicine* **3**: 1–43

Nitecki M H, Nitecki D V (eds.) 1993 *Evolutionary Ethics*. State University of New York Press, Albany, NY

Pinker S 1997 *How the Mind Works*. Norton, New York

Richards R J 1987 *Darwin and the Emergence of Evolutionary Theories of Mind and Behavior*. University of Chicago Press, Chicago

Sahlins M 1976 *The Use and Abuse of Biology*. University of Michigan Press, Ann Arbor, MI

Spencer H 1897 *The Principles of Ethics*. Appleton, New York

Stent G S (ed.) 1980 *Morality as a Biological Phenomenon*. University of California Press, Berkeley, CA

Tinbergen N 1951 *The Study of Instinct*. Oxford University Press, London

Trivers R 1971 The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–71

Wilson E O 1975 *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA

D. V. Nitecki and M. H. Nitecki

Ethical Codes, Professional: Business Codes

Business codes are formal or informal standards of behavior that give structure and direction to economic action and behavior. They codify the values of a corporation, an industry, or a whole sector, and thereby create accountability and certainty of expectation. Conceived of as self-binding acts, business codes provide constraints both to collective organizations (e.g., firms) and individual actors (e.g., management) within the economy. In Western economies, as well as in the emergence of a global economy, they play an important role as governance structures because of, and insofar as they make possible, the expansion of economical and societal cooperation.

1. History and Actuality

Voluntary standards of behavior in the economy as a means of self-binding and constraint can be found in early history. Their emergence is connected to the guilds of craftsmen and merchants in the Middle Ages, where they regulated behavior towards the other members of the guilds, customers, co-workers, and employees. In the beginning they were informal and derived from tradition, common rites, and the customs belonging to a trade. With the expansion of long-distance trading these standards were laid down and were adopted by traveling merchants. Thus they became increasingly formal which meant that they became enforceable by third parties. These early business codes served to ensure contracts concerning

future delivery and payment of goods and were enforceable by guild courts and the threat of ostracism (Mitchell 1969).

To the extent that national and international business laws were established and enforceable by government bodies during the nineteenth century, private standards of behavior based on the threat of ostracism and loss of reputation gradually started to lose their importance. But in the 1970s a renaissance of private standards of behavior emerged from two main sources. The first was the internationalization and globalization of markets and production networks, which were and still are unaccompanied by developments of political, juridical, and cultural institutions and their enforceability. Such business codes include maintaining human rights in cooperative relationships, for example with suppliers and partners, and they define common perspectives in areas where different values standards as well as cross-cultural values conflicts hinder the establishment of transcultural teams.

The second source was the organizational decentralization of corporations, which led to an increase in 'gray areas' and decreasing possibilities of utilizing direct monitoring and control mechanisms. Here business codes have to define a common set of values in order to give the company identity and a behavioral orientation. To reach this aim it needs to form vision, mission, and value statements, compliance standards, and rules for dealing with the company's stakeholders (customers, employees, suppliers, environment) (Weaver 1993). Such standards of behavior can be enforced by sanctions and incentives within the company. With regard to customers and suppliers the companies' reputation and brand names are important enforcing mechanisms.

2. *Minimum Standards, Problem Standards, and Corporate Standards*

Business codes are characterized by self-binding, possibilities of private enforcement, and the identity of the actors. Self-binding by voluntary constraints of action, certainty of expectation in contracts by forceability (sanctions, incentives, reputation), and the creation of corporate identity by means of codified and realized values in business routine make standards of behavior a functional equivalent to national regulations and laws. Whenever national or supernational regulations and their enforcement do not exist or are not feasible, regardless of the identity of an actor, private standards of behavior become a dominating issue of economic and societal governance. They are therefore known as 'soft laws' (Fatouros 1994). By creating certainty of expectation for economic action in societies which are characterized by division of labor, business codes directly influence the state of welfare in any given society. With these reliable and trustworthy rules more cooperation and more transactions are made possible than without them.

The distinction between minimum standards, problem standards, and corporate standards enables us to group modern business codes along with the kinds of cooperations and transactions that are steered by them.

2.1 *Minimum Standards*

Minimum standards are instruments serving the governance of global business transactions. There are, for example, 'Guidelines for Multinational Enterprises' and 'Corporate Governance Guidelines' (OECD 1991/1998), a 'Code of Conduct on Transnational Corporations' (UN 1994), and 'Guidelines on the Treatment of Foreign Direct Investment' (World Bank 1992), to name just a few. They define minimum requirements for such areas as working conditions, involvement of unions, safety at work, consumer protection, competition, technology transfer, investment in developing countries, corruption, and human rights.

Minimum standards emerge from bargaining processes between national governments, national and international organizations of employers and unions, and any nongovernmental organizations (NGOs) involved, directed by a recognized international organization (Fatouros 1994). Their status is precarious. These agreements are obligatory for the signing parties although they cannot be enforced by law. Whether it is possible to implement them therefore depends on the voluntary efforts of the contractors as well as on whether or not it is possible to take action against rule breakers or to damage their reputation in a cost-generating way. Their strength lies in the fact that they provide standards which possess some chance of realization while there is no suitable international law available or accessable. Their weak points are the difficulty of implementing and controlling them. Nevertheless, they are a significant component of economic globalization because they are a functional equivalent to the still underdeveloped international business law. Hence, they create a certain formal and political certainty within the world economy which is a crucial demand for worldwide prosperity.

2.2 *Problem Standards*

Problem standards define proposals of behavior that help to solve certain problems, and go far beyond the rules of the game asserted through the minimum standards. One example from the area of ecological sustainability is the 'responsible care' process of the chemical industry's international associations and companies. From the area of social responsibility the

efforts to abolish child labor and forced labor according to the Council on Economic Priorities' standard 'Social Accountability 8000' may be cited. Problem standards are highly ambitious compared to the potentialities and mentality of many companies. Hence they are accepted by only a few companies, mostly by those which have elaborated them, usually in cooperation with NGOs. This may happen in associations or chambers of commerce or in loose assemblies of those companies for whose business actions the problem in question is instantly relevant. Their strength lies in their model character and in the fact that they make practical experiences concerning these standards possible. Their weakness lies in mostly very difficult controlling and in possible competitive disadvantages for those companies that do not succeed in overcoming the costs of implementing and keeping the standards through prices.

2.3 Corporate Standards

Corporate standards codify the constitutive and action-directing values of a corporation. As codes of conduct or codes of ethics they are components of corporate governance, that is, of the leading, steering, and controlling mechanisms of an economic organization. They are also expressions of the distinct philosophy and policy of the corporation. The values of a corporation can be classified as performance values (e.g., profit, quality), cooperative values (e.g., loyalty, team spirit), communication values (e.g., respect, transparency), and moral values (e.g., integrity, fairness) (Wieland 2000). A firm gains identity and distinctiveness as soon as a specific values mix drawn from these four classes of values endows it with a particular profile. Particularity here means that it fits best with the promises of the company, with its tradition and history, and the aspired direction of future developments. As a rule, these business codes provide regulations for coping with suppliers and customers, shareholders and employees, as well as rules of law and of environment protection. They are part of an overall ethics management system by which the business codes are operated and pursued in corporate practice. Normally this is executed by means of internal and external communication, by training, by ethics officers, and by being incorporated in labor contracts, and the system is controlled by monitoring audits (Weaver et al. 1999).

The systems themselves can be divided into law-driven and value-driven ones (Paine 1994). While the first are compliance programs which focus on the observation of the rules of law and are organized by the legal department of a corporation, the latter as value management systems aim at a comprehensive development of the corporate culture specifically concerning relations with the totality of stakeholders. Codes of ethics therefore never make any statements about existing matters of fact, and they never stand in for the moral consensus of all parties that are interested in the company. It is just the other way around. They define the values of an organization against all stakeholders and declare them conditional for further interaction. As conditions of cooperation with and within an organization these values are shaped like commitments of performance. The company commits itself to these values and promises to continually work on the fulfillment of its commitment. This makes it obligatory for all stakeholders also to bind themselves to the values. In the last decade of the twentieth century, about 80 percent of the big US companies had such codes (Center for Business Ethics 1992). In Europe this must be estimated considerably lower, although there are no exact figures available.

Apart from those reasons for the renaissance of business ethics mentioned in the outset another important factor is legal developments. Because of the introduction in the USA of organizational liability in case of deviant behavior of employees, incentive systems have been installed which make it financially worthwhile to document before the court the implementation and communication of a code of ethics, so that in case of an individual breach of law, the organization cannot be held responsible for it (United States Sentencing Commission 1995). Legal developments similar to this can also be observed in Europe. In this context business codes aim to achieve a higher level of transparency and thereby enhance the means of control in the corporation. The strength of standards of behavior in corporations is that they are highly specific and integrate the involved actors in maintaining them by the implementation of incentives. However, they cause problems for any form of control and by the fact that they have to be introduced into the operative actions of the corporation in a relevant way.

3. Functions and Effect

Although all three kinds of business codes refer to different types of cooperation and transactions they are identical in respect of their functions and effectiveness. These can be differentiated as follows:

(a) Business codes are a functional equivalent to national deficits of regulation by way of directing companies to self-binding and self-control.

(b) Business codes determine and define the identity of those corporations that bind themselves to standards of behavior. These corporations are distinguished from other organizations in the public arena of realization and communication.

(c) Business codes are the basis, as well as a criterion of selection, for corporate decisions. Hence, they define the range of liability a company claims beyond existing law.

(d) Business codes internally and externally indicate the behavioral preferences of an organization and at

the same time communicate a challenge to all its stakeholders to bind themselves to them.

(e) Business codes provide certainty of expectation in business relations and make possible the building up of mutual trust by creating a check against informational uncertainty and opportunistic behavior.

(f) Their prevalence on the one hand relies on economic benefits which arise from the acceptance of standards of behavior. On the other hand one runs the risk of losing economic and social reputation in any case of violation of the legal and moral obligations committed to.

(g) The aspects mentioned—self-binding, identity, responsibility, certainty of expectation, building up trust, reputation—constitute a system of mutual effects which lead to an expansion of economic and societal cooperation and to accompanying gains from trade.

(h) These effects can only be reached by actually maintaining the standards of behavior within the business routine of the company. For that purpose they have to be implemented and integrated into the operative processes of the organization and its management. One of the most important instruments in this respect are ethics management systems.

See also: Aristotle (384–322 BC); Economics and Ethics; Ethics and Values; Ethics for Biomedical Research Involving Humans: International Codes; Kant, Immanuel (1724–1804); Kantian Ethics and Politics; Research Conduct: Ethical Codes; Virtue Ethics; Weber, Max (1864–1920); Wisdom, Psychology of

Bibliography

Center for Business Ethics at Bentley College 1992 Installing ethical values in large corporations. *Journal of Business Ethics* **11**: 863–7

Fatouros A (ed.) 1994 *Transnational Corporations: The International Legal Framework*. Routledge, London

Mitchell W 1969 *An Essay on the Early History of Law Merchant*. Burt Franklin Press, New York

Paine L S 1994 Managing for organizational integrity. *Harvard Business Review* **72**: 106–17

United States Sentencing Commission (ed.) 1995 *Corporate Crime in America: Strengthening the 'Good Citizen' Corporation*. United States Sentencing Commission, Washington, DC

Weaver G R 1993 Corporate codes of ethics: purpose, process, and content issues. *Business and Society* **32**(1): 44–58

Weaver G R, Trevino L K, Cochran Ph L 1999 Corporate ethics programs as control systems: influences of executive commitment and environmental factors. *Academy of Management Journal* **42**(1): 41–57

Wieland J 2000 The ethics of governance. *Business Ethics Quarterly* **11**(1): 73–87

Zadek S, Pruzan P, Evans R (eds.) 1997 *Building Corporate Accountability*. Earthscan, London

J. Wieland

Ethical Codes, Professional: Engineering Codes

Engineering codes of ethics are sets of rules adopted by engineering associations and engineers' professional bodies. They usually include (a) fundamental principles (general guidelines) demarcating the engineering profession; (b) professional rules for engineering practice in dealings with clients, superiors, colleagues (particularly competitors), the profession, and the general public; and (c) procedural rules governing conduct in cases of moral conflict. Codes vary in their degree of legal validity, and range from general guidelines in the form of statements of aims and aspirations, to concrete rules with the status of association bylaws that are binding on all members, on to the vocational regulations of public bodies. Furthermore, many codes do not cover all three elements (a), (b), and (c); as such they are either limited to general principles, restricted to the precepts of professional conduct, or fail to include appropriate procedural rules with a protective function which can be enforced in situations of moral conflict. In particular, many codes fail to clarify the relationship between an orientation to public welfare (a) and the duty to inform the public (b).

1. Tradition

The first codes of ethics for engineers, drawn up in the United States from 1912 onwards, initially consisted only of professional rules of conduct for dealings with clients and competitors (the professional ethos). Although these codes were promptly supplemented by principles stating that public welfare was to be held paramount in the practice of the engineering profession, this step forward was counteracted by the ban on criticism in public of engineering projects and their execution, and by the lack of procedures for the supervision and protection of the engineers affected. Guidelines on environmental sustainability and nature conservation were also lacking in this first phase of development (in 1908, J. Cooke, the president of the US engineering association, foundered both in his goal of combating air pollution by means of binding rules and in his plan to establish judicial proceedings for cases of moral conflict; in 1932, two members of the American Society of Civil Engineers (ASCE) were expelled after they breached the provisions of the ASCE code by disclosing irregularities in dam con-

struction, resulting in the punishment of those responsible; see Layton 1971). While a wealth of different codes were adopted in the United States over the following decades (see Flores 1980, Unger 1982; for an international overview of ethical codes see VDI and Hubig 2000), developments in Europe were rather more restrained. This can be attributed partly to the fact that other organizations were responsible for the control and supervision of technical equipment in Europe (e.g., the technical control board TÜV and its predecessors in Germany) and that the European engineers' associations and the institutes concerned had drawn up legally binding rules for technological design (e.g., DIN, the German Institute for Industrial Standardisation, and VDI, the German Association of Engineers). Furthermore, the more advanced labor and social welfare laws (including protection against dismissal) in force in the European states perform functions that need to be fulfilled by different means in the United States.

In the 1970s, subsequent to a number of prominent scandals in the domain of technical safety (notably the Bay Area Rapid Transit [BART] case, in which engineers who called attention to unsafe aspects of the BART system design lost their jobs despite the fact that their intervention was justified; see Friedlander 1974), it became increasingly clear that codes of ethics needed to include regulations to protect whistleblowers. By the same token, regulations governing conduct in cases of conflict—including the establishment of supportive bodies such as appellate bodies, courts of arbitration, etc.—were needed to prevent engineers being forced into the role of lonely 'heroes' or 'victims' (Alpern 1987, MacCormac 1987, Lenk 1987). The IEEE played a substantial role in these developments. At the beginning of the twenty-first century, codes of ethics are gaining prevalence internationally, but still lag notably behind the much more advanced discussion in the United States: their provisions may be restricted to professional issues, for example, to the domain of general principles (VDI and Hubig 2000).

2. *The Specific Professional Responsibility of the Engineer*

Engineers' codes of ethics have to solve the problem of how the universal moral responsibility which the engineer automatically assumes in his or her capacity as a private citizen can be placed in relation to actual practice in the engineering profession. This first appears to be a problem of application: the engineer is bound to universal moral principles and has to verify whether his or her action is permitted, required, or prohibited by these principles. The engineer's actions may be constrained by the principles themselves. Alternatively, specific internal-technical (instrumental) rules (e.g., functionality and product safety) and (strategic) rules on the maximization of utility and minimization of harm (e.g., use of resources, environmental pollution, long-term societal and environmental impact, etc.) may guide his or her action. Moral conflicts arising here have a similar structure as in other occupational fields.

In addition, some more recent codes take into consideration the specific role responsibility of the engineer. This professional responsibility arises from the engineer's specific knowledge and skills—competencies that cannot be expected of those working in other occupational fields, or of the users of technology, in particular. In addition to bearing the responsibility for the direct consequences of his or her actions in an instrumental-technical or strategic respect, the engineer is obliged to apply his or her specific knowledge and skills to other societal areas by (a) contributing to the early detection of new problematic issues in technological progress and the use of technology, and warning and advising accordingly; (b) calling attention to new alternatives arising in the course of technological development; (c) offering honest criticism of legal provisions with respect to their application and practicability and calling attention to alternatives; (d) using their technical expertise to help draw up standards for technological design and contribute to the discussion of technology-related assessment criteria, and finally (e) preventing foreseeable misuse in the lay public's dealings with technology and limiting the scope for deliberate abuse of technical products and procedures (e.g., by mechanisms to prevent misuse by unauthorised persons—see draft VDI code 2000).

3. *The Values Underlying Engineering Codes of Ethics*

In view of traditional and cultural differences in the basic ethical orientations (value pluralism), most codes of ethics are based on fundamental principles which appear justifiable from various perspectives, and which express a kind of moral common sense. The decisive values and value orientations are thus expressed largely in general and formulaic terms. This may well lead to problems. On the one hand, the potential for conflict which may arise as a result of differing interpretations of the set phrases remains hidden; on the other hand, the fact that the values are immanently contradictory and likely to lead to conflict is often overlooked when general lists of values are drawn up. Rules applying to the narrower domain of professional practice (see point (b) above) focus on the values of loyalty, fairness, solidarity, and contractual fidelity. Here, potential for conflict is most likely to arise as a result of competing commitments to the parties involved: loyalty to clients, superiors, the law,

etc.; solidarity with colleagues, the profession, the users of technology, etc. The fundamental principles of the majority of codes (see point (a) above) refer to public welfare and, more recently, to environmental quality and nature conservation. Reference is also made to individual health, safety and welfare, and universal moral rights (human dignity, human rights). Conflicts between these general value orientations have a similar structure as those in domain (b). Immanent conflicts in matters of public welfare result from the difficulty of identifying collective preferences which are compatible with the conditions of liberalism on the one hand (e.g., Arrow 1963) and from differing interpretations of total utility as either average utility or aggregated utility on the other. Conflicts between the basic values are also apparent in the debate on sustainability in technological design, for example. Further potential for conflict arises when individual and public welfare have to be weighed up against one another in cases of rationalization or rationing, for example, or in the familiar conflicts of interpretation on human rights matters. Unless the resolution of such conflicts is regulated by fixed value hierarchies within the codes (determining, for example, that loyalty to the law takes priority over loyalty to clients), it will only be possible to settle such value conflicts by discursive means. The engineer is not a privileged partner in such discourses, however, and his or her responsibility here lies in the willingness to engage in such discourses (solely the draft VDI code (2000) takes steps in this direction).

4. *Institutionalization of the Engineer's Ethical Responsibility*

In the complex processes of technology generation and use, it is only in the rarest of cases that the engineer as an individual bears sole responsibility for the consequences of his or her direct actions, the role he or she occupies, or way he or she performs this role. The individual engineer possesses only a certain amount of specialist knowledge and has only limited authority to act. However, this does not exonerate the individual from a share of the responsibility for the results of complex actions in which he or she was involved, or from a share of the responsibility for the institutional actions of the organizations (companies, associations) to which he or she belongs, or the decision-making processes in which he or she participates. Individual actions are embedded into the institutional framework in various ways in different codes of ethics. Some codes stipulate institutionalized procedures to relieve, support, and protect the individual engineer in cases of moral conflicts that he or she does not have the power to resolve, and designate appellate and advisory bodies, hotlines, disciplinary procedures, and incentive schemes to which the engineer has recourse in cases of conflict. If no such measures are stipulated, the provisions of the code may be embedded into the institutional framework in the form of a supplementary bylaw. Only then is the association able to implement the code and regulate its members' conduct accordingly (e.g., the IEEE Member Conduct Committee, which interprets the provisions of the code in conjunction with an ethics committee, or incentive systems such as the Carl Barus Award for Outstanding Service in the Public Interest). Provision is also made for training programs and support in legal disputes.

5. *The Legal Status of the Codes*

Codes which take the form of statements of general aspirations and are formulated in terms of fundamental principles are not directly legally binding. Even in this form, however, they have unequivocal legal relevance. First, they provide a point of reference for the filling of unlegislated areas left open by statutory regulations, and thus constitute an important complement to the orientation function of the law. Second—and even more importantly—codes play an important role in the interpretation of gray concepts in labor law, vocational regulations, and the provisions of administrative and civil law when such gray legal concepts as 'public morals,' 'endangering the health of others,' 'significant material assets,' 'state of the art,' 'duty of care for product safety,' etc., need to be more clearly defined in terms of certain standards of responsibility. These standards can be derived from the norms of the engineering profession if these are established in a corresponding code of ethics. Engineers can then refer to the code when making decisions, and the courts can follow the guidelines of the codes when interpreting gray legal concepts.

A code has a greater degree of legal validity if it takes the form of an association bylaw which is binding on the association's members. This makes it possible to protect members from competing demands and expectations and to take disciplinary action in cases of misconduct. Arbitral or association jurisdiction (as is the case in sports clubs, for example) is of course required for a code with the status of an association bylaw to be enforced. A code attains the greatest degree of legal validity as part of the vocational regulations of a professional association which is also a public body (e.g., Federal Chamber of Engineers [*Bundesingenieurkammer*] 1998). All those practicing the profession in question are bound to the code by means of compulsory membership of the association. By giving vocational regulations precedence to all contractual obligations under civil law, but not to labor law, the state surrenders part of its jurisdiction to the profession, thus offsetting the limits of its own powers.

In the United States, in particular, a further link between ethical codification and the law can be observed: under the terms of the Federal Sentencing Guidelines for Organizations (FSGO), the question of whether or not an organization or corporation has subscribed to a code of ethics and corresponding enforcement program ('effective ethics program') in the form of a bylaw is taken into account in sentencing. Under the guiding principle of 'good corporate citizenship,' the state acknowledges that it is not itself in a position to direct corporate action by means of the existing framework regulations, and determines the sanctions to be imposed by considering the corporate will to cooperate in an ethical approach to technology. In Great Britain, the 'Public Interest Disclosure Bill' assures the disclosure of moral conflicts and designates appellate bodies and entities.

Finally, some more recent codes of ethics include an additional guideline which goes some way to determining the engineer's position with regard to legal systems. The 'qualified obedience to the law' required here moderates the effect of external legal regulations. In the problematic case of a formal interpretation of the law (e.g., with respect to new and as yet unforeseen technologies), precedence is given to the intention of the law. Furthermore, problematic legal regulations are to be checked with respect to their compatibility with civil liberties and human rights. In conflicts such as this, the individual engineer is in acute need of the support of an association that has adopted such a code.

See also: Engineering Psychology; Environmental Health and Safety: Social Aspects; Ethical Codes, Professional: Business Codes; Ethics for Biomedical Research Involving Humans: International Codes; Professionalization/Professions in History; Safety, Economics of; Technological Innovation; Technology and Social Control; Technology, Anthropology of; Workplace Safety and Health

Bibliography

Alpen K D 1987 Ingenieure als moralische Helden. In: Lenk H, Ropohl G (eds.) *Technik und Ethik*. Reclam, Stuttgart, pp. 177–93

Arrow K 1963 *Social Choice and Individual Values*. Yale University Press, London

Baum R J 1980 The limits of profession responsibility. In: Flores A (ed.) *Ethical Problems in Engineering*. Rensselaer Polytechnic Institute, New York, Vol. 1, pp. 48–53

Baum R J (ed.) 1980 *Ethical Problems in Engineering*. Center for the Study of the Human Dimensions of Science and Technology, Rensselaer Polytechnic Institute, New York, Vol. 2

Deiseroth D 1997 *Berufsethische Verantwortung des Ingenieurs*. Lit, Münster

Flores A (ed.) 1980 *Ethical Problems in Engineering*. Rensselaer Polytechnic Institute, New York, Vol. 1

Friedlander G D 1974 The case of the three engineers vs. BART. *IEEE Spectrum* October: 69–76

Hubig C 1995 *Technik- und Wissenschaftsethik. Ein Leitfaden*. Springer, Berlin, Heidelberg, New York

Johnson D C (ed.) 1991 *Ethical Issues in Engineering*. Prentice Hall, Englewood Cliffs, NJ

Layton E T 1971 *The Revolts of the Engineers. Social Responsibility of the American Engineering Profession*. Hopkins, Cleveland, OH

Lenk H 1987 *Ethikkodizes für Ingenieure*. In: Lenk H, Ropohl G (eds.) *Technik und Ethik*. Reclam, Stuttgart, pp. 194–221

Lenk H, Ropohl G V (eds.) 1987 *Technik und Ethik*. Reclam, Stuttgart

Liedtke R, Meihorst W, Wendelin-Schröder U (eds.) 2000 *Der engineer-Eid. Ethisch-naturphilosophisch-juristische Perspektiven*. Scientic Nova, Bretten

MacCormac E B 1987 Das Dilemma der Ingenieurethik. In: Lenk H, Ropohl G (eds.) *Technik und Ethik*. Reclam, Stuttgart, pp. 222–44

Schweizerische Akademie der Technischen Wissenschaften (SATW) 1991 *Ethik für Ingenieure/Technische Wissenschaftler*. SATW, Schriften 16, Zürich

Unger S H 1982 *Controlling Technology. Ethics and the Responsible Engineer*. John Wiley & Sons, New York

VDI, Hubig C (eds.) 2000 *Ethische Ingenieurverantwortung. Handlungsspielräume und Perspektiven der Kodifizierung*. VDI, Düsseldorf

C. Hubig

Ethical Dilemmas: Research and Treatment Priorities

1. Ethical Dilemmas in Moral Theory

1.1 Causes and Kinds of Dilemmas

Dilemmas appear in situations where (a) moral principles seem difficult to apply to the case at hand, or (b) where various moral principles clash and/or do not guide action.

As to the first case, ethical dilemmas arise due to an epistemic uncertainty like risk assessment or evaluation of the ethical implications of scientific and technological change and/or due to conflicting obligations in light of the new technological advances that support different courses of action. Two kinds of ethical dilemmas can be distinguished: an agent faces conflicting obligations which he cannot both fulfill, and neither obligation is overridden: he ought to do A and he ought to do B, but cannot do both A and B. For example, a father is unable to rescue both of his children from a burning house. He has an obligation to rescue his first child, and he has an obligation to rescue his second child, but neither obligation outweighs the other. Or, moral reasons speak equally for

and against an option with respect to different features of that option: and agent ought to do C and ought not to do C. A doctor, for instance, who has an obligation to provide medical assistance, faces such a dilemma if his help implies a questionable prolongation of a patient's life that might be difficult to assess. No matter how the person in moral dilemmas acts there will be moral losses.

In the second case, an ethical dilemma is due to disagreement about which moral convictions and principles like 'maximize the expected utility' or 'treat a human being as an end in itself' apply to the case. Thus, there is a deeper difficulty about the application of moral principles to practice, if there is disagreement about the principles or if there are several principles that are applicable to the respective case and yield different solutions. There is no standard way of weighing the contending principles in order to decide which predominate in specific cases. Such dilemmas constantly arise in medical practice where it is difficult to assess whether life-saving treatment regards the patient as an end in itself, and should therefore be provided, or whether it disregards patient interests and values.

1.2 Epistemic Dilemmas

Ethical theorists have concentrated on the question about which there is widespread disagreement whether ethical dilemmas can be solved. Some have tried to tackle the problem of moral principles that, if jointly applied, lead to inconsistencies. Two principles are inconsistent, if they cannot be both fulfilled.

Two moral principles in particular are alleged as being unable to guide action in light of seemingly conflicting obligations. (a) The so-called 'agglomeration principle' entails that an agent S is obligated to do A, and is obligated to do B, and therefore is obligated to do both A and B: $[(OA \& OB) \rightarrow O (A \& B)]$. (b) The principle of deontic logic 'ought implies can' infers from the premise that an agent S is obligated to do both A and B that the agent *can* do both A and B: $[(O (A \& B) \rightarrow c (A \& B)]$.

These two principles contradict the standard definition of ethical dilemmas according to which an agent is obligated to do A, and is obligated to do B, but *cannot* do both A and B. Some theorists have tried to differentiate between *prima facie obligations* and *all things considered obligations* in order to show that not every obligation requires to be acted upon. A so-called *deliberative ought* (B. Williams in Gowans 1987) is supposed to help us decide which obligation overrides a seemingly conflicting obligation by referring to stronger reasons (P. Foot in Mason 1996). Dilemmas are thus described to be merely 'epistemic.' Moral principles are shown to be consistent by distinguishing between preliminary and final (R. M. Hare in Gowans 1987) or general and specific (A. Donagan in Gowans 1987) principles. According to R. M. Hare (in Gowans 1987), for example, conflicting prima facie duties are selected on a 'critical level' with the help of actutilitarian considerations. A generally accepted rule might thus be overridden in light of the case at hand. A. Donagan (in Gowans 1987) proposed to specify principles with additional premises in order to provide exceptions to such principles if they seem to be justified. A promise, for example, can be broken if a more urgent and morally defensible action calls for it.

1.3 Are Dilemmas Genuine?

Even though many ethical dilemmas might be solved in such a way, there still remain cases in which no 'deliberative ought' can show which prima facie obligation outweighs the other. Furthermore, it can be maintained that even obligations that are supported by less weighty reasons might still remain in force and do not cease to obligate. There are three arguments that try to support the hypothesis that ethical dilemmas can be insoluble. (a) So-called residual emotions like regret that agents express after a choice between conflicting obligations are purported to show that the obligation not acted upon remains in force like an unfulfilled desire (B. Williams in Gowans 1987). (b) Conflicting oughts are considered to stem from plural sources of value, e.g., obligations, rights, utility, perfectionist ends, and commitments. Hence, if values (and also moral principles) are plural, dilemmas seem to be unavoidable (T. Nagel in Gowans 1987). (c) Feelings of guilt that agents display after a difficult choice are taken to motivate the avoidance of conflicts in the future. They show that dilemmas are real (R. B. Marcus in Gowans 1987; in Mason 1996). As to the solution of dilemmas the considerations of ethical theorists are also constantly challenged in light of the practical problems that technological research, for example, poses.

2. Ethical Dilemmas in Scientific Research

2.1 Risk and Technology Assessment, Professional Responsibility, and Moral Status

Ethical dilemmas arise in many areas of research, particularly where the development and use of new technologies challenges the application of moral concepts, or leads to conflict between moral principles and demand for their extension or further specification. Thus, for two conflicting courses of action different moral arguments, rules or principles can be equally compelling, or their moral costs are difficult to assess. This pertains to problems of risk and technology assessment, the scope of professional responsibility, and to questions of autonomy and the moral status of those affected by these technologies. Ethical dilemmas typically arise in the controversies surrounding life-

saving technologies at the beginning and end of life, life-enhancing technologies to improve the quality of life, the technologies of reproduction, and genetic engineering.

Technological and risk assessment can be dilemmatic in the following way: in the early stage of a technological development, when it is still relatively easy to control its direction, there is an inevitable lack of knowledge to exercise reasonable control; by the time more experience and better understanding of the risks involved is gained, the technology has often developed so much social momentum that it is almost impossible to control it (Collingridge 1980).

The responsibility of scientists can be said to consist of the pursuit of truth in their own area of competence, for wielding social power appropriately, and for making their results generally accessible. Professional responsibilities can lead to moral conflict if, for example, research results and/or the effects of technologies have a potential to endanger others. For example, scientists can have a duty to pursue truth, and they can equally have a duty to avoid harm, and they might not be able to fulfill both. The justification of scientific investigation can only be provided from within scientific research, since expertise is needed to evaluate it. This can undermine social control of research (Erwin et al. 1994).

The application of moral principles to practice also sheds light on the question who or what has autonomous moral status with regard to fetuses, those in permanently vegetative states, members of future generations, animals, and inanimate nature. For example, is research on human embryos permitted in order to foster medical therapy and relieve suffering or do we also have obligations to embryos? Particular moral uncertainty is thus engendered by contexts that demand divided allegiances. For a more detailed treatment of professional responsibility and control, see *Censorship in Research and Scholarship*; *Research Conduct: Ethical Codes*; *Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives*; *Ethics Committees in Science: European Perspectives*; *Research: Ethnical Aspects of Long-term Responsibilities*; *Objectivity of Research: Ethical Aspects*.

2.2 Ethical Dilemmas in Biotechnology

The far-reaching potential advances in gene therapy and genetic engineering for humans (The Human Genome Project), and the implications for humans of cloning, have given rise to ethical dilemmas scientists have to face. The following examples can illustrate how progress in genetic engineering generates dilemmas between conflicting obligations and/or conflicts due to risk assessment (Chadwick 1992).

The identification of human genes, for example, can lead to the following conflict: on the one hand genetic knowledge enhances therapies for hereditary disease, on the other hand, it poses problems about the potentially exploitative use of resources and genetic information. The attempt to sequence the entire human genome raises ethical questions whether the risks of exploitation outweigh the benefits of knowledge.

Genetic alterations passed on to future generations through so-called germline therapy raise further problems regarding consent while, at the same time, 'improving' human genetic potential for future generations. Do we have an obligation to present generations to relieve suffering by seeking treatments for genetic disease or do future generations have a right to an unmodified genetic inheritance? Equally, the advantages of genetic screening can be outweighed by the costs of stigmatization on the basis of a person's genetic make-up. Such individuals might find themselves unemployable or uninsurable. There also arises the question whether the very existence of genetic screening can exert pressure on individuals with regard to their reproductive decisions, thus impairing their autonomy. How can genetic public health be fostered without practicing eugenics? Arguments about the moral urgency of relieving suffering also conflict with arguments about human dignity that discredit the production of 'designer babies' (Annas and Elias 1992).

The incorporation of foreign genes into the genome of an organism is commonly discussed in connection with animals and plants. One dilemmatic issue concerns the interests of the host organism (particularly in the case of animals), the consequences for human health and for other species, and the risks of releasing genetically engineered organisms into the environment.

A related issue has been the matter of justification of treatment. Animal experimentation poses the question whether animals have moral status to the detriment of life-enhancing research results for humans. As subjects of genetic engineering, for example, farm animals have suffered from unintended deleterious effects, while research animals have suffered the consequences of being intentionally bred for propensity to develop debilitating diseases. A further important issue is related to the question whether an agent who had moral status can cease to have that status. The human cases of the brain dead, anencephalic infants, and those in a permanently vegetative state are cases in point. If so, then the case of xenotransplantation or the transplanting of animal organs into humans is affected. The interlocking questions of moral standing, justification of treatment, and loss of moral considerability can thus cause conflicts as to which consideration should be given more weight. Ethical dilemmas in research are thus a challenge to those within and those outside research, to debate whether research practices and their effects are right and just. For further treatment see *Animal Rights in Research*

and Research Application; *Bioethics: Examples from the Life Sciences*; *Euthanasia*; *Genetic Counseling: Historical, Ethical, and Practical Aspects*; *Reproductive Medicine: Ethical Aspects*; *Research Subjects, Informed and Implied Consent of*.

2.3 Limits to the Application of Moral Principles

Moral rules and principles do not only have to be adapted to the specific circumstances at hand. They also have to be questioned in light of practice. However, in many of the cases that feature in research ethics, often times we do not agree in our moral principles or the moral theories by which they are justified. Even if the application of principles to cases can be specified and thus epistemic dilemmas resolved, there are limits as to what the application of principles can do to resolve dilemmas: the basic problem of dilemmas remains at least in one respect. There is no agreement as to how to balance between various conflicting moral principles in order to decide which predominate in a respective case. Intuitions vary widely, and various normative theories like utilitarianism, Kantianism, contractualism, virtue theory, and feminist ethics usually remain at odds with each other.

It seems desirable to take into account as many considerations as possible on the level of principles themselves in order to highlight the various ethical implications of a case at hand. Not only questions of utility, but also questions of integrity and care can thus be given more weight. Further work, however, needs to be done at the level of ethical theory itself to set out how various principles derived from different normative theories can be balanced.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Health Care, Rationing of; Research Conduct: Ethical Codes; Research Ethics: Research

Bibliography

Annas G J, Elias S (eds.) 1992 *Gene Mapping: Using Law and Ethics as Guides*. Oxford University Press, New York

Beauchamp T L, Childress J F 1994 *Principles of Biomedical Ethics*, 4th edn. Oxford University Press, New York

Chadwick R F 1992 *Ethics, Reproduction and Genetic Control*. Routledge, London

Collingridge D 1980 *The Social Control of Technology*. St. Martin's Press, New York

Erwin E, Gendin S, Kleiman L (eds.) 1994 *Ethical Issues in Scientific Research*. Garland, New York

Gowans C W (ed.) 1987 *Moral Dilemmas*. Oxford University Press, New York

Mason H E (ed.) 1996 *Moral Dilemmas and Moral Theory*. Oxford University Press, New York

Nida-Rümelin J (ed.) 1996 *Angewandte Ethik*. Aröner, Stuttgart, Germany [1996 *Applied Ethics*]

Sinnott-Armstrong W 1988 *Moral Dilemmas*. Blackwell, Oxford, UK

Stocker M 1990 *Plural and Conflicting Values*. Clarendon, Oxford, UK

M. Betzler

Ethical Issues in the 'New' Genetics

Genetics, both as a field and as a word, was born in the early twentieth century. The 'new' genetics refers the vast increase in the power of genetics that followed the increased understanding of the biochemical basis for genetics in deoxyribonucleic acid (DNA). That understanding has made it possible to 'read' the sequence of DNA in organisms, including humans; to discover, in some cases, the physiological significance of that sequence; and to contemplate changing the sequence in order to attain desired ends. These heightened powers have both created new concerns and exacerbated older ones.

Concern about implications of the new genetics for human societies grew with scientists' ability to decipher and manipulate DNA. The late 1960s saw discussions of human cloning, followed in the 1970s and early 1980s by safety concerns about recombinant DNA and social concerns about genetic discrimination. When the Human Genome Project was proposed for substantial funding in the United States in the late 1980s, the Project's first director, Dr. James Watson, suggested that 5 percent of the funding be set aside to study the 'ethical, social, and legal implications' of human genetics (Cook-Deegan 1994). The subsequent financial commitment led to the rapid expansion of 'ELSI' studies across many disciplines and the publication of a vast number of articles and books on these topics.

Three general caveats are useful in surveying the resulting discussions. First, although new genetic technologies will sometimes provide great predictive power, oftentimes their predictions will be weak. The percentage of people with a particular set of genetic variations (or 'genotype') who develop a particular 'phenotype' (trait or condition), a percentage known as 'penetrance,' often will be low. Second, although the technologies are quite novel, their implications almost always have parallels in the social effects of other modern technologies. Third, although much of the discussion of the consequences of new genetic technologies has focused on individuals or families, these technologies usually also have implications for broader human groupings. With those caveats in mind, this article will discuss the ethical, social, and legal effects of the new genetics in seven areas: health care, establishing identity, predicting the future, re-

vealing the past, genetic manipulation, ownership and control of genetic material and information, and cultural understandings.

1. Issues Arising from Medical Successes

Writing on social issues of genetics focuses on the dangers of human genetics. It rarely looks at the implications of the hoped-for uses of genetics to prevent and treat disease. The vast sums, private and public, spent on research into genetics are not being committed purely in a search for knowledge, but in the expectation that the research will bring immense medical benefits. The diseases in question may not be limited to 'genetic diseases'; the tools of the new genetics are allowing unprecedented understanding, at the molecular level, of how human cells and human pathogens function. The most powerful lesson of the first 25 years of the biotechnology industry has been that the human body and its functions are more complicated than expected, but those complications are steadily being tackled by the new genetics. A steady stream of new treatments, derived from increased knowledge of genes and their associated proteins, is flowing towards approval.

The results might include a significant extension of average human lifespan; an increase in pharmaceutical costs; and changes, up and down, in the demand for various medical services. The social implications could be substantial on pension plans, voting blocks, and health care systems. Some more focused medical consequences of genetic research such as individualizing patient therapies as a result of genetic tests (called 'pharmacogenomics') or some kinds of proposed gene therapy or tissue replacement will require that treatments be created just for one patient, straining the existing mechanisms for drug development, manufacturing, approval, and financing. Any new treatment will raise questions of availability, both within a country and between countries. The new genetics has created a uniquely high expectation of medical progress; its successes therefore may raise all these issues in dramatic form.

2. Establishing Identity

The new genetics raises three issues centering around identity: forensic identification, personal identity, and ethnic identity.

2.1 Forensic Identification

The new genetics has already led to major changes in forensic identification. Any tissue from a person that contains DNA can be tested for a pattern of identifying markers, sections of the genome that vary from one person to another. Those markers can be compared with markers analyzed from tissue found in connection with a crime (National Research Council 1992). Powerful techniques for analyzing DNA mean only tiny amounts of genetic material are needed, derived from flesh, blood, semen, or even cells found in saliva. A negative test is very powerful. If the markers are significantly different, the tissues cannot be from the same person. Initially, there was concern that a positive match might not be compelling because ethnic groups might share common patterns of markers. These fears have been assuaged through more information on the distribution of the markers. (National Research Council 1996).

Forensic use of DNA is not essentially different from the use of fingerprints, blood groups, or dental records. Like those techniques, it has had to prove itself in judicial cases. Like the earlier techniques, it can never be conclusive. Good defense counsel will always investigate the possibilities of an innocent explanation for the presence of the DNA or for mistakes, contamination, or fraud in the process of collection and analysis.

The real issues about the forensic use of DNA now concern not individual cases, but the collection and retention of DNA or DNA records for identification purposes. Many countries or states are creating DNA repositories with DNA samples from convicted, or even merely accused, criminals. These samples can be used to seek matches in criminal cases or for the identification of human remains. If markers from the samples are analyzed in advance, the results can be put into a database and used to seek 'cold matches' to samples or tissue of unknown origin. These repositories raise two different issues. The first involves their creation. Should people be forced to give samples? If so, who should be required to provide samples—convicted sex offenders, convicted felons, those arrested, or the whole population? A more subtle, but equally important, issue concerns *what* the repository should contain—a DNA sample or only the marker analysis. The markers used for forensic identification are not associated with any traits. If the markers are kept, no information about a person other than identity can be discussed. If samples of whole DNA are kept, however, they could be analyzed for DNA associated with a wide variety of characteristics, raising serious privacy concerns.

2.2 Personal Identity

The furor about the possible use of somatic cell nuclear transfer technology to clone humans involves many issues (National Bioethics Advisory Commission 1997), most of which have already been raised by *in vitro* fertilization. The issue unique to cloning is a

concern about human identity. Would a person with the same genome as another living individual be a copy of the first person?

Monozygotic twins have always shared the same genome. Although their physical similarity is strong, no one doubts that they have individual identities and separate personalities. Cloning by somatic cell nuclear transfer could not produce individuals *more* alike than monozygotic twins. The clone and the source of the clone's DNA would most likely be much less alike. They would develop in the wombs of different women, subject to different environmental influences *in utero*. Once born, they would be subject to different environments, often in different decades. The degree to which identical human genotypes born from different wombs produce similar phenotypes would be an interesting research question, if it could ethically be studied. The popular perception of identity would likely prove grossly exaggerated.

2.3 Ethnic Identity

Ethnic groups often share both a common culture and a substantial degree of common ancestry. This genealogical connection implies a genetic connection; particular variations in DNA sequences, or patterns of those variations, will be more common in some populations than in others. For example, each of the four ABO blood groups, which are genetically determined, are found in all (or almost all) of the world's populations, but in different proportions in different groups. Other variations in DNA sequence may be found at high levels in some groups but rarely in most others (Cavalli-Sforza et al. 1994).

This kind of variation can be used for anthropological research; it could also be used to try to define a person's ethnic identity. Of course, no known DNA variations are found in every member of an ethnic group and never found in nonmembers. Given the human history of frequent intermixing, adoption, and conversion, no such variations are likely. The ethical implications of the naïve use of such inherently inaccurate methods for defining ethnicity will depend on the circumstances, including whether they were adopted by the group itself or imposed on it.

3. Predicting the Future

The most discussed issues arising from the new genetics come from its ability to predict individuals' futures as a result of associations between genetic variations and physical and behavioral traits. The strength of the predictions will vary dramatically. Some conditions, such as Huntington's disease, follow inevitably from possession of a particular genetic variation. For other conditions or traits, a person with a particular genetic variation will have a somewhat increased but still small likelihood of having that condition or trait. The ethical significance of such predictions will vary largely based on who uses what predictions for which ends.

3.1 Prenatal Selection to Avoid Disease

The new genetics brings the ability to test the DNA of the parents and of the embryos and fetuses they create. Where the genotypes are strongly or completely associated with particular phenotypes, whether the children might (when the parents are tested) or would (when the embryos or fetuses were tested) show those traits can be confidently predicted. This kind of prenatal testing has been used for several decades to avoid serious genetic diseases such as sickle cell anemia and Tay–Sachs disease. In some cases prospective parents have been tested to alert to them to whether such a disease is a possibility for their children. This carrier testing is simple and relatively uncontroversial.

Testing of fetuses is possible only with genetic material from the fetus, currently retrieved from the amniotic fluid or from the chorionic villi well into a pregnancy. Both procedures are expensive and somewhat risky. After a direct genetic test of the fetus, the parents could choose to abort or could prepare themselves better for the birth of a child with the particular genetic condition. Testing of fetuses for genetic conditions has been more controversial. Some oppose it because it often leads to abortion, which they condemn as murder. Advocates for the disabled argue that this testing implies that the lives of people with those genetic conditions are not worth living. In the United States, abortion, at least through the second trimester, is legal for any reason, so there has been little discussion of outlawing abortion based on genetic tests; many countries do regulate the reasons for abortion and must face this issue. Even in the United States, a question remains as to whether prenatal genetic testing for such conditions should be encouraged or discouraged.

A recent technical advance may change discussions about prenatal genetic testing. 'Pre-implantation genetic diagnosis' makes it possible to test an embryo before it is implanted. This procedure can only be done with *in vitro* fertilization. One cell is detached from an embryo and the DNA in that cell tested. Based on the results, the parents can decide whether to implant that embryo or not. No fetus would be aborted; instead, some embryos would just not be implanted.

3.2 Prenatal Selection for Genetic 'Enhancement'

Pre-implantation genetic diagnosis increases the possibility of parents selecting children based on genetic traits that would 'enhance' their children. The idea of

using genetic testing to select 'enhanced' genetic traits for offspring has raised many concerns (Parens 1998). Some fear that allowing parents to choose traits for their children would deny those children their right to 'an open future' not of their parents' choosing. Others worry that the rich will be able to buy unfair genetic advantages for their offspring. Still others worry that parental selection will lead to a homogenization of the human gene pool, as parents opt for children with similar traits. Finally, some fear that these parental choices could lead to a self-perpetuating genetic caste system. These results can only follow if genetic variations are identified that strongly predict 'success' and if many parents are willing to undergo pre-implantation genetic diagnosis. Neither is certain.

One might also question whether parents choosing to enhance their children's lives through genetics is fundamentally different from efforts to do the same through environment. Rich parents send their children to 'good' schools. Parents may push their children from an early age to excel in certain activities—Wolfgang Mozart and Tiger Woods are both successful examples. Arguably, a major goal of parenting is to 'deny' one's children certain unhappy or illegal futures. The argument that enhancement through genotype selection is inherently more threatening than other kinds of enhancement is not obvious. (Sex selection is a special case of 'genetic enhancement,' though one that is also done without genetics, by sonography.)

3.3 State-sponsored Prenatal Genetic Selection

The discussion thus far has focused on parental choices about their children's genetic make-up, but those decisions need not be made by parents. Governments have intervened to prevent people with 'inferior' genes from having children; such eugenics laws could come again in more precise forms. 'Eugenics' is discussed in detail in other articles and will not be reviewed here except to note two things. First, state actions to encourage prenatal genetic selection without compulsion can also be 'eugenics.' Second, the ethics of government-sponsored eugenics depends on striking a balance between government public health powers and individual procreative liberty that remains in debate.

3.4 Postnatal Genetic Testing

The possibilities of genetic testing do not end at birth. Genetic tests can help diagnose diseases or predict future disease risks. Like all medical interventions, their value depends on their appropriate use (President's Commission 1983, Institute of Medicine 1993). Postnatal genetic tests provide information that can lead to a medically useful interventions, that can help the patient make life plans, and that might just satisfy a patient's desire to know her future. Genetic testing also has costs. Test results may pose psychological problems for those tested or affect the patient's relationships with family members. They may also have implications for the patient's dealings with the broader society, through employment or insurance discrimination. In the absence of useful interventions, the benefits from genetic testing depend greatly on the circumstances and personality of each individual. Where tests, genetic or otherwise, do not have powerful implications for medical intervention, the strong influence of the individual's personality and circumstances on the value of testing argues for a particularly good process of informed consent, especially when the patient's understanding of the links between genetic variations and disease may be weak. Such testing is particularly suspect for children, if the benefits of such testing could be put off until they reach adulthood and make their own informed decision.

Of course, many nongenetic medical tests predict higher or lower disease risks. The diagnosis of a disease, whether through genetic or nongenetic methods, can trigger many of the same costs as prediction. Genetic diseases do implicate family members to an unusual extent, but shared environments or diets also put family members at shared risk.

3.5 Genetic Discrimination

Great concern has been expressed about the possible uses of genetic tests to discriminate against people based on genetically predicted susceptibilities. Employment, life insurance, and, in the United States, health insurance are the fields where genetic discrimination is most feared. The likelihood of genetic discrimination depends on two different sets of factors: the strength of the prediction possible from genetic information and the social structure governing the relevant fields.

For someone with the allele that leads inevitably to the fatal condition Huntington's disease, genetic prediction is quite powerful. Fortunately, such powerful genetic predictions are rare; it is not clear that employers or insurers would find significant an increased risk of diabetes from 5 percent to 8 percent. Similarly, discrimination in health insurance is mainly a concern in the United States; all other rich countries guarantee all, or almost all, their residents health coverage. Even in the United States, the great majority of Americans with health insurance obtain it through methods that are not susceptible to genetic discrimination. Medicare, Medicaid, and employer-provided health insurance are not medically underwritten—they must accept everyone eligible without using any

medical exclusions. There is little good evidence that genetic discrimination in employment or insurance exists in the United States, although there is clear evidence of public fear of such discrimination.

Whether such discrimination turns out to be common or rare, the question remains what to do about it. Thus far, no general federal law bans genetic discrimination in insurance or in employment (although the opaque Americans with Disabilities Act may prohibit some employment discrimination based on genetic characteristics). Most US states have limited the use of genetic information in health insurance, many have limited their use in employment decisions. No states have forbidden their use in life insurance (Bobinski 2000). The United Kingdom, where the National Health Service makes private health insurance of limited importance, has set up a regulatory authority to pass on whether or not life insurers may require tests for particular genetic conditions, recently allowing the use of Huntington's disease testing. How well statutes limiting genetic discrimination will work remains unclear. If statutes banning genetic discrimination can function with low costs, one might justify them on the ground that they increase the research on and clinical use of human genetics by reducing public fears, whether or not those fears are justified (Greely 2001).

3.6 *Prediction of Behavioral Traits*

The successes of the new genetics have largely been with physical traits or conditions. Public interest in genetics, though, seems to lie disproportionately in behavioral traits. Some human behavioral traits have been linked to genetic variations, including some forms of mental retardation. For the most part, though, efforts to link genetic variations with human behavioral variations have, thus far, had little success. If such predictions were possible, the ethical implications could be substantial (Carson and Rothstein 1999).

Strong links between genetic variations and behavior would undercut notions of personal responsibility for the involved behaviors. At the broadest level, they could affect society's view of the extent of individual free will. More narrowly, they could affect society's view of, for example, 'genetically determined' criminal behavior or sexual orientation, although whether the result would be more or less tolerance for the behavior is not clear. These kinds of genetic predictions could also lead to interventions in individual's lives—people predicted to act violently might be put into custody before they committed any offense. The successful association of behaviors with genetic variations would boost the interest of both parents and governments in prenatal genetic selection.

Societies make predictions about future individual behavior in many ways. Except prenatally, it is not clear that doing a genetic test to determine, for example, musical ability, would offer any advantages over a direct test of such ability. Genetic tests would have an advantage only when the prediction is strong and the behavior is not easily observed directly. At this point, it is not clear to what extent genetic variations will prove able to lead to strong behavioral predictions.

4. *Uncovering the Past*

DNA can reveal aspects of the past as well as the present and future. Sometimes it raises in new contexts issues of the privacy of historical figures in matters such as Thomas Jefferson's relationship with his slave, Sally Hemmings. Other historical uses of DNA raise broader questions.

4.1 *Individual Ancestry*

Genetic variations can be used in a straightforward way to establish close biological relationships. Blood groups or isoforms of particular proteins have long been able to provide such information; direct DNA testing can increase accuracy. The ethical implications of such paternity or maternity testing varies according to its use. If done with consent, it seems unproblematic. If done without consent, it might be justified, as in establishing child support obligations, but would require consideration. One problem occurs where genetic testing, done for some other reason, such as attempting a diagnosis of a child's disease, reveals unsought information about 'false paternity' that be unwelcome or even dangerous.

4.2 *Group Histories*

Analysis of patterns of genetic variation among different human populations can reveal how closely related two populations are. This information then becomes evidence about human history and migrations. It is not conclusive evidence, but joins with linguistic, archaeological, anthropological, historical, and other sources to improve understanding of the human past (Cavalli-Sforza et al. 1994). The evidence may not always be welcome. It may contradict a population's own beliefs about its history and origins. In some situations, that information alone might destabilize or disrupt the entire culture. It could also have modern political implications where ancestral origins, and the length of time in possession of particular territories, could affect land disputes. Whether a historical researcher has ethical obligations to avoid doing research that could provoke such consequences is unclear. Similarly unclear is whether there is an ethical obligation to seek the informed

consent of the entire group to such genetic research that might have consequences for the group (Greely 1997). Such consent has not generally been thought necessary for other historical investigations affecting contemporary peoples; some might argue that 'genetic history' is different, either because of the nature of the materials being used—DNA samples from living and sometimes dead members of the population—or the different degree of certainty perceived for such 'scientific' evidence.

5. *Manipulating Genes*

The new genetics may make it possible not just to read person's genes but to change them. That ability would create additional concerns.

5.1 *Somatic Cell Gene Therapy*

Somatic cell gene therapy involves the placement of a human gene into a living person's somatic cells—cells that do not produce the eggs and sperm that in turn produce the next generation. Somatic cell gene therapy would aim to cure a disease only in the patient, not in the patient's descendants. It was initially conceived as introducing a properly functioning copy of a gene into a person who had a genetic disease as a result of inheriting only improperly functioning copies. Different types of somatic cell gene therapy have since been investigated for the treatment of diseases that are not primarily caused by inherited genes, such as AIDS and cancer. Over 100 clinical trials of somatic cell gene therapy have taken place; very few have, thus far, shown any success.

The genetic aspects of somatic cell gene therapy have been largely uncontroversial. In essence, the gene therapy is merely another drug delivery system, a different way to get a normal human protein to the right place in the body. Somatic cell gene therapy therefore stands in the same position as most experimental therapies. Like such therapies, it has prompted concerns that desperate patients are not truly giving informed consent and that the possible benefits of the treatment are exaggerated. Gene therapy may face the latter problem to a greater extent than most experimental treatments because of the exaggerated public view of the power of anything genetic.

5.2 *Germ Line Gene Therapy*

Germ line gene therapy is much more controversial (Nelson 2000). It would introduce 'normal' human genes into the eggs or sperm of parents, or into the fertilized egg or early embryo of the offspring. The goal would be to change the eventual child's genetic inheritance. This could be done in order to avoid a genetic disease or in order to introduce an 'enhancing' genetic variation. There have been no trials of human germ line gene therapy; indeed, there is an informal moratorium in the scientific community on trying such experiments in humans. Both its feasibility and its value are unclear.

New genes have been successfully introduced into the germ lines of other mammals, but with very low efficiency. At the same time, pre-implantation genetic diagnosis allows parents to choose embryos based on their genetic variations, as long as the parents themselves produced the desired variations. If not, donated eggs or sperm would be a much safer and easier way to introduce the desired genes than somatic cell gene therapy. Germ line gene therapy may turn out to be most important as a barrier to somatic cell gene therapy. If germ line gene therapy were banned, researchers using somatic gene therapy might need to make the difficult showing that the transplanted genes could not 'infect' the patient's germ cells and thus constitute inadvertent germ line gene therapy.

5.3 *Chimeras*

The manipulation of genes permits the creation of 'chimeras,' creatures that are genetically a mix of two species, including organisms that could not possibly mate. In agriculture, this mixing of genes from very different organisms has been perceived by many as unethical and potentially dangerous. Unless one takes a strong view of the sanctity of sharp lines between species, it is hard to make a strong ethical argument against such mixes. On the other hand, one might argue that a gene in the setting of new species might develop different, possibly harmful, functions than in its home species.

Chimeras involving human genes provoke other worries. Intentionally moving nonhuman genes into humans could be seen as lessening the recipient's humanity; moving human genes into nonhumans could be seen as investing nonhumans with some aspects of humanity. The former has not been attempted; the latter is routine. Most of the products of the biotechnology industry are made by creatures that are genetically partially human and partially nonhuman. Human genes are transferred into useful host cells such as yeast that then produce large amounts of the corresponding human protein. Of course, an organism with 6,000 yeast genes and one human gene does not seem very human; concerns about this kind of chimera require a strong belief in the essentially human, or sacred, nature of any human DNA sequence. Moving more human genes or moving human genes into more closely related organisms such as chimpanzees could provoke more serious concern about blurring the definition of humanity.

5.4 Artificial Genes or Genomes

It is possible to create 'new' genes, not found in any existing species, that would create new or modified proteins. One branch of the biotechnology industry specializes in this effort, through what is called 'directed evolution.' The technology takes related genes from different species, recombines them or makes mutations in them, and then examines how well the resulting protein functions. Similarly, scientists could construct, piece by piece, a novel genome. Placed inside a cell, it might be able to generate a new living organism. This issue has been discussed to some extent in the context of the now-abandoned 'minimal genome' project (Cho et al 1999). The safety of making new genes, proteins, or is an obvious concern. A deeper concern would be whether humanity should 'play God' by intentionally creating such new entities. Humans have long created new organisms by cross-breeding between species and by selective breeding within species. New genes are created whenever a gene is mutated. The speculative possibilities outlined here are more significant, and less accidental, interventions. It is not clear whether that makes them ethically different.

6. Ownership and Control

6.1 Research

Rules governing the conduct of human genetics research control the research subject's genetic materials and information. Those rules generally require the informed consent of the subject to the research. Research subjects, however, rarely have any control over the subsequent uses of materials or information derived from them. The California Supreme Court decision in *Moore v. Regents of the University of California* (1990) ruled that the plaintiff could not assert a property interest in a cell line derived from cells taken from his body, although it allowed him to try to prove that the physician-researcher had not gotten his informed consent. Although the *Moore* case is the only decision on this point, current practice generally follows it and gives research subjects no control over who can use their genetic materials and information, for what purposes, or for how long. Current practices also discourage any sharing of financial benefits of the research with the subjects on the theory that the hope of such benefits could be an 'undue inducement' to them to take part in the research.

A new kind of human genetics research raises other issues about the control of research subject information and materials. Associational research seeks to find weaker connections between genetic variations and disease by searching for correlations between the genotypic and phenotypic data of large groups of people. This kind of research requires the creation of a resource made up of health records and DNA samples of hundreds of thousands of people. These resources would be too expensive to create to study just one illness. Instead, the resources would be available to test a wide range of hypotheses. This kind of 'genotype/phenotype resource' is most advanced in Iceland, where the government has given a private firm, deCODE Genetics, the license to create a health records database with clinical medical information on all 275,000 residents of the country.

In the case of Iceland, the legislation also provides that consent is presumed; unless an Icelander opts out of the database, his medical records will be included. The Icelandic plan has raised numerous ethical concerns, including the use of presumed consent, the absence of discussion of the risks and benefits of specific research done on the resource, the degree that privacy would be protected, the access (if any) of other researchers to the company's data, the propriety of having a for profit company control this information, and the financial fairness of the agreements between Iceland and deCODE (Greely 2000). It is not clear on what terms this kind of research will ultimately proceed, in Iceland or elsewhere.

6.2 Patents

When the US Supreme Court, in *Diamond v. Chakrabarty*, approved a patent on a genetically altered bacterium, it started a continuing debate over the patenting and genetics (*Diamond v. Chakrabarty* 1980). To many patent lawyers, DNA is just another organic molecule and patenting of forms or portions of it falls within long-established principles concerning the composition of matter patents on molecules not found in pure or refined form in nature (Eisenberg 1990). Others object to some, or all, genetic patents on a variety of grounds. Those objections can be put into two categories: fundamental and technical.

Some people have fundamental objections to patents on genes, on human genes, or on genetically modified life forms. One set of objections, focusing on genes, contends that they are 'discoveries,' not 'inventions' and so should not be patented. Others point out that genes were made by God, nature, or evolution—in any event, something other than the 'inventors' filing the patent claims—and should not be claimable as someone's intellectual property. Another objection is that human genes are the common heritage of mankind and should be held in common. Still others oppose gene patents as part of broader opposition to biotechnology. Proponents of patents put forward answers to each of these objections; throughout the industrialized world, the proponents have won. Patents on genes, human genes, and to a somewhat

lesser extent genetically modified life forms have been commonly allowed in the United States, Europe, and Japan. It is worth noting, though, that patents usually expire 20 years from the date of application. The first wave of gene patents will soon expire; presumably all patents on human genes will have expired by 2021, 20 years from the publication of 'the human genome.'

Technical arguments continue to rage within the field. The biggest fight has largely been over the patent requirement of 'utility': how much, or little, knowledge about the uses of a genetic sequence is sufficient to convey what kind of patent rights? This controversy became open in the early 1990s when the National Institutes of Health filed patent applications on expressed sequence tags ('ESTs'), small sections of DNA from inside genes of unknown sequence and function. The fight has continued over patent applications on single nucleotide polymorphisms, one base pair differences in the genome. It has also included how much information about a gene and its function is necessary in order to grant a patent.

6.3 Privacy

Protecting the privacy of genetic information is one way to control its use (Annas et al. 1995, Rothstein 1997). But such 'genetic privacy' legislation raises its own set of problems.

The first problem is definitional—what is the 'genetic information'? Information about a person's genetic variations can come from DNA tests, from other biochemical tests, from a naked eye examination, or from family history. A narrow definition, focusing on the results of 'DNA tests,' would miss such powerful information as the 50 percent chance that the child of a Huntington's disease patient carries the Huntington's allele. A broad definition, covering everything from which an inference can be drawn about a person's genotype, would cover almost all medical information. A very high cholesterol level is strong evidence that a person carries two alleles for familial hypercholesteremia. An average or low level is strong evidence that he or she does not.

If almost all medical information leads to inferences about genotypes, it may not make sense to try to protect 'genetic information' separately from medical information. And it may not be possible to 'protect' a person's medical information from the many institutions that have legitimate uses for it, including physicians; hospitals; and those employers, insurers, and governments who pay health care bills. At the same time, the passage of legislation expressly to protect 'genetic' information sends the public the inaccurate message that genetic information is much more powerful and important than other medical information. Legislation to protect genetic privacy has been introduced in the United States Congress but has not yet been passed. Several states have passed such legislation, but it is too early to tell how effective that legislation will prove.

7. Cultural Consequences

The most far-reaching, but hardest to predict, ethical implications of the new genetics may lie in its effects on society's beliefs. Three areas stand out.

First, the new genetics demonstrates graphically that all life is related. Over a third of the genes found in the single-celled brewer's yeast have recognizable relatives in the human genome. Some stretches of DNA appear in close to identical form in humans, mice, and fruit flies. Indeed, there appear to be few, if any, 'human genes'—just human variants of genes shared by all mammals, vertebrates, multicellular creatures, or life forms generally. Darwin's thesis that all earthly life is related by descent from a common ancestor can be seen in these similarities in DNA. It is unclear what cultural significance this will have. It should not promote vegetarianism, because the new genetics shows that carrots and corn, like sheep and cattle, are our relatives. It might, however, promote greater respect for nonhuman life.

Second, the new genetics shows that all humans are closely related. Our DNA differs, on average, at one spot in 1,000. In the regions of the genome that code for protein, the differences are one base pair in 10,000. Humans from opposite ends of the earth are far more similar to each other genetically than chimpanzees from the same band. Genetic theories were used to provide support for a 'scientific' racism in the first part of the twentieth century. The new genetics should provide evidence against such racism.

Third, genetics may shift the balance in the cultural debate between nature and nurture as the source of human characteristics (Degler 1991). Dr. James Watson, co-discoverer of the structure of DNA, was famously quoted as saying 'We used to think that our fate was in the stars. Now we know, in large measure, our fate is in our genes.' DNA may appear to be proof that individuals are powerfully shaped by inherited forces. In fact, the new genetics paints a more complicated picture. Genes play a role in the development of many traits or diseases, but the environment or luck may also be essential. The general population seems to hold a much stronger belief in the power of genes. For that reason, the new genetics could end up promoting a more closed and fatalistic view of human life and abilities than either current society holds or than science would support (Nelkin and Lindée 1995). That reaction may prove to be the most significant ethical challenge of the new genetics.

See also: Aging and Health in Old Age; Aging, Theories of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Differential Aging; Lifespan Development, Theory of; Old Age

and Centenarians; Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

Bibliography

Annas G J, Glantz L H, Roche P A 1995 *The Genetic Privacy Act and Commentary*. Boston University School of Health, Boston, MA

Bobinski J A 2000 Genetic Information, Legal, ERISA Preemption, and HIPAA Protection. In: Murray T H, Mehlman M J (eds.) *Encyclopedia of Ethical, Legal, and Policy Issues in Biotechnology*. John Wiley & Sons, New York

Carson R A, Rothstein M A (eds.) 1999 *Behavioral Genetics: The Clash of Culture and Biology*. Johns Hopkins University Press, Baltimore, MD

Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

Cho M K, Magnus D, Caplan A L, McGee D 1999 Ethical considerations in synthesizing a minimal genome. *Science* **286**: 2087–90

Cook-Deegan R M 1994 *The Gene Wars: Science, Politics, and the Human Genome*. Norton, New York

Degler C N 1991 *In Search of Human Nature: The Decline and Revival of Darwinism in American Social Thought*. Oxford University Press, New York

Diamond v. Chakrabarty 1 1980 444 US 397

Eisenberg R S 1990 Patenting the human genome. *Emory Law Journal* **39**: 721–45

Greely H T 1997 The control of genetic research involving the 'groups between.' *Houston Law Review* 1397–1430

Greely H T 2000 Iceland's plan for genomics research: Facts and implications. *Jurimetrics Journal* **40**: 153–91

Greely H T 2001 Genotype discrimination. *Pennsylvania Law Review* **149**: 1483–1505

Institute of Medicine, US National Academy of Sciences 1993 *Assessing Genetic Risks: Implications for Health and Social Policy*. National Academy Press, Washington, DC

Moore v. Regents of the University of California 1990 51 Cal.3d 120

National Bioethics Advisory Commission 1997 *Cloning Human Beings*. NBAC, Rockville, MD

National Research Council 1992 *DNA Technology in Forensic Science*. National Academy Press, Washington, DC

National Research Council 1996 *DNA Technology in Forensic Science: An Update*. National Academy Press, Washington, DC

Nelkin D, Lindee M S 1995 *The DNA Mystique: The Gene as a Cultural Icon*. WH Freeman, New York

Nelson R 2000 Gene therapy, ethics, germ cell gene transfer. In: Murray T H, Mehlman M J (eds.) *Encyclopedia of Ethical, Legal, and Policy Issues in Biotechnology*. John Wiley & Sons, New York

Parens E 1998 *Enhancing Human Traits: Ethical and Social Implications*. Georgetown University Press, Washington, DC

President's Commission for the Study of Ethical Problems in Medicine and Biomedical Research 1983 *Screening and Counseling for Genetic Conditions: A Report on the Ethical, Social, and Legal Implications of Genetic Screening, Counseling and Education Programs*. The Commission, Washington, DC

Rothstein M A (ed.) 1997 *Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era*. Yale University Press, New Haven, CT

H. T. Greely

Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives

Committees charged with the responsibility for assuring that research involving humans as subjects are called research ethics committees (RECs). Although these committees were designed originally to review biomedical research, their domain has been extended to cover research in the social and behavioral sciences as well.

The Declaration of Helsinki establishes as the international standard for biomedical research involving human subjects the requirement that each research '… protocol should be submitted for consideration, comment, guidance, and where appropriate, approval to a specially appointed ethical review committee …' (World Medical Assoication 2000, Article 13). In most of the world this committee is called the Research Ethics Committee (REC). In the United States, federal law assigns to the committee the name Institutional Review Board (IRB) and the authority and responsibility for approving or disapproving proposals to conduct research involving human subjects (Robertson 1979b). In Canada, the Research Ethics Board (REB) has similar authority to approve or disapprove research proposals, not merely to offer 'consideration, comment and guidance' (Medical Research Council of Canada 1987, p. 48).

1. History

The Nuremberg Code (Superintendent of Documents 1950) and the original Declaration of Helsinki (World Medical Association 1964) made no mention of committee review; these documents placed on the investigator all responsibility for safeguarding the rights and welfare of research subjects. The first mention of committee review in an international document was in the Tokyo revision of the Declaration of Helsinki (World Medical Association 1975).

In the United States, the first federal document requiring committee review was issued on November 17, 1953. Entitled 'Group Consideration for Clinical Research Procedures Deviating from Accepted Medical Practice or Involving Unusual Hazard,' its guidelines applied only to research conducted at the newly opened Clinical Center at the National Institutes of Health (NIH) (Lipsett et al. 1979). We know very little about committee review in other institutions in the 1950s other than that it existed at least in some medical schools. In 1961 and again in 1962, questionnaires were sent to American university departments of medicine. Approximately one-third of those responding reported that they had committees and one-quarter either had or were developing procedural documents (Curran 1970).

On February 8, 1966, the Surgeon General of the United States Public Health Service (USPHS) issued the first federal policy statement requiring research institutions to establish the committees which subsequently became known as RECs (Curran 1970). This policy required recipients of USPHS grants in support of research involving human subjects to specify that 'the grantee institution will provide prior review of the judgment of the principal investigator or program director by a committee of his institutional associates. This review shall assure an independent determination: (a) of the rights and welfare of the … individuals involved, (b) of the appropriateness of the methods used to secure informed consent, and (c) of the risks and potential medical benefits of the investigation.'

The evolution of the federal government's charges to the committee and of its recognition of the need for diversity of its membership was reflected in several revisions of its policy between 1966 and 1969 (Veatch 1975, Levine 1988); these will be discussed further shortly.

2. *Purpose*

The purpose of the REC is to assure that research involving human subjects is designed to conform to relevant ethical and legal standards. Historically, the REC's primary focus was on safeguarding the rights and welfare of individual research subjects, concentrating on the plans for informed consent and the assessment of risks and anticipated benefits. In 1978, the US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (National Commission) added a requirement that the REC assure equitableness in the selection of research subjects (Levine 1988, pp. 67ff). The National Commission was concerned primarily with protecting vulnerable subjects from bearing a disproportionately large share of the burdens of research. Subsequently, as participation in some types of research became perceived as a benefit, RECs also assumed responsibility for assuring disadvantaged persons equitable access to such benefits (C. Levine 1988).

A source of continuing controversy is whether the REC has an obligation to approve or disapprove the scientific design of research protocols (Levine 1988, pp. 20–5). Those who argue that they do or should have such an obligation point out that the leading ethical codes establish an ethical requirement for good scientific design. Moreover, they argue, the REC's obligation to determine that risks to subjects are reasonable in relation to anticipated benefits necessarily relies on a prior determination that the scientific design is adequate, for if it is not, there will be no benefits and any risk must be considered unreasonable.

Opponents to assigning such an obligation to the REC, while conceding these two points, argue that the REC is not designed to make expert judgments about the adequacy of scientific design. RECs are generally competent to appraise the value of the science—what the Nuremberg Code calls 'the humanitarian importance of the problem to be solved'—but not the validity of the methods or the results (Freedman 1987, Veatch 1975). In general, responsibility for assessment of scientific validity is and ought to be delegated to committees designed to have such competence—e.g., scientific review committees either within the institution or at funding agencies such as the NIH (Levine 1988, pp. 21–3).

3. *Membership*

The Surgeon General's 1966 memo called for prior review by 'a committee of [the investigator's] associates,' what was commonly called 'peer review.' As of 1968, 73 percent of committees were limited in membership to immediate peer groups; viz., scientists and physicians (Curran 1970, p. 443).

On May 1, 1969, USPHS Guidelines were revised to indicate that a committee constituted exclusively of biomedical scientists would be inadequate to perform the functions now expected of such a committee: 'the membership should possess … competencies necessary in the judgment as to the acceptability of the research in terms of institutional regulations, relevant law, standards of professional practice, and community acceptance.'

Regulations of the US Department of Health and Human Services (DHHS), first promulgated in 1974 and since revised several times, maintain the spirit of the 1969 policy and in addition require gender diversity; at least one nonscientist (e.g., lawyer, ethicist, member of the clergy) and at least one member who is not otherwise affiliated with the institution (commonly and incorrectly called a 'community representative'). Persons having conflicting interests are to be excluded; this concern is also reflected in the Declaration of Helsinki's requirement of a 'committee independent of the investigator and sponsor.'

According to Robert Veatch, the REC is an intermediate case between two models of the review committee (1975): The 'interdisciplinary professional review model,' made up of diverse professionals such as doctors, lawyers, scientists, and clergy, brings professional expertise to the review process, while the 'jury model … reflects the common sense of the reasonable person.' In the jury model 'expertise relevant to the case at hand is not only not necessary, it often disqualifies one from serving on the jury.' Veatch concedes that in order to perform all of its functions the REC requires both professional and jury skills. However, he argues that the presence of professionals makes it more difficult for the REC to be responsive to the informational needs of the reasonable person or to be adept at anticipating community acceptance.

John Robertson (1979a) would correct the 'structural bias' of professional domination by introducing a 'subject surrogate,' an expert advocate for the subjects' interests. DHHS regulations require that a REC which 'regularly reviews research that involves a vulnerable category of subjects, … shall include one or more individuals who are primarily concerned with the welfare of these subjects.' For research involving prisoners, for example, regulations require that at least one member of the REC be either a prisoner or a prisoner representative. There is unresolved controversy over whether persons with AIDS should be appointed to membership on all RECs which review research in the field of HIV infection (C Levine et al. 1991).

4. *Locale*

In the United States the first RECs were established in the institutions in which research was conducted. The 1966 Surgeon General's policy statement required a committee of 'institutional associates.' In 1971 the FDA promulgated regulations which required committee review only when regulated research was conducted in institutions; hence their name, Institutional Review Committee (IRC). Regulations proposed in 1973 by the Department of Health, Education and Welfare, forerunner of DHHS, also reflected a local setting in their term, Organizational Review Board (ORB). In 1974 the National Research Act established a statutory requirement for review by a committee to which it assigned the name, Institutional Review Board (IRB), a compromise between the two names then extant.

RECs are required to comply with federal regulations when reviewing activities involving FDA-regulated 'test articles' such as investigational drugs and devices, and when reviewing research supported by federal funds (Robertson 1979b). Moreover, all institutions that receive federal research grants and contracts are required to file 'statements of assurance' of compliance with federal regulations. In these assurances virtually all institutions voluntarily promise to apply the principles of federal regulations to all research they conduct regardless of the source of funding.

These points notwithstanding, each REC has a decidedly local character. Most have local names such as Human Investigation Committee, or Committee for the Protection of Human Subjects. Each is appointed by its own institution and each lends its own interpretation to the requirements of federal regulations. For example, at one university medical students are forbidden to serve as research subjects while at another, involvement of medical students as research subjects is sometimes required as a condition of approval (Levine 1988, pp. 80–2).

The National Commission recommended that RECs should be 'located in institutions where research … is conducted. Compared to the possible alternatives of a regional or national review … local committees have the advantage of greater familiarity with the actual conditions' (1978, pp. 1–2). The National Commission envisioned the local REC as an ally of the investigator in safeguarding the rights and welfare of research subjects as well as a contributor to the education of both the research community and the public.

FDA's change in regulations in 1981 to require REC review of all regulated research regardless of where it was done created a problem for the many physicians who were conducting investigations in their private offices, many of whom had no ready access to RECs. In response, private corporations developed 'noninstitutional review boards' (NRBs) (Herman 1989). Although there are theoretical reasons to question the validity of NRB review, they appear to be performing satisfactorily (Levine and Lasagna 2000).

In 1986, FDA began to waive the requirement for local REC review for some protocols designed to evaluate, or to make available for therapeutic purposes, investigational new drugs, particularly those intended for the treatment of HIV infection. In such cases RECs were offered the option of accepting review by a national committee as fulfilling the regulatory requirement for REC review. Such practices have caused some commentators to question the strength of the government's commitment to the principle of local review.

Internationally, there is much less commitment to the importance of local review. The International Ethical Guidelines for Biomedical Research Involving Human Subjects, promulgated by the Council for International Organizations of Medical Sciences, require REC approval for all research involving human subjects and recognize the validity of review at a regional or, 'in a highly centralized administration,' a national level (1992). In many European countries, RECs are regional (McNeill 1989).

Several commentators have expressed concern that in the United States the local institution has too much power in the field of protection of human research subjects. Robertson, for example, alerts us to 'the danger … that research institutions will use [RECs] to protect themselves and researchers rather than subjects' (1979a); others point to the close associations between RECs and risk-management offices in many institutions as evidence that RECs are being used in this manner.

5. *Criticisms*

Before 1962, 'a general skepticism toward the development of ethical guidelines, codes, or sets of procedures concerning the conduct of research' pre-

vailed in the medical research community (Curran 1970). In the 1970s several biomedical scientists were harshly critical of the REC system claiming that it tended to stifle creativity and impede progress (Levine 1988, pp. 345–8); survey research, however, showed that only 25 percent of biomedical researchers agreed with the statement that: 'The review … is an unwarranted intrusion on the investigator's autonomy—at least to some extent' (National Commission 1978, p. 75). Behavioral and social scientists were considerably less accepting of review, claiming that their research activities were much less likely than those of the biomedical scientist to harm subjects. Some argued that since all they did is talk with subjects, review was an unconstitutional constraint on their freedom of speech (Levine 1988, p. 359).

According to Peter Williams, RECs do an inadequate job of assuring that risks will be reasonable in relation to anticipated benefits (1984). This is inevitable for three reasons: (a) Federal regulations on this standard are written in vague language in contrast to the more clear direction provided for protecting subjects' rights. Moreover, since the regulations permit consideration of the long-range effects of applying knowledge as benefits but not as risks, they create a bias in favor of approval. (b) The membership of the committee, dominated as it is by professionals, is likely to place a higher value than laypersons would on the benefit of developing new knowledge. (c) Groups confronted with choices involving risks may be either more or less cautious or 'risk aversive' than the average of individuals within the group; this is known as the 'risky shift' or 'group polarization phenomenon.' Williams (1984) and Veatch (1975) believe that in the context of RECs, the groups are likely to be more tolerant of higher levels of risk than they would be as individuals.

Several commentators have proposed that RECs could enhance their effectiveness by sending members to the sites of the actual conduct of research to verify compliance with protocol requirements (Robertson 1979a) or to supervise consent discussions (Robertson 1982). Others respond that while such activities should be done when there are reasons to suspect problems in specific protocols, routine monitoring activities might be detrimental to the successful functioning of the committee by eroding its support within the institution (Levine 1988, pp. 341ff).

6. *Evaluation*

Critics of the REC system claim that there is little or no objective evidence that REC review prevents the conduct of inadequate research. For example, a national survey of RECs revealed that the rate of rejection of protocols is less than one in one thousand (National Commission 1978, pp. 1–216). Supporters of the system respond that the actual rejection rate is much higher if one includes protocols withdrawn because investigators refuse to modify them as required by RECs. Moreover, rejection rates may be a poor indicator of the REC's quality; protocols may be improved in anticipation of the REC's requirements and investigators, fearing rejection, may decide not to submit proposals they think might be rejected by the committee.

It is very difficult to evaluate the REC's performance objectively; satisfactory subjective evaluations can be made only by experienced REC members and administrators (Levine 1988, pp. 358–9). Jerry Mashaw concludes in his excellent theoretical analysis of RECs that: 'If [the REC] is to do its core job well, we must live with its inevitable incompetence at other tasks. Moreover, we must also live with the rather vague regulatory standards and with the continuing inability of the federal funding agencies to know for sure whether [RECs] are functioning effectively. If we would have wise judges and paternalistic [skilled in protecting subjects' rights and welfare interests] professionals, we can neither specifically direct nor objectively evaluate their behavior' (1982, p. 22).

See also: Civil Liberties and Human Rights; Ethics Committees in Science: European Perspectives; Ethics for Biomedical Research Involving Humans: International Codes; Free Will and Action; Objectivity of Research: Ethical Aspects; Privacy of Individuals in Social Research: Confidentiality; Psychiatry: Informed Consent and Care; Research Conduct: Ethical Codes

Bibliography

Annas G 1991 Ethics committees: From ethical comfort to ethical cover. *Hastings Center Report* **21**(3): 18–21

Council for International Organizations of Medical Sciences 1992 *International Ethical Guidelines for Biomedical Research Involving Human Subjects*. Council for International Organizations of Medical Sciences, Geneva

Curran W J 1970 Government regulation of the use of human subjects in medical research: The approaches of two federal agencies. In: Freund P A (ed.) *Experimentation with Human Subjects*. George Braziller, New York. pp. 402–54

Freedman B 1987 Scientific value and validity as ethical requirements for research: A proposed explication. *IRB: A Review of Human Subjects Research* **9**(6): 7–10

Herman S S 1989 A noninstitutional review board comes of age. *IRB: A Review of Human Subjects Research* **11**(2): 1–6

Levine C 1988 Has AIDS changed the ethics of human subjects research? *Law, Medicine & Health Care* **16**: 167–73

Levine C, Dubler N N, Levine R J 1991 Building a new consensus: Ethical principles and policies for clinical research on HIV/AIDS. *IRB: A Review of Human Subjects Research* **13**(1 and 2): 1–17

Levine R J 1988 *Ethics and Regulation of Clinical Research*. 2nd edn. Yale University Press, New Haven, CT

Levine R J, Lasagna L 2000 Demystifying central review boards: Current options and future directions. *IRB: A Review of Human Subjects Research* **22**(6): 1–4

Lipsett M B, Fletcher J C, Secundy M 1979 Research review at NIH. *Hastings Center Report* **9**(1): 18–21

Mashaw J L 1982 Thinking about institutional review boards. *Whistleblowing in Biomedical Research: Policies and Procedures for Responding to Reports of Misconduct*. The President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research, Washington, DC, pp. 3–22

McNeill P M 1989 Research ethics committees in Australia, Europe, and North America *IRB: A Review of Human Subjects Research* **11**(3): 4–7

Medical Research Council of Canada 1987 *Guidelines on Research Involving Human Subjects*. Medical Research Council of Canada, Ottawa, Canada

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1978 *Institutional Review Boards: Report and Recommendations*. DHEW, Washington, DC, Publication No (OS) 78–0008

Robertson J A 1979a Ten ways to improve IRBs. *Hastings Center Report* **9**(1): 29–33

Robertson J A 1979b The law of Institutional Review Boards *UCLA Law Review* **26**: 484–549

Robertson J A 1982 Taking consent seriously: IRB intervention in the consent process *IRB: A Review of Human Subjects Research* **4**(5): 1–5

Superintendent of Documents, US Government Printing Office 1950 Trials of war criminals before the Nuremberg military tribunals under Control Council Law 10. *Military Tribunal 1, Case 1, United States v. Karl Brandt et al., October 1946–April 1949* **1**: 1–1004; **2**: 1–352, US Government Printing Office, Washington, DC

Veatch R M 1975 Human experimentation committees: Professional or representative? *Hastings Center Report* **5**(5): 31–40

Williams P C 1984 Success in spite of failure: Why IRBs falter in reviewing risks and benefits. *IRB: A Review of Human Subjects Research* **6**(3): 1–4

World Medical Association 1964 Declaration of Helsinki: Ethical principles for medical research involving human subjects. *18th WMA General Assembly, Helsinki, Finland, June 1964*

World Medical Association 1975 Declaration of Helsinki: Ethical principles for medical research involving human subjects (amended). *29th WMA General Assembly, Tokyo, Japan, October 1975*

World Medical Association 1983 Declaration of Helsinki: Ethical principles for medical research involving human subjects (amended). *35th WMA General Assembly, Venice, Italy, October 1983*

World Medical Association 1989 Declaration of Helsinki: Ethical principles for medical research involving human subjects (amended). *41st WMA General Assembly, Hong Kong, September 1989*

World Medical Association 1996 Declaration of Helsinki: Ethical principles for medical research involving human subjects (amended). *48th WMA General Assembly, Somerset West, Republic of South Africa, October 1996*

World Medical Association 2000 Declaration of Helsinki: Ethical principles for medical research involving human subjects (amended). *52nd WMA General Assembly, Edinburgh, Scotland, October 2000* http://www.wma.net/e/policy/17c.pdf

R. J. Levine

Ethics and Psychiatry

Ethics concerns what is right and what is wrong. It is often used synonymously with morality, but sometimes a distinction is made between ethics, the systematic approach to understanding right and wrong, and morality, the forces which govern right conduct. Ethics is central to human experience, and plays a vital role in the profession of psychiatry. In a time of cultural change, the appeal to ethics can help to resolve conflicts of both civic life and self-understanding. Often ethical principles are used in an attempt to find a common ground or a 'higher ground' for resolving disputes or to find a basis for acting in a dilemma.

Ethical codes and traditions are important in defining the norms of professional conduct and even what it means to be a profession. Insofar as ethics deals with intellectual attempts to understand right and wrong (reason) as well as the affective impact of struggling with moral issues (emotion), psychiatric thinking offers an important perspective on ethics. This article deals with both ethics in psychiatry and psychiatry's contributions to ethics.

1. *Types of Ethical Theories*

There is in ethics a constant tension between rule-based (deontological) approaches and approaches which appeal to principle (teleological). Deontological ethical theories maintain that there are rules of action with moral validity independent of the consequences that require adherence, such as keeping a promise, maintaining justice, or following a commandment of God or the state. Teleological ethical theories are based on the principle of utility (utilitarianism) or beneficence, which proposes that the ultimate moral standard is the general happiness of all people concerned, or the greatest net balance of good over evil ('the greatest good for the greatest number'). Situation ethics represents another form of ethical reasoning, called the consequentialist approach, in which an action is judged to be morally right or wrong by assessing the consequences of the action. Much of clinical medicine is consequentialist in that the goodness of actions is judged by their outcomes.

2. *Ethics of Virtue and Character*

Both teleological and deontological approaches to ethics have one serious shortcoming. They depend on the activities of the rational mind of a motivated moral agent to apprehend right and wrong. Moral philosophers and psychiatrists alike appreciate that not everyone is motivated to be moral; that even good people sometimes do bad things; and that people are capable of rationalizing self-interest (Hauerwas 1977, 1981, Pincoffs 1971, 1980).

A delineation of obligations, however essential, is not the whole picture of ethics. Virtue, character and integrity are every bit as important in understanding right conduct. For the professions especially, integrity in concrete situations is the hallmark of virtuous life.

3. *A Brief History of Medical Ethics*

The ethical traditions of the medical profession originate in the fourth century BC with the Oath of Hippocrates. This oath, sworn by healer–craftsmen, has come to symbolize the ethical ideals of the medical profession. In the nineteenth century, first the British Medical Association, and subsequently the American Medical Association (AMA) and professional organizations in other countries began to adopt formal codes of ethics modeled on the principles of the Hippocratic Oath.

By the mid-1960s, the values underlying the doctor–patient relationship began to change. The most notable factor was the emergence of new technologies, which began to change the way we thought about medicine and even what it meant to be human: genetic engineering, organ transplantation, advanced life support, safe abortion and birth control, and medical and surgical treatments based on clinical research. No longer was medical care primarily palliative. No longer could it be assumed that the patient would necessarily want what the doctor might offer (Dyer 1997a, 1997b).

Also in this era, several social movements began to change the way society thought about assumed relationships, with an increased emphasis on the right of the autonomous person to self-determination. Informed consent and autonomy came to be the keys to medical ethics and medical decision-making. Medical ethics in this era became bioethics as it came to be realized that the decisions in medicine were decisions in which society had a stake.

By the 1970s, economics increasingly became a part of thinking about ethics. Allocation of resources became a concern, and the question was often asked, 'Could health costs be contained without rationing care?' The principle of justice eventually superseded beneficence and autonomy in ethical discourse, with an emphasis on fair distribution of medical resources.

3.1 *The Ethics of Advertising*

Most physicians are not inclined to seek patients though advertising. Healthcare organizations, however, may make their appeal through advertising, especially in a market-driven healthcare economy. The 1957 version of the AMA *Principles* prohibited 'solicitation' of patients, by which was meant 'obtaining patients by deception', or making false claims (AMA 1957). After the Supreme Court decision Federal Trade Commission (FTC) vs. AMA (Greenhouse 1982), professional organizations such as the AMA can no longer formally constrain physicians from advertising.

Advertising has two basic goals: (a) dissemination of information and (b) product differentiation. For physicians, providing information about services offered and fees charged is consistent with patient benefit, but attempts to differentiate the product are suspect. Licensing, accrediting and credentialing organizations assume the role of maintaining standards. Beyond the activities of individual practitioners, the larger ethical question for advertising is truthfulness. Notable is the advertising of unhealthy products such as tobacco products and alcohol. Much advertising is antithetical to professional goals of patient benefit (Dyer 1995, 1997a, 1997b).

4. *The Place of Ethics in the Definition of a Profession*

What differentiates a profession from a trade? What makes medicine a profession? The emphasis on the commerce of technology in medicine in recent years may obscure the centrality of an ethical attitude toward the patient, which has traditionally been considered to be the defining feature of a professional life.

A profession may be defined by (a) its knowledge, technology and expertise, or (b) its ethics or values. The habits of modern thought might lead us to believe that this is an either–or choice. Clearly for psychiatry, as for most of medicine, allopathic technologies such as drugs and procedures have become important and perhaps even central in some people's minds. The recent primary care thrust has reopened consideration of a more holistic approach to healing, including talking. Ethics is more fundamental to professional definition than technology. Technology is useful as long as it serves ethical ends but not as an end in itself. Knowledge, technology and expertise are not commodities to be bartered in the marketplace, but skills which may be used to ethical ends under the purview of professional values.

5. *The Hippocratic Oath and the Hippocratic Tradition*

The Hippocratic Oath is a remarkable document, not so much because it answers the ethical questions posed by modern medicine, but because it frames those questions. It articulates principles, notably the principle of patient benefit, which is often helpful to physicians in sorting out where their allegiances lie. Taken together with its corollary, the principle of nonmaleficence (*Primum non nocere*, first do no harm),

the principle of beneficence puts the patient at the center of ethical decision-making. The Hippocratic Oath provides an ethical perspective that calls into questions many of the assumptions of modern culture and transcends the pressures of political expediency.

5.1 The Contemporary Relevance of the Hippocratic Oath

A reading of the Hippocratic Oath reveals many phrases which responsible physicians might not want to follow literally. Medicine was delineated as a higher calling, a covenant with the gods. The Oath provided a relationship with a like-minded family of individuals who shared certain values and self-understanding. It defined the physician's life in a beneficial relationship to those served. Its paternalism was questioned when society emphasized individualistic autonomy, but the Oath's emphasis on justice foreshadowed a larger concern for community interests.

The Hippocratic tradition has at various times and places served as an ethical reference external to immediate expediency, by which physicians may gauge their behavior and to which they can appeal (or by which they may be judged by their colleagues). For example, the World Psychiatric Association censured the practices of psychiatrists in the former Soviet Union and other eastern European countries for the detention of political dissidents. Such mislabeling with psychiatric diagnoses is called the political abuse of psychiatry, for the profession is used for ends other than patient benefit (Bloch 1991, Bloch and Reddaway 1977). Such extreme abuses make a case, which is not always appreciated, for professions being autonomous organizations self-regulated by a code of ethics rather than serving the interests of society which may be politically conflicted. The Hippocratic Oath is an ever-present reminder that acting in the interest of someone other than the patient creates a conflict, a conflict which is often experienced internally by the physician.

5.2 Honor in Medicine

The Hippocratic Oath speaks of honor, holiness and purity. The good name and respect of both the practitioner and the profession are to be earned through consistent application of the ethical principles, and the code provides guidance. Good conduct becomes 'characteristic' of the professional, and a code of conduct organizes one's thinking about good conduct both symbolically and in very practical matters.

5.3 The Fiduciary Principle

The doctor–patient relationship is spoken of sacredly as a fiduciary relationship, a relationship based on trust. That trust must be earned, at the least by the hard work required to acquire the skills necessary to apply the art. But it must also be earned by the consistent application of attentive response to the patient's needs, what we call 'responsibility'. Unlike law or business, where the trustee may act for the client, in medicine the physician must act with the patient and with the patient's consent, implying a partnership. That action may be paternalistic, but if that concern verges on control, the modern patient may well lose trust and confidence.

5.4 Hippocratic Oath for Psychiatrists

Maurice Levine (1972) articulated a Hippocratic Oath for psychiatrists with some unique insights about the values of this profession. His code articulates the importance of self-knowledge or at least self-awareness and constant self-scrutiny. He stresses the need to recognize in ourselves the feelings that patients can stir up, particularly seductive patients, and the importance of getting consultation (or supervision) not just in training but whenever it might be necessary. This is a useful reminder at times when many treatments are briefer and more biological, because many of the complaints of unethical conduct received by the American Psychiatric Association (APA) come from patients whose doctors did not sufficiently deal with their own feelings and acted out in treatment situations. Self-reflection is an important ethical tradition, not only in psychiatric and psychoanalytic circles, but throughout Western culture dating back to the Socratic admonition to 'Know thyself.'

6. Contemporary Codes of Ethics (AMA and APA)

When the AMA was founded in 1847, it adopted a code based largely on the Hippocratic principles of beneficence and honor. This code (*The Principles of Medical Ethics*) has been revised every few decades (most recently in 1957 and 1980). When the APA adopted its first code in 1973, it decided to use the AMA Principles *With Annotations Especially Applicable to Psychiatry*. The principles should be read teleologically, not deontologically as a list of rules. The requirement of honorable behavior of the physician goes to the physician's character and virtue and cannot be reduced to a list of rules. The physician must understand the principles in terms of a higher calling and act in accord with the dictates of conscience.

The APA (1995) may impose four possible sanctions for misconduct by a psychiatrist: admonishment (informal warning); reprimand (formal censure); suspension (not to exceed five years); and expulsion. In this sense, this code is minimalist but essential, serving as a tool for professional discipline. A psychiatrist should know what is specified in the code because he

or she might be subject to sanctions. The modern physician should also understand the principles that underlie the code in order to be able to act in situations which have not been explicitly spelled out. The following discussions illustrate the constant tension between rule and principle and the importance of the physician striving for virtuous action.

6.1 Sexual Misconduct and Boundary Violations

Sexual contact with a patient is unethical. This is one of the least ambiguous sections of the ethical code. It is a tradition that goes back to the Oath of Hippocrates, which speaks of such conduct as 'mischief'. It is important for psychiatrists because the intimacy of the treatment activates strong feelings and fantasies in the doctor–patient relationship, the discussion of which may be essential to healing. In psychiatry especially, the nature of such feelings the patient may have for the doctor (transferences, i.e., derived from significant relationships in the past and activated in the treatment) and feelings the doctor may have for the patient (countertransferences) receive close scrutiny. They are no less important for other physicians, other therapists, or other professionals. It is sometimes argued that the reason sexual contact is proscribed is because of the power differential, an argument which applies equally to employers and employees, supervisors and supervisees, teachers and students. Most basically, the importance of trust in the therapeutic relationship requires forbearance. As a rule, sexual contact with a patient is proscribed.

Those more legalistically inclined might wonder if erotic feelings might be acted on if the professional relationship were terminated. Transferences endure over time, so the interests of the patient–client could never be served by crossing this boundary. The most recent version of the APA *Annotations Applicable to Psychiatry* spells this out: 'Sexual activity with a current or former patient is unethical.' Other examples of boundary violations, which rest on similar considerations, are business relations with a patient, using the professional relationship for other contacts, or profiting from information gained from the therapeutic relationship.

6.2 The Boundary between Ethics and Law

Ethics and the law share a similar concern for right and wrong. But the ethical requirements of professionals do not coincide identically with the requirements of all citizens. The Oath of Hippocrates held physicians to a standard that was not required of all citizens. The AMA *Principles* and APA *Annotations* recognize this tension and expect physicians to 'respect the law' and 'seek changes in those requirements which are contrary to the best interests of the patient.' Civil disobedience against unjust laws is a respected ethical tradition. A physician is a citizen whose professional role requires adherence to principles such as patient-benefit above all.

6.3 Confidentiality

Confidentiality, understood as secrecy, in the doctor–patient relationship is one of the most fundamental principles of professional ethics. Confidentiality has its roots in the confidence or trust that is placed in the physician and on which the treatment depends. It has become increasingly difficult to safeguard confidentiality. Insurance companies, managed care organizations, other providers, courts and sometimes families claim a right to know what is going on in patients' treatments. The possibility of misuse of information from insurance data banks is a large issue. Confidentiality is one of the most fragile tenets in the contemporary era.

At times, principles may conflict. For example, the *Tarasoff* decision obligates a psychotherapist who has knowledge that a patient may intend harm to a third party to notify that third party of the potential risk. Mandatory reporting of suspected child abuse is another example in which the interest of the state in protecting children legally supersedes the physician's duty to maintain secrecy. In a larger context the erosion of confidentiality threatens to undermine the possibility of a kind of therapy that requires developing openness to another person confident that what is disclosed is only for the patient's benefit.

6.4 The 'False Memory' Controversy

Much conflict has developed around the issue of therapists directing therapy to recover memories of childhood sexual abuse, which may not have occurred. In particular, there have been instances reported in which therapists used suggestive methods such as hypnosis in order to recover memories of past sexual abuse that were believed to be repressed. Alleged perpetrators of such abuse have been prosecuted and even jailed. Therapists' zeal about these issues may lead to unethical behavior, even with good intentions, with destructive results for their patients and families (Chodoff 1999).

6.5 The Double Agent Problem

Double agents work both sides. There are situations in which the physician works for someone other than the patient, and there may be a conflict between the patient and the physician's employer resulting in a divided allegiance and compromised confidentiality. The classic case of this is the physician working for the military, trying to get soldiers back into combat. This is ethically analogous to treating the psychotic pris-

oner incompetent to face execution. The student health psychiatrist doing an evaluation for the dean or the psychiatrist doing a pre-arraignment examination have loyalties that go beyond the patient's best interest and which require at the least disclosure of the purpose of the examination and how the information will be used. Increasingly physicians working for corporations, managed care organizations and government hospitals have interests which radically transform the nature of the relationship between doctor and patient.

6.6 *Fees, Billing, and Reimbursement*

The financial arrangements for medical service strike at the heart of the ethical issues in how medicine is valued. Inevitably there must be some sort of exchange for the services rendered, and inescapably there must be some sort of feelings about the monetary transaction. In the dyadic doctor–patient relationship that is a very direct transaction, though it may be buffered by insurance companies, billing and collection processes. Patient benefit and fee-for-service have traditionally gone hand in hand, and the fee (alongside honor and duty, of course) has helped focus the physician's attention on whose benefit should receive attention. One of the general criticisms often leveled against traditional (Hippocratic) medical ethics is that it does not deal with the physician's larger obligations to society and particularly to the distribution of medical services.

One financial arrangement in particular that sometimes questioned is the practice of charging for missed sessions. The APA *Annotations* has specifically addressed this and has said, 'It is ethical for the psychiatrist to make a charge for a missed appointment when this falls within the terms of the specific contractual agreement with the patient' (Sect. 2, paragraph 6). The idea of patients having responsibility for the treatment and ultimately for their health and wellbeing is not something usually considered in medical ethics, but it goes beyond the idea of medicine as technology that the doctor delivers and stresses the patient's investment and commitment to the healing process.

Another financial arrangement that is addressed throughout the AMA *Opinions* and APA *Annotations* is fee-splitting. Physicians should receive remuneration for the work they do, and they should obtain patients by referral based on the quality of their work and the reputation they earn in the eyes of satisfied patients and other physicians. Any incentive such as splitting the fee with a referring physician is considered unethical. It offers a temptation to place monetary gain above the best interest of the patient.

6.7 *Informed Consent*

An issue that has been receiving increasing attention recently is the importance of preserving patient autonomy and control in the context of medical decision-making and participation in research. Recent debate has occurred around the issue of whether patients with psychiatric illnesses are competent to give consent in such situations. The danger in such debate is that psychiatric patients can be further stigmatized with policies that are overprotective and confuse psychopathology with a generalized lack of competence. Efforts to educate patients regarding the nature of medical decisions and research projects should protect their rights by helping them to make decisions for themselves (Appelbaum 1998).

7. *The Ethics of Managed Care*

The current revolution in healthcare financing challenges many traditional assumptions of medical ethics. The most fundamental change is the departure from a dyadic doctor–patient relationship. Doctors in the new environment are being transformed into 'providers,' patients into 'consumers,' and any number of 'third parties' claim an interest in what goes on in that relationship.

In the economic model of marketplace transactions, much more is at stake than how the payment is to be allotted. The metaphor of healthcare as an industry radically transforms the healing process to physiological interventions that take place in a delimited period of time. Physicians become technicians, and patients become recipients of technology, with both likely to be strangers to one another. Medicine is transformed from a human service into a commodity.

In the interest of efficiency or economy these transformations might seem warranted, but the practices we are currently witnessing in the name of managed care must themselves be subjected to ethical scrutiny. Certainly much of what is going on is not ethically justifiable, nor does it serve the ends it was purported to serve. The transformation of medicine from a profession to a trade by the Federal Trade Commission was justified as a cost-saving measure on the grounds that increasing consumer choice (through advertising) would lower costs. The value placed on autonomy set the stage for a transformation in which consumers have very little choice about the healthcare they receive. Healthcare has become an investment opportunity where vast amounts of money have been siphoned away from service delivery into the pockets of executives and shareholders of mega-corporations. The market solution has been a nonsolution to this point in that it has only succeeded in lowering costs without adequately addressing quality of service or the distribution and allocation of services.

The ethical physician is inevitably in tension with a system designed to limit or deny the care given to his or her patients. While recognizing that resources are finite, the ethical physician must struggle to do the best

possible for individual patients without compromising integrity. Many of the practices of managed care are fraudulently unethical, such as gag rules prohibiting physicians from discussing treatment options with their patients. Such policies do not enhance consumer choice, but diminish patient autonomy and oppose informed consent, placing the physician in the position of a double agent without the possibility of disclosure. Many practices of managed care are more subtly unethical, such as misleading advertising or incomplete disclosure. Some practices, such as limiting who may serve on panels, may be in restraint of trade. They are certainly not pro-competitive. Managed care is an ethically unstable response to the healthcare financing dilemma and must be subjected to thoroughgoing ethical evaluation. A just society must demand accountability for its citizens.

8. *Conclusion*

Looking back to the longstanding traditions of medical ethics, we are now in a position to better appreciate the place of the Hippocratic Oath in contemporary medicine. Though it is sometimes criticized for being anachronistic, its true value is apparent. The Hippocratic tradition in medicine provides a perspective of enduring values by which the shortcomings of modern economic expediency may be judged.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Culture as a Determinant of Mental Health; Deceptive Methods: Ethical Aspects; Ethical Dilemmas: Research and Treatment Priorities; Ethics Committees in Science: European Perspectives; Ethics for Biomedical Research Involving Humans: International Codes; Psychotherapy: Ethical Issues; Suicide: Ethical Aspects

Bibliography

American Medical Association 1957 *Opinions and Reports of the Judicial Council*. AMA, Chicago

American Medical Association 1997 *Code of Medical Ethics (150th Anniversary Edition); Current Opinions with Annotations of the Council on Ethical and Judicial Affairs*. AMA, Chicago

American Psychiatric Association 1993 *The Principles of Medical Ethics with Annotations Especially Applicable to Psychiatry*. American Psychiatric Press, Washington DC

Appelbaum P S 1998 Missing the boat: Competence and consent in psychiatric research. *American Journal of Psychiatry* **155**: 1486–8

Bloch S 1991 The political misuse of psychiatry in the Soviet Union. In: Bloch S, Chodoff P (eds.) *Psychiatric Ethics*, 2nd edn. Oxford University Press, Oxford (UK), New York and Melbourne (Australia), pp. 493–516

Blosch S 1999 Misuse and abuse of psychiatry. An overview. In: Blosch S, Chodoff P, Green S A (eds.) *Psychiatric Ethics*, 3rd edn. Oxford University Press, Oxford, UK, pp. 493–516

Bloch S, Reddaway P 1977 *Psychiatric Terror: How Soviet Psychiatry is Used to Suppress Dissent*. Basic Books, New York

Chodoff P 1999 Misuse and abuse of psychiatry: an overview. In: Bloch S, Chodoff P, Green S A (eds.) *Psychiatric Ethics*, 3rd edn. Oxford University Press, Oxford, UK pp. 49–66

Dyer A R 1988 *Ethics and Psychiatry: Toward Professional Definition*. American Psychiatric Press, Washington, DC

Dyer A R 1995 Advertising. In: Reich W (ed.) *Encyclopedia of Bioethics*, 2nd edn. Macmillan, New York, pp. 76–9

Dyer A R 1997a Ethics, advertising, and assisted reproduction: the goals and methods of advertising. *Women's Health Issues* **7**: 143–8

Dyer A R 1997b Ethics of human genetic intervention: a postmodern perspective. *Experimental Neurology* **144**: 168–72

Greenhouse J 1982 Justices uphold right of doctors to solicit trade. *The New York Times* 24 March, p. 10

Hauerwas S 1977 *Truthfulness and Tragedy: Further Investigations into Christian Ethics*. University of Notre Dame Press, Notre Dame, IN

Hauerwas S 1981 *Community and Character*. University of Notre Dame Press, Notre Dame, IN

Levine M 1972 *Psychiatry & Ethics*. G Braziller, New York

Pincoffs E 1971 Quandary ethics. *Mind* **75**: 552–71

Pincoffs E 1980 Virtue, the quality of life and punishment. *The Monist* **63**: 23–7

M. N. Miller and A. R. Dyer

Ethics and Values

In the twentieth century, moral philosophy and the theory of value came to be explicitly subdivided into two elements: metatheory, which concerns questions about meaning and justification; and normative theory, which concerns more substantive questions about what is good, right, or valuable. In earlier centuries, these two kinds of question were seldom distinguished. More recently, dissatisfaction with any sharp separation of these areas of inquiry has grown: how are we to judge the adequacy of any account of meaning without asking whether it does some justice to widespread substantive convictions? The distinction nonetheless will make the presentation below more manageable. Its focus will almost exclusively be on so-called 'analytic,' Anglo-American philosophy and its claimed antecedents, although fundamental work on ethics and the theory of value has of course been done in other traditions.

1. Ethics

1.1 Metaethics

There is no agreed-upon distinction between, say, ethics and morality, and the terms are used more or less interchangeably (though Williams 1985 associates 'morality' especially with duty-based systems). Both terms show in their origin a relation to social customs—ethos and mores—and both have a descriptive as well as a normative sense. In the former sense, we can ask 'What was the Greek warrior's ethic?' or 'How has morality evolved in Western Europe over last two hundred years?' In the latter sense, we can ask 'What is the ethical thing to do when one accidentally learns of a deep confidence?' or 'Is it moral for us to carry out animal experimentation?' (see *Conventions and Norms: Philosophical Aspects*).

The descriptive sense seems fairly straightforward, at least once one has noticed that there are wider and narrower uses (see Gibbard 1990). On the wide use, a morality is virtually any code of conduct involving attitudes of approval and disapproval, regardless of its particular content or social standing: a warrior ethic might favor unstinting cruelty toward one's enemies; a personal morality might glorify selfishness or esthetic accomplishment over all else, even if society finds this abhorrent. On the narrow use, morality is taken to involve norms of conduct with a distinctive purport, sometimes characterized in terms of a moral point of view, as contrasted with, say, an esthetic or prudential point of view (see Baier 1958). Morality, narrowlyunderstood, is said to involve some or all of: impartiality, concern for the well-being of others, treating like cases alike, willingness to cooperate, a notion of personal agency and responsibility, associated feelings of impersonal anger, guilt, or deservingness. In the broad sense, we can ask of virtually anyone or any society, 'What is his (or their) morality?'; in the narrow sense, we can ask, 'Does this individual (or this society) have a morality, and, if so, when did it emerge?' Much of the interest of philosophers and social scientists in individual moral development or in the historical development of morality concerns morality in the narrow sense. For example, there is an active debate over whether various ancient civilizations recognized morality, as distinct from enlightened prudence or religious requirement (see Annas 1993).

An important moment came when Prichard (1912) convinced virtually everyone that showing certain principles of conduct to be advantageous would not suffice to show they had distinctively moral standing. Recent interest in the possible evolutionary origins of morality also typically concerns morality in the narrow sense: 'debunking' views tend to emphasize that the gene- and kin-oriented mechanisms of natural selection are not impartial; 'vindicating' views tend to emphasize that these mechanisms could have equipped us with psychological resources—empathy, cooperativeness, concern for fairness—that can be mobilized culturally on behalf of genuinely moral practices (for examples, see Thompson 1995).

Philosophical analysis of moral language in the normative sense also has focused on the narrow use of the term, though this has not eliminated controversy. Central to the disputes has been the question whether moral claims are capable of truth or falsity, the question of 'cognitivism.' Noncognitivists have maintained that the primary semantic function of moral claims is not to state facts, but to express attitudes or commands intended to influence behavior, and so they are incapable of being strictly true or false.'Noncognitivism' is in some ways an unfortunate term, since moral discourse clearly involves a great deal of cognition and conceptual content, even according to those who claim its primary function is to express attitudes (see Stevenson 1944, Hare 1952, Gibbard 1990, Blackburn 1984). For an attitude to count as distinctively moral, it must represent itself as responsive to certain features rather than others (those characterized by a moral point of view), and as regulative for certain emotions (e.g., impersonalanger). Recently so-called 'minimalist' theories of truth have gained currency, according to which the labels 'true' or 'false' can appropriately be applied to sentences of a discourse so long as they enter seamlessly into familiar logical relations, patterns of inference, etc. (Horwich 1990). Since moral statements unquestionably do this—'Stealing is wrong' can be the premise of a normal logical inference, but a command like 'Don't you dare steal!' cannot—there has been a tendency to concede on all sides of the debate that moral statements can be said to be 'truth-apt' or 'propositional' in the minimal sense. This might or might not suffice to earn the label 'realism' about moral properties (see Blackburn 1993).

The burning issue has become: even if propositional, do moral statements have an 'internal' or conceptual connection with motivation, such that making a moral judgment involves an element of endorsement not found in ordinary factual judgments? One can perfectly appropriately assert, for example, 'This rock is granite' without expressing any favorable or unfavorable attitude toward that state of affairs. If one cannot similarly assert 'This situation is immoral' with perfect sincerity and linguistic propriety while having no attitude of disapproval, then there remains a fundamental semantic difference between moral and 'merely descriptive' discourse.

Defenders of 'propositionalism' face a choice: either insist that this expressive function belongs to the pragmatics of moral discourse rather than its semantics, or devise an account of the propositional meaning of moral statements that incorporates motivation 'internally.'

Subjectivists pursue the latter path. A simple subjectivism holds that moral statements are tantamount

to claims about what one, or one's society, approves or disapproves. This view results in a kind of moral relativism: which moral judgments may be applied to an individual or a society depends upon which norms or principles he, she, or they accept. Relativist views have enjoyed a certain popularity, since they seem laudably cosmopolitan, although they typically account poorly for moral disagreement (except as a factual debate over which norms are accepted) and often verge on incoherence, since they may include a nonrelativistic moral recommendation of tolerance in response to moral difference (but see Harman and Thomson 1996 for a sophisticated debate on relativism). Contemporary subjectivists therefore take a more objectivist approach, inspired in part by an analogy between moral qualities and secondary qualities, such as color (McDowell 1985). Color judgments show a high degree of objectivity and intersubjective agreement, so much so that they are often used in scientific observation and social coordination (e.g., 'red' for 'stop'). Yet colors are nonetheless based in part on a subjective response, and those without a certain human-typical perceptual apparatus are incapable of seeing color. Moral judgments might be similar. Those without a certain human-typical motivational apparatus might be incapable of moral experience, but since virtually all of us have (or could develop) these motivations, moral judgments can also be highly objective and intersubjective. So we might think of moral properties neither as independent features of the world, nor as mere projected feelings on our part, but as a pairing of features of situations and characteristic (or ideal-typical) human responses to those features—rather in the way dangerousness might be seen as a pairing or risk and fear—and therefore as response-dependent.

In the moral case, the response—like fear—would have a motivational aspect, perhaps arising from moral training in the virtues rather than sheer human physiology. For someone equipped with moral character, then, moral experience and judgment are necessarily accompanied by a motivating response. Does this sort of subjectivism lead to relativism—different patterns of moral training (say) yield different, yet locally apt, moral judgments? Not quite, since we can take our moral discourse, like our actual color discourse, to be anchored to our actual responses (by a semantic device known as 'rigidification'; see Wiggins 1998). The difficulty now is that those in different moral traditions may simply turn out not to have a common moral vocabulary with us.

Nonsubjectivists typically reject the semantic or 'internal' tie to motivation, and claim that moral judgments can be made without linguistic impropriety even by someone for whom they are normally motivationally inert. Such 'externalist' approaches recognize that in our ordinary life we presuppose certain typical motivations—so that it will be somewhat misleading to utter 'This situation is immoral' if one does not intend thereby to indicate or express any negative attitude toward it. But similarly, if, in normal circumstances, one utters 'I believe that she's coming at noon,' one's listeners can suppose that one thereby endorses their believing likewise, though one does not *contradict* oneself in saying, 'I believe that she's coming at noon, but don't take my word for it.' Expressions of belief, and of moral evaluation, play certain roles in our social lives owing to various prevailing or presupposed background conditions, but these do not all become part of the very meaning of these words—or perhaps, the distinction between what 'belongs to meaning' and what 'belongs to pragmatics' is so soft that we should not allow ourselves to be driven to surprising conclusions like relativism or rigidification by the so-called 'motivation problem.' Externalists are divided into naturalists, who think the properties attributed in moral discourse are within the ambit of ordinary empirical knowledge, and non-naturalists, who think of moral properties as in some sense *sui generis*.

The first need to explain how morality can be so controversial, if moral questions are in principle resolvable by matters of natural fact; the second need to explain how moral properties can be known or referred to, if they are not causally related to our experience. Some naturalists outright identify moral concepts with certain natural concepts; others give a broadly functional characterization of the role or meaning or moral terms—the truisms about morality and its social role—and leave conceptually 'open' the question which particular acts, practices, etc. play the relevant role (for some varieties, see Brandt 1979, Railton 1986, Boyd 1987, Brink 1989, Lewis 1989, Jackson and Pettit 1996). Of course, one possibility is that nothing can be found that will play all or most of the role ascribed to moral properties, resulting in an error theory of moral discourse rather like atheism about religion (Mackie 1977). Non-naturalists argue that morality is not alone in showing the need to give an account of how we are able to talk about or have knowledge of 'nonempirical' phenomena—mathematics poses similar problems (see Ross 1930, Wiggins 1988).

G. E. Moore, a non-naturalist externalist, famously diagnosed a 'naturalistic fallacy' in moral philosophy, but the import of this claim has been widely misunderstood (Moore 1903). He did not doubt, and few in the twentieth century have doubted, that moral evaluation supervenes on nonmoral, natural properties—roughly, those properties accessible to empirical experience and investigation. It is, according to this view, a mistaken picture of value to think that two situations could be identical in all empirical features yet one correctly judged good and the other bad. In this sense, then, 'naturalism' is not at issue in these metaethical debates, and a kind of transition from 'is' to 'ought' is possible: we are able to learn whether acts are good or bad, once we have mastered moral

concepts, by learning of their empirical features. Moore's point is that moral concepts are conceptually distinct from the natural concepts or criteria we might use to apply them—we do not *contradict* ourselves by asking whether happiness is good in itself, even if, as a matter of fact, it is (Frankena 1939). This is perhaps to be contrasted with supernaturalist thinking about ethics, according to which some supernatural feature— such as autonomous approval by a deity—is necessary for moral standing. Most contemporary moral philosophy would not sustain the conclusion that 'If God is dead, everything is permitted.'

Oscar Wilde has Mrs. Cheveley say, after being reprimanded by Gertrude Chiltern, 'Do you know, Gertrude, I don't mind your talking morality a bit. Morality is simply the attitude we adopt towards people whom we personally dislike' (*An Ideal Husband*, Act II). Such a 'debunking' depends on a contrast between the normative purport of morality and the normative status it actually possesses or deserves. If 'x is wrong' simply meant 'I personally don't like x', there would be no presumption that it would be bothersome to be told that what one is doing is wrong. Morality might simply dress up personal dislikes in judgmental clothing. But giving an account of the meaning of moral terms—a metatheory—involves saying just what this dressing-up consists of. (A collection of related essays and a more extensive survey is to be found in Darwall et al. 1996.)

1.2 Normative Ethics

It is difficult to imagine that moral properties have nothing to do with likes and dislikes, or broadly, with what matters to people or to sentient beings generally. Thus there does appear to be something profoundly subjective about it. Yet it strives to develop objective forms of evaluation on this subjective base. We can see normative moral theories as seeking to develop a synthesis of subjective and objective elements that: (a) presents a credible account of moral rightness and value, one which fits reasonably well with our firmest moral judgments on reflection; (b) is psychologically and socially realistic ('ought implies can'; a condition that is growing in importance in philosophy as we learn more about cognitive and motivational psychology and social dynamics); (c) involves no suspect metaphysical or epistemological assumptions (see *Parfit* 1984, Sinnott-Armstrong and Timmons 1996); and (d) enables us to locate a compelling basis for moral evaluation that would explain how those judgments could credibly be justified, criticized, and improved—since we are all aware of moral judgments that seemed obviously right or self-evident (such as the permissibility of slavery or the moral irrelevance of animals), that we have subsequently come to reject. A normative moral theory should illuminate the deeper springs of moral concern (if such can be found), even if it calls for some normative rethinking on our part. This set of concerns is sometimes called seeking wide reflective equilibrium (Rawls 1971; see *Reflective Equilibrium*).

We find three major living traditions in normative moral theory, one claiming as its spiritual ancestor Aristotle, the second David Hume, and the third Immanuel Kant.

On Aristotelian accounts, the fundamental question of ethics is 'What is the good life?,' and the answer is thought to take the form: the virtuous life (for some Aristotelian approaches, see Aristotle 1989, McDowell 1985, Williams 1985, MacIntyre 1981). Virtues are robust traits of character, drawing on normal human faculties and motivations but shaped by proper moral training, that enable an individual to see what is best or required, and act appropriately, in a wide range of circumstances. Virtue thus involves cognitive, affective, and conative elements. Moreover, virtues are thought to be mutually-reinforcing ('the unity of the virtues'), and to yield a life that is itself good for the virtuous person to lead. Typically, one cannot formulate strict principles of action that govern the virtuous agent: moral conduct essentially involves the exercise of judgment (shaped by the 'moral experience' of the virtuous agent), yet this need not preclude fairly objective recommendations about specific situations—much as we think that persons of sensitivity, discernment, and knowledge tend to converge in esthetic judgment even if esthetic evaluation cannot be captured in terms of strict rules. Although virtue theories take agents and their motives as the primary locus of moral evaluation, the view need not simply be individualistic. Aristotle, for example, thought the virtuous life to be possible only in the right sort of social and political setting, and only if one plays one's appropriate role in that setting (see *Virtue Ethics*).

Humean accounts, though they may make room for virtue, typically do not assign it so fundamental a role. They seek to understand morality as a scheme of social cooperation and regulation that enables individuals and groups to achieve sought-after ends that they could not attain individually, or on the strength of native instinct. In Hume's own view (see Hume 1888, a related view is found in Adam Smith 1976), we are natively equipped with many of the sentiments necessary for such cooperation and regulation—empathy (what Hume called 'sympathy'), an intrinsic concern for kith and kin, a willingness to defer to custom, etc.—but others within this tradition (e.g., Bentham) suppose humans to be egoistic hedonists. Each strand solves the 'motivational problem' for morality in a distinctive way—Hume by the reshaping of native sentiment toward greater impartiality and willingness to cooperate through social conventions and 'artificial' in addition to 'natural' virtues, Bentham by the use of socially backed 'sanctions' for changing the incentives individuals face. Both of these strands recognize the need for a common, objective standard for moral

assessment: 'the general interest.' An influential form of this general approach is utilitarianism, which has taken a vast number of particular forms. Essential to it are two features: (a) a notion of a person's good or well-being that can be assessed prior to judging moral worthiness—a theory of intrinsic nonmoral value (which need not be hedonistic, but could be perfectionist or involve the satisfaction of informed desires, personal or impersonal); (b) a way of evaluating acts or practices across individual lives and over time by comparing tendencies to promote intrinsic nonmoral value, impartially assessed (which characteristically is thought of as maximizing, but need not be).

Such value-based theories, often called consequentialist or teleological (end-oriented), hold that the good is prior to the right (see *Consequentialism Including Utilitarianism*). Morality partakes of the subjective because intrinsic goodness is grounded in what does, or could, matter to people (or to sentient beings generally) for its own sake. Yet morality is objective, because it seeks to ground distinctively moral judgments on an unbiased and impartial assessment of how that act or practice affects the long-term realization of intrinsic value in our lives.

On Kantian accounts, there is an attempt to locate a ground of rightness that is not in this way wholly dependent on such tendencies to promote nonmoral good. In some sense, the end does not entirely justify the means, and certain constraints on right action, or prerogatives for individual choice, exist that are not functions solely of actual or typical consequences. Indeed, there is on Kantian views some resistance to the idea that we can characterize 'intrinsic nonmoral good': happiness derived from using another inappropriately as a means, e.g., torturing an innocent, is not, they claim, a good at all. Normative theories that thus claim that the right is prior to the good are sometimes called deontological (duty-based). Although Kant's own view is a bit hard to categorize at its foundation, views of Kantian inspiration have become the dominant and best developed form of philosophical deontology. (Other forms include: Lockean natural rights theories, e.g., Nozick 1974, see *Libertarianism*; and rational intuitionism, e.g., Ross 1930. For a discussion, see Scheffler 1982.) Humean moral theorists are able to give a rationale for encouraging nonconsequentialist moral norms or motivations by a 'second-order' consequentialist argument: this is necessary to avoid individually or collectively self-defeating behavior (see Parfit 1984). But how are we to give a rational justification of moral actions and duties, or constraints and prerogatives, if we are willing to fly directly in the face of consequences for what we most value? The Kantian approaches this question from the standpoint of a responsible agent asking 'What have I reason to do?' (see *Practical Reasoning: Philosophical Aspects*). From this standpoint, we recognize that desires and goals do not automatically give reasons to act, since they are themselves up for scrutiny—and not merely because some are stronger or more basic than others.

A desire for revenge, however powerful or however much a part of human nature, ought to be fought, rather weighed in an individual or social 'utility function.' We know from experience that we are free enough from the control even of 'natural' desires and interests to do this. If one claims that one's act is rationally permitted, one in effect licenses the maxim on which it is based as legitimate for anyone in similar circumstances. Like ideal legislators, we thus authorize ourselves to act by 'making law,' aware of the condition that we ought to be—even when we would prefer otherwise—subject to the very same law (the imperative is categorical). This universalizability condition on our wills brings an element of objectivity and generality to our decisions, based not on actual or probable consequences, but on a notion of ourselves as autonomous (literally, 'self-law-giving'; see *Autonomy, Philosophy of*), seeking the moral law (see Kant 1964; also see *Kantian Ethics and Politics*). Some contemporary Kantians prefer to understand this in terms of a hypothetical social contract: we show that our action is permitted or required by rules that we would all agree to (or find reasonable) in a situation of initial freedom and equality. Some acts or practices that would have good consequences—promoting social peace by treating individuals paternalistically in a secretive way, or permitting inequalities that raise total utility but do not tend to benefit all members of society—might thereby be ruled out (Rawls 1971, Scanlon 1998; for a related view, expressed in terms of the conditions for nondistorted social communication, see Habermas 1990; also see *Contractarianism*). Other contemporaries inspired by the notion of autonomy have seen in this doctrine a radical affirmation of individual human freedom. Jean-Paul Sartre, for example, gave a distinctively existentialist account of 'choosing for all mankind,' holding that we can lapse into bad faith if we attempt to escape this absolute responsibility for our acts by pointing to things 'beyond our control,' such as 'regrettable circumstances,' 'natural urges,' or even 'moral principle' itself (Sartre 1946).

One might see these three traditions as each emphasizing a different, yet important root of morality: the notion of the moral life as the best or fittest sort of life for humans (Aristotle); the notion of the moral life as one of benevolence and freedom from prejudice (Hume); and the notion of the moral life as one of respect for one's own freedom and rationality as well as that of others (Kant). Can all these elements be combined into one normative theory? No overarching synthesis has been found, and any view that seeks to be at all systematic cannot, so to speak, guarantee that it is optimal in three dimensions at once. Kant, aware there could be no guarantee, nonetheless took it to be an essential condition of the rationality of moral conduct that autonomous self-legislation would in the

end yield both good consequences overall and a good life for humans—happiness and virtue would be united as the 'highest good,' if not in this life, then in the afterlife—but later Kantians have largely given up this doctrine of 'rational faith' and argued that Kant's account can stand even without postulating that these conditions are met.

An alternative response to this sense of multiple roots of morality is to claim that no systematic account of normative ethics should be attempted. Different circumstances, roles, personal histories, and so on make different sorts of moral consideration more or less appropriate, and this is an ineluctable fact of moral life. A relation with one's spouse or child is seldom best regulated by 'the general interest' or 'universalizable maxims,' a legislator cannot generally rely on personal virtue to tell her how to shape a complex piece of environmental law, and in our relations with animals we cannot regulate our conduct as if we were operating in a reciprocal realm of rationality. Experience and wisdom, and caring, not theory, enable us to act and deliberate well in a complex moral world (see Larmore 1987). This view was given some impetus by, and finds continuing support in, a growing literature on moral development, where individuals who score highly on traditional tests of moral-theoretical sophistication may do poorly in practice in comparison with those capable of greater personal caring, empathy, and adaptiveness (see Gilligan 1982, Damasio 1994). Defenders of normative theory insist that systematic ambition is an essential antidote to arbitrariness or narrow-mindedness, even though it must of course be combined with suitable motivation and sensitivity. Moral particularists (Dancy 1993) and some feminists (see Card 1991) counter that the systematizing impulse can itself become a form of moral insensitivity and arbitrariness.

This is probably a healthy dialectic in normative ethics. It is important to emphasize, however, that resistance to normative moral theorizing need not be combined with a resistance to metaethical theorizing: moral particularism is often defended on the ground of an account of the meaning of moral concepts or the ground of moral justification. Thus, on one influential view (Ross 1930), morality consists of many mutually independent *prima facie* duties and permissions, the bearing of which on particular situations is not governed by general rules. Ross grounded his view in a non-naturalist intuitionism in metaethics, and numerous contemporary defenses of Ross-like normative theories or other forms of particularism draw upon the idea that moral concepts are response-dependent (see Dancy 1993) or that fundamental values are plural and incommensurable (see Stocker 1990).

2. *Theory of Value*

In the first half of the twentieth century, theory of value (axiology) was a flourishing, central area of philosophy, represented by such major figures as G. E. Moore (1903), John Dewey (1929), C. I. Lewis (1962), and Ralph Barton Perry (1954). It has since gone so thoroughly out of style that it is difficult for a contemporary philosopher to recognize the themes or structure of a book like Perry's *Realms of Value*, with its chapters on *Motor-Affective Psychology, Gradation, Integration, and Mutation, The Science of Science*, and *The History of History*. This may partly be the result of a familiar tendency within disciplines to subdivide into specialities: ethics, esthetics, rational choice, law, epistemology—all of these are domains of evaluation, yet they seldom are united together under a common rubric. Yet, like specialization generally, it involves some loss of perspective. The value-theoretic aspects of some subdisciplines within philosophy became, at least until fairly recently, relatively invisible (epistemology, for example) or radically de-emphasized ('philosophical esthetics'). This has also given the mistaken impression that the so-called 'fact/value' distinction is disciplinary—ethics on one side, science on the other. But value theorists would be quick to point out that values (but not simply moral values) as well as facts are fundamentally involved in all areas of human practice and inquiry.

There are excellent reasons for focusing on value as such. There is independent interest in investigating the nature and structure of value. Some questions about objectivity, cognitivism, and realism are familiar from metaethics, but others, such as the distinctions among intrinsic, extrinsic, inherent, and instrumental value, or the holistic vs. atomistic character of value, or the possibility of commensurability or interpersonal comparison, merit sustained scrutiny (see also *Value Pluralism*).

Moreover, that moral philosophy in the twentieth century has been overwhelming preoccupied with questions of rightness and obligation (deontic or directive categories), rather than questions of value (axiological or evaluative categories), has given us a lopsided picture of morality and of the 'motivation problem' for ethics. Perhaps an ought judgment expresses some motivation on the speaker's part, but a judgment that a way of life would be valuable seems to be more dependent on the actual or potential motivation of those to whom it is applied. Further, evaluative judgments in general seem more directly responsive to facts of psychology, sociology, and history than directive judgments (see Wiggins 1998, Williams 1985). The emphasis on deontic judgment has also tended to divorce philosophical ethics from value-involving subjects where duty plays a less conspicuous role. Kant and Hume, for example, saw quasiesthetic evaluative judgment as at the foundation of morality.

Finally, a certain lost opportunity for philosophy for mutual enlightenment between 'analytic' and 'continental' philosophy occurred during the twentieth century. Moore (1903) observed that the philosopher

whose outlook seemed closest to his own was Brentano (1902). Theorists within the phenomenological tradition in philosophy and psychology, such as Kohler (1938) and Heidegger (1962), explored our experience of value in life in ways quite comprehensible from the standpoint of, say, Dewey (see *Phenomenology: Philosophical Aspects*). Frankfurt School (see Habermas 1990) criticisms of an overly narrow construal of value can be read alongside a growing body of Anglo-American literature critical of current notions of value in philosophy and economics, especially in rational choice theory and cost-benefit analysis (see Maclean 1986, Anderson 1993; also see *Economics and Ethics*).

See also: Aristotle (384–322 BC); Conventions and Norms: Philosophical Aspects; Dewey, John (1859–1952); Hume, David (1711–76); Kant, Immanuel (1724–1804); Kantian Ethics and Politics; Kohlberg, Lawrence (1927–87); Law and Morality: A Continental-European Perspective; Law and Morality: An Analytic Perspective; Moral Reasoning in Psychology; Normative Aspects of Social and Behavioral Science; Norms

Bibliography

Anderson E A 1993 *Value in Ethics and Economics*. Harvard University Press, Cambridge, MA

Annas J 1993 *The Morality of Happiness*. Oxford University Press, New York

Aristotle 1989 *The Nichomachean Ethics* [trans. W D Ross], rev. edn. Oxford University Press, Oxford, UK

Baier K 1958 *The Moral Point of View*. Cornell University Press, Ithaca, NY

Blackburn S 1984 *Spreading the Word*. Oxford University Press, Oxford, UK

Blackburn S 1993 *Essays in Quasi-Realism*. Oxford University Press, New York

Boyd R 1987 How to be a moral realist. In: Sayre-McCord G (ed.) *Essays on Moral Realism*. Cornell University Press, Ithaca, NY

Brandt R B 1979 *A Theory of the Good and the Right*. Oxford University Press, New York

Brentano F 1902 *The Origin of Knowledge of Right and Wrong*, trans. C Hague. Constable, London

Brink D O 1989 *Moral Realism and the Foundations of Ethics*. Cambridge University Press, Cambridge, UK

Card C (ed.) 1991 *Feminist Ethics*. University Press of Kansas, Lawrence, KS

Damasio A R 1994 *Descartes' Error: Emotion, Reason, and the Human Brain*. G. P. Putnam, New York

Dancy J 1993 *Moral Reasons*. Blackwell, Oxford, UK

Darwall S, Gibbard A, Railton P 1996 *Moral Discourse and Moral Practice*. Oxford University Press, New York

Dewey J 1929 *The Quest for Certainty*. Minton, Balch, New York

Frankena W K 1939 The naturalistic fallacy. *Mind* **48**: 464–77

Gibbard A 1990 *Wise Choices, Apt Feelings*. Harvard University Press, Cambridge, MA

Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA

Habermas J 1990 *Moral Consciousness and Communicative Action* [trans. Lenhardt C, Nicholsen S W]. MIT Press, Cambridge, MA

Hare R M 1952 *The Language of Morals*. Oxford University Press, Oxford, UK

Harman G, Thomson J J 1996 *Moral Relativism and Moral Objectivity*. Blackwell, Cambridge, MA

Heidegger M 1962 *Being and Time*, trans. Macquarrie J, Robinson E. Harper, New York

Horwich P 1990 *Truth*. Blackwell, Oxford, UK

Hume D 1888 In: Selby-Bigge LA (ed.) *A Treatise of Human Nature*, 2nd edn. 1740. Oxford University Press, Oxford, UK

Jackson F, Pettit P 1996 Moral functionalism and moral motivation. *Philosophical Quarterly* **46**: 82–6

Kant I 1964 *Groundwork of the Metaphysics of Morals* (1785), trans. Paton H J. Harper & Row, New York

Kohler W 1938 *The Place of Value in a World of Facts*. *Liveright*, New York

Larmore CE 1987 *Patterns of Moral Complexity*. Cambridge University Press, Cambridge, UK

Lewis C I 1962 *An Analysis of Knowledge and Valuation*. Open Court, La Salle, IL

Lewis D 1989 Dispositional theories of value. *Proceedings of the Aristotelian Society* Suppl. **63**: 113–37

MacIntyre A 1981 *After Virtue*. University of Notre Dame Press, Notre Dame, IN

Mackie J L 1977 *Ethics: Inventing Right and Wrong*. Penguin, Harmondsworth, UK

Maclean D (ed.) 1986 *Values at Risk*. Rowman and Allanheld, Totowa NJ

McDowell J 1985 Values and secondary qualities. In: Honderich T (ed.) *Morality and Objectivity*. Routledge & Kegan Paul, London

Moore G E 1903 *Principia Ethica*. Cambridge University Press, Cambridge, UK

Nozick R 1974 *Anarchy, State and Utopia*. Basic Books, New York

Parfit D 1984 *Reasons and Persons*. Clarendon Press, Oxford, UK

Perry R B 1954 *Realms of Value*. Harvard University Press, Cambridge, MA

Prichard H A 1912 Does moral philosophy rest on a mistake? *Mind* **21**: 21–37

Railton P 1986 Moral realism. *Philosophical Review* **95**: 163–207

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Ross W D 1930 *The Right and the Good*. Clarendon Press, Oxford, UK

Sartre J-P 1946 *L'existentialism est un humanisme*. Nagel, Paris

Scanlon T 1998 *What We Owe Each Other*. Harvard University Press, Cambridge, MA

Scheffler S 1982 *The Rejection of Consequentialism*. Oxford University Press, Oxford, UK

Sinnott-Armstrong W, Timmons M (eds.) 1996 *Moral Knowl edge?* Oxford University Press, New York

Smith A 1976 In: Raphael D D, Macfie A L (eds.). *The Theory of Moral Sentiments* (1759). Clarendon Press, Oxford, UK

Stocker M 1990 *Plural and Conflicting Values*. Oxford University Press, New York

Stevenson C L 1944 *Ethics and Language*. AMS Press, New York

Thompson P (ed.) 1995 *Issues in Evolutionary Ethics*. State University of New York Press, Albany, NY

Wiggins D 1998 *Needs, Values, Truth*, 3rd edn. Clarendon Press, Oxford, UK
Williams B 1985 *Ethics and the Limits of Philosophy*. Harvard University Press, Cambridge, MA

P. Railton

Ethics Committees in Science: European Perspectives

Research in medicine, including methods and results of other disciplines like biology or psychology, is for a major part, undertaken with the aim to improve the understanding of normal and abnormal structures, functions, and regulations of man. The results may contribute to a higher effectiveness of diagnosis and treatment of disease. Research on man may include the use of personal data, of cells or tissues of human origin, and interactions with, or interventions on, a person. It is obvious that research, understood in such a broad sense, can interfere with the autonomy, the dignity, and the fundamental rights of a person participating in a scientific project. In order to protect these fundamental rights the Declaration of Helsinki of the World Medical Association (1996) recommends, that an 'experimental protocol' should be 'transmitted for consideration, comment, and guidance to a specially appointed committee.' These committees are usually known under the term 'ethics committee,' a term covering a lot of very different bodies based on a variety of regulations, from formal laws to codes of deontology of doctors. These regulations vary from country to country. An older publication 'Ethics Bodies' (1992) gives an overview on the situation in 28 states of the USA. A recent publication (Deutsch and Taupitz 2000) reflects the approach in several states from various continents and different systems of regulations. It will be the basis for the following considerations.

It is generally accepted that the above mentioned Declaration of Helsinki had a most important influence on constitution and function of ethics committees, although it has nowhere reached the degree of a formal law. Nevertheless, it has been introduced in a very different way into the system of regulations of the states. In contrast to this, the declaration is obviously not observed in the USA.

Promoted by the declaration, ethics committees have been established by a broad variety of organizations, for example, university hospitals, general hospitals, medical associations, or pharmaceutical companies. This development very often preceded activities of parliaments or administrations. Introducing legal regulations, parliaments, and administrations, for their part, tried to respect and to overtake, as far as possible, with respect to constitution and internal right, the experience achieved by the already existing committees.

The first legal basis for ethics committees, in most cases, have been the drug laws. These specialized prescriptions have been expanded in an analogous way on all the other kinds of medical research on man, including medicinal products. Consequently, in some countries, the parliament has covered the whole of research on man by laws, decrees, or similar kinds of legitimation.

The protection of persons participating in biomedical research is generally considered being the principal obligation of ethics committees. The term 'protection' is understood as a prevention of physical damage as well as the protection of fundamental rights of the person concerned. The main point under this view of protection is to assure that the person participating in a research project gives his or her free, informed consent. This consent has to be given for a special project and has to be documented. Other points of consideration are data-protection and an adequate assurance of persons undergoing the research. In several countries the committees also have the duty to prevent the researcher himself or herself to be too audacious in undertaking a scientific project. Ethics committees have to prove the scientific quality and the ethic acceptance of research projects. There are a lot of points for such an examination, namely, minimization of risk by using methods with known risks whenever possible; acceptable relation between risk and potential direct benefit for the person concerned; and for the improvement of knowledge with the leading idea 'the lower the expected direct potential benefit, the lower the acceptable risk and burden,' a careful recruitment of participating persons with respect to the special problems of vulnerable persons. So far persons who are not able to consent, shall be included in the project, the informed consent has to be given by the legal representative in accordance with the national legislation. This representative has to receive all the information that is given to a person able to consent. Other points of examination are documentation of the results; use of collected data following the research protocol with guarantee for the privacy of the person; and protection of all persons (especially vulnerable persons) against force or undue influence on this decision.

In some countries the ethics committees have to prove the quality of the responsible researcher, the adequate equipment of the research institution, the foreseen benefit for the person concerned, as well as for the society and any kind of remuneration for researchers and persons participating in a project. In a number of countries, and this will be one of the most important tasks for the future, ethics committees have to monitor the research and to receive periodical reports of the ongoing project. Upcoming results or serious adverse events my lead an ethics committee to change its former vote.

The committees should intensify the discourse on ethical questions including statements of so-called national ethics committees. The unreflected acceptance of ethical guidelines, promoted by national ethic bodies, is neither wanted nor expected.

In decision-making the ethics committee has to respect national or federal laws regulating drugs, medicinal products, or the system of assurances. If there are special laws regulating medical research, like the '*Loi Huriet*' in France, the Medical Research Council Act in the Republic of South Africa, or the Federal Common Rule in the USA, these regulations have to be observed as well. Administration, legislation or 'guidelines for ethics committees' formulated by a competent body, may oblige ethics committees to fulfil the requirements of The Declaration of Helsinki.

In the Republic of South Africa, the ethics committee of the Medical Research Council instituted by law, is responsible for regulating and the control of research, as far as it is undertaken for, or with support of, the Medical Research Council. This council has the right to publish guidelines for the function of ethics committees. In the case of a conflict between these guidelines and the regulation of a local ethic committee the prescriptions of the Medical Research Council have priority. In New Zealand, the ethics committee, instituted by the Health Research Council, itself founded by law in 1990, has the right to elaborate ethical guidelines for the Research Council itself and to elaborate standards for other ethics committees. The principal task of this ethics committee is the examination and admission of local ethics committees, which have to be accredited in a formal way and must follow the so-called national standard.

Membership and composition of ethics committees vary in such a wide range, that a special consideration is not given with one exception: the composition has to be 'multi-disciplinary,' laymen should be included, both sexes have to be represented in an adequate manner.

Ethics committees need some kind of administration and financing. This is done by preference, for the national ethics bodies by Ministers of Social Affairs, of Health, or of Research. Local ethics committees are mostly supported by the instituting organization, for example, university hospitals. In some countries, for example, the Federal Republic of Germany, the states' laws prescribe that ethics committees have to be instituted by the medical associations and by faculties of medicine.

Generally, ethics committees are responsible for all research undertaken in the institution, by which they have been instituted. There may be some exceptions. In the Republic of South Africa only research projects that are supported by the Medical Research Council, have to be presented to such a committee. In Israel, the Director General of the Minister of Health has the right to reserve examination and approval of special kinds of research to the national ethics committee. In these cases, the decision of this committee is binding for local ethics committees. In some countries, the institution of ethics committees is restricted to states or community hospitals. Private hospitals are independent in their decision to introduce, or not introduce, such bodies.

The duty to seek the decision or the advice of an ethics committee is regulated in a different, but simple and understandable way. There are legal prescriptions, there is the legally based code of deontology, like in the Federal Rebublic of Germany or, in most cases, it is the administration, which requires such a step. In this field, considerations like approval by the responsible administration or financing play a major role. The conclusion is justified that the award of a grant depends on a favorable vote of an ethics committee.

In contrast, it is very difficult to define the legal or binding character of such a vote. Votes are considered as expertise, as anticipating expertise, as advice, as help and orientation, or as an approval. The character of advising is reflected by the official name of the national French ethics committee as '*Comité consultatif*.' Positive or favorable votes are considered in all countries as a basic condition for undertaking a research project. Also, in the case of a negative vote the researcher, at least theoretically, has the right to decide on the undertaking of his project. This situation is, again in theory, given, for example, in the Federal Republic of Germany. In reality, it is the ethics committee that decides by its vote whether a research project is done or not. During a research project, in some countries ethics committees have the right to intervene. They can, with respect to, for example, serious adverse events, change their decision or, in the worst case, withdraw a positive vote.

There is a growing interest on whether there is some kind of hierarchy of the ethics committee within a country. As already mentioned, some national ethics committees, for example, in Israel, in New Zealand, or in the Republic of South Africa, have the right to prescribe guidelines for local ethics committees, or to accept general points of view for special fields of research to be followed by local ethics committees.

In some countries, higher authorities may intervene or correct the decision of local ethics committees. A researcher who does not agree with the decision of the local ethics committee can call the national ethics committee in Denmark. The national Committee may come to another decision that will be binding. In France, a negative vote has to be presented to the Minister of Health, who has the right to suspend it. The Minister of Health in Spain has independent of the vote of the relevant ethics committee, the possibility to stop a research project or to introduce a change with respect to national laws.

There is only little knowledge on formation or continued education of members of ethics committees or of methods of control of an ethics committee as a

decision-making body. As an implemented system there is only the monitoring of IRBs by the 'Department of Health and Human Services' or by the 'Center for Drug Evaluation and Research' of the Food and Drug Administration in the USA.

Since 1997 the World Medical Association has tried to introduce a fundamental revision of its 'Declaration of Helsinki.' The 52 general assembly of the WMA adopted a revised version in October 2000. It is an open question whether this new version will gain the same outstanding importance as the former one. Beside this, there is a growing tendency in many states to introduce legal regulations of all kind of research by so-called 'research laws.' Such a law, based on a national legislation or on international conventions, for example, guidelines of the European Union or the 'Convention on Human Rights and Biomedicine' of The Council of Europe would bind all researchers, whereas the Declaration of Helsinki and prescriptions linked to it, are only compulsory for medical doctors. Ethics committees may exist in other fields of science other than medicine. Published details of such bodies are not known.

See also: Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives; Ethics for Biomedical Research Involving Humans: International Codes; Objectivity of Research: Ethical Aspects; Research: Ethnical Aspects of Long-term Responsibilities; Research Ethics, Cross-cultural Dimensions of; Research Ethics: Research; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects; Research Subjects, Informed and Implied Consent of

Bibliography

Deutsch E, Taupitz J (eds.) 2000 *Freedom and Control of Biomedical Research*. Springer, Berlin

Le Bris S 1992 *National Ethics Bodies*. Council of Europe, Strasbourg, Austria

World Medical Association 1996 Declaration of Helsinki—Recommendations Guiding Physicians in Biomedical Research involving Human Subjects. Ferney-Voltaire Cedex, France, no. 17

E. Doppelfeld

Ethics for Biomedical Research Involving Humans: International Codes

There are three leading international codes of ethics for biomedical research involving human subjects: the Nuremberg Code (reprinted in Levine 1988), the World Medical Association's (WMA) Declaration of Helsinki (2000) and the International Ethical Guidelines for Biomedical Research Involving Human Subjects developed by the Council for International Organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization (WHO) (1993); these will be referred to respectively as 'Nuremberg,' 'Helsinki' and 'CIOMS Guidelines.'

At the close of World War II, in Nuremberg, Germany, certain leaders of the Nazi regime were put on trial for having committed 'crimes against humanity.' During the course of one of these trials, the so-called doctors' trial, the first international code of research ethics, the Nuremberg Code, was written. Subsequent international codes were designed to improve on or to interpret the efforts of their predecessors. WMA began its project in 1953 because it recognized '… a need for professional guidelines designed by physicians for physicians (as opposed to the Nuremberg Code, which was formed by jurists for use in a legal trial).' The result of this project was its code of research ethics which was adopted at the World Medical Assembly in Helsinki, Finland in 1964. CIOMS introduced its Guidelines in 1982 by stating that they were designed to provide guidance for the correct application of the principles of Helsinki, particularly when the research is initiated by researchers and sponsors in technologically developed countries and carried out in developing countries. However, in the revised Guidelines, published in 1993, there are several substantial departures from Helsinki's requirements. At the time of this writing, CIOMS Guidelines are once again being revised. (For further discussion of the subject matter of this paragraph, see Levine 1996.)

1. *Ethical Universalism or Pluralism?*

Discussions of international codes of ethics generally reflect a tension between ethical universalism and cultural pluralism. Universalists hold that ethical principles are identical in all human societies and in all historical periods. It is as if ethical principles exist in nature and are waiting to be discovered. The reason that they appear to change from time to time is that our ability to discern and understand them changes (one hopes for the better). The pluralistic position, by contrast, is that ethics are invented (not discovered) in the course of conversations conducted within particular societies; the results will necessarily reflect the histories and traditions of the cultures in which the conversations are conducted (Levine 1996).

The writers of Nuremberg and Helsinki each attempted to create universal documents. By contrast, the CIOMS Guidelines, which although they recognize some ethical principles as 'transcendent,' also acknowledge the validity of some degree of cultural pluralism. The writers of the CIOMS Guidelines aspire to 'global applicability,' a document that provides adequate ethical guidance for all known contemporary

societies with the understanding that revisions will be required from time to time as ethics change in response to changing circumstances and understandings (Levine 1996).

International codes of research ethics have no legal authority except to the extent that parts or all of them are incorporated in the laws of nations or other political entities having the authority to enact and enforce laws. This point notwithstanding, these documents have important moral status in that those who wish to criticize the ethics of various research activities generally rely on one or more of these documents to ground their judgments.

2. *The Content of the Ethical Codes*

Three ethical principles generally are recognized as 'basic' or 'fundamental' in the ethics of biomedical research and practice. These principles, which were first articulated in their most widely used form by the US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (National Commission), are 'respect for persons,' 'beneficence,' and 'justice.' The first—and to this date, only—explicit reference to these principles in the international codes is in the 1993 revision of the CIOMS Guidelines. However, the National Commission found in Nuremberg and Helsinki implicit statements of respect for persons and beneficence (National Commission 1978, Levine 1988); concern for justice is also implied in the introduction to Nuremberg. Whether or not these principles are universally valid is the topic of a controversy beyond the scope of this article (Levine 1996, Macklin 1992).

The purpose of the ethical codes is to provide guidance on how research involving human subjects should be carried out so that it is responsive to the requirements of the basic ethical principles. Because they are considered relevant to virtually all fields of human behavior, the principles themselves are stated at such a level of abstraction that they do not provide clear direction on what behaviors are required or forbidden. Ethical norms (or rules) are incorporated in ethical codes to clarify what behaviors are prescribed or proscribed by the document. Ethical norms are either substantive or procedural. A substantive norm specifies behaviors that are required (or forbidden) because they are morally right (or wrong). Some procedural norms specify procedures that should be performed to determine the most appropriate behavior when the behavior has not been specified by a substantive norm. Other procedural norms are designed to provide documentary evidence that research has been conducted in accord with the relevant substantive norms (Levine 1988).

An illustration. The principle of respect for persons requires that human persons must never be treated only as means to the ends of others. The substantive norm that requires informed consent is a specification of one way in which this principle is to be made operational in the conduct of research involving human subjects. The procedural norm that requires review and approval by an institutional review board affords a method for determining what specific bits of information must be divulged to prospective subjects in a particular research protocol. Another procedural norm that requires the signing of a consent form provides documentary evidence that the behavior required by the substantive norm has been accomplished.

Ethical codes written by ethical universalists tend to have a high proportion of substantive norms. Those that espouse cultural pluralism tend to have a larger proportion of procedural norms. Substantive norms predominate in Nuremberg and Helsinki while the CIOMS Guidelines contain a relatively large number of procedural norms along with abundant commentary on how the norms should be interpreted in various specific circumstances.

All of the substantive norms articulated in the leading international codes of ethics are variations on and elaborations of seven general normative statements (Levine 1988). The first three of these are most closely related to the principle of beneficence which requires that those who perform research should maximize possible benefits and minimize possible harms and, further, that they should not deliberately injure any human being.

(a) There should be good research design. This norm is linked inextricably to the purpose of doing research—to contribute to the development of generalizeable new knowledge. If the research is not designed adequately, there is little or no chance that it will make the desired contribution to the development of knowledge. As stated in the CIOMS Guidelines (1993, p. 38), '… scientifically unsound research is *ipso facto* unethical in that it may expose subjects to risk or inconvenience to no purpose.'

(b) There should be competent investigators. One purpose of this norm is identical to that of the first; competent investigators are much more likely than incompetents to succeed in producing beneficial new knowledge. In addition, there must be at least one member of the team of researchers who has the clinical competence to monitor the health of the subjects with a view toward prompt intervention as indicated to preserve the health of the subject by (e.g.) ending the individual's participation in the research or by administration of an antidote to some toxic substance. As stated in Helsinki (Article 15): 'Medical research involving human subjects should be conducted only by scientifically qualified persons and under the supervision of a clinically competent medical person. The responsibility for the human subject must always rest with a medically qualified person …'

(c) There should be a favorable (or reasonable) balance of risks and benefits. This is a judgment that

must be made by the investigator and affirmed by the research ethics committee before the prospective subject is presented with an invitation to participate. The risks of interventions and procedures that hold out the prospect of direct benefit to the subject are justified much the same as they are in the conduct of medical practice—in terms of the anticipated benefit to the subject. The risks and benefits of interventions and procedures that do not hold out the prospect of direct benefit for the individual subject (nonbeneficial procedures) must be justified in terms of the anticipated benefit to society. Nuremberg (Principle 6) requires that: 'The degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved.' Plans to employ nonbeneficial procedures that present more than minimal risk in vulnerable populations require special justification. If in the course of conducting the research, the balance of risks and benefits ceases to be favorable, the investigator is required to intervene appropriately. Nuremberg requires: 'During the course of the experiment the scientist … must be prepared to terminate the experiment … if he has probable cause to believe … that a continuation … is likely to result in injury, disability or death … .' (Principle 10) Helsinki does not call for separate analysis of the risks and benefits of individual components of research as described above. Rather, it calls for the evaluation of entire research protocol as either beneficial or nonbeneficial. 'The physician may combine medical research with medical care, only to the extent that the research is justified by its potential prophylactic, diagnostic or therapeutic value.' (Article 28) This is a remnant of Helsinki's historical distinction between therapeutic and nontherapeutic research; reliance on that distinction undermines the validity of the entire document (for further discussion of the problems created by this illogical distinction, see Levine 1999).

The next two of the general substantive norms are most related to the principle of respect for persons:

(d) There should be informed consent. The first principle of Nuremberg set forth the definition of informed consent that, with minor variations, has been incorporated in all subsequent codes and regulations.

> The voluntary consent of the human subject is absolutely essential. This means that the person involved should have legal capacity to give consent; should be so situated as to be able to exercise free power of choice, without the intervention of any element of force, fraud, deceit, duress, overreaching, or other ulterior form of constraint or coercion; and should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision. This latter element requires that before the acceptance of an affirmative decision by the experimental subject there should be made known to him the nature, duration, and purpose of the experiment; the method and means by which it is to be conducted; all inconveniences and hazards reasonably to be expected; and the effects upon his health or person which may possibly come from his participation in the experiment.

(Principle 1) In this way, the prospective subject is invited to accept the researcher's goals as if they were his or her own at least for the duration of his or her participation. The result is that the subject is not treated as *merely* a means to the researcher's goals.

(e) There should be respect for privacy and maintenance of confidentiality. 'Privacy' and 'confidentiality' are often incorrectly used as synonyms. 'Privacy' refers to the 'freedom of the individual to pick and choose for himself the time and circumstances under which, and most importantly, the extent to which his attitudes, beliefs, behavior and opinions are to be shared with or withheld from others' (Kelman 1977). 'Confidentiality' refers to a mode of management of private information; if a subject shares private information with (confides in) an investigator, the investigator is expected to refrain from sharing this information with others without the subject's authorization or some other justification. CIOMS Guidelines require:

> 'The investigator must establish secure safeguards of the confidentiality of research data. Subjects should be told of the limits to the investigators' ability to safeguard confidentiality and of the anticipated consequences of breaches of confidentiality' (Guideline #12).

The final two general substantive norms are most closely related to the principle of justice which requires fairness in the distribution of both the burdens and benefits of research.

(f) Selection of subjects should be equitable. From the 1940s through the mid-1980s, the prevailing attitude among commentators on and policy makers in the field of research ethics was that participation in research was dangerous to the subjects and that researchers were likely to exploit the socially and economically less powerful individuals who were most likely to serve as research subjects. For this reason, the ethical codes and regulations of that time focused on the protection of the prospective subject from bearing an unfair share of the burdens of research. The criterion for determining whether an individual (or a class of individuals) was in need of 'special protections' was 'vulnerability.' In the 1970s, vulnerability tended to be equated with 'limited capacity to give informed consent' and special procedural and substantive norms were developed to protect the rights and welfare of children, prisoners, the institutionalized mentally infirm, and fetuses (Levine 1998). The category 'vulnerable' has been expanded subsequently to include others including 'subordinate members of hierarchical groups, persons with limited financial

resources and persons living in developing countries or communities' (CIOMS 1993, p. 30). Since the mid-1980s, research participation increasingly has been recognized as beneficial to both individuals and the class of persons they represented. This change began as a response to the AIDS activists who in 1986 pointed out that participation in a placebo-controlled clinical trial of AZT was not to be considered burdensome; it was, they argued, beneficial in that it was the only way people could have access to the only drug that anyone believed might be directed at the cause of their disease. Since then, ethical codes and regulations have begun to interpret the requirement for equitable selection of subjects as requiring also that vulnerable persons be assured of receiving a fair share of the benefits of participation in research.

(g) There should be compensation for research-induced injury. CIOMS Guidelines require: 'Research subjects who suffer physical injury as a result of their participation are entitled to such financial or other assistance as would compensate them equitably for any temporary or permanent impairment or disability.' They note, however, that 'In some societies the right to compensation is not acknowledged' (CIOMS 1993, pp. 35–6). One such society is the United States.

3. *Exploitation of Subjects in Developing Countries*

In the past, industrial sponsors and investigators from technologically developed countries occasionally conducted research designed to evaluate their therapeutic innovations in developing countries. This was done to lower the costs of the research and to take advantage of the less complex and sophisticated regulatory systems typical of developing countries. This was unfair in that these innovations, once proved safe and effective, were generally not available to the residents of developing countries owing to economic factors.

With the aim of preventing such exploitation CIOMS Guidelines require special justification for conducting research in developing countries (e.g., the research must be responsive to the health needs and priorities of the 'host' country) and special obligations for the sponsors (e.g., they must engage in capacity-building in the host country to enable it to conduct independent ethical and scientific review). The 1993 version of the CIOMS Guidelines requires that any product demonstrated safe and effective in such research should be made 'reasonably available' to the residents of the country at the conclusion of the research.

Helsinki requires that a new therapy it must be '… tested against those of the best current prophylactic, diagnostic, and therapeutic methods' (Article 29). This standard, which is designed to prevent exploitation of all patients including those residing in developing countries, may also have the undesired effect of obstructing the development for use in developing countries inexpensive alternatives to expensive therapies employed in developed countries. The latter are generally the 'best current … therapeutic methods.' An alternative standard which would not have this disadvantage currently is being considered by CIOMS: 'the highest attainable and sustainable' standard (Levine 1999, 2000). 'Highest attainable' means that one should provide the best one can offer under the circumstances of the research rather than settle for the standard prevailing in the host country (often no therapy at all). The upper limit of what should be provided is defined by 'sustainability.' It would be unethical to develop a therapy that could not be made available to residents of the host country (sustained) after the research has been completed. Such research would not have been responsive to the health needs and priorities of the host country.

See also: Bioethics: Examples from the Life Sciences; Bioethics: Philosophical Aspects; Ethical Dilemmas: Research and Treatment Priorities; Pecuniary Issues in Medical Services: Ethical Aspects; Privacy of Individuals in Social Research: Confidentiality; Reproductive Medicine: Ethical Aspects; Research Conduct: Ethical Codes; Research Ethics: Research; Research Subjects, Informed and Implied Consent of

Bibliography

Bankowski Z, Levine R J (eds.) 1993 CIOMS Round Table Conference: Ethics and Research an Human Subjects: International Guidelines, Proceedings of the XXVIth CIOMS Conference, Geneva. CIOMS, Geneva

Kelman H C 1977 Privacy and research with human beings. *Journal of Social Issues* **33**: 169–99

Levine R J 1988 *Ethics and Regulation of Clinical Research*, 2nd edn. Urban & Schwarzenberg, Baltimore, MD

Levine R J 1996 International codes and guidelines for research ethics: A critical appraisal. In: Vanderpool H Y (ed.) *The Ethics of Research Involving Human Subjects: Facing the 21st Century*. University Publishing Group, Frederick, MD, pp. 235–59

Levine R J 1999 The need to revise the Declaration of Helsinki. *New England Journal of Medicine* **341**: 531–4

Levine R J 2000 Global issues in clinical trials. In: Osterweis M, Holmes D E (eds.) *Global Dimensions of Domestic Health Issues*. Association of Academic Health Centers, Washington, DC, pp. 119–30

Macklin R 1992 Universality of the Nuremberg code. In: Annas G J, Grodin M A (eds.) *The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation*. Oxford University Press, New York, pp. 240–76

United States National Commission for the Protection of

Human Subjects of Biomedical and Behavioral Research 1978 *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Washington, DC, DHEW Publication No. (OS) 78-0012

World Medical Association *Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects*. Adopted by the 18th WMA General Assembly Helsinki, Finland, June 1964 and amended by the 29th WMA General Assembly, Tokyo, Japan, October 1975; 35th WMA General Assembly, Venice, Italy, October 1983; 41st WMA General Assembly, Hong Kong, September 1989; 48th WMA General Assembly, Somerset West, Republic of South Africa, October 1996; and the 52nd WMA General Assembly, Edinburgh, UK, October 2000

R. J. Levine

Ethics in Organizations, Psychology of

The concern over questions of ethics in organizations represents a new area of research. Its focus is the ethically justifiable moral responsibility for the structures and processes in organizations. This article will first attempt to elucidate what is meant by the concepts of organization, morality, and ethics in order to arrive at a definition of ethics in organizations. Contractual theory forms the link between ethics and organizations. It, however, requires an underpinning in fundamental ethical theories. Then there will be a brief discussion of exemplary instances concerning ethics in organizations. Finally, a description of the different forms taken by the enforcement and implementation of ethical standards in organizations will be given.

1. *The Concept of the Organization*

Organizations can be understood as legal entities for which, on the basis of a contract, people work (Coleman 1990, see *Organization: Overview*). Organizations act as legal entities through the actions of their employees. For this reason, organizations can also be characterized as secondary agents. In addition to having voluntary members who work for it, an organization may also have compulsory members. Such compulsory membership may be based upon national laws and regulations, as in the case, of students in compulsory schools, prison inmates, military inductees, forced laborers, or the inmates of psychiatric institutions. The discretion of members of an organization as representatives of the organization (i.e., those in a superior position, such as management or professional staff) over other members of the organization (i.e., subordinates) are created on the basis of contractual and legal relationships. The relationship between a person and an organization has three aspects: (a) people act towards the organization as a legal entity, for example, when a person enters into a contract of employment with the organization; (b) by virtue of this contract, people take on a specific role in the organization and then act for the organization, for example, as a 'superior'; and (c) in the context of the organization people also act for themselves, for example, when they try to prevent the promotion of a colleague in order to be promoted themselves.

2. *Morality and Ethics in Organizations*

Moral judgments and actions (see *Religion: Morality and Social Control*) are permanent components of the activities of individuals in organizations. One's own actions and the actions of others are judged as good and bad, just and unjust, fair and unfair. This can be accompanied by specific emotions such as outrage, embarrassment, or regret. In addition, people discuss what would have been a morally appropriate action in a given situation. While doing so, they attempt to give grounds for their conclusions (Gethmann 1989). At times, however, moral utterances are used as weapons in order, for example, to give others a bad conscience, to exclude others from access to particular resources, or to mobilize others to support one's own cause. Thus, whereas morality is part of our everyday practice, ethics is concerned with the conceptual analysis of moral linguistic expressions (meta-ethics) as well as with the rational justification of norms on the basis of ethical theories (normative ethics). Ethics is thus a considered reflection on morality. Rationally justified moral demands and norms can be characterized as moral and ethical demands.

One important meta-ethical question for ethics in organizations is the problem as to whether it would make sense to ascribe to organizations not only a juridical, but also a moral and ethical responsibility (Curtler 1986). Speaking in favor of a morally and ethically justifiable responsibility of organizations is the fact that organizations can be understood as agents, and that the ends of these agents are not in every case reducible to the ends of the members of the organization. The question, however, is whether one should start from the assumption that organizations are to be understood as ends in themselves rather than as the means to the ends pursued by individual people. However, speaking against the view that organizations have an ethically justifiable moral responsibility is the fact that organizations have neither feelings of pain, nor a sensitivity for the suffering of others, empathy, or a conscience. Lenk and Maring (1992) come to the conclusion that people cannot exculpate themselves morally and ethically simply by pointing to the responsibility of the organization. Alternatively, the proof of the morally and ethically unacceptable actions of individual people does not automatically mean an exculpation of the organization.

The problems of normative ethics in organizations are the following: What objectives may organizations

legitimately pursue vis-à-vis the members of the organization? What objectives may people in the organizations pursue? What conditions must organizational rules satisfy to be ethically acceptable? What should the interactions between people in organizations and between people and the organization be for them not to violate any ethical rules? Ethical problems such as these can occur in both profit and nonprofit organizations.

3. *Normative Ethical Theories*

Organizations are based on contracts. These contracts are morally and ethically permissible as long as they satisfy the following conditions (Garrett and Klonoski 1986, pp. 88–91): (a) the contract does not come about as the result of coercion, but rather is based upon the voluntary consent of all contracting parties; (b) the contracting parties know the terms of the contract; (c) no side may make intentionally false statements vis-à-vis the other parties about the underlying circumstances of the contract or consciously fail to disclose relevant facts to them; and (d) the contract cannot obligate the contracting parties to carry out morally and ethically unacceptable actions. From an ethical perspective, people should not enter into any contracts with organizations which could obligate them to carry out morally and ethically unacceptable actions. Since organizations can act only through the actions of people, it should also be possible in this way to rule out morally and ethically unacceptable actions by organizations. In practice, one of the difficult problems is the question of determining how dissimilar negotiating positions should be viewed; the side with the weaker negotiating position may face coercion to enter into the contract on the conditions of the stronger side.

However, theoretical contractual considerations are not fundamental from an ethical point of view, for they presuppose that there are criteria available by means of which one can distinguish between morally and ethically unacceptable actions and morally and ethically neutral actions. In addition, the justification for complying with contracts can also not be based solely on a contract; for, if the foundation of contracts were only personal interests, there would be no reason not to breach a contract at precisely the moment when one's personal interests could be better served by breaching the contract.

The focus of utilitarianism (see Glover 1990, see *Utilitarianism: Contemporary Applications*) is on general good. Utilitarianism postulates that in a choice from among various possible courses of action, the moral and ethical courses of action are those whose empirical consequences result in the greatest good of all parties involved. Accordingly, it does, in general, make good sense to comply with contracts. If, however, in a particular situation it would be better for the overall good of all of the parties involved not to comply with a contract, one has, in the view of utilitarianism, an ethical duty to breach the contract. Thus, utilitarianism cannot underpin ethically contractual demands in organizations. Nevertheless, for behavior that is not covered by contracts, utilitarianism can tell people how they should conduct themselves toward others in organizations.

The respecting of basic rights (see *Human Rights, History of*) as an unconditional moral and ethical duty is the object of the so-called categorical imperative of Immanuel Kant (1964, p. 96), who demanded that one never treat oneself and others only as a means, but always also as an end in itself. This is expressed in the demand for the respect of a person's dignity, freedom, and right to life. The other basic rights, such as, for example, the right to privacy, property, freedom of conscience, or the right to freedom of expression can be derived from these fundamental human rights. One criticism expressed about Kant is that he merely postulates the categorical imperative as a 'fact of reason' without providing any further grounds for it.

Discourse ethics (Habermas 1990, Apel 1999) takes as its point of departure the assumption that only a rational clarification of conflicts of interests can be an ethical clarification of conflicts of interests. According to discourse ethics, a rational clarification occurs whenever, through the exchange of sound arguments, an agreement is reached which can be consented to by all affected parties without coercion. Thus it is important to create conditions which enable the people involved to enter into a debate concerning the norms in dispute. Since conflicts can only be clarified rationally through debate between the affected parties, the preconditions for debate also constitute the basis of ethics. An important ethically relevant precondition for debating is that those debating may not exclude other people able to act and speak on their own account from debating with them, for consent to a debate is rational when it is based solely upon the recognition of the argument's better grounds. And it is from these preconditions that the so-called principle of universalization of discourse ethics follows. It states that a norm is only just if and when all affected parties are able to accept the consequences and side effects expected to arise from the universal compliance with the norm for the satisfaction of the interests of each individual. Any skeptic wishing to give reasons for the ethical thesis that there are no universally binding rational norms would have to begin to argue. In doing so the skeptic has in fact already become involved in a rational argument. If, however, he or she has become involved in a rational argument, the skeptic has thus already accepted the preconditions of argument and the principle of universalization it implies.

Thus, discourse ethics is, in contrast to utilitarianism or Kant's categorical imperative, rationally grounded. If one follows discourse ethics, a person is ethically bound by contracts only as long as, and to the

extent that, it can be assumed that all contracting parties fulfill their commitments. If a party does not fulfill its commitments and if the compliance with the contract is thus attended with unreasonable disadvantages for the other parties, then the ethical duty to adhere to the contract is also suspended.

4. Exemplary Instances

In this section a rough outline will be given of some exemplary instances concerning ethics in organizations.

To regulate admittance, an organization frequently applies a procedure for the selection of personnel. These procedures must not violate the applicants's right to rational self-determination and freedom from emotional and social harm. To ensure transparency, before the selection procedure is carried out, applicants should be informed of its objectives, the approach, and the possible consequences. The selection process is to be conducted tactfully and respectfully. The diagnostician is obligated not to disclose certain facts to third parties. This is to protect applicants from potential social stigmatization. The question of the justice of selection strategies is posed by the problem of whether people are hired merely as the result of how well they do in a diagnostic procedure. Should other criteria be decisive factors, the consequence would be a deviation from the allocation of the most qualified applicants to positions. Such a strategy could in the long run, however, have a detrimental effect on the common weal.

Within the organization employees are subordinated to the instructions of a manager on the basis of the employment agreement. They thus give up a part of their autonomy, trusting that the superior will neither misuse it for the benefit of the organization nor for his or her own personal benefit. From this there ensues an ethical and moral responsibility of the superior for the care and welfare of his or her employees. In the view of discourse ethics, the superior must implement this duty so that it becomes progressively unnecessary. This means that he should transparently organize the conditions within his or her area of responsibility and should manage the employees in such a way, that in the framework of their actions in the organization, they can win back, piece by piece, their self-responsibility.

Putting pressure on someone to act against conscience reduces a person to the status of a pure means for the predetermined ends of others and in so doing violates the basic moral and ethical right to rational self-determination. If the superior becomes convinced that a member of the organization has, on the basis of honorable motives, refused an activity for reasons of conscience and has seriously examined his or her matter of concern, the manager should make every effort to entrust that member of the organization with another activity.

Micropolitics is understood as the secret machinations by which individual members of organizations, or special groups in organizations, act for the attainment of private advantages (Velasquez et al. 1983). From the viewpoint of moral and ethical basic rights, the use of micropolitical tactics are to be rejected because the possibilities of another person achieving rational self-determination are restricted by means of tricks and deceit. A consequence of the discourse ethics view is the demand that people in organizations should collaborate on creating and preserving structures which make possible the open resolution of conflicts.

The following criteria have been postulated for the creation of quality working conditions in organizations (Levine et al. 1984, Luczak et al. 1989, p. 58): (a) at work people should find working conditions that are feasible, tolerable, unrestricted, and not dangerous; (b) in terms of job content, the work assignment, working environment, and pay should meet standards of social adequacy; (c) in addition to this, work should be accompanied by a certain amount of latitude in terms of activities and scope of operations; (d) the work should provide an opportunity to acquire skills; (e) it should give the worker satisfaction and preserve and further develop their personality, i.e., their self-confidence and willingness to take on responsibility; (f) the workers should, in addition, have the possibility of participating in the organization of their working conditions; and (g) finally, the effects of the working conditions should be socially compatible (e.g. pro-family). Apart from this, demands are made for environmentally friendly working conditions.

As many case studies prove, safety insufficiencies and risks are frequently well-known in factories, planning offices, and laboratories long before there are deaths and disasters. Often the members of the organization are aware of such insufficiencies. If they do not get any response to their warnings in the organization, some contribute to the protection of the public by making these risks public (James 1989).

The due-process doctrine (Carroll 1992, pp. 421–8) concerning the termination of employment, demands that discharges ought to take place in accordance with codified regulations. Thus, notice of termination should not be given arbitrarily, but only when certain facts can be shown to be the case. The affected parties are to be informed of the grounds for the termination. They should have the opportunity to comment on these grounds and thus effect, if appropriate, the rescinding of the termination.

5. Forms of Enforcement and Implementation

There are starting points available for rendering morally and ethically acceptable conduct practically possible by and in organizations at the macro, intermediate, and micro levels.

At the macro level of many national constitutions, government agencies are committed to the protection of fundamental human rights or to compliance with the principle of equal treatment. It is, however, disputed in many countries whether, for the guaranteeing of these principles, the state or the administration of justice ought also to intervene or be permitted to intervene legally in the relationship of private organizations and members of organizations. In addition, a problem of substantive law is the extent to which it is not just promulgated, but actually applied as well.

At the intermediate level, independent unions and a right to strike guaranteed by the government have played an important role. Unions, however, frequently do not only pursue moral and ethical objectives, but above all the particular interests of their members. Professional associations with a high public reputation, such as the professional associations of scientists, engineers, physicians, architects, or psychologists (Lindsay 1996), can also influence practices at the intermediate level by means of moral and ethical conduct guidelines for their members in organizations (see *Ethical Codes, Professional: Business Codes*). Associations in which organizations have joined together can likewise introduce moral and ethical standards on the basis of a voluntary self-commitment on the part of the organizations. The motivation behind the introduction of standards, however, is frequently not only the facilitation of morally and ethically acceptable conduct in organizations, but also the improvement of reputation in the public eye as well as the avoidance of statutory regulations. Although such self-interested motives are legitimate, they must not result in a biasing of the standards.

On the micro level, the public denunciation of organizations on the basis of unethical practices can be an effective means of enforcing moral and ethical standards in organizations. If individuals with a high degree of moral sensitivity have a very strong position in an organization, as is sometimes the case for private entrepreneurs, the result can be that they try to make their moral views binding for the objectives, rules, and processes in the entire organization. Even individual members of an organization without any special position in the organization can trigger changes as conscientious objectors or 'whistle-blowers.'

Regarding the implementation within the organization of external legal and ethical standards, two ideal-typical strategies are currently being discussed, namely the so-called compliance approach and the integrity approach (Paine 1994). In the compliance approach externally predetermined standards are translated into internal active and restrictive regulations, compliance with which is enforced within the organization by means of controls and sanctions. In the integrity approach an effort is made through an organizational development process at all levels of the organization to activate the moral sensitivity of the members of the organization. The danger in the compliance approach is that, due to the strong outside control, members of the organization become morally desensitized and develop a mentality of getting around the guidelines. In the integrity approach, members of an organization cannot invoke written rules to defend themselves against morally and ethically inadmissible expectations. What is important is a sensible combination of the compliance and the integrity approaches.

6. *Summary*

Organizations can be understood as legal entities for which, on the basis of a contract, persons work. Moral judgments and actions are permanent components of the activities of individuals in organizations. Ethics is a considered reflection on morality. Ethics in organizations deals with the ethically justifiable moral responsibility for the structures and processes in organizations. People cannot exculpate themselves morally and ethically simply by pointing to the responsibility of the organization. However, the proof of the morally and ethically unacceptable actions of individual people does not automatically mean an exculpation of the organization. Discourse ethics can underpin the ethical commitment to transparent contracts entered into voluntarily. In addition, the basic-rights approach provides concrete examples and discourse ethics provides formal rules for the determination of ethically unacceptable obligations to which one should not commit oneself in a contract. Exemplary instances were discussed, namely ethical guidelines for personnel selection, ethical problems in the superior–subordinate relationship, the freedom of conscience in organizations, the ethical problematic of micropolitics in organizations, the right to make public abuses in organizations (whistle blowing), as well as ethical rules for dismissal from organizations. Different forms taken by the enforcement of ethical standards in organizations at the macro, intermediate, and micro levels were presented. Finally, a sensible combination of the compliance and the integrity approaches regarding the implementation of moral and ethical standards in organizations was proposed.

See also: Economics and Ethics; Ethical Codes, Professional: Business Codes; Human Rights, History of; Organization: Overview; Religion: Morality and Social Control; Utilitarianism: Contemporary Applications

Bibliography

Apel K O 1999 The problem of justice in a multicultural society. The responses of discourse ethics. In: Kearney R, Dooley M (eds.) *Ethics in Question*. Routledge, London

Carroll A B 1993 *Business & Society*, 2nd edn. South Western Publishing, Cincinnati, OH

Coleman J S 1990 *Foundations of Social Theory*. Belknap, Cambridge, MA
Curtler H (ed.) 1986 *Shame, Responsibility, and the Corporation*. Haven, New York
Garrett T M, Klonoski R J 1986 *Business Ethics*, 2nd edn. Prentice Hall, Englewood Cliffs, NJ
Gethmann C F 1989 Protoethics: Towards a formal pragmatics of justificatory discourse. In: Butts R E, Brown J R (eds.) *Constructivism and Science*. Kluwer, Dordrecht, The Netherlands
Glover J (ed.) 1990 *Utilitarianism and Its Critics*. Macmillan, New York
Habermas J 1990 Discourse Ethics—notes on a program of philosophical justification. In: Habermas J (ed.) *Moral Consciousness and Communicative Action*. MIT Press, Cambridge, MA
James G G 1989 In defense of whistle blowing. In: Shaw W H, Barry V (eds.) *Moral Issues in Business*, 4th edn. Wadsworth, Belmont, CA
Kant I 1964 [trans. Paton H J] *Groundwork of the Metaphysics of Morals*. Harper and Row, New York
Lenk H, Maring M 1992 Ecology and ethics: Notes about technology and economic consequences. In: Ferré F (ed.), *Research in Philosophy and Technology. Vol. 12: Technology and the Environment*. JAI Press, Greenwich, CT
Levine M F, Taylor J C, Davis L E 1984 Defining quality of working life. *Human Relations* **37**: 81–104
Lindsay G 1996 Psychology as an ethical discipline and profession. *European Journal of Psychology* **1**: 79–88
Luczak H, Volpert W, Raeitel A, Schwier W 1989 *Arbeitswissenschaft* [*Ergonomics*], 3rd edn. RKW Verlag, Cologne, Germany
Paine L S 1994 Managing for organizational integrity. *Harvard Business Review* **72**: 106–17
Velasquez M, Moberg D J, Cavanagh G F 1983 Organizational statesmanship and dirty politics: Ethical guidelines for the organizational politician. *Organizational Dynamics* **12**: 65–80

G. Blickle

Ethnic and Racial Social Movements

Ethnic and racial social movements (hereafter E/R social movements) are goal-directed actions sustained by groups whose identity is recognized by the presence of racial or ethnic markers. These markers typically include skin pigmentation, ancestry, language, and history of discrimination, conquest, or other shared experience. For simplicity (and to avoid invoking unscientific assumptions about race), many researchers prefer the label *ethnic mobilization*. Ethnic mobilization includes activities that range from small-scale, sporadic events to well-developed campaigns and/or civil war.

Several conceptual distinctions have facilitated the understanding of the emergence, growth, and decay of E/R social movements. Some (but not all) scholars distinguish *ethnic mobilization* from *ethnic solidarity*. Solidarity is characterized as the conscious identification (and loyalty) with a particular race or ethnic population, measured by attitudes or organizational involvement. This distinction has allowed Minkoff (1995) to analyze the effects of the growth of racial solidarity organizations on the rate of racial protest activity and vice versa.

Social movement theories distinguish various forms of movements by their duration, target, tactics, violence, and audience. These distinctions yield three broad categories of E/R movements: (a) territorial sovereignty movements that demand regional autonomy, separatism, diaspora settlements, and/or secession; (b) protests that demand expansion of a group's civil and economic rights or demand an end to discrimination; (c) collective attacks ranging from genocide, ethnic cleansing, and mob violence, to symbolic threats. Studies of E/R movements traditionally analyze these forms separately (Horowitz 1975, Banton 1983). For example, race riots are distinguished from other types of ethnic conflict by expressions of violence that might involve hundreds or thousands of persons in activity that lasts several hours or more (Spilerman 1970, Myers 1997). Protests for civil rights typically include marches, demonstrations, sit-ins, and other similar tactical forms (McAdam 1982). To the extent that E/R social movements seek to eradicate and replace existing geographical and administrative state boundaries, they provoke violence on the part of insurgents and state authorities. This is particularly true for separatist or secession movements demanding formal withdrawal from a central political authority (or, in the case of irredentist/diaspora movements, claiming territorial sovereignty for formerly dispersed populations) (Hechter 1992).

Most studies focus on one of these forms of E/R social movement within a single country. Other researchers have used comparative research designs to test the emergence of various forms in different settings (Gurr 1993). These designs allow researchers to pursue questions about the diffusion of E/R social movements across national boundaries.

Two key questions drive research efforts in this area. First, how does ethnic identity become transformed into social movements? Second, what explains the emergence and persistence of E/R movements?

1. *The Transformation of Group Identity into Social Movements*

To transform subjective group identity into goal-directed ethnic mobilization, a group must overcome the seemingly natural resistance of individuals to sustain collective action. By invoking norms of group loyalty, social sanctions sometimes transform inert groups into meaningful and purposive social action. Consequently, members are continuously constructing group boundaries (Barth 1969, Hechter 1987a). Social

movements tend to develop when these boundaries are challenged and group identities are defended.

When does sporadic ethnic activity become a social movement? Answering this question is difficult because no thresholds have been established in the empirical literature. In contrast to isolated ethnic events, these social movements involve sustained efforts to evoke major changes in political goals, such as policies regarding subordinate groups, redefinition of citizenship boundaries or rights, and/or efforts to exclude, subjugate, or control a population or region. Thus, social movements tend to be more enduring and continuous than are other forms of ethnic mobilization.

Historical evidence suggests that nation-building activities have often provoked enduring E/R social movements. For example, Smith (1979, p. 34) identifies a sequence of events that encourage separatism, beginning with initial state-building processes including creation of a centralized bureaucracy and diffusion of national educational institutions. A related line of research has emphasized the role of elite mobility into leading institutions also shaping E/R movements. If ethnic elites find their mobility blocked, ethnic mobilization around claims of minority discrimination will arise (Williams 1994).

An extreme form of violent social movements against a target population is genocide or ethnic cleansing. Historical examples often include claims of ethnic or racial purity that required exclusion or extermination of some other group. Most recently, this form of violence has arisen as former states (or modern empires) fragment or dissolve entirely and attempt to forge new and ethnically homogeneous identities (Jalali and Lipset 1992). The consequence is often a combination of pogroms, terrorist movements, disenfranchisement, and other methods of physical attack such as lynchings, rape, or civil war.

2. *The Dynamics of Movements: Emergence, Growth, and Decay*

Several prominent theoretical traditions have been offered to explain the emergence, growth, and decay of E/R social movements. Each is distinguished by an emphasis on one or more processes of changing economic, political, and/or ideological environments as key features shaping the trajectory of these movements.

2.1 Economic Factors

2.1.1 Internal colonialism theory. Internal colonialism theory suggests that a combination of uneven industrialization and cultural differences among regions in core nations cause ethnic grievances to become the basis of enduring political contention. In this view, the sources of ethnic solidarity include uneven regional development that reinforces or creates inequality, dependence on external or international investment, and an occupational structure that is highly segregated along ethnic lines. Furthermore, according to this argument, a high level of ethnic solidarity and a division of labor segmented along ethnic lines provokes ethnic conflict in developed regions, rather than in impoverished areas (Hechter 1975).

2.1.2 Competition theory. Some counterevidence suggests that states experiencing a decline in ethnic inequality experience movements expressing claims for autonomy and political rights for minorities, while patterns of ethnic violence occurs in states that exclude ethnic minorities from the political process (Smith 1979, p. 35). Competition theories of race and ethnic relations offer an explanation for these findings. According to competition theory, declining inequality among regions (or groups) promotes competitive conflict among race and ethnic groups (Olzak and Nagel 1986). This is because declining inequality and intergroup contact release forces of competitive exclusion and conflict (Barth 1969). In this view, E/R social movements result from conditions of niche overlap (rather than from niche segregation). For example, competition theorists argue that ethnic conflict arises when ethnic groups within nations come to compete in the same labor markets and increase their access to similar sets of political, economic, and social resources (Nielsen 1985, Olzak 1992). Split labor market theories further suggest that ethnic conflict intensifies when groups command different levels of wages and conflict becomes divided along both economic and ethnic lines (Bonacich 1972). Mobilization based upon race or ethnic identity occurs as dominant groups attempt to reassert their dominance over newly competing groups, or as formerly disadvantaged ethnic groups challenge the existing power structure.

2.1.3 Rational choice theory. This perspective emphasizes shifts in the calculus of the costs and benefits attached to ethnic mobilization. According to this view modern ethnic movements occur with regularity because they have unique properties that allow them to overcome the free rider problem that encourages nonparticipation. According to this view, because ethnic groups are able to form dense social networks, group solidarity is high, minimizing costs of mobilization. Simultaneously, ethnic groups can efficiently

apply systems of monitoring behavior, insuring loyalty, and sanctioning members (Hechter 1987a). Building on rational choice models, Fearon and Laitin (1996) and Weingast (1998) have linked the strategic aspects of ethnic identity to violence, as elites build on existing ethnic loyalties. Additionally, Bhavhani and Backer (2000) argue that the presence of genocidal norms (that sanction in-group members who decline participation in ethnic mayhem) tends to increase the scale of ethnic violence. One testament to the power of these norms is that the intensity of ethnic killing can remain high, despite a history of intergroup cooperation and trust among groups.

2.2 Political Factors

These perspectives emphasize the role of shifts in political constraints and opportunity structures that influence the trajectory of E/R social movements. Horowitz (1975) examines the centrifugal force of ethnic political parties, which maintain ethnic loyalties through institutional arrangements and patronage based on ethnic loyalties. Some social scientists argue that the decline of authoritarian regimes coincides with the resurgence of ethnic or nationalist movements because the retreat of strong, repressive authorities leaves a power vacuum (Gurr 1993). As the former military and administrative structures recede, local level elites mobilize ethnic loyalties and take advantage of this vacuum. Along similar lines, McAdam (1982) provides evidence from the USA that suggests that shifts in political opportunities (either positive or negative) drove the rates of protest activity during peak periods of civil rights insurgency. State repression may subdue such movements, but this effect is often temporary.

2.3 Cultural/Ideological Factors

2.3.1 Nationalism. All forms of E/R movements articulate demands that invoke one or more cultural themes of nationalism, rights of self-determination, expansion of human rights, and basic rights of sovereignty (Smith 1979, Hechter 1987b). Sovereignty claims usually refer to shared experiences of 'a people,' which can be real or imagined (Anderson 1991). Moreover, the acceptance of the legitimacy of nationalism has parallels in the diffusion of a worldwide human rights movement, which acts as an orienting frame for states and E/R social movements.

2.3.2 Diffusion. Social movement tactics, organizations, and claims have diffused across national borders, creating an international network of E/R social movements that share common themes. Self-and-other definitions based on racial markers, linguistic, or historic patterns of subordination justify and legitimate these claims in virtually all of the world's states. For instance, Olzak and Tsutsui (1999) find that although integration into a world system of power and domination facilitates E/R movements, these effects vary considerably across countries. In addition, processes of diffusion have intensified the international scope of these movements, transforming them into highly contentious international issues.

See also: Civil Rights Movement, The; Ethnic Conflicts; Identity: Social; Minorities; Multiculturalism: Sociological Aspects; Race Identity; Racial Relations; Residential Segregation: Sociological Aspects; Social Movements and Gender; Social Movements, Sociology of

Bibliography

Anderson B 1991 *Imagined Communities*. Verso, London

Banton M 1983 *Racial and Ethnic Competition*. Cambridge University Press, Cambridge, UK

Barth F 1969 *Ethnic Groups and Boundaries*. Sage, Los Angeles

Bhavnani R, Backer D 2000 Localized ethnic conflict and genocide. *Journal of Conflict Resolution* **44**(3): 283–306

Bonacich E 1972 A theory of ethnic antagonism. *American Sociological Review* **37**: 547–59

Fearon J, Laitin D 1996 Explaining interethnic cooperation. *American Political Science Review* **90**: 715–35

Gurr T R 1993 *Minorities at Risk*. Institute of Peace Press, Washington, DC

Gurr T R, Moore W H 1997 Ethnopolitical rebellion. *American Journal of Political Science* **41**: 1079–103

Hechter M 1975 *Internal Colonialism*. University of California Press, Los Angeles

Hechter M 1987a *Principles of Group Solidarity*. University of California Press, Berkeley, CA

Hechter M 1987b Nationalism as group solidarity. *Ethnic and Racial Studies* **10**: 415–26

Hechter M 1992 The dynamics of secession. *Acta Sociologica* **35**: 267–83

Horowitz D L 1975 *Ethnic Groups in Conflict*. University of California Press, Berkeley, CA

Jalali R, Lipset S M 1992 Racial and ethnic conflicts: A global perspective. *Political Science Quarterly* **107**(4): 585–606

McAdam D 1982 *Political Process and The Development of Black Insurgency*. University of Chicago Press, Chicago

Minkoff D 1995 *Organizing for Equality*. Rutgers University Press, New Brunswick, NJ

Myers D 1997 Racial rioting in the 1960s. *American Sociological Review* **62**: 94–112

Nielsen F 1980 The Flemish movement in Belgium after World War II. *American Sociological Review* **45**: 76–94

Nielsen F 1985 Ethnic solidarity in modern societies. *American Sociological Review* **50**: 133–45

Olzak S 1983 Contemporary ethnic mobilization. *Annual Review of Sociology* **9**: 355–74

Olzak S 1992 *The Dynamics of Ethnic Competition and Conflict*. Stanford University Press, Stanford, CA
Olzak S, Nagel J (eds.) 1986 *Competitive Ethnic Relations*. Academic Press, Orlando, FL
Olzak S, Tsutsui K 1999 Status in the world system and ethnic mobilization. *Journal of Conflict Resolution* **42**: 691–720
Smith A 1979 Towards a theory of ethnic separation. *Ethnic and Racial Studies*. **2**: 21–37
Spilerman S 1970 The causes of racial disturbances. *American Sociological Review* **35**: 627–49
Weingast B R 1998 Constructing trust: the politics and economics of ethnic and regional conflict. In: Soktan K, Uslaner E, Uslaner V (eds.) *Institutions and Social Order*. University of Michigan Press, Ann Arbor, MI, pp. 163–200
Williams R 1994 The sociology of ethnic conflicts. *Annual Review of Sociology* **20**: 49–79

S. Olzak

Ethnic Cleansing, History of

'Ethnic cleansing' is a term that has come into common usage only with the recent war in former Yugoslavia. It describes the process of the forced deportation of peoples or nations from their homes and homelands. Even though violence and murder often take place in the course of ethnic cleansing, it is not the same historical phenomenon as genocide, which involves the intent to kill off part or all of a people. For a variety of reasons, some scholars find the term problematic. Nevertheless, it is useful not only for describing the attacks on national groups in former Yugoslavia, but for understanding similar instances of the violent expulsion of nations from their homelands throughout the twentieth century.

1. *The Term 'Ethnic Cleansing'*

The war of Yugoslav succession gave the term 'ethnic cleansing' its currency in the popular press and its meaning for political leaders, academic commentators, and the International Tribunal in the Hague. The term burst into public consciousness in the West in May 1992, and was used with increasing frequency during the following summer to describe Serbian attacks against Bosnian Muslims in Bosnia–Herzegovina (Gutman 1993). In this context, the use of the term 'ethnic cleansing' indicated that Serbs and Bosnian Serbs assaulted the Muslims with the intention of driving them out of targeted territories that the Serbs wanted to make ethnically Serbian. Subsequently, the term was also applied to Serbian attacks on Croats during the 1991–1992 Serbian–Croatian war, Croat attacks against Bosnian Muslims in 1992–1993, and Croat attacks against Serbs in the Krajina during the summer of 1995. All of these attacks were motivated by the desire to seize and secure these territories for their own national groups.

Most recently, the term 'ethnic cleansing' has been applied to the crisis in Kosovo, both in connection with the Serbian offensive against the Kosovar Albanians in the winter and spring of 1998–1999, which saw nearly a million Kosovars driven from their homes, and as a way to describe the Kosovar Albanian attacks on the Serbs since the conclusion of NATO bombing in May 1999. The origins of the term also are related to Serbian–Albanian relations in Kosovo. The term 'ethnic cleansing' was first used by Serbs in the mid-1980s to describe what they claimed were Kosovar Albanian attempts to drive the remaining Serbs of Kosovo from the province by means of economic pressure and physical threats (Vujacic 1993).

From the beginning of its use in the west to describe the events in Yugoslavia, the term was criticized on a number of grounds. Some commentators thought it was little more than a euphemism for genocide (see *Symbolism in Anthropology*). Others were critical of the way genocide and ethnic cleansing were used interchangeably (Hayden 1996). In fact, distinctions between ethnic cleansing and genocide have been usefully drawn to understand the differences between the character of the Serbian attacks on the Bosnian Muslims, in particular, and genocides carried out earlier in the century. The intention to murder all or part of a nation, ethnic, or religious group, is critical to the definition of genocide developed by Rafael Lemkin during World War II and made part of international law in Article II of the United Nations Convention on the Prevention and Punishment of the Crime of Genocide, December 9, 1948 (Melson 1992). The intent of ethnic cleansing is to drive a nation, ethnic, or religious group out of a designated piece of territory, not necessarily to murder them.

The problem remains that ethnic cleansing is not free of genocidal implications. The Serbian massacre of some 7,300 Muslim men in Srebrenica should be characterized as genocide, even if the ultimate goal was to drive the Muslims from the territory. In this connection, the Croatian massacres of Bosnian Muslims in the Lasva Valley in the spring of 1993 were classified by the International Tribunal in the Hague as 'ethnic cleansing,' rather than 'genocide.' Critical was the intent to drive the Muslims from the region rather than the intention of mass murder.

Some scholars identify the mass rape associated with ethnic cleansing in Bosnia as an attack on the reproductive potentialities of the Bosnian Muslim nation and therefore a form of genocide (Allen 1996, Stiglmayer 1994). It should in no way diminish the horrific quality of the crime to suggest that its purpose was fundamentally to drive Muslims from towns and villages seized and claimed by Serbs. It was therefore a part of the ethnic cleansing of the region rather than genocide. The mass rape in Bosnia led the Inter-

national Tribunal in 1998 to classify it, for the first time in its long history, as a 'crime against humanity.'

The word 'cleansing' is sometimes criticized by scholars because it ostensibly sanitizes a process that is otherwise brutal and murderous. But in the Slavic form of the word, for example *ciscenje* (Serbo–Croatian), *oczyszczanie* (Polish), or *chistki* (Russian), cleansing can refer to violent purges and is often used in a political context. The German word for cleansing and purges, *Saeuberung*, was used in pre-Nazi and Nazi eugenics and racial theorizing to indicate the need for 'purifying' the German race. In both Slavic and German, cleansing has a double meaning: the individual has been dirtied by alien elements and must be cleansed as a result, and the nation has been infiltrated by ethnic 'others,' who must be excised from the social body. Violence is implied in both. The definition of who belongs to an ethnic group is clearly determined by subjective criteria (see *Ethnic Groups/ Ethnicity: Historical Aspects*). The existence of the group itself often is a creation of the persecutor.

2. *Ethnic Cleansing and Modernity*

As quickly as 'ethnic cleansing' was accepted in public discourse as an appropriate description for the attacks in former Yugoslavia, scholars began applying the term to similar episodes in the past and present. Some analysts looked at ethnic cleansing as a phenomenon deeply rooted in human society going back to the beginning of history (Bell-Fialkoff 1996). The Assyrians, the Babylonians, and the Greeks are claimed to be the first ethnic cleansers. From this point of view, the causes of ethnic cleansing in Bosnia more often than not are traced to supposed 'ancient hatreds' between the peoples of the region.

A second approach looks at ethnic cleansing, like genocide, as a fundamentally modern phenomenon, one associated with the development of the modern state (see *State, History of*) and the goals of its political leaders (Ahmed 1995, Baumann 1989). Especially the 'high modernist' aspirations of the late nineteenth and twentieth-century state, which seeks to control, homogenize, and manipulate its population, lead to episodes of ethnic cleansing (Scott 1998). The political ideology of the Serbian leadership and its relationship to the control of state power are at the core of the problem in the Balkans, not 'ancient hatreds.' The political manipulation of recent conflicts, like those during World War II in Yugoslavia, set the stage for ethnic cleansing in the 1990s.

The modern state also commands the means to engage in ethnic cleansing. Not only can it register and identify the ethnic other, it uses modern means of communications to initiate and justify its ethnic cleansing campaigns. Propaganda and the mass media are critical to creating the appropriate images of enemies within and justifying forceably deporting the other. If earlier in the century newsreels and radio conveyed the messages of hatred and prejudice, television became critical to the late twentieth-century catastrophes in Yugoslavia and Rwanda. The state also uses modern means of warfare and transport to move populations from their homelands. Machine guns make it easy for small contingents of troops to guard (and sometimes to murder) large numbers of ethnic others. The sealed railway freight car might well serve as a symbol for ethnic cleansing, especially in Europe. Tens of millions of people were forced from their homes in hellish rail deportations; hundreds of thousands died from hunger, thirst, exposure, or disease in the process of transportation alone.

The major episodes of ethnic cleansing in the twentieth-century world reinforce these ideas. As a crucible for 'industrial killing,' World War I set the tone for the rest of the century (Bartov 1996). Yet even during the Balkan Wars of 1912–1913, some half a million people were driven from their homes or fled their persecutors. The Young Turk triumvirate that led the Ottoman Empire during World War I was responsible for the ethnic cleansing and genocide of the Armenians in 1915. Approximately, 750,000 to 1,200,000 Armenians died as a result. The modernizing Turkish leader, Mustapha Kemal, drove the Greeks from Anatolia in 1922–1923, in a clear case of ethnic cleansing. Hundreds of thousands of Greeks died in the process, many in refugee camps in Greece and the islands from hunger and disease. Hitler and the Nazis engaged in a form of ethnic cleansing that quickly degenerated into genocide and the Holocaust (see *Holocaust, The*). The Third Reich arrogated to itself a huge territorial realm (*Lebensraum*) where it would bring in German settlers and forceably expel the allegedly inferior local ethnic inhabitants. The case is complicated by the fact that Nazi rhetoric was consistently eliminationist, while the government's policies until World War II were oriented towards forced deportation and population transfer. Gypsies, Poles, and Russians were involved in Nazi plans of racial destruction, as were the Jews, their primary target.

Stalin and the Soviet leadership engaged in a particularly Bolshevik form of ethnic cleansing (Martin 1998). During World War II alone, nearly a million people from the Northern Caucasus and Crimea—Chechens, Ingush, Karachaevtsy, Balkars, and Crimean Tatars, among others—were driven out their home territories to Central Asia. These 'punished peoples' were slated for elimination as nations, if not necessarily as individuals. Still, so many of them died in the process that their survivors speak of genocide. During and after the war, some 11.5 million Germans were ethnically cleansed from their previous homes in Eastern and East Central Europe, primarily Poland, Czechoslovakia, and Hungary. Roughly two million died or were killed in the process. Poles were ethnically cleansed from Ukraine and Ukrainians from Poland (Ther 1998). During the separation of India and

Pakistan, in 1947, some 10 million people, Hindus, Sikhs, and Muslims, were driven from their homes or fled persecution. The list goes on: Rwanda, Yugoslavia, Nagorno-Karabakh, and East Timor, among others.

3. *A Taxonomy of Ethnic Cleansing*

In these cases and others, the impulse of political leaders is to employ modern integral nationalism (see *Nationalism: General*) to achieve, maintain, and/or build political power. Nationalism claims superiority for the genetic material of the dominant nation, and it counters alienation and anomie by trumpeting a sense of collective belonging. It promotes the perception of shared national values, culture, and experiences. However, modern nationalism also fosters unity and common purpose among the dominant nation by denigrating the qualities of the subordinate ones. Part of nationalist rhetoric is the identification and stereotyping of the 'other,' an ethnic, religious, and/or national minority. The other cannot be tolerated because of their inherent inferiority and treacherousness. These minority groups are targeted for persecution and forced deportation, which almost always takes place in the context of violence.

Those who are left behind become even more dangerous than those who are forced out: thus the impetus in ethnic cleansing is to drive out every member of the minority, leaving no exceptions. There is a totalistic impetus to ethnic cleansing as a result. Sometimes, in the case of the Armenians, the Jews, and the Tutsi, ethnic cleansing turns into genocide or overlaps with genocide. But almost everywhere, the 'unmixing of peoples' (Brubaker 1996) involves the application of extreme violence. Even in those cases where population exchange is mandated by international treaties—Lausanne in 1923, Potsdam in 1945, and Dayton in 1995 are the most famous examples—they usually follow upon, rather than prevent or ameliorate, ethnic cleansing.

Ethnic cleansing almost always takes place under the cover of war) or in a period of transition from war to peace. The force of international opinion or journalistic inquiry is easily neutralized. Paramilitaries are involved in ethnic cleansing even more than regular army units. Strategic objectives for displacing a people are used as a way to justify the actions. Sometimes, the territory inhabited by the targeted minority is designated as strategically located. Sometimes, the argument is that the nations themselves are allied with foreign enemies. During war, normal civil laws are often suspended; men are accustomed to bloodletting; refugees and the dead and wounded are a normal part of the landscape. Soldiers follow orders, and few are punished for attacking the 'other.'

Yet the violence of ethnic cleansing has a different character from that of war. In war, men in military formations are aligned against one another and fight with guns and machines in a more or less reciprocal exchange. In ethnic cleansing, it is usually an armed perpetrator who attacks an unarmed civilian. Those who are conducting the 'cleansing' have complete power over the other, producing a variety of brutal and sadistic practices that have little relationship to the normal rules of military engagement. They burn down homes, rape women and girls, beat helpless victims, and rob indiscriminately. In the process of dislodging peoples from their homes and their homelands, every cruelty—maiming, branding, torture—has been used.

The violence of ethnic cleansing is therefore directed almost exclusively against civilians. This usually means women, children, and the elderly. The gendered quality of ethnic cleansing is notable. Unarmed women and girls are attacked and brutalized in almost every case. The history of rape, mass rape, and rape murder in Bosnia–Herzegovina, where some 50,000 rapes are said to have taken place, was not unique to ethnic cleansing, as it has occurred throughout the world. Females are targeted both as the biological source of the future of ethnic groups, and as the bearers and teachers of national cultural values. They are often the primary victims of ethnic cleansing, as men are the major victims of war.

It is never sufficient in ethnic cleansing simply to expel a people from the designated territory. Their physical and cultural monuments and artifacts are targeted for destruction, as well. Serbs routinely blew up mosques and desecrated Muslim graveyards; Turks tried to destroy the remnants of Armenian and Greek civilizations in Anatolia; Jewish gravestones were used to pave roadways or factory floors, as were the gravestones of the Chechens and Ingush. The names of the peoples, of their towns, and of their regions, were routinely changed, removed completely from sight, chipped away from stone engravings, and even banned from speech. There should be no physical traces left of their previous existence. They are written out of history, or histories are rewritten to denigrate their roles in the region they formally inhabited.

4. *The Future of Ethnic Cleansing*

There is every reason to believe that ethnic cleansing will continue to take a fearsome toll on innocent human beings in the twenty-first century. Part of the problem is that international institutions have a very difficult time intervening in sovereign countries for humanitarian reasons alone. Most of the great catastrophes of ethnic cleansing and genocide in the twentieth century—the Armenian genocide, the Holocaust, Rwanda—were well known to the world community. Yet little if anything was done to stop them.

The recent NATO actions in Bosnia and Kosovo

demonstrate just how difficult it is to gain international consensus for intervention. The Russian involvement in Chechnya might well end in ethnic cleansing; only the most tepid of protests are registered on behalf of the hundreds of thousands of innocent refugees from the conflict. Even when intervention takes place, it is very difficult to separate nations peaceably who have experienced the ravages of ethnic cleansing. One of the fallacies of those who justify ethnic cleansing is that peace is advanced by forceably creating homogeneous nation-states. Instead, long-term animosities and national traumas are engendered that can explode into violence and war.

Especially in the former communist world, there are numerous countries with weak state structures, underdeveloped civil societies, and struggling economies. Where political elites are willing to play the 'nationalist card' in order to mobilize populations on behalf of the modernizing nation-state, ethnic cleansing might well occur. Under similar circumstances, countries like Turkey, Afghanistan, Pakistan, and Indonesia could also be vulnerable to attacks on minority nations.

See also: Anti-Semitism; Ethnic Conflict, Geography of; Ethnic Conflicts; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Ethnic Groups/Ethnicity: Historical Aspects; Genocide: Anthropological Aspects; Genocide: Historical Aspects; Holocaust, The; Race: History of the Concept; Racism, History of; Xenophobia

Bibliography

Ahmed A S 1995 Ethnic cleansing: A metaphor for our time? *Ethnic and Racial Studies* **18**(1): 1–25

Allen B 1996 *Rape Warfare: The Hidden Genocide in Bosnia–Herzegovina and Croatia*. University of Minnesota Press, Minneapolis, MN

Bartov O 1996 *Murder in our Midst: The Holocaust, Industrial Killing, and Representation*. Oxford University Press, New York

Baumann Z 1989 *Modernity and the Holocaust*. Cornell University Press, Ithaca, NY

Bell-Fialkoff A 1996 *Ethnic Cleansing*, 1st edn. St. Martin's Press, New York

Brubaker R 1996 *Nationalism Reframed: Nationhood and the National Question in the New Europe*. Cambridge University Press, Cambridge, UK

Cigar N 1995 *Genocide in Bosnia: The Policy of 'Ethnic Cleansing'*, 1st edn. Texas A&M University Press, College Station, TX

Gutman R 1993 *A Witness to Genocide*. Macmillan, New York

Hayden R M 1996 Schindler's fate and Balkan tragedy of the 1990s: Genocide, ethnic cleansing, and population transfers. *Slavic Review* **55**(4): 727–48

Kuper L 1981 *Genocide: Its Political Use in the Twentieth Century*. Yale University Press, New Haven, CT

Lemberg H 1992 'Ethnische Säuberung': Ein Mittel zur Lösung von Nationalitätenproblemen? *Aus Politik und Zeitgeschichte: Beilage zur Wochenzeitung Das Parlament* B **46**(92): 27–38 ['Ethnic cleansing': A means for the solution of nationality problems?' *From Politics and Contemporary History: A Supplement to the Weekly, The Parliament*]

Martin T 1998 The origins of soviet ethnic cleansing. *Journal of Modern History* **70**(4): 813–61

Melson R 1992 *Revolution and Genocide: On the Origins of the Armenian Genocide and the Holocaust*. University of Chicago Press, Chicago

Naimark N 2001 *Fires of Hatred: Ethnic Cleansing in Twentieth Century Europe*. Harvard University Press, Cambridge, MA

Scott J C 1998 *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT

Stiglmayer A 1994 *Mass Rape: The War against Women in Bosnia–Herzegovina*. University of Nebraska Press, Lincoln, NE

Ther P 1998 *Deutsche und polnische Vertriebene: Gesellschaft und Vertriebenenpolitik in der SBZ/DDR und in Polen 1945–1956* [*German and Polish Deportees: Society and Deportee Policy in the SBZ/GDR and in Poland 1945–1956*]. Vandenhoeck & Ruprecht, Göttingen, Germany

Vujacic V 1993 Communism, Nationalism, and Democracy in Russia and Serbia: 1985–1993, draft manuscript

N. M. Naimark

Ethnic Conflict, Geography of

Ethnic conflict is a worldwide phenomenon. Much of it is territorially based, entailing disputes over the control of space. It occurs over a wide range of spatial scales, from the interstate level to the urban neighborhood. Violence takes place at the extreme, though a significant amount of ethnic conflict is characterized by nonviolent behavior. Causal factors can range from material welfare concerns to identity issues. A range of territorially based solutions or at least attempts at conflict regulation have been put forward.

1. Spatial Scales

It is possible, from a geographical perspective, to classify ethnic conflicts into three categories—the interstate, the intrastate and the micro-scale or intra-urban. However it must be stressed that there are powerful linkages between the various scales. Events at the interstate scale can reverberate down into the individual states and may even have consequences at the level of urban neighborhoods. Likewise, conflicts at the urban micro-scale can impact at the larger spatial scales, even as far as triggering ill-feeling or worse, in the 'international' arena.

1.1 Ethnic Conflict at the Interstate Scale

The geographical distributions of many ethnic groups do not conform neatly to existing state boundaries (the

Basques and the Kurds provide clear examples of this). The misfit between state and ethnic group boundaries may generate claims. As Donald Horowitz (1985) has put it 'if irredentism is conceived as a movement to retrieve ethnic kinsmen and their territory across borders, the common disjunction of group boundaries and territorial boundaries offers scope for irredentas aplenty.' Horowitz further notes that a decision to attempt to forcibly retrieve ethnic group members across a border is, in the main, a government decision. This contrasts with secession, which is an ethnic group decision to break away from the state that the group currently finds itself penned into. Secession may involve an attempt to form a separate state, or it may mean breaking away from one state and joining another.

Stanley Lieberson (1972) has stressed that an ethnic group has the possibility of reducing or eliminating any disadvantage it may suffer, by opting for political separation. As he points, out such a strategy is not available to other disadvantaged groups in a stratified society (for instance those disadvantaged on grounds of gender, age or economic position). Thus, as Lieberson outlines it, the most fundamental difference between ethnic and other forms of stratification lies in the fact that the former is nearly always the basis for the internal disintegration of the existing boundaries of the state. Since the objective is most likely to be the formation of a new state, the best way to describe such secessionist groups is to refer to them as 'ethnonationalists.' Such attempts at the disruption of existing states will not be greeted favorably by existing governments. Thus secession is likely to be pursued and to be resisted by the use of violence.

Interstate ethnic conflict also manifests itself through the roles of 'external national homelands' (Brubacker 1996) and ethnic diasporas. With the former a state may feel an obligation towards its fellow ethnics in one or more other states. This obligation may translate into attempts to influence those other states' policies towards the homeland state's co-ethnics, or to the offering of immigration and citizenship privileges for returning members of the ethnic diaspora. Involvement, however, may extend to irredentist claims. Ethnic diasporas, on the other hand, may express concerns about circumstances in their original (real or mythical) home countries. In this case elements in the diaspora communities will press their 'own' government to adopt policies that are perceived to be advantageous to fellow ethnics in the 'old country' or ethnic homeland. Irish Americans vis-à-vis Ireland, Jewish Americans vis-à-vis Israel and Arab Americans vis-à-vis Palestine are clear instances of this. Indeed some of these groups may well adopt stances that are more extreme that those taken by their fellow ethnics 'back home,' or as Samuel P. Huntington has put it, they may be 'more Catholic than the Pope' (Huntington 1997). He also observes that diaspora communities become particularly active in the age of electronic communication: 'the commitments of diasporas are reinvigorated and sometimes polarized by constant contact with their former homes.'

Ethnic conflict at the interstate level is also evident in what Samuel Huntington labels 'fault line wars.' While these wars erupt at the interfaces between 'civilizations' (labeled by Huntington 'Western,' Islamic,' African' etc.), they frequently involve localized ethnic groups, acting, as it were, as the standard bearers of their respective global-scale groupings. Of course, some of these fault-line wars occur at the intrastate level. In either case they are struggles for control of people and of territory.

Population transfers play a significant role in many interstate ethnic conflicts. If borders cannot be adjusted, peoples may be 'adjusted' instead. Sometimes the ethnic territories are 'purified' by agreement, as with the transfer of Greeks and Turks in the early 1920s. Frequently, however, transfers are achieved by expulsion or—the ultimate savagery—by genocide. Finally, it can be noted that some transfers are associated with what John McGarry (1998) calls 'demographic engineering' on the part of states. Here 'agents' may be moved in, being populations allotted special roles on behalf of the state concerned; on the other hand some groups perceived as 'enemies' may be moved out, enemies being defined as groups whose present locations pose problems for the authorities and an obstacle to their goals (these goals generally being ones of territorial control). As with all types of population transfer, these movements can occur both within and between states.

1.2 Ethnic Conflict at the Intrastate Scale

Gurr and Harff (1994), in their book on ethnic conflict in world politics place ethnic groups into four categories—ethnonationalists, indigenous peoples, communal contenders and ethnoclasses. Ethnonationalists are relatively large and regionally concentrated ethnic groups, which live within the boundaries of a single state or straddle several adjacent ones. They are likely to be seeking a greater degree of autonomy or even independent statehood. Indigenous peoples may also be seeking some degree of autonomy, but are particularly concerned about the discrimination and exploitation experienced at the hands of the more technologically advanced peoples who, by and large, control them. Communal contenders, unlike ethnonationalists, do not seek separation or secession, rather, they seek to share power in the governance of the state they reside in.

As with communal contenders, ethnoclasses seek equality. Unlike them, however, they are usually spatially dispersed, rarely having a well-defined territorial base (for example, where ethnic entrepreneurs operate in specialized, spatially dispersed economic niches). Consequently, in their case, strategies aimed

at secession from the existing state or at seeking some degree of autonomy within it are not relevant.

Thus, while some ethnic group conflicts impinge at the inter-state level, many are contained within the one state and do not spill over 'international' boundaries. The roles played by dominant groups in these situations are crucial—they set the context for the other ethnic groups present. What Hennayake (1992) calls majority ethnonationalism can generate a reaction in a situation he refers to as interactive ethnonationalism. In a similar vein Brubacker (1996) writes of 'nationalizing states' where dominant elites promote the language, culture, demographic position, economic flourishing or political hegemony of their own group, creating an environment disadvantageous to the cultural aspirations and material interests of the other ethnic groups present within the state. Indeed here issues of inclusion and exclusion are met, seen at their most general in the distinction between 'civic' and 'ethnic' nationalism. With the former, the nation is ethnically inclusive of all those who subscribe to the nation's political creed. With the latter the claim is made that an individual's deepest attachments are inherited, not chosen (Ignatieff 1993). National identity is defined by ethnic identity—*ein Volk*: *ein Staat*.

The maintenance of political and social cohesion in multi-ethnic states will be possible under three contrasting circumstances: first, where we find civic nationalism to be dominant, second, where one ethnic group is dominant and third, where an external power exerts hegemony. This latter situation has been particularly prevalent where imperial powers have operated—from the British, French, German, Austro-Hungarian and the Russian empires of the nineteenth and early twentieth centuries to what Michael Ignatieff (1993) refers to as the Soviet and American joint imperium after World War II. However with the collapse of all but one of these, the lid of the pressure cooker has been lifted and apparently solid states have fragmented (the Soviet Union and Yugoslavia being the most notable instances). This has led to the (re-) emergence of previously suppressed or slumbering ethnic groups. Such a reawakening has threatened the fabrics of many states, generating movements that call for varying degrees of separation (from secession to regional autonomy).

Ethnic struggles at the intrastate level, from time to time, have been characterized both by what may be called 'census wars' and by 'symbolic strife.' With reference to the former Donald Horowitz (1985) has coined the phrase 'winning the census'—here there are concerns about existing (ethnic) demographic balances and future trends. As he puts it 'numbers are an indicator of whose country it is.' Census watching becomes an obsession; arguments occur regarding the accuracy of census enumerations, and in some cases ethnic numeral balances are so sensitive that the census is abandoned (as happened, for instance, in Lebanon).

Matters of symbolism also take on a huge importance in states with internal ethnic conflict. Raymond Breton (1984) argues that the symbolic order is a key component in constructing and maintaining a collective entity. A dominant ethnic group will attempt to impose its symbolic order on the state—flags, official language, anthems, etc. Other ethnic groups may resist this process, with the consequence that many symbols become contested and contribute to disunity. Ethnonationally contested spaces, in particular, are notable for their symbolic discensus. Of course symbols may well help create unity within a particular ethnic group, but in terms of inter-ethnic relations they divide.

Within-state ethnic conflict may arise between indigenous or well-rooted populations, on the one hand, and recent, less well-rooted immigrants on the other. If the immigrant flow is overwhelming (numerically or technologically) the indigenous ethnic group(s) will find themselves marginalized (as was the case in North America). However many immigrant flows have to accommodate themselves to the pre-existing social and political environment. In this case, conflict will be over equality of treatment—the immigrant will not expect (or presumably wish) to take over the state that receives them.

1.3 Ethnic Conflict at the Micro-scale

Ethnic conflict gains its greatest intensity at the small scale—here individuals and small groups interface with each other as part of the daily round. Rural, relatively low population density environments can experience intense conflict (as well as day-to-day coexistence), but it is usually in the urban, big city environments that ethnic antagonisms can achieve a peculiarly focused, 'molecular' nastiness.

In the urban context, foreign origin migrants attempt to establish themselves as do those from within the state, some of which may also have their origins in ethnically differentiated regions. While conflict is not necessarily an inevitable outcome, it is of frequent occurrence, as the migrants compete with each other and with the receiving society for scarce resources of housing, employment, educational opportunities and so on. In these circumstances many ethnic newcomers experience a degree of residential segregation, part based on the wishes of the immigrant and migrant ethnics themselves, part based on the somewhat hostile response of those who see themselves as the 'hosts.'

From a bottom-up perspective, ethnic groups may be seen to cluster together residentially for reasons of physical defense, from a wish to distance themselves from the embarrassment of contact with ethnocentric others, and, indeed from a wish to have a base for organizing politically and thereby gaining a say in the decision making in the wider society through judicious use of the electoral system. In the more extreme

conflict environments, the segregated residential cluster can provide a base for organizing physical attack against a wider society that is seen to be oppressive. This is where the urban guerrilla is most likely to thrive. Finally segregation can provide a context for cultural preservation, one where the ethnic culture can be tended and, indeed, transmitted to future generations. Some of this clustering is for very positive reasons, but conflict can also be a key generating factor (Boal 1987).

From the viewpoint of the encompassing society, segregation also has its merits—it provides a means of containing 'alien' populations, it insulates host culture from what may be perceived as undesirable influences, it provides a means of manipulating and minimizing the electoral impact of the immigrant ethnics and, at the extreme, it sets up the ethnic clusters as readily definable targets for attack. Beyond the urban scale, Nurit Kliot (1986) refers to intermingled ethnic clusters in Lebanon as 'hostage situations.' The same situation can apply at the urban scale.

Much ethnic conflict at the urban scale is relatively low key, entailing competition for access to the resources of everyday life and frictions over cultural matters. However, some cities become cockpits of ethnonational strife. In this situation, local conflicts and competition become embroiled in issues of sovereignty, and are thereby greatly intensified. Meron Benvenisti (see Bollens 2000) has labeled such cities 'polarized,' while Joël Kotek (1999) has applied the term 'frontier'. According to him, frontier cities are above all disputed places because they are subject to contradictory and opposing sovereignty claims. For Kotek, a frontier city is 'a territory for two dreams.'

The urban encapsulation of ethnonational conflict may be seen as the city merely mirroring wider conflicts over sovereignty, secession, irredentist claims and desires for regional autonomy. More accurately, cities displaying ethnonational encapsulation are, in reality, actually key players in the conflicts in their own right. Examples, such as Jerusalem, Belfast, Brussels, Montreal, Nicosia and Sarajevo make the point. Indeed some of these cities are the epicenters of the wider conflict and are the focus of particularly intense ethnonational disputation. Jerusalem stands out in this regard.

2. *Conflict Regulation and Resolution*

Solutions to ethnic conflicts can take a number of forms—territorial approaches, dominance approaches and mutuality approaches. Territorial solutions entail actions that increase the ethnic homogeneity of specified territorial spaces. The most extreme approaches involve the removal—by genocide or by forced population transfer—of one or more ethnic groups, producing what might be called 'purified' spaces. More acceptably, territorial approaches can involve the creation of separate states for each ethnonational group (partition or secession), or for cantonal or federal arrangements within the one state. Of course processes of integration/assimilation can also remove difference within a given state. If these occur by consent, then inter-ethnic relations are likely to be positive; if they occur in a coercive manner the consequences will be alienation and the consequent storing up of future conflict.

Dominance is where hegemonic control is exerted by the most powerful group present in a multi-ethnic environment. It is unlikely to provide a long term solution to inter-ethnic conflict—rather it suppresses it, leading to a situation where one is likely to find reluctantly acquiescent ethnic minorities at best, considerable alienation at worst.

Mutuality is where ethnic groups in conflict reach a point where they are prepared to enter into a process of mutual recognition and acceptance. Consociational/power sharing arrangements within the state will be the outcome. In this situation we would find what has earlier been referred to as civic nationalism. Space is shared in a process of mutual accommodation (see discussions in O'Leary and McGarry 1995 and in Boal 1999).

See also: Conflict Sociology; Ethnic Cleansing, History of; Ethnic Conflicts; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Ethnocentrism; Ethnonationalism: Cultural Concerns; Racial Relations; Racism, History of; Racism, Sociology of; Tribe;

Bibliography

Boal F W 1987 Segregation. In: Pacione M (ed.) *Social Geography: Progress and Prospect*. Croom Helm, London, pp. 90–128

Boal F W 1999 Seeking the common ground. *Geopolitics* **4**: 239–61

Bollens S A 2000 *On Narrow Ground: Urban Policy and Ethnic Conflict in Jerusalem and Belfast*. State University of New York Press, Albany, NY

Breton R 1984 The production and allocation of symbolic resources: an analysis of the linguistic and ethno-cultural fields in Canada. *Canadian Review of Sociology and Anthropology* **21**: 123–44

Brubacker R 1996 *Nationalism Reframed*. Cambridge University Press, Cambridge, UK

Connor W 1994 *Ethnonationalism: The Quest for Understanding*. Princeton University Press, Princeton, NJ

Gurr T R, Harff B 1994 *Ethnic Conflict in World Politics*. Westview Press, Boulder, CO

Hennayake S K 1992 Interactive ethnonationalism: an alternative explanation of minority ethnonationalism. *Political Geography* **11**: 526–49

Horowitz D 1985 *Ethnic Groups in Conflict*. University of California Press, Berkeley, CA

Huntington S P 1997 *The Clash of Civilizations and the Remaking of World Order*. Simon and Schuster, London

Ignatieff M 1993 *Blood and Belonging: Journeys into the New Nationalism*. BBC Books, London
Kaplan D H 1992 Nationalism at a micro-scale: educational segregation in Montreal. *Political Geography* **11**: 259–82
Kliot N 1986 Lebanon: a geography of hostages. *Political Geography Quarterly* **5**: 199–220
Kotek J 1999 Divided cities in the European cultural context. *Progress in Planning* **52**: 227–37
Lieberson S 1972 Stratification and ethnic groups. In: Richmond A H (ed.) *Readings in Race and Ethnic Relations*. Pergamon, Oxford, UK, pp. 199–209
McGarry J 1998 Demographic engineering: the state-directed movement of ethnic groups as a technique of conflict regulation. *Ethnic and Racial Studies* **21**: 613–38
O'Leary B, McGarry J 1995 Regulating nations and ethnic communities. In: Breton A, Galleotti G, Salmon P, Wintrobe R (eds.) *Nationalism and Nationality*. Cambridge University Press, Cambridge, UK
Waterman S 1987 Partitioned states. *Political Geography Quarterly* **6**: 151–70

F. W. Boal

Ethnic Conflicts

1. What is Ethnic?

Ethnicity is ancient and ubiquitous, and commentaries on ethnic differences have been highly variable over time and place. The term has been used variously to signify 'nation,' 'race,' 'religion,' or 'people,' but the central generic meaning is that of collective cultural distinctiveness. For the present we shall avoid the popular but awkward and potentially misleading 'ethnic group' in favor of the more convenient term ethny. An ethny here is a culturally distinctive collectivity, larger than a kinship unit, whose members claim a common origin or descent. The prototype is a local endogamous population sharing cultural traits that differentiate it from other collectivities. From such groupings, more extended ethnies develop by nepotism, extended endogamy, fictive kinship, descent myths, political enclosures, economic linkages, and territorial expansion. For modern large-scale ethnies the 'symbolic' boundaries can be quite vague and elastic but the essential retained qualities are ascribed membership (by birth) and cultural identity (cf. Williams 1994, pp. 52–3, 57–8).

An elementary but critical distinction, often ignored in scholarly discussions, is that ethnicity can refer either to boundary-markers—an ethny's distinctive culture or lines of social closure—or to the content of the issues (or 'stakes') in ethnic confrontations. Thus, an 'ethnic conflict' can mean that two or more distinctive ethnies are fighting to control scarce resources (oil, gold, timber, diamonds, water, land, fishing grounds). The contenders are ethnic but the stakes are not. But in other cases, the objects of rivalry or violent conflict are themselves ethnic: language use, religious practices, marriage customs, domestic law, ceremonies and holidays, and so on. Especially likely to lead to severe conflict are situations in which rigidly-bound ethnies are rivals for political control of centralized states. Many so-called ethnic conflicts are struggles over non-ethnic goods, but genuine 'conflicts of identity' are those in which boundaries are rigid and salient and the objects of contention are cultural.

Much scholarly disputation has centered upon whether ethnicity is primordial or instrumental. Primordial ethnicity is seen as closely tied to kinship and descent, rigidly bound, enduring, emotionally charged. Instrumental (situational) ethnicity is thought to be ambiguous, changeable, driven by considerations of advantage or disadvantage in the pursuit of immediate interests. Indeed, it has long been recognized that ethnic boundaries are often permeable and changeable—because of territorial intermingling, continuous variations in cultural traits, interethnic interactions, intraethnic diversity, and state interventions (Levine and Campbell 1972, Chap. 7). But the inclusive reality is that ethnies are both primordial and circumstantial—some are fluid, others rigid; some endure over centuries, others are short-lived. Over the long run, much change can be observed. But in the short run, of one lifetime or a few generations, strong ethnic boundaries are often associated with great inequalities of social, economic, and political status, with strongly felt grievances, and with passionate commitments, solidarities, and conflicts. Today's world of vast migrations and rapid economic and political changes often results in change and merging of ethnies, and individuals frequently have multiple ethnic identities. Nevertheless, there is no prospect that ethnicity will disappear: it might be said, to paraphrase V. Pareto that those who seek to totally abolish ethnicity are engaged in cutting holes in the water. Because membership is an ascribed status, intra-ethnic relations tend to be diffuse and particularistic; for the same reason ethnic politics tends to be exclusivistic. While ethnies thus look backward into origins and history, they also look sidewise to persons who share in communal distinctiveness, and forward into a future of shared fate.

Struggles over definitions in this field have a long and complex history. Because the objects of interest are inherently complex, the search for the One True Definition will obviously fail.

This consideration also applies to definitions of our other key term, conflict, which is loosely used in ordinary discourse. In the present review conflict refers to social behavior, not to psychological processes or cultural contradictions; it consists of a struggle in which an opponent seeks to neutralize, defeat, injure, or eliminate another. It is not synonymous with competition, regulated contestation or rivalry. In particular, the distinctive character of violent conflict

must be recognized. Violence is a qualitatively distinct form of conflict (Williams 1994, p. 54, Brubaker and Laitin 1998, p. 426).

2. Recent Changes in Research and Theories

In the first decades after World War II, as overly simple theories of innate racial differences were discredited, the dominant explanations of ethnic conflict focused upon prejudice and discrimination—first in terms of individual personality and then as features of social location and of institutional norms and practices. Later research emphasized macrosocial factors and conditions, including state structures and policies as well as interstate relations.

For understandable reasons, much research on ethnicity until recently focused upon domestic (national) populations and truly comparative studies were few. Accordingly, many theories have been concerned with intra-societal processes such as economic competition, internal colonialism, labor markets, assimilation, and prejudice and discrimination. Overlapping in time, however, and increasingly prominent have been historical and comparative studies: of ethnic nationalism and conflict, of multiethnic societies, of international linkages and interventions, and of democide and genocide.

As attention has been drawn increasingly to worldwide perspectives, earlier preoccupations with psychological factors and anthropological case studies have receded in favor of systematic comparison and statistical analyses of large data-sets. Frequent and severe ethnic warfare and large-scale immigration and flows of refugees from political repression and civil wars have stimulated both types of research. Since the 1960s, the increasing use of multivariate statistical modeling of large data-sets has produced many substantial empirical generalizations. The worldwide compilation of data by Gurr (1993) indicates that contention among ethnies for control of central states is the most deadly form of communal conflict—even more so than ethnonationalist movements or the struggles of indigenous peoples. Analysis of many protests and violent mass actions in the former Soviet Union during 1965–89, showed that nonviolent protests tended to come from mobilization of well-educated urban populations having substantial organizational resources (Beissinger 1992). Protracted ethnic civil wars cluster in weak but repressive states in the Third World and in the peripheries of industrialized countries. Numerous comparative studies show that ethnic conflicts are more severe in autocratic than democratic states; democracies have more numerous but less deadly protests. An exhaustive compilation and analysis of available records show that genocides and democides are greatest in non-democratic polities (Rummel 1994, see Horowitz 1997, Johassohn and Bjorson 1998).

The most recent research has recognized the severe limitations of case studies and of static correlational analyses, as well as the hazards of simplified 'one-factor' theorizing. There is increased emphasis on multivariate studies and upon the timing and sequence of relevant events and conditions. Correspondingly, more studies are using the techniques of event history analysis, qualitative comparative analysis, and ingenious combinations of narrative accounting together with newer statistical techniques. Comparisons are increasingly regarded as essential to reliable generalizations across time and societal contexts. There is renewed emphasis upon the need for disaggregation of data along with the equally urgent specification of concepts to identify the different types of states, of kinds of ethnicity, of opposition and competition and rivalry, of grievances, of the diverse types of conflict and of violence, of processes of mobilization and de-mobilization. Particular attention is now directed to the ways in which ethnic relations are affected by new transnational processes.

3. Characteristics of Violent Ethnic Conflict

Just as ethnies are diverse, so are the forms of ethnic conflict. The main types are: turmoil, internal war, democide and genocide. Turmoil includes strikes, demonstrations, mutinies, sabotage, rioting, and terrorism. Internal war involves armed collective violence as in *coups d'etat*, separatist rebellions, civil or guerrilla wars, and revolutions. Democides are state-organized or incited mass killings of political opponents and dissidents; genocides are organized efforts to destroy a whole cultural, religious, or racial population (Rummel 1994). These types represent different combinations of: (a) the extent of mass mobilization; (b) the degree of centralized organization and control; and (c) the amounts and kinds of violence. Usually turmoil is low on all three conditions; revolutions involve maximal levels of all; democides and genocides maximize the slaughter of unarmed civilians. Although ethnic wars may not be more difficult to terminate than other civil wars (Licklider 1993), the ethnic factor tends to create perceptions of total and irreversible threat and thus to encourage extraordinary ferocity and brutality, as attested by the frequent observation of torture and mutilation of victims (Liberia, Sierra Leone, Rwanda, Kosovo, Sri Lanka). In any case, once violent conflicts are under way, they develop their own dynamics—attracting opportunists, creating desires for revenge, shifting alliances and balances of forces, transforming issues. Initial conditions may not predict outcomes, and limited conflicts that are easy to start can escalate into protracted warfare.

Complexity and fragmentation frequently characterize ethnic and regional partisans. A clear example is presented by Basque nationalist movements, which were criss-crossed by class, rural-urban, linguistic,

religious, and ideological cleavages so that they were both 'complex and paradoxical' (Shafir 1995, p. 102); even the small militant underground organization, ETA, itself experienced repeated schisms. In Sri Lanka, the Tamil resistance to the central state repeatedly divided into opposing factions (Tambiah 1986). In the southern rebellion against the Sudanese government, tribal and local allegiances as well as ideological differences have plagued the movement and its armed forces. In Liberia, state collapse resulted from factional civil warfare, and in Somalia a central state splintered into clan fighting.

In the scholarly literature, there is a growing emphasis on the distinctive characteristics of violence over nonviolent forms of opposition and protest (Williams 1994, p. 62, Tambiah 1996, p. 292, Brubaker and Laitin 1998, pp. 426–7). Indeed Brubaker and Laitin contend that the study of violence should be treated as an autonomous phenomenon in its own right (1998, p. 426). Nevertheless, a rich store of case studies provide instructive data, as in the growing literature on Rwanda and Burundi.

4. *Sources of Ethnic Conflict*

Around the world at the close of the twentieth century ethnic-related conflicts were frequent and deadly; nearly all wars in the 1990s were intra-state and most of these involved ethnic cleavages. How are these remarkable facts to be explained?

Some popular interpretations are not convincing. Collective ethnic violence is not primarily due to 'modernization' (urbanization, commercialization, industrialization): some of the most lethal ethnic-related conflicts have been in less modernized countries such as Rwanda, Burundi, Sudan, and Zaire. Nor is the partial end of the Cold War, allegedly 'unleashing' suppressed ethnic antagonisms, a satisfactory account; the superpowers actually exacerbated many ethnic confrontations and the upsurge in ethnic conflicts began in the 1950s and 1960s, long before the diminution in superpower rivalries. Furthermore, scepticism is warranted of the notion that the rise of ethnic warfare results from the release of so-called ancient hatreds. Long-ago, wrongs are effective instigators of today's conflicts only when reactivated and politicized, often as the product of elites' manipulation (Gagnon 1994, 1995). Of course, the deadly effects would not occur unless there were some ethnic identities and allegiances to which political appeals could be made. But peaceable neighbors do not suddenly attack one another merely because their ancestors once fought. Such 'general laws' as that people tend to more often favor others when the others are fellow ethnies than when they are not, tells us little about specific cases in which a wide variety of other conditions may be involved. So generalized propositions about how initial conditions of disadvantage and discrimination can lead to group-level grievances still leave us searching for additional and/or intervening conditions that convert discomfort into collective grievances. Undoubtedly sociobiological characteristics (including propensities for group formation and conformity) constitute a fundamental substratum for ethnic solidarities and conflicts (van den Berghe 1981). Likewise, resource scarcity with the frequent accompaniment of incompatible claims (to water, land, any other resources) must always be taken into account.

But more specific and more proximate conditions and dynamic processes provide the more convincing explanations for the numerous violent ethnic confrontations we now observe. For example, a theoretically significant observation is that severe ethnic conflicts are most common in recently independent states that were formerly colonies of major powers or are successor states resulting from the disintegration of multiethnic polities. These countries typically are multiethnic, relatively poor, with weak but repressive states. Under these conditions, the state is a main source of wealth and prestige and rivalry for political control is likely to be intense while ethnic membership frequently is crucial in political contention. Economic scarcity and social pluralism thus combine with state centrality to encourage communal politics. Because the stakes are high and the outcomes are seen as irreversible, resorting to violence is tempting. In the great majority of such severe conflicts, lack of a common civic culture and of positive inter-ethnic interdependence, combined with struggles for control of the state, create intense threats and countermobilizations. Ethnicity is rendered highly salient and provides an attractive base for political entrepreneurs. A common result is that ethnic rivals come to see each other as rigid groupings and membership as crucial to all life-chances. Every move in that direction escalates and polarizes. Not surprisingly, then, conflicts among communal contenders for power are much more deadly than conflicts involving ethnonationalists or indigenous peoples (Gurr 1993, pp. 98–9). State expansion and impositions upon ethnies are likely in rentier states in which governing elites have sources of income independent of the potential electorate, as in Nigeria and the former Zaire. In collapsed states—as in Lebanon, Somalia, or Liberia—central authorities lose control over contenders for power and multiple-armed struggles result. Partisan external support to states and rebels often has prolonged and intensified ongoing ethnic-related violent conflicts.

In Gurr's formulation (1993), the main factors favoring conflict include strong identity, inequalities and grievances, political opportunity structures allowing mobilization, provocative state policies, and international contagion and diffusion. Many of the ethnic conflicts reviewed by Esman (1994), Gurr (1993), and Horowitz (1985) develop into protracted warfare because one party believes that another seeks total and irreversible domination: 'they will enslave

our children and our children's children,' 'they intend to kill us all.' Diverse case studies, from Northern Ireland, to Sri Lanka, to Lebanon, to Rwanda, and others have emphasized that fear of extinction is far more common in protracted ethnic conflicts than previously recognized. In a drastic oversimplification, it can be said that the most dangerous situations cluster in multiethnic societies governed by centralized but weak and repressive states that fail to provide voice or autonomy to aggrieved ethnic and regional dissidents. Thus, strongly-bound and relatively large ethnies seek control in political systems that do not favor compromise and power-sharing. Particularly explosive are: (a) situations of lost autonomy in unstable polities; (b) abrupt impositions of cultural and economic deprivations by a dominant ethnic state, (c) intrastate conflicts attracting partisan external interventions.

5. *Why is There Not More Ethnic Conflict?*

At any given time only a small percentage of ethnies are politically mobilized. (cf. Levinson 1991–96, Gurr 1993). Of all ethnic disputes, only a small percentage result in collective violence and even fewer in civil war. Most of the world's ethnies are not fighting each other or other states within which they reside. Of the some 1600 major cultural groupings, less than 300 have recently mobilized in politicized protest or rebellion and only some 30–40 wars are ongoing in a particular year. To explain the impressive record of peaceful coexistence, of mutual accommodation, and even of active cooperation must be the task of future research and theory-building. Nevertheless some general observations even now seem justified. First, the reverse of the factors here adduced to help account for conflict can aid our understanding of ethnic peace. For example, we should look for relatively 'open' political systems with non-ethnic parties, lack of imposed segregation and discrimination, relatively equitable distribution of opportunities and rewards, and the presence of an array of conflict-resolution devices and arrangements.

The most obvious circumstance accounting for the failure of ethnies to mobilize for conflict is their members' relative satisfaction with or accommodation to the status quo. If ethnic identity carries with it no acute grievances or severe threats, there is little incentive to disturb existing relationships. This is an important but banal and intellectually uninteresting case. A second case is more provocative: ethnies that have deep grievances but do not rise up in conflict. Often their conflict potential is negated by weakness: lack of resources, internal dissension and social fragmentation, lack of a committed elite, small size, and geographic dispersion. (It must be noted, however, that weakness is a relative matter and some apparently weak ethnies do attack much stronger opponents.) A third factor may be labeled simply as fear—fear of the consequences of protest or rebellion. The apprehension may be that other ethnies will respond in oppositional and hostile ways, or that a repressive and ruthless state with abundant resources will severely punish a rebellious ethny. External aid, so often important, may be lacking or perceived as dangerously unreliable. For these and other reasons, even highly aggrieved ethnies may avoid the more violent forms of protest. To more fully understand the necessary and sufficient conditions for ethnic peace and cooperation clearly requires additional fine-grained analyses of the many specific cases that now exist. Meanwhile the evidence and analyses that are rapidly accumulating provide favorable prospects for advances in knowledge in this crucial field of deep human concerns.

See also: Conflict Sociology; Discrimination: Racial; Ethnic Conflict, Geography of; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity/Ethnicity and Archaeology; Ethnic Identity, Psychology of; Ethnicity, Sociology of; Gender, Class, Race, and Ethnicity, Social Construction of; Identity Movements; Identity: Social; Race Identity; Racial Relations; Racism, History of; Racism, Sociology of; Residential Segregation: Sociological Aspects

Bibliography

Beissinger M R 1992 [1991] Protest mobilization among Soviet nationalities. Presented at Annual Meeting of the American Sociological Association, 86th, Cincinnati; rev. 1992. Peace Studies Program, Cornell University

Brubaker R, Laitin D D 1998 Ethnic and nationalist violence. *Annual Review of Sociology* **24**: 423–52

Enloe C H 1980 *Ethnic Soldiers: State Security in Divided Societies*, Athens: University of Georgia Press, Athens, CA

Esman M J 1994 *Ethnic Politics*. Cornell University Press, Ithaca, NY

Gagnon Jr V P 1994/95 Ethnic nationalism and international conflict: The case of Serbia. *International Security* **19**: 130–66

Goodwin J 1997 The libidinal constitution of a high-risk social movement: Affectual ties and solidarity in the Huk Rebellion, 1946–1954. *American Sociological Review* **62**: 53–69

Gould R V 1991 Multiple networks and mobilization in the Paris Commune, 1871. *American Sociological Review* **56**: 716–29

Gurr T R 1993 *Minorities at Risk: A Global View of Ethnopolitical Conflicts*. United States Institute of Peace, Washington, DC

Horowitz D L 1985 *Ethnic Groups in Conflict*. University of California Press, Berkeley/Los Angeles

Horowitz I L 1997 *Taking Lives: Genocide and State Power*. Transaction Publishers, New Brunswick, NJ

Jonassohn K, Bjorson K S 1998 *Genocide and Gross Human Rights Violations in Comparative Perspective*. Transaction Publishers, New Brunswick, NJ

Levine R A, Campbell D T 1972 *Ethnocentrism: Theories of Conflict, Ethnic Attitudes, and Group Behavior*. Wiley, New York

Levinson D (ed.) 1991–96 *Encyclopedia of World Cultures*. G K Hall, Boston, MA, Vols 1–4

Licklider R (ed.) 1993 *Stopping the Killing: How Civil Wars End*. New York University Press, New York
Marvell G, Oliver P 1993 *The Critical Mass in Collective Action*. Cambridge University Press, Cambridge, UK
Rummel R J 1994 *Death by Government: Genocide and Mass Murder since 1900*. Transaction Publishers, New Brunswick, NJ
Shafir G 1995 *Immigrants and Nationalists: Ethnic Conflict and Accommodation in Catalonia. The Basque Country, Latvia, and Estonia*. State University New York Press, Albany, NY
Tambiah S J 1986 *Sri Lanka: Ethnic Fratricide and the Dismantling of Democracy*. University Chicago Press, Chicago
Tambiah S J 1996 *Leveling Crowds: Ethnonationalist Conflicts and Collective Violence in South Asia*. University of California Press, Berkeley
van den Berghe P L 1981 *The Ethnic Phenomenon*. Elsevier, New York/Oxford
van den Berghe P L (ed.) 1990 *State Violence and Ethnicity*, 1st edn. University Press Colorado, Niwot, CO
Williams R M 1994 The sociology of ethnic conflicts: Comparative international perspectives. *Annual Review of Sociology* **20**: 49–79

R. M. Williams

Ethnic Conflicts and Ancient Hatreds: Cultural Concerns

The view that ethnic conflicts today can be traced back to ancient group hatreds is known as essentialism, sometimes also called 'primordialism.' It is very popular in nonscholarly circles, but has very few proponents in the scholarly world.

Essentialism relies on two interconnected arguments, sometimes made together. It refers to 'primordial' or 'ancient animosities' as a cause of contemporary conflict. The animosities are said to be based on inherent differences of race, religion, or culture, going back into ancient times; and it is further contended that individuals acquire the characteristics of their races, religions, or cultures. A second essentialist argument is that ethnicity inheres in human beings, meaning that all of us inevitably search for, or can easily be made to care for, our ancestry and blood. Either way, conflict results, for a rational calculus is superseded by the emotional ties of blood, or by ancient hatreds. Human beings live out not only the positive attributes of the collectivity to which they belong, but also its prejudices with respect to other groups. Again and again in history, intrinsic group differences activate prejudices and trigger violence.

More than any other view discussed below, this perspective dominates the portrayals of ethnic conflict in newspapers and magazines. Almost any popular account of conflict between Hindus and Muslims, Serbs and Croats, Arabs and Jews, Whites and Blacks, Catholics and Protestants, Hutu and Tutsi is marked by phrases such as 'old animosities,' 'tribalism,' and 'ties of blood' (Kaplan 1993). In the academic world, too, this view has had its exponents. Clifford Geertz (1973) was among the first scholars after the Second World War to popularize the term 'primordialism.' More than Geertz, however, Walker Connor has been a leading and consistent advocate of the essentialist view. Man, argues Connor, is a 'national, not rational animal' (Connor 1994). So powerful is the search for origins in human beings that both Hitler and John Jay, says Connor, were driven to emphasize common ancestry to build their nations. 'Blood binds more firmly than business,' argued Adolph Hitler; and the Americans, contended John Jay, are 'a people descended from the same ancestors,' 'despite,' adds Connor, 'the presence of settlers of Dutch, French, German,.... Scottish and Irish extraction' on American soil in the eighteenth century.

Few academic scholars subscribe to the essentialist view today. On the whole, essentialism tends *not* to make a distinction between ethnic (or national) identity on the one hand and ethnic (or national) conflict on the other. Though it is a problem not confined to essentialism alone, as explained below, essentialists do tend to conflate the two. It may well be that for building community feelings, leaders tend to, or have to, emphasize ancestry. Conflict, however, does not necessarily follow. It is one thing to call up an imagined common ancestry *across* ethnic groups, in which case the attempt would be to build bridges. John Jay's example above would illustrate a bridge-building invocation of an assumed ancestral commonality. It is, however, something quite different to summon the common ties of blood explicitly against an ethnically distinguishable group, in which case the result could well be ethnic hatred and violence. Hitler's attempt would clearly belong to the latter category, not the former.

That both John Jay and Hitler had to invoke common ancestry implies neither that man is a national animal, nor that he is a rational agent—he could be both. Nor do such attempts inevitably reignite ancient animosities. Ethnic, or national, identity can well be a source of meaning and security, without implying hatred for another ethnic or national group. These two aspects of ethnic or national identity—positive and negative—are by now well understood and clearly distinguished (Taylor 1997).

Those who came to be known as 'instrumentalists' launched the first systematic attack on the essentialist view. Indeed, going back to the late 1960s and 1970s, 'essentialism vs. instrumentalism' was the first big theoretical axis around which scholars sought to formulate their views and understand the virulence of ethnic animosities and violence (Young 1995). The key proposition of instrumentalism rests on the purely instrumental use of ethnic identity for political or economic purposes by the elite, regardless of whether the elite themselves believe in ethnicity (Hardin

1995).The two views are presented as being fundamentally at odds; one focuses on the intrinsic power of ethnic differences, the other concentrates on their instrumental value, political or economic.

The emphasis of on how leaders strategically manipulate ethnicity for the sake of power has an intuitive appeal because the behavior of many, if not all, political leaders can be cited in support. But this view also runs up against a fundamental difficulty. The elite may indeed gain power by mobilizing ethnic identity, without actually believing in it, and may therefore behave instrumentally. But if the masses were only instrumental about ethnic identity, why would ethnicity be the basis for mobilization at all? Why do the leaders decide to mobilize ethnic passions in the first place? Why do they think that ethnicity is the route to power, not the economic interests of the people? And if economic interests coincide with ethnicity, why choose ethnicity as opposed to economic interests as a central symbol for mobilization?

In principle, an instrumental resolution of these problems exists. Ethnicity can serve as a 'focal point,' which is defined as set of symbols so obvious, unique and easily comprehensible by the members of a group that it can facilitate convergence of individual expectations, and hence be useful as a mobilization strategy. The idea of 'focal points' comes from Schelling's seminal treatment of the coordination problem in bargaining. In the famous Schelling example:

> When a man loses his wife in a department store without any prior understanding on where to meet if they get separated, the chances are good that they will find each other. It is likely that each will think of some obvious place to meet, so obvious that each will be sure that the other is sure that it is obvious to both of them (Schelling 1963).

Ethnicity, in other words, can be viewed as one such focal point for mobilization; it is not valued for its own sake. Its potential for mobilization may be deployed to extract goods and services from the modern sector, or to establish power, by some leaders (Bates 1974).

However, the idea of a 'focal point' is also not sufficient to explain ethnic mobilization, for it does not distinguish between different kinds of collective action and what their respective costs might be. Ethnic *mobilization* for political action is not the same as ethnic *coordination* for economic and social activities. By providing a social occasion, festivals may indeed bring people together even if not every one appreciates the ritual meaning of celebration or mourning; and by forming mutually converging trust, geographically spread ethnic kinsmen are also known to have supplied credit in long-distance trade without a prior explicit contract between trading partners.

But the analogy of a focal point cannot be extended to group action when the costs of participation for the masses are very high. By its very nature, ethnic mobilization in politics is group action not only in favor of one's group *but also often against some other group*. An increase in the rights and power for one group often means a diminution in the ability of some other group(s) to dictate terms, or a sharing of power and status between groups, where no such sharing earlier existed. Ethnicity in intragroup, social or economic transactions is thus very different from ethnicity in intergroup political conflicts. The former illustrates the value of ethnicity as a focal point; the latter presents problems of a different order. When an individual provides credit to ethnic brethren without an explicit contract, incarceration, injury, or even death is not the likely cost he bears in mind. Such costs are not unlikely in ethnic or national conflicts. For something to be manipulated by a leader when death, injury, or incarceration is a clear possibility, it must be valued as a good by a critical mass of people, if not by all (Horowitz 1985). A purely instrumental conception of ethnicity cannot explain why ethnic identities are mobilized by leaders at all (Elster 1988).

Does this mean that essentialism is right after all about the primordial character of hatreds. Do such hatreds indeed play an important role in contemporary ethnic conflicts? Are there any other approaches to answering this question beyond essentialism and instrumentalism?

Since the mid-1980s to early 1990s, the study of ethnicity has been profoundly influenced by constructivism (Varshney 1997). Part of the constructivist inspiration has come from postmodernism, but postmodernists are not the only constructivists. At varying levels, scholars as diverse and 'unpostmodern' as Eric Hobsbawm and Ranger (1983), Linda Colley (1992), and Benedict Anderson (1983) have shared the constructivist view, and demonstrated how so many identities that we take for granted today were quite recently constructed in history.

The central claim of constructivism is that the formation of ethnic or national identities is a modern phenomenon. Does this claim by any chance imply that there were no Hindus, Muslims, Jews, Christians, Tibetans, and Han Chinese in premodern times?

Such a claim would obviously be quite absurd, and that is not the constructivist contention either. The claim simply is that identities in premodern times tended on the whole to be face-to-face and operated on a small scale. Ordinary people rarely interacted beyond their local environments. Conflict, when it emerged, was managed locally and identities were considerably flexible. Extralocal communities did not include 'the people'; such larger communities consisted primarily of the ecclesiastical elite and the court-based aristocracy and nobility.

Modernity changed the meaning of identities by bringing the masses into a larger, extralocal, framework of consciousness. It made identities and communities wider and more institutionalized. In what has become a classic constructivist argument, Colley shows how shared Protestantism, opposition to

France, and the benefits of empire managed to dissolve the bitter historical disputes between the Scots and the English, and led to the construction of a British identity in the eighteenth and nineteenth centuries. In Benedict Anderson's *Imagined Communities*, one of the most influential texts in the field today, the emphasis is on how modern technology and a modern economic system—the printing press and capitalism, to be more precise—made it possible to have imaginations about large, popular, and secular communities based on language, which overtook the premodern, extralocal, religious communities of the clergymen on the one hand and the aristocratic dynasties on the other.

The constructivist emphasis is on the construction of group categories by the knowledge-elite, its promotion by centers of power, and its effects on 'the people.' For example, the Census, a modern instrument of categorization, would typically ask the masses whether they were (a) Hindu, (b) Muslim, or (c) Christian, even if in their self-image the masses felt their culture borrowed from all three and their identities were an intersection of (a), (b), and (c). Fuzzy identities, even when real, were not registered as such and were instead split into lucid, modernist boundaries. Furthermore, if public policy, based on such census categorizations, allocated patronage, public office, or state grants, then the very act of census categorization would begin to create larger identities.

Such categorization, according to constructivists, was not innocent or scientific. The power elite created some categories, not others. The selectivity was based on their preconceptions of what the building blocks of a society were, or on a calculation of what divisions would maintain their power. Thus, only those narratives acquire staying power that get promoted by the elite.

On postcolonial societies, the principal constructivist claim is that the major contemporary ethnic cleavages were a creation of the colonial power and, given the immense power of colonial masters, such divisions have endured, and will last for a long time. Of course, the power of the colonial masters was not unlimited. They did require local alliances and the new categories were established with co-optation of the indigenous elite in the colonial framework of power. In a sense, thus, the argument is about the nature of the colonial *system*, not about the evil intentions of the colonial masters. 'Modern colonialism,' writes one of the best practitioners of the constructivist genre, 'instituted enduring hierarchies of subjects and knowledges—the colonizer and the colonized, the Occidental and the Oriental, the civilized and the primitive, the scientific and the superstitious, the developed and the underdeveloped. …(T)he colonial rulers enacted their authority by constituting the "native" as their inverse image…not because of the colonizer's bad faith but due to the functioning of colonial power…' (Prakash 1994).

Let us take a concrete example of this abstract claim. On India's Hindu–Muslim conflict, constructivists argue, there is no universally acceptable theory about the origins, rise, and spread of Hindu–Muslim antagonisms; rather, there have only been 'discourses' or 'narratives.' In the hands of the British, a primordial antagonism between Hindus and Muslims dating back centuries became the 'master narrative,' even though there was enough evidence of Hindu–Muslim coexistence. Primordial antagonism was not the 'truth' about Hindus and Muslims. It was constructed and promoted as such by the British, partly because it suited them to split India into its two largest religious groups, and partly because the 'natives,' argued the British, could not constitute a 'modern' nation—they could think only in terms of premodern religious communities. Hindus and Muslims may have existed before the British came to India, but being a Hindu or Muslim did not refer to large, political identities. It only signified small, personal, and village-based cultural identities.

In what ways are these arguments a challenge to essentialism? Do they advance our understanding of ethnicity or ethnic conflict? What might their pitfalls be?

To explain the contemporary power of ethnic cleavages, essentialism either referred to old animosities or to the pull of ancestry. It was unable to see how the episodes suggesting old animosities might have been selectively retrieved by the knowledge elite, ignoring the many instances of cooperation and coexistence. The argument, one might parenthetically note, is also different from instrumentalism. For instrumentalists, ethnicity is basically a mask for a core of interests, and as interests change, the masks also do. Constructivists show why some identities, striking roots in popular consciousness, endure and do not easily change. To call such identities a mask is to undervalue their endurance and not understand their historical roots. The fact that they are constructed does not mean that they are not *deeply* constructed. Identities do not often, or automatically, change as interests do.

Has constructivism won the battle of arguments, defeating essentialism (and instrumentalism)? There are significant unresolved problems in the constructivist claim as well. After all, the basic constructivist argument for ethnic conflict in developing countries is that the colonially created master narratives, emphasizing cleavages between communities, led to conflict. The explanation provided for the formation of ethnic *identity* is thus also extended to ethnic *conflict*. Whatever one may say about the constructivist explanation for modern identity-formation, its extension to conflict does not make sense.

To illustrate this point, let us return to the example of Hindu–Muslim conflict. Is Hindu–Muslim conflict ubiquitous in India, or is it concentrated in some pockets? One could also ask the question more

generally: does ethnic conflict tend to get evenly spread through the length and breadth of a country, or does it develop local or regional concentrations? If it is true that ethnic violence tends to be unevenly distributed and pockets of violence coexist with vast stretches of peace, then invoking the ubiquitous power of a system-wide, nationwide, master narrative as an explanatory device will not do. Either there are several master narratives in a country, one hegemonic in some regions, others powerful elsewhere; or some other, locally based, factors intervene between the potential power of narratives to cleave society and generate violence, and the actually observed patterns of violence. To the detriment of constructivism, it turns out that ethnic conflict indeed tends to be locally concentrated, not evenly geographically spread, even though ethnic identities may be ubiquitous. This is as true of India, as of the US, and Northern Ireland, countries with such remarkable divergences otherwise.

Essentialism thus may have lost out to other ways of looking at ethnic conflicts, especially its claim that contemporary ethnic conflict is based on ancient hatreds which the scholars of ethnicity simply don't find acceptable any more. But we should note that other alternatives are also riddled with some serious problems of evidence and explanation. It may be that as knowledge of ethnic consciousness evolves further, a new approach that somehow combines the strength of each approach and seeks to eliminate the weaknesses of each will come about. It is hard to tell what this syncretistic approach would look like.

See also: Conflict and War, Archaeology of; Conflict: Anthropological Aspects; Ethnic Cleansing, History of; Ethnic Conflicts; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity, Sociology of; Gender, Class, Race, and Ethnicity, Social Construction of; Internal Warfare: Civil War, Insurgency, and Regional Conflict; War: Anthropological Aspects; War, Sociology of

Bibliography

Anderson B 1983 *Imagined Communities*. Verson, London

Bates R 1974 Ethnic competition and modernization in contemporary Africa. *Comparative Political Studies* **6**

Colley L 1992 *Forging the Nation 1707–1837*. Yale University Press, New Haven, CT

Connor W 1994 *Ethnonationalism*. Princeton University Press, Princeton, NJ, pp. 199–201

Elster J 1988 *Nuts and Bolts*. Cambridge University Press, Cambridge, UK, p. 118

Geertz C 1973 The integrative revolution: primordial sentiments and civil politics in the new states. In: (ed.) *Interpretation of Cultures*. Basic Books, New York

Hardin R 1995 *One for All: The Logic of Group Conflict*. Princeton University Press, Princeton, NJ

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Horowitz D 1985 *Ethnic Groups in Conflict*. University of California Press, Berkeley, CA

Kaplan R 1993 *Balkan Ghosts: A Journey Through History*. St. Martin's Press, New York

Prakash G (ed.) 1994 *After Colonialism*. Princeton University Press, Princeton, NJ

Schelling T 1963 *The Strategy of Conflict*. Oxford University Press, New York, p. 54

Taylor C 1997 Nationalism and modernity. In: McKim R, MacMahan J (eds.) *The Morality of Nationalism*. Oxford University Press, New York

Varshney A 1997 Postmodernism, civic engagement and ethnic conflict: a passage to India. *Comparative Politics* October

Young C (ed.) 1995 *The Politics of Cultural Pluralism*. University of Wisconsin Press, Madison, WI

A. Varshney

Ethnic Groups/Ethnicity: Historical Aspects

1. Is 'Ethnicity' Ancient or Modern?

From one point of view, ethnic groups, and the conflicts between them, are as old as humankind. The ethnopsychiatrist Devereux (1975) therefore used examples from Homer, Herodotus, Sophocles, Plutarch, and Xenophon to the Bible to develop his conceptual approach to *Ethnic Identity: Its Logical Foundations and Its Dysfunctions*. Among a total of 56 groups on whose basis Devereux develops his theories are ancient Hebrews, Spartans, Sedang, Arunta, Mohave, Hipa, Yuma, and numerous examples from modern Europe, especially from the time of fascism.

From another vantage point, ethnic identity is a feature of modernization. The sociologist Glazer (1975), for example, noted in his essay on the *Universalization of Ethnicity* that while processes of modernization could be expected to weaken primordial ethnic identities, in fact, mass societies had created a new need for 'some kind of identity—smaller than the State, larger than the family, something akin to a 'familistic allegiance.' Hence the production of ethnic difference could be expected to intensify and increase under modern conditions.

What speaks for the first assumption is the impressive array of social conflicts involving differences in the sense of belonging and of group membership. There is also the history of the word 'ethnos' which goes back to antiquity and contains an interesting doubleness of standing for people in general as well as for people who are different from the speaker (the Greek *ethnikos* was a translation of Hebrew *goy*), making 'ethnos' available as a term of in-group self-description and of out-group ascription.

What speaks for the second assumption is the fact that in the modern period, with its intensified migration and urbanization processes, ethnic identifications could free themselves from all other moorings in the modern period, that ethnic identification has proliferated steadily and that, at the end of the twentieth century, it seems to have engulfed the whole globe. While the ancient Greek word 'ethnos' has remained the root of many related words (including 'ethnic'), the modern period has contributed a whole vocabulary for the perception and discussion of ethnic matters. The words include 'race' (which emerged in the context of the Spanish Inquisition) and the twentieth-century neologisms 'racism,' 'genocide,' 'ethnicity,' and 'identity.'

2. *Brief History of the Vocabulary of Ethnicity*

The word race, derived from Italian *razza*, Spanish and Castilian *raza*, and Portuguese *raca*, became widespread in the fifteenth and sixteenth centuries. An English example from 1570, referring to the 'race & stocke of Abraham,' supports Leo Spitzer's theory that the obscure ultimate roots of race may lie in the word 'generation.' Verena Stolcke has shown that the word 'race' could mean 'the succession of generations (*de raza en raza*) as well as all the members of a given generation.' In addition to the word's relation of 'generation,' in fifteenth century Spain *race* connoted 'quality' and 'nobility of blood' (an aristocratic sense) as well as 'taint' and 'contamination' (a meaning opposite to that of 'nobility'). Race in its negative sense appeared in the doctrine of purity of blood (*limpieza de sangre*), 'understood as the quality of having no admixture of the races of Moors, Jews, heretics, or *penitenciados* (those condemned by the Inquisition).' The more modern word race thus shared the ambiguity inherent in the ancient *ethnos*.

Twentieth-century fascists used the word racism (coined in analogy to other modern '-isms' such as communism) as a positive term to describe the importance they assigned to race. The term racism came into general use only in the 1930s, and it acquired its pejorative sense when it became associated with intellectual critiques of fascism. Hirschfeld's (1938) remarkable antifascist book *Racism* marked the turning point.

Another neologism to emerge in the period of the struggle against National Socialism was genocide, introduced by Lemkin (1944) in the book *Axis Rule in Occupied Europe*. Lemkin studied the new, eliminationist Nazi occupation policies (most especially toward Jews, but also toward other nationality groups in occupied countries) and found that these measures intended to destroy 'the essential foundations of the life of national groups, with the aim of annihilating the groups themselves.' The attacks on political and social institutions, on 'culture, language, national feelings, religion, and the economic existence of national groups' were steps toward the 'destruction of their personal security, liberty, health, dignity, and even the lives of the individuals belonging to such groups.' As examples of 'genocide legislation,' Lemkin reproduced the texts of numerous laws, including those that established German as the official language in Luxembourg, banned the Cyrillic alphabet in Croatia, and placed Kosovo under Albanian rule. Lemkin's coinage 'genocide' was disseminated widely when the United Nations adopted it in 1945 in an indictment of Nazi leaders for having 'conducted deliberate and systematic genocide—namely, the extermination of racial and national groups.'

The English word ethnicity was once recorded in 1772, in an obsolete instance where it denotes 'heathenish superstition.' In the sense in which we now use it, the word was revived only during World War II, at a time that race had become compromised by racism. Warner reintroduced 'ethnicity' as social category, parallel to sex, age, and religion, by which human beings can be differentiated from each other. In *The Social Life of a Modern Community*, co-authors Warner and Lunt (1941) argued that an 'individual was classified as belonging to a specific ethnic group if (a) he considered himself or was considered by the…community as a member of the group, and (b) if he participated in the activities of the group.' Warner and Lunt introduced the noun when they stressed that 'the concept of ethnicity is not based simply on place of birth.' Just like 'ethnos' and 'race,' 'ethnicity' was an ambiguous term. On the one hand, Warner's 'ethnicity' was an inclusive category, by which every inhabitant of a community could be classified. On the other hand, it could be a term that excluded dominant groups and refer to all inhabitants of the New England community he studied except 'the natives, or Yankees.' Warner used the word 'ethnicity' again in later publications; and Riesman (1953) employed it in an essay. However, the term became more popular only in the 1970s, especially when it was used in the titles of Glazer and Moynihan's (1975) collection *Ethnicity: Theory and Experience* and of Greeley's (1974) study *Ethnicity in the United States*.

The word 'ethnic' has been in circulation for a somewhat longer time. Going back to the noun ethnos, the word was used to refer not just to people in general, but to 'others.' In English usage, the meaning shifted from 'non-Israelite, Gentile' to 'non-Christian.' Thus, the word retained a quality of defining another people, and often negatively, as 'heathens.' Only in the mid-nineteenth century did the more familiar meaning of 'ethnic' as referring to a secular group of people emerge. However, the English language has retained the pagan memory of 'ethnic' in the sense of ethnic as 'other,' as 'nonstandard.' The *Oxford English Dictionary*, thus, traces three strands of meaning that coalesced in the modern sense of 'ethnic' (as in 'ethnic

group'): (a) 'pertaining to nations not Christian or Jewish; Gentile, heathen, pagan' (the oldest sense of the word, documented since the fifteenth century); (b) 'peculiar to a race or nation' (going back to mid-nineteenth-century uses in the sense of 'racial'); and (c) 'ethnic minority (group), a group of people differentiated from the rest of the community by racial origins or cultural background' (popular since World War II).

3. *Approaches to Ethnic Groups, Ethnic Identity, and Ethnicity: A Sampling*

The ethnic group became the subject of scholarly discussion with Weber's (1978) foundational definition in *Economy and Society* written before World War I and first published posthumously in 1922. Weber saw ethnic groups as 'those human groups that entertain a subjective belief in their common descent because of similarities of physical type or of customs or both, or because of memories of colonization and migration; this belief must be important for the propagation of group formation; conversely, it does not matter whether or not an objective blood relationship exists.' He emphasized the sense of commonality (*Gemeinsamkeit*, 'ethnic membership' in English translation) that does not in itself constitute a group but that 'facilitates group formation.' Weber offered a broad international perspective and discussed such ethnic groups as American Negroes, Serbs and Croats, German-speaking Alsatians, Poles in Prussian Upper Silesia, Baltic Germans, and French Canadians. Weber used 'racial' and 'ethnic' interchangeably and wrote that the 'question of whether conspicuous 'racial' differences are based on biological heredity or on tradition is usually of no importance as far as their effect on mutual attraction or repulsion is concerned. Weber also discussed ethnic groups under the term 'nationalities,' the word dominant in the period of World War I and the Versailles treaty.

In his focus on the 'subjective belief,' Weber's concept of the ethnic group implies a psychological component. When Sigmund Freud reflected on the subjective sense of belonging, he arrived at a formulation that was to lead to the coinage of another key term. In a 1926 B'nai Brith address, Freud wondered about his subjective sense of Jewishness—which was independent of religious faith or ethnic pride—and saw it as the result of what he called 'the secret familiarity of identical psychological construction' (*Heimlichkeit der gleichen inneren Konstruktion*). When Erikson (1950) attempted to translate Freud's phrase in *Childhood and Society*, he offered the term identity as an English equivalent for Freud's notion. As Gleason has shown, it was a word that took: for instance, Seligman and Johnson's (1930–1935) *Encyclopedia of the Social Sciences* still had no entry for identity and discussed fingerprinting techniques under 'identification,' whereas Sills's (1968) *International Encyclopedia of the Social Sciences* carried a substantial essay.

Simmel (1908) described an individual as defined by an always varying web of only partly overlapping, multiple group affiliations. He also focused on the figure of the 'stranger' who is 'near and far at the same time.' Simmel's stranger is 'the person who comes today and stays tomorrow' without ever becoming completely part of the group within which he lives. Simmel viewed the European Jew as the exemplary 'stranger.' Simmel's attention to multiple group memberships and to the special role of an outsider who is also an insider gave rise to a focus on marginality in the study of social structures and to the question how 'ethnicity' relates to other, simultaneously existing group affiliations?

In his essay '*Human Migration and the Marginal Man*,' Park (1928) applied Simmel to issues of migration. Park viewed marginality and the race-mixing as the rule and not the exception in human history, and saw the 'marginal man' as the ideal figure for the study of 'the processes of civilization and of progress.' The marginal man thus became a key figure for an understanding of modernity, urbanism, and cosmopolitanism; and marginality and assimilation became central terms for the study of ethnicity.

In his often-cited book *Assimilation in American Life: The Role of Race, Religion, and National Origin*, Gordon (1964) outlined such pluralist and assimilationist concepts of ethnic integration as 'melting pot' (the term derived from Israel Zangwill's play of 1908), 'Anglo-conformity,' and 'pluralism' (the term the philosopher Horace M. Kallen adopted from William James and ethnicized). Others have added such terms as 'ethnogenesis' or 'ethnic differentiation' to Gordon's helpful array of terms.

Devereux (1975) took as his point of departure the question how ethnic identity can be related to other forms of identity. He emphasized what he called the 'dissociative' nature of ethnic behavior that was not actually prompted by past tradition but was informed by a contrastive and presentist strategy of opposition of one ethnic group against another. Ethnicity, seen this way, 'is logically and historically the product of the assertion that 'A is an X because he is not a Y'—a proposition which made it remarkably easy to identify and homogenize Xness, exaggerating the differences between the Xs and Ys. 'X does not equal Y' was thus the fundamental ethnic formula.

Writing after World War II, Devereux located the fascist potential of ethnic movements in this reduction of complex authentic identity to contrastive one-dimensionality in the name of ethnicity. 'Ethnicity' could demand that it overrule all other group affiliations with which it might be in conflict. Devereux saw the danger of ethnicity as an 'all-or-nothing proposition,' for in the complex ensemble that constituted

individual identity, ethnicity could turn out to be one feature that is superimposed—Devereux calls it 'hypercathected'—over all other features of a personality, including the important characteristic of shared membership in the human race. This was a strange logical operation that could produce catastrophic and disastrous actions—including ethnic cleansing and genocide.

In his important essay on *Symbolic Ethnicity: The Future of Ethnic Groups and Cultures in America*, Gans (1979) called attention to the ways in which modern ethnic identification may work by external symbols rather than by continuous activities that make demands upon people who define themselves as 'ethnic' so that a gentle ethnic diversification could take place in an assimilative framework.

These observations were important for approaches that looked at ethnic processes and conflicts not only as residues of primordial differences, but as new constructions and inventions. Three widely received books supported this direction of inquiry, Berger and Luckmann's (1966) general *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*, Hobsbawm and Ranger's (1983) *Invention of Tradition*, focusing especially on Europe and Africa, and South Asia specialist Anderson's (1983) *Imagined Communities: Reflections on the Origins and Spread of Nationalism*. The constructivist turn opened up the whole interrelationship of ethnicity and culture to new inquiries, as the focus shifted from the issue of how ethnic culture was preserved or succumbed to assimilation to the question of how ethnic groups emerged and how ethnicity was constantly reimagined and reinvented.

In his landmark introduction to *Ethnic Groups and Boundaries*, Barth (1969) drew comparatively on wide-ranging field work on such groups as the Pathans in Swat, the Yao in China, or the Lapps in Norway in order to develop an influential approach to ethnicity. He suggested that we 'ask ourselves what is needed to make ethnic distinctions emerge in an area.' Barth argued:

> When one traces the history of an ethnic group through time, one is not simultaneously, in the same sense, tracing the history of "a culture": the elements of the present culture of that ethnic group have not sprung from the particular set that constituted the group's culture at a previous time, whereas the group has a continual organizational existence with boundaries (criteria of membership) that despite modifications have marked off a continuing unit.

Barth saw the essence of ethnicity in mental, cultural social and not necessarily territorial boundary-constructing processes which functioned as cultural markers between groups. For Barth it was 'the ethnic *boundary* that defines the group, not the cultural stuff that it encloses.' Barth directed attention to processes of boundary maintenance and argued that 'members of all ethnic groups in a poly-ethnic society' could 'act to maintain dichotomies and differences.'

In *Urban Ethnicity*, Cohen (1974) investigated the emergence of ethnic distinctions against the background of power relations and insisted on ethnicity as a category that may apply to dominant elites as well as marginal groups, with examples ranging from Sierra Leone to the city men of London.

The conceptual richness of the pioneering scholarship and of provocative work from the 1960s to the early 1980s supported a growing body of empirical work on ethnic groups around the globe. Questions of ethnic persistence and emergence, of power relations and agency, of assimilation and pluralism, of marginality, ethnic invention, symbolic ethnicity, and boundary construction, permeated scholarship in many disciplines at a time in which intellectual attention as riveted by issues of ethnicity, class, and gender. The proliferation of scholarly work reached such proportions that by 2000, there were hundreds of books whose titles contained the relatively new word 'ethnicity.'

4. *Exemplary Scholarly Engagements with Ethnicity*

Starting in the 1970s, studies in many disciplines placed a new emphasis on ethnicity. Emigration and immigration, slavery and freedom, and ethnic mergers and conflicts had been the subjects of many earlier historical and sociological works. But such topics now assumed a new position of centrality in the various disciplines. In the US, for example, the pioneering work by Marcus Lee Hansen, Oscar Handlin, or John Higham inspired many historians—among them Rudolph Vecoli, John E. Bodnar, Kathleen Conzen, and Matthew Jacobson, to turn to topics of immigration or related to the ethnic heterogeneity of the US. Of all American ethnic groups, African Americans have particularly been the subject of historical research, ranging from the early work by Carter G. Woodson, W. E. B. Du Bois, and John Hope Franklin to scholarship by Nathan I. Huggins, Leon Litwack, or David Levering Lewis.

The *Journal of American Ethnic History* is one of many specialized journals that have accompanied the establishment of black and ethnic studies departments in the US. Thernstrom and Orlov's (1980) *The Harvard Encyclopedia of American Ethnic Groups* gives a sense of the whole field, both in its thematic entries and in its accounts of 107 American ethnic groups. In Germany, a handbook such as *Ethnische Minderheiten in der Bundesrepublik Deutschland: Ein Lexikon*, (Schmalz-Jacobsen and Hansen 1995) offered an overview of German ethnic groups and issues, with a certain focus on such linguistic minorities as Danes and Sorbs. Noiriel, *Le creuset français: Histoire de l'immigration,*

XIXe–XXe siècles (1988) offered an account of emigration and immigration in France. In *Ethnic Groups in Conflict*, Horowitz (1985) analyzed explosive ethnic fault lines around the globe.

5. Conclusion

While ethnic identification seems fluid and ready to emerge unpredictably in new circumstances, it is also remarkable how many areas of the world have been enmeshed in intermittent ethnic conflicts for a very long time. As long as the sense of belonging—be it ancient or modern in origin—persists among members of groups and as long as conflicts arise between different groups, there will be ample material for ethnic scholarship.

See also: Ethnic Cleansing, History of; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Ethnonationalism: Cultural Concerns; Eugenics, History of; Race: History of the Concept; Race Identity; Racism, History of; Simmel, Georg (1858–1918); Weber, Max (1864–1920)

Bibliography

Anderson B 1983 *Imagined Communities: Reflections on the Origins and Spread of Nationalism*. Verso, London

Barth F 1969 *Ethnic Groups and Boundaries*. Bergen and Little, Brown & Co., Boston

Berger P L, Luckmann T 1966 *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Doubleday, Garden City, NY

Cohen A 1974 *Urban Ethnicity*. Tavistock Publications, London

Devereux G 1975 Ethnic identity: its logical foundations and its dysfunctions. In: de Vos G, Romanucci-Ross L (eds.) *Ethnic Identity: Cultural Continuities and Change*. Mayfield, Palo Alto, CA

Erikson E 1950 *Childhood and Society*. Norton, New York

Gans H J 1979 Symbolic ethnicity: the future of ethnic groups and cultures in America. In: *On the Making of Americans: Essays in Honor of David Riesman*. University of Pennsylvania Press, Philadelphia, PA

Glazer N 1975 The universalization of ethnicity: peoples in the boiling pot. *Encounter* **44**(2): 8–17 [reprinted in Glazer N 1983 *Ethnic Dilemmas 1964–1982*. Harvard University Press, Cambridge, MA, London, pp. 233–53

Glazer N, Moynihan B P 1975 *Ethnicity: Theory and Experience*. Harvard University Press, Cambridge, MA

Gordon M M 1964 *Assimilation in American Life: The Role of Race, Religion, and National Origin*. Oxford University Press, New York

Greeley A M 1974 *Ethnicity in the United States*. John Wiley, New York, London

Handlin O 1951 *The Uprooted: The Epic Story of the Great Migrations That Made the American People*. Little Brown, Boston

Hirschfeld M 1938 *Racism*. Victor Gollancz, London

Hobsbawm E, Ranger T 1983 *Invention of Tradition*. Cambridge University Press, Cambridge, UK

Horowitz D L 1985 *Ethnic Groups in Conflict*. University of California Press, Berkeley, CA

Journal of American Ethnic History 1981 Transaction Periodicals Consortium, New Brunswick, NJ

Lemkin R 1944 *Axis Rule in Occupied Europe. Laws of Occupation, Analysis of Government, Proposals for Redress*. Proposals for Redress, Carnegie Endowment for International Peace, Washington, DC

Noiriel G 1988 *Le creuset français: Histoire de l'immigration, XIXe–XXe siècles*. Seuil, Paris

Park R E 1928 Human migration and the marginal man. *American Journal of Sociology* **33**(6): 881–93

Riesman D 1953 Some observations on intellectual freedom. *American Scholar* **23**: 9–25

Schmalz-Jacobsen C, Hansen G (eds.) 1995 *Ethnische Minderheiten in der Bundesrepublik Deutschland: Ein Lexikon*. C. H. Beck, Munich

Seligman E R A, Johnson A 1930–1935 *Encyclopedia of the Social Sciences*. Macmillan Company, New York

Sills E 1968 *International Encyclopedia of the Social Sciences*. Macmillan and Free Press, New York

Simmel G 1908 In: Hughes E C (ed.) *Conflict and The Web of Group Affiliations*. Macmillan and Free Press, New York

Spitzer L 1933 Wortgeschichtliches. *Zeitschrift für Romanische Philologie* **53**(3–4): 300–1

Stolcke V 1994 Invaded women: gender, race, and class in the formation of colonial society. In: Hendricks M, Parker P (eds.) *Women, "Race," and Writing in the Early Modern Period*. Routledge, London and New York

Thernstrom S, Orlov R (eds.) 1980 *The Harvard Encyclopedia of American Ethnic Groups*. Harvard University Press, Cambridge, MA and London

Warner W L, Lunt P 1941 *The Social Life of a Modern Community*, Yankee City Series, Vol. 1. Yale University Press, New Haven, CT

Weber M 1978 In: Roth G, Wittich C (eds.) *Economy and Society*. University of California Press, Berkeley, CA

W. Sollors

Ethnic Identity/Ethnicity and Archaeology

Many would think that the study of ethnicity is a recent topic in archaeology but this is only partially true. Archaeology has always been involved in the identification of ethnic groups in the past, but it has done this using different terms. In the early years of archaeology ethnic groups were called 'peoples' and 'nations.' These terms were later substituted for that of 'archaeological cultures' from the start of the twentieth century (Díaz-Andreu 1996). It is only since the 1980s that the growing influence of anthropology upon archaeology has persuaded the latter to begin to talk about ethnicity and ethnic groups.

Ethnicity will be defined here as that aspect of a person's self-conceptualization and his or her conceptualization by other individuals that results from identification with one or more broader groups in opposition to others, on the basis of perceived cultural differentiation or common descent (Jones 1997). Not all archaeologists, however, would agree with this definition. It does not focus on the ethnic group in itself as something objective and describable, as many archaeologists would like. Far from it, the definition provided here stresses perception and, although it does not intend to deny commonalities among the members of the same group, it does not consider them essential for the characterization of an ethnic group.

The definition above considers ethnic identity as multidimensional, for it argues that individuals can identify with one or more broader groups. It proposes, therefore, that multiple ethnic affiliations may coexist and overlap in the same subjects. Each person is (or is potentially) associated with various groups that would enter into the definition of ethnicity. A Mancunian, for example, can feel him/herself as Mancunian, as a person from Lancashire, as English, British, European or a world citizen. However, because of our political education, if asked which ethnic identity they have, individuals will provide a political answer, for only one ethnic identity will be given, usually the one that coincides with their national identification. This reaction has been expected from each citizen since the end of the eighteenth century. However, this might not necessarily be what people feel and what people felt in the past.

Two more aspects may be added to the proposed definition of ethnicity. The first is that ethnic identity can be described as situational. Each of these complementary and enriching ethnic identities, can individually come into play in various situations, depending on where people are, what the situation is and what political ideology they hold. Each individual is active regarding his/her ethnic identities and the daily negotiations will entail that one or several of them are made manifest. Secondly, ethnic affiliations are fluid. They are and have been in a continuous process of flux, as is reflected by the appearance and disappearance of group names throughout the several centuries referred to in written sources.

Of the two theoretical approaches currently predominant in the study of ethnicity, the primordialist and the instrumentalist, the conceptualization of ethnicity provided above fits with the latter. The primordialist perspective considers ethnicity as something inherent, acquired by birth. Thus, the instinctive attachment of the individual to his or her territory and culture is emphasized and the ethnic group is regarded as the natural universal unit of human organization. The instrumentalist perspective, however, views ethnicity as a fluid and situational group identity. It is through practice that individuals perceive and negotiate their ethnic allegiances. This perspective stresses the social aspect of ethnicity as a regulatory device for social behavior and social organization within each ethnic group and between different ethnic groups. Archaeology has traditionally perceived ethnic identity from a primordialist stance and it is only in the last two decades that the instrumentalist viewpoint has made its appearance.

Archaeological discussions of ethnicity have a long history. Throughout much of archaeology's history, material culture has been understood in a very straightforward manner as a basis for inferring ethnicity. The presence of specific types of material culture and the absence of others have been considered as diagnostics to define groups. For example, the characteristic prehistoric incised beaker pot was seen to indicate the presence of the Beaker culture in an area and, similarly, the presence of Samian pottery attested to the presence of Romans. This way of looking at ethnicity has been dominant from the early days of archaeology, although it was only clarified theoretically at the turn of the nineteenth century and early twentieth century with the culture historical approach. Culture history first appeared in Austro-German anthropology (Zwernemann 1983). It was adopted in archaeology a decade later, mainly through the work of the German archaeologist Gustaf Kossinna (Trigger 1980, p. 44). The pre-eminence of German archaeology at this time favored the spread of the theory elsewhere. However, whereas anthropology dropped the term 'culture' in the 1920s, and instead began to use 'tribe,' finally substituting it by 'ethnic group' (Jenkins 1997), archaeology has only very recently been rethinking its use of 'culture' and beginning to explicitly discuss, problematize, and theorize ethnicity.

Within the culture historical perspective ethnic groups were called cultures (or civilizations, in the case of nonprimitive societies). The first definition of culture was proposed by Childe. As he put it, 'we find certain types of remains—pots, implements, ornaments, burial rites, house forms—constantly recurring together. Such a complex of regularly associated traits we shall term a 'cultural group' or just a 'culture' (Childe 1929, p. v–vi) (see *Classification and Typology (Archaeological Systematics)*). In the views on culture/ethnicity of Childe and other culture historians there was no place for fluidity and negotiation, neither was there a place for perception. One could say that, in practice, archaeologists considered peoples almost as a byproduct of objects. For example, changes in material culture were often interpreted as the result of invasions, i.e., as the substitution of one ethnic group by another. The following statement is typical of the archaeology of the period:

> These two groups of hoards [in the British Late Bronze Age] may be taken to indicate first, definite invaders … and second, their fusion with the inhabitants and the absorption, modification, or discarding of their exotic types (Kendrick & Hawkes 1932, p. 135).

Despite the emphasis on invasions and migrations, individuals (who were presumably involved) were perceived as passive spectators of the more important process. Hence, Childe, when discussing the archaeology along the river Danube could say: 'whether these Moravian peasants actually came themselves up the Danube is immaterial; their culture did' (Childe 1927, p. 85).

Culture history archaeologists viewed cultures or ethnic groups as organic beings, as living personalities. Their features were conceptualized like those of an individual. A culture was born, developed, and eventually declined, before moving on to another cycle of emergence, rise, and decay. The impact of these ideas are evident in the quasi-obsession with origins at this time, exemplified by such titles as *The Dawn of European Civilization* (Childe 1925) and *The Beginnings of Egyptian Civilization* (Brunton 1929).

In many ways processual archaeology or New Archaeology continued some of the trends seen in culture history into the 1960s and 1970s (Jones 1997, pp. 107–9; see *Theory in Archaeology*). The aim of processual archaeology was to develop methods to understand the formation of the archaeological record and to refine approaches to inferring past behavior from observations of that record. This led to a growing interest in, for example, site formation processes (see *Geoarchaeology*), ethnoarchaeology (see *Ethnoarchaeology*) and the use of analogy. There was thus an emphasis on the use of a rigorous hypothetico-deductive method (involving the testing of models and the use of inferential generalizations about the past). Within this perspective, culture was seen as a system of interrelated components (Binford 1968, pp. 16–23). It was, however, later to be criticized, in the same way that functionalist anthropology was, for its rather fixed views of human behavior.

As distinct from archaeology, the subjectivist or emic perspective for the study of ethnicity was increasingly being accepted in anthropology, at least beginning in the 1970s (Jones 1997). The general recognition of the role of perception in the definition of identity, and in particular of ethnicity, is partly a consequence of the influential work of Fredrik Barth. Barth (1969) radically changed the study of ethnicity with his proposal that ethnicity was something that could not be objectified, but was defined by diagnostic socially relevant factors for membership and the (successful) ascription of the members to their ethnic group. Although he pointed out that all ethnic groups have their own overt signals (diacritical features that people look for and exhibit to show identity), together with basic value orientations (the standards of morality and excellence by which performance is judged), he considered that the critical features that defined ethnicity were self-ascription and ascription by others, and that there was 'no simple one-to-one relationship between ethnic units and cultural similarities and differences. The features that are taken into account are not the sum of "objective" differences, but only those which the actors themselves regard as significant' (Barth 1969, pp. 13–14). Barth's proposal ruled out the possibility of there being an objective and objectified way to designate ethnicity, and to use material culture as a critical factor in the study of ethnicity, and instead perception took on an essential role. Barth's ideas were later refined by other authors. Among them the work by Ronald Cohen should be emphasized. Cohen (1978) went even further than Barth when he stressed the situational and fluid character of ethnicity, whose influence was only possible as far as it was a meaningful element in social interaction. Although neither Barth nor Cohen, nor any of the other authors arrived at the contention expressed above—that ethnic identity is multidimensional, and it is likely that multiple ethnic identifications coexist in the same person—Cohen (1978) almost reached this conclusion. Some of his examples clearly pointed in that direction. He even stated that 'the same person can be categorized according to different criteria of relevance in different situations' (Cohen 1978, p. 388), but instead of conceding that more than one of these characterizations could be ethnic, he continued to consider ethnicity as monolithic.

Barth and Cohen were not dealing with prehistoric peoples, and never considered the issue of whether ethnicity was an appropriate concept to be applied to the past. There have been some claims that the concept is not applicable to precapitalist societies. Eriksen (1993, p. 80), for example, argued that outside Europe ethnicity appeared only after peoples' contact with capitalism. There is some truth in Eriksen's view, as some new ethnic groups were indeed created by colonizers (Cohen 1978, p. 383, Gulliver 1969, p. 15), or even by anthropologists (Cohen 1978, pp. 380–3). However, colonizers seem to have created these new ethnic groups on the basis of pre-existent ones, either amalgamating, separating or reorganizing them.

Colonizers may not have been successful in all their new creations, but in some they undoubtedly were. In the case of anthropologists, Cohen acknowledges that some of them had to provide a name for a 'tribe' (a concept which Cohen sees as equivalent to ethnic unit) 'even when the group faded imperceptibly into other named groups more or less similar and was broken up into named subgroupings that had strong we/they feelings dividing them.' (Cohen 1978, pp. 38–1.) Following the definition of ethnicity provided above, one should interpret these anthropological categorizations, on the one hand, as reflecting the multiple levels of ethnic affiliations among these groups. On the other, they should be understood as an example of how anthropologists themselves used their own sense of a 'we/they' division in order to classify 'the other' in their own (nationalistic) language. Their nationalistic perspective would lead them to deny all groupings as ethnic and hence to the need to decide which of them was the only one—the proto-ethno-nation—that

should be considered as their object of study. As was indicated above, nationalism has left a definite mark on the way that ethnicity has been understood, mainly with respect to its one-dimensional characterization.

In archaeology the first major break with the processual paradigm occurred only in the early 1980s, much later than in anthropology, to which it was clearly related. The seminal book in this process was Ian Hodder's *Symbols in Action* (Hodder 1982b), which was, not coincidentally, an ethnoarchaeological study (see *Ethnoarchaeology*). Hodder depicted the manipulation of material culture by several groups in East Africa as active in creating social relationships, and stressed the need to find ways of interpreting the meaning(s) of material culture. Wider discussions of this topic were, however, well underway by then, as shown by many of the contributions to *Symbolic and Structural Archaeology* (Hodder 1982a). The major aim of this body of research was the study of living cultures 'in order to shed some light on the analysis and interpretation of cultures in prehistoric archaeology'(Hodder 1982a, p. 1). Interestingly, Hodder's ethnoarchaeological work originated from an overtly processual hypothesis, which was to test the relationship between resources, interaction and cultural similarity (1982a, p. 8), and it was assumed that cultural similarity would 'reflect' degrees of interaction (Hodder 1982a, p. 9). However, observations made during Hodder's ethnographic fieldwork, however, showed that while the use and distribution of certain items such as drinking cups, wooden eating bowls and stools followed the expected patterns, those of other items, such as spears and calabashes, did not. Hodder concluded that 'the extent to which cultural similarity relates, for example, to interaction, depends on the strategies and intentions of the interacting groups and how they use, manipulate and negotiate material symbols as part of those strategies' (Hodder 1982a, p. 185). These interacting groups were ethnic (tribal), but other subgroups within those societies, such as younger men and women, were also seen to follow their own strategies in their relations with the dominant older men.

Despite Hodder's important contribution on ethnicity, archaeology remained reluctant to analyse this issue of ethnic identity for several years. In 1989 Shennan published a much-cited article in this area but in it he referred mainly to culture instead of ethnicity (Shennan 1989). This was not the case in the Polish journal, *Archaeologia Polona*, which, in 1991, dedicated an entire issue to ethnicity and archaeology. The articles published here, together with others that have subsequently appeared (e.g., Dìaz-Andreu 1998, Dietler 1994, Funari 1999), show that the debate was becoming more intense. This process was further fostered by the 1997 publication of Jones's book, *The Archaeology of Ethnicity*, which provides archaeologists with an excellent overview of the state of the art.

See also: Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Ethnoarchaeology; Identity: Social

Bibliography

Barth F 1969 Introduction. In: F Barth (ed.) *Ethnic Groups And Boundaries: The Social Organization Of Culture Difference*. Universitetsforlaget, Oslo

Binford L R 1968 Archaeological perspectives. In: Binford S R, Binford L R (eds.) *New Perspectives in Archaeology*. Aldine, Atherton Chicago, New York, pp. 1–32

Brunton G 1929 The beginnings of Egyptian civilization. *Antiquity* **3**: 456–67

Childe V G 1925 *The Dawn of European Civilization*. Kegon Paul, London

Childe V G 1927 The Danube thoroughfare and the beginnings of civilization in Europe. *Antiquity* **1**: 79–91

Childe V G 1929 *The Danube in Prehistory*. Clarendon Press, Oxford, UK

Cohen R 1978 Ethnicity: problem and focus in Anthropology. *Annual Review of Anthropology* **7**: 379–403

Díaz-Andreu M 1996 Constructing identities through culture. In: Graves-Brown P, Jones S, Gamble C (eds.) *Cultural Identity and Archaeology*. Routledge, London, pp. 48–61

Díaz-Andreu M 1998 Ethnicity and Iberians. The archaeological crossroads between perception and material culture. *European Journal of Archaeology* **2**: 199–218

Dietler M 1994 Our ancestors the Gauls: Archaeology, ethnic nationalism and the manipulation of Celtic identity in modern Europe. *American Anthropologist* **96**: 584–605

Eriksen T H 1993 *Ethnicity and Nationalism. Anthropological Perspectives*. Pluto, London

Funari P P A 1999 Etnicidad, identidad y cultura material: Un estudio del Cimarrûn Palmares, Brasil, Siglo XVII. In: Zarankin A, Acuto F A (eds.) *Sed non Satiata. Teoría Social en la Arqueología Latinoamericana Contemporānea*. Ediciones del Tridente, Buenos Aires, Argentina, pp. 77–96

Gulliver P H 1969 Introduction. In: Gulliver P H (ed.) *Tradition and Transition in East Africa: Studies of the Tribal Element In the Modern Era*. Routledge & Kegan Paul, London, pp. 5–38

Hodder I (ed.) 1982a Symbolic and structural archaeology *New Directions in Archaeology*. Cambridge University Press, Cambridge, UK

Hodder I 1982b *Symbols in Action. Ethnoarchaeological Studies of Material Culture*. Cambridge University Press, Cambridge, UK

Jenkins R 1997 *Rethinking Ethnicity: Arguments and Explorations*. Sage Publications, London

Jones S 1997 *The Archaeology of Ethnicity*. Routledge, London

Kendrick T D, Hawkes C F C 1932 *Archaeology in England and Wales*. Methuen, London

Shennan S 1989 Introduction: archaeological approaches to cultural identity. In: Shennan S (ed.) *Archaeological Approaches to Cultural Identity*. Unwin Hyman, London, pp. 1–32

Trigger B G 1980 *Gordon Childe. Revolutions in Archaeology*. Thames and Hudson, London

Zwernemann J 1983 *Culture History and African Anthropology. A Century of Research in Germany and Austria.* Uppsala Studies in Cultural Anthropology 6, University of Uppsala, Uppsala, Sweden

M. Díaz-Andreu

Ethnic Identity, Psychology of

Ethnic identity is a dynamic, multidimensional construct that refers to one's identity, or sense of self, as a member of an ethnic group. An ethnic group can be thought of as a subgroup within a larger context that claims a common ancestry and shares one or more of the following elements: culture, race, religion, language, kinship, and place of origin. Ethnic identity is a central defining characteristic of many individuals, particularly those who are members of minority or lower status groups (Liebkind 1992, Phinney 1990).

Ethnic identity has traditionally been most salient in immigrant-receiving countries like the USA and Australia, but it has become an increasingly important issue throughout the world, as social and political changes have increased the amount of contact among people from different ethnic groups and, in some cases, have led to ethnic conflict. Social scientists have attempted to understand ethnic identity from a variety of perspectives. This article examines the development of ethnic identity and the ways in which ethnic identity changes over time and contexts. It concludes with a discussion of the implications of ethnic identity.

1. *The Formation of Ethnic Identity*

1.1 *The Beginnings of Ethnic Identity*

An ethnic identity is developed and modified as individuals become aware of other groups and of the ethnic differences between themselves and others, and attempt to understand the meaning of their ethnicity within the larger setting. Ethnic identity may have its origins in the infant's attachment to caregivers and the fear of strangers that is experienced early in life. The family provides the initial bases for feelings about what is normal and comfortable, and these feelings are generally supported in the immediate community. In homogeneous settings, familiar practices are not recognized as ethnic.

Once children encounter others who are different in appearance, language, behaviors, or attitudes, they begin to categorize themselves and others. Frances Aboud (1987) has shown that children learn their ethnic self-label between the ages of four and seven years, although the age varies with the group and the amount of contact with other groups. By eight or ten years of age, children develop an understanding of ethnic constancy (the fact that their ethnicity does not change over time or superficial changes). Cognitive developmental level influences the way in which children understand ethnicity. Young children understand ethnicity in literal and concrete terms, defined by food, customs, and language. With increasing cognitive competence, children begin to develop a group consciousness and to understand ethnicity in terms of norms and values (Quintana 1998).

Socialization plays an important role in the content and meaning children attach to their own ethnicity. Children acquire knowledge about their ethnic culture through the process of enculturation (Bernal and Knight 1993). Children's feelings about their ethnicity are influenced from an early age by the family and community. When families provide strong positive images of their ethnicity, children's early feelings about their ethnic group are likely to be positive. A vital ethnic community also provides a context in which children can form a positive sense of their group.

Children are influenced as well by messages from other groups and the larger society. When a group is held in lower esteem or is disparaged by others, these negative messages may become internalized. Children may then hold conflicting or negative feelings about their ethnicity and express the desire to belong to another group. However, in childhood ethnic identity remains largely unexamined; that is, children have not thought through for themselves the meaning of their ethnic group membership (Phinney 1989).

1.2 *The Search for a Meaningful Ethnic Identity*

Beyond childhood, the process of ethnic identity development becomes more complex. During adolescence, identity formation is a critical developmental task, particularly in complex modern societies. Like ego identity, the process of ethnic identity has been conceptualized in terms of stages, with an individual moving from the unexamined attitudes of childhood, through a moratorium or period of exploration, to a secure achieved ethnic identity at the end of adolescence (Phinney 1989). The moratorium period begins for adolescents as a result of their increasing cognitive ability, their growing understanding of social and historical processes, and in some cases, personal experiences of discrimination. As ethnic identity becomes a salient issue, many youths, especially those from ethnic groups with lower status or power, question the meaning of their ethnicity and its place in the larger context (Phinney 1989).

This search for a meaningful ethnic identity can lead to a wide range of emotional reactions, both positive and negative. Individuals may become deeply involved in learning about their ethnicity. This process may lead to exaggerated claims about their group. It may

result in constructive and creative actions aimed at affirming the value and legitimacy of one's group and establishing a positive basis for a group identity (Tajfel and Turner 1986).

Conversely, the process of examining and dealing with the social implications of one's ethnicity may lead to feelings of insecurity, confusion, or anger. Individuals who are confused or ambivalent about their ethnic identity are likely to react more strongly to perceived threats to their ethnicity (Worchel 1999). They may experience strong negative feelings toward other groups, particularly in cases where a group has been the victim of negative or exploitative intergroup behaviors or where group members feel that their needs and aspirations are being thwarted. Such attitudes can lead to violence when legitimate means of ethnic affirmation are seen as ineffective. Much interethnic violence results from feelings of threat to one's ethnic identity. This process is assumed to be relevant primarily to disadvantaged minority groups, but it can apply as well to members of a dominant group who feel threatened by changing demographics, for example as a result of large-scale immigration.

1.3 The Achievement and Internalization of Ethnic Identity

The optimal outcome of the ethnic identity formation process is the achievement of a secure and confident sense of one's ethnicity. This mature sense of self as an ethnic group member is assumed to include positive feelings about one's group and to be a source of personal strength and positive self-evaluation (Phinney 1989, Tajfel and Turner 1986). Feeling secure about one's own ethnic identity is also assumed to be associated with more positive attitudes toward other groups. An achieved ethnic identity may be related to the ability to assume the perspective of other groups, to adopt a multicultural perspective, and to see the place of one's own group in a larger perspective.

2. Ethnic Identity Across Time and Contexts

Although developmental processes underlie the formation of ethnic identity, ethnic identity may be renegotiated throughout life in response to individual, contextual, and historical changes. Ethnic identity varies in the short term over differing contexts, with strong ethnic feelings emerging in settings where ethnicity is highly salient, such as traditional ethnic celebrations, and receding in settings which de-emphasize ethnicity. Many bicultural individuals experience variation in the strength of their group identities as they move between ethnic and nonethnic contexts, such as home, school, and workplace, and change their language and behaviors to suit the context.

2.1 Ethnic Identity and Immigration

Group identity issues are raised most dramatically when individuals and groups move across national and cultural borders, or when political borders change around stable groups. As migrants, refugees, guest workers, and other travelers come in contact with other cultures, questions of identity are raised both for the mobile groups and for the societies where they settle. These processes are highly dependent on the environmental context, including the size and structure of ethnic communities, the history and current status of ethnic groups in the larger society, and attitudes of other members of the society toward diversity.

For immigrants, changes in ethnic identity over time depend on a number of influences, including age at time of immigration and generation of immigration (that is, whether one is an immigrant, child of immigrants, grandchild of immigrants, etc.). Adult immigrants typically retain a strong identity with their culture of origin without identifying strongly as members of the larger society. Children who migrate at a young age adapt more quickly than their parents to a new setting and are likely to become bicultural, identifying with both the ethnic group and the larger society.

2.2 Retention of Ethnic Identity

For later generations of immigrants, that is, the children and grandchildren of immigrants, and for stable ethnic groups within larger societies, the retention of ethnic identity depends both on factors within the ethnic group and on the broader societal or historical context. Ethnic group members differ in the extent to which they desire to retain their culture and the degree to which they wish to become part of the larger society. Societies differ widely in their relations with minority groups, their policies toward ethnic groups, and the attitudes of members of society toward immigrants and minorities.

As a result of these varied influences, ethnic groups and their individual members may adopt different group identity modes, similar to modes of acculturation (Berry 1990). When individuals or groups wish to retain their ethnic identity and maintain their ethnic language and customs, and when ethnic institutions support such activities, ethnic identity can remain strong over long periods. If they also develop a sense of belonging to the larger society, they become bicultural or integrated. If they retain a strong ethnic identity and do not become part of the larger society, either by choice or because of nonacceptance, they may be described as separated. On the other hand, individuals and groups who give up their ethnic identity and become identified with the larger society are assimilated. Finally, those who give up their ethnic

identity but do not become part of the larger society are seen as marginalized.

The ways in which these patterns play out over time is highly variable. Ethnic communities may be maintained over many generations and can be revitalized when circumstances permit. Groups that are culturally or phenotypically different from the larger society, or that experience prejudice and discrimination, may remain separated over many generations, as in the case of gypsies in Europe or African Americans in the USA. For these groups, ethnic identity typically remains strong and salient over time. In contrast, when distinctive cultural or phenotypic markers are absent and when there is little desire to maintain one's ethnic culture, assimilation is likely in the third and later generations, as has been the case for most European Americans in the USA (Waters 1990).

3. *Implications of Ethnic Identity*

Because of the centrality of ethnic identity for group members, together with social and political processes that support or threaten such an identity, ethnic identity can have positive or negative implications for the individual and for the larger society.

Most psychological research has shown that ethnic identity is positively associated with psychological well-being, but the relationship is complex, and the impact is likely to depend on the particular aspect of ethnic identity being considered. A widely used measure of ethnic identity (Roberts et al. 1998) suggest two distinct components of ethnic identity: an affective component, comprising a sense of belonging and positive feelings about one's group; and a developmental component, involving the extent to which one has explored and resolved ethnic identity issues (Phinney 1990). The affective component shows a consistent though modest correlation with self-esteem; those who feel more positive about their own ethnicity have higher self-esteem and lower levels of depression. Conversely, internalized negative feelings about one's own group are related to negative feelings about oneself. The developmental component has a weaker relationship with psychological well-being, perhaps because of the variability of attitudes that occur during development. Nevertheless, the attainment of a secure and confident identity appears to be the basis of positive feelings about oneself and others.

Feelings of belonging to the larger society also contribute to well-being. Research with immigrants has shown a bicultural identity to be the most adaptive, and a marginal identity to be the least adaptive (Berry 1990). When individuals and groups have the freedom to express their values and practice their ethnic customs, ethnic identity can be the basis of powerful positive feelings. Societies that thwart the desire of ethnic group members to affirm and express their group identity may experience turmoil and violence (Worchel 1999).

The study of ethnic identity is relatively recent in psychology. The knowledge that we have is based on a limited amount of research, and there are many areas in need of study. Longitudinal studies are needed to provide deeper insight into the processes by which ethnic identity changes over time. Links to other aspects of identity need to be explored, to determine the extent to which ethnic identity is similar to or different from national, racial, cultural, or other group identities. Research has focused on relatively few groups and situations; future research should extend the range of groups studied and the national and regional settings that provide contexts for ethnic identity development and expression. Increased collaboration among researchers from different countries would be valuable in exploring whether the current findings apply to the wide diversity of ethnic groups in the world. Finally, because the study of ethnic identity remains fragmented, there is need for conceptual models that can integrate research across different settings. Because of the growing importance of ethnicity throughout the world, it is imperative for the behavioral sciences to gain a better understanding of ethnic identity and its implications for society.

See also: Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Gender, Class, Race, and Ethnicity, Social Construction of; Social Psychology: Research Methods; Urban Ethnography

Bibliography

Aboud F 1987 The development of ethnic self-identification and attitudes. In: Phinney J, Rotheram M (eds.) *Children's Ethnic Socialization: Pluralism and Development*. Sage, Newbury Park, CA, pp. 32–55

Bernal M, Knight G (eds.) 1993 *Ethnic Identity: Formation and Transmission Among Hispanics and Other Minorities*. State University of New York Press, Albany, NY

Berry J 1990 Psychology of acculturation. In: Berman J (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, vol. 37, pp. 201–34

Liebkind K 1992 Ethnic identity: Challenging the boundaries of social psychology. In: Breakwell G (ed.) *Social Psychology of Identity and the Self-concept*. Academic Press, London, pp. 147–85

Phinney J 1989 Stages of ethnic identity development in minority group adolescents. *Journal of Early Adolescence* **9**: 34–49

Phinney J 1990 Ethnic identity in adolescents and adults: A review of research. *Psychological Bulletin* **108**: 499–514

Quintana S 1998 Children's developmental understanding of ethnicity and race. *Applied and Preventive Psychology* **7**: 27–45

Roberts R, Phinney J, Masse L, Chen Y, Roberts C, Romero A 1998 The structure of ethnic identity in young adolescents from diverse ethnocultural groups. *Journal of Early Adolescence* **91**: 301–22

Tajfel H, Turner J 1986 The social identity theory of intergroup behavior. In: Worchel S, Austin W (eds.) *Psychology of Intergroup Relations*. Nelson-Hall, Chicago, IL, pp. 7–24
Waters M 1990 *Ethnic Options: Choosing Identities in America*. University of California Press, Berkeley, CA
Worchel S 1999 *Written in Blood: Ethnic Identity and the Struggle for Human Harmony*. Worth, New York

J. S. Phinney

Ethnicity: Anthropological Aspects

The concept of 'ethnicity' derives from the ancient Greek *ethnos*, referring to human collectivities living and acting together. It is typically translated today as 'people' or 'nation.' Since the 1900s, the linked concepts of ethnicity and ethnic group have passed into everyday discourse. They have become central to the politics of group differentiation and advantage in the culturally diverse social democracies of Europe and North America. With notions of 'race' in disrepute since 1945, ethnicity has become a rallying cry in the bloody re-ordering of the post-Cold War world: 'ethnic cleansing' stands shoulder to shoulder with earlier euphemisms such as 'racial hygiene' and 'the final solution.' In the markets of the affluent west and north, to call a commodity 'ethnic' is to imply the value of 'authenticity.' In the social sciences, ethnicity—including related topics such as 'race,' 'nationalism,' and 'difference'—has become a rapidly expanding specialism.

1. Corporate Groups and Cultural Differences

One of anthropology's earliest ambitions as a comparative scientific enterprise was to explain and understand human diversity. This aspiration has remained central to the social and cultural anthropological imagination—and to the limited, but significant, unity of purpose that divergent national and intellectual anthropological traditions have succeeded in maintaining—ever since. The different approaches to this quest which have characterized the discipline's twentieth-century history have, for most of that period, been rooted in two interrelated anthropological axioms. On the one hand, there is an understanding of human social life as based in membership in definite, bounded corporate groups. On the other, the basic unity of humankind is understood as crisscrossed by patterns of cultural differentiation.

These fundamental concepts emphasize similarity (corporate groups) and difference (cultural differentiation), the two basic dimensions of all processes of social identification. In groups, we are both together (as us) and apart (from them). Although the two notions, and their implication in each other, remain constant, how they are defined, and the relationship between them, has changed since the nineteenth century. This process can be understood as a gradual shift from documenting the social structure and customs of 'tribes,' to attempting to understand the complex dynamics of 'ethnic groups.'

During the colonial and immediately postcolonial eras, anthropology's theoretical preoccupation with corporate groups was manifest in the assumption that 'primitive peoples' were characteristically organized into tribal groups. In one move, the concept of 'the tribe' distanced tribal society from civilized society, in both common-sensical and analytical discourse, and provided the anthropologist with a theoretical model of the nature of 'noncivilized' social organization which could organise ethnographic data and function as a framework for cross-cultural comparison. At the same time as the difference between 'them' and 'us' was established, the basic similarity between different sorts of 'them' was proclaimed. This 'tribe-group' model was not, however, an explicit analytical framework; it was, rather, implicit in most ethnographic studies, neither examined nor questioned.

2. The Anthropology of Ethnicity

One of the earliest anthropologists to doubt the usefulness of the notion of the tribe was Leach (1954), who argued, in his study of highland Burma, that tribes, as discrete bounded entities, were essentially models developed by outsiders, for their own purposes, not locally meaningful principles of everyday social organization. Locals might talk about themselves as if there were clear-cut collective identities and entities, but everyday interaction and organization revealed a more complex pattern of overlap, variation, and fluidity.

By the 1960s, the notion of 'the tribe' was gradually being replaced in British social anthropology by the, perhaps less colonial-sounding, 'ethnic group.' However, the underlying assumptions about corporate groups and cultural difference had probably not changed much. Within American anthropology, the increasing use of an ethnicity model, over the same period, can also be understood as part of a long-term, incremental shift in the discipline's axiomatic metatheoretical model of difference: from 'race,' to 'culture,' to 'ethnicity' (Wolf 1994).

2.1 Ethnic Groups and Boundaries

The paradigm shift towards a social constructionist model of ethnic groups followed the publication of the Norwegian symposium edited by Barth (1969). The subtitle of this collection of essays—*The Social Organization of Culture Difference*—tells us a great deal

about both anthropological understandings of ethnicity (cultural differentiation), and Barth's take on it (as social organization). In the 'Introduction,' he argued against the presumption that the social world was made up of 'distinct named groups.' Ethnicity, the boundaries of ethnic groups, and indeed their ontological status as collectivities, could not be accepted uncritically as fixed aspects of the social reality in question. Instead, Barth insisted that ethnic identity—and its production and reproduction in routine social interaction—must be treated as essentially problematic features of social life, as emergent properties of everyday interaction. The ethnographer must examine the practices and processes whereby ethnicity and ethnic boundaries are constructed socially.

Barth and his colleagues shifted the anthropological emphasis from the static evocation of tribal identity as a feature of social structure, to a recognition of ethnic identity as a dynamic aspect of social organization: ethnic identification is something that people can and do manipulate, depending on the needs and possibilities of the context. People are not 'at the mercy' of their ethnicity, nor do they do things simply because they are ethnically 'X' or 'Y.' Ethnicity is not an imperative. In acknowledging this, *Ethnic Groups and Boundaries* marked a shift from an interest in the uncivilised peoples of the colonies and ex-colonies, to a more equitable concern with the complex heterogeneity of all societies.

Barth's emergent alternative framework had several key features. First, the analysis of ethnicity starts from the definition of the situation held by social actors. Second, the focus of attention then becomes how ethnic boundaries are maintained or changed in the structured interaction between 'us' and 'them' which takes place across boundaries. Third, the ethnicity of actors is not necessarily fixed; it is defined situationally. Fourth, ethnic identity depends on ascription, by members of the ethnic group in question and by outsiders with whom they interact. Fifth, ethnicity is not a matter of 'real' cultural differentiation; differences are in the eye of the beholder, the 'cultural stuff' which had been hitherto believed to determine group identification is somewhat irrelevant. Finally, ecological issues are influential in producing and reproducing ethnic identity: economic competition for scarce resources plays an important role in the generation of ethnicity.

2.2 Ethnicity as Socially Constructed

This social constructivist model of ethnicity can be traced back to the ideas of sociologists such as Max Weber, Everett Hughes, and Erving Goffman. Although *Ethnic Groups and Boundaries* was the model's most systematic, extended, and influential statement, similar arguments can also be found in the work of contemporary North American anthropologists such as Moerman (1965) and Cohen (1969). It has dominated discussions of ethnicity within European and North American anthropology since.

This basic paradigm was developed by Handelman (1977). Acknowledging the importance of boundaries, and the situational negotiability of ethnic membership, he argued that Barth overemphasised the centrality of social interaction in establishing boundaries, at the expense of acknowledging the role of culture. For Handelman, the cultural content of ethnicity is an important aspect of its social organization. A dichotomy between the cultural and the social is misleading. He argues further that ethnicity is incorporated—made socially significant—in differing degrees of intensity and territoriality, on which depends its salience in individual experience. Moving from 'the casual to the corporate,' Handelman distinguished the ethnic set, the ethnic category, the ethnic network, the ethnic association and the ethnic community. Ethnicity can, for example, organize everyday life without ethnic groups featuring locally as meaningful social forms.

What we can call 'the basic anthropological model of ethnicity'—it has become a 'normal science paradigm' of sorts—has many virtues. In stressing social process and the practices of actors, it encouraged a move away from the various strands of determinism of the 1970s and early 1980s. The emphasis upon social construction discourages the intrusion of biological conceptions of ethnicity or 'race' into social analysis. Finally, the importance that is attached to the views and self-perceptions of social actors themselves undermines the taken-for-granted ethnocentrism of much social science. These are important gains, and in all of these respects the approach may be said to have anticipated some of the most valuable aspects of postmodernism and poststructuralism.

However, the approach has suffered at the hands of some of those who have most wholeheartedly embraced it. In stressing concepts such as 'group,' 'category,' and 'boundary,' and processes of 'maintenance,' Barth—despite his intentions to the contrary—contributed to the further reification of the ethnic group as a perduring corporate entity. In particular, the spatial and physical overtones of the notion of the 'boundary' allow the superficial or the careless to celebrate the situational flexibility of ethnicity without taking on board the more difficult questions about the nature of collective social forms in which Barth is interested. In their enthusiasm for processes of boundary maintenance—negotiation and strategizing—many of those whose analyses owe much, if not everything, to Barth have taken for granted what he did not, i.e., the 'group' that is the object of the research.

Three other, related, criticisms are worth mentioning. First, although the locations where Leach, Barth, and others did their fieldwork—South-East Asia and the Middle East, for example—do allow for flexibility

and a degree of choice with respect to identification, the same is not true everywhere. In many other contexts—such as Northern Ireland or Kosovo—there may be much less room for maneuver. Second (although this is changing), many anthropologists, in their celebration of the positivities of everyday life, still appear reluctant to recognize power imbalances. The third criticism concerns the relative lack of theoretical attention which has been given to the relationships between ethnicity, on the one hand, and other principles of social identification, particularly class and stratification, on the other.

None of these problems, however, are inescapable theoretical weaknesses of the general anthropological approach. Recently, a solution to some of them has been put forward in the application to the study of interethnic relations of the distinction between groups and categories (Jenkins 1997). It is important to distinguish between two analytically distinct processes of ascription: group identification and social categorization. The first occurs inside the ethnic boundary (group members identify themselves), the second outside and across it (a collectivity and its members are categorized by nonmembers). Social categorization, in particular, is intimately bound up with power relations. It reflects the capacity of one group successfully to impose its categories of ascription upon another set of people, and the resources which the categorized collectivity can draw upon to resist that imposition. Acknowledging the significance of this distinction places relationships of domination and subordination on the theoretical center-stage.

2.3 Ethnicity as Personally Meaningful

If ethnic identity is a flexible matter, dependent on situation, it is legitimate to ask what it really means to the individuals concerned. This question has generated discussion about the nature of ethnic identity and is a perennial topic in the textbooks. Is ethnicity a fundamental, primordial aspect of human existence and self-consciousness, essentially unchanging and unchangeable in the demands it makes upon individuals and the bonds which it creates between the individual and the group? In other words, is ethnicity an aspect of 'human nature'? Or is it, to whatever extent, defined situationally, strategically or tactically manipulable, and capable of change at both the individual and collective levels? Is it, indeed, wholly socially constructed, as the post-Barth conventional wisdom might suggest?

This debate takes its place alongside a range of essentially philosophical controversies about human nature, the human condition, and the capacity of humans to intervene in their own lives, to determine or to be determined. But the 'primordial model' also has other historical roots, in the Romantic reaction to Enlightenment rationalism. Finding some of its more respectable voices in Herder and Hegel, this is much more than an academic point of view. Tragically, it has provided the intellectual charter for terrible—to the point of murderous—ethnic chauvinism, nationalism, and genocide. The ideology of primordialism naturalizes ethnic groups and justifies chauvinistic ethnic sentiments. Academically speaking, it can, for example, be identified at play in the ongoing debate about the modernity and nature of nationalism (e.g., Smith 1994).

Within anthropology, the name that is most often identified with a primordial model of ethnicity is Geertz (1973). Writing originally in 1963, Geertz was concerned to understand the obstacle that 'primordial attachments'—known variously as tribalism, parochialism, or communalism, and encompassing more than ethnicity—posed to the development of the modern political sentiment of citizenship in emergent post–colonial 'new states.' This was an important recognition—if one was needed—that whatever else it is about, ethnicity is also about affect and the sometime power of culture to move people.

So on the one hand we have negotiable identities, while on the other there are primordial attachments. The apparent conflict between the 'primordialist' and 'situationalist' perspectives is, however, not as clear-cut as it seems. It is hard to resist the conclusion that the debate has generated more heat than light. Both positions arguably have at least as much in common as not, and the protagonists are certainly usually misrepresented, sometimes to a remarkable degree.

To look, for example, at what Geertz actually said, he recognized explicitly the role of culture in defining the primordial 'givens,' and that primordial attachments are historically and situationally variable in their power to move people. He is also clear that what matters is that ties of blood, language, and culture are seen by actors to be ineffable and obligatory; that they are seen as natural. He is concerned with the terms in which attachments are understood and mobilized locally; with what people believe. Finally, Geertz further argues that these putative 'primordial attachments,' far from being necessarily ancient, may actually be stimulated and quickened by the political modernization of nation-building.

Barth, for his part, has never neglected the power and stability of ethnic identifications. As much concerned with the persistence of ethnic boundaries as with their flux, what he said was that under certain, not uncommon, circumstances ethnic change can happen, not that it must. Furthermore, his point of view is often presented fossilized in its 1969 incarnation. Since then he has explored the importance of ongoing, and historically relatively stable, 'streams of tradition' or 'universes of discourse,' within the constraints of which ethnic identities are produced and reproduced in practice. Most recently, he has acknowledged the need to add micro- and macrolevels of analysis to the intermediate level upon which his

original arguments concentrated. This means looking at what ethnicity means for individuals.

There are other reasons for becoming impatient with this debate. In the evocative words of one teaching text, it offers a contrast between 'ethnicity in the heart' and 'ethnicity in the head' (Banks 1996). This alerts us to the need to acknowledge affect and emotion in our considerations of ethnicity. But there is no necessary contradiction between instrumental manipulation, on the one hand, and sentiment, on the other. They may, in fact, go hand in hand. They may also, of course, conflict. But so may different sentiments or opposing instrumental goals.

There is, however, good reason to reject totally any strongly primordialist view. Too much ethnographic evidence exists of the fluidity and flux of ethnic identification and of the differing degrees to which ethnicity organizes social life in different settings, for any other position to be sensible. However, nor, on the other hand, can we deny the longevity and stubbornness, under particular circumstances, of ethnic attachments. But to acknowledge this point does not necessitate either embracing a primordial model, or abandoning the idea of social construction. It is not stretching the point to regard ethnic differentiation—the social construction of 'us' and 'them,' marked in cultural terms—as a ubiquitous feature of human sociability, and hence of all human societies. Arguments about whether or not ethnicity is 'situational' or 'primordial' confuse the ubiquity of a social phenomenon such as ethnicity with 'naturalness' (implying determinism and presocial powers of causation). To suggest that ethnicity is ever present, as one of the 'givens' of human social life, is not to endorse the primordialist point of view.

Within anthropology, the primordial intellectual position is unconvincing largely because it is something of a straw man. The scholarly debate about 'instrumentalism' and 'primordialism' exaggerates the differences between, for example, Geertz and Barth, and it does neither justice. More widely, it is tempting to suggest that the notion of the 'primordial' should be banished from the social-science lexicon altogether. On the other hand, however, it would be irresponsible to forget the debate altogether. Crude primordialism is essentially a common-sense view, with enormous power in the world. In sheep's clothing it occasionally makes an academic showing, indirectly, in the guise of sociobiology and discussions about territoriality. These views cannot be simply ignored, and an anthropological perspective—rooted in social constructionist assumptions—has much to offer as a critique of the naturalization of chauvinism and ethnic conflict.

2.4 *A Basic Anthropological Model of Ethnicity*

Despite the criticisms and debates summarized here—and in unacknowledged harmony with post-1980s discourses about 'difference,' globalization, transnationality, and so on—the 'basic anthropological model of ethnicity' remains in place as disciplinary conventional wisdom. In its foundational emphasis upon social construction, its recognition of the fluidity as well as the solidity of identity, and its refusal to biologize or naturalize what are sociocultural phenomena, it remains the cornerstone of any anthropological approach to the topic. It can be summarised thus:

(a) ethnicity is about cultural differentiation (bearing in mind that similarity and difference are inextricably connected),

(b) although ethnicity is centrally concerned with culture—shared universes of meaning—it is also rooted in, and to a considerable extent the outcome of, social interaction,

(c) ethnicity is no more fixed or unchanging than the culture of which it is a component or the social situations in which it is produced and reproduced, and

(d) ethnicity as a social identity is collective and individual, externalized in social interaction and internalized in personal self-identification.

However, that there is a relatively consensual model, combined with the anthropological enthusiasm for detailed ethnography rather than systematic theory, may lead to some things being taken for granted. In particular, there is a perpetual need to struggle against the tendency to reify culture and ethnicity. Although we talk about them in these terms, neither culture nor ethnicity is 'something' that people 'have' or to which they 'belong.' They are, rather, complex repertoires which people experience, use, learn and 'do' in their daily lives, within which they construct an ongoing sense of themselves and an understanding of their fellows.

The reification of ethnicity may also lead to its being seen as typically—or even only—an attribute of the Other. We need to remind ourselves all the time that each of us participates in an ethnicity—perhaps more than one—just like them, just like the Other or 'the ethnic minorities.' National identity may, however, be easier to recognize in ourselves than ethnicity. Nationalism and the construction of national identity are explicit projects of the state (and, if nothing else, we have passports). In the 'modern world,' the contours and contents of national identity, and the contexts of its uses and justifications, may be more visible than ethnicity. It is, therefore, no surprise that the anthropology of nationalism is a recent growth area (Eriksen 1993).

Although anthropologists may not have to remind themselves that national identity—as a variation on the theme of ethnicity—is socially constructed, there is a need to keep this idea firmly in the public eye. The timeliness of Wolf's warning (1994) that 'race,' 'culture,' and 'people' are 'perilous ideas,' makes the social constructionist, anthropological model of ethnicity as relevant at the turn of the twenty-first century

as it was in the 1960s. Anthropology's comparative global reach and local-level research focus, its emphasis upon culture and social construction, and its capacity to see individual trees as well as the collective wood, relativizes notions about ethnicity and undermines the naturalization of ethnic or national identity. That remains an important contribution to our understanding of the modern social world.

See also: Ethnic and Racial Social Movements; Ethnic Conflicts; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity, Sociology of; Identity in Anthropology; Identity: Social

Bibliography

Banks M 1996 *Ethnicity: Anthropological Constructions*. Routledge, London

Barth F (ed.) 1969 *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*. Universitetsforlaget, Bergen, Norway

Cohen Y A 1969 Social boundary systems. *Current Anthropology* **10**: 379–403

Eriksen T H 1993 *Ethnicity and Nationalism: Anthropological Perspectives*. Pluto, London

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Handelman D 1977 The organization of ethnicity. *Ethnic Groups* **1**: 187–200

Jenkins R 1997 *Rethinking Ethnicity: Arguments and Explorations*. Sage, London

Leach E R 1954 *Political Systems of Highland Burma: A Study of Kachin Social Structure*. Athlone, London

Moerman M 1965 Ethnic identification in a complex civilization: who are the Lue? *American Anthropologist* **67**: 1215–30

Smith A D 1994 The problem of national identity: ancient, medieval and modern? *Ethnic and Racial Studies* **17**: 375–99

Wolf E R 1994 Perilous ideas: race, culture, people. *Current Anthropology* **35**: 1–12

R. Jenkins

Ethnicity, Crime, and Immigration

Ethnicity is often conflated with race, from which it must be distinguished. Being an immigrant is a legal status conferred upon persons who come as permanent residents to a country other than their native land. Crime is deviant behavior that is in breach of the law. In all Western-style democracies, racial, ethnic, and immigrant minorities are an object of penal repression which is disproportionate vis-à-vis their percentage of the total population. This article reviews the explanations proposed to account for this fact. For detailed treatment of their various elements and for further reference, see *Discrimination: Racial*; *Ethnic Conflicts*; *Ethnicity, Sociology of*; *Integration: Social*; *Underclass*; *Migration: Sociological Aspects*; *Minorities*; *Ethnic and Racial Social Movements*; *Multiculturalism: Sociological Aspects*; *Racial Relations*; *Racism, Sociology of*; *Conflict Sociology*; *Race Identity*.

1. Identifying and Measuring the Problem

All Western democracies collect data on immigrants and more generally on noncitizens living within their respective boundaries, although these data cannot record the significant number of illegal immigrants. It is thus possible for these countries to collect justice statistics on the very general basis of who is a national citizen and who is not when having a brush with the penal system. However, these statistics are not further broken down according to racial and ethnic origin. Even if state immigration departments keep tabs on where immigrants are coming from, this kind of information is not recorded by other departments. Consequently, justice statistics are generally not discriminant with respect to ethnicity and race. In continental Europe, such discriminant statistics would seem a throwback to pre-Second World War racism and bring back the yellow stars nightmare. In other countries such as Canada, ethno-racially discriminant statistics are believed to violate basic civil rights, with the exception of data on Aboriginal people, which have traditionally been collected because First Nations are endowed with a special legal status. However, there are two countries where justice statistics sensitive to race and ethnicity are collected. The first is the US, where such data have been collected for a long period of time, and the other is the UK, where a 1995 Home Office letter to Chief Constables required them to introduce a system of ethnic monitoring of suspects and offenders. The UK also conducts victimization surveys that discriminate between ethno-racial groups either as victims or offenders.

It is then crucial to stress that despite its intensity, the debate on crime, race, ethnicity, and immigration does not in most countries rely on facts, but on spotty data generated by academic research and biased media impressions. In continental Europe or Canada this situation is not about to change.

2. Basic Distinctions: Race, Ethnicity, and Related Categories

Several distinctions must be made in order to discuss the issues properly. First, ethnicity must be distinguished from race. *Race* refers to outward physical attributes such as the color of the skin, whereas

ethnicity is comprised of cultural traits, and is a factor of *identity* for those who share these traits. Second, the status of being an *immigrant* is a legal construct and differs in this regard from race and ethnicity which are cultural constructs. The fact that being an immigrant is primarily a legal status in Western societies implies that there are forms of immigration which are illegal. In contradiction, asking whether one belongs legally or not to a racial or an ethnic group is generally a spurious question. However, race and ethnicity can be construed as a legal status, as it was in Nazi Germany, in South Africa during the period of Apartheid, and in Rwanda, where one's identification papers indicate whether one is a Hutu or a Tutsi. As is shown by these examples, attempts to enshrine in law differences with regard to race or ethnicity have often resulted in human tragedies. One exception to this assertion is granting a specific legal status to *Aboriginal peoples* in countries such as Australia, Canada, New Zealand, the US, and in South America. The purpose of this measure—which was very imperfectly achieved—was to recognize that the Aboriginal peoples were the first nations that populated these countries and that they were entitled as such to certain privileges.

The importance of making the preceding distinctions is embodied in one cardinal rule for addressing the issues raised in this article: the wider are the categories used—e.g., being white or black-skinned—and the lesser the probability to shed light on an issue (Georges-Abeyie 1989).

3. *Further Basic Distinctions: Disparity and Discrimination*

There is one last distinction that does not refer to groups of persons but that defines the most important issue discussed in relation to these groups. Disparity should not be confused with discrimination. *Disparity* is a descriptive notion and refers to a difference, which reveals itself through disproportion or over-representation. For instance, it is well known that African Americans are disproportionately represented in the US prison population with regard to their number in the total population. *Discrimination* refers to the differential treatment of groups without respect to the behavior or qualifications of individuals belonging to such groups (Walker et al. 1996, p. 16). Discrimination can be used to explain disparity. It varies in scope from cases of *individual* prejudice to differential treatment which is applied *systematically* at all thresholds within a system. *Systemic*, as opposed to *systematical* discrimination, occurs when disparity is unintendedly generated by features integrated within a system of rules. For instance, very few Asian immigrants were hired in Western police forces because of obsolete height requirements that they could not meet.

4. *The Main Issues*

Until the 1990s, the issues most debated concerned racial and ethnic minorities in their relation to crime. In recent years, immigration also became an issue through the problem of illegal immigrants. Because the US and the UK are the only countries collecting data that are sensitive to ethnic and racial differences, this article will focus mostly on US and UK data.

4.1 *Disparity or Discrimination*?

The evidence of disparity is overwhelming, particularly with respect to the most repressive component of the penal system, that is, incarceration.

(a) Although African Americans comprise 12 percent of the US population, they account for 31 percent of the federal and for 51 percent of the state prison population (Walker et al. 1996, p. 205).

(b) In 1992, Black people constituted 5 percent of the population of the UK but made up 18 percent of the prison population (Dhokalia and Sumner 1993, p. 28).

(c) In 1988, nearly one in three inmates in the Canadian western provinces were of Aboriginal ancestry. Aboriginal people make up almost 10 percent of the federal inmate population, although they make up less than 2 percent of the Canadian population (Law Reform Commission of Canada 1991, p. 67, note 166).

These figures on the over-representation of racial or ethnic minorities are the same with respect to the operation of other components of the criminal justice system, most noticeably the police. In a 1975 experiment conducted in two precincts of San Diego, nearly half of all people stopped and questioned by police were African Americans, even if they respectively represented only 17.5 percent and 4.8 percent of the population of the two precincts in the study (Walker et al. 1996, p. 101). The evidence is similar for the UK (W. Skogan 1990, *The Police and Public in England and Wales: A British Crime Survey Report*. Home Office Research Study No. 117, HMSO, London). Racial profiling still is a persistent feature of policing.

Such being the evidence for disparity, what of its explanation by ethnic or racial discrimination? Two quotes from the work of Robert Reiner epitomize the present predicament of research in this respect. Reiner wrote in 1989 that despite methodological caveats, 'the quantity and quality of (the) evidence is such as to render any doubts about discrimination fanciful' (Reiner 1989, p. 8). He also wrote in 1992 that it was 'inconceivable that the (statistical) approach could ever conclusively establish racial discrimination' (R. Reiner, quoted in D. Cook and B. Hudson (eds.), *Racism and Criminology*, Sage, London, 1993, p. 8). Therein lies the paradox. On the one hand, the evidence of disparity is so overwhelming that it

seemingly does not allow for explanations other than intended discrimination. On the other hand, most statistical attempts to demonstrate the presence of discrimination are inconclusive, when all variables are duly controlled. For instance, African Americans and Hispanics were respectively found to be six and four times more likely to be shot by police in Chicago between 1974 and 1978. However, when the research controlled for the risk status, defined in terms of arrests in the context of the perpetration of a serious violent crime, the racial disparity vanished, and it was white perpetrators that were more likely—but not by much—to be shot by police (A. Geller and K. J. Karales 1981 *Split Second Decisions: Shootings of and by Chicago Police*. Chicago Law Enforcement Study Group, Chicago, p. 119). Such erosion of the discrimination factor in controlled statistical research occurs most of the time (Hagan 1988).

What then is the explanation of racial and ethnic disparity in criminal justice?

(a) There is no question that discrimination plays a role. However, it is difficult to assess its extent, which is revealed only by complex statistical analysis highly dependent on data quality (Blumstein et al. 1983, Hood 1992).

(b) Factors which are in themselves racially and ethnically neutral also account for disparity. Judges tend to incarcerate youths when their family cannot provide any support for their rehabilitation. The number of broken African American homes being high, more young African Americans are consequently put in jail. There are many of these neutral factors, ranging from the way in which an immigrant community tries to integrate (Smith 1997, p. 759), to special forms of drug addiction (crack vs. powder cocaine) that are repressed with disproportionate severity (Tonry 1995, p. 188).

(c) Nevertheless, the most thorough and methodological research generally concludes that it is the differential involvement of minorities in the perpetration of crime that mainly explains disparity (Smith 1997). This conclusion is frustrating because it only pushes the problem one step further: why are some ethnic and racial minorities more involved in crime?

4.2 *Intra- or Intervictimization*

Another volatile issue is whether perpetrators victimize members of their own racial or ethnic community or whether they cross the ethno-racial divide. The answer to this question varies from one country to the other.

(a) In the US, white offenders victimize the white community in more than 90 percent of cases. However, the percentage of white victims from African American offenders ranges between 40 percent for all violent crime and 60 percent for robbery (Walker et al. 1996, p. 49).

(b) In the UK, white offenders cross the ethno-racial divide much more frequently, 68 percent of Asian and 50 percent of Afro-Caribbean victims having identified their aggressor as white in recent British crime surveys (1988 and 1992; Smith 1997, p. 715). Afro-Caribbeans also cross this divide, 23 percent of Asian victims having identified their aggressor as Afro-Caribbean.

The difference between the US and the UK in terms of the tendency of whites to attack other communities may in part be explained by the fact that the Afro-Caribbean community is less than 1 percent of the total UK population, whereas African Americans account for 12 percent of the US population.

4.3 *Issues Related to Illegal Immigration*

Although illegal immigration is an old phenomenon, the issues that it raises have taken a particular urgency after 1990 but they have not been thoroughly researched yet. Three issues can be singled out. First, the criminal networks involved in smuggling illegal immigrants have grown immeasurably since the end of the Cold War and the explosion of civil strife in former Eastern bloc countries. Second, although it is directly related to the previous one, the issue of the transnational exploitation of prostitutes, often recruited in refugee camps, is a pressing one, for which the police have no immediate answer. There is finally the question of how to react to illegal immigrants who have settled for a number of years in their adopted country, where they have begun to be socially integrated, but where they have failed to meet the legal requirements to obtain legal citizenship.

5. *Conclusion*

This article has focused mainly on the overpunishment of racial and ethnic minorities and on its possible explanations. There is a concomitant phenomenon which falls outside the scope of this article, but that deserves to be mentioned: it is the underpunishment of members of the white majority when they victimize minorities (Radelet 1989). Finally, it ought to be stressed that in the present context of globalization, the migratory fluxes will increase, thus sharpening the edge of the issues that we have been discussing. As long as most Western countries will resist collecting data sensitive to race and ethnicity, their policies on integration will be developed in a blind vacuum.

See also: Crime and Ethnicity (Including Race); Crime, Sociology of; Immigration and Migration: Cultural Concerns; Immigration: Consequences for Fiscal Developments in the Receiving Population; Immigration: Public Policy

Bibliography

Blumstein A, Cohen J, Martin S E, Tonry M H (eds.) 1983 *Research on Sentencing: The Search for a Reform on Sentencing*. National Academy Press, Washington, DC, Vol. 2

Dhokalia N, Sumner M 1993 Research, policy and racial justice. In: Cook D, Hudson B (eds.) *Racism and Criminology*. Sage, Thousand Oaks, CA, pp. 28–44

DuBois W E B (ed.) 1904 *Some Notes on Negro Crime. Particularly in Georgia—Proceedings of the Ninth Atlanta Conference for the Study of Negro Problems*. Atlanta University, Atlanta, GA [Reprinted Octogon Books, New York, 1968]

Georges-Abeyie D E 1989 Race, ethnicity and the spatial dynamic: Toward a realistic study of black crime, crime victimization and criminal justice processing of blacks. *Social Justice* **16**: 35–54

Hagan J 1988 *Structural Criminology*. Polity Press, Cambridge, UK

Hood R 1992 *Race and Sentencing: A Study in the Crown Court*. Clarendon Press, New York

Hudson B A (ed.) 1996 *Race, Crime and Justice*. Dartmouth, Aldershot, UK

Law Reform Commission of Canada 1991 *Aboriginal Peoples and Criminal Justice*. Law Reform Commission of Canada, Ottawa, Report 34

Radelet M L 1989 Executions of whites for crimes against blacks: Exceptions to the rule? *The Sociological Quarterly* **30**: 529–44

Reiner R 1989 Race and criminal justice. *New Community* **16**: 5–21

Reiner R 1992 *The Politics of the Police*. University of Toronto Press, Toronto, ON

Reiss A J 1971 *The Police and the Public*. Yale University Press, New Haven, CT

Smith D J 1997 Ethnic origins, crime and criminal justice. In: Maguire M, Morgan R, Reiner R (eds.) *The Oxford Handbook of Criminology*. Oxford University Press, Oxford, UK, pp. 703–59

Tonry M 1995 *Malign Neglect—Race Crime and Punishment in America*. Oxford University Press, London

Walker S, Spohn C, DeLone M 1996 *The Color of Justice*. Wadsworth Publishing Company, Belmont, CA

J.-P. Brodeur

Ethnicity, Race, and Health

This article provides an overview of the role of race and ethnicity in health. Racial and ethnic statuses are primarily social categories that can reflect differential exposure to health-related risks or resources. Accordingly, across a broad range of societies, disadvantaged racial and ethnic populations have elevated rates of disease and death. Racial disparities in health are persistent over time. Data suggest that racial differences in health largely reflect racial variations in socioeconomic circumstances, but individual and institutional discrimination often play an incremental role. Understanding racial and ethnic variations in health requires greater attention to data quality issues and to assessment of both the resources and vulnerabilities to which population groups are exposed.

1. *What is Race and Ethnic Status?*

Historically, race has been viewed as capturing biological variation within human populations. However, the concept of race never rested on a firm, scientific foundation. Racial taxonomies pre-dated modern scientific theories of genetics and carefully executed genetic studies. Thus, race emerged as a sociopolitical construct, useful not only to classify phenotypic variation but also to justify the exploitation of groups defined as inferior. Historically, the conceptualization of race has been influenced by cultural and political considerations, and race has served important ideological functions in society (Duster 1984). For example, medical research in the nineteenth century attempted to document that blacks were biologically inferior to whites and therefore more susceptible to a broad range of diseases (Krieger 1987). Importantly, such data provided a scientific justification for slavery and economic exploitation.

Understanding and addressing the role of race and ethnicity in health requires a clear understanding of what race is, and the identification of the specific risk factors and resources linked to racial and ethnic status that may be responsible for variations in health. A growing body of evidence indicates that the traditional conceptualization of race is flawed. Scientific data reveal that 95 percent of human genetic variation exists within racial groups (Lewontin 1982). That is, there is more genetic variation *within* races than between them and, although there are patterns to the distribution of genetic characteristics, our racial categories do not capture biological distinctiveness. Our current racial categories are more alike than different in terms of biological characteristics and genetics, and there are no specific, scientific criteria to classify the human population unambiguously into discrete biological categories with rigid boundaries (Lewontin 1982). Racial taxonomies are arbitrary, and race is more of a social category than a biological one.

Although ethnic groups can share a range of phenotypic characteristics due to their shared ancestry, the term 'ethnicity' is typically used to highlight cultural and social characteristics instead of biological ones. Ethnic groups share some combination of common geographic origins, family patterns, language, common values, cultural norms, religious traditions, literature, music, dietary preferences, and

employment patterns. Thus, the current 'racial' categories really capture ethnic variation; that is, they reflect common ancestry, history, or culture. Some have suggested that the term ethnicity should replace 'race' in health research, and the term race should be abandoned (Cooper 1994). However, although not useful as a biological category, race is an important social category. It is what sociologists call a master status—a central determinant of social identity and obligations, as well as of access to societal rewards and resources. Given that racial categories have reflected racism and its consequences (oppression, exploitation, and inequality) historically, collecting the appropriate data on race can facilitate ongoing monitoring of the magnitude of differentials and enhance understanding of their causes and the development of effective interventions to address them.

Relatedly, although variations in health are evident for many ethnic groups, much of the marked variation is evident for minority racial and ethnic groups. The term minority is used, not to represent the numerical size of a population, but rather to reflect the stratification and access to power and resources of that group. The current minority groups in the United States, for example, reflect the intersection of racial and ethnic status combined with economic disadvantage. Only those population groups that experienced prejudice and discrimination historically and have been unsuccessful in assimilating are minority groups.

2. *Racial and Ethnic Variations in Health*

Across a broad range of societies, racial and ethnic statuses predict variation in health. The United States has routinely reported health statistics by race, and the role of race and ethnic status is readily evident in US data. National mortality data reveal that African Americans (or blacks) have death rates that are higher than whites for 13 of the 15 leading causes of death in the United States (NCHS 1991). The differences in death rates are small for some conditions, such as atherosclerosis, but at least twice as high for deaths from diabetes, kidney disease, homicide, septicemia, prenatal conditions, and HIV/AIDS. These pervasive disparities across such a broad range of health conditions are striking. In US data, Hispanics and American Indians have lower mortality rates than whites for cardiovascular disease and cancer, but higher rates for some causes of death, such as diabetes and cirrhosis of the liver. In contrast, the Asian American population has mortality rates that are lower than those of whites.

Until recently, the United Kingdom did not focus on racial/ethnic variations in health. However, analyses of death rates by country of birth suggest that there may be important racial/ethnic variations in health status in England (Balarajan and Raleigh 1995). These patterns are complex and vary by both the specific ethnic group and health outcome under consideration. For example, death rates from coronary heart disease and stroke are higher than the national average for persons from the Indian subcontinent and the African commonwealth. In contrast, Caribbean immigrants have mortality rates lower than the national average for coronary heart disease but higher for hypertension and stroke. Similarly, persons from the Indian subcontinent and the Caribbean have higher-than-average mortality for cancers of the liver, gall bladder, and cervix. Death rates from breast and lung cancer are lower in these three major immigrant groups, but Asians have high rates for oral cancer, and Caribbeans have high mortality rates from prostate cancer.

The experience of the United States and England is not unique. In a comprehensive review of racial and ethnic differences in disease internationally, Polednak (1989) reveals that racial and ethnic status predicts variations in health across a broad range of societies. Disadvantaged racial/ethnic groups, such as the Maori in New Zealand, blacks in South Africa, and Aboriginal populations in Australia, have higher rates of disease than the more advantaged groups within their societies. For example, for the 1991–6 period in South Africa, compared to the survival experience of white infants, the infant mortality rate was 5.2 times higher for blacks, 4.2 times higher for Coloureds and 1.3 times higher for Asians (Sidiropoulos et al. 1997). Recent Australian data reveal that the rate of low birthweight infants for indigenous mothers is twice that of their nonindigenous peers, and the neonatal mortality rate (deaths of liveborn infants within 28 days of birth) is 2.6 times higher for babies born to Aboriginal mothers compared to the rest of the population (McLennan and Madden 1999). These elevated mortality rates persist throughout the life course, with Aboriginal males and females having a mortality rate between ages 35 and 44 that is at least seven times higher than that of the nonindigenous population. Moreover, there is considerable variation by ethnic status in multiple societies, with some ethnic groups having higher frequencies of some disease conditions but lower frequencies of others (Polednak 1989). For example, among European groups, the French tend to have low rates of heart disease, the Poles low rates of stroke, and the Finns low rates of breast and colorectal cancer.

2.1 *Persisting Racial Disparities in Health*

One important feature of racial differences in health is the persistence of the racial gap in health status over time. For example, despite declining death rates from stroke in England between 1970 and 1983, there was the persistence of elevated rates of stroke deaths for persons from the Caribbean and Africa (Balarajan and Raleigh 1995). Table 1 shows that, in 1950, blacks

Table 1
Mortality rates by race 1950—97

	All-cause mortality Deaths per 100,000 population			Infant mortality Deaths per 1,000 live births		
Year	White (W)	Black (B)	B/W Ratio	White (W)	Black (B)	B/W Ratio
1950	800.4	1,236.7	1.5	26.8	43.9	1.6
1960	727.0	1,073.3	1.5	22.9	44.6	2.0
1970	679.5	1,044.0	1.5	17.8	32.6	1.8
1980	559.4	841.5	1.5	10.9	22.2	2.0
1990	492.8	789.2	1.6	7.6	18.0	2.4
1997	456.5	705.3	1.5	6.0	14.2	2.4

Source: NCHS, 1999

in the United States had an overall death rate that was 1.6 times higher than that of whites. Mortality rates have declined for both racial groups over time, but in the late 1990s, blacks had a death rate that was still 1.6 times that of whites (NCHS 1999). For several causes of death (heart disease, cancer, cirrhosis of the liver, and diabetes), the black/white ratio was wider in the late 1990s than in 1950. There were multiple paths to this pattern of persistent racial differentials (Williams 1999). For some outcomes, such as heart disease, declining rates are evident for both racial groups, but the black/white difference widened due to more rapid improvements in the health of the white population than of the black population. For cancer and diabetes, there were stable or declining rates for whites, but increasing rates for blacks. For other outcomes, such as stroke and unintentional injury, the black/white gap was virtually identical in the late 1990s to what it was in 1950. However, influenza and pneumonia stand out as the only causes of death where the black/white difference in mortality has narrowed over time due to more rapid improvements in the health of the black compared to the white population. We need an enhanced understanding of the reasons for this success so that we could develop interventions to replicate it.

Infant mortality data tell a similar story. The right panel of Table 1 indicates that, compared to the 1950 infant death rate, the 1997 infant mortality rate was three times lower for blacks and more than four times lower for whites (NCHS 1999). However, the relative difference in rates between the two racial groups was larger in 1997 than in 1950. A black infant was 1.6 times more likely than a white infant to die before his or her first birthday in 1950. This ratio increased to 2.4 in 1997.

3. *Economic Status: A Central Determinant of Health*

Although racial and ethnic status do not tell us much about underlying biological risks for disease, they do reflect an important part of the economic inequality and historic injustice that has existed in particular societies. Racial differences in economic circumstances are central to understanding health status variations.

3.1 *Racial Differences in Socioeconomic Status*

There are important differences in power and socioeconomic status (SES) across major racial and ethnic groups. Table 2 presents levels of per capita income and poverty rates by race in South Africa from a World Bank-funded national survey on Living Standards and Development conducted in 1993. Racial disparities in economic circumstances are striking. Whites in South Africa have a per capita income (R 32,076) that is almost 12 times that of blacks, five times that of Coloureds, and two and a half times that of Asians (Whiteford et al. 1995). In a similar vein, compared to the poverty rate of whites of 2 percent, the rate of poverty is 29 times higher for blacks, 10 times higher for Coloureds, and more than three times higher for Asians. For all racial groups, the poverty rate is higher for children than for the total population.

In the United States, blacks, Hispanics, and American Indians have lower rates of income, education, and occupational status than whites. The Asian population in the United States is overrepresented at both ends of the SES distribution. Compared to whites, Asians have higher levels of median income, but some subgroups of the Asian population, such as the Hmong, Laotians, and Cambodians, have lower levels of income and higher rates of poverty than the black or the American Indian population. This highlights the heterogeneity that exists within many of the major racial and ethnic categories, and the importance of attending to the variation within these subgroups. For multiple health outcomes and multiple racial/ethnic groups, an overall statistic masks the variation that exists for subgroups within that population.

Data for the US also reveal that racial differences in income understate the true magnitude of racial differences in economic circumstances. Income captures the flow of economic resources into the household. Wealth

Table 2
Selected socioeconomic characteristics and race, South Africa 1993

SES Indicator	African	Colored	Asian	White
1. Per capita income (Rand)	2,717	6,278	12,963	32,076
2. Poverty				
Percent poor, all	57.2	19.8	6.8	2.0
Percent poor, children	63.9	23.1	9.6	2.1

Source: Witheford et al. 1995

Table 3
United States life expectancy, at age 45, by family income (1980 dollars)

	Females			Males		
Family income	White	Black	Difference	White	Black	Difference
All	*36.3*	32.6	*3.7*	*31.1*	*26.2*	*4.9*
1. Less than $10,000	35.8	32.7	3.1	27.3	25.2	2.1
2. $10,000–14,999	37.4	33.5	3.9	30.3	28.1	2.2
3. $15,000–24,999	37.8	36.3	1.5	32.4	31.3	1.1
4. $25,000 or more	38.5	36.5	2.0	33.9	32.6	1.3

NCHS 1998; NCHS 1997.

reflects the accumulated economic reserves that can cushion a decline in income and facilitate long-term planning. Racial differences in wealth are considerably larger than those for income. For example, in the United States, the median net worth of whites ($44, 408) is almost ten times larger than that of blacks, and eight times larger than that of Hispanics. Moreover, measures of wealth make an incremental contribution in predicting variations in health over and above measures of education and income (Williams and Collins 1995).

3.2 Race, SES, and Health

Research reveals that SES differences between the races account for much of the racial/ethnic differences in health. For example, adjusting black/white disparities in health in the USA for SES sometimes eliminates, but always substantially reduces, these differences. Table 3 illustrates these patterns. It presents US life expectancy at age 45 by family income for black and white males and females. At age 45, white males have a life expectancy that is almost five years longer than their black counterparts; and white females outlive their black counterparts by 3.7 years. Several important trends are noteworthy. First, for both racial groups, income is strongly linked to life expectancy. Consistently, persons of lower levels of income have lower life expectancy than their more economically favored peers. Second, the effect of SES on health is not confined to the low end of the income distribution. Persons of each higher level of income experience better health than those at the level immediately preceding theirs, at least up to the highest category used here (which is about the median).

Third, the difference in health status by SES for both black and white men is larger than the overall racial difference. African American men in the highest income group live 7.4 years longer than those in the lowest income group; the comparable number for whites is 6.6 years. The overall racial difference for men was 4.9 years. Although SES predicts variation in health for both black and white women, for women, the overall racial difference is somewhat larger than the differences between SES groups. The life expectancy data also reveal an independent effect of race, even when blacks and whites are comparable in income. At every level of income for both men and women, African Americans have lower levels of life expectancy than their white peers. This pattern has been observed across multiple health outcomes, and for some indicators of health, such as infant mortality, the racial gap becomes larger as SES increases.

3.3 Racism: An Added Burden for Non-dominant Populations

There are several factors linked to historical or contemporary racism that can account for these differences. Racism refers to organized institutional structures and policies, built on an ideology of superiority that allocates systematically fewer societal resources to groups regarded as inferior. Racial residential segregation in the United States and South Africa has been a prime example of a societal structure that importantly restricted socioeconomic opportunity and mobility for groups regarded as inferior (Cell 1982, Jaynes and Williams 1989). Residential segregation determined housing conditions, education and employment opportunities, and thus truncated

economic mobility. Because of conditions created by residential segregation, SES indicators are not equivalent across race. There are racial differences in educational quality in elementary and high schools. Similarly, the income return for a given level of education varies by race. For example, data from the United States reveal that, at every level of education, blacks and Hispanics, working full time in the labor force, have lower levels of income than whites (Williams and Collins 1995).

We noted earlier that there were large racial differences in wealth. These racial differences in wealth are evident at all levels of income, and are largest at the lowest income level (Williams and Collins 1995). For example, persons in the lowest quintile of income in the United States, the net worth of whites is $10,257 compared to $1 for blacks and $645 for Hispanics. For most US households, housing equity is a major source of wealth. Thus, contemporary racial differences in wealth reflect the institutional discrimination that was embodied in government housing policy and the practices of banks and the real estate industry a generation ago (Jaynes and Williams 1989). There are also large differences in purchasing power across race. For example, on average, African Americans pay higher prices for a broad range of goods and services (such as groceries, housing, and insurance) because of the higher costs in the residential contexts in which they reside. Residential segregation has also led to the concentration of poverty and the non-equivalence of neighborhood environments across race. Compared to whites of equivalent income, blacks in the USA live in communities lower in SES and resources, such that, in the 171 largest cities in the USA, the worst urban residential context for whites is better than the average context of black communities (Sampson and Wilson 1995).

Second, the health status of adults is influenced not only by current SES but also by exposure to adverse economic conditions over the life course. Non-dominant racial groups are more likely than their economically favored counterparts to have experienced economic deprivation, elevated rates of illness, and deficits in medical care during infancy and childhood. Some research suggests that the health status of adults may be affected adversely due to childhood diseases and living conditions, as well as to nutritional deprivation in the intrauterine environment or during infancy and childhood (Elo and Preston 1992). Thus, the persistence of racial differences in adult health, even in the face of adjustment for SES, could reflect, at least in part, racial differences in early-life economic and health adversity.

Third, racial discrimination and prejudice may constitute added risk factors that adversely affect health. There is evidence of systematic discrimination in the quantity and quality of services across racial groups, including medical care (Williams 1999). Studies in the United States reveal that, even after adjustment for SES, health insurance, and clinical status, whites are more likely than blacks to receive a broad range of specific medical procedures. These disparities exist even in contexts where differences in economic status and insurance coverage are minimized, such as the Veteran's Administration Health System and the Medicare program.

Research also suggests that subjective experiences of unfair treatment may be an important type of stress that can affect health adversely. Several studies reveal that experiences of discrimination are related adversely to physical and mental health (Krieger 1999, Williams and Williams-Morris 2000). Some recent data suggest that exposure to experiences of discrimination makes an incremental contribution to explain racial differences in self-reported measures of health over and above SES (Williams 1999). Other data indicate that a widely held societal stigma of inferiority can create specific anxieties, expectations, and reactions that can adversely impact socioeconomic performance as well as mental health status (Williams and Williams-Morris 2000, Steele 1997). First, across multiple societies, groups viewed as inferior have lower educational performance. Second, members of negatively stereotyped groups may internalize societal stereotypes such that the stereotype becomes a self-fulfilling prophecy leading to poor performance. Alternatively, the threat of poor performance can lead to anxieties that affect academic performance adversely. Third, some limited research suggests that African Americans who endorse negative cultural images of their group tend to have higher levels of depressive symptoms and substance use.

3.4 Changes in Racial Inequalities in Health

Data from the United States suggest that patterns of racial inequality change over time in response to larger socioeconomic changes. For example, although it was noted earlier that the black/white gap in overall mortality was identical in the late 1990s to what it was in 1950, over this time period there were times of both widening and reduced racial differentials in health. For example, the Civil Rights Movement improved the economic circumstances of blacks and narrowed the black/white gap in income during the 1960s and early 1970s. Correspondingly, mortality rates for blacks declined, both on a percentage and an absolute basis, more rapidly than those of whites between 1968 and 1978 (Williams and Collins 1995). Similarly, during this time period, the life expectancy increases for black males and females were larger on both a relative and an absolute basis than those for whites. However, the economic status of blacks relative to whites stagnated in the mid-1970s and worsened during the 1980s. Similarly, the black/white gap in health worsened for multiple indicators of health status during this decade. The black/white gap in life expectancy, infant mortality, and excess deaths in-

creased between 1980 and 1991 (Williams and Collins 1995). For example, for five consecutive years after 1984, the life expectancy at birth for blacks progressively declined from 1984 levels, while that of whites continued to increase progressively. These data are consistent with research, which shows that the health of populations is affected by a broad range of economic factors, including growth, stability, and inequality (Brenner 1995). For example, mortality from all causes tends to increase about two to three years after the low point of a recession. These data highlight that both race and SES have powerful influences on health because, as master statuses, they shape the experience of and exposure to almost all psychosocial and environmental risk factors for health.

4. Data Quality

Race is routinely used in both the health policy and the research literature in a routine, atheoretical, and uncritical manner, with little attention given to the measurement difficulties that are attendant on the assessment of racial categories. However, there is growing attention to reliability and validity problems linked to the assessment of race that have implications for the quality of available data on racial differences in health (Hahn 1992, Williams and Collins 1995).

4.1 Problems with the Numerator

In the United States, it is well documented that the numerators for officially reported death rates and some other health events are underreported due to the discrepancy between interviewer-observed race and respondent-reported race. This problem is minimal for data on blacks and whites, but substantial for American Indians, Asians, and Hispanics. Persons who self-identify with these groups are often misreported as white on death certificates and in other health statistics contexts. This numerator problem leads to an underestimation of health events for these racial/ethnic groups.

4.2 Problems with the Denominator

Another problem affecting the quality of racial data is the accuracy of the denominators used in calculating rates of health events. In many countries, data collected by the official census are used as the basis for denominators. In the United States, Census Bureau evaluations of the extent of the undercount based on administrative data and demographic trends reveal that, while the overall undercount for the United States population is small, it is considerably larger for blacks than for whites, and is especially large for middle-aged black males. Importantly, the use of a denominator that is undercounted inflates the obtained rate in exact proportion to the undercount in the denominator. Thus, most officially reported mortality and morbidity rates for middle-aged black males are about 15 percent too high. Currently, demographic analysis estimates of Census undercount are available only at the national level, and it is likely that the omission of black males from households (the major cause of the undercount of blacks) varies by geographic area and could thus be considerably larger for some states or municipalities.

4.3 The Challenge of Defining Race and Ethnicity

Data from the US also suggest that there is considerable inconsistency in reporting racial, and especially ethnic, status over time. Given that many persons are multiracial or multiethnic in ancestry, self-identity appears to be importantly linked to the particular context of measurement. Relatedly, there is considerable debate over the optimal terminology to be used for particular racial groups, the number of racial groups for which data should be collected, the boundaries between racial and ethnic status, and whether new categories should be developed for persons who are of mixed ancestry. For a growing number of persons in society, racial and ethnic status is not singular and static, but multiple and dynamic (Evinger 1995).

Complicating the definitional problems linked to race is the fact that categories of race and ethnicity are often assessed differently across societies, over time within particular societies, and by different agencies and institutions even within the same society. In the USA, for example, the standard birth and death certificates were revised in 1989 to include identifiers for Hispanic status. But the wording of the question assessing Hispanic status varies from state to state, and the data on Hispanic ethnicity is not completely reported in all states. Similarly, many health surveys and federal data-collection systems in the USA use one question to assess race and a separate one to assess Hispanic origin, while other agencies use one combined question to assess both racial and ethnic status. Importantly, differing wording of racial and ethnic questions will provide varying estimates of the size of racial/ethnic populations.

5. Resources and Acculturation

Health-enhancing resources exist even in the most disadvantaged racial and ethnic populations. Much research on the health of the disadvantaged has focused historically on pathology and vulnerability. However, the available health data clearly suggest that health-enhancing factors are also resident within contexts of economic deprivation. For example, despite high rates of poverty and low rates of health insurance coverage, Mexican immigrants still have a

relatively good health profile. Similarly, although African Americans are disadvantaged on almost every indicator of physical health status compared to whites, they have surprisingly good levels of mental health (Williams and Williams-Morris 2000). Rates of suicide are markedly lower for blacks than whites, and blacks tend to have equivalent or lower rates of psychiatric illness compared to whites. These data suggest the need to identify cultural strengths and health enhancing resources that exist within disadvantaged communities, and that may shield these populations from at least some of the negative exposures in their social context. Research has nominated the quality of family and extended kin relationships, as well as the role of religious institutions and participation, as key health-enhancing resources that may reduce some of the negative effects of stress on the health of disadvantaged populations. More research is needed on the mechanisms and processes by which these factors operate, and the ways in which they combine with existing risk factors to affect health.

In many societies, many minority racial and ethnic groups are immigrant populations. The health profiles of these groups may reflect their immigrant status importantly and may change with increasing length of stay in the host society. In the United States, immigrants of all the major racial/ethnic groups have lower rates of infant and adult mortality than their native-born counterparts (Williams 1999). Importantly, these patterns persist even when immigrant groups, such as Mexicans in the United States, are lower in SES and have limited access to medical care. At the same time, other data suggest that, across a broad range of health status indicators such as infant mortality, low birth rate, cancer, high blood pressure, adolescent pregnancy, and psychiatric disorders, the health status of immigrants worsen with length of stay in the USA (Williams and Collins 1995). The mechanisms and processes responsible for these negative effects of acculturation are not fully understood. Several behaviors that adversely affect health status appear to increase with length of stay. It has also been suggested that there may be health-protective aspects of indigenous cultures and belief systems that are lost with increasing time in a new society.

Culture is a construct that is widely invoked in discussions of racial and ethnic variations in health. However, it is often treated as a residual explanation for unexplained differences in health. There is a critical need for the theoretical identification and the empirical assessment of the specific cultural beliefs and behaviors that are presumed to affect health. Moreover, researchers must avoid labeling as cultural what are more structural differences in the social position of different groups. Culture is dynamic and its development and persistence must be understood within the context of powerful reciprocal influences that are operative as social groups interact with the constraints and opportunities of their specific environment.

6. *Concluding Recommendations*

Future research on racial/ethnic variations in health must assess comprehensively racial and ethnic status. This should include the assessment of the considerable ethnic heterogeneity that exists within all of the widely used racial categories. Second, researchers must recognize that there is no generic health model that applies equally to all racial or ethnic minority populations. Research and health policy regarding race should be attentive to the distinctiveness of each racial group, including using the terms that are broadly recognized and preferred by a substantial proportion of respondents of particular racial groups. The fluidity of racial and ethnic categories requires periodic monitoring and updating measures of racial and ethnic status. Importantly, whenever racial and ethnic data are reported, more attention must be given to the interpretation of these data. For example, it is always necessary to indicate why race or ethnicity is being utilized, the limitations of racial and ethnic data, and how the findings should be interpreted. Given the strong association between racial/ethnic status and SES, the presentation of data on racial differences should routinely stratify them for SES within racial groups. Failure to attend to these issues of presentation and interpretation may misspecify complex health risks and even perpetuate negative racial stereotypes. Finally, future research must conceptualize and operationalize increasingly the specific factors linked to race that might affect health.

See also: Conflict Sociology; Critical Race Theory; Culture as a Determinant of Mental Health; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Gender, Class, Race, and Ethnicity, Social Construction of; Health and Illness: Mental Representations in Different Cultures; Population Composition by Race and Ethnicity: North America; Race: Genetic Aspects; Race: History of the Concept; Race Identity; Racial Relations; Socioeconomic Status and Health; Socioeconomic Status and Health Care

Bibliography

Balarajan R, Raleigh V S 1995 *Ethnicity and Health in England*. Department of Health, London

Brenner M H 1995 Political economy and health. In: Amick B C III, Levine S, Tarlov A R, Walsh D (eds.) *Society and Health*. Oxford University Press, New York, pp. 211–46

Cell J 1982 *The Highest Stage of White Supremacy: The Origin of Segregation in South Africa and the American South*. Cambridge University Press, New York

Cooper R S 1994 A case study in the use of race and ethnicity in public health surveillance. *Public Health Reports* **109**(1): 46–52

Duster T 1984 A social frame for biological knowledge. In: Duster T, Garret K (eds.) *Cultural Perspectives on Biological Knowledge*. Ablex Publishing, Norwood, NJ, pp. 1–40
Elo I T, Preston S H 1992 Effects of early-life conditions on adult mortality: A review. *Population Index* **58**: 186–212
Evinger S 1995 How shall we measure our nation's diversity? *Chance* **8**(1): 7–14
Hahn R A 1992 The state of federal health statistics on racial and ethnic groups. *Journal of the American Medical Association* **267**(2): 268–71
Jaynes G D, Williams R M 1989 *A Common Destiny: Blacks and American Society*. National Academy Press, Washington, DC
Krieger N 1987 Shades of difference: Theoretical underpinnings of the medical controversy on black/white differences in the United States, 1830–1870. *International Journal of Health Services* **17**: 259–78
Krieger N 1999 Embodying inequality: A review of concepts, measures, and methods for studying health consequences of discrimination. *International Journal of Health Services* **29**(2): 295–352
Lewontin R C 1982 *Human Diversity*. Scientific American Books, New York
McLennan W, Madden R 1999 *The Health and Welfare of Australia's Aboriginal and Torres Strait Islander Peoples*. Australian Bureau of Statistics, Commonwealth of Australia
NCHS (National Center for Health Statistics) 1991 *Vital Statistics of the United States, 1988. Vol 2. Mortality, Part A*. Public Health Services, Washington, DC
NCHS (National Center for Health Statistics) 1997 US Decennial life tables for 1989–91. *United States Life Tables* **1**(1): 12–29
NCHS (National Center for Health Statistics) 1998 *Health, United States, 1998 With Socioeconomic Status and Health Chartbook*. US Department of Health and Human Services Hyattsville, MD
NCHS (National Center for Health Statistics) 1999 *Health, United States, 1999 With Health and Aging Chartbook*. US Government Printing Office, Hyattsville, MD
Polednak A P 1989 *Racial and Ethnic Differences in Disease*. Oxford University Press, New York
Sampson R J, Wilson W J 1995 Toward a theory of race, crime, and urban inequality. In: Hagan J, Peterson R D (eds.) *Crime and Inequality*. Stanford University Press, Stanford, CA, pp. 37–54
Sidiropoulos E, Jeffery A, Mackay S, Forgey H, Chipps C, Corrigan T 1997 *South Africa Survey 1996/97*. South African Institute of Race Relations, Johannesburg, South Africa
Steele C M 1997 A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist* **52**(6): 613–29
Whiteford A, Posel D, Kelatwang T 1995 *A Profile of Poverty, Inequality and Human Development*. Human Sciences Research Council, Pretoria, South Africa
Williams D R 1999 Race, SES, and health: The added effects of racism and discrimination. *Annals of the New York Academy of Sciences* **896**: 173–88
Williams D R, Collins C 1995 US socioeconomic and racial differences in health. *Annual Review of Sociology* **21**: 349–86
Williams D R, Williams-Morris R 2000 Racism and mental health: The African American experience. *Ethnicity and Health* **5**: 243–68

D. R. Williams

Ethnicity, Sociology of

Ethnicity was conspicuous in Athens, Rome, and Alexandria, and it was also a factor in the Indian caste system, the millet of Muslim societies, and in other instances of segregated groups. In these premodern cases, ethnic relations were taken for granted, and imposed, by superior authorities, traditions, abrupt power relations, or the individuals involved themselves. It is principally in modern settings that processes like the emergence of the nation-state—as the primary focus of identification—and the growth of a meritocratic–individualistic civilization have seemed to leave no room for continuous ethnic allegiances. It is also then that the persistence and increasing saliency of ethnicity have led scholars to delve into the phenomenon as a sociological issue. Max Weber (1978) spoke of ethnicity to designate collectives characterized by religious, historical, linguistic, or racial traits. Since the 1960s, following the political awakening of Blacks and other groups in the USA, 'ethnic group' became popular among sociologists while some of them proposed 'imagined communities' (Anderson 1983). Such groups respond to a variety of syndromes, and may also be factors of political change—especially in democratic settings—when they contribute to the assertion of multiculturalism and eventually evolve, in this era of globalization, into 'transnational diasporas.'

1. *A Kind of Collective Identity*

Above all, ethnicity consists of a reference to a collective identity, i.e., how and how far individuals: (a) feel committed to people whom they see as fellow-members of their group, (b) perceive that group as conveying singular, 'original' and 'unique' values, norms, or symbols, and (c) accordingly, distinguish themselves from nonmembers (see also Eriksen 1993). This notion describes an essentially subjective phenomenon and, beyond the structures and interests wherein it may concretize 'objectively,' it involves dilemmas more often than unambiguous assertions. Hence, the same identity may be phrased in different (possibly antagonistic) terms, over time (i.e., diachronically) or among different circles in the group (i.e., synchronically). A classic example is the Jewish identity. Traditional Jewishness has for long been the dominant trend among Jews, up to the advent of modernity, and it still defines Judaism by (a) its commitment to Jews' welfare anywhere, (b) its insistence on the religious faith and its constraints, and (c) its messianic aspiration to a repositioning vis-a-vis the non-Jew through Redemption. Many non-Orthodox and free-thinkers are, in contrast, (a) committed to a community which they perceive as (b) carrying universalistic values under the cover of

parochial customs, while at the same time, (c) also seeing themselves an integral, though distinct, segment of non-Jewish societies. The nationalists, as for them, speak of (a) patriotic commitment to (b) a notion of a Jewish nation that is, (c) 'territorialized' in the 'ancestral land,' as the basis of the 'normalization' of the Jewish condition among nations. These different versions pertain to a same identity only because they answer the same questions: the commitment to Jews, the perception of their singularity, and an understanding of the relation to non-Jews (Ben-Rafael 1998) (see also *Identity: Social*; *Race Identity*; *Identity Movements*). Such different versions of the same identity articulate distinct symbols or attach contrasted meanings to the same symbols. In both cases, they broadly, though not exclusively, draw these symbols, which are theirs, from the same 'store'—myths of descendancy, time-honored customs, and sanctified objects or places. As a rule, these symbols also include language, even though its relation to ethnicity is unpredictable in modern settings where political and cultural imperatives impose the public use of official languages and downgrade ethnic codes to the role of secondary vernaculars. It remains that linguistic elements—a regular register, a restricted vocabulary, typical expressions or a characteristic accent—tend to signal social intercourse involving an ethnic dimension and mark, under its different versions, a group's contrastive identity (Fishman 1989).

The different markers attached to the various versions of a same collective identity represent diverse manners of stating the group's 'uniqueness.' The comparison of these versions should reveal a common denominator that could be viewed as their deep structures. This denominator should make explicit the nature of the collective claims of the members of the group from society—whether or not these claims are consensual among them. The pertinence of this methodology is less obvious, however, when it comes to the more arduous comparative discussion of different identities, that is, of independent formulations of 'uniqueness.' In this respect, the analysis is to widen its scope to the social and cultural context, and its search for a common criterion of evaluation is to switch its focus from 'identity' to 'identification.'

2. *The Parameters of Identification*

More than the markers themselves, it is indeed their saliency in the social practice that indicates how far individuals refer to ethnicity in their daily life. This aspect points to the vitality of ethnicity, that is, the degree that ethnics identify with their particular identity through their participation in a wider society. Ethnic identification stands thus at the opposite of assimilation which depicts the abandonment by individuals of their original features and their increasing likening to people who embody, in their behavior and attitudes, the culture prevailing in society. Unlike acculturation, assimilation also includes collective identities, and researchers who have studied these processes show the role played here by three factors, namely, inequality, the dominant culture, and ethnics' own aspirations.

2.1 *Social Inequality*

The Chicago school has initiated a tradition in ethnic research which focuses on inequality in metropolitan areas; Robert Park, William Thomas, Florian Znaniecki, or Louis Wirth, for instance, emphasized race relations and immigrant groups. Later on, Leo Kuper explained ethnicity on the basis of domination relations, while Edna Bonacich investigated labor market participation. They and others also focused on ethnics' unequal access to social and educational opportunities, self-actualization, and empowerment (Giddens 1991). All in all, they show that ethnic groups tend to concentrate, and remain, in inferior strata as the outcome of two categories of factors: discrimination and weak human capital. By discrimination one means normally-'irrelevant' pressures of exclusion exerted against a group from outside. It is rationalized, especially where ethnicity is associated with race (Banton 1988), by ethnocentric and racist perceptions among the stronger group. While these attitudes may be diffused by people who feel threatened themselves, they sustain practices of inferiorization and instill deprivation that is found in ethnic identification, a manner for retaining self-dignity (see *Prejudice in Society*; *Egalitarianism, Sociology of*; *Race Relations in the United States, Politics of*). On the other hand, human capital refers to formal education, cultural predisposition, or models of behavior that account for individuals' capability to seize social opportunities. It is when members of a group exhibit weak human-capital assets that they are particularly vulnerable to discrimination. The confinement to lower status signifies that the group is of little attractiveness to other people, and this relative isolation accounts for restricted exposure to the prevailing culture (see *Inequality*). Hence, the group may be left to its parochial symbols and styles in the margins of society. By the same token, the contrary is true of mobility. Whether it concerns a small or large cohort of the group, mobility implies acculturation and lower saliency of ethnicity as a result of education, personal relations with 'others,' and participation in middle-class frameworks (see *Mobility: Social*). The extent, however, that acculturation leads to assimilation still depends on other factors.

2.2 *The Dominant Culture*

The culture which is predominant in society is another pertinent factor. The 'void of community,' it is often

contended, created by modernity induces individuals—mobile and nonmobile alike—to value the retention of community. Instrumental thinking which pervades all areas of life brings ethnics to aspire to self-expression and to assert their particular identities (Taylor 1994). More specifically, it is also contended that the 'dominant culture' is a contingency that may display diverse perspectives toward ethnicity and influence it in different manners. By this notion, one means the beliefs, values, and symbols that identify the society, according to those—the elite or larger social segments—which constitute its leading political forces. These forces give expression to the dominant culture by defining the legal frame of the social insertion of ethnics and weighing on the public opinion through media and policies. Hence, in Germany, the notion of *Volk* has for a long time stated the supremacy of the nation over the individual, and its anchorage in a given history and culture (Brubacker 1992). As a consequence, naturalization laws obstructed the acquisition of German nationality for non-Germans who reside in the country, even for decades. In the USA where an assimilationist notion of 'melting-pot' has been gradually substituted by a pluralistic concept of 'salad-bowl,' groups are entitled to retain their particularism while becoming Americans and adjusting to its culture. In France, the 'republican nation' is understood as a homogeneous whole—a 'community of citizens'—which expects new groups to integrate society on these terms (see also *Nation: Sociological Aspects*). Hence, these dominant cultures display distinct perspectives vis-a-vis ethnics' social evolution, i.e., assimilation, pluralism, or segregation.

2.3 The Group's Aspirations

The ethnics also play a part here, and it is at this point that identity is introduced as a factor of identification. In the same material and cultural circumstances, indeed, it is according to the contents of their identity that members of some groups may tend to be 'assimilationist' where others are 'retentionist' (Glazer and Moynihan 1963). Kinship ties, a common country of origin, or a collective religion are often depicted as 'primordial' bonds carrying special strength (Geertz 1963). More specifically, religion has often shown its power to retain the dedication of believers, and sustain the cohesion of communities, in front of assimilatory forces (Smith 1992). Even when challenged by modernity and secularism, it might be rephrased so as to retain relevance. At the limit, fundamentalism, which is one such rephrasing, acknowledges the predominant role of politics in modern societies to formulate the essentials of the faith in terms of a 'total' political project. Acting then as a quasi-Jacobinian force, fundamentalism, which itself stands out on behalf of a religious-ethnic collective, leaves little room in its utopia for any form of particularism (Eisenstadt 1997). It remains that as a case of ethnic community, and all other factors being equal, a religious group generally exemplifies the extreme end of 'retentionism,' the opposite of those who view assimilation as the realization of their aspiration.

3. The Diversity of Profiles

All three factors thus say 'something' of their own about ethnic identification/assimilation: (a) the concentration of a group in a lower class distances members from the surrounding culture, at the benefit of retentionism, while, in contrast, social mobility sustains acculturation; (b) a dominant culture encourages retentionism when it is segregative or pluralistic, and assimilationism when it requests cultural uniformization; (c) according to their own original perspectives, the orientations prevailing in the group influence its members whether to evince or to underrate their collective identity. Viewed together, these independent expectations widely account for the various profiles and extents that groups constitute 'contrastive entities' in society. This question puts in perspective material conditions and interests, cultural contexts and identity, and thus accounts for the maintenance, erasing, or reshaping of ethnic boundaries, e.g., what Stuart Hall calls ethnic 'cut-and-mix.' Empirical cases illustrate the syndromes yielded by this comparative methodology. American-Indian communities are an example of (a) an underprivileged category that suffered from discrimination and carried little human capital; (b) a case where the dominant culture was segregationist; and (c) a group whose basic orientations sustained the retention of its legacy and identity. The small number of acculturated but rarely assimilating mobile elements remains a part of their community best depicted as *enclave*. A low-class group such as *Afro-Americans* (until the mid-twentieth century) may, however, aspire to assimilation. Its weak collective identification leaves its few mobile elements free to aspire to assimilation. This, however, is *usurpation* for a segregationist dominant culture, which pushes them to form a 'black bourgeoisie' midway their original community and the dominant group. The *Quebecois*, a widely low-class population for decades, most often aspired to remain distinct from the assimilationist English-Canadian culture. Many stood for *autonomy*, even independence, while mobile elements split into supporters of assimilation, offered by the dominant culture and the retentionism of identity favored by the group's basic aspirations. Regarding *Brazil's Blacks*, again a disadvantaged group, no taboo, either from the dominant culture or the group, prevents assimilation. This is of little significance for the majority as only a few mobile individuals may contemplate integrating the pre-

dominantly white middle-class, leaving thereby their fellow-ethnics in *crypto-segregation*. In a context where mobility is more general in the group, *South African Indians* experienced *middle-class closure* under Apartheid, as a result of both the dominant culture's segregationism and their own reluctance to give up their particularism altogether. Ethnic markers and the concomitant adoption of traits indicating acculturation thus circumscribed the boundaries of quite a cohesive group. A model that may be labeled *margins to the center* concerns the mobile assimilationist Australian-Asian groups who are maintained at a distance by a pluralistic dominant culture. They are similar to the former case but emphasize less their own legacy, and more the markers of acculturation. Still different, the mobile *French Jews* confront a determinedly assmilationist dominant culture. Many effectively assimilate but others still share the group's original retentionism by maintaining homogeneous *networks*. The voluntary character of their allegiance accounts for the variety of their versions of collective identity. Where mobility has erased inequality and deepened acculturation, and the dominant culture and the ethnic group are both open to each other, one may finally speak of genuine *assimilation*, as illustrated by numberless groups ranging from Italians in France to Norwegians in the US.

4. *Multiculturalism, Globalization, and Transnational Diasporas*

In summary, the phrasing of their collective identity defines ethnics' self-perceived 'uniqueness' as a group; their identification/assimilation draws out their social boundaries. The identity explains the nature and consequent rationality of ethnics' claims and interests in terms of public goods, illustrating what Boudon (1993) calls 'normative rationality.' Identification accounts for the ethnics' determination to pursue these goods, as it fluctuates in the context of material and cultural conditions as well as according to the contents of their collective identity, which is a reminder of what Boudon calls 'subjective rationality' (see *Rationality in Society*). It is in this latter dynamic respect that beyond 'how society influences ethnic groups?' arises 'how ethnicity influences society?' This issue is mainly pertinent in a modern democracy, where nearly any group able to 'mobilize' its members has access to political opportunities, even when disapproved by the elite or public opinion. Parties or pressure groups presenting themselves on behalf of ethnics may gain influence, even recognition, if powerful enough to impose their participation and shrewd enough to exploit the rivalries characteristic of the political scene. Moreover, a successful group brings to ethnic politics other dormant allegiances (see also *Ethnic Conflicts*; *Ethnic and Racial Social Movements*). The risks implied by ethnic politics for the groups involved are indeed limited by the unconditional welfare rights that are anyone's in democracy, and which warrant ethnics a 'costless' liberty of political action. In turn, legitimate ethnic politics increases the systemic incoherence and fragmentation of the setting, substantiating the notion of multiculturalism (Rex 1996) by challenging the dominant culture, where it is assimilationist in its very ideological premises (see *Multiculturalism: Sociological Aspects*). Such developments are still amplified by 'globalization' processes which include new means of communication, an expansion of worldwide mass media, and improved air transport. By drastically reducing the significance of geographic distances, these circumstances stimulate new flows of emigration from all parts of the world toward wealthy countries. They also signify that integrating a new society does not necessarily imply for immigrants, emotional, cultural, and even social detachment from their countries of origin. From the onset, immigrants may institute direct contact with their families in the homeland as well as with friends and relatives living in other places. They are able to create a space of communication, including regular visits to and from the homeland and elsewhere, where the retention of languages and symbols of origin remain pertinent. On the other hand, in the context of rising multiculturalism in many of these societies which also often happens to be welfare states, such new groups also rapidly acquire the skills necessary for participation in open ethnic politics. Hence, while this global era is marked by a cultural uniformization forwarded by the influence of the West (Bauman 1998), one of the distinctive features that Western societies now share in common is their internal diversity. All these, however, do not exclude that these immigrants are not moved, as well, by the aspiration to integrate their 'choice society' by acquiring its legitimate language and culture. What these conditions indicate is the emergence of 'transnational groups' (Soysal 1994) attempting to illustrate additive bilingualism, dual culture, and at the limit, 'dual home.' While by the sequels of their original culture, they add to the diversity of their actual society (see *Migration: Sociological Aspects*), by the imprints of their experience in new cultural and social landscapes they produce the heterogeneity of their Diasporas. It is this duality that makes transnational diasporas a kind of bridge between choice societies and homelands that not only connect states but also populations.

See also: Conflict Sociology; Ethnic and Racial Social Movements; Ethnic Cleansing, History of; Ethnic Conflict, Geography of; Ethnic Conflicts; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity/Ethnicity and Archaeology; Ethnicity: Anthropological Aspects; Gender, Class, Race, and Ethnicity, Social Construction of; Inequality: Comparative Aspects; Race: History of the Concept; Race

Relations in the United States, Politics of; Racial Relations; Racism, History of; Racism, Sociology of

Bibliography

Anderson B R O'G 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Banton M P 1988 *Racial Consciousness*. Longman, London

Bauman Z 1998 *Globalization: The Human Consequences*. Columbia University Press, New York

Ben-Rafael El 1998 Quasi-sectarian religiosity, cultural ethnicity and national identity. In: Krausz E, Tulea G (eds.) *Jewish Survival: The Identity Problem at the Close of the 20th Century*. Transaction, New Brunswick, NJ, pp. 33–64

Boudon R 1993 Towards a synthetic theory of rationality. *International Studies in the Philosophy of Science* **7**: 5–20

Brubacker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Eisenstadt S N 1997 *Les Antinomies de la Modernité*. L'Arche, Paris

Eriksen T H 1993 *Ethnicity and Nationalism: Anthropological Perspectives*. Pluto Press, London

Fishman J A 1989 *Language and Ethnicity in Minority Sociolinguistic Perspective*. Multilingual Matters, Clevedon, UK

Frazier E F 1957 *The Negro in the United States*. Macmillan, New York

Geertz C 1963 The Integrative revolution. In: Geertz C (ed.) *Old Societies and New States: the Quest for Modernity in Asia and Africa*. Free Press, New York

Glazer N, Moynihan D P 1963 *Beyond the Melting Pot: the Negroes, Puerto Ricans, Jews, Italians and Irish of New York City*. MIT Press, Cambridge, MA

Rex J 1996 *Ethnic Minorities in the Modern Nation State: Working Papers in the Theory of Multiculturalism and Political Integration*. MacMillan Press, London

Smith A D 1992 Chosen peoples: why ethnic groups survive. *Ethnic Racial Studies* **15**(3): 436–56

Soysal Y N 1994 *Limits of Citizenship: Migrants and Postnational Membership in Europe*. University of Chicago Press, Chicago

Taylor C 1994 *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton, NJ

Weber M 1978 Ethnic groups. In: Roth G, Wittrich C (eds.) *Economy and Society: an Outline of Interpretive Sociology*. University of California Press, Berkeley, CA, Vol. 1, pp. 389–95

E. Ben-Rafael

Ethnoarchaeology

Ethnoarchaeology is the study of variability in material culture and its relation to human behavior and organization among living peoples in order to strengthen archaeological inference. The term first entered the literature in 1900 when the US anthropological archaeologist, Jesse Walter Fewkes, referred to himself as an 'ethno-archaeologist.' Since then, there have been a number of synonyms coined including 'action archaeology,' 'living archaeology,' and 'archaeological ethnography.' Most archaeologists prefer the term 'ethnoarchaeology' for this form of research.

1. History of Ethnoarchaeology

Its beginning is closely tied to the onset of anthropological archaeology in the USA around 1880. Most of the archaeological work was coordinated by the Bureau of Ethnology (later, the Bureau of American Ethnology) and was carried out in the American Southwest. Stimulated by the general theory developed by Lewis Henry Morgan (1877), archaeology was seen as the extension of ethnology into the past. Early workers include Fewkes, who worked with the Hopi Puebloan people (1900) and Frank Hamilton Cushing, who worked with the Zuni Puebloan people (1886). The concerns were upon the development of 'Puebloan society' through a study of its antecedents using archaeology aided by ethnology of the living Puebloan peoples.

With the publication of Morgan's final contribution in 1881, the Bureau hired an architect to document ruins and contemporary communities (Longacre 1999a). Victor Mindeleff and his younger brother spent more than a decade doing fieldwork in the US Southwest and used the data they collected among Puebloan peoples as an aid to interpreting archaeological sites (e.g., Mindeleff 1900). Both Cushing and Fewkes used design symbolism on pottery as an aid to archaeological inference.

Although from a contemporary perspective these early attempts at exploring behavior and organizational aspects of past societies seem simplistic, the questions they were concerned with have a modern tone. They did pioneer the methods of ethnoarchaeology, and they did use the results to strengthen their conclusions. At about the same time, in other parts of the world similar approaches begin to appear (e.g., Mitchell 1881).

A new general theory became dominant in US anthropology after 1910, and this had a profound effect on the nature of archaeology in the USA. It is most closely associated with Franz Boas and is often referred to as '*historical particularism*' (Darnell 1969). The emphasis on reconstructing culture history through a focus on culture 'traits' resulted in defining prehistoric cultures and placing them in time and space. The basic outline of prehistoric cultural development in the USA began to emerge as a result, but without the social or behavioral concerns of the previous era.

For nearly 50 years, anthropological archaeology continued to be dominated by this theoretical ap-

proach. As a result, archaeology became increasingly out of step with general anthropology, leading to great tensions within the field. The Kluckhohn–Taylor attack on the lack of modern directions in archaeology (Kluckhohn 1973, Taylor 1948) resulted in rapid changes in the field. By the early 1950s, explicit concern with social institutions in the past were returning to archaeology through settlement pattern studies pioneered by Gordon R. Willey (1953, 1956). The first attempt to reconstruct prehistoric social organization appeared at this time (Martin and Rinaldo 1950). Not surprisingly, a return to ethnoarchaeology appears then as well, after a hiatus of 50 years.

One of the first calls for such a return was made by the English archaeologist O. G. S. Crawford (1953), who based it upon his work in East Africa. He felt that a study of modern pottery and other materials among the living societies in the area could enrich archaeological interpretation. In 1956, Kleindienst and Watson (1956) called for 'action archaeology,' and, encouraged by Robert Braidwood, who was working in Iraq and believed that a study of modern villages, agriculture, and tools would help with an interpretation of the region's prehistory, Patty Jo Watson undertook ethnoarchaeological fieldwork in Iran (1979). The first ethnoarchaeological monograph of this new era was Raymond H. Thompson's *Modern Yucatan Maya Pottery Making* (1958).

The pace of ethnoarchaeology quickened after 1960, in response to the rise of the New Archaeology. Lewis Binford launched his fieldwork among the Nunamiut Eskimo in the 1970s (Binford 1978a). At the same time, Margaret Hardin began her lifelong interest in ceramic ethnoarchaeology through fieldwork among Tarascan Indians in Mexico (Hardin 1970) and, more recently, among Zuni potters in New Mexico (1989). Warren DeBoer and Donald Lathrap began their studies among the Conibo-Shipibo in South America (DeBoer and Lathrap 1979), as did Ian Hodder in East Africa (Hodder 1977) and Dean Arnold in Mexico and Peru (Arnold 1975, 1983, 1989). Kramer followed her earlier village ethnoarchaeology in Iran with a major project in India, where she focused on urban pottery production and distribution (Kramer 1982, 1997). Longacre began his field project among the Kalinga people of northern Luzon, Philippines in 1973, work that continues today. Brian Hayden launched the Coxoh Ethnoarchaeological Project in 1977, fieldwork that focused upon both stone tools and ceramics in the Maya Highlands (Hayden and Cannon 1984, Hayden 1987).

There are very few peoples left in the world who still make and use stone tools, especially chipped stone implements. Not surprisingly, there are relatively few ethnoarchaeological studies of such peoples. Hayden's work mentioned above as well as Richard Gould's project in Australia (1968) and recent studies by Stephen Brandt in Ethiopia on stone scrapers made and used in the processing of cow hides (personal communication) are examples. World globalization is reducing the availability of traditional crafts such as pottery making or stone tool production for ethnoarchaeological research.

Although chipped stone tools are extremely rare among modern hunter-foragers, we do have examples of ethnoarchaeological studies of bow and arrow hunters who use arrow points of metal or wood. One excellent example is the work of Griffin (1997), who has done fieldwork among Agta peoples of northeastern Luzon, in the Philippines. Most of the hunting is done by males among the Agta, but some hunting with the bow and arrow is done by skilled females as well.

Binford's ethnoarchaeological fieldwork among the Nunamiut has provided important contributions to our understanding of site formation processes (1978a, 1978b, 1979, 1980). His detailing of accumulations of faunal remains presents information that is useful in interpreting animal bone assemblages in prehistoric sites. Another important ethnoarchaeological example is continuing work among the Hadza peoples of east Africa by Bunn and O'Connell and their colleagues, focused on the implications of animal bone distributions in Hadza camps (O'Connell 1990, Bunn et al. 1991, Bunn 1993).

2. *Contemporary Approaches in Ethnoarchaeology*

Contemporary ethnoarchaeology is stimulated by several robust theories, including behavioral archaeology (Schiffer 1976) and the postmodernist inspired 'interpretive' anthropology (Hodder 1987). One area of great contribution by ethnoarchaeology lies in formation process research. Basic data, such as records of pottery use lives in ongoing societies, collected by multiple projects, have provided the basis for what has come to be called 'accumulations research' (Varien and Mills 1997). An understanding of subtle surface alterations reflecting use of pots has resulted in many insights to help archaeologists interpret prehistoric data.

One example of this is the pioneering work by Skibo (1992), which explores residue analyses with pottery collected among the Kalinga. Another area of important contribution to archaeological inference from ethnoarchaeology centers on the relationship between degree of standardization of particular products and the degree of specialization of the producers. If there is indeed a relationship, then prehistorians could look for the beginnings of standardized pottery, for example, as a correlate of the presence of craft specialists. Thus, early processes of state formation might be explored through ceramic standardization analyses. But is there really a link between standardization and specialization?

Ethnoarchaeological research would seem to be the ideal approach to explore the degree of linkage. One might begin with a field investigation among people where each household made its own pots. The Kalinga are a good example of this mode of production, and in many villages each household has at least one potter at work. Examining the degree of standardization, in this case metrical standardization, with a large sample of cooking pots reveals variation at about 12 percent. That is, in any dimension or in any ratio (height to aperture, etc.), Kalinga pottery varies about 12 percent. This provides a baseline to compare the pottery produced on a household basis with pottery made by more specialized producers, such as a group of craft specialists making pottery for market.

Several other field sites in the Philippines have now been used to provide the degree of metrical variation in pottery produced by specialized producers. The neighborhood of Paridijon, in the small city of Gubat in Sorsogon Province in southeastern Luzon, has around 70 potters working full-time (as weather and parenting duties permit) to produce thousands of kitchen pots, flower pots, and stoves per week. Measuring thousands of cooking pots from Paridijon reveals metrical variation hovering around 6 percent, roughly half that of Kalinga pottery. Using these data, a study was conducted using prehistoric cooking pots from Grasshopper Pueblo, a prehistoric community in Arizona to assess the mode of production. The Grasshopper pottery varied at 12 percent metrically, suggesting that the household mode of production was in place in this fourteenth-century community (Longacre et al. 1988).

Additional field sites have been explored that suggest that the link between specialization and standardization is not simply the unintended by-product of rote production as has been suggested, but rather is a deliberate attempt by specialized producers to make standardized products for their consumers. As a corollary of this, the older, more skilled potters make pots that are far more standardized than those made by beginner potters, even reaching an astounding 3 percent metrical range of variation (Longacre 1999b).

Identifying general principles such as these, which link variation in material culture to specific aspects of human behavior and organization, result in strengthening the power of archaeological interpretation. Ethnoarchaeology is one source for the identification of such regularities; experimental archaeology is another. A combination of the two would seem to hold special promise as a source of such principles.

Exploring performance characteristics (Schiffer and Skibo 1998) provides one example. The cooking pots produced by the craft specialists in a neighborhood of San Nicolas, a small city in Ilocos Norte Province in northwestern Luzon, are shiny and black in color. The potters explain that the black color makes their pots stand out and thus easily found by customers in the markets. This is an example of a visual performance characteristic. Consumers identify the pots from these potters as more durable than those from competing potters.

After the pots are formed and dried, the surface is covered with a red slip with a high iron content. The pots are then polished and then fired in a 'bonfire' at about 750 degrees centigrade. When they are fired, they are removed from the fire and immersed in a pile of rice chaff, and this produces a very smoky reducing atmosphere that turns the pots black. To test whether or not such treatments affect the performance characteristics of the pots, a collection of pots was commissioned for a series of experiments in the lab. Pots that had no slip or blackening were collected, as were pots that had only the red slip exterior. Pots that were blackened but that had no slip, as well as pots that were produced as the potters normally do were also collected.

Experiments were conducted on the collection at the University of Arizona in the Laboratory of Traditional Technology. Strength was measured using the three ball on one tester (Neupert 1993). Heating effectiveness was also measured by recording the length of time it took to bring an amount of distilled water to 90 degrees centigrade. These were critical performance characteristics in assessing the durability and cooking effectiveness of the pots in the sample.

The results indicate that the traditional blackened pots with red slip are not much stronger than any other variant, but that their heating effectiveness is much higher than the others. So in addition to the visual performance characteristics of the black pots, they cook more efficiently than pots with alternative surface treatments.

This approach, combining ethnoarchaeology and experimental studies, holds great promise for improving our understanding of decisions made by potters in the prehistoric past. It is a direction of study that can profitably be applied to other forms of material culture as well. This would seem especially true for the study of stone tools. As noted above, there are few people left in the world who still make and use stone tools. But there are a number of expert flint knappers today who can make replicas of stone tools that can be used in experiments under controlled laboratory conditions. Expert knappers who pioneered such an approach include Don Crabtree and François Bordes. At the turn of the twenty-first century, whole laboratories devoted to experimental archaeology are appearing, such as the Laboratory of Traditional Technology founded by Michael Brian Schiffer at the University of Arizona and the Center for Research into the Anthropological Foundations of Technology established at Indiana University by Nicolas Toth.

Traditional ethnoarchaeology continues to be carried out all over the world, leading to new insights about the relationships between variability in material

culture and the behaviors of the people responsible. Such work continues to be useful to archaeologists who employ such findings in their own research. Scott Van Keuren (1999) has drawn on Margaret Hardin's earlier (1977) work among Trascan Indian potters to infer the work of individual potters in a collection of prehistoric Puebloan pottery from Arizona. This, in turn, has led Van Keuren to explore aspects of prehistoric social organization in the area, an ultimate outgrowth from ethnoarchaeology.

New ethnoarchaeology is being carried out worldwide and promises important contributions. Mark Neupert (1999) has just completed a study of elite factionalism in a small city in the Philippines and the attempts on the part of the elites to capture the allegiance of craft specialists. The subsequent split of the potters is reflected in the different clay sources utilized by each faction. Another excellent example of contemporary ethnoarchaeology is the Phuket Project, a long-term research endeavor by Pamela Rogers and Richard Engelhardt in southwest Thailand. They have been working in the Phuket Islands with the Chaw Lay people, a group of maritime foragers producing shell middens that have been extensively studied and documented. They have developed a model of maritime adaptation of coastal peoples, and are testing the model with archaeological data from the islands of Hong Kong (Rogers 1999). Although opportunities for traditional ethnoarchaeology are diminishing, work continues, and we can expect new insights into artifact variability and site formation processes to result.

See also: Anthropology, History of; Ceramics in Archaeology; Hunter–Gatherer Societies, Archaeology of

Bibliography

Arnold D E 1975 Ceramic ecology of the Ayacucho Basin, Peru: Implications for prehistory. *Current Anthropology* **16**: 183–206

Arnold D 1983 Design structure and community organization in Quinua, Peru. In: Washburn D K (ed.) *Structure and Cognition in Art*. Cambridge University Press, Cambridge, UK

Arnold D 1989 Patterns of learning, residence, and descent among potters in Ticul, Mexico. In: Shennan S (ed.) *Archaeological Approaches to Cultural Identity*. Unwin Hyman, London

Binford L R 1978a *Nunamiut Ethnoarchaeology*. Academic Press, New York

Binford L R 1978b Dimensional analysis of behavior and site structure: Learning from an Eskimo hunting stand. *American Antiquity* **43**: 330–61

Binford L R 1979 Organization and formation processes: Looking at curated technologies. *Journal of Anthropological Research* **35**: 255–73

Binford L R 1980 Willow smoke and dogs' tails: Hunter-gatherer settlement systems and archaeological site formation. *American Antiquity* **45**: 4–20

Bunn H T 1993 Bone assemblages at base camps: A further consideration of carcass transport and bone destruction by the Hadza. In: Hudson J (ed.) *From Bones to Behavior: Ethnoarchaeological and Experimental Contributions to the Interpretation of Faunal Remains. Occasional Paper 21*. Center for Archaeological Investigations, Southern Illinois University Carbondale, Carbondale, IL, pp. 156–68

Bunn H T et al. 1991 Bone distribution on a modern East African landscape and its archaeological implications. In: Clark J D (ed.) *Cultural Beginnings: Approaches to Understanding Early Hominid Life-ways in the African Savanna*. R. Habelt, Bonn, Germany, pp. 33–54

Crawford O G S 1953 *Archaeology in the Field*. Praeger, New York

Cushing F H 1980 Preliminary notes on the origin, working hypothesis, and primary researches of the Hemenway Southwestern Archaeological Expedition. In: *Proceedings of the 7th International Conference of Americanists*, Berlin, pp. 151–94

Darnell R 1969 The development of American anthropology, 1879–1920: From the Bureau of American Ethnology to Franz Boas. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA

DeBoer W, Lathrap D 1979 The making and breaking of Shipibo-Conibo ceramics. In: Kramer C (ed.) *Ethnoarchaeology: Implications in Global Perspective*. Smithsonian Institution Press, Washington, DC

Fewkes J W 1900 Tusayan migration traditions. In: *Nineteenth Annual Report of the Bureau of American Ethnology, 1897–98*. Bureau of American Ethnology, Washington, DC, pt. 2, pp. 577–633

Griffin P B 1997 Technology and variation in arrow design among the Agta of northeastern Luzon. In: Knecht H (ed.) *Projectile Technology*. Plenum Press, New York

Hardin M 1970 Design structure and social interaction: Archaeological implications of an ethnographic analysis. *American Antiquity* **35**: 332–43

Hayden B (ed.) 1987 *Lithic Studies Among the Contemporary Highland Maya*. University of Arizona Press, Tucson, AZ

Hodder I 1977 The distribution of material culture items in Baringo district, western Kenya. *Man* **12**: 239–69

Hodder I (ed.) 1987 *The Archaeology of Contextual Meanings*. Cambridge University Press, Cambridge, UK

Kleindienst M, Watson P J 1956 Action archaeology: The archaeological inventory of a living community. *Anthropology Tomorrow* **5**: 75–8

Kluckhohn C 1973 The conceptual structure in Middle American studies. In: Hay C L (ed.) *The Maya and Their Neighbors*. Cooper Square Publishers, New York

Kramer C 1982 *Village Ethnoarchaeology, Rural Iran in Archaeological Perspective*. Academic Press, New York

Kramer C 1997 *Pottery in Rajasthan: Ethnoarchaeology in Two Indian Cities*. Smithsonian Institution Press, Washington, DC

Longacre W A 1999a Why did the BIA hire an architect? *Journal of the Southwest* **41**: 359–69

Longacre W A 1999b Standardization and specialization: What's the link? In: Skibo J M, Feinman G M (eds.) *Pottery and People: Dynamic Interactions*. University of Utah Press, Salt Lake City, UT

Longacre W A, Kvamme K, Kobayashi M 1988 Southwestern pottery standardization: An ethnoarchaeological view from the Philippines. *The Kiva* **53**: 110–21

Martin P S, Rinaldo J B 1950 Sites of the Reserve Phase, Pine Lawn Valley, Western New Mexico. *Fieldiana: Anthropology* **38**

Mindeleff C 1900 Localization of Tusayan clans. *Nineteenth Annual Report of the Bureau of American Ethnology, 1897–98*. Government Printing Office, Washington, DC, pt. 2, pp. 635–53
Mindeleff V 1891 A study of Pueblo architecture in Tusayan and Cibola. In: (ed.) *Eighth Annual Report of the Bureau of American Ethnology, 1886–87*. Bureau of American Ethnology, Washington, DC, pp. 3–228
Neupert M 1994 Strength testing archaeological ceramics: A new perspective. *American Antiquity* **59**: 409–23
Neupert M 1999 Potters and politics: An ethnoarchaeological study of ceramic manufacture in Paradijon, Gubat, Sosogon, the Philippines. Ph.D. thesis, University of Arizona, Tucson, AZ
O'Connell J F et al. 1990 Reanalysis of large mammal body part transport among the Hadza. *Journal of Archaeological Science* **17**: 301–16
Schiffer M B 1976 *Behavioral Archaeology*. Academic Press, New York
Skibo J M 1992 *Pottery Function: A Use-alteration Perspective*. Plenum, New York
Taylor W W 1948 A study of archaeology. *Memoirs of the American Anthropological Association* **69**
Thompson R H 1958 *Modern Yucatecan Maya Pottery Making*. Society for American Archaeology
Van Keuren S 1999 *Ceramic Design Structure and the Organization of Cibola White Ware Production in the Grasshopper Region, Arizona*. Arizona State Museum, Tucson, AZ
Varien M D, Mills B J 1997 Accumulations research: Problems and prospects for estimating site occupation span. *Journal of Archaeological Method and Theory* **4**: 141–91
Watson P J 1979 *Archaeological Ethnography in Western Iran*. The University of Arizona Press, Tucson, AZ
Willey G R 1953 *Prehistoric Settlement Patterns in the Virú, Peru*. Bureau of American Ethnology, Washington, DC
Willey G R (ed.) 1956 *Prehistoric Settlement Patterns in the New World*. New York

W. A. Longacre

Ethnobiology

1. What is Ethnobiology?

Traditional societies, often referred to as indigenous or tribal people, usually accumulate significant environmental knowledge through their everyday experience with nature and natural resources. The intimate relationship between their livelihood and nature commonly gives rise to an ecocentric perspective, reflected in attitudes towards plants, animals, water, and soil, in which nature and society are viewed as an indivisible whole. Such perspectives are concretized in iconography and imagery of the sculptural forms, and the associated ethics are transmitted through tales and legends (Vatsayan 1993).

One of the concerns of ethnobiologists and related researchers is to document and assess the value of ethnobiological information, especially in regard to rapidly depleting biodiversity and other forms of ecosystem degradation. Unlike modern human societies whose impact upon ecosystems may be seen as exogenous transformations, traditional societies tend to influence the functioning of ecosystem from within (Ramakrishnan et al. 1998) in order to derive a variety of benefits. These benefits are of three kinds: (a) economic—traditional crop varieties and lesser known plants and animals of food value, medicinal plants, and so forth harvested from the wild (Ramakrishnan 1992); (b) ecological/social—manipulation of biodiversity for coping with uncertainties in the environment and global change, for controlling soil water regimes and hydrology, soil fertility management through soil biological processes, and for efficient organic residue management; and (c) ethical—cultural, spiritual, and religious belief systems centered around the concept of the sacred species, sacred groves, and the sacred landscapes.

2. The Context

Ethnobiology started off as an appendage to classical taxonomy and systematic biology, essentially listing species collected from the wild and used by traditional societies. Whilst there is much accumulated literature of a descriptive nature on the food and medicinal species used by traditional societies from different parts of the world, which in itself is important (National Academy of Sciences 1975, Jain 1991, Berlin 1992, Hladik et al. 1993), it is only recently that ethnobiology has begun to look at the dynamics of the relationships existing between individual species and populations, and between ecosystems and landscapes.

A pioneer of ethnobiology, Conklin (1954) distinguished two distinct shifting agricultural (swidden) systems among the Hanunoo forest farmers in the Philippines: (a) the 'partial' performed by migrants from outside and using swidden as a simple means of producing a crop and (b) the 'integral' reflecting the traditional, year-round, community-wide, largely self-contained and ritually sanctioned way of life. Subsequently, a plethora ethnobiological research has followed, exemplified by the work of Kunstadter and Chapman (1978) which classifies swidden systems in Thailand on the basis of the use of resources and the relationships between cultivation and fallow periods in these agroecosystem types. Ramakrishnan (1992) has further elaborated this typology for north-east India, taking into consideration ecological, socio-economic, and cultural considerations. 'Traditional Ecological Knowledge' (TEK) emerges from such work and constitutes a rapidly evolving area of research (Johannes 1989), with serious implications for our understanding of the human dimensions of ecology. That TEK is often embedded in local belief

systems blurs the heretofore boundaries between ethnobiology and ethnology (Ramakrishnan et al. 1998).

There is an increasing realization that traditionally valued species could play a more important resource role for society at large, including food and medicinal resources (National Academy of Sciences 1975). More recently, however, interest in TEK has moved in the direction of understanding the interconnections that exist between ecological and social processes, determining the functional attributes of ecosystems/landscapes (Ramakrishnan 1992), and exposing the long-term interactions between nature and society that give rise to ecosystems that might otherwise be seen as 'natural' (Alcorn 1984). The way in which traditional societies perceive and manipulate biodiversity in the landscape, both in space and time, to ensure ecosystem stability and resilience, and have evolved ecologically sound biotechnologies of land management, such as those geared for soil fertility and soil water enhancement, to cite two examples, are now deemed important for environment and development and, increasingly, for global change science (Ramakrishnan et al. 1996). At the rate at which global change is occurring, a major proportion of all species on earth will be lost over the next century, and yet it is those species that we need to build a secure future (Raven 1998). TEK offers improved understanding of ecosystem maintenance and, hence, species preservation as well as holistic and cost-effective approaches to sustainable development.

3. *Economic Benefits*

There have been various attempts to evaluate the economic value of tropical forest biodiversity used by traditional societies, apart from the fact that these species provide basic subsistence. Such evaluations are partial, largely concerned only with the direct uses, and rely upon various assumptions about and mechanisms for attributing market values. Examples include an annual value of US$16–22 per ha for the forests near Iquitos, Peru, in Amazonia (Padoch and de Jong 1989), and US$117–244 per ha for non-timber forest products used by forest dwellers from different parts of India (Chopra 1993).

3.1 *Medicinal Plants and Alternate Medical Systems*

The ethnobiology of medicinal plants from the tropics has been rediscovered, complete with an understanding of how traditional uses reflect alternative medical systems. The traditional system of medicine in India, *Ayurveda*, provides many medicines from the Neem tree, *Azadirachta indica*. The recent commercial patenting of the active principle Margosan-O derived from this tree, which has been extracted successfully, chemically stabilized, and patented by Western industry for its pesticidal properties, is an elegant example of commercial exploitation of traditional knowledge. The term exploitation is appropriate, given that this species has been part of the Indian traditional systems of medicine since antiquity, and yet the economic rewards of the patent do not return to the people on whose knowledge the patent is partially based.

Alternate medical systems currently form the basis of primary healthcare for a very large number of the world's population. The role of these indigenous practices in addressing the healthcare needs of developing countries, integration of these alternate systems into the national planning process, and investments for capacity building to preserve and perpetuate this traditional knowledge and for renewed research to build upon this knowledge base, increasingly is recognized (Bodeker 1994). In spite of a variety of myths that marginalized traditional healthcare in the past (Fig. 1), some developing countries, such as India and others from Asia, have taken steps to strengthen research on these remedies and to integrate the holistic practices into the national healthcare. These events challenge the belief that Western medicine is the only effective way of dealing with health problems and seek to provide complementary therapy.

TEK health systems have served some 2,800 culturally distinct communities living in India composed of over 50 million 'tribals' belong to over 600 ethnic communities (Dube 1998). About 76 species of animals—16 invertebrates and about 60 vertebrates—are known to be part of the tribal medicine in India alone. Over 9,500 wild plants are recognized to be of ethnobotanical value of which over 7,500 as medicines, 3,900 as edibles, 500 for fibre, 400 as fodder, 300 as pesticides; many of them have the potential for commercial exploitation (Ministry of Environment & Forests 1994). The Ayurveda system, noted above, can be traced to *Vedic* (Hindu religious texts of antiquity) times (circa 1500 BC). With vegetable products dominating this ancient medical system, diet, and drugs derived from plants were considered to be part of a holistic approach to treatment of diseases. Reserpine yielding *Rauwolfia serpentina* for hypertension and a gum that is hyoplipidemic drug yielding, *Commiphora wightii*, are just two examples of the untapped wealth of resources available in this system. Apart from this, there are a variety of traditional medical practices, including the *Unani* system largely restricted to the northern part of India and the *Siddha* system in the southern region. With such a rich tradition in India alone and equally diverse systems elsewhere in the tropical world, the opportunities to develop refined drugs to feed the global market are vast. These cheap herbal remedies are part of a thriving alternate system of medicine, with minimal side effects in clinical experience (Valiathan 1998).

(a) Facilitation procedures: induction of early fruiting in Durian (*Durio zibethinus*) through grafting techniques; facilitating fruit production and their collection at lower levels of the tree such as Citrus and the Asiatic jackfruit as in Asian and Central American tropics; or even deflecting them away from the trunk, but at a lower level through a bamboo blade wedge inserted into the fruit producing stem of tall growing banana in Thailand; shaping square-section bamboos that are easier to handle for scaffolding, by placing mould of four wooden plates around supple new shoots, as practiced in the Philippines

(b) Increasing production: surrounding the tuber of yam with hollow banana stem, before planting, to encourage tuber growth through a humid environment around it, as done in Madagascar; girdling the stem of Cassava plant about 10 cm above the soil or grafting done between two species of Cassava done by Indonesian farmers to promote root growth and thus increase tuber production; fusion of seedlings of many fruit trees such as citrus, durian, mango, jackfruit, and starfruit done in Thailand and Indonesia to stimulate tree growth and production; watering the fruit directly through capillary action of a wick, as done in arid regions of Sri Lanka for enlarging the size of the gourds; using whole trees as physical barriers to harvest water from mists, using the Garoe tree (*Ocotea foetens*) or *Agave fourcroydes*, as done in Canary Islands or using Olive trees as in Oman; tying stones by many in the Asian tropics, at the tip of the developing snake gourd, to produce straight large fruits; pantropical technique of rapid multiplication of many tree species through pole cuttings; placing clumps of soil between leaves of young trees of sugar palms in Bangladesh to increase plant growth and sap production

Figure 1
Traditional manipulation techniques used by rural communities (Halle 1996)

3.2 Lesser Known Plants of Food Value

Humans have used a few thousand plant species and a large number of animals as a source of food collected from the wild. Of these, only a small fraction has been domesticated, and still fewer numbers are used extensively. Today, about 20 plant species largely sustain the entire world population with food. Alarms have been raised, however, that this reliance may be imprudent and concern mounts to consider an increase in the use of lesser known plants to meet food demands of a growing world population, especially those with less access to the dominant crops, to cover the possible risk involved in our dependence on those species, and to explore pathways for agriculture less dependent on high energy input, modern agriculture alone. Many of the lesser crop species are cultivated by traditional societies as part of multispecies complex agroecosystems, others are collected from the wild (Swift et al. 1996). These considerations prompted a panel of experts of the US National Academy of Sciences (1975) to evaluate a range of under-exploited tropical plants of promising economic value, and select list of a few dozen species that, in their opinion, hold promise for development. Similarly, another set of promising leguminous species were identified for their value as food, fodder, timber, gum, green manure, and soil reclamation (National Academy of Sciences 1979).

3.3 Traditional Tropical Phytopractices

Moving from individual to collective species, Halle (1996) identifies a variety of manipulation techniques and practices applied to individual or groups of plants in order to improve the quality and quantity of products obtained from them (Fig. 2). He concludes that these traditional technologies from diverse geographical regions share some common features based on an intimate knowledge of plant growth that offer possibilities for transfer to other systems of plant management.

4. Ecological/Social Implications

4.1 Interconnections between Ecological and Social Processes

Historically, scientific research on biodiversity in ecosystem functioning omit humans in the definition of the ecosystem boundary, although the synergism between human activity, such as burning, and the resulting landscape have long been known. Traditional societies, intimately linked to the environment for their livelihoods, afford the opportunity to examine

Traditional medicines are of value only when their active ingredients are known, and they are purified (*isolated?*) for mass production

Based on screening done by various drug development programmes, minimal therapeutic benefit is to be found in traditional medicines

Traditional health systems may have some value for treating chronic, low-level conditions but are of no value to treat acute or emergency care

Little scientific knowledge is available on the safety and efficacy of traditional medicines, and research efforts should be directed to toxicity and efficacy research

The global value of traditional medical knowledge is limited to providing leads for developing Western medicines and for tropical biodiversity conservation

Figure 2
Myths contributing to the marginalization of traditional healthcare (Bodeker 1994)

their part in the functioning of the ecosystem, especially in the tropics. In such contexts, the role of biodiversity in ecosystem functioning portends to change substantially as the social, economic, and cultural dimensions determining ecosystem properties are revealed. Such a shift in understanding the interconnections between ecological and social processes has implications for sustainable management of natural resources in developing countries in which local communities have a participatory role.

In this new approach, the village is seen as an ecosystem, with all its ramifications for agriculture, animal husbandry, and the domestic sector embedded with the forest and forest-related activities, such as hunting and gathering of food, fodder, fuel wood and medicinal plants collection, and forest-linked traditional farming practices, such as shifting agriculture (Ramakrishnan 1992, Swift et al. 1996). This approach integrates humans within the ecosystem. It implies that much ecosystem structure and function must be understood as an integrative process and that conserving biodiversity is crucial for the immediate survival of people and ecosystems.

4.1.1 Traditional agroecosystems. Traditional societies maintain a variety of complex, multispecies agroecosystems, operated under varied levels of production intensity. These range from casually managed shifting agricultural systems through an array of rotational fallows, agroforestry systems, compound farms, traditional cash cropping systems, and crop rotation systems at the middle intensity levels, to modern high input agriculture. The ecological complexity of these agroecosystems are due to biodiversity (subspecific and species level—crop and associated biodiversity) both in space and time.

4.1.2 The concept of home garden. Of all the traditional agroforestry systems, the 'home garden' of the tropics, variously called kitchen, house, or forest garden, deserves special attention. They display much variation in their structural and functional attributes, depending on the ecological and social settings in which they occur. Imitating a natural forest with a highly stratified and compacted set of economically important trees, shrubs, and herbs in small plots of 0.5 to 2 ha, these gardens can have over a hundred species (Millat-e-Mustafa 1998). The household is able to obtain many of its requirements from this system year round, such as food products, firewood, spices, ornamentals, and medicinal plants. Detailed economic and energy output/input analysis done on the north-east Indian systems (Ramakrishnan 1992) suggest that, generally speaking, these systems are highly efficient.

Contrary to suggestions that home gardens are confined only to more fertile soils, studies in Cherrapunji in north-east India indicate that they can also be a TEK response to declining soil fertility and site degradation from shifting agriculture (Ramakrishnan 1992, Millat-e-Mustafa 1998). In this case, more fertile conditions are humanly induced. Indeed, TEK-based changes in garden diversity, structure, and management have been reported from the upland central region of Mexico as a response to industrialization and population increase (Gliessman 1990). Such adaptations are a response to labor constraints, available off-farm employment opportunities, and insurance against possible loss of outside income.

Much more needs to be known about the organization and functioning of these unique human-managed ecosystems. A redeveloped system built upon TEK is critical to meet with sustainable livelihood concerns and provide cash income to traditional societies in the biodiversity-rich regions

of the world, especially in the context of global environmental change.

4.1.3 Agroecosystem biodiversity manipulations. It is not only the mere presence of biodiversity and the functional role it has for traditional societies that is significant, but the manner in which these societies manipulate this biodiversity for ecosystem functional attributes and landscape integrity. In the shifting agriculture of the hills of north-eastern India, for example, the number of species in a mixed cropping system declines drastically with shortening of the agricultural cycle (Ramakrishnan 1992). The farmer also shifts emphasis from cereals under a long 30-year agriculture cycle, to tuber and vegetable crops under a shorter 5-year cycle, which are more nutrient-use efficient species. Even on the same slope, such crops are emphasized on the top of the slope and the less efficient ones are largely placed towards the base, reinforcing the diversity of species while improving production and risk aversion. Through mixed cropping involving a large number of species in space and time, and traditional weed management strategies, shifting agricultural farmers in north-east India, as elsewhere, ensure effective checks on nutrient loss during the cropping phase. The traditional weed management practice, where about 20 percent of the weed biomass is left undisturbed in the plot by the shifting agriculture farmer is also a practice common to the Mayan agriculture in Mexico, with implications for agroecosystem sustainability (Altieri and Liebman 1988).

Traditional societies also ensure the integrity of the systems at the level of the landscape. The Apatani tribe in north-east India have evolved elaborate wet rice cultivation systems that are adapted to gradients in soil fertility and water availability, linked with traditional management options possible within the landscape (Ramakrishnan 1992). Widening plots by digging adjacent higher ground down to an irrigable level seems to be successful responses to population increase and new market opportunities, as also demonstrated for Tara'n Dayaks of West Kalimantan in Indonesia (Brookfield and Padoch 1994).

The lessons learnt from such manipulations of biodiversity done both in space and time as adaptation to a variety of environmental changes involving population pressure, land degradation, biological invasion, and climatic uncertainties, are important for biodiversity management in resource-rich regions of the tropics (Ramakrishnan et al. 1996). These are being used to design strategies for agroecosystem redevelopment (Swift et al. 1996).

4.1.4 Pathways for agroecosystem redevelopment. TEK is embedded in the complex, multispecies agroecosystems maintained by their various custodians and offers a basis for more environmentally sustainable systems of resource production. The objective of TEK-based redevelopment efforts is to maximize production focused on *in situ* conservation of agro-biodiversity (Swift et al. 1996) and on strengthening internal processes for relative stability and resilience within the system (Altieri and Liebman 1988, Gliessman 1990, Ramakrishnan 1992, Woomer and Swift 1994). To achieve these goals, TEK focuses on cropping patterns based with such attributes, among others, as (a) optimization of production along a nutrient gradient where it occurs, (b) synchrony in nutrient release from the soil and uptake by the crop, (c) efficient recycling of biomass residues, and (d) weed management rather than weed control.

This approach to developing agroecosystems stands in contrast to the 'green revolution pathway' based on high intensity management predicated on external energy subsidies. To reach it often requires an 'incremental pathway' (Swift et al. 1996) which builds upon traditional systems in a step by step manner sensitive to ecological, social, and cultural constraints under which traditional societies function. This pathway builds towards a 'contour' one in which the full range of agroecosystem principles are employed to reach ecologically attuned but high levels of production.

4.1.5 Manipulating natural/crop biodiversity through keystone species. The role of socially selected, ecological keystone species in conserving and enhancing biodiversity and manipulating functions in forest ecosystems has not been explored adequately. Keystone species provide critical ecosystem functions that help to preserve biodiversity. Often they are also socially or culturally valued as well. Therefore, they can be used not only to manage pristine ecosystems (Ramakrishnan 1992), but also to build biodiversity in degraded natural and human-managed ecosystems through appropriate rehabilitation strategies (Wali 1992, Lamb and Tomlinson 1994) that ensure people's participation (Ramakrishnan et al. 1994). For example, in areas of north-east India where the shifting agricultural cycle is less than 5 years and the landscape is highly degraded, a legume of lesser known food value, *Flemingia vestita*, is socially valued and used to support a 1–2 year fallow and even non-fallow system of cultivation. By fixing 250 kg of nitrogen per hectare per year, this keystone species ensures sustainability of these low-input agroecosystems, under conditions of extreme pressure on the land and low soil fertility (Ramakrishnan 1992). The adjacent successional forests are also enhanced by a socially selected keystone species. Nepalese alder (*Alnus nepalensis*), a nitrogen-fixing species, and many bamboo species (*Dendrocalamus hamiltoni*, *Bambusa tulda*, and *B. khasiana*) with ability to conserve nitrogen, phosphorus, and potassium play a key role.

The interphase between ecological and social processes, exemplified in socially selected keystone species, is critical for biodiversity management in the context of rapid change, but has only just started to receive research attention. Such keystone species may also be found within the sacred groves of traditional societies, as in the case of *Englehardtia spicata*, *Echinocarpus dasycarpus*, *Sysygium cuminii*, and *Drimycarpus racemosus* which conserve high levels of nitrogen, phosphorus, and potassium in otherwise highly infertile soils in north-east India. They are also found at the landscape level within indigenous 'forest reserves' set aside within the forest-agriculture landscape as found among the Tara'n Dayaks of West Kalimantan, Indonesia (Padoch 1993) and elsewhere (Alcorn 1984, Ramakrishnan et al. 1994). The species within these reserves are protected by strictly enforced, community rules of use, with implications for sustainability.

5. *Ethical Considerations. Conserving the Sacred*

In some cases, refugia in a given region are given the status of 'sacred groves,' complete with cultural beliefs, institutions, and practices that create 'social fencing' for relict ecosystems. Often these groves occur within derelict landscape, otherwise heavily affected by humans. To this concept can be added those of the 'sacred species' and 'sacred landscape.' The former are keystone or resource critical species as described above (e.g., sacred Basil, *Ocimum sanctum*, and Neem, *Azadirachta indica*). Sacred landscapes scale up from species and groves to large areas intermeshed natural and human-made ecosystems. The Ganga watershed of the Central Himalayan Garhwal region in India, for example, consists of a variety of landscapes sacred to Buddhists (Ramakrishnan et al. 1998), while sacred mountains exist elsewhere, such as the holy hills of the Dai of Xishuangbann in Yunnan province in China (Messerli and Ives 1997). Such landscapes and related uses are ancient in time and worldwide in distribution (Monasterio 1994).

The state-of-art ecological knowledge on conserving this sacred cultural heritage is now synthesized (Ramakrishnan et al. 1998) with a view towards biodiversity conservation and seeking new paradigms for sustainable management of natural resources. Ethnobiology has, thus, evolved through TEK towards a holistic view in which ecology, economics, and ethics combine. A major aim is to recast the circular connections among them, now broken (Bormann and Kellert 1991), in order to provide alternative understanding and means of addressing questions of environment and development.

6. *Conclusions*

Ethnobiology has moved from an important descriptive stage of study to an even more important understanding of the linkages between ecological and social processes as they affect biodiversity and ecosystem functioning. This understanding, in turn, is being used to address biodiversity conservation and sustainable land-use management. Ethnobiology also lends itself to system-level assessments, linking its process studies to traditional agroecological management strategies, such as water harvesting systems, which can be used to produce for human consumption while triggering ecological rehabilitation. It is heartening to witness the growing recognition of the historic and existing interaction of various cultures with their surrounding ecosystems and landscapes and the extent to which these interactions have shaped them. Thus, ethnobiology increasingly promotes synergism between science and world views arguing against the segregation of the biosphere into the 'natural' and the 'people,' and for holistic concepts, such as the 'biocultural region' (e.g., Walton and Bridgewater 1996). Based on this movement, ethnobiology is becoming increasingly relevant in the area of sustainable management of natural resources, with concerns for livelihood and development of traditional societies based on a value system that they understand, appreciate, and therefore engage.

See also: Alternative and Complementary Healing Practices; Folk Religion; Healing; Medical Geography; Medical Sociology; Religion: Psychiatric Aspects; Shamanism

Bibliography

Alcorn J B 1984 *Huastec Mayan Ethnobotany*. University of Texas Press, Austin, TX

Altieri M A, Liebman M (eds.) 1988 *Weed Management in Agroecosystems: Ecological Approaches*. CRC Press, Boca Raton, FL, p. 354

Berlin B 1992 *Ethnobiological Classification: Principles of Categorization of Plants and Animals in Traditional Societies*. Princeton University Press, Princeton, NJ

Bodeker G 1994 Traditional health knowledge and public policy. *Nature and Resources* **30**: 5–16

Bormann F H, Kellert S R (eds.) 1991 *Ecology, Economics, Ethics: The Broken Circle*. Yale University Press, New Haven, CT, p. 233

Brookfield H, Padoch C 1994 Appreciating agrodiversity: A look at the dynamism and diversity of indigenous farming practices. *Environment* **36**: 6–11, 37–45

Chopra K 1993 The value of non-timber forest products: An estimation for tropical deciduous forests in India. *Economic Botany* **47**: 251–7

Conklin H C 1954 An ethnoecological approach to shifting cultivation. *Transactions of the New York Academy of Sciences. Ser. 2* **17**: 133–42

Dube S C (ed.) 1998 *Antiquity to Modernity in Tribal India, Vol. 1*. Inter-India Publications, New Delhi, p. 358

Gliessman S R 1990 Integrating trees into agriculture: The home garden agroecosystem as an example of agroforestry in the tropics. In: Gliessman S R (ed.) *Agroecology: Researching the Ecological Basis for Sustainable Agriculture*. Springer-Verlag, New York, pp. 160–8

Halle F 1996 Discovering traditional tropical phytopractices. *Nature and Resources* **32**: 31–8
Hladik C M, Hladik A, Linares O F, Pagezy H, Semple A, Hadley M 1993 *Tropical Forests, People and Food: Biocultural Interactions and Applications to Development*. UNESCO-MAB Book Series 13, UNESCO, Paris and Parthenon Pub. Group, Carnforth, UK, p. 852
Jain S K (ed.) 1991 *Contribution to Indian Ethnobotany*. Scientific Publishers, Jodhpur, India
Johannes R E 1989 *Traditional Ecological Knowledge: A Collection of Essays*. IUCN, Gland, Switzerland, p. 64
Kunstadter P, Chapman E C 1978 Problems of shifting cultivation and economic development in northern Thailand. In: Kunstadter P, Chapman E C, Sabhasri S (eds.) *Farmers in the Forest: Economic Development and Marginal Agriculture in Northern Thailand*. East-West Center, University Press of Hawaii, Honolulu, pp. 3–23
Lamb D, Tomlinson M 1994 Forest rehabilitation in the Asia–Pacific region: Past lessons and present uncertainties. *Journal of Tropical Forest Science* **7**: 157–70
Messerli B, Ives J D (eds.) 1997*Mountains of the World: A global Priority*. Parthenon, Carnforth, UK, p. 495
Millat-e-Mustafa M 1998 Overview of research in home garden systems. In: Rastogi A, Godbole A, Shengji P (eds.) *Applied Ethnobotany in Natural Resource Management—Traditional Home Gardens*. International Centre for Integrated Mountain Development, Kathmandu, Nepal, pp. 13–37
Ministry of Environment & Forests 1994 *Ethnobotany in India: A Status Report*. Ministry of Environment & Forests, Government of India, New Delhi, p. 68
Monasterio M 1994 Traditional prehistoric ecotechnologies for the management of biodiversity in Latin America. *Biology International* **32**: 12–22
National Academy of Sciences 1975 *Underexploited Tropical Plants with Promising Economic Value*. National Academy of Sciences, Washington, DC, p. 189
National Academy of Sciences 1979 *Tropical Legumes: Resources for the Future*. National Academy of Sciences, Washington, DC, p. 331
Padoch C 1993 Fruits of diversity. *Pacific Discovery* **46**: 30–45
Padoch C, de Jong W 1989 Production and profit in agroforestry: An example from the Peruvian Amazon. In: Browder J O (ed.) *Fragile Lands of Latin America: Strategies for Sustainable Development*. Westview, Boulder, CO
Ramakrishnan P S 1992 *Shifting Agriculture and Sustainable Development: An Interdisciplinary Study from North-Eastern India*. UNESCO-MAB Series, Paris and Parthenon, Carnforth, UK, p. 424 (republished by Oxford University Press, New Delhi, 1993)
Ramakrishnan P S, Campbell J, Demierre L, Gyi A, Malhotra K C, Mehndiratta S, Rai S N, Sashidharan E M (eds.) 1994 *Ecosystem Rehabilitation of the Rural Landscape in South and Central Asia: An Analysis of Issues*. UNESCO (ROSTCA), New Delhi, India, p. 29
Ramakrishnan P S, Das A K, Saxena K G (eds.) 1996 *Conserving Biodiversity for Sustainable Development*. Indian National Science Academy, New Delhi, India, p. 246
Ramakrishnan P S, Saxena K G, Chandrasekara U 1998 *Conserving the Sacred: For Biodiversity Management*. UNESCO, New Delhi, India, p. 480
Raven P H 1998 Biodiversity and the human prospect. In: Chou C H, Shao K T (eds.) *Frontiers in Biology: Biotechnology and Sustainable Agriculture*. Academia Sinica, Taipei, pp. 1–9
Swift M J, Vandermeer J, Ramakrishnan P S, Anderson J M, Ong C K, Hawkins B 1996 Biodiversity and agroecosystem function. In: Mooney H A, Cushman J H, Medina E, Sala O E, Schulz E-D (eds.) *Functional Roles of Biodiversity: A Global Perspective*. Wiley, Chichester, UK, pp. 261–98
Valiathan M S 1998 Healing plants. *Current Science* **75**: 1122–67
Vatsayan K 1993 *Prakriti*. Indira Gandhi National Centre for the Arts, New Delhi, India, p. 64
Wali K (ed.) 1992 *Ecosystem Rehabilitation. Vol. 2. Ecosystem Analysis and Synthesis*. SPB Academic, The Hague, p. 388
Walton D W, Bridgewater P 1996 Of gardens and gardeners. *Nature and Resources* **32**: 15–9
Woomer P L, Swift M J (eds.) 1994 *The Biological Management of Tropical Soil Fertility*. Wiley, Chichester, UK, p. 243

P. S. Ramakrishnan

Ethnocentrism

Ethnocentrism is a term that passed from the social sciences into broader English discourse during the twentieth century. In current usage, it means culturally biased judgment, i.e., applying the frame of reference provided by one's culture to an object, action, person, or group of a different culture. The term as introduced by William Graham Sumner in his seminal treatise *Folkways* (Sumner 1906) was explicitly sociological and included biased judgment as one of several features in a theory linking group solidarity with intergroup antagonism as a universal characteristic of human societies. Sumner's theory influenced research on ethnic relations and nationalism during the first half of the twentieth century. Later, the critique of Sumner by Merton and Rossi (1957) gave attention to the concept of ethnocentrism as problematic in social research, and LeVine and Campbell (1972) used the term as a rubric under which to review theories of intergroup relations and attitudes and to construct a framework for comparative research on the subject. By the end of the century, however, interest in ethnocentrism as a sociological concept had waned, even as the adjective 'ethnocentric' was commonly used as a descriptor or epithet in discussions of intercultural attitudes.

Sumner's brief but dense and multifaceted discussion of ethnocentrism in *Folkways* remains a primary source of theory on this subject. Although Sumner acknowledged previous work by the Austrian sociologist Ludwig Gumplowicz, and Coser (1956) has cited Marx and other earlier theorists for certain aspects of ethnocentrism, Sumner himself may have coined the term ethnocentrism (as well as 'ingroup' and 'outgroup') and was responsible for endowing it with wide-ranging theoretical implications for the social sciences:

> The conception of 'primitive society' which we ought to form is that of small groups scattered over a territory. … A group of groups may have some relation to each other … which draws them together and differentiates them from others.

> Thus a differentiation arises between ourselves, the we-group, or in-group, and everybody else, or the others-group, out-groups. The insiders in a we-group are in a relation of peace, order, law, government and industry, to each other. Their relation to all outsiders, or others-groups, is one of war and plunder, except so far as agreements have modified it … .
> … *Sentiments in the in-group and towards the out-group*. The relation of comradeship and peace in the we-group and that of hostility and war towards others-groups are correlative to each other. The exigencies of war with outsiders are what makes peace inside, lest internal discord should weaken the we-group for war. These exigencies also make government and law in the in-group, in order to prevent quarrels and enforce discipline. Thus war and peace have reacted on each other and developed each other, one within the group, the other in the intergroup relation. … Sentiments are produced to correspond. Loyalty to the group, sacrifice for it, hatred and contempt for outsiders, brotherhood within, warlikeness without—all group together, common products of the same situation. These relations and sentiments constitute a social philosophy. It is sanctified by connection with religion. … Virtue consists in killing, plundering, and enslaving outsiders.
> … *Ethnocentrism* is the technical name for this view of things in which one's own group is the center of everything, and all others are scaled and rated with reference to it. Folkways correspond to it to cover both the inner and the outer relation. Each group nourishes its own pride and vanity, boasts itself superior, exalts its own divinities, and look with contempt on outsiders. Each group thinks its own folkways the only right ones, and if it observes that other groups have other folkways, these excite its scorn. Opprobrious epithets are derived from these differences. 'Pig-eater', 'cow-eater', 'uncircumcised', 'jabberers', are epithets of contempt and abomination. The Tupis called the Portuguese by a derisive epithet descriptive of birds which have feathers around their feet, on account of trousers. … For our present purpose the most important fact is that ethnocentrism leads a people to exaggerate and intensify everything in their own folkways which is peculiar and which differentiates them from others. It therefore strengthens the folkways (Sumner 1906, pp. 12–13).

This passage shows both the scope of Sumner's discussion—covering attitudes and stereotypes as well as social relationships, organizational features and warfare—and his commitment to the theory that ingroup–outgroup dynamics generate internal solidarity and cultural coherence as well as hostility and warfare. Sumner claimed these dynamics and the concomitant attitudes to be universal in human societies, although he was aware that not all peoples were warlike. Social scientists before and after Sumner discovered many of the same phenomena in their own research, but his influential passage became a point of reference for subsequent social theory and research on war and peace, ethnic relationships, and intergroup conflicts and attitudes.

Ethnocentrism as a process simultaneously promoting ingroup solidarity and outgroup hostility has been described at different levels, in many places and historical periods, not only by sociologists and anthropologists but also by political scientists, historians, and journalists. It is as familiar now, particularly in ethnic politics, as it was in Sumner's time. But Sumner's theoretical claims have been criticized and largely rejected by sociologists and anthropologists. Merton and Rossi (1957), in formulating the sociological theory of reference groups, pointed out that individuals often belong to, or wish to belong to, more than one group, making the ingroup–outgroup boundary ambiguous and the outgroup often an object of admiration rather than contempt. They argued that the ethnocentrism syndrome as Sumner formulated it should be seen as a special case of human intergroup relations rather than universal, as he claimed.

Anthropologists have been similarly critical. Describing many parts of the world where ethnic groups do not have bounded territories and individuals and groups change ethnic affiliation, anthropologists cast serious doubt on generalizations presuming that human groups have clear boundaries and stable membership (Leach 1954, Barth 1956, 1969, Goody 1956, Moerman 1965, LeVine and Campbell 1972, pp. 81–113). Furthermore, the threat-dependent ethnic and national identities that Sumner assumed were products of a natural human tendency have increasingly been seen by anthropologists as deliberately constructed by leaders to gain adherents and seize power. Finally, it now seems likely that the intense nationalism and threatening military competition of the Western world in 1906 predisposed Sumner and his readers to accept ethnocentrism as a universal fact of human social life in advance of the comparative evidence. With better evidence and a different experience of international relations, we are less inclined to accept his theory.

Thus the sociology of Sumner's ethnocentrism syndrome, viewed in the light of recent evidence and newer perspectives, seems a dated fallacy of early twentieth century social science. The central problem has been that without clear and stable boundaries between ingroup and outgroup, the question of whether ingroup solidarity and outgroup hostility are consistently interdependent cannot be unequivocally answered. Group boundaries, if they ever were clear, have become more and more blurred as international migration and communication have intensified. Sumner's concept, as he generalized it, is not an accurate guide to the real world, but the question arises, if Sumner's formulation of the ethnocentrism syndrome has no merit as social theory, why does it seem so familiar as ethnographic description? How it is that, despite globalization, the contemporary world produces so many examples of conflict between ethnic groups differing in language, religion, and other customary practices, with mutual disparagement and threat-dependent solidarity escalating to intergroup violence? Answering this question through social research requires the investigation of the ethnocentrism syndrome as a contingent historical phenomenon, examining in detail how the ideological and organizational patterns to which Sumner called at-

tention have emerged in particular cases and how long they have lasted.

The term ethnocentrism remains in use to refer to culturally biased judgment, the act of applying the frame of reference from one's own culture to someone or something belonging to another culture. The *Dictionary of Anthropology* (Barfield 1997, p. 55) defines ethnocentrism as 'the belief that one's own culture is superior to others, which is often accompanied by a tendency to make invidious comparisons. In a weaker form, ethnocentrism is the tendency to look at other cultures through the filter of one's own cultural presuppositions.' As in Sumner's examples in the passage quoted, ethnocentric judgments are usually derisive epithets, in which outgroup practices that deviate from ingroup norms form the basis for an insulting characterization: 'pig-eaters,' 'uncircumcised,' 'jabberers.' This can reflect a naïve form of ethnocentrism, called phenomenal absolutism by Segall et al. (1966), in which a person unreflectively takes his own culture's values as objective reality and automatically uses them as the context in which he judges less familiar objects and events. Alternatively, it can reflect a judgment made with full knowledge of cultural variability, but with an equally or even more scornful attitude. An ethnocentric judgment can be treated as a cognitive phenomenon, but it has emotional component as well. In its cultural form it is usually called an ethnic stereotype. Such stereotypes were intensively studied in American social psychology before and after World War II, up to the mid-1950s (Allport 1954). Since then, there has been relatively little research on ethnic stereotypes, but the concepts of ethnocentrism and stereotypes are still widely used in the mass media, in programs of training in intercultural communication, and in public discussions of racial prejudice and ethnic relations. The teaching of anthropology, through publications as well as in the classroom, is often justified by its role in reducing the ethnocentric attitudes of Western populations.

See also: Cultural Relativism, Anthropology of; Ethnic Conflict, Geography of; Ethnic Conflicts; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Groups, Sociology of; Nationalism: General; Nationalism, Sociology of; Prejudice in Society; Xenophobia

Bibliography

Allport G W 1954 *The Nature of Prejudice*. Addison-Wesley, Cambridge, MA

Barfield T (ed.) 1997 *The Dictionary of Anthropology*. Blackwell, Oxford, UK

Barth F 1956 Ecologic relationships of groups in Swat, North Pakistan. *American Anthropologist* **58**: 1079–89

Barth F (ed.) 1969 *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*. Little, Brown, Boston, MA

Coser L A 1956 *The Functions of Social Conflict*. Free Press, Glencoe, IL

Goody J R 1956 *Social Organization of the LoWiili*. Her Majesty's Stationery Office, London

Leach E R 1954 *Political Systems of Highland Burma*. Bell, London

LeVine R A, Campbell D T 1972 *Ethnocentrism: Theories of Conflict, Ethnic Attitudes and Group Behavior*. Wiley, New York

Merton R K, Rossi A 1957 Contributions to the theory of reference group behavior. In: Merton R K (ed.) *Social Theory and Social Structure*. Free Press, Glencoe, IL

Moerman M 1965 Ethnic identification in a complex civilization: Who are the Lue? *American Anthropologist* **67**: 1215–30

Segall M H, Campbell D T, Herskovits M J 1966 *The Influence of Culture on Visual Perception*. Bobbs-Merrill, Indianapolis, IN

Sumner W G 1906 *Folkways*. Ginn, New York

R. A. LeVine

Ethnogenesis in Anthropology

Ethnogenesis is a process by which a distinct ethnic group emerges with a common sociocultural consciousness of belonging. In the 1960s Singer (1962) analyzed the process among African-Americans and Sturtevant (1971) subsequently introduced this term into the anthropological study of native Americans. Since then it has been widely utilized for examining the consciousness of belonging among various populations.

Although Sturtevant did no more than point out how ethnogenesis could be used to analyze the historical establishment of group distinctiveness, the concept has been regenerated during the 1990s with new implications. It is used to focus on the interaction between global and local histories, especially on local initiatives to deal with the impact of globalization. Ethnogenesis therefore refers to cultural and political struggles by indigenous and enslaved populations to create enduring identities at times of often violently imposed radical change resulting from European domination—such as demographic collapse, forced relocation, or enslavement. The idea of ethnogenesis refutes the established view of societies in the Third World as isolated, static, and ahistorical. Instead it pays attention to their creativity and ways of making history.

1. Theoretical Background

Currently the focus on issues of ethnicity has shifted from study of particular ethnic groups to looking at the dynamic process of ethnogenesis. This was made possible by the continued development of new theoretical understandings of ethnicity. Following Barth's contribution (1969), the subjective and interactive

aspects of ethnicity began to gain attention. Despite the pre-eminence accorded the objective and primordial aspects of ethnicity by earlier perspectives, Barth did not regard ethnic identity as constituting an *a priori* reality, but as being an interactive process of self-definition and classification by others.

Actors categorize themselves and others in relation to how they interact with one another. It is when they use an ethnic identity to categorize themselves that Barth sees an ethnic group emerging. He rejected the prevailing idea that geographical and cultural isolation or separation is essential to the preservation of ethnic groups and the maintenance of cultural diversity. Instead he insisted that ethnic differentiation forms an overall social system; what is to be studied, consequently, is not the internal relations of the individual ethnic group, but the way the different ethnic groups establish boundaries. This perspective led to the view that ethnicity involved a formative process.

Another shift in the 1980s was the emergence of new perspectives that sought to deconstruct ethnicity as an essential entity. Deconstructionism in the social sciences exercised its influence to show how seemingly essential categories such as sex, generation, and race are far from being natural and *a priori*, but are an invention of modern society. Hobsbawm and Ranger showed that what had been taken to be traditional elements in symbols, customs, myths, and identity were actually invented as part of a dynamic process in the course of modern nation-building. According to this view, ethnic identity is also a modern invention that has been shaped in a political and historical way within the frame of the nation-state. This deconstructionist perspective, together with a subjective active view of ethnicity, provided a basis for the social sciences to theorize the phenomenon of ethnogenesis.

2. *Ethnogenesis and Ethnocide*

The concept of ethnogenesis has developed from studies of societies in the New World. Here the indigenous population was conquered and oppressed in the centuries after 1492. In addition, an enormous number of people with different cultural-linguistic backgrounds were transported from Africa and set to work as slaves. This is a process of ethnocide.

Ethnocide means the extermination of a culture. It involves the systematic destruction of cultural identities, ecosystems, political authorities, and religious rituals and belief systems. Common forms of ethnocide included the forced conversion to the religion of the colonizers, the banning of the use of native languages and cultural practices, and the forced relocation from the homeland to reservations.

Although it differs from genocide, which means the physical destruction of a whole people, the distinction is not always clear, and the two phenomena may occur together. Following the arrival of Columbus in the New World in 1492, the Arawak and Carib peoples he encountered, whose population is estimated to have numbered around 10 million, were almost entirely wiped out within 100 years. Virtually no indigenous people now remain in the Caribbean, although some were able to flee to the American continent where they merged with other survivors to generate new cultures.

In the century after the discovery, the indigenous population of South America, estimated to have been around 40 million, declined to about 440,000. In Australia in the nineteenth century the situation was much the same, although there were not as many people to be killed. The people of Tasmania, with an estimated population of 23,000 in 1803, were exterminated before the end of the century. Other people, on the verge of cultural extinction, have attempted to establish distinct cultural identities in order to reorganize their societies and secure their existence. This is a process of ethnogenesis. There is therefore a direct relationship between ethnocide and ethnogenesis. Where ethnocide takes place, one will usually also find ethnogenesis.

In the present day, peoples who were enslaved or marginalized under colonial domination are able to revive their cultural identities and reclaim their right to existence. This has gained momentum with the setting-up of statutory and voluntary organisations in the international context.

3. *Mechanisms of Ethnogenesis*

Minority populations which were marginalized in the world politico-economic system needed a shared consciousness of belonging. This required definitions of group boundaries, and processes of cultural creation, revitalization, and appropriation.

The creation of a religious ideology and practice was one means of integrating different groups into one unit. The enslaved African populations of the Caribbean, for instance, had originally come from a variety of sociocultural backgrounds with different languages, religions, value systems, and social structures. Despite these differences, they invented rituals, oaths, and a common set of religious values. This bricolage of different African elements became marked with the emergence of a Maroon identity in Jamaica in the middle of the seventeenth century.

In 1655, Britain took Jamaica from Spain by force. African slaves took advantage of the resulting turmoil to escape from the plantations into the forests and mountains. These deserters comprised a medley of ethnic origins, from Senegambia and the Gold Coast to the Congo and Angola, and even to the east coast of Africa. They were able to establish their own political authority and military alliances in the seventeenth and eighteenth centuries. Such polities created and mobilized common religious ideologies and practices

based on rituals of spirit possession, which produced a shared consciousness and generated a common identity.

Ethnic identity may also be developed through a revitalization of precolonial symbols and ritual practices, as was the case with the Seminoles of Northern Florida. In the eighteenth century, when British, French, and Spanish forces were competing in the region, a distinct Seminole ethnicity emerged from the variety of Muskogean and other peoples with different cultural-linguistic backgrounds, who had colonized this region. They established new chiefdoms which closely followed the traditional Muskogee patterns. Among Muskogean peoples, who had previously been only loosely related with each other, there was the powerful division of Kawita, which overpowered the others and standardized and transformed a variety of politico-religious cultures under its hegemony. The early Seminole settlements still acknowledged Kawita's suzerainty. This meant giving a central role to rituals of the mythical creation of sacred fire, originally controlled by Kawita's traditional chiefs whose authority was thereby revitalized in the newly created Seminole societies.

Ethnogenesis may also take place through appropriation. From the late fifteenth century, indigenous peoples in the Third World were in most cases conquered by overwhelmingly superior European powers and forced to accept the religion, language, and way of living of the colonizers. In doing so, however, they often created a new ethnic identity by the way they transformed and adapted the ritual symbols of the Europeans. For instance, the Kapon and Makushi of Venezuela and Guyana emerged when the Christian saints and deities brought by the European colonizers were reinterpreted according to the traditional cosmology of indigenous peoples who used these ritual symbols to forge new cultural identities.

4. *Ethnogenesis in the Twentieth Century*

Although the process of ethnogenesis can be clearly seen in small-scale societies in the Third World following colonization, it has also developed in the process of modern nation-building. The ideology of the nation-state expressly aims to transform different ethnies into one nation with a single national culture. The principle is often different from the practice. The ruling group, in most cases the majority group but sometimes a numerical minority having dominant power, has generally monopolized the political-economic interests and imposed its culture on the others in the name of the national culture. In response, marginalized peoples have attempted to generate a new ethnic identity in order to claim their sociocultural and politico-economic rights within the frame of the nation-state. Modern ethnogenesis may be classified in three main ways.

The first type is integration. Those peoples who originally had formed separate and independent polities without any consciousness of belonging together were given the same name and treated as one group by European explorers, missionaries, anthropologists, and administrators. They accepted this name and formed a new identity in accordance with the imposed boundary, in order to resist or negotiate with the colonial regime, especially in modern nationalist movements that sought political independence. In the process, a new ethnicity emerged. This type of ethnogenesis can be seen among the Aborigines in Australia, the Inuit in the Arctic regions and the San in the Kalahari, where the central government has adopted policies of multiculturalism to secure minority group rights—the government needs fixed sociocultural units in order to assist political autonomy and cultural preservation.

The second type is amalgamation, which occurred frequently in African societies at the time of independence. In Kenya after World War II, for example, an active African nationalism was organized by two major ethnic groups, the Kikuyu and the Luo. Political elites from smaller ethnic groups united under the hegemony of these two groups and gradually formed new cultural identities. Thus the Luyia emerged in the 1940s to bring together 17 Bantu-speaking peoples, among them the Maragoli, Bukusu, and the Wanga, who lived in what was then the North Kavirondo district of western Kenya. The word Luyia, meaning the communal space in which people sit around the fire, was appropriated for the the idea of mobilizing a consciousness of brotherhood among those from western Kenya who studied at Makerere University in Kampala.

The same sort of amalgamation took place among the 10 or so Eastern Cushitic-speaking ethnic groups in the Rift Valley. This ethnogenesis was promoted by those who studied at the prestigious Alliance high school in Nairobi in the 1940s. They invented the name of a new ethnic group, the Kalenjin—literal meaning 'I'm telling you' (in the plural sense)—and it started to spread among influential members of the home communities in the 1950s. Aidan Southall has termed this type of ethnogenesis 'supertribalization.'

The third type is separation. During the period of colonial domination, some small indigenous ethnic groups were subsumed into bigger neighboring groups or were put together in one large unit, either voluntarily or forcibly. Later they have attempted to revive their former cultural identity and secede from the constituted group, to create an independent ethnic group for themselves. The Suba, mainly living in the small islands of Lake Victoria, had been Luo-ized for a century, and their language was beginning to become extinct. But after the multiparty system was reintroduced into the Kenyan political scene in 1991, which caused the nation-state to become more ethnicized, the Suba political elite began to dif-

ferentiate Suba from Luo and promoted the Suba culture and language. In so doing they are again forming themselves into a distinct ethnic group.

In the contemporary world system indigenous peoples and ethnic minorities who were once colonized and marginalized have begun to assert their socio-cultural and politico-economic rights. Because re-fashioned ethnic identities play a large part in such assertions, ethnogenesis is becoming a key concept in understanding society in the twenty-first century.

Bibliography

Barth F (ed.) 1969 *Ethnic Groups and Boundaries*. Universitetsforlaget, Oslo, Norway

Henry M 1962 *Studies in Siberian Ethnogenesis*. University of Toronto Press, Canada

Hill D (ed.) 1996 *History, Power, and Identity: Ethnogenesis in the Americas*. University of Iowa Press

Hobsbawm E, Ranger T O (eds.) *The Invention of Tradition*. Cambridge University Press

Matsuda M 1998 *Urbanisation from Below*. Kyoto University Press, Kyoto, Japan

Roosens E 1989 *Creating Ethnicity: The Process of Ethnogenesis*. Sage

Singer L 1962 Ethnogenesis and Negro-Americans today. *Social Research* **29**: 419–32

Smith A 1986 *Ethic Origins of Nations*. Blackwell, Oxford, UK

Sollors W 1989 *The Invention of Ethnicity*. Oxford University Press

Southall A 1970 The illusion of tribe. In: Gutkind P C W (ed.) *The Passing of Tribal Man in Africa*. Brill, Leiden

Sturtevant W 1971 Creek into Seminole. In: Leacock E, Lurie N (eds.) *North American Indians in Historical Perspective*

M. Matsuda

Ethnography

1. Introduction

'Ethnography' has turned into a complicated concept. To understand why, it is useful to go back to a time when it was thought to be simpler. Traditionally, the term was used to label what anthropologists did when they did their research.

By 'anthropologist' is meant here those anthropologists—sometimes called 'ethnologists'—who work with living peoples to explore and document their culture. Their research involved long-term involvement in a community called 'participant observation,' discussions of topics of interest with community members called 'informal interviews,' and a record of the experience called 'field notes.' Ethnography also labeled the book-length report that an anthropologist wrote, using intuitive methods of analysis called 'immersion in the data.' The ethnography, the book, was also called a description of the *culture* of the community, although the culture focus was more characteristic of work in the USA than in Europe.

This straightforward definition of ethnography must be understood in its historical context. Anthropological research, as practiced in the early to middle twentieth century, usually involved an Anglo-American or European working in a small isolated community in a remote area of the world, with a group of people who often lived without modern amenities, had little or no formal education, and existed at a modest if not impoverished economic level. The anthropologist took up residence in the community, and set out to learn who the people in the community were and how they lived. Depending on their interests, anthropologists investigated topics as diverse as economic and household tasks, social organization and religious beliefs, family life and local markets—virtually anything that made up daily life. Since they also lived in the community, anthropologists visited with people, attended communal events, and worked together with community members on common projects, including the ethnography itself.

As one learned the community's way of living, one said one was 'doing ethnography.' Ethnography, then, named the process of learning what for the anthropologist was a new and different way of talking, thinking, and acting.

When the work was done, a monograph was usually written—a report of this exploration into the community's way of life. One could point to that book and say it was an 'ethnography' as well. Further, one could also say that the book was a description of the group's culture.

Traditionally, then, ethnography named both a research process characteristic of anthropology and a research product, often a book-length description, but at any rate a representation of the culture of the community in which the research had been done. (Agar 1996 introduces this traditional view and contextualizes it in light of more recent discussions.)

Because anthropologists set out to learn ways of thinking, talking, and acting whose nature they could not specify before they started their research, ethnography did not look much like traditional science. In fact, the scientific status of ethnography has been an issue since its beginning.

Traditional science sought universal laws; ethnography went after local particulars. Traditional science emphasized control of the research process; ethnography featured adaptability as the nature of community emerged. Traditional science evaluated a sample by size and strategy; ethnography, by competence, social position, and relationship to subject. Traditional science preserved initial concepts through the course of the research; ethnography developed new ones. Traditional science rested on linear models; ethnography, on models more systemic and processual. Tra-

ditional science represented data with numbers; ethnography more often used words.

Over the last twenty years it has become clear that ethnography cannot be adequately modeled or evaluated by traditional science. Some still argue that it should be. Some argue that it is a different kind of science. Some argue that it should be classed with the humanities. The issue is treated in Sect. 2 of this article.

A second problem arose with the nature of ethnography as a product, a description of a culture that resulted from the research. Anthropologists assumed that small communities represented a single culture, one that was internally consistent and traditional, passed from generation to generation in essentially unaltered form. The 'culture' concept served as description, explanation, and generalization. The fact that a person did something was data for the description of that culture. What was done could be explained by membership in that culture, and it could be generalized to the single culture assumed to be relevant.

The culture concept that guided earlier ethnographies is now in a shambles. The earlier assumptions about culture, such as its traditional and consistent nature, have been shown to have been in error from the beginning (Wolf 1982). Further, contemporary life makes the assumption that a person is a member of only one culture look hopelessly naive. While the nature of ethnography as a research process endures, it is no longer clear how to describe, explain, or generalize what is learned, nor is it clear what form a report—the ethnographic product—should take. The old concept of culture is no longer adequate to the task. Sect. 3 of this article deals with the problem.

Until this point ethnography has been discussed in the context of anthropology. Because of its longstanding concern with the exploration of previously unknown ways of living, anthropology is the academic discipline with the most elaborate history of ethnographic research. Ethnographic work outside anthropology also exists. For example, predisciplinary efforts such as Tacitus' description of the Germans in Roman times are often cited as a precursor. And ethnographic work in the discipline of sociology also has a long tradition, beginning with work based on Weber's theory of social action, continuing with the symbolic interactionism of the Chicago school and with the development of grounded theory and ethnomethodology in the 1960s.

By now the historical link between anthropology and ethnography has turned complicated in at least two ways. First of all, ethnography is used in a variety of different disciplines—ranging from history to speech communication to organization and management to public health—as a *new* approach, as well as in kindred social sciences such as political science and psychology. Second, the concept now coexists with such related terms as qualitative research, grounded theory, and phenomenology, terms that are also used widely in a number of different research contexts. These multiple uses of 'ethnography' and related terms have further blurred the edges of how ethnography should be defined, a problem dealt with in Sect. 4 of this entry.

In traditional cultural anthropology, then, ethnography had a clear meaning—the process of learning a community's way of life and reporting the results of that learning under the label 'culture.' Beginning in the late twentieth century, many of the assumptions underlying traditional ethnography were called into question—questions amplified by the changing nature of the world we all live in now. Finally, the increasing popularity of ethnography and related nontraditional approaches to research in numerous academic and organizational settings further complicates any attempt to define the term.

This article, then, is built less on a definition than it is on a sample of key issues around the question of what a definition might look like. 'Ethnography' is now a cover term for several debates around the nature of human understanding; debates for which, at the time of writing, no end is in sight. In fact, this might be the major contribution of the term to current social research, a contribution with more historical importance than its use as a label for a discipline-internal activity. If nothing else, this entry alerts the readers new to ethnography and culture that they are stepping into a series of issues rather than a recipe for doing research. Many new to ethnography import outdated notions of ethnography and culture from traditional anthropology into their work, a move that replicates historical mistakes and distorts their results.

2. *Science or Humanity?*

Traditional anthropology has long lived with the contradiction of being *both* a science *and* a humanity. In fact, ethnography makes this contradiction particularly visible. On the methodological level, ethnography seeks new patterns in 'found' data, displaying its kinship with philosophy, history, journalism, and second-language learning, fields typically found in the 'humanities' branch of a university. On the level of theory, ethnography usually deals with issues related to sociology, psychology, economics, and political science, disciplines categorized as 'social sciences.'

The idea of a 'science of society,' as first articulated by Spencer and Comte, linked 'social science' explicitly with the Western Enlightenment tradition. In the classic formulation of this approach, one deduces hypotheses that should be true if a particular theory is true, where a hypothesis is a statement of a relationship among two or more variables. One then operationalizes the variables by specifying what observations one might make or what questions one might ask in order to measure them—i.e., assign a numerical value where

differences in the number represent relevant differences in the phenomena of interest. One selects a sample of a population on some principled basis, the point being to say at the end how confidently one can generalize from the sample to that population. Finally, one specifies a decision procedure so that one can say at the end whether the hypothesis is accepted or rejected. Often a decision procedure involves a statistical test with a cutoff point of statistical significance, significance technically meaning, 'there is no more than a 5 percent (or 1 percent, or some other figure) chance that I'm wrong when I say that the variables in the hypothesis are related in ways beyond what I'd expect by chance.'

Numerous studies in anthropology have revealed the power and value of traditional science in the context of ethnography. (For a good example of methodology from this angle of vision, see Bernard 1988.) Further, other social sciences recognize the importance of testing hypotheses 'cross-culturally,' linking those fields to anthropology. And finally, in a metaphorical sense, much of anthropology serves the purpose of challenging assumptions by comparison among different societies, a 'hypothesis test' of a rather fundamental sort. In fact, one can say that social research done only in one society that generalizes to all of humanity is bad science.

However, much of ethnography does *not* fit the tradition of social science. Consider two example problems. The first—scientific research is objective and replicable. By this is meant that well-designed research, done in the same way in any place by any trained person, will yield the same results. In contrast, ethnographers argue that, for some group X and some ethnographer Y, both the ethnographic process and the ethnographic product will vary as a function of Y as well as a function of X. In other words, an ethnography can take different shapes depending on the biographical and cultural background of the ethnographer. The ethnographer is, in a nutshell, part of the data. It is now a cutting-edge methodological issue—more than one ethnography of X is possible, but not all ethnographies of X are acceptable. The problem is to learn how to specify the difference between the two.

A second problem—the scientific emphasis on control. In traditional science, control is key, from theory through data collection and analysis to decision criteria on whether to accept or reject a hypothesis. When the variable of interest, the dependent variable, changes in value, it must be due to the independent variables and nothing else. Ethnography has a defined starting point, but as it progresses a study is emergent. New concepts are learned that suggest new questions; new knowledge of communication styles suggests new ways of asking them; something someone does opens new avenues of inquiry; new knowledge of social types within the group alters sampling designs; analysis consists of putting new pieces together into patterns rather than breaking them down into measurable variables. The research process is nonlinear, dialectical rather than sequential. An emphasis on control from the outset and throughout the research would eliminate these possibilities.

Many anthropologists now reject the scientific tradition and consider ethnography a humanity rather than a social science (see Geertz 1973 for a classic statement of this position). But the humanities offer no straightforward model for ethnography, either. To offer just a few examples, a look at journalism reveals genre constraints whose *a priori* structural requirements may be as limiting to emergent research as is traditional science. Models of second language learning deal awkwardly, if at all, with the massive amounts of background knowledge implicated in syntax and lexicon, and it is just this knowledge that an ethnographer hopes to learn and represent. Narrative approaches from both history and literature are promising, since much ethnographic data is narrative in form. However, narrative is built on chronological sequence, and ethnography abstracts and generalizes from such sequences.

Ethnography, then, is caught between agreement with general scientific goals and a poor fit with traditional social scientific practice, and it is caught on a research process akin to the humanities that calls for systematic approaches to modeling and validation that those fields do not provide.

What is needed is an epistemology that fits ethnography, a clear sense of how knowledge is obtained and what it looks like in the end. Epistemologies borrowed from either traditional science or traditional humanities are suggestive and helpful, but at the same time inadequate. Three conclusions are clear. First, a proper epistemology will not come from a simple classification—i.e., ethnography is a kind of social science or ethnography is a kind of humanity. Second, the proper epistemology is, as of now, far from established. And third, the development of such an epistemology will be a contentious issue well into the future.

3. *Culture*

The topic of culture is dealt with elsewhere in this encyclopedia. However, the concept is relevant here because 'culture' traditionally has been used to characterize the results of ethnographic research.

In the words of Roy D'Andrade, studying culture today is like studying snow in the middle of an avalanche (cited in Carrithers 1992, an introduction to various views of anthropological research). Traditionally, anthropologists assumed that the small communities they studied constituted a single culture. What people said or did could be described, explained, and generalized by attribution to that single culture, bounded in space and replicated over time.

Any contemporary ethnography now shows that such assumptions no longer hold. Indeed, numerous 'cultures'—global to national to regional to local—might or might not be relevant to the description, explanation, and generalization of a particular ethnographic moment. There are cultures of generation and gender, of nationality and religion, of ethnicity and region, of occupation and leisure-time pursuit, media inspired and driven by tradition, and in the process of creation. Some cultures span the globe, some distribute unevenly across it, and some take shape only in a particular location. Some people enact a particular culture; some resist it; some ignore it. Some 'craft a self' that is a blend of different cultures (the phrase is from Kondo 1990, an excellent example of narrative ethnography). Popular phrases now tag almost anything with 'culture': the 'culture of violence', or the 'culture of capitalism', or the 'culture of poverty.'

Culture can no longer be used as it was in traditional anthropology. Indeed, its use in traditional anthropology was inaccurate even then. The work of Eric Wolf (1982) showed that traditional assumptions were incorrect, even in the nineteenth century when anthropology first appeared as an institutional discipline. Differences among people still lend themselves to ethnographic investigation. But the multiple and simultaneous relevance of different 'cultures,' and the multiple positions of actors vis-à-vis those cultures at different points in time, have turned description, explanation, and generalization into a most difficult problem. Certainly it is now the rare case when any overview of group life can be modeled under a single 'culture' label (Hannerz 1992, Kuper 1999).

The culture problem took another twist with a book entitled *Writing Culture: The Politics and Poetics of Ethnography* (Clifford and Marcus 1986). *Writing Culture* asked why ethnographies were written the way that they were. What genre constraints did students unconsciously absorb as they read the works of their predecessors? And how did these constraints set limits on the images of the 'other,' the world in which they lived, and the role of the ethnographer who did the work?

The book inspired a series of debates that continue to this day. A few key issues can be mentioned here. First, what were the 'politics' of the ethnographer/community relationship as it developed over time? Any ethnography builds on collaborative relationships, more so than traditional images of social science allow. Why doesn't the writing reflect that collaborative process, with control of the final product distributed across all the participants?

Second, any community lives in a world of layered contexts, from local realities up to the global political economy. Traditional ethnographies tend to feature the local situation as if it hangs isolated in space and time. Clearly it doesn't. Regional, national, and global histories help explain how the local community came about, as well as the systems of power that hold it in place. Why don't traditional ethnographies include such material? And what is the proper way to obtain and analyze it, since some crucial data lie outside the awareness of community members who are interviewed and observed?

Third, who is this ethnographer? He or she has a biography, social identities, a history of training, a set of ideological biases, and a vested interest in doing the work and satisfying whoever pays for it. Traditional writing minimized the presence of an ethnographer, if it recognized it at all. In part, the writing reflected notions of traditional science, as discussed earlier. In an objective and replicable study, the identity of the researcher has no influence on the outcome. Now the question is, how and how much of the ethnographer to include in the writing of the results?

These and other issues around the 'politics' of the research implicate the 'poetics' of the writing, as suggested by the subtitle of *Writing Culture*. Traditional ethnographic writing does not handle issues of politics and broader national and global contexts and ethnographer presence. What kind of writing does? Narrative? Legal briefs? Postmodern collage? Or, never mind writing, how might an ethnographic product be represented in a film or in a museum exhibit? Here, as with previous topics discussed in this entry, the questions around 'writing culture' have opened up a variety of new possibilities that simply did not exist in traditional anthropology.

The relevance of multiple 'cultures' and the call to include previously neglected 'political' issues have converted the old concept of culture into a list of issues that, for the near future, will remain under debate.

4. *The Growing Ethnographic Family*

Until recently, ethnography remained the specialty of anthropologists and sociologists. Now interest in ethnographic approaches to research has increased in a number of fields. Anthropologists often react with dismay to these developments, arguing that critical aspects of ethnography and the culture concept get lost in the transfer. More often than not they are right to do so. At the same time, outside anthropology various models of research are being developed that have a 'family resemblance' to ethnography, to use Wittgenstein's terminology. The existence of this new 'family,' together with the indeterminate nature of ethnography within anthropology already discussed in this entry, make for an even more complicated situation.

Consider the term 'qualitative,' the most widely used term to label this family of research. Strictly speaking, 'qualitative' labels only a type of data, propositional rather than numeric in form. Yet 'qualitative' may or may not also imply differences in

methods, analysis, or epistemology. At one extreme, a traditional social scientist might add open-ended questions to a survey, code them with a few concepts derived from prior theory, and announce that 'qualitative research' had been done.

An ethnographer, who would also be called a 'qualitative' researcher, would criticize these data on several grounds. The open-ended questions are too narrow and the researcher invented them. They represent only interview data without any participant observation. No other contextual data are available to constrain interpretations. Codes are taken from prior theory rather than being developed dialectically with the texts themselves. For an ethnographer, these 'qualitative' data are simply alternative measures of prior variables characteristic of traditional science, measures that have little or nothing to do with ethnography.

'Quantitative,' on the other hand, supposedly labels traditional scientific research, but the same ambiguities operate with this term as well. An ethnographer, for example, might use census data as part of their investigation of a community. They might also use prior survey results to link local detail with larger populations. Or they might use a 'natural experiment' approach to organize ethnographic data to test covariation among elements of a discovered pattern. An ethnographer uses quantitative data following a different epistemology when compared to traditional social science, as previously discussed. But 'qualitative' versus 'quantitative' is not a distinction adequate to capture such differences in research.

A second strategy to clarify the nature of the new 'family' is to divide it into types. Introductory textbooks on qualitative research lay out subdivisions such as ethnography, grounded theory, and phenomenology. Such distinctions confuse historical sources of the family with the family itself. Phenomenology, and later developments such as hermeneutics, helps clarify the philosophical foundations of the family. Grounded theory isolates and makes explicit a part of the research process—the development of summary categories from uncontrolled data. Ethnography, as shown earlier, takes from anthropological traditions the notion that long-term intense involvement—participant observation—is a precondition of exploring a community. But these three historical sources are contributions to this new family of research, not separate ways of doing it. Dividing the family into its historical sources does not clarify what it is and what it is becoming, and 'qualitative' won't serve as an adequate label for the new family of research, either.

For the author of this article, the new family is, in fact, an adoption of basic ethnographic epistemology. However, the scope and depth of data and the method of analysis vary widely across different disciplines and problem orientations. Many, probably most, approaches, do *not* feature participant observation as a *sine qua non* of the research.

For ethnographers, this new family poses a problem. Ethnography, as it developed in anthropology, is clearly a historical source and a current member of the family. Most family members are recognizable as instances of ethnographic logic in action. At the same time, for anthropologists, the ethnographic tradition defines the prototype for the family, and departures from this tradition count as a decline in quality. Narrow focus and lack of participant observation are two frequent criticisms of such departures. For many ethnographers, then, the growth of the new family is viewed negatively, as the appropriation of a logic without the disciplinary tradition that developed it. This negative view is complicated because, within contemporary anthropology, traditional ethnography is now viewed as inadequate on several grounds as well (Gupta and Ferguson 1997). The paradox generates much debate among anthropologists.

5. *Conclusion*

Fifty years ago this article would have been much simpler but much less interesting to write. 'Ethnography' labeled both the way that anthropologists did their research and the report that an anthropologist wrote, a report also called a description of the culture of a community. That definition, by the late twentieth century, had mutated into numerous unresolved questions about the nature of one human setting out to understand some others in our diverse, rapidly changing world. These same questions are now of interest in a variety of fields—other social sciences and humanities, artificial intelligence, biology, business/management, and public health, to name a few examples. Ethnography, as practiced in anthropology, has the most elaborate history, the greatest amount of experience with such questions, and its insistence on participant observation remains unique. But what it will become, separately and as a part of the new research family described in the preceding section, remains to be seen.

See also: Case Study: Logic; Case Study: Methods and Analysis; Clinical Assessment: Interview Methods; Ethnology; Field Observational Research in Anthropology and Sociology; Fieldwork in Social and Cultural Anthropology; Health Research, Qualitative; Participant Observation; Qualitative Methods, History of; Qualitative Methods in Geography; Urban Anthropology; Urban Ethnography

Bibliography

Agar M H 1996 *The Professional Stranger: An Informal Introduction to Ethnography*, 2nd edn. Academic Press, San Diego, CA

Bernard H R 1988 *Research Methods in Cultural Anthropology*. Sage, Newbury Park, CA

Carrithers M 1992 *Why Humans Have Cultures: Explaining Anthropology and Social Diversity*. Oxford University Press, Oxford, UK
Clifford J, Marcus G E (eds.) 1986 *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA
Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York
Gupta A, Ferguson J (eds.) 1997 *Anthropological Locations: Boundaries and Grounds of a Field Science*. University of California Press, Berkeley, CA
Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York
Kondo D K 1990 *Crafting Selves: Power, Gender and Discourses of Identity in a Japanese Workplace*. University of Chicago Press, Chicago
Kuper A 1999 *Culture: The Anthropologist's Account*. Harvard University Press, Cambridge, MA
Wolf E R 1982 *Europe and the Peoples Without History*. University of California Press, Berkeley, CA

M. H. Agar

Ethnology

Ethnology started in the eighteenth century as a systematic attempt to acquire and compare information on those non European populations who did not possess written records of their history and cultural heritage. In the historiography of the social sciences, ethnology represents an early stage in the development of the anthropological disciplines. The meaning of the term, however, is not limited to this historical designation. In contemporary anthropology, ethnology can also refer to a methodology of cross-culturally comparing data on multiple populations. Ethnology in this sense is both distinct from and dependent on ethnography as the primary data gathering by fieldwork conducted in individual societies. In recent decades, the term ethnology has acquired yet another meaning. In a number of European countries, ethnology is prefixed with European and replaces earlier approaches to the study of folklore and the investigation of national culture history.

1. *Ethnology as the Formative Period of Anthropology*

Academic circles at the German universities of Goettingen and Halle first used the term 'ethnology' in the late eighteenth century. Coined as a derivation of the Greek word 'ethnos,' meaning 'a people,' the term 'ethnology' in its most general meaning indicates a scholarly interest in how aggregations of human beings are distinct from each other in terms of material culture, language, religion, moral ideas, or social institutions. Early developments in ethnology also included speculative theories on assumed interrelations between cultural and biologic group differences. In the nineteenth century, ethnology emerged as an exclusively historical line of inquiry, initially supported and brought to public attention by the founding of learned societies in a number of European capitals. This in turn prepared the ground for the inclusion of ethnology as an academic discipline in the curricula of universities first in England, then in Germany and the United States of America (see Schmidt 1924). Ethnology was first considered to be a subfield of anthropology. In the late nineteenth and well into the twentieth century, the umbrella term of anthropology was applied to a much broader domain of scholarly interest than today and included physical anthropology, linguistics, and archeology along with ethnology. Because what has become known as cultural anthropology in the United States of America, and as social anthropology in Great Britain and the Commonwealth countries, grew out of earlier ethnological concerns, ethnology is also to some extent regarded as the predecessor of Anglo–American anthropology.

1.1 *Evolutionism, Diffusionism, and the Historical Method*

In 1888, German-born natural scientist Franz Boas, who during the following decades would initiate and establish cultural anthropology as an academic discipline in the United States, lectured on 'The Aims of Ethnology' before an audience of a German immigrants' civic society in New York City. He identified himself as an ethnologist and defined ethnology's agenda as follows:

> 'The task of ethnology is the study of the total range of phenomena of social life. Language, customs, migrations, bodily characteristics are subjects of our studies. Thus its very first and most immediate object is the study of the history of mankind; not that of civilized nations alone, but that of the whole of mankind, from its earliest traces found in the deposits of the ice age, up to modern times.' (Boas 1940, p. 628)

Coming to ethnology in the late nineteenth century, Boas found himself at the cusp of an important paradigm shift that changed the face of this still young discipline. Up to this point, ethnology had been driven by the impetus to discover the hidden laws of human cultural development. Scholars hoped to do this by aggregating data on as many cultures as possible—both past and present—and by tabulating and classifying them. Findings were supplied either by archeology, itself an emerging discipline, or came from travel accounts, mission reports, and increasingly

also from more sophisticated ethnographic surveys of indigenous groups in Africa, Asia, and the Americas. Most of the nineteenth-century ethnologists attributed the manifestly observable cultural diversity of mankind to what they considered to be different levels of cultural development: the more 'primitive' groups were classified as being still far down on the lowest rungs of the steep ladder of ascent to civilization. At the same time, to the gaze of their more progressed contemporaries these 'savages' personified the earliest evolutionary stages of humanity. Therefore, this approach personified among others, by Edward B. Tylor in England (1832–1917), Lewis Henry Morgan (1818–81) in America, and to some extent by Adolf Bastian (1826–1905) in Germany, was labeled—not by its protagonists, but by later opponents and historiographers—'evolutionism.'

As increasingly more exact and detailed information on non European cultures was made available, many evolutionist generalizations became untenable. At the turn of the twentieth century and already anticipated by Boas in the lecture cited above, a new theory took over that shifted the emphasis away from determining the universal rules of cultural development. Instead, evidence of migration and of contact between groups, and the resulting exchanges and transfers of cultural elements were now taken to be the key to the understanding of human history and cultural diversity. The underlying assumption was that human culture emerged from a single or a small number of origins and the ensuing dispersal of cultural elements came about by migration and diffusion—hence the label of 'diffusionism' that was attributed to this new school. It had its strongest and methodologically innovative proponents among museum ethnologists in Germany (Graebner 1911) who used similarities in the shape and make of pottery vessels, weaponry, ornaments, or other artifacts as a basis to map geographical areas where cultures were assumed to have a shared origin. This enabled German-speaking ethnologists, among them Leo Frobenius (1873–1939) and Pater Wilhelm Schmidt (1868–1954), at the beginning of the twentieth century to lay out a worldwide system of so-called 'Kulturkreise,' areas of shared cultural traits. With their evolutionist predecessors, representatives of diffusionism shared a tendency for sweeping generalizations and a disregard for the cultural meanings associated with the artifacts under scrutiny. They were not interested in the social integration of groups or in the cultural context of customs they used as the basis for their theoretical speculations, rather they treated them as isolates, much as if they were specimens of material culture in a collection's glass case. Also, while insisting on their endeavour being historical in nature, they established a frozen chronology of past cultures rather than dating events and gaining insights into historical processes.

Franz Boas (1858–1942) who had been trained by Bastian in Germany and was appointed a professor at Columbia University in New York City in 1899 rejected not only evolutionism—which had an American stronghold in the Smithsonian Institution's Bureau of American Ethnology in Washington DC—but German diffusionism as well. He founded a new approach, 'the historical method,' that could perhaps best be described as a moderate type of diffusionist ethnology. He and his students, however, shied away from projective theory and, indeed, from generalizations on a grand scale. They built up an enormous database from fieldwork among contemporary North American indigenous populations; their concept of culture areas paid careful attention to proven historical relationships between native American tribal groups and also integrated environmental factors. With this emphasis on 'critical empiricism' (Kuper 1988, p. 143), the Boasians, as his students came to be called who later would become university professors in their own right at the most prestigious American universities, ended the era of speculative ethnology. In Great Britain, evolutionism proved to be more tenuous and was not superseded by diffusionist ideas until the 1920s (see Kuper 1983). However, both in the United States and in England, the term 'ethnology' was gradually eclipsed by the newer designations of cultural anthropology—in America—and its British counterpart of social anthropology, with a focus on culture and on social structure respectively.

2. *Ethnology as Cross-cultural Comparison*

Cross-cultural comparison had already been an indispensable tool for the historical reconstruction of cultural evolution as well as of cultural diffusion, but it was not until the advent of the functionalist school that cross-cultural comparison emerged as the principal methodology of social anthropology. 'By comparing a sufficient number of diverse types we discover uniformities that are more general, and thus may reach to the discovery of principles or laws that are universal in human society,' British social anthropologist Alfred R. Radcliffe-Brown put it succinctly in 1921 (Radcliffe-Brown 1951). Relinquishing the generalizing thrust of earlier social anthropology, later approaches used comparison to inquire into the interplay of variable and invariable factors shaping social relations in a given culture by contrasting it with others like it, often in a controlled sample of regionally equivalent cultural groupings. Meanwhile, in the United States of America, the statistical analysis of nonmaterial cultural data had been elevated to an advanced 'scientific' level by the development of the Human Relations Area Files at Yale University, a project that since 1937 had abstracted and systematized information on cultures around the world from ethnographic reports, and successively created a huge cross-cultural survey (see Moore 1961).

2.1 *Ethnography, Ethnology, Anthropology*

By the end of World War II, the meaning of 'ethnology' had contracted to define but one of the three scholarly practices that form the anthropological enterprise. Ethnography was understood as data-gathering in a single society, usually in a spatially and temporally bounded situation, such as when the fieldworker spends 1 year as a participant observer in a local community and later writes up the findings in a stand-alone, case study-type monograph (also called ethnography). Ethnology, by contrast, utilizes the information collected in a number of ethnographies in order to engage in systematic cross-cultural comparisons between two or more societies. In a report on the status of anthropology in higher education submitted to UNESCO in 1954, French anthropologist Claude Lévi-Strauss defined ethnography, ethnology, and anthropology as three consecutive stages in the production of anthropological knowledge, each building up on the results of the other. While ethnology, to his mind, achieves synthesis by the systematic comparison of ethnographic data across cultures, only anthropology as the third and final stage aspires to theory building with a generalizing impetus, attempting to come to universal statements on the nature of humanity (see Levi-Strauss 1975).

2.2 *The Crisis of Comparativism in Anthropology*

In cultural anthropology in 1960s America, new developments predicated a significant shift away from comparative work in anthropology. Most notably, Clifford Geertz's championing of 'thick description' suggested 'not to generalize across cases but to generalize within them' (Geertz 1973, p. 26) by in-depth immersion in one culture and ethnographic representation of one individual case-study. Criticism of cross-cultural comparativism became increasingly vocal. Comparative methodologies were attacked for their tendency to cut up cultural systems into trans-culturally comparable bits and pieces, with contextual meanings and culturally specific relations getting lost in this process. By treating the cultures under comparison as discrete and separate, cross-cultural comparison in the past tended to deny or downplay influences and linkages between cultures. Also, cross-cultural comparison has been criticized for fixing cultures in time and thus robbing them of their historicity (see Fabian 1983). Another important critique concerns the anthropologist conducting the comparison of at least two, usually multiple non Western cultures and assumes the position of objective observer who is purportedly above or beyond culture. Critics argue that cross-cultural comparison as an operation tends to make invisible power and privilege shaping the hierarchical relations between observer and observed, or between comparer and compared.

In recent decades, many anthropologists have avoided the pitfalls of comparativism by doing away with explicitly cross-cultural perspectives alltogether. At the same time, ethnography stopped being conceived of as merely providing 'raw materials' for comparison and theory building, but as a mode of producing anthropological knowledge became unexpectedly central for the disciplinary identity and epistemology of anthropology. However, anthropologists never completely abandoned comparativism. In recent years, there is a definite move to reassess comparative methodologies in anthropology. New approaches are introduced to strengthen what is considered the inherently critical potential of comparison (see Marcus and Fischer 1986).

3. *European Ethnology*

The disciplinary landscape of European countries is far from uniform. Especially in the social sciences and the humanities, the names of academic disciplines, their boundaries, and their designated domains of scholarly inquiry differ from one country to the next, sometimes even between individual universities and research institutions. Historical developments account for this high degree of variation. A closer look, however, reveals some remarkably consistent patterns. A European phenomenon that in the past has been most strongly articulated in the German-speaking countries is the so-called 'Volkskunde/Völkerkunde split' in anthropology (Gerholm and Hannerz 1982, p. 22) that emerged in the nineteenth century, resulting in two separate anthropological disciplines coexisting in academic teaching and research. Traditionally, those disciplines grouped under the *Volkskunde* heading were concerned with the historical investigation and the cultural construction of a national heritage. Conversely, those disciplines belonging to the *Völkerkunde* category focused on nonEuropean 'primitive' peoples. While the latter scholarly line of inquiry became dominant in those societies marked by colonial expansionism, the former interest led to the establishment of folklore and folklife studies dedicated to the philology and history of national and regional cultures in most of Eastern, Central and Northern Europe.

Since the late 1960s, and intensifying with the transformations in Eastern Europe starting in 1989, both sections of the anthropological enterprise have undergone a marked modernization, in some countries leading to a rapprochement of the disciplines. In Nordic countries, ethnology, formerly a museum science concerned with the material culture of non-European small-scale societies, was transformed into an internationally acclaimed social anthropology, while the label ethnology was appropriated by a historically informed and sociologically acute 'anthropology at home' that has little in common with the folklife studies it emerged from. In the German-

speaking countries of Europe, conversely, the term ethnology was initially adopted by *Völkerkunde*, while many university departments and programs previously called *Volkskunde* acquired new names—cultural studies, European ethnology, cultural anthropology, among others. Their name change signified a shift towards the problem-oriented study of contemporary German, Swiss, or Austrian society informed by new developments in social theory as well as in international anthropology (Greverus 1978). In France, the 'ethnologie contemporaine de la France' represents a new orientation of anthropological inquiry into French society with a strong ethnographic flavor (see Cuisenier and Segalen 1986). In the post-socialist and postcommunist countries of Eastern Europe, new orientations being established could build on the fact that prior to 1989, the discipline of ethnography had integrated scholars working within and outside of Europe. Abandoning the name ethnography that dates back to the 1929 rejection of purportedly 'bourgeois' ethnology by Moscow and Leningrad anthropologists, in the Czech Republic and in Slovakia, at Polish and at Hungarian universities, to cite a few examples, ethnology, cultural anthropology, or combinations of both have been chosen as new designations for anthropological inquiries at home and abroad. Also, in Southern Europe, regional anthropologies are developing that combine research interests in their home countries and regions with the most advanced theoretical developments in international anthropology (for a European survey see Giordano et al. 1990, Giordano and Greverus 1992). These new directions on the European continent have been entering into a dialog with a Europeanist specialization developing in American cultural anthropology and British social anthropology since the 1970s.

As early as 1955, a Europe-wide conference of ethnologists and folklorists in Arnhem, Netherlands, recommended to use the name 'European ethnology' internationally for all disciplines concerned with the anthropological study of European societies. This suggestion was taken up by the founders of a European association, the Société Internationale de Folklore et d'Ethnologie, in 1964 and a Europe-wide journal *Ethnologia Europaea* in 1967. Still, it is obvious that inconsistencies in disciplinary names will persist between and within European societies. However, at the turn of the new millennium in many continental European societies, ethnology not only continues to be widely used as a disciplinary designation, but has been adopted and even reinvented to suggest a new direction taken in the anthropological investigation of European cultures past and present (Kaschuba 1999). As social and cultural anthropology are redefining themselves as the analysis of cultural complexity in a global framework, European ethnology promises to bring this new anthropological approach to the study of European societies and their cultures of late modernity, and might ultimately prove capable of overcoming the old split between intra- and extra-European fields of research in continental anthropology.

See also: Anthropology; Boas, Franz (1858–1942); Case Study: Methods and Analysis; Ethnography; Fieldwork in Social and Cultural Anthropology

Bibliography

Boas F 1940 *Race, Language and Culture*. MacMillan, New York

Cuisenier J, Segalen M 1986 *Ethnologie de la France*. Presses Universitaires de France, Paris

Fabian J 1983 *Time and the Other. How Anthropology Makes Its Object*. Columbia University Press, New York

Geertz C 1973 *The Interpretation of Cultures. Selected Essays*. Basic Books, New York

Gerholm T, Hannerz U 1982 Introduction: The shaping of national anthropologies. *Ethnos* **47**(I-I): 5–35

Giordano G, Greverus I M (eds.) 1992 Anthropologizing Europe Continued. *Anthropological Journal on European Cultures* **1**(2): 7–112

Giordano G, Greverus I M, Stummann F J (eds.) 1990 Anthropologizing Europe. *Anthropological Journal on European Cultures* **1**(1): 7–139

Graebner F 1911 *Methode der Ethnologie. Carl Winter's*. Universitätsbuchhandlung, Heidelberg, Germany

Greverus I M 1978 *Kultur und Alltagswelt. Eine Einführung in Fragen der Kulturanthropologie*. C H Beck, Munich, Germany

Kaschuba W 1999 *Einführung in die Europäische Ethnologie*. C H Beck, Munich, Germany

Kuper A 1983 *Anthropology and Anthropologists. The Modern British School*. Routledge, London and New York

Kuper A 1988 *The Invention of Primitive Society. Transformations of an Illusion*. Routledge, London and New York

Levi-Strauss C 1975 Anthropologie. *Diogène* **90**: 3–30

Marcus G E, Fischer M M J 1986 *Anthropology as Cultural Critique. An Experimental Moment in the Human Sciences*. University of Chicago Press, Chicago

Moore F W (ed.) 1961 *Readings in Cross-Cultural Methodology*. Human Relations Area Files Press, New Haven, CT

Radcliffe-Brown A R 1951 The comparative method in social anthropology. *Journal of the Royal Anthropological Institute* 15–22

Schmidt W 1924 *Werden und Wirken der Völkerkunde*. Josef Habbel, Regensburg, Germany

G. Welz

Ethnomethodology: General

Ethnomethodology is a mode of inquiry devoted to the study of the practical methods of common sense reasoning used by members of society in the conduct of everyday life. The significance of this seemingly mundane subject matter resides in the fact that

practical reasoning is what enables societal members to make sense of the circumstances in which they find themselves, find ways of acting in relation to those circumstances, and thereby contribute to the production and maintenance of an intelligible social world. Ethnomethodology, as the study of such reasoning practices, is thus concerned with the very foundations of social order.

It should be clear from this brief definition that ethnomethodology is not primarily a theory of social life, nor is it a methodology for the study thereof. It is in the first instance a discipline of inquiry devoted to a distinctive order of phenomena, one that is sometimes understood to be orthogonal to the concerns of mainstream social science. However, these phenomena have important theoretical, methodological, and substantive ramifications, and in the course of elucidating them ethnomethodology has had a major impact across a range of social science disciplines. Although it originated within sociology, ethnomethodology's sphere of influence extends to anthropology, cognitive science, communication, linguistics, psychology, and the philosophy of the social sciences.

Although a few studies appeared in the late 1950s and early 1960s, the nascent approach crystallized and attracted widespread attention in 1967 with the publication of Harold Garfinkel's *Studies in Ethnomethodology*. This was a time of great intellectual ferment in the social sciences, and ethnomethodology contributed in no small way to the revolutions that challenged prevailing theoretical perspectives and methodological approaches. In the domain of sociological theory, ethnomethodology helped end the dominance of the structural functionalist paradigm associated with Talcott Parsons, and led to a major reconceptualization of the theory of action and its relationship to social structure. It was also an important impetus behind the broad-based social constructionist movement that revolutionized theorizing about subjects ranging from deviance to gender. In methodology, ethnomethodological studies challenged the scientistic pretensions of positivistic research methods, and fostered greater sensitivity and self-reflection among methodologists of various stripes. Finally, ethnomethodology inspired numerous research initiatives that revitalized a wide range of social science subfields, including the study of language and social interaction, the inner workings of bureaucratic and people-processing institutions, and the construction of formal scientific knowledge.

1. *Intellectual Origins*

The origins of ethnomethodology can be traced back to seminal research conducted in the late 1940s by Harold Garfinkel in the course of his graduate work in sociology at Harvard University. Garfinkel was a student of Talcott Parsons and he took Parsons' voluntaristic theory of action as a basic frame of reference, but he was also deeply influenced by the phenomenological writings of Alfred Schütz and the teachings of Aaron Gurwitsch. (Sometime later, the ordinary language philosophy of Ludwig Wittgenstein would become another source of inspiration.) The phenomenological enterprise directed Garfinkel's attention to certain fundamental pretheoretical problems posed by Parsons' theory of action, problems which were not adequately addressed within the Parsonian framework. What would later come to be known as 'ethnomethodology' is in large part the result of Garfinkel's sustained effort to confront these problems by empirical means (see *Parsons, Talcott (1902–79)*; *Phenomenology in Human Science*).

Parsons had sought to develop a theory of social action that would solve the Hobbesian problem of order while retaining purposive human agency within its framework. At the same time, he wished to avoid the limitations and pitfalls of strictly utilitarian thinking, which cannot account for the ends toward which action is oriented, and which treats intrinsic rationality as the sole standard governing the selection of means. Parsons' solution centered around the role of moral values in social action—he proposed that, through socialization, such values become internalized and incorporated into the actor's personality, where they guide both the selection of ends of action and the normative means by which they are sought. Insofar as such values are also institutionalized within a society and are shared by societal members, patterned activity and social cohesion will be the natural result.

Parsons' theory places primary emphasis on the motivational wellsprings of action, and in so doing it has relatively little to say about how concrete actions are managed and coordinated in real time. Thus, as Heritage (1987, p. 228) has observed, it is less a theory of action *per se* than a theory of the dispositions that give rise to action. Correspondingly, Parsons' theory fails to give serious consideration to the knowledge and forms of reasoning that actors themselves bring to bear in the course of producing and recognizing actual conduct. This insensitivity to actors' *endogenous* perspectives is exacerbated by Parsons' decision to define rational action in terms of an *exogenous* standard based on consistency with the application of scientific/logico-empirical methods. Any deviation from this standard is regarded as the embodiment of ignorance and error, in which case actors' own explanations of their actions may be dismissed in favor of a 'scientific' explanation cast in the language of norms and values.

Garfinkel recognized that Parsons' motivational solution to the Hobbesian problem of order implicates an analytically prior *cognitive* problem of order, one that cannot be resolved without due consideration to the common sense knowledge and practical reasoning employed by actors themselves. How do actors analyze their circumstances, determine which features are

relevant for present purposes, and grasp what those features 'mean?' How do they know what forms of behavior will be recognized as normatively appropriate under those circumstances? How in other words is the fundamental intelligibility of action and circumstance accomplished? It is questions like these that Garfinkel sought to answer.

2. *Garfinkel's Classic Studies*

Although Garfinkel's central analytic questions were inspired by phenomenology, his way of addressing them was distinctively his own. He eschewed highly interpretive methods involving introspection or free-wheeling *verstehen* in favor of a rigorously empirical approach based on close observation of social behavior. Since the methods of reasoning of interest are those embodied in social activities, they are publicly available to both lay members of society and professional analysts and are, at least in principle, eminently observable.

Analyzing such methods of reasoning-in-action, however, is intrinsically difficult because they are so deeply taken for granted. As an omnipresent resource for the management of social life, practical reasoning is not ordinarily an object of conscious reflection in its own right. Its very familiarity renders it elusive and resistant to analysis.

Garfinkel's ingenious solution to this problem entailed seeking out extraordinary situations in which the sense-making process is foregrounded, exaggerated, and hence rendered conspicuous. For the most part, these situations were engineered by Garfinkel and his associates through a series of experiments and quasi-experimental demonstrations. The situations often involved social actors encountering anomalous events—typically, particular social actions—that are incongruous with a pre-existing or default 'definition of the situation.' In one such experiment, subjects were engaged in a game of tic-tac-toe by experimenters who would erase and reposition the subjects' mark before placing their own mark on the grid. In another experiment, subjects were told to ask yes/no questions of a counselor but, unbeknownst to them, the 'answers' were given according to a random schedule. Elsewhere the focus shifted to incongruities that arise naturalistically rather than experimentally, and to the practices of social actors rather than observers—an example is the study of 'Agnes,' a person whose masculine genitalia and biography were at odds with her claim to female status. Across these examples, incongruity and its management served to throw processes of practical reasoning into sharp relief.

The results revealed that interactants would strive to render the situation as coherent and intelligible, invoking a wide range of background assumptions, contextual knowledge, and other elements of common sense—often in the form of natural language accounts—to this end. Three patterns of response were recurrent. Subjects would either (a) 'normalize' the anomalous event by reinterpreting it as consistent with the prior definition of the situation, (b) 'demonize' the event by treating it as a motivated and morally suspect departure from normality, or (c) reconstitute the environing situation so as to make it congruent with what is ostensibly taking place.

These outcomes demonstrate forcefully that interactants make sense of a given action by considering not only the intrinsic properties of the action itself, but also its relationship to the broader social environment in which it is embedded. The meaning and import of any particular action is thus thoroughly context-dependent. This would be a trivial matter if contexts were autonomous and stable—'the' context of an action could then serve as an archimedian vantage point from which to disambiguate that action. Contexts, however, do not have such transcendent properties—the relevant context at any given moment is derived primarily from the actions through which it is composed, and even a single incongruous action can dramatically alter the context that is understood to be in play. Practical reasoning is thus a dynamic and reflexive process whereby the sense of particular actions and environing social situations are ongoingly adjusted and reconciled with one another. By implication, the experiential reality and temporal persistence of any social situation—whether a game of tic-tac-toe, a course of counseling advice, or a person's sexual status—rests upon a foundation of such reasoning practices.

3. *Theoretical Implications*

Garfinkel's findings amount to a major reconceptualization of the fundamental locus of social order. Most social science theories view social life as organized by structural entities (e.g., social institutions, cultural symbol systems, structures of race, class, and gender, etc.) that stand outside of the flow of events in everyday life and exert a more or less determining influence on the course of those events. Such theories embody what might be thought of as a *top-down* conception of social order. Garfinkel, in contrast, offers a thoroughly *bottom-up* conception, and this theoretical inversion is a natural consequence of his decision to treat social order as a cognitive rather than a Hobbesian problem—not a problem of how conflict is avoided and solidarity maintained, but a problem of how the social world, whatever its character, becomes intelligible and accountable to its members. From this vantage point, every orderly feature of social life is an ongoing contingent accomplishment, the result of members' concerted effort to make those features recognizable to one another and the basis for subsequent action.

This theoretical inversion has further implications for the theory of action, and in particular for the role of social norms (as well as rules, conventions, etc.) in

the conduct of action (Heritage 1984, 1987, Wilson 1971). In the traditional Parsonian view, norms regulate action by specifying what courses of action are appropriate under given circumstances. This view presupposes that situations, norms, and actions are independent entities, with each situation standing outside of the actions contained within it, and predefined norms constraining those actions to unfold along situationally appropriate lines. However, Garfinkel's findings about the nature of practical reasoning suggest that this viewpoint is fundamentally misguided. Far from being independent entities, situations, actions, and norms stand in a co-constitutive or reflexive relationship to one another. Correspondingly, actors are knowledgeable agents at the very center of this process, with the capacity to alter or transform the 'definition of the situation' through their actions, and to decide upon the sense and applicability of the norms deemed relevant to that situation. Thus, the experiential reality of social norms, like every other organized feature of social life, rests upon a foundation of practical reasoning whereby action is produced and rendered intelligible in normative terms. This does not mean that norms are inconsequential for social organization, but their primary significance is constitutive rather than regulative—norms play a crucial role as a resource for imputing meaning and motivation to situated behavior (see *Action, Theories of Social*; *Norms*).

4. *Contemporary Research Initiatives*

Following Garfinkel's classic studies, ethnomethodological research developed in a number of fruitful directions. While some researchers continued to elucidate the generic properties of practical reasoning (e.g., Pollner 1987), most shifted toward examining such reasoning as it is applied in various specialized domains of social life. At the same time, the predominant research methods employed by ethnomethodologists underwent a corresponding shift away from quasi-experimental designs and toward more naturalistic methods involving the direct observation of ordinary conduct. The resulting corpus of studies resists easy summary, but three prominent lines of work will be distinguished.

4.1 *People-processing Institutions*

Ethnomethodologists first explored various organizational environments involved in people processing: schools, public welfare offices, police departments, and so on (e.g., Cicourel 1968, Sudnow 1965, Wieder 1974). These studies repeatedly documented the inadequacy of codified rules, formal procedures, and informal norms as explanations of organizational conduct. Rules and allied phenomena fail to capture the elaborate judgmental work necessary to implement the rules in concrete circumstances, and more generally to perform competently the tasks intrinsic to each setting. These previously unexamined professional competencies and discretionary judgements became the focus of close scrutiny and sustained analysis, shedding new light on the complex inner workings of people-processing institutions.

The findings have important implications for our understanding of organizational decision-making and its products. The institutions under examination routinely generate official designations of persons (e.g., 'criminal,' 'juvenile delinquent,' 'qualified welfare recipient,' etc.) and their actions (e.g., 'burglary,' 'assault,' etc.). Such designations do not result from workers mechanically applying clearly defined criteria to each case, but involve various *ad hoc* considerations guided by common sense knowledge of what outcome would be normal and reasonable under the circumstances. Insofar as the resulting designations form the basis for calculations of official statistics (on rates of crime, mortality, poverty, etc.), these studies also cast doubt on the validity of official statistics and their value for social scientific research (see *Bureaucracy, Sociology of*).

4.2 *Science and Technology*

Ethnomethodologists have also ventured into the laboratory to explore the highly technical competencies that underlie the creation of scientific and mathematical knowledge, and the production and use of technological artifacts. This line of research is most closely associated with the later work of Garfinkel himself and his immediate associates (e.g., Livingston 1986, Lynch 1993). It is reminiscent of Edmund Husserl's later writings on the European sciences in that a primary objective is to elucidate the unexplicated foundations of scientific knowledge. However, where Husserl's enterprise was essentially philosophical in character and retained a phenomenological concern with transcendental consciousness, the ethnomethodological approach proceeds empirically by examining publicly available details of situated 'workbench' practices.

Much research in this vein is concerned with the genesis and reproduction of technical discoveries ranging from physical objects to mathematical theorems. Such discoveries are arrived at, accountably validated, and made intersubjectively available through complex courses of practical reasoning and embodied activity that receive scant attention in scientific texts. This may seem surprising, given that scientists are supposed to document their own methods so as to permit replication, but scientists' methodological descriptions—like all abstracted accounts

of situated action—are necessarily incomplete. Ethnomethodologists thus seek to recover the mundane praxiological foundations of discovering work, and in so doing they re-specify discovered objects as locally produced and naturally accountable achievements. Furthermore, the very concept of discovery is transformed by this research. If physical and mathematical objects are unavoidably intertwined with situated practices, then the discovery of any such object is at the same time the discovery of a substrate of practices through which that object may be accountably produced and reproduced within concrete situations.

Ethnomethodological studies of science and technology hold the promise of yielding findings that represent recognized contributions to the various technical disciplines being studied. Indeed, the practical value of ethnomethodology is increasingly recognized in computer science, software engineering, and human–computer interaction (Suchman 1987, Button 1993) (see *Technology, Anthropology of*; *Truth and Credibility: Science and the Social Study of Science*).

4.3 Talk and Social Interaction

Perhaps the most widespread contemporary variant of ethnomethodology is what has come to be known as *conversation analysis*. This burgeoning field was developed by Harvey Sacks, originally Garfinkel's student and colleague, in collaboration with Emanuel Schegloff and Gail Jefferson (e.g., Sacks 1992, Atkinson and Heritage 1984).

Conversation analysis (henceforth CA) involves the study of practical reasoning as it is put to use in the conduct of spoken interaction. The domain of interaction—what Erving Goffman referred to as *the interaction order*—is more general than any of the specialized institutional domains investigated by ethnomethodologists, for interaction lies at the heart of virtually all of these institutions and extends as well to informal encounters between persons. Moreover, in the history of the human species, interaction developed long before other societal institutions came into being, and is in this respect 'the primordial site of sociality' (Schegloff 1988).

CA differs from other lines of ethnomethodological research not only in substance but in methodology. Conversation analysts rely exclusively on audio- and video-recordings of interactional data, and transcripts that capture the details of interaction as it actually occurs. Such data have numerous advantages—they can be examined repeatedly, analyzed at an unprecedented level of detail, and reproduced in published works so that readers can independently assess the validity of analytic claims.

The resulting research enterprise has generated an impressive array of interlocking and cumulative findings on a wide range of subjects. These include the organization of turn taking, action sequences, lexical choice, the relationship between talk and nonvocal activities, and the collaborative management of various interactional activities (e.g., giving advice, delivering good and bad news, telling troubles, etc.). More recently, researchers have applied the analytic resources of CA to various phenomena that intersect with, and can be informed by, the study of talk-in-interaction. These include how talk is organized in various institutional settings, and how it serves as a medium for the accomplishment of occupational tasks such as medical examinations, classroom lessons, journalistic interviews, trial examinations, and so on (e.g., Boden and Zimemrman 1991, Drew and Heritage 1992). Researchers have also begun to explore how the study of talk-in-interaction can illuminate linguistic phenomena such as grammar (Ochs et al. 1996), as well as medical disorders such as aphasia that manifest themselves at the level of speech (Goodwin 1995, Heeschen and Schegloff 1999).

See also: Conversation Analysis: Sociological; Parsons, Talcott (1902–79); Science and Technology Studies: Ethnomethodology

Bibliography

Atkinson J M, Heritage J (eds.) 1984 *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, Cambridge, UK

Boden D, Zimmerman D H (eds.) 1991 *Talk and Social Structure*. Polity Press, Cambridge, UK

Button G (ed.) 1993 *Technology in Working Order: Studies of Work, Interaction and Technology*. Routledge, London

Cicourel A V 1968 *The Social Organization of Juvenile Justice*. Wiley, New York

Drew P, Heritage J (eds.) 1992 *Talk at Work*. Cambridge University Press, Cambridge, UK

Garfinkel H 1967 *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ

Goodwin C 1995 Co-constructing meaning in conversations with an aphasiac man. *Research on Language and Social Interaction* **28**: 233–60

Heeschen C, Schegloff E A 1999 Agrammatism, adaptation theory, and conversation analysis: on the role of the so-called telegraphic style in talk-in-interaction. *Aphasiology* **13**: 365–405

Heritage J 1984 *Garfinkel and Ethnomethodology*. Polity Press, Cambridge, UK

Heritage J 1987 Ethnomethodology. In: Giddens A, Turner J (eds.) *Social Theory Today*. Polity Press, Cambridge, UK, pp. 224–72

Livingston E 1986 *Ethnomethodological Foundations of Mathematics*. Routledge, London

Lynch M 1993 *Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science*. Cambridge University Press, Cambridge, UK

Ochs E, Schegloff E A, Thompson S A 1996 *Interaction and Grammar*. Cambridge University Press, Cambridge, UK

Pollner M 1987 *Mundane Reason: Reality in Everyday and Sociological Discourse*. Cambridge University Press, Cambridge, UK
Sacks H 1992 *Lectures on Conversation*. Blackwell, Oxford, UK
Schegloff E A 1988 Goffman and the analysis of conversation. In: Drew P, Wootton A (eds.) *Erving Goffman: Exploring the Interaction Order*. Polity Press, Cambridge, UK, pp. 89–135
Suchman L A 1987 *Plans and Situated Actions: The Problem of Human Machine Communication*. Cambridge University Press, Cambridge, UK
Sudnow D 1965 Normal crimes. *Social Problems* **12**: 255–76
Wieder D L 1974 *Language and Social Reality*. Mouton, The Hague
Wilson T P 1971 Normative and interpretive paradigms in sociology. In: Douglas J (ed.) *Understanding Everyday Life*. Routledge, London, pp. 57–79

S. E. Clayman

Ethnonationalism: Cultural Concerns

Claims to national identity based on race, kinship, language, or common culture are described as 'ethnonationalist.' Such ethnonationalist claims have been widespread throughout the modern era. They sometimes extend beyond the construction of identity to the reproduction of enmity, demands that members place the nation ahead of other loyalties, and attempts to purge territories of those defined as foreign. As a result, ethnonationalism is often associated with ethnic violence and projects of ethnic cleansing or genocide. However, ethnic solidarity is also seen by many as basic to national identity as such, and thus to the notion of the nation-state. While this notion is as much contested as defended, it remains influential.

In such usage, ethnonationalism is commonly opposed to civic nationalism. The latter is understood as the loyalty of individual citizens to a state based purely on political identity. Habermas (1998) has theorized this as constitutional patriotism, stressing the extent to which political loyalty is to a set of institutional arrangements rather than a prepolitical culture or other extrapolitical solidarity. Ethnonationalism, in such usage, refers precisely to rooting political identity and obligation in the existence of a prepolitical collective unit—the nation—which achieves political subjectivity by virtue of the state. The legitimacy of the state, in turn, is judged by reference to the interests of the nation.

The contrast of ethnic to civic nationalism is heavily influenced by that of Germany to France (Kohn 1967, Alter 1989). The contrast has been enduring, and has resulted in different understandings of citizenship. France has been much more willing, for example, to use legal mechanisms to grant immigrants French citizenship, while Germany—equally open to immigration in numerical terms—generally refuses its immigrants German citizenship unless they are already ethnic Germans (Brubaker 1992). Other countries vary on the same dimension, but it is important to recognize that the difference is one of proportion and ideological emphasis (Calhoun 1997, Sassen 1999). As Smith (1986, p. 149) has remarked, 'all nations bear the impress of both territorial and ethnic principles and components, and represent an uneasy confluence of a more recent "civic" and a more ancient "genealogical" model of social cultural organization.' Not all scholars accept the distinction or hold it to be sharp; those who do use it often attribute ethnonationalism to countries that are 'late modernizers' (Bendix 1964, Nairn 1998, Schwarzmantel 1991).

Two enduring debates have shaped social science scholarship on ethnonationalism. First, is ethnonationalism an ancient (or historically nonspecific) phenomenon, possibly rooted in 'primordial' social relations, or is it distinctively modern? Relatedly, is it vanishing, enduring, or recurrent? Second, is ethnicity basic to nationalism in general, perhaps even an explanation for nationalism, and is ethnonationalism thus its 'normal' form? Much nationalist ideology has claimed definitive ethnic roots; social scientists are more divided on the question. Beyond these broad questions of orientation, research focuses on a variety of issues from the cultural content of ethnonationalism to explanations of its occurrence, forms of public performance, dynamics of leadership and mobilization, and reasons for violence.

1. *Modernity vs. Primordiality*

A long-running debate in the literature on nationalism pits arguments that it is an extension of ancient ethnicity (Smith 1986, Armstrong 1982, Hutcheson 1994) against those who argue that it is essentially modern (Gellner 1983, Hobsbawm 1990, Greenfeld 1992). Majority scholarly opinion tends toward the latter view, though explanations differ. 'Modernists' variously see nationalism rooted in industrialization (Gellner 1983), state-formation (Tilly 1990, Mann 1993), the rise of new communications media and genres of collective imagination (Deutsch 1966, Anderson 1991), and the development of new rhetorics for collective identity and capacities for collective action (Calhoun 1997). While many favor specific factors as primary explanations, most recognize that several causes are interrelated.

Many nationalists but few scholars see nationalism as ubiquitous in history and simply the 'normal' way of organizing large-scale collective identity. Most social scientists point rather to the variety of political and cultural forms common before the modern era—empires and great religions, for example—and the transformations wrought by the rise of a new kind of

intensive state administration, cultural integration, popular political participation, and international relations. Many of these social scientists argue that nations and nationalism in their modern sense are both new. In particular, they would argue that ethnicity as a way of organizing collective identity underwent at the least a substantial reorganization when it began to be deployed as part of ethnonationalist rhetoric in the modern era. Others, however, including notably Anthony Smith (1986, 1991, 1998) and John Armstrong (1982) argue that there is more continuity in the ethnic core of nations, though they too would agree that modernity transformed—if it did not outright create—nationalism.

The attraction of a claimed ethnic foundation to nations lies largely in the implication that nationhood is in some sense primordial and natural. Nationalists typically claim that their nations are simply given and immutable rather than constructions of recent historical action or tendentious contemporary claims. Much early scholarly writing on nations and nationalism shared in this view and sought to discover which were the 'true' ethnic foundations of nationhood (Skurnowicz 1981, Zacek 1969). It is no doubt ideologically effective to claim that a nation has existed since time immemorial or that its traditions have been passed down intact from heroic founders. In no case, however, does historical or social science research support such a claim. All nations are historically created.

Noting this, one line of research emphasizes the manipulation of popular sentiments by the more or less cynical production of national culture by intellectuals and state building elites. Hobsbawm and Ranger (1983), for example, have collected numerous examples of the ways in which apparently definitive cultural markers of ethnonational identity can in fact be traced to specific acts of creation embedded in political (or sometimes marketing) projects rather than reflecting pre-existing ethnicity. The Scots tartan kilt is a famous example, dating not from the mists of primordial Highland history but from eighteenth century resistance to Anglicization (Trevor-Roper 1983). Likewise, nineteenth century Serbian and Croatian intellectuals strove to divide their common Serbo-Croatian language into two distinct vernaculars with separate literary traditions. But as this example makes clear, it is not obvious that because the 'traditions' of nationalism are 'invented' they are somehow less real or valid (see also Hobsbawn 1990). Anderson (1996, p. 6) finds the same fault with Gellner: 'Gellner is so anxious to show that nationalism masquerades under false pretences that he assimilates "invention" to "fabrication" and "falsity," rather than to "imagining" and "creation".'

Hobsbawm and Ranger imply that long-standing 'primordial' tradition would somehow count as legitimate, while by contrast various nationalist traditions are of recent and perhaps manipulative creation. Many ideologues do claim origins at the dawn of history, but few scholars have doubted that cultural traditions are constantly renewed. What so-called 'primordialists' have argued is that certain identities and traditions—especially those of ethnicity—are *experienced* as primordial (Geertz 1963). Sociologically, thus, what matters is less the antiquity of the contents of tradition, than the efficacy of the process by which certain beliefs and understandings are constituted as unquestioned, immediate knowledge. This has more to do with current bases for the reproduction of culture than with history as such. Ethnicity or cultural traditions are bases for nationalism because they effectively constitute historical memory, because they inculcate it as 'prejudice,' not because the historical origins they claim are accurate (prejudice means, following Gadamer (1981), not just prior to judgment, but constituting the condition of judgment). Moreover, all traditions are 'invented' (or at least in a more diffuse sense, created); none are truly primordial. This was acknowledged, though rather weakly, even by some of the functionalists who emphasized the notion of primordiality and the 'givenness' of cultural identities and traditions (see Eisenstadt 1966, 1973, Geertz 1963, Gellner 1964). All such traditions also are potentially contested and subject to continual reshaping, whether explicit or hidden. Some claims about nationality may fail to persuade because they are too manifestly manipulated by creators, or because the myth that is being proffered does not speak to the circumstances and practical commitments of the people in question.

Notions of nations as acting subjects are distinctively modern, part of a new way of constructing collective identity (Steiner 1988). This said, there is no scholarly agreement about when nationalism began. Greenfeld (1992) dates it from the English Civil War, Anderson (1991) from Latin American independence movements, Alter (1989) from the French Revolution, and Breuilly (1993) and Kedourie (1993) both from German Romanticism and reaction to the French Revolution. Calhoun (1997) suggests that rather than trying to identify a single point of origin, scholars should see nationalism as drawing together several different threads of historical change. As a discursive formation, it took on increasingly clear form through the early modern period and was fully in play by the Napoleonic era.

2. *Nationalism as an Ethnic Phenomenon*

Most prominent twentieth-century analysts of nationalism have sought to challenge the explanation of nationalism by ethnicity. Kohn (1944) and Seton-Watson (1977) stress the crucial role of modern politics, especially the idea of sovereignty. Hobsbawm (1990) treats nationalism as a kind of second-order political movement based on a false consciousness which ethnicity helps to produce but cannot explain

because the deeper roots lie in political economy not culture. The dominant approach in contemporary scholarship treats nationalism largely as an ideological reflection of state formation (Tilly 1975, 1990, Hall 1995, Mann 1993). Gellner (1983) emphasizes industrialization, and also stresses the number of cases of failed or absent nationalisms: ethnic groups which mounted either little or no attempt to become nations in the modern senses. This suggests that even if ethnicity plays a role it cannot be a sufficient explanation (though one imagines the nineteenth-century German Romantics would simply reply that there are strong, historic nations and weak ones destined to fade from the historic stage). Hayes (1931) argues for seeing nationalism as a sort of religion. Hechter (2000) analyzes it in terms of strategic individual action aimed at maximizing mostly economic and political benefits. Kedourie (1993) approaches nationalism as an ideology, and attempts to debunk nationalism by showing the untenability of the German Romantic cultural-ethnic claims. Indeed, in their different ways, all these thinkers have sought to debunk the common claims that nationalists themselves make to long-established ethnic identities.

Against this backdrop, Smith acknowledges that nations cannot be seen as primordial or natural, but nonetheless argues that they are rooted in relatively ancient histories. Smith argues that the origins of modern nationalism lie in the successful bureaucratization of aristocratic *ethnie*, which were able to transform themselves into genuine nations only in the West. In the West, territorial centralization and consolidation went hand in hand with a growing cultural standardization. Nations, Smith thus suggests, are long-term processes, continually re-enacted and reconstructed; they require ethnic cores, homelands, heroes, and golden ages if they are to survive. 'Modern nations and nationalism have only extended and deepened the meanings and scope of older ethnic concepts and structures. Nationalism has certainly universalized these structures and ideals, but modern "civic" nations have not in practice really transcended ethnicity or ethnic sentiments' (Smith 1986, p. 216).

The ethnic similarities and bonds that contribute to the formation of nations may indeed be important and long-standing, but in themselves they do not fully constitute either particular nations or the modern idea of nation. While some critics of ethnic explanations of nationalism emphasize the influence of state formation or other 'master variables,' a number assert that nations are created by nationalism—by this particular form of discourse, political rhetoric, or ideology—not merely passively present and awaiting the contingent address of nationalists (Kedourie 1993, Gellner 1983, Anderson 1991, Chatterjee 1986).

An emphasis on pre-existing ethnicity—even where this is rightly identified—is unable to shed much light on why so many modern movements, policies, ideologies, and conflicts are constituted within the discourse of nationalism. Indeed, as Gellner (1983, pp. 8–18, 61) has suggested, the very self-recognition of ethnicities or cultures as defining identities is distinctively modern. Walker Connor (1994, p. 103) uses a similar point to distinguish ethnic groups as 'potential nations' from real nations: 'While an ethnic group *may*, therefore, be other-defined, the nation *must* be self-defined.'

Explanations of ethnonationalism, thus, need to address the contemporary conditions that make it effective in people's lives, attempts to orient themselves in the world, and actions. Such conditions are of course subject to change and nationalist constructions are apt to change with them. Thus Indian nationalists from the nineteenth century through Nehru were able to make a meaningful (though hardly seamless or uncontested) unity of the welter of subcontinental identities as part of their struggle against the British. The departure of the British from India changed the meaning of Congress nationalism, however, as this became the program of an Indian state, not of those outside official politics who resisted an alien regime. Among other effects of this, a rhetorical space was opened up for 'communal' and other sectional claims that were less readily brought forward in the colonial period (Chatterjee 1994). Similarly, the proliferation of nationalisms in Eastern Europe attendant on the collapse of communist rule involved a 'reframing' of older national identities and nationalist projects; the nationalisms of the 1990s were neither altogether new nor simply resumptions of those that predated communism. The opposition between primordiality and 'mere invention' leaves open a very wide range of historicities within which national and other traditions can exert real force.

In 1871, in one of the most famous essays ever written on nationalism, Ernst Renan (1990, p. 11) grasped the importance of the tensions masked in nationalist invocations of history:

> Forgetting, I would even go so far as to say historical error, is a crucial factor in the creation of a nation, which is why progress in historical studies often constitutes a danger for [the principle of] nationality. Indeed, historical enquiry brings to light deeds of violence which took place at the origin of all political formations, even those whose consequences have been altogether beneficial. Unity is always effected by means of brutality.

The common contrast between France and Germany—the literature's paradigmatic opposition of civic to ethnic nationalism—thus, is between two different styles of invoking history and ethnicity, not radically between nonethnic and ethnic claims. French schoolchildren learn that their commonality is not merely ethnic but achieved in the collective action of the Revolution. German nationalist historians put forward stronger claims for the primacy of common culture and ethnicity partly because their narratives

must help schoolchildren 'forget' that Germans spent most of their history as members of separate polities (often combative and not all very uniform culturally), even while they celebrate the roles of Bismarck and others in unifying Germany.

Like all forms of nationalism, ethnonationalism is partly a matter of narrative construction, the production (and reproduction and revision) of narratives locating the nation's place in history (Bhabha 1990). As Anderson (1991) puts it, nations move through historical time as persons move through biographical time; each may figure in stories like characters in a novel. This is one reason why the continuity of ethnic identities alone does not adequately explain nationalism: The narrative constructions in which it is cast change and potentially transform the meaning of whatever ethnic commonalties may exist. Ironically, the writing of linear historical narratives of national development and claims to primordial national identity often proceed hand in hand. Indeed, the writing of national historical narratives is so embedded in the discourse of nationalism that it almost always depends rhetorically on the presumption of some kind of preexisting national identity in order to give the story a beginning. A claim to primordial national identity is, in fact, a version of nationalist historical narrative.

3. Conclusion

The modern literature on nationalism is centrally structured by debates over the extent to which nationalism is based on and/or explicable by ethnicity. The rhetoric employed by nationalists around the world commonly presents nationality as an extension of ethnicity, but academic analysis often challenges this. Researchers point to the extent to which national identities are based on new cultural creativity, not ancient traditions, or on selective appropriation of ancient traditions. They stress the role of state formation, industrialization, and economic change, of geopolitical pressures and rivalries, of the availability of a globally distributed rhetoric of nationalism, and of the interests of those who promote nationalism. Even if ethnicity is not the primary explanation of nationalism, it still often plays a role—at the very least in ideology. Movements and policies shaped by ethnonationalism are powerful sources of both collective identity and collective conflict in the contemporary world. They do not seem to be declining in significance, but rather to be among the main forms through which people grapple with the challenges and opportunities posed by globalization.

See also: Cultural Evolution: Ethnogenesis; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Ethnicity, Race, and Health; Ethnocentrism; National Character; Nationalism and Expressive Forms; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: East Asia; Nationalism, Historical Aspects of: South Asia; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of

Bibliography

Alter P 1989 *Nationalism*. Edward Arnold, London

Anderson B 1991 *Imagined Communities*, rev. edn. Verso, London

Anderson B 1996 Introduction. In: Balakrishnan G (ed.) *Mapping the Nation*. Verso, London

Armstrong J 1982 *Nations before Nationalism*. University of North Carolina Press, Chapel Hill, NC

Bendix R 1964 *Nation-Building and Citizenship*. University of California Press, Berkeley, CA

Bhabha H (ed.) 1990 *Nation and Narration*. Routledge, London

Breuilly J 1993 *Nationalism and the State*, rev. edn. University of Chicago Press, Chicago

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Cambridge University Press, Cambridge, UK

Brubaker R 1996 *Nationalism Reframed*. Cambridge University Press, Cambridge, UK

Calhoun C 1997 *Nationalism*. Open University Press, Milton Keynes, UK

Chatterjee P 1986 *Nationalist Thought and the Colonial World: A Derivative Discourse?* Zed Books, Atlantic Highlands, NJ

Chatterjee P 1994 *The Nation and Its Fragments: Studies in Colonial and Post-Colonial Histories*. Princeton University Press, Princeton, NJ

Connor W 1994 *Ethnonationalism*. Princeton University Press, Princeton, NJ

Deutsch K W 1966 *Nationalism and Social Communication: An Inquiry into the Foundations of Nationality*, 2nd edn. MIT Press, Cambridge, MA

Eisenstadt S 1966 *Modernization, Protest, and Change*. Prentice-Hall, Englewood Cliffs, NJ

Eisenstadt S 1973 *Building States and Nations*. Sage, Beverly Hills, CA

Gadamer H-G 1981 *Truth and Method*. Seabury, New York

Geertz C 1963 *Old Societies and New States*. New York

Gellner E 1964 *Thought and Change*. Weidenfeld and Nicholson, London

Gellner E 1983 *Nations and Nationalism*. Blackwell, Oxford, UK

Greenfeld L 1992 *Nationalism: Five Paths to Modernity*. Harvard University Press, Cambridge, MA

Habermas J 1998 *The Inclusion of the Other*. MIT Press, Cambridge, MA

Hall J 1995 Nationalisms, classified and explained. In: Periwal S (ed.) *Notions of Nationalism*. Central European University Press, Budapest, Hungary, pp. 8–33

Hayes C J H 1931 *The Historical Evolution of Modern Nationalism*. RR Smith, New York

Hechter M 2000 *Containing Nationalism*. Oxford University Press, New York

Hobsbawm E 1990 *Nations and Nationalism Since 1780: Programme, Myth, Reality*. Cambridge University Press, Cambridge, UK

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Kohn H 1944 *The Age of Nationalism*. Harper and Row, New York
Mann M 1993 *Sources of Social Power*. Cambridge University Press, Cambridge, UK, Vol. 2
Nairn T 1998 *Faces of Nationalism: Janus Revisited*. Verso, London
Noiriel G 1998 *The French Melting Pot*. University of Minnesota Press, Minneapolis, MN
Renan E [1871] 1990 What is a nation? In: Bhabha H (ed.) *Nation and Narration*. Routledge, London, pp. 8–22
Sassen S 1999 *Guests and Aliens*. University of Chicago Press, Chicago
Schwarzmantel J 1991 *Socialism and the Idea of the Nation*. Harvester Wheatsheaf, Hemel Hempstead, UK
Seton-Watson H 1977 *Nations and States*. Westview Press, Boulder, CO
Skurnowicz J S 1981 *Romantic Nationalism and Liberalism: Joachim Lelewel and the Polish National Idea*. New York
Smith A 1983 *Theories of Nationalism*. Duckworth, London
Smith A 1986 *The Ethnic Origins of Nations*. Blackwell, Oxford, UK
Smith A 1991 *National Identity*. Penguin, London
Smith A 1998 *Nationalism and Modernism*. Routledge, London
Steiner G 1970 The French Revolution and history. In: Best P (ed.) *The French Revolution*. University of Chicago Press, Chicago
Tilly C (ed.) 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ
Tilly C 1990 *Coercion, Capital and European States, AD 990–1990*. Blackwell, Cambridge, MA
Trevor-Roper H 1983 The invention of tradition: The Highland tradition of Scotland. In: Hobsbawm E, Ranger T (eds.) *The Invention of Tradition*. Cambridge University Press, Cambridge, UK, pp. 15–42
Zacek J F 1969 Nationalism in Czechoslovakia. In: Sugar P F, Lederer I J (eds.) *Nationalism in Eastern Europe*. University of Washington Press, Seattle, WA

C. Calhoun

Etymology

'Etymology' is the branch of linguistic science that treats the history of words and their components, with the aim of determining their origin and their derivation. It may also be defined as word history. In this sense one may speak of the etymology of a word.

1. Relation to Historical Linguistics

The term is attested in English from the late fourteenth century, when as in earlier periods the study was aimed at determining the 'true meaning' of words, often with fanciful results. Accurate results were achieved only in the nineteenth century as historical linguistics developed. From that time, historical treatment of the lexicon has gradually come to be dealt with separately from historical treatment of the grammar. Historical linguistics deals with phonology, inflectional morphology, and syntax, while etymology confines its attention to the lexicon and derivational morphology; there has been some overlap. Etymological dictionaries may include entries on affixes like the past participle suffix *-ed*, as in *hoped*, probably because many of these are used as adjectives as well, e.g., in *long hoped* …, but dictionaries generally omit inclusion of the past tense suffix *-ed* and other inflectional affixes.

Although etymology omits treatment of grammatical elements, it relies heavily on historical grammar for phonology and morphology. It may also draw on syntax, as in the treatment of compounds. Advances in etymology have accordingly gone hand-in-hand with advances in historical linguistics.

2. Gradual Development as an Independent Branch

In the early decades of the nineteenth century etymological studies were directed at comparing prior forms in previous periods of the language and related forms in other dialects of that language, often in connection with historical essays, monographs, and grammars. Jacob Grimm (1785–1863) in his Germanic grammar of 1821 set side by side cognates like Greek *treĩs*, Latin *trēs*, Gothic *þṙeis*, Old High German *drī*, and proposed rules for indicating the corresponding consonants, e.g., |**gk t. goth th. ohg d**|. He also recognized discrepancies, such as the differing correspondences with Latin *t* in *frāter*, Gothic *brōþar* and Latin *pater*, Gothic *fadar*. But he did not account for the difference between the medial consonants of the two sets, nor for the vowels. Similarly Friedrich Diez (1794–1876) in his etymological dictionary of the Romance languages (1858) merely brought together related words centering on Italian.

Following etymological investigations as by August Pott (1802–87) (Pott 1836), August Fick (1833–1916) published a comparative dictionary of the Indo-European languages (Fick 1871) half a century after Grimm's grammar. August Schleicher (1821–68) had introduced the practice of providing reconstructed/starred forms (Schleicher 1862). Etymological dictionaries now provided the etymon and history of words rather than comparative lists. Etymology has subsequently been treated independently of historical grammar, as may be exemplified by Walter Skeat's (1835–1912) etymological dictionary of the English language (1909) as well as the comparable German work in 1883 by Friedrich Kluge (1856–1926) (Kluge 1989). They were brought repeatedly up to date, as Kluge's has been with more than twenty editions.

3. Explanation of all Elements of Words

By the end of the nineteenth century etymological essays, monographs, and dictionaries aimed at accounting for all the phonological and morphological changes in the development of words and stating their oldest attainable form. Methodology had become precise in accordance with increased understanding of phonology and morphology. In the fourth decade of the century linguists controlling the increasing knowledge of phonetics accounted for the unchanged *p t k* in Germanic words like *spew*, Latin *spuō*, *stand*, Latin *stāre* by noting their position after fricatives; attention then was directed not only at individual sounds but also at combinations of them. In 1863 Hermann Grassmann (1809–77) (Grassmann 1863) accounted for the discrepancy exemplified in the initial *b* of Sanskrit *bandhati* and that of NE *bind* and similar patterns through establishing that an aspirate was deaspirated in Sanskrit and Greek when a second aspirate occurred in the following syllable. The Indo-European root was accordingly posited as **bhendh-* and the Germanic *b-* was in accordance with the general change posited by Grimm. Attention then was directed at entire words, not merely at individual sounds and their immediate neighbors. After Karl Verner (1846–96) accounted for the voiced *d* in Gothic *fadar* and the voiced rather than voiceless fricatives of other such words by pointing to the position of their accent, as in the Greek cognate *patḗr* (Verner 1875), attention was directed at suprasegmentals as well as segmentals. Subsequently etymological studies treated all the formal elements of words—individual sounds, sounds in combination, and accentuation.

The entry for *father* in Skeat's dictionary, brought up to date in 1909, may serve as an example. After a brief indication of the meaning, 'a male parent,' Skeat provides the Middle English and Old English form *fader*; thereupon he adds the cognates in other Germanic dialects and finally those in some of the Indo-European dialects, concluding with the reconstructed Proto-Indo-European **pater-*. He then lists a few derivatives, e.g., *father-hood*. The entry includes one formal comment, that 'the spelling *fader* is almost universal in Middle English (and) *father* is due to dialectal influence, which changed *-der* to *-ther*.' The entry may illustrate the concentration on the form of individual words characteristic of many etymological dictionaries.

4. The Impact of Expanded Linguistic Studies

While greater control of phonology was being achieved, increasing attention was given to meaning and change of meaning. The results may be illustrated by the entry for the German cognate of *father* in the Kluge dictionary, which has been revised many more times than that of Skeat. After providing formal information, much like the entry of Skeat, the Kluge entry has an additional explanatory statement on the meaning: 'The word refers clearly to the social status of the father (as master of the house) though a more precise etymology is not certain. Beside it there exists in most languages a familiar word, an imitative word (doubtless of child's language) of the type Gothic *atta*, NE *daddy*, French and New High German *Papa*, etc.).' This is followed by a list of cognates in the other present-day Germanic languages and by references to *Patriot*, *Patrizier*, *Vetter*. The entries for these are similar in pattern. They indicate that their etymon may be found under that for *Vater*. The entry for *Vetter* (cousin) also refers to other relationship terms which in turn contain further such references. Even though the kinship system is not provided as such, it may be inferred through the cross-references.

Such entries illustrate that a great deal of attention had been given to the position of words in society; the attention was fostered by the developing fields of dialect geography and sociolinguistics in the latter part of the nineteenth century. There was also more specific concern for the history of nonnative or, as they are generally known, borrowed words, as may be noted from the references under *Vater*. In addition, words were examined by semantic classes, such as the numerals, or wider areas, such as clerical, military, and political terms. The impact of all such relationships was then identified for more precise historical treatment. Conversely the information drawn from such relationships was applied to determine in greater detail the social conditions they revealed.

5. Various Kinds of Imported Words and their Modifications

Influences from other languages may be introduced in a number of ways. Words may be adopted with little change. These are known as 'borrowings.' Examples from the early contacts between Romans and Germanic speakers are words referring to improved transportation and measurement, such as *street*, Latin *strāta* (*via*) 'paved road,' *mile* Latin *mīlle* '1000,' and *inch*, Latin *uncia* 'twelfth.' As in these words, modifications result from differences in pronunciation of individual sounds or in inflectional endings. Modifications may also be made in meanings. The Latin adverb *alibī* 'elsewhere' has been taken over as a noun in English, as has the Latin verb *affīdāvit* 'he has pledged his faith.'

Rather than adoption, a foreign word may be translated, utilizing native elements. Several processes may be involved in such importations. If the foreign word or its elements are rather literally translated, the results are known as 'loan translation' or calques. Examples may be cited from the time when classical

culture and religion were introduced into northern Europe. The word *heathen* 'inhabitant of the heath or countryside' is a translation of the Latin *pāgānus*. Similarly, *gospel* Old English *gōd-spell* 'good story' is a translation of Greek *euaggélion*, also imported as *evangel*.

If the translation is not as exact as in these words, it is referred to as a 'loan rendition.' Examples are prominent in German, in which *Gewissen* corresponds to Latin *conscientia*, *conscience*, *Sauerstoff* to *oxygen*. If the foreign word is taken as a model but not reproduced with similar elements, it is referred to as a 'loan creation,' for example German *Auferstehung*, Latin *resurrectio*, *resurrection*. The meaning of foreign words may be taken over for native words in 'semantic loans.' Words like *God*, *heaven*, *hell*, *Lord* were modified in this way with the introduction of Christianity.

After foreign words or elements are introduced into a language their components may be paired with native components, resulting in 'hybrids.' The native inflectional suffixes of English like the genitive *-s* and the participial *-ing* are used with non-native nouns and verbs; on the other hand, non-native suffixes like *-able* are used with many native English words, as in *doable*.

As another modification, foreign words may be modified to fit the native pattern resulting in 'back formations.' Widely cited examples are *pea* from Old English *pise*, Latin *pisum*, where the *-s* was taken as plural marker, and *cherry*, based on French *cerise*.

These processes involved in etymological explanation may illustrate that etymology is closely related to historical grammar.

6. Relationships Between Etymology, Historical Grammar, and Culture

Etymological investigations contribute information on the phonological or morphological status of languages concerned at a given time. For example, the word *wine* adopted from Latin at the beginning of our era informs us that the initial consonant of *vīnum* was still pronounced as a resonant. When *vine was* later introduced into English, it had been shifted to a fricative. Similarly the borrowing of Latin *catīnus* as *kettle* informs us that vowel mutation in the West Germanic languages took place only later.

Such information is especially important in providing information on the early periods of languages. The importation of Proto-Indo-European words into Proto-Finno-Ugric, e.g., **medhu-* 'honey,' PFU **mete* among other borrowings, and that of PIE **pek'u-* 'cattle' into Proto-North-Caucasian **päHäKwV* among other borrowings, has been taken as evidence that the speakers of Proto-Indo-European were located along the Volga in the fifth millennium BC.

The words affected also inform us of the social and cultural relations between the two sets of speakers. English provides paradigmatic information. The words taken over from Latin at the beginning of our era such as those cited above belong to the everyday vocabulary but suggest greater technical advances for the Latin speakers.

A half century later many words were taken over from Scandinavian, such as *bank*, *car*, *sky* as well as the pronouns *they*, *them*, *their*. Adoption of these words indicates that both sets of speakers shared approximately the same culture.

The sets of words adopted from French and the classical languages after 1000 AD belong in general to advanced facets of society, such as political, clerical, military and, other such spheres. They suggest introduction of items from speakers of a more complex society.

We may regard these three types of borrowings as patterns for interpreting the findings regarding language relationships generally. Importations among the languages of Europe today are comparable to those between Scandinavian and English. Influence of Arabic on Spanish from the tenth century, as of Arabic and Persian on Turkish, is comparable to that of French on English. Historical examination of the lexicon to determine the etymology of its items then not only illuminates the development of languages but also contributes to our understanding of the culture and social situation of their speakers.

See also: Evolution and Language: Overview; Historical Linguistics: Overview; Lexical Semantics; Lexicology and Lexicography; Lexicon; Sociolinguistics

Bibliography

Diez F 1858 *Etymologiches Wörterbuch der romanischen Sprachen*. Marcus, Bonn, Germany

Fick R 1871 *Vergleichendes Wörterbuch der indogermanischen Sprachen*. Vandenhoeck and Ruprecht, Göttingen, Germany

Grassmann H 1863 Über die Aspiraten und ihr gleichzeitiges Vorhandensein im An- und Auslaute der Wurzeln. *Zeitschrift für vergleichende Sprachforschung* **12**: 81–138

Kluge F 1989 *Etymologisches Wörterbuch der deutschen Sprache*, 22nd edn. Walter de Ceryter, Berlin

Malkiel Y 1976 *Etymological Dictionaries. A Tentative Typology*. University of Chicago Press, Chicago

Malkiel Y 1993 *Etymology* Cambridge University Press, Cambridge, UK

Pott A 1833–6 *Etymologische Forschungen auf dem Gebiete der Indogermanischen Sprachen*. Meyer, Lemgo, Gemany

Schleicher A 1862 *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*, 2 Vols. Böhlau, Weimar, Germany

Skeat W W 1909 *An Etymological Dictionary of the English Language* (Revised Edition). Clarendon, Oxford

Verner K 1875 Eine Ausnahme zur ersten Lautverschiebung. *Zeitschrift für vergleichende Sprachforschung* **23**: 97–130

W. P. Lehmann

Eudaemonism

Following up on the ancient concepts of *eudaimonia* (lit. 'having a good demon'), the notion of 'happiness' has been given a great variety of different meanings. Its semantic space reaches from 'the favor of fate, the advantageous lot, the happy circumstance,' all referring to accidental, external circumstances, to completely achieved goals and the 'good life.' Across this breadth, thus, the term happiness includes that which is outside of the reach of human beings, individual destiny (as in the ancient figure of *Fortuna*, the goddess of fate, who distributes her offerings blindly), as well as a durable state that can be influenced and, in principle, achieved by human action. Precisely the determination of such action, guided by the pursuit of happiness, has long been the basis of ethics and moral philosophy and of their debates about action guided by virtue, which then in turn was to determine and safeguard happiness.

Eudaimonia (*Glückseligkeit*, *prospérité*, happiness) is thus one of the central concepts of ethics and moral philosophy. It has been cast in a variety of ways and points into a number of different directions. Happiness was, for once, conceptualized in subjective terms, with reference to the self (see *Self: History of the Concept*), to the passions and affects, to the realization of one's objectives, or of one's overall life. Second, happiness can be cast in relation to the other, as friendship, love, recognition, as ethically founded intersubjectivity. And finally, there is a conception of happiness that takes all other conceptions for incomplete as long as there is no connection to the community and its political order. Even though this conception found its strongest expression in ancient philosophy, it has by far not disappeared from current discourse. In contrast, it may possibly even be claimed that this conception with its emphasis on justice as a precondition for happiness and the 'good life' presents, just as much as the reference to the self and to the other, an inescapable point of concern and a 'utopian' horizon of the philosophy of happiness.

In sum, then, the conceptualization of happiness shows a temporal tension between fulfillment across all of the life-span, on the one hand, and fulfillment in the sudden and epiphanic moment, on the other. Further, it is marked by a transcendental-religious tension that stretches from revelation, redemption and salvation to joy. It carries with it a political-social-ethical tension between an individualistic view of self-realization and the conception of a good political order. Finally, a tension with regard to a theory of action is unavoidable, a tension that situates the singular human being between the capability to act autonomously and destiny. All these tensions can be traced back, in a variety of ways, to the heritage of Greek reason.

1. *The 'Classical' Tradition*

In the Nicomachean Ethics, engaging with Plato's conception (Wolf 1996), Aristotle links the concept of *eudaimonia* to the three human life-forms: to pleasure, from which most human beings gain their view about what happiness is; to the service of the state; and to life dedicated to philosophy. Because the essence of human life lies in the pursuit of reason, neither profane enjoyment nor politics and the art of governing, but *theorein* is the highest in the hierarchy of human activities, in which *eudaimonia* can be realized. 'Happiness is activity (*energeia*) in accordance with virtue' (X, 7) and 'the highest good reachable for human beings presents itself as the activity of the soul in the sense of its essential prowess (*areté*)' (Aristotle 1934, I, 6, see Heidegger 1992). *Theorein* as an activity is not aiming at the cognition of just anything, but of that which is eternal. As such it is among the human activities the one that is the most persistent of all, independent of changes in the external circumstances. This activity is full of pleasure, it is so even most among all activities—on the one hand, because it is continuous; on the other, because it participates in the eternal that it gets to know. *Theorein* is not confined to any moment; such activity stretches across an entire life. Furthermore, to be active in the sense of *theorein* is a purpose for itself, the final purpose, that which is pursued for its own sake, since it does not produce anything that lies outside of itself; it is self-sufficient. Such autarchy, however, is possible only if the polis has provided the required means so that human activity can develop free of the concern for survival. For Aristotle, happiness is an eminently social and political matter, since under conditions of lack of resources no accomplished life is possible, and it is also not thinkable without friendship, without the relations to others. Thus, happiness is finally possible only within a just political community that is organized on principles of reason.

Aristotle, thus, linked happiness to human essence in need to be actualized, to the relation to others in friendship and to institutionally supported forms of living and acting together. With the disruption of the order of the polis as well as of Hellenic philosophy, in contrast, for the Stoa and for Epicureism (and as well for skepticism), *eudaimonia* becomes a concept that refers to purposes that now reside in that which is available to human beings. *Eudaimonia* is, then, the striving for whatever is available and, vice versa, emphasizes the diminution of demands and the avoidance of pain. The thought that happiness opens itself

towards the humans only in a Stoic relation of radical distance to all that which is not at hand is alien to Epikur; nevertheless he, too, sees happiness neither in *theorein* nor in a goal-directed activity that finds its fulfillment in ethical and theoretical practice. Despite all differences between the ancient Greek schools of thought, they also show commonalities, such as in particular in the connection of virtue and happiness. Ancient Greek philosophies, as Kant stated, follow the same method in rejecting any validity to the assertion that virtue and happiness could be two separate elements of the highest good. They 'consequently sought the unity of the principle in accordance with the rule of identity,' even if they differ in the choice of the fundamental concept. 'The Epicurean said: to be conscious of one's maxim leading to happiness is virtue; the Stoic said: to be conscious of one's virtue is happiness. For the first *prudence* was equivalent to morality; for the second, who chose a higher designation for virtue, *morality* alone was true wisdom' (Kant 1993a, A 200, pp. 239–40/book II, chapter 2, p. 229).

2. *Eudaimonia in the Philosophy of Early Modernity*

Conceptions of happiness in relation to individuals and to society changed profoundly during the period known as early modernity; these changes were part of broader reproblematizations in the moral and social philosophy of the time. Political thinking of this period introduces happiness as a component and as a means of politics and the symmetrical opposition of passion and reason enters into a conceptual relation with virtue and politics. Happiness becomes one of the key concepts of ethical-political thinking during the seventeenth and eighteenth century, a thinking that at the same time reorders the relations between the concepts for affects and those for morality. No longer the affective drives of the human beings are the object under consideration, but rather human actions and the consequences of these actions. Affects as such are considered neither as good nor evil, it is rather individual selfishness and self-love (*amour-propre*) which moves into the center of concern. This focus generates the new problématique of moral philosophy: If it is the case that human beings pursue their happiness on the basis of their drives and affects, then it is the task of moral philosophy to determine how—despite that—a good social order can exist and be maintained.

The new emphasis on the individuals and their interests went along with a reconceptualization of the political. From now on, it was precisely their striving for happiness that made human beings inclined to subject themselves to the laws of the common order. This reasoning emerges in classical form in Thomas Hobbes's (1588–1679) *Leviathan*. Hobbes first determines 'felicity' as the attempt to completely accomplish the objectives of human desires. Since, however, all human beings hope to reach this goal, the one potentially turns into a danger for the other. To avoid such a situation of *bellum omnium contra omnes*, of 'war of every man against every man,' human beings assemble to form a state that itself embodies no longer any conception of happiness or of the good life, but that guards to prevent any situation from arising in which the various strivings for happiness collide in a violent manner.

At the end of this period, during which happiness remains the focal point of theoretical constellations, an attempt is made to indeed erect a polity that is committed to support its members' striving for happiness. The American Declaration of Independence even declares the 'pursuit of happiness' to be a human right (Jefferson 1776). A historically novel compromise is here elaborated between a classic-substantive and an individualistic-instrumental conception of the state: the state respects the autonomy of the individual and the plurality of their views, but does not entirely renounce the proposition of a general guiding idea about the 'good life.' The individuals' striving for happiness—to the protection of which the state has been created, exactly in the sense of the theories of delegation—in its sum produces the common good. This common good, however, is now in a particular way mixed with elements of a Protestant ethic, in which inner-worldly happiness, for example, as economic success, gains divine approval. This unique privileging of the pursuit of happiness as a human right, which cannot be found in European declarations, is rooted in a particular relation between the state and the citizens. It is nevertheless close to European conceptions in another respect, namely in understanding the citizen predominantly as an economic subject, a *bourgeois* in contrast to a *citoyen*, whose pursuit of property needs to be protected by the state, a protection that is considered as being the primary task of the state.

The utilitarianism of the eighteenth and nineteenth century similarly still places happiness in the center of moral and social philosophy (see *Utilitarian Social Thought, History of*). The ethics of utilitarianism considers individual—as well as general—utility as the purpose of action and sets thus 'good' synonymous with 'useful.' Jeremy Bentham (1748–1832) calls for 'the greatest happiness of the greatest number' as the collective goal; by promoting the well-being of all, the well-being of the individual is promoted as well. Morals and legislation are that art to regulate human actions such that to bring forth 'the greatest possible quantity of happiness' (*An Introduction to the Principles of Morals and Legislation* 1789). Drawing on Bentham, John Stuart Mill (1806–73) considers the pursuit of the greatest possible happiness of all as the supreme goal as well (*Utilitarianism* 1861). On the one hand, utilitarism maintained a conception according

to which the pursuit of happiness is the most important motive of human action, and it also kept the link between happiness and politics. What was added were amendments that had become historically unavoidable, namely the view that every human being determines their happiness in their own way and strive for it, and that everybody has the same right to do so. On the other hand, however, what was thus created was an *aporia* between the common good and the diversity of the human pursuits of happiness. This *aporia* was to be dealt with by assuming the comparability of the various individualized and subjectivized pursuits. Measurability and quantification were the tools to make the diverse compatible, since the state's task to pursue the greatest happiness of the greatest number demanded the possibility of adding up the manifoldness of forms of individual happiness. Quantification of happiness is the attempt to maintain a measure for collective happiness by at the same time making only a minimum of substantive assumptions about happiness as such. Utilitarianism had inevitably to fail when radically subjectivized conceptions of happiness were to be compared to each other, measured and sublated in an overall concept of utility.

3. *The Eviction of Happiness from Philosophy: Immanuel Kant*

In continental moral and political philosophy, the most systematic attempt to clarify the relation between action and happiness was undertaken by Immanuel Kant—it was at the same time an attempt with far-reaching consequences for later philosophy. It had long been an objective of moral philosophy to demonstrate that there is no, and could not possibly be any, contradiction between ethical demands on the one side, and the pursuit of happiness (or of utility or interest) on the other. Thus, it had merged virtue and happiness into one, and stated that virtuous action promotes happiness at the same time. Kant radically breaks this connection of interest, virtue, and happiness.

Happiness remains for Kant as well a concept that has a basic influence on human action: 'Happiness is "the satisfaction of all our desires"; *extensive*, in regard to their multiplicity; *intensive*, in regard to their degree; and *protensive*, in regard to their duration' (1974a, II, B 834, p. 677 [p. 518]), '*happiness* is the state of a rational being in the world in the whole of whose existence *everything goes according to his wish and will*, and rests, therefore, on the harmony of nature with his whole end as well as with the essential determining ground of his will' (1993a, A 224–5). This concept, however, is not suitable as a foundation of moral philosophy, according to the following reasoning.

First, happiness can hardly be the true purpose of nature, since in such a case, human action would not be guided by practical reason, steering the will, but subjected to the instincts (1993b, BA 5, p. 20). Second, the concept of happiness cannot be determined unequivocally, since it rests on experience, which is never closed, on the one hand (1993b, BA 46, p. 47, 1993a, A 46, p. 133), and since, on the other hand, the 'idea of happiness' requires an absolute whole, a present and future maximum of well-being. The contradiction between the principled openness of experience and the idea of the accomplishment of happiness cannot be sublated within time. Furthermore, the conception of a reachable maximum of happiness is impossible, for two reasons. No sum of happiness can ever be added up, since only similar perceptions can be added up; feeling, however, is highly variable according to the manifoldness of being touched in the very intricate state of life. In addition, the nature of the human being can never halt and be satisfied in possession and enjoyment. The inclinations change, they grow, and they always leave an even larger emptiness behind them than the one one hoped to fill. (The dynamics of needs and their limitless character (*apeiron*) was already of concern for Hellenic philosophy.) The fact that each human being may have a specific view of happiness, and which furthermore may undergo change in the course of life, puts a further limit to any general determination of happiness. Happiness is 'a subjective basic feeling of enjoyment of dissatisfaction' (1993a, A 46, p. 133); a subjective feeling however, cannot be the foundation of general rules for action. As a consequence, thus, it is not possible to demand the agreement of the pursuit of happiness with moral requirements—which precisely was the former assumption: that namely the pursuit of happiness realizes ethical requirements since happiness resides only in morally virtuous action.

Such subjectivization separates happiness radicality from virtue and ethicality and subordinates it, as an indeterminate inclination, to the categorical imperative. 'Act in such a way that the maxim of your will could always serve as the principle of a general law.' It is no longer the pursuit of happiness that makes an action moral, but only the respect of the moral law and the obedience to duty, since it is only in them that human freedom is anchored. The categorical imperative refers the concept of ethicality back to 'the idea of freedom' (1993b, BA 102–3, p. 84). Free will is nothing but 'autonomy, that is, the property of the will to be its own law' (1993b, BA 97–8, p. 81). Happiness is accidental and not at our disposal, but freedom is effective 'independent of external causes that determine it' (1993b, BA 97–8, p. 81, BA 99–100, pp. 82–3). A human being does not act morally because he or she hopes for happiness—'If *eudaemonia* (the principle of happiness) instead of *eleutheronomia* (the principle of freedom for inner legislation) is set as the foundation, the consequence will be euthanasia (the soft death) of all morality' (1997, AIX, p. 506)—but because he or she follows the principle of freedom of inner legislation. This entails that morality has been

robbed of the foundation on which it could ground itself, that is, the hope for the remuneration for moral action that would arrive at some point. And by implication the pursuit of happiness loses the justification that was derived from its link to morality. Simmel (1989, p. 253) called this the 'mutual independence of ethicality and happiness.' 'If our ethical action,' Simmel wrote, 'were only a detour on the path to happiness, then the point of our freedom would show itself again interwoven with the dependence on the powers of being, without whose favors no happiness could be complete.'

The expulsion of the concept of happiness from the reach of moral philosophy has its counterpart in the eviction of happiness out of the realm of the political. For Kant, freedom has taken the place of happiness as the goal of politics—a concept driven further by romanticism, although setting out from the non-*eudaimonistic* ways of reasoning in transcendental philosophy. Kant's categorical imperative and the ideal of autonomy inherent in it expel the concept of *eudaimonia* from the main lines of philosophical discussion.

4. *Redemption, Happiness, and Political Order*

The emerging social sciences have started out from this rupture and have largely accepted it. In his theory of religiosity and morality, Durkheim re-interprets happiness and recognizes it only in religious experience, in moments of 'collective effervescence,' of self-transcendence and loss of self. Happiness is seen as the experience of being-one-with-oneself or of being submerged in the world, as a 'feeling of strengthening' (Durkheim 1981, p. 291), as the subjective 'oceanic feeling,' which Freud (1994) considers without any comprehension. Simmel, too, considers spiritual salvation (*Seelenheil*) as a state in which complete harmony with the world has been achieved, a state that shows the full liveliness and inner touchability of the perceived, the immediate being (Simmel 1995). 'Happiness, as a relation of transcendence, thus has two fundamental forms. It exists as the experience of either *oneness* or *otherness*, as the dissolution of all difference, or as the apprehension of, and unconditional surrender to, absolute difference' (Ferguson 1992, pp. 11–12). Such experience turns into an articulation of social integration, which at the same time rests on individual feeling and, again, connects the individual with the other and with a 'holy' collectivity, the ideal community of justice.

Within the main line of critical theorizing, that is, in Theodor W. Adorno and Max Horkheimer, the subjectivization of happiness is further radicalized. The foundational character of the aporias that result from the dualism of individual and society is rejected. Instead, the joint attack of liberalism (*bourgeois* ideology) and capitalism is held responsible for the destruction of the link between freedom and happiness.

Happiness is privatized in 'administered society' and under the reign of 'culture industry,' and it becomes the expression of societally determined false consciousness. Due to the horror of the circumstances, individual happiness becomes nothing but the dreadful arrangement of a petty-bourgeois idyll with its swollen sentiment. The *bourgeois* promise of happiness arising from property and possession—already denounced by the romantic critique of the revolution as the abandoning of the shaping of the world according to the ideas of reason in favor of the demands for material happiness on the part of the bourgeoisie—becomes under conditions of inhumanity a curse. The promise of the commodity turns into pale appearance; in the totality of objectivizable socioeconomic relations, the connection between the individual, happiness, and society is interrupted and the promise of happiness only admitted as aesthetic experience (Adorno). But, within the same critical tradition, and by appropriating critically Freud's reflections on the pleasure principle and the reality principle, Herbert Marcuse (1955/1966) aims at developing a social theory that elaborates on a conception of possible happiness and permits the critique of both, the societal circumstances as well as the empirically existing conceptions of happiness. 'The concept (of happiness) denotes a more-than-private, more-than-subjective condition; happiness is not in the mere feeling of satisfaction but in the reality of freedom and satisfaction' (Marcuse 1955/1966, pp. 103–4). The concept of happiness has to be liberated from bourgeois conformism and relativism; even the singular human beings cannot be 'judge of their own happiness' under these circumstances. It is precisely the task of critical analysis to show the 'possibility of a happier real constitution of humanity' beyond the existing 'false happiness.'

In another line of critical thinking, Walter Benjamin turns political time, directed towards happiness, into the key element of his philosophy of history. 'The image of happiness … is tainted by time through and through … In other words, the imagination of happiness resonates inalienably with the one of redemption' (1991a, p. 693). Happiness is here not in the (past) event, but in its possibility, to which a 'weak messianic power' is given; it is recognizable only in its being missed historically (1991a, p. 694). 'The order of the profane has to erect itself with the support of the idea of happiness' (1991b, p. 203)—and in the *Angelus Novus*, Benjamin finds the image for history in which the past, that which is complete (misfortune and misery), and the future, that which has remained incomplete (happiness), get reverted and in which misfortune and misery turn into something incomplete hoping for redemption (1991c, pp. 588–9). Ernst Bloch, too, reinstates the old connection between the individual, passion and society; happiness becomes

again an inalienable part of human being. 'Because hope's grounding in the human drive for happiness is too difficult to be destroyed; and too evidently has it always been the engine of history' (1978, I, p. 515).

Happiness, thus, even though pushed into the private realm since Kant, persistently returns into the political realm—as 'redemption,' 'the principle of hope,' as 'emancipation,' 'undamaged life' and, in the version of existentialist philosophy, as 'authenticity.'

During the past few years, happiness has not only become the object of dubious empirical-quantitative research on levels of satisfaction, but also again a theme of social philosophy, partly by referring to classical liberalism, partly also by returning to ancient traditions. Under terms such as justice, the good life, and self-realization, the question about the societal conditions under which happiness can unfold is again raised. The attempt is, thus, renewed to locate happiness not only in an 'accomplished' relation of the individual to her- or himself, but also in relation to others and to a just societal order.

Hannah Arendt already held a view of political happiness, of 'public happiness,' that could not be equated with the obligation of the state to care for the welfare of its citizens (such a view of *bonheur*, in contrast, is, in her view, coresponsible for the loss of freedom in the course of the French Revolution). In the American Revolution, however, the connection between 'public happiness and political freedom,' which is an 'inalienable part of the structure of a republican community,' could be maintained at least for some period (Arendt 1994, p. 178). The new connection between happiness and freedom made here occurs in the sign of plurality and a public sphere as guardians against both dangers; that namely substantive conceptions of happiness could suffocate freedom or that abstract and atomized freedom empties out any conception of happiness at all.

Drawing on the classical contract theories by John Locke and Jean-Jacques Rousseau and situating himself—often in a rather decontextualized way—in relation to Immanuel Kant, John Rawls develops a 'theory of justice' (1971) for modern, pluralistic and democratically constituted societies. For Rawls's theory, the 'right' enjoys primacy over the 'good,' since value-pluralistic societies will be able to agree only over a minimum set of foundations for what is right, but not over what is good. Rawls, thus, renounces any attempt to formulate a substantive theory of the good life. The starting point of his considerations is the hypothetical construction (the 'intuitive thought') of a society as a fair system of cooperation between free and equal persons. Since these considerations provide the general political framework for the 'good life as justice and fairness,' a human being can for Rawls be called 'happy … during those periods when he is successfully carrying through a rational plan and he is with reason confident that his efforts will come to fruition' (Rawls 1971, p. 550). Rawls ties his theory of justice to a critique of utilitarianism. For him the well-being of people does not depend on the degree to which individual preferences are realized. It is important only that people are provided with the basic means to the same amount. As Martha Nussbaum, however, points out, Rawls did not consider what people do with these goods, to what they are enabled, and whether the space of enablements that thus opens up is not marked by striking inequalities.

Nussbaum (1990, 1993) herself resorts to Aristotle to sketch those political foundations that are to enable all people to lead a good life. She does not advocate any liberal relativism, but rather a universalism, based on an anthropological conception, that refers to needs and capacities which all people across all divides of culture, class, and sex have in common. Such basic needs constitute moral demands; they are an inalienable part of ethics and moral philosophy. At the same time, Nussbaum's perspective aims at pushing back the primacy of reason in moral philosophy, since moral principles had hitherto mostly been anchored in human reason. Her ethics of virtue attempts to counteract the insufficient regard for the particularity of situations in moral theories that are based on general principles. In addition, it aims at integrating emotional dimensions into ethics. To define the concept of virtue more precisely, and not merely to understand it by recourse to rules, a 'superior instance' needs to be found; for Nussbaum this is a theory of the good. Starting out from the 'basic structure of the human form of life' (mortality, corporeality, cognitive abilities, practical reason, sociality, etc.), she develops an 'open list' of the basic capacities of human beings and of the good. She herself characterizes the list as 'vague,' which means that it is open for concretization in each specific cultural context, and then puts specific demands towards politics. The objective of measures taken by the state is that they enable the citizens to choose a good form of leading their lives, to develop their capacities and to live in accordance with practical reason and with their social ties.

In his attempt at clarifying the relations between a moral life and a good life, Seel (1995) similarly defends a 'universalism of the good,' which is not linked to particular forms of society. This universalism, however, is no essentialism that claims the existence of values independent of human volition or raises anthropological facts to the status of values or norms. Seel develops a position that he calls 'reflexive *eudaimonism*' and asks what it entails, from the viewpoint of different people, 'to sensibly lead a succeeding life' (p. 10), defining a 'good life' formally on the basis of moral respect, social recognition and political legitimacy. The incarnation of good human life is a 'life, open to the world and self-determined, the moral and legal protection of which includes a respect for all individuals who are able to have a (however determined) good life' (p. 11).

'Happiness' and 'the good life' are, thus, always also tied to sociocultural circumstances. It remains the task of philosophy as well as of the social sciences to think the—political—conditions under which a good life and happiness stay open as a possibility. Happiness as the relation to oneself, to others, and to the community belongs then to reflection, to experience, and to the dream (Mauzi 1969).

See also: Arendt, Hannah (1906–75); Aristotelian Social Thought; Aristotle (384–322 BC); Bloch, Mark Léopold Benjamin (1886–1944); Critical Theory: Frankfurt School; Existential Social Theory; Hobbes, Thomas (1588–1679); Justice: Philosophical Aspects; Kant, Immanuel (1724–1804); Kantian Ethics and Politics; Morality (East and West): Cultural Concerns; Philosophy, Psychiatry, and Psychology; Psychoanalysis: Overview; Self: History of the Concept; Utilitarian Social Thought, History of; Virtue Ethics; Well-being (Subjective), Psychology of

Bibliography

Arendt H 1963/1994 *On Revolution*. Penguin, Harmondsworth, UK

Aristotle 1934 *The Nicomachean Ethics*. Harvard University Press, Cambridge, MA

Benjamin W 1991a Über den Begriff der Geschichte. In: Tiedemann R, Schweppenhäuser H (eds.) *Gesammelte Schriften*. Suhrkamp, Frankfurt am Main, Germany, Vol. I2, pp. 691–704

Benjamin W 1991b Theologisch-politisches Fragment. In: Tiedemann R, Schweppenhäuser H (eds.) *Gesammelte Schriften*. Suhrkamp, Frankfurt am Main, Germany, Vol. I2, pp. 203–4

Benjamin W 1991c Das passagen-werk. In: Tiedemann R, Schweppenhäuser H (eds.) *Gesammelte Schriften*. Suhrkamp, Frankfurt am Main, Germany, Vol. V(1–2)

Bloch E 1978 *Das Prinzip Hoffnung*. Suhrkamp, Frankfurt am Main, Germany, Vols. 1–2

Durkheim E 1912/1981 *Les formes élémentaires de la vie réligieuse*. Alcan, Paris

Ferguson H 1992 *Religious Transformation in Western Society: The End of Happiness*. Routledge, London

Freud S 1994 *Das Unbehagen in der Kultur*. Fischer, Frankfurt am Main, Germany

Heidegger M 1992 Platon: Sophistes. In: *Gesamtausgabe*, Vol. 19. Klostermann, Frankfurt am Main, Germany [*Plato's Sophist*. Rojcewicz R, Schuwer A trans. 1997 Indiana University Press, Bloomington, IN]

Kant I 1781/1974a Kritik der reinen vernunft. In: Weischedel W (ed.) *Werkausgabe*. Suhrkamp, Frankfurt am Main, Germany, Vols. III/IV [*Critique of Pure Reason*, Politis V (ed.) trans. based on Meiklejohn 1993 Everyman, London]

Kant I 1790/1974b Kritik der Urteilskraft. In: Weischedel W (ed.) Werkausgabe. Suhrkamp, Frankfurt am Main, Germany, Vol. X

Kant I 1788A/1993a Kritik der praktischen Vernunft. In: Weischedel W (ed.) *Werkausgabe*. Suhrkamp, Frankfurt am Main, Germany, vol. VII, pp. 107–302

Kant I 1785A/1993b Grundlegung zur Metaphysik der Sitten. In: Weischedel W (ed.) *Werkausgabe*. Suhrkamp, Frankfurt am Main, Germany, Vol. VII, pp. 11–102

Kant I 1798/1997 Die Metaphysik der Sitten. In: Weischedel W (ed.) *Werkausgabe*. Suhrkamp, Frankfurt am Main, Germany, Vol. VIII

Marcuse H 1955/1966 *Eros and Civilization*. Beacon Press, Boston

Mauzi R 1960/1969 *L'idée du bonheur dans la littérature et la pensée françaises au XVIIIe siècle*. Colin, Paris

Nussbaum M C 1990 Aristotelian social democracy. In: Douglas R B, Mara G, Richardson H (eds.) *Liberalism and the Good*. Routledge, New York, pp. 203–52

Nussbaum M C 1993 Non-relative virtues: An Aristotelian approach. In: Nussbaum M, Sen A (eds.) *The Quality of Life*. Clarendon Press, Oxford, UK

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Harvard, MA

Seel M 1995 *Versuch über die Form des Glücks. Studien zur Ethik*. Suhrkamp, Frankfurt am Main, Germany

Simmel G 1989 Die Lehre Kants von Pflicht und Glück. In: Rammstedt O (ed.) *Gesamtausgabe*. Suhrkamp, Frankfurt am Main, Germany, Vol. 7, pp. 247–54

Simmel G 1995 Die Religion. In: Rammstedt O (ed.) *Gesamtausgabe*. Suhrkamp, Frankfurt am Main, Germany, Vol. 40, pp. 39–118

Wolf U 1996 *Die Suche nach dem guten Leben. Platons Frühdialoge*. Rowohlt, Reinbek, Germany

H. Friese

Eugenics as an International Movement

The term eugenics, derived from the Greek *ευγενεσ* (eugenes) was first coined by English mathematician and geographer Francis Galton (1822–1911). In his book *Inquiries into Human Faculty and Its Development* eugenics refers to one born 'good' in stock, hereditarily endowed with noble qualities (Galton 1883, p. 24). Eugenics was the human counterpart of scientific animal and plant breeding.

In his book, Galton lays out all the dimensions that came to characterize eugenics as an ideology and social/political movement during the first half of the twentieth century. These were as follows. (a) Adherence to the methods of selective breeding as applied to the improvement of the human species. (b) A strong belief in the power of heredity in determining physical, physiological, and mental (including personality) traits. (c) An inherent ethnocentrism and racism that included belief in the inferiority of some races or groups and superiority of others (a view extended to ethnic groups and social classes as well). (d) Belief in the power of science, rationally employed, to solve social problems, including ones so seemingly intractable as pauperism, crime, violence, urban decay,

prostitution, alcoholism, and various forms of mental disease, including manic depression and 'feeblemindedness' (retardation).

It took some time after Galton first propounded idea of eugenics for eugenics organizations and activities to emerge around the world, but by the time of World War I (1914) eugenics had become a truly international movement. Thus, the standard historical view of eugenics as a largely British, American, and German phenomenon is inaccurate. Eugenics developed in a variety of national contexts with a wide range of ideological and political programs (Adams 1990, pp. 4–6 and 217–31). The earliest eugenics movements were founded in Germany in 1904, in Britain in 1907, and in the United States in 1908–1910. Other eugenics movements appeared around the world shortly thereafter: in Western Europe between 1910 and 1920: France, Spain, the Scandinavian countries, Finland, Russia, Latin America (Cuba, Brazil, Mexico), in Canada, Africa (particularly South Africa), and Asia (China and Japan). While eugenics clearly originated in its modern form, that is, viewed as a branch of genetics, in the west, its spread around the globe gave it an international flavor that retained some common core of scientific and political beliefs, while developing considerable local differences. Eugenicists themselves were consciously internationalists, arguing that like all science, eugenics knew no national boundaries. Despite this international profile, however, it was in western countries such as the United Sates and, of course, Germany under the Nazi regime (1933–1945), that eugenics as an intellectual and social movement reached its greatest strides and from eugenicists' point of view, enjoyed its greatest ideological and political success.

The international character of eugenics is reflected in the three major international congresses, at which eugenics enthusiasts gathered literally from around the world. The first was held in London in 1912, while the second and third were held in New York (the third congress was held at the American Museum of Natural History) just following the sixth International Congress of Genetics, in 1932). Papers and photographs of exhibits from the second and third congresses fill several impressive volumes.

Although eugenics is often thought of as a politically conservative movement, globally its supporters came from widely divergent political backgrounds. Some, like Charles B. Davenport, Harry H. Laughlin, and Madison Grant in the US, and Eugen Fischer in Germany, were reactionary, while others, such as Britain's J. B. S. Haldane, the United States's H. J. Muller, or Russia's Alexander Serebrovsby, were influential left-wing eugenicists who held socialist or even communist views (Paul 1984, p. 567). Left-wing eugenicists were not necessarily egalitarians, however, and few envisioned a society in which all people were truly 'equal.' Their model was a meritocracy based on inherited biological worth.

1. *Background: Degeneration Theories and the Rise of Eugenics (1890–1914)*

In both the west and east the late nineteenth century saw, for different reasons, a widespread belief that the human species was in various phases of deterioration and decline (Paul 1995). Much of this pervasive pessimism came from observations in the social and political arenas. British eugenicists were particularly concerned with the high fecundity and presumed inherited mental degeneracy of the urban working class, particularly those labeled as 'paupers' or (in Britain) the 'residuum.' By contrast, American eugenicists were more concerned with the number of feebleminded who appeared to be filling the prisons and insane asylums in the wake of urbanization and industrialization after 1865 (the US Civil War) and with the genetic deficiencies of the 'new immigrants' around the turn of the century. In Germany various racial and ethnic groups were thought to harbor an undue proportion of mentally ill, psychotic, psychopathic, and psychiatric members. In France and Brazil where ideas of the inheritance of acquired characteristics held more sway than in most other countries, public health measures, rather than reliance on selective breeding, were seen as the most effective means to combat degeneracy. In Asia, general theories of degeneration spread with the encroaching influence of western imperialism from the 1860s onward. The inability of the Asian populations to ward off imperial penetration or intra-Asian rivalries was seen as partly a result of scientific and technological backwardness and of physical and cultural weakness in the face of the western onslaught.

2. *Eugenics and Scientism*

Eugenics throughout the world was viewed as the application of rational, and 'modern' scientific principles to the solution of age-old social problems. Eugenicists and many of their supporters had a high regard for science and technology, symbols in the western countries of the triumph of enlightenment and rationality, and of modern capitalist control over and exploitation of, nature for human gain. The goal of science was to predict and control, an ideology extended to society at large. The period in which eugenics developed was globally one of great economic, social, and political instability: two major world wars, the dissolution of the great empires of Europe in the wake of rebellions abroad, rapid industrialization in the west, especially the growth of heavy industry, and urbanization, all of which contributed to the rise in such problems as alcoholism, prostitution, crime, mental illness, and social unrest. In the major western countries organized labor gained great strength, leading to increased and bloody conflicts in the period

1890–1940. Communism made major gains during the period, with the Bolshevik Revolution in Russia, the organization of Workers' Councils that took over several cities in Germany after the Armistice in 1918; and with the expansion of such groups in the United States as the International Workers of the World (the IWW, or 'Wobblies') who carried out a successful general strike in Seattle and the northwest in 1919.

Outside of Europe, similar, sometimes related, historical events led to similar responses. Latin America underwent convulsive economic and social change in the wake of the decline of Spanish and rise of North American colonialism. In Asia, the struggle in China and Japan toward 'modernization,' inter-imperial rivalries culminating in the Sino–Japanese War of 1894–95 and the Russo–Japanese war of 1905, all stemmed from growing expansionism related to trade, and control of raw materials and markets between east and west and within rising industrial powers, like Japan, in Asia itself. As in the west, science and technology in Asia were associated with modernism and economic advancement.

As a supposed new 'science,' eugenics meshed well with modernist ideology. Eugenics was based on what was thought to be the cutting edge research in a new science (genetics), was mathematical, experimental, rigorous, and predictive, and purported to apply the principles of science to the solution of previously intractable social problems. Worldwide, it provided an incentive for academically-trained practioners (geneticists) to join the new managerial class that was beginning to apply highly-trained expertise to everything from organizing factory production lines, offices, and schools, to 'scientific' agriculture and not least, human reproduction. Applied to society, eugenics became part of the ideology of 'social engineering' that was replacing the older notion of *laissez-faire* associated with social Darwinism. It was not enough to let nature take its course: scientifically-trained social engineers had to step in and direct that change along national lines. Eugenically, this change involved the guidance of human reproduction to ensure that only those with good heredity passed on their genes to future generations. Eugenics was viewed as the 'control of human evolution.' Eugenicists argued that to control heredity at the source was to be efficient—to prevent problems before they occurred, rather than let them happen and then deal with the consequences. In various forms—either as proposed breeding and sterilization programs, or maternal and other public health proposals, the efficiency argument dominated much of eugenic social rhetoric. In the United States this collage of views and ideas was swept up into a somewhat heterogeneous movement known as Progressivism. Progressives advocated the use of science to solve social problems, and embraced the efficiency movement of Frederick Winslow Taylor (1856–1915), which had been developed in the United States initially in an industrial context. Similar strains and arguments appeared in different forms in eugenics movements around the world.

Since other articles focus on details of various eugenics movements around the world (see *Eugenics in Europe*; *Eugenics as the Basis of Population Policy*; *Eugenics, History of*; *Eugenics in Asia*) this article will focus on the general nature and extent of international eugenics activities, comparing the similarities and differences among the various national movements.

3. *British and American Eugenics, 1880–1940*

British and American eugenics shared a number of intellectual and social concerns, as well as personal contacts. The major importer of eugenics to the United States, Charles B. Davenport (1866–1944), spent part of the year 1899–1900 with Galton and his protégé, Karl Pearson (1857–1935) at University College, London, learning statistics and biometrics and gaining initiation into Galton's 'secular religion of eugenics.' American eugenics thus grew directly out of the vision that Galton, and later Pearson, developed. At the same time, American eugenics developed its own particular twists, especially its early acceptance of particulate, discontinuous and Mendelian concepts of heredity, a relationship not shared by their British counterparts until the later 1920s.

Mendelism notwithstanding, American and British eugenicists shared the basic assumption that such traits as alcoholism, criminality, physical and mental 'defects,' and intelligence, were all biologically inherited, with the worse forms concentrated in the lower classes and among certain races. Horrified by the significantly higher birth-rate of the lower classes, eugenicists in the United States and Britain argued that the observed decline in productivity and 'national efficiency' was due to the increasing number of less-productive social defectives that were being born and were having to be cared for by the state. Eugenics was a scientific approach to improving social efficiency. In Britain, Pearson's largely theoretical orientation to eugenics led him to warn eugenicists to avoid politics. Partly as a result, British eugenicists had much less legislative success than their counterparts in the United States.

In several ways the American movement followed a different trajectory from the movement in Britain: (a) American eugenicists became strong advocates of psychometrics, the burgeoning new school of I.Q. testing, which took it for granted such tests measured an innate, unchanging mental function. While British eugenicists also assumed that mental ability was inherited, they did not employ the I.Q. (or other similar) tests to the same extent as in the United States; (b) Leaders of American eugenics were from the start anxious to play a role in the public arena—in what today would be called formation of public policy.

Eugenicists such as Harry H. Laughlin (1880–1943), Superintendent of the Eugenics Research Office at Cold Spring Harbor, NY translated eugenic principles into social action. Eugenicists lobbied for passage of the Johnson–Reed Act (Immigration Restriction Act) in the United Sates congress in 1924, while Laughlin designed a 'Model Sterilization Law' and eugenicists lobbied for the passage of eugenical sterilization laws in a number of states during the 1920s and 1930s. The first eugenical sterilization laws were passed in the United States (the first was Indiana, 1907) though by the mid-1930s such laws were in place in a number of western countries (Norway, Sweden, Finland, Germany, and Canada). Between 1907 and 1965, some 60,000 people in the United States—mostly in state institutions for the feeble-minded, psychiatric hospitals, and prisons—were sterilized. Like their British counterparts, American eugenicists were highly motivated by the ideas of national efficiency and rational planning (Vecoli 1960, Allen 1986, pp. 255–60).

4. *Eugenics on the Continent: Germany, Scandinavia, and France*

On the continent, eugenics in Germany and Scandinavia shared many features in common, while that in France took its own, unique direction. The German and Scandinavian movements were based strongly on what has been termed notions of 'hard heredity' (i.e., that most mental and moral attributes are rigidly determined by genes at conception), and employed basic concepts of Mendelian genetics. While the ideology of Nordic and Aryan superiority was present in both movements, it represented only one strand, and were not even the most prominent strand prior to the ascendancy of the National Socialists in Germany in 1933. Ideologues of the German movement, such as Fritz Lenz (1887–1976) and Eugen Fischer (1874–1967) were racial purists, but during the Weimar years kept a 'subdued' profile on overtly racial issues. Because racial purity was a cornerstone of the Nazi regime, however, eugenics after 1933 became openly racist and pro-Nordic.

A number of common ideological convictions characterized German eugenics from its pre-1900 days, right down through the 'final solution.' (a) The conviction that *laissez-faire* capitalism was at heart dysgenic because it allowed individuals to reproduce at will, with no concern for the future well-being of society. (b) The belief that Germany's population was degenerating, as evidenced by the debacle of World War I, increases in labor unrest, and social problems such as alcoholism, criminality, prostitution, mental illness, and retardation. (c) A strong belief in the value of a managed society based on scientific expertise. At the same time several features distinguish German eugenics from its counterparts elsewhere. (a) More German eugenicists were drawn from the medical profession than in either England or the United States (Proctor 1988). (b) During the Weimar Republic and later the Nazi regime, eugenics received far more government funding than in most other countries. German geneticists and eugenicists all held posts on important government advisory councils dealing with matters of eugenics and national health (Weindling 1985, pp. 308–9). (c) A major feature of German eugenics, of course, was the subsequent role it came to play in the National Socialist government and the Holocaust, between 1933 and 1945.

Some historians have argued, cogently, that eugenics became so prominent and virulent in Germany because of the humiliation of defeat in World War I and the exacerbating terms of the Treaty of Versailles. Among the factors leading to military defeat, eugenicists argued, was the nation's national degeneracy, the influences of racial impurity, declining birthrates among the intelligentsia, and economically prosperous, and the antieugenic policies associated with socialist egalitarianism (Mosse 1964, pp. 237–40, Meineke 1963, pp. 30–4). Worsening economic conditions due to reparations payments and loss of the mineral-rich Ruhr Valley, gave, among average Germans, added weight to arguments for improving national efficiency by eliminating what the Nazis termed 'useless eaters' or 'lives not worth living.' Eugenics became the scientific rationale for euthanasia and genocide.

All three Scandinavian countries, including Finland, had well-entrenched eugenics movements by the 1920s, those in Norway and Sweden being the largest and most politically successful. Scandinavian eugenics shows much of the character of the other western movements: an emphasis on efficiency, commitment to legislative action, and a strong pro-Aryan and Nordic stance (Broberg and Roll-Hansen 1996). In the period between 1910 and 1925, Scandinavian eugenics was characterized by a strong belief in the superiority of the Nordic races, and thus displayed a nationalist and racist undertone. These took the form of social legislation, particularly the passage of eugenical sterilization laws (Denmark's law was passed in 1919, Norway's in 1934, and Sweden's and Finland's in 1935). The rates of sterilization varied considerably: between 1935 and 1960s Sweden sterilized over 62,000 people—an astonishing figure for a country whose total population at the time was less than 7 million (Roll-Hanson 1996, p. 263). Eugenics achieved its greatest political success in Sweden, though the most vocal international voice from Scandinavia was Jon Alfred Mjøen (1860–1939), a pharmacist turned eugenic activist. Although shunned by many in his own country for his racist views, Mjøen was viewed as a major theoretician by eugenicists in Sweden and Denmark. Nonetheless, Scandinavian eugenics as a whole never adopted the virulent Nordic racism found among Nazi elements in Germany. As advanced

welfare states, the Scandinavian countries also couched their eugenics proposals in terms of 'national efficiency.' Hit hard by the same economic problems faced by other countries during the interwar period, and especially after 1929, the Scandinavian countries were seeking to reduce state expenditures wherever possible, and the supposedly hereditarily defective were a likely target.

5. *Eugenics in France*

The French, as other Europeans, harbored a growing concern about what was perceived as the degeneration of society, in France's case associated with a declining birth rate and loss of military hegemony on the continent after the Franco–Prussian war of 1870. French eugenics also embraced as strongly as elsewhere the idea of rational management by scientifically trained experts, especially in areas of social policy. Like other countries during the same period, France was beset with a variety of economic, social, and political problems that called for sweeping changes in both public attitudes and government policy (Thomson 1960, pp. 173–86). Before, but especially after, World War I, class struggle accelerated, in such publicly visible forms as large-scale strikes (Tilly et al. 1975, pp. 71–6). Like Germany, eugenics in France was strongly supported by members of the medical, including public health, professions. French eugenics differed from that in England, the United States, and Germany, however, in three very important ways. (a) The presence of a strong Enlightenment tradition of radical social theory and practice laid greater emphasis than elsewhere on environmental, as opposed to biological, determinants of human behavior. (b) Concomitant with their emphasis on the environment, though not directly deriving from it, French biologists maintained a strong commitment to neo-Lamarckian concepts of heredity, that is, the direct inheritance of environmentally induced variations. Consequently, eugenicists in France did not emphasize Mendelian genetics and other more strictly biological factors such as Darwinian natural selection. (c) The strong influence of the Catholic church, especially with regard to matters of reproduction, caused French eugenicists for the most part to avoid advocating measures such as contraception, sterilization, or abortion to deal with actual or potential defectives (especially after the Papal Bull, *Casti connubi* in 1930, which condemned eugenical sterilization). The effect of the church position was to widen an already-existing division within the eugenicists' ranks: between liberal and conservative and between pro- and antisterilization forces. Thus, French eugenicists never attained the unity necessary to lobby effectively for specific eugenic measures at the legislative levels.

6. *Eugenics in Russia*

The development of a distinct eugenics movement in Russia was apparent in the period just before the onset of the Bolshevik Revolution (1917). Russian eugenicists drew more of their inspiration and direction from the movements in the United States and Germany than elsewhere. Although eugenics, along with what might be called more strictly human genetics, was widely discussed in the Soviet Union, the eugenics movement there was confined almost wholly to the decade between 1920 and 1930 (Graham 1977, p. 1145). Increasingly viewed as racist and elitist in the political sphere, eugenics was disbanded even before the rise of neo-Lamarckism and the Lysenkoist movement in the 1930s and 1940s. Yet many enthusiastic and committed Bolsheviks saw no conflict between eugenical principles and socialism (Graham 1977, p. 1146). Indeed, many younger communists found eugenic ideals completely compatible with Marxism, providing a rational and materialist base to their visionary hopes for the scientific reconstruction of society.

Like the movements in England, the United States, and Germany, Russian eugenics at first was closely allied with Mendelian genetics. Although Russian eugenicists eventually moved away from negative toward more positive programs, as in Europe and the United States, they began with the deep forebodings of racial degeneration. As in other countries, in Russia after the revolution, eugenicists propounded a technocratic vision of social engineering by scientifically trained experts. It was only because of a strong Marxist opposition to technocracy that such a vision did not persist as long in Russian eugenics as it did elsewhere.

Like their counterparts elsewhere, soviet eugenicists hoped to apply both positive and negative eugenic measures to improvement of the human species. Based on a survey of the state sterilization laws in the United States, Russian eugenicists such as Nikolai Kol'stsov (1872–1940) argued for large-scale sterilizations of the unfit to improve the genetic character of the Russian people. Since birth control or sterilization were viewed as both morally questionable as well as dysgenic, negative eugenics, especially sterilization, never developed a wide following in the USSR.

A more viable option for pursuit of eugenic goals after 1917 was emphasis on the widespread belief in neo-Lamarckian inheritance. Like biologists and eugenicists in France, many Soviet geneticists and eugenicists adhered to some form of the inheritance of acquired characteristics. At the time, neo-Lamarckian views were accepted by many biologists around the world. Pioneer student of human genetics, Solom G. Levit (1894–c.1938), argued that neo-Lamarckism alone allowed for the rapid improvement in human health and social behavior that eugenics called for. If the gene was as fixed an entity as classical Mendelian genetics suggested, there was little chance of creating

the new soviet human being by eugenical means. Moreover, eugenicists were accused of propounding a hard-line concept of heredity to the total neglect of social and economic factors (Graham 1977, p. 1150). Eugenicists seemed to be reducing human beings to genetic machines whose consciousness was determined by their biology, not by their socioeconomic circumstances. As a result of these growing doubts, and the advent of the debate between Lysenkoists and Mendelians, by 1930 eugenics in the Soviet Union had been labeled reactionary and technocratic, and the major journal, *The Russian Journal of Eugenics*, was disbanded. More clearly than perhaps in other countries the fate of eugenics in Russia was uniquely tied to the turbulent period of history in which it rose and fell.

7. *Eugenics in Latin America*

The existence of significant eugenics movements in non-European and non-Anglo-Saxon contexts has been ignored by historians of science until quite recently (Stepan 1991). Eugenics movements in Latin America rose to prominence between the world wars and commanded considerable attention within the worldwide eugenics community. Interest in eugenics stimulated some of the first genetics courses in Latin American universities, and was often an important factor in child welfare, maternal health, infectious disease, and immigration legislation in the various countries (Stepan 1990, p. 110). Eugenics was also discussed vigorously at various Latin American scientific and public health conferences, as well as being the focus of two specific Pan-American conferences of its own (Cuba in 1927 and Argentina in 1934). A special eugenics society, the *Fédération International Latine des Sociétés d'Eugénique*, was founded in 1936 under the leadership of Corrado Gini, a well-known Italian eugenicist and president of the Italian society of Genetics and Eugenics. The *Fèdèration* held one international congress, in Paris in 1937. Moreover, leaders of eugenics movements in other countries expressed keen interest in the developments in Latin America. Not only Gini in Italy, but also Davenport and Laughlin in the United States, offered considerable encouragement to fledgling societies in Cuba and Mexico.

One of the most important differences between Latin American and western eugenics was the political economy of neocolonial 'underdevelopment' in which most Latin American countries existed in the first decades of the twentieth century. Although just beginning to experience industrialization and urbanization after World War I, Brazil, for example, was still largely an agricultural country with an impoverished rural population that was illiterate and dominated by the Catholic church. Much of the land was held as plantations by local family oligarchies and foreign investors who produced more for export than local consumption. The country was ruled by a tiny elite largely educated abroad. Biology, such as existed, was housed only in medical schools and agricultural stations. Historian Nancy Stepan has pointed to the highly mixed racial composition of most South American countries as a factor differentiating the Brazilian eugenics movement from that in Europe or the United States (Stepan 1990, p. 11).

Four specific factors helped to make eugenics attractive in the period after 1920 in countries such as Brazil (Stepan 1990, pp. 112–4). (a) A rising new nationalism and sense of independence gained from the war experience. The collapse of Germany, whose science and culture had been so much admired in Brazil and other Latin American countries before the war, generated a new sense of independence about what Latin Americans could accomplish on their own. (b) There was increased recognition of 'the social question,' that is, the appalling poverty and poor health of the working and peasant classes. After 1915 the combination of poverty, massive migrations, and immigration from abroad (particularly Italy) produced a radicalization of politics. A large-scale strike in 1917 was met with brutal repression by the government and industry leaders. Under these circumstances, eugenics began to play the same role as it had in the United States and Britain: a means of breeding out the biological defectives who were claimed to be the cause of such disruptions. (c) Growing enthusiasm for science, and particularly biological science, as a means of economic and social progress, especially in the area of public health, had been imported into Latin America from Europe and the United States. As in both France and Germany, eugenics in Latin America was easily incorporated by an expanding medical profession eager to assert its own power and expertise in generating new or reformed public health policy. (d) Finally, the widespread racial intermixing (between Indian, African, and European) that characterized most Latin American countries raised the specter of race suicide and impending degeneracy.

After starting from a more anthropological and public health perspective, by the mid-1920s a new generation of eugenicists had begun to apply Mendelian genetics to formulate Latin American eugenics policies. Greatly stimulated by the visit of American geneticist Theodosius Dobzhansky (1900–1975) to Brazil in 1943, eugenicists in Latin America, unlike their US or German counterparts, used genetics to argue that miscegenation, like crossing different strains of animals and plants, could produce superior, not inferior or 'mongrel,' lines. With such divisions in their ranks, however, eugenicists in Latin America could not speak with one voice. Yet, despite strong antiracist views advocated by Mendelian eugenicists, racism, and marginalization of the continent's black

and Indian populations continued with full force in the face of largely foreign-dominated economic, political, and social expansion.

The Latin American experience illustrates the quite different path that eugenics could take in an environment far removed from the centers of scientific advancement, one in which scientific ideas were derived largely from outside. The strong orientation of most Latin American eugenics to personal and public hygiene, as well as to nutrition and medical care, clearly reflects the existence of widespread poverty, malnutrition, and disease (hookworm, malaria, yellow fever) among the vast and largely rural populations. With economic and health problems of such magnitude, issues of race and race-mixture were less central among Latin American eugenicists than among their counterparts in the industrialized world.

8. *Eugenics in Asia*

Scholars studying the history of modern Asian science have pointed out that eugenics had a role to play in many of those societies. Moreover, it was not merely western eugenics imported and imposed on Asian society *carte blanche*. In both China and Japan, eugenics fitted well with movements for the increased promotion of science and technology associated with the 'modernization.' Mendelian genetics was introduced to both China and Japan in the early decades of the twentieth century through Western scientists who traveled to Asia, as well as by Asian scientists who came to work in laboratories in the West. Unlike their Western counterparts, eugenicists in both China and Japan relied less on a formal, organized set of genetic theories, promoting eugenics primarily as a 'modern' way of dealing with traditional problems.

For Chinese 'progressives,' eugenics became a way of grounding the calls for social and economic reform in science, and thus in the principle of rational, social control. Eugenics promoted a naturalistic view of society, based on human biology. Chinese adopted the 'soft' approach to heredity, one in which biological inheritance could be significantly molded by environmental factors. Not exactly relying on Lamarckian notions of inheritance of acquired characteristics, during the 1920s and 1930s, Chinese eugenicists emphasized the importance of education, public health, and other social measures, along with reproductive control. There was relatively little in the way of statistical or genetic research carried out by eugenicists in Republican China. As elsewhere, eugenics in China was particularly supported by the rising professional middle classes.

In Japan, during the early decades of the twentieth century, eugenics was linked closely with movements for birth control, sex education, and public health. In this context, conditions such as venereal disease, tuberculosis, alcoholism, and leprosy were all considered inherited and eugenicists argued they should fall under the same regulations as those governing public health matters. Here too, as in the West, eugenics was seen as a scientific means of combating social and moral degeneracy. Japanese eugenicists were successful in 1935 in getting the Imperial Diet to pass a Racial Eugenics Protection bill which favored the sterilization of hereditarily mental defectives. A sterilization law based on the German 1933 law (in turn based on model laws in the United States) was drafted in 1936, but a fully comprehensive eugenics law was not adopted until 1940.

Ironically, although eugenics was on the wane in the West after World War II, it became more prominent in China and Japan during that same period. Various laws and regulations in China focused on limiting family size, culminating in the 1978 'One Child' law that has raised considerable controversy both inside and outside of China. In Japan, a full-scale Eugenical Protection Law was passed in 1948, allowing for involuntary sterilization (by 1996 over 16,000 such operations had been carried out).

9. *Conclusion*

Although the term eugenics was used worldwide, eugenic movements in various countries were shaped by local economic, social, political, and cultural factors, and thus differed from one another in substantial ways. It has become clear that, as a movement, eugenics was not monolithic. Although tied to rigidly hereditarian Mendelian genetics in the United States, Scandinavia, and somewhat later in England and Germany, eugenics was more aligned with neo-Lamarckian views of inheritance in France, Latin America, and the Soviet Union. In their orientation toward various social issues and proposed solutions, eugenicists also varied considerably from country to country. In some, such as the United States and Germany, concern with racial issues played a prominent role especially in Germany after the Nazi takeover in 1933. In the United States and Germany, concern with the ill-effects of race-crossing was widespread while in countries such as Brazil, race mixture was considered by some eugenicists to be a source of biological improvement. For British eugenicists, the supposed degeneration apparent among the lower classes was the counterpart of the supposed inferiority of racial and ethnic groups in the United States. In all three cases, it was the economically dispossessed who formed the focal point of eugenicists' fears, and the targets of their programs.

Eugenics movements also differed widely in the social policies that were brought forward. In those countries such as France, Brazil, and to some extent the Soviet Union, where a more neo-Lamarckian view

prevailed and where the medical profession was more actively involved in eugenics, proposals tended to focus on environmental reforms, particularly with regard to public health. In those countries such as the United States and Britain, and especially Germany after 1933, where more strictly Mendelian views were predominant, segregation and/or sterilization of supposed 'defectives,' as well as immigration restriction, were more commonly proposed solutions.

It should be pointed out that the scientific basis for eugenical claims—that complex behavioral, personality, and mental traits are inherited in a 'hard-wired' fashion, and particularly that they follow Mendelian patterns of inheritance—has never found strong empirical support. Most eugenicists had a simplistic understanding of Mendelian genetics, and persistently ignored environmental inputs into determining social traits. Various geneticists in the West, such as Thomas Hunt Morgan (1866–1945) or Herbert Spencer Jennings (1868–1947) strongly criticized the genetic claims on which eugenics was based, but their public critiques came late in the movement and appear to have done little to stop it.

Underlying all the differences, eugenics movements worldwide shared a concern for efficiency, the ideology of scientific management, the role of middle-class experts in establishing social and political policy, and a strong faith in science. Underlying all of this, however, in almost all eugenics movements, was a powerful undercurrent of apprehension about the supposed degenerating biological quality of the human species. Eugenics programs, from California to the Baltic Sea, were aimed at increasing the biological worth of the human species by applying the latest biological and medical knowledge. That such knowledge was often incomplete or flawed ultimately brought the eugenic tree to bear a bitter fruit.

See also: Ethical Issues in the 'New' Genetics; Eugenics as the Basis of Population Policy; Eugenics, History of; Eugenics in Asia; Eugenics in Europe; 'New' Eugenics, The: Role in Society.

Bibliography

Adams M B (ed.) 1990 *The Wellborn Science. Eugenics in Germany, France, Brazil, and Russia*. Oxford University Press, New York

Allen G E 1975 Genetics, eugenics, and class struggle. *Genetics* **79**: 29–45

Allen G E 1986 The eugenics record office at Cold Spring Harbor, 1910–1940: An essay in institutional history. *Osiris* **2**(Second Series): 225–64

Allen G E 1991 Old wine in new bottles: From eugenics to population control in the work of Raymond Pearl. In: Rainger R, Benson K, Maienschein J (eds.) *The Expansion of American Biology*. Rutgers University Press, New Brunswick, NJ, pp. 231–61

Broberg G, Roll-Hansen N (eds.) 1996 *Eugenics and the Welfare State: Sterilization Policy in Denmark, Sweden, Norway, and Finland*. Michican State University Press, East Lansing, MI

Galton F 1883 *Inquiries into Human Faculty and its Development*. Macmillan, London

Graham L R 1977 Science and values: The eugenics movement in Germany and Russia in the 1920s. *American Historical Review* **82**: 1133–64

Kevles D J *In the Name of Eugenics*. Alfred A. Knopf, New York

Ludmerer K 1972 *Genetics and American Society: A Historical Appraisal*. Johns Hopkins University Press, Baltimore, MD

Meineke F 1963 *The German Catastrophe*. Beacon Press, Boston

Mosse G L 1964 *The Crisis of German Ideology. Intellectual Origins of the Third Reich*. Grosset and Dunlap, New York

Paul D 1984 Eugenics and the Left. *Journal of the History of Ideas* **45**: 567–90

Paul D 1995 *Controlling Human Heredity: 1865–Present*. Humanities Press, International Atlantic Highlands, NJ

Pickens D 1968 *Eugenics and the Progressives*. Vanderbilt University Press, Nashville

Proctor R 1988 *Racial Hygiene: Medicine under the Nazis*. Harvard University Press, Cambridge, MA

Roll-Hansen N 1996 Norwegian Eugenics: Sterilization as social reform. In: Broberg G, Roll-Hansen N (eds.) *Eugenics and the Welfare State: Sterilization Policy in Denmark, Sweden, Norway and Finland*. Michigan State University Press, East Lansing, MI

Schneider W H 1990 The eugenics movement in France, 1890–1940. In: Adams M B (ed.) *The Wellborn Science. Eugenics in Germany, France, Brazil, and Russia*. Oxford University Press, New York, pp. 69–109

Stepan N 1990 Eugenics in Brazil, 1917–1940. In: Adams M B (ed.) *The Wellborn Science. Eugenics in Germany, France, Brazil, and Russia*. Oxford University Press, New York, pp. 110–52

Stepan N 1991 *The Hour of Eugenics. Race, Gender, and Nation in Latin America*. Cornell University Press, Ithaca, NY

Thomson D 1960 *Democracy in France: The Third and Fourth Republics*. Oxford University Press, Oxford, UK

Tilly C, Tilly L, Tilly P 1975 *The Rebellious Century, 1830–1930*. Harvard University Press, Cambridge, MA

Vecoli R 1960 Sterilization: A progressive measure? *Wisconsin Magazine of History* **43**: 190–202

Weindling P 1985 Weimar eugenics: The Kaiser Wilhelm Institute for anthropology, human heredity, and eugenics in social context. *Annals of Science* **42**: 303–18

Weiss S F 1990 The race hygiene movement in Germany, 1904–1945. In: Adams M B (ed.) *The Wellborn Science. Eugenics in Germany, France, Brazil, and Russia*. Oxford University Press, New York

G. E. Allen

Eugenics as the Basis of Population Policy

Population policy refers to control of human reproduction by legislation or other official means for a given population of people (state, nation, social class, etc.). It involves setting goals for such population

parameters as total size, maintenance of stability, and rates of growth. Population policy includes programs such as the privately funded contraception program introduced into India in the 1950s by the Rockefeller Foundation, and publicly-funded programs such as those promoted throughout the Third World by the United States Agency for International Development and by China's official policy limiting families to two children. Although we are accustomed to thinking of population policies being aimed at reducing birth rate, policies such as the Catholic Church's opposition to contraception that increase birth rate, also represents a form of population policy.

Population policy should be distinguished from what is known as family planning or birth control. The latter refers to decisions within a given family regarding when to start having children, how many children to have, their spacing apart, etc. Organizationally, in the West family planning is represented by groups such as Planned Parenthood, but occurs through a variety of individual channels in most cultures. Population policy and family planning have areas of overlap in that population policy can only be effected at the level of the individual reproductive process. Both population policy and family planning/birth control include efforts to increase as well as decrease birth rate.

In Western Europe and the United States in the twentieth century population policy has often been associated with eugenics. The term 'Eugenics,' derived from the Greek *εὐγενής*, or *eugenes*, and was first coined by English statistician Francis Galton (1822–1911) in his book *Inquiries into Human Faculty and Its Development* to refer to one born 'good in stock, hereditarily endowed with noble qualities' (Galton 1883). As an intellectual and social movement in the early twentieth century, eugenics came to mean, in the words of one of its strongest American supporters Charles B. Davenport (1866–1944), 'the science of human improvement by better breeding' (Davenport 1910). For both Galton and Davenport, better breeding implied improving the hereditary quality of the human species by the scientific principles of heredity. Eugenics was the human counterpart of scientific animal and plant husbandry. It seemed ironic to eugenicists that human beings paid such careful attention to the pedigrees of their farm and domestic stock, while ignoring the pedigrees of their children.

Eugenicists wanted to develop population policies that dealt not only with quantity but also the quality of children being born. In this sense they sought to increase the reproductive rate of that segment of the population deemed to be biologically (meaning genetically) well-endowed (positive eugenics), and limit or prevent the reproduction of that segment of the population deemed to be biologically inferior or degenerate (negative eugenics). Eugenicists were concerned about what they saw as the increased reproductive rate of inferior (meaning working class and lower socioeconomic) classes, compared to superior people (meaning the middle and upper classes). While they were concerned with the inheritance of physical abnormalities such as hemophilia, color-blindness, brachydactyly (shortened digits), and debilitating diseases such as Huntington's chorea (all recognized to be genetic prior to 1910), eugenicists were ultimately more interested in human behavioral and personality traits—including mental ability, mental illness, 'moral degeneracy,' and criminality, to list only a few—all of which they claimed were genetically determined to a significant degree. Eugenics represented one of the earliest theories in the twentieth century to be based on biological, or genetic determinism (the idea that genes determine more of what a person is like, both in personality as well as in physical traits, than environment; eugenics took the side of 'nature' in the long-standing nature-nurture debate).

Although the term 'eugenics' and the idea behind it were Galton's, he himself did not found a movement either to develop eugenics as a 'science' or to put its principles into practice. Eugenics movements did not begin to arise in various countries of Europe or the United States until the first decade of the twentieth century, nor did they become generally effective in promoting social and political programs nationally or internationally until after 1910. The earliest eugenics movements were founded in Germany in 1904, in Britain in 1907, and in the United States in 1908–10. Other eugenics movements appeared subsequently around the world: in Western Europe (France, Norway, Spain), Russia, Latin America (Cuba, Brazil, Mexico), in Canada, and Asia (Japan). However, it was in the United States, Britain, and Germany where eugenics as an intellectual and social movement reached its greatest strides and from eugenicists' point of view, greatest ideological and political results.

One of the acknowledged leaders of eugenics on an international level was American Charles B. Davenport. Although somewhat eccentric, Davenport was a highly respected biologist, with a Ph.D. from Harvard (1891), where he subsequently served on the faculty for a decade (1891–1900). Between 1900 and 1904 he taught at the University of Chicago before obtaining funds from the Carnegie Institution of Washington to establish his own research laboratory, the Station for the Experimental Study of Evolution (SESE) at Cold Spring Harbor, Long Island (New York). In 1909 Davenport convinced the Harriman family of New York to establish a second institution, the Eugenics Record Office (ERO), also at Cold Spring Harbor to serve as a center for eugenical research, a clearing-house for exchange of eugenic information, a repository for eugenic data, and an organizational center for eugenic education and propaganda. Heading the ERO was Davenport's chief administrative aide, Harry Hamilton Laughlin, formerly a teacher of agricultural sciences at the Kirksville Normal School in Missouri.

In its claim to be scientific, eugenics drew heavily on the newly discovered work of Gregor Mendel, the Moravian monk who, in 1866, had published a paper on breeding patterns in peas. Although ignored for the next 35 years, in 1900 Mendel's paper was recognized for its path breaking approach to the problems of heredity by three investigators, more or less independently. By the time of World War I the Mendelian principles of heredity were being applied with great success to many varieties of animals and plants, including humans. It was in this atmosphere that eugenics flourished. It was hailed as a scientific solution to recurrent social problems. If alcoholism, criminality, pauperism, and feeblemindedness were all determined by genes, then the way to eliminate these problems would be to prevent the reproduction of those individuals thought to carry those genes. As eugenic ideas spread, other centers of eugenics activity arose in the United States, England, and in Europe: the Eugenics Research Association (ERA) and the American Eugenic Society (AES), both in the US, the Eugenics Education Society (England), and the Gesellschaft für Rassenhygiene (Society for Racial Hygiene) in Germany, among many others.

1. *Eugenic Research: Its Aims and Problems*

Since eugenicists could not carry out planned matings in the manner of laboratory geneticists, the family pedigree chart was the main analytical means for displaying and analyzing data on the heredity of one or another trait. Information was collected from among a group of relatives, and it was displayed as a family pedigree chart, emphasizing that traits ran in families and thus how heredity influenced many behavioral, personality, and mental traits (see Table 1). Although Davenport was acknowledged in many ways as the intellectual and scientific leader of American eugenics, his research was often naive, careless, and overstated. Even Galton and his protégé Karl Pearson, felt that despite Davenport's enthusiasm, he was 'not a clear, strong thinker' (Kevles 1985). His claim that complex mental and behavioral traits were almost exclusively controlled by one or two Mendelian genes, ignored the complexity of human mental development and claimed that thalassophilia, or 'love of the sea,' was a sex-linked Mendelian recessive appearing in the families of prominent US naval officers (Davenport 1919). That the condition must be sex-linked was clear to Davenport, since in pedigree after pedigree only males in the various families he studied became naval officers. When one of his friends, a professional psychiatrist, criticized him for lumping complex human behaviors into single categories such as 'insanity,' Davenport dismissed the criticism as uninformed (Kevles 1985). Such simplistic Mendelian arguments were then extended from individuals to large groups as a way of explaining the differences between racial, ethnic, and national groups.

Table 1
Some *lists of proposed human races* (compiled from Coon 1962, Cohen 1991, and Molnar 1997)

Author	N of races	Races proposed
Linnaeus (1735)	6	europaeus, asiaticus, afer, americanus, ferus, monstruosus
Buffon (1749)	6	laplander, tartar, south asiatic, european, ethiopian, american
Blumenbach (1795)	5	caucasian, mongolian, ethiopian, american, malay
Cuvier (1828)	3	caucasoid, negroid, mongoloid
Deniker (1900)	29	
Weinert (1935)	17	
Von Eickstedt (1937)	38	
Biasutti (1959)	53	
Coon (1962)	5	congoid, capoid, caucasoid, mongoloid, australoid

Because of their concern with the apparent rise in 'feeblemindedness' eugenicists embraced the nascent psychometric movement (intelligence testing) starting as early as 1912. In that year Henry H. Goddard administered a version of the French Binet-Simon test to immigrants at Ellis Island. Although the Binet-Simon test was intended to measure only an individual's mental functioning at a given point in time, Goddard and a host of psychometricians in the United States, England, and Germany considered that it also measured innate, or genetically determined intelligence. Goddard coined the term 'feeblemindedness' (which he claimed was genetic) to refer to those people who scored below a given level (70) on his tests, and claimed that it 'was a condition of the mind or brain which is transmitted as regularly and surely as color of hair or eyes' (Kevles 1985). It was Goddard who carried out the famous study of the Kallikak family, showing what was presumed to be the hereditary effects of feeblemindedness and degeneracy (Goddard 1912). Most of the early psychometricians such as Lewis Terman and Robert Yerkes were members of the American Eugenics Society, while both Yerkes and Terman sat on the Society's Committee on Psychometrics. For eugenicists, the new mental tests were seen as a precise, quantitative tool for measuring an otherwise elusive, but fundamental human trait. Like their British professional, middle-class counterparts, psychometricians felt that the scientifically-trained expert could at last solve with rational and scientific methods the social problems that muddle headed politicians and social workers had been unable to solve during the preceding century.

A major problem with eugenic research was that it quickly fell behind the work being done by geneticists in laboratories throughout the world. By the 1920s many geneticists had come to recognize that genes (collectively, known as the organism's genotype) interact with one another (a phenomenon known as epistasis) and with the environment to produce any given trait (the organism's phenotype). The interaction of genes with one another and environmental factors (for example, temperature, drugs, diet) was found to

alter the gene's expression in profound ways, yielding a variety of alternative phenotypes from the same genotype. Even the position of a gene along the length of its chromosome turned out to affect the way that gene is expressed. This malleability, or plasticity, of the genotype-phenotype relationship meant that even knowing the genotype of the offspring at conception did not provide a certain means of predicting the phenotype. Eugenicists' understanding of Mendelian heredity seems to have become frozen in time somewhere around 1910–5. While they coat-tailed on much of the exciting new work in genetics in the 1920s and 1930s, it is apparent that many eugenicists did not really understand it fully, or take to heart its implications for their social programs.

2. *Eugenics and Legislation of Reproductive Policy*

In the United States Harry H. Laughlin was particularly active in lobbying for the passage of a number of state eugenical sterilization laws in the 1920s and 1930s (Chase 1977, Kühl 1993). Indeed, Laughlin drew up a 'Model Sterilization Law' that served as a prototype from which each state could derive its own modified version (Laughlin 1920). Although the earliest such laws had been passed before World War I, the majority was established during the interwar period (Reilly 1992). Eugenical sterilization was aimed specifically at those individuals in mental or penal institutions who, from family pedigree analysis, were considered likely to give birth to 'socially defective' offspring. Sterilization could be ordered only after a patient had been examined by a eugenics committee, usually composed of a lawyer or family member representing the individual, a judge, and a doctor or other eugenic 'expert' (Reilly 1992).

Between 1907 (when the first such law was put into effect in Indiana) and 1941, over 38,000 eugenical sterilizations were performed in the United States, with California leading all other states (Reilly 1992). By the early 1960s, the number had risen to some 60,000 (Robitscher 1973). The most famous of the sterilization cases in the US was that of Carrie Buck in Virginia. When Virginia passed its eugenical sterilization law in 1924, a test case was arranged the next year to determine if the law was constitutional. Buck *vs* Bell was tried in 1925 in the Virginia Circuit Court (Lombardo 1985). When the lower court ruled in favor of the law, an appeal was sent to the Supreme Court of the United States, where Justice Oliver Wendell Holmes wrote the majority report. In upholding the lower court ruling, Holmes made his oft-repeated assertion: 'Three generations of imbeciles are enough.' More to the legal point, Holmes claimed that: 'The principle that sustains compulsory vaccination is broad enough to cover cutting the Fallopian tubes' (Reilly 1992). Laughlin's model sterilization law also served as the basis for a similar law in Germany passed by the National Socialist government in 1933 (Chase 1977). For this effort, as well as for his enthusiastic support of Nazi eugenic programs, Laughlin was awarded an honorary doctorate from Heidelberg University on the occasion of its 550th anniversary celebration in 1936: three years after the Nazis had come to power and two years after their sweeping sterilization act had gone into effect (Allen 1987).

3. *Criticism of Eugenics, 1925–50*

Almost from the beginning many of the basic premises of eugenics were brought under critical scrutiny by biologists, medical doctors, social workers, and lay persons from all walks of life. Among the first and most important critics from the biological side was Thomas Hunt Morgan, the *Drosophila* geneticist at Columbia University and, in 1933 the first recipient of the Nobel Prize for work done in genetics, and one of the earliest members of the Committee on Animal Breeding of the American Breeders' Association (a Committee that had taken on itself organizing the first eugenic activities in the US). In a personal letter sent to his close friend Davenport in 1915, Morgan resigned from the Committee, explaining that the claims being made on behalf of eugenics were exaggerated beyond what any scientific evidence could support (Allen 1978). Although Morgan did not criticize eugenical work in print at the time, from 1925 on, starting with his book *Evolution and Genetics* (Morgan 1925) he publicly chastised eugenicists for lumping many mental/behavioral conditions together under a rubric like 'feeblemindedness' and treating it as if it had a single underlying cause. Especially critical, he noted, was the fact that personality and social traits were the most malleable of our characteristics, and hence the most difficult in which to tease apart the relative contributions of heredity and environment:

> The important point … is that mental traits in man are those that are most often the product of the environment which obscures to a large extent their inheritance, or at least makes very difficult their study (Morgan 1925).

The pedigree charts on which Mendelian analyses were based, were also strongly criticized by Morgan and others as giving the impression of tracing genetically determined transmission when, of course, they equally well could be interpreted as showing culturally-determined effects:

> The pedigrees that have been published showing a long history of social misconduct, crime alcoholism, debauchery and venereal disease are open to the same criticism from a genetic point of view; for it is obvious that these groups of individuals have lived under demoralizing social conditions

> that might swamp a family of average persons. It is not surprising that, once begun from whatever cause, the effects may be to a large extent communicated [socially] rather than inherited [genetically] (Morgan 1925).

Pedigree charts do not separate 'nature' from 'nurture.'

Other geneticists followed suit. Raymond Pearl, a geneticist and supporter of general eugenical principles at Johns Hopkins University, also voiced public criticism of eugenicists claims, especially those regarding the supposed biological inferiority of various racial and ethnic groups (Pearl 1927). Along with his colleague Herbert Spencer Jennings, Pearl felt that propagandists like Laughlin, Madison Grant, and others made claims that went far beyond any reasonable evidence (Barkan 1992). Jennings devoted the better part of the ensuing decade (after 1924) to lecturing all over the United States about the dangers of oversimplified and racially based eugenical claims. H. J. Muller, a student of Morgan's, and later himself to become a Nobel laureate in genetics, delivered a searing attack on 'old style' eugenics at the Third International Eugenics Congress in New York City in 1932 (Muller 1932). Muller, who still harbored strong eugenical beliefs, argued that until the economic and social environment could be equalized it would be impossible to know how much of any individual's 'social inadequacy' was due to heredity and how much to environment (see also Ward 1935, and for a review, Paul 1984). Much of feeblemindedness might be due to poor maternal diet during pregnancy or during the early years of childhood, or to maternal drug and alcohol use; although these effects might appear at a superficial glance to be genetic, they had nothing to do with defective germ plasm. The same point was made also by influential Danish plant breeder Wilhelm Johannsen, who wrote:

> There is no reason to assume that the weak and sickly would represent the genetically inferior stock: they could be individuals of the same [hereditary] value as children from higher social classes who are better cared for (Hansen, in Broberg and Roll-Hansen 1995).

English mathematician/geneticist and sometime eugenics critic Lancelot Hogben made one of the clearest statements at the time about the over-simplified concept of genetics that informed much of the eugenics movement:

> No statement about a genetic difference has any scientific meaning unless it includes or implies a *specification of the environment* in which it manifests itself in a particular manner (Ward 1935).

In 1934, a special blue-ribbon committee of the British government published the Brock Report (named for its chairman, Laurence B. Brock, head of the Joint Committee on Mental Deficiency), outlining difficulties in determining the relative contribution of genetics to mental deficiency. Among others, the Committee included Ronald A. Fisher, himself a staunch eugenicist but also one of the world's foremost statisticians and population geneticists. The Brock Report agreed with the criticisms of eugenics put forward a few years earlier by Lionel Penrose, who had demonstrated that the category of 'feeblemindedness' was ill-defined; was used carelessly and indiscriminately to cover a host of mental problems; was caused in many cases by disease and social deprivation; and for the most part little was known of the genetic cause of mental disability (Kevles 1985). As the Brock Report stated:

> the more closely individual records are examined the more difficult it becomes to fix on one cause to the exclusion of others, or to say with certainty that the genetic endowment of any individual is such that it must produce a given result (Kevles 1985).

An even more comprehensive and devastating critique of the whole range of eugenic claims was published by the American Neurological Association in the mid-1930s, just a few years after the appearance of the Brock Report (Myerson et al. 1936). Work for the book was funded by the Carnegie Foundation through the New York Academy of Sciences. Like the Brock Report the findings of the Myerson Committee found that by the mid-1930s abundant evidence was available throwing into doubt such well-accepted eugenical beliefs as: (a) the vast number of mental and moral/personality traits are genetically determined; (b) environment (upbringing) plays a relatively minor role in the development of mental and personality traits; and (c) sterilization is the most effective means of reducing the frequency of socially inadequate behavior in future generations.

The Myerson Committee supported many of the general conclusions of the Brock Report. Citing the work of Morgan, Jennings, and others in the US, and that of J. B. S. Haldane in England the authors argued that we do not know 'enough about the mental diseases and the effects of the social structure to manipulate in a successful way the procreation of the race' (Myerson et al. 1936). The level of genetic knowledge was too slight to formulate any social policies or carry out any specific manipulations guaranteed to alter the genetic basis of future generations.

More specifically, Myerson et al. found that eugenicists consistently downplayed the role of the environment in the development of mental and behavioral problems, even when it seemed obvious that social and economic conditions must be playing a highly significant role. H. S. Jennings at Johns Hopkins University made the same point, emphasizing that 'environment' included the internal, cellular, as well as external environment in affecting how genes were expressed. Even if a person is clearly shown to have a genetic defect, as in hereditary thyroid deficiencies, Jennings argued, the condition can be treated

with dietary and/or drug supplements, which can restore normal function. A whole host of human anatomical and physiological conditions that were clearly genetic did not need to be debilitating or limiting to the individual, according to Jennings, because in many cases we possess the technological means to restore, or at least compensate for, the genetic defect. Genetics, according to Jennings, was not destiny (Jennings 1927). For eugenicists, however, it was, and this remained one of the underlying assumptions of hereditarian thinking from the past to the present.

On the eve of World War II (literally one week before), another strong criticism of eugenical thinking emerged as the 'Geneticists' Manifesto,' a document signed in early August, 1939, by a group of geneticists assembled in Edinburgh at the VIIth International Congress of Genetics. Among the major initiators were J. B. S. Haldane and Lancelot Hogben (England), and Hermann Joseph Muller (United States). The Manifesto put forth Muller's argument that from a genetics point of view as long as vast disparities in wealth and access to resources continued to exist in Western societies, no claims could be made about the relative contribution of heredity over environment to human mental capacity and social behavior (Chase 1977). They condemned racialist thinking, especially the attempt to rank races with respect to one another on the basis of genetic worth. The geneticists who initiated and signed the Manifesto were among the most well-known and respected workers in the field during the twentieth century.

In the period after World War II overt criticisms of eugenics and eugenical thinking became more widespread. Revelation of Nazi atrocities committed in the name of eugenics played an important role in this change of attitude. In addition, hard-line eugenicists such as Davenport, Laughlin, Grant, and others on the American scene were retired or deceased by the end of the war, and were replaced by what historian Daniel J. Kevles has dubbed moderate, or 'reform eugenicists.' Although it might be questioned how really 'reformed' these eugenicists were, it is clear they were more knowledgeable about the nuances of what traits genes might and might not control, less overtly racist, and more modest in their claims for what eugenics might accomplish. Partly because of the criticism, it was difficult for the field of human genetics to get a fresh start after the war, and many geneticists shied away from dealing with human heredity (Kevles 1985, pp. 193ff., Ludmerer 1972). Textbooks and other writings on human genetics were careful to distinguish themselves from old style eugenic thinking (Ludmerer 1972). Laurence H. Snyder's *Human Genetics* (1951), for example, stated clearly: 'We should hesitate at accepting without careful analysis the claims and recommendations that are present in a large portion of the eugenics literature.' An understanding of the historical development of the eugenics movement, Snyder went on to say, would suggest 'why professionally trained geneticists, on the whole, have given relatively little support to eugenic theories and programs' (Ludmerer 1972). Anthropologist and prolific popular writer Ashley Montague echoed the same feeling in his 1959 *Human Heredity* when he wrote that much unfortunate public policy—by which he meant immigration restriction and state sterilization laws—had been based on 'the ill-considered, prejudiced, and unscientific judgments of the late Harry Laughlin … Madison Grant … and Charles Benedict Davenport' (Montague 1959, pp. 25–6). And, in 1960, the American Medical Association came out with a report, *A Reappraisal of Eugenic Sterilization Laws*, that called into question virtually all the assumptions of the old eugenic movement.

Despite these severe criticisms, eugenics continued to thrive in the Scandinavian countries and especially in Nazi Germany (Proctor 1988). After passage of the sterilization laws in 1934, the Nazi government enacted the first of the infamous Nüremberg Laws in 1935, dismissing Jewish and half-Jewish citizens from state jobs (including all university positions) and prohibiting marriage between Aryans and Jews and other 'non-Nordic' groups. Doctors were required to report to the Reich Health Department cases of what they deemed to be hereditary diseases among their patients. After the large-scale deportation to concentration camps, many doctors became involved in a variety of genetic and other studies, including painful and deadly experiments, on twins. All in all, the Nazis carried eugenic principles further than in any other country of Western Europe or North America.

4. *Later Development of the American Eugenics Movement*

The period of greatest growth and political influence of eugenics ideology in the United States occurred between 1920 and 1935. By the mid-1930s American eugenics began to experience a decline in general popularity and in political effectiveness (Haller 1963, Ludmerer 1972, Paul 1995). Scholars have disagreed on the extent to which the movement went into 'decline' in the United States, or the reasons for its change of fortune. Clearly, both the depression of 1929–33 and reports of Nazi eugenics in the American press played some part in a general disaffection with eugenical principles. In the depression people without jobs became 'paupers' and 'social inadequates' overnight with no change in their genetic make-up, while in Germany the wholesale sterilization of those deemed biologically inferior demonstrated the logical conclusion to which eugenic principles could lead.

An additional factor may have been the recognition that American eugenicists were out of touch with the most recent findings of laboratory genetics. Davenport and Laughlin's simple unit-character concept did not

square with recent experimental data suggesting that most traits were produced by the interaction of many genes, and that the evidence for a genetic basis of complex human social behaviors was almost non-existent. Nonetheless, it is the opinion of historian of science Diane Paul, among others, that neither the scientific criticisms nor the pre-war use of eugenics by the Nazi government had much influence on the movement's decline. More specifically, G. E. Allen argues that a major cause for the decline of eugenics was that the movement had served the purposes of the economic elites (Rockefellers, Harrimans, Kelloggs, Carnegies) whose philanthropies had supported it. That function was to try and bring science to bear on the solution of social problems in a period rocked by violent labor-capital confrontations, and by the rise of a variety of social problems engendered by rapid industrialization and urbanization (Allen 1986). Whatever the exact causes, what has been labeled 'mainline' or 'old-style' eugenics was losing favor in the US and England especially, by the late 1930s; symbolically, the Carnegie Institution of Washington withdrew funding for the Eugenics Record Office by late 1939, and it closed on December 31 of that year (Allen 1986). Secretary of the American Eugenics Society, Frederick Osborn, retired railway executive and nephew of American Museum of Natural History President, Henry Fairfield Osborn (a wealthy-elite New York family who supported eugenics in a variety of ways), has been described as a 'moderate' among eugenicists, perhaps a misleading term but nonetheless signaling the different twist that he gave to eugenics in the postwar period (Mehler 1988). That approach did not involve a rejection of eugenic ideology or goals, but a toning down of the overtly racist claims and simplistic genetic analyses put forward by Davenport, Laughlin, Grant, and others of their generation. The face of eugenics had begun to change, even if the principles at its core remained largely the same.

5. *Eugenics and Population Control*

As early as 1926 Raymond Pearl of Johns Hopkins University claimed that eugenicists had focused too much attention on the quality of offspring from particular (usually 'degenerate') families, and had lost sight of the equally important issue of the rapid birth rate of the biologically 'degenerate' at home and abroad. In a series of scientific and popular articles he began to articulate the ideology of population control. While critical of much of the current eugenical propaganda, Pearl did not disavow basic eugenical principles. He simply emphasized the importance of quantitative as well as qualitative considerations.

In the aftermath of World War II the United States emerged as a contender for global power, inheriting the mantle of international economic and political involvement that had previously been carried by the British and, to a lesser extent, the French. In 1952 President Dwight D. Eisenhower formed a blue-ribbon panel, headed by David Rockefeller of the Chase National (later Chase-Manhattan) Bank, to evaluate the future prospects for US industrial and commercial expansion around the world. In the same year, the Rockefeller Foundation organized a conference in Williamsburg, Virginia, under the direction of John D. Rockefeller III, to study what Pearl had highlighted as the 'population problem.'

The blue-ribbon panel concluded that the single greatest obstacle to US industrial expansion (meaning access to cheap raw materials, labor, and ultimately markets) was the rapid population growth occurring in most Third World countries. Such population increases not only required the retention of more of countries' raw materials (including agricultural products) at home, but also fueled local revolutionary movements aimed at eliminating foreign control. Not surprisingly, the Williamsburg conference called for the formation of an organization to look into methods for controlling population growth. Out of this meeting came the proposal for establishing the Population Council, with Rockefeller funds and John D. III himself as President. Frederick Osborn, still Secretary of the American Eugenics Society at the time, soon took over as Council President, running both organizations out of his office in Rockefeller Center in New York.

One of the first major projects of the Council was the funding of investigations of various forms of birth control, including the IUD (intra-uterine device), and eventually, the first birth-control pill (Watkins 1998). Initial trials in Puerto Rico beginning in April 1956, led to field studies in other Third World countries such as India and Pakistan. One of the major incentives used to promote these studies was the economic benefit that would accrue to families that had fewer children. That such pilot population control movements encountered considerable resistance in the countries where they were introduced, suggested that local people saw the matter quite differently. The programs provided a lesson in the cultural and social components of reproductive life. For many in agricultural, Third World countries, large families were the only way to insure some manner of economic survival. Children worked on farms at an early age, and thus more than earned their 'keep,' while large families often provided the only form of old-age insurance in countries that lacked programs for the elderly. This lesson is still one to be learned by leading political and many environmental and conservation groups, who continue to advocate population control as the major agenda for Third World countries. Today, for example, the United States Agency for International Development, a branch of the State Department, devotes a significant portion of its budget to its Population Office. Programs emerging from

USAID support population control programs in over 20 Third World countries. Eugenic ideology is still providing as a basis for population policy throughout much of the industrialized world.

See also: Bioethics: Examples from the Life Sciences; Bioethics: Philosophical Aspects; Eugenics as an International Movement; Eugenics, History of; Eugenics in Asia; Eugenics in Europe; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease-related Characteristics; Genetic Testing and Counseling: Educational and Professional Aspects; 'New' Eugenics, The: Role in Society; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations

Bibliography

Adams M B (ed.) 1990 *The Wellborn Science. Eugenics in Germany, France, Brazil and Russia*. Oxford University Press, NY

Allen G E 1986 The Eugenics Record Office at Cold Spring Harbor, 1910–40: An essay in institutional history. *Osiris 2nd. Series 2*: 225–64

Allen G E 1991 Old wine in new bottles: from eugenics to population control in the work of Raymond Pearl. In: Rainger R, Benson K, Maienschein J (eds.) *The Expansion of American Biology*. Rutgers University Press, New Brunswick, NJ

Barkan E 1992 *The Retreat of Scientific Racism. Changing Concepts of Race in Britain and the United States between the World Wars*. Cambridge University Press, NY

Chase A 1977 *The Legacy of Malthus*. Alfred A. Knopf, NY

Davenport C B 1910 *Eugenics*. Henry Holt, NY

Davenport C B 1919 *Naval Officers: Their Heredity and Development*. Carnegie Institution of Washington, Washington, DC

Galton F 1883 *Inquiries into Human Faculty and Its Development*. Macmillan, London

Haller M 1963 *Eugenics. Hereditarian Attitudes in American Thought*. Rutgers University Press, New Brunswick, NJ

Hassencahl F 1970 *Harry H. Laughlin, 'Expert Eugenics Agent' for the House Committee on Immigration and Naturalization, 1921 to 1931*. Ph.D. Dissertation. Case Western Reserve University, University Microfilms, Ann Arbor, MI

Jennings H S 1924 Heredity and environment. *Scientific Monthly* **19**: 225

Jennings H S 1927 Public health progress and race progress—are they incompatible? *Science* **66**: 45–9

Kevles D J 1985 *In the Name of Eugenics*. Alfred A. Knopf, NY

Kuhl S 1993 *The Nazi Connection. Eugenics, American Racism and German National Socialism*. Oxford University Press, NY

Lombardo P A 1985 Three generations no imbeciles: New light on Buck v. Bell. *New York University Law Review* 1985 **60**: 30–62

Ludmerer K 1972 *Genetics and American Society: A Historical Appraisal*. Johns Hopkins University Press, Baltimore, MD

Mehler B 1987 *A History of the American Eugenics Society, 1921–40*. Ph.D. Dissertation. University of Illinois, Champaign-Urbana, IL

Morgan T H 1925 *Evolution and Genetics*. Princeton University Press, Princeton, NJ

Muller H J 1933 The dominance of economics over eugenics. *Scientific Monthly* **37**: 40–7

American Neurological Asociation 1936 *Eugenical Sterilization: A reorientation of the problem*. Macmillan, NY

Paul D 1995 *Controlling Human Heredity, 1865–Present*. Humanities Press International, Atlantic Highlands, NJ

Pearl R 1927 The biology of superiority. *American Mercury* **1**: 257–66

Proctor R 1988 *Racial Hygiene: Medicine under the Nazis*. Harvard University Press, Cambridge, MA

Reilly P 1992 *The Surgical Solution. A History of Involuntary Sterilization in the United States*. Johns Hopkins University Press, Baltimore, MD

Roll-Hansen N 1989 Eugenic sterilization: A preliminary comparison of the Scandinavian experience to that of Germany. *Genome* **31**(2): 890–5

Vecoli R 1960 Sterilization: A progressive measure? *Wisconsin Magazine of History* **43**: 190–202

Weiss S 1990 The race hygiene movement in Germany, 1904–45. In: Adams M (ed.) *The Wellborn Science, Eugenics in Germany, France, Brazil and Russia*. Oxford University Press, NY

Zenderland L 1998 *Measuring Minds*. Cambridge University Press, NY

G. E. Allen

Eugenics, History of

Eugenics is the theory and practice of improving human heredity, typically through control of reproduction. Although the concept is very old, stretching back at least to the ancient Greeks, the modern movement dates to the nineteenth century when Francis Galton, a British polymath and first cousin of Charles Darwin, decided it was time for humans to learn from the breeders of horses and cattle and take human evolution into their own hands. In 1883, Galton termed this project 'eugenics,' from the Greek *eugenes* for 'well-born.' He defined it broadly as 'the science of improving stock, which is by no means confined to questions of judicious mating, but which … takes cognisance of all influences that tend in however remote degree to give the more suitable races or strains of blood a better chance of prevailing speedily over the less suitable than they otherwise would have had.' A conventional distinction contrasts 'positive' eugenics, which promotes breeding by those possessed of superior hereditary qualities, with 'negative' eugenics, which deters breeding by those with physical, mental, or moral defects.

1. Eugenics before Galton

Abandonment or outright killing of deformed infants was practiced in the ancient world, although both the rationale for and the extent of infanticide are matters

of controversy (Roper 1913, Riddle 1992, pp. 10–15). Infanticide of defective infants was approved by both Plato and Aristotle; indeed, the latter thought it should be mandated by law.

Eugenics is central to Plato's *Republic*. Written in the fifth century BC, it describes a utopia in which the family is replaced by a community of wives and children. Mating within the Guardian class is tightly controlled. According to Socrates, the Rulers would never allow unregulated unions. Such a practice would be 'a profanation in a state whose citizens lead the good life' (Cornford 1941, p. 157). In a dialogue whose theme will be often reprised, Socrates constructs an analogy between human and animal breeding. Just as we breed from the best of our sporting dogs and game birds, he argues, we should ensure that among our own flock, there be as many unions of the best, and as few of the inferior, as possible. Plato's analogy, which was further elaborated in the *Laws*, influenced the seventeenth-century Dominican Monk, Tomasso Campanella. His *City of the Sun* (1623) describes a utopia where children are bred for the good of the commonwealth, and whose inhabitants are scornful of those who neglect the breeding of their own kind.

2. *From Galton to the 1930s*

2.1 *Charles Darwin and Evolutionary Thought*

Campanella notwithstanding, proposals to control human heredity faded with the advent of Christianity. They re-emerged following publication of Darwin's *Origin of Species* (1859). Darwin argued that evolution occurred through a process of fierce competition in which organisms with advantageous traits would tend to survive and reproduce their kind, while the ill-adapted would leave few if any progeny. He and his contemporaries viewed this process as a progressive force that tended toward the constant physical and mental improvement of organic beings. But many Victorians worried that natural selection was no longer at work in their own species since modern medicine, charity, and other humanitarian measures prevented the culling of the physically and mentally weak. Moreover, it seemed as though the worst elements of society now were breeding at a much faster rate than the best. Darwin expressed a common middle-class anxiety about the 'differential birthrate' when he remarked in *The Descent of Man* (1871) that if various checks 'do not prevent the reckless, the vicious, and the otherwise inferior members of society from increasing at a quicker rate than the better class of men, the nation will retrograde, as has occurred too often in the history of the world. We must remember that progress is no invariable rule.' Darwin himself did not advocate any practical measures to counter the anti-selective forces of modern society. But in 1865, his cousin Galton published the first modern eugenics manifesto, 'Hereditary talent and character.'

2.2 *Galton's Argument*

Galton aimed to prove that human mental, moral, and temperamental traits, and not just physical ones, were transmitted from parents to offspring. These traits included a gregarious personality, a truthful or courageous character, a liking for alcohol or drugs or gambling, a tendency to pauperism or crimes of violence or fraud. He was particularly concerned to prove the heritability of traits making for success.

His method was simple, if from a contemporary perspective naive. Consulting biographical dictionaries, he established that fathers who had achieved eminence in various fields were much more likely than members of the public at large to have had fathers who were themselves eminent; in short, that distinction ran in families. Of course Galton knew that individuals also inherit social advantages, and he conceded that this might explain success in certain fields. No admirer of the landed aristocracy, which still dominated parliament, the army, and civil service, Galton noted that 'many men who have succeeded as statesmen would have been nobodies had they been born into a lower rank of life' (quoted in Jones 1998, p. 7). But in the 'open fields' of science, literature, and the law, individuals without talent would surely fail, no matter how distinguished their families, while conversely, the talented would succeed, no matter how unfavorable their background. Unfortunately, it seemed that those with the most talent were the least prolific. Galton argued that there was an urgent need to reverse this trend. However, he offered no specific measures to accomplish this end, only the conviction that if everyone could be made to see the importance of breeding, 'some way or other' would be found to get the job done.

2.3 *Changing Views of Heredity*

Galton believed that the hereditary material was transmitted intact from one generation to another. However, his contemporaries were 'Lamarckians,' who assumed that characteristics acquired by organisms during their lifetime (primarily through the use and disuse of particular organs) could be transmitted to their offspring.

The 'Lamarckian' or 'soft' view of heredity blurs the distinction between nature and nurture, since it assumes that inherited traits can be suppressed or redirected by changing the conditions of life. Thus, even if social problems such as criminality, promiscuity, or pauperism resulted from poor heredity, the solution could reside in improved nutrition, education, or other social reforms. But if heredity is impervious to

environmental changes, the only way to eliminate the problems it caused would be through selective breeding.

Galton lacked experimental evidence for his assumption that heredity was fixed. But in 1883, the German cytologist August Weismann (1834–1914) was able to distinguish germ cells, which are present in the gonads and give rise to sperm and ova, from somatic cells, which are present in all the other tissues of the body. Weismann argued that only the somatic cells could be affected by the environment, whereas the germ cells would be transmitted intact from generation to generation. Thus, no matter how hard individual parents worked to improve their minds or bodies, or whatever social reforms might be instituted, no child would be born healthier, or more equable, virtuous, musical, athletic, or smarter as a result.

2.4 Eugenics Becomes Popular

Galton's work initially received a lukewarm reception, especially in the popular press. But after the turn of the century, eugenics caught on. Between 1890 and 1930, eugenics movements of some form developed in over 30 countries, encompassing Central and South as well as North America, and Asia as well as Europe (Adams 1990). The greatest impact, however, was felt in the United States, Canada, Britain, Germany, and Scandinavia—predominantly Protestant countries in which Weismann's 'hard' view of heredity prevailed.

Eugenics societies drew most of their members from the professional middle classes: doctors, lawyers, teachers, social workers, ministers. Numerous demographic studies seemed to demonstrate that workers and agricultural laborers had more children than the middle and upper-middle classes, with those living in the worst conditions producing the largest families (actual patterns of fertility are a matter of dispute; see Szreter 1996).

Of particular concern was the ostensibly rapid increase in the number of 'feebleminded.' In 1910, the psychologist Henry H. Goddard proposed a three-tiered grading system for mental deficiency. Those with a mental age of one or two he termed 'idiots,' those with a mental age of three to seven 'imbeciles,' and those with a mental age between nine and 12 'morons' (Zenderland 1999). Members of the 'high-grade feebleminded' or 'moron' class posed the greatest threat to society since they could not easily be differentiated from normal persons. Unlike idiots and imbeciles, they were often physically fit and attractive, and thus had no trouble either finding mates or producing children. But lacking foresight and self-restraint, and unable to distinguish right from wrong, they easily fell into lives of pauperism, alcoholism, sexual promiscuity, and crime. Exhibits at museums, expositions, and state fairs described the disastrous consequences of allowing reproduction by these defectives to go unchecked.

On the basis of pedigree studies demonstrating the passage of traits in several generations of the same family (such as Goddard's 'Kallikaks') feeblemindedness was judged to be heritable. According to the best science of the day, feeblemindedness resulted from a single recessive gene. It thus initially seemed that the trait could be eliminated if its possessors were prevented from breeding. By the 1920s, that optimistic assessment was shown to be wrong, since most of the offending genes would be hidden in apparently normal carriers. Policies such as eugenic segregation or sterilization, which targeted only those affected, would thus work much more slowly than had been thought. But most eugenicists reasoned that even a small reduction was better than none (Paul and Spencer 1995).

Modern scholarship has challenged the once-conventional identification of eugenics with political conservatism and support of the social status quo. Many eugenicists were indeed reactionary. But eugenics was also deployed in support of a variety of progressive causes. In China, eugenics flourished during the Republican Era, where it was invoked by modernizers in their efforts to counter tradition (Dikötter 1998). In Europe and North America, pacifists claimed that war destroyed the strongest and bravest of men, thus weakening the race (Crook 1994). Although many Fabian and even some Marxian socialists worried that the worst were outbreeding the best, they turned this concern into a case for social and economic change, arguing that only in a society that abolished economic inequality would it be possible to distinguish genetic value from environmental good fortune (see Kevles 1995, Paul 1998, Grossmann 1995). These Left eugenicists criticized the racial and class bias that informed the 'mainstream' eugenics movements. In their view, the real problem was the high reproductive rates among the least prudent and foresighted of every class.

While many eugenicists were socially conservative, sex reformers invoked eugenics in support of access to birth control, legalization of divorce and abortion, and radical changes in gender roles, including expansion of educational and employment opportunities for women (who were often drawn to the eugenics cause; for example, nearly half the members of the British Eugenics Education Society were female: see Soloway 1990, p. 1280, Hall 1998; on Weimar Germany see Grossmann 1995; on Spain see Cleminson 1994). Birth-control advocates argued that access to contraception would allow poor women, of inferior heredity, to voluntarily limit their births. According to Margaret Sanger, 'funds that should be used to raise the standard of our civilization are diverted to maintenance of those who should never have been born.' (Opponents countered that contraception would be used disproportionately by the middle-class.) Even 'free love' or abolition of marriage was promoted on

the grounds that marriage was often based on class or other genetically-irrelevant factors (see Paul 1995, pp. 91–2, Hansen 1996).

2.5 Eugenic Measures

Until recently, studies of eugenics focused almost exclusively on Britain and the US. Since the 1980s, however, the focus has greatly broadened to include Canada, Latin America, Germany, Russia, France, Spain, Scandinavia, Australia, China, and Japan, among other countries and regions. As a result, historians now understand that eugenics took quite different forms in different places. In the Latin countries, where Lamarckianism was strong, race-mixing common, and Catholicism the dominant religion, concern with biological degeneration did not result in calls for sterilization of defectives. In France, the 'puericulture' movement, spearheaded by Adolphe Pinard, stressed the value of breast feeding and of improved diet and more rest for pregnant women, as well as a premarital examination law to prevent dysgenic unions among those individuals with hereditary defects. In these countries, environmental and hereditary improvement were generally thought to be mutually reinforcing, rather than at odds (on France, see Schneider 1990; on Latin America, see Stepan 1991).

Even in North America and the Protestant countries of Europe, eugenics encompassed a wide range of activities. In Britain, where the modern movement originated, there was no effort to pass coercive legislation (which would have been politically impractical, in part due to the strength of the Labour Party, which viewed eugenics as hostile to workers). Following Galton's lead, the Eugenics Education Society emphasized the need to convince their compatriots of the value of controlling reproduction. Thus, eugenics consisted primarily of education and propaganda, such as the Pauper Pedigree Project, which compiled family trees of typical degenerate families (Mazumdar 1992).

Coercive measures were adopted elsewhere. Beginning in 1907, when Indiana authorized the compulsory sterilization of 'criminals, idiots, rapists, and imbeciles,' sterilization laws were passed in over 30 American states, two Canadian provinces, all the Scandinavian countries, Estonia, Germany, and Japan. In the US, opposition from the Catholic Church, concern that the procedure (vasectomy in men, tubal ligation in women) would promote sexual promiscuity and venereal disease, and especially legal challenges to the constitutionality of state laws, originally limited the number of sterilizations (Reilly 1991). To address the legal issues, Harry Laughlin, Superintendent of the Eugenics Record Office at Cold Spring Harbor, New York (the center of the organized eugenics movement in the US), drafted a model law. Its use by the state of Virginia was tested in the 1927 case of *Buck v. Bell*. In his famous decision in that case, Justice Oliver Wendell Holmes wrote: 'It is better for all the world if, instead of waiting to execute degenerate offspring for their crime, or to let them starve for their imbecility, society can prevent those who are manifestly unfit from continuing their kind … . Three generations of imbeciles is enough.' In the aftermath of the case, a dozen states revised existing laws or enacted new ones. (However, more than half the 60,000 legal sterilizations were performed in California.)

In the US, concerns about deterioration had been heightened by the results of the mental testing program initiated by the army during World War I. Psychologists tested more than 1.75 million recruits and concluded that nearly half the white draft was feeble-minded, with blacks and recent immigrants scoring worst. The testing program thus exacerbated anxiety about the quality of immigrants from Southern and Eastern Europe, whose numbers increased dramatically beginning in the 1880s. The newcomers from Russia, Poland, Italy, and the Balkans were anyway resented for their cultural and linguistic differences, support for big-city political machines, and willingness to work for low wages. Eugenicists added that they were also of poor stock, and that mixing with older-stock Americans would produce biological degeneration. In the first decade of the century, Theodore Roosevelt had warned repeatedly that old-stock Americans were losing a desperate 'war of the cradle' with lesser breeds, a theme that recurred frequently in both academic and popular literature. In 1924, President Calvin Coolidge signed into law the Immigration Restriction Act, which restricted entrants from any European country to two percent of the foreign-born of the same national origin recorded in the 1890 census, that is, before the massive influx of new immigrants.

American developments were noted in Germany, where a draft law permitting sterilization with the consent of the person involved or their guardian had been drafted at the end of the Weimar Republic. In *Mein Kampf*, Adolf Hitler had characterized National Socialism as 'applied biology.' The Nazi government quickly passed a compulsory law, which extended to a wide range of ostensibly genetic conditions. The law was often praised by non-German commentators for its procedural safeguards (although these were often disregarded in practice). Approximately 400,000 Germans were eventually sterilized under the Nazi regime (Bock 1986, pp. 230–46).

In 1939, the Nazis instituted a euthanasia program aimed at eliminating the 'useless eaters' in mental hospitals. Neither the sterilization nor euthanasia policies were racially-oriented. But soon after the Nazis came to power, they enacted the first of the Nuremberg laws, which forbade Germans to marry Jews and others of 'inferior blood.' Other actions targeting Jews, Gypsies, the offspring of German

mothers and black French soldiers, and other reviled groups followed, culminating in the Holocaust. The Nazis also instituted a number of 'positive' eugenics measures, including the *Lebensborn* program, which allowed single and married women who passed a racial test to give birth in special SS maternity homes.

Events in Germany were watched closely in Scandinavia, where many eugenicists were critical of Nazi racism, but impressed by the sterilization law. Denmark, Sweden, Norway, Finland, and Iceland passed laws of their own, virtually without dissent (Broberg and Roll-Hansen 1996). Support for sterilization reflected fears about degeneration that were shared by elites elsewhere in Europe and in America. It also reflected the Social Democratic emphasis on communal responsibilities, as well as the view that a state that has taken on the responsibility of caring for its citizens must limit the numbers of those who will burden it.

Opposition to compulsory sterilization did not necessarily imply wholesale opposition to eugenics. In fact, most critics took for granted that mentally deficient individuals should be persuaded or prevented from breeding. The issue was how best to achieve this end. Skeptics variously charged that the laws promoted promiscuity, were ineffective, were based on unfounded assumptions about the heredity of mental defect or criminality, and were inevitably biased in their application. A common concern was that sterilization would provide a cheap substitute for the more desirable policies of segregation, training, and community supervision.

Even the Catholic Church—the most powerful foe of sterilization—was not opposed to eugenics *tout court* (much less committed to the view that procreation is a private matter). In his 1930 encyclical *Casti Connubi*, Pope Pius XI conceded that eugenics must 'be accepted, provided lawful and upright methods are employed within the proper limits,' although he denounced sterilization as an illegitimate usurping by the state of power over the bodies of its subjects. In the same encyclical, the Pope condemned contraception and abortion, and noted that private individuals are also barred from tampering with their bodies in order to prevent procreation.

3. *The Decline of Eugenics*

In the aftermath of World War II, eugenics fell from favor. Historians once attributed its decline in large part to the progress of genetics. In the conventional view, which still informs much popular writing on the subject, it became increasingly clear that traits were usually the product not of one but of several genes, that these genes acted in complex ways with each other and with the environment, and that selection against undesirable traits could not in fact eliminate them. As a result, support for programs of selective breeding inevitably waned. From this standpoint, eugenics was unable to withstand its encounter with scientific fact. However, most recent historiography emphasizes that the scientific developments said to have eroded support for eugenics occurred in the 1910s, long before it became disreputable, and that the researchers responsible for these discoveries were themselves eugenics enthusiasts.

Eugenics' decline seems to be linked most closely to a shift in the social climate resulting from revelations of Nazi atrocities—crimes that had been justified by the ostensible biological need to protect and purify the German race. The result was a backlash against eugenics, including its underlying hereditarianism. In the 1950s, genetic explanations of individual and group differences fell from favor. Although the reign of 'cultural determinism' was short-lived (Nelkin and Lindee 1995, pp. 34–7), eugenics' reputation never recovered. It was now equated with its most barbarous manifestations, viewed as inherently racist and reactionary, as well as scientifically shoddy.

In fact, not all eugenics was racist. Even in Germany, the movement originally had been dominated by technocratic elitists, not racial purists, and some prominent eugenicists were Jewish (Weiss 1988, Weingart et al. 1988, Weindling 1989). In the US, the subjects of the pedigree studies, such as Goddard's 'Kallikaks,' were all white and Anglo-Saxon (Rafter 1988). Nor was eugenics invariably supportive of the status quo, as its popularity among Fabian socialists, architects of the Scandinavian Social Democratic welfare states, and a host of sex reformers indicates. Moreover, many eugenicists, especially in Britain, opposed coercion. Following Galton's lead, they instead supported educational and propagandistic measures. Thus racism, class-bias, social conservatism, and violations of individual liberty—while prominent aspects of the history of eugenics—were not constitutive of it (Wikler 1998). However, the new historiography, which stresses eugenics' complexity and diversity, has so far had little impact on popular perceptions of the subject.

See also: Darwinism: Social; Ethical Issues in the 'New' Genetics; Eugenics as an International Movement; Eugenics as the Basis of Population Policy; Eugenics in Asia; Eugenics in Europe; Evolutionism, Including Social Darwinism; Galton, Sir Francis (1822–1911); Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Engineering; Genetic Screening for Disease-related Characteristics; Genetic Testing and Counseling: Educational and Professional Aspects; National Socialism and Fascism; Nazi Law; 'New' Eugenics, The: Role in Society

Bibliography

Adams M B (ed.) 1990 *The Well-Born Science: Eugenics in Germany, France, Brazil, and Russia*. Oxford University Press, New York

Bock G 1986 *Zwangsterilization im Nationalsozialismus*. Westdeutscher, Opladen, Germany
Broberg G, Roll-Hansen N (eds.) 1996 *Eugenics and the Welfare State: Sterilization Policy in Denmark, Sweden, Norway, and Finland*. Michigan State University Press, East Lansing, MI
Cleminson R 1994 Eugenics by name or nature? The Spanish anarchist sex reform of the 1930s. *History of European Ideas*. **18**: 729–40
Cornford F M 1941 *The Republic of Plato*. Oxford University Press, London
Crook P 1994 *Darwinism, War, and History*. Cambridge University Press, Cambridge, UK
Dikötter F 1998 *Imperfect Conceptions: Medical Knowledge, Birth Defects and Eugenics in China*. Columbia University Press, New York
Grossmann A 1995 *Reforming Sex: The German Movement for Birth Control and Abortion Reform, 1920–1950*. Oxford University Press, Oxford, UK
Hall L 1998 Women, feminism, and eugenics. In: Peel R A (ed.) *Essays in the History of Eugenics*. Galton Institute, London
Hansen B F 1996 Something rotten in the state of Denmark: Eugenics and the ascent of the welfare state. In: Broberg G, Roll-Hansen N (eds.) *Eugenics and the Welfare State: Sterilization Policy in Denmark, Sweden, Norway, and Finland*. Michigan State University Press, East Lansing, MI
Jones G 1998 Theoretical foundations of eugenics. In: Peel R A (ed.) *Essays in the History of Eugenics*. Galton Institute, London
Kevles D J 1995 *In the Name of Eugenics: Genetics and the Uses of Human Heredity*. Harvard University Press, Cambridge, MA
Mazumdar P M H 1992 *Eugenics, Human Genetics, and Human Failings: The Eugenics Society, Its Sources and Its Critics in Britain*. Routledge, London
Nelkin D, Lindee M S 1995 *The DNA Mystique: The Gene as Cultural Icon*. Freeman, New York
Paul D P 1984/1998 Eugenics and the left. In: *The Politics of Heredity: Essays on Eugenics, Biomedicine, and the Nature–Nurture Debate*. SUNY, Albany, NY
Paul D B 1995 *Controlling Human Heredity: 1865 to the Present*. Humanity Press, Amherst, NY
Paul D P, Spencer H G 1995 The hidden science of eugenics. *Nature* **374**: 302–4
Rafter N H (ed.) 1988 *White Trash: The Eugenic Family Studies, 1877–1919*. Northeastern University Press, Boston
Reilly P R 1991 *The Surgical Solution: A History of Involuntary Sterilization in the United States*. Johns Hopkins University Press, Baltimore, MD
Riddle J M 1992 *Contraception and Abortion from the Ancient World to the Renaissance*. Harvard University Press, Cambridge, MA
Roper A G 1913 *Ancient Eugenics*. Oxford University Press, Oxford, UK
Schneider W 1990 *Quality and Quantity: The Quest for Biological Regeneration in Twentieth-century France*. Cambridge University Press, Cambridge, UK
Soloway R 1990 *Demography and Degeneration: Eugenics and the Declining Birthrate in Twentieth Century Britain*. University of North Carolina Press, Chapel Hill, NC
Stepan N L 1991 *The Hour of Eugenics: Race, Gender, and Nation in Latin America*. Cornell University Press, Ithaca, NY
Szreter S 1996 *Fertility, Class and Gender in Britain 1860–1940*. Cambridge University Press, Cambridge, UK
Weindling P 1989 *Health, Race, and German Politics between National Unification and Nazism*. Cambridge University Press, Cambridge, UK
Weingart P, Kroll J, Bayertz K 1988 *Rasse, Blut und Gene. Zur Geschichte der Eugenik und Rassenhygiene in Deutschland*. Suhrkamp, Frankfurt, Germany
Weiss S F 1988 *Race, Hygiene, and National Efficiency: The Eugenics of Wilhelm Schallmayer*. University of California Press, Berkeley, CA
Wikler D 1998 Eugenic values. *Science in Context* **11**: 455–70
Zenderland L 1999 *Measuring Minds: Henry Herbert Goddard and the Origins of American Intelligence Testing*. Cambridge University Press, New York

D. Paul

Eugenics in Asia

Eugenics, a term coined in 1883 by the English scientist Francis Galton, had its heyday during the first half of the twentieth century. Its proponents claimed that breeding principles such as assortative mating and artificial selection could improve the quality of human offspring. 'Positive eugenics' would ensure that individuals with above-average abilities would breed at a higher rate than ordinary people. 'Negative eugenics' would restrict the reproduction of inferior people, while those having subnormal abilities would have to be physically prevented from perpetuating their 'infirmities.' Mainline eugenicists assumed that intellectual capacity and behavioral traits were inherited and could not be significantly enhanced by education. Supporters included not only scientists, in particular geneticists, but also social reformers, journalists, and other members of the emerging liberal professions. Until recently, however, it was often assumed that eugenics only appealed to extreme right-wing thinkers or organisations in Germany or the United States, as the multifarious dimensions and extraordinary appeal of eugenics to individuals of very different social backgrounds, political convictions, and national affiliations was ignored.

Far from being a politically conservative and scientifically spurious set of beliefs which remained confined to the Nazi era, eugenics belonged to the political vocabulary of many modernizing forces between the two World Wars. It was not only supported by scientific societies, pressure groups, and political institutions in such different countries as France, Brazil, and Sweden, but also developed a large following in Asia, in particular in China and Japan.

1. *China*

Eugenics in Asia is sometimes dismissed as a derivative manifestation of a more 'authentic' dis-

course in 'the West,' a misleading interpretation which can only impoverish our understanding of the complexities of cultural and intellectual encounters between population groups around the world. When the history of eugenics is explored in a variety of cultural, social, and political contexts, it resists any reductive explanation, an observation which is all the more true as the less familiar shores of China and Japan come under investigation. As in European cultures, popular hereditarian notions—encapsulated in such sayings as 'Like begets like'—helped the emergence and acceptance of eugenic ideas in Asia. Well before Darwin's time, experts in China had attained a sophisticated level of knowledge about selection and breeding, particularly in botany and in goldfish. On the other hand, many medical publications discussed in great detail ways of improving the physical quality of a line of descent while also highlighting the medical dangers to future offspring of careless reproduction. Patrilineal culture, in which reproduction was represented as a potentially dangerous process that should be carefully regulated in order to enhance the possibilities of engendering healthy male descendants, permeated traditional medical literature in imperial China. The individual, seen in relationship to the lineage, was held to be responsible not only for his or her own reproductive behavior, but also for the health of future offspring.

Interest in eugenics, however, only appeared in China in the wake of the reform movement of the 1890s, which openly sought to challenge imperial institutions and orthodox ideology. In search of wealth and power after the country's disastrous defeat against Japan in 1894–5, reform-minded scholars used Darwinian theories—inspired by Herbert Spencer and T. H. Huxley—to propose a radically new political vision based on faith in progress. Slogans such as 'survival of the fittest' and 'struggle for existence' became popular in China, increasingly represented as a 'race' which ought to be strengthened in its confrontation with imperialist countries. Public health, individual hygiene, even sexual behavior were targeted by reformers such as Kang Youwei and Liang Qichao as areas in need of state regulation in order to ensure the 'survival of the race' in a context of international competition. Only after the collapse of the empire in 1911 and its substitution by a republic, however, did eugenics develop a much wider following.

Eugenics in republican China (1911–49), as in other countries during the interwar period, was not so much a clear set of scientific principles, as a 'modern' way of talking about social problems in biologizing terms: politicians with mutually incompatible beliefs and scientists with opposed interests could all selectively appropriate eugenics to portray society as an organic body which had to be guided by biological laws. Eugenics gave scientific authority to social fears and moral panics and provided an overarching rationale for welfare policies in China. Powered by the prestige of science, it allowed modernizing elites in China to represent their prescriptive claims about social order as objective statements irrevocably grounded in the laws of nature. Eugenics promoted a biologising vision of society in which the reproductive rights of individuals were subordinated to the rights of an abstract organic collectivity.

Eugenic discourse had different meanings to different social groups in different contexts: it was appropriated and used in a variety of ways to accommodate radically divergent purposes and interests, being flexible enough to unite individuals or constituencies with contradictory agendas. One of the major dividing lines within eugenic discourse was that between a hard approach to heredity, based on genetics in Britain and Germany, and a soft approach combining an emphasis on the environment with hereditarian explanations.

The soft approach to eugenics was common in republican China as well as in other countries as diverse as Brazil and France. Professional groups in China campaigned for better public education in social hygiene and sexual health in the name of race improvement. Moreover, unlike researchers in countries which pursued a hard Mendelian approach, supporters of eugenics in China did not produce significant biological research or statistical studies. An emphasis on the virtues of education accounts for the relative absence in republican China of formal institutions, official organs, or professional organisations centerd around the promotion of eugenic ideals. Rather than emanating from a solid organisational foundation, eugenic discourse was supported by a variety of voices in the social field: it attracted popular journalists, social reformers, medical writers, sex educators, university professors, and political ideologues, all attempting to promote medical knowledge and reproductive health for the sake of a more eugenic nation.

Eugenics was actively supported by members of the professional classes, in particular intellectuals, university teachers, sociologists, and biologists, and cut across most political positions, from the fascist core of the Guomindang to the communist theories of Chen Duxiu. Cheap textbooks on heredity and genetics explained to the public the dangers of racial degeneration. Primers, self-study manuals, cheap pamphlets, and 'ABC' introductions to mainline eugenics were published throughout the 1920s. Hereditary principles, endowing sex with a biological responsibility for future generations, were thought to further highlight the need for social discipline. In the 1930s, eugenic discourses became increasingly common in medical circles, 'degeneration' and 'racial hygiene' being the catchwords of the day. There is little evidence, however, to show that the widespread interest in eugenic ideas was translated into actual practice before the advent of a socialist regime, although some isolated cases, such as the execution of

lepers by the police in Guangdong province in the 1930s, point to a more radical approach in questions of public health during the Nanjing decade.

The Second World War, followed by the civil war between the Communist Party and the Guomindang from 1945 to 1949, was the main reason why eugenic ideas did not achieve institutional form. The Committee for the Study of Population Policies, organized by the Ministry of Social Affairs in 1941, was the first official attempt to approach eugenics systematically. It recommended the segregation of physically and mentally handicapped persons from the normal population for what was called 'cultural advancement and racial rejuvenation.' Because people were recognized as being unequally endowed, the report advocated a differential birth-rate: 'Thus viewed,' said the report, 'some individuals may have children, others not.' The committee—whose members included Chen Da and Pan Guangdan—also encouraged the use of sterilization for the racial rejuvenation of the country and recommended that '(s)teps should be taken to segregate persons of hereditary defects, physical or mental, from the normal population' (Dikötter 1998).

After the communist victory in 1949, the class bias of eugenics was systematically condemned, while genetics was denounced as 'bourgeois science.' As in the Soviet Union, Lysenko's doctrine became dominant in the 1950s and 1960s while Mendelian laws of inheritance and T. H. Morgan's chromosome theory were rejected for ideological reasons. This trend was only reversed after the accession of Deng Xiaoping to power in 1978. Medical experts and population specialists were put into powerful positions as official policies aimed at the limitation of births were initiated by the government. Besides the one-child family program, by which the government only in exceptional circumstances allows parents to have more than one or occasionally two children, an important component of these policies has been the improvement of the quality of newborn babies.

The one-child family policy has effectively prepared the terrain for a better acceptance of eugenic legislation among large sectors of the urban population: many families which are allowed to have one child only are keen to avoid 'defective' births themselves. A prolific medical discourse has responded to the general public's concern for healthy offspring, a virtuous child, or even a genetically improved line of descent. Numerous pamphlets of scientific vulgarization thus dispense advice on the art of engendering a prodigy child, suggesting that the eugenist vision of the government is shared to a great extent by a population which is anxious to avoid what are called 'inferior births.'

As the result of such consensus, local regulations have been passed virtually unopposed in a number of provinces and cities since the late 1980s. On November 25, 1988, the Standing Committee of the National People's Congress of Gansu Province passed the country's first law prohibiting mentally retarded people from having children. According to the law, those mentally retarded people whose condition was either inherited or a consequence of marriage between close relatives were not allowed to have children; no mentally retarded people were allowed to get married unless they had undergone sterilization surgery; those who had married before the promulgation of the law were also to undergo sterilization surgery; and pregnant women suffering from mental retardation were required to have their pregnancies terminated. Individuals who violated the law by allowing mentally retarded people to have children would be punished by both administrative and economic means. A similar law was enacted by Liaoning province in February 1990, while Zhejiang province passed a law on the preservation of 'eugenic health' in June 1992 to combat the increasing incidence of 'cretins,' who were estimated to number more than 300,000. The Henan eugenics law of 1992 requires that if one partner in a married couple suffers from a 'chronic mental disorder' such as schizophrenia or manic depression, that partner shall be sterilized. The implementation of such sterilization laws at the provincial level is advanced by a system of incentives and disincentives. People who refuse to be sterilized or anyone who impedes the sterilization of another may be subject to penalties. For example, Henan's eugenics law requires prompt reporting of those diagnozed as 'needing to be sterilised' who refuse to accept the operation. Differential fines for noncompliance with 'eugenic measures' are imposed for rural and urban residents.

Under pressure from a variety of lobbies, mainly family planning experts and geneticists, the provincial laws finally culminated in the People's Republic of China passing eugenic legislation at the national level at the beginning of 1995. The law aims to prevent 'inferior births' from becoming a burden on the state and society. Renamed Maternal and Infant Health Law after protests against a preliminary draft entitled Eugenics Law (*Youshengfa*), it supports the systematic 'implementation of premarital medical check-ups' in order to detect whether one partner in a couple suffers from a 'serious hereditary,' venereal, or reproductive disorder, a 'relevant mental disorder,' or a 'legal contagious disease': it asserts that those 'deemed unsuitable for reproduction' should be urged to undergo sterilization or abortion, or to remain celibate, in order to prevent 'inferior births.' During pregnancy, medical testing will be compulsory and can be followed by termination if the foetus is found to have a serious disorder.

In the absence of any detailed studies by social-anthropologists and population experts, concrete information on the implementation of eugenic laws is missing. Foreign observers, however, have underlined that the implementation of eugenic legislation not only undermines the rights of the person vis-à-vis the state, but it is based on controversial if not antiquated

theories of human heredity (O'Brien 1996, Lancet 1995, Nature 1994).

2. Japan

In Japan, eugenics developed gradually after modernizing intellectuals encountered evolutionary theories in the works of Charles Darwin and Herbert Spencer in the 1880s. In the wake of the Meiji revolution (1868), science and technology were perceived to be the key to national defence in an era of intense imperialism. As in many other countries, racial improvement became a popular notion among leading intellectuals who viewed the world in terms of a 'struggle for existence' among competing 'races.' Belief that individual reproductive decisions should be controlled by the state in the name of collective health flourished during the first half of the twentieth century. In the 1910s, works on eugenics by Umino Kōtoku, Nagai Hisomu, and Saitō Shigesaburō appeared, closely related to an increased interest in birth control, sex education, and public health.

Specialized journals were also established in the 1920s, both the *Yūzenikkusu* (Eugenics) and the *Yūsei Undō* (Eugenics Movement) being crucial in spreading eugenic discourse to a wider audience. As in China, belief in the power of heredity led many followers of the eugenics movement to view venereal disease, tuberculosis, leprosy, and alcoholism as heritable conditions which should be strictly administered by the state in order to prevent the increase of 'inferior offspring.' Eugenics became part of such widely discussed themes as evolution, degeneration, civilization, and modernity, and touched on a wide variety of emerging fields like maternity, psychiatry, criminology, public health, and sex education.

In Japan, too, the orthodox Mendelian roots of eugenics were often disregarded in favor of a more flexible approach to eugenics which emphasized belief in the transmission of acquired characteristics. Contrary to China, however, eugenic ideas provided a rationale for welfare policies with deep institutional and legal ramifications. A first attempt to implement eugenic legislation took the form of an appeal to the Imperial Diet in the early 1920s (Matsubara 1998). Proposed by the New Women's Federation, the petition sought to legally restrict the marriage of individuals with venereal diseases as a means of strengthening the nation. Despite support from leading intellectuals as well as members of the Imperial Diet, the petition was not adopted. Only in 1934–5 was a Racial Eugenics Protection Bill introduced by the Imperial Diet, which proposed the sterilization of persons with mental and physical diseases deemed to be hereditary, including serious criminals who were claimed to transmit their 'bad traits' to their descendants. The Japanese Association of Racial Hygiene, on the other hand, campaigned for a eugenics law limiting sterilization to patients with hereditary diseases. Headed by Nagai Hisomu, the group drafted a sterilization bill in 1936 based on the German Sterilization Law of 1933. Only in 1940, however, was a National Eugenic Bill passed, its enactment leading to heated controversy among government representatives and members of the medical professions. Objections came both from nationalists defending the need of the state to increase its population and from geneticists questioning the scientific validity of eugenics. While the Bill came into effect a year later, the government was compelled to prevent the enforcement of eugenic sterilization: by the end of the Second World War, only 454 operations were performed under the eugenics law.

After Japan's defeat in the Second World War, questions of 'racial survival' surfaced again in the context of occupation by the Allied Powers (Matsubara 1998). As abortion was legalized, a broad consensus emerged within medical and political circles around the need to prevent not only 'inferior descendants,' but also the reproduction of patients deemed to have 'infectious diseases' such as leprosy and non-hereditary mental disorders in the name of 'eugenic protection.' Supported by conservative advocates as well as socialist politicians, the Eugenics Protection Law (*Yūsei hogohō*) of 1948 gave legitimacy to eugenic sterilisation without the consent of the patients considered. The law remained the foundation for eugenic sterilisations throughout the second half of the twentieth century: involuntary operations reached over 16,500 cases by 1996, while a further 1,500 leprosy sufferers segregated in sanatoriums had been sterilized with consent by 1992. In 1996, the law was finally revised and renamed Maternal Protection Law, while the eugenic clauses on sterilization and abortion were deleted. Even after these legal changes, however, the government refused to critically examine the practice of compulsory sterilization on eugenic grounds, despite the severe criticisms from a variety of civil rights groups (Norgren 1998).

3. Conclusion

The legacy of eugenics is still highly visible today in many countries, including in the less researched parts of the world outside 'the West.' As the examples of China and Japan demonstrate, eugenic ideas elaborated during the first half of the twentieth century continue to carry scientific and legal credibility. While crude eugenic programs sponsored by the state may have become less frequent today, eugenic ideas are prominent in contemporary debates about new medical techniques for isolating and manipulating genes. During the heyday of eugenics, assumptions about 'race,' gender, and class pervaded genetic research: with the revolution in human genetics since the end of

the twentieth century, social and political factors continue to exert formidable influences on scientific research. Eugenic laws have been passed at the provincial and national level in China, a country which has curbed the reproductive freedoms and civil rights of individuals on behalf of the 'public good.' Even in democratic countries like Japan, however, marginalized people may be treated in a discriminatory way, as social prejudice and economic interest influence the genetic information made available to families, employers, hospitals, insurance companies, or welfare systems. As much as 'ethnicity' is a concept which is sometimes used as a thinly disguised substitute for the outdated notion of 'race,' discredited eugenic ideas have thrived on rather than suffered from the revolution in human genetics in some parts of the world: East Asia is no exception.

See also: Darwinism: Social; Eugenics as an International Movement; Eugenics as the Basis of Population Policy; Eugenics, History of; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease-related Characteristics; 'New' Eugenics, The: Role in Society

Bibliography

Dikötter F 1998 *Imperfect Conceptions: Medical Knowledge, Birth Defects and Eugenics in China*. Columbia University Press, New York

Matsubara Y 1998 The enactment of Japan's sterilization laws in the 1940s: A prelude to postwar eugenic policy. *Historia Scientiarum* **8**(2): 187–201

Nature 1994 China's misconception of eugenics. **367**: 1–2

Norgren T 1998 Abortion before birth control: The interest group politics behind postwar Japanese reproductive policy. *Journal of Japanese Studies* **24**: 59–94

O'Brien C 1996 China urged to delay 'eugenics' law. *Nature* **383**: 204

Sumiko O, Bartholomew J R 1998 Eugenics in Japan: Some ironies of modernity, 1883–1945. *Science in Context* **11**(3–4): 545–65

Suzuki Z 1975 Geneticists and the eugenics movement in Japan. *Japanese Studies in the History of Science* **14**: 157–64

The Lancet 1995 Western eyes on China's eugenic law. **346**: 131

F. Dikötter

Eugenics in Europe

Eugenics were practiced as early as in the Classical Greek era. Spartans were known for throwing abnormal newborns into the Kaiada, a wild ravine. Plato in the *Republic* (459D) stated that rulers should permit the best men to mate with the best women for procreation, while in the *Laws* he compared animal breeding to that of humans, rulers playing the part of shepherds in the amelioration of their breeds. Inspired by Plato, Tomasso Campanella wrote of a eugenic utopia, the *City of the Sun* (1623). Actually, Francis Galton (1822–1911), a cousin of Charles Darwin and a man of many talents (explorer, anthropologist, statistician, student of heredity), was the modern founder of eugenics. This term derived from the Greek 'eugenes,' meaning 'from good birth.' Galton first spoke of eugenics in 1883, one year after Darwin's death, defining it as the study of the elements, controlled by society, which may improve the physical or mental racial qualities of future generations.

Galton started with the study of human intelligence. In two articles written in 1865 and expanded into a book, *Hereditary Genius*, in 1869, he investigated the hereditary origin of mental ability by accumulating data on important men and their relatives. Darwin was impressed by his half-cousin's work and wrote: 'We now know, through the admirable labours of Mr. Galton that genius … tends to be inherited.' However, Darwin was ambivalent on eugenics. While being convinced from Galton's writings on the inheritance of intelligence, he did not advocate a programme of genetic amelioration. He thought that one of the higher qualities of humans is 'sympathy,' namely sentimental closeness to other human beings, especially those suffering, and this was contrary to all attitudes of 'social Darwinism,' which aimed at the improvement of society by letting the laws of natural selection operate against the poor and the needy. Darwin never promoted or approved of such social policies. Eugenics and social Darwinism achieved their prominence after Darwin's death.

1. England

Galton assembled data indicating that many human characteristics, physical as well as social (ranging from pauperism to mental illness), were inherited. In the International Health Exhibition of 1884 Galton established an Anthropometric Laboratory and assembled metrical data. He also remarked that, in England, individuals with lower intelligence seemed to have more offspring than those with higher intelligence and thus in future generations intelligence (being inherited) was condemned to decline. Lower classes were formed by more fertile but less intelligent individuals. Adequate measures would become necessary to prevent the decline of society. Darwin noted that higher fertility combined in lower classes with higher mortality, so a population's intelligence would not change, but he did remark on cases (as in ancient Greece and in Spain) where, according to him, such a

reduction took place and explained the decline of these powerful nations.

We should distinguish two different sorts of social Darwinism, the holistic and the individualistic. The individualistic considers the competition between individuals within a society or population, while the holistic the competition between races, states, and societies. The first is linked to proposals for eugenic measures within society, the second to racism.

Two students of Galton continued his tradition of combining anthropology and eugenics, the biometricians Karl Pearson (1857–1936) and W. F. R. Weldon (1860–1906). Pearson became Professor at University College, London, and headed the Galton Eugenics Laboratory (which included the Eugenics Record Office) and the Biometrical Laboratory. Both Pearson and his friend Weldon were Darwinists and eugenicists but anti-Mendelians. In 1907 the Eugenics Education Society was founded, with seven branches in different English towns plus one in Australia. Galton was convinced to accept the honorary presidency, while the son of Darwin, Major Leonard Darwin, became president (from 1911 to 1928). Many important scientists, professors (among them John Maynard Keynes, later famous as an economist), social radicals as well as conservative public figures, such as George Bernard Shaw, Havelock Ellis, H. G. Wells, Harold Laski, Winston Churchill were in favor of eugenics or became active members of the Society. In 1926 it was renamed The Eugenics Society. In 1912 the First International Congress of Eugenics was organized in London with the participation of Societies or representative committees from Sweden, Norway, Russia, Switzerland, Germany, Poland, Italy, France, and other countries.

The history of the eugenic movement in England is connected strongly with that of the USA; they behaved as communicating vessels. The concern caused by the greater fertility of individuals with lower mental capacities was great, especially after observations were reported suggesting that the majority of recruits for the Boer War did not perform well in mental tests. The French psychologist, Alfred Binet, invented a test, which was perfected by his colleague Theodore Simon. The Binet–Simon test was used to estimate the 'intelligence quotient' (IQ) and investigate its inheritance. In 1913, with an overwhelming majority (only three votes were cast against it, among them that of the liberal Josiah Wedgwood), the British House of Commons approved the Mental Deficiency Bill, in which the different categories of feebleminded were recognized. Measures of segregation were compulsory only for extreme cases, to avoid their procreation. Sterilization was not mentioned. English society was reluctant to impose extreme eugenic measures. England did not include minorities (as later was the case in Germany) or immigrants (as was the case in the USA). The concern was for the urban poor strata, the middle and upper classes found repulsive any extreme repressive measure. Leonard Darwin avoided including them in discussion. Catholics were opposed to eugenic action as were the Labour party, Liberals, and Conservatives (Churchill was afraid that eugenic measures could undermine population growth), while the eugenic group was not homogeneous. Pearson never joined the Society, which soon became dominated by Mendelians. Important scientists, such as the American geneticist T. H. Morgan and the English scholars J. B. S. Haldane, Julian Huxley, and Lancelot Hogben, were opposed to extreme measures.

On the other hand, socialists and members of the Fabian society tended to be favorable to eugenic practices. In 1926, Pearson started the journal *Annals of Eugenics*. Two International Congresses of Eugenics were held in New York, the second in 1923 and the third in 1932, the latter attracting a very small audience. In 1933 Ronald Fisher (1810–1962), a mathematical geneticist and statistician, who demonstrated that Mendelian genetics was compatible with Darwinian selection, even essential to explain the evolutionary process, was appointed to the Galton Eugenic Professorship, succeeding Pearson. Fisher was a strong supporter of eugenics, and a major part of his epoch-making book, *The Genetical Theory of Natural Selection* was devoted to eugenics.

Lionel S. Penrose, a polymath physician and human geneticist, who embarked on the study of the genetics of mental deficiency, succeeded Fisher in 1945 to the Galton Eugenics Professorship. Penrose soon realized that in a few cases of mental deficiency, the genetic etiology was clear and due to a simple gene. In many cases, however, there was an interaction between environmental influences and genetic predispositions. Penrose was critical of the work of the psychologist Cyril Burt on the inheritance of IQ. He disliked eugenics. Penrose might be considered as a turning point in the shift from eugenics to human genetics. Actually, in 1954 he changed the title of *Annals of Eugenics* to *Annals of Human Genetics* and dropped the word Eugenics from the name of his chair. This shift was strengthened by the postwar revelation of eugenic and racist German atrocities. In 1950, UNESCO issued a strong declaration on the concept of human races and against racism signed by well-known and respected evolutionists such as the Americans Herman Muller, and Theodosius Dobzhansky, and the Englishman, Julian Huxley.

The growth of knowledge about human genetics was impressive. Blood groups, enzyme studies, and the cytological etiology of several abnormal syndromes are many of the subjects that marked the flourishing of this field, first in Britain, later in Europe and the USA. With the development, after World War II, of human population and statistical genetics, genetic counseling became feasible. The advent, in the 1970s/1980s, of molecular techniques made possible early detection of carriers of genes causing serious abnormalities and pathologies, and allowed early

abortions. However, this part of the history, although following naturally from the ideas and preoccupations of eugenics, belongs to a different subject.

2. *France*

The case of France is different. In France, the belief in Lamarckian inheritance of acquired characters (while Darwinism was not widely accepted) influenced the development of eugenics. Two major themes played a decisive role, the depopulation of France, and the degeneration of its population. For the first time, in 1854, deaths exceeded births and, although this was not repeated, in the following years the birth rate steadily declined. This combined with the disastrous outcome of the 1870 Franco-Prussian war, was of great concern to demographers, politicians and the common people. Unlike France, the population of Germany increased. The German general von Moltke was reported to have said that France was losing a battle every day because of its declining birth rate.

Depopulation did not permit the neo-Malthusian movement, active in the last quarter of the nineteenth century, to play an important role. The natalist movement, on the contrary, was very popular and influential (the demographer Jacques Bertillon was one of its founders and its most vocal representative). Natalists preached control of the family but emphasized guidance towards its growth. While neo-Malthusians tended to be socialists and libertarian, natalists tended to be Catholics and in 1920 succeeded in having a law passed banning the sales and advertisement of birth control devices. In 1939, a second law encouraged the creation of large families.

France differed further from England in terms of the development of 'solidarism,' an ideology combining individual freedom with state intervention on behalf of social justice. Cooperation rather than competition was emphasized. This was an alternative to English or American social Darwinism. Important public figures participated in the movement, which was the expression of the official philosophy of the French Third Republic. Leon Bourgeois, ex-prime minister, and member of the Senate, identified with the solidarity movement, and became honorary president of the French Eugenics Society. Another member of the solidarity movement, the biologist Edmont Perrier (1844–1921), became president of the French Consultative Committee to the First Eugenics Congress in London, and later the first president of the French Eugenics Society.

A rather general consensus, popularized in the literature of that time, was that the French population was tending to degenerate. Emile Zola, in his novels about the Rougon-Macquart family, insisted on this trend of decline and decadence. The degeneration was attributed to several causes, of which alcoholism, tuberculosis, and venereal diseases were considered to be the most prominent. An environmental amelioration (in a Lamarckian spirit) could rescue, if not the patients, at least their progeny (since it was believed that all three conditions had a strong hereditary component).

An amateur anthropologist, Georges Vacher de Lapouge (1854–1936), ostracized from the French scientific institutional system, promoted racist and eugenic ideas, observing that a very small number of males of absolute perfection and worthy of perpetuating the race could inseminate all females! Lapouge was more appreciated in the USA and Germany than in his own country.

Adolphe Pinard (1844–1934), professor at the Medical faculty in Paris, and a well-known obstetrician, together with his student, demonstrated that pregnant women having proper shelter and food, and not being obliged to work during the last three months of pregnancy delivered newborns weighing on average 300 g more than those working during these last months. Pinard's aim was to improve the next generation by improving physical conditions and the art of raising children, what he called 'puericulture.' Pinard first distinguished an extrauterine puericulture (the raising of the newborn and the baby) and an intrauterine one (improving the conditions of the pregnant women). Later, in 1899, he extended his concept of puericulture, to apply before procreation. This reflected his concern for the hereditary component. Then in spirit, puericulture joined eugenics. Even so, the emphasis was on education, and supportive of premarital examination. Women and men should choose their mates for procreation; mothers should stay near their children. The teachings of Pinard had great influence in France at the beginning of the twentieth century, his theories were taught in schools, while his book was translated in Romania, Spain, Belgium, and Italy (Latin and Catholic countries).

The French delegation to the First Eugenics Congress in London in 1912 was the starting point for the founding of the French Eugenic Society in December of the same year. Perrrier was elected President; Pinard, Louis Landouzy (dean of the medical Faculty in Paris), and Frederic Houssay (1860–1920, professor at the Faculty of Science) Vice Presidents; Eugene Apert (1868–1940, pediatrician), Secretary, and Lucien March (1859–1933, demographer, head of the Statistique Génèrale), Treasurer. Another important member was Charles Richet (1850–1935), professor of physiology and Nobel laureate for his work on anaphylaxis, and science popularizer, who replaced Landouzy after his death. Richet believed that the influence of the environment was small and to improve the quality of the population he strongly supported negative eugenic measures. The natalist and politician Adolphe Landry, in the same year, stated the contrary position, that sterilization and marriage restrictions were repugnant to the need for liberty and

French individualism. The number of founding members of the Society exceeded 100. The Society was very active before World War I, holding meetings every month and publishing a journal, first by itself, and later as part of the bulletin of the Ecole d'Anthropologie in Paris. During the war, the Society stopped its activities, which were resumed in 1920. Heavy human losses during the war (about a million) affected the thinking of many of its members; thus, after the war the emphasis switched to positive eugenics combined with a Lamarckian outlook. Social hygiene became the prime concern, and institutions dealing with public health were founded and strengthened. The Rockefeller Foundation financed an antituberculosis programme in France. However, some eugenicists did not put aside the idea of a premarital examination and a proposal for such legislation was first raised in 1920 by Georges Schreiber. In 1926 Pinard presented to the Chamber of Deputies a bill on premarital examination. This bill was amended repeatedly (Pinard's version was very mild), resulting in a delaying as its vote. The last amendments and a new bill proposal in the Senate, in 1932, were not voted for. Opposition from the Catholics (especially Edouard Jordan) helped to achieve this demand. A Papal encyclical, issued in the last days of 1930 and dealing with Christian marriage (*Casti Conubi*), condemned eugenic practices. Thus, only under the Vichy regime was the premarital examination implemented, by a decree in December 1942, which was amended the following year. An influential personality in the 1930s was the physician and aristocrat Just Sicard de Pauzoles (1872–1968). He defended class-based eugenics (lower classes were multiplying faster, but their genetic quality was poor) and a method for estimating human genetic capital in monetary terms. He was one of the few eugenicists who survived World War II scientifically.

In the 1930s, the depression and the arrival of numerous immigrants, triggered the expression of racist attitudes. France had a long tradition in this matter. Both monogenists (believing in a single origin of all human 'races' such as Buffon and Gobineau) and polygenists (Voltaire, Le Bon, Taine) expressed racist opinions. Joseph Alfred de Gobineau (1816–82) is considered a prominent racist. Ernest Renant (1823–92) stated that 'a very small quantity of noble blood infused into a population suffices to ennoble it.' Gustave Le Bon (1841–1931) tended to identify the hierarchy of races with those of gender and class, believing that the lower social strata corresponded to primitive man. Lastly, Hippolyte Taine's (1828–93) beliefs could be interpreted as a critique of the enlightment (man does not exist, but *men*) and as revealing racist attitudes. The Dreyfus affair, at the beginning of the twentieth century, displayed the attitude of a significant part of the French society towards those considered alien. In the 1930s the skull index started being replaced by the frequency of ABO blood groups. Immigrants should be accepted, according to some scientists, on the basis of the similarity of the frequency of blood groups in the population of their origin to that of French people.

During the Vichy period, German Nazism influenced France: von Verscheuer and Fischer spoke in Paris on the German racial legislation. Well-known racists were placed in key positions. However, the Eugenic Society was disbanded. An important figure during this period, Alexis Carrel (1873–1944), a physician with a successful career in the USA and Nobel laureate for his work in experimental life-support of organs, in his book, *Man, the Unknown*, supported euthanasia for the insane and those guilty of criminal acts. The Vichy regime agreed to create a foundation, proposed by Carrel and headed by him, dealing with eugenics. After the war, in November 1945, the French provisional government, by an ordinance, upheld the Vichy regime decree on premarital examination.

3. *Germany*

The corresponding term for eugenics in Germany was *Rassenhygiene*. Eugenic thought originated in this country partly from social Darwinism, and partly from a concern to increase the population size of the German nation and partly from a racist ideology related to the concept of the Aryan or Nordic race. Racism had a long tradition in Germany. It was imported from England by George II, king of England and elector of Hanover, who formed a cultural bridge between England and Germany. He founded the University of Göttingen early in the eighteenth century, where during the last third of it Professor Johann Friedrich Blumenbach taught racism. Eugenicists in Germany formed a heterogenous group, some accepting and others rejecting the racist ideology. Those opposing proposed the term *Rassehygiene*, the use of the singular for *Rasse* indicating the entire human population. At the end of the nineteenth century, Germany experienced social problems resulting from its fast industrialization. In this atmosphere of social change, of increased criminality, alcoholism, and prostitution, the German middle and upper classes favoured a new '*Sozialpolitik*' in the form of race hygiene, thus solving social problems in an economical way. Physicians played a major role in this, and they were inspired by the Haeckelian version, the selectionist variety of social Darwinism. The influence of August Weismann caused the Lamarckian alternative (the welfare state) to decline. Two major founding figures must be mentioned at this point, Alfred Ploetz (1860–1940) and Wilhelm Schallmayer (1857–1919). Ploetz studied economics and medicine. He was an enthusiast of Germany's Teutonic past and coined the term *Rassenhygiene* as a German synonym for eugenics. He was concerned with the improvement of the

hereditary quality of the population as well as its increase to an optimal size. He believed in a germ plasm selection—that is, selection before birth—in order to improve the population genetically without resorting to inhumane practices. In 1904 he founded the journal *Archiv für Rassenkunde und Gesellschaftsbiologie* (a journal in which not only leading biologists published but also some racists), and in 1905 the society for Race Hygiene. Schallmayer, in opposition to Ploetz, rejected racist ideology. He was a physician influenced by the writings of Karl Marx, and a Darwinist. Being deceived by the possibility of curing mental diseases, he adopted a eugenic approach, supported marriage restrictions for the insane, the feebleminded, alcoholics and others, but opposing state legislation for them.

The Society for Race Hygiene was founded in 1905 by Ploetz, the psychiatrists Anastasius Nordenholz (his brother-in-law), Ernst Ruedin (1874–1952), and the ethologist Richard Thurnwald (1869–1954). It was dominated by physicians, and formed local groups in Berlin and in Munich. Among the practical measures it proposed, in 1910, were the establishment of a counterbalance to the state protection of the feebleminded and the weak, the isolation of them, and the imposition of marriage restrictions so that they would not procreate as well as the control of immigration from 'inferior elements.' The German people, it seems were not ready to accept mandatory sterilization. Some eugenicists recommended the creation of work colonies, which were probably precursors of the concentration (labor) camps.

In the face of deterioration of the international climate before World War I, in which France and Russia would be allies against Germany, special emphasis was given to birth polity and the rejection of neo-Malthusian proposals. After the war, the Weimar Republic (1918–33) succeeded the Wilhelmine era. During this period, the same preoccupations continued, namely race hygiene, population degeneration, and its size increase. After the death of Schallmayer and the retirement of Ploetz, Fritz Lenz (1887–1976) played a major role. A physician, influenced by Weismann, and a member of the Munich chapter, he became known as the co-author of a treatise on human genetics and eugenics. He was a racist and a sexist, believing that races had physical characteristics which could change under the influence of natural selection, but most importantly were their spiritual characters, which permitted race valuation. The Aryan race (especially its Nordic or German strand) reached the highest point in his scale. Jews were not considered to be very different from the German Aryans, but this did not prevent him from later collaborating willingly with the Nazis. He considered himself a student of Ploetz. Weismann's influence directed him to adopt non-Lamarckian practices for population improvement. He was conservative politically but also a strong supporter of meritocracy. Ruedin and other members of the Munich branch sympathized with Lenz's views. In general, the Munich branch succeeded in dominating that of Berlin, which was more progressive politically.

Scientific journals and the creation of institutions (research centers, university positions) gave increased status to eugenics. Lenz became professor of race hygiene in Munich in 1923, and Ruedin director of a research institute for psychiatry in Munich in 1931. The Rockefeller Foundation financed the creation of the Kaiser Wilhelm Institute for Anthropology, Human Heredity and Eugenics in Berlin in 1927 headed by Eugen Fischer (1874–1967). Fischer, a docent in anatomy at Freiburg, investigated in 1908 the bastards produced between Boer men and Hottentot women in South Africa and their progeny, and in 1913 published a book about them. He stated that such mixing of blood produced inferior individuals. Later he was considered as the expert on the products of racial mixing. The economic crisis and depression of 1929 put an end to the welfare state by which the Weimar Republic had tried to counter Bolshevism. As a means of alleviation of state charges, the implementation of eugenic measures was considered seriously: the creation of work colonies for the unfit was discussed again and a law on sterilization was drafted. Although there was general support for this law, the 1930 Papal encyclical prevented its enactment. It was implemented in the Nazi period (1933–45), when extreme racist practices were combined with eugenic ones. Despite the fact that racism was not always connected to eugenics, it is clear that compulsory sterilization and euthanasia were measures first proposed by eugenicists, and these constituted the inspiration for mass killings. Actually Nazis, as they characteristically put it, considered these measures to be 'applied biology.' Eugenisists who did not support racist practices (such as Ostermann and Muckermann) were removed from their posts. The Society for Race Hygiene came under state control, Ruedin was appointed as its head and given the authority to have officials and heads of the branches. The membership of the Society was restricted to German Aryans, and excluded Jews. Eugen Fischer was appointed head of the Berlin Institute of Anthropology and in 1933 was named rector of the University of Berlin. Another important figure was Otmar von Verschuer (1896–1969), a human geneticist, who fist occupied the position of head of the human genetics division under Fischer (Lenz occupying that of eugenics). Subsequently, he held a professorship at the University of Frankfurt and finally, in 1942, succeeded Fisher on the latter's retirement. Lenz, who previously regarded Jews as being nearly equal to Germans, kept silent under the Nazis and changed the text of his treatise in a subsequent edition to accord with the partly line. Along with many others, he collaborated willingly with the regime. About one-third of the physicians became members of the Nazi party, and several took

an active part in sterilization and 'euthanasia' projects. Although few scientists were involved directly in such projects (mainly, SS and party members were employed), all seemed to be aware of what was going on. In July 1933 a law permitting mandatory sterilization for the feebleminded was enacted, as well as for those suffering from mental illnesses and other 'hereditary' afflictions. A committee formed by Lenz, Ploetz and Ruedin was the highest expert authority and made the final decision on every case. From 1934 to 1939 (when apparently this practice was discontinued) between 200,000 and 400,000 individuals, probably 350,000, were sterilized without their consent, or even their knowledge. At least 367 women and 70 men died from postsurgical complications after such operations. In 1937, 385 children from German mothers and coloured fathers from among French colonial troops who occupied the Rhineland, were sterilized. In 1938 euthanasia by starvation was practiced in the mental hospital in Helborn. Euthanasia was applied in order not to waste money and valuable work on those designated as 'useless eaters.' About 100,000 individuals were eliminated in this way.

In 1934, several university professors and lawyers helped to draft the Nuremberg laws for the protection of the racial purity of the German race. Marriages and extramarital affairs of Germans with Jews were prohibited. Fischer, at the end of a meeting, thanked Hitler for giving geneticists the opportunity, through these laws, to make the results of their research useful to the general public! A law for euthanasia was drafted by Lenz, but apparently 'euthanasia' was practiced 'secretely' in labor camps for Jews, gypsies, some Slavs, and, of course, homosexuals, and socialists. The Holocaust was planned carefully by Hitler and Himmler. Some physicians experimented on those destined to die, or collected their organs and dispatched them to the Institute in Berlin, among them the well-known Dr Josef Mengele (1911–84), an assistant of von Verschuer at that time. It is estimated that at least 70,000 mental patients, 5–6 million Jews (mostly German, eastern and Balkan Jews) and an unknown number of gypsies, Slavs, and others were gassed. Thus it is not surprising that in Germany, after the war, population genetics was a taboo subject, and until the 1980s only one chair of evolution and population genetics existed (at the University of Tübingen).

4. *Switzerland and the Scandinavian Countries*

Along with some of the states of the USA, Switzerland was of the first countries to pass a law for compulsory sterilization, in 1927. Even as late as the early 1950s, recidivist rapists were given a choice, either to be castrated and go free, or suffer life imprisonment. In the 1930s, influenced by the USA, and especially by Germany, Scandinavian countries passed sterilization laws. Denmark permitted voluntary sterilization in 1929 and extended it to coerce the feebleminded in 1934, Norway legalized sterilization in 1934, Sweden the same year, Finland in 1935, and Iceland in 1938. A Baltic state, Estonia, passed such a law in 1936. These laws should be seen as part of welfare state policies rather than connected with the racial cleansing of Nazi Germany. However, it seems that such laws continued to be implemented well after the end of World War II, and the sterilization was performed without permission or even the knowledge of those being subjected to it. Mental defectives and asocial subjects were treated in the same way. In Denmark between 1929 and 1945 some 3,800 persons were sterilized.

5. *Russia*

Eugenics was introduced from England in 1875. Geneticists and biologists played the major roles, joined by physicians. Two young biologists, Nikolai Konstantinovich Kol'tsov (1892–1940) and Yuri Alexandrovich Filipshenko (1882–1930) were instrumental. After the Revolution and the civil war (1917–22) several opportunities were open to young scientists. Kol'tzov was named professor at the two universities in Moscow, while Filipshenko, a year after his thesis (1917), became professor at Petrograd University and the following year director of the Laboratory of Genetics and Experimental Biology.

Kol'tsov had created a privately funded Institute of Experimental biology in 1916. In a rhetorical declaration he stated that socialism was bound to earthly life and eugenics were an attempt to achieve perfect order in relations between people, thus (following Galton) he claimed it was similar to a religion, giving meaning to life. Using eugenics, he secured from N. Semashko, his friend and Commissar of Public Health, funds for his Institute. Semashko and others adapted German 'social hygiene' to a Soviet-style enterprise. In 1920 Kol'tsov founded the Russian Eugenics Society, which he controlled, since most of its members were his students and collaborators. He was therefore elected as president and served at the society's bureau with two other members, the psychiatrist T. I. Yudin and the anthropologist V. V. Bunak. Bunak, his student the geneticist A. S. Serebrovsky (1892–1948) and M. V. Volotskoi (1893–1944), were all Kol'tsov's protégés. In 1922 the Society started *The Russian Eugenics Journal*. Finally, at the Brussels' meeting in 1922, the Society became a member of the International Commission of Eugenics. Several eugenic societies were created in other cities in the Soviet Union and became affiliated with that in Moscow; one was created by Filipshenko in Petrograd, others in Odessa, Saratov, and Sverdlovsk. Filipshenko approached the Academy of Sciences and established a Bureau of Eugenics under its auspices, to which he appointed his students Diakonov, Lusis, Lepin,

Zuitin, Dob(r)zhansky, Kerkis, and Medvedev. The Bureau issued a *Bulletin* (1921–25); it later became the Bureau of Genetics and Eugenics (1925–27).

Since both Filipshenko and Kol'tsov were adepts in Mendelian and later in Morgan's genetics, eugenics in Russia had close links to classical genetics. In the mid 1920s Pavlov, a well known Lamarckian, was convinced by Kol'tsov and his wife that his data could be better interpreted by Mendelian genetics. This was a major victory for genetics in Russia. The research agenda of eugenics was to establish genealogies of talented individuals; the investigation of the heredity of criminality, alcoholism, schizophrenia, and manic depression, as well as of several bodily abnormalities and malfunctions (goiter, hernia). It also included demographic and statistical inquiries in general population studies.

In an attempt to provide a local origin to eugenics, Russians attributed it to the neglected writer, Florinsky. Subsequently they tried to render it 'socialist' or 'Bolshevik' to differentiate it from 'bourgeois' or 'capitalist' eugenics. With this attempt, Russian geneticists tried to protect their field from persecution and to secure state funds. To label eugenics as a scientific religion or an international movement did not help much; they should have presented material results in public health, or expectations of beneficial results. Young scientists devoted to Marxism and having an influence on the party helped in this respect. One of them, M. V. Volotskoi, became secretary of the Moscow society, another was A. S. Serebrovsky. In 1923 Volotskoi proposed a negative eugenic measure: sterilization of defective individuals. Vasectomy was used widely in the USA, as it was easily performed and was not a castration. In women, salpingectomy was a more complicated operation. This proposal met with heavy criticism, formulated by Filipshenko, and was finally rejected on moral grounds and because during those years the USSR witnessed an important population implosion due to the wars and famine. Deaths exceeded births. In one of the eugenic societies a proposal was approved also to eliminate abortions.

Supporters of Lamarckism were numerous in Russia, especially among those following the party line. Inheritance of acquired characters was thought to fit with Marxism and permit the improvement of society through concrete social actions. The Viennese biologist and strong supporter of this kind of inheritance, Paul Kammerer, was considered a hero, his books were translated and circulated widely, and the Communist Academy invited him to head a laboratory. His suicide was presented as part of a capitalist plot. Volotskoi originally supported Lamarckism. A physician, Soloman G. Levit (1884–1938) joined Volotskoi in this. Filipshenko, in the same year (1925), by himself and together with H. J. Muller criticized Lamarckism and demonstrated that it was scientifically unacceptable. He pointed out (as did Haldane in England) that if Lamarckism was right, proletarians were at a great disadvantage, bearing the unfavourable influences of generations of slavery. Serebrovsky tried to reconcile the negative attitude towards eugenics with this confrontation. He argued that, while genetics was a science, eugenics was a social enterprise influenced by the ideology of the class of its supporters; therefore each class created its own eugenics, but, whatever the case might be, bourgeois or socialist eugenics, it should be based on the objective science of genetics. Thus Serebrovsky distinguished '*antropogenetika*,' human genetics (*anthropos* means '*man*' in Greek), a scientific field, from eugenics, which was not a science. In 1927 the American geneticist, H. J. Muller, student of Morgan and a communist, demonstrated that irradiation by X-rays was mutagenic. This implied that genetic traits could be changed by an environmental agent and be manipulated for eugenic purposes (at that time it was not known that most of the mutations are harmful). This helped the conversion of Levit to Mendelism under the influence of his fellow Marxist, Serebrovsky. Serebrovsky made a new proposal in 1928, to provide a Bolshevik form of eugenics, the use of genetically recommended sperm in artificial inseminations to produce a great number of children. He remarked that 'one talented and viable producer could have up to 1,000 children' and 'that various women would be proud of the successes and achievements in this … production of new human beings.' Serebrovsky had worked in a zootechnical station on a project for amelioration of poultry and was aware of the technique, devised by Ilia I. Ivanov (1870–1932), for the artificial insemination of sheep, cows, and eventually of poultry. Serebrovsky's *antropotechnika* was obviously inspired from *zootechnika*. Vacher de Lapouge in France had proposed a similar measure in 1896, and later Muller, inspired by Serebrovsky, also proposed the eugenic use of artificial insemination. It was positive eugenics: the dissemination of valuable inherited traits. The proposal met with satirical comments.

The beginning of the Stalinist era (1929–32) with its great purges and the Bolshevization of the Academy of Sciences marked the end of the eugenics movement. In Leningrad, Filipshenko withdrew his interest in the movement from the mid-1920s and in 1930 he was relieved of his university duties and his department was disbanded. He died suddenly in the same year, from meningitis. In 1930, the Russian Eugenics Society was disbanded and its journal suspended. The Eugenics Section of the Kol'tsov Institute was abolished. Filipshenko's successor, N. I. Vavilov, the well-known plant geneticist, and a student of Bateson, regrouped the geneticists that were left, together with plant specialists. Technocrats were suspected of having bourgeois influences; a new generation of Marxists, especially those having a proletarian background, replaced them. Kol'tsov, in anticipation of difficulties to come, made some changes: He turned to medical genetics and included Marxists on his staff. His

Institute had two important groups, one headed by Serebrovsky, another by S. S. Chetverikov. Serebrovsky left the Institute in 1927, and Chetverikov was arrested and sent to exile in 1929. In 1930, Kol'tsov was removed from his teaching duties; and his department was disbanded and replaced by a Genetics Department headed by Serebrovsky. In 1933, Kol'tsov re-established his genetics division by bringing in a new geneticist of proletarian origin, N. Dubinin. In 1928, Kol'tsov created a Society for the Study of Racial Pathology and the Geographical Distribution of Diseases and became its president. From the beginning of the 1930s, eugenics was no longer considered a legitimate field, therefore former eugenic researchers moved to medical research.

Levit became director of the Biomedical Institute in 1930. He was working at this Institute from 1928 and had established an Office of Human Heredity and Constitution. In 1930–1, on Rockefeller Fellowships, Levit and Agol worked in Muller's laboratory in Texas. On his return, Levit continued to work at the Institute, and in 1934, at a conference alongside Muller, Kol'tsov and others, an expansion of medical genetics, was called for. In 1935, the Institute was renamed the Maxim Gorky Research Institute of Medical Genetics, and A. R. Luria headed its psychology division. Muller, a convinced communist, visited Russia for the first time in 1922, and returned in 1932. During his first stay he advanced genetic studies, bringing with him *Drosophila* cultures, and visiting Kol'tsov Institute and Serebrovsky's poultry station. Muller was interested in eugenics from 1910. In 1932 he accepted an offer from Vavilov to head a laboratory in Leningrad. In 1935 he published his book *Out of the Night: A Biologist's View of the Future*, where he expounded his eugenics views and the artificial fertilization programme. It seems that he arrived in the USSR in order to implement his eugenics programme. In 1936, he sent Stalin a copy of his book and a letter supporting his plan, arguing that eugenics would ensure the triumph of socialism. This letter had extremely negative consequences. A student of Muller's was arrested. The speakers at a session of the Lenin All-Union Academy of Agricultural Sciences, Vavilov, Serebrovsky and Muller defended genetics. Muller did not follow the advice he was given, to avoid mentioning human genetics. At this meeting Serebrovsky was heavily attacked by Lysenkoists. It seems that Stalin had read the transcription of Muller's talk and disliked it. Muller had to leave Russia for fear of being arrested too. After that, a Kol'tsov student, N. K. Beliaev, was arrested and shot. The translator of Muller's book into Russian was also shot, while Levit lost his position as director of the Gorky Institute, was arrested in 1938 and executed. In 1938, Kol'tsov was removed from his position and died in 1940 from a heart attack. The next day, his wife committed suicide. In 1938 Lysenko was elected as a member of the Academy of Sciences (the charges against Kol'tsov were formulated to secure the election of Lysenko, Kol'tsov also being a candidate for the same position) and the Soviet government and party gave him full support from 1940 to 1965. In 1940, Vavilov was arrested, along with his close friends and colleagues, G. D. Karpeshenko and G. A. Levitsky. All died in prison during the war (1942–3). During this quarter century, genetics became an illegal topic, but it did survive in a few centers which Lysenko could not control. It reappeared after Lysenkoism was repudiated in 1965; but not the eugenics movement, which the Nazis atrocities had thoroughly disqualified.

See also: Eugenics as an International Movement; Eugenics, History of; Eugenics in Asia; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Testing and Counseling: Educational and Professional Aspects; 'New' Eugenics, The: Role in Society; Primate Families, Structure of

Bibliography

Adams M B 1990 *The Wellborn Science: Eugenics in Germany, France, Brazil, and Russia*. Oxford University Press, New York

Broberg G, Roll-Hansen N (eds.) 1996 *Eugenics and the Welfare State: Sterilization Policy in Denmark, Sweden, Norway, and Finland*. Michigan State University, East Lansing, MI

Graham L R 1993 *Science in Russia and the Soviet Union; A Short History*. Cambridge University Press, Cambridge, UK

Kevles D J 1985 *In the Name of Eugenics: Genetics and the Uses of Human Heredity*. Harvard University Press, Cambridge, MA

Macrakis K 1993 *Surviving the Swastika: Scientific Research in Nazi Germany*. Oxford University Press, Oxford, UK

Muller-Hill B 1988 *Murderous Science: Elimination by Scientific Selection of Jews, Gypsies, and others in Germany, 1933–1945* [Trans. G. R. Fraser]. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Olff-Nathan J (ed.) 1993 *La Science sous le Troisième Reich*. Seuil, Paris

Paul D B 1995 *Controlling Human Heredity, 1865 to the Present*. Humanities Press, New Jersey

Proctor R N 1988 *Racial Hygiene: Medicine under the Nazis*. Harvard University Press, Cambridge, MA

Schneider W H 1990 *Quality and Quantity: The Quest for Biological Regeneration in Twentieth-century, France*. Cambridge University Press, Cambridge, UK

Todorov T 1993 *On Human Diversity, Nationalism, Racism, and Exotism in French Thought* [trans. Porter C]. Harvard University Press, Cambridge, MA

Tort P (ed.) 1996 *Dictionnaire du Darwinisme et de l'Evolution*, 3 Vols. Presses Universitaires de France, Paris

Weindling P 1989 *Health, Race and German Politics between National Unification and Nazism, 1870–1945*. Cambridge University Press, Cambridge, UK

C. Krimbas

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Europe, Archaeology of

European archaeology is the study of the material culture left by inhabitants of the European continent anywhere from three million years ago until the beginning of the Middle Ages. These people were as mixed ethnically as they were culturally and made their homes in Europe's diverse microregions and microclimatic zones. In their wake, they left a large legacy of material traces and it is from these traces that archaeologists can reconstruct the complex patterns of their lives. As a science, European archaeology got its start during the Renaissance, when the classical world was rediscovered. But it was not until the end of the nineteenth century, when typology, classification, systematic excavation, and relative chronology emerged, that archaeology became an academic discipline. By the twentieth century, European archaeology's major tenets had been established as had conflicting schools of archaeological thought (see archaeological theory). When the Iron Curtain fell in 1989, European archaeology experienced a new rupture. Terrain that was once impenetrable and colleagues and research that for decades had been unreachable suddenly became available. This exchange has both incited rich dialogue and stirred stimulating controversy. Spectacular recent finds such as the Iceman in the Austrian Alps and technological advances such as the ability to analyze prehistoric DNA are revolutionizing archaeology and promise to bring exciting new breakthroughs in the twenty-first century.

This article is divided into five parts, each of which outlines the major chronological and cultural shifts in European pre- and protohistory. Also included are major finds that significantly altered or enriched knowledge of the past.

1. *Europe's Origins—the Paleolithic and the Mesolithic*

The first European hominids belonged to the group *Homo erectus* and migrated from Africa about 1.8 million years ago. Cyclical changes in ice sheets caused them to move frequently. Archaeologically, this period, the oldest in hominid history, is known as the Lower Paleolithic. The second period, the Middle Paleolithic, occurred when *Homo erectus* was replaced by *Homo sapiens neanderthalensis* about 130,000 years ago. Unlike their ancestors, these new hominids buried their dead. Several Neanderthal graves have been found with grave goods, indicating that these hominids had some sort of belief system with notions of an after-life. The Neanderthals hunted, gathered, and scavenged.

The Middle Paleolithic ended when *Homo sapiens neanderthalensis* were rapidly replaced by *Homo sapiens sapiens*, the first modern humans. *Homo sapiens sapiens*, whose anatomic structure is the same as ours, emerged in Europe around 40,000 years ago, and soon spread throughout the continent. Their movement was probably facilitated by the steady disappearance of the ice sheet and marks the beginning of the Upper Paleolithic. During this time, stone technology advanced and archaeologists have identified four varieties of flint tools which they believe represent distinct societies of hunter–gatherers. They are: Aurignacian, Gravettian, Solutrean, and Magdalenian. Cave paintings found at Upper Paleolithic sites offer proof of artistic activity and the ability of *Homo sapiens sapiens* to think symbolically. The paintings mainly depict animals, however abstract patterns of lines and squares are also present. Two sites—Altamira in northern Spain and Lascaux and Cosquer in southern France dating from c. 30,000 BC—offer particularly vivid examples. Equally elaborate are the carved female figurines depicted with exaggerated breasts, stomachs, buttocks, and thighs. They have been found all over Europe, the best example being the Venus of Willendorf.

The Mesolithic is the first period of the Holocene and comes at the end of the Ice Age. Its chronology varies according to region, but it dates from about 10,000–5,000 BC. As the ice sheet disappeared, hunter–gatherers began to inhabit forest land and open tundra that had earlier been unlivable. Warmer temperatures also stimulated great changes in fauna. Mammoths became extinct, reindeer moved north and were replaced by forest-dwelling creatures like red deer, roe deer, and wild boar. In a few regions, a complex settlement and economic pattern emerged which involved a base camp and a number of small, specialized activity camps. In some cases, like in Ertebölle, coastal resources were exploited, while in Star Carr, hunting red deer, roe deer, and aurochs became a specialized task. At Lepenski Vir, located at the Iron Gate of the Danube Gorge, finely worked flint objects, rounded stone sculptures, and trapezoid houses demonstrate the aesthetics of these hunter–gatherers. Overall, the Mesolithic is characterized by an ability to adapt to and exploit diverse environmental climates for personal gain. It is marked by a series of technological innovations, most notably the microlithization of flint tools. These tools were used as projectile points and as components in other tools. In some regions, such as southern Scandinavia and the eastern part of the Baltic Sea, a hunting and gathering economy continued long after farming was introduced in Europe.

2. *The Emergence of Farming and Social Complexity: The Neolithic and the Copper Age*

The Neolithic is often identified by the appearance of domestic animals and crops. This huge change revolu-

tionized European history. It is accompanied by an increase in the size and complexity of settlements as well as the use of pottery and polished stone tools, and in some regions, the erection of elaborate monuments. During the Neolithic, people began to view the world in a whole new way. Farming revolutionized ways of life and subsistence. The human diet shifted from an almost total dependence on wild animals to a diet based primarily on cereals and domesticated animals. Sedentary life had demographic consequences and as populations increased, society became more complex and stratified.

Early Neolithic communities came to Europe from the Near East and Anatolia around 7,000 BC, and first settled in the southeastern part of the continent in Greece and the Balkans. As in Anatolia, subsistence was based on emmer, wheat, sheep, goats, pigs and cattle. Farmers inhabited open settlements and lived in free-standing, rectangular houses. Some settlements were inhabited for long periods and the constant building and rebuilding led to the emergence of large settlement mounds (tells). From these southeastern settlements, Neolithic communities dispersed across Europe. The movement went in two general directions: To the north and north-west and to the west along the Mediterranean coast. The process of colonization varied tremendously, as did the adoption of domesticates. In southern Scandinavia and the British Isles, for example, it occurred around 4,000 BC, while in central Germany, it occurred about 1,300 years earlier. The western route of Neolithic colonization led along the Mediterranean coast to southern Iberia and southern France, both of which were colonized around 5,500 BC. These routes have been identified closely with the spread of Impressed Ware Pottery. From the Mediterranean, domestication and farming moved northward along the Atlantic coast into the Parisian Basin. Here, the southern route met the northern route. This route, which followed the Danube through Central Europe, is associated with Linear Pottery and showed advances in the use of long-houses, incised decorated pottery and the shoe-last adze for clearing forests. By 5,300 BC, the Neolithization of continental Europe was complete. The most significant pottery emerged from the Chasseen culture in France, the Michelsberg culture between the Rhine delta and the Alps and the Trichterbecher culture in northern and eastern Europe. Along the Atlantic coast, from Portugal in the south, the Orkney Isles and Scandinavia in the north, and Poland in the east, large megaliths in the form of standing stones, stone circles, henges, and cursus monuments have been found. During the fifth millennium BC, copper metallurgy was introduced into southeastern Europe, specifically in the Balkans and the Carpathian Basin. This period, which lasted from 4,500 to 2,500 BC, is hence known as the Copper Age.

Economic changes such as the introduction of the wheel and plow, wool production, and the increased use of animals for specialized tasks occurred. Together, this is known as the Secondary Product Revolution. At this time, more emphasis was placed on creating cemeteries than on erecting domestic structures. Cemeteries were a means of establishing group identity and instilling a sense of belonging. A spectacular example of this is in Varna on the Black Sea coast of Bulgaria. Here, individual graves featured sophisticated gold and copper work. By 3,000 BC, the explosion of new innovations ruptured traditional lifestyles across the continent and change occurred rapidly. As wealth became portable, emphasis shifted from places to people and from collective to personal possessions. A symbol of this change is Corded Ware pottery, which is decorated with impressed cords. Such pottery has been found in male graves located under circular mounds across Northern and Central Europe.

In addition to the pottery, men were found in single graves accompanied by stone battleaxes. This has both been interpreted as a sign of the importance of fighting and drinking and of individualization within a stratified society. A second type of pottery, handless drinking-cups know as Bell-Beakers, appeared concurrently and slightly later in the expansive area from Scotland to Sicily and Spain to Central Europe. Graves were flat graves and in some regions also individual round mounds and, along with Bell-Beakers, contained daggers and archery equipment such as triangular barbed-flint arrowheads and wrist guards of fine stone. The vast distribution of the Bell-Beakers across temperate Europe and the Mediterranean indicates an increasingly mobile lifestyle that arrived just as traditional social structures disintegrated.

3. *The Emergence of Elites: The Bronze Age*

During the second millennium BC, bronze became the universal medium of ordinary objects and also of prestige. Horses and wheeled vehicles, textiles, amber necklaces, and metal dress-pins for clipping wool clothing emerged as visible signs of wealth. Trade in bronze expanded and bronze became popular for manufacturing weapons, ornaments and tools. Most importantly, metallurgy shifted as copper became regularly alloyed and tin began to replace arsenic. Although copper is relatively rare as a naturally occurring material, tin is even rarer and the creation and circulation of bronze objects depended on regular supplies of both. This led to the rise of leading centers of metalwork, each of which made technical advances in bronze casting that were widely diffused and adopted. As techniques of war changed, so did the types of weapons used. New bronze weapons and designs spread across the continent.

In Central Europe, in a region known as the Carpathian Basin, where the Únětice culture flourished, large hoards of bronze objects and magnificent burial mounds filled with gold and bronze objects have been uncovered. The influence of these communities stretched from Denmark to Italy to Greece, and to some degree, the British Isles. In the western Mediterranean, similar developments around 2,200 BC spawned the establishment of new centers along the coast of the Iberian Peninsular. Here, hilltop settlements containing rectangular houses and streets were surrounded by stone walls. These settlements lasted throughout the second millennium and are called El Agrar. The proximity of urban cultures in the Eastern Mediterranean also provided alternative centers of trade and exposure to novel eating and drinking habits, clothing, furniture, means of transport, prestige ornaments, and weaponry.

Demand for bronze by the populations of the northpontic steppes spurred trade and soon communication and exchange networks were established. North of the Caspian Sea, new grave architecture that included timber-built underground chambers known as Timber Grave culture was found. Eastward, the closely-related Andronovo culture perfected the use of horses and created a horse-driven chariot complete with a bridle-bit with spiked cheek pieces made of antler. At the timber-grave cemetery of Sintashta in the southern Urals, shaft graves similar to those of Mycenae were found containing wheeled chariots and horses.

When the palatial societies of the eastern Mediterranean collapsed after 1,300 BC, dramatic changes swept across Europe. This period is known as the late Bronze Age and is characterized by hill-forts, flat cemeteries, scrap hoards, and votive finds from rivers. The most striking sign of the changes is the almost-universal shift from burying the dead in graves to cremating them and putting their ashes in pits or funerary urns. The period lasts until c. 800 BC and has been named the 'Urnfield Period.' It is a transitory time, a phase in between the Bronze and the Iron Age, in which economic, social, military, and religious changes that are of pivotal importance in the formation of historic Europe occurred. For the first time, swords, spearheads, armor, safety pins, and certain ornaments from central and northern Europe appeared in the Mediterranean. Hilltop forts and stockades on lower ground became popular around 1,100 BC and lasted until c. 550 BC. At Biskupin, near Bydogszcz in Poland, a fortified settlement island with thirteen parallel rows of houses divided by roads and encircled by a rampart were excavated. The layout resonates communal order and indicates the presence of a coercive power. Throughout the period, exchange networks allowed goods to travel over vast distances and for the first time salt from the Austrian Alps and other places emerged as a valuable trade commodity. Hallstatt, the richest cemetery of the period, sits on top of a salt mine. Spiritual life also took on great importance and the bird became a beloved symbol. Figurines in the shape of birds have been found, as have numerous pots, cauldrons, and cult wagons with bird applications. Numerous rock carvings with engravings of humans, animals, boats and a variety of symbols were found in Sweden (Tanum), the Alps at Val Camonica north of Brescia, and on Monte Bego near Bordighera. The abundance of swords, daggers, spearheads, arrowheads, and armor found indicate the importance of war and the extensive choices available to the warrior.

4. From Chieftain to 'State' Organization in Celtic Europe: The Iron Age

By 600 BC, iron became the standard material for tools and weapons, while bronze was used for high quality art products. The explosion of trading ports around the Mediterranean provided a strong incentive for European development. Mediterranean goods were sought avidly and the most valuable ones were acquired by the wealthy as proof of their exalted status. The valleys of the Rhone and Saône provided easy access to the region from Burgundy to southern Germany. This west Hallstatt zone was pivotal as it spanned the upper courses of the Seine, Saône, Rhine, and Danube rivers and directly linked the continent north of the Alps. Greek-led trade intensified through the colony of Massalia in the sixth century and the elite in this northern zone controlled trade of exotic Mediterranean products, including prestige equipment for wine-drinking rituals of the Greek symposium. Large valuables, such as the human-sized bronze crater from Vix, were imported and were exchanged for local products and slaves. The native elite maintained full control of these exchanges and likely held lavish feasts and gift exchanges among important chieftains, including the famous chief from Hochdorf. At graves, social prestige was demonstrated by the inclusion of costly sets of wine-drinking equipment and other exotic gear with the dead. Funerary rituals took place in large-scale religious structures surrounding burial mounds. At Glauberg in Hessen, a kilometer-long ceremonial path leading to a grave from the early La Tène period was recently discovered. Hilltops also emerged as prime locations for defended residences for the nobility. At Mount Lassois in Burgundy and at the Heuneburg in Germany, huge quantities of Mediterranean luxury goods were found. The Heuneburg was surrounded by a mud-brick wall built on dry stone footings with forward-projecting rectangular bastions. This technique, common enough in the Greek world, was totally alien to temperate Europe, and is a striking example of how avidly the European elite adopted Mediterranean achievements, even when they were impractical for the

climate. In the east, in Slovenia, a second group of less wealthy communities known as the east Hallstatt zone arose from the eighth to the fifth centuries. They were characterized by fortified settlements such as Magdalenska Gora, Vače, and Stična and large cemeteries with aristocratic graves furnished with sheet bronze body armor and decorated situlae.

The rise of a new warrior elite and the drying up of the market for luxury goods in the early fifth century destabilized the east and west Hallstatt systems and by the middle of the fifth century caused their demise. A new, northern elite arose and relied on native craftsmen to produce luxury goods. Numerous late fifth century elite tombs have been found containing locally-made masterpieces of Celtic-style art. Schools of creative metalworkers emerged during this period and their work led to the creation of the first truly Pan-European art style, known as Celtic or La Tène art. A rapid population increase, evident in the growth in the number of cemeteries, ensued and shook the social equilibrium. Over time, the system collapsed and resulted in a widespread migration. Late fifth- and early fourth-century cemeteries which extend south into the Po Valley and east along the Danube into Hungary trace the path of this migration. Some early groups also crossed the great Hungarian plains and settled in Transylvania. In the Marne–Moselle region, elite burials continued at Somme-Bionne, Reinheim, and Waldalgesheim. But overall, the number of cemeteries decreased, indicating migration. In the tombs, young men were usually found accompanied by spears, swords and shields. Wealthier men had bronze helmets and some possessed two-wheeled chariots. It is these warriors, called 'Gauls' or 'Celts,' who, in tens of thousands, made their way across Europe seeking new lands to conquer and settlements to plunder. The Celtic migrations between 450 to 200 BC can be traced archaeologically across Europe through the La Tène culture, which is characterized by art depicting mobility, upheaval, and turbulence.

The northern limit of La Tène culture lay at the mouth of the Rhine and across southern Germany and northern Italy to the Carpathians. Further east, the Celts were well-established in the lower Danube region. In the early fourth century, they began moving eastward into Transylvania and in the third century they pushed into southern Thrace and Greece. In the north-east, Celtic groups crossed the Prut and Dnestr, exploiting the power vacuum left by the decline of the Scythians. One of the few surviving pieces of silver made during the disrupted Celtic period is the Gundestrup cauldron found in Jutland. The puzzle of this elaborate cauldron encapsulates the complexity of these migrations. The techniques used to make it are typically Augustian Thracian. Some of the figures on it, such as a warrior wearing a bird-crested helmet look Celtic, while other motifs are reminiscent of steppe shamanism. The cauldron was probably carried north as booty by raiding Germans.

5. *Center and Periphery: Barbarian Europe and the Roman Empire*

For Rome, the Rhine River marked an ethnic divide between the Celts to the south and the Germans to the north. In reality, tribes on both sides of the river had integrated to such a degree that few ethnic differences existed. Wealth derived from constant access to particularly fertile land and did not build up in the hands of any single group. Status could, however, be acquired by acts of bravery and leadership in raids. Power at this time lay with the elite and was measured by entourage and the ability to bestow patronage. The elites formed the tribal council. Caesar overran Gaul fairly easily because the Gauls were a settled people with a well-developed agricultural base and tribes centrally located on permanent defended settlements known as oppida. Oppida were the first urban centers in temperate Europe and have been found from France to Serbia. At Manching in southern Germany, the settlement was densely built up with regularly laid out timber buildings aligned along straight, ordered streets. A wide range of craft skills were practiced and coins were minted. Commodities such as wheel-turned pottery, glass beads, bracelets, and ironwork of every kind was done on an industrial scale. In oppida like Bibracte, cargoes of wine amphorae were exchanged with the local aristocracy for slaves and other commodities. This stability made Gaulish society soft and vulnerable. Germans were the opposite. Mobility and warfare were central to their lives and they had few southern luxuries. Since they had no oppida to protect, they could simply retreat and disappear. The Roman Empire desperately needed a continuous supply of raw materials and manpower. As its frontiers continued to expand, valuable metals and other commodities arrived. Rome's greatest dependence lay in its need for slaves to keep the Roman system working. The scale of the slave trade was enormous. Judging from the distribution maps of artifacts in Free Germania, Romans traded coins extensively, but also traded sets of metal vessels used for wine-drinking rituals, buckets, situlae, and jugs. In this zone, a number of rich burial sites known as Lübsow were found. The graves contained extraordinary luxury goods, indicating that the dead were members of the elite, who, by virtue of their rank, were able to acquire a set of the most costly Roman items. Such local elite served as middle men between the Roman world and the Baltic region. This system of socioeconomic zones continued until the middle of the second century AD, when population expansion, social instability, and folk movements in Central Europe built up and pressured tribes along the frontier. They then threatened the stability of the Empire. The territory occupied by the warrior burials coincides with the region from which the Burgundians and Vandals are thought to have emerged. The first recorded indications of the growing problem came in AD 162 when a frontier tribe, the

Chatti, attempted to migrate south into Roman territory. A few years later, the Longobardi and Marcomanni crossed the Danube into transdanubian Hungary. This was followed by even more massive movements by other tribes. These events, and the Marcomannic War fought by Rome between 166 and 180 AD, mark the beginning of a wave of migrations which within three centuries destroyed the Empire and laid the groundwork for the formation of medieval Europe.

Bibliography

Barker G (ed.) 1999 *Companion Encyclopedia of Archaeology*, 2 Vols. Routledge, London and New York

Clottes J 1996 *The Cave Beneath the Sea: Paleolithic Images at Cosquer*. Abrams, New York

Die Gletschermumie aus der Kupferzeit: Neue Forschungsergebnisse zum Mann aus dem Eis 1999. Schriften des Südtiroler Archäologiemuseums 1. Folio-Verlag, Bozen

Gamble C 1999 *The Palaeolithic Societies of Europe*. Cambridge University Press, Cambridge, UK

Hänsel B (ed.) 1998 *Man and Environment in Bronze Age. The Bronze Age: The First Golden Age of Europe*. Oetker-Voges Verlag, Kiel, Germany

Herrmann F-R, Frey O-H 1996 *Die Keltenfürsten vom Glauberg: Ein frühkeltischer Fürstengrabhügel am Hang des Glauberges bei Glauburg-Glauberg, Wetteraukreis*. Archäologische Gesellschaft in Hessen, Landesamt für Denkmalpflege, Wiesbaden, Germany

Hoops J 2001 *Reallexikon der Germanischen Altertumskunde*, 17 Vols. Walter De Gruyter, Berlin and New York

Kristiansen K 1998 *Europe Before History*. Cambridge University Press, Cambridge, UK

Lichardus J, Lichardus-Itten M 1985 *La protohistoire de l'Europe. Le Neolithique et le Chalcolithique entre la Mediterranee et la Mer Baltique*. Nouvelle Clio Ibis, Paris

Müller-Karpe H 1998 *Grundzüge früher Menschheitsgeschichte*, 5 Vols. Konrad Theiss Verlag, Stuttgart, Germany

P. F. Biehl

Europe: Sociocultural Aspects

1. The Term 'Europe'

In ancient Greece, the term Europe evoked legendary or divine characters from the mythological universe, rather than a geographic site. The Greek geographer Strabo (58 BC–AD 25), who lived under the Roman emperors Augustus and Tiberius, gave the first detailed description of the favorable climatic and natural conditions in which European populations lived. Strabo contended that these populations were intellectually superior to those of other continents. The Roman Empire however was never particularly Eurocentric. In those days, it was seen as a tricontinental entity, socially and culturally composite, whose hub was the Mediterranean, which the Latin knew as 'Mare Nostrum.'

The term Europe took on a more political and sociocultural connotation during the Carolingian period (seventh to eighth centuries). According to a medieval anonymous chronicle, the soldiers led by Charles Martel who fought victoriously against the Arabs at Poitiers (732) were 'European.' At this time, the notion of Europe was assimilated with the Frankish Kingdom, and the idea of Europe as a political-spiritual entity developed along with the expansion of Arab–Muslim settlements north of the Strait of Gibraltar.

One of the recurrent themes of various subsequent concepts of Europe as an 'imagined community' (Anderson 1983) can be identified *in statu nascendi* from the latter observation. They are characterized almost without exception by a scheme in which the notion of Europe appears together with another contrastive term. In most cases, it is a dichotomy or bipolarity, rather than a comparison between what Europe is reckoned to be and what is supposedly its opposite. This 'European self-awareness,' as de Rougemont (1990) calls it, becomes even more tangible at the time of geographical discoveries and the first colonial conquests.

From an anthropological point of view, Europe discovers itself by discovering and conquering the rest of the world. Between 'Europe' and 'Anti-Europe,' which from time to time could be Arab–Muslim culture or Asia, Africa, or the Orient, there is a reflective relationship in which the former identifies its basic traits through a presumed diametrical difference with the latter. Balthazar Gracian (1601–1658) coins the famous phrase 'Europe is the admirable face of the world.' Following the same train of thought, the *Grand Dictionnaire historique de Louis de Moreri* (published in 1674) illustrates the European traits that have allowed them to overpower the people of other continents. It is quite surprising to read statements in this publication that seem to foreshadow the stereotypes of German 'Völkerpsychologie,' as well as American studies on 'modal personality' and 'national character.' Similar arguments were expressed more briefly and clearly by Carl von Linné (1707–1778); the Swedish naturalist describes the 'Homo Europaeus' as 'subtle' and 'creative' while Asians are 'melancholic' and 'miserly' and Africans are 'sly' and 'lazy.'

During the *ère des philosophes*, almost all the eighteenth century's great thinkers—Montesquieu, Voltaire, Jean-Jacques Rousseau, Ferdinando Galiani, Melchior-Gaspar de Jovellanos, Antoine de Condorcet, Edward Gibbon, Edmund Burke—pondered over a 'European peculiarity.' Their dissertations share a common denominator: a scheme, usually bipolar, in which at least one other term of comparison appears besides Europe. Thus, the main question for most of the above-mentioned authors is: how is it that European nations are the only ones that have reached such levels of civilization, while the

'other' societies are still savage, primitive, or barbarian? Leaving aside the answers, almost all eighteenth century texts on Europe draw either implicit or explicit parallels and hierarchies between 'us' and 'them,' between 'Europe' and 'Anti-Europe.'

During the nineteenth century, the notion of Europe took on an increasingly hegemonic connotation. The century was characterized by a bold optimism that turned out to be an awful self-deception. In historical terms, European 'élites' surprisingly had underestimated the changes that had been going on beyond their continent. They had overlooked that there had been the American 'Great Transformation,' or the Japanese development model spurred by the emperor Mutsuhito (1852–1912), who had opened his country to foreign commerce. It is no surprise that the term Europe was being increasingly used as a synonym of 'civilization' and 'progress' (Schmitt 1974). From a European point of view, the Old Continent was the only one that could establish a 'civilizing movement' to 'break through the darkness that enfolds whole populations' as Leopold II (1835–1909), King of the Belgians, stated in a solemn speech in 1876. For this typical representative of the time, European colonial expansion in the heart of the 'Anti-Europe' at that moment, namely Africa, was 'a crusade worthy of this century of progress.'

Nineteenth-century certainties were curtailed radically by the World War I, which the German philosopher Karl Löwith (1897–1973) keenly called 'the first European civil war' (Löwith 1990); by the critical analysis of the nationalistic and totalitarian frenzy between the two World Wars; by World War II traumas; and finally by the melancholic awareness of crumbling colonial empires. Europe as a nation became less triumphalist and much more defensive. After 1945, Europe was perceived as suffering from a chronic economic, cultural, and above all political and military fragility. This intrinsic deficiency could be checked only by a strong synergy between all the continent's components sharing the same fate. In this respect, the French sociologist Edyan Morin (1987) has fittingly spoken about Europe as a '*Schicksalsgemeinschaft*,' a term first used by the Austro-Marxist theorist Otto Bauer (1881–1938). From a sociological point of view, the 'European fragility' concept implies a 'menace' coming from external hostile forces. Each community, especially if it is a '*Schicksalsgemeinschaft*,' is strengthened or crystallized if under pressure from an actual or presumed enemy. The part of Europe that after the World War II was not under Soviet control always reverted to the 'Anti-Europe' theme to emphasize who the 'menace' was. From 1945 until 1989, while Europe was establishing its first joint institutions, the leading 'Anti-Europe' 'menace' was the Soviet and communist sphere.

The fall of the Berlin Wall (1989) and the subsequent disintegration of the opposing blocs coincided with the birth of a new European institution, the European Union (EU). A 'fragile' and 'menaced' Europe is still a common perception on the Old Continent despite this new highly specialized and official legal apparatus. Behind this defensive attitude some authoritative, though overly severe, 'Euro-skeptic' critics detect the danger of the creation of a 'European fortress' that wishes only to defend an acquired economic standing threatened by supposed external 'menaces.' At the end of the twentieth century, the bipolarity between 'Europe' and 'Anti-Europe' still exists. Now 'Anti-Europe' is mainly the planet's boundless south: its migratory waves towards the Old Continent, its high demographic rates, and political-religious conflicts.

2. *What Europe is Not: Some Misleading and Dangerous Ideas*

During the twentieth century, notions of Europe with powerful ideological implications have been spread occasionally; yet, they can hardly endure historical, sociological, and anthropological analysis. The idea of 'European civilization' as a 'monolithic' entity was most popular and politically relevant. 'European civilization' was seen as the bulwark of Christianity. This concept of Europe was an ideological backbone of antidemocratic political movements. Although under many aspects admittedly in conflict with Christian doctrine and ecclesiastic institutions, German Nazism and Italian Fascism envisioned themselves as the upholders of 'European civilization' and strenuous paladins of 'Christianity' particularly against Soviet Bolshevism, which was stigmatized as 'Asian,' hence 'alien' to Europe. Other regimes with similar ideological backgrounds, such as authoritarian and dictatorial governments and movements in Central and East European countries as well as Francoism in Spain and Salazarism in Portugal, resorted to analogous ideas.

Europe as a 'united civilization' embodied in the 'Christian Occident' is an idea that can still be traced in present-day European extreme rightists, such as the French *Front National* and neo-Nazi and neo-Fascist fringes in Germany, Austria, and Italy. These movements resort to similar concepts to assert their hostility towards immigration, especially from Islamic and/or non-European countries. It is a dangerous idea, since it implies the pursuit of cultural, denominational, and maybe even ethnic homogeneity based upon the 'exclusion strategy' and the 'purity myth,' which caused so many of Europe's tragedies during the twentieth century. This vision of Europe is misleading since it is far from the continent's past as well as present social realities.

3. *Europe: A System of 'Historical Regions'*

The German philosopher Ernst Bloch (1885–1977) advises us not to turn a highly differentiated social

entity as Europe into a sum of 'cultural areas' because this would turn 'History…into a circus (with) three or more (separate) rings' (Bloch 1985). Europe must be considered a system of strictly (inter)dependent yet structurally diverse 'historical regions.' Hungarian historian Szücs (1983) was the first one to apply a strictly scientific method to the problem of Europe's internal frontiers, to pose the question of its historical divisions, and consequently of its sociostructural differences.

Szücs stresses that from the early Middle Ages, Europe began to differentiate socioeconomically giving rise to two different poles: '*Europa occidentalis*' and '*Europa orientalis*.' The border could be a line running from the Elbe River (socioeconomic boundary line between western and eastern Germany) to the Leitha River (for many centuries a previous frontier between Austria and Hungary). A second economic and sociostructural division, which includes all societies involved in the Great Schism (1054), came to pass between the eleventh and fourteenth century. Thus, there are three 'historical regions' with very diverse socioeconomic, political, and cultural characteristics. The first one includes the territory from the Atlantic to the above-mentioned Elbe–Leitha ideal line: more or less Carolingian Europe, today believed to be the quintessence of Western Europe. The central 'historical region' east of the Elbe–Leitha line but west of the eastern and southern borders of the original kingdoms of Hungary and Poland (Grand Duchy of Lithuania included) approximately overlaps with what is known today as mid-Eastern Europe, also called *Mitteleuropa*. Everything east of the Baltic–Black Sea line or south of the Carpathians is East Europe proper.

Szücs' arguments are relevant for the social sciences because they reveal how processes dating back to a remote past have created enduring economic, sociostructural, and cultural differences in Europe; however, his hypothesis of at least three 'historical regions' must be broadened and specified. It is useful to add Wallerstein's analysis of Europe incorporated in the 'world- system,' yet divided into 'core' and peripheries' (Wallerstein 1974). A new system to organize and manage the economy arose and spread between 1450 and 1640. Max Weber labeled this new system rational capitalism. According to Wallerstein (1974), as well as to the French historian Braudel (1979) amongst others, expansionist movements are inherent to this new model of economic and social behavior. The capitalist economy's worldwide expansion implies new relationships between economic partners, that is, societies, states, regions, towns, etc. These new relationships entail a new international division of labor that will permeate social structures and cultural models of all the societies involved.

The international division of labor within the 'world-system' is also based on a systematic separation between core and peripheries. Europe at first, and then the whole world, came to be divided into a number of regions. Some of these underwent a historical acceleration, quickly modernizing most of their socioeconomic structure, while others suffered centuries of socioeconomic stagnation if not growing impoverishment. The following classification highlights roles of Europe's 'historical regions' within the new international division of labor.

3.1 Northwestern Europe

The first core of the 'world-system' ensued from this region. Steady decline of the feudal system, early establishment of an extended net of self-governing towns (*freie Reichstädte*, *comuni*, etc.), striking capital accumulation followed by important technological changes culminating in the 'industrial revolution,' the rise of strong territorial states with more culturally homogeneous populations, colonial expansion outside Europe: these are only some of the most evident processes marking the socioeconomic and cultural structure of this 'historical region.' From the onset, it included southern England, France, Flanders, Holland, and the westernmost part of Germany. Northern Spain and Northern Italy, other areas of Germany and most of Switzerland, Austria, Bohemia, and Scandinavia joined the core in subsequent periods.

3.2 Mediterranean Europe

This 'historical region,' for centuries the continent's socioeconomic and cultural hub, was turned into a vast periphery by the new international division of labor. It provided the core with foodstuffs and raw materials for the growing textile industry. From a socioeconomic point of view, the area comprising Portugal, Spain, and Italy was ruled by a system of latifundistic wheat monoculture and/or stock-farming with an absentee owner aristocracy. The vast properties were administered by large tenants who sublet to smaller ones, who in turn would sublease to the smallest peasants. This long renting-chain led to a devastating fragmentation. Too many people without a productive activity lived exclusively off the land and the backs of those who actually tilled it.

Even though the banking system, rational capitalism's pillar, was developed in Renaissance Italy, and the two Iberian States were the first to rely on imperial colonies, Mediterranean Europe's structure was rooted in agrarian societies and economies from the sixteenth century up to most of the twentieth century. The rising bourgeoisie of this 'historical region' did not play a specific role, nor did it have its own specific economic behavior. Its ideal was to imitate the aristocracy. This matched the economic and social behavior of such categories of tenants as the *señoritos* in Andalusia, La Mancha, and Extremadura, or the *gabelloti* and *galantuomini* in Sicily and

southern Italy, who were inclined to a passive existence based on the principle of 'dignified idleness.'

3.3 Central-east Europe

This 'historical region,' with few exceptions, lies between the two ideal boundaries traced by Szücs. It is Europe's second vast periphery and has been one of the core's major suppliers of foodstuffs. The typical socioeconomic scenery of Europe's second periphery is in western and eastern Prussia, the Baltic countries, the Hungarian plain, and the Romanian Danube plain besides Poland's so-called eastern territories (most of which were lost after World War II). In these boundless areas were the landed estates of Central-east Europe's aristocracy: major Polish and Hungarian noble families, Baltic overlords of Germanic ancestry, and Romanian boyars. The institution of serfdom was the rule in these estates nearly up to the end of the nineteenth century. In fact, rather than being abolished, serfdom relationships were revived in what historians call the 'second serfdom.' Great landowners introduced new farming methods in order to increase crops, mainly cereals, bringing down harsher labor and life circumstances on peasants. Central-east European 'second serfdom' was also accompanied by the formation of new groups of great landowners with a strong 'corporate spirit' and staunch political and socio-economic ambitions such as the Prussian *Junker*, Polish *Szlachta*, and Hungarian 'gentry.'

3.4 Southeast Europe

This 'historical region' includes mainly the Balkans (except Romania). The Ottoman Empire which dominated it for almost 500 years became a part of the 'world-system' only during the eighteenth century, that is, during its manifest decadence. However, even under Ottoman domination the Balkans was a poor periphery, on the precarious border between the Ottoman and Austro–Hungarian empires. Therefore, the circumstances of this 'historical region' should not be compared to other more flourishing socioeconomic areas of the Ottoman Empire such as parts of Syria or Egypt. To make matters worse, in the Balkans there was a degeneration of the patrimonial system based on 'tax-farming.' The Balkans' destiny within the 'world-system' was to become a 'periphery of the periphery.'

3.5 East Europe

In Europe's third 'historical region,' according to Szücs' classification, there is mainly Russian territory west of the Urals, except the areas south of the Carpathians. During its first centuries this enormous country was an 'external area' as regards to the 'world-system,' with a distinctive socioeconomic structure. A communitarian land property regime and a peculiar patrimonialism by which an imperial 'official' 'arrogated' property administrated on behalf of the tsar, were the socioeconomic structure's cornerstone. Historians compared the situation engendered by such a structural peculiarity, which lasted up to the Russian Revolution (1917), to a 'bog of misery' (Riasanovsky 1984). However, East Europe's gradual integration in the 'world-system,' beginning in the eighteenth century, remained incomplete. Russia continued to be a European 'external area' and never became a real periphery or a core-state, as the other 'historical regions' already described.

3.6 '*Historical Regions*' and Present Socio-economic Gaps

The establishment of the 'world-system' and the growth of a new international division of labor, which have molded the European core and peripheries, are the primary components of an epochal historical change whose socioeconomic and cultural consequences are still visible to this day.

The gaps that appeared between Europe's 'historical regions' during the formation of the 'world-system' are still plain to see. The north–south gap mirrors past historical sociostructural and economic differences between the core and the Mediterranean periphery. To date, even the integration of these two 'historical regions' within the European Union has not eliminated the gap. The so-called 'enlargement towards the East' of the European Union is a problematic issue due specifically to east–west discrepancies (and not only to 50 years of socialism) that embody old boundaries between the European nucleus and the central–east periphery. The Balkan situation, which is characterized by feeble industrialization and incomplete modernization, is due to the fact that for centuries this area was a periphery of the periphery. Finally, present-day Russia with its distinct 'development model' torn between capitalistic borrowings and (pre)Socialist communitarist revivals, is adhering to its typical 'tradition' of lingering on the borders of the 'world-system.' The core–periphery gaps are detectable not only interregionally but also within states. In fact, states that straddle 'historical regions' have strongly 'dualistic' societies, economies, and cultures. For example, Italy is divided into north and south; Spain has discrepancies between Catalonia and Spanish Basque Country on one side, and the rest of the country on the other; in former Czechoslovakia there were differences between Bohemia and Slovakia; in former Yugoslavia, Slovenia, and Croatia were more connected to the core while Serbia, Bosnia, and Macedonia belonged to the southeast periphery.

3.7 Perceived Gaps and Peripheral Knowledge

Core–periphery gaps go beyond structural data. They are also 'lived,' therefore 'perceived' and 'constructed'

realities. In European peripheries, especially among the intellectual élites, there is an awareness of the 'marginality' if not the 'backwardness' of one's country or its 'historical region.' However, there is also a strong belief that one's 'nation' or 'people' has outstanding 'excellence' and 'virtues.' 'Peripheral élites' have been coping with the discrepancy between their nation's perceived socioeconomic 'inferiority' and its presumed spiritual and moral 'superiority' since just before the formation of their respective nation-states. European peripheries' ruling classes' interpretation of this apparent contradiction has been remarkably ambivalent as regards to relationships with the core as a socioeconomic and cultural complex. Quoting Bendix's terminology, the core is looked upon as a 'reference society' with which there is both an 'identification' and a 'distinction' relationship (Bendix 1978). Emulation of the core is encouraged while dissociation is maintained as well. Over the past 150 years, almost every European periphery underwent major modernization processes, emulating the core, which included all aspects of society. However, these processes were accompanied by 'archaistic' counter-trends (agrarian populist movements, nationalisms, etc.) which instead glorified what was deemed to be the 'past,' 'tradition,' 'culture,' 'identity' of one's group. Modernization drives and nativistic revivals were both at work.

A modernization drive basically means 'Europeanization,' particularly of politics (establishment of parliamentary institutions and legal systems), the economy (industrialization), culture (more theaters, art museums, avant-garde arts, etc.), and even everyday life (new habits, conventions, and ephemeral fads from the capitals of 'reference societies'—Paris, Vienna, or Berlin). Nativistic revival means stressing one's 'difference' to be distinguished from the core, and revitalizing 'imagined' social institutions, guardians and bearers of 'primordial virtues,' such as the village, rural community, medieval municipality, or guilds.

4. *Europe and the Sociocultural Order of Nation-states*

Since the nineteenth century, capitalist economic expansion in 'world-system' Europe has been accompanied by a proliferation of nation-states. They were established (except in the United Kingdom) through a combination or a separate use of two main models conceived in France and Germany first (Brubaker 1992). Notwithstanding some differences, both concepts are based on the doctrine by which each 'nation' is entitled to its own territorial state and each state must include only one 'nation.'

Because of this ideal, the last two centuries have been marked by repeated efforts to make single national territories more and more ethnically and culturally homogeneous. Already from the twelfth to the seventeenth century, Western Europe was overrun by several homogenization waves in the shape of religious conflicts and persecutions. Even in the more culturally homogeneous nation-states of Western Europe such as France, however, assimilation processes continued to the end of the nineteenth century (Weber 1976). In Central and East Europe by contrast, where the doctrine was applied only after the downfall of the multicultural empires, the processes of 'ethnocultural recomposition' aimed at 'ethnic purity' have been implemented only during the nineteenth and twentieth century, through a series of boundary revisions, forced assimilations, expulsions, aimed and planned immigrations, deportations, ethnic wars, and genocides.

Four main periods can be identified in the various processes of 'ethnic separation' that over the last two centuries concerned the European nation-states, especially in the central and eastern parts of the continent. The first period was predominantly in the Balkans, immediately after the creation of the first nation-states in the nineteenth century. Vast population groups of Turkish origin or merely of Islamic faith were forced to leave the region. As administrators of the Ottoman Empire, some indeed represented the hated occupiers, but members of social strata that had little in common with the ruling class were also expelled. During the great 'Crisis in the Orient,' which led to the bloody Russian–Turkish war, from 1875 to 1878, a million and a half people were repatriated.

The second virulent phase took place between 1913 and 1925. It was characterized by the forced transfer of whole minority groups, yet it was internationally recognized and guaranteed. In diplomatic language, it was euphemistically termed 'population exchange.' Substantial groups of Albanians from Kosovo and western Macedonia were transferred to Turkey after the Balkan Wars (1913) mainly because of their religion. Particularly after the creation of the Kingdom of Yugoslavia, the Islamic populations were replaced with Serbian, Montenegrin, Croatian, and Slovenian people. The 'population exchange' between Greece and Turkey was even more dramatic. The Treaty of Lausanne in 1923 ratified a series of reciprocal expulsions and hasty repatriations caused by the Greek military catastrophe during the campaign in Asia Minor. After the defeat, Greece was overrun by refugees from the coasts of western Anatolia, plus the Greeks from the Black Sea area and the Caucasus who had fled the new Bolshevik regime's repressions. A country with 4.5 million inhabitants faced the arrival of 1.3 million refugees. At the same time, the 'population exchange' provided for the departure of Turks and Albanians.

The third phase of 'ethnic homogenization' includes the decade between 1940 and 1950, marked by the Nazi policy of annihilation, transfer, and expulsion of whole ethnic groups, and by Stalinist deportations and

purges. Along with the holocaust of the so-called 'transnational minorities,' that is, Jews and Roma, there were massive population movements in Central and East Europe. Eleven and a half million Germans were expelled from the *Ostgebiete* (eastern territories) while three million Poles, particularly from the regions absorbed by the Soviet Union after World War II, were settled in Silesia and in the southern portion of eastern Prussia. Thus, Poland became an almost monoethnic country.

The treaties between Czechoslovakia and Hungary and between the latter and Yugoslavia, which also provided for reciprocal 'population exchanges,' date back to the same period. Finally, Stalin consolidated his conquests in the western part of the Soviet Union through a policy of 'planned mobility.' This involved the deportation of populations considered 'accomplices of the enemy,' therefore 'traitors of the great patriotic war' (Latvians, Lithuanians, Estonians, etc.), to Siberia or central Asia. It also involved replacing them with more 'reliable' immigrants, mainly of Slavic origin.

The fourth virulent phase of 'ethnic homogenization' is the wave of 'ethnic separations' that has devastated Central and East Europe over the past 15 years. It can be traced back to Socialist Bulgaria with its approach to minority problems. In the second half of the 1980s, the Bulgarian solution was to expel and/or forcibly assimilate 'ethnic Turks.' The phase continued with the disintegration of the three countries born after World War I through a multiethnic and multinational 'logic'; Yugoslavia, the Soviet Union, and Czechoslovakia. The resulting new and old states are all based on the nation-state principle. This latest wave of 'separations,' which aims at the homogenization of territories with highly differentiated ethnic structures and claimed by one or more 'nations,' continued in the 1990s in Bosnia and Kosovo, and may not yet have come to an end, particularly in the Balkans and former Soviet Europe.

5. *'Postnational' Europe in the Making?*

Over the last 50 years of the twentieth century, there were major population movements in Europe; this is the most outstanding characteristic of this period. These migrations for economic (job seekers) or political, that is, humanitarian (asylum seekers), reasons have affected Western Europe more than Central-east Europe. Until 1989, Soviet hegemony in this area allowed only a limited and often planned mobility, even though immigrant laborers were not totally lacking (Vietnamese in Bulgaria, former Czechoslovakia, German Democratic Republic, and the USSR as well). Within the EU and its previous community institutions a gradual, albeit slow, frontier abatement amongst member states has made transnational mobility easier. From the end of the 1950s, unprecedented numbers of laborers from the Mediterranean and Southeast peripheries (Italy, Spain, Portugal, Greece, and former Yugoslavia) reached the core (France, Germany, Austria, Switzerland, Benelux, United Kingdom, and Scandinavia). Here they found jobs in industry and in fields that are not sought after by the 'autochthonous' population because they are considered low status activities. From the 1970s on, these immigrants were joined by an increasing amount of laborers from the Mediterranean area bordering on the Old Continent (Turkey and Maghreb). Although to date some countries such as Germany and Switzerland challenge this denomination, northwestern European countries became 'immigration societies' in this period, notwithstanding strict measures introduced in the early 1970s regulating foreigners' permanent entry.

The circumstances in Euro-Mediterranean former 'emigration societies' (Italy, Spain, Portugal, Greece), which in the meantime have become 'immigration countries' instead, are slightly different from Northwestern Europe's situation. Migration is mainly from so-called Third World countries (Sub-Saharan Africa, Southeast Asia, China, the Indian subcontinent, Latin America, etc.), and from the continent's most underdeveloped areas (Albania). From the end of the 1980s, however, what became known as 'mass unemployment' worsened in all Western Europe. Therefore, recruitment of foreign workers fell drastically while political and humanitarian immigration increased.

At the same time, large numbers of migrants reached the core from non-European regions and former Socialist bloc countries after the fall of the Berlin Wall (1989) and the subsequent opening of frontiers. Facing a difficult socio-political transformation and an unstable economy, Central-east Europe is not the preferred destination for these migrants, who regard it only as a stopover on their way to Western Europe. Therefore, former 'Real Socialism' countries become a temporary 'buffer zone.'

The above-mentioned migratory circumstances have called into question the sociocultural order created by nation-states that is based on the ideal of a mono-ethnic territory. The growing presence of diasporas, transnational groups, creolized communities, and deterritorialized collectivities in Europe's social and cultural setting, plus minorities' emphatic claims and demands (linguistic, religious, and ethnic), have fired discussions on nation-state adequacy and the need for 'postnational' organizations with new types of 'transnational' or 'multiple' citizenships, as Ferry (1992), Pierré-Caps (1995), and Kleger (1997) foresee.

However, such projects will be for the twenty-first century. For the time being the nation-state is the political and sociocultural standard, both internationally and regionally, from which different types of autonomism, Euro-regionalism, federalism, 'devolution,' and separatism (Catalonia, Galicia, Spanish

Basque Country, Corsica, Flanders, Walloon Provinces, Scotland, 'Padania,' etc.) continue to draw inspiration when they seek an organizational form for their ambitions.

See also: Eastern European Studies: Culture; Eastern European Studies: Economics; Eastern European Studies: Politics; Eastern European Studies: Society; European Integration; Regional Geography; Regional Integration; Western European Studies: Culture; Western European Studies: Society

Bibliography

Anderson B 1983 *Imagined Communities. Reflections on the Origin and Spread of Nationalism*. Verso, London

Bendix R 1978 *Kings or People. Power and the Mandate to Rule*. University of California Press, Berkeley, CA

Bloch E 1985 *Tübinger Einleitung in die Philosophie*. Suhrkamp, Frankfurt am Main, Germany

Braudel F 1979 *Civilisation matérielle, économie et capitalisme, XVe-XVIIIe siècle. Le temps du monde*. A. Colin, Paris, 3 Vols

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Ferry J M 1992 Pertinence du postnational. In: Lenoble J, Dewandre N (eds.) *L'Europe au soir du siècle. Identité et démocratie*. Editions Esprit, Paris, pp. 39–57

Kleger H (ed.) 1997 *Transnationale Staatsbürgerschaft*. Campus Verlag, Frankfurt, Germany

Löwith K 1990 *Der Mensch inmitten der Geschichte. Philosophische Bilanz des 20. Jahrhunderts*. J. B. Metzler, Stuttgart, Germany

Morin E 1987 *Penser l'Europe*. Gallimard, Paris

Pierré-Caps S 1995 *La multination. L'avenir des minorités en Europe centrale et orientale*. Editions Odile Jacob, Paris

Riasanovsky N V 1984 *A History of Russia*. Oxford University Press, Oxford, UK

de Rougemont D 1990 *Vingt-huit siècles d'Europe. La conscience européenne à travers les textes—d'Hésiode à nos jours*. Bartillat, Paris

Schmitt C 1974 *Der Nomos der Erde im Völkerrecht des Jus Publicum Europaeum*. Duncker and Humblot, Berlin

Szücs J 1983 *Vázlat Európa három történeti régiójáról (The Three Historical Regions of Europe)*. Magvetö, Budapest, Hungary

Wallerstein I 1974 *The Modern World-System. Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

Weber E 1976 *Peasants into Frenchmen: The Modernization of Rural France, 1870–1914*. Stanford University Press, Stanford, CA

C. Giordano

European Integration

European integration is the product of the selective pooling of national sovereignty, or ultimate jurisdiction over a body politic, by postwar European nation-states (see *Sovereignty: Political*). Many regional organizations exist to promote and foster cooperation in Europe, above all the European Union (EU, formerly known as the European Economic Community, European Community or Common Market). From its origins in the 1950s, the EU expanded from six to 15 member states, with most of the rest of Europe seeking to join by the end of the twentieth century (see Table 1). The EU maintained close links to a variety of other organizations which promote or foster European cooperation, such as NATO, the Organization for Security and Cooperation in Europe (OSCE), and the European Bank for Reconstruction and Development. But the EU was by far the most powerful, important, and successful manifestation of European integration.

Table 1
The European Union's member states and applicants

Member states	Year of accession
Belgium, France, (West) Germany, Italy, Luxembourg, Netherlands	1957
Denmark, Ireland, United Kingdom	1973
Greece	1981
Portugal, Spain	1986
Austria, Finland, Sweden	1995
Applicants	
Bulgaria, Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Romania, Slovakia, Slovenia, Turkey	As of 2000

1. European Integration and Political Science

European integration poses daunting challenges to political science as a discipline. The EU's supranational character, particularly the legal superiority of its legislation over its member states' national laws, is replicated in no other international organization (see *International Organization*). The EU defies traditional conceptions of nation-states as atomized, self-sufficient units that resist interdependence and engage in alliances strictly on an *ad hoc* basis (see *Interdependence: Political*; *International Relations: Theories*). The Union is based on international treaties, but ones that function as a sort of constitution for a quasi-federalist system of government (Weiler 1999). Nonetheless, the EU sustains more than it supersedes the nation-state in Europe. It is plausible to argue that it has acted to strengthen—and not weaken—postwar European states by fostering economic growth and recovery, providing a bulwark against the Soviet bloc, and enhancing Europe's role in international affairs (Milward 1992).

European integration has been subject to frequent stops and starts. The EU rarely has shaken free of

tensions between distinct national interests and shared economic interdependence. However, a series of revisions to the EU's Treaties—four in 15 years after the mid-1980s—had the general effect of more closely binding together the Union's member states, strengthening its common institutions, and adding to its policy remit. It also encouraged the emergence of EU studies as a subdiscipline in its own right.

2. *European Policies and Politics*

Over time, the politics of the EU have become far more like the domestic politics of its member states, and less like the international politics of relations between states (Wallace and Wallace 2000). The primary reason for the emergence of the EU as a polity in its own right has been political will to create an internal or 'single' European market, free of barriers to internal trade and constituting the world's largest single capitalist market. The ambition to create an internal market was reflected in the EU's founding treaties, but began to be realized only after the Single European Act (SEA) effectively relaunched European integration in the late 1980s, following a period of deep stagnation and 'Eurosclerosis.'

Throughout this period, the European Court of Justice remained the ultimate arbiter of legal disputes, and many of its decisions advanced European integration as a political project. The EU's hybrid executive/civil service, the European Commission, retained exclusive powers to propose legislation. However, the SEA (ratified in 1987) substantially altered the Union's institutions and rules governing relations between them. Before the late 1980s most EU legislation required the unanimous agreement of the Council of Ministers, the EU's main legislative body offering representation to all member states. After the SEA, renewed political commitment to the internal market combined with extended majority voting on the Council—with each member state given a weighted number of votes based roughly on population—transformed European integration. Increased majority voting significantly empowered the Commission, which also achieved considerable autonomy in a number of policy sectors, including agriculture and competition policy. The volume of EU legislation mushroomed after 1987 and the Union developed a range of new policies, mostly in areas linked to the internal market such as environmental protection, research, and the development of its poorer regions. It even developed ambitions to create a single currency to take the internal market, and European integration, a major step further (see McNamara 1998).

Meanwhile, EU decision making became increasingly complex and protean. Council deliberations remained mostly closed, a dizzying variety of different decision rules applied to different policies, and polls reflected popular apathy or cynicism with the EU. The growing perception that European integration needed to be democratized allowed the European Parliament (EP) to gain in strength and stature. In 1979, it was directly elected for the first time. By the early 1990s, it had been made a legal colegislator with the Council in many policy areas (including the internal market). However, EP elections attracted low voter turnouts and the Parliament remained an ill-disciplined institution that engaged in effective collective action only rarely. The EU was far from being a parliamentary democracy. Its lack of either a government or opposition made it difficult to apply majoritarian methods of democratic control (see *Accountability*: *Political*).

EU Treaty revisions in the 1990s made the Union, on balance, an even more complex institution. The Maastricht Treaty (ratified in 1993) upgraded the EP's powers but also created two new 'pillars' to foster intergovernmental cooperation between member states on foreign and justice/home affairs policies, creating new (almost purely unanimous) decision procedures and sidelining the EU's supranational institutions. Neither pillar enjoyed much success, with the EU's performance during the Bosnian war in the mid-1990s particularly criticized despite the Maastricht Treaty's sanction of a new 'Common Foreign and Security Policy.' Very limited progress on justice and home affairs cooperation eventually led to a more powerful role for the EU's institutions in this area, but according to an entirely new set of decision rules not previously used in any other EU policy sector.

The Maastricht Treaty also set a firm timetable and entry criteria for monetary union. Most governments that wished to join the single currency were forced to slash their public budgets and embrace austerity when Europe was suffering through its worst recession since the 1930s. The drive to monetary union, along with the EU's failure in Bosnia, further dampened popular enthusiasm for European integration.

The Amsterdam Treaty (ratified in 1999) mostly tinkered at the margins of the EU's constitutional settlement, despite intense pressures to enlarge the EU to take in new democracies in Central and Eastern Europe. Qualified majority voting was extended to very few new policy areas even though it was impossible to imagine that a future EU of 25–30 member states could achieve unanimity very often. The inevitable effect was to raise doubts about the future political unity of a much larger and more diverse EU, particularly one that offered membership to new Central and Eastern European states which were far less modern or wealthy than its existing members.

3. *The Study of European Integration*

The study of early post-War institution building in Europe was dominated by functionalists (or 'neo-functionalists,' see Haas 1958), broadly in line with the rise of structural functionalism as a general theory of

social science. Neofunctionalism assumed that European integration was driven primarily by the need for states to cooperate in noncontroversial, technical sectors in order to grow their economies. It predicted that integration in selected areas of policy would 'spill over' into separate but related sectors, for both functional and political reasons, thus giving integration a self-sustaining character. It was assumed that the European experience of integration would be replicated in other parts of the world.

The stagnation of European integration in the late 1960s and 1970s led neofunctionalism, and EU studies more generally, to fall into disrepute. The reinvigoration of integration in Europe in the 1980s prompted an enormous increase in academic interest in the EU. Eventually, debates between advocates of liberal theories of international relations (Moravcsik 1998) and institutionalist theories of political organization (Armstrong and Bulmer 1998) came to dominate EU studies (see Puchala 1999). Liberal intergovernmentalism theorized that European integration was the outcome of the pursuit of economic advantage by rational state agents, with the EU's supranational institutions only granted as much power as was needed to seal bargains between them. Institutionalists countered that European integration was highly path-dependent, produced many unanticipated consequences, and was uncontrollable by states in important respects (see *Institutionalism*). Few advocates of either theory sought to disavow neofunctionalism altogether, particularly given the development of the EU's own system of law, powers of enforcement, and money.

Recent scholarship on European integration reflects consensus that the EU has evolved into a distinctive polity with multiple levels which is like no other in existence (see Marks et al. 1996). On the one hand, its rise to become one of the most important sources of public policy in Europe has prompted the application of conceptual tools associated with policy analysis in national (especially federal) contexts (Peterson and Bomberg 1999). On the other, the *sui generis* nature of the EU creates an '*N* of 1' problem: the EU is not a readily comparable case for most general theories or methodologies of social science. One effect is that scholarship on the EU is often subject to bitter criticism by nonspecialists, in part because it is viewed as nonscientific and atheoretical. However, a related problem is that European integration generally and the EU specifically are so complex as to resist simple generalizations and often comprehension by anyone but devoted students.

4. *Future Theory, and Research*

In some respects, the EU is becoming less *sui generis*, particularly given new regional economic projects in the Americas and Asia (see *Regional Integration*). The Union is likely to become the subject of more normative theory and research, as its perceived success makes it a model for other experiments in regional cooperation. The creation of the Euro has attracted a further group of scholars—monetary economists—to EU studies, and the Union's inevitable enlargement makes it a primary focus for scholars specializing in Central and Eastern Europe. International relations scholars preoccupied with military or 'strategic studies' are also likely to find the EU impossible to ignore as it seeks to develop its own military capability in the early twenty-first century.

In theoretical terms, scholarship on globalization increasingly offers insight into what the EU does, how and why (see *Globalization: Political Aspects*). Most EU policies are designed either to defend European interests in an increasingly global economy, or soften the nefarious effects of globalization. On balance, the continuing structural shift in power from states to markets encourages continued European integration. The EU is the most successful experiment in international cooperation in modern history. However, the enlargement of the EU is likely to transform the Union fundamentally, in ways that challenge the notion that ever closer European unity is inevitable.

Bibliography

Armstrong K A, Bulmer S J 1998 *The Governance of the Single European Market*. Manchester University Press, Manchester, UK

Haas E B 1958 *The Uniting of Europe: Political, Social, and Economic Forces, 1950–7*. Stanford University Press, Stanford, CA

Marks G, Hooghe L, Blank K 1996 European integration from the 1980s: state-centric v. multi-level governance. *Journal of Common Market Studies* **34**: 341–78

McNamara K R 1998 *The Currency of Ideas: Monetary Politics in the European Union*. Cornell University Press, Ithaca, NY

Milward A S 1992 *The European Rescue of the Nation-state*. Routledge, London

Moravcsik A 1998 *The Choice for Europe*. Cornell University Press, Ithaca, NY

Peterson J, Bomberg E 1999 *Decision-making in the European Union*. St Martin's Press, New York

Puchala D J 1999 European integration and supranational governance. *Journal of Common Market Studies* **37**: 317–31

Wallace H, Wallace W (eds.) 2000 *Policy Making in the European Union*, 4th edn. Oxford University Press, Oxford, UK

Weiler J H H 1999 *The Constitution of Europe*. Cambridge University Press, Cambridge, UK

J. Peterson

European Religion

The concept of Europe—whether considered geographically or culturally—is not self-evident. While the Western, coastal edge is clear enough, the distinctiveness of Europe within the Eurasian land mass

is less so. The first part of this article will consider the relevance of the religious factor to the definitional question; it will also outline the major confessional divisions *within* the continent. The second section will look at the patterns of religious life in different parts of Europe, underlining the contrasts between 'hard' and 'soft' indicators of religion in this respect. The final section deals with the bewildering variety of church–state relationships in modern Europe.

1. *The Religious Boundaries of Europe*

The complexities involved in the concept of Europe are clearly outlined by Davies (1996). In the early modern period, the idea of Europe gradually disentangled itself from the concept of 'Christendom,' a process in which the religious factor played a decisive role. As the nations of Europe endured centuries of religious war, they were disinclined to remind themselves of what they supposed to have in common: the notion of Europe provided a more neutral alternative. A longer historical perspective reveals, however, that Christendom had itself been divided long before the early modern period: between an Orthodox East (centered first in Constantinople and then in Moscow) and a Catholic West (centered in Rome), a dispute which came to a head for a mixture of theological and political reasons in the eleventh century.

The central question of Europe's self-understanding is closely connected to this division: should this or should it not include Russia, an Orthodox, economically backward, and slavophile state? This article will include Russia, but will at the same time emphasize the distinctiveness of the Orthodox part of the continent. Such distinctiveness takes the form of a markedly different cultural heritage. The nations associated with Western Christianity have, for example, experienced the Renaissance, the Reformation, the scientific revolution, the Enlightenment, and the Romantic revival (and take these pivotal movements for granted). Countries further east have been through an equally formative, but entirely different historical evolution. For them the ebb and flow of Islam has been of greater significance. Indeed, it is possible to argue that the concept of Europe finally gives way to something else as Orthodoxy gives way to Islam, on a line which has fluctuated over centuries and continues, from time to time, to play a crucial and not always peaceful role in the establishment of identities. The 1990s disputes between Russia and Chechnya or between Armenia and Azerbaijan are cases in point.

Given its long-term cultural significance, the boundary between Catholic and Orthodox is of more importance in understanding Europe's religious life than the much more recent division between communist and noncommunist Europe. With this in mind, it is useful to introduce the terms 'East' and 'central' Europe when dealing with the former communist countries. East Europe (European Russia, Romania, Bulgaria, Serbia, and Macedonia) belong to the Orthodox tradition, whereas the central European countries (the Baltic states, Poland, the former East Germany. the Czech Republic, Slovakia, Hungary, Slovenia, and Croatia) developed within Western Catholicism. It is only in the postwar period that they have been separated from the economic and cultural traditions of the West. The position of Greece is anomalous in this respect; it is the only Orthodox country within the European Union. In a much longer timescale, the contribution of Greek culture to the understanding of what it means to be European in the Western sense can hardly be overestimated.

The border between 'East' and 'central' Europe is, however, as problematic as any other, for on it are found the Greek Catholic or Uniate churches—those, in other words, which have been received under the jurisdiction of Rome, but which retain their own ritual, practice, and canon law. In the emergent identities of this part of Europe, moreover (notably in the Ukraine and in Belarus), the Uniate churches are of considerable importance. Further south and in a rather different way, Bosnia-Herzogovina is uncomfortably pincered between Orthodox Serbia and Catholic Croatia, a situation rendered all the more complex by the presence of a sizable Muslim population dating from the time of the Ottoman occupation.

So much for the east–west boundaries of Europe. The subsequent divisions of the west into areas or nations which are primarily Catholic or Protestant (or combinations of the two) are inseparable historically from the emergence of the nation state as the dominant form of Europe's political life. The processes by which such divisions occurred are highly complex, involving economic, social, political, as well as religious issues. Which of these led to the others and how the whole thing was set in motion in the first place is the subject of an ongoing debate among historians, themselves of different persuasions. What remains indisputable, however, is an unprecedented upheaval in the ordering of Christian society in the sixteenth century; an upheaval which included the emergence of separate political entities or nation states, some of which expressed their independence from papal interference in the form of a state church.

Part of the ambiguity regarding the whole historical process lies in the understanding of the term 'Reformation.' Does this imply innovation and the breaking of new ground? Or does it involve a return to and rediscovery of primitive excellence? Were those who endorsed the theological changes taking place at this time looking primarily for radical change or for conservative independence? Motives were mixed. The more conservative interpretation, however, was bound to appeal to those political rulers anxious to establish independence from external authority, but with a careful eye on stability within. Both were possible

within the Lutheran concept of a 'godly prince.' Sometimes the prince had jurisdiction over a whole kingdom or kingdoms. Such was the case in Scandinavia where Lutheranism became embodied in the state churches of Northern Europe. Elsewhere the process was far more local and concerned relatively small patches of land. The German case exemplifies the latter, leading to patterns which are not only extant but influential some 400 years later.

The Reformation took different forms in different places. In addition to Lutheranism, parts of Europe—notably the Swiss, the Dutch, the Scots, some Germans, some Hungarians and Czechs and a small but significant minority of French people—were attracted first by Ulrich Zwingli, but then by Jean Calvin, towards a more rigorous version of Protestantism. Calvinism was both more radical and more restrained: radical in the sense of a new kind of theology based on the doctrines of predestination and redemption, but restrained in terms of its stringent moral codes. The effect of this particular combination on the subsequent economic development of Europe has provided inexhaustible material for an ongoing debate among historians and sociologists alike.

Broadly speaking West Europe divided itself into a Catholic South (Spain, Portugal, Italy, France, but including Belgium and Ireland), a Protestant North (Scandinavia and Scotland), with a range of 'mixed' countries in between (England and Wales, Northern Ireland, The Netherlands, and Germany). Central Europe exemplifies similar categories, though the geography is rather more complicated. Lithuania, Poland, Slovakia, Slovenia, and Croatia are firmly Catholic; Estonia and Latvia are Lutheran and relate closely to their Scandinavian neighbors (a commonality strongly re-emphasized as the Baltic Republics regained their political independence); Hungary and the Czech Republic, finally, are rather more mixed (primarily Catholic but with significant Protestant minorities). In other words, boundaries gradually emerged all over Europe dividing one nation from another, one region from another, and one kind of Christianity from another. Boundaries, moreover, imply dominance as well as difference. Majorities and minorities were, and still are, created depending on the precise location of the line in question. One of the most arbitrary in recent years has been the line that divides Northern Ireland from the Irish Republic.

2. *Patterns of Religious Life in Europe*

The European Values System Study Group (EVSSG) provides a principal source of data for this section (Barker et al. 1992). Using careful sampling techniques, the EVSSG aims at an accurate mapping of social and moral values across Europe. The first study was carried out in 1981, the second in 1990 and a third—unavailable at the time of writing—in 1999. Initially only some countries in West Europe were involved, a coverage which quickly extended to most of Europe and beyond. The following discussion starts with West Europe; East and central Europe will be dealt with subsequently.

There are, broadly speaking, five religious indicators within the EVSSG data: denominational allegiance, reported church attendance, attitudes towards the church, indicators of religious belief, and some measurement of subjective religious disposition. These variables have considerable potential: they can be correlated with one another and with a wide range of sociodemographic data. In this respect the survey shows commendable awareness of the complexity of religious phenomena and the need to bear in mind more than one dimension within a nation's religious life. What emerges in practice, however, with respect to these multiple indicators is a clustering of two types of variable: on the one hand, those concerned with feelings, experience, and the more numinous religious beliefs; on the other, those which measure religious orthodoxy, ritual participation, and institutional attachment.

It is, moreover, the latter (the more orthodox indicators of religious attachment) which display, most obviously, an undeniable degree of secularization throughout Western Europe. In contrast, the former (the less institutional indicators) demonstrate considerable persistence in some aspects of religious life. Bearing this in mind—it seems more accurate to suggest that West Europeans remain, by and large, unchurched populations rather than simply secular. For a marked falling-off in religious attendance (especially in the Protestant North) has not resulted yet in a parallel abdication of religious belief. In short, many Europeans have ceased to belong to their religious institutions in any meaningful sense, but they have not abandoned, so far, their underlying religious aspirations (Davie 2000).

Two short parentheses are important in this connection. The first may seem obvious, but the situation of believing without belonging (if such it may be called) should not be taken for granted. This widespread, though fluctuating, characteristic within European religion in the late twentieth century must be examined, probed, and questioned. The second point illustrates the need for caution. At the same time it introduces the contrasting situation in parts of East and central Europe where believing without belonging was not the norm. Indeed, prior to 1989, the two variables were (in some places at least) reversed, for the nonbeliever quite consciously used Mass attendance as one way of expressing disapproval of an unpopular regime. The Polish case is the most obvious illustration of this tendency (Michel 1998).

Returning to the West European data and beginning to probe more deeply, there is further evidence of consistency in the shapes or profiles of religiosity which have been obtained across a wide variety of

European countries. Whatever the overall level of belief in any given country, for example, the following variables—belief in God, in a soul, in heaven, in life after death, in the devil, in hell and in sin—almost always come out in the same order. Correlations between religious indices and socioeconomic variables confirm the existence of socioreligious patterning across national boundaries. The correlation with age is particularly striking, and raises the future shape of European religion. Indeed, it prompts the most searching question of the study: is it the case that West Europe is experiencing a permanent generational shift with respect to religious behavior, rather than a manifestation of the normal life-cycle? The EVSSG findings seem to indicate that this might be so. Data from the 1990 restudy reinforce this point.

So much for the similarities across West Europe. What about the differences? The first and most obvious of these lies between the notably more religious, and Catholic, countries of Southern Europe and the less religious countries of the Protestant North. This variation holds across almost every indicator; indeed, they are inter-related. Levels of practice, for example, are markedly higher in Italy, Spain, Belgium, and Ireland (closer in its religious life to continental Europe than to Britain) than they are elsewhere. Not surprisingly, one effect of regular Mass attendance is a corresponding strength in the traditional orthodoxies through most of Catholic Europe.

There are, however, exceptions to this rule, and at this point it is necessary to point to the limitations in the EVSSG data. There is no way of telling from the data why a particular country should be similar to or different from its neighbors. France, for example, displays a different profile from the other Catholic countries (with markedly lower levels of practice), a contrast that cannot be explained without reference to the specificities of French history. Other exceptions to a European pattern, or patterns, should be looked at in a similar light; notably, the countries which do not conform to the believing without belonging framework. Conspicuous here are the two Irelands in both of which practice as well as belief remains unusually high. Within central Europe, the Polish case reflects a similar pattern. As ever, there are historico-political reasons for these exceptions.

A crucial point concludes this section. It concerns the patterns of practice and belief in East and central Europe and their possible future(s). In making this assessment, it is important once again to remember the distinctiveness of Orthodoxy, a form of belief which contains within itself a greater degree of diversity than is possible in the Western churches (it is more inclusive). It is also different in its relationship to political power (see below). The profiles of religion in East (and indeed central) Europe are, moreover, not only varied but volatile—in some places the indices rise sharply after 1989, in others less so. With this in mind it is probably too soon to answer the key question. Will Eastern Europe follow the West into a state of relative religious indifference (increasingly unusual in the modern world) or will this part of Europe emerge a distinct alternative, following its own religious trajectory? Only time will tell.

3. *Church–State Relationships*

A broad brush profile of religion in West Europe provides the background for a more detailed discussion of church and state relationships; these in turn constitute the parameters within which religious life on a national level takes place. It is helpful to set a discussion of church–state relationships within the following framework, taken from Martin's seminal work on the evolution of secularization in the Western world (Martin 1978). Following Martin, Europe is a unity by virtue of having possessed one Caesar and one God (the legacy of the Judaeo-Christian tradition); it is a diversity by the existence of nations. The patterns of European religion derive from the ongoing tensions and partnerships between religion and the search for national integrity and identity. These tensions and partnerships have resulted in a bewildering variety of church–state relationships within the continent as a whole.

In Western Europe, such relationships range from the state churches of Northern Europe, through the pillarized cultures in The Netherlands and Belgium, to the considerable diversity contained within the United Kingdom and Ireland (including a Calvinist national church in Scotland, an Anglican establishment in England, assorted nonconformity in Wales, and a technically secular but heavily Catholic state in Ireland). Germany is a biconfessional state in which both the Lutheran and Catholic churches remain important public bodies. Latin Europe includes the privileged Catholic churches of Italy, Spain, and Portugal and the rather different situation in France (the only European nation, for example, with a rigorously secular school system in addition to the state itself). In this part of Europe (and especially in France), anticlericalism has a historically important role—itself diminishing as the Catholic hegemony declines.

In East and central Europe, it is necessary to start with the autocephalous churches of Orthodox Europe, closely aligned with political power over centuries (of which Russia is the prime example). In central Europe there are examples (notably Poland) of the Catholic Church becoming the effective carrier of an *alternative* memory in the communist period; elsewhere—in Hungaria and Czechoslovakia—the churches were more compromised. The Lutheran churches (in, for example, Estonia or the former East Germany) all but collapsed altogether—notwithstanding the outstanding courage of individual believers. In all these cases, the postcommunist period brings a new set of chal-

lenges in terms of church–state relationships and the associated privileges.

Indeed, in a whole variety of ways, the Europe that is emerging as the twenty-first century dawns is a rapidly changing place. From a religious point of view, one of the most significant evolutions is undoubtedly the increasing representation of faiths other than Christian. Analytical concepts must evolve accordingly. European religion is giving way to the 'religions of Europe'; a continent which now houses a significant representation of Muslims, Sikhs, Hindus, and Buddhists in addition to the Jewish communities which have played such a crucial role in Europe's recent history. Martin's tension between unity and diversity remains paramount, though in forms that are peculiar to late modern rather than early modern European society.

A final section of the population should be mentioned in conclusion: the sizable minority of Europeans who declare that they have no religion and are not affiliated to any religious institution. A striking but overlooked factor within this significant minority is its internal diversity. Not only does it vary from the cautiously agnostic to the aggressively atheist, it is unevenly present in different European nations and in different socioeconomic groups. From a sociological point of view, the lack of material concerning this significant minority is striking.

See also: Catholicism; Europe: Sociocultural Aspects; Orthodoxy; Protestantism and Gender; Religion, History of; Religion, Sociology of; Western European Studies: Religion

Bibliography

Barker D, Halman L, Vloet A 1992 *The European Values Study 1981–1990: Summary Report*. Gordon Cook Foundation, Aberdeen, UK

Davie G 2000 *Religion in Modern Europe: A Memory Mutates*. Oxford University Press, Oxford, UK

Davies N 1996 *Europe: A History*. Oxford University Press, Oxford, UK

Martin D 1978 *A General Theory of Secularization*. Blackwell, Oxford, UK

Michel P 1998 East Europe. In: Wuthnow R (ed.) *The Encyclopaedia of Politics and Religion*. Routledge, London

G. Davie

European Union Law

The European Union was established by the Treaty on European Union, signed at Maastricht on February 7, 1992 (Official Journal of the European Community – OJ – 1992 C 224/1). It relates to three areas commonly called the pillars of the Union. The first consists of the Communities, which are often thought to be a single Community but are, in fact, three Communities: the European Community (EC) established as the European Economic Community by the Treaty of Rome of March 25, 1957 (United Nations Treaty Series – UNTS – Vol. 298, p. 11), the European Coal and Steel Community (ECSC) established by the Treaty of Paris of April 18, 1951 (UNTS Vol. 261, p. 140), and the European Atomic Energy Community (EURATOM) established by a separate Treaty of Rome of March 24, 1957 (UNTS Vol. 298, p. 167). While these Communities are invested with supranational powers, the second and third pillars of the European Union are constituted by policies and forms of intergovernmental cooperation in the fields of 'common foreign and security policy' and 'co-operation in the fields of justice and home affairs,' renamed 'police and judicial co-operation in criminal matters' by the Treaty of Amsterdam of October 2, 1997 (OJ 1997 C 340/1). The three Communities were set up by six founding member states (Belgium, France, Germany, Italy, Luxembourg, and the Netherlands) but later accessions (Austria, Denmark, Finland, Great Britain, Greece, Ireland, Portugal, Spain, and Sweden) have made the Community grow to 15 member states. A further extension of the Community towards central and eastern Europe is foreseeable. The success of this integration is almost entirely due to the transfer of sovereign powers to the three original communities and in particular to the European Community. The following article will therefore essentially deal with European Community law.

1. A Community Based Upon the Rule of Law

As the Court of Justice has repeatedly held, the European Community is a 'Community based on the rule of law, in as much as neither its member states nor its institutions can avoid a review of whether the measures adopted by them are in conformity with the basic constitutional charter, the Treaty' (Case C-2/88 Imm. – *Zwartveld* – [1990] European Court Reports (ECR) I – 3365 para. 16). In comparison with previous models for the peaceful cooperation and integration of states, the Community approach is outright revolutionary. While political considerations and diplomacy still have a great impact on the settlement of conflicts between member states, national sovereignty is no longer unlimited. States have to respect Community law lest violations are ascertained by the Court of Justice. Access to the Court is not limited to the member states and the institutions of the Community, legal proceedings may also be commenced by national courts of law which can request a preliminary ruling concerning the interpretation or the validity of Com-

munity Acts (article (art.) 234 EC). The enforcement of Community law is thus indirectly entrusted, beyond national governments, to a great number of national judges. While a change of government policy usually threatens to undermine the application or even the existence of international obligations which states have assumed under traditional treaties, Community law is much more stable due to its decentralized enforcement. An agreement of some, or even all, member state governments not to comply with certain Community acts would fail sooner or later as long as it is not embodied in an amendment of the basic Treaty which, however, has to be ratified by the parliaments of all member states. The repercussions on national policies are clear: member states are compeled to conduct their policies within the Community institutions. The pursuit of a separate policy outside that framework will often meet with opposition from Community law and the agencies called upon to enforce it. Within the Community institutions it is, however, almost impossible to pursue a policy of segregation. Thus, the rule of law established by the Treaty of Rome has narrowed the political margin of discretion left to the member states and has channeled their policies towards integration.

2. *Institutions*

Under art. 7 EC the tasks of the Community are carried out by five institutions: the European Parliament, the Council, the Commission, the Court of Justice, and the Court of Auditors. In addition, arts. 8 and 9 EC provide for a European Central Bank and a European Investment Bank.

In the political process, only the European Parliament, the Council, and the Commission are involved. While the traditional model of the Western constitution provides for a separation of legislative and executive powers between parliament and government, the respective roles of the three European Institutions are more complicated. The Commission may originally have been conceived as a kind of European executive which had to implement the integration program of the Treaty. But this task implies wide margins of political discretion, concerning both the priorities and the content of the measures to be adopted. Moreover, the Commission is the only institution which is entitled under the Treaty to initiate legislative proceedings. It is therefore a political heavyweight in the integration process. The governments of the member states have the right to nominate one or, in the case of bigger member states, two members of the Commission who are, however, completely independent in the performance of their duties (arts. 213 s. EC).

It is up to the Council to represent the interests of member states within the Community. The Council consists of a representative of each member state. While each representative may veto a decision where unanimity is required, the votes of the Council members are weighted in accordance with the size of the member state where the Treaty requires a decision taken by qualified majority (art. 205 EC).

The European Parliament was initially named the Assembly and had very little weight in the legislative process, its rights being limited to mere consultation. However, various amendments to the Treaty have improved the legal situation of the Parliament and thereby increased its political influence. As of 2000, its members are elected by direct universal suffrage (art. 190 EC), which however grants a much greater relative weight to the votes stemming from small member states like Luxembourg than to those originating in big member states like Germany. While the rights of the Parliament are still limited to mere consultation in some areas, such as immigration policy (art. 67 EC), most legislative acts now have to be approved by both the Council and the Parliament in the so-called codetermination procedure under art. 251 EC. The Parliament still lacks the power to initiate legislation by its own motions.

3. *Legislative Powers of the Community*

The Community has no inherent powers, it can only act within the limits of the powers conferred upon it by the member states in the Treaty (art. 5 (1) EC). However, the effect of that limitation is mitigated by three factors: first, the Court of Justice interprets the empowering provisions of the Treaty in a wide manner; second, the Treaty contains provisions such as art. 308 EC that confer upon the Community an open-ended power; and third, by the doctrine of implied powers, the Court of Justice has extended the Community's powers even beyond the limits imposed by the Treaty to international commitments insofar as they help to protect the uniform application of Community legislation vis-à-vis third states (Case 22/70 – ERTA – [1971] ECR 263 para. 28).

The exercise of Community competence is further limited by the principles of subsidiarity and proportionality (arts. 5 (2) and (3) EC). The legal significance of the principle of subsidiarity is rather limited as compared with its sweeping use in political discussions. It does not relate to the objectives of a Community action which have to be established in conformity with the Treaty by the political institutions. Subsidiarity pertains only to the question of whether such objectives can better be achieved by the Community than by the member states, the so-called comparative efficiency test. It is highly questionable whether the Court of Justice is in a better position than the political institutions of the Community to decide upon that issue. This may be different in the case of the

principle of proportionality which prohibits the Community from taking action that goes beyond what is necessary to achieve the objectives of the Treaty. In the context of the basic freedoms (see Sect. 6), the Court of Justice has in fact frequently assessed the proportionality of national restrictions imposed by member states.

4. *Sources of Community Law*

It is common to distinguish primary and secondary, or derivative, Community law. The primary law of the Community consists of the founding treaties indicated in the introductory paragraph of this article and some additional international conventions. Moreover, the general principles common to the laws of the member states are sometimes classified as part of the primary Community law. These general principles are explicitly referred to in art. 288 (2) EC in the context of the noncontractual liability of the Community for damage caused by its institutions or by its servants; the Court of Justice considers the general principles as a source of law even outside that field and has recourse to them in other matters, too. However, if the distinction between primary and secondary sources of law indicates a ranking, it would appear that only certain general principles such as the fundamental rights can actually count among the primary sources of Community law. Whatever the status of the general principles may be, they allow for the flexibility that every legal system needs in order to adjust to new and unforeseen situations.

Article 249 EC lists five types of secondary sources of Community law: regulations, directives, decisions, recommendations, and opinions. While recommendations and opinions are not binding, most Community acts are adopted either in the form of a regulation or a directive. Regulations are directly applicable in all member states which means that the Community can impose obligations on, and grant rights to, the citizens of the member states without the intervention of national legislators. As opposed to that, a directive is binding only upon each member state to which it is addressed, and only as to the result to be achieved; to have any effect on the citizens of the Community it must be implemented by member states in accordance with the forms and methods of their legal systems (art. 249 (3) EC). It soon turned out that member states are sometimes slow or reluctant to implement directives. In a long line of cases the Court of Justice has therefore extended their effects: national laws, even if enacted prior to the directive, have to be interpreted so as to conform with it (Case 14/83 – *von Colson* – [1984] ECR 1891 para. 26); insofar as the directive is sufficiently precise and deals with 'vertical' relations involving the defaulting member state it may even be applied directly by the national courts and authorities (Case 8/81 – *Becker* – [1982] ECR 53 para. 24 s.); while such a direct effect is not admitted in 'horizontal' relations (Case C-91/92 – *Faccini Dori* – [1994] ECR I-3325 paras. 22 seq.), the private party suffering loss from the non-implementation of the directive may invoke the defaulting member state's government liability (see Sect. 5). Although case law has approximated the directive to the regulation, the differences are still conspicuous: the regulation creates uniform law whose appearance is the same in all member states, the directive provides for a harmony of solutions at best, but even after that harmonization it is still difficult to assess the precise state of the law in other member states in the area at issue.

It is generally acknowledged, and has been affirmed on various occasions by the Court of Justice, that Community law is a separate legal system distinct from both international law and the national laws of the member states. Ever since the leading case *van Gend and Loos* was decided in 1963, the Court has also stressed the need for a uniform application of Community law in all member states (case 26/62, [1963] ECR 1). Two consequences follow from this premise: Community law has to be interpreted autonomously, i.e., without reference to definitions established in national law; moreover, as recognized in *van Gend and Loos*, Community law must be given direct effect and precedence over national law in so far as the Community provisions are sufficiently clear and unambiguous. In the course of time the conditions for direct effect were watered down. At the end of the twentieth century, direct effect is the norm rather than the exception; it applies to many Treaty provisions, to regulations, and even to many directives (see above in this Sect.). By the doctrine of direct effect the Court of Justice has obliged the national courts to enforce Community law, a duty which has been accepted over time by the national judiciaries.

5. *Ensuring Compliance with Community Law*

Both the Community institutions and the member states must comply with Community law. The Community institutions are subject to legal review in respect of their acts and their failure to act (arts. 230, 232 EC). In case of a violation of its laws committed by the Community or by its servants the Community may incur liability under art. 288 (2) EC. With regard to the member states, the Treaty ensures their compliance with Community law in a much less comprehensive way. Under arts. 226 and 227 EC the Commission and other member states are entitled to bring a violation before the Court of Justice. In practice, only the Commission has made use of this right. The Treaty does not provide for a direct right of action of individuals who complain of a violation of Community law committed by a member state, nor does it

establish the liability of member states for such violations. Both gaps have been filled by judge-made law. By the aforementioned doctrine of direct effect the Court has enabled and instructed the national courts to disregard national provisions which violate Community law. Private parties may therefore invoke Community law in national courts, which can ask the Court of Justice for a preliminary ruling under art. 234 EC if they have doubts about the interpretation of the relevant Community provisions. In addition, the Court of Justice has established the member state's liability for violations of Community law. Insofar as the respective member state has manifestly and gravely disregarded the limits of its political discretion it has to compensate private individuals who have suffered loss as a consequence of that violation. This concerns in particular the cases of a continuous breach of the obligation to implement a directive (cases C-46/93 and 48/93 – *Brasserie du Pêcheur vs. Germany and R. vs. Secretary for Transport, ex parte Factortame* – [1996] ECR I – 29; cases C-187 et al. – *Dillenkofer vs. Germany* – [1996] ECR I – 4845).

6. *Principles of the Internal Market*

Among the various Community policies listed in art. 3 EC the establishment of the Common Market which was later renamed the Internal Market is of central significance. With regard to other policies, the Treaty either limits Community activities to 'a contribution' to efforts made by the member states (e.g., health protection, education, consumer protection), contains vague indications as to the direction of a Community policy (economic and social cohesion, research, trans-european networks), or empowers the Community only to conduct a policy in a given area without indicating any targets (agriculture, transport, environment, energy). As opposed to these policies, the establishment of a Common or Internal Market is not only unequivocal objective stated in arts. 2, 3, and 14 EC, but the basic principles of such a market are also spelt out in clear and operative rules throughout the Treaty. Thus, art. 14 (2) EC defines the Internal Market as 'an area without internal frontiers in which the free movement of goods, persons, services and capital is ensured.' These basic freedoms are guaranteed in further provisions of the Treaty which the Court of Justice has interpreted as producing direct effects. National provisions which subject certain markets and activities to regulations in the interest of consumers, of the safety, of the environment, etc. are therefore regarded as restrictions of the basic Community freedoms and will be approved only if mandatory requirements of the public interest so command (case 120/78 – *Cassis de Dijon* – [1979] ECR 649 paras. 8, 9). By elaborating, in a long line of cases, both the basic rule of free movement and the exception in favor of national restrictions the Court of Justice effectively paved the way for the deregulation of many markets in Europe in the 1980s and 1990s. While political and economic discussions on deregulation were often deadlocked, a single judgment of the Court of Justice would announce a new deal for further negotiations. The case law being inspired by a strong stance against all kinds of discrimination on grounds of nationality (art. 12 EC), the Court has effectively opened national markets to competitors from other member states.

The basic freedoms are directed against restrictions imposed by states. Their implementation extends the scope of private liberties to engage in certain activities, but individuals and companies are not compelled actually to exercise them. Market actors may instead fear the intensification of competition and therefore try to maintain by private agreement the barriers to market access which were previously ordered by national legislation. The risk of such anticompetitive agreements and practices had been foreseen by the founders of the Community who explicitly provided for the prohibition of anticompetitive agreements and abuses of monopoly powers in arts. 81 and 82 EC. The enforcement of competition law which is entrusted to the Commission and to National Competition Authorities must be regarded as complementary to the basic freedoms: while the latter purport to open up the markets by removing national restrictions, the former are meant to protect that competition against private restrictions.

While the basic freedoms have been implemented to a large extent, the Internal Market is far from being a reality in many sectors. This is due to a great number of remaining and sometimes hidden obstacles. Thus, while nationality requirements for certain jobs have been abolished all over Europe, it is still lawful for member states to require a particular type of qualification which, in the absence of harmonization in the field of education, can hardly be acquired by foreigners in other member states; consequently, access to those professions and trades is still closed for foreign applicants. Another impediment to the implementation of the Internal Market consists in the diversities relating to the legal framework of transactions, in particular in civil and commercial law; they deter many small traders from making use of their basic freedoms. While the Community is empowered to approximate the laws of the member states for the purposes of the Internal Market (art. 95 EC), it makes slow progress in that area.

7. *Fundamental Rights*

The EC Treaty does not contain a list of unalterable rights. While some fundamental rights of that type are guaranteed by single provisions scattered over the

Treaty such as prohibitions of discrimination or the basic freedoms, their scope is limited. It is therefore possible that secondary Community legislation infringes private rights which are protected by fundamental rights granted by national constitutions. In respect of such situations the German constitutional court indicated that it would subject secondary Community legislation to the basic rights granted by the German constitution as long as the Community had not developed a charter of basic rights of its own. A constitutional dispute between the Community and some member states appeared inevitable. As a response the European Court of Justice started to acknowledge the fundamental rights flowing from the constitutional traditions common to the member states and from the European Convention on Human Rights (UNTS Vol. 213, p. 221) as an integral part of the general principles of law protected by the Court (case 11/70 – *Internationale Handelsgesellschaft* – [1970] ECR 1125 at 1134). An attempt of the Community to accede to the European Convention on Human Rights failed when the Court ruled that the Community lacks the relevant power without an amendment to the Community treaties (opinion 2/94 – ECHR – [1996] ECR I – 1759). At present, a written affirmation of fundamental rights is absent in Community law, while the Treaty on European Union provides in art. 6 (2) that the Union shall respect fundamental rights as guaranteed by the European Convention on Human Rights and as they result from the constitutional traditions common to the member states. This assertion is of doubtful significance and does not go beyond the case law of the European Court of Justice.

See also: European Integration

Bibliography

Cartou L, Gruber A, Clergerie J-L, Rambaud P 2000 *L'Union Européenne*. Dalloz, Paris

Grabitz E, Hilf M, Borchardt K-D *Das Recht der Europäischen Union*. 4 Vols. Beck, Munich, Germany

Hartley T 1998 *The Foundations of European Community Law*, 4th edn. Oxford University Press, Oxford, UK.

Kapteyn P J G, VerLoren van Themaat P 1995 *Inleiding tot het recht van de Europese Gemeenschappen*, 5th edn. Kluwer, Deventer, The Netherlands

Lenaerts K, van Nuffel P 1999 *Europees Recht*. Maklu, Antwerpen, Belgium

Mégret J, Louis J-V, Vignes D 1970 *Le Droit de la Communauté Economique Européenne*. Bruylant, Brussels, Belgium

Mengozzi P 1997 *Il diritto comunitario e dell'Unione europea*. Trattato di diritto commerciale e di diritto pubblico dell'economia, diretto da Galgano F. Cedam, Padova, Italy, Vol. 15

Oppermann T 1999 *Europarecht*, 2nd edn. Beck, München

Tesauro G 2001 *Diritto comunitario*, 2nd edn. Cedam, Padova, Italy

von der Groeben H, Thiesing J, Ehlermann C-D 1999 *Kommentar zum EU-/EG-Vertrag*, 5th edn. Nomos, Baden-Baden, Germany

Weatherill S, Beaumount P 1999 *EU Law*, 3rd edn. Penguin, London

Wyatt D, Dashwood A 1993 *European Community Law*, 3rd edn. Sweet and Maxwell, London

J. Basedow

Euthanasia

1. Introduction

In the original Greek, 'euthanasia' meant 'good dying' (*eu thanatos*). First used by Greek poets from about 400 BC onwards, the term was adopted by Greek and Roman philosophers and historians to denote a death that comes quickly and without great and prolonged suffering. The term 'euthanasia' was not used for the intervention, medical or otherwise, to ease a patient's death or to end his or her suffering. Such intervention cannot, however, have been uncommon in antiquity, since assisting a patient's suicide was explicitly ruled out in the so-called Hippocratic oath. Whoever swore this oath (probably the members of a small Pythagorean sect of physicians) committed him- or herself to 'give no deadly medicine to anyone if asked, nor suggest any such counsel.'

In modern times, the meaning of the concept has considerably changed. Though its core meaning of 'good dying' was preserved, it no longer refers to the quality (however defined) of the process of dying as such, but to the cooperation of someone else (the physician, the priest, or others) in making death easier for the dying. Francis Bacon, in a passage of his *De dignitate et augmentis scientiarum* (1623), was the first to distinguish two kinds of assistance to which, according to him, a dying patient has a claim, *euthanasia interior*, the 'spiritual preparation of the soul' for death, and *euthanasia exterior*, 'whereby the dying may pass more easily and quietly out of life' (Bacon 1857, p. 387). Bacon's point was that the purely spiritual preparation provided by the priest was not enough.

What should be introduced was what we now call palliative medicine, the systematic effort to relieve pain and suffering of the dying by medical means. Euthanasia in this sense, Bacon insisted, was not only a natural part of the duties of the physician but also a requirement of plain humanity. Though psychological and spiritual assistance in dying is by now recognized as one of the central elements in the care for the dying (often, however, neglected in practice), the modern concept of euthanasia includes the 'easing of death'

only in so far as this is effected by the application or nonapplication of medical means.

2. Five Types of Euthanasia

The principal meanings of the term 'euthanasia' in ethical and legal discussion correspond to the various forms medical assistance in dying can take.

(a) Relieving suffering by palliative care, e.g., by giving pain-killing drugs and suppressing symptoms such as vomiting or shortness of breath. Palliative care for the dying is rarely referred to as 'euthanasia' in the absence of the more typical practices (b) and (c). However, the term 'indirect euthanasia' is generally applied to cases in which medical interventions aimed at relieving suffering (such as administering drugs or operating on a patient to make the dying process more tolerable) cause the patient's death.

(b) Forgoing life-sustaining medical treatment (either by not initiating treatment or by terminating it at some point) where this implies that the patient dies at an earlier date than he or she would if treatment were continued and where the intention is wholly or primarily to reduce the patient's suffering. This form of euthanasia is commonly referred to as 'passive euthanasia,' irrespective of whether withholding or withdrawing life-sustaining treatment is explicitly requested by the patient.

(c) Actively hastening death to shorten suffering, either by killing the patient ('active euthanasia'), by assisting the patient to kill himself ('assisted suicide'), or by putting the patient in a prolonged state of unconsciousness (e.g., by giving him or her a strong dose of narcotics) with the intention to withdraw life-sustaining treatment while the patient is in this state ('terminal sedation').

Interest in euthanasia of these three types has grown in proportion to the growth of modern medicine and its capacities to maintain life-functions in states of terminal disease and in reaction to a widespread tendency of physicians to preserve life in situations of imminent death at the cost of severe suffering. Political impulses towards making euthanasia a patient right continue to come primarily from the euthanasia and 'right-to-die' societies that have developed all over the world (the British 'Voluntary Euthanasia Legalization Society' was founded in 1935, the 'Euthanasia Society of America' in 1938). They vary, however, in the forms of euthanasia they want to see legalized by law and professional codes. All societies, including the Hospice movement, advocate the extension and improvement of psychological and palliative care for the dying as well as passive euthanasia where this is explicitly requested by the patient. The differences between them concern the extent to which they advocate active euthanasia (as well as the related forms of assisted suicide and terminal sedation). While for the Hospice movement (Twycross 1995) it has become a matter of principle to reject active euthanasia and assisted suicide, most euthanasia societies demand the legalization both of (physician-) assisted suicide and active euthanasia under what might be called the 'paradigmatic euthanasia conditions': (a) the patient is in an irreversible state of terminal illness, (b) the patient suffers intolerably, (c) the patient explicitly wishes to die, (d) this wish is not only momentary but constant. A number of societies, however, favor assisted suicide over active euthanasia as a more unambiguous expression of the patient's autonomous will. At present, an officially approved and legal practice of active euthanasia exists in The Netherlands, an officially approved and legal practice of physician-assisted suicide in the state of Oregon, USA.

The common element of these three variants of euthanasia is that the primary and ultimate aim of giving euthanasia is the relief of suffering, either from physical pain or from other inconveniences of the patient's condition. Shortening the life of the patient is either intended as a means to that end (active euthanasia, passive euthanasia, assisted suicide) or accepted as a side effect (indirect euthanasia). This common element is lacking in practices (d) and (e).

(d) Withholding life-sustaining treatment from a patient in a persistent state of unconsciousness (as in so-called *coma vigile*) with the resulting (and mostly intended) death of the patient. In cases of this sort, the intention cannot be to relieve the patient's suffering because an unconscious patient cannot experience suffering. (There is, however, a residual unclarity on this point. A minority of neurologists think that there is a small chance that consciousness in these patients survives.) Euthanasia in these cases is generally justified either by the patient's presumed will not be kept alive in a persistent state of unconsciousness (as documented, e.g., by an advance directive), by the insistence of the family distressed by a treatment they often see as degrading and contrary to human dignity, or, in advanced stages, by considerations of medical futility (studies have shown that the probability of an 'awakening' from a so-called 'vegetative state' is about 1 percent after one year). Relevant jurisdiction in the US and the UK shows a tendency to allow the withdrawal of life-sustaining treatment after several years of 'persistent vegetative state' with the approval of the family.

(e) Ending the lives of adults and children seen as a burden to society or to parents. In German speaking countries, the German equivalent to the term 'euthanasia' inevitably evokes the National Socialist practice of killing children with malformations and the inmates of psychiatric hospitals and is normally replaced, in reference to practices (a)–(d), by *Sterbehilfe* (assistance in dying). Moreover, memories of the NS-'Euthanasie' explain at least part of the violent resistance, especially in German speaking countries, to philosophical justifications of infanticide (such as by Peter Singer, Singer 1993, Chap. 7). The motives

behind the killing, with the cooperation of their physicians, of 70,000 patients of psychiatric institutions from 1940 to 1941 were partly economic (one criterion of selection was whether patients were able to work), partly an ideology of racial purity.

After stopping the active killing in 1941 a further 90,000 patients were exterminated by 'softer' methods such as undernourishing (Benzenhöfer 1999, p. 129). Calling these killings 'euthanasia' was a deliberate euphemism, as was the term '*lebensunwert*' (unworthy of living) used by the authors Binding and Hoche in a seminal book as early as 1920 and suggesting a poor subjective quality of life (the opposite of 'a life worth living') but in fact expressing an objective judgement about the usefulness of a life to society. Ideas of a selective killing of 'useless' lives have their roots in the Social Darwinism of the late nineteenth century, though there are important earlier philosophical models. A rigid eugenic policy involving compulsory abortion and exposition of 'weak' children (probably on the Spartan model) were a central characteristic of Plato's *Republic*. A curious blending of euthanasia in the sense of relieving suffering and euthanasia in the sense of getting rid of people seen as a burden occurs in Thomas Morus' *Utopia* (1516), where priests and state officials persuade the terminally ill to kill themselves rather than to 'survive their own death.'

3. *Ethical Problems*

While euthanasia in sense (a) is almost universally accepted and euthanasia in sense (e) almost universally rejected by medical ethics and medical law, ethical and legal discussion of euthanasia in each of the other meanings is highly controversial and of a particular emotional quality. There are at least four controversial issues.

(a) How to resolve the conflict between the *principle of patient autonomy* and the *best interest principle* in euthanasia situations. According to the principle of patient autonomy, medical treatment (except in emergency situations) must be based on the informed consent of the patient. If the patient is unconscious or otherwise unable to give valid consent, physicians are required to take into account the presumed consent of the patient, as indicated by written or oral advance directives, by the views expressed by the patient about how he or she would like to be treated in relevant situations and what else can be inferred from the patient's beliefs and world-view. But the physician is also bound by the best interest principle which says that he or she should further the well-being of the patient in the best possible way. This is generally taken to entail an obligation to relieve suffering even without an explicit request. Generally, this principle is taken to be weighty enough to justify shortening the life of a patient with terminal illness if the intention of the physician is exclusively or primarily to relieve suffering.

Problems arise whenever the two principles come into conflict, as when a patient refuses an operation which, in all probability, would save his or her life at an acceptable level of subjective quality of life, whereas following his refusal would lead to instant death. There is considerable consensus among medical ethicists that a medical paternalism (i.e., treating the patient against his or her will, see Van De Veer 1986, Chap. 5) is objectionable even in cases of markedly 'unreasonable' patient wishes except when there is evidence that the patient is not sufficiently informed about his or her prospects, that his or her will is distorted by neurosis, psychosis, or excessive affect, or proves to be unstable under information. Though, on the best interest principle, the physician is called upon to persuade the patient to revise his or her decision, it is widely agreed that the physician should not resort to deception to change the patient's will, since this undermines trust and infringes patient autonomy no less than compulsory treatment would.

(b) Whether actively discontinuing treatment (e.g., by turning the switch of a respirator) is normatively equivalent to strict inaction (e.g., by 'freezing' oxygen pressure on a level that is insufficient for the patient to survive). Ethicists and jurists predominantly take both modes of action to be normatively equivalent. They see no morally significant difference between 'active' and 'passive' discontinuing nor between not starting treatment and terminating treatment. (Discontinuing treatment with the result or the intention of a patient's death in a euthanasia situation generally counts as 'passive' even if involving positive action because the state of the patient is merely restored to the condition in which it was before treatment.) There is, however, a considerable psychological difference. Physicians and nurses find it much easier to let a patient die 'passively' because causality is attributed to external factors instead of one's own behavior.

(c) Whether active killing should be allowed to doctors (or to others) in euthanasia situations. Voluntary active euthanasia, i.e., actively killing a patient to end his or her suffering on the patient's explicit wish is an attractive option because of substantial doubts whether all states of suffering in the last phase of a patient's life (e.g., in some forms of end-stage cancer) can be effectively relieved by discontinuing treatment and palliative measures. Furthermore, many of the ethical arguments against active euthanasia on the level of principles suffer from the fact that they attribute much more normative weight to the distinction between activity (killing) and passivity (letting die) than is thought acceptable, and indeed desirable, in other medical contexts where the physician has not only a duty not to do harm but also a duty to treat. In particular, the question arises how the acceptance of indirect euthanasia (which also involves active intervention) by most legislatures can be made consistent with the rejection of voluntary active euthanasia. The crucial difference is that death is intended as a *means* to

reducing suffering in the latter, while it is accepted as a *side effect* in the former. This difference is regarded as highly relevant by the doctrine of 'double effect' of Catholic moral theology, but not generally by legal systems.

More compelling arguments against legalizing active euthanasia are pragmatic considerations of potential misuse and abuse. Establishing a practice of active euthanasia risks that the autonomy of the patient is overruled by the wishes of relatives, the standards of the medical profession, and the interests of society. In The Netherlands, the only country where active euthanasia is openly practiced, along with a publicly approved practice of voluntary euthanasia (about 2,300 cases per year), there goes a secret practice of nonvoluntary euthanasia (about 1,000 cases per year) which deliberately disregards the procedural rules laid down by the Royal Dutch Medical Association. Though far from proving that a practice of voluntary euthanasia necessarily has a practice of nonvoluntary euthanasia in its wake, the opponents of the legalization of active euthanasia have a point in quoting the 'Dutch experience' (see Thomasma and Marshall 1998) as a warning. Surely, physician-assisted suicide (which is only rarely practiced in the Netherlands) is, wherever it is possible, preferable as an alternative to voluntary euthanasia, not least because it asserts the autonomy of the patient (see Quill 1993).

4. Social Research on Euthanasia

Social science studies about the practice of passive and active euthanasia have begun to be undertaken only recently, partly after the model of the Dutch Remmelink report which was done in 1990–1 at the request of the Dutch government in preparation of euthanasia legislation. This nationwide study included interviews with 405 physicians (carried out by 30 experienced physicians and five other academics) and a questionnaire study with the physicians of a sample of 7,000 recently deceased persons. The main result of this study was a surprisingly high incidence of euthanasia in The Netherlands prior to regulation. Fifty-four percent of the physicians interviewed confirmed that they had practiced active euthanasia or assisted suicide on request of patients and a further 34 percent said that they could conceive of situations in which they would be prepared to do so (van der Maas et al. 1991, p. 669). Twenty-seven percent said that at some time they had ended a patient's life without his or her explicit request and a further 32 percent said that they would prepared to do so under appropriate circumstances. Five years later the study was repeated showing that the situation in The Netherlands had changed only slightly (van der Maas et al. 1996). Substantially lower figures were found in an Australian and a British questionnaire study in which 29 percent (12 percent) of all doctors who returned a completed questionnaire confirmed that they had taken active steps to end a patient's life. A large proportion of the British respondents (46 percent) said, however, that they would consider taking active steps to bring about death if it was legal to do so (Kuhse and Singer 1988, p. 623, Ward and Tate 1994, p. 1332). The importance of cultural factors is highlighted by the contrast between figures for The Netherlands and for Germany, where active euthanasia is still largely a taboo and where questionnaire studies show that the proportion of physicians practicing active euthanasia is maximally 5 percent (Kirschner and Elkeles 1998, p. 251).

See also: Bioethics: Examples from the Life Sciences; Death and Dying, Psychology of; Death and Dying, Sociology of; Death, Anthropology of; Ethics for Biomedical Research Involving Humans: International Codes; Eugenics, History of; 'New' Eugenics, The: Role in Society

Bibliography

Bacon F 1857 Of the dignity and advancement of learning, books II–IV. In: Spedding J (eds.) *The Works of Francis Bacon*, Longman, London

Beauchamp T L (ed.) 1995 *Intending Death. The Ethics of Assisted Suicide and Euthanasia*. Prentice Hall, Upper Saddle River, NJ

Benzenhöfer U 1999 *Der gute Tod? Euthanasie und Sterbehilfe in Geschichte und Gegenwart*. Beck, Munich

Hastings Center 1987 *Guidelines on the Termination of Life-Sustaining Treatment and the Care of the Dying*. Indiana University Press, Bloomington, IN

Keown J (ed.) 1995 *Euthanasia Examined. Ethical, Clinical and Legal Perspectives*. Cambridge University Press, Cambridge, UK

Kirschner R, Elkeles T 1998 Ärztliche Handlungsmuster und Einstellungen zur Sterbehilfe in Deutschland. *Gesundheitswesen* **60**: 247–53

Kuhse H, Singer P 1988 Doctors' practices and attitudes regarding voluntary euthanasia. *Medical Journal of Australia* **148**: 623–7

Maas P J van der, Delden J van, Pijuenborg L, Looman C W N 1991 Euthanasia and other medical decisions concerning the end of life. *Lancet* **338**: 669–74

Maas P J van der et al. 1996 Euthanasia, physician-assisted suicide, and other medical practices involving the end of life in the Netherlands, 1990–1995. *New England Journal of Medicine* **335**: 1699–1705

Quill T 1993 *Death and Dignity*. W W Norton, New York

Singer P 1993 *Practical Ethics*, 2nd edn. Cambridge University Press, Cambridge, UK

Thomasma D C, Marshall P A (eds.) 1998 *Asking to Die: Inside the Dutch Debate About Euthanasia*. Kluwer, Dordrecht, The Netherlands

Twycross R G 1995 Where there is hope, there is life: A view the hospice. In: Keown J (ed.) *Euthanasia Examined. Ethical, Clinical and Legal Perspectives*. Cambridge University Press, Cambridge, UK

Van De Veer D 1986 *Paternalistic Intervention. The Moral Bounds on Benevolence*. Princeton University Press, Princeton, NJ

Ward B J, Tate P A 1994 Attitudes among NHS doctors to requests for euthanasia. *British Medical Journal* **308**: 1332–4

D. Birnbacher

Evans-Pritchard, Sir Edward E (1902–73)

Modern anthropology, born in the imagination of the educated classes of Western Europe and the USA in the late nineteenth century, owes its parentage to several ways of witnessing and recording the world beyond their immediate horizon. There were already high quality travelers tales, scientific expeditions mounted to investigate natural history (including direct observations of humanity), and political reports on the troublesome borders of empire. Early professional anthropologists experimented within this framework, and various emphases developed in specific national traditions. In the UK, by the second half of the twentieth century a distinctive combination of personal engagement, systematic research aims within the broad field of social science, and relevance to the affairs of state produced the discipline that became proud to carry the label 'British social anthropology.' Edward Evans-Pritchard was a key architect of this discipline's style and success, both intellectually and institutionally, and achieved perhaps a surprising degree of academic recognition for anthroplogy across the humanities and social sciences. The particular timing of his education and his career meant that he was able to take full advantage of the opportunities of the high colonial period in Africa for his primary research. The expansion of the British University system after the Second World War meant that he was able to promote the institutional development of his discipline. His synthesis was not, however, primarily dependent upon 'participant observation' of culture and human life as such (as had been emphasized by Malinowski) nor upon the need for defining a 'natural science of society' (which had so concerned Radcliffe-Brown). Evans-Pritchard's intellectual ambitions for anthropology sprang rather from an imaginative interest in long-term history and in philosophy, an interest traceable back to his student days and earliest publications.

Edward Evans-Pritchard was born at Crowbridge in Sussex on 21 September 1902, the son of an Anglican clergyman. He attended Winchester College, from which he entered Exeter College at Oxford University in 1921. His family background was less conventional, however, than this account might indicate, as his father was from Caernarvon and spoke Welsh, while his mother's family came from Liverpool. Nor was his time as an undergraduate at Oxford spent very conventionally; he found his obligatory studies of history, mainly about kings and battles, somewhat constricting. Exeter College was, however, already established as a focus for anthropological activities, as a result of the enthusiasms of R. R. Marett, University Reader in Social Anthropology and later Rector of the College, and Evans-Pritchard met many anthropologists there. It is extremely likely that he also attended lectures, almost certainly on Roman Britain but also possibly on philosophy, by the young R.G. Collingwood, then regarded as an unconventional and slightly 'extra-curricular' lecturer but very popular among students. Evans-Pritchard was a little later to review Collingwood's *Roman Britain* (for details see Beidelman 1974); warmly endorsing what we would call today its 'socio-cultural' approach to the everyday life and practices of the ordinary folk.

Drawn by the promise of working with experienced field anthropologists, Evans-Pritchard moved on to study at the London School of Economics and Social Science in 1924. C. G. Seligman was already teaching there, and Bronislaw Malinowski arrived the same year. C.G. Seligman, with his wife Brenda, had been engaged in anthropological work for the government of the Anglo-Egyptian Sudan since 1910, but they were beginning to seek others to continue their project, and Evans-Pritchard seized his opportunity. After completing a required survey in the southern Blue Nile region in 1926 he was allowed a free hand in embarking on his own project to study the Azande kingdoms of the equatorial forests bordering the Congo basin in the south-west. This field research, together with a series of visits between 1930 and 1936 to the very different region of the upper Nile flood-plains inhabited by the Nuer, whose reputation for being a 'difficult' and elusive people far outweighed anything else known about them, laid the foundation for Evans-Pritchard's later career and his shaping of the discipline of social anthropology. He did have the experience of teaching at King Fuad University in Cairo from 1930 to 1934, and at this time published a series of critical, theoretical papers which foreshadowed some of his mature work (for details see Beidelman 1974). A steady output of ethnographic articles mainly on the Azande and the Nuer also appeared through the 1930s, and indeed in the case of the Azande continued to appear throughout his life. But it was the two monographs appearing at the end of the decade which confirmed Evans-Pritchard's outstanding position in anthropology and enabled him, after the War, to re-launch his discipline. They were *Witchcraft, Oracles, and Magic among the Azande* (1937) and *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People* (1940), both published by the Clarendon Press in Oxford, who successfully pleaded the need to obtain a subsidy in each case from the Sudan Government.

While Evans-Pritchard's best-known ethnographic work was carried out in the Sudan, he also undertook field research in several other countries, partly as a result of the War. He was commissioned in the Sudan

Defence Force in 1940, and fought alongside locally recruited militia in Anuak country on the Ethiopian border. In 1942 he was posted to Libya, where he served in the British Military Administration following the liberation of that country from Italian occupation. His investigations of the Bedouin tribes of the interior bore fruit in his book the *Sanusi of Cyrenaica* (1949), an explicitly historical treatment of some of the ideas about segmentary social relations which he had worked out for the Nuer. By the time the War was over, Evans-Pritchard had achieved a greater ethnographic range than any of his British predecessors. There was also an innovative quality to his analytical writings. It was during the period from the late 1930s to the late1940s too that other key events in his life took place. In 1939 he married Ioma Heaton-Nicholls, daughter of the South African high commissioner in London, with whom he raised a lively family of five children. In 1944 he became a Roman Catholic, and returning to Britain at the end of the War, was appointed in 1945 to a Readership in Cambridge and then in 1946 to the Chair of Social Anthropology in Oxford, in succession to A. R. Radcliffe-Brown. He immediately established a high profile position, both for himself and for the now rapidly-expanding subject, becoming president of the Royal Anthropological Institute (1949–51), a founder (1946) and later first life president of the Association of Social Anthropologists of the Commonwealth, and a Fellow of the British Academy (1956). He retired in 1970, and his outstanding life's contribution was recognized with a knighthood in 1971. He died at his home in Oxford, on 11 September 1973. Even before his death, Evans-Pritchard had become something of a living legend, a 'household name' in anthropology. His works helped to define his subject to successive generations of students, and indeed still do so.

Evans-Pritchard's major achievement was to establish social anthropology after the War as a respected member of the humane disciplines, engaging on more equal terms than previously (at least in Britain) with the concerns of linguists, philosophers, historians, theologians, and political theorists. In achieving this distinctive character (which sustains the subject still) some distance had to be created between the 'new' social anthropology on the one hand and the 'old' museums together with 'physical' or biological anthropology on the other. Moreover, a clear identity had to be forged in respect of sociology, which in its new post-War vigor was seen as a richer cousin and tremendous rival (and indeed still is, especially in claims to public funding). Here, the emphasis lay on anthropology's qualitative appreciation of *relational structures and processes* in social life: in the principles behind the uses of language and the creation of meaning in interaction; behind the choreography of kinship and marriage, and the patterning of religious and symbolic practice; and behind the workings of authority, power, and political strife.

The Azande analysis (Evans-Pritchard 1937) immediately struck chords; for example, Collingwood referred to it in his *Principles of Art* (Collingwood 1938). After the War, the book became widely influential in the debate opening up within philosophy over the socially and historically placed foundations of what could be understood as knowledge. It also resonated with the more flexible approaches being taken by historians towards the social context of 'witchcraft' phenomena in the European and American past. Because of its pace and well-constructed storyline, the book easily moved into teaching syllabuses and has always been popular with students. An abridged edition appeared in 1976, mainly to serve teaching courses, and sales are still healthy. The Azande and their doggedly 'rational' and clever way of monitoring and justifying the operation of their oracles (still quite surprising to many readers) have helped to define what modern anthropology is about. They have certainly entered almost every textbook on the subject published since the War. Ironically, however, the appeal of the Azande has more recently converged with new and easy forms of cultural relativism which lead students to ask 'Why shouldn't the Azande believe in witchcraft, if it suits their way of life, and who am I to criticize them?' This is not Evans-Pritchard's line; on the contrary, he shows how most ordinary Azande themselves are not at all comfortable with the presence of witchcraft. It is those in powerful and privileged positions who focus on the idea, as the source of suffering and death, and turn their accusations on the relatively helpless. Sometimes-cruel kings and princes ultimately control the oracles. Even at the household level only senior men have access to them, while women constitute a large proportion of the accused. Evans-Pritchard's analysis is explicitly set in a political and historical context that offers a more nuanced reading than that of simple relativism, and ultimately, it can be argued that this extra depth is what gives the text its lasting qualities.

The Nuer, as a people, and *The Nuer* as Evans-Pritchard's first book on them, next entered the life and language of academic anthropology (Evans-Pritchard 1940). This was the first time that 'œcology' had explicitly been selected for analysis in a field monograph. It was also the first analytical field study of the political principles behind patterns of hostility and alliance between apparently unrelated groups of mobile herders, here cast in a strong idiom of moral obligation invoked through the situational rhetoric of paternal descent. Also for the first time, structured conceptions of space, and of time, were shown to have a social dimension and to constitute an overall framework of shared understandings within which political action was cast and could be 'read' by an observer. The book offered quite a different vision of 'social structure' from the empiricism of Radcliffe-Brown, and, significantly, has since been interpreted as a forerunner of the abstract 'structuralism' associated

especially with Lévi-Strauss; Louis Dumont (in his introduction to the French translation, 1975) even claimed that it had not been properly appreciated by English-speaking anthropologists. *The Nuer* nevertheless served as model and inspiration to a wide range of field studies of 'segmentary' political systems both in Africa and elsewhere. With its successors *Kinship and Marriage among the Nuer* (Evans-Pritchard 1951a, written as a companion text and explicitly focusing on kinship rather than politics) and *Nuer Religion* (Evans-Pritchard 1956) it provoked a considerable industry of 're-analysis', criticism, and comment. The secondary debates flourished in American and British anthropology for some decades until fresh field research began to alter the position of 'the Nuer' within anthropology by asking questions about them and their social world from a different standpoint than that of Evans-Pritchard's founding texts, taking into account aspects of social interconnection across the wider region of the southern Sudan and the longer-term as well as post-colonial social history of North East Africa as a whole and its on-going struggles.

The Nuer, along with the collection on *African Political Systems* co-edited with Meyer Fortes which appeared in the same year, marked the birth of the subdiscipline of field-based political anthropology and a new dialogue with the growing subject of political science. The final volume in Evans-Pritchard's trilogy devoted to the Nuer also helped create a subdiscipline, in this case the field-based anthropology of religion, and also established a new base-line for conversations with a neighboring discipline. *Nuer Religion* (1956) was written quite provocatively as a challenge to academic theology, requiring its specialists as well as a wider readership in religious studies to take tribal systems of belief and practice seriously. Evans-Pritchard's presentation of Nuer trust in *kwoth*, the Spirit of the above, has the schematic character of a creed, but at the same time a lived and shared reality. He portrays for us in almost biblical tone an elemental confrontation between God and Man, demanding recognition across the gulf between the historical religions and their sacred texts on the one hand and the worlds they often denigrate on the other as erring paganism or blind custom. The work is presented as social analysis, emphasizing how diversity of perception and representation reflects the relativity of points of view within Nuer society as established in his first book, and it draws the line (explicitly) at entering the subjectivity of religious experience. A turn towards this very subjectivity, however, has been at the heart of more recent work in the anthropology of religion, opening up avenues to the anthropological study of the major world religions in their lived reality.

Evans-Pritchard's use of ethnographic data from his field research in advancing fresh analysis was innovative and brilliant. To appreciate his 'theories' or 'methods' we need, indeed, to engage with his concrete analyses. However, Evans-Pritchard did set out in the 1950s what he clearly intended as a new horizon for social anthropology. He made explicit his life-long sympathy for the historian's perspective and his ambition to establish social anthropology as 'a kind of historiography,' a comparative study of the forms of social life which did not caricature those forms as mere empirical data but recognized their moral character and their participation in the same historical world as that of the observer (see for example Evans-Pritchard 1951b, 1962). He specifically evoked R.G. Collingwood's vision of what 'history' could be (Collingwood 1946) in sketching his own ambitions for social anthropology, and also followed Collingwood's example in drawing up a new academic lineage for his disciplinary field as he saw it.

It has sometimes been suggested that Evans-Pritchard's style of anthropology could be glossed as 'the translation of culture,' and indeed this was adopted as the title of one of the main volumes of essays offered to him on his retirement in 1970 (Beidelman 1971). However, as a gloss on his overall repositioning of anthropology this phrase sits a little uncomfortably, deriving more from late twentieth-century ideas of 'culture' difference than from the disciplined comparison of social forms towards which Evans-Pritchard strove, taking his inspiration from the French sociological school of Durkheim, and especially perhaps from the historically comparative work of Marcel Mauss. Among other initiatives in the 1950s, Evans-Pritchard sponsored a series of translations of classic essays from the *Année Sociologique*, starting with Mauss's essay on *The Gift* (1954), which remains a foundation text in the teaching of anthropology, especially economic anthropology, and is clearly about more than the 'translation of culture.' Evans-Pritchard was certainly concerned with translation, but essentially as a craftsman in handling the subtleties of language as such, as for example in his elegant paper on *sanza*, the way in which Azande employ proverbs in double-talk (1962), or in his treatments of poetry and oral literature. A project he undertook with colleagues in the 1960s was the launching of the Oxford Library of African Literature, a series that flourished over the following two decades, and had considerable influence in developing anthropology's openness to art, performance, narrative, and the uses of textual material generally.

The lasting significance of Evans-Pritchard's legacy lies partly in its successful move away from simple observer's empiricism and a narrowly-based definition of the subject-matter of anthropology, towards an engagement with key debates in history, politics, literary criticism, philosophy, theology, and religious studies (for evaluations, see, for example, Barnes 1987, Douglas 1980). At the same time, these subjects have come increasingly to recognize the value of empirical inquiry into the lives, opinions, memories, and practices of people living 'in the world,' while the feminist critique and new fields such as gender studies,

cultural, and media studies have adopted many ideas from anthropology. They have also helped to give a new lease of life to the idea of 'ethnography' as a reflexive engagement with the world.

Meanwhile, anthropology has moved firmly from the 'tribal periphery' towards the comparative study of the major civilizations, including the practice of the world religions; it is now tackling modern economic ideas and international development, urban life, and global processes. A generation of anthropologists born long after Evans-Pritchard's era have naturally sought to refashion and refresh the subject in the fast-changing world of the late twentieth century, and various lines of criticism have been directed towards his work. Along with his contemporaries, he has been taken to task for not keeping more of an intellectual distance from the colonial situation in which anthropology had such a privileged position. He has been criticized for being too impressionistic in his ethnographic portrayals, for not adopting a more differentiated view of the peoples he was studying, and not attending more carefully to the specifics of their lives. He has been accused of using a patronizing tone and perpetuating character stereotypes. On the question of religion, he has been accused of presenting Nuer beliefs as too biblical and the people as too pious. He has also had his fair share of accusations over 'male bias.' Details of his Nuer ethnography (though rarely of his accounts of the Azande) have been combed over and reanalyzed in order to support various more 'rigorous' accounts based on modern social or psychological theories. Evans-Pritchard himself rarely responded directly to these criticisms. From the mid-1960s on, in common with others whose lives had become involved in the Sudan in one way or another, he was shocked and angry about the deepening civil war in the country and its effect on people he remembered, and who remembered him. Perhaps partly because of this he had no taste for picking over his earlier writings. He developed various enthusiasms for both radical and conservative causes, and seemed to enjoy his growing reputation for eccentricity (see Lienhardt's memoir, 1974). Age added a more of a waspish tone to his famous wit and camaraderie, while the debates over his work were left to others.

With hindsight, and the turn of another century, it can be seen that Evans-Pritchard's work remains a source of inspiration despite the broadside against 'functionalism' and imperial collusion. He was 'accused' of working for the Sudan government, but at the time he was aware of the contradictions of his position and engaged in debate with the colonial authorities (Johnson 1982). He put very vividly those paradoxes of multiple belonging or 'positionality' which the modern student ponders, focusing in a still helpful way on the relativity of 'self and other' in the making of personal and group identities, as we might rephrase the relevance of segmentary theory today. He placed 'dialogue' both between the Azande themselves and between himself and his informants firmly at the heart of the analysis of knowledge and belief that he carried out among them. He also respected, explicitly, the privacy and inner religious consciousness of the Nuer whose world became in so many other ways public property through his work. It is clearly true that in many respects Evans-Pritchard failed to take 'colonialism' into account in his analyses, or to allow for historical change; but then again, looking back, he is foremost among those who worked to re-orient anthropology towards history. His treatment of field research was never as an end in itself; field research was only one of many methods, including many of those scholarly methods standard to history and the social sciences, to be drawn into anthropological analysis.

Moreover, Evans-Pritchard's anthropology was broader, in practice, than many imagine, including for example the sensitive use of informants' own accounts about relations between men and women among the Azande—long before gender or sexuality had become an academic topic (1974). Nor was he confined to work in 'social' anthropology alone; he made quite extensive collections of ethnographic objects, musical recordings, and photographs, now mostly deposited in the Pitt Rivers Museum in Oxford. In Britain, social anthropology has sought to reestablish debates with the study of material culture, art and music on the museum side and also with medical and biological anthropology. Evans-Pritchard's work remains a source of questions relevant to these new debates. His achievement of a distinctive intellectual tradition within the humanities for 'social anthropology' in the mid-twentieth century provided a momentum from which the discipline in the broader sense is still benefiting. His memory continues to be honored, most recently with the establishment of an annual series of Evans-Pritchard Lectures at All Souls College, Oxford, in 1999.

See also: Anthropology; Anthropology, History of; Functionalism, History of; Political Anthropology; Structure: Social; Witchcraft

Bibliography

Barnes J A 1987 Edward Evan Evans-Pritchard. *Proceedings of the British Academy* **73**: 447–89

Beidelman T O (ed.) 1971 *The Translation of Culture: Essays to E. E. Evans-Pritchard*. Tavistock, London

Beidelman T O (ed.) 1974 *E. E. Evans-Pritchard, a Bibliography*. Tavistock, London

Collingwood R G 1938 *The Principles of Art*. Clarendon Press, Oxford, UK

Collingwood R G 1946 *The Idea of History*. Clarendon Press, Oxford, UK

Douglas M 1980 *Evans-Pritchard*. Harvester Press, Fontana, Brighton, UK

Dumont C L 1975 Preface to the French edition of The Nuer. In: Beattie J H M, Lienhardt R G (eds.) *Studies in Social Anthropology*. Clarendon Press, Oxford, UK
Evans-Pritchard E E 1937 *Witchcraft, Oracles, and Magic among the Azande*. Clarendon Press, Oxford, UK. Abridged edition issued in 1976
Evans-Pritchard E E 1940 *The Nuer: a Description of the Modes of Livelihood and Political Institutions of a Nilotic people*. Clarendon Press, Oxford, UK
Evans-Pritchard E E 1949 *The Sanusi of Cyrenaica*. Clarendon Press, Oxford, UK
Evans-Pritchard E E 1951a *Kinship and Marriage among the Nuer*. Clarendon Press, Oxford, UK
Evans-Pritchard E E 1951b *Social Anthropology*. Cohen & West, London
Evans-Pritchard E E 1956 *Nuer Religion*. Clarendon Press, Oxford, UK
Evans-Pritchard E E 1962 *Essays in Social Anthropology*. Faber and Faber, London, Chaps. 1, 9
Evans-Pritchard E E 1974 *Man and Woman among the Azande*. Faber and Faber, London
Fortes M, Evans-Pritchard E E (eds.) 1940. *African Political Systems*. Oxford University Press, London, UK
Johnson D H 1982 Evans-Pritchard, the Nuer, and the Sudan Political Service. *African Affairs* **81**: 231–46
Lienhardt R G 1974 Evans-Pritchard: a personal view'. *Man* n.s. **9:** 301–3
Mauss M 1954 *The Gift: Forms and Functions of Exchange in Archaic Societies*. Cohew and West, London

W. James

Event History Analysis: Applications

'Event history analysis' studies a collection of units, each moving among a finite (usually small) number of states. An example is individuals who are moving from unemployment to employment. In this article, various examples of event history applications are presented, the concepts and advantages of event history analysis are demonstrated, and practical complications are addressed.

1. *Examples of Continuous-time, Discrete-state Processes*

Event history analysis is the study of processes that are characterized in the following general way: (a) there is a collection of units (which may be individuals, organizations, societies, etc.), each moving among a finite (usually small) number of states; (b) these changes (or events) may occur at any point in time (i.e., they are not restricted to predetermined points in time); and (c) there are time-constant and/or time-varying factors influencing the timing of events. Examples are workers who move between unemployment and employment; men and women who enter into consensual unions or marriages; companies that are founded or closed down; governments that break down; people who are mobile between different regions or nation states; consumers who switch from one brand to another; prisoners who are released and commit another crime; people, who show certain types of verbal and nonverbal behaviors in social interactions; students who drop out of school; incidences of racial and ethnic confrontation, protest, riot and attack; people who show signs of psychoses or neuroses; patients who switch between the states 'healthy' and 'diseased,' and so on. In event history analysis, the central characteristics of the underlying stochastic process is mirrored in the specific way theoretical and mathematical models are built, data are collected, and the estimation and evaluation of models is done.

2. *Different Observation Plans*

In event history analysis the special importance of broader research design issues has been stressed. In particular, different observation plans have been used to collect information on the continuous-time, discrete-state substantive process. These various schemes produce different types of data that constrain the statistical analysis in different ways: cross-sectional data, event count data, event sequence data, panel data, and event history data.

The cross-sectional observation design is still the most common form of data collection in many fields. A cross-sectional sample is only a 'snapshot' taken on the continuous-time, discrete-state substantive process (e.g., the jobs of people at the time of the interview). Event history analysis has shown that one must be very cautious in drawing inferences about explanatory variables on the basis of such data because, implicitly or explicitly, social researchers have to assume that the process under study is in some kind of equilibrium. Equilibrium means that the state probabilities are fairly trendless and the relationships among the variables are quite stable over time. Therefore, an equilibrium of the process requires that the inflows to and the outflows from each of the discrete states be equal over time to a large extent. This is a strong assumption that is not very often justified in the social and behavioral sciences (see Blossfeld and Rohwer 1995).

A comparatively rare type of data on changes of the process are event count data (Andersen et al. 1993). They record the number of different types of events for each unit in an interval (e.g., the number of racial and ethnic confrontations, protests, riots or attacks in a period of 10 years). Event sequence data provide even more information. They record the sequence of specific states occupied by each unit over time (e.g., see *Event Sequence Analysis: Methods*). The temporal data most often available to the scientist are panel data. They provide information on the same units of analysis at a series of discrete points in time. In other words, there is only information on the states of the units at

predetermined survey points and the course of the process between the survey points remains unknown. Panel data certainly contain more information than cross-sectional data, but involve well-known distortions created by the method itself (e.g., panel bias, attrition of sample). In addition, causal inferences based on panel approaches are much more complicated than has been generally acknowledged (Blossfeld and Rohwer 1995). With regard to the continuous-time, discrete-state substantive process, panel analysis is particularly sensitive to the length of the time intervals between the waves relative to the speed of the process. They can be too short, so that too few state transitions will be observed, or too long, so that it is difficult to establish a time-order between the events.

The event oriented observation design records the complete sequence of states occupied by each unit and the timing of changes among these states. For example, an event history of job mobility consists of more or less detailed information about each of the jobs and the exact beginning and ending dates of each job. Thus, this is the most complete information one can get on the continuous-time, discrete-state substantive process. Such event history data are often collected retrospectively with life history studies. Life history studies are normally cheaper than panel studies and have the advantage that they code the data into one framework of codes and meaning. But retrospective studies also suffer from several limitations (see Blossfeld and Rohwer 1995). In particular, data concerning motivational, attitudinal, cognitive, or affective states are difficult (or even impossible) to collect retrospectively because the respondents can hardly recall the timing of changes in these states accurately. Also the retrospective collection of behavioral data has a high potential for bias because of its strong reliance on autobiographic memory. To reduce these problems of data collection, modern panel studies (e.g., the Panel Study of Income Dynamics (PSID) in the USA, the British Household and Panel Survey (BHPS) in the UK, or the Sozio-oekonomisches Panel (SOEP) in Germany) use a mixed design which provides traditional panel data and retrospectively collects event history data for the period before the first panel wave and between the successive panel waves.

3. *Event History Analysis*

Event history analysis, originally developed as independent applications of mathematical probability theory in demography (e.g., the classical life table analysis and the product-limit estimator of Kaplan and Meier (1958), reliability engineering, and biostatistics (e.g., the path-breaking semiparametric regression model for survival data of David Cox (1972) has gained increasing importance in the social and behavioral sciences (see Tuma and Hannan 1984, Blossfeld et al. 1989, Lancaster 1990, Andersen et al. 1993, Blossfeld and Rohwer 1995) since the early 80s. Due to the development and application of these methods in various disciplines, the terminology is normally not easily accessible to the user. Central definitions are therefore given first.

3.1 *Basic Terminology*

Event history analysis studies transitions across a set of discrete states, including the length of time intervals between entry to and exit from specific states. The basic analytical framework is a state space and a time axis. The choice of the time axis or clock (e.g., age, experience, marriage duration, etc.) used in the analysis must be based on theoretical considerations and affects the statistical model. Dependent on practical and theoretical reasons, there are event history methods using a discrete (Yamaguchi 1991, Vermunt 1997) or continuous time axis (Tuma and Hannan 1984, Blossfeld et al. 1989, Courgeau and Lelièvre 1992, Blossfeld and Rohwer 1995). Discrete-time event history models are, however, special cases of continuous-time models and are therefore not further discussed here. An episode, spell, waiting time, or duration—terms that are used interchangeably in the literature—is the time span a unit of analysis (e.g., an individual) spends in a specific state. The states are discrete and usually small in number. The definition of a set of possible states, called the state space, is also dependent on substantive considerations. Thus, a careful, theoretically driven choice of the time axis and design of state space are crucial because they are often serious sources of mis-specification.

The most restricted event history model is based on a process with only a single episode and two states (an origin state j and a destination state k). An example may be the duration of first marriage until the end of the marriage, for whatever reason (separation, divorce, death). In the single episode case each unit of analysis that entered into the origin state is represented by one episode. Event history models are called multistate models, if more than one destination state exists. Models for the special case with a single origin state but two or more destination states are also called models with competing events or risks. For example, a housewife might become 'unemployed' (meaning entering into the state 'looking for work'), or start being 'full-time' or 'part-time employed.' If more than one event is possible (i.e., if there are repeated events or transitions over the observation period), the term multi-episode model is used. For example, an employment career normally consists of a series of job shifts.

3.2 *Censoring*

Observations of event histories are often censored. Censoring occurs when the information about the

duration in the origin state is incompletely recorded. An episode is fully censored on the left, if starting and ending times of a spell are located before the beginning of an observation period (e.g., before the first panel wave) and it is partially censored on the left, if the length of time a unit has already spent in the origin state is unknown. This is typically the case in panel studies, if individuals' job episodes at the time of a first panel wave is known but no further information about the history of the job is collected. Left censoring is normally a difficult problem because it is not possible to take the effects of the unknown episodes into account. It is only without problems, if the assumption of a Markov process is justified (i.e., if the transition rates do not depend on the duration in the origin state).

The usual kind of censoring, however, is right-censoring. In this case the end of the episode is not observed but the observation of the episode is terminated at an arbitrary point in time. This type of censoring typically occurs in life course studies at the time of the retrospective interviews or in panel studies at the time of the last panel wave. Because the timing of the end of the observation window is normally determined independently from the substantive process under study, this type of right censoring is unproblematic and can easily be handled with event history methods (Kalbfleisch and Prentice 1980—type I censoring). Finally, episodes might be completely censored on the right. In other words, entry into and exit from the duration occurs after the observation period. This type of censoring normally happens in retrospective life history studies in which individuals of various birth cohorts can only be observed over very different spans of life. To avoid sample selection bias, such models have to take into account variables controlling for the selection, for example, by including birth cohort dummy variables (see Yamaguchi 1991).

3.3 The Transition Rate

The central concept of event history analysis is the transition rate. Because of the various origins of event history analysis in the different disciplines, the transition rate is also called the hazard rate, intensity rate, failure rate, transition intensity, risk function, or mortality rate:

$$r(t) = \lim_{t' \to t} \Pr(t \leqslant T < t' | T \geqslant t)/(t'-t) \qquad (1)$$

The transition rate provides a local, time-related description of how the process evolves over time. It can be interpreted as the propensity (or intensity) to change from an origin state to a destination state, at time t. But one should note that this propensity is defined in relation to a risk set ($T \geqslant t$) at t, i.e., the set of units that still can experience the event because they have not yet had the event before t.

3.4 Statistical Models

The central idea in event history analysis is to make the transition rate, which describes a process evolving in time, dependent on time (t) and on a set of covariates, x:

$$r(t) = g(t, x) \qquad (2)$$

The causal interpretation (see *Causation (Theories and Models): Conceptions in the Social Sciences*) of the transition rate requires that we take the temporal order in which the processes evolve very seriously. In other words, at any given point in time, t, the transition rate $r(t)$ can be made dependent on conditions that happened to occur in the past (i.e., before t), but not on what is the case at t or in the future after t.

There are several possibilities to specify the functional relationship g (see Blossfeld and Rohwer 1995). For example, the exponential model, which normally serves as a baseline model, assumes that the transition rate can vary with different constellations of covariates x, but that the rates are time-constant. The piecewise constant exponential model allows the transition rate to vary across fixed time periods with period-specific effects of covariates. There are also different parametric models which are based on specific shapes of the time-dependence (e.g., the Gompertz-Makeham, Weibull, Sickle, Log-logistic, Log-normal, or Gamma model). If the time-shape of the base-line hazard rate is unspecified, and only possible effects of covariates x are estimated, the model is called a semi-parametric or partial likelihood model (Cox 1972). Finally, there are also models of unobserved heterogeneity in which the transition rate can be made dependent on the observed covariates x, the duration t, and a stochastic error term.

4. Parallel and Interdependent Processes

The most important scientific progress permitted by event history analysis is based on the opportunity to include explicitly measured time-varying covariates in transition rate models (see Blossfeld and Rohwer 1995). These covariates can change their values over process time in an analysis. Time-varying covariates can be qualitative or quantitative, and may stay constant for finite periods of time or change continuously. From a substantive point of view, time-varying covariates can be conceptualized as observations of the sample path of parallel processes. These processes can operate at different levels. In sociology, for example, (a) there can be parallel processes at the level of the individual's different domains of life (e.g., one may ask how upward and downward moves in an individual's job career influence his/her family trajectory); (b) there may be parallel processes at the level of some few individuals interacting with each other

(e.g., one might study the effect of the career of the husband on his wife's labor force participation); (c) there may be parallel processes at the intermediate level (e.g., one can analyze how organizational growth influences career advancement or changing household structure determines women's labor force participation); (d) there may be parallel processes at the macro level (e.g., one may be interested in the effect of changes in the business cycle on family formation or career advancement); and (e) there may be any combination of such processes of type (a) to (d). For example, in the study of life-course, cohort and period effects, time-dependent covariates at different levels (see below) must be included simultaneously (multi-level analysis).

In dealing with such systems of parallel processes, the issue of reverse causation (see *Causation (Theories and Models): Conceptions in the Social Sciences*) is often addressed in the methodological literature (see, e.g., Kalbfleisch and Prentice 1980, Tuma and Hannan 1984, Blossfeld et al. 1989, Yamaguchi 1991, Courgeau and Lelièvre 1992). Reverse causation refers to the (direct or indirect) effect of the dependent process on the independent covariate process(es). Reverse causation is often seen as a problem because the effect of a time-dependent covariate on the transition rate is confounded with a feedback effect of the dependent process on the values of the time-dependent covariate. However, Blossfeld and Rohwer (1995) have developed a causal approach to the analysis of interdependent processes that also works in the case of interdependence, too. For example, if two interdependent processes, Y_t^A and Y_t^B, are given, a change in Y_t^A at any (specific) point in time t' may be modeled as being depend on the history of both processes up to, but not including t'. Or stated in another way: what happens with Y_t^A at any point in time t' is conditionally independent of what happens with Y_t^B at t', conditional on the history of the joint process $Y_t = (Y_t^A, Y_t^B)$ up to, but not including, t ('principle of conditional independence'). Of course, the same reasoning can be applied if one focuses on Y_t^B instead of Y_t^A as the 'dependent variable.' Beginning with a transition rate model for the joint process, $Y_t = (Y_t^A, Y_t^B)$, and assuming the principle of conditional independence, the likelihood for this model can then be factorized into a product of the likelihoods for two separate models: a transition rate model for Y_t^A which is dependent on Y_t^B as a time-dependent covariate, and a transition rate model for Y_t^B, which is dependent on Y_t^A as a time-dependent covariate. From a technical point of view, there is therefore no need to distinguish between defined, ancillary, and internal covariates (see, e.g., Kalbfleisch and Prentice 1980) because all of these time-varying covariate types can be treated in the estimation procedure. Estimating the effects of time-varying processes on the transition rate can easily be achieved by applying the method of episode splitting (see Blossfeld and Rohwer 1995).

5. *Unobserved Heterogeneity*

Unfortunately, researchers are not always able to include all the important covariates into an event history analysis. These unobserved differences between subpopulations can lead to apparent time dependence at the population level and additional identification problems. There exists a fairly large literature on mis-specified models in general. In particular, there have been several proposals to deal with unobserved heterogeneity in transition rate models (so-called frailty models) (see, e.g., Blossfeld and Rohwer 1995).

6. *Illustrative Applications*

The following three applications illustrate the technical strength of event history models in different substantive contexts.

6.1 *A Multi-level Analysis of Life-course, Cohort, and Period Effects in Career Mobility*

Any standard statistical analysis using age, year of birth (as a cohort measure), and chronological year of measurement (as a period measure) leads to an identification problem (age = year of measurement–year of birth). The technical solution to that problem often is to assume that at least one of these measures is constant over time or to apply arbitrary statistical constraints. Blossfeld (1986) in an event history study of career mobility of German men approached this problem from a more substantive point of view. He suggested to rely more on explicit measures for the theoretically important macro effects of cohort and period and used 14 time series from official statistics indicating the long-term development of the labor market structure in Germany. Based on factor analysis, he first reduced the 14 highly correlated time series to two orthogonal macro factors which represented the 'monotonic trend in the level of modernization' and the 'cyclic changes in labor market conditions' and then matched these macro data to individual event career data to conduct a multi-level analysis. To represent the changing conditions under which cohorts enter the labor market, Blossfeld used the factor scores of these two factors at the year men entered the labor force. To introduce period as a time-dependent covariate, he used the method of episode splitting. In accordance with this method, the original job episodes were split into sub-episodes every year so that the factor scores of the two macro factors could be updated in each job episode for all employed men every year. Modeling cohort and period effects this way not only provides more direct and valid measures of the explanatory concepts, but also results in the

disappearance of the problem of non-estimable parameters above. It was, therefore, not necessary to impose (in substantive terms normally unjustifiable) constraints on the parameters to identify life-course, cohort, and period effects.

6.2 *Parallel Processes at the Level of the Individual: Women's Rising Educational and Career Investments, and Changes in the Process of Family Formation*

Using event history models, Blossfeld and Huinink assessed the hypothesis of the 'new home economics' that women's rising educational and career investments will lead to a decline in marriage and motherhood (see Blossfeld and Rohwer 1995, p. 143). Because the accumulation of educational and career resources is a lifetime process, they analyzed it as a dynamic process over the life course. To model the changes in the accumulation of qualifications over the life course, Blossfeld and Huinink updated a time-varying covariate 'level of education' at the age when a woman obtained a particular educational rank in the hierarchy. They also generated a time-dependent dummy variable indicating whether or not a woman is attending the educational system at a specific age. Finally, for the dynamic modeling of job-specific investments in human capital over the life course, Blossfeld and Huinink have developed a specific time-dependent labor force experience measure, which they called 'career resources.' The time-dependent increments and final levels of this measure were designed to reflect the 'quality of each job' and the respective 'employment durations.' The time-dependent measure also took into account that career resources are declining when women interrupted their employment career. Blossfeld and Huinink applied the method of episode splitting to include these various time-dependent dimensions of education and career together with additional control variables in the rate equation. The results of their longitudinal analysis were surprising: contrary to the expectations of the economic approach to the family, women's entry into marriage obviously was independent of women's level of education and their career resources. Equally surprisingly, the effect of women's level of education on the rate of their first birth was even positive. An explanation for the positive effect on the timing of first birth is easily demonstrated with an event history model: women who remain in school longer and attain high qualifications are first shifting their fertility decisions to higher ages, and then, they are subject to pressure not only from the potential increase in medical problems connected with having children late, but also from societal age norms. Thus the higher the educational attainment level the stronger the catching-up effect. This application demonstrates that the standard empirical support for the economic theory of the family claimed on the basis of cross-sectional data might be misleading. Cross-sectional data do not permit a differentiation between the time-dependent effect of accumulation of human capital over the life course and the time-dependent effect of participation in the educational system.

6.3 *A Causal Approach to Interdependent Systems: The Effects of Pregnancy on Entry into Marriage for Couples Living in Consensual Unions*

Finally, the application of a causal approach to interdependent systems is demonstrated by Blossfeld and Rohwer in an analysis of the effect of pregnancy on entry into marriage in Germany (see Blossfeld and Rohwer 1995, p. 157). This study shows that the effect of a fertility event on entry into marriage is highly time dependent: It starts to rise at some time shortly after conception (thus there is some time lag involved between cause and effect), increases during pregnancy to a maximum at around seven to eight months after conception, and then decreases again reaching a level even below the starting level at around six months after the birth of the child. Thus, the dynamic interdependence of fertility and nuptiality mainly occurs in a very specific phase of the life course and the unfolding effect of fertility on nuptiality is highly dynamic over time. Of course, standard statistical applications based on cross-sectional observation can be hardly used to analyse such situations in a meaningful way. However, in event history analysis, there is an easy way to approach that problem. One can use a series of time-dependent dummy variables each indicating a month since the occurrence of conception (the causal event). They show that the effects of the time-dependent dummy variable at first increase, reach a maximum, and then decrease again. Thus, starting with conception, there is an increasing normative pressure to avoid an illegitimate birth that increases the marriage rate, particularly for people who are already 'ready for marriage.' But with an increasing number of marriages, the composition of the group of still unmarried couples shifts towards couples being 'less ready for marriage' or being 'not ready for marriage,' which, of course, decreases the fertility effect again. This is an important result not only in substantive terms; the methodological lesson for event history analysis is also revealing. One should be very careful in modeling qualitative parallel processes with only one dummy variable (e.g., in the current application using only the time of conception, or the time of birth for the switch of the time-dependent covariate) because the effect itself might be highly time-dependent. For example, Blossfeld and Rohwer show that the fertility effect disappears completely, when the time-dependent covariate is switched at the time of birth because then a period with a low marriage rate up to the time of discovery of conception and a period with a high marriage rate during pregnancy is aggre-

gated and compared with a period of low rate after the birth of the child.

See also: Event-history Analysis in Continuous Time; Event-history Analysis: Selectivity; Longitudinal Data: Event–History Analysis in Discrete Time

Bibliography

Andersen P K, Borgan Ø, Gill R D, Keiding N 1993 *Statistical Models Based on Counting Processes*. Springer-Verlag, New York

Blossfeld H-P 1986 Career opportunities in the Federal Republic of Germany: a dynamic approach to the study of life-course, cohort, and period effects. *European Sociological Review* **2**: 208–25

Blossfeld H-P, Hamerle A, Mayer K U 1989 *Event History Analysis*. Erlbaum, Hillsdale, NJ

Blossfeld H-P, Rohwer G 1995 *Techniques of Event History Modeling. New Approaches to Causal Analysis*. Erlbaum, Mahwah, NJ

Courgeau D, Lelièvre E 1992 *Event History Analysis in Demography*. Clarendon, Oxford

Cox D R 1972 Regression models and life-tables. *Journal of the Royal Statistics Society* **34**: 187–220

Kalbfleisch J D, Prentice R L 1980 *The Statistical Analysis of Failure Time Data*. Wiley, New York

Kaplan E L, Meier P 1958 Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**: 457–81

Lancaster T 1990 *The Econometric Analysis of Transition Data*. Cambridge University Press, Cambridge, UK

Tuma N B, Hannan M T 1984 *Social Dynamics. Models and Methods*. Academic Press, Orlando, FL

Vermunt J K 1997 *Log-linear Models for Event Histories*. Thousands Oakes, Newbury Park, CA

Yamaguchi K 1991 *Event History Analysis*. Sage, Newbury Park, CA

H.-P. Blossfeld

Event-history Analysis in Continuous Time

1. Introduction

Event history analysis deals with data obtained by observing individuals over time focusing on events occurring for the individuals. Thus, typical outcome data consist on *times of occurrence* of events and on the *types of events* which occurred. Frequently, an event may be considered as a *transition* from one state to another and, therefore, *multistate models* will often provide a relevant modeling framework for event history data. Multistate models are discussed from several points of view in the books by Andersen et al. (1993), Blossfeld and Rohwer (1995), and Courgeau and Lelièvre (1992); see Hougaard (1999) and Commenges (1999) for recent surveys.

There are two broad purposes of event history modeling: *analysis* where the interest is in the statistical modeling of each individual transition, including its possible dependence on internal or external covariates, and *synthesis* where the combination of several transitions is described via suitable summary measures.

2. Survival Data

The simplest multistate model is the two-state model for survival data with one transient state '0: alive' and one absorbing state '1: dead,' see Fig. 1. In general, an *absorbing* state is a state from which further transitions cannot occur while a *transient* state is a state which is not absorbing. The observation for a given individual will here in the simplest form consist of a random variable, say T, representing the time from a given origin (time 0) to the occurrence of the event 'death.' The distribution of T may be characterized by the probability distribution function $F(t) = \text{Prob}(T \leqslant t)$ or, equivalently, by the survival distribution function $S(t) = 1 - F(t) = \text{Prob}(T > t)$. It is seen that $S(t)$ and $F(t)$, respectively, correspond to the probabilities of being in state 0 or 1 at time t. If every individual is assumed to be in state 0 at time 0 then $F(t)$ is also the *transition probability* from state 0 to state 1 for the time interval from 0 to t. In continuous time the distribution of T may also be characterized by the *hazard rate function*

$$\alpha(t) = -\mathrm{d}\log S(t)/\mathrm{d}t = \lim_{\Delta t \to 0} \frac{\text{Prob}(T \leqslant t + \Delta t | T \geqslant t)}{\Delta t},$$

that is

$$S(t) = \exp\left(-\int_0^t \alpha(u)\mathrm{d}u \right)$$

Thus, $\alpha(\cdot)$ is the *transition intensity* from state 0 to state 1, that is, the instantaneous probability per time unit of going from state 0 to state 1.

In general, event history analysis deals with inference for transition intensities and transition probabilities in multistate models. This includes estimation and hypothesis tests for these quantities and analysis of regression models where these quantities are related to (possibly time-dependent) explanatory variables observed for the individuals under study. Most frequently, multistate models are defined by their transition intensities from which transition probabilities may or may not be derived depending on the modeling assumptions.

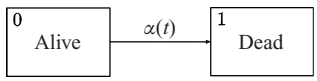

Figure 1
The two-state model for survival data

A typical feature of event history analysis is the inability to observe complete event histories, for example, by the end of the observation period all individuals under study may not have reached an absorbing state. In survival analysis this would correspond to individuals still being alive by the end of the study and this kind of *incomplete observation* is known as *right-censoring*. Furthermore, all individuals may not have been observed from the same time origin. This kind of incomplete observation where individuals are only observed conditionally on not having reached an absorbing state by the time of initiation of the study is known as *left-truncation*. Restricting attention to right-censoring a crucial problem is whether the available incomplete data enables one to make valid inference on parameters in the multistate model for the complete data. The condition for this is known as *independent* right-censoring and the interpretation is that a sample observed after independent right-censoring is 'representative' for the population without censoring. This means that individuals who are censored should have neither lower nor higher risk of future events than individuals who are not censored (see, e.g., Andersen et al. 1993, Chap. III, Kalbfleisch and Prentice 1980, p. 120).

Right-censoring may be caused by a number of different mechanisms. Suppose, for example, that one is interested in the disease specific mortality for patients suffering from a given disease like skin cancer and followed from their times of diagnosis (time 0). Then right-censoring may be caused by:

(a) a patient still being alive by the time of closure of the study,

(b) a patient being lost to follow-up,

(c) a patient dying from causes other than skin cancer.

In case (a) there are usually no reasons to believe that censoring should not be independent since the time of closure of a study can most often be assumed independent of the process observed. In case (b) one may be suspicious that patients lost to follow-up could be individuals who generally feel well (or the opposite) and, therefore, do not want to take part in a study any more and this could lead to *dependent* censoring. Situation (c) is more tricky since here the population without censoring is quite hypothetical: it is hardly reasonable to think of a population of skin cancer patients for whom it is impossible to die from other causes. This fact has consequences for how to choose an appropriate model in this situation and a relevant model to consider could be a *competing risks model*, see Fig. 2.

The problem is that even though it may be possible to make valid inference for the single cause-specific hazard $\alpha_1(t)$ of dying from skin cancer by using methods for survival data and treating deaths from other causes as right-censored observations then, in order to evaluate the survival probability for skin-cancer patients, the intensity of dying from other causes is also needed. In Sect. 3 we return to the relationship between the two cause-specific hazards and the transition probabilities in Fig. 2. The conclusion is that a *partial model specification* where the hazard, $\alpha_2(t)$, of dying from other causes is not modeled may suffice for some, but not all, purposes and a multistate model like the one depicted in Fig. 2 can be a sensible model to study in spite of the fact that focus may be on a single transition. Partial model specification is further discussed in Sect. 5.

3. *Multistate Models*

In this section, we will present a number of different multistate models including the two-state model for survival data and the competing risks model introduced in Sect. 2. We will begin with a few definitions. Some of the definitions are formulated loosely but they may all be made rigorous within the framework of so-called *marked point processes* (Andersen et al. 1993, Chap. III, Arjas and Haara 1984).

A multistate process is a stochastic process $(X(t), t \in \mathscr{T})$ with a finite state space $\mathscr{S} = \{1, \ldots, p\}$ and with right-continuous sample paths: $X(t+) = X(t)$. Here, $\mathscr{T} = [0, \tau]$ or $[0, \tau)$ with $\tau \leqslant +\infty$. The process has *initial distribution* $\pi_h(0) = \text{Prob}(X(0) = h)$, $h \in \mathscr{S}$. A multistate process $X(\cdot)$ generates a *history* $\mathscr{X}_t$ consisting of the observation of the process in the interval $[0, t]$. Relative to this history we may define *transition probabilities* by

$$P_{hj}(s, t) = \text{Prob}(X(t) = j | X(s) = h, \mathscr{X}_{s^-})$$

for $h, j \in \mathscr{S}$, $s, t \in \mathscr{T}$, $s \leqslant t$, and *transition intensities* by the derivatives

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} P_{hj}(t, t + \Delta t)/\Delta t$$

which we shall assume exist. Some transition intensities may be 0 for all t. Graphically, multistate models may be illustrated using diagrams like Fig. 1 or 2 with boxes representing the states and with arrows between the states representing the possible transitions, that is, the nonzero transition intensities (Hoem 1976, Andersen et al. 1993, Sect. I.3). A state $h \in \mathscr{S}$ is *absorbing* if for all $t \in \mathscr{T}$, $j \in \mathscr{S}$, $j \neq h$, $\alpha_{hj}(t) = 0$; otherwise h is *transient*. The *state probabilities* $\pi_h(t) = \text{Prob}(X(t) = h)$ are given by

$$\pi_h(t) = \sum_{j \in \mathscr{S}} \pi_j(0) P_{jh}(0, t)$$

Notice that both the $P_{hj}(\cdot, \cdot)$, the $\pi_h(\cdot)$, and the $\alpha_{hj}(\cdot)$ depend on both the probability measure Prob and on the history. Sometimes one is interested in considering an extended history which also includes observed *covariates*. If only time-fixed covariates Z are studied then the observed history is $\mathscr{F}_t = \mathscr{X}_t \vee \mathscr{Z}_0$ whereas

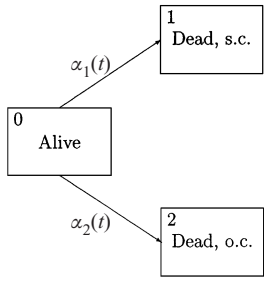

Figure 2
Competing risks model for skin cancer (s.c.) mortality (o.c. ~ other causes)

time-dependent covariates $Z(t)$ give rise to an extended history of the form $\mathscr{F}_t = \mathscr{X}_t \vee \mathscr{Z}_t$ where in both cases $\mathscr{Z}_t$ is the history generated by the covariates in $[0, t]$. We shall here focus on the *purely endogeneous* case where $\mathscr{Z}_t \subset \mathscr{X}_t \vee \mathscr{Z}_0$, i.e. the covariates are either all time-fixed or the random development of the time-dependent covariates is fully specified by the history of the process itself (see, however, Sect. 5).

3.1 The Two-state Model for Survival Data

This model, illustrated in Fig. 1, has $p = 2$ states and only one possible transition from state 0 to state 1. The corresponding transition intensity $\alpha_{01}(t)$ is given by the hazard rate function $\alpha(t)$, while $\alpha_{10}(t) = 0$ for all t, that is, state 1 is absorbing. The initial distribution is degenerate in 0: $\pi_0(0) = 1$. The process is *Markovian*, cf. Sect. 5.1. Covariates may be entered into the model using a regression model for $\alpha(\cdot)$.

3.2 The Competing Risks Model

This model has one transient state '0: alive' and a number, k, of absorbing states, state h, $h = 1,\ldots, k$ corresponding to 'death from cause h.' Thus, there are $p = k+1$ states. The model is illustrated for $k = 2$ in Fig. 2. The transition intensities $\alpha_{0h}(t)$ for $h = 1,\ldots,k$ are given by the cause-specific hazard functions

$$\alpha_h(t) = \lim_{\Delta t \to 0} \text{Prob}(\textit{Dead from cause } h \textit{ by } t+\Delta t | T \geqslant t)/\Delta t$$

where T is the survival time. The initial distribution is degenerate in 0, the only transient state of the model, that is, $\alpha_{hj}(t) = 0$ for all $h \neq 0$ and all j. The transition probabilities are given by

$$P_{00}(0, t) = S(t) = \text{Prob}(T > t) = \exp\left(-\int_0^t \sum_{h=1}^{k} \alpha_h(u)\mathrm{d}u \right)$$

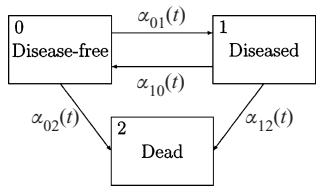

Figure 3
The illness–death or disability model

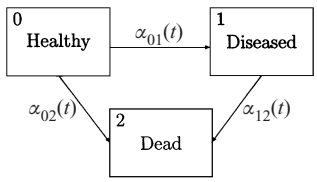

Figure 4
Unidirectional illness–death model

and

$$P_{0h}(0, t) = \int_0^t S(u-)\alpha_h(u)\mathrm{d}u,\ h = 1,\ldots, k$$

like the simple two-state model ($k = 1$) the competing risks model is Markovian and covariates may be included into the model via regression models for the cause-specific hazards.

3.3 The Illness–Death model

This model is illustrated in Fig. 3, which by now should be almost self-explanatory. Often the time t is the age of the individual, and usually individuals will be assumed to be in state 0 at $t = 0$. However, individuals will not always be observed from $t = 0$ as shall be further discussed in Sect. 5 below regarding observational patterns. The mortality $\alpha_{12}(t)$ of the diseased (the *lethality*) may sometimes depend on duration d since entry to state 1 in addition to the dependence on 'age' t. (Notice that, despite the notation, $\alpha_{12}(t)$ then depends on the *random* time of the most recent transition into 1.) If $\alpha_{12}(t)$ does not depend on d the process is Markovian, otherwise it is a *semi-Markov process*, which is clearly purely endogeneous.

In Fig. 3 we have indicated the possibility of *reversibility*: the transition back from state 1 to 0 is possible. It will turn out that the simple unidirectional model in Fig. 4, is rather easier to analyse statistically. Thus, the transition probabilities in this model have simple explicit expressions

$$P_{00}(s, t) = \exp\left(-\int_s^t (\alpha_{02}(u) + \alpha_{01}(u))\mathrm{d}u \right)$$

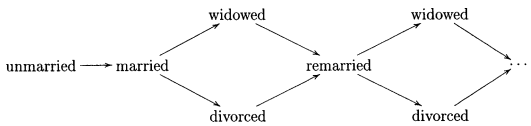

Figure 5
A model for marriage and dissolution

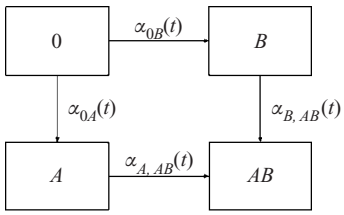

Figure 6
Interaction between life history events

and (in the Markovian case)

$$P_{01}(s, t) = \int_s^t P_{00}(s, u-)\alpha_{01}(u)P_{11}(u, t)\mathrm{d}u$$

where

$$P_{11}(s, t) = \exp\left(-\int_s^t \alpha_{12}(u)\mathrm{d}u \right)$$

For this reason, and because of the desire to have a more detailed description, the reversible illness–death model is sometimes replaced by a model with one state for each new 'disease.' An example from nuptiality is indicated by Fig. 5, where the possible transition from each state to death is omitted from the figure. (Only finitely many states are assumed!)

3.4 Interaction Between Life-history Events

This Markov model, illustrated in Fig. 6 describes the joint behavior of two life events A and B; if $\alpha_{0B} = \alpha_{A,AB}$ but $\alpha_{0A} \neq \alpha_{B,AB}$, A is called *locally dependent* on B but B is not locally dependent on A. The temporal order of events allows for this *asymmetric* concept of dependence which yields more information for drawing causal inference than the standard symmetric association concepts from cross-sectional studies. Similar duration dependence as in the illness–death process might be added. A model of this type was discussed by Aalen et al. (1980) for a study of interaction between menopause and a certain chronic skin disease (see also Blossfeld and Rohwer 1995 and Courgeau and Lelièvre (1992).

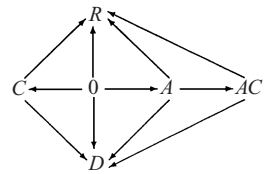

Figure 7
A model for events following bone marrow transplantation: acute and chronic graft-vs.-host disease, relapse and death

3.5 Bone Marrow Transplantation

A model combining most of the above features has been studied in detail (e.g., Klein et al. 1993, Keiding et al. 2000) as describing some of the possible states of a leukemia patient following bone marrow transplantation, see Fig. 7.

Patients have been given various kinds of therapy to temporarily keep the disease down, they are said to be in *remission*. In our context these patients are followed since bone marrow transplantation ($t = 0$), initially considered in state 0. Two different types of complications are considered: acute graft-vs.-host disease (A), chronic graft-vs.-host disease (C), and a special state AC is defined for those patients acquiring both A and C. Patients are followed until relapse of the leukemia (R) or death (D) while still in remission. Relapsed patients are not followed further in this context. If all transition rates depend only on time t since transplantation, we have again a Markov process, but various kinds of duration dependence (semi-Markov process models) may also be relevant.

4. Counting Process Representation, Likelihood

Assume that multistate processes $X_i(t)$ such as described in Sect. 3 are observed over intervals $[0, \tau_i]$ for individuals $i = 1,\ldots, n$. In this section, we will assume that τ_i is a fixed (i.e., nonrandom) time of termination of observation for individual i. Random right-censoring (Sect. 2) is treated in Sect. 6. Since $X_i(t)$ is constant between transitions, it is equivalent to record $X_i(0)$ and the *counting processes*

$$N_{hj}^i(t) = \#(\text{direct transitions } h \to j \text{ in } [0, t] \text{ for } i)$$

described by the times T_{hj}^{ik} of these transitions, where

$$0 < T_{hj}^{i1} < \ldots < T_{hj}^{iN_{hj}^i(\tau_i)} \leqslant \tau_i$$

Let $N_{hj}(t) = \sum_{i=1}^n N_{hj}^i(t)$.

It will turn out to be useful to also introduce $Y_h^i(t) = I\{X_i(t-) = h\}$ and $Y_h(t) = \#$ (individuals 'at risk' in state h at time $t-$) $= \sum_{i=1}^n Y_h^i(t)$. Note that, for

$t > \tau_i$, $N^i_{hj}(t) = N^i_{hj}(\tau_i)$ and $Y^i_h(t) = 0$; these can thus be considered to be defined on $(0, \infty)$.

For individual i denote the initial distribution $(\pi^i_h(0))$, the density of time-fixed covariates $f(Z_i)$, and the transition intensities $\alpha^i_{hj}(t)$, then the *likelihood* is

$$\prod_{i=1}^{n} f(Z_i)\pi_{X_i(0)} \prod_{h \neq j} \prod_{k=1}^{N^i_{hj}(\tau_i)} \alpha^i_{hj}(T^{ik}_{hj}) \exp\left(-\int_0^{\tau_i} \alpha^i_{hj}(t) Y^i_h(t) \mathrm{d}t \right).$$

It is very common to condition on Z_i and on the initial values $X_i(0)$ (the distribution of which may often be degenerate anyway), and consequently omit the factors $f(Z_i)\pi_{X_i(0)}$ from the likelihood. We shall do so without further comment in the sequel.

Recall from above that the notation $\alpha^i_{hj}(t)$ represents possible dependence of the transition intensity on the whole history $\mathscr{X}^i_t$ of the process. Thus, $\alpha^i_{hj}(t)$ may well contain covariates and other random elements, as already exemplified.

5. *Statistical Model Specification*

As indicated in the introduction the first purpose of event history analysis is to gain insight into the dynamics of the processes by quantifying transition intensities and perhaps assessing their dependence on covariates, possibly using various stratifications. Sometimes additional functionals are useful, particularly various types of *transition probabilities* obtained by integrating certain functions of the transition intensities. A final purpose may be *prediction*, both as illustration of the dynamics and for concrete practical purposes.

5.1 *Markov Processes*

The most important class of models is the (continuous time) *Markov process* $X(t)$ on a finite state space $\mathscr{S} = \{1, \ldots, p\}$ where the dependence of $\alpha^i_{hj}(t)$ on the history $\mathscr{X}_t$ introduced at the beginning of Sect. 3 is only via the current state of $X(t)$ (and possibly via time-fixed covariates). Statistical models usually are obtained by specifying the class of transition intensities $(\alpha^i_{hj}(t))$ for each individual i.

5.1.1 Constant and piecewise constant transition intensities. The simplest class of models is obtained by keeping the transition rates *constant*: $\alpha^i_{hj}(t) = \alpha^i_{hj}$. *Piecewise constant* intensities

$$\alpha^i_{hj}(t) = \alpha^{i(k)}_{hj}, \quad t_{k-1} < t \leqslant t_k, \quad t_0 = 0$$

form the next step up, and both of these choices are of widespread use, particularly in large studies in econometrics, epidemiology, sociology, and demography.

For ease of notation focus on the case with constant intensities and assume first that all individuals have the same transition intensities, $\alpha^i_{hj} = \alpha_{hj}$. The likelihood then simplifies to

$$\prod_{h \neq j} \alpha_{hj}^{N_{hj}(\tau)} e^{-\alpha_{hj} S_h}$$

where $\tau = \max_i \tau_i$ and

$$S_h = \sum_i \int_0^{\tau_i} Y^i_h(t) \mathrm{d}t$$

Since this likelihood looks like one resulting from observing $N_{hj}(\tau)$ events in a Poisson process with intensity α_{hj} observed over the interval $(0, S_h)$, except that here S_h is random, this likelihood, and even the model, are often associated with Poisson's name. Maximum likelihood estimation is elementary, yielding

$$\hat{\alpha}_{hj} = N_{hj}(\tau)/S_h$$

the classical *occurrence/exposure rate*. Asymptotic inference may be obtained from the information

$$-D^2 \log L = N_{hj}(\tau)/\alpha^2_{hj}$$

yielding a variance estimate

$$\mathrm{Var}(\hat{\alpha}_{hj}) \sim \alpha^2_{hj}/N_{hj}(\tau) \sim N_{hj}(\tau)/S^2_h$$

Inclusion of time-fixed covariates conveniently is achieved log-linearly

$$\alpha^i_{hj} = \alpha_{0hj} e^{\beta'_{hj} Z_i}$$

The Poisson likelihood easily accommodates this in what is often termed Poisson regression models, which may be treated using standard generalized linear model software.

Transition probabilities for the constant and piecewise constant Markov process models are explicit functions of the transition intensities (e.g., Chiang 1968), allowing direct 'plug-in' maximum likelihood estimation, as well as calculation of standard error estimates via the delta method.

5.1.2 Other parametric models for the transition intensities. Although the piecewise constant model is often sufficient to describe the dependence of intensities on time, other possibilities exist. Certain mathematical functions of time may generate the

model, such as the Gompertz–Makeham model for mortality

$$\alpha(t) = \alpha + \beta\gamma^t$$

but except for mortality studies in actuarial and some demographic contexts such parametric models are used rather little. One reason for this may be the powerful development of methodology for 'nonparametric' statistical inference, where $\alpha_{hj}(t)$ is left unspecified.

5.1.3 Freely varying ('nonparametric') transition intensities. Assume first that the transition intensities are the same for all individuals but that they are allowed to vary freely with time: $\alpha^i_{hj}(t) = \alpha_{hj}(t)$. Statistical inference is then conveniently phrased in terms of the counting process approach pioneered by Aalen (1976, 1978) (see Andersen et al. 1993 for a detailed exposition). Estimators (which may be given a nonparametric maximum likelihood interpretation) of the integrated intensities

$$A_{hj}(t) = \int_0^t \alpha_{hj}(u)\mathrm{d}u$$

are obtained as the *Nelson–Aalen estimators*

$$\hat{A}_{hj}(t) = \int_0^t \frac{J_h(u)}{Y_h(u)} dN_{hj}(u) = \sum_i \sum_{k:0 < T^{ik}_{hj} < t} \frac{1}{Y_h(T^{ik}_{hj})}$$

$J_h(u) = I\{Y_h(u) > 0\}$, with variance estimators

$$\hat{\sigma}^2(\hat{A}_{hj}(t)) = \int_0^t \frac{J_h(u)}{Y_h(u)^2} dN_{hj}(u) = \sum_i \sum_{k:0 < T^{ik}_{hj} < t} \frac{1}{Y_h(T^{ik}_{hj})^2}$$

An elaborate mathematical theory based on stochastic integrals and martingales is available to study exact and asymptotic properties of these estimators.

When estimates are desired of the transition intensities $\alpha_{hj}(t)$ themselves rather than their integrals, *smoothing* techniques are necessary.

An important feature of the nonparametric approach is its elegant generalization (due to Aalen and Johansen 1978), to estimating *transition probabilities*. The basic tool is the (matrix) *product-integral*. Let I be the identity matrix and G a matrix-valued function. The corresponding product–integral is defined as

$$\Pi_0^t(I + G(ds)) = \lim_{\max|t_v - t_{v-1}| \to 0} \prod (I + G(t_v) - G(t_{v-1}))$$

where $0 = t_0 < t_1 < \ldots < t_n = t$ is a partition of $[0, t]$. In particular, if G is continuous and scalar

$$\Pi_0^t(1 + G(ds)) = e^{G(t) - G(0)}$$

and if G is a scalar step function

$$\Pi_0^t(1 + G(ds)) = \prod_{k=1}^{K} (1 + \Delta G(t_{(k)}))$$

where $t_{(0)} = 0$ and $0 < t_{(1)} < \ldots < t_{(K)} \leqslant t$ are the jump times of G and

$$\prod G(t_{(k)}) = G(t_{(k)}) - G(t_{(k-1)})$$

Define

$$\alpha nn(t) = -\sum_{j \neq h} \alpha_{hj}(t)$$

and the intensity matrix function

$$A(t) = ((\int_0^t \alpha hj(u)du))$$

then the matrix $P(s, t) = ((P_{hj}(s, t)))$ of transition probabilities

$$P_{hj}(s, t) = \mathrm{Prob}(X_i(t) = j | X_i(s) = h)$$

is given by

$$P(s, t) = \Pi_s^t(I + A(\mathrm{d}u))$$

The *Aalen–Johansen* estimator of $P(s, t)$ is obtained by plugging the matrix of Nelson–Aalen estimators $(\hat{A}_{hj}(t))$ into this formula

$$\hat{P}(s, t) = \Pi_s^t(I + \hat{A}(du))$$

For the simple two-state model for survival data $\hat{P}_{00}(0, t)$ reduces to the classical Kaplan and Meier (1958) estimator $\hat{S}(t) = \prod T_i \leqslant t(1 - dN_{01}(T_i)/Y_0(T_i))$ of the survival function $S(t)$.

As documented in detail by Andersen et al. (1993, Sect. IV. 4), there is a well-developed theory, again based on stochastic integrals and martingales, about the asymptotic properties of the Aalen–Johansen estimator.

5.1.4 Semiparametric regression Markov models. For Markov models with several states, there will often be too little empirical basis for estimating freely

varying transition intensities between all states, so that more parsimonious models are required. One possibility (demonstrated in detail by Klein et al. 1993 and Keiding et al. 1999 for the bone marrow transplant example above) is to postulate that transition intensities *into* a state are proportional: $\alpha_{hj}(t) = \alpha_{0j}(t)e^{\beta_h}$ for all (relevant) h. The statistical analysis of this model is equivalent to a Cox regression model with time-dependent covariates indicating the departure state h.

The Cox model (proposed by Cox (1972a, 1972b) for survival data and for renewal processes, respectively), may also be used for including covariates not depending on time, this preserving the Markov property. The consequent generalization of the Aalen–Johansen estimator was studied by Andersen et al. (1991b, 1993, p. 512 ff.) and implemented for the bone marrow transplant problems just mentioned.

5.2 Beyond Markov Processes

The most important deviations from the Markov property in practice are various kinds of *duration dependence*, where transition intensities depend on other time origins that $t = 0$, typically the time at entry to the present state. There are two main approaches to handling these.

As long as transition intensities depend only on one time origin each (e.g., all intensities depend only on duration in the present state), a model for the multistate process may be obtained by combining independent submodels for each transition intensity. These may, in turn, be modeled as constant or piecewise constant or by non- or semiparametric models, and as long as there is a unidirectional flow in the model, transition probabilities are still straightforward explicit functionals which may be estimated by plugging in the intensity estimates. Variance calculations may, however, become less direct.

More elaborate models will include several time origins (such as both age, disease duration, duration since marriage dissolution), often in piecewise constant intensity models or semiparametric regression models such as the Cox model. An example concerning mortality of diabetics with or without nephropathy was studied by Andersen (1988), see also the consequences for Danish life insurance premiums (Ramlau-Hansen et al. 1987).

The heterogeneity between individuals cannot always be accounted for by *observed* covariates, and at least some residual heterogeneity will have to be modeled as random variation. An important class of models here is the multiplicative *frailty models*, obtained by postulating random (usually transition-specific) *frailties* W^i_{hj}, and conditional transition intensities

$$\alpha^i_{hj}(t|W^i_{hj}) = \alpha_{0hj}(t)W^i_{hj}e^{\beta' z_i}$$

5.3 Hypothetical Calculations in Multistate Models

As mentioned earlier, there is often considerable interest in studying the consequences of the estimated transition intensities by calculating summary measures such as transition probabilities. When a full model has been estimated, this can be done not only of the model observed 'in this world.' Rather, the consequences of an assumed (or fitted) multistate model may be usefully further illustrated by calculating transition probabilities in hypothetical models obtained by changing some of the parameters. An elaborate example of this was given by Keiding et al. (2000) for the bone marrow transplantation context, but similar calculations have been performed in the competing risks model ever since the first discussion by Bernoulli 1760 of the effect on population mortality of removing smallpox through vaccination. The interpretational justification of such calculations was discussed heatedly, particularly in the 1970s (see Gail 1975, Prentice et al. 1978, and Kalbfleisch and Prentice 1980, Chap. 7).

Hoem (1969b) derived a simple expression for the modified transition probability in the following particular Markov process situation. Assume the state space $\mathscr{S} = \{1,\ldots, p\} = \mathscr{K} \cup \mathscr{R} = \{1,\ldots k\} \cup \{k+1,\ldots, p\}$, where all states $h \in \mathscr{R}$ are absorbing.

Let $P_{hj}(s, t)$ be the transition probabilities corresponding to transition intensities $\alpha_{hj}(t)$ and $Q_{hj}(s, t)$ these corresponding to intensities

$$\rho_{hj}(t) = \begin{cases} \alpha_{hj}(t), & j \in \mathscr{K} \\ 0, & j \in \mathscr{R} \end{cases}$$

Then if $\alpha_{h\mathscr{R}}(t) = \sum_{j \in \mathscr{R}} \alpha_{hj}(t)$ is independent of h (non-selective occurrence of the events in $\mathscr{R}$), we get

$$P_{hj}(s, t) = Q_{hj}(s, t) \exp\left(-\int_s^t \alpha_{h\mathscr{R}}(u)\mathrm{d}u \right)$$

for $h \in \mathscr{K}$, $j \in \mathscr{K}$, $h \neq j$, $s < t$. A partial converse of this result by Cohen (1972) is that if

$$P_{hj}(s, t) = Q_{hj}(s, t) \sum_{j \in \mathscr{K}} P_{hj}(s, t)$$

for all $h \in \mathscr{K}$, $j \in \mathscr{K}$, $s < t$, then $\alpha_{h\mathscr{R}}(t)$, does not depend on $h \in \mathscr{K}$.

5.4 Partial Model Specification

We have so far assumed that the multistate model was completely specified through statistical modeling of all transition intensities and a specific probability mechanism for the combination of these into transition probabilities. In a series of papers, (Pepe 1991, Pepe et al. 1991, Pepe and Mori 1993, Couper and Pepe 1997), Pepe and co-workers have developed estimates of certain functionals in multistate models without as-

suming a full probability structure. One example is the *prevalence* of a transient condition indicated by state c defined as

$$P_{0c}(0, t)/\sum_{j \in T} P_{0j}(0, t)$$

where T is the set of transient states and 0 is a fixed 'initial' state. The idea is to estimate numerator and denominator separately by simple linear combinations of Kaplan–Meier estimates. Easily applicable variance estimates are then available which in one recent application (Klein et al. 2000) showed that the precision of the Pepe approach was close to the more elaborate (and restrictive) complete Markov model.

Another example of partial model specification occurs when the model contains time-dependent covariates which are not purely endogenous. In fact, for time-fixed covariates we just (cf. Sect. 4) conditioned on their observed values without specifying their distribution $f(Z_i)$, but for time-dependent covariates such a conditioning is more tricky. Formally, the likelihood will contain factors for the stochastic development of $Z_i(t)$ given the history $\mathscr{F}_{t^-} = \mathscr{X}_{t^-} \vee \mathscr{Z}_{t^-}$ and the likelihood described in Sect. 4 is not the full likelihood but only a partial likelihood for the parameters for the transition intensities $\alpha^i_{hj}(t)$. This means that inference for the transition *intensities* may be based on this partial likelihood whereas the transition *probabilities* typically will depend also on the parameters in the model for $Z_i(t)$.

Thus, if the model contains time-dependent covariates which are not purely endogenous then transition probabilities cannot be estimated using only a partial model specification. A joint model for the multistate process $X_i(t)$ and the time-dependent covariates $Z_i(t)$ is needed. When $Z_i(t)$ only takes a finite number of values this joint model could, again, be a multistate model where $Z_i(t)$ is now endogenous (e.g., Andersen 1986, Andersen et al. 1991b). An example of a more general joint model was presented by Wulfsohn and Tsiatis (1997).

6. *Observational Patterns*

As emphasized in the introduction, event history models rarely are observed completely. Some patterns of incomplete observation are more easily handled than others, and this final section aims at introducing some of the more important classes.

6.1 *Delayed Entry and Right Censoring*

Most of the statistical methods introduced in this article depend directly or indirectly on the likelihood representation of Sect. 4.

Two patterns of incomplete observations are particularly easily tractable, because they lead to only minor modification of the likelihood: *Delayed entry* where individual i enters at some time V_i and *right censoring* where nothing is known about i after some time U_i. Both V_i and U_i may be random although only dependent on the previous history of the process (see, e.g., Andersen et al. 1993 for precise specification of this and further discussion). The reason for the particular tractability of these mechanisms is that the 'at risk' indicator $Y^i_h(t) = I\{X_i(t-) = h\}$ just needs to be amended to

$$Y^i_h(t) = I\{X_i(t-) = h,\ V_i < t \leqslant U_i\}$$

the statistical methods then all go through.

6.2 *Interval Censoring*

An important class of incomplete observational patterns consists in the times of some (but often not all) transitions not being known exactly but only up to an interval, e.g. between visits to a clinic or between censuses. (The *number* of transitions usually is assumed known precisely.) A classical approach in demography (Hoem 1969a) is to approximate the 'exposure'

$$S^i_h = \int_{V_i}^{U_i \wedge \tau_i} Y^i_h(t)\mathrm{d}t$$

another to impute values in the observation interval for the time at risk. Systematic studies of nonparametric maximum likelihood estimation under interval censoring of the healthy $\rightarrow$ diseased transition in the unidirectional illness–death model are by Frydman (1992, 1995), Joly et al. (1999), and Gaüzère (2000). Kay (1986) and Andersen et al. (1991a) exemplified interval-censored observation in the reversible illness–death model, using piecewise constant intensity models.

6.3 *Conditioning in Multistate Models*

Many observational patterns in event history analysis may be described by conditional distributions in simpler models, which often describe 'direct' observations that are practically unobtainable. A prime example is the left truncation just described, another *right truncation* with widespread use in studies of AIDS patients whose development is often observed conditional on having contracted the disease before the study entry. For general discussions see Kalbfleisch and Lawless (1989), Keiding and Gill (1990), and Gross and Huber-Carol (1992). A more elaborate

application of such retrospective observational plans obtained by conditioning in an underlying 'prospective' Markov process model was documented by Aalen et al. (1980) for the four state interaction of life history events example described above (Fig. 6). These authors relied heavily on the concise but important general framework of Hoem (1969b). We shall here briefly outline how this methodology works for a simple example of retrospective incidence estimation, obtained from the Markov illness–death model without recovery illustrated in Fig. 4. Assume that we study a random sample of individuals alive at same fixed age u: for those who had by then contracted the disease the age at which this happened is recorded. The observed multistate model has state space $\mathscr{K} = \{0, 1\}$ and transition probabilities

$$Q_{hj}(s, t) = \text{Prob}\{X(t) = j | X(s) = h, X(u) \neq 2\}$$

for $h, j \in \{0, 1\}$ and $0 < s < t < u$. We get

$$Q_{hj}(s, t) = P_{hj}(s, t) \frac{P_{j\mathscr{K}}(t, u)}{P_{h\mathscr{K}}(t, u)}$$

where $P_{hj}(s, t)$ are the transition probabilities in the original illness–death process and $P_{h\mathscr{K}} = P_{h0} + P_{h1}$. Hoem (1969b) used the term *purged* for the conditional Markov process on $\mathscr{K}$ with transition probabilities $Q_{hj}(s, t)$. The transition intensity of the purged process is

$$\lambda_{01}(t) = \alpha_{01}(t) \frac{P_{11}(t, u)}{P_{1\mathscr{K}}(t, u)}$$

and it may be proved that if the mortality of the diseased is never smaller than that of the healthy, that is, $\alpha_{02}(t) \leqslant \alpha_{12}(t)$ for all $t \leqslant u$, then $\lambda_{01}(t) \leqslant \alpha_{01}(t)$, with equality if and only if $\alpha_{02}(t) \equiv \alpha_{12}(t)$. This documents the intuitively obvious result, that the retrospective study will underestimate the disease incidence because of *survivor selection*. Andersen and Green 1985 used such methodology to study robustness of diabetes incidence estimates in a situation where diabetics were only observed conditionally on not emigrating before a certain age.

6.3.1 Competing risks revised. The controversial calculation in the competing risks model (Fig. 2) is the transition probability $P_{01}(0, t)$, the probability that event 1 has happened before time t. In addition to the direct calculation (in epidemiology often called the 'cumulative incidence rate'), it is (surprisingly) common to calculate the hypothetical probability in a model obtained by setting $\alpha_2(t) \equiv 0$ (cf. the classical smallpox application by Bernoulli 1760) quoted above). Pepe and Mori (1993) discussed these as well as the third possibility, calculating the *conditional* transition probability $\text{Prob}(X(t) = 1 | X(0) = 0, X(t) \neq 2)$; the analysis is very similar to that just shown for the retrospective incidence estimation.

6.3.2 The prevalent cohort study. An important sampling frame for the illness–death model without recovery is the *prevalent cohort study* where a cross-sectional sample of diseased is taken at a fixed calendar time. Keiding (1992) (see also Lund 2000), discussed the conditions for correct inference on mortality $\alpha_{12}(t)$ based on follow-up of the diseased, studied under left-truncation, and compared with inference based on the *length-biased* durations which include the retrospective time from disease onset, as well as to the *forward recurrence time* from sampling to death, assuming stationarity.

A comprehensive substantive discussion of biases in prevalent cohorts was given by Brookmeyer (1998). Retrospective estimation of incidence based on the disease onset information of the survivors and independent lethality information was exemplified by Keiding et al. (1989), see also Ogata et al. (2000).

6.4 Some Other Partial Information Designs

Sometimes interval censoring is extreme: in a cross-sectional study it is for all individuals only known whether or not an event has happened at age of sampling. Such *current-status data* were discussed in detail by Diamond and McDonald (1992), Keiding (1991), and Keiding et al. (1996), and there is an elaborate recent mathematical-statistical development in this area (see, e.g., Groeneboom and Wellner 1992, Lin et al. 1998). For *time to pregnancy* data, Keiding et al. (1999) proposed using the *current duration* elapsed so far; under suitable stationarity conditions this distribution can be considered a backward recurrence time.

7. Concluding Remarks

We have attempted to provide an exposition of current problems and solutions in event history analysis in continuous time. The probability models are continuous-time nonhomogeneous Markov processes and generalizations, the main statistical techniques are log-linear intensity models as well as non and semiparametric counting process-based models. It is emphasized that event history analysis often also includes a

'synthesis' phase and that considerable care (and consequent methodological innovation) is motivated by the many indirect observational patterns.

Bibliography

Aalen O O 1976 *Statistical inference for a family of counting processes*. Ph.D. thesis, University of Copenhagen, Denmark

Aalen O O 1978 Nonparametric inference for a family of counting processes. *Annals of Statistics* **6**: 701–26

Aalen O O, Borgan Ø, Keiding N, Thormann J 1980 Interaction between life history events: Nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics* **7**: 161–71

Aalen O O, Johansen S 1978 An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* **5**: 141–50

Andersen P K 1986 Time-dependent covariates and Markov processes. In: Moolgavkar S H, Prentice R L (eds.) *Modern Statistical Methods in Chronic Disease Epidemiology*. Wiley, New York, pp. 82–103

Andersen P K 1988 Multistate models in survival analysis: A study of nephropathy and mortality in diabetes. *Statistics in Medicine* **7**: 661–70

Andersen P K, Borgan Ø, Gill R D, Keiding N 1993 *Statistical Models Based on Counting Processes*. Springer, New York

Andersen P K, Green A 1985 Evaluation of estimation bias in an illness–death-emigration model. *Scandinavian Journal of Statistics* **12**: 63–8

Andersen P K, Hansen L S, Keiding N 1991a Assessing the influence of reversible disease indicators on survival. *Statistics in Medicine* **10**: 1061–7

Andersen P K, Hansen L S, Keiding N 1991b Non- and semi-parametric estimation of transition probabilities from censored observations of a non-homogeneous Markov process. *Scandinavian Journal of Statistics* **18**: 153–67

Arjas E, Haara P 1984 A marked point process approach to censored failure data with complicated covariates. *Scandinavian Journal of Statistics* **11**: 193–209

Bernoulli D 1760 Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour le prévenir. *Historie avec le Mémoires, Académie Royal des Sciences* 1–45

Blossfeld H, Rohwer G 1995 *Techniques of Event History Modeling*. Erlbaum, Mahwah, NJ

Brookmeyer R 1998 Biased sampling of cohorts in epidemiology. In: Armitage P, Colton T (eds.) *Encyclopedia of Biostatistics*. Wiley, New York, pp. 338–50

Chiang C L 1968 *Introduction to Stochastic Processes in Biostatics*. Wiley, New York

Cohen J E 1972 When does a leaky compartment model appear to have no leaks? *Theoretical Population Biology* **3**: 404–5

Commenges D 1999 Multi-state models in epidemiology. *Lifetime Data Analysis* **5**: 315–27

Couper D, Pepe M S 1997 Modelling prevalence of a chronic condition: Chronic graft-versus-host disease after bone marrow transplantation. *Statistics in Medicine* **16**: 1551–71

Courgeau D, Lelièvre E 1992 *Event History Analysis in Demography*. Clarendon Press, Oxford, UK

Cox D R 1972a Regression models and life tables. *Journal of the Royal Statistical Society B* **34**: 187–220

Cox D R 1972b The statistical analysis of dependencies in point processes. In: Lewis P A W (ed.) *Stochastic Point Processes*. Wiley, New York, pp. 55–66

Diamond I D, McDonald J W 1992 Analysis of current-status data. In: Trussel J, Hankinson R, Tilton J (eds.) *Demographic Applications of Event History Analysis*. Clarendon Press, Oxford, UK, pp. 231–52

Frydman, H 1992 A non-parametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *Journal of the Royal Statistical Society Series B* **54**: 853–66

Frydman, H 1995 Semiparametric estimation in a three-state duration dependent Markov model from interval-censored observations with application to AIDS data. *Biometrics* **51**: 502–11

Gaüzère F 2000 *Approche non-paramétrique pour un modèle* 3 *états avec censures par intervalles—application à la dépendance*. Ph.D. thesis, Université Victor Segalen Bordeaux 2

Gail M 1975 A review and critique of some models used in competing risk analysis. *Biometrics* **31**: 209–22

Groeneboom P, Wellner J A 1992 *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel

Gross S T, Huber-Carol C 1992 Regression models for right truncated left censored survival data. *Scandinavian Journal of Statistics* **19**: 193–213

Hoem J M 1969a Fertility rates and reproduction rates in a probabilistic setting. *Biométrie-Praximétrie* **10**: 38–66

Hoem J M 1969b Purged and partial Markov chains. *Skandinavisk Aktuar Tidsskrift* **52**: 147–55

Hoem J M 1976 The statistical theory of demographic rates. A review of current developments. *Scandinavian Journal of Statistics* **3**: 169–85

Hougaard P 1999 Multi-state models: A review. *Lifetime Data Analysis* **5**: 239–64

Joly P, Commenges D 1999 A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to AIDS. *Biometrics* **55**: 887–90

Kalbfleisch J D, Lawless J F 1989 Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association* **84**: 360–72

Kalbfleisch J D, Prentice R L 1980 *The Statistical Analysis of Failure Time Data*. Wiley, New York

Kaplan E L, Meier P 1958 Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**: 457–81

Kay R 1986 A Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics* **42**: 855–65

Keiding N 1991 Age-specific incidence and prevalence: A statistical perspective. *Journal of the Royal Statistical Society A* **154**: 371–412

Keiding N 1992 Independent delayed entry. In: Klein J P, Goel P K (eds.) *Survival Analysis: State of the Art*. Kluwer, Dordrecht, pp. 309–26

Keiding N, Begtrup K, Scheike T H, Hasibeder G 1996 Estimation from current-status data in continuous time. *Lifetime Data Analysis* **2**: 119–29

Keiding N, Gill R D 1990 Random truncation models and Markov processes. *Annals of Statistics* **18**: 582–602

Keiding N, Hartvig H, Tvede M, Juul S 1999 Estimating time to pregnancy from current durations in a cross-sectional sample. Research Report 99/7, Department of Biostatistics. University of Copenhagen, Denmark

Keiding N, Holst C, Green A 1989 Retrospective estimation of diabetes incidence from information in a current prevalent population and historical mortality. *American Journal of Epidemiology* **130**: 588–600
Keiding N, Klein J P, Horowitz M M 2000 Multistate models and outcome prediction in bone marrow transplantation. Research Report 00/1, Department of Biostatistics. University of Copenhagen, Denmark
Klein J P, Keiding N, Copelan E A 1993 Plotting summary predictions in multistate survival models: Probabilities of relapse and death in remission for bone marrow transplantation patients. *Statistics in Medicine* **12**: 2315–32
Klein J P, Keiding N, Shu Y, Szydlo R M, Goldman J M 2000 Summary curves for patients transplanted for chronic myeloid leukemia salvaged by a donor lymphocyte infusion: The current leukemia free survival curve. *British Journal of Haematology* **109**: 148–52
Lin D Y, Oakes D, Ying Z 1998 Additive hazards regression with current status data. *Biometrika* **85**: 289–98
Lund J 2000 Sampling bias in population studies—how to use the Lexis diagram. *Scandinavian Journal of Statistics* **27**: 589–604
Ogata Y, Katsura K, Keiding N, Holst C, Green A 2000 Empirical Bayes age-period-cohort analysis of retrospective incidence data. *Scandinavian Journal of Statistics* **27**: 415–32
Pepe M S 1991 Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* **86**: 770–8
Pepe M S, Longton G, Thornquist M 1991 A qualifier *Q* for the survival function to describe the prevalence of a transient condition. *Statistics in Medicine* **10**: 413–21
Pepe M S, Mori M 1993 Kaplan–Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine* **12**: 737–51
Prentice R L, Kalbfleisch J D, Peterson A V, Flournoy N, Farewell V T, Breslow N 1978 The analysis of failure time data in the presence of competing risks. *Biometrics* **34**: 541–54
Ramlau-Hansen H, Jespersen N C B, Andersen P K, Borch-Johnsen K, Deckert T 1987 Life insurance for insulin-dependent diabetics. *Scandinavian Actuarial Journal* : 19–36
Wulfsohn M S, Tsiatis A A 1997 A joint model for survival and longitudinal data measured with error. *Biometrics* **53**: 330–9

P. K. Andersen and N. Keiding

Event-history Analysis: Selectivity

1. Introduction

The general purpose of the analysis of event history data is to explain why certain individuals are at a higher risk of experiencing the event(s) of interest than others. This can be accomplished by using special types of methods which, depending on the field in which they are applied, are called failure-time models, life-time models, survival models, transition-rate models, response-time models, event history models, duration models, or hazard models. Here, the terms event history model and hazard model are used interchangeably.

In hazard models, the risk of experiencing an event within a short time interval is regressed on a set of covariates. Two special features distinguish hazard models from other types of regression models: they make it possible to include censored observations in the analysis and to use time-varying explanatory variables. Censoring is, in fact, a form of partially missing information: On the one hand, it is known that the event did not occur during a given period of time, but, on the other hand, the time at which the event occurred is unknown. Time-varying covariates are covariates that may change their value during the observation period. The possibility of including covariates which may change their value in the regression model makes it possible to perform a truly dynamic analysis.

In the context of the analysis of survival and event history data, the problem of unobserved heterogeneity, or the bias caused by not being able to include particular important explanatory variables in the regression model, has received a great deal of attention. This is not surprising because this phenomenon, which is also referred to as selectivity or frailty, may have a much larger impact in hazard models than in other types of regression models: Unobserved heterogeneity may introduce, among other things, downwards bias in the time effects, spurious effects of time-varying covariates, spurious time-covariate interaction effects, as well as dependence between competing risks and repeatable events. This may be true even if the unobserved heterogeneity is uncorrelated with the values of the observed covariates at the start of the process under study. Several model-based approaches have been proposed to correct for unobserved heterogeneity.

2. State, Event, Duration, and Risk Period

In order to understand the nature of even history data and the purpose of event history analysis, it is important to understand the following four elementary concepts: state, event, duration, and risk period. These concepts are illustrated below using an example from the analyses of marital histories.

The first step in the analysis of event histories is to define the relevant states which are distinguished. The states are the categories of the 'dependent' variable the dynamics of which we want to explain. At every particular point in time, each person occupies exactly one state. In the analysis of marital histories, four states are generally distinguished: never married, married, divorced, and widow(er). The set of possible states is sometimes also called the state space.

An event is a transition from one state to another, that is, from an origin state to a destination state. In

this context, a possible event is 'first marriage,' which can be defined as the transition from the origin state, never married, to the destination state, married. Other possible events are: a divorce, becoming a widow(er), and a non-first marriage. It is important to note that the states which are distinguished determine the definition of possible events. If only the states married and not married were distinguished, none of the above-mentioned events could have been defined. In that case, the only events that could be defined would be marriage and marriage dissolution.

Another important concept is the risk period. Clearly, not all persons can experience each of the events under study at every point in time. To be able to experience a particular event, one must occupy the origin state defining the event, that is, one must be at risk of the event concerned. The period that someone is at risk of a particular event, or exposed to a particular risk, is called the risk period. For example, someone can only experience a divorce when he or she is married. Thus, only married persons are at risk of a divorce. Furthermore, the risk period(s) for a divorce are the period(s) that a subject is married. A strongly related concept is the risk set. The risk set at a particular point in time is formed by all subjects who are at risk of experiencing the event concerned at that point in time.

Using these concepts, event history analysis can be defined as the analysis of the duration of the non-occurrence of an event during the risk period. When the event of interest is 'first marriage,' the analysis concerns the duration of nonoccurrence of a first marriage, in other words, the time that individuals remained in the state of never being married. In practice, as will be demonstrated below, the dependent variable in event history models is not duration or time itself but a rate. Therefore, event history analysis can also be defined as the analysis of rates of occurrence of the event during the risk period. In the first marriage example, an event history model concerns a person's marriage rate during the period that he/she is in the state of never having been married.

3. *Basic Statistical Concepts*

The manner in which the basic statistical concepts of event history models are defined depends on whether the time variable T, indicating the duration of non-occurrence of an event, is assumed to be continuous or discrete. Even though it seems logical to assume T to be a continuous variable, in many situations this assumption is not realistic. First, it may happen that T is not measured accurately enough to be treated as strictly continuous. This occurs, for example, when the duration variable in a study on the timing of the first birth is measured in completed years instead of months or days. Second, the events of interest can sometimes only occur at particular points in time. Such an intrinsically discrete T occurs, for example, in studies on voting behavior.

Suppose that we are interested in explaining individual differences in women's timing of the first birth. In that case, the event is having a first child, which can be defined as the transition from the origin state, no children, to the destination state, one child. This is an example of what is called a single non-repeatable event, where the term single reflects that the origin state no children can only be left by one type of event, and the term non-repeatable indicates that the event can occur only once. Below, situations in which there are several types of events (multiple risks) and in which events may occur more than once (repeatable events) are presented. In the first birth example, it seems most appropriate to assume the time variable to be a continuous variable although it is, of course, measured discrete, for instance, in days, months, or years after a woman's 15th birthday.

Suppose T is a continuous random variable indicating the duration of nonoccurrence of the first birth. Let $f(t)$ be the probability density function of T, and $F(t)$ the distribution function of T. As always the following relationships exist between these two quantities,

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t+\Delta t)}{\Delta t} = \frac{\partial F(t)}{\partial t}$$

$$F(t) = P(T \leqslant t) = \int_0^t f(u)\,d(u)$$

The survival probability or survival function, indicting the probability of nonoccurrence of an event until time t, is defined as

$$S(t) = 1 - F(t) = P(T \geqslant t) = \int_t^\infty f(u)\,d(u)$$

Another important concept is the hazard rate or hazard function, $h(t)$, expressing the instantaneous risk of experiencing an event at $T = t$, given that the event did not occur before t. The hazard rate is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t+\Delta t \mid T \geqslant t)}{\Delta t} = \frac{f(t)}{S(t)}$$

in which $P(t \leqslant T < t+\Delta t \mid T \geqslant t)$ indicates the probability that the event will occur during $[t \leqslant T < t+\Delta t]$, given that the event did not occur before t. The hazard rate is equal to the unconditional instantaneous probability of having an event at $T = t$, $f(t)$, divided by the probability of not having an event before $T = t$, $S(t)$. It should be noted that the hazard

rate itself cannot be interpreted as a conditional probability. Although its value is always non-negative, it can take values greater than one. However, for small Δt, the quantity $h(t)\Delta t$ can be interpreted as the approximate conditional probability that the event will occur between t and $t+\Delta t$.

Above $h(t)$ was defined as a function of $f(t)$ and $S(t)$. It is also possible to express $S(t)$ and $f(t)$ completely in terms of $h(t)$, that is,

$$S(t) = \exp\left(-\int_0^t h(u)\,d(u)\right)$$

$$f(t) = h(t)\,S(t) = h(t)\exp\left(-\int_0^t h(u)\,d(u)\right)$$

This shows that the functions $f(t)$, $F(t)$, $S(t)$, and $h(t)$ give mathematically equivalent specifications of the distribution of T.

4. Log-linear Models for the Hazard rate

When working within a continuous-time framework, the most appropriate method for regressing the time variable T on a set of covariates is through the hazard rate. This makes it straightforward to assess the effects of time-varying covariates—including the time dependence itself and time-covariate interactions—and to deal with censored observations. Censoring is a form of missing data that is explained in more detail below.

Let $h(t|\boldsymbol{x}_i)$ be the hazard rate at $T=t$ for an individual with covariate vector $\boldsymbol{x}_i$. Since the hazard rate can take on values between 0 and infinity, most hazard models are based on a log transformation of the hazard rate, which yields a regression model of the form

$$\log h(t|\boldsymbol{x}_i) = \log h(t) + \sum_j \beta_j \chi_{ij} \qquad (1)$$

This hazard model is not only log-linear but also proportional. In proportional hazard models, the time-dependence is multiplicative (additive after taking logs) and independent of an individual's covariate values. Below it will be shown how to specify non-proportional log-linear hazard models by including time-covariate interactions.

The various types of continuous-time log-linear hazard models are defined by the functional form that is chosen for the time dependence, that is, for the term $\log h(t)$. In Cox's semi-parametric model, the time dependence is left unspecified (Cox 1972). Exponential models assume the hazard rate to be constant over time, while piecewise exponential model assume the hazard rate to be a step function of T, that is, constant within time periods. Other examples of parametric log-linear hazard models are Weibull, Gompertz and polynomial models.

As was demonstrated by several authors (Laird and Oliver 1981, Vermunt 1997, pp. 106–17), log-linear hazard models can also be defined as log-linear Poisson models, which are also known as log-rate models. Assume that we have—besides the event history information—two categorical covariates denoted by A and B. In addition, assume that the time axis is divided into a limited number of time-intervals in which the hazard rate is postulated to be constant. In the first birth example, this could be one-year intervals. The discretized time variable is denoted by Z. Let h_{abz} denote the constant hazard rate in the zth time interval for an individual with $A=a$ and $B=b$. To see the similarity with standard log-linear models, it should be noted that the hazard rate, sometimes referred to as occurrence-exposure rate, can also be defined as $h_{abz} = m_{abz}/E_{abz}$. Here, m_{abz} denotes the expected number of occurrences of the event of interest and E_{abz} the total exposure time in cell (a, b, z).

Using the notation of hierarchical log-linear models, the saturated log-linear model for the hazard rate h_{abz} can now be written as

$$\log h_{abz} = u + u_a^A + u_b^B + u_z^Z + u_{ab}^{AB} + u_{az}^{AZ} + u_{bz}^{BZ} + u_{abz}^{ABZ} \qquad (2)$$

in which the u terms are log-linear parameters which are constrained in the usual way, for instance, by means of ANOVA-like restrictions. Note that this is a non-proportional model because of the presence of time-covariate interactions. Restricted variants of model described in equation (2) can be obtained by omitting some of the higher-order interaction terms. For example,

$$\log h_{abz} = u + u_a^A + u_b^B + u_z^Z$$

yields a model that is similar to the proportional log-linear hazard model described in equation (1). In addition, different types of hazard models can be obtained by the specification of the time-dependence. Setting the u_z^Z terms equal to zero yields an exponential model. Unrestricted u_z^Z parameters yield a piecewise exponential model. Other parametric models can be approximated by defining the u_z^Z terms to be some function of Z. And finally, if there are as many time intervals as observed survival times and if the time dependence of the hazard rate is not restricted, one obtains a Cox regression model. Log-rate models can be estimated using standard programs for log-linear analysis using E_{abz} as a weight vector (Vermunt 1997, p. 112).

Unobserved heterogeneity or selectivity may bias the results obtained from the hazard models discussed

Table 1
Hazard rates illustrating the effect of unobserved heterogeneity

Time point	$A = 1$			$A = 2$			Ratio between $A = 2$ and $A = 1$
	$W = 1$	$W = 2$	observed	$W = 1$	$W = 2$	observed	
0	0.010	0.050	0.030	0.020	0.100	0.060	2.00
10	0.010	0.050	0.026	0.020	0.100	0.045	1.73
20	0.010	0.050	0.023	0.020	0.100	0.034	1.50
30	0.010	0.050	0.019	0.020	0.100	0.027	1.39

so far in various ways. The best-known implication is the negative bias of the time dependence. Other implications are that covariate effects may be biased (underestimated) and that one may find spurious time-covariate interactions.

We will illustrate the effects of unobserved heterogeneity with a small example. Suppose that the population under study consists of two subgroups formed by the two levels of an observed covariate A, where for an average individual with $A = 2$ the hazard rate is twice as high as for someone with $A = 1$. In addition, assume that within each of the levels of A there is (unobserved) heterogeneity in the sense that there are two subgroups $W = 1$ and $W = 2$, where $W = 2$ has a 5 times higher hazard rate than $W = 1$. Table 1 shows the assumed hazard rates for each of the possible A–W combinations at four time points. As can be seen, the true hazard rates do not change over time given A and W. The reported hazard rates in the columns labeled 'observed,' which were obtained by setting up a simple life table, show what happens if we do not observe W. First, it can be seen that both for $A = 1$ and $A = 2$ the observed hazard rates decline over time while the true rates were time constant. This illustrates the fact that unobserved heterogeneity biases the estimated time dependence in a negative direction. Second, while the ratio between the hazard rates for $A = 2$ and $A = 1$ equals the true value 2.00 at $t = 0$, it declines over time (see last column). Thus, when estimating a hazard model with these observed hazard rates, we will find a smaller effect of A than the true value of (log) 2.00. Third, in order to describe fully the pattern of observed hazard rates, we need to include a time-covariate interaction in the hazard model: the covariate effect changes (declines) over time or, equivalently, the (negative) time effect is smaller for $A = 1$ than for $A = 2$.

5. *Censoring*

A subject that always receives a great amount of attention in discussions on event history analysis is the problem of censoring. An observation is called censored if it is known that is did not experience the event of interest during some time, but it is not known when it experienced the event. In fact, censoring is a specific type of missing data. In the first-birth example, a censored case could be a woman which is 30 years of age at the time of interview (and has no follow-up interview) and does not have children. For such a woman, it is known that she did not have a child until age 30, but it is not known whether nor when she will have her first child. This is, actually, an example of what is called right censoring. Another type of censoring that is more difficult to deal with is left censoring. Left censoring means that we do not have information on the duration of nonoccurrence of the event before the start of the observation period.

As long as it can be assumed that the censoring mechanism is not related to the process under study, dealing with right censored observations in maximum likelihood estimation of the parameters of hazard models is straightforward. Let δ_i be a censoring indicator taking the value 0 if observation i is censored and 1 if it is not censored. The contribution of case i to the likelihood function that must be maximized when there are censored observations is

$$\begin{aligned}\mathscr{L}_i &= h(t_i|\boldsymbol{x}_i)^{\delta_i} S(t_i|\boldsymbol{x}_i)\\ &= h(t_i|\boldsymbol{x}_i)^{\delta_i} \exp\left(-\int_0^{t_i} h(u|\boldsymbol{x}_i)\,du\right)\end{aligned}$$

This likelihood function is, however, only valid if the censoring mechanism can be ignored for likelihood-based inference. The presence of unobserved heterogeneity that is shared by the process of interest and the censoring process will lead to a violation of the assumption that censoring is ignorable.

6. *Time-varying Covariates*

A strong point of hazard models is that one can use time-varying covariates. These are covariates that may change their value over time. Examples of interesting time-varying covariates in the first-birth example are a woman's marital and work status. It should be noted that, in fact, the time variable and interactions between time and time-constant covariates are time-varying covariates as well.

The saturated log-rate model described in equation (2), contains both time effects and time-covariate interaction terms. Inclusion of ordinary time-varying covariates does not change the structure of this hazard model. The only implication of, for instance, covariate B being time varying rather than time constant is that in the computation of the matrix with exposure times E_{abz} it has to taken into account that individuals can switch from one level of B to another.

The presence of unobserved heterogeneity may seriously bias the effects of time-varying covariates. More precisely, the effects of time-varying covariates may be partially spurious as a result of the presence of unobserved risk factors influencing both the covariate process and the dependent process. An example of an association that may be partially spurious is the association between a woman's labor participation and the first birth rate. In several studies, it has been found that women who are not employed have higher first birth rates than women who are employed. This association may, however, be partially the result of unobserved causes that work and fertility behavior have in common, such as certain gender role attitudes.

7. *Multiple Risks*

Thus far, only hazard rate models for situations in which there is only one destination state were considered. In many applications it may, however, prove necessary to distinguish between different types of events or risks. In the analysis of the first-union formation, for instance, it may be relevant to make a distinction between marriage and cohabitation. In the analysis of death rates, one may want to distinguish different causes of death. And in the analysis of the length of employment spells, it may be of interest to make a distinction between the events voluntary job change, involuntary job change, redundancy, and leaving the labor force.

The standard method for dealing with situations where—as a result of the fact that there is more than one possible destination state—individuals may experience different types of events is the use of a multiple-risk or competing-risk model. A multiple-risk variant of the hazard rate model described in equation (1) is

$$\log h_d(t|\boldsymbol{x}_i) = \log h_d(t) + \sum_j \beta_{jd}\chi_{ij}$$

Here, the index d indicates the destination state or the type of event. As can be seen, the only thing that changes compared with the single type of event situation is that we have a separate set of time and covariate effects for each type of event.

Again the presence of unobserved heterogeneity may distort the results obtained from the hazard model. More precisely, if the different types of events have shared unmeasured risk factors, the results for each of the types of events have shared unmeasured risks factors, the results for each of the types of events is only valid under the observed hazard rates for the other risks. In fact, the resulting dependence among risks is comparable to what in the field of discrete choice modeling is known as the violation of the assumption of independence of irrelevant alternatives.

8. *Repeatable Events and Other Types of Multivariate Event Histories*

Most events studied in social sciences are repeatable, and most event history data contains information on repeatable events for each individual. This is in contrast to biomedical research, where the event of greatest interest is death. Examples of repeatable events are job changes, having children, arrests, accidents, promotions, and residential moves.

Often events are not only repeatable but also of different types, that is, we have a multiple-state situation. When people can move through a sequence of states, events cannot only be characterized by their destination state, as in competing risks models, but they may also differ with respect to their origin state. An example is an individual's employment history: an individual can move through the states of employment, unemployment, and out of the labor force. In that case, six different kinds of transitions can be distinguished which differ with regard to their origin and destination states. Of course, all types of transitions can occur more than once. Other examples are people's union histories with the states living with parents, living alone, unmarried cohabitation, and married cohabitation, or people's residential histories with different regions as states.

Hazard models for analyzing data on repeatable events and multiple-state data are special cases of the general family of multivariate hazard rate models. Another application of these multivariate hazard models is the simultaneous analysis of different life-course events. For instance, it can be of interest to investigate the relationships between women's reproductive, relational, and employment careers, not only by means of the inclusion of time-varying covariates in the hazard model, but also by explicitly modeling their mutual interdependence.

Another application of multivariate hazard models is the analysis of dependent or clustered observations. Observations are clustered, or dependent, when there are observations from individuals belonging to the same group or when there are several similar observations per individual. Examples are the occupational careers of spouses, educational careers of brothers, child mortality of children in the same family, or in medical experiments, measures of the sense of sight of

both eyes or measures of the presence of cancer cells in different parts of the body. In fact, data on repeatable events can also be classified under this type of multivariate event history data, since in that case there is more than one observation of the same type for each observational unit as well.

The hazard rate model can easily be generalized to situations in which there are several origin and destination states and in which there may be more than one event per observational unit. The only thing that changes is that we need indices for the origin state (o), the destination state (d), and the rank number of the event (m). A log-linear hazard rate model for such a situation is

$$\log h_{od}^{m}(t|\boldsymbol{x}_i) = \log h_{od}^{m}(t) + \sum_j \beta_{jod}^{m} \chi_{ij}$$

The different types of multivariate event history data have in common that there are dependencies among the observed survival times. These dependencies may take several forms: the occurrence of one event may influence the occurrence of another event; events may be dependent as a result of common antecedents; and survival times may be correlated because they are the result of the same causal process, with the same antecedents and the same parameters determining the occurrence or nonoccurrence of an event. If these common risk factors are not observed, the assumption of statistical independence of observation is violated, which may seriously distort the results.

9. *Model-based Approaches to Unobserved Heterogeneity*

As described in the previous sections, unobserved heterogeneity may have different types of consequences in hazard modeling. The best-known phenomenon is the downwards bias of the duration dependence. In addition, it may bias covariate effects, time-covariate interactions, and effects of time-varying covariates. Other possible consequences are dependent censoring, dependent competing risks, and dependent observations. This section presents two types of methods that have been proposed for correcting for unobserved heterogeneity: random-effects and fixed-effects methods.

9.1 *Random-effects Approach*

The random-effects approach is based on the introduction of a time-constant latent covariate in the hazard model (Vaupel et al. 1979). This unobserved variable is assumed to have a multiplicative and proportional effect on the hazard rate, i.e.,

$$\log h(t|\boldsymbol{x}_i, \theta_i) = \log h(t) + \sum_j \beta_j \chi_{ij} + \log \theta_i$$

Here, θ_i denotes the value of the latent variable for subject i. In the parametric random-effects approach, the latent variable is postulated to have a particular distributional form. The amount of unobserved heterogeneity is determined by the size of the standard deviation of this distribution: The larger the standard deviation of θ_i or $\log\theta_i$, the more unobserved heterogeneity there is. Vaupel et al. (1979) proposed using a gamma distribution for θ_i, with a mean of 1 and a variance of $1/\gamma$, where γ is the unknown parameter to be estimated. Another option would be to assume $\log\theta_i$ to come from a normal distribution with mean zero and standard deviation σ.

Heckman and Singer (1982) demonstrated that the results obtained from continuous-time hazard models can be sensitive to the choice of the functional form of the distribution of the random effect. Therefore, they advocated using a nonparametric characterization of this distribution by means of a finite set of mass points whose number, locations, and weights are empirically determined. The nonparametric random-effects model proposed by Heckman and Singer is, actually, a latent class or finite mixture model. As in latent class analysis, the population is assumed to be composed of a finite number of exhaustive and mutually exclusive groups formed by the categories of an unobserved variable. Suppose W is a categorical latent variable with W^* categories, and w is a particular value of W. The model of Heckman and Singer can be formulated as follows:

$$\log h(t|\boldsymbol{x}_i, \theta_w) = \log h(t) + \sum_j \beta_j \chi_{ij} + \log \theta_w$$

Here, θ_w denotes the (multiplicative) effect on the hazard rate for latent class w. The contribution of the ith subject to the likelihood function in the case of a single non-repeatable event is

$$\mathscr{L}_i = \sum_{w=1}^{W^*} \pi_w h(t_i|\boldsymbol{x}_i, \theta_w)^{\delta_i} S(t_i|\boldsymbol{x}_i, \theta_w)$$

where π_w is the proportion of the population belonging to latent class w. In the terminology used by Heckman and Singer (1982), the number of latent classes (W^*), the latent proportions (π_w), and the effects of $W(\theta_w)$ are called the number of mass points, the weights, and the mass points locations, respectively.

The most important drawback of the parametric and nonparametric random-effects approaches to unobserved heterogeneity is that the latent variable is

assumed to be independent of the observed covariates. This is, in fact, in contradiction to the omitted variables argument that is often used to motivate the use of these types of mixture models. If one assumes that particular important variables are not included in the model, it is usually implausible to assume that they are completely unrelated to the observed factors. In other words, by assuming independence among unobserved and observed factors, the omitted variable bias, or selection bias, will generally remain (Chamberlain 1985, Yamaguchi 1986, 1991, p.132). To overcome the limitations of the standard random-effects approaches, Vermunt (1997) proposed a more general nonparametric latent variable approach to unobserved heterogeneity. It differs from Heckman and Singer's method in that different types of specifications can be used for the joint distribution of the observed covariates, the unobserved covariates, and the initial state. This makes it possible to specify hazard models in which the unobserved factors are related to the observed covariates and to the initial state.

9.2 Fixed-effects Approach

A second method for dealing with unobserved heterogeneity involves adding cluster-specific effects, or incidental parameters, to the model (Chamberlain 1985, Yamaguchi 1986). In fact, a categorical variable is included in the hazard model indicating to which cluster a particular observation belongs: observations belonging to the same cluster have the same value for this 'observed' variable while observations belonging to different clusters have different values. This approach to unobserved heterogeneity, which is called the fixed-effects approach, can only be applied with multivariate survival data, that is, when there is more than one observation for the largest part of the observational units. Note that actually the unobserved heterogeneity is transformed into a form of observed heterogeneity capturing the similarity among observations belonging to the same cluster.

The advantage of using fixed-effects methods to correct for unobserved heterogeneity is that they circumvent two objections against random-effects methods presented above: No functional form needs to be specified for the distribution of the unobserved heterogeneity and the unobserved heterogeneity is automatically related to both the initial state and the time-constant covariates.

The major limitation of the fixed-effects approach is that since each cluster has its own incidental parameter, no parameter estimates can be obtained for the effects of covariates having the same value for the different observations belonging to the same cluster: Only the effects of observation-specific or of time-varying covariates can be estimated. Another problem is that the incidental parameters cannot be estimated consistently, since by definition they are based on a limited number of observations regardless of the sample size. This inconsistency may be carried over to the other parameters if the parameters are estimated by means of maximum likelihood (Yamaguchi 1986).

See also: Event History Analysis: Applications; Event-history Analysis in Continuous Time

Bibliography

Allison P D 1984 *Event History Analysis: Regression for Longitudinal Event Data*. Sage, London

Blossfeld H P, Rohwer G 1995 *Techniques of Event History Modeling*. Lawrence Erlbaum, Mahwah, NJ

Chamberlain G 1985 Heterogeneity, omitted variable bias, duration dependence. In: Heckman J J, Singer B (eds.) *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, Cambridge, UK

Cox D R 1972 Regression models and life tables. *Journal of the Royal Statistical Society B* **34**: 187–203

Heckman J J, Singer B 1982 Population heterogeneity in demographic models. In: Land K, Rogers A (eds.) *Multidimensional Mathematical Demography*. Academic Press, New York

Laird N, Oliver D 1981 Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* **76**: 231–40

Lancaster T 1990 *The Economic Analysis of Transition Data*. Cambridge University Press, Cambridge, UK

Tuma N B, Hannan M T 1984 *Social Dynamics: Models and Methods*. Academic Press, New York

Vaupel J W, Manton K G, Stallard E 1979 The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**: 439–54

Vermunt J K 1997 Log-linear models for event history histories. *Advanced Quantitative Techniques in the Social Sciences Series*. Vol. 8. Sage, Thousand Oaks, CA

Yamaguchi K 1986 Alternative approaches to unobserved heterogeneity in the analysis of repeatable events. In: Tuma N B (ed.) *Sociological Methodology 1986*. 213–49. American Sociological Association, Washington, DC

Yamaguchi K 1991 Event history analysis. *Applied Social Research Methods*, Vol. 28. Sage, Newbury Park, CA

J. K. Vermunt

Event-related/Evoked Potentials

The term event-related potential (ERP) refers to changes in the brain's electrical activity locked in time to an externally defined event. ERPs have been used since the early 1960s, when they were usually referred to as 'evoked potentials,' to investigate sensorimotor and cognitive function in healthy and clinical popu-

lations. This article outlines how ERPs are acquired, what is known about the origin of ERP signals in the brain, and the sorts of question to which the ERP method can be applied. A comprehensive introduction to these and related issues can be found in Rugg and Coles (1995).

1. *How are ERPs Obtained?*

ERPs are small in relation to features of the brain's electrical activity (the EEG) that are not time-locked to a specific event. Therefore it is necessary to employ processing techniques that can extract the ERP 'signal' from the 'noise' in which it is embedded. The most commonly employed technique is signal averaging, whereby a number of EEG segments, each time-locked to the same class of event, are averaged to give a waveform in which consistent features of the segments (the ERP) are retained while features that vary across segments are attenuated (see Fig. 1). Depending on the relative sizes of the ERP and the background activity (the 'signal-to-noise ratio'), the number of segments required for averaging might vary from as few as 10 to more than 1,000. It is important to note that the validity of the signal averaging procedure rests on the assumption that the ERP does not vary across segments to any significant extent. If this assumption is violated (e.g., because of habituation over successive segments or just poor time-locking to the eliciting event), the averaged waveform will not be representative of the time-locked activity elicited in the single segments from which it is derived.

As is shown in Fig. 1, the ERP is conventionally depicted as a plot of voltage against time, so that the ERP waveform represents fluctuations over time in the potential difference between a recording site on the scalp and some 'reference' site such as the earlobe or nosetip. These fluctuations represent the time-varying changes in neural activity that were triggered by the eliciting event. The depiction of ERPs in this fashion emphasizes the good time resolution that can be achieved with the method, and in the days when intrumentation limitations meant that ERPs could be

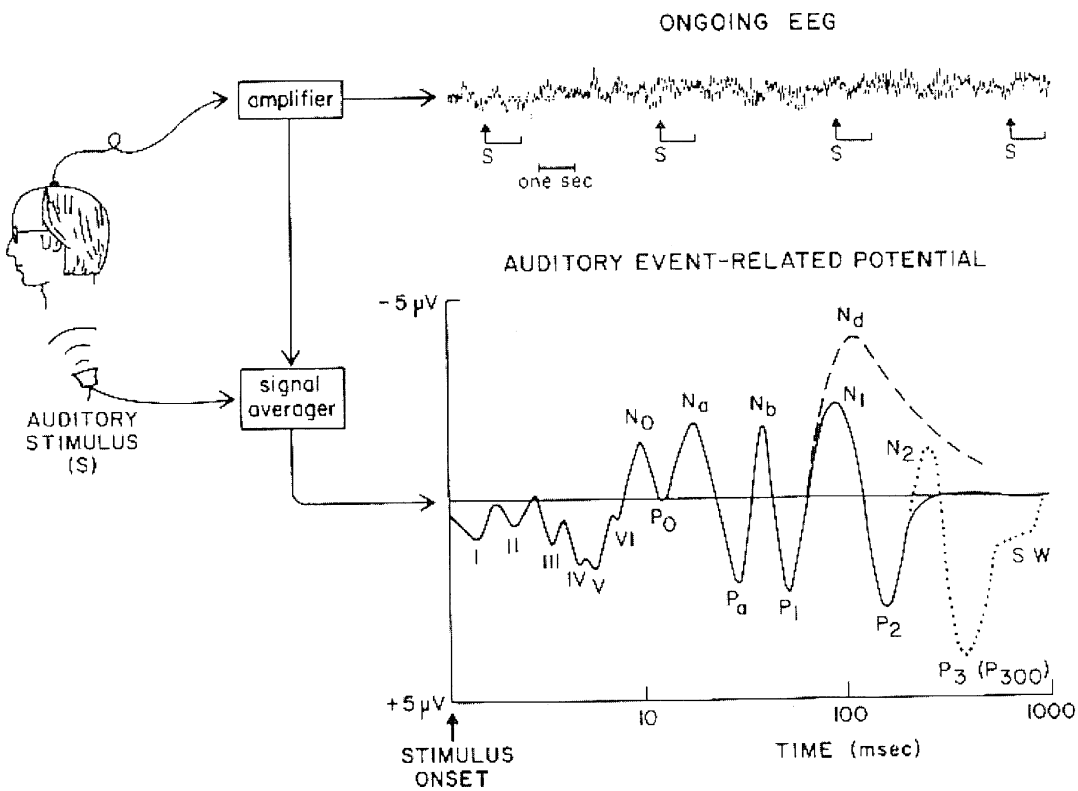

Figure 1
Idealized ERP waveform elicited by a brief auditory stimulus. The logarithmic time axis allows display of very early latency 'brainstem responses' (waves I-VI), 'midlatency' responses (No-Nb), 'vertex' potentials (P1-N2), and task-related endogenous components (Nd, N2, P300 and slow wave) (from Hillyard and Kutas 1983: with permission, from the *Annual Review of Psychology*, Volume 34 © 1983 by Annual Reviews. www.AnnualReviews.org)

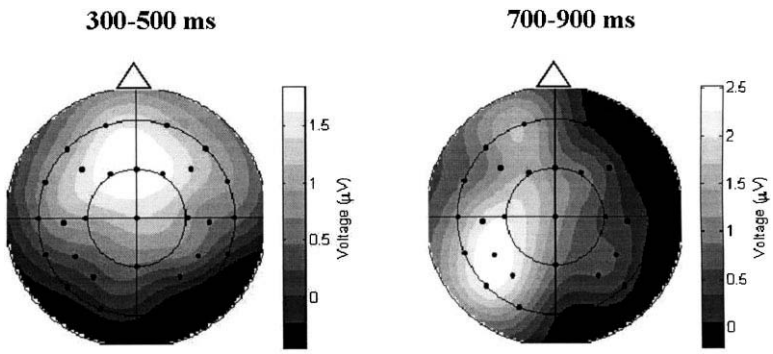

Figure 2
Maps of the distribution over the scalp of the difference in voltage between ERPs elicited by correctly classified old and new pictures in a recognition memory test. The maps are shown for two time regions—300–500 ms and 700–900 ms—following the onset of the stimuli. They depict an idealized head as viewed from above. The black dots represent the locations from which recordings were made. The scalp distributions of the differences between old and new pictures differ markedly in the two time regions illustrated, indicating that the differences were generated by different neural populations

recorded from only a few sites at a time was sufficient to allow appreciation of the data. Technological advances mean that it is now commonplace to record from 100 or more electrodes simultaneously. The rich spatial information provided by such 'dense-array' recordings can best be appreciated by depicting data in the form of scalp 'maps' rather than as time-varying waveforms (see Fig. 2).

2. *Generation and Source Localization*

Because the electrical currents generated by single neurons are miniscule, scalp electrical fields must arise from neural populations in which the activity of the individual neurons is sufficiently synchronous to allow their currents to summate and be detectable at the scalp. It is generally believed that, under most circumstances, the duration of neuronal action potentials ('spikes') is too short to allow the currents associated with these potentials to summate in this manner. It is assumed therefore that most scalp electrical activity is generated by postsynaptic potentials, the duration of which is sufficient to allow summation even if their onset in different neurons is not perfectly synchronous.

Crucially, if activity in a neuronal population is to be detectable at the scalp, it must not only overlap in time, but also have certain spatial characteristics. Specifically, the neurons comprising the population must be arranged geometrically in a way that allows the currents in each neuron to flow in the same direction, so that the currents add to rather than cancel each other. Such a configuration is known as an 'open field' and generally involves the alignment of neurons in parallel. A good example of this sort of arrangement is given by the pyramidal cells of the neocortex, which are oriented at right angles to the cortical surface. Neurons in random or circular arrangements give rise to 'closed' fields, in which the currents generated by each neuron tend to cancel each other. Because the neurons in many brain structures have closed field configurations (for example, the basal ganglia), ERPs can never provide more than a partial record of the neural activity triggered by the event to which they are time-locked.

Although the principles underlying the generation of event-related scalp electrical fields are understood quite well, the localization within the brain of the regions responsible for a given field (the field's 'sources') is a formidably difficult problem. The difficulty arises because the shape of a field on the scalp does not predict the configuration of its underlying sources in any simple way, and there is no mathematical method for uncovering this relationship unambiguously (indeed, from a mathematical perspective the problem is 'ill-posed,' i.e., it does not have a unique solution). Because of this, in all but a handful of cases it has not been possible to localize the generators of ERP phenomena with any great accuracy. Unlike researchers using other neuroimaging methods, ERP researchers have therefore tended to describe their data not in terms of the activity or function of specific brain regions or systems, but with respect to features of the ERP waveforms themselves and their presumed psychological correlates.

3. *ERP Components and their Classification*

Although it is most natural to think of ERPs as representing brain activity which follows a given event (event-following potentials), ERPs can also be recorded during intervals preceding an event (event-preceding potentials). An example of event-preceding

potentials are those that precede a motor act ('motor' or 'readiness' potentials), providing valuable information about movement preparation and execution.

As can be seen from Fig. 1, an ERP waveform following a sensory event consists of a series of peaks and troughs, or 'deflections.' Early in the waveform, the timing ('latency') of these deflections shows relatively little variability across different individuals, and within an individual the size ('amplitude') and latency of the deflections are determined largely by the physical characteristics of the stimulus (characteristics such as intensity, duration, etc.). These early deflections—sometimes referred to as 'exogenous' potentials—reflect initial sensory processing of the eliciting event, and they are employed routinely in the clinical assessment of sensory function (Chiappa 1997). For example, subtle damage to the optic nerve can result in delay in the latency of exogenous visual ERPs.

As time since stimulus onset progresses, features of the ERP waveform are determined less by physical features of the stimulus and more by the psychological state of the subject and the nature of the processing accorded the stimulus. Indeed, from around 250–300 ms or so post-stimulus onset, psychological factors largely predominate, such that the same physical stimulus will elicit markedly different waveforms depending on such factors as its relative probability of occurrence, whether it signals the need to make a response, etc. Emphasizing their dependence on internal, psychological factors rather than external ones, late-onset ERP deflections are often referred to as 'endogenous' potentials.

The deflections observed in ERP waveforms are easy to measure and provide convenient descriptive labels; for example, 'P1' and 'N2' refer respectively to the first positive and second negative deflection in a waveform; 'N250' refers to a negative deflection peaking around 250 ms post-stimulus onset, etc. It is not the case, however, that each deflection represents the time course of a single source of neural activity within the brain. Because activity recorded at the scalp represents the sum of all the sources active at a given time, the peak of an ERP deflection need not reflect the moment when any one source is maximally active. For this reason a distinction is made between ERP deflections—the surface features of waveforms—and ERP components—contributions to the waveform arising from specific sources or source configurations, the spatial and temporal overlap of which yields the recorded waveform. An important part of ERP analysis involves the decomposition of recorded waveforms into their constituent components.

4. *Applications of ERPs*

As already mentioned, exogenous ERPs are widely used in the assessment of sensory function. Efforts have also been made to use endogenous components to assess cognitive function and its impairment in various clinical groups, albeit with only limited success as aids to clinical diagnosis. From the perspective of psychology and related disciplines, the ERP method is of interest largely to the extent that it can be used to further understanding of cognitive functions and their neural underpinnings. ERP findings can be informative in these respects for several reasons, only the most important of which are outlined here (see Rugg and Coles 1995). First, neural activity associated with stimulus processing can be measured with a temporal resolution in the millisecond range. Thus, the maximum time required by the nervous system to discriminate between different classes of stimulus (e.g., novel vs. familiar faces) can be determined directly, even when volunteers are not required to discriminate between or respond to the stimuli. Second, ERP waveforms can be formed 'off-line,' after experimental trials have been sorted into different conditions according to the subject's behavior. It is therefore easy to compare records of brain activity associated with different classes of response (e.g., remembered vs. forgotten items in a memory test). Finally, ERPs can be used to investigate whether different experimental conditions engage distinct cognitive operations. On the assumption that different patterns of neural activity imply the engagement of different cognitive operations, it follows that when two ERP scalp distributions differ (as in Fig. 2), it can be concluded that the experimental conditions associated with each distribution engaged at least partially separate cognitive operations. Thus, differences in the spatial characteristics of ERP fields can be used to make inferences at a psychological level, even if the location of the neural activity responsible for these differences is unknown.

These applications of the ERP method are illustrated below with respect to two central areas of psychological enquiry. For further information on these and other areas of psychology and cognitive neuroscience in which ERPs have made a contribution see Kutas and Dale (1997) and Rugg and Coles (1995).

5. *Selective Attention*

The question of whether human selective attention acts 'early' or 'late' during stimulus processing has proved difficult to resolve on the basis of behavioral evidence alone. ERP data clearly show however that when processing load is high, selective attention exerts an effect early on during processing. This conclusion derives from the finding that, in all sensory modalities investigated, attended stimuli elicit exogenous ERP components that are larger in amplitude than the components elicited by unattended stimuli. In the auditory modality, for example, attention has been reported to influence components occurring as early as 20 ms post-stimulus onset, which reflect the earliest cortical stages of auditory processing (Woldorff and

Hillyard 1991). Such findings have been interpreted as evidence that attention acts in part by selectively biasing stimulus processing before full perceptual analysis has occurred (Luck et al. 2000).

6. Language

Whereas the distinction between syntax and semantics is fundamental from a formal linguistic perspective, the question whether semantic and syntactic processes are distinct in terms of their underlying cognitive operations is the subject of debate. In 1981 it was shown that the terminal words of semantically 'anomalous' sentences (e.g., 'he took his coffee with cream and socks') elicited a negative-going ERP deflection (the 'N400' component) that was much enhanced relative to ERPs elicited by nonanomalous endings (e.g., 'it was his first day at work'; Kutas and Hillyard 1980). Subsequent work showed that this effect was present in the auditory modality also, and could be found for words at any position within a sentence. It was also shown that the amplitude of the N400 component depended on a word's semantic expectancy, rather than whether or not it was anomalous (e.g. check elicits a larger N400 than stamp following 'he mailed the letter without a check/stamp'). Later, it was demonstrated that syntactic anomalies (e.g., 'the broker persuaded to sell the stock' vs. 'the broker hoped to sell the stock') also elicit characteristic ERP deflections, but with quite different distributions over the scalp to the N400 and, for one such deflection (the 'P600' component or 'syntactic positive shift'), a difference in polarity (Osterhout and Holcomb 1992). These deflections also do not depend on anomaly *per se*; was in the 'garden-path' sentence 'the lawyer charged the defendant was lying' elicits an enhanced P600 relative to the same word in 'the lawyer charged that the defendant was lying,' presumably reflecting the need for the listener or reader to engage in 'syntactic repair' at this point in the sentence. These and related findings have been taken as evidence in support of the idea that the derivation of syntactic and semantic information during sentence comprehension depend on distinct neural and cognitive processes (Hagoort and Brown 1999).

7. Relation to Other Methods

The ERP method is only one of several methods permitting the noninvasive measurement of human brain activity. In particular, the method of functional magnetic resonance imaging (fMRI) has recently come to prominence as a powerful way to identify event-related neural correlates of sensorimotor and cognitive function (D'Esposito et al. 1999). Whereas fMRI has far superior spatial resolution to that possible with ERPs, the temporal resolution of fMRI is markedly inferior. Since an understanding of brain function requires that neural activity be located accurately in time as well as in space, ERPs are likely to continue to play an important role in cognitive neuroscience for some time to come. Indeed, the development of methods for the integration of ERP and fMRI data would represent a major step forward for functional imaging (McCarthy 1999).

See also: Cognitive Neuroscience; Electrical Stimulation of the Brain; Functional Brain Imaging; Magnetoencephalography

Bibliography

Chiappa K H 1997 *Evoked Potentials in Clinical Medicine*, 3rd edn. Lippincott Williams & Wilkins, Philadelphia, PA

D'Esposito M, Zarahn E, Aguirre G K 1999 Event-related functional MRI: Implications for cognitive psychology. *Psychological Bulletin* **125**: 155–64

Hagoort P, Brown C M 1999 The neurocognition of syntactic processing. In: Brown C M, Hagoort P (eds.) *The Neurocognition of Language*. Oxford University Press, Oxford, UK, pp. 271–316

Hillyard S A, Kutas M 1983 Electrophysiology of cognitive processing. *Annual Review of Psychology* **34**: 33–61

Kutas M, Dale A 1997 Electrical and magnetic readings of mental functions. In: Rugg M D (ed.) *Cognitive Neuroscience*. Psychology Press, Hove, UK, pp. 169–242

Kutas M, Hillyard S A 1980 Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* **207**: 203–5

Luck S J, Woodman G F, Vogel E K 2000 Event-related potential studies of attention. *Trends in Cognitive Sciences* **4**: 432–40

McCarthy G 1999 Event-related potentials and functional MRI: A comparison of localization in sensory, perceptual and cognitive tasks. *Electroencephalography and Clinical Neurophysiology Suppl.* **49**: 3–12

Osterhout L, Holcomb P 1992 Event-related potentials elicited by syntactic anomaly. *Journal of Memory and Language* **31**: 785–806

Rugg M D, Coles M G H (eds.) 1995 *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition*. Oxford University Press, Oxford, UK

Woldorff M G, Hillyard S A 1991 Modulation of early auditory processing during selective listening to rapidly presented tones. *Electroencephalography and Clinical Neurophysiology* **79**: 170–91

M. D. Rugg

Event Sequence Analysis: Methods

Event Sequence Analysis (SA) Methods are methods seeking patterns that recur across a number of sequences. They are applied to data on social sequences like occupational careers or life courses, as well as to

historical sequences such as might emerge in policy adoption or in the development of organizational structures and to cultural sequences like the various orders of elements in different versions of a widespread ritual or story. The most common techniques for seeking these patterns are standard sequence comparison algorithms widely available in biological software.

1. Sequence Analysis and its Origins

Sequence analytic methods are part of a general turn towards the use of exploratory techniques in social science. As such they stand with techniques like cluster analysis, multidimensional scaling, and factor analysis. All of these methods seek radical simplification of the data space in the hope that analyzable patterns can be found. They contrast with the many causal techniques that take high dimensionality of the data space for granted and analyze that space by seeking to find the impact of a set of 'independent' dimensions on one 'dependent' dimension of the data.

This turn to exploratory techniques has a number of sources. One of these has been the stockpiling of immense data sets analyzable in many diverse ways. Another has been the more or less exploratory use of causal methods enabled by commodification; where hundreds of regressions precede those actually reported, the idea of classical hypothesis testing becomes questionable. Exploratory analytic techniques have generally been more common in applied social science rather than in academic work, having their widest use in market research. Within the academy, they have seen wider use in psychology than in any other social or behavioral science. While not unknown in political science and sociology (see, e.g., the various works of E. O. Laumann, such as Laumann and Knoke 1987), they are virtually absent from ceconomics.

The turn to sequence analysis in particular, however, proceeded from a critique of causal conceptions articulated by Abbott (1983, 1988, 1997) and Abell (1987). Both writers criticized what they called the 'variables' tradition for assuming that the social world is best imagined in terms of independent units of analysis whose properties 'cause' one another rather than in terms of chains of actions and events. Abell's argument was the more abstract, being rooted in a theory of social action based on rational choice. Abbott's was more empirical, being driven by the problem of comparing histories across social units. Both writers sought methods that would identify and compare similar chains of events across varieties of social sequence data.

Sequence analytic strategies vary in the amount of narrative detail they retain. Essentially, the more narrative detail retained the less powerful the modes of generalization or pattern search available. Strongest in retaining detail is David Heise's (1989) event structure analysis (ESA), which is essentially an interactive program for establishing the essential narrative dependencies in a complex story. The analyst inputs events and answers questions posed by the program about which events depend on which others and how. The program then produces a minimal representation of these dependencies. While it is very useful for establishing rigorous narrative patterns, ESA has not seen widespread use as a technique for generalizing about sequence patterns (for an example, see Griffin 1993).

Abell's (1993) method for narrative analysis retains less empirical detail. It begins with a 'narrative table.' The table is essentially a matrix of actors across actions, implicitly like the Heise pattern, but ignoring the specific content of events to focus on the network 'shape' of the connections between events. Narratives are similar if they can be reduced homomorphically to the same network structure. Abell's method has not seen substantial use, perhaps because his notion of similar narrative structure is not seen by empirical analysts as of immediate utility, despite its theoretical elegance. (But see the comments by various writers on Abell's techniques in *Journal of Mathematical Sociology* **18**, Numbers 2 and 3.)

Abbott's approach simplifies sequence data far more than do those of Abell or Heise, insisting on unilinear sequences of simple events. These sequences are then compared using the widely available pattern-matching algorithms that have emerged in computer science and biology since the 1970s (for a review, see Abbott and Tsay 2000). As might be expected given its simplicity, Abbott's method has seen the widest use of SA methods. It is variously referred to as sequence analysis, optimal matching analysis, or optimal alignment, the latter two being names for the standard algorithms used to measure resemblance between sequences.

Abbott's method (hereafter abbreviated OM for its most common name, optimal matching) presupposes a set of 'sequence data,' that is, a set of strings (sequences), each one of which is a vector of elements representing a succession of events in some discrete order. In career data these might be sequences of jobs held in each month over some period. In cultural analysis, they might be orders of gestures in rituals. Events can recur or return an indefinite number of times, and there is no necessary constraint of equal length on the sequences.

Simply described, Abbott's method consists in using dynamic programming algorithms to define distances between all possible pairs of sequences in a dataset. These distances are then analyzed by clustering or scaling to produce some typology of sequence types. This typology may be used as a simple description or as a dependent or independent variable in more standard causal techniques. Thus stated, Abbott's formulation of sequence analysis for the social sciences is equivalent to what is known in computer science as

the multiple string alignment problem, on which there is an extensive literature (see Gusfield 1997, Chap. 14).

Algorithms for multiple string alignment are typically based on simple algebra establishing 'Levenshtein distance' between two strings, the minimum number of edit operations (insertions, deletions, replacements) necessary to transform one string into the other. Although in some applications more complex operations have been investigated (see papers in Sankoff and Kruskal 1983), the string comparison literature has focused on the simplest three operations.

Finding the minimal edit distance between two sequences is a matter of inspection unless varying weights are placed on varying operations. Typical weighting schemes involve a fixed insertion and deletion weight coupled with replacement weights that vary in terms of what replaces what. Under such weighting schemes, the minimal edit distance problem must be solved by a simple dynamic recursion usually attributed to the classic paper of Needleman and Wunsch (1970).

Sequence analysts may also be interested in the question of whether a set of sequence data has some common subpattern, i.e., whether there is a characteristic subsequence that may appear in many or most sequences. This problem, known technically as the multiple local alignment problem, has been solved by a Gibbs Sampling algorithm due to Lawrence et al. (1993). Abbott has applied it to the analysis of rhetorical subsequences in sociological articles (Abbott and Barman 1997).

2. *Methodological Issues*

Two kinds of methodological issues have arisen in the area of SA methods. In the main these concern OM methods, as these are the most widely applied. The first such issues are those proper to the method itself, the second those concerning its utility. Like all methods, OM has its own internal assumptions and difficulties. Issues of data coding have vexed many users, although there is some indication that the methods, like most exploratory methods, are less vulnerable to the vagaries of coding than methods taking the data as given. One coding issue in particular has seemed worrisome, that of coding the underlying temporality. Some have used regular time intervals; others have collapsed continuous episodes. Another important issue has been the setting of weights of replacement and insertion/deletion. A wide variety of schemes have been used, but relatively few analysts have studied the impact of varying schemes. It is clear that all these areas need more experimentation than has presently been done. This is also true in terms of validation. Most typologies and clusterings in the OM literature have been accepted as is, without further validation. A few papers have used Monte Carlo validation methods.

The other major issues with OM methods in particular, as with SA methods more generally, concerns their utility. There are two versions of this concern, one small-scale, the other more general. The small-scale issue involves what SA methods can really add to event history or other methods investigating hazard rates for particular kinds of events. These are the standard 'causal' methodologies that are applied to sequence data. In areas where such methods are well-developed (examination of careers, labor force involvement, and marital status), SA methods seem to standard analysts to have little to add. In other areas, however, particularly in the analysis of sequential cultural structures like stories or rituals, event history methods themselves have no utility. It is not yet clear whether SA methods can provide standard-method career analysts with results that will surprise them. (Like any paradigm, event history methods to some extent enforce criteria for results that favor their own types of results.) All the same, the eventual success of SA methods will depend on finding regularities that event history methods cannot find. A clear discussion (and example) of such regularities appears in Stovel et al. (1996).

The more general question about SA's utility seems to have a clearer answer. There is little question that the broad project of employing pattern-matching techniques in social science data is a sound one. Partisans of SA methods entered the area largely to carry out a theoretical agenda that began as an attack on standard methodologies. But in turning to the computer science and biological literature, they found one of the most rapidly advancing areas of modern scientific methodology. Seeking patterns directly, rather than through hypothesized models, is not only good science, but also may be one of the only means of addressing the monumental data sets of the future.

3. *Future Developments*

To date, most applications of SA have involved a fairly limited substantive range. The most common application has been to career data. Some have worked with particular job histories, others with broader occupational classes. Still others have combined family data into these patterns. A number have worked with the 'careers' of larger social organizations—professions, counties, and nations. Still other applications have involved patterns of activity—at travel agencies, libraries, decision-making groups, and the like. A final group of studies has considered cultural artifacts—stories, dances, academic articles and so on. As these examples show, the method has been typically applied where standardized patterns are strongly expected.

Most early SA was done using fairly simple computer packages adapted for the purpose. At the beginning of the twenty-first century standard bio-

logical packages are being adapted for social use (e.g., CLUSTAL by scholars at the University of Strasbourg). Applications of SA can be expected to grow more common once the actual algorithm (not just the weighting scheme) can be tailored to the type of sequence regularity expected in the data.

The extraordinary flowering of pattern-matching methods in biology and computer science will undoubtedly lead to greater application of these methods within social analysis. They are already widely used in market research, of course, and therefore constitute a central part of what is by far the most extensive and comprehensive form of social scientific analysis at present. To be sure, the causal paradigm against which SA methods were developed remains strong within the empirical social sciences. But in the past, widespread availability of new methods has led to wholesale adoption even in the face of earlier paradigms. Indeed, the present causal methods spread precisely in this way. Given the rise of pattern-matching as a general strategy of scientific inquiry, SA and related methods will almost certainly spread. They will also constitute the entering wedge of a much broader set of descriptive approaches to social reality that can be expected to transform social inquiry in the next decades.

See also: Classification: Conceptions in the Social Sciences; Event History Analysis: Applications; Growth Curve Analysis; Person-centered Research; Statistical Pattern Recognition

Bibliography

Abbott A 1983 Sequences of social events. *Historical Methods* **16**: 129–47

Abbott A 1988 Transcending general linear reality. *Sociological Theory* **6**: 169–86

Abbott A 1997 Of time and space. *Social Forces* **75**: 1149–82

Abbott A, Barman E 1997 Sequence comparison via alignment and Gibbs sampling. *Sociological Methodology* **27**: 47–87

Abbott A, Tsay A 2000 Sequence analysis and optimal matching methods in sociology. *Sociological Methods and Research* **29**: 3–33

Abell P 1987 *The Syntax of Social Life Oxford*. Oxford University Press, Oxford, UK

Abell P 1993 Some aspects of narrative method. *Journal of Mathematical Sociology* **18(2–3)**: 93–134

Griffin L J 1993 Narrative, event structure analysis, and causal interpretation in sociology. *American Journal of Sociology* **98**: 1094–133

Gusfield D 1997 *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, Cambridge, UK

Heise D 1989 Modeling event structures. *Journal of Mathematical Sociology* **14**: 139–69

Laumann E O, Knoke D 1987 *The Organizational State*. University of Wisconsin Press, Madison, WI

Lawrence C E, Altschul S F, Boguski M S, Liu J S, Neuwald A F, Wooton J C 1993 Detecting subtle sequence signals. *Science* **262**: 208–14

Needleman S B, Wunsch C D 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**: 443–53

Sankoff D, Kruskal J B 1983 *Time Warps, String Edits, and Macromolecules*. Addison-Wesley, Reading, MA

Stovel K, Savage M, Bearman P 1996 Ascription in achievement. *American Journal of Sociology* **102**: 358–99

A. Abbott

Everyday Life, Anthropology of

1. The Study of Everyday Life

From the 1960s and onward the concept of 'the study of everyday life' became an increasingly popular notion in many social sciences, such as anthropology, history, and sociology, as well as in the expanding field of cultural studies.

It has come to signal a special interest in the commonplace and the seemingly trivial routines and activities of daily life as well as a focus on 'ordinary people' as creative actors rather than passive consumers or objects of domination. The different ways in which this seemingly simple concept has been put to work illustrate changing concerns in social and cultural research, as well as conflicts between different research paradigms. Like few other similar concepts, the study of everyday life has been used not only as a research perspective or analytical tool but as a research ideology. The study of everyday life often developed as an alternative to or critique of mainstream traditions.

The use of the concept as an alternative approach means that it is constantly being constructed in opposition to something else. It can be organized as the study of the life and thoughts of ordinary or common people against that of elites, intellectuals, or experts. It can be defined as the routines, habits, and phenomena which tend to be neglected because their 'everydayness' makes them invisible or self-evident, unlike colorful rituals, celebrations, or other activities with high visibility or prestige. Here it is 'the everyday' as contrasted to 'the Sunday best' in social life, working with polarities like mundane/sacred, informal/formal, repetitive/unique, trivial/special. Everyday life can further be seen as the arena of expressions and meanings threatened by invasion from 'above,' by institutions like the state, the market or just 'the system.' This perspective favors a view of everyday life as resistance against domination. The focus on resistance is often linked to a view of everyday life as a site of cultural creativity. Finally, 'the study of everyday life' often defines a research orientation rather than an empirical field. It is then seen as a certain style of doing research, usually with a focus on qualitative and ethnographic methods—exploring the hidden worlds of daily life. The argument is often that by focusing on seemingly banal phenomena new and sometimes surprising analytical

insights into major social or cultural issues may be produced.

The study of everyday life has emerged as an arena of intense cross-disciplinary dialogues, but it is also an interest which also has developed in a constant interchange with new public and political ideas about the importance of everyday life (see Gullestad 1991) (see *Common Sense, Anthropology of*).

2. The Making of a Concept

The development of everyday life as a research topic is thus a complex one. It has entered the social sciences along many paths and in this process been charged with very different theoretical and ideological meanings. There are two nineteenth-century traditions which can be seen as starting points for the twentieth-century interest in the study of everyday life.

First, the works of those cultural historians, who in the late nineteenth century turned to the everyday of 'ordinary people' to create an alternative to the classic history of 'kings and elites.' Second, the interest among early social researchers and reformers exploring the lives of urban poor or rural peasants, starting with Henry Mayhew's classic work *London Labor and the London Poor* of 1861.

2.1 Sociological Approaches

It was this latter interest in questions of social welfare that influenced the interwar ethnographic studies of the so-called Chicago School of American sociology. An early example is Nels Anderson's book *The Hobo: The Sociology of Homeless Men* of 1923.

This interest in urban everyday life of the Chicago School was developed further by the new generations of postwar American sociologists. Behind the new label 'The Sociology of Everyday Life' very different theoretical traditions and research perspectives were found. One was the symbolic interactionism of Erving Goffman's (1959) pioneering work *The Presentation of Self in Everyday Life* with its qualitative, ethnographical approach to the seemingly trivial or invisible routines of 'the everyday,' as it surfaced in patterns of social interaction. For early phenomenological contributions to this sociology of everyday life see, for example, Schutz (1973) and Douglas (1971). The third new approach was ethno-methodology (see Garfinkel 1984), mainly devoted to understanding the construction of common-sense knowledge in everyday life. Today, the sociology of everyday life usually stands for a qualitative and ethnographic approach (see Glassner and Hertz 1999).

In France, the sociology of the everyday developed along very different lines. Here the studies of popular mentalities in the Annales School of social history and especially the works of Fernand Braudel created an early analytical interest in the role of everyday life for understanding society. The sociologist and rather unorthodox Marxist Henri Lefevbre started to explore the concept of the everyday already in the interwar period and published a number of books on the topic between 1947 and 1991 (see, e.g., Lefebvre 1971). His interest in everyday life was connected with the rapid modernization France underwent especially after the Second World War with the creation of a new consumer based and standardized everyday life. He used the term 'everydayness' (*le quotidien*) to describe what he saw as the banality and conformity of modern everyday life and the ways in which it was colonized by modern capitalism. But Lefebvre's approach was not only centered around questions of alienation and domination. His treatment of the problems of 'everydayness' also stressed the ways in which social scientists underestimated the role of the everyday, not only as a site of domination but also as an arena with potentials for transcendence and change—the peak experiences or special moments embedded in it (see Shields 1999).

Lefebvre's discussions of everyday life had a strong philosophical profile, which may be one of the reasons his work did not come to inspire empirical studies in the same way as the work of another French sociologist, Michel de Certeau. His book (1984) has been very influential in shaping the approaches to the study of everyday life in many disciplines during the 1980s and 1990s. Lefebvre and de Certeau both shared an interest in the uses of space in everyday life, but the similarities end there. Unlike Lefebvre, de Certeau focused on everyday life as an arena of cultural creativity and resistance, and above all on consumption as creative cultural production.

2.2 Anthropological Approaches

In anthropology the actual concept 'study of everyday life' was not at first used very frequently, in spite of the fact that anthropology often was seen as 'the science of the everyday' with its ethnographic and qualitative explorations of life in different settings over the world. One reason for the lack of an explicit use of the concept of everyday life was that it was seen as a cultural category belonging to the modern Western world. In non-Western societies this polarity between the everyday and its counterparts, however defined, did not seem very relevant. In the broad anthropological approach to the study of culture there was no need for a subcategory labeled everyday culture.

In many ways the interdisciplinary interest in studies of everyday life among historians and sociologists from the 1960s and onward meant an anthropologization of both approaches and theoretical frameworks. It was not only the ethnographic methods and the metaphors of 'discovering' and 'exploring' everyday life as a hidden but exotic culture which accounted for this anthropologization. The anthropological connection also had to do with studies of everyday life

emphasizing a cultural perspective on society. It was in the habits, rituals, routines, and traditions of daily life that 'culture' was embedded, as invisible to the actors as 'water to the fish,' to quote a popular metaphor from the 1970s.

A much more explicit use of the category of the everyday was found in the anthropological branch of European ethnology, which from the 1960s and onward redefined itself as a discipline in which the study of everyday life in European settings, past and present, was central. This transformation may serve as an example of how the concept came to be given strong symbolic and ideological charges (see Frykman and Löfgren 1996).

As European ethnologists in the 1960s oriented their research towards modern industrial society rather than vanishing peasant traditions the basic concept of folk culture was exchanged for the study of the everyday and 'the ordinary people.' These new concepts not only signaled a break with earlier generations but also worked as a boundary marker against neighboring disciplines. Ethnologists studied everyday life with a qualitative perspective unlike sociologists and historians, the concept of everyday life was used to show that the ethnological interest in society was not concerned with formal institutions and macrostructures. In reality the concept was thus not about the contrast between everyday and non-everyday phenomena, but a way of stressing a research interest focusing on neglected groups, activities, and spheres.

3. *Everyday Life and Cultural Creativity*

Another research tradition in which the notion of everyday life became very important was the emerging field of Cultural Studies. As in European ethnology this interest in everyday life during the 1980s and 1990s was linked closely to the new concept of cultural creativity.

Just as the notion of everyday life, the concept of creativity functioned as a counter argument in cultural research. Groups and settings, which at first glance seemed characterized by passivity, simply reproducing dominant patterns, were in the new studies coming out as culture builders and bricoleurs: busy reworking, elaborating, and thickening their everyday lives with new meanings and routines. This approach is very striking in de Certeau's work as well as in much of the studies made by the so called Birmingham School of Cultural Studies. A good example is Willis' (1990) book *Common Culture: Symbolic Work at Play in the Everyday Cultures of the Young*. Here everyday creativity becomes not only a counter argument against the presupposition that creativity is an elite resource, a cultural capital of the gifted, of artists, writers, intellectuals, etc, but also against the notion that modern mass consumption is a passifying, homogenizing force in people's lives.

When consumption was redefined as 'cultural production' or 'symbolic production,' the creativity of everyday life moved in to occupy a central position in this production process. This emphasis on cultural creativity as resistance must be seen as a reaction against an earlier discourse on the seduction of mindless consumers: locked up in the iron cage held together by market forces. Creativity became, in some ways, the weapon of the weak—a positive strategy of resistance. More rarely the concept was used to describe strategies of dominance and oppression: other terms were used for the creative and imaginative ways of control and manipulation.

The pairing of cultural creativity and everyday life can thus be seen as an attempt to redress the balance, to develop a more actor-oriented approach in studies of mass consumption. In this process, however, the 'when, where, how, why, and who' of cultural creativity in everyday life was narrowed down and produced certain paved roads: a tendency to over-research some aspects and arenas and overlook others (see Löfgren 1997).

This slanting of the two concepts was not a very conscious process. It started with the focus on consumption as creativity, a creativity much focused on symbolic production and expressive forms in what Featherstone (1991) has called the aesthetization of everyday life. This very selective emphasis was made more narrow by a marked concentration to specific social settings, above all youth cultures with their constant experimentation with identities and reworking of mass media and mass consumption styles, images and commodities.

With a selective focus like this cultural creativity became a very positive concept. It was seen as enrichening, elaborating, 'thickening' everyday life, a process through which people make their daily existence colorful, unique, specific, distinct' and above all positive. The perspective usually was reserved for those defined as the underdogs of the modern world: consumers, workers, women, teenagers, colonial, and postcolonial subjects, and this is where the concept is linked closely to ideas of counter-hegemony, to the tactics of resistance in the world of mass consumption, or in postcolonial processes of globalization.

4. *Problems and Potentials*

Just like everyday life creativity became a concept mainly for the age of late modernity. The academic interest in the two concepts mirror a public and political concern too—especially in North European and North American settings—which had its roots in the counter-cultures of the 1960s and 1970s and the interests in alternative life-styles as well as the democratization of cultural expression. During this period the idea of the importance of everyday life became a powerful political symbol, condensing and unifying a

number of sometimes conflicting values (see Gullestad 1991). Everyday life was 'where the action was,' the location of 'real or authentic life.' By studying and engaging this arena silent groups and interests could be given a voice.

Such political and popular interests influenced academic research. Some activities and people easily came to be seen as more everyday than others. Another danger of this perspective was that everyday life was too easily seen as something 'down there,' in a way that seldom reflected over the problem of what the opposite of the everyday was supposed to be. It was forgotten that everyday life existed not only in retirement homes or on the streets but also in the corridors of power, that there were everyday practices not only in immigrant suburbs but also in scholarly research projects. Researchers who sometimes said that they wanted to 'get out into everyday reality' missed the point that what is interesting about the everyday is that it is ever-present.

Taking such ambiguities and possible ideological overtones into account, research interest in the study of the everyday has on the whole been important both in revitalizing and redirecting scholarly attention to the role of the seemingly trivial activities and ideas embedded in the commonplace. During the last decades of the twentieth century, the study of everyday life has been constituted as a very productive cross-disciplinary field, developing innovative research strategies for destabilizing or problematizing the taken-for-grantedness of much daily life.

The study of the everyday has also been a strategy to avoid paved road in research and develop back-door entrances to major issues. Two examples may illustrate such strategies. The first is taken from the French sociologist Kaufmann's ethnographies of daily life, his study of the conflicting values of modern marriages analyzed through the routines of doing the laundry in a sample of French families (Kaufmann 1998). Through extensive interviews with wives and husbands about their laundry habits Kaufmann throws new light on the constant negotiations typical of modern marriages, negotiations about divisions of labor, definitions of love and individual space, as well as changing perceptions of the moral economy of the family.

Similar examples of such a use of back-door entrances to major issues can be found in those studies of nationalism, which do not take ideology or official rhetoric of major institutions as their starting point, but want to understand the power of 'the national' through the modes in which it is anchored and materialized in everyday life. Such studies, for example, Billig (1995), illustrate the powerful ways in which the nation–state has been made visible and tangible in a wealth of everyday situations and settings, how citizens have been touched by the state.

Studies like these work as a reminder that the seemingly insignificant may be a good starting point for exploring the workings of society. Studies of everyday life have produced techniques and methods for understanding how a lot of cultural energy may be condensed or crystallized in a trivial situation, an unperceived object or a prosaic routine. They have also led to a much greater reflexivity about the manners in which the study of the everyday is conditioned by the researcher's own embeddedness in their own everyday practices.

See also: Collective Beliefs: Sociological Explanation; Collective Memory, Anthropology of; Domestic Life and Material Culture; Ethnology; Family, Anthropology of; Folklore; Hegemony: Anthropological Aspects; Home in Anthropology; Play, Anthropology of; Symbolic Interaction: Methodology; Tradition, Anthropology of

Bibliography

Billig M 1995 *Banal Nationalism*. Sage, London

de Certeau M 1984 *The Practice of Everyday Life*. University of California Press, Berkeley, CA

Douglas J (ed.) 1971 *Understanding Everyday Life*. Routledge and Kegan Paul, London

Featherstone M 1991 *Consumer Culture and Postmodernism*. Sage, London

Frykman J, Löfgren O (eds.) 1996 *Force of Habit: Exploring Everyday Culture*. Lund University Press, Lund, Sweden

Garfinkel H 1984 *Studies in Ethnomethodology*. Polity Press, Oxford, UK

Glassner B, Hertz R (eds.) 1999 *Qualitative Sociology as Everyday Life*. Sage, Thousand Oaks, CA

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY

Gullestad M 1991 The transformation of the notion of everyday life. *American Ethnologist* **18**(3): 480–99

Kaufmann J-C 1998 *Dirty Linen. Couples as Seen Through Their Laundry*. Middlesex University Press, Middlesex, UK

Lefebvre H 1971 *Everyday Life in the Modern World*. Harper & Row, New York

Löfgren O 1997 Scenes from a troubled marriage: Swedish ethnology and material culture studies. *Journal of Material Culture* **2**(1): 95–113

Schutz A 1973 *Structures of the Life-World*. Northwestern University Press, Evanston, IL

Shields R 1999 *Lefebvre, Love and Struggle: Spatial Dialectics*. Routledge, London

Willis P 1990 *Common Culture*. Westview Press, Boulder, CO

O. Löfgren

Evidentials in Linguistics

The term 'evidentials,' introduced originally in descriptions of languages where suffixes express the source of subjective knowledge—as by hearsay, or by

dream—was extended progressively to a repertoire of grammatical devices coding different kinds of evidence on which statements are based: (a) markers indicating the source of the speaker's information or knowledge, and (b) epistemic markers coding the speaker's attitude toward knowledge. In this broad sense, evidentials qualify the source in different ways (hearsay, inference, appearance) and specify the truth value of an utterance (Chafe and Nichols 1986). Expanding the notion of epistemic modality to the status of the speaker's knowledge (Palmer 1986, p. 51), evidentials cover most of what linguists treat now under the notion of evidentiality: a strict grammatical distinction between direct (typically 'unmarked') and indirect evidence, the latter being based on verbal report and inference (Willett 1988, p. 57). However, this broader conception of evidentiality is not without problems.

All languages can express the speaker's (or the writer's) various modes of access to information, as well as the degree of the speaker's responsibility for that information. In most of the Indo-European languages, information about the source of the conveyed content is not grammaticized, but may be provided by various optional devices, either syntactic ('They say that'…, 'I heard that'…) or lexical (adverbs: 'allegedly,' 'reportedly'…; or particles such as Bulgarian *kaj*, *už* …). Many languages do, however, encode in their grammars specific forms for expressing direct experience, reportedness, hearsay, inference, etc. For instance, some Balkan and Middle Eastern languages provide paradigms (mediative forms derived from the perfect), which refer to the information source or to the speaker's attitude toward the information, without specifying explicitly its source or commenting on its truth. These forms are opposed to the indicative forms, which are neutral with respect to the source. In these languages, the speaker's choice of the form is subordinate to the register (colloquial dialogue, historical narrative, distantiated narrative, etc.), so that the specific interpretation of the information mediated through report or inference depends on the register (Guentchéva 1996b).

In other languages the speaker must specify the way in which the information was acquired, by choosing a grammatical marker indicating whether the information is based on direct experience, or on reports or inference (something seen, told, heard, inferred, including dreams, tales, and myths). In Tibetan (Tournadre 1996), special markers serve to indicate that the situation has been seen (*-song*), is common knowledge or has been heard from someone else (*-pare'*), or is inferred (*-sha'*):

(1) mätrokungkar-la sāyom-ki'
Metrokungkar-LOC earthquake-ERG
khābgpa Mangpo
house many
Lo-*'song*/*-pare'*/*-sha'*
destroy-AUX
'In Metrokungkar, the earthquake destroyed many houses'

Moreover, the hearsay morpheme *-sa* may combine with any of these verb forms, to indicate the source of the speaker's informant's information. The system is even more complex, since these markers interact closely with person and with the semantics of lexical verbs.

The three major issues are: (a) are there pure evidential systems? (b) what is the core meaning of evidentiality? and (c) what is the relation of evidentiality to modality?

Tuyuca (Brazil and Columbia) is often presented as a 'classical evidential system.' If a Tuyuca speaker wishes to say 'he played soccer,' he must choose an appropriate suffix (Barnes 1984, p. 256):

(2a) díiga apé-**wi** (I saw him play)
(2b) díiga apé-**ti** (I heard the game and him, but I didn't see it or him)
(2c) díiga apé-**yi** (I have seen evidence that he played: his distinctive shoe print)
(2d) díiga apé-**yigi** (I obtained the information from someone else)
(2e) díiga apé-**hĩyi** (It is reasonable to assume that he did)

At first sight, this system appears to be based on the contrast between information obtained through direct sensory experience (visual in (2a) and nonvisual in (2b)), and information obtained through indirect experience (apparent, inferred in (2c), secondhand report in (2d)). The assumed evidential in (2e) seems to hover on the boundary between direct and indirect evidence, being used only when no information about the state or event is being or has been received. Furthermore, this analysis does not take into account the fact that the visual *-wi* of (2a) can co-occur with the 'resultative' morpheme *-ri*: *wáarigi nū-wi*. While this construction may be translated as 'he went,' the literal gloss is 'he was here earlier, and later I couldn't find him.' This is very similar to the inferential meaning of the perfect in Western languages. Clearly, this kind of inference should be differentiated from that in (2c).

Wintu provides an instructive example of another kind. According to Schlichter (1986, p. 54), Wintu has an evidential deictic system establishing a strict grammatical distinction between a direct 'visual experience' (unmarked) and an indirect experience that is 'an alternative to the temporal orientation of "Western" civilization.' It has four evidentials:

(3a) heket wira waca:-bi-**nt**^{h}**e**·-*m*
someone come cry-Imperfective-NONVISUAL SENSORIAL-DUB = 3 person subject
'Someone is coming crying (I hear)'

(3b) Č'epkal neɪ ba·-bi-**nt**^{h}i-da
bad we eat-Imperfective-SENSE-we
'We've been eating bad things (I sense)'

(4) Minel kir-**ke**·-m die COMPLETIVE
ASP-HEARSAY–DUB = 3 person subject
'He has died. (I'm told)'

(5) Ničaj ewin suke-**re**·
nephew here stand-INFERENTIAL
'My nephew must have been here (see tracks)'
(6) *?lmtop· nuqu·-·el*
berries ripe-EXPECTATIONAL
'The berries must be ripe (it's that time of year)'

The non-visual *-nt^{h}er* in (3), deriving from a passive form of the perception verb *mut-* 'hear, feel, sense, perceive,' involves a fact known to the speaker through a sense other than vision, or by any kind of intellectual experience or 'sixth sense.' It interacts closely with person, agentivity, and aspect. Occurring most often in third person, as in (3a), it only expresses the speaker's perception of the situation; but when occurring in first person (the subject marker *-da* is required), it involves the speaker's experiential role, as in (3b). Furthermore, it is always in complementary distribution with the derivational suffix *-a*, which indicates that 'the grammatical subject is a voluntary participant or agent of the situation.' It seems also to have developed a surprisal or 'mirative' meaning. The suffixes *-ke·* and *-re·* are true evidentials, indicating, respectively, information known through hearsay (4), and information based on inference (5). Finally, the suffix *-ʔel* in (6) is usually used to specify the speaker's 'experience with similar situations, regular patterns, or repeated circumstances common in human life.' It thus has an 'experiential' meaning, though it can also function as a hearsay marker in myths.

The comparison between Tuyuca and Wintu shows that their evidential systems do not cover the same semantic space. It demonstrates clearly that the languages grammaticize different distinctions, and that the diverse processes in the grammatical categorization of each language are based on different conceptualizations.

Commenting on the search for a core meaning, Anderson (1986, p. 274) points out that not all forms that have been called 'evidential' share a common set of 'evidential functions.' He proposes four parameters to identify a true evidential: (a) it shows 'the kind of justification for a factual claim which is available to the person making that claim', (b) it is not itself the main predication, but rather 'a specification added to a factual claim about something else', (c) the indication of evidence is its primary meaning, not just a pragmatic implication, and (d) it never occurs as, a derivational morpheme nor as an element in a compound. Anderson's model and other recent cross-linguistic analyses show that the basic principle governing the unity of evidentiality in a particular language and across languages is the widespread occurrence of mediated information. Evidentiality is the grammatical manifestation of a two-dimensional operation binding simultaneously the event's own spatial and temporal coordinates (event's time reference) and the speaker's spatial and temporal reference (speech time reference). This is not to say that the two time references co-occur, but that the binding of the two reflect a scalar mapping of temporal relations which allows the speaker to express different degrees of disengagement with respect to the content of the utterance (Déscles and Guentchéva 2000).

From a typological point of view, the semantic space of evidentiality is articulated according to two fundamental meanings: inferential and reportive (including an unspecified third person, hearsay, rumor, common knowledge, tales, and myths). Cross-linguistic data suggest that attested evidential systems are based mainly on a binary opposition which distinguishes mediated (and morphologically overt expression) from nonmediated information (most often carried by the non-overt member of the morphological opposition). When a language has only one evidential marker, its main function is the reportive one. If the one marker covers both reportive and inferring information, its interpretation is left to the context. When a language has two markers, they may distinguish mediated and nonmediated information (Turkish, for instance), or reportive and inferring information.

Reportives are found in a wide variety of languages. In some, a true reportive evidential is derived from a verb of saying and is used in dialogue as well as in tales and myths. In Makah (Jacobsen 1986, p. 17), the reconstructed passive form *-wa:t* of the verb *wa·* 'say' indicates hearsay and this also occurs in the narration of tales and myths:

(7a) *xu·bitdib$^{?}$**u***
'He was snoring'
(7b) *xu·bitdibit**wa·d***
'He was snoring (I was told)'

Clearly, (7b) is a mediated message. The denoted situation is part of the original utterance attributed to someone other than the speaker; i.e., it combines the term reference of the original utterance with the speaker's time reference. This is not to say that *wa:**t*** always implies an antecedent speech act of another speaker; it may merely be a mental construct of overhearing, common knowledge, tales, etc.

Although the relation between hearsay and indirect discourse is very close, they are independent. Indirect discourse usually involves overt expression of *verba dicendi*, and deictic transposition which 'shifts' person, tense, and mood. Alternatively, it may be a special mophosyntactic device, as in Maricopa (Gordon 1986, p. 86). Cliticized to the invariant form verb *'ish* or *ish* 'say,' *-'a* in (8a) conveys heard information in an independent clause, while in (8b), *-k* introduces 'a complement whose verb is not marked with switch-reference suffixes' and indirect discourse interpretation:

(8a) *Bonnie-sh chuy-k-'ish-'a*
Bonnie-Sj marry-**k**-say + sh-hearsay
'(They said, I hear tell) Bonnie got married'
(8b) *Bonnie-sh chuy-k uu'ish-k*
Bonnie-Sj marry-**k** say + Pl-Asp
'They say Bonnie got married'

As noted earlier, in some languages, the overt inferential markers arise from perfects, while in others they arise from perception verbs. In the latter case, the message does not focus on the source of the evidence but, rather, on the fact that the evidence is not reported verbally. Occurring in the utterance, an inferential evidential indicates that the factual content of the message is mediated by sensory perception, but also by experience or intuition. Clearly, the speaker may construct an inferential reasoning about a supposedly produced event.

Wintu provides two grammatically different inferential markers which reflect inferential reasoning: *-re* in (5) and *-ʔel* in (6). With *-re* the speaker indicates that an earlier plausible event is inferred from traces in evidence. The grammar thus encodes an abductive reasoning, which involves (Peirce 1974, Vol. II, p. 636; Vol. V, p. 171): (a) a law allowing an explanatory inference of hypothesis based on observed facts, and (b) the observed facts that are traces sufficient to relate to the hypothesis, which is merely taken to be plausible. With *-ʔel* the speaker indicates that it is inferred from experience 'the existence of phenomena such as we have observed in cases which are similar' (Peirce 1974, Vol. II, p. 636); the grammar encodes an inductive reasoning.

The (ad)mirative meaning of evidential constructions (expressing surprise or irony) is an extension of the reportive or the inferential. It arises when the speaker confronts an unexpected discovery and expresses the contrast between the expected and discovered situation ('I didn't know that S (but now I do)').

While evidentiality has often been identified with epistemic modality, analyses of data from different languages show that neither the concept of 'truth' nor the strength of a speaker's commitment to his or her statement's truth value fully accounts for evidentials and evidentiality. The translation of evidential constructions with modals (English 'must,' French *devoir*) in European languages is not always an adequate one (Chvany 1999, p. 437). Moreover, in some languages modal verbs appear in mediative forms.

Among recent attempts to assign a uniform treatment of evidentiality, an account in terms of a dual time-reference (to speech time and event time) appears to be the most promising and relatively straightforward. Interpretation may, however, vary according to gender or stylistic register. The speaker's disclaimer of responsibility is organized along a continuum of degrees of disengagement with respect to the content of the utterance.

See also: Inferences in Discourse, Psychology of; Knowledge, Anthropology of; Tense, Aspect, and Mood, Linguistics of; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Anderson L 1986 Evidentials, paths of change, and mental maps: Typologically regular asymmetries. In: Chafe W, Nichols J (eds.) *Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes*. Ablex, Norwood, NJ, Vol. XX, pp. 273–312

Barnes J 1984 Evidentials in the Tuyuca verb. *International Journal of American Linguistics* **50**: 255–71

Chafe W, Nichols J (eds.) 1986 *Evidentiality: The Linguistic Coding of Epistemology*, Advances in Discourse Processes. Ablex, Norwood, NJ, Vol. XX

Chvany C 1999 Book review of Z. Guentchéva (ed.). *Journal of Pragmatics* **31**: 435–39

Déscles J-P, Guentchéva Z 2000 Enonciateur, locuteur et médiateur. In: Ericson Ph, Monod-Becquelin A (eds.) *Rituels du Dialogue*. Université de Paris-Nanterre, Paris

Gordon L 1986 The development of evidentials in Maricopa. In: Chafe W, Nichols J (eds.) *Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes*. Ablex, Norwood, NJ, Vol. XX, pp. 75–88

Guentchéva Z (ed.) 1996a *L'énonciation médiatisée*. Peeters, Louvain, Belgium

Guentchéva Z 1996b Le médiatif en bulgare. In: Guentchéva Z (ed.) *L'énonciation médiatisée*, Peeters, Louvain, Belgium, pp. 47–70

Jacobsen W H 1986 The heterogeneity of evidentials in Makah. In: Chafe W, Nichols J (eds.) *Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes*. Ablex, Norwood, NJ, Vol. XX, pp. 3–23

Palmer R F 1986 *Mood and Modality*. Cambridge University Press, Cambridge, UK

Peirce C 1974 *Collected Papers*, Vols. I–VI. Belknap Press, Cambridge, MA

Schlichter A 1986 The origins and deictic nature of Wintu evidentials. In: Chafe W, Nichols J (eds.) *Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes*. Ablex, Norwood, NJ, Vol. XX, pp. 46–60

Tournadre N 1996 Comparaison des systèmes médiatifs de quatre dialectes tibétains (tibétain central, ladakhi, dzongkha et amdo). In: Guentchéva Z (ed.) *L'énonciation médiatisée*. Peeters, Louvain, Belgium, pp. 195–213

Willett T 1988 A cross-linguistic survey of the grammatization of evidentiality. *Studies in Language* **12**: 51–97

Z. Guentchéva

Evolution and Language: Overview

The complex relationship between evolution as a general theory and language is discussed here from two points of view. The first concerns the isomorphism of the basic structure of evolutionary theory in biology and linguistics. The many languages in the world fall into coherent groups of successively deeper level and wider membership, e.g., the Romance languages and the Germanic languages at a more superficial level are both themselves subgroups of the wider Indo-European language family. Similarly in biology groupings

like the orders of primate, the Felidae (catlike animals), and others are subgroups of the larger class of mammals. The existence of such hierarchies in both fields is accounted for by a similar dynamic process of change and differentiation through time. The second major topic is the unique and basic role of language in the evolution of communication.

1. *The Parallel between Biological and Linguistic Evolution*

The essential resemblance, or homology of structure, between biology and linguistics is based on the equating of languages to species and dialects to varieties. In both fields somewhat similar criteria are employed to distinguish these levels and in both cases they have a penumbra of vagueness. In biology membership in the same species is defined by the possibility of fertile interbreeding. In linguistics, a language is defined by the test of mutual intelligibility. The vagueness in both cases derives from the fact that we are striving to capture by definitions the distinctions of two levels, species and variety, language and dialect, when in fact in both cases we are dealing with a dynamic process. The variety of today is the species of tomorrow and the dialect of today is the language of tomorrow.

1.1 *Transformationism Versus Creationism*

The transformational meaning of evolution arises from the considerations involved in any class of phenomena in which the investigator is confronted by the existence of a variety of kinds or species. In such instances there are two fundamentally different types of theory to account for the existence of distinct kinds. According to the creationist view the distinctness of species is explained as issuing from distinctive creative acts. Species can be created or destroyed but they cannot give rise to other and new species since this would change their essential characteristics. When confronted with fossil evidence that the species of different geological epochs were, in general, similar although distinct, the creationist necessarily denied that any lines of development connected them. The theory thus adopted was called 'catastrophism,' namely that by a succession of cataclysms earlier species had been destroyed and were replaced by more recent ones through new creative acts.

The evolutionary theory of transformation of species on the contrary maintained that there are no fixed bounds to variation. Therefore earlier species give rise to later ones by the process of developmental evolutionary change. The fact that in life forms species fall into coherent larger groupings, the genera, and that groups of genera then form distinct families and so on in an ordered hierarchy receives a distinctive transformationist explanation.

According to transformationist theory, those species that belong to the same genus are the differentiated descendants of a single ancestral species, their resemblance being explained by common descent. The resemblances among genera that belong to the same family are explained in turn by the theory that the species ancestral to each genus are descendants of a still earlier form ancestral to the whole family, and similarly with larger and taxonomically higher groupings. Fossil forms are then either ancestors of existing groupings or additional lines of descent that have become extinct without leaving descendants.

The model of evolution that emerges is that of a branching tree. The varieties of today are like twigs that, as they sprout, become the species of tomorrow. The theory of evolution as transformation applies both to linguistic and biological evolution. This resemblance was noted by Darwin himself in *The Descent of Man* (1871, p. 40) and had indeed been noted earlier by Lyell (1863), the founder of modern geology. As Darwin states: 'the formation of different languages and of distinct species and that both have been developed through a gradual process is curiously parallel.'

In linguistics the creationist view is represented by the Biblical account of the Tower of Babel, according to which all language differences were created at the same time by the confusion of tongues. By not noting that languages fall into coherent groups like the Germanic languages etc., and that these in turn function as units within higher level units in an ordered hierarchy, the Biblical account which reigned supreme until nearly the end of the eighteenth century was inferior to its creationist counterpart in Linnaean biology, which did note the coherence in biological classification but simply ascribed this coherence to a divine plan to produce such groupings.

Perhaps because the Babel theory more egregiously failed to account for such groupings, the creationist account was superseded earlier in linguistics than biology. The event that marks most clearly the triumph of transformationism over creationism in linguistics was the recognition that the resemblance among the languages we now call Indo-European is to be explained as a result of common inheritance with differential changes from an extinct ancestral form of speech which in later terminology was called Proto-Indo-European. The acceptance of this theory is usually but somewhat arbitrarily dated from a statement of Sir William James, an eminent British official in India, in 1786. In fact, in regard to the Finno-Ugric languages this explanation had been advanced and accepted some years earlier. Thus the recognition of transformation in linguistics substantially antedates its acceptance in biology.

Probably the chief factors in the relatively earlier acceptance of this type of explanation in linguistics as

compared to biology is the vastly more rapid rate of change in language. The commonsense objection that actual changes in species had never been observed in historic times and that, to all appearances, they were fixed appears to have been the most powerful factor in the rejection of transformationism by biologists in the period preceding Darwin. Darwin in turn appealed to languages as an exemplification of the reality of the process of transformation he proposed in regard to life forms (cf. Alter 1999).

Language within the realm of individual experience and unaided by records from an earlier period seems fixed on the whole, but not to the same degree. Older people can recall vocabulary items that were current in their youth but are now no longer heard.

These changes, apparently small within the lifetime of a single person, display powerful cumulative effects in periods of time within the scope of human records. Thus Anglo-Saxon (Old English) exhibits differences from modern English over a period of about 1,200 years, comparable to that between English and German at the present time. If Anglo-Saxon and modern English were spoken contemporaneously they would undoubtedly rank as separate languages. It was possible then to note in this historically well-documented instance the change of one language into another. Moreover there are earlier records of German also in the form of Old High German and Old Low German. These languages are both far more similar to Anglo-Saxon than modern English is to contemporary German. Thus one could actually observe a convergence as one went back in time, the mirror image as it were of the hypothetical divergence in time from a single common ancestor.

In general, languages ancestral to existing groups of genetically related languages, which would correspond to fossils in biology, take us back to periods before the existence of written records. However, in the well-known instance of Latin, there are abundant records of a language which through a series of locally different variants has given rise to the existing diversity of Romance local dialects and standard languages. In view of these well-known facts, it was not too audacious a step to assume that in similar fashion an ancestral language had once existed bearing the relation to the existing branches of Indo-European that Latin does to Romance speech forms. It thus becomes a mere historical accident that Latin survived in written form, whereas Proto-Indo-European, which was spoken before the existence of writing, could not be known through such evidence.

The transmission of physical characteristics by the genetic mechanism corresponds to the transmission of language by learning from one generation to the next. In both cases variations arise, some of which are perpetuated. In both instances geographical isolation whether complete or imperfect is the usual factor in the perpetuation of different variants, although linguistic isolation of social groups in the same territory also occurs. Difficulties of determining where a variety ends and species begins, difficulties that were important factors in Darwin's disillusionment with creationist theory, resemble the linguistic difficulty in defining language as against dialect. Ultimately these variant descendants become sufficiently distinct as to be indisputably ranked as separate languages or species.

The status of transformationalist explanations of linguistic similarities was further enhanced by the largely successful reconstruction of Proto-Indo-European where, as is usually the case, the ancestral language was not known from written records. This enterprise received a great impetus through the discovery that in certain respects, especially the sound system, changes were not haphazard but rather showed a surprising degree of regularity.

The enterprise of reconstructing Proto-Indo-European and the laying of the foundations of the comparative method that underlies it was essentially carried out by the so-called Neogrammarian School of the 1870s and the 1880s.

1.2 Cladistics in Linguistics and Biology

In the course of laying out more explicit methods of historical comparison and reconstruction the Neogrammarian School developed the notion that only those changes which were innovations, not those which were retentions, were relevant in classification. The period in which what marked one subgroup as a development distinct from that of another was defined by the different set of changes that each was undergoing from a common original. Retentions were simply inherited features that persisted from the common period.

This was indeed the very point made by Willi Hennig (1966) the founder of cladistics in modern biology, some 70 years later. However, in linguistics as in biology, it is not easy in many instances to distinguish innovations from retentions. For example, for a long period it was held that the Italic and Celtic branches of Indo-European shared an important innovation that helped to define a separate Italo-Celtic branch of Indo-European, namely the formation of a passive voice in -r. However, when Tokharian in Chinese Turkestan was discovered from documents as a new branch of Indo-European in the 1930s, it was found to have a passive in -r but hardly anything that would otherwise affiliate it with Italic or Celtic.

1.3 Analogy and Homology in Linguistic and Biological Evolution

Just as transformationism is a dynamic interpretation in time of the hierarchy based on differences and

shared similarities in languages, so the creationist view of fixed species is implied in the use of properties for classification with typological rather than genetic relevance. Once again there is a parallelism between biology and linguistics. In the terminology of biology, we wish to distinguish between homologies that are similarities which are the outcome of common descent, and analogies that result from convergence usually through similarity of function and are irrelevant to historical classification.

In language it is possible to define two kinds of criteria that define the difference between homologies and analogies. The former are those that involve sound and meaning simultaneously, while the latter concern sound in abstraction from meaning or meaning in abstraction from sound. Thus the resemblance between English *nose* and German *Nase* involves both sound and meaning. If supported by a reasonably large number of other resemblances of the same kind, these forms are then said to be 'cognates,' i.e., related. Such cognates are the basis for drawing the conclusion of a common origin of English and German.

An example of resemblances of sound without meaning is the existence of tonal systems, i.e., systems in which pitch difference distinguishes semantically different meaningful elements if we simply take these systems as such abstracting the sounds from the meanings of the words being compared. An example of meaning without sound is the existence of sex gender in two different languages where we disregard the sounds by which the differences between masculine and feminine are expressed.

The latter criteria, sound without meaning and meaning without sound are called typological by linguists; these criteria define types of languages e.g., tonal vs. non-tonal, gender vs. non-gender. The use of typological criteria for genetic linguistic classification requires the assumption that certain features like the essential as opposed to the accidental features of the creationists are fixed and define the species; or in this instance the language family. Thus if we state that tonalism is a requirement for membership in a particular linguistic stock we are assuming that a tonal language can never become a non-tonal language. If this were true monogenesis would be refuted. We would initially have two or more languages, each of which would have tonality or non-tonality as an essential attribute and would be incapable of either acquiring tonality where it was lacking or losing it where it was present. This is the analogue of the notion of fixed species.

Once it is realized that a classification based on gender or non-gender is similar in principle to a biological classification into flying and non-flying animals, we see that a Semitic language which ceases to have triconsonantal roots does not cease to be a Semitic language by descent, any more than a bat ceased to be a mammal when it acquired the property of flight.

1.4 Differences between Linguistic and Biological Evolution

Thus far the discussion has stressed the basic parallel between linguistics and biological evolution. However, there are significant differences, mostly in the contrast between genetic and social transmission. There is not only lineal transmission of languages to one's biological offspring but also lateral transmission in that speakers are capable of bilingual, or indeed multilingual behavior. Thus a child may acquire a language that is not the native language of its parents. Moreover, unlike the genome which is fixed at birth, speakers may borrow words and grammatical forms which are part of the inherited language of another group of speakers with whom they have come in contact, and which may in turn be transmitted to their own offspring as part of their native language acquired from their parents. The closest analogue to this in biological evolution is hybridization. In the animal kingdom hybrids are usually infertile but in the plant kingdom this is not invariably so.

1.5 Language, Natural Selection, and the Notion of Progress in Language

The theory of evolution in its Darwinian form not only espoused transformationist as against the creationist notion of fixed species, but also suggested as its main mechanism natural selection in order to provide a plausible account of the transformation of species. The concept of the transformation of species was, of course, not new. Lamarck, for one, had already espoused it at least 50 years before Darwin.

Now as Darwin states repeatedly, the descendants of a particular life form are, through the agency of natural selection, likely to be more efficient though this of course is relative to a particular environment. There is, however, still implicit in the notion of natural selection the theory of long-term advance in the series of life forms.

It is true, moreover, that the paleontological evidence as we go back in time becomes increasingly confined to forms of relatively simple and undifferentiated structure and of a more limited range of adaption to the environment. It was precisely one of the recommendations of the theory of natural selection that it helped to make this temporal advance intelligible.

This aspect of evolutionary theory has also been applied to language. The new linguistic forms that arise continually in any linguistic community may then be likened to variations, only the fittest of which survive and become incorporated in the language. Lyell (1863) in his elaborate comparison of language to biological species, noted also that the spread of specific languages over wide areas resembles the

biological processes of adaptive radiation and the extinction of species. Natural selection as applied to language would then involve an intralinguistic struggle for survival among forms within the language and an interlinguistic struggle among languages in which some spread and produce descendants while others perish.

Regarding the intralinguistic struggle it appears that many changes are functionally indifferent, and that while certain changes make for greater efficiency there are certain outcomes of normal processes of change, such as grammatical irregularities and semantic ambiguities, which are functionally negative. Taking linguistic change as a whole there seems to be no discernible movement towards greater efficiency such as might be expected if in fact there were a continuous struggle in which superior linguistic innovations won out as a general rule.

Similarly it can be seen that one language succeeds another not because it is more advanced as a language, but for extralinguistic reasons of the military, economic, or cultural superiority of its speakers. This is not to say that there are not important differences between languages that have been the object of literary cultivation and those that have not, and those that possess an extensive technical vocabulary and those which do not. However, such differences are but reflections of non-linguistic differences. They affect nothing basic in the language itself. Any language is capable of literary expression or technical expansion if non-linguistic circumstances encourage or require it.

The question that has just been discussed relates to another way in which languages of the present do not conform to one of the basic aspects of evolutionary theory in biology. A theory which regarded all species as interconnected, but which posited some mammalian form as the primitive ancestral type for all life whence descended in one line all the other vertebrates and in another the ancestor of all non-vertebrate phyla, and in which plants and Protozoa first appeared in the most recent period, would be a strange evolutionary theory. In addition to the notion of transformation of species that of progress or advance of some kind is inherent in the notion of natural selection. In general we are dealing with the increase in the range and variety of adjustment to the environment. Among the developments that may be considered advances there are, on the perceptual side the ability to respond to finer differences in stimuli, to stimuli from a greater distance, and to new ranges of stimuli, e.g., a new sense. On the motor, or effector side there are the ability to live in a greater range of temperature or moisture, speed of movement, and the ability to make finer manipulatory adjustments to objects in the environment. In the interface between perception and response come the coordination of responses to stimuli such as the development of a central nervous system, of social cooperation and intelligence in general, and finally speech itself.

This gives us many facets of comparison and we cannot simply arrange all species on a linear scale of more advanced and less advanced. It is clear that time, indeed vast periods of time are required for the development of more advanced species in the sense just sketched. This is borne out in general by the paleontological record which shows that on the whole more advanced species appear later than less advanced although the less advanced persist as viable adaptions.

In the nineteenth century and the earlier part of the twentieth century the belief persisted that languages could in a similar way be arranged from least to more advanced. This was generally expressed by the notion of successive types. The most prominent and persistent of these schemes was the threefold division into isolating, agglutination, and inflective. The earliest and most rudimentary type was the isolating in which, ideally, each word was completely simple and could not be divided into parts as do many words in English, e.g., truth-ful-ly. In the agglutinative stage formerly isolated meaningful units were merely glued together mechanically without any modification of their parts. In the final stage, inflection, there were modification and adaptations of the parts of a word so that the word was a complex organic structure.

Generally only Indo-European was considered to embody the highest stage. The ethnocentrism of this scale is obvious and there were such anomalies as the fact that the closest language to an ideal isolating one was Chinese with its ancient literature and civilization. Further, certain American Indian languages turned out to have a more complex word structure than Indo-European. Moreover in historic times Indo-European languages seemed to be losing their complex inflectional structure and to approach more and more an isolating type, a retrograde movement from the viewpoint of this theory. These considerations led ultimately to its abandonment.

As a further criticism it may be noted that only one very limited aspect of languages was being considered, namely how words were formed from smaller meaningful units, the aspect of language called morphology. In fact it is not clear in what ways an isolating structure is any way less adapted to the functions of language than an agglutinative or inflective. It can also be seen that in this now abandoned traditional treatment the subject of evolution is almost solely in the context of simplicity evolving into complexity and limited to the morphological structure of words. The significance of simplicity versus complexity has certainly been overrated. Irregular alternations, a supposedly glorious property of the inflectional type, have no evolutionary superiority. Irregular alternations are, by definition, functionless. The variation between 'go' and 'wen-' is useless since the difference in meaning between present and past is already adequately expressed by the '-t' of the past in 'wen-t.' A past 'go-ed' would perform the same work and without the burden of learning this irregularity–a real, if hardly noticed,

difficulty for the native speaker and a more conspicuous one for an individual who learns English as a second language. That this is a point of inefficiency is witnessed by the universal tendency towards analogic change, that is, the imposition of regularity. This is hindered by sheer frequency of use that tends to perpetuate irregularities in the most frequent words as in the inflection of 'to be' and the irregular plurals 'feet,' 'teeth,' etc.

While the analogic process mentioned above tends toward regularity, universal processes of linguistic change produce new irregularities. The irregular plurals 'feet,' 'teeth,' 'mice,' etc., came into existence from earlier regular forms. In Anglo-Saxon we had *fot* 'foot' and *fot-i* 'feet.' The first vowel of the plural was fronted by the influence of the plural '-i' expressed by a front vowel ('i' as in 'pit') producing a vowel in the first syllable similar to 'ö' of present day German, i.e., *föti*. This vowel was unusual in English and the rounding was redundant, as the plural was already expressed by the '-i' since the singular was *fot*. Possibly in consequence of this the front rounded vowel was unrounded and the final vowel dropped. Thus we finally had singular *fot*, plural *fet* giving modern English 'foot/feet' which generations of children venturing on an analogical plural 'foots' have been unable to dislodge.

Recapitulating in terms of the earlier consideration of evolutionary advance, it becomes clear that it is not complexity as such which is more advanced than simplicity but any change in efficiency relative to a function to be performed. The traditional criterion for simplicity in the way words are put together from smaller meaningful parts (morphology) is here of minor significance. Similarly in biology a hoof is more efficient than toes in relation to locomotion but a hoof is simpler in form than toes. In other instances we have growth of complexity as functionally relevant, e.g., in the differentiation of teeth in relation to grinding (molars), cutting (incisors), etc.

2. *Language and Communication*

The basic function of language is communication. This implies that language should be considered in the more inclusive framework of the evolution of communication. To discuss the evolution of language by considering language in isolation from communication in general is like discussing the evolution of the bow without regard to its relation to other weapons.

2.1 *The Three Stages of Language Evolution*

The evolution of language from the viewpoint of communication can be divided into three stages: pre-language, language proper, and post-language. Although this is the chronological order of their first appearance, each continues to exist after the initiation of the next stage.

2.2 *The Age and Definition of Language Proper*

The most commonly held view is that true language first developed in our own species *homo sapiens sapiens* ca. 50,000 years BP. No doubt non-language communication existed previously as it does in many species both in the genus *homo* and outside of it. There is no satisfactory account, conjectural as it must be, of steps leading from pre-linguistic communication to language proper. For many the problem of the origin of language is the development of such an account.

We can however consider what distinguishes language proper from non-language, which we may call pre-language in this context. This sets up the requirements of what must come into existence before we can talk of language as distinct from pre-language. The two major requirements are duality of patterning and an infinity of possible sentences. The term duality, (first used by Hockett 1958), refers to the existence of two levels in language. The first is the level of sound units (phonemes) which are relatively limited in number, roughly 10 to 70, and which are in themselves and in isolation meaningless. In combination they form meaningful units of various levels from the smallest, the morpheme e.g., 'hand,' 'the,' 'streng-' as in 'strength' through words, simple or complex morphemically 'hand,' 'the,' 'streng-th' and phrases, up to the sentence level. The sentence is a unit that can occur in isolation to express a coherent thought, e.g., *you* 'forgot to put out the cat.' This analyzability of complex sequences of sound to produce sentences as expressions of coherent thought can be called grammaticality, and the rules when codified form what is called grammar. As a second major property that distinguishes true language from pre-language there is the infinitude of grammatically well-formed sentences that can be formed in conformity to the rules of grammar. There is no finite length for a possible sentence.

Pre-language is then not true language because it lacks these properties of duality and infinity. Pre-language continues to function in human societies as gesture and expressive cries although the latter are often conventionalized, e.g., 'ouch!' The advantage brought by grammatical rules which organize the finite set of distinctive sounds or phonemes is not only a vast increase in the number of lexical elements constituting the vocabulary but the possibility of combining in sentences categories not found together in reality, e.g., lies, hypotheses, and past or future states of affairs.

It seems unlikely that genuine language developed from gestural elements. In addition to the values inherent in a grammatical organization on the dual level of phonemes and morphemes, certain advantages

of the sound medium will help explain why the first true language is spoken language. The use of vocal organs, an overlaid function, did not require the development of a new organ through the slow mechanism of genetic change. The voice is always available, involves little exertion, and does not interfere with other activities except, to a minor degree, eating. Above all, it allows the hand to be utilized. Moreover it can be used by day or night and is perceptible in all directions.

The development of brain lateralization and an increase in complexity and intelligence in the brain are also a major concomitant. Probably this should be viewed as reciprocal. The development of the brain made language possible but the development of the brain itself was to a great extent the result of the use of language.

2.3 Post-language

Despite all of its advantages language in its physical aspect lacks permanence and range. The first advance in what we may call post-language is the invention of writing ca. 10,000 BP in the Middle East. It gives permanence to speech and overcomes distance by the sending of written missives. Relatively recent inventions such as telegraphy, radio, and the computer all give greater range and the possibility of channelized and hence more private communication. All that these developments have in common is that they are isomorphic with spoken language at the sentence level and often at lower levels such as the word or the morpheme. By this is meant that there are rules of one-to-one correspondence between entire sentences in spoken language and some sequence of written or other symbols without necessary one-to-one correspondence on the lowest level of basic phonologic elements (phonemes). Because of the isomorphism on the sentence level of all these systems with spoken language, any inefficiencies that adhere to the semantic and grammatical systems of spoken language are unaffected.

In its semantic aspect certain disadvantages rise from its method of definition of semantic units. This is because the meaning is learned normally not by definition but in context and implicitly rather than explicitly. The two main properties of meaning in natural languages that arise are ambiguity and vagueness. By ambiguity is meant the existence of alternative and different meaning, i.e., homonymy. Vagueness is the lack of definite boundaries to the set of cases in which a term applies. Imagine that the speakers of English are confronted with a man without a single hair on his head. They will presumably agree that he is bald. Now take a man with a full head of hair and remove the hairs one by one. There will be lack of agreement as to the point at which the statement that he is bald is true and, no doubt, an interval within which some or all speakers will be uncertain. We might agree to define the man as bald if he has fewer than a certain number of active hair follicles. In everyday discourse many terms have an ambiguous set of meanings each of which is vague. Actually much ambiguity is harmless. For example the use of the term 'case' both in grammar and law will presumably mislead no one and such instances are even a conservation of vocabulary resources. Vagueness can in principle be eliminated by an agreement on boundaries. However the absence of boundaries corresponds to our own lack of precise knowledge. Do we wish to count active hair follicles every time that we are about to employ the word 'bald'?

Far more insidious are the closely similar but distinct meanings of such words as 'function' in the social and biological sciences. The needs of philosophical and scientific discourse cannot always in the long run be satisfied by the traditional implicit definitions. The first step in departing from the traditional use of terms is the use of definitions by postulation but employing the semantic and grammatical structure of natural language. The physicist defines 'force' in the professional discourse of physicists. Sometimes, however, such half-measures prove inadequate and, as in the case of mathematics and formal logic, an entire system with its own syntactic rules and symbolization of terms is created by postulation.

The development in technologically advanced societies of such formalized systems that overcome in good part the ambiguity and vagueness of meaning and the morphological irregularities of natural language does not mean the supersession of language. It may rather be interpreted within the overall context of the evolution of communication, as part of the process by which each kind of communication becomes more efficiently served by an instrument that more adequately fulfills some specific function. The constant availability and flexibility of language in expanding to cover new needs suggests that it will not be replaced in person-to-person interaction in the foreseeable future, if ever.

2.4 Ordinary Language as the Ultimate Level of Explanation

Language plays a unique role among means of communication that aside from all other considerations assures its future. If we wish to explain a symbol we must do it in some sign system. If the term is still not understood it must be explained in some lower-level system, and so on. But at some point this process must reach an end. Such a system provides the level of ultimate explanation. Language serves this function, hence its generality as compared with the limited subject matter of other signs. For example one explains the symbolic use by the Navajo of 'red' as referring to 'north' by means of a word for 'north' in Navajo or

some other language and it is then assumed to be understood.

Ontogenetically, too, such systems or post-language symbols (e.g., traffic lights) are learned after language and by the use of language, but never vice versa. We would not entrust a pre-language infant with crossing the street alone. Humankind is a symbolic animal who constantly devises new post-language signs. For example, without explicit argument or legal enactment there arose in the United States in the 1990s the convention of a black rim for glass pots containing ordinary coffee and an orange-tannish hue for decaffeinated coffee where publicly served.

2.5 The Evolution of Societal Specialization in Communication

In general, the greater the economic productivity and density of population, the less likely that speech communities will differentiate into linguistically distinct local variants. Further, the felt needs of wider communication will result in the development of standard languages and lingua francas. Such developments may be accelerated by the concentration of political power that will expand the role of some particular form for convenience in administrative use and/or to enhance solidarity by the common use of some preferred speech form. Such developments generally lead to widespread bilingualism or multilingualism in which each language fulfills some specific function.

Greater societal differentiation is reflected in, and has as one of its important aspects, changes in the social aspects of communication. In folk societies the majority of individuals are roughly of equal status in regard to communication. In industrialized societies specialized senders such as journalists, broadcasters, and authors of books send to far more people than those from whom they receive communications. No doubt the Internet is a further stage in this process, increasing the density of communication and the proportion of persons who both send and receive multiple communications.

It can be seen then that it is not so much language that has evolved after its initial invention as communication in general. Within this process, ordinary spoken language has a central and unique position as the source of post-language developments and as the ultimate level of explanation. In this broad sense there is a correlation between the evolution of communication and the evolution of culture. Language, evolves by begetting that which is not language, but transcends it even while ultimately being dependent on it and having language as its source.

See also: Brain, Evolution of; Communication and Social Psychology; Evolution: Diffusion of Innovations; Evolutionary Social Psychology; Linguistics: Overview; Neurogenetics and Behavior; Social Evolution: Overview; Sociality, Evolution of; Sociobiology: Philosophical Aspects; Verbal Communication, Psychology of

Bibliography

Alter S G 1999 *Darwinism and the Linguistic Image*. Johns Hopkins University Press, Baltimore, MD

Cavalli-Sforza L L 1988 Reconstruction of human evolution: Bringing together genetic, archaeological and linguistic data. *Proceedings of the National Academy of Sciences* **85**: 6002–6

Cavalli-Sforza L L, Menozzi P, Piazza A. *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

Darwin C 1871 *The Descent of Man and Selection in Relation to Sex*, 2 vols. John Murray, London

Durham W H 1991 *Coevolution: Genes, Culture and Human Diversity*. Stanford University Press, Stanford, CA

Greenberg J H 1957 Language and evolutionary theory. In: Greenberg J H (ed.) *Essays in Linguistics*. Chicago University Press, Chicago, pp. 56–65

Hennig W 1966 *Phylogenetic Systematics*. University of Illinois Press, Urbana, IL

Hockett C 1958 *A Course in Modern Linguistics*. Macmillan, New York

Hoenigswald H M, Wiener L (eds.) 1987 *Biological Metaphor and Cladistic Classification*. University of Pennsylvania Press, Philadelphia

Lyell C 1863 *The Principles of Geology*

Lyell C 1863 *Geological Evidences of the Antiquity of Man*. John Murray, London

Meggers B J 1959 *Evolution and Anthropology: A Centennial Appraisal*. Anthropological Society of Washington, Washington, DC

J. H. Greenberg

Evolution: Diffusion of Innovations

This article discusses the model of the diffusion of innovations and describe this model's key variables. First, however, a brief history of diffusion research is presented. The paradigm for diffusion research can be traced to the rural sociology research tradition, which began in the 1940s. Rural sociology is a subfield of sociology that focuses on the social problems of rural life. One rural sociology study in particular influenced the methodology, theoretical framework, and interpretations of later students in the rural sociology tradition, and in other diffusion research traditions. Ryan and Gross (1943) investigated the diffusion of hybrid seed corn among Iowa farmers. Hybrid seed, the result of agricultural research, was made available to Iowa farmers in 1928. The hybrid vigor of the new seed increased corn yields on Iowa farms, hybrid corn varieties withstood drought more effectively than the open-pollinated seed they replaced, and hybrid corn

was better suited to harvesting by mechanical corn pickers. By 1941, about 13 years after its first release, the innovation was adopted by almost 100 percent of Iowa farmers. Ryan and Gross studied the relatively rapid diffusion of hybrid corn in order to obtain lessons learned that might be applied to the diffusion of other farm innovations. However, the intellectual influence of the hybrid corn study reached far beyond the study of agricultural innovations, and outside of the rural sociology tradition of diffusion research. Since the 1960s, the diffusion model has been applied in a wide variety of disciplines such as education, public health, communication, marketing, geography, sociology, and economics (Pelto 1973, Rosen 2000, Van den Ven et al. 1989, Zaltman et al. 1973). Diffusion studies in these various disciplines have ranged from the rapid diffusion of the Internet to the nondiffusion of the Dvorak keyboard (in typewriters and computers). The innovations that have been studied in diffusion research have mainly been technological innovations.

Diffusion is the process by which (a) an innovation (b) is communicated through certain channels (c) over time (d) among the members of a social system (Rogers 1995). Diffusion is a special type of communication concerned with the spread of messages that are perceived as new ideas, and necessarily represents a high degree of uncertainty to the individual. The four main elements in the diffusion of new ideas are (a) the innovation, (b) communication channels, (c) time, and (d) the social system.

1. *The Innovation*

An innovation is an idea, practice, or object that is perceived as new by an individual or other unit of adoption. The characteristics of an innovation, as perceived by the members of a social system, determine its rate of adoption. Some innovations diffuse relatively slower, and other innovations diffuse faster. Why do certain innovations spread more quickly than others? The characteristics which determine an innovation's rate of adoption are: (a) relative advantage, (b) compatibility, (c) complexity, (d) trialability, and (e) observability.

1.1 *Relative Advantage*

This is the degree to which an innovation is perceived as better than the idea it supersedes. The degree of relative advantage may be measured in economic terms, but social prestige, convenience, and satisfaction are also important factors. It does not matter so much if an innovation has a great deal of objective advantage. What does matter is whether an individual perceives the innovation as advantageous. The greater the perceived relative advantage of an innovation, the more rapid its rate of adoption will be.

1.2 *Compatibility*

This is the degree to which an innovation is perceived as being consistent with the existing values, past experiences, and needs of potential adopters. An idea that is incompatible with the values and norms of a social system will not be adopted as rapidly as an innovation that is compatible. The adoption of an incompatible innovation often requires the prior adoption of a new value system, which is a relatively slow process.

1.3 *Complexity*

This is the degree to which an innovation is perceived as difficult to understand and use. Some innovations are readily understood by most members of a social system; others are more complicated and will be adopted more slowly. New ideas that are simpler to understand are adopted more rapidly than innovations that require the adopter to develop new skills and understandings.

1.4 *Trialability*

This is the degree to which an innovation may be experimented with on a limited basis. New ideas that can be tried on the installment plan generally will be adopted more quickly than innovations that are not divisible. An innovation that is trialable represents less uncertainty to the individual who is considering it for adoption, who can learn by doing.

1.5 *Observability*

This is the degree to which the results of an innovation are visible to others. The easier it is for individuals to see the results of an innovation, the more likely they are to adopt it. Such visibility stimulates peer discussion of a new idea, as friends and neighbors of an adopter often request innovation-evaluation information about it.

2. *Communication Channels*

The second main element in the diffusion of new ideas is the communication channel. Communication is the process by which participants create and share in-

formation with one another in order to reach a mutual understanding. A communication channel is the means by which messages get from one individual to another. Mass media channels are more effective in creating knowledge of innovations, whereas interpersonal channels are more effective in forming and changing attitudes toward a new idea, and thus in influencing the decision to adopt or reject a new idea. Most individuals evaluate an innovation, not on the basis of scientific research by experts, but through the subjective evaluations of near-peers who have adopted the innovation. So the process is essentially social in nature, being driven by individuals talking to others about the new idea, as they give meaning to the innovation.

3. *Time*

The third main element in the diffusion of new ideas is time. The time dimension is involved in diffusion in three ways.

3.1 *Innovation-decision Process*

First, time is involved in the innovation-decision process. The innovation-decision process is the mental process through which an individual (or other decision-making unit) passes from first knowledge of an innovation to forming an attitude toward the innovation, to a decision to adopt or reject, to implementation of the new idea, and to confirmation of this decision. An individual seeks information at various stages in the innovation-decision process in order to decrease uncertainty about an innovation's expected consequences.

3.2 *Innovativeness*

The second way in which time is involved in diffusion is in the innovativeness of an individual or other unit of adoption. Innovativeness is the degree to which an individual or other unit of adoption is relatively earlier in adopting new ideas than other members of a social system. There are five adopter categories, or classifications of the members of a social system on the basis on their innovativeness: (a) innovators, (b) early adopters, (c) early majority, (d) late majority, and (e) laggards.

3.2.1 Innovators. These are the first 2.5 percent of the individuals in a system to adopt an innovation. Venturesomeness is almost an obsession with innovators. This interest in new ideas leads them out of a local circle of peer networks and into more cosmopolite social relationships. Communication patterns and friendships among a clique of innovators are common, even though the geographical distance between the innovators may be considerable. Being an innovator has several prerequisites. Control of substantial financial resources is helpful to absorb the possible loss from an unprofitable innovation. The ability to understand and apply complex technical knowledge is also needed. The innovator must be able to cope with a high degree of uncertainty about an innovation at the time of adoption. While an innovator may not be respected by the other members of a social system, the innovator plays an important role in the diffusion process: that of launching the new idea in the system by importing the innovation from outside of the system's boundaries. Thus, the innovator plays a gatekeeping role in the flow of new ideas into a system.

3.2.2 Early adopters. These are the next 13.5 percent of the individuals in a system to adopt an innovation. Early adopters are a more integrated part of the local system than are innovators. Whereas innovators are cosmopolites, early adopters are localites. This adopter category, more than any other, has the greatest degree of opinion leadership in most systems. Potential adopters look to early adopters for advice and information about the innovation. This adopter category is generally sought by change agents as a local missionary for speeding the diffusion process. Because early adopters are not too far ahead of the average individual in innovativeness, they serve as a role-model for many other members of a social system. The early adopter is respected by his or her peers, and is the embodiment of successful, discrete use of new ideas. The early adopter knows that to continue to earn this esteem of colleagues and to maintain a central position in the communication networks of the system, they must make judicious innovation-decisions. The early adopter decreases uncertainty about a new idea by adopting it, and then conveying a subjective evaluation of the innovation to near-peers through interpersonal networks.

3.2.3 Early majority. This is the next 34 percent of the individuals in a system to adopt an innovation. The early majority adopt new ideas just before the average member of a system. The early majority interact frequently with their peers, but seldom hold positions of opinion leadership in a system. The early majority's unique position between the very early and the relatively late to adopt makes them an important link in the diffusion process. They provide interconnectedness in the system's interpersonal networks. The early majority are one of the two most numerous adopter categories, making up one-third of the members of a system. The early majority may deliberate for some time before completely adopting

a new idea. 'Be not the first by which the new is tried, nor the last to lay the old aside,' fits the thinking of the early majority. They follow with deliberate willingness in innovations, but seldom lead.

3.2.4 Late majority. This is the next 34 percent of the individuals in a system to adopt an innovation. The late majority adopt new ideas just after the average member of a system. Like the early majority, the late majority make up one-third of the members of a system. Adoption may be the result of increasing network pressures from peers. Innovations are approached with a skeptical and cautious air, and the late majority do not adopt until most others in their system have done so. The weight of system norms must definitely favor an innovation before the late majority are convinced. The pressure of peers is necessary to motivate adoption. Their relatively scarce resources mean that most of the uncertainty about a new idea must be removed before the late majority feel that it is safe to adopt.

3.2.5 Laggards. These are the last 16 percent of the individuals in a system to adopt an innovation. They possess almost no opinion leadership. Laggards are the most localite in their outlook of all adopter categories; many are near isolates in the social networks of their system. The point of reference for the laggard is the past. Decisions are often made in terms of what has been done previously. Laggards tend to be suspicious of innovations and change agents. Resistance to innovations on the part of laggards may be entirely rational from the laggard's viewpoint, as their resources are limited and they must be certain that a new idea will not fail before they can adopt.

3.3 Rate of Adoption

The third dimension in which time is involved in diffusion is in rate of adoption. The rate of adoption is the relative speed with which an innovation is adopted by members of a social system. The rate of adoption usually is measured as the number of members of the system that adopt the innovation in a given time period. As shown previously, an innovation's rate of adoption is influenced by the five perceived attributes of an innovation.

4. The Social System

The fourth main element in the diffusion of new ideas is the social system. A social system is defined as a set of interrelated units that are engaged in joint problem-solving to accomplish a common goal. The members or units of a social system may be individuals, informal groups, organizations, and/or subsystems. The social system constitutes a boundary within which an innovation diffuses. How the system's social structure affects diffusion has been studied. A second area of research involved how norms affect diffusion. Norms are the established behavior patterns for the members of a social system. A third area of research has had to do with opinion leadership, the degree to which an individual is able to influence informally other individuals' attitudes or overt behavior in a desired way with relative frequency. A change agent is an individual who attempts to influence clients' innovation-decisions' in a direction that is deemed desirable by a change agency.

A fourth type of research involves the types of innovation-decisions (whether individual adoption decisions or organizational decisions, and whether they are made by an authority or by consensus). Another type of research, conducted mainly by anthropologists, analyzed the consequences of innovation.

A crucial concept in understanding the nature of the diffusion process is the critical mass, which occurs at the point at which enough individuals have adopted an innovation that the innovation's further rate of adoption becomes self-sustaining. The concept of the critical mass implies that outreach activities should be concentrated on getting the use of the innovation to the point of critical mass. These efforts should be focused on the early adopters, the 13.5 percent of the individuals in the system to adopt an innovation after the innovators have introduced the new idea into the system. Early adopters are often opinion leaders, and serve as role-models for many other members of the social system. Early adopters are instrumental in getting an innovation to the point of critical mass, and hence, in the successful diffusion of an innovation.

5. Opinion Leaders

The two-step flow model of mass communication suggests that communication messages flow from a source, via mass media channels, to opinion leaders, who in turn pass them on to followers (Rogers 1995). This two-step flow model of mass communication can be utilized to diffuse an innovation. The model focuses attention on the inter-media interface between mass media channels and interpersonal communication channels. Mass media channels are primarily creators of awareness-knowledge of innovations, while interpersonal networks are more important in persuading individuals to adopt or reject innovations.

The two-step flow model highlights the importance of opinion leaders in diffusing innovations, and especially in reaching the critical mass. Opinion leaders informally influence other individuals' attitudes or

overt behavior in a desired way with relative frequency, so they are critical to the successful diffusion of innovations.

6. New Communication Technologies

We live in the age of a communication revolution. New and developing communication technologies, and, ultimately, equipment and systems, are influencing the communication industry and American society greatly. Take the Internet, for example. The Internet is a network of interconnected computer systems. By using the Internet as a data highway, an individual can communicate with colleagues and friends at great distance very cheaply, and gain direct access to information in other computer systems. Therefore, diffusion of the Internet has been exceedingly rapid. From 1990 to 2001, the number of individuals adopting the Internet in North America rose from approximately 5 million to 200 million. A huge percentage of the population is interconnected via the Internet. The critical mass in the diffusion of the Internet occurred in about 1989 or 1990, after 20 years of very slow increase in the rate of adoption of computer networks.

Most schools and local libraries are connected to the Internet, and can provide Internet access to individuals who otherwise might not own, or have access to, a computer. The potential for using computer-mediated communication for outreach activities is enormous, and the nature of the diffusion process may be speeding up and becoming less spatially restricted.

See also: Communication: Geographic Aspects; Diffusion: Anthropological Aspects; Diffusion: Geographical Aspects; Diffusion, Sociology of; International Communication: History; Mass Communication: Technology; Technological Innovation

Bibliography

Pelto J 1973 *The Snowmobile Revolution: Technology and Social Change in the Arctic*. Cummings, Menlo Park, CA

Rogers M 1995 *Diffusion of Innovations* 4th edn. Free Press, New York

Rosen E 2000 *The Anatomy of Buzz: How to Create Word of Mouth Marketing*. Doubleday/Currency, New York

Ryan B, Gross C 1943 The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology* **8**: 15–24

Van de Ven H, Angle H L, Poole M (eds.) 1989 *Research on the Management of Innovation: The Minnesota Studies*. Ballinger/Harper & Row, New York

Zaltman G, Duncan R, Holbeck J 1973 *Innovations in Organizations*. Wiley, New York

E. M. Rogers

Evolution, History of

The term 'evolution' is widely used to denote the development through time of societies, cultures, and more especially of living species. It is often contrasted with the view that these entities were divinely created as we see them today, and is routinely (but increasingly controversially) associated with the idea of progress. This article outlines the various models of evolution which have been suggested to account for the development of life and of social organization, and then shows how the theories were formulated and popularized. Particular attention is paid to the work of Charles Darwin, whose theory of biological evolution by natural selection is seen as the most influential expression of the basic idea of natural development.

1. Models of Evolution

The term 'evolution' is derived from the Latin *evolutio*, denoting the unrolling of a scroll. In the eighteenth century it was applied to the growth of the embryo, then often described as the mere expansion of a preformed miniature. This application was retained up to the nineteenth century, by which time it was clear that the development of the embryo consisted of the progressive appearance of more complex structures. The philosopher Herbert Spencer generalized the term by applying it to any process of natural progressive development, thereby creating the still-common belief that all evolution must be progressive. In particular, Spencer applied the term to the development of societies and of life on earth.

Charles Darwin did not often use the term when describing his theory of the origin of species, but by the end of the nineteenth century this had become the most common version of 'evolution.' But Darwin's theory did not imply an inevitable progression toward complexity, and much controversy has surrounded the association between evolution and progress. Darwin's theory of natural selection now dominates biology and is being extended to many other domains, but it is but one of a number of mechanisms which have been suggested to explain evolution, each of which has its own associated implications.

1.1 The Ladder of Progress

Perhaps the most popular model of evolution, and one which reveals the link with embryology, is of a linear hierarchy of stages of development from inception to maturity. Such a model simply assumes that all change is progressive and that there is only one direction of development. When applied to societies or species, the model can be adapted by supposing that many lines of

development each ascend the same hierarchy but at different rates, so that the slow developers illustrate the ancestral stages already passed through by those at the top of the scale.

1.2 Evolutionary Trees

Although the ladder model is still popularly applied to the evolution of life on earth, and was once accepted in the social sciences, biologists and sociologists now realize that there is no single goal toward which evolution is developing—neither the human race nor Western civilization can be treated as the most mature phase of development. Evolution is better represented as a tree or bush with many branches, each of which is constantly subdividing. The end product of one branch cannot be seen as the goal toward which all the others are ascending, and it is impossible for the end product of one branch to illustrate exactly the ancestral form of another branch. The living apes are not immature humans, because even if the common ancestor from which humans and apes have evolved would be classified as an ape, it would no longer be identical to any of the living ape species. If there is progress on the branching-tree model, it has to be defined in more sophisticated terms so that there are many different ways to become more complex.

Darwin argued that branching evolution explains how we are able to classify species into groups within groups (species, genera, families, etc.). Two species descended from a common ancestor will still share many characters derived from that ancestor, and will share some characters with even more distant relatives. Darwin and his modern followers assume that each new character is developed only once, so all species exhibiting the character are descended from a common ancestor which also had the character. This assumption will be invalidated if there are cases where two lines of evolution evolve the same character independently (convergence or homoplasy). It will also be invalidated if there are even small-scale examples of ladder-like evolution where several branches are constrained to evolve along the same pattern of development (parallelism).

1.3 Adaptation

The branching model had already been applied to the evolution of languages before Darwin and his followers used it to transform biology. Darwin's key insight was that there was no natural mechanism that could force evolution along a single predetermined path. Evolution is driven not by progress but by the necessity for each species to adapt to its environment in an ever-changing world. Branching occurs because a single population can sometimes become divided by a geographical barrier, after which each isolated population will develop independently in response to the different adaptive pressures it encounters. Progress to a more complex structure is at best a by-product of adaptation and is never inevitable in any one branch; indeed, many species are the products of degenerative evolution. In the animal kingdom, two branches of evolution can seldom if ever rejoin via hybridization (although this does happen in plants). Here, there is a clear contrast with social and linguistic evolution, where mergings and borrowings are frequent.

1.4 Darwinism and Lamarckism

Two mechanisms of adaptive evolution have been suggested: natural selection (Darwinism) and the inheritance of acquired characters (Lamarckism). Darwin assumed that the species consists of a population of organisms which exhibit some degree of individual variation and that these variations are inherited—we now explain them as genetic differences produced by mutations. He argued that if the population is exposed to a new environment, some variants will by chance be able to cope better with the new conditions; they will survive and reproduce, transmitting their favorable characters to the next generation. Any that are less fitted to the environment will be unlikely to breed and may even die, so their characters are eliminated from the population. Over many generations this process of natural selection (Spencer called it the 'survival of the fittest') will change the species to adapt it to the environment.

Some time before Darwin, J. B. Lamarck proposed that the inheritance of characters acquired during an organism's lifetime could accumulate to give adaptive transmutation. An acquired character is produced by the organism's behavior, which in the wild is usually a response to the environment. Thus, the ancestral giraffes, striving to reach the leaves of trees as an alternative source of food, were able to lengthen their necks, just as a weightlifter can acquire better-developed muscles. If such characters are inherited, the next generation of giraffes would be born with longer necks and would then stretch them further. This mechanism was once popular in biology, and corresponds quite closely to how humans transmit new ideas to their descendants. But where teaching and learning provides a natural mechanism for ideas to be transmitted, modern genetics has shown that there is no way in which acquired biological characters can be inherited. Lamarckism has been banished from biology: The giraffe got its long neck because those who by chance inherited the genes producing that character reproduced faster than those who carried the genes for shorter necks. The genetic differences are the product of mutations which are random as far as the requirement of the species are concerned.

1.5 Nonadaptive Evolution

There are also several mechanism of biological evolution which suppose that not all developments are

adaptive. The possibility that saltations (evolutionary jumps or macromutations) might found new populations was once taken seriously even by geneticists. The theory of orthogenesis supposed that the species was somehow programmed to vary continuously in a particular direction, forcing groups of species to evolve in parallel. Such ideas are now largely discounted in biology, although the possibility that there might be constraints on the range of possible variation is taken more seriously. In the social sciences, the question of whether all cultural and social differences serve a practical purpose has been much debated.

2. *Pre-Darwinian Evolutionism*

Darwin's was not the first theory of evolution, although the publication of his *Origin of Species* in 1859 triggered the conversion of most scientists to evolutionism. Several earlier versions of the idea were proposed, and although rejected by conservative thinkers and most scientists, they did much to prepare the way for the reception of Darwinism.

2.1 *Creation and Design*

The early Church Fathers did not take the Genesis story of creation in the Bible literally, and it was the Protestant theologians of the seventeenth century who established the world view which survives in modern 'young earth' creationism. Archbishop James Ussher published his much-ridiculed calculation that the earth was created in 4004 BC. In such a climate of opinion a form of natural theology flourished in which it could be argued that God had designed and created all of the modern species just as we see them today, each with the structures and instincts appropriate for its intended habitat. John Ray and many other naturalists stressed the adaptation of species to their environment, and argued that similarities could be used to classify species because they expressed the rational order of God's plan of creation—a view adopted by Carolus Linnaeus, who founded the modern system of biological classification (Greene 1959). Others exploited the ancient concept of the 'chain of being,' in which the plan of creation was a linear hierarchy with the human species at the top of the animal series. This was promoted without any sense of a progression through time, although the influence of the chain was retained in some of the early theories of progressive evolution.

2.2 *Evolution in the Enlightenment*

The more radical thinkers of the eighteenth-century Enlightenment recognized that the Earth was much older than Ussher's estimate, and that the planet had changed considerably in the course of its history. Some explored the possibility that life had been produced by spontaneous generation from nonliving matter, and began to doubt that species remained unchanged through time. The French naturalist Georges Leclerc, Comte de Buffon, was especially influential in this respect, as were the materialist philosophers Denis Diderot and the Baron d'Holbach (Roger 1998).

Toward the end of the eighteenth century, Erasmus Darwin (grandfather of Charles) published a theory of transmutation in his *Zoonomia*, arguing that all living things were derived from a simple origin and had become more complex through their striving to adapt to the environment. These views were echoed in a more sophisticated form in J. B. Lamarck's *Zoological Philosophy* of 1809. Lamarck accepted that there was a basic progressive force driving living things to ascend the chain of being over many generations, but he accepted that no linear pattern is visible in the present animal and plant kingdoms, explaining the divergence as a consequence of life's constant need to adapt to an ever-changing environment. He invoked the inheritance of acquired characters (described above) to explain this process, and his name has been associated with that mechanism ever since (Corsi 1988).

2.3 *The Fossil Record*

Historians used to think that Lamarck was marginalized by his great rival, Georges Cuvier, who pioneered the study of fossil bones to reconstruct the history of life. Cuvier showed that there was a succession of populations in the course of the earth's history, but insisted that species disappear abruptly from the record through the action of geological catastrophes. He rejected transmutation, but did not openly endorse the idea of divine creation—although the theory of a succession of creations was popularized by Cuvier's British followers, including William Buckland. Although they had extended the Earth's history, Cuvier and Buckland were convinced that the human race was a recent creation. Their work thus endorsed a view of human history in which there was no time for social evolution, many scholars still believing that language and the arts of civilization were taught to the earliest humans by their Creator.

Later historical work has shown that although the theory of successive creations was endorsed by the scientific establishment, more radical thinkers were attracted to Lamarck's theory (Desmond 1989). Darwin, who developed his theory in the 1830s, concealed it for fear of being associated with this radical movement. The situation changed significantly following the anonymous publication of Robert Chambers' *Vestiges of the Natural History of Creation* in 1844. Chambers argued that an ideology of progress was just what the middle class needed to support its demands for reform, and provided a scientific foun-

dation for this ideology based on the idea of progressive evolution. Although he presented evolution as the unfolding of a divine plan, his book's open recognition that humans are transformed animals, and that the mind is a product of the expansion of the brain caused an outcry from conservative thinkers. Over the following decade, however, *Vestiges* popularized the idea of progressive evolution, creating a space in which Darwin's more scientific theory could have an impact.

3. The Development of Darwin's Theory

Charles Darwin was converted to evolutionism by his discoveries on the voyage of HMS *Beagle* (1831–6) and returned to England determined to understand how the process worked. He soon discovered the mechanism of natural selection, and continued to work on the various aspects of his theory in relative secrecy. Much controversy has surrounded Darwin's discovery, because in addition to the many scientific factors influencing his thinking, he acknowledged the impact of Thomas Malthus' principle of population, a product of the ideology of *laissez-faire* economics.

3.1 Scientific Influences

While studying at Cambridge, Darwin was attracted to William Paley's *Natural Theology* of 1802, which restated the argument that species are adapted to their environment through the wisdom and benevolence of the Creator. On the voyage of the *Beagle*, his observations of South American geology converted him to Charles Lyell's uniformitarian position, which accounted for all geological change in terms of the action of observable forces over vast periods of time. Biogeographical studies, especially those of the birds of the Galapagos Islands, forced him to rethink the idea of Divine Creation, because he saw that individual islands had each acquired a distinct species of immigrant families, such as the finches. On his return to England he soon accepted a complete evolutionary philosophy and began to search for a mechanism (Browne 1995).

Darwin saw that evolution must be a branching process and that the ultimate cause was the adaptation of isolated populations to different environments. He studied animal breeding and realized that breeders produce new varieties by selecting out individual animals with the required character and breeding only from these. Artificial selection draws on the natural and largely random variability of the population and allows only individuals with a selected character to breed, thereby enhancing that character in future generations. Darwin was soon looking for a natural equivalent of this process.

3.2 Malthus and Population

A vital step in Darwin's thinking came when he read Malthus' *Essay on Population*, which argued that populations breed so rapidly that they outstrip their food supply. Drawing on hints in Malthus, Darwin realized that population pressure must generate a struggle for existence in which many individuals die. He was now able to argue that any better-adapted variant would be more successful in the struggle and would be more likely to survive and reproduce, while poorly adapted variants would die, thus establishing the theory of natural selection described in Sect. 1.4. This gave Darwin the basis on which to erect a comprehensive account of evolution. The role played by Malthus' ideologically loaded principle has, however, led to frequent charges that Darwin's theory is itself a projection of *laissez-faire* values on to nature, with the consequence that the theory would be used to justify the competitive ethos of Victorian capitalism (Young 1985). Much controversy also surrounds the degree of Darwin's commitment to the idea of progress. Although the selection theory undermines the ladder model of progress, historians now accept that Darwin's thinking was influenced by the assumption that competition is the driving force of a more general form of progress. However, it is clear that Darwin's commitment to branching evolution went far beyond the conventional social values later linked to his theory, while his application of the population principle undermined Malthus' own ideology, which was an attempt to discredit the optimism of social progressionists.

Darwin let only a few close contacts know of his theory, while continuing to work on exploring its scientific implications. He came to appreciate that, even in a stable environment, specialization would allow natural selection to continue the process of divergence by which species separate from close relatives. Eventually, he began to write up his ideas for publication, but was interrupted in 1858 by the arrival of a paper by Alfred Russel Wallace describing a similar (although not identical) theory. The joint Darwin–Wallace papers were published in that year, while Darwin rushed to complete the *Origin of Species*.

4. Darwinism in Biology

The *Origin of Species* sparked an intense debate which soon led to the conversion of most scientists and educated people to evolutionism (Hull 1973). Despite initial opposition by conservative thinkers, the belief that the human race was the outcome of a progressive trend built into nature was accepted. In science, too, creationism was replaced by the expectation that all living species are the products of natural processes, and here, too, there was a widespread assumption that such processes guaranteed progress, even if super-

imposed on a branching model of development. The success of Darwinism needs to be assessed with care because the popularity of this developmental model of evolution meant that the more radical implications of Darwin's thinking were evaded (Bowler 1988). Far from welcoming the theory of natural selection, most nineteenth-century scientists, including some of Darwin's leading supporters, were lukewarm about it. Opposition grew toward the end of the century, with many alternative mechanisms of evolution being tried out.

4.1 The History of Life

Much of the first generation of evolutionists' efforts focused on reconstructing the course of the history of life on earth from anatomical, embryological, biogeographical, and fossil evidence (Bowler 1996). Darwin himself treated this project with caution, but under the influence of the German biologist Ernst Haeckel, a whole generation of evolutionists strove to complete it. They had some successes, but the problems were much greater than they had anticipated and the project was sidelined by new developments in biology at the beginning of the twentieth century. Darwin had stressed the imperfection of the fossil record, but in a few important areas new discoveries helped to show how the major steps in evolution had taken place. Elsewhere, anatomical and embryological evidence was used to reconstruct evolutionary relationships, on the assumption that new characters were developed only once. In embryology, Darwin's assumption that early stages of development would illustrate affinities was overtaken by Haeckel's recapitulation theory, according to which earlier stages could actually illustrate ancestral adult forms. This way of thinking was a product of the progressionist way of thought, presupposing a privileged line of development toward maturity. Only in biogeography did it become clear that evolution was a highly irregular process, depending on the hazards of migration in a world where geological forces were constantly remolding the topography.

4.2 Natural Selection and the Alternatives

Historians have focused on the debates over natural selection, but this is more a product of our modern preoccupation with that mechanism than of its significance in late-nineteenth-century biology. One much-discussed theme centers on Darwin's concept of heredity, which was quite unlike the model of unit-character inheritance promoted by Mendelian genetics. It has even been argued that, without the concept of nonblending unit characters, natural selection was fatally flawed. It is true that Darwin was much concerned by Fleeming Jenkins' 1867 review which argued that if male and female parental characters blend in the offspring, the advantages of a well-favored 'sport' or mutation would be diluted rapidly. But as A. R. Wallace pointed out, this objection is irrelevant if there is a continuous range of variation in the population, which is the case for most characters (Gayon 1998).

In fact, many of the objections to natural selection were intended only to show that evolution must be a more purposeful process than any mechanism based on random variation would allow. One major alternative which now became popular was the Lamarckian theory of the inheritance of acquired characters, in which the animals' own behavior directs their variation. Some objections to natural selection were aimed at the whole idea of adaptive evolution and were intended to show that some internally driven force directed the organisms' variation. The theory of orthogenesis supposed that variation generated linear evolutionary trends, while the theory of saltations assumed that macromutations could somehow found new breeding populations (Bowler 1983).

5. Human Origins and Social Evolution

Many religious thinkers objected to the idea that the human soul was the product of evolution from an animal ancestry. Much of this initial opposition was overcome by stressing the purposeful nature of evolution, allowing the appearance of humanity to be presented as the culmination of a divine plan (Moore 1979). Darwin's theory of evolution threatened this assumption, and in his *Descent of Man* of 1871 he argued that humans had evolved from African apes which had stood upright as an adaptation to a new lifestyle on the open plains. Most late-nineteenth-century accounts of human evolution, however, stressed the expansion of the brain as the main driving force of the process. At the same time, the idea of progressive evolution was applied to the history of human cultures and societies. Some evolutionists, including Spencer, stressed competition as the driving force of progress, leading to the charge that they were promoting a 'social Darwinism' designed to legitimize capitalist values. While Darwin's theory was certainly associated with this movement, it was by no means the only scientific justification offered.

5.1 Cultural Evolutionism

Coincidentally with the Darwinian revolution, archaeologists undermined the belief that the human race was a recent creation, and exposed a vast period of prehistory in which our ancestors had used only stone tools. It was this initiative rather than any influence from Darwinism which led anthropologists such as

Edward B. Tylor to equate modern 'savages' (i.e., people with relatively unsophisticated technology) with the ancestral stages through which civilized humans had passed in prehistoric times on their march toward a more mature culture. Cultural evolutionism was a product of the nineteenth-century developmental viewpoint, based on the ladder model of progress (Bowler 1989a). Tylor himself did not accept that humans had evolved from apes, but other cultural evolutionists, especially John Lubbock, linked the two modes of evolution and argued that 'savages' were biologically as well as culturally primitive—living examples of the earlier stages in human biological and mental evolution. Herbert Spencer's model of social evolution, while stressing ostensibly the divergent nature of evolution, still presented 'lower' races as surviving primitives incapable of matching the mental powers of the Anglo-Saxons.

5.2 *Social Darwinism*

Spencer is most often remembered, however, as the social philosopher who linked the idea of progress to the ideology of free-enterprise individualism. As an exponent of *laissez-faire*, and the coiner of the term 'survival of the fittest,' it has been easy for critics to portray Spencer's enthusiasm for struggle as the motor of progress as a direct application of the Darwinian selection theory to society, hence the image of him as the founder of 'social Darwinism.' Spencer worried about the elimination of the unfit, but in biology he defended Lamarckism, and his enthusiasm for struggle came more from a conviction that competition would stimulate individual self-improvement (which in a Lamarckian scheme could be passed on to future generations). Darwin's theory is thus only one of the scientific factors which stimulated the ideology of 'progress through struggle'—remembered preferentially because natural selection has now become the only mechanism by which that metaphor can be put into practice in biology (Bannister 1979). In the later nineteenth century, there was more emphasis on struggle between nations and races, which bypasses the central feature of natural selection within populations.

6. *Modern Darwinism*

By the end of the nineteenth century, the attempt to reconstruct the course of biological and social evolution had run out of steam. The social sciences abandoned the evolutionary model and any suggestion that human behavior is predetermined by biological factors. Biologists began to study the actual processes of heredity and variation, leading to the emergence of Mendelian genetics. The geneticists destroyed Lamarckism, and although at first hostile to the selection theory, they were eventually convinced that genetic mutation was the source of the random variation that is the raw material of natural selection. By the mid-twentieth century, the 'Modern Synthesis' of Darwinism and genetics had come to dominate biology but remained controversial elsewhere.

6.1 *The Modern Darwinian Synthesis*

In the last decade of the nineteenth century, the statistician Karl Pearson developed a mathematical model of natural selection acting on the variations in a wild population (Gayon 1998). But many biologists were now exploring the idea that new species were produced by saltations or macromutations, and it was partly from this tradition that the 'rediscovery' of Gregor Mendel's laws of particulate inheritance emerged. Ignored since the 1860s, these laws postulated the existence of unit characters transmitted unchanged from one generation to the next. In 1900, Mendel's laws were proclaimed as the basis for a new science of heredity, which William Bateson soon called 'genetics.' But Pearson was already hostile to Bateson's saltationism, while Bateson rejected Pearson's commitment to continuity and adaptation. Genetics was thus perceived as a new alternative to Darwinism, although it destroyed the credibility of Lamarckism by suggesting that there is no plausible way in which acquired characters can affect the gene (Bowler 1989b).

T. H. Morgan's studies of genetic mutations showed that most macromutations are fatal, while small mutations simply add to the genetic variation in the population. Soon, Morgan conceded that natural selection would affect the reproductive success of the various genes. In the 1920s and 1930s, R. A. Fisher, J. B. S. Haldane, and Sewall Wright brought together the statistical study of populations with the new genetics to create a science of population genetics based on the genetical theory of natural selection (Provine 1971, Mayr 1982). Naturalists such as Julian Huxley and Ernst Mayr now recognized that the adaptation of local populations to their environment was the most likely explanation of how species originate. The new selection theory was integrated with these wider developments to create what Huxley called the 'Modern Synthesis.'

6.2 *Later Developments*

The Darwinian synthesis has dominated biology since the mid-twentieth century, although there have been numerous developments within it and some efforts to limit its applicability. Some modern thinkers now argue that the concept of natural selection can be applied to almost all areas where change is observed, sweeping away the whole idea of a universe which is

the product of divine purpose. Most controversial are efforts to apply selection to human beings, especially in the area of social behavior. The science of sociobiology has been successful in explaining the social behavior of animals in terms of genetics and reproductive success, via the concept of the 'selfish gene.' But E. O. Wilson's efforts to apply sociobiology to human behavior by suggesting that we are driven by instincts implanted by natural selection have been greeted with much suspicion by social scientists, and have also been dismissed as a revived 'social Darwinism' by the political left (Segerstrale 2000).

See also: Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Darwin, Charles Robert (1809–82); Darwinism; Evolution: Optimization; Evolutionary Social Psychology; Evolutionism, Including Social Darwinism; Malthus, Thomas Robert (1766–1834)

Bibliography

Bannister R C 1979 *Social Darwinism: Science and Myth in Anglo-American Social Thought*. Temple University Press, Philadelphia, PA

Bowler P J 1983 *The Eclipse of Darwinism: Anti-Darwinian Evolution Theories in the Decades around 1900*. Johns Hopkins University Press, Baltimore, MD

Bowler P J 1988 *The Non-Darwinian Revolution: Reinterpreting a Historical Myth*. Johns Hopkins University Press, Baltimore, MD

Bowler P J 1989a *The Invention of Progress: The Victorian and the Past*. Basil Blackwell, Oxford, UK

Bowler P J 1989b *The Mendelian Revolution: The Emergence of Hereditarian Concepts in Modern Science and Society*. Athlone, London

Bowler P J 1989c *Evolution: The History of an Idea*. University of California Press, Berkeley, CA

Bowler P J 1996 *Life's Splendid Drama: Evolutionary Biology and the Reconstruction of Life's Ancestry, 1860–1940*. University of Chicago Press, Chicago

Browne J 1995 *Charles Darwin: Voyaging*. Jonathan Cape, London

Corsi P 1988 *The Age of Lamarck: Evolutionary Theory in France, 1790–1830*. University of California Press, Berkeley, CA

Desmond A 1989 *The Politics of Evolution: Morphology, Medicine and Reform in Radical London*. University of Chicago Press, Chicago

Gayon J 1998 *Darwinism's Struggle for Survival; Heredity and the Hypothesis of Natural Selection*. Cambridge University Press, Cambridge, UK

Greene J C 1959 *The Death of Adam: Evolution and its Impact on Western Thought*. Iowa State University Press, Ames, IA

Hull D L 1973 *Darwin and his Critics: The Reception of Darwin's Theory by the Scientific Community*. Harvard University Press, Cambridge, MA

Mayr E 1982 *The Growth of Biological Thought*. Harvard University Press, Cambridge, MA

Moore J R 1979 *The Post-Darwinian Controversies: A Study of the Protestant Struggle to Come to terms with Darwin in Britain and America, 1870–1900*. Cambridge University Press, Cambridge, UK

Provine W B 1971 *The Origin of Theoretical Population Genetics*. University of Chicago Press, Chicago

Roger J 1998 *The Life Sciences in Eighteenth-century French Thought*. Stanford University Press, Stanford, CA

Ruse M 1979 *The Darwinian Revolution*. University of Chicago Press, Chicago

Segerstrale U 2000 *Defenders of the Truth: The Battle for Science in the Sociobiology Debate and Beyond*. Oxford University Press, Oxford, UK

Young R M 1985 *Darwin's Metaphor: Nature's Place in Victorian Culture*. Cambridge University Press, Cambridge, UK

P. J. Bowler

Evolution, Natural and Social: Philosophical Aspects

In its most minimal sense, *evolution* simply means gradual change by some natural process. In this sense, belief that species of plants and animals have evolved is at least as old as Aristotle's contrary belief in the fixity of the species. The theory of evolution acquires more content as the nature of the natural process involved is specified.

Charles Darwin (1859) provided the broad outlines of the modern theory. There are traits that are inherited. There is some process that produces natural variation in these traits. The traits may affect the ability of the organism to reproduce, and thus the average number of individuals bearing the traits in the next generation. Therefore, those traits that enhance reproductive success increase in frequency in the population, and those that lead to reproductive success below the average decrease in frequency. The three essential factors in Darwin's account are (a) natural variation, (b) differential reproduction, and (c) inheritance. The questions of the existence, extent, and nature of these factors are empirical scientific questions.

Could these processes, as they naturally occur, account for large changes over time? Darwin argued that they could by calling attention to the large changes that had come about from the selective breeding of plants and animals. Selective breeding is just a particularly severe form of differential reproduction, imposed by the breeder. Darwin emphasized this argument by calling his theory a theory of Natural Selection:

> I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection, in order to mark its relation to man's power of selection. But the expression often used by Mr. Herbert Spencer, of the Survival

> of the Fittest, is more accurate, and is sometimes equally convenient. We have seen that man by selection can certainly produce great results, and can adapt organic beings to his own uses, through the accumulation of slight but useful variations, given to him by the hand of Nature. But Natural Selection, as we shall hereafter see, is a power incessantly ready for action, and is as immeasurably superior to man's feeble efforts, as the works of Nature are to those of Art. (Darwin 1859 *Origin of the Species*, 6th edn., Chap. III)

The status of Spencer's description has been a matter of controversy. On superficial reflection on the slogan, *survival of the fittest*, more than one commentator has asked: 'If fitness is fitness to survive, is not the theory of survival of the fittest simply a tautology?' At this point we can give a preliminary examination of this idea in a highly idealized setting. Let us take *Darwinian fitness* just to mean expected number of offspring. Let us suppose that the population is effectively infinite, so that the expected number of offspring for a type can be taken to equal the average number of offspring. Let us suppose that reproduction is asexual and transmission of a type is perfect. If we had one generation every time unit, these assumptions specify a dynamics of population proportions called the discrete time *replicator dynamics*. Taking generations close together, or taking overlapping generations, it is possible to idealize to a system of differential equations which represent the continuous time *replicator dynamics* (Taylor and Jonker 1978, Hofbauer and Sigmund 1988). This is an idealized model of differential reproduction. In this idealized model, we can ask whether it can be proved that the fittest type will go to fixation in the population.

To get any kind of affirmative answer, we must make more assumptions. Suppose that the fitness of a type is fixed. It does not depend on the frequency of types in the population. The environment is considered constant. Suppose that all the types of interest are initially present in the population and that one of them has the greatest Darwinian fitness. Then it is possible to prove that the dynamics will converge to a state of fixation of the fittest strategy, a state where its proportion of the population is 100 percent. Convergence means just that the population gets arbitrarily close to fixation in an arbitrarily long time. In this sense, survival of the fittest can be mathematically demonstrated. In this sense it is a tautology.

But the demonstration depends not only on the conception of Darwinian fitness, but also essentially on a battery of idealizations and assumptions. If the Darwinian fitness of a type is not fixed, but dependent on the frequency of other types in the population, then—even with all the other idealizations in place—the dynamics may not converge to any state at all. (It may cycle or even exhibit chaotic behavior.) If there is a fitter type not present in the population, then differential reproduction provides no way for it to get in. The model can be modified in various ways to allow for mutation, which may put the type in question into the population, but then the theorem is no longer true. If the population is finite, then only probabilistic conclusions can be reached. These depend on factors other than simply Darwinian fitness, such as the distribution number of progeny for a type and the size of the population. Moreover, in a finite population mutation cannot keep every possible type in play, but rather must be modeled as a chance process that introduces some new types from a very large space of possible types. With the slightest move towards realism the tautology goes away.

In the first half of the twentieth century, Darwinian theory was enriched by Mendelian genetics. Gregor Mendel, an Augustinian monk, had discovered basic principles of genetics by a series of experiments in plant breeding carried out in the garden of the monastery (Mendel 1866). Mendel first bred pure strains of peas that differed in seven selected traits, for instance, green vs. yellow seeds. He crossed these, self-fertilized the hybrids, and studied the frequency of the traits in the hybrids. He found a ratio of 3:1, from which he concluded that each individual carries two genetic units for the trait, one from each parent, chosen at random from the parent's two genetic units. (This is haploid reproduction.) Thus from hybrid parents, carrying color genes $\langle gy \rangle$ we would expect to get homozygotes $\langle gg \rangle$ and $\langle yy \rangle$ one-quarter of the time each and heterozygotes $\langle gy \rangle$ one-half of the time. This 1:2:1 ratio then explains the observed 1:3 ratio on the hypothesis that one of the possibilities is *dominant*, so that this possibility is expressed by the heterozygotes as well as by the appropriate homozygote. Mendel opened the door to genetic models of inheritance.

Mendelian genetics introduces a new source of variation in such sexually reproducing haploid species. Recombination of genes creates rapid variation. But the variation created has its limits. Can recombination of fixed genetic units account for the origin of new species and the whole sweep of Darwinian evolution? Darwinians thought not, and at the beginning of the twentieth century Darwinism and Mendelism were enemies. Only when mutation is included in the picture, so that more fundamental sorts of variation are possible on a longer timescale, is it evident that Mendelism and Darwinism can be thought of as reinforcing each other. This 'modern synthesis' of evolution and genetics was created in the first half of the twentieth century (Fisher 1930, Haldane 1932, Dobzhanski 1937, Wright 1968, 1969, 1977, 1978).

Mendelian genetics also enriches the account of differential reproduction. We now need to consider reproduction at more than one level. Individuals reproduce but so do genes. Fitness on one level is connected to that on the other, but in a way that can be complicated. The most obvious case is one in which expression of genes works in such a way that an individual who is a heterozygote is then fitter than either of the homozygotes. A famous example is that of sickle

cell anemia. The two possibilities of the gene are sickle, s, or not, n. Heterozygotes, ⟨s,n⟩ are protected against malaria; ⟨n,n⟩ are not, and ⟨s,s⟩ die from sickle cell anemia. But heterozygotes cannot take over the population, because of the mechanics of genetic reproduction. Here, frequency-dependent selection re-emerges at the level of the gene.

Mendelian genetics is a mathematical model (or a set of mathematical models). In 1953 Watson and Crick opened the door to understanding the physical basis of genetics at the molecular level. The dynamics of genetic replication, recombination, mutation, and expression can and are being studied at this level. Already the study of molecular genetics has provided many surprises such as the action of retroviruses that can transcribe their own information into the DNA of the host cell, the fact that bacteria can exchange genetic information, and the existence of plasmids—rings of DNA in bacteria that exist and replicate independently of the bacterial chromosome. With the theory of largely adaptively neutral mutations (Kimura 1968, 1983), molecular genetics has also provided the basis of a molecular evolutionary clock which can be used to reconstruct evolutionary history.

The history of the theory of biological evolution has been—and will continue to be—a story of the gradual amplification of Darwin's insights by empirical research. The three essential factors in Darwin's account—natural variation, differential reproduction, and inheritance—remain at the center of the theory, but in a form that Darwin could not have imagined.

By social evolution, we mean not the evolution of sociality, which may be one aspect of biological evolution, but rather a social process in some way analogous to the process of biological evolution. But what is the analogy to fix upon? Different thinkers have had different aspects of biological evolution in mind (and sometimes, different conceptions of the nature of the biological process). In its most minimal sense, social and cultural evolution can just be thought of as social and cultural change. So interpreted, the theory of social evolution would be such a truism as to lack all interest. Are there deeper and more informative analogies that are worth exploring?

Some thinkers have thought of ideas that can spread through, or disappear from, a population as analogous to genes (Campbell 1965, Cavalli-Sforza and Feldman 1981, Lumsden and Wilson 1981, Boyd and Richerson 1985). These *memes* (Dawkins 1976) can recombine in various ways. (A comprehensive discussion of this topic would have to range from nineteenth century associationalist psychology to modern genetic programming.) They can replicate themselves as they spread from mind to mind. They can be thought of as having a sort of Darwinian fitness in terms of the replicator process.

Others have focused on adaptive dynamics as the appropriate analogy. Darwinian evolutionary models fall, more or less, within a broad class of adaptive models which also includes various kinds of learning models. In psychological theories of reinforcement learning, if it assumed that there is some 'reward' that is reinforcing, and the theory models the dynamics of behavior under various learning schedules. (The determination of what sort of thing is reinforcing is empirical, but extrinsic to the theory.)

We might then model various dynamical processes of social learning or adaptation instead of trying to model social evolution so as to have a close resemblance to biological evolution.

One common way for individuals to adapt in a social setting is to imitate others whom they observe to be successful according to some generally shared criterion of success. It is a remarkable fact that the replicator dynamics, introduced by Taylor and Jonker (1978) as a mathematical model of differential reproduction, has been derived from a number of different models of learning by imitation in a large population (Binmore et al. 1995, Björnerstedt and Weibull 1996, Schlag 1998). The replicator dynamics also has deep connections with classical psychological models of positive reinforcement learning (Borgers and Sarin 1997).

The replicator dynamics is a member of the broader class of adaptive dynamics according to which all behaviors with greater than average success increase their population proportion, which is a subset of the still broader class according to which the most successful behavior increases its population proportion. These broader classes of adaptive dynamics may also come into play in theories of social evolution (Weibull 1995, Samuelson 1997).

Populations may not be large, and encounters may not be random but rather structured. Learning by imitation in such settings may lead to different results than it would lead to in large random mixing populations, just as biological evolution may be affected by population structure and population size (Hamilton 1964, Axelrod and Hamilton 1981, Frank 1998, Sober and Wilson 1998, Epstein 1998, Lindgren and Nordahl 1994, Nowak and May 1993, Grim 1996, Pollack 1989, Anderlini and Ianni 1997). It is of particular interest that population structure can favor evolution or learning of cooperative behavior in interactions where it would not survive in large random mixing populations.

Classical reinforcement learning is the kind of learning that demands the least thought. A more sophisticated learner might use some kind of inductive reasoning to form beliefs about the aspects of the environment that, together with her actions, influence payoffs and then act optimally according to these beliefs. If the relevant aspects of the environment include the actions to be taken by other agents, then the reasoning involved may be complicated, depending on the strategic sophistication of the agents and the character of the payoffs. (I am trying to predict what she will do. Is she trying to predict what I will do? If so

what will she predict and what will she do and what is my best response? But might she be imagining that I am thinking this way? And so forth.)

Here there is a whole spectrum of learning models, whose appropriateness depends on the inductive and strategic sophistication of the agents being modeled (Fudenberg and Levine 1998).

If small probabilistic perturbations are added to deterministic dynamics of learning or evolution, then a new kind of equilibrium concept is appropriate, that of a *stochastically stable* equilibrium. In some cases, nature of the underlying deterministic dynamics does not matter, and the stochastically stable equilibrium is determined by the payoff structure (Foster and Young 1990, Kandori et al. 1993). A small population with local interaction may reach stochastically stable equilibrium much faster than a large population with random interactions (Ellison 1993).

The stage of development of the theory of social evolution is comparable to the earlier stages of the theory of biological evolution. There is no analog of molecular biology in view for social evolution. Rather, there are various mathematical models of learning that can be put to empirical test in special domains of application (Suppes and Atkinson 1960, Cavalli-Sforza et al. 1982, Hewlett and Cavalli-Sforza 1986, Roth and Erev 1995, van Huyck et al. 1995, Friedman 1996, Camerer and Ho 1998, 1999).

The processes of biological evolution and social evolution can interact (Cavalli-Sforza and Feldman 1981, Boyd and Richerson 1985, Crawford 1989). The biological evolution of humans as a social species sets the stage upon which social evolution is carried out. Social structure can feed back to biological evolution by imposing or modifying selective pressures. Basic cultural norms, for instance, the prevailing system of agriculture in a society, can be influenced by the biological evolution of other species. As both Darwin and Mendel observed, they can impose selective pressures on domesticated plants and animals. When cultural changes allow humans to survive in new or modified environments, selective pressures on humans themselves can be modified, as emphasized by Baldwin (1902). As humans continue to transform the planet, the biological effects of human cultural evolution become more and more extensive.

See also: Conventions and Norms: Philosophical Aspects; Cultural Evolution: Overview; Darwin, Charles Robert (1809–82); Darwinism; Darwinism: Social; Evolution, History of; Evolutionary Selection, Levels of: Group versus Individual; Evolutionary Theory, Structure of; Evolutionism, Including Social Darwinism; Game Theory: Noncooperative Games; Prisoner's Dilemma, One-shot and Iterated; Social Behavior (Emergent), Computer Models of; Social Evolution: Overview; Social Evolution, Sociology of; Sociobiology: Philosophical Aspects

Bibliography

Anderlini L, Ianni A 1997 Learning on a torus. In: Bicchieri C, Jeffrey R, Skyrms B (eds.) *The Dynamics of Norms*. Cambridge University Press, New York

Axelrod R, Hamilton W D 1981 The evolution of cooperation. *Science* **211**: 1390–6

Baldwin J M 1902 *Development and Evolution: Including Psychophysical Evolution, Evolution by Orthoplasty, and the Theory of Genetic Modes*. MacMillan, New York

Binmore K J, Gale J, Samuelson L 1995 Learning to be imperfect: the ultimatum game. *Games and Economic Behavior* **8**: 56–90

Björnerstedt J, Weibull J 1996 Nash equilibrium and evolution by imitation. In: Arrow K, Columbotto E, Parlman M, Schmidt C (eds.) *The Rational Foundations of Economic Behavior*. St. Martin's Press, New York, pp. 155–71

Borgers T, Sarin R 1997 learning through reinforcement and the replicator dynamics. *Journal of Economic Theory* **77**: 1–14

Blume A, DeJong D V, Kim Y G, Sprinkle G 1998 Experimental evidence on the evolution of meaning in messages in sender–receiver games. *American Economic Review* **88**: 1323–40

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Camerer C, Ho T 1998 Experience weighted attraction learning in coordination games: probability rules, heterogeneity and time-variation. *Journal of Mathematical Psychology* **42**: 305–26

Camerer C, Ho T 1999 Experience-weighted attraction learning in normal form games. *Econometrica* **67**: 827–74

Campbell D 1965 Variation and selective retention in sociocultural evolution. In: Barringer H R, Blanksten G, Mack R M (eds.) *Social Change in Developing Areas: A Reinterpritation of Evolutionary Theories*. Schenkman, Cambridge, MA

Cavalli-Sforza L L, Feldman M W 1981 *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press, Princeton, NJ

Cavalli-Sforza L L, Feldman M W, Chen K H, Dornbush S 1982 Theory and observation in cultural transmission. *Science* **218**: 19–27

Crawford V P 1989 Learning and mixed strategy equilibria in evolutionary games. *Journal of Theoretical Biology* **140**: 537–50

Darwin C 1859 *On the Origin of the Species by Means of Natural Selection*. J Murray, London

Darwin C 1874 *The Descent of Man and Selection in relation to Sex*. John Murray, London

Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford, UK

Dobzhanski T 1937 *Genetics and the Origin of Species*. Columbia University Press, New York

Ellison G 1993 Learning, local interaction and coordination. *Econometrica* **61**: 1047–71

Epstein J 1998 Zones of cooperation in demographic prisoner's dilemma. *Complexity* **4**: 36–48

Fisher R A 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK

Foster D, Young P 1990 Stochastic evolutionary game dynamics. *Theoretical Population Biology* **38**: 219–32

Frank S A 1998 *Foundations of Social Evolution*. Princeton University Press, Princeton, NJ

Friedman D 1996 Equilibrium in evolutionary games: some experimental results. *Economic Journal* **106**: 1–25

Fudenberg D, Levine D K 1998 *The Theory of Learning in Games*. MIT Press, Cambridge, MA

Grim P 1996 Spatialization and greater generosity in the stochastic prisoner's dilemma. *Biosystems* **37**: 3–17
Haldane J 1932 *The Causes of Evolution*. Longmans, Green and Co, New York
Hamilton W D 1964 The genetical evolution of social behaviour. *Journal of Theoretical Biology* **7**: 1–52
Hewlett B S, Cavali-Sforza L L 1986 Cultural transmission among Aka pygmies. *American Anthropologist* **88**: 922–34
Hofbauer J, Sigmund K 1988 *The Theory of Evolution and Dynamical Systems*. Cambridge University Press, Cambridge, UK
van Huyck J, Battalio C, Mathur S, Ortmann A, van Huyck P P 1995 On the origin of convention: evidence from symmetric bargaining games. *International Journal of Game Theory* **24**: 187–212
Kandori M, Malaith G J, Rob R 1993 Learning, mutation, and long run equilibria in games. *Econometrica* **61**: 29–56
Kimura M 1968 Evolutionary rate at the molecular level. *Nature* **217**: 624–6
Kimura M 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK
Lindgren K, Nordahl M G 1994 Evolutionary dynamics in spatial games. *Physica D* **75**: 292–309
Lumsden C J, Wilson E O 1981 *Genes, Mind and Culture*. Harvard University Press, Cambridge, MA
Mendel G 1866 *Versuche über Pflanzenhybriden*. Verhandlungen des Naturforscher-Vereins in Brünn. Tr. as [*Experiments in Plant Hybridisation*] 1960. Harvard University Press, Cambridge, MA
Nowak M, May R 1993 The spatial dimensions of evolution. *International Journal of Bifurcation and Chaos* **3**: 35–78
Pollack G 1989 Evolutionary stability of a viscous lattice. *Social Networks* **11**: 175–212
Roth A E, Erev I 1995 Learning in extensive form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* **8**: 164–212
Samuelson L 1997 *Evolutionary Games and Equilibrium Selection*. MIT Press, Cambridge, MA
Schlag K-H 1998 Why imitate and if so, how? A boundedly rational approach to many multi-armed bandits. *Journal of Economic Theory* **78**: 130–56
Sober E, Wilson D S 1998 *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, Cambridge, MA
Suppes P, Atkinson R C 1960 *Markov Learning Models for Multiperson Interactions*. Stanford University Press, Stanford, CA
Taylor P D, Jonker L-B 1978 Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* **40**: 145–56
Wright S 1968, 1969, 1977, 1978 *Evolution and the Genetics of Populations*. University of Chicago Press, Chicago, 4 Vols.
Weibull J W 1995 *Evolutionary Game Theory*. MIT Press, Cambridge, MA

B. Skyrms

Evolution of Cognition: An Adaptationist Perspective

Human and animal cognition is complex and functionally organized. The theory that best accounts for the existence of organized, functional complexity in the design of living things is Darwin's theory of evolution by natural selection. Because developmental processes have reproductive consequences for organisms, cognitive development is also a target of natural selection, leading to the evolution of adaptive developmental trajectories. Hence, developmental and evolutionary accounts of cognition are complementary and not mutually exclusive. The adaptive organization of cognition should be studied from a functional, ecological perspective, taking the species' natural history into account, and using a variety of methods, including laboratory studies of cognition, neuropsychological dissociations, developmental studies, and phylogenetic/comparative research.

1. The Functional Complexity of the Brain and its Processes

One of the characteristic features of living things is their complex, functionally organized structure—living things appear to be designed to solve the problems confronting them in ways that are often ingenious, elegant, and economical. Humans are no different. Indeed, the human brain is probably the paramount example of a complex, functionally organized structure in the biological world. Viewed macroscopically, there is nothing particularly remarkable about the human brain. A little more than a liter in volume, it simply looks like a wrinkled mass of tissue. However, the brain, as is now generally recognized, is a cognitive organ. It represents, stores, and processes information, and it is in the information processing abilities of the brain that its functional complexity resides. But how did the cognitive architecture of the brain arise?

One of the great puzzles of biological life is the origin of the kind of complex, functional organization that the brain so clearly illustrates. According to the laws of physics, organized systems should tend toward disorder; but biological systems are not only highly ordered, they seem to increase in complexity with time (Dawkins 1986). Ordered structures can persist, however, if they reproduce themselves faster than they fall apart. By reproducing themselves, biological entities initiate a process—evolution through natural selection—that Darwin recognized could account for ever increasing refinement of complex, functional organization. Features of an organism's design that confer an advantage in acquiring resources for reproduction will enhance their own spread through differential reproduction. The slow accretion of such incremental improvements in the design of organisms can eventually result in the evolution of the ingenious, elegant and economical structures (called adaptations) that are so common in living things.

Because organisms in all their complexity develop from single cells, some might argue that the complex

functional organization of the brain requires a developmental explanation, rather than an evolutionary one. However, the development of the brain and learning are exceedingly complicated processes, just as much in need of an evolutionary explanation as the final product, adult cognition. Development can evolve, precisely because it is repeated from generation to generation with variation in developmental trajectories, where some lead to fitness-enhancing improvements and some to less fit products. Natural selection will act on these developmental trajectories to produce ever more adaptive ones.

Moreover, an evolutionary view of cognitive development leads one to expect the same economy and efficiency of design in developmental mechanisms (Elman et al. 1996). From an evolutionary point of view, what is important is whether an adaptation develops reliably, that is, whether the trait is reproduced reliably, in the sorts of environment typical of the species. If this can be accomplished by exploiting the information structure of the environment, then there is reason to believe that natural selection would favor the use of such information (see *Modularity versus Interactive Processing, Psychology of*). Evolution through natural selection is not synonymous with the 'genetic coding' of functional architecture. Darwin, after all, was able to formulate his theory of natural selection without any knowledge of genetic coding.

A corollary of this view is that what often appears to be cognitive 'flexibility' can, in fact, be the result of selection for developmental efficiency. Evolved mechanisms may be designed to tolerate some kinds of variation in developmental inputs, particularly variation that is not functionally relevant to the adaptive problem the mechanism solves. Consider language, for example. Modern English, Japanese, and Quechua are all evolutionarily novel inputs to the mechanisms for acquiring and processing language, yet each of these modern languages solves the problem of communication effectively, for which language abilities were presumably selected. Hence, viewed from one perspective, the developmental mechanisms are flexible in being able to accommodate a wide range of novel inputs, but viewed from another angle, they have functional design constraints that all modern languages satisfy. From an adaptationist point of view, what is important is that adults develop reliably functional competence in linguistic communication (see *Language Acquisition*).

Rather than offering mutually exclusive accounts mental structure, development and evolution complement each other. Development provides a proximate explanation for how mental complexity is constructed during the course of a lifetime, while evolution through natural selection explains why such complex developmental processes arose over evolutionary time. Recently, there has been considerable interest in how complexity can emerge through interactions of various kinds between evolutionary and developmental processes. Such emergent phenomena admit of various kinds of explanation. Some theorists prefer to locate the explanation for emergent complexity in the local interactions that produce it. However, it is unclear whether complexity that arises in the absence of natural selection is organized functionally in a way that provides an adaptive fit between the organism and its environment. For this reason, many theorists prefer to focus on natural selection as the cause of functionally organized complexity. Adaptations are not merely complex traits, but are also organized functionally. This functional organization is observed at various levels of biological organization: biochemical, physiological, anatomical, and cognitive, with developmental mediation playing a greater role at higher levels of organization.

Regardless of the level of organization that is involved, what matters is that a trait in some way enhances its own reproduction. In the course of competing for resources, such as food, shelter, and mates, the individual is faced with many different challenges that vary with the ecological and social environments encountered. It is by virtue of its ability to solve such adaptive problems that an adaptation furthers its own reproduction, and hence is said to have the function of solving that problem (Williams 1966). Mobile species, in particular, are highly dependent upon cognitive adaptations that can interpret and integrate information in order to respond adaptively to a heterogeneous environment. The animal kingdom is filled with many striking examples of cognitive adaptation, such as the ability of Clark's nutcrackers (*Nucifraga columbiana*) and other food-caching birds to recall the location of thousands of widely scattered food stores after a delay of several months (Gallistel 1990, pp. 155–8; see also *Animal Cognition*).

2. *Consequences of Cognitive Evolution*

The evolutionary account of adaptation has three important consequences for cognitive functioning. The first is that traits must be evaluated from a practical, ecological perspective (see *Decision Making (Naturalistic), Psychology of*). Traits with no impact on an organism's reproductive success cannot evolve through natural selection. From an evolutionary point of view, one would only expect cognitive adaptations for the sorts of problem that could conceivably have an impact on reproductive success. Typically, this will mean traits that have a real impact on the organism's interactions with the physical, biological, and social world upon which successful reproduction ultimately depends. Natural selection, therefore, gives rise to practical cognitive mechanisms that can solve the real world problems confronting the organism. While the ability to solve evolutionarily novel tasks may indeed be beneficial, it is important to keep in mind that

performance on artificial tasks, such as laboratory tasks often used in cognitive psychology, may depend on the degree to which these tasks reflect aspects of adaptive problems faced by organisms during their evolutionary history.

The second important consequence is that, in order to understand the adaptive organization of cognition, one needs to consider the past history of the species. Evolution is a slow, historical process. Complex adaptations take many generations to evolve, so the evolved design of an organism will reflect the sorts of problem that confronted ancestral members of the species and not necessarily the problems confronting the species currently. Hence, it is often useful to make a distinction between the proper and actual function of a cognitive adaptation (Millikan 1984). The proper function of an adaptation is the function that it served in ancestral environments, by virtue of which it persisted and proliferated over successive generations. The actual function of an adaptation is the use that it serves currently in a particular individual. The actual function of a mechanism may or may not correspond to its proper function. Consider object segregation in vision. The properties of light have changed very little, if at all, since the mechanisms of object segregation began to evolve in ancestral vertebrates. Hence, there is every reason to believe that object segregation in modern humans functions much as it always has: the proper and actual functions of the mechanisms are identical. Compare the situation with language (see *Language Acquisition*). The best available evidence suggests that languages evolved as a spoken form of communication (Pinker 1994). Written language, on the other hand, shows evidence of being a relatively recent cultural invention with an origin dating to only a few thousand years ago. Reading and writing employ some of the same cognitive processes as spoken language, but this does not make reading and writing part of the proper function of these mechanisms. Reading, writing, and speech are all part of the actual function of linguistic processes, while speech alone constitutes the proper function of the mechanism. The distinction between proper and actual functions is important to bear in mind, because the structure of an evolved mechanism is best explained by its proper function, not its actual function, and mechanisms which may appear to be maladaptive in present environments might not have been so in past ones.

The third important consequence is that cognitive adaptations are exhibited universally among individuals with an identical history of selection. Although individual variation is crucial for the evolutionary process, natural selection systematically sifts through this variation, retaining those variants that are relatively beneficial and displacing those variants that are less adapted to their environment. Eventually all variation with reproductive consequences is eliminated from the species, leading to universal, species-typical cognitive adaptations, and a residue of variation with no systematic reproductive consequences. Hence, there are three basic products of the evolutionary process that are expressed in the individual: adaptations, by-products, and random effects (Buss et al. 1998). Adaptations are complex, functionally organized structures that evolved as the accumulated product of natural selection acting on variants solving a specific problem (see). Adaptations often appear as though they were designed to solve specific problems, and are therefore said to show evidence of special design (Williams 1966). By-products are incidental features of adaptations that played no role in an adaptation's selective history. Because they derive their structure from adaptations, by-products may also appear complex and functionally organized. Random effects are random features of an organism with no functional organization, no history of systematic selection, and tending to show individual differences in their expression.

3. *Methods and Evidence in the Study of Cognitive Evolution*

The hallmark of adaptation by natural selection is functional specialization, or special design. In general, the principles of adaptive design are similar to the principles of good engineering: efficiency, precision, reliability, and so on. Just as the features of a well-engineered machine can be distinguished from the products of chance, so can the features of evolved adaptations (Williams 1966). To demonstrate the existence of an adaptation it is necessary to demonstrate design for a specific adaptive function, but not just complex organization *per se*, because it is important to distinguish adaptations from by-products. Some of the sources of evidence for special design are the following.

3.1 *Laboratory Studies of Cognition*

It is useful to study an organism in its natural environment in order to understand the adaptive problems that it faces and the selection pressures operating on it. In this respect, observational studies are very important, but they have their limitations when it comes to elucidating the nature of the psychological mechanisms underlying observable behavior. For this purpose, controlled laboratory studies using methods from cognitive psychology are ideal (see *Inferences, Psychology of*). For example, in order to test the proposal that human resource acquisition is sensitive to variability in returns, as predicted by evolutionary theories of foraging behavior (Stephens 1981). Rode et al. (1999) employed a standard procedure used by cognitive psychologists studying people's decision making. They offered people a choice between two gambles with the same expected return, but with more or less variability in returns. By introducing a minimal requirement that people had to

satisfy in order to win any money, Rode et al. (1999) were able to manipulate whether people selected the more or less variable option in a manner predicted by Stephens' (1981) risk sensitive foraging theory, but not by standard models of human decision-making. These laboratory studies complement nicely the observational field studies performed by human behavioral ecologists studying traditional foraging peoples (e.g., Smith and Winterhalder 1992).

3.2 Neuropsychological Dissociations

If the mind of an organism is composed of different, functionally specific adaptations, then these adaptations should, in principle, be dissociable: a particular competence may be disabled, either temporarily or permanently, while others are left largely intact. Such dissociations can be an important source of evidence for special design, because while it can be difficult for the observer to identify distinct cognitive abilities in the behavior of normally functioning individuals, the absence of a specific competence can bring its nature and boundaries into strong relief, against a backdrop of intact skills. Many such dissociations have been identified in humans, and the study of patients with particular kinds of neurological impairment has been a valuable source of data for testing hypotheses about the functional building blocks of the mind. For example, the study of autism, in which social cognition is impaired relative to nonsocial competences, has provided evidence for the existence of a variety of cognitive mechanisms whose hypothesized evolved function is to interpret and predict the behavior of others (Baron-Cohen 1995; see *Theory of Mind*).

3.3 Developmental Studies

One reason among many for studying cognitive development is that potentially it can illuminate the functional components that underlie adult competences. Many such competences, although they appear to function seamlessly in adults, are in fact composed of separate components whose individual operation can be observed more readily in children. For example, many mammalian species orient spatially using the large-scale structure of the environment (Gallistel 1990; see *Animal Cognition*). Adult humans use many spatial cues to orient themselves during navigation, obscuring the separate subcomponents of this competence. However, this core mammalian competence is observed more readily in young children (Hermer and Spelke 1994; see *Infant Development: Physical and Social Cognition*).

3.4 Phylogenetic/Comparative Approaches

Even if one's interest is in human cognition, there are still valuable insights to be gained from the study of nonhuman species. Such studies provide yet another way of isolating the adaptive components of cognition. For example, it has been claimed that the human adult number competence—specifically, the ability to count and perform other manipulations with numbers—is derived from a core number competence, observable in infants (Carey and Spelke 1994 see). However, such a core competence has also been found in Rhesus monkeys (*Macaca mulatta*) (Hauser et al. 1996), suggesting that the competence seen in human adults has other components as well.

4. Conclusion: Evolutionary Theory as Computational Theorizing

In his classic book on vision, Marr (1982) stressed the importance of beginning investigations of complex systems with a computational theory—a task analysis of the problem to be solved (see *Vision, Psychology of*): What are the available inputs? What are the desired outputs? What are the constraints on the system? The answers to these questions serve as an important guide when formulating hypotheses about the cognitive mechanisms that actually solve the task. Although Marr's theory has been highly influential in vision research where there is some intuitive appreciation of the functions of mechanisms in the visual system, his theory has not been applied as readily in other areas of psychology. The merging of Marr's approach with insights from evolutionary theory offers promise for many areas of psychology outside perception. Evolutionary biology is a rich source of task analyses for a variety of adaptive problems that require cognitive solutions, in many practical domains. Moreover, cognitive psychology is a rich source of theoretical constructs for describing the psychological mechanisms that can solve such problems. Hence, evolutionary theorizing and cognitive psychology are well suited to play complementary roles in the study of the mind's design.

See also: Human Cognition, Evolution of; Intelligence, Evolution of; Psychological Development: Ethological and Evolutionary Approaches

Bibliography

Barkow J, Cosmides L, Tooby J (eds.) 1992 *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, New York

Baron-Cohen S 1995 *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, Cambridge, MA

Buss D, Haselton M, Shackelford T, Bleske A, Wakefield J 1998 Adaptations, exaptations, and spandrels. *American Psychologist* **53**: 533–48

Carey S, Spelke E 1994 Domain-specific knowledge and conceptual change. In: Hirschfeld L, Gelman S (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, New York, pp. 169–200

Cummins D, Allen C (eds.) 1998 *The Evolution of Mind*. Oxford University Press, New York

Dawkins R 1986 *The Blind Watchmaker*. Norton, New York
Elman J, Bates E, Johnson M, Karmiloff-Smith A, Parisi D, Plunkett K 1996 *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press, Cambridge, MA
Gallistel C R 1990 *The Organization of Learning*. MIT Press, Cambridge, MA
Hauser M D, Macneilage P, Ware M 1996 Numerical representations in primates. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 1514–7
Hermer L, Spelke S 1994 A geometric process for spatial reorientation in young children. *Nature* **370**: 57–9
Marr D 1982 *Vision*. Freeman, New York
Millikan R 1984 *Language, Thought, and other Biological Categories*. MIT Press, Cambridge, MA
Pinker S 1994 *The Language Instinct*. Morrow, New York
Pinker S 1997 *How the Mind Works*. Norton, New York
Rode C, Cosmides L, Hell W, Tooby J 1999 When and why do people avoid unknown probabilities in decisions under uncertainty? Testing some predictions from optimal foraging theory. *Cognition* **72**: 269–304
Smith E A, Winterhalder B (eds.) 1992 *Evolutionary Ecology and Human Behavior*. de Gruyter, New York
Stephens D W 1981 The logic of risk-sensitive foraging preferences. *Animal Behaviour* **29**: 628–9
Williams G C 1966 *Adaptation and Natural Selection*. Princeton University Press, Princeton, NJ

L. Fiddick and H. C. Barrett

Evolution: Optimization

A claim for 'optimization' has one meaning to the various kinds of scientists who study evolution. It is a claim that a trait has evolved or is evolving 'optimal' characteristics that enable it to perform better than counterparts with different characteristics. A biologist studying, say, the body size of a species might claim it to be the body size that results in a higher number of offspring than that produced by any other plausible body size. An economist studying, say, the pattern of resource use in a given company might claim it to be the pattern that maximizes the company's market share. What unifies these examples is that there is an implied process of selection among competing entities, with the competitor who performs better at a given task receiving a greater reward.

The focus here is on optimization in biological evolution; further information on the other contexts in which optimization may be important can be found in Schoemaker (1991) and references therein.

1. *The Term 'Optimization'*

When a biologist makes a claim about optimization (e.g., Tyler and Gilliam 1995), it is most often a claim about natural selection among individuals within populations; such a claim generally (but not necessarily) excludes selection at the level of the population or species. More controversially, it usually excludes 'random' or nondirectional forces, such as genetic drift (random sampling of gametes and individuals in populations), from having any significant affect on the trait's evolution.

A claim about optimization also always has an explicit or implicit assumption about trait inheritance, a common one being that the trait values of parents and offspring are identical or nearly so. This, along with the reproductive advantage of an optimal trait, implies that the trait will become fixed in the population, all other things being equal.

2. *The Controversy Over 'Optimization'*

What *does not* engender controversy among evolutionary biologists is the overall theoretical conception of the process of optimization. Most evolutionary biologists agree that one can make predictions of optima for traits and that natural selection could plausibly act to cause evolution in the direction of these optima. In addition, most of these biologists agree that at least some aspects of the process of optimization occur in nature (see Travis 1989 for more details).

What *does* engender controversy can be better understood by describing the two most important contexts in which optimization is invoked.

2.1 *Continuously-distributed Traits and 'Optimization'*

The first context involves traits having a more-or-less continuous distribution, usually with one central peak. Many traits have such a distribution; often it is statistically indistinguishable from a normal distribution. So, for example, Van Valen and Mellin (1967) showed in their study of human babies in New York City that birth weight has such a bell-shaped distribution. Optimizing selection is invoked in such examples to explain the peaked character of the trait distribution, the hypothesis being that a central trait value performs better than any other trait value. In their study, Van Valen and Mellin showed that the probability of perinatal death is minimal at a central weight (about 3.6 kg for males and about 3.8 kg for females). Ostensibly, this is because more deviant weights are associated with specific physiological deficits that are congenitally or environmentally based, their consequence being a higher risk of death. Subsequent work has revealed that the length of the gestation period also affects the probability of perinatal death (see Schluter and Nychka 1994.)

What then explains the presence of the continuous trait distribution? After all, if there is one best birth weight, why isn't there at most a small range of birth weights around this best weight? More generally, why do any traits undergoing trait optimization have continuous trait distributions? It is in answering this question that controversy has arisen; there are at least

three different claims as to why such distributions occur.

The first claim is that the genetic 'details' of trait determination impede the action of optimizing natural selection. What this means is that the nature of trait determination and the Mendelian 'breakdown' of the trait when reproduction occurs act together to generate a continuous trait distribution in the next generation. So, for example, imagine that natural selection is such that only individuals with the optimal trait live to the start of reproduction, the result being that there is only one type of trait remaining in the population at that time. If one assumes that the optimal trait does not breed true and is, say, the result of an individual being a heterozygote at all or most of the loci affecting the trait, then newborn offspring will manifest a continuous trait distribution. This is a necessary consequence of the normal segregation of alleles during gamete production and the subsequent union of gametes to form the offspring of the next generation.

The underlying presumption of this claim about genetic details is that they are an unchangeable impediment to the action of natural selection even when the intensity of selection is strong. This could happen even if the optimal trait is not associated with being a heterozygote at many loci. One possible (but not necessary) consequence is that a stable amount of genetic variation can be maintained in a population, as a result of the balance between optimizing natural selection, which will tend to eliminate genetic variation, and the genetic details of the trait (including mutation), which can act to preserve or increase it. Such a balance has been invoked as an explanation for the high levels of genetic variation seen for some traits in some populations (see Lande 1988 and Turelli 1988 for further discussion.)

The second claim as to why continuous trait distributions occur is that natural selection is not and never will be strong enough or constant enough to eliminate or nearly eliminate nonoptimal traits from the population. However, the presumption here is that this would occur if selection were stronger; genetic details would not get in the way.

The third claim is that it is just a matter of time until the continuous distribution disappears and an optimal or near-optimal trait predominates in the population, the notion being that the current data reveal only an intermediate state of a long process. From this viewpoint, natural selection is powerful and sustained enough to cause trait optima to evolve and genetic details will not ultimately prevent this from happening, as the process of mutation will ultimately provide the appropriate kind of genetic variation to allow an optimal trait to become fixed. So, for example, a mutation that when homozygous produces an optimal trait might arise in a population in which the trait has been manifested only by heterozygotes. The true-breeding optimal trait may increase in frequency and displace the heterozygotes (see Hammerstein 1996 and Eshel and Feldman 2001 for further discussion of the dynamics and outcome of this kind of long-term evolution.)

2.2 *Fixed Traits and 'Optimization'*

The second context in which optimizing selection is invoked concerns traits that are constant in a given species. For example, functional anatomists speak of *the* horse's foot as having evolved for the task of running on uneven ground at high speed (e.g., see Hildebrand 1985) and behavioral ecologists speak of *the* dance language of honey bees (e.g., see Wilson 1971). In such instances, the trait is regarded as having a single state, which is often taken to be the result of past optimization; in other words, the trait is the way it is because it performs better than all reasonable alternatives and this superior performance has resulted in the past elimination of alternative forms of the trait. The underlying attitude is very often that optimality 'must be,' given the importance of the trait to the organism. From this viewpoint, the potential for understanding how the trait is optimal is limited only by the industry and ingenuity of the investigator (cf., Cain 1989). Such an 'adaptationist' approach is reasonable enough in theory although it is easy to see the potential for circularity in practice because discrepancies between the predictions of a model used to describe the optimal trait and the data can always be attributed to an incomplete or flawed model.

2.3 *Types of Optimality Models*

All optimality models make a prediction about the trait of *an* individual. Beyond this common aspect, there is as much variety in the features of optimality models as there is in the traits studied. However, one can recognize two general classes of such models. The first class concerns traits for which one can determine the performance of a trait by reference only to the functioning of the trait itself. Traits still affect one another in that an optimal trait will outperform all others and thereby cause their selective elimination from the population. But the performance of any trait is the same in isolation as it is in a population. There are many optimality models of this kind in every area of evolutionary biology. For example, when constructing an optimality model of the shape of a horse's foot, a functional anatomist will likely regard the performance of the foot as independent of the performance of any other individual's foot in the population. Similarly, an ecologist might assume that the evolutionary consequence of a given number of offspring produced by an individual is independent of the numbers produced by other individuals.

In contrast, the second class of optimality model involves traits whose performance is inherently related to the performance of other individuals because it is dependent on the frequency or density of various traits in the population. Such dependency arises naturally for many traits, especially for those relating to the sexual production of offspring. A classic example is the evolution of the sex ratio, the relative ratio of males and females in the population. The evolutionary fate of a female producing, say, all or nearly all daughters depends very much on whether all of the other females in the population do this as well, as opposed to producing, say, equal numbers of daughters and sons. In the former instance, the population could go extinct; in the latter instance, even a female producing only daughters may survive. If the tendency to produce given numbers of males and females is autosomally inherited and sons and daughters are equally costly in terms of energy, one can show that a female-biased sex ratio will be optimal if the population is subdivided and females tend to migrate more than males (Hamilton 1967). However, in an infinite randomly-mating population such a sex ratio is not optimal; instead, a female producing equal numbers of daughters and sons will be optimal as long as equal numbers of males and females are not produced by the population as a whole. Finally, when equal numbers are produced, there is no optimal sex ratio; all sex ratios are evolutionarily equivalent. This variety of outcomes is typical for interactive traits.

Traits like the sex ratio are often analyzed with a model in which one determines the trait which when common prevents the invasion of rare mutants into the population. Such a trait is taken to be optimal in the sense that individuals possessing it outperform all other traits when the population has this configuration (this does not necessarily imply that the trait has an advantage for all other configurations.) This evolutionarily stable strategy (ESS) approach has received considerable attention since about 1980, with extensive elaboration of the theory in formal game-theoretic terms and extensive application of models to the data (see Charnov 1982, Maynard Smith 1982, and especially Hines 1987 for examples and further details.)

For other traits there is no necessary choice about what type of optimality model to use and the choice depends upon the investigator's prior assumptions about whether and how individuals interact with one another.

3. *Optimization and the Debate about the Adaptationists' Claim*

To proponents of the adaptationist approach, such as Parker and Maynard Smith (1990), the danger of circularity in claims of optimality is real but of little concern if one is careful and rigorous when deriving the model and if one is scrupulous when testing it. To opponents of this approach, such as Gould and Lewontin (1979), it is inherently flawed because the hypothesis of optimality is never challenged or tested, as a failed optimality model is simply replaced with another optimality model. One alternative they suggest is to formulate predictions from nonadaptive or nonoptimal hypotheses, which can compete with the predictions of the optimality model. For example, one might claim that a trait has not evolved in response to the present environment, but instead evolved in a past environment; a plausible ancestral trait would then be used as a predictor of the trait and its predictive performance would be compared with that of an optimality model. Adherents of this 'pluralist' approach often claim that if the predictions of one or more of these nonoptimal competing hypotheses explain the data as well as the optimality model does, the former should be given precedence as an explanation for the trait (cf. Lewin 1980).

There are substantive differences between these viewpoints and as a result, proponents and opponents of the adaptationist approach have traded claims and counterclaims for decades; this is a central debate in evolutionary biology (see Orzack and Sober 1994, 2001). Adaptationists generally regard natural selection to be the only important influence on trait evolution, while pluralists regard natural selection to be just one of several such influences, the others being genetic drift, constraints engendered by genetic details, and phylogenetic inertia (the carryover of traits from a species' ancestors; see Harvey and Pagel 1991 for further details). In an important sense, however, adaptationist and pluralist claims about trait evolution have a common feature—both start with a *premise* as to the truth or falsity of the adaptationist approach. To adaptationists, the near-certainty of optimality follows necessarily from what they view as the basic and unquestionable Darwinian dynamic of variation and selection; to pluralists, the near-certainty of nonoptimality follows necessarily from what they view as the inherent genetic and developmental complexity of the organism.

One alternative to the reliance on either of these premises is to assess the validity of the adaptationist approach via an ensemble test in which the successes and failures of specific optimality models at predicting trait values and explaining trait variation are tallied (see Orzack and Sober 1994 for further details.) One might regard the adaptationist claim to be true if, say, 90 percent of traits examined could be adequately explained by an optimality model. This ensemble test faces difficult but not insurmountable problems that relate to how traits and models are chosen to be part of the ensemble. Nonetheless, such a test has several virtues, an important one being that the truth or falsity of the adaptationist approach is not a premise; instead, either possibility is a potential *conclusion* of the ensemble test. To this extent, a central debate in evolutionary biology could be resolved on the basis of

data, instead of being left unresolved because of the contradictory premises investigators use as a basis for their analyses.

See also: Adaptation, Fitness, and Evolution; Body, Evolution of; Brain, Evolution of; Darwin, Charles Robert (1809–82); Evolution, History of; Evolutionary Selection, Levels of: Group versus Individual; Evolutionary Theory, Structure of; Human Genome Diversity and its Value; Microevolution; Natural Selection; Optimization in Evolution, Limitations of

Bibliography

Cain A J 1989 The perfection of animals. *Biological Journal of the Linnaean Society* **36**: 3–29

Charnov E L 1982 *The Theory of Sex Allocation*. Princeton University Press, Princeton, NJ

Eshel I, Feldman M W 2001 Optimality and evolutionary stability under short-term and long-term selection. In: Orzack S, Sober E (eds.) *Adaptationism and Optimality*. Cambridge University Press, New York, pp. 161–90

Gould S J, Lewontin R C 1979 The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society London, Series B* **205**: 581–98

Hamilton W D 1967 Extraordinary sex ratios. *Science* **156**: 477–88

Hammerstein P 1996 Darwinian adaptation, population genetics and the streetcar theory of evolution. *Journal of Mathematical Biology* **34**: 511–32

Harvey P H, Pagel M D 1991 *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, UK

Hildebrand M 1985 Walking and running. In: Hildebrand M, Bramble D M, Liem K F, Wake D B (eds.) *Functional Vertebrate Morphology*. Harvard University Press, Cambridge, MA, pp. 38–57

Hines W G S 1987 Evolutionarily stable strategies: A review of basic theory. *Theoretical Population Biology* **31**: 195–272

Lande R 1988 Quantitative genetics and evolutionary theory. In: Weir B S, Eisen E J, Goodman M M, Namkoong G (eds.) *Proceedings of the Second International Conference on Quantitative Genetics*. Sinauer, Sunderland, MA, pp. 71–84

Lewin R 1980 Evolutionary theory under fire. *Science* **210**: 883–7

Maynard Smith J 1982 *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK

Orzack S H, Sober E 1994 Optimality models and the test of adaptationism. *American Naturalist* **143**: 361–80

Orzack S H, Sober E 2001 *Adaptationism and Optimality*. Cambridge University Press, Cambridge, UK

Parker G A, Maynard Smith J 1990 Optimality theory in evolutionary biology. *Nature* **348**: 27–33

Schluter D, Nychka D 1994 Exploring fitness surfaces. *American Naturalist* **143**: 597–616

Schoemaker P J H 1991 The quest for optimality: A positive heuristic of science? *Behavioral and Brain Sciences* **14**: 205–45

Travis J 1989 The role of optimizing selection in natural populations. *Annual Review of Ecology and Systematics* **20**: 279–96

Turelli M 1988 Population genetic models for polygenic variation and evolution. In: Weir B S, Eisen E J, Goodman M M, Namkoong G (eds.) *Proceedings of the Second International Conference on Quantitative Genetics*. Sinauer, Sunderland, MA, pp. 601–18

Tyler J A, Gilliam J F 1995 Ideal free distributions of stream fish: A model and test with minnows, *Rhinicthys atratulus*. *Ecology* **76**: 580–92

Van Valen L, Mellin G W 1967 Selection in natural populations 7. New York babies (Fetal Life Study). *Annals of Human Genetics* **31**: 109–27

Wilson E O 1971 *The Insect Societies*. Harvard University Press, Cambridge, MA

S. H. Orzack

Evolution: Self-organization Theory

Self-organization theory holds that the distinctive characteristics of living systems require explanation in terms of life's emergent properties at different levels of organization. The properties of living systems cannot be understood simply in terms of the Darwinian processes of change in genotypes, associated modification in the fitness of phenotypes, and selective pressures from changing environments. They require explanation in terms of organizing principles expressed at different levels of biological organization, from cells through whole organisms to communities and ecosystems.

1. Historical Background

Self-organization theory is based on the evidence that living systems express distinctive types of intrinsic dynamic order. In the eighteenth century the founder of modern biology, Carolus Linnaeus, regarded his taxonomy as a prolegomenon to a general theory of biology based on the Enlightenment principle which observed that regularities of structure in nature can be explained by underlying generative laws of form. This same idea informed the comparative morphological studies of Goethe, Cuvier, St. Hilaire, Reichert, and Owen in the eighteenth and nineteenth centuries, despite differences of emphasis among them.

A revolution in biological thought occurred when Darwin replaced intrinsic principles of order in organisms by historical and functional contingencies. For Darwin, constraints on morphology came from descent, specifically descent from ancestors whose morphologies were the result of adaptation to particular habitats, realized through the selection of fitter hereditary variants (Mayr 1988).

Within embryology the problem of biological form remained: how does the morphology of the adult organism emerge through growth and transformation

of the egg? During the twentieth century this problem was transformed by the rise of genetics into a new question: what are the changes in gene activity that underlie morphological change in developing organisms? In evolutionary studies a crucial question became, what are the changes in genes that underlie the evolution of different species? Developmental and evolutionary genetics are now converging to provide detailed answers to these questions.

What remains problematic is explaining relationships across levels of biological organization, from patterns of gene activity, to phenotypes (morphology and behavior), to community organization, and to higher-level dynamics such as patterns of species extinctions. The difficulty is that there is a hierarchy of organizational levels. Collapsing these levels down to genes and their products does not provide explanations of the different structures of cells, tissues, organs, and organisms that emerge during development and in the course of evolution. It is here that principles of self-organization in biology have re-emerged in a somewhat different guise from more classical conceptions, sharpening the issues defined earlier and working systematically at different levels of biological order.

The unifying term for this field of study is the sciences of complexity (Cowan et al. 1994). The field focuses on the emergent properties of complex systems during the course of development and evolution (Solé and Goodwin 2000). One of the descriptive terms for these emergent properties is 'order for free' (Kauffman 1995), meaning that the order arises spontaneously, can occur at any level, and is not a result of natural selection. In fact, natural selection can do no more than stabilize emergent order if it is adaptive, a principle of evolutionary theory which is sometimes forgotten. This new perspective offers the possibility of integrating the historical and functional explanatory framework of neo-Darwinism with the 'intrinsic principles' approach of nonlinear dynamics, the heart of the sciences of complexity. The theory of self-organization now provides the possibility of recasting evolutionary principles in consistently dynamic terms, resolving the tensions that have existed between development and evolution, ontogeny and phylogeny.

2. *Development, Behavior, and Life-cycles*

Developing organisms can be understood as self-organizing systems whose morphology emerges from the dynamics of 'morphogenetic fields.' These fields are defined by the coordinated properties of cells and their interactions via signaling molecules, the extracellular matrix, adhesion molecules, cell movement, and so on. The particular patterns that emerge are often not obvious at all from the properties of individual cells and require higher-level concepts to be understood.

This principle can be illustrated by the classic example of the cellular slime mold, *Dictyostelium discoideum*, whose vegetative phase consists of single amoebae that live on bacteria in damp, rotting vegetation, growing and dividing independently of one another. Under conditions of starvation and dessication their metabolism and behavior change so that they begin to interact through pulsatile signals of a metabolite (cyclic adenosine monophosphate) that is released periodically by individual cells and diffuses between them. A cell receiving such a signal emits a pulse itself and moves toward its source, producing dynamic patterns of movement that result in the aggregation of thousands of cells. During this process of aggregation, spiral waves of moving cells are observed. The pattern cannot be explained in terms of individual cell behavior; it requires the higher-level concept of an 'excitable medium' which serves as a morphogenetic field for the slime mold (Goodwin 1994).

Initially uniform patterns of aggregating cells break up into streams, another purely dynamic process resulting from an intrinsic instability that breaks the symmetry of the initial aggregate (Höfer et al. 1995). The cells form a shapeless mound of several thousand similar cells around the cell or cells that initiated the signaling process.

Vasiev et al. (1997), who have modeled this whole developmental sequence in the slime mold, point out that 'this complicated process could take place without the necessity for transcription of new genes. In real life, however, it is known that many different genes are transcribed during aggregation and mound formation.' For instance, after stream formation, molecules are produced which make cells sticky, reinforcing and stabilizing this pattern. The mound then self-organizes into an elongated shape called a slug, a few millimeters in length. The slug assembles under the influence of continuous periodic signals that come from the same region of cells, which becomes the leading end of the migrating slug. The migration of the slug is a coherent movement of the whole mass of several thousand cells—in other words, the coordinated dynamics of an excitable medium involving three-dimensional waves of activity that propagate through the slug from anterior to posterior (Bretschneider et al. 1999).

During migration of the slug, two types of cell differentiate, involving differential gene activities: the front one-third of the cells of the slug develop stalk characteristics while the posterior two-thirds become the spores. Differential gene activity is a major feature of this process. The slug stops migrating and organizes itself into a fruiting body consisting of a stalk supporting a spherical mass of spores a few millimeters above the surface. When conditions become moist again, resulting in bacterial growth, the spores are released and amoebae emerge to start the life cycle again, growing on the bacteria.

This example illustrates the point that to understand

a species' life cycle it is necessary to understand the dynamics of self-organizing patterns as well as the role of gene activity, which can both initiate and stabilize the morphogenetic changes characteristic of the species. During evolution, changing patterns of gene activity will occasionally produce coherent changes of morphology and behavior, which will persist if they are adaptive. What is becoming clear is that gene activity can stabilize, consolidate, or initiate developmental events, but these processes can only be understood in terms of the self-organizing, coherent dynamics of whole life cycles.

3. *Patterns of Behavior in Social Insects*

Entomologists have recognized for many years that communities of insects display high levels of coherent order in their activities and social organization (Hölldobler and Wilson 1990). However, these usually cannot be understood simply by observing the behavior and interactions of individuals. For example, several species of ant exhibit a rhythmic pattern of activity–inactivity in the brood chamber, where the workers are tending the queen and the brood, whereas individual ants have no tendency to show such rhythmic behavior. On the contrary, their activity patterns in isolation or at low density have been characterized as chaotic, in the technical sense of the term (Cole 1991). Another example comes from termite colonies, which construct very complex nests with coherently organized passageways, vaults, and ceilings. However, observation of activity in termite 'construction gangs' reveals a pattern of behavior that is anything but coherent. Individuals move about, join groups working on different building sites scattered about the construction area, stop suddenly for several minutes, and then scatter randomly to new sites. In these cases, how does coherent order arise from such disorder?

Once again, explanations of pattern require an analysis that goes beyond genes, individual behavior, and natural selection, to models that introduce a new level of complexity in the dynamics of the group. Rhythmic behavior of workers in brood chambers can be shown to arise from chaotic behavior of individuals as a result of stimulatory interactions between ants when they exceed a critical density, as usually occurs in brood chambers (Solé et al. 1993). Furthermore, the range of parameters (e.g., sensitivity of ants to one another and density) that result in rhythmic activity is large, so that this behavior pattern is robust to significant genetic and environmental disturbance, as stable evolutionary characters must be. Similarly, it has been shown (O'Toole et al. 1999) that the coherent architectural results of apparently disorganized termite behavior arise from a dynamic process called 'self-organized criticality' (Bak 1988). The attraction of termites to occupied building sites, the sudden immobility at high density (but variable size) of construction gangs, and subsequent scattering to new sites form the basis of a process ensuring that all sites grow at roughly the same rate. With such synchrony the nest components fit together coherently to give organized architectural constructions. Again, this dynamic is stable to parameter perturbation. 'Order for free' must be robust if it is to persist in evolution.

4. *Emergent Properties of Ecosystems and the Biosphere*

A crucial issue in the study of natural ecosystems concerns the relationship between complexity and stability: are complex ecosystems more or less stable than simple ones? Complexity may be defined, following May (1973), using the notion of 'degree of connectedness': ecosystems whose species are richly interconnected are more complex than sparsely connected ones. May's theoretical study, using conventional population dynamics to model ecosystems and using linear stability analysis, showed that there is a fairly sharp transition between two states. Below a critical degree of complexity (connectivity) ecosystems are stable; above, they are unstable. Detailed analysis of field studies by Keitt and Marquet (1996) showed that there is an interesting inverse relationship between number of species in an ecosystem and their connectivity, supporting May's conclusion and suggesting an underlying dynamic order. Furthermore, these investigators examined the frequencies of species extinction events of different sizes in ecosystems that were getting more complex as a result of the introduction and establishment of new, exotic species. Their results revealed a regular 'power law relationship' that is suggestive of an organized dynamic.

A study by Ellner and Turchin (1995) took stability analysis a step further in identifying a special emergent property of ecosystems. They examined an extensive set of ecological time series within an ecosystem and concluded that they showed chaotic dynamics. Their conclusion supported a conjecture from purely theoretical studies of co-evolving systems, that interacting populations co-evolve 'at the edge of chaos' (Wolfram 1986). Langton (1990) then noticed that there are sets of interaction rules that result in highly ordered patterns of stable behavior and others that result in extremely complex activities without any repetitive order ('chaos'). The rules can be ordered logically and it turns out that in between the ordered and the chaotic regimes there is a realm of complexity whose behavior suggested a potential for 'emergent computation' (Langton 1990). Langton suggested that real biological evolution might occur in such a domain of complex dynamics, the edge of chaos, where new order can arise as evolving populations move from order to chaos and on into new order.

Langton's provocative conjecture has spawned a number of more sophisticated computer models of co-

evolving populations, among them one by Kauffman and Johnsen (1991). What this model simulates is a dynamic pattern of species extinction sizes of the type described by Pimm (1984) and the time series behavior examined by Ellner and Turchin (1995). A more accurate fit to the data has been obtained by Kauffman (1995) using an extended version of the original model.

A further development of self-organization theory arises from a study of extinction statistics obtained from the fossil record. Solé et al. (1997) showed that extinction has the property of 'self-similarity,' a power-law governing the relationship between probability of extinction and size of extinction event (small extinctions are common and large ones are rare). A model of ecosystem dynamics that fits these observations has been constructed (Solé et al. 1996, 1998), suggesting a clear distinction between the dynamics of microevolution (adaptive modifications of species) and macroevolution (patterns of extinction events). Competitive interactions between species and adaptive processes clearly occur in evolution, but they do not explain the large-scale pattern of evolutionary change (extinction sizes and probabilities). The latter arises from the collective dynamics of ecological networks as an emergent property that is not reducible to the competitive interactions and adaptive modifications of species. Micro- and macroevolution are then distinct, requiring different explanations.

Finally, unexpected regularities are observed at the level of the whole biosphere in variables such as the oxygen and carbon dioxide content of the atmosphere, the mean surface temperature of the earth, and the concentrations of various ions (calcium, sodium, potassium, chloride, phosphate, sulfate, etc.) in the oceans. Purely geophysical processes are unable to account for the stability of these quantities over geological time periods (millions of years) and the evidence now favors the intimate involvement of biological processes as crucial contributors to global biosphere dynamics. This radical shift of focus for evolutionary thinking comes from the work of James Lovelock (1979, 1991) and the study of microbial communities by Lynn Margulis (1987), who together gave birth to the Gaia Hypothesis. A whole new class of models is under development to account for the capacity of the living earth to regulate variables crucial for life by processes that give rise to emergent stability properties (Harding 1999, Lenton and Watson 2000).

See also: Chaos Theory; Cultural Evolution: Overview; Emergent Properties; Evolution, History of; Evolution: Optimization; Genes and Culture, Coevolution of; Infant and Child Development, Theories of; Lifespan Development: Evolutionary Perspectives; Lifespan Development, Theory of; Optimization in Evolution, Limitations of; Psychological Development: Ethological and Evolutionary Approaches; Self-organizing Dynamical Systems

Bibliography

Bak P 1988 Self-organized criticality. *Physical Review Letters A* **38**: 364–73

Bretschneider T, Vasiev B, Weijer C J 1999 A model for *Dictyostelium* slug movement. *Journal of Theoretical Biology* **199**: 125–36

Cole B J 1991 Is animal behaviour chaotic? Evidence from the activity of ants. *Proceedings of the Royal Society London B* **244**: 253–9

Cowan G A, Pines D, Meltzer D (eds.) 1994 *Complexity: Metaphors, Models and Reality*. Santa Fe Institute Studies in the Sciences of Complexity. Addison-Wesley, Reading, MA, Vol. 19

Ellner S, Turchin P 1995 Chaos in a noisy world: New methods and evidence from time-series analysis. *American Nature* **145**: 343–75

Goodwin B C 1994 *How The Leopard Changed Its Spots*. Weidenfeld and Nicolson, London

Harding S P 1999 Food web complexity enhances community stability and climate regulation in a geophysiological model. *Tellus* **51B**: 815–29

Höfer T, Sherratt J A, Maini P K 1995 *Dictyosteium discoideum*: Cellular self-organization in an excitable biological medium. *Proceedings of the Royal Society London* **259**: 249–57

Hölldobler B, Wilson E O 1990 *The Ants*. Harvard University Press, Cambridge, MA

Kauffman S A 1995 *At Home in the Universe*. Oxford University Press, New York

Kauffman S A, Johnsen S 1991 Coevolution at the edge of chaos: Coupled fitness landscapes, poised states and coevolutionary avalanches. *Journal of Theoretical Biology* **149**: 467–80

Keitt T M, Marquet P A 1996 The introducted Hawaiian arifauna reconsidered evidence for self-organized critically? *Journal of Theoretical Biology* **182**: 161–7

Langton C 1990 Computation to the edge of chaos: Phase transitions and emergent computation. *Physica* **42D**: 12–37

Lenton T, Watson A 2000 Redfield revisited: What regulates the oxygen content of the atmosphere? *Global Geochemical Cycles* **14**: 249–68

Lovelock J E 1979 *Gaia: A New Look at Life on Earth*. Oxford University Press, Oxford, UK

Lovelock J E 1991 *Healing Gaia: Practical Medicine for the Planet*. Harmony, New York

Margulis L 1987 Early life: The microbes have priority. In: Robson J M (ed.) *Gaia: A Way of Knowing*. Lindisfarne Press, Hudson, New York, pp. 98–109

May R 1973 *Stability and Complexity in Model Ecosystems*. Princeton University Press, Princeton, NJ

Mayr E 1988 *Toward a New Philosophy of Biology*. Harvard University Press, Cambridge, MA

O'Toole D V, Robinson P A, Myerscough M R 1999 Self-organised critically in termite architecture: A role for crowding in ensuring ordered nest expansion. *Journal of Theoretical Biology* **198**: 305–27

Pimm S 1984 The complexity and stability of ecosystems. *Nature* **307**: 321–6

Pimm S 1991 *The Balance of Nature*. Chicago University Press, Chicago, Chap. 8

Solé R V, Bascompte J, Manrubia S C 1996 Extinctions: Bad genes or weak chaos? *Proceedings of the Royal Society London B* **263**: 1407–13

Solé R V, Goodwin B C 2000 *Signs of Life: How Complexity Pervades Biology*. Basic Books, New York

Solé R V, Manrubia S C, Benton M, Bak P 1997 Self-similarity of extinction statistics in the fossil record. *Nature* **388**: 764–7
Solé R V, Manrubia S C, Perez-Mercader J, Benton M, Bak P 1998 Long-range correlations in the fossil record and the fractal nature of macroevolution. *Advances in Complex Systems* **1**: 255–66
Solé R V, Miramontes O, Goodwin B C 1993 Oscillations and chaos in ant societies. *Journal of Theoretical Biology* **161**: 343–57
Vasiev B, Siegert F, Weijer C J 1997 A hydrodynamic model for *Dictyostelium discoideum* mound formation. *Journal of Theoretical Biology* **184**: 441–50
Wolfram S 1986 *Theory and Applications of Cellular Automata*. World Scientific, Singapore

B. Goodwin

Evolutionary Approaches in Archaeology

Since the 1970s, archaeologists in the US and the UK have sought to understand the material record of human existence in terms explicitly drawn from Darwinian evolutionary theory. All such approaches begin from the assumption that human behavior, like other aspects of the human phenotype, is shaped by variation generation (i.e., mutation if the focus is on biologically inherited traits), selection, and random processes (e.g., drift, founder effect). Beyond this basic area of agreement, views diverge considerably on how evolutionary processes shape the archaeological record and how archaeological methods should be tailored to facilitate their study. In some respects, the debates about evolution in archaeology parallel those in evolutionary biology.

1. *Evolution*

'Cultural evolution' is often used by archaeologists to refer to a progressive historical trend, with progress defined in ethnocentric terms, such as greater social and political complexity. Anthropologists of the mid-twentieth century, including Leslie White and Julian Steward, brought such progressive evolutionism into the anthropological mainstream, and the 'New Archaeology' of the 1960s and 1970s was influenced by these efforts. Dunnell (1980) has argued, however, that 'cultural evolution' of the progressive variety, which can be traced back to the nineteenth-century writings of Spencer, Tylor, and Morgan, has little in common with scientific evolution, which was developed within biology following publication of Darwin's *Origin of Species* in 1859. According to Dunnell (1980, p. 44), 'cultural evolution' is not scientific because it provides no mechanisms to explain history, instead positing plausible empirical generalizations that summarize certain ethnographic observations recorded during European expansion across the globe.

Dunnell (1982) and others (Lyman and O'Brien 1997, Schennan 1991) find antecedents of a scientific evolutionary archaeology not in 'cultural evolution' of the progressive variety but rather in culture historical studies of the early through mid-twentieth century. What culture history had going for it that cultural evolution did not was a 'decidedly materialist feel' (Lyman and O'Brien 1997, p. 43). 'Materialism,' which is used in this context as a synonym for 'population thinking' (Mayr 1959), is the view that historical change arises out of individual variation. This was Darwin's great insight (Lewontin 1974). Before Darwin, explanations of change were bound by an essentialist worldview. With habits of thought derived ultimately from Aristotelian essentialism, early evolutionists could only conceive of change as taking place when things are transformed into other things through individual striving or the unfolding of potentials inherent in the individual's underlying essence. Caterpillars are transformed into butterflies; monarchies are transformed into parliamentary democracies. Darwin freed evolution from the bounds of essentialism when he argued that 'descent with modification' could arise as a purely mechanistic consequence of variants being reproduced and dropping out of populations at different rates. Inasmuch as they purport to be Darwinian, all of the evolutionary approaches summarized below adhere at least implicitly to this fundamental understanding of the mechanism that produces change through history.

2. *Evolutionary Approaches in Archaeology*

Taking inspiration from Darwin's phrase 'descent with modification,' Durham (1992) distinguishes between cultural evolutionary studies focusing on descent and those focusing on modification. Out of favor during the heyday of the 'New Archaeology,' studies of cultural descent were revived during the 1980s in works such as Renfrew's (1987) study of Indo-European origins, Kirch and Green's (1987) study of cultural phylogeny in Polynesia, and Flannery and Marcus's (1983) study of Zapotec prehistory. Other historical sciences, such as historical linguistics and biological anthropology contribute along with archaeology toward the common goal of reconstructing cultural phylogenies.

When biological, linguistic, and material similarities are summarized in the form of a cultural phylogeny, an underlying assumption is that the pattern was created by the mechanisms outlined by Darwin (which Durham (1992) subsumes under 'modification'). But how would Darwinian mechanisms work to generate the material remains that make up the archaeological record? The ways that archaeologists have answered

this question can be conveniently grouped under five headings, which provide a framework for the following survey of evolutionary approaches in archaeology.

2.1 Evolutionary Archaeology

The approach instigated by Dunnell (e.g., 1978, 1980) is often called evolutionary archaeology (EA) by both its proponents and its critics (e.g., Boone and Smith 1998). Evolutionary archaeologists contend that culture constitutes an inheritance system that parallels genetic inheritance. Moreover, culturally transmitted traits, like genetic traits, may affect an organism's probability of survival and reproduction, which leads to frequency change over time in cultural traits just as it does for genetic traits. The complete human phenotype is encoded by both genetic and cultural instructions, and artifacts and other cultural manifestations therefore owe their existence and diversity to the same evolutionary processes (selection and drift) that create biological diversity (e.g., Dunnell 1989). Humans perceive and solve problems, but the importance of these intentional efforts is that they generate variation, which then becomes available for selection (Rindos 1989).

David Rindos' (1984) work on the origins of agriculture incorporated many aspects of the EA approach first articulated by Dunnell (1980). Rindos viewed the emergence of domesticates as a Darwinian evolutionary process in which selective pressures shaped genetic evolution of the plants at the same time they were shaping cultural characteristics in coevolving human populations. Rindos argued that the same coevolutionary processes govern the mutualistic relationships between humans and their domesticated plants as those which govern other animal-plant relationships, such as that between ants and acacia trees that they harvest for food. A distinctive (though perhaps not unique) feature of the human-plant coevolutionary relationship is that two distinct modes of inheritance, one cultural and one genetic, are responsible for transmission of the human behavioral traits that are being shaped by selection. Nonetheless, Rindos subsumed frequency change in both genetic and cultural traits within the same set of coevolutionary models. Hart (1999) has recently applied and extended Rindos' framework in a study of the evidence for maize agriculture in the Eastern Woodlands.

Theoretical works in the EA tradition have elaborated themes originally introduced by Dunnell, such as the style-function dichotomy. According to Dunnell (1978), functional traits can be understood in an evolutionary context to be those traits that are fixed by selection (i.e., they have positive fitness values), whereas stylistic traits are those for which frequency change can be attributed to drift. Dunnell contends that archaeological frequency seriation requires neutral traits, and this has been borne out by mathematical models and simulation (Neiman 1995, Lipo et al. 1997). Results of this research have been used to develop a series of inferences regarding intergroup interaction in the Woodland period Midwest (Neiman 1995) and lower Mississippi Valley (Lipo et al. 1997).

Left unanswered in Dunnell's original works is the question of whether biological reproductive success is necessary for a cultural trait to exhibit positive fitness and thus be favored by selection. Leonard and Jones (1987) introduce the term 'replicative success' to refer to any trait that increases in frequency relative to alternative traits, but they maintain that cultural traits that do not affect biological reproductive success cannot be termed functional (i.e., subject to selection). Barton and Clark (1997, pp. 12–13) summarize the conflicting positions of evolutionary archaeologists on this issue.

A theoretical issue dealt with at greater length in EA (e.g., O'Brien and Holland 1990, Neff and Larson 1997) is the supposed circularity of 'fitness' when more fit individuals are identified as such only because they enjoy greater success at passing on genetic or cultural information. As in evolutionary biology, the 'propensity definition' of fitness backed up by engineering analysis of artifact design are seen as means to circumvent this circularity (O'Brien and Holland 1990). These arguments lead to one of the major methodological prescriptions of EA, namely that investigation of evolutionary history should focus on concrete phenomena (artifacts at various scales) and proceed by advancing and testing hypotheses about the forces that shaped observed artifact design. A focus on concrete phenomena, such as artifacts, is also suggested by Dunnell's (1980) discussion of empirical sufficiency.

Substantive applications within EA, not surprisingly, have tended to emphasize the explanation of variation at the level of the artifact. Examples are found in a volume edited by Teltser (1995), where the case studies include one on ceramic source-usage variation through time, one on the design of agricultural fields, one on artifactual change following the arrival of Europeans in the New World, and one on ceremonial architecture in Polynesia. In these studies, the history of artifact design changes is explained by reference to constraints and opportunities posed by the environment. The models of neutral trait distributions developed by Neiman (1995) and Lipo et al. (1997) have also emphasized variation at the level of the artifact.

2.2 Evolutionary Ecology

The evolutionary ecology (EE) approach to the analysis of archaeological remains emerged in the early 1970s (citations in Broughton and O'Connell 1999).

Archaeologists identifying themselves with this approach (e.g., Boone and Smith 1998) stress the fact that human organisms have been designed by past selection of genetic variation to seek fitness-enhancing ways of responding to environmental conditions. Thus, cultural change and diversity are explained primarily by phenotypic adjustment to variable opportunities and constraints posed by the environment. More specifically, 'adaptive phenotypic variation involves the interaction between genetically or culturally evolved cognitive mechanisms and variable environmental conditions' (Boone and Smith 1998, p. S144). To take a now paradigmatic example, snowmobiles replaced snowshoes among the Cree in less than one human generation, a change that evolutionary ecologists explain by reference to Cree calculation of energy return rates under various transportation options and subsequent decisions to adopt the most efficient one.

The premise that humans have been designed by natural selection to seek fitness-enhancing behaviors warrants the use of optimization models, which is a key defining feature of the EE approach to archaeology. That is, given a set of alternative behavioral strategies, evolutionary ecologists assume that humans will follow the strategy that tends to maximize biological fitness in a given environment. Broughton and O'Connell (1999) point out that this underlying premise leads most naturally in archaeology to an emphasis on subsistence and settlement. For instance, usage of prey species, about which the optimal foraging models of EE make clear predictions, can be inferred more or less directly from archaeological data. Winterhalder and Goland (1997) have used the prey-choice framework to explore how different plants, including eventual domesticates, might have been brought into the diet of hunter-gatherers. This account converges with and complements the EA account of agricultural origins offered by Rindos (1984). In some cases, predictions about prey choice can also be extended to predictions about change in tools used to obtain various prey species, thus extending the applicability of EE into the realm of technology (Broughton and O'Connell 1999). Winterhalder (1997) shows how optimization models of evolutionary ecology apply to another key area of archaeological interest, exchange.

Although the inference of behavior from archaeological remains is quite direct when the focus is on subsistence, EE models, by definition, are always stated in behavioral terms, and this means that methods for reconstructing behavior from archaeological remains are crucial components of the EE program. EE proponents recognize that this constraint ties their research to other programs aimed at providing a rigorous basis for behavioral inference in archaeology. Thus, it is argued that ethnoarchaeology (O'Connell 1995) and behavioral archaeology (Boone and Smith 1998, Broughton and O'Connell 1999) can render the archaeological record interpretable in behavioral terms, while EE furnishes the theory to explain why particular behaviors have appeared when and where they have in human history. Schiffer (1999), the main architect of behavioral archaeology, approves of the EE focus on behavior, but considers the optimization models of EE useful for explaining only a small part of the range of human behavioral variability.

A somewhat polemical debate between proponents of EE and proponents of EA emerged in the late 1990s. Evolutionary archaeologists contend that diversity and change observed in the archaeological record can be explained by reference to the effects of selection and drift on cultural inheritance. They reject the EE reliance on phenotypic adjustment on the grounds that it ties explanation to individual intent rather than mechanisms (Dunnell 1989, Lyman and O'Brien 1998). EE proponents contend, on the contrary, that many if not most instances of historical change happen too fast to be explainable by reference to the long-term process of natural selection working on culturally inherited traits. They cite the Cree snowmobile case in support of the necessity to consider phenotypic adjustment: the long-term criterion of differential biological success, they argue, simply cannot explain the adoption of snowmobiles in less than one human generation (Boone and Smith 1998, Sects. 146–7).

The EE and EA camps also dispute whether behavior should constitute the object of evolutionary explanation in archaeology. Evolutionary ecologists advocate a hypothetico-deductive approach in which optimization models are the basis for predictions about prehistoric human behavior, and they argue that, with methods of behioral archaeology and ethnoarchaeology, these predictions are testable. From this point of view, the preoccupation with artifacts and their distributions in time and space is a serious limitation of the EA approach (e.g., Boone and Smith 1998, pp. S154–6, Broughton and O'Connell 1999, pp. 158). EA proponents (e.g., Dunnell 1989, Lyman and O'Brien 1998), however, argue that evol utionary models stated in behavioral terms are inherently untestable because the behaviors about which they make predictions are unobservable and can only be inferred from the archaeological record with a leap of faith.

2.3 Approaches based on Dual Inheritance Theory

Dual inheritance theory (DIT) (Boyd and Richerson 1985, Durham 1992) addresses the ways in which the principle of selective retention coupled with the special properties of cultural transmission lead to expectations about the evolution of adaptive and not so adaptive cultural practices, artifacts, and institutions. The theory explicitly incorporates complexities of cultural transmission that are suggested by social learning research. Boyd and Richerson (1985) con-

struct models of how various forms of bias in the transmission process lead to different evolutionary results; Durham (1992) prefers to call the forces of bias 'cultural selection.'

Two main predictions emerge from DIT. First, cultural evolution, which is aided by guided variation and biased transmission (Durham's (1992) 'primary value selection'), can converge on solutions that are adaptive for the human organism faster than evolution that depends exclusively on natural selection of genetic variation. In this case, fitness maximization adequately predicts human behavior. Second, DIT predicts that differential survival of cultural variants may also generate inherited behavior patterns that run counter to the predictions of genetic fitness maximization (Bettinger 1991, Shennan 1991). Indirect bias and frequency-dependent bias (which Durham (1992) calls 'secondary value selection') further enhance the potential for cultural manifestations to be maladaptive in a biological sense.

DIT shares with EA the view that culture constitutes an inheritance system evolving with some degree of independence. However, the EA conviction that variation must be independent of the force of selection for a true Darwinian process to work is violated by guided variation and biased transmission, which are central to DIT. For their part, dual inheritance theorists argue that natural selection is inadequate, on its own, to explain cultural change and diversity (Richerson et al. 1998). Instead, decision making—the ways in which cultural transmission interacts with learning to produce variation that can then be shaped by selection—has to be modeled explicitly if the archaeological record is to be understood in evolutionary terms.

DIT is thus aligned with EE regarding the importance of decision making in cultural evolution. Moreover, the optimality models of EE are regarded as valuable sources of hypotheses that can serve as starting points for archaeological investigation (Bettinger and Richerson 1996). At the same time, however, the independent dynamics of an evolutionary process based on cultural transmission leads to skepticism regarding simple optimality interpretations of human behaviors (Bettinger 1991, Richerson and Boyd 1992, Shennan 1991). People may tend to acquire beliefs, values, practices, and artifacts that are adaptive in the biological sense, but this assumption is limited in its potential to account for the full range of human variation.

Efforts to put DIT to work in the service of archaeology have tended to emphasize means for recognizing archaeologically the various modes of cultural transmission. Bettinger (1991, pp. 203–7), for instance, suggests that the complex of changes associated with the European Upper Paleolithic transition may be explainable as indicators that indirect bias and frequency-dependent bias for the first time came to play important roles in human cultural evolution. Bettinger argues that climatic change and population growth would have favored local subsistence specialization, which in turn would have favored mechanisms (like indirect and frequency-dependent bias) for efficiently acquiring locally adaptive behavioral traits as a complex.

2.4 *Approaches based on Evolutionary Psychology*

Like dual inheritance theorists, Mithen (1989, 1997a, 1997b) believes that simple adaptive models of human behavior such as those that make up EE are insufficient to account for human behavioral variation. However, rather than exploring the implications of cultural transmission mechanisms, Mithen explains departures from adaptive expectations largely by reference to the fact that humans now and in the recent past have lived in environments very different from those in which human cognitive capacities evolved. This viewpoint has been explored most thoroughly within evolutionary psychology. A major implication of this research, accepted by Mithen (1997a, 1997b), is that the human brain consists of various specialized devices designed by evolution to solve specific problems faced by hominids in the past. Mithen (1997a) contends that this modular view invalidates models, such as Boyd and Richerson's, that only indirectly consider the content of cultural transmission.

Mithen's methodological prescriptions converge with those of many other brands of archaeology, especially behavioral archaeology (see Sect. 2.2). The archaeological record is conceived as the aggregate result of active individuals endowed with common psychological propensities making decisions in unique historical contexts, and archaeology's task is to relate the short-term individual behavior to the gross character of the archaeological record (Mithen 1989, pp. 491–2). Evolution enters the picture to the extent that behavioral patterns inferred from the record are eventually interpreted as adaptations, or as the expression of human propensities that are no longer adaptive due to their existence in novel, inappropriate environments, or as the unintended consequences of pleiotropic adaptive traits.

Mithen(1997a, 1997b) also believes that archaeology can contribute to evolutionary psychology by providing better understanding of the evolution of human cognitive capacities (see *Cognitive Archaeology*). As an example of the mutualistic relationship of cognitive archaeology and evolutionary psychology, Mithen (1997a) interprets religious ideas as evolutionary byproducts of increased 'cognitive fluidity' (i.e., increased integration of the brain's modular cognitive capacities) that characterizes the brains of modern humans. Since religious ideas are not in themselves adaptive and may often be maladaptive, the brain has no specific cognitive capacity for transmitting religious ideas. Early modern humans therefore had to develop techniques, including ritual and

art, to ensure religious transmission. Thus, the flowering of Upper Paleolithic art is interpreted as the archaeological expression of evolutionary changes in cognitive architecture associated with the emergence of modern humans.

Proponents of other evolutionary approaches in archaeology have mainly ignored the cognitive archaeology work, despite some severe critiques launched by Mithen (1997a) in the other direction. From the EA perspective, inference of cognitive processes from artifacts would probably be considered informed speculation at best. Evolutionary ecologists, for their part, might observe that optimization models that lie at the core of EE are preferable to antiquated functionalist arguments used to attribute adaptive significance to behavioral patterns inferred model-free from the archaeological record.

2.5 'Memes' and Cultural Virus Theory

In the evolutionary approaches considered so far, artifacts are viewed either as reflections of human behavior (EE, DIT, and evolutionary psychology) or as phenotypic characteristics of individuals (EA). An alternative perspective, considered by some 'the most extreme variant of cultural selectionism' (Maschner and Mithen 1996, p. 7), builds on the meme concept introduced by Dawkins (1976). Memes are cultural replicators, analogous to genes. The only purpose of any replicator, whether it is a gene or a meme, is to make copies of itself. Memes cannot self-replicate, but, like the genetic instructions in viruses, they can harness the capacities of other organisms to get themselves replicated. The last characteristic has led Ben Cullen (1993, 1996), the main proponent of this perspective in archaeology, to propose naming it 'cultural virus theory' (CVT).

To the extent that opportunities for memes to be replicated are limited by the amount of brain space available in a human population, memes 'compete,' just as other replicators (i.e., genes) compete. But the differential persistence of memes need have little to do with the biological reproductive success of human organisms in which they reside. Instead, differential persistence of memes arises purely out of differential success at replicating from brain to brain. The reasons why some memes enjoy relative success at replication can be found in the differential physical effects they have upon the world: under some circumstances, a cooking pot in which water boils in five minutes will have its design memes copied more often than one that requires 10 minutes. Evolutionary ecologists might argue that selection of memes in this case arises in the human cognitive capacity (which was designed by natural selection of genetic variation) to recognize more optimal solutions to problems such as how to boil water. From the meme perspective, however, artifacts, not humans, are the 'interactors' in the evolutionary process. Artifacts are packages of the physical effects of memes, and the success or failure of artifacts determines frequency change in their design memes. Archaeological assemblages, from this point of view, represent ecological assemblages of artifacts as interactors.

Although CVT has been largely ignored by less 'extreme' advocates of evolutionary approaches in archaeology, the artifact-oriented approach of EA can be considered consistent with this perspective. The engineering design approach of O'Brien and Holland (1990) evaluates the fitness of artifacts, thus implying that artifacts are the interactors in the evolutionary process. Likewise, models for predicting the pattern of trait distributions in the archaeological record (e.g., Lipo et al. 1998; Neiman 1995) necessarily ignore the biological reproductive cycles of humans and lead to predictions about the distributions of traits counted on artifacts.

Like DIT and approaches based on evolutionary psychology, memetics (or CVT) predicts the existence of cultural manifestations that are irrelevant to biological success or even patently maladaptive from the perspective of human survival and biological reproduction. However, as the example of the cooking pot mentioned above illustrates, expectations about artifactual evolution can converge with predictions of an optimality argument. More generally, human proclivities and decision-making capacities, fashioned by natural selection in the past, provide selective criteria that help determine the fate of culturally encoded designs

See also: Archaeology and Philosophy of Science; Cognitive Archaeology; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Evolution, History of; Evolutionary Social Psychology; Evolutionism, Including Social Darwinism; Memes and Cultural Viruses; Theory in Archaeology

Bibliography

Barton C M, Clark G A 1997 Evolutionary theory in archaeological explanation. In: Barton C M, Clark G A (eds.) *Rediscovering Darwin: Evolutionary Theory and Archaeological Explanation*, Archaeological Papers of the American Anthropological Association, No. 7, Arlington, VA

Bettinger R L 1991 *Hunter-gatherers: Archaeological and evolutionary theory*. Plenum, New York

Bettinger R L, Richerson P J 1996 The state of evolutionary archaeology: evolutionary correctness, or the search for common ground. In: Maschner H D G (ed.) *Darwinian Archaeologies*. Plenum, New York

Boone J L, Smith E A 1998 Is it evolution yet? A critique of evolutionary archaeology. *Current Anthropology* **39**: S141–73

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. The University of Chicago Press, Chicago

Broughton J M, O'Connell J F 1999 On evolutionary ecology, selectionist archaeology, and behavioral archaeology. *American Antiquity* **64**: 153–65
Cullen B 1993 The Darwinian resurgence and the cultural virus critique. *Cambridge Archaeological Journal* **3**(2): 179–202
Cullen B 1996 Cultural virus theory and the eusocial pottery assemblage. In: Maschner H D G M (ed.) *Darwinian Archaeologies*. Plenum, New York
Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford, UK
Dunnell R C 1978 Style and function: a fundamental dichotomy. *American Antiquity* **43**: 192–02
Dunnell R C 1980 Evolutionary theory and archaeology. In: Schiffer M B (ed.) *Advances in Archaeological Method and Theory*, Academic Press, New York Vol. 3
Dunnell R C 1982 Science, social science, and common sense: the agonizing dilemma of modern archaeology. *Journal of Anthropological Reseacrch* **38**: 1–25
Dunnell R C 1989 Aspects of the application of evolutionary theory in archaeology. In: Lamberg-Karlovsky C C (ed.) *Archaeological Thought in America*. Cambridge University Press, Cambridge, UK
Durham W H 1992 Applications of evolutionary culture theory. *Annual Review of Anthropology* **21**: 331–55
Flannery K V, Marcus J (eds.) 1983 *The Cloud People: Divergent Evolution of the Zapotec and Mixtec Civilizations*. Academic Press, New York
Hart J P 1999 Maize agriculture evolution in the Eastern Woodlands of North America: A Darwinian perspective. *Journal of Archaeological Method and Theory* **6**(2): 137–80
Kirch P V, Green R C 1987 History, phylogeny, and evolution in Polynesia. *Current Anthropology* **28**: 431–56
Leonard R D, Jones G T 1987 Elements of an inclusive evolutionary model for archaeology. *Journal of Anthropological Archaeology* **6**: 199–219
Lewontin R C 1974 Darwin and Mendel: The materialist revolution. In: Neyman J (ed.) *The Heritage of Copernicus: Theories Pleasing to the Mind*. MIT Press, Cambridge, MA
Lipo C P, Madsen M E, Dunnell R C, Hunt T 1997 Population structure, cultural transmission, and frequency seriation. *Journal Anthropological Archaeology* **16**: 301–33
Lyman R L, O'Brien M J 1997 The concept of evolution in early twentieth century Americanist archaeology. In: Barton C M, Clark G A (eds.) *Rediscovering Darwin: Evolutionary Theory and Archeological Explanation*, Archaeological Papers of the American Anthropological Association, No. 7, Arlington, VA
Lyman R L, O'Brien M J 1998 The goals of evolutionary archaeology. *Current Anthropology* **39**: 615–52
Maschner H D G, Mithen S 1996 Darwinian archaeologies: An introductory essay. In: Maschner H D G (ed.) *Darwinian Archaeologies*. Plenum, New York
Mayr E 1959 Typological thinking versus population thinking. In: Meggers B J (ed.) *Evolution and Anthropology: A centennial Appraisal*, The Anthropological Society of Washington, Washington, DC
Mithen S 1989 Evolutionary theory and post-processual archaeology. *Antiquity* **63**: 483–94
Mithen S 1997a Cognitive archaeology, evolutionary psychology, and cultural transmission, with particular reference to religious ideas. In: Barton C M, Clark G A (eds.) *Rediscovering Darwin: Evolutionary Theory and Archeological Explanation*, Archaeological Papers of the American Anthropological Association, No. 7, Arlington, VA
Mithen S 1997b *The Prehistory of the Mind*. Thames and Hudson, London
Neff H, Larson D O 1997 Methodology of comparison in evolutionary archaeology. In: Barton C M, Clark G A (eds.) *Rediscovering Darwin: Evolutionary theory in Archeological Explanation*, Archaeological Papers of the American Anthropological Association, No. 7, Arlington, VA
Neiman F D 1995 Stylistic variation in evolutionary perspective: Inferences from decorative diversity and interassemblage distance in Illinois Woodland ceramic assemblages. *American Antiquity* **60**: 7–36
O'Brien M J, Holland T D 1990 Variation, selection, and the archaeological record. In: Schiffer M B (ed.) *Archaeological Method and Theory*, University of Arizona Press, Tucson, AZ Vol. 2
O'Connell J F 1995 Ethnoarchaeology needs a general theory of behavior. *Journal of Archaeological Research* **3**: 205–55
Renfrew C 1987 *Archaeology and Language: The Puzzle of Indo-European Origins*. Cambridge University Press, London
Richerson P J, Bettinger R L, Boyd R 1998 Comment on Lyman R L, O'Brien M J, The goals of evolutionary archaeology: History and explanation. *Current Anthropology* **39**: 638–9
Richerson P J, Boyd R 1992 Cultural inheritance and evolutionary ecology. In: Smith E A, Winterhalder B (eds.) *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, New York
Rindos D 1984 *The Origins of Agriculture: An Evolutionary Approach*. Academic Press, Orlando, FL
Rindos D 1989 Undirected variation and the Darwinian explanation of cultural change. In: Schiffer M B (ed.) *Archaeological Method and Theory*, University of Arizona Press, Tucson, AZ Vol. 1
Schiffer M B 1999 Behavioral archaeology: Some clarifications. *American Antiqquity* **64**: 166–8
Shennan S J 1991 Tradition, rationality, and cultural transmission. In: Preucel R W (ed.) *Processual and Postprocessual Archaeologies: Multiple ways of Knowing the Past*, Center for Archaeological Investigations, Occasional Paper No. 10. Board of Trustees, Southern Illinois University, Carbondale, IL
Teltser P A (ed.) 1995 *Evolutionary Archaeology: Methodological Issues*. The University of Arizona Press, Tucson, AZ
Winterhalder B 1997 Gifts given, gifts taken: The behavioral ecology of nonmarket, intragroup exchange. *Journal of Archaeological Research* **5**(2): 121–68
Winterhalder B, Goland C 1997 An evolutionary ecology perspective on diet choice, risk, and plant domestication. In: Gremillion K J (ed.) *People, Plants, and Landscapes: Studies in Paleoethnobotany*. University of Alabama Press, Tuscaloosa, AL

H. Neff

Evolutionary Epistemology

Traditionally, epistemology concerns justifying beliefs. What right do we have to think that our beliefs are true? Evolutionary epistemology is a form of naturalized epistemology. In naturalized epistemology, the

justification provided turns on an appeal to something about the natural world, in this case the evolutionary process. Organisms that do not react appropriately to their environments are not likely to survive to reproduce. In some organisms this reaction can appropriately be termed 'knowledge.' Those organisms that are better at learning about their environments (which includes other organisms) are more likely to survive to reproduce than their less adept congeners. Thus, knowledge of those features of the environment that are relevant to survival and reproduction is to some extent justified. *Homo sapiens* as a species depends for its survival on playing the knowledge game better and more extensively than any other species. Evolutionary epistemologists propose to justify at least some of the beliefs that human beings hold by reference to survival value (for the most seminal paper in evolutionary epistemology, see Campbell 1974).

1. *Gene-based Evolutionary Theory*

In order to understand evolutionary epistemology, one must understand evolution, in particular the distinction between genotypes and phenotypes (see *Genotype and Phenotype*). The genotype of an organism consists in the sequence of bases in its DNA, or rarely, in its RNA. Everything else is part of its phenotype, from its blood type to its behavior. Variations in the genotype of an organism lead, through embryological development, to different phenotypic traits. These genetic variations are caused, totally caused. Among these causes are unequal crossover, increased temperature, and the presence of numerous mutagens. However, to count as Darwinian evolution, no correlation must exist between the variations an organism might need and those it might get; for example, it cannot be 'Lamarckian.' For Lamarckian evolution to occur, some feature of an organism's environment must change the phenotype of that organism in such a way that it is better adapted to the environmental feature that produced this change in the first place. This phenotypic change must then somehow be transmitted to the organism's genes to be passed on to subsequent generations via the genetic material. A mother dog giving fleas to her puppies does not count as Lamarckian inheritance.

Some of the traits exhibited by an organism make it better able to survive and reproduce. As a result, the genes that code for these traits are passed on differentially from generation to generation. However, as organisms become better-adapted to their environments, their environments frequently move out from under them. A few species succeed in surviving through the millennia largely unchanged, but most survive only by evolving. One feature of gene-based selection in biological evolution is that it can occur quite rapidly, as in viruses and bacteria, or quite slowly, as in most larger animals. However, in every case of Darwinian evolution, the relevant traits are inherited literally via the genetic material. Initially, the 'units' used in evolutionary biology were traits because evolutionary biologists had no access to the genetic material. As this access has increased, traits have been replaced by genes as units of inheritance. Unfortunately, the genetic material can be divided up in numerous ways. As a result, biologists concentrate on the 'information' contained in the genetic material.

Advocates of evolutionary epistemology in its most literal sense argue that the preceding story equally applies to some behavioral traits and predispositions. These traits and dispositions are as much phenotypic traits as are hair texture and skin color. Knowledge of the world in which organisms live gets programmed into an organism's genes. For example, fear of heights and the sucking reflex are in some sense programmed into the genes of human beings. The evolution of behaviors and behavioral predispositions has all the usual characteristics of gene-based biological evolution. As traits, behaviors can be adaptive. The inheritance involved is genetic, strictly Darwinian, not Lamarckian, and it can occur at a variety of rates from very slow to very fast.

The majority of epistemologists are not very impressed with evolutionary epistemology, in part because it is naturalistic, and only normative epistemologies are worth the title of epistemology. But a second weakness is that the only beliefs that evolutionary epistemology in this strict sense can even purport to justify are beliefs about entities and processes that the organism can perceive. Evolutionary epistemology might lend some justification for how human beings distinguish species of large organisms, but it cannot justify any beliefs that we have about the millions of tiny microorganisms that fill every nook and cranny of the world, let alone quantum indeterminacy.

2. *Individual Learning*

In 'eusocial organisms,' such as bees and termites, nearly all of the behaviors exhibited by these creatures evolved in a strictly Darwinian fashion. Relatively little learning takes place in the span of a single generation. But many organisms can learn from interacting with their environments. In a second form of evolutionary epistemology, knowledge does not get programmed into our 'genes,' but into our 'brains.' Gene-based selection in biological evolution can handle features of the environment that change relatively slowly, e.g., changes in temperature resulting from continental drift, but some changes are so rapid that organisms with longer generation times cannot adapt to them. One solution to this problem is the

evolution of the 'central nervous system.' Organisms with highly developed central nervous systems can learn about their environments and react to changes in them with greater speed. This information does not get programmed into an organism's genes but into its brain. 'Operant learning' is the simplest form of individual learning. Each kind of organism has a range of behaviors that it tends to emit spontaneously. Feedback from its environment increases or decreases the frequency of these behaviors. Sequential environmental responses shape the behavioral repertoires of these organisms.

The preconditions for operant learning are certainly programmed into the genes of those organisms that are capable of individual learning, and arose through gene-based natural selection. However, the specific learned repertoires cannot be passed on to successive generations. Each organism must learn anew. For better or for worse, calculus is not programmed into the genes of any human being. Each generation must learn it from scratch. If over many generations the same learning regimen is repeated and produces some evolutionary advantage, what was once a learned behavior can to some extent become programmed into the genes. Within a single life-history of an organism, the mechanisms that produce increased knowledge can be viewed as a 'selection process.' However, in this selection process, 'brain states' take the place of genes. Behavioral units can vary from something as brief and discrete as pressing a lever to something as complex and protracted as mowing a lawn. We have yet to discover the size and shape of the brain states that are responsible for the retention of information.

A second solution to the problem posed by differences in reaction times is the 'immune system.' In order to keep up with the rapid rate of reproduction of such antigens as viruses, immune systems evolved in many larger organisms. Within the life of a single organism, selection in the immune system occurs at the cellular level. Antibodies are produced in huge numbers, each with its own reaction site. If any of these antibodies happens upon an antigen that contains a protein that matches its reaction site, this antibody proliferates and, in most cases, eliminates the antigen. Once again, the 'information' garnered through successive invasions by antigens cannot be passed on to later generations. After each invasion, a few memory cells lie in wait for any subsequent invasions. That is all. Although some workers in the early years of modern immunology thought that the process might contain some elements of Lamarckian inheritance, it does not.

3. *Social Learning*

Classic epistemology dealt with individual, isolated cognizers. How much can we learn about our environments in total isolation from all other people? How certain could Robinson Crusoe be, once marooned on his island, of any new knowledge that he obtained? After all, he had no one with whom to check his discoveries. But even the Robinson Crusoe example is not pure enough, because Crusoe brought with him language and a fund of socially acquired knowledge. What about someone who began life totally isolated from all other human beings? In the first place, such an isolated baby would rapidly die, but how about the science fiction case? How about a baby adopted by wolves or some other unlikely creature? How much could it know? How justified would its knowledge be? The answer to both questions is very little. Social organisms can learn more than solitary organisms and be more secure in this knowledge because of culture (see *Cultural Evolution: Theory and Models*).

Most discussions of evolutionary epistemology do not deal with strict gene-based biological evolution or individual learning such as operant behavior, but 'social' learning. In social learning, knowledge acquired by one organism in its lifetime can be passed on to other organisms. This knowledge is transmitted from organism to organism, not through genes, not through the passing on of any brain tissue, but through 'imitation' and 'communication'. As a result, social learning need not proceed vertically, as defined by the direction that genes take in passing from parent to offspring, but horizontally (see Cavalli-Sforza and Feldman 1981). Education in the human species is an example of social learning. Through repeated efforts, teachers succeed in getting at least some of their pupils to be able to solve problems in calculus. Although the ability to learn in this way surely has a genetic component, social learning does not involve transmitting information genetically. As a result, it cannot count literally as Darwinian or Lamarckian. Of course, nothing keeps genes and culture from coevolving. In fact 'gene-culture coevolution' is likely to be of primary importance in human evolution (see Durham 1991, for additional reading, see Boyd and Richerson 1985, Cavalli-Sforza and Feldman 1981, Eldredge and Grene 1992, and Lumsden and Wilson 1981).

4. *Memetics*

At this stage, advocates of evolutionary epistemology as social learning make a significant move. They cease treating bits of knowledge, behavior, and the like as analogous to 'characters' but as analogous to 'genes.' Just as genes get passed on differentially in gene-based biological evolution, 'memes' get transmitted differentially in 'memetic' evolution (Blackmore 1999, Dawkins 1976, and Plotkin 1994). Just replacing genes with memes may not sound like a very profound change, but the implications are significant. Once again, memetic evolution is not in the least Lamarckian. Literally, memetic evolution is not Lamarckian

because nothing is being passed on 'genetically.' Metaphorically, it is not Lamarckian because what is being transmitted is not analogous to 'characters' but to 'genes.' It is the inheritance of acquired memes, not genes or even characters. The reason for emphasizing that memetic evolution is in no cogent sense Lamarckian, is that critics of memetic evolution think that they can dismiss this perspective simply by terming it 'Lamarckian.'

Critics of memetics provide a long list of objections to the memetic form of evolutionary epistemology. First, supposedly memetic evolution occurs so much more quickly than does gene-based biological evolution. But the same variations in speed of transmission can be found in both. Certainly viruses can spread as quickly as any meme. Second, critics complain that the analogy between genes and memes is imperfect. Genes are supposedly absolutely distinct, all of the same size and kind, and do not interact with each other. Supposedly memes exhibit none of these characteristics. Hence, the meme–gene analogy breaks down. But one need not know very much about genetics to know that genes possess none of the characteristics attributed to them by critics of memetics. Some genes are reasonably discrete; many are not, and much of the genetic material cannot be subdivided meaningfully into discrete genes. Genes vary tremendously in size, and numerous different sorts of genes exist (e.g., regulator genes). Also genes interact with each other in extremely complex ways in development. In addition, many elements of knowledge communication possess all of the preceding characteristics reasonably well developed; for example, e-mail addresses.

The chief problem with treating conceptual entities as memes is the difficulty of distinguishing *phenotypic characteristics* in this metaphorical usage. Behaviors count as characteristics in literal evolutionary epistemology. They are analogous to 'genes' in the metaphorical usage. Then what counts as phenotypic characters in the metaphorical usage? Memes get transferred from brains to air waves, to computer chips, to computer screens, to brains, etc. All of this counts as replication because information is being transmitted with relatively little loss. However, every once in a while, 'replication' is interrupted by 'translation.' Knowledge is connected causally to what it is knowledge of: you believe, for example, that barking dogs do not bite, so as you approach a barking dog, you put this belief to the test. The result will modify your belief system. The dog itself is not a meme.

As in other cases of selection, the notion of information plays a central role. One important feature of memetic replication is transmission with relatively little loss of information. One important feature of translation into the phenotype is significant loss of information. Instructions on how to make a Stradivarius violin can be transmitted from agent to agent. Antonio Stradivari might have written down these instructions, and even allowed an apprentice to watch him make a violin. If he had, this apprentice might have been able to make violins of the same caliber as did his master. He might have passed down this combination of knowledge and skill to the present. However, that is not what happened. All we have are Stradivari's violins. As repeated failures have shown, making violins based just on what we can learn from the finished product (reverse engineering) is extremely difficult, if not impossible. Future work in memetics is needed to present adequate analyses of 'information' and the 'genotype-phenotype' distinction.

5. *A General Analysis of Selection*

Early versions of evolutionary epistemology involved reasoning analogously from gene-based selection in biological evolution to knowledge acquisition. The analogical character of this activity has been the source of endless criticism. Evolutionary epistemology is 'merely' an analogy. In these arguments, the peculiarities of gene-based selection have played an inordinately important role in biological evolution. Any departure from gene-based selection in biological evolution, no matter how trivial, is viewed as a lethal disanalogy. However, an alternative view has been suggested periodically. Instead of reasoning analogously from one sort of selection process to another, present a 'general analysis of selection' that is equally applicable to all sorts of selection processes, whether gene-based selection in biological evolution, the reaction of the immune system to antigens, individual learning, social learning, memetics, etc. (Toulmin 1972). Although the father of evolutionary epistemology, Donald Campbell (1974) was one of the strongest proponents of justifying knowledge by reference to gene-based selection in biological evolution. In his later writings he adopted this second program. Evolutionary epistemology is no longer an analogy, but it is a special case of selection. Just as the processes that result in functional organization have been isolated and analyzed as a special sort of process, selection processes warrant an equally careful treatment.

Starting with Lewontin's (1970) classic statement, successive attempts have been made to specify the essence of selection processes. According to Lewontin, the three necessary elements for selection to occur are phenotypic variation (different individuals in a population have different morphologies, physiologies, and behaviors), differential fitness (different phenotypes have different rates of survival and reproduction in different environments), and heritable fitness (there is a correlation between parents and offspring in the contributions of each to future generations). According to Hull's (1980) more general characterization, the selection process is made up of three subprocesses related in very specific ways: (a) replication of information in successive generations with only minimal

modification, (b) environmental interaction that causes the preceding replication to be differential, and (c) some source of variation. The goal is to show that gene-based selection in biological evolution, the reaction of the immune system to antigens, individual learning, social learning, and other processes as well, all exhibit the basic structure of selection processes. However, such general analyses of selection do not provide any epistemological warrant. If evolutionary epistemology is to remain 'epistemology,' then some other source of epistemological warrant must be provided (Hull 2001).

6. Conclusion

Evolutionary epistemology has changed significantly from its slim beginnings. Initially, it dealt only with knowledge that is literally gene-based, e.g., a tendency to group organisms into discrete species. Evolutionary epistemology in this strict sense justified certain beliefs. However, this warrant was neither very strong nor very extensive. Evolutionary epistemology in this sense is limited only to those environmental interactions that are in the range of an organism's sense organs and that matter with respect to an organism's survival and reproduction. Evolutionary epistemology in this strict sense can provide no justification whatsoever for a belief in quantum indeterminacy.

However, the sort of evolutionary epistemology that has received the greatest attention is metaphorical in character. Knowledge is in no sense programmed into the genes to be transmitted to future generations. Instead the transmission is social. The entities being transmitted are in a broad sense memes. Numerous objections have been raised to this metaphorical evolutionary epistemology. First, so it is alleged, memes are not all that similar to genes. However, these objections stem from an unreal view of both genes and memes. The real problem is identifying something that can count as the phenotype in memetic evolution. In the absence of such a distinction, memetic evolution cannot be characterized as either Darwinian or Lamarckian. Once again, a mother dog giving her puppies fleas is not Lamarckian inheritance in either a literal or a metaphorical sense.

Whatever warrant metaphorical evolutionary epistemology provides arises from its analogical connections to literal evolutionary epistemology. Because many species of organism must be able to distinguish members of their own species to reproduce and survive, other beliefs more distantly related to reproduction and survival gain some analogical warrant. One problem with both sorts of evolutionary epistemology is that they do not distinguish between 'success' and 'truth.' Many false beliefs have been proven to be very successful with respect to survival and reproduction, e.g., the belief that God is on the side of one's own group in times of war. For this and many other reasons, a majority of professional epistemologists reject evolutionary epistemology as not being nearly good enough. Of course, professional epistemologists reach the same conclusion with respect to all other epistemological justifications as well, save their own.

A third sort of evolutionary epistemology has gained increased popularity through the years. Instead of reasoning analogically from gene-based to meme-based evolution, a general analysis of selection is produced that equally applies to all sorts of selection processes from the reaction of the immune system to antigens to social learning. As promising as this research program might well be, it has one serious drawback. It supplies no epistemological warrant, at least not in any of the traditional senses. Just because a belief arose through a selection process, it does not follow that this belief is true. It all depends on the character of the selection process. For example, a belief that survives the sort of serious testing that goes on in science is likely to be true, while biblical exegesis is not likely to increase its warrant.

However, during this same period, what counts as 'epistemology' has expanded greatly. Epistemologists are no longer limited just to justifying knowledge. They are engaged in all sorts of other activities as well. For our purposes, the issue is whether or not providing a general analysis of selection can justify certain sorts of knowledge and, if not, whether it counts as 'epistemology' in any of its more recent senses.

See also: Adaptation, Fitness, and Evolution; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Darwin, Charles Robert (1809–82); Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolution of Cognition: An Adaptationist Perspective; Evolution: Optimization; Evolutionary Theory, Structure of; Intelligence, Evolution of; Social Evolution: Overview

Bibliography

Blackmore S 1999 *The Meme Machine*. Oxford University Press, Oxford, UK

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Campbell D 1974 Evolutionary epistemology. In: Schilpp P A (ed.) *The Philosophy of Karl R. Popper*. Open Court Publishers, La Salle, IL

Cavalli-Sforza L L, Feldman M W 1981 *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University, Princeton, NJ

Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford, UK

Durham W H 1991 *Coevolution: Genes, Culture, and Human Diversity*. Stanford University Press, Stanford, CA

Eldredge N, Grene M 1992 *Interactions: The Biological Context of Social Systems*. Columbia University Press, New York, NY

Hull D L 1980 Individuality and selection. *Annual Review of Ecology and Systematics* **11**: 311–32

Hull D L 2001 In search of epistemological warrant. In:

Campbell D T (ed.) *Selection Theory and Social Construction*. State University of New York Press, Stony Brook, NY
Lewontin R C 1970 The units of selection. *Annual Review of Ecology and Systematics* **1**: 1–18
Lumsden C J, Wilson E O 1981 *Genes, Mind, and Culture: The Coevolutionary Process*. Harvard University Press, Cambridge, MA
Plotkin H 1994 *Darwin Machines and the Nature of Knowledge*. Harvard University Press, Cambridge, MA
Toulmin S 1972 *Human Understanding*. Princeton University Press, Princeton, NJ

D. L. Hull

Evolutionary Selection, Levels of: Group versus Individual

Throughout history, including the history of the social sciences, groups have been compared to individual organisms in the harmony and coordination of their parts. Aristotle compared the various classes of society to a single organism and religious communities are often described in organismic terms: the body of the church, united under the head of Christ. Almost every founder of the social sciences shared this group-level perspective, at least to a degree (Wegner 1986).

In more recent times the social sciences have become dominated by a more reductionistic perspective, sometimes referred to as 'methodological individualism,' in which 'groups and social organizations have no ontological reality—that where used, references to organizations, etc., are but convenient summaries of individual behavior' (Campbell 1994, p. 23). In addition, evolutionary biologists reached a consensus during the 1960s that groups are highly unlikely to have evolved into adaptive units comparable to individual organisms (Williams 1966). These two developments, which occurred largely independently of each other, appeared to relegate the concept of groups as organisms to the antique shop of history.

Developments in evolutionary biology since the 1960s have led to a new consensus that is represented by the term 'multilevel selection' (e.g., Frank 1998, Michod 1999, Sober and Wilson 1998). Natural selection acts on a hierarchy of biological units, from genes within individuals to groups within a metapopulation (a population of groups). Individual organisms are an upper level of this biological hierarchy, social groups that have become so functionally integrated that the whole is more conspicuous than the parts. If individuals are themselves groups, the concept of groups as individuals can no longer be regarded as theoretically improbable. In fact, an especially strong case can be made for human groups as units of natural selection (Boehm 1997). The group-level perspective is back in evolutionary biology, with profound implications for the human social sciences. However, groups can never be assumed axiomatically to function as adaptive units. Special conditions are required and opposing forces always exist, including subversion from within. The modern multilevel view can draw upon the originators of the social sciences for insight, but is not a return to the grandiose and often naïve groupism of the past.

1. Evolution and the Fundamental Problem of Social Life

Prior to Darwin, supernatural agents provided the most ready explanation of purpose and order at all levels, from celestial bodies, to human society, to the actions of individual people and other creatures. Darwin provided the first successful scientific theory of adaptations, which remains today the only theory in an ultimate sense. Human artifacts have a purpose and did not evolve by natural selection in a proximate sense, but the ability of humans to make artifacts with a purpose itself presumably evolved by natural selection. All intentional processes are probably rooted in a process of blind variation and selective retention (Campbell 1960).

Evolution explains adaptations but it is not nearly as permissive as supernatural explanations. In particular, group-level adaptations are not as easy to explain as individual-level adaptations. To see why, consider the evolution of a nonsocial adaptation, such as cryptic coloration. Imagine a population of insects that varies in the degree to which they match their background. Every generation, the most conspicuous individuals are detected and eaten by predators while the most cryptic individuals survive and reproduce. If offsprings resemble their parents, then the average individual will become more cryptic with every generation. Anyone who has beheld an insect that looks exactly like a leaf, right down to the veins and simulated herbivore damage, cannot fail to be impressed by the power of natural selection to evolve breathtaking adaptations at the individual level.

Now consider the same process for a social adaptation, such as members of a group warning one another about approaching predators. Imagine a flock of birds that vary in their tendency to scan the horizon for predators and to utter a call when one is spotted. It is not obvious that the most vigilant individuals will survive and reproduce better than the least vigilant. If scanning the horizon detracts from feeding, the most vigilant birds will gather less food than their more oblivious neighbors. If uttering a cry attracts the attention of the predator, the sentinel places itself at risk by warning others. If these suppositions are correct, then birds that do not scan the horizon and that remain silent when they do see a predator will

survive and reproduce better than their more vigilant neighbors.

These two examples show that the evolutionary concept of adaptation does not always conform to the intuitive concept, especially at the group level. It is easy to imagine a bird flock as an adaptive unit and to predict its properties. We would expect members of the flock to adopt the creed 'all for one and one for all.' We might expect sentries to be posted at all times to detect predators at the earliest possible moment and to relay the information to feeding members of the flock. Unfortunately, individuals who possess these behaviors do not necessarily survive and reproduce better than individuals who enjoy the benefits and do not share the costs. Since Darwin's theory explains adaptations only on the basis of differential survival and reproduction, it appears unable to explain groups as adaptive units. This can be called the fundamental problem of social life. Groups function best when their members produce benefits for one another, but it is difficult to translate this kind of social organization into the currency of biological fitness. Evolutionary theory has difficulty explaining any kind of group as an adaptive unit, including those that might be found in our own species.

2. *Darwin's Solution to the Fundamental Problem*

Darwin was aware of the fundamental problem of social life and proposed a solution. Suppose there is not just one flock of birds but many flocks. Furthermore, suppose that the flocks vary in their proportion of vigilant callers. It is true that vigilant callers do not have a fitness advantage within a single flock, but groups of vigilant callers will be more successful than groups whose members do not look out for one another. Darwin used this reasoning to explain examples of apparent altruism in nonhuman species, such as the bee's suicidal sting, and also human moral virtues that appear designed to promote group welfare. In his own words: 'At all times throughout the world tribes have supplanted other tribes; and as morality is one important element of their success, the standard of morality and the number of well-endowed men will thus everywhere tend to rise and increase' (Darwin 1871, p. 166).

Darwin's solution is elegant and perhaps even obvious in retrospect. If adaptations evolve by differential survival and reproduction, it makes sense that the evolution of adaptive groups requires the differential survival and reproduction of groups. Adaptation at each level of the biological hierarchy requires a corresponding process of natural selection at that level. However, Darwin's solution only shows that adaptive groups can evolve in principle. Conditions are required that may or may not exist in the real world. In addition, the adaptations favored by group selection can be opposed at other levels. In the case of our birds, group selection favors vigilance but selection within groups favors nonvigilance. If we wish to explain bird flocks as adaptive units, not only must we demonstrate a process of between-group selection, but we also must show that it overpowers within-group selection.

Darwin's solution has another disturbing implication, which demonstrates that the biological concept of adaptation remains limited, even in its multilevel form. We have already seen that adaptations at one level of the biological hierarchy can be highly maladaptive at other levels. A male lion that displaces another male and systematically kills his rival's offspring so that he can proceed to mate with the females is behaving adaptively, in a narrow biological sense of the word. However, this adaptation is good only for the individual male; not for the vanquished male, the females, the group, the species, or the ecosystem. It is possible to imagine a process of group selection that favors noninfanticidal males. After all, groups with noninfanticidal males might grow larger by producing more cubs, enabling them to outcompete groups with infanticidal males. However, this would not eliminate conflict so much as elevate it up the biological hierarchy, to the level of between-group rather than within-group competition. The most that group selection can do is evolve groups that are like organisms in the harmony and coordination of their parts. We already know what individual organisms do to one another and nothing different can be expected of groups. Harmony among groups would require a process of natural selection operating at a yet higher level.

3. *Evolutionary Theory's Wrong Turn*

Group selection has been portrayed as a process that can occur but which also must contend with other forces that pull in other directions. As previously mentioned, a consensus emerged among evolutionary biologists in the 1960s that group selection is such a weak force that it can be ignored for most purposes. Even though it is theoretically possible for groups to evolve into adaptive units, it almost never happens in the real world. Of course, group selection could not be rejected without alternative theories that explain apparent group-level adaptations in more individualistic terms. The theories that became the foundation for the evolutionary study of social behavior were kin selection (Hamilton 1964, Maynard Smith 1964), in which individuals evolve to benefit copies of their genes present in genetic relatives,
and reciprocal altruism (Trivers 1971, Maynard Smith 1982, Axelrod and Hamilton 1981), in which individuals indirectly benefit themselves by helping others. With these theories, 'biologists seemed to have little

further need for the metaphor of society as an organism' (Konner 1999, p. 31).

The rejection of group selection was hailed by evolutionary biologists as a great accomplishment but subsequent developments have shown it to be a massive wrong turn from which the field is only starting to recover (see Sober and Wilson 1998 for a detailed historical and conceptual analysis of this period). We can see why by returning to our bird example. Suppose that uttering a cry does not increase the risk of being attacked by an approaching predator. On the contrary, it tells the approaching predator that it has been spotted by the crier and that a less vigilant member of the group should be targeted. If these are the facts of the matter, then the evolution of so-called warning cries could be explained entirely by within-group selection. Individuals who call survive and reproduce better than individuals in the same group who do not call. The function of the adaptation is not to warn other members of the group but to communicate with the predator in a way that actually endangers other members of the group. We would be right to reject group selection in this case. But now suppose that our original story was correct; crying is selectively disadvantageous within groups and evolves only because groups of criers fare better than groups of noncriers. It is still the case that the average crier in the metapopulation is more fit than the average noncrier—that is just another way of saying that the crying behavior evolves—but the evolutionary force that causes the average crier to be more fit is group selection, not selection among individuals within groups.

It follows that distinguishing levels of selection requires several steps. First we must compare the fitnesses of individuals within groups. Then we must compare the fitnesses of groups in the metapopulation. Finally we must combine these effects to determine the bottom line of what evolves. It turns out that many evolutionary models do not follow this procedure. Instead, whatever evolves in the model is vacuously defined as the product of 'individual (or gene) selection,' despite the presence of groups and counteracting selection pressures within and among groups. A good example is n-person evolutionary game theory, which has become the framework for studying reciprocal altruism (Dugatkin and Reeve 1998). The term 'n-person' signifies that a large population is broken into groups of size n within which social interactions occur. The fitness of an individual is given by a payoff matrix and is determined by its own behavior and the behavior of other members of its group. Our bird flock example could easily be modeled with n-person game theory, in which the fitness of an individual depends on whether it is a caller and on the number of other callers in its group. Despite the presence of groups and fitness differences within and among groups that are plain to see in the payoff matrix, the standard procedure in evolutionary game theory is to average the fitness of the alternative behavioral types across groups and to say that the type with the highest average fitness evolves by individual-level selection. Another example is kin selection theory, which implicitly assumes that a large population is broken into groups of genetic relatives, within which social interactions occur.

Elliott Sober and I call the vacuous practice of defining anything that evolves the product of individual or gene-level selection 'the averaging fallacy' (Sober and Wilson 1998). Avoiding the averaging fallacy has two major implications for evolutionary theories of social behavior. First, kin selection and reciprocal altruism become examples of multilevel selection theory rather than alternative theories. When the groups are identified in these theories, it becomes clear that the prosocial behaviors being modeled (e.g., cooperation and altruism) are selectively disadvantageous within groups and evolve only because groups with prosocial members outperform other kinds of groups. Second, not only does multilevel selection theory include kin selection and reciprocal altruism as special cases, but it goes further to explain other cases that are beyond the imagination of these theoretical frameworks. This is fortunate, because kin selection and reciprocal altruism have always seemed poorly suited to explain the kinds of prosocial behaviors that take place in human groups.

4. *Ancestral Human Groups as Adaptive Units*

Our species evolved in small groups that are roughly approximated by the hunter-gatherer societies of today. Prior to the resurgence of multilevel selection theory, evolutionary biologists found it difficult to envision human groups as adaptive units. Genetic relatedness among members of ancestral human groups was only moderate and the principle of reciprocal altruism has always been difficult to extend beyond pairwise relationships. However, modern multilevel selection theory enables human groups to be seen as potent units of selection.

Extant hunter-gatherer societies around the world are remarkably egalitarian. It is not that hunter-gatherers lack selfish impulses, but rather that selfish impulses are effectively controlled by other members of the group. This form of guarded egalitarianism has been called 'reverse dominance' by anthropologist Chris Boehm (1993). In many animal groups, the strongest individuals are usually able to dominate their rivals, taking a disproportionate share of the resources. This is within-group selection, pure and simple. In human groups, an individual who attempts to benefit himself or herself at the expense of others is likely to encounter the combined retaliation of the rest of the group. In most cases even the strongest individual is no match for the rest of the group, so self-serving acts are effectively curtailed. Boehm's survey

of hunter-gatherer societies includes many examples of reverse domination, ranging from ridicule to ostracism, to assassination, of those who attempt to impose their will upon others. Hunter-gatherer societies are first and foremost moral communities with a strong sense of right and wrong that organizes the practices of the group.

The concept of a human group as a moral community shows how much has been missed by the concept of genetic relatedness. In a kin selection model, behavioral uniformity within groups can be achieved only by genetic uniformity. In a moral community, behavioral uniformity can be achieved (or at least approached) by a shared belief in what constitutes right behavior and an ability to discourage wrong behavior effectively. Social norms can create a degree of behavioral uniformity within groups and differences among groups that could never be predicted from their genetic structure and which radically shifts the balance between levels of selection in favor of group selection.

The concept of a human group as a moral community also shows how much has been missed by the concept of reciprocal altruism. Consider a moral community whose members believe that it is right to help others in proportion to need rather than the likelihood of return benefit. Individuals who abide by the norm are rewarded, those who violate the norm are punished, and the group (let us say) prospers compared to groups whose members restrict helping to those who will return the favor. It is true that giving according to need appears less altruistic when enforced by social controls, but social control itself emerges as a group-level adaptation when analyzed in multilevel terms (Boyd and Richerson 1992, Sober and Wilson 1998). Causing others to provide a public good is itself a public good, as economists have long realized (Heckathorn 1993).

Moral communities have both a cultural and an innate psychological component. The specific content of social norms can be modified on the basis of experience and socially transmitted from one person to another. On the other hand, the emotional reactions associated with rightdoing and wrongdoing are almost certainly innate and evolved by genetic multilevel selection. Finally, it should be obvious that wrongdoing remains as much a part of human nature as rightdoing, ready to surface in at least some individuals whenever the maintenance of social norms becomes lax. For this reason alone, human groups can never be axiomatically assumed to function as adaptive units.

Many of these principles that emerge from multilevel selection theory are supported by social psychology experiments that simulate social dilemmas (e.g., Ostrom et al. 1994). In a typical experiment, members of a newly created ten-person group are each given 10 dollars, which they can keep or anonymously put into a central pot where it will be doubled and redistributed to all members of the group. If everyone cooperates they can double their money. However, a single individual who defects keeps his or her 10 dollars and gets 18 dollars back from the central pot, providing a powerful incentive to cheat. If everyone cheats the benefits of group life are lost and no one increases their money. If a single individual cooperates in a group of cheaters, he or she gets a paltry two dollars for his or her investment of 10 dollars, providing a powerful incentive not to cooperate. This is the fundamental problem of social life in a nutshell.

Most people in this experiment initially display a moderate degree of generosity, putting about half their money into the pot, but there is variation around the average. As soon as they realize that others are cheating, the more generous members of the group become both angry and stingy. Of course, there is no outlet for anger within the framework of the experiment so the level of giving quickly drops to zero. In a modified version of the experiment, the members are allowed to contribute their money, not only to the first pot, but also to a second pot that is used to punish the cheaters. Those who contribute to the second pot are second-order public good providers. They enforce contributing to the first pot at their own expense. Nevertheless, when this outlet for righteous indignation is provided within the framework of the experiment, the second pot overflows with contributions!

Generalizing from these experiments, we can provisionally envision a human nature that is reluctant to give in ways that can easily be exploited but is willing and even eager to contribute to building and maintaining a social system that realizes the advantages of social life while preventing exploitation from within. Of course, a well-protected social system is required because at least some people, and perhaps all people some of the time, are inclined to exploit the generosity of others. Furthermore, the whole purpose of a group unified by a well-protected social system may be to exploit other groups. This complicated mix of rightdoing and wrongdoing, and behaviors that change their moral status depending upon the level at which they are viewed, may be the legacy of thousands of generations of multilevel selection operating on human populations that were subdivided into small foraging groups.

5. *Modern Human Groups as Adaptive Units*

A theory that explains small-scale human societies as (largely) adaptive units is an enormous advance over previous theories that seemed restricted to groups of genetic relatives and narrow reciprocators. However, more work is required to explain the nature of human societies at a scale above a few hundred or at most a few thousand individuals. Large-scale societies did not even exist before the advent of agriculture, which is a

short (although not negligible) period of time as far as genetic evolution is concerned (see Durham 1991, Wilson 1994 for discussions of genetic adaptations to modern environments). Nevertheless, this has not prevented cultural processes from creating a parade of new social forms that succeed or fail on the basis on their properties.

Superficially, large-scale human societies appear much less egalitarian than hunter–gatherer groups but the apparent inequalities can be interpreted in two very different ways. On the one hand, social control mechanisms are probably strongest in small groups in which everyone knows and depends on everyone else. Many inequalities that exist in large-scale societies are therefore exactly what they seem: some individuals profiting at the expense of others within the society. These inequities should not be interpreted as group-level adaptations but rather as individual-level adaptations with consequences that are often dysfunctional at the society level. On the other hand, purely from the group-level functional standpoint, societies must become differentiated and hierarchical as they increase in size. Thirty persons can sit around the campfire and arrive at a consensual decision. Thirty million persons cannot. It is therefore an open question whether extreme status differences and other seeming inequalities in large-scale societies represent pure and simple domination or design features that enable the society to function at a large scale, especially in competition with other societies. There can be little doubt that the scale of a society itself is a group-level adaptation. Larger societies tend to replace smaller societies unless their larger size is offset by problems of coordination and internal conflicts of interest. It is possible to imagine human history as a process of multilevel cultural evolution in which smaller groups coalesce into larger groups, just as for long-term biological evolution.

6. *Conclusion*

Despite its turbulent history, the concept of human groups as organismic units can be partially justified by multilevel selection theory. The implications extend the length and breadth of the social sciences; from cognition as a group-level process (Wilson 1997), to individual differences in prosocial and antisocial behaviors (Wilson et al. 1996), to historical and cultural studies (MacDonald 1994), to modern economic policy (Bowles and Gintis 1998). Social scientists have every reason to question the dominance of methodological individualism and to revisit the perspective of the originators of their own disciplines in the light of modern evolutionary theory.

See also: Adaptation, Fitness, and Evolution; Darwin, Charles Robert (1809–82); Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolution: Optimization; Evolutionary Epistemology; Evolutionary Theory, Structure of; Natural Selection

Bibliography

Axelrod R, Hamilton W D 1981 The evolution of cooperation. *Science* **211**: 1390–6

Boehm C 1993 Egalitarian society and reverse dominance hierarchy. *Current Anthropology*. **34**: 227–54

Boehm C 1997 Impact of the human egalitarian syndrome on Darwinian selection mechanics. *American Naturalist* **150**: S100–21

Bowles S, Gintis H 1998 Is equality passé? *Boston Review* **23**(6): 4–26

Boyd R, Richerson P J 1992 Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* **13**: 171–95

Campbell T D 1960 Blind variation and selective retention in creative thought and other knowledge processes. *Psychological Review* **67**: 380–400

Campbell D T 1994 How individual and face-to-face-group selection undermine firm selection in organizational evolution. In: Baum J A C, Singh J V (eds.) *Evolutionary Dynamics of Organizations*. Oxford University Press, New York, pp. 23–38

Darwin C 1871 *The Descent of Man, and Selection in Relation to Sex*. Murray, London

Dugatkin L A, Reeve H K (eds.) 1998 *Game Theory and Animal Behavior*. Oxford University Press, Oxford, UK

Durham W H 1991 *Coevolution: Genes, Culture and Human Diversity*. Stanford University Press, Stanford, CA

Frank S A 1998 *Foundations of Social Evolution*. Princeton University Press, Princeton, NJ

Hamilton W D 1964 The genetical evolution of social behavior, I and II. *Journal of Theoretical Biology* **7**: 1–52

Heckathorn D D 1993 Collective action and group heterogeneity: Voluntary provision vs. selective incentives. *American Sociological Review* **58**: 329–50

Konner M 1999 Darwin's truth, Jefferson's vision. *American Prospect* **45**: 30–8

MacDonald K B 1994 *A People that Shall Dwell Alone: Judaism as a Group Evolutionary Strategy*. Praeger, Westport, CT

Maynard Smith J 1964 Group selection and kin selection. *Nature* **201**: 1145–6

Maynard Smith J 1982 *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK

Michod R E 1999 *Darwinian Dynamics*. Princeton University Press, Princeton, NJ

Ostrom E, Gardner R, Walker J M 1994 *Rules, Games, and Common-Pool Resources*. University of Michigan Press, Ann Arbor, MI

Sober E, Wilson D S 1998 *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, Cambridge, MA

Trivers R L 1971 The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–57

Wegner D M 1986 Transactive memory: A contemporary analysis of the group mind. In: Mullen B, Goethals G R (eds.) *Theories of Group Behavior*. Springer-Verlag, New York, pp. 185–208

Williams G C 1966 Adaptation and Natural Selection: A critique of some current evolutionary thought. Princeton University Press, Princeton, NJ
Wilson D S 1994 Adaptive genetic variation and human evolutionary psychology. *Ethology and Sociobiology* **15**: 219–35
Wilson D S 1997 Incorporating group selection into the adaptationist program: A case study involving human decision making. In: Simpson J, Kendrick D (eds.) *Evolutionary Social Psychology*. Erlbaum, Mahwah, NJ, pp. 345–86
Wilson D S, Near D, Miller R R 1996 Machiavellianism: A synthesis of the evolutionary and psychological literatures. *Psychological Bulletin* **119**: 285–99

D. S. Wilson

Evolutionary Social Psychology

In 1908, McDougall's *Social Psychology* adopted an evolutionary perspective, following Darwin's (1872) suggestion that human social behaviors (including smiles, sneers, and other emotional expressions) were based in dispositions that helped our ancestors survive and reproduce. Evolutionary theory is founded on the idea of natural selection—that characteristics increasing chances of survival and reproduction proliferate because their bearers have more successful offspring. Characteristics well-suited to particular environmental demands are called adaptations. Evolutionary social psychology involves a search for evidence of adaptive mechanisms linked to human interaction.

Although people vary among themselves, both within and between cultures, certain regularities recur in human social interactions the world over. Members of all human groups communicate using complex spoken language, and share emotional expressions such as smiles and sneers, for example. Some social behavior patterns are unique to humans, some we share with many other species—the strong bonds between mammalian mothers and offspring, for instance. Similarly, other animals with slow-developing young often have biparental care of offspring, and other animals that live in small groups tend to have dominance hierarchies.

1. *General Principles Applying Across Species*

Evolutionary theorists search for principles applying across animal species. *Differential parental investment* refers to differences between parents in the amount of resources invested in offspring (Trivers 1972). As either sex increases investment in offspring, they become more selective about choosing mates. Members of the other sex consequently compete to be chosen. In mammals, females have an initially high investment (due to internal gestation and nursing). Because females in most mammalian species contribute substantially to offspring, they are in a position to choose among males demonstrating relatively superior health, attractiveness, or dominance. Hence, mammalian males, compared to females, tend to fight more among themselves, to be physically larger, and to possess more ornaments such as antlers. *Sexual selection* occurs when features are transmitted to offspring because they confer advantages in attracting the opposite sex (as in peacocks' feathers), or in competing with the same sex (as in rams' horns).

A central principle is *inclusive fitness maximization*—the idea that animals are designed not simply to survive and produce their own offspring, but to increase the probability their genes make it into future generations. Inclusive fitness explains many instances of apparently self-sacrificing behavior in animals. Ground squirrels who spot predators are more likely to make alarm calls if neighbors are closely related (Sherman 1981). By slightly increasing their own risk, humans and other animals can gain a net genetic benefit—saving other individuals with copies of their genes.

1.1 *Application to Human Social Behavior*

Evolutionary theorists mine these broad principles for insights about human behavior, and to make predictions about circumstances under which people will be more or less likely to manifest aggression, altruism, or courtship behaviors. For example, men are slightly larger than women and reach reproductive maturity at a later age. Women commit substantially fewer homicides or assaults than men, and male violence is predominantly directed at other males. Across animal species, these are hallmarks of differential parental investment and polygyny. Of 849 societies in the cross-cultural area files, polygyny (one male and multiple females) was either usual or occasional in 708, polyandry (one female and multiple males) was found in only four, and the remaining 137 were exclusively monogamous (Daly and Wilson 1983). Further, the circumstances associated with these variations follow principles found in other species (Daly and Wilson 1983, Crook and Crook 1988). For example, polyandry often involves brothers (who share genetic interests) living in resource-poor environments. Steep polygyny (as in harems) is found in circumstances of relative abundance, steep social hierarchy, and occasional famine (so a female's offspring have a better chance of surviving if she chooses an already married man from a wealthy powerful family over a single poor man). Although polygyny was an option across societies, most marriages are monogamous, because only highly successful men attracted multiple wives. Hence, ancestral males invested most of their reproductive effort in the offspring of a small number of

women (often one). As a consequence, human males are selective about long-term mates, and females will compete with one another for desirable males (Buss 1999, Kenrick and Trost 1993).

2. Cross-cultural Data

Because all human groups involve some common demands, evolutionary theorists assume that people the world over, despite cultural and ecological differences, share common human characteristics. Cross-cultural data reveal some general patterns consistent with principles of differential parental investment and sexual selection (Kenrick and Trost 1993). Adult males have killed one another many times more frequently than adult females in societies ranging from hunter-gatherers in remote jungles to urbanites in modern Asia and Europe, as far back as historical data are available (Daly and Wilson 1988). Likewise, women across societies place more emphasis than men on potential mates' resources and status (Buss 1999).

Other data on mate selection reflect a unique feature of human biology—females reach sexual maturity earlier than males, show a fertility peak in the twenties and early thirties, and then completely cease reproduction despite many remaining years of life. Consistent with this difference in reproductive lifespan, women are attracted to mates several years older than themselves, whereas men change relative preferences as they age. Young men are attracted to women their own age or older; older men prefer women much younger than themselves (Kenrick et al. 1996). These patterns are found around the world, with somewhat stronger sex differences in societies in which women begin reproducing earlier (Kenrick and Keefe 1992).

3. Experimental Data

Evolutionary principles can be used to generate experimental predictions. Experiments are not designed to test basic assumptions about natural selection, but to use the general principles as heuristics. For example, one team of researchers used inclusive fitness theory to generate predictions about helpfulness towards individuals varying in genetic relatedness (Burnstein et al. 1994). Experimental scenarios described helping situations ranging from small errands to rescuing someone in a burning building. For everyday helping, norms of politeness buffered discrepancies based on kinship and reproductive potential (grandmother was more likely to be helped than younger sister, for example). In life or death situations, however, help was focused more on closely related individuals with high reproductive potential (younger sister compared to grandmother or cousin, for example). Another series of studies examined how sex differences in mate selection criteria affect perceptual processes related to judgments of oneself and potential mates (Gutierres et al. 1999, Kenrick et al. 1994). Women rated commitment to male partners lower after exposure to socially dominant, but not physically attractive, men, whereas men's commitment showed the opposite effect. Conversely, women's self-ratings of their mating desirability were diminished more by physically attractive than by socially dominant women, with men's self-ratings showing the opposite effect (diminished by dominant, but not physically attractive men).

4. History and Future Directions

McDougall's (1908) *Social Psychology* assumed human social 'instincts.' The position was much misunderstood and abandoned without an empirical legacy. Decades later, European ethologists began successfully applying evolutionary models to social behaviors of nonhuman animals. In 1975, E. O. Wilson made a controversial suggestion that 'sociobiology' (the evolutionary analysis of animal social behavior) would eventually encompass the social sciences. This stimulated social psychologists to consider empirical applications to their field (e.g., Cunningham 1981). Part of the controversy about the approach stemmed from misunderstandings about genetic determinism. Human social behaviors vary across contexts and cultures, and social scientists assumed sociobiology implied rigid instincts insensitive to environments. Some protested that evolutionary hypotheses were untestable, or required researchers to study genetic biology or archaeology. Evolutionary theorists clarified that environmentally insensitive instincts based on deterministic genes are misconceptions, hardly relevant to invertebrates, much less to mammals (e.g., Alcock 1998, Kenrick 1995). The brains of complex animals contain not rigid programs, but environmentally sensitive psychological mechanisms (Buss and Kenrick 1998, Kenrick et al. 1998, Tooby and Cosmides 1992). Questions about testability were addressed with empirical data and competing evolutionary hypotheses (e.g., Graziano et al. 1997, Gangestad and Thornhill 1997). The position remains controversial, but has moved towards mainstream empirical science (e.g., Buss 1999, Crawford and Krebs 1998, Simpson and Kenrick 1997). Evolutionary psychologists assume the perspective could enrich research on any topic in social behavior, and that other areas of psychology including personality, psychopathology, and cognitive neuroscience are all linked by this perspective. Although the implications for mating behaviors and sex differences have been most thoroughly explored, evolutionary models have implications for understanding kinship and family relationships, intergroup relations, and social cognition. Evolutionary psychologists believe the Darwinian perspective is as relevant to the study of

human beings as it has been to the study of all other living species.

See also: Comparative Psychology; Gender Differences in Personality and Social Behavior; Genes and Behavior: Animal Models; Primate Socioecology; Primates, Social Behavior of; Social Psychology: Research Methods; Sociality, Evolution of; Writing, Evolution of

Bibliography

Alcock J 1998 *Animal Behavior*, 6th edn. Sinauer Associates, Sunderland, MA

Burnstein E, Crandall C, Kitayama S 1994 Some neo-Darwinian decision rules for altruism: Weighing cues for inclusive fitness as a function of the biological importance of the decision. *Journal of Personality and Social Psychology* **67**: 773–89

Buss D M 1999 *Evolutionary Psychology: The New Science of Mind*. Allyn and Bacon, Boston, MA

Buss D M, Kenrick D T 1998 Evolutionary social psychology. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York

Cosmides L, Tooby J 1992 Cognitive adaptations for social exchange. In: Barkow J H, Cosmides L, Tooby J (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, New York

Crawford C, Krebs D L 1998 *Handbook of Evolutionary Psychology: Ideas, Issues, and Applications*. Erlbaum, Mahwah, NJ

Crook J H, Crook S J 1988 Tibetan polyandry: Problems of adaptation and fitness. In: Betzig L, Borgerhoff-Mulder M, Turke P (eds.) *Human Reproductive Behaviour: A Darwinian Perspective*. Cambridge University Press, Cambridge, UK

Cunningam M R 1981 Sociobiology as a supplementary paradigm for social psychological research. In: Wheeler L (ed.) *Review of Personality & Social Psychology*. Sage, Beverly Hills, CA

Daly M, Wilson M 1983 *Sex, Evolution, and Behavior*, 2nd edn. Willard Grant Press, Boston

Daly M, Wilson M 1988 *Homicide*. de Gruyter, New York

Daly M, Salmon C, Wilson M 1997 Kinship: The conceptual hole in psychological studies of social cognition and close relationships. In: Simpson J A, Kenrick D T (eds.) *Evolutionary Social Psychology*. Erlbaum, Mahwah, NJ

Darwin C 1872 *The Expression of Emotion in Man and Animals*. Murray, London

Gangestad S W, Thornhill R 1997 Human sexual selection and developmental stability. In: Simpson J A, Kenrick D T (eds.) *Evolutionary Social Psychology*. Erlbaum, Mahwah, NJ

Graziano W G, Jensen-Campbell L A, Todd M, Finch J F 1997 Interpersonal attraction from an evolutionary perspective: Reactions to dominant and prosocial men. In: Simpson J A, Kenrick D T (eds.) *Evolutionary Social Psychology*. Erlbaum, Mahwah, NJ

Gutierres S E, Kenrick D T, Partch J J 1999 Beauty, dominance, and the mating game: Contrast effects in self-assessment reflect gender differences in mate selection. *Personality & Social Psychology Bulletin* **25**: 1126–34

Kenrick D T 1995 Evolutionary theory versus the confederacy of dunces. *Psychological Inquiry* **6**: 56–61

Kenrick D T, Gabrielidis C, Keefe R C, Cornelius J S 1996 Adolescents' age preferences for dating partners: Support for an evolutionary model of life-history strategies. *Child Development* **67**: 1499–511

Kenrick D T, Keefe R C 1992 Age preferences in mates reflect sex differences in human reproductive strategies. *Behavioral and Brain Sciences* **15**: 75–133

Kenrick D T, Neuberg S L, Zierk K L, Krones J M 1994 Evolution and social cognition: Contrast effects as a function of sex, dominance, and physical attractiveness. *Personality & Social Psychology Bulletin* **20**: 210–17

Kenrick D T, Sadalla E K, Keefe R C 1998 Evolutionary cognitive psychology. In: Crawford C, Krebs D (eds.) *Evolution and Human Behavior: Ideas, Issues, and Applications*. Erlbaum, Mahwah, NJ

Kenrick D T, Trost M R 1993 The evolutionary perspective. In: Beall A, Sternberg R J (eds.) *Perspectives on the Psychology of Gender*. Guilford Press, New York

Sherman P W 1981 Kinship demography and Belding's ground squirrel nepotism. *Behavioral Ecology and Sociobiology* **8**: 251–59

Simpson J A, Kenrick D T (eds.) 1997 *Evolutionary Social Psychology*. Erlbaum, Mahwah, NJ

Tooby J, Cosmides L 1992 The psychological foundations of culture. In: Barkow J H, Cosmides L, Tooby J (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, New York

Trivers R L 1972 Parental investment and sexual selection. In: Campbell B (ed.) *Sexual Selection and the Descent of Man 1871–1971*. Aldine, Chicago, IL

Wilson E O 1975 *Sociobiology: The New Synthesis*. Belknap Press of Harvard University Press, Cambridge, MA

D. T. Kenrick

Evolutionary Theory and Education

1. Introduction

In recent years, Darwin's (1859) theory of evolution has guided the theoretical and empirical research of an increasing number of social and behavioral scientists, an approach that is often called evolutionary psychology. The goals here are to consider how evolutionary theory and research in evolutionary psychology can be used to understand children's academic development and to explore related educational issues. Before these issues can be fully appreciated, an overview of the evolutionary approach to cognition and development is needed.

2. Evolution and Cognition

One basic assumption of evolutionary psychologists is that natural selection has resulted in the evolution of cognitive competencies that facilitated the survival

and reproduction of our ancestors (Cosmides and Tooby 1994). It is further assumed that most of these competencies are modular and domain specific, that is, they are supported by neural and cognitive systems that are designed to process only certain types of information. For example, there are dedicated neural and cognitive systems that process basic language sounds (e.g., 'ba,' 'pa') and different systems that process other types of information, such as the visuospatial information involved in navigating in the environment. There are, of course, more general cognitive systems that coordinate and integrate the workings of these specialized systems (Smith and Jonides 1999), but research in evolutionary psychology tends to be focused on domain-specific cognitive modules, such as those associated with language.

The issues of modularity, domain specificity, and the number and organization of any associated cognitive systems are currently debated. Nonetheless, it is clear that there is some degree of inherent and modular structure to the human brain and mind, in keeping with evolutionary theory. The associated neural and cognitive systems appear to be designed to process information corresponding to the domains of folk psychology, folk biology, and intuitive physics (Geary 1998), although there are other modules as well (e.g., for basic numerical abilities; Geary 1995). The cognitive modules associated with folk psychology include language, theory of mind (e.g., being able to make inferences about the intentions of other people), and competencies that allow people to interpret the body language and facial expressions of other people. These skills allow people to monitor and regulate dyadic social interactions and to establish and maintain social relationships. The competencies associated with folk biology include the ability to classify flora and fauna in the local ecology, and learn about the associated growth and behavioral patterns (Atran 1998, Keil 1992). This folk biological knowledge allows people in preindustrial cultures to classify and categorize local species, hunt some of these species, and use plants as medicines, for food and social rituals. Intuitive physics refers to the neural and cognitive systems that engage the physical world, and enable people to navigate in 3-dimensional space; remember the location of objects in the environment, and use objects (e.g., stones) to make tools (Shepard 1994).

The inherent structure and functioning of these modules appears to be skeletal in nature (Gelman 1990). Early in life, the associated neural and cognitive systems direct attention to and the initial processing of domain-specific information, but the normal development of these systems requires input from the environment. Environmental input, in turn, shapes the development of these cognitive modules so that they are adapted, during childhood, to local conditions—nature provides the skeletal structure of evolved cognitive domains and this structure is fleshed out with experience. For example, it appears that children in all cultures are biologically prepared to process and respond to the sounds of all human languages, but the language that eventually emerges is the specific language to which they are exposed (Kuhl et al. 1997). In other words, the neural and cognitive systems that respond to language sounds—and later enable the comprehension and production of human language—are inherent, but the normal development and functioning of these systems requires exposure to language.

3. *Evolution and Development*

A long period of development, as is found in humans, has a clear risk—death before the age of reproduction—and thus would only evolve if there were benefits that outweighed this risk. Comparative studies suggest that one purpose, and an important adaptive benefit, of delayed maturation is the accompanying ability to refine the physical, social and cognitive competencies that support survival and reproduction. As an example, a long developmental period is found in all social mammals and the length of this period increases with the increases in the complexity of the species' social system (Joffe 1997). These patterns suggest that one purpose of childhood is to practice and refine sociocognitive competencies, such as language and other social skills. In short, delayed maturation allows children to practice and refine the physical, social, and cognitive skills associated with the survival (e.g., hunting) and reproduction (e.g., parenting skills) of our ancestors.

Play, social interactions, and exploration of the environment and objects appear to be the mechanisms through which these emerging competencies are practiced and refined during development. Child-initiated social play, exploration, and so forth are intimately linked to cognitive and neural development, in that these activities provide experiences with the social, biological, and physical world. These experiences, in turn, interact with the inherent, but skeletal structure of cognitive modules and ensure their normal development and adaptation to local conditions. In this view, children are biologically prepared to learn about other people, and the biological and physical world and are inherently motivated to seek out experiences that will facilitate this learning.

4. *Implications for Education*

A basic assumption of evolutionary psychology is that modern humans evolved domain-specific cognitive abilities and behavioral strategies to deal with conditions in the environments of our ancestors, but these abilities and strategies may not always be well-suited to contemporary conditions. In fact, much formal education is 'unnatural' in that much of what children

are expected to learn in school involves tasks never encountered by our ancestors (Geary 1995, Rozin 1976). The basic goals of schools and schooling are thus to organize the activities of children so that they acquire competencies, such as the ability to read, that are important in the wider culture but have no evolutionary history. There follows discussion on some of the basic issues that arise from this perspective of education.

4.1 Evolution and Academic Development

Geary (1995) referred to language and other evolved forms of cognition as biologically primary abilities, and skills that build upon these primary abilities but are principally cultural inventions (e.g., reading) as biologically secondary abilities. The mechanisms by which evolved systems are adapted to produce secondary competencies are not yet fully understood, but appear to involve the co-optation of primary systems for secondary learning and access to knowledge implicit in these primary systems (Geary 1995, Rozin 1976).

As an example of the former, consider the relation between language, a primary ability, and reading, a secondary ability. The acquisition of reading-related abilities (e.g., word decoding) appears to involve the co-option of primary language and language-related systems, among others (e.g., visual scanning; Rozin 1976). Wagner et al. (1994), reported that individual differences in the fidelity of kindergarten children's phonological processing systems, which are basic features of the language domain, are strongly predictive of the ease with which basic reading skills (e.g., word decoding) are acquired in first grade. In other words, the evolutionary pressures that are selected for phonological processing systems, such as the ability to segment language sounds, were unrelated to reading, but these systems are used, or co-opted, when children learn how to read.

As an example of the latter, consider that the development of geometry may have been initially based on access to knowledge implicit in the primary systems that support navigation in the physical world. In the development of the basic principles of classical geometry, Euclid apparently 'started with what he thought were self-evident truths and then proceeded to prove all the rest by logic' (West et al. 1982, p. 220). For example, the implicit understanding that the fastest way to get from one place to another is to go 'as the crow flies,' was made explicit in the formal Euclidean postulate, 'a line can be drawn from any point to any point (In Euclidean geometry, a line is a straight line)' (West et al. 1982, p. 221). From an evolutionary perspective, the former reflects an implicit understanding of how to quickly get from one place to another and is knowledge that is built into the neural and cognitive systems that support navigation. The latter was discovered, that is, made explicit, by Euclid. Once explicit, this knowledge was integrated into the formal discipline of geometry and became socially transmittable and teachable.

4.2 Motivation to Learn

One very important implication of the evolutionary perspective is that the motivation to acquire school-taught secondary abilities is based on the requirements of the larger society and not on the inherent interests of children. Given the relatively recent advent of near universal schooling in contemporary societies, there is no reason to believe that the skills that are taught in school are inherently interesting or enjoyable for children to learn. In other words, one important difference between primary and secondary cognitive abilities is the level and source of motivation to engage in the activities that are necessary for their acquisition. This does not, however, preclude the self-motivated engagement in some secondary activities.

Even though reading is a secondary ability that involves the co-optation of primary language systems, many children and adults are motivated to read. The motivation to read, however, is probably driven by the content of what is being read rather than by the process itself. In fact, the content of many stories and other secondary activities (e.g., video games, television) might reflect evolutionarily relevant themes that motivate engagement in these activities, such as social relationships and social competition. Furthermore, the finding that intellectual curiosity is a basic dimension of human personality (Goldberg 1993) suggests that there will be a number of intellectually curious individuals who will pursue secondary activities. Euclid's investment in formalizing and proving the principles of geometry is one example. However, this type of discovery typically reflects the activities and insights of only a few individuals, and the associated advances spread through the larger society only by means of informal (e.g., newspapers) and formal education. The point is, the motivation to engage in the activities that will promote the acquisition of secondary abilities is not likely to be universal.

4.3 Instructional Activities

The basic structure of primary abilities is inherent (Gelman 1990), that is, the supporting neural and cognitive systems automatically orient children to relevant features of the environment (e.g., other people) and process the associated information (e.g., facial expressions or language sounds). As noted above, children are inherently motivated to seek out

experiences, for example, through social play, that ensure the appropriate development of these primary systems. In contrast, there is no inherent structure supporting the acquisition of secondary abilities, nor are most children inherently motivated to engage in the activities that are necessary for secondary learning. While this conclusion might seem self evident, it runs counter to many assumptions about children's learning in contemporary education; for example, that children are inherently motivated to learn secondary abilities and will do so through activities that involve play and social discourse.

Thus, from the evolutionary perspective, one essential goal of schooling is to provide content, organization, and structure to the teaching of secondary abilities, features that have been provided by evolution to primary abilities. Moreover, it cannot be assumed that children's inherent interests, such as social relationships, and preferred learning activities, such as play, will be sufficient for the acquisition of secondary abilities, even though they appear to be sufficient for the fleshing out of primary abilities. Instruction must therefore involve engaging children in activities that facilitate the acquisition of secondary abilities, whether or not children are inherently interested in engaging in these activities. This does not mean that play and social activities cannot be used to engage children in some forms of secondary learning. It does, however, mean that it is very unlikely that the mastery of many secondary domains (e.g., reading or algebra) will occur with only these types of primary activities.

In fact, research in cognitive and educational psychology indicates that some forms of secondary learning will require activities that differ from those associated with the fleshing out of primary abilities (see Geary 1995, for related discussion). These would include, among others, direct instruction, where teachers' provide the goals, organization and structure to instructional activities and explicitly teach basic competencies, such as how to sound out unfamiliar words or manipulate algebraic equations. The mastery of secondary domains also requires extensive exposure to the material, distributed over many contexts and oftentimes over many years, as well as extensive practice in using any associated procedures (e.g., to solve mathematics problems). Extensive exposure and practice also appear to be needed for the development of primary abilities, but this exposure and practice automatically occur as children engage in social discourse, play, and exploration. In contrast, most children will not automatically engage in the practice needed to master secondary domains, and, as a result, this practice needs to be built into instructional activities. For some domains, such as in the biological and physical sciences, mastery will also require many 'hands on' activities, as in conducting experiments, although more traditional methods will be needed as well (e.g., learning basic facts and principles, such as the theory of evolution).

In closing, although not enough is known to draw firm conclusions about which instructional practices can most effectively adapt primary cognitive systems for the secondary learning, the following principles can be used to guide future educational research. First, the process of evolution has provided the basic neural and cognitive structure to primary abilities, such as language, but for secondary abilities, such as reading, this basic structure and organization must come from instructional practices. Second, children are inherently motivated to engage in the types of activities, such as social play, that will facilitate the development of primary abilities, but it is not likely that these same activities will be sufficient for the acquisition of secondary abilities. This is because the brain and mind are inherently designed to be sensitive to and respond to primary activities, as they are related to the development of primary abilities. It cannot be assumed that the brain and mind are equally responsive to the activities that are needed to master secondary competencies, nor can it be assumed that children are inherently motivated to engage in these activities. Finally, primary abilities are universal, but secondary abilities are culturally derived. Thus, educational research must be an on-going process designed to determine the most effective means of instruction for the ever-changing array of secondary competencies needed to function in contemporary society.

See also: Cognitive Development: Child Education; Cultural Evolution: Overview; Developmental Behavioral Genetics and Education; Evolutionary Theory, Structure of; Human Cognition, Evolution of; Lifespan Development: Evolutionary Perspectives; Modularity versus Interactive Processing, Psychology of; Psychological Development: Ethological and Evolutionary Approaches

Bibliography

Atran S 1998 Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences* **21**: 547–609

Cosmides L, Tooby J 1994 Origins of domain specificity: The evolution of functional organization. In: Hirschfeld L A, Gelman S A (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, New York, pp. 85–116

Darwin C 1859 *On the Origin of Species by Means of Natural Selection*. John Murray, London

Geary D C 1995 Reflections of evolution and culture in children's cognition: Implications for mathematical development and instruction. *American Psychologist* **50**: 24–37

Geary D C 1998 *Male, Female: The Evolution of Human Sex Differences*. American Psychological Association, Washington, DC

Gelman R 1990 First principles organize attention to and learning about relevant data: Number and animate-inanimate distinction as examples. *Cognitive Science* **14**: 79–106

Goldberg L R 1993 The structure of phenotypic personality traits. *American Psychologist* **48**: 26–34
Joffe T H 1997 Social pressures have selected for an extended juvenile period in primates. *Journal of Human Evolution* **32**: 593–605
Keil F C 1992 The origins of an autonomous biology. In: Gunnar M R, Maratsos M (eds.) *Modularity and Constraints in Language and Cognition: The Minnesota Symposia on Child Psychology*. Erlbaum, Hillsdale, NJ, Vol. 25, pp. 103–37
Kuhl P K, Andruski J E, Chistovich I A, Chistovich L A, Kozhevnikova E V, Ryskina V L, Stolyarova E I, Sundberg U, Lacerda F 1997 Cross-language analysis of phonetic units in language addressed to infants. *Science* **277**: 684–6
Rozin P 1976 The evolution of intelligence and access to the cognitive unconscious. In: Sprague J M, Epstein A N (eds.) *Progress in Psychobiology and Physiological Psychology*. Academic Press, New York, Vol. 6, pp. 245–80
Shepard R N 1994 Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin and Review* **1**: 2–28
Smith E E, Jonides J 1999 Storage and executive processes in the frontal lobes. *Science* **283**: 1657–61
Wagner R K, Torgesen J K, Rashotte C A 1994 Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology* **30**: 73–87
West B H, Griesbach E N, Taylor J D, Taylor L T 1982 *The Prentice-Hall Encyclopedia of Mathematics*. Prentice-Hall, Englewood Cliffs, NJ

D. C. Geary

Evolutionary Theory, Structure of

1. *The Structure of Evolutionary Models*

Much of evolutionary theory is represented at the beginning of the twenty-first century through mathematical models, especially through the models of population genetics, of the evolution of states of a given system, both in isolation and interaction, through time. This is done by conceiving of the model as capable of a certain set of states—these states are represented by elements of a certain mathematical space, the state space. (Generally speaking, 'models' and 'systems' always refer to ideal systems; when the actual biological systems are being discussed, they are called 'empirical' or 'natural' systems.) The variables used in each mathematical model represent distinct measurable or potentially quantifiable, physical magnitudes. Classically, any particular configuration of values for these variables is a 'state' of the system, the 'state space' being the collection of all possible configurations of the variables.

The theory itself represents the behavior of the system in terms of its states: the rules or laws of the theory (i.e., laws of coexistence, succession, or interaction) can delineate various configurations and trajectories on the state space. A description of the structure of the theory itself therefore only involves the description of the set of models, which make up the theory.

Construction of a model within the theory involves assignment of a location in the state space of the theory to a system of the kind defined by the theory. Potentially, there are many kinds of systems that a given theory can be used to describe—limitations come from the dynamical sufficiency (whether it can be used to describe the system accurately and completely) and the accuracy and effectiveness of the laws used to describe the system and its changes. Thus, there are two main aspects to defining a model. First, the state space must be defined—this involves choosing the variables and parameters with which the system will be described; second, coexistence laws, which describe the structure of the system and laws of succession, which describe changes in its structure, must be defined.

Defining the state space involves defining the set of all the states the system could possibly exhibit. Certain mathematical entities—in the case of many evolutionary models, these are vectors—are chosen to represent these states. The collection of all the possible values for each variable assigned a place in the vector is the state space of the system. The system and its states can have a geometrical interpretation: the variables used in the state description (i.e., state variables) can be conceived as the axes of a Cartesian space. The state of the system at any time may be represented as a point in that space, located by projection on the various axes.

The family of measurable physical magnitudes, in terms of which a given system is defined, also includes a set of parameters. Parameters are values that are not themselves a function of the state of the system. Thus, a parameter can be understood as a fixed value of a variable in the state space—topologically, setting a parameter amounts to limiting the number of possible structures in the state space by reducing the dimensionality of the model.

Laws, used to describe the behavior of the system in question, must also be defined in a description of a model or set of models. Laws have various forms: in general, coexistence laws describe the possible states of the system, while laws of succession describe changes in the state of the system.

Let us discuss in more detail the description of the evolutionary models that make up evolutionary theory. The main items needed for this description are the definition of a state space, state variables, parameters, and a set of laws of succession and coexistence for the system. Choosing a 'state space' (and thereby, a set of state variables) for the representation of genetic states and changes in a population is a crucial part of population genetics theory.

Paul Thompson suggests that the state space for population genetics would include the physically possible states of populations in terms of genotype frequencies. The state space would be 'a Cartesian *n*-

space where 'n' is a function of the number of possible pairs of alleles in the population' (Thompson 1983 p. 223). We can picture this geometrically as n axes, the values of which are frequencies of the genotype. The state variables are the frequencies for each genotype. Note that this is a one-locus system; that is, we take only a single gene locus and determine the dimensionality of the model as a function of the number of alleles at that single locus.

Another type of single locus system, used less commonly than the one described by Thompson, involves using single gene frequencies, rather than genotype frequencies, as state variables. Some of the debates about 'genic selectionism' center around the adequacy of this state space and its parameters for representing evolutionary phenomena (see Lloyd 1988/1994, Chap. 7). With both genotype and gene frequency state spaces though, treating the genetic system of an organism as being able to be isolated into single loci involves a number of assumptions about the system as a whole. For instance, if the relative finesses of the genotypes at a locus are dependent on other loci, then the frequencies of a single locus observed in isolation will not be sufficient to determine the actual genotype frequencies. Assumptions about the structure of the system as a whole can thus be incorporated into the state space in order to reduce its dimensionality. Lewontin (1974) offers a detailed analysis of the quantitative effects of dimensionality of various assumptions about the biological systems being modeled.

Interactions between genotypes, and between one locus and another cannot be represented in a single locus model (with the exception of frequency-dependent selection), for the simple reason that they involve more than one locus (Lewontin 1974, pp. 273–81, Wimsatt 1980, pp. 226–9).

Although all single locus models should, in some sense, be grouped together, they are not all exactly the same model—each particular model has a different number of state variables, depending upon the number of alleles at that locus. Bas Van Fraassen suggests calling the general outline for each model its 'model type' (1980, p. 44). Because a model type is simply an abstraction of a model, constructed by abstracting one or more of the models parameters, a single model can be an instance of more than one model type.

Along similar lines, we may say that each model type be associated with a distinctive state space type. In the preceding example, the single locus model is to be taken as an instance of a general state space type for all single locus models, i.e., the different single locus model types are conceived as utilizing the same state space type. Alternatives, such as two-locus models, must be taken as instances of a different state space type.

'Parameters' are the values that appear in the succession and coexistence laws of a system that are the same for all possible states of the defined system. For instance, in the modification of the Hardy–Weinberg equation that predicts the frequencies of the genotypes after selection, the selection coefficient, s, appears as a parameter in the formula

$$p^2AA + 2pqAa + (1-s)q^2aa = 1 - sq^2$$

There are a variety of methods of establishing the value at which a parameter should be fixed or set in the construction of models for a given real system, e.g., simulation techniques can be used to obtain estimates of biologically important parameters. In some contexts, maximum likelihood estimations may be possible. Parameters can also be set arbitrarily, or ignored. This is equivalent to incorporating certain assumptions into the model for purposes of simplification (Levins 1968, pp. 8, 89).

One expects the values of parameters to have an impact on the system being represented; but variations in parameter values can make a larger or smaller amount of difference to the system. For instance, take a deterministic model that incorporates a parameter for mutation, μ. The rates of change of this model are virtually insensitive to realistic variations in the value of μ. Yet the selection parameters play a crucial role in this same model. A very small amount of selection in favor of an allele will have a cumulative effect strong enough to replace other alleles (Lewontin 1974, p. 267).

Population size is another case in which the value assigned to the parameter has a large impact on the model results. The parameter for effective population size, N, can play a crucial role in some models, because selection results can be quite different with a restricted gene pool size (Mayr 1967, pp. 48–50). In many of the stochastic models involved in calculating rates of evolutionary change, the resulting distributions and their moments can depend completely on the ratio of the mean deterministic force to the variance arising from random processes (Lewontin 1974, p. 268). This variance is usually proportional to $1/N$ and is related to the finiteness of population size. Thus, change in the value of the single parameter, N, can completely alter the structure represented by the theory.

The choice of parameters can also make a major difference to the model outcome. Theoreticians have choices about how to express certain aspects of the system or environment. The choice of parameters used to represent the various aspects can have a profound effect on the structure, even to the point of rendering the model useless for representing the empirical system in question. Group selection models provide a case in which choice of parameters not only alters the results of the models, but also leads to the near disappearance of the phenomenon being modeled (Wade 1978).

Some authors, when discussing genetic changes in populations, speak of the system in terms of a phenotype state space type. This makes sense, because the phenotype determines the breeding system and the action of natural selection, the results of which are reflected in the genetic changes in the population. In

his analysis of the present structure of population genetics theory, Lewontin traces a single calculation of a change in genetic state through both genotypic and phenotypic descriptions of the population. That is, according to Lewontin, population genetics theory must map the set of genotypes onto the set of phenotypes, give transformations in the phenotype space, and then map the set of phenotypes back onto the set of genotypes. We would expect, then, that descriptions of state in population genetics would be framed in terms of both genotypic and phenotypic variables and parameters. But this is not the case—the description can be in terms of either genotypic or phenotypic variables, but not both. Dynamically, then, it seems as if population genetics must operate in two parallel systems: one in genotype state space, one in phenotype state space (Lewontin 1974).

Lewontin explains that such independence of systems is illusory 'and arises from a bit of sleight-of-hand in which phenotype and genotype variables are made to appear as merely parameters that need to be experimentally determined, constants that are not themselves transformed by the evolutionary process' (1974, p. 15).

Now consider a few particular aspects and forms of the 'laws' used in these models. The most obvious differences, in laws as well as in state space types, are between deterministic and stochastic models. Coexistence laws describe the possible states of the system in terms of the state space. In the case of evolutionary theory, these laws would consist of conditions delineating a subset of the state space that contains only the biologically possible states. Laws of succession describe changes in the state of the system. In the case of evolutionary theory, dynamic laws concern changes in the genetic composition of populations.

The laws of succession select the biologically possible trajectories in the state space, with states at particular times being represented by points in the state space. The law of succession is the equation of which the biologically possible trajectories are the solutions. The Hardy–Weinberg equation is the fundamental law of both coexistence and succession in population genetics theory. Even the dynamic laws of the theory are usually used to assess the properties of the equilibrium states and steady state distributions. The Hardy–Weinberg law is a very simple, deterministic succession law that is used in a very simple state space. As parameters are added to the equation, we get different laws, technically speaking. For example, compare the laws used to calculate the frequency p' of the A allele in the next generation. Plugging only the selection coefficient against a recessive into the basic Hardy–Weinberg law, we get the recursion for the dominant allele frequency as $p' = p/(1-sq^2)$. Addition of parameters for the mutation rates yields a completely different law, $p'' = 1-\mu p' + vq'$, where μ is the rate of mutation away from the dominant allele, and where v is the rate of mutation towards it. We could consider these laws to be of a single type—variations on the basic Hardy–Weinberg law—which are usually used in a certain state space type. The actual state space used in each instance depends on the genetic characteristics of the natural system, and not usually on the parameters. For instance, the succession of a system at Hardy–Weinberg equilibrium and one that is not at equilibrium but is under selection pressure, could both be modeled in the same state space, using different laws.

A theory can have either deterministic or statistical laws for its state transitions. Furthermore, the states themselves can be either statistical or non-statistical. In population genetics models, gene frequencies that often appear in the set of state variables, thus the states themselves are statistical entities. In general, a law is deterministic if, when all of the parameters and variables are specified, the succeeding states are uniquely determined. In population genetics, this means that the initial population and parameters are all that is needed to get an exact prediction of the new population state.

In sum, the structure of evolutionary theory can be understood by examining the family of models it presents. In the case of population genetics theory, the set of model types—stochastic and deterministic, single locus or multilocus—can be understood as a related family of models. The question then becomes defining the exact nature of the relationships among them.

2. *Confirming and Testing Models*

The most obvious way to support a claim of the form 'this natural system is described by this model,' is to demonstrate the simple matching of some part of the model with some part of the natural system being described. For instance, in a population genetics model, the solution of an equation might yield a single genotype frequency value. The genotype frequency is a state variable in the model. Given a certain set of input variables (e.g., the initial genotype frequency value, in this case), the output values of variables can be calculated using the rules or laws of the model. The output set of variables (i.e., the solution of the model equation given the input values of the variables) is the 'outcome' of the model. Determining the fit between model and natural system involves testing how well the genotype frequency trajectory calculated from the model (the outcome of the model) matches that measured in the relevant natural (or experimental) populations. It can be evaluated by determining the fit of one curve (the model trajectory or coexistence conditions) to another (taken from the natural system); ordinary statistical techniques of evaluating curve-fitting are used.

Also, numerous assumptions are made in the construction of any model, and testing a model sometimes involves confirming some or all of these assumptions independently. These include assumptions about which factors influence the changes in the system, what the ranges for the parameters are, and what the mathematical form of the laws is. On the basis of these assumptions, the models take on certain features. Many of these assumptions thus have potential empirical content. That is, although they are assumptions made about certain mathematical entities and processes during the construction of the theoretical model, when empirical claims are then made about this model, the assumptions may have empirical significance.

For instance, the assumption might be made during the construction of a model that the population is panmictic, i.e., that all genotypes interbreed at random with each other. The model outcome, in this case, is still a genotype frequency trajectory, for which ordinary curve-fitting tests can be performed on the natural population to which the model is applied. But the model can have additional empirical significance, given the empirical claim that a natural system is a system of the kind described in the model. The assumption of panmixia, as a description of the population structure of the system under question, must be considered part of the system description that is being evaluated empirically. Evidence to the effect that certain genotypes in the population breed exclusively with each other (i.e., evidence that the population is not, in fact, panmictic) would undermine empirical claims about the model as a whole, other things being equal, and even if genotype frequencies provide good instances of fit. In other words, the assumption that genotypes are randomly redistributed in each generation is intrinsic to many population genetics model types. Hence, although the assumption that the population is panmictic often is taken for granted in the actual definition of the model type—i.e., in the law formula—it is interpreted empirically and plays an important role in determining the empirical adequacy of the claim.

By the same token, evidence that the assumptions of the model hold for the natural system being considered will increase the credibility of the claim that the model accurately describes the natural system. Technically, we can describe this situation as follows. From the point of view of empirical claims made about it, the model has three parts: state variables, empirically interpreted background assumptions, and those aspects and assumptions of the model that are not directly empirically interpreted. Because the possible outcomes of the model (along with the inputs) are actually values of the state variables, we can understand the input and outcome of the model as a sort of minimal empirical description of the natural system. If a model is claimed to describe accurately a natural system, at the very least, this means that the variables in which the natural system is described change according to the laws presented in the theoretical model. Under many circumstances, such models, in which only the state variables are empirically interpreted, are understood to be 'mere' calculating devices, because the isomorphism with the natural system is so limited.

Moreover, various aspects of the model—for instance, the form of its laws, or the values of its parameters—rely on assumptions made during model construction that can be interpreted empirically. The assumption of panmixia, discussed above, is an empirically interpreted background assumption of many population genetics models. Because it is empirically interpreted, the presence or absence of panmixia in the natural population in question makes a difference to the empirical adequacy of the model.

Finally, there may be aspects of the model that are not interpreted empirically at all. For instance, some parameters appearing in the laws might be theoretically determined and might have no counterpart in the natural system against which the model is compared.

Given that there can be aspects of the model structure that are not directly tested by examining the fit of the state variable curve, direct testing of these other aspects would give additional reason to accept (or reject) the model as a whole. In other words, it is taken that direct testing provides a stronger test than indirect testing, hence there is a higher degree of confirmation if the test is supported by empirical evidence. Direct empirical evidence for certain empirically interpreted aspects of the model that are not included in the state variables (and thus are confirmed only indirectly by goodness-of-fit tests), therefore provides additional support for the application of the model.

The above sort of testing of assumptions involves making sure that the empirical conditions for application of various parts of the model description actually do hold. In other words, in order to accept an explanation constructed by applying the model, the conditions for application must be verified. The specific values inserted as the parameters or fixed values of the model are another important aspect of empirical claims that can be tested. In some models, mutation rates, etc., appear in the equation—part of the task of confirming the application of the model involves making sure that the values inserted for the parameters are appropriate for the natural system being described. Finally, there is a more abstract form of independent support available, in which some general aspect of the model, for instance, the interrelation between the two variables, or the significance of a particular parameter or variable, can be supported through evidence outside the application of the model itself.

Lastly, variety of evidence, of which there are three kinds, is an important factor in the evaluation of empirical claims. I discuss three kinds of variety of

evidence here: (a) variety of instances of fit; (b) variety of independently supported model assumptions; and (c) variety of types of support, which include fit and independent support of aspects of the model.

First, let us consider variety of fit, that is, variety of instances in which the model outcome matches the value obtained from the natural system. Traditionally, variety of evidence has meant variety of fit. There are two issues that must be distinguished when considering the variety of fit. One involves the range of application of a model or model type, while the other involves the degree of empirical support provided for a single application. Both of these issues involve the notion of independence, which needs to be clarified before continuing.

One point of any claim to variety of evidence is to show that there is evidence for the hypothesis in question from different or independent sources. The notion of independence here is problematic. It is not the sort of independence that is found in the frequency interpretation of probability theory, where the independence depends only on accepting a particular reference class. Rather, in the context of scientific theories, independence is relative to whatever theories have already been accepted.

In evolutionary theory, independence is usually relative to some assumption about natural kinds. Often, empirical claims are made to the effect that a model is applicable over a certain range of natural systems. Inherent in these claims is the assumption that all of the natural systems in the range participate in some key feature or features that make it possible and/or desirable to describe them all with the same model type. Thus, there is some assumption of a natural kind (characterized by the key feature or features) whenever range of applications arises. The scientist, in making a claim that the model type is applicable to such and such systems, is making an empirical assumption about the existence of the key feature or features in the range of natural systems under question. Hence, for evolutionary models, testing for variety of fit depends on accepting an assumption about what constitutes different or independent instances of fit; this in turn, amounts to accepting a particular view of natural kinds. Part of what variety of evidence does, in a bootstrapping effect, is to confirm this original assumption about natural kinds. More technically, variety of evidence confirms the sufficiency of the parameters and state space.

Take the case in which a model type is claimed to be applicable over a more extended range than that actually covered by available evidence. This extrapolation of the range of a model can be performed by simply accepting or assuming the applicability of the model to the entire range in question. A more convincing way to extend applicability is to offer evidence of fit between the model and the data in the new part of the range. Provision of a variety of fit can thus provide additional reason for accepting the empirical claim regarding the range of applicability of the model.

For instance, a theory confirmed by ten instances of fit involving populations of size 1,000 (where population size is a relevant parameter) is in a different situation with regard to confirmation than a theory confirmed by one instance of each of ten different population sizes ranging from 1 to 1,000,000. If the empirical claims made about these two models asserted the same broad range of applicability, the latter model is confirmed by a greater variety of instances of fit. That is, the empirical claim about the latter model is better confirmed, through successful application (fits) over a larger section of the relevant range than the first model. Variety of fit can therefore provide additional reason for accepting an empirical claim about the range of applicability of a model. Variety of fit can also provide additional reason for accepting a particular empirical claim, i.e., one of the form, 'this natural system is accurately described by that model.' That is, variety of evidence can serve to increase confidence in the accuracy of any particular description of a natural system by a model.

Variety of fit is only one kind of variety of evidence. An increase in the number and kind of assumptions tested independently, i.e., greater variety of assumptions tested, would also provide additional reason for accepting an empirical claim about a model. The final sort of variety of evidence involves the mixture of instances of fit and instances of independently tested aspects of the model. In this case, the variety of types of evidence offered for an empirical claim about a model is an aspect of confirmation.

According to the view of confirmation sketched above, claims about models may be confirmed in three different ways: (a) through fit of the outcome of the model to a natural system; (b) through independent testing of assumptions of the model, including parameters and parameter ranges; and (c) through a range of instances of fit over a variety of the natural systems to be accounted for by the model, through a variety of assumptions tested, and including both instances of fit and some independent support for aspects of the model.

See also: Cultural Evolution: Theory and Models; Demographic Models; Evolution, History of; Evolution: Self-organization Theory; Evolutionary Selection, Levels of: Group versus Individual; Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems; Population Dynamics: Theory of Stable Populations

Bibliography

Levins R 1968 *Evolution in Changing Environments*. Princeton University Press, Princeton, NJ

Lewontin R C 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York
Lloyd E A 1988/1994 *The Structure and Confirmation of Evolutionary Theory*. Princeton University Press, Princeton, NJ
Mayr E 1967 Evolutionary challenges to the mathematical interpretation of evolution. In: Moorehead P S, Kaplan M M (eds.) *Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution*. Wistar Institute Press, Philadelphia, pp. 47–54
Thompson P 1983 The structure of evolutionary theory: A semantic perspective. *Studies in History and Philosophy of Science* **37**: 215–29
Van Fraassen B C 1980 *The Scientific Image*. Clarendon Press, Oxford University Press, Oxford, UK
Wade M 1978 A critical review of the models of group selection. *Quarterly Review of Biology* **53**: 101–14
Wimsatt W 1980 Reductionist research strategies and their biases in the units of selection controversy. In: Nickles T (ed.) *Scientific Discovery: Case Studies*. Reidel, Dordrecht, The Netherlands, pp. 213–59

E. Lloyd

Evolutionism, Including Social Darwinism

Evolutionism, the view that present structures and behaviors develop naturally over long periods of time, has taken two forms over the past two centuries. Nineteenth century 'classical evolutionism' (Sanderson 1990) was essentially endogenous, progressionist, and unilinear, that is, it pictured a developmental unfolding of pre-existing tendencies from lower to higher stages along similar lines. In the *Origin of Species* (1859) Charles Darwin instead stressed mechanism (natural selection) over cosmic laws; external (exogenous) pressures over internal tendencies; and a divergent or branching model of change. Whereas classical evolutionists viewed change as the logical outcome of universal laws, Darwin explained it in terms of responses to specific, historical circumstances. Social evolutionist and social selectionist theories apply these biological models to human society and behavior, directly and by analogy. Although social evolutionism and social selectionism are conceptually distinct from one another, and from social organicism and hereditarianism, critics of biologized social theory charge that all are varieties of social Darwinism.

1. Classical Evolutionism: The Nineteenth Century

Evolution in biology first appeared in Erasmus Darwin's *Zoonomia* (1794–6), Jean Baptiste Lamarck's *Philosophie Zoologique* (1809), and Robert Chambers' *Vestiges of the Natural History of Creation* (1844). Lamarck, the most important of the three, argued that traits acquired during an individual's lifetime may be passed to future generations through use-inheritance, a view many biologists embraced into the twentieth century. Although Lamarck held other views equally alien to modern biology (for example, that organisms consciously will changes in their bodily structure), Lamarckianism provided the basis for the social evolutionism that emerged after 1850.

1.1 Social Evolutionism

Varieties of social evolutionism emerged separately during the late eighteenth and early nineteenth centuries, initially in the work of the Scottish philosophers John Millar and Adam Ferguson and of the French theorists Marquis de Condorcet and A.R.J. Turgot. An anti-Enlightenment reaction after 1815 produced somewhat different versions of evolutionism in the work of Claude Henri St. Simon and Auguste Comte. Georg F. Hegel and Karl Marx contributed a third strain of evolutionism destined to play a major role in shaping events in the twentieth century.

The chief representative of the classical social evolutionism of the late nineteenth century was the British philosopher Herbert Spencer. In *Social Statics* (1851) and an essay *Progress* (1857) Spencer described evolution as a redistribution of matter and motion whereby all natural objects (including living organisms) develop from the simple (homogeneous) to the complex (heterogeneous), finally achieving equilibrium with their environment. In *First Principles* (1862) he launched a multivolume *Synthetic Philosophy*, surveying all knowledge from the perspective of evolution. In *The Principles of Biology* (1864–7) he substituted the phrase 'survival of the fittest' for Darwin's 'natural selection,' thus assuring readers that evolution is progressive.

1.2 Evolutionism in Anthropology

Although Spencer gained little following among British or American university professors, other writers won evolutionism academic respectability in the emerging social sciences. In *Ancient Law* (1861), the British jurist Sir Henry Maine traced the movement from 'status' to 'contract' as the legal basis of society from the ancient to the modern era. In *The Origin of Civilisation and the Primitive Condition of Man* (1870) the English naturalist and politician Sir John Lubbock presented prehistoric cultures as part of human social development. In *Researches into the Early History of Mankind* (1865) and later works, Edward B. Tylor, the founder of British anthropology, proposed that certain items of culture are carried as 'survivals' into later social stages. In *Ancient Society*

(1877), a study of American Indians, the American anthropologist Lewis Henry Morgan argued that human history proceeds from savagery to barbarism to civilization, each identified by its stage of technology.

The classical evolutionists' terminology registered their commitment to progress ('barbarism,' 'civilization,' 'fittest') but each insisted that progress was a mixed blessing. Although a champion of *laissez faire*, Spencer conceded that industrialism was 'extremely detrimental' to workers (Peel 1971, p. 216) and criticized the militarism and imperialism of late Victorian England. Tylor stated unequivocally that material progress did not automatically result in moral and political gain. None argued that all societies move without exception through identical stages but only that they develop along roughly similar lines—a loose unilinearism at best.

1.3 Darwinism

In the *Origin of Species* (1859) Charles Darwin focused on speciation rather than evolution in general, arguing that all species develop through a process of natural selection in a struggle for existence. Since individuals vary within species, the better adapted tend to survive and produce more offspring. New species result from changes in conditions and splits in populations. Selection operates on individuals not groups or races, selecting variations, not varieties. Although small and gradual, selected variations do not disappear due to reversion to an original type as critics contended. Individualism and gradualism remained the hallmarks of twentieth century Darwinism.

Darwin unlike Spencer did not view evolution as an *a priori* deduction from general laws. But vestiges of an older evolutionism lingered. Darwin's use of the term evolution and his eventual adoption of Spencer's phrase survival of the fittest suggested a progressionism alien to his own theory. His doctrine of 'sexual selection' echoed Lamarckian agency. Unanswered questions compounded the problem. Physicists argued that the earth was not sufficiently old to allow changes numerous enough to create so many species. Biologists insisted that minor variations in organisms would cancel out. To this last charge Darwin had no reply, being unaware of the genetic theory of mutations formulated by Gregor Mendel in 1869 but unknown in scientific circles until 1900. These vestigial assumptions and unanswered questions left an opening for pseudo-Darwinisns who publicly praised the *Origin* but ignored or muted natural selection.

1.4 Social Darwinism

The result was doubly ironic. The first irony was that Darwin helped legitimate an older tradition of evolutionism even though the vast majority of scientists, clergy, and laypersons rejected his central doctrine. Widely accepted by the 1870s, this pseudo-Darwinian evolutionism was a teleological developmentalism closer in spirit to Larmarck than to Darwin. The second irony was that Darwin's name—in the pejorative phrase 'social Darwinism'—became widely associated with attempts to relate social development to biology or even to 'natural law,' whether or not the theory in question owed any debt to Darwin. This epithet first surfaced about 1880 in the writings of socialists in Italy and France, and by the 1890s appeared in political debates in the USA and Britain.

Darwin's double-edged reputation explains the apparent determination of classical evolutionists to distance themselves from Darwinism. Although Maine, Morgan, and Tylor occasionally used the terms 'struggle,' 'fitness,' and 'natural selection' these passing references showed only that Darwin and his contemporaries shared a common milieu and, in some cases, a common debt to the British economist Thomas Robert Malthus, whose *Essay on the Principle of Population* (1798) was one source of Darwin's theory. In coining the phrase 'survival of the fittest,' Spencer attempted to reconcile the *Origin of Species* with his developmentalism. Tylor, in the preface to the second edition of *Primitive Culture*, insisted that the study followed 'on its own lines' (Sanderson 1990, p. 30) rather than being derived from Darwin or Spencer.

The celebrated opposition of Yale professor William Graham Sumner and the Smithsonian scientist Lester Frank Ward in the 1880s underscored these ironies. A leading advocate of *laissez faire*, Sumner based his economics on the Malthusian struggle of people against nature. In the early 1880s, he assigned his students Spencer's *Study of Sociology* and occasionally stated that the alternative to the survival of the fittest was the 'survival of the unfittest,' meaning that economic competition maximized success in the collective struggle against nature. His use of Spencer, still considered subversive at Yale, almost cost him his job, while his rhetoric brought immediate criticism and eventually a reputation as America's leading 'social Darwinist.' The irony was that Sumner distrusted all metaphysics and had little use for Spencer's evolutionism. In the face of criticism, he quickly dropped references to the 'unfittest' and remained ignorant of Darwin's work until late in his career. *Folkways* (1906), Sumner's sociological masterpiece, was neither evolutionist nor social selectionist.

Lester Ward, the founder of American sociology, argued in *Dynamic Sociology* (1883) and *The Psychic Factors of Civilization* (1893) that, whereas blind struggle and selection are laws of nature, the appearance of the human mind allows conscious direction of previously wasteful forces. A prominent neo-Lamarckian, Ward made use-inheritance the basis of his faith in education and advocated a 'sociocracy' in which social scientists direct public policy. Although

historians later termed Ward a 'reform Darwinist,' in tribute to his evolutionist arguments for positive government, 'social Lamarckian' is a more appropriate description.

2. Challenge and Crisis 1890–1920

Renewed interest in natural selection after 1890 altered the discussion. Attempts to demonstrate natural selection statistically, begun by Francis Galton in the 1870s, paid off in William Weldon's study of crabs (1894), the first demonstration of natural selection in nature. The Austrian biologist August Weismann argued that the egg and sperm are immune from environmental influences, making natural selection the sufficient cause of evolution. Although flawed by later standards, this 'neo-Darwinism' (as opponents and then supporters soon labeled it) undermined the tacit compromise many pseudo-Darwinians had effected between Larmarck and Darwin. The rediscovery of Mendel's laws in about 1900 by Hugo DeVries and others provided a missing key to understanding inheritance.

Ironically, however, these discoveries resulted in a temporary 'eclipse of Darwinism,' despite the now well-established belief in evolution. A school of neo-Larmarckians arose to challenge Weismann. Other anti-Darwinians defended orthogenesis, the view that evolution results from internally generated variations that are nonadaptive and immune to selection, the result potentially being the extinction of a species. Mendel's mutations, although later a building block of modern Darwinism, were seen initially as a sufficient explanation of evolution, not materials for selection. An acrimonious battle between Mendelians and biometricans, a group of statistically minded Darwinians, muddied the waters, the central issue being whether variation was sporadic and discontinuous or gradual and continuous.

2.1 Sociology and Anthropology

In American and British sociology, social and selectionist evolutionism continued to flourish. In *The Principles of Sociology* (1896), Franklin Giddings of Columbia University (New York) described social evolution as a threefold process of aggregation, association, and selection, a view perpetuated in *An Introduction to the Theory of Social Evolution* (1913) by a Giddings student, F. Stuart Chapin. In *Societal Evolution* (1915) and *The Science of Society* (1927) Sumner's protégé and successor Albert G. Keller argued that folkways and mores (Sumner's terms) were the result of variation, selection, and cultural transmission. Leonard T. Hobhouse, professor of sociology at London University, explored the evolution of mind in human history in *Mind in Evolution* (1901), *Morals in Evolution* (1906), and *Social Evolution and Political Theory* (1911).

But reaction was already underway. In Europe, the sociologists Max Weber and Émile Durkheim in different ways treated societies and cultures as discrete entities not to be evaluated by comparison with others or ranked hierarchically. Giddings himself gradually abandoned evolutionism for a rigorously objective sociology grounded in statistics, a battle led during the 1920s by his students Chapin and William Fielding Ogburn. Since the probabilistic thinking at the heart of this 'new statistics' was itself a legacy of Darwinism, evolutionism can be said to have contained the seeds of its own demise.

In anthropology Franz Boas and his disciples meanwhile attacked classical evolutionism for its use of the comparative method, for underestimating the role of cultural diffusion, and for ranking cultures from lower to higher. In the interwar years, Boasian cultural relativism informed such popular classics as Margaret Mead's *Coming of Age in Samoa* (1928) and Ruth Benedict's *Patterns of Culture* (1934).

2.2 Evolutionism and Social Reform

Evolutionism also shaped political debate from the 1880s to 1920. In *Cooperative Commonwealth* (1884), Laurence Gronlund combined Marx and Spencer to advocate socialism. In *Looking Backward* (1888), Edward Bellamy popularized a tradition of evolutionary Utopianism that later included Charlotte Perkins Gilman's *Herland* (1915). After 1890, selectionist language shaped discussion of race and gender. Science, wrote Tulane professor William Smith in *The Color Line* (1905), showed that the heights of civilization had been achieved through the 'Selection … of favored individuals and favored races' (Bannister 1979, p. 193). Whereas in *Physics and Politics* (1872) the British journalist Walter Bagehot described international competition in terms of cultural selectionism, a new breed of militarists and imperialists depicted a physicial and often bloody struggle. After 1900, eugenicists increasingly invoked Darwinian rhetoric of struggle and selection to support measures that ranged from stricter marriage laws to sterilization of the 'unfit.'

Although these new manifestations of social Darwinism (as it was now widely termed) may appear at odds with the temporary eclipse of Darwinism in biology, the appearance is deceptive and the label inaccurate. Racism and militarism were established traditions in nineteenth century thought and continued to flourish independently of any Darwinian references. Although Darwin himself sometimes echoed the racist views of contemporaries, he insisted that neither Lamarckian adaptation nor natural selection explained racial differences. Prominent racialists and even eugenicists were often outspoken anti-

Darwinians, for example Arthur de Gobineau, author of *The Inequality of the Human Races* (1853–5) and American eugenicist Charles B. Davenport, an early Mendelian. Many self-proclaimed Darwinians opposed racism and militarism. 'Peace eugenicists' stressed war's dysgenic function in killing the ablest of a nation's male population.

Darwinian metaphors, in fact, typically prefaced proposals to avoid the consequences of uncontrolled natural selection, drawing on long-standing distrust of this aspect of Darwin's theory. American and British reformers argued for government regulations and welfare measures on the ground that *laissez faire* was nothing more than a Darwinian struggle. Others called for more extreme forms of order imposed by the state: in colonial policy, sterilization laws, legalized racial segregation, or, for some socialists, in public ownership and an end to private property.

A few prominent social Darwinists were recognized extremists whose Darwinism was barely skin deep, for example Friedrich von Bernhardi author of *Germany and the Next War* (1914). When war came, the fact that Germanic names such as Ernst Haeckel and Weismann were prominently associated with extremes of Darwinism gave the charge new power. Even Friedrich Nietzche, a philosopher who rejected Darwinism, was included in the ranks of the social Darwinists. In the interwar years, and again during WW II, the phrase social Darwinism proved to be a powerful weapon against evolutionism and selectionism in the social sciences and in popular discourse.

3. *1920s–1990s: Revival and Controversy*

In the interwar years, developments in biology paved the way for a revival of evolutionism and selectionism. Combining Mendel, Darwin, and population genetics, biologists viewed evolution as acting less on individual organisms than on gene frequencies within populations, mutations providing the materials for selection rather than being the sole source of evolutionary change. This 'modern synthesis' was articulated in Theodosius Dobzhansky's *Genetics and the Origin of Species* (1937), Ernst Mayr's *Systematics and the Origin of Species* (1942), Julian Huxley's *Evolution, The Modern Synthesis* (1942), and George Gaylord Simpson's *Tempo and Mode in Evolution* (1944). In 1953, James Watson and Francis Crick successfully constructed a molecular model of DNA. In works such as Crick's *Of Molecules and Men* (1966), Jacques Monod's *From Biology to Ethics* (1969) and Francois Jacob's *The Logic of Life* (1973), these molecular biologists reduced all life to the impersonal operations of coding, copying, selection, and survival.

The newly professionalized field of ethology, the comparative study of human and animal behavior, furthered the rebiologizing of social theory. Influenced by the earlier work of Konrad Lorenz as summarized in *On Aggression* (1966), ethologists and anthropologists provided accounts of territoriality and aggression among humans in such best-selling books as Robert Ardrey's *The Territorial Imperative* (1966), Desmond Morris's *The Naked Ape* (1967), and Lionel Tiger's *Men in Groups* (1969).

Neuropsychologists meanwhile explored the link between physical states of the brain and human behavior, while behavioral geneticists identified mutations that affect behavior. In 1964, William Donald Hamilton coined the terms 'inclusive fitness' and 'kin selection' to explain patterns of animal behavior that contribute to group genetic fitness (the flocking of birds, for example, or sterile cases of insects) in Darwinian terms without resorting to 'group selection.'

3.1 *Revival in Anthropology*

In the social sciences, evolutionism and Darwinian selectionism re-emerged independently. In anthropology, social evolutionism in the 'classical' mode first resurfaced in V. Gordon Childe's, *Man Makes Himself* (1936) and *What Happened in History* (1942), and in Julian Steward's *Theory of Cultural Change* (1955). Unlike their nineteenth century predecessors, these social evolutionists were consistently materialist; eschewed an unfolding model of social change in favor of ordinary causal explanations; stressed exogenous as well as endogenous factors; and employed the comparative method judiciously. After 1960, a younger generation of anthropologists continued to work in this materialist tradition, among them Marshall Sahlins and Elman Service in *Evolution and Culture* (1960) and Marvin Harris in *Cannibals and Kings* (1977) and *Cultural Materialism* (1979).

Other anthropologists called for a return to social selectionism. Writing in 1965, Donald Campbell distinguished theories that explain the 'mechanism' of change from those that merely describe its 'fact and course,' calling his version of the former the 'variation-and-selective-retention model' (Sanderson 1990, p. 170). Although cultural variations are random, Campell argued, the cumulative effects are increasingly complex, better-adapted sociocultural forms. In *Cultural Transmission and Evolution* (1981), an ambitious example of cultural group selectionism, L. Luca Cavilli-Sforza and Marcus W. Feldman of Stanford proposed that sociocultural variations are often directed and purposeful, making their case with sophisticated mathematical modeling. In both versions, the concept of fitness was purely cultural, not biological as with sociobiology.

3.2 *Sociological Neoevolutionism*

Evolutionism and selectionism emerged belatedly and more reluctantly in sociology, a bastion of opposition to all forms of biologized social theory since the 1920s.

From the 1960s to the 1980s, sociologists Robert Nisbet and Anthony Giddens led the attack against classical and contemporary evolutionist theories (Nisbet 1969, Giddens 1984). However, in *Social Change in the Industrial Revolution* (1959), Neil Smelser offered an evolutionist explanation of modernity based on a model he developed with Talcott Parsons. Parsons extended the analysis in *Societies* (1966) and *The System of Modern Societies* (1971), the fullest statements of sociological 'neoevolutionism,' so labeled because of Parsons' efforts to distance himself from the classical evolutionism of Spencer. Avoiding all reference to cosmic law, Parsons explained that societies, as goal-oriented systems, adapt to changing environment through 'structural differentiation' and 'adaptive upgrading.' Although Parsons denied that his system was progressionist, idealist (in the sense of privileging values and ideas as causative factors), or unilinear, critics argued that his theory was teleological and ethnocentric. By the early 1970s, Parsonian neoevolutionism virtually disappeared, although a decade later there were signs of a 'Parsons revival' in Jeffrey C. Alexander's *Theoretical Logic in Sociology* (1983, Vol. 4) and Niklas Luhmann's *The Differentiation of Society* (1982).

Selectionism proved an even harder sell in sociology. Historically, Marxist sociologists rejected selectionist analogies despite the efforts of Friedrich Engels and others to give Marxism a Darwinian twist. Among non-Marxists, Max Weber rejected social selection (*soziale Auslese*) as having no explanatory value. Methodologically, sociologists could not see how Darwinian theory related to industrialization, stratification, social mobility, and other problems they addressed. In 1998, however, British sociologist Walter G. Runciman, in the wake of the success of sociobiology, called for a sociological rethinking of the 'selectionist paradigm' (Runciman 1998).

After the mid-1970s, the best-known and most controversial application of Darwinian selectionism to human society was sociobiology. In *Sociobiology: The New Synthesis* (1975) Harvard entomologist and population geneticist Edward O. Wilson presented an extreme version of genetic selectionism wherein the organism is, in a widely quoted phrase, 'only DNA's way of making more DNA.' Altruism serves this end, Wilson argued. Although he devoted only one chapter of *Sociobiology* to human behavior, Wilson expanded the ethical and political implications in his Pulitzer prize-winning *On Human Nature* (1978). Sociobiology was further popularized in David P. Barash's *Sociobiology and Behavior* (1977), Richard Dawkins's *The Selfish Gene* (1976), and Richard Alexander's *Darwinism and Human* Affairs (1979).

Criticism quickly followed. A Boston-based Sociobiology Study Group, Marxist in orientation, combined some reasonable questions regarding Wilson's biological reductionism with *ad hominem* attacks linking him to Nazi racial science and other forms of social Darwinism. In *Use and Abuse of Biology* (1976), Marshall Sahlins, having repudiated his earlier support of cultural evolutionism, launched a similar attack on the concept of 'kin selection.'

By the early 1980s, sociobiologists were exploring the interaction of genes and culture, soon termed coevolutionism, the best-known examples being Wilson and Charles J. Lumsden's *Genes, Mind, and Culture* (1981) and Robert Boyd and Peter J. Richerson's *Culture and the Evolutionary Process* (1985). Although Wilson argued that culture is important in its own right, one critic characterized coevolutionism as 'sociobiology in drag' (Sanderson 1990, p. 175) since the entire process appeared biased toward the biological.

In the 1980s, biologists resumed debate of evolutionism and selectionism. In 1968 the Japanese population geneticist Motoo Kimura launched a 'neutral theory' of the evolution of proteins, expanded in *The Neutral Theory of Molecular Evolution* (1983). Kimura demonstrated that protein formation and gene function are controlled by a regular, continuous process relatively immune to natural selection. Meanwhile, other biologists argued that entire species might be units of selection rather than outcomes of it *and* that speciation occurs sporadically and cataclysmically, not continuously and gradually, a theory of 'punctuated equilibrium' first proposed by Harvard professors Stephen Jay Gould and Niles Eldredge in 1972.

During the 1990s the evolutionist divide deepened. Anti-Darwinians under the banner of 'intelligent design' seized upon the debate among biologists to mount a new crusade. In *Darwin on Trial* (1991) Phillip E. Johnson mounted scientific arguments sufficiently sophisticated to draw a lengthy rebuttal from Gould (1992 *Scientific American* **267**:118–21). In *Darwin's Black Box* (1996), Michael J. Behe, a biologist at Lehigh University, argued that gradual, undirected evolution could not explain cellular biochemistry's 'irreducibly complex' adaptive systems.

Darwinians also pushed their arguments to new extremes. Under the banner 'evolutionary psychology' academic contributors to John Barkow et al. The Adapted Mind (1992) called for renewed study of the human mind within a Darwinian framework, an argument popularized in Matt Ridley's *The Origins of Virtue* (1997) and Steven Pinker's *How the Mind Works* (1997). British zoologist Richard Dawkins followed *The Selfish Gene* with *The Blind Watchmaker* (1986), *River Out of Eden* (1995), and *Climbing Mount Improbable* (1996), in which he explained how natural selection could produce the improbable perfection evidenced in the human eye.

Just as earlier theorists were branded and dismissed as neo-Darwinists and social Darwinists, so these new selectionists were termed 'ultra-Darwinists' or 'ultra-social Darwinists' (*Wilson Quarterly* 1996). Although the Darwinian modern synthesis continued to enjoy widespread support among mainstream biologists,

cultural and political opposition to biological and social evolutionism remained almost as strong as a century earlier.

See also: Darwin, Charles Robert (1809–82); Feminist Theory and Women of Color; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Radical Lesbian; Neurogenetics and Behavior; Sociobiology: Overview; Spencer, Herbert (1820–1903)

Bibliography

Bannister R C 1979 *Social Darwinism: Science and Myth in Anglo-American Social Thought*. Temple University Press, Philadelphia

Bowler P J 1983 *The Eclipse of Darwinism: Anti-Darwinian Evolution Theories in The Decades Around 1900*. Johns Hopkins University Press, Baltimore

Bowler P J 1984 *Evolution, The History of an Idea*. University of California Press, Berkeley, CA

Bowler P J 1988 *The Non-Darwinian Revolution*. Johns Hopkins University Press, Baltimore

Crook D P 1994 *Darwinism, War, and History: The Debate Over The Biology of War From The Origin of The Species to The First World War*. Cambridge University Press, Cambridge, UK

Degler C N 1991 *In Search of Human Nature: The Decline and Revival of Darwinism in American Social Thought*. Oxford University Press, New York

Depew D J, Weber B H 1995 *Darwinism Evolving: Systems Dynamics and the Genealogy of Natural Selection*. MIT Press, Cambridge, MA

Giddens A 1984 *The Constitution of Society*. University of California Press, Berkeley, CA

Hawkins M 1997 *Social Darwinism in European and American Thought, 1860–1945: Nature as Model and Nature as Threat*. Cambridge University Press, Cambridge, UK

Hofstadter R 1944 *Social Darwinism in American Thought, 1860–1915*. University of Pennsylvania Press, Philadelphia

Jones G 1980 *Social Darwinism and English Thought: The Interaction Between Biological and Social Theory*. Harvester Press, Brighton, UK

Kaye H L 1986 *The Social Meaning of Modern Biology: From Social Darwinism to Sociobiology*. Yale University Press, New Haven, CT

Nisbet R A 1969 *Social Change and History: Aspects of Western Theory of Development*. Oxford Universiy Press, New York

Peel J D Y 1971 *Herbert Spencer: The Evolution of a Sociologist*. Heinemann Educational, London

Pittenger M 1993 *American Socialists and Evolutionary Thought, 1870–1920*. University of Wisconsin Press, Madison, WI

Runciman W G 1998 The selectionist paradigm and its implications for sociology. *Sociology* **32**: 163–88

Sanderson S K 1990 *Social Evolutionism: A Critical History*. Blackwell, Cambridge, MA

Wilson Quarterly 1996 The still-passionate debate over Darwinism. *Wilson Quarterly* **20**: 111–12

R. C. Bannister

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Ex

Exchange in Anthropology

The terms 'gift' or 'reciprocal exchange' are used to describe the forms of exchange that anthropologists traditionally have described and analysed. These forms of exchange coexist with market exchanges such as barter, buying and selling, and money lending. Exchanges of the latter type have been the traditional concern of economists but anthropologists have also produced detailed ethnographic descriptions of markets, especially the periodic market systems found in rural areas around the world. There has been much debate in the anthropological literature about the defining characteristics of gift exchange systems, how they are distinguished from commodity exchanges, and of the manner in which they articulate with the market exchanges both locally and globally.

This article focuses on nonmarket exchanges and consists of four parts: the first examines exchange as a general concept and defines some key terms; the second looks briefly at the history of anthropological theories on the subject; the third presents some illustrative examples; and the fourth examines some attempts to understand the historical transformation of gift exchange systems in the context of globalization.

1. *The Concept of Exchange*

An exchange can be defined as a *transaction* that is *reciprocated*. A *transaction* is the process whereby one transactor transfers an object to another transactor. This can be represented formally, as

$$A \xrightarrow{x} B \qquad (1)$$

where A is the giver, B the receiver, and x the object transferred.

Reciprocation is the process whereby the initial transaction is reversed such that the receiver becomes the giver. This can be represented formally as

$$A \xleftarrow[y]{} B \qquad (2)$$

where A is now the receiver, B the giver, and y the object reciprocated. The reciprocation can be *simultaneous* in time or *delayed*. When it is *simultaneous* it takes the classic form of an exchange

$$A \underset{y}{\overset{x}{\rightleftarrows}} B \qquad (3)$$

where A gives B the object x in exchange for the object y. When the reciprocation is *delayed* in time it is called a delayed exchange. Exchanges, whether simultaneous or delayed, define three distinct relations—*value*, *property*, *and social*—the significance and meaning of which will vary depending on the social context.

Value relations refer to those that hold between the objects and can be expressed formally as x:y. These relations may be quantitative or qualitative and involve either the same object or a different object. If the qualitative dimension is held to be more significant, as it is in many culturally specific contexts, the size, shape, color, age, and history of an object is more important than the number given. Most gift exchanges involve valuations of this kind. They are analogous to that between playing cards where we find, for example, that a ten of clubs has a higher qualitative rank than a five of clubs rather than being twice its quantitative size.

Property relations refer to those relations that hold between the object and the person giving it; they can be expressed formally as A:x and B:y. Market exchanges involve private property rights that give the possessor the legal right to alienate the object in question. The exchanges that anthropologists study often involve different forms of nonprivate property rights of a personal, inalienable kind. Objects of this kind can be detached and given away but the sentimental bond between the giver and the object is often deemed not to be broken. An example here is the exchange of hair that young children sometimes make to affirm a friendship.

Social relations refer to those that hold between the transactors to an exchange. The transactors can be kin, friends, or strangers. The purpose of an exchange may be to reproduce existing relations, to break them, or to create new relations; in other cases the purpose of an exchange may be to obtain an object of utility rather than to create social relations.

Anthropological studies of exchange have tended to concentrate on those exchanges that involve qualitative value relations rather than quantitative ones, on those concerned with inalienable property relations rather than private property, and on those concerned with the establishment of social relations rather than the quest for an object of utility. This is the theore-

tical domain that the theory of reciprocity and the gift has tried to illuminate.

2. Anthropological Theories of Exchange

Anthropological theories of exchange arose out of the attempt to present a synthetic account of the pioneer ethnographic research of scholars such as Boas (1897) and Malinowski (1922) at the turn of the twentieth century. These theories, in turn, were modified and developed by subsequent ethnographic research in different historical and political contexts.

Boas (1897) described the spectacular *potlatch* of the Kwakiutl people of northwest America. This is a system of competitive gift giving among chiefs where one chief struggles to outdo a rival by out-giving him. The competition sometimes ended with chiefs destroying large amounts of property.

Malinowski (1922) described the equally spectacular *kula* system of the people of the Milne Bay area of Papua New Guinea. Here, people went on overseas expeditions to exchange high ranking armshells for necklaces.

The incremental giving in the *potlatch* system bore formal resemblance to market transactions involving interest and Boas himself used this term to describe it. The *kula*, for its part, had formal similarities with barter exchange. Malinowski, though, was of the opinion that conventional economic theory had no role to play here for this was something radically different to commodity exchange.

This theme was developed in Mauss's (1925) classic essay *The Gift*, one of the first syntheses of the comparative literature. His greatest contribution was to pose a new question that focussed on social relations between transactors rather than value relations between objects. Whereas the economists posed the question of price determination, Mauss posed the question of transactor obligation. He argued that we should not enquire into the determinants of the exchange rate of armshells for necklaces in a *kula*, but into the question of the obligations transactors are under to give, receive, and repay gifts.

This question led anthropologists to seek the answer not in the sphere of production or consumption, as the economists were wont to do, but in the sphere of reproduction and political authority, i.e., in the sphere of kinship and chieftainship.

Polanyi (1944), for example, emphasized the key role of reciprocity and redistribution and argued that the former had its basis in the 'give and take' of kinship relations and the latter in centralized political relations where the chief pooled gifts given to him and redistributed them in grand rituals. Polanyi's principal concern, however, was the 'great transformation' that led to a self-regulating market economy.

Lévi-Strauss's (1969) theory of kinship, which was developed around the same time as Polanyi's theory, is an attempt to answer Mauss' question about the obligation to give, receive, and repay. He argued that a woman given in marriage was the 'supreme gift' and that the incest taboo, whose universality he saw as deriving partly from nature, was the supreme rule of the gift. His argument has been challenged by feminist anthropologists who have shown that marriage systems are patriarchal, cultural constructions with no basis in nature.

Lévi-Strauss distinguished three forms of gift exchange—restricted, delayed, and generalized—and showed how they are the result of three different types of marriage—respectively, bilateral cross-cousin marriage, patrilateral cross-cousin marriage, and matrilateral cross-cousin marriage. Restricted exchange has the general form of simultaneous exchange, delayed exchange is where the reciprocated transaction occurs a generation later, and generalized exchange is a chain of one-way transactions that may feed back on itself to form a circle.

Sahlins (1972) developed this idea by showing how exchanges involving objects, and the form of reciprocity underlying them, are related to kinship distance; he also showed that political rank could transform the principles of reciprocity into a system of redistribution whereby gifts are channeled upwards to a chief who subsequently redistributes them to maintain his status.

Gregory (1982) has attempted to synthesize this anthropological tradition of thought with that of the political economists (Smith, Ricardo, Marx, and Sraffa) on commodities in order to develop a conceptual framework for understanding the particular situation in Papua New Guinea where gift exchange has effloresced under the impact of colonization and globalization.

Interest in the theory of the gift has continued to flourish in the last quarter of the twentieth century but it has taken new directions. Godbout and Caillé (1998) return the debate to one of Mauss' original concerns, the political implications of the gift for French capitalism and European society more generally. Anthropological theory, especially the work of Weiner (1992) and Godelier (1999), has begun to problematize those things that are kept but not given, i.e., things that are neither gifts nor commodities. Examples here include family heirlooms and the crown jewels. The theoretical and ethnographic study of these objects raises questions about lineages, inheritance, religion, and cultural identity.

3. Some Illustrative Examples

The theory of reciprocity and redistribution, and the particular theories of value, property, and social relations informing it, are best illustrated by means of some examples that illustrate the problems of taking a market-centric approach to gift exchange.

Restricted exchange is a type of reciprocity based on 'sister exchange.' Figure 1 illustrates this. Here, the

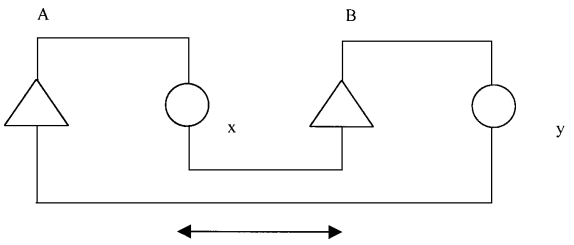

Figure 1

transactors are two men, A and B, who exchange their sisters x and y. Exchanges of this type are formally equivalent to barter but, from a social point of view, radically different. This is because the property relation between brother and sister cannot be reduced to a private property relation. The women are, from a male point of view, equivalent in some sense but the value relation cannot be reduced to that of price. Finally, an alliance is created here between two exchanging groups. Repeated exchanges of this type between the male descendants of these men create cross-cousin marriage systems, which are still of great importance in south India and many other parts of the world. In some of these places, such as the Sepik area of Papua New Guinea, the two exchanging groups are exogamous marriage moieties who maintain their alliances by seasonal exchanges of yams. The ideology is that these yams must be balanced, such that if a 2-meter long yam is received a yam of the same length must be returned. According to some theorists, yams serve as the symbolic substitutes for the women. This is called 'balanced reciprocity' and it associated with marriage of the sister-exchange type.

Delayed exchange, in Lévi-Strauss's terms, is where a man, 1A, gives his sister, x, away and receives back his sister's daughter, y, as a wife for his son, 2A, a generation later. Figure 2 illustrates this. Taking the perspective of the man 2A, it is clear that he marries a woman who is his father's sister's daughter, i.e., his patrilateral cross-cousin. Marriages of this kind are the preferred type in some areas of the highlands of Papua New Guinea. Interestingly, this is also where the famous *moka* system of incremental gift exchange is found which has led some to speculate that the *moka* gifts, too, are symbolic substitutes for women. In a *moka* system the increment, today measured in pigs, shells, and money, is called *moka* and it creates a new gift rather than being the 'interest' necessary to cancel an old debt. Thus, if 100 pigs are given and 120 are repaid at a latter date, this return gift includes 100 pigs, which cancels the old debt, and an increment of 20 pigs, the *moka*, which creates a new debt. The effect of competitive incremental gift giving is to create what Strathern (1971), the ethnographer of *moka*, calls 'alternating disequilibrium' because now one trans-

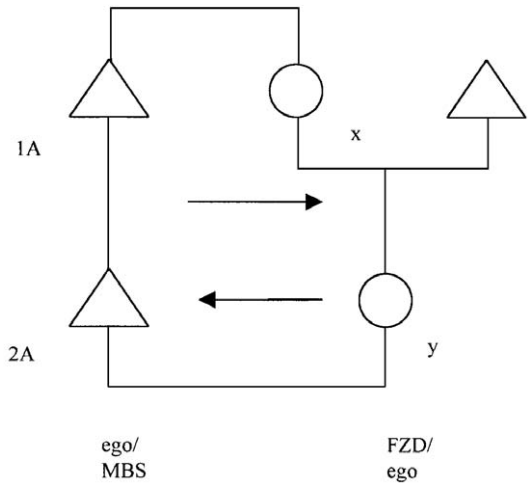

Figure 2

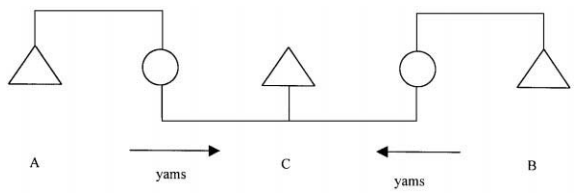

Figure 3

actor, now the other, is indebted because of the last incremental gift. The concept has no meaning in an exchange where the increment takes the form of interest. This is because interest cancels debts rather than renews them.

Redistribution, the process whereby objects are pooled by a chief and then redistributed, raises the question about the means by which the chief accumulates the objects. In places such as the Trobriand Islands in Papua New Guinea this is done by means of polygyny along with an ideology that obliges brothers to give yams, and husbands to receive them, at the annual harvest. Figure 3 illustrates this. Here, the two brothers, A and B, of the wives of the chief, C, give yams to the chief. He stores the yams in a large prestigious storehouse, which acts as a symbol of his status. On certain ceremonial occasions these are then redistributed. Redistribution transforms the ideology that it is 'better to give rather than to receive' into one which says that this is only true when a chief gives.

4. *Historical Transformations*

An early, and ingenious, attempt to understand the historical relationship between gifts and commodities

was Bohannan (1959) who sought to explain the impact of money on the Tiv economy of West Africa. For him, the problem was to understand the process by which the 'multicentric' gift economy was destroyed by the 'unicentric' commodity economy. His theory stressed the differences between the qualitative rank order of value relations characteristic of gift economies with the quantitative value relations characteristic of commodity economies. The Tiv classified the pounds, shillings, and pence of the colonial money system by quality rather than quantity; but the implicit quantitative order underlying notes and coins, he argued, 'flattened' the gift economy.

Whatever the merits of this explanation in Africa, the theory fails elsewhere, not because of any logical flaws, but because it asks the wrong question. In places like Papua New Guinea the problem to be explained was the efflorescence of gift exchange under the impact of colonization and the global market. Such was the problem Gregory (1982) sought to explain. More recent approaches of greater generality, such as the work of Appadurai (1986), focus on consumption and the 'social life of things,' the idea that things have a biography and that that they change their social form as they move from one context to another. For example, a pig may be raised as a pet, given away in a *moka* gift exchange, then sold on the market a number of times, reused in a gift exchange, and eventually given away as pork in a pig kill ceremony to be finally eaten. Other works have examined the morality of money in different social settings (Parry and Bloch 1989).

Barter, a category economists consigned to the prehistory of economic development, is emerging as a category of ethnographic and theoretical importance (Humphrey and Hugh-Jones 1992). Anthropologists working in the postsocialist states have noticed its re-emergence in these states where it seems to be related to the disintegration of the economies in these areas.

Ethnographic research over the twentieth century has presented many problems for theorists of exchange and many significant theoretical advances have been made. The twenty-first century will no doubt pose new problems for ethnographic description and analysis.

See also: Economic Anthropology; Economic Sociology; Exchange: Social; Money: Anthropological Aspects; Money, Sociology of; Trade and Exchange, Archaeology of

Bibliography

Appadurai A 1986 Introduction: Commodities and the politics of value. In: Appadurai A (ed.) *The Social Life of Things: Commodities in Cultural Perspective*. Cambridge University Press, Cambridge, UK

Boas F 1897 *Kwakiutl Ethnography*. Chicago University Press, Chicago

Bohannan P 1959 The impact of money on an African subsistence economy. In: Dalton G (ed.) *Tribal and Peasant Economies: Readings in Economic Anthropology*. The Natural History Press, Garden City, NY

Godbout J T, Caillé A 1998 *The World of the Gift* (trans. Winkler D). McGill-Queen's University Press, Montreal, PQ

Godelier M 1999 *The Enigma of the Gift*. Chicago University Press, Chicago

Gregory C A 1982 *Gifts and Commodities*. Academic Press, London

Humphrey C, Hugh-Jones S 1992 *Barter, Exchange and Value: An Anthropological Approach*. Cambridge University Press, Cambridge, UK

Lévi-Strauss C 1969 *The Elementary Structures of Kinship*, rev edn. Eyre & Spottiswoode, London

Malinowski B 1922 (1961) *Argonauts of the Western Pacific*. Dutton, New York

Mauss M 1925 (1954) *The Gift: Forms and Functions of Exchange in Archaic Societies* (trans. Cunnison I). Routledge and Kegan Paul, London

Parry J, Bloch M (eds.) 1989 *Money and the Morality of Exchange*. Cambridge University Press, Cambridge, UK

Polanyi K 1944 *The Great Transformation*. Farrar & Rinehart. Rinehart, New York

Sahlins M 1972 *Stone Age Economics*. Aldine-Athetone, Chicago

Strathern A 1971 (1975) *The Rope of Moka*. Cambridge University Press, Cambridge, UK

Weiner A B 1992 *Inalienable Possessions*. University of California Press, Berkeley, CA

C. A. Gregory

Exchange: Social

Social exchange theory has its origins in the field of sociology in the work of George C. Homans and Peter Blau. Homans' (1958) famous paper entitled 'Social Behavior as Exchange' laid the foundations of this orientation to the study of behavior. He subsequently expanded these ideas in *Social Behavior: Its Elementary Forms*, published in 1961 (1974). The primary focus of his work is what he referred to as 'sub-institutional' or elementary behavior, the actions of individuals in direct interaction with one another that form the bedrock of groups and organizations. He contrasted elementary behavior with institutionalized forms of behavior, such as that involved in conforming to norms or to role prescriptions for appropriate behavior. For Homans, institutionalized behavior and elementary forms of behavior were most often distinct. He argued that elementary behavior could sometimes 'crack the institutional crust,' forcing changes in the institutionalized ways of doing things (e.g., the changes in social behavior and institutions forged by rebellions and revolutions).

1. Major Theoretical Formulations: Blau and Homans

Blau (1964), in his famous book, *Exchange and Power in Social Life*, developed a much more extensive

treatment of the links between microlevel social behavior and the groups, organizations, and institutions it constitutes. A major difference between the perspectives of Homans and Blau is the latter's recognition of 'emergent' processes at more complex levels of social organization. For Blau the whole is more than the sum of its parts. Relationships between the parts or elements of the structure create emergent processes that evolve from the interaction of the parts, but are not reducible to properties of the individual elements. Homans, in contrast, adopted a reductionistic orientation to explanation using exchange principles arguing that many of the 'laws' of social organization could be understood in terms of the behavior of the interacting parts, typically individuals. For this reason, according to Homans, psychological principles ultimately laid the foundation for the laws of elementary social behavior.

While Homans's efforts were aimed at understanding elementary behavior, Blau's aim was much more macro in focus. He intended to formulate a theory of social structure and institutions based upon a sound microfoundation. He presents his book on social exchange as a prolog to a theory of social structure. The origin of Blau's vision of social behavior was not behaviorism, as in Homans's case; rather, it was a more sociological version of microeconomics, which he labeled 'social exchange theory,' borrowing the term from Homans. (For a lengthy critique of Blau's use of microeconomic reasoning see Heath 1976, among others.) In many ways Blau's work was one of the first major theoretical efforts to extend the logic of microeconomics to the analysis of social behavior, a task that much later won Gary Becker the Nobel Prize in Economics. Homans instead based his analysis of the social behavior he called 'exchange' on principles of behavioral analysis derived from the work of the behaviorists famous at the time he wrote. A major influence upon his theoretical work was the empirical research of B. F. Skinner, his well-known colleague at Harvard.

Both Homans and Blau addressed indirect exchange processes, though such processes form a much more significant component of Blau's theoretical formulation. Homans acknowledged the importance of the indirect ties that actors have who are in direct exchange with the same sets of individuals. For example, employees of the same employer have indirect exchange relations with each other through their direct contacts with their employer. Blau extends the analysis of indirect exchange, acknowledging the role it plays in larger and more complex collectivities. In particular, Blau views indirect exchange as the basis for coordinated action in complex social structures (and the substructures of which they are composed). This coordination of action through indirect exchange in larger collectivities necessitates mediation by common values, which can be the focus of opposition and conflict, especially as the social structures involved become more complex. Blau discusses group formation, cohesion, integration, differentiation, opposition, conflict, and dissolution as forces of structural change and stability in complex social structures. These forces play a major role in his theory of social exchange as he endeavors to provide a comprehensive microfoundation for macrosociology. Homans and Blau were both engaged in distinctive efforts to produce a coherent theoretical alternative to Parsonian functionalism, the dominant paradigm in sociology until the late 1960s.

2. *The Roots of Social Exchange Theory*

Somewhat independently, Homans and Blau developed their theories of social exchange even though each had been well-versed in functionalism and Parsons's social systems approach to the analysis of society. The vestiges of both traditions are clearly evident in Blau's treatment of social exchange, though much less so in the work of Homans who labored skillfully to ground his theory of exchange explicitly on Skinnerian principles of behavioral analysis, repudiating the need for more macrolevel concepts and principles. Blau's goal was to produce a theory of social structure by clarifying the nature of the social processes and associations that generate the more complex structures of organizations, communities, and societies. He explicitly attempted to link the sociology of everyday life popularized in the work of Simmel and Goffman with the broad theories of society characteristic of Weber and Parsons. Unlike Homans, however, his focus on the interpersonal was primarily as the basis for theorizing about more complex forms of association.

2.1 *Psychological Roots*

Besides functionalism and Parsons's version of systems theory other influences on the development of social exchange theory came from an eclectic set of sources including pragmatism, utilitarianism, behaviorism (for Homans), and conflict theory (in the case of Blau's work). Turner (1986) provides a fairly accurate and detailed account of the various roots of social exchange theory in his discussion of the development of this theoretical orientation and its impact in the field of sociology. In addition to these sources of influence, the major works of several psychologists and anthropologists provided further impetus to the development of the exchange paradigm in sociology. The dominant contribution from psychology is the work of John Thibaut and Harold Kelley (1959), *The Social Psychology of Groups*. This volume presents a lengthy treatment of the forms of interdependence among individuals engaged in various types of social interaction. In matrix format it lays

out the rewards and costs attached to different behaviors under varying conditions of interdependence. The framework is used to make predictions about behavior given different types of control over the outcomes of the other actor(s) in the situation. For example, Thibaut and Kelley (1959) distinguish behavior control from fate control in social relationships using their basic orientation to interdependence. Their work formed the basis of much experimental research over the next two decades, particularly on cooperation, competition, negotiation, and coalition formation. It also dovetailed nicely with the emergence of game theory in the social sciences as a tool for understanding strategic interactions.

2.2 Anthropological Roots

The early work of anthropologists like Levi-Strauss ([1922] 1969), Malinowski (1922), and Mauss (1925) influenced exchange theory in an entirely different way. Sociologists drew upon the fieldwork of these scholars to provide examples of the significance of social exchange, often contrasted directly with economic exchange. Malinowski (1922), for example, in his comprehensive analysis of the Kula Ring among the Trobriand Islanders, argues that the exchange of arm bands for necklaces across the islands over time was an important source of social solidarity, the significance of which far exceeded the economic value of the items exchanged. Blau (1964), and subsequently Richard Emerson (1972, 1981), developed conceptions of indirect exchange and generalized exchange based to some extent upon Malinowski's extensive fieldwork. Mauss's (1925) complex analysis of the gift also provided input to the development of theory concerning generalized exchange and reciprocity (see also Molm, 1997, on reciprocal or non-negotiated exchange).

3. Exchange Theory: Empirical Research

Homans and Blau popularized exchange theorizing in sociology in the late 1960s and early 1970s, but sustained empirical research on the topic did not begin until the mid-1970s largely as a result of the influence of Emerson's (1972) more formal theoretical work, based upon his earlier treatment of power-dependence relations (1962). Ironically, Emerson's work on power-dependence relations also influenced Blau's (1964) discussions of power and power gaining strategies. Emerson built a conception of social exchange around his fundamental insights concerning power in social relations. He, like Blau, made power and inequality central to his treatment of exchange processes. Similarly, like Blau, he viewed his theory of social exchange as the beginning steps toward formulating a general theory of social structure. In Emerson's case, however, the structures of primary interest were the networks formed as a result of the connections among a set of actors engaged in social exchange with one another. Two of Emerson's distinct contributions to exchange theory in sociology are his fundamental insight into the relational nature of power and his extension of exchange theory to analyze the social networks created by exchange relations.

Before expanding upon these two major contributions to the development of exchange theory it is worth noting that Emerson's first efforts to formulate a theory of social exchange had more in common with Homans's work than that of Blau, despite their common goal in building a theory of social structure. In the two papers written in 1967 (and subsequently published in 1972) in which Emerson laid out the foundations of his theory of exchange, he formulated a model of the actor built on principles of behavior analysis adopted from Skinner via Homans. Key concepts include reward (positive reinforcement), cost (rewards foregone), alternative exchange relations, dependence, and balance. Propositions are formulated based upon the basic reinforcement notions of behavior first introduced into exchange theory by Homans.

3.1 Emerson's Theory of Exchange Relations

A central focus of Emerson's theoretical framework is the exchange relation as the unit of analysis. As Turner (1986) notes, this move makes exchange theory less vulnerable to charges of tautology and reductionism. Emerson's goal is to present a theory of the formation, maintenance, and termination of exchange relations and the forms of social structure created by connected exchange relations (e.g., networks of exchange). Less critical in this theoretical formulation is the precise nature of the values of the actors involved or the cognitive structure of their interests. Assuming actors enter exchange relations for reasons of reward (or avoidance of negative reinforcement in alternative relations), the focus shifts to understanding aspects of the exchange relations and the structure of the networks or groups in which they are embedded. The emphasis on the social relation as the unit of analysis distinguishes Emerson's work from that of Homans and Blau and allows Emerson to present a more coherent theory of social structure.

3.2 Power and Dependence in Exchange

The primary legacy of Emerson's work on social exchange is his conception of how power is determined by the dependence of one actor upon another in an exchange relation and how dependence is structured by the nature of the connections among actors in networks of exchange opportunities that deny or

provide access to exchange alternatives. The key definition of power in Emerson's formulation is: The power of actor A over actor B in the Ax:By exchange relation (where x and y represent resources of value) increases as a function of the value of y to A, and decreases proportional to the degree of availability of y to A from alternative sources (other than B). These two factors (resource value and availability) determine the level of B's dependence upon A and thus A's power over B. That is, the power of A over B is a direct function of B's dependence upon A in the A:B exchange relation. The more dependent B is upon A, the more power A has over B. Embedding this relationship in a network of exchange opportunities creates the basis for a structural theory of power in exchange networks.

Exchange relations are connected to the extent that exchange in one relation affects or is affected by the nature of the exchange in another relation. The connection according to Emerson's work can either be positive or negative. A negative connection means that exchange in one relation reduces the amount or frequency of exchange in another exchange relation involving one of the same parties (e.g., the A–B and B–C exchange relations are negatively connected at B if exchange in the A–B relation reduces the frequency or amount of exchange in the B–C relation). A connection is positive if the amount or frequency of exchange in one relation increases the amount or frequency of exchange in an exchange relation involving at least one of the parties to both exchanges (e.g., the A–B relation is positively connected to the B–C relation if exchange in the A–B relation increases the frequency or amount of exchange in the B–C relation). These are modal cases; however, exchange in two relations may also be 'mixed,' involving both positive and negative exchange connections relating to different aspects (or dimensions) of exchange. Exchange in more complicated networks often involves both positive and negative connections (see Cook et al. 1983).

3.3 Exchange Networks

Emerson (1972, 1976, Cook and Emerson 1978, etc.) expanded this important direction of development in exchange theory in subsequent publications formulating what is now known as 'exchange network theory.' A similar perspective was developed later by Markovsky et al. (1988) that came to be known as 'network exchange theory.' It derives from Willer's (e.g., Willer and Anderson 1981) 'elementary theory' of social relations that has an affinity with exchange concepts, but is not derived from Emerson's power-dependence perspective. These traditions of work in recent research have much in common, despite the remaining subtle differences in assumptions about actor behavior, network connections, and bases of power. In all of these programs of research the main focus of the early empirical work was determining the structural determinants of the distribution of power in networks of varying configurations.

Early debates about the value of exchange theory in the social sciences centered on the role of rationality, tautology, and reductionism (see Emerson 1976). These debates gave way during the 1980s and 1990s to more sophisticated arguments in the field concerning theoretical strategy and the best ways to develop microlevel models of exchange that inform macrolevel processes. More complex theoretical formulations emerged during this period based on the results of careful empirical work, much of it experimental, testing the fundamental propositions derived from exchange theory. Emerson's formulation provided the impetus for much of this empirical work primarily because he formalized the theory to make it more precise and therefore testable. Subsequent formulations adopted a similar theoretical strategy (e.g., Markovsky et al. 1988, Bonacich 1998, Bienenstock and Bonacich 1992, Friedkin 1992, Skvoretz and Willer 1993, Yamaguchi 1996).

3.4 Key Assumptions of Exchange Theory

The key assumptions of exchange theory summarized recently by Molm (1997, see also Molm and Cook 1995, p. 210) include: (a) behavior is motivated by the desire to increase gain and to avoid loss (or to increase outcomes that are positively valued and to decrease outcomes that are negatively valued); (b) exchange relations develop in structures of mutual dependence (both parties have some reason to engage in exchange to obtain resources of value or there would be no need to form an exchange relation); (c) actors engage in recurrent, mutually contingent exchanges with specific partners over time (i.e., they are not engaged in simple one-shot transactions); and (d) valued outcomes obey the economic law of diminishing marginal utility (or the psychological principle of satiation). Based on these core assumptions various predictions are made about the behavior of actors engaged in exchange and the effects of different factors on the outcomes of exchange. The power-dependence principle, in addition, allows for the formulation of predictions concerning the effects of increasing the value of the resources involved in the exchange and the availability of resources from alternate sources.

4. Power Inequalities

Empirical work since the early 1980s has established several primary facts about exchange in networks of social relations. Power inequalities in networks of exchange are determined by access to valued resources in the network. The networks structure this access by

providing connections or links to certain actors and restricting access to others. The structure of the network determines the accessibility of resources and thus the nature of the distribution of power in the network (see, e.g., Cook and Emerson 1978, Cook et al. 1983, Markovsky et al. 1988, Whitmeyer 1999).

In exchange relations characterized by power inequalities the use of power varies depending upon whether the actors control positive or negative outcomes. Extending the theory to deal with the exercise of coercive power within social exchange relations, Molm (1997) has produced new insights into the nature of power relations in situations in which the actors have not only reward power, but also punishment power. Comparing the behavioral effects of the use of different bases of power in social exchange relations, Molm demonstrates that the results that hold for the exercise of reward power do not hold for the use of punishment. Specifically, punishment power is used much less frequently, in part because it carries with it the risk of withdrawal from the relationship on the part of the target of the punishment to the extent that the relationship can be exited. (Molm focuses primarily on exchange relations that are voluntary; that is, both parties have some degree of control over their involvement in the exchange and can exit if they so desire.) However, when punishment power is used it tends to be very effective in modifying behavior to make the relations more mutually rewarding in the long run. According to Molm (1997, p. 268): 'When punishment is administered contingently and consistently, coercion is a powerful means of getting what one wants.' However, coercive power is used rarely.

4.1 Coercive Power

Integrating the analysis of coercive power with reward power in social exchange relations allows for the investigation of a wider range of social situations in which more complex exchanges occur. Frequently, those involved in ongoing exchange relations have control over both rewards and punishments (even if it is only the withholding of rewarding behaviors). In addition, Molm's (1997) work has made a major contribution to the analysis of the strategic use of structurally based power advantages and disadvantages. As she correctly notes, the missing link in Emerson's (1962, 1972) original work on power-dependence relations was the nature of the precise mechanisms that relate structural determinants of power with the actual use of power by those in positions of power. In this effort norms of fairness or justice and attitudes toward risk play a key role. Conceptions of fairness constrain the use of power under some conditions, especially the use of coercive power, and aversion to risk makes some actors unwilling to use the structural power at their disposal for fear of loss.

4.2 Risk and Uncertainty

Recently additional empirical research has been conducted that extends exchange theories to deal with the effects of important factors such as uncertainty and risk on the nature and structure of social exchange. Facing uncertain environments, actors involved in exchange are more likely to seek to form committed exchange relations (Cook and Emerson 1978, Kollock 1994, Lawler and Yoon 1996). Commitment reduces the extent to which actors seek exchange with alternatives and thus serves to reduce the power inequalities both within the exchange relation and within the network in which the relation is embedded. As Kollock's (1994) work demonstrates, uncertainty not only results in commitment formation as a means of reducing uncertainty, it also tends to be correlated with perceptions of trustworthiness of the actors involved in the exchange relations.

Yamagishi et al. (1998) report that trust emerges in exchange relations under conditions of high uncertainty as actors form commitments to exclusive exchange in an attempt to avoid the possibility of exploitation by unknown actors who enter the exchange opportunity structure. Under conditions of low uncertainty actors are much more likely to continue to 'play the market' and to avoid forming commitments to specific partners in order to maximize their access to valued resources. (Uncertainty in these experimental studies referred to the likelihood of being exploited by a new partner in an exchange network of opportunities that changes over time.) In the current work on trust, uncertainty and vulnerability are often defined as two of the key elements of situations in which trust considerations are paramount (e.g., Heimer 2001).

5. Relational Cohesion and Social Exchange

Lawler and his collaborators (e.g., Lawler and Yoon 1993) have developed a new theory of relational cohesion based on principles of social exchange. The focus of this work is to examine the conditions under which social exchange relations emerge out of opportunities for exchange and are linked to the emergence of positive emotions about the exchange relation. These positive emotions and subsequently conceptions of cohesion or solidarity, Lawler argues, develop based upon positive evaluations of the outcomes of exchanges between actors and the frequency of their exchange. Low frequency and unfavorable (or less favorable) outcome exchanges are much less likely to lead to commitment to the relation, to positive feelings about the exchange, and to feelings of cohesiveness or solidarity (i.e., what Lawler terms a 'we-feeling'). The significance of this new line of research is that it returns to some of the earlier anthropological work that examines the links between exchange and solidarity in social relations. It also integrates into existing

theoretical conceptions of exchange theory important considerations of the emotional bases of exchange and commitment.

6. Applications of the Theory

Applications of exchange theory and power-dependence concepts extend from dyadic relations like dating and longer-term relationships including marriage (e.g., Sprecher and Metts 1999) to often short-term relations characteristic of various practitioner–client interactions (such as physician–patient meetings under managed care, see Grembowski et al. 1998). In these contexts the focus of the analysis is often on the nature of the relationship and the factors that predict satisfaction (or dissatisfaction) and longevity (or termination) of the relation. Two dimensions that are frequently investigated in this respect are the degree of power imbalance in the relationship and the extent to which the relationship is embedded in a network of other important exchange relationships. Power imbalance in partnerships often creates additional tension in the relation as actors try to manage the resource differences between the parties to the exchange and attempt to maintain commitments in the face of alternatives. Within an organizational structure such as that produced under managed care in the health care delivery system, choice of provider is often limited exacerbating patient concerns with their dependence upon the available physicians. Mechanisms are put into place to provide patients in these settings access to redress if the power of the physician is abused. These include increased professional and utilization reviews by colleagues, administrators, and those who insure the patients. In all cases in which a power imbalance exists in an exchange relation there is the potential for exploitation, thus widely differing mechanisms have been created that serve to provide some monitoring to insure the equitable and fair treatment of those who are most dependent and thus vulnerable to the abuse of power. A prime example is the oversight demanded frequently in nursing homes that must deliver care to one of the most dependent populations in the USA, the infirm aged. Aspects of both power-dependence theory and exchange theory have proven useful in the study of power processes in many contexts.

At the more macro level, exchange networks now represent an interesting framework for analyzing the Internet revolution in business, as well as the reorganization of the workplace and the rapid increase in interpersonal connections over the World Wide Web. As this encyclopedia is going to press the world of business relations across the globe is being transformed, as B2B (business to business) commerce becomes the dominant mode of conducting transactions at the macro level. Business to client (B2C) relations has now been superseded in significance by the global transformation of corporate relations. These relations represent complex networks of exchange and trade that include not only economic, but also political and social dimensions of transnational significance. Future research will likely see the expansion of exchange analyses of power in global networks of trade embedded in complex social and political realities.

See also: Groups, Sociology of; Integration: Social; Interactionism: Symbolic; Labor, Division of; Malinowski, Bronislaw (1884–1942); Mauss, Marcel (1872–1950); Networks: Social; Power in Society; Structure: Social; Trust, Sociology of; War: Anthropological Aspects

Bibliography

Bienenstock E J, Bonacich P 1992 The core as a solution to exclusionary networks. *Social Networks* **14**: 231–43

Bienenstock E J, Bonacich P 1993 Game-theory models for exchange networks: Experimental results. *Sociological Perspectives* **36**: 117–35

Blau P M 1964 *Exchange and Power in Social Life*. Wiley, New York

Bonacich P 1998 A behavioral foundation for a structural theory of power in exchange networks. *Social Psychology Quarterly* **61**: 185–98

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Cook K S, Emerson R M 1978 Power, equity and commitment in exchange networks. *American Sociological Review* **43**: 721–39

Cook K S, Emerson R M, Gillmore M R, Yamagishi T 1983 The distribution of power in exchange networks: Theory and experimental results. *American Journal of Sociology* **89**: 275–305

Cook K S, Gillmore M 1984 Power, dependence, and coalitions. In: Lawler E J (ed.) *Advances in Group Processes* (Vol. 1, pp. 27–58). JAI, Greenwich, CT

Cook K S, Yamagishi T 1992 Power in exchange networks: A power-dependence formulation. *Social Networks* **14**: 245–65

Emerson R M 1962 Power-dependence relations. *American Sociological Review* **27**: 31–41

Emerson R M 1972 Exchange theory, Part I: A psychological basis for social exchange; Exchange theory, Part II: Exchange relations and network structures. In: Berger J, Zelditch M Jr, Anderson B (eds.) *Sociological Theories in Progress*. Houghton Mifflin, Vol. 2, pp. 38–87

Emerson R M 1976 Social exchange theory. *Annual Review of Sociology* **2**: 335–62

Emerson R M 1981 Social exchange theory. In: Rosenberg M, Turner R (eds.) *Social Psychology: Sociological Perspectives*. Basic Books, New York, pp. 30–65

Friedkin N E 1992 An expected value model of social power: Predictions for selected exchange networks. *Social Networks* **14**: 213–29

Grembowski D, Cook K S, Patrick D, Roussell A 1998 Managed care and physician referral: A social exchange perspective. *Medical Care Research and Review* (March)

Heath A F 1976 *Rational Choice and Social Exchange: A Critique of Exchange Theory*. Cambridge University Press, Cambridge, UK

Heckathorn D 1983 Extensions of power-dependence theory: The concept of resistance. *Social Forces* **61**: 1206–31
Heimer C 2001 Solving the problem of trust. In: Cook K S (ed.) *Trust in Society*. Russell Sage Foundation, New York, pp. 40–88
Homans G C 1958 Social behavior as exchange. *American Journal of Sociology* **62**: 597–606
Homans G C [1961] 1974 *Social Behavior: Its Elementary Forms*, 2nd edn. Harcourt, Brace, & World, New York
Kollock P 1994 The emergence of exchange structures: An experimental study of uncertainty, commitment and trust. *American Journal of Sociology* **100**: 313–45
Lawler E J, Yoon J 1993 Power and the emergence of commitment behavior in negotiated exchange. *American Sociological Review* **58**: 465–81
Lawler E J, Yoon J 1996 Commitment in exchange relations: Test of a theory of relational cohesion. *American Sociological Review* **61**: 89–108
Lawler E J, Yoon J 1998 Network structure and emotion in exchange relations. *American Sociological Review* **63**: 871–94
Levi-Strauss C [1922] 1969 *The Elementary Structures of Kinship* (Rev. edn.). Beacon, Boston
Lovaglia M J, Skvoretz J, Willer D, Markovsky B 1995 Negotiated exchanges in social networks. *Social Forces* **74**: 123–55
Malinowski B 1922 *Argonauts of the Western Pacific*. Routledge, London
Markovsky B 1987 Toward multilevel sociological theories: Simulations of actor and network effects. *Sociological Theory* **5**: 101–17
Markovsky B 1995 Developing an exchange network simulator. *Sociological Perspectives* **38**: 519–45
Markovsky B, Skvoretz J, Willer D, Lovaglia M J, Erger J 1993 The seeds of weak power: Extending network exchange theory. *American Sociological Review* 197–209
Markovsky B, Willer D, Patton T 1988 Power relations in exchange networks. *American Sociological Review* **53**: 220–36
Mauss M 1925 *Essai sur le Don en Sociologie et Anthropologie*. Presses Universitaires de France, Paris
Molm L D 1991 Affect and social exchange: Satisfaction in power-dependence relations. *American Sociological Review* **56**: 475–93
Molm L D 1997 *Coercive Power in Social Exchange*. Cambridge University Press, Cambridge, UK
Molm L D, Cook K S 1995 Social exchange and exchange networks. In: Cook K S, Fine G A, House J S (eds.) *Sociological Perspectives on Social Psychology*. Allyn and Bacon, Needham Heights, MA, pp. 209–35
Molm L D, Peterson G, Takahashi N 1999 Power in negotiated and reciprocal exchange. *American Sociological Review* **64**: 876–90
Skvoretz J, Lovaglia M J 1995 Who exchanges with whom: Structural determinants of exchange frequency in negotiated exchange networks. *Social Psychology Quarterly* **58**: 163–77
Skvoretz J, Willer D 1993 Exclusion and power: A test of 4 theories of power in exchange networks. *American Sociological Review* **58**: 801–18
Sprecher S, Metts S 1999 Romantic beliefs: Their influence on relationships and patterns of change over time. *Journal of Social and Personal Relationships* **16**: 834–51
Thibaut J W, Kelley H H 1959 *The Social Psychology of Groups*. Wiley, New York
Turner J 1986 *The Structure of Sociological Theory*, 4th edn. Dorsey, Chicago
Whitmeyer J M 1994 Social structure and the actor: The case of power in exchange networks. *Social Psychology Quarterly* **57**: 177–89
Whitmeyer J M 1999 Interest-network structures in exchange networks. *Sociological Perspectives* **42**: 23–47
Willer D 1992 Predicting power in exchange networks: A brief history and introduction to the issues. *Social Networks* **14**: 187–211
Willer D, Anderson B 1981 *Networks, Exchange and Coercion: The Elementary Theory and Its Applications*. Elsevier, New York
Willer D, Skvoretz J 1997 Games and structures. *Rationality and Society* **9**: 5–35
Yamagishi T, Cook K S 1993 Generalized exchange and social dilemmas. *Social Psychology Quarterly* **56**: 235–48
Yamagishi T, Cook K S, Watabe K 1998 Uncertainty, trust and commitment formation in the United States and Japan. *American Journal of Sociology* **104**: 165–94
Yamagishi T, Gillmore M R, Cook K S 1988 Network connections and the distribution of power in exchange networks. *American Journal of Sociology* **93**: 833–51
Yamaguchi K 1996 Power in networks of substitutable and complementary exchange relations: A rational-choice model and an analysis of power centralization. *American Sociological Review* **61**: 308–32

K. S. Cook

Executive Branch, Government

'Executive' has two meanings, one of which is a subset of the other. It can either be one branch of government (to be distinguished from the legislative and the judiciary branches), including the head of state, the government, political staff, the civil service, the armed forces, and all other government employees engaged in the implementation of laws (down to policemen, school teachers, etc.). This is the broadest definition. The second meaning of 'executive' is that of the political part of the executive branch. There is, however, no easy answer where precisely to draw the boundaries between the political part and the remainder of the executive. Consequently, the term 'executive' means different things not only between countries and periods but often also between authors writing about the same countries or periods.

1. Origin

Executives in the modern sense emerged through the piecemeal splitting-off of state functions from a traditionally undivided central government (mostly a monarch) (Shepard 1931, King 1975). In order to limit the government's power, judicial functions were transferred to courts and legislative functions to parliaments. This process began in twelfth- and thirteenth-century England. It had many national

variations and in Europe was not completed before the twentieth century. The constitutional doctrine of the separation of powers—as developed first and foremost by Locke, Montesquieu, and Madision—provided a normative justification for the separation of legislative, judicial, and executive institutions which was seen to guarantee liberty and justice (Vile 1967).

In practice the state functions were never separated as neatly as envisaged by normative theorists. The executive has retained important legislative functions, in particular drafting legislation and issuing government decrees and ordinances (Carey and Shugart 1998). With the full democratization of polities and political parties establishing themselves as the main mechanism to structure elections and to co-ordinate incumbents, executives have gained almost a monopoly in lawmaking in parliamentary systems. In presidential systems this is not true to the same extent, but executives also exercise large influence on legislation. Although recognized early in the literature, this gap between the division of labor between the branches of government as laid down in the constitution and political practice has received much attention in empirical studies of parliamentary systems.

2. *Types of Executives*

The shape of the executive and executive–legislative relations traditionally have been crucial for the classification of regime types. In the contemporary world the most important regime types are (a) presidential systems, (b) parliamentary or cabinet systems, and (c) semipresidential systems. In presidential systems the chief executive is elected directly for a fixed term and appoints and dismisses the members of the government. These cannot be dismissed by parliament. In a parliamentary or cabinet system the executive, whether appointed by a head of state or directly created by parliament, remains fully accountable to parliament. Although the prime minister typically has a gate-keeper function, at least formally controlling access to and dismissal from cabinet, formal decision-making is collective for important matters. Less important matters fall in the jurisdiction of individual cabinet ministers. A semipresidential system combines a directly elected and powerful head of state with a prime-minister and government who are accountable to parliament.

There is, however, no full agreement of how precisely to define these regime types or whether the semipresidential type indeed constitutes a distinct regime type (Sartori 1997, Lijphart 1999, Elgie 1999, Strøm 2000). In addition to these widespread regime types, those of assembly government and collegial (or 'directorate') government can be identified. In assembly government, decisions are indeed made in parliament and the cabinet is not more than the assembly's executive secretariat. With the exception of the French Third Republic, it has existed in its pure form only for short periods, often after a revolutionary regime change. There is only one case of collegial (or 'directorate') government—Switzerland. There, the cabinet is elected by parliament for a fixed term, that is, it is not accountable to the assembly, and it is fully collegial, with even the chairmanship rotating among its members.

Nondemocratic regimes typically claim to conform to one of the regime types presented here, though they may give other actors, such as the 'state party,' special constitutional status.

While these five regime types have belonged to the arsenal of political science for many decades, a new demand for constitutional engineering (which came with the new waves of democratization of the 1970s and particularly the 1980s and 1990s) and the new theoretical research agenda of neo-institutionalism have created new interest in them. The questions which are common to constitutional engineers and neo-institutional theorists are: what types of output are these regime types likely to generate in terms of the maintenance of democracy, the accountability of public officials, and public policy? The best research has demonstrated that each regime type is likely to maximize some effects, but at the costs of others. Hence, the appropriateness of regime types and of specific institutions within the confines of a regime type depend on the most important challenges a polity faces (Weaver and Rockman 1993, Linz and Valenzuela 1994, Moe and Caldwell 1994, Sartori 1997, Lijphart 1999, Strøm 2000).

3. *The Study of Executives*

Research on the political executive is a large and increasingly diverse field, including studies on presidents and prime ministers, cabinets, the top layer of bureaucracy, their interaction, and their interaction with other actors.

Classic institutionalism had not much to say about the executive *per se*. The executive mainly was characterized through its relations with other state institutions, which in turn were important for the classification of regime types. The behavioral revolution inspired many studies in the elite research tradition, which compiled and analyzed data on the social background and career patterns of people in executive functions (e.g., Blondel 1980, Blondel and Thiébault 1991). While certainly valuable to demonstrate cross-national differences and the impact of social and political change of great magnitude, it has remained doubtful that inferences can be made from this data on the behavior of incumbents. At best, plausible hypotheses have been proposed, for example, inferring the relative ineffectiveness of incumbents from short office periods. The bulk of research on the executive, however, always has been historic–qualitative, building on *mémoires*, media reports, the occasional interview, private papers of former members of the

executive, and other archival sources. Notwithstanding some impressive individual scholarship, this research traditionally has suffered from three problems. First, with the exception of the US presidency, research on the executive is rare, in particular given the relevance of its subject and compared to the attention other political phenomena have received. Second, compared to other research fields, fewer attempts at original and systematic data collection have been made in executive studies. This partly reflects properties of the research objects, namely small numbers (of chief executives, cabinets, cabinet ministers, and top bureaucrats) and limited access (because they are busy people with greater secrecy requirements). Finally, much research on the executive has also been quite idiosyncratic and more rooted in national peculiarities than, for instance, research on electoral behavior or parliaments.

Although these three points are still valid to some extent, executive research has made important achievements since the 1980s. First, the number of studies in the historic-qualitative tradition has increased. Second, the best work has become much more rigorous, building, for instance, on systematic interviews with former cabinet ministers and theory-guided case studies of executive decisions (e.g., Burch and Holiday 1996, Rhodes and Dunleavy 1995; for research on the US presidency see Edwards et al. 1993). Finally, executive studies have become more theory-guided and comparative. The fact that country expertise is probably more crucial here than in any other subdiscipline is reflected by much cross-national collaborative work. Chief executives traditionally receive most attention (e.g., Rose and Suleiman 1980, Jones 1991), but more recently systematic comparative research on Western Europe has also highlighted cabinet decision-making, the impact of individual ministers (Blondel and Müller-Rommel 1993, Laver and Shepsle 1994), the mechanisms of coalition governance (Müller and Strøm 2000), and the impact of parties on government decision-making (Blondel and Cotta 2000). However, comparative work which includes executives from both sides of the Atlantic is still rare (Weaver and Rockman 1993, Weller et al. 1997).

New frontiers of research are defined by both real world and theoretical developments. Thus, executive studies have begun to explore how European Union member states meet the challenge of multilevel governance, that is, the participation of all levels of the national executive in the decision-shaping and decision-making in EU institutions (e.g., Kassim et al. 2000). A rapidly increasing body of literature is concerned with the changes of the executive below its top political level, in particular the introduction of New Public Management and 'agencification' (for a general picture see Peters 1996). At the forefront of theory-inspired research, agency theory has been applied to executive–legislative relations (e.g., Epstein and O'Halloran 1999, Saalfeld 2000) and intra-executive relations (e.g., Andeweg 2000).

See also: Democracy; Leadership in Organizations, Sociology of; Leadership: Political; Parliamentary Government; Political Elites: Recruitment and Careers; Power: Political; Presidential Government; Presidential Personality

Bibliography

Andeweg R B 2000 Ministers as double agents? The delegation process between cabinet and ministers. *European Journal of Political Research* **37**: 377–395

Blondel J 1980 *World Leaders*. Sage, London

Blondel J, Cotta M (eds.) 2000 *The Nature of Party Government: A Comparative European Perspective*. Macmillan, Houndmills, UK

Blondel J, Müller-Rommel F (eds.) 1993 *Governing Together: The Extent and Limits of Joint Decision Making in Western European Cabinets*. Macmillan, Houndmills, UK

Blondel J, Thiébault J-L (eds.) 1991 *The Profession of Government Ministers in Western Europe*. Macmillan, London

Burch M, Holliday I 1996 *The British Cabinet System*. Prentice Hall, London

Carey J M, Shugart M S (eds.) 1998 *Executive Decree Authority*. Cambridge University Press, Cambridge, UK

Edwards G C, Kessel J H, Rockman B A (eds.) 1993 *Researching the Presidency: Vital Questions, New Approaches*. University of Pittsburgh Press, Pittsburgh, PA

Elgie R (ed.) 1999 *Semi-Presidentialism in Europe*. Oxford University Press, Oxford, UK

Epstein D, O'Halloran S 1999 *Delegating Powers: A Transaction Cost Politics Approach to Policy Making under Separate Powers*. Cambridge University Press, Cambridge, UK

Jones G W (ed.) 1991 *West European Prime Ministers*. Cass, London

Kassim H, Peters B G, Wright V (eds.) 2000 *EU Policy Coordination: The National Dimension*. Oxford University Press, Oxford, UK

King A 1975 Executives. In: Greenstein F I, Polsby N W (eds.) *Handbook of Political Science*. Addison-Wesley, New York, Vol. 5

Laver M, Shepsle K A (eds.) 1994 *Cabinet Ministers and Parliamentary Government*. Cambridge University Press, Cambridge, UK

Lijphart A 1999 *Patterns of Democracy: Government Forms and Performance in Thirty-six Countries*. Yale University Press, New Haven, CT

Linz J J, Valenzuela A (eds.) 1994 *The Failure of Presidential Democracy*. Johns Hopkins University Press, Baltimore, MD

Moe T M, Caldwell M 1994 The institutional foundations of democratic government: A comparision of presidential and parliamentary systems. *Journal of Institutional and Theoretical Economics* **150**: 171–195

Müller W C, Strøm K (eds.) 2000 *Coalition Governments in Western Europe*. Oxford University Press, Oxford, UK

Peters B G 1996 *The Future of Governing: Four Emerging Models*. University Press of Kansas, Lawrence, KA

Rhodes R A W, Dunleavy P (eds.) 1995 *Prime Minister, Cabinet and Core Executive*. Macmillan, Houndmills, UK

Rose R, Suleiman E N (eds.) 1980 *Presidents and Prime Ministers*. American Enterprise Institute, Washington, DC

Saalfeld T 2000 Members of parliament and governments in Western Europe: Agency relations and problems of oversight. *European Journal of Political Research* **37**: 353–76
Sartori G 1997 *Comparative Constitutional Engineering*, 2nd edn. Macmillan, Houndmills, UK
Shepard W J 1931 Executive. In: Seligman E R A, Johnson A (eds.) *Encyclopaedia of the Social Sciences*. Macmillan, New York
Strøm K 2000 Delegation and accountability in parliamentary democracies. *European Journal of Political Research* **37**: 261–89
Vile M J C 1967 *Constitutionalism and the Separation of Powers*. Oxford University Press, Oxford, UK
Weaver R K, Rockman B A (eds.) 1993 *Do Institutions Matter?* Brookings Institution, Washington, DC
Weller P, Bakvis H, Rhodes R A W (eds.) 1997 *The Hollow Crown: Countervailing Trends in Core Executives*. Macmillan, Houndmills, UK

W. C. Müller

Exhaustible Resources, Economics of

This article reviews research into how an economy uses natural resources, the inputs to production and human enjoyment received as part of nature's bounty. The resources examined are termed exhaustible because, in principle, their stocks are subject to exhaustion as a result of human consumption. This is clearly evident for coal, oil, and gas, and concern that humankind will run out of these resources has provided much of the impetus for the development of natural resource economics, as a field of study. Populations of fish and other species are also depletable, in the sense that their stocks can be extinguished. Through growth these resources are capable of providing continuous flows of consumption, however, and for this reason are called renewable. Resources incapable of growth are called nonrenewable.

The fields of natural resource and environmental economics are closely connected, not just by common intellectual roots, but by a simple physical identity. Natural resource economics is concerned with the study of how naturally occurring commodities, fossil fuels, metallic minerals, fish, wood, and water, are used as inputs by an economy. Environmental economics is concerned with the emission of waste or residual products; substances that cause air pollution, water pollution, and litter from the economy into the environment. Because matter can neither be created nor destroyed, the mass of natural resource inputs withdrawn from the environment and put to work in an economy must be matched in the long run by an equal mass of waste products back into the environment. This connection makes it difficult, and sometimes arbitrary, to separate natural resource and environmental economics as subdisciplines.

This discussion is organized by individual resources. Nonrenewable resources are examined first, and the discussion here focuses mainly on fossil fuels. This is followed by a general introduction to topics in renewable resources, with specific subsections on fisheries, forestry, and water resources. (The examination of water is confined to groundwater, and excludes issues of water rights, surface water management, and water pollution. These topics arguably belong to the fields of agricultural and environmental economics.) Space limitations have forced the omission ofsome important areas of research from this survey. These are noted in Sect. 3, which also points to areas where research activity is presently most active.

1. *Nonrenewable Resources*

A natural resource is nonrenewable if it does not grow naturally, and if the sum of future consumption services it can yield is therefore bounded by the existing stock. (Nonrenewable resource theory is treated in detail in Dasgupta and Heal 1979, Chaps. 6–8, 10–12, and 15 and in Fisher 1981, Chaps. 2 and 4. A recent review of nonrenewable resource scarcity is Krautkraemer 1998.) Natural gas, petroleum, and coal are nonrenewable resources because the act of consuming a unit reduces the stock, and hence future consumption opportunities, by a like amount. Metals such as copper are also nonrenewable, although consumption must be carefully defined in such cases to account for recycling. (Recyclable natural resources are still nonrenewable, because it is impossible to completely recover scrap.)

Public concern regarding nonrenewable resource scarcity is nothing new. Concern is particularly urgent when the resources in question are the energy inputs so vital to an industrial economy and exhaustion, or even a sharp price increase, would have large macroeconomic effects. The fact that consumption must eventually go to zero leads to an obvious question: when? Common measures express scarcity as the ratio of the existing resource stock to current consumption; if consumption continued at current levels, this would represent the number of years remaining until exhaustion.

Measuring scarcity in this way is misleading for two reasons. First, as explained shortly, the basic economic model of nonrenewable resource use predicts that resource prices will rise continuously as depletion proceeds. This will cause resource users to seek substitutes, reducing consumption and mitigating scarcity. Second, there are several physical measures of the existing resource stock, and their economic relevance varies from measure to measure. The largest possible measure is crustal abundance—the physical stock in some portion of the Earth's crust. This measure surely overstates ultimate consumption, as

much of the planet's endowment is technically impossible or extremely expensive to extract. A second measure is current or proven reserves—known deposits that can be profitably extracted at current prices and extraction costs. As prices rise, proven reserves will expand, so this measure surely understates ultimate consumption. 'Ultimately recoverable' reserves are proven reserves plus deposits that are currently not profitable but are expected to be in the future. Such measures depend on prices and costs, which presumably reflect scarcity. (These measures do provide a way to trace the inventory of minerals over time. The reserves-to-consumption ratio for petroleum increased over the second half of the twentieth century, indicating that discovery and cost saving innovation outpaced extraction over this period.) Economists generally prefer to assess resource scarcity from prices and costs directly. Aboveground prices, extraction costs, and mineral rights prices all reflect scarcity to some degree, as the following theoretical discussion indicates.

A nonrenewable resource deposit is an asset. To induce an owner to hold it, the deposit must generate a return that is competitive with other assets. Abstracting from differences in risk, asset market equilibrium implies that a resource stock must yield a rate of return equal to the real interest rate. Nonrenewable resource deposits do not yield a flow of consumption service, nor do they grow or depreciate physically. Thus, the only return an owner can receive from holding it rather than consuming it immediately comes from an expected increase in the asset's price. Because this capital gain is the only return a nonrenewable deposit can earn, in expectation it must be equal to the real interest rate. (As explained later, this must be amended if mineral rights ownership is monopolized to some degree.) The level of this increasing price path is determined by a requirement that prices rise to choke off demand as the resource stock approaches depletion. More precisely, demand for the resource must equal zero on the date when the resource is exhausted. (This will occur on a specific date, if there is a finite price above which the demand equals zero. It will occur only in the limit if price increases cause demand to approach zero asymptotically.) This general theory was developed by Harold Hotelling in 1931.

This result applies to prices for unextracted resources, or mineral rights, but it has direct implications for prices of extracted minerals, which are more commonly observed. If mineral extraction is competitive, the price of a unit of the extracted resource will equal the price of a mineral right plus the marginal cost of extraction. Accordingly, the Hotelling model makes the following prediction: the expected 'aboveground' price of a nonrenewable resource less marginal extraction costs will increase at the rate of interest. As exhaustion nears, the aboveground price approaches the 'choke price,' the lowest price at which quantity demanded equals zero. The exhaustible nature of such resources is socially important, as it implies that price will increase continually. The actual event of exhaustion is not a particular concern, however, because it occurs only after prices have risen high enough to reduce demand for the resource to zero.

This simple result holds only for a resource that exists in deposits of homogeneous quality. In practice, resource deposits typically vary in mineral concentrations, cost of extraction, or other attributes. Under common conditions, profit-minded mining firms will extract the most valuable deposits first, as to do otherwise would not minimize the present value of extraction costs. Suitably amended, the Hotelling rule predicts that deposits will be extracted in order of their quality or grade, with the highest grade extracted first. Within each grade, price minus marginal extraction cost will increase at the rate of interest. The lowest grade extracted is exhausted when price rises to the 'choke price.' Over time, the mineral rights price for the grade being extracted tends to rise, but mineral rights prices fall when extraction switches from higher to lower quality grades. Once the possibility of variations in quality and extraction costs is introduced, it is clear that some deposits may never be extracted, because their costs of recovery exceed their value to consumers.

Attempts to test the Hotelling model have focused on the simple uniform grade case, and the hypothesis that mineral rights prices will rise over time at approximately the rate of interest. This research, reviewed later, has found limited support at best. In hindsight, this lack of success is not terribly surprising. Hotelling's theory predicts that the entire price path for a resource will shift in response to unanticipated changes in interest rates, extraction technologies, input prices, demand conditions, knowledge about the extent of deposits, tax policy, degrees of monopoly power, and other factors. Attempts to test the theory generally have not controlled for changes in these factors. It is instructive to illustrate some of the model's predictions in this regard. An unanticipated increase in the rate of interest reduces the value of returns from future extraction relative to returns from extraction in the present. The predicted response is a reallocation of mineral extraction from future to present. This causes an abrupt drop in price and hastens the date of exhaustion. As extraction proceeds, price minus marginal extraction cost increases at the (now higher) rate of interest.

Changes in other factors will shift price paths in different ways. An increase in extraction cost should cause the aboveground price to jump, but by less than the cost change. Price starts higher and, as before, must reach the choke price at the date of exhaustion. Because price is higher on each date, consumption is reduced and the period of time over which extraction takes place is extended. Discovery of a new deposit should cause a discontinuous drop in the mineral

rights price, subsequent price increases following the Hotelling rule, and a later date of exhaustion. Among the many possibilities for unanticipated shifts in demand, one case has been singled out for detailed study—discovery of a renewable energy source that is highly substitutable for fossil fuel energy. Invention of a nuclear fusion reactor or perfection of a low cost photovoltaic device are potential examples. In this case the Hotelling model predicts an immediate reduction in the aboveground rights price followed by resumption of price increases, with exhaustion occurring when the aboveground price reaches the cost of the renewable substitute.

Shifts from competition to monopoly ownership of energy resources have been studied in detail to gain insights on how OPEC (the Organization of Petroleum Exporting Countries) has affected energy markets since the early 1970s. OPEC, acting as a cartel, seeks to control petroleum prices by coordinating the production decisions of its member nations. The simple Hotelling model predicts that, under plausible conditions, monopolization will cause an abrupt jump in price followed by relatively modest price increases until the choke price is reached upon exhaustion. The overall period of extraction is longer with monopoly than with competition. While OPEC is far from a perfect monopoly, these prescriptions are broadly consistent with energy price patterns and observed OPEC behavior since the 1970s.

An important policy issue arises when there is an absence of clear ownership claims. Crude petroleum and natural gas exist under pressure in subsurface natural deposits. Extraction from one portion of a deposit can cause part of the stock to flow from other regions toward the point of extraction. This creates a management problem when portions of the land overlying the deposit are owned by different individuals. Each landowner has an incentive to extract rapidly and thereby capture part of the stock presently underlying a neighbor's property. Because all landowners have the same incentive, the result is a rush to extract, and thus there is inefficient rapid depletion. An efficient policy response is unitization, whereby the entire reservoir is placed under unified management and each landowner receives a share of the net proceeds.

Another important policy issue arises when resource use causes environmental damage. If damage results from the extraction process, as may occur with open-pit mining and offshore oil production, the correct policy response is to charge the extracting firm an amount that reflects this damage. This gives the producer an incentive to reduce such damage. The effect on resource markets is to raise price when the charge is imposed, slow the rate of production, and render some deposits uneconomic. Environmental damage sometimes cumulates with the total amount extracted. When fossil fuels are extracted and combusted, carbon dioxide is released. Carbon dioxide does not decay rapidly, but rather accumulates in the atmosphere, where it is believed to contribute to climate change. In this case, the appropriate charge for extraction must reflect the fact that emitting a unit of the gas will cause damages that persist for a long time. Also, the charge will generally increase over time as concentrations grow.

The 'strong' prediction of a uniformly increasing price path has been tested in a variety of ways, which are outlined here. But, as mentioned earlier, a proper test would require accounting for the many variables that can shift extraction and price paths. If these factors change over time, comparing price changes to the interest rate does not provide a valid test.

Neither extraction costs nor mineral rights prices are easily observed. Some studies have instead focused just on aboveground prices, while others have attempted to estimate marginal cost statistically. Much of this research has found that prices of many resources are falling over time. A falling resource price could be consistent with the Hotelling model temporarily if deposits were being found or extraction costs being reduced, but price must eventually turn back upward as depletion approaches. This suggests the possibility of U-shaped price paths for some resources and some early evidence supported this for petroleum. Subsequent evidence, however, was contradictory. Studies using estimated extraction costs to construct series of mineral rights prices (the proper object of study) have also failed to support the simple prediction of price increases.

The Hotelling Valuation Principle provides an alternative test. The Principle states that the average asset value of a unit of a nonrenewable resource reserve should equal the current price net of extraction cost. This implies, remarkably, that the average reserve value is independent of future prices and extraction costs. The value of a firm's reserve is estimated from financial data and the implication is tested by comparing the average value of firms' reserves to resource prices net of extraction costs. Again not suprisingly, tests of the strong prediction tend to be rejected.

2. *Renewable Resources*

Renewable resources are exemplified by forests, fish populations, and water. In each case the stock available for consumption at any point in time is, or can be, regenerated by a natural process. Because such resources can grow, the cumulative future consumption service that a renewable resource stock can provide is not bounded. (The general theory of renewable resources is developed in Dasgupta and Heal 1979, Chap. 5 and in Fisher 1981, Chap. 3. Brown 1999, provides a perceptive and entertaining review of management issues related to renewable resources.)

A renewable resource stock is also an asset, so it must generate a return competitive with other assets to

induce an owner to hold it. The phenomenon of renewability causes the behavior of associated markets to differ fundamentally from what we observe for nonrenewable resources, however. Physical growth, as with biomass in a population of fish and in forests, represents a return to the owner. Some renewable resource stocks can also yield flows of consumption services for their owners, for example, forests can provide habitat for wild game and aesthetic benefits for those who enjoy looking at them or walking in them. A larger population of fish generally benefits the harvester by lowering search costs. Hence, the returns earned by renewable resource stocks are multidimensional, complicating the equal returns condition.

An important caveat is that equalization of returns across assets follows from the behavior of owners who rationally seek to maximize their returns. As we shall see, some important renewable resource stocks, particularly fish, tend to be unowned. In these cases the equalization of returns condition can serve as a guide for efficient management, but it will not describe how unowned, unregulated resources are used in practice.

2.1 Fisheries

A fish population is a living, mobile resource, capable (within limits) of regenerating itself if part of the stock is harvested for human consumption. If the stock performed no economic function other than supporting growth, an efficient manager would seek a stock size that equates the proportionate rate of growth to the real rate of interest. The long run efficient annual harvest would then equal the stock's annual growth. (In this simple model, a higher interest rate implies a smaller equilibrium stock and a larger annual harvest. The first implication tends to hold true in more complex settings. For simplicity, we are abstracting from the possibility of ongoing price changes or from price changes related to changes in the volume of harvests.)

As this simplified discussion shows, the natural population growth process must play an important role in any economic analysis of the fishery. Most economic treatments have adopted the simple logistic description of population growth and have specified a production technology in which the factors of production are of two sorts: (a) the inputs engaged in fishing effort, for example, crew, vessels, and gear, and (b) the stock of fish itself. A larger stock aids harvesting due to search costs—less effort need be expended to catch a ton of fish where the population is plentiful. This cost reduction is an additional return the stock earns, beyond the return it generates by supporting growth. Accordingly, the efficient stock in the presence of search costs is greater than would be implied by simply equating the proportionate rate of growth to the interest rate. The value of avoiding extinction provides another reason for a larger stock, since this generally lowers the probability of extinction. (Brown 1999, p. 22 notes that larger stocks of elephants may impose costs, rather than benefits, on some members of society by terrorizing farmers and destroying crops.)

The management strategy implied by equalizing rates of return will maximize the present value of net economic returns from harvesting the resource. Alternative management goals have often been pursued in practice. One such goal is maximizing sustainable yield (MSY), and another is management to protect stocks from falling below biologically critical levels. Economists have generally been critical of MSY as a management goal, since it amounts to maximizing output without considering either the value of that output or the cost of producing it. The economic arguments against MSY are illustrated by the fact that a doubling of fish prices, or a halving of fishing input costs, would have no effect on the management strategy this goal prescribes.

Generally, there is no private owner who controls how the population is used. This is due partly to the fact that fish are mobile and difficult or impossible to monitor. Hence, enforcing ownership to all or part of a population would be impossible or prohibitively costly. This explanation for the typical lack of ownership in fisheries is reinforced by the observation that the few instances of partial private ownership tend to be cases where the species are sedentary and harvesting is easily detected (for a general discussion of communal management of renewable resource, see Brown 1999, pp. 38 ff). The term 'open-access' is used to describe such unowned fisheries if they are unregulated. While the equal rates of return rule remains valuable for policy, the self-interest of asset owners cannot be relied upon to achieve efficient management where ownership is lacking.

Economic reasoning can characterize how an unowned fishery will be exploited by profit-minded harvesters. The structure of this analysis resembles a predator–prey system in ecology. Individual firms (boats) combine fishing inputs in a profit maximizing way, given the fish population available for them to exploit. Each regards its own catch as too small to have any appreciable effect on the stock it encounters. Positive profits attract new boats to the fishery, increasing fishing pressure. Added fishing pressure drives the stock down, increasing harvesting costs and reducing profits. When excess profits have been eliminated, entry stops and the industry is in long run economic equilibrium. The stock itself earns no economic return—fishing revenue just covers the cost of labor, vessels, and gear. While the population is clearly capable of generating a positive economic return in the form of growth and reduced search costs, this return is not realized in open access equilibrium.

This model has implications for harvesting activity in an open access fishery. Higher fish prices or lower prices of fishing inputs will increase fishing effort and reduce the stock; the harvest could either increase or decrease, depending on the initial stock level and the shape of the biological growth function. Populations exhibiting particular biological growth characteristics can face the possibility of extinction due to open access harvesting. (The presence or absence of critical depensation, which is characterized by a minimum viable population size below which the population cannot survive, is key consideration.) This is a rare outcome, but it becomes more likely if fish prices are high, fishing input prices are low, and search costs are unimportant.

Recognition that a lack of ownership is a cause of inefficiency points the way to more effective regulatory approaches. The Individual Transferable Quota (ITQ) is the most important regulatory innovation proposed by economists. It operates by transforming a fishery-wide restriction on catch into quantitative harvest rights that can be bought, sold, and owned by individual harvesters. One such right entitles its owner to harvest, for example, one ton of a particular species of fish in a designated area during a specified period of time. These rights are of obvious value, so command a positive price. ITQ prices represent an economic payment for use of the stock, and force harvesters to recognize that the stock is a valuable input. This induces efficiency in combining use of the stock with other fishing inputs in the harvesting process. ITQ fishery policies have been implemented in New Zealand, Iceland, Greenland, the Netherlands, and Australia.

Fisheries utilizing ITQs generally experience greater returns for harvesters, and no nation adopting the ITQ approach has subsequently abandoned it (Brown (1999) Table 1 and pp. 34 ff). In practice ITQs are often allocated to those who can demonstrate a historic involvement in the subject fishery, and this no doubt accounts for some of their political popularity. An alternative allocation would have government claiming initial ownership of ITQ rights and then auctioning them to harvesters. Economic theory, *per se*, gives no reason for preferring one initial allocation to another, and actual allocations are driven more by considerations of political feasibility and equity. (The ITQ operates by charging harvesters a price for each unit of the stock captured. One could charge the price directly, without the ITQ device, by imposing a fee or tax on fish landings. Policy economists proposed this long ago, but were ignored for political and equity reasons.)

The ITQ approach requires first setting an overall target catch. The correct economic criterion for choosing the target is maximization of the present value of net receipts, suitably adjusted to recognize that some values, for example, the value nonusers of the resource may place on avoiding extinction, are not matched by monetary payments. Traditionally, the goal of regulation was to avoid endangering stocks from a purely biological perspective. This goal was often approached by restricting fishing to only part of the year, by prohibiting the use of particularly effective fishing techniques or equipment, and by limiting the number of fishing boats allowed to operate. Economists have leveled two sorts of criticisms at such regimes. First, they often do not generate any economic return for society. While they do protect stocks, restricting fishing times or the kinds of gear used simply raises harvest costs. If costs had been raised by imposing a harvest fee, the resulting stock, catch and harvester profits could have been similar and the fee revenue would represent a net economic benefit for society. Second, harvesters will find ways around such regulations and this can magnify inefficiencies. When regulators limit the number of boats in a fishery, harvesters tend to invest in larger, faster boats with larger crews to increase fishing effort despite the regulation. When regulators limit the time available to fish, harvesters fish more intensively during the open season, raising costs, lowering fish quality, and expanding effort back toward its pre-regulation level.

2.2 Forestry

The study of renewable resources dates to a fundamental forest management problem first addressed in the eighteenth century. An early example of both intertemporal and biological components in economic models, this classic problem considers a decision facing an owner of a growing stand of trees: when to cut and harvest them for timber. The first proper economic treatment of this problem is attributed to Martin Faustmann in 1849, but in part because the solution is subtle, the optimal rotation age of a tree has since received a great deal more attention in economics. Other, more modern issues in forestry are outlined later, following a brief description of this basic model.

A tree is an asset whose return to the owner is assumed to derive only from the simple fact that it grows. (The logic of this problem is equivalent whether looking at a stand of trees of the same age or one tree.) If the price of timber remains constant, the rate of growth in value equals the rate of growth of biomass. A wealth-maximizing owner will harvest when annual growth in the tree's biomass (the benefit to leaving it standing) is equal to the sum of two distinct opportunity costs. This first cost is the rate of return that could be earned cutting and selling the resource and earning interest on the proceeds. The second cost is the land rent and represents the return that could be earned by the land the tree occupies. At the wealth-maximizing harvest age, the net rate of return earned

by the standing biomass (the value of growth minus land rent) just equals the interest rate.

The forest's biological growth function plays a key role in determining the harvest age. When a tree is young and growing quickly, the return to leaving it standing exceeds the relevant opportunity costs. As the resource matures and the growth rate falls to the level of the combined opportunity costs, the time has arrived to cut. The reasoning at the heart of this solution is that part of the cost of waiting to harvest a stand of trees is that an older tree takes up space that a younger, faster growing tree could occupy.

Investigating how this optimal harvest age responds to changes in the price of timber, replanting costs, and the market interest rate yields useful insights. When replanting is costly (in the absence of replanting costs, the optimal rotation age does not depend on the price of timber; this, perhaps surprising, result derives from the fact that all of the costs and benefits are proportional to this price) a higher price of timber will shorten the optimal harvest age and a higher replanting cost will lengthen it. These results follow from the fact that a higher price or lower replanting cost will increase the maximum value of forest land, and therefore increase the land rent. An increased rent, by raising the opportunity cost of letting the forest stand, induces the owner to harvest earlier. A higher rate of interest tends to reduce the harvest age. Intuitively, this occurs because the forest's net rate of return (growth minus land rent) tends to fall as the forest matures. It will, therefore, reach equality with the interest rate at a younger age when the interest rate is higher. To summarize, a higher interest rate causes the forest to be cut sooner.

These results have implications for the supply of timber. An increase in the price of timber increases the value of land in forestry and therefore is likely to increase acreage in forest. This increase reduces the average quantity of timber produced on each acre, however, due to the shortened harvest age. (This is true if the optimal harvest age is less than the age that maximizes sustainable yield.) As a result, the net effect of a price increase has an ambiguous effect on timber supply. The exact opposite is found for replanting costs. An increase in the interest rate unambiguously reduces timber supply as a higher interest rate reduces both land rents in forestry and the harvest age, and both effects reduce supply.

Other management goals, particularly maximizing sustainable yield, are sometimes advanced as criteria for publicly owned timber. Economists have criticized MSY in this context for essentially the same reasons they have criticized it for fisheries. While the determination of optimal harvest age is a classic problem in resource economics, other issues of forest management have become arguably more important. These include the land use decision, especially at the margin between uses in forest and agriculture, management under imperfectly defined property rights, and the external (nontimber) value of forests. These subjects are briefly discussed in the remainder of this section.

There are, of course, competing uses for a parcel of land and efficiency requires choosing the use with the greatest discounted net benefit. Importantly, the costs and benefits appropriate for this criterion include some that may be external to the owner. Such external benefits range from the services of carbon sequestration, habitat for a large fraction of the world's species, and recreational or general aesthetic value. (CO_2 captured by a growing forest is 'sequestered' from the atmosphere, where it would contribute to a potential greenhouse effect. The value of biodiversity is in both potential medical services and less tangible human preferences.) In general, if a standing forest provides benefits not captured by the landowner, and if those benefits increase with the age of the stand, then the private owner's harvest age will be too short and there will be an inefficient bias toward nonforest land uses. This observation suggests a role for government policy related to land use and harvest age on both private and public lands.

Ownership of forest land is often incomplete. In many areas of the world the ubiquitous property rights regime is 'common property' or 'communal management,' where a village or otherwise organized group of people share ownership of land used alternatively for agriculture, fuel wood, forest products, and other services. Without collective action to regulate individual behavior, each member of the community has an incentive to clear more forest land for agriculture, and to allow livestock to overgraze common pasture, etc. An individual farmer considers the private cost of converting common forest to agriculture, for example, the reduced availability of fuel wood for him personally, but ignores the cost borne by every other member of the community. This incentive is magnified by the realization that failure to appropriate a piece of the commons today only decreases the probability that it will be available for appropriation in the future.

The difference between common property resource management and open access (a concept developed in the fisheries section) relies on the effectiveness of informal controls. Effective common property management requires a set of rules or norms, and a binding enforcement mechanism that promotes socially efficient resource use. While the requisite conditions are not well understood, informal solutions to commons problems should be more common when bargaining, enforcement, and information about resource use are inexpensive relative to the gains from efficient management. This seems more likely when the resource is small and scarce, and perhaps where the number of actors is small. Another seemingly important ingredient is a long-established community that shares traditional norms of behavior and a common technology. Empirical work suggests that actual common

property arrangements internalize some, but not all, of the relevant social costs.

2.3 Groundwater

Groundwater exists in subsurface reservoirs, or aquifers, that usually are replenished by percolation of water from the surface. As with oil and gas, groundwater is mobile and does not respect the boundaries of individual parcels of land that overly it. Absent of regulation, one can therefore expect groundwater stocks to be overused.

Water in an aquifer is an asset and the problem of efficient management can again be guided by the principle that rates of return be equalized. As water in storage is depleted, the distance it must be pumped to the surface increases, raising costs. Water in storage thus earns a return by reducing pumping costs. It can also earn a return associated with the fact that withdrawing groundwater can cause the overlying land surface to subside, damaging structures or otherwise reducing the land's value. Similarly, withdrawing groundwater near an ocean shoreline can cause saltwater to intrude and degrade the quality of water in an aquifer. In both cases, leaving more water in storage reduces the damage, or makes it less likely, adding new components to the return that water in storage yields.

As with fish and unowned petroleum reservoirs, groundwater will not be efficiently managed if left to the choices made by individuals withdrawing from a common aquifer. Policies similar to ITQs in fisheries, or unitization with oil and gas are viable management instruments in principle. They have not generally been adopted, however, owing in part to the unique legal institutions that apply to property rights for water.

3. Conclusions and New Directions

Several important topics were not discussed due to a constraint on space, and some trends in the study of exhaustible resources are worthy of mention. These are briefly broached these here, and the interested reader is directed to more detailed discussions in Brown (1999), Krautkraemer (1998), and Dasgupta and Heal (1979).

The study of nonrenewables has largely shifted from optimal extraction and scarcity measurement to the environmental impacts of fossil fuel consumption. The traditional concern was that energy would become expensive as coal, petroleum, and natural gas reserves were depleted. Prices of these inputs have not risen consistently, but today it is recognized that energy consumption is expensive in a sense historically not properly accounted for. A ton of coal in the ground provides few external costs or benefits, but a ton of coal burned spreads damages widely in the form of acid rain and greenhouse gases. This link between nonrenewable resources and environmental policy is central to the field today.

There are several directions in which the theory of fisheries and fishery regulation is being extended at present. First, most economic models of fisheries do not incorporate state-of-the-art ecological models of the resource. Fisheries ecology has moved far beyond the simple logistic models of population dynamics that are still a mainstay of fisheries economics. Second, economic analyses of fishery regulation have, with a few recent exceptions, made no attempt to study the motives and the behavior of the regulator. Third, economic models of fisheries tend to ignore spatial considerations in the distribution of fish populations or fishing activity. As this very active area of natural resource economics develops, it seems likely that each of these topics will receive attention.

While the optimal harvest age is the textbook forestry topic, the total stock of forests is more sensitive to the factors that determine land use. This is particularly true as it becomes clear that the conversion of land to agricultural uses drives much of global deforestation. Recent study has focused on how ownership security, agricultural and trade policy, and the functioning of related input markets affect land use decisions. Empirical attempts to measure the external value forests provide, through the services of carbon sequestration and biodiversity habitat, are met with great interest.

This article has ignored a large and important literature incorporating exhaustible resources into the theory of economic growth. This literature has traditionally treated exhaustible resources as capital necessary for production but substitutable with other inputs. Whether exhaustible resources represent 'limits to growth' is certainly an issue that commands attention far outside the field of economics.

Nearly all of the basic models described here ignore the fact that the world is an uncertain place. An owner of a resource is assumed to know what will happen to future prices and costs. How poor an approximation of behavior is it to replace random variables with their expected values? The degree to which people are averse to, and able to insure against, risk, are key aspects of investment behavior that shed light on resource management problems. Optimal resource use is also likely affected if relevant events are irreversible, a point distinct from risk aversion. Irreversibility is key to the study of species extinction and climate change in the modern literature.

See also: Agricultural Sciences and Technology; Ecotourism; Environmental and Resource Management; Environmental Economics; Environmental Planning; Environmentalism: Preservation and Conservation; Sex Differences in Pay; Sustainability Transition: Human–Environment Relationship; Sustainable Development; Sustainable Transportation; Urban Policy: North America

Bibliography

Brown G M 1999 *Renewable Natural Resource Management and use Without Markets*. Working paper, Department of Economics, University of Washington, Seattle, WA

Dasgupta P S, Heal G M 1979 *Economic Theory and Exhaustible Resources*. Cambridge University Press, Cambridge, UK

Faustmann M 1968 [1849] On the Determination of the Value which Forest Land and Immature Stands Possess for Forestry, [Gane M, English trans.], Oxford Institute Paper 42

Fisher A C 1981 *Resource and Environmental Economics*. Cambridge University Press, Cambridge, UK

Gordon H S 1954 The economic theory of a common property resource: The fishery. *Journal of Political Economy* **62**: 124–42

Hotelling H 1931 The economics of exhaustible resources. *Journal of Political Economy* **39**: 137–75

Krautkraemer J A 1998 Nonrenewable resource scarcity. *Journal of Economic Literature* **36**: 2065–67

Schaefer M B 1954 Some aspects of the dynamics of populations important to the management of commercial marine fisheries. *Bulletin of the Inter-American Tropical Tuna Commission* **1**: 25–56

E. Balsdon

Existential Social Theory

Modernity, as the continuing project of human self-understanding and human self-making, is both the historical foundation of social theory and an existential perspective on the world. All modern social theory, thus, insofar as it reflects upon the experiential character of social life, is to some degree existentialist.

1. *Max Weber's Fundamental Contribution to Existential Social Theory*

The centrality of an existential perspective to social theory is nowhere clearer than in the work of Max Weber (1864–1920), who must be accounted its most formidable advocate. The existential character of Weber's perspective is most evident in his essays on politics and science and, more especially, in his methodological writings. These stress both the centrality of meaning to a proper understanding of social action, and the essential freedom of humans in relation to the values that ultimately confer meaning on events. In a radical fashion Weber insists that meaning flows from an existential 'orientation to the world' and does not inhere in any objectively given reality. Weber's sociology, thus, is concerned primarily with the typification, on a world historical scale, of an extraordinarily varied range of such meaning complexes, their historical contextualization, and the existential orientations within which they were formed and enacted.

Weber's social theory also established critical insight into the existential problems of modern society. On the one hand, he argues that modernity can be understood in terms of a specific set of values which he typifies as the 'spirit of capitalism' or 'modern rational capitalism'; while, on the other hand, he argues that modernity is characterized above all by the 'disenchantment of the world,' which involves nothing less than the annihilation of value as such. Not only are these two very different conceptions, but the notion of disenchantment seems to contradict his philosophical-methodological premise that all human activity be inherently open to meaning through its relation to extra-rational ends. This duality in Weber's characterization of modern life in fact brings to light the seriousness of the struggle in modern society to create and maintain a sense of individual human worth and dignity in the face of powerful processes of rationalization. He traces this theme substantively to the Reformation and the 'unprecedented inner loneliness' which was its religious expressions and the historical foundation of the anxiety which has become a general condition of modern existence. In this approach it is clear that Weber was influenced by Friedrich Nietzsche (1844–1900) and Søren Kierkegaard (1813–55). Indeed, a remark of Nietzsche's might well stand as an epigraph for his existential sociology;

> Man first implanted value into things to maintain himself—he created the meaning of things, a human meaning! Therefore he calls himself: 'Man,' that is; the evaluator … Evaluation is creation.

2. *Kierkegaard and the Origins of European Existential Thought*

But it is to Kierkegaard that we must turn as the originating source of the varied forms of modern existential thought. Not only in his work do we find the first modern account of human reality focused directly on the elusive and difficult character of 'existence,' we also discover, linked to it, a significant social theory of modernity.

Kierkegaard's writings provide a rich exploration of all the forms of existence which have become possible in the modern world. Rather than analyze these positions in terms of an abstract analytic framework—itself presumed to be outside of any of their existential determinants—he presents them as the self-expressed orientations of a number of pseudonymous authors. This semi-fictional method draws author, subject matter, and reader into an existential complicity and Kierkegaard uses this technique to mount a powerful attack on all modern (especially Hegelian) forms of scientific abstraction and objectivity which have been misleadingly applied to problems of existence.

This method, inaugurated in the celebrated *Either/Or* (1987 [1843]), is neither a form of pure 'subjectivism' nor does it result in chaos of arbitrary life-

views. In fact the authorship reveals a specific, albeit complex, structure which conditions all subjective contents of experience, whatever position is adopted by the protagonist. This structure—a kind of existential map of modern society—is characterized by Kierkegaard in terms of 'existence-spheres' within which a number of positions are inter-related. Thus, the aesthetic mode of existing is organized around the polarities of pleasure and unpleasure, and is characterized by successive phases of enthusiasm and boredom; the ethical sphere is ordered in terms of the distinction of good and evil, and emerges in a more concrete selfhood of resolution and guilt; and the religious sphere crystallizes and takes form in the opposition of faith and sin. The fluid interchange and dialogic structure of the authorship, in opposition to any systematic theory of experience leaves this existential experimenting incomplete and unresolved.

Kierkegaard, under his own name, but in the form of a lengthy book review rather than a direct essay, also provided a suggestive existential theory of the social and cultural tendencies of his own time (*Two Ages* 1978 [1845]). What he called the Present Age (as distinct from a previous Revolutionary Age) is characterized by superficiality, bourgeois philisitinism, empty conventionalism in public life, and a general process of leveling in which all social differences are obliterated in favor of a new 'mass' culture. The Present Age, like Hegelian philosophy, has become abstract and empty, obsessed with public opinion and meaningless 'chatter'; it is the mere play of forms which lack the seriousness of any of the positions explored in the pseudonymous authorship. The Present Age is not at all 'interesting,' in the significant sense of engaging the self and enriching its interrelational structure with authentic life-contents; everything happens as if one were a spectator of a remote and detached world. Yet it is just the character of existence that it is and remains interesting, that the pervasive boredom of the Present Age, at the same time discloses the possibility of a new reality and prompts the continual renewal of groundless but not hopeless existential experimenting. The pseudonyms, thus, choose boredom over conventionality, propelling themselves into new and unpredictable projects of existence.

The existential value of 'authenticity,' it should be noted, is not associated in Kierkegaard's seminal writings with the romanticism of 'self-positing' and the self-realization of a specific inner 'personality,' but with the decisive character of 'choosing.' This is not conceived in a utilitarian market sense as deciding among available alternatives but, rather, is constituted as a transformative 'leap' into new forms of existence. To 'choose' is simultaneously to be chosen, to surrender to a value outside the self, only thus can it genuinely become part of the self; it is like falling in love, a possibility which the rationalizing conditions of the Present Age make daily more difficult.

Kierkegaard's extraordinary writings were for a long time all but ignored and it is only relatively recently that his work has been read again as a creative and significant contribution to modern social theory. His influence was felt more directly in the philosophical tradition. First, through Karl Jaspers (1883–1969), himself a friend of Max Weber's and also an ardent reader of Nietzsche, who, in addition to his major philosophical works, in *Man in the Modern Age* popularized Kierkegaard's social criticism of the *Two Ages* and in an original fashion opened up a vast field of psychopathology to existential investigation (Jaspers 1963). Second, through Martin Heidegger (1889–1976) whose *Being and Time* (1962 [1927]), which is hardly conceivable without Kierkegaard's *Concept of Anxiety* (1980 [1844]), made existence once again a central preoccupation for contemporary philosophy.

Jaspers represents a turn away from nihilistic interpretations of existentialism back to an appreciation of the positive insights of Kierkegaard and Nietzsche. His view of existence is explicitly social in character; 'Since man only exists in and through society … we must study his nature by studying society.' And, although in Jaspers view, 'owing to the turmoil of modern life, what is really happening eludes our comprehension,' he proposes a conception of modern society as a dynamic structure rather than an undifferentiated crowd; 'there are only diverse masses which form, dissolve, and reform … articulated masses are mutable, diversified, transitory expressions of some specific historical outcome of human existence.' But, as 'even an articulated mass always tends to become unspiritual and inhuman,' modern existence is generally inauthentic.

Heidegger marks a turn from nihilism towards primitivism; an approach to human existence as primordial Being. For Heidegger authentic Being is always preconceptual and unthematized, it is founded upon a rejection of the world of modernity rather than an encounter with, or within, it. In the extraordinary elaboration of this project he introduces a new language of Being in which all subsequent existential reflection took place. In this language he emphasizes and makes his own Kierkegaard's rejection of every misdirected form of objectifying thought: 'Philosophy cannot base itself on the illusion of making man a disinterested spectator of himself.'

The openness of existential reflection to new realities and new interpretations of reality is evident both from its originating variety and in the ideological diversity of its followers. Whatever the political, religious, or ethical tendency favored by its major proponents, however, each has been distinguished by fearless radicalism. In its still best known version, indeed, existentialism and radicalism become synonymous. What had begun as a pitiless anatomy of despair was transformed by Jean-Paul Sartre (1905–80) into a new, optimistic but equally uncompromising vision of

liberation. Reading Heidegger in relation to Hegel he produced in *Being and Nothingness* (1956 [1943]) an impressive restatement of the entire project of modernity as the discovery of human autonomy. Here the commitment to freedom as the fundamental prerequisite for any valued form of human being is restated with unsurpassed determination. The result is less a secular version of a Christian notion of existence (Kierkegaard) than the consecration of a particular kind of Romanticism (Fichte). This is characteristic also of Albert Camus (1913–60), whose *The Rebel* (1954) returned to nineteenth-century Russian literature for an understanding of nihilism as the most consistent and fearless of existential views of the human world. Only the free act could be considered authentically human, and as conformity to any existing convention was tainted with coercion, acts of rebellion bore the only indubitable sign of freedom and, therefore, of authenticity.

Sartre increasingly sought for his existentialism a larger and more dramatic stage; history and the collective subject, rather than the small individual acts of everyday life, supplied the setting for the working out of the great refusal; the negation of everything intolerable in modern life. The attempted synthesis of existentialist and Marxist traditions resulted, in *Critique of Dialectical Reason* (1976 [1960]), in a social theory paradoxically 'founded' on 'anti-foundational' existential insights. Through this Sartre sought to establish praxis relevant to contemporary life rather than an explanation of its various features.

His views, thus, remain rooted in existential insight; 'The only concrete basis for the historical dialectic is the dialectical structure of individual acts.' The radical freedom central to *Being and Nothingness* is somewhat muted here—reciprocity and otherness, mediated through a series of 'objective' forms (dyad, triad, serial group, fused group, statutory group, organization, class), are conceived as pre-given and as essential aspects of, and constraints upon, every human project. However the subversive genius of its empty form is also invoked: 'For freedom is nothing other than a choice which creates for itself its own possibilities.' The human remains the being for whom 'The upsurge of freedom is immediate and concrete.'

Existential social theory seeks to realize two distinct and seemingly incompatible values; the absolute and untrammeled freedom of the subject; and the actualization of authentic selfhood. Much of the development of existentialism is an attempt to reconcile these two values and to imagine a social world in which they could be harmonically interrelated. The urgency and difficulty of such a task is the source of the pathos, which distinguishes all serious existential social theory.

See also: Marxist Social Thought, History of; Modernity; Modernity: History of the Concept; Modernization and Modernity in History; Nietzsche, Friedrich (1844–1900); Sociology, History of; Weber, Max (1864–1920); Weberian Social Thought, History Of

Bibliography

Camus A 1954 *The Rebel* [English translation by Bower A]. Knopf, New York

Heidegger M 1962 [1927] *Being and Time* [English translation by Macquarrie J, Robinson E]. Harper, New York

Jaspers K 1933 *Man in the Modern Age* [English translation by Paul E, Paul C]. Routledge, London

Jaspers K 1963 *General Psychopathology* [English translation by Hoening J, Hamilton M]. University of Chicago Press, Chicago, IL

Kierkegaard S 1978 [1845] *Two Ages* [English translation by Hong H V, Hong E H (eds.)]. Princeton University Press, Princeton, NJ

Kierkegaard S 1980 [1844] *The Concept of Anxiety* [English translation by Hong H V, Hong E H (eds.)]. Princeton University Press, Princeton, NJ

Kierkegaard S 1987 [1843] *Either/Or* [English translation by Hong H V, Hong E H (eds.)]. Princeton University Press, Princeton, NJ

Langiulli N 1997 *European Existentialism*. Transaction, New Brunswick, Canada and London

Nietzsche F 1986 *Human, All Too Human* [English translation by Hollingdale R J]. Cambridge University Press, Cambridge, UK

Sartre J-P 1956 [1943] *Being and Nothingness* [English translation by Barnes H]. Philosophical Library, New York

Sartre J-P 1976 [1960] *Critique of Dialectical Reason* [English translation by Sheridan-Smith A]. New Left Books, London

Weber M 1948 *From Max Weber: Essays in Sociology* [English translation by Gerth H H, Wright Mills C (eds.)]. Routledge and Kegan Paul, London

H. Ferguson

Expectations, Economics of

'Expectations' in economics refers to the forecasts or views that decision makers hold about future prices, sales, incomes, taxes, or other key variables. The importance of expectations is due to their often substantial impact on the current choices of firms and households, and hence on current prices and the overall level of economic activity.

1. Expectations in the History of Thought

Modern economic theory recognizes that the central difference between economics and natural sciences lies in the forward-looking decisions made by economic agents. Therefore expectations are a basic building

block of economic theories. For example, in consumption theory the paradigm life cycle and permanent income approaches stress the role of expected future incomes. In investment decisions present value calculations are conditional on expected future prices and sales. Equity prices, interest rates, and exchange rates all clearly depend on expected future prices.

A central aspect of economic theories is that expectations influence the time path of the economy, and conversely one might reasonably hypothesize that the time path influences expectations. The current standard methodology for modeling expectations is to assume *rational expectations* (RE), which is in fact an equilibrium in this two-sided relationship.

RE modeling is a recent key step in a long line of dynamic theories which have emphasized the role of expectations. The earliest references to economic expectations or forecasts date to the ancient Greek philosophers and the Bible. Systematic economic analyses in which expectations play a major role began as early as Henry Thornton's treatment of paper credit, published in 1802, and Émile Cheysson's 1887 formulation of a framework which had features of the 'cobweb' cycle. The role of expectations was given some attention by the classical economists, but their method of analysis was based on the stationary state in which perfect foresight prevails. Expectations were equated with actual outcomes, which downplayed their significance.

Alfred Marshall is credited with the notion of 'static expectations' of prices. The *'cobweb' model* of a market with a production lag was one of the first formal models with expectations in the 1930s. In the same decade, the *temporary equilibrium* approach, initiated by the Stockholm school, explicitly introduced expectations of future prices influencing current demands and supplies. John Muth (1961) was the first to formulate the notion of rational expectations and did so in the context of the cobweb model.

In macroeconomic contexts the importance of the state of long-term expectations of prospective yields for investment and asset prices was emphasized by John Maynard Keynes (1936) in his General Theory. Keynes stressed the central role of expectations for the determination of output and employment, but did not have an explicit model of how expectations are formed. He even sometimes suggested that attempting to forecast very distant future events can virtually overwhelm rational calculation. In the 1950s and 1960s expectations were introduced into almost every area of macroeconomics, including consumption, investment, money demand and inflation using *adaptive expectations* or related schemes.

Rational expectations made the decisive appearance in macroeconomics in the work of Robert E. Lucas Jr. and Thomas J. Sargent in the beginning of 1970s. Many of the key contributions are collected in Lucas and Sargent (1981) and Lucas (1981). The RE hypothesis became widely used in the 1970s and 1980s and it is currently the benchmark paradigm in both micro- and macroeconomics. Some recent research has gone beyond RE by developing models of *learning behavior* with explicit theories of data collection and forecasting.

In this article we review developments in the modeling of expectations formation, with an emphasis on rational expectations and learning behavior. RE modeling is the subject of many books, e.g., Sargent (1987) and Farmer (1999). Evans and Honkapohja (2001) is a treatise on the learning approach.

2. *Traditional Models*

We will illustrate the modeling of expectations with some well-known simple models. The first example is the cobweb model. Consider a single competitive market in which there is a time lag in production. Demand is assumed to depend negatively on the prevailing market price:

$$d_t = m_I - m_p p_t + v_{1t} \tag{1}$$

while supply depends positively on the expected price:

$$s_t = r_I + r_p p_t^e + v_{2t} \tag{2}$$

where m_p, $r_p > 0$ and m_I and r_I denote the intercepts. We have introduced shocks to both demand and supply. v_{1t}, v_{2t} are exogenous *iid* (identically and independently distributed) random variables with mean zero and constant variance.

The interpretation of the supply function is that there is a one-period production lag, so that supply decisions for period t must be based on information available at time $t-1$. For simplicity we make the representative agent assumption that all agents have the same expectation. Extensions to heterogeneous expectations have been analyzed in the literature.

We assume that markets clear. The observed price is then obtained by equating s_t and d_t, which leads to

$$p_t = \mu + \alpha p_t^e + \eta_t \tag{3}$$

where $\mu = (m_I - r_I)/m_p$ and $\alpha = -r_p/m_p < 0$. $\eta_t = (v_{1t} - v_{2t})/m_p$ so that we can write $\eta_t \sim iid(0, \sigma_\eta^2)$, i.e., η_t is an *iid* random variable with mean zero and variance σ_η^2.

Equation (3) is an example of a temporary equilibrium relationship. It illustrates the central role of expectations by showing how the current market-clearing price depends on expected prices. Developments since the Stockholm School and Keynes can be seen as different theories of expectations formation, i.e., how to close the model so that it constitutes a fully specified dynamic theory. We now describe briefly some of the most widely used schemes with the aid of this example.

2.1 Static Expectations

Naive or *static expectations* were used widely in the early literature. In the context of the cobweb model they take the form

$$p_t^e = p_{t-1} \tag{4}$$

Once this is substituted into Eqn. (3) one obtains $p_t = \mu + \alpha p_{t-1} + \eta_t$, which is a stochastic process known as an autoregressive process of first order (AR(1)). This and related stochastic processes are studied in standard textbooks on time series analysis and econometrics.

In the early literature there were no random shocks, yielding a simple difference equation $p_t = \mu + \alpha p_{t-1}$. This immediately led to the question of whether the generated sequence of prices converged to the stationary state over time. The convergence condition is, of course, $|\alpha| < 1$. Whether this is satisfied depends on the relative slopes of the demand and supply curves. In the stochastic case this condition determines whether the price converges to a stationary stochastic process. (Loosely speaking, a stationary stochastic process is a non-explosive process with statistical properties that are unchanging over time.)

2.2 Adaptive Expectations

The origins of the *adaptive expectations* hypothesis can be traced back to Irving Fisher. It was formally introduced in the 1950s by Phillip Cagan, Milton Friedman, and Marc Nerlove. In terms of the price level the hypothesis takes the form

$$p_t^e = p_{t-1}^e + \lambda(p_{t-1} - p_{t-1}^e) \tag{5}$$

For the cobweb model it can be shown that both expectations and prices converge to stationary stochastic processes, provided the stability condition $|1-\lambda(1-\alpha)| < 1$ is met.

Adaptive expectations can equivalently be written as a distributed lag with weights declining exponentially at rate $1-\lambda$. Besides adaptive expectations other distributed lag formulations were used in the literature to allow for extrapolative or regressive elements.

Adaptive expectations played a prominent role in macroeconomics in the 1960s and 1970s. For example, inflation expectations were often modeled adaptively in the analysis of the expectations augmented Phillips curve.

3. Rational Expectations

The RE revolution begins with the observations that adaptive expectations, or any other fixed weight distributed lag equation, may provide poor forecasts in certain contexts and that better forecast rules may be readily available. The optimal forecast method depends on the stochastic process of the variable being forecast and this implies interdependency between the forecasting method and the economic model. On this approach we write

$$p_t^e = E_{t-1} p_t \tag{6}$$

for the cobweb example. $E_{t-1}p_t$ denotes the mathematical expectation of p_t conditional on variables observable at time $t-1$ (including past data).

RE is an equilibrium concept. The actual stochastic process followed by prices depends on the forecast rules used by agents, so that the optimal choice of the forecast rule by any agent is conditional on the choices of others. An RE equilibrium imposes the consistency condition that each agent's choice is a best response to the choices by others. In the simplest models we have representative agents and these choices are identical.

For the cobweb model we have $p_t = \mu + \alpha E_{t-1} p_t + \eta_t$. Taking conditional expectations E_{t-1} of both sides yields $E_{t-1}p_t = \mu + \alpha E_{t-1} p_t$ so that expectations are given by $E_{t-1}p_t = (1-\alpha)^{-1}\mu$ and we have $p_t = (1-\alpha)^{-1}\mu + \eta_t$. This is the unique way to form expectations which are 'rational' in the model (3).

Two related observations should be made. First, under RE the appropriate way to form expectations depends on the stochastic process followed by the exogenous variables, η_t in the example. If these are not *iid* processes then the RE will themselves be random variables, and they often form a complicated stochastic process. Second, neither static nor adaptive expectations are in general rational, except in special cases.

Next, we consider some key aspects of RE models.

3.1 Forward-looking Character

In the cobweb model expectations pertain only to current prices. Most economic models have forward-looking elements, so that expectations about future periods appear. A simple example is the *Cagan model of inflation* which postulates that the demand for money depends linearly on expected inflation:

$$m_t - p_t = -\psi(p_{t+1}^e - p_t) + \zeta_t, \quad \psi > 0 \tag{7}$$

where m_t is the log of the money supply at time t, ζ_t is an *iid* money demand shock with zero mean, p_t is the log of the price level at time t, and p_{t+1}^e denotes the expectation of p_{t+1} formed in time t. Solving for p_t we obtain the equation

$$p_t = \alpha p_{t+1}^e + \beta m_t + v_t \tag{8}$$

where $\alpha = \psi(1+\psi)^{-1}$, $\beta = (1+\psi)^{-1}$, and $v_t = -(1+\psi)^{-1}\zeta_t$. Note that $0 < \alpha < 1$ and $\beta = 1-\alpha$.

The same formal model arises in other contexts, e.g., the standard model of asset pricing with risk neutrality

takes the form of Eqn. (8) with different values for α and β. Under RE the model is

$$p_t = \alpha E_t p_{t+1} + \beta m_t + v_t \tag{9}$$

Assume that m_t follows some general exogenous stochastic process. One can show that the expression

$$p_t = v_t + \beta \sum_{i=0}^{\infty} \alpha^i E_t m_{t+i} \tag{10}$$

solves the model, provided that the sum converges. Given an exogenous stochastic process for m_t, techniques are available for computing the forecasts $E_t m_{t+i}$ and hence the RE solution (10).

An important feature of the solution (10) is its forward-looking character. The current price level is a sum of expected values of future money stocks. If, for example, there is a change in monetary policy which is anticipated to lower the money stock from some specified future date $T > t$, this will be reflected in the price level already at time t. Previously unexpected changes in policy will affect p_t from the moment they become known.

3.2 Solutions to Linear RE Models

Equation (9) is a simple example of a linear RE model. More general models allow for a dependence on lagged values of the endogenous variable p_t, expectations formed at different times and over various horizons, and more general exogenous processes. Generalizations to multivariate frameworks are particularly important in practice.

A number of techniques are available for obtaining solutions under RE. Undetermined coefficient methods are often particularly convenient. Frequently, as in the model (9) with $|\alpha| < 1$, there is a unique nonexplosive solution, given nonexplosive exogenous variables. In this case a method called the Blanchard–Kahn technique can be used to compute this solution.

4. Econometric Issues with Rational Expectations

4.1 Testing for Rationality

If the RE hypothesis is correct, then expectations obey very strong assumptions. The law of mathematical conditional expectations, from probability theory, states that $E(E(p \mid I) \mid S) = E(p \mid S)$ if S and I are alternative information sets with $S \subset I$. Here $E(p \mid \Omega)$ denotes the mathematical expectation of p conditional on an information set Ω. This result immediately implies $E(e \mid S) = 0$, where $e = p - E(p \mid I)$.

This result, called the *orthogonality principle*, implies that the RE forecast error e must be uncorrelated with every variable in the information set.

If data on expectations, for example from surveys, are available, then this provides a straightforward way to test for rationality. Suppose we have time series data on, say, the average one-period ahead forecast f_t of p_t, where the forecast f_t is made based on I_{t-1}, the information available at time $t-1$. Under RE $f_t = E(p_t \mid I_{t-1})$ and we therefore test whether $E(e_t \mid I_{t-1}) = 0$, where $e_t = p_t - f_t$. In practice, the test is implemented by a regression:

$$e_t = \phi' z_{t-1} + u_t \tag{11}$$

where z_{t-1} is a vector of variables in I_{t-1}. Under the null hypothesis of RE the restrictions $H_0: \phi = 0$ hold and u_t is serially uncorrelated. H_0 can be tested using the standard F-test under some additional assumptions. The intuition for the tests is straightforward. If $\phi \neq 0$ then forecast errors are systematically correlated with z_{t-1} and the information in z_{t-1} is not being fully exploited.

Tests can also be conducted using forecast data with multiple period horizons and/or panel data, although these lead to econometric complications. In practice, tests of rationality using survey data on forecasts typically lead to rejections of rationality, although the interpretation of these results is controversial. Supporters of RE often argue that the individuals surveyed are not the relevant group of agents or that the expectations are measured with error. However, rejections of rationality also raise the possibility that agents do deviate from full RE owing to learning, forecasting costs, and/or strategic uncertainty. See Pesaran (1987) for further discussion of econometric tests of the rational expectations hypothesis.

4.2 The Lucas Critique

During the 1950s and 1960s, applied macroeconometric models explicitly or implicitly assumed that expectations were given by adaptive expectations or by some related lag scheme. The estimated systems were used for policy analysis by evaluating the effects of alternative policies. This implicitly assumed that the estimated coefficients reflect a structure that is invariant to alternative policies. Lucas (1976) criticized this approach, arguing that if expectations are rational, then the coefficients relating target variables to policy instruments will change when the policy process changes.

As an example, consider the Cagan model (9) with $\beta = 1 - \alpha$ and assume that money supply follows the rule

$$m_t = \mu + \rho m_{t-1} + w_t \tag{12}$$

where $|\rho| < 1$ and w_t is an (unforecastable) mean zero *iid* shock. In this case the solution (10) reduces to

$$p_t = \delta + \gamma m_t + v_t \tag{13}$$

where

$$\delta = (1-\alpha\rho)^{-1}\alpha\mu \quad \text{and} \tag{14a}$$

$$\gamma = (1-\rho\alpha)^{-1}(1-\alpha) \tag{14b}$$

The 'traditional' approach to policy would estimate δ and γ from historical data and use the estimated version of Eqn. (13) to evaluate the effects of alternative policy sequences m_t on the price level. However, it is clear from Eqn. (14) that γ is not invariant to the policy rule since, for example, a change in the policy parameter ρ affects the value of γ. Thus, under RE, the response of target variables to policy variables depends on the form of the policy rule.

The effect of the 'Lucas critique' has been to shift attention to estimation of deep parameters (such as α in the above example), which do remain invariant to policy changes and to use these estimates to evaluate alternative policy rules under RE.

4.3 Estimation and Testing of RE Models

When data on expectations are not available, it is still often possible in the context of structural models to both estimate the models under the assumption of RE and to test the RE assumption. In this case the RE assumption is being tested jointly with the model. A simple example is the Cagan model (9), with $\beta = 1-\alpha$, together with the money supply rule (12). Equations (12) and (13) can both be estimated, but Eqns (14a,b) imply that there are nonlinear cross-equation parameter restrictions that are imposed by the RE hypothesis. The system can be estimated under these overidentifying restrictions. The restrictions can also be statistically tested, in effect testing RE jointly with the model.

Another strand that has arisen recently is the estimation and testing of Euler equations. These are dynamic equations, relating, e.g., current consumption to expected future consumption and interest rates, which arise as optimality conditions from dynamic utility maximization. Under RE testable overidentifying restrictions again arise.

5. Multiple Equilibria and RE

Many RE models can have *multiple equilibria*. The existence of rational speculative bubble solutions besides the fundamental equilibrium illustrates this phenomenon. The standard model of asset pricing is formally identical with the Cagan model (8) if we set $v_t \equiv 0$, $\beta = 1$, $\alpha = (1+r)^{-1}$, where r is the real net interest rate. Here m_t represents the dividend at t and p_t is the price of the asset. Then Eqn. (10) is called the 'fundamental solution,' i.e., the present value of anticipated future dividends. A 'rational bubble solution' is the sum of the fundamental solution and a bubble, which is any process B_t satisfying $E_t B_{t+1} = \alpha^{-1}B_t$. The *bubble solutions* are 'explosive' since $\alpha < 1$, and macroeconomists are often content to assume away 'explosive bubbles,' although rational bubbles that burst periodically can be constructed. In models with carefully elaborated microfoundations, explosive solutions can often be ruled out theoretically. Nonetheless, the empirical issue of whether rational (or almost rational) bubbles affect asset prices is still debated.

Other models have multiple RE solutions, none of which are explosive. As a simple example, consider the model $p_t = \alpha E_t p_{t+1} + \beta m_t + v_t$ with $|\alpha| > 1$. (This case can arise from linearized versions of the overlapping generations model of money under some specifications of preferences.) If m_t is *iid* with mean $\bar{m}$, from Eqn. (10) there is the RE solution $p_t = (1-\alpha)^{-1}\alpha\beta\bar{m} + \beta m_t + v_t$. However, there are also solutions of the form

$$p_t = \alpha^{-1}p_{t-1} - \beta\alpha^{-1}m_{t-1} - \alpha^{-1}v_{t-1} + \varepsilon_t \tag{15}$$

where ε_t is an arbitrary process, observable at time t and satisfying $E_{t-1}(\varepsilon_t) = 0$.

These solutions are well-behaved nonexplosive processes. The variable ε_t is often called a 'sunspot' variable and the corresponding equilibrium a '*sunspot solution*.' The term sunspot is used to indicate the arbitrary nature of the variable. The solution depends on ε_t only because everyone believes that it does. It does not appear in the basic model structure. Guesnerie and Woodford (1992) provide a review of the results on sunspot equilibria and *endogenous fluctuations*.

It has been demonstrated formally that sunspot solutions can exist in overlapping generations models in which the microfoundations are carefully constructed. More recently it has been shown that, with increasing returns to scale and monopolistic competition, sunspot solutions can exist in variations of common business cycle models. These models raise the possibility that business cycle fluctuations may be due to random but rational fluctuations in household and firm expectations as a '*self-fulfilling prophecy*.' The empirical significance of self-fulfilling prophecies remains an open issue.

There are numerous other examples of multiple equilibria under RE. One example is provided by a variation of the Cagan model $p_t = \alpha E_t p_{t+1} + \beta m_t$, where now money supply is assumed to follow a feedback rule $m_t = \bar{m} + \xi p_{t-1} + u_t$. (Here also for convenience $v_t = 0$). This leads to the equation

$$p_t = \beta\bar{m} + \alpha E_t p_{t+1} + \beta\xi p_{t-1} + \beta u_t \tag{16}$$

For many parameter values this equation yields two RE solutions of the form

$$p_t = k_1 + k_2 p_{t-1} + k_3 u_t \tag{17}$$

where the k_i depend on the original parameters $\alpha, \beta, \xi, \bar{m}$. In some cases both of these solutions are stochastically stationary.

Other types of multiplicity of RE solutions arise in nonlinear models. Many nonlinear models can be put in the general form

$$y_t = F(y^e_{t+1}) \tag{18}$$

where random shocks have here been left out for simplicity. If y_t is increasing in y^e_{t+1}, i.e., $F' > 0$, there may be multiple steady states that satisfy $\bar{y} = F(\bar{y})$, i.e., that occur at the intersection of the graph of $F(.)$ and the 45° line. This possibility can arise in models with increasing returns to scale in production, externalities, or monopolistic competition. y is often a measure of output or aggregate economic activity and the low steady states represent inefficient '*coordination failures*' (Cooper 1999). Other specifications of Eqn. (18) have multiple perfect foresight equilibria taking the form of regular cycles in addition to a steady state.

Nonlinear models can also exhibit sunspot equilibria, taking the form of a finite state Markov process. For example, in models with two steady states it can be shown that there are RE solutions, depending on a two-state Markov sunspot process, in which the solution alternates between values close to the perfect foresight steady states.

6. *Learning*

The RE approach presupposes that economic agents have a great deal of knowledge about the economy. Even in the simple examples, in which RE are constant, computing these constants requires the full knowledge of the structure of the model, the values of the parameters and that the random shock is distributed *iid*. In empirical work economists, who postulate RE, do not themselves know the parameter values and must estimate them econometrically. A more plausible view of rationality might thus be that the agents also act like statisticians or econometricians when doing the required *forecasting*. This insight is the starting point of the *adaptive learning* approach to modeling expectations formation.

Taking this approach immediately raises the question of its relationship to RE. In many cases learning can provide at least an asymptotic justification for the RE hypothesis. For example, in the cobweb model (3), if agents estimate an unknown constant expected value by computing the sample mean from past prices one can show that expectations will converge over time to the RE value. In more general models convergence to RE can occur if agents use the appropriate econometric functional form and run regressions in the same way that an econometrician might.

Another major advantage of the learning approach arises when there are multiple equilibria. Consider the above example with a monetary feedback rule in which there are two solutions of the form (17). For the pure RE approach this is a conundrum. Which solution should we and the agents choose? Similarly, in nonlinear models of the form $y_t = F(y^e_{t+1})$ with multiple steady states, cyclic solutions or sunspots, one can ask which equilibria are stable under learning. The adaptive learning approach provides answers to all these questions, as discussed in the next section.

Finally, the transition under learning to RE may itself be of interest. The process of learning adds dynamics that are not present under strict rationality and they may be of empirical importance. In the cases just described these dynamics disappear asymptotically. However, there are various situations in which one can expect learning dynamics to remain important over time. For example, if the economy undergoes structural shifts from time to time then agents will need periodically to relearn the relevant stochastic processes.

6.1 *Statistical Approach to Learning*

In the adaptive learning approach economic agents behave like statisticians or econometricians when forecasting economic variables needed in their decision making. The economy is taken to be in a temporary equilibrium in which the current state of the economy depends on expectations. The learning approach to expectations formation makes the forecast functions and the estimation of their parameters fully explicit. A novel feature of this situation is that the expectations and forecast functions influence future data points.

As an illustration, consider again the cobweb model (3). Assume that agents believe that the stochastic process for the market price takes the form $p_t = constant + noise$, i.e., the same functional form as the RE solution. The sample mean is the standard way for estimating an unknown constant, and in this example it is also the forecast for the price. Thus agents' expectations are given by $p^e_t = \frac{1}{t}\sum_{i=0}^{t-1} p_i$. Combining this with Eqn. (3) leads to a fully specified stochastic dynamic system. It can be shown that the system under learning converges to the RE solution if $\alpha < 1$.

If the economic model incorporates exogenous or lagged endogenous variables, it is natural for the agents to estimate the parameters of the perceived process for the relevant variables by means of least-squares regressions. As an illustration suppose that an observable exogenous variable w_{t-1} is introduced into the cobweb model, so that Eqn. (3) now takes the form

$$p_t = \mu + \alpha p^e_t + \delta w_{t-1} + \eta_t \tag{19}$$

It would now be natural to forecast the price as a linear function of the observable w_{t-1}. In fact, the unique RE solution is of this form:

$$p_t = \bar{a} + \bar{b} w_{t-1} + \eta_t \tag{20}$$

where $\bar{a} = (1-\alpha)^{-1}\mu$ and $\bar{b} = (1-\alpha)^{-1}\delta$ and the corresponding rational forecast is $p_t = \bar{a} + \bar{b}w_{t-1}$.

Under *least-squares learning* agents would forecast according to

$$p_t^e = a_{t-1} + b_{t-1}w_{t-1} \tag{21}$$

where a_{t-1} and b_{t-1} are parameter estimates obtained by a least-squares regression of p_t on w_{t-1} and an intercept, using the data available through time $t-1$. It can be shown that (a_t, b_t) converges to the unique REE $(\bar{a}, \bar{b})$ if $\alpha < 1$.

6.2 Stability Under Learning

In general, when expectations are modeled by least-squares learning there is convergence to the REE as $t \to \infty$ provided that a stability condition is met. The stability condition can usually be obtained by the *expectational stability* approach.

To illustrate this approach for the cobweb model (19), suppose that expectations are based on the *perceived law of motion* (PLM) $p_t = a + bw_{t-1} + \eta_t$ and hence given by $p_t^e = a + bw_{t-1}$, where (a, b) may not be the REE values. Substituting this into Eqn. (19) we obtain the corresponding *actual law of motion* (ALM):

$$p_t = (\mu + \alpha a) + (\delta + \alpha b)w_{t-1} + \eta_t \tag{22}$$

This yields a mapping from the PLM parameters (a, b) into the ALM parameters $T(a, b) = (\mu + \alpha a, \delta + \alpha b)$. Only at the REE values does one have $T(a, b) = (a, b)$. Expectational stability (*E-stability*) looks at whether the REE is the stable outcome of a process in which the parameters of the PLM are adjusted slowly toward the parameters of the ALM that they induce. Formally, this adjustment is described by a differential equation and E-stability corresponds to local stability of the REE under these dynamics. For details, see Evans and Honkapohja (2001).

In the cobweb model the E-stability condition is given by $\alpha < 1$. Even in fairly complicated models it is often straightforward to compute E-stability conditions and these appear generally to govern convergence of statistical learning rules.

Local stability under adaptive learning, or E-stability, provides a selection criterion in models with multiple RE equilibria. For example, in the Cagan model with policy feedback (16) there are two RE solutions of the form (17), but only one of these is E-stable and hence stable under least-squares learning. The other solution, although rational, is therefore unlikely to arise in practice. Similarly, in the nonlinear model $y_t = F(y_{t+1}^e)$, only the steady states with slope $F'(\bar{y}) < 1$ are E-stable. This model can also have sunspot solutions and it has been shown that an appropriate form of adaptive learning can converge to rational sunspot equilibria, provided the corresponding E-stability condition is satisfied.

6.3 Other Approaches to Learning

There are some other recent approaches to modeling expectation formation. '*Eductive learning*' and '*rational learning*' are closer to RE than adaptive learning. In the former agents engage in a process of reasoning using common-knowledge assumptions and the learning takes place in logical or notional time. Rational learning takes place in real time, but retains the RE equilibrium assumptions at each point in time.

In contrast, the adaptive learning approach assumes that agents possess a form of *bounded rationality* which may, however, approach rational expectations over time. Besides the statistical formulation outlined above, alternative models of bounded rationality learning have been studied. Concepts from *computational intelligence*, such as genetic algorithms, neural networks and classifier systems, have been used as models of adaptive learning in economics.

Finally, we note that the literature has also considered learning in situations with structural shifts and/or misspecification of the perceived stochastic process. Certain learning rules, such as constant gain algorithms, can track structural changes more accurately than least-squares algorithms, although they often fail to converge fully to an RE equilibrium if the economic structure is in fact constant. Such procedures can lead to new forms of persistent *learning dynamics* which may be empirically important.

See also: Asset Pricing: Derivative Assets; Bounded Rationality; Keynes, John Maynard (1883–1946)

Bibliography

Cooper R 1999 *Coordination Games, Complementarities and Macroeconomics*. Cambridge University Press, Cambridge, UK

Evans G W, Honkapohja S 2001 *Learning and Expectations in Macroeconomics*. Princeton University Press, Princeton, NJ

Farmer R E 1999 *The Economics of Self-fulfilling Prophesies*, 2nd edn. MIT Press, Cambridge, MA

Guesnerie R, Woodford M 1992 Endogenous fluctuations. In: Laffont J-J (ed.) *Advances In Economic Theory: Sixth World Congress*. Cambridge University Press, Cambridge, UK, Vol. 2

Keynes J M 1936 *The General Theory of Employment, Interest and Money*. Macmillan, London

Lucas R E Jr 1976 Econometric policy evaluation: A critique. In: Lucas R E Jr (ed.) *Studies in Business Cycle Theory*. MIT Press, Cambridge, MA (originally published in 1976)

Lucas R E Jr 1981 *Studies in Business Cycle Theory*. MIT Press, Cambridge, MA

Lucas R E Jr, Sargent T J (eds.) 1981 *Rational Expectations and Econometric Practice*. University of Minnesota Press, Minneapolis, MN

Muth J F 1961 Rational expectations and the theory of price movements. *Econometrica* **29**: 315–35

Pesaran M H 1987 *The Limits to Rational Expectations*. Blackwell, Oxford, UK

Sargent T J 1987 *Macroeconomic Theory*, 2nd edn. Academic Press, New York

G. W. Evans and S. Honkapohja

Experiential Psychotherapy

1. Introduction

The term *experiential* is defined in the *Concise Oxford Dictionary* (1990) as 'involving or based on experience' and experience is defined as the 'actual observation of, or practical acquaintance with, facts or events' or 'to feel or be affected by (an emotion, etc.).' A distinction between two ways of knowing, knowledge by acquaintance and knowledge by description, first made by St. Augustine, and later emphasized in the epistemologies of William James and Bertrand Russell, helps describe the essence of experiential therapy. Experiential therapy is an approach to therapy that focuses on promoting knowledge by acquaintance. Here a person does not come to know something about him- or herself conceptually but rather has an emotional experience of it. In experiential therapy the clients' experiencing process is kept as a continuous point of reference for all therapist responses and change is seen as occurring by the promotion of new in-session experience.

Although the term experiential therapy first arose to refer to Whitaker and Malone's approaches to psychodynamically oriented family and individual therapy (Whitaker and Malone 1953), this approach has subsequently grown out of a variety of 'third force,' humanistic, approaches. These include client-centered (Rogers 1959), Gestalt (Perls et al. 1951), and existential therapy (May and Yalom 1995), as well as other approaches that emphasize working with the clients in-therapy experiencing process (Greenberg et al. 1998). Experiential therapy consists of those approaches that, within the context of a facilitative human relationship, focus on the creation of new meaning by the symbolization of experience in awareness. The two major foci of experiential therapy practice are the personal presence of the therapist in the provision of an emphatic and facilitatively genuine therapeutic relationship, and increasing clients' awareness of internal experience by focusing them on their moment by moment subjective experience.

2. View of Human Nature and Functioning

Experiential therapy adopts an existential view of human nature, one that sees existence as preceding essence. People are seen as active agents in the construction of their own realities and choice is seen as the final arbiter in human functioning. Individuals are seen as being born morally neutral with a penchant both for health and sickness, and good and bad. What is essential, however, is that people are seen as having innate worth, the ability to know the difference between good and evil, and the capacity to choose. Free will thus is the foundational principle and therapy involves facilitation of the actualization of the potential for healthy functioning. Health is seen as arising from the ability to symbolize experience in awareness and to integrate different parts of the self to create equilibrium and coherent meaning.

Experiential theory of personality proposes a process model of the self. The self is seen as a dynamic experiential system that is in a continual process of self-organization. Experiential theory is constructivist in nature, and adopts a dynamic and dialectical view of functioning. It emphasizes that meaning is created by human activity, in dialogue with others, that change is an inherent aspect of all systems and that this occurs by a synthesis of parts. Meaning construction is viewed as being constrained by a bodily felt emotional experience but as influenced by language and culture and ultimately as created by a synthesis of experience and symbol (Gendlin 1962, Greenberg et al. 1993, Greenberg and Pascual-Leone 1995). Emotional experience is seen as both creating and being created by its conscious symbolization and expression. This view casts people as creators of the self they find themselves to be. In addition, experiential therapy, by adopting a relational view of functioning, sees the self as coming into existence at the boundary between inside and outside, between organism and environment, by a synthesis of bodily experience, symbol and interpersonal validation.

Emotional experience, although seen as a basically healthy resource, is viewed as capable of either providing healthy adaptive information based on its biologically adaptive origins or, in certain instances, as having become maladaptive through learning and experience. The most basic process for the individual in therapy is thus one of developing awareness of emotion and discriminating which emotional responses are healthy and can be used as a guide, and which are maladaptive, and need to be changed.

2.1 Goals

The general goals of treatment are to promote more fluid and integrative self-organizations. Therapy focuses on the whole person, i.e. it is person centered rather than problem or symptom focused, but within this wholistic focus, the underlying determinants of different types of self-dysfunction are also an important point of focus. Thus both a change in manner of functioning of the whole self and changes in particular problems in self-organization are viewed as important. For example, a client may be seen as changing the manner of functioning both by becoming more empathic to the self and by being able to symbolize bodily felt experience, as well as changing

by resolving a specific problem such resolving unfinished business with a significant other. Thus treatment in addition to promoting self-acceptance and a strengthening of the self also aims at solving particular problems of self-organization that emerge in treatment. The problems that are focused on emerge in a collaborative fashion over the course of treatment.

3. *Basic Principles of Practice*

Experiential therapy recognizes both the power of the understanding relationship and the importance of different in-therapy tasks in promoting different types of therapeutic change. The quality of the bond between participants as well as collaboration on the tasks and goals of therapy are seen as essential in creating a good therapeutic alliance. The bond is warm, respectful, empathic, and validating, while the goals and tasks focus on increasing awareness, deepening experience, and resolving specific in-session emotional problems.

The relational bond is seen as involving three main healing ingredients: first, a more transcendent aspect, the human *presence* of the therapist, witnessing and validating the other's humaness; second, a set of more explicit facilitative *attitudes* that create a safe working environment; and third, a set of specific interpersonal *behaviors* that facilitate growth and provide new interpersonal experience. Buber's (1958) 'I–thou' relationship, involving such elements as presence, commitment to dialogue, and nonexploitiveness, and the Rogerian triad (Rogers 1957) of empathy, positive regard, and congruence together describe the general nature of the relationship. Empathy is seen as a complex cognitive/affective process of imaginative entry into the world of the other and involves understanding, in addition to being with, the other and the promotion of deeper experience. This process helps the client to feel connected, to regulate affect, and to construct new meaning.

A relational bond of this type is seen as both confirming the client as an authentic source of experience and as providing the optimal context for helping the client to attend to and become aware of prereflective experience and to communicate and explore it without fear of evaluation. The facilitative relationship, in addition to being curative in and of itself, also provides a safe environment for working on particular problems. Finally, not only does the relationship serve as a confirming environment and as a context for specific forms of intrapsychic work, but it is itself also a medium for specific corrective interpersonal experiences. Thus certain forms of work on the relationship between client and therapist are also seen as mutative.

In addition to the provision of a 'healing' relationship, facilitating work on particular therapeutic tasks is also seen as a core ingredient of experiential therapy (Greenberg et al. 1993). The most central task of experiential therapy is that of deepening the clients experiencing. This involves focusing clients on their internal experience, helping them to symbolize it in words, and create new meaning. The therapist promotes different internal processes at different times to aid experiential processing. The processes facilitated range from, symbolizing a bodily felt sense, to evoking memories, to letting an intense feeling form, to expressing feelings, to reflecting on experience to create new meaning. The client in addition is encouraged to engage in such activities as psychodramatic enactments, or exercises in imagination to help address particular emotional issues. Specific tasks such as resolving problematic reactions, or resolving splits between parts of the self, are worked on in therapy. Resolving problematic reactions for example involves re-evoking the situation in a vivid manner in order to re-experience what occurred in order to develop a new view of self and situation. Resolving splits involves the promotion of separation and contact between two parts of the self in a dialogue to facilitate self-acceptance. In these tasks the client is seen as an active agent engaging in an exploratory process, and the tasks are viewed as being done *by* clients rather than done *to* them.

3.1 *A Process Theory*

An important distinguishing characteristic of experiential therapy is that it offers a process theory of how to facilitate knowledge of acquaintance, rather than a content theory of personality or psychopathology. A process theory of this type specifies both the moment by moment steps in the client's process of change and the therapist interventions that will facilitate these steps. The emphasis in each step always is on how to promote the direct sensing of what is concretely felt in the moment to create new meaning.

There is an explicit assumption that within each individual there is a flow of experiencing to which the person can refer in order to be informed about the personal meaning of particular experiences. The main principle of the experiential method is to have people check whatever is said or done against their own concretely felt experience. Change is seen as emerging from a growing awareness of previously unsymbolized experience and the bringing of this experience into dialectical interaction with words or symbols and other aspect of experience to create new meaning.

The key to experiential therapy is to have clients experience or become acquainted with any manner of content in a new way such that this new experience will produce a change in the way they view themselves, others, and the world. Experiential therapy thus adds to the other approaches the emphasis that symbols, schemes, and even behavior must interact with the body based, experiential, level of existence in order to produce change. It thus offers a process theory of how body and symbol interact, and a set of methods for promoting this process.

Experiential therapy theorists have specified patterned sequences of change processes and events, the explicit connections of which are spelled out by their experiential theories of change. Three such characteristic sequences are described below. One sequence offered by Gendlin (1996) includes three fundamental client change events: (a) the client focuses on a directly felt meaning; (b) the client allows feelings, words, and pictures to arise from this inward focusing, and attends to the generated feeling; (c) the client receives a new felt meaning that emerges from the ensuing shift in body experience. Another experiential sequence (Mahrer 1989) consists of four client basic change events: (a) the client attains a level of strong feeling; (b) the client welcomes and appreciates the accessed inner experiencing; (c) the client becomes the inner experiencing in the context of earlier life scenes; (d) the client becomes and behaves according to the inner experiencing in the context of imminent future life scenes.

A third sequence specified by Greenberg and Paivio (1997) involves six basic change events: (a) the client experiences the problematic bad feelings in the session; (b) the client accesses, allows, and receives deeper core emotions and needs in the session; (c) the patient and therapist together explore whether the core emotion is adaptive or maladaptive; (d) if judged to be adaptive, the core feelings are used as a guide, and if judged to be maladaptive, alternative adaptive emotions and needs are accessed; (e) the core maladaptive emotions and associated beliefs are challenged from within by the newly adaptive emotions and needs; (f) new meaning is created based on the new experience that emerges from the dialectical interaction of adaptive and maladaptive parts of self.

In these processes, the therapist is seen as an expert on how to facilitate new steps in the client's experience rather than an expert on the content of the client's experience. Therapists thus avoid interpretations of the content of a client's experiences that tell the patient why he or she does things and that are theory-driven rather than experience-near. Responses that are conceptual or explanatory, or are expressed as fact, or that give the message that truth comes from the therapists' professional knowledge are avoided in favor of phenomenologically refined exploration of the client's experience. Experiential therapy thus attempts to eliminate any interpretations that are based on the therapist's theory of how people are, or should be, or that attempt to reveal hidden 'truths.'

The experiential approach rather places great emphasis on, and respect for, what the patient experiences, and pays special attention to what the patient experiences in the relationship with the therapist. Experiential therapy involves consistent listening from within the patient's frame of reference. Hence sustained empathic inquiry is a central part of the practice. However, although this empathic emphasis is necessary for good psychotherapy, it is not necessarily sufficient for the best psychotherapy. The most effective psychotherapy also requires therapist practices of a technical nature, such as phenomenological focusing, experimentation, and behavioral observation (Gendlin 1996, Perls et al. 1951). Diagnostic understanding of the individual, and also understanding the social, cultural, and institutional forces affecting the individual, are also required.

Notwithstanding the fact that what the patient experiences is the indispensable essence of psychotherapy, and that it is imperative that this be the subject of respectful ongoing inquiry by therapist and patient, what the patient does not experience is also an indispensable and critical component of what happens in therapy. Important factors that are kept from the patient's conscious experience that do not become figural in awareness also need to be the subject of phenomenological exploration. In experiential therapy there is a special emphasis on bringing into awareness the processes that regulate this process of awareness/unawareness (Polster and Polster 1973, Yontef 1993). This work is codirected by therapist and patient and is not based on any alleged higher truth of the therapist. It is based on a joint empathic exploration that may include observation, experimentation, and dialogue.

4. *Conclusion*

In the experiential approach, an empathic focus on the patient's actual experience is seen as indispensable, but the therapist is also seen as making contributions in addition to sustained empathic inquiry. Therapists complement the empathic inquiry with a variety of therapeutic interventions that can help the patient learn how efficiently to focus awareness and distinguish actual phenomenological experience, that can highlight the processes essential to the patient's characterological organization by sophisticated and competent understanding that includes understanding theories of healthy functioning and psychopathology.

Experiential therapy has been demonstrated to be effective in treating depression, anxiety, childhood maltreatment, and marital distress in addition to some personality disorders and a variety of other problems in living (Greenberg et al. 1994, 1998).

See also: Constructivist Psychotherapies; Person-centered Psychotherapy; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported

Bibliography

Buber M 1958 *I and Thou*. Scribner, New York

Concise Oxford Dictionary, 8th edn. 1990 Clarendon Press, Oxford, UK

Gendlin E T 1962 *Experiencing and the Creation of Meaning*. Free Press of Glencoe, New York
Gendlin E T 1996 *A Focusing Approach to Psychotherapy*. Guilford Press, New York
Greenberg L, Elliott R, Lietaer G 1994 Research on experiential psychotherapy. In: Bergin A E, Garfield S L (eds.) *Handbook of Psychotherapy, Behavior Change*. Wiley, New York
Greenberg L, Paivio S 1997 *Working with Emotion in Psychotherapy*. Guilford Press, New York
Greenberg L, Pascual-Leone J 1995 A dialectical constructivist approach to experiential change. In: Neimeyer R A, Mahoney M J (eds.) *Constructivism in Psychotherapy*. American Psychological Association, Washington, DC, pp. 169–91
Greenberg L, Pascual-Leone J 1997 Emotion in the creation of personal meaning. In: Power M, Brewin C (eds.) *Transformation of Meaning*. John Wiley, Chichester, UK
Greenberg L S, Rice L N, Elliot R 1993 *Facilitating Emotional Change: The Moment by Moment Process*. Guilford Press, New York
Greenberg L, Watson J, Lietaer G (eds.) 1998 *Handbook of Experiential Therapy*. Guilford Press, New York
Mahrer A R 1989 *How to Do Experiential Psychotherapy: a Manual for Practitioners*. University of Ottawa Press, Ottawa, Canada
May R, Yalom I 1995 Existential psychotherapy. In: Corsini R, Wedding D (eds.) *Current Psychotherapies*, 4th edn. Peacock, Itasca, IL, pp. 262–92
Perls F, Hefferline R F, Goodman P 1951 *Gestalt Therapy*. Dell, New York
Polster E, Polster M 1973 *Gestalt Therapy Integrated*. Brunner/Mazel, New York
Rogers C R 1957 The necessary and sufficient condition of therapeutic personality change. *Journal of Consulting Psychology* **21**: 95–103
Rogers C R 1959 A theory of therapy, personality and interpersonal relationships, as developed in the client-centered framework. In: Koch S (ed.) *Psychology: a Study of a Science*. McGraw-Hill, New York, Vol. 3, pp. 184–256
Whitaker, C A, Malone T P 1953 *The Roots of Psychotherapy*. Blakiston, New York
Yontef G M 1993 *Awareness, Dialogue and Process: Essays on Gestalt Therapy*. Gestalt Journal Press, Highland, NY

L. Greenberg

Experiment, in Science and Technology Studies

Experiment was a major concern of science and technology studies (STS) during the last quarter of the twentieth century. This is because certain views of experimentation as a privileged way of making knowledge about nature were important focuses of dissent during the specialty's early career. Various versions of the so-called 'received view' of science prevalent in Anglo-American philosophy of science in the period following World War II, with their roots in logical empiricism, held up experiments as the crucial empirical elements of scientific knowledge. Whether the logic employed in their modeling of scientific reasoning was falsificationist or conformationist, versions of the 'received view' represented theories as conceptual structures that coordinate with the world through independent experimental tests of their predictions. This article will survey some of the major developments that have emerged to modify those older views.

1. Development of New Approaches to Experiment: Challenges to the 'Received View'

In the 1950s and early 1960s, one of the central assumptions of the 'received view' was seriously challenged. The presumed independence of experimental and observational data from contamination by the very theories that the data were intended to judge was challenged by W. V. O. Quine (Quine 1953), Norwood Russell Hanson (Hanson 1958), and Thomas S. Kuhn (Kuhn 1970 [1962]). Quine showed, using an argument similar to one used by Pierre Duhem (1954), early in the twentieth century, that empirical data always underdetermine the theories that can rest upon them. In Duhem's hands, the argument rested on a position of 'conventionalism,' whereby the relationship between items of discrete empirical data and the theories employing those data was seen to be underdetermined; in particular, Duhem argued that there could be no such thing as a 'crucial experiment' that could decide definitively between two alternative theories. A result inconsistent with one of the theories need not be attributed to an inadequacy in the target theory itself, but might instead be attributed to error in a subsidiary theory (such as one relating to observational instruments) that had not initially been in question. Possibilities of that sort amounted to decisions, according to Duhem, and were consistent with a conventionalist view of scientific theories, whereby pure convention determined which in a network of theories would be preserved and which amended so as to fit experimental results.

Hanson and Kuhn went further, with their arguments in favor of the 'theory-ladenness of observation.' They held that a pure observation language was in principle impossible, and that, accordingly, the results of experiments and observations of nature always implicated theoretical concepts in their very formulation. This position blurred the distinction between empirical data on the one hand and the theories that are justified by those data on the other. Both Kuhn and Hanson likened empirical data to the psychologist's gestalt experiments, wherein, for example, the 'same' image can sometimes be seen as a duck and sometimes as a rabbit; the difference, they claimed, was that in science there is no possibility of resting one's description on an account of the underlying set of lines. Instead, the scientist sees a duck or a rabbit, with no further reduction to a common

observational base being possible. All empirical data are already interpreted; that is, seen in a particular way (cf. Fleck 1979 [1935], an acknowledged source for Kuhn).

2. *Sociological and Historical Exploitations of New Theoretical Arguments*

These challenges to the 'received view' of science provoked much debate in the 1960s and into the 1970s. In its wake, however, there arose attempts to exploit the ideas of such as Hanson and Kuhn in the service of sociological projects aimed at making sense of scientific knowledge seen as a social product. In the early days of the 'sociology of scientific knowledge' (SSK), the Duhem–Quine 'underdetermination' thesis became a standard in-principle justification for sociological studies of the acceptance of knowledge claims by scientific communities (e.g., Mulkay 1979, Knorr-Cetina and Mulkay 1983). This was so because it was argued that the inadequacy of empirical evidence to necessitate any particular theoretical belief in science left room for social explanations regarding which belief among the manifold logical possibilities was the one actually chosen. At the same time, however, a more fundamental theoretical challenge to the self-sufficiency of 'received view' accounts of the role of experiment in science was being developed by H. M. Collins, drawing on a sociological reading of Wittgenstein (Winch 1958, cf. Bloor 1983).

Wittgenstein had argued that rules do not contain rules for their own application; such a position, in the form of 'finitism,' also applies to practices of classification (see, for a summary of its relevance to STS, Barnes et al. 1996). The basic claim amounts to observing that every new situation, every new occurrence of some kind of event, is in its own way unique; it therefore requires indeterminate judgments of similarity and difference to decide whether a new event or instance is or is not a legitimate member of a given category. In articles from the mid-1970s onwards, and culminating in a book of 1985 (Collins 1985), Collins developed empirical case studies that exploited Wittgenstein's arguments to show how the participants in actual scientific controversies resolve problems of the ambiguity of experimental outcomes. In work on claims to the detection of high-flux gravitational radiation in the 1970s, as well as in the transfer of skills relating to the successful replication of TEA-laser technology, Collins developed the notion of the 'experimenters' regress,' which he saw as an application in the domain of the sociology of science of Wittgenstein's 'indeterminacy of rules.' In the case of an experimental result, a judgment must be made that involves deciding whether the experiment or experimental run was competently conducted. If yes, then the outcome can be used to assess an associated theoretical claim (which would then open up Duhem–Quine issues), whereas if not, then the outcome will be irrelevant to the claim. Collins showed that, routinely, such assessments of experimental competence were central to scientific controversies. The ultimate criterion for judging the competence of an experiment, according to Collins, was whether the experimental outcome was correct or not. In the case of experiments the results of which were taken to depend on unknown or controversial aspects of the natural world, therefore, controversy over nature was coextensive with controversy over which experiments were fit to judge concerning that feature of nature. This, then, is the essential paradox designated by the term 'experimenters' regress': it is only possible to know whether an experiment was competently performed if it is known what the correct outcome should be, but (in an empiricist epistemology) the correct outcome can only be known as a consequence of the performance of competent experiments.

The sort of empirical studies of scientific practice conducted by Collins also yielded fruitfully to analysis in terms of the Duhem–Quine thesis. Trevor Pinch's *Towards an Analysis of Scientific Observation* (Pinch 1985) introduced the idea of 'externality' in understanding the role and fate within the scientific community of experimental work. Pinch noted that knowledge-claims made on the basis of experimental studies could themselves be couched in alternative ways that maximized or minimized the likelihood that the competence and reliability of the experiments themselves would be questioned. An experimentally based assertion with high externality would be one that used the experimental outcome to make far-reaching inferences about some domain of existing knowledge; an assertion with low externality would depart but little from a bald statement of the local circumstances of the experimental situation, and could provoke but little controversy. Andrew Pickering (Pickering 1984) examined similar professional dynamics in the work that established the existence and characteristics of quarks in the 1970s.

Together, the Duhem–Quine thesis, finitism, and the experimenters' regress had opened up, by the mid-1980s, a vast terrain within which sociologically informed study of the experimental dimensions of scientific knowledge could be conducted. A landmark study that fully exploited the new possibilities was Steven Shapin and Simon Schaffer's *Leviathan and the Air-Pump* (Shapin and Schaffer 1985). This historical study directly challenged the older 'received view' conception of the role of experiment in making scientific knowledge, and did so by investigating a classic, oft-cited historical example usually held to exemplify that view. The subject chosen in the book was the air-pump work of Robert Boyle in the 1660s. Boyle's work had been presented (Conant 1970 [1948]) in one of the *Harvard Case Histories in Experimental Science*, dating from the late 1940s and 1950s. These were publications that, importantly, had shaped the

teaching and scholarship in the history of science for many years, and which were consistent with the usual conception of experiments as unequivocal arbiters of theoretical ideas. Shapin and Schaffer thus accepted Boyle's air-pump experiments, for the sake of argument, as an exemplary model of scientific procedure. They then attempted to show how a close examination revealed that Boyle's experiments and their associated knowledge claims were far from unequivocal—were, in fact, strongly contested—and that the interpretation of them advanced by Boyle and his allies could also be understood as part of a political vision of the proper social constitution and conduct of a philosophical community. In order to achieve this purpose, Shapin and Schaffer used the criticisms of Boyle advanced by the philosopher and mathematician Thomas Hobbes, a man whose view of the proper way to make natural knowledge differed significantly from that of Boyle and whose associated conception of the place and legitimacy of knowledge in a polity also differed markedly from Boyle's. Hobbes's detailed criticisms of Boyle's air-pump work revealed an alternative set of values and commitments regarding the best way to achieve social stability and consensus, a way that invalidated experimental work as an effective means of securing knowledge claims. An alternative approach to similar issues (Dear 1985, 1987, 1995) focuses on conventions of knowledge making that were less local than those examined by Shapin and Schaffer, and which support a view of much more mediated connections with social organization. Dear's main argument is that the modern 'experiment' as a historically specific, contrived event was a development of seventeenth-century Europe, and had close filiations with disciplinary approaches employed in mathematical sciences that stressed operationalism.

3. *Laboratory Studies and Naturalistic Approaches*

By the mid-1980s, in addition to the work already discussed, a number of so-called 'laboratory studies' were being published. These studies also had a strong influence on views of experiment in science, but, in contrast to the approaches considered above, they tended to eschew, often explicitly, an overt engagement with the 'received view' philosophies of science. Instead, the methodological and/or epistemological approaches that were central to laboratory studies drew from such traditions as those of anthropology, ethnomethodology, and phenomenology. The three crucial and exemplary lab studies from this period were Latour and Woolgar's *Laboratory Life* (Latour and Woolgar 1979/1986), Karin Knorr-Cetina's *The Manufacture of Knowledge* (Knorr-Cetina 1981), and Michael Lynch's *Art and Artifact in Laboratory Science* (Lynch 1985). These three have in common a naturalistic stance towards the work of the laboratory, rejecting presuppositions as to the character of that work. Thus, they have no *a priori* interest in studying laboratory work as a means of learning how 'reliable,' or 'true,' or 'epistemologically privileged' knowledge results from laboratory practices. Instead, they treat their particular laboratories as sites where activities are carried out, and attempt to describe and account for the nature of those activities. A notable feature of the rapidly changing intellectual landscape on which these laboratory studies were performed is the existence, in the first edition of *Laboratory Life*, of the subtitle *The Social Construction of Scientific Facts*, and its replacement in the second edition, seven years later, by the less doctrinaire *The Construction of Scientific Facts*. The change signals the way in which, in the years surrounding 1980, the rhetoric of SSK, with its insistence on 'social construction' as the crucial mechanism by which scientific knowledge was made, dominated science studies. By 1986, this insistence was already looking rather hackneyed and uninformative in some quarters.

The place of experiment in the tradition of laboratory studies is less clear-cut than it was in the mainstream SSK tradition, because the former were not reacting directly to the epistemological highlighting of experiment characteristic of the philosophical 'received view.' Thus, experiments were less likely to be seen as obviously individuated events, instead falling within an overall passage of scientific activity. The highlighting of particular experiments as especially significant cognitive elements of such passages now became contingent on the behavior and utterances of the scientists being studied, rather than being self-evidently central to an understanding of scientific work.

A collection of papers originating in a conference of 1985 on 'The Uses of Experiment' (Gooding et al. 1989, see also Le Grand 1990 for a similar collection) attempted to shed light on experiment as a category used by scientists rather than as an epistemologically significant item singled out by philosophers. The contributors made use of the new work in science studies to lay out, chiefly by means of empirical studies of contemporary and past science, the multifarious ways in which experiment has been used as a rhetorical and practical tool in the establishment of knowledge-claims and in the consolidation of scientific communities of practice. One especially notable concept deployed in this volume, centrally in the contribution by Simon Schaffer (Schaffer 1989), was that of experimental 'transparency.' This concept designated situations in which an experimental apparatus becomes so unproblematic and theoretically unquestioned that the phenomena to which it affords access are seen directly, as it were, as if no theoretically justified experimental device intervened between observer and phenomena. In such cases, in effect, the apparatus has changed from being an experimental set-up to being a scientific instrument. For example,

such, was the fate of the mercury barometer in the seventeenth century, which began life as an experiment intended to examine the pressure of the air and the possibility of a vacuum but which, by the end of the century, had become an instrument for measuring atmospheric pressure as an aid to understanding the weather. These kinds of conceptualization also hark back to the much earlier work of the French philosopher Gaston Bachelard, who spoke of the artificial creation of phenomena by apparatus, using the term *phénoméno-technique* (e.g., Bachelard 1938, see Gaukroger 1976 for a lucid introduction).

The second half of the 1980s also saw the appearance of empirical studies of modern physical experimentation that adopted attitudes quite different from those of the advocates of SSK. Peter Galison, in his book *How Experiments End* (Galison 1987), examined several episodes in post-WWII high-energy physics in the United States. The book's intention was to describe the ways in which experimental closure was brought about in the episodes discussed, but in a way explicitly opposed to that of such as Collins; that is, Galison's aim was to show that social mechanisms were insufficient to explain the outcome of the experimental work. Allan Franklin also published work on modern experimental physics with broadly the same aim (Franklin 1986, 1990). In both cases, the role of an independent physical world in creating sometimes-unexpected outcomes to experiments was emphasized, to reduce the role of social actions and social structures in determining outcomes: in Franklin's case, at least. It was argued that social considerations become irrelevant in light of the overwhelming influence of objective experimental outcomes and their correct rational assessment. Franklin's position, however, appears largely to ignore the epistemological problems that fueled SSK.

Discussions of experiments and the best ways of understanding their role in science also went well beyond issues of their direct epistemological function. Historical studies, pre-eminently Shapin and Schaffer's *Leviathan and the Air-Pump*, were by now having an impact on wider conceptions of the ideological structure and social meaning of science. The most important example of that impact was Yaron Ezrahi's *The Descent of Icarus* (Ezrahi 1990). This took Shapin and Schaffer's presentation of Robert Boyle's kind of experimental science as the model for science ever since the seventeenth century, and adapted their view of its political meaning to an account of the development of modern liberal ideology and, subsequently, modern liberal states from the eighteenth to the twentieth centuries. In Ezrahi's view, the legitimacy of experimental knowledge claims in science became the accepted norm for knowledge claims in the political arena as well. Shapin and Schaffer had stressed the role of 'witnessing' in Boyle's kind of experimental science, whereby many people, either directly or through the reading of detailed reports ('virtual witnessing'), came to attest the truth of an asserted experimental outcome and thereby its associated factual knowledge claim. Ezrahi suggested that a similar visibility, or transparency, of political actions and the reasons for them became the political ideal, along with the notion of the accountability of political leaders. Ezrahi's ideal-type of such a polity, the United States in the twentieth century, thus placed great stress on the possibility of citizens seeing what their government was doing and in assessing the grounds for and wisdom of those actions (see also Latour 1993 for different political extrapolations).

4. *Latour on Experiments and their Mobility*

Other aspects of experimentation were also investigated in the 1980s and early 1990s, most notably in the work of Bruno Latour. Latour's naturalistic approach to scientific practice, which had been explicitly anthropological in inspiration in *Laboratory Life*, extended his representation there of laboratory science as an enterprise centrally concerned with the production of 'literary inscriptions' (any kind of trace inscribed on paper). This extension took the form, in his *Science in Action* (Latour 1987), of an analysis of the structure of experimental papers. That structure, he argued, reveals the agonistic logic of experimental procedures through a kind of shadowboxing, whereby the paper's authors attempt to anticipate and head off skeptical criticisms of their claims, such as might be put forward by a determined opponent. Thus, the experimental paper (and, by implication, the experimental protocol itself) takes the form of one side of a dialogue aimed at silencing the critic through anticipation of all plausible objections. Part of the power of this analytical approach supposedly resides in Latour's stress on semiotic analysis (Latour and Bastide 1986), which connects with his emphasis on taking into account 'natural actants' in accounts of scientific activity (cf. also Latour 1988). Thus semiotic actors in the experimental account which are normally taken to refer to things in the natural world (such entities as subatomic particles, or microbes, etc.) are treated on a par with those semiotic actors normally taken to refer to things in the social world (pre-eminently, the scientists performing the experiment). Thus, they are understood in exactly the same way. Latour therefore claims that there is no justification for treating social actors as real while treating 'natural actors' as simply social constructs, which is what the SSK approach advocates. Problems seem to arise in Latour's approach, however, when the question of shifting between the analytical level of the semiotic and the analytical level of the putatively 'real' is considered; Latour is obscure on the point.

One especially influential idea in Latour (Latour 1983, 1988) concerns a way of understanding how experimental outcomes can be mobile; that is, how a

successful experiment conducted in a specific laboratory is made to be successful at other locations. This issue has been referred to as the 'problem of constructivism' in Golinski (1998); although Latour would not count himself as a 'constructivist' in the SSK sense, his similar stress on the work that goes into creating an experimental 'fact' leaves him with the same problem. The difficulty that Latour purports to solve is to explain how, if experimental facts are made in situations whose specificity and locality ('natural' and 'social') are taken as crucial to explaining the particular outcome, that outcome is seen to be reproduced in significantly different settings. The traditional ('received view') answer is that the experimental outcome is simply a reflection of the way that nature works, and an implicit principle of the uniformity of nature therefore implies that the same experiment, in the sense of the same material set-up and procedure, will yield the same result wherever it is conducted. The constructivist, however, because of a conviction that 'social' elements (however defined in detail) are essential to producing the outcome, appears to have difficulty in explaining this mobility of experiment (cf. Collins 1985 on the TEA-laser). Specifically, Latour explains the apparent paradox of experimental success outside the laboratory setting, using Louis Pasteur's field trials of his anthrax vaccine as an example, by denying that the laboratory has ever really been abandoned. Instead, when, for instance, Pasteur managed to create a vaccine found to work on animals in carefully controlled laboratory conditions, his demonstration of its effectiveness on actual farm animals involved, according to Latour, transferring his laboratory to the farm. It was not that Pasteur's vaccine worked in the laboratory and on the farm; the vaccine only ever worked in the context of the laboratory trial. Pasteur's trick was in turning the farm into a laboratory while maintaining the illusion that it was still a farm. Hence, experiments can be spread, on this view, only via networks of laboratories or laboratory-equivalents; they do not in fact work everywhere.

5. *Experimental Systems*

In the 1990s, there appeared a number of new conceptualizations of experiment, including an important collection of essays concerning the life sciences, *The Right Tools for the Job* (Clarke and Fujimura 1992), and significant discussions of various experimental practices and regimes in Peter Galison's *Image and Logic* (Galison 1997). These new conceptualizations had in common a more expansive and complex view of the ways in which experiment forms a part of elaborate systems of scientific practice. Andrew Pickering's *The Mangle of Practice* (Pickering 1995) developed from and built upon his earlier SSK-influenced work by attempting to understand scientific practice as a pragmatic interaction between experimental procedures, theoretical models, and physical reality. The constraints on scientific ideas resulting from the interaction between these theoretical and material practices had become in Pickering's work an important and central feature that rejected any attempt at pure sociological reductionism (although whether this has ever truly existed in STS is a debated point). Pickering saw scientific work as an ongoing enterprise aimed at stabilization: theory and experiment cooperate, at least for a time, in producing mutually consistent, intelligible interactions between human beings and the world. By placing experiment within this more complex structure of scientific practice, Pickering thus endeavored to move beyond the sociological insights of science studies and to regain a connection with the natural reality with which science has always claimed to deal.

Similar pragmatic moves had also appeared in an important study by the historian Robert Kohler. In his *Lords of the Fly* (Kohler 1994), Kohler studied the early decades of *drosophila* (fruit-fly) research in the United States, centered on the laboratory of Thomas Hunt Morgan at Columbia University and subsequently at the California Institute of Technology. Kohler described the operations of Morgan's fly lab in terms of the exploitation of the capabilities of the materials, centrally including the flies, with their particular breeding and variation characteristics. The whole taken together, and crucially including material practices (given some precedence in this study over intellectual practices), was characterized by Kohler variously in terms of 'experimental practice,' 'experimental production,' or as a 'system of experimental mass production.' The term 'experimental system' was also used by Hans-Jörg Rheinberger in an ambitious study, *Toward a History of Epistemic Things* (Rheinberger 1997). Rheinberger, following the same general tendencies as Kohler and others, and with an important nod to Fleck, represented his project as being directed against the old philosophical tendency to view experiments as isolated epistemological elements, instead focusing his attention on the integration of experiment within a multifaceted system of practices which together constitute the modern scientific enterprise (cf. also Knorr-Cetina 1999, especially pp. 39–43). Consequently, experiment has lost much of its former philosophical mystique.

See also: Empiricism, History of; Experimental Design: Large-scale Social Experimentation; Experimentation in Psychology, History of; Medical Experiments: Ethical Aspects; Quasi-Experimental Designs; Single-case Experimental Designs in Clinical Settings; Truth and Credibility: Science and the Social Study of Science; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Bachelard G 1938 *La formation de l'esprit scientifique*. J. Vrin, Paris

Barnes B, Bloor D, Henry J 1996 *Scientific Knowledge: A Sociological Analysis*. University of Chicago Press, Chicago

Bloor D 1983 *Wittgenstein: A Social Theory of Knowledge*. Macmillan, London

Clarke A E, Fujimura J H (eds.) 1992 *The Right Tools for the Job: At Work in Twentieth-century Life Sciences*. Princeton University Press, Princeton, NJ

Collins H M 1985 *Changing Order: Replication and Induction in Scientific Practice*. Sage, London

Conant J B 1970 [1957] Robert Boyle's experiments in pneumatics. In: Conant J B (ed.) *Harvard Case Histories in Experimental Science*. Harvard University Press, Cambridge, MA, Vol. 1

Dear P 1985 *Totius in verba*: Rhetoric and authority in the early Royal Society. *Isis* **76**: 145–61

Dear P 1987 Jesuit mathematical science and the reconstitution of experience in the early seventeenth century. *Studies in History and Philosophy of Science* **18**: 133–75

Dear P 1995 *Discipline and Experience: The Mathematical Way in the Scientific Revolution*. University of Chicago Press, Chicago

Ezrahi Y 1990 *The Descent of Icarus: Science and the Transformation of Contemporary Democracy*. Harvard University Press, Cambridge, MA

Fleck L 1979 [1935] *Genesis and Development of a Scientific Fact*. Trans. In: Bradley F, Trenn T J (eds.). University of Chicago Press, Chicago

Franklin A 1986 *The Neglect of Experiment*. Cambridge University Press, Cambridge, UK

Franklin A 1990 *Experiment, Right or Wrong*. Cambridge University Press, Cambridge, UK

Galison P 1987 *How Experiments End*. University of Chicago Press, Chicago

Galison P 1997 *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, Chicago

Gaukroger S 1976 Bachelard and the problem of epistemological analysis. *Studies in History and Philosophy of Science* **7**: 189–244

Golinski J 1998 *Making Natural Knowledge: History of Science after Constructivism*. Cambridge University Press, Cambridge, UK

Gooding D, Pinch T, Schaffer S (eds.) 1989 *The Uses of Experiment*: *Studies in the Natural Sciences*. Cambridge University Press, Cambridge, UK

Hanson N R 1958 *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press, Cambridge, UK

Knorr-Cetina K D 1981 *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Pergamon Press, Oxford, UK

Knorr-Cetina K D 1999 *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press, Cambridge, MA

Knorr-Cetina K D, Mulkay M J (eds.) 1983 *Science Observed: Perspectives on the Social Study of Science*. Sage, London

Kohler R E 1994 *Lords of the Fly: Drosophila Genetics and the Experimental Life*. University of Chicago Press, Chicago

Kuhn T S 1970 [1962] *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

Latour B 1983 Give me a laboratory and I will raise the world. In: Knorr-Cetina K D, Mulkay M J (eds.) 1983 *Science Observed: Perspectives on the Social Study of Science*. Sage, London

Latour B 1987 *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge, MA

Latour B 1988 *The Pasteurization of France* (Trans. Sheridan A, Law J). Harvard University Press, Cambridge, MA

Latour B 1993 *We Have Never Been Modern* (Trans. Porter C). Harvard University Press, Cambridge, MA

Latour B, Bastide F 1986 Writing science–Fact and fiction. The analysis of the process of reality construction through the application of socio-semiotic methods to scientific texts. In: Callon M, Law J, Rip A (eds.) *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. Macmillan, London

Latour B, Woolgar S 1979/1986 *Laboratory Life: The [Social] Construction of Scientific Facts*. Sage, London

Le Grand H E (ed.) 1990 *Experimental Inquiries: Historical, Philosophical and Social Studies of Experimentation in Science*. Kluwer, Dordrecht, The Netherlands

Lynch M 1985 *Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory*. Routledge & Kegan Paul, London

Mulkay M J 1979 *Science and the Sociology of Knowledge*. George Allen & Unwin, London

Pickering A 1984 *Constructing Quarks: A Sociological History of Particle Physics*. University of Chicago Press, Chicago

Pickering A 1995 *The Mangle of Practice: Time, Agency and Science*. University of Chicago Press, Chicago

Pinch T 1985 Towards an analysis of scientific observation: The externality and evidential significance of observational reports in physics. *Social Studies of Science* **15**: 3–36

Quine W V O 1953 Two dogmas of empiricism. In: Quine W V O (ed.) *From a Logical Point of View: Nine Logico-philosophical Essays*. Harvard University Press, Cambridge, MA

Rheinberger H J 1997 *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford, CA

Schaffer S 1989 Glass works: Newton's prisms and the uses of experiment. In: Gooding D, Pinch T, Schaffer S (eds.) 1989 *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge University Press, Cambridge, UK

Shapin S, Schaffer S 1985 *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, Princeton, NJ

Winch P 1958 *The Idea of a Social Science and Its Relation to Philosophy*. Routledge & Kegan Paul, London

P. Dear

Experimental Design: Bayesian Designs

This article provides an overview of experimental design using a Bayesian decision-theoretic framework. Scientific experimentation requires decisions about how an experiment will be conducted and analyzed. Such decisions depend on the goals and purpose of the experiment, but certain choices may be restricted by

available resources and ethical considerations. Prior information may be available from earlier experiments or from conjectures which motivate the investigation. The Bayesian approach provides a coherent framework where prior information and uncertainties regarding unknown quantities can be combined to find an experimental design that optimizes the goals of the experiment.

1. Introduction

Experimentation plays an integral part in the scientific method. Conjectures and hypotheses are put forth based on the current state of knowledge. Experimental data may be collected to address unknown aspects of the problem. Finally, analysis of experimental results may lead to one hypothesis being favored over others or may lead to new questions and investigations, so that the process is repeated, with the accumulation of additional knowledge about the scientific process under investigation.

In some fields of scientific inquiry, physical models can be used to describe the outcome of an experiment given certain inputs with complete certainty. In the majority of applications, one cannot describe the scientific phenomena perfectly, leading to a distribution of possible outcomes, which can be described by probability models. For example, in comparing a new therapy to an existing treatment, individuals receive one of two treatments. The outcome or response is an indicator of 'success' or 'failure' of the given treatment, which can be modeled using Bernoulli distributions with unknown success probabilities. In psychological or educational testing, the outcome may be scores on a battery of tests. The outcomes or responses may depend on many other factors besides the assigned treatment, such as age, education, gender, or other explanatory variables. Hypotheses or quantities of interest are often phrased in terms of parameters θ of statistical models which relate the distribution of outcomes Y to levels of explanatory variables.

In conducting an experiment, there are many design issues to resolve, including deciding which treatments to study, which factors to control, and what aspects of an experiment to randomize. Other aspects of experimental design—such as how many experimental units are needed, how many observations should be allocated to each treatment, or what levels of other input or control variables should be used—have traditionally fallen under the umbrella of statistical design (see *Experimental Design: Overview*). Because of costs, ethics, or other limitations on resources or time, sample sizes are usually restricted, and efficient use of available resources is critical. The purpose of optimal experimental design is to improve statistical inference regarding the quantities of interest by the optimal selection of values for design factors under the control of the investigator, within, of course, the constraints of available resources. Decision theory (see *Decision Theory: Bayesian*) provides a mathematical foundation for the selection of optimal designs. Prior information from earlier related experiments, observational studies, or subjective beliefs from personal observations, can be valuable in deciding how to allocate treatments efficiently, leading to more informative experiments. The Bayesian approach to experimental design provides a formal way to incorporate such prior information into the design process.

2. Bayesian Optimal Designs

The statistical aspects of an experiment e may be formally described by the sample space Ω (possible outcomes for the response Y), the parameter space Θ, and a probability model $p_e(y \mid \theta)$ that represents the distribution of observable random variables Y indexed by a parameter θ, an element of the parameter space Θ. Sample sizes, treatment levels, number of treatments, levels of explanatory variables, or other aspects of the design to be selected are implicitly contained in $p_e(Y \mid \theta)$. The primary goal(s) or terminal decision(s) of an experiment may include, but are not limited to, estimating θ or other quantities of interest that are functions of θ, predicting future observations, selecting among competing models, or testing other hypotheses.

Lindley presented a two-part decision theoretic approach to experimental design, which provides a unifying theory for most work in Bayesian experimental design today. Lindley's approach involves specification of a suitable utility function reflecting the purpose and costs of the experiment; the best design is selected to maximize expected utility. In this framework, an experiment e is selected from the possible collection of experiments E (the first decision problem). After choosing an experiment e, outcomes Y are observed. Based on the observed data Y and experiment e, a terminal decision d is selected from possible decision rules D, which addresses the terminal goal(s) of the experiment. A utility function in the form $U(d, \theta, e, Y)$ encodes the costs and consequences of using experiment e and decision d with data Y and parameter θ. Assuming that the goals of an experiment and terminal decision can be formally expressed through a utility function, the Bayesian solution is to find the best design and best decision rule that maximize expected utility.

While the process of experimentation followed by inference/decision making proceeds in time order, it is easier to solve the optimal decision problem in reverse time order. The terminal-stage decision problem involves finding the best decision rule d given the observed data Y under experiment e that maximizes the *posterior* expected utility,

$$\max_d \int_\Theta U(d, \theta, e, Y) p(\theta \mid Y, e)\, \mathrm{d}\theta = U(e, Y) \quad (1)$$

Here, the expectation or averaging over θ accounts for uncertainty regarding the unknown θ. The expectation is taken with respect to the posterior distribution of θ, which properly reflects uncertainty in θ at the terminal stage after Y has been observed under experiment e.

As the experiment e must be selected before data Y are observed, the second-stage optimization problem involves finding the best experiment e that maximizes the *preposterior* expected utility. The preposterior expected utility is obtained by integrating the result in Eqn. (1) over possible outcomes in the sample space Ω,

$$U(e) = \int_{\Omega} U(e, Y)p(Y|e)\,\mathrm{d}Y = \int_{\Theta}\int_{\Omega} U(e, Y)p_e(Y|\theta)p(\theta)\,\mathrm{d}\theta\,\mathrm{d}Y \quad (2)$$

This integral is with respect to $p(Y|e)$, the marginal distribution of the data under experiment e, which is obtained by integrating $p_e(Y|\theta)$, over possible *prior* values for θ, described by prior distribution $p(\theta)$. The Bayesian solution to the experimental design problem is provided by the experiment e^* which maximizes $U(e)$:

$$U(e^*) = \max_e \int_{\Omega} \max_d \int_{\Theta} U(d, \theta, e, Y) = p(\theta|y, e)p(y|e)\,\mathrm{d}\theta\,\mathrm{d}Y \quad (3)$$

This general formulation can be used to find optimal designs for a single experiment, and can be extended to optimal selection of a sequence of experiments and sequential decision making (Lindley 1972).

3. *Choice of Utility Functions*

It is important that utility functions be tailored to the goals of a given problem. Optimal designs for discriminating between two different models may be quite different from designs for prediction. In a one-way analysis of variance model, the best design for comparing k treatments to a control group, is not necessarily the optimal design for estimating the effects of $k+1$ treatments, as these experiments have different goals. While taking an equal number of observations in each of the $k+1$ treatment groups is a possibility, other arrangements may provide more information, particularly when data from previous experiments are taken into account. For example, several previous studies may be available using an existing therapy, but limited information may be available on a new treatment. Differential costs of treatment also need to be considered, which may lead to other allocations of sample sizes or choice of experiments. Ethical considerations may also be incorporated that may constrain the assignment of treatments (Kadane 1996).

Many of the papers in Bayesian design have based utility functions on Shannon information or quadratic loss. Motivation for these choices is discussed below. The reader is encouraged to refer to Chaloner and Verdinelli (1995) for additional details on these criteria, references, and relationships of other Bayesian utility functions to standard optimality criteria.

3.1 *Shannon Information*

Shannon information is appropriate for inference problems regarding θ or functions in θ, without specification of particular hypotheses. The expected utility function is based on the expected change in Shannon information or equivalently the Kullback-Leibler divergence between the posterior and prior distributions. As the prior distribution does not depend on the design, this simplifies to the expected Shannon information of the posterior distribution,

$$U(e) = \int_{\Theta}\int_{\Omega} \log\{p(\theta|U, e)\}p_e(y|\theta)p(\theta)\,\mathrm{d}Y\,\mathrm{d}\theta \quad (4)$$

thus the design goal is to find the design that maximizes the information provided by the experiment. In normal linear models with normal prior distributions, this leads to a criterion related to the well-known D-optimality from classical design,

$$U(e) \propto \log|(X_e^T X_e + R)/\sigma^2| \quad (5)$$

where X_e is the design matrix for experiment e and R/σ^2 is the prior precision matrix (inverse of the prior covariance matrix).

If prediction of future observations is important, the expected gain in Shannon information for a future observation Y_{n+1},

$$U(e) = \int \log(p(Y_{n+1}|Y, e)p(Y_{n+1}|Y, e)p(Y|e)\,\mathrm{d}Y\,\mathrm{d}Y_{n+1} \quad (6)$$

may be relevant. In normal linear models with prediction at a new point x_{n+1} this leads to

$$U(e) \propto \log[\{x_{n+1}^T(X_e^T X_e + R)^{-1}x_{n+1} + 1\}^{-1}/\sigma^2] \quad (7)$$

a Bayesian version of c-optimality.

3.2 *Quadratic Loss*

Point estimation based on quadratic loss leads to the expected utility function

$$U(e) = -\int_{\Theta}\int_{\Omega} (\theta + \hat{\theta})^{-T} A(\theta - \hat{\theta}) \quad (8)$$

where A is a symmetric non-negative definite matrix. The matrix A can be used to weight several different

estimation problems where interest may be in estimating individual components of θ or linear combinations of θ. Under the normal linear model and normal prior distribution this results in

$$U(e) = -\sigma^2 \mathrm{tr}\{A(X_e^T X_e + R)^{-1}\} \tag{9}$$

a Bayesian generalization of the A-optimality design criterion.

3.3 Other Utility and Loss Functions

Utility functions may be based on other loss functions besides quadratic loss. While a quadratic loss function may be appropriate in many cases, there are times when underestimation of a quantity incurs greater losses than overestimation. In such situations, asymmetric loss functions are more appropriate for design and inference (Clyde et al 1996). Discussion of other utility functions related to prediction, hypothesis testing and model discrimination and applications can be found in Chaloner and Verdinelli (1995).

3.4 Multiple Objectives

An experiment may often have several, possibly competing objectives which cannot be easily characterized by only one of the standard optimality criteria, and several design criteria may be appropriate. Weighted combinations of utility functions are still valid utility functions, so optimal designs for multiple objectives can be handled within the maximizing expected utility framework. For examples see Verdinelli (1992), Verdinelli and Kadane (1992), Clyde and Chaloner (1996). A difficulty with this approach is that utilities may be expressed in different scales which must be accounted for in the choice of weights. Equivalently, one can find the optimal experiment under one (primary) optimality criterion, subject to constraints on the minimal *efficiency* of the experiment under criteria for other objectives. These approaches are a potential way to address robustness of experimental designs under multiple objectives, models, and prior distributions.

4. Prior Distributions

Prior elicitation is an important step in designing Bayesian experiments, as well as analysis. Clyde et al (1996) use historical data from previous experiments to construct a hierarchical prior distribution to use for design of future experiments. Kadane (1996) considers many of the practical issues in subjective elicitation for clinical trials. Tsai and Chaloner (2001) describe a design problem where prior distributions are elicited from more than 50 clinical experts

While some researchers may agree to using prior information to help design an experiment, they may want to have their final conclusions stand on their own, such as in a frequentist analysis. One of the goals of the experiment may be to convince a skeptic, who has different beliefs from the designer of the experiment, that a treatment is effective. In such cases the prior distribution used in constructing the posterior distribution $p(\theta \mid Y, e)$ in the expected utility Eqn. (1) for the terminal decision problem may be different from the prior distribution $p(\theta)$ used in finding the best design in Eqn. (2). Etzioni and Kadane (1993) consider the problem of design when the designer and terminal decision maker have different prior beliefs corresponding to different decision rules.

5. Calculations

Calculation of expected utility, as in Eqns. (1) and (2), requires evaluation of potentially high-dimensional integrals, combined with difficult optimization problems, and in part, has limited the use of Bayesian optimal design in the practice. In normal linear models for many of the standard utility functions presented, the terminal decision rule can be solved in closed form and integrals can be computed analytically, leaving the optimization problem of finding the best experiment e^*. Except in special cases, numerical optimization must often be used to find the optimal design.

One can often relax the problem in the following way. Finding optimal 'exact' designs is often a difficult problem (similar to the traveling salesman problem). Designs can often be viewed in terms of a collection of support points (treatment levels) with weights that indicate the number of observations to be assigned at each support point. An exact design is one where the number of observations at each support point is an integer. A 'continuous' design is obtained by allowing the solution for the weights to be any real number. Rather than finding the optimal exact design, in the relaxed problem, the class of experiments is enlarged to include continuous designs. Mathematically, finding the optimal design in the continuous problem is easier to solve, and methods for checking optimality are often feasible. If the solution corresponds to an exact design, one can then show that it is the globally optimal design. While continuous designs cannot be used in practice, rounding of continuous designs provides exact solutions that are often close to optimal.

For nonlinear models, generalized linear models and other 'nonlinear' design problems (i.e., interest in nonlinear functions in an otherwise linear model), expected utility generally cannot be calculated in closed form and must be approximated. Many of the results for nonlinear design rely on asymptotic normal approximations in calculating expectations (Chaloner and Verdinelli 1995). Integrals may also be approxi-

mated by numerical quadrature, Laplace integration or Monte Carlo integration. With advances in computing resources, simulation-based optimal design is now an option, although methods often have to be designed for each specific application (see Müller 1999 for an overview of this area). As utility functions in many problems have been selected for their computational tractability, simulation-based design may soon open the way for greater use of scientific-based utility functions that better match the goals of the experiment. Hierarchical models are becoming increasingly important in modelling latent and random effects and accounting for subject-to-subject variability, for example. Advances in Bayesian computation, such as Markov chain Monte Carlo, now mean that inference in such models can be carried out in real problems. It is becoming easier to accommodate hierarchical and other complex models in the design of experiments with simulation-based optimal design schemes. Clyde et al (1996) compare analytic approximations and Monte Carlo-based design schemes for a hierarchical binary regression model using an asymmetric loss function.

6. *Applications*

Pilz (1991) covers Bayesian design and estimation in linear models, although from a rather mathematical viewpoint. Atkinson (1996) reviews both classical and Bayesian optimal design for linear and nonlinear designs, and presents recent applications such as design for clinical trials. Chaloner and Verdinelli (1995) provide a thorough review of literature on Bayesian designs and applications, which includes experimental designs for linear models such as regression and analysis of variance models, factorial and factional factorial experiments, variance component models, mixtures of linear models, hierarchical models, nonlinear regression models, binary regression models, designs for clinical trials and sequential experimentation. The article includes several worked-out examples, including design for one-way analysis of variance. The article by Clyde et al (1996) explores simulation-based design for a hierarchical logistic regression model, and illustrates how to construct prior distributions based on previous experiments. Other examples of simulation-based design appear in Müller (1999). For an example related to social sciences, see *Experimental Design: Large-scale Social Experimentation*.

The area of Bayesian design and analysis for clinical trails in both sequential and nonsequential designs is an exciting and active area, with many practical developments. For a survey of literature on sequential design, see *Sequential Statistical Methods*. The volume edited by Kadane (1996) describes a complete case study and discusses many of the critical issues in design and analysis of clinical trials, considering ethics, prior elicitation, randomization, treatment allocation, utilities, and decision making. Tsai and Chaloner (2001) describe the design of two large clinical trials, where prior distributions are based on eliciting prior opinions from over 50 clinicians. The design problem involves finding a sample size so that consensus of opinion is met with high probability, where consensus means that all clinicians would prescribe the same treatment based on their posterior opinions after the trial. Other recent examples include Rosner and Berry (1995), Simon and Freudman (1997).

While the increased use of Bayesian design in applications is especially encouraging, specialized software is often required. There is a growing need for reliable, user-friendly software for Bayesian design so that the methods can be more accessible and see greater applicability.

7. *Example*

The following design problem is concerned with choosing a sample size for an experiment to confirm results from an earlier study. All individuals in the first study had been diagnosed with breast cancer. Each individual had a biopsy of tumor tissue analyzed for expression levels or a protein implicated in progression of the disease; measured protein expression levels were scored between 0 and 8. All 135 individuals received a chemotherapeutic agent, with the goal of achieving pre-operative tumor regression. The measured outcome for each subject was a binary indicator of tumor regression. The researchers expected that the probability of tumor regression would increase as protein expression levels increased, and were surprised that the estimated probabilities reached a maximum around an expression level of 5, and then declined with higher levels. A logical regression model (see *Multivariate Analysis: Discrete Variables* (*Logistic Regression*)) was used to relate expression 1evels to clinical outcome (indicator of tumor regression). The linear predictor in the logistic regression model included a term with a parameter θ, such that for $\theta \geqslant 1$ the probability of tumor regression increases with expression level, while for $\theta < 1$ the probability of tumor regression increases up to a level of 5, then decreases. From a statistical perspective the evidence in favor of $\theta < 1$ was weak with a Bayes Factor of 1.65 ('not worth a bare mention' on Jeffreys' scale of interpretation for Bayes Factors (Kass and Raftery 1995). The goal of the second study is to investigate whether this decline was a chance occurrence. If the decrease is real, then this may suggest that a higher dose is necessary at higher expression levels, leading to other experiments.

Spezzaferri (1998) presented a utility function for model discrimination which can be used in designs for testing $H_0{:}\theta \geqslant 1$ versus $H_1{:}\theta < 1$. Spezzaferri's criterion can also incorporate parameter estimation, but

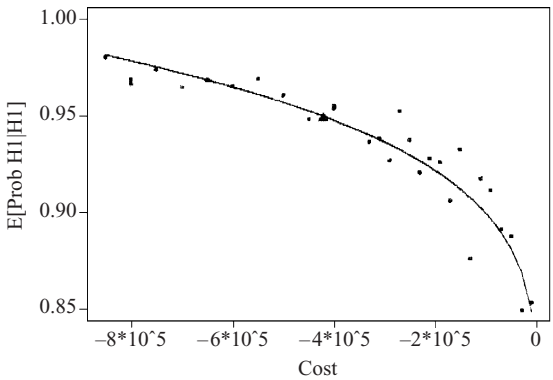

Figure 1
Smoothed expected utility for model discrimination vs. costs; the points indicate simulated expected utilities which exhibit Monte Carlo variation

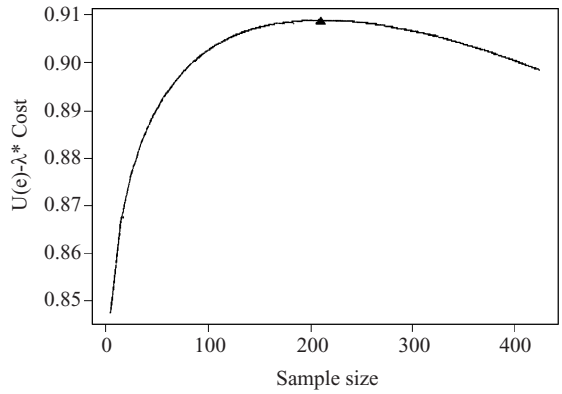

Figure 2
Combined utility for model discrimination with a penalty for costs; the triangle indicates the combined utility at the optimal design

for simplicity only model discrimination is considered. The expected utility function for model discrimination can be represented as

$$U(e) = \int_{\Omega} Pr(\theta < 1 \mid Y, e) p(Y \mid e, H_1) \mathrm{d}Y$$

where $p(Y \mid e, H_1)$, is the marginal distribution of the new data under $H_1: \theta < 1$.

The experimental design involves selecting the total sample size N_e, as individuals with specific protein expression levels cannot be selected in advance due to costs of obtaining the results. The expected utility versus negative costs ($-\$2000 \times N$) is illustrated in Fig. 1, and was constructed using smoothing of utilities generated using Monte Carlo experiments (Müller 1999). While analytic calculations are intractable, calculation of utilities for this model are straightforward via simulation. The marginal distribution for Y given an experiment with sample size N must incorporate averaging over different configurations of expression levels among the N subjects and the unknown parameters in the logistic regression model. Configurations of size N are generated using a multinomial distribution, with a probability vector π. The expression level distribution from the first study is used to construct a conjugate Dirichlet prior distribution for π. Data from the previous experiment is used to construct the prior distribution for θ and other parameters in the logistic model, and construct the predictive distribution of the new data given a vector of new expression levels. For each set of stimulated data under the new study $Pr(\theta < 1 \mid Y, e)$ is calculated, which incorporates both the old data (in the prior distribution) and the new data. For each sample size N, protein expression levels and outcomes are repeatedly generated from their predictive distributions, and the average $Pr(\theta < 1 \mid Y, e)$ over experiments e with the same sample size N_e provides an estimate of the expected utility for e. These averages were smoothed in order to eliminate Monte Carlo variation in their estimates and are plotted in Fig. 1.

For the above criterion, expected utility will increase with the total sample size, and as formulated the design problem has an infinite solution. Each point on the curve (Fig. 1) is a combination of $U(e)$ and cost. Expected utility has a maximum at 1, which is obtained as N goes to infinity (and with infinite costs). The experiment which minimizes costs is given by $N = 0$, where the posterior probability that $\theta < 1$ given the data from the previous study was 0.85. The researchers want to minimize total costs, but at the same time find a sample size such that the expected utility is above 0.95 (corresponding to a Bayes factor above 3, which would provide strong evidence in favor of H_1). Combining costs and model discrimination, the combined expected utility function can be written as

$$U(e) - \lambda(\$2000 * N_e) \tag{10}$$

where λ reflects the trade-off between for discriminating between the two hypotheses and costs of subjects. In general, it is difficult to specify these trade-offs explicitly, as the different components are often measured in different units. Verdinelli and Kadane (1992) illustrate choosing λ when there are trade-offs between two components in a utility function. If the expected utility for model discrimination is selected as 0.95, then the corresponding cost is \$421,327 (the point represented by the triangle in Fig. 1), and λ corresponds to the negative of the slope of the curve at this point ($\lambda = 0.0001961/\$2{,}000$). Figure 2 shows the combined utility given by Eqn. (10) using this value of λ. The maximum (marked by the triangle) corresponds

to an optimal sample size of 211. For utilities which combine more than two objectives, graphical solutions may not be feasible, but a constrained optimization approach can be used to determine the trade-off parameters (Clyde and Chaloner 1996).

8. Summary

Bayesian experimental design is a rapidly growing area of research, with many exciting recent developments in simulation-based design and a growing number of real applications, particularly in clinical trials. By incorporating prior information, the Bayesian approach can lead to more efficient use of resources with less costly, and more informative, designs. Utility functions can be explicitly tailored to the given problem and can address multiple objectives. Models, prior distributions, and utility functions have often been chosen to permit tractable calculations. However, with increased computing power, simulation-based methods for finding optimal designs allow for more realistic model, prior and utility specifications. Optimal designs have often been criticized because the number of support points in the design equals the number of parameters (in nonlinear Bayesian design this is not always the case), which does not permit checking the assumed model form. Model uncertainly is almost always an issue in inference, and certainly at the design stage. Bayesian model averaging has been extremely successful in accounting for model uncertainly in inference problems, and has great potential for use in Bayesian experimental design for constructing more robust designs.

Bibliography

Atkinson A-C 1996 The usefulness of optimum experimental designs (discussion pp. 95–111). *Journal of the Royal Statistical Society, Series B, Methodological* **58**: 59–76

Chaloner K, Verdinelli I 1995 Bayesian experimental design: A review. *Statistical Science* **10**: 273–304

Clyde M, Chaloner K 1996 The equivalence of constrained and weighted designs in multiple objective design problems. *Journal of the American Statistical Association* **91**: 1236–44

Clyde M, Müller P, Parmigiani G 1996 Inference and design strategies for a hierarchical logistic regression mode. In: Berry D, Stangl D (eds.) *Bayesian Biostatistics*. Marcel Dekker, New York, pp. 297–320

Etzioni R, Kadane J B 1993 Optimal experimental design for another's analysis. *Journal of the American Statistical Association* **88**: 1404–11

Kadane J B 1996 *Bayesian Methods and Ethics in a Clinical Trial Design*. Wiley, New York

Kass R E, Raftery A E 1995 Bayes factors. *Journal of the American Statistical Association* **90**: 773–95

Lindley D V 1972 *Bayesian Statistics—A Review*. SIAM, Philadelphia, PA

Müller P 1999 Simulation based optimal design (with discussion). In: Berger J, Bernardo J, Dawid A, Smith A (eds.) *Bayesian Statistics 6*. Clarendon Press, Oxford, UK, pp. 459–74

Pilz J 1991 *Bayesian Estimation and Experimental Design in Linear Regression Models*. Wiley, New York

Rosner G L, Berry D A 1995 A Bayesian group sequential design for a multiple arm randomized clinical trial. *Statistics in Medicine* **14**: 381–94

Simon R, Freedman L S 1997 Bayesian design and analysis of 2×2 factorial clinical trials. *Biometrics* **53**: 456–64

Spezzaferri F 1988 Nonsequential design for model discrimination and parameter estimation. In: Bernardo J, DeGroot M, Lindley D, Smith A (eds.) *Bayesian Statistics 3*. Clarendon Press, Oxford, UK, pp. 777–83

Tsai C-P, Chaloner K 2001 Using prior opinions to examine sample size in two clinical trials. In: Gatsonis C, Kass R E, Carlin B, Carriquiry A, Gelman A, Verdinelli I, West M (eds.) *Case Studies in Bayesian Statistics*. Springer-Verlag, New York, Vol. V

Verdinelli I 1992 Advances in Bayesian experimental design (with discussion). In: Bernardo J, Berger J, Dawid A, Smith A (eds.) *Bayesian Statistics 4*, Clarendon Press, Oxford, UK, pp. 467–81

Verdinelli I, Kadane J B 1992 Bayesian designs for maximizing information and outcome. *Journal of the American Statistical Association* **87**: 510–15

M. A. Clyde

Experimental Design: Compliance

Subject noncompliance with, or nonadherence to assigned treatment in randomized experiments designed to evaluate the relative efficacy of two or more treatments or interventions may seriously jeopardize the interpretation of results from such studies. Essentially, the noncompliant subject self-selects his or her treatment and/or treatment level, thereby disturbing the covariate balance (Rosenbaum 1995, pp. 14–15) and accompanying safeguard against confounding factors afforded by randomization (see *Experimental Design: Randomization and Social Experiments*; *Experimental Design: Large-scale Social Experimentation*). Attributing subject outcomes to the treatments under study when subjects are noncompliant is stymied by confounding factors introduced by that noncompliance. For example, in a study to evaluate the relative efficacy of mental health interventions on the quality of life for breast cancer patients (Helgeson et al. 1999), women were asked to attend eight sessions of either an informational seminar or peer support group during the course of their adjuvant chemotherapy and/or radiation treatments. In this case, women's lowered attendance at their assigned interventions may have been related to how well they were tolerating their drug or radiation

treatments, which at the same time may have affected their quality-of-life outcomes, so that the effects of the interventions are obscured.

1. Difficulties with Noncompliance

The complication of noncompliance necessitates the clear distinction between treatment *assignment* and treatment *exposure*. In randomized experiments with perfect compliance, these are the same; but when some subjects are noncompliant, assignment and exposure are different for those noncompliers. This distinction represents the crux of the problem induced by noncompliance: if a study is analyzed without regard to differences in assignment and exposure (e.g., using an intention-to-treat analysis) then the research question that is answered may not be the one that is posed (see *Causal Inference and Statistical Fallacies*).

Boruch (1997) takes a strong stand for analyzing studies as they are randomized (i.e., without accounting for noncompliance), since compliance behavior during a randomized trial is likely to approximate that same behavior in the general population were the interventions under study to be applied more broadly. Indeed, substantial noncompliance in itself may speak to the efficacy of assignment to the interventions under study. Nevertheless, understanding the relative efficacy of exposure to interventions and possible relationships between compliance behavior and subject outcome warrants a deeper examination of compliance.

Consideration of the statistical problems induced by subject noncompliance has taken place primarily in the field of clinical trials in cases where subjects take less of their assigned drug treatment and/or supplement their treatments with medications or activities other than the ones under study. A direct translation of the compliance problem from the medical setting to the behavioral or education setting comes from considering attendance at sessions of experimental interventions as a 'dose' or 'treatment level.' Alternatively, non-compliance with behavioral interventions might be as simple as subjects not taking advantage of an offered program, such as a family's decision to keep their child in public school even after they have been selected by lottery to attend a private school (see Barnard et al. 1999).

Regardless of the manifestation of noncompliance, questions regarding causal effects of interventions under study remain. Holland (1986) provides an extensive overview of causal inference in statistics, making use of the counterfactual model, the seeds of which date at least to Fisher (1918) and Neyman (1923), and which has been refined over several years by Rubin (1978, 1980; see *Causal Inference and Statistical Fallacies*). This model, relating exposure and outcome, is summarized here to provide a framework for illustrating the noncompliance problem and for describing methods for describing existing methods for addressing that problem (Dawid 2000 argues for causal modeling without using counterfactual outcomes, though addressing noncompliance is greatly facilitated by their use; see discussions following this article).

2. The Counterfactual Model

The foundation of the counterfactual model is the notion of potential outcomes, which are the collection of outcomes that would be observed for a subject if that subject were exposed to all possible treatments. The words counterfactual and potential suggest quantities that are not observed—outcomes 'counter to the fact,' or that 'might be observed.' Consider N subjects, and let T_i denote the treatment assignment and X_i the treatment exposure for subject i. For simplicity, let T_i and X_i assume binary values, 0 and 1, representing a control and active treatment, respectively.

The specific quantities of interest for causal inference are subject outcomes under each treatment: $Y_i(0)$ is the outcome that would be observed for subject i if he or she were exposed to treatment 0, and $Y_i(1)$ is the outcome that would be observed for the same subject if he or she were instead exposed to treatment 1. Typically, only one of $Y_i(0)$ or $Y_i(1)$ is observed for each subject. The observed outcome for subject i, Y_i (or $Y_i(X_i)$), is linked to the counterfactuals via the consistency relation

$$Y_i = Y_i(1)X_i + Y_i(0)[1 - X_i], \tag{1}$$

that is, the observed outcome *is* the potential outcome under the observed treatment exposure.

It should be noted that $Y_i(0)$ and $Y_i(1)$ are reduced from their general form as functions of the entire vector of treatment exposures via the *stable unit treatment value assumption* (*SUTVA*) of Rubin (1980). This assumption asserts that each subject's potential outcomes depend only upon his or her own treatment exposure and not those of any other subjects. In clinical trial settings, where subjects typically do not interact, SUTVA may be a reasonable assumption. In studies of behavioral or educational interventions, however, where subjects may interact with each other extensively, this assumption must be examined more carefully, and the causal conclusions that are based on it may be tenuous (Gitelman 1999).

The potential outcomes just defined, and SUTVA, are sufficient for defining an individual causal effect (ICE). This is simply the effect, for subject i, of exposure to treatment 1 vs. exposure to treatment 0

$$Y_i(1) - Y_i(0). \tag{2}$$

Alternative individual causal effects such as the ratio of these individual potential outcomes or the difference

of their logs could be considered, though the causal parameter obtained by taking the expectation of the difference in Eqn. (2) argues for using this simpler form. The average causal effect (ACE) is the individual causal effect averaged over the space of subjects

$$\theta = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]. \quad (3)$$

The probability distribution over which these expectations are taken is induced by the process by which subjects enter, or are selected for, the study.

In general, θ is nonidentifiable without further assumptions. Rubin (1978) showed, for instance, that θ is identifiable when treatment exposure is 'ignorable,' meaning that each subject's treatment exposure is stochastically independent of his or her collection of potential outcomes (this independence is expressed using the symbol $\perp\!\!\!\perp$, following Dawid (1979)

$$\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp X_i. \quad (4)$$

Under randomized treatment assignment and complete compliance, treatment exposure ignorability holds, since $X_i \equiv T_i$ for all subjects i. In this case

$$\begin{aligned}\theta &= E[Y(1)] - E[Y(0)] \\ &= E[Y(1)|X = 1] - E[Y(0)|X = 0] \\ &= E[Y|X = 1] - E[Y|X = 0] \\ &= \alpha \end{aligned} \quad (5)$$

where the second equality follows from ignorability and the third from the consistency relation in Eqn. (1). α is called the apparent treatment effect, and it can be estimated by the difference in sample means of the group of subjects exposed to each treatment, perhaps appropriately scaled by a standard error estimate. Notice that α is typically identifiable regardless of treatment exposure ignorability—it is its equivalence to θ that is afforded by ignorability.

If treatments are randomly assigned, but some subjects are noncompliant, then treatment exposure is probably *not* ignorable. In that case, the apparent treatment effect is not a good estimate of the average causal effect. Specifically, there may be some characteristic of the subject (e.g., in a study comparing educational interventions, his or her academic ability) that predisposes him or her to be noncompliant with treatment assignment. This very characteristic may also be reflected in his or her potential outcomes under that treatment, and therefore his or her outcome cannot be attributed solely to the treatment in question.

One option when faced with noncompliance is to estimate the *intention-to-treat* effect, and use it as an estimate of the average causal effect. The intention-to-treat effect is given by

$$\alpha_{\text{itt}} = E[Y|T = 1] - E[Y|T = 0] \quad (6)$$

which can be estimated by the difference in sample means of the two treatment groups. This effect explores the relationship between assignment and outcome. Alternatives to this approach that seek to illuminate the relationship between exposure and outcome are described next.

3. Simple Noncompliance

In the case of binary treatment assignment and binary treatment exposure, noncompliance with treatment assignment manifests as 'treatment switching.' That is, a subject assigned to the active treatment is noncompliant if he or she switches to the inactive treatment (for instance by not taking the treatment at all). Similarly, a subject assigned to the control treatment is noncompliant by switching to the active treatment. This is essentially the situation described in Barnard et al. (1999), wherein families participated in a lottery to receive funding and other support for private school attendance. Some families, even though they 'won' the lottery, decided to keep their children in public schools—hence, these families are the 'noncompliers.'

3.1 Modification of the Counterfactual Model

The counterfactual outcomes described previously depend upon treatment exposure only, whereas to account for noncompliance they must be altered to depend also on treatment assignment. Following Angrist et al. (1996), let $X_i(0)$ and $X_i(1)$ denote potential or counterfactual exposures for subject i; these are the treatments subject i would be exposed to (i.e., based on his or her compliance behavior) under assignment to treatments 0 and 1, respectively. Let $Y_i\ (T_i,\ X_i(T_i))$ denote the potential outcome of subject i, under treatment assignment T_i and exposure $X_i(T_i)$. Expression of both the potential exposures and potential outcomes again rely on Rubin's SUTVA. In the binary assignment and binary exposure case, each subject has four potential outcomes, one for each assignment/exposure pair

$$\{Y_i(0,0),\ Y_i(0,1),\ Y_i(1,0),\ Y_i(1,1)\} \quad (7)$$

For instance, $Y_i(1, 0)$ is the outcome subject i would experience had he or she been assigned to the active treatment and been noncompliant with that treatment, thereby receiving the control treatment.

As in the case of the simpler potential outcomes $Y_i(0)$ and $Y_i(1)$, the outcomes in Eqn. (7) are related to the observed outcome via a consistency relation

$$Y_i = \sum_{t=0}^{1} \sum_{x=0}^{1} Y_i(t, x) I_{\{T_i = t,\ X_i = x\}} \quad (8)$$

Where $I_{\{\cdot\}}$ is the indicator function. As before, the observed outcome is the potential outcome that corresponds to the observed treatment assignment/exposure pair. Nonidentifiability is more drastic in this setting, however, as there are now four counterfactual outcomes to consider for each subject instead of just two.

The causal parameters of interest are the expected values (with expectations taking over the space of subjects) of the potential outcomes in (7); namely

$$E[Y(0,0)],\ E[(Y(0,1)],\ E[Y(1,0)],\quad \text{and } E[Y(1,1)]$$

With noncompliance, exposure is not ignorable, and so it is not necessarily the case that $E[Y(0, 0)]$ equals $E[Y(1, 0)]$ or that $E[Y(0, 1)]$ equals $E[Y(1, 1)]$. That is, the average subject outcome of a noncompliant subject assigned to the treatment arm will not, in general, be the same as that of a *compliant* subject assigned to the control arm. Methods to address this complication are described next.

3.2 Bounds for ACE Using an Instrumental Variable

A variable T is an *instrumental variable* if the distribution of another variable, Y, depends on T only through a third variable, X. Instrumental variables have been popularized in the econometrics literature (see *Instrumental Variables in Statistics and Econometrics*); they have been used to address noncompliance in other settings (e.g., Angrist et al. 1996). Ideally, T and X should be correlated, and T and Y uncorrelated conditionally on X. Then T can be used as a surrogate or *instrument* for X. An instrumental variable, then, is a variable that is partly defined by a conditional independence relationship, known as the *exclusion restriction*, or *instrumental variable assumption*

$$T \perp\!\!\!\perp Y|X \tag{9}$$

and partly by the required correlation between T and X. Angrist et al. (1996) translate this assumption into the language of potential outcomes, using assignment as an instrument for exposure

$$Y_i(T_i, X_i(T_i)) = Y_i(X_i(T_i)) \tag{10}$$

that is, a subject's potential outcomes do not depend on the treatment assignment directly, but only through that assignment's effect on the treatment exposure. Angrist et al. (1996) show the equivalence of the counterfactual and structural (or simultaneous) equation models (see *Simultaneous Equation Estimation: Overview*).

The instrumental variable assumption in Eqn. (10) has the effect of reducing the number of potential outcomes from four, as in Eqn. (7) to two, simply $Y_i(0)$ and $Y_i(1)$. For each subject, however, each of $Y_i(0)$ and $Y_i(1)$ can be observed in one of two ways, either by the subject's compliance with assigned treatment or their noncompliance with the opposing treatment. By using assignment as an instrument for exposure the average causal effect, θ can be bounded by taking minimums or maximums over these two ways of observing each treatment exposure (of course, these bounds also depend upon the responses having bounded expectations). Balke and Pearl (1994) produce the tightest possible bounds for θ (under binary assignment, binary exposure, and binary outcome) using the instrumental variable assumption and linear programming results.

Angrist et al. (1996) use the instrumental variable assumption, combined with a monotonicity assumption, to estimate θ rather than bound it. This monotonicity assumption applies to the counterfactual exposures—it is assumed that a subject will only be noncompliant by switching from the active treatment to the inactive control, but not vice versa, whereby $Y_i(1, 0)$ is eliminated from the collection in Eqn. (7). Because it relates exposure under an active treatment to that under a control treatment, this monotonicity assumption only is applicable in trials that compare an active treatment (or treatments) to a control.

4. Noncompliance as Underdosing

A more complicated noncompliance situation occurs when noncompliant subjects take a lower-than-intended dose of their assigned treatment (or attend fewer-than-intended sessions of their assigned intervention), rather than switching between treatments. For simplicity, treatment assignment remains binary in this discussion. The counterfactual exposures, $X_i(0)$ and $X_i(1)$, now take categorical or even continuous values, however, and there is no switching between treatments except perhaps in the case of a subject taking a zero dose of an active treatment. In a randomized dose-response experiment, subjects are randomly assigned to different doses of a treatment (i.e., dose is ignorable); but with noncompliance the subjects *choose* those doses (i.e., compliance 1evels), and so the resulting treatment outcomes cannot be attributed to dose or compliance levels alone.

In a direct extension of the simple compliance case discussed previously, Goetghebeur and Molenberghs (1996) model counterfactual compliance-response curves (similar to dose-response curves) for categorical exposure levels and binary outcomes, Their model relies on a different monotonicity assumption (this one asserts that increased compliance cannot lead to decreased probability of a successful outcome); an instrumental variable type of assumption and a constant global odds ratio assumption. A subsequent contribution, Goetghebeur et al. (1998), relaxes this

last assumption somewhat. An essential difficulty with these models, however, is that due to the need for modeling counterfactual (i.e., unobservable) outcomes, rather strong, untestable assumptions are needed. For example, the constant global odds ratio assumption is untestable, and it dictates an artificial relationship between the counterfactual compliance-response curves.

Efron and Feldman (1991) take the approach of modeling compliance behavior as a fixed attribute of the subject, rather than as a response to treatment assignment. As a fixed attribute, each subject's compliance behavior remains the same, regardless of the treatment to which he or she is assigned—this eliminates the problem of having to address counterfactual exposure levels. Compliance-response curves are estimated with this approach, and with additional modeling assumptions (e.g., linearity of the dose-response relationship), portions of the 'true' dose-response curves are recovered, Again, the true dose-response curves can be fully estimated only if dose level is ignorable.

Robins (1994) explores the slightly different setting of sequential randomized trials, in which subjects are assigned to treatment regimes in series, interspersed with assessment of their medical condition. In this work it is assumed that, conditional on all observable covariates, the treatment exposure regime is ignorable.

5. *Addressing Noncompliance in Practice*

When subjects are noncompliant with assignment to experimental interventions, a study loses the solid footing afforded by randomized treatment assignments. The essential question of interest regarding the effect of exposure to interventions is now compromised by the subject's selection of their intervention and/or intervention level. The counterfactual model provides a foundation for examining the relationship between exposure and outcome, though careful consideration of the assumptions of both that model and the additional modeling components needed for identifying causal parameters of interest are crucial and may depend upon the type of study.

When faced with compliance problems, or the possibility of compliance problems, the researcher may be best served by careful planning of intervention delivery with an eye toward averting attendance problems or other lapses in the integrity of the intended treatment (Meyer and Fienberg 1992). Furthermore, by collecting as many relevant covariates as possible, differential, subject-selected treatment exposure can be alleviated, in part by adjusting for these covariates using, for example, propensity score matching (Rosenbaum and Rubin 1983). If sufficient subject covariates are observed, it may be reasonable to assume that treatment exposure level is ignorable conditional on these covariates, whereby the causal parameters of interest can be identified. For compliance problems that cannot be addressed *a priori*, however, the bounding (e.g., Balke and Pearl 1994) and modeling (e.g., Angrist et al. 1996, Barnard et al. 1999) methods described herein provide useful options.

Bibliography

Angrist J D, Imbens G W, Rubin D B 1996 Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**: 444–72

Barnard J, Frangakis C, Hill J, Rubin D 1999 School choice in NY City: A Bayesian analysis of an imperfect randomized experiment. In: Kass R E (ed.) *Proceedings of the Fifth Case Studies in Bayesian Statistics Workshop*, Vol. 5. Springer-Verlag, New York

Boruch R F 1997 *Randomized Experiments for Planning and Evaluation, A Practical Guide*. Sage, Thousand Oaks, CA

Balke A, Pearl J 1994 Counterfactual probabilities, computational methods, bounds and applications. In: Lopez de Mantaras R, Poole D (eds.) *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Mateo, CA

Dawid A P 1979 Conditional independence in statistical theory. *Journal of the Royal Statistical Society*, *Series B* **41**: 1–31

Dawid A P 2000 Causal inference without counterfactuals. *Journal of the American Statistical Association* **95**: 407–24

Efron B, Feldman D 1991 Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association* **86**: 9–26

Fisher R A 1918 The causes of human variability. *Eugenics Review* **10**: 213–20

Gitelman A I (1999) Treatment integrity concerns in comparative education studies. Ph.D. thesis, Carnegie Mellon University

Goetghebeur E, Molenberghs G 1996 Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association* **91**: 928–34

Goetghebeur E, Molenberghs G, Katz J 1998 Estimating the causal effect of compliance on binary outcome in randomized controlled trials. *Statistics in Medicine* **17**(3): 341–56

Helgeson V S, Cohen S, Schulz R, Yasko J 1999 Education and peer discussion group interventions and adjustment to breast cancer. *Archives of General Psychiatry* **56**: 340–47

Holland P W 1986 Statistics and causal inference. *Journal of the American Statistical Association* **8**: 945–70

Meyer M M, Fienberg S E 1992 *The Case of Bilingual Education Strategies*. National Academy Press, Washington, DC

Neyman J 1923 On the application of probability theory to agricultural experiments. Essay on principles. *Koczniki Nauk Roiniczych*, *Tom* **X**: 1–51. [Reprinted in *Statistical Science* **5**: 463–80]

Robins J M 1994 Correcting for noncompliance in randomized trials using structural nested mean models. *Communications in Statistics A* **23**: 2379–412

Rosenbaum P R 1995 *Observational Studies*. Springer, New York

Rosenbaum P R, Rubin D B 1983 The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55
Rubin D 1974 Estimating causal effects of treatment in randomized and non-randomized studies. *Journal of Experimental Psychology* **66**: 688–701
Rubin D 1977 Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics* **21**: 1–26
Rubin D B 1978 Bayesian inference causal effects: The role of randomization. *The Annals of Statistics* **6**: 34–58
Rubin D B 1980 Discussion of 'Randomization analysis of experimental data: The Fisher randomization test,' by D. Basu. *Journal of the American Statistical Association* **75**: 591–93

A. I. Gitelman

Experimental Design: Large-scale Social Experimentation

An experiment is a deliberately planned process to collect data that enable causal inferences and legitimize treatment comparisons. Random assignments of subjects to treatments (including control groups) allow statistically significant differences among treatment outcomes to be attributed solely to the treatment differences. Without strong additional assumptions, observational studies and quasi-experiments cannot justify drawing causal inferences, because only experiments, by definition, assign the treatments randomly (see *Experimental Design: Quasi-experimental Designs*; *Experimental Design: Randomization and Social Experiments*; *Experimental Design: Compliance*).

Public welfare relies on experiments for many purposes, including to approve medicines, medical procedures, new foods, and product reliability. Large-scale social experiments differ from most of these experiments, even from many that involve human subjects, in that the treatments in social experiments are social programs, the subjects are humans, and the responses measure outcomes for these people (or families) as they make decisions under different possible programs.

Large-scale public policy experiments in the United States generally are considered to have started with the New Jersey Negative Income Tax Experiment 1968. Orr (1999) puts the number of social experiments (not all large scale) launched per decade in the US at six in the 1960s, at 49 and 84 in the next two decades, and then at 56 for 1991–95.

These experiments were fielded to gather scientific information so policy makers could know as accurately as possible the costs and benefits of possible social programs. Most social experiments have been initiated and funded by agencies (national, state, or at other political levels) concerned with the welfare of large populations. While these large experiments were costly and time consuming, they found sponsors because their costs were dwarfed by the costs of actual programs.

1. Some Examples

Much has been written about social experiments. Greenberg and Shroder (1997) summarize the duration, cost, treatments, measured outcomes, sample sizes, design, funding sources, developers, evaluators, and results of over 100 of them. Orr (1999), Fienberg et al. (1985), and others in the references here, review some of the main studies and list further sources. This article describes several of the better-known public policy social experiments, each of which monitored human subjects for several years as they made personal or family economic decisions while eligible for benefits provided by an experimental policy to which they were randomly assigned.

1.1 The New Jersey Negative Income Tax Experiment (NJ–NIT)

The NJ–NIT experiment (Fienberg et al. 1985, Greenberg and Shroder 1997, Orr 1999, Hausman and Wise 1985) was inspired by the thesis of Ross (1966) while she was a PhD student in economics at MIT. Ross argued that only a social experiment could resolve disputes over how work patterns of the poor might change if the government were to adopt a Negative Income Tax (NIT) program to supplement their incomes. The Office of Economic Opportunity in the Department of Health, Education, and Welfare (HEW) funded the initial Income Maintenance Experiment (IME) to begin in 1968. Mathematica Policy Research, commissioned to design and field it, randomized a total of 725 experimental households into nine treatment groups, and another 632 households into a control group, with welfare and near welfare households in four New Jersey cities being enrolled for three years each. The nine experimental treatments in this response surface design involved two main variables, one providing a range of income guarantees, and the other providing a range of tax rate reductions to families as incentives to earn additional income. The NJ–NIT objectives were to estimate, mainly via regression analyses, how work hours, rates of high school graduation, and other outcomes depended on these factors.

Difficulties encountered in the NJ–IME included significant nonstationary participant behavior at the outset caused by participants needing time to become familiar with their program incentives, and toward the end caused by some gaming behaviors. The NIT design deliberately correlated the random treatment assignments with family incomes in order to reduce experimental costs. That strategy complicated the

analyses and risked compromising the experiment's randomization. Other design problems included differences between treatment and control participants in their initial income reporting, and using of total income as an eligibility criterion at enrollment with a resulting underrepresentation of working wives.

1.2 The Health Insurance Experiment (HIE)

The need for a health insurance experiment was recognized during the Nixon administration when for the first time the government was considering possible forms of a national health insurance program. Congressional leaders held widely divergent views about possible costs of a national program, and especially about the elasticity of demand for health care. If demand were completely inelastic, previously insured individuals would continue to purchase medical services at the same rate. That would make government sponsorship more affordable. While no experiment could provide all needed information, policymakers funded the HIE because it would narrow the range of disagreements, and help avoid disastrous errors in a national program.

The HIE was fielded in 1974 as the income maintenance experiments wound down. The Office of Economic Opportunity, and later HEW, asked the Rand Corporation to design and conduct the HIE, ultimately detailed in 'Free for All' (Newhouse 1993) by its principal investigator.

The HIE was designed to assess the demand for healthcare and many other questions, including whether health benefits might derive from a national health insurance program. Nearly 3,000 families, each for three to five years, were randomized to one of 14 insurance plans, including an HMO group. Insurance plans had two main dimensions, with varying coinsurance rates that families faced, and with varying deductible limits (after reaching the limits families were exempted from any further expenses for the years remaining). Anticipating that the world might differ considerably after the results of an eight year experiment became available, the HIE treatments included no particular proposal, but instead were chosen to provide a range of rates and benefits that could be used reliably to extrapolate HIE results to future legislative proposals.

Demand in the HIE was ultimately found to be elastic. Perhaps primary among HIE findings have been the elasticity estimates that legislators still use when comparing the costs of new health insurance proposals (Newhouse 1993, Orr 1999).

1.3 JOBSTART, a Demonstration

The rate of new social experiments continues to increase as the technology for doing them matures, and as new research groups and firms develop the skills to design, conduct, and analyze them. The newer experiments tend to be smaller and more focused than the first ones, many having been designed to demonstrate the effectiveness of an existing program, and to identify areas for improvement. These 'demonstration' experiments are simpler to field than those, like the IME and the HIE, that are designed to provide information on a wide range of unspecified programs.

Perhaps politicians are partial to demonstrations because of their lower costs and because they focus on concrete legislation. Demonstration experiments cost less because some of the participant funding comes through an existing program's benefits, and because focus is exclusively on one treatment and a control group. A complication is that control group subjects, if matched to treatment subjects, will be eligible for the same program benefits and so must be allowed to access them. One way to counter this is to have the treatment to be an encouragment for eligible subjects to take advantage of an underused program.

JOBSTART was a demonstration experiment to evaluate the Job Training Partnership Act (JTPA) of 1982 (Greenberg and Shroder 1997, Cave et al. 1993). Sponsored jointly by the US Department of Labor and by several private foundations, JOBSTART was designed and evaluated by Manpower Demonstration Research Center (MDRC) between 1985 and 1992, and administered by the Local Service Delivery Areas in 13 geographically disparate US sites. Subjects were randomized into treatment and control groups, with about 1,150 subjects in each. All subjects were JTPA eligible (a national program), being economically disadvantaged, high school dropouts aged between 17 and 21 years, and all were poor readers. Treatment interventions included two summers of vocational and educational training, support services, job placement assistance, and incentive payments. Subjects were monitored for four years.

JOBSTART found that the treatment group had a dramatically higher rate of high school completion. Women who had no children before the demonstration, but who gave birth afterward, were less likely to receive AFDC payments. The opposite result was found for women who started with children. No major long-term differences were found for employment, earnings, or welfare measures.

1.4 The Housing Assistance Supply Experiment (HASE)

Economic theory predicts an elastic supply response to increased demand, so that benefits received by individuals do not necessarily rise in proportion to the dollars provided. Even so, almost all large-scale social experiments have measured demand only, mainly because a supply response is extremely difficult to

produce in an experiment. HASE is a notable exception (Lowry 1983). While Housing Allowance Demand Experiments had been fielded to evaluate new housing policies (Bradbury and Downs 1981), evaluating the supply response to alternative housing programs requires a long-term saturation experiment in which all eligible members of an entire community participate in an experimental program. A supply experiment takes longer because the supply side needs adequate time to respond to new demand, generally much longer than individuals need to respond to new opportunities.

The Department of Housing and Urban Development sponsored HASE to learn how the supply side would respond if individuals and families received increased government assistance for making home improvements. HASE ran from 1972 to 1982 in Green Bay, Wisconsin and in South Bend, Indiana, selected as two small, stable communities in which a meaningful supply response might be stimulated. Long-term benefits were guaranteed to all eligible families for 15 years, well beyond the length of the experiment, to give the construction community sufficient incentive to relocate workers and businesses into these regions.

Supply experiments are extremely ambitious, but the questions they address are crucial.

2. *Alternatives to Social Experiments*

Less expensive alternatives must be considered before undertaking a social experiment. These include expert (and nonexpert) opinions, theoretical models, surveys, observational studies, quasiexperiments (see *Experimental Design: Quasiexperimental Designs*), and, if they exist, appropriate 'natural experiments.' Because these alternatives usually are cheaper, faster, and easier than an experiment, they will be preferred whenever they can provide valid predictions. Even if they cannot produce valid predictions, their consideration is vital to designing an efficient social experiment.

These alternatives all have drawbacks. Expert opinions and theoretical models are no better than the experts or models. While surveys can sample randomly from a target population, they cannot assign treatments randomly, and surveyed subjects rarely would know how they would respond to a hypothetical future program. Nonexperimental data, quasiexperiments, and observational studies (both synonyms for nonrandomized studies) might be used to extrapolate from data on existing programs to predict outcomes of untried programs. Such predictions, however, are especially unreliable for major program changes, and self-selection in such data may bias results. Natural experiments might occur if a comparable country or region adopts a policy similar to one under consideration, but they usually do not exist, and if one does occur, results still must be extrapolated to another country or population.

As none of these alternatives has randomly assigned treatments, selection bias will be present and causal inferences will be suspect. Drivers who use seat belts have lower traffic fatality rates, but those rates are confounded with the same drivers being generally safer drivers. Similarly, individuals who choose health insurance plans with less generous coinsurance rates generally consume fewer health services. Is that only because some of them, knowing they are healthier, choose less generous insurance plans? If so, their behavior in a future health insurance program that includes everyone will not be predictable from their past utilizations. A health insurance experiment offers an alternative option to using observational data for predicting the costs of proposed national health insurance programs.

3. *Risks of Social Experiments*

Being costly and time-consuming, social experiments have been used sparingly. They might not provide useful information for many years, risking their irrelevance. Human subjects can refuse to enroll, and enrollees can drop out before completion. If that happens with regularity, randomization fails, and so the experiment fails (see *Errors: Nonsampling Errors*).

If large experiments become too visible, their subjects may feel that the situation is artificial and act differently.

Drastic political, social, legislative, or economic changes can occur during an experiment and invalidate it. However, a control group can be protective in such situations, to the extent that the shock affects all treatments and controls equally. Orr (1999), who as a government economist helped spawn and plan several early public policy experiments, reviews such concerns about social experiments, and he discusses how experiments can protect themselves. Besides the issues mentioned here, he discusses whether and how experiments have affected policy, their credibility, time considerations, communication issues, their generalizability, relevance of their results, the policy environment, and policy links.

4. *Design Issues*

Those who design social experiments make numerous crucial choices. They must choose treatments that span the policies that are likely to be considered without picking an overly large range that provides too little information on the policies that ultimately matter. How many sites are needed? Too many sites are difficult to manage, and for a fixed budget, increased

management costs must be paid for by decreased sample sizes. Too few sites restricts generalizability. Should sites be chosen randomly? Probably not if the number of sites must be kept small, because then randomization provides little basis for inference.

More subjects are better, but too many subjects make an experiment unfeasibly expensive. Without strong reasons to do otherwise, it helps to keep the percentage of subjects allocated to each treatment the same in every site. Balanced samples, which means matching the key characteristics of the sample across sites and treatments, have various optimal properties in experimental design, and they enjoy considerable face validity (Morris and Hill 2000).

Enrollment participation must be long enough for individuals to learn how to take advantage of experimental programs (and so to reach steady state), and perhaps to discount the times at the beginning and/or the end of the experiment when subjects may act differently than they would in a continuing government program. However, overly long enrollments waste time because early results are better for policy, and because, assuming positive correlations within individuals, one individual provides less information per unit time than two independent individuals. These transitory issues may make it necessary to field a longer experiment, or at least to guarantee treatment benefits beyond the actual measurement period.

Decisions about what data to collect, about the design of questionnaires, and about how much and how frequently to interview subjects may be as crucial as other statistical design decisions. Interviewing subjects too often risks overstimulating them ('Hawthorne effects'), so that subjects become overly aware of participating in an experiment and behave artificially. Respondent burden can cause dropouts, careless responses, and selective nonresponse. Interviewing subjects too infrequently sacrifices potentially important data. Experimenters can give a better understanding of respondent burden by subjecting themselves to the same interviewing processes which experimental subjects face.

While large-scale social experiments encounter difficulties beyond those of smaller experiments, they will have greater financial resources to deal with them. Some potential difficulties that concerned HIE designers included: creating Hawthorne effects by frequent interviewing; activating participants by obtaining a medical screening examination at enrollment; stimulating health expenditures by having to pay participation incentives to families; and not being sure of the best time horizon (Newhouse et al. 1979). The HIE budget allowed for four (balanced) 'subexperiments' within the main experiment to measure these effects. In one subexperiment, half of the HIE subjects were chosen at random for weekly interviews, and half for bi-weekly. Similarly, 60 percent of the HIE subjects were randomly selected for initial screening exams, and the rest not; and 70 percent were randomly assigned for three-year enrollments, and the other 30 percent for five years. This allowed testing for these effects and, if found, adjusting for them.

5. *Further Reading*

Orr (1999) provides an overview and first-hand examples of the design and implementation of social experiments. Greenberg and Shroder (1997) summarize 143 completed social experiments in the United States, and 75 others that were ongoing at the time of their publication. Each summary describes the target population, policies tested, experimental design, sites, findings, sources of further information, and public access to the data. Campbell's fundamental work in social experimentation is summarized in Campbell and Russo (1999). Boruch (1997) provides reference material on randomization in field experiments and Neuberg (1989) discusses anomalies of social control experimentation.

See also: Sample Surveys: The Field

Bibliography

Aigner D J, Morris C (eds.) 1979 Experimental design in econometrics. *Journal of Econometrics* **11**(1)

Boruch R F 1975 On common contentions about randomized field experiments. In: Boruch R F, Riecken H W (eds.) *Experimental Tests of Public Policy*. Westview, Boulder, CO, pp. 108–45

Boruch R F 1997 *Randomized Experiments for Planning and Evaluation: A Practical Guide*. Sage, Thousand Oaks, CA

Bradbury K, Downs A (eds.) 1981 *Do Housing Allowances Work?* Brookings Institution, Washington, DC

Campbell D T, Russo M J 1999 *Social Experimentation*. Sage, Thousand Oaks, CA

Cave G, Doolittle F, Bos H, Toussaint C 1993 *JOBSTART: Final Report on a Program for School Dropouts*. Manpower Demonstration Research Corporation, New York

Ferber R, Hirsch W Z 1982 *Social Experimentation and Economic Policy*. Cambridge University Press, Cambridge, UK

Fienberg S E, Singer B, Tanur J M 1985 Large-scale social experimentation in the United States. *A Celebration of Statistics: The ISI Centenary Volume*: 287–326, Springer-Verlag, New York

Greenberg D, Shroder M 1997 *The Digest of Social Experiments*, 2nd edn. Urban Institute Press, Washington, DC

Hausman J, Wise D (eds.) 1985 *Social Experimentation*. University of Chicago Press, Chicago

Lowry I S (ed.) 1983 *Experimenting with Housing Allowances: The Final Report of the Housing Assistance Supply Experiment*. Oelgeschlager, Gunn & Hain, Cambridge, MA

Morris C N 1979 A finite selection model for experimental design of the health insurance study. *Journal of Econometrics* **11**: 43–61

Morris C N, Hill J L 2000 *The Health Insurance Experiment: Design Using the Finite Selection Model. Public Policy and Statistics: Case Studies from RAND*. Springer, New York

Neuberg L G 1989 *Conceptual Anomalies in Economics and Statistics: Lessons from the Social Experiment*. Cambridge University Press, Cambridge, UK

Newhouse J P 1993 *Free for All? Lessons from the RAND Health Insurance Experiments*. Harvard University Press, Cambridge, MA

Newhouse J P, Marquis K H, Morris C N, Phelps C E, Rogers W H 1979 Measurement issues in the second generation of social experiments: The health insurance study. *Journal of Econometrics* **11**: 117–29

Orr L L 1999 *Social Experiments: Evaluating Public Programs with Experimental Methods*. Sage, Thousand Oaks, CA

Ross H 1966 *A Proposal for a Demonstration of New Techniques in Income Maintenance* (mimeo). Data Center Archives, Institute for Research on Poverty, University of Wisconsin, Madison, WI

C. N. Morris and C. W. F. Chiu

Experimental Design: Overview

Many scientific discoveries are made by observing how a change in a stimulus that is presented to a subject or object affects the response (measurement) given by that subject or object. In an experiment, the investigator has direct control over which stimuli are presented to which subjects in which time periods. This control, when exercised correctly, enables the investigator to deduce a 'cause and effect' relationship, that is, the investigator can deduce that a given change in a stimulus causes a given change in the measured response. The plan of how an experiment is to proceed is called the 'design of the experiment.' The art of experimental design is the art of devising an experimental plan which maximizes the information that can be obtained on the effects of the stimuli.

1. *Terminology*

Experimentation is used in almost every branch of science, with the result that the terminology used in experimental design is not quite standardized. For example, in some fields, the subject or object which is to be presented with a stimulus and then to be measured is called a 'unit' or 'experimental unit.' The stimulus, itself, may be called the 'treatment' or the 'level of a factor' or the 'level of an independent variable.' In factorial experiments, where a subject is presented with a combination of different types of stimuli (such as a particular light intensity together with a particular noise level), the combination may be called a 'treatment combination' but, for simplicity, the term 'treatment' may be taken to mean either a single stimulus or a combination of stimuli, depending on the context.

In some experiments, measurements are made on each subject over several time periods. These are known as 'repeated measurements.' The terminology concerning the associated designs again differs between disciplines. The term 'repeated measurements design' may refer solely to a design involving repeated measurements on a subject to whom a single treatment has been administered or it may include designs in which the treatment is changed before each measurement. The latter type of design is also known as a 'within subjects design' or a 'block design' or a 'cross-over design.' All of these designs are grouped under the heading of 'split-plot designs' by some authors, while others reserve this last term for a design with two types of treatment, one of which is held constant throughout the repeated measurements on a subject and the other of which changes before each measurement see Sect. 9.2.

2. *The Purpose of Experimental Design*

The pioneer in statistical experimental design was Sir R. A. Fisher (Fisher 1951) who was concerned with maximizing the amount of information about agricultural crop production. In the Social and Behavioral Sciences, the questions of scientific interest are very different, but the art of good experimental design is similar.

Every experiment has a budget (time and money) and a limit as to the number of subjects that can be recruited. Also, every experiment has inherent variability; subjects differ from one another in fundamental ways, technicians differ in how they read measuring instruments, subjects and instrument readings vary over time, and so on. Variability translates into uncertainty and uncertainty reduces the amount of useful information. Information gained from an experiment may be viewed either from a sampling perspective or from a Bayesian perspective. In the former, hypothesis tests and confidence intervals are generally used, and the most informative experimental designs yield the most powerful tests and the shortest intervals, whilst producing unbiased results. (E.g., *Estimation: Point and Interval*; *Hypothesis Testing in Statistics*). For a Bayesian analysis, the most informative designs are those which maximize the expected utility (e.g., *Experimental Design: Bayesian Designs*).

The purpose of designing an experiment is (a) to maximize the amount of information that can be gained within a given budget or (b) to minimize the budget required to obtain a given amount of information. Maximizing information is done by controlling and reducing the effect of extraneous variables and by the removal of confounding variables and bias, as described in Sects. 3.2 and 3.3.

3. *Features of Good Design*

3.1 *Comparison*

Experiments, by their nature, tend to be comparative. Questions of interest tend to be of the type 'does this treatment elicit a "better" response than that treatment?' and, if so, 'how much better?.' Even if a single treatment appears to be the only one of interest,

information about its effect on a subject is usually of no value without a comparison with other treatments. For instance, Moore and McCabe (1999, chap. 3) cite an example of a medical study that showed a gastric freezing technique to be a good treatment for ulcer pain, but in later comparative experiments it was shown that the pain relief using freezing was no better than the relief achieved using the same technique but with no freezing solution. Apparently, the pain relief was due to nothing more than doctors showing concern for their patients (a placebo effect) and the freezing technique was abandoned. A good design for evaluating the effect of a single treatment will, therefore, always include a second treatment, called a control or control treatment, for comparison purposes. In some experiments, the control is the treatment in current use, and in other experiments it is the 'absence of a stimulus.' An experiment with more than one treatment needs no control since the experimental treatments can be compared among themselves. Nevertheless, a control can often add extra information. For example, in an experiment on how different types of background music (such as 'classical,' 'rock,' 'rap,' etc.) affect the time taken to learn a new task, a control treatment might be the absence of background music.

3.2 Control of Extraneous and Confounding Variables

A good experimental design allows a particular set of stimuli to be compared with each other with high accuracy. Therefore, any other variable (or factor) that causes the experimental measurements to vary at best reduces the efficiency of the experiment and at worst completely masks the true effect being investigated. For example, in an experiment to investigate the effect of employing different types of memory aid in memorization, the IQ of the subjects may play an important role. The variability of IQ of subjects in the experiment would then contribute to the variability in the measured memorization scores. One way of controlling the effect of such an extraneous variable is to hold the variable fixed during the experiment. For instance, IQ could be held more-or-less fixed by using as subjects only people with a tested IQ within a certain range. Although it reduces variability and avoids masking the effects of the memory aids, this strategy limits the applicability of the conclusions of the experiment since the results would apply only to the people in the population with IQ within this range. A preferable method of controlling extraneous variability is to use a matched design (see below).

An extraneous variable whose effect is completely muddled with that of the factor(s) of interest is called a confounding variable. In the above example, if all the subjects with high IQ were to be tested using the first memory aid and all those with low IQ tested using the second, then, if high IQ is correlated with memorization ability, the first memory aid will inevitably appear to be the better, regardless of its true merits.

The masking effect of a confounding variable can be reduced by randomization (e.g., *Experimental Design: Randomization and Social Experiments*). The simplest form of randomization leads to a 'completely randomized design' (see Sect. 7.1); subjects are recruited from the general public as randomly as is possible (see Sect. 5) and then assigned to the stimuli at random in such a way that every subject has the same chance of being assigned to any one of the stimuli. In such a design, it is likely, although not guaranteed, that each stimulus will receive roughly the same distribution of values of the extraneous variable. If a completely randomized design is used in the above example, the memory aids should receive roughly similar ranges of subject IQs.

In all experiments there are confounding variables that have small effects which are ignored by the experimenter and confounding variables that are accidentally overlooked. The use of randomization helps to spread out the effects of these variables so that the response given to any one stimulus is less likely to be inflated upward or downward due to extraneous factors. For a discussion of random assignment using a random number table or a computer random number generator, see for example, Dean and Voss (1999, chaps. 1, 3), (e.g., *Random Numbers*).

A more foolproof method of ensuring similar distributions of IQ levels for each memory aid in the above example is to divide the subjects into groups so that subjects within the same group have similar IQs, and then to make the random assignment of subjects to stimuli within each group separately. The division into groups provides the control necessary for removing the variability and masking due to the extraneous IQ variable, while maintaining the applicability to the general population. This type of design is called a 'matched design' or 'block design' (see Sect. 7.2). Matched designs and completely randomized design are examples of 'between subjects designs' (Sect. 7).

A further possibility is to measure each subject under a sequence of different treatments—a 'crossover design' or 'within subjects design' (see Sect. 8). The use of such a design in the above example would completely control the extraneous variable since it would ensure that the distribution of IQs assigned to each memory aid is identical. However, new extraneous variables have now been introduced, such as fatigue on the part of the subject over the course of the experiment. If the effects of the new extraneous variables are likely to be small, they can be ignored, but if they are large, then the variables should be controlled by making sure that subjects are assigned the memory aids in different orders. The orders can be assigned at random for each subject separately—a

'block design,' or by deliberately making sure that each stimulus is viewed by the same number of subjects in each time period—a 'latin square design.'

3.3 Removal of Bias

Since each experiment is run with a particular purpose in mind, experimenters tend to have inbuilt, although perhaps subconscious, biases towards or against certain treatments. A random assignment of subjects to treatments and a random ordering of observations ensures that experimenter bias cannot consciously or unconsciously favor one treatment above another.

Subjects' own biases towards the treatments can affect their responses. It may or may not be possible or ethical to conceal the true nature of the experiment from the subjects. It may be possible, however, to mask from both the subjects and the person(s) running the experiment which of the stimuli is the experimental treatment(s) and which is the control (a 'double blind experiment'). Leach (1991 Sect. 24, Appendix 1) discusses ethical issues of concealing information from subjects and lists the guidelines published by the British Psychological Society. Kirk (1982 Sect. 1.5) gives references to ethical guidelines put out by the American Psychological Association, the American Sociological Association, and other bodies.

4. Planning an Experiment

Guides to planning experiments can be found in many texts; for example, Myers (1979, chap. 1), Cox (1958) Dean and Voss (1999, chap. 2), Leach (1991 Sects. 5–9). A protocol, which gives in great detail, step by step, how the experiment is to proceed, is usually prepared well in advance. The protocol includes details about subject selection, measurement, and data collection methods, preparation of materials, preparation of subjects, and a draft statistical analysis.

A pilot experiment, in which a small number of observations is collected, is often run early in the planning stage. Although these observations will usually be thrown away when the main experiment commences, the pilot experiment gives an opportunity to check that the experimental procedure will work as planned and that the required analyses are possible. It also gives indications of unexpected important confounding variables and the likely accuracy of the results. It allows problems to be detected and corrected before they arise in the main experiment.

5. Selection of Subjects

Ideally, the subjects taking part in the experiment are selected at random from the population to which the conclusions of the experiment will be applied (the 'target population'). Since this is not always possible, the experimenter may be forced to use volunteers who will inevitably come from a subset of the population. The results of the experiment may or may not then apply to the entire target population.

The number of subjects required to achieve the goals of powerful hypothesis tests and short confidence intervals, or of maximum expected utility, can be calculated using statistical formulae (e.g., see *Estimation: Point and Interval*; *Hypothesis Testing in Statistics*; *Experimental Design: Bayesian Designs*). The required number of subjects depends upon the design selected for the experiment, on the comparisons of interest, and on the variability of the responses of subjects when assigned the same treatment under identical experimental conditions. In general, 'between subjects designs' require many more subjects than 'within subjects designs.' Between subjects designs are used in fields where subjects can be assigned only a single treatment (as in the evaluation of different teaching methods) and/or where subjects are sufficiently plentiful to off-set the subject-to-subject variability. Within subjects designs are preferred in fields where subjects are scarce or highly variable.

The length of the sequence of treatments that can be presented to any one subject depends upon the nature of the experiment. For example, in experimentation involving multiple visits to a laboratory, subject tolerance can be as low as two or three visits. In experiments in which treatments can be changed rapidly with no long term effect on the subject, a much longer sequence of treatments can be used, requiring fewer subjects in total. There are exceptional circumstances in which an experiment is conducted on a single subject (see Wilson 1995, Kratochwill and Levin 1992), but such experiments cannot give conclusive information about the population as a whole and are not used in standard experimentation.

6. Models and Analysis

The model links the dependent (response) variable(s) to all of the factors (independent variables) that could influence the response, such as the various stimuli, the subjects, time periods, extraneous variables that were used in determining the design, and other important variables that can be taken into account only during the analysis (called 'concomitant variables' or 'covariates': see, e.g., Maxwell and Delaney 1990, chap. 9). All extraneous variables that were ignored in the design are grouped together in a single 'error variable.' A measurement on a subject during a time period before the experiment begins is called a 'baseline measurement' and can be used to increase the accuracy of the results (see, e.g., Jones and Kenward 1989, chap. 2, Sect. 4.4; Cotton 1998, chaps. 9, 10).

If the distributions of the error variables are identical and approximately normal distributions, then standard analysis of variance techniques can be used (e.g., *Analysis of Variance and Generalized Linear*

Models). Analysis of designs in which treatments are assigned the same numbers of subjects and the same numbers of time periods are the easiest to analyse and interpret. They are also the least sensitive to assumptions about normality of the error distributions and equal variances (e.g., *Errors in Statistical Assumptions, Effects of*; *Statistical Analysis, Special Problems of: Transformations of Data*). When the errors do not follow a normal distribution, other types of analysis are needed such as analysis of generalized linear models (e.g., *Analysis of Variance and Generalized Linear Models*), nonparametric analysis (e.g., *Nonparametric Statistics: Rank-based Methods*), and categorical data analysis (e.g., see Ratkowski et al. 1993, chap. 7, Jones and Kenward 1989, chap. 3, Crowder and Hand 1990, chap. 8).

Repeated measurements on a subject under a particular treatment may require a time series or regression analysis, (see Crowder and Hand 1990), (e.g., *Linear Hypothesis: Regression (Basics)*). Bayesian techniques in analyzing cross-over designs are mentioned by Jones and Kenward (1989, pp. 80, 235) (e.g., *Experimental Design: Bayesian Designs*). In the case of correlated responses per subject, multivariate analysis is used for normally distributed errors (see, e.g., Winer et al. 1991, chap. 4, Johnson and Wichern 1992, Maxwell and Delaney 1990, chaps. 13, 14, Crowder and Hand 1990, chap. 4, Jones and Kenward 1989, chap. 7, Myers 1979, chap. 18), (e.g., *Multivariate Analysis: Overview*).

7. Between Subjects Designs

7.1 Completely Randomized Designs

In a completely randomized design, each subject is assigned to just one treatment. The assignment is done completely at random so that each subject has exactly the same chance of being assigned to each possible treatment (stimulus or combination of stimuli). Often a restriction is applied so that each stimulus receives the same number of subjects.

Completely randomized designs are simple to use and simple to analyse. They are most suited to situations where subjects are plentiful, where subjects' responses would not be too variable if they were all given the same stimulus under the same experimental conditions, and where experimental conditions can be held constant throughout the experiment.

If there are extraneous variables which add variation to the responses and which can be measured during the experiment (such as age or IQ), their effects can be removed during the analysis (analysis of covariance) (e.g., Winer et al. 1991 chap. 10). However, where it is possible for the extraneous variables to become confounding variables in an unfortunate randomization, a matched pairs design, block design, or within subjects design would be preferred (see Sects. 7.2 and 8).

As in all designs, it is possible to take repeated measurements in completely randomized designs. Each subject is measured for some number of time intervals after administration of the single treatment. Cotton (1998) calls such a design a 'multigroup split-plot design.'

7.2 Matched Pairs and Block Designs

If there are just two treatments of interest, the variability in the responses (observations) due to differences in the subjects themselves can be reduced by pairing the subjects so that, within each pair, the subjects are alike as possible. For each pair separately, the two subjects are assigned at random to the two treatments (a matched pairs design).

When there are t treatments with $t > 2$, the subjects are matched into groups (or blocks) of size t and, within in each group, the subjects are assigned at random to the t treatments. This type of design is usually known as a 'randomized block design.'

Block designs are also appropriate when the extraneous variation is due to variables unrelated to the subjects. Experimental conditions cannot always be held constant throughout the experiment, as for example, changes in the weather, use of different testing centers, use of different laboratory technicians, etc. To combat this, the subjects would be put into groups of size t (not necessarily matched) and, apart from the assignment to different treatments, all subjects within a group would be tested under the same experimental conditions.

If the number of treatments, t, to be compared is large (as is often the case in a factorial experiment where the treatments are combinations of several stimuli), there may not be enough subjects to form groups of t similar subjects, nor may it be possible to hold conditions constant for t measurements. In this case, an incomplete block design can be used, where each group of $s(< t)$ subjects is assigned at random to a preselected subset of s treatments (see, e.g., Winer 1971, chap. 9, Dean and Voss 1999, chaps. 11, 13).

8. Within Subject Designs

In a within subjects design, each subject is essentially matched with himself or herself and assigned a sequence of some or all of the treatments. The order or presentation is decided using randomization for each subject separately. The comparison of any two treatments is made for each subject ('within' each subject) and then averaged over all the subjects. This has the advantage that subject to subject variability does not play a part in the comparison of the treatments.

In experiments where stimuli are presented to each subject in quick succession, 'carry-over effects' can be a problem. For example, a subject asked to work in bright light followed by normal light may perceive the normal light to be darker than if it had been preceded

by dim light. These carry-over effects can be mitigated by separating the trials by a period of time, called a 'washout period,' in which the subject is asked to do something completely different in some control state. If a long enough washout period cannot be arranged, then the experiment usually is counterbalanced (see Sect. 8.3). Carry-over effects are also known as 'residual effects' and the effect of the treatment administered in the current time period is called the 'direct treatment effect.'

8.1 Cross-over Designs

In the simplest within subjects design, any randomization of the order of treatments for any subject is accepted. However, if time period effects or carry-over effects are thought to be important and are to be included in the model, then it is desirable to exercise control over which treatment sequences are used. In most cases, the carry-over effect from a given treatment will be assumed to be the same no matter which treatment follows it. If the treatments interact, then this assumption may not be valid and larger designs and more complicated analyses are needed.

When there is a small number of treatments, say two or three, and the subject can be measured over several time periods, then each subject can be assigned some or all of the treatments more than once. For three time periods and two treatments (one assigned to one period and one assigned to two periods), there are six possible treatment sequences; with four time periods and two treatments (each assigned to two periods), there are fourteen possible treatment sequences; with four time periods and four treatments (each assigned to one period), there are twelve possible sequences, and so on. If there are sufficient subjects, all the possible sequences can be used an equal number of times. If the number of subjects is small, however, it may not even be possible to use each sequence once. Information must then be used about which set of sequences provides the best design. Among the possibilities are 'variance balanced designs' (Sect. 8.2) and 'latin squares' (Sect. 8.3).

In general, if it is possible to avoid them, two-period designs are not recommended. Not only can the carry-over effects not be estimated independently of the treatment by period interaction, but also three-period designs are considerably more efficient (Jones and Kenward 1989, Sect. 4.16).

8.2 Balanced Cross-over Designs

Cross-over designs that allow differences between all pairs of treatments to be estimated with the same precision are called 'variance balanced.' Variance balanced designs include cross-over designs that use all possible treatment sequences, and counterbalanced latin square designs (see Sect. 8.3). Variance balanced designs, efficient for comparing all pairs of treatments, are tabulated by Ratkowski et al. (1993, chap. 5) and also by Jones and Kenward (1989, pp. 212–4, 223–4). The latter authors also list designs efficient for comparing test treatments with a control.

The treatment given in the last period of a cross-over design can be repeated in an extra period. Such designs, called 'extra period designs,' allow the carry-over from the last treatment to be measured, thus increasing the precision of the treatment comparisons.

Balanced designs typically require a large number of subjects when the number of treatments is large and so cannot always be used. An alternative is to use an efficient 'partially balanced' design. These designs allow treatment differences to be estimated with two or three different precisions that are fairly close in value. When the treatments are factorial in nature, the effects of the individual factors (main effects) and the interactions between the factors are usually of primary interest. Variance balance is then desirable for the comparisons of the levels of each factor separately (see Jones and Kenward 1989, pp. 222–8, Ratkowski et al. 1993, chap. 6, for tabulated designs).

8.3 Latin Squares

A latin square design is ideal for any experiment in which it is possible to measure each subject under every treatment and, in addition, it is necessary to control for changing conditions over the course of the experiment. A latin square is a design in which each treatment is assigned to each time period the same number of times and to each subject the same number of times (see Dean and Voss 1999, chap. 12). If there are t treatments, t time periods, and mt subjects then m latin squares (each with t treatment sequences) would be used.

Carry-over effects are controlled by using latin squares that are 'counterbalanced' (Cotton 1993). This means that, looking at the sequences of treatments assigned to all the subjects taken together, every treatment is preceded by every other treatment for the same number of subjects. Counterbalanced latin squares exist for any even number of treatments and for some odd numbers of treatments (e.g., $t = 9$, 15, 21, 27; see Jones and Kenward 1989 Sect. 5.2.2, for references). For other odd numbers, a pair of latin squares can be used which between them give a set of $2t$ counterbalanced sequences.

If a carry-over effect is expected to persist for more than one time period, then the counterbalancing must be extended to treatements occurring more than one time period prior to the current treatment.

9. Other Designs

9.1 Nested or Hierarchical Designs

It is not unusual for extraneous variables to be 'nested.' For example, if subjects are recruited and tested

separately at different testing centers, the subjects are 'nested within testing center.' If subjects are animals such as mice or piglets, then the subjects are naturally nested within litters, which are nested within parent, which may be nested within laboratory. The nesting information can be used in matched designs, since the nesting forms natural groupings of like subjects. For within subjects designs, the nesting information can be used during the analysis for examining the different sources of extraneous variation (e.g., *Hierarchical Models: Random and Fixed Effects*). Designs in which different levels of nesting are assigned different treatment factors are called 'split-plot designs' (see Sect. 9.2).

A second type of nesting is a nesting structure within the treatment factors being examined. Examples given by Myers (1979) include memorization of words within grammatical class; time taken to complete problems within difficulty levels. Models and analyses used in such experiments must reflect the nested treatment structure.

9.2 Split-plot Designs

An experiment with more than one type of stimulus (factor) can be run as a split plot design with a level of one or more factors being applied to a subject throughout the course of the experiment (as for a between-subjects design), and the levels of the other factor(s) being changed for each time period (as for a within subjects design). Such designs are sometimes called 'mixed designs.' The stimuli applied as the within subjects design will generally be measured more accurately than those applied as the between subjects design, since subject to subject variability enters into the comparison of the latter.

Split-plot designs are useful when it is difficult to change the levels of one of the factors. For example, Dean and Voss (1999, chap. 19) cite an example of an optokinetic experiment on the drift of focus of a subject's eyes from the center of a rotating drum measured under two different lighting conditions. The change of lighting conditions was a time-consuming process, whereas it was simple to change the speed of rotation of the drum. Consequently, each subject was assigned a single lighting condition throughout an entire viewing session, and during the viewing session the subject was assigned a sequence of different speeds.

10. Optimality and Efficiency of Designs

As pointed out by Cotton (1998), designs that are best (i.e., optimal) for one purpose are not necessarily best for another purpose and compromises may need to be made. The optimal design for investigating the effects in one model may be totally unsuitable for a different model. For example, in a cross-over design with two treatments and two time periods, a set of counterbalanced latin squares provides an optimal design for estimating direct treatment effects and carry-over effects. This design does not, however, allow estimation of both a carry-over effect and an interaction between treatments and time periods. Thus, if both of these effects are required in the model, then more than two time periods must be used in the experiment (e.g., *Statistical Identification and Estimability*).

As a general rule, the most balanced designs are optimal when interest lies equally in all treatment comparisons. The following features are typical characteristics of balance; every treatment is assigned the same number of subjects, every treatment is observed in every time period the same number of times, every treatment is preceded by every other treatment (including itself, if possible) for the same number of subjects and in the same number of time periods. When balance is not achievable, computer programs for generating optimal designs are commercially available. For other settings, where comparison of all treatments is not the main goal of the experiment, more sophisticated algorithms are needed see, for example, Atkinson and Donev (1992). For a discussion of design optimality from a Bayesian perspective, see for example, *Experimental Design: Bayesian Designs*.

Bibliography

Atkinson A C, Donev A N 1992 *Optimum Experimental Design*. Oxford Science Publications, Oxford, UK

Breakwell G M, Hammond S, Fife-Shaw C (eds.) 1995 *Research Methods in Psychology*. Sage, London

Cotton J W 1993 Latin square designs. In: Edwards L K (ed.) *Applied Analysis of Variance*. Dekker, New York

Cotton J W 1998 *Analyzing Within-subjects Experiments*. Erlbaum, Marwah, NJ

Cox D R 1958 *Planning of Experiments*. Wiley, New York

Crowder M J, Hand D J 1990 *Analysis of Repeated Measures*. Chapman and Hall, London

Dean A M, Voss D T 1999 *Design and Analysis of Experiments*. Springer Verlag, New York

Edwards L K (ed.) 1993 *Applied Analysis of Variance in Behavioral Sciences*. Dekker, New York

Fisher R A 1951 *The Design of Experiments*, 6th edn. Oliver and Boyd, Edinburgh, UK

Harris P 1986 *Designing and Reporting Experiments*. Open University Press, Milton Keynes, UK

Johnson R A, Wichern D W (eds.) 1992 *Applied Multivariate Statistics*, 3rd edn. Prentice Hall, Englewood Cliffs, NJ

Jones B, Kenward M G 1989 *Design and Analysis of Cross-over Trials*. Chapman and Hall, London

Keppel G 1982 *Design and Analysis: A Resercher's Handbook*, 2nd edn. Prentice Hall, Englewood Cliffs, NJ

Kirk R E 1982 *Experimental Design: Procedures for the Behavioral Sciences*, 2nd edn. Brooks/Cole, Belmont, CA

Kratochwill T R, Levin J R 1992 *Single-case Research Design and Analysis: New Directions for Psychology and Education*. Erlbaum, Hillsdale, NJ

Kuehl R O 1994 *Statistical Principles of Research Design and Analysis*. Duxbury, Belmont, CA

Leach J 1991 *Running Applied Psychology Experiments*. Open University Press, Milton Keynes, UK
Maxwell S E, Delaney H D 1990 *Designing Experiments and Analyzing Data. A Model Comparison Perspective*. Wadsworth, Belmont, CA
Moore D S, McCabe G P 1999 *Introduction to the Practice of Statistics*, 3rd edn. Freeman, New York
Myers J L 1979 *Fundamentals of Experimental Design*, 3rd edn. Allyn and Bacon, Boston
Ratkowski D A, Evans M A, Alldredge J R 1993 *Cross-over Experiments. Design, Analysis and Application*. Dekker, New York
Senn S 1993 *Cross-over Trials in Clinical Research*. Wiley, New York
Wilson S L 1995 Single case experimental designs. In: Breakwell G M, Hamond S, Fife-Shaw C (eds.) *Research Methods in Psychology*. Sage, London
Winer B J 1971 *Statistical Principles in Experimental Design*, 2nd edn. McGraw-Hill, New York
Winer B J, Brown D R, Michels K M 1991 *Statistical Principles in Experimental Design*, 3rd edn. McGraw-Hill, New York

A. M. Dean

Experimental Design: Randomization and Social Experiments

1. *Definition, Functions, and Rationale*

In a randomized field trial, a sample of individuals or of entities, such as schools or hospitals, are randomly assigned to one of two or more interventions. These interventions may be different programs or variations on the same program, different intensities of social service, different mixes of therapies, and so on. The interventions may include a 'control' condition in which no services, or services that are ordinarily available, are provided (see *Experimental Design: Overview*).

For example, to understand whether a new employment and training program works well, one randomly allocates some eligible individuals to the new program. All or some of the remaining eligible individuals are randomly allocated to the alternative intervention. This alternative may be a normally available training service, i.e., a control condition, or a program with lower levels of service, or a different service.

This random assignment of individuals to a new employment and training program and to an alternative program permits one to make a fair comparison of the wage rates achieved by the people who were assigned to each group. The value that is added by a new program can be estimated well because the people who were assigned to the conventional services and to the new program are equivalent on account of the random assignment. Furthermore, one can take into account the chance statistical differences among individuals who were assigned to each group, a fact that is important in science and law (Kaye and Freeman 1994).

The logic that underlies a randomized trial is as follows. To estimate the relative effect of a particular intervention, one has to estimate the status of individuals (or entities) had they not had the intervention. At times, a precise forecast can be made about how individuals or entities would fare in the absence of the intervention. Also at times, one may be able to construct what appears to be a fair comparison group that represents the state of people (or institutions) in the absence of the particular intervention. Under these circumstances, quasi-experiments or observational studies might then be employed to generate evidence about the relative effectiveness of programs. These approaches try to make a fair comparison of the outcome of alternative interventions based on sophisticated statistical models of how people and entities behave, and on conscientious attention to competing explanations about why a difference in the effects of interventions might appear. See Maynard and Chalmers (1997) and *Quasi-Experimental Designs*.

Often, the researcher cannot make good forecasts, construct defensible *ad hoc* comparison groups, or account for competing explanations for a difference among interventions in a quasi-experiment. A randomized trial may than be warranted because it assures that the prediction is fair: the comparison is fair simply because people or entities are randomly allocated into two or more equivalent groups. It assures that competing explanations for people's (or entities') behaviors are taken into account because both groups are subject to the same influences, apart from the interventions.

Social and behavioral scientists and statisticians use phrases other than 'randomized field trial' to describe this approach to generating evidence. These phrases include randomized experiment, randomized trial, randomized clinical trial (RCT) in medical research, and controlled experiment. Different phrases are used to identify trials in which entities such as schools or hospitals are randomly assigned to different interventions. These descriptors include cluster-randomized trials, group randomized experiments, macro-experiments, and place-based trials (Murray 1998, Boruch and Foley 2000).

2. *Examples of Randomized Field Trials*

Good examples of randomized trials are not difficult to locate, although such trials are in the minority of studies on the effects of interventions. Boruch (1997) gives many such examples. In Switzerland, for example, randomized trials have been carried out to understand the effects of heroin therapy vs. conventional treatment of drug addicts and to estimate

the effects of community service vs. imprisonment. In Germany, community rehabilitation programs in prisons have been assessed vs. conventional prison programs in randomized trials. In the US, randomized trials have been carried out to learn the effects of arresting people for misdemeanor domestic assault, new approaches to courts' handling of certain cases, intensive police patrol strategies, and other interventions. A main outcome variable in such experiments is recidivism. See the special issues of *Crime and Delinquency*, 2000, for details on these and other trials.

Measures of academic achievement and related outcomes have also been important in randomized trials to estimate the effects of new approaches to tutoring in England and the United States (Fitz-Gibbon 1990, Fuchs et al. 1997); reducing the size of elementary school classrooms in the United States (Mosteller 1999); radio-based mathematics education in Nicaragua (Jamison et al. 1980); cultural enrichment programs for children in the barrios in Colombia (McKay et al. 1978); and cooperative group learning in elementary grades in Australia (Gillies and Ashman, 1998).

Randomized trials have also been mounted to test programs that were purported to reduce individuals' dependence on public welfare. The United States has been the site for many studies on training and employment, housing, and welfare reform. Greenberg et al. (1997) summarized over 140 such trials, including a Canadian study of tax rates and guaranteed annual income and a Dutch experiment on intensive job assistance services for the unemployed.

Many randomized field trials have been used to test behavioral and social approaches to health care, including reducing HIV risk exposure or the risk of sexually transmitted diseases. See, for instance, reports on tests of husband-oriented approaches to enhancing the use of contraceptive devices in Ethiopia by Terefe and Larsen (1993), AIDs education in the Philippines (Alplasca et al. 1995), among others.

3. History

Claude Bernard's *Introduction to the Study of Experimental Medicine*, first published in 1865 in France, outlined a framework for controlled comparisons. Bernard, however, said nothing about the topic of randomized trials.

Scientists' use of randomization in order to construct groups that do not differ systematically, apart from chance and effects of interventions, is more recent. In the late 1890s, for instance, researchers in Denmark used randomized groups to understand whether certain diphtheria serum worked (Hrobjartsson et al. 1997). During the same period, psychophysical experiments in German laboratory studies employed random assignment (see Stigler 1986). For history on the idea of randomized experiments in health and other sectors and of comparisons, see www.rcpe.ac.uk/cochrane, Maynard and Chalmers (1997), and Marks (1997).

Large-scale randomized controlled trials (RCTs) in medicine appear to antedate randomized field trials (RFTs) in the social sector. Contemporary RCTs have their origins in trials during the 1940s and 1950s, undertaken by Bradford Hill and others in the United Kingdom to generate high-quality evidence on the effectiveness of streptomycin in treating tuberculosis (Marks 1997). Later, Jonas Salk, Paul Meier, and others in the United States successfully mounted randomized trials to discern the effects of vaccine on the incidence of poliomyelitis (Chalmers 1999).

Some contemporary history of randomized trials in the social sector is given in Boruch (1997) and Oakley (2000). See both books for references to the examples that follow. In criminological research, the Manhattan Bail Bond experiment of the 1960s is a classical case. It involved randomly eliminating bail for some people to understand whether bail bond requirements actually work to assure that people appear for a trial. More recent and remarkable trials are described in the University of Maryland's report to the US Congress (http://www.nijrs.org and http://preventingcrime.org).

Advances in the economic arena include the graduated work on incentive trials of the 1970s, tests of employment and training programs in the 1980s, and trials on welfare programs in the 1990s. Fienberg et al. (1985) discuss these and other large-scale economic and policy experiments, and how substantive theory is important in the experiments' designs. Many of these studies are abstracted in Greenberg et al. (1997). See the website for the Social Research Demonstration Corporation for information on related work in Canada (http://www.policyresearch.gc.ca).

In the 1960s, the Perry Preschool/High Scope trial on preschool education is a precedent that is well known by people who want to understand the effects of early childhood intervention programs (Boruch 1997). Taroyan and others (2001) have mounted a similar trial in England. No reports on a large-scale randomized trial on the effects of reducing classroom size were reported until the 1990s. The Tennessee STAR trial is a remarkable exception covered in a special issue of *Education Evaluation and Policy Analysis* (2000), Mosteller (1999), and Finn and Achilles (1990).

4. Ethics and Law

The ethical standards for research on human subjects in many countries have been influenced heavily by the Nuremberg Code and the Helsinki Declaration on medical research. Elements of these codes have been adopted, at times, to assure that the rights of people

are protected in social, criminological, and educational research. The codes are pertinent to randomized field trials as they are to other kinds of studies.

Such codes require, for instance, that individuals be informed that their engagement in a study is voluntary, and that they can decline to participate or, having agreed to participate, that they can withdraw. Codes regard it as essential to obtain the participant's informed consent, based on a well-specified statement of possible risks and benefits to the participant. Moreover, an institutional review board is often required to examine the experiment's design in order to assure that certain ethics' standards are met.

In the United States, for example, the US Code of Federal Regulations (45 CFR 46) requires that most research organizations in the United States abide by this kind of standard. Canada, the United Kingdom, and the Nordic countries have similar codes. Such governmental rules can be found at times on worldwide web sites such as www.helix.nih.gov. Research journals and books also provide important information. See, for instance, see Sieber (1992) and articles in two special issues of *Crime and Delinquency*, 2000, **46**(2–3), for specific references.

Many professional associations in social research have adopted codes of conduct that enlarge on the international codes and government standards. For instance, members of the Board of the American Criminological Association have (a) reiterated the need to attend to the ethics of research on human subjects and (b) emphasized the ethical responsibility of service providers, as well as researchers, to generate good evidence through randomized trials about the effectiveness of interventions in justice research. The Board also reminded readers that it is unethical for professionals in the service delivery arena, such as the courts, to require that individuals receive treatments whose effectiveness is unknown (*Crime and Delinquency*, 2000, **46**(3)).

One of the paradoxes of codes of ethics for researchers is that codes of research ethics have not been applicable to professionals who experiment in an uncontrolled way by 'trying something out.' Despite the absence of unequivocal evidence on a treatment's efficacy, for example, a physician may use the treatment on a patient without any ethics board review and without the informed consent that is required in a randomized trial. Similarly, a teacher who tries out new ways of teaching may do so without oversight by any institutional review board.

Ethical standards are often reified in law. That is, legislatures, diets, parliaments, and other law-making bodies can influence decisions about whether randomized trials are run. In the United States, for instance, welfare laws that were enacted during the 1990s permitted various states of the US to depart from national rules. This permission to depart from national law was in the interest of testing unconventional approaches to enhancing the education and employment status of low-income families. Such 'waivers' are an important vehicle for allowing ordinary societal interest in improvement, to deploy and test the innovation.

More generally, the designer of a randomized field trial can use at least four simple strategies to tailor the trial's design so as to conform to good standards of ethics and law. In the first place, if a purportedly effective service is in short supply, relative to the demand for the service, random assignment of people to the service can often be justified. That is, many people regard a lottery as the fairest way to allocate a scarce resource.

Second, people recognize that the eligibility of individuals for a particular intervention is an important ingredient in their design of a trial. For example, some criminal offenders clearly need to be jailed. Other offenders clearly ought not be put into jail. For still others, the effectiveness of jail would not be known. The randomized study would focus then only on those for whom the benefit of jail to the individual is unpredictable. That is, one designs the trial so as to include only those for whom the intervention's effectiveness is uncertain.

A third strategy to the ethical design of randomized field trials involves a wait-list approach. That is, equally eligible people are randomly allocated into a group that receives treatment early and others are randomized to receive treatment later. The experiment then compares the state of two groups after the first has completed treatment and before the second group has begun.

A fourth strategy recognizes that, in some cases, random assignment of individuals to treatments is unethical (or infeasible) but recognizes that the random assignment of entities to different interventions is ethical. The trial's designer, for instance, may find that local juvenile facilities are unwilling, on ethical grounds, to assign individuals randomly to different treatments in their facility. The trialists may also find, however, that the same facilities are willing to try out interventions whose effectiveness is unknown, so long as the entire facility is assigned (randomly) to the intervention. Boruch and Foley (2000) and Murray (1998) reviewed randomized trials of this kind. Hospitals, factories, schools, geopolitical jurisdictions, police precincts, and other entities are the units of random allocation and analysis.

5. *Institutional Capacity for Randomized Trials*

A randomized field trial requires resources. To judge from experience in different countries, an essential resource is people who are willing to support and contribute to studying the effects of innovations such as policy makers and practitioners. The Tennessee Legislature's support of the State's trial on reducing the sizes of classes in schools is a case in point. See the

special issue of *Educational Evaluation and Policy Analysis*, 2000, **20** (Nov. 10).

Doing a randomized field trial also requires understanding of managerial issues in field research and how to handle the cultural and other political-institutional factors that must be taken into account. Organizations and teams that are capable of running the experiment are important. Over the past 20 years, organizations such as Manpower Demonstration Research Corporation, Abt Associates, and Mathematica Policy Research have emerged in the United States to conduct large-scale experiments. In Canada, the Social Research and Demonstration Corporation has done so. Other organizational arrangements have been developed in universities to conduct small- to medium-scale trials. In the United Kingdom, for instance, see Fitz-Gibbon (1990) on school-based interventions. Martin Killias and his colleagues in Switzerland and Rudiger Ortmann and his colleagues in Germany built teams to mount randomized trials in the crime prevention arena. (See the special issue of *Crime and Delinquency*, 2000, **46**(2).)

6. *Complications and Issues*

When a simple randomized trial involving two interventions is well designed and deployed, the statistical analysis and inferences are usually straightforward. Trials are not always simple. Nor are all such studies implemented without problems.

Trialists know how to randomize and check the integrity of the randomization. See Boruch (1997) and *Experimental Design*: *Overview*. Nonetheless, randomization can be corrupted especially when the trial is run in an unfamiliar context. If the randomization is subverted, it is often difficult if not impossible to take this into account in a satisfactory way.

Further, trialists capitalize on contemporary statistical methods and software to design a trial that maximizes the experiment's statistical power under different scenarios. This assures that real differences among programs can be discerned. In novel contexts, however, it may not be possible to use this technology easily because little is known about the target groups, programs, and context. More important, such power analyses help to discount small sample size and other design factors as a reason for finding no remarkable differences among programs. The trialist must usually get beyond this, addresssing questions such as: did the interventions and control conditions really differ in character or operation? Is there something wrong with the theory underlying the new program or one's understanding of the problems it was designed to resolve? See *Experimental Design: Compliance*.

The advice to 'analyze the units as you have randomized them' is wise in that this approach preserves the benefits of randomization. That is, when individuals have been randomly allocated, they are then the units of analysis. When institutions have been randomly assigned to alternative programs, then these are the units of statistical analysis. In some trials, matters are more complicated. Schools may be randomly assigned to interventions; teachers within schools may or may not be randomly assigned to children within classrooms and schools. Designing these more complicated trials depends on the knowledge base in education and in statistics, including understanding of hierarchical and mixed statistical models.

Accumulating and synthesizing results of randomized trials is obviously important in building the knowledge base about the effectiveness of social innovations. Advances in meta-analysis and systematic reviews can help to do this. There are complications of course. In any given area, randomized trials vary in quality and trustworthiness and so ought not be given equal weight. Despite the threats to validity of nonrandomized trials, some of these trials produce evidence that is as defensible, under certain assumptions, as the evidence in randomized trials. Learning how to put the two together depends on new statistical methods such as propensity scoring, access to original microrecords, and substantive knowledge about programs, target populations, and contexts. See *Meta-analysis*: *Overview* and Sect. 7.

7. *International Information on Randomized Field Trials*

Until the early 1990s, there existed no central source of information on randomized trials undertaken by health care researchers. In 1993, the international Cochrane Collaboration (Maynard and Chalmers 1997) was created to prepare, maintain, and disseminate systematic reviews of studies on the effects of health care programs. As part of its mission to produce systematic reviews, the Collaboration evolved into an important major source of information on randomized trials. The Cochrane Library is accessible through subscription on the worldwide web site (ww.cochrane.ac.uk).

Until the year 2000, no organization had been invented to provide ready access to randomized trials in education, crime and justice, social work and social welfare, and other social sectors. The international Campbell Collaboration was developed as a younger sibling to the Cochrane Collaboration. The Campbell Collaboration's mission, in part, is to identify, maintain, and make accessible, randomized field trials done in all social sectors. The collaborators use this registry resource to generate systematic reviews of the effectiveness of interventions (http://campbell.gse.upenn.edu).

See also: Fieldwork: Ethical Aspects; Meta-analysis in Practice; Psychological Treatments: Randomized Controlled Clinical Trials

Bibliography

Alplasca M, Sieged D, Mandel J S, Santana R, Paul J, Hurdes E S, Monzon O T, Hearst N 1995 Results of a model AIDS prevention program for high school students in the Philippines. *AIDS Supplement* **1**: 7–13

Bernard C 1957/1865 *An Introduction to the Study of Experimental Medicine* [trans. Greene H C]. Dover Publications, New York

Boruch R F 1997 *Randomized Experiments for Planning and Evaluation: A Practical Guide*. Sage Publications, Thousand Oaks, CA

Boruch R F, Foley E 2000 The honestly experimental society. In: Bickman L (ed.) *The Legacy of Donald T. Campbell*. Sage Publications, Thousand Oaks, CA, pp. 193–238

Chalmers I 1999 Why transition from alteration to randomization was made: Letter to the Editor. *British Medical Journal* **319**(20): 1372

Fienberg S E, Singer B, Tanur J M 1985 Large scale social experimentation in the United States. In: Atkinson A C, Fienberg S E (eds.) *A Celebration of Statistics: The ISI Centenary Volume*. Springer-Verlag, Berlin, pp. 287–326

Finn J D, Achilles C M 1990 Answers and questions about class size: a state wide experiment. *American Educational Research Journal* **27**(3): 557–77

Fitz-Gibbon C T 1990 Success and failure in peer tutoring experiments. In: Goodlad S, Hirst B (eds.) *Explorations in Peer Tutoring*. Basil Blackwell Ltd, Oxford, UK, pp. 26–57

Fuchs D, Fuchs L S, Mathes P G, Simmons D C 1997 Peer-assisted learning strategies: making classrooms more responsive to diversity. *American Educational Research Journal* **34**(1): 174–206

Gillies R M, Ashman A F 1998 Behavior and interactions of children in cooperative groups in lower and middle elementary grades. *Journal of Educational Psychology* **90**(4): 746–57

Greenberg D, Shroder M, Onstott M 1997 *Digest of Social Experiments*. Urban Institute Press, Washington, DC

Hrobjartsson A, Gotzche P C, Gluid C 1997 The controlled clinical trial turns 100, Fibieger's trial of serum treatment of diphtheria. *British Medical Journal* **317**: 1243–5

Kaye D H, Freeman D A 1994 Reference guide on statistics. In: Federal Judicial Center (ed.) *Reference Manual on Scientific Evidence*. Federal Judicial Center, Washington, DC, pp. 331–412

Jamison D, Searle B, Suppes P 1980 *Radio Mathematics in Nicaragua*. Stanford University Press, Stanford, CA

Marks H M 1997 *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990*. Cambridge University Press, Cambridge UK

Maynard A, Chalmers I (eds.) 1997 *Nonrandom Reflections on Health Services Research: On the 25th Anniversary of Archive Cochrane's Effectiveness and Efficiency*. British Medical Journal (BMJ) Publishing Group, London

McKay J H, McKay A, Sinnestera L, Gomez H, Lloreda P 1978 Improving cognitive ability in chronically deprived children. *Science* **200**(4): 3270–8

Mosteller F 1999 How does class size relate to achievement in schools?. In: Mayer S E, Peterson P E (eds.) *Earning and Learning: How Schools Matter*. Brookings Institution Press, Washington DC and Russell Sage Foundation, New York, pp. 117–30

Murray D M 1998 *Design and Analysis of Group Randomized Trials*. Oxford University Press, New York and Oxford, UK

Oakley A 2000 *Experiments in Knowing*. Polity Press, Cambridge, UK

Sieber J 1992 *Planning Ethically Responsible Research: A Guide for Students and Institutional Review Boards*. Sage Publications, Thousand Oaks, CA

Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belnap Press of Harvard University Press, Cambridge, MA

Taroyan T, Roberts I, Oakley A 2000 Randomisation and resource allocations: A missed opportunity for evaluating health care and social interventions. *Journal of Medical Ethics* **26**: 319–22

Terefe A, Larson C P 1993 Modern contraception use in Ethiopia: does involving husbands make a difference? *American Journal of Public Health* **83**(1): 1567–71

R. F. Boruch

Experimental Economics

Why do scientists perform experiments? Ostensibly, they want to test their operating understanding of the principles or theory that give coherence to the phenomena that interest them. They are interested in knowing if and how their conceptual schemes work, when tried. Experimental economics asks whether, and under what conditions, our models of decision, strategic interaction, and markets work in the sense that the implications of the principles are consistent with what we observe. But experimentalists everywhere—including economics—soon discover that their theoretical (conceptual) framework is inadequate to embrace all of their observations, and they (should) ask why? At this point experimentation takes on a life of its own. To the extent that experimentalists get interested in that life, learn to like it, and pursue it, they become phenomenologists seeking a deeper understanding of behavior. Such phenomena can lead to a fundamental restructuring of theory; until the restructuring occurs, commonly the phenomena exist as a way of thinking about the world that provides an engine for generating hypotheses that only indirectly connect with theories that might have motivated the initiating experiments. This is a potent source of experimental knowledge driven by empirical/theoretical probes of behavior and its interpretation outside the boundaries of standard theory.

This article provides an introduction to laboratory methodology in economics, an overview of a small selection of results indicating examples of what has been learned from this research with an emphasis on how such learning can be linked with data and theory outside of traditional economics. Also the use of the laboratory as a test-bed for economic systems design is briefly discussed. Where possible, references will be to citations and discussion in Davis and Holt (1993), Friedman and Sunder (1994), Kagel and Roth (1995), and Smith (1991).

1. What is an Economics Experiment?

Every microeconomy (laboratory or field) can be represented by three elements. (Smith 1991, pp. 255–71). The first is the environment, specifying agent characteristics that motivate action. In neoclassical economics this included agent endowments, preferences, and technology (costs). In laboratory experiments monetary rewards are most often used to induce a specific value/cost environment and to motivate decision. The institution defines the language (messages) of agent communication, the rules (property rights) governing the exchange of information, allowable actions (moves), and how messages/actions lead to valued outcomes. Laboratory experiments use instructions and decision forms (often computer controlled) to define the institution. Finally, there is the observed behavior of agents/subjects given the environment and institution. Here are three examples of laboratory microeconomies

1.1 Trust Game (*Fig. 1*)

1.1.1 Environment. The alternative payoffs are shown in the outcome parentheses for (DM1, DM2), where DM_i is decision-maker $i = 1, 2$. Thus, at the bottom of the decision tree, conditional on reaching node 2, if DM2 moves down, the payoffs are (DM1, DM2) = (\$0, \$40).

1.1.2 Institution. Subjects, for example, in groups of at least 12, are randomly and anonymously paired, and randomly assigned the roles (1 or 2). DM1 moves first, choosing right, in which case the game ends with payoffs (\$10, \$10), or down. If down, DM2 chooses right yielding (\$15, \$25) or down (\$0, \$40). Each player sees the moves in sequence of the other. No communication is permitted (except the moves); neither DM knows with whom he/she is paired. Variations on this institution might include players not seeing each other's moves; being able to communicate or play might be repeated with the same unknown person. The experimental protocol—rules and procedures—define the institution in an experiment.

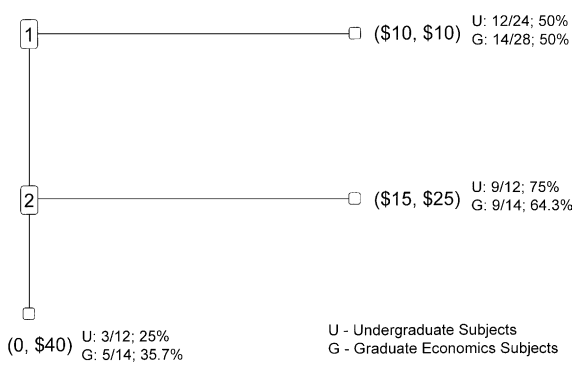

Figure 1
Trust game: moves and payoffs for (decision-maker 1, decision-maker 2); game ends at payoff node

1.1.3 Behavior. Shown next to each payoff outcome is the (proportion; percentage) of all subjects ending at the indicated payoffs for each of two subject population samples:

(a) Undergraduates (U) with various majors; the typical subject.

(b) Graduate (G) students in economics, from across the US and Europe in their third or fourth year of study, attending a workshop in experimental economics.

For a discussion of bargaining experiments and further references see Davis and Holt (1993, pp. 241–75), Friedman and Sunder (1994, pp. 56–60), Kagel and Roth (1995, pp. 253–348), and McCabe et al. (1996).

1.2 Auction Experiments

1.2.1 Environment. Subjects, for example, in groups of four, are each assigned one value, $V_i^t \varepsilon$ [0, \$10], chosen independently with replacement, where each value is equally likely to be drawn for each i in each auction period, $t = 1, 2, \ldots, T$, where T is unknown to the bidders. V_i^t is the maximum-willingness-to-pay of i for one unit of the abstract good in period t, since if i wins the auction he/she earns $V_i^t - (\text{price paid})_i^t$. At each t one unit is offered for sale to the four competing bidders at zero supply price.

1.2.2 Institution (sealed high-bid auction). At t each i submits a bid, $b_i^t \geqslant 0$, the high bidder wins and pays a price equal to the amount bid. Hence, if $i = 1$ is the high bidder, his/her earnings are $V_1^t - b_1^t$.

1.2.3 Behavior. Subjects tend to submit bids that are approximately proportional to their assigned value, i.e., $b_i^t = \hat{\alpha}_i V_i^t$, where $\hat{\alpha}_i$ is i's estimated bid coefficient based on all observations (b_i^t, V_i^t). If all bidders are risk neutral, the Bayes–Nash theory predicts $\alpha_i = \frac{N-1}{N} = \frac{3}{4}$, with $N = 4$, for all i. If agents are risk averse, their bid coefficient is $\alpha_i > \frac{3}{4}$. This is because, so it is argued from the theory, that they are willing to accept less profit, $V_i^t - b_i^t$, to increase the chances of being the high bidder. Almost all bidders (over

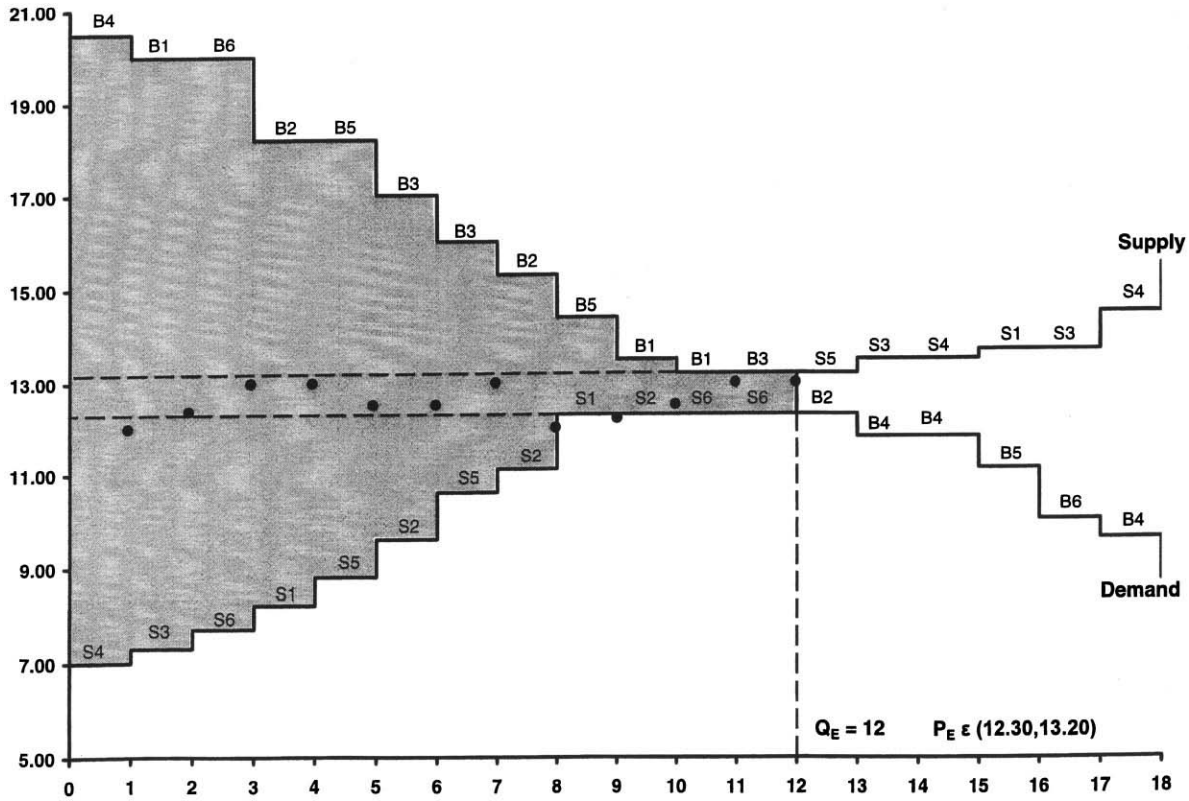

Figure 2
Oral double auction experiment: supply and demand, plot of contract prices

90%) in high-bid experimental auctions bid as if they are risk averse. See Davis and Holt (1993, pp. 275–308), Friedman and Sunder (1994, pp. 107–9), Kagel and Roth (1995, pp. 501–85), and Smith (1991, pp. 509–661) for references and discussion of experimental auctions.

1.3 Double Auction Stationary Supply and Demand Experiment

1.3.1 Environment. M buyers are assigned one or more positive values in period 1. Each buyer earns $V_i - P_i$ for each unit with value V_i that is purchased at price P_i; each of N sellers earns $P_j - C_j$ on each unit sold at price P_j, where $C_j > 0$ is one of the costs assigned to seller j in period 1. The same values and costs might be assigned in period 2 and all subsequent periods, defining a stationary replicating economic environment.

An example is provided in the experiment of Fig. 2 in which there are six buyers, each with a capacity to buy up to three units. Buyer 1 has a three-unit demand schedule. His three maximum willingness-to-pay limit prices are at \$20, \$13.50, and \$13.20 as indicated by the three units designated by B1 in Fig. 1. If B1 buys a first unit, he earns \$20 minus the purchase price; a second unit earns \$13.50 minus its purchase price and so on. Other buyers have limit price demand schedules as shown. When all limit prices are pooled and sorted from highest to lowest this defines the market demand schedule of prices and corresponding maximum purchase quantities.

Similarly, sellers are assigned minimum willingness-to-accept limit selling prices. For example, seller S4 can sell up to three units at unit costs \$7.00, \$13.50, and \$14.50, respectively. When all seller limit prices are sorted from lowest to highest we get the market supply schedule, shown in Fig. 2.

Each buyer and seller only knows his or her private local supply or demand schedule. As Hayek would say, information is dispersed; each agent knows his or her own particular circumstances of time and place which we summarize in the form of the indicated limit values and costs in Fig. 2 (Hayek 1945, cited in Smith 1991, p. 234). Supply is to order in the environment shown in Fig. 2 so that there is no carryover of units

not sold and therefore not supplied. In a multiple period competitive market experiment the supply and demand schedules so defined become flows per period. Note that any P_E in the interval (\$12.30, \$13.20) is a competitive equilibrium that clears the market of 12 units. The theoretical gains from exchange are represented by the shaded area of Fig. 1. Keep in mind that no subject in the experiment knows this, and none sees before them a chart like Fig. 1. This information is unnecessary.

1.3.2 Institution. We also need an institution that defines the language of the market and its operating rules. In the example of Fig. 2 we use the bid-ask double auction rules common in financial and commodity markets. Any buyer (seller) can announce a bid (ask) price to buy (sell). Agents are recognized one at a time with bids and asks written into a spread sheet in sequence on an overhead. Once a bid (ask) is established a new bid (ask) must be higher (lower). Hence, the bid-ask spread is required to narrow. At any time any buyer (seller) can accept the standing ask (bid) to form a contract. For example, seller S1 might sell her first unit to buyer B6 at price \$12. The seller earns a profit of \$12 − \$8.20 = \$3.80 and the buyer a profit of \$20 − \$12 = \$8. This process continues for, say, 5 minutes or until no new contracts are forthcoming. The first market period ends, and the market opens for a new trading period. There are many software programs allowing these rules to govern trading via computer terminals.

1.3.3 Behavior. With M, $N = 3$ or more, the sequence of contract prices tends to be volatile at first, then converge, within a few trading periods, to near the competitive equilibrium price and quantity that equalizes the quantities demanded and supplied (Smith 1991, Davis and Holt 1993, Chap. 3). But in Fig. 2 the subjects (Department of Energy administrators) were especially good traders, and prices stayed very close to the P_E interval during the first period. The contract prices for each trade are plotted as dots in the order of their occurrence: the first at \$12, then \$12.40, etc., the 11th and 12th at \$13. Although 12 units traded the market was 97 percent efficient (not shown), just short of 100 percent. Efficiency is the percentage of the shaded area that is collected as gains from exchange (profit) in any period. What was the source of inefficiency? Buyer B2 bought his third unit (the 13th market demand unit in order; value \$12.30) from one of the lower cost sellers at the price \$12.20 on the 9th contract as plotted. If a demand unit below P_E trades, this displaces one of the higher value demand units that must trade if the market is to be 100 percent efficient. Full efficiency is achieved only if the highest 12 demand units are satisfied by the 12 lowest cost supply units, with all other units not trading. Otherwise, further gains from exchange can be realized. In this case buyer B2 could resell his unit to the higher value buyer he displaced, yielding a profit for each. In the experiment shown, no provision was made for such retrading.

Double auction markets are studied in Davis and Holt (1993, pp. 125–72), Friedman and Sunder (1994, pp. 18–20), Kagel and Roth (1995, pp. 368–72), and Smith (1991, Parts I and II).

2. *What Have we Learned from Experiments?*

2.1 Personal vs. Impersonal Exchange: Reconciling Cooperative and Noncooperative Behavior

The results reported in the trust game of example 1.1 above contrast with those in the single unit auction, example 1.2, and the double auction example 1.3, posing an anomaly. Why do half of the subjects choose cooperative strategies in anonymous two-person games, such as Fig. 1, while, universally, subjects behave noncooperatively in anonymous interaction first price auctions and in double auctions (Fig. 2)?

To see the problem, consider the predictions of theory in the game of Fig. 1. Agents are assumed to be self-interested maximizers who choose dominant strategies, and who assume that any paired counterpart does the same. Thus, if player 2 gets a chance to move, she receives \$40 by moving down, \$25 by moving right. Player 1, expecting player 2 to choose her dominant strategy (\$40), will not play down. Thus, the noncooperative equilibrium (NE) of the game is for player 1 to move right yielding (\$10, \$10).

An alternative theory, prominent in social and evolutionary psychology, postulates that there are reciprocity types in the population who follow reputational lifestyles in social interactions, normally cooperating with other like persons. They also reciprocate when others offer cooperation to them. All social exchange among friends, associates, and in-group members involves trading favors, services, and gifts, with deep roots across all cultures including hunter-gathers. The sense of indebtedness for a favor is expressed in the common phrase, 'I owe you one.' For such types a move down by DM1, followed by a move right by DM2, constitutes an exchange. Each incurs a potential opportunity loss (Player 1, \$0 − 10 = −\$10; player 2, \$25 − 40 = −\$15) for a potential gain of 50 percent for player 1 and 150 percent for player 2 relative to NE. The exchange requires trust by DM1 and trustworthiness by DM2. Experimentalists originally used anonymous matching protocols to control for such pair-bonding. Therefore one of the

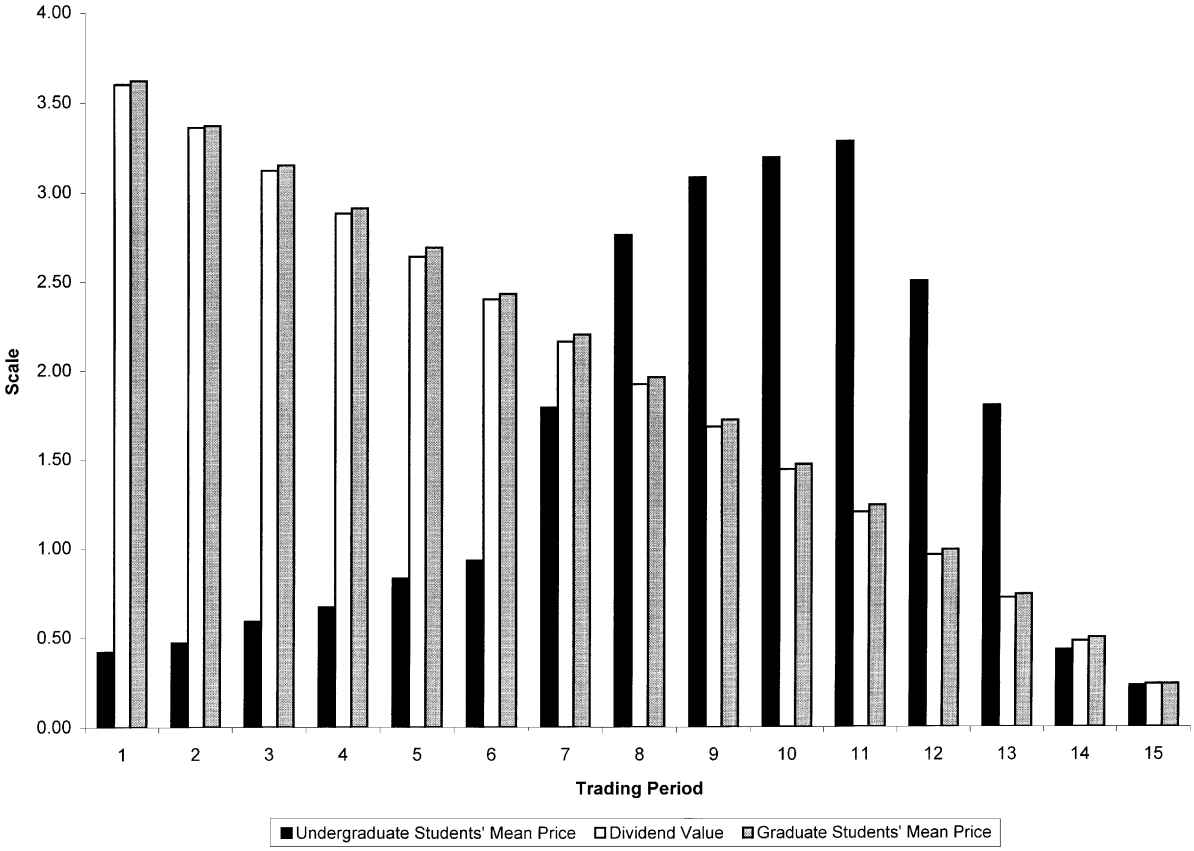

Figure 3
Dividend (fundamental) value: mean double auction prices by period for two subject groups

important discoveries was that this protocol controls very poorly for such phenomena.

Even graduate students in economics with knowledge and training in game theory do not predominantly apply the NE model in Fig. 1. One interpretation is that the reciprocity model is so strong and innate that it overcomes the effect of such training.

Thus, in an important sense, reciprocity behavior in Fig. 1, and noncooperative behavior in experimental markets, such as Fig. 2, are not anomalous: both achieve efficient–maximum gain—outcomes. In the former, an example of personal exchange, coordination on efficient outcomes, requires shared attention, intentionality detection, and 'mindreading' (Baron-Cohen 1995) by reciprocity types; in impersonal exchange in markets, coordination through a pricing institution achieves efficient outcomes which are not part of the intention of agents who, noncooperatively, seek only their own gain. One cannot speak of an identifiable 'other mind' to be read. In Smith (1998) it is argued that reciprocity, in the extended family and the tribal societies of our hunter-gatherer ancestors, would have laid the exchange foundation for specialization and wealth creation millennia before the emergence of long-distance exchange based on barter, and, ultimately, something that would be called 'money.' (Also see North 1990 on personal vs. impersonal exchange and the emergence of long-distance trade.)

2.2 Institutions Matter

The rules governing social and economic interactions matter because they affect the state of agent information, perception, and incentives, which in turn affect decisions. Because information on value (cost) is dispersed among agents the frequency and content of the messages exchanged are important elements in equilibrium price discovery. This is demonstrated at the elementary level in single-object auctions. In the English 'open outcry' auction (E), if each bidder knows only his own value, then after the initial bid

each bidder has an incentive to raise the standing last bid if it is below the bidder's valuation for the item, but never to raise his own standing bid. Consequently, the bids increase until the highest value bidder eventually raises the standing bid until it is equal to or just above the second highest value bidder, and he or she is awarded the item at the price equal to the last bid. The award is thus efficient, and experiments tend to confirm the high efficiency and price predictions of this theory. The E auction, however, is vulnerable to collusion, since a subset of friends/associates may refrain (reciprocally) from raising each other's bids even in the absence of formal agreement. In the absence of collusion, no one is motivated to invest in information about the values or bidding strategies of others (which would be socially wasteful) because bids and withdrawals are public information, although the identity of those taking the actions may not be public.

In the first price sealed bid auction (F), the high bidder also wins and pays his bid price, but each bidder bids once, blind to the bids of others. In F, bids are also related to value, but also to what a bidder may know, guess or believe about the valuations of others, and, theoretically, to his own attitude toward risk. Information on the bids or values of others is informative, and bidders are motivated to invest in acquiring such information. Knowing the highest bid of your competitors enables you to win with a slightly higher bid, and leave no money on the table.

A variation of the F auction is the second price sealed bid auction (S), in which the highest bidder wins but pays the seller a price equal to the second highest bid submitted. Whoever wins is put in the position of being a price-taker, i.e. the amount of the winning bid determines the probability of winning but has no effect on the profit from bidding. Therefore it is a dominant strategy to bid your value, thus maximizing expected profit; if all do this, the high value bidder wins at a price equal to the second high value. This logical characteristic of S auctions is not immediately transparent to unsophisticated bidders, although many learn over time to bid their value without necessarily being able to articulate why.

In the Dutch auction (D) the price starts high and descends until the first bidder stops the clock (in Dutch flower auctions) or yells 'mine' (in Spanish fish auctions): the auction is over with the first (public) bid. It is a theorem that the D and F auctions are isomorphic: it is rational to stop the clock at the price you would bid in the F auction. Behaviorally, in D auction experiments, on average, subjects tend to let the clock run below the level at which they would bid in an F auction (Cox et al. 1983 in Smith 1991, pp. 580–94). Note that collusion is difficult in D auctions because no one knows what might be the highest of your competitor's bids because it is never observed.

Theoretically, under the Vickrey assumptions (e.g., risk-neutral bidders), the above auctions all yield the same efficient allocations and expected price. With risk averse bidders, the E and S auctions are still equivalent and efficient, the D and F auctions are equivalent, but not efficient, and yield higher average prices than E and S. This is an example in which the environment interacts with the institution: with risk aversion bidders' behavior in D and F diverges from that in E and S. Behaviorally, in lab experiments, E and S auctions are roughly equivalent, and highly efficient, with comparable average prices. The D and F auctions are not equivalent; F prices are higher, and have higher efficiency: in both D and F bidders bid as if they were predominantly risk averse. (For surveys see Davis and Holt 1993, Friedman and Sunder 1994, Kagel and Roth 1995, and Smith 1991.)

2.3 Stock Market Bubbles are Common in the Laboratory

Finance theory teaches that the value of an equity share is determined by its fundamental value: the expected discounted value of its future yield (or dividends). Consider a simple environment for testing this hypothesis: $N = (9 \text{ or } 12)$ subjects are each given an endowment of cash and shares; at the end of each of 15 trading periods a dividend is drawn from the four alternatives (0, \$0.08, \$0.28, \$0.60), each with a probability of one-quarter of yielding an expected dividend of \$0.24 per period. Hence, during the first period each individual has a right to 15 draws from this dividend distribution, so that fundamental value is $15 \times (\$0.24) = \$3.60$; during the second period the right is to 14 draws, and value is $14 \times (\$0.24) = \$3.36$, and so on. This information and these calculations are displayed to all subjects to guard against any misunderstanding (see Smith et al. 1988 in Smith 1991, pp. 339–71).

In Fig. 3 the white bars plot the declining theoretical (dividend) value, from \$3.60 in period 1 to \$0.24 in period 15. A typical experiment using inexperienced undergraduate subjects is indicated by the black bars that plot the average double auction contract price by period: average prices start below fundamental value, rise over time to above this value, then crash a few periods before the end, approximating the fundamental value by periods 14–15. Perhaps 100 such experiments have been reported in the literature documenting this bubble/crash behavior for inexperienced subjects, and the fact that only when a group is brought back for a third session do prices finally approximate the fundamental value. The interpretation is that common information is not sufficient to yield common fundamental value expectations; it is not enough because each person still has uncertainty about the behavior of others. It is through a common experiential process that subjects resolve this behavioral uncertainty and approach common expectations.

But suppose that advanced graduate students participate in this experiment. A subset of 22 subjects from the sample who participated in the trust game, whose results are depicted in Fig. 1, also participated in an asset trading experiment, yielding mean prices represented by the gray bars in Fig. 3: mean prices deviate no more than five cents from fundamental value. Hence, in this market game, advanced students in economics are rational, and expect others in their group to be rational. Common information and knowledge of others' types is sufficient to yield common rational expectations. For a comprehensive survey of asset trading, see Sunder, pp. 445–500 in Kagel and Roth 1995).

2.4 Rewards and Decision Costs Matter

The examples of environments in Sect. 1 all use salient cash rewards for the subjects; i.e., payment is directly related to earnings from decisions. Subject moves in Fig. 1 have substantial payoff consequences, while in market experiments buyers are motivated to buy at low, and sellers to sell at high, prices (see Davis and Holt 1993, pp. 24–26). There are many studies over the past 40 years showing that reward level matters. There are also cases showing that increased rewards do not improve performance (Smith and Walker 1993). Even when the central tendencies of the data support theoretical predictions, increased reward can cause the variance of the observations to decline as the data clusters more closely around theoretical predictions. Where the data do not support the predictions of a theory, increased payoffs may cause the data to move closer to those predictions, although statistically the predictive hypothesis is still rejected (see Smith and Walker 1993 for a survey). Psychologists sometimes interpret this to mean that rewards have 'little or no effect.' (Certainly the 'effect' may not rescue the theory.) Alternatively, when theory fails, a common critique from economists is that the payoffs, in particular the payoff deviations from the predictions, were inadequate (Kagel and Roth 1995, pp. 66–7). This is the methodological problem of flat maxima: even with substantial rewards, as in an auction experiment, bidding above or below the optimum may reduce payoff by only a small amount.

Notice that the findings that the variance, or distance, of the observations from the predictions declines with rewards, and the interpretations, that the rewards have little effect, or change only slightly with deviations from optimality, are all extra theoretical issues. The theory predicts the unique optimum, with zero deviation/variance, however gently rounded is the payoff function. Thus, the *ex post hoc* claim that the opportunity cost of deviating from optimality is 'too low' is implicitly a claim that other resource costs, pertinent to a decision, are not included in the theory. If the rewards are adequate, then such costs would be overcome and the decision would be closer to the theory. Economic theories of optimal behavior ignore decision costs—the costs of thinking, calculating, attention, monitoring, deciding, etc. Stakes are not a formal part of the theory, and payoff adequacy tends to surface only as an afterthought when a theory performs poorly in tests.

This problem illuminates the key methodological issue in any experiment: an experiment is a joint test of a theoretical hypothesis and everything you choose to do (procedures, relative and level of payoffs, information provided, etc.) to implement the test. This is why, whenever a theory fails replicable testing, it is most important to re-evaluate both the theory and the testing apparatus (laboratory or field) as a system. Otherwise, one can always interpret the test as rejecting the auxiliary hypotheses used in the test, and not the theory: every time you double the stakes, the test either succeeds or fails; if it fails one concludes that the stakes were inadequate; if it succeeds one concludes that the theory is correct. This defines a process in which the theory cannot be falsified. Yet in traditional economics all basic theory development tends to abstract from the issues that arise in implementing a test. In the above example, the pattern of observed deviations from predictions begs for some formal extension of theory that takes rewards into account. (For a discussion of methodology in experimental economics, see Smith 1991, pp. 783–801.)

We briefly outline a simple such extension based on a model of subjects' decision cost (Smith and Szidarovszky 1999). Consider an interactive game among N decision-makers, with payoffs $\pi_i(y_1,\ldots, y_N)$ given the decision choices, y_i, for each i. Let $y_i = x_i^*$ define a Nash equilibrium in the standard paradigm of unbounded, costless, rationality. Now let $y_i = x_i^* - \varepsilon_i(z_i)$ where $\varepsilon_i(z_i)$ is agent i's error deviation from equilibrium, depending on the cognitive effort, $z_i \geqslant 0$. More effort is productive, implying that $\varepsilon_i(z_i)$ is decreasing in z_i. But reduced error increases payoff. From the perspective of the decision-maker, reward net of cognitive cost, $C_i(z_i)$, is $\pi_i(x_1^* - \varepsilon_1(z_1),\ldots, x_i^* - \varepsilon_i(z_i),\ldots, x_N^* - \varepsilon_N(z_N))$-$C_i(z_i)$. This defines a modified equilibrium in which effort cost is balanced against its value in reducing error. For optimal $z_i > 0$, we get the testable implication that increasing payoffs reduces average error, and the variance of error.

This model also explains what can go wrong when increasing the stakes has an insignificant effect on performance, as is often found across the range of experiments: the decision task has very high effort cost and/or effort has low productivity in reducing error. Then, decision costs may dominate the effect of larger rewards. Also it can be seen why subject experience might be as, or more, important than reward level. Experience allows decisions to become routine, with reduced attentional effort. Learning initially to play a Bach sonata may be arduous even for a skilled pianist,

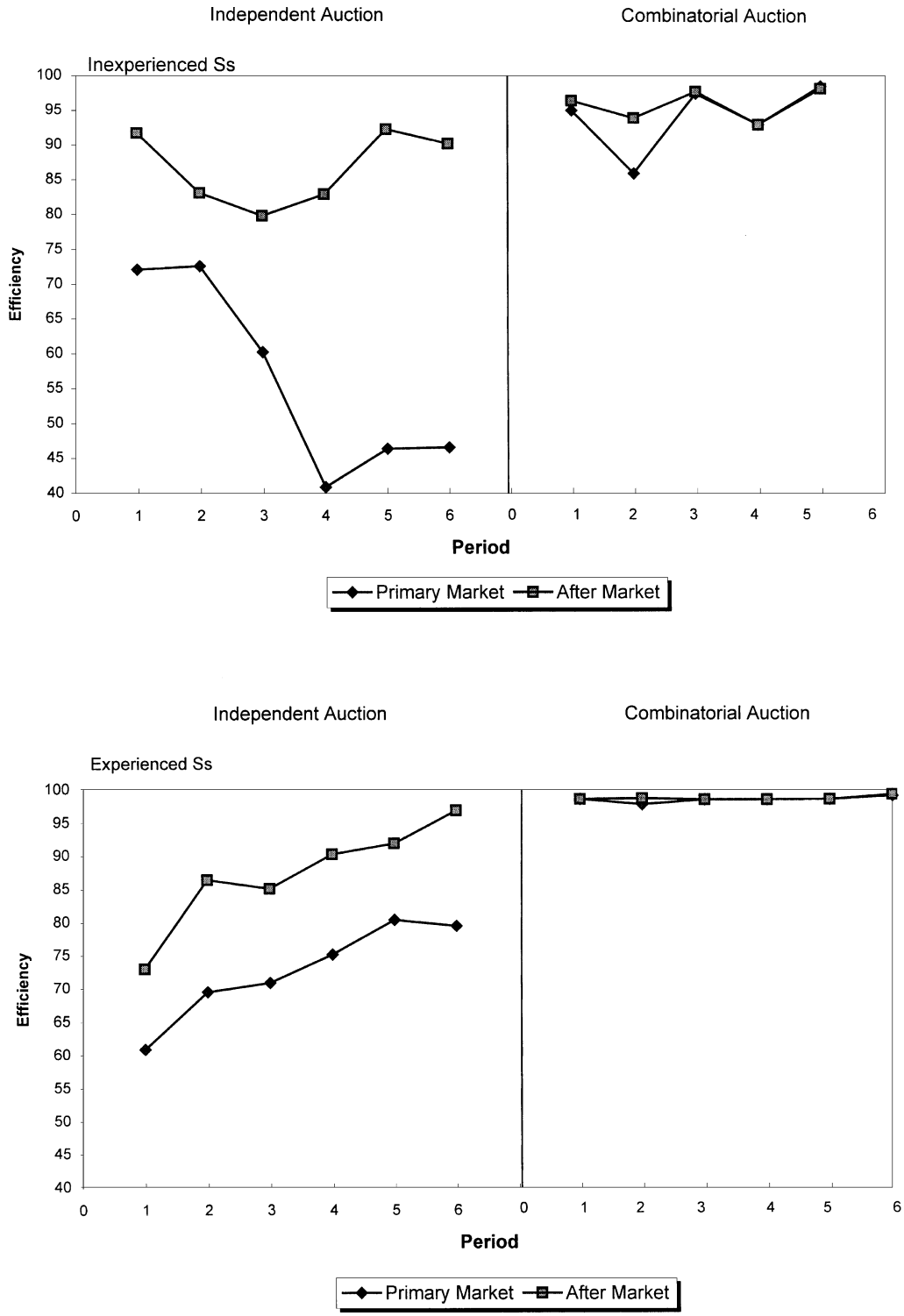

Figure 4
Independent auction compared with combinatorial auction in primary market and in subsequent after market trading for inexperienced and experienced subjects

who performs with little effort and without conscious thought once it is mastered. Thus, rewards may sometimes be more effective for inexperienced than experienced subjects (see Jamal and Sunder 1991, cited in Friedman and Sunder 1994, p. 220).

3. How Can Experiments be Applied in the Field?

Beginning over 30 years ago experimental methods have been employed to study a wide variety of phenomena with direct application outside the laboratory (Smith 1991, pp. 511–14). The worldwide privatization/deregulation movement over this period has created a great demand for 'test bedding,' the use of the laboratory to test the performance properties of new kinds of markets and to train participants in the field. (For some discussion and further references see Davis and Holt 1993, pp. 296–304, Kagel and Roth 1995, pp. 119–21, Friedman and Sunder 1994, pp. 32–7, Smith 1991, pp. 662–702, Rassenti and Smith 1998).

Thus, laboratory experiments have been used to inform the design of pollution credit trading, natural gas production, delivery and transmission, spot markets in electricity, and recently the auction of rights to the spectrum.

These constitute examples of smart computer assisted markets, in which optimization algorithms are applied to the decentralized messages submitted by agents. Such systems combine the information advantages of decentralized private property rights and ownership with the coordination advantages of central processing. The control center is not a decision-making unit; it merely applies coordinating, allocation, and pricing algorithms to the messages of decentralized owners. The center itself can be organized as a rule-governed joint venture of the private users in which entry is required to be open for new users who invest in the expansion of capacity. The first smart computer assisted market was a combinatorial auction that selected combinations of elemental items, valuable only in packages, that were responsive to the package bids of independent agents. (see Rassenti 1981, cited in Smith 1991, p. 677, Rassenti, Smith and Bulfin 1982, cited in Smith 1991, pp. 662–77). Fig. 4 compares the efficiency of a combinatorial auction, in the panels on the right, with an independent auction of the components of the multiple discrete units that comprise the packages.

Each auction period consists of two markets, a primary market represented by either the independent (uniform price sealed bid) auction or the combinatorial auction of packages, followed by an open book after market allowing traders to post bids to buy or sell for elements or combinations of elements. The after market enables bidders to obtain valued elements missing from packages they attempted to piece together in the independent auction, or packages they failed to obtain in the combinatorial auction. Six elementary resources were available, each in fixed supply. Six bidders each had induced values for four, five, or six packages of the six elements. For example, a bidder had a value of 887 for the package (1, 0, 1, 0, 1, 0) consisting of one unit each of resources A, C, and E. If all three elements are obtained, the bidder receives 887. If any element is missing the value is zero. The upper panel results are for two groups of inexperienced subjects, while the lower panel uses experienced subjects. With inexperienced subjects, comparing the results at the end of each primary market for the independent and combinatorial auctions, efficiency in the latter was larger than that in the former in every period. The same holds for the after market comparisons, except the improvement from primary to after market is much greater in the independent auction. This is because almost all of the high efficiencies in the combinatorial auction were achieved in the primary auction, which, for experienced subjects, is 98–99 percent efficient.

See also: Economic History, Quantitative: United States; Experimental Design: Overview; Game Theory; Game Theory: Noncooperative Games; Institutional Economic Thought; Market and Nonmarket Allocation

Bibliography

Baron-Cohen S 1995 *Mindblindness*. MIT Press, Cambridge, MA

Davis D D, Holt C A 1993 *Experimental Economics*. Princeton University Press, Princeton, NJ

Friedman D, Sunder S 1994 *Experimental Methods*. Cambridge University Press, Cambridge, UK

Kagel J H, Roth A E 1995 *Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ

McCabe K, Rassenti S, Smith V 1996 Game theory and reciprocity in some extensive form experimental games. *Proceedings of the National Academy of Sciences of the USA* **93**: 13421–8

North D C 1990 *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, UK

Rassenti S, Smith V 1998 Deregulating electric power: market design issues and experiments. In: Chao H, Hunting H G (eds.) *Designing Competitive Electricity Markets*. Kluwer, Boston

Smith V 1991 *Papers in Experimental Economics*. Cambridge University Press, Cambridge, UK

Smith V L 1998 The two faces of Adam Smith. *Southern Economic Journal* **65**: 2–19

Smith V, Szidarovszky F 1999 Monetary rewards and decision cost in strategic intentions. Unpublished manuscript

Smith V, Walker J M 1993 Monetary rewards and decision cost in experimental economics. *Economic Inquiry* **35**: 245–61

V. L. Smith

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Experimental Laboratories: Biobehavioral

Until the nineteenth century, research on human and animal behavior was conducted and financed by individuals exhibiting a personal rather than a professional interest in these topics. In the human domain, educated persons documented spectacular actions, habits, and customs from various cultures, and ventured systematic inquiries to explore normal and abnormal behavior in individuals and populations. These inquiries included single case studies (e.g., diaries on infant behavior), and compilations of statistical data (e.g., crime rates). In the animal domain, naturalists observed the endemic fauna and joined expeditions to study the wildlife abroad. Eventually, experimental methods were introduced in physiology, psychology, and biology. Cabinets in private homes served for experiments with humans. Also within the confines of private estates, animals were kept for experimental purposes in cages and yards.

Since the nineteenth century, medical and natural science research laboratories were established at universities. The Departments of Physiology at the German Universities of Leipzig (inaugurated in 1865) and Berlin (1877) were milestones for experimental research in the life sciences. Shortly thereafter, psychological laboratories were established to conduct experiments on humans and animals. Sahakian (1975, p. 138) lists 68 psychological laboratories that were founded in Europe, North America, and South America by the end of the nineteenth century. Among the first and most noteworthy were the research facilities of the German universities of Leipzig (established in 1880) and Göttingen (1881), and, in the USA, at Harvard University (approximately 1890). Most experiments on humans in psychological laboratories were conducted with single individuals. However, there were also experimental approaches addressing the behavior and cognitions of groups, or of individuals within groups.

Zoos, primarily established for the public, gave further opportunities for scientific animal observation. However, the city zoos provided only narrow and artificial habitats. Therefore, animal stations were set up to promote research in more natural settings. The controversies spawned by evolution theory led to the establishment of primate centers whose goal was to investigate the intelligence of apes. The most renowned research facilities were the primate stations in Tenerife, Spain (1912), and in Orange Park, Florida (1930).

Recent progress in computer technology not only revolutionizes experimental techniques, but also permits to overcome the spatial limits of the traditional laboratory. Work places, architectural sites, political and economic situations can be simulated, and the behavior of humans can be studied in virtual settings. Moreover, experimental tasks are presented in the Internet, and responses are collected from worldwide samples of participants.

Given this variety of institutions for behavior research, four types of laboratories will be examined: (a) Psychological laboratories for experiments with individual humans, (b) psychological laboratories for small animals, (c) animal stations, and (d) experiments with groups of humans. In addition, (e) the present tendencies of collecting experimental data via the Internet are outlined.

1. Psychological Laboratories for Experiments with Single Humans

Psychological laboratories are equipped for two purposes. Firstly, for the recording of behavior and for the monitoring of mental and emotional states; secondly, for the systematic control of environmental and individual factors which influence and determine human mental and emotional states. Experiments are designed to relate measures of behavior and internal states as dependent variables to preceding or concomitant environmental and individual factors as independent variables. Typically, these kinds of experiments are conducted with individuals. The data from these samples are then accumulated over successive experimental sessions.

As one class of dependent variables, behavioral data primarily comprise speed and quality of performance in achievement tasks. The achievements tested include learning (e.g., reproduction of texts), perception (e.g., estimating distances), problem solving (e.g., finding the shortest route in a labyrinth), and sensorimotor actions (e.g., reacting to a light signal). Besides assessing performance, behavioral records also serve as indications for habitual or transient mental and emotional states (e.g., fatigue, positive affect, and surprise). An example is observations of facial expressions indicating emotional reactions. Moreover, behavioral records reflect stable personality traits (e.g., self-efficacy, anxiety, and withdrawal).

Another class of dependent variables are self-reports. Self-reports of participants express judgments and mental processes (e.g., logical reasoning), mental structures (e.g., impressions of paintings), and emotional states (e.g., optimism). A third class of dependent variables is psycho- and neurophysiological data. Measures of the autonomic nervous system (e.g., cardiovascular and gastrointestinal activity) and the endocrine system (e.g., adrenaline and cortisol secretion) serve as evidence for stable and transient states. More directly related to cognitive and sensorimotor processes are cerebral functions (e.g., event-related brain potentials).

Several unobtrusive methods exist for collecting experimental data. These include free protocols, video and voice recordings. However, most methods are obtrusive, for instance performance tests, which entail

the operation of technical devices (e.g., reaction keys for measuring the speed of manual reactions). The obtrusive methods include various kinds of psycho- and neurophysiological recordings where participants are placed in or connected with special appliances ranging from blood pressure meters to magnetic resonance tomographs.

The instructions and tasks provided for the participants are the crucial aspects of the independent variables. These tasks may be tests of memory and learning, or of cognitive and motor skills. They may initiate logical judgments and actions, and may arouse emotions. Instructions not only serve to explain the requirements of the experimental tasks, but are also designed to affect the participants' motivation, attitudes, and moods. A plethora of devices exists for various experimental designs such as the presentation of visual or acoustic stimuli and for the control of environmental variations, which permit to study learning, movement, communication, and self-regulation. The appliances should guarantee precise and stable conditions over repeated trials. They also enable the scientist to control the variations in experimental conditions, such as the time of exposure to visual stimuli as measured in milliseconds.

The establishment of psychological laboratories for humans benefited from growing expertise in pertinent experimental techniques and the feasibility of exact measurements. Many experimental investigations depended on the progress made in precision mechanics. In the early stages of experimental psychology, most research laboratories employed a mechanical engineer. The advent of electronic elements also revolutionized experimentation, and has brought electronic engineers into the psychological laboratories. Increasing demands and standardization requirements permitted the commercial production of the devices used in psychological research. Thereafter, more and more technical research instruments were purchased and maintained by commercial companies, and the technical work force within the research laboratories diminished. This process continued when laboratory computers became available which could be adapted to various experimental purposes by specialized software.

For many decades, most psychological laboratories for humans were part of the university departments of psychology. Most of them were rather small, consisted of a few rooms, and were located alongside offices and classrooms. Some were insulated for sound or light protection to permit better control over sensory stimulation. Some were electrically shielded to enable electrophysiological data collection. Laboratories in psychological departments were rarely set up as central, permanent research units. Neither were they equipped for a variety of research tasks. Most of them served a specialized form of research (e.g., perception psychology). They were maintained by individual researchers and small work groups, and were often transferred to other purposes once a project or a contract was terminated. However, psychology departments had a policy of employing at least one experimentalist in their staff who raised funds and organized a laboratory for scientific research and academic training.

As brain research gained in recognition, special institutions of neuroscience with a high emphasis on medical and biological expertise have established laboratories for the study of human cognition and behavior. A similar increase in number and scientific impact has occurred in regard to research institutes devoted to cognitive science, computer science, and multimedia applications. The newer institutions operate on long-term projects, and are often sponsored by corporations rather than individuals. Consequently, they tend to outperform university departments of psychology in attracting funds for more spacious laboratories with costly equipment such as high-resolution magnetic resonance imaging and high technology environments.

2. *Psychological Laboratories for Small Animals*

Two issues have dominated in laboratory research with small animals: the laws governing the acquisition of behavior, and the neural basis of behavior. The first issue was primarily investigated in learning studies which varied environmental stimuli eliciting responses, drive states, the reinforcement of behavior (i.e., reward and punishment), and the role of practice. This research has promoted special training devices such as mazes and boxes with food dispensers, electric grids, etc., where animals could operate buttons or change locations to obtain rewards or to avoid punishments. In neuropsychological experiments the effects of brain lesions and the patterns of electrical brain waves have been investigated. Functional centers of an animal's brain were surgically removed, or nervous pathways were disconnected. After the animal had recovered from surgery, its capacity for learning, discrimination, etc., was tested in order to ascertain the impact of the lesions. Alternatively, electrodes were implanted in functional centers, and subsequently the activation of various brain areas was determined during maze exploration, water intake, etc. Advances in research technology have also revolutionized the work in the animal laboratory. Above all, new techniques for monitoring ongoing nervous processes and for analyzing the structure of nervous tissues have facilitated the understanding of the neural basis of behavior. Moreover, the influence of endocrine secretions on behavior has been extensively studied.

These investigations were primarily performed with rodents (mainly mice and rats) and birds (mainly pigeons), and sometimes with domesticated animals such as cats and dogs, or small fish. Animals of these

species are fairly easy to handle. They can be kept in separate cages and aquariums, and moved to the experimental locations. They are small enough to fit into a table-size apparatus, and large enough for precise surgery and observation. Moreover, their learning history, their diet, their environment, and other vital factors can be controlled experimentally. After the completion of the experiment they can be sacrificed in order to verify the location of lesions and electrodes.

Engineers associated with the laboratory first manufactured the technical devices employed in the training and testing of small animals. In recent decades, they have been produced commercially. Cages and surgical instruments had always been available from medical and biological research. Special locations outside the actual laboratory were often reserved for surgical procedures, microscopy, etc.

Until 1970, animal labs for the study of learning and other psychological functions were part of psychology and biology departments. Medical schools maintained such units for brain research. Typically, these laboratories were located near offices and classrooms. In biology and medicine rather than in psychology, animal laboratories were established as permanent institutions, parallel to other animal labs which investigated basic physiological functions, pathologies, and pharmacological treatments. Caretakers, who feed the animals, clean the cages, etc., complement the scientific and technical staff.

Apparently, small animals often served as substitutes for humans, since they were easier to control, and, above all, less protected by ethical and legal norms. Since 1970, both the public concern and the legislation regarding the protection of animals have militated against invasive and severely restrictive measures in animal experiments. In consequence, laboratories for small animals have diminished in number and size.

3. *Animal Stations*

Animal stations may consist of several departments, which work on different research projects. In the biomedical branch, there is an emphasis on immunology, virology, and pharmacology, as well as on surgical and post-surgical treatment. In the biogenetic branch, reproduction is the primary object of investigation. From the psychological point of view, studies on intelligence, communication, and social organization are most salient. Regarding intelligence, many studies have focused on the problem solving abilities of animals (e.g., the construction and use of tools). Communication includes vocal messages, gestures, and bodily, most notably facial expressions. In going beyond their natural repertoire, some animals have been trained to use a symbolic language (e.g., American Sign Language as used by the deaf). In the social domain, hierarchies, and roles (e.g., leadership and subordination in herds, selection of partners in mating), helping behavior and aggression (e.g., sharing and defending a territory, social support), and their impact on health (psychosocial stress) are major research topics. Due to recent methodological advances, behavioral studies are often performed in collaboration with brain research. Moreover, intriguing new findings have deepened our understanding of the genetic basis of animal behavior.

Traditionally, monkeys (orangutans, chimpanzees, macaques, etc.) have been the preferred species in animal stations because of their genetic resemblance to humans. Primates exhibit high intelligence in instrumental tasks and develop considerable conceptual and linguistic skills when trained in sign language. They build families and groups with stable social structures. However, monkeys are unable to overcome their natural limitations regarding articulation. Besides primates, many other species are investigated scientifically. For instance, some researchers study how singing birds communicate over long distances in elaborated vocal codes. Also highly popular among scientists are dolphins.

Animal stations are more spacious than the laboratories for small animals (see Sect. 2). Rather than being transported from a narrow cage to an experimental device, the animals live and are studied in larger cages or in fenced natural resorts. The stations are typically located outside of educational centers. If associated with a zoo, however, they may be close to a city. An example is the Animal Behavior Research Team, a joint organization of West Chester University and the Philadelphia zoo. Although the animals are held in captivity, their living conditions bear a close resemblance to the animals' natural habitats. Some centers combine research on captive animals with wildlife observations. An example is the Kewalo Basin Marine Mammal Laboratory of the University of Hawaii. This institute studies whales in the Hawaiian waters as well as the intellectual potential of dolphins housed in a basin.

In terms of budget, facilities, and staff size the larger animal stations have achieved a high status among scientific institutions. The Yerkes Primate Research Center in Atlanta, Georgia, maintains a permanent staff of 50 scientists who work in collaboration with more than 100 visiting researchers. The Primate Center in Göttingen, Germany, employs a staff of more than 70 scientists. An equal or even greater number of technical and administrative staff members, including caretakers, are employed by these respective organizations. Animal stations are funded on a long-term basis by government agencies and private foundations. Moreover, they raise funds by short-term grants and contracts, and they also receive personal donations. The first and prototypical stations were financed by private foundations: in Europe, the German Samson

foundation paid for the Primate Station at Tenerife, Spain; in the USA, the Carnegie and the Rockefeller foundations, respectively, enabled the establishment of Yale University's Primate Center in Orange Park, Florida. After World War II, national institutions assumed responsibility for the larger stations. For instance, the Primate Center in Göttingen, Germany, is organized as a private company, subsidized by the German Federal Government and the Federal State of Lower Saxony. The Primate Research Center in Orange Park, Florida, named after its founder Robert Yerkes, was sold for 1$ by Yale University to Emory University in 1956. In 1965, the center moved to Atlanta, Georgia. Now it is among the eight institutions sponsored by the American National Institute of Health within its primate research program.

4. *Experiments with Groups of Humans*

Experiments with groups of humans serve the purpose of demonstrating principles and testing propositions of the social sciences. Popular paradigms include: the individual in the social context (i.e., changes in achievement, creativity, attitudes, etc., in the presence of other individuals and social models, under social pressure, etc.), group organization (e.g., leadership, division of labor, communication structure); group performance (e.g., problem solving, decision making), cooperation and tension in groups (e.g., group coherence, morale, deviant behavior). In order to enhance the control over experimental conditions and the assessment of individual data, the number of individuals in a group is usually limited to numbers of two to twelve.

The effect of observers, competitors, etc. on individual behavior (e.g., problem solving, learning) has been investigated in standard laboratory settings (see Sect. 1). In studying communication within groups, multi-channel systems permit the exchange of written or oral messages (e.g., computer networks). Communications and other social interactions (e.g., creative group work, aggressive behavior in dyads) are carefully observed and recorded (e.g., on videotape). In some laboratories, observers and recording devices are concealed behind a one-way mirror.

Research facilities for field experiments in social psychology were mostly provided by psychological departments. Numerous field experiments were conducted in real-life settings (e.g., in firms or youth clubs). A noteworthy model for a research unit, which organized such field experiments, was the Center of Group Dynamics, established in 1945 at the Massachusetts Institute of Technology in Boston.

In the tradition of game theory, subjects were confronted with experimental tasks simulating political and economic decision-making. Provided with an appropriate computer model of the situation (e.g., strategic bargaining, alliance formation), cooperative or competitive groups could gather pertinent information and take measures to stabilize or change the system. The impact of these measures was subsequently computed, and participants were confronted with the consequences of their actions. Thus, decision processes were examined for several consecutive periods of virtual time. Such simulation studies were mainly conducted in social science and economic departments.

5. *Globalization of Laboratory Tests*

When experimental tasks are presented by electronic media and responses are transmitted to a central processor; there is no need to restrict the experimental situation to a single workplace. Therefore, many experiments can now be conducted via the Internet. A special site for psychological experiments with humans is provided on behalf of the American Psychological Society (http://www.psych.hanover.edu/APS/exponnet.html). The main topics represented in this service are sensation and perception, learning and cognition, social psychology, psycholinguistics, and decision-making. Internet experiments attract large numbers of participants. However, the worldwide distribution of participants and the multimedia context cause some problems in psychological experimentation with humans, most notably a lack of commitment to the purpose of the experiment.

See also: Animal Cognition; Animal Rights in Research and Research Application; Behavioral Neuroscience; Biomedical Sciences and Technology: History and Sociology; Cognitive Psychology: Overview; Decision Making, Psychology of; Experiment, in Science and Technology Studies; Experimental Laboratories: Social/Psychological; Experimentation in Psychology, History of; Facial Expressions; Laboratory Experiment: Methodology; Laboratory Studies: Historical Perspectives; Motivation and Actions, Psychology of

Bibliography

Boring E G 1950 *A History of Experimental Psychology*. Appleton-Century-Crofts, New York

Cartwright D, Zander A (eds.) 1953 *Group Dynamics*. Row and Petersen, Evanston, IL

Sahakian W 1975 *History and Systems of Psychology*. Wiley, New York

Selten R 1991 *Game Equilibrium Models*, 4 Vols. Springer, Heidelberg, Germany

Vaitl D 1998 Psychophysiology in Europe. *Journal of Psychophysiology* **12**: 172–83
Yerkes R M 1943 *Chimpanzees. A Laboratory Colony*. Yale University Press, New Haven, CT

W. Schönpflug

Experimental Laboratories: Social/Psychological

What is an experimental laboratory in social psychology? A lab is defined as a place that contains equipment necessary for scientific work. Under this definition the two key words are 'equipment necessary.' In practice, the contents of a lab are dictated by the kind of research one does and the equipment necessary to facilitate that research. The social psychologist studying communication patterns within a married couple needs a different lab than a social psychologist studying the psychophysiological correlates of aggressive behavior or a social psychologist testing a behavioral implication of attribution theory. Sometimes a social psychology lab is merely a small room with a computer. As long as that space, and the equipment it houses, allows the research to be conducted, then the room can be called a lab. Much as Pygmalion carved his ideal mate out of ivory, a social psychologist creates an ideal research environment from the equipment necessary to conduct research.

1. *The Computer*

There is one piece of equipment that has become indispensable in just about all social psychology labs—the computer. At minimum, the computer facilitates the editing of manuscripts. Writing experts claim that the ability to edit and rewrite one's prose is critical for clear, crisp writing. Further, a sophisticated user of a word processing program can format a manuscript in the manner required by a journal editor and easily construct a reference list from an electronic database. An Internet connection permits electronic mail where, for instance, collaborators can exchange working drafts of articles, and authors can submit manuscripts for publication electronically as well as distribute reports to other interested parties. In this sense, the computer has replaced the typewriter as the machine used to write and is replacing the copier as the machine used to disseminate the written product of one's research.

There are, of course, more advanced uses of the computer in social psychological research. For instance, the computer can be useful in data collection. Unlike the traditional paper-and-pencil questionnaire, administering a survey by computer permits automatic tailoring of questions depending on the answers the respondent provides to previous questions. Software can help in solving and administering complicated sampling designs, and can facilitate elaborate laboratory procedures (such the presentation of research stimuli at a rapid rate or the timing of a participant's response).

Some researchers have started placing their studies on the world wide web, to complement their lab and field data collection (e.g., see http://www.social-psychology.org/expts.htm. Though few researchers allow anyone access to their study (for one reason, it would not be clear how to establish the relevant population to generalize the research findings), the world wide web makes it convenient to collect data at multiple sites from any number of computer platforms. This permits data collection anywhere in the world at relatively little expense.

Computers are also used to analyze data. A simple spreadsheet program can be used for descriptive statistics and simple inferential tests, whereas specialized software is needed for more computationally intensive data analytic techniques such as bootstrapping and empirical bayes algorithms. Even the analysis of some forms of qualitative data (such as narratives) can be facilitated through the use of special software programs designed to convert the original data into a theoretically linked database to facilitate analysis.

The computer also aids lab management. There exists planning software that helps organize interdependent parts of a research project on to a time line. Such time lines are useful not only for lab management but also for grant proposals. Scheduling and calendaring software can facilitate the scheduling of lab facilities and lab meetings.

How many computers should a lab contain? What kind of computer should a lab have? These questions are difficult to answer in the abstract because so much depends on the research question. For example, one project may involve 200 hand-held computers because the study is on 200 married couples who answer questions each evening at home for a month directly on the hand-held computer. These couples visit the lab weekly to download their data. On the other hand, a researcher who studies social interaction through computer simulation may only need one computer. There is obviously a wide range of needs and, as stated above, the needs of the research drive the equipment required in the lab.

2. *What is Your Research Question?*

It is useful to consider three factors when designing a lab in social psychology: the context the researcher needs to conduct the research, the manipulations the researcher will perform, and the measurements the

researcher will make. Each of these three issues are subsumed by the following more general question: what is your research question? If your research question is how Mayans conceptualize kinship terms, then you would need a special setting, special materials, and special measures.

2.1 What Context Does the Research Require?

What context does the research require? For example, if one wants to study social behaviors in a bar, then it might be useful to design a lab setting that resembles a bar (complete with mood music, barstools, and liquor cabinets). If one wants to study jury deliberation, then it is useful to design a lab setting that resembles an actual jury deliberation. These labs can be constructed to reproduce, or mimic, a real-world setting, and hence increase the external validity of the research.

However, it is also possible to study social behavior in an artificial setting, an analog setting that evokes similar psychological processes as the real-world setting but is easier to create and easier to control. An example is the classic Milgram (1975) study on obedience. Milgram claimed that his obedience findings could help us understand how citizens in a totalitarian government may be pressured to follow superiors, but his research setting was actually low in external validity with respect to totalitarian governments. Instead, Milgram's research setting was successful at evoking the psychological processes that make it difficult to refuse requests or refuse to follow orders. Milgram's lab setting involved an elaborate stage production that the experiment was about the effects of punishment on learning. However, the actual purpose of the study was to examine the social interaction between the participant (who believed he was playing the role of the 'teacher' in the study on learning) and the experimenter. Thus, a carefully constructed artificial setting still allows one to study real psychological variables, perhaps in a more controlled way, but the price one pays is a reduction in external validity.

Sometimes the researcher must assume the role of a stage manager in a theatrical company. The researcher attempts to create the setting that produces the reaction in the participant required by the research question. This may frequently include the use of stage props, as in the Milgram study that used a machine to administer (bogus) electric shock to the 'learner.' The role of stage manager should not be underestimated. Social psychology experiments frequently are criticized for being low in external validity, but a study that engages the participant in the task is high in another important feature–mundane realism (see Aronson et al. 1985 for a discussion of validity and mundane realism).

Some research questions can be studied in a relatively simple setting. For instance, research investigating people's facial reactions to disturbing photographs would need a means for displaying the pictures (a slide projector, a computer, a photo album), so an appropriately lit room would be a sufficient setting.

2.2 Tools for the Manipulation

One of the defining features of an experiment is that it includes a manipulation, also known as an independent variable. The goal is to create at least two conditions that differ only on the particular feature that is manipulated. Random assignment to condition facilitates this goal.

Laboratory equipment may be needed to conduct the manipulation. For instance, in work on eyewitness identification the research question may require that the composition of the lineup be varied to include foils that are either high in physical similarity to the suspect or low in physical similarity. Thus, if the researcher uses a photolineup, then the lab should include a broad range of photographs so that many different lineups could be constructed.

A computer can be useful in carrying out experimental manipulations. For instance, if the manipulation requires varying the duration a stimulus is presented, then the computer can be used to control such durations. Experimental stimuli may involve presentation of video or animation, and a computer can facilitate the creation, editing, and presentation of such material.

Sometimes the manipulation requires the use of drugs. For instance, research on arousal may require the participant in one condition to receive a dose of epinephrine whereas a participant in another condition to receive a dose of a neutral, saline solution. Such research requires equipment necessary to store and administer the drugs, and in some cases appropriate medical professionals on site.

2.3 Measurement Devices

An important feature of empirical work is that something is measured. Even qualitative research has measurement, such as the recording of a narrative. The lab must include the tools necessary to make such measurements.

Standard measurement devices include an audio recorder, a video recorder, a computer with interactive software, and pencils for participants to write responses. For some research questions it is necessary to measure the electrophysiology of the brain or autonomic nervous system, which requires specialized equipment to make the recordings, as well as to store and analyze samples (e.g., measurement of cortisol level).

Some research questions require the use of standardized scales (such as personality tests, intelligence tests,

scales to measure depression). Some of these scales are copyrighted, so researchers must contact the developers of the scale for permission to use. Again, which scales are needed in the lab depend on the research question. If the research question involves examining the relation between intelligence and aggressive behavior, then scales to assess intelligence would need to be on hand.

3. *Managing a Lab*

Much like large offices, some experimental labs may require a lab manager to handle the day-to-day functioning of the lab. The manager's tasks include scheduling the use of lab resources, monitoring research budgets, and staying abreast of manuscript and grant proposal deadlines. Sometimes a lab requires more specialized personnel, such as when electrophysiological recording equipment is used. Lab personnel also handle maintenance of lab equipment.

4. *Summary*

New researchers frequently ask, 'What kind of equipment should I include in my lab?' As argued in this article, one should reply with the following questions: What is your research question? What setting do you need to create to conduct your research? What equipment do you need to conduct your manipulations? What equipment do you need to make your measurements? How will you handle the mundane chores in your lab?

5. *Further Reading*

For details on how to conduct experiments and an explanation of experimental methodology see Aronson et al. (1992). For an explanation of how to analyze data from an experiment see Keppel (1982). For suggestions on how to write an article describing empirical work in social psychology see Bem (1986) and the *Publication Manual of the American Psychological Association* (see http://www.apa.org).

See also: Biomedical Sciences and Technology: History and Sociology; Experimental Laboratories: Biobehavioral; Laboratory Experiment: Methodology; Laboratory Studies: Historical Perspectives

Bibliography

Aronson E, Brewer M, Carlsmith J M 1985 Experimentation in social psychology. In: Lindzy G, Aronson E (eds.) *Handbook of Social Psychology*, 3rd edn. Random House, New York, Vol. 1, pp. 441–86

Aronson E, Ellsworth P, Carlsmith J M, Gonzales M 1992 *Methods of Research in Social Psychology*. Random House, New York

Bem D J 1986 Writing the empirical journal article. In: Zanna M, Darley J (eds.) *The Compleat Academic: A Practical Guide for the Beginning Social Scientist*. Random House, New York, pp. 171–201

Keppel G 1982 *Design and Analysis: A Researcher's Handbook*. Prentice-Hall, Englewood Cliffs, NJ

Milgram S 1975 *Obedience to Authority: An Experimental View*. Harper & Row, New York

R. Gonzalez

Experimentation in Psychology, History of

From the very beginning of their studies, psychology students in Western countries learn that the discipline specializes in quantitative research rather than 'armchair philosophy.' Methodological and statistical textbooks teach how to measure individual or group characteristics such as intelligence, neuroticism, or attitudes. Moreover, introductory and advanced textbooks spell out rules of experimentation for tracing the causes of the psychological characteristics that are measured or evaluating attempts at improving them. Standardized experimentation in particular constitutes the core of psychology's scientific identity.

The typical psychological experiment compares one or more treated experimental groups to an untreated control group. As natural groups may vary in many more respects than only the hypothesized cause or treatment to be tested, experimenters must compose their own groups equal in all respects but the factor concerned. The ideal way to create comparable groups is to assign subjects to groups on the basis of chance and thus cancel out unwanted between-group differences. From the 1940s in America and from the 1960s in Europe, generations of psychologists have been taught that the *exemplary* experiment is an experiment in which randomly composed experimental and control groups are compared. Statistical inference testing is employed for establishing the significance of any differences in the results.

1. *The Origins of the Exemplary Design*

Psychology textbooks typically discuss the experiment comparing randomly composed experimental and control groups (from now on 'the random group experiment') as being 'the' scientific experiment. They implicitly suggest or explicitly state that this form of experiment was borrowed directly from the natural sciences (more on the definition of an experiment in psychology textbooks can be found in Winston and Blais 1996). Yet, as has been demonstrated by

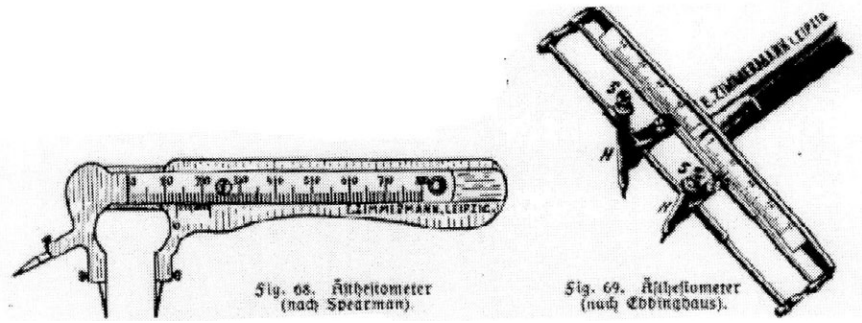

Figure 1
'Aesthesiometers' for pinpricks experiments. The one with three pins was especially designed for experiments with control trials (from Schulze 1913, p. 5)

researchers of science, in natural sciences such as physics and chemistry there are various ways of experimenting, and these are mostly learned during apprenticeship and via exemplars rather than through isolated methodology books and courses (Kuhn 1962). Moreover, comparing random groups is not a key experimental practice in such sciences (Hacking 1983, Galison 1997).

Historical research suggests that comparing randomly composed groups is firmly rooted in psychology's *own* experimental traditions, traditions which have changed together with changing social relations. The discipline's present-day definition of what constitutes an ideal experiment was brought about bit-by-bit when methodological practices from nineteenth-century 'psychophysical' experimenters trying to discover the given laws of human perception were adopted and adapted by early twentieth-century educational psychologists looking for technically useful data. These educational psychologists then passed the random group design on to other branches of psychology and social sciences (Danziger 1990, Winston and Blais 1996, Dehue 1997).

1.1 Experimentation in Nineteenth-century Psychophysics

Psychological experimentation began as an endeavor of mid-nineteenth-century gentlemen-scholars to discover the given design of human nature. The word 'experiment' could refer to all kinds of research other than naturalistic observation. Experimentation involved manipulating human sensation and perception within artificial situations. An important founder of psychological experimentation was the German scholar Gustav Fechner, who during the years 1845 to 1849 lifted his arms with a weight in each hand over 67,000 times. After each lifting, Fechner meticulously recorded whether he had been able to distinguish the heavier weight from the lighter. In this way he established the 'just noticeable difference' or 'threshold of perception' as well as the mathematical relationship between the weights as measured and the weights as perceived.

In Fechner's time, the claim to measure anything as subjective as human sensation was novel and impressive. In the second half of the nineteenth century, innumerable 'threshold experiments' were conducted in biology, physiology, and particularly the newly established psychological laboratories at European and American universities. Countless extra weights were lifted, and Fechner's methods of 'psychophysical' experimentation were employed for testing other senses as well. A popular variation was touching the skin with two pins at specific distances and trying to establish the minimum distance that can still be perceived (see Fig. 1). The different sensitivities of foreheads, chins, shoulders, arms, fingers, hips, bellies, legs, knees, and ankles were discussed. Neophytes would complain about sensitive and scaly skins and apologetically argue that some 6,000 pricks should be enough to draw a conclusion (Dehue 1997).

Like Fechner, the first psychophysical researchers acted as their own experimental subjects. Soon, however, the idea arose that the stimuli should be applied in random order, and that random control trials were needed. Researchers in Germany and subsequently the USA argued that the perception of distances might be biased by expectations on the next distance to come. Such expectations might be induced by awareness of the actual distances in previous trials or inadvertent regularities in previous series of trials. Moreover, it was argued, the sheer knowledge that there are always two pins might arouse the idea of perceiving a distance even if the actual distance is below the threshold of perception. For these reasons, it was decided that it takes *two* to experiment. Assistants were employed to apply the stimuli in random order as well as randomly carry out control trials in which the skin was touched at one point only without knowledge of the 'observers' (later on the individuals providing the stimuli would

be called the 'experimenters' and the observers became the 'subjects,' see Danziger 1990 on the social meaning of such terminological changes). These were the first experiments employing randomization as a methodological means. In the 1880s, the American scholars Charles S. Peirce and Joseph Jastrow, moreover, introduced the first *artificial* randomizers. Distrusting the assistants' ability to establish truly random orders, they used playing cards or dice (for a more extensive discussion and references, see Dehue 1997).

If the stimuli are indicated as 'S,' the responses as 'R,' and the randomly applied control trials (in which the actual double prick was withheld) as '*noS*,' then the design of these experiments was something like S-R-*noS*-R-S-R-S-R-*noS*-R-S-R-*noS*-R-S-R-S-R (etc.). In later educational experimentation, the randomly inserted control trials (*noS*) were no longer conducted with the same subjects as the experimental trials (S) but with randomly composed control *groups*. This change implied the constitution of the random group design. In order to explain why that happened, some extra information should first be given on the methods of early psychophysics as well as social changes taking place outside the laboratories.

1.2 The 'Transfer of Training' Phenomenon

Considerable concern about individual expectations and mistaken assessments became the hallmark of psychophysics. For the sake of replicability, utmost precision had to be exercised. The initiators of particular experiments extensively described the required laboratory setting, the instruments to be used, the handling of the instruments, as well as the required behaviors of assistants and observers both during the trials and in their spare time (Morawski 1988, Benschop and Draaisma 2000).

The researchers of just noticeable differences, however, were confronted with a rather obvious extra source of bias. After many trials, the *skill* of the experimental subjects improved, and this artificially lowered the threshold of perception. Switching to the other arm or leg did not solve the problem because the undesirable training seemed to transfer automatically to other parts of the body. The transfer phenomenon thwarted the experiments but also greatly intrigued the experimenters. Many of them became so fascinated by it that from a source of bias 'transfer of training' became a research topic of its own. At the end of the century, experiments were being conducted in American and European laboratories to establish the effects of training in all kinds of sensory, motor, and cognitive skills. Subjects were not only pinpricked. They were also asked to put needles through small holes with one hand and then checked for improvement with the other hand, or they were trained to tap their right big toe and tested for their ability with their left big toe (Bray 1928).

One has a mental picture of distinguished-looking men in suits sitting on wooden laboratory chairs and fiddling with needles or concentrating on their bare feet. But most importantly for the historical origination of the random group design, in these transfer of training studies the design of the threshold experiments had to be adapted. Once the stimuli had become trainings the original design of S-R-*noS*-R-S-R-S-R-*noS*-R-S-R-*noS*-R-S-R-S-R (etc.) was no longer feasible. Trainings cannot vary randomly as they cannot be haphazardly undone. Thus the standard design was limited to a particular 'cutout': (*noS*-) R-S-R. First a response (now an ability) was tested *without* preceding stimulus (now a training), next the training was given, and subsequently the ability was tested again. The transfer of training researchers worked with experimental groups whose abilities were compared before and after training.

1.3 Experimentation in Twentieth-century Educational Psychology

While psychophysical experimenters were absorbed in their painstaking experiments, the world outside the laboratories was rapidly changing. In early-twentieth-century Western countries, the common ideology of *laissez-faire* politics was widely replaced by that of planned social progress. Social amelioration gradually turned from a local charity matter into a concern of centralized government. The concomitant emergence of bureaucratic administration was to generate significant changes in the kind of knowledge about human functioning and human relations that was produced.

Whereas the authority of nineteenth-century local officials depended on their personal trustworthiness, the newly appointed administrators had to justify their decisions in terms of impersonal data and overt standards. Moreover, progress had to come about as economically as possible. The expanding class of bureaucratic officials turned to professional social researchers for knowledge that could serve the needs of administrative standardization and efficiency. With the shift from aristocratic to bureaucratic domination the former gentleman-scholar trying to discover the given laws of nature was increasingly replaced by the administrative researcher looking for technically useful data (Danziger 1990, Ross 1991, Porter 1995).

As far as psychology was concerned, the change was first brought about by psychophysical researchers taking an interest in educational administration. Around the turn of the twentieth century a debate blew up within American educational administrative circles on the utility of teaching subjects such as Latin and formal mathematics. School administrators wanted to abolish such subjects because of their apparent redundancy while proponents argued that 'formal discipline' strengthens general mental capaci-

ties. The issue reminded some American psychologists of the transfer of training phenomenon. Reasoning that formal discipline basically is a matter of transfer, they slightly adapted the former transfer experiments. They let subjects estimate weights or mark misspelled words, then trained them in estimating other weights or marking other misspelled words, and subsequently subjected them to the first tests again (*NoS*-R-S-R).

It may seem odd from a present-day perspective, but at the time a great deal of value was ascribed to the outcomes of such experiments. They seemed to provide an ideal basis for solid and impersonal decision-making. In the heat of the debate charges of experimental bias were soon made. In 1907, John E. Coover, a psychology student and former school principal, published an article on the issue (together with his teacher Frank Angell). The article argued that no valid causal conclusions can be drawn if the usual design of test-training-test (*noS*-R-S-R) was used (Coover and Angell 1907). With this strategy, the authors emphasized, it is impossible to tell whether the outcomes are caused by the training or by any other factor. The first test, for instance, might serve as a training too.

In experiments done for Coover's 1905 master's thesis, and described in the 1907 article, the solution was used of comparing 'experimental reagents,' who had been given a training, with 'control reagents,' who had not. With Coover's design the usual sequence of (*noS*-) R-S-R was, so to speak, cut in two. The experimental group underwent the second part of S-R (training and test) and the control group only the first half of (*noS*-) R (a test without preceding training). In other words, the classical control *trial* in which the stimulus was withheld (*noS*), was now applied to special control *subjects* (more on the early introduction of the control group is given in Dehue 2000)

1.4 Creating Equal Groups

In the 1910s, experimental educational research began to thrive. Educational experimenters acquired contracts from school administrators and left their laboratories for experimentation in real schools. 'Treatments' to be assessed now included methods of memorizing, ways of teaching, educational measures such as punishing and praising, and all kinds of classroom circumstances. The idea of using a control group also spread. In the 1916 volume of the *Journal of Educational Psychology*, 14% of the articles reported the use of control groups (Danziger 1990, pp. 113–15).

In the schools, however, life was much messier than in the former locations of psychophysical experimentation. It was much harder to protect experimental results against charges of arbitrariness. Methodological scrutiny became ever more important. It soon turned out that a price had to be paid for comparing the results of groups rather than comparing subjects to themselves. Contaminating factors between the pre-test and the post-test were replaced by invalidating differences between the groups.

At first it became customary to handle the new problem by hypothesizing which factors could bias the results, then subjecting subjects to preliminary tests for these factors, and next forming groups with equal test results. For instance, for testing the efficacy of a new way of teaching, equal groups were created by first measuring the pupils' intelligence and next composing equal groups as to mental capacity. With this 'matching' procedure, however, the choice of suspect factors depended very much on the researchers' personal imagination. Possibly the sex of the children also makes a difference, or their social background, or any other factor. Moreover, matching was very time and money consuming. For the combined reasons of saving money and cancelling out subjectivity—the twin principles of administrative rationality—experimental psychologists again invoked the methodological use of chance.

In 1923, the American psychologist William McCall published a manual entitled *How to Experiment in Education*. In his introduction, McCall estimated that increased efficiency in education could save a full year of teaching per person, and worked out that psychological experimentation would save '$134,680,000,000,000 for the next 100 generations of Americans.' He then he advanced a way of making his profession even more profitable to society, which was to compose the groups by chance. McCall did not take the task of randomization lightly. He rejected the procedure of writing numbers on pieces of paper, because papers with larger numbers contain more ink and are therefore likely to sink further to the bottom of a container. He nevertheless maintained that 'any device which will make the selection truly random is satisfactory' (McCall 1923, pp. 41–2).

Randomization, formerly used for creating random orders of stimuli, was now advanced as a cheap way of discarding unwanted differences between experimental groups. The random group design was first worked out in US educational psychology in the 1920s (on its later introduction in medicine, see Marks 1997).

2. A Guiding Rather Than Practicable Principle

While most psychology textbooks ascribe the random group design to natural science, some credit the British biometrician Ronald Fisher with it. Fisher developed the design in the 1930s for agricultural research. His methodological handbook, *The Design of Experiments* (1935), was widely received indeed. However, psychologists did not learn to use randomly composed groups from Fisher. From the 1920s onwards, educational experimenters had struggled with the complexity of data resulting from experiments simultaneously testing various experimental groups and comparing them

to one control group. Fisher offered statistical techniques for analyzing such complex between-groups variance and assessing the significance of the results (on Fisher's influence on psychology, see Gigerenzer and Murray 1987). His warnings that, for *statistical* reasons, these techniques could only be validly used if the groups were randomly composed added extra authority to psychologists already promoting the employment of random groups. In the 1940s, a range of handbooks on research design and statistics, mostly written by educational psychologists, passed on Fisher's instructions to other branches of psychology. One of these methodological professionals was Quinn McNemar who would later observe that Fisher's ways quickly 'became epidemic just as though psychologists had never previously planned an experiment' (McNemar 1960, p. 297).

At first, the combination of the random group design and significance testing seemed to provide an unquestionable experimental paradigm. Yet, in the context of experiments with humans, comparing random groups became an ideal mostly in the sense of a guiding principle rather than a viable strategy. Moral, legal, or practical reasons often prohibit people being given a treatment purely on the basis of chance. For example, the educational effects of prizes cannot be investigated by granting them at random, and differences in ethnic backgrounds among schools cannot be cancelled out by creating new classes on the basis of chance. Moreover, psychologists became involved with administrative research projects outside their traditional domain. From the 1930s, they were enlisted by other social scientists to measure the educational effects of radio talks and movies on issues such as voting, immigration, bootlegging, or war. In such contexts, it became all the more difficult to let the assignment of individuals to treatments depend purely on chance.

Renowned experiments by the US Army constitute a telling example of the unfeasibility of random assignment even with adult subjects living under command in a similar way to schoolchildren. During the Second World War, the Army's 'Morale Division' was charged with improving the soldiers' spirit. A film series entitled *Why We Fight* was made, and the Division's experimental section, chiefly staffed with psychologists, was asked to investigate the films' efficacy. The plan to evaluate the efficacy of these movies with the help of randomly composed groups did not work, as most soldiers became suspicious when some of them were summoned to see a movie while others were not. Random assignment thus appeared to elicit an unexpected new kind of bias. Although upon pre-testing the existing army units turned out to differ in many consequential respects, the units had to be used as experimental and control groups (Hovland 1949; for a history see Buck 1985).

After the war, the army experimenters were ensured of jobs at the best US universities (Herman 1995, p. 137). They widely propagated the standards of proper experimentation. Methodological treatises began to appear still advancing the random group design as the paragon form of experiment, but simultaneously indicating that unlike subjects in agricultural experimentation, people often cannot be assigned haphazardly to experimental conditions. Alternative designs and accompanying statistical techniques were developed.

Among the first methodologists to write in this vein was the young psychologist Donald Campbell, who also started his career in the army as an attitude and propaganda researcher. With the assistance of the statistician Julian Stanley, Campbell published a chapter entitled 'Experimental and Quasi-Experimental Designs for Research on Teaching.' In 1966 the chapter was reprinted as a separate 83-page booklet with the briefer but more-embracing title of *Experimental and Quasi-Experimental Designs for Research* (Campbell and Stanley 1966).

The book was the starting point of Campbell's career as a renowned methodologist of social science. In the decades to come, innumerable publications and thick textbooks on 'true' and 'quasi' experimental designs by Campbell and many other authors in America and Europe were to follow. These works testify to the lasting high regard but limited applicability of random group experiments as well as their sociotechnical inspiration. As one of them summarized these three aspects in a striking comparison: 'Most people after all do not drive around in Rolls Royces or Cadillacs, even if they represent some peak of automobile engineering perfection' (Bulmer 1986, p. 170).

See also: Experiment, in Science and Technology Studies; Experimental Design: Overview; Experimental Laboratories: Social/Pyschological; Experimenter and Subject Artifacts: Methodology; Laboratory Experiment: Methodology; Medical Experiments: Ethical Aspects; Psychophysics

Bibliography

Benschop R, Draaisma D 2000 In pursuit of precision. The calibration of minds and machines in late nineteenth century psychology. *Annals of Science* **57**(1): 1–25

Bray C W 1928 Transfer of Learning. *Journal of Experimental Psychology* **11**: 443–67

Buck P 1985 Adjusting to military life: the social sciences go to war, 1941–1950. In: Smith M R (ed.) *Military Enterprise and Technological Change: Perspectives on the American Experience*. MIT Press, Cambridge, MA

Bulmer M 1986 Evaluation research and social experimentation. In: Bulmer M (ed.) *Social Science and Social Policy*. Allen and Unwin, London

Campbell D T, Stanley J C 1966 *Experimental and Quasi-experimental Designs for Research*. R McNally, Chicago

Coover J E, Angell F 1907 General practice effect of special exercise. *American Journal of Psychology* **18**: 328–40

Danziger K 1990 *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge University Press, New York
Dehue T 1997 Deception, efficiency and random groups. Psychology and the gradual origination of the random group design. *Isis* **88**: 653–73
Dehue T 2000 From deception-trials to control-reagents. The introduction of the control group about a century ago. *American Psychologist* **55**(2): 264–268
Fisher R A 1935 *The Design of Experiments*. Oliver & Boyde, Edinburgh, UK
Galison P L 1997 *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, Chicago
Gigerenzer G, Murray D J 1987 *Cognition as Intuitive Statistics*. Lawrence Erlbaum, Hillsdale, NJ
Hacking I 1983 *Representing and Intervening. Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, New York
Herman E 1995 *The Romance of American Psychology: Political Culture in the Age of Experts*. University of California Press, Berkeley, CA
Hovland C I, Lumdsdaine A A, Sheffield F E 1949 *Experiments on Mass Communication*. Princeton University Press, Princeton, NJ
Kuhn T S 1962 *The Structure of Scientific Revolutions*. Chicago University Press, Chicago
Marks H M 1997 *The progress of experiment. Science and therapeutic reform in the United States, 1900–1990*. Cambridge University Press, New York
McCall W A 1923 *How to Experiment in Education*. Mcmillan, New York
McNemar Q 1960 At random: sense and nonsense. *American Psychologist* **15**: 295–300
Morawski J G (ed.) 1988 *The Rise of Experimentation in American Psychology*. Yale University Press, New Haven, CT
Porter T M 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ
Ross D 1991 *The Origins of American Social Science*. Cambridge University Press, New York
Schulze R 1913 *Aus der Werkstatt der experimentellen Psychologie und Pädagogik* [From the workshop of experimental psychology and pedagogy]. Vogtländer Verlag, Leipzig, Germany
Winston A S, Blais D J 1996 What counts as an experiment? A trans-disciplinary analysis of textbooks, 1930–1970. *American Journal of Psychology* **109**: 599–616

T. Dehue

Experimenter and Subject Artifacts: Methodology

'Experimenter artifacts' and 'subject artifacts' refer to systematic errors that can be attributed to uncontrolled experimenter and subject variables, respectively. The term 'experimenter' is used broadly in this context to refer not just to scientists who perform laboratory experiments, but to researchers working in any area of empirical investigation. Discussed here are the definition and history of artifacts in the social and behavioral sciences, the nature and control of experimenter and subject artifacts, and the delicate balance between ethical accountability and the avoidance of artifacts in research with human subjects.

1. *The Nature of Artifacts in Scientific Inquiry*

The sociologist H. H. Hyman wisely cautioned researchers not to equate ignorance of error with the lack of error. Because all scientific investigation is subject to error, it is prudent to be aware of this, to expose the sources in an attempt to control them, and to estimate the magnitude of such errors when they occur (Hyman 1954). In the social and behavioral sciences, researchers study the conduct of participants (or 'subjects') to further our understanding of the complexities of human behavior. Some of the complexity of subjects' behavior, however, results from specifiable but uncontrolled extraneous variables, often social in nature. The effects of such variables are conceptualized as a set of systematic errors (or 'artifacts') to be isolated, quantified, taken into account in interpreting the results, and, when possible, eliminated. Thus artifacts are not merely inconsequential effects in the research design, but may jeopardize the validity of the results. To be sure, today's artifact could become tomorrow's independent variable, as conditions discounted as 'nuisance' variables at one time may be redefined as substantive variables in their own right in some future study. For example, the placebo effect in medical research, where an inert substance or nonspecific treatment produces a 'healing' effect, can be defined as an artifact, but can also be exploited as a potential 'treatment' for chronic pain disorder (Turner et al. 1994).

Experimenter and subject artifacts, because they are conceptualized as systematic errors, are further distinguished from random errors. To explain the difference, Rosnow and Rosenthal (1997) envisioned an imaginary grocer who weighed the same bunch of grapes a number of times in a row. In an ideal world, no matter how precise the measurements, they would always come out a bit differently. Some of them would slightly overestimate the correct weight of the grapes, while other measurements would slightly underestimate the correct weight. Because the various overestimates and underestimates occur haphazardly (rather than in a patterned fashion), we can assume that the random fluctuations (or random errors) will cancel one another over the long run. Thus we can simply average the measurements to estimate the correct weight of the grapes. But suppose the grocer always weighed the grapes with his thumb on the scale, thereby introducing a systematic error by always tacking a few extra ounces onto the true weight. This is essentially the nature of artifacts, inasmuch as they tend to push measurements in the same direction, up or down, systematically causing the 'average value' to be consistently too large or too small. The problem for

researchers is to determine the direction in which a particular artifact might push the results, to try to improve the accuracy of the measurements, and when possible, to eliminate the source of systematic error and therefore the artifact.

The discovery of artifacts in the social and behavioral sciences has a long history, but it was not until the 1950s and 1960s that investigation of experimenter and subject artifacts began in earnest (see Rosnow 1981 for a detailed discussion of the historical context and wider implications of this work). Claims about the existence of experimenter and subject artifacts were sometimes met with skepticism, but by the late 1960s there was ample evidence to show the magnitude of the artifact problem in social and behavioral research (see Rosenthal and Rosnow 1969). Seminal investigations conducted by R. Rosenthal and M. T. Orne, working independently of one another, provided the initial impetus behind a flood of relevant work in the 1960s and 1970s. On the assumption that typical subjects often try to penetrate the nature of the study in which they are participating, Orne proposed the expression 'demand characteristics of the experiment' (which he derived from K. Lewin's concept of *Aufforderungscharaktere* to describe the 'demand value' of uncontrolled task-orienting cues in an experiment that exert an influence on subjects' ideas about its purpose. About the same time, Rosenthal showed that the experimenter's tacit hypothesis or expectancy can sometimes become a self-fulfilling prophecy of the subjects' experimental responses. Well before this work appeared, however, there had been intriguing clues about the possible existence of artifacts caused by uncontrolled variables when sentient subjects are the objects of investigation.

2. *Earlier Clues about Artifacts*

An early indication that artifacts may be lurking in the investigative procedures used in animal research involved a horse named 'Clever Hans.' At the turn of the twentieth century, an eminent German psychologist, O. Pfungst (1911), carried out a six-week investigation of this remarkable horse owned by a mathematics teacher named von Osten. Hans, who responded by tapping his hoof according to a code taught to him by von Osten, was purported to be able to reason and to answer questions put to him in German, including solving arithmetic problems. It was unlikely that von Osten had fraudulent intent because he readily allowed others (even in his absence) to question Hans, and also did not profit financially from the horse's talents. Thus it was possible to rule out intentional cues from von Osten as the reason for the horse's cleverness. In a series of elegant experiments, Pfungst demonstrated that Hans's accuracy diminished when: (a) he was fitted with blinders so he could not see his questioners; (b) the distance between Hans and his questioners was increased; or (c) the questioner did not know the answer. These results implied that Hans's performance was due to something other than his capacity to reason. Pfungst found that Hans was responding to subtle cues given by his questioners, not just intentional cues, but unwitting movements and mannerisms. This classic research provided an object lesson in the susceptibility of behavior (even animal behavior) to unconscious suggestion.

A half century later, another influential development fostered the suspicion that human subjects behaved in special ways because they knew they were 'subjects' of investigation. This proposition, called the 'Hawthorne effect,' grew out of a series of human factors experiments begun in the 1920s by a group of industrial researchers at the Hawthorne works of the Western Electric Company in Cicero, Illinois (Roethlisberger and Dickson 1939). The experiments were intended to study how workers' productivity and job satisfaction might be affected by specific workplace conditions (lighting, temperature, rest periods, and so on). The conclusion reached, however, was that increases in productivity were due not to the manipulated experimental conditions, but to the uncontrolled effects of being singled out for participation and to adventitious cues intrinsic in the experimental situation. More recently, this research has been taken to task because of ambiguities and irregularities that make it difficult to interpret; it is also charged that secondary accounts of the research were tarnished by sweeping generalizations embroidered by overly zealous authors (see, e.g., R. Gillespie's chapter in Morawski 1988, and also Parsons 1974 for an interesting behavioral interpretation of previously unreported circumstances of the Hawthorne research). Nonetheless, the 'Hawthorne effect' entered into the vocabulary of the social and behavioral sciences as implying a kind of placebo effect in social research with human subjects (see Sommer 1968). As a consequence, researchers were advised to employ 'Hawthorne control groups,' which received special attention or a meaningless task. Presumably, by examining obtained differences in performance between the Hawthorne control group and a no-treatment control group, any Hawthorne effects should be detectable.

Another important development, which went largely unnoticed for many years, involved a conceptual advance. S. Rosenzweig (1933) published an insightful critique of uncontrolled aspects of psychology experiments. In particular, he identified three distinct sources of artifacts, which he theorized were associated with (a) the experimenter's 'observational attitude'; (b) the subjects' motivational states; and (c) the experimenter's demeanor when interacting with the subjects. Rosenzweig sketched procedures that he thought might obviate these 'errors,' such as the use of 'simple deceptions' to distract subjects from the true purpose of the experiment. He cautioned, however, that it might be unclear whether the experimenter or the

subject was the 'true deceiver,' because subjects typically carry on a train of thought in which they attempt to guess the purpose of the experiment, and then model their own behavior accordingly. Interestingly, early psychologists referred to their subjects as 'reagents,' implying that they resembled the ingredients that chemists mixed in clean test tubes to produce predictable reactions. Rosenzweig argued that the 'test tubes' (i.e., research settings) and 'reagents' (the subjects) of psychological researchers were sullied by the needs, anxieties, self-deceptions, and intentions of human beings who knew very well that their behavior was being scrutinized as part of a scientific investigation.

3. *The Role of Experimenter Artifacts*

Rosenthal (1966) divided experimenter artifacts into two general classes, called 'noninteractional' and 'interactional.' The noninteractional kind operate, so to speak, in the mind, in the eye, or in the hand of the scientist, but do not directly affect the subjects' responses. Within this class are three subtypes, known as observer, interpreter, and intentional effects. 'Observer effects' refer to overestimates or underestimates of results during the observation and recording phase of an experiment, which are controlled by the use of mechanical recording devices or by independent replications to discover if similar results are observed. 'Interpreter bias' refers to an error in the interpretation of data, which is ostensibly controlled by allowing other scientists to analyze and interpret the original data for themselves (see also MacCoun 1998). 'Intentional effects' imply that the data were fabricated or fraudulently manipulated. Such distortions are devastating to science because they undermine the integrity of the literature on which the advancement of science depends; safeguarding against dishonesty is presumed to be the ethical responsibility of every scientist.

Interactional artifacts, the second general class, are those due to extraneous experimenter-related variables that can directly influence the reactions of the research subjects. These include 'biosocial effects' (biases due to biological or social attributes of researchers, such as the sex, age, ethnicity, and race of the researcher), 'psychosocial effects' (e.g., the researcher's personality and comportment), 'situational effects' (biases attributable to uncontrolled aspects of the setting), and modeling effects (i.e., the example set by the investigator in the role of a 'scientist'). Replications across settings, subjects, and experimenters are used to isolate interactional artifacts, although procedural and sampling differences may not be easy to interpret. Moreover, scientists, like all humans, are susceptible to the biases imposed by human limitations of perception and cognition.

One additional interactional artifact is well known in the social and behavioral sciences, called the 'experimenter expectancy effect' by Rosenthal, its discoverer. It has spawned a large number of studies (Rosenthal and Rubin 1978), with various applications in education, the law, and clinical practice (see Blanck 1993). An 'experimenter expectancy bias' means that the experimenter's hypothesis leads unintentionally to behavior toward the subjects that, in turn, increases the likelihood that the hypothesis will be confirmed (reminiscent of the Clever Hans phenomenon). One solution to this problem is to keep experimenters unaware of ('blind' to) which subjects are to receive the experimental treatment and which the control treatment. If the experimenters do not know what treatment the subject has received, then they cannot accurately communicate expectancies about the nature of the particular treatment. The necessity of keeping experimenters blind is well recognized in randomized drug trials, for example. In fact, no drug trial is taken completely seriously unless it has followed elaborate 'double-blind procedures,' in which neither the subjects nor the experimenters know who is in the experimental and who is in the control group.

4. *The Role of Subject Artifacts*

In focusing on subject artifacts, researchers have generally been interested in the role motivations that guide the research subject's behavior. Orne (1962) described the role of the 'good subject,' that is, the participant who is particularly attentive to, and compliant with, demand characteristics. In one fascinating demonstration study, Orne asked the subjects to add hundreds of thousands of two-digits numbers—which they kept doing until finally the experimenter gave up! Even when the subjects were told to tear each worksheet into a minimum of 32 pieces before going on to the next, they persisted. Orne explained this behavior as the 'role enactment' of individuals who reasoned that, no matter how trivial and inane the assigned task seemed to them, it must have some important scientific purpose or they would not have been asked to participate in the first place. Thus, Orne argued, they complied with the demand characteristics of the experiment in order to 'further the cause of science.'

Not all research participants are willing to play the 'good subject' role, but Rosenthal and Rosnow (1975) reported that many volunteers for research participation are likely to conform to this expectation. Volunteers for research are also usually better educated, higher in social class status, more intelligent, more approval-motivated, and more sociable than nonvolunteers. These and other demographic and personality characteristics may result in 'volunteer subject bias' in that the results of studies using subjects who specifically volunteered may not be simply generalizable to the population at large. For example, standard-

izing the norms of an intelligence test on typical volunteer subjects would produce biased estimates of true values in a more general population. Rosenthal and Rosnow described some strategies for improving recruitment procedures and the generalizability of results, and also logical procedures for estimating the direction of volunteer subject biases.

Other subject motivations may be operating, however, including what M. J. Rosenberg called 'evaluation apprehension,' referring to an anxiety-toned state in which subjects worry that the experimenter plans to evaluate an aspect of their psychological competence. As a result, they develop their own hypotheses about how to win the experimenter's approval or avoid disapproval. Experiments in which evaluation apprehension appears likely to occur are those that contain an element of surprise or have an aura of mystery. The more explicit the cues, the more control the experimenter has in granting positive evaluation, or the less effortful the subjects' responses, the greater may be the resulting response bias due to subjects' evaluation apprehensions (see Rosenberg's discussion in Rosenthal and Rosnow 1969). One solution to this problem is to ensure the confidentiality of subjects' responses, on the assumption that individual subjects will then be less apprehensive, and more forthcoming in their responses. In research on sensitive topics, such as studies of sexual behavior, self-reported data nevertheless may be suspected of measurement error and participation bias (see Catania et al. 1990 for detailed discussion).

Other subject artifacts have been studied in the field of assessment, including a variation on evaluation apprehension known as 'social desirability bias.' In this case, the respondent to a personality inventory answers items in ways that seem likely to elicit favorable evaluation. One way to assess this problem is by means of a questionnaire that was specifically designed to tease out this response bias (Crowne and Marlow, 1964). Various other subject biases in assessment have been identified, and particular strategies have been developed to isolate or avoid those problems, such as the development of 'lie scales' to detect prevarication. Orne (see his chapter in Rosenthal and Rosnow 1969, pp. 143–79) suggested a strategy to ferret out demand characteristics by using 'quasi-control subjects' in addition to the usual control groups. Quasi-controls are individuals who are asked to step out of the traditional subject role; they reflect on the experiment in which they participated and suggest how their own experimental behavior was influenced by extraneous variables as opposed to only the controlled independent variable. It is now argued, however, that demand characteristics also exert an influence on behavior beyond the experimental setting, and should be recognized as a pervasive influence on all human interaction (see Orne and Bauer-Manley 1991, for the role of demand characteristics in the context of psychotherapy).

5. *Artifacts and Ethics in Collision*

Until World War II, the popular conception of science as an 'endless frontier' unencumbered by moral constraints remained largely unshaken. Currently, however, investigators in even the most benign areas of the social and behavioral sciences are required to comply with stringent ethical guidelines laid down in professional and statutory codes overseen by review boards (see Rosnow et al. 1993). This situation can at times lead to knotty moral dilemmas as researchers try to strike a balance between ethical accountability and the technical demands of sound scientific practice (see Blanck et al. 1992, Kimmel 1996, Schuler 1980). The effort to control for artifacts may, for example, come into collision with ethical imperatives such as informed consent and full disclosure. By way of illustration, in clinical trials using double-blind placebo-control designs, complying will full disclosure implies that the potential participants will be told they could be in a condition that will be deprived of possible benefits of the experimental treatment. Once they become aware of this fact, however, some subjects might surreptitiously share doses in the hope that this will increase the likelihood that each of them will receive at least some access to the experimental treatment. Such a situation could render any subsequent comparison of the experimental group and the 'untreated' control group meaningless. A strategy that can be used to avoid this problem, if there is an alternative treatment that is better than a placebo, is to provide the control group with this treatment rather than a placebo. This approach avoids the ethical bind of depriving control patients of any effective treatment, keeps the design double-blind, and focuses the research on other comparisons of practical and scientific importance (e.g., efficacy, relative cost, and side effects). As moral dilemmas proliferate in the sciences, the challenge faced by researchers is to expand our existing knowledge with valid data and yet abide by an evolving social contract that is responsive to current ethical sensitivities (see Rosnow 1997 for further discussion).

Bibliography

Blanck P D (ed.) 1993 *Interpersonal Expectations: Theory, Research and Applications*. Cambridge University Press, New York

Blanck P D, Bellack A S, Rosnow R L, Rotheram-Borus M J, Schooler N R 1992 Scientific rewards and conflicts of ethical choices in human subjects research. *American Psychologist* **47**: 959–65

Catania J A, Gibson D R, Chitwood D D, Coates T J 1990 Methodological problems in AIDS behavioral research: Influences on measurement error and participation bias in studies of sexual behavior. *Psychological Bulletin* **108**: 339–62

Crowne D P, Marlow D 1964 *The Approval Motive*. Wiley, New York

Hyman H H 1954 *Interviewing in Social Research*. University of Chicago Press, Chicago

Kimmel A J 1996 *Ethical Issues in Behavioral Research: A Survey*. Blackwell, Oxford, UK
MacCoun R J 1998 Biases in the interpretation and use of research results. *Annual Review of Psychology* **49**: 259–87
Morawski J G (ed.) 1988 *The Rise of Experimentation in American Psychology*. Yale University Press, New Haven, CT
Orne M T 1962 On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implication. *American Psychologist* **17**: 776–83
Orne M T, Bauer-Manley N K 1991 Disorders of self: Myths, metaphors, and demand characteristics of treatment. In: Strauss J, Goethals G R (eds.) *The Self: Interdisciplinary Approaches*. Springer Verlag, New York, pp. 93–106
Parsons H M 1974 What happened at Hawthorne? *Science* **183**: 922–32
Pfungst O 1911 *Clever Hans* (*The Horse of Mr. von Osten*). H. Holt, New York, [reissued 1965 by Holt, Rinehart & Winston, New York]
Roethlisberger F J, Dickson W J 1939 *Management and the Worker*. Harvard University Press, Cambridge, MA
Rosenthal R 1966 *Experimenter Effects in Behavioral Research*. Appleton-Century-Crofts, New York
Rosenthal R, Rosnow R L (eds.) 1969 *Artifact in Behavioral Research*. Academic Press, New York
Rosenthal R, Rosnow R L 1975 *The Volunteer Subject*. Wiley, New York
Rosenthal R, Rubin D B 1978 Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences* **3**: 377–86
Rosenzweig S 1933 The experimental situation as a psychological problem. *Psychological Bulletin* **68**: 337–54
Rosnow R L 1981 *Paradigms in Transition: The Methodology of Social Inquiry*. Oxford University Press, New York
Rosnow R L 1997 Hedgehogs, foxes, and the evolving social contract in psychological science: Ethical challenges and methodological opportunities. *Psychological Methods* **2**: 345–56
Rosnow R L, Rosenthal R 1997 *People Studying People: Artifacts and Ethics in Behavioral Research*. W.H. Freeman, New York
Rosnow R L, Rotheram-Borus M J, Ceci S J, Blanck P D, Koocher G P 1993 The institutional review board as a mirror of scientific and ethical standards. *American Psychologist* **48**: 821–86
Schuler H 1980 *Ethische Probleme psychologischer Forschung*. Hogrefe, Göttingen, Germany
Sommer R 1968 Hawthorne dogma. *Psychological Bulletin* **70**: 592–95
Turner J A, Deyo R A, Loeser J D, Von Korff M, Fordyce W E 1994 The importance of placebo effects in pain treatment and research. *Journal of the American Medical Association* **271**: 1609–14

R. L. Rosnow

Expert Memory, Psychology of

The objective of this article is to provide an overview of some of the research on cognitive expertise and the differences between experts and novices in complex cognitive skills. When such skills were first studied, it was thought that expertise in them would be a matter of particularly adept fundamental 'reasoning' processes (e.g., Newell and Simon 1972—especially regarding the Logic Theorist and the General Problem Solver). Instead, what has been revealed in subsequent research is that the primary basis of expertise resides in knowledge, knowledge organization, and complex associated perceptual processes. For the interested reader, this 'knowledge turn' in the understanding of expertise has been documented in a number of landmark volumes (Chi et al. 1988, Ericsson and Smith 1991, Ericsson 1996, Feltovich et al. 1997a, 1997b, Hoffman 1992).

In the remainder of the article, some of the fundamental phenomena that have been discovered about human expertise will be outlined. An argument will be advanced that *the development of expertise is largely a matter of reorganizing knowledge and cognitive processes to do tasks efficiently and effectively*. Experts restructure their inner workings—knowledge and procedures—for efficient application to their work-a-day environments (see also Ericsson and Lehmann 1996).

1. Larger Cognitive Units

As people gain increasing experience working in an environment, they cognitively organize the environment into larger units. This is the classic and one of the best-established phenomena of expertise, supported by a long line of studies that starts with the game of chess.

In the 1960s and early 1970s, De Groot (1965) and Chase and Simon (1973) studied master level and less expert chess players. In the basic task from these studies, subjects were shown a chess board with pieces representing game states from real games. Subjects were shown the boards for only five seconds and were then asked to reproduce the board they had seen.

Experts, after a brief glance, were able to reproduce much more of the configuration than novices. In the Simon and Chase studies, expert advantage was about 4–1. In the De Groot studies, expert performance was nearly perfect (for 25 piece boards), while novices were able to reproduce about five items—or about the number of items that can be held in short-term memory. In related studies, the same investigators showed that after presentation of a board, experts were able to choose better next moves than novices.

The general explanation for expert superiority in the chess studies involves 'chunking' in perception and memory. With experience, experts acquire a large 'vocabulary' or memory store of board *patterns* involving groups of pieces or 'chunks.' A chunk is a perceptual or memory structure that bonds a number of more elementary units (e.g., chess pieces) into a larger organization. When experts see a chess setup from a real game, they are able to recognize these

familiar patterns. They can then associate these patterns with moves stored in memory that have proved effective in the past. Novices have not been exposed to game configurations enough to have developed these kinds of patterns. Hence they deal with the board in a piece-by-piece manner. Similarly, when *experts* are presented with chess boards containing *randomly* placed pieces that do not enable the experts to take advantage of established patterns, they perform at the same level as novices.

These basic chess phenomena have been replicated many times—in chess, but also in numerous other fields (e.g., the games of bridge, Engle and Bukstel 1978; GO, Reitman 1976; and electronics, Egan and Schwartz 1979). In most studies of this sort, it is really the chunk *size* that has been shown to be larger for experts. Both the novice *and* the expert are subject to the same limitations of short-term memory (7 ± 2 units, Miller 1956). However, expert chunks are larger. Hence, whereas a chess novice sees a number of isolated chess pieces, the expert recognizes about the *same* number of *larger* units—for example, one might be a 'king defense configuration'—each composed of a number of chess pieces.

The fact that expert organizational units are larger is interesting; it indicates an expert adaptation useful for expanding the functional size of short-term or working memory. However, it is not the general finding that is *most* interesting from these kinds of studies. What is perhaps more revealing is the *nature and organization* of the 'chunks' or memory structures experts use, discussed next.

2. *Abstracted, Functional Organization*

Some studies, using variants of the standard Simon and Chase chessboard paradigm, have examined the actual nature of expert and novice 'chunks' or knowledge structures. Table 1 presents a general overview of the differences that emerge.

Chase and Simon (1973) examined the nature of chunks experts used in reconstructing a briefly presented board. Expert chunks were found largely to comprise attack and defense configurations, tied to more general strategic aspects of the game. It is not clear how novice units were organized.

Studies of bridge (Charness 1979, Engle and Bukstel 1978) and electronics (Egan and Schwartz 1979), following the Chase and Simon paradigm, showed similar results among experts. In the bridge studies, experts and novices were presented depictions of four-handed bridge deals and were required to reproduce these deals. Experts reproduced the cards by suit across hands, remembering cards of the same suit from three hands and inferring the fourth, an organization useful in playing the game of bridge. Novices recalled the cards by order of card rank within hands, perhaps a carryover from other card games, for example, poker. In the electronics studies, subjects were shown an electronic circuit diagram, which they were then requested to reproduce. Experts grouped individual diagram components into major electronic components (e.g., amplifiers, filters, rectifiers), while novice organization was largely based on the spatial proximity of basic diagram symbols.

Similar results have been shown by studies from other fields, using somewhat different methodologies. Voss and co-workers (Spilich et al. 1979) studied baseball aficionados and more casual baseball observers. Subjects were presented a colorful description of a half-inning of baseball and were then to recall the inning. Expert recall was structured by major goal-related sequences of the game such as advancing runners, scoring runs, and preventing scoring. Novice recall involved less integral components, for example, the weather and the crowd mood, and did not capture goal-directed, sequential activity as well.

Two studies of computer programming are also noteworthy. In a study of expert and novice ALGOL programmers, McKeithen et al. (1981) presented a list of 21 commands in the ALGOL language to ALGOL experts, students after one ALGOL course, and students at the beginning of an ALGOL course. Subjects were given 25 recall trials after initial learning of the list. Recall organization by pre-ALGOL students included organization by surface features of commands (e.g., commands with the same beginning letter or same length of command name), and groups of commands forming natural language segments (e.g.,

Table 1
Organizational bases for 'chunking'

	Expert	Novice
Chess	Attack, defense, etc.	?
Bridge	Suits-across hands	Rank-within hands
Programming	'Algorithms'	Surface features (e.g., orthography) Commands
Electronics	System components	Proximity
Baseball	Scoring runs, defense	Misc. (e.g., weather, crowd)
Physics	Physics laws	Problem objects

STRING IS NULL BITS) which have no conceptual meaning within ALGOL. Experts, however, grouped commands that formed mini ALGOL algorithms (e.g., formation of loops) or represented types of ALGOL data structures. Students after an ALGOL course produced groupings that were a mixture of surface-related and conceptual ALGOL organizations. A similar study (Adelson 1981) presented a list containing three intact computer programs, scrambled together and out of order, to expert and novice programmers who were required to recall the list. Over recall trials, experts reconstructed the original three major algorithms, while novice recall organization was according to syntactic similarities in program statements, regardless of which program they had come from.

Other findings have come from work in physics (Chi et al. 1981) and medicine (Feltovich et al. 1984). In the basic task from the physics study, problems from chapters in an introductory physics text were sampled and placed on individual cards. Expert (professors and advanced graduate students) and novice (college students after their first mechanics course) physics problem solvers sorted problems into groups they 'would solve in a similar manner.' The finding was that experts created groups based on the major physics laws (conservation and force laws) applicable in solution. Novice groupings were organized by salient objects (e.g., springs, inclined planes) and features contained in the problem statement itself. Similarly, in studies of expert and novice diagnoses within a subspecialty of medicine, it was found that expert diagnosticians organized diagnostic hypotheses according to the major physiological issue relevant in a case, while novice hypotheses were more isolated and seemed more dependent on particular patient cues.

Zeitz (1997) has reviewed these and more recent studies of this type, investigating what she calls experts' use of 'Moderately Abstracted Conceptual Representations' (MACRs), representational abstractions of the type we have just discussed. She proposes numerous ways that such abstraction aids the efficient utilization of knowledge and reasoning by experts. These include: (a) the role of abstracted representations in retrieving appropriate material from memory (e.g., Chi et al. 1981); (b) the schematic nature of MACRs in integrating information, suggesting what information is important, activating appropriate prior knowledge, providing guidance for appropriate action, and providing support/argumentation for a line of approach (e.g., Phelps and Shanteau 1978, Schmidt et al. 1989, Voss et al. 1983); (c) aiding productive analogical reasoning (e.g., Gentner 1988); and (d) providing guidance for the elimination of unimportant information and unnecessary detail (e.g., Patel and Groen 1991).

In summary, research conducted in the last quarter of the twentieth century indicates that expert knowledge organization develops in the direction of increased value for *doing* the work of the expert's field. Novices also impose organization—but it is less useful—for example, organization by items named in a situation, proximity of entities to others, or superficial analogy.

3. *Automaticity*

Practice is almost a defining property of expertise. Experts are generally individuals with extreme amounts of practice on a circumscribed set of tasks of a work environment. For example, some expert radiologists who have been studied estimated that they have analyzed more than half a million radiographs (X-rays) in their careers (Lesgold et al. 1988).

A general theme of research on automaticity is that the character of cognitive operations changes with practice. Operations that are initially slow, serial, and demand conscious attention become fast, less deliberate, and can run in parallel with other processes (Schneider and Shiffrin 1977). With enough practice, one can do several things at once.

There are many examples in the literature of the effects of practice on automaticity. For example, expert typists can type and recite nursery rhymes at the same time (Shaffer 1975). Skilled abacus operators can answer routine questions ('What is your favorite food?') without loss of accuracy or speed in working with the abacus (Hatano et al. 1977). After six weeks of practice (one hour per day), college students could read unfamiliar text while simultaneously copying words read by an experimenter—without decrement in reading speed or comprehension (Spelke et al. 1976).

Automaticity is important to expertise. In complex skills with many different cognitive components, some of the more basic ones (e.g., fundamental decoding/encoding of input) must be automated if higher levels skills such as reasoning, inference, and integration are ever to be proficient (e.g., Lesgold and Resnick 1982). There is also a possible interaction between automaticity of processes and the usability of available knowledge. Investigators (e.g., Feltovich et al. 1984, Jeffries et al. 1981) have suggested that a major limitation of novices is in their inability, situationally, to utilize knowledge that they in some sense possess. Problems in knowledge usability may be associated with overload or inefficiency in using working (or short-term) memory. The usable knowledge of experts, in turn, may devolve from the subordination of many task components to automatic processing, which increases capacity for more controlled management of memory and knowledge application (cf. Perfetti and Lesgold 1979).

Although many years of experience are a necessary requirement for the development of expertise, extensive experience alone is not sufficient. The seminal work of Ericsson indicates that long-term practice needs to be thoughtful, reflective, and self-critical; that

is, it must constitute ongoing 'deliberate practice' (e.g., Ericsson and Charness 1994, Ericsson and Lehmann 1996). It is not sufficient just to 'go through the motions.'

4. More Recent Developments

Recent research on expertise has examined the contexts in which expertise develops and is practiced. An impetus for this research is to understand how the contexts of expertise provide the support (including that supplied by other talented co-workers and sponsors) and opportunities (e.g., access to the unusual, tricky cases in one's field) that enable an individual's development toward expertise to progress (e.g., Agnew et al. 1997, Stein 1997). Another area of investigation involves how the nature of expertise varies with changing contexts, including the nature of expertise when work environments change so frequently that it is impossible to practice any one, circumscribed set of skills over a long period of time (Feltovich et al. 1997a, 1997b, Hoffman et al. 1997, Prietula et al. 2000).

5. Conclusions

A general argument has been made that expert knowledge structures and procedures are reorganized in directions that enable effective application to task demands of a working environment. Most of these changes are adaptations that enable utilization of large amounts of information in the context of limited internal processing resources—in particular those imposed by the small capacity of short-term or working memory.

Grouping or chunking on information structures and procedure components functionally increases the size of working memory and its efficiency. More information can be considered for each 'unit' of working memory. Selectivity, discrimination, and *abstraction* insure that only the most useful information is thrown into competition for resources. *Automaticity* is a means of restructuring some procedures so that working memory is largely circumvented—freeing resources for other cognitive chores.

It is a tension between high information flow and limited internal resources that encourages the development of strategies for the efficient use of information and knowledge. Expertise is largely a progressive adaptation by a limited cognitive system to the great demands posed by important, complex cognitive tasks.

See also: Chess Expertise, Cognitive Psychology of; Expert Systems in Cognitive Science; Expertise, Acquisition of; Knowledge Representation; Lifespan Theories of Cognitive Development; Medical Expertise, Cognitive Psychology of; Problem Solving and Reasoning, Psychology of; Science and Technology Studies: Experts and Expertise; Sports as Expertise, Psychology of

Bibliography

Adelson B 1981 Problem solving and the development of abstract categories in programming languages. *Memory and Cognition* **9**: 422–33

Agnew N M, Ford K M, Hayes P J 1997 Expertise in context: Personally constructed, socially selected, and reality-relevant. In: Feltovich P J, Ford K M, Hoffman R R (eds.) *Expertise in Context: Human and Machine*. AAAI/MIT Press, Menlo Park, CA, pp. 219–44

Bransford J D, Stein B S, Shelton T S, Owings R A 1981 Cognition and adaptation: The importance of learning to learn. In: Harvey J H (ed.) *Cognition, Social Behavior and the Environment*. Lawrence Erlbaum, Hillsdale, NJ

Charness N 1979 Components of skill in bridge. *Canadian Journal of Psychology* **33**: 1–16

Chase W G, Simon H A 1973 Perception in chess. *Cognitive Psychology* **4**: 55–81

Chi M T H, Feltovich P J, Glaser R 1981 Categorization and representation of physics problems by experts and novices. *Cognitive Science* **5**: 121–52

Chi M T H, Glaser R, Farr M J 1988 *The Nature of Expertise*. Lawrence Erlbaum, Hillsdale, NJ

DeGroot A D 1965 *Thought and Choice in Chess*. Basic Books, New York

Egan D E, Schwartz B J 1979 Chunking in recall of symbolic drawings. *Memory and Cognition* **7**: 149–58

Engle R W, Bukstel L 1978 Memory processes among bridge players of differing expertise. *American Journal of Psychology* **91**: 673–89

Ericsson K A 1996 *The Road to Excellence: The Acquisition of Expert Performance in Arts and Sciences, Sports and Games*. Lawrence Erlbaum, Mahwah, NJ

Ericsson K A, Charness N 1994 Expert performance: Its structure and acquisition. *American Psychologist* **49**: 725–47

Ericsson K A, Lehmann A C 1996 Expert and exceptional performance: Evidence on maximal adaptations on task constraints. *Annual Review of Psychology* **47**: 273–305

Ericsson K A, Smith J 1991 *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press, Cambridge, UK

Feltovich P J, Ford K M, Hoffman R R 1997a *Expertise in Context: Human and Machine*. AAAI/MIT Press, Menlo Park, CA

Feltovich P J, Johnson P E, Moller J H, Swanson D B 1984 LCS: The role and development of medical knowledge of diagnostic expertise. In: Shortliffe E H, Clancey W J (eds.) *Readings in Medical Artificial Intelligence: The First Decade*. Addison-Wesley, Reading, MA, pp. 275–319

Feltovich P J, Spiro R J, Coulson R L 1997b Issues of expert flexibility in contexts characterized by complexity and change. In: Feltovich P J, Ford K M, Hoffman R R (eds.) *Expertise in Context: Human and Machine*. AAAI/MIT Press, Menlo Park, CA, pp. 125–46

Gentner D 1988 Metaphor as structure mapping: The relational shift. *Child Development* **59**: 47–59

Hatano G, Miyake Y, Binks M B 1977 Performance of expert abacus operators. *Cognition* **5**: 57–71

Hoffman R R 1992 *The Psychology of Expertise: Cognitive Research and Empirical AI*. Springer-Verlag, New York

Hoffman R R, Feltovich P J, Ford K M 1997 A general framework for conceiving of expertise and expert systems in context. In: Feltovich P J, Ford K M, Hoffman R R (eds.) *Expertise in Context: Human and Machine*. AAAI/MIT Press, Menlo Park, CA, pp. 543–80

Jeffries R, Turner A A, Polson P G, Atwood M E 1981 The processes involved in software design. In: Anderson J R (ed.) *Cognitive Skills and Their Acquisition*. Lawrence Erlbaum, Hillsdale, NJ

Lesgold A M, Resnick L B 1982 How reading difficulties develop: Perceptives from a longitudinal study. In: Das J P, Mulcahy R, Walls A E (eds.) *Theory and Research in Learning and Learning Disabilities*. Plenum, New York

Lesgold A M, Rubinson H, Feltovich P J, Glaser R, Klopfer D, Wang Y 1988 Expertise in a complex skill: Diagnosing x-ray pictures. In: Chi M T H, Glaser R, Farr M J (eds.) *The Nature of Expertise*. Lawrence Erlbaum, Hillsdale, NJ, pp. 311–42

McKeithen K B, Reitman J S, Reuter H H, Hirtle S C 1981 Knowledge organization and skill differences in computer programmers. *Cognitive Psychology* **13**: 307–25

Miller G A 1956 The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* **63**: 81–97

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Patel V L, Groen G J 1991 The general and specific nature of medical expertise. In: Ericsson K A, Smith J (eds.) *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press, Cambridge, UK, pp. 93–125

Perfetti C A, Lesgold A M 1979 Coding and comprehension in skilled reading. In: Resnick L B, Weaver P (eds.) *Theory and Practice of Early Reading*. Lawrence Erlbaum, Hillsdale, NJ

Phelps R H, Shanteau J 1978 Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance* **21**: 209–19

Prietula M J, Feltovich P J, Marchak F 2000 Factors influencing analysis of complex cognitive skills: A framework and example from industrial process control. *Human Factors* **42**: 56–74

Reitman J 1976 Skilled perception in GO: Deducing memory structure from inter-response times. *Cognitive Psychology* **8**: 336–56

Schmidt J A, McLaughlin J P, Leighton P 1989 Novice strategies for understanding paintings. *Applied Cognitive Psychology* **3**: 65–72

Schneider W, Shiffrin R M 1977 Controlled and automatic human information processing: 1. Detection, search and attention. *Psychological Review* **84**: 1–66

Shaffer L H 1975 Multiple attention in continuous verbal tasks. In: Rabbitt P M A, Dornic S (eds.) *Attention and Performance V*. Academic Press, New York

Spelke E, Hirst W, Neisser U 1976 Skills of divided attention. *Cognition* **4**: 215–30

Spilich G J, Vesonder G T, Chiesi H L, Voss J F 1979 Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior* **18**: 275–90

Stein E W 1997 A look at expertise from a social perspective. In: Feltovich P J, Ford K M, Hoffman R R (eds.) *Expertise in Context: Human and Machine*. AAAI/MIT Press, Menlo Park, CA, pp. 181–94

Voss J F, Greene T R, Post T A, Penner B C 1983 Problem solving skill in the social sciences. In: Bower G (ed.) *The Psychology of Learning and Motivation*. Academic Press, New York, Vol. 17, pp. 165–213

Zeitz C M 1997 Some concrete advantages of abstraction: How experts' representations facilitate reasoning. In: Feltovich P J, Ford K M, Hoffman R R (eds.) *Expertise in Context: Human and Machine*. AAAI/MIT Press, Menlo Park CA, pp. 43–65

P. J. Feltovich

Expert Systems in Cognitive Science

Expert Systems (ES), also called Knowledge-based Systems (KBS), Intelligent Agent Systems or in short Knowledge Systems (KS), are computer programs that exhibit a similar high level of intelligent performance as human experts. By performing inference processes on explicitly stated knowledge, these computer programs can solve problems within a given domain that would previously have required a human expert.

Figure 1 shows the classical architecture of an expert system. According to its classical characterization, an expert or knowledge system consists of (a) a knowledge base where knowledge is encoded in some fixed representational form, (b) the search system (also referred to as inference system or inference engine), which derives solutions for the posed problems and questions from the knowledge base, (c) the knowledge acquisition system, which allows human experts to enter their knowledge into the knowledge system, and (d) the user interface, possibly including an explanation component, by which the users of the knowledge system may present their problems and questions and obtain solutions from the system. The knowledge base contains symbols and symbol structures (Newell and Simon 1976) to embody the knowledge about some given domain, given tasks, as well as their interdependences. The search system performs inference processes by manipulating these symbols. The symbol manipulation processes may be performed by the rules of formal logic or by heuristic processes. With the assistance of the knowledge acquisition system and the knowledge engineer, who knows the representational formats of the knowledge system and assumes the role of a scribe, the knowledge of a human expert can thus be coded into the knowledge system. Knowledge acquisition is, on the one hand, driven by the target knowledge which the human expert wants to publish in the expert system. On the other hand, it is constrained by the formalism of the representation and inference systems. Successful knowledge acquisition is therefore dependent on productive interactions between the knowledge engineer and the domain expert as well as upon the adequacy of the representation and search formalisms of the expert system.

The static and dynamically generated symbol structures as well as the inference processes of expert

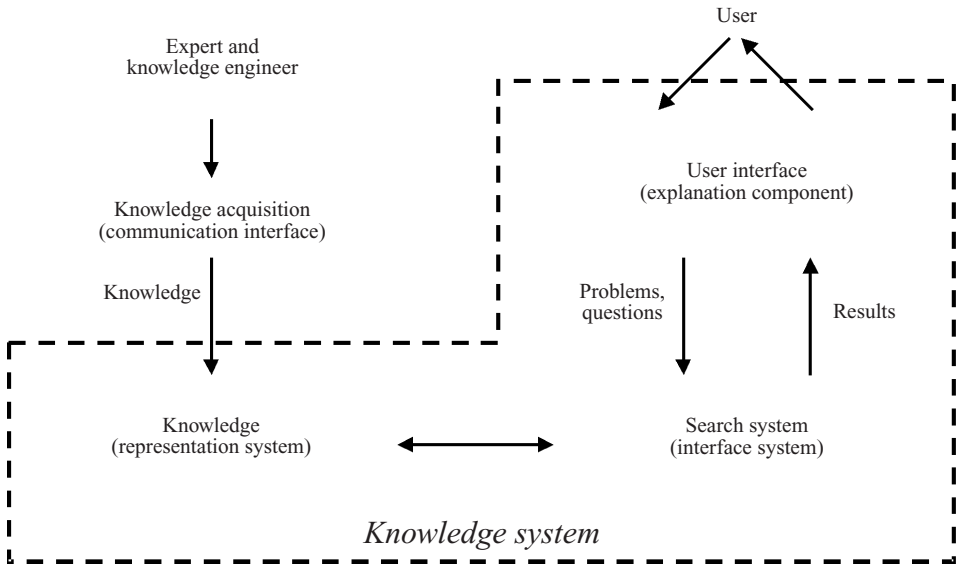

Figure 1

systems are highly complex. This already applies to the MYCIN system (Buchanan and Shortliffe 1984), which is one of the earliest expert systems to perform at the level of a human expertise. MYCIN was developed at Stanford University in the mid-1970s and provided medical consultations about the diagnosis and treatment of certain infections (bacterial blood infections and meningitis) by reasoning about the symptoms and general characteristics of a patient in combination with the laboratory results of body fluid analyses. MYCIN's knowledge base contained about 500 high-level production rules. A production rule encodes knowledge by IF–THEN rules, which describe which conditions must be satisfied (specified by the IF segment) in order to derive some conclusion or to perform some action (specified by the THEN segment). One of MYCIN's production rules had the following knowledge encoded:

IF the infection is primary bacteremia
AND the site of the culture is one of the sterile sites
AND the suspected portal of entry is gastrointestinal tract
THEN there is suggestive evidence (0.7) that the infection is bacteroid.

To obtain a diagnosis, MYCIN would build a sequence (or a tree) of several such rules, so that the preconditions of the rules (i.e., the IF parts) of some given inferencing episode would be provided as the conclusions (i.e., the THEN part) of the subsequent inferencing episode. In other words, MYCIN performed a backward construction of a kind of proof tree. This means that the tree is constructed by first considering a set of possible conclusions. Second, evidence for these conclusions is collected and assessed. By backward chaining rules in this fashion, MYCIN could recognize about 100 causes of bacterial infections and recommend effective drug prescriptions. The chaining of the rules was thereby guided by the search system and the user's request to find a diagnosis and a therapy. In its reasoning, MYCIN also used probability considerations, which were encoded by so called certainty weights (e.g., suggestive evidence: 0.7). During its time of application, the MYCIN system was very successful and performed at the level of the best experts in its specific domain of medical expertise.

Other expert systems at the time of the first generation systems (mid-1970s to mid-1980s) were similarly successful in different application domains, such as chemistry, geology, mathematics, engineering, speech, and language understanding as well as for different tasks such as diagnosis, configuration, and comprehension tasks. Together with MYCIN, these systems have become well known as legacy expert systems. These legacy systems include Dendral (Lindsay et al. 1980), a chemistry system, which examines a spectroscopic analysis of an unknown molecule and predicts the molecular structures that account for the particular analysis, INTERNIST/CADUCEUS (Miller et al. 1982, Wolfram 1995), a diagnosis system for internal medicine, PROSPECTOR (Duda and Reboh 1984), a classification and geological site evaluation system, MACSYMA (Engelman 1971, Moses 1975) which assists mathematicians, scientists, and engineers in solving complex algebraic problems, and XCON/R1 (McDermott

1982), a knowledge system for configuring computers. Hearsay (Erman et al. 1980) is a computer program for speech understanding and ELIZA (Weizenbaum 1965) is a computer program that mimics the answers of a client-centered psychotherapist in a dialog with her or his client. For most of these systems, the area of applicability was limited to specific tasks and specific content domains.

The successes and shortcomings of these first-generation systems spurred an in-depth analysis and a better understanding of knowledge representations, search systems (cf., Clancey 1985) and knowledge acquisition, yielding the more advanced second generation systems, with corresponding knowledge models and development tools (Boose and Gaines 1989). The second-generation systems (Breuker and Wielinga 1989, Marcus and McDermott 1989) are usually termed knowledge-based systems rather than expert systems. In the subsequent third generation, knowledge systems are no longer stand-alone systems, that may only interact with a single user at a given time, but instead are systems that interact, negotiate and communicate in multiagent systems (Huhns and Singh 1997) are accessible over the internet (Ambite and Knoblock 1997), represent and reason with common-sense knowledge (Lenat and Guha 1990), and which have brought their developers and users back into the loop (Fischer et al. 1994) to be cooperatively involved in solving problems and evolving new knowledge.

Over the period of three decades, the relatively specific goal of replicating the functionality of human expertise for specific tasks in narrowly defined domains has progressed to the broad field of knowledge systems, which is beginning to integrate the academic research on human knowledge and intelligence, application-oriented developments of the computer and engineering sciences, and the business applications of knowledge systems. Now in the twenty-first century, there are many significant business applications of knowledge systems, for instance in the form of knowledge-management and workflow-management systems, or multiagent systems in intra- and internets that are applied for electronic commerce.

Knowledge systems are viewed as artificial agents that have (assigned or attributed) goals, knowledge, and may interact with other agents in a rational manner. As an alternative to this anthropomorphic view of knowledge systems, similar systems may also be characterized not only as a new sophisticated medium that allows humans to publish static knowledge (such as in a traditional book) but also knowledge that becomes active and via the computer performs actions in an environment that may be remote to the environment of its author (i.e., the human expert of a given subject domain). The human expert may nevertheless remain responsible for these actions. The area of expert system research has become much larger over the years, and philosophical and methodological variety has been added. In combination with the mobility and omnipresence of computers and related new information technologies, such as multimedia systems, inter- and intranets as well as telematics, the field of expert system research continues to enjoy a broad interest from different disciplines and application areas. Expert systems, intelligent artificial agents, or knowledge-management systems, in short knowledge systems, have thus not only become an integral part of modern work environments with businesses attempting to leverage their knowledge by artificial reasoning systems, but may indeed be one of the key factors that is shaping the knowledge society of the twenty-first century.

1. *Changes in Focus and Emphasis Over Time*

In the development of knowledge systems there have been two substantial changes in scientific research, development, empirical tests and practical applications in medicine and business since the 1970s. These two changes define three different generations of expert systems. In the pioneering years, expert systems were developed by rapid prototyping. These systems were typically rule-based systems with a rule interpreter and a set of rules for a given application domain. The developers of the first expert systems were interested in what they saw to be a practical problem. They took professional reasoning in medicine or other areas and practical problem solving, as it had been researched in psychology as their central examples of intelligent behavior. For obtaining the information that was required to build an operational computer program, the developers then adopted interview techniques, protocol analysis (Ericsson and Simon 1984) and information processing psychology (Bruner et al. 1956). Although these systems were fairly successful in a narrow task and application domain, such expert systems were brittle and could hardly be maintained or updated with additional knowledge. The attempts to reuse the inference engine of these systems for different fields of applications (Hayes-Roth et al. 1983) also indicated the rigidity and the lack of potential to accommodate changes in these systems. The hope of building new expert systems by simply reusing the inference engine of a successfully developed expert system and eliciting the rules of a different domain from respective experts did not yield many successes but instead demonstrated the drawbacks of rapid prototyping. By the method of rapid prototyping, the knowledge base and the inference engine had become attuned to and intertwined with one another, thus limiting the reuse of either of the two components. The inference engines were therefore not of much use for the efficient development of expert systems in other application domains. This experience led to the conclusion that knowledge acquisition is the bottleneck in expert system development. Knowledge

acquisition needed to be more than simply asking human experts for their professional guidelines and operating procedures. This research had shown that the power of an expert system was due to the knowledge it possessed rather than to the particular formalisms and inference schemes which it employed. Therefore, knowledge acquisition had to be viewed as a complex modeling activity, where task and domain models had to be developed as the blueprints for expert systems that could subsequently be implemented.

The second-generation knowledge systems were consequently developed in a top-down fashion, where the modeling of an application domain preceded the implementation of the system itself. Top-down implementation and standard methods of software engineering were now utilized for building knowledge systems. The activities which are involved in the knowledge acquisition for second-generation expert systems are similar to developing a cognitive model of human problem solving, decision making or comprehension. The main difference between knowledge acquisition for expert systems and cognitive modeling concerns the empirical information and data which are analyzed. Whereas data from psychological laboratories are employed for cognitive modeling, professional real-world domains and the respective professional performance are analyzed when developing a knowledge-based system.

The modeling approach to knowledge systems yielded a taxonomy of generic tasks, task structures and problem-solving methods (Breuker and Wielinga 1989, Chandrasekaran and Johnson 1993, Benjamins and Fensel 1998) which include heuristic classification (the implicit task model of MYCIN), cover and differentiate, or skeletal-plan refinement. By applying the task structure of heuristic classification, a diagnosis is obtained by abstracting the available data, associating the abstract data with abstract solutions and refining an abstract solution to a specific solution (i.e., the specific diagnosis which best suits the specific data). In skeletal plan refinement, an abstract description of the application of several operators and a description of the application condition becomes instantiated to a specific problem. Such task structures were collected in libraries of expert models that are maintained in a uniform notation (Breuker and van de Velde 1994). Although knowledge systems using first- and second-generation technology continue to be successful in small- or medium-sized applications, the second-generation research unveiled the frame problem (Clancey 1991) and also problems that concern the handling of complexity. The frame problem shows that any sound knowledge system is a closed system and therefore cannot accommodate those novel circumstances which were not foreseen (explicitly or implicitly) by its designers. Depending on the degree of stability of an application area, unforeseeable changes may occur more or less frequently (new diseases or medication in medicine; new manufacturing materials in engineering, etc.). With the words of Karl Popper one might say that any knowledge is preliminary and can be falsified by future observations. The competence of a knowledge-based system is therefore bound by the predictability or rate of unforeseen changes of the given application area, in other words by the quality of the predictions which the designers have indirectly (and mostly unknowingly) built into the system. The processes and conceptual structures of reasonably sized application fields have furthermore been shown to be too complex, so that a knowledge system cannot calculate an answer in due time. Although heuristic processing is frequently used to bypass this problem, the quality of such responses is uncertain.

The third-generation knowledge systems are consequently focused on issues of stability and change over time and across various communities of practice. Knowledge sharing, knowledge reuse and the knowledge evolution that occur over the years in any field of application have become central topics of research. The technologies and theories that are applied as well as the underlying philosophical assumptions have become more diverse. There are knowledge systems that represent a large body of common sense knowledge such as the CYC system (Lenat and Guha 1990) and are represented in formal logic, as well as multiagent systems, where each agent is a small expert system that is able to communicate with other agents. In addition, there are cooperative knowledge systems where the seeding, evolutionary growth and reseeding of a knowledge base (Fischer et al. 1994) is used to accommodate the changes that occur in an environment and where a human user is back in the loop of solving problems and evolving new knowledge. Not all systems subscribe to the assumptions of the symbol level hypothesis (Newell and Simon 1975) but may instead view a knowledge system as a medium for communication among humans. Despite their diversity in architecture and philosophical assumptions, ranging from the representational view of an objective world to the more relativistic assumptions of situated actions (Suchman 1987, Clancey 1997), the ideas and techniques of such systems are already playing a pivotal role in the knowledge society and the emerging global markets of the future. In sum, research on knowledge systems started out with the assumption of human expertise being best modeled by intelligent computation and progressed through a phase which emphasized modeling and knowledge representation to the issues of context, interaction, change, and multiagent systems.

2. *Intellectual Context*

Expert system research stands in the intellectual tradition of the cognitive revolution (cf., Gardner 1987) that substantially changed the research

paradigms of the behavioral and social sciences in the middle of the twentieth century. By conceptualizing thinking as information processing instead or researching behavioral conditioning, the area of behaviorism was succeeded by information processing psychology. As the newly emerging field of artificial intelligence demonstrated, computer systems were consequently believed to execute complex cognitive functions. Language understanding (Kintsch 1974, Schank and Abelson 1977) and language production were similarly in reach of becoming mechanized by symbol manipulation processes. The assumption of humans and computer systems having, in principle, the same mental capacities, as human information processing is crisply stated by Newell and Simon's (1975) symbol level hypothesis: 'a physical symbol has the necessary and sufficient means for general intelligent action.' A physical symbol system is thereby a machine such as a computer that operates on symbol structures. The development of knowledge systems for real-world applications amounts to a practical test of this hypothesis (cf., Schmalhofer et al. 1995). From a scientific point of view, expert system research can be understood as an empirical test of the main hypothesis of the cognitive revolution and the cognitive sciences. Although the assumptions of the physical symbol system hypotheses have been criticized in general, it was not before the late-1980s that alternative views were developed and formalized precisely enough for being considered serious competitors.

The paradigm of situated cognition and parallel distributed processing (subsymbolic or connectionist models) challenged the representational assumptions (or the expert systems of the first and second generation). Situated cognition asserted that the world would be its own best representation and that representations do not come into existence prior to action and will therefore not guide the actions of a natural system in the manner in which knowledge systems do. Instead, human representations would be more like ossified structures that result from and document previous interactions of an agent with its environment. The connectionist approach, on the other hand, emphasized that human information processing, as it is indicated by the neural structures of the brain and the activity patterns that become visible by brain-imaging procedures (fMRI, etc.), does not occur in a serial and discrete manner, which is characteristic of symbol processing systems, but instead occurs in a parallel and continuous manner. The software systems that were developed under these paradigms thus differ in their fundamental assumptions, architectures, and representational assumptions. As such systems attempted to achieve similar but somewhat different functionality to the traditional expert systems, they became competitors as well as a supplementation for the weaknesses of the traditional expert systems. They also yielded an additional fruitful method for modeling human intelligence.

Knowledge system research is also closely related to cognitive modeling, and generally to research in cognitive psychology, where knowledge systems are built so that they would explain and mimic human processes, as they are reified in the results of psychological experiments (cf., Schmalhofer 1998). Cognitive architectures, as they have been proposed by Anderson (1993), Kintsch (1998), and others, provide general but different frameworks for describing complex human information processes in a precise manner. Cognitive psychology has adopted many of the representational assumptions which resulted from the expert and knowledge systems research as psychological constructs for explaining human behavior. Examples are semantic nets and production systems which represent the declarative and procedural knowledge in human information processing.

The discussion of how symbols are and become grounded in the environment (Harnad 1990) has further led to the distinction of knowledge systems as an agent architecture and knowledge systems as a document architecture or medium for communicating (rather than emulating human intelligence). Another scientific debate, initiated by the philosopher John Searle, asserted that the manipulation of symbols is not sufficient for intelligence but instead requires intrinsic intentionality (and that the symbol structures need to be grounded in its environment). Brooks (1991) suggested that intelligent systems need not have representations. By developing systems whose behavior was grounded in the environment he could develop an intelligent creature that did not contain an explicit knowledge representation. Minsky (1990), on the other hand, sees 'symbolness' as a matter of degree. When comparing connectionist and symbolic formalisms, Minsky concludes that mental architectures need both kinds.

Minsky (1986) has proposed modeling mind as a society of agents, composed of simpler agents all the way down. The society of mind blurs together all of the issues of communication and computation. In a society model the interactions among high-level agents may be more akin to communication between separate beings while symbolic interactions between low-level agents may be more akin to computation. This model also challenges another sharp distinction, namely the distinction between symbols used externally between agents (such as in natural language) and the symbols used internally within the mind of each agent. Indeed, people build knowledge bases collaboratively. They discuss and agree about the meanings of symbols in the knowledge bases. They design the behavior of the systems around the agreed upon meanings of the symbols. Viewed in this way, knowledge bases are like blackboards or paper. They are a medium, a place for writing symbols. They augment human memory and human processing with external memory and external processing. Whereas traditional expert systems are very knowledge centric, in the real world knowledge is

distributed. In a university or in a business firm, knowledge is distributed across different schools or different departments. The knowledge to perform a task is therefore often spread out among the people who are involved in these tasks.

3. Issues and Problems

Putting knowledge into computers raises many fundamental questions, some of which have already been treated by philosophers for more than 2,000 years: What is knowledge? Where does it come from? How is it created? How is it held by computers? How is it updated, disseminated, and managed? Through their practical proof of existence, expert systems gave a very specific answer to these questions that is based on a representational theory of mind and more specifically on the physical symbol system hypothesis. As Stefik (1995, p. 21) points out, the essence of this hypothesis is that 'intelligent systems are subject to natural laws and that the natural laws of intelligence are about symbol processing.' A symbol can be read, recognized, and written by a recognizer. A physical symbol system is a machine such as a computer that operates on symbol structures. What a symbol stands for is, however, not determined by a symbol's encoding or by its location because designation is not a physical property of the markings but occurs relative to an observer, namely in the mind of the observer (cf., Kidd 1994). Such a connection between a phenomenon at a specific place and time and a distant symbol is called a *causal coupling*. Because of the existing distance, it is important to distinguish the symbol system from its environment. From a psychological perspective one may ask under what conditions two people may disagree about what a symbol represents. When certain patterns are said to be physical symbols, there must always be a recognizer that identifies the symbols. 'Physical' means that a symbol has a tangible existence. But giving an account of what a symbol represents requires a vantage point, usually outside the symbol system. The tangible knowledge is thus to a certain degree subjective. Although the explicit representations and reasoning methods used in knowledge systems make knowledge tangible, some knowledge may also be intangible and bound to individuals (cf., Nonaka and Takeuchi 1995).

4. Emphasis in Current Research and Future Directions

After research on expert and knowledge systems since the 1970s, there are now at least three distinct facets in research. *Technological-oriented research* is focused on the development on domain-specific knowledge representations and formal ontologies (Gruber 1994), multiagent systems (Huhns and Singh 1997), and collaborative environments for knowledge work (Domingue and Scott 1998). An ontology is thereby defined as an agreement about a shared conceptualization, which includes frameworks for modeling domain knowledge and agreements about the representation of particular domain theories. The denotations of entities in some universe of discourse (e.g., the denotations of classes, relations, functions, or objects) are thereby associated with human-readable text describing what they mean. Formal axioms may furthermore constrain the interpretation and enforce a well-formed use of these denotations. A shared ontology (cf., van Heijst 1995) will thus help to develop a shared virtual world (e.g., the world-wide web, or the intra-/extranets of a business enterprise) in which knowledge systems can ground their beliefs and actions. These knowledge systems may jointly form multiagent systems and have a basis for cooperating and competing with one another (Schmalhofer and van Elst 1999). Knowledge systems, intelligent agents, or Webbots may therefore live and be grounded in a (virtual) environment. The actions performed by one agent constrain and are constrained by the actions of other agents. The issues of organizations are thereby becoming more important: Organizations are systems that constrain the actions of member agents by imposing mutual obligations and permissions. Typically, each agent has incomplete knowledge and capabilities for solving a problem, has its own goals and viewpoint, knowledge thus becomes decentralized, and computation becomes asynchronous. To allow agents to interoperate, communication languages, such as KQML (Finin et al. 1997) have been developed.

The application of knowledge systems is focused on the management of knowledge in businesses and industries. Knowledge-management systems provide the mechanisms to distribute the right knowledge to the right people at the right time. Intelligent agents and multiagent systems carry much promise for the application of knowledge systems in electronic commerce. There is now a larger emphasis on the human and organizational issues related to knowledge. For the emerging knowledge society, knowledge systems will undoubtedly play a pivotal role: Imagine thousands of individuals and groups preparing and making available by gift or sale portions of a large, common distributed knowledge base or knowledge, in the same way as people write textbooks. Knowledge systems will probably become a central component of the blending of computers, televisions, and books into new, integrated forms that are changing the ways in which we work, play, and learn. For the psychological research on human cognition and human intelligence, knowledge systems finally provide the promise of unifying the various empirical and experimental findings across different research paradigms (Newell 1990). This research will continue to unravel the similarities and differences between artificial and

biological intelligence, resulting in a deeper understanding of the logical and biological essence of intelligence.

See also: Chess Expertise, Cognitive Psychology of; Cognitive Science: Overview; Connectionist Approaches; Expertise, Acquisition of; Knowledge Representation; Medical Expertise, Cognitive Psychology of; Science and Technology Studies: Experts and Expertise; Situated Cognition: Origins

Bibliography

Ambite J L, Knoblock C A 1997 Agents for information gathering. *IEEE Expert* **12**: 2–4

Anderson J R 1993 *Rules of the Mind*. Erlbaum, Hillsdale, NJ

Benjamins V R, Fensel D 1998 Special Issue on Problem Solving Methods. *International Journal of Human–Computer Studies* **49**(4): 305–650

Boose J H, Gaines B R 1989 Knowledge acquisition for knowledge-based systems: notes on the state-of-the-art. *Machine Learning* **4**: 377–94

Breuker J A, van de Velde W (eds.) 1994 *The CommonKADS Library of Expertise Modeling*. IOS, Amsterdam

Breuker J A, Wielinga W J 1989 Model driven knowledge acquisition. In: Guida P, Tasso G (eds.) *Topics in the Design of Expert Systems*. North-Holland/Elsevier, Amsterdam, pp. 265–96

Brooks R A 1991 Intelligence without representation. *Artificial Intelligence* **47**: 139–59

Bruner J S, Goodnow J J, Austin G A 1956 *A Study of Thinking*. Wiley, New York

Buchanan B G, Shortliffe E H 1984 *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA

Chandrasekaran B, Johnson T R 1993 Generic tasks and task structures: history, critique and new directions. In: David J M, Krivine J P, Simmons R (eds.) *Second Generation Expert Systems*. Springer, Berlin

Clancey W J 1985 Heuristic classification. *Artificial Intelligence* **27**: 289–350

Clancey W J 1991 The frame of reference problem in the design of intelligent machines. In: VanLehn K (ed.) *Architectures for Intelligence. The Twenty-second Carnegie-Mellon Symposium on Cognition*. Erlbaum, Hillsdale, NJ, pp. 357–423

Clancey W J 1997 *Situated Cognition: on Human Knowledge and Computer Representations*. Cambridge University Press, New York

Domingue J, Scott P 1998 KMI Planet: putting the knowledge back into media. In: Eisenstadt M, Vincent T (eds.) *The Knowledge Web: Learning and Collaborating on the Net*. Kogan, pp. 173–84

Duda R O, Reboh R 1984 Artificial intelligence and decision making: the PROSPECTOR experience. In: Reitman W (ed.) *Artificial Intelligence Applications for Business*. Ablex, Norwood, NJ, pp. 111–47

Engelman C 1971 The legacy of MATHLAB 68. In: *Proceedings of the Second Symposium on Symbolic and Algebraic Manipulation, Los Angeles*

Ericsson K A, Simon H A 1984 *Protocol Analysis. Verbal Reports as Data*. MIT Press, Cambridge, MA

Erman L D, Hayes-Roth F, Lesser V R, Reddy D 1980 The HEAR-SAY-II speech-understanding system: integrating knowledge to resolve uncertainty. *Computing Surveys* **12(2)**: 213–53

Finin T, Labrou Y, Mayfiled J 1997 KQML as an Agent Communication Language. In: Bradshaw J (ed.) *Software Agents*. MIT Press, Cambridge, MA

Fischer G, McCall R, Ostwald J, Reeves B, Shipman R 1994 Seeding, evolutionary growth and reseeding: supporting the incremental development of design environments. In: *Proceedings of CHI-94: Human Factors in Computing Systems, Boston, MA, April 24–28 1994*. ACM, New York, pp. 292–8

Gardner H 1987 *The Mind's New Science: a History of the Cognitive Revolution*. Basic Books, New York

Gruber Th 1993 Towards principles for the design of ontologies used for knowledge sharing. In: Guarino N, Poli R (eds.) *Formal Ontologies in Conceptual Analysis and Knowledge Representation*. Kluwer, Boston

Harnad S 1990 The symbol grounding problem. *Physica* **42**: 335–46

Hayes-Roth F, Waterman D A, Lenat D B 1983 *Building Expert Systems*. Addison-Wesley, Reading, MA

Huhns M N, Singh M P 1997 *Readings in Agents*. Morgan Kaufman, San Francisco

Kidd A 1994 The marks are on the knowledge worker. In: *Proceedings of CHI-94: Human Factors in Computing Systems, Boston, MA, April 24–28 1994*. ACM, New York, pp. 186–91

Kintsch W 1974 *The Representation of Meaning in Memory*. Erlbaum, Hillsdale, NJ

Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK

Lenat D B, Guha R V 1990 *Building Large Knowledge-Based Systems. Representation and Inference in the Cyc Project*. Addison-Wesley, Reading, MA

Lindsay R, Buchanan B G, Feigenbaum E A, Lederberg J 1980 *DENDRAL*. McGraw-Hill, New York

Marcus S, McDermott J 1989 SALT: a knowledge-acquisition language for propos-and-revies systems. *Artificial Intelligence* **39**: 1–37

McDermott J 1982 R1: A rule-based configurer of computer systems. *Artificial Intelligence* **19**: 39–88

Miller R A, Pople H E, Myers J D 1982 Internist-1: an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine* **307**: 468–76

Minsky M 1986 *Society of Mind*. Simon and Schuster, New York

Minsky M 1990 Logical vs. analogical or symbolic vs. connectionist or neat vs. scruffy. In: Winston P H, Shellard S A (eds.) *Artificial Intelligence at MIT: Expanding Frontiers*. MIT Press, Cambridge, MA, pp. 218–43

Moses J 1975 *A MACSYMA Primer*. Mathlab Memo No. 2, Computer Science Laboratory. Massachusetts Institute of Technology, Cambridge, MA

Newell A 1982 The knowledge level. *Artificial Intelligence* **18**: 87–127

Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

Newell A, Simon H A 1976 Computer science as empirical enquiry: symbols and search. *Communications of the ACM* **19**: 113–26

Nonaka I, Takeuchi H 1995 *The Knowledge-creating Company*. Oxford University Press, Oxford, UK

Schank R C, Abelson R 1977 *Scripts, Plans, Goals and Understanding*. Erlbaum, Hillsdale, NJ

Schmalhofer F 1998 *Constructive Knowledge Acquisition: A*

Computational Model and Experimental Evaluation. Erlbaum, Mahwah, NJ
Schmalhofer F, Reinartz R, Tschaitschian B 1995 A unified approach to learning in complex real world domains. *Applied Artificial Intelligence* **9**: 127–56
Schmalhofer F, van Elst L 1999 An oligo-agents system with shared responsibilities for knowledge management. In: Fensel D, Studer R (eds.) *Knowledge Acquisition, Modeling and Management*. Springer, Berlin, pp. 379–84
Stefik M 1995 *Introduction to Knowledge Systems*. Morgan Kaufman, San Francisco
Suchman L 1987 *Plans and Situated Actions: The Problem of Human–Machine Communication*. Cambridge University Press, Cambridge
van Heijst G 1995 The role of ontologies in knowledge engineering. Doctoral dissertation, University of Amsterdam
Weizenbaum J 1965 ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* **9**(1): 36
Wolfram D A 1995 An appraisal of Internist-1. *Artificial Intelligence in Medicine* **7**: 93–116

F. Schmalhofer

Expert Systems in Medicine

Medical expert systems (MESs) make predictions or treatment recommendations, acting like a doctor. They have been studied and typically perform near the level of a doctor. For a variety of reasons, which are discussed, MESs are not in widespread use.

1. Introduction

Medical expert systems are a class of artificial intelligence systems that make predictions or treatment recommendations as is done by a medical expert. The idea that a computer system could act as a medical expert is an exciting concept. Images of universal access to this expert via the Internet are part of this fantasy, as is the hope that such a system would provide state-of-the-art (perhaps someday even flawless) patient care. Much progress has been made, but important challenges remain.

It is important to differentiate MESs from other systems, such as decision support systems (see *Decision Support Systems*). A MES actually makes a prediction or treatment recommendation, as would a human expert (i.e., a clinician). This is the key difference from a decision support system, which is designed to help the clinician make a better decision than he or she would in the absence of the MES. The MES can actually stand alone with its own prediction or recommendation. At the conclusion of the MES encounter, the clinician may accept or reject the output of the MES, but in theory, the MES recommendation with no further refinement could be compared to that of the human expert. Thus, there would really be three potential predictions or recommendations: the human expert's alone, the MES alone, and the human expert's (possibly revised) recommendation after viewing the prediction made by the system. With a decision support system, only the first and last recommendations could be compared.

Physical representations of MESs vary greatly. While in general the representational form is of little concern, it occasionally may affect MES accuracy or ease of use. For example, a table may oversimplify relative to a graph, text, or computerized algorithm, resulting in a loss of accuracy because the table may collapse levels of continuous variables into ranges. The patient with an age of 40 may have a prognosis different from the patient of age 50, yet they would be treated identically if the table grouped patients as 'Age 40–50.' Note that computerization of the MES is not necessarily required, unless ease of use or accuracy are affected. For example, tedious manual calculations may become inaccurate if not computerized, though having to have a computer at the patient's bedside may complicate ease of use. The bottom line, however, is quality of the recommendations, which is most often measured as predictive accuracy. For this reason, the reader is referred to *Decision Support Systems* for technology issues, since they are not particularly unique to MESs.

The focus on accuracy of the MES recommendations and predictions is not meant to dismiss issues of innovation diffusion or political championship. Clearly, a perfectly good MES, which does not make the clinician more efficient and/or more effective, will never be widely adopted. The system must be at least as effective as the human expert. An MES that is more efficient but makes more mistakes than its human counterpart will not, and indeed should not, be utilized in a litigious medical environment where quality of care is scrutinized. However, a system which is of comparable effectiveness but higher efficiency would be particularly welcome in a capitated provider system where the clinician is not directly compensated for each patient encounter or procedure. Alternatively, a system that is more effective than the human expert might still be enthusiastically accepted even if efficiency is impaired. The reason for this is that some patients will seek the most effective care for their condition and look for the clinician or provider who employs the most effective technology if it produces superior patient outcomes. Of course, MESs which increase both efficiency and effectiveness will have an easy time with diffusion, but examples of such MESs are rare.

2. Classes of Medical Expert Systems

There are two major classes of medical expert systems in development or use today. The most popular are *predictive* MESs. This type of MES strictly makes

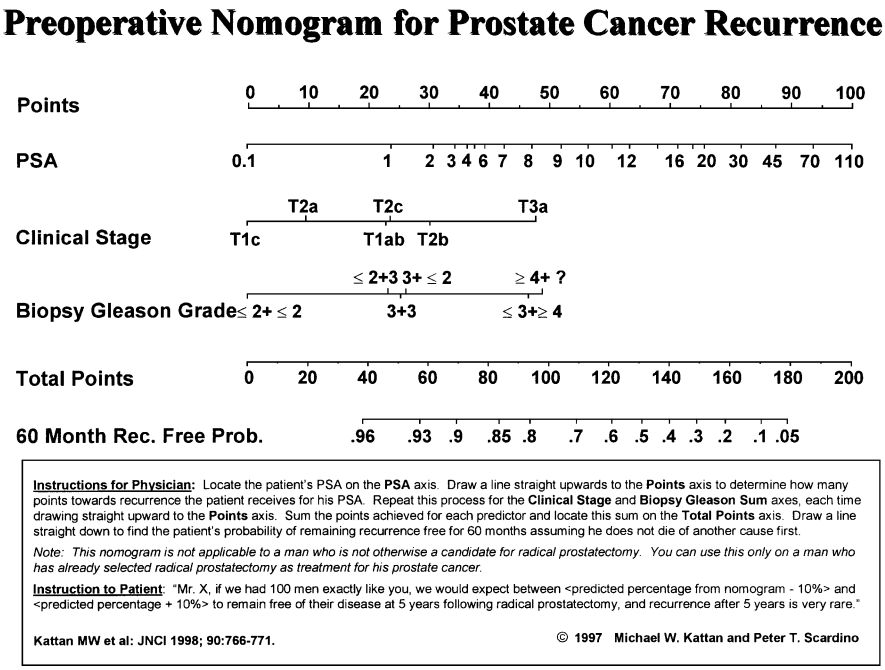

Figure 1
Example of a nomogram

predictions of some medical endpoint or outcome that is not immediately known. Diagnostic systems would be included under this category because these systems predict what disease the patient currently has, which is not immediately known until a (perhaps expensive, time-consuming, or painful) gold standard evaluation is performed. Generally, such MESs rely on readily available predictor variables, and the outcome being predicted usually requires an invasive procedure or follow-up time before it can be known with certainty. Thus, the clinical values usually associated with being able to predict the endpoint are that of (a) lead time that a future event will or will not occur or (b) possibly sparing the patient from a painful, delayed, or expensive procedure. The second most popular MES class is *prescriptive*. Prescriptive MESs make treatment recommendations and possibly partially implement them. A distinction between predictive and prescriptive MESs is that the outcome is not affected by the predictive MES but may be affected by the prescriptive MES since a treatment is attached to the latter recommendation.

2.1 Predictive Medical Expert Systems

Numerous examples of predictive MESs exist. These are essentially prediction tools, and appear in a variety of forms. Some of the more common forms are charts, tables, graphs, or 'nomograms' which are scales with numbers. Figure 1 is a nomogram that serves as a predictive MES for prostate cancer (Kattan et al. 1998). The reason for the graphical rather than tabular depiction is to avoid, as much as possible with printed media, categorizing the continuous variables, which would have to be done if a table were used. Alternatively, predictive MESs may be computerized as traditional software or enabled to run on dedicated handheld calculators.

The most desirable feature of a predictive MES is predictive accuracy. For a new predictive MES to become popular, it must predict more accurately than the traditional method of prediction in that clinical setting, which may be an older prediction tool or simply clinical judgment. With the latter, a substantial body of research has been devoted to studying human prediction and its limitations (Hogarth 1988). It has been shown quite clearly that human beings, even clinical experts, have biases when it comes to prediction. For example, humans do not remember all cases in the past with equal ability; certain cases stand out, and these are given more weight when trying to predict outcomes of future cases. Also, humans tend to predict the outcome they want to occur, rather than the outcome with the highest actual probability. For these reasons and others, predictive MESs often outperform human experts, even when the experts have more information available than that utilized by the predictive MES. The likely reason for this seems to be that humans do not properly weight the additional

information when mentally revising the predictive MES output.

Predictive MESs are typically created based on statistical analyses which form the underlying prediction model. The statistical analysis may have been performed using traditional statistical procedures (such as multiple regression analysis) or with computationally intensive procedures (such as neural networks). Typically, the actual outcome is modeled by attempting to predict it using other variables contained in a dataset. Alternatively, one could model the judgments of a human expert. With this approach, what the expert decided is the outcome being modeled, rather than what actually occurred with the patient. However, the resulting model would not have much chance of being more accurate than the expert(s) modeled, so this type of MES would only succeed in the setting where (a) a premier expert were modeled and (b) the predictive MES was needed for application outside that original setting. An example of this could be a treatment recommendation, where what the expert decided serves as the gold standard, and where success or failure of the expert's recommendation might be difficult to judge. In settings where the prediction task is deterministic (i.e., 100 percent predictive accuracy is possible with existing data), statistical predictions models may not be the preferred basis for the MES. Here, traditional MES chaining algorithms (Ignizio 1991) could be implemented in lieu of a statistical prediction formula. However, interesting deterministic medical settings are extremely rare.

The studies that have compared human experts with predictive MESs have been quite convincing. Ignoring the issue of potential publication bias, human experts only rarely outperform a predictive MES, and quite often predict less accurately than a predictive MES. This has been illustrated in a variety of medical settings, usually involving clinical experts. Dawes et al. (1989) present one of the more comprehensive reviews of this literature. In most cases, it appears that the predictive MES is not routinely embedded in electronic medical record system software. Rather, a separate data entry step is required for the human expert to obtain the predictive MES prediction. Despite this slight decrease in efficiency of the expert, numerous applications exist where this tradeoff is apparently worthwhile. An example is the commercially successful APACHE software for predicting the probability of death for patients admitted to intensive care (Knaus et al. 1995).

2.2 *Prescriptive Medical Expert Systems*

Prescriptive MESs are more comprehensive than predictive MESs since the former make treatment recommendations that incorporate predictions in their recommendation. With a prescriptive MES, a judgment is made regarding appropriate care, not just anticipated outcome. Many prescriptive MESs at least partially implement their recommendations, such as a drug dosing application. In this setting, the prescriptive MES prescribes a treatment as well as monitors its effectiveness, possibly making adjustments and revising its predictions based on feedback within the same patient. The common example from this class is drug dosing, where numerous repeated predictions at the individual patient level may be made (i.e., the level of drug to deliver). The system manages those predictions over time based on its performance within the individual patient (i.e., how well the last dose level controlled the medical condition and what future dose changes need to be made).

As with predictive MESs, the literature is quite consistent concerning prescriptive MESs. Properly developed and validated prescriptive MESs seem to perform at a level that is at least as accurate as a clinical expert, and typically exceed that of a clinical novice (Raschke et al. 1996). Also, as with predictive MESs, clinicians in the presence of a prescriptive MES tend to perform better than they did prior to the prescriptive MES. However, prescriptive MESs are available only in limited domains, where the medical process has been well studied and understood. Furthermore, it seems likely that this literature has the strong potential to suffer from publication bias, in that poorly performing prescriptive MESs are underreported. In addition, results of a prescriptive MES are somewhat more difficult to validate and analyze. The reason for this is that usually the patient can only be treated by the MES or by the clinician, but not by both. With a predictive MES, the same patient can usually have his outcome predicted by both the MES and the clinician who is blind to the MES prediction. With a prescriptive MES, patients must be randomized to receive treatment from either the MES or the clinician. Since the same patient can no longer serve as his own control, the sample size requirements and study design become more complicated in studies of prescriptive MES performance. Furthermore, patients may have difficulty with consenting to be randomized to treatment by either a computer or a doctor. An alternative study design is to randomly provide the clinician with the MES recommendation and observe whether outcomes are improved when the clinician has this recommendation, but interpretation of this design is also complicated because it is not a true test of the prescriptive MES performing alone.

3. *Future Development of Medical Expert Systems*

Predictive MES development is likely to continue to flourish. These systems have been well established as valuable, and much research is directed to the development and comparison of alternative systems.

Adoption of these systems takes time, but the Internet has facilitated this with promotion of particular systems as well as offering end user and physician interaction with these systems. For example, women can now estimate their risk of being diagnosed with breast cancer using a National Cancer Institute web page. While it seems that prescriptive MESs also should play a larger role in healthcare delivery, this process will be slower. Since they are more comprehensive, they are harder to develop and validate. Nonetheless, these systems have good potential to improve efficiency and effectiveness of patient care.

An important challenge for prescriptive MESs is to incorporate patient preferences into medical decision-making. Quality of life can be as important as quantity of life in particular medical situations. It is clear that the goal of a medical test or treatment cannot be survival with complete disregard to quality of life. The medical decision is not complex when, for example, two treatments have comparable survival but differing quality of life: the treatment with higher quality of life would be preferred. However, it is not uncommon to find medical options where one treatment has better survival but worse quality of life than the other treatment. Obviously, the magnitudes of the survival and quality of life differences become an issue. A minor decrease in quality of life might be worth a major difference in survival, but the simultaneous consideration of these measures is nonetheless required.

The current paradigm for simultaneous consideration of quantity and quality of life, when comparing treatment or testing options, is that of clinical decision analysis. This technique incorporates outcome probabilities and patient preferences to compute quality adjusted survival times. This is typically done by assuming that the patient progresses through a series of health states, each with its own utility (measured on a scale where 0 equals death and 1 equals perfect health). Quality adjusted survival is computed by weighting the survival time in a health state by its utility, and summing those products across the life expectancy of the patient. The appeal of this approach is that the preferred treatment can be judged even when treatments differ with respect to quantity and quality of life. Via the quality adjusted survival computation approach, magnitudes of survival probabilities and quality of life preferences are considered.

Though the decision analytic approach is an apparently rational one, many challenges to its implementation remain. The preferred method of measuring utility varies greatly. Unfortunately, there is little data on longitudinal utility assessment in the literature, so questions of how well patients adapt to their health are largely unanswered. This body of literature is conflicting, with some studies indicating that patients rate future health states lower than when patients rate the health state upon entering it. Also of concern is the patient's attitude towards risk, or similarly, how he or she discounts future health. For these reasons, it is not clear the individual patient should even want to maximize his quality adjusted life years without first incorporating some of these factors. Until these theoretical issues are resolved, or until the decision analytic approach is demonstrated to improve outcomes in a randomized clinical trial, adoption of this method as a basis for prescriptive MESs will be hampered.

4. *Future Evaluation of Medical Expert Systems*

In light of the difficulty associated with judging what an optimal treatment should provide when both quality and quantity of life are important, alternative evaluation methods of medical expert systems will become necessary. Survival and quality of life by themselves are often limited endpoints, much the same way intermediate endpoints do not necessarily translate into final endpoints. Thus, judging whether an MES does better than status quo can be challenging. Work is just beginning in the definition of more global endpoints, such as regret-free survival. Here, a patient is followed until death or regret of treatment choice (either indicates treatment failure), and when applied in the context of a randomized trial, would appear to offer a measure which also incorporates quantity and health-related quality of life but with possibly fewer technical implementation and interpretation problems.

5. *Conclusions*

Predictive (predict outcomes) and prescriptive (make treatment recommendations) MESs have potentially valuable roles in the delivery of medical care. Validation of these systems is critically necessary and difficult, but numerous examples of successful medical expert systems are available.

See also: Categorization and Similarity Models; Decision Research: Behavioral; Decision Support Systems; Medical Expertise, Cognitive Psychology of

Bibliography

Dawes R M, Faust D, Meehl P E 1989 Clinical versus actuarial judgment. *Science* **243**: 1668–74

Hogarth R M 1988 *Judgement and Choice: The Psychology of Decision Making*, 2nd edn. Wiley, New York

Ignizio J P 1991 *An Introduction to Expert Systems*. McGraw-Hill, New York

Kattan M W, Eastham J A, Stapleton A M F, Wheeler T M, Scardino P T 1998 A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *Journal of the National Cancer Institute* **90**: 766–71

Knaus W A, Harrell F E, Lynn J, Goldman L, Phillips R S, Connors A F Jr, Dawson N V, Fulkerson W J Jr, Califf R M, Desbiens N et al. 1995 The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine* **122**: 191–203

Raschke R A, Gollihare B, Peirce J C 1996 The effectiveness of implementing the weight-based heparin nomogram as a practice guideline. *Archives of Internal Medicine* **156**: 1645–9

M. W. Kattan

Expert Testimony

Expert testimony is presented in legal proceedings when a judge or jury needs assistance evaluating a material fact in a court proceeding. Qualified by education, skill, or technical expertise, an expert provides fact finders (i.e., judges or juries) with specialized information to educate them about case-relevant issues that are not within the common knowledge of fact finders. In this article, expert testimony in common law and civil law systems is reviewed and the empirical research on the effects of expert testimony on fact finders' decisions are examined.

1. *Expert Testimony in Common Law Systems*

In common law systems, one or both parties may proffer experts in a legal proceeding. Judges then make decisions about whether the proffered expert testimony will be admitted into evidence. Although each country has its own unique law regarding the admissibility of expert testimony, there are several striking similarities across several countries, including the USA, Australia, England, and Canada (Gatowski et al. 1996).

In each of these countries, judges may only admit expert evidence that is relevant to a factual issue in contention. Stated another way, expert evidence must be probative (i.e., make some fact in contention more or less probable). In addition, judges must weigh the probative value vs. the prejudicial value of the expert evidence. Even if the evidence is probative, it may be inadmissible if the value of this relevant information is outweighed by the risk that the evidence may confuse, mislead, or unfairly bias the jury. To be admitted at trial, the expert evidence must also be helpful; that is, the evidence must assist the trier of fact in making a decision. If the expert evidence is within the common knowledge of the fact finder, then the evidence will not be helpful and should not be admitted.

A recent trend in case law on the admissibility of expert evidence is increased attention to the reliability or scientific validity of the information that experts are providing in court. Recent decisions in both the USA and Canada have addressed concerns about the proliferation of unreliable or junk science in court. These decisions have delineated specific criteria that judges must use to assess the reliability of expert evidence, such as potential error rates, the use of appropriate methodology, and the general acceptance of the evidence by other experts in the field. Courts in England and Australia have not provided specific criteria for evaluating the reliability of expert evidence, however, they have ruled that expert evidence must be based on legitimate science or expertise (Gatowski et al. 1996).

2. *Expert Testimony in Civil Law Systems*

Although there are often provisions for judges in common law systems to find their own court-appointed experts, this option is rarely exercised (Gross 1991). In civil law systems, however, experts are usually appointed by the judge presiding in a case, in which the issues to be resolved are beyond the understanding of the court. Moreover, the laws regarding expert testimony in these systems focus on the procedures used to appoint an expert, the definition of the scope of the expert's duties, and whether the expert's testimony or opinion is binding upon the presiding judge. For example, French courts maintain lists of experts that meet certain qualifications, although judges are free to appoint an expert that is not on the list (Taylor 1996). In French, German, and Korean courts, experts are charged with a specific task or tasks and their findings are presented to the court in a written report rather than in oral testimony (Bastuck and Gopfert 1994, Lee 1997, Taylor 1996). In Korea and France, the findings of the expert are not binding upon the judge (Lee 1997, Taylor 1996).

3. *Empirical Research on Expert Testimony*

Psychological researchers have primarily used trial simulation techniques to examine the effects of expert testimony on fact-finder judgments. These trial simulations involve presenting mock jurors (who may be college undergraduates, members of the community, or jurors who have reported to jury duty) with written, audiotaped, or videotaped trial stimuli. The trial stimuli vary the presence or absence of expert testimony or some characteristic of the expert testimony.

3.1 *Empirical Research on Adversarial Expert Testimony*

Research on the effects of expert testimony on jury decisions became common in the 1980s. Most of this research has examined whether the testimony of an expert called by one of the parties (an adversarial

expert) affected juror or jury decisions. For example, early research examined the question of whether expert testimony influenced juror judgments of guilt. Trial simulation studies demonstrated that when jurors heard expert testimony presented by the defense they were more likely to return defense verdicts than were jurors who did not hear expert testimony (e.g., Schuller and Cripps 1998). In contrast, jurors who heard expert testimony presented by the prosecution were more likely to convict the defendant than were jurors who did not hear prosecution expert testimony (e.g., Brekke and Borgida 1988).

After establishing that expert testimony influenced juror decisions, researchers began to explore whether certain variables moderate the effects of expert testimony on juror judgments. For example, expert testimony has greater influence on jury decisions when it appears early rather than late in the trial (Brekke and Borgida 1988, Schuller and Cripps 1998). There is some speculation that early presentation is advantageous because it allows the expert evidence to serve as a framework for understanding other trial evidence. Several studies have also demonstrated that expert testimony is more influential if the expert explicitly links the scientific research to the facts of the case under consideration (Brekke and Borgida 1988, Schuller 1992).

Eventually researchers began to study whether expert testimony improves the quality of juror decisions. Cutler and colleagues (Cutler et al. 1989) presented jurors with a trial simulation in which they manipulated factors affecting the reliability of eyewitness identifications (e.g., presence of a weapon, disguised perpetrator, retention interval). They also manipulated the presence of expert testimony describing how these factors influence the accuracy of identifications (e.g., people are less likely to recognize a perpetrator who wore a disguise). In the absence of expert testimony, jurors' decisions were unaffected by factors known to decrease the reliability of eyewitness identifications. After hearing expert testimony on the topic, jurors became sensitive to these factors and used this information in making decisions. Expert testimony's sensitization effect is more likely to occur when the expert explicitly links the scientific research and the facts of the case (Kovera et al. 1997).

3.2 Empirical Research on Nonadversarial Expert Testimony

There may be important differences between the effects of adversarial expert testimony and the testimony provided by a nonadversarial or court-appointed expert. Some have argued that because a nonadversarial expert is proffered by the court rather than one of the parties in a case, jurors may be less likely to critically evaluate the evidence provided by that expert. Indeed testimony from a court-appointed expert does not have the same sensitizing effect that adversarial expert testimony has. In one study, nonadversarial expert testimony on the factors influencing the reliability of eyewitness identifications caused jurors to become skeptical of the reliability of all eyewitness identification rather than just those identifications made under unreliable conditions (Cutler et al. 1990). This pattern of effects suggests that jurors are not systematically processing the expert evidence presented by the court-appointed expert, therefore, they do not understand the subtleties of that testimony (e.g., that only some eyewitness identifications are unreliable, not all).

Another concern raised about the use of court-appointed or nonadversarial experts is that their testimony will be prejudicial because jurors may accord it undue weight. Trial simulation research comparing juror evaluations of nonadversarial and adversarial expert testimony demonstrated that jurors find both types of experts equally credible and that their testimony is given similar weight by jurors when making trial judgments (Brekke et al. 1991). Moreover, jurors seem to remember evidence presented by the adversarial expert better than they remembered evidence presented by the nonadversarial expert, providing additional evidence that jurors do not scrutinize nonadversarial expert testimony as closely as they scrutinize adversarial testimony.

3.3 Can Fact-finders Distinguish Between Valid and Flawed Expert Evidence?

Because most of the social scientific research on expert evidence has relied on trial simulations in which mock jurors make decisions, little experimental evidence addresses how judges evaluate expert testimony. This void in the research literature is particularly problematic because many countries that allow expert testimony have judges, not juries, serving as fact-finders during trials. Moreover, most research in this area has assumed that it is desirable for expert testimony to influence the decisions of fact finders. This assumption neglects the possibility that flawed expert evidence may be presented to the court, requiring that judges or jurors be able to differentiate between flawed and reliable expert evidence.

A recent trend in the empirical research is to question whether fact finders have the ability to make competent judgments about the quality of expert evidence. In one study, researchers asked judges to evaluate the admissibility of expert testimony that varied in its scientific validity after reading a description of a psychological experiment that was fundamentally flawed (e.g., was missing a control group) or was scientifically valid (Kovera and McAuliff 2000). Judges' decisions to admit the expert evidence at trial and their judgments of evidentiary reliability were unaffected by the methodological

quality of the expert evidence. Trial simulation studies also show that jurors' trial judgments are unaffected by variations in methodological quality (Kovera et al. 1999) and that a scientifically informed cross-examination that highlights the methodological flaws does not sensitize jurors to the problems with flawed studies (Kovera et al. 1999). Further research is needed to determine whether attorneys' pretrial motions objecting to the admissibility of expert testimony will help judges and whether judicial instruction or opposing experts will help jurors make more informed decisions about the quality of expert testimony.

See also: Expert Witness and the Legal System: Psychological Aspects; Forensic Psychology; Judges; Juries; Scientific Evidence: Legal Aspects

Bibliography

Bastuck B, Gopfert B 1994 Admission and presentation of evidence in Germany. *Loyola of Los Angeles International and Comparative Law Journal* **16**: 609–27

Brekke N, Borgida E 1988 Expert psychological testimony in rape trials: A social-cognitive analysis. *Journal of Personality and Social Psychology* **55**: 372–86

Brekke N J, Enko P J, Clavet C, Seelau E 1991 Of juries and court-appointed experts: The impact of nonadversarial versus adversarial expert testimony. *Law and Human Behavior* **15**: 451–75

Cutler B L, Dexter H R, Penrod S D 1989 Expert testimony and jury decision making: An empirical analysis. *Behavioral Sciences and the Law* **7**: 215–25

Cutler B L, Dexter H R, Penrod S D 1990 Nonadversarial methods for sensitizing jurors to eyewitness evidence. *Journal of Applied Social Psychology* **20**: 1197–207

Gatowski S I, Dobbin S A, Richardson J T, Nowlin C, Ginsburg G P 1996 The diffusion of scientific evidence: A comparative analysis of admissibility standards in Australia, Canada, England, and the United States, and their impact on social and behavioural sciences. *Expert Evidence* **4**: 86–92

Gross S R 1991 Expert evidence. *Wisconsin Law Review* **6**: 1113–232

Kovera M B, Gresham A W, Borgida E, Gray E, Regan P C 1997 Does expert psychological testimony inform or influence juror decision-making? A social cognitive analysis. *Journal of Applied Psychology* **82**: 178–91

Kovera M B, McAuliff B D 2000 The effects of peer review and evidence quality on judge evaluations of psychological science: Are judges effective gatekeepers? *Journal of Applied Psychology* **85**: 574–86

Kovera M B, McAuliff B D, Hebert K S 1999 Reasoning about scientific evidence: Effects of juror gender and evidence quality on juror decisions in a hostile work environment case. *Journal of Applied Psychology* **84**: 362–75

Lee E I 1997 Expert evidence in the Republic of Korea and under the US Federal Rules of Evidence: A comparative study. *Loyola of Los Angeles International and Comparative Law Journal* **19**: 585–632

Schuller R A 1992 The impact of battered woman syndrome evidence on jury decision processes. *Law and Human Behavior* **16**: 597–620

Schuller R A, Cripps J 1998 Expert evidence pertaining to battered women: The impact of gender of expert and timing of testimony. *Law and Human Behavior* **22**: 17–31

Taylor R F 1996 A comparative study of expert testimony in France and the United States: Philosophical underpinnings, history, practice, and procedure. *International Law Journal* **31**: 181–213

M. Kovera

Expert Witness and the Legal System: Psychological Aspects

Due to the diversity of legal arenas and procedures internationally, this article will focus on the use of expert witness in the United States. The points raised here will apply to varying extents to other judicial systems.

1. What is an 'Expert' Witness?

An expert witness is one who, by virtue of education, training, and experience, is able to assist the trier of fact (i.e., judge or jury) in understanding scientific or other specialized knowledge that is beyond the ken of the average person. For example, an expert might discuss methods or materials for building bridges that laypersons or nonprofessionals might not normally understand, such as principles of engineering or the stress that particular alloys can withstand.

The evidence and subject matter on which experts may rely and testify about differ from lay witnesses in at least two fundamental ways. First, lay witnesses may testify only as to what they have done or observed directly and cannot offer conclusions or opinions based on those matters. In contrast, experts may offer opinions and interpret the facts at issue so as to assist the judge or jury in understanding technical or scientific materials that they might not otherwise be able to grasp sufficiently. For example, a lay witness might be limited to describing a person's appearance or testifying that an individual talked about suicide, but an expert in psychology might be able to discuss the scientific literature related to depression and its treatment, and draw conclusions about cause and prognosis in a given case.

Second, experts are often permitted to review and describe various forms of evidence that might be deemed inadmissible with lay witnesses, such as statements within learned treatises or other types of materials usually classified as hearsay. For such evidence to be admissible, it should be of a type that experts in the respective area normally rely upon in formulating conclusions or opinions. For example,

psychologists may routinely draw inferences from the written notes and reports of other professionals to analyze an individual's psychological state at a certain point in time.

To the extent that an expert is qualified to make a unique contribution to the understanding of the fact-finder, that testimony may be admitted. However, if an expert's testimony is deemed to be within the scope of laypersons' understanding, it may be barred. For example, if the expert is merely reiterating another individual's statements absent any additional interpretation, or describes conditions or events that are within the common experience of most people, that testimony may be excluded. Expert testimony must also be deemed relevant to the issue at hand. Thus, a world authority might not be allowed to testify on the psychological effects of trauma if trauma is not at issue in the case. Expert testimony may be barred completely or partially in a given case. For example, an expert might be allowed to testify about the presence of a disorder but not its cause.

Although the expert's defined role of 'assisting the jury' may sound neutral, in practice, one or the other side in a legal case usually retains a particular expert in order to advance that side's position. This does not mean that experts are necessarily partisan advocates who provide whatever testimony will help the side that has retained them. A forthright expert operating in a fair and scientific manner may well reach a position consistent with that the retaining attorney hopes or plans to put forth. However, in part because experts often become involved in issues or questions that are complex or ambiguous or that have been the subject of research with conflicting results, and also because, unfortunately, some experts are less than forthright, it is very common for both sides to retain their own expert and for these experts to reach and present opposing conclusions. Such 'battles of experts' have been a source of much criticism and lead to the perception, sometimes correct and sometimes not, that expert testimony has simply been bought from 'hired guns.'

2. *The Role of the Expert Witness*

Expert witnesses began to appear in England during the thirteenth century in the form of specially commissioned juries of experts who heard or investigated matters related to their trade, or expert individuals who consulted directly with judges in areas where scientific or technical expertise was desired. As the judicial process evolved over the past 800 years to its current adversarial form, expert witnesses became increasingly less likely to be appointed by the court as fact-finders or neutral advisors (Landsman 1995). Today, expert witnesses are usually retained by, and testify for, specific parties to a legal issue, although under some circumstances judges may seek additional consultation or advice separate from the legal proceedings when deciding an issue. An expert witness may also be selected by a judge, or the parties in dispute may agree upon an expert. In addition, psychologists who were originally involved in a treating role may be asked to testify about the individual for whom they provided services.

In the early history of expert witnesses, physicians were most likely to address issues related to mental health. However, as psychology and allied disciplines established themselves as independent specializations, such matters increasingly became the domain of these fields. Today, expert witnesses in psychology testify in a wide variety of civil and criminal cases and in various adjudication situations such as civil commitment and different matters heard in family or juvenile court. The standards for granting an individual expert status vary in different jurisdictions by statute, precedent, tradition, and judge, and no universal standard exists. Additionally, the rules regulating the use of experts are subject to abrupt change as a result of new legislation or court decisions. However, some general characteristics and trends can be discussed.

In the criminal setting, a psychologist may testify as to a defendant's competency to stand trial or participate in various facets of the investigation, prosecution, and disposition of a case (e.g., competency to confess, to waive legal rights, or to enter a 'guilty' plea). A psychologist may offer opinions about the defendant's mental state at the time of the offense, in particular when an insanity defense is raised, or offer testimony related to psychological conditions or diagnoses that might mitigate or contribute to the defendant's responsibility (e.g., brain damage leading to poor impulse control or the effects of alcohol and drugs). Psychological research on various aspects of trial proceedings (e.g., the degree to which jurors follow a judge's instructions) and evidence (e.g., the reliability of eye-witness testimony) may also be introduced during trial as each side attempts to make its case. A psychologist may also become involved during the sentencing phase of a criminal trial, where the prospects for rehabilitation, the prediction of recidivism, and the severity and appropriateness of different sentencing options (e.g., incarceration vs. treatment) are at issue. For example, a psychologist may testify that an individual is, and will remain, a serious threat to society due to a high likelihood of future violent behavior.

In civil cases, psychologists may be called on to help establish or refute claims of psychological injury and distress. Testimony in these cases may involve the presence or absence of injury or disorder, the cause of disorder, and the effects or damages involved with a condition, such as employment prospects, needed treatment and associated expenses, pain and suffering, and decreased involvement in, or enjoyment of, pleasurable activities.

Psychologists often serve as expert witnesses in cases where civil commitment or competency is at issue. Civil commitment refers to the involuntary hospitalization or restriction of an individual's liberty in order to prevent harm to self or others as a result of mental illness or incapacity. Competency hearings may involve such issues as the ability to manage one's own finances, create a will, or enter into binding agreements.

Within the family court system, psychologists may address the best interest of children in custody disputes, and participate in the investigation and disposition of allegations of child abuse or neglect. In custody disputes, it is not uncommon for judges to have the parties agree upon a single expert for a particular issue, such as the adequacy of parents' child-rearing capacities. For a detailed review of the various roles of psychologists in the court, see Melton et al. (1997). Monahan and Walker (1994) provide an overview of relevant case materials.

3. *The Expert in the Legal Process*

What occurs from beginning to end, when an expert becomes involved in a legal case, can vary widely, and hence a typical or common scenario is presented here. Initially, an expert is retained by an attorney and begins a review of the case. This may include an examination of various documents and reports, relevant scientific literature, and, in some cases (but not all), an examination of the individual or individuals whose psychological state is at issue. Based on this review, the psychologist, if possible, will then draw conclusions and present these in some manner (e.g., orally or in writing) to the attorney. In some situations, even if the attorney does not wish to maintain further involvement with the expert and does not want that expert to testify, these opinions must be disclosed to the other side; in other cases no disclosure is mandated. If the attorney does want to use the expert as a potential witness, that expert's opinions usually must be disclosed at some point before trial. The attorney usually must file some description, based on the expert's input, of the expert's qualifications, the materials reviewed and depended upon, and the expert's conclusions. The nature and level of detail required in the disclosure varies greatly in different jurisdictions.

Following disclosure, various steps may be taken by the other side to learn more about the expert and the expert's methodology and opinions. The opposing attorney may file interrogatories, or written questions to which the expert, via the retaining attorney, must provide sworn written answers. In many jurisdictions, attorneys may depose opposing experts. Discovery deposition involves giving sworn testimony under questioning by lawyers. During deposition, experts may be asked about such issues as their experience, training and education, credentials, the nature and bases for any conclusions they may draw as well questions about various professional, theoretical, and technical issues. Much or all of the questioning will be conducted by the opposing attorney although the retaining lawyer may also ask some questions, usually at the end of the process, either to clarify or 'fix' what might seem to be earlier problematic answers or to establish certain strong points to gain leverage for settlement. Deposition does not occur in front of the judge or jury, but the expert is under oath and a court reporter makes an official transcript.

The purpose of these discovery activities is to avoid 'trial by ambush' or surprise, by allowing attorneys the opportunity to understand and prepare for the testimony that the expert is anticipated to provide. In addition, more complete understanding of the strengths and weaknesses of each side of the case may well facilitate agreement between the parties and foster a pretrial settlement.

If the case gets to trial (which in fact is unusual because a high percentage of cases are settled before that time), the expert's court appearance begins with a swearing in. Before the expert states any opinions, the retaining lawyer will usually ask questions to establish the witness's credentials and expertise. At some point the attorney then asks the judge to admit or qualify the witness as an expert. Prior to the judge's ruling, the opposing attorney may then challenge the psychologist's qualifications. The lawyer may seek restrictions on the scope and nature of the expert's testimony or to have the expert excluded from testifying altogether.

What follows next (assuming that the expert has been qualified) is the direct examination. The retaining attorney will ask the expert questions designed to bring out such matters as the evidence reviewed, the bases for opinions, and the opinions derived. The opposing side then gets to conduct a cross examination, during which the opposing attorney may ask questions designed to cast doubt on the expert's credibility, methods, opinions, and profession. It is important to note that on direct examination, the retaining attorney may not ask leading questions (i.e., questions that imply or contain the answer), whereas such practice is allowed during cross-examination, which can allow for a more confrontational approach by the opposition attorney. Cross-examination may be followed by redirect, during which the retaining attorney attempts to 'rehabilitate' any problems caused during cross-examination by asking questions that may clarify or reinforce the expert's testimony and credibility.

4. *The Admissibility of Expert Testimony*

Any discussion of the admissibility of expert testimony, including that involving mental health experts, requires coverage of a recent landmark decision.

In the case of *Daubert, vs. Merrell Dow Pharmaceuticals, Inc.* (1993) the United States Supreme Court established a set of standards that now govern the admissibility of expert testimony in the Federal court system. Many states have also fully or partially adopted the Daubert standards. Prior to *Daubert*, two standards of admissibility governed the admissibility of expert testimony, *Frye* and the Federal Rules of Evidence (1974, 1984) which sometimes seemed to contradict one another and assigned different functions to the judge.

In *Frye vs. United States* (1923), the defendant in a criminal case sought to have an expert provide testimony that, based on the experimental use of blood pressure measurements, the defendant was truthful in his assertion of innocence. The court ruled the expert's testimony inadmissible because it was not based on techniques that were generally accepted within the relevant scientific community. The 'general acceptance' standard set down by *Frye* created certain problems because it combines both excessive conservatism in relation to novel but valid techniques with excessive acceptance of techniques that, while invalid or at least highly questionable or insufficiently tested, are endorsed by the majority (or at least a 'respected minority') of practitioners of a given field. When following *Frye*, the judge does not so much decide the admissibility of testimony based on its seeming scientific merits, but defers to the judgment of the scientific community by considering the degree to which the methodology underlying that testimony or conclusions is accepted. For example, if a review of the scientific literature seems to raise very serious doubts about a method or is generally negative, but the method is still embraced by a substantial proportion of practitioners, then the latter consideration rather than the former should be decisive under *Frye*.

It is easy to see how this standard can create tremendous problems when considering expert psychological testimony. There may be a wide gulf between the practices of psychologists and the scientific data on which they base (or fail to base) their methods and opinions. In many areas of interest to the courts, the scientific research may have yielded mixed results, or be lacking altogether, and yet, for a variety of reasons, practitioners may still act in a manner that belies that evidence or engage in activities in the absence of sound science. Even in the circumstance where a sound base of scientific knowledge does exist, it may not pertain directly to the specific questions at issue.

In the *Daubert* decision, the Supreme Court stated explicitly that the judges' role as gatekeeper under the Federal Rules of Evidence superceded the 'general acceptance' criteria specified in *Frye*, although it did not dismiss all elements of *Frye* entirely. Admissibility was to be determined by the judge's evaluation of the scientific merits of an expert's methodology. In addition, the text of the decision set forth a list of criteria upon which such decisions could be based, which are not exhaustive nor given in any hierarchical order or priority. These criteria included:

(a) General acceptance in the relevant scientific field (as per *Frye*).

(b) The degree to which the theoretical basis of the testimony is testable and falsifiable.

(c) Publication of peer-reviewed literature.

(d) The known error rate where diagnostic or predictive tasks are involved.

Daubert touched off a firestorm of debate in the legal and scientific communities, some of which focused on the overall effect that it would have on the quality of the expert advice the courts would allow into evidence. On the one hand, *Daubert* liberalizes the standards as an expert may present evidence that is not necessarily endorsed by the broader scientific community. This serves to widen the range of potential topics for consideration by the court, and recognizes that science is constantly evolving. On the other hand, by requiring judges to critically evaluate the scientific status of testimony, *Daubert* has the potential to function as a barrier to weak or junk science entering into the legal process. Within psychology, the response to *Daubert* has involved, to date, articles predicting the possible effects of the ruling, proscriptions, and recommendations for the psychologist working within the framework provided by *Daubert*, and debate about the degree different techniques, diagnoses, and procedures meet (or fail to meet) the *Daubert* standard.

There have been dozens of cases in which the admissibility of forensic psychology and social science evidence has been decided under the *Daubert* standard, including both civil and criminal cases. The types of evidence proffered to date have included (broadly) testimony on psychological syndromes; behavior profiles and diagnostic techniques related to sexual abuse/victimization; eyewitness identification; theory and experimental data related to civil rights and discrimination; the prediction of dangerousness and causal or mitigating factors in violent behavior; issues of mental state or condition; and 'truth-telling' techniques. In a substantial percentage of these cases, the testimony offered has been rejected either in whole or in part.

The bases for rejecting a psychologist's testimony in part or fully have included a lack of relevant scientific research, inconclusive or negative findings in the research literature, failure to adhere to well-established methods or procedures, and inadequate credentials. A *Daubert* challenge to an expert's status can be raised at different points during litigation depending on a series of factors. For example, a hearing may be held well in advance of trial or at the time of trial.

Although some areas of psychological research are more vulnerable to a *Daubert* challenge than others, the frequency with which testimony has been judged inadmissible is of concern to forensic psychologists, attorneys, and the parties who are involved in litigation where the expert testimony of psychologists is of

significance or may be absolutely critical to the case. Legitimate claims (and the defenses against illegitimate ones) based on psychological testimony are substantially, and perhaps fatally, undermined when psychological experts fail to practice in a manner consistent with good science and their testimony is excluded.

A problem that arises under *Daubert* is that an expert witness possesses knowledge or expertise that the judge often does not. Judges may lack the training that is required to fully evaluate the expert's field or scientific methodology in general. Thus, although it seems an important advance to require experts to back their opinions with more than the assertion that 'I am the expert, and therefore my opinion is valid,' the pragmatics for achieving the desired scrutiny still leave certain gaps. Further, because the list of considerations is not complete nor arranged in order of priority, the standards are clearest in the almost self-evident situation in which all scientific indicators are positive or negative. It is in the more ambiguous situation in which indicators are inconsistent and yet scientific status may still be strong (or weak) overall that more guidance is needed and decisions to date have sometimes been rather inconsistent. For further references and review of the effect of *Daubert* on the admissibility of psychological evidence, see Schuman and Sales (1999) and Wettstein et al. (1995).

5. *Recent Trends and Developments*

In recent years there has been significant professionalization in the field of psychology and law, which has included the formation of numerous professional organizations and the publication of various books, professional journals, ethical standards, training guidelines, and other scholarly materials related to the theory and practice of forensic behavioral science. Training programs and internships with a forensic component are increasingly available and a number of programs offering joint law and psychology degrees have been established. With this specialization has come a growing body of research that seeks to directly address legal issues (e.g., the prediction of violence or detection of malingering) and various aspects of human behavior within the courtroom (e.g., factors influencing jury decisions or the credibility of eyewitness testimony).

Although standards vary in different jurisdictions, there appears to be an increased acknowledgment of the gap between the level of certainty required by the court and the state of the art within many areas of psychology. Stimulated in part by calls for increased accountability (e.g., Ziskin 1970, Faust et al. 1991), and mandated by rulings such as *Daubert*, experts are now frequently required to demonstrate that there is a scientific basis for their methods, predictions, and conclusions. The pronouncement 'We know from clinical experience that …' is no longer accepted in many legal arenas as an adequate justification for expert status or for the particular aspects of expert testimony.

The rules governing the use of expert witnesses are subject to change at any time in response to new precedent or legislation, but given the above trends, it seems reasonable to hope that the above efforts will continue to strengthen the scientific bases that will allow psychologists to increasingly provide true assistance to the trier of fact, and aid in the fair and just resolution of legal cases.

See also: Courtroom Consultation: Psychology; Expert Memory, Psychology of; Expert Testimony; Forensic Psychiatry; Forensic Psychology; Science and Law; Science and Technology Studies: Experts and Expertise; Scientific Evidence: Legal Aspects; Statistics as Legal Evidence

Bibliography

Daubert vs. Merrell Dow Pharmaceuticals, Inc. 1993 113 S. Ct. 2786

Faust D, Ziskin J, Hiers J B Jr 1991 *Brain Damage Claims: Coping with Neuropsychological Evidence*. Law and Psychology Press, Los Angeles

Federal Rules of Evidence (1974, 1984) West, St. Paul, MN

Frye vs. United States 1923 293 F. Supp 1013 (D.C. Cir.)

Landsman S 1995 Of witches, madmen, and products liability: An historical survey of the use of expert testimony. *Behavioral Sciences & the Law* **13**: 131–58

Melton G B, Petrila J, Poythress N G, Slobogin C 1997 *Psychological Evaluations for the Courts: A Handbook for Mental Health Professionals and Lawyers*, 2nd edn. The Guilford Press, New York

Monahan J, Walker L 1994 *Social Science in Law: Cases and Materials*, 3rd edn. The Foundation Press, Westbury, NY

Schuman D W, Sales B D (eds.) 1999 Special theme: Daubert's meanings for the admissibility of behavioral and social science evidence. *Psychology, Public Policy and Law* **5**

Wettstein R M, Ewing C P, Tomkins A J (eds.) 1995 Behavioral science evidence in the wake of Daubert. *Behavioral Science and the Law* **13**

Ziskin J 1970 *Coping with Psychiatric and Psychological Testing*. Law and Psychology Press, Los Angeles

K. V. Heard and D. Faust

Expertise, Acquisition of

Acquisition of expertise is a significantly expanding area of interest in psychological and educational research. Understanding the nature of outstanding performance is of relevance in many contexts in daily and professional life. Scientific analyses of character-

istics of experts and theories about the acquisition of expertise thus are of great importance. In this entry, an overview is presented on main concepts, theories, and acquisition models in research on expertise.

1. Expertise: Core Concepts

1.1 Definitions

'Expertise' denotes the outstanding performance of an individual in a particular domain (e.g., medicine, physics, chess, music). 'Experts' thus are persons who, by objective standards and over time, consistently show superior performance in typical activities of a domain. Expert performance is often illuminated by comparison with individuals with limited performance. Such individuals are called 'novices.' In contrast to giftedness, expertise is usually considered as *acquired* special skill *in* or knowledge *of* a particular subject through practical experience.

1.2 Scientific Orientation

De Groot's (1965) seminal work on chessmasters pioneered the scientific orientation of research on expertise. The most striking difference between grandmasters and weaker players was revealed in a memory task, in which subjects were presented chess positions for a few seconds and asked to reconstruct them immediately. The experts' superior recall was explained with specific perceptual structures they held in memory which were closely related to their domain-specific knowledge.

De Groot's interpretation of these findings directed the future research on expertise from the perspective of information processing theory and cognitive psychology. The focus on the analysis of cognitive correlates of expert performance (perception, memory, knowledge, problem-solving) has since been maintained (Chi et al. 1988, Ericsson and Smith 1991) and has only recently been supplemented by research on the contexts in which expert performance is situated. Three main research topics can be derived from this orientation: (a) modeling of cognitive structures and mechanisms that characterize expert information processing; (b) description of acquisition of expertise; (c) study of instructional means to support acquisition of expertise.

1.3 Methods

Two kinds of empirical research methods are most prominent in studies on expertise: the contrastive approach and retrospective analyses.

In studies using the quasi-experimental contrastive approach (Voss et al. 1986), experts are compared with novices; sometimes subjects with an intermediate level of expertise ('semi-experts') are investigated, too. Maximizing the difference in level of expertise helps to point out the effects of expertise. Contrasting characteristics of the groups can be identified, which help to establish theories of expertise and to draw instructional implications for its acquisition. To a certain degree, the cross-sectional comparison can be interpreted as a developmental model 'from novice to expert.'

In retrospective analyses, two variants can be differentiated. The classical one is the biographical analysis of the development of distinguished experts. The aim to enlighten individual conditions for outstanding success is impaired by the danger of solely anecdotal reports. Ideally, different sources are investigated, or the method is combined with the contrastive approach. The same is true for the second variant, in which contrastive groups retrospectively report about their expertise career—in particular, about the contexts of pursuit with the domain. The main idea with using retrospective reports of experts is that they are the most valid and authentic data source. Evidence exists that retrospective verbal reports can complement cross-sectional methods sensibly.

2. Expertise: Research Results

2.1 Domain Specificity

Experts in virtually any complex domain have intensively practiced for 10 years or more (according to the 10-year rule of necessary preparation), and thus have extensive knowledge at their disposal. Not even the most 'talented' individuals can attain outstanding performance without such a period of preparation; most experts have even spent considerably longer.

A main consequence of the 10-year rule is that expertise is extremely domain-specific. Expertise cannot easily be transferred to another domain. For example, experts' memory is not in general superior to that of novices, though excellence in memorizing new information is the most striking characteristic of experts.

2.2 Expert Memory

As outlined, the central point of departure of research on expertise was evidence for experts' excellent memory for domain-specific information. In contrastive comparisons of memory performance, large effect sizes can be observed. These differences cannot be explained by general memory factors: no differences occur in control tasks such as digit-span. Thus, it is obvious that domain-specific practice determines the differences. The concept of 'pattern recognition' (Chase and Simon 1973) explains experts' ability to recognize

relevant patterns very quickly. It is based on the integration of perception and knowledge through practice. Components of successful pattern recognition are the mechanisms of 'chunking' and 'skilled memory.'

The mechanism proposed in the chunking concept is the integration of small knowledge units into larger chunks, which are labeled with indices that are subsequently used during recall instead of separate knowledge units. Simulations suggest that experts have some 50,000 chunks available. The mechanism proposed in the skilled memory concept relies on the availability of an explicit hierarchical organization of memory that allows new information to be quickly stored in and recalled from long-term memory.

2.3 Expert Knowledge

Experts' superior memory performance is closely related to their knowledge. Experts excel by both a great amount of knowledge and advantageous knowledge organization in order to make functional and efficient use of their knowledge.

The domain-specificity of expertise implies that the analysis of domain knowledge plays a major role in research on expertise. However, it is a fallacy to equate 'expert knowledge' and 'declarative domain knowledge.' Though experts have much domain knowledge available, they excel in flexibly using this knowledge, modifying it, or restraining from applying it within certain contextual circumstances. Recent educational studies showed that providing students with much declarative knowledge often leads to *inert knowledge* rather than to expert knowledge.

The relevance of knowledge for expert performance is so obvious that in studies on expertise only rarely is the concept of knowledge under investigation explicitly defined. However, many different types and qualities of knowledge can be differentiated, each of them with distinguished functionality (De Jong and Ferguson-Hessler 1996). Most important for the acquisition of expertise is the distinction between declarative knowledge (know-what) and proceduralized knowledge (know-how). Proceduralization of knowledge is one of the basic mechanisms for explaining the acquisition of expertise.

3. Acquisition of Expertise: General Assumptions

3.1 Proceduralization

Many expert actions are highly automated. With growing expertise, knowledge increasingly becomes proceduralized. In his model, Anderson (1982) stated that skill acquisition mainly consists of changing declarative knowledge through practice into proceduralized knowledge. This is achieved through knowledge compilation and rule tuning. Skill acquisition as a basic component of expertise can be described in three phases: declarative, compilation, and tuning. Students first acquire much declarative knowledge, which is later proceduralized and associated with action sequences. Then the skill is automatized and tuned through repeated practice. Anderson assumes innate cognitive structures for the acquisition of knowledge in the declarative phase, which are not elaborated in detail. However, it is plausible that the roles of dispositions and experience vary in different phases. This is influential for educational efforts to foster the acquisition of expertise.

3.2 Changing Roles of Dispositions and Experience

The changing roles of dispositions and experience are analyzed in the theory of ability determinants of skilled performance (Ackerman, 1986). Similar to Anderson's model, three phases are distinguished in this theory (cognitive, associative, and autonomous). In the cognitive phase, students are burdened with high demands: understanding task instructions, comprehending relevant goals, expressing strategies, and memorizing knowledge. For mastering these demands, general dispositional abilities play an important role, unless consistent task characteristics allow knowledge proceduralization. If tasks remain inconsistent, general dispositions remain important. During the associative phase, strategies get proceduralized, and compilation of knowledge leads to an increased speed of information processing. Skills are increasingly automatized during the autonomous phase, in which conscious attention for actions is no longer necessary. In ill-defined domains with inconsistent tasks, the autonomous phase is reached only rarely.

The roles of dispositions and experience thus depend on the nature of the domain. Given inconsistent demands, general abilities are important for performance; if demands are consistent, the importance of practice and experience increases. The analysis of individual differences in dispositions are important in order to understand the way in which individuals notice, accumulate, process, and apply information and in order to design learning environments adequately.

3.3 Development of Generic and Specific Expertise

Beyond the emergence of experience-based proceduralized knowledge, further elaborations of acquisition of expertise are possible. Hatano and Inagaki (1986) distinguished the acquisition of routine expertise and adaptive expertise. Routine expertise can be described through automatization of actions, adaptive expertise through increasing flexibility of actions. Adaptive expertise can emerge only when routine expertise is available. Similarly, Patel and Groen (1991) distinguish generic expertise and specific expertise. Generic expertise denotes the acquisition of

declarative knowledge, whereas specific expertise is characterized by action competence based on a sufficient amount of generic expertise. The development of specific expertise passes through several stages. Beginners have declarative knowledge available; intermediates have already compiled their knowledge into simple procedures; generic experts dispose of domain-specific schemata and scripts; and specific experts have enriched these with case experience.

4. Acquisition of Expertise: Models

4.1 Stage Models

Some models of acquisition of expertise roughly delineate stages of this process. The merit of stage models is that they foster an understanding of components that are important at different moments of the acquisition process. Stage models of acquisition of expertise have existed for almost a century, an early one being Cleveland's (1907) five-stage model in the domain of chess. Through the domain-specificity of this model, relevant cognitive processes during the acquisition of expertise are described rather precisely.

In sum, stage models give hints about typical developments but remain on a molar level. More detailed models are necessary if the acquisition of expertise is to be supported by instruction.

4.2 Deliberate Practice

More detailed approaches theorize about learning processes. Understanding these processes is important, because extensive practice within a domain is a necessary, but not sufficient, condition for acquisition of expertise. Using retrospective methods, it was found that today's experts differed from other individuals early in their career: they practiced more efficiently, had more committed teachers, and showed higher achievement demands. Ericsson et al. (1993) found that experts were more involved in effortful training activities over a long period of time that had the sole purpose of improving performance. Such activities are called 'deliberate practice.' Spontaneously, individuals only rarely engage in deliberate practice, although they recognize that it would improve their performance; they prefer regular activities that are motivated by inherent enjoyment (play) or external rewards (work). Therefore, it is important that teachers persist in deliberate practice, and they can foster the motivation for deliberate practice by offering explicit teaching goals, feedback, and opportunities for gradual improvement through repetition and correction of errors.

The most important reason for guidance by expert teachers is that in all complex domains over time a body of organized experience, in the form of knowledge and produced artifacts, has been accumulated. Through teachers this body is shared with students, because teachers can foresee future skill demands. They know with what method and to what degree of mastery simpler tasks have to be learned to serve as a solid foundation for more complex skills. In other words: students are supported in becoming fully enculturated in a community of expert practice without having to remake all the experiences earlier generations have already made.

4.3 Encapsulation of Knowledge in Cases

Changes of knowledge during practice are at the core of the encapsulation theory (Schmidt and Boshuizen 1993). This model, developed in the domain of medicine, postulates that acquisition of expertise leads to an integration of declarative and experiential knowledge in 'encapsulated knowledge.' Through professional activity and experience with real cases, declarative knowledge about diseases is developed into knowledge structures called 'illness scripts.' These are generalized over cases, but nevertheless are closely related to application contexts, because they are based on episodic experiences with real cases. The use of illness scripts leads to quick action without the need to activate declarative knowledge effortfully. The declarative knowledge remains available if necessary but can be neglected in most cases, because the knowledge encapsulated in clinical experience is sufficient.

Empirical evidence for the encapsulation model is based on findings that medical experts improve their performance if presented with additional case data about patients—such information is useless for novices, who are dependent upon abstract declarative knowledge. Additionally, experts' descriptions of clinical phenomena of diseases contain much more information about cases than do those of novices. Connecting declarative knowledge with case experiences is a crucial component of acquisition of expertise in medicine.

4.4 Experience of Relevant Episodes

The experience of episodes and their conscious reflection plays an important role for the acquisition of expertise. In the model of dynamic memory, Kolodner (1983) explains how episodes are represented in memory, and how episodic experiential knowledge can be applied. Knowledge is regarded as 'episodic definitions' that include the subjective relevance and perception of episodes, as well as knowledge about applicability and application errors. Since experts' episodic definitions are superior to those of novices, the acquisition of expertise can be interpreted as continuing refinement of episodic definitions based on experience from repeated application of knowledge.

Episodic definitions are represented in Episodic Memory Organization Packets (E-MOPs). An E-MOP can be described as a generalized episode; it contains

both information that is common across episodes, and deviations of particular episodes from this general information. Thus E-MOPs include applicable episodic knowledge in which the subjects' own experiences are integrated.

Two classes of events trigger learning from experience: generalization across episodes, and analysis of errors. After the occurrence of an error in a particular episode, the deviations of this episode from the generalized one are explicitly stored. Such knowledge about errors can later be used to avoid further errors. In educational theory, the explicit use of errors during learning has been discussed controversially. However, because errors inevitably occur in complex domains, dealing with them is very important, in order to enable subjects to cope with new errors.

5. *Acquisition of Expertise: Instructional Consequences*

5.1 *Case-based Learning*

Both the model of encapsulation and the theory of dynamic memory imply consequences for instructional support of the acquisition of expertise based on the implementation of case-based reasoning during learning.

As learning by experience with cases alters knowledge structures, instructional consequences are evident: the acquisition of expertise can be supported instructionally by fostering reflective application of knowledge through the presentation of complex learning environments in which real application situations occur. Case-based learning is a preferable mode of learning to reach these goals.

Case-based learning stresses the similarity between learning situation and application situation and implies a number of advantages: (a) by dealing with complex initial case problems, students get a notion of the relevance of the 'to-be-learned'; (b) cases are highly authentic and they are embedded in relevant situative contexts—thus they offer a possibility to make experiences in complex near-to-reality episodes of learning; and (c) multiple perspectives on the same subject matter help to avoid oversimplifications and to enhance the transferability of the to-be-learned.

5.2 *Cognitive Flexibility Through Multiple Perspectives*

The advantage of multiple perspectives is the focus of the cognitive flexibility theory (Spiro et al. 1989). Employing multiple perspectives aims at making knowledge more transferable. Students should deal with concepts at different times, in different contexts, with different purposes, and in different roles. Thus, oversimplification and too narrow ties to specific contexts are avoided.

Cognitive flexibility theory stresses the importance of multiple contexts in which the knowledge to be acquired is embedded. The theory is particularly relevant for research on expertise, because it mainly deals with advanced knowledge acquisition in ill-structured domains. Such domains can be described by: (a) complexity of concepts and cases; and (b) irregularity and large variability of cases. Instruction following the theory of cognitive flexibility aims to induce multiple and, as a consequence, flexible representations of the knowledge, which can be applied in many different contexts. An instructional means to induce this is the technique of landscape criss-crossing, in which the 'conceptual map' is explored in many different ways. As a consequence, many facets of the concepts are learned, so that they can be applied in a variety of contexts. Expert flexibility can thus be enhanced.

6. *Future Research Issues*

The outline of general assumptions and theoretical and instructional models about acquisition of expertise shows that much is already known in that area. Some restrictions were mentioned; to overcome them is a challenge for the future. It is easily predictable that the following issues will play a major role in forthcoming research on acquisition of expertise.

6.1 *Longitudinal Studies*

Research on acquisition of expertise lacks concomitant longitudinal studies. Factors other than cognitive structures and processes could thus be revealed. For example, reasons have to be identified *why* some individuals are motivated to work and deliberately practice for many years, whereas many others are not.

6.2 *Analysis of Ill-defined Domains*

For many years, research on expertise has focused on subjects' performance in solving well-defined problems that frequently can be worked out by routines. Only preliminary steps are taken concerning the flexibility of expert practitioners working in communities of practice in ill-defined domains. In such domains, a constructivistic epistemology might be more appropriate, in which, for example, the veridicality of knowledge is doubted.

6.3 *Generalizability Across Domains*

As a consequence of the domain-specificity of expertise, an important question has been widely neglected: the question about generalizability of empirical results across domains. Seeking commonalities between experts of different—yet not *too* different—

domains will be a demanding task within the innovative field of comparative research on expertise.

See also: Chess Expertise, Cognitive Psychology of; Culture-rooted Expertise: Psychological and Educational Aspects; Education: Skill Training; Knowledge (Explicit and Implicit): Philosophical Aspects; Medical Expertise, Cognitive Psychology of; Science and Technology Studies: Experts and Expertise; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Ackerman P L 1986 Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence* **10**: 101–39

Anderson J R 1982 Acquisition of cognitive skill. *Psychological Review* **89**: 369–406

Chase W G, Simon H A 1973 Perception in chess. *Cognitive Psychology* **4**: 55–81

Chi M T H, Glaser R, Farr M J (eds.) 1988 *The Nature of Expertise*. Erlbaum, Hillsdale, NJ

Cleveland A A 1907 The psychology of chess and learning to play it. *American Journal of Psychology* **18**: 269–308

De Groot A D 1965 *Thought and Choice and Chess*. Mouton, The Hague, The Netherlands

De Jong T, Ferguson-Hessler M G M 1996 Types and qualities of knowledge. *Educational Psychologist* **31**: 105–13

Ericsson K A, Krampe R T, Tesch-Römer C 1993 The role of deliberate practice in the acquisition of expert performance. *Psychological Review* **100**: 363–406

Ericsson K A, Smith J (eds.) 1991 *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press, Cambridge, UK

Hatano G, Inagaki K 1986 Two courses of expertise. In: Stevenson H W, Azuma H, Hakuta K (eds.) *Child Development and Education in Japan: A Series of Books in Psychology*. Freeman, New York, Chap. 17, pp. 262–72

Kolodner J L 1983 Towards an understanding of the role of experience in the evolution from novice to expert. *International Journal of Man–Machine Studies* **19**: 497–518

Patel V L, Groen G J 1991 The general and specific nature of medical expertise: A critical look. In: Ericsson K A, Smith J (eds.) *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press, Cambridge, UK, Chap. 4, pp. 93–125

Schmidt H G, Boshuizen H P A 1993 On acquiring expertise in medicine. *Educational Psychology Review* **5**: 205–21

Spiro R J, Feltovich P J, Coulson R L, Anderson D K 1989 Multiple analogies for complex concepts: Antidotes for analogy-induced misconception in advanced knowledge acquisition. In: Vosniadou S, Ortony A (eds.) *Similarity and Analogical Reasoning*. Cambridge University Press, Cambridge, UK, Chap. 18, pp. 498–531

Voss J F, Fincher-Kiefer R H, Green T R, Post T A 1986 Individual differences in performance: The contrastive approach to knowledge. In: Sternberg R J (ed.) *Advances in the Psychology of Human Intelligence*. Erlbaum, Hillsdale, NJ, Vol. 3, Chap. 8, pp. 297–334

H. Gruber

Explanation, Cognitive Psychology of

An explanation is a judgment about why an event occurred. Explanations are drawn for numerous and varied events (e.g., the behavior of others, the weather, the movement of planets, natural disasters). Much of the psychological research and theory on explanation comes from work on Attribution Theory, which is the study of how people make judgments about others and themselves. Despite its name, attribution theory is not a single theory, but a large collection of minitheories and investigations. This article summarizes some of the major perspectives, theories, and research topics within this large body of work on the psychological study of explanation.

The study of explanation is an integral and well-researched topic in psychology, and understandably so. Explanations are ubiquitous in everyday life (e.g., Hsiang-Ling did well on the exam because she has wonderful quantitative skills), they influence affect (e.g., I'm angry at Joe because he intentionally broke my favorite vase), and they determine social behaviors (e.g., I helped Margaret because I thought she was really trying her best to succeed). It should be noted, however, that there are important differences between the lay person's goals in generating explanations and the scientist's goals in studying explanations. Lay people typically desire explanations that are understandable, informative, accurate, consistent with beliefs and expectations, and personally satisfying. In contrast, scientists are interested in the psychological processes that determine what kinds of explanations people generate, as well as the affective, cognitive, and behavioral consequences of drawing different kinds of explanations.

1. Heider, Founder of the Psychological Study of Explanation

Interest in explanation processes and consequences can be traced at least as far back as the fourth century BC, with Aristotle's book on The Rhetoric. However, if a particular individual had to be singled out as the founder of the scientific study of explanation, it would surely be Fritz Heider. Heider's insights (Heider 1944, 1958) laid the foundation for much of the research and theory in the study of explanation. Consider three of his important ideas. First, Heider suggested that people want to understand the stable causes of a given behavior so that they can predict future behavior. He wrote, 'It is an important principle of common-sense psychology ... that man grasps reality, and can predict and control it, by referring transient and variable behavior and events to relatively unchanging underlying conditions, the so-called dispositional properties of his world' (1958, p. 79). Thus, just as astronomers may seek to explain the orbit of planets

so that they can predict their future positions in space, people seek to explain the actions of individuals so that they can predict the future actions of those individuals. Second, Heider emphasized the distinction between personal and situational causes. This is a fundamental distinction in the study of explanation, and it seems to capture a great deal of common sense reasoning as well. For example, if an athlete performs poorly, one might decide that something about the athlete is the cause or one might decide that something about the field conditions is responsible. Third, Heider suggested that people tend to focus on personal causality and underestimate the influence of the situation. Heider (1944) wrote: 'Often the momentary situation which, at least in part, determines the behavior of a person is disregarded and the behavior is taken as a manifestation of personal characteristics' (p. 361). This tendency is what later came to be called the correspondence bias (see below). Heider's insights provided the impetus for numerous theories and investigations. Two of the most well known theories are Kelley's Covariation Model and Jones and Davis's Correspondent Inference Theory.

2. *Intuitive Scientists and Logical Explanations*

The theories of Kelley, and Jones and Davis were formulated during a time when many psychologists viewed people as logical, rational, 'intuitive scientists.' Thus, it is not surprising that both theories emphasize logical principles that a sophisticated person could use to make rational judgments. This logical perspective has been very influential in the study of explanation, although, as described below, it is not the only perspective that theorists have found useful.

2.1 *Kelley's Covariation Model*

Suppose an acquaintance of yours, Torsten, says to you, 'You have to try Kjelleren, that new restaurant on Storgata. It's great!' This recommendation can be explained in several ways. Something about the restaurant may be the cause, such as its fine food or pleasant atmosphere. Something about Torsten may be the cause. Perhaps he is easily impressed and raves about every restaurant he tries.

Kelley (1967) suggests there are three types of information that you should consider. First, you should consider consensus information. Consensus pertains to how other people react with regard to the target object. If others also rave about Kjelleren, consensus is high. If they dislike Kjelleren, consensus is low. Second, you should consider distinctiveness information. If Torsten recommends every restaurant he samples, his recommendation on this occasion is not very distinctive (distinctiveness is low). But if he rarely recommends restaurants, his comment is distinctive (distinctiveness is high). Finally, you should consider consistency information. If Torsten's reaction has been positive every time he has been to Kjelleren, consistency is high. But if he disliked Kjelleren previously, consistency is low.

Different combinations of these three types of information suggest different conclusions. For example, suppose others do not seem to like Kjelleren (low consensus), Torsten often recommends restaurants (low distinctiveness), and Torsten often recommends Kjelleren (high consistency). This suggests that something about Torsten is the cause of his recommendation. In contrast, suppose others also like Kjelleren (high consensus), Torsten rarely recommends restaurants (high distinctiveness), but Torsten consistently recommends Kjelleren (high consistency). Now it seems that something about Kjelleren is the cause.

Of course, sometimes people lack the information, ability, or motivation to perform such a sophisticated analysis. In this case, one may use the discounting principle, which states that 'the role of a given cause in producing a given effect is discounted if other plausible causes are present' (Kelley 1972, p. 8). For example, suppose one sees a person who appears anxious. If it is known that an anxiety-provoking event is present, one should discount the person's personality as a cause of the anxiety displayed on this occasion.

2.2 *Jones and Davis's Correspondent Inference Theory*

Suppose the following three behaviors are observed. First, Elena is leaving a funeral and she appears sad. Second, Jason declines a promotion in the company where he works. Third, at a party Edward is arguing vehemently that it is every person's responsibility to conserve energy, recycle, and protect endangered wildlife. What should one conclude about the personalities of these individuals? Jones and Davis's Correspondent Inference Theory (Jones and Davis 1965, Jones 1990) suggests that two pieces of information are important to consider. First, one should consider the expectedness or normativeness of the behavior. When people behave in ways that are normative for the situation (e.g., sad Elena leaving the funeral), their behavior is relatively undiagnostic of their personalities. Second, one should consider the clarity of the behavior with regard to its implications (there are numerous reasons why Jason might have declined the promotion). When many explanations are possible, one cannot be certain which is the correct one.

However, when behavior is atypical and has clear implications, a personality inference can be drawn. In the case of Edward, although one could be mistaken, a reasonable inference is that his behavior reflects an underlying disposition, specifically, that he is an environmentalist. This is a correspondent trait inference. In Jones's words, 'a correspondent inference is

a straightforward extrapolation from the behavior observed: the behavior is seen as corresponding to or reflecting an underlying disposition of the actor ... for example, to call a person hostile after observing a hostile act would be to draw a correspondent inference' (1990, pp. 46–7).

3. *Cognitive Misers and Top-of-the-head Explanations*

As mentioned above, the theories of Kelley and Jones and Davis emphasize logical principles that a sophisticated person could use to make rational judgments. Research suggests that people can use such logical principles, and they may do so when they have the time and motivation (perhaps during a jury deliberation). However, people may not typically use such logical principles when they formulate explanations. Some theorists suggest that although people can be 'intuitive scientists' who use logical rules, they can also be 'cognitive misers' who do not expend many cognitive resources unless accuracy is at a premium. Research from this perspective reveals that people often draw relatively unsophisticated 'top-of-the-head' judgments that are heavily influenced by whatever causal force is the most obvious. For example, when an actor interacts with others who are moving or brightly illuminated (McArthur and Post 1977), the actor's behavior is attributed more to situational factors.

4. *Story Understanders and Theory-driven Explanations*

Yet a third perspective is that explanations can be viewed as part of the process of understanding information in that explanations often rely on background information and common sense reasoning (see Schemas, frames and scripts, in cognitive psychology). For example, if one learns that Charles, who is health conscious, asked to be moved to a different table in a restaurant, one might surmise that he wanted to sit farther from people who were smoking. On the other hand, if one learns that Diederik, who was sitting in a coffee shop reading a new novel, asked to be moved to the table farthest from the crowd, one might speculate that he wanted to find a quiet place to read. In everyday life we often use information drawn from prior knowledge and previous experiences, and this information influences explanations as well (e.g., Read 1987). Research from this perspective tends to be more specific in that it focuses on particular knowledge structures (e.g., the restaurant script) rather than content-independent rules that can be applied to any behavior. However, it has an advantage in that the explanations that are produced are more natural ('Diederik wanted a quiet place to read his book' rather than 'Something about Diederik is the cause'). Of course these perspectives need not be mutually exclusive, and some theories make use of more than one (e.g., Hilton and Slugoski 1986).

5. *Bias in Explanation*

Given the emphasis on logical principles in the early work on explanation, it is perhaps not surprising that deviations from completely rational explanations have captured the interest of explanation theorists (see *Decision Biases, Cognitive Psychology of*). Two of the most well-known biases are the correspondence bias and the self-serving bias.

5.1 *The Correspondence Bias*

The correspondence bias, also known as the fundamental attribution error, (see Gilbert and Malone 1995, Jones 1990 for reviews) refers to peoples' tendency to infer that personality corresponds to behavior. For example, if Anne behaves in a friendly manner, we may infer that she has a friendly personality, and if Neil behaves in an aggressive manner, we may infer that he has an aggressive personality. Of course, often personality does correspond to behavior, and so the correspondence bias need not be an error. However, research suggests that even when there is good reason to think that a behavior is substantially situationally driven, people still tend to infer that behavior reflects personality. For example, even though we know that actors and actresses are only playing roles, we may still infer that their personalities correspond to the behavior of the characters that they play.

5.2 *The Self-serving Bias*

The work described thus far has addressed explanations for the behavior of other people. However, considerable research has also been conducted on explanations about oneself, particularly on how people explain their own successes and failures. One of the notable findings of this research is that many people are self-serving in their explanations in that they tend to take more credit for their successes than their failures (see Bradley 1978 for a review). For example, if Robert runs well in a 400 m race, he might indicate that his high ability is the cause. But if he runs poorly, he might blame his performance on the field conditions. Interestingly, this self-serving tendency extends to explanations about the successes and failures of in-group members (e.g., Hewstone and Ward 1985). So, for example, Robert might also assign his team-mates relatively more credit for their successes and relatively less blame for their failures. However, as described below, this self-serving tendency is not found in all populations.

6. *Process Models of Explanation*

Much of the research on explanation processes has emerged from the Jones and Davis tradition. Just as their Correspondent Inference Theory sought to describe how people draw inferences about personality, so most process models seek to describe the mechanism by which people use information about behavior and situational forces when drawing inferences about personality. Several models suggest that dispositional inferences are drawn through a series of stages that differ in their degree of automaticity (see). For example, Gilbert et al. (1988) suggest that people first categorize behavior (Susan appears anxious), next they characterize personality (Susan has an anxious personality), and then they correct for the influence of the situation (perhaps some anxiety-provoking situation is making Susan anxious; she may not be an anxious person after all). Gilbert suggests that the third stage requires more effort than the first two, so when cognitive resources are scarce, people may fail to correct their characterizations completely and so display correspondence bias. Numerous investigations provide support for this view. Although sequential models of explanation have been the norm, a parallel distributed processing approach (see *Cognition, Distributed*) may become more common in the future.

7. *Cultural Differences*

Although the majority of research on explanation has been conducted with North American or European participants, interest in cross-cultural research has been waxing, and as researchers in different cultures find it easier to collaborate, productivity in cross-cultural research will probably continue to increase. One cross-cultural finding is that people in individualist cultures (e.g., US, Great Britain) prefer dispositional explanations for behavior, but people in collectivist cultures (e.g., China, India) often do not, and sometimes collectivists prefer situational explanations (e.g., Miller 1984). It should be noted, however, that research of this type has been conducted in relatively few nations; more research is needed. Less research has been conducted on the correspondence bias, but the extant work suggests that this bias may be cross-cultural, although more robust in individualist cultures (e.g., Choi and Nisbett 1998). There are also cultural differences in the degree to which explanations are self-serving or group serving (e.g., Hewstone and Ward 1985), but these differences do not seem to fall as easily along individualist/collectivist lines.

8. *Explanation Consequences*

The study of the consequences of explanation is itself a substantial area of study. Research suggests that both emotional states and achievement motivation are influenced by explanations for success and failure (e.g., Anderson et al. 1996). For example, if Veronica receives a failing grade on a literature paper and concludes that her failure was due to a lack of effort, she may be motivated to work harder next time. On the other hand, if she concludes that she has no talent for writing, she may feel depressed and helpless. Indeed, the study of such explanations is a major topic in the clinical study of depression (see, Abramson et al. 1989). Moreover, not only are internal and uncontrollable explanations for failure likely to lead to depression, but people who are depressed are also more likely to draw such conclusions than the self-serving explanations characteristic of nondepressed individuals.

See also: Causation: Physical, Mental, and Social; Comprehension, Cognitive Psychology of; Explanation-based Learning, Cognitive Psychology of; Functional Explanation: Philosophical Aspects; Interpretation and Translation: Philosophical Aspects; Rational Choice Explanation: Philosophical Aspects; Verstehen und Erklren, Philosophy of

Bibliography

Abramson L Y, Metalsky G I, Alloy L B 1989 Hopelessness depression: A theory-based subtype of depression. *Psychological Review* **96**: 358–72

Anderson C A, Krull D S, Weiner B 1996 Explanations: Processes and consequences. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology: Handbook of Basic Principles*. Guilford Press, New York

Bradley G W 1978 Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology* **36**: 56–71

Choi I, Nisbett R E 1998 Situational salience and cultural differences in the correspondence bias and actor–observer bias. *Personality and Social Psychology Bulletin* **24**: 949–60

Gilbert D T, Malone P S 1995 The correspondence bias. *Psychological Bulletin* **117**: 21–38

Gilbert D T, Pelham B W, Krull D S 1988 Inference and interaction: When person perceivers meets persons perceived. *Journal of Personality and Social Psychology* **54**: 733–40

Heider F 1944 Social perception and phenomenal causality. *Psychological Review* **51**: 358–74

Heider F 1958 *The Psychology of Interpersonal Relations*. Wiley, New York

Hewstone M, Ward C 1985 Ethnocentrism and causal attribution in Southeast Asia. *Journal of Personality and Social Psychology* **48**: 614–23

Hilton D J, Slugoski B R 1986 Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review* **93**: 75–88

Jones E E 1990 *Interpersonal Perception*. Macmillan, New York

Jones E E, Davis K E 1965 From acts to dispositions: The attribution process in person perception. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 2

Kelley H H 1967 Attribution theory in social psychology. In: Levine D (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, Vol. 15

Kelley H H 1972 Attribution in social interaction. In: Jones E E, Kanouse D E, Kelley H H, Nisbett R E, Valins S, Weiner B (eds.) *Attribution: Perceiving the Causes of Behavior*. General Learning Press, Morristown, NJ

McArthur L Z, Post D L 1977 Figural emphasis and person perception. *Journal of Experimental Social Psychology* **13**: 520–35

Miller J G 1984 Culture and the development of everyday social explanation. *Journal of Personality and Social Psychology* **46**: 961–78

Read S J 1987 Constructing causal scenarios: A knowledge structure approach to causal reasoning. *Journal of Personality and Social Psychology* **52**: 288–302

D. S. Krull and C. S. Anderson

Explanation: Conceptions in the Social Sciences

Explanation is an important goal of any scientific inquiry. Many social scientists would say that it is the most important one. Nevertheless, the expression 'explanation' as it is used within the social sciences is still a vague and ambigous term with at least five different connotations: (E1) the explanation of an event or a phenomenon, (E2) the explanation of a scientific law or theory, (E3) the explanation of the functioning of a system, (E4) the explanation of the meaning of an entity (e.g., an expression, an idea, a behavior), and (E5) the explanation of an action in the sense of its (moral) justification. In each case, an explanation is a statement or account which makes what is to be explained clearer than it was before and promotes its understanding. But how this is to be achieved for the different variants of the concept of explanation and what are the characteristics of a good or adequate explanation in each case is a controversial topic not only in the social sciences but in science in general (cf. Stegmüller 1983). The focus of discussion has been, from the beginning, on the first variant E1, i.e., the explanation of events or phenomena. This variant and the problems involved will be the main subject of this article.

1. *Explanation-seeking Questions*

Explanations can be understood as answers to certain questions, the so-called *explanation seeking questions*. In the social sciences, a very important type of explanation seeking questions refers to actions of a certain kind performed by a certain person. An explanation of individual actions is not only of interest in itself, but it is also required for a proper understanding of many *macrosocial* phenomena that are construed as the aggregate result of individual actions of large numbers of persons. As Little (1991) puts it, all social processes, causal influences, systemic interactions, etc., are ultimately embodied in the actions of individual actors within a specific social and natural environment. This may be one reason why most publications about explanation in the social sciences primarily, or even exclusively, deal with the explanation of human actions.

Suppose a person p performed an action of kind a. There are at least the following eight questions relevant to the explanation of this singular action a of person p (cf. Tuomela 1977):

(Q1) Why did p do a?

(Q2) What was the reason of p for doing a?

(Q3) What was the motive (emotion, feeling, sensation, personality characteristic) which led p to do a?

(Q4) Which of the wants (desires, decisions, duties, obligations, norms) of p brought about his or her doing a?

(Q5) Which of the beliefs (thoughts, considerations, knowledge) of p brought about his or her doing a?

(Q6) What was the purpose (intention, aim, end, goal) of p for doing a?

(Q7) How is it possible that p did a?

(Q8) What does p's doing a mean (express)?

Answers to these questions are not independent from each other. Q1 could be answered by referring to causes and reasons for action a, whereas Q2 seems to accept only the latter. Q4 and Q5 which already mention certain kinds of reasons in the formulation of the question are included in Q2. Thus, an answer to Q4 and Q5 can be transformed into an answer to Q2. An answer to Q3 can be interpreted as an answer to Q1 or Q2. Tuomela assumes that answers to questions Q1–Q5 are statable as:

(A1) p did a because of c.

Here c gives the causes or reasons for p's doing a. All answers to Q6 can be paraphrased in the form

(A2) p did a in order to do g.

Here g is p's goal action for the purpose of which a was performed. An illustration of A2 would be: p saved money in order to enjoy life after retirement.

Tuomela claims that answers of the type A2 can be further reduced to answers of the type A1 and that the expression 'because of' in A1 signifies a causal relation between a and c. Answers to *how possible questions* of type Q7 are incomplete variants of A1. Explanations of individual actions by means of sociological variables such as minority status, social class, or social position of an agent are good examples of answers to questions of type Q7.

Answers to questions of type Q8 are of a totally different nature. They belong to category E4, i.e., the explanation of the meaning of an entity, in that they reveal an additional understanding of action a. The so-called *speech acts*, that are linguistic actions, may

illustrate this type: if action *a* of person *p* consists of the utterance 'I will do that!,' an appropriate answer to the question 'What does *p*'s saying "I will do that!" express?' might be 'This utterance of *p* is a promise, not only an announcement or a prediction.'

Answers of type A1 seem to be fundamental in the explanation of human actions. Differentiating between types of psychological explanations allows a more comprehensive treatment of component *c* of A1.

2. Types of Psychological Explanations

Bunge and Ardila (1987) suggest the following classification of psychological explanations: (T1) tautological explanations, (T2) teleological explanations, (T3) mentalist explanations, (T4) metaphorical explanations, (T5) genetic explanations, (T6) developmental explanations, (T7) environmental explanations, (T8) evolutionary explanations, (T9) physiological explanations, and (T10) mixed explanations. Each type shall be briefly explained (in the sense of E4) and illustrated by two simple examples.

T1: *Tautological explanations* refer to basic capabilities or mental faculties: person *p* is able to anticipate (or anticipates) events because of *p*'s forethought capability. Person *p* is able to think about (or thinks about) herself because of *p*'s self-reflective capability.

T2: *Teleological explanations* refer to goals or purposes (cf. component *g* in A2): person *p* sold his shares in order to maximize his profit. Person *p* studied medicine in order to become a doctor.

T3: *Mentalist explanations* refer to mental events: person *p* developed a neurosis because *p* suffered from an intrapsychic conflict. Person *p* works so hard because *p* is highly motivated.

T4: *Metaphorical explanations* refer to analogies with physical or social processes or with animals or machines: person *p* processes information like a parallel processing computer does. Person *p*'s drivings follow the same principles as a steam boiler.

T5: *Genetic explanations* refer to the genetic endowment of a person: person *p* has a high intelligence because her or his parents also had a high intelligence. Person *p* suffers from Alzheimer's disease because *p* carries two apolipoproteine E type 4 alleles.

T6: *Developmental explanations* refer to stages or levels of biological or psychological development or to events in the persons's past: person *p* does not understand the lever laws because *p* is not yet in the stage of formal operations. Person *p* suffers from agoraphobia because *p* has experienced a severe trauma in early childhood.

T7: *Environmental explanations* refer to external conditions and factors: person *p* shows a high response rate because *p* is working under a variable ratio reinforcement schedule. The agoraphobic symptoms of *p* are weakened because *p* is massively exposed to the threatening stimuli.

T8: *Evolutionary explanations* refer to the survival value of a behavior or behavior tendency, its selective advantages or disadvantages: male persons show sexual jealousy in order to ensure paternity in putative offspring by preventing encroachment by competing males, and this behavior tendency has evolved because of the selective advantage of a greater reproduction rate of these persons (cf. Buss 1997). Persons have a high pain threshold under duress because of its survival value.

T9: *Physiological explanations* refer to physiological, especially neurophysiological and endocrinological, processes and mechanisms: introverts are more arousable than extraverts because introverts have a lower threshold for activation in the ascending reticular activating system than extraverts. The depressive person *p* experienced an elevation of her or his mood because *p* took a cyclic antidepressant which increased the chemical neurotransmitter serotonin.

T10: *Mixed explanations* are combinations of two or more of the above mentioned types of explanation. Mostly, explanations are not pure cases of one type, but combinations of at least two types, i.e., they are mixed explanations. Cases of one type usually provide only a partial answer to the respective explanation seeking question. In answers of the type A1 to explanation seeking questions and in most of the given examples of the different types of psychological explanations, the relation term 'because' occurs. It links that which is to be explained to that which explains. A characterization of this relation is provided by the so-called *models of explanation*.

3. Models of Explanation

Explanation is one of the main subjects of the philosophy of science. So far, many different models of explanation have been constructed, most of them incompatible with one another (cf. Salmon 1989). Each model attempts to answer at least two questions: (a) What is (the structure of) an explanation? and (b) What is a good (proper, appropriate, adequate) explanation? Some answers to these questions will be briefly outlined.

3.1 The Deductive-nomological Model of Explanation

According to this classical view, an explanation is an *argument* which shows that the phenomenon to be explained can be inferred from certain other facts by means of specified general laws. This type of argument may be schematized as a deductive inference of the following form (cf. Hempel 1965):

(D-NE)	$L_1, L_2, \ldots, L_r$	General laws
	$C_1, C_2, \ldots, C_k$	Statements of antecedent conditions
	————	
	E	Description of the empirical phenomenon to be explained

E as the description of the phenomenon to be explained is called the *explanandum* (*sentence*). The statements of antecedent conditions, which make assertions about particular facts, and the general laws together form the *explanans*, i.e., that which explains. Explanations of this kind are called explanations by deductive subsumption under general laws, or *deductive-nomological explanations*. (The root of the term 'nomological' is the Greek word 'nomos' for law.)

D-NE gives an answer to the question 'What is (the structure of) an explanation?' A comparison between this structure and the one underlying the examples of the different types of psychological explanations reveals that most of the latter are stated in an elliptical form, viz., 'E because of c' (cf. A1). The component c refers only to a subset of the antecedent conditions $C_1, C_2, \ldots, C_k$ that are required for a proper explanation of the explanandum E, and the general laws, if there are any at all, are totally omitted. Thus, explanations of the form 'E because of c' do not count as proper explanations.

To give an answer to the second question, i.e., 'What is a good explanation?,' requires the formulation of *conditions of adequacy*. In the case of the model of deductive-nomological explanation, there are four such conditions (cf. Hempel 1965):

(R1) The explanandum must be a logical consequence of the explanans, i.e., the explanandum must be logically deducible from the information contained in the explanans.

(R2) The explanans must contain general assumptions or hypotheses, and these must actually be required for the derivation of the explanandum.

(R3) The explanans must have empirical content, i.e., it must be capable, at least in principle, of test by experiment or observation.

(R4) The sentences constituting the explanans must be well confirmed.

It is not easy to find examples of deductive-nomological explanations in the social and behavioral sciences that display the proper structure and satisfy the conditions R1–R4. One example may be: the phenomenon to be explained is Peter's frequent aggressive behavior in the presence of his mother. The explanation is given with reference to the theory of operant behavior. The explanandum statement E can be formulated as 'Peter shows a high frequency of aggressive behavior if his mother is present.' The principle of positive reinforcement 'If a positive reinforcer is presented contingent on the occurence of a behavior, then the frequency of this behavior will increase' can be the law statement L. At least two statements of antecedent conditions are required, viz., A_1 'Contingent on Peter's aggressive behavior, his mother regularly shows attentive behavior,' and A_2 'His mother's attentive behavior is a positive reinforcer for Peter.'

In this case, the explanandum E seems to be deducible from the explanans, i.e., the conjunction of L, A_1, and A_2. Thus, R1 as the first condition of adequacy seems to be satisfied. But what about the conditions R2–R4? That depends on the status of L. If L would have empirical content and could be regarded as well confirmed, R2–R4 would be satisfied too. With regard to L, two problems arise: (a) According to its critics, L is no empirical law, but only a logical consequence of the definition of the concept of a positive reinforcer (the much discussed problem of circularity of the empirical 'law of effect'); (b) even if this were not the case, L is not a well confirmed universal law because its validity depends upon conditions which are not explicitly mentioned in L. One important condition is the following: (C1) there exists no other behavior in the respective situation contingent on the occurrence of which a more effective positive reinforcer is presented. If there exists only one such other behavior in the case of Peter and his mother—and this is not excluded by the antecedent conditions A_1 and A_2—a quite different law statement would become relevant, viz., the 'law' of parallel reinforcement. If there exist even two or more of the behaviors excluded by C1 which seems to be the usual case in unrestricted natural environments, it is difficult to find an appropriate law statement at all.

This example is representative of the problems that conflict with any attempt to apply the model of deductive-nomological explanation to the social sciences. Universal deterministic laws as they are required by the model are hard to find in this domain. Nevertheless, Hempel (1965) who is the best known proponent of the model defended its significance for the social sciences and even for history. He considered any deviation from the standards set by the model as leading to insufficient and incomplete explanations. Nowadays, few social scientists are willing to share his view. But the idea of a *nomological* social science with strict laws and explanations of the deductive-nomological type is still alive (cf. McIntyre 1996).

The trust in the possibility of a change in the direction of a nomological social science is based on a certain feature of scientific explanations, viz., that any explanation of an event is an explanation of that event *under a certain description*. The explanandum statement E in a deductive-nomological explanation does not represent the event to be explained as such, but the event under a certain description. It might be that the

current situation in the social sciences is the consequence of an unfavorable choice of modes of description—that, at least, is the argument. Other choices which would lead to a redescription of the subject matter of the social sciences could, perhaps, reopen the way to a truly nomological social science. Unfortunately, McIntyre (1996) who is a confirmed supporter of this view has not produced any laws, offered any new social theories, or even given any redescriptions so far.

3.2 The Statistical-relevance Model of Probabilistic Explanation

It may be difficult to find genuine nomological laws in the social sciences, but it is easy to find statements which describe *statistical regularities* between events. '85 percent of the persons who have a panic disorder and who undergo an exposure therapy experience relief from their symptoms' is an example. Concrete percentages or probabilities do not occur in every statistical statement. Instead, fuzzy terms like 'nearly all,' 'most,' 'the majority,' 'more than half,' etc., are used. Hempel (1965) had recognized this situation very early and broadened the scope of his theory of explanation by proposing his much discussed *model of inductive-statistical explanation*.

Referring to the above example of a statistical law, an explanandum E could be 'Person p experiences relief from his or her symptoms of panic.' Appropriate antecent conditions would be A_1 'Person p has a panic disorder' and A_2 'Person p undergoes an exposure therapy.' The statistical law in conjunction with these two antecedent conditions constitutes the explanans in this case. The inductive-statistical explanation of E could be expressed in two ways: (a) the inductive probability of the explanandum given the explanans amounts to 0.85. (b) The explanans inductively supports the explanandum in the degree 0.85 (cf. Schurz 1995/96).

The objections against Hempel's model of inductive-statistical explanation led Salmon (cf. 1989) to the formulation of his *statistical-relevance model of probabilistic explanation*. Whereas in Hempel's model an explanation of an event is an inductive argument that confers upon the event to be explained a high inductive probability, an explanation of an event according to Salmon's model is an assemblage of factors that are statistically relevant to the occurrence of the event to be explained, accompanied by an associated probability distribution.

Salmon and Salmon (1979) gave the following example to clarify this notion: Albert, an American teenager, is convicted of stealing a car. A sociologist who attempts to explain that act of delinquency will presumably look for certain factors that are known to be relevant to delinquent behavior, e.g., sex, age, marital status of parents, religious background, socio-economic status, residence (rural/urban/suburban), etc. Each of those factors, as well as many others, is assumed to be statistically relevant to delinquency in the sense that the probability of delinquency is different for boys than for girls, different for urban dwellers than for rural youngsters, different for children whose parents have remained married than for those from broken homes, etc. If any of these factors does not make some difference in the probability of delinquency, it will be discarded as irrelevant, and hence as lacking in explanatory value.

To explain Albert's delinquent behavior, the population of American teenagers to which Albert belongs is partitioned into a series of subclasses in terms of such relevant factors as those already mentioned. For each subclass in the resulting partition, the probability for delinquent behavior on the part of members of that subclass will be determined. Then, Albert will be assigned to one such subclass in terms of his actual attributes, e.g., 17-year-old male, suburban dweller, middle-class undivorced parents, etc., and the probability will be cited for that subclass. If the number of relevant factors which enter into the partition is large, the probability obtained in a given case may be quite low. For Salmon, that is not crucial. What matters is that the relevant factors be taken into account in partitioning, and that the probabilities associated with the various subclasses be accurate.

This model is confronted with two problems. (a) The requirement of a homogeneous relevant partition of the reference class (in the example: the population of American teenagers) is not easy to meet. In most cases in which explanations of single events are looked for in the social sciences, a comparable probability distribution will not be available. (b) The fact that an event counts as appropriately explained even if the probability value associated with the respective subclass of the partition may be very low comes into conflict with deeply rooted preconceptions about proper explanations. With regard to Tuomela's explanation seeking questions, the best that could be achieved by this kind of explanatory efforts is an answer to a question of type Q7, i.e., an answer to a how possible question.

These problems of the statistical-relevance model of explanation led others to challenge its claim to be a model for the explanation of single events. Schurz (1995/96) comes to the conclusion that the genuine explananda of statistical laws are not single events but sample frequencies. In his discussion of the explanatory value of causal modeling techniques in the social sciences, Woodward (1989) assumes that the statistical information in the delinquency example may contribute to the explanation of facts about population level parameters but not to the explanation of why any particular boy became a delinquent.

Even if this critique of Salmon's model were to be accepted, it would still be of value to the social sciences, since it provides a standard for a deeper statistical-

Table 1
The canonical form for causal explanations of specific events (cf. Humphreys 1989, pp. 286 f.)

Request for explanation:
What is the explanation of Y in S at t?
Appropriate explanation:
Y in S at t [occurs, was present] because of Φ despite Ψ

Notes
'Y' is a term referring to a property or change in property.
'S' is a term referring to a system.
't' is a term referring to a time.
'Φ' is a (nonempty) list of terms referring to contributing causes of Y.
'Ψ' is a (possibly empty) list of terms referring to counteracting causes of Y.

causal understanding of social phenomena as, for example, juvenile delinquency, unemployment, or terrorism in certain reference classes. This standard could also be applied to the increasing use of meta-analytic methods for explanatory purposes in the social sciences when several explanatory variables are involved at the same time (cf. Cook et al. 1994).

3.3 The Model of Aleatory Explanations

A more recent model of causal explanations of specific events is the model of aleatory explanation (see Table 1) introduced by Humphreys (cf. 1989). He agrees with Salmon that causal explanations are possible within the realm of chancy, or aleatory, phenomena. But in his model, in contrast to Salmon's, no probability value is assigned to an explanation. The demands on the knowledge base are much lower than in Salmon's model, e.g., no complex probability distribution on the basis of a homogeneous relevant partition of a reference class has to be available.

What is required for a proper explanation of a specific event Y in S at t are two lists of causes of Y, i.e., a list of *contributing* and a list of *counteracting causes*. In Humphreys' model, for something to be a cause it must invariantly produce its effect. But causes in this model are probabilistic causes, and they produce changes in the value of the *chance* of the effect. A contributing cause of Y produces an increase, a counteracting cause of Y a decrease in the value of the chance of Y.

This model seems to be applicable in various contexts within the social sciences (cf. Henderson 1993). It takes account of the fact that social phenomena are usually the result of multiple causal influences. It does not presuppose the existence of complete lists of all influences which, positively or negatively, affect a given outcome. Aleatory explanations are conjunctive. They can be improved by including additional probabilistic causes which may come up in further research.

The model has already been applied to explanations within a theory of social interaction in small groups (cf. Westmeyer 1996). But its importance goes well beyond that particular context. Take, for example, the interactionistic theory-frame of Magnusson (1980) under which many theories of social interaction can be subsumed. One fundamental assumption of this theory is that changes in the probabilities for behaviors of a person in a certain situation are a function of changes in the factors within this situation. Such determining within-situation factors are the physical situation, behaviors of the person concerned, and behaviors of the interaction partners.

In terms of Humphreys' model, 'Y' would refer to a change in the probability of a behavior of the person concerned; 'S' would refer to the setting in which the intrasituational dynamic interaction between the person and the interaction partners takes place; 't' would refer to a certain point or period in time at which the change in the probability of the behavior occurs; 'Φ' would refer to the set of all those determining within-situation factors, the changes of which facilitate (contribute to) the occurrence of Y; and 'Ψ' would refer to the set of all those determining within-situation factors, the changes of which inhibit (act counter to) the occurrence of Y. Contributing and counteracting causes could be identified by methods of event sequence or time series analysis.

3.4 The Embedding Theory of Explanation

Most of the models and notions of explanation dealt with so far are strongly connected to the so-called *statement view* of scientific theories that assumes theories to be sets of laws or theoretical assumptions. A very different picture is given by the *structuralist view* which is a variant of the model-theoretic approach to the understanding of scientific theories. Here, theories are reconstructed as complex set-theoretical structures. The structuralist program has already demonstrated its fruitfulness in the analysis of various theories from the social sciences (cf. Stegmüller et al. 1982, Westmeyer 1992).

From this point of view, explaining a phenomenon means to explain it *within the context of a certain theory*. This requires to show (a) that the phenomenon can be construed as an element of the set of intended applications of the theory, and (b) that the theory can be successfully applied to that element. In short, what has to be done is to embed the phenomenon into the theory. This leads Bartelborth (1996) who gives the most detailed account of this idea to speak of the *embedding theory of explanation*. For a concrete example of this type of explanation in the social sciences see Westmeyer (1996). He shows that it is possible to combine the embedding of a phenomenon into a certain theory with the application of Humphreys' model of aleatory explanation.

The embedding theory is the most promising variant of the *unification approach* (cf. Kitcher 1989) to the

problem of explanation. 'Explaining as unifying' is a notion which increasingly attracts the attention of philosophers of science. It paves the way for an integrative treatment of explanation and understanding (cf. Schurz and Lambert 1994) which is especially of interest to those social scientists who are used to construe explaining and understanding as basically different scientific activities. It is in accordance with a coherence theory of truth and fully compatible with basic epistemological tenets of the more recent constructionist and constructivist approaches to the social sciences. These characteristic features of the embedding theory give rise to the expectation that this theory will play a leading role in future discussions of the issue of explanation in the social sciences.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Cognitive Modeling: Research Logic in Cognitive Science; Constructivism/Constructionism: Methodology; Event Sequence Analysis: Methods; Explanation, Cognitive Psychology of; Functional Explanation: Philosophical Aspects; Historical Explanation, Theories of: Philosophical Aspects; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Rational Choice Explanation: Philosophical Aspects; Theory: Conceptions in the Social Sciences

Bibliography

Bartelborth T 1996 *Begründungsstrategien* [*Strategies of Justification*]. Akademie Verlag, Berlin

Bunge M, Ardila R 1987 *Philosophy of Psychology*. Springer-Verlag, New York

Buss D M 1997 Evolutionary foundations of personality. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA

Cook T D, Cooper H, Cordray D S, Hartmann H, Hedges L V, Light R J, Louis T A, Mosteller F 1994 *Meta-analysis for Explanation: A Casebook*. Russell Sage Foundation, New York

Hempel C G 1965 *Aspects of Scientific Explanation*. Free Press, Albany, New York

Henderson D K 1993 *Interpretation and Explanation in the Human Sciences*. State University of New York Press, New York

Humphreys P W 1989 Scientific explanation: The causes, some of the causes, and nothing but the causes. In: Kitcher P, Salmon W C (eds.) *Scientific Explanation*. University of Minnesota Press, Minneapolis, MN

Kitcher P 1989 Explanatory unification and the causal structure of the world. In: Kitcher P, Salmon W C (eds.) *Scientific Explanation*. University of Minnesota Press, Minneapolis, MN

Kitcher P, Salmon W C (eds.) 1989 *Scientific Explanation*. University of Minnesota Press, Minneapolis, MN

Little D 1991 *Varieties of Social Explanation*. Westview Press, Boulder, CO

Magnusson D 1980 Personality in an interactional paradigm of research. *Zeitschrift für Differentielle und Diagnostische Psychologie* **1**: 17–34

McIntyre L C 1996 *Laws and Explanation in the Social Sciences*. Westview Press, Boulder, CO

Salmon W C 1989 Four decades of scientific explanation. In: Kitcher P, Salmon W C (eds.) *Scientific Explanation*. University of Minnesota Press, Minneapolis, MN

Salmon W C 1998 *Causality and Explanation*. Oxford University Press, New York

Salmon M H, Salmon W C 1979 Alternative models of scientific explanation. *American Anthropologist* **81**: 61–74

Schurz G 1995/96 Scientific explanation: A critical survey. *Foundations of Science* **3**: 429–65

Schurz G, Lambert K 1994 Outline of a theory of scientific understanding. *Synthese* **101**: 65–120

Stegmüller W 1983 *Erklärung–Begründung–Kausalität* [*Explanation–Justification–Causality*], 2nd edn. Springer-Verlag, Berlin

Stegmüller W, Balzer W, Spohn W (eds.) 1982 *Philosophy of Economics*. Springer-Verlag, New York

Tuomela R 1977 *Human Action and its Explanation*. D. Reidel, Boston, MA

Westmeyer H (ed.) 1992 *The Structuralist Program in Psychology: Foundations and Applications*. Hogrefe & Huber Publishers, Seattle, WA

Westmeyer H 1996 A concept of explanation for social interaction models. In: Hegselmann R, Mueller U, Troitzsch K G (eds.) *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*. Kluwer, Dordrecht

Woodward J 1989 The causal mechanical model of explanation. In: Kitcher P, Salmon W C (eds.) *Scientific Explanation*. University of Minnesota Press, Minneapolis, MN

H. Westmeyer

Explanation-based Learning, Cognitive Psychology of

In explanation-based learning a domain or a task is explicated or demonstrated to a learner. For problem solving, this can take the form of worked-out problems. Worked-out problems can be used for 'analogy-based' problem solving of new problems, or can form the basis for 'self-explanations': student generated 'annotations' of the problem that relate the problem to theoretical principles, make abstractions from the given problem, or have a metacognitive function.

1. Introduction

In learning, a distinction can be made between acquiring conceptual (declarative) knowledge (principles, systems, and processes from a domain) and operational (procedural) knowledge (domain-specific tasks, such as solving problems). Both types of knowledge can be acquired in learning that has a more discovery or experiential character or by explanation-based learning (see e.g., Glaser 1966, Zhu and Simon 1987).

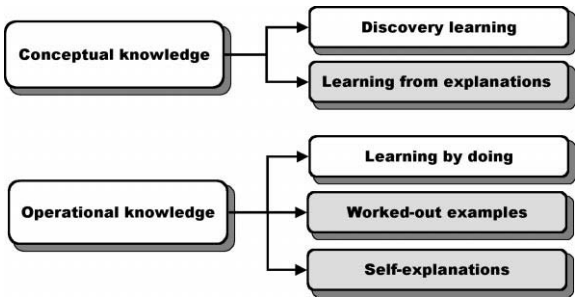

Figure 1
Overview of types of learning

In *discovery or experiential learning*, theory or skills are not introduced directly by the teacher or the learning material but have to be induced from the environment by learners themselves. For conceptual knowledge, this can take for example the form of scientific discovery learning with simulations (e.g., de Jong and van Joolingen 1998). For operational knowledge, experiential learning is embodied in 'learning by doing.' After having acquired the basic conceptual knowledge, learners start to solve problems by themselves (see e.g., Anderson et al. 1981).

Explanation-based learning comes very close to our traditional, classroom-related idea of teaching and learning and involves a teacher (or a learning material) who explains or shows the learner what to know or how to act. For conceptual knowledge, explanations lay out the basic elements and relations between these elements (concepts and relations) from the domain and may contain descriptions of new concepts in terms of known principles, validation of new concepts by logical derivations, presentation of examples and nonexamples, explicitation of causality, etc. These explanations take the form of text or visuals of many kinds (e.g., graphs, diagrams, animations). For operational knowledge, 'explanations' can take the form of step-by-step instantiated procedures. For problem solving these are called worked-out examples. These worked-out examples can be extended by 'annotations' that further explain to learners the related theoretical principles or generalizable problem-solving procedures and strategies. Learners can be provided with these annotations by the teacher or the learning material, but they can also generate explanations themselves, in which case they are called 'self-explanations.'

Figure 1 summarizes the different types of learning. The explanation-based types of learning are indicated in gray in the figure. 'Learning from explanations' refers to learning conceptual knowledge from a teacher or from instructional material. The literature on learning (processes) in this area can be found in studies on studying text, hypertext, and other types of 'expository' (multimedia) learning material. In this article, there is an emphasis on learning operational knowledge from worked-out examples and from self-explanations.

2. *Learning from Worked-out Examples*

Traditionally, textbooks (on e.g., physics) are built up in such a way that first theory (conceptual knowledge) is explained, then example problems are presented, and subsequently students have to solve problems themselves. Anderson et al. (1981) studied students solving geometry proof problems and found that students made heavy use of examples. In solving a novel problem, they try to find similarities and relate characteristics of the new problem to the characteristics of the old one. Pirolli and Anderson (1985) studied students learning recursion in programming. They report that, in an initial stage, 18 of 19 students writing programs containing recursion used a previous example, and students spent 34.3 percent of their time examining this example. In this 'learning by analogy,' students first modify the solutions of examples given and then create 'new problem-solving operators which can apply to future problems' (p. 242). Pirolli and Anderson (1985) further point out that students only rely on analogical problem solving in novel or difficult situations, such as when learners encounter an 'impasse' in the problem-solving process. When they are more experienced they have built a knowledge base that contains 'compiled knowledge,' which is knowledge that is dedicated to solving specific kind of problems and that may lead to fast and errorless problem solving.

A theory that supports the idea that worked-out examples should be used in initial stages of learning is 'cognitive load theory' (Sweller and Cooper 1985). This theory takes as a starting point that proficient problem solving requires *schemas* and *rules*. Schemas are cognitive constructs that help to recognize problem types and trigger the adequate rules from memory. Learning to solve problems therefore involves the processes of 'schema acquisition' and 'rule automation' (Ward and Sweller 1990). If learners, in an initial stage, have not yet acquired schemas they have to rely on general (weak) problem-solving procedures such as 'means–ends analysis' to find rules. Means–ends analysis is a general problem-solving procedure in which comparisons between the current problem state and the goal state (solution) guide the search for operators that help to bring the current state to the goal state. These weak problem-solving procedures require large cognitive resources because problem solvers have to 'simultaneously consider and make decisions about the current problem state, the goal state, differences between states, and problem solving operators that can be used to reduce such differences' (Ward and Sweller 1990, p. 3). With worked-out problems learners can concentrate on problem states and their associated moves, thus reducing cognitive

load and preventing potential cognitive 'overload.' In several studies, Sweller and associates found beneficial effects of learning with worked-out examples compared to open problems.

3. *Learning from Self-explanations*

Anderson et al. (1981) pointed to the fact that learning from worked-out problems and using analogies in solving new problems only works on short-term notice because students do not remember examples that well and that the matching of the two problems is very 'syntactic' and thus superficial in nature. In a similar vein, Zhu and Simon (1987) found that students solving quadratic algebraic expressions (e.g., x^2+ax+b) sometimes matched a new problem on one condition with the example (e.g., that the sum of the numerical factor should equal the coefficient of the linear term of the quadratic) but then failed to look at another necessary condition (e.g., that the product of the factors should equal the constant term) (see Zhu and Simon 1987, p. 145). These findings suggest that analogy-based problem solving may easily lead to using inadequate problem-solving procedures. The learning process in using worked-out examples may thus have a more reproductive character, leading to more superficial knowledge without the necessary abstractions needed for transfer. A solution can be found by providing students with annotated examples in which annotations link the examples to theoretical principles and transferable problem-solving knowledge. These explanations of the problem-solving process can also be generated by students themselves ('self-explanations'). According to Chi and vanLehn (1991) 'A self-explanation is a comment about an example statement that contains domain relevant information over and above what was stated in the example line itself' (p. 69). Theories of 'self-explanation' try to combine the best of both worlds. Together with providing basic information or part-explanations, students are required to reformulate information themselves. This stimulates the integration of the new information and prior knowledge and excites generalization. At the same time, too high a cognitive load is avoided.

Of the studies listed by Chi and vanLehn (1991), several found that students learn successfully from examples while other studies indicate that learning from examples does not lead to the ability to solve new problems. Chi and vanLehn (1991) reason that in the successful studies students applied self-explanations. Several studies indeed found this relation between success and self-explanations. Chi et al. (1989) and Ferguson-Hessler and de Jong (1990) found that good problem solvers generate more self-explanations while studying a text and example problems than poor problem solvers. Ferguson-Hessler and de Jong (1990) found that good problem solvers tended to explain principles and problem-solving steps to themselves, whereas poor problem solvers took the information for granted.

It is, however, not only the amount of self-explanations but also the quality of them that matters. Chi and vanLehn (1991) state that there are two ways to generate self-explanations. One is a more deductive way in which connections are made with theoretical concepts and principles (e.g., from the text). Bielaczyc et al. (1995) mention that in self-explanations students link features of the examples to concepts that were introduced in the preceding text. The second type of self-explanation that Chi and vanLehn identified follows an inductive route and concerns making abstractions or generalizations from a specific example given. Chi and vanLehn (1991) found, for example, that self-explanations generated by students while studying worked-out physics problems mainly concerned justifying the action in the problem's solution, by extending the conditions of the action and following up the consequences of the action.

Similarly, Bielaczyc et al. (1995) mention identifying the processes and rationale underlying actions in the example (LISP-programming) problem as a self-explanation strategy. This may explain differences found in the use of examples by good and poor problem solvers in the Chi et al. (1989) study. They found that good problem solvers used the examples as prompts or references (so using abstractions) whereas the poor problem solvers used them as analogies in solving new problems (which indicates more superficial comparisons). In addition to these two types of self-explanations, students also use explanations to 'plan, monitor, and evaluate their comprehension and strategy use' (Bielaczyc et al. 1995, p. 249). Also, for this type of self-explanation Ferguson-Hessler and de Jong (1990) found differences between good and poor problem solvers. Good problem solvers had meta-knowledge and were able to assess when they understood something; poor problem solvers, to the contrary, generally overestimated their understanding of the domain.

4. *Instructional Consequences*

Studies cited above have shown advantages to learning with worked-out problems but also clearly suggest that favorable effects will only be reached when students process these worked-out examples in a 'deep' way by generating self-explanations. A way to foster this is to train or stimulate students to generate self-explanations. Bielaczyc et al. (1995) found that students who received an explicit training in generating self-explanations indeed improved on this and gained higher performance than a control group that did not receive this instruction. The training that Bielaczyc et al. gave was very domain-specific, but Chi et al. (1994) also found beneficial results from providing students with quite general hints to generate explanations for themselves.

See also: Cognitive Theory: ACT; Cognitive Theory: SOAR; Deductive Reasoning Systems; Discovery Learning, Cognitive Psychology of; Explanation, Cognitive Psychology of; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning; Reasoning with Mental Models; Scientific Reasoning and Discovery, Cognitive Psychology of; Verstehen und Erklren, Philosophy of

Bibliography

Anderson J R, Greeno J G, Kline P J, Neeves D M 1981 Acquisition of problem-solving skills. In: Anderson J R (ed.) *Cognitive Skills and their Acquisition: Carnegie Symposium on Cognition (16th: 1980: Carnegie-Melton University*. Erlbaum, Hillsdale, NJ, pp. 191–230

Bielaczyc K, Pirolli P L, Brown A L 1995 Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction* **13**: 221–52

Chi M T H, Bassok M, Lewis M W, Reimann P, Glaser R 1989 Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science* **13**: 145–82

Chi M T H, van Lehn K A 1991 The content of physics self-explanations. *The Journal of the Learning Sciences* **1**: 69–105

Chi M T H, de Leeuw N, Chiu M-H, Lavancher C 1994 Eliciting self-explanations improves understanding. *Cognitive Science* **18**: 439–77

de Jong T, van Joolingen W R 1998 Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research* **68**: 179–202

Ferguson-Hessler M G M, de Jong T 1990 Studying physics text; differences in study processes between good and poor performers. *Cognition and Instruction* **7**: 41–54

Glaser R 1966 Variables in discovery learning. In: Shulma L S, Keislar E R (eds.) *Learning by Discovery. A Critical Appraisal*. Rand McNally, Chicago, pp. 13–26

Pirolli P L, Anderson J R 1985 The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology* **39**: 240–72

Sweller J, Cooper G A 1985 The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction* **2**: 59–89

Ward M, Sweller J 1990 Structuring effective worked examples. *Cognition and Instruction* **7**: 1–39

Zhu X, Simon H A 1987 Learning mathematics from examples and by doing. *Cognition and Instruction* **4**: 137–66

T. de Jong

Explanatory Style and Health

Explanatory style refers to how an individual habitually explains the causes of bad events. This cognitive personality variable is described as optimistic when bad events are explained with pervasive and long-lasting causes that involve the self ('I am an unworthy person'). A pessimistic explanatory style is one in which bad events are explained with circumscribed and transient causes that do not involve the self ('It was just one of those things'). The notion of explanatory style emerged from the attributional reformulation of learned helplessness theory, where it is proposed as a distant influence on the specific expectation that one's behaviors are related (or not) to important outcomes. An optimistic explanatory style leads individuals to believe that their behaviors affect future outcomes, whereas a pessimistic explanatory style encourages individuals to believe that they are helpless, that nothing they do matters. The expectation of helplessness versus efficacy is the theoretically critical link between explanatory style and outcomes. Explanatory style has most frequently been investigated in relationship to depression, where an extensive research literature shows a pessimistic style to be a consistent correlate of depression and, when coupled with negative events, to be a risk factor for subsequent depression (see *Depression*). But explanatory style is related as well to other outcomes involving maladaptive passivity, including physical illness. Research establishing an association between a pessimistic explanatory style and poor health converges with both laboratory experiments and survey studies linking stress and negative emotions to illness. More generally, investigations of explanatory style and poor health are one example of a growing interest in mind–body interactions and their impact on well-being.

1. Measurement of Explanatory Style

Explanatory style took off as a line of research when measures of this individual difference began to be developed. The first measure of explanatory style to be developed was the Explanatory Style Questionnaire (ASQ; Peterson et al. 1982). The ASQ presents respondents with hypothetical good and bad events involving themselves (e.g., 'you go out on a date, and it goes badly'). They are asked to imagine the event happening to themselves, and then to write down the event's 'one major cause' if it happened to them. Then they use rating scales to indicate the degree to which the cause falls along dimensions reflecting pessimism versus optimism.

A second way of measuring explanatory style is a flexible content analysis method: the Content Analysis of Verbatim Explanations (CAVE; Peterson et al. 1992). This strategy stemmed from the observation that causal explanations, identical to those given by respondents to the ASQ, were abundant in spontaneous writing or speaking. Was it possible to identify these explanations in verbatim material, extract them, and then to rate them along the dimensions reflecting pessimism versus optimism? Indeed, this could all be done, and explanatory style so assessed proves reliable, consistent, and valid. Potential research subjects neither able nor willing to participate in typical

research can be studied with the CAVE technique, as long as they have left behind suitable material containing causal explanations about themselves. To date, the CAVE technique has been used to measure explanatory style from psychotherapy transcripts, interviews, open-ended questionnaires, political speeches, sports stories, religious texts, and even song lyrics. The methodological virtue of this approach to studying explanatory style is that it allows longitudinal research to be conducted retrospectively. If suitable verbal material can be located from early in the lives of individuals whose long-term fate is known, then the CAVE technique allows studies to be done very quickly that nonetheless span decades. For obvious reasons, studies of explanatory style and physical well-being have frequently relied on the CAVE.

2. *Explanatory Style and Health*

Approximately one dozen studies have established a correlation between explanatory style (measured with the ASQ or the CAVE) and physical health. Taken together, the results of these studies converge with other investigations of positive thinking and well-being. Optimism, rendered as explanatory style or in other ways, predicts good health operationalized in a number of ways, from self-reports to physician ratings of general well-being to doctor visits to survival time following a heart attack to immunological efficiency to longevity.

How strong is the association between optimism and good health? Most of the relevant studies report correlation coefficients in the 0.20–0.30 range, which are moderate in size and typical of correlations in psychological research. Research participants have included males and females, some initially healthy and others initially quite ill. Adults across the life span have been included. Many of these studies have been longitudinal, spanning weeks to decades. At least some of these longitudinal studies have statistically controlled for initial levels of health and potential confounds involving tendencies to complain.

Optimistic explanatory style apparently impacts health at a number of junctures. It can make the initial onset of illness less likely; it can minimize the severity of illness; it can speed recovery; it can make relapse less likely. Most of the relevant studies, by virtue of rather stark correlational designs, do not allow conclusions about when health is impacted. Future investigations of a more fine-grained nature are needed.

3. *Questions about the Link between Explanatory Style and Health*

Can a lifetime of good health be encouraged by the early cultivation of optimism? Researchers at the University of Pennsylvania have begun an intervention program that involves teaching grade-school children to be more optimistic. Results to date suggest that optimism training makes subsequent episodes of depression less likely. Perhaps studies of physical well-being can be included in this ongoing investigation.

We can similarly ask if optimistic ways of thinking acquired once an individual falls ill will boost good health. Again, studies of optimism and depression suggest that this question is worth pursuing. Cognitive therapy that targets negative ways of thinking alleviates depression and prevents its recurrence. By implication, cognitive therapy for the seriously ill might pay health dividends.

Certainly, the most basic question that arises from research into optimism and health involves the mechanisms linking the two. The link appears to be overdetermined, and there are several plausible routes.

For example, there may be an *immunological pathway*. Researchers have shown that an optimistic explanatory style is positively correlated with the vigor with which the immune system responds to an antigen challenge. Several other similar lines of investigation have looked at how explanatory style is linked to the progression of AIDS. Results here are mixed, but there are hints that optimistic explanatory style predicts survival time and that this effect is mediated in part by immunological factors (see *Psychoneuroimmunology*).

There may also be an *emotional pathway* between explanatory style and health. As noted earlier, an extensive research literature has shown optimistic explanatory style to be incompatible with depression, and other studies have linked depression to poor health and early death. At least part of this latter path may be immunological.

There are probably several *cognitive pathways* between optimism and health as well. Optimistic explanatory style is not an isolated belief but rather part of a complex knowledge system that can influence physical well-being in numerous ways. For example, individuals with an optimistic explanatory style see the world as less filled with hassles than do their pessimistic counterparts; this tendency is in turn linked to better health. For another example, optimistic individuals are more likely to see health outcomes as something over which they can exert control; again, this tendency is linked to better health.

Another explanation of why optimistic thinking is related to physical well-being points to a *social pathway*. People with a pessimistic explanatory style are often socially isolated, and social isolation is a consistent predictor of poor health (see *Social Support and Health*). In general terms, the individual's social context can set the stage for the optimism–health link. Epidemiologist Leonard Sagan (1987) put forth the intriguing argument that the dramatic increase in life expectancy in the Western world over the centuries was due not to breakthroughs in medical or public health practices but instead to cultural diffusion of the

originally radical notion that an individual is a discrete self able to have an effect on the world. Once the idea of individual agency was invented, legitimized, and disseminated, the findings discussed here became possible.

Although the link between optimism and health is complexly determined, the most typical and most robust mechanism is a mundane *behavioral pathway*. An optimistic explanatory style is associated with a variety of 'healthy' practices: exercising, drinking in moderation, avoiding fatty foods, and the like (see *Health Behaviors*. People with an optimistic explanatory style are more likely than those with a pessimistic explanatory style to respond to colds with appropriate actions: staying home, taking it easy, and increasing the intake of fluids (see *Health: Self-regulation*).

A recent study of optimistic explanatory style and physical well-being looked at more than 1,000 individuals over almost 50 years (Peterson et al. 1998). Pessimistic individuals had an increased likelihood of an early death, and the large sample size made it further possible to investigate associations between explanatory style and death due to different causes. The interesting finding was that pessimistic individuals were most likely to die accidental and/or violent deaths. This effect was particularly pronounced for men.

Deaths like these are not random. 'Being in the wrong place at the wrong time' may be the result of an incautious and fatalistic life-style entwined not only with pessimism but also with the male gender role. This study did not reveal what the deceased research participants were doing when they died accidentally or violently, there is good reason to suspect that their behavior was somehow implicated, if only by affecting the settings they habitually entered or not.

Follow-up investigations suggest that the link between explanatory style and traumatic accidents involves a preference for potentially risky settings or activities (or at least a tendency not to avoid these) as well as the experience of negative emotions while in these settings. Negative emotions can make accidents more likely by producing passivity, limiting decision making, distracting individuals from dangers, and so on.

A final unresolved issue is whether the health benefits of explanatory style are due to optimism in its own right or simply to the absence of pessimism. Some writers have referred to the latter possibility as 'the power of non-negative thinking.' The issue cannot be resolved so long as research juxtaposes optimistic and pessimistic explanatory styles, treating them as mutually exclusive.

Regardless, explanatory style is linked to good health and long life, but it is worth emphasizing that styles of thinking are just one ingredient in physical well-being. Genetic predispositions, lack of health care, traumatic events, and environmental hazards such as toxins and germs take a toll on health regardless of how people might think. Researchers as well as the general public need to think about the determinants of physical health in a sophisticated manner. Psychological factors such as explanatory style are part of the equation but so too are a myriad of other factors.

See also: Control Behavior: Psychological Perspectives; Coping across the Lifespan; Depression, Hopelessness, Optimism, and Health; Self-efficacy; Self-efficacy and Health

Bibliography

Abramson L Y, Seligman M E, Teasdale J D 1978 Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology* **87**: 49–74

Peterson C, Bossio L M 1991 *Health and Optimism*. Free Press, New York

Peterson C, Maier S F, Seligman M E 1993 *Learned Helplessness: A Theory for the Age of Personal Control*. Oxford University Press, New York

Peterson C, Schulman P, Castellon C, Seligman M E 1992 CAVE: Content analysis of verbatim explanations. In: Smith C P (ed.) *Motivation and Personality: Handbook of Thematic Content Analysis*. Cambridge University Press, New York, pp. 383–92

Peterson C, Seligman M E, Vaillant G E 1988 Pessimistic explanatory style is a risk factor for physical illness: A thirty-five year longitudinal study. *Journal of Personality and Social Psychology* **55**: 23–7

Peterson C, Seligman M E, Yurko K H, Martin L R, Friedman H S 1998 Catastrophizing and untimely death. *Psychological Science* **9**: 127–30

Peterson C, Semmel A, Von Baeyer C, Abramson L Y, Metalsky G, Seligman M E 1982 The explanatory style questionnaire. *Cognitive Therapy and Research* **6**: 287–99

Reivich K 1995 The measurement of explanatory style. In: Buchanan G M, Seligman M E (eds.) *Explanatory Style*. Erlbaum, Hillsdale, NJ, pp. 21–47

Sagan L A 1987 *The Health of Nations: True Causes of Sickness and Well-being*. Basic Books, New York

C. Peterson

Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)

In statistical regression, one wants to predict a response variable from values of explanatory variables. For example, one might try to predict personal income as a function of age, education, location, gender, and other relevant variables. As the number of explanatory variables grows, the prediction problem becomes more difficult—this is known as the 'Curse

of Dimensionality' (COD). Traditional regression methods, such as multiple linear regression, did not encounter this problem because they made the strong model assumption that the mathematical form of the relationship between the response and explanatory variables was linear. But modern statisticians prefer to use nonparametric methods, in which fewer model assumptions are made, since this allows flexible fits of data in complex applications. Therefore, statisticians recently have invented a number of computer-intensive statistical techniques that attempt to achieve both flexibility and some degree of insensitivity to the COD. This article describes the main ideas in that endeavor.

1. The Curse of Dimensionality

The Curse of Dimensionality (COD) was coined by the mathematician Richard Bellman in the context of approximation theory. For statisticians, the term has a slightly different connotation, and refers to the difficulty of finding predictive functions for the response variable Y using explanatory variables $\boldsymbol{X} \in \mathbb{R}^p$ when p is large.

There are three nearly equivalent formulations of the COD, each of which offers a useful perspective on the problem:

(a) When p is large, nearly all datasets are sparse.

(b) The number of possible regression functions increases faster than exponentially in the dimension p of the space of explanatory variables.

(c) When p is large, nearly all datasets show multicollinearity (and its nonparametric generalization, concurvity).

These problems can be reduced if data are collected using a careful statistical design, such as Latin hypercube sampling. However, except in simulation experiments, this is difficult to do.

The first formulation is the most common. It refers to the fact that if one takes five random points in the unit interval, they tend to be close neighbors, while five points in the unit square and the unit cube are increasingly separated. Thus as dimension increases, a fixed sample size provides less local information. One way to quantify this is to ask what is the side-length of a p-dimensional cube that is expected to contain half of the data that are randomly (uniformly) distributed in the unit cube? The answer is $(.5)^{1/p}$, and this increases towards 1 as p gets large. Thus, the expected number of observations in a fixed region of $\mathbb{R}^p$ goes to zero, implying little information in the data about the local relationship between Y and $\boldsymbol{X}$. Therefore, one requires very large sample sizes before nonparametric regression is feasible in higher dimensions.

The second formulation of the COD has an information-theoretic quality. There are many possible models one can fit when p is large, and it is difficult for a finite sample to successfully discriminate among those alternatives in selecting and fitting the best model. To see this, suppose one considers only polynomial models of degree 2 or less. For $p = 1$ there is a single explanatory variable, and the seven possible models are:

$$
\begin{aligned}
Y &= \beta_0 + \varepsilon \\
Y &= \beta_0 + \beta_1 X_1 + \varepsilon \\
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon \\
Y &= \beta_1 X_1 + \varepsilon \\
Y &= \beta_0 + \beta_2 X_1^2 + \varepsilon \\
Y &= \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon \\
Y &= \beta_0 + \beta_2 X_1^2 + \varepsilon
\end{aligned}
$$

where ε denotes random noise in the measurement of Y. When $p = 2$, there are 63 possible models (including terms of the form $X_1 X_2$), and a simple combinatorial argument shows that, in general, there are $2^{1+p+p(p+1)/2}$ polynomial models of degree 2 or less. Since realistic applications typically need to explore much more complicated functional relationships than just low-degree polynomials, one needs vast amounts of data to distinguish among so many possibilities.

The third formulation is more subtle. In traditional multiple regression, multicollinearity occurs when two or more explanatory variables are correlated, so that the data mostly lie in an affine subspace of $\mathbb{R}^p$ (e.g., close to a line or a plane within a volume). When this happens, there are an infinite number of models that fit the data almost equally well, but these models can be very different with respect to predictions of future values that do not lie in the subspace. As p increases, the number of possible subspaces gets large very rapidly, and thus, just by chance, any realistically finite dataset will tend to concentrate along a few of them. The problem is aggravated in nonparametric regression—the analogue of multicollinearity occurs when data concentrate along a smooth manifold within $\mathbb{R}^p$, and there are many more smooth manifolds than affine subspaces.

The COD poses both practical and theoretical difficulties for researchers who want to let the data select an adequate model. These problems have been deliberately addressed in a body of statistical literature that began in the mid-1980s, and the remainder of this entry sketches the kinds of partial solutions that have been obtained. Further discussion of the COD and its consequences for regression may be found in Hastie and Tibshirani (1990) and Scott and Wand (1991).

2. Methods of Nonparametric Regression

Among the many new methods in nonparametric regression, we focus on six that have become widely used. These are the additive model (AM), alternating conditional expectation (ACE), projection pursuit

regression (PPR), recursive partitioning regression (RPR), multivariate adaptive regression splines (MARS), and locally weighted regression (LOESS).

2.1 The Additive Model (AM)

The AM has been developed by several authors; Buja et al. (1989) describe the background and early development. The simplest AM has the form

$$E[Y_i] = \theta_0 + \sum_{j=1}^{p} f_j(X_{ij}) \tag{1}$$

where the functions f_j are unknown (but have mean zero).

Since the f_j are estimated from the data, the AM avoids the traditional assumption of linearity in the explanatory variables; however, it retains the assumption that explanatory variable effects are additive. Thus, the response is modeled as the sum of arbitrary smooth univariate functions of the explanatory variables, but not as the sum of multivariate functions of the explanatory variables. One needs a reasonably large sample size to estimate each f_j, but under the model posted in (Eqn. 1), the sample size requirement grows only linearly in p.

The backfitting algorithm is the key procedure used to fit an additive model; operationally, the algorithm proceeds as follows:

(a) At the initialization step, define functions $f_j^{(0)} \equiv 0$; also, set $\theta_0 = \bar{Y}$.

(b) At the ith iteration, estimate $f_j^{(i+1)}$ by

$$f_j^{(i+1)} = Sm(\boldsymbol{Y} - \theta_0 - \sum_{k \neq j} f_k^i \mid X_{1j}, \ldots, X_{nj})$$

for $j = 1, \ldots, p$.

(c) Check whether $| f_j^{(i+1)} - f_j^{(i)} | < \delta$ for all $j = 1, \ldots, p$, where δ is the convergence tolerance. If not, return to step 2; otherwise, use the $f_j^{(i)}$ as the additive functions f_j in the model.

This algorithm requires a smoothing operation (such as kernel smoothing or nearest-neighbor averaging) indicated by $Sm(\cdot \mid \cdot)$. For large classes of smoothing functions, the backfitting algorithm converges to a unique solution.

One can slightly generalize the AM by allowing it to sum functions that involve a small number of prespecified explanatory variables. For example, if one felt that income prediction in the AM would be improved by adding a bivariate function that depended on both a person's IQ and their years of education, then a bivariate smoother could be used to fit such a term in the second step of the backfitting algorithm.

2.2 Alternating Conditional Expectations (ACE)

An extension of the AM permits transformation of the response variable as well as of the p explanatory variables. This depends on the ACE algorithm, developed by Breiman and Friedman (1985), which fits the model

$$E[g(Y_i)] = \theta_0 + \sum_{j=1}^{p} f_j(X_{ij}) \tag{2}$$

where all conditions are as given for (Eqn. 1), except g is an unspecified function, scaled to satisfy the technically necessary constraint that $\mathrm{var}[g(Y)] = 1$ (otherwise, the zero transformation would be trivially perfect).

Given variables Y_i and $\boldsymbol{X}_i$, one wants g, θ_0, and $f_1, \ldots, f_p$ so that $E[g(Y_i) \mid \boldsymbol{X}_i] - \theta_0 - \sum_{j=1}^{p} f_j(X_{ij})$ is independent error. Formally, one solves

$$(\hat{g}, \hat{f}_1, \ldots, \hat{f}_p) = \mathrm{argmin}_{(g, f_1, \ldots, f_p)} \left\{ \sum_{i=1}^{n} \left[g(Y_i) - \sum_{j=1}^{p} f_j(X_{ij}) \right]^2 \right\}$$

where $\hat{g}$ satisfies the unit variance constraint. Operationally, one proceeds as follows:

(a) Estimate g by $g^{(0)}$, obtained by applying a smoother to the Y_i values and standardizing the variance. Set $f_j^{(0)} \equiv 1$ for all $j = 1, \ldots, p$.

(b) Conditional on $g^{(k-1)}(Y_i)$, apply the backfitting algorithm to estimate $f_1^{(k)}, \ldots, f_p^{(k)}$.

(c) Conditional on the sum of $f_1^{(k)}, \ldots, f_p^{(k)}$, obtain $g^{(k)}$ by applying the backfitting algorithm (this interchanges the role of the explanatory and response variables). Standardize the new function to have unit variance.

(d) Test for convergence: is $| e^{(k)} - e^{(k-1)} | \leqslant \delta$, where δ is a prespecified small value and

$$e^{(k)} = n^{-1} \sum_{i=1}^{n} \left[g^{(k)}(Y_i) - \sum_{j=1}^{p} f_j^{(k)}(X_{ij}) \right]^2.$$

If convergence is sufficient, set $\hat{g} = g^{(k)}$, $\hat{f}_j = f_j^{(k)}$; otherwise go to step 2.

Steps 2 and 3 calculate smoothed expectations, each conditional upon functions of either the response or the explanatory variables; this alternation gives the method its name.

The ACE analysis finds sets of functions for which the linear correlation of the transformed response variable and the sum of the transformed explanatory variables is maximized. Thus ACE is more closely related to correlation analysis and the multiple correlation coefficient than to regression. Since ACE does not aim directly at regression, it has some undesirable features, for example, it treats the response and explanatory variables symmetrically, small changes can lead to radically different solutions (cf. Buja and Kass 1985), and it does not reproduce model transformations.

To overcome some of these drawbacks in ACE, Tibshirani (1988) invented a variation called AVAS which imposes a variance-stabilizing transformation

in the ACE backfitting loop for the explanatory variables. The basis for this modification is technical, but it improves the theoretical properties of ACE for regression applications.

2.3 Projection Pursuit Regression (PPR)

The AM considers sums of functions taking arguments in the natural coordinates of the space of explanatory variables. When the underlying function is additive with respect to variables formed by linear combinations of the original explanatory variables, the AM is inappropriate. PPR was designed by Friedman and Stuetzle (1981) to handle such cases.

Heuristically, suppose there are only two explanatory variables and that the regression surface looks like a piece of corrugated aluminum. If that surface is oriented so that the corrugations are parallel to one of the axes of the explanatory variables (and hence perpendicular to the other axis), then the AM is an excellent technique. But if the aluminum surface is turned so that the corrugations are not parallel to a natural axis, the AM will fail but PPR would work.

PPR employs the backfitting algorithm and a conventional numerical search routine, such as Gauss–Newton, to fit a model of the form

$$E[Y_i] = \sum_{k=1}^{i} f_k(\alpha_k' \boldsymbol{X}_i) \tag{3}$$

where the $\alpha_1,\ldots,\alpha_r$ determine a set of r linear combinations of the explanatory variables. These linear combinations are analogous to those used in principal components analysis (cf. Flury 1988). An important difference is that these vectors need not be orthogonal, but are chosen to maximize the predictive accuracy of the model as assessed through cross-validation.

Specifically, PPR alternately calls two routines. The first conditions upon a set of pseudovariables that are linear combinations of the original variables; these are used in the backfitting algorithm to find an AM that sums functions whose arguments are the pseudovariables (which need not be orthogonal). The second routine conditions upon the estimated AM functions, and then searches for linear combinations of the original variables that maximize the fit. Alternating iterative application of these methods converges, very generally, to a unique solution.

PPR can be hard to interpret when $r > 1$ in (Eqn. 3). If r is allowed to grow without bound, PPR is consistent, in the sense that as the sample size increases, the fitted regression function will converge to the true regression function.

Unlike the AM, PPR is invariant to affine transformations of the data; this is appealing when the variables used in the research are measured in the same units and have similar scientific rationales. For example, PPR might sensibly be used in biometry to predict a person's weight when the measurements are different long-bone lengths, but it would be less advisable to use PPR when one explanatory variable is ulna length and another is the thickness of the layer of subcutaneous fat.

PPR is very similar to neural net methods. The primary difference is that neural net techniques usually assume that the functions f_k are sigmoidal, whereas PPR allows more flexibility. Zhao and Atkeson (1992) show that PPR has similar asymptotic properties to standard neural net techniques.

2.4 Recursive Partitioning Regression (RPR)

RPR methods have become popular since the advent of CART (Classification and Regression Tree) methodology, developed by Breiman et al. (1984). This article is not concerned with classification problems, and so the basic RPR algorithm we consider fits a model of the form

$$E[Y_i] = \sum_{j=1}^{M} \theta_j I_{R_j}(\boldsymbol{X}_i)$$

where the $R_1,\ldots,R_M$ are rectangular regions that partition $\mathbb{R}^p$, and $I_{Rj}(\boldsymbol{X}_i)$ is an indicator function taking the value 1 if and only if $\boldsymbol{X}_i \in R_j$, and otherwise is zero. Here θ_j is the estimated value for all responses whose explanatory variables lie in R_j.

RPR is designed to be very good at finding local low-dimensional structure in functions that have high-dimensional global dependence. It is consistent; also, it has a powerful graphic representation as a decision tree, which increases interpretability. However, many elementary functions are awkward for RPR, for example, it would approximate a simple linear regression by a stairstep function. In higher dimensions it is usually difficult to discover when the fitted piecewise constant model approximates a standard smooth function.

To be concrete, suppose one were trying to use RPR to predict income from gender and years of education. The algorithm would search all possible intermediate splits in the training sample observations for both explanatory variables, and perhaps decide that the most useful first split is at 12.5 years. For training sample cases with more than high school education, the algorithm repeats the search and perhaps finds the next useful split is at 16.5 years, corresponding to college graduation. Similarly, for those with no education beyond high school, the algorithm searches the cases and may decide to split on gender. This search process repeats in each subset of the training data formed by previous divisions, and eventually no potential split achieves sufficient reduction in variance to justify further partitioning. At that point RPR fits the average of all the cases within the most refined

subsets as the predictions θ_j, and represents the sequence of splits in a decision tree. (Note: Some RPR algorithms are more complex, for example, CART will 'prune back' the final tree, removing splits better balance fit in the training sample against predictive error.)

2.5 Multivariate Adaptive Regression Splines (MARS)

Friedman (1991) describes a method that combines PPR with RPR, using multivariate adaptive regression splines. This procedure fits a model that is the weighted sum of multivariate spline basis functions (tensor-spline basis functions), and takes the form

$$E[Y_i] = \sum_{k=0}^{q} a_k B_k(\boldsymbol{X}_i)$$

where the coefficients a_k are determined in the course of (generalized) cross-validation fitting. The constant term follows by setting $B_0(X_1,\ldots, X_n) \equiv 1$, and the other multivariate splines are products of univariate spline basis functions

$$B_k(x_1,\ldots, x_n) = \prod_{s=1}^{r_k} b(x_{i(s,k)}|t_{s,k}) \quad 1 \leqslant k \leqslant r$$

Here the subscript $i(s, k)$ indicates a particular explanatory variable, and the basis spline in that variable has a knot at $t_{s,k}$. The values of q, the $r_1,\ldots, r_q$, the knot sets, and the appropriate explanatory variables for inclusion are all determined adaptively from the data.

MARS admits a decomposition similar to that obtained from an analysis of variance; it can be represented in a table and similarly interpreted. MARS is designed to perform well whenever the true function has low local dimension. The procedure automatically accommodates interactions between variables and variable selection.

2.6 Locally Weighted Regression (LOESS)

Cleveland (1979) proposed a locally weighted regression technique. Rather then simply taking a local average, LOESS fits a model of the form $E[Y] = \boldsymbol{\theta(x)}'\boldsymbol{x}$ where

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \text{argmin}_{\theta\in\mathbb{R}^p} \sum_{i=1}^{n} w_i(x)(Y_i - \boldsymbol{\theta}'\boldsymbol{X}_i)^2 \qquad (4)$$

and w_i is a weighting function that controls the influence of the ith observation according to the distance and direction of $\boldsymbol{X}_i$ from $\boldsymbol{x}$.

LOESS has good consistency properties, but can be inefficient at discovering some relatively simple structures in data. Although not originally designed for high-dimensional regression, its use of local information makes it quite flexible. Cleveland and Devlin (1988) have generalized LOESS to use polynomial regression instead of the linear regression $\boldsymbol{\theta}'\boldsymbol{X}_i$ in (Eqn. 4), but that does not usually have a strong impact on performance.

3. Comparisons

Currently, our understanding of comparative nonparametric regression performance consists of a scattering of theoretical and simulation results. The key sources of information are:

(a) Donoho and Johnstone (1989) make asymptotic comparisons in terms of the $\mathscr{L}^2$ norm criterion

$$\| \hat{f} - f \| = \int_{\mathbb{R}^p} [\hat{f}(x) - f(x)]^2 \phi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

where p is the dimension of the space and ϕ is the density of the standard normal distribution (i.e., they use a weighted mean integrated squared error (MISE) criterion). They find that projection-based regression methods (PPR, MARS) perform significantly better for radial functions, whereas kernel-based regression (LOESS) is superior for harmonic functions. Radial functions are constant on hyperspheres centered at $\mathbf{0}$, whereas harmonic functions vary periodically on such hyperspheres—a ripple in a pond is radial; an Elizabethan ruff is harmonic.

(b) Friedman (1991) reports simulation studies of MARS alone, and related work is described by his discussants. Friedman examines several criteria; the main ones are scaled versions of MISE and predictive-squared error (PSE), and a criterion based on the ratio of a generalized cross-validation error estimate to PSE. From our standpoint, the most useful conclusions are that:

(i) When the data are pure noise in five and 10 dimensions, for sample sizes of 50, 100, and 200, MARS and AM are roughly comparable and unlikely to find spurious structure.

(ii) When the data are generated from the additive function of five variables

$$Y = 0.1\exp(4X_1) + \frac{4}{1+\exp(-20X_2+10)} + 3X_3 + 2X_4 + X_5$$

with five additional noise variables and sample sizes of 50, 100, and 200, MARS had a slight but clear tendency to overfit, especially at the smallest sample sizes.

All of Friedman's simulations, except the pure noise condition, have high signal-to-noise ratios.

(c) Tibshirani (1988) gives theoretical reasons why AVAS has superior properties to ACE (but notes that consistency and behavior under model misspecification are open questions). He describes a simulation experiment that compares ACE to AVAS in terms of weighted MISE on samples of size 100, finding that AVAS and ACE are similar, but that AVAS performs better than ACE when correlation is low.

(d) Barron (1991, 1993) shows that in a somewhat narrow sense, the mean integrated squared error of neural net estimates for a certain class of functions has order $\mathcal{O}(1/m)+\mathcal{O}(mp/n)\ln n$, where m is the number of nodes, p is the dimension, and n is the sample size. This is linear in the dimension, evading the COD; similar results were subsequently obtained by Zhao and Atkeson (1992) for PPR, and it is likely that the result holds for MARS, too. These results may be less applicable than they seem; Barron's class of functions are those whose Fourier transform $\tilde{g}$ satisfies $\int \mid w \mid \mid \tilde{g}(w) \mid \mathrm{d}w < c$ for some fixed c, and this excludes such standard cases as hyperflats, and becomes smoother as dimension increases.

(e) De Veaux et al. (1993) compared MARS and a neural net on two functions, finding that MARS was faster and more accurate in terms of MISE.

The broad conclusions include (a) model parsimony is increasingly valuable in high dimensions; (b) hierarchical models based on sums of piecewise-linear terms are relatively good; and (c) for any method one can find datasets on which it excels and datasets on which it fails badly.

These short, often asymptotic, explorations do not provide sufficient understanding for users to make an informed choice among regression techniques. Instead, a practitioner might hold out a portion of the data, apply many regression techniques to the rest of the data, and use the models that are built to predict the hold-out sample. Whichever method achieves minimum prediction error is probably the most appropriate for that application.

Bibliography

Barron A R 1991 Complexity regularization with applications to artificial neural networks. In: Roussas G (ed.) *Nonparametric Functional Estimation*. Kluwer, Dordrecht, The Netherlands, pp. 561–76

Barron A R 1993 Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**: 930–45

Breiman L, Friedman J H 1985 Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80**: 580–98

Breiman L, Friedman J, Olshen R A, Stone C 1984 *Classification and Regression Trees*. Wadsworth, Belmont, CA

Buja A, Hastle T, Tibshirani R 1989 Linear smoothers and additive models. *Annals of Statistics* **17**: 453–510

Buja A, Kass R 1985 Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80**: 602–7

Cleveland W 1979 Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association* **74**: 829–36

Cleveland W, Devlin S 1988 Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**: 596–610

De Veaux R D, Psichogios D C, Ungar L H 1993 A comparison of two nonparametric estimation schemes: MARS and neural networks. *Computers in Chemical Engineering* **17**: 819–37

Donoho D L, Johnstone I M 1989 Projection-based approximation and a duality with kernel methods. *Annals of Statistics* **17**: 58–106

Friedman J H 1991 Multivariate additive regression splines. *Annals of Statistics* **19**: 1–66

Friedman J H, Stuetzle W 1981 Projection pursuit regression. *Journal of the American Statistical Association* **76**: 817–23

Flury B 1988 *Common Principal Components and Related Multivariate Models*. Wiley, New York

Hastie T J, Tibshirani R J 1990 *Generalized Additive Models*. Chapman and Hall, London

Scott D W, Wand M P 1991 Feasibility of multivariate density estimates. *Biometrika* **78**: 197–205

Tibshirani R 1988 Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association* **83**: 394–405

Zhao Y, Atkeson C G 1992 Some approximation properties of projection pursuit networks. In: Moody J, Hanson S J, Lippmann R P (eds.) *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann, San Matco, CA, pp. 936–43

D. L. Banks

Exploratory Data Analysis: Univariate Methods

Since about 1970 the phrase 'exploratory data analysis' (EDA) has referred to the approach and techniques promoted by John W. Tukey and a number of collaborators. EDA emphasizes relatively simple numerical and graphical techniques for looking at a set of data, to see what 'appearances' are present, and removing 'partial descriptions' of those appearances, in order to look further. These steps are intended to precede any attempt at drawing inferences from the data, a process that Tukey labeled 'confirmatory data analysis.' Statisticians have used exploratory techniques since they began to work with data, but extensive developments in statistical inference (roughly during the first half of the twentieth century) had relegated exploration to a secondary role as 'descriptive statistics.' Tukey sought to restore the balance between exploratory and confirmatory data analysis. In the process, mainly in his book *Exploratory Data Analysis* (1970–71, 1977), he introduced a variety of new techniques. Within a few years basic EDA techniques, especially displays, were available in major packages and libraries of statistical software. By the

end of the century a number of those techniques had become part of statistical instruction at all levels, including elementary school.

Other efforts, parallel to Tukey's, led to alternative, and generally more complex, exploratory methods. The leading examples are correspondence analysis, developed by Benzecri and the French school (see *Scaling: Correspondence Analysis*); multivariate scaling methods (see *Scaling: Multidimensional*); and semi-parametric approaches to prediction and classification (see *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)*).

1. Main Themes of EDA

At the core of the EDA approach lie four main themes: display, re-expression, resistance and residuals.

(a) *Displays* (most of them graphical) allow the analyst to see the behavior of the data and, as analysis proceeds, the behavior of residuals and various diagnostic measures. EDA emphasizes the frequent use of displays, to ensure that unexpected features are not overlooked. The most common displays include the stem-and-leaf display, the boxplot, scatterplots of residuals, and the spread-vs.-level plot.

(b) *Re-expression* involves applying a single mathematical transformation (such as the square root or the logarithm) to each data value. The aim is to simplify, and thus make more effective, the analysis of the data. Depending on the structure of the data, the simplification may come from promoting symmetry, constancy of variability, straightness of relationship, or additivity of effect. EDA asks, at nearly the first opportunity, whether the scale in which the data are presented is satisfactory. If not, it freely tries a likely transformation.

The re-expressions most often used in EDA come from the family of functions known as power transformations, which take y into y^p (almost always with a simple value of p such as $\frac{1}{2}$, -1, or 2), together with the logarithm (which, for data analysis purposes, fits into the power family at $p = 0$) . For example, in a study of the levels of nicotine in the blood and urine of smokers and nonsmokers, the analysis gained simplicity by re-expressing the nicotine concentrations in the logarithmic scale.

(c) *Resistance* provides insensitivity to misbehavior or unusual behavior in data. An analysis or summary is resistant if an arbitrary change in any small part of the data produces only a small change in the analysis or summary. EDA emphasizes resistance because experience suggests that 'good' data seldom contain less than about 5 percent gross errors.

The notion of robustness is akin to resistance, but not the same. Robustness generally implies insensitivity to departures from assumptions surrounding an underlying probabilistic model. (Resistance is sometimes mentioned as one aspect of 'qualitative robustness.')

In summarizing the location of a sample, the median is highly resistant. (In terms of efficiency, it is not so highly robust because other estimators achieve greater efficiency across a broader range of distributions.) By contrast, the mean is highly nonresistant. A number of exploratory techniques for more structured forms of data provide resistance because they are based on the median.

(d) *Residuals* are what remain of data after a summary or fitted model has been subtracted out, according to the schematic equation

$$\text{residual} = \text{data} - \text{fit}.$$

For example, if the data are the pairs (x_i, y_i) and the fit is the line $\hat{y}_i = a + bx_i$, then the residuals are $r_i = y_i - \hat{y}_i$.

EDA regards a careful examination of the residuals as an essential step in analyzing a set of data. This emphasis reflects the tendency of resistant analyses to provide a clear separation between dominant behavior and unusual behavior in the data. When the bulk of the data follows a consistent pattern, that pattern determines a resistant fit. The residuals then contain any drastic departures from the pattern, as well as the customary chance fluctuations. Unusual residuals suggest a need to check on the circumstances surrounding those observations. As in more traditional practice, the residuals can warn of systematic difficulties with the data—curvature, non-additivity, and non-constancy of variability.

The examples and discussion that follow illustrate displays (via stem-and-leaf displays and boxplots), resistance (median polish), residuals, and re-expression.

2. Example 1: Stem-and-Leaf Displays (Florida voting)

The initial results of the 2000 US presidential election focused intense attention on the state of Florida, where the difference in total votes cast for George W. Bush and Albert Gore was so small that state law required a recount and thus delayed the outcome. Among the unusual features of the Florida voting was the relatively large number of votes cast in Palm Beach County for Patrick Buchanan, the candidate of the Reform Party (3407 out of 430,762 votes). Critics argued that most of those votes had been intended for Gore, but were recorded for Buchanan because of a problem with the type of ballot.

A number of analyses compared the results in Palm Beach County against those in other counties. Without explaining the reasons for those choices, one analyst selected the data from the 30 counties in which Bush received more than 30,000 votes and calculated the ratio of the number of Buchanan votes to the number of Bush votes, concluding that the probability of the ratio in Palm Beach County by pure chance was extremely small.

```
0|9              Dade
1|33456679
2|13445677
3|0222778
4|023
5|5              Marion
6|2              Santa Rosa
7|9              Palm Beach
```

Figure 1
Stem-and-leaf display of the percentage of votes cast for Patrick Buchanan in the 30 Florida counties in which George W. Bush received more than 30,000 votes. The unit for the leaves is 0.1 percent

```
 0|9                  Dade
 1|334456679
 2|1134456666777
 3|00222677889
 4|0023367788
 5|1133456
 6|12225
 7|1569
 8|79
 9|
10|4                  Holmes
11|06                 Washington, Gulf
12|
13|
14|
15|
16|3                  Liberty
17|4                  Calhoun
```

Figure 2
Stem-and-leaf display of the percentage of votes cast for Buchanan in all 67 Florida counties. The unit for the leaves is 0.1 percent

For those 30 counties Fig. 1 shows a stem-and-leaf display of the percentage of votes for Buchanan (out of the total number of votes for four candidates: Bush, Gore, Buchanan, and Ralph Nader). In overall outline this display resembles a histogram with an interval width of 0.1 percent, but it shows more detail of the data values by using the 0.01 percent digit of each percentage instead of merely enclosing a uniform amount of area. This feature makes it easy to go back from a part of the display to the individual data values (and their identities) in the listing of the data—a valuable step in EDA.

Several alternatives and refinements permit the stem-and-leaf display to accommodate a wide variety of data patterns. The interval width (that is, the range of data values represented by one line in the display) may be 2, 5, or 10 times a power of 10; and stray observations may be listed separately at each end so as not to distort the scale. The use of positive and negative stems readily handles sets of residuals.

In Fig. 1 Palm Beach County's 0.79 percent for Buchanan is noticeably high (indeed, another EDA technique would flag it as 'outside' and hence deserving investigation). But the 0.55 percent for Marion County and the 0.62 percent for Santa Rosa County are also noticeable at the high end. The layout of the stem-and-leaf display makes it easy to include tags that identify a few data values at either end.

The analyst's choice of counties with more than 30,000 votes for Bush naturally raises a question about the extent of support for Buchanan in the other 37 counties of Florida. Figure 2 shows a stem-and-leaf display for all 67 counties. In the context of the whole state, Palm Beach County no longer stands out; seven counties had a percentage for Buchanan higher than 0.79 percent, and two of those were above 1.6 percent. Those seven counties, however, are small; most cast fewer than 8,000 total votes for the four candidates. Also, geographically, they are all located in north Florida, far from Palm Beach County. Thus, among counties with larger numbers of voters, it still appears that Palm Beach County produced an unusually large percentage for Buchanan. A stem-and-leaf display of the vote totals (not shown) reveals a clear gap around 40,000. Modifying Fig. 1 to include all counties with totals greater than 40,000 would add three counties, whose percentages for Buchanan were 0.21 percent, 0.26 percent, and 0.47 percent.

Even after this modification the position of Palm Beach County in Fig. 1 does not form the basis for a conclusion. It does invite further examination. Other analysts responded, taking into account the voting in counties throughout the US, precinct-level voting in Palm Beach County, and voting in other Florida elections. For example, even if one uses only data from Florida counties, a scatterplot of the percentage for Buchanan against the percentage of Bush (or against the percentage of Gore) shows that Palm Beach County is very far from the other counties.

3. *Example 2: Boxplots (childhood immunization)*

In the United States the Centers for Disease Control and Prevention use the National Immunization Survey (NIS) to monitor the extent to which children 19 to 35 months of age have received the recommended vaccinations. A telephone survey, the NIS uses random-digit dialing to locate households that contain such children, in each of 78 Immunization Action Plan

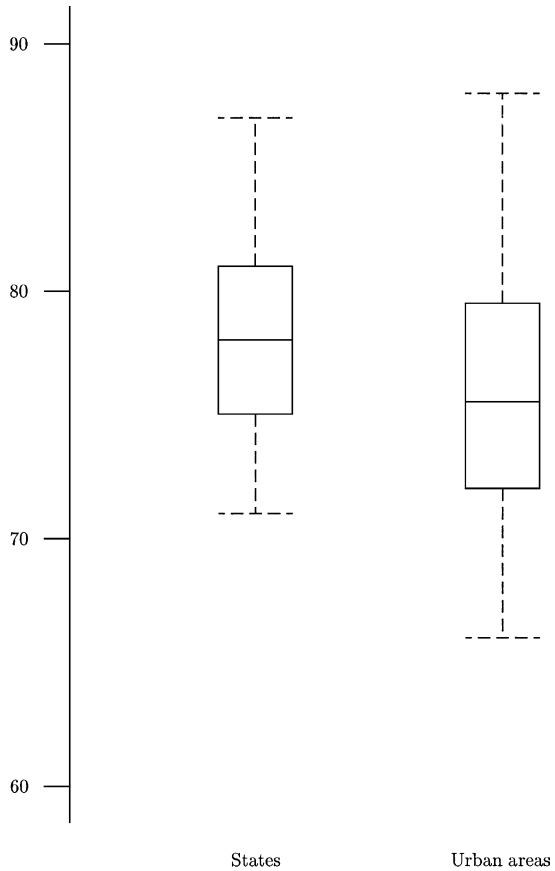

Figure 3
Parallel boxplots of the percentage of children, age 19 to 35 months, who are up-to-date on the 4:3:1 series of vaccinations in the 50 states and 28 urban areas, U. S. National Immunization Survey, 1997

(IAP) areas that cover the US. It interviews the children's parents or guardians, and also (by mail) contacts the children's vaccination providers to obtain data from the child's medical record. In a year the NIS completes household interviews for approximately 400 children in each IAP area, and it obtains provider data for about two-thirds of them.

For each IAP area, each state, and the whole US, the NIS produces estimates of vaccination coverage (the percentage of children who are up-to-date) for several individual vaccinations and for selected series of shots, such as the 4:3:1 series (4 or more doses of diphtheria and tetanus toxoids and pertussis vaccine, 3 or more doses of poliovirus vaccine, and 1 or more doses of measles-containing vaccine), which the CDC has used in allocating funds to states.

The CDC selected the 28 urban IAP areas for a variety of reasons, including concern for lower rates of vaccination coverage. For one overall comparison, Fig. 3 shows parallel boxplots of the estimates of 4:3:1 coverage in 1997 for the 50 states and the 28 urban areas.

In a boxplot (also called a 'schematic plot') the box extends from the lower hinge (an approximate quartile) to the upper hinge and has a line across it at the median. The dashed lines show the extent of the data, except for observations that are likely strays (defined according to a rule based on the hinges). Such strays (when present) appear individually, in order to focus attention on them. The general intent is to indicate the median, outline the middle half of the data, and show the range, with more detail at the ends if needed.

The message of Fig. 3 is that 4:3:1 vaccination coverage is typically lower in the urban areas than in the states—by about 2 percentage points according to the hinges and median. But the range is greater in the urban IAP areas, from 66 percent in Houston and 67 percent in El Paso, Texas to 88 percent in Boston, Massachusetts.

4. *Example 3: Median Polish, Residuals (hotel occupancy)*

In 1983, as several new hotels were under construction in Boston, operators of existing major hotels were concerned about the impact on their business. Table 1 shows the occupancy rates for 11 major Metropolitan Boston hotels, month by month, for January through June 1983. One approach to exploring such a two-way table uses an additive fit of the form

$$\hat{y}_{ij} = m + a_i + b_j$$

with a row effect, a_i, for each hotel and a column effect, b_j, for each month. The EDA technique known as median polish arrives at such a fit by iteratively subtracting row medians and column medians, first from the table of data and then from the table left by the previous iteration. The common value, m, is determined by centering the effects so that the median of the a_i is 0 and the median of the b_j is 0. The use of the median gives the fit resistance: isolated behavior of the data in a few cells of the table will tend to have only a minor impact on the fit.

For the hotel occupancy rates Table 2 shows the two-way table of residuals, $y_{ij} - \hat{y}_{ij}$, bordered by the hotel effects, the month effects, and the common value. The month effects (ranging from -22.0 to $+15.0$) show occupancy rates in Boston hotels increasing substantially from January to June. Among the hotel effects the Marriott Long Wharf (at 21.5) stands out as having higher occupancy than the other hotels, and the Colonnade and the Park Plaza (at -12.0 and -12.5, respectively) were substantially lower. A stem-and-leaf display of the residuals (not shown) calls attention to three sizable negative residuals (-20.0,

Table 1
Occupancy rates of major Metropolitan Boston hotels, 1983 (percent)

Hotel	Jan	Feb	Mar	Apr	May	June
Bostonian	56	65	71	86	87	93
Colonnade	30	24	34	73	74	80
Copley Plaza	52	62	80	84	83	84
Hyatt	52	71	72	89	86	90
Lenox	38	45	55	71	78	78
Marriott Long Wharf	77	84	86	95	94	96
Meridien	54	63	61	73	81	82
Park Plaza	40	42	47	64	72	72
Parker House	57	60	60	78	82	85
Ritz-Carlton	45	53	52	75	83	82
Sheraton-Boston	49	37	60	78	85	82

Table 2
Result of applying median polish to the hotel occupancy rates: table of residuals, bordered by effects and common value

Hotel	Jan	Feb	Mar	Apr	May	June	Hotel effect
Bostonian	−0.0	0.0	1.0	−0.0	−3.0	0.0	9.0
Colonnade	−5.0	−20.0	−15.0	8.0	5.0	8.0	−12.0
Copley Plaza	−0.5	0.5	13.5	1.5	−3.5	−5.5	5.5
Hyatt	−3.5	6.5	2.5	3.5	−3.5	−2.5	8.5
Lenox	−3.0	−5.0	0.0	0.0	3.0	0.0	−6.0
Marriott	8.5	6.5	3.5	−3.5	−8.5	−9.5	21.5
Meridien	7.0	7.0	0.0	−4.0	0.0	−2.0	−0.0
Park Plaza	5.5	−1.5	−1.5	−0.5	3.5	0.5	−12.5
Parker House	9.0	3.0	−2.0	−0.0	−0.0	0.0	1.0
Ritz-Carlton	0.0	−1.0	−7.0	0.0	4.0	0.0	−2.0
Sheraton	2.0	−19.0	−1.0	1.0	4.0	−2.0	0.0
Month effect	−22.0	−13.0	−8.0	8.0	12.0	15.0	69.0

−19.0, and −15.0) and one sizable positive residual (+13.5). These indicate that the Colonnade (beyond its generally low occupancy) had unusually low occupancy in February and March, that the Sheraton-Boston had unusually low occupancy in February, and that the Copley Plaza had unusually high occupancy in March.

5. *Example 4: Everyday Re-expressions*

In a number of applications from a variety of fields, a re-expression has become so customary that no one mentions it. By the time the data or results are presented, they are already in the transformed scale. Examples include established relationships in science and engineering, as well as contemporary data (see also Hoaglin 1988).

For earthquakes the Richter scale expresses the strength in logarithmic units (base 10). The magnitude is defined as the logarithm of the amplitude recorded by a particular type of seismograph (at a specified distance from the epicenter).

In measuring the intensity of sounds, the decibel scale is logarithmically related to sound pressure.

Measurements of the fuel consumption of automobiles often involve determining the number of gallons of gasoline that a car used on a standard test course. Thus, the ratings in miles per gallon are the result of a reciprocal transformation.

6. *Literature*

The first published presentation of exploratory data analysis was the preliminary edition (1970–71) of the book by John W. Tukey, whose 1977 edition represents the definitive account of the subject. Mosteller and Tukey (1977) also includes substantial discussion of exploratory attitudes and techniques.

The two volumes of the *Statistician's Guide to Exploratory Data Analysis* (Hoaglin et al. 1983, 1985) discuss the rationale and development of a number of EDA techniques, explain and illustrate connections with classical statistical theory, and introduce new techniques.

Hoaglin et al. (1991) extend the basic attitudes and techniques of EDA to factorial analysis of variance.

Broader discussions of the roles of exploratory and confirmatory data analysis in scientific inquiry appear in Tukey (1972) and Tukey (1980).

Bibliography

Hoaglin D C 1988 Transformations in everyday experience. *CHANCE* 1, **4**: 40–5

Hoaglin D C, Mosteller F, Tukey J W 1983 *Understanding Robust and Exploratory Data Analysis*, Wiley Classics Library edition 2000. Wiley, New York

Hoaglin D C, Mosteller F, Tukey J W 1985 *Exploring Data Tables, Trends, and Shapes*. Wiley, New York

Hoaglin D C, Mosteller F, Tukey J W 1991 *Fundamentals of Exploratory Analysis of Variance*. Wiley, New York

Mosteller F, Tukey J W 1977 *Data Analysis and Regression*: A Second Course in Statistics. Addison-Wesley, Reading, MA

Tukey J W 1970–71 *Exploratory Data Analysis*, limited preliminary edition. Addison-Wesley, Reading, MA [available from University Microfilms]

Tukey J W 1972 Data analysis, computation, and mathematics. *Quarterly of Applied Mathematics* **30**: 51–65

Tukey J W 1977 *Exploratory Data Analysis*. Addison-Wesley, Reading, MA

Tukey J W 1980 We need both exploratory and confirmatory. *The American Statistician* **34**: 23–5

D. C. Hoaglin

Expression and Interaction: Microfoundations of Expressive Forms

All expressions such as art, entertainment, religion, science, law, and subcultural ways of life have microfoundations, meaning that they are the outcomes of people interacting with one another. Artists are born into a family, have friends, enemies, teachers, rivals, business partners, co-workers, or supporters who influence what they produce. Even the most gifted of them receive, especially at the beginning, some form of guidance or inspiration to create by mentors if not through formal training. What is more, all expressive endeavors are received by other human beings if they are to have some social meaning. If an expression is not heard, viewed, read, or otherwise appreciated by anybody, then it has no social existence. That expression, which involves the collaboration of many actors, has been taken for granted mainly by social scientists who insist on studying it like other social phenomena without divorcing it from its social context. The emphasis has been placed on how and to what extent such collaboration shapes various forms of expression. The theoretical roots of this approach can be traced to a branch of social psychology called symbolic interactionism. The relationship between expression and interaction has long been the focus of study in the tradition of symbolic interactionist theory.

1. *The Intellectual Legacy of Symbolic Interactionism*

Charles Horton Cooley and George Herbert Mead are two of the early theorists of symbolic interactionism who studied the processes through which new members of a group are socialized into their respective communities. Cooley ([1902] 1964) coined the term 'looking-glass self' to describe how an individual learns to see oneself from another's point of view. Mead (1934) expanded the analysis of social learning beyond two-person interactions and showed that children gain increasingly complex understanding of the response of a variety of other people. Cooley and Mead's emphasis on microsettings in which socialization of new group members occurs inspired later researchers to study how, just like the human self in general, the self of an artist is shaped through social interaction. It is, however, the work of Erving Goffman that offers one of the best examples of studies in the symbolic interactionist approach that focus increasingly on the centrality of the relationship between expressive symbols and interaction.

1.1 *The Work of Erving Goffman and His Followers*

Goffman introduced the dramaturgical model of interaction in his earlier work *The Presentation of Self in Everyday Life* (1959). He viewed social life as a 'theatrical' performance in which actors give creative interpretations to their roles. They employ 'props' and 'scenery' to support the impression of their self that they seek to convey to others. Much of such impression management takes place in public domains and individuals may withdraw to their private settings as a 'backstage' where they can prepare themselves for their next performance. Examining props and masks as expressive symbols manipulated for an audience, Goffman at the same time paid particular attention to the ways in which the conceptions of the self are shaped by the attitudes and reactions of other actors.

In his later work *Frame Analysis* (1974), Goffman abandoned his reliance on theatrical metaphor and shifted attention to the settings of recurrent human actions. Social interaction is made meaningful by frames, which tell an observer how to define the situations. Frames provide a way of organizing experience. Without them, the social world is a jumble of meaningless facts. Goffman showed how human behavior is subject to extraordinarily subtle frame manipulations.

Goffman's insight is directly observable in a number of studies, a representative example being Manning's

research on police work (1977). Analyzing interactions between the police and citizens as actor and audience in Goffman's sense, Manning examines ways in which the police develop a finely-tuned capacity to sort out 'police work' from other events of little relevance. In the drama of police work, affective or emotive elements of citizens' messages are systematically recoded by the logical code of the police. Works in the tradition of symbolic interactionism such as Manning's emphasize the role of social interaction in defining, creating, and transmitting meanings.

1.2 Identity Work

One of the key concepts in the symbolic interactionist approach is identity, which is produced and confirmed by social interaction. According to Goffman, a performance can be said to have succeeded if an actor has confirmed a certain self-identity to oneself and to the audience. In much the same spirit, Snow and Anderson examined the 'identity work' of the homeless in a city in Texas (Snow and Anderson 1993). They found that the homeless manage interactions so that they can project a certain image of themselves. Some embrace the homeless role and take pride in their freedom and ability to survive in the harsh conditions, while others distance themselves from such people, insisting that they are not like them and therefore do not need the services of relief agencies. Still others live in the fantasies of their past or future prosperity. Divergent as they are, the homeless are like any other person, performing impression management in their interaction with others in order to control the meanings they present to them.

Another striking example of identity work is provided by Allison in her study of a hostess club in Tokyo (1994). A hostess pours drinks, lights cigarettes, and makes flattering conversation with businessmen. The language and behavior at the table are sexually charged, yet sex is confined to conversation. Allison argues that sexuality that is all talk not only structures the identity of women but that of men. Night-time entertainment paid by corporations homogenizes male workers. Through a playful sexuality that titillates men's egos and makes them feel self-assured, a particular construct of 'maleness' is developed that serves to align men with the workplace. Going to hostess clubs, therefore, is not primarily for or about sex. It concerns, rather, the confirmation of men's identity as workers.

Notions of the self are like other cultural constructions in that they vary across cultures. Recent literature on self in Japan has emphasized the situatedness, multiplicity, and shifting nature of self (see, for example, Bachnik and Quinn 1994 and Kondo 1990). The unitary 'I' presupposed in the Western discourse is made anew each time according to the particularities of a given interaction.

In sum, from the symbolic interactionist point of view and many others that go well beyond its boundaries, the human self is susceptible to diverse outside influences so that interactions with others powerfully shape the course of individual action. Such human interactions, moreover, give rise to various forms of expression. Researchers have shown how an impression favorable to oneself, a proper identity as a homeless, or an identity as an efficient Japanese businessman has been expressed through theatrical performance, language and storytelling, or sexually-charged conversation and behavior in a hostess club. The legacy of symbolic interactionism is apparent in three aspects of the relationship between expression and interaction examined in detail below. They are discussed in the following order: how expressions are produced, defined, and received as people interact with each other.

2. Production of Expressive Forms and Interaction

Novice producers of various expressive forms usually acquire skills and motivations to produce from interactions with fellow workers and experienced mentors. What is more, even before novices are trained to create, they must often be taught that there is indeed an audience for their work. Without such knowledge, they may not be motivated to learn how to express themselves through creative work. Once they understand that audiences can recognize their work, they begin experimenting with their style. Expressive processes first of subculture and next of art are discussed in turn.

2.1 Subculture

The culture of a group is structured through an ongoing negotiation between members of the group. Rather than assuming the content and meaning of culture *a priori*, many studies of subculture follow symbolic interactionist theory and examine how microlevel group processes give rise to a set of symbols, meanings, and behavioral norms—often a whole way of life—that bind the group members. A representative case is Fine's analysis of group culture or 'idioculture' that emerged in American little league baseball teams (1987). Specifically, he identifies five elements that are necessary for a cultural item to enter the life of a group that interacts intensively over a long period of time: the item must be known, usable, functional, appropriate, and repeatedly triggered.

Although the American subculture studies customarily avoid political questions, subculture often has explicitly political content. What is critical for understanding such subculture is to examine the relationship between a subculture and the social class within which

it emerges and the relationship between a subculture and the dominant culture. When subculture is a site of political struggle, the central problem becomes understanding how interacting members of the social class connect and negotiate their lived experiences to subcultural representations. British Cultural Studies led by Stuart Hall, Tony Bennett, and associates at the Birmingham Centre emphasize the political context of subculture.

British Cultural Studies is often said to take much from Marxism and structuralism, but it is clear that it also incorporates many elements of symbolic interactionism. Using case studies and ethnographic methods, the emergence of subculture is analyzed as a collective solution to the problems faced by members of the social class such as lack of educational and employment opportunities. An illustrative study is Hebdige's research on 'style' developed by postwar English youth, which includes dress, talk, rituals, and music (1979). Hebdige details the way in which the white working-class youth deal with their problems by developing a style that define their identity and prove their membership in a group. They create a new style by borrowing items from existing styles and putting them together in novel ways. Hebdige's work illustrates that working-class youth possess few other means than creating their own style in order to attain some collective status within a society and to exercise some control over their lives.

One of the interesting phenomena concerning subculture is the development of a tension that members of a group experience between staying faithfully committed to the aesthetic goals of the group and compromising them for financial and reputational gain in the commercial world of the dominant culture. This issue is examined by Lachmann in his study of graffiti writers (1988). Lachmann analyzes how graffiti writers of New York created organizational and ideological bases for the allocation of fame, and why they eventually lost such bases. According to Lachmann, the vast majority of graffiti writers, called taggers, compete on the basis of the quantity of tags, while the minority become muralists who seek fame through their artistic styles. The latter meet at 'writers' corners' to allocate fame for style among skilled colleagues. However, when the corners were destroyed by police harassment, some writers were recruited to paint graffiti for sale in galleries, which resulted in fragmentation of the graffiti subculture. Lachmann's study shows how oppositional aesthetic standards of graffiti writers developed through social interaction with their mentors, colleagues, and audiences, and how such standards were weakened by co-optive interventions of the larger society.

Of course, subcultural expression is not limited to Western culture. Sato 1991 has described the teenaged *bosozoku* drivers (the 'violent driving tribe'), who in tens, sometimes in hundreds, careen motorcycles and cars on the busy streets of large Japanese cities with deafening sounds. Many are dressed carefully in their elegant and bizarre costumes of kamikaze party uniforms combined with American punk, surfer, or rock-and-roll gear. Their cars and bikes are customized beyond recognition of their original appearance. The majority of *bosozoku* drivers are from middle-class families and seldom get involved in other unlawful activities. Analyzing the symbols, fashions, and customs of this youth subculture, Sato argues that such deviant lifestyles satisfy the cravings of the affluent youth for excitement and new experience in what is commonly regarded to be a harmonious and homogeneous society.

2.2 Art World

Values, norms, beliefs, and attitudes of a subgroup develop through interaction. However, it is not only such 'implicit culture' but also 'recorded culture' (Crane 1994) including information, entertainment, science, technology, law, education, and art that are produced as the results of social interaction. Artists such as painters, sculptors, novelists, and poets are generally considered to produce their works individually. Such a view of solitary creation has the danger of oversimplifying an activity that is usually much more collaborative. For example, novelists often receive a variety of help from editors when writing novels. Many paintings of the past were produced in the names of principal painters who supervised their teams of apprentices, assistants, and students.

Another popular image of an artist is someone who is engrossed in self-expression, oblivious of common thoughts and beliefs and activities of mundane everyday life. However, artists too have social existence in that they live and work within the constraints of institutions and conventions. Those who are rejected by existing institutions, audiences, and patrons are forced to face the problem: either they give up creative activities altogether or become determined to live as innovators. Some of the works of those that rejected conventions and became liberated from creating for the demands of others came to be appreciated by later generations of people, although they and their work were not typically understood by people of the time.

Social interactions thus sometimes permit, enable, and encourage producers of cultural works to engage in creative activities. Other times they constrain and discourage. One of the best studies of collaborative creators is provided by Becker who describes 'art worlds' as more or less institutionalized subcultural community (1982). Each world centers around one of four types of artists (integrated professionals, mavericks, folk artists, or naive artists) and the extent to which the community is organized differs according to the type of artist. The most institutionalized art world is that of integrated professionals, who produce

canonical works following conventions current in the time.

Participants of a professional art world including artists, audience, critics, and support personnel share knowledge of conventional ways of dog things in that art. Such conventions are the terms on which participants cooperate. For example, there are conventions regarding materials to be used, abstractions to be employed to convey particular ideas or experiences, proper length of a musical or dramatic event, proper size and shape of a painting or sculpture, and relations between artists and audience specifying the rights and obligations of both sides. Because things do not have to be decided anew each time, conventions allow decisions to be reached quickly, and thus make it possible for artists and support personnel to coordinate their activities with ease and efficiency.

Becker's analysis of an art world consisting of participants cooperating through the medium of conventions demystifies art. From this point of view, art is not a spontaneous creation of a genius, but is rather a well-planned and orchestrated action of a variety of social actors. Moreover, because art is a collective process, without interaction among participants, it may not successfully result in public expression. The view of art as a collective process points to the danger of crediting only one artist for a particular work.

An art world as envisioned by Becker is a self-contained unit that is in some ways extremely isolated from other segments of the society or from other art worlds. An art world is decontextualized from the political, historical, and ideological movements of the time. Because of their intense focus on interpersonal factors to their relative neglect of attention to forces that influence creative processes external to an art world, it has been aptly pointed out that Beckerian microlevel studies concentrate less on the production of culture and more on the culture of production (Peterson 1994).

The application of the art-world approach may not be limited to the analysis of high culture or Western culture. Japanese traditional arts such a Noh play, samisen music, a mountain ascetic kagura dance (Fukushima 1995), as well as popular theater (Ukai 1994) have all been analyzed with particular attention paid to face-to-face mentoring through which performers acquire the critical body postures and movements.

2.3 Interaction with Other Social Actors

Various members of an art world including fellow artists, mentors, support staff, assistants, students, audiences, critics, dealers, buyers, producers, corporations in culture industries, and mass media influence expressive activities of an artist. However, it is not only interactions with members of the art world that shape the careers of artists. Nonmembers sometimes exert equally strong influence on artists and their work.

For example, marital and family experience may differentially affect the chances of men and women to become creators and audiences of cultural works. This is because, in general, an individual must be isolated from the flurry of daily activities in order to engage in a form of art. It is difficult for adult women who are busy with housework and child care to attain such isolation. Adult men, on the other hand, are usually free to abandon housework and indulge in the moment of solitude. The relationship between gender and isolation, however, works in the other way for children. Parents and teachers tend to permit girls to be quiet and withdrawn, but not boys who are encouraged to engage in activities outside. Rogers shows in her study of literary production and reception (1991) that such differences have consequences on men and women with the former over-representing as writers.

It has also been pointed out that the quality of the relationships with spouses and other significant partners may influence the quality of artistic activities. In his study of *avant-garde* visual arts in the SoHo section of Manhattan, New York, Simpton has found that there was a tendency for successful artists to have supportive partners and unsuccessful artists to have overprotective and smothering partners (Simpton 1981).

3. Expressive Forms Defined Through Interaction

Art works themselves have typically been given a less central place in sociological studies of art than in humanistic studies. Art works have traditionally been treated as 'black boxes' whose meanings and interrelationships do not require analysis. Focusing on art objects has been problematic for some sociologists, partly because this was considered to bring them dangerously close to making value judgments about them. Contemporary social science disciplines, however, are questioning this approach (see, for example, Griswold 1986).

Social scientists have become increasingly aware that labels such as art and nonart, fine art and popular art, professional art and naive art, and art and craft are produced through social interaction. It has been shown that there is nothing inherent in objects, paintings, dances, or music that determines them as art or nonart. They become art or nonart when they are defined as such by social actors. Aesthetic standards of art works are set by artists, spokespersons, and influential members of the public. As in all other social fields, social meanings of art works arise in the interactions among the actors. This is why functional items such as toilets, hats, typewriters, and

teapots are sometimes displayed in art museums as works of art. They become art when they are made nonfunctional both socially and technically by artists and displayed as such by museum owners (Becker 1982). Another example is contemporary jazz music, whose market has become so fragmented with regard to age, race, and musical idiom that the term jazz has become increasingly difficult to define (Rockwell 1983). It is suggested that jazz can now be more accurately defined in terms of its support system including club and concert circuits, record labels, and critical apparatuses than in terms of its content.

The distinction between art and craft is in fact not clear-cut. Craft workers are generally considered to follow the conventions primarily of utility and technical virtuosity, while artists are thought to place emphasis on uniqueness and beauty. However, Neapolitan has found that craft workers also have considerable pride in the beauty of their work, and furthermore, place high value on preservation of traditional values and ways of doing things (1986). In addition, there are craftpersons who claim to work to express their inner feelings in their work. These people appear to be more fine art than craft oriented.

In his study of restaurant cooks, Fine argues that although the expressive side of production is often in conflict with other organizational goals such as efficient and profitable operation, aesthetic value is central to workers (1992). Just as artistry is constrained by market and institutional imperatives, factory work has a creative component. Not only is art like work then, but work is like art. Both art and nonart workers negotiate and renegotiate their standards of production in their chosen field where aesthetic autonomy and institutional control interpenetrate.

4. *Reception of Expressive Forms and Interaction*

A common assumption is that it is possible to distinguish between the artists and the audiences for whom works are intended. In many sociological studies of culture, however, a sharp distinction between artists and audiences disappears, since both are participants in the creative process. Although audiences usually devote less time to any particular piece of work than artists or other professional members, the contribution that audiences make to art works is not slight. It is ultimately audiences that determine which artist's works is to be displayed in a gallery or which musician is to hold a concert by participating or not participating in an event.

4.1 *Period Eye and the Role of Audience in Creation*

If a piece of work is considered to be made anew each time an audience experiences and appreciates it, only those properties of the work that the audience takes note of can be said to exist in his or her experience. From this point of view, work of art is whatever that audience perceives. What audiences know and are capable of appreciating are extremely important.

Knowledge and skills used in appreciating art works are in part determined by experience and training the audience has had. Art works aimed at a certain segment of the population is usually made based on the assumption of what that particular segment knows. Therefore, concerts attended by an inner circle of students will be different from those held for the lay public, just as books written for adults are not the same as those intended for children.

The knowledge and skills audiences rely on to appreciate art works also vary according to time and place. A cognitive style is rooted in society so that people of particular period and culture have their own ways of appreciating art works. Baxandall has called this 'period eye' (1972). Audiences use skills of the period when appreciating art works, and artists respond to this. For example, fifteenth-century Italian merchants were trained in mathematical understanding of lines and shapes. Such understanding constituted a significant portion of the merchants' intellectual life, as many played games, told jokes, and otherwise took pride in their prowess in it. They were naturally sensitive to pictures which used the same devise. Painters were thus encouraged to display mathematical skills in works for merchants, who were their good patrons. Baxandall asserts that it was in fact for these skills that merchant patrons paid the painters. Most present-day audiences have lost the ability to respond to the elements of fifteenth-century Italian paintings.

4.2 *Symbolic Manipulation of Signs*

A clear distinction is also difficult to make between producers and receivers of subcultural expressions. Members of a subcommunity take turns in engaging in and appreciating expressive endeavors, and rewards are allocated within the subculture. The view of reception as an active process has been proposed most forcefully by theorists of British Cultural Studies and those inspired by them (see, for example, Cruz 1994, Morley 1986, and Press 1991). Reception may be at the end of a communication process, but it is by no means automatic. Asserting active and creative interpretations by audiences, researchers have explored the ways in which consumers make critical use of elements of popular culture.

For example, in the aforementioned study of British youth subculture, Hebdige shows how the working class youth challenge the symbolic order of the society by reinterpreting mass-produced materials in a way oppositional to the dominant culture. By displaying a mismatch between what the materials are intended for in the dominant culture and how they are used by the

resistant youth, they challenge the inevitability and naturalness of class, gender, and other stereotypical categorizations found in the society. Mass-produced goods become tools for motivated communication. Reception can be a site of political struggle.

4.3 Reputation

Finally, contrary to the idea that there is some quality inherent in the work of art that determines its value, the reputations of artists and works are also in part outcomes of social processes. In the case of visual arts for example, art dealers, critics, auction houses, museums, government, corporate, and individual patrons, and finally, mass media all play roles in determining the value of works and artists (see, for example, Moulin 1987 and White and White 1965). Of course, artists themselves have a large stake in protecting and projecting their reputations, and they can achieve this by such means as keeping adequate records of their works and making arrangements for their proper custodianship. In addition, whether or not there is a survivor dedicated to preserving the value of their works will influence the durability of their reputations.

In their study of the once relatively popular form of art, etching, Lang and Lang (1988) ask why the reputations of many women etchers who were critically acclaimed during their lifetimes are forgotten while those of men have survived. They find the primary reason in different efforts expended by the spouse and family on behalf of the deceased to preserve and enhance the artist's reputation. The Langs show that the surviving wives typically worked diligently to link their husbands' works to influential networks. However, the likelihood of there being survivors with similar financial or emotional stake in perpetuating female etchers' reputations was lower. Even if married, many women usually outlived their husbands and sometimes their children. Much of the work of female etchers has been lost to posterity, not because of the second-rate quality of their art but because of their marital and family relations.

In sum, social scientists have attempted to study art and artists like any other human beings and actions that cannot escape being affected by interactions with surrounding networks of people. If demystification of art and artists has called into question the conventional distinction made between varying qualities of art, art and nonart, artists and their collaborators, and artists and audience, future research can no longer continue to shy away from analyzing art works themselves. At the same time, more research will make efforts to link different levels of analyses, since an art world, just like human communities in general, is not immune to profound changes recent rapid innovation in information technology will bring. It will become increasingly important to examine how changes in the social world including the electronic media revolution directly or indirectly influence interactions and expressions at the microlevel.

See also: Cultural Assimilation; Cultural Diversity, Human Development, and Education; Cultural Expression and Action; Cultural History; Cultural Psychology; Cultural Variations in Interpersonal Relationships; Ecology, Cultural; Expressive Forms and the Evolution of Consciousness; Expressive Forms as Generators, Transmitters, and Transformers of Social Power

Bibliography

Allison A 1994 *Nightwork: Sexuality, Pleasure, and Corporate Masculinity in a Tokyo Hostess Club*. University of Chicago Press, Chicago

Backnik J M, Quinn C J Jr (eds.) 1994 *Situated Meaning: Inside and Outside in Japanese Self, Society, and Language*. Princeton University Press, Princeton, NJ

Baxandall M 1972 *Painting and Experience in Fifteenth-Century Italy*. Oxford University Press, Oxford, UK

Becker H S 1982 *Art Worlds*. University of California Press, Berkeley and Los Angeles

Cooley C H (1902) 1964 *Human Nature and the Social Order*. Schocken, New York

Crane D (ed.) 1994 *The Sociology of Culture: Emerging Theoretical Perspectives*. Blackwell, Oxford, UK and Cambridge, MA

Cruz J 1994 Testimonies and artifacts: elite appropriations of African-American music in the nineteenth century. In: Cruz J, Lewis J (eds.) *Viewing, Reading, Listening: Audiences and Cultural Reception*. Westview Press, Boulder, CO

Fine G A 1987 *With the Boys: Little League Baseball and Preadolescent Culture*. University of Chicago Press, Chicago

Fine G A 1992 The culture of production: aesthetic choices and constraints in culinary work. *American Journal of Sociology* **97**: 1268–94

Fukushima M (ed.) 1995 *Shintai no kochikugaku: Shakaiteki gakushukatei toshiteno shintaigiho* (*Structuring of Bodies: Body Techniques as Processes of Social Learning*). Hitsujishobo, Tokyo

Griswold W 1986 *Renaissance Revivals: City Comedy and Revenge Tragedy in the London Theatre 1576–1980*. University of Chicago Press, Chicago

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, New York

Goffman E 1974 *Frame Analysis: An Essay on the Organization of Experiences*. Harper and Row, New York

Hebdige D 1979 *Subculture: The Meaning of Style*. Methuen, London

Kondo D K 1990 *Crafting Selves: Power, Gender, and Discourses of Identity in a Japanese Workplace*. University of Chicago Press, Chicago

Lachmann R 1988 Graffiti as career and ideology. *American Journal of Sociology* **94**: 229–50

Lang G E, Lang K 1988 Recognition and renown: the survival of artistic reputation. *American Journal of Sociology* **94**: 79–109

Manning P 1977 *Police Work*. MIT Press, Cambridge, MA

Mead G H 1934 *Mind, Self, and Society: From the Standpoint of a Social Behaviorist*. University of Chicago Press, Chicago

Moulin R 1987 *The French Art Market: A Sociological View* [trans. Goldhammer A]. Rutgers University Press, New Brunswick and London

Morley D 1986 *Family Television: Cultural Power and Domestic Leisure*. Comedia, London
Neapolitan J 1986 Art, craft, and art/craft segments among craft media workers. *Work and Occupations* **13**: 203–16
Press A L 1991 *Women Watching Television: Gender, Class, and Generation in the American Television Experience*. University of Pennsylvania Press, Philadelphia, PA
Rogers M F 1991 *Novels, Novelists, and Readers: Toward a Phenomenological Sociology of Literature*. State University of New York Press, Albany, NY
Peterson R A 1994 Cultural studies through the production perspective: progress and prospects. In: Crane D (ed.) *The Sociology of Culture: Emerging Theoretical Perspectives*. Blackwell, Oxford and Cambridge, MA
Rockwell J 1983 *All American Music: Composition in the Late Twentieth Century*. Knopf, New York
Sato I 1991 *Kamikaze Biker: Parody and Anomy in Affluent Japan*. University of Chicago Press, Chicago
Simpton C R 1981 *Soho: The Artist in the City*. University of Chicago Press, Chicago
Snow D A, Anderson L 1993 *Down on Their Luck: A Study of Homeless Street People*. University of California Press, Berkeley and Los Angeles
Ukai M 1994 *Taishu engeki eno tabi: Nanjo Masaki no ichinen nikagetsu* [*A Journey to Popular Performance: A Year and Two Months of Nanjo Masaki*]. Miraisha, Tokyo
White H C, White C A 1965 *Canvases and Careers*. Wiley, New York

Y. Ogasawara

Expressive Forms and the Evolution of Consciousness

Expression, from the Latin *exprimere* (to push out, or to squeeze), will be used in this context to refer to behavior that appears to be motivated by the desire of organisms to enact salient internal states in public form; in other words, to express some important aspect of their being so as to make it accessible to oneself and to others. Taking a comparative perspective, one might see a continuity of expressive forms going back to such performances as the purring of a cat, the wagging of a dog's tail, or a peacock's display of tail feathers signaling its desirability as a healthy mate.

In human groups, expressive practices take on a variety of forms ranging from music and dance to fashions in food and in clothing, and they are essential components of any culture. It is true that, as Dante Alighieri wrote over six centuries ago in *De Monarchia* '…in every action…the main intention of the agent is to express his own image; thus it is that every doer, whenever he acts, enjoys (*delectatur*) the doing; because everything that is desires to be, and in action the doer unfolds his being, enjoyment naturally follows, for a thing desired always brings delight…Therefore nothing acts without making its self manifest' (Alighieri 1921/1317, *I*, 13, author's translation). Yet some actions are more expressive of the individuality of the actor than others. In fact most 'instrumental' activities such as work, childcare, and housework are overdetermined by bodily needs or social expectations, and thus afford little opportunity to bring out the unique characteristics of the individual. Thus, parallel to instrumental activities there arise 'expressive' forms, mostly in the leisure sphere of life, the purpose of which is primarily to make the self evident, and—as Dante observed—to provide enjoyment in the process. But this does not mean that such forms are arbitrary or fortuitous. Although expressivity is treated often as if it had no function and no adaptive purpose, it is more useful to begin exploring this phenomenon without making such an assumption, and to see what can be learned by starting instead with the premise that expressive forms, like other widespread human institutions, have been shaped by selective forces impinging on the human condition over time.

1. *The Origins of Expressive Forms*

Sometime between 500 and 1,000 centuries ago, after the brain had passed a threshold of complexity that allowed for speech to develop, our ancestors began to emancipate their mental life from the genetic instructions that had controlled their thoughts and emotions according to the script laid down by natural selection during the previous millennia. The mind emerged from the brain as people started to share their experiences, leading to a quantum jump in the information available to each individual.

Every living being needs to get reasonably accurate information about its environment in order to survive. In different species a vast number of sensory modalities have evolved to provide such a function, from incredibly sensitive olfactory organs to vision and to ecolocation. In addition, many species have developed systems of communication that allow information to be transmitted from one individual to another. Bees do it by wiggling their bodies in flight, ants by releasing chemical signals, and prairie dogs by whistling warning calls that identify sources of danger. This makes it possible for a perceptual datum observed by a single organism to be immediately available to the rest of the group, even when none of the other members have direct access to the source of information. The significance of sensory organs, and even more of communication, is that they allow individuals to interact with the environment vicariously, creating a space between themselves and potential dangers or opportunities. Eyesight makes it possible to detour around obstacles without bumping into them every time. Instead of getting ill before learning which berries are poisonous, one can avoid the pain by heeding a warning.

In his pioneering writings on 'evolutionary epistemology,' the psychologist Donald Campbell described the evolution of systems of knowledge, starting with sensory organs (Campbell 1974). As language developed, information was no longer dependent on the actual presence of a sensory stimulus. With the help of words, it became possible to evoke in the mind objects and events that had happened in the past, or that never happened in reality. Eventually, single events were fashioned into songs and stories: then myths emerge as the first knowledge systems. Their function was to justify a global attitude towards reality, connecting and explaining the otherwise fragmented everyday experiences, so as to allow individuals to act as if they understood what life was about. After myths, religions, political ideologies, philosophical systems, and theories have played the same role. According to Campbell, the latest step in this evolutionary process has been the emergence of the scientific method, which lifts vicarious experiencing to an entirely new level: In the laboratory, one can learn about reality without having to incurr undesirable consequences. If we make mistakes, theories can die in our stead.

As a result of these developments, the human mind has become increasingly shaped by data that were given meaning and that were transmitted by other human beings; in other words, by cultural information. It is useful to think of units of knowledge passed down through culture as *memes* (Dawkins 1976) to highlight several similarities they share with genes: memes control behavior, they are selected differentially and those that are selected are transmitted to the following generation. Of course, there are also glaring differences between genes and memes, yet focusing on the similarities is revealing and helpful in explaining human behavior.

1.1 The Emergence of Consciousness

It is likely that prior to the spread of memes our ancestors had no reason to experience their individuality as separate from the ongoing stream of events. Whatever they saw or felt prompted directly a visceral reaction dictated by the genetic program without the conscious mediation of a 'self.' The contents of the mind were inevitable and true. But as more and more of the information processed in the mind related to virtual as opposed to sensory data, it became possible to question the truth of one's knowledge—especially after the first 'urban revolutions' taking place around 6000 BC allowed people from different cultures to live together in proximity, thus challenging each others' assumptions about the world.

One consequence of doubt induced by the proliferation of memes has been the necessity for the brain to develop a mechanism for evaluating competing information, a faculty generally known as self-reflective consciousness. Having to decide which experiential datum to trust leads to the conclusion that there is *someone* who evaluates and decides; hence there develops the notion of a self as the arbiter of experience. From an evolutionary perspective, the Cartesian cogito could be rephrased as 'I wonder, therefore I am.'

The self, thus, separates individuals from the social and environmental matrix in which they had been embedded seamlessly. In reflecting on what this self is, one becomes aware of individual differences, due to the genetic and memetic uniqueness of each mind. Each person learns to identify with the peculiar configuration of traits inherited either through genes, or through the memes learned in the particular sociocultural milieu of early development. In cultures that share a 'Western' historical trajectory a differentiated individuality has taken an increasingly central importance; in those of an 'Eastern' tradition the integration of the person in the social context is seen as relatively more important. In every culture, however, the self is experienced as both differentiated and unique, and as integrated and related to others, even though the relative importance of these dimensions may be reversed from one culture to the next, depending on the overall pattern of adaptative responses (i.e., the economic, political, social, and cultural institutions that a given culture develops over time).

2. The Expression of Individuality

Reflecting on its own contents, consciousness beholds its uniqueness. But this subjective experience is not quite 'real' even to the experiencing person, and quite unsubstantial to anyone else. Hence the necessity to express this potential identity so that it may acquire a tangible reality. Uniqueness, however, is not the only trait of the self. It has been said that while in some respects each person is like no one else in the world, in other respects they are more like some people than others, and in still other respects they are like everyone else. Thus, expressive forms rarely embody only the actor's unique characteristics, but tend to express also their belonging to a community or human type.

If one considers the arts, for instance, which are among the most universal expressive forms, it is not accidental that each of them is the medium for one of the seven 'intelligences' that Gardner (1983) claims are supported by different neurological brain circuitry, and that are distributed unequally in the population. These include musical, linguistic, bodily-kinesthetic, visual, numerical, intrapersonal and interpersonal intelligences (plus possibly some other less well documented ones). It is easy to see the correspondence between the first four and the art-forms of music, literature, dance or athletics, and the visual arts. But also mathematics, introspection, and sociopolitical involvement, when reflecting the peculiar identity and

strengths of a person, can be seen as an expressive form in its own right.

Taking the case of music as an example, it is reasonably well documented that some children start life with above-average sensitivity to sound. For them it is easier to express individuality through making music. According to Pantev et al. (1998), compared with average people, 25 percent more of the auditory cortex of musicians is devoted to processing music, suggesting a more complex cortical discrimination and representation of sound patterns. By contrast young artists are able to represent mentally shapes and figural relationships much better than nonartists, excelling at such tasks as spatial visualization (Getzels and Csikszentmihalyi 1974).

Most artists discover their 'gift' accidentally early in life. All children draw and paint, but the sketches and cartoons of future artists attract more attention than those by other children. As a result, they draw even more. Slowly, this special trait begins to be seen by the child and by others as a central feature of the self. At that point causality becomes reciprocal: drawing builds the self, and the self requires drawing to express itself. There is no reason to doubt that other artistic careers follow the same pattern.

While musicians, artists, and athletes are among the few who make a career of their expressive skills, each person experiences the need to display their individuality, and does so on a less formal basis. Often, as is the case with cosmetics, hairstyle, and dress, what is being expressed is not any actual trait of the person, but rather an ideal image of the self that the person aspires to. Although some scholars, such as Goffman (1974), interpret these forms as nothing but attempts intended to *impress*an audience, there is no reason not to see in them also a need to *express*, in other words, to see them as genuine gestures essential to the person's psychological integrity.

Music is perhaps the most universal medium for externalizing one's subjective states. Especially in adolescence, playing and listening to music helps give form to the bewildering emotions that alternate so rapidly each day at that stage of life. In a study of urban American families, when asked which objects in the home were the most important, the majority of teenagers (46 percent) mentioned stereo sets—more than any other object—as opposed to 18 percent of their parents and only 6 percent of their grandparents. The following is a typical answer from a teenager as to why the stereo is so important: 'Because when I'm not real happy and gay, I turn it on and it makes me happy again…it helps me recover…from bad days' (Csikszentmihalyi and Rochberg-Halton 1981, p. 71).

A similar effect is achieved through some forms of writing, such as diaries and poetry. Words are not linked neurologically to feelings as music is, thus they are less concrete embodiments of emotion. They compensate for this by leaving a more permanent trace, which makes it easier to reflect and build cognitively on what one has written. Again as with music, poetry and other forms of subjective writing are practiced most widely in late adolescence, when identity formation becomes a salient issue (Erikson 1950). Persons become transparent to themselves in part by writing down what transpires inside.

'Lifestyle choices' also act as media for externalizing certain images of the self. Of these, we may mention four recurrent types: the heroic, the holy, the homey, and the hedonic. Together, they account for a large proportion of the forms people use to express their self through behavior patterns. The heroic lifestyle has a long tradition, going back to Homeric Greece and beyond. In our times it is manifested in sports, in high-risk activities such as climbing the highest peak on each continent in sequence, and less flamboyantly in workaholism or other feats of endurance. What may be called the holy lifestyle expresses itself in self-sacrifice, abnegation, helping others. These gestures of denial, if indeed a manifestation of the self, are presumably as pleasurable to the actor as they appear unpleasant to those watching from the outside. A homey lifestyle expresses a self bound to kin and hearth; it manifests itself through domesticity and what the French call *le petit bonheur*. Finally, the hedonic form is the one chosen by those who identify themselves with a sensate life, whether through their knowledge of fine vintages or through rowdy behavior at the neighborhood saloon.

Expressions of individuality are also achieved through the arrangement of artifacts in one's living space. The symbolic ecology of the home is almost evenly split between objects that signify the unique traits of the owner's self, and those that refer to its integration with the social context (Csikszentmihalyi and Rochberg-Halton 1981). Examples of the former are musical instruments, sports equipment, tools, cameras, camping gear, favorite books. By looking at them, using them, and showing them off, a person can literally objectify important dimensions of the self.

3. *Expressiveness and Social Integration*

If what makes a person unique ordinarily is invisible, so are those dimensions of their being which are shared with others. Because the relationship of the self to the social milieu is not self-evident, in every culture there have arisen behavioral practices that express in concrete form the ties that bind the individual to the community and to the natural environment, thus helping to reify the second essential dimension of what it means to be a person.

3.1 *Musical Expression*

Music, which is processed in the right auditory cortex of the brain (in contrast to listening to words, which

activates the left side) and simultaneously by the limbic system where emotions are produced (Carter 1998, p. 148), is one of the most ancient and powerful means for expressing subjective states. The reason for this is that certain combinations of sounds are believed to stimulate pleasure centers in the brain, providing experiences similar to those produced by 'a cocktail of recreational drugs' (Pinker 1997, p. 528). This neurological resonance between sounds and feelings has been exploited by every culture, generating an enormously rich variety of musical forms ranging from the beating of drums around the campfire to rock festivals and symphonic concerts.

Communal festivities generally are focused on music, singing, and dancing; it is by experiencing the harmony of one's voice joined to that of others that Durkheim (1965/1915) believed our ancestors reached that state of 'collective effervescence' which proved to them the existence of a collective entity greater than the individual, and hence laid down the belief in the supernatural which eventuated in religion. A typical example of music used to achieve this feeling of communion is Turnbull's description of how the Pygmies of the Congo play their *molimo* trumpets and sing in order to express a relationship with the forest they inhabit and which provides for all their needs: 'So when something goes wrong, like illness or bad hunting or death, it must be because the forest is sleeping and not looking after its children. So what do we do? We wake it up. We wake it up by singing to it, and we do this because we want it to awaken happy … when our world is going well then also we sing to the forest because we want to share our happiness' (Turnbull 1962, p. 92).

Other uses for music as a vehicle for expressing integration are work-songs which establish a common rhythm of physical activity and attentional focus among laborers. The historian William McNeil has explored in great detail the history of music in the service of warfare: from the great horns of ancient warriors to pipes and drums, sounds have always been used to intimidate the enemy or mark the step of marching armies.

One question raised by the ubiquitous availability of recorded sound is whether the emotional impact of music is going to decline with time. Moving through the day from music on the alarm-clock radio to the car, to the muzak in the elevator, to the tunes in the supermarket—it is likely that the mind will reach a point of saturation which will make it insensitive to beautifully produced sound.

3.2 Visual Forms

As is the case with music, the brain seems to have inbuilt preferences for certain patterns of visual stimuli. Gestalt psychologists and estheticians have examined what makes lines, shapes, colors, and forms enjoyable to look at (Arnheim 1969,Gombrich 1979). According to some neurologists, the primary function of the visual cortex is to construct consistencies in the visual field, with the aim of obtaining useful knowledge about the world. Responding to 'the brain's incessant demand for constancy' (Zeki 1997, p. 86), visual artists search for the most essential and lasting features of the environment, and represent them so that they can be shared. From earliest times, cultures have exploited the potentialities of visual data to communicate the individual's link with the community.

Among these are body-paintings, or carvings incised on a person's face or torso. According to Lévi-Strauss (1961), who studied some of the most isolated Amazonian tribes of South America, such tattoos and paintings locate the individual who wears them within the kinship network by providing the equivalent of a social map worn on the body, thus transforming the biological animal into a social person. Preliterate cultures have developed an enormous variety of practices equivalent in purpose, ranging from feathered head-dresses to masks, from Nuer spears to Maori canoes, from Hopi katchina dolls to the carvings on Melanesian homes. Through traditional visual symbols the person who owns such artifacts is linked to the community, to ancestors, and to the gods themselves.

Dress is used everywhere as favorite vehicle of expression. In traditional cultures, folk costumes can identify a person's native village as well as their status. Uniforms are an indication of the person's allegiance to a particular religious or ideological group. The Italian language neatly illustrates the way uniforms can express both the differentiation and integration of the self: the single noun 'uniform' can be translated into that language by two terms with opposite connotations: *divisa*, which indicates the fact that the clothes differentiate the person from most others; and *uniforme*, which like its English equivalent brings attention to the person belonging to a group. So 'the Boy Scout uniform' could be translated as *divisa* if one wanted to stress the difference between scouts and everyone else, or as *uniforme* to stress the appearance of conformity with other scouts. In reality it is almost impossible to imagine an expressive device that would only serve to differentiate a person; it would also point to the group that person belonged to. A teenager may want to show their individuality by piercing tongue, nose, nipple, and other body parts in a pattern never attempted before; nevertheless, their display would inevitably associate them with the subset of teens who go in for such body art.

Fashion in clothes also serves the same purposes as a uniform (or *divisa*). Men and women who keep track of trends in *haute couture* do so in order to make sure that their appearance will reflect belonging to the right social set, while setting themselves apart from the masses. Similarly teenagers in urban gangs signal their affiliation through their clothing; a baseball cap worn

with the visor forward could cost one's life in a neighborhood where the visor must be turned backward. Specialized symbols—such as flags, coats of arms, the sign of the cross worn on Crusaders' tunics, or swastikas on the SS's sleeves—also express group affiliation, and often imply a life-and-death commitment to a cause.

In the objects people choose for their homes, those that signify belonging to a kinship, ethnic, or elective group are among the most cherished. In a study, typical urban Americans, the largest number of objects (1,207) were mentioned as being special because they reminded the respondents of immediate family members (vs. 1031 objects referring to unique qualities of the self). Furniture, textiles, glassware, canvases, and statuettes that crowd the home tend to derive their value not so much because of their price-tags but because of the relationships they embody (Csikszentmihalyi and Rochberg-Halton 1981). The self of the owner is made manifest through a quilt sewn by a grandmother, a sofa where the children were weaned, a landscape purchased during the honeymoon, a Bible with the fading names of one's ancestors and their birth dates.

Visual expression achieves its highest form in the arts of architecture, sculpture, painting, and contemporary derivatives like photography, film, and videos. The status of 'art' is claimed by objects that provoke in the viewer experiences that are rarely felt in everyday life—such as awe, sorrow, surprise, mystery, and transcendence. In practice, what is to be considered art is decided by a cultural elite who control the public expression of information—Popes and prelates in the Renaissance, in our time the 10,000 aficionados of the 'painted world' in Manhattan of whom Tom Wolfe wrote. Nevertheless, objects that achieve the status of art tend to be those that help bring out in the viewer valued elements of the self that usually are unavailable to consciousness.

Taking advantage of the power of visual art to attract the attention of viewers, individuals and institutions have used its forms to broadcast messages intended to advance their special interests. For example in Western Europe the Church was able through the iconography of the Annunciation, Nativity, Holy Family, Crucifixion, Deposition, Resurrection, Last Judgment, and similar themes with universal appeal repeated in thousands of frescoes, paintings, and sculptures, to reify in visual form the tenets of its doctrine. Later sovereigns and wealthy merchants became adept at inserting their own propaganda in the great art of their times (Freedberg 1989). Currently advertising and marketing exploit the power of visual expression for frankly commercial ends.

3.3 Literary Forms

At first, extrasomatic recording of data, that is, the use of writing to complement word-of-mouth as a way of sharing information, was employed primarily to record taxes, laws, and other instrumental transactions. The first available inscriptions from China or Mesopotamia list bushels of grain, numbers of pigs, and the lapidary decrees of rulers. Eventually, however, all around the world people discover the potential of writing for expressing more subtle aspects of reality—deeds, desires, fears, and feelings—often attributed to gods, mythical beings, or imaginary heroes (Csikszentmihalyi 1990).

Like the other forms of expression, literature makes manifest both the individual uniqueness of the writer, as well as their affiliation with a particular culture. It has been argued, for instance, that the Sagas of Iceland were an essential means for the inhabitants of the frozen outposts to keep their identity and hope in the isolated wilderness (Johnson 1930). During the recently ended decades of Soviet occupation, the population of Eastern Europe similarly turned to poets and novelists to preserve its national identity and a vision of the future.

3.4 Bodily Expression

For people lacking written language and elaborate visual representation, the movements of the body (often accompanied by music) provide the richest form of expression. Dance until recently was primarily a ritual medium demonstrating virility, femininity, sexuality, hospitable or hostile attitudes sanctioned by the culture's world-view. On every continent ritual dances often lasting several days relied on complex choreography to remind the participants of the sacred mysteries of the culture. As with other art forms, during the last few centuries dance largely has left its ritual context behind, and developed instead its own esthetics. This resulted in much more varied and complex movements and choreographies, although arguably accompanied by an impoverishment in the substance being communicated.

The face is without doubt the part of the body most adapted to express inner states. Infants orient themselves automatically towards faces, and especially the eyes, which in many cultures are known as 'windows to the soul.' Except for persons suffering from prosopagnosia, most people can recognize basic emotions like anger, fear, or surprise from facial expressions. Apparently, Australian aborigines or New Guinea tribesmen have no trouble identifying these expressions from photographs of American people (Ekman et al. 1969). Of all forms of expression, the articulation of facial muscles is the most widespread. Practically each moment we externalize our subjective states, as well as providing a running commentary on relationships by showing empathy or rejection. Actors and mimes carry this universal ability into highly regarded art forms.

Athletic performances also have a very long tradition of serving expressive functions. In the Olympic

Games, the athlete's prowess was a highly admired sign of individual excellence, but even more it was an indication of the greatness of the community to which the victor belonged. In the Mayan ball games, teams competed for several days in a contest that echoed mythical battles among gods and often ended in the ceremonial slaying of the losing players. While professional sports still retain some elements of the sacred and the communal, as shown by the quasitribal antics of spectators at soccer and football games, more and more they express only the team owner's purchasing power.

4. The Evolutionary Adaptiveness of Expressive Forms

There is no question that the ability to express oneself and to read others' expressions is extremely useful as a means of both self-knowledge and communication. In the first place, the information thus gained helps individuals to achieve psychic integrity and facilitates interaction. Without means of exploring one's inner life and sharing it with others, life feels arid and constrained. According to the sociologist Schwartz (1987), a major cause of juvenile unrest in American suburbs is the lack of opportunities for teenagers who live constantly hemmed in by adult institutions to express themselves in forms appropriate to their needs.

It is also important to note that, as Dante intimated, the process of making the self manifest is profoundly enjoyable. Almost any action, as long as it is perceived by the actor to be the genuine expression of an inner reality, is experienced as worth doing for its own sake. Thus, the ubiquitous expressive forms that exist in every culture and define its character—arts, poetry, music, ritual, athletic events—are performed voluntarily, often without expectation of any reward, just because they are enjoyable to do.

However, according to a strictly biological perspective, a form must contribute to the inclusive fitness of the individual for it to qualify as being selected by evolution. Unless a behavior helps a person's genes to be reproduced in subsequent generations, there is no call to invoke the evolutionary paradigm. Do expressive forms meet this definition? Most scholars would give a negative answer. Pinker, for instance, adamantly states that the arts are not biologically adaptive (Pinker 1997 p. 526). In a dissenting opinion, Blakmore argues that actors, poets, dancers, athletes, singers, and other specialists in expression do indeed have a head start reproducing their genes. She suggests that as memes and genes began to coevolve, individuals who had a greater facility in using the emerging symbols became more successful and thus had a better chance attracting mates. Their offspring in turn would be more likely to inherit the symbolic facility, have better access to mates, and so on (Blackmore 1999).

However, the importance of expressive forms does not depend on whether they help some germ lines survive at the expense of others. As evolution has increasingly been about the transmission of memes, it is now possible to pass on songs, paintings, verses, and dances from one mind to another, without having to implicate genetic transmission in the process. This makes it possible for consciousness to be enriched beyond the constraints of its biological limitations, while at the same time opening up radically new opportunities for enjoying the expression of its own complexity.

See also: Consciousness and Sensation: Philosophical Aspects; Discourse and Identity; Expression and Interaction: Microfoundations of Expressive Forms; Expressive Forms as Generators, Transmitters, and Transformers of Social Power; Knowledge (Explicit and Implicit): Philosophical Aspects; Knowledge, Sociology of

Bibliography

Alighieri D 1921/1317 *De Monarchia*. Rostagno, Florence, Italy
Arnheim R 1969 *Visual Thinking*. University of California Press, Berkeley, CA
Blackmore S 1999 *The Meme Machine*. Oxford University Press, Oxford, UK
Campbell D T 1974 Evolutionary epistemology. In: Schlipp P A (ed.) *The Philosophy of Karl Popper*, Vol. 1. Open Court, La Salle, II pp. 413–63
Carter R 1998 *Mapping the Mind*. University of California Press, Berkeley, CA
Csikszentmihalyi M 1990 Literacy and intrinsic motivation. *Daedalus* **119**(2): 115–40
Csikszentmihalyi M, Rochberg-Halton E 1981 *The Meaning of Things*. Cambridge University Press, New York
Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford, UK
Durkheim E 1965 *The Elementary Forms of the Religious Life*. Free Press, New York
Ekman P, Sorensen E R, Friesen W V 1969 Pan-cultural elements in the facial expression of emotion. *Science* **164**: 86–8
Erikson E 1950 *Childhood and Society*. Norton, New York
Freedberg D 1989 *The Power of Images*. University of Chicago Press, Chicago
Gardner H 1983 *Frames of Mind: A Theory of Multiple Intelligences*. Basic Books, New York
Getzels J W, Csikszentmihalyi M 1976 *The Creative Vision*. Wiley, New York
Goffman E 1974 *Frame Analysis*. Harper & Row, New York
Gombrich H 1979 *The Sense of Order*. Cornell University Press, Ithaca, NY
Johnson S 1930 *Pioneers of Freedom: An Account of the Icelanders and the Icelandic Free State, 879–1262*. Stratford, Boston
Lévi-Strauss C 1963/1955 *Tristes Tropiques*. Atheneum, New York
Pantev C et al. 1998 Increased auditory cortical representation in musicians. *Nature* **392**(6678): 811–14
Pinker S 1997 *How the Mind Works*. Norton, New York

Schwartz G 1987 *Beyond Conformity and Rebellion*. University of Chicago Press, Chicago
Turnbull C 1962 *The Forest People: A Study of the Pygmies of the Congo*. Simon & Schuster, New York
Zeki S 1998 Art and the brain. *Daedalus*. **127**(2): 71–103

M. Csikszentmihalyi

Expressive Forms as Generators, Transmitters, and Transformers of Social Power

Social power, in its core meaning, combines three elements: it is intentional, relational, and causal. Someone is exerting power when s/he has a more or less explicit purpose, involving one or more persons, and succeeds at least partly in bringing about the intended result. Thus, power has two aspects: the social relation ('power over') and the means–ends relationship ('power to'). In practice these are closely connected.

Power is being exerted directly, by means of threats, orders, information, established consent. Or indirectly by influencing other actors' frames of action, definition of the situation, goals, identities (Lamont 1989). Power may also function solely based on norms and expectations surrounding leading positions, directing subordinates' inference of intentions of power holders, without their being explicitly announced.

A special type of power relations are conceived as 'hegemonies,' for example ideologies or visions of the world legitimizing one or a few dominant groups. Conversely, groups or classes aspiring for power not only depend on material resources, but must assume intellectual and cultural leadership to challenge dominant ideas (Gramsci 1971). Mostly conceived and diffused intentionally by adherents and ideological leaders, such visions are at the same time linked to deep, unacknowledged cultural assumptions. However, in order to achieve a pertinent social impact, 'unconscious' ideas must be brought to fore in organized action.

Generally, power is not based on physical force alone, but depends on communication, including the metalevel of 'talking about talking.' Thereby expressive forms assume a central place in the exercise of power. A threefold connection exists between power and communication. Being the basic social institution, language is a precondition for the establishment of other institutions, and thus of stable power relations (Searle 1995). Moreover, power presupposes complex decoding abilities in receivers. Alter must interpret and act on the signals emitted by ego, or the expectations implicit in his or her role. Thus, clarity and intersubjectivity are central virtues of expressive forms generating power. Finally, the power of expressive forms is connected to their rhetorical qualities in a broad sense. By virtue of a felicitous form, they have a seductive appeal to the eye and the ear, thereby shaping our perceptions of the world. Consequently, a constant tension exists between, on the one hand, the need for clarity in order for power to be exercised, and, on the other hand, the ambiguities inherent in messages with a strong appeal to listeners, readers and spectators.

Expressive forms (a) convey information, (b) produce labels and classifications, (c) depict nonexistent, possible worlds, (d) may be construed as rituals, (e) or broader lifestyles, and (f) form the base of generalized signal systems. In social relationships, these forms may be activated by (a) definitions of the situation, (b) persuasion by virtue of rhetorical efficiency, (c) emanation, whereby power of one instance is vested in another person or group, and (d) expression of collective identities.

Exercise of power on a wide scale through expressive forms presupposes not only use of symbols, but also control over the production of vocabularies and signs, as well as their material carriers, such as adornments, clothing, weapons, buildings. This implies the ability to influence the work of artists, artisans, architects, writers, professors, journalists, and editors who serve as producers and judges of aesthetic quality.

1. *Power in Small Groups and Intimate Relations*

In his descriptions of the strategies of everyday social interaction, Erving Goffman (1959), following up on the work of W. I. Thomas, points to the importance of communicative devices in establishing a definition of the situation. The person who succeeds in imposing his definition is accorded a privileged position in the ensuing interaction process. In its simplest form, definition of the situation implies evoking the relevant set of roles, and the rules accompanying them. Definitions may be extended from the role to characterization of the person filling it. Flattery, insult, irony, and double communication influence the self-image and the actions of the other person by creating confusion or diminishing control of the self. Rhetorical elegance is pivotal in producing these effects.

Gender relations highlight many of these issues. Generally, couples are characterized by reciprocal power relations: even if one of the parties is in a dominant position, the other is not powerless. The battle of the sexes is to a large extent fought by symbolic means aiming at undermining self-control or self-esteem in the other. Elaborate strategies are poetry—the beauty of words opening doors otherwise closed—and passionate calls for action, exemplified by Icelandic saga women egging their men to defend their honor. In addition to rhetorical strategies, symbolic power of women over men resides body images, shaped by jewelry, hairdressing, clothing, and manipulation of the face and the body itself. Symbolic

power of men over women is usually somewhat less focused, residing in signs of strength rather than beauty, but occasionally combined with insistent poetic appeals.

In the relationship between parents and children, a substantial part of the socialization process consists in the transmission of signs, linguistic, and symbolic structures. A close connection exists between the learning of rules and grammars of symbolic forms, and that of moral and social norms. Narratives, fairy tales, nursery rhymes, and lullabies enhance the communicative competence of the child, while at the same time they serve as vehicles for learning of social expectations. Here, the complex effects of power in parents and other guardians come to the fore, constituting subjects in a double sense—as dependent group members and autonomous persons.

In families, as well as work groups expressive leadership is of prime importance for integrating common tasks, or easing the burden of common hardships. Power in the form of expressive leadership is assumed by the use of family myths, proverbs, slogans, and work songs.

2. *Public Power: From Decoration to Constitution of Legitimacy*

In larger groups, the simplest relationship between power and aesthetics is that of decoration or status marker. Powerful individuals and groups demonstrate their power by symbols and lavish embellishments setting them apart from the lives of ordinary people. This is seen not least in feudal and autocratic regimes. The decorative function should not be viewed as an epiphenomenon. The aesthetic expressivity of the powerful calls forth fascination among subordinates, which in itself reinforces their power. Further reinforcement emerges from the legitimizing effects of exquisite clothing, music, well-modulated speech, and impressive buildings. Hardly any person can enter the Louvre or Taj Mahal without being struck by awe and admiration. Beauty and solemnity emanate on those who are in command of it, and convey that they deserve their powerful status. These mechanisms are aptly summarized in Nietzsche's 'aristocratic equation': 'Good = noble = powerful = beautiful = happy = loved by God' (Nietzsche 1992, Sect. 7).

Symbolic messages carry information about specific rights and duties in social arenas, drawing the border between public and private territory or between profane and sacred spheres. The ability to manipulate symbols designating types of arenas is a notable source of power, because they to a large extent determine what is acknowledged as acceptable roles and modes of action. Related effects are found in palaces, parliamentary buildings, courthouses, or churches. It would be simplistic to assume that aesthetics mainly function via direct effects on behavior, for example, that a given architectural design of the courthouse yields a higher number of confessions. The main effect of design is the reinforcement of a definition of the situation already established, concentrating attention on the relevant processes taking place in a given arena. Uniforms used by public officials, soldiers, police forces have similar effects. In early modern societies, military uniforms were given great aesthetic value, obviously to generate admiration and deference in spectators. Spectacular shows of military drill accompanied by music and drumming (McNeill 1995) indicated an irresistible force springing out of the groups' readiness to act in concert, under one will, in a virtually mechanical way. Modern uniforms have fewer aesthetic features, but make stronger reference to the bearer's membership in a bureaucratized organization, reflecting the duty to perform certain tasks on behalf of the authorities. Even if uniforms are not threatening in themselves, they demonstrate that behind the person wearing it stands a much larger group, whose resources may be mobilized if necessary.

Uniforms usually signify both the larger group ('Spanish cavalry,' 'US Air Force') and the rank within the group. Symbols of rank, such as crown regalia, may also be confined to unique roles or persons. French kings in the seventeenth and eighteenth centuries exercised monopoly not only over official signs of power, but also over informal objects such as clothing, furniture, and the use of special colors. The king dictated the style of the court, while at the same time his own possessions were to be so exquisite that nobody could compete with them, by virtue of being prohibitively expensive, or due to regulation of the production through royal monopoly (Auslander 1996). This specificity of aesthetic objects put the king in a position where he could not be rivaled in any respect, not only in what concerned his formal power over the citizens. Poetry celebrating the great deeds of power holders have a similar impact.

Aesthetic monopoly has a direct connection to the wider legitimacy of power. In addition to the purely aesthetic effects, Max Weber (1968) assumes that legitimacy is enhanced through the construction of social myths and narratives, intended to justify the position of the rulers. In such diverse societies as the Kwakiutl, Aztec, and Nazi Germany, ascending groups managed to reinterpret and reconstitute existing cosmologies to make them fit their own needs (Wolf 1999). In a similar vein Pierre Bourdieu (1991) conceives symbolic power as the ability to define what is common sense, and thus affect other actors' knowledge and perceptions of the world. Roland Barthes (1972) points to a source of naturalization of everyday mythology in the conflation of denotation and connotation.

Establishing legitimacy on a broader basis requires elaborate symbolic systems based on delegation, mainly in one of two forms. Persons in powerful roles

may delegate their competencies to third parties, who are entrusted to perform their duties on behalf of the power holder, be they tax collectors, police forces, or royal academies. Alternatively, delegation in a wider sense is connected to symbolic systems, such as legal codes or money systems. In the case of money, the power holder issues currency to be used by the population. Delegation takes the form of a guarantee from the central power that the money circulating is genuine, and that it will be accepted by other actors within its jurisdiction. These mechanisms of delegation are dependent on the recognition and eventual legitimacy of the central power holder, be it a king or a more anonymous state apparatus. They cannot be operative without writing (Goody 1987) and derived symbolic systems: heraldic insignia, tax forms, badges of identification, stamps, uniforms, money bills, that by virtue of their form convey to the public the unambiguous message that they represent the power holder and are guaranteed by it. Such mechanisms are most widespread in modern societies, but they hold an important position in highly developed premodern societies as well; for example in late Aztec culture, the design of codes regulating clothing and the production of status symbols was a privilege of the throne, directly regulated by the ruler Motecuzoma (Wolf 1999).

More diffuse, but no less important to modern states than the explicit delegation of power, is the national culture surrounding it. A modern society can hardly function without a channel of legitimacy imprinting a collective identity on its population. This may be seen as a more elaborate version of the myths of legitimation surrounding the rulers. Basis for national cultures are foundational myths and folk art, refined and shaped into tales conveying the dominating self-images of the people: how 'we' became 'us' through conquest, acts of liberation, adoption of a constitution. This process is expanded into national historiography and cultural traditions within art and literature (Hohendahl 1989), school textbooks, as well as the national and popular press.

3. *Social Groups and Classes—Identity Formation Through Inclusion and Exclusion*

In his theory of stratification, Weber (1968) distinguished between classes and status groups. Classes crystallize around common economic interests and similarities of life chances, whereas status groups are based on common lifestyles. Expressive forms serve status group interests by being part of mechanisms of inclusion and exclusion. The more groups compete over status or resources, the greater tendency to group closure. A related point was made by Émile Durkheim (1961), who in his sociology of religion underlined the impact of naming and labeling, and thus of classification of social states or persons, into sacred and profane, groups of worthy and unworthy.

The definition of groups takes place in two steps; drawing boundaries (Lamont 1989) and creating collective identities. The most elementary case is the use of signs or uniforms to signal group membership. In addition to the possible effects on third parties, the uniform permits to distinguish between allies and enemies, and thus enhances the possibilities for rational operation in the battle group. This is true of team sports as well as military operations. In groups without uniforms, standards, flags, and the rhythmic shouting of slogans serve as rallying points, vividly described for example, in Snorri Sturlusson's narration of the last battle of Saint Olaf. Albeit more silently, public monuments commemorating kings, poets, or local notabilities contribute to the constitution of a collective identity by representing a common heritage or a common cause.

Dress codes are widespread signals of group membership, also in societies without official prescriptions on clothing. Even if less visible to the public, tastes in home decoration function as status markers for owners and their peers who are invited in. The same is true of tastes in art, music, literature, as well as vacations and general lifestyles (Bourdieu 1984). Through networks of communication tastes are diffused into the public sphere, visualized through furniture catalogs, record reviews, press coverage of art auctions. As a parallel to the judgment of artistic achievements, specific judges of taste acquire a visible role, by laying down what constitutes acceptable style. Manuals on courtly etiquette are known as far back as the twelfth century, and from the seventeenth century special magazines devoted to the 'right' style made their appearance (Auslander 1996).

Rhetorical elements and the fascination they create in listeners and spectators, have a pivotal role in the construction and diffusion of collective identities. This is true for eminent political orators from Demosthenes to Montgomery (1983), as well as charismatic religious leaders. Related mechanisms are found in modern advertising. Even if it is foremost aimed at selling specific products, very often it does so by selling lifestyles, carefully tailored to a variety of target groups. The ambiguity of advertising—simultaneously communicating puffery and fact—has developed into complex aesthetics, where irony and metacommunication play an increasingly significant role (Parmentier 1994). Creating semiautonomous semiotic worlds, the ad for the product becomes part of the product itself, with the power of lifting mediocre products into the world of exquisite taste.

Besides narrating fascinating fictional stories, popular literature, film, and TV soaps transmit interpretations of the real world—what are viable values and central concerns in life, how are conflicts solved, what are significant alternatives to the present state of the world, to mention only a few. To a large extent they are also edifying stories, drawing boundaries between the 'ins' and the 'outs,' and clarifying what is required

to be among the ins. Albeit on an artistically lower level, in this they are not very different from the great works of the nineteenth century, be they by Austen, Balzac, or Dickens.

Collective identities are also reinforced by exclusion. Exclusionary devices are found is definitions of the situation and in ways of talking and classifying people and events. High walls around the estates in upper-class living areas articulate a clear message: 'Stay where you belong.' Alternatively, exclusion may take place through metaphorical characteristics and 'labeling' of persons, whether of a positive or negative type.

Moreover, exclusion is enacted by the interpretation of consumption patterns. The classic example is that of conspicuous consumption (Veblen 1899), underlining the distinction between 'us' and 'them,' and making 'them' feel inferior due to their modest consumption pattern. Dress codes is one of the most widespread phenomena of this type. On a more general level, the creation of separate class styles by representatives of the dominating culture may subdue aspirations of mobility in the lower classes (Auslander 1996).

Exclusion may also be triggered by self-selection due to low familiarity with social codes. One effect of shifting fashions is the weakening of self-confidence among those who stand outside the group of consumption pioneers, and are unable to appreciate the new codes of elegance. The same point has been made on high culture and tastes for modern art in general. Some forms of cultural consumption (abstract painting etc.) evoke a feeling of inferiority in most people, because they presuppose aesthetic competencies not possessed by the general public (Bourdieu 1984).

4. *Institutionalization of Expressive Forms*

The exercise of power being dependent on communication, power holders have an interest in codifying systems of meaning as clearly as possible. The need for institutionalization of expressive forms becomes the more pressing in modern, populous societies.

Language is the most obvious case; some dialects and sociolects are regularly elevated to the official norm. National languages are further refined by writers, academies, and professional linguists. In the wake of Dante's *Divine Comedy*, the Tuscan dialect became the dominant variant of the Italian language. In France, the sociolect of the court in Paris was elevated to the national norm and further refined by the Académie Française from the seventeenth century. On the popular level, conscription for military service was among the most important sources of linguistic homogeneity in Europe. The normalization of one dialect/sociolect not only creates national homogeneity, but also enhances the cultural dominance of those groups that come close to this linguistic form as their native language (Bourdieu 1991).

Cultural dominance at the level of linguistic pragmatics is explored by Michel Foucault (1984) under the conceptual heading of 'discourse.' The discourse in this meaning is a broad 'way of talking' structured by several sets of factors: certain themes are proscribed (desire, power, madness, truth). The discourse is unified through assumptions about canonical texts and authorship, while at the same time rituals, doctrines, and discourse communities limit the number of relevant speakers.

In his theory of social institutions John Searle (1995) describes how vast social systems of power resources, such as the institution of money, are based on symbolization analogous to those found in speech acts. Performative speech acts, constituting social facts solely by expressing them ('Hereby I declare the Parliamentary session opened,' 'I declare you man and wife') are essential to the workings of social institutions. Even if the act is purely linguistic, the result is the creation of a binding institutional fact. These speech acts take on a highly formalized and expressive form. One source of formalization is the fact that they are repeated again and again by the speakers—when Mary and John marry they are neither the first nor the last couple to appear before the magistrate. The speech acts also have to be recognized as such by hearers; the couple, along with third parties, need to know that they were properly married when the ceremony was taking place.

Expressive forms are even more crucial in religious rituals. The ritual as an elaborate speech act (Tambiah 1985) establishes a connection to an otherworldly instance, by callings to God or the gods, material or symbolic sacrifices, and reenactment of earlier contacts with the deity. Rituals institute and make visible the divine power over life. Their effects are closely related to the strictness of the expressive form, both as concerns the effects on the deity and on the congregation of believers. Moreover, the enactment of the ritual confers power on worldly persons or instances. It emanates to the priest, who establishes the connection to the otherworldly realm, and onto the church or other religious organization of which he is member. They acquire a high degree of control over the lives of members—as well as an income from their pecuniary contributions.

In modern and protomodern societies, secular knowledge is produced and diffused by specific institutions having a pivotal role in the socialization of the population. Elementary training in schools to a large extent consists of indoctrination of recognized cultural achievements, mostly with greatest weight on literature, but also in other arts. At the higher level, universities produce judges of taste: editors, curators, critics, interpreters, who function as mediators between the direct producers and their public. These intellectuals contribute to the canonization of established works, while at the same time recording and judging new, emerging forms. Thus, schools and

universities are central in the production of the established aesthetic sensibility and the diffusion of works deemed socially valuable, at the cost of other cultural patterns.

Moreover, governance relies on the production and distribution of knowledge, not only that of the natural sciences, but also the humanities and behavioral sciences (Foucault 1991). Textbooks, manuals, and discursive practices build up a scientific mode of reasoning in the population, giving access not only to useful knowledge, but to a more general rational and 'disciplined' way of thinking, which is a precondition for mastering the complexities of modern societies. At the same time, the exercise of power is to an increasing degree dependent on support from the public mediated by the press, in which images produced by the media have an essential role. The transmission of news becomes globalized, television setting agendas on an ever larger scale. The Gulf War in 1991 was the first full-blown media war, with worldwide TV audiences as crucial participants in the events. It made Jean Baudrillard ask whether the war was actually fought out in the media only. Even if the question is dismissed as purely rhetorical, it demonstrates undeniably that the border between image and reality is becoming blurred.

5. *Hegemonies and Cultural Constellations*

As pointed out in the introduction, hegemony and related macrophenomena should primarily be treated as intentional. However, cases where nonintentional structures acquire a position similar to that of power deserve discussion. In relations between minorities and majorities, even when no discriminatory behavior takes place, forms of pseudodiscrimination may occur. A minority group with a separate language is forced to take up the majority language, in order to fully participate in society. The same may be true if groups using the same language have distinctly different codes of interpretation, as is often the case with social class, gender, or ethnic groups. Likewise, the absence of cultural resources and role models may contribute to keeping one group in an inferior position, even if it is intended by nobody. Finally, a relatively powerless minority may be experienced as threatening, and thus influence the majority to exaggerate the need for precautionary steps to avoid confrontations that on the average are rather insignificant.

More intriguing is the position of the mass media. Some of its most important social consequences are aggregate effects that only partially may be characterized as intended results, while at the same time being very pertinent to the functioning of modern societies. Take first the press and the news media. Their basic inclination is simply to transmit news. But the transmission is necessarily selective. In the enormous diversity of social events taking place every day, only few events and facts are singled out as being relevant to the social agenda. This depends on a combination of expected sales figures and professional criteria. Important aggregate effects emerge as journalists normally rely on each other when selecting the 'news that's fit to print.' Thereby, reciprocal expectations in the media sector influence not only the level of information, but the general pattern of tastes and social norms as well.

To some extent this is also true for the culture industry. The seductive character of popular culture has been attacked on the assumption that its power of fascination gives the mass media a grip on the mind, making them an 'opium for the people.' In postwar sociology, this point was made most energetically by the Frankfurt School. Even if the immediate goal of the culture industry is not the exercise of power as such, by directing peoples' attention away from basic social conflicts and into unrealistic daydreaming, the culture industry is assumed to reinforce the legitimacy of established power groups and thus pave the way for a quasitotalitarian regime (Horkheimer and Adorno 1972).

6. *Questioning Power: Resistance and Social Change*

A core assumption in the criticism of the culture industry is that resistance has been appropriated by power holders. But power from above is invariably met by resistance from below, even if their forms may differ. Processes of legitimization of power are never perfect; hence, countercultures have a tendency to appear in all societies.

In totalitarian societies resistance is mainly covert. In Eastern Europe under communist rule productions of classical plays could take on a special significance if they contained hidden references to the forces in power. Jokes about power holders were a permanent part of everyday life. Under other oppressive regimes, music has been one of the privileged forms of resistance, a prominent example being black culture in America. Even if they change nothing in the day-to-day operations of power, these symbols make it easier to breathe, and contribute to the maintenance of collective identities.

A specific neutralization of power structures is institutionalized in carnivals and feasts where everyday power relations temporarily are bracketed and turned upside down. The powerful are relegated to common positions, while the 'queens' and 'kings' of the feast are selected among ordinary participants. In literature, the archetypal version is the comedies of Aristophanes, ridiculing the brave citizens of Athens. From the mid-ninenteenth century, criticism of bourgeois society became a regular feature of the arts in the Western world.

More direct challenges to existing power structures come from religious and social protest movements.

Religious conviction may serve as a source of obstinate resistance to the secular power of the prince or the bureaucratic state. This was a central element in the founding of the United States as a nation. Social movements develop their own aims and their own signs of collective identities. One important source of their mode of operation lies in the openness of symbols and concepts used by the established power, which permits their meaning to be constantly adjusted and reinterpreted (Snow et al. 1986). A telling example is the slogan of '*Liberté, égalité, fraternité*' from the French revolution. When first put to use, the concept of political equality was reserved for a small fraction of the population—mature men with a certain property. During the subsequent two centuries the concept was broadened, first to include larger groups of property owners, then all men regardless of their property, then also women, and later citizens of lower age, down to 18 years and in some countries even down to 16 years.

Art and literature, moreover, are used by social movements in order to strengthen collective identities and enhance their own power of action, while at the same time envisaging the long and difficult struggle and depicting future goals. The European labor movement put great effort into creating a wholly separate cultural domain for its members, with their own literature, publishing houses, theaters, and sports clubs. Regional movements typically make use of literature from 'their' region (Draper 1989). Likewise, to the modern feminist movement it has been of prime importance to reissue, upgrade and shed new light on female writers and artists of the past.

More diffuse countermovements and subcultures, such as youth cultures, rely heavily on expressive forms to challenge established authority systems, while stressing the here and now rather than long term political goals. Instead of functioning as a vehicle, expressive form becomes a goal in itself. Specific dress codes, musical styles, and linguistic expressions constitute the core of the protest. Viewed from the social establishment, the hippie or punk styles are felt as extreme, and that is exactly the point. In this respect, protest culture has common features with *avant garde* art, aiming at shocking the general public, by defying mainstream aesthetics, and thereby the accepted lifestyles. This has been analyzed most extensively in protest movements among British working-class youth, not least directed against the culture and discipline of the class-based school system (Hall and Jefferson 1976). But youth protest is not limited to issues of class, as the phenomenon is ubiquitous in the modern world.

An intriguing dialectic exists between these protest cultures, the media, and commercial interests, often making them victims of their own success. In their search for 'news,' the media quickly pick up new trends and aesthetics, thus amplifying the voice of protest. Reaching a larger audience, the protest gets mixed up with commercial interests, such as the record industry and the fashion industry. Protest against power becomes appropriated by powerful instances, gets watered out, and in diluted form the aesthetics of protest may even be taken over by well-established consumer groups.

During the 1990s the Internet developed as a new channel for popular protest. By enhancing the possibilities to reach a critical mass for small and scattered groups, identities and abilities of collective action are reinforced. International alliances among aboriginal peoples is an example close at hand, extremist groups barred from operating in the open is another. More generally, the Internet presents unprecedented opportunities for putting pressures on individuals and groups by the distribution of petitions, spread of information and rumors. Small groups with few direct resources are the ones most likely to build out their power in cyberspace.

7. *Challenges to Further Research*

The most significant recent development in the study of power in the last decades stems from a growing awareness of the pivotal position of communication. The work of Foucault and Bourdieu probably has been the most influential in this respect. Despite being anchored in social action, the communicative impulses in their theories are to a surprising degree marked by a structuralist bias, albeit in different versions. In Bourdieu's conception, internalization is excessively strong; actors being guided by a *habitus* which hardly leaves any room for free choice. In Foucault the opposite problem is dominating: individuals are seen as disciplined by external forces, which taken together create a virtually impenetrable social structure.

These biases reflect two still unresolved problems in social theory. The first is connected to the understanding of values, preferences, and social autonomy in individuals. To be efficient, socialization presupposes the long-term exercise of power over individuals. But how can this process result in individual autonomy and a capacity for moral action? Second, if social structure is simply seen as the outcome of myriad social actions over long time, how is it conceivable that societies may function as relatively stable environments? These questions are haunting social theory in general, and they become even more acute when focusing on the significance of communication, and thereby expressive forms, for the exercise of power.

See also: Collective Beliefs: Sociological Explanation; Collective Identity and Expressive Forms; Commemorative Objects; Counterculture; Durkheim, Emile (1858–1917); Foucault, Michel (1926–84); Gender and Feminist Studies in Anthropology; Gender, Feminism, and Sexuality in Archaeological

Studies; Gender History; Gender Ideology: Cross-cultural Aspects; Hegemony: Cultural; Male Dominance; Matrifocality; Myths and Symbols: Organizational; Poverty: Measurement and Analysis; Power in Society; Ritual; Ritual and Symbolism, Archaeology of; Symbolic Boundaries: Overview; Symbolism (Religious) and Icon; Weber, Max (1864–1920)

Bibliography

Auslander L 1996 *Taste and Power. Furnishing Modern France*. University of California Press, Berkeley, CA

Barthes R 1972 [1952] *Mythologies*. Hill and Wang, New York

Bourdieu P 1984 [1979] *Distinction*. Cambridge University Press, Cambridge, UK

Bourdieu P 1991 *Language and Symbolic Power*. Polity Press, Cambridge, UK

Draper R (ed.) 1989 *The Literature of Region and Nation*. Macmillan, London

Durkheim E 1961 [1912] *The Elementary Forms of Religious Life*. Collier, New York

Foucault M 1984 [1970] The order of discourse. In: Shapiro M (ed.) *Language and Politics*. New York University Press, New York

Foucault M 1991 Governmentality. In: Burchell G, Gordeon C, Miller P (eds.) *The Foucault Effect*. Harvester Wheatsheaf, London

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY

Goody J 1987 *The Interface Between the Written and the Oral*. Cambridge University Press, Cambridge, UK

Gramsci A 1971 *Selections From the Prison Notebooks*. Lawrence & Wishart, London

Hall S, Jefferson T (eds.) 1976 *Resistance Through Rituals*. Hutchinson, London

Hohendahl P U 1989 [1985] *Building a National Literature. The Case of Germany 1830–1870*. Cornell University Press, Ithaca, NY

Horkheimer M, Adorno T W 1972 [1947] *Dialectics of Enlightenment*. Herder and Herder, New York

Lamont M 1989 The power culture link in a comparative perspective. *Comparative Social Research* **11**: 131–50

McNeill W H 1995 *Keeping Together in Time: Dance and Drill in Human History*. Harvard University Press, Cambridge, MA

Montgomery H 1983 *The Way to Chaeronea. Foreign Policy, Decision Making and Political Influence in Demosthenes' Speeches*. Universitetsforlaget, Bergen, Norway

Nietzsche F 1992 [1887] *Genealogy of morals*. In: Kaufmann W (ed.) *Basic Writings of Nietzsche*. Modern Library, New York

Parmentier R J 1994 *Signs in Society. Studies in Semiotic Anthropology*. Indiana University Press, Bloomington, IN

Searle J R 1995 *The Construction of Social Reality*. Free Press, New York

Snow D, Rochford E B, Worden S K, Benford R D 1986 Frame alignment process, micromobilization and movement participation. *American Sociological Review* **51**: 464–81

Tambiah S J 1985 *Culture, Thought, and Social Action*. Harvard University Press, Cambridge, MA

Veblen T 1899 *The Theory of the Leisure Class*. Macmillan, New York

Weber M 1968 [1922] *The Theory of Economic and Social Organization*. University of California Press, Berkeley, CA

Wolf E R 1999 *Envisioning Power*. University of California Press, Berkeley, CA

F. Engelstad

External Memory, Psychology of

Memory was originally defined as the capability of humans and animals to retain past experiences (e.g., childhood events). When neurophysiologists discovered areas in the brain which enable recall and recognition, these biological structures were also designated as memory (e.g., the auditory memory in the cortex). The rise of computer technology has led to the development of electronic devices for information storage and retrieval which were likewise called memories (e.g., random access memory, RAM). Moreover, the concept of memory has been generalized to all social and artificial systems serving the purpose of information storage and retrieval: artifacts ranging from Mesopotamian cuneiform scripts and Indian quipus to printed volumes and magnetic tapes, other persons such as secretaries or prompters, as well as souvenirs and relics such as mummies, i.e., ingeniously preserved corpses.

Information stores should be distinguished from memory aids. The former represent meanings in specific codes (e.g., phonological language, two-dimensional drawings); they distinguish meanings by modifying the encoded representations. Some representations are analogous, as in sculptures and photographs whose shapes, colors, etc., vary to permit the identification of different persons, places, or things. Other representations are symbolic, as in writing, where words, digits, etc., describe different events, addresses, and the like. In contrast, memory aids remind one of persons, events, etc., because an association with these components has been established. An example of a memory aid is a knot in a handkerchief. A person may have associated the intention to take his or her medicine with this knot. Upon seeing the handkerchief, the knot will remind the person of this intention. However, any other intention (e.g., mailing a letter) may equally become associated with the same knot. Therefore, memory aids need not change to discriminate between specific meanings.

Some external memories are unique constructions of individuals, whereas others are shared by any number of persons. From a sociological point of view, Halbwachs (1992) has interpreted the world of memorials and documents, of relics and contemporary

witnesses as one collective memory, which emerges from and constitutes specific cultures. From a philosophical point of view, prints, tapes, and other artificial media can be regarded as manifestations of an exterior mind beyond individual cognition. Thus, the topics (e.g., biographies, geometry) and formats (e.g., genealogical trees, numeric systems) of these media can be analyzed as contents of the exterior mind (Koch and Krämer 1997). From a psychological point of view, external media extend and support the internal memory. Individuals may utilize extended memory systems comprising internal and external components to augment their knowledge and to increase the efficacy of their actions (Schönpflug and Esser 1995).

The comparison of external and internal memories yields intriguing insights. For this purpose, it is useful to examine their functional features. When external memories supplement the internal memory, the control within the arising extended memory system becomes a crucial issue. The functional features of memories and the issue of control in extended memory systems will be considered in turn.

1. *External Memories vs. Internal Memory: Functional Features*

Noteworthy in the comparison of external and internal stores are the following six features: capacity, permanence, speed of encoding, speed of retrieval, usability, and availability. Generally, effectiveness and efficiency have to be rated separately. Effectiveness relates to maximal performance and is predominantly defined by the criteria of capacity, permanence, and access speed. Efficiency pertains to the tradeoff between demands and expenses, and is also determined by usability and availability. Depending on task characteristics and personal aspirations, people may refrain from maximal performance and decide to save their efforts.

Technically advanced information stores have reached high levels of effectiveness. In 2000, for example, a hard disk the size of a coin may store 35 gigabits of information, while providing access within a few milliseconds (for overviews on electronic storage devices, see Garcia-Moina et al. (1999), http://www.storage.ibm.com). However, technically advanced and highly organized information stores (e.g., public libraries, compact disks) may be difficult to use (owing to technical requirements, administrative regulations, etc.). They may even become unavailable for a user (because they are out of reach, temporarily closed, etc.).

Technically less sophisticated information stores may be inferior to more advanced devices with respect to their effectiveness. In low-performance situations, however, simple external stores are preferred because of their efficiency. An example is personal notes (Harris 1980). Notes on small sheets of paper are not very effective; they must be brief, may easily become illegible, or may get lost. Nevertheless, in some tasks a written note may suffice (e.g., for telephone numbers to be dialed immediately and only once). Compared with a hard disk, a small piece of paper may be more readily available. Furthermore, it may be more usable as people may feel more comfortable writing on and reading from paper than accessing and retrieving information from an electronic device.

Similarly, some external memories surpass internal memories in some respects, but in many situations the latter may be superior. The internal memory may possess considerable capacity and permanence. It enables a person to learn several languages, to acquire orientation in numerous towns, etc., and makes the lifelong retention of this information possible. However, personal memory is limited in capacity and permanence, especially for some disabled or elderly individuals. Distinctive features of personal memory are high usability and availability. With regard to encoding and retrieval, personal memory often functions automatically and within parts of a second (e.g., retaining a word from a conversation, recalling a name, etc.), but on other occasions people may need considerable time and effort for memorizing and remembering information. Moreover, one's personal memory is constantly at one's disposal without further provisions—at least in times of wakefulness and mental health.

It is debatable whether all human knowledge is fit for external storage. As stated above, external storage requires verbal, pictorial, or other forms of representation. However, there seem to exist profound insights, intuitions, and skills which cannot be stated explicitly. For instance, explanations of art may miss substantial aspects from the experience of the artist. It is equally questionable whether grammar books convey the full expertise necessary for the consummate command of a language.

2. *Control in the Interaction Between Internal and External Memory*

Typically, persons store and retrieve information pertaining to their current concerns, i.e., their needs, interests, and ongoing activities. They monitor their concerns and their actual state of knowledge, and then employ storage devices to supplement their knowledge in such a way that it serves their current concerns. This control process is applied to both internal and external memories (e.g., when memorizing or copying documents for a forthcoming lawsuit).

In an extended memory system, control may be transferred to other agents. Thus, other persons or technical systems (e.g., clerks, surveillance systems) initiate information storage and retrieval. The external agents monitor current concerns and actual know-

ledge, and they make priority decisions about the supplementary information to be stored and retrieved. In order for such external control to be satisfactory for a person, his or her concerns and knowledge must be taken into consideration. For instance, an assistant to a managerial board should be aware of the company's business and of what the board members already know. He or she should provide supplementary materials that support the position of the board. In the implementation of large and complex external stores (such as database systems and encyclopedias), it becomes a crucial issue to monitor and consider carefully the knowledge and the current concerns of users. Otherwise, entries may be compiled which lie outside the user's comprehension and interests, or may even lead to misinformation and frustration. Similar dissatisfaction may result from maladapted search routines.

Another aspect of control in extended memory systems is management, i.e., the efficient distribution of information over internal and external components, followed by an efficient search and retrieval. Management presupposes the adequate representation of internal and external components and their contents. Given the simple case of a single person with a written note, the person should remember the location and the general content of the note to be able to use it when needed (e.g., a shopping list in the supermarket). With regard to more complex systems, users must memorize procedures for registration and access, routines for storage and retrieval, lists of files, etc., or else should learn what external information is needed to operate a storage device, and where and how it can be obtained.

Moreover, source knowledge should be distinguished from target knowledge. Target knowledge comprises all information related to primary concerns (e.g., biographical events, price lists). Source knowledge refers to information that is instrumental for preserving and reconstructing target knowledge (e.g., the location and content of an album, mailing addresses). Although some people exhibit a primary interest in sources of information, including their operating characteristics (e.g., archivists, Internet experts), standard users acquire source knowledge only for the purpose of improving the retention of target knowledge.

When the acquisition of source knowledge is required for preserving target knowledge, it is of paramount importance to economize the cognitive load on the internal memory. While transfer of target information to an external store reduces the load on internal memory, utilizing the external store imposes the load of retaining source information on the internal memory. How do the benefits of saving some load by transferring target information to an external store compare with the costs of increasing the load resulting from source information? It may be uneconomical when persons memorize and store the same information externally. On the other hand, people may refrain from external storage altogether in order to obviate the load from source information.

Another strategy is to retain minimal information internally, and to transfer most information to an external store. This can be done by condensation, as in schemata of objects and macrostructures of coherent texts. Schemata are cognitive structures implying slots to be filled (e.g., the schema of a city includes slots for 'location', 'size', etc.). Schematic knowledge can economically be retained in the internal memory (e.g., 'Kilkenny is a city'), whereas details (e.g., 'is situated in Ireland', 'population of 70,000') can be transferred to external stores. On demand, the slots can be filled in the internal memory when queries are initiated in order to retrieve information from the external store. Similarly, macrostructures form integrated accounts from texts (e.g., 'Gandhi fought for the liberation of India'). Individuals who are asked to study a lengthy text on the history of India, and are permitted to use a database system, typically memorize such macrostructures, and enter explicating details (e.g., years, places) into the database. When tested for retention, they first recall the macrostructures, and then elaborate by relying on the database.

3. *Prospects*

Most certainly, technological and organizational progress will continue to enhance the effectiveness and efficiency of external information stores. A simultaneous evolution of human brains yielding a substantial improvement of personal memories cannot be predicted for the near future. Therefore, the importance of external storage in private and professional life will continue to increase. Technical devices will complement personal memory and compensate for its deficits. On the other hand, the internal memory will remain an indispensable component enabling control and economical utilization of extended memory systems.

See also: Attention; Attention: Models; Attention, Neural Basis of; Encoding Specificity in Memory; Macrostructure in Discourse Comprehension, Psychology of; Memory Retrieval; Mental Representations, Psychology of; Metamemory, Psychology of

Bibliography

Garcia-Moina H, Ullman J, Widom J 1999 *Database System Implementation*. Prentice-Hall, Upper Saddle River, NJ

Halbwachs M 1992 *On Collective Memory*. University of Chicago Press, Chicago

Harris J E 1980 Memory aids people use: two interview studies. *Memory and Cognition* **8**: 31–8

Koch P, Krämer S (eds.) 1997 *Schrift, Medien, Kognition. Über die Exteriorität des Geistes* [*Scripture, Media, Cognition. On the Exteriority of Mind*]. Stauffenburg, Tübingen, Germany

Schönpflug W, Esser K B 1995 Memory and its graeculi: Metamemory and control in extended memory systems. In: Weaver Ch A, Mannes S, Fletcher Ch (eds.) *Discourse Comprehension: Strategies and Processing Revisited*. Erlbaum, Hillsdale, NJ, pp. 245–56

W. Schönpflug

External Validity

1. Definition and Challenges

External validity (along with statistical conclusion validity, internal validity, and construct validity) provides a framework to conduct and interpret empirical studies of causal relationships. External validity assesses 'to what populations, settings, treatment variables, and measurement variables can [the effect] be generalized?' (Campbell and Stanley 1966, p. 5). By definition, external validity is about rules for inference; they decrease uncertainty that a cause–effect relationship holds under particular conditions. External validity is not defined by the *extent* to which a cause–effect relationship generalizes across conditions, any more than internal validity is defined by low p values or large effect sizes. After all, when we know that a causal relationship does *not* generalize, the information is useful to refine theory or to prevent the useless expenditure of resources, just as internally-valid information on absence of an effect is useful for these reasons. Sometimes researchers apply the term 'external validity' to correlational studies, for example, validating a measure by applying it to populations other than those originally studied. Although this more general usage of external validity is consistent with recent reformulations (e.g., Cronbach 1982), this article will deal only with the external validity of causal relationships. Historically, researchers have faced three challenges in assessing external validity, which new developments have helped to address.

1.1 The Challenge of Inductive Inference

External validity requires inductive inference, relying on extrapolation beyond the conditions and findings of a single study. Induction is never fully justified on logical grounds (Campbell and Stanley 1966) (see *Generalization: Conceptions in the Social Sciences*). By contrast, internal validity is established through a more deductive framework whereby causal relationships can be both corroborated and falsified within a single study (see *Internal Validity*). Both internal and external validity depend on information that originates outside the single study, especially regarding judgements about what is already known. However, external validity is considerably more dependent on information from outside the study, and this poses a challenge: can external validity rise above the level of the educated guess?

1.2 Perceived Tradeoffs Between Internal and External Validity

Internal validity is compromised if research does not involve controls and the measurement of outcomes. But external validity can often be compromised by the very presence of design controls, by obtrusive measures, and by restricting the study to contexts where respondents are willing to receive any treatment to which they might be assigned (e.g., Barker 1968, Brunswik 1955). Researchers who are concerned with the artificiality of controlled studies have, over time, divided into those who set less priority on the study of causes, in general, and those who would maximize external validity by studying causal relationships in field settings. Some researchers put a higher priority on internal validity, arguing that there is little point in generalizing a causal conclusion about which there is legitimate doubt; moreover, even in field research the study context loses verisimilitude whenever controls are needed to help assess cause (Campbell and Stanley 1966, Cook and Campbell 1979). Others maintain that external validity deserves priority because the research should be optimally relevant to practice (Cronbach 1982). The challenge then arises: what is the best balance in a given study? Cannot both internal and external validity be optimized somehow?

1.3 The Challenge of Complex Interactions

Causal relationships in real-world settings are complex, and statistical interactions of variables are assumed to be pervasive (e.g., Brunswik 1955, Cronbach 1982). This means that the strength of a causal relationship is assumed to vary with the population, setting, or time represented within any given study, and with the researcher's choices about treatments and measurement of outcomes. Without a sensitive assessment of such interactions, true effects can be obscured or causal claims can be overgeneralized to a wider range of people, settings, times, treatments, or outcome constructs than is warranted. Unfortunately, the number of possible interactions is endless, posing problems to the analyst that are insuperable, at least in theory (Cook 1993). The challenge is: how do we cope with this complexity?

The practical task is to assess the most plausible interactions within a given research area. Interactions occur often between treatment, populations, measures and context, posing the most plausible *threats to external validity* (Campbell and Stanley 1966, Cook and Campbell 1979). Because external validity specifies the conditions under which an internally valid relationship can be reproduced, threats to external validity necessarily invoke internal validity.

The interaction of testing and treatment suggests that an effect size varies by the conditions of measurement.

The interaction of selection and treatment hypothesizes that an effect size varies by population studied.

The interaction of setting and treatment describes effects that vary by setting.

The interaction of history and treatment deals with the extent to which a causal relationship replicates across different times.

Other interactions between features of the study and treatment sometimes occur, but their plausibility depends on the individual study context and on what is generally known about these study features (see *Internal Validity*).

2. External Validity Involves Inferences from Samples

2.1 What Do We Generalize About?

The statistician would say that researchers generalize about a sample they have observed of a population, usually of persons or other units being studied. However, external validity treats inferences about treatments, measures, and settings just as seriously as inferences about human population groups. It therefore assumes that studies have to sample from these domains, as well as from human populations (Cook 1993). A given study therefore represents not only a sample from a population of units, but also a sample from a universe of potential settings, a universe of potential treatments, and a universe of potential measures. The sampling frame becomes explicit whenever an investigator makes choices about whom to study, what treatment to administer, or which measures upon which to base inferences. Beyond the individual study, bodies of research (literature reviews, meta-analyses, and programs of related studies) are now understood to be samples from the universe of potential studies. The available sample of findings both limits and guides inferences about external validity.

The sampling frame for external validity considers not only categories, but constructs (Cronbach 1982). Construct validity is therefore intimately tied to external validity. This becomes clear when we note that a population of persons is akin to a category of persons and that specifying the treatment variables involved in a causal relationship requires identifying the entities, constructs, classes or categories to which generalization can be justified (Cook and Campbell 1979).

Methodologists have adopted the following notation from Cronbach (1982). Upper case represents the universe or domain, while lower case notation represents an element in a specific study.

U represents the population of units (persons, sites, communities, organizations, states, nations, or other settings and aggregates) about which a conclusion is sought. In contrast, *u* refers to the sample of units in a study.

T represents the plan and the set of realizations for a treatment, experimental manipulation, program, or other intervention, while *t* is the single realization of *T* for a given *u*. A naturally-occurring event can also be denoted *t*, such as an earthquake or economic recession, and *T* the universe of similar events.

O represents the admissible procedures and conditions for obtaining data (and the data themselves). In causal research, it is often used to refer to the possible effect, thus using *O* to represent the outcome construct and *o* the actual measure of effect.

S represents the context or setting, including the culture and times in which the study takes place. Some researchers use *s* to denote an individual study's setting and large *S* to denote the accrual of such settings (Campbell and Stanley 1966). Others maintain that *S* is always confounded with *U* and *T*, and cannot be deliberately sampled in the same way as *U*, *T*, and *O*.

UTOS is the domain about which a research question is asked. Once *U*, *T*, *O*, and *S* are specified concretely, that constitutes a description of the casual question being researched. On the other hand, an individual *utoS* is the combination of *u*, *t*, and *o* that constitutes an instance on which data are collected. Thus, it is the operational or sampling level of a study; it is how the researcher proposes to answer the question about *T* as a cause, *O* as an effect, *U* as a population of persons, and *S* as the cultural and historical context in which the hypothesis is tested.

Within *UTOS*, replications are of three kinds: (a) exact replication using fresh samples of *U*, *T*, and *O*; (b) studies by other investigators trying to repeat the *UTO* procedures according to the best description available; and (c) studies by other investigators who choose different procedures to represent the same *U*, *T*, and *O* based on their own judgement about what constitutes instances of each. In all three cases, the interest lies in ascertaining whether a class of causal relationships was replicated. For example, the policy maker may want to know whether a treatment for injection drug users (IDUs) in New York City will produce a reduction in behaviors that lead to HIV transmission. Replication of the treatment across fresh samples of IDUs in New York and other northeastern cities increases confidence about treatment effectiveness.

At the level of *UTO*, the instances are joint distributions of units, treatments, and observations. However, some combinations are more common than others. For instance, in laboratory experiments in social psychology certain classes of treatment studied there (e.g., about factors affecting group performance) are confounded with a population almost exclusively composed of college sophomores. By studying categories within these joint distributions, called sub-

UTOS, we may ask whether a causal relationship holds with different categories within *UTOS*. These lower-order categories often include characteristics of the populations, variations in the content or level of a treatment or dimensions of some set of measured responses. For example, one may study attitude changes in whites and blacks, or in cases that experience a treatment at high and moderate levels of intensity and duration. Sub-*UTOS* provide a key to external validity because patterns of variation within a study allow us to ascertain some of the conditions under which a causal relationship holds.

2.2 What Do We Generalize To?

Cronbach (1982) distinguished two basic types of external validity inferences: those concerning reproducibility of findings in *UTOS*, and those that go beyond *UTOS* to new domains termed **UTOS*. The asterisk indicates that causal relationships in these domains have not yet been studied, or have not been studied jointly.

Inferences about *UTOS* are both retrospective and prospective, because they rely on an existing corpus of replications and a definition of constructs that incorporates the existing range of replications. The studies are already in existence. Generalization from *utoS* to *UTOS* is not entirely retrospective, since inferences sometimes need to be made about levels of variables that were not directly observed, for example, when fourth-grade and sixth-grade children are studied, but not those in fifth grade, or when four hours of therapy are given vs. eight hours. Moreover, a critical understanding of variation in *UTOS* always depends on construct validity and theory-building, which usually go beyond the observed *utoS*. Nevertheless, at some point causal hypotheses undergo a test that then extends external validity to new variations within *UTOS*, or expands *UTOS*.

By comparison, inferences about **UTOS* are much more heavily prospective and even speculative, because by definition causal relationships in **UTOS* have not been studied directly. This poses a problem because new knowledge often stimulates a desire to draw conclusions about causal relationships outside *UTOS*, whether the finding involves an important relationship in basic social science, or the effect of a social program or policy. Take the example of HIV prevention: results may impress policy makers who are looking for ways to combat HIV/AIDS in other populations. Will the treatment for IDUs employed in the 1980s, when modified for use (**T*) with women at risk in family planning clinics (**U*), in public housing communities of the twenty-first century (**S*), also lead to reductions in the (very different) behaviors that put them at risk of HIV transmission (**O*)? This requires asking whether the cause–effect relationship generalizes to **UTOS*. The answer is not definitive until cause-probing studies of **UTOS* are conducted.

2.3 Generalization as Decision Making Under Uncertainty

There is always some uncertainty about causation in the social sciences, since plausible alternative explanations may be discovered for even the best replicated cause-effect relationship (see *Internal Validity*). However, uncertainty increases about the conditions under which causal relationships hold, as investigators move from an observed set of *utoS*, to unobserved variations within a well-defined *UTOS*, to **UTOS* (Cronbach 1982). Bayesian statisticians offer a useful way to cope with uncertainty about generalization: first incorporate information to assess the prior likelihood that a finding will generalize, assess the degree of uncertainty about the prior likelihood, and assess the likely consequences of a decision about generalization that is correct or incorrect (Cronbach 1982). Where the consequences are important (expenditure of public resources, or a new direction for theory), new information is needed. After a decision is taken, some additional information will become available: either a test of causal relationships or descriptive and experiential information that may help to further reduce uncertainty.

Generalization is more than educated guessing when it is systematic guessing. The next section outlines principles that reduce uncertainty in generalizing about *UTOS*, by planning studies or using the variation in existing studies; and in generalizing to **UTOS*, through noncausal empirical studies, expert judgement, and theory-building.

3. Studies to Reduce Uncertainty About UTOS and *UTOS

Inferences about external validity can proceed only so far in the context of a single study. Many of the problems can be addressed more adequately through literature reviews and programs of studies, whether planned or unplanned. Multiple studies allow us to improve the *quality* of inductive inference about external validity. They also permit us to recognize potential tradeoffs between internal and external validity and to *optimize* both (where feasible). Finally, they improve our ability to *disentangle the web of interactions* among variables and so identify more of the conditions under which a causal relationship varies in magnitude and even sign.

3.1 Practical Limitations on Representative Statistical Sampling in utoS

Representative statistical sampling of *utoS* would be ideal, but it is not realistic for most causal studies in the social sciences. These operate under conditions that prohibit a sampling frame that could represent the universe of all possible versions of a cause (*T*) and

effect (*O*). Concerning *S*, we cannot sample historical occasions at random in order to probe the temporal generality of a causal proposition, and expense usually prohibits choosing a random sample of all relevant social situations in order to implement a treatment therein. And to the degree that voluntary consent is required, we are prohibited from studying certain *u* in certain ways.

4. Alternative Methods and Principles

A pragmatic alternative approach is to use the less sophisticated methods and principles typically used for generalizing from operational representations of a cause or effect construct. Social scientists routinely generalize from *t* to *T* and from *o* to *O* without random sampling, and they seem to feel very confident about the inferences they draw because the *constructs* make logical sense. Can the principles and methods they use also apply to generalizing from *u* to *U* and from *s* to *S*, and even to **UTOS* (Cook 1993)?

4.1 The Principle of Proximal Similarity

Proximal similarity (Campbell 1988) refers to instances or samples of people, treatments, measures, and settings that look like the populations or categories to which one would like to generalize. A limited set of matching variables is needed that, on the basis of prior theory and other forms of consensus in the social science community, identify the most prototypical attributes of the entity to which generalization is sought. This theory of generalization depends on: (a) superficial correspondences rather than formal sampling, (b) the ability to specify prototypical attributes of a given *UTOS* (Berkowitz and Donnerstein 1982), and (c) on the ability to purposively select research operations that represent the targets of inference at least on the sampled, prototypical variables and any correlates thereof (Cook 1993). Such matching is central to the process researchers now routinely use in research planning when they want to select, not just cause and effect operations, but also measures of control, moderating and mediating variables.

For generalization in *UTOS*, prototypical attributes need to be included in the existing or planned cause-probing studies (Cook 1993). For **UTOS*, the research community uses prior information and best judgement to determine the plausibility or implausibility of generalization (Cronbach 1982). Qualitative and quantitative descriptive studies about **U* may become available. For example, focus groups may tell us how Hispanics experience heart attack symptoms; a **T* is identified, similar to *T* from existing studies, that is likely to increase Hispanics' recognition of symptoms (Leviton et al. 1999).

In popular language, we might exemplify the principle of proximal similarity by the following: if you want to generalize about ducks and cannot randomly choose instances of ducks from the class of ducks, then you should at least select instances for study that look like our culturally shared image of ducks, that quack like ducks and that waddle like ducks. But so defined, 'ducks' could still be confused with small geese or juvenile swans. To deal with this we need either a theory of classification based on more valid characteristics that are not easily observable (e.g., DNA) or we need additional principles.

4.2 The Principle of Discriminant Validity

The constructs in *UTOS* will be well-defined when cognate concepts can be identified and differentiated from them. So, one has to be explicit about how a duck differs from a goose (or liking from love, or how one school reform program differs from another). For generalization about *UTOS* one has to demonstrate empirically that the causal proposition one wants to defend holds for ducks when they are differentiated (discriminated) from swans. Both theory and measurement are needed to support the discrimination required.

To differentiate what can be generalized in **UTOS*, boundary variables are helpful (Fromkin and Streufert 1976), defined as 'critical differences between the research setting and criterion setting which could alter the existence or magnitude of the relationship between two or more variables' (Bernardin and Villanova 1986, p. 43). For example, cause-probing studies of performance appraisal are usually conducted in laboratories, which are somewhat artificial compared to the business environment. To explore external validity, Bernardin and Villanova (1986) surveyed 125 personnel administrators about the most common characteristics of their performance appraisal systems, supplementing this information with other surveys and the available literature. Some laboratory studies were ruled out as too dissimilar to these modal instances, while others were tentatively ruled in.

4.3 The Principle of Heterogeneous Irrelevancies

Across multiple measures, sites, or studies, much irrelevant variation can occur in *UTOS*. Thus, a treatment might be implemented in different ways; an outcome might be conceptualized and measured in different ways; the samples of respondents might differ in age, gender, or racial composition even though they all belong in the category to which one wants to generalize. This heterogeneity is usually considered an impediment, and it can certainly reduce statistical power. However, it also provides the opportunity to test whether a causal relationship holds across the irrelevancies actually sampled (Cook 1993, Kruglanski and Kroy 1976). It is therefore a test of whether the relationship is robust despite irrelevancy in how

measurement occurs, a treatment is operationalized, a sample is drawn and settings are selected.

A deliberate strong test of external validity will introduce as much irrelevant variation in *UTOS* as possible. In this sense, the *extent* of generalization does, in fact, assist our inferences about external validity. For example, we know that various populations at risk of HIV/AIDS (*U*) respond well to messages delivered by people similar to themselves (*T*), although the venues vary tremendously (street, classroom, gay bar, homeless shelter), and the specific risk behaviors (*O*) vary as well. The risk behaviors form a larger class, as do the populations and treatments, and the findings replicate across cultures and decades (*S*). For generalizing to **UTOS*, this robust type of finding is also helpful. Given the consistent findings, what is the prior likelihood that a new group at risk of HIV/AIDS (**U*) would also change risk behaviors when exposed to an exemplar of *T*? It is certainly not zero.

4.4 The Principle of Causal Explanation

One of the major goals of science is to explain phenomena, including causal relationships. Identifying the how or why of a relationship can help those with the requisite knowledge to recreate whatever causal forces originally created the effect in circumstances that are quite different. For instance, knowing why TV violence causes violent behavior in children might help modify how violence is portrayed in the future. This is why causal explanation is held to be the major mechanism for moving from *utos* to **UTOS*, for generalizing to as yet unstudied circumstances (Cronbach 1982). Thus, we judge social and behavioral theories by their capacity to specify constructs that moderate or mediate a causal relationship. However, it is rare to find studies that have manipulated all of these causal-explanatory contenders or even a goodly portion of them. Some are usually measured, though, and so it becomes important to use measurement and data analysis to establish which components are most likely to be causally responsible for an effect and so are essential to any practical generalizations that will be made from the data.

4.5 The Principle of Interpolation and Extrapolation

Few social or behavioral theories succeed in specifying the levels of a treatment that will usually bring about a desired effect. Parametric studies that vary many levels on the independent variable and then examine response curves are rare: instead, just two or three levels are manipulated, leaving us unclear about how the outcome variable would have behaved at points between the levels measured. In principle, such interpolation is easy. Dose-response studies, or other plans to vary the level of *T*, provide a means to assess this issue, as does meta-analysis when levels of *T* vary across studies. Moreover, if a similar effect size is obtained across two levels, it is often reasonable to assume it will also hold at non-sampled points between them. For generalizing about *UTOS*, it is desirable to sample a wide range of treatment levels and as many points in-between as are feasible. In the same way, we might sample a wide range of persons (e.g., on age or educational attainment), time periods (e.g., years after discovery of the AIDS epidemic), and levels of outcome variables that might otherwise be truncated.

Extrapolation of effects that are beyond the ranges available in studies is more problematic, because the shape of the curve or the nature of the effect itself may change at more extreme values on the independent variable. Thus, water turns to ice at zero degrees centigrade and to vapor at 100 degrees. Between these points, it is water. But few social science relationships are understood this well. We simply do not know where qualitative transformations come about in many social relations (for instance, how much life stress is needed to create clinically significant depression). Extrapolation to points closer at hand (*UTOS*) is generally less uncertain than extrapolation to more distant points (**UTOS*).

See also: Content Validity; Generalization: Conceptions in the Social Sciences; Internal Validity

Bibliography

Barker R G 1968 *Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior*. Stanford University Press, Stanford, CA

Berkowitz L, Donnerstein E 1982 External validity is more than skin deep. *American Psychologist* **37**: 245–57

Bernardin H J, Villanova P 1986 Performance appraisal. In: Locke E A (ed.) *Generalizing from Laboratory to Field Settings*. Lexington Books, Lexington, MA

Brunswik E 1955 Representative design and probabilistic theory in a functional psychology. *Psychological Review* **62**: 193–217

Campbell D T 1988 *Methodology and Epistemology for Social Science: Selected Papers*. University of Chicago Press, Chicago

Campbell D T, Stanley J C 1966 *Experimental and Quasi-experimental Designs for Research*. Rand McNally, Chicago

Cook T D 1993 A quasi-sampling theory of the generalization of causal relationships. In: Sechrest L B, Scott A G (eds.) *Understanding Causes and Generalizing About Them. New Directions for Program Evaluation* **57**: 39–82

Cook T D, Campbell D T 1979 *Quasi-experimentation*. Rand McNally, Chicago

Cronbach L J 1982 *Designing Evaluations of Educational and Social Programs*. Jossey-Bass, San Francisco

Fromkin H L, Streufert S 1976 Laboratory experimentation. In: Dunnette M D (ed.) *Handbook of Industrial and Organizational Psychology*. Rand McNally, Chicago, pp. 415–65

Kruglanski A W, Kroy M 1976 Outcome validity in experimental research: A reconceptualization. *Representative Research in Social Psychology* **7**: 166–76

Leviton L C, Finnegan J R, Zapka J G 1999 Formative research methods to understand patient and provider responses to heart attack symptoms. *Evaluation and Program Planning* **22**: 385–98

L. C. Leviton

Extinction of Species

There is much evidence that we are into the opening phase of a mass extinction of species (Ehrlich and Ehrlich 1981, Myers 1990, Pimm et al. 1995, Raven 1990, Wilson 1992). A mass extinction can be defined as an exceptional decline in biodiversity that is substantial in size and generally global in extent, and affecting a broad range of taxonomic groups over a short period of time (Jablonski 1986). In this sense, the present mass extinction—if it remains unchecked by conservation action of appropriate scope and scale—will rival and conceivably surpass the numbers of species lost during the five mass-extinction episodes of the prehistoric past (Myers 1996).

Earth's stock of species generally is estimated to total a minimum of 10 million. Some scientists believe the true total could well be 30 million, possibly 50 million and conceivably 100 million (Ehrlich and Wilson 1991). Of the conservative estimate of 10 million species, around 90 percent usually are considered to be terrestrial; and of the terrestrial species, roughly 80 percent or more than seven million are believed to occur in the tropics, with roughly five million in tropical moist forests (Ricklefs and Schluter 1992). As we shall see, these tropical forests are not only the richest biome biotically but constitute the biome where habitat depletion is occurring fastest—'depletion' including not only outright destruction but gross disruption (fragmentation, etc.), with loss of many ecosystem attributes including food webs. So this is the prime locus of the mass extinction underway, and it is the main focus of the present analysis.

Consider again the higher estimates for the planetary spectrum of species, between 30 million and 100 million. Since many if not most of the additional species are believed to occur in tropical forests, the true planetary total is not only a matter for speculation. If the real total is 30 million species, the extinction rate will be much higher than the rate postulated on the basis of only 10 million species. But for the sake of being cautious and conservative, let us accept a total of 10 million species. As noted, at least half of these species live in tropical forests, even though remaining forests cover only 6 percent of Earth's land surface.

Tropical forests are being destroyed at a rate of at least 150,000 square kilometres per year (Myers 1992). In addition to this outright destruction, an expanse as large again is being grossly disrupted through overheavy logging and slash-and-burn cultivation, with much degradation and impoverishment of ecosystems and species' life-support systems. But in the interests of being cautious and conservative again, let us consider only outright destruction. The current annual loss represents 2 percent of remaining forests; and the amount increased by nearly 90 percent during the 1980s. If present patterns and trends of forest destruction persist, leading to still more acceleration in the annual rate, the current destruction of two percent of remaining forests may well have increased by another 50 percent during the 1990s. In a recent year more forest was burned in Amazonia and Borneo than ever before.

How shall we translate a 2-percent annual rate of forest destruction into an annual species extinction rate? An analytic approach is supplied by the theory of island biogeography, being a well-established theory with much empirical evidence from on-ground analyses around the world (MacArthur and Wilson 1967). The theory states that the number of species in an area rises or falls as approximately the fourth root of the increase or decline in the area, with actual values varying between the third and fifth roots. So when a habitat loses 90 percent of its original extent, it can generally support no more than 50 percent of its original species.

The calculation depends critically upon the status of the remaining 10 percent of habitat. If this relict expanse is split into many small pieces (as is often the case with remnant tracts of tropical forests), a further 'islandizing effect' comes into play, reducing the stock of surviving species still more. It is not clear how severe this additional depletion can be. Informed estimates suggest the 50 percent can readily be reduced to 40 percent, more generally to 30 percent, sometimes to 20 percent, and occasionally even to 10 percent. Similarly, if most species concerned occur in small, local endemic communities, the percentage loss of species can readily approach the percentage loss of area. Moreover, isolated remnants of forest become prone to additional depauperizing processes such as 'edge effects', also to dessicating effects due to local climate change (Myers 1992). So the 90/50 calculation for island-biogeography extinctions should be viewed as a minimum estimate.

The most broadscale application of island biogeography to tropical deforestation and species extinctions has been presented by one of the authors of the original theory, Edward Wilson of Harvard University. He estimates a late-1980s annual loss of 27,000 species in the forests (Wilson 1992). As he emphasizes repeatedly, this estimate is extremely optimistic. If we employ a more 'realistic' reckoning by qualitatively incorporating a number of other bioecological factors such as alien introductions, overhunting, and diseases, the annual total will become larger. (Indeed certain

other analyses propose that 60,000–90,000 species are lost in tropical forests each year.) In addition, a good number of species, albeit undetermined in even preliminary terms, is presumably disappearing in other parts of the world. Suppose we accept a bare minimum estimate of 30,000 species eliminated worldwide per year. This means we are witnessing a rate at least 120,000 times higher than the 'natural' rate of extinctions before the advent of the human era, considered to be perhaps one species every four years (Raup 1991).

True, the tropical forest calculation is a broad-brush affair, viewing the forests as a single homogeneous expanse even though they are highly heterogeneous. So the island-biography approach needs to be complemented by a local-scale assessment, available in the form of a 'hotspots' analysis. Hotspots are areas that (a) feature exceptional concentrations of species with exceptional levels of endemism, and (b) face exceptional threats of imminent habitat destruction. There are 25 such hotspots identified which contain 44 percent of all Earth's plant species and 35 percent of all species in four vertebrate groups confined to 1.4 percent of Earth's land surface (Myers et al. 2000). So far as we can determine, they contain an even larger proportion of Earth's animal species with highly localized distributions. Most of these hotspots have already lost the great bulk of their biodiversity habitats, and it is reasonable to suppose that many thousands of species are disappearing annually in hotspot areas alone.

So much for the current extinction rate. What of the future? Through detailed analysis backed by abundant documentation, Wilson (1992) considers we face the prospect of losing 20 percent of all species within 30 years and 50 percent or more thereafter. Another expert, Peter Raven of the Missouri Botanical Garden, calculates that half of all species exist in those particular tropical forests that, in the absence of adequate conservation measures, will be reduced to less than one tenth of their present expanse within the next three decades. In accord with island biogeography, Raven (1990) concludes—and he stresses this is a conservative prognosis—that one quarter of all species are likely to be eliminated during the next 30 years. Worse, 'fully half of total species may disappear before the close of the 21st century.' Another two biodiversity analysts, Paul and Anne Ehrlich (1992), assert that 'If the current accelerating trends (of habitat destruction) continue, half of Earth's species might easily disappear by 2050.' These estimates are in line with those of several other biodiversity specialists. If we accept, then, that half of all species are likely to be eliminated within the foreseeable future, this means in turn that perhaps one third of the problem is confined to the 25 hotspots.

The loss of large numbers of species will be far from the only outcome of the present biotic debacle, supposing it proceeds unchecked. There is likely to be a significant disruption of certain basic processes of evolution (Myers 1996). The forces of natural selection and speciation can work only with the 'resource base' of species and subunits available—and as we have seen, this crucial base is being grossly reduced. To cite the graphic phrasing of Michael Soule, 'Death is one thing; an end to birth is something else.' Given what we can discern from the geologic record, the recovery period, i.e., the interval until speciation capacities generate a stock of species to match today's in abundance and variety, will be protracted. After the late Cretaceous crash, between 5 and 10 million years elapsed before there were bats in the skies and whales in the seas. Following the mass extinction of the late Permian when marine invertebrates lost roughly half their families, as many as 20 million years were needed before the survivors could establish even half as many families as they had lost.

The evolutionary outcome this time around could prove yet more drastic. The critical factor lies with the likely loss of key environments. Not only do we appear set to lose most if not virtually all tropical forests. There is progressive depletion of tropical coral reefs, wetlands, estuaries, and other biotopes with exceptional abundance and diversity of species and with unusual complexity of ecological workings. These environments have served in the past as pre-eminent 'powerhouses' of evolution, meaning they have thrown up more species than other environments. Virtually every major group of vertebrates and many other large categories of animals have originated in spacious zones with warm, equable climates, notably the Old World tropics and especially their forests. The rate of evolutionary diversification—whether through proliferation of species or through emergence of major new adaptations—has been greatest in the tropics, especially in tropical forests. Tropical species, notably tropical forest species, appear to persist for only brief periods of geologic time, which implies a high rate of evolution.

As extensive environments are eliminated wholesale, moreover, the current mass extinction applies across most if not all major categories of species. The outcome will contrast sharply with the end of the Cretaceous, when not only placental mammals survived (leading to the adaptive radiation of mammals, eventually including man), but also birds, amphibians, and crocodiles among many other nondinosaurian reptiles. In addition, the present extinction spasm is eliminating a large portion of terrestrial plant species, by contrast with mass-extinction episodes of the prehistoric past when terrestrial plants have survived with relatively few losses—and have thus supplied a resource base on which evolutionary processes could start to generate replacement animal species forthwith. If this biotic substrate is markedly depleted within the foreseeable future, the restorative capacities of evolution will be the more diminished.

All this will carry severe implications for human societies extending throughout the recovery period

estimated to last at least five million years, possibly several times longer. Just five million years would be twenty times longer than humankind itself has been a species. The present generation is effectively imposing a decision on the unconsulted behalf of at least 200,000 follow-on generations. It must rank as the most far-reaching decision ever taken on behalf of such a large number of people in the course of human history. Suppose that Earth's population maintains an average of 2.5 billion people during the next five million years (rather than the six billion we have today), and that the generation time remains 25 years. The total affected will be 500 trillion people. Just 1 trillion is a large number: check the length of time made up by 1 trillion seconds.

A final thought, this one as regards our policy responses. The Global Environment Facility under the World Bank has provided half a billion dollars over 3 years to assist biodiversity. This is quite the largest such dispensation ever made. But compare it with what is at stake. Every year, just the commercial value to just the rich nations of just the present array of plant-derived pharmaceuticals is 100 times greater. More significant still, the worldwide amount spent annually on 'perverse' subsidies, i.e., subsidies that inadvertently foster overloading of croplands, overgrazing of rangelands, profligate burning of fossil fuels, wasteful use of water, overcutting of forests and overharvesting of fisheries (to cite but a few examples of activities that also reduce biodiversity) is 40 times greater again (Myers and Kent 2000).

Biodiversity loss is indeed a profound problem. It is also a glorious opportunity. As is obvious, species extinction is irreversible. And we face the prospect of eliminating species in their millions, perhaps half of all that share this planet with us, by virtue of what we do (and don't do) during the immediate future. Time is of the super-essence. Suppose we allow ourselves until, say, the year 2010 to take the vital decisions that will affect our planetary home for hundreds, thousands, and even millions of years. After the year 2010, the processes of habitat destruction will surely have worked up so much momentum that they will be hard to slow down, let alone to halt or even reverse. Till 2010 there are roughly 3,500 days. We lose one percent of our maneuvering room every five weeks!

Bibliography

Ehrlich P R, Ehrlich A H 1981 *Extinction: The Causes and Consequences of the Disappearance of Species*, 1st edn. Random House, New York

Ehrlich P R, Ehrlich A H 1992 The value of biodiversity. *Ambio* **21**: 219–26

Ehrlich P R, Wilson E O 1991 Biodiversity studies, science and policy. *Science* **283**: 758–62

Jablonski D 1986 Causes and consequences of mass extinctions: a comparative approach. In: Elliott D K (ed.) *Dynamics of Extinction*. Wiley, New York, pp. 183–229

MacArthur R H, Wilson E O 1967 *The Theory of Island Biogeography*. Princeton University Press, Princeton, NJ

Myers N 1990 Mass extinctions: what can the past tell us about the present and the future? *Global and Planetary Change* **82**: 175–85

Myers N (ed.) 1992 *Tropical Forests and Climate*. Kluwer Academic, Dordrecht, The Netherlands

Myers N 1996 The biodiversity crisis and the future of evolution. *The Environmentalist* **16**: 37–47

Myers N, Kent J 2000 *Perverse Subsidies: How Tax Dollars Can Undercut the Environment and the Economy*. Island Press, Washington, DC

Myers N, Mittermeier R-A, Mittermeier C G, da Fonseca G, Kent J 2000 Biodiversity hotspots for conservation priorities. *Nature* **403**: 853–58

Pimm S L, Russell G J, Gittleman J L, Brooks T M 1995 The future of biodiversity. *Science* **269**: 347–54

Raup D 1991 *Extinction: Bad Genes or Bad Luck?* Norton, New York

Raven P H 1990 The politics of preserving biodiversity. *BioScience* **40**(10): 769–74

Ricklefs R, Schluter D (eds.) 1992 *Species Diversity: Geographical and Historical Aspects*. University of Chicago Press, Chicago, IL

Wilson E O 1992 *The Diversity of Life*. Belknap Press of Harvard University Press, Cambridge, MA

N. Myers

Extraversion

Extraversion is one of the most-studied individual differences in psychology. It has been the focus of research attention for as long as personality psychology has existed; the trait or some variation of it has been included in virtually all comprehensive models of personality (see *Personality Structure*). Extraversion (also spelled extroversion) encompasses a set of characteristics including active engagement with the world, self-confidence, sociability, and the need for external stimulation. Yet, the defining characteristics and processes hypothesized to underlie the trait have varied considerably across theorists. Thus, the behavioral outcomes of extraversion such as leadership ability and the propensity to engage in impulsive behaviors might be related to one type of extraversion but not to others.

1. Early History of Extraversion

Precursors of the extraversion dimension appeared in the writings of the ancient Greeks and Romans. Hippocrates and Galen described a set of personality types that were based on the balance of four humors in the body, and these categories resemble the opposing poles of the modern extraversion–introversion dimension. In the early twentieth century, Wundt extended

this ancient model, suggesting that the personality types that Hippocrates and Galen identified are not discrete categories, but continuous dimensions (for a historical overview of these models, see Eysenck and Eysenck 1985). Specifically, the personality dimension that is now thought of as extraversion–introversion was said to reflect a person's propensity to have quickly changing emotions (extraversion) versus slowly changing or stable emotions (introversion). Thus, in Wundt's formulation, when situations change, extraverts' emotions change too, whereas introverts are left introspecting and reflecting.

Early theorists suggested that extraverts and introverts process information about the world in very different ways (for a review of early models, see Guilford and Braly 1930). Otto Gross, for example, proposed two personality types that were based on reactions following exposure to a stimulus. According to Gross, 'primary' functions are elicited immediately upon stimulation, and 'secondary' functions 'persist after stimulation and permit organization and systematization of sense-impressions' (Guilford and Braly 1930, p. 97). The 'shallow–broad' personality type was thought to have relatively high levels of primary function compared to secondary function, whereas the 'deep–narrow' type was thought to have relatively high levels of secondary function compared to primary function. Thus, the shallow–broad type (corresponding to the extravert) was predisposed to short-lived reactions to immediate situations, whereas the deep–narrow type (corresponding to the introvert) was more likely to introspect and brood about past events.

Citing work by Gross and others, Jung ([1921] 1971) proposed what is probably the first trait that was named extraversion/introversion. According to Jung, the causal mechanism underlying extraversion can be traced to individual differences in orientation towards the external, objective world. Like Gross's shallow–broad type, Jung's extravert is characterized by immediate reactions to the objects and stimuli that he or she experiences. The introvert, on the other hand, is oriented toward the subjective feelings that an object or stimulus creates, just as the deep–narrow type is dominated by secondary processes. Extraverts react to the immediate features of the surrounding environment, whereas introverts are influenced by the subjective feelings that such stimuli arouse. Thus, extraverts were thought to be better able to adapt to the changing external environment and were seen as less prone to introspection.

Beginning in the 1930s, Guilford used factor analytic techniques to examine whether the different individual difference characteristics that researchers had been studying reflect a single unidimensional trait of extraversion. In a series of studies, Guilford found that extraversion measures were indeed multidimensional, containing two different 'extraverted' axes (see *Personality Assessment*). He and his colleagues developed a series of scales to assess the various dimensions, culminating in the Guilford–Zimmerman Temperament Survey (Guilford and Zimmerman 1949). Their hierarchical questionnaire and model of personality included 11 primary traits and a number of higher-order factors. Two of the higher-order factors are relevant to the extraversion dimension: a social activity factor that includes 'general activity,' 'ascendance,' and 'sociability' (see *Personality and Social Behavior*); and an impulsivity factor that includes the lack of 'restraint' and lack of 'thoughtfulness' (see *Personality and Risk Taking*). Following the lead of early personality theorists, Guilford labeled the impulsivity factor rather than the social activity factor as extraversion.

2. *Hans Eysenck and the Arousal Model*

Around the same time that Guilford was refining his measures, Hans Eysenck was developing his own comprehensive model of personality. His early work focused on identifying dimensions that underlie different forms of neuroses. According to Eysenck, extraversion emerged consistently as the second major factor that could account for variance in symptoms and behavioral tasks once a neuroticism factor was removed. Like Guilford, Eysenck believed that extraverts were primarily impulsive individuals and he was initially unwilling to include sociability as an important facet of the trait. Eventually, after substantial factor analytic work, Eysenck concluded that sociability and impulsivity were correlated and together they formed the higher-order trait of extraversion.

At the same time that Eysenck was constructing and refining his measures of extraversion, he was also developing an influential psychophysiological theory of extraversion. Originally, Eysenck proposed a theory of extraversion that was based on individual differences in excitation and inhibition. However, this theory was found to be insufficient and was quickly replaced with the arousal theory of extraversion (for a historical review of Eysenck's models, see Eysenck and Eysenck 1985). According to the arousal theory, individual differences in extraversion result from differences in levels of general arousal, which in turn are linked to activity in the ascending reticular activation system (ARAS). Because introverts are thought to have low thresholds for ARAS activation, they are hypothesized to become more aroused when the ARAS is activated. Extraverts, on the other hand, are thought to have higher thresholds, and thus lower internal arousal. Eysenck also proposed that too much or too little arousal was detrimental to certain types of performance and was subjectively unpleasant. Thus, he could test the arousal hypotheses explicitly by examining performance and subjective enjoyment under conditions of varying stimulation. Extraverts should perform better than introverts and experience

more enjoyment under arousing conditions because extraverts tend to be chronically underaroused in certain neural systems. To explain extraverts' social behavior, Eysenck suggested that social situations are more arousing than nonsocial situations (see *Personality and Social Behavior*). Thus, extraverts tend to seek these situations, whereas introverts try to avoid them (for a critical review of the arousal theory, see Matthews and Gilliland 1999).

3. Extraversion and Reward-sensitivity

Soon after Eysenck proposed the arousal hypothesis, Gray (1970) presented a modified version of the theory. Like Eysenck, Gray believed that a few major psychophysiological systems could explain most of the major variation in personality. However, Gray believed that evidence from studies of brain functioning did not support the extraversion and neuroticism dimensions that Eysenck proposed, and Gray proposed two dimensions that were rotated in factor analytic space from Eysenck's original dimensions. Specifically, Gray argued that the Behavioral Inhibition System (BIS) controls responses to conditioned punishment and nonreward, and variation in the strength of this system leads to individual differences in anxiety. A second system, the Behavioral Activation System (BAS) controls responses to conditioned reward and nonpunishment and leads to individual differences in impulsiveness. Instead of attributing individual differences in personality to arousal, Gray suggested that extraversion and neuroticism result from individual differences in reward and punishment sensitivity. Extraversion reflects the relative strength of the BAS versus the BIS (extraverts have stronger BAS functions), and neuroticism reflects the combined strength of the BAS and the BIS (neurotics have strong BIS and BAS).

Ironically, although Gray labeled the dimension corresponding to the BAS as 'impulsivity,' many theorists who adopted Gray's ideas downplayed this impulsivity label, focusing instead on a separate theoretical consequence of the BAS—high pleasant affect. For example, Tellegen (1985) suggested that extraversion is a broad dimension that reflects positive emotionality rather than impulsivity or sociability. Tellegen linked positive emotionality to Gray's BAS, suggesting that reward-sensitive individuals were more likely to experience positive affect, and that the broad extraversion factor resulted from this underlying individual difference. Tellegen suggested that impulsivity should remain separate from this reward-sensitivity/positive emotionality dimension. Watson and Clark (1997) extended Tellegen's argument, showing that the different facets of extraversion were often more strongly correlated with positive affect than they were with each other (suggesting that the various characteristics of extraverts were linked by their common associations with positive emotions); and Lucas et al. (2000) found that only the extraversion facets that reflect positive emotionality are associated with the higher-order trait. They suggest that characteristics of extraversion such as warmth, sociability, and dominance are the results of extraverts' greater positive emotionality. DePue and Collins (1999) provided an updated reward sensitivity model based on dopamine systems in the brain.

4. Extraversion and the Big Five

At the same time that Guilford, Eysenck, and others were developing and revising self-report questionnaire measures of the constructs in their models of personality, other researchers were investigating an alternative approach to developing comprehensive models of personality. Beginning with Allport and Odbert (1936), researchers turned to natural language descriptors of personality to identify important personality dimensions (see *Personality Structure*). According to the *lexical hypothesis*, all important dimensions of personality should be represented in everyday language, and thus a systematic examination of personality descriptors should allow personality psychologists to identify these dimensions. Raymond Cattell's early attempts to find the primary dimensions of natural language descriptors resulted in 16 separate traits, but a higher-order trait that resembles extraversion can be identified in his personality measures (Cattell et al. 1980). In other researchers' analyses, factor analyses of personality descriptors often converged on five major factors (the 'Five Factor Model'). The first and largest factor that emerges from these analyses (variously labeled 'confident self-expression,' 'surgency,' 'assertiveness,' 'extraversion,' and 'power') clearly reflects the extraversion dimension.

5. Modern Conceptualizations of Extraversion

Most modern theories of extraversion include one or more of the following clusters of characteristics: warmth, gregariousness, and affiliation; assertiveness and ascendance; high levels of activity; and the tendency to experience pleasant emotions. Occasionally, theorists also emphasize extraverts' ambition or their tendency to seek excitement (Watson and Clark 1997). A number of questionnaire measures are available to assess extraversion, including Costa and McCrae's (1992) Revised NEO Personality Inventory Extraversion scale and Goldberg's International Personality Item Pool (Goldberg, 1999). Goldberg's large set of items includes many variations that allow researchers to assess diverse conceptualizations of the extraversion trait. The Eysenck Personality Inventory and Eysenck Personality Questionnaire are also still used widely (see *Personality Assessment*).

6. Future Directions

Although much extraversion research has been based on exploring how general self-report items cluster together, a deeper understanding can be gained by including on-line recording of feelings and behavior, as well as biological measures (see *Personality Psychology: Methods*). For example, researchers have examined the heritability of the trait (see *Personality Development and Temperament*) and have linked reports of extraverted behaviors to specific neural systems, focusing on the neurotransmitters, hormones, and brain locations that appear to be responsible for extraverted behavior (for a review, see Depue and Collins 1999). This multimethod approach is necessary so that researchers can move beyond the simple analysis of how global self-report items cluster. By including on-line recording of behavior, as well as personality questionnaires and biological measures, we shall be better able to understand the major types of extraversion (e.g., the type based on impulsive sociability vs. the type based on high reactivity in the positive affect system), and the resultant behaviors of each.

See also: Arousal, Neural Basis of; Neuroticism; Personality and Risk Taking; Personality and Social Behavior; Personality Assessment; Personality Psychology; Personality Psychology: Methods; Personality Structure; Personality Theories

Bibliography

Allport G W, Odbert H S 1936 Trait-names: A psycho-lexical study. *Psychological Monographs* **47**(1, Whole No. 211)

Cattell R B, Eber H W, Tatsuoka M M 1980 *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Institute for Personality and Ability Testing, Champaign, IL

Costa P T, McCrae R R 1992 *Revised NEO Personality Inventory (NEOPI-R) and Five Factor Inventory (NEO-FFI) Professional Manual*. Psychological Assessment Resources, Odessa, FL

DePue R A, Collins P F 1999 Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and Brain Sciences* **22**: 491–569

Eysenck H J, Eysenck M W 1985 *Personality and Individual Differences*. Plenum, New York

Goldberg L R 1999 A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. Available without cost at http://ipip.ori.org/ipip/

Gray J A 1970 The psychophysiological basis of introversion–extraversion. *Behaviour Research and Therapy* **8**: 249–66

Guilford J P, Braly K W 1930 Extroversion and introversion. *Psychological Bulletin* **27**: 96–107

Guilford J P, Guilford J S, Zimmerman W S 1949 *The Guilford–Zimmerman Temperament Survey: Manual*. Sheridan Psychological Services, Beverly Hills, CA

Jung C G [1921] 1971 *Psychological Types*. Princeton University Press, Princeton, NJ

Lucas R E, Diener E, Grob A, Suh E M, Shao L 2000 Cross-cultural evidence for the fundamental features of extraversion. *Journal of Personality and Social Psychology* **79**: 452–68

Matthews G, Gilliland K 1999 The personality theories of H. J. Eysenck and J. A. Gray: A comparative review. *Personality & Individual Differences* **26**: 583–626

Tellegen A 1985 Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In: Tuma A H, Maser J D (eds.) *Anxiety and the Anxiety Disorders*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 681–706

Watson D, Clark L A 1997 Extraversion and its positive emotional core. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 767–93

R. E. Lucas and E. Diener

Eye Movement, Control of (Oculomotor Control)

1. Eye Movements

Five types of eye movements are generally recognized. These movements fall into two broad categories: gaze-stabilization movements and gaze-shifting movements. Gaze-stabilization movements rotate the line-of-sight to compensate for movement of the head and body, stabilizing the visual world on the retina. In the absence of these movements, visual acuity is severely compromised because visual images move across the retina whenever the observer moves through the visual environment. In almost all vertebrates two neural systems work together to achieve gaze stabilization: the vestibulo-ocular system and the opto-kinetic system. Gaze-shifting movements, unlike gaze-stabilization movements, shift the fovea into alignment with objects of interest in the visual world. The three classes of movements that achieve gaze shifts are saccades, smooth pursuit eye movements, and vergence eye movements. Saccades are rapid shifts of the line-of-sight from one point in the visual world to another. Smooth pursuit eye movements allow the line-of-sight to track moving objects, slowly rotating the eye to keep a visual stimulus which is moving across the visual field aligned with the fovea. Together saccades and smooth pursuit movements are referred to as versional movements because they cause both eyes to move in a similar direction and at the same speed. Vergence eye movements, which occur only in binocular animals like humans, shift the lines-of-sight of the two eyes with regard to each other (either converging or diverging the angle between the two eyes) so that a visual stimulus can be aligned with the same point on each retina, even if that stimulus moves towards or away from the observer. In the nineteenth century Herring proposed that only vergence move-

ments involve the independent control of both eyes. For all other movements, he proposed that the two eyes always receive identical movement commands. This hypothesis, which still informs our basic view of oculomotor control, is known as Herring's law of equal innervation.

2. The Musculature of the Eyes

In humans, all eye movements are rotations of the eye (never linear motions or translations) and are accomplished by six extra-ocular muscles that operate in three antagonistic pairs. One pair, the medial and lateral rectus muscles, is located on either side of each eyeball and controls the horizontal orientation of the eye. A second pair, the superior and inferior rectus muscles, controls the vertical orientation of the eye and a third pair, the superior and inferior oblique muscles, controls rotations of the eye around the line-of-sight. Although most people are unaware of them, rotations of the eye about the line of sight are quite common, though usually less than 10° in amplitude.

3. Three Cranial Nerve Nuclei Control Tension in the Extra-Ocular Muscles

The tension generated in each muscle is regulated by the activity of pools of alpha-motor neurons located in the oculomotor brainstem. The pools of neurons regulating the superior rectus, inferior rectus, medial rectus, and inferior oblique muscles each lie within separate subdivisions of the oculomotor, or third cranial nerve, nucleus. Neurons innervating the superior oblique muscle lie within the trochlear, or fourth cranial nerve, nucleus. Neurons innervating the lateral rectus muscle lie within the abducens, or sixth cranial nerve, nucleus. Because these six pools are the only source of input to the extraocular muscles, they serve as a final common pathway through which all eye movement control must be implemented. Physiological studies of these neurons made while monkeys produced each of the five types of eye movements described above indicate that all of these motor neurons participate in the generation of all of the five classes of eye movements; there is no specialization amongst the alpha-motor neurons for any particular class of movement.

Engineering models of the eye and its muscles developed in the 1970s (largely by David Robinson and colleagues at Johns Hopkins University, USA; Robinson 1981) indicate that each of these motor neurons must generate two types of muscle force to accomplish any of the five classes of movements: a pulsatile burst of force produced during the movement, which regulates the instantaneous velocity with which the eye rotates, and a long-lasting increment or decrement in maintained force, which persists after the movement is complete. This maintained step in force holds the eye stationary by resisting the elastic forces of the muscles and connective tissues surrounding the eye. Forces which act to passively pull the eye back towards a straight ahead position. Physiological experiments have demonstrated that each motor neuron produces both of these two types of force. In summary then, the control circuits associated with gaze stabilizing and gaze shifting movements all must activate the same set of alpha motor neurons and all of these circuits must produce signals which regulate both the velocity and position of the eyes in a manner appropriate for that type of eye movement.

4. Gaze Stabilization Movements

Gaze stabilization movements achieve image stabilization when the head rotates in space by rotating the eyes at a velocity equal to that of the head but in an exactly opposite direction. Gaze stabilization thus requires very precise information about the speed and direction of head rotation across a broad range of head speeds. At high speeds of head rotation the vestibulo-ocular system obtains information from the semicircular canals of the inner ear which accurately encodes the velocity of head rotation. However, as the velocity of head rotation decreases, the vestibular system begins to underestimate velocity. At low speeds the optokinetic system obtains information about the rate at which the whole visual field slides across the retina, information which accurately encodes the velocity of head rotation. In this system as the velocity of head rotation increases head velocity is underestimated. Thus either system working alone could not fully compensate for all movements of the head and body, but working together these two systems can provide an accurate estimate of head velocity that can be used by the oculomotor control circuitry to counter-rotate the eyes in the head across a broad range of head speeds. Of course, regardless of the speed of eye rotation, as the eyes move towards far eccentric positions in the orbit, it eventually becomes impossible for the eyes to move any farther. The gaze stabilization systems respond to this challenge by very quickly moving the eyes in the opposite direction; this gives rise to a repetitive pattern of quick reset movements and slow stabilization phases usually called nystagmus.

4.1 Vestibulo-ocular Response

In the vestibulo-ocular system (often called the vestibulo-ocular reflex or VOR) the rotational velocity of the head is transduced directly by the semicircular canals of the inner ear (even in total darkness) and enters the brain via the vestibular nucleus. Robinson and colleagues recognized that because this signal

already encodes the velocity of head rotation, in order to compute the desired direction and velocity for a gaze stabilization movement it would only be necessary for the brain to take the inverse of this signal. Recalling that all oculomotor alpha-motor neurons produce both a pulsatile burst of force that regulates the velocity of the eye movement and a long-lasting step-like increment or decrement in maintained force that holds the eye stationary, Robinson proposed that the vestibular signal could be channeled directly to the alpha-motor neurons to produce the pulsatile burst of force. He then went on to suggest that changes to the long-lasting step in force required after each eye rotation could be computed from this pulse, or velocity, signal by the mathematical operation of integration. In the 1980s the lesion of a discrete brain area, the nucleus prepositus hypoglossi, was shown to eliminate only the long-lasting force change from the motor neurons without effecting eye velocity during movements, suggesting that the nucleus prepositus hypoglossi is a neural integrator employed by the oculomotor system (Leigh and Zee 1991).

4.2 Optokinetic Response

In the optokinetic response (often called optokinetic nystagmus or OKN) head velocity is computed based on the global motion of the retinal image. This computation, determining the speed and direction of image motion, is performed in the pretectum and yields a signal that also specifies the rotational velocity of the eyes necessary to achieve image stabilization. This visually derived signal passes to the vestibular nucleus where it has been shown to combine with the vestibular signal used by the VOR to yield a neural estimate of head velocity that is accurate across the entire range of speeds for which gaze stabilization can be accomplished. This combined signal is then processed by the same circuitry that Robinson and his colleagues described for the vestibulo-ocular system. Gaze stabilization thus appears to involve two sensory systems that are combined and passed to a single control circuit that governs eye rotation.

5. Gaze Shifting Movements

In some sense, gaze-shifting movements oppose gaze stabilization. While the VOR and OKN stabilize images on the retina to maximize visual acuity, gaze shifting movements redirect the line-of-sight, temporarily reducing the acuity with which some or all the visual world is perceived.

5.1 Saccades

The saccadic system, in order to achieve a precise gaze-shift, must supply the alpha-motor neurons with a command that controls the amplitude and the direction of the desired change in the line-of-sight. Current evidence indicates that this command can originate in either of two brain structures: the superior colliculus of the midbrain or the frontal eye fields of the neocortex (Sparks 1986). Both of these structures contain laminar sheets of neurons that code all possible saccadic amplitudes and directions in a topographic map-like organization and both project to brainstem regions that supply control signals to the oculomotor neural integrators and the motor neurons described above. Activation of neurons in these collicular and cortical maps has been shown to precede the onset of saccadic eye movements by about 30 milliseconds. One group of theories proposes that these cortical and collicular signals govern a brainstem feedback loop (incorporating the neural integrator) which accelerates the eye to a high velocity and keeps the eye in motion until the desired eye movement is complete while other theories make similar predictions but place this feedback loop outside the brainstem. In either case, it seems clear that the superior colliculus and frontal eye fields are important sources for these signals because if both of these structures are removed, no further saccades are possible. The superior colliculus and frontal eye fields, in turn, receive input from many areas within the visual and parietal corticies, the basal ganglia, and brain structures involved in audition and somatosensation. These areas are presumed to participate in the processes that must precede the decision to make a saccade, processes like attention and target selection.

5.2 Smooth Pursuit

In the smooth pursuit system, signals carrying information about target motion are extracted by motion processing areas in the visual cortex and then passed to the dorso-lateral pontine nucleus of the brainstem. There, neurons have been identified which code either the direction and velocity of pursuit eye movements, the direction and velocity of visual target motion, or both. These signals are passed to the cerebellum where neurons have been shown to encode specifically the velocity of pursuit eye movements. These neurons, in turn, make connections with cells known to be upstream of the nucleus prepositus hypoglossi (the oculomotor integrator described above). As in the other oculomotor systems, the brainstem integrator appears to compute the long-term holding force from this signal and then to pass the sum of these signals to the motor neurons.

5.3 Vergence

A number of different cues are used by primates to determine the distance to a visual stimulus and from this information to compute the convergence angle necessary to fixate a target. The relative positions of a stimulus falling on the two retinas (the binocular

disparity of a stimulus) can be used to compute how much additional convergence or divergence is required to cause a stimulus to fall on the same position in each retina. The state of the lens in the eye can also be used to determine the distance to objects both in and out of focus and thus to compute the required vergence angle. For that matter, even perspective cues can be used to compute a desired vergence angle. In any case, all of these cues are used to activate brainstem circuitry which appears to be very similar to the brainstem circuitry employed by the saccadic system. This brainstem circuit appears to generate a vergence velocity command similar to the saccadic velocity command which is in turn integrated and passed on to the alpha motor neurons. Of course, unlike the commands for versional movements, in this case each eye receives a very different command for both velocity and holding force.

6. Summary

While many important questions remain the focus of active research, the oculomotor system is a comparatively well understood set of neural mechanisms that serves as an excellent model for understanding movement control in general.

See also: Eye Movements and Perception; Eye Movements in Reading; Vestibulo-ocular Reflex, Adaptation of the

Bibliography

Berthoz A, Melvill Jones G (eds.) 1985 *Mechanisms in Gaze Control: Facts and Theories. Reviews of Oculomotor Research.* Elsevier, New York, Vol. 1

Buttner-Ennever J A (ed.) 1988 *Neuroanatomy of the Oculomotor System. Reviews of Oculomotor Research.* Elsevier, New York, Vol. 2

Carpenter R H S 1988 *Movements of the Eyes*, 2nd edn. Pion, London

Carpenter R H S (ed.) 1991 *Eye Movements. Vision and Visual Dysfunction.* CRC Press, Boston, MA, Vol. 8

Collewijn H 1981 *The Oculomotor System of the Rabbit and its Plasticity.* Springer, New York

Fuchs A F, Kaneko C R S, Scudder C A 1985 Brainstem control of saccadic eye movements. *Annual Review of Neuroscience* **8**: 307–37

Glimcher P W 1998 Eye movements. In: Zigmond M J et al. (eds.) *Fundamental Neuroscience.* Academic Press, New York

Melvill Jones G 1991 The vestibular contribution. In: Carpenter R H S (ed.) *Eye Movements.* CRC Press, Boston, MA, pp. 13–44

Kowler E (ed.) 1990 *Eye Movements and their Role in Visual and Cognitive Processes. Reviews of Oculomotor Research.* Elsevier, New York, Vol. 4

Leigh R J, Zee D S 1991 *The Neurology of Eye Movements*, 2nd edn. F A Davis, Philadelphia, PA

Lisberger S G, Morris E J, Tychsen L 1987 Visual motion processing and sensory-motor integration for smooth pursuit eye movements. *Annual Review of Neuroscience* **10**: 97–129

Miles F A, Wallman J (eds.) 1993 *Visual Motion and its Role in the Stabilization of Gaze. Reviews of Oculomotor Research.* Elsevier, New York, Vol. 5

Raphan T, Cohen B 1978 Brainstem mechanisms for rapid and slow eye movements. *Annual Review of Physiology* **40**: 527–52

Robinson D A 1981 Control of eye movements. In: Brooks V B (ed.) *The Nervous System. Handbook of Physiology, Part 2.* Williams and Wilkins, Baltimore, MD, Vol 2, pp. 1275–1320

Sparks D L 1986 Translation of sensory signals into commands for saccadic eye movements: Role of primate superior colliculus. *Physiological Reviews* **66**: 118–71

Wurtz R H, Goldberg M E (eds.) 1989 *The Neurobiology of Saccadic Eye Movements. Reviews of Oculomotor Research.* Elsevier, New York, Vol. 3

P. W. Glimcher

Eye Movements and Perception

The visual world contains more information than can be perceived in a single glance. At any moment in time the visual array that falls on the eyes subtends approximately 180 degrees of visual angle horizontally and 130 degrees vertically, but perception is limited by the anatomical and physiological characteristics of the visual system (see *Visual Perception, Neural Basis of*). The distribution and the properties of rods and cones, for example, limit perception of color and fine spatial detail to the center of gaze and perception of dim light to the visual periphery. Consequently, one's experience of the visual world depends critically on the ability to move the eyes. Eye movements allow one to direct the gaze at objects of interest in the environment and to maintain fixation on an object even when it moves or when the head and body move. Not all eye movements are alike, however; rather, the eyes move in several different ways to achieve different goals. The scientific study of eye movements dates back until at least the mid to late nineteenth century with the pioneering work of von Helmholtz (1866), Dodge (1906), and Huey (1900). Technological advances in the 1970s allowed for very precise measurements of the temporal, spatial, and dynamic properties of eye movements and for the development of sophisticated methods of perceptual investigation. This article reviews the major kinds of eye movements that people make and their consequences for visual perception. For a detailed treatment of the physiological mechanisms underlying the control of eye movements, see *Eye Movement, Control of (Oculomotor Control)*.

1. Types of Eye Movements

One counterintuitive fact about the eyes is that they are almost constantly moving. Although one is generally not aware of it, the eyes frequently make very

small involuntary movements often called physiological nystagmus. Consequently, the optical image that falls on the retina is constantly jittering. Rather than being an impediment to perception, however, these tiny movements are actually a necessity. If the eyes are somehow constrained from moving or equivalently if the image that falls on the back of the eyes is stabilized by some external means then after a few seconds perception of the visual image fades away and nothing is seen at all! Thus, in some sense, without eye movements visual perception would be impossible.

While nystagmus keeps a fresh image of the world on the retina, in many situations it is desirable to minimize retinal image displacement in order to extract high-resolution information from objects in the environment. One way this is achieved is via vestibular eye movements. Vestibular eye movements keep the eyes directed on an object in the world despite movements of the head and body. As the head moves in one direction, the eyes rotate in the opposite direction just enough to keep the fixated object in the center of gaze. Without this ability even small movements of the head would cause the optical images of objects in the world to be smeared across the retina, making their recognition difficult.

Optokinetic eye movements often occur in conjunction with vestibular eye movements and they serve the same purpose: maintaining fixation on an object in the visual world. The optokinetic reflex refers to the fact that when a large portion of the visual field moves across the retina, the eyes reflexively fixate one object in the world and track it slowly for some period of time before moving rapidly back in the opposite direction to begin fixation on another object in the visual world.

Smooth pursuit eye movements are used to track the position of a single moving object in the visual field. This enables one to examine the moving object for an extended time, aiding object identification. However, the eyes are able to keep up with the moving object only as long as it does not move too quickly (usually 20–30 degrees per second).

Vergence eye movements allow one to fixate objects at different distances so that their optical images fall in the center of gaze. The two eyes converge so that they both point at the object in the world, and the degree of convergence required is a cue to depth. Vergence eye movements also allow one to track objects moving in depth.

The major purpose of the eye movements described so far is to maintain the optical image of an object in the world in central gaze. However, another reason for moving the eyes is to acquire new information. This is necessary because the ability to resolve spatial detail (visual acuity) is best at the center of vision and drops off precipitously with increasing eccentricity. Thus, saccadic eye movements are made in order to bring new objects into the center of gaze. Saccades are made while reading, while viewing pictures, and while exploring the visual environment. Via saccades the point of gaze is rapidly and abruptly shifted from one position in space to another. It takes approximately 150–200 milliseconds (ms) to plan a single saccade, but the movements themselves are extremely rapid. Movement time depends on saccade length, but the average saccadic movement lasts approximately 30 ms. Between movements the eyes fixate for a variable period of time on an object of interest in the environment before moving on. The average fixation duration is approximately 250–300 ms. In sum, saccadic eye movements occur approximately three times each second when one is scanning the environment.

2. *The Perception of a Stable Visual World Across Saccades*

One of the classic problems in visual perception concerns how it is that people ordinarily perceive the visual world as unified, stable, and continuous despite eye movements. Research addressing this question has concentrated primarily on the perceptual consequences of saccades. Although saccades facilitate the exploration of the visual environment they also create several problems for visual perception. One problem is that during a saccade visual stimulation is swept across the back of the eyes at a very high speed, producing a blur or a smear that is not ordinarily perceived because of saccadic suppression. Saccadic suppression is largely a consequence of visual masking: the clear, bright, long-duration fixations that precede and follow each saccade mask the perception of the low-contrast blur that is on the retina in much the same way that a bright flash of light would mask the perception of any low-contrast, brief-duration stimulus. Because of saccadic suppression visual information about the world is registered in isolated glimpses that are separated in time.

A second problem is that the retinal positions of objects in the world change suddenly and dramatically when the eyes move, so that an object that appears in the periphery in one fixation suddenly may appear at the center of gaze in the next fixation. Thus, saccades result in visual input that is temporally discontinuous and rapidly changing. Despite this, the visual world appears to be unified, stable, and continuous, with objects maintaining their positions in space. This quality of perception suggests that the perceptual system computes the spatial coordinates of objects in the world and maintains this information across eye movements to produce the experience of a stable and continuous visual world. How this is accomplished has intrigued psychologists and vision researchers for over a century (see Bridgeman et al. 1994 for a review).

Much of the research addressing this question has concentrated on why objects in the world appear to maintain a constant direction with respect to the viewer even though their retinal positions change

when the eyes move (see also *Neural Representations of Direction (Head Direction Cells)*). This phenomenon is usually called visual direction constancy. Traditionally, theories of visual direction constancy have assumed that some kind of oculomotor information about eye position is used in conjunction with retinal (i.e., visual) information to allow the viewer to compensate for the changes in visual stimulation that occur during an eye movement. Some have claimed that this oculomotor information consists of a copy of the efferent commands that are sent to the eyes to move them while others have argued that it consists of inflowing afferent information about eye position from muscle spindles in the oculomotor muscles. Hybrid solutions have been proposed as well in which direction constancy is achieved via an interaction between outflowing gamma efferent signals and inflowing afferent signals. Evidence for these theories comes largely from demonstrations that direction constancy breaks down when there is a mismatch between oculomotor information and visual stimulation. For example, people with oculomotor paralysis report that the perceived world seems to jump when they attempt to move their eyes, and passive displacements of the eyes (produced by pressing on the side of the eye, for example) produce a perception of retinal image movement.

These traditional theories of direction constancy have been called into question by experiments that have examined the spatial coding of objects across changes in eye position. Perceptual research examining people's ability to localize visual flashes that are presented near the time of a saccade has shown that any oculomotor information that exists is too slow and too imprecise to account for the stability that is perceived across saccades. Other studies have shown that stimulus displacements that occur during saccades are often unnoticed, which suggests that the compensatory oculomotor signal accompanying the saccade must be rather inaccurate. Visual direction constancy is also greatly influenced by the presence of visual landmarks in the environment; for example, the spatial localization errors that occur when the eyes are paralyzed are experienced only in the dark and not in normally illuminated visual surrounds. Similarly, the illusory displacements of the visual world that are generated by passive displacements of the eyes are greatly reduced in the presence of a stable visual background. Based on the results of such studies, it appears that visual direction constancy relies not on oculomotor information alone, but rather on some combination of oculomotor information and relative visual position information, with the visual information dominating the oculomotor information in many circumstances.

More recent research has addressed the role of memory in the construction of a stable and continuous visual environment (see Irwin 1996 for a review). At one time it was hypothesized that the visible contents of successive eye fixations were spatially reconciled across eye movements (using the same sources of information that produce direction constancy) and superimposed according to environmental coordinates in a low-level memory buffer to produce a unified representation of the visual environment. However, extensive research has failed to find evidence for such a memory buffer. In fact, very little information appears to be remembered from one eye fixation to the next. Thus, the experience of a stable and continuous visual environment appears to arise from the contents of individual eye fixations, rather than from a detailed mental representation that is constructed across eye movements.

See also: Eye Movement, Control of (Oculomotor Control); Eye Movements in Reading; Neural Representations of Direction (Head Direction Cells); Vestibulo-ocular Reflex, Adaptation of the; Visual Perception, Neural Basis of

Bibliography

Bridgeman B, Van der Heijden A H C, Velichkovsky B M 1994 A theory of visual stability across saccadic eye movements. *Behavioral and Brain Sciences* **17**: 247–92

Dodge R 1906 Recent studies in the correlation of eye movement and visual perception. *Psychological Bulletin* **3**: 85–92

Hallett P E 1986 Eye movements. In: Boff K R, Kaufman L, Thomas J P (eds.) *Handbook of Perception and Human Performance*. Wiley, New York

Huey E B 1900 On the psychology and physiology of reading. *American Journal of Psychology* **11**: 283–302

Irwin D E 1996 Integrating information across saccadic eye movements. *Current Directions in Psychological Science* **5**: 94–100

Rayner K 1998 Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* **124**: 372–422

von Helmholtz H 1866 *Handbuch der physiologischen Optik* [trans. Southall J P C]. 1925 *Treatise on Physiological Optics*. Optical Society of America, New York

D. E. Irwin

Eye Movements in Reading

Reading is arguably the most important skill that human beings acquire during their school years. An integral part of the reading process, both for beginning readers and skilled readers, is the eye movements that occur when we comprehend text. In this article, some basic facts about eye movements during reading are discussed. Developmental changes in eye movements are also discussed, as are the following topics: the perceptual span (or span of effective vision), the

integration of information across fixations, and the control of eye movements.

1. Basic Facts About Eye Movements During Reading

Although it seems as though we move our eyes smoothly across the text when we read, this is an illusion since our eyes make a series of movements (called *saccades*) that last about 20–35 ms. Between the saccades, our eyes are relatively still for about 200–250 ms (these periods when the eyes are stationary are called *fixations*). No information is obtained during saccades because the eyes are moving so quickly across the text; all new information is thus encoded during the fixations. Reading is therefore a bit like a slide show where each new slide (with a different view of the text) appears every quarter of a second or so. The saccade is typically 7–9 letter spaces and is not affected by the size of the print as long as it is not too small or too large. The appropriate metric to use when discussing eye movements therefore is letter spaces, and not visual angle (generally, 3–4 letter spaces are equivalent to 1 degree of visual angle). The other primary characteristic of eye movements is that about 10–15 percent of the time readers move their eyes back in the text (*regressions*) to look at material that has already been read.

Eye movements during reading are necessary because of acuity limitations in the visual system. With respect to the fixation point, any line of text can be divided into three regions: foveal, parafoveal, and peripheral. In the foveal region (extending 1 degree of visual angle to the left and right of fixation), acuity is sharpest and the letters can be easily resolved. In the parafoveal region (extending to 5 degrees of visual angle on either side of fixation) and the peripheral region (everything on the line beyond the parafoveal region), acuity drops off markedly so that our ability to identify letters is not very good even in the near parafovea. The purpose of eye movements in reading therefore is to place the fovea on that part of the text to be processed next. Whereas a majority of the words in a text are fixated during skilled reading, many words are skipped so that foveal processing of each word is not necessary. For example, content words are fixated about 85 percent of the time while function words are fixated about 35 percent of the time. Function words are fixated less frequently than content words because they are generally quite short and there is a relationship between word skipping and word length: as length increases, the probability of fixating increases; whereas 2–3 letter words are fixated less than 25 percent of the time, words that are 8 letters or longer are almost always fixated (and often fixated more than once).

When reading passages of text that are hard or unfamiliar, the eye movements of skilled readers are somewhat different from when they read text that is easy or familiar. As text difficulty increases, there are systematic effects on eye movements: fixation durations increase, saccade lengths decrease, and regression frequency increases. An important fact about eye movements in reading is that the ease or difficulty associated with understanding the text strongly influences the durations of fixations, the length of saccades, and the frequency of regressions.

It is important to note that the values presented above for fixation duration, saccade length, and regression frequency are averages and there is considerable variability in all of the measures. Thus, although the mean fixation duration is 200–250 ms and the mean saccade length is about 8 letter spaces, for individual readers these values might be somewhat higher or lower. Thus, the mean fixation duration for individual readers ranges between 200 ms and 300 ms and the average saccade length ranges from 6 to 10 letters. This between-reader variability (which also exists for regression frequency) is perhaps not as important as the fact that there is considerable within-reader variability. In other words, although a reader's mean fixation duration is 250 ms, the range of individual fixation durations can be from under 100 ms to over 500 ms within a passage of text. Likewise, the variability in saccade length can range from 1 letter to over 15 letter spaces (though such long saccades typically follow regressions).

While fixation durations are typically examined when global aspects of reading are considered, when more detailed analyses involving individual target words are necessary, measures other than the mean fixation duration are usually examined. If readers always made one and only one fixation on a word then the fixation duration on the word would be the unit of processing. However, words are sometimes fixated more than once, and sometimes they are skipped. Thus, using the mean fixation duration to infer the processing time for a word is inadequate because it underestimates the time the eyes are actually on the word (that is, a 200 ms fixation and a 150 ms fixation on the same word would yield a mean fixation duration of 175 ms). Likewise, using only words that are fixated once (*single fixation duration*) is problematic because many words are skipped. Thus, the two most frequently used measures of processing time for a word are *first fixation duration* (the duration of the first fixation on a word independent of the total number of fixations) and *gaze duration* (the duration of all fixations on a word prior to an eye movement to another word).

2. Developmental Changes in Eye Movements

When children learn to read, they must learn to systemically move their eyes through text. Children in

first grade typically fixate for 300–400 ms and only move their eyes a few characters. Thus, they fixate on virtually every word in the text. However, as reading skill develops, eye movements become more like those of skilled readers. By the time children are in fourth or fifth grade, the basic characteristics of their eye movements are like those of skilled readers. While their fixation durations and saccade lengths are similar to those of skilled readers, they do make more regressions than skilled readers. Children who experience severe difficulties learning to read (*dyslexic* readers) do not follow this systematic progression in the development of eye movement characteristics. Indeed, adult dyslexic readers make longer fixations, shorter saccades, and more regressions than skilled readers.

3. The Perceptual Span

How much information does a skilled reader acquire on each fixation and what is the size of the perceptual span (or area of effective vision) on each fixation? In order to investigate this question, the *eye-contingent display change* paradigm (McConkie and Rayner 1975, Rayner 1975, Rayner and Bertera 1979) was developed. In this paradigm, a reader's eye movements are monitored (generally every millisecond) by a highly accurate eye-tracking system. The eyetracker is interfaced with a computer which controls the display monitor from which the reader reads and changes in the text are made contingent on the location of the reader's eyes. Generally, the display changes are made during saccades and the reader is not aware of the changes.

There are three primary types of eye-contingent paradigms: the *moving window*, *foveal mask*, and *boundary* techniques. With the moving window technique, on each fixation a portion of the text around the fixation point is available to the reader. However, outside of this window, the text is replaced by other letters, or by Xs (see Fig. 1). When the reader moves his or her eyes, the window moves with the eyes. Thus, wherever the reader looks, there is readable text within the window and altered text outside the window. The rationale with the technique is that when the window is as large as the region from which a reader can normally obtain information, reading will not differ from when there is no window present. The foveal mask technique is very similar to the moving window paradigm except that the text and replaced letters are reversed. Thus, wherever the reader looks, the letters around the fixation are replaced by Xs while outside of the mask area the text remains normal (see Fig. 1). Finally, in the boundary technique, an invisible boundary location is specified in the text and when the reader's eye movement crosses the boundary, an originally displayed word or letter string is replaced by

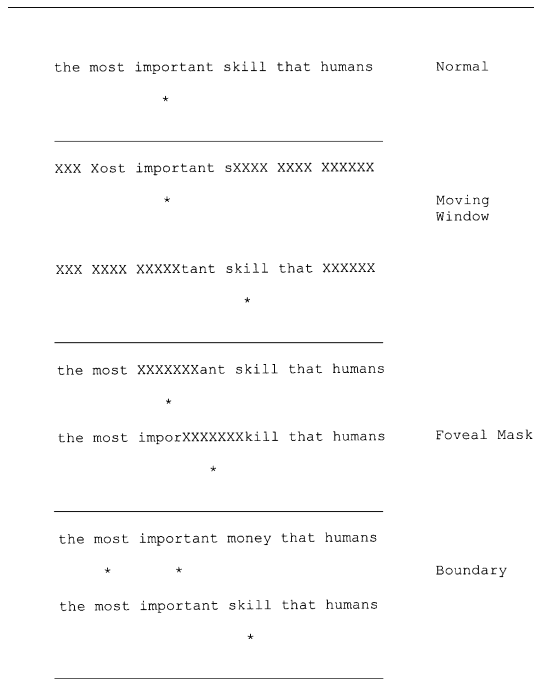

Figure 1
Examples of the moving window, foveal mask, and boundary paradigms. The first line shows a normal line of text with the fixation location marked by an asterisk. The next two lines show an example of two successive fixations with a window of 15 letter spaces (7 letter spaces to each side of fixation) and the other letters replaced with Xs (and spaces between words preserved). The next two lines show an example of two successive fixations with a 7-letter foveal mask. The bottom two lines show an example of the boundary paradigm. The first line shows a line of text prior to a display change with fixation locations marked by asterisks. When the reader's eye movement crosses an invisible boundary (the last letter is important), an initially displayed word (money) is replaced by the target word (skill). The change occurs during the saccade so that the reader does not see the change

a target word (see Fig. 1). The amount of time that the reader looks at the target word is computed as a function of (a) the relationship between the initially displayed stimulus and the target word, and (b) the distance that the reader was from the target word prior to launching a saccade that crossed the boundary location.

For readers of English (and other alphabetic writing systems printed from left-to-right), the span extends from the beginning of the currently fixated word, or roughly 3–4 letter spaces to the left of fixation, to about 14–15 letter spaces to the right of fixation. The span is thus asymmetric—it extends further to the

right of fixation than to the left of fixation. For readers of languages printed from right-to-left (such as Hebrew), the span is asymmetric but in the opposite direction from English so that it is larger left of fixation than right. While the span is asymmetric, it is the case that no useful information is acquired below the line of text that is currently being read as readers apparently focus their attention only on the line being fixated.

Although the perceptual span extends about 14–15 letter spaces to the right of fixation, the area from which words can be identified on a given fixation (the word identification span) generally does not exceed 7–8 letter spaces to the right of fixation. However, neither the perceptual span nor the word identification span is fixed as both can be modulated by word length. For example, if three short words occur in succession, readers are able to identify all of them. If the upcoming word is highly constrained by the context, readers acquire more information from that word than from unpredictable words, and if the fixated word is difficult to process, readers obtain less information from the upcoming word.

Orthography also influences the size of the perceptual span. Hebrew readers have a smaller span than English readers, and Japanese and Chinese readers have even smaller spans. Hebrew is a more densely packed language than English, and Japanese and Chinese are both more densely packed than Hebrew; densely packed refers to the fact that it takes more letters per sentence in English than Hebrew, for example. Finally, reading skill influences the size of the perceptual span. Beginning readers (at the end of second grade) have a smaller span than skilled readers and adult dyslexic readers have smaller spans than skilled readers.

4. *Integration of Information Across Eye Fixations*

It is clear from research using both the moving window and the boundary technique that there is a *preview* benefit from having the word to the right of fixation available on an eye fixation, and that information obtained about the parafoveal word on fixation n is combined with information on fixation $n + 1$ to speed identification of the word when it is subsequently fixated. What kind of information is integrated across consecutive eye fixations in reading?

Experiments using the boundary paradigm described above have varied the orthographic, phonological, morphological, and semantic similarity between an initially displayed stimulus and a target word in attempts to determine the basis of the preview effect. A number of major findings have emerged from these studies. First, there is facilitation due to orthographic similarity as having a preview of *chest* facilitates the processing of *chart*. However, the facilitation is not strictly due to visual similarity because changing the case of letters from fixation to fixation (so that *ChArT* becomes *cHaRt* on the next) has little effect on reading behavior. Second, the facilitation is in part due to abstract letter codes associated with the first few letters of an unidentified parafoveal word; although there may be some facilitation from other parts of the word to the right of fixation, the bulk of the preview effect is due to the beginning letters. The effect is not simply due to spatial proximity because there is facilitation from the beginning letters of words when readers are asked to read sentences from right to left, but with letters within words printed from left to right. Third, there is facilitation due to phonological similarity, since *beech* facilitates *beach* and *shoot* facilitates *chute*, with less facilitation in the latter than in the former case. Fourth, although morphological factors can influence fixation time on a word, they are not the source of the preview benefit. Finally, there is no facilitation due to semantic similarity since *song* as the initial stimulus does not facilitate the processing of *tune*. It thus appears that orthographic (abstract letter) and phonological codes are the basis of the preview benefit and used in integrating information across fixations.

5. *Eye Movement Control: Where to Fixate Next*

There are two components to the issue of how eye movements are controlled during reading: (a) where to fixate next and (b) when to move the eyes. It appears that there are separate mechanisms involved in these decisions and they will accordingly be discussed separately.

Word length seems to be the primary determinant of where to fixate next in skilled reading. When length information about the upcoming word is not available, readers move their eyes a shorter distance than when such information is available. Also, the length of the word to the right of fixation strongly influences the size of the saccade. Furthermore, there is a *landing position* effect such that readers tend to fixate about halfway between the beginning and the middle of words. However, there is also a *launch site* effect such that where readers land in a word is strongly influenced by where the saccade came from. Thus, whereas the most frequent landing position is near the middle of the word, if the prior saccade was launched some distance (8–10 letters) from the target word then the landing position shifts to the left of center. Likewise, if the prior saccade was launched close (2–3 characters) to the beginning of the target word, the landing position shifts to the right of center. Given this, the optimal strategy would be to fixate near the middle of each successive word. However, since short words can often be identified when they are just to the right of fixation,

they are often skipped. Factors such as these result in the landing position distribution being spread somewhat.

6. Eye Movement Control: When to Move the Eyes

What determines when we move our eyes? The amount of time a reader fixates on a word or segment of text reveals something about the cognitive processes associated with comprehending that word or segment. One reason for this is that information gets into the processing system very early in a fixation (thus leaving a lot of time for processes associated with word recognition and other necessary comprehension processes). Experiments using the foveal mask paradigm in which the onset of the mask is delayed following a saccade have demonstrated that reading proceeds quite normally if the reader has 50 ms to process the text prior to the onset of the mask. If the mask occurs earlier, reading is disrupted. Although readers typically acquire the visual information needed for reading during the first 50 ms of a fixation, they can extract information at other times during a fixation as needed.

There have been many demonstrations that various lexical, syntactic, and discourse factors influence fixation time on a word. In particular, all of the following variables influence fixation time on a word: (a) word frequency, (b) contextual constraint, (c) semantic relationships between words in a sentence, (d) anaphora and co-reference, (e) lexical ambiguity, and (f) syntactic disambiguation. Thus, for example, readers look longer (the first fixation duration and gaze duration are both longer) at words that do not occur as frequently in the language than they look at more frequent words, and they look at words that are highly predictable for less time than they look at words that are not predictable from the preceding context. In essence, the amount of time that readers look at words is strongly influenced by the ease or difficulty associated with understanding that word. Other factors also influence how long readers look at words. When reading a sentence such as 'While Mary was mending the sock fell off her lap' (referred to as garden-path sentences), readers fixate for a long time on *fell*. This is because they typically parse *the sock* as a direct object and then don't know what to do with *fell* when they get to it. Thus, they need to reanalyze the sentence. In this particular example, readers would not only fixate for a long time on the disambiguating word (*fell*), they would also make a regression back to the earlier part of the sentence where the misanalysis began.

In essence, the variability in fixation durations that was mentioned at the outset of this article is largely due to the ease or difficulty of processing the fixated words. Because the oculomotor system is apparently closely linked to the part of the brain responsible for processing language, eye movements are a very good way to study various questions related to language and comprehension. Some computational models of eye movements in reading (see Legge *et al.* 1997, Reichle *et al.* 1998) have been developed that nicely capture many of the properties of skilled reading.

See also: Eye Movement, Control of (Oculomotor Control); Eye Movements and Perception; Reading, Neural Basis of; Reading Nonalphabetic Scripts; Reading Skills; Speed Reading; Vision for Action: Neural Mechanisms; Vision, High-level Theory of; Vision, Low-level Theory of

Bibliography

Crowder R G, Wagner R K 1992 *The Psychology of Reading*, 2nd edn. Oxford University Press, New York

Frazier L, Rayner K 1982 Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* **14**: 178–210

Just M A, Carpenter P A 1980 A theory of reading: From eye fixations to comprehension. *Psychological Review* **87**: 329–54

Just M A, Carpenter P A 1987 *The Psychology of Reading and Language Comprehension*. Allyn and Bacon, Boston

Legge G E, Klitz T S, Tjan B S 1997 Mr. Chips: An ideal-observer model of reading. *Psychological Review* **104**: 524–53

McConkie G W, Rayner K 1975 The span of the effective stimulus during a fixation in reading. *Perception and Psychophysics* **17**: 578–86

Pollatsek A, Bolozky S, Well A D, Rayner K 1981 Asymmetries in the perceptual span for Israeli readers. *Brain and Language* **14**: 174–80

Rayner K 1975 The perceptual span and peripheral cues in reading. *Cognitive Psychology* **7**: 65–81

Rayner K 1998 Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* **124**: 372–422

Rayner K, Bertera J H 1979 Reading without a fovea. *Science* **206**: 468–9

Rayner K, Pollatsek A 1989 *The Psychology of Reading*. Prentice-Hall, Englewood Cliffs, NJ

Reichle E D, Pollatsek A, Fisher D L, Rayner K 1998 Toward a model of eye movement control in reading. *Psychological Review* **105**: 125–57

K. Rayner

Eyelid Classical Conditioning

Classical conditioning as Pavlov (1927) discovered it involves the creation of an association between a previously neutral stimulus (a bell, a lab coat) and a reflex-evoking stimulus (meat powder) which causes an unconscious reflexive response (salivation). With repeated pairings the neutral stimulus (becoming the

'conditional stimulus' as Pavlov knew it or the conditioned stimulus as we know it) comes to evoke a learned response (the conditioned response) which is similar to but not exactly like the reflexive response elicited by the meat powder (the unconditioned stimulus). A picture of Pavlov's salivating dogs is a common enough image to the public in cartoons and science fiction. Strictly speaking, Pavlovian conditioning is classical conditioning of an autonomic or visceral response.

1. Somatic Conditioning

In contrast, classical conditioning of a somatic response involves a striated muscle reflex rather than smooth or cardiac muscles. The classic example is Twitmeyer's patellar tendon reflex, familiar to most of us in a physical exam: a tap to the knee with a physician's hammer (the unconditioned stimulus) causes a knee jerk reflex (the unconditioned response). The conditioning situation is similar to Pavlov's autonomic conditioning, with a previously neutral stimulus such as a sound acquiring the eliciting properties of the unconditioned stimulus—the sound alone comes to elicit a knee jerk. The difference from Pavlov is that somatic conditioning of the patellar tendon reflex is a skeletal muscle response. For various reasons mostly related to its application, the patellar tendon reflex is awkward to implement. Theoretically, however, any reflexive motor response can be classically conditioned. The eyelid reflex provides a convenient, easily accessible, easily implemented means of studying classical conditioning of a somatic muscle response (e.g., Gormezano et al. 1986, Schneiderman et al. 1962, Woodruff-Pak and Steinmetz 2000). First widely used in studies of classical conditioning in humans, an eye blink response is typically elicited with a puff of air directed at the cornea (the unconditioned stimulus) or shock to the face. Either stimulus reliably elicits an eye blink. Animal studies of eye blink conditioning often use an air puff when the subject is restrained, but under conditions where the animal is freely moving a face shock may be used instead. Studies using human eye blink conditioning favor using an air puff.

2. Essential Properties of Classical Conditioning

It is important to note that classical conditioning using Pavlov's salivation or the eye blink paradigms are structured in the same way and have the same essential properties found in all classical conditioning. That is, in each paradigm there are identifiable conditioned and unconditioned stimuli, unconditioned reflexes and learned conditioned responses. The relationships between these stimuli and responses are also the same. Thus, for example, the conditioned stimulus must precede the unconditioned stimulus (forward conditioning) by at least about 100 ms. The learned conditioned response is very similar to but not identical to the evoked reflexive response. In both types of classical conditioning, acquisition shows a typical sigmoid curve and reaches an asymptotic level. In general, strong stimuli result in faster learning. Subsequent training where the unconditioned response is withheld (extinction procedure) causes diminution of the learned response, spontaneous recovery at the beginning of subsequent days of testing, and eventually the absence of conditioned responses. Additional training that pairs the conditioned and unconditioned stimuli rapidly restores the conditioned response. These are only a few of the salient properties that are shared in both types of classical conditioning.

3. Learning Conditioned Responses

In studies with animals, optimal learning is usually reached in about 130 trials when each trial has the onset of the tone conditioned stimulus preceding the onset of the air puff unconditioned stimulus by 250 ms. The interval between trials might average 30 s. Intertrial intervals of 9 s or less do not support conditioning. In studies with humans, optimal learning is usually reached in about 50 trials when each trial has the onset of the tone conditioned stimulus preceding the onset of the air puff unconditioned stimulus by 500 ms. The interval between trials might average 60 s. Humans must additionally be distracted during conditioning, for example, by watching silent movies. Recent studies show, however, that humans who learn can accurately describe the stimulus contingencies.

Learned responses are defined as eye blinks that occur after the onset of the tone conditioned stimulus and before the air puff unconditioned stimulus on paired trials. A further criterion is that the response must be 0.5 mm or larger to be counted as a response. On test trials where only the tone conditioned stimulus is given, conditioned responses are defined as any 0.5 mm or larger response after the tone until the end of the trial, usually a period of 250 ms after the air puff would have begun. In both cases, reflexive alpha responses that occur too quickly after the tone conditioned stimulus should be excluded as not being true conditioned responses. Typically for rabbits any response latency less than 25 ms is considered to be an alpha response. For humans, any response latency less than 100 ms might be considered to be an alpha response. A common criterion for learning is the first time that eight conditioned responses occur in nine trials (89 percent responding). An additional criterion for stable performance (overtraining) requires continued performance at some level, for example 70 percent responding with learned responses. With rabbits 100 percent responding is often achieved. With humans the rate is usually not perfect. Once learned,

the association shows good retention over months and perhaps years.

Traditionally, the unconditioned response was thought to be stable and impervious to influence by the training stimuli. However, reflex facilitation is seen before conditioned responses appear. Reflex facilitation occurs when the size of the reflexive response to the air puff unconditioned stimulus increases in the presence of the tone conditioned stimulus. It has also been shown that there is sensitization of the air puff unconditioned response. This sensitization is seen as an increase in the size of the air puff unconditioned response after repeated exposure to tone conditioned stimuli and air puff unconditioned stimuli that have never been presented together.

4. *Motor Response in Eyeblink Conditioning*

The fact that the motor response is so well defined is one of the distinguishing features between operant and classical conditioning. The actual motor response in eye blink conditioning involves sensations of the air puff from receptors on the cornea and the region of the face around the eye (Prince 1964). This sensation involves the trigeminal nerve (cranial nerve V) and its projections to the trigeminal nuclei. The trigeminal nuclei most involved are the principal (sensory) and descending (inferior or spinal) trigeminal nuclei. The former is characterized as being involved in the sense of touch and position, while the latter is characterized as being involved in light touch and pain, although these designations overlap. The reflex arc involves projections from the trigeminal nuclei to cranial motor nuclei that project motor neurons to the muscles involved in the eye blink. In its simplest terms the principal motor nucleus for the eye blink is the facial cranial nucleus (cranial nerve VII) which innervates the orbicularis occuli muscles that effect the eye blink. The eye blink involves more than just this connection but in humans this is the pathway usually considered. In animals other pathways may receive most of our attention. In common experimental subjects such as the rabbit and cat which have a 'third eye lid'—the nictitating membrane—the cranial motor nuclei most involved are the abducens (cranial nucleus VI) and the accessory abducens nuclei, which innervate the rectus bulbus muscle in the rabbit. The nictitating membrane response for the rabbit involves retraction of the eyeball into the orbit, which presses Hardner's gland, which in turn causes nictitating membrane extension. In the rabbit, nictitating membrane movement is a passive phenomenon. In the cat, nictitating membrane extension is an active process, where there is a projection to the superior oblique muscles that control the nictitating membrane. In addition to these connections, the eye blink response (especially in the case of the nictitating membrane in the rabbit) involves cranial projections to the extraocular eye muscles through the oculomotor (cranial nucleus III), trochlear (cranial nucleus IV), and abducens (cranial nucleus VI) innervations. Simultaneous activations of these extraocular muscles pull the eye ball into the socket.

5. *Measuring the Eyeblink Response*

From the foregoing it should be obvious that there are a couple of ways to measure the response to a puff of air directed at the eye. The most direct method is to measure the movement of the eyelids or nictitating membrane. This is accomplished by using a transducer—a device to translate movement into an electronic signal. Most commonly, this is accomplished with a minitorque potentiometer with a lever attached to the axle. Although it has been done, researchers are generally hesitant to directly attach a potentiometer lever to the eyelids of human subjects, for fear of poking the subject in the eye. This is especially a concern when the subject's movements are unpredictable, as in Alzheimer's patients. In animal subjects this measurement is made by placing a small loop of suture into the eyelid or nictitating membrane for attachment of the potentiometer lever or by temporarily gluing the potentiometer lever directly to the eyelid. The great advantage of using a potentiometer is that the actual movement is recorded. Since an eye blink presumably performs a protective feature, monitoring the actual eyelid movement gives an indication of its functionality.

An alternative to directly measuring the movement is to record EMG activity from the eyelid. This can be done directly by inserting recording electrodes into the orbicularis muscles, or indirectly by placing surface electrodes over the skin of the muscles. Both methods have been used in animals, but in humans surface electrodes are favored because they are noninvasive. It should be noted, however, that surface electrodes also pick up information about eye position because the anterior chamber of the eye is charged. In any event, the signal obtained through EMG measurements is imprecisely related to the actual movement. One method is to treat EMG as if it were a neural signal and to send it through a height discriminator. The actual shape of the eye blink is surprisingly reasonable with such a crude measure. The greatest advantage, however, is that the latency of the eye blink is very accurate. However, since movement is easier for us to relate to, EMG signals are usually converted electronically into signals that mimic the output of a potentiometer. The advantage is that the measurement appears to be more familiar. The disadvantage is that the electronics involved in the conversion delay the latency measurement depending on the speed of the eye blink (and therefore it is not a constant delay) and the overall travel of the blink is distorted by the filtering characteristics of the conversion.

Probably the most favored method for recording eye blinks in humans involves an indirect measure.

Here, a light is reflected off the eye and the returning signal gives an indication of the position of the eyelid. Two methods are commonly employed. The first method uses visible light emissions from an incandescent source (for example, a flashlight bulb) or a light emitting diode (LED). The advantage of using visible light is that it can be used to continuously monitor the position of the eyelid. The one drawback of this method is that it is affected by ambient light. In particular, 60 Hz noise from fluorescent lighting can be picked up with the light detector. The second method uses infrared (IR) emissions from an IR-LED. The advantage of IR emissions is that they are little affected by ambient light. The disadvantage of IR emissions is that IR emission is heat which can be uncomfortable or even harmful to the eye. To counteract the heating problem, IR emissions are typically turned on and off rapidly to limit the build up of heat. This technique requires electronic circuitry to only measure eyelid activity during the period when the IR device is turned on, a technique which introduces higher frequency noise problems.

There is no perfect measure of the eye blink response. The methods outlined here indicate some of the trade-offs that must go into making a decision about acquiring the data.

6. *A Model System for the Biology of Learning and Memory*

One of the old issues about conditioning is whether the subject needs to make the actual physical response in order to make the association. At one time there were reports in the operant autonomic conditioning literature that paralyzing the skeletal muscles with curare did not prevent the formation of learning, but this could not be replicated later. Instead, somatic mediation (voluntary skeletal movements or relaxation) is the prime factor in supposedly autonomic conditioning. Recently, however, experiments with reversible inactivation of the cranial motor nuclei show that classical conditioning is learned without the subject actually making the eye blink response during acquisition training.

Pavlov himself appreciated the advantages of classical conditioning for studies of the neural bases of learning and memory. The advantages include the experimenter's control over the stimuli for the association and the experimenter's ready identification of the motor responses (Gormezano et al. 1986, Schneiderman et al. 1962, Prince 1964). Classical conditioning is a model system for understanding the biology of learning. Unfortunately, Pavlov failed to localize learning and memory, in large part because he took a 'top-down' approach, starting with the cerebral cortex, which most people incorrectly assume is the repository of even the simplest of memories. The detached spinal cord is capable of forming primitive associations that are classically conditioned. Recently, studies that have taken a 'bottom-up' approach starting with the circuitry for the basic reflexive response have shown that the association for classical conditioning of the somatic muscle response involved in eyelid conditioning is localized to the cerebellum as a type of motor learning (e.g., Thompson et al. 1997).

Today, classical conditioning of the eyelid response is increasingly favored for providing a quick, reliable test of cognitive functioning in human clinical settings (Woodruff-Pak and Steinmetz 2000, Green and Woodruff-Pak 2000). The reasons for this include the clear structure of the paradigm with identifiable control over the stimuli and responses, the extensive behavioral studies on the properties of classical conditioning, and recent advances in identifying neural structures responsible for or participating in classical conditioning.

See also: Autonomic Classical and Operant Conditioning; Classical Conditioning and Clinical Psychology; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Operant Conditioning and Clinical Psychology; Pavlov, Ivan Petrovich (1849–1936)

Bibliography

Gormezano I, Prokasy W F, Thompson R F (eds.) 1986 *Classical Conditioning: III. Behavioral, Neurophysiological, and Neurochemical Studies in the Rabbit*. Erlbaum, Hillsdale, NJ

Green J T, Woodruff-Pak D S 2000 Eyeblink classical conditioning: Hippocampal formation is for neutral stimulus associations as cerebellum is for association-response. *Psychological Bulletin* **126**: 138–58

Pavlov I P 1927 *Conditioned Reflexes*. Oxford University Press, Oxford, UK

Prince J H (ed.) 1964 *The Rabbit in Eye Research*. Charles C Thomas, Springfield, IL

Schneiderman N, Fuentes I, Gormezano I 1962 Acquisition and extinction of the classically conditioned eyelid response in the albino rabbit. *Science* **136**: 650–2

Thompson R F, Bao S, Berg M S, Chen L, Cipriano B C, Kim J J, Thompson J K, Tracy J A, Krupa D J 1997 Associative learning. In: Schmahmann J D (ed.) *The Cerebellum and Cognition*. Academic Press, San Diego, CA

Woodruff-Pak D S, Steinmetz J E (eds.) 2000 *Eyeblink Classical Conditioning*, *Vols*. 1 (*Animal*) *and* 2 (*Human*). Kluwer Academic Publishers, Norwell, MA

D. Lavond

Eyewitness Memory: Psychological Aspects

Of the dozens of events we experience every day, many are later recalled and discussed. However, on rare occasions, one of these may also be the subject of

intense interest, investigation, and analysis by others. These events are often ones in which someone is observed (usually unfamiliar to the observer) engaging in an interaction or activity, such as the perpetration of a crime, that is of interest to law enforcement authorities. The verbal reports by the observer of such an activity form the content of what has historically been called *eyewitness testimony*.

1. Scientific Approaches to the Study of Eyewitness Testimony

The scientific investigation of verbal descriptions of brief events was initiated by experimental psychologists in Germany and North America. Surprisingly, the same cautions and warnings that these researchers expressed 100 years ago concerning the low reliability of eyewitness testimony appear frequently in today's research journals and the criminal justice system. That is, although much has been learned about the variables related to eyewitness testimony, researchers have provided relatively little information as to how to improve its reliability and accuracy. Instead, the lion's share of their efforts has served to maintain as much reliability as possible in the face of potentially contaminating investigative procedures that may follow a witnessed event; for example, the use of suggestive questioning and biased identification procedures.

The types of information of interest to investigators may relate to the actions, words, and characteristics of the perpetrator (including voice and movement) as well as to the details of the context and the roles played by other people. A focus upon the acoustic qualities of voice is sometimes referred to as earwitness testimony (Thompson et al. 1998). Witnesses may also be victims and, as such, their recollections may have been influenced by the demands of the perpetrator and the degree of threat leveled at them. Indeed, in comparison to our recall of more mundane everyday events, it is the surprising, threatening, and emotional impact of criminal acts that renders them of special interest to eyewitness testimony researchers. As a result, one focus of their research has been the level of arousal experienced by witnesses to and victims of crime, particularly when they are children (Goodman and Bottoms 1993). For historical reviews of the growth of the field and the results of early eyewitness testimony research, the reader is referred to Cutler and Penrod 1995, Loftus 1979, Ross et al. 1994, Shepherd et al. 1982, Thompson et al. 1998, and Wells 1993.

1.1 Methodological Limitations to the Study of Eyewitness Testimony

Current understanding of eyewitness testimony derives from two sources: real-world observations of witnesses/victims and research simulations of criminal acts. All would agree that the recollections of real victims and witnesses is the most relevant if we wish to understand the quality and quantity of eyewitness testimony. All would also agree that to assess the accuracy of eyewitness testimony it is necessary to have independent knowledge of the events as they occurred. Unfortunately, in most real-world events this ground truth can never be known because independent records of events in natural contexts are rarely made. To circumvent this problem and to be assured of access to the ground truth, eyewitness researchers often present simulated events to witnesses (staged or on videotape) that can be later recalled. When scientists control the events to be experienced by witnesses, they also have invaluable control over most other variables relevant to the scenario. For example, researchers can manipulate the type and duration of the event, the specific features of the perpetrator(s), the age, gender, and race of the witnesses, the presence of weapons or threat, and the types of recall and identification tests. In real crimes of course, all of these are outside the control of the researcher and assessment of the credibility of a witness's allegation or a defendant's statement about an event falls to triers-of-fact (judges and jurors) in the courtroom.

Depending upon the researcher's creativity, a crime simulation may be very realistic; however the very features that distinguish eyewitness events from mundane real-world events are, for the most part, missing in the laboratory simulations; specifically, surprise, threat, and high emotional involvement. Research participants must give their informed consent before taking part in scientific experiments and ethical considerations preclude the possibility of their being exposed to highly emotional, disturbing, or life-threatening events. For these reasons, virtually all of the simulation work on eyewitness testimony is open to the criticism that it incorporates, to a greater or lesser degree, insufficient ecological validity; that is, the results may not necessarily be generalizable to eyewitness events in the real world. Nonetheless, with appropriate safety and ethical precautions, threat, arousal, and fear have been manipulated in laboratory research (Thompson et al. 1998).

2. Nature of Eyewitness Testimony

The vagaries of human perception, memory, and decision-making were well known to philosophers, scientists, police, and judges long before experimental psychologists initiated their research. However, psychologists did make a unique contribution to this knowledge base: systematic data collection that allowed measurement of the magnitude and frequency of human errors in the recollections of brief events. Most researchers in the 1970s and 1980s were sufficiently impressed by the frequency of these errors that much energy was directed to their documentation.

Whereas none would anticipate perfect performance in either the perception or recall of such events, an unexpected finding was the influence a variety of nonperceptual and nonmemorial variables had upon performance, for example, those of social influence and the form of questions asked of witnesses.

2.1 Recall of Details and Perpetrator Identification

Two categories of eyewitness information of special interest to researchers are the details of the crime and of the perpetrator. It is recognized that the measurement of accuracy of recalled details is highly dependent upon the types of questions used to elicit responses. For example, police investigators regularly interviewed witnesses and victims, yet their interviews were usually not based upon techniques known to enhance recall in cognitive and social psychology. As of 1983 this situation began to change as researchers applied cognitive principles to police interviewing. One result was a specific interview protocol known as the cognitive interview (Fisher and Geiselman 1992) that has well demonstrated its superiority in gathering more correct information from both adults and children without increasing the numbers of errors.

The misidentification of an innocent person is a particularly dramatic kind of eyewitness error because of its serious consequences for a person's freedom. Indeed, legal opinion in most countries has emphasized the dangers of convictions based upon eyewitness testimony evidence alone (Cutler and Penrod 1995) and media attention given erroneous convictions has raised the issue's profile in the public mind. As increasing numbers of demonstrably faulty convictions were studied, it became readily apparent that simple manipulations of photos and test instructions could have large effects upon the identification decisions of eyewitnesses. For example, Loftus (1979) showed that presenting a photo with a unique quality or different alignment in a photospread significantly raised the number of misidentifications of the person depicted in the photo. Analyses of both live and photo lineups in real cases have shown many to be poorly constructed and when so constructed, heavily biased against the police suspect (Wells and Bradfield 1998). Further, minor wording changes in the instructions given witnesses before they view a lineup dramatically raise the rate of false positive errors or selections of an innocent person. Indeed, false positive rates of 50 percent or more are not uncommon and Wells and Bradfield (1998) recently reported that when instructions encouraged witnesses to believe falsely that the perpetrator was in the photospread, all witnesses selected someone.

Similarly, social pressure from an authority figure or from other witnesses can influence lineup decisions. In particular, information may be subtly conveyed from investigators to witnesses by the kinds of questions asked of the witnesses and the investigator's responses to answers given. Wells and Bradfield (1998) have shown that when research participants were informed falsely that their identification choices had been correct, their descriptions of the perceptual qualities of the event itself were dramatically altered, for example, how much time and attention they had directed to the perpetrator, the ease and confidence of their identification, and their willingness to testify in court.

2.2 Types of Variables Studied

Factors of interest to researchers may be classified as either estimator or system variables (Wells 1993). In the real world, *estimator variables* are not under the control of either the criminal investigator or the criminal justice system and their effects upon eyewitness identification accuracy may only be estimated, for example, the age, sex, and race of the witnesses and the lighting conditions at the scene of the crime may be related to reliability of eyewitness testimony but they are inherent to the crime itself. Knowledge of the effects of these variables may assist us to better characterize the average performance levels that are obtained by specific types of people in particular environmental and viewing conditions. On the other hand, system variables are under the control of the criminal justice system, for example, the size, type, and quality of the lineup or photospread, the instructions given the witnesses, and the temporal interval between the crime and the identification task. Wells (1993) has argued vigorously that researchers should dedicate more effort understanding system variables because the accuracy and reliability of eyewitness testimony can be improved by manipulating features of the system itself.

A striking example of this kind of improvement may be seen in the recommendation (Lindsay and Wells 1985, Wells 1993) that lineup or photospread members be presented one at a time to an eyewitness rather than together. Although simultaneous presentation of lineup and photospread members has been normative in North America and the UK for many years, doing so is well known to encourage the use of a relative judgment strategy by which witnesses attempt to choose the person who best fits their memory of the perpetrator. The difficulty with relative judgments is that every lineup or photospread necessarily includes someone who looks most like the perpetrator and, therefore, someone will virtually always be chosen. Mistaken identifications will necessarily comprise a subset of these choices. Indeed, the high levels of false positives reported for identification tests are a likely consequence of just such a strategy.

In contrast, a sequential lineup presents each member one at a time and the goal is to force witnesses to rely upon an absolute judgment strategy. For each lineup member shown, the witness must indicate whether this person matches the representation of the

perpetrator in memory. Once a person or photo has been presented, it may not be viewed again. To further reduce the opportunity to make relative judgments, witnesses are not informed of the number of people in the lineup. Thus, for the witness who attempts to use a relative judgment strategy, there remains the possibility that there may yet be someone who is even more like the perpetrator among the as yet unseen lineup members. In research comparing sequential to simultaneous lineup performance the two presentation procedures yield identical hit rates (correct identifications of the target) for target-present lineups (when the guilty person is included). However, in target-absent lineups (in which the perpetrator is not present), significantly fewer false positives are made with the sequential than the simultaneous lineup. Thus, when police investigators employ a sequential procedure, innocent persons are better protected from misidentifications than with the simultaneous presentation.

3. *Relationship between Accuracy and Confidence*

Analyses of jurors' decision-making processes have indicated that of all the factors influencing their judgments of credibility of a witness, the confidence expressed by the witness is most influential. When witnesses claim to be absolutely certain of their statements their evidence is given more weight than that from other witnesses who expresses less certainty. Assignment of greater evidential weight to the first than to the second group of witnesses would be reasonable if subjective confidence was predictive of accuracy. Although it is clear that the triers-of-fact do believe there is a strong relationship between accuracy and confidence, the research evidence has been equivocal. Whereas a minority of studies has demonstrated moderate to strong relationships between the two variables, the majority have not. Indeed, the obtained relationship has been generally so low as to provide virtually no predictive value from confidence to accuracy. However, as research evidence accumulates from experiments with higher ecological validity, it is possible that the earlier research will be seen as having underestimated the strength of the accuracy-confidence relationship (Lindsay et al. 1998).

Witnesses' subjective confidence is an estimator variable because it reflects a witness's characteristics, such as age and gender. Nonetheless, its expression by a witness may be seen also as a system variable because it can be altered by the criminal justice system. For example, witnesses may be encouraged by others to assign high confidence to their in-court opinions, often much higher than they provided at the time of the identification. As a result, many researchers have argued that some standardized assessment of confidence should be taken by the investigating officer at the time of a witness's identification decision (Wells et al. 1998).

4. *Effects of Post-Event Information and Cognitive Activities*

Many factors influence a person's recall of a brief event and because human memory is assumed to involve both constructive and reconstructive processes, information considered or received by the witness following the event may also contribute to the specifics of what is ultimately recalled (see *Reconstructive Memory*, *Psychology of*). For example, in postevent misinformation studies, participants view a video event, then hear a narrative about it that contains incorrect information about details in the film (e.g., the getaway car was blue rather than green). Later, they are asked to recall details from the original video they viewed. The typical finding is that participants often incorporate information from the narrative by recalling (or recognizing) details that are consistent with the misleading information. Going well beyond distortion of minor details, research participants have also constructed complete but false autobiographical events as a result of similar suggestive misinformation techniques. The ease with which such memories may be manipulated or constructed has contributed to the development of an entire new field of false memory research, a field whose topics often overlap with those of eyewitness testimony research (see *False Memories*, *Psychology of*).

Other post-event mental activities such as rehearsal, verbal coding, and image generation can similarly contribute to altered recollections of the event or person. The Wells and Bradfield (1998) research dramatically demonstrated these kinds of changes as do the detrimental effects of both postevent verbal (Schooler and Engstler-Schooler 1990) and conceptual (Read 1995) rehearsal of events and people. These kinds of retrospective reconstructions or reframing of events are likely to form the basis of much additional research in the field.

5. *Impact of Eyewitness Testimony Research on the Criminal Justice System*

The fruits of eyewitness testimony research since the 1970s may be seen in the publication of two significant documents. The first was a scientific brief commissioned by the American Psychology-Law Society (Wells et al. 1998) concerning best practice guidelines for the collection of eyewitness identification evidence. The second document (NIJ 1999) surveyed eyewitness testimony evidence more broadly and was the outgrowth of an ongoing US National Institute of Justice project reviewing convictions of men for whom DNA testing of forensic evidence could not be completed at the time of their convictions, but for whom such testing could now be accomplished. The first summary of the project (Connors et al. 1996) described 28 men who had been incarcerated for 10–20 years but were

exonerated by DNA testing. In 90 percent of these cases one or more eyewitnesses had falsely identified the person convicted. Many more incarcerated individuals have been similarly exonerated since the Connors et al. (1996) paper was published (Wells et al. 1998). It appears that the cautions expressed 100 years ago by early researchers finally have had an impact upon investigative practices.

See also: Confessions: Psychological and Forensic Aspects; Elaboration in Memory; Expert Testimony; Expert Witness and the Legal System: Psychological Aspects; False Memories, Psychology of; Memory: Autobiographical; Mood-dependent Memory

Bibliography

Connors E, Lundregan T, Miller N, McEwan T 1996 *Convicted by Juries, Exonerated by Science: Case Studies in the use of DNA Evidence to Establish Innocence after Trial*. National Institute of Justice, Alexandria, VA

Cutler B L, Penrod S D 1995 *Mistaken Identification: The Eyewitness, Psychology, and the Law*. Cambridge University Press, New York

Fisher R P, Geiselman R E 1992 *Memory-enhancing Techniques for Investigative Interviewing*. Charles C. Thomas, Springfield, IL

Goodman G S, Bottoms B L (eds.) 1993 *Child Victims, Child Witnesses*. Guilford Press, New York

Lindsay D S, Read J D, Sharma K 1998 Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely (as they do across real-world witnesses). *Psychological Science* **9**: 215–8

Lindsay R C, Wells G L 1985 Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentations. *Journal of Applied Psychology* **70**: 556–64

Loftus E F 1979 *Eyewitness Testimony*. Harvard University Press, Cambridge, MA

National Institute of Justice (NIJ) 1999 *Eyewitness Evidence: A Guide for Law Enforcement*. US Department of Justice, Washington DC

Read J D 1995 The availability heuristic in person identification: The sometimes misleading consequences of enhanced contextual information. Applied *Cognitive Psychology* **9**: 91–121

Ross D F, Read J D, Toglia M P 1994 *Adult Eyewitness Testimony: Current Trends and Developments*. Cambridge University Press, New York

Schooler J W, Engstler-Schooler T Y 1990 Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology* **22**: 36–71

Shepherd J W, Ellis H D, Davies G M 1982 *Identification Evidence: A Psychological Evaluation*. Aberdeen University Press, Aberdeen, UK

Thompson C P, Herrmann D, Read J D, Bruce D, Payne D G, Toglia M P 1998 *Eyewitness Memory: Theoretical and Applied Perspective*. Erlbaum, Mahwah, NJ

Wells G L 1993 What do we know about eyewitness identification? *American Psychology* **48**: 553–71

Wells G L, Bradfield A L 1998 'Good, you identified the suspect': Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology* **83**: 360–76

Wells G L, Small M, Penrod S, Malpass R S, Fulero S M, Brimbacombe C A E 1998 Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior* **22**: 603–47

J. D. Read

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

F

Face Recognition Models

The human face is a highly meaningful stimulus that provides us with diverse information for adaptive social interaction with people. Our ability to recognize faces is remarkably accurate and long lasting. We are also able to categorize people along a number of visual dimensions including sex, race, and age and can readily interpret facial expression. The challenges associated with encoding and interpreting this information have become evident over the last two decades as psychologists, computer scientists, and cognitive scientists have endeavored to formulate computational models of these processes. The resultant models give insight into the complexity of the problems solved by the human brain in perceiving, representing, and remembering faces. In this article, computational approaches to modeling the perception, categorization, and recognition of human faces will be presented. The properties of the human face as a visual stimulus are described first, followed by definitions of the relevant tasks we perform with faces. The steps involved in modeling these tasks are reviewed next, and representative approaches for modeling individual tasks are discussed. Finally, the article closes with a few open questions in face recognition modeling.

1. *The Human Face as a Visual Stimulus*

The human face is a complex three-dimensional object defined by the structure of the skull and by the shape, texture, and pigmentation of the overlying skin and tissue. All faces share a basic set of features (e.g., eyes, nose, and mouth, etc.) arranged in a well-defined configuration (eyes above the nose, etc). Individual faces comprise virtually limitless variations on this standard theme. To recognize an individual from a face, we must attend to the information that makes the face unique. To categorize a face we must extract and encode the information that a face shares with an entire category of faces (e.g., male faces), but which distinguishes the category from competing categories (e.g., female faces) (see *Face Recognition: Psychological and Neural Aspects*).

2. *The Tasks*

'Face recognition' models commonly encompass a range of tasks, including recognition, identification, verification, categorization, and the analysis of facial expression (see *Facial Expressions*). Face recognition refers to the judgment of whether or not a particular face is 'known.' Face identification refers to the retrieval of information about the 'owner' of the face, such as a name or context of encounter. Face verification refers to a decision about whether a particular face image belongs to a particular individual. Is this person John Doe? Face verification is a common goal of face algorithms developed for security systems.

3. *Modeling: A Step by Step Approach*

Face recognition models involve: (a) preprocessing algorithms to encode facial 'features' and (b) the application of this information to solve particular tasks.

3.1 *Preprocessing Algorithms*

3.1.1 Aligning faces. All models that involve the analysis of a three-dimensional object from a two-dimensional image begin with the process of aligning the images into a common coordinate system. This facilitates feature extraction and comparison. Most current face recognition models operate effectively only with frontal images, tolerating only minimal changes in viewpoint. The alignment procedure employed in different models varies both in precision and in the degree of automaticity with which it is accomplished (i.e., by hand or by a computer algorithm). At the most basic level, alignment involves image translation, rotation, and scaling procedures implemented to assure that the eye levels are equivalent and that the centers of the foreheads correspond. More precise alignment is possible with *morphing* techniques that 'warp' individual faces into the 'average face' (Craw and Cameron 1991). To morph a face into another face (e.g., the average face), control points are located on the two faces (usually by hand). These consist of facial landmarks (e.g., corners of the eyes) and supplemental points (e.g., equally spaced points along the eyebrows). Using these points as a guide, each face is warped into the shape of the average face, yielding a correspondence of the control points across all faces. This alignment enables a separable encoding of the two-dimensional shape of the face and the image intensity information. Automated solutions to this correspondence problem have been implemented using all of the pixels or surface samples of

the face rather than just a subset (Beymer and Poggio 1996, Blanz and Vetter 1999). These algorithms employ elaborated optic flow computations and work well on sets of faces for which correspondence is relatively easy to establish, (e.g., faces without hair that are pre-aligned with the translation method). Though difficult to achieve, when successful, complete alignment provides a powerful basis for synthesizing faces with arbitrary shapes and faces composed of intensity composites of other faces (Blanz and Vetter 1999).

A different approach to alignment is represented by the work of Lades et al. (1993) who developed a face recognition algorithm base on the dynamic link architecture. This algorithm combines alignment with identification. The model operates by placing a deformable grid over the target face, sampling the face at the grid vertices. The sampling is done with a series of oriented Gabor wavelets, designed to emulate the orientation specific neurons of visual cortex. The connectors between the vertices are allowed to deform elastically, enabling a resampling of the image until the best fit is obtained. The deformation parameters of this fit serve as the face representation, which is matched to the faces in the database to identify the best match.

3.1.2 Encoding and representing faces. The information in the aligned faces must be quantified in a way that enables recognition, identification, verification, categorization, and the analysis of expression in the model. What are the features of the face? We commonly think of the features of a face as its eyes, nose, and mouth. Descriptions of these features, such as those an eyewitness might provide, are inadequate for communicating enough information about an individual face to distinguish it from competing candidates. Geometrical measures, e.g., distance between eyes, have proved similarly inadequate (Laughery et al. 1981). More recent models have employed relatively raw perceptual codes, including roughly aligned images and three-dimensional surfaces, including pigmentation information. Another code, common since the advent of morphing technology, involves a two-component separable encoding of the two-dimensional face shape and the image intensities. The 'shape' part of this code is defined as the deformation of the control points from the control points in the average face. The 'shape-free' part of the code consists of a 'shape standardized' two-dimensional array of image intensities created by warping an individual face into the shape of the average face.

In current computational and psychological models of face recognition, further analysis of these perceptual codes is carried out using a principal component analysis (PCA) (Sirovich and Kirby 1987; for a review see Valentin et al. 1994). In the US Government's tests of automatic face recognition algorithms between 1994–7, five of the seven algorithms tested used PCA. PCA is a statistical method for describing a set of correlated variables using a smaller number of uncorrelated or orthogonal variables. The uncorrelated variables are called eigenvectors or principal components (PCs), denoted u_i and play the role of 'features' for describing the faces. PCs can be considered features in the sense that any individual face, f, can be expressed as a linear combination of the PCs, $\Sigma_i w_i u_I$, where the weights are the dot products, $w_I = u_I^T f$, between the faces and PCs. Because PCA is applied usually to images/surfaces, the PCs are also images/surfaces. Thus, individual faces can be synthesized as a linear combination of the PC images/surfaces. In geometrical terms, the PCA creates a multidimensional space in which the PCs define the axes of the space and individual faces are points in the space. The coordinates of a face in the space are the weights that specify the face's value on each PC feature. Note also, that three-layer back propagation networks can extract facial features similarly when they are trained to reconstruct faces through a bottleneck of hidden units. The hidden units of these auto-encoders have been shown to derive rotated versions of the PCs space (Cottrell et al. 1987).

PCA has appeal as a psychological model of face perception and memory for several reasons. First, it is consistent with psychological theories that posit a 'face space' metaphor for human face memory (Valentine 1991). By this metaphor, faces can be thought of as points in a multidimensional space, with the distance between faces a measure of their similarity. At the center of the face space is the average or 'prototype' face. The prototype, a central concept in psychological studies of face recognition, is invoked to explain the role of face typicality in predicting recognition performance. Typical faces, thought to be close to the prototype, are recognized less accurately than distinctive faces. This occurs presumably due to the greater density of faces close to the prototype, causing more confusion among typical faces than among distinctive faces. The prototype is also used as a reference face in creating automatic caricatures. Caricatures can be created by 'moving a face' away from the average in the face space. This results in a more distinctive and recognizable version of the same face.

Second, the features that emerge from PCA are derived from the experience of the model. The role of experience in face recognition performance has been established perhaps most clearly in the phenomenon of the 'other-race effect'—the finding that people recognize faces of their own-race more accurately than faces of other-races. This effect is predicted when the PCA is applied to a majority of faces of one race, and a smaller number of faces of other races. Because PCA derives its features from the statistical structure of the input faces, the resultant features are most appropriate for describing the majority race of faces. Consequently, less distinct encodings of minority race faces

result because these faces are not well characterized by the features extracted primarily from the majority race of faces.

4. Tasks

4.1 Recognition

The quality of the stimulus representation determines the difficulty of the recognition or classification task. With a PCA-based representation, face recognition models can be implemented in a relatively simple way. A face is considered 'known' when an image of the individual was part of the input used to create the PCA space. The most common 'recognition' algorithm implements both recognition and identification. A target face is projected into the space and the distance to all other faces in the space is assessed. The nearest neighbor is chosen as the identity of the target face. Recognition can be implemented by setting a threshold distance, beyond which a target face is declared 'unknown.' An alternative and computationally more expedient algorithm for recognition assesses the representation error incurred by projecting the target face into the space. A threshold tolerance for error is used to determine whether a target face is known or novel.

4.2 Categorization

To categorize faces by sex, race, or age, individual exemplar faces must be assigned to different categories based on visually accessible facial features. Face categorization has been approached with supervised connectionist or neural network classifiers, such as the perceptron (see *Perceptrons*). The models use examples to learn the mapping between face representations and categories. Numerous sex categorization models have been implemented and have been found to perform at or near human performance levels. A similarly structured race classifier has also been implemented, though the imbalance of experience most people have for the faces of different races must be implemented also to model human performance accurately. Finally, little work has been done on categorizing faces by age, though two complementary models of facial aging make use of morphing and caricaturing techniques, respectively. The former simulates aging by morphing individual faces toward the average of older faces (Burt and Perrett 1995). The latter simulates aging by caricaturing the three-dimensional head structure relative to a mean of young adult faces. Surprisingly, this results in an aged face (O'Toole et al. 1997).

4.3 Facial Expression Analysis

Models for categorizing faces by expression have been implemented in ways similar to sex and race classification models, but with somewhat less success. These models operate by mapping images of faces on to expression categories using supervised learning techniques. Representations have varied from aligned images, to PCAs of faces preprocessed by the Gabor wavelet filters described previously. Performance has been found to be well above chance, though still short of human performance on a similar task with similar stimuli. Facial expression analysis is currently a very active area of research and more published work on this problem is expected in the near future.

5. Open Questions

Despite the clear successes of face recognition models in the 1980s and 1990s, the problem of recognizing faces from different viewpoints remains an unsolved challenge for models. Though part and parcel of the larger unsolved inverse optics problem of computer vision, the domain of faces may be more accessible due to the specific nature of face recognition as a within category problem. Some promising lines of research have begun and may soon yield new insights into this difficult problem (Edelman 1999).

See also: Linear Algebra for Neural Networks; Object Recognition: Theories; Recognition Memory, Psychology of; Vision, Low-level Theory of; Visual Perception, Neural Basis of

Bibliography

Beymer D, Poggio T 1996 Image representations for visual learning. *Science* **272**: 1905–9

Blanz V, Vetter T 1999 A morphable model for the synthesis of 3D faces. *ACM SIGGRAPH Proceedings*, pp. 187–94

Burt D M, Perrett D I 1995 Perception of age in adult Caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society London B* Bio **259**: 137–43

Cottrell G W, Munro P, Zipser D 1987 Learning internal representations of gray scale images: An example of extensional programming. In: *Proceedings of the 9th Annual Cognitive Science Society*, Erlbaum, Hillsdale, NJ

Craw I, Cameron P 1991 Parameterizing images for recognition and reconstruction. In: Mowforth P (ed.) *Proceedings of the British Machine Vision Conference*. Springer-Verlag, London

Edelman S 1999 *Representation and Recognition in Vision*. MIT Press, Cambridge, MA

Lades M, Vorbrueggen J C, Buhmann J, Lange J, von der Malsburg C, Wiskott R P, Konen R W 1993 Distortion invariant face recognition in dynamic link architectures. *IEEE Transactions on Computers* **42**: 300–11

Laughery K, Rhodes B, Batten G 1981 Computer-guided recognition ad retrieval of facial images. In: Davies G, Ellis H, Shepherd J (eds.) *Perceiving and Remembering Faces*. Academic Press, London

O'Toole A J, Vetter T, Volz H, Salter E M 1997 Three-dimensional caricatures of human heads: Distinctiveness and the perception of facial age. *Perception* **26**: 719–32

ff

ffi

ff

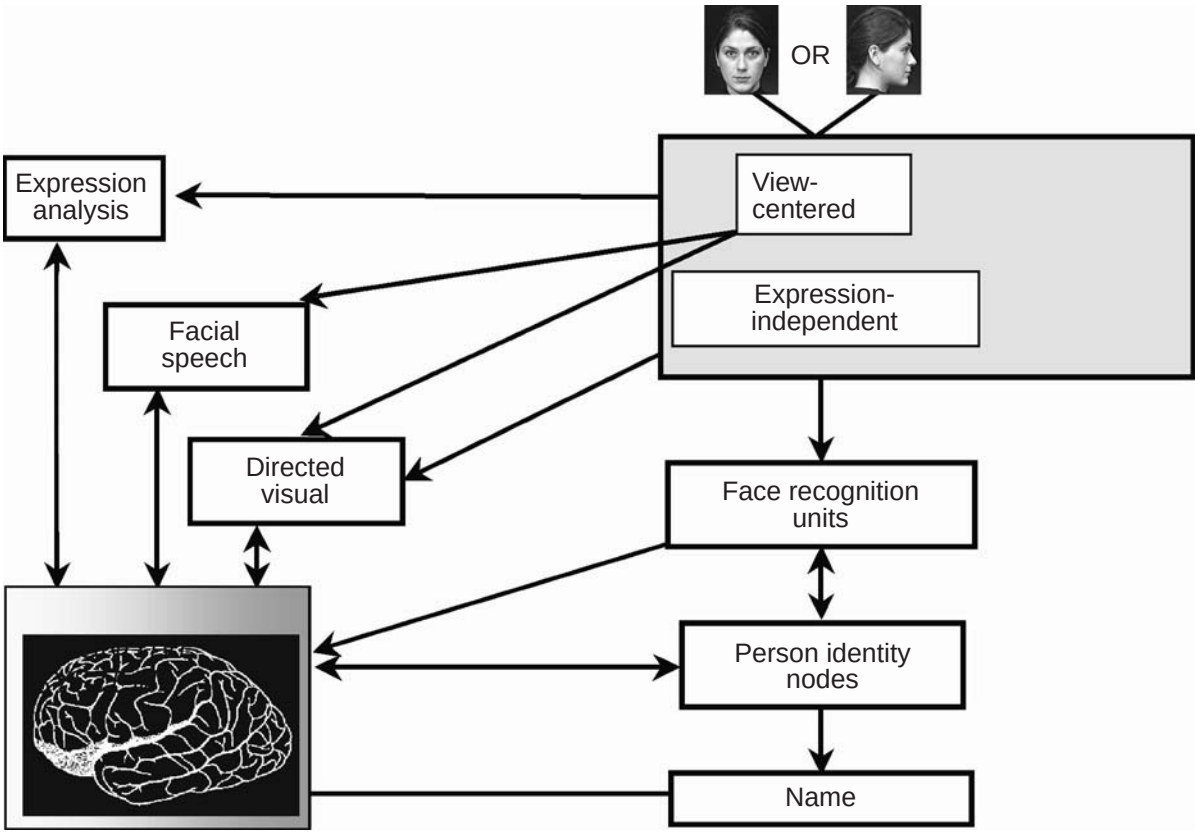

Figure 1

ff

ff

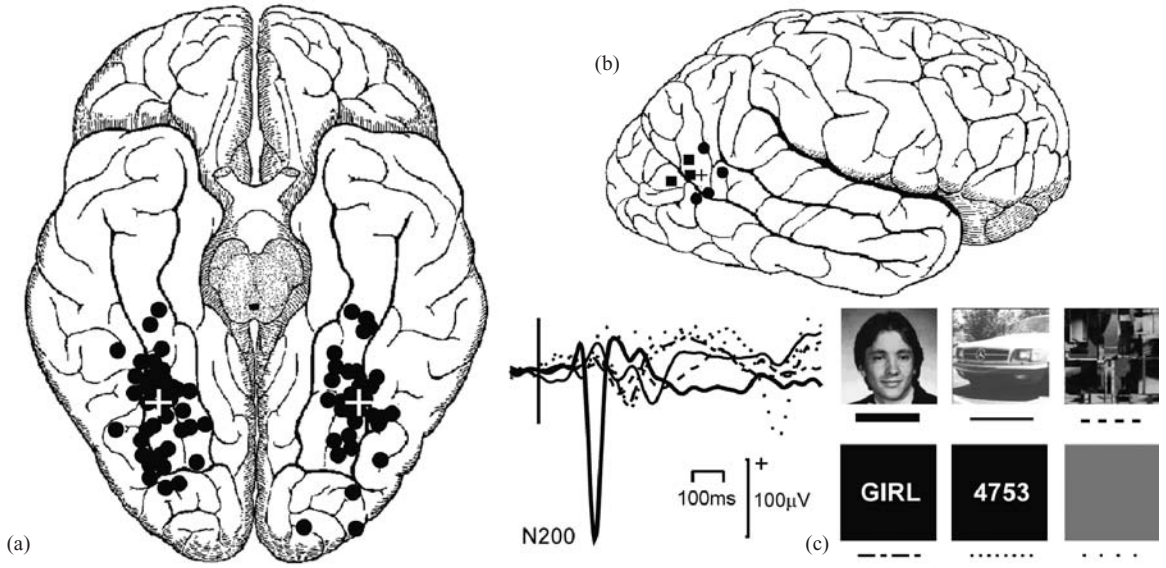

Figure 2

ffi

ff ff

ffi

(a) (b) (c) (d) (e)

Figure 3

ff

ff

ff

ff

ffi

ff

ff

Farah M J, Rabiowitz C, Quinn G E, Liu G T 2000 Early commitment of neural substrates for face recognition. *Cognitive Neuropsychology* **17**: 117–23
Gauthier I, Logothetis N K 2000 Is face recognition not so unique after all? *Cognitive Neuropsychology* **17**: 125–142
Haxby J V, Hoffman E A, Gobbini M I 2000 The distributed human neural system for face perception. *Trends in Cognitive Science* **4**: 223–33
Humphreys G W, Donnelly N, Riddoch M J 1993 Expression is computed separately from facial identity, and it is computed separately for moving and static faces: neuropsychological evidence. *Neuropsychologia* **31**: 173–181
Meadows J C 1974 The anatomical basis of prosopagnosia. *Journal of Neurology, Neurosurgery and Psychiatry* **37**: 489–501
Milders M V, Perrett D I 1993 Recent developments in the neuropsychology and physiology of face processing. *Baillieres Clinical Neurology* **2**: 361–88
Morton J, Johnson M 1991 CONSPEC and CONLEARN: A two-process theory of infant face recognition. *Psychological Reviews* **98**: 164–81
Puce A, Allison T, McCarthy G 1999 Electrophysiological studies of human face perception, III. Effects of top-down processing on face-specific potentials. *Cerebral Cortex* **9**: 445–58
Sergent J, Ohta X, McDonald X 1992 Functional neuroanatomy of face and object processing. *Brain* **115**: 15–36
Yin R K 1969 Looking at upside-down faces. *Journal of Experimental Psychology* **81**: 141–45

A. Puce

Facial Expressions

1. Introduction

This article describes one aspect of human communication and behavior—facial expressions. It describes the characteristics of a facial expression and its biological wiring, and then offers a selective history of research on facial expressions that emphasizes the nature vs. nurture debate over the origins of facial expressions of emotions. The article concludes with some current research issues and future directions of facial expression research.

2. Defining Facial Expressions

Of all the forms of human communication, which includes the written and spoken word, body language, and so forth, facial expressions are recognized as among the most salient and influential. Researchers reserve the term 'facial expression' for those recurring configurations of facial muscle movements that communicate some thought, emotion, or behavior. This is because not all recurring facial muscle configurations express specific messages. For example, some facial muscle actions that accompany spoken words—such as raising one's eyebrows when emphasizing a particular word—may modify those words, but are not messages in and of themselves (e.g., Ekman 1991).

The face can express various thoughts. For example, a person who raises the outer corner of one eyebrow may convey sophisticated skepticism. A person whose eyebrows are pulled up in the middle may convey sympathy for another. A wink can convey that one is kidding. Flashing both eyebrows upward may convey a greeting. Or, lowered eyebrows may convey uncertainty (Eibl-Eibesfeldt 1989). Researchers agree for the most part that these types of facial expressions are learned like language, displayed under conscious control, and their meanings are culturally specific that rely on context for proper interpretation (e.g., Birdwhistell 1970). Thus, the same lowered eyebrow expression that would convey 'uncertainty' in North America might convey 'no' in Borneo (Darwin 1872/1998).

The face can also express emotions. For example, humans express the emotion of happiness by raising lip corners into what is commonly called a smile. Humans can express sadness by frowning. Besides happiness and sadness, other emotions that seem to have specific facial expressions include anger, disgust, fear, and surprise, and to a lesser extent contempt, embarrassment, interest, pain, and shame (e.g., Ekman 1993, Izard 1991). What makes the facial expression of these aforementioned emotions different from other facial expressions is that there is evidence that these emotions are expressed and interpreted the same across all cultures (e.g., Ekman 1993, Izard 1971). This 'universal' production and perception across cultures suggests that those emotions and their specific facial expressions might be determined genetically, rather than socially learned. However, this claim is not without controversy (e.g., Russell 1994).

3. Neuroanatomy of Facial Expression

The idea that facial expressions can be both determined genetically, as in the case of some of the emotions, and learned socially, as in the case of all other facial expressions, is supported by an examination the neuroanatomy of the face. There appears to be two distinct neural pathways that mediate facial expressions, each originating in a different area of the brain; one area for the voluntary, willful facial actions (the cortical motor strip), and the second area for the more involuntary, emotional facial actions (subcortical areas; reviewed by Rinn 1984). This dual origin hypothesis is supported by clinical observations of patients who are paralyzed on one side of their face. When these patients were asked to pose a smile, they could only smile on half their face. Yet when these same patients felt the spontaneous emotion of enjoyment after being told a funny joke, they were able to smile on both sides of their face. Likewise, patients with lesions of the subcortical areas of the brain such

as the basal ganglia have difficulty showing spontaneous, emotional facial expressions; however, these patients are able to move their facial muscles on command. These facial action observations are so reliable that they serve as diagnostic criteria for brain lesions.

4. *History of Facial Expression Research*

The turbulent history of the systematic study of facial expressions began with the publication of Darwin's book *The Expression of the Emotions in Man and Animals* (1872/1998). In this book, Darwin proposed that humans across all cultures have particular and distinct facial expressions for particular emotions, and that these expressions are produced involuntarily as a result of that emotion. Darwin defined emotions as behavioral and physiological reactions that have helped humans and animals to survive the various life challenges they faced throughout their evolutionary history. For example, the fear reaction assisted humans and animals escape danger, the anger reaction assisted humans and animals to fight rivals, and so forth. Those who possessed these emotional reactions were more likely to live to reproductive age and therefore pass their genes to the next generation (see *Emotions, Evolution of*).

What Darwin argued (and elaborated by others, e.g., Ekman 1991, Izard 1991, Plutchik 1991) is that social animals, such as humans, must communicate these emotions to others in the group because emotions express imminent behavior, such as striking out in anger, fleeing in fear, and so on (see *Emotions, Psychological Structure of*). The facial expression of anger thus becomes the visual signal of this intention to strike. This signal allows others in the group to avoid this person, and thus avoid a potential fight (although others argue that these expressions would deprive an individual of a competitive advantage like the element of surprise prior to an attack, e.g., Fridlund 1994). These facial expressions of emotion were seen as vestiges of an entirely nonverbal human communication system that must have existed in extinct human forms such as Neanderthal, because only modern humans have the throat structures necessary to produce articulate speech. A further clue to this prehistoric human communication comes from the genetically closest living relatives of humans, the chimpanzees, who have a repertoire of facial expressions of emotion that parallel, but are not identical to, human facial expressions of emotion. Thus, current human facial expression of emotion represents more the communication methods of the past genetic history of the species, rather than its present conditions (Brown 1991).

However, much of the empirical work that followed Darwin's book failed to support his notion that there were particular facial expressions for particular emotions. For example, when subjects were startled by firecrackers, embarrassed, or disgusted by having to decapitate a live rat, their most common facial expression shown across all these situations was a smile, even though these subjects were not experiencing positive emotion. Likewise, observations of people outside North America by social scientists cast further doubt; for example, the smile was observed as an expression of uncertainty in Africans but as an expression of sadness in Japanese women (see *Culture and Emotion*). Findings such as these—although there were scattered findings to the contrary—caused scholars to conclude that facial expressions did not provide accurate information as to emotional state (Birdwhistell 1970). Thus, by the early 1960s, social science seemed to conclude that all facial expressions—including facial expressions of emotion—were culturally relative, socially learned, and that there were no universals.

Despite this conclusion, two theorists revived Darwin's ideas about the evolutionary origins of facial expressions of emotion (Plutchik 1991, Tomkins 1962/1963). These researchers took photographs of people posing prototypical emotions such as anger, disgust, fear, happiness, and so on, and found that observers would agree as to which expression represented which emotion. Other researchers found similar results with various European, South American, African, and Asian cultures (e.g., Izard 1971). Proponents of the social learning/cultural relativism perspective counter-argued that the populations upon which this evidence for universality was based were mostly educated, and thus could have learned from various forms of media which expressions represented which emotions (e.g., Birdwhistell 1970). To parry this argument, researchers conducted similar studies with visually isolated peoples whom had limited contact with Westerners, and thus could not have learned these expressions from the media (e.g., the Sadong of Borneo, and Fore in New Guinea). These researchers found for the most part the same pattern of universal expression and recognition of facial expressions of emotion as in the Westernized peoples (e.g., Ekman 1993). Follow up research using a variety of methodological alterations to this basic paradigm found patterns consistent universality throughout the 1970s and 1980s (e.g., Izard 1991). Finally, parallel evidence in favor of universality came from observations of children who were born blind and deaf, and who could not have seen these facial expressions to learn how to express them. These children showed similar expressions of emotion as their sighted counterparts (Eibl-Eibesfeldt 1989).

However, proponents of Darwin's idea were still stuck with the findings that peoples of different cultures sometimes showed different expressions for a given emotion than North Americans. Ekman proposed that the reason this happened was that different cultures learned different rules to regulate their facial

expression of emotion—what he called 'display rules' (Ekman 1993). For example, Japanese culture has a display rule that prohibits expression of anger or disgust to higher status people, unlike North American culture. Researchers found that both groups showed facial expressions of disgust when viewing a gruesome film alone. But when in the presence of a high status person, the Japanese group hid their disgust feelings with a smile, whereas the Americans still showed disgust expressions. This concept of display rules seemed to account for why people smiled to such seemingly different events as the death of a loved one, confusion, uncertainty, startle, sexual excitement, disgust, and so on (Ekman 1991). Based on these findings, Ekman (1993) proposed his neurocultural theory of emotions. This theory argued that certain basic human emotions generated particular patterns of physiology and facial expressions, that these facial expressions were universal across all cultures, but that their ultimate expression was modified, exacerbated, suppressed, or masked by social learning processes dependent upon personal, family, or cultural display rules (see *Adulthood: Emotional Development*).

5. *Current Facial Expression Research*

By the early 1990s, a consensus seemingly emerged in the field of psychology that Darwin was correct after all—that some facial expressions of emotion were universal. This was not a peaceful consensus; social scientists who placed the uniqueness of culture at the forefront of any understanding of emotion were not convinced of universality, based on the observations described earlier (e.g., Russell 1994). Experimental psychology itself issued two challenges to universality in the early 1990s. One challenge suggested that all facial expressions were simply communicative gestures, that is, they are not the result of internal emotional states, but only the result of the social motives of the person within a particular context (the 'behavioral ecology' view; Fridlund 1994). The behavioral ecology view found that facial expression, particularly smiling, was related not to felt emotion, but to the presence of others. Proponents of universality counter-argued that not all smiles are the same. They demonstrated that only one type of smile, called the enjoyment smile, is related to the positive emotional experience of enjoyment, as measured through self-report or pattern of brain activity. This enjoyment smile looks different from other smiles in that only enjoyment smiles feature *orbicularis oculi* action (the muscles surrounding the eye that give a 'crow's feet' appearance) along with zygomatic major action (the typical lip corner raising). Failing to note the distinction between enjoyment and other smiles may have been why other researchers found no relationship between smiling and positive emotion (Ekman and Rosenberg 1998).

A second challenge to universality attacked the concepts behind what was meant by universality, as well as the methods used to document universality. These methodological problems—such as biased response forms and preselected facial expressions—when added together may have conspired to bias observers' judgments, causing them to artificially agree on which facial expression represented which emotion (Russell 1994). Prompted by these criticisms, ensuing experimental research corrected many of the proposed methodological shortcomings, and has so far reconfirmed support for the universality of facial expressions of emotion (Ekman 1994).

Although the issue of the biological vs. social origins of facial expressions of emotion is not fully resolved, what has been impressive is the amount of current research generated by the findings on universality of facial expressions of emotion (Ekman 1993). First, researchers have shown that people who pose and hold these universal facial expressions of emotion begin to experience the particular emotion they are posing—although researchers have debated the exact role of a facial expression in reflecting vs. causing an emotion (e.g., Buck 1988). Regardless, this phenomenon has enabled researchers to document physiologically specific patterns of arousal for specific emotions, and in more than one culture. Second, these universal facial expressions of emotion have been employed in studies of brain activity, leading researchers to discover that there are centers in the human brain that respond specifically to these expressions (e.g., the amygdala responds to fear expressions; Whalen 1998). Third, this work on the universal facial expressions has prompted researchers to examine their origins in children (e.g., Izard and Malatesta 1987). This work has shown that children as young as 12 months of age react differently to their mothers' expressions of fear versus to happiness; a mother's fear expression will stop a child's risky behavior, whereas a mother's happy expression will not (see *Emotions, Children's Understanding of*).

These universal expressions of emotion have also shown utility as markers of social and psychological functioning. For example, the presence of enjoyment smiles on the part of a person who has survived the death of their romantic partner predicts successful coping with that traumatic loss. Schizophrenic patients tend to show different, and sometimes fewer or more disorganized facial expressions than normal patients (reviewed by Ekman and Rosenberg 1998). Mothers show different sorts of smiles to their difficult compared to their nondifficult children. The facial expression of disgust or contempt, but not anger, predicts marital divorce (Gottman 1994). Researchers found that these unbidden facial expressions of emotion can occur for very brief flashes, called 'micro-expressions,' that under certain circumstances can betray deception (Ekman 1991). Thus, current research on facial expressions has moved away from

documenting the existence of these emotional facial expressions and has moved toward examining the implications of the presence or absence of these facial expressions and their corresponding emotions on human social development, interaction, relationships, and psychopathology.

6. *Future Directions of Facial Expression Research*

Advances in technology will aid facial expression research by allowing researchers to quickly, validly, and reliably observe facial expressions. This will be helpful to the field because current work on facial expression is extremely time and labor intensive, or suffers from other experimental concerns. For example, visible scoring systems, that require close examination of videotape, can take 60 minutes to analyze one minute of behavior (e.g., Ekman and Rosenberg 1998, Izard and Malatesta 1987). Electromyographic techniques, that use electrodes on the face to measure the faint electrical impulses produced by muscle contractions, suffer from concerns about the salience of electrodes on a person's face affecting the behavior of that person (e.g., Fridlund 1994). In the future, computer based analysis programs will be developed to assess the specific muscle movements associated with facial expressions at a much faster, and more reliable way, than these older methods, without causing awareness on the part of the person being analyzed. This will have the effect of making research on facial expression more accessible to more researchers, which can only help the field progress more quickly than in the past.

There are many questions 'facing' future facial expression research, and space limitations permit only a description of a few. First, researchers will try to clarify the stimuli and processes by which social information elicits an emotion and its expression (the 'appraisal' process; Scherer et al. 2000), as well as the process by which people learn to control their emotional facial expressions. This inquiry might provide researchers with a gateway to understanding better the role of expression in the experience and management of emotions. It will also lead to understanding how perceptions of facial expressions may account for differences in social competence and functioning, or 'emotional intelligence' of adults and children. An offshoot of this work would explore whether professionals and lay people can be trained to improve their accuracy at interpreting emotional expressions, and the implications this has for their relationships. Second, researchers will move toward investigating more interactional research designs that place the facial expression of emotion back into the social context which it is typically embedded, to measure the consequences of such expression in the real world. Third, given that much of the previous work has been with posed expressions, future work would employ more spontaneous facial expressions. Fourth, with the assistance of technology that allows noninvasive observation of the working brain (e.g., Positron Emission Tomography or functional Magnetic Resonance Imaging), researchers will continue to use facial expressions of emotion to map where in the brain the expressions are perceived as well as where they are generated (e.g., Whalen 1998). Fifth, an examination of these first four future directions will inevitably lead to a better understanding of individual differences in production, control, and recognition of facial expressions, of which there is little work at present. Finally, this work would need to be expanded to cultures other than Europe or North America to assess the relative universality of the process of emotion, its antecedents, attempts to control, and its effect on facial expression (e.g., Ekman 1993).

Research on facial expressions has both paralleled and driven changes in the general assumptions in the field of psychology. The finding that people of all cultures seemed to agree on which facial expressions represented which emotions pushed psychology toward re-examining the biological bases for behavior. But research on facial expressions will continue to be controversial because it exposes the strong feelings of those who believe in the power of social situations to mold all human behavior, expressive or not, and those who believe in the biological origins of some of those behaviors. Thus, debates over facial expressions are really debates about human nature—a debate that has tormented social science from time immemorial. What research on facial expression has done is to help move this debate away from an argument over political beliefs about human nature, and toward an argument over observable data.

See also: Adulthood: Emotional Development; Culture and Emotion; Emotion and Expression; Emotion: History of the Concept; Emotion in Cognition; Emotion, Neural Basis of; Emotional Inhibition and Health; Emotions and Health; Emotions, Children's Understanding of; Emotions, Evolution of; Emotions, History of; Emotions, Psychological Structure of; Emotions, Sociology of; Face Recognition Models; Face Recognition: Psychological and Neural Aspects; Infancy and Childhood: Emotional Development; Psychological Therapies: Emotional Processing

Bibliography

Birdwhistell R L 1970 *Kinesics and Context*. University of Pennsylvania Press, Philadelphia, PA
Brown D E 1991 *Human Universals*. McGraw-Hill, New York
Buck R 1988 *Human Motivation and Emotion*. Wiley, New York
Darwin C 1872/1998 *The Expression of the Emotions in Man and Animals*. Oxford University Press, New York
Eibl-Eibesfeldt I 1989 *Human Ethology*. de Gruyter, New York

Ekman P 1991 *Telling Lies*. Norton, New York
Ekman P 1993 Facial expression and emotion. *American Psychologist* **48**: 384–92
Ekman P 1994 Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin* **115**: 268–87
Ekman P, Rosenberg E L (eds.) 1998 *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, New York
Fridlund A J 1994 *Human Facial Expression: An Evolutionary View*. Academic Press, San Diego
Gottman J 1994 *Why Marriages Succeed or Fail*. Simon & Schuster, New York
Izard C E 1971 *The Face of Emotion*. Appleton-Century Crofts, New York
Izard C E 1991 *Human Emotions*. Plenum Press, New York
Izard C E, Malatesta C Z 1987 Perspectives on emotional development. I. Differential emotions theory of early emotional development. In: Osofsky J D (ed.) *Handbook of Infant Development*, 2nd edn. Wiley, New York, pp. 494–554
Plutchik R 1991 *The Emotions: Facts, Theories, and a New Model*. University Press, New York
Rinn W E 1984 The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin* **95**: 52–77
Russell J A 1994 Is there universal recognition of emotion from facial expression? A review of cross-cultural studies. *Psychological Bulletin* **115**: 102–41
Scherer K R, Schorr A, Johnstone T 2000 (eds.) *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, New York
Tomkins S S 1962/1963 *Affect, Imagery, Consciousness (Vol. 1, The Positive Affects, Vol. 2, The Negative Affects)*. Springer, New York
Whalen P J 1998 Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science* **7**: 177–88

M. G. Frank

Faction: Political

In its broadest construction, a political faction is any part of a political whole. The term has been applied to phenomena ranging from sets of people whose policy preferences tend to align to membership groups that undertake collective action.

In a most famous formulation Madison (Hamilton et al. 1961) defined a faction as a partisan political division of any size, although he qualified the definition to refer to such groups whose aims were 'adverse to the rights of other citizens, or to the permanent and aggregate interests of the community.' Madison viewed factions as the inevitable consequence of political liberty and social diversity, and urged the 'extended sphere' of a federal republic to contain 'the violence of faction.' Madison's treatment of the 'mischief of faction' is often given as evidence for the American founders' distaste for such commonplace features of representative democracy as political parties and lobbying groups.

For current analytical purposes, Madison's definition of faction is overly broad, in encompassing political parties and interest groups, and overly restrictive in parsing motives. A good, if still broad, definition is that a political faction is a subsidiary part of a political institution, of a political party, a lobbying organization, or a legislature. Unlike political parties within a polity or within a legislature, factions commonly have no legal standing in the institution of which they are part, although they might play an important and enduring role in organizing it. Subgroups that have legal status within institutions are usually designated in other ways, as are the caucuses in the United States Congress and the international political parties in the International Typographical Union. Sometimes analysts identify as factions groupings that are evanescent and happenstantial, for instance a voting coalition on a particular legislative bill or party nomination. Most, however, limit the term to groupings that are enduring and widely recognized. The factional affiliations of Liberal Democrats standing for election to the Japanese Diet, for instance, were routinely reported in the newspapers. The factions within the Democratic Party in West Virginia, while not reported in the press, nevertheless persisted through changes in political administration and party leadership.

In any complex society, a tendency toward faction within political parties and interest groups would seem to be inevitable. Members of parties and groups might be united on the central purposes of the organization but still divided on secondary questions. For decades, the Democratic Party in the United States was loosely united on issues of economic policy but deeply divided into northern and southern factions on issues of race and federal powers, to the point where party allegiances were often superseded by a conservative coalition, an informal alliance of Republicans and southern Democrats, in the US Congress. A common purpose in clericalism still left the Christian Democrats, Italy's confessional party, vulnerable to factions arising from region and class.

Although they are more difficult to observe, similar factional divisions occur in many lobbying groups. At several points in its history, the American Farm Bureau Federation harbored two or three identifiable factions that reflected regional commodity interests. In the 1920s, the Anti-saloon League foundered on a factional division over the question of whether the League ought to emphasize the enforcement of Prohibition in the United States or the promotion of the cause of prohibition abroad.

Considerable evidence exists, however, that some political parties and some interest groups are more prone to faction than others. Political parties in Japan, Italy, and France in the Fourth Republic were

famously factionalized, while parties in Norway, Sweden, and the German Federal Republic were not.

Several conditions appear to promote a tendency toward factional strife in political parties. One factor is electoral domination: factionalization appears more likely in parties that face no real electoral challenge. The Italian Christian Democrats (DC) and the Japanese Liberal Democrats (LDP) are striking both for their enduring factional divisions and their long runs of electoral success—the DC participated in every Italian government from World War II into the 1990s, and the LDP controlled the Japanese Diet continuously from 1955 to 1993. In the American southern states, the electoral threat posed by the mountain Republicans produced a Democratic Party in North Carolina and Tennessee that was appreciably more cohesive than the Democratic Party in states like South Carolina, Florida, Mississippi, and Arkansas, where the Republican Party constituted barely even a nuisance. While counterexamples of dominant parties not rent by faction abound—the Social Democrats in Sweden, the Republicans in Vermont—a considerable body of theory, beginning with the work of Riker (1962) in fact predicts that parties that enjoy more than minimal majorities will be prone to factional disintegration (see *Minimum Winning Coalition, in Politics*).

A second influence on the propensity toward factionalization is the structure of political cleavage within the society (Lipset and Rokkan 1967). In Belgium, class divisions cut across parties that are organized primarily on the basis of language and religion. In Italy, economic and regional differences fueled factions within a confessional party whose raison d'etre is the establishment of the Roman Catholic Church. In Sweden, by contrast, religious homogeneity and a dominant secularism caused religious disputes to recede well behind the class divisions that are the basis for the political parties (see *Cleavages: Political*).

A third influence on faction is the electoral system. The electoral laws place an upper bound on the number of political parties that can be sustained, a limit equal to one more than the number of candidates to be elected from the constituency (Cox 1997). Accordingly, the electoral laws govern the ease with which factional divisions can progress to outright schism and the formation of new political parties. In the Low Countries of Belgium and the Netherlands, highly proportional representation rules coupled with low election thresholds have produced party systems in which parties and factions morph each into the other with remarkable fluidity. Japan's system of multi-member districts with a single nontransferable vote fostered a half-dozen factions that slated candidates and oversaw the distribution of spoils. The hard limit of a system of single-member districts and plurality voting in the United States forced northern and southern Democrats to stick together, despite their factional animosities. In the United States, the advent of nominating primaries—creating, effectively, single-member constituencies with plurality voting *within* parties—often limited the number of enduring party factions to two, as in North Dakota on the Republican side or in Louisiana on the Democratic side. In North Dakota, in fact, the more liberal wing of the Republican Party, the Nonpartisan League, split off and joined the Democrats in the 1950s, when two-party competition finally became viable.

The ease or difficulty of schism also contributes to the propensity of interest groups to be ridden with factions. As Hirschman (1970) put it, when exit is costly, members resort to voice. Interest groups that rely primarily on the expressive value of the group's purposes to motivate members seem especially prone to faction (Wilson 1974). Doctrinal purity matters when members are motivated primarily by doctrine. The National Organization for Women, for example, suffered through disputes between a faction that wished to pursue conventional lobbying for the Equal Rights Amendment and a faction that wished to promote social movement activities like direct action. Interest groups that rely upon expressive benefits, moreover, tend also to be prone to schism, because 'purposive' inducements to group involvement are so easy to provide. In the later part of the twentieth century, the 'public interest groups' in the United States multiplied like Protestant denominations, rapidly and schismatically. A dissident faction of the Sierra Club, for example, bolted the organization to form Friends of the Earth; a few years later, a dissident faction of Friends of the Earth created the Environmental Policy Center. In contrast, interest groups that attract members with relatively expensive material benefits like insurance, or interest groups that are able to secure membership through coercion, make it costly for dissidents to exit, channeling what might have been schism into faction. Factional fights within labor unions typically continue intramurally rather than extramurally; the costs a new organization would have to bear to win the right to represent workers in collective bargaining preclude exit. The American Medical Association (AMA) has endured despite tensions between specialists and general practitioners, between doctors in large practices and in small practices, between academics and clinicians, bound by the range of AMA services and by the state powers that have been delegated to the medical societies to regulate medical employment.

The potential for faction in interest groups has long been seen as an important limit on the power and influence of lobbying groups in American politics. Pluralist political scientists identified 'overlapping memberships,' that is, conflicting internal interests, as an important limit on the demands that interest groups might make. During the energy policy debates of the 1970s, for instance, the National Petroleum Refiners Association took no position on crude oil price

regulation, immobilized by conflicts between the small independent refiners and the 'majors.' Likewise, the potential for internal disagreement sidelined many of the large business associations, like the US Chamber of Commerce and the National Association of Manufacturers, in the debates over the extension of the Reciprocal Trade Agreements Act in the 1950s. 'Quasiunanimity,' Bauer et al. (1972) concluded, was a 'premise of action' in lobbying groups. A group that was too extreme in its demands risked disabling internal controversy.

In political parties, factions emerge and recombine to produce change in party systems, as has occurred recently in both Japan and Italy. But factional divisions also hobble parties in the achievement of their policy goals. The most dramatic example is surely the decades-long obstruction of civil rights legislation, a Democratic Party priority since the Truman administration, by the Party's southern Dixiecrat faction. But if factionalism is joined with limited party competition, many argue, the consequences are the more severe. When parties are isolated from the judgment of the electorate, factions dispute not policy direction but the division of the spoils. The extreme factionalization of the Liberal Democratic Party in Japan and the Christian Democratic Party in Italy contributed to the inefficiency, particularism, and corruption of those two dominant parties. But at least the Japanese and Italian factions were relatively enduring and organized. As Key (1950) argued of the Democratic Party in the American southern states, where commonly it was 'every man for himself' in elections, competition between ephemeral factions within a hegemonic party is no substitute for competition between two or more political parties. Political parties develop a stake in their policy reputations that factions commonly do not. Popular control in a democracy requires the clear electoral choices that political parties provide and party factions cannot.

See also: Cleavages: Political; Electoral Systems; Factionalism; Interest Groups; Interest Groups, History of; Interest: History of the Concept; Lobbying; Minimum Winning Coalition, in Politics; Party Systems; Political Parties

Bibliography

Bauer R A, Pool I deS, Dexter L A 1972 *American Business and Public Policy*. Aldine-Atherton, New York

Cox G W 1997 *Making Votes Count*. Cambridge University Press, New York

Hamilton A, Madison J, Jay J 1961 In: Rossiter C (ed.) *The Federalist Papers*. New American Library, New York, no. 10

Hirschman A O 1970 *Exit, Voice, and Loyalty*. Harvard University Press, Cambridge, MA

Key V O Jr 1950 *Southern Politics*. Knopf, New York

Laver M, Schofield N 1990 *Multiparty Government*. Oxford University Press, New York

Lipset S M, Rokkan S 1967 Cleavage structures, party systems, and voter alignments. In: Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments*. Free Press, New York

Mayhew D R 1986 *Placing Parties in American Politics*. Princeton University Press, Princeton, NJ

Ramseyer J M, Rosenbluth F McC 1993 *Japan's Political Marketplace*. Harvard University Press, Cambridge, MA

Riker W 1962 *The Theory of Political Coalitions*. Yale University Press, New Haven, CT

Wilson J Q 1974 *Political Organizations*. Basic Books, New York

J. M. Hansen

Factionalism

Factionalism refers to dissension between rival subgroups—factions—within a larger social unit. Factionalism can take many forms and has been observed in all parts of the world. It is a basic political process dynamically related to social change.

1. Factions and Organizations

Factionalism, regardless of where it is classified on the conflict continuum, is conflict between factions. Factions are coalitions of persons or subgroups that compete over specific issues within a larger organization or community. The central focus of a faction is the leader who coordinates its activities and recruits its members. Ties between leader and followers are usually personal, although some followers may recruit others on behalf of the leader. The issues about which factions compete are diverse. They generally concern scarce resources the control over which provide power chances, such as economic assets, office and new laws. But they may also involve honor, ideology and behavioral norms. Loosely structured, factions are non-corporate groups that generally dissolve when the particular issues that gave rise to them are resolved. But if the issues remain unresolved, factions may acquire a range of cultural trappings such as property, symbols, ideology and bureaucratic organization. They can then evolve into permanent corporate groups, such as ritual moieties, political parties or other formal associations, for which the term faction is no longer appropriate.

Factionalism takes place within the framework of an established social entity, whether village, school, political party, office, club, kin group, etc., that have clear norms of behaviour. These norms generally include notions of unity, consensus, and cooperation.

The covert maneuvering of factions bent on achieving their own, often self-serving aims, contradict these norms. Hence factionalism has a pejorative connotation. It is seen as subverting organizational rules and goals. Faction members are consequently viewed as disloyal, obstructive persons whose pursuit of narrow, short-term advantage endangers the wider, long-term goals of the organization. The divisiveness inherent in factionalism also hampers the day to day affairs of the organization or community that depend on cooperation. Furthermore, factionalism jeopardizes the good name and image of unity and harmony it seeks to project to the outside world.

2. *The Study of Factionalism*

The study of factions and factionalism developed slowly, shadowing theoretical shifts in the social sciences. Although Linton (1936) long ago suggested that factions presented an interesting but unexplored field, little was done until the 1950s. Political anthropology until then was dominated by the functionalist paradigm elaborated in *African Political Systems* (Fortes and Evans-Pritchard 1940). This viewed politics as maintaining order through consensus, harmony, and balanced opposition. The political groups on which functionalists focused were enduring units, corporate groups. Conflict, if examined, was viewed as reinforcing the social structure. Loosely structured, temporary coalitions such as factions patently did fit into this conception of politics. The theoretical hegemony of Africanist political anthropologists began to be challenged in the 1950s. Not surprisingly, Raymond Firth, given his interest in individual choice, had for some time been uncomfortable with the functionalist paradigm. He and colleagues, who had observed factionalism in Indian communities, were the first to examine factionalism theoretically (Firth 1957). They treated factions as informal counterparts of more formal political formations whose members were recruited according to structurally diverse principles. They also noted that factions tended to become activated on specific occasions and not as regularly recurring features. Other studies of factions and factionalism swiftly followed (Siegel and Beals 1960, Boissevain 1964, 1974, Nicholas 1965, Bailey 1969, Thoden van Velzen 1973, Bujra 1973, Alavi 1973). Most employed a transactionalist perspective, viewing political activity as an arena in which entrepreneurs transact personal relations for political and economic gain. The study of factionalism culminated in the late 1970s with the volume edited by Silverman and Salisbury (1977). This is still considered the definitive work on the subject. Following its publication and the demise of functionalist social science—in which transactionalists played a significant role—academic debate moved on to puzzles related to symbolic, cognitive, and discursive approaches to politics. Factions and factionalism have become accepted concepts whose characteristics are no longer debated. They are also proving useful to related disciplines (see Brumfiel and Fox 1994).

3. *Issues*

Much of the work on factionalism in the 1960s, although using a transactionalist perspective, continued to be strongly influenced by functionalism. Factionalism was seen as occurring as a result of rapid sociocultural change. Change was seen as coming from the surrounding environment, not as the result of tensions inherent in the community or society in which factionalism arose. Factionalism was generally viewed as occurring because the system's equilibrium was disturbed, and it operated to restore its dynamic equilibrium. Factions were regarded as structurally similar and in balanced opposition. These viewpoints were debated and successfully challenged during the 1970s.

Many of the societies in which factionalism was observed were in fact not subject to rapid social change. But if factionalism was not always a 'result' of change, it seemed always to be 'about' change: changes in power resources, ideology, rules and/or ways of doing things. Factions competed about who was to be boss, about which normative concepts were to be followed, about whose will was to prevail, and thus about which rules were to be followed. Moreover, a closer examination of rival factions revealed that far from being similar or evenly matched, they differed in their access to resources, strategy, tactics, internal organization, ideology, and social composition.

4. *Factions: Structure, Symmetry, and Balance*

Factions often form in opposition to or in defense of some issue or some pre-existing source of power and authority within a community or organization. The distribution of such resources is binary. Some persons have more and some have less. Those with more normally constitute the local establishment that coalesces around a leader or a dominant personage such as a headman, mayor, parish priest, or club president. They and their supporters form the 'establishment' faction. Those who are dissatisfied with the establishment's exercise of its power constitute a category from which a rival or 'opposition' faction can be recruited.

Reports also mention conservative and progressive factions. Because the local establishment normally defends the 'status quo,' from which it derives its superordinate position, it is often labeled as conservative. Since the opposition faction challenges the established defenders of the 'status quo' it is labelled progressive. These labels often do not reflect reality, as when a progressive opposition faction defeats its rival and becomes the dominant, establishment faction.

There is a further reason why opposition factions come to be regarded as progressive. Since they are weaker, they are perhaps more receptive to new resources if and when these become available in the wider society. With these they can challenge their rivals. In rapidly changing societies these may include new government and commercial offices, new laws and new ideologies. Such new resources rapidly tend to change the balance of power. Factionalism thus does not necessarily result from the availability of new resources, as some authors have suggested. Rather, new resources are used in ongoing competitions for power and prestige and tend to escalate the conflict. The use of new resources is not random. A faction will use new resources when it seems likely that they will strengthen its position. It is then labelled progressive, in the sense of favoring change of the 'status quo.'

There is evidence that conflict groups—whether faction, ritual moiety, or political party—differ organizationally. When the opposition consists merely of a category of persons disgruntled with the dominant power elite, they are obviously less organized than the establishment. The internal structure of the local establishment faction, whose members are used to networking to maintain their position, will normally have a more developed exchange circuit than the opposition faction. But if a conflict between the two persists over time, the opposition faction may well become better organized than its rival may. Good organization is a valuable resource and one of the ways a weaker faction can successfully challenge its rival. It is thus more open than the establishment faction to organizational innovation. Because of its superior resources, the dominant faction also tends to be more wasteful. It does not need to husband its resources to the extent that its weaker rival does.

Like most coalitions, factions have core and peripheral members. Core members cluster around the leader and have multiple links to each other. The peripheral members are often linked only to the leader or to a single member of the core group. Where there is a strong core, the faction often acquires some of the characteristics of corporate groups noted previously.

Factions are not necessarily ideologically neutral, as some authors have suggested. The differences between establishment and opposition, like those between conservative and progressive, are not random. They have ideological implications. There is evidence that opposition factions recruit more support than their rivals from weaker or even marginal social categories do. Since opposition leaders usually lack the network ties and material resources that establishment leaders use to recruit followers, they cannot afford to be too particular about the nature of their support. The strength of a faction is usually a function of its size. Just as they often turn to new ideologies and tactics—which because they are new are often viewed as socially unacceptable and subversive—opposition leaders also recruit supporters from among those who are less influential or are regarded as social or morally inferior. Followers are followers.

Opposition factions consequently often develop or adopt an overarching ideology or symbol to bind their heterogeneous members into a unity. They also often align themselves with political parties that defend the interests of their socially weaker supporters. Most often, opposition factions seek links to parties that embody an emancipatory ideology. Their establishment rivals develop relations with political parties representing vested interests in the wider society. Political parties, on the other hand, also consciously make use of local factions to recruit support at the grass roots level.

5. *Conclusion*

If factionalism is not necessarily a product of social change, it appears to be always about change. Factions are coalitions that compete for power to determine, and thus to change, what is to be accepted as normal. Rival factions, because they have different access to power chances, are not evenly matched, structurally similar groups. Their asymmetry is fundamental to understanding the nature of factionalism and its dynamic for long term change.

Structural asymmetry and competition for power are also characteristic of class-based conflict groups. This suggests that the line of cleavage between faction, class or party cuts across moral categories and socio-economic classes, not at right angles, as most functionalist and class-oriented analysts postulate, but diagonally. Where the line approaches the vertical, forming conflicting coalitions with a generally even spread across socio-economic classes, it is reasonable to speak of factionalism, in the case of face-to-face groups, and party conflict in the case of conflict on a broader scale. Where the line of cleavage approaches the horizontal, forming conflict groups that are more clearly differentiated according to socio-economic criteria, the term class conflict seems appropriate. But in every case the axis of cleavage must be determined by empirical investigation. It should not be taken for granted.

See also: Charisma: Social Aspects of; Conflict and Conflict Resolution, Social Psychology of; Conflict: Anthropological Aspects; Conflict/Consensus; Conflict: Organizational; Conflict Sociology; Faction: Political; Groups, Sociology of; Issue Networks: Iron Triangles, Subgovernments, and Policy Communities; Leadership, Psychology of; Solidarity, Sociology of

Bibliography

Alavi H 1973 Peasant classes and primordial loyalties. *Journal of Peasant Studies*. **1**: 232–62

Bailey F G 1957 *Caste and the Economic Frontier*. Manchester University Press, Manchester, UK
Bailey F G 1969 *Stratagems and Spoils*. Basil Blackwell, Oxford, UK
Boissevain J 1964 Factions, parties and politics in a Maltese village. *American Anthropologist*. **66**: 1275–87
Boissevain J 1974 *Friends of Friends: Networks, Manipulators and Coalitions*. Basil Blackwell, Oxford, UK
Brumfiel E M, Fox J W (eds.) 1994 *Factional Competition and Political Development in the New World*. Cambridge University Press, Cambridge, UK
Bujra J 1973 The dynamics of political action: A new look at factionalism. *American Anthropologist*. **75**: 132–52
Firth R 1957 Introduction to factions in Indian and overseas Indian societies. *British Journal of Sociology*. **8**: 291–5
Fortes M, Evans-Pritchard E E (eds.) 1940 *African Political Systems*. Oxford University Press, London
Linton R 1936 *The Study of Man*. D. Appleton-Century, New York
Nicholas R W 1965 Factions: A comparative analysis. In: Banton M (ed.) *Political Systems and the Distribution of Power*. Tavistock, London
Siegel B, Beals A R 1960 Pervasive factionalism. *American Anthropologist*. **62**: 395–417
Silverman M, Salisbury R F (eds.) 1977 *A House Divided? Anthropological Studies of Factionalism*. Memorial University, Newfoundland, Canada
Thoden van Velzen H U E 1973 Coalitions and network analysis. In: Boissevain J, Mitchell J C (eds.) *Network Analysis: Studies in Human Interaction*. Mouton, The Hague, The Netherlands

J. Boissevain

Factor Analysis and Latent Structure, Confirmatory

Confirmatory factor analysis (CFA) is a quantitative data analysis method that belongs to the family of structural equation modeling (SEM) techniques. CFA allows for the assessment of fit between observed data and an *a priori* conceptualized, theoretically grounded model that specifies the hypothesized causal relations between latent factors and their observed indicator variables. Because population-level equivalence between data and model cannot be proven with sample data, CFA should be viewed as a mainly disconfirmatory technique. That is, CFA facilitates the statistical rejection—or, at best, a very tentative retention—of a specific theory regarding the factor(s) responsible for the observed relations in the data. If, on the other hand, the investigator's intentions are a mostly ungrounded exploration of relations suggested by the data, classical exploratory factor analysis is the more appropriate approach. In this entry, typical steps in a CFA are introduced theoretically and via example: from model specification and identification, to parameter estimation, data-model fit assessment, and potential model modification. Applied and methodological references are provided for a more in-depth study of CFA and SEM techniques in the social and behavioral sciences.

1. Overview

The term 'factor analysis' describes a host of methods, all of which have the purpose of facilitating a better understanding of the latent, unobserved variables (factors) that underlie a set of directly measurable and observed variables. These factors are often believed to represent constructs, psychological or otherwise, that have a direct bearing on the measured variables; as such they are assumed to motivate (and in turn be inferable from) the pattern of correlations or covariances among those observed variables. In the late 1960s, works by Karl Jöreskog (e.g., 1966, 1967) articulated a method for confirmatory factor analysis (CFA), an application of normal theory maximum likelihood estimation to factor models with specific theoretical latent structures. Such structures could include the *a priori* specification of the number of factors, their orthogonality or obliquity, and which variables had zero and nonzero relations with those factors. This distinguishes CFA from well-known exploratory factor analysis (e.g., Gorsuch 1983, Mulaik 1972) wherein the number and nature of the factors emerge from the observed variables' data through a mathematical algorithm, largely blind to any substantive theory. Most crucial in Jöreskog's CFA work was the provision for a formal statistical χ^2-test of the fit between the pattern of relations among the measured variables and the theorized factor model, thereby facilitating the disconfirmation or tentative confirmation of an hypothesized factor model. Soon after, Jöreskog and others put forth a more general framework for the integration of measured and latent variables into complex causal networks, serving as the foundation for what is often known as structural equation modeling (SEM). CFA, which may be considered a special case within the more general SEM framework, is the focus of the current entry.

2. CFA Model Specification, Identification, and Parameter Estimation

Suppose an educational researcher wishes to investigate the possibility of a low positive relation between reading and mathematics ability for fifth grade students, which is measured by standardized tests such as the Stanford Achievement Test or the Iowa Test of

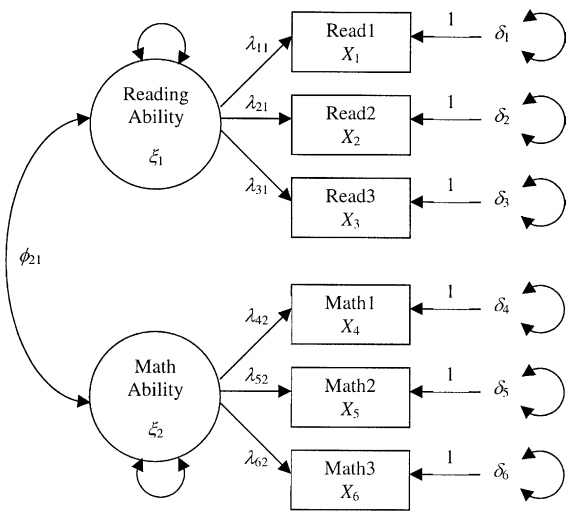

Figure 1
Hypothetical CFA model of reading and mathematics ability

Basic Skills. The model shown in Fig. 1 might be hypothesized. Measured variables X_1 through X_6, shown in rectangles, are believed to be caused by the latent factors ξ_1 and ξ_2, shown in circles. Here, ξ_1 and ξ_2 represent true latent (unobserved) reading and mathematics ability, respectively, with X_1 through X_3 being standardized reading test measures (Read1 through Read3) and X_4 through X_6 being standardized mathematics test measures (Math1 through Math3). Table 1 includes a simulated variance–covariance matrix for the observed variables X_1 through X_6 based on test data from $n = 1{,}200$ fifth graders. Our example's focus is the noncausal covariance between reading and mathematics ability, ϕ_{21}. In general, a covariance is indicated by a two-headed arrow connecting the two constructs, and—because a variance is a covariance of a variable (observed or latent) with itself—it, too, is depicted by a two-headed arrow from the variable to itself.

A factor's hypothesized causal impact on its measured indicator variables is symbolized by an arrow from the factor to the variable with magnitude λ_{ij}, where i denotes the observed variable and j denotes the latent factor. Note that such a model explicitly posits the factors as causing the variables, rather than the variables causing the factors; the latter type of model, in which the factor is characterized as emergent rather than latent, is much less common and beyond the scope of this entry. In many cases, there is no arrow from a factor to a variable, such as from ξ_1 to X_4; this implies that Reading Ability has no theoretical causal bearing on the Math1 variable. Finally, to the extent that the factors do not perfectly explain each variable, a residual term, δ, is included as an influential contributor (with its variance shown by a two-headed arrow from δ to itself). This residual might consist of variable-specific measurement error as well as other influences. Thus, each observed variable is the sum of two parts, that attributable to the common factor(s) and that residual part specific to the variable.

The causal relations of the hypothesized model shown in Fig. 1 may be expressed as a system of six regression-like structural equations:

$$X_1 = \lambda_{11}\xi_1 + 0\xi_2 + \delta_1 \quad (1)$$
$$X_2 = \lambda_{21}\xi_1 + 0\xi_2 + \delta_2 \quad (2)$$
$$X_3 = \lambda_{31}\xi_1 + 0\xi_2 + \delta_3 \quad (3)$$
$$X_4 = 0\xi_1 + \lambda_{42}\xi_2 + \delta_4 \quad (4)$$
$$X_5 = 0\xi_1 + \lambda_{52}\xi_2 + \delta_5 \quad (5)$$
$$X_6 = 0\xi_1 + \lambda_{62}\xi_2 + \delta_6 \quad (6)$$

Table 1
Simulated data and selected parameter estimates for the reading and mathematics ability model in Fig. 1

Variance/covariance matrix ($n = 1200$)	X_1	X_2	X_3	X_4	X_5	X_6
	129.96					
	79.75	192.65				
	694.20	871.11	12038.48			
	307.03	391.75	3402.04	9876.38		
	230.99	415.74	2476.89	4815.24	12126.41	
	37.85	53.36	416.71	740.98	656.84	135.722
Standardized factor loadings and indicator reliability estimates						
ξ_1	0.70*	0.73*	0.79*	0	0	0
ξ_2	0	0	0	0.75*	0.60*	0.85*
R^2	0.49	0.53	0.62	0.56	0.35	0.73
Data-model fit indices						
$\chi^2 = 16.98$, $df = 8$, $p = 0.030$				CFI = 0.996	SRMR = 0.016	RMSEA = 0.031

Note. $^*p < 0.05$

Equivalently, these equations can be represented in matrix form, as in

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix} \quad (7)$$

That is,

$$\mathbf{X} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \quad (8)$$

where $\mathbf{X}$ is a column vector of observed variables, $\mathbf{\Lambda}$ is a matrix of factor loadings, $\boldsymbol{\xi}$ is a vector of latent constructs, and $\boldsymbol{\delta}$ is a column vector of residuals.

The implication of Fig. 1 and the accompanying structural equations is that the population variance–covariance matrix for the X variables, $\mathbf{\Sigma}$, is a function of (1) the λ_{ij} loadings in matrix $\mathbf{\Lambda}$, of (2) the variances and covariance between the latent factors in a matrix $\mathbf{\Phi}$, and of (3) the variances and covariances among the residuals in a matrix $\mathbf{\Theta}_\delta$ (note that in Fig. 1 all residual covariances are zero as implied by the absence of two-headed arrows between the δ terms). More specifically, if all model parameters (loadings, variances, and covariances) are contained in a single column vector $\boldsymbol{\theta}$, the population variance-covariance matrix of the observed variables that is implied by the model and its parameters, $\mathbf{\Sigma}(\boldsymbol{\theta})$, is given by

$$\mathbf{\Sigma}(\boldsymbol{\theta}) = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Theta}_\delta \quad (9)$$

A vector of parameter estimates $\hat{\boldsymbol{\theta}}$ can be derived so that the model-implied variance–covariance matrix $\mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$ is as similar as possible to the observed variance–covariance matrix $\mathbf{S}$ provided that model identification has first been ensured. To this end, each parameter in a model must be expressible as a function of the variances and covariances of the observed variables. When a system of such relations can be uniquely solved for the unknown parameters, the model is just-identified. When multiple such expressions exist for one or more parameters, the model is over-identified; in this case, a best-fit (although not unique) estimate for each parameter is derived. If, however, at least one parameter cannot be expressed as a function of the observed variables' variances and covariances, the model is under-identified and some or all parameters cannot be estimated on the basis of the data alone. This under-identification might be the result of the researcher attempting to impose a model that is too complex relative to the number of variances and covariances of the observed variables. Additionally, empirical under-identification might arise when unfortunate estimates for select parameters (e.g., values of zero for factor covariances) render subsets of model parameters inestimable. Fortunately, in most CFA applications, it suffices to ensure that (1) the number of parameters to be estimated, p, does not exceed the number of variances and covariances of the observed variables, c, and (2) each latent factor has an assigned unit of measurement. To accomplish the latter condition for the model in Fig. 1, we set the factor variances to unity (alternatively, for each of the two factors we could have specified one of the factor loadings to equal unity, thereby setting each factor's units equal to that observed variable). The model in Fig. 1 is over-identified with $c = 6(7)/2 = 21$ non-redundant observed (co)variances and $p = 13$ parameters to be estimated: one covariance between the two latent factors, six variances of the error terms associated with the observed variables, and six factor loadings.

Given that a model is just- or (preferably) over-identified, sample estimates can be obtained through a variety of estimation methods. These include maximum likelihood and generalized least squares, both of which assume multivariate normality and are asymptotically equivalent, as well as asymptotically distribution-free estimation methods that generally require a substantially larger sample size. These methods iteratively minimize a function of the discrepancy between $\mathbf{S}$ and $\mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$, where $\mathbf{S}$ is the unrestricted variance-covariance matrix of the observed X variables and $\mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$ is the model-implied variance-covariance matrix reproduced from the iteratively changing parameter estimates. The standardized maximum likelihood estimates of key parameters in the reading and mathematics ability model are presented in Table 1. Before focusing on our example's main parameter estimate (ϕ_{21}), however, we should consider whether or not there is any evidence suggesting data-model misfit, any statistical—and theoretically justifiable—rationale for modifying the hypothesized model, or any indication of factor unreliability.

3. *Data–Model Fit Assessment and Model Modification*

One of the advantages of CFA is the ability it offers to assess the quality of the fit of the data to the model. A multitude of measures exists that assist the researcher in deciding whether to reject or tentatively retain an *a priori* specified over-identified model (see Marsh et al. 1988, Tanaka 1993). In general, measures to assess the fit between the variances and covariances observed in the data and those implied by the model can be classified into three categories: absolute, parsimonious, and incremental. Absolute fit indices are those that improve as the discrepancy between $\mathbf{S}$ and $\mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$ decreases, that is, as the number of parameters, p, approaches the number of nonredundant observed variances and covariances, c. Examples of such measures include the model χ^2 statistic that tests the

stringent null hypothesis H_0: $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$, the standardized root mean-square residual (SRMR) that roughly assesses the average standardized discrepancy between observed and model-implied variances and covariances, and the goodness-of-fit index (Jöreskog and Sörbom 1996) designed to evaluate the amount of observed variance–covariance information that can be accounted for by the model.

Parsimonious fit indices take into account not just the overall absolute fit, but also the degree of model complexity required to achieve that fit. Indices such as the adjusted goodness-of-fit index (Jöreskog and Sörbom 1996), the parsimonious goodness-of-fit index (Mulaik et al. 1989), and the root mean-square error of approximation (RMSEA, Steiger and Lind 1980) indicate greatest data–model fit when data have reasonable absolute fit and models are relatively simple. Finally, incremental fit indices such as the normed fit index (Bentler and Bonett 1980) and the comparative fit index (CFI, Bentler 1990) gauge the data–model fit of an hypothesized model relative to that of a more restrictive baseline model with fewer parameters.

The three types of fit indices together help the researcher to converge upon a decision regarding the CFA model's acceptability. In our example, the significant $\chi^2 = 16.98$ ($df = c - p = 8$, $p < 0.05$ indicates that the observed variance–covariance matrix would occur rarely if our model correctly depicted the true population relations. This absolute fit statistic, however, is notoriously sensitive to very small and theoretically trivial model mis-specifications (e.g., slight amounts of error covariance) under large sample conditions. As such, other fit indices are generally preferred for model evaluation. According to Hu and Bentler (1999), CFI values of 0.96 or greater together with SRMR values less than 0.09 (or with RMSEA values less than 0.06) point to acceptable data-model fit. The indices in Table 1 suggest no appreciable data–model inconsistency given the Hu and Bentler cut-off criteria.

After the data–model fit has been assessed, a decision about that model's worth must be reached. Acceptable fit indices usually lead to the conclusion that no present evidence exists warranting a rejection of the model or the theory underlying it. This is not to say that the model and theory have been confirmed, much less proven as correct; rather, the current factor model remains as one of possibly many that satisfactorily explain the relations among the observed variables. On the other hand, when fit indices indicate a potential data–model misfit, one might be reluctant to dismiss the model entirely. Instead, attempts are often made to modify the model *post hoc* so that acceptable fit indices can be obtained. Such modifications could include the addition of cross-loadings, allowing a given variable to load on multiple factors, or of error covariances in which variables' residuals are allowed to covary in order to reflect some potential variable relation above and beyond that motivated by the factor structure itself.

Most CFA software packages will facilitate such model 'improvement' by providing modification indices (Lagrange multiplier tests) indicating what changes in the model could reap the greatest increase in absolute fit, that is, decrease in the model χ^2 statistic. While such indices constitute a potentially useful tool for remedying incorrectly specified models, it seems imperative to warn against an atheoretical hunt for the model with the best fit. Many alternative models exist that can explain the observed data equally well; hence, attempted modifications must be based on a sound understanding of the specific theory underlying the model. Furthermore, when modifications and reanalyses of the data are based solely on data–model misfit information, subsequent fit results might be due largely to chance rather than true improvements to the model. Modified structures therefore should be cross-validated with an independent sample whenever possible. If a new sample is not available but the initial sample is large enough, one can randomly split the sample into calibration and validation subsamples and compute Cudeck and Browne's (1983) cross-validation index. When the initial sample is too small, Browne and Cudeck (1989) also offered a single sample alternative, an estimate of the expected value of the cross-validation index. From the analysis of the model in Fig. 1, none of the modification indices suggested changes to the model that would result in a significant improvement in data–model fit (i.e., a significant decrease in the model χ^2 statistic); thus, we did not report them in Table 1.

Before drawing conclusions regarding the relation between reading and mathematics ability, the question of the 'quality' (i.e., reliability) of the factors should be addressed. Traditionally, this has been accomplished (a) by focusing on the reliability of scores from individual indicator variables, R^2 (i.e., the proportion of variability in an observed variable that can be accounted for by the underlying factor) or (b) by assessing the reliability of scores from linear composites of the indicator variables (e.g., Miller 1995, Raykov 1997). More recently, Hancock and Mueller (2001) suggested a measure (coefficient H) computed from the standardized factor loadings that can be used to assess the reliability of a latent construct itself as reflected by scores from multiple observed variables. One of coefficient H's advantages over traditional construct reliability measures is that it is never less than the best indicator variable's reliability (R^2), thereby drawing information from all indicators in a manner commensurate with their own ability to reflect the construct. For the current example, $H = 0.79$ and $H = 0.82$ for the reading and mathematics ability constructs, respectively, while the R^2 values for the respective factor's indicator variables in Table 1 range between 0.49–0.62 and 0.35–0.73. Thus, the two factors exhibit reasonable and satisfactory levels of

construct reliability since about 80 percent of their variance is explainable by their respective indicators. Table 1 also lists the standardized factor loadings, that is, the estimates of the λ_{ij} parameters in the $\mathbf{\Lambda}$ matrix that indicate the strength and direction of the *a priori* specified causal influences of the latent factors on the observed variables. All factor loadings are positive and significantly different from zero.

Finally, our primary research question can be addressed. The CFA estimate of the correlation between the two latent constructs reading and mathematics ability is 0.51 and significantly different from zero. This estimate indicates a low to moderate positive association between the two constructs of interest, as hypothesized.

4. Conclusion

Confirmatory factor analysis has become established as an important analysis tool for many areas of the social and behavioral sciences. It belongs to the family of structural equation modeling techniques that allow for the investigation of causal relations among latent and observed variables in a priori specified, theory-derived models. The main advantage of CFA lies in its ability to aid researchers in bridging the often-observed gap between theory and observation. For example, an instrument might be developed by creating multiple items for each of several specific theoretical constructs (Fig. 1). Instead of analyzing data with an exploratory factor analysis (where each item is free to load on each factor) and potentially facing a solution inconsistent with initial theory, a CFA can give the investigator valuable information regarding the fit of the data to the specific, theory-derived measurement model (where items load only on the factors they were designed to measure), and point to the potential weakness of specific items. CFA is best understood as a process, from model conceptualization, identification and parameter estimation, to data-model fit assessment and potential model modification. As opposed to exploratory methods, CFA's strength lies in its disconfirmatory nature: models or theories can be rejected, but results might also point toward potential modifications to be investigated in subsequent analyses.

Numerous CFA examples exist in the various disciplines covered in this encyclopedia. The reader might consult (a) in education: O'Grady (1989), Stevens (1995); (b) in psychology: Asmundson et al. (2000), Byrne (1993); (c) in sociology: Alsup and Gillespie (1997), Mulvey et al. (1992); and (d) in health: Galassi et al. (1992), Windle and Dumenci (1999). Some of the more widely used and state-of-the-art SEM software packages for conducting CFA are LISREL (Jöreskog & Sörbom 1996), EQS (Bentler 1997), and AMOS (Arbuckle 1999). Popular textbooks on CFA and more general SEM include Bollen (1989), Hayduk (1987), Loehlin (1998), and Mueller (1996). Social science journals that publish many CFA and SEM applications and methodological developments include: *Educational and Psychological Measurement, Multivariate Behavioral Research, Journal of Experimental Education, Psychological Methods, Sociological Methodology, Sociological Methods and Research, and Structural Equation Modeling: A Multidisciplinary Journal.*

See also: Factor Analysis and Latent Structure: Overview; Latent Structure and Casual Variables

Bibliography

Alsup R, Gillespie D F 1997 Stability of attitudes toward abortion and sex roles: A two-factor measurement model at two points in time. *Structural Equation Modeling: A Multidisciplinary Journal* **4**: 338–52

Arbuckle J L 1999 *AMOS 4.0*. SmallWaters Corporation, Chicago

Asmundson G J G, Frombach I, McQuaid J, Pedrelli P, Lenox R, Stein M B 2000 Dimensionality of posttraumatic stress symptoms: A confirmatory factor analysis of DSM-IV symptom clusters and other symptom models. *Behaviour Research and Therapy* **38**: 203–14

Bentler P M 1990 Comparative fit indexes in structural models. *Psychological Bulletin* **107**: 238–46

Bentler P M 1997 *EQS 5.7*. Multivariate Software, Inc., Encino, CA

Bentler P M, Bonett D G 1980 Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* **88**: 588–606

Bollen K A 1989 *Structural equations with latent variables*. Wiley, New York

Browne M W, Cudeck R 1989 Single sample cross-validation indexes for covariance structures. *Multivariate Behavioral Research* **24**: 445–55

Byrne B M 1993 The Maslach Burnout Inventory: Testing for factorial validity and invariance across elementary, intermediate, and secondary teachers. *Journal of Organizational and Occupational Psychology* **66**: 197–212

Cudeck R, Browne M W 1983 Cross-validation of covariance structures. *Multivariate Behavioral Research* **18**: 147–67

Galassi J P, Schanberg R, Ware W B 1992 The Patient Reactions Assessment: A brief measure of the perceived quality of the patient–provider medical relationship. *International Journal of Psychology* **27**: 639

Gorsuch R L 1983 *Factor Analysis*, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, NJ

Hancock G R, Mueller R O 2001 Rethinking construct reliability within latent variable systems. In: Cudeck R, du Toit S, Sörbom D (eds.) *Structural Equation Modeling: Present and Future—A Festschrift in Honor of Karl Jöreskog*. Scientific Software International, Lincolnwood, IL, pp. 195–216

Hayduk L A 1987 *Structural Equation Modeling with LISREL: Essentials and Advances*. Johns Hopkins University Press, Baltimore, MD

Hu L, Bentler P M 1999 Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* **6**(1): 1–55

Jöreskog K G 1966 Testing a simple structure hypothesis in factor analysis. *Psychometrika* **31**: 165–178

Jöreskog K G 1967 Some contributions to maximum likelihood factor analysis. *Psychometrika* **32**: 443–82
Jöreskog K G, Sörbom D 1996 *LISREL 8*. Scientific Software International, Chicago
Loehlin J C 1998 *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*, 3rd edn. Lawrence Erlbaum Associates, Mahwah, NJ
Marsh H W, Balla J R, McDonald R P 1988 Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin* **103**: 391–410
Miller M B 1995 Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal* **2**: 255–73
Mueller R O 1996 *Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS*. Springer-Verlag, New York
Mulaik S A 1972 *The Foundations of Factor Analysis*. McGraw-Hill, New York
Mulaik S A, James L R, Van Alstine J, Bennett N, Lind S, Stilwell C D 1989 Evaluation of goodness-of-fit indexes for structural equation models. *Psychological Bulletin* **105**: 430–45
Mulvey P W, Miceli M P, Near J P 1992 The Pay Satisfaction Questionnaire: A confirmatory factor analysis. *Journal of Social Psychology* **132**: 139–42
O'Grady K E 1989 Factor structure of the WISC-R. *Multivariate Behavioral Research* **24**: 177–93
Raykov T 1997 Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement* **21**: 173–84
Steiger J H, Lind J M 1980 *Statistically Based Tests for The Number of Common Factors*. Paper presented at the annual meeting of the Psychometric Society. Iowa City, IA
Stevens J J 1995 Confirmatory factor analysis of the Iowa Test of Basic Skills. *Structural Equation Modeling: A Multidisciplinary Journal* **2**: 214–31
Tanaka J S 1993 Multifaceted conceptions of fit in structural equation models. In: Bollen K A, Long J S (eds.) *Testing Structural Equation Models*. Sage, Newbury Park, CA, pp. 10–39
Windle M, Dumenci L 1999 The factorial structure and construct validity of the Psychopathy Checklist-Revised (PCL-R) among alcoholic inpatients. *Structural Equation Modeling: A Multidisciplinary Journal* **6**: 372–93

R. O. Mueller and G. R. Hancock

Factor Analysis and Latent Structure: IRT and Rasch Models

In education and the social sciences, we often ask subjects to respond to a set of items (questions, statements, or tasks) on survey forms, self-report inventories, and mental tests, that are coded as discrete—often dichotomous—variables. In many settings it is natural to think, in analogy with factor analysis (see *Factor Analysis and Latent Structure: Overview*; *Factor Analysis and Latent Structure, Confirmatory*), that there is one or more continuous latent variables for each subject—such as political agency, extrovertedness, or ability in some area of mathematics—that can be measured or estimated using the positive and negative responses to these items. Item Response Theory (IRT), and specializations such as the Rasch Model, refer to a set of statistical modeling and estimation tools for making inferences about one or more continuous latent variables from this kind of multivariate discrete data.

1. Introduction: the Rasch Model

As an example, consider a test of transitive reasoning in school children (e.g., Sijtsma and Verweij 1999). An item may consist of three rods, A, B, and C, of differing lengths. The child is given evidence that Length(A) < Length(B), and Length(B) < Length(C), and is asked to deduce a relationship between Length(A) and Length(C); his or her answer is scored 1 if it is correct and 0 if it is incorrect. Let us say that J such items, involving different types of objects and attributes (balls of varying weight, disks of varying area, etc.) and different numbers of objects, make up a test given to N school children. If the scored responses were continuous we might build a two-way additive ANOVA model in which child i's response to item j, X_{ij}, is the sum (or difference) of the child's achieved transitive reasoning level, and the item's innate difficulty, as determined by the attribute on which objects are compared and the number of objects being compared.

However, since the responses are dichotomous $X_{ij} = 0$ or 1, we need a model specifying the *probability* that $X_{ij} = 1$, in terms of the child's transitive reasoning ability and the item's difficulty. One of the earliest such models, developed by the Danish mathematician Georg Rasch (Rasch 1960), specifies

$$P_j(\theta_i) \equiv P[X_{ij} = 1 \mid \theta_i, \beta_j] = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}, \quad i = 1, \ldots, N, j = 1, \ldots, J \quad (1)$$

where θ_i represents child i's transitive reasoning level, and β_j represents the *difficulty* of the item: as the latent variable θ_i increases, the probability of getting the item right increases, and as the difficulty parameter β_j increases, the probability of getting the item right decreases. The Rasch model, together with similar models using different forms for $P_j(\theta_i)$ developed by Jane Loevinger (Loevinger 1948) and Frederic M Lord (Lord 1952), became the core of what is now item response theory.

On the logit scale, the Rasch model has exactly the two-way additive ANOVA structure we sought:

$$\text{logit}\, P_j(\theta_i) = \ln \frac{P_j(\theta)}{1 - P_j(\theta_i)} = \theta_i - \beta_j.$$

We further assume local independence among item responses, i.e., all X_{ij} are conditionally independent,

given θs and βs. Letting x_{ij} denote the observed values of X_{ij}, we may write the likelihood for the Rasch model as

$$P[X_{ij} = x_{ij}, i = 1, \ldots, N, j = 1, \ldots, J \mid \theta_1, \ldots, \theta_N; \beta_1, \ldots, \beta_J] = \prod_{i=1}^{N} \prod_{j=1}^{J} P_j(\theta_i)^{x_{ij}} [1 - P_j(\theta_i)]^{1-x_{ij}}. \quad (2)$$

An up-to date account of the Rasch model and its various extensions is given by Fischer and Molenaar (1995), to whom we also refer for historical notes and references.

The Rasch model in Eqn. 1 has the interesting invariance property that the odds ratio comparing two items j and k is

$$\frac{P_j(\theta_i)}{1 - P_j(\theta_i)} \cdot \frac{1 - P_k(\theta_i)}{P_k(\theta_i)} = \exp \beta_k - \beta_j$$

independent of θ_i. Hence in our example, the 'difficulty' of two items involving different transitive reasoning settings can be compared, in principle, using any convenient sample of subjects regardless of their transitive reasoning level (in practice, subjects still contribute at least 'small sample bias' to the estimated βs). Similarly, the odds ratio comparing two subjects is independent of the item difficulty parameters. These two invariance properties are instances of 'specific objectivity,' a property Rasch considered important in defining a measurement model and from which Eqn. 1 can be derived (see Fischer and Molenaar 1995). The Rasch model is also closely related to the Bradley–Terry model of paired comparisons. Consider two items j and k for which $\beta_j < \beta_k$. For any subject i who got only one item right, it follows from Eqn. 1 that the probability that the easier item was correct is

$$P[X_{ij} = 1 \mid X_{ij} + X_{ik} = 1, \theta_i, \beta_j, \beta_k] = \frac{\exp(\beta_k - \beta_j)}{1 + \exp(\beta_k - \beta_j)} \quad (3)$$

independent of θ_i; this is exactly the Bradley–Terry model with parameters $\phi_j = -\beta_j$. The lack of dependence on θ_i on the right can also be used to construct empirical tests of the Rasch model, since the probability in Eqn. 3 can be estimated without assuming the Rasch model and should be the same in any subpopulation of subjects in which the Rasch model applies.

2. Estimation in the Rasch Model

Two estimation methods are traditionally associated with the Rasch model, joint maximum likelihood (JML) and conditional maximum likelihood (CML). One can maximize Eqn. 2 jointly in the θs and βs, by numerically solving the 'normal equations' obtained by setting all partial derivatives of the logarithm of Eqn. 2 equal to zero; the maximizing values $\hat{\theta}_i$ and $\hat{\beta}_j$ are called joint maximum likelihood estimators (JMLEs). However, the JMLEs turn out to be *inconsistent*, or asymptotically biased: suppose we are primarily interested in estimating β_js; it need not be the case that each $\hat{\beta}_j$ converges to β_j as we increase the subject sample size N, keeping the number of items J fixed. Indeed, as N increases, more θ_is are also added to the model: intuitively, some information must be expended on estimating them, and there is not enough information left over to improve the estimates of the β_is. One solution to this problem is to exploit the fact that in the Rasch model, a sufficient statistic for each θ_i is subject i's total number correct score, $X_{i+} = \Sigma_{j=1}^{J} X_{ij}$. If the likelihood in Eqn. 2 is divided by the joint likelihood for the observed x_{i+}s, the resulting conditional likelihood contains only the β_j parameters, and the conditional maximum likelihood estimators (CMLEs) $\hat{\beta}_j$ obtained by maximizing this new likelihood are consistent (asymptotically unbiased) as N grows and J stays fixed. For this reason, CML estimates are usually preferred over JML estimates. Additional details are provided by Andersen (1980), whose earlier work demonstrated the inconsistency problem with JML, and established consistency of CML for the Rasch model.

Another approach to solving the inconsistency problem with JML is to assume the θ_i are independent random effects following a common (discrete or continuous) distribution $f(\theta \mid \lambda)$. Integrating over θ_i for each subject yields the marginal likelihood

$$\prod_{i=1}^{N} \int \prod_{j=1}^{J} P_j(\theta_i)^{x_{ij}} [1 - P_j(\theta_i)]^{1-x_{ij}} f(\theta_i \mid \lambda) \, \mathrm{d}\theta_i \quad (4)$$

and maximizing this with respect to β_is and λ (using an E-M algorithm, for example) yields what are called maximum marginal likelihood (MML) estimates $\hat{\beta}_j$ and $\hat{\lambda}$. These MML estimates are also consistent (asymptotically unbiased). The method based on Eqn. 4 can be interpreted as an empirical Bayes method; it also links the Rasch model directly with other latent variable approaches such as factor analysis, where the latent variable θ_i is treated as an unobserved random variable, or equivalently as completely missing data (see *Factor Analysis and Latent Structure: Overview*).

Finally, it can be shown that under Eqn. 4, the marginal distribution of the data follows a log-linear model for the 2^J table of the form

$$\ln p(x_{i1}, \ldots, x_{iJ}) = \alpha - \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{k=0}^{J} \gamma_k \, \chi(x_{i+} = k)$$

where $\chi(S) = 1$ if statement S is true, and 0 otherwise. This is the log-linear model of quasisymmetry; the connection between the Rasch model and the quasi-

symmetry model has been independently discovered innumerable times in the literature. Lindsay et al. (1991) give a reasonably complete account of the connection with quasisymmetry and consequences for the identifiability of the distribution $f(\theta|\lambda)$ in a semiparametric formulation of Eqn. 4.

3. Parametric IRT

Other parametric IRT models have been developed for situations in which the Rasch model does not fit, while retaining the ideas of dichotomous ($X_{ij} \in \{0, 1\}$) or at least polytomous ($X_{ij} \in \{0, 1, \ldots, K-1\}$) item scores, a unidimensional (or low-dimensional) latent variable θ_i characterizing subjects, a low-dimensional parameter β_j characterizing items, and local (conditional) independence of all item responses X_{ij} given all parameters. An accessible review of developments in parametric IRT in the 1960s, 1970s and 1980s is given by Hambleton et al. (1991).

A simple generalization of the Rasch model that is often useful for dichotomously-scored items,

$$P_j(\theta_i) = \frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]} \tag{5}$$

is called the two-parameter logistic (2PL) model. In the 2PL model, the slope parameters α_{jk} are called 'discrimination parameters,' and they control the Fisher information for estimating the θs. Again, the location parameters β_j play the role of 'difficulty parameters'; they also determine the value of θ maximizing the Fisher information. The three-parameter logistic (3PL) model adds a nonzero lower asymptote to Eqn. (3), to model random or exchangeable response behavior by low-θ subjects. Further generalizations, discussed in the next several paragraphs, are treated at length by Fischer and Molenaar (1995) and van der Linden and Hambleton (1997).

Additional covariates and other structures can be incorporated into these models as regression terms. For example, in the linear logistic test model (LLTM), the difficulty parameters β_j in the Rasch model are rewritten as linear combinations of K basic parameters η_k with weights q_{jk}, and

$$\text{logit}\, P_j(\theta_i) = \theta_i - \sum_{k=1}^{K} q_{jk}\eta_k.$$

The matrix $Q = [q_{jk}]$ is usually obtained as a regression model matrix from *a priori* analysis of the items; for example, in the transitive reasoning example described in Sect. 1, the q_{jk} might indicate what kind and how many objects were used for each item. By contrast, multidimensional compensatory IRT models decompose the unidimensional θ_i parameter into an item-dependent linear combination of underlying traits, e.g.,

$$\text{logit}\, P_j(\theta_i) = \sum_{d=1}^{D} \alpha_{jd}\theta_{id} - \beta_j.$$

For example, mathematics word problems may involve both a math component, θ_{i1}, and a verbal component θ_{i2} in different proportions in different problems, as determined by the discrimination parameters α_{jd}. In some settings, the α_{jd} may be fixed *a priori*, like the q_{jk} in the LLTM model, and in other settings they may be estimated analogously to factor loadings in Factor Analysis. Multiplicative or conjunctive IRT models combine unidimensional models for components of response in a product $P_j(\theta_i) = \prod_d^D = 1 P_{jd}(\theta_{id})$ where $P_{jd}(\theta_{id})$ are parametric unidimensional dichotomous response functions. The usual interpretation is that the $P_{jd}(\theta_{id})$ are probabilities of correctly performing component skills or subtasks, which must be done in conjunction in order to generate a correct response to the item itself; this model is sometimes referred to as the multicomponent latent trait model (MLTM).

Probit versions of all the above models, in which the logit function is replaced with $\Phi^{-1}(p)$, the inverse standard normal distribution function, are also in use. Aside from a trivial rescaling of the parameters, there is little difference in practice between the probit and logit models; the choice depends mostly on computational expedience.

IRT model can also be extended to handle polytomous or multicategory items, $X_{ij} \in \{0, 1, 2, \ldots, K-1\}$, in much the same way that logistic regression models are extended polytomous data: in the 'graded response model' (GRM) the cumulative logit, logit $P[X_{ij} > c|\theta_i, \beta_j]$, is linear in θ_i; in the 'sequential model' (SM) the continuation-ratio logit, logit $P[X_{ij} > c\,|X_{ij} > c-1, \theta_i, \beta_j]$, is linear in θ_i; and in the 'partial credit model' (PCM) the adjacent-category logit, logit $P[X_{ij} = c+1|X_{ij} \in \{c, c+1\}, \theta_i, \beta_j]$, is linear in θ_i. Again, probit reformulations of these models are straightforward. When the discrimination parameters are all non-negative, the cumulative response curves in any of these models is nondecreasing (Hemker et al. 1997), a condition referred to as 'monotonicity.' By contrast, IRT formulations of direct-response probabilistic unfolding models, also known as proximity models or parallelogram models, have unimodal response curves that peak near the location β_j of the item.

As was the case for the Rasch model discussed in Sect. 2, JML estimates for IRT models are relatively easy to set up but generally inconsistent (asymptotically biased); CML methods can be used when simple sufficient statistics are available. MML methods based on applying an E-M algorithm to Eqn. 4, with $P_j(\theta_i)$ redefined according to one of the above

First level: $$X_{ij} \sim P(\theta_i; \beta_j)^{X_{ij}}[1 - P(\theta_i; \beta_j)]^{(1-X_{ij})},$$

$$i = 1, \ldots, N;\ j = 1, \ldots, J$$

Second level: $$\theta_i \sim f_i(\theta|\lambda_f), \quad \text{each } i$$

$$\beta_j \sim g_j(\beta|\lambda_g), \quad \text{each } j$$

Third level: $$\lambda_f \sim \phi_f(\lambda_f)$$

$$\lambda_g \sim \phi_g(\lambda_g)$$

Figure 1
Two-way hierarchical structure for N individuals and J dichotomous response variables. Factors in the first level are independent and are multiplied together to form a likelihood for the response data matrix $[X_{ij}]$. Factors in the second and third levels are also independent, and covariates and special structure for the latent variables and item parameters may be introduced in these levels

parametric forms, are most common, however. The recent development of Markov Chain Monte Carlo (MCMC) integration methods for Bayesian statistics (see *Markov Chain Monte Carlo Methods*) provides an alternative to these methods. Patz and Junker (1999) sketch a general MCMC methodology for parametric IRT models, and Johnson and Albert (1999) survey a range of related applications from a working Bayesian statistician's perspective.

The IRT toolbox has been pushed forward in recent years by the needs of both large-scale educational assessment surveys (e.g., Johnson et al. 1994) and cognitively based assessment models (e.g., Nichols et al. 1995). To accommodate finer structure in the latent space, as well as to incorporate subject and item covariates in the model, it is useful to reformulate the IRT framework as a two-way hierarchical Bayes structure for N individuals and J response variables, as in Fig. 1. In addition to the item response functions $P_j(\theta_i) = P(\theta_i; \beta_j)$ in Fig. 1, f_i and g_j are prior distributions on the latent variables θ and item parameters β respectively (both of which may be multidimensional), and λ_f and λ_g represent sets of hyperparameters needed to specify these prior distributions. The model in Fig. 1 is expressed for dichotomous items but can be generalized easily to polytomous items or combinations of item types, incomplete data, and so forth (e.g., Patz and Junker 1999). The MML formulation discussed in the previous paragraph is obtained by taking $f_i(\theta)$ to be a single population distribution not dependent on j for the latent trait, and taking g_j to be flat priors not depending on j for the item parameters.

4. *Nonparametric IRT*

Nonparametric IRT refers to several related methodologies for working with item response data without fully committing to any particular well-known family of parametric item response theory models. Some topics in nonparametric IRT are considered in van der Linden and Hambleton (1997) and Boomsma et al. (2001), and several current issues are surveyed in a special issue of *Applied Psychological Measurement* (Junker and Sijtsma 2001).

Modern nonparametric scaling methods include the 'essential independence' (EI) approach of Stout and his students and colleagues (e.g., Stout et al. 1996), and the Mokken scaling approach (e.g., Hemker et al. 1995); both methods are related to Cronbach's alpha bound on reliability. Mokken scaling is very good at finding groups of items that are highly discriminating among subjects; Stout's methods tend to identify groups of items that narrowly satisfy LI/EI, M, and U better. Ramsay (e.g., Ramsay 1996) has developed computational tools for estimating item response functions nonparametrically, and for visualizing the dimensionality of θ in terms of the surface in the J-dimensional unit cube generated by the joint likelihood for $X_{i1}, \ldots, X_{iJ}$ as θ_i varies.

5. *Some Applications*

Parametric and nonparametric IRT methods are routinely used to assess the quality of individual items and sets of items in educational measurement work (Linn 1989) and to elucidate experimental design and statistical estimation issues when performances of examinees who sit for different versions of the same exam must be compared; such work goes under the name test equating. A similar problem arises in the scoring of sequentially-designed, computerized adaptive tests (e.g., Sands et al. 1997). IRT modeling has also elucidated research into the assessment of sociological bias (differential item functioning) on standardized test items (Holland and Wainer 1993) and the detection and diagnosis of subject outliers (Meijer 1996). Related work includes the study of rater effects (e.g., Patz et al. 2000) and matrix-sampling and the incorporation of group effects and other item and subject covariates in educational assessment survey work (Johnson et al. 1994).

In the broader social sciences and related areas, IRT has also played a substantial role. The Rasch model has long been applied in social survey work; see, for example, Duncan (1985). The Rasch model and other

IRT models can also be applied to model subject heterogeneity in estimating a closed population from multiple-recapture census data (see *Censuses: History and Methods*); Fienberg et al. (1999) recently surveyed this methodology and compared Bayesian IRT models with standard log-linear models in this context. Applications to multiple outcomes of designed experiments and to panel data are indicated in the volumes by Fischer and Molenaar (1995) and van der Linden and Hambleton (1997). Sijtsma and Verweij (1999) apply nonparametric IRT to scale construction in developmental psychology; and Nichols et al. (1995) collect many applications of IRT and related Bayesian inference network models (see *Bayesian Graphical Models and Networks*; Mislevy 1996) in cognitive diagnosis. Applications of parametric and nonparametric IRT to psychiatric scales include Gibbons et al. (1985), Santor et al. (1995) and Kim and Pilkonis (1999). In biostatistics Legler and Ryan (1997) use IRT to model multiple physical outcomes in the study of birth defects.

6. Summary

Item response theory (IRT) has grown from its roots in postwar mental-testing problems, through intensive use in educational measurements in the 1970s, 1980s, and 1990s, to become a mature statistical toolkit for modeling of multivariate discrete response data using subject-level latent variables. Applications of IRT can be found throughout the social sciences and related areas, from education, psychology, economics, and demography to medical research.

Most parametric IRT models would be recognizable by modern statisticians as mixed-effects multivariate generalized linear models, but IRT has benefited from interaction with all parts of the statistical community. Almost any assessment phenomenon—from between-subject dependence due to institutional or sociological factors, to behavioral aspects of raters, to the analysis of item responses in terms of requisite subject or item features—can be expressed in the hierarchical mixture/Bayes modeling framework of Fig. 1, because of its conceptual simplicity. Recent advances in computation, and MCMC methods in particular, have made it possible to estimate a vastly wider variety of these models than would have been imaginable even in the early 1990s however the more complicated models typically also make higher sample-size demands. Nonparametric IRT approaches also rely on computationally intensive methods, including spline and kernel smoothing and bootstrap techniques, to estimate and test probability inequalities, stochastic ordering properties, and similar features of the models. Questions motivating both parametric and nonparametric IRT modeling inevitably involve identifying the phenomena that are worth detailed modeling, and seeing if the computational and data collection machinery can be pushed to be informative about reasonable models of these phenomena.

Hambleton et al. (1991) provide a fairly straightforward introduction to IRT in educational testing; Andersen (1980) presents the Rasch model in the context of other statistical models for the social sciences; and Fischer and Molenaar (1995) and van der Linden and Hambleton (1997) provide modern technical accounts. A broad range of current research issues in IRT is collected in Boomsma et al. (2001).

Bibliography

Andersen E B 1980 *Discrete Statistical Models with Social Science Applications*. North-Holland, New York

Boomsma A, Snijders T A B, van Duijn M A J (eds.) 2001 *Essays on Item Response Theory*. Springer-Verlag, New York

Duncan O D 1985 Some models of response uncertainty for panel analysis. *Social Science Research* **14**: 126–41

Fienberg S E, Johnson M S, Junker B W 1999 Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of Royal Statistical Society: Series A* **162**: 383–405

Fischer G H, Molenaar I W (eds.) 1995 *Rasch Models: Foundations, Recent Developments, and Applications*. Springer-Verlag, New York

Gibbons R D, Clark D C, von Ammon Cavanaugh S, Davis J M 1985 Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research* **19**: 43–55

Hambleton R K, Swaminathan H, Rogers H J (eds.) 1991 *Fundamentals of Item Response Theory*. Sage, Newbury Park, CA

Hemker B T, Sijtsma K, Molenaar I W 1995 Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement* **19**: 337–52

Hemker B T, Sijtsma K, Molenaar I W, Junker B W 1997 Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psymka* **62**: 331–47

Holland P W, Wainer H 1993 *Differential Item Functioning*. Erlbaum, Hillsdale, NJ

Johnson V E, Albert J 1999 *Ordinal Data Modeling*. Springer-Verlag, New York

Johnson E G, Mislevy R J, Thomas N 1994 Theoretical background and philosophy of NAEP scaling procedures. In: Johnson E G et al. (eds.) *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading*. OERI, US Dept. of Ed., Washington, DC, Chap. 8, pp. 133–46

Junker B W, Sijtsma K 2001 Nonparametric IRT in action: An overview of the special issue. *Applied Psychological Measurement* **25**:

Kim Y, Pilkonis P A 1999 Selecting the most informative items in the IIP scales for personality disorders: An application of item response theory. *Journal of Personality Disorders* **13**: 157–74

Legler J M, Ryan L M 1997 Latent variable models for teratogenesis using multiple binary outcomes. *Journal of the American Statistical Association* **92**: 13–20

Linn R L (ed.) 1989 *Educational Measurement*, 3rd edn. Macmillan, New York

Lindsay B, Clogg C C, Grego J 1991 Semiparametric estimation in the Rasch model and related exponential response models,

including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86**: 96–107
Lord F M 1952 A theory of test scores. *Psychometric Society*, New York
Loevinger J 1948 The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin* **45**: 507–30
Meijer R R 1996 Person-fit research: An introduction. *Applied Measurement in Education* **9**: 3–8
Mislevy R J 1996 Test theory reconceived. *Journal of Educational Measurement* **33**: 379–416
Nichols P D, Chipman S F, Brennan R L (eds.) 1995 *Cognitively Diagnostic Assessment*. Erlbaum, Hillsdale, NJ
Patz R J, Junker B W 1999 Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics* **24**: 342–66
Patz R J, Junker B W, Johnson M S submitted The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*
Ramsay J O 1996 A geometrical approach to item response theory. *Behmka* **23**: 3–17
Rasch G 1960 Probabilistic models for some intelligence and attainment tests. University of Chicago Press, Chicago
Sands W A, Waters B K, McBride J R 1997 *Computerized Adaptive Testing: From Inquiry to Operation*. American Psychological Association, Washington, DC
Santor D A, Zuroff D C, Ramsay J O, Cervantes P, Palacios J 1995 Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychiatry Assessment* **7**: 131–9
Sijtsma K, Verweij A C 1999 Knowledge of solution strategies and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement* **23**: 55–68
Stout W, Habing B, Douglas J, Kim H R, Roussos L, Zhang J 1996 Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement* **20**: 331–54
van der Linden W J, Hambleton R K (eds.) 1997 *Handbook of Modern Item Response Theory*. Springer-Verlag, New York

B. W. Junker

Factor Analysis and Latent Structure: Overview

The terms factor analysis and latent structure refer to two aspects of essentially the same problem. Both are concerned with statistical problems in which some of the variables are latent, meaning that they are unobservable. Information about the latent variables has, therefore, to be obtained indirectly through indicators, also known as manifest or observable variables. The terminology of the subject reflects the diverse origins of the subject accumulated over almost a century. 'Factor' in this context is a synonym for latent variable. From a theoretical perspective, the only distinctive thing about a latent variable model is the presence of unobservable variables. In all other respects the normal methods of statistics apply.

Indeed, a latent variable problem can be regarded as a standard statistical problem in which the data on the latent variables are missing. There is no essential difference, for purposes of analysis, between a problem in which some of the data are never obtained and one in which some are lost.

1. Introduction

Latent variables arise mainly, but not exclusively, in the social sciences. This is because social science often deals in concepts which are constructs rather than the directly measurable variables which are typical of physical sciences. The earliest example, and still one of the most important, is general intelligence or 'g.' This goes back to Spearman (1904) and was constructed to describe the variation among individuals in what appeared to be common to a wide range of mental tests. Psychology and sociology abound in such latent variables. Attitudes as well as abilities are all spoken of in the discourse of these subjects as things which occur in varying amounts and which, therefore, appear in the theory as quantitative variables. In economics, variables like business confidence play a similar role. It could be justly claimed that the aspiration of these subjects to be regarded as sciences depends on the success with which latent variables can be quantified.

Latent variables can be classified into several types. Many, like intelligence, are conceived of as continuous in which case we are looking for a scale on which individuals can be located. In other contexts it is more appropriate to think of the latent variable as categorical. In that case individuals are supposed to belong to one of several classes which may or may not be ordered. What is true of the latent variables is, of course, true for the manifest variables and the only essential difference between the various methods is in the types of variables for which they are appropriate.

A convenient way of displaying the relationship between the methods is to introduce the fourfold classification of Table 1. We classify the manifest and the latent variables as metrical or categorical. In the former case we mean that they take values on some

Table 1
Classification of latent variable methods

	Manifest variables	
	Metrical	Categorical
Latent variables		
Metrical	Factor analysis	Latent trait analysis
Categorical	Latent profile analysis	Latent class analysis

continuous or discrete scale of measurement; in the latter, that they fall into categories.

This classification is not exhaustive. It does not, for example, include cases where both kinds of variable occur among the manifest (or latent) variables. Neither does it distinguish between continuous and discrete metrical variables or between ordered and unordered categorical variables. Nevertheless, it does cover the most widely used methods on which this overview focuses.

2. Historical Perspective

Factor analysis was invented by Spearman (1904) for the specific purpose of operationalizing the notion of general intelligence. He remained wedded to the view that there was a single dominant latent variable of human ability which he named 'g' which accounted for most of the variation in observed performance. Thurstone (1947) emphasized the multifactorial approach and, for that purpose, generalized Spearman's treatment to allow for several factors. Factor analysis was ahead of its time in two senses. First, it was a sophisticated multivariate technique introduced long before the statistical technology required was available. In consequence, its development was somewhat idiosyncratic and its separation from the statistical mainstream gave it an alien appearance which persists to this day and still causes suspicion and misunderstanding. Second, it is a computer intensive technique whose full potential could only be realized with the arrival of the powerful electronic computers since the 1980s. This lack distorted the early development of the subject by the need to concentrate on the search for computational shortcuts.

Latent structure analysis, which covers the remaining cells of Table 1, was introduced by Lazarsfeld after the Second World War, with a view to applications in sociology. A book-length treatment, by Lazarsfeld and Henry (1968), has held the field for many years as the definitive source of information but has now been superseded by a new generation of books of which Heinen (1996) is both recent and comprehensive.

Latent structure analysis came on the scene half a century after factor analysis, and in a different disciplinary context. For that reason there was, until very recently, little cross-fertilization between the two fields. When viewed from a level of sufficient generality, the methods are essentially the same.

In the postwar era more powerful methods became available but the whole field largely has remained a set of distinct subspecialisms each with its own literature and technology. It is over 40 years since Anderson (1959) pointed out that all latent methods shared a common conceptual framework but his insight has been slow to yield its full potential. The classification of methods as given in Table 1 was made the basis of a common approach in Bartholomew (1984) the culmination of which is to be seen in Bartholomew and Knott (1999).

3. The Basic Conceptual Framework

First, we describe the key ideas in nonmathematical terms and then, in the following section, express them more precisely. The central idea on which the treatment is based is that of a probability model in which all variables are represented by random variables. This is so natural to the modern statistician as to pass without comment, but this was not so at the time when factor analysis was being developed. Even today the notion is not always clear. A statistical model is, essentially, a statement about the joint distribution of a set of random variables. Model-building is the process of specifying that distribution.

The discussion is based on a simple example. Consider the case of arithmetical ability and suppose that this varies from one individual to another along a one-dimensional scale but cannot be observed directly. Test items are chosen which depend on this particular skill, and are administered to a large sample of subjects. Our problem is to see whether our hypothesis about the existence of a single latent variable is correct and, if so, to provide a means of locating subjects on that scale on the evidence of their test score.

A simple intuitive approach, which has been used in the educational world for generations, is to add up, or average, the scores for any individual and treat the average as an approximation to the value of the underlying latent variable for that individual. From this it is a short step to a rudimentary model according to which the observed score is equal to the 'true score' plus a random deviation. The deviations will tend to cancel out on addition. All latent variable models can be regarded as elaborations of this basic idea.

The first step is to specify how the observed score should vary when the value of the latent variable is fixed. The next is to deduce what the joint distribution of the scores ought to be when there is no variation. The final step is to introduce the distribution of the latent variable(s)—the true scores. Together these give the overall distribution of the scores which can then be compared with the observed distribution. If the two are in agreement the model is supported, if not it must be replaced with something better.

The second stage in the model building exercise is more problematical. By definition, a latent variable can never be observed so its distribution cannot be observed either. The choice of prior distribution is thus a matter of convention. If it is chosen to be normal, that is equivalent to agreeing to adopt a scale which renders the distribution normal.

All of the models covered by Table 1 have these two elements: (a) a statement of how the scores vary when

the latent variable(s) is fixed and (b) a convention about the form of the distribution of the latent variable in the population from which the sample of those tested has come. Once these two things are known probability theory provides the means to go in the reverse direction and deduce how the latent variable would vary if the observed score were to be fixed.

4. Formal Statement of the Conceptual Framework

The last section set out the key ideas underlying all latent variable models. Here they are expressed more formally in mathematical terms. The nonmathematical reader who is interested only in specific models may pass immediately to the next section.

Denote the set of indicators by the p-vector $\mathbf{x}$. Its elements $x_1, x_2, \ldots, x_p$ may be continuous or categorical. The latent variables are denoted by the q-vector $\mathbf{z}$. In the example above there was only one latent variable, arithmetical ability, but in general, there may be several.

Any model must specify, completely or partially, the joint probability distribution of $\mathbf{x}$ and $\mathbf{z}$ denoted by $f(\mathbf{x}, \mathbf{z})$. The distinctive thing about this kind of model is that the values of $\mathbf{z}$ can never be observed. For this reason everything must be expressed in terms of the distribution of the observable variables $\mathbf{x}$. The relationship between the two is

$$f(\mathbf{x}) = \int f(\mathbf{x}, \mathbf{z})\, d\mathbf{z} \tag{1}$$

Here and subsequently equations will be written as if the variables were all continuous—hence the integral in (1). This is purely for convenience and the translation needed to cover other cases does not affect the logic of the argument. The next step is to write

$$f(\mathbf{x}, \mathbf{z}) = f(\mathbf{z}) f(\mathbf{x} \mid \mathbf{z}) \tag{2}$$

This decomposition provides us with the two elements to which the description of the last section referred—namely a statement of how the manifest variables would vary if $\mathbf{z}$ were known, given here by $f(\mathbf{x} \mid \mathbf{z})$, and the prior distribution of $\mathbf{z}$. If both parts of (2) can be specified, $f(\mathbf{x})$ can be determined and thus the way is open to estimate any parameters of the model from a likelihood function constructed from $f(\mathbf{x})$. To scale $\mathbf{z}$, that is to find value for $\mathbf{z}$ for an individual with a given $\mathbf{x}$, the posterior distribution given by

$$f(\mathbf{z} \mid \mathbf{x}) = f(\mathbf{z}) f(\mathbf{x} \mid \mathbf{z}) / f(\mathbf{x}) \tag{3}$$

is needed. A scale value, or factor score, for a particular $\mathbf{z}$ can then be found by using some measure of location of its posterior distribution such as $E(z_i \mid \mathbf{x})$.

This program is deceptively simple. It merely maps out some of the consequences of regarding $\mathbf{x}$ and $\mathbf{z}$ as random variables. To turn it into a model we have to specify $f(\mathbf{z})$ and $f(\mathbf{x} \mid \mathbf{z})$. Here we face a fundamental problem because the decomposition of (3) is not unique—there are infinitely many possible pairs each leading to the same $f(\mathbf{x})$ and hence indistinguishable empirically from one another.

The choice of $f(\mathbf{z})$ is resolved by appealing to the fact established above that $\mathbf{z}$ is a latent variable and hence that it can be constructed to have any convenient distribution. For instance, as a matter of convention one may choose the zs to be standard normal and mutually independent.

The core of the model-building exercise centers on the choice of the conditional distribution $f(\mathbf{x} \mid \mathbf{z})$. This has to be considered in the context of individual problems but there is one further simplification which applies to all models. This simplification involves making explicit what is already implicit in our formulation. The fact that the $\mathbf{x}$s are supposed to be indicators of $\mathbf{z}$ implies that the xs will be mutually correlated—because of their common dependence on some, at least, of the zs. Hence, if $\mathbf{z}$ were to be fixed, the source of those correlations would vanish. Conditional on $\mathbf{z}$ the xs would then be independent. If this were not so it would be inferred that there was at least one other z, not already included, giving rise to the correlations remaining. Sufficient zs therefore have to be included in the model to eliminate the correlations after conditioning of $\mathbf{z}$, It then follows that

$$f(\mathbf{x} \mid \mathbf{z}) = \prod_{i=1}^{q} f_i(x_i \mid \mathbf{z}) \tag{4}$$

This is often referred to as the assumption of conditional (or local) independence but it is not an assumption in the usual sense. It is, rather, a statement of what is meant by saying that $\mathbf{z}$ explains the interdependence of the xs. It is, essentially, a definition of q as the smallest integer for which an equation of the form (4) can be constructed.

How this may be done in particular cases will be shown in the following section, but there is an important general principle which offers a fruitful way forward. Distributions of the exponential family are widely used in statistical modeling because they lead to estimates with highly desirable properties such as sufficiency. They provide a rationale for many of the most important models used in statistics, such as generalized linear models (see *Neural Networks and Related Statistical Latent Variable Models*).

Essentially the same properties can be utilized here. They are set out in the chapter on the General Linear Latent Variable Models in Bartholomew and Knott (1999). The essential idea is to make what is called the 'natural' or 'canonical' parameter of the one-parameter exponential family distribution a linear function of the latent variables. If that is done it turns out that the posterior distribution of $\mathbf{z}$ given $\mathbf{x}$ of (3) depends

on the xs only through q linear combinations of the xs which we may write

$$X_j = \sum_{i=1}^{p} \alpha_{ij} x_i \quad (j = 1, 2, \ldots, q) \tag{5}$$

This is a remarkable result which justifies the common practice of 'estimating' latent variables from sums, or weighted sums, of the indicators. Technically speaking the Xs are sufficient for $\mathbf{z}$ in the sense that, if we know $\mathbf{X}$, there is no further information in the data about $\mathbf{z}$.

5. Special Cases

5.1 Factor Analysis

This is the oldest and most widely used latent variable method. It is usually written

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{z} + \mathbf{e} \tag{6}$$

where $E(\mathbf{z}) = E(\mathbf{e}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{e}) = \boldsymbol{\Psi}$, a diagonal matrix, $\mathrm{cov}(\mathbf{z}, \mathbf{e}) = \mathbf{0}$ and $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$. An equivalent way of writing it, which conforms more closely to the general treatment above, is to specify the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ as

$$\mathbf{x} \,|\, \mathbf{z} \sim N(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{z}, \boldsymbol{\Psi}) \tag{7}$$

which makes it clear that $\mathbf{z}$ influences $\mathbf{x}$ only through its mean which is linear in $\mathbf{z}$. The marginal distribution of $\mathbf{x}$ is $N(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})$ and it is from this that the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ must be estimated. The matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ is sometimes referred to as the covariance structure.

Fitting the model consists in determining that $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ which make the fitted and observed covariances of the xs as close as possible. This can be done without invoking the distributional assumptions at all since the covariance structure derived from (6) is the same, regardless of the assumptions about the form of the distributions of $\mathbf{z}$ and $\mathbf{e}$. However, without some such assumption nothing can be said about goodness of fit or the sampling variation of the estimators.

Posterior analysis allows the distribution of $\mathbf{z}$ given $\mathbf{x}$ to be determined, which for the factor model is

$$\mathbf{z} \,|\, \mathbf{x} \sim N(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), (\mathbf{I} + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})) \tag{8}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. As predictions of $\mathbf{z}$ given $\mathbf{x}$ we can use the mean values of this distribution, replacing the parameters by sample estimates.

5.2 Interpretation and Rotation

Interpretation of factors involves procedures which are common to all linear latent variable models. They are most familiar in the context of factor analysis so, although treated here, their generality must be emphasised. The problem may arise in two forms. If there is some idea of what factors to expect (as in the arithmetical ability example above) the question is how to recognize them. If, on the other hand, the analysis is purely exploratory, further guidance is needed on how to assign meaning to any factor that is found. The essential idea is to look at the relationships between the latent and manifest variables. This is facilitated by noting that the element λ_{ij} of $\boldsymbol{\Lambda}$ may be interpreted as the correlation between x_i and z_j. Manifest variables which are highly correlated with z_1 say, have a lot in common with z_1 which, in turn is highly influential in determining those xs. It must then be asked what it is that those xs have in common with one another which they also share with z_1. Interpretation is often facilitated by a process known as rotation. Thus far the treatment has supposed that there was only one solution to the problem of fitting the model. In fact there are infinitely many models which all predict the same covariance structure. These solutions can be generated from one another by a procedure known (because of its geometrical interpretation) as rotation. Each such solution represents an equivalent way of describing the latent space, but some may be easier to interpret than others. For example, a solution for $\boldsymbol{\Lambda}$ which has, at most, one nonzero value in each row implies that the xs have been divided into disjoint groups, each group depending on a single latent variable. That latent variable is then interpreted in the light of what that subset has in common.

5.3 The Latent Class Model

If the latent space consists of a finite (usually small) set of classes and if the manifest variables are also categorical the bottom right cell of Table 1 applies. The commonest case is where the manifest variables are binary. If there are c classes the prior distribution $f(\mathbf{z})$ becomes a discrete probability distribution over the classes, η_j being the probability of belonging to class j. In this case these probabilities can be treated as unknown parameters to be estimated. For convenience, the two values which each x_i takes are coded as 1 and 0, the conditional distributions $f_i(x_i \,|\, z)$ consists of two probabilities: $\Pr\{x_i = 1 \,|\, z\}$ and $\Pr\{x_i = 0 \,|\, z\}$. The natural choice for the distribution of x_i is the Bernoulli distribution written as

$$f_i(x_i \,|\, j) = \{\pi_i(j)\}^{x_i}\{1 - \pi_i(j)\}^{1-x_i} \qquad x_i = 0, 1 \quad j = 1, 2, \ldots, c \tag{9}$$

The joint probability distribution of $\mathbf{x}$ is then

$$f(\mathbf{x}) = \sum_{j=1}^{c} \eta_j \prod_{i=1}^{p} \{\pi_i(j)\}^{x_i}\{1 - \pi_i(j)\}^{1-x_i} \tag{10}$$

This can be used to construct the likelihood function from which the parameters may be estimated.

Posterior analysis for this case is concerned with allocating individuals to latent classes after **x** has been observed and this is done easily by substituting into (3). This gives us the probability that an individual with observed vector **x** belongs to each of the classes.

The number of parameters in this model increases rapidly with c and serious problems of identifiability arise if there are more than four or five classes. Also the standard errors of the parameter estimates increase rapidly as c increases.

5.4 Latent Trait Models

These models were devised, primarily, for use in educational testing where the latent trait refers to some ability. There is, thus, usually one latent variable, representing the ability, and many indicators. These are often binary, corresponding to the response to the item being 'right' or 'wrong.' This has become a specialized field, often referred to as Item Response Theory (IRT) with a literature and notation of its own (see *Factor Analysis and Latent Structure*: *IRT and Rasch Models*). The model may also be used in other fields, such as sociology, where it may be more appropriate to introduce several latent variables.

A latent trait model with binary xs is similar to a latent class model. The prior distribution is now continuous and will usually be taken to be standard normal. The response variables will still be Bernoulli random variables but now they depend on the continuous latent variable(s). Since the Bernoulli distribution is a member of the exponential family the appropriate form for $\pi_i(z)$ turns out to be

$$\operatorname{logit}\pi_i(\mathbf{z}) = \alpha_{i0} + \alpha_{i1}z_1 + \alpha_{i2}z_2 + \cdots + \alpha_{iq}z_q \quad (11)$$

Other versions of the model in which the logit on the left hand side of (11) is replaced by $\mathbf{\Phi}^{-1}(\cdot)$, the inverse standard normal distribution function, are widely used and give similar results to (11) but they lack the sufficiency properties. If $j = 1$ and the α_{ij}s are the same for all i, the model is a random effects version of the Rasch model (see, e.g., Bartholomew 1996). In the latter the trait values are taken as parameters rather than random variables so, in the strict sense of classical inference, the Rasch model is not a latent variable model.

5.5 Other Models

The latent profile models of Table 1 have been omitted because they have found relatively little practical application. It has also been shown by Molenaar and von Eye (1994) that they are virtually indistinguishable from factor models because they have an equivalent covariance structure. Neither have we considered hybrid models in which manifest and latent variables may be mixed type—discrete or continuous. The advantage of our general approach is that it lends itself to the treatment of all such models. This approach has been exemplified by Moustaki (1996) and is capable of further development.

6. Relationship between Latent Variables

In much sociological research the interest is not so much in the latent variables as individual entities, but in the relationships between them. The interrelationship of manifest variables is a major subject of investigation in social science using techniques such as path analysis, regression, log-linear analysis, and graphical models. It is natural to extend these models to include latent variables. This has given rise to linear structural relations modeling which can be implemented using widely available software such as LISREL, AMOS, and EQS (see *Factor Analysis and Latent Structure, Confirmatory*; *Latent Structure and Casual Variables*).

In essence, such models have two parts: (a) a measurement model which supposes that each latent variable is linked to its own set of indicators through a factor model and (b) a structural model which specifies the (linear) relationships among the latent variables. Such a model is fitted by matching the observed and theoretical covariance matrices. A much more general framework which allows a wider range of models is provided by Bartholomew and Knott (1999). Although these methods are very widely used, serious questions have been raised about the identifiability of the models (Anderson and Gerbing 1988, Croon and Bolck 1997, Bartholomew and Knott 1999). These authors have suggested that it may be better to separate the measurement and structural parts of the analysis. This can be done by constructing 'estimates' of the latent variables and then exploring their interrelationships by the traditional methods used for manifest variables.

7. Future Developments

The advent of massive computer power has changed the practice of multivariate analysis radically, and of latent variable analysis in particular. The limiting factor is no longer computing power but of getting data of sufficient quality and quantity to fit the very complicated models which the theory provides and computers can handle. The precision with which models with many parameters can be estimated is

often very low unless the sample size runs into thousands. This makes it all the more important to estimate the sampling variability of the estimates of the parameters on which the interpretation depends. Traditionally this has been done by finding asymptotic standard errors but these can be very imprecise. However, it is now possible to supplement these results by resampling methods such as the bootstrap.

There are other fields in which latent variable models are used which currently exist in isolation. An obvious generalization is to latent time series. Some work has been done for the case where the latent process is a Markov chain. In this area the term 'hidden' is used instead of 'latent' which helps to conceal the family connections (see *Neural Networks and Related Statistical Latent Variable Models*) (for an introduction see MacDonald and Zucchini 1997). An application in a more traditional time series context will be found in Harvey and Chung (2000). Also, there is work by economists on unobserved heterogeneity as it is called which, essentially, involves the introduction of latent variables into econometric models.

Bibliography

Anderson T W 1959 Some scaling models and estimation procedures in the latent class model. In: Grenander U (ed.) *Probability and Statistics*. Wiley, New York

Anderson J C, Gerbing D W 1988 Structural equation modeling in practice: a review and recommended two-step approach. *Psychological Bulletin* **103**: 411–23

Bartholomew D J 1984 The foundations of factor analysis. *Biometrika* **71**: 221–32

Bartholomew D J 1996 *The Statistical Approach to Social Measurement*. Academic Press, San Diego, CA

Bartholomew D J, Knott M 1999 *Latent Variable Models and Factor Analysis*, 2nd edn. Arnold, London

Croon M, Bolck A 1997 On the use of factor scores in structural equations models. *Technical Report 97.10.102/7*. Work and Organization Research Centre, Tilburg University

Harvey A, Chung C-H 2000 Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society A* **163**: 303–39

Heinen T 1996 *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Sage, Thousand Oaks, CA

Lazarsfeld P F, Henry N W 1968 *Latent Structure Analysis*. Houghton Mifflin, New York

MacDonald I L, Zucchini W 1997 *Hidden Markov and other Models for Discrete-valued Time Series*. Chapman and Hall, London

Molenaar P C W, von Eye A 1994 On the arbitrary nature of latent variables. In: von Eye A, Clogg C C (eds.) *Latent Variables Analysis: Applications for Developmental Research*. Sage, Thousand Oaks, CA

Moustaki I 1996 A latent trait and latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology* **49**: 313–34

Spearman C 1904 General intelligence, objectively determined and measured. *American Journal of Psychology* **15**: 201–93

Thurstone L L 1947 *Multiple Factor Analysis*. University of Chicago Press, Chicago

D. J. Bartholomew

False Memories, Psychology of

A false memory is a mental experience that is mistakenly taken to be a veridical representation of an event from one's personal past. Memories can be false in relatively minor ways (e.g., believing one last saw the keys in the kitchen when they were in the living room) and in major ways that have profound implications for oneself and others (e.g., mistakenly believing one is the originator of an idea or that one was sexually abused as a child). False memories arise from the same processes as do true memories and hence their study reveals basic mechanisms of memory. This article describes empirical research investigating false memories and a theoretical approach—the source monitoring framework—for integrating the findings and guiding further investigation.

1. Selected Early Research

Psychologists have long been interested in memory distortions. A classic example from the 1930s is Bartlett's report of studies in which he told people a folk tale from an unfamiliar culture and asked them to recall it from memory. He noted that the memory errors people made brought the story more in line with their own culturally determined expectations. Another is Carmichael, Hogan, and Walter's demonstration that how forms are labeled influences how they are remembered. For example, people's reproductions from memory of two circles connected by a straight line were more likely to include a curved line if the form had been labeled as 'eyeglasses' than if it had been labeled as 'dumbbell.' Still another is work in the late 1950s and early 1960s by Deese (1959) and by Underwood (1965) demonstrating that people are very likely to falsely remember an item (e.g., needle) if they had earlier been presented with related items (e.g., thread, haystack, sharp).

In the 1970s, studies from a number of laboratories highlighted the fact that such memory distortions do not only occur for unfamiliar materials such as folk tales, or lists of words or pictures—they are a byproduct of our everyday understanding of, and memory for, information and events. For example, Bransford and Johnson's (1973) work on comprehension and memory demonstrated that recalling information often depends on engaging constructive processes by which information is related to prior knowledge or schemas at encoding. For example, memory for a sentence such as 'the haystack was

important because the cloth ripped' was much better when the idea of a parachute had been recently activated than when it had not. However, this work also showed that people are more likely to falsely claim that stories included information that was only inferred based on prior schemas or knowledge. People who heard a story that included the information that 'the spy threw the secret document into the fireplace just in time because 30 seconds longer would have been too late' were later likely to claim that 'the spy burned the document.' Thus distorted 'remembering' of information that might be, but is not necessarily, true (e.g., the spy could have been hiding the document in a fireplace that was not lit) is the potential downside of the intelligent, active processing of information.

Such findings led Johnson and Raye (1981) to suggest that memory distortions like those described above reflect errors arising from imperfect reality monitoring processes that are an integral part of remembering: that is, people sometimes confuse the information they generate during the initial encoding, the retention interval, or subsequent remembering of an event for information that came from the event. Reality monitoring is a special case of the more general ongoing function of source monitoring—making attributions about the origin of activated information in mental experience (Johnson et al. 1993). Thus, source monitoring errors include both confusions between internal and external sources and between various external sources (e.g., attributing something that was imagined to actual perception, an intention to an action, something read in a tabloid to a television program, an incident that occurred in place A or time A to place B or time B).

Of course, how serious any of these memory distortions are depends on what is at stake. A clear case where it matters is eyewitness testimony. In the 1970s, E. Loftus launched an influential research program investigating the impact on event memory of information that is suggested between the time of a witnessed event and the time people are required to 'testify.' In one study, people saw a film of an automobile accident and were asked some questions, including 'About how fast were the cars going when they smashed into each other?' Other people were asked the same question except that the words 'smashed into' were replaced with the word 'hit.' On a memory test a week later, the people in the first group were more likely to mistakenly say 'yes' to a question about whether they had seen broken glass in the film. To take another example from the Loftus laboratory, after seeing a film, participants were asked 'How fast was the white sports car going when it passed the barn while traveling along the country road?' Other participants were asked the same question without mention of the barn. In fact, there had been no barn in the film, but those people who had the barn mentioned in the question were later more likely to say that they had seen a barn in the film. Both of these examples likely involve source monitoring errors—the former a confusion between a self-generated inference based on the supplied information that the cars 'smashed into' each other and the latter a confusion between what was read and what was seen.

In sum, generation, elaboration, and integration of information across individual experiences from different sources reflects associative, imaginative, and reasoning processes that are necessary for all higher-order, complex thought. But, this very capacity for creativity makes us vulnerable to having false memories.

2. *Source Monitoring*

Perhaps the most comprehensive theoretical account of false memories to date is provided by the source monitoring framework (SMF) proposed by the present author and her colleagues. According to the SMF, it is not the case that memories are found (or not found), but rather that mental experiences are attributed to memory (or not) by ongoing judgment processes. The SMF further highlights several key aspects of these memory attributions:

(a) Memory attributions are based on various qualitative characteristics of the mental experience. For example, perceptual, spatial, temporal, or emotional details typically are taken as evidence that a mental experience reflects a true memory.

(b) Memory attributions are influenced by the embeddedness of the mental experience. Embeddedness depends on such factors as the availability of supporting memories, consistency with knowledge and beliefs (e.g., plausibility), and coherence of the information, and agreement with the reports of others about the event.

(c) Memory attributions are made according to flexible criteria (which qualities are considered and how they are weighted, how much evidence of any given type is needed). Hence, what may be taken for a memory under one set of circumstances might not be under another.

(d) Goals, beliefs, and motivational and social factors influence what characteristics are looked for, how much embedding occurs, and which criteria are applied.

According to the SMF, false memories occur because the mental experiences arising from events of different sorts (e.g., imagination and perception) overlap in characteristics (they are imperfectly differentiated) and because the processes that make judgments about these mental experiences are also imperfect (i.e., they not only operate on imperfectly differentiated data, they are not always fully engaged, they sometimes overweight nondiagnostic evidence or employ inappropriate criteria, they are subject to social influences, etc.). As do most theoretical accounts of false memories, the SMF assumes that the as-

sociative, constructive, elaborative, and integrative processes that are central to human perception and thought contribute information that may become a candidate for misattribution. However, the SMF has emphasized that these processes alone do not produce false memories. Rather, false memories arise when the products of these processes are taken *to be* memories. That is, it is not the fact that associations and other elaborations occur that produces false memories—for example one can make an inference and later attribute it to an inference, or imagine an event and later correctly attribute it to imagination. Survival would be difficult, if not impossible, if there were no differences between the memory records of things that happened and things we imagined, or between activities we observed or participated in and activities we only read or heard about. Fortunately, the imperfect system people have is usually good enough, and perhaps even has some advantages over a perfect system (e.g., life satisfaction may be higher when we remember things as better than they were, social relations may be better when we agree on a common account of events, or generalization between similar situations may be faster when we are not concerned about where inferences come from).

Some theorists (e.g., Gardiner, Jacoby, and Kelley) have proposed that there are two distinct types of memory experience, 'Knowing and Remembering,' or 'Familiarity and Recollection,' and have shown that errors often result from misattribution of familiarity which can sometimes be corrected with recollection. Within the SMF, these would represent ends (or a coarse categorization) of a continuum of mental experiences. Also, in the SMF, both true and false memories can be associated with a sense of familiarity and with a sense of recollection.

3. *Investigating True and False Memories*

Evidence for the propositions in Sect. 2 comes from laboratories in many countries working in many domains (the SMF can be thought of as a way of unifying this evidence), including work on familiarity-based attributions, cryptomnesia, the phenomenal qualities of memories, suggestibility, rehearsal, schema effects in memory, the impact of emotion, motivational and social effects in memory, interpersonal reality monitoring, and imagination. Several illustrative findings are briefly outlined here.

3.1 *Memory Characteristics*

When participants are asked to make ratings of the qualitative characteristics of their memories, memories for real events tend to have more perceptual and emotional detail than memories for imagined events. Moreover, when they are asked to justify why they believe their own (or someone else's) memory they invoke these characteristics or are influenced by them. Manipulating these characteristics influences the accuracy of source monitoring. For example, if participants have imagined a word in another person's voice they are later likely to claim that person said the word. If they have seen a magnifying glass, they are likely to say they have seen a physically similar item that they only imagined (e.g., a lollipop). Furthermore, features are 'borrowed' or cumulated across modalities. For example, hearing the sound of a dog barking increases subsequent claims that an imagined dog was seen. In some experiments, people are as likely to claim to 'remember' their false as their true memories and to be as confident (or more so) about them. However, in several experiments using more detailed rating categories (e.g., asking for specific ratings of visual and auditory detail, emotional qualities) false memories were on average somewhat less vivid or detailed than true memories.

3.2 *Evidence and/or Criteria used in Source Monitoring*

Attribution processes are subject to task and social demands. If people are induced to, or have the opportunity to more carefully evaluate their memories, the probability of false memories decreases. Ways to decrease false memories include avoiding leading questions, asking people to explicitly consider the possible sources of their memories rather than simply asking about one source, having them rate the qualitative characteristics of their memories, and public as opposed to private remembering. Presumably, these manipulations decrease the degree that people will rely on a single type of information (e.g., familiarity or perceptual detail) and increase consideration of, and retrieval of, additional potentially useful information.

3.3 *Rehearsal and Imagination*

Thinking about events can increase the vividness of imagined events, the probability people will claim imagined events occurred, and their confidence that an event occurred. Imagining events can also make similar perceived events seem to have occurred more frequently. Furthermore, inducing subjects to generate images of suggested information increases later source misattributions.

3.4 *Attention*

Emotional self-focus or distraction during encoding or remembering can result in decreases in source accuracy without necessarily decreasing memory that an event

occurred. Such findings emphasize that memory 'strength' (in the sense of probability of recognition or recall) is not a perfect predictor of source memory. People can have strong memories for information for which they have only vague source cues, or strong-but-wrong source cues.

4. *Are Some Memories Immune to Distortion?*

In 1977 Brown and Kulick proposed that there is a class of 'flashbulb memories'—accurate, long lasting, and detailed—for highly significant, emotional events (e.g., where one was when one heard that Kennedy was assassinated or that the Challenger space shuttle exploded). However, subsequent research has shown that, like ordinary memories, these memories are subject to distortion.

The question of whether some memories might be immune to distortion resurfaced in the 1980s and 1990s, prompted in part, by high profile legal cases that involved either children's testimony in sexual abuse cases (e.g., New Jersey vs. Michaels), or adults' testimony about recovering previously repressed memories that they were sexually abused (e.g., Ramona vs. Ramona; Martinelli vs. Diocese of Bridgeport), or witnessed crimes (e.g., People vs. Franklin). Such cases greatly broadened the interest among the research community in the processes of memory distortion and the number of published papers related to memory distortion increased dramatically.

A central issue in discussions about these cases has been to what extent the findings and principles arising from laboratory research apply to real life, especially to traumatic memories. Some clinicians and authors of popular self help books encourage individuals experiencing psychological problems to freely imagine what might have happened to them as children and to think about why such events might have been likely. Although intended to help clients, the potential consequences of such practices are clearly problematic from the perspective of the SMF. Such practices encourage the client to generate details that may not be accurate, embed false information in a web of supporting knowledge and beliefs, and to adopt weak criteria for what constitutes evidence for a memory. At the same time, the conditions of laboratory studies cannot mirror those of real-life traumatic situations. Thus some have suggested that experimental results and some theoretical ideas do not always apply. They claim that certain classes of traumatic memory, including childhood sexual abuse, are unlikely to be false because the associated emotion is evidence of their authenticity.

Although there are obvious ethical limits to the types of laboratory situations that researchers can set up, researchers have demonstrated that false memories can be created for entire, complex, salient, and emotionally significant events—inducing both children and adults to believe that they remember autobiographical events that never happened. The general procedure in some of these studies is to obtain reports of actual events from a relative of the participant and then query the participants about these events. Embedded in the list of real events is a false event that did not happen to the participant. After reading the accounts and, especially after being encouraged to think about the events, some participants claim to remember the false event. For example, Loftus and Pickrell induced adult participants to remember being lost in a shopping mall as a child. Hyman and colleagues induced participants to remember being taken as a child to the hospital for an ear infection. These false memories can contain quite specific details not provided in the relative's account. Similarly, Ceci and his colleagues have found that some children will provide compellingly vivid accounts of complex events (e.g., having their finger caught in a mousetrap), particularly after repeated questioning about the event. Spanos and colleagues demonstrated that adult participants could also be led to report memories from infancy. For example, participants were told that, because they have good visual skills, they were likely to have been born in a hospital that hung mobiles over their cribs and then they were 'age regressed' to recover the memories. Groups that were hypnotized and groups that were not both reported a high incidence of 'infant memories.' Spanos and colleagues suggest that memories of past-life identities, UFO abductions, and satanic ritual abuse are similarly likely to be created in a social context where authoritative or trusted sources make such events seem plausible, create the expectation that the individual has had such experiences, and use techniques such as generating imagery or repeated questioning to facilitate recovery of such hidden memories. Ofshe has also argued that induction of hypnotic or dissociative states, in combination with specific suggestions from authority figures, can result in false memories and provided a chilling account of a criminal investigation in which leading questioning techniques induced a man to confess to the sexual abuse of his children.

Of course, it is quite difficult to isolate particular factors contributing to false autobiographical memories in these relatively complex, emotionally salient situations. Nevertheless, overall the results of these studies appear to reflect the processes proposed and they increase confidence in the generality of laboratory findings. Repeated questioning or thinking about an event increases the details that are remembered or confidence in the memory. Encouraging participants to embed a 'memory' in other supporting personally relevant details increases false memories. Individuals with high imagery ability seem to be more susceptible to induced false memories, presumably because they embellish more or create representations that are more like perceptions. In some studies, individuals who

score high on tests of hypnotizability or dissociative experiences are also more susceptible, perhaps because these individuals are high in imagery ability or easy to coax into using lax criteria for assuming something is a memory.

Studies that induce false autobiographical memories highlight the fact that remembering often takes place in a social context that can be a source of false information, can provide occasions for repeating and embellishing 'memories,' and can coax people to ignore the doubt that is an important cue during normal source monitoring. Therapy is itself a social context that potentially can have these characteristics. In 1995, Poole in the US and Lindsay in Canada, in collaboration with UK researchers Memon and Bull, reported results of a survey of therapists in the US and UK: 25 percent of the sampled therapists who work with adult female clients believe 'that recovering memories is an important part of therapy, think they can identify clients with hidden memories during the initial session, and use two or more techniques to help such clients recover suspected memories of [childhood sexual abuse].' They note that such techniques may be used with caution by many therapists, but that other therapists may be unfamiliar with the cognitive literature on processes of memory distortion, and may underestimate their own influence. As professional books and journals and, especially, the popular press, have provided wider exposure to theoretical issues and research findings, therapists are likely becoming better informed for making judgments about which techniques to use. Importantly, however, not only therapists, but also police officers, lawyers, social workers, parents, and many other trusted authorities (e.g., talk show hosts, Internet sites) are in a position to influence how people search and evaluate their memories. There is no reason to believe that memories cannot be accurate (they often are quite accurate in both laboratory and more naturalistic studies), and appropriate cueing and some conditions of social remembering can facilitate memory as well. Nevertheless, the potential for memories to be false is clear and the mechanisms of distortion are generally understood. Recent work has been directed at developing interview techniques for reducing false memories in both children and adults.

5. *The Prefrontal Cortex and Source Monitoring*

Based on the evidence that memory involves constructive and reconstructive processes (including the self-generation of retrieval cues, setting criteria, and evaluation of activated information), one would expect the region most associated with reflective or executive processes—the prefrontal cortex (PFC)—to be important for episodic memory, especially distinguishing between true and false memories. And, indeed, damage to PFC often produces memory deficits, including deficits in self-generated retrieval necessary for autobiographical recall, increases in false recognition of related distracters, and, in some cases, clinically significant confabulation. Source memory deficits (and resulting false memories) are more likely in children (whose frontal lobes are slow to develop) and in older adults (who are likely to show increased neuropathology in PFC with age). PFC dysfunction may play a role in schizophrenia, which sometimes includes severe source monitoring deficits in the form of delusions. Furthermore, in cognitive tasks given to neurologically intact young adult participants, increasing the cognitive load (having the participant simultaneously engage in a secondary task) disrupts performance more on tasks requiring source identification than on tasks simply requiring old/new recognition. Presumably, the extra load taxes the recruitment, organization, and integration of cognitive component processes necessary for source identification that are subserved by PFC.

A few studies have directly examined the brain activity associated with true and false memories. These studies show that there is overlap in the brain regions activated for correct recognition of old items and recognition of semantically related lures, as would be expected based on the behavioral evidence of their confusion. However, evidence also suggests that the similarity of the brain activity associated with true and false memories depends on how people are tested (e.g., with targets and lures randomly intermixed or in separate lists). This is consistent with the idea that what people are looking for as they are remembering will influence the qualities of their memories.

6. *Conclusions*

Autobiographical memories are narratives that are influenced by expectations before events ever happen and by rumination after the fact. These narratives are also influenced by other experiences (photographs, other people's accounts, and even unrelated events) and our goals and motives at the time of remembering. False memories arise from the same encoding, rehearsal and source monitoring (memory attribution) processes that produce true memories; thus one can never be absolutely sure of the truth of any particular memory. Of course, in many contexts, minor (and even major) distortions are of little practical consequence. There may even be some advantages to certain kinds of false memories (e.g., remembering a vacation as having been more pleasant than it was). However, there are other contexts in which it does matter whether memory is accurate—when distorted memories result in consequences that are clearly undesirable from the personal (e.g., mistakenly believing one was a victim), social (e.g., giving erroneous testimony about someone's actions), or professional (e.g., appropriating someone else's ideas) perspectives. Furthermore, certain conditions of psychopathology or

brain damage result in clinically significant increases in false memories and beliefs—confabulations and delusions. Hence understanding the mechanisms underlying true and false memories can potentially help reduce their occurrence under everyday circumstances (e.g., in therapy, in police interrogations, in scholarship) and help specify more completely and assess cognitive dysfunction in clinical populations.

See also: Comprehension, Cognitive Psychology of; Confessions: Psychological and Forensic Aspects; Constructivism in Cognitive Psychology; Elaboration in Memory; Emotion in Cognition; Eyewitness Memory: Psychological Aspects; Inferences in Discourse, Psychology of; Interference and Inhibition, Psychology of; Memory and Aging, Cognitive Psychology of; Memory: Autobiographical; Memory Development in Children; Memory for Meaning and Surface Memory; Memory Retrieval; Mood-dependent Memory; Reconstructive Memory, Psychology of; Schemas, Frames, and Scripts in Cognitive Psychology

Bibliography

Bartlett F C 1932/1995 *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, New York

Bottoms B, Goodman G (eds.) 1996 *International Perspectives on Child Witnesses: Psychological Research and Law*. Sage, Thousand Oaks, CA

Bransford J D, Johnson M K 1973 Considerations of some problems of comprehension. In Chase W (ed.) *Visual Information Processing*. Academic Press, New York, pp. 383–438

Brown R, Kulik J 1977 Flashbulb memories. *Cognition* **5**: 73–99

Bruck M, Ceci S J 1999 The suggestibility of children's memory. *Annual Review of Psychology* **50**: 419–39

Carmichael L, Hogan H P, Walter A A 1932 An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology* **15**: 73–86

Ceci S J, Crotteau Huffman M L, Smith E, Loftus E F 1994 Repeatedly thinking about non-events. *Consciousness and Cognition* **3**: 388–407

Christianson S-A (ed.) 1992 *The Handbook of Emotion and Memory: Research and Theory*. Erlbaum, Hillsdale, NJ

Conway M A (ed.) 1997 *Recovered Memories and False Memories*. Oxford University Press, Oxford, UK

Deese J 1959 On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology* **58**: 17–22

Hyman I E, Husband T H, Billings F J 1995 False memories of childhood experiences. *Applied Cognitive Psychology* **9**: 181–97

Johnson M K, Hashtroudi S, Lindsay D S 1993 Source monitoring. *Psychological Bulletin* **114**: 3–28

Johnson M K, Raye C L 1981 Reality monitoring. *Psychological Review* **88**: 67–85

Johnson M K, Raye C L 1998 False memories and confabulation. *Trends in Cognitive Sciences* **2**: 137–45

Lindsay D S, Read J D 1994 Psychotherapy and memories of childhood sexual abuse: A cognitive perspective. *Applied Cognitive Psychology* **8**: 281–338

Loftus E F 1997 Creating false memories. *Scientific American* **277**: 70–5

Loftus E F, Pickrell J E 1995 The formation of false memories. *Psychiatric Annals* **25**: 720–5

Ofshe R J 1992 Inadvertent hypnosis during interrogation: False confession due to dissociative state: Mis-identified multiple personality and the satanic cult hypothesis. *International Journal of Clinical & Experimental Hypnosis* **40**: 125–56

Pezdek K, Banks W P (eds.) 1996 *The Recovered Memory/False Memory Debate*. Academic Press, San Diego, CA

Poole D A, Lindsay D S, Memon A, Bull R 1995 Psychotherapy and the recovery of memories of childhood sexual abuse: US and British practitioners' opinions, practices, and experiences. *Journal of Consulting and Clinical Psychology* **63**: 426–37

Read J D, Lindsay, D S (eds.) 1997 *Recollection of Trauma: Scientific Evidence and Clinical Practice*. Plenum, New York

Schacter D L (ed.) 1995 *Memory Distortion: How Minds, Brains, and Societies Reconstruct the Past*. Harvard University Press, Cambridge, MA

Spanos N P, Burgess C A, Burgess M F, Samuels C, Blois W O 1998 Creating false memories of infancy with hypnotic and non-hypnotic procedures. *Applied Cognitive Psychology* **13**: 201–18

Tulving E, Craik, F I M (eds.) 2000 *The Oxford Handbook of Memory*. Oxford University Press, New York, Chap. 10, 12, 14, 24

Underwood B J 1965 False recognition produced by implicit verbal responses. *Journal of Experimental Psychology* **70**: 122–9

M. K. Johnson

Familial Studies: Genetic Inferences

The twentieth century may be described as the century of genetics. It began with the rediscovery of Mendel's work by Johannsen (1903) and ended with the completion of the Human Genome Project (HGP). *Nature* (2001, **409**) published the *Initial sequencing and analysis of the human genome* by The Genome International Sequencing Consortium and *Science* (2001, **291**) published similar sequencing by Celera. The twenty-first century, particularly its first half, will see a spate of work related to gene manipulation and genetic disorders. We can confidently look forward to substantial developments in our understanding of our species.

We fear that even in the post-HGP era, some of the twentieth-century errors relating to inferences from genetic analyses of familial data will be repeated and some more will be added. Indeed, McGuffin et al. (2001) say 'The most solid genetic findings about individual differences in human behavior come from quantitative genetic research such as twin and adoption studies that consistently converge on the conclusion that genetic variation makes a substantial contribution to the phenotypic variation for all behavioral domains.' The conceptual and mathematical errors in behavior genetic models, therefore, need to

be known widely. Behavioral traits are also described as 'complex' traits (McGuffin et al. 2001, Carroll 2001). Feldman and Lewontin (1975) define a complex trait as one whose frequency of occurrence cannot be analysed in terms of simple genetic hypothesis. We prefer Mather's (1949) description of such a trait as 'polygenic' as it indicates clearly the basic genetic hypothesis. This article is concerned with the inheritance of a polygenic trait. It will: (a) explain the nature of the statistical model used for the statistical analysis of a polygenic trait, (b) show that heritability is not defined in the presence of assortative mating, (c) explain Galton's ideas about nature–nurture contributions and their extension in the twentieth century, (d) clarify some confusions about correlation and regression, (e) discuss heritability, and (f) show that behavior geneticists use incorrect formulas in model fitting. As an example, we choose Devlin et al. (1997) because it was published in *Nature*, and show that the formulas they used were incorrect. We state why Fisher's (1918) formulas are wrong and point out the deficiencies of Jinks and Fulker (1970). We explain why phenotypic, that is, observed variance of a trait cannot be partitioned into genetic and environmental components. Finally, we state how the research in genetic disorders should progress.

1. *Mathematical Models in Biology and Medicine*

There is a long tradition of using mathematical models in medicine and related areas. For example, Cullen (1983) divides his book *Mathematics for the Biosciences* into four major areas: (a) physiology and medicine, (b) ecology and population, (c) fisheries and oceanography, and (d) genetics. Roach (1984) in *Mathematics in Medicine and Biomechanics* has chapters on blood modeling, oxygen transport, nerve modeling, etc. Mathematical models have also been used in epidemiology, neurophysiology, cardiovascular simulation, tubercular skeleton, body fluids, kidneys, etc. All these models are type I models and have proved very useful in medical research. Recently, type II models to estimate the heritability of a trait have been introduced in familial aggregation studies. We give a brief description.

A Type II model is also known as the Components of Variance model. Its basic structure is: observed value = constant + linear function of *random* variables + residual. Residuals are assumed to have 0 mean and σ^2 variance (the assumption of normality of residuals is required for tests of significance). Here, we sample the variables at random from all the variables that affect a disorder. The problem is that we may not know all variables that affect it. Therefore, random selection of variables is very important. Our interest lies not in the mean effect of a factor but in the contribution it makes to the total variance. Thus, the aim is to divide the total (phenotypic) variance of a disorder into the sum of variances due to factors. Numerical analyses for type I and type II models are similar and this may cause confusion in the minds of some nonspecialist researchers. We emphasize that no treatment of a disorder can be deduced from a type II model.

2. *Heritability*

In genetics, narrow heritability (h^2) and broad heritability (H^2) are strictly defined. h^2 = additive variance/phenotypic variance and H^2 = genetic variance/phenotypic variance. Additive variance is the variance of additive values and genetic variance is the variance of genetic values. The genetic value of an allele at a locus is defined as the regression of the phenotypic values on genotype and the additive value is defined as the regression of genetic values on genotype. Thus, both additive value and genetic value at a locus *are statistical constructs*. We can calculate phenotypic variance directly from phenotypic values. We cannot, however, calculate additive or genetic variance directly, as neither additive nor genetic values are known (Capron et al. 1999). Geneticists have devised various methods for estimating h^2 and H^2. For example, if we know *beforehand* that a trait is under genetic control, then assuming random environment and random mating with respect to a trait, h^2 = corr. (mid-parent, child).

Jensen (1969) used the concept of heritability for analysis of IQ data. Since then an enormous literature on heritability has built up. Layzer (1999) questions the assumption that phenotypic measurement = genetic component + an environmental component, as 'every phenotypic trait is a *function* of genetic and nongenetic variables.' Feldman and Lewontin (1975) said: 'As we show below, the partition of the causes of variation is really illusory and the analysis of variation can not really separate the variation that is the result of genetic segregation. *The genetic variation depends on the distribution of environments and the environmental variance depends on the distribution of genotypes*' (italics added).

We agree with these criticisms but the nature of our criticism is slightly different. We believe that heritability is not defined when the population mates assortatively with respect to the trait. Fisher (1918) showed that assortative mating will introduce association between similar alleles and, thereby, destroy independent segregation of alleles. This association will also destroy the linearity of regression on which the definition of additive values is based. We do not know how to define additive value at a locus when a population is at equilibrium under assortative mating. Moreover, it can be shown that assortative mating will also create association between additive and dominance deviations at different loci. Thus, in the presence of assortative mating narrow heritability is not

defined. Fisher (1951) did not approve of the concept of heritability.

Nonrandom environment also causes problems. The trait is then associated with environmental variables. It can be shown, if assortative mating continues, it will create association between genetic alleles and environmental factors. The additive and genetic values cannot now be defined. Fisher may have known this and assumed random environment (see Kempthorne 1969). Moreover, few of us are capable of identifying all aspects of environment. To equate environment in genetic studies with the number of cars, or the number of rooms, or the number of books in a family home, etc. is simplistic (see Sect. 10). If a researcher wishes to study the effect of books, cars, or rooms on a trait, they should use a Type I model. Heritability analysis is a genetic cul-de-sac. It suggests no remedy for a disorder. With high H^2 one can only argue incorrectly, as Jensen (1969) did, that intervention programmes are useless.

Concerning comparisons between two groups using heritability, we agree with Feldman and Lewontin (1975) that 'the concept of heritability is of no value for the study of differences in measures of human behavioral characters between groups' (see also Thoday 1973).

3. *Type II Model Fitting on Twin Data*

This section is in two parts: (a) Galton's ideas, and (b) later attempts to divide the phenotypic variance into nature and nurture components.

3.1 *Galton's Nineteenth-century Ideas*

Galton was not aware of Mendel's work. His main interest lay in Eugenics, that is, the improvement of the white 'racial' stock. He invented the concept of correlation. He was well-off and endowed the Galton Laboratory for National Eugenics to propagate eugenics. Karl Pearson was its first Director. He devised the commonly used formula for product moment correlation following Galton's concept of corelation. Galton thought that the difference between concordances of monozygotic (MZ) and dizygotic (DZ) twins on a trait gave an indication of genetic (nature) effects on that trait. Galton's views have had a powerful effect in genetics and social sciences. His *law of familial regression* has been used to 'prove' that a trait is determined genetically.

Galton's method of concordance is explained in textbooks on genetics (e.g., Snustad et al. 1997 and Hartl and Clark 1997, etc.). In exercise 15-6 Levitan (1988, p. 361) asks: 'What can be said about the relative roles of heredity and environment in the following studies?' and lists a number of studies with respective MZ and DZ concordances. The first on the list is tuberculosis (TB). The ratio of MZ and DZ concordant twins are 202/381 and 187/843. Perhaps he expects the students to compare the two concordances and conclude that heredity plays a major role in causing TB. This, indeed, was the 'scientific' view of the researchers in the field of TB based on analysis of familial data until the late 1950s (Capron et al. 1999). We now know that this is not the case.

Vogel and Motulsky (1996) give a similar table (6.22, p. 237) with larger sample sizes. They also calculate the statistics 'MZ higher than DZ.' Its highest value, 2.9, is for leprosy and the second highest, 2.56, is for TB. They say, 'In Western Europe, for example, leprosy disappeared during the seventeenth and eighteenth centuries without any therapy, only due to improvement in living conditions. *There was probably little or no influence of genetic changes*' (p. 237, italics added). They then claim that 'Analysis of discordance can shed some light on genetic versus environmental factors in disease.' They are obviously wrong. The only reasonable conclusion from these data is that the concordance method has no value in genetics. Note that heritability, as defined above, cannot be estimated from concordances.

3.2 *Nature–Nurture Methods in the Twentieth Century*

Researchers in twin studies soon realised the deficiencies of Galton's method and devised new methods to estimate nature–nurture contributions. Holzinger (1929) gave an index $H^2 = (r_{mz} - r_{dz})/(1 - r_{dz})$ in terms of intraclass correlations between MZ and DZ twins. Nichols (1965) proposed an 'improved' index $\mathrm{I} = 2(r_{mz} - r_{dz})/r_{mz}$. In these formulas r_{mz} and r_{dz} are the phenotypic correlations between MZ and DZ twins, respectively. Obviously, neither of these formulas estimates h^2 or H^2. Jensen (1967) realised this and published his formula for estimating broad heritability, H^2 from twin data in the PNAS. He began by stating a complex 'generalized formula' but did not say where it came from. That formula has *no* theoretical basis. After 13 steps he stated the formula $H^2 = (r_{mz} - r_{dz})/(1 - \rho_{dz})$ where ρ_{dz} is the genetic correlation between DZ twins. Comparison of Jensen's and Holzinger's formulas shows that Jensen had replaced the phenotypic DZ correlation r_{dz} with the genetic DZ correlation ρ_{dz}. We do not know the value of the genetic correlation ρ_{dz}. Jensen (1967) got himself into a circular situation.

How did he remove the circularity? He claimed that we could find ρ_{dz} ($= \rho_{oo}$, the genetic correlation between offsprings) by using the formula $\rho_{oo} = (1 + \rho_{pp})/(2 + \rho_{pp})$ where ρ_{pp} is the genetic correlation between parents on the trait, and gave Li (1955, Chap. 13) as reference for the formula. It does not appear there and has no genetic basis. Clearly, circularity remains. One now needs an estimate of h^2 to obtain ρ_{pp}

from r_{pp}. When Jensen's formula was used to calculate heritability in England's largest study of twins, Vetta (1977) was able to inform the researchers that this formula is not correct. Their response was 'It follows that the formula ought not to be perpetuated' (Adams et al. 1977).

The quest for a formula for estimating heritability from twin data continues. Vogel and Motlusky (1996, p. 769) said 'Twin data can be utilised as an alternative way to get heritability estimates.' As to how it can be done, they said 'An empirical way … is to calculate alternative estimates from the same (twin) data to determine how well they coincide.' They made a number of unrealistic assumptions and developed formulas for finding h_1^2 and h_2^2. They then claimed that H^2 lies between h_1^2 and h_2^2. We cannot commend their formulas.

4. *Correlation and Regression Tell Us Nothing about Cause and Effect*

The formulas given in the last section use MZ and DZ correlations. Geneticists borrowed the concepts of correlation and regression from statistics. Confusion regarding their signifcance permeates the scientific literature. For example, Lawrence and Jinks (1973) said: 'This tendency for like to beget like, or in statistical terms, for the stature of parents and offspring to be positively correlated, is a sure sign of genetic determination.' The concept of regression to the mean originated with Galton's (1869) law of filial regression. He noticed that the children of tall parents were, on the whole, not as tall. He called it regression to the mean. He thought that a trait showing regression of progeny mean on mid parental value must be genetically determined. Jensen (1969) discussing some results on IQ says: 'None of these findings is at all surprising from the standpoint of a genetic hypothesis, of which an intrinsic feature is Galton's "law of filial regression."' Eysenck (1975) wrote 'There is no mention (in Kamin's book, 1974), for instance, of the remarkable regression effects which provide such a striking proof of inheritance, as well as making possible estimates of heritability' (see also Eysenck 1976).

This confusion has now *entered* the area of familial aggregation. We give two examples. Tsuang et al. (2000) found that 'In the Huntington's disease probands with psychosis, the onset of psychosis correlated with the onset of the neurological symptoms of the Huntington's disease.' From this correlation they concluded: 'Patients with Huntington's disease and psychotic symptoms may have a familial disposition to develop psychosis.' Perusse et al. (2000) used a familial correlation model and concluded 'that familial/genetic factors are more important …' Behavior geneticists Neale and Carden (1992, p. 99) said: 'In other words, genotypes and environments are not measured directly but their influence is inferred through their effects on the covariances of relatives.'

It is remarkable that despite Fisher's (1924) efforts, confusion persists. Fisher, discussing the phenomenon that children of tall parents are, on average, less tall said: 'This phenomenon has been called "regression to the mean" a use of the term regression quite different from the *technical meaning of the term* explained previously' (italics added). He suggested some reasons for this (see also Vetta 1975). Statisticians know that correlation or regression, in absence of other evidence, tell us nothing about cause and effect. Fisher put it elegantly (p. 196), 'To begin with, these statistics (regressions) do not prove that daughters inherit their height from their fathers; if they prove that they will equally prove that fathers inherited their height from their daughters, which is absurd. What they prove is that the heights of fathers and daughters are influenced by the same causes and it is generally agreed on *quite other grounds*, that the important cause of this similarity is that daughters have a great deal in common with their fathers in their hereditary constitution' (italics added). Thus, familial correlations, covariances or regressions on a trait should not be used to estimate genetic contribution, that is, h^2 or H^2 unless one is absolutely certain, on quite other grounds, that the trait has a genetic component.

Some behavior geneticists use path coefficients (Wright 1921). A path coefficient is a standardized regression coefficient. We do not discuss path analysis. We do not believe that a regression coefficient with or without the help of a diagram can assist us in resolving the problem of cause and effect.

5. *Development of Modern Genetics*

Modern genetics was developed by Fisher, Wright, Haldane, and Dobzhansky, among others, in the early part of the twentieth century. Fisher was a brilliant geneticist, as well as a brilliant statistician. His 1918 paper *Correlation between relatives on the supposition of Mendelian inheritance* is a comprehensive paper on quantitative genetics. It is difficult to read and almost impossible to understand. It is the basis of most of the model-fitting work in behavior genetics. It is, therefore, important that one understands the assumptions he made and the consequences of their failure.

6. *Fisher's (1918) Correlation Formulas are Incorrect*

Vetta (1976) showed that Fisher's (1918) correlation formulas are incorrect. He gave two reasons that may have led to the error. A third reason is that Fisher's basic assumption for his model was that the contribution of terms of third and higher degree of smallness are negligible as compared with the con-

tribution of terms of second degree of smallness (which are associated with variances). This assumption is valid only under random mating and not under assortative mating. The terms of third degree of smallness do make a contribution to covariances but the contribution of the terms of fourth and higher degree of smallness is, indeed, negligible. P-c and sib correlation formulas obtained by taking the contribution of terms of third degree of smallness into account meet the criteria discussed above.

7. *Fisher's Model in Behavior Genetics*

Behavior genetics is a relatively young science (Hirsch 1967). One of the major problems was the existence of genotype-environment interaction for most behavioral traits. There was no method of estimating it and Jinks and Fulker (1970) devised one. Their method was hailed as a breakthrough. It was used by Jensen (1970) and was extended to the multivariate case by Eaves (1972). In 1975 Vetta found that there was an algebraic error in their paper. When this error is corrected their method is useless. This was made known immediately to Professor Jinks and his colleagues. Despite this, Fulker and Eysenck (1979) said 'We can test directly for some form of genotype-environment interaction.' This is not true. In the rest of their paper Jinks and Fulker used Fisher's (1918) *incorrect* formulas to analyze data on some behavioral traits. Their paper is probably the most cited paper in behavior genetics. Eysenck (1979) says 'This book is the first to base itself entirely on these new methods.' Some behavior geneticists regard it as 'seminal' (Martin et al. 1989, p. 5). We disagree.

Most behavioral traits involve assortative mating. Behavior geneticists use incorrect formulas when they fit assortative mating models. Devlin et al. (1997) supplement Bouchard and McGue's (1981) correlations with correlations from twin studies published after 1981. They said: 'IQ heritability is not well characterized.' On the basis of past analyses Herrnstein and Murray's (1994) believed 'that IQ heritability is at least 60 percent, and is probably closer to 80 percent …' (p. 470). They predicted that this will give rise to dysgenic trends and a cognitive caste will emerge. Devlin et al. (1997) aimed to show that a cognitive caste will not emerge. Using incorrect formulas, they obtained estimates of 0.34 and 0.48 for narrow and broad heritability respectively. They argued that, in view of these smaller estimates, a cognitive caste will not emerge. Actually, a cognitive caste will still emerge but over a longer period of generations, if IQ is a genetic trait. McGue (1997, p. 417) said that Devlin et al.'s analysis follows 'a long line' of analyses. Behavior geneticists keep on analyzing essentially the same data and obtain estimates of heritability of IQ to suit the occasion.

Devlin et al. (1997) wrote expected covariance formulas for different kinships. Each formula had a genetic and an environmental component. The genetic part of their covariance formulas for dizygotic twins and sibs is $1/2\ (1+r\sigma^*_A)\ \sigma^2_A$. The genetic part of their parent-child covariance formula is $1/2\ (1+r)\ \sigma^2_A$. In these formulas, σ^2_A is the additive variance of the trait, σ^*_A is its standardized additive variance, and r is the phenotypic coefficient of assortative mating. As stated earlier, in the absence of dominance these genetic parts should be equal. Therefore

$$(1+r\sigma^*_A)\ \sigma^2_A = (1+r)\ \sigma^2_A \qquad (1)$$

Simplifying, $r\sigma^*_A = r$. This gives, $\sigma^*_A = 1$. But σ^*_A is always < 1 as it is the additive part of the phenotypic variance (except for the trivial case when only additive variance is present). These formulas are not correct. Similarly, their formulas expected covariances of dizygotic twins, siblings reared together, and siblings reared apart are not correct (Table 1, p. 469). Their model is, therefore, useless.

8. *Statistics and Behavior Genetics*

A recently borrowed concept from statistics is 'shared environment.' McGuffin et al. (2001) said that the second striking finding of behavior genetics is that the contribution of environment 'tends to be of the nonshared type, that is, environmental factors make people different from, rather than similar to, their relatives.' What is 'nonshared' environment? Neale and Cardon (1992, p. 15) said 'The environment between families is sometimes called the *shared environment*.' Thus, within family environment it is the 'nonshared environment.' Behavior geneticists in their analyses of variance of observed data may have found that, for some traits, within-family variance > between-family variance. We do not believe that this justifies their claim.

A statistician working on a designed two-factor component of variance model will have three types of data; (a) observed values, (b) levels of factor I, and (c) levels of factor II. A behavior geneticist only has observed values from an observational study. He makes up for the lack of data (b) and (c) by calculating kinships covariances. He then equates calculated covariances to expected covariances, under assortative mating. His expected covariance formulas are invariably wrong. He then makes various claims which, in our view, cannot be justified.

9. *High Heritability and Change in the Population Mean of a Trait*

The concept of heritability is surrounded with many misconceptions. One such misconception is: if the heritability of a trait is high then environmental changes have little or no effect on its mean. When with high H^2, the mean value of a trait changes appreciably,

some geneticists are at a loss to explain the change. For example, see Vogel and Motulsky's (1996, p. 239) discussion of the 'recent increase in stature.' There has also been a secular trend in IQ during since the 1970s. A secular trend in presence of high heritability deserves serious consideration. Fisher (1918) showed that a polygenic trait is normally distributed. The mean and variance of a normal distribution are, however, independent, that is, the factors which affect variance do not affect mean. Therefore, H^2, the ratio of two variances, does not affect the mean. This is true in industrial production. We do not believe it is true for a genetic trait. The factors that contribute to H^2, that is, genes, must also affect the mean of the trait. Genes, however, do not determine secular trend. We should, therefore, consider alternate explanations, for example, genes only provide possibilities; the trait has no genetic component; all genes for the trait are 'fixed,' therefore, the variance is due to 'environmental' factors, etc.

10. Can Genetic and Environmental Effects be Separated?

A designed statistical experiment could separate effects of factors on a disorder. Designed experiments are not possible on human beings. Moreover, we shall not be able to sample from all genotypes and all environments. In the absence of designed experiments, the genetic and environmental variances *cannot* be separated. Fisher (1918) did separate the genetic and environmental variances relating to height by assuming random environment. We do not believe that this assumption is viable. It is noteworthy that Fisher never really returned to the subject of this paper. Actually, the concept of 'evolution by adaptation' implies complex relationships between genotypes and environments. The history of human evolution may be divided into two periods: the first period when Homo Sapiens tried furiously to adapt to the then existing environment and the second, when they tried to control it. We are now far into the second period. We do not know the bygone environment to which our bodies adapted. We agree with Fisher (1930) that a species is defined only *in* an environment. Change this environment a little, the species will try to adapt. Change it too much and the species may not be defined. Therefore, research aimed at finding the genetic and environmental components of human behavior has no scientific content.

11. The Future?

McGuffin et al. (2001) envisage a new science of behavior genomics. We suspect that as the unscientific nature of behavior genetic analysis becomes known, researchers will eschew heritability analysis. HGP has made the identification of a genetic disorder easier. If, for example, a large number of individuals suffering from a disorder have mutation at a locus as compared with the normal type, this provides some evidence of the genetic nature of the disorder. Heritability analysis is useless as it relates to a population and not an individual. To find remedies for genetic disorders, type I models are useful. Venter (2001, *The Independent*, February 12) succinctly summarizes our view when he says, HGP indicates 'to me that we are not hard wired. The idea that there is a simple deterministic explanation—that is: we are the sum total of our genes—makes me as a scientist, want to laugh and cry.'

See also: Behavioral Genetics: Psychological Perspectives; Depression; Genetic Factors in Cognition/Intelligence; Genetic Studies of Behavior: Methodology; Genetic Studies of Personality; Genetics of Complex Traits Through the Life Cycle; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Mental Illness, Genetics of; Schizophrenia

Bibliography

Adams R, Ghodsian M, Richardson K 1977 *Nature* **266**: 279

Bouchard T J Jr, McGue M G 1981 Familial studies of intelligence: A review. *Science* **212**: 1055–9

Capron C, Vetta A R, Duyme M, Vetta A 1999 Misconceptions of biometricals IQists. *Current Psychology of Cognition, Cahiers de Psychologie Cognitive* **18**: 115–60

Carroll S B 2001 Communication breakdown? Review of *The Century of the Gene*. *Science* **291**: 1264

Cullen M R 1983 *Mathematics for the Biosciences*. Prindle, Webber and Schmidt, Boston

Devlin B, Daniels M, Roedre K 1997 The heritability of IQ. *Nature* **388**: 468–70

Eaves L J 1972 Computer simulation of sample size and experimental design in human psychogenetics. *Psychological Bulletin* **77**: 144–52

Eysenck H J 1975 Review: The science and politics of IQ. *Bulletin of the British Psychological Society* **28**: 352

Eysenck H J 1976 Genetic determination and regression effects. *Bulletin of the British Psychological Society* **29**: 406

Eysenck H J 1979 *The Structure and Measurement of Intelligence*. Springer-Verlag, Berlin

Feldman M W, Lewontin R C 1975 The heritability hang up. *Science* **190**: 1163–8

Fisher R A 1918 The correlation between relatives on the supposition of Mendelian inheritance. *Translations of the Royal Society Edinburgh* **52**: 399–433

Fisher R A 1924 The biometrical study of heredity. *Eugenics Review* **16**: 189–210

Fisher R A 1930 *The Genetical Theory of Natural Selection*, 2nd edn. Clarendon Press, Oxford, UK (rev. edn. 1958 Dover Publications, Broadway, UK)

Fisher R A 1951 Limits to intensive production in animals. *British Agricultural Bulletin* **4**: 217–8

Fulker D W, Eysenck H J 1979 Nature and nurture: Heredity. In: Eysenck H J (ed.) *The Structure and Measurement of Intelligence*. Springer, Berlin, pp. 102–32
Galton F 1869 *Hereditary Genius: An Inquiry into its Laws and Consequences*. Macmillan, London
Hartl D L, Clark A G 1997 *Principles of Population Genetics*. Sinauer Associates, Canada
Herrnstein R J, Murray C 1994 *The Bell Curve*. Free Press, London
Hirsch J 1967 *Behavior-genetic Analysis*. McGraw-Hill, New York
Holzinger K J 1929 The relative effect of nature and nurture on twin differences. *Journal of Educational Psychology* **20**: 241–8
Jensen A 1967 Estimation of the limits of heritability of traits by comparison of monozygotic and dizygotic twins. *Proceedings of the National Academy of Science of the United States* **58**: 115–74
Jensen A 1969 How much can we boost IQ and scholastic achievement? *Harvard Educational Review* **39**: 1–123
Jensen A 1970 IQs of identical twins reared apart. *Behavior Genetics* **1**: 133–48
Jinks J L, Fulker D W 1970 Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin* **73**: 311–49
Johannsen W 1903 *Über Eblichkeit in Populationen und in Reinen Linien*. Gustav Fischer, Jena (summary and conclusions trans. in Peters J A 1959 *Classic Papers in Genetics*. Prentice-Hall, Englewood Cliffs, NJ)
Kamin L 1974 *The Science and Politics of IQ*. Erlbaum, Hillsdale, NJ
Kempthorne O 1969 *An Introduction to Genetic Statistics*. Wiley, New York
Lawrence M J, Jinks J L 1973 Quantitative genetics. In: Sheppard P M (ed.) *Practical Genetics*. Blackwell Scientific, Oxford, UK
Levitan M 1988 *Textbook of Human Genetics*. Oxford University Press, Oxford, UK
Li C C 1955 *Human Genetics*. McGraw-Hill, New York
Martin N G, Boomsman D I, Neale M C 1989 Foreword. *Behavior Genetics* **1**: 5–7
Mather K 1949 *Biometrical Genetics: The Study of Continuous Variation*. Dover Publications, NewYork
McGue M 1997 The democracy of the genes. *Nature* **388**: 417–8
McGuffin P, Brien R, Plomin R 2001 Toward behavioral genomics. *Science* **291**: 1232
Neale M C, Cardon L R 1992 *Methodology for Genetic Studies of Twins and Families*. Kluwer, Dordrecht, The Netherlands
Nichols R C 1965 The National Merit twin study. In: Vandenberg G S (ed.) *Methods and Goals in Human Behavior Genetics*. Academic Press, New York
Perusse L, Rice T, Province M A, Gagnon J, Leon A S, Skinner J S, Wilmore J H, Rao D C, Bouchard C 2000 Familial aggregation of amount and distribution fat and their responses to exercise training in the HERITAGE family study. *Obesity Research* **8**: 140–50
Roach G F (ed.) 1984 *Mathematics in Medicine and Biomechanics*. Shive Publications, Cheshire
Snustad D P, Simmons M J, Jenkins J B 1997 *Principles of Genetics*. Wiley, New York
Thoday J M 1973 Review of educability and group differences. *Nature* **245**: 218–20
Tsuang D, Almqvist E W, Lipe H, Strgar F, DiGiacomo L, Hoff D, Eugenio C, Hayden M R, Bird T D 2000 Familial aggregation of psychotic symptoms in Huntington's disease. *American Journal of Psychiatry* **157**: 1955–9
Vetta A 1975 A note on regression to the mean. *Social Biology* **22**: 86–8
Vetta A 1976 Correction to Fisher's correlations between relatives and environmental effects. *Nature* **263**: 316–7
Vetta A 1977 Estimation of heritability from IQ data on twins. *Nature* **266**: 279
Vetta A, Capron C 1999 Review of *How the Mind Works* by Steven Pinker. *Cahiers de Psychologie Cognitive* **18**: 105–11
Vogel F, Motulsky A G 1996 *Human Genetics*. Springer, Berlin
Wright S 1921 Systems of mating. I–V. *Genetics* **6**: 111–23

A. Vetta and C. Capron

Families and Households, Behavioral Demography of

The last half of the twentieth century has seen the study of household and family develop into a major subspeciality in demography, rivaling in scope and interest more traditional subfields, such as mortality–morbidity, fertility, migration, and urbanization. The study of households and families introduces a new order of complexity into demography since (a) the focus is now on a set of interrelated individuals rather than on the isolated individual or on a population aggregate described by summary measures, and (b) the balancing equations for these residential and kinship units involve several events other than those in the fundamental demographic equation (population change = births − deaths + in-migration − out-migration). Changes in households and families must be related also to marriage, cohabitation, divorce, separation, home-leaving of young adults, as well as other forms of exit and entrance to the group (e.g., the entry of an elderly parent or some other relative). The household may be viewed as a resultant of many individual behaviors.

There are more fundamental complications relating to household and family identity: it is clear when a particular individual begins (birth) and ceases to exist (death), and there is individual continuity over the lifetime. By contrast, it often is not at all clear when a particular household or family begins or ends. Households, unlike individuals, can split apart into two or more smaller households (fission), or can merge to form larger and more complex units (fusion). This fundamental problem has been not solved so much as finessed by extensive data-collection and analysis of *individuals in terms of their family and household statuses and the events changing their statuses*.

But progress has been considerable, particularly with the advent of large-scale social surveys containing event and status history data on marriage, family, and household (individual event-histories are much more common than histories of households or families).

These data have allowed the description and analysis of individual family life-histories and changing household statuses, based on real cohort data. These newer studies stand in contrast to the synthetic approach of the original 'family life-cycle' studies, based on census data (cross-sectional) and sometimes misleading. Recent work has documented a proliferation of family and household behavioral sequences, overturning the stereotypical view of the 'family life-cycle,' based on averages or medians (a development anticipated by Höhn 1987).

Another important breakthrough is the work of Zeng et al. (1997), who have developed accounting equations for a number of specific family household types, equations that can be applied using routine census and vital registration data. The model produces detailed household projections by size and type of household that are consistent with population projections by age, sex, and marital status.

Review articles and summary volumes on the demographic approach to households include: Burch (1979), Bongaarts (1983), Bongaarts et al. (1987). Keilman et al. (1988) and van Imhoff et al. (1995). Work in behavioral household demography, of course, has not occurred in disciplinary isolation, and should be viewed in the context of parallel or overlapping efforts in family sociology, household economics, anthropology, and demographic history.

1. *The Definition of 'Household'*

The central idea of the household concept is that of *coresidence*: a household is a group of persons who 'live-together' day-to-day, or an individual 'living alone.' In demographic usage, the operational definition of this concept has tended to follow census practice, with households defined in terms of persons occupying a separate housing or dwelling unit (house or apartment), or, less often, the persons who participate in a common domestic economy. The housing unit approach provided a convenient basis for census enumeration (first identify housing units and then enumerate the people usually residing in each), and later, for modern large-scale sample surveys.

1.1 *Households and Families*

Whereas the *household* is defined in terms of *coresidence*, the *family* is defined in terms of *kinship*, which in turn is based on marriage, descent from a common ancestor, or social fiction (adoption or other forms of 'fictive kinship'). Ryder (1987) has aptly termed these, respectively, the residential, the conjugal, and the consanguineal dimensions of intimate relationships among individuals (see also van Imhoff et al. 1995). This distinction is not always made in everyday language, however, and the concepts and reality overlap: a household consisting of persons related by kinship is also a family, but such a household seldom contains all of an individual's close kin, the so-called 'extended family.' Modern census practice often used the term *family* to refer to a group of related persons living together (more properly called a *census family*), a usage encouraged by and reinforcing a Western cultural tendency to view the small, nuclear family group as 'the family'—other kin are 'relatives.' Historical and comparative studies of household and family have forced a clarification and refinement of census concepts to take into account the greater complexity of household and family patterns, as have more recent developments such as cohabitation or the joint custody (and split coresidence) of children following divorce. In general, the single word *family* is ambiguous, and should always be related to context, and qualified as necessary with such terms as *nuclear, extended, census, economic*, etc.

1.2 *Large Households*

The majority of households are relatively small, consisting of groups such as a couple and their children, a small group of students sharing an apartment, or an individual living alone. Such households often are referred to as *private households* in contrast to *institutional households* such as boarding schools, prisons, chronic-care hospitals, mental institutions, convents, or seminaries. There is no inherent limit to the size of private households, but their average size seldom exceeds six, and in urban-industrial societies is closer to three.

In modern census practice, private households can contain boarders, lodgers, or servants, but an arbitrary numerical limit on such members is set, beyond which the household is considered institutional.

1.3 *Limitations and Criticisms*

The modern household concept does not deal easily with social realities such as residential *compounds*, clusterings of separate but closely related (in terms of kinship or economic activity) households, often physically distinct, occupying a separate farm or walled or fenced space. An example from a polygamous society would be the compound containing a man and his several wives and children, the latter occupying separate huts and eating separately. Laslett (1982) has suggested the term *houseful* for European cases where several smaller nuclear families (some related, some agricultural workers or servants) occupied a well-defined farmstead.

In contemporary urban societies, census practice reveals virtually nothing about relationships among separate households even when such relationships are frequent and close. For example, an older couple and two grown children, all occupying separate apartments in the same neighborhood, or even the same building, would be reckoned simply as three households, with no census item even hinting at their close ties. Relationships are counted within but not across households. It is of interest that this census practice was codified around the middle of the twentieth century, at a time when social scientists, especially family sociologists, were emphasizing (in retrospect exaggerating) the isolation of the nuclear family in modern urban societies. These limitations of the census have been rectified in part by large social surveys including information on numbers and types of kin, on kin location, and on kin interaction.

A related problem has been noted with use of the phrase *living alone* to refer to one-person households. Accurate enough in terms of census housing-unit concepts, its connotations can be misleading with respect to the degree of social integration or isolation of the occupants.

These and other concerns have led some to more or less reject the household concept as ethnocentric or culture-bound, and too limited, artificial, and abstract. But an increasingly popular view of scientific methodology (see, e.g., Giere 1999) reminds us that all scientific concepts (models, theories) are abstract—therefore, limited and artificial—and are better judged in terms of their analytic usefulness for the study of particular problems. The household clearly is the relevant unit for much consumer behavior (from buying a house or renting an apartment to routine purchases of food or clothing), demographic decision making (e.g., migration), and much of our day-to-day intimate social interaction and support, including the intimacy of couples and the continuing care of infants and young children. Thus, housing analysts and planners still are major users of household projections, as opposed to ordinary population projections. And the household remains a major focus for family sociology and for microeconomics (notably the Chicago school of the 'new household economics'), even as they have broadened their view to encompass such economic realities as intergenerational transfers and such psychological realities as 'altruism.'

Even so, the household concept needs refinement and extension to take more account of the fluidity of household membership and its 'fuzziness,' as more persons can realistically claim membership in two households: children in joint custody arrangements, 'commuter marriages,' Caribbean-style 'visiting relationships,' and ownership of second homes. Residence will increasingly be seen as a multidimensional concept, with different households fulfilling different household functions for the same individual. The census convention of a one-to-one correspondence between individuals and households will not be abandoned entirely, but formal and behavioral demographic analysis will of necessity go beyond it.

2. *Late Twentieth-century Trends and their Demographic Determinants*

The generalization is sometimes advanced that twentieth-century households have tended to become smaller and less complex. The generalization regarding size almost certainly is valid, that regarding complexity less clearly so. Households have tended to become smaller in large measure because on average they contain fewer minor children, consequent to widespread fertility decline. If the total fertility rate is two or below, the typical nuclear family household (husband, wife and own children) will contain three or four members or fewer. Only populations that retain traditional high fertility levels, and perhaps other features of traditional family structure, still report average household sizes of five or six persons and above.

Other factors are at work. In highly urban-industrial populations, smaller average households also have resulted from such factors as the rise of independent living of women (lone mothers, earlier home-leaving of women) and the elderly (widows and widowers), and the splitting of households following divorce. Also relevant in some societies has been the decline of live-in servants and of boarders or roomers. But these latter trends, unlike fertility decline, differ considerably across societies, depending on cultural traditions relating to coresidence with nonnuclear kin. In Japan, for example, the norm requiring taking in elderly parents remains strong. In Italy, adult children tend to remain in the parental home until marriage. Paradoxically, these familistic patterns are facilitated by low fertility, with a small nuclear family leaving space for young adults and elderly who might otherwise be crowded out. Mitchell et al. (1989), for example, show that in Canada number of siblings is an important determinant of age at leaving the parental home. These examples underline the point that household-formation behaviors feed back on one another.

In many western societies, rising rates of divorce and of nonmarital cohabitation have led to considerable fluidity in household formation, particularly among young adults and among older persons following divorce or widow(er)hood. Looking to the future, one might anticipate even greater flexibility in household arrangements, including among the elderly, as individuals develop broader repertoires of behavior based on early experience as children and young adults.

The older stereotype identifying large households with complex households of extended kin must give way to the view that smaller households can also be more rather than less complex, given greater degrees

of freedom to add members due to reduced domestic demands of smaller numbers of minor children. Like many other demographic behaviors, household formation and dissolution now show more heterogeneity in most dimensions except size.

In most societies, coresidence of kin has been lessened by high rates of migration, and the consequent location of kin in different, sometimes distant, communities. In urban-industrial societies, this includes urban-to-urban labor migration as well as retirement-related migration of the elderly. In developing societies, a major factor is rural-urban migration and, in some cases, high rates of overseas migration. Kinship contacts persist, but they are attenuated by distance.

3. *Behavioral Theories of Household Dynamics*

As with other demographic behaviors, explanation of household dynamics must be partly formal, partly behavioral. But unlike other demographic behaviors, accession to or relinquishment of a household status is not a unitary behavior. In traditional terms, an individual may become a household 'head' by setting up his or her own household, by being widowed or divorced, by the departure of one's last child (if already widowed or divorced), and so forth. It is this fact that has made the fitting of so-called 'household headship' curves relatively intractable: they are the resultant of too many different behaviours (Corner1987). A theory of household formation must, therefore, in some sense be a theory of marriage, divorce, fertility, mortality, home-leaving, etc.—all combined in some meaningful way.

3.1 *Microeconomic Theories of Household Formation*

In this field, as in demography generally, economics has had considerable theoretical influence, notably from the works of the 'Chicago School' and from empirical studies of household formation, especially the rise of one-person households. Becker (1981) viewed marriage and household formation in terms of a traditional sexual division of labor, with husband and wife coming together for practical economic advantage in a 'household economy.' Women tended to remain in the home performing domestic tasks, while men entered the labor market for wages, both pursuing comparative advantages associated with differential earning potential. Davis (1984) called this arrangement the 'breadwinner' system, and associates it with the Victorian era. Prior to that, on family farms or in small enterprises, both husband and wife worked to support the family economically.

Clearly, such a system no longer prevails in Western urban-industrial societies, as women have entered the labor market in force, pursuing careers throughout their adult lifetimes. In some contexts and especially among young adults, women often appear to have the advantage in terms of employment and income, as work requiring physical strength has declined in importance. Women are not as dependent as formerly on men for their economic well-being, and are more able to establish their own households. At the same time, the costs of maintaining desired levels of consumption have risen, so that having more than one wage-earner per household is advantageous. Thus, both members of a couple are apt to work, bringing in more income but foregoing the real economic advantages of Becker's household division of labor, even if not along traditional lines. There still are economic rationales for marriage and coresidence, but they are somewhat different than in earlier times.

A central notion in economic analysis of the household is that there are economics of scale to be gained by coresidence—in the provision of housing, utilities (heat, lighting, etc.), food and meal preparation, and consumer durables such as stoves, refrigerators, or automobiles. A *leitmotiv* in the economic literature on households is that rising levels of living have led to the purchase of greater privacy and independence, with a sacrifice of economies of scale associated with the sharing of living quarters.

In their influential article (*Privacy, Poverty and Old Age*, Beresford & Rivlin (1966)) began a research tradition of empirical research on 'living alone' in one-person households. The aim was to explain the rise in the 'propensity' to live alone, that is, the rise in age-sex-marital status specific rates of living alone, notably among older women and young adults of both sexes. The central explanatory idea was that high real income was being used to purchase more household 'privacy,' viewed as a luxury in the technical economic sense, but a luxury that had always been desired. In keeping with the tenor of economic thinking of the times, any possible shift in 'tastes' or values relating to household arrangements was played down.

Further research tended to qualify a pure 'income' explanation. Clearly, individuals would not live alone if they literally couldn't afford it. But older women, for example, lived alone in greater numbers partly due to new social programs (social security, government pensions, old-age assistance), created in response to demand or felt need, based in turn on new ideas about individual independence. The larger explanation was social and political as well as economic. Similarly, among young adults, the pursuit of independence from parents often has been intense, with youth leaving affluent homes to live in relative squalor, a pattern documented in path-breaking research based on the Survey of Labor and Income Dynamics (SLID), one of the first major longitudinal surveys (Hill & Hill 1976).

Many of these household dynamics can be related to a general breakdown of traditional patterns of age-sex

stratification (the 'decline of the patriarchy'), in the family as well as society at large. Women, older persons, children, and young adults have sought and been accorded rights and perquisites formerly reserved to the middle-aged male. This has led to greater competition *within* households for scarce household goods such as personal space, quiet, privacy, and freedom of movement (Burch 1985).

In the West, these developments clearly are bound up with a strong cultural strain towards individualism and with its full expression in situations of affluence. Whether others societies will follow is unclear (McDonald 1992). Similarly, in the absence of affluence individualism may be sacrificed, as coresidence is sought for economic reasons—the *estrategias de sobrevivencia* emphasised by the Latin American social demographic school.

In addition to economic and cultural factors, early research by Sweet (1977) and by Kobrin (1976) focused attention on a demographic supply factor in the choice of household statuses: to the extent that coresidence continued to be seen as a family matter, even an act of relative intimacy, one's options for coresidence were limited by the number and availability of suitable kin. Sweet succinctly noted that an older widow cannot live with a child if she has no living children.

3.2 Demographic Theories of Household Formation

Ermisch (1988) has developed and empirically tested what might be called a two-part model of household formation, one in which 'demographic influences on household formation and composition can be separated from their economic influences …'.(p. 24). In this approach, demographic events of marriage, birth, death, and home-leaving lead to the formation of 'minimal household units' (childless unmarried adults, lone parents with children, childless married couples, married couples with children). These 'minimal household units' then distribute themselves into distinct household arrangements in response to socioeconomic factors. The model is not meant to imply that household formation literally occurs in two temporally ordered stages, but is seen only as an analytic convenience.

Burch & Matthews (1987) have proposed a decision-making model that is nearly the opposite of Ermisch's —with couples or individuals giving priority to a choice of household living arrangement, and then deciding on demographic behaviors (marriage, fertility, etc.) accordingly.

Only detailed descriptions of decision-making processes over individual life courses would clarify the order of demographic and household decision-making in reality. Until now, adequate survey data are limited to events, not the decision processes that underlie them. And, of course, some changes in household status are not chosen by the individual: spouses or partners die or separate; grown-up children leave home; circumstances dictate taking in an elderly parent. The question then is one of inertia v. choice of a different arrangement.

Whether an integrated theory of household dynamics is possible is a moot question. But some contemporary philosophers of science (Giere 1999) would urge us not to worry on that score. Theory is seen as consisting of a large collection of theoretical models pertaining to the household, some useful for some analyses, others useful for others—theory as a toolkit, rather than as 'truth.' And the household dynamics literature is now rich with such tools.

3.3 Household Formation and 'Social Capital'

Since the 1960s, theories of demographic behavior have been subject to strong—some would say excessive—influences from neoclassical microeconomics, with various demographic choices viewed as resulting from rational and forward-looking individual assessments of costs and benefits ('lifetime utilities'). Sociologists and psychologists, among others, have objected that the economic perspective provides a narrow and unbalanced view of human behavior and social structure, and have tried to extend the core microeconomic model to deal with human realities such as emotions, social norms, and social and cultural systems (leading economists of demographic behavior also attempted such broadening, as in Becker's concern with altruism, or Easterlin's with 'psychic costs' of fertility control). A recent attempt has been the introduction into the study of fertility of concepts of 'social interaction,' 'social influence,' and 'diffusion' (Bongaarts & Watkins 1996, Montgomery & Casterline 1996).

But there have been few efforts (and no successes) at constructing a theoretical scheme that does not retain microeconomic ideas at its core—perhaps a reflection that when all is said and done, a good deal of human behavior and social structure is at base purposive and rational. Clearly this is the case with households, which serve a number of purely utilitarian functions: provision of shelter, heat, physical security, sleep; storage of personal property; and the purchase, preparation, and consumption of food. But households also meet nonmaterial needs, as the locus of various forms of companionship and intimacy, social support, and the formation of 'social capital.'

A programmatic article on this concept by Astone et al. (1999) seems destined to shape behavioral research in household demography for some time to come. These authors call for a reorientation of family and household demography around the concept of *social capital*—defined as resources of an individual that derive from his or her social ties (p. 2)—with that concept reinserted into the context of social exchange theory, from which it emerged. 'The formation of sexual partnerships, the birth and rearing of children, and both intra-generational and inter-generational

transfers constitute major forms of investment in social capital in virtually all societies' (Astone et al. 1999, p. 203). Clearly, their concern is with both households and families (including extended kinship networks) as these have been defined earlier.

This perspective raises a number of interesting questions regarding contemporary household patterns in urban-industrial societies' in general whether they have moved too far in the direction of solitary living—what Kuznets (1978) referred to as 'the apartness of adults,' with associated sacrifices of both economic and social-psychological benefits of coresidence. More specific questions might include the following: Do persons in one-person households suffer from a lack of enduring social ties, or are the four walls of a house or apartment relatively unimportant as social boundaries? Do children suffer from growing up in households with only two adults, as compared with those containing aunts or uncles, or grandparents—do they grow up with a more limited repertoire of behaviors; or, are the limitations of the household offset by increased social interaction with adults at 'daycare.' Are households in which one member of a couple (whether a woman or a man) remains at home performing 'domestic' tasks better off in terms of 'social capital' formation even if household income is lower?

The study of these and other questions pertaining to household and family structure might be strengthened by focus on issues widely viewed as 'social problems.' That is, the development of household and family demography might benefit not just from theoretical reorientation as suggested by Astone et al. (1999) but also by greater attention to real-world social problems.

See also: Family and Kinship, History of; Family Size Preferences; Family Systems and the Preferred Sex of Children; Family Theory and the Realities of Childbearing Behavior; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Intergenerational Relations; Family Theory: Role of Changing Values; Historical Demography

Bibliography

Angel R, Tienda M 1982 Determinants of extended household structure. *American Journal of Sociology* **87**: 1360–83

Astone N M, Nathanson C A, Schoen R, Kim Y J 1999 Family demography, social theory and investment in social capital. *Population and Development Review* **25**: 1–32

Becker G 1981 *A Treatise on the Family*. University of Chicago Press, Chicago

Beresford J C, Rivlin A M 1966 Privacy, poverty, and old age. *Demography* **3**: 247–58

Bongaarts J 1983 The formal demography of families and households: An overview. *IUSSP Newsletter* **17**: 27–42

Bongaarts J, Burch T K, Wachter K W (eds.) 1987 *Family Demography: Methods and their Application*. Clarendon, Oxford, UK

Bongaarts J, Watkins S C 1996 Social interactions and contemporary fertility transitions. *Population and Development Review* **22**: 639–82

Burch T K 1979 Household and family demography: A bibliographic essay. *Population Index* **45**: 173–95

Burch T K 1985 Changing age-sex roles and household crowding: A theoretical note. In: *Proceedings, International Population Conference, Florence, 1985*, International Union for the Scientific Study of Population, Liège, Vol. 3, pp. 253–62

Burch T K 1995 Theories of household formation: progress and challenges. In: van Imhoff E, Kuijsten A C, Hooimeijer P, van Wissen L J C (eds.) *Household Demography and Household Modelling*. Plenum Press, New York, pp. 85–108

Burch T K, Matthews B J 1987 Household formation in developed societies. *Population and Development Review* **13**: 495–511

Corner I 1987 Household projection methods. *Journal of Forecasting* **6**: 271–84

Davis K 1984 Wives and work: the sex role revolution and its consequences. *Population and Development Review* **10**: 397–418

Ermisch J 1988 An economic perspective on household modelling. In: Keilman N, Kuijsten A C, Vossen A (eds.) *Modelling Household Formation and Dissolution*. Clarendon Press, Oxford, pp. 23–40

Giere R N 1999 *Science Without Laws*. University of Chicago Press, Chicago

Hill D, Hill M 1976 Older children and splitting off. In: Duncan G J, Morgan J N (eds.) *Five Thousand American Families—Patterns of Economic Progress, Vol. IV, Family Composition Change*. Institute for Social Research, Ann Arbor, pp. 117–54

Höhn C 1987 The family life cycle: needed extensions of the concept. In: Bongaarts J, Burch T K, Wachter K W (eds.) *Family Demography: Methods and their Application*. Clarendon, Oxford, UK, pp. 65–80

Keilman N, Kuijsten A C, Vossen A (eds.) 1988 *Modelling Household Formation and Dissolution*. Clarendon Press, Oxford, UK

Kobrin F 1976 The fall in household size and the rise of the primary individual in the United States. *Demography* **13**: 127–38

Kuznets S 1978 Size and age structure of family households: exploratory comparisons. *Population and Development Review* **4**: 187–223

Laslett P 1982 Introduction. In: Laslett P, Wall R (eds.) *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK, pp. 1–90

McDonald P 1992 Convergence or compromise in historical family change. In: Berquó E, Xenos P (eds.) *Family Systems and Cultural Change*. Clarendon Press, Oxford, pp. 15–30

Mitchell B, Wister A, Burch T K 1989 The family environment and leaving the family home. *Journal of Marriage and the Family* **51**: 605–13

Montgomery M R, Casterline J B 1996 Social learning, social influence, and new models of fertility. In: Casterline J B, Lee R D and Foote K A (eds.) *Fertility in the United States: New Patterns, New Theories*. Supplement to *Population and Development Review* **22** (S): 151–75

Ryder N 1987 Discussion. In: Bongaarts J, Burch T K, Wachter K W (eds.) *Family Demography: Methods and their Application*. Clarendon Press, Oxford, UK, pp. 345–66

Sweet J A 1977 Demography and the family. *Annual Review of Sociology* **3**: 363–405
van Imhoff E, Kuijsten A C, Hooimeijer P, van Wissen L J C (eds.) 1995 *Household Demography and Household Modelling*. Plenum Press, New York
Zeng Y, Vaupel J, Wang Z 1997 A multidimensional model for projecting family households—with an illustrative numerical application. *Mathematical Population Studies* **6**: 187–216

T. K. Burch

Families and Households, Formal Demography of

'Family and household demography' differs from traditional demography in that it explicitly recognizes and studies *relationships between individuals*. In traditional demography, the core processes of birth, death, marriage, divorce, and migration are typically studied as occurring to individuals in isolation. In contrast, family and household demography studies how these processes occur to multiple persons interacting with each other. In doing so, additional relevant processes like entering or leaving co-residence are also studied. In short, relationships between individuals are the central focus of family and household demography. This article outlines the main subfields within the broad area of family and household demography, as well as the most important formal concepts.

1. Classification by Subject Matter

Three types of relationships between individuals can be distinguished (Ryder 1987):

(a) *Consanguineal* relationships run via the fundamental parent–child relation. Any two persons sharing a common ancestor, i.e., linked by a sequence of birth events, are said to be related by blood.

(b) *Co-residence* relationships are between persons living together. A foolproof definition of co-residence does not exist (think, for instance, of student housing, lodgers, or part-time households), but 'sharing roof and manger' is not bad for a start.

(c) *Conjugal* relationships are between individuals who are married to each other. In fact, living together in marriage is a special form of co-residence. To many people, the bond of marriage is so special that it is (almost or fully) put on the same footing as the consanguineal bonds mentioned under (a). In modern societies, the traditional marriage is gradually replaced by cohabitation (consensual unions). In these societies, 'being married' is more and more replaced by the more neutral 'living together as a couple,' which includes both formal marriages and consensual unions, and frequently also includes homosexual cohabitation. The United Nations. Statistical Division (1998) explicitly recommends: 'Couples living in consensual unions should be regarded as married couples' (p. 67).

The broad field of family and household demography studies all these links between individuals, in their own right but also in relation to other demographic processes (e.g., migration of household, mortality patterns in kin groups). Within the broad field, several subfields can be identified, according to the type of relationship on which they concentrate:

(a) *Demography of kinship* studies links of type (a) and (c).

(b) *Demography of households* studies links of type (b), but also of type (c) since 'living together as a couple' is a special form of 'living together' (see also *Families and Households, Behavioral Demography of*).

(c) *Demography of the family* studies the simultaneous links of type (a) and (c), on the one hand, and of type (b), on the other: a family is a group of co-resident kin.

(d) *Demography of marriage* studies links of type (c).

2. Data Issues

The data sources for household and family demography are the same as for demography in general: censuses, population registers, surveys (in kinship studies, genealogies are also sometimes used). However, *definitional problems* are particularly strong in household and family data.

There are two sets of definitional problems. The first relates to criteria for establishing *whether or not a certain relationship exists* between two persons. Examples of such problems include:

(a) Two persons (e.g., a mother and a student child) live together during weekends only. Do they belong to the same household?

(b) Two persons form a household. How to judge whether they 'live together as a couple'?

(c) Two students live in an apartment, occupying separate rooms but sharing kitchen and bathroom. Do they count as one or as two households?

The second group of definitional problems relates to criteria for *classifying households and/or family into categories*, like those used in published household statistics. Examples include:

(a) A woman lives with her 30-year old son. Should this household be termed a lone-parent household?

(b) A lone mother cohabits with a new partner. Is this a 'couple with children' household, or rather a 'one-parent family plus other(s)'?

Such definitional degrees of freedom result in actual definitional differences, the consequences of which are potentially very large (Keilman 1995). Therefore, demographers must be well aware of the precise definitions underlying their household and family data. In descriptive studies, definitional differences can seriously distort international or intertemporal

comparisons. In modeling exercises, it is crucial that model parameters are derived from data that are definitionally consistent with the variables in the model. In order to stimulate international comparability of household and family data, the United Nations has specified detailed, extremely useful recommendations for data collection, definition, and tabulation in population and household censuses (United Nations, Statistics Division 1998).

3. *Unit of Analysis*

A major problem in the formal demography of families and households is whether the group or the individual should be taken as the primary unit of analysis. Almost by definition, the behavior of the individuals within the group are highly interdependent. Therefore, analyzing the behavior of the group cannot be done without proper account of the characteristics of its members; conversely, analyzing the behavior of the individual cannot be done without proper account of the characteristics of the group (including those of the other individuals in the group).

Today, most family and household demographers agree that it is most convenient to work at the level of *individuals*. The main reason for this is that, unlike individuals, groups do not have a clearly defined origin and destination in time since members can enter and leave all the time. Thus, modern practice is to define a group in terms of its members and the relationships between them, rather than defining individuals in terms of the group to which they belong.

Because of strong interdependence of within-group behavior, proper modeling of individual family or household members' behavior requires sufficient information on the characteristics of the other individuals involved. In microlevel data, this implies the inclusion of *pointers* or reference variables to other individuals in the dataset, with sufficient detail on the type of the relationship to that other individual. A main advantage of such data is that it can provide very rich descriptive information on different family and household types, as opposed to standard census breakdowns by just a few main categories.

Naturally, when using sample survey data, proper account should be given to the representativeness of the data at different units of analysis. For example, a random sample of individuals is *not* representative at the household level, since large households will be over-represented. Weighting schemes must be used to correct this and other potential distortions.

4. *Models in Family and Household Demography*

A model is a simplified, quantitative description of reality. In family and household demography, models provide a formal description of the quantitative relationships between family or household characteristics and their demographic determinants and consequences (Bongaarts 1983).

At the heart of any modeling exercise lies the specification of the *state space*: the representation of the components of the system of interest. At the level of the *individual*, the state space consists of a number of characteristics or attributes (e.g., age, sex, position in the household, age of spouse, number of children), each of which can take a certain value. At the population level, the state space consists of all possible combinations of attribute values: it is a breakdown of all individuals comprising the population by relevant characteristics.

The ultimate purpose of model building is to obtain an understanding about this multidimensional breakdown of the total population. A classical demographer is satisfied with a breakdown by age and sex. A marriage demographer would want to add at least 'marital status' as an additional variable. A kinship demographer would want to add a complete breakdown by number of living children, grandchildren, siblings, parents, and so on. A household demographer would want to have, next to age and sex, variables like household size and household type, in a classification that is sufficiently detailed for the research or policy problem at hand. All these demographers have in common that they want to understand why an observed population breakdown is how it is; how particular changes in behavior (for example, total fertility level) would change this breakdown; or how the currently observed breakdown will develop over time (forecasting).

The particular problem in household and family models is that the relevant state space is potentially very large and complicated. This has forced family and household demographers to sacrifice all kinds of descriptive, behavioral, and/or conceptual detail. Evidently, the result is that existing household and family models reflect a wide range of different modeling strategies. Three important classes of these strategies are discussed in the sections below.

4.1 *Severing the Link*

As said before, family and household demography is all about relationships between individuals. A particularly important type of 'sacrifice' frequently made by family and household demographers is to discard most of the information on the characteristics of the individuals *with whom* a particular relationship exists. For example, in the case of a marriage model, a woman aged 40 and married to a man aged 42 is classified as being a married woman aged 40: the information about the age of the husband is discarded. Without discarding, the total population would require a table of $2 \times 100 \times (1 + 100) = 20{,}200$ cells (2 sexes, 100 age groups, 1 for nonmarried, 100 for married by age of spouse); with discarding, a table of $2 \times 2 \times 100 = 400$ cells is sufficient (2 sexes, 2 for

nonmarried vs. married, 100 age groups). In other words, a frequently used modeling strategy is to collapse all information on all individuals to whom person x has relevant relationships into just a few attributes of person x, thus in fact 'severing' the explicit link to those related individuals.

An example of this 'severing the link' strategy is provided by marital-status models of the multistate type (see *Multistate Transition Models in Demography*). Here, individuals are classified by sex s (m for men, f for women), age x, and marital status i, where i can take the following values: N for never married, M for married, D for divorced, W for widowed. The development over time of the population is described in terms of *events*, which are changes in the attributes of the individual members of the population. For example, next year's 31-year old married female population is given by (ignoring migration):

$$\begin{aligned} P(f, 31, M, t+1) &= P(f, 30, M, t) \\ &+ E(f, 30, N, M, t) + E(f, 30, D, M, t) \\ &+ E(f, 30, W, M, t) - E(f, 30, M, D, t) \\ &- E(f, 30, M, W, t) - E(f, 30, M, dead, t) \end{aligned}$$

where $P(s, x, i, t)$ denotes the population at the start of year t of sex s, age x, in marital status i, and $E(s, x, i, j, t)$ denotes the number of individuals of sex s and age x who during year t experience an event from position i to position j. The various numbers of events depend on the size of the relevant population at risk of experiencing the event, and on the rate at which individuals in the population at risk experience the event. For example, the number of women who marry for the first time at age 30 in year t, i.e., $E(f, 30, N, M, t)$, equals the product of two factors: the number of never-married 30-year old women $P(f, 30, N, t)$, and the first-marriage rate for 30-year old women $M(f, 30, N, M, t)$.

Given a full set of age-, sex-, origin-, and destination-specific rates, the population by age, sex, and marital status can be projected forward in age and time. Also, a marital-status life table can be calculated. However, a crucial characteristic of this model is that marital-status events for one sex are essentially independent of the demographic situation for the other sex. As a result, in this model it is not guaranteed that the resulting population flows for men and women will be mutually consistent. For example, the number of married men having a divorce should equal the number of married women having a divorce, but there is no mechanism in the model to ensure this equality. This problem is known as the *two-sex problem* in nuptiality models (Keilman 1985a). It is a direct consequence of the fact that the explicit link between marriage partners is missing. There are several solutions to the two-sex problem, each of which introduces fresh problems:

(a) To restore the link between marriage partners. This greatly expands the state space, since married persons must be simultaneously classified by own age *and* by the age of their spouse. It also greatly complicates the modeling of numbers of events, since rates must be defined for all combinations of male *and* female ages and marital statuses (see also *Population Dynamics: Two-sex Demographic Models*).

(b) A slightly different way of restoring the link between marriage partners is to recast the model in terms of *microsimulation*. See Sect. 4.2.

(c) To concentrate on one sex (usually: women) only. This is convenient, but obviously unrealistic.

(d) To introduce an adjustment mechanism, by which aggregate numbers of events (i.e., aggregated over all ages) are forced to satisfy certain requirements. For example, if the model initially produces 1,000 male divorces and 800 female divorces, male divorces are adjusted downward and female divorces upward to result in, say, 900 divorces for each sex. Such procedures are called *consistency algorithms* (Keilman 1985b, Van Imhoff 1992). Although such algorithms can be given a satisfactory behavioral interpretation, they nevertheless have an artificial ring and are certainly rough approximations only.

(e) A combination of the previous two 'solutions' leads to consistency algorithms that are one-sex dominant. For example, in a female-dominant marital status model, the total number of male divorces is determined by the female married population and female divorce rates.

Marital status models are by no means the only models in which 'severing the link' leads to consistency problems. All household and family models which are of the multistate type suffer from essentially the same problem.

4.2 Macro- vs. Micromodels

In principle, an algebraic representation of a model can be manipulated to study the model's properties. If the model is not too complicated, it could be analyzed in this way. Because family and household models easily get quite complex, analytic models are rare in family and household demography. A good example of an analytic model (although valid under a number of quite restrictive assumptions only) is the kinship model of Pullum (1982).

For more complex models, analytic manipulation becomes infeasible, so that *numerical simulation* methods have to be used to study the implications of the model. Here, two fundamentally different strategies can be distinguished: *macro*simulation vs. *micro*simulation (Van Imhoff and Post 1998; see also *Microsimulation in Demographic Research*).

In macromodels, numbers of events (e.g., 'first marriages to women aged 25') are obtained by applying rates (e.g., 'first-marriage rate for 25-year old women') to a *group* of individuals of a certain size (e.g., 'number of never-married women aged 25'). In

micromodels, events are obtained by applying probabilistic decision rules (Monte Carlo experiments) to *individual persons*, separately for a number of individual persons. To follow the example of first marriages: for each individual never-married woman aged 25 in the database, draw a random number between 0 and 1; if this random number is less than the first-marriage probability, the woman is deemed to marry, otherwise she remains unmarried.

In macrosimulation, the calculations are carried out in terms of the cells in the aggregate cross-classification table: for each cell, the model should evaluate how the number it contains will change over time. Microsimulation, on the other hand, does its calculations in terms of the individual records: for each individual, the attribute vector is updated according to the specifications of the model and the results of the Monte Carlo experiments. The storage and retention of information in microsimulation occurs via a *list* of individuals and their attributes; in macrosimulation this is done via the aggregate cross-classification *table*. If there are K attributes and M_i categories for attribute $i = 1 \ldots K$, the table at the macrolevel consists of $M_1 \times M_2 \times \ldots \times M_K$ cells; in contrast, at the microlevel a total population of N individuals can be described by a matrix with $N \times K$ cells. For most applications where a relatively large number of attributes is considered, the size of the aggregate table is much larger than the size of the list. This is particularly the case in household and family models, where characteristics of the individuals to whom relationships exist need to be included. In macromodels this would lead to incredibly large tables.

In micromodels, contrary to macromodels, it is quite easy to keep links between individuals intact, simply by including in the individual records of the database some reference numbers (pointers) to other records (persons) in the database. As a result, the consequences of an event that is simulated for one person can be easily determined and updated for the other persons involved. Thus, the two-sex problem or more generally the consistency problem discussed in the previous section is more or less 'automatically' solved in micromodels. The other side of the coin is that, for events that involve creating a *new* link between individuals, a matching problem has to be solved (e.g., if a woman is simulated to marry, a concrete husband must be identified).

Another main advantage of microsimulation is that it can provide very *rich output*. The output of a microsimulation model consists of a database with individual data that can be aggregated in an almost infinite number of ways. Compare this with the situation in macromodels, where the aggregation scheme is fixed once the model (i.e., the state space) has been specified. Apart from detailed cross-sectional counts, the microsimulation database can also be used to construct longitudinal information, e.g., in the form of individual biographies.

In short, microsimulation offers three main advantages that are of particular relevance in family and household demography: it can handle a large state space; relationships between individuals can easily and explicitly be retained; and it provides richer output.

This does not automatically imply that microsimulation is the ideal methodology for family and household modeling. As always, these advantages are bought at a cost, and the cost can be heavy. The most important drawbacks of microsimulation are twofold: first, the model results are subject to random variation, and the degree of this randomness gets larger as more explanatory variables are included; second, it takes a lot of time and effort to construct a concrete microsimulation application.

4.3 Dynamic vs. Static Models

Populations change over time, and so do household and family structures. Thus, if a model is to describe this change, is must always contain the *time* element in one way or another. In this sense, all demographic models are dynamic by definition. However, there are many ways in which time can be included into a model. Merely adding an index t to all model variables and parameters hardly warrants the term 'dynamic.' A truly dynamic model should not only specify what the system looks like 10 years from now, but also how the system is supposed to get from here to there. In other words, the *processes* that underlie the changes in the system variables should be explicitly included in the model. In a truly dynamic model, the focus is on the *events* of formation and dissolution of relationships between individuals. The LIPRO household model belongs to this category (Van Imhoff and Keilman 1991, Van Imhoff 1995).

In contrast, *static* models do little more than compare household and family structures at different points in time, thus essentially treating the underlying processes of household and family formation and dissolution as a black box. The classical headship ratio models belong to this category. The headship ratio is the proportion of household heads in a population category, e.g., '80 percent of all 40-year old males are head of a household.' If from a population forecast (see *Population Forecasts*) we know that the number of 40-year old males will increase from 1,000 to 1,500, and we extrapolate that the headship ratio for 40-year old males will increase from 80 percent to 90 percent, then it follows that the number of households headed by a 40-year old male will increase from 800 to 1,350. With headship ratios for all ages and both sexes, a full population forecast can in this way be easily converted into a forecast of the number of (private) households. Several generalizations of the classical headship ratio model exist (Kono 1987). However, they all suffer from the disadvantage that the headship ratio is a static concept: the model cannot tell us why the current ratio of 80 percent will increase to 90 percent.

In a way, the dynamic vs. static issue is gradual rather than fundamental. For instance, if a demographic forecaster hypothesizes that marriage rates will fall by 10 percent between now and 10 years into the future, one could argue that the marriage component of the model is static since the demographic forecaster does not specify *how* the marriage rates are going to fall. Similarly, the headship ratio method is generally termed static since the changes in age-specific headship are not explicitly specified; however, the changes in the age-specific population size to which these headship ratios are applied *are* explicitly modeled, so the headship ratio model is dynamic at least to some extent.

In fact, leaving some model variables static rather than 'truly' dynamic is an effective strategy to reduce the complexity of the model. The decision of where to stop making variables dynamic is governed by the fundamental tradeoff between information intensity on the one hand, and the capacity to make meaningful predictions on the other. This tradeoff must be faced in every effort of modeling human behavior.

The household model used by Statistics Netherlands for its official household forecasts for The Netherlands is a nice example of a carefully constructed mixed dynamic-static household model (De Beer 1995). This model consists of two layers. The first layer is a multistate marital status model, producing forecasts of the population by age, sex, and marital status. In the second layer, each combination of age, sex, and marital status is distributed over various *household positions* (e.g., living alone, with partner, as lone parent, as dependent child, etc.), using household position proportions which are extrapolated into the future. The rationale for this strategy is that the available statistical data are sufficiently reliable to allow a dynamic approach to nuptiality, but do not warrant a dynamic treatment of household formation and dissolution. The resulting model is dynamic to the extent that household events are related to nuptiality behavior.

See also: Family Theory: Economics of Marriage and Divorce; Household Archaeology; Household in Anthropology; Kinship in Anthropology

Bibliography

Bongaarts J 1983 The formal demography of families and households: An overview. *IUSSP Newsletter* **17**: 27–42

De Beer J 1995 National household forecasts for The Netherlands. In: Van Imhoff E, Kuijsten A, Hooimeijer P, Van Wissen L (eds.) *Household Demography and Household Modeling*. Plenum, New York/London

Grebenik E, Höhn C, Mackensen R (eds.) 1989 *Later Phases of the Family Cycle*. Clarendon Press, Oxford, UK

Keilman N W 1985a Nuptiality models and the two-sex problem in national population forecasts. *European Journal of Population* **1**: 207–35

Keilman N W 1985b Internal and external consistency in multidimensional population projection models. *Environment and Planning A* **17**: 1473–98

Keilman N W 1995 Household concepts and household definitions in Western Europe: Different levels but similar trends in household developments. In: Van Imhoff E, Kuijsten A, Hooimeijer P, Van Wissen L (eds.) *Household Demography and Household Modeling*. Plenum, New York/London

Keilman N W, Kuijsten A, Vossen A (eds.) 1987 *Modelling Household Formation and Dissolution*. Clarendon Press, Oxford, UK

Kono S 1987 The headship rate method for projecting households. In: Bongaarts J, Burch T K, Wachter K W (eds.) *Family Demography*. Clarendon Press, Oxford, UK

Pullum T W 1982 The eventual frequencies of kin in a stable populations. *Demography* **19**: 549–65

Ryder N 1987 Discussion. In: Bongaarts J, Burch T K, Wachter K W (eds.) *Family Demography*. Clarendon Press, Oxford, UK

United Nations. Statistical Division 1998 *Principles and Recommendations for Population and Housing Censuses, Revision* 1. United Nations, Department of Economic and Social Affairs, Statistics Division, Statistical Papers Series M No. 67/Rev.1, New York

Van Imhoff E 1992 A general characterization of consistency algorithms in multidimensional demographic projection models. *Population Studies* **52**: 159–69

Van Imhoff E 1995 LIPRO: A multistate household projection model. In: Van Imhoff E, Kuijsten A, Hooimeijer P, Van Wissen L (eds.) *Household Demography and Household Modeling*. Plenum, New York/London

Van Imhoff E, Keilman N 1991 *LIPRO 2.0: An Application of a Dynamic Demographic Projection Model to Household Structure in The Netherlands*. Swets & Zeitlinger, Amsterdam/Lisse

Van Imhoff E, Post W 1998 Microsimulation methods for population projection. *Population* **52**: 889–932

E. van Imhoff

Families as Educational Settings

1. Introduction

Children's education is viewed here in broad terms, reflecting not only academic standing in the formal setting of school, but also general cognitive development. In that vein, family influences on children's educational progress are myriad, as family provides the primary context for children's development and also reflects the broader societal and cultural contexts within which the family is embedded. Both change and continuity in family influences are noted during infancy, early childhood, and the school years.

2. Early Research

Much of the early research in family influences on children's educational progress examined ways in which academic outcomes were related to demo-

graphic characteristics such as parent marital status and education, ethnicity, birth order, and socioeconomic status. This early work revealed consistent, moderate relationships between most of these factors and children's cognitive development. In particular, more positive educational outcomes were found in families where the parents' marriage remained intact, where parents were better educated and had higher income, and for children who were first- rather than later-born. Psychologists and educators have noted that such group comparisons, which have been labeled 'social address' (Bronfenbrenner 1979), are less informative than research which identifies the processes underlying these group differences. Four categories of process variables are summarized here: parental beliefs, instruction from parents and siblings, structuring of the home environment and family life, and parents' behaviors outside the home.

3. *Parental Beliefs*

Parents' beliefs affect children's educational outcomes both through their influences on parental behaviors and on children's developing beliefs. For instance, a parent will speak to her 6-month-old daughter more often if she believes the infant can benefit from 'conversations' with an adult partner. Parents' beliefs also shape the developing beliefs of their children, many of which serve a motivational role for the child. For instance, a child whose father conveys that a failure was due to lack of effort rather than lack of ability will be more likely to persevere on a future task. Three types of parental beliefs are described here: parents' educational aspirations for their children, attributional beliefs about the reasons underlying academic outcomes, and beliefs about child rearing.

3.1 *Parents' Educational Aspirations*

Most parents have expectations for their children's performance in school and eventual educational attainment. Both aspirations regarding the child's future education level and projections of what specific grades a child will obtain are frequently highly optimistic. Thus, for instance, many parents who themselves have failed to complete high school expect that their own children will be college graduates.

It is clear that children's competencies inspire parents' aspirations (i.e., parents have higher expectations of more competent children). In general, the research literature also indicates an independent effect of parent expectations on children's achievement: Parents with more ambitious aspirations have, on average, children who excel academically and who are more likely than other children to pursue higher education (e.g., Okagaki and Frensch 1998). Parents who value education and who have high aspirations for their children are likely to foster achievement striving in their children and to provide a supportive learning environment in the home. Another way in which high parental expectations foster children's educational development is through the degree of challenge parents provide in learning situations.

3.2 *Parents' Attributional Beliefs*

The classic work of Rotter (1966) and Weiner (1979) demonstrated that individuals vary in their beliefs about the reasons underlying success and failure, and that future performance is related to these beliefs. When failure is attributed to an internal controllable factor such as effort, the individual is more likely to persist than if the failure is attributed to lack of ability or an external factor such as teacher preference or luck. Conversely, a belief that success is due to the internal factors of high ability and hard work is likely to foster persistence. Attributional beliefs are thus posited to influence children's outcomes largely through their motivational effect on children's persistence in the face of failure and/or challenging tasks.

Much of the research in parents' attributions has entailed cross-cultural investigations of ethnic or national differences in children's academic achievement. This work shows that Asian parents are more likely to attribute their children's academic outcomes to effort than are parents of European cultural heritage; these beliefs are then presumed to lead to academic superiority of Asian students over European American students, particularly in math. Although considerable evidence has supported the notion that children's beliefs influence their achievement striving, particularly in the face of failure, it is unclear that beliefs explain cross-national differences in achievement. European parents are more likely than European Americans to espouse ability attributions, yet children's math achievement in a number of European countries is higher than in the USA. Cultural variations in beliefs are undoubtedly confounded with other cultural and national differences such as schooling characteristics.

3.3 *Parents' Beliefs about Child-rearing*

Parents' beliefs about child-rearing also contribute to the educational qualities of the home environment. These beliefs include parents' views of optimal ways to socialize children (e.g., discipline practices, how and to what extent autonomy is fostered) and age-related expectations of children's abilities. A substantial body of literature indicates more positive cognitive and social outcomes for children whose parents show warmth while exhibiting firm yet flexible control, a style called democratic parenting by Baldwin (1955), and later described by Baumrind (1973) as authoritat-

ive parenting. Authoritative parents rely primarily on reasoning as a discipline tool rather than punishment and view their children's good behavior as the primary goal of their parenting rather than obedience. Thus, parents who value their children's autonomy, and whose knowledge and beliefs are consistent with an authoritative parenting style are more likely to foster educational progress in their children, as marked by school grades, low incidence of discipline problems, and healthy self-esteem. However, some recent research indicates cultural variation in the relationship between parenting style and children's outcomes, both in the importance of autonomy encouragement and the possible detrimental effects of harsh discipline.

Parents vary tremendously in their beliefs about children at different ages. For instance, some parents expect that 18-month-olds can be potty-trained, whereas other parents are nonplussed when the three-year-old continues to have 'accidents.' By the same token, some parents expect their three-year-olds to master rudimentary concepts of number and quantity that other parents would not expect before the age of seven. Although parents may not be conscious of many of these beliefs, they nevertheless shape everyday activities in the household, family rules, and the messages children are given—both blatant and subtle—regarding their abilities. For example, parents who believe that an eight-year-old is capable of planning are more likely than other parents to expect their child to pack his/her school lunch, select weather-appropriate clothes, and budget time to include homework, soccer practice, and leisure in a given evening.

Research in this area has indicated that a U-shaped function operates; expectations that are either too low or too high are not optimal for the child. According to Vygotsky's concept of scaffolding (1978), parents foster children's cognitive progress when they provide graduated cues that enable the child to function at a level slightly beyond what the child could perform independently. A parent who persistently has low expectations is unlikely to challenge the child to progress further. On the other hand, expectations that are too high (e.g., attempting to teach algebra to an average kindergartner) are equally unlikely to enhance the child's development.

4. *Instruction by Parents and Siblings*

Parent beliefs affect child outcomes primarily through their impact on parent behaviors. One type of parent behavior that shapes the educational quality of the home is parent instruction. In addition, siblings are an important source of information for young children. Instruction from parents and siblings are discussed here together. From birth onward, parents and siblings are a very important source of children's knowledge. The parent's early language use guides the development of the infant's ability to distinguish among phonemes. The toddler observes an older sibling's interactions with friends and imitates that behavior later in other social situations. A young adolescent acquires concepts of social stereotypes through his parents' conversation at the dinner table.

Family influences are readily evident in a child's acquisition of specific skills and knowledge about the world. Thus, for example, a child whose parents are swimmers will probably learn to swim, and the child of musicians is likely to know more about music than peers from nonmusical families. Family also provides the primary context for children's cognitive development in less obvious ways. Through everyday conversations and activities with family members, children acquire concepts of schemas (e.g., what a visit to the doctor's office entails), grammar, and mathematical concepts (e.g., the early distinction between 'one' and 'more;' how to fairly divide the last piece of cake). Children learn memory strategies (e.g., where can I put my show-and-tell item so I won't forget it tomorrow morning), meta-cognitive knowledge (I can't concentrate on my homework while big brother is listening to that music), and problem solving (How should I clean up this spilled sugar?).

Some instruction from family members is verbal; other 'instruction' is nonverbal, communicated through body language and mood, or is modeled through behavior that the child later imitates. Instruction that takes place in the home is markedly different from that found in the school setting, in that it is largely unplanned, informal, and context-grounded, rather than the planned, formal, and context-independent instruction that occurs in school. Most instruction in the home occurs spontaneously based on the specific events, feelings, and needs of a given moment in the family life. Because it is so highly context-dependent, this instruction from siblings and parents provides the basis for most of the child's knowledge and skills of how to survive in the day-to-day world.

5. *Structuring of the Home Environment*

In addition to their beliefs and instruction, parents shape the educational qualities of the home environment in important ways through their choices regarding the physical qualities of the home and the child's activities. The well-established relationship between socioeconomic status and children's academic success has been attributed in part to substantial material differences in the home environments of children. Space, availability of developmentally appropriate games and toys, warm and responsive caregivers, and exposure to cognitively stimulating adults and new experiences have all been found to be related to later IQ or school achievement. A child's needs change with age. For instance, a supportive

home environment for infants includes a responsive mother and appropriate play materials, whereas in mid-childhood emotional climate of the home and child's participation in the family are associated with higher school achievement (Bradley et al. 1988).

In addition to the physical and social characteristics of the home environment, the lifestyle choices that parents make for their children determine on a day-to-day basis the educational experiences of the children. These lifestyle choices include how children spend their time outside of school; parental supervision of homework; the amount of television the child is permitted to watch; the extent to which play with peers is monitored; and so forth. All of these factors are related to children's school achievement. Differences in the educational outcomes of children of divorce versus intact marriages is now attributed in part to reliable differences in the home environments and daily activities of these groups of children. Single parents, who on average must survive on a lower income than couples, have fewer material and emotional resources to devote to their children. Thus children in single parent households are more likely than peers in two-parent households to have little adult supervision during after-school hours, to interact with parents who are stressed, and to shoulder heavy domestic responsibilities at an early age. These qualities of the home environment rather than parental marital status have the strongest influence on children's cognitive and emotional outcomes.

6. *Parents Outside the Home*

The child's development occurs within a multilevel ecological framework (Bronfenbrenner 1979). In addition to the microsystem level, where the child is involved in all interactions (e.g., family, school, church), other settings influence the child's development. At the mesosystem level, microsystem interactions occur without the child's presence (e.g., a parent attending a school open house). The exosystem, in turn, is comprised of those settings which never include the child, but which nonetheless influence the child. Two of these broader settings are discussed here: parents' employment (an exosystem influence) and parents' involvement in the child's school (a mesosystem influence).

6.1 *Parental Employment and Children's Development*

Early investigations in parent employment and children's outcomes were mostly group comparisons aimed at determining whether or not maternal employment had a detrimental effect on children. These investigations indicated no differences in quantity or quality of parenting systematically related to maternal employment status. More recent research has identified process variables related to parent employment that have both positive and negative influences on children.

According to scarcity or stress theory, parents have a limited amount of time and energy, and parental employment is deleterious to children's developmental outcomes to the extent to which it decreases parents' time, energy, and patience. Stress theory reflects a strong trend in the USA and many other industrialized nations: Namely, in spite of the impressive technological advances of the past century, most parents work harder than ever (Schor 1991). Particularly as maternal employment rates have increased, the everyday lives of children have become harried. Thus one consequence of parental employment for children is the resulting stress associated with parental employment-related absence, whether that absence is actual or psychological, as stressed parents are preoccupied, tired, and impatient. A second consequence of demographic changes in employment patterns is that children spend a greater amount of time with other caretakers. Therefore, parents' selection of day care for infants and preschoolers, and their choices of after-school activities for school age children have real impact on children's developmental outcomes.

Parental employment can also benefit children's educational progress. On the one hand, parents gain in self-esteem, competency, and social support (as well as income) through their employment; these benefits for parents may translate into benefits for children. Mothers who are employed in professional positions serve as positive role models for daughters. In addition, parents' work characteristics are related to the quality of their parenting. Parents who enjoy their work, who are socially rewarded for it, and who are intellectually challenged in the work setting are more likely to show warmth and greater flexible but firm control with their children than parents who have less satisfying employment (Greenberger et al. 1994).

6.2 *Parents' Involvement in Schooling*

Formal education is an important part of the everyday lives of children and—along with family—is the most important source of children's educational progress in most countries. Therefore, parents' attitudes toward school and involvement in school activities have long been studied as an important factor related to children's educational progress. Parent involvement in children's formal education is positively related to their children's achievement: Parents of higher-achieving children tend to have more positive attitudes toward school, attend school functions, volunteer at their children's schools, and know teachers better than parents of lower-achieving children. Some theorists have pointed out that families of higher socioeconomic status are more comfortable with the cultural milieu of

schools, and thus find it easier to approach teachers and school administrators than families of lower socioeconomic status. These barriers to parent involvement are even greater when parents represent an underprivileged minority group and have had previous contacts with school officials that were negative (Clark 1983).

7. Current Research Directions

The home is one setting embedded within broader societal, cultural, and generation settings. Much current research examines how the family is changing due to changes in demographic factors such as parent employment, poverty, divorce and remarriage, and migration. Current research also emphasizes cultural variations in family life, both in terms of similarities and differences across groups, and in understanding what is optimal for children's healthy development (e.g., Okagaki and Frensch 1998). Finally, current research seeks to better understand family as an educational setting within a developmental framework, studying both the ways in which family influences change across childhood and the long term effects of early enhancement or insult.

See also: Education and Employment; Family and Schooling; Family as Institution; Family Processes; Nontraditional Families and Child Development; Parenting in Ethnic Minority Families: United States; Parents and Teachers as Partners in Education; Socialization in Adolescence; Socialization in Infancy and Childhood

Bibliography

Baldwin A L 1955 *Behavior and Development in Childhood*. Dryden Press, New York

Baumrind D 1973 The development of instrumental competence through socialization. In: Pick A D (ed.) *Minnesota Symposium on Child Psychology*. University of Minnesota Press, Minneapolis, MN, Vol. 7, pp. 3–46

Bradley R H, Caldwell B M, Rock S L 1988 Home environment and school performance: A ten-year follow-up and examination of three models of environmental action. *Child Development* **59**: 852–67

Bronfenbrenner U 1979 *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press, Cambridge, MA

Bronfenbrenner U 1986 Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology* **22**: 723–42

Clark R 1983 *Family Life and School Achievement: Why Poor Black Children Succeed or Fail*. University of Chicago Press, Chicago

Connors L J, Epstein J L 1995 Parent and school partnerships. In: Bornstein M H (ed.) *Handbook of Parenting. Applied and Practical Parenting*. Erlbaum, Mahwah, NJ, Vol. 4, pp. 437–58

Greenberger E, O'Neill R, Nagel S K 1994 Linking workplace and homeplace: Relations between the nature of adults' work and their parenting behaviors. *Developmental Psychology* **30**: 990–1002

Hart B, Risley T R 1995 *Meaningful Differences in the Everyday Experience of Young American Children*. Brooks, Baltimore, MD

Okagaki L, Frensch P A 1998 Parenting and children's school achievement: A multiethnic perspective. *American Educational Research Journal* **35**: 123–44

Rotter J B 1966 Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs* **80**: No. 609

Schor J B 1991 *The Overworked American*. Basic Books, New York

Scott-Jones D 1984 Family influences on cognitive development and school achievement. *Review of Research in Education* **11**: 259–304

Vygotsky L S 1978 *Mind in Society*. Harvard University Press, Cambridge, MA

Weiner B 1979 A theory of motivation for some classroom experiences. *Journal of Educational Psychology* **71**: 3–25

B. Kurtz-Costes

Family and Consumer Sciences: United States

Family and Consumer Sciences is the study of relationships among individuals, families, and communities and the social, economic, political, biological, physical, and aesthetic environments in which they function. An early widely accepted definition stated:

> … in its most comprehensive sense (it) is the study of the laws, conditions, principles and ideals which are concerned on the one hand with man's immediate physical environment and on the other hand with his nature as a social being, and is the study specifically of the relations between those two factors (Lake Placid Conference 1902).

Themes which characterize the field include (a) holistic, systems approach to the study of individuals, families, and communities, (b) interdisciplinarity as the field draws knowledge from the social, behavioral, and life sciences, from technology and engineering, and from the arts and humanities, and (c) integration and synthesis of concepts as they are applied to well-being. A continuum from discovery and integration of knowledge through its application in practical settings also is a defining characteristic of the field.

Most of the work of the profession fits this characterization; however, degrees of emphasis on particular elements and the desire to communicate that emphasis through nomenclature has led to the use of a variety of names. This is most evident in the naming of colleges in higher education institutions in the United States. Human ecology, human sciences, human environ-

mental sciences, human resources, as well as family and consumer sciences are used to designate this field of study in the academy.

Family and consumer sciences is mission-oriented. As a profession, field of study, and critical science, the mission of family and consumer sciences is 'the empowerment of families to function interdependently and the empowerment of individuals to perform family functions' (Green 1996). The mission-oriented nature of family and consumer sciences is reflected in statements about areas of leadership for the profession, to wit:

(a) improving individual, family, and community well-being;

(b) impacting the development, delivery and evaluation of consumer goods and services;

(c) influencing the development of policy; and

(d) shaping societal change; thereby enhancing the human condition ('The Conceptual Framework for the 21st Century' 1994).

1. Definition and Body of Knowledge

Family and consumer sciences is the study of relationships among individuals, families, and communities and the reciprocal relations of family to its human-created and natural environments. Specific elements in the body of knowledge have varied somewhat over time. A task force of individuals from many of the organizations linked to family and consumer sciences began updating the specification of the body of knowledge in 2000. The work of this group is incorporated in Fig. 1, which depicts the core concepts in family and consumer sciences (Baugher et al. 2000).

1.1 Basic Human Needs

A primary expectation of families in most societies is that they work toward meeting basic human needs. Abraham Maslow's conceptualization summarizes these needs as physiological needs (food, shelter, clothing), safety (security and the absence of fear), love and belongingness (human emotions connecting individuals in human systems), self-esteem (sense of competence and respect for self), self-actualization (achieving one's human potential) (see Paolucci et al. 1977).

Starting from the center of this model and moving outward, three interdependent concepts are represented: individual well-being, family well-being, and community vitality. Subsumed within these concepts, but not delineated, are multiple specific areas of research within family and consumer sciences. These concepts also represent the mission-oriented nature of family and consumer sciences. Together, the three areas reflect the ultimate concern of family and consumer sciences to enhance the quality of life.

1.2 Individual Well-being

Individual well-being refers to the promotion of a person's physical and mental health, knowledge and

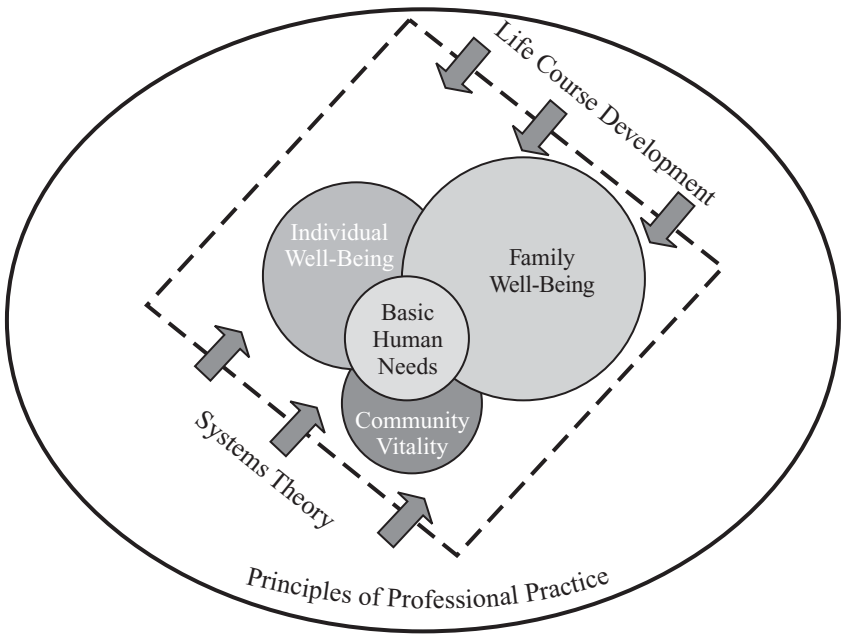

Figure 1
Core concepts in family and consumer sciences

skills, values, and civic responsibilities. Research in this domain may be as basic as studies to understand human metabolic processes and human genetic make-up, personality traits, or self-esteem as reflected in the selection of apparel. Application of this concept is reflected in numerous family and consumer sciences programs to increase knowledge and skills and foster more effective personal functioning.

1.3 Family Well-being

Family well-being is concerned with the development and dissemination of knowledge designed to empower families to meet the basic human needs of their members and engage in processes that build sustainable viable communities. Roles, responsibilities, and functions of the family have predominated the concerns of family and consumer sciences rather than structure (Green 1996). This is reflected in the definition of family adopted by the American Association of Family and Consumer Sciences. Family is defined as 'two or more persons who share resources, share responsibility for decisions, share values and goals, and have a commitment to one another over time' (Bivens et al. 1975, p. 26).

In a statement on the 'intellectual ecology' of family and consumer sciences Green (1996) identifies the substantive core of family and consumer sciences as family functions. The family functions framework addresses the role of families in meeting the needs of its members (i.e., basic human needs). In the developed economy and contemporary society of the United States, there is great interdependence of families and other systems in producing the goods and services to meet these needs. Accreditation documents of the American Association of Family and Consumer Sciences have delineated these aspects of human needs in relation to the near environment that are of concern to family and consumer sciences (see Council for Accreditation 1995). The following concepts would generally be found in such a listing:

(a) Family systems, including economic, social, cultural, political, and demographic factors; family dynamics and roles;

(b) Interdependence of families and communities;

(c) Human development across the life course;

(d) Resource development and management, including production, consumption, disposition, management and decision-making;

(e) Human nutrition and food in the development, health and wellness of individuals, families, and communities;

(f) Design and changing technology in relation to the development of home, community, and work environments that are supportive of consumers and families;

(g) Role of apparel and textiles in meeting the physical, social, and aesthetic needs of consumers and families.

1.4 Community Vitality

Community vitality fosters interdependency and collaboration designed to enhance the well-being of individuals and families. This concept may apply to those who reside in a specific locality or to groups who perceive themselves as a community based on shared interests and beliefs. Technology makes it possible to sustain community vitality in cyberspace. Community vitality can encompass the global community.

1.5 Systems Theory

A system is defined as interrelated elements having functional unity within a larger system. 'Ecosystem' is the regular interdependency of organisms within their environment. The basic concepts found in systems theory are development, relationships, interdependence, and reciprocity. The fundamental theoretical framework that gives family and consumer sciences its dynamic and integrative nature is family ecosystems. Paolucci et al. (1977) define family ecosystems as follows:

> Family members, their external environments as perceived by them, and the web of human transactions carried out through the family organization constitute the basic elements of the family ecosystem. One fundamental characteristic of the family ecosystem is that it is made up of a collectivity of interdependent but independent parts working together to achieve a common purpose. Each element (organism and environment) is interrelated (p. 15).

Human ecology theory is congruent with family ecosystems, and human ecology theory is often identified as the theoretical foundation of family and consumer sciences. Bubolz and Sontag (1993) identify the uniqueness of human ecology theory as its 'focus on humans as both biological organisms and social beings in interaction with their environment' (p. 419). The use of the human ecological perspective is found in many social science disciplines and applied fields.

1.6 Life-course Development

The concept of life course refers to a sequence of socially defined, age-graded events and roles that individuals enact over time (Elder 1998). Life-course thinking appears in most of the behavioral sciences scholarly literature after the 1960s.

As understood in life-course theory, human development is a co-active process in which social, cultural, psychological, and biological influences interact over time in the context of culture and social structures (Elder 1998). The concept of life-course development for individuals has been extended in family and consumer sciences to families, which also are observed to have experiences and characteristics described as life course events. An understanding of

life-course development is important in family and consumer sciences because environments in which families function, and families themselves, change over time.

1.7 The Professional Context

As shown in the model, principles of professional practice are a part of the core concepts of family and consumer sciences. Brown and Paolucci (1979) identify three criteria for identity as a profession:

(a) A profession is oriented toward providing a service;

(b) The service involves intellectual activity, including practical judgment, which requires that the professional master theoretical knowledge related to the work;

(c) Members in the profession seek to assure that work within the profession is morally defensible both in nature and in the quality of performance.

Family and consumer sciences includes courses and experiences to facilitate a professional orientation with regard to ethics, valuing diversity, critical thinking skills, working in partnerships, and advancing policies that support the well-being of individuals, families, consumers, and communities.

1.8 Synergistic/Integrative Approach

Family and consumer sciences claims as one of its distinguishing characteristics a 'synergistic, integrative nature' (Council for Accreditation 1995). The overlapping circles, dotted lines, and arrows in Fig. 1 depict this interplay among concepts. Addressing issues of wellness, for example, in the context of multiple factors rather than as isolated elements is the essence of the synergistic concept. For example, a dietitian needs to understand that obesity and impending cardiovascular disease are not just a matter of nutrition. Obesity is probably related to cultural food patterns, economic resources, and family relationships.

The integrative focus entails recognizing the interrelatedness of individuals, families, and consumers with their environments. In the example, the absence of sidewalks and parks with walking trails in a community may be a barrier to the state health agency's goal of reducing obesity through fitness programs.

2. Specializations and Family and Consumer Sciences

Like a chameleon, family and consumer sciences is perceived differently by observers depending on the context and the 'colors' highlighted in that context. In the mid-twentieth century, Lee and Dressell (1963) observed three conceptions of the field, as follows:

(a) a single field with a broad general perspective and a number of specialties;

(b) a unified field with subspecialties embedded in the home and family;

(c) a collection of disciplines with no unifying theme or 'anchor.'

Families may be recognized as the basic social unit of society, but technology and changing social norms increasingly influence how family functions are carried out. In economically developed countries, such as the United States, the household produces only a small portion of the goods and services it consumes. The family has become increasingly a managerial unit.

As family roles changed and new technology, products, and services emerged, specializations developed related directly to the family or derived from the functions of the family (e.g., providing food, shelter, nurturance, clothing, and reproducing children and culture) (Bailey and Firebaugh 1986). Some specializations developed around certain human issues (e.g., nutrition, dependent care of the young and the elderly, housing). Other specializations focus on the application of knowledge in business, industry and agencies (e.g., restaurant management, interior design, marriage and family therapy). Certain specializations may or may not be found in family and consumer sciences academic units. At the turn of the twenty-first century, examples of the three conceptions of family and consumer sciences can still be found in academic units in the United States.

3. History

Family and Consumer Sciences evolved from home economics. Home economics emerged at the turn of the twentieth century in the context of the progressive reform movement. The organizers of the home economics movement were educated women and men who deplored the problems associated with urban crowding, child labor, malnutrition, immigration, and lack of education; but they also believed that the application of scientific knowledge could improve the daily lives of people and the policies of public agencies and employers (Stage 1997, Vincenti 1982). They organized courses to improve nutrition and the management of resources in public schools and neighborhood organizations in Boston and New York City. They launched campaigns to reform corrupt practices of landlords, employers, and businesses. They advocated for responsible 'municipal housekeeping' in which government would improve living conditions and citizens would be involved in governing bodies.

The educational innovations of the 1860s, such as the creation of land-grant universities through the Morrill Act of 1862 (whose express purpose was to educate the children of the common folk), offered

opportunities for higher education where they were not previously available. Home economics became a major point of entry for women into public higher education. The 'second Morrill Act' was passed in 1892, authorizing institutions of higher education for black citizens in the Southern States. The home economists who worked in the Extension Service and in the Southern schools and colleges worked in a segregated system with few resources (Harris 1997). Despite these constraints, their efforts greatly improved the daily lives of their constituents and offered career opportunities to those with formal education.

In 1899 began a series of meetings at the Lake Placid Club at Morningside, New York, for the expressed purpose of uniting efforts in the study of the home under one name and to organize a body of knowledge (see Brown 1985 and Lake Placid Conference Proceedings 1902–8). The proceedings record lively dialogue about nomenclature and content. Home economics was adopted as the name of the discipline. Ellen H. Swallow Richards, a chemist and the first woman graduate and first woman faculty member at the Massachusetts Institute of Technology, emerged as a zealous leader for academic programs, demonstration and outreach projects, research, and professionalization of the field (Hunt 1942, Stage 1997). Ellen H. Swallow Richards is generally identified as the 'founder' of the home economics movement.

The growing strength of home economics in terms of students enrolled, organizational structure, and public policy advocacy is documented in Pundt (1980). She describes the first decade after the founding of the American Home Economics Association in 1909 as 'continuing the struggle of the new discipline to define itself and find a rightful place among the more orthodox disciplines of a college or university …' (Pundt 1980, p. 25).

Ideas about the modernization of family roles and the application of science in the home took a central place in home economics during the first quarter of the century. Behavioral sciences greatly influenced home economics teaching about child development and parenting, and child rearing was integrated into the home economics curriculum (Grant 1997). Principles from industrial engineering were applied to work simplification and design in the kitchens of America. Home economists were called upon during the 1930s to help families meet the devastation of the Great Depression. An outcome of these efforts was that nutrition and dietetics achieved a secure place in home economics curriculum and outreach programs (Babbitt 1997).

Dramatic changes in social roles, attitudes, and organization at mid-century and beyond prompted a series of conferences—New Directions (1959) and New Directions II (1975), the French Lick Conference (1961), the Eleventh Lake Placid Conference (1973), and Home Economics Defined (1979)—as the field engaged in a 'search for identity' (Brown 1985). The focus of most of these conferences and 'white papers' was the purpose and subject matter central to the field (Simerly et al. 2000).

In 1975, a definition of home economics was published in New Directions II:

> The study of the reciprocal relations of family to its natural and man-made environments, the effects of these singly or in unison as they shape the internal functioning of families, and the interplay between the family and other social institutions and the physical environment (Bivens et al. 1975, pp. 26–27).

The Scottsdale Conference held in 1993 examined the mission, breadth, scope, and name of the field (Simerly et al. 2000, Stage and Vincenti 1997). The outcome of the meeting was the recommendation of the name Family and Consumer Sciences for the field, which was subsequently ratified by the sponsoring organizations. A systematic statement of the basic beliefs, assumptions, professional practice, and areas of focus was developed (The Conceptual Framework for the 21st Century 1994).

4. *Research Methods*

Research in family and consumer sciences covers the entire spectrum from empirical studies using survey research methods, experimental laboratory studies, historical analysis, qualitative research such as case studies and personal narrative, needs assessments, and evaluation studies. Both 'basic research' (i.e., research conducted for the sake of testing theory, or to study phenomena with no expectation of a utilitarian outcome) and 'applied research' (i.e., research directed toward the solution of specific problems) are found in family and consumer sciences (Touliatos and Compton 1988). Intervention projects based on experimental design are used to test the efficacy of programs aimed at behavioral change.

Certain research methods tend to be associated with specializations within family and consumer sciences. For example, nutrition, food science, and textile research tends to be conducted in laboratories using basic or experimental research methods. While animal models have long been used to study diet-related diseases, studies with human subjects are becoming more prevalent. Population studies based on the methods of epidemiology are employed in assessing the extent of dietary diseases.

Areas in textiles and apparel such as fashion merchandising and international trade adopt a variety of research methods ranging from historical analysis, psychometrics, and secondary data analysis. Textile scientists concerned with the environmental impacts of processing conduct field studies and operate pilot plants to investigate their research questions.

Demography is concerned with human population and changes over time, thus it plays a major role in

social sciences and history. Family demography has traditionally been used to study macrolevel family patterns, but as life-course events are increasingly matters of choice, the importance of both macro and microlevel understanding is recognized (Teachman et al. 1999). Demographic methods are used to study housing, family economics, marriage, and fertility.

Research in family studies relies heavily on cross-sectional surveys and quantitative analysis. Data for these studies may be collected through interviews or self-administered questionnaires. Longitudinal analysis based on panel studies of the same subjects provide insights about families and their behavior over time. Acock (1999) observes that advances in computer technology and software facilitate the analysis of quantitative data which was not possible as recently as the 1970s.

Observational studies either in clinical or laboratory settings or the 'naturalistic' setting of the home are used to study family interaction. Qualitative methods are increasingly accepted as legitimate for the study of family and consumer issues. Qualitative family research uses content analysis of in-depth interviews, documents and other artifacts, and case studies (Rosenblatt and Fischer 1993). Qualitative studies are especially useful in the study of meanings, perceptions, and other subjective phenomena.

The mission-oriented nature of family and consumer sciences creates a close affinity for action research. The primary objectives of action research are solving problems and improvement of practice, but increased knowledge is also a potential outcome (Touliatos and Compton 1988). Participatory action research incorporates the indigenous knowledge of the community. Valuing such knowledge is important in facilitating societal change.

5. *Outreach and Application*

The Hatch Act, passed in 1887, provided funds to the States to establish agricultural experiment stations at the land-grant universities. The goal was solving practical problems. The Smith–Lever Act was passed in 1914, creating the Cooperative Extension Service. This provided the institutional vehicle for disseminating the research results from the universities and encouraging citizens to adopt practices to improve their living standards (see Babbitt 1997, Firebaugh and Redmond 1996, Harris 1997). The Cooperative Extension Service was a unique creation of higher education in the United States. It continues to have the most access to the most people through its network of county-based education programs.

The tradition of taking knowledge to the people has been modified through the years and adopted by family and consumer sciences programs, and many related social and behavioral sciences, in all types of higher education institutions. Community–university collaboration in identifying and addressing the problems facing families and youth in the United States is creating effective programs and solutions to many of these problems (Chibucos and Lerner 1999) and increasing the value of outreach scholarship within higher education (Lerner and Simon 1998). Limitations of space preclude a comprehensive review of the multitude of models and programs in which family and consumer sciences is in partnership with other disciplines and with constituents in the application of knowledge to address quality of life issues of the twenty-first century.

6. *The Future*

Family and consumer sciences has evolved with growing sophistication in research methods, proliferation of specializations within the field, use of new technologies for dissemination of knowledge, and advocacy for policies to protect families and consumers. The field, however, has remained focused on its mission. Fidelity to its mission and responsiveness to change are expected in the future.

Application of knowledge to persistent practical problems has characterized family and consumer sciences. The list of accomplishments is vast: the school lunch program, delineation of requirements for nutrients, standardization of clothing sizes, best practices models in dependent care, Truth in Lending and Truth in Saving legislation, to name a few. New issues present themselves in the twenty-first century. Most of these issues are complex and require multiple perspectives. More collaborative projects across the disciplines and with community groups can be expected in the future. New guidelines in federal and foundation funding, designed to foster a closer integration of research and application, will encourage family and consumer sciences researchers and practitioners to engage in more collaborative outreach efforts.

Increased availability of information through the Internet and rising education levels of the population are expected to change the focus of professional practice from an orientation of being an 'expert' to being a 'facilitator.' The holistic family ecosystems theoretical framework and the integrative, synergistic orientation of family and consumer sciences will provide a comparative advantage for practitioners working in the 'information society.'

Deliberations about the nature of family and consumer sciences as a discipline are sure to continue. Conceptualizations about the body of knowledge (Baugher et al. 2000), discourse on the nature of the discipline (Brown and Paolucci 1979), and attempts to clarify and unify the discipline and the profession (Green 1996, Simerly et al. 2000) are echoes of the discussions at the Lake Placid Conferences. Such examination of family and consumer sciences will foster its future vitality and relevance as a field of study and a profession.

See also: Community Health; Community Sociology; Consumer Culture; Consumer Economics; Consumer Psychology; Consumption, Sociology of; Family and Kinship, History of; Family and Schooling; Family, Anthropology of; Family as Institution; Family Bargaining; Family Business; Family Health; Family Law; Family Processes; Life Course: Sociological Aspects; Religion: Family and Kinship; Segregation Indices

Bibliography

Acock A C 1999 Quantitative methodology for studying families. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*, 2nd edn. Plenum Press, New York, pp. 263–89

Babbitt K R 1997 Legitimizing nutrition education: The impact of the Great Depression. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 145–62

Bailey L, Firebaugh F M 1986 *Strengthening Home Economics Programs in Higher Education*. College of Home Economics, The Ohio State University, Columbus, OH

Baugher S, Anderson C I, Green K B, Nickols S Y, Shane J, Jolly L, Miles, J 2000 Body of knowledge of family and consumer sciences. *Journal of Family and Consumer Sciences* **92**(3): 29–32

Bivens G, Fitch M, Newkirk G, Paolucci B, Riggs E, St. Marie S, Vaughn G 1975 Home economics–New directions II. *Journal of Home Economics* **67**(3): 26–7

Brown M 1985 *Philosophical Studies of Home Economics in the United States: Our Practical-Intellectual Heritage, Vol. I*. Michigan State University, East Lansing, MI

Brown M, Paolucci B 1979 *Home Economics: A Definition*. American Home Economics Association, Washington, DC

Bubolz M M, Sontag M S 1993 Human ecology theory. In: Boss P G, Doherty W J, LaRossa R, Schumn W R, Steinmetz S K (eds.) *Sourcebook of Family Theories and Methods: A Contextual Approach*. Plenum Press, New York, pp. 419–48

Chibucos T R, Lerner R M 1999 *Serving Children and Families Through Community-University Partnerships: Success Stories*. Kluwer Academic, Norwell, MA

Council for Accreditation 1995 *Accreditation Documents for Undergraduate Programs in Family and Consumer Sciences, 1995 Edition*. American Association of Family and Consumer Sciences, Alexandria, VA

Elder G H, Jr. 1998 The life course and human development. In: Lerner R M (ed.) *Handbook of Child Psychology*. Wiley, New York, pp. 939–91

Firebaugh F M, Redmond M B 1996 Flora Rose: A leader, innovator, activist, and administrator. *Kappa Omicron Nu Forum* **9**(2): 7–16

Grant J 1997 Modernizing mothers: Home economics and the parent education movement, 1920–1945. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 55–74

Green K B 1996 Our intellectual ecology: A treatise on home economics. In: Simerly C, Light H, Mitstifer D I (eds.) *A Book of Readings: The Context for Professionals in Human, Family and Consumer Sciences*. American Association of Family and Consumer Sciences, Alexandria, VA, pp. 1–7

Harris C 1997 Grace under pressure: The Black Home Extension Service in South Carolina, 1919–1966. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 203–28

Hunt C L 1942 *The Life of Ellen H. Richards*. American Home Economics Association, Washington DC

Lake Placid Conference on Home Economics 1901 *Proceedings of the First, Second, and Third Conferences*. American Home Economics Association, Washington, DC

Lake Placid Conference on Home Economics 1902 *Proceedings of the Fourth Annual Conference*. American Home Economics Association, Washington, DC

Lake Placid Conference on Home Economics 1903 *Proceedings of the Fifth Annual Conference*. American Home Economics Association, Washington, DC

Lake Placid Conference on Home Economics 1904 *Proceedings of the Sixth Annual Conference*. American Home Economics Association, Washington, DC

Lake Placid Conference on Home Economics 1905 *Proceedings of the Seventh Annual Conference*. American Home Economics Association, Washington, DC

Lake Placid Conference on Home Economics 1906 *Proceedings of the Eighth Annual Conference*. American Home Economics Association, Washington, DC

Lake Placid Conference on Home Economics 1907 *Proceedings of the Ninth Annual Conference*. American Home Economics Association, Washington, DC

Lake Placid Conference on Home Economics 1908 *Proceedings of the Tenth Annual Conference*. American Home Economics Association, Washington, DC

Lee J A, Dressell P 1963 *Liberal education and home economics*. Bureau of Publications, Teachers College, Columbia University, New York

Lerner R M, Simon A 1998 *University-Community Collaborations for the Twenty-First Century: Outreach Scholarship for Youth and Families*. Garland, New York

Paolucci B, Hall O A, Axinn, N 1977 *Family Decision Making: An Ecosystems Approach*. Wiley, New York

Pundt H 1980 *AHEA: A History of Excellence*. American Home Economics Association, Washington, DC

Ralston P A 1996 Flemmie P. Kittrell: Her views and practices regarding home economics in higher education. In: Simerly C, Light H, Mitstifer D I (eds.) *A Book of Readings: The Context for Professionals in Human, Family and Consumer Sciences*. American Association of Family and Consumer Sciences, Alexandria, VA, pp. 181–90

Rosenblatt P C, Fischer L R 1993 Qualitative family research. In: Boss P G, Doherty W J, LaRossa R, Schumm W R, Steinmetz S K (eds.) *Sourcebook of Family Theories and Methods: A Contextual Approach*. Plenum Press, New York, pp. 167–77

Simerly C B, Ralston, P A, Harriman L, Taylor, B 2000 The Scottsdale initiative: Positioning the profession for the 21st century. *Journal of Family and Consumer Sciences* **92**(1): 75–80

Stage S 1997 Ellen Richards and the social significance of the home economics movement. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 17–33

Stage S, Vincenti V B 1997 *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY

Teachman J D, Polonko K A, Scanzoni J 1999 Demography and families. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*, 2nd edn. Plenum Press, New York, pp. 39–75

The conceptual framework for the 21st century 1994 *Journal of Home Economics* **86**(4): 38
Touliatos J, Compton N H 1988 *Research Methods in Human Ecology/Home Economics*. Iowa State University Press, Ames, IA
Vincenti V B 1982 Toward a clearer professional identity. *Journal of Home Economics* **74**(3): 20–25
Vincenti V B 1997 Chronology of events and movements which have defined and shaped home economics. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 321–30

S. Y. Nickols

Family and Gender

The relationship between gender and the family has been a contested issue in the social sciences. Much of this controversy arises because even to speak about gender is, deliberately or inadvertently, to raise the issue of equality between men and women. Moreover, the issue of equality between the genders, in turn, is intimately linked to the paradoxical treatment of the family in classical liberal political thought (Eisenstein 1981).

1. The Classical Liberal Legacy—Patriarchal Antipatriarchalism

Classical liberal theory is one of the constitutive elements of modern society. Liberal political ideas emerged against the background of the collapse of the feudal social order and the emergence of a society based on market relations. Liberalism simultaneously promoted the ideals of freedom and equality, and strict separation between a public and a private realm. However, classical liberalism's assertion that principles of equality and freedom should be limited to the public realm has led many later thinkers to see the transformation of power in private family households as the last remaining objective of an incomplete revolution.

The most celebrated exponent of classical liberalism, John Locke, developed his doctrine in opposition to the doctrine of patriarchalism—a Latin term meaning the rule of the father. The patriarchalist Sir Robert Filmer had justified the absolute rule of the monarch by analogy to the 'divine and natural' authority of fathers over their households. In Filmer's social world, no distinction was drawn between the public and the private capacity of the king. King James I, for example, was expected to finance his own government from the revenue generated by his own estates, crown lands, and customs.

Locke opposed Filmer precisely because Filmer conflated the political and the familial. In Locke's view each institution had a separate purpose. Family, as an institution dealing with the rearing of dependent children, was organized around the inequality of parent and child. In a market society, politics, by contrast, was concerned with the coordination of independent individuals; that is to say, it was organized around the principle of equality. Thus, Locke rejects patriarchal doctrine, acclaiming the principle of public equality between reasoning adults, while at the same time endorsing patriarchy as the appropriate logic of the private realm of the family.

Locke and his contemporaries excluded women from the public realm because they assumed that rationality was an exclusively male characteristic. Locke recognized women's role in procreation and the preparation of children for political life but assigned these activities to the private sphere. Furthermore, Locke believed that a 'wife's subordination to her husband had a "Foundation in Nature" and that the husband's will must prevail in the household as he is naturally "abler and the stronger"' (Pateman 1989, p. 121).

Classical liberalism's legacy, among other things, is a continuing tension between the antipatriarchal demand for equality associated with rule by consent in public life and the creation of a separate private sphere still based on patriarchal prerogatives. This tension has given rise to a demand for equality between the genders, even in the private sphere of the family. Indeed, the family has often been seen as a key institution and, in some cases, the very locus of patriarchal rule (Hartmann 1981). However, the demand for gender equality has different variants depending on different analyses of the significance of the family for the subordination of women.

2. Liberal Feminism—Escaping the Doll's House

Liberal feminists accept broadly the classical liberal idea that citizenship is based on participation in public institutions, especially the (labor) market. Furthermore, liberal feminists also to a great extent accept that what happens at home is a private matter to be negotiated between husband and wife. Consequently, theorists of this persuasion argue that the key to women's subordination is the restriction of women to the private sphere.

Liberal feminism urges women to be like Nora, the heroine of Henrik Ibsen's play *The Doll's House*, and throw off domesticity and participate fully in the labor market, competing with men for the economic and social rewards available through paid work. It assumes that the traditional domestic division of labor, where the man is the provider and the woman the homemaker, follows from the fact that the homemaker is more dependent economically on the provider. It is the

dependence of the housewife on her husband, and not her gender, that is responsible for the sexual division of labor. The implication of this doctrine is that once women achieve economic independence, gender equality in the home will follow automatically.

From the perspective of liberal feminism, women's legal rights to property, including property in their own person, is the first step in the emancipation of women. Liberal feminism broadly accepts the proposition that contemporary advanced industrial societies are meritocratic, and that women as a group are not innately less talented or less diligent than men but are denied opportunities simply because of their sex. It shares with neoclassical economics the belief that discrimination distorts the market mechanism and produces undesirable outcomes, while disagreeing with the neoclassical economists' presumption that discrimination is an infrequent occurrence. Liberal feminism views wage discrimination and sex segregation of occupations in the labor market as a mechanism designed to exclude women from the labor market. Liberal feminist doctrine lies behind much of the legislation on antidiscrimination and equal opportunity. Discrimination against women, claims liberal feminism, often arises from stereotyped expectations.

These gender stereotypes and prejudices are often learned in the family. Like structural functionalist theory in sociology, liberal feminism sees the family as an institution that specializes in socialization. Children learn gender roles in the family, using role models available to them that are 'reinforced' by the wider society. Characteristically, liberal feminism implies that the inequitable influence of gender roles can be reversed by simply giving girls better role models, establishing equal opportunity in education, and encouraging high achievement aspirations. The mother's role in the traditional domestic division of labor models poor achievement aspiration to daughters but it is not itself a source of constraint. The fundamental gender neutrality of the domestic division of labor, which liberal feminism presumes, can be justified on two distinct bases, either on the claim that the current organization of the family is an economically rational specialization of roles, or on the basis of the exchange theory of power.

Households, according to economist Gary Becker (1981), allocate their resources of time and money rationally between the spheres of market work, nonmarket work, and leisure. Individuals specialize in those tasks for which they have a 'comparative advantage'; a decision based on relative productivities. Becker claims that small biological differences, such as lactation, give women an advantage in child rearing. Men become more productive at market work because gender specialization leads them to invest in their capacity to earn (human capital), through education and on-the-job experience. This leads to a situation where men's wages are generally higher than women's, and any extra allocation of men's time to housework would oblige them to forgo a greater amount of income. By contrast, the 'shadow price' or opportunity cost of women's time spent in unpaid work at home is lower, because the cost of income forgone is smaller. This somewhat circular reasoning leads to the claim that men have a 'comparative advantage' in paid work, while women have a 'comparative advantage' in unpaid work. Because each member specializes in activities where their advantage (productivity) is greatest, and they pool and exchange the rewards of these separate activities, the household is said to benefit from 'gains of trade.' Assuming an equitable distribution of these gains, the household that specializes is collectively better off than the one that does not. The asymmetry in the allocation of domestic work in the family is seen as a freely chosen and rational response to the biological differences plus inequalities in the labor market. The contribution of discrimination to gender inequality in the labor market is de-emphasized, since neoclassical economic theory implies that discrimination should erode in competitive markets. The allocation of domestic tasks will become symmetrical if the barriers to women's economic emergence are removed, and their labor market situation becomes similar to that of men.

Sociological exchange theory does not require the assumption that everyone in the household strives to increase the household's joint welfare. On the contrary, it assumes that whoever brings the most valuable resources to a marriage can expect to be able to bargain for some kind of compensation in the form of release from menial activities. In the traditional family, the wife provides domestic labor in return for economic support. The fact that housework and child care are done by women follows, according to exchange theory, from her inability to contribute socially recognized resources, such as money, to the relationship. In other words, responsibility for housework and child care falls to her because she contributes significantly less money rather than because she is a woman. Improve the labor market position of women, argue liberal feminists, and the inequities in the domestic division of labor will disappear.

If liberal feminists allow for any 'stickiness' in patterns of family organization this is attributed to a process of lagged communication and the learning of new sex role expectations. Women must acquire the techniques of assertiveness that allow them to negotiate their relationships with their partners. Men must learn to accept new domestic responsibilities and become 'new fathers.'

3. *Other Forms of Feminism*

Other forms of feminism view the family as a key site in the determination of gender equality. Nonliberal

feminists have drawn attention to how the privatization of family life contributes to the subordination of women in two key respects. First, women's unpaid work around the home is private, unseen, and devalued. Moreover, the pattern of the distribution of resources is hidden from view. Second, by definition, privacy screens family relations from view, allowing domination, including physical forms of coercion, to displace rule by consent. In addition, modern family relations have been seen as the crucible in which gender identity is forged.

Since the 1940s it has been popular to view the family as a 'haven in a heartless world'—a secluded refuge from the ruthlessly impersonal world of market relations. However, according to its critics, this liberal view of the family as refuge relies on two propositions: the separation of home and work, and the idea that only market work is work.

Nonliberal feminists challenged the presumption of economics (before the development of Becker's new home economics)—that the family household is purely a site of consumption—pointing to the valuable stream of goods and services produced in the home. In place of the separation of home and work, feminists substituted the distinction between paid and unpaid work. In the traditional division of labor, the male provider returns from market work ('a hard day at the office') to seek comfort and rest at home. For his homemaker spouse, however, the home is the site of her work.

Just how nonmarket work (performed in private homes) became virtually invisible, even to social scientists, would make a fascinating study. It is possible, however, to make explicit some of the social conditions that make this illusion possible. In premodern times, households were largely self-sustaining (consuming mostly goods produced at home) and most work was nonmarket work. However, as soon as market exchange becomes the dominant principal of economic organization, most things start to be bought with money. A labor market develops where people buy and sell the capacity to work for a stipulated amount of money (a wage). Money becomes the universal measure of value. Since mothers and housewives receive no pay for their work, it is seen as valueless, as not actually work at all. By the middle of the twentieth century the idea of housework as an instrumental labor process had disappeared so thoroughly that most celebrated social theorists of the time thought that wives and mothers specialized in 'expressive' behavior (Parsons and Bales 1955).

In contrast to the gender-neutral theories of the allocation of unpaid work favored by liberal feminists, other varieties of feminism place great emphasis on the fact that unpaid work is allocated on the basis of gender, so that housework and child care continue to be 'women's work.' Women, not men, interrupt their careers to raise a family. Indeed, men with families are more devoted to careers than single men. Studies of the time devoted to unpaid work show that women still do an overwhelming proportion of laundry, cleaning, cooking, and the physical care of children. The same studies show that men continued to specialize in mowing the lawn and polishing the car, and their chief contact with children is in the context of play.

Even when both partners are in full-time employment the sex segregation in unpaid work tasks does not diminish. The idea that gender relations within the family make it difficult to reassign the sex-stereotyped domestic responsibilities lies behind the anxiety about women performing a 'second shift' (Hochschild and Machung 1989).

Although responsibility for key domestic tasks is allocated to women, all members of the family enjoy the benefits of domestic work—live in clean dwellings, eat home-cooked meals, wear laundered clothes, receive emotional nurturance, etc. Although society as a whole depends on the reproduction of acculturated individuals, the costs of the activity of mothering, such as loss of lifetime earnings, reduced autonomy, and greater exposure to poverty, predominantly are borne privately by mothers.

The recognition of these facts about the distribution of domestic labor brings under suspicion the key presumption of the 'rational' specialization thesis, namely joint family welfare. These suspicions are reinforced by research into the relationship between women's share of family income and consumption of market goods. If families work to maximize their common (unitary) utility and pooled resources, then their consumption will be determined by total family income alone, and not by who earns or controls income. However, there is mounting evidence that, independent of the amount of total family income, increasing women's share of family income changes family expenditure patterns. The increasing women's share of family income results in greater expenditure on goods and services preferred by women (women's and children's clothing, restaurant expenditures) (Lundberg et al. 1997). Taken together, this evidence suggests that neither the sexual division of labor nor the distribution of market and nonmarket goods and services is consensual.

This raises the prospect that the family as an institution is constituted by systematically unequal exchange between the genders. Indeed, significant numbers of feminists believe that men as a social group actively or passively resist the reallocation of responsibility for domestic tasks because, as beneficiaries of this unequal exchange, it is not in their interests to change (Delphy and Leonard 1992).

Recognition of resistance leads feminists to explore the basis of capacity to resist, or to say the same thing in different words, to seek out the sources of male power. While liberal feminists tend to see men's dominance of the public sphere as the key to their power, radical feminists have often sought the basis of male power in family relations.

A simple, and long neglected, form of male power is physical coercion. The rise of second-wave feminism toward the end of the 1960s was accompanied by a growing recognition of the shocking extent of domestic violence. The significance of this kind of coercion is probably greater than the recorded incidence of violent acts, because in most social systems it is the threat of violence (terror) rather than direct acts of physical coercion that intimidates people into acquiescence. Susan Brownmiller used a similar logic when she claimed that rape is 'a conscious process of intimidation by which all men keep all women in a state of fear' (1975, p. 15). Emphasizing the crime of rape focused attention on the fact, a legacy of classical liberal thinking, that in most jurisdictions rape within marriage was not a crime in the middle of the twentieth century.

Domestic violence, and rape within marriage in particular, is so unexpected (even shocking) because they contradict the notion that family relations are relations between consenting equals. Much of the debate around sexual harassment and date rape also centers on the issue of consent. In classical liberal terms, it has become natural for us to allow the standards of liberty and equality developed as principles applying only to the public sphere to spill over to the private sphere. Many writers argue that the contemporary family is based on the progressive democratization of intimate relations.

Most social scientists believe that few systems of power can rely on coercion alone, and that the routine exercise of power must be seen as somehow legitimate. Sociologists believe that power is most successful when it becomes domination, that is, when it operates with the apparent consent of the dominated. Among the ideas that can be used to legitimate domestic inequality, a few have received the most consistent theoretical attention. The first is to present it as consensual and beneficial. Opponents of the 'rational' specialization thesis argue that this doctrine serves as an apologia for patriarchalism at home. The second focus has been the devaluation of women's work. Devaluation generally consists of diminishing the perceived importance of something, failing to recognize its true worth. The extreme case of devaluation is complete invisibility—which was a notable feature of the treatment of household productive activities as 'expressive' rather than 'instrumental.' The third idea that has received attention is the association of domestic work with femininity and its obverse, the association of masculinity with market work. Gender identity, many claim, is a major support for patriachalism in the family and a significant obstacle to the achievement of gender equity.

The idea that the activities typically associated with women have been devalued has led to both attempts to revalue women's work, for example, through comparable worth campaigns, and attempts to reconsider the relationship between equality and difference. At one level this has involved positive emphasis on 'women's values' (sensitivity, caring, and nurture) in opposition to the male values of competition, self-centeredness, aggression, and violence.

The celebration of difference has often been accompanied by an intellectual exploration of the socially constructed nature of gender identity which has taken priority over an analysis of family relations. Theories about language, and more broadly discourse, have been used to argue that signs used to represent things are arbitrarily self-referential. Self-referential because the meaning or one word is defined in terms of its relationship to other words—for example, 'light' is the opposite of 'dark,' while 'many' is defined against the background of the concept of 'one.'

It follows from this approach that the most powerful thing a social analyst can do is to deconstruct the discursive practices through which the social world is portrayed. The political objects of these more recent postmodern varieties of feminism have been representations of women and men, their bodies and relations in the various media of cultural transmission. Feminist theory has moved away from a preoccupation with things to emphasis on the salience of words and other cultural symbols.

Of course, the method of deconstruction can and has also been applied to the texts of feminist theory itself. This has made it problematic to assume that a political movement, with common objectives, can be rooted in a single category called 'women.' There has been an emerging sensitivity to the idea of difference among women, especially between 'women of color' and 'white middle-class feminists.' Hooks (1984) has argued that, for black American women, the family has been a source of solidarity and resistance to white racism and therefore less oppressive to black women than white American women. This emphasis on difference has been combined with a hostility derived from postmodernism to grand narratives or 'totalizing' conceptions. Generally, this has resulted in a political agenda that celebrates multiple identities, plastic sexuality, and diversity of household types ranging beyond the currently known groupings of family household.

From this perspective, sexual identity is an arbitrary cultural or social product. Gender identity, however, is seen as among the elements that contribute, in the deepest sense, to the conception of what one is. In 1978 Nancy Chodorow drew on a variant of Freudian psychodynamic theory (object relations theory) to explain gender differentiation and the different personality traits of adult men and women. Chodorow argued that because women are currently responsible for the primary care of children, this situation creates deeply internalized, unconscious desires in sons and daughters that reproduce male dominance and female mothering. When women 'mother,' all children have an initial attachment to and identification with a female. Because boys have a carer of a different sex,

achieving gender identity encourages emotional separation from the mother. Individuating from a female carer encourages boys to define their masculine gender identity that emphasizes difference from and rejection of femaleness, identification with emotionally absent (non-nurturing) males. For girls, raised by a same-sex carer, gender identity and individuation are not as mutually reinforcing. Thus, males grow up with an internalized need to emphasize separation and with misogynistic tendencies to define themselves in opposition to whatever is female; females grow up with a desire for emotional connection, including mothering. Chodorow draws the heartening conclusion that if men were to 'mother,' and therefore become more like women, this would change the masculine psyche. Chodorow's proposal to create greater gender equity ultimately turns on reforming the nurturing relations in the family.

Interest in the formation of gender identity, especially masculinity, has increased since Chodorow wrote *The Reproduction of Mothering*, using a different framework for understanding the process which places more emphasis on the role of culture and less on family relations. The most prominent alternative approach is social constructionism. In social constructionist variants, 'gender is not fixed in advance of social interaction, but is constructed in interaction' (Connell 1995, p. 35). It encourages study of the construction of masculinity in everyday life, acknowledges the existence of significantly different masculinities, and emphasizes the contradictory and dynamic nature of gender. This framework presupposes that Western culture typically valorizes the masculine over the feminine. The sexual division of labor at home has been viewed as a scaffold for the construction of gender identities.

4. Conclusion

The relationship between families and gender continues to be the subject of dispute in the social sciences. The 'linguistic turn' in the study of gender relations has directed attention to the most general cultural determinants of the gender identity of family members. The economic significance of nonmarket work has directed fresh attention to the process of home production. The classical liberal values of freedom and equality have proven themselves to be irrepressible, and although the split between the public sphere and the private sphere has endured, it has proven difficult to limit the demands for equality purely to the public sphere.

See also: Dissolution of Family in Western Nations: Cultural Concerns; Divorce and Children's Social Development; Divorce and Gender; Divorce, Sociology of; Families as Educational Settings; Family and Kinship, History of; Family and Psychiatric Illness/Family Therapy; Family, Anthropology of; Family as Institution; Family Bargaining; Family Law; Family Processes; Family Systems and the Preferred Sex of Children; Family Theory: Feminist–Economist Critique; Family Theory: Role of Changing Values; Family Therapy, Clinical Psychology of; Fatherhood; Feminist Legal Theory; Feminist Theory: Psychoanalytic; Gender and Feminist Studies in Anthropology; Gender and Reproductive Health; Gender and the Law; Gender Ideology: Cross-cultural Aspects; Gender-related Development; Gender Role Stress and Health; Lone Mothers in Affluent Nations; Lone Mothers in Nations of the South; Love and Intimacy, Psychology of; Marital Interaction: Effects on Child Development; Marriage and the Dual-career Family: Cultural Concerns; Masculinities and Femininities; Men's Health; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Parenthood and Adult Psychological Developments; Parenting: Attitudes and Beliefs; Parenting in Ethnic Minority Families: United States; Partnership Formation and Dissolution in Western Societies; Repartnering and Stepchildren; Rural Industrialization in Developing Nations and Gender

Bibliography

Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Brownmiller S 1975 *Against Our Will: Men, Women and Rape*. Penguin, Harmondsworth, UK

Chodorow N 1978 *Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. University of California Press, Berkeley, CA

Connell R W 1995 *Masculinities*. Allen & Unwin, Sydney

Delphy C, Leonard D 1992 *Familiar Exploitation: A New Analysis of Marriage in Contemporary Western Societies*. Polity Press, Cambridge, UK

Eisenstein Z R 1981 *The Radical Future of Liberal Feminism*. Longman, New York

Hartmann H I 1981 The family as the locus of gender, class and political struggle: The example of housework. *Signs* **6**(3): 366–94

Hochschild A, Machung A 1989 *The Second Shift: Working Parents and the Revolution at Home*. Viking, New York

Hooks B 1984 *Feminist Theory: From Margin to Center*. South End Press, Boston

Lundberg S J, Pollak R A, Wales T J 1997 Do husbands and wives pool their resources? Evidence from the United Kingdom Child Benefit. *Journal of Human Resources* **32**(3): 463–80

Parsons T, Bales R F 1955 *Family, Socialization and Interaction Process*. Free Press, Glencoe, IL

Pateman C 1989 *The Disorder of Women: Democracy, Feminism and Political Theory*. Polity Press, Cambridge, UK

M. Bittman

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Family and Kinship, History of

This article discusses the emergence and development of the historical study of the family and interprets them in the larger context of social history. It addresses the following issues: the emergence of the field of family history; household structure; developmental perspectives; kinship; the family's interaction with the process of industrialization; family strategies; and future directions in the field of family history.

1. *Emergence of the Field*

Recent historical research on the family has revised some widely held myths about family life in the past, as well as generalizations about the impact of the grand processes of social change on the family and society. Family history has complex roots in both the historical demography of the early 1960s and the 'new social history' of the same period. Particularly in the United States, it has shared with the latter a commitment to reconstructing the family patterns of ordinary people and viewing them as actors in the process of change. Contemporary historians of the family have sought to reintroduce human experience into historical research and to emphasize the complexity of historical change (see Hareven 1971, 2000a, Tilly 1987). Hence, research in the history of the family explores previously neglected dimensions of human experience such as growing up, courting, getting married, bearing and rearing children, living in families, becoming old and dying. The formidable goal has been to understand the family in various contexts of change and at different points in historical time. In short, the historical study of the family represents an effort to understand the interrelationship between 'individual time,' 'family time,' and 'historical time' (Elder 1978).

Before systematic historical research on the family began, various social science disciplines had generated their own myths and grand theories about continuities and changes in family behavior in the past. Sociologists in particular argued that, in preindustrial societies in the west, the dominant household form had contained an extended family, often involving three co-resident generations, and that the 'modern' family—characterized by a nuclear household structure, family limitation, the spacing of children, and population mobility—was brought about by industrialization. Associated with these generalizations was also the popular myth that industrialization broke up traditional family patterns and community life. By providing a perspective on change over time, as well as on family behavior within specific social and cultural contexts in discrete time periods, historical research on the family has led to the rejection of these assumptions and to the resulting questioning of the role of industrialization as a major watershed in this field for American and European history (Wrigley 1972, Goode 1963).

Over the three-and-a-half decades of its existence, family history has moved from a limited view of the family as a static unit at one point in time to an examination of the family as a process over the life course of its members; from a study of discrete domestic structures to the investigation of the nuclear family's relations with the wider kinship group; and from a study of the family as a separate domestic unit to an examination of its interaction with the worlds of religion, work, education, correctional and welfare institutions, and with processes such as migration, industrialization, and urbanization (Vinovskis 1977, Hareven 1982).

As research in this field developed, new findings and approaches have led to the revision and augmentation of the pioneers' approaches. Efforts to explore decision-making processes within the family have led to an investigation of strategies and choices that individuals and family groups make. The life-course approach has added an important developmental dimension by focusing on age and cohort comparisons in ways that link individual and family development to historical events. Research also expanded chronologically to cover antiquity, and geographically to encompass southern Europe, eastern Europe, the Mediterranean, Japan, China, and Latin America, in addition to Western Europe, North America, and Japan. The cumulative impact of studies in the history of the family has been to revise simplistic views of both social change and family behavior (Reher 1997, Andorka 1994, Plakans and Wetherell 1999, Todorova 1997, Halpern et al. 1996, Hayami and Uchida 1972, Kuroso 1996, Lee and Campbell 1997, Lee and Feng 1999, Borges 1992, Kuznesof 1997).

The emergence of the history of the family as a special area of inquiry received its major impetus from the publication of Philippe Aries's *Centuries of Childhood* (1960 in French and 1962 in an English translation). Aries argued that childhood as we know it originated only in the early modern period and that its discovery was closely linked to the emergence of the 'modern' or conjugal family, in which parents' private relationships with their children were more important 'than the honor of a line, the integrity of an inheritance, or the age and permanence of a name' (Aries 1962, p. 393). By linking the 'discovery' of childhood to transformations in family and social structures as well as economic and demographic changes, Aries inspired a whole generation of scholars.

Along another line of research, historical demographers in France in the early 1960s provided family historians with a powerful 'new weapon' for the analysis of demographic processes. Louis Henry and Pierre Goubert had developed a family reconstitution technique in the 1950s. Using genealogies, and marriage, baptismal, and death records from parish registers, this method enabled demographers to recon-

struct aggregate patterns of fertility, nuptiality, and mortality for vast numbers of people and, in some instances, over several generations (Henry 1953).

From the 1960s on, historical demography and family history in France developed into two parallel but interrelated streams. One continued to concentrate on demographic analysis, along the lines of Henry and Goubert; the other, influenced by Aries, anthropology, and the French social history tradition, integrated demographic analyses with patterns of family and sexuality, linking community and social and cultural variables with *mentalité*, as exemplified in the work of Emmanuel LeRoy Ladurie (1978), André Burguière (1987), Jean-Louis Flandrin (1976), and Martine Segalen (1991), among others. Family reconstitution subsequently became a powerful tool in the hands of the Cambridge Group for the History of Population and Social Structure.

Established in 1964, the Cambridge Group adapted the family reconstitution method to English parish registers, while also pursuing analysis of seventeenth-century nominal household registers. Wrigley found in the decline in fertility in the seventeenth century and in its recovery in the eighteenth century, evidence that rural births and marriages responded to changing economic conditions. He concluded that the demographic transition involved a change 'from a system of control through social institution and custom to one in which the private choice of individual couples played a major part in governing the fertility rate' (Laslett 1977, Wrigley and Schofield 1981).

Such demographic analyses for France and England, and subsequently Sweden, Austria, Germany, and The Netherlands, revealed that, in the preindustrial period, people married later than had been generally assumed, couples practiced some form of family limitation and child spacing as early as the seventeenth century, households were predominantly nuclear rather than extended, and preindustrial populations experienced extensive geographic mobility (Laslett 1965, Goubert 1977). Late marriage served as a method of family limitation. It was also closely related to the expectation that a newly wed couple would establish a separate household and be self-supporting, while continuing to contribute to their families of origin. Linking late age at marriage to the nuclearity of the household, John Hajnal (1965) developed his model of the 'West European Marriage Pattern,' which served as the basic model for the analysis of West European families until it came under criticism. Hajnal claimed that marriage patterns in historic Europe could be differentiated with a line drawn approximately from St. Petersburg on the Baltic Sea to the Adriatic Sea. West of this line marriages tended to be relatively late (late 20s) for both sexes, and substantial portions of the population did not marry at all. East of this line, marriage for both sexes was relatively early (late teens) and marriage was nearly universal. He identified the 'western' pattern as the 'European' pattern and suggested that the 'eastern' pattern was shared by many developing countries (in the 1960s).

2. *Household Structure*

Using nominal census records, Laslett found evidence of continuity in nuclear household structure in England, at least since the sixteenth century. A landmark volume published by Laslett and Wall (1972), containing articles on household structure for various countries, including Japan, concluded that preindustrial households in Western Europe and North America affirmed these findings and dispelled the previously held assumption that industrialization brought about a nuclear family form. Although Laslett's early work emphasized the persistence of a nuclear 'family,' the major unit discussed was the household, not the family. A household is not identical with a family, since the domestic group may include nonrelatives as well. Nor is a 'family' restricted to the household, since extended family ties transcended the household unit (Goody 1972, Hareven 1982).

Subsequent studies showed that co-residence with extended kin tended to increase, not decrease, after the 'industrial revolution,' because of the need of newly arrived migrants to industrial cities to share housing space (Laslett and Wall 1972, Hayami and Uchida 1972, Anderson 1971). David Herlihy (1987) and Richard Smith (1979), respectively, traced a nuclear household structure in medieval Tuscany back to the twelfth century, as well as in England; and, in the medieval village of Montaillou, LeRoy Ladurie (1978) found a variety of forms of family structures in the *domus* (household), ranging from ordinary nuclear families to various complex configurations. Household structure varied over the life course of its members, in relation to the economy and migration.

Such early discoveries also led to the formation of new generations and stereotypes: by combining the findings about the nuclearity of the household with the Hajnal thesis on the Western marriage pattern, Laslett developed a typology of the 'western family,' which he characterized as having a simple nuclear household; a relatively late child-bearing age; a relatively narrow age gap between husband and wife, with a relatively high proportion of wives older than their husbands; and the presence in the household of 'life cycle' servants unrelated to the family with whom they were residing. By contrast, he characterized the South European and Mediterranean family type as having complex household structures and early marriage, in which nuclear households are formed by the breakup of extended ones, rather than through marriage (Laslett 1977). These typologies, along with the Hajnal thesis, later came under criticism. David Kertzer refuted the thesis that links early female marriage to joint household organization as a characteristic of the

'South European' family pattern. He showed variability in household structures in Italy, depending on types of landholding and sharecropping. Similarly, Haim Gerber, on the basis of his analysis of urban Turkish households in the seventeenth century, questioned stereotypes of the traditional 'Middle Eastern family' (Kertzer and Hogan 1989, Gerber 1989).

In New England, when interpreted in the context of economic and social institutions (such as landholding, inheritance, religion, community structure, and religion), demographic patterns served as the backbone of complex family and community analyses. In his analysis of household and family structure in colonial Andover, Massachusetts, Philip Greven revealed the power relations between fathers and sons and the limits of the sons' autonomy based on inheritance. He found that households were primarily restricted to members of the nuclear family and nonrelatives, but a 'complex web' of family connections reinforced by proximity of residence, often on the same land, permeated the community (Greven 1970). Demos, in his study of colonial Plymouth, Massachusetts, used family reconstitution, the analysis of nominal census records and a great variety of other sources to reconstruct family and community patterns (1970). He also found a predominance of nuclear households, but these differed considerably from contemporary ones. Families were larger then and included nonrelatives, such as servants, apprentices, and boarders and lodgers in the households. Age configurations were also different; in some families, the oldest child in the household would be an adult while the youngest was still at the breast.

A major critique of early studies of household structure was their reliance on 'snapshots' in the household schedules listed in the census at one point in time. On the basis of his reconstruction of peasant households in the villages of Heidenreichstein in northern Austria in 1763, Berkner (1972) provided evidence that household structure changed several times over the course of the family's life, as it moved through its cycle—from a stem family to a nuclear one, and later to a stem family again, in relation to inheritance. Households were not static. At the same time, Demos' reconstruction of the stages of the life cycle in Plymouth demonstrated dynamic movements of individuals within and across families and households, in relation to the religious and cultural prescriptions of Puritan society (Demos 1970).

The presence of 'life cycle' servants—young men and women in their teens who lived and served in other people's households in the transitional period between leaving home and marrying—which Laslett found in English villages, as well as the presence of other unrelated individuals in the household, proves the remarkable flexibility of households in the past. Nuclear households included nonrelatives such as servants, apprentices, boarders and lodgers. Similarly, in nineteenth-century American urban households, Modell and Hareven (1973) found a high proportion of boarders and lodgers, who provided important substitutes for children who had left home. Households expanded and contracted over the life course and in accordance with the family's needs.

3. *Developmental Perspectives: The Life Course*

The discovery of boarders and lodgers in nineteenth-century households and of the overall fluidity of household structure led historians of the family to search for a framework that would allow them to capture the movement of individuals through various household forms over their lives and the changes in the composition of the family and the household under various historical conditions. The family cycle model, which was used originally to identify changes in family structure, became unsatisfactory, because the stages of the family cycle were derived from contemporary American middle-class families, and were *a priori*. As Glen Elder explained, the family cycle identifies stages of parenthood rather than the more dynamic aspects of individual transitions into and out of various family roles (Elder 1978).

Dissatisfaction with the family cycle model led historians and sociologists to develop the life-course approach, which has introduced a dynamic dimension into the historical study of the family, and has moved analysis from a simplistic examination of stages of the family cycle to individuals' and families' timing of life transitions in relation to historical time (Hareven 1978, Elder 1978). According to the life-course paradigm, the pace and definition of the timing of life transitions, such as starting work, leaving home, and getting married, are determined by their social and cultural context. On the familial level, timing involves the synchronization of individual life transitions with collective family ones (Elder 1978). For example, using both demographic evidence and prescriptive literature, Modell, Funstenberg, and Herschberg found a major historical change in American society over the past century, from the 'erratic' timing of life transitions governed by collective family needs to more individualized timing in conformity with age norms (Modell et al. 1976).

Life-course research also illuminates the links between behavior and perception. While the actual timing of life transitions can be reconstructed from demographic records, its meaning to the individual and family members undergoing these transitions hinges on the examination of qualitative, subjective sources (Neugarten and Hagestadt 1976, Bourdelais 1993, Hareven and Masaoka 1988, Hareven 2000b). A life-course perspective has also made an important contribution to the study of kinship by directing attention to the changing configurations of kin with whom individuals 'travel' together over their lives. Such configurations are formed and re-formed; they

change in their composition, in their relationship to the individual and the nuclear family, and in relation to each other over the life course.

4. Kinship

During the second decade of the development in the field of family history, kinship ties outside the household received increasing attention. Rather than being isolated, nuclear households were embedded in extended kinship ties outside their confines. Members of the nuclear family were engaged in various forms of mutual assistance, collaboration, and rituals with extended kin. Even though aging parents did not generally reside in the same household with their married adult children, they lived in the vicinity, often on the same land (Greven 1970, Berkner 1972, Smith 1979, Sabean 1998).

For nineteenth-and twentieth-century urban populations, Anderson (1971) and Hareven (1982), respectively, have documented the central role of family members and more distant kin in organizing migration from rural areas to industrial cities, in facilitating settlement in urban communities, and in helping migrants adapt to new working and living conditions. In her study of late-nineteenth-and-twentieth-century textile workers in Manchester, New Hampshire, Hareven found that kinship networks in communities of origin were reinforced by the back-and-forth migration of individual members, and the transfer of resources. Following 'chain migration' routes, villagers who went to work in industrial communities spearheaded migration for other relatives by locating housing and jobs. Those who remained in the communities of origin often took care of aging parents and other relatives who stayed behind (Anderson 1971, Hareven 1982). Hareven found that despite the scale of the modern factory system, workers' kinship networks cushioned the adaptation of immigrant workers to new industrial working conditions. Kin acted as brokers between the workers and their industrial employers. Rather than disrupting kinship ties, migration often strengthened them and led to the development of new functions for kin in response to changing economic and employment conditions in response to the requirements of the industrial system (Hareven 1982).

5. The Family's Interaction with the Process of Industrialization

An examination of the family's interaction with the process of social and economic change enables us to understand more precisely what occurred internally within the family and how such changes were accomplished on a societal level. It provides new insights into the process of industrialization and urbanization, into how labor markets functioned, and how industrial work processes and labor relations were organized. Historical research on the family has led to the rejection of simplistic models of social change, especially of modernization theory. How the family both initiates change and adapts to it and how it translates the impact of larger structural changes to its own sphere are crucial questions for the understanding of the process of social change.

The important principle underlying these questions is a view of the family as an active agent. The family planned, initiated, or resisted change; it did not just respond passively. Historical research over the past two decades, highlighting the family's role as an active agent, has led to revising questions such as 'What was the impact of industrialization on the family?' to 'What was the family's impact on the industrial system?' (Hareven 1982, c.f. Goode 1963). The question has been further cast into a more complex form: Under what circumstances was the family more able to control its destiny and to affect the larger social processes, and under what circumstances did the family succumb to declining markets, changing modes of production, business cycles, and other external forces?

The family had a central role in charting the transitions to various modes of industrial production by following diverse routes: from a rural economy to a proto-industrial or industrial one, or from a rural economy directly to an industrial one. In all these cases, while following its own priorities, when possible, the family facilitated the advance of industrialization by releasing the labor force needed for the newly developing factories and by organizing migration to industrial centers (Smelser 1959, Tilly and Scott 1978, Quataert 1985). The family's interaction with the industrial system has received attention from studies of family production systems in artisanal and proto-industrial settings and the study of family and work in the context of large industrial enterprises.

Proto-industrialization was a household-based form of production in the countryside and in some urban areas in England, France, Belgium, Switzerland, and Austria, from the seventeenth to the nineteenth centuries that preceded the 'industrial revolution.' It was characterized by the production of goods in local cottages for a capitalist employer who sold the products in external markets. Rudolph Braun (who first identified these patterns in Switzerland) and other historians developed a model of demographic behavior they claimed was typical of 'proto-industrial' families. This consisted of earlier age at marriage, higher fertility, and delays in children's departure from the home, all of which were conducive to maximizing a family labor force. Marriage formed the key to the spread of the cottage industry, because it led to the formation of a new nuclear family unit, which was also a 'unit of labor' (Braun 1974, Mendels 1972, Medick 1976).

In the 1980s and 1990s, the demographic model of proto-industrialization has come under criticism. On the one hand, scholars have questioned whether demographic behavior associated with proto-industrialization had been sufficiently widespread to warrant such generalizations. They found that demographic behavior in response to new economic conditions 'differed according to the environment (i.e., land ownership, social structure, and inheritance) and the exact nature of the economic change which took place' (Gutmann and Leboutte 1984, p. 589). Regional variation, interaction with the local agricultural structure, as well as the type of cottage industry (whether weaving or spinning), made a difference (Mitterauer 1992).

Like research on the family in proto-industrial regimes, the study of family in the factory system has provided considerable evidence documenting the family's role as an active agent in its interaction with the process of industrialization. Such research effectively undermined the theory of 'social breakdown' that had haunted sociology and social history. According to this notion, migration from rural to urban centers had led to the uprooting of people from their traditional kinship networks. The pressures of industrial work and urban life caused a disintegration of the family unit, and adaptation to industrial life stripped migrants of their traditional culture (Parsons 1955, Wirth 1938).

Historians and sociologists alike have challenged the theory of social breakdown, by providing contrary evidence. Neil Smelser, in his pioneer study of the family in the early stages of industrialization in Britain, documented the recruitment of entire family groups as work units in the early textile factories. Fathers contracted for their children, collected their wages, and were responsible for their performance in the factory. Entire families depended on the factory as their employer; the factories, in turn, depended on the recruitment of family groups to maintain a continuous labor supply (Smelser 1959). Anderson documented a continuity of recruitment in family units and in the vital role of kin among workers in the textile industry in Lancashire, England, at least through the middle of the nineteenth century. The practice of family members working together in the factory was carried over into the United States, as the factory system developed there, and was still present among immigrant workers in the twentieth century. The very success of the early industrial system depended on the recruitment of labor along kinship lines (Anderson 1971, Hareven 1982). As mentioned above, Hareven's study of family and work in the Amoskeag Mills in Manchester, New Hampshire, has documented the active role of the family and the kinship group in relation to the industrial corporation. It showed that the family type most 'fit' to interact with the factory system was not an 'isolated' nuclear family, but rather one embedded in extended kinship ties (Hareven 1982).

The functions that kin fulfilled in their interaction with the factory system were not merely archaic carryovers from rural society but a selective use of pre-migration patterns of kin assistance patterns in response to the needs dictated by industrial conditions. The encounter of immigrant workers with the modern factory system led neither to the abandonment of pre-migration family traditions, nor to a rigid adherence to them. Rather, the workers adapted their customs and social organization to the new conditions they confronted. In doing so, they addressed the factory system on its own terms. The family selected those aspects of its pre-migration culture that were most useful in coping with industrial conditions and adapted them to new needs. These patterns of selectivity called into question the linear view of social change advanced by modernization theory. Family behavior did not modernize at the same pace as workers' conduct in the factory. The family was both a custodian of tradition and an agent of change (Hareven 1982).

6. *Family Strategies*

To better understand the family's role in these processes of social change, it is necessary to examine how families charted their strategies in relation to external opportunities and constraints. Family strategies involved both the decisions individuals or families made and the actual timing of such decisions in relation to opportunities and constraints. Strategies included, at times, calculated trade-offs in order to find employment, achieve solvency, buy a house, facilitate children's education or their occupational advancement, control or facilitate a child's marriage, save for the future and provide for times of illness, old age, and death. Strategies were part of a larger life plan. As people encountered new circumstances, they modified and reshaped their plans and strategies in the context of their own culture and traditions (Hareven 1990).

As research on family strategies developed further, the concept also underwent important revisions. Under the influence of feminist historians, the question was raised: 'Whose strategies are family strategies?' This question is still part of an important research agenda about the process of decision making within the family, particularly concerning women's participation. Recent research suggests, though, that there may have been a considerable variation along gender-lines and among family members in the forging of strategies.

7. *Future Directions*

The multidimensional research effort in family history has produced impressive results and identified new

lines of investigation. Research still needs to pursue established topics that have not received sufficient attention, including the family's relationship to social space and a more systematic study of the family's relationship to religion, the state, and the legal system.

There is a need for more systematic studies of the 'edges' of family life, such as solitary individuals, orphans and foundlings, and the process of family breakdown through divorce or death and of family reconstitution through remarriage. A second direction is the forging of more systematic linkages between demographic behavior and household structure with internal family dynamics; a closer integration of the study of the household with nonresident kin; a more careful linkage of kinship and household patterns with various processes such as work and migration; and the relationship between demographic and family patterns with cultural dimensions.

In addition to this specific agenda, the larger questions of historical changes in the family over time still need to be explored systematically in several directions: First, if industrialization did not cause the emergence of 'modern' family behavior in Western Europe and the United States, what did cause it? One plausible explanation that has been advanced is the commercial revolution of the sixteenth and seventeenth centuries. This does not explain, however, the existence of nuclear household patterns in the Middle Ages. While Stone (1977), like Aries, dated the emergence of the 'modern family' in the late seventeenth and early eighteenth centuries, Shorter placed it in the late eighteenth, and early nineteenth centuries. Carl Degler (1980) dated the emergence of the American modern family in the late eighteenth or early nineteenth centuries as well. Following in Aries's footsteps, Stone, Shorter (1976) and Degler have focused on the rise of 'affective individualism' as the major criterion of 'modern' family life. They generally agreed that the 'modern family' is privatized, nuclear, domestic, and based on emotional bonding between husband and wife, and between parents and children. The correlates of its emergence were the weakening influence of extended kin, friends, and neighbors on family ties and an isolation of the family from interaction with the community. There is some disagreement and, at times, lack of clarity about which class initiated these changes. Aries, Stone, and, more implicitly, Degler viewed the bourgeoisie and the gentry as the vanguard, while Shorter assigned a crucial role to peasants and workers.

Related to the emergence of the family as a private, intimate entity is the widely accepted interpretation that this transformation was based on the family's surrender of earlier-held functions of production, welfare, education, and social control to other institutions and its withdrawal into domestic privacy has become one of the standard clichés of family history and sociology. We need to understand the process by which these functions were transferred to other institutions. A systematic exploration of the family's relationship to public agencies and institutions—specifically institutions of education, welfare, and social control, after the transfer of its functions—will help historians escape the trap of viewing the family as an isolated domestic entity.

Historical studies in Europe in the 1990s have emphasized an integrated view of the family both as a private entity and an object of the state. *The History of Private Life*, under the general editorship of Philippe Aries and Georges Duby, which focuses primarily on French society, deserves particular attention. The fourth volume, *From the Fires of Revolution to the Great War*, edited by Michelle Perrot, presents the experiential history of the family in relation to public life. The authors view the family's private life and sexuality as inseparable from the state (Perrot 1990).

As further research continues to emerge, it becomes necessary to develop a comprehensive model of change in family behavior that does justice to its complexity. The most important dimension still absent from studies of long-term changes in the family involves more systematic distinctions by social classes, rather than generalizing for the entire society on the basis of the middle class. We need a more detailed understanding of the historical process by which other classes adopted middle-class family behavior, if indeed that was the case, and of what class differences have survived such a process. How did patterns of family behavior that first emerged in the middle class transfer to other classes in society and by what processes? Equally important, at the end of that process, had rural and working-class families really adopted middle-class family forms, or did differences persist underneath apparent similarities?

The question about when and how change takes place needs to be raised again. Various strands of change in family history still need to be traced individually to the point of their culmination as visible transformations. The process is rendered more difficult by the fact that not all the strands undergo change at the same pace. Since the family is not a monolith, different members within it may initiate change or accept change at different points. For example, women were agents of change in certain areas of family life, such as in the practice of family limitation, while men were innovators in others. Even children were innovators in certain areas. They may have brought literacy into the family, new perceptions of behavior they learned in school or new work habits and new technologies. Thus, when we examine the family's role in the process of social change, we need to differentiate among the respective roles of members within the family, in interacting with these processes. This is an important task for future decades.

See also: Divorce and Gender; Family as Institution; Gender History; Industrialization, Typologies and History of; Kinship in Anthropology; Labor History;

Labor, Division of; Life Course in History; Sociology, Epistemology of; Work: Anthropological Aspects; Work, History of

Bibliography

Anderson M S 1971 *Family Structure in Nineteenth-century Lancashire*. Cambridge University Press, Cambridge, UK

Andorka R 1994 The historical demography of a proper Hungarian village: At any in the eighteenth and nineteenth centuries. *Journal of Family History* **19**: 311–31

Aries P 1962 *Centuries of Childhood: A Social History of Family Life*. Baldick R (trans.). Knopf, New York

Berkner L K 1972 The stem family and the developmental cycle of the peasant household: An eighteenth century Austrian example. *American Historical Review* **77**: 398–418

Borges D E 1992 *The Family in Bahia, Brazil, 1870–1945*. Stanford University Press, Stanford, CA

Bourdelais P 1993 *La Nouvelle Age de la vieillesse: Histoire du vieillessment de la Population*. O. Jacob, Paris

Braun R 1974 Early industrialization and demographic change in the Canton of Zurich. In: Tilly C (ed.) *Historical Studies of Changing Fertility: Quantitative Studies in History*. Princeton University Press, Princeton, IN

Burguiére A 1987 The formation of the couple. *Journal of Family History* **12**: 39–56

Degler C N 1980 *At Odds: Women and the Family in America from the Revolution to the Present*. Oxford University Press, New York

Demos J 1970 *A Little Commonwealth: Family Life in Plymouth Colony*. Oxford University Press, New York

Elder G H Jr 1978 Family history and the life course. In: Hareven T K (ed.) *Transitions: The Family and the Life Course in Historical Perspective*. Academic Press, New York

Flandrin J-L 1976 *Families in Former Times: Kinship, Household, and Sexuality*. Cambridge University Press, Cambridge, UK

Gerber H 1989 Anthropology and family history: The Ottoman and Turkish families. *Journal of Family History* **14**: 409–21

Goode W 1963 *World Revolution and Family Patterns*. Free Press, New York

Goody J 1972 Evolution of the family. In: Laslett P, Wall R (eds.) *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK

Goubert P 1977 Family and province: A contribution to the knowledge of family structures in early modern France. *Journal of Family History* **2**: 223–36

Greven P 1970 *Four Generations: Population, Land, and Family in Colonial Andover, Massachusetts*. Cornell University Press, Ithaca, NY

Gutmann M, Leboutte R 1984 Rethinking protoindustrialization and family. *Journal of Interdisciplinary History* **14**: 587–607

Hajnal J 1965 European marriage patterns in perspective. In: Glass D V, Eversley D E C (eds.) *Population in History*. E Arnold, London

Halpern J M, Kaser K, Wagner R A 1996 Patriarchy in the Balkans: Temporal and cross-cultural approaches. *History of the Family: An International Quarterly* **1**: 425–43

Hareven T K 1971 The history of the family as an interdisciplinary field. *Journal of Interdisciplinary History* **2**: 273–99

Hareven T K (ed.) 1978 *Transitions: The Family and the Life Course in Historical Perspective*. Academic Press, New York

Hareven T K 1982 *Family Time and Industrial Time: The Relationship Between the Family and Work in a New England Industrial Community*. Cambridge University Press, New York

Hareven T K 1990 A complex relationship: Family strategies and the process of economic and social change. In: Friedland R, Robertson A F (eds.) *Beyond the Marketplace: Rethinking Economy and Society*. Aldine de Gruyter, New York

Hareven T K 2000a *Families, History, and Social Change: Life-Course & Cross-Cultural Perspectives*. Westview, Press, Boulder, CO

Hareven T K 2000b Formal timing and subjective timing: Historical changes and the cultural construction of the life course. In: Bideau A, Bourdelais P, Brunet G (eds.) *Structures Par Âge et Âges de la Vie: Deux Visions du Monde?* Cahier de Annales de Demographie Historique, Paris

Hareven T K, Masaoka K 1988 Turning points and transitions: Perceptions of the life course. *Journal of Family History* **13**: 271–89

Hayami A, Uchida N 1972 Size of household in a Japanese country throughout the Tokugawa era. In: Laslett P, Wall R (eds.) *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK

Henry L 1953 Une richesse démographique en fiche: Les registres paroissiaux. *Population* **8**: 281–90

Herlihy D 1987 The family and religious ideologies in medieval Europe. *Journal of Family History* **12**: 3–18

Kertzer D I, Hogan D 1989 *Family, Political Economy, and Demographic Change: The Transformation of Life in Casalecchi, Italy, 1961–1921. Life Course Studies*. University of Wisconsin Press, Madison, WI

Kurosu S 1996 Leaving home in a stem family system: Departures of heir and non-heirs in pre-industrial Japan. *History of the Family* **1**: 329–52

Kuznesof E A 1997 Who were the families of 'natural' children in nineteenth-century Rio de Janeiro, Brazil? A comparison of baptismal and census records. *The History of the Family: An International Quarterly* **2**: 171–82

Laslett P 1965 *The World We Have Lost*. Methuen, London

Laslett P (ed.) 1977 *Family Life and Illicit Love in Earlier Generations: Essays in Historical Sociology*. Cambridge University Press, Cambridge, UK

Laslett P, Wall R (eds.) 1972 *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK

Lee J, Campbell C 1997 *Fate and Fortune in Rural China: Social Organization and Population Behavior in Lioning 1774–1873*. Cambridge University Press, Cambridge, UK

Lee J, Feng W 1999 *One Quarter of Humanity: Malthusian Mythology and Chinese Realities, 1700–2000*. Harvard University Press, Cambridge, MA

LeRoy Ladurie E 1978 *Montaillou: The Promised Land of Error*. Bray B (trans.). Braziller, New York

Medick H 1976 The proto-industrial family economy. *Social History* **3**: 291–315

Mendels F 1972 Proto-industrialization: The first phase of the industrialization process. *Journal of Economic History* **32**: 241–61

Mitterauer M 1992 Peasant and non-peasant family forms in relation to the physical environment and the local economy. *Journal of Family History* **17**: 139–59

Modell J, Furstenberg F, Hershberg T 1976 Social change and transitions to adulthood in historical perspective. *Journal of Family History* **1**(1): 7–32

Modell J, Hareven T K 1973 Urbanization and the malleable household: Boarding and lodging in American families. *Journal of Marriage and the Family* **35**: 467–79

Neugarten B, Hagestadt G 1976 Age and the life course. In: Binstock R, Shanas E (eds.) *Handbook of Aging and the Social Sciences*. Van Nostrand Reinhold, New York
Parsons T 1955 The American family: Its relation to personality and to the social structure. In: Parsons R, Bales R (eds.) *Family, Socialization, and Interaction Process*. Free Press, Glencoe, IL
Perrot M (ed.) 1990 *A History of Private Lives*. Vol. 4, *From the Fires of Revolution to the Great War*. Goldhammer A (trans.). Beklnap Press of Harvard University Press, Cambridge, London
Plakans A, Wetherell C 1999 Fertility and cultural diversity: An exploratory essay concerning the borderlands problem in the nineteenth century Baltic area. *Continuity and Change* **1**: 33–56
Quataert J 1985 Combining agrarian and industrial livelihood: Rural households in the Saxon Oberlausitz in the nineteenth century. *Journal of Family History* **10**: 145–62
Reher D S 1997 *Perspectives on the Family in Spain, Past and Present*. Oxford University Press, Oxford, UK
Rogers J 1993 Nordic family history: Themes and issues, old and new. *Journal of Family History* **18**: 291–314
Sabean D W 1998 *Kinship in Neckarhausen, 1700–1870*. Cambridge University Press, Cambridge, UK
Segalen M 1991 *Fifteen Generations of Bretons: Kinship and Society in Lower Brittany, 1720–1980*. Cambridge University Press, New York
Shorter E 1976 *The Making of the Modern Family*. Basic Books, New York
Smelser N 1959 *Social Change in the Industrial Revolution: An Application of Theory to the British Cotton Industry*. University of Chicago Press, Chicago
Smith R 1979 Kin and neighbors in a thirteenth-century Suffolk community. *Journal of Family History* **4**: 219–56
Stone L 1977 *The Family, Sex, and Marriage in England, 500–1800*. Harper and Row, New York
Tilly C 1987 Family history, social history, and social change. *Journal of Family History* **12**: 319–30
Tilly L, Scott J 1978 *Women, Work, and Family*. Holt, Rinehart and Winston, New York
Todorova M 1997 *Imagining the Balkans*. Oxford University Press, New York
Vinovskis M 1977 From household size to the life course: Some observations on recent trends in family history. *American Behavioral Scientist* **21**: 263–87
Wirth L 1938 Urbanism as a way of life. *American Journal of Sociology* **44**: 1–24
Wrigley E 1972 The process of modernization and the industrial revolution in England. *Journal of Interdisciplinary History* **3**: 225–60
Wrigley E, Schofield R S 1981 *The Population History of England, 1541–1871*. Harvard University Press, Cambridge, MA

T. K. Hareven

Family and Psychiatric Illness/Family Therapy

The term family therapy refers to the involvement of family members in the treatment of the patient. One participating member, as in the case of a spouse or partner, is sufficient to justify the use of the term. Family members included in the therapy may share the same household with the patient or live apart, and may be seen by therapists with or without the patient. Family therapy was introduced as a means of altering family relationships that were believed to contribute to the development of the patient's illness. Recently, interventions have been mounted to help relatives cope better with long lasting illnesses without any implication that they have played a part in their causation.

1. *Theories of Family Causation of Schizophrenia*

Schizophrenia is one of the most severe mental illnesses and although relatively uncommon, affecting 1 percent of the population, can cause lifetime disability. Its causes are still unknown, although there is strong evidence for a contribution from inheritance (see *Mental Illness*, *Genetics of*). This evidence was scanty in the 1940s when psychoanalysts began to identify features of parental behavior that they linked to the development of schizophrenia. One of the earliest papers was by Hajdu-Gaines (1940), based on his psychoanalysis of four women with schizophrenia. He characterized their mothers as cold, rigorous, and sadistic and their fathers as soft, indifferent, and passive. Fromm-Reichman (1948), when discussing the problems patients with schizophrenia pose in psychoanalysis, coined the term 'schizophrenogenic mother,' which became common currency among psychiatric professionals. Coming from a similar background, Lidz et al. (1957) studied 14 families with a member suffering from schizophrenia and concluded that the parents' marriages were either unbalanced in terms of power ('marital skew') or exhibited an emotional distance between the partners ('marital schism'). Bateson et al. (1956) developed a theory that the irrational thinking of people with schizophrenia was induced by exposure to contradictory messages from their parents. They worked on the theory and came up with the term 'the double bind' that was an early formulation of catch 22. The double bind was defined carefully as a contradiction within the verbal content of a message, or between the words used and the feelings expressed. In addition, the recipient of the message is not allowed to comment on the contradiction (the doubling of the bind). Like Fromm-Reichman's term, double bind rapidly infiltrated the psychiatric vocabulary, and indeed went beyond into usage by the lay public. The basis of its popular appeal was in fact its greatest weakness as an explanation for the origin of schizophrenia, namely the fact that it was recognized immediately as occurring in everyday interchanges.

All the foregoing theorists were American, but Britain produced its own charismatic figure, R. D. Laing, who achieved popular fame during his lifetime. His formulation of schizophrenia was that it was a

completely understandable response to the family's efforts to drive one of its members mad (Laing and Esterson 1964). He believed the family selected one of its number as a scapegoat to bear the brunt of the family's emotional tangles. His ideas also spread far beyond his colleagues to be taken up by young people in particular, who found the role of being their parents' victim immensely appealing. A film based on his theory, called *Family Life*, was a popular success.

All four theories described above are flawed for the same reason: they were developed from clinical experience with highly selected groups of patients, and were not subjected to scientific tests of their validity by their originators. This would have entailed developing reliable measures of the phenomenon described, e.g., cold, rejecting mothers; double bind messages; and comparing their occurrence in representative samples of families of people with schizophrenia and in control families. One pair of theorists did just this: Lyman Wynne and Margaret Singer. Like Bateson's group, they focused on distorted communication in the family, only they developed an extensive schema for rating varieties of what they termed 'communication deviance.' Using this method to rate transcripts of subjects' speech they were able to distinguish parents of people with schizophrenia from parents of people with other psychiatric disorders and from normal subjects with uncanny accuracy (Singer and Wynne 1966). The results were so striking that they demanded replication and Hirsch and Leff (1971) attempted to do this using a British sample. They learned to make ratings of communication deviance reliably and assessed transcripts of parents of patients with schizophrenia and of parents of patients with neuroses. The former scored significantly higher than the latter but there was substantial overlap in the scores of the two groups. Thus, the parents of people with schizophrenia do seem to have an increased tendency to communicate in an unclear fashion, but this finding by no means proves that this lack of clarity induces schizophrenia in their offspring. An equally plausible explanation is that deviant communication in the parents is a manifestation of the genes they share with the ill family member. As a consequence of shared inheritance, 10 percent of parents of people with schizophrenia develop the illness themselves.

The theories presented address different parental qualities: their emotional responsiveness, the balance of power in their relationship, their ability to communicate. Each group of theorists has tended to formulate schizophrenia in terms of the particular abnormality they have focused on in the parents, rather than starting with a generally agreed concept of the nature of the illness. For example, Bateson's group conceptualized schizophrenia as a disturbance of logical thinking, while Laing saw it as an attempt to escape from the role of victim. In fact schizophrenia produces pervasive disturbances in the entire mental life of the sufferer: thinking, feeling, perception, and communication may all be affected. None of the theories accounts for all the varied manifestations of this puzzling condition.

2. *Family Therapy*

2.1 *The Introduction of Family Therapy*

Family therapy developed initially in relation to schizophrenia, fuelled by the widely held view among psychiatric professionals that parents induced the illness in their offspring. Experienced psychoanalysts brought their expertise to bear on families caring for a person with schizophrenia. This in itself was an unconventional use of psychoanalysis, which was traditionally centered on the relationship between the analyst and a single client. Although scientific evaluations of this adaptation of psychoanalysis are almost nonexistent, the pioneers realized after some years that they were making little impact on the illness. Rubinstein (1974), one of the most respected therapists in the field, wrote that 'One of our earliest preconceptions, which proved to be unwarranted with the passage of time and our increase in experience, was that schizophrenic behavior could easily be modified if we treated the family conjointly, and if we helped the members sort out some of their distorted interpersonal dynamics.' The disillusionment of therapists with this work can be documented by calculating the decline over time in the number of papers on schizophrenia in one of the premier journals dedicated to family therapy. In the first volume of *Family Process*, published in 1962, exactly half the papers were concerned with schizophrenia. By 1975, the proportion of papers dealing with this condition had dropped to 10 percent, and by 1990 to 6 percent.

While psychoanalytic therapists turned their attention to families with other types of problems, a few charismatic individuals continued the work with schizophrenia, developing their own approaches and techniques, which often departed radically from the tenets of psychoanalysis. Therapists who are convinced of the value of their work rarely submit their efforts to scientific evaluation and no rigorous studies have been conducted of these individualistic therapists.

2.2 *Shift from Schizophrenia to Other Diagnostic Groups*

The focus of family therapy shifted to disorders of children, such as bedwetting, tantrums, and school refusal. These disturbances proved to be much more responsive to a family approach than schizophrenia, partly because there is not likely to be a biological basis to these behavioral problems, and partly because

children are much more under the influence of their parents than young adults. This point is underscored by the results of family therapy for eating disorders. This has beneficial results when the patient is a child or adolescent, but is less effective for adult patients, who do better with individual therapy (Russell et al. 1987).

2.3 The Influence of Systems Theory

While psychodynamic approaches to family work continue to have their adherents, they have been eclipsed by the rise of systemic family therapy. This was based on systems theory as enunciated by Bertalanffy (1968). The family is conceptualized as a system that attempts to maintain its integrity by responding to change in one part of the system with compensatory changes in other parts. Homeostasis, then, is the overriding concern of the family system (Jackson 1957). When a child develops a disturbance such as bedwetting, this is seen as compensating for a disturbance between other members of the family, say a conflictual relationship between the parents. When the parents bring the child for treatment, the child's symptoms are taken to signify dysfunction in the family system as a whole. For this reason the individual presenting for treatment is often termed the 'designated patient,' with the implication that this person has been selected by the family (unconsciously) as a messenger to the health professional that the whole family is in need of help. There is an echo here of Laing's view of the person with schizophrenia as the scapegoat for the rest of the family.

2.4 Family Therapy Approaches

Structural family therapy was developed by Minuchin and co-workers and entailed a basic assumption about the way in which a normative family should function (Minuchin 1974). The therapist takes an active role in trying to mold the family as closely as possible to this ideal image. As it involved a view of the father as the head of the household, it was criticized as gender biased and ethnocentric. In response to this, structural therapists now request families to state how they would like to function, and then help them to achieve the necessary changes, thus incorporating cultural sensitivity.

Strategic therapy, by contrast with systemic therapy, takes the presenting symptom at face value and involves a narrow focus and a time-limited intervention. An important innovation of strategic therapists is 'reframing,' in which the problem as presented by the family is reformulated, often changing its negative connotations into positive ones. This can free the family from established patterns of behavior and enable them to respond to the 'problematic' member in a different way.

Systemic family therapy received a boost from a group working in Italy who became known as the Milan School. The key therapists were Selvini Palazzoli, Boscolo, and Cecchin. They introduced innovative techniques such as paradoxical injunctions, and focused on patients with eating disorders and with schizophrenia (Selvini Palazzoli et al. 1978). They developed an interviewing technique of circular questioning, which has become recognized by family therapists as a useful tool. An example would be the therapist asking one member to describe the relationship between two other family members. This will produce responses from all concerned and will prompt the therapist to formulate the next question. This process makes evident to the family the circularity of interaction, as hypothesized by systems theory. The fame of the Milan therapists was such that clients came to them from all over Italy. This made follow-up very difficult and no convincing outcome research has been conducted on their patients.

2.5 Research on Family Therapy

In fact systemic formulations about the family's contribution to mental illness are difficult, though not impossible, to test scientifically. A number of instruments have been devised to measure the ways in which family members function when together, but these are not entirely satisfactory in grappling with the problem, since when several people are interacting they produce a vast amount of data. As an example, Mishler and Waxler (1967) recorded interactions between two parents and a single offspring. The recording, transcribing, coding, and scoring of the interviews required 129 hours of staff time for every hour of interview. For this reason greater success has been achieved by focusing on specific aspects of family interactions. The introduction of videotaping of family sessions, allowing replay, pausing, and slow motion, facilitated the study of interactions in great detail. One of the most productive lines of research has developed from the use of a measure of relatives' emotional response to a sick family member, known by the term Expressed Emotion (EE).

3. Relatives Expressed Emotion

This assessment instrument was introduced to investigate one of the consequences of deinstitutionalization (e.g., see *Deinstitutionalization, Psychiatry of*). In the UK it was found that patients discharged from psychiatric hospitals to live with close relatives were more likely to be readmitted than those who went to live with more distant kin or with landlords. George Brown, a sociologist, had a hunch that it was to do with the emotional relationship between patients and their close relatives, and with Michael Rutter, a child

psychiatrist, developed a technique to measure these. It is based on a semistructured interview with the relative enquiring about the patient's behavior and symptoms in the previous three months. The interview is audiotaped and ratings are made later from the tape. These depend partly on what the relative says and partly on the way their voice is used to convey emotion. The key ratings are of critical comments, reflecting how angry the relative is with the patient, hostility, representing rejection of the patient, overinvolvement, comprising overprotection and overidentification, and warmth.

3.1 Relatives EE and Schizophrenia

The initial studies were of patients with schizophrenia, since this was the most common diagnosis in patients discharged from long-term hospital care (Leff and Vaughn 1985). It was found that relatives' negative emotions—critical comments, hostility, and over involvement—were each associated with a greater likelihood of the patients to relapse. Consequently, they were combined into an index of Expressed Emotion (EE) and relatives were classified as high or low EE. In most parental households both parents were congruent in being either high or low EE. When they differed in their ratings, the highest rating was taken as characterizing the household. By contrast with the ratings of negative emotion, warmth was associated with a better outcome for the patients. Thus, it appeared that the family members could influence the course of schizophrenia either in a positive or a negative way.

The early studies of EE aroused a great deal of interest since they demonstrated that family members' emotional relationship with the patient could be measured reliably, and furthermore that these relationships appeared to determine the outcome of the patient's illness. It should be emphasized that, unlike the theories of causality reviewed above, it was not claimed that high levels of EE antedated the appearance of schizophrenia or played any part in its etiology. If there was any causal influence, it was on the subsequent course of the illness once it was established. Many groups have replicated the link between EE and the outcome of schizophrenia, with the result that it is now one of the best evidenced associations in psychiatry (Butzlaff and Hooley 1998). These replications have been carried out in a number of languages and cultures, including India, China, and Japan, showing that the association is independent of variations in family structure and function.

3.2 Family Interventions for Schizophrenia

The research on relatives' emotional responses and schizophrenia led to a revival of interest in working with the families of such patients. However the approach was quite different from that of the earlier generation of therapists. Instead of viewing the family as pathogenic and in need of treatment, the new attitude was one of respect for the carers and a genuine desire to work with them as allies in the struggle against the illness. This in itself was revolutionary since for decades family members had been subject to vilification and rejection by mental health professionals. Instead of the patients being scapegoated by their families, as Laing had claimed, it was the relatives who were treated as scapegoats by the professionals. Relatives were denied information about the patient's illness on the grounds of confidentiality, which unnecessarily hindered their attempts to understand what was happening to the patient. Partly as a response to this treatment, relatives formed national organizations, such as the National Schizophrenia Fellowship in the UK and the National Alliance for the Mentally Ill in the US. This development raised the profile of relatives, gave them access to political power, and increased their confidence in dealing with professionals (e.g., see *Mental Illness: Family and Patient Organizations*).

Controlled trials of family interventions for schizophrenia began to appear in print in 1978 and over the next two decades more than a dozen such studies were published (Lam 1991). They all employed the same design, namely a comparison of family intervention plus antipsychotic medication against medication alone. This design reflects the important assumption that antipsychotic medication is an essential basis for the treatment and maintenance of a person with schizophrenia, but that working with the family to improve the management of the condition confers an added benefit. Indeed, this series of studies provided strong evidence for that proposition. Patients' relapse rate over one year was reduced from an average of 50 percent with medication alone to 20 percent or less by the addition of family interventions.

Despite a multiplicity of names, the interventions share much in common. They all start with an educational program which conveys to the family information about the causes of schizophrenia, the symptoms, the likely outcome, and the treatment and management. This is followed by instruction in problem solving, improving communication, reducing the intensity of negative emotions, utilizing the family's natural support networks, and developing realistic expectations for the patient. The general aim of the approach is to build on the family's strengths and to facilitate their mutual co-operation in coping with the illness.

3.3 Testing Systems Theory

Some programs were delivered by sessions in the family's own home, a radical departure from traditional family therapy, in which treatment often took place in a special clinic room equipped with a one-way

mirror to allow the whole team to monitor the interactions between therapist and family. Home visits are an effective way of engaging families, as well as enabling the therapists to observe families in their natural surroundings. These advantages have to be weighed against the extra staff time taken up by travel. Other programs set up groups for relatives that were run by professionals on the same lines as the family sessions. Voluntary organizations for families often establish self-help groups, but when they invite professionals along it is to provide information and not to institute therapy. Indeed they reject the term 'family therapy' as implying that relatives are pathological and in need of treatment themselves.

One American therapist, Bill McFarlane, based his program on multifamily groups that included patients as well as relatives. He has produced some evidence that these are more effective than groups comprising only relatives (McFarlane et al. 1993). Relatives groups provide the opportunity to test one of the basic tenets of systems theory, namely that altering one part of the system can effect change in the entire system. In the course of evaluating family interventions for schizophrenia, Leff et al. (1989) compared the effectiveness of family sessions with a relatives group from which patients were excluded. They found that relatives who attended the group showed a significant reduction in EE and this was accompanied by a lowered relapse rate for the patients. Hence, altering one component of the system, the relatives, led to the desired change in another component, the patient.

3.4 Relatives EE and Other Illnesses

Since the original research on schizophrenia, the measurement of EE has been applied to a variety of other conditions, both psychiatric and physical. In the realm of psychiatry, high levels of EE have been found in relatives of patients with manic-depressive psychosis, depressive neurosis, dementia, learning difficulties, alcoholism, eating disorders, and post-traumatic stress disorder, and in families in which children have suffered physical or sexual abuse. High EE has been found to be linked with outcome in all these conditions, with the exception of dementia and learning difficulties. In the domain of physical illness, high EE attitudes have been recorded in relatives of patients with epilepsy, heart disease, diabetes, inflammatory bowel disease, and Parkinson's disease. From this it appears that high EE attitudes can develop in response to any long-standing or recurrent illness, be it psychiatric or physical. However, an association with outcome is much stronger for psychiatric than for physical disorders. Regarding the latter, only for diabetes has a link been demonstrated and the evidence is equivocal.

It seems from this series of studies that the emotional attitudes of family members have a much greater impact on people suffering from disorders with an emotional basis than on people with physical illnesses where there is a demonstrable pathology. The exception might seem to be schizophrenia, in which brain abnormalities have been demonstrated (e.g., see *MRI (Magnetic Resonance Imaging) in Psychiatry*). However, a high proportion of patients with schizophrenia experience depressive symptoms and they are responsive to the same adverse events in their life as are people with a depressive illness. Hence, the underlying brain disorder does not negate their sensitivity to the social environment.

4. Couple Therapy for Depression

Marital therapy, as it used to be called before marriage went out of fashion, has a long history, and developed to help couples resolve their relationship difficulties or to part amicably. In the UK, a voluntary organization, Relate, trains counselors in the skills needed for couple therapy. However, there is not usually an identifiable psychiatric illness in either member of the partnership. In recent years, a small number of studies have tested the value of couple therapy for people with a depressive illness. In most of these, couple therapy has been compared with cognitive therapy, a treatment of known effectiveness for depression that focuses on the individual patient. In general, couple therapy has been found to be as effective as cognitive therapy for improving depression, but has stronger effects on the couple's relationship.

A recent study tested couple therapy against the best possible regime of antidepressant medication for depressed patients who were in a stable relationship. Both treatments were continued for one year and then stopped, after which the subjects were followed up for a further year. This design was chosen because most treatments in use improve depressive symptoms: the challenge is to reduce the liability for the symptoms to recur. The couple therapy was given by two systemic therapists who focused on the relationship between the partners as being the means by which the depression was maintained. The patients who received couple therapy showed a greater improvement in depression than those assigned to medication, and this advantage for couple therapy persisted over the second year after treatment had been stopped (Leff et al. 2000). Hence, changing the relationship between patient and partner was more effective both for treatment and prevention than altering the biochemistry in the patient's brain.

5. The Family as a Therapeutic Resource

Since family therapy was first introduced, there has been a major shift in professional attitudes from holding relatives responsible for the patient's illness to

viewing them as allies in a struggle to help the patient overcome symptoms and disabilities. Alongside this sea change, western families have been altering in their composition, structure, and function, requiring a greater flexibility in professional inputs. Several schools of family therapy have developed, each with their own distinctive concepts of the family, and have introduced novel techniques. Not all of these have proved their worth, but the emerging evidence of family influences on the course of a wide range of psychiatric disorders is a spur to test existing methods of working with families and to develop innovative ones.

See also: Cognitive and Interpersonal Therapy: Psychiatric Aspects; Family Therapy, Clinical Psychology of; Fatherhood; Mental Illness: Family and Patient Organizations; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy: Ethical Issues

Bibliography

Bateson G, Jackson D D, Haley J, Weakland J H 1956 Toward a theory of schizophrenia. *Behavioral Science* **1**: 251–64

von Bertalanffy L 1968 *General Systems Theory*. Braziller, New York

Butzlaff R L, Hooley J M 1998 Expressed emotion and psychiatric relapse: A meta-analysis. *Archives of General Psychiatry* **55**: 547–52

Fromm-Reichmann F 1948 Notes on the development of treatment of schizophrenics by psychoanalytic psychotherapy. *Psychiatry* **11**: 73–8

Hajdu-Gaines I 1940 Contributions to the etiology of schizophrenia. *Psychoanalysis Review* **27**: 421–38

Hirsch S R, Leff J P 1971 Parental abnormalities of verbal communication in the transmission of schizophrenia. *Psychological Medicine* **1**: 118–27

Jackson D 1957 The question of family homeostasis. *Psychiatric Quarterly* **31**: 79–90

Laing R D, Esterson D 1964 *Sanity, Madness and the Family*. Tavistock, London

Lam D H 1991 Psycho-social family intervention in schizophrenia: A review of empirical studies. *Psychological Medicine* **21**: 423–41

Leff J, Berkowitz R, Shavit N, Strachan A, Glass I, Vaughn C 1989 A trial of family therapy vs. a relatives group for schizophrenia. *British Journal of Psychiatry* **154**: 58–66

Leff J, Vaughn C 1985 *Expressed Emotion in Families: Its Significance for Mental Illness*. Guilford, New York

Leff J, Vearnals S, Brewin C R, Wolff G, Alexander B, Asen E, Dayson D, Jones E, Chisholm D, Everitt B 2000 The London intervention trial: An RCT of antidepressants versus couple therapy in the treatment and maintenance of depressed people with a partner: Clinical outcome and costs. *British Journal of Psychiatry* **177**: 95–100

Lidz T, Cornelison A R, Fleck S, Terry D 1957 The intrafamilial environment of the schizophrenic patient II. Marital schism and marital skew. *American Journal of Psychiatry* **114**: 241–8

McFarlane W R, Dunne E, Lukens E, Newmark M, McLaughlin-Toran J, Deakins S, Horen B 1993 From research to clinical practice: Dissemination of New York State's family psychoeducation project. *Hospital and Community Psychiatry* **44**: 265–70

Minuchin S 1974 *Families and Family Therapy*. Tavistock, London

Mishler E G, Waxler N E 1967 Family interaction patterns and schizophrenia: A multi-level analysis. In: Romano J *The Origins of Schizophrenia*. Excerpta Medica International Congress Series No. 151, Excerpta Medica, Amsterdam

Rubinstein D 1974 Techniques in family psychotherapy of schizophrenia. In: Cancro R, Fox N, Shapiro L E (eds.) *Strategic Intervention in Schizophrenia*. Behavioral Publications, New York

Russell G F, Szmukler G I, Dare C 1987 An evaluation of family therapy in anorexia nervosa and bulimia nervosa. *Archives of General Psychiatry* **44**: 1047–56

Selvini Palazzoli M, Boscolo L, Cecchin G, Prata G 1978 *Paradox and Counterparadox*. Jason Aronson, New York

Singer M T, Wynne L C 1966 Communication styles in parents of normals, neurotics and schizophrenics. *Psychiatry Research Reports* **20**: 25–38

J. Leff

Family and Schooling

There is an obvious difficulty in considering relationships of the family to other institutions at a time when it is changing rapidly and radically in structure, functions, and internal dynamics. While some investigations have addressed aspects of those changes, much research on family–school relationships does not take account of them. Perhaps this research is more properly regarded as providing information on homes, which can encompass a variety of family types, rather than on families.

Up to relatively recently, the family took major responsibility for the care, education, and training of children. In it, children acquired skills to deal with day-to-day tasks, including those of work. During the Industrial Revolution, as work moved into factories and became more specialized and skilled, and knowledge accumulated, homes could no longer provide children with the required knowledge and skills, and the need for an institution dedicated to formal learning arose. Increases in migration, industrialization, and urbanization served to underline the need, and so nations came to accept education and training in school as their responsibilities.

Children still spend most of their waking time outside school, however, and it would be surprising if the continuous influence of the family, both before children enter school and when they are attending school, did not play a major role in their learning. Although one would expect families and schools to play complementary and mutually reinforcing roles in children's education, they may not always do so

because of differences in their goals, structure, demands, activities, values, and the competencies they foster. When these differences are large, children may experience a serious dysfunction in socialization when they go to school, which will affect their ability to adapt and to learn. To deal with this problem, concerted efforts have been made in several countries in recent years to increase parents' involvement in the formal education of their children.

This article provides a summary of the findings of research that indicate that children's home circumstances are related to school learning. Proposed explanations of the findings are presented. Finally, programs to increase parental involvement in the education of children living in families and communities characterized by high rates of educational failure are considered.

1. *The Findings of Research on Family Characteristics and Children's School Learning*

Research on the family factors that influence children's development, much of it in the context of policies of providing equality of educational opportunity, has been carried out since 1900 in several disciplines: education, psychology, sociology, economics, and behavior genetics. There has been little integration of the work across disciplines.

Findings of the empirical studies can be summarized in eight general conclusions (Kellaghan et al. 1993, White 1982): (a) level of socioeconomic status (or social class) is positively but not very strongly related to children's performance on measures of scholastic ability and achievement. Children from high socioeconomic status homes tend to perform better than children from low socioeconomic status homes; (b) differences in cognitive skills between children from varying socioeconomic and ethnic backgrounds are in evidence before children go to school; (c) the gap between children from different backgrounds widens as children progress through school; (d) the length of time a student spends in the education system is a function of the socioeconomic status of the student's family. This is so even for students of the same level of general scholastic ability; (e) in countries in which secondary schools are differentiated, major selection in terms of socioeconomic background occurs at the point of transfer from primary to secondary school. Students from high socioeconomic backgrounds tend to transfer to an academic type of secondary school which will lead to third-level education, while children from low socioeconomic backgrounds tend to go to technical, vocational, or 'short-cycle' schools; (f) variables which describe the ethos and teaching style of families predict students' academic achievement better than do measures of the socioeconomic status of families. Thus, what parents do is more important than their status; (g) children's learning benefits from parental involvement. Such involvement is more likely to occur if parents construe their role as involving activities related to children's education, and if they believe that involvement will have a positive influence on their children and will be welcomed by schools (Epstein 1991, Hoover-Dempsey and Sandler 1995); (h) many studies reflect a low negative correlation between family size and educational abilities, achievements, and attainment. Increased spacing between children tends to reduce the decrement associated with increase in family size, while the effects of size are more marked in homes in which the father has a low occupational level.

Methodologies have been employed which reject the 'objective' quantitative approach of empirical studies, adopting the view that since acts imply purposeful constructions by actors, they can only be understood from the actor's point of view. In this context, ethnographic and case study procedures have been used to explore individuals' 'interpretations' of home-school relationships (see Marjoribanks 1991). The approach is supported by the findings of psychological research that perceptions of family supports are better predictors of personal adjustment than are actual supportive behaviors provided by network members (Pierce et al. 1996).

Research on the effects of variation in family structure and function on children's school learning does not provide clear-cut findings, which is not surprising, given the complexity and variety of the conditions being investigated. It is, for example, difficult to design studies that will isolate the effects of family circumstances (e.g., divorce, maternal employment, blended families, lone parenting) from the effects of many conditions frequently associated with these circumstances. One consistent finding, however, is that lone parenting is often associated with school problems for children when the parent (usually a mother) is young, financially poor, and poorly educated (Kellaghan et al. 1993).

Changes in family structure, which are associated with a decrease in the monitoring and supporting of children's behavior by parents and other adults, and in which the cultivation of one's own well-being replaces interest in others, are perceived by some commentators to have a negative impact on children's educational development (see Coleman 1987). The changes are represented as reflecting an erosion of 'social capital' (see Sect. 2).

2. *Explaining the Family's Influence on Children's Development*

Using a concept analogous to financial capital, the term 'cultural capital' has been used to conceptualize some of the influences of families, and the communities

in which they are embedded, on children's academic development. It exists in three forms: in cognitive and noncognitive competencies derived from past experiences (particularly familial ones) which actively organize future experience; in an 'objectified' state in cultural goods (pictures, books, instruments); and in an institutionalized state (educational qualifications). Language forms an important part of cultural capital since, in addition to being a means of communication, it provides, together with a richer or poorer vocabulary, a system of categories, which enables one to decipher and manipulate complex logical and aesthetic structures (Bourdieu 1986).

Separate strands of research in education (Kellaghan et al. 1993) and in psychology (Moos and Moos 1994) have identified a number of behaviors and conditions in the family that contribute to cultural capital: (a) modeling (in use of complex language; in planning and organization to ensure that time and space are well structured and used; intellectual-cultural orientation in activities; moral-religious emphasis); (b) providing motivation and reinforcement (encouraging and rewarding school-related activities and independence in decision-making); (c) holding high academic aspirations and expectations; (d) providing direct instruction (guiding and supporting academic work; helping with homework); and (e) ensuring that the activities engaged in are developmentally appropriate.

A number of points have been made about cultural capital: (a) its value will vary with the 'markets' in which it can be used advantageously; (b) capital required for success in school is defined by the dominant social groups in society; (c) ability or talent is the product of an investment of time and cultural capital; and (d) the notion of cultural capital helps explain the unequal scholastic achievement of children originating from different social classes. If you bring the right kind of capital to school, you will do well. If you do not, you are likely to find yourself at an academic, social, and ideological disadvantage in most schools (Bourdieu 1986).

In separate, but mutually supporting research in psychology and sociology, interpersonal relationships have been identified as another area in which the family's influence on children's development can be discerned. In psychology, the focus is on cognitive aspects of support (that incorporate expectations about how others will respond when assistance is needed) and on actual supportive behavior (Pierce et al. 1996). Sociologists, on the other hand, speak of 'social capital,' which is perceived to inhere in the structure of relations between and among individuals and organizations. Three forms have been identified: shared obligations and expectations; useful information that is passed between members of a social network; and a value or norm consensus between adults, coupled with a willingness to monitor behavior and enforce norms that control and shape children's activities. These conditions are seen as contributing to the development in children of basic attachments, attitudes, willingness to make certain efforts, and concepts of self, all of which work to children's advantage when they go to school. A variety of individuals and organizations can help children with these developmental tasks: siblings, grandparents and other members of the extended family, neighbors, churches, and various social and cultural groups (Coleman 1987, Garbarino et al. 1997).

These explanations of the factors in social and familial contexts that affect school learning do not help us understand how it is that children from the same family can differ considerably in their success at school. The answer would seem to lie in the fact that, although certain aspects of children's genetic endowment and environment (inside and outside the home) are shared by children in a family, other aspects are unique to each child. It may be that non-shared influences (e.g., the specific treatment of a child by parents and siblings) contribute more to variation in children's characteristics than family variables that are common to all children (e.g., socioeconomic status, general parenting style) (Rowe 1994). The complexity of the interaction between genetic and environmental factors becomes obvious if it is the case that compared to their less 'able' counterparts, genetically more 'able' parents create more stimulating environments, and genetically more 'able' children take a more active role in shaping their environment.

3. *Home–School Partnerships*

Recent efforts to improve home–school relationships may be attributed to concern about achievement levels in schools, sometimes arising out of the results of comparative international studies of achievement; the limited success, despite considerable investment, of efforts, such as 'Headstart' in the USA and similar programs elsewhere, to raise significantly the educational achievements of children living in areas described as economically and socially disadvantaged; and realization of the contribution that factors outside the school in homes and communities, particularly in the early years, make to the development of competence in children, a realization that was inspired by evidence relating to the importance of early childhood experiences, and of the effects of home background on achievement.

In the USA, there is a long tradition of involving parents and communities in children's formal education in, for example, parent–teacher associations. In Europe, the situation is more complex because of differences between countries in legal arrangements and ideology. However, in many countries, recent efforts, in some cases supported by legislation, go well beyond traditional parent–teacher associations in the role they propose for parents, involving them in

curriculum decision-making, and in teaching and learning activities in the classroom (e.g., as classroom assistants) or in the home (e.g., in developing children's skills in problem solving and language, and in supervision of homework). Many of the efforts are directed towards families in 'disadvantaged' areas, where children face economic, educational, and social problems, and are considered to be at risk of educational failure, school dropout, and intergenerational poverty. As is the case with family-support initiatives in general, programs are designed to improve the developmental capacity and educational options available to families. They are based on a variety of models and approaches, have few theoretical frameworks, and vary in the extent to which empirical evidence is available on their effectiveness.

Most programs can be described as instrumental-technical, in that they are designed to provide parents with a range of skills that are considered relevant to their children's education. Alternative approaches are based on an interpretative paradigm, which, recognizing the uniqueness of parents' experience, attempt to facilitate them in interpreting and understanding that experience. Yet other approaches go a step further, again helping parents interpret their situation, but to do so critically in the context of presumed determining social structures and ideologies (e.g., Morgaine 1992).

One of the mechanisms used in the USA (Becker et al. 1997) and in the UK (Bastiani and Wyse 1999) to involve parents requires them to sign home-school contracts. Such agreements describe, among other things, the reciprocal responsibilities of schools, parents, and students in relation to student learning and school discipline.

A number of problems associated with efforts to promote partnerships have been identified: (a) despite rhetoric regarding choice, partnership, and parents' role as 'primary educators,' many efforts mirror the traditional home-school power relationships that they were designed to alter; (b) in many cases, schools limit the involvement of parents, denying them any real input into the construction of 'shared understandings'; (c) many approaches are based on an implicit view of parents as being inadequate in parenting skills, or as being 'hard to reach'; (d) many parents are untouched by the partnership, and remain distanced and alienated; (e) when involvement is based on contracts which require agreement before a student will be admitted to a school, only families with a certain level of commitment may enrol their children; (f) the extent of parental influence depends on parents' social, cultural, and material resources; (g) finally, strategies generally fail to reach fathers (Vincent and Tomlinson 1997).

While evaluations of individual initiatives to improve family involvement have reported a variety of effects on parents and children, it is difficult to generalize about the effectiveness of home–school partnerships because of variation in the goals of programs, in the situations in which programs are delivered, in the delivery of services, and in the recipients of services.

4. Conclusion

Research has gone a considerable way in identifying family characteristics that are associated with children's successful school learning. We may expect future research to be more interdisciplinary, to focus more on individuals' perceptions of family–school relations, to attend to the role of individuals' unique nonshared attributes and experiences, and to support the development of parent involvement programs that are based on the concept of equality of the partners.

See also: Divorce and Children's Social Development; Families as Educational Settings; Family as Institution; Learning and Instruction: Social-cognitive Perspectives; Nontraditional Families and Child Development; Parenting: Attitudes and Beliefs; Parenting in Ethnic Minority Families: United States; Parents and Teachers as Partners in Education

Bibliography

Bastiani J, Wyse B 1999 *Introducing Your Home-School Agreement*. Royal Society for the Encouragement of Arts, Manufacture and Commerce, London

Becker H J, Nakagawa K, Corwin R G 1997 Parent involvement contracts in California's charter schools: Strategy for educational improvement or method of exclusion? *Teachers College Record* **98**: 511–36

Bourdieu P 1986 The forms of capital. In: Richardson J G (ed.) *Handbook of Theory and Research for the Sociology of Education*. Greenwood Press, Westport, CT, pp. 241–58

Coleman J S 1987 Families and Schools. *Educational Researcher* **16**(6): 32–8

Epstein J L 1991 Effects on student achievement of teachers' practices of parent involvement. In: Silvern S B (ed.) *Advances in Reading/Language Research. Literacy through Family, Community, and School Interaction*. JAI Press, Greenwich, CT, Vol. 5, pp. 261–76

Garbarino J, Kostelny K, Barry F 1997 Value transmission in an ecological context: The high-risk neighborhood. In: Grusec J E, Kuczynski L (eds.) *Parenting and Children's Internalization of Values. A Handbook of Contemporary Theory*. Wiley, New York, pp. 307–32

Hoover-Dempsey K V, Sandler H M 1995 Parental involvement in children's education: Why does it make a difference? *Teachers College Record* **97**: 310–31

Kellaghan T, Sloane K, Alvarez B, Bloom B S 1993 *The Home Environment and School Learning. Promoting Parental In volvement in the Education of Children*, 1st edn. Jossey-Bass, San Francisco

Marjoribanks K 1991 Families, schools, and students' educational outcomes. In: Fraser B J, Walberg H J (eds.) *Educational Environments: Evaluation, Antecedents and Consequences*, 1st edn. Pergamon, Oxford, UK, pp. 75–91

Moos R H, Moos B S 1994 *Family Environment Scale Manual*. Consulting Psychologists Press, Palo Alto, CA
Morgaine C A 1992 Alternative paradigms for helping families change themselves. *Family Relations* **41**(1): 12–17
Pierce G R, Sarason B R, Sarason I G 1996 *Handbook of Social Support and the Family*. Plenum Press, New York
Rowe D C 1994 *The Limits of Family Influence. Genes, Experience and Behavior*. Guilford Press, New York
Vincent C, Tomlinson S 1997 Home–school relationships: 'The swarming of disciplinary mechanisms?'. *British Educational Research Journal* **23**: 361–77
White K R 1982 The relation between socioeconomic status and academic achievement. *Psychological Bulletin* **91**: 461–81

T. Kellaghan

Family, Anthropology of

At intervals during the last 50 years, anthropologists have pointed out the deficiencies of the concept of the family as an analytical tool (e.g., Fortes 1958, Yanagisako 1979, La Fontaine 1985). One of the more recent, Robertson, starts with an account of the importance of the idea of the family in Western culture; later, he draws attention to social scientists' 'fixation on the family' (1991, p. 154). These two phenomena are undoubtedly connected. An account of anthropological understandings of the family must distinguish between its significance in Western society and its use in the comparative analysis of social life. Comparing the two may then expose the fundamental ethnocentricity of the concept and its weakness in cross-cultural analysis.

1. *The Family in Western Society*

The significance of the family in Western societies is obvious in daily life. Children in primary school are asked to describe their families as a writing exercise and politicians vie with one another to be 'the party of the family.' Christianity, the majority religion of Western society, celebrates the Holy Family and its priests proclaim the duty of Christians to emulate it in their domestic life. The terms used by members of a family for one another are not applied to any other relatives and the relationships they designate are seen as unique. The family encloses the private life of individuals and, as the building block of society, is given equal weight with the public institutions of the state, since the division into public and private is one of the fundamental social distinctions in Western culture. Yet this important institution is also the subject of controversy. There are arguments over the past and future of the family, over whether it is universal or merely an adaptation to the modern world, over the possibility of its decline and the effect on it of political and economic measures. The family is not taken for granted.

Despite its importance, the word 'family' has no clearly defined meaning. In twentieth century England, for example, in ordinary speech it may evoke any one of a cluster of meanings, depending on context (Harris 1983, p. 31). As well as parents and children, 'family' can mean: a network of kin ('all the family'); a succession of related people ('the family has lived here for 400 years'); children ('they have not brought the family with them'); birth status ('he comes from a good family'); household ('children leave the family as they grow up'); relatives generally ('she has no family in London'); etc. It is difficult to claim that any of these meanings is the 'true' meaning and others merely metaphorical usages, although some people might claim that the unit of parents and children is the primary sense of the term. However, there is evidence to show that this narrower meaning is relatively recent in English history. Davidoff and Hall (1987) trace its importance to the rise of the industrially based middle-class in the nineteenth century.

Three elements are *interrelated* in complex ways in the various uses of the word family: marriage, relatedness, or kinship and domestic or household organization. Each usage of the term carries different combinations of these elements. The 'elementary' or 'nuclear' family, as social scientists label parents and children, combines all three and it is this form that carries such symbolic weight in twentieth-century Western culture. At some point in the history of anthropology each of these three features has been used separately, and effectively, in cross-cultural comparison (see *Marriage*; *Kinship in Anthropology*; *Household in Anthropology*). The three principles are quite distinct in the abstract but separating them in analysis is not easy. However, it is the confusion between, for example, kinship and residence or marriage and parentage that causes problems in using 'the family' as a technical term.

In the West marriage is the foundation of 'the family.' As a legally established relationship the marital tie may be contrasted with the 'biological' links between parents and children or between siblings, which are referred to as blood ties. In Western societies marriage entails co-residence, although this is not a universal consequence of marriage (e.g., Fortes 1949). Where the norm is for a newly-married couple to set up their own household as in much of North America and Western Europe, this qualifies the newly-married couple to be referred to as a 'family' in anticipation of the children that are expected to follow. The death of a spouse or a divorce are individual breaks in the expected pattern but remarriage 'reconstitutes' the family once more.

The centrality of relatedness to the notion of family can be seen in the fact that not all residential units qualify to be called families. Individuals living on their own, boarding schools, prisons, and colleges are never referred to as families, nor are groups of unrelated people sharing a flat. In the case of the last three, this

may indicate that children are essential to a family, but children's residential homes are not considered families either. It is the nature of the ties among them that qualifies people who live together to be called a family.

The relationships internal to a family are thought of as emotional bonds, permanent ties that reflect ideals of sharing and altruism by contrast with the short-lived rational contracts of the market or the self-serving nature of political alliances. The authority of parents over children and the role of the father as 'head of the family' are believed to derive from the reproductive functions of the family and to reflect the 'natural' differences of generation and gender within it. The assumption that living together is necessary for the family to fulfill its reproductive function underlies the interchangeability of 'family' and 'household.'

However, Western cultures also make implicit distinctions between families and households. Where children are expected to leave the parental household at adulthood, this may, like the death or divorce of parents, be said to 'break up' the original family. Nevertheless, married adults continue to be related to their parents and siblings by genealogical ties that may be referred to collectively as belonging to a 'family,' although one which is not a household. Husband and wife belong, as offspring, to distinct 'families' that the marriage can be seen as linking. Individuals are thus members of at least two 'families': one in which they are children and (usually) siblings and one in which they are parents and spouses. Sociologists have labeled these the family of origin and the family of procreation, which emphasizes the different methods by which the focal individuals are recruited, birth and marriage; the labels incorporate the folk view. From a more analytical perspective the network of genealogical ties linking individuals can be seen as cross-cut by living arrangements that group them into different households.

Other usages of the term 'family' used in a variety of contexts show different combinations of kinship and residence. The term extended family may be used to refer to all individuals linked genealogically but not living together, giving this use of the term 'family' a purely kinship connotation, unlike its use to mean close kin who do live together. Alternatively, a large number of related households in one community may be referred to as an extended family, making their common locality a significant factor. Finally, 'extended family' may mean large households containing three or more generations, which can still be found in some rural communities in Europe. All these uses of the term are also current in academic works (e.g., Janssens 1993, 1fn).

The range of meanings that attach to the concept of the family make clear its great importance in Western culture. However, its complexity also raises a series of problems when it is used to analyze other societies. Simplifying the concept by identifying one or other of its component elements as fundamental has not solved the problems in cross-cultural analysis but has preserved the idea that 'the family' is a useful concept.

2. *The Family in Anthropological Analysis*

The issues with which the anthropology of the family have been associated reflect Western notions. At the outset the focus was on marriage. In the nineteenth century, the scholars who developed anthropology regarded monogamous lifelong marriage as characteristic of 'civilization,' the culmination of human evolution. The Western family was a late development not to be found among peoples who represented earlier stages of the evolutionary process. These latter were characterized by plural marriage and a failure to understand the nature and significance of biological links. Nineteenth century evolutionism has long been recognized as lacking a sound empirical basis for its conjectures, but the idea that the present form of 'the family' is a product of social changes over time was not entirely abandoned. It survived in modified form in the view that the present form of 'the family' is a product of modernization, of capitalism, or of industrial society and in the myth that Europe's peasant forebears all lived in extended families.

By the middle of the twentieth century, some anthropologists argued instead that the family could be said to be universal. The basis for this contrary view was the assumption that 'the family,' so far from being associated with marriage in a particular cultural form, was based on a universal, natural function: reproduction. It was this biological 'fact' that formed the basis of the family, a view widespread in the societies from which anthropologists came. 'The family,' a couple with their children, might be incorporated in various different domestic arrangements, but was itself a biological 'given.' Whether any individual society recognized this or not depended on their culture, a view not unlike that of the evolutionists. Malinowski, an early proponent of this functional view, nevertheless cautioned that '…the essential features of the individual family…depend upon the general structure of a given society…' (1913, p. 6). More boldly, Murdock (1949) proclaimed the nuclear family to be universal, a co-residential unit with 'sexual, economic, reproductive, and educational' functions. His conclusions depended on the application of the label 'family' to a wide variety of social units that were not strictly comparable. The procedure showed all too clearly the looseness of the term.

Anthropological arguments over the incest taboo paralleled those over whether the family was a biological or social entity (La Fontaine 1990, pp. 20–43). The view that the incest taboo reflected a natural avoidance of sexual relations between related people assumed that incest referred to sexual relations within the 'nuclear family'; these relationships were such that

sexual attraction naturally did not arise. Other anthropologists pointed to documented cases of sexual prohibitions of an identical nature involving relationships outside the 'nuclear family,' as was also the case in Western society. Incest taboos, considered as total systems of prohibitions, varied widely cross-culturally, and could not therefore be considered biologically determined or 'natural' (Goody 1956). The later development of sociobiology resulted in claims that the taboo could also be observed to exist among the higher primates and was therefore genetic, rather than a feature of human family relations. This and other forms of biological determinism were vigorously contested (Schneider 1965, Sahlins 1976); in particular it was pointed out that paternity cannot be identified among higher primates and among human beings it is always socially determined. Few social or cultural anthropologists now accept that the incest taboo is universal, but the ideas that among close kin it is only rarely broken and that incest is 'unnatural' remains. At the end of the twentieth century, evidence that breaches of the incest taboo were infrequent but not rare in Western societies became public knowledge there. This offered proof that the 'taboo' was neither genetic nor unbroken. The reality of incest was not easily accepted, whether by the public at large or by social scientists, showing that the assumptions underlying it and also entailed in the concept of the family were deeply rooted in Western culture.

Knowledge of the variation in incest prohibitions came from the study of kinship systems that expanded rapidly in the middle decades of the twentieth century. The development had a further impact on views of the family. A solution to the theoretical problems raised by the data on domestic organization was to distinguish between 'the family' as a reproductive unit and the household or domestic group. One of the earliest to do this was Fortes (1958) who used the term domestic group, referring to the 'family' as the reproductive nucleus within it. His distinction emphasized a number of contrasts between the concepts: 'the family' consists of parents and children, whereas the domestic group might include more than two generations; the family is recruited through marriage and kinship ties, whereas many different kinds of tie may draw individuals into living together. The family is concerned with social and physical reproduction, the domestic group 'is essentially a householding and housekeeping unit' (Fortes 1958, p. 8). According to Fortes, the changes in the size and structure of domestic groups that can be observed in a particular society do not reflect breaches of any 'rule of residence' as had previously been thought, but derive partly from the effect on household constitution of the political and economic structure but also from the developmental cycle of growth and maturation in the reproductive nucleus. These processes and the cultural reactions that they entail are the driving force behind household changes.

Fortes' distinction also drew attention to production as a household activity. Property rights and labor needs were shown to influence the form of the domestic group as much as the developmental cycle of the reproductive core. Households were shaped by the demands for different kinds of labor associated with their subsistence activities and the association of related people within different households concerned economic viability as well as kinship obligations. Goody went further and showed that interpersonal relations in the 'reproductive core' were also affected by these factors (Goody 1956). Marxist anthropology subsequently attempted to construct different modes of domestic production and reveal the power structures entailed in them, but these largely ignored reproductive processes.

Distinguishing 'family' from 'household' avoided rather than solved the problems inherent in the concept of the family (La Fontaine 1985). The notion of the family as a reproductive *unit* conflates childbearing and child-rearing with co-residence so that in effect the family is actually defined by residence, a return to the Western folk concept (Smith 1973). It is not surprising that some anthropologists have subsequently reverted to using family and household indistinguishably. The distinction also maintained the segregation of domestic kinship from the network of kin as a whole, which was particularly misleading where spouses were chosen from among them.

The French anthropologist Lévi-Strauss' revolutionary approach to kinship disposed of some of these questions. He argued that exchange of goods and of women, not biological reproduction, was the basis of social life. Hence, marriage rules are integral to the wider social structure, not merely the basis of the family or a means of recruitment to households. Systems where marriage with relatives was the ideal showed how the patterns of distribution of women in marriage constituted social structures. His 'atom of kinship' (1958 (1963)) contained the same elements as 'the nuclear family,' but rearranged so as to emphasize the relation between a man and his wife's brother created by marriage. Moreover its elements, pairs of opposites, reflected the oppositions in systems of thought, the deep structure of culture.

Under the rather more materialist influence of feminism, anthropologists in the 1970s showed how the 'natural' processes of childbirth seemed to be very widely used to symbolize women and legitimize their subordination to men. Within the domestic sphere, also associated with women, the division of labor was seen as reflecting the natural capabilities of men and women; the products of their labor were shared with their children. However, the distribution of goods and the allocation of labor within the household actually reflected its internal power structure rather than a pooling of family resources or a 'natural' division of family labor. The realities of power were presented as unquestionable by being associated with the natures of

men and women. Concepts of nature were thus shown to be more than mental constructs; they were also intimately bound up with the social distribution of power and authority, which they helped to reproduce. The 'family' was revealed as a similar social representation legitimizing household organization (Harris 1981).

Both of these approaches stimulated research into the meanings of folk concepts and into local systems of knowledge (Barnes 1973, Harris 1981, Carsten 1997). The social constructions of the person in folk theories made it clear that whatever the common physical nature of human beings, it is perceived through these culturally variable understandings. The image of shared blood or common physical substance are only two among many ways of representing relatedness and the concept of the family is a folk concept among others.

While Levi-Strauss' structuralist approach undoubtedly revealed the significance of cultural understandings of the world, he was not interested in distinguishing between the meanings drawn from the analysts' own Western culture and those of the societies they were studying. Recognition of this ethnocentricity led to a wholesale rejection of such oppositions as domestic and political, natural, and social that underpinned the notion of the family and kinship (Schneider 1984, Carsten 1997, p. 20). Others accepted that certain distinctions might be culturally specific, but argued that the topic of relatedness could be approached through the folk concepts of the society studied (Carsten 1995, p. 225). This left 'the family' as a Western folk concept, limited in its usefulness to an understanding of the West.

A final strand in late twentieth century anthropological thinking on 'the family' has been contributed by research in Europe and on issues of public concern (La Fontaine 1990, Simpson 1998). While there is a general acceptance that the concept of the family is analytically unhelpful, the term continues to be used. Thus, Segalen and Zonabend, having noted that under the influence of anthropology the 'overfluid' concept of the family has been replaced by analytical concepts such as household and kinship (1987, p. 112), then use 'family' unreflectively and do not discuss why or in what sense. Rather than refer to kinship networks established after divorce, Simpson retains the term family, though qualifying it as 'the unclear family.' He argues for the study of divorce, not as a social problem or a symptom of general breakdown, but as the reorganization of interpersonal relationships in new forms of kindred. His rejection of 'the family' as a theoretical model is explicit, but his continued use of the term implies, as does that of Segalen and Zonabend, an attachment to certain connotations of it that have yet to be explored. The concept of the family may have been discarded from the repertoire of many anthropologists but it is too soon to regard its relegation to the scrap-heap as final.

See also: Dissolution of Family in Western Nations: Cultural Concerns; Divorce and Children's Social Development; Divorce and Gender; Divorce, Sociology of; Family and Gender; Family and Kinship, History of; Family and Psychiatric Illness/Family Therapy; Family as Institution; Family Law; Family Processes; Family Systems in Europe; Family Theory: Feminist–Economist Critique; Family Therapy, Clinical Psychology of; Fatherhood; Female Genital Mutilation; Feminist Legal Theory; Gender and Feminist Studies in Anthropology; Gender and Reproductive Health; Gender Role Stress and Health; Household in Anthropology; Household Production; Kinship in Anthropology; Kinship Terminology; Lone Mothers in Affluent Nations; Love and Intimacy, Psychology of; Male Dominance; Marital Interaction: Effects on Child Development; Marriage; Marriage and the Dual-career Family: Cultural Concerns; Matrifocality; Men's Health; Nation-states, Nationalism, and Gender; Parenthood and Adult Psychological Developments; Parenting: Attitudes and Beliefs; Parenting in Ethnic Minority Families: United States; Partnership Formation and Dissolution in Western Societies; Repartnering and Stepchildren

Bibliography

Barnes J 1973 Genitor: genitrix: nature: culture. In: Goody J (ed.) *The Character of Kinship*. Cambridge University Press, Cambridge, UK

Carsten J 1995 The substance of kinship and the heat of the hearth: feeding, personhood, and relatedness among Malays in Pulau Langkawi. *American Ethnologist* **22**: 223–41

Carsten J 1997 *The Heat of the Hearth*. Clarendon Press, Oxford, UK

Davidoff L, Hall C 1987 *Family Fortunes*. University of Chicago Press, Chicago

Fortes M 1949 Time and social structure: an Ashanti case-study. In: Fortes M (ed.) *Essays Presented to A. R. Radcliffe-Brown*. Clarendon Press, Oxford, UK

Fortes M 1958 Introduction. In: Goody J (ed.) *The Developmental Cycle in Domestic Groups*. Cambridge University Press, Cambridge, UK

Goody J 1956 A comparative approach to incest and adultery. *British Journal of Sociology* **7**: 286–305

Goody J 1983 *The Development of the Family and Marriage in Europe*. Cambridge University Press, Cambridge, UK

Harris O 1981 Households as natural units. In: Young K, Wolkowitz C, McCullagh R (eds.) *Of Marriage and the Market*. CSE Books, London

Harris C 1983 *The Family in Industrial Society*. George Allen and Unwin, London

Janssens A 1993 *Family and Social Change*. Cambridge University Press, Cambridge, UK

La Fontaine J S 1985 Anthropological perspectives on the family and social change. *The Quarterly Journal of Social Affairs* **1**(1): 29–56

La Fontaine J S 1990 *Child Sexual Abuse*. Polity Press, Cambridge, Oxford, UK

Levi-Strauss C 1958 (1963) *Structural Anthropology*. Basic Books, New York

Malinowski B 1913 *The Family Among the Australian Aborigines*. University of London Press, London
Murdock G P 1949 *Social Structure*. Macmillan, New York
Robertson A 1991 *Beyond the Family*. Polity Press, Cambridge, UK
Sahlins M 1976 *The Use and Abuse of Biology*. University of Michigan Press, Ann Arbor, MI
Scheider M A 1984 *A Critique of the Study of Kinship*. University of Michigan Press, Ann Arbor, MI
Schneider D M 1965 Kinship and biology. In: Coale A J, Levy M J, Fallers L A, Scheider M A, Tomkins S S (eds.) *Aspects of the Analysis of Family Structure*. Princeton University Press, Princeton, NJ
Segalen M, Zonabend F 1987 Social anthropology and the ethnology of France: The field of kinship and the family. In: Jackson A (ed.) *Anthropology at Home* [trans. Benthall J]. ASA Monographs 25. Tavistock, London
Simpson B 1998 *Changing Families*. Berg, Oxford, UK
Smith R T 1973 The matrifocal family. In: Goody J (ed.) *The Character of Kinship*. Cambridge University Press, Cambridge, UK
Yanagisako S J 1979 Family and household: The analysis of domestic groups. *Annual Review of Anthropology* **8**: 161–205

J. S. La Fontaine

Family as Institution

The family is one of the foundational social institutions in all societies, although the definition of 'the family' varies from place to place and from time to time. Thornton and Fricke (1989, p. 130) suggest an inclusive definition that provides a good starting point: 'family [is] a social network, not necessarily localized, that is based on culturally recognized biological and marital relationships.' In most times and places, families were responsible for production, distribution, and consumption of commodities, for reproduction and socialization of the next generation, for co-residence and transmission of property. And families generally still are. This article examines the structure of the contemporary family and changes in that structure. It describes alternative family forms and it looks at the processes producing families and dissolving them.

1. *Family Mode of Social Organization*

Under the family mode of social organization, kin groups pooled resources, including their labor, specialized in particular tasks, coordinated their activities, and connected to the larger community as a unit. This family mode of social organization is often associated with agricultural production, but it appears in a wide range of economic environments (Thornton and Fricke 1989).

The family mode of social organization has altered, however, with other, far-reaching social changes, including the rise of the market economy, vast increases in productivity with concomitant increases in real income (Fogel 2000), urbanization, changes in ideology toward greater individualization (Lesthaeghe 1983), and changes in the structure of education. All of these changes have shifted decision making and social control away from the family, toward the individual or toward other social institutions. As families have less control over the time and resources of children, they are less able to influence marriage choices—whether, when, and whom to wed. As more people support themselves through wage-based employment rather than through work on a family farm or small business, families have less stake in the property and family connections that a potential marriage partner brings and young adults acquire more autonomy in marriage choices (Caldwell et al. 1983). Urbanization and electronic communication have made the family a less important source of companionship and entertainment now than when most people lived on farms or in villages (Burch and Matthews 1987).

In developed industrial societies such as the USA, the family retains responsibility for reproduction, socialization, co-residence, and transmission of property across generations. It is the main unit of consumption and often also produces considerable amounts of goods and services. Families provide care and support for both the young and the old. Although older adults receive financial transfers and access to medical care from the government in many societies, family members still provide the vast majority of their help and support (Logan and Spitze 1996) and children are almost entirely dependent on their families for financial, emotional, and instrumental support.

2. *Structure of the Family*

In the USA and many industrialized societies, the structure of the family looks quite different than it did in the mid-twentieth century. In fact, fewer people live in families as traditionally defined and more live in nonfamily households. The rise in nonfamily living can be traced to earlier nestleaving by young adults (Goldscheider and Goldscheider 1993), to delayed marriage and to nonmarriage, to continued high rates of marital disruption with lower rates of remarriage (Cherlin 1992), and to increases in independent living at older ages (Michael et al. 1980). In 1998, 15 percent of all people lived in nonfamily households, 10 percent alone (US Bureau of the Census 1999, Table 16), compared to 6 percent in nonfamily households in 1950 (US Bureau of the Census 1955).

3. *Marriage*

In the USA, men and women are delaying marriage into their mid-to-late twenties, often entering a cohabitation first. Divorce rates are high and stable, but rates of remarriage have fallen, so that a larger

proportion of adults are unmarried at the beginning of the twenty-first century than in the past. In 1970, unmarried people made up 28 percent of the adult population. In 1996, 40 percent of all adults were unmarried. Seventy-one percent of women born in the early 1950s had married by age 25, compared to 54 percent of those born in the late 1960s (Raley 2000). In fact, the shift away from marriage has been so dramatic for blacks that now a majority of black men and women are not married, compared to about a third of white men and women (Waite 1995).

Similar changes in marriage patterns have taken place in most European countries; recent cohorts are marrying at older ages and over a wider range of ages than in the past. In addition, European countries differ substantially in marriage ages. The Nordic countries of Sweden, Denmark, and Iceland show the highest average ages at marriage for women (around age 29) and the Eastern European countries of Bulgaria, the Czech Republic, Hungary, and Poland the lowest (around age 22). Since societies with relatively high age at marriage also tend to be those in which many people never marry, this diversity suggests that marriage is a more salient component of family in some European countries than others (Kiernan 2000).

Countries in Europe also show a great deal of variation in the proportion of women in marital unions. Marriage is most common in Greece and Portugal, where over 60 percent of women aged 25 to 29 are married, and least common in the Nordic countries, Italy, and Spain where a third or less are.

Age at marriage also seems to be rising in many countries in Africa, including the northern African countries of Egypt, Sudan, Morocco, and Tunisia, and in Kenya, Mauritania, Nigeria, Senegal, Togo, Uganda, and Zimbabwe. Countries in northern Africa also show a rise in age at first birth that is not apparent in the rest of Africa, suggesting an increase in nonmarital childbearing in those countries in which age at marriage is rising (van de Walle 1993). The proportion of women who are currently married is 55 percent in Morocco, 61 percent in Kenya, 58 percent in Haiti (both formal and informal unions), 68 percent in Indonesia, and 63 percent in Bangladesh (Macro International 2000).

4. *Cohabitation*

Declines in marriage are closely linked to increases in cohabitation, although it is difficult to untangle the nature of the association. In the USA cohabitation has become an increasingly common step in the courtship process; only seven percent of the women born in the late 1940s cohabited before age 25 compared to 55 percent among those born in the late 1960s (Raley 2000). Most couples begin their intimate life together by cohabiting rather than by marrying, so that the form of the union has changed more than its existence. But even when we consider both marriage and cohabitation, young adults are less likely to have formed a union now than in the past. Among young women born in the early1950s about a quarter had not formed a union by age 25, compared to a third of those born in the late 1960s (Raley 2000).

Although a number of European countries have experienced similar increases in cohabitation, some have experienced much more and some much less. Cohabitation is strikingly common in the Nordic countries of Denmark, Sweden, and Finland, and France also shows fairly high levels, with about 30 percent of the women aged 25–29 in cohabiting unions. A group of countries that includes The Netherlands, Belgium, UK, West and East Germany, and Austria shows moderate levels of cohabitation—from 8 to 16 percent of women aged 25–29 are in this type of union. And in the Southern European countries and Ireland cohabitation is rare with less than three percent cohabiting among women aged 25–29 (Kiernan 2000).

In many European countries, women typically are in either cohabitational or marital unions by their mid-to-late twenties. However, over 60 percent of Italian women and 50 percent of Spanish are single, neither cohabiting nor married at these ages, compared to around one in three Portuguese and Greek women. In the Nordic countries and France, about a third of women aged 25–29 are cohabiting, a third are married, and a third are single. Marriage is much more common than cohabitation in all other European countries (Kiernan 2000).

5. *Unmarried Childbearing*

Changes in marriage have played a central role in increases in unmarried childbearing, which has reached historically unprecedented levels (Bachrach 1998); in 1996 32.4 percent of all births and 44 percent of all first births in the USA occurred to women who were not married (Ventura et al. 1999). But over a quarter of unmarried mothers are cohabiting with the child's father at the time of the birth so their children are living in 'intact,' if unmarried, families (Bumpass et al. 1995).

The proportion of births to unmarried women depends on the share of all women who are married, the fertility of unmarried women, and the fertility of married women. Marital fertility is relatively low in the USA and many European countries. Where nonmarital fertility is also low, as it is in Spain, Italy, and Japan, total fertility is substantially below replacement, but those low-fertility countries with fairly high levels of nonmarital childbearing, like the USA and France, tend to have higher total fertility (Rindfuss et al. 2000).

Unmarried childbearing varies substantially among racial and ethnic groups in the USA. The percentage of births to unmarried women is highest for black women (69 percent), and lowest for Chinese Americans (7 percent), with whites intermediate between

these two extremes at 26 percent. Rates of unmarried childbearing also vary a good deal within Hispanic origin groups, with rates for Puerto Rican women approaching rates for blacks (59 percent), whereas rates for Cuban-origin women (24 percent) approximate those of non-Hispanic whites (Ventura et al. 1999).

As we might expect from differences in the proportion of women who are single and in the proportion in cohabitational unions, unmarried childbearing also varies dramatically across countries in Europe. The Nordic countries, which have a high rate of cohabitation, also have quite high percentages of births to unmarried women; in Norway, for example, 46 percent of births at the end of the twentieth century occurred to unmarried women, most of whom were cohabiting. More than half of births in Sweden and almost half in Norway occur to unmarried women. At the other extreme, only 10 percent of births in Spain, 8 percent in Italy and 3 percent in Greece are to unmarried women. Countries such as Ireland, Belgium, Germany, Portugal, and The Netherlands fall in between (Rindfuss et al. 2000). The social implications of unmarried childbearing depend on both the extent to which these births take place within stable, socially recognized cohabiting couples, like in the Nordic countries, and the extent to which social welfare programs cushion the financial impact on the family of having a single parent. Where few supports for single parents exist, as in the USA, children in these families do less well than children raised in two-parent families, but even under these circumstances, most children raised by single parents are happy and successful as adults (Amato and Booth 1997).

6. *Marital Disruption and Union Dissolution*

A substantial proportion of all marriages end in divorce or separation due to marital discord. The divorce rate, which reflects the number of divorces in a year relative to the number of married people, rose continuously for more than a century in the USA and many similar industrialized countries, then leveled off at a fairly high level in about 1980 (Goldstein 1999). In the USA, the best estimates suggest that around half of all marriages will be disrupted (Cherlin 1992). The marriages most likely to end include those with no children, with children from a previous union or older children (Waite and Lillard 1991), marriages begun at a young age, and marriages between partners with relatively low levels of education (Martin and Bumpass 1989).

Although high divorce rates make marriages seem unstable, other types of unions are much more likely to dissolve. Cohabitational unions show quite high chances of disruption, with a quarter ending in separation within three to four years compared to only five percent of marriages, according to one study (Wu and Balakrishnan 1995). Many cohabitations become marriages, but these show lower stability than marriages not preceded by cohabitation (Lillard et al. 1995).

7. *Alternative Family Structures*

The married, two-parent family has been the most common family form in the USA and other industrialized countries for some centuries. But even at the height of the married couple family, many people lived in other types, most often due to the death of one member of the couple before all the children were grown (Watkins et al. 1987). When death ended many marriages relatively early in life, remarriage and stepfamilies were common, as were single-parent families caused by widowhood. The rise of cohabitation and nonmarital childbearing have meant that unmarried-couple families and never-married mother families are now common alternative family forms.

8. *Sex*

In spite of the sexual revolution, marriage circumscribes the vast majority of sexual relationships. Almost all married men and women are sexually active and almost all have only one sex partner—their spouse. Unmarried men and women have much lower levels of sexual activity than the married, and frequently have no sex partner at all. Cohabiting couples are at least as sexually active as married couples, but are much less likely to be sexually exclusive (Laumann et al. 1994). Thus, the married couple remains the locus of the vast majority of sexual activity.

9. *Working Families*

Perhaps as dramatic and far-reaching as the alternations in the structure of the family are the changes in the way its members use their time. In the early 1960s in the USA, among those in the prime working ages most married couples followed the male breadwinner/female homemaker model; 56 percent had only one earner. The dual-income family was uncom mon—both spouses worked full-time in 21 percent of married couples. By 1997, only a quarter of married couples had one earner. In 44 percent of married couples both spouses worked full-time, and in another 24 percent one worked full-time and one part-time. The shift toward the dual-worker family was even more dramatic for couples with children (Waite and Nielsen 2001). Even by the beginning of the 1970s, most children living with a single parent did not have a parent at home full-time; now most children in married-couple families do not either.

Many public commentators and some scholars of the family argue that the family is in decline and under siege from legal, economic and social change (Popenoe 1993). As evidence, they point to low marriage rates,

high divorce rates, low marital and high nonmarital fertility. And this evidence is compelling, as far as it goes. But most adults are married (although it is a second marriage for many), most have children, most rate their marriage as very happy and place a high value on family life, and most have only one sex partner—their spouse. Most single parents raise happy and successful children. This evidence suggests that the family remains a key social institution, even in its altered state.

See also: Divorce and Gender; Divorce, Sociology of; Families as Educational Settings; Family and Gender; Family and Kinship, History of; Family Bargaining; Family Processes; Family Theory: Economics of Childbearing; Family Theory: Economics of Marriage and Divorce; Family Theory: Role of Changing Values; Gender-related Development; Kinship in Anthropology; Marriage; Marriage and the Dual-career Family: Cultural Concerns

Bibliography

Amato P R, Booth A 1997 *A Generation at Risk: Growing Up in an Era of Family Upheaval*. Harvard University Press, Cambridge, MA

Bachrach C 1998 The changing circumstances of marriage and fertility in the United States. In: Moffitt R A (ed.) *Welfare, the Family and Reproductive Behavior: Research Perspectives*. National Academy Press, Washington, DC, pp. 9–32

Bumpass L L, Raley R K, Sweet J A 1995 The changing character of stepfamilies: Implications of cohabitation and nonmarital childbearing. *Demography* **32**: 425–36

Burch T K, Matthews B J 1987 Household formation in developed societies. *Population and Development Review* **13**(3): 495–511

Caldwell J C, Reddy P H, Caldwell P 1983 The causes of marriage change in South Asia. *Population Studies* **37**: 343–61

Cherlin A J 1992 *Marriage, Divorce, and Remarriage*. Harvard University Press, Cambridge, MA

Fogel R W 2000 *The Fourth Great Awakening and the Future of Egalitarianism*. University of Chicago Press, Chicago

Goldscheider F K, Goldscheider C 1993 *Leaving Home Before Marriage: Ethnicity, Familism, and Generational Relationships*. University of Wisconsin Press, Madison, WI

Goldstein J R 1999 The leveling of divorce in the United States. *Demography* **36**: 409–14

Kiernan K 2000 European perspectives on union formation. In: Waite L, Bachrach C, Hindin M, Thomson E, Thornton A (eds.) *Ties that Bind: Perspectives on Marriage and Cohabitation*. Aldine de Gruyter, New York, pp. 40–58

Laumann E O, Gagnon J H, Michael R T, Michaels S 1994 *The Social Organization of Sexuality*. University of Chicago Press, Chicago

Lesthaeghe R 1983 A century of demographic and cultural change in Western Europe: An exploration of underlying dimensions. *Population and Development Review* **9**(3): 411–35

Lillard L A, Brien M J, Waite L J 1995 Pre-marital cohabitation and subsequent marital dissolution: Is it self-selection? *Demography* **32**: 437–58

Logan J R, Spitze G D 1996 *Family Ties: Enduring Relations between Parents and Their Grown Children*. Temple University Press, Philadelphia, PA

Macro International 2000 www.macroint.com/dhs/pubs/fs/nav.cfm?fn=bd93e.cfm

Martin T C, Bumpass L L 1989 Recent trends in marital disruption. *Demography* **32**: 509–20

Michael R T, Fuchs V R, Scott S R 1980 Changes in the propensity to live alone: 1960–1976. *Demography* **20**: 433–47

Popenoe D 1993 American family decline, 1960–1990: A review and reappraisal. *Journal of Marriage and the Family* **55**: 527–55

Rindfuss R R Benjamin K Morgan S P 2000 The changing institutional context of low fertility. Working paper, Carolina Population Center, University of North Carolina, Chapel Hill

Raley R K 2000 Recent trends in marriage and cohabitation. In: Waite L, Bachrach C, Hindin M, Thomson E, Thornton A (eds.) *Ties that Bind: Perspectives on Marriage and Cohabitation*. Aldine de Gruyter, New York, pp. 19–39

Thornton A, Fricke T E 1989 Social change and the family: Comparative perspectives from the West, China and South Asia. In: Mayone Stycos J (ed.) *Demography as an Interdiscipline*. Transaction, New Brunswick, NJ, pp. 128–61

US Bureau of the Census 1955 Household and family characteristics: March 1954. *Current Population Report*. US Government Printing Office, Washington, DC

US Bureau of the Census 1999 Household and family characteristics: March 1998 (Update). *Current Population Report*. US Government Printing Office, Washington, DC

van de Walle E 1993 Recent trends in marriage ages. In: Foote K A, Hill K H, Martin L G (eds.) *Demography Change in Sub-Saharan Africa*. National Academy Press, Washington, DC, pp. 117–52

Ventura S J, Martin J A, Curtin S C, Mathews T J 1998 Report of final natality statistics, 1996. *Monthly Vital Statistics Report*, Vol. 46, No. 11, Suppl. National Center for Health Statistics, Hyattsville, MD

Ventura S J, Martin J A, Curtin S C, Mathews T J 1999 Births: Final data for 1997. *National Vital Statistics Report*, Vol. 47, No. 18. National Center for Health Statistics, Hyattsville, MD

Waite L J 1995 Does marriage matter? *Demography* **32**: 483–508

Waite L J, Lillard L A 1991 Children and marital disruption. *American Journal of Sociology* **96**: 930–53

Waite L J, Nielsen M 2001 The rise of the dual-worker family: 1963–1997. In: Hertz R, Marshall N (eds.) *Women and Work in the Twentieth Century*. University of California Press, Berkeley, CA

Watkins S C, Menken J A, Bongaarts J 1987 Demographic foundations of family change. *American Sociological Review* **52**: 346–58

Wu Z, Balakrishnan T R 1995 Dissolution of premarital cohabitation in Canada. *Demography* **32**(4): 521–32

L. J. Waite

Family Bargaining

1. Introduction

The concept of bargaining refers to a particular approach to decision-making in situations of conflict. Such conflicts could be resolved in other ways: dictatorially, for instance, through the use of force or authority, or democratically through majority vote. Bargaining, on the other hand, refers to the attempts

by the different parties to a dispute to threaten, cajole, or persuade others to settle it in their favor. Bargaining takes different forms because it occurs in different contexts. This article reviews different social science approaches to the analysis of bargaining in the specific context of family-based households and the insights they provide into its distinctive features. Two broad approaches will be distinguished: the neoclassical and the 'institutional.'

2. *Neoclassical Economics and Household Decision-making*

Neoclassical economists tend to explain human behavior through abstract mathematical models and econometric approaches to data analysis. They therefore place considerable value on 'parsimony,' or the ability to explain complex phenomena through a very limited set of variables. This is evident in early models of household decision-making which assumed it to be the responsibility of the household head, acting as 'benevolent dictator': 'benevolent' because he was assumed to take account of the welfare of all members in his decisions and a 'dictator' because he was able to impose his decisions on them (such models implicitly assumed a male household head, an assumption which is often but not invariably appropriate). However, while this dual assumption considerably simplified attempts to model household decision-making, it ruled out the possibility that power and conflict might have any role to play in the process (Becker 1981)

Two bodies of empirical evidence challenged this representation of household decision-making. The first showed that access to resources by individual members influenced the allocation of intrahousehold resources, suggesting that household decision-making processes were susceptible to the economic leverage exercised by different members (Thomas 1993, Hoddinott and Haddad 1995). The second documented phenomena such as domestic violence as well as systematic inequalities in household welfare along age and gender lines (Momsen and Townsend 1987), further undermining the assumption of conflict-free, welfare-maximizing, decision-making. Bargaining models of household decision-making emerged in response to such evidence. Their starting premise was that individuals cooperated to form households because the resulting benefits exceeded those obtainable from living alone. Where members had conflicting views about certain decisions, they might seek to bargain over the final outcomes. However, not all members had equal bargaining power and bargaining models largely were concerned with establishing the determinants of differentials in bargaining power within the household. These were seen to depend on the kind of 'threats' that different members could make in order to influence the decision-making process. The ultimate threat available was the threat of 'exit,' which in the context of the household, implied withdrawal from its membership.

However, the credibility of such a threat depended on the kind of resources that members were able to command as individuals rather than as household members. Initially, the 'threat' position was equated with access to individual economic resources, for instance, labor market earnings. It was expanded subsequently to include aspects of the social context which were likely to impinge on the threat positions of individuals as social groups. For female members, for instance, individual threat positions might be influenced by sex ratios in the relevant marriage market, laws concerning alimony/child support settlements, women's ability to return to their natal homes after marital breakdown, and the cultural acceptability of outside work.

Quantitative tests of bargaining models supported the assumption that individual economic characteristics, such as wages and incomes, as well as broader structural factors, such as legal provision relating to divorce settlements, did influence decision-making outcomes within the household (Hoddinott and Adam 1998, McElroy 1990, Schultz 1990, Jones 1986). They confirmed the feminist argument that, by and large, male household members exercised greater bargaining power than females, a reflection of their higher earning capacity as well as gender asymmetries in the legal structure. They also confirmed anthropological findings that men used their stronger bargaining position to claim a disproportionate share of household resources for their own consumption as well as more leisure (Dwyer and Bruce 1988). They pointed to the importance of women's access to paid employment for helping to shift the balance of decision-making power within the household. However, the insights provided by economic models have been limited by their focus on the economic determinants of bargaining power and their failure to consider its social dimensions. Yet it could be argued that it is its social dimensions which most help to distinguish the process of bargaining within the family from bargaining in any other arena of social life.

3. *Institutional Approaches and Family Bargaining*

Institutional approaches to household bargaining refer to attempts by nonneoclassical economists, sociologists, and anthropologists to account for the specificities of the household as an institutional form. These see households as an institutional response to the human need for long-term stable environments in which to bear and bring up children, to care, and be cared for, through sickness, disability, and old age and to plan for the future in a world characterized by uncertainty. What gives the household its institutional advantage in achieving these goals is the close inter-

twining of emotions and interests that characterize its relationships. Despite the different forms households take in different parts of the world, its relationships are generally familial. Its members are related by blood or marriage, they have lived together over considerable periods of time, they care for each other, and in any case, trust each other more than they trust strangers.

However, cooperative behavior within the household is not left entirely to the spontaneous impulse of members. It is also underwritten by a series of 'implicit contracts' that spell out the claims and obligations of different members to each other. Given the importance of the activities carried out within the family to the reproduction of social life in the wider society, these contracts typically embody social norms and beliefs about the meaning of the family in different cultural contexts. And given the importance of gender relationships in the key activities of the family, contracts between women and men within the family typically embody social norms and beliefs about gender differences in the wider society.

As long as household members have more to gain from their membership of the household than from going it alone, they have an incentive to abide by the terms of intrahousehold contracts, including its rules about decision-making. However, when these gains are less certain, how household members respond, whether or how they seek to achieve more favorable outcomes, will depend on what they are stand to gain from such action compared with what they stand to lose. Institutional approaches to cooperation and conflict within the household differ from neoclassical approaches in two important ways. First of all, they explain the rationale for cooperation within the family in affective and contractual terms rather than purely instrumentalist ones. Household members do not make decisions and seek to bargain as unrelated individuals, differentiated only by their resource endowments, but as members of families, with ascribed roles and responsibilities. Familial connections also mean that what individual members value about their membership of particular households, and hence what they stand to lose, cannot be captured by a materialist calculus alone. Nelson (1996), for instance, suggests that along with concerns with material well-being, household members will also seek to balance their need for 'affiliation,' the need to love and belong, with the value they attach to 'agency,' the capacity to define and act in their own interests. These different dimensions of well-being often involve trade-offs so that an individual may gain greater 'agency' by withdrawing from the household but have to sacrifice affectivity or material well-being.

At the same time, there may be situations where members may feel strongly enough about a decision to seek to influence its outcome. Because such action can, in certain cases, lead to open conflict, the likelihood of it happening will be determined partly by the extent to which the member, or members, in question feel able to jeopardize their membership of the household, in other words, on the strength of their fall-back positions. The second important difference between institutional and neoclassical approaches to bargaining relates to how they conceptualize differences in fall-back positions. While the primary focus of neoclassical approaches has been on differences in individual economic resources, institutional analysis emphasizes the much wider range of constraints and possibilities which reflect the household as a specific institutional form and create social differentials among individual members, or categories of members. These points have been articulated most clearly in the literature dealing with the gender dynamics of household decision-making and will be elaborated by drawing from this literature.

4. *Gender and the Structure of Constraint*

Neoclassical approaches to household economics tend to treat differences in resource endowments as the product of individual efforts or as 'given' by genetic features like the sex of individuals. Institutional approaches, by contrast, suggest that these differences are influenced significantly by social rules and norms which prescribe the distribution of resources, the division of labor and responsibility for decision-making within the household. While it is largely men who are favored by these rules and norms, gender inequalities vary considerably across societies, leading to wide variations in whether, and how, women have influence in household decision-making. For instance, most inheritance laws and customs favor male members of family but vary considerably in how they treat women. In some societies, women can own land and property in their own right. In others, they have rights of access or use, but only as members of particular households. In yet others, they are effectively propertyless. Similarly, social norms tend to assign primary responsibility for domestic work to women in most societies, but vary considerably in whether they permit women to engage in economic activities as well. In societies that practice female seclusion, for instance, women are proscribed culturally from undertaking public forms of employment. They are consequently likely to be far more dependent on male members of the household, and their capacity to influence household decision-making processes correspondingly curtailed, compared with societies where women have a socially-sanctioned role in production.

Finally, there are also cross-cultural variations in the access that women enjoy to extrahousehold social networks which will also influence the extent to which they are dependent on their immediate household members for provision and security. For example, in the Indian context, the ability of women to retain contacts with their natal families after marriage has been found to be an important factor in determining

the degree of agency they are able to exercise within marriage.

These variations in the contractual relations of the family create very different sets of gender interests in different social contexts and very different possibilities for women, as the subordinate category, to pursue their own interests. In contexts where women have recognized access to resources and economic opportunities, as in parts of sub-Saharan Africa, the relationship between spouses is often characterized by an overt bargaining element. Attempts to infringe on women's economic autonomy can result in open conflict, often ending with women leaving their husbands to set up their own households. By contrast, in many parts of South Asia, where women are denied access to critical productive resource or the ability to engage in economic activities, there is less scope for overt bargaining with more powerful male members. Conflicts are likely to be suppressed and dissent expressed covertly through what have been described as the 'weapons of the weak': deception, manipulation, and subversion.

5. *Gender and the Structure Of Preferences*

Neoclassical approaches to household decision-making do not pay a great deal of attention on what motivates people in their behavior. It is assumed that each member will have their individual preferences, depending on what gives them 'utility' or satisfaction. Because these are subjectively determined, they will be idiosyncratic to individuals and distributed randomly across the membership. However, evidence that women tend to behave in systematically different ways from men within the household, that they are more likely to put the collective interests before their own and to allocate resources under their jurisdiction to collective, rather than individual forms of consumption, has been given different explanations.

One set of explanations focus on differences in interests and fallback positions. Thus, although women often appear to give value to the collective well-being over their own as a result of prevailing ideologies of maternal altruism, Whitehead's (1981) work in Ghana leads her to suggest that such altruism often contained a dimension of self-interest. Women's fortunes are bound up with the fortunes of the household collectivity to a larger extent than those of men and their longer-term self-interests better served through altruistic forms of behavior which helped to preserve household solidarity. According to Agarwal (1997), evidence from South Asia that women tend to subordinate their own well-being to the well-being of others reflected their awareness of the weakness of their fall-back positions and a resulting caution in acting overtly in their own self-interest. An alternative set of explanations suggest that these differences in behavior reflect actual differences in the values and priorities, differences which in turn reflect social definitions of gender identity. Social norms and beliefs do not simply determine gender differences in resource endowments; they also justify and legitimate these differentials by ascribing very different aptitudes and dispositions to men and women, differences which are internalized by them as aspects of their gender identities.

For instance, studies in the US and Europe suggest that the processes of socialization through which men and women acquire their gender identities tend to stress agency and 'separative' notions of selfhood for men, and affiliation and 'connected' notions of selfhood for women (Chodorow 1978). In the South Asian context, it has been argued that gender identities are linked to differentiated notions of individual self-worth (Sen 1990). Women are seen to accept the lesser status accorded to them by a society and to define their own sense of well-being in terms of the well-being of other family members.

If men and women within the family have different values and preferences, bargaining can take the form of trade-offs between different dimensions of individual well-being noted earlier. Thus, a wife may agree to love, honor, and obey her husband in exchange for his promise to love, honor, and protect her: both value affiliation but the wife trades some amount of agency in return for material security. Alternatively, a husband may accept greater agency for his wife (and perhaps less for himself) by agreeing to her paid employment but only on condition that she does not neglect the housework and hence his material comforts. However, these negotiations and trade-offs will not be conducted on equal terms because the gender differentials in values and priorities do not have symmetrical implications for the bargaining process. If women do indeed attach greater value to affiliation or less value to their own self-worth, it is likely to work to men's advantage in situations of conflict. First of all, women will be less willing to disrupt household solidarity by open expression of dissent. Second, even if they do decide to engage in explicit bargaining, they will find it harder to hold out for their own self-interest or to threaten to use the exit option, if things do not go their own way.

6. *Conclusion*

Institutional approaches to family bargaining differ from neoclassical bargaining approaches because they focus on households as sets of social relationships, rather than a collection of individuals, on the structural determinants of bargaining power rather than the individual, and on the role of norms and values in shaping preferences rather than taking preferences as given. They consequently offer important insights into the specificities of households as social institutions and into what makes bargaining within the family different from bargaining in other arenas of life. Because

families are constructed around relations of unequal, but intimate, interdependence, power within the family is rarely exercised in a naked form. Violence, and the threat of violence, exist, but more often, men are able to resort to the authority vested in them in their capacity as household heads or as primary breadwinners to protect their own interests. The association of authority with male roles within the family explains why decision-making roles are not reversed automatically in households with unemployed husbands and employed wives or where women earn more than their husbands.

Equally, however, despite their stronger fall-back position, men within the family do not always get their own way. Bargaining processes within the family encompass forms of 'threats' as well as 'persuasions' that cannot be reduced to the economics of fall-back positions. They reflect the fact that the family is an intensely personal arena of life and that family relationships are suffused with feelings and emotions to a much greater extent than any other institution in society. Because families are based on ideologies of affectivity and shared interests, bargaining within the family rarely takes the form of open self-interest. Instead, it is conducted as negotiations over meanings and interpretations of, for instance, what constitutes the collective interests of family members, what their proper roles and responsibilities should be, and the basis of claims on collective household resources.

This means that, despite their subordinate position, women have a valuable resource on which they can draw in devising their bargaining strategies: their intimate knowledge of other family members, of their feelings and emotions, as well as their values and priorities. This knowledge can be translated into various forms of 'threats,' including persistent complaining, ridicule, withdrawal into silence, and withholding sex from husbands. Alternatively, they may employ more persuasive tactics. These may be discursive in form, for instance, appeals to shared interests in children's welfare as the justification for pursuing some preferred course of action. Or they may be more practical, involving strategies of 'wielding and yielding,' trading concessions in some areas in some areas to win concessions in others.

However, the fact that, despite the weakness of their fall-back positions, women are able to engage in a variety of bargaining maneuvers to achieve their priorities, should not be taken to negate the importance of improvements in their fall-back positions. Such improvements can bring about qualitative transformation in the asymmetries of the bargaining process. They can, for instance, lead to more open forms of bargaining between equals in place of the covert tactics that make up the 'weapons of the weak.' They can also open up for negotiation and contestation of aspects of household inequality that were previously considered nonnegotiable, for instance, in the domestic division of labor.

Research is needed to establish which are the most critical opportunities and constraints which make up women's fall-back position in different societies, how changes in these opportunities and constraints come about, and what role can be played by policy and public action to facilitate the right kind of changes. Women's paid activities has been found to be one of the most consistent predictors of their capacity to exercise agency within the household, both in advanced industrialized countries as well as in poorer ones. However, their ability to take advantage of employment opportunities is itself a matter for bargaining and negotiation, particularly in societies which practice female seclusion. The fact that women have, nevertheless, been able to do so suggests that other factors also came into play in negotiating these opportunities. Economic resources are important, but they may require changes of the kind brought about by enabling policies and social movements before their transformatory potential can be realized.

In addition, research is needed on the pathways by which changes in the wider structures of constraints and opportunities transform the unequal interdependencies of the family. It is possible that they will vary considerably according to local beliefs and values. In cultures that are strongly individualist in their outlook, women may use improvements in their fall-back position to seek greater independence, perhaps exercising their exit options. On the other hand, in cultures that emphasize collective values, they may prefer to bargain for greater equality in decision-making, the democratization of household structures. Attempts to evaluate the transformatory potential of different kinds of change in women's fall-back position will 'misread' the nature of the process if they are premised on an inappropriate set of values.

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict/Consensus; Domestic Life and Material Culture; Economic Development and Women; Exchange in Anthropology; Exchange: Social; Family and Gender; Family and Kinship, History of; Family Theory: Feminist–Economist Critique; Feminist Economics; Gender, Economics of; Group Decision Making, Cognitive Psychology of; Household in Anthropology; Household Production; Infancy and Childhood: Emotional Development; Kinship in Anthropology; Land Rights and Gender; Love and Intimacy, Psychology of; Marriage and the Dual-career Family: Cultural Concerns; Rural Industrialization in Developing Nations and Gender; Temperament and Human Development

Bibliography

Agarwal B 1997 'Bargaining' and gender relations: Within and beyond the household. *Feminist Economics* **3**(1): 1–51

Becker G 1981 *A Treatise on the Family*. Harvard University Press, MA

Chodorow N J 1978 The reproduction of mothering: Psychoanalysis and the sociology of gender. University of California Press, Berkeley, CA
Dwyer D, Bruce J 1988 *A Home Divided: Women and Income in the Third World*. Stanford University Press, Stanford, CA
Hoddinote J, Adam C 1998 Testing Nash bargaining household models with time series data: Divorce law reform and female suicide in Canada. *Food Consumption and Nutrition Division, Discussion Paper 52*. International Food Policy Research Institute, Washington, DC
Jones C 1983 The mobilisation of women's labor for cash crop production: A game theoretic approach. *American Journal of Agricultural Economics* **65**(5): 1049–54
Kabeer N 2000 *The Power to Choose: Bangladeshi Women and Labor Market Decisions in London and Dhaka*. Verso, London and New York
Kandiyoti D 1988 Bargaining with patriarchy. *Gender and Society* **2**(3): 274–90
McElroy M B 1990 The empirical content of Nash-bargained household behavior. *Journal of Human Resources* **25**(4): 559–83
Momsen J H, Townsend J 1987 *Geography of Gender in the Third World State*. University of New York Press, Albany, NY
Nelson J A 1996 *Feminism, Objectivity and Economics*. Routledge, London
Schultz T P 1990 Testing the neoclassical model of family labor supply and fertility. *Journal of Human Resources* **25**(4): 599–634
Sen A K 1990 Gender and cooperative conflicts. In: Tinker I (ed.) *Persistent Inequalities*. Oxford University Press, Oxford, UK
Thomas D 1993 The distribution of income and expenditure within the household. *Annales D'Economie et de Statistique* **29**: 109–35
Whitehead A 1981 I'm hungry, Mum: The politics of domestic budgeting. In: Young K, Wolkowitz C, McCullagh R (eds.) *Of Marriage and Market: Women's Subordination in International Perspective*. CSE Books, London

N. Kabeer

Family Business

Family business is an interdisciplinary field involving all professionals who advise, study, and work in family firms addressing their structural and transitional problems. Family firms are companies of various sizes controlled by one or more owners tied by family relationship or solid alliances (see Corbetta 1995, Aronoff et al. 1996, Neubauer and Lank 1998). Even if this control-based definition is widely accepted, some authors propose others based on the owners' managerial involvement. Other authors do not consider either firms controlled only by the founder or those controlled by unrelated partners as family businesses. Another group tends to identify family businesses with small and medium-size enterprises where the overlapping of family and firm is particularly high. Whichever the accepted definition may be, it includes various types of family firms that can be classified using different dimensions. Family business specialists' fields of origin are equally varied and include such disciplines as business, law, psychology, sociology, and history.

A study carried out on the articles published in *Family Business Review* from 1988 to 1997, shows that over 70 percent of the academic articles were written by authors in the field of business management (Dyer and Sànchez 1998). In this discipline, family business literature mainly overlaps with studies on entrepreneurship and management issues in small and medium enterprises (SMEs) (see *Entrepreneurship*). However, it distinguishes itself from such studies on at least three levels. First of all, family business literature investigates family firms' evolution from succession to the founding owner up to following stages in which the firm may have become very large and even public but it may still be controlled by one or few families tied by family relationship or solid alliances (Gersick et al. 1997). Second, family business management literature deals with the entrepreneur's characteristics and entrepreneurship training and developing processes in a peculiar way.

On one hand, these topics are always related to owning families' characteristics and dynamics, and on the other hand much more attention is devoted to the issue of how to pass on entrepreneurial capabilities rather than on how to start a new venture. Finally, family business literature deals with topics that are not investigated by the streams mentioned previously, such as managing the ownership group, governing the family firms and managing the family–firm relationship.

Family businesses are dynamic systems that include two subsets, the family and the firm, which are characterized by their own evolutionary patterns and are linked to each other by many dynamic relations. Hence a scholar may be considered a scholar at family business when they deal with the whole system: a psychologist interested in parents–children relationships addresses family business issues only when they consider how these relationships' dynamics impact the way the business operates. A business management scholar interested in strategy deals with family business only when investigating the relationships between the owning families' structure and evolution and business continuity.

1. *Origins and Development of Family Business*

Even if some family business articles appeared in the 1950s, 1960s, and 1970s, only at the beginning of the 1980s was family business born as a field acknowledged by a number of North American scholars and practitioners. A fundamental contribution was made by a management researcher (Richard Beckhard, Management and Organization Behavior Professor at

MIT Sloan School of Management for more than 20 years), by a psychology researcher (Harry Levinson, professor at the Psychiatry Department at Harvard Medical School), and by a consultant (Léon Danco, CEO of the Center for Family Business in Cleveland). In the mid-1990s they have been recognized publicly as family business pioneers and interviewed in the 1994 *Family Business Review* issue (Vol. 2) dedicated to the field's history; in these interviews they all admit to how difficult it was to make the scientific community accept the existence of family business as a special discipline.

Until then business and family (and related studies) had been dealt with as two separate, autonomous fields: it was difficult to convince scholars in either that it would be useful to get acquainted and cooperate. Moreover, the first studies in family business were case descriptions without any interpretative framework, which made it even more difficult to open a scientific debate. Finally, most family business professionals were practitioners at that time; as such, they were not well respected as researchers by the scientific community. Further, they were not very committed to publishing in order to share their knowledge and experiences with others.

Generally speaking, family firms traditionally have been considered—and not only by academics—as a residual of the past to overcome, beginning from the study by Berle and Means (1932) who considered separation between ownership and control as a fundamental condition of development for modern economies. Not surprisingly, the media were used to dealing with family businesses only to highlight how disputes among family owners or nepotistic behaviors negatively impacted the business.

In 1984 something of decisive importance in the development of family business occurred: a group of North American scholars and practitioners, coordinated by Barbara Hollander, gave rise to the Family Firm Institute (FFI). By means of its various services, FFI has become a special forum for knowledge sharing and development. Particularly, annual conferences and the *Family Business Review*—founded by Ivan Lansberg in 1988—have been fundamental tools that helped develop theory and practice. In the leading article of the first issue, Lansberg himself provided significant data about how early the stage of the field's development was: according to a study conducted on the Social Science Citation Index from 1977 to 1987, only 53 articles had been published in the US on family business issues, as compared with the 813 on entrepreneurship.

In 1990, the Family Business Network (FBN) was founded in Europe. Despite its different membership policies (there are more family business members and less professionals), FBN fulfills a similar mission by spreading family business knowledge and fostering a scientific and cultural debate on this topic.

The increasing number of monographic issues on family business published in such reviews as *Organizational Dynamics* (summer 1983), *Business History* (October 1993), and *Entrepreneurship Theory and Practice* (N. 1, 1994) shows that rising attention has been devoted to this new discipline by scholars in other fields as well.

In the US and in other parts of the world, some universities followed the increasing interest on the part of scholars and other stakeholders towards family business. Nowadays more than 100 universities and business schools in the world offer specific research and training programs on family business issues (Aronoff 1998). Most of these programs have been developed in family business centers, which—even if they are often based in management departments—usually get contributions from various disciplines.

Different factors have persuaded individuals and institutions to invest resources in exploring family business. Among these is the diffusion of data about the importance of family businesses in modern economies, evidence that family firms can grow big and cannot just be identified with SMEs, and consciousness of the difficulties for family firms to last beyond the first generation. Depending on the adopted definition, family firms represent from 30 percent to 70 percent of the GDP in the main industrialized countries, and above all, they seem to be the enterprises with the highest rate of growth in employment.

Looking into the future, family business seems to be in a delicate phase. In order to gain importance in the academic context this new discipline needs to be acknowledged from an institutional point of view as well; for this reason, a key challenge—at least in Europe—is to increase the number of universities and business schools that create chairs in family business. Probably this will happen mainly in management departments, as it occurred in the closely related field of entrepreneurship.

Concerning the practitioners' sphere, the family business field is growing very rapidly, with many advisors coming from a variety of professions and entering the field. The main problem in the near future is the difficulty in defining the minimum standards necessary to qualify people as family business professionals. The FFI is working on a research project that started with identification of four major content areas—Behavioral Science, Finance, Law, and Management Science—and is going on with the definition of fundamentals that competent family business advisors, consultants, and educators should know to work effectively in those subfields.

2. *Research Areas and Achieved Results*

Family businesses are a composite world including different archetypes. Many authors have proposed classification criteria. According to a widely accepted distinction there are family firms owned and managed either by a single owner or by two or more siblings or

by cousins' teams. A fourth archetype consists of those family firms where equity has been opened up to partners other than the founding family (Corbetta 1995). Usually, changing from one type to the others the company grows in size. But notable exceptions make it impossible to significantly correlate company size with ownership structure. The distinction mentioned here cannot be confused with the generation-based perspective, for example, first generation family enterprises can be controlled either by a single owner or by a sibling partnership or by a cousin consortium.

For every archetype identified, a variety of research topics can be defined, such as:

(a) family structure and management;
(b) ownership structure and management;
(c) company governance;
(d) company management;
(e) relations among family, ownership, governance and management; and
(f) transitions into another type of family business.

Other general topics of research concern the comparison among different types of family businesses, between family and nonfamily businesses (in terms of values, culture, strategies, and performances), among family enterprises operating in different countries, the relationship between enterprises and consultants, and the role of women.

In terms of necessary competencies, many family business scholars believe that most research topics need interdisciplinary team effort; in particular, researchers in management, psychology, and sociology should cooperate to capture the distinctive facets of family business.

2.1 Studies Before the 1980s

The first family business studies mostly concerned founder succession issues, that is either transition from the controlling owner archetype to another or recycling of the same archetype with different actors (Danco 1975). Such studies typically had the following characteristics:

(a) they concerned small and medium sized firms with simple management structures;
(b) they took into consideration only the simplest family situations—one founder, one heir—overlooking more complicated but also more widespread circumstances, for example, those of founding partners' teams or those where the transition occurs between multiple founders and multiple heirs;
(c) they centered their attention on the two people involved in leadership transition—the founder and the heir—without considering other family and business members;
(d) they almost exclusively dealt with the issue of passing on managerial roles overlooking ownership changes; and
(e) they considered the family as an obstacle to the effective and efficient management of the firm and suggested a clean distinction between family and enterprise as an indispensable condition for business continuity.

Even if many of these studies, which were clearly influenced by some streams of rationalism (Hollander and Elman 1988), have been revisited or even become outdated, they have actually achieved some important results. First of all, they have opened up a new, interesting stream by explicitly recognizing the differences—in terms of roles, goals and needs—and the mutual influence between the family and the business. Moreover, these studies suggested some interpretative frameworks still valid today such as: succession considered as a multiphase process and not an event; the difficulties of this transition coming from the variety of roles—parent, owner, and manager—simultaneously played by the founder; the importance of training for people; the need to consider the firm's continuity as a fundamental goal to be achieved to be able to consider generational transitions as successful ones.

Many of the first scholars were focused on small and medium sized enterprises, which made them also study managerialization processes in relation to company growth patterns.

2.2 The Subsequent Studies

In the following years the family business field was consolidated; its boundaries were clarified and new topics were identified. In particular, several contributions concerned the following:

(a) the definition of family businesses and their classification, going beyond complete overlapping with small and medium sized enterprises;
(b) the definition of the main dynamic relationships between family and business using interpretative frameworks based on different stages of the firms, people, and families' life cycles (Davis and Tagiuri 1982, Beckhard and Dyer 1983, Gersick et al. 1997);
(c) succession processes, widening the range of considered situations and variables; and
(d) new topics such as management of extended families, family business governance, and strategy.

Concerning the first topic, numerous studies have been focused either on the role and way of implementing family protocols or on the way of functioning of specific bodies like the family council. With regard to the second topic, attention has been centered on the structure and the functioning of boards of directors (Ward 1991, Neubauer and Lank 1998). On the third topic, interesting contributions have been published on internationalization processes and strategic management in general (Ward 1988).

During this period, scholars have reached a greater consensus about the definition of family business and about the classification of different family firms' archetypes. The field has experienced significant development and has involved scholars from other

disciplines (e.g., finance and corporate governance), which has generated new research opportunities.

One very important achievement is that the family is no longer considered an obstacle to the firm's development. Longitudinal studies of family business cases and quantitative researches definitely have shown that the owning family can be a strong point for the firm as it provides such key factors as long-term orientation, education to entrepreneurship, values, commitment and personal sacrifice, reputation, speed in decision-making, and lower transaction costs (Gallo 1995). Social scientists' studies are considered an important basis to build up theories and models that can be helpful for family business success in the long run. Theories on family life cycles and adult development have been used to analyze family firms' stages of development, which are seen as the joint result of family, ownership, and business dynamics (Gersick et al. 1997). Family systems' theories are used to suggest how to better manage interpersonal relations, communications, and to handle conflicts through their predictable stages in the family business context (Barnes 1988, Kaye 1991). Interpretative models have recently been proposed considering the family just as important as the company as a determinant of family business success; so, both institutions are included in the model at the same level of detail, as purposive social systems transforming resources and constraints into outcomes by interpersonal and resource transactions (Scannel Trent 1999). The same studies have investigated family firms' diffusion by surveying households instead of companies.

Concerning succession, which is still the hottest topic in family business literature, research findings highlight the need to plan the transition: the importance of successors' personal development both inside and outside their own family firms; the critical role of the entrepreneur and of their ability to delegate; the importance of communication; and the useful contribution of nonfamily members either as outside directors or as tutors and mentors to the younger generation. Efforts have been made to broaden the focus of research: a variety of situations has been addressed in addition to the 'simple' one featuring one entrepreneur with only one heir; ownership transitions and their financial implications have emerged as relevant succession topics besides the classical one concerning company leadership. Efforts have also been made to offer deeper insights on the 'classical' issue of succession to the founder. Numerous studies have built upon the existing literature by exploring the complexities that characterize this type of succession as well and that had been overlooked so far. Relationship between founder and the whole organization, and how organizational culture can be created and passed on so as to enable company continuity have been investigated (Schein 1983). The broad range of stakeholders involved besides the founder and the successor involved has been analyzed, highlighting how each of them may resist coping with the generational transition issue and by this means create obstacles to the whole process (Lansberg 1988). The issue of how entrepreneurial attitudes and capabilities can be learnt and taught has been further explored (Aronoff et al. 1996).

Finally, relevant insights have been provided on family business governance, particularly on the structure and the functioning of boards of directors and on their contribution to company development.

2.3 Future Developments

A recent article that appeared in the *Family Business Review* has identified some trends in family business studies (Aronoff 1998). Succession will remain one of the most interesting areas in the broader perspective of generational transition. A new research topic could concern possible roles for retiring entrepreneurs. Generational transitions will increasingly lead to both owners and family managers' teams; conditions under which such teams can effectively work will be one of the hottest topics in the nearest future. Companies' strategic context is significantly changing because of the impact of new technologies and globalization. Family businesses' capability to compete in such a context will make a relevant research topic as well. Another key issue will be the impact of increasing managerial and financial sophistication on both the culture and behaviors of family businesses. The other topics mentioned by Aronoff concern the roles of women, consultants, and educators.

Interesting developments may also come from the increasing focus on big family businesses that characterizes some studies in progress. Certainly this will lead not only to identifying and offering more sophisticated solutions to the already well-known problems of smaller family businesses, but also to investigating different ways to govern and manage large family fortunes in terms of both personal attitudes and necessary skills.

Studies on family businesses' developmental stages have not yet deeply explored transitions from one stage to another. Relevant insights on this topic could be provided by joint efforts of management and psychology researchers.

Finally, other topics are likely to undergo important developments. We can mention, for instance: comparisons among different archetypes of family firms, among family businesses from different countries, and between family and nonfamily enterprises in terms of resources, strategies, and performances; nonfamily managers' characteristics and roles; family businesses' selling processes; and the impact of some variables on family firms' continuity. Research efforts could also address family values and ownership structures.

Research topics still seem to be driven by both family enterprises and family business professionals, necessary because they are the main sponsors of

research projects. On the other hand, the strong link with family businesses is helpful in making the relationship between theory and practice mutually beneficial. Obviously scholars should avoid focusing on short-term needs and commit themselves to identifying long-term research goals and strategies.

3. Methodological Issues

Practitioners, as it was stated before, have played a predominant role in the formation of family business as a discipline. At least until the early 1980s, as a consequence, research methodologies were not emphasized. Even those scholars who have been in the field from the very beginning were much more concerned in defining boundaries and topics rather than in carrying out methodologically acceptable research projects. In the first issue of the *Family Business Review*, in 1988, only 14 percent of articles were research based. Notably, 64 percent of the articles published in the 10th issue of the same review reported qualitative or quantitative research findings (Dyer and Sànchez 1998).

In any case, family business research still has to cope with at least three specific problems. The first and most important one is that family business, as should be evident, requires that contributions be provided by professionals from different disciplines, particularly social sciences such as sociology, psychology, and business management. On one hand, it is not easy that a single person dominates both; at the same time, interdisciplinary collaboration is not well developed, probably because of the high costs of joint research due to the very different starting points in terms of knowledge, languages, and methodologies. On the other hand, since in many fields radical advances come from collaboration between authors in different disciplines and professions, such collaboration should be strongly encouraged.

The second problem concerns information availability. Entrepreneurs and their families are known to be very reserved and protective of their privacy. Moreover, most family businesses do not have to disclose much information about their structure or about their performances. Researchers are trying to overcome this problem by collaborating with institutions such as banks that can provide family business databases or by designing research projects based on primary sources and making use of interviews, questionnaires, and longitudinal case studies.

The third relevant problem regards the specific need of this field to encourage collaboration between scholars and practitioners. Increasing interest in family business among younger Ph.D.'s is definitely improving the quality of research. Some difficulties could arise as young scholars often focus their attention on topics that are not considered particularly relevant by family businesses, which may hinder communication between academics and other stakeholders. In other fields such as entrepreneurship, such incommunicability has led to very disappointing results (Brockhaus 1994).

Finally, given the interdisciplinary nature of the family business field, relevant developments might result from the application of theoretical frameworks and analytical tools developed in other knowledge areas such as the system dynamics approach (which might be useful in explaining the dynamics of the family-business system) or the resource-based view (which might lead to a better understanding of the relationships between the business and the family's resources and the firm's performances).

See also: Business History; Entrepreneurship; Entrepreneurship, Psychology of; Ethnic Groups/Ethnicity: Historical Aspects; Family as Institution; Firm Behavior; Retail Trade

Bibliography

Aronoff C E 1998 Megatrends in family business. *Family Business Review* **3**: 181–5

Aronoff C E, Astrachan J H, Ward J L 1996 *Family Business Sourcebook II*. Business Owners Resources, Marietta, GA

Barnes L B 1988 Incongruent hierarchies: daughters and younger sons as company CEOs. *Family Business Review* **I**: 9–21

Beckhard R, Dyer Jr. W G 1983 Managing continuity in the family-owned business. *Organizational Dynamics* **12**(1): 5–12

Berle A, Means G C 1932 *The Modern Corporation and Private Property*. Macmillan, New York

Brockhaus R H 1994 Entrepreneurship and family business research: Comparisons, critique and lessons. *Entrepreneurship Theory and Practice* **19**(1): 25–38

Corbetta G 1995 Patterns of development of family businesses in Italy. *Family Business Review* **4**: 255–65

Danco L A 1975 *Beyond Survival: A Business Owner's Guide for Success*. University Press, Cleveland, OH

Davis J A, Tagiuri R 1989 The influence of life stage on father–son work relationships in family companies. *Family Business Review*, Spring

Donnelley R G 1964 The family business. *Harvard Business Review* **42**(3): 93–105

Dyer Jr. W G, Sànchez M 1998 Current state of family business theory and practice as reflected in family business review 1988–1997. *Family Business Review* **4**: 287–95

Gallo M A 1995 *Empresa Familiar*. Textos y Casos, Editorial Barcelona

Gersick K E, Davis J A, McCollom M, Lansberg I 1997 *Generation to Generation: Life Cycles of the Family Business*. Harvard Business School Press, Boston

Hollander B S, Elman N S 1988 Family-owned businesses: an emerging field of inquiry. *Family Business Review* **II**: 145–64

Kaye K 1991 Penetrating the cycle of sustained conflict. *Family Business Review* **1**: 21–44

Lansberg I 1988 The succession conspiracy. *Family Business Review* **2**: 119–43

Martinelli A 1994 Entrepreneurship and management. In: Sunclser N J, Swedlord R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ

Neubauer F, Lank A G 1998 *The Family Business. Its Governance for Sustainability*. Routledge, New York

Scannel Trent E (ed.) 1999 Family business from the household perspective. *Family Business Review Special Issue* **3**: 197–281
Schein E H 1983 The role of the founder in creating organizational culture. *Organizational Dynamics* **12**(1): 13–28
Ward J L 1988 The special role of strategic planning for family businesses. *Family Business Review* **2**: 105–17
Ward J L 1991 *Creating Effective Boards for Private Enterprises: Meeting the Challenges of Continuity and Competition.* Jossey-Bass, San Francisco

G. Corbetta

Family Health

The very diversified topic of family health is currently receiving increasing attention from different disciplines and different directions within psychology. This is especially evident in the areas of clinical and family psychology, gender-specific research, and health psychology (see *Health Psychology*). In this article, important theories and empirical results from these fields are summarized in a systematic overview which is divided into the following sections: (a) the bidirectional connections between family dimensions and the health of family members (e.g., the mother), (b) systemic concepts like 'psychosomatic families' or 'family mental health,' (c) family resources in terms of coping capabilities or 'salutogenetic' and regenerative variables of family life, and (d) differential aspects of family health, that is, differences between the developmental stages of the family life course, differences between family structures, and main results from gender-specific research.

1. Introduction

The article will first of all locate the topic of 'family health' within some main psychological traditions, before a systematization of the topic is proposed and selected empirical results are described briefly.

The field of family health can be seen at the crossroads between the disciplines of family psychology and health psychology. The connections between health and family factors are numerous, varied, and important. Nevertheless, the fields of 'family health psychology' and 'family systems medicine' are still in the early stages of development, although family psychology and health psychology are established disciplines (Akamatsu 1992).

Despite the fact that the family is the basic social context in which health behaviors are learned, clinical and health promotion theory and research traditionally have focused primarily on individual behavior. In contrast, family psychology has been studying interindividual relationships or 'systems' from its beginnings. Yet, within health psychology, family systems considerations are currently gaining recognition.

As health is a very important value for most individuals and families, it has always been of major interest within the discipline of family psychology, though not called 'family health.' Important contributions in research on family and health can be accredited to family stress research. In the beginning, this field of study focused on the familial crises brought about by major life events. At the same time, studies on the psychological processes of individuals were also 'pathogenetically' oriented. Antonovsky's 'salutogenetic' approach offered an alternative to the pathogenetical perspective.

Antonovsky (1979) builds on earlier developments of Humanistic Psychology and the theory and practice of prevention, and he focuses on health-promoting factors and processes as well as on the characteristics of extremely healthy people or systems. In contrast to earlier studies on health and on the family, the current disciplines of family and health psychology are paying more and more attention to these salutogenetic views and models (see Olson and Stewart 1991).

The early notions of optimal family functioning were inferred from experiences with families or family members in treatment. As the practice of family therapy became institutionalized, integrative models of family health were developed increasingly. Generally speaking, theories and models of family psychology that can be regarded as unequivocally salutogenetic, either (a) conceptualize a health ideal of families, or (b) consider health processes within the context of individual or familiar resources that contribute to the state of health of the whole family or its individual members.

In the following sections, health is conceptualized as a state of total physical, psychological, and social well-being, and not only the absence of disease and disability (see Antonovsky 1979). As this definition contains a physical, a psychological, and a social component, it meets the requirements of a 'biopsychosocial' perspective in health psychology (see Schwarzer 1997).

Psychological definitions of family point out that a family is a special group of persons living in a close personal relationship, in the past, in the present, and in the future. It is an economic unit bound together by emotional ties (see Schneewind 1991).

In the following, the two topics of family and health are combined and a systematic overview of the field of 'family health' is offered.

2. Family Health: A Systematic Overview and Selected Empirical Results

The very diversified subject of family health was dealt with by totally different directions in theory and research, such as stress research or feministic research.

Table 1
Systematic overview of essential facets of family health in theory and research

1. *The family and the health of individual family members or dyads*
   (a) Family stress and the health of individual family members
   (b) The family as context for socialization
   (c) Intergenerational theory of personal development within the family system: the family of origin in its relevance for the current health of its members
   (d) Influences of a chronically or acutely sick individual upon other family members
2. *The family as resource and health-promoting institution*
   (a) Family coping; family variables as 'stress buffers'
   (b) The couple and family as direct resource and source of satisfaction
3. *Differential aspects of stress and health within the family*
   (a) Gender-specific family psychological research regarding health and stress
   (b) Different family structures, e.g., two-parent vs. divorce- and stepfamilies
   (c) Health and stress in different phases of the family life course; intergenerational differences in health variables
4. *Systemic view of family health*
   (a) Family health in the form of 'family competence' or 'family mental health'
   (b) Family ideals
   (c) Health-relevant characteristics of dyads within the family or of family structure
   (d) Systemic interventions

Note: The numbering 1 to 4 is parallel to the order in the Sects. 2.1 to 2.4

Following the systematic overview in Table 1, some of these different aspects of family health will now be described briefly.

2.1 The Family and the Health of Single Family Members

Studying the links between family characteristics and the health of individuals, one has to distinguish whether (a) the effects of certain family characteristics or situations on system members are being considered, (b) the effects of certain individuals on the family, or whether (c) interaction processes and the resulting effects are being taken into account.

Effects of familial characteristics or situations are studied intensively by family stress research. Within this field, the family is viewed as a source of stress or a mediating factor. 'Stress' is considered either as 'life events,' family transitions, caregiving, or economic strains or stressors resulting from family interactions. Furthermore, extremely adverse familial influences such as violence and abuse are examined. All these forms of familial stress can affect psychological and physical health. Yet the extent of negative effects is highly dependent on coping processes and on individual and familial resources (see Sect. 2.2).

The study of familial influences on child development leads one into the very wide field of socialization within the family. Generally, the family is regarded as the main matrix of personality development, laying the foundations for mental health in childhood. From a clinical perspective, there are certain familial factors which put a child's development at risk. Some of these risk factors are psychological disorders of a parent, parental delinquency, strong marital conflicts, and poor educational level of the parents. These variables have been found to correlate with emotional and behavioral disorders and academic failure during childhood. Moreover, childhood problems such as 'internalizing disorders' (e.g., depression) and 'externalizing disorders' (e.g., conduct problems) that stem in part from poor caregiving practices, are again strong predictors of mental disorders in adulthood.

What distinguishes effective from ineffective caregiving? Scientists have identified two fundamental dimensions of caregiving that are particularly important for children's adjustment. The first is, how much warmth, nurturance, and acceptance (vs. hostility and rejection) caregivers convey to children. The second is how much control, structure, and involvement (vs. permissiveness and detachment) caregivers display towards their children. Schneewind and Ruppert (1998), for example, demonstrate by longitudinal data that the family's atmosphere with the primary dimensions of positive emotionality, activation, and level of organization is a relevant predictor of family members' emotional, personality, and social development as well as the child's locus of control (see Schneewind 1995). Furthermore, studies have found that the family's emotional expressiveness also plays a role in the development of social competence in children. Intergenerational family systems theory of personal development within the family views well-being within the context of a multigenerational model. The quality and dynamics of these significant relationship patterns are conceptualized as having far-

reaching implications, not only for psychological functioning, but also for physical well-being, ability to cope with stress, and willingness to engage in positive health-related behavior. All of these assumptions are verified empirically. For practical reasons, nearly all of the studies rely on retrospective data, and most of them address only two generations. These studies, which consider the relationships between important psychological dimensions of the family of origin and the current health of family members, find, for example, moderate correlations between the degrees of autonomy and coherence, which were experienced by grown-up members of their family of origin, and their current psychological adaptation. Studies on the role of parental separation and divorce for child development have shown evidence of impairments of a child's social and personality development, mainly during the first year after the parent's separation. But the effects measured have been proven to be minimal, and the differences between children of intact families and divorced families almost vanish as time goes by.

The study of effects of individual characteristics on the family has a theoretical tradition in the field of family therapy, despite the transactional thinking of this systems approach. In earlier family therapy theories, a family in which an individual member became schizophrenic or developed a psychosomatic illness was known as a 'schizophrenic' or 'psychosomatic family.' Virginia Satir's experiential model of family therapy stresses the importance of the individual's self-esteem to the family's overall level of health. There are also empirical verifications of these interrelationships between self-esteem and perceived family functioning (e.g., Heaven et al. 1996).

This direction of effects from an individual's state of health on the couple or familial relationship has also been examined in cases of acute or chronic illness of a child, a parent, or a spouse. Using longitudinal data, Booth and Johnson (1994), for example, demonstrated that health impairments of a spouse diminish the quality of the couple's relationship quality, especially on the part of the sick person's partner. Nevertheless, the couple's or family's coping processes are very relevant for the system's adaptation to the situation.

2.2 *The Family as a Resource and Health-promoting Institution*

Although the family and its patterns of interaction can be an enormous stressor, scientists also know about the positive and regenerative aspects of the family. These 'family resources' are fundamental for the functioning of every society. They include a large variety of aspects, ranging from material resources, over social support, dimensions of recovery, and familial coping competencies, to 'daily uplifts' within the family's interaction.

An essential part of the health-promoting aspects of the family lies in the regenerative and protective effects of a good marriage or couple relationship (e.g., Gove et al. 1983). Even family status alone (single, married, separated, divorced, or widowed) shows significant connections with emotional well-being and life expectancy (e.g., see Cotten 1999, Gove et al. 1983). Despite a higher percentage of overweight persons among the married, living together with a spouse is on the whole health-promoting, especially for men. But these differences (between the groups and between the sexes) are diminishing currently. Furthermore, relationship quality is a moderating variable in the correlation between marital status and health, so that 'good,' satisfying marriages are more health-promoting (see Ross et al. 1990).

The correlation between socioeconomic status of families and mental and physical health appears consistently in the literature. Thus, the economic benefits within 'dual-income' families can be seen as a resource, as well as the emotional benefits which many women experience when they are working. These benefits could 'buffer' the stress of the 'double burden.'

Individual stress research has shown that the way people cope with stressors such as critical 'life events' is even more important than their occurrence (e.g., see Antonovsky 1979). Furthermore, there is evidence that also on the family level the quality of coping (e.g., with acute or chronic illness), is extremely relevant. Is the family able to 'buffer' the potentially negative effects of the stressor, or does the family use dysfunctional strategies which only serve to worsen the situation? Adequate coping is a resource which can lead to the establishment of other resources, which then can be useful in comparable or totally different situations in the future.

For couples or parents, the emotional support of a spouse is evidently very important (see *Social Support and Health*). Process analyses in the daily life of couples show that only in cases where a relatively high level of spouse support was evident does a decline in work-related stress occur during the evenings. Interestingly, allowing the partner to withdraw has proven to be one of the forms of spouse support (see Repetti 1989).

Furthermore, high quantity as well as quality of marital interactions has been connected with less risk-taking behaviors, such as less alcohol or drug use (see Wickrama et al. 1995). Similar findings regarding families demonstrate that the lifestyle of persons who live in families is more solid and healthier.

2.3 *Differential Aspects of Family Health*

Especially in the case of traditional role sharing, we expect gender differences in the experiences of family life and—in relation to these differences—in health-relevant variables.

Studies on the situation of the so-called 'dual income' families show that it is the women who experience a 'double burden' and difficulties in coordinating work and family life. In addition, for women the experienced inequality is more strongly connected with their couple satisfaction (see Kluwer et al. 1997). While women more often experience family stress in the form of their spouses' behaviors (e.g., not enough support), men more often feel stressed by their partners' criticism. Furthermore, as women tend to be supporting and do the larger portion of the household tasks, they have less time to 'unwind' after a stressful work day than their husbands (see Repetti 1989).

In summary, studies in Western nations have found that, although men show more risk-taking behavior compared to the more preventive behavior of women (Schwarzer 1997), women nevertheless have less favorable scores of physical and psychological health. This fact is to be seen in the context of the women's work and family roles, and the overload that is so often connected with these (e.g., see Barnett and Marshall 1991). This strain from work, parental, and couple domains has negative effects on women's well-being and health, and on the functioning of the family.

The lower health scores of older persons that were found by a large number of cross-sectional and longitudinal studies (e.g., see Schneewind and Ruppert 1998) can also be linked with family variables. Apart from biological factors, the stress of family transitions, such as disruptions of marital relationships through divorce or death, also bear aversive emotional and immunological consequences.

Another differential aspect of family health is the family structure or type. While it has often been thought that family structure (i.e., intact, divorced, stepfamily) was a key determinant of well-being, especially among children and adolescents, recent investigations have concluded that family process variables, such as conflict dynamics, cohesiveness, and support, are more important (e.g., Heaven et al. 1996; see above). Furthermore, cultural, social, and economic factors have widespread influences on the health of families.

Differential examinations of the different phases within the family life course show that the level of family stress is highest during the adolescent stage of family life, while family cohesiveness and satisfaction is lowest (see Aldous and Klein 1988).

2.4 *Systemic View of Family Health*

Scientists who are dealing with an overall, systemic concept of family health try to meet the demands of the systemic approach. They look at the topic in a more holistic manner and identify the structures and processes of the 'whole,' the family system which is embedded in other contexts (e.g., society). Conceptions that consider the health of the whole family are, for instance, the psychological constructs of 'family competence' or 'family mental health' (e.g., see Pruchno et al. 1994), as well as all family ideals.

In terms of 'family health' measures, there are some self-report assessment instruments and additional measures which are designed to quantify the quality of familial functioning from the perspective of an external clinical observer. Questionnaires that consider an overall systemic construct, for example, 'family mental health,' usually found their operationalizations on sum scores of the individual family members as well as on discrepancy scores between them. Families with higher discrepancies in their perceptions of family reality normally have a lower score on individual well-being. Obviously, converging realities in the subjective views of the system members are functional, whereas larger divergences are a characteristic of clinical families.

Another concept which is of interest, not only from a systemic, family-centered perspective, but also from the perspective of health psychology, is the one of 'family health cognitions': but until now, the empirical basis regarding the impacts of family cognitions on the use of health services and on personal health actions is weak and needs further attention.

An example for a family ideal is the 'balanced' family type within the 'Circumplex model' (Olson 1989). These families have relatively high scores in cohesion (but without being 'enmeshed'), and relatively high scores in their capability to adapt to new situations (without being 'chaotic'). These characteristics are connected with better communication, more favorable coping, and higher family and personal satisfaction and health.

A further systemic perspective of family health concerns health-relevant characteristics of the system structure and the characteristics of dyads within the family system. Such a link between systemic–structural characteristics and family health is represented by the variables of generational boundaries and parental coalitions. Clear—yet not too rigid—boundaries between the parental subsystem on the one hand and the children's subsystem on the other have proven to be a characteristic of healthy family systems. Similarly, coalitions between the parents, the 'architects of the family,' are more functional than strong transgenerational coalitions (e.g., mother–daughter).

Furthermore, a systemic view of the whole family's health is practiced in all family interventions, that is, in family-oriented prevention or health-promotion programs and in family therapies. The insight that the health of the whole family should be considered when one member has a symptom was the birth of family therapy. Family health-promotion programs help families to sustain or enhance the social, emotional, and physical well-being of the family system or its members (see Schneewind 1991). These different kinds of family interventions have become well institutional-

ized. Nevertheless, researchers should pay more attention to evaluations of these different family interventions.

3. Conclusions

Family life experiences affect deeply the competence, resilience, and well-being of everyone. The family shapes the quality of our lives, but—to a certain extent—we also shape the quality and health of our families, especially when we become parents.

Research is paying closer attention to salutogenetic family-centered variables of coping and health. Empirical evidence supports this trend. Yet, in the field of research a great amount of work still lies ahead in order clearly to establish an overall systemic view of the family, whereas in the domain of family therapy this perspective has been the major characteristic from the very beginning.

As patterns of marriage and family life are currently changing in the Western world, scientists should be aware that changes in the strength of connections between family factors and health variables can occur. In the case of the association between family status and health variables, a decline in correlations has already been observed.

Although the family is the basic context in which health behaviors are learned, further research studies should well consider the interactions between the family and other social contexts, like peer or work relationships. Influences from the side of these other persons and systems could, for instance, be one possible explanation for the fact that until now studies have found only rather weak links between 'family health cognitions' and health-related behaviors.

All in all, knowledge of the frames for a healthy development from and in families is very relevant for different branches of psychology. The application of this knowledge is especially important for family interventions, such as prevention programs, but also for family politics.

See also: Childhood Health; Family and Gender; Family and Psychiatric Illness/Family Therapy; Family, Anthropology of; Family Systems in Europe; Gender Role Stress and Health; Public Health; Social Support and Stress; Stress and Health Research

Bibliography

Akamatsu T J 1992 Family health psychology: Defining a new subdiscipline. In: Akamatsu T J, Parris Stephens M A, Hobfoll S E, Crowther J H (eds.) *Family Health Psychology*. Hemisphere, Washington, DC, pp. 239–50

Aldous J, Klein D M 1988 The linkages between family development and family stress. In: Klein D M, Aldous J (eds.) *Social Stress and Family Development*. Guilford Press, New York, pp. 3–19

Antonovsky A 1979 *Health, Stress and Coping*. Jossey-Bass, San Francisco, CA

Barnett R C, Marshall N L 1991 The relationship between women's work and family roles and their subjective well-being and psychological distress. In: Frankenhaeuser M, Lundberg U, Chesney M (eds.) *Women, Work and Health. Stress and Opportunities*. Plenum Press, New York, pp. 111–36

Booth A, Johnson D R 1994 Declining health and marital quality. *Journal of Marriage and the Family* **56**: 218–23

Cotten S R 1999 Marital status and mental health revisited: Examining the importance of risk factors and resources. *Family Relations* **48**: 225–33

Gove W R, Hughes M, Style C B 1983 Does marriage have positive effects on the psychological well-being of the individual? *Journal of Health and Social Behavior* **24**: 122–31

Heaven P, Searight H R, Chastain J, Skitka L J 1996 The relationship between perceived family health and personality functioning among Australian adolescents. *The American Journal of Family Therapy* **24**: 358–66

Kluwer E S, Heesink J A M, van de Vliert F 1997 The marital dynamics of conflict over the division of labor. *Journal of Marriage and the Family* **59**: 635–53

Olson D H 1989 Circumplex model and family health. In: Ramsey C N Jr (ed.) *Family Systems in Medicine*. Guilford Press, New York, pp. 75–94

Olson D H, Stewart K L 1991 Family systems and health behaviors. In: Schroeder H E (ed.) *New Directions in Health Psychology Assessment*. Hemisphere, New York, pp. 27–64

Pruchno R, Burant C, Peters N D 1994 Family mental health: Marital and parent–child consensus as predictors. *Journal of Marriage and the Family* **56**: 747–58

Repetti R L 1989 Effects of daily workload on subsequent behavior during marital interaction: The roles of social withdrawal and spouse support. *Journal of Personality and Social Psychology* **57**: 651–9

Ross C E, Mirowsky J, Goldsteen K 1990 The impact of the family on health: The decade in review. *Journal of Marriage and the Family* **52**: 1059–78

Satir V 1983 *Conjoint Family Therapy*. Science and Behavior Books, Palo Alto, CA

Schneewind K A 1995 Impact of family processes on health beliefs. In: Bandura A (ed.) *Self-efficacy in Changing Societies*. Cambridge University Press, New York

Schneewind K A 1991 *Familienpsychologie* (*Family psychology*), 2nd edn. Kohlhammer, Stuttgart, Germany

Schneewind K A, Ruppert S 1998 *Personality and Family Development: An Intergenerational Longitudinal Comparison*. Lawrence Erlbaum Associates, Mahwah, NJ

Schwarzer R (ed.) 1997 *Gesundheitspsychologie. Ein Lehrbuch* (*Health Psychology: A Textbook*) 2nd rev. edn. Hogrefe Verlag für Psychologie, Göttingen, Germany

Wickrama K A, Conger R D, Lorenz F O 1995 Work, marriage, lifestyle, and changes in men's physical health. *Journal of Behavioral Medicine* **18**: 97–111

J. Weiß

Family Law

1. Introduction

Geertz described law as much a part of the culture of a society as its poetry, painting, or music (Geertz 1983). This profound and illuminating approach to

the patchwork of norms, rules, and institutions by which we regulate our daily lives is nowhere more appropriate than when we consider the regulation of family matters. Here law is at its most social as we move forward to new ways of organizing our households and relationships in a multicultural context. The study and practice of family law has attracted socially minded lawyers and has provided a seedbed for the development of the sociological approach to law now known as sociology of law.

In this short article, I will first lay out the various ways of conceptualizing the role of law in family life as the framework within which the individual, the family, and the state define their mutual obligations, and second examine the implications of these models for the framing and operation of the legal assumptions and instruments which regulate family matters.

As Aubert has pointed out, both law and social science are concerned with social interaction: with rights and expectations, with obligations and sanctions, and with predictability (Aubert 1963). But the same terms are used for different purposes. The lawyer uses a normative framework; she gives directions to her client and expects action to follow. The social scientist, on the other hand, observes and describes, seeking to identify and to explain patterns of behavior. The lawyer interested in family work is often concerned with how people behave in family settings, as well as with the strict applications of legal rules. The social scientist who looks at family law is often policy-aware and interested in the practical implications of the work, as well as in the theoretical analysis. The study and development of family law in many different societies, including Japan, Italy, France, the UK, the USA, Australia, and Poland has been imbued with collaboration between social scientists and lawyers (Murayama 1999, Ronfani 1991, Bastard and Cardia Voneche 1995, Commaille and Singley 1997, Maclean and Eekelaar 1998, Weitzman 1991, Funder 1996, Maclean and Kurczewski 1994). The result is a rich understanding of law in family life.

2. *Families, Politics and the Law*

What do we want from law in the context of family life—as opposed to business life, or economic life, or international relations?

The way we model our answer will depend on our view of the family, of the law, and of the state. Let us take three examples of the interweaving of politics, family life, and the law: an old fashioned liberal democracy, an old democracy emerging from totalitarian rule, and a new democracy emerging into a multicultural society. From these examples, we shall derive our models.

In the predominantly secular UK, discussion of the boundaries between the obligations of individuals to each other as individual citizens or as members of family groups (however these are defined) usually is couched in economic rather than in moral terms. In times of economic stringency, there is a tendency to retrench public spending and to stress the obligations of families to support their members, and also to extend the ties which carry a duty to support. For example, the obligations of parents who do not share a household with their children will be highlighted and enforced. Such moves are countered by the claims of weaker family members, usually women and children but also the older generation, to establish their right to call on the public purse as individuals, irrespective of their family relationships. In the UK, family forms have changed and diversified as a result, in part, of the increased opportunities to form separate households when housing became more widely available in the late 1960s and early 1970s. We have experienced a divorce explosion and a rise in cohabitation. More older people are forming separate households, more women are working despite limited child care, and more men are unemployed.

'Supporting families' has become a political slogan attracting universal support in the UK, but for a variety of reasons. Family life is seen as an answer to the need for increased social control, particularly of young delinquents, and also as the alternative to public provision of shelter and support. The conservative administration of the Thatcher era was active in passing legislation to increase the visibility and enforceability of private obligations as an adjunct to the rolling back of the welfare state. Thus, the UK saw the paradox of a government committed to 'privatizing' the family and to withdrawing from intervention in family matters but compelled to construct a regulatory framework in order to do so. The Child Support Act of 1989 stands at the pinnacle of this policy. This Act attempted to reverse the widely held assumption that when a couple separate and the breadwinner re-partners and produces a second family, their resources should be mainly available to the household in which they now live and not be directed back to the first set of children. The Act firmly stated that first children must have first call on their absent parent's resources. To achieve this change, the government, which was pledged to reducing the role of the 'nanny state,' was required to take draconian powers of investigation and enforcement.

On the other hand, in Poland between 1945 and 1990, the family had a unique role as the only institution supported both by the ruling Communist Party and by the Catholic Church, for different reasons but with a similar purpose. The family was seen as the institution through which moral values are best transmitted and reinforced, and as a powerful agent of social control. In times of intellectual and spiritual alienation, the family became of primary importance as the place where trust predominated over suspicion and where communication could be frank and open.

When there were shortages, the family functioned as an effective searching, queuing, and allocating organization, providing specialist's support for the group in bringing home goods in short supply, or in helping with access to jobs and housing. Following this period of intense interdependency, it has been suggested that family members may wish to withdraw from this level of closeness and move towards the pure relationships of coupledom, parenthood, and friendship defined by Giddens (Reith Lectures 1999). But if the public welfare provision of care for children and care for the sick and elderly break down, then the family may be needed to replace these functions. Despite the emphasis on individual rights and liberties in the free market after the end of authoritarian rule in Poland, family values remain strong and are reinforced by the Catholic Church, whose wish to restrict divorce and abortion has had a profound impact on family law since 1990. Divorce has been removed to the higher level of the courts, which means a long wait and increased costs to the petitioner. And at the same time, the long struggle in the *Sejm* to restrict the right to abortion succeeded in January 1993 by a narrow margin.

For our third example, let us look to South Africa. Those who prepared the new Rainbow Constitution, which signals the passing of the apartheid regime, have struggled with the question of whether to constitutionalize the rights of the individual to marry the person of his choice and to found a family. The Constitutional court acknowledges that a number of international instruments clearly establish the right to form and live in a family. But although many constitutions expressly defend the family (e.g., Germany, Pakistan, China), many do not (e.g., Austria, India, Singapore, Taiwan). The South African government decided not to constitutionalize family matters, stating that the very nature of the family, particularly in multicultural and multifaith societies, is so diverse that it is best to leave the fundamental rights and liberties in relation to family life to the basic principles of freedom, security, and choice. 'Then through legislation, the development of precedent in courts, social custom and practice, the different forms of family life will manifest themselves and be appropriately protected. The minute you constitutionalize the family, the courts are obliged to establish a prototype of what is meant by a family, and families take such diverse forms in South Africa that this could impose a straight jacket on future developments.' (Eekelaar and Nhlapo 1999).

3. *Modeling Family Law*

We have looked at three very different examples of the interrelationship between the political situation, the institution of the family, and family law. Now we will try to draw on these variations to develop a set of models which incorporate the key elements and illuminate our own choices. We can categorize the family law system according to the level of intervention accorded to the state in family life.

3.1 *The Interventionist Model*

In societies where the state has a strong and coherent message to communicate, based on a political or religious ideology or world view, we tend to find an interventionist model of family law. The function of the family may be to raise good socialist citizens, as in the former communist countries of Eastern Europe, or to raise good catholic citizens, as in Southern Europe. In the former, divorce and abortion were freely available, gender equality was well developed, citizens were equal in the eyes of the state, and labor productivity required women to be available for work. In the latter, divorce and abortion were unavailable. If the family is seen as a box with walls of varying degrees of porosity, in interventionist societies the box is very transparent and porous, allowing the regulatory system to enter. In this model, the family is less able to function as an institution of civil society but, instead, serves the dominant ideology, whether secular or sacred. Such a system has difficulty providing for a multicultural society and traditionally obscures ethnic differences. Such systems ironed out the differences in family life between the Roma, Turks, Pomacks, and Bulgarians in Bulgaria (Maclean and Kurczewski 1997).

3.2 *The Residualist or Minimalist Model*

The residualist model lies at the other extreme of our continuum. In this model, the family is seen as the desirable or competent source of nurture, caring, social control, and perhaps moral education. In such a model, the walls of the family box are dense to opaque, and the state does not easily penetrate the private family world. The function of the law is to do no more than protect the weaker members of the family group from violence or neglect. This may be left to the criminal law rather than requiring a special form of intervention through family law. Such a model is not acceptable to feminists, as nonintervention in effect favors the status quo and does not redress the power imbalance characteristic of patriarchal societies (see *Gender and the Law*).

3.3 *The Facilitative Model*

There is, however, a middle position which may be termed the Facilitative model of family law. Under this approach, family law is required to respect the human rights outlined above and (Eekelaar and Nhlapo 1999) to be based on the underlying principles

of freedom, security, and choice, to enable each individual to develop to the fullest and to accommodate the varieties of faith, ethnicity, and sexual orientation. Such a model would encompass acceptance of cohabitation, same sex marriage, assisted reproduction, abortion, divorce, and remarriage, and maybe polygamy and polygyny.

4. *The Way Ahead*

Whichever model of family law underpins the regulatory system of a particular country or state, there do seem to be some common directions for change. In the 1980s and 1990s, writers on family law identified what became known as the rolling back of family law, that is, the increased reliance on individual decision-making and the removal of a number of matters related to personal relationships from legal intervention. It became less common for legal rules to regulate choice of marriage partner (i.e., the old restrictions from the Judaeo–Christian tradition on marrying the dead wife's sister which conflict with the Moslem requirement to take on responsibility for a dead brother's wife). In many societies, no-fault divorce became a matter of personal choice. But although the couple relationship is coming to be seen more as a matter of choice, and to lie outside the boundaries of legal regulation, the relationship between parents and children becomes more exposed and subject to increasing legal intervention. Where children live in households with parents, the need for protective legislation remains, but the need for regulation of how the household sustains itself is minimal (see *Children and the Law*).

But as couple relationships become divorced from the parenting function, the need for legal regulation of the obligations of parents increases. If two people produce a child, and either never live together or separate, and either or both go on to have other children with other partners, the potentially conflicting obligations between this network of parents and children requires a regulatory structure. We referred earlier to the difficulty in the UK when the Child Support Act in 1991 required a change in the distribution of the assets of the absent parent, requiring him to give priority to the children born first, and not to children currently sharing the new household. This was a hard switch to make, and it aroused great public outcry.

But there is a further aspect of the parent–child relationship which may also give rise to legal regulation in the future. With an aging population, and the rolling back of welfare provisions, we may be looking to adult children to offer more support to their elderly relatives (see *Law and Aging*). After the establishment on a firmer footing of child support, we may need to consider legislating for parent support. Such a development manifests a view of family law as no longer concerned with moral codes of behavior or political ideologies, but as trying through basic human rights principles to deal as best it can with multicultural and multifaith societies. Family law is developing to deal with personal obligations such as material obligations to support and enable.

At the same time, as the thrust of legislative intent focuses on obligations, rather than on rights, so too there is a move away from lawyers and courts to enforce these obligations. We see a resurgence of interest in paralegal activity and in finding alternative methods of dispute resolution. Mediation and conciliation, which require the individuals concerned to take responsibility for arrangements made in family matters, are expected to reduce levels of hostility and to produce longer lasting settlement. Empowerment of the individual is thought to be well served by these means, together with improved ability to communicate, which will be increasingly important as parents continue to take long-term responsibility for their children while no longer sharing a household.

Finally, as family arrangements become more attuned to individual choices in societies with a more tolerant view of different ways of arranging personal life, the ability to negotiate and to adjust becomes more important than the ability to fight for a particular right. If we continue to move in this direction, then it is probable that family law will become less concerned with adjudication or settlement of disputes, and more concerned with the business of managing change. A separation or divorce will involve less acrimony, and there will be clear directives on how to allocate responsibility for children. Legal intervention in families may come to resemble the technical financial advice of the accountant, rather than the fighting spirit of the old divorce lawyer 'taking the husband for all he's got' or 'making sure she doesn't get away with a penny.' The family law of the past was prescriptive, derived from the allocation of property between marrying groups. The family law of the future will be grounded in human rights and will facilitate the maximization of individual choices about personal living arrangements. It will ensure protection for the vulnerable family member, who is likely to be defined by age and relationship to the labor market, rather than by gender.

See also: Children and the Law; Family, Anthropology of; Family as Institution; Family Bargaining; Family Processes; Gender and the Law; Law and Aging; Law and Society: Sociolegal Studies; Motherhood: Economic Aspects; Property: Legal Aspects of Intergenerational Transmission

Bibliography

Aubert V (ed.) 1963 *Sociology of Law*. Penguin, Harmondsworth, UK

Bastard B, Cardia Voneche L 1995 Professional intervention in divorce. *International Journal of Law and the Family* **9**: 275–85
Commaille J, Singley F (eds.) 1997 *The European Family*. Kluwer, Dordrecht
Eekelaar J, Nhlapo T (eds.) 1999 *The Changing Family*. Hart, Oxford, UK
Funder K 1996 *Remaking Families*. AIFS, Melbourne, Australia
Maclean M, Eekelaar J 1998 *Parental Obligation*. Hart, Oxford, UK
Maclean M, Kurczewski J (eds.) 1994 *Families, Politics and the Law*. Oxford University Press, Oxford
Maclean M, Kurczewski J (eds.) 1997 *Family Law and Family Policy in the New Europe*. Hart, Oxford, UK
Murayama M 1999 Does a lawyer make a difference. *International Journal of Law Policy and the Family* **13**: 52–77
Naoumova S 1997 Ethnicity and Family. In: Maclean M, Kurczewski J (eds.) *Family Law and Family Policy in the New Europe* Hart, Oxford, UK
Ronfani P 1991 Towards a Dejuridification and a Dejurisdictionalisation of the Family in Family Law and Social Policy. Onati Proceedings (13)
Weitzman L J, Maclean M 1991 *The Economic Consequences of Divorce*. Oxford University Press, Oxford

M. Maclean

Family Planning Programs: Development and Outcomes

1. Demographic Momentum

One of the most important developments of the twentieth century was a huge increase in global population and a massive change in the distribution and age structure of the population. In the second half of the century, differences in birth rates were an important factor dividing the world into developed and developing nations. For a long time there has been compelling evidence that many couples all over the world want fewer children. In most Developing nations with high birth rates, national family planning programs have been established. Some of these programs have been highly successful, but others failed due to controversy, lack of political will, or inappropriate technical advice.

In the twenty-first century the world will divide along a new socio-economic fault line, also largely driven by differences in birth rates. Countries that achieve replacement level fertility by about 2010 are likely to progress economically. Those that do not will face continued rapid, and in some countries unsustainable, population growth. In much of Africa, for example, half the present population is below the mean age of marriage, and even if the fresh cohorts of young people are able to have smaller families, there will be so many new parents that the population will continue to grow. This phenomenon of population, or demographic, momentum occurred in China in the 1960s and 1970s. As a result of rapid population growth, many more Chinese brides were entering the fertile years than were leaving them by death or reaching the menopause. In 1979 China instituted a one-child policy, but even today the country still has one million more births than deaths each month—and therefore a net growth to China's population of one million each month. Since the one-child policy began, China has added more people than live in the whole of the USA.

In order to overcome past failures in family planning, countries that are being overtaken by a new wave of demographic momentum will need to adopt bold new policies. This will be difficult.

2. Why Do People Have Children?

Unlike most other animals, human beings have frequent sex; most of it unassociated with ovulation and often taking place thousands of times more frequently than is required for fertilization. This behavior probably evolved as a social bond linking parents in the arduous task of bringing up a family. In preliterate societies, pregnancies were naturally spaced by hidden physiological mechanisms that suppress ovulation during breastfeeding. In such communities, women averaged 4–6 live-born children in a lifetime, of which approximately half died before they could reproduce. In the modern world, family size has increased because (a) women often only breastfeed for a few months or not at all, (b) puberty occurs several years earlier than it used to, and (c) infant mortality has plummeted. These are the factors that drove the twentieth century's 'population explosion,' and these factors will continue to operate in the twenty-first century.

The simple fact that human beings have sexual intercourse much more frequently than most animals is obvious, yet often overlooked by governments when they are setting family planning policies, or specialists analyzing demographic changes. Many policy-makers continue to focus on factors that correlate with family size, such as education or income, instead of those that cause changes in family size, which, in addition to the age when intercourse begins and patterns of breastfeeding, are the use of contraception and abortion.

Economists often speak of people making rational choices over the costs and benefits of a particular act, such as buying a washing machine. However, as a result of frequent sexual relations between couples, human fertility is in effect perpetually turned on, and such rational choices are not possible. We buy a washing machine when we want one, but if this worked like human fertility, the washing machine would arrive whether it was ordered or not. In the modern world it is possible to have a small family only if the couple have access to contraceptives and abortion. Tragically, as the twenty-first century begins, hundreds of millions of people still find contraceptives geographically, socially, or financially inaccessible, and for even more

women, abortion is available only in dangerous and commonly financially or sexually exploitative circumstances.

3. Changing Attitudes

In 1900 the world was deeply hostile to the public support of fertility regulation. In the US and parts of Europe, the sale of contraceptives was illegal, while in the UK sales were 'under the counter.' Throughout Europe, North and South America, and in all the European colonies, abortion was illegal. Yet doctors and clergymen, who often opposed contraception in public, were beginning to limit the number of children they had. Condoms, spermicides, intrauterine devices (IUDs), and voluntary sterilization were all known in 1900, but were of poor quality and were difficult to acquire. The theory behind the Pill was described in the 1920s, but research in contraception was actively discouraged. When Margaret Sanger opened a family planning clinic in New York in 1916, it was closed by the police. When Marie Stopes started family planning clinics in England in 1921, she had to face stiff criticism and even arson.

The pivotal point in the development of family planning programs came near the middle of the twentieth century. The 1950s saw the introduction of oral contraceptives, the more widespread use of IUDs, and some Scandinavian and eastern European countries introduced liberal abortion laws. Family planning clinics run by non-governmental agencies began to be important in a few developed nations. In 1965, the US Supreme Court struck down the last of the anti-contraceptive laws passed almost 100 years earlier. In 1966 the British reformed their abortion law, and in 1973 the USA Supreme Court (Roe v. Wade) made safe abortion legal throughout the country.

In developing nations, non-governmental organizations also took the lead. In 1952 the International Planned Parenthood Foundation (IPPF) was founded to tie these efforts together. In the 1960s, US foundations and the Swedish International Development Agency (SIDA) began to offer money and know-how to help family planning programs in developing nations. The US Agency for International Development (USAID) followed soon afterwards, and finally a number of developed nations joined in offering support.

But there were also problems. Much of the academic community clung too long to outdated paradigms and until1959 the US National Institutes of Health (NIH) was forbidden to conduct contraceptive research. The Vatican State, which is part of the UN system, successfully blocked the World Health organization (WHO) from offering governments help in family planning until the 1960s. The British representative at the WHO Assembly was ordered by the Foreign Office to not discuss family planning, but eventually India, Trinidad and Tobago, and the Netherlands (with British support) did succeed in getting a resolution permitting WHO assistance to respond to requests for family planning technical assistance.

In 1968, many Catholics were stunned when Pope Paul published the Encyclical *Humanae vitae* and upheld the 1600-year-old tradition of St Augustine that sexual intercourse was sinful unless it was open to the possibility of procreation. (In a clear demonstration of the fact that theological opinions on reproduction are not necessarily constant, *Humanae vitae* permitted the use of periodic abstinence, or the safe period, which St Augustine in the fourth century had explicitly condemned as immoral.) Today, almost every developed nation, with the exception of the Republic of Ireland and Malta, offers its citizens access to modern family planning. Pills, condoms, safe and effective IUDs, increasing access to voluntary sterilization for either sex, and above all, access to safe abortion has revolutionized the lives of women in developed nations. For the first time in history, they could fit childbearing into a variety of chosen lifestyles and no longer had to spend every month worrying whether they were going to see their next menstrual period. Nearly every developing country has also established some sort of national family planning program. However, the hostility and controversy that had characterized the slow and painful process of family planning in the first half of the twentieth century has not gone away, and many family planning programs in developing nations are deeply flawed and fail to offer the choices required to truly control family size.

4. Family Planning Programs

The development of international family planning movement can be divided conveniently into three phases, each one symbolized by an international population conference organized by the UN.

In the 1950s and 1960s, a few timid, and sometimes frankly unrealistic, efforts at family planning began. In 1951 the WHO responded to a request by the Indian government by advising use of periodic abstinence. (At that time, the Vatican City, as a member of the WHO governing body vetoed more practical methods.) India adopted a formal governmental population family planning program in 1952, but efforts were made to use IUDs on a large scale failed because the health personnel or facilities to support the method were not available. Very different programs began in South Korea (1963) and Taiwan (1959). Here, several methods including voluntary sterilization were offered and safe abortion, although initially illegal, was widely available. New choices were added to the menu of contraceptives as couples showed they wanted them.

These initial efforts were largely brought about by a small number of charismatic, hard-working individuals. They probably had a powerful impact because

they were responding to a deep and long-hidden desire for fertility regulation in groups all around the world. In the US William Draper Jr. led the Population Crisis Committee, badgered Congress, and played a key role in putting the UNFPA together. Reimert Ravenholt became an extraordinarily effective leader of the early USAID programs. Alan Guttmacher led Planned Parenthood/World Population in the US and changed the face of domestic and international family planning. Ottesen Jensen in Sweden and Madam Kato in Japan moved mountains of opposition. Fernando Tamayo, Jae Mo Yang, Dapne Chung, Lady Rama Rao, L. P. Chow and others appreciated the damage rapid population growth was doing to their societies and economies, and launched family planning programs in their respective developing nations. Those most closely involved in putting programs together, whether from developed or developing nations, were universally impressed by the unmet demand for family planning and the opportunity that family planning provided to meet individual, as well as community needs. Many of these early pioneers were also particularly skilled at turning the controversy, which associated their efforts to their own advantage.

The United Nations Fund for Population Activities (UNFPA) was launched in 1969, and the first UN population conference was held in Bucharest, Romania in 1974. Demographers and economists warned against the dangers of rapid population growth. The Bucharest meeting was dominated by the developed nations who emphasized the need for family planning programs. But a number of developing nations, including India, expressed resentment at what they perceived to be the demographic emphasis of the meeting.

However, when it comes to sex and reproduction, people do not always say what they mean, or mean what they say. Although India and China adopted the rhetoric that 'development is the best contraceptive,' almost as soon as the meeting ended China introduced the one-child family and India embarked on a top-down family planning campaign that ended with episodes of coercive sterilization, which brought down the government of Indira Gandhi. The problem in China was that Chairman Mao had denied there was such a thing as a population problem, and this delayed the introduction of a realistic family planning program. As noted, China ended up with an enormous imbalance of young people. No one in China liked the one-child policy but neither has anyone ever come up with a plausible alternative strategy. Had China started a voluntary family planning program just a few years earlier, there is every reason to believe it would have been as successful as Korea, which saw the TFR fall from 6.0 in 1960 to 1.7 in 1990. Indeed, during the 1970s a number of countries, such as Colombia and Thailand, followed the example of Taiwan and South Korea. They forged ahead offering several methods of contraception, they developed non-medical channels of distribution and they confronted the public health issues of unsafe abortion. Whereas it had taken the USA almost 60 years (from 1842 to 1900) for the total fertility rate (TFR) to fall from 6.0 to 3.5, Taiwan, Columbia, and Thailand all made the same transition in a few years. In 1984 the world community met for a second population conference, this time in Mexico City. Whereas the US had been criticized in 1974 for emphasizing population problems, in1984 the Reagan administration asserted that population growth was not a serious problem and followed this up with policies in the US. It was now the turn of the developing nations to call for more support of family planning programs.

At the Cairo International Conference on Population and Development in 1994, yet another set of political forces surfaced. Population issues had fractured into five different schools of thought. Women's advocates focused on the needs of the individual, emphasized broader aspects of reproductive health as well as access to family planning, and downplayed demographic forces, often portraying family planning programs as coercive. Those emphasizing socio-economic development focused on the inequalities between North and South and the need to redistribute the world's wealth more equally. Voices from the South harked back to the Bucharest conference, and again used the slogan that 'development is the best contraceptive' to diminish any role for purely population issues. The Vatican continued its hard-line stand against any form of 'artificial contraception' and condemned the use of condoms, even by monogamous couples where a man might have acquired HIV through a blood transfusion. The group who had launched family planning in the 1960s were reduced in numbers, but they maintained a traditional concern about rapid population growth and emphasized that if the unmet need for family planning was met in a respectful way then fertility would fall rapidly. Most groups at Cairo (other than the Vatican) considered the meeting a considerable success because consensus appeared to have been achieved on so broad a range of issues.

The program of action that came out of Cairo underscored the heavy burden of social and cultural injustice that falls on many women in the world. The conference looked at sexually transmitted diseases and AIDS, it condemned female genital mutilation, and it set family planning in a broad context of reproductive health care. With the passage of time, however, the Cairo Program appears more blurred and less satisfactory. One of the weakest parts of the Cairo document were the targets set for finance, which were much too low. A broad agenda not backed up by the necessary budget is a dangerous situation. It is a sad fact that since the Cairo conference, the sum total of human suffering in the reproductive health field has significantly increased. The WHO has raised the estimates of women dying from pregnancy, childbirth,

and abortion from 500,000 to 575,000 a year—more than one a minute. In the countries most heavily affected by HIV/AIDS, reproductive health has deteriorated more in the past five years than in any similar period during the twentieth century. (See *HIV and Fertility*; *Mortality and the HIV/AIDS Epidemic*.)

5. *Meeting Individual Family Planning Needs*

A great deal is known about how family planning choices can be made available in a poor country. Experience suggests that a 15 percent rise in contraceptive prevalence is associated with approximately one birth less per family. The wider the choice of contraceptive methods and the greater the variety of distribution channels the more rapidly contraceptive prevalence rises. In many countries prevalence has been going up at more than 2 percent per annum. Family planning is a choice and not a therapy, and therefore contraceptives can be distributed responsibly within a community, often outside any clinic-based health system. Unfortunately, the expansion of family planning also continues to be limited by restrictive medical policies. For example, injectable contraceptives could perfectly well be distributed at a subsidized cost by traditional practitioners, injection doctors, and others who provide medicines in rural and shantytown communities. The subsidized sale of contraceptives has proved an effective way of distributing pills and condoms and is almost always more cost-effective than distribution through medical clinics.

Family planning methods vary in their cost-effectiveness. Condoms are relatively expensive, while voluntary sterilization and intrauterine devices often cost the least per couple-year of protection. Probably the most realistic and acceptable way of satisfying the community's need for family planning would be as follows. The government health services would focus on IUDs and voluntary sterilization. Pills, condoms, and injectables would be available in government clinics, but the primary mode of distribution would be through the commercial sector or through subsidized marketing, involving a variety of outlets, from kiosks on the pavement, through chemical sellers and pharmacies, to traditional practitioners and finally trained physicians. Private physicians could offer sterilization and IUDs. The consumers would pay what they can afford and the government would make up the shortfall. Such a scheme has worked exceedingly well in South Korea and Taiwan and needs to be repeated in countries like India or Nigeria.

6. *Abortion*

The Cairo conference was the first UN meeting to mention the health consequences of unsafe abortion. It concluded that 'prevention of unwanted pregnancies must always be given the highest priority and every attempt should be made to eliminate the need for abortion.' These are goals that most reasonable people support, but it will take a long time to achieve them given the current low levels of funding.

No country has achieved a low birth rate without a considerable number of abortions—whether legal or illegal. Conversely, wherever women have access to safe abortion, fertility has declined to a lower level than where such a choice is not available. Indeed, nearly all countries that have access to safe abortion have reached or passed below replacement-level fertility. Technically, abortion in the first 3 months of pregnancy can be safely preformed in any clean room. Manual vacuum aspiration (MVA) using hand-held plastic equipment costing less than $20 has proved an appropriate technology in locations from rural Vietnam to urban San Diego, California. The technique can be used to treat patients with botched abortions as well as to terminate an early pregnancy. In Bangladesh, MVA is available for menstrual regulation, approved by the government as a family planning method through approximately 12,000 providers, and it is one reason why the TFR has fallen from almost 7 in 1960 to 3.3 in 1998.

Abortion is an ethical problem, and all modern democratic societies that separate church and state have to find ways to accommodate a variety of sincerely held views on this critically important topic. Some groups may have to learn the hard lesson that toleration is preferable to confrontation.

7. *HIV/AIDS*

By the early decades of the twenty-first century, AIDS will have killed 30 to 40 million people—more than died as civilians and combatants in World War II. Current UN population projections assume that the global population in the twenty-first century will be 400 million less than it would otherwise have been as a result of the spread of HIV/AIDS. This does not imply that 400 million people will die from this disease, but that the approximately 30 million projected to die early in the next millennium will not themselves have children, and the downstream effects of these deaths will be considerable. In most of the worst hit countries of Africa AIDS will slow population growth but it will not halt it.

8. *The Future*

The developing world is approximately halfway through the demographic transition that began in the 1960s. Family planning programs have played a significant, (and some would argue, a major) role in this important change. As the twenty-first century opens, the largest cohort of young women in history is

entering the fertile years. Contraceptive prevalence is rising. Unfortunately, at this time of unprecedented need, attention is being taken away from family planning. There is an increasing demand for subsidized contraceptives to meet the needs of many millions of very poor people, but there is no clear strategy in the donor community to provide the essential commodities required to fulfill this need.

In 1998 the UN population division issued new projections for global population. Demographic projections are 'what if' statements. The UN projections assume a continued fall in fertility, and in regions such as Africa, these are based on the assumption that the needed contraceptives and services will be available. This may not happen, in which case there will either be a 4-fold rise in abortions, or else millions of unintended pregnancies, in which case even the high UN projections will be exceeded. Of the small amount transferred from rich nations to the poor in overseas aid, only between 1 and 2 percent goes to international family planning. With the added burden of AIDS and lack of a coordinated global policy to ensure that essential commodities are available, the world faces a genuine crisis in family planning and reproductive health.

9. Summary

A great deal is known about how to provide effective family planning choices. When people are offered several methods of contraception and access to safe abortion, fertility always falls. However, family planning has a controversial history. Currently, poor countries sometimes lack the will to make realistic services available and rich countries are reluctant to put enough money into needed programs. As a result the world looks destined to become increasingly divided between high fertility and low fertility countries.

See also: Economic Development and Women; Family Planning Programs: Feminist Perspectives; Family Planning Programs: Quality of Care; Family Size Preferences; Family Theory: Economics of Childbearing; Family Theory: Role of Changing Values; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; HIV and Fertility; Land Rights and Gender; Mortality and the HIV/AIDS Epidemic; Population Dynamics: Momentum of Population Growth; Population, Economic Development, and Poverty; Population Policies, Politics of; Population Policy: International; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender.

Bibliography

Campbell M 1998 Schools of thought: An analysis of interest groups influential in international population policy. *Population and Environment* **19**: 487–512

Cleland J 1994 Different pathways to the demographic transition. In: Graham-Smith F (ed.) *Population: The Complex Reality*. The Royal Society, London, pp. 229–48

Kulczycki A, Potts M, Rosenfield 1996 Abortion and fertility regulation. *The Lancet* **347**: 1663–8

Potts M 1997 Sex and the birth rate: Human biology, demographic change, and access to fertility regulation methods. *Population and Development Review* **23**: 1–39

Potts M, Walsh J 1999 Making Cairo work. *The Lancet* **353**: 315–8

Potts M 2000 The unmet need for family planning. *Scientific American* (January) **282**: 88–93

M. Potts

Family Planning Programs: Feminist Perspectives

1. Population Paradigms

Since the 1960s and even earlier in some cases, governments and private organizations in many poorer countries have developed and implemented family planning programs. By 1985, 37 governments had established national family planning programs, and an additional 33 countries had significant nongovernmental programs with the direct or indirect support of governments (Bongaarts et al. 1990). These efforts were largely driven by fears of rapid population growth, and the reduction of fertility was their highest priority. By the 1990s, however, critics of these programs succeeded in causing, what many have called, a revolution in the rationale for these programs. The human right to autonomy in matters of sexuality and reproduction was proposed by feminists, as the cornerstone for these efforts. As a result of this critique, 180 countries achieved consensus on a new approach to population policies and family planning programs at the United Nations International Conference on Population and Development (ICPD), held in Cairo in 1994.

1.1 Family Planning Programs and the Erstwhile Population Approach

To fully understand the significance of the Cairo accord and the influence feminist thinking had on it, it is necessary to look at the philosophies and practices it replaced. The 1960s, with the advent of new con-

traceptive technologies and an increase in fears of a population explosion, saw a significant increase in support for family programs in both industrialized and Third World nations. In 1967, the United Nations Fund for Population Activities (later renamed the United Nations Population Fund) was established to support population and family planning programs. At the UN population conference in Bucharest in 1974, many policy makers—mainly from Northern nations—viewed unchecked population growth as the root cause of poverty, environmental degradation, and social unrest. Many developing countries resisted this line of thinking, suggesting instead that economic development was a more straightforward solution to societal ills, than family planning. While womens' issues gained some attention, they were not seen to be central. Meanwhile, on the ground, family planning programs continued, largely unaffected by the debate at the international forum (Demeny 1985).

At the 1984 Mexico City conference, demographic concerns continued to be the driving force in discussions. However, in a reversal of its previous policy, the powerful US delegation now asserted that population growth had a neutral, not harmful, effect on economic development. Furthermore, anti-abortion politics gained center stage as the US launched an attack on abortion rights (Demeny 1985). Still, family planning programs on the ground continued to be guided by demographic goals and to use mechanistic measures such as reaching target numbers of 'contraceptive acceptors.' These programs' principal strategy was the distribution of contraception, chiefly to married women. Perhaps called population programs more accurately than family planning programs, these efforts usually did not address other health needs, even reproductive health needs.

2. *The Feminist Critique of Population and Family Planning*

As these programs proliferated throughout the 1970s and 1980s, women around the world began focusing on the negative effects of family planning programs that were based on demographics. While not denying the benefits of family planning, feminists rejected the depiction of women as targets of programs designed to limit births (Pitanguy 1995). The feminist critique concentrated on several aspects of program practices, including their vulnerability to coercion, their low quality, and their sole concentration on fertility regulation to the exclusion of other womens' health needs.

The vulnerability to coercion was a major element of the feminist critique. In some countries, health care providers were under pressure to achieve targets or be subject to penalties, and workers pressed people into accepting contraception or sterilization against their will. With success of family planning programs measured in terms of numbers of acceptors, some agencies and governments devised incentive and disincentive schemes to induce couples to participate in family planning programs. For some time, there was a debate about the fine line separating incentives from coercion. Feminists argued that, while milder than sheer force, the incentive practice is coercive; incentives are aimed primarily at the poor, for whom an economic inducement can be a vital necessity, not a free choice (Hartmann 1995). The Cairo Programme of Action later disavowed incentives.

Furthermore, women—and women were the objects of these campaigns in much greater numbers than men—were given little contraceptive choice. The focus was on sterilization or other modern, provider-controlled, methods of contraception such as the injectable Depo-Provera. The risks of these methods, feminists argued, were largely ignored and even exacerbated as a result of poor screening procedures and the lack of follow-up examinations. Safer methods, such as the condom and diaphragm were neglected, even though barrier methods have the ability to help prevent the spread of sexually transmitted diseases. Indeed, apart from fertility regulation, family planning programs provided little else in the way of services. Furthermore, male contraceptive methods were ignored in the research sphere and male participation was neglected in family planning programs, except for sterilization campaigns in some countries (Dixon-Mueller 1993).

Feminists insisted that family planning programs be comprehensive in nature with the limiting of births being only one item in a wide range of women's reproductive health needs. Other items included skilled care during pregnancy; delivery and postpartum care; prevention of infertility; counseling for sexual dysfunction; safe abortion; prevention and management of sexually transmitted diseases; reproductive tract infections, and reproductive system cancers; and treatment of postmenopausal health problems (HERA n.d.).

Furthermore, feminists demanded that women's health advocates should be represented at all levels of policy and decision making. They also maintained that men's responsibilities in the reproductive realm should not be ignored (Women's Declaration on Population Policies 1992).

Perhaps most central to the feminist critique is the assertion that women have the ability to make appropriate reproductive choices. In the feminist view, reproductive choices belong at the individual level, not with the health practitioner and especially not with the state. Noncoercive family planning programs aim to prevent unwanted fertility—as defined by couples, not by the government or other outside parties—through the provision of competent services that respect dignity and privacy. The essence of the feminists' perspective is the view that reproductive rights are human rights.

Another fundamental premise of the feminist perspective on family planning programs, sees women's empowerment as a desirable goal in its own right, rather than a means to a demographic end. Feminists also emphasize gender equity, including mens' responsibilities in sexual and reproductive matters (García-Moreno and Claro 1994).

Articulating these views, women in many countries around the globe began advocating wholesale transformation of the approach to population concerns and the basis for family planning programs. By the time of the 1994 ICPD, an international women's health movement was organized and had developed the expertise and the professional skills needed for influencing policy makers and shaping the paradigm shift on population and family planning (Klugman 2000). 'The Cairo Programme of Action sums up the agenda of priorities which women from all over the world, through their leadership networks, built gradually during the years of preparation for the conference' (Berquó 1997).

3. *The Cairo Programme of Action and the 'Feminist Imprint'*

The Cairo conference ushered in a new era of population policy and family planning programming. Its resulting Programme of Action reflects a clear 'feminist imprint' (Petchesky 1995). The conference document presents a new perspective, in which high rates of population growth are understood as an interdependent and aggravating factor—rather than the cause—of problems such as poverty and environmental degradation. Even more critically, it places women's wellbeing at the center of population policy and points to the human rights of individuals to determine and plan family size. Unlike documents from previous population conferences, the ICPD Programme of Action includes chapters on 'Reproductive Rights and Reproductive Health,' and 'Gender Equality, Equity and the Empowerment of Women.'

Echoing the feminist critique, the comprehensive nature of reproductive health is elucidated in section 7.2 of the Programme of Action, as follows:

> Reproductive health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity, in all matters relating to the reproductive system and to its functions and processes. Reproductive health therefore implies that people are able to have a satisfying and safe sex life and that they have the capability to reproduce and the freedom to decide if, when and how often to do so. Implicit in this last condition are the right of men and women to be informed and to have access to safe, effective, affordable and acceptable methods of family planning of their choice, as well as other methods of their choice for regulation of fertility which are not against the law, and the right of access to appropriate health-care services that will enable women to go safely through pregnancy and childbirth and provide couples with the best chance of having a healthy infant.

While previous population conferences had always seen women's development as a tool for achieving lowered fertility rates, the Cairo Programme of Action places women's empowerment as a worthy goal in its own right. 'The empowerment and autonomy of women,' the agreement reads, 'and the improvement of their political, social, economic and health status is a highly important end in itself' (*Programme of Action*: 4.1). In another clear echo of the feminist critique, the document goes far in its call for gender equality (*Programme of Action*: 4.1):

> The power relations that impede women's attainment of healthy and fulfilling lives operate at many levels of society, from the most personal to the highly public. Achieving change requires policy and programme actions that will improve women's access to secure livelihoods and economic resources, alleviate their extreme responsibilities with regard to housework, remove legal impediments to their participation in public life, and raise social awareness through effective programmes of education and mass communication.

4. *Critiquing the Critique*

The feminist critique has its detractors and the outcome of the Cairo Conference was not universally applauded. Vocal critics were mainly religious fundamentalists, especially the leadership of the Roman Catholic Church (Shannon 1994). A much quieter reaction was registered among some demographers and others whose concerns about high rates of population growth led to their dismay at what they saw as the dismantling of population policies (Dixon-Mueller and Gemain 2000). For example, Malcom Potts maintained that the 'flaw in the Cairo process was to move from a well-tested middle ground of supplying family planning to adopting a much broader agenda' (Potts 1995). Jason Finkle and Alison McIntosh claimed that women's groups 'co-opted' the ICPD. The conference, they maintained, was 'not a step forward for the population movement, which for many years has given primacy to efforts to limit population growth' (Finkle and McIntosh 1996).

Even some supporters of the reproductive health paradigm adopted in Cairo found some faults in its processes and outcomes. In a review of responses 'that tend to temper the initial feeling of euphoria surrounding the triumph of the feminist agenda,' Huda Zurayk writes that while a large number of women from Third World countries were present in the ICPD process, 'the strings were definitely being pulled by the well-organized Western feminist groups' (Zurayk 1997). Others, however, point to the significant con-

tribution of Third World—particularly Brazilian, Indian, Egyptian, Nigerian, Mexican, and South African—women to ICPD discourse.

5. *Beyond Cairo: The Feminist Critique Moves Forward*

After the Cairo Conference, the feminist critique expanded to include a vision of policies, governing not only family planning programs but also the broader panoply of issues related to reproduction and sexuality. One year after the ICPD, in 1995, the United Nations held another of its major international conferences, this time the focus was on women and the venue was Beijing. Again the ensuing conference document, the Beijing Platform for Action, reflected feminist input. While refraining from using the phrase 'sexual rights,' the agreement nonetheless endorsed the basic tenets of such rights, and placed them within a human rights framework. 'The human rights of women include their right to have control over and decide freely and responsibly on matters related to their sexuality, including sexual and reproductive health, free of coercion, discrimination and violence,' reads paragraph 96 of the *Plan of Action*.

There is little question that feminist analysis has had significant influence on policy worldwide, principally through the actions of women's nongovernmental organizations. Feminist scholars are now pressing for a similar impact on the discipline of demography, where gender issues have remained marginal. While ideological resistance presents a considerable obstacle, 'intellectually, the incorporation of gender issues in demography promises to enhance the science by offering a deeper understanding of social process and the application of multilevel methodologies' (Presser 2000).

Feminists are now focusing on implementation of the ICPD. Even with the advances in international agreements, many observers have pointed out that change on the level of rhetoric is one thing, and change in the everyday lives of women is another. Studies in many countries show varying levels of success in making reality conform to aspirations. Feminists continue to point to reluctant donor nations, recalcitrant cultural conventions, and wayward practitioners as some of the many obstacles to realizing quality family planning programs that are based on the expressed needs and desires of individual women and men.

See also: Economic Development and Women; Family Health; Family Planning Programs: Development and Outcomes; Family Planning Programs: Quality of Care; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Gender and Health Care; Gender and Reproductive Health; Household Production; Land Rights and Gender; Population Policies, Politics of; Population Policy: International; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender

Bibliography

Berquó E 1997 The ICPD programme of action and reproductive health policy development in Brazil. In: *Proceedings of the International Population Conference*. International Union for the Scientific Study of Population, Liège, Belgium

Bongaarts J, Parker M W, Phillips J F 1990 The demographic impact of family planning programs. *Studies in Family Planning* **21**(6): 299–310

Correa S, Petchesky R 1994 Reproductive and sexual rights: A feminist perspective. In: Sen G, Germain A, Chen L (eds.) *Population Policies Reconsidered: Health, Empowerment and Rights*. Harvard University Press, Boston

Demeny P 1985 Bucharest, Mexico City, and beyond. *Population and Development Review* **11**(1): 99–106

Dixon-Mueller R 1993 *Population Policy and Women's Rights: Transforming Reproductive Choice*. Praeger, Westport, CT

Dixon-Mueller R, Germain A 2000 Reproductive health and the demographic imagination. In: Presser H, Sen G (eds.) *Women's Empowerment and Demographic Processes: Moving Beyond Cairo*. Oxford University Press, Oxford, UK

Finkle J, McIntosh A 1996 Cairo revisited: Some thoughts on the implications of the ICPD. *Health Transition Review* **6**: 110–4

Garcia-Moreno C, Claro A 1994 Challenges from the Women's Health Movement: Women's rights versus population control. In: Sen G, Germain A, Chen L (eds.) *Population Policies Reconsidered: Health, Empowerment, and Rights*. Harvard University Press, Boston

Hartmann B 1995 *Reproductive Rights and Wrongs: The Global Politics of Population Control*. South End, Boston

HERA: Health, Empowerment, Rights and Accountability. *Women's Sexual and Reproductive Rights and Health Action Sheets*. HERA, New York

Jain A, Bruce J 1994 Quality: The key to success. In: Mazur L (ed.) *Beyond the Numbers*. Island, Washington, DC

Klugman B 2000 Empowering women through the policy process: The making of health policy in South Africa. In: Presser H, Sen G (eds.) *Women's Empowerment and Demographic Processes: Moving Beyond Cairo*. Oxford University Press, Oxford, UK

Petchesky R P 1995 From population control to reproductive rights: Feminist fault lines. *Reproductive Health Matters* **6**: 152–61

Pitanguy J 1994 Feminist politics and reproductive rights: The case of Brazil. In: Sen G, Snow R (eds.) *Power and Decision: The Social Control of Reproduction*. Harvard University Press, Boston

Pitanguy J 1995 From Mexico to Beijing: A new paradigm. *Health and Human Rights* **1**(4): 454–60

Potts M 1995 Cairo's skewed consensus. *Population Today* **23**(11)

Presser H 2000 Demography, feminism, and the science-policy nexus. In: Presser H, Sen G (eds.) *Women's Empowerment and Demographic Processes: Moving Beyond Cairo*. Oxford University Press, Oxford, UK

Shannon D 1994 All roads led to Cairo. *Conscience* **15**(4): 3–10

Women's Declaration on Population Policies 1992 In: Mazur L (ed.) *Beyond the Numbers*. Island, Washington, DC

Zurayk H 1997 Reproductive health in population policy: A review and a look ahead. In: *International Population Conference*. International Union for the Scientific Study of Population (IUSSP), Liège, Belgium

C. Barroso

Family Planning Programs: Quality of Care

The term 'quality of care' was introduced into the family planning lexicon around 1990. Since then, the concept has been widely accepted as an important dimension of family planning programs by managers, donors, health activists, and researchers. However, the actual progress in improving quality of care rendered by these programs has been limited for various reasons: lack of demonstrated large-scale experience in improving quality in different settings; limited number of studies to document the effects of improved quality; misperception of high cost; and lack of indicators to measure and monitor quality on a regular basis. A number of field studies are currently underway which will remedy this situation. This article reviews the concept of quality of care and the progress made in the field of family planning over the past decade.

1. *The Concept of 'Quality of Care'*

The concept of quality has been difficult to pin down precisely. Bruce (1990) defined quality as an important but neglected 'dimension' that all family planning services have. Whether the quality of a given service is good or bad is a matter of judgment. Bruce identified six elements in the quality of care framework as 'critical' aspects of services from the perspective and experience of clients. These are: choice of methods; information given to users; technical competence of providers; interpersonal relations between providers and clients; follow-up or continuity mechanisms; and appropriate constellation of services. While the use of these elements will assess from client's perspectives, their actual articulation in the framework are not based on any systematic or random surveys of clients' attitudes toward and expectations from services. Furthermore, clients may not be able to assess the adequacy of elements such as the technical competence of providers or what information they should have received at the time of interaction. The main purpose of this framework has been to draw attention to a neglected dimension and to generate interest among managers and donors toward improving the quality of care.

One criticism of Bruce's framework has been that clients themselves may not necessarily define these elements as critical in their decision to use family planning services. In this sense, quality needs to be distinguished from other concepts, such as availability, accessibility, and cost of services—all of which may also influence an individual's decision concerning the use of services. Sometimes there is a tendency to discuss all these factors together; but that is not useful for understanding and improving any particular dimension of services. While availability simply refers to the physical availability in an area, accessibility to a service delivery point (SDP) is usually defined and measured in terms of the available mode of transportation and the time taken in reaching that SDP. The cost of utilizing services is typically defined and measured by financial cost involved in getting services once the client reaches that SDP. At times, total cost also includes other costs that the client may bear—social, psychological, and the opportunity costs. Quality, on the other hand, reflects what happens to the client once they reach the SDP.

Often the two terms—'quality of services' and 'quality of care'—are used interchangeably. Nevertheless, a distinction must be made between them. The quality of services refers to attributes of the family planning services, e.g., the number and types of contraceptive methods, staff, and equipment available at an SDP. In comparison, quality of care refers to the way clients are treated by the system providing these services. For example, choice can be offered to clients only if the SDP stocks various methods; on the other hand, availability of methods at an SDP does not guarantee that a choice will be offered to a particular client. In this sense, quality of care places additional emphasis on the interpersonal interactions between providers and clients.

1.1 *Improving Quality*

A distinction must be made between determinants, indicators, and impact of quality. The determinants of quality include factors such as national population policies, resources allocated to services, and other social and cultural factors. The impact of quality, by contrast, can be considered in terms of improvements in clients' knowledge and behavior with respect to contraception, fertility, and health. As mentioned above, quality is a dimension that all programs have. In this sense, there is always room for improvement, irrespective of the current level of quality. There is no real maximum or upper limit. As far as the minimum level of quality is concerned, such standards should be set locally and explicitly through a process of discussion, which should seek to address the following four interrelated questions:

1. What quality of care does a program *intend* to offer?
2. What quality of care is a program *prepared* to offer?

3. What quality of care is *actually offered* to clients?
4. What quality of care is *received by* the clients?

The process of quality improvement can start with any question. The process is illustrated here by considering the element of choice. The first question deals with the explicit statement at the policy level about clients' right to choose among appropriate methods. The statement then has to be translated into practice through asking the second question, i.e., by considering the type and number of methods that are approved by the program in a country and by considering how appropriate the logistic system to deliver these methods to SDPs is. The next step in the process will be to ascertain whether the intended methods are actually available at an SDP and whether the providers are prepared to offer these methods, i.e., whether they have the requisite training and equipment. This type of assessment at the SDP level will inform the program managers about the extent to which the SDPs are prepared to offer a choice.

Next comes the assessment of the interactions between providers and clients, and this can be done from two perspectives, namely (a) whether the provider has actually offered a choice to clients, and (b) whether the client has received the choice. A particular program may, for example, decide to include three methods, improve the logistic system to get these methods and other required equipments to SDPs, and train providers to offer these methods, i.e., the program intends and is prepared to offer three methods. When a client walks into the SDP, the provider may even offer a choice between two methods; but the client's experience may be that she was not offered the third method of her choice even though all the procedures were in place and even though she had no medical contraindication.

Thus, it is important to think of quality at three levels—policy, SDP, and client. Quality at the policy level and quality of services at the SDP level can be considered as determinants of quality of care actually received by clients. What a particular client receives is the most important basis for assessing, monitoring, and improving quality of care. The identification of gaps between intention, preparedness, services offered, and care received through self-assessment and through systematic data collection is important for an eventual improvement of the quality of care received by clients over time.

The management literature suggests that the commitment of top management is critical in improving quality of care. But how one can generate commitment to quality of care among top management of family planning programs remains an unresolved issue. Unlike in the private sector, a commitment to improve quality in public sector programs with no competition cannot be generated by the motivation of maximizing profit or by the fear of going out of business if consumers do not like the product. Nevertheless, a combination of various approaches can make it happen.

One way to generate commitment to quality of care is through an organized movement of consumers to demand services of good quality. Another way could be for international donors and institutions to lobby for good quality of care from the consumer's-right perspective. A third possibility is for the top management to realize that improvements in quality will help them to achieve their stated goals of providing services in the first place. A fourth way is for the research community to generate empirical evidence to demonstrate the effect of improved quality of clients' knowledge and their reproductive behavior and health.

1.2 Monitoring Quality

There is considerable interest in creating indicators to measure and monitor quality. This interest stems from the observation that what gets measured and monitored also gets improved. Various efforts have been made to create indicators of quality. A breakthrough came with the methodological invention—Situation Analyses—to collect data from SDPs by actually visiting them. The technique usually involves collection of data about services on the day of a team's visit to a SDP through four questionnaires. The inventory questionnaire collects data about equipment and commodities available at the SDP. Information about provider's background, training, experience, attitudes, and knowledge is collected by interviewing providers working at the SDP. Information about quality provided and received is collected by observing client–provider interactions and by interviewing clients after they finish receiving services. These surveys conducted throughout many developing countries, especially in Africa, have provided important and useful information about the service environments. Surveys conducted over time on a representative sample of SDPs are also useful in measuring, monitoring, and improving quality. However, there is no agreed and tested set of indicators of quality. Additional efforts continue to be made.

Regular monitoring of commodities, equipment, and staff is equivalent to monitoring quality of services, which is sometimes included in what is called 'quality assurance.' Quality assurance is, however, not the same as quality of care, and it may be more difficult to measure and monitor the latter regularly. The problem is created by the small number of clients who use the SDP on the day it is visited by the data collection team. In some surveys, many SDPs visited did not have a client, and quite a few had just one or two clients. Biases are also introduced simply by the presence of an 'observer' at the time of interaction. Finally, it is difficult to assess the effects of variations among SDPs and providers on the overall quality of

care indicators developed on the basis of only a few observations. Some of these issues are also being addressed by ongoing research.

2. Impact of Quality

Clients deserve services of good quality, and they deserve to be treated with dignity and respect while accessing these services. From this perspective, there is no need to demonstrate the effect of improving quality on the reproductive behavior and health of clients. However, since family planning programs in developing countries have also been justified and evaluated in terms of their contribution to fertility decline, an assessment of the effect of improved quality on some indicators of fertility becomes critical in the process of improving quality.

Women who are very motivated (or desperately want) to control their fertility would do so even without the public sector services or even if these services were of poor quality. At the other extreme, women who do not want to regulate their fertility would not use services provided even if they were of the best quality and were available next door at no cost. Improved quality is hypothesized to affect the behavior of women in between these two extremes—those who are ambivalent about their fertility intentions, and those who do not go to services or discontinue the use of contraception because they feel that they receive poor quality or bad treatment. Improvements in quality of care thus can be expected to contribute to the reduction of unintended (unplanned and unwanted) pregnancies and not so much to the reduction in intended pregnancies.

Empirical studies to demonstrate the effect of improved quality on women's reproductive behavior are limited. It has been suggested that a focus on decreasing discontinuation of contraceptives is a better strategy than a focus on attracting new users and not taking good care of their needs. There is also some indication that improved quality in terms of providing a choice among methods increases contraceptive use by attracting new users, as well as by improving continuity of use. Moreover, improving the contents of information exchange between clients and providers improves the initial selection of the method and also improves continuity of contraceptive use. Additional studies are underway in Pakistan, the Philippines, Senegal, Zambia, and other countries, and they will shed further light on the effect of improved quality on clients' knowledge and behavior.

3. Future Directions

Quality was a neglected dimension of family planning services until about a decade ago. Clients have the right to, and they deserve, services of good quality. Therefore, program managers need to pay attention to improving quality of services. Empirical studies to demonstrate the effect of improved quality on clients' knowledge and behavior can only provide additional justification for improving quality; they cannot replace the main rationale of improving quality—clients' right to good quality.

The original framework of quality assessment was developed for contraceptive services provided through the public sector and through its fixed SDPs. Moreover, the framework is primarily applicable to first-time users of these services. This does not mean that services provided through the private sectors are of good quality or are monitored regularly. This also does not mean that repeat users of a SDP receive services of good quality. There are some efforts to extend the framework to other services and also to incorporate the experiences of repeat users. Such efforts need to be continued. Above all, there is a continued need to initiate processes in developing countries to improve quality of care received by clients.

It is possible to improve quality. It is true that improvements in certain elements (e.g., improving choice of contraceptives by adding new methods) will imply additional costs to the program. It is also true that poor quality resulting in the underutilization of services is costly. Moreover, an improvement in certain other elements is unlikely to cost much. This is especially true for items such as treating clients with respect and dignity and providing them with adequate and accurate information.

Providers often go through initial and refresher training programs. It should not be costly to modify the content and philosophy of these training programs so that providers learn how to solicit information from clients about their reproductive intentions, prior contraceptive experience, and to learn about other pertinent social factors which might influence the selection and use of a particular contraceptive method. It should also be possible to modify training programs so that providers learn how to offer clients adequate and accurate information about how to use the method selected and its adverse effects, how to manage those effects, when to come back for a checkup or a new supply, protection against transmission of disease, and the possibility of switching between methods or sources of supply whenever the method or source initially selected does not remain suitable for the client. Such a transformation, however, would require a shift from using family planning programs as a tool to reduce aggregate fertility and population growth to helping or empowering clients to achieve their own stated reproductive goals in a healthful manner.

See also: Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants; Regionalism and Cultural Expression

Bibliography

Bruce J 1990 Fundamental elements of the quality of care: A simple framework. *Studies in Family Planning* **21**(2): 61–91

Fisher A A, Mensch B, Miller R, Askew I, Jain A K, Ndeti C, Ndhlovu L, Tapsoba P 1992 *Guidelines and Instruments for a Family Planning Situation Analysis Study*. The Population Council, New York

Jain A K 1989 Fertility reduction and the quality of family planning services. *Studies in Family Planning* **20**(1): 1–16

Jain A K (ed.) 1992 *Managing Quality of Care in Population Programs*. Kumarian Press, West Hartford, CT

Jain A K 1999 Should eliminating unmet need for contraception continue to be a program priority? *International Family Planning Perspectives* **25**: S39 plus suppl. S

Jain A K, Bruce J, Mensch B 1992 Setting standards of quality in family planning programs. *Studies in Family Planning* **23**(6): 392–5

Mensch B, Arends-Kuenning M, Jain A K 1996 The impact of quality of family planning services on contraceptive use in Peru. *Studies in Family Planning* **27**(2): 59–75

Mensch B S, Arends-Kuenning M, Jain A K, Garate M R 1997 Avoiding unintended pregnancy in Peru: Does the quality of family planning services matter? *International Family Planning Perspectives* **23**: 21–27

Miller A, Ndhlovu L, Gachara M M, Fisher A 1991 The situation analysis study of the family planning program in Kenya. *Studies in Family Planning* **22**(3): 131–43

A. K. Jain

Family Processes

Although there is no univocal or consensual definition of the meaning of family processes, it most often refers to the interactions by which families make up and maintain their unity, manage their conflicts, achieve their socioeconomic integration, and model the social and emotional personality of their members (Sussman et al. 1999). These processes have been analyzed from various theoretical perspectives: structurofunctionalism, symbolic interactionism, development framework, system theory, conflictualism, rational choice, ecological analysis, and so forth (Klein and White 1996). This article will briefly review some of the major results and issues raised by these analyses.

1. The Constitution of Couples and Family

Several trends characterize the process of constitution of couples and family in modern societies. First, there is the widespread custom of romantic marriage, where the partners themselves—rather than parents or tradition—choose their spouse and their civil status.

However, various authors have shown that this individualization of the process of mate selection is accompanied by strong homogamy: people often choose their partner within their own social, cultural, ethnic, or religious background. The consequence of this homogamy—which is not so obvious in the middle of the social pyramid than in its extremes— is that love marriage does not disrupt social divisions, but conforms to and possibly reinforces them, as in the case of arranged marriages (Aldous 1996). It should, however, be noted that initial homogamy often alters during the family life course because of the disparity of social mobility between the spouses.

Former linear and standardized sequences of becoming part of a couple (i.e., courtship, engagement, marriage), punctuated by strong rites of passage, are often replaced by recursive and individualized paths, which are not culturally well-established. This is true for both the life course of one particular couple and the frequent succession of several marital stories: Family careers diversify and jostle together. Far from signifying an obsolescence of marital union, this individualization of the process seems to meet the couple's increasing concern for personal identity and stability within a mobile, polyreferential, anonymous, and segmented society. In the same way, different forms of today's family ties—cohabitation, one-parent families, step-families, and so forth—can be viewed simultaneously as much in terms of their appropriateness to a complex society as of a factor of stress or deficiency for the person (cf. Kaufmann 1993, Singly 1997).

Although it is individualized, the process of constitution of the couple is also shaped by omnipresent scientific and popular models of 'the' good marriage, models which partly replace the role of juridical norms and religious beliefs.

The transition from couple to family has formed the subject of more contradictory theoretical modeling. Economic theories, like Becker's (1981), model this transition as a balance between the important emotional meaning of the child for potential parents from all social backgrounds, and the cost it involves (loss of profit, restriction to autonomy, contribution and support to education, etc.). Therefore, the number of children born in a family, or their 'quality' (i.e., their social endowment), mostly depends on the parent's material resources. Without denying the growth—within secularized societies—of the importance of intentionality, of evaluation in terms of costs and profits, and concern for family planning, sociological theories are reluctant to limit the meaning of the child to its relational dimension. They add to it a statutory motivation for parents (the child as a resource of prestige, as a means to gain power, as a resource against the outside world) and a motivation regarding the child's needs (care, protection, attendance, etc.). According to the spouses' social assets, these different types of motivation can either be in competition with each other (the attraction for the relationship can be balanced by the fear of losing one's self-identity, professional power, and daily activities) or in con-

vergence. The situation of competition occurs more frequently as the cultural level of the wife increases. This often leads to a delay in producing a child, even when strongly desired, or even to this being renounced it completely by the couple (cf. Teachman et al. 1999).

2. *Types of Functioning*

Once the families are constituted, how do they work? Early analysis on this theme (e.g., Burgess et al. 1949) have contrasted traditional families (institution)—strongly institutionalized, with clear differentiation of roles between spouses, hierarchical organization of the relationships, and with instrumental goals (economic survival, social integration) prevailing within the couple—with modern families (companionship), marked by the predominance of expressive goals, of equality as the norm, where institutionalization and the differentiation of family member's social identities are softened.

But the growing variety of structural types of families has led several researchers to also emphasize the diversity of the types of functioning of contemporary families. Different typologies have been put forward.

First, focusing on primary goals, Farber contrasts the 'child-oriented families'—where emphasis is placed on children's social integration, on the differentiation of domestic tasks, and where constraint of socialization prevails on the couple's personal interests—with the 'home-oriented families,' where the emphasis is placed on comfort and mutual support in the family, and a flexibility of regulations, and to 'parent-oriented families,' more concerned with the parent's socioprofessional integration, who place parental constraints over those of children, and promote a certain lack of differentiation of roles.

Second, and more centered on modes of cohesion, the typologies of Roussel (1989) contrast 'association families' (strong emphasis on the autonomy of the spouses, a little differentiation of roles, essentially individualized goals) with 'enmeshed families' (emphasis on consensus and full sharing in the couple, predominance of emotional goals of the couple, relative flexibility of roles), and 'alliance families' (strong institutionalization, predominance of the family's social integration as a goal, strong differentiation of roles).

Lastly, other authors have focused on types of normative regulations. For Kantor and Lehr, 'open families' are characterized by organization standards (roles, behavior, etc.) that evolve with contextual changes, whereas 'random families' show almost no intersituational coherence in their norms of functioning, and 'closed families' are characterized by a rigidity of their rules and incapacity to adapt to the evolution of the context and to take into account external information. A similar interpretation can be found by Reiss, who contrasts 'consensus-sensitive families' with 'environment-sensitive families.' Olson's circomplex model, crossing dimensions of cohesion and regulation, seeks to show that extreme types of functioning (i.e., too rigid or too enmeshed) are more at risk than balanced types. These types of functioning vary according to the social status of the spouses: an increase in economic and cultural assets is linked to a more distinct emphasis on individual autonomy, a flexibility of regulation, and wide receptiveness of the group towards relationships and external information (see Kellerhals et al. 1994 for a synthesis).

3. *Distribution of Roles, Power, Justice*

Within the study of these modes of functioning, the question of role and power distribution has obviously occupied a central position. In the 1950s, the dominant idea was that small groups—that is, the nuclear family—generally engender a differentiation between instrumental roles (contribution of external resources, choice of the group's orientation) and expressive roles (organization of the internal production and the management of relationships), that brings about a higher efficiency of effort and turns out to be essential to the success of the socialization process (Parsons and Bales 1955). However, under the triple influence of feminism, the diversification of lifestyles, and the revisiting of ethnological analysis, the universality of this model has been contested and its functional character challenged by research showing that within a romantic marriage, the sharing of the expressive function and the equality of involvement is a determining factor of relational and educational success (Aldous 1977).

This 'functionalist consensus' was then overtaken by a rational utilitarian perspective—the resource theory (Blood and Wolfe 1960)— that understands the exercise of power in the family as a reflection of the inequality of socioeconomic resources between the spouses. According to this theory, the actor with higher outside resources exercises—despite possible egalitarian norms—a greater internal decision-making power; moreover, the comparative significance of the spouses' resources dictates their respective contribution to educational and domestic tasks. Criticism of this theory has been mostly concerned with the fact that (a) resources, far from being only economic, can also be charismatic, can concern expert knowledge, or be normative; (b) the weight of the latter vary according to the cultural context; and (c) the power of the other depends on the ego's alternatives. This triple determination can explain the permanency of sexism or domestic violence, as much as the unequal distribution of tasks in contemporary families. However, contemporary formulations of the problem emphasize the fact that by increasing the wife's potential in-

dependence and economic power, professional integration of women also increases her motivation and ability to speak for herself and require some consideration, while these elements will in turn reinforce the actor's power in negotiations (cf. Szinovacz 1987, Gelles 1995). The mechanistic exchange model is thus taken over by a sociocognitive formulation, where processes by which identity is constituted play an important part.

This transition from one paradigm to another has had repercussions particularly on the theorization of feelings of justice (fairness) in the family. The extent of inequality suffered by couples on a long-term basis tends to show that the simple equity theory—according to which feelings of injustice (unfairness) appear when the balance between one actor's contributions and rewards is different from another—is not always appropriate here. In fact, criteria of fairness vary according to status (social rank, age, sex), to the primary goals of the group, and to its type of cohesion. Three very different concepts of justice can thus be drawn (Kellerhals et al. 1997). The 'voluntarist' concept considers that justice comes when all concerned parties have been duly referred to and have been able to make an explicit or implicit agreement concerning the distribution of tasks or rewards in the family. Agreement between each other's will creates fairness. The 'finalist' concept defines the occurrence of fairness when the distribution of rights and duties brings positive outcomes to the individual or the group, with no primary consideration for merit or agreement. Finally, the 'comparative' concept considers that fairness occurs when people with the same status are treated in a similar way.

As we can see, these concepts of justice enable individuals to allow or make sense of very strong objective inequalities. They result from a collective construction of meaning, rather than from a static comparison between debts and bonds. This collective construction varies according to the type of cohesion of the group (normative families with a strong sense of unity tend to favor the 'comparative' concept, whereas more associative and negotiator families prefer the 'voluntarist' concept), and to social status ('comparative' concepts are more common at the bottom of social hierarchy, whereas 'voluntarist' ideas prevail at the top). Lastly, the spouses' concepts are not necessarily identical, so much so that the fairness/unfairness debate often seems to refer more to the confrontation of ideals, rather than to the objective disparity between contribution and retribution.

4. *The Educational Relationship*

Every family has to comply with an educational task that usually aims at providing children with social competencies (cognitive skills, social skills, and sense of autonomy/belongingness) essential to their social integration. However, in doing so, the parents' educational strategies vary greatly. They have often been characterized by their degree of permissiveness or control, on the one hand, and by their degree of warmth, commitment, and support (vs. lack of concern or hostility), on the other. It is on that basis that Baumrind, for instance, distinguishes between an authoritarian style (weak support and nurturance, quite strict and punitive authority, little rational communication), as opposed to a permissive style (high support and weak control, either rational or punitive), to a harmonious style where parents try to have influence on their children by egalitarian negotiation, and an authoritative style (high degree of support and control) which favors rational communication and the child's autonomy. Many researches show that an average degree of control and support are favorable to a good acquisition of social competencies and good self-esteem, whereas too much authority and punishment—or conversely too much protection—lead to blocks or rejections (cf. Rollins and Thomas 1979).

The parent's social status (education, occupation, income) affects educational strategies. Kohn (1977), for instance, shows that self-control, autonomy, and curiosity are given more value at the top of the hierarchy, whereas immediate conformity tends to be emphasized at the bottom. Concerning educational methods, it has been found that coercive methods are slightly more used in poorer backgrounds, with emphasis on motivation increasing along the social scale (Gecas 1979). According to Kohn (1977), these disparities reflect the occupational situation of the parents: The more their occupation necessitates initiatives, autonomy, and symbolization, the more their educational methods will value these elements, and vice versa.

Subsequent studies have improved the rather mechanistic aspect of these analyses by giving more importance to a mutual, joint, and socially constructed character of socialization. First, observation of infants has shown how their reaction can modify the mother's attitude. Moreover, the mutual ability of children and adults to establish synchronies, turn-taking, or behavior state matching proved to be an important factor of the attachment to, and the transmission of, norms (Peterson and Rollins 1987). Lastly, sociocognitive aspects, expectations relating to the 'good child,' and common knowledge relating to interpretations of deviant behavior model the adult's response to the child's actions and doings. This diversity of perspectives leads naturally to viewing the educational process as a site of interinfluence of the family microsystem, the institutional mesosystem, and the society macrosystem (Bronfenbrenner 1986). This 'ecological' perspective has notably shown how much the quality of the relationship between husband and wife affects the mother's relationship with her child (second-order effect). In the same way, it has also been proved that

educational style is an integral part of the family's type of cohesion: 'associative' families are likely to insist on autonomy and creativity, whereas merged families concentrate mostly on mimicry and conformity (Kellerhals et al. 1994). Then again, mechanistic models have been taken over by more symbolic interactionist perspectives.

5. *Families and Kinship Network*

Because families have so many tasks to accomplish, there is a question as to whether they generally face these problems on their own or not. In this regard, the idea of the modern family withdrawing into itself has not stood up to several studies on the theme of family solidarity which show that the kin group's support in the family management—purchase of property, child care, moral support, and so forth— is often essential, due to its flexibility, versatility, and speed (Pitrou 1992). However, these kinship ties are often characterized by a poor number of effective links, a clear primacy of vertical links (filiation) over horizontal links (sister/brother link), the predominance of the mother's kin, and by a much higher involvement of women. Kinship solidarity is, therefore, much affected by geographical distance, divorce, and professional occupations of both spouses, three points which limit availability.

Moreover, because kinship relationships often rely on affinity rather than on duty or status, and because the number of recomposed families following a divorce is multiplying, the kinship tie looks increasingly like a network (weak bond connections and absence of an important center point) rather than like a group concentrically focused on forebears and endowed with a strong collective consciousness (cf. Coenen-Huther et al. 1994).

6. *Adjustment and Divorce*

The study of family processes logically leads to research concerning reasons for the success or failure of the family faced with its tasks; in other words, concerning factors of adjustment. But what criteria of adjustment can be used?

Measuring marital and family adjustment in regard to the propensity for divorce creates confusion between the question of relational difficulties and economic or cultural obstacles preventing separation. Situations where marital stability coexist with disappointment often occur. However, it is important to try to define what are the factors of divorce. Before and during the 1960s, marital separation was usually interpreted with deficit models. Poor adjustment has therefore been viewed as a result of deficiency in the socialization to the other: couples getting married and having children very young, and also those whose parents have failed to pass on stable or harmonious models of marital union, have a higher rate of divorce than others. It has also been shown (Cherlin and Furstenberg 1994) that probability of divorce is higher in second unions, where there is often a lack of clear cultural models to guide relationships between generations.

Poor adjustment has also been related to a deficiency of symbolic or economic resources. The emphasis on happiness, communication, and consumption often found in contemporary family's plans demands financial resources that are not always available. Consequently, it is understandable that couples with low income and poor education separate more often, the distance between objectives and means being so large.

Poor adjustment has finally been viewed as the result of pronounced cultural differences between spouses, which finds expression in a lack of shared codes and values. A higher rate of divorce has therefore been noticed among mixed religion or nationality couples, or among heterogamous couples in age or social level (cf. Faust and McKibben 1999). However, once again, we should be aware that the social meaning of difference is probably even more important than the difference itself: hypergamous couples (where the man occupies a higher position than the woman does), for instance, are far more stable than hypogamous couples (where the situation is reversed), who are not so well accepted socially. But these different deficit models fail to explain the majority of contemporary divorces. More global, but also less verifiable, interpretations have been suggested, like Berger's, for whom too much expectation regarding the couple and the family explains the extent of disillusion. Other analyses (Popenoe, Bellah, Roussel) insist on the incompatibility of individualistic concern for personal genuineness with constraints related to community commitment; others also have placed the accent on the existing hiatus between egalitarian models of marital relationship and very unequal marital practices. In all these approaches, divorce appears to be more like an intrinsic property of contemporary couples, rather than linked to clearly identifiable individual deficiencies (Roussel 1989).

Given the biases of the 'divorce' indicator, there have been attempts—since initial studies by Burgess and Cotrell—to measure marital adjustment with a great many multidimensional scales (implying, e.g., degree of satisfaction, cohesion, consensus, and emotional expression) and to examine factors that affect this marital quality. These are due to the marriage's previous history (homogamy, socialization to appropriate role models), to the spouses' current resources (integration and community support, economic level, educational level), and to the nature of their interaction (quality of communication, appropriateness to roles) (Gelles 1995). One can believe that these different 'resources' make it possible to avoid extreme

forms of cohesion (enmeshment or disengagement) and of normative regulation (rigid or chaotic) which themselves limit the family's ability to cope with stress, to adapt to external changes, and to manage disruptions (cf. Olson et al. 1983, Boss 1988).

See also: Divorce and Gender; Divorce, Sociology of; Domestic Violence: Sociological Perspectives; Families and Households, Behavioral Demography of; Families as Educational Settings; Family and Gender; Family as Institution; Family Bargaining; Fatherhood; Generations, Relations Between; Kinship in Anthropology; Marriage; Marriage and the Dual-career Family: Cultural Concerns; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Socialization, Sociology of; Status and Role, Social Psychology of

Bibliography

Aldous J 1977 Family interaction patterns. *Annual Review of Sociology* **3**: 105–35

Aldous J 1996 *Family Careers*. Sage, Thousand Oaks, CA

Becker G 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Blood R O, Wolfe D H 1960 *Husbands and Wives: The Dynamics of Family Life*. Free Press, New York

Boss P G 1988 *Family Stress Management*. Sage, Newbury Park, CA

Bronfenbrenner U 1986 Ecology of the family as a context for human development: Research perspectives. *Development Psychology* **22**(6): 723–42

Burgess E W, Locke H J, Thomes M 1949 *The Family: From Institution to Companionship*. American Book, New York

Cherlin A J, Furstenberg F F 1994 Stepfamilies in the United States: A reconsideration. *Annual Review of Sociology* **20**: 359–81

Coenen-Huther J, Kellerhals J, von Allmen M 1994 *Les réseaux de solidarité dans la famille*. Réalités Sociales, Lausanne, Switzerland

Faust K A, McKibben J N 1999 Marital dissolution: Divorce, separation, annulment and widowhood. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*, 2nd edn. Plenum, New York

Gecas V 1979 The influence of social class on socialization. In: Burr W R et al. (eds.) *Contemporary Theories about the Family*. Free Press, New York, Vol. 1, pp. 365–404

Gelles R J 1995 *Contemporary Families*. Sage, Thousand Oaks, CA

Kaufmann J C 1993 *Sociologie du couple*. Presses Universitaires de France, Paris

Kellerhals J, Modak M, Perrenoud D 1997 *Le sentiment de justice dans les relations sociales*. Presses Universitaires de France, Paris

Kellerhals J, Troutot P Y, Lazega E 1994 *Microsociologie de la famille*. Presses Universitaires de France, Paris

Klein D M, White J M 1996 *Family Theories—An Introduction*. Sage, Thousand Oaks, CA

Kohn M L 1977 *Class and Conformity: A Study in Values*. University of Chicago Press, Chicago

Olson D H, Hamilton I, McCubbin 1983 *Families: What Make them Work?* Sage, Beverly Hills, CA

Parsons T, Bales R F 1955 *Family: Socialization and Interaction Process*. Free Press, New York

Peterson G W, Rollins B C 1987 Parent–child socialization. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*. Plenum, New York, pp. 471–507

Pitrou A 1992 *Vivre sans famille? Les solidarités familiales dans le monde d'aujourd'hui*. Privat, Toulouse, France

Rollins B C, Thomas D L 1979 Parental support, power and control technics in the socialization of children. In: Burr W R et al. (eds.) *Contemporary Theories about the Family*. Free Press, New York, Vol. 1, pp. 317–64

Roussel L 1989 *La famille incertaine*. Odile Jacob, Paris

Singly F de 1997 *Le soi, le couple et la famille*. Nathan, Paris

Sussman M B, Steinmetz S K, Peterson G W (eds.) 1999 *Handbook of Marriage and the Family*, 2nd edn. Plenum, New York

Szinovacz M E 1987 Family power. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*. Plenum, New York

Teachman J D, Polonko K A, Scanzoni J 1999 Demography and families. In: Sussman M B, Steinmetz M K, Peterson G W (eds.) *Handbook of Marriage and the Family*, 2nd edn. Plenum, New York

J. Kellerhals

Family Size Preferences

Preferred or desired family size is the number of children wanted in one's lifetime. Family size preferences are viewed as measures of the demand for children which, in combination with the supply of children and contraception determines the number of children born. Family size preferences may change over the life course, especially in response to childbearing and childrearing experience.

Family size preferences or desires are conceptually and empirically distinct from family size ideals, norms, intentions, and expectations. Fertility ideals and norms represent what is desirable for a group of people or a typical group member, rather than what is desirable for every person in the group. Family size intentions or expectations reflect not only family size preferences but also constraints on one's ability to achieve desired or preferred goals.

1. Conceptualization and Measurement

1.1 Desired, Intended, Expected, and Ideal Family Size

McClelland assessed the construct desired family size as a measure of the demand for children, that is, 'the

number of children parents would have if there were no subjective or economic problems involved in regulating fertility' (1983, p. 288). He concluded that family size desires do, indeed, represent the theoretical construct, demand for children.

Family size intentions and expectations reflect not only family size preferences but also perceptions of the likelihood that preferences can be translated into outcomes. The gap between intended or expected family size and desired family size may reflect influence of partner's desires, perceived fecundity and fertility control, or other known constraints on fertility behavior or outcomes.

Questions about a general ideal family size produce responses that fit the character of a social norm rather than that of a personal preference. For some individuals—those who have internalized societal ideals or who believe their own situation is typical—ideals may be identical to desires (Mason 1983).

1.2 Measures of Family Size Preferences

Two types of questions are most commonly used to measure family size preferences. The first asks respondents how many children they would want to have in their lifetime, if they could choose the exact number. For childless respondents, this question measures desires prospectively; for parents, the question includes a retrospective assessment of previous births with a prospective assessment of potential future births. Retrospective assessments of birth wantedness are subject to social desirability bias or rationalization and therefore produce upwardly biased estimates of desired family size. Another version of this question attempts to give respondents permission to report previous births as unwanted by asking something like: 'If you could start life over again, how many children would you want to have?' Such questions may also place conditions on the question, specifying that the respondent's life would be in other respects about the same as it had been so far. It is not clear that these two forms of the lifetime question produce different responses.

The second type of question asks respondents how many more children they would like to have; for childless respondents, the answer would presumably be the same as for the first type of question, but for parents, retrospective assessments of previous births are excluded. To obtain an estimate of desired family size, some analysts simply add the respondent's current number of children to the additional number desired. This strategy assumes that all previous births were wanted, producing an even stronger upward bias than does the first type of measure. Several methods have been proposed for estimating the number of unwanted previous births from parity-specific values of additional desired children, and thereby to produce a more accurate measure of desired family size (Bongaarts 1998). Unfortunately, these methods do not provide estimated family size preferences for individuals.

An alternative measure of family size preferences asks respondents to make choices between pairs of family sizes, producing a scale reflecting desires for large versus small families. Scales have also been developed to incorporate preferences for sex of children in family size preferences and vice-versa (Coombs 1979). The interdependent IN and IS scales reflect the fact that family size desires may incorporate desires for particular numbers of boys or for particular numbers of girls or some balance between the two.

1.3 Non-numerical Responses

In high-fertility societies, small but significant proportions of respondents do not provide a numerical response to questions about desired family size, reporting instead that the number of children they have is 'up to God' or that they will accept whatever number of children they have. These responses raise the question of whether individuals form preferences about things they believe they cannot or should not try to influence. Such respondents do respond to questions about family size ideals, preferences for one family size over another, and unwanted or mistimed previous births (McClelland 1983, Riley et al. 1993).

2. Her, His, and Their Family Size Desires and Intentions

Most data on family size desires or intentions come from interviews with women. The assumptions underlying women-only fertility surveys are that (a) women and men share the same conditions of life and will therefore want and intend the same number of children, (b) because women bear children and are primary childrearers, their desires and intentions are those that determine fertility, and (c) women's desires and intentions will incorporate those of their partners, reflecting resolution of any initial disagreements. These assumptions appear to be tenable in the aggregate; when data are collected from women and men, their respective family size desires and intentions are more similar than different (Mason and Taj 1987). On the other hand, data collected from couples consistently show that substantial proportions of couples disagree about desires or intentions for further childbearing or completed family size (Becker 1996).

3. Family Size Preferences, Intentions, and Births

The relationship between family size desires and births depends on relationships between preferences and intentions, between intentions and behavior, and

between behavior and outcomes. Couple disagreement is a primary source of gaps between desires and intentions, that is, the gap between individual desires and intentions to have a child is explained in large part by a spouse's different preference (Thomson 1997). Another key source of gaps between desires and intentions is the perception of control over fertility. Those who believe themselves to be infecund may desire more children than they intend; those who do not perceive themselves to be able to control fertility may intend more children than they desire.

Family size preferences or intentions may also change over the life course and are particularly likely to be influenced by experiences with previous births (Namboodiri 1983). In spite of these expected gaps, measures of family size desires, particularly the desire for additional children, have moderately strong effects on fertility behavior and outcomes (McClelland 1983). Lee (1980) has demonstrated that a 'moving target' model of desired and achieved fertility produces even stronger correspondence between family size desires and period fertility.

Davidson and Jaccard (1979) identified many of the key conditions under which a strong effect of fertility desires or intentions on births should be expected. First, the strength of the effect depends on the extent to which the ultimate goal (birth or no birth) is achieved through a sequence of behaviors and events, including sexual intercourse, contraception, conception, and abortion. Fertility desires and intentions should have stronger effects on the behaviors at the beginning of the sequence than on the ultimate outcome. Second, the effect will be stronger over shorter versus longer periods of time because the opportunity for new experiences that change desires or intentions is reduced. Third, effects of desires or intentions on behavior and outcomes should be stronger for individuals who place a high value on cognitive consistency, who believe they have control over their actions, and who generally engage in planful behavior than for those who do not. Finally, desires and intentions and behaviors and outcomes should be measured at the same level of specificity in order to expect a strong relationship; desired or ideal family size should not, for example, be expected to have strong effects on contraceptive use or even the timing of a particular birth.

The relationship between childbearing desires or intentions and births depends to a considerable extent on couple agreement. Couple disagreement leads to birth rates between those for couples who both desire or intend a child and those who do not. This means that both partners' desires have at least some influence on births (e.g., Thomson 1997, Thomson and Hoem 1998, Dodoo 1998).

Estimated relationships between fertility desires and intentions or between intentions and behavior also depend on measurement quality. The IN-scale discussed above provides more accurate measures of preferences and therefore has larger estimated effects on fertility behavior than does a single most preferred number of children (Coombs 1979). Measures of the desire for children that incorporate dimensions of attitude strength have stronger effects on childbearing intentions and behavior than do yes-no-don't know responses to direct questions about desire for children (e.g., Davidson and Jaccard 1979). Similarly, intentions measured with some degree of certainty or uncertainty are better predictors of outcomes than a simple dichotomy (Thomson 1997).

See also: Children, Value of; Demographic Transition, Second; Family as Institution; Family Processes; Family Systems and the Preferred Sex of Children; Family Theory and the Realities of Childbearing Behavior; Family Theory: Economics of Childbearing; Fertility Change: Quantum and Tempo; Fertility Control: Overview; Fertility of Single and Cohabiting Women; Fertility: Proximate Determinants; Reproductive Rights in Affluent Nations

Bibliography

Becker S 1996 Couples and reproductive health: A review of couple studies. *Studies in Family Planning* **27**: 291–306

Bongaarts J 1998 Fertility and reproductive preferences in post-transitional societies. Policy Research Division Working Paper No. 114, Population Council, New York

Coombs L C 1979 Reproductive goals and achieved fertility: A fifteen-year perspective. *Demography* **16**: 523–34

Davidson A R, Jaccard J J 1979 Variables that moderate the attitude-behavior relation: Results of a longitudinal survey. *Journal of Personality and Social Psychology* **37**: 1364–76

Dodoo F N 1998 Men matter: Additive and interactive gendered preferences and reproductive behavior in Kenya. *Demography* **35**: 229–42

Lee R D 1980 Aiming at a moving target: Period fertility and changing reproductive goals. *Population Studies* **34**: 205–26

Mason K O 1983 Fertility norms. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries, Vol. 1, Supply and Demand for Children*. Academic Press, New York

Mason K O, Taj A M 1987 Differences between women's and men's reproductive goals in developing countries. *Population and Development Review* **13**: 611–38

McClelland G H 1983 Family-size desires as measures of demand. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries, Volume 1, Supply and Demand for Children*. Academic Press, New York

Namboodiri N K 1983 Sequential fertility decision making and the life course. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries, Volume 1, Supply and Demand for Children*. Academic Press, New York

Riley A P, Hermalin A I, Rosero-Bixby L 1993 A new look at the determinants of nonnumeric response to desired family size: The case of Costa Rica. *Demography* **30**: 159–74

Thomson E 1997 Couple childbearing intentions and births. *Demography* **34**: 343–54

Thomson E, Hoem J M 1998 Couple childbearing plans and births in Sweden. *Demography* **35**: 315–22

E. Thomson

Family Systems and the Preferred Sex of Children

1. Introduction

Most human societies exhibit some degree of gender inequality, but there is enormous variation in its nature and manifestations. In some societies, women face only mild or moderate constraints on functioning as socially and economically autonomous persons in the public life of the community; in others, women are largely subordinate to males. This has implications for any old-age support parents can expect from daughters, which, in turn, affects the aspirations of couples regarding the sex of their children. In most regions of the world, parents express a preference for a gender-balanced family—with, perhaps, mild preferences for children of a particular gender. However, in some regions parents exhibit a significant degree of son preference. These preferences manifest themselves in decision making at the margin in many ways—ranging from investing more in educating sons to the extreme outcome of excess mortality among daughters.

To understand why some societies have strong son preference while others do not, we need to examine the gap between sons' and daughters' ability to contribute to the physical, emotional, and financial well-being of their parental household—as this is the unit primarily responsible for making and implementing decisions about childbearing and childrearing. This gap is determined largely by kinship systems, which specify clearly defined roles for different categories of family members. In most societies, it is possible for women to maintain mutually supportive relations with their parental household even after marriage, making for little difference in the value of girls and boys to their parents. However, in other societies women's links with their parental household are largely severed at marriage and sons are enjoined to care for their parents—making it clearly more attractive for parents to rear sons instead of daughters. Gender-appropriate roles are also decided, to a lesser extent, by other norms—such as religious beliefs and injunctions, which can, for example, discourage women from participating in the labor market and thus reduce their ability to help their parents financially.

In this entry we review the broad features of family systems in different cultural regions of the developing world to describe how far daughters are in a position to contribute to parental well-being—both in absolute terms, and relative to what is expected from sons. We then explore how these sex preferences are reflected quantitatively in survey data on family-building desires and in measures of excess female child mortality.

2. Family Systems

The strongest son preference is evident in societies which are rigidly patrilineal and patrilocal. Patrilineality implies that group membership and productive assets are passed on through the male line. Patrilocality implies that it is normative for a woman to take up residence with her husband's family after marriage—often outside her parental village. In these societies, in which both inheritance and marriage rules are heavily weighted in favor of men, women have little economic independence or autonomy. Furthermore, since daughters move away from their birth homes after marriage, parents can have little expectation of subsequent physical or emotional sustenance from them since they are not living in close proximity. Even if women are educated and employed, the fruits of their labor are monopolized by their husband's family, and parents cannot claim a share of their daughters' earnings. On the other hand, a son is expected to provide his parents with economic security and insurance. He is also expected to bring in a daughter-in-law who will take care of his parents' physical needs in their old age. In these societies, women also gain in prestige and status if they bear sons to continue the family line, so both women and men have a strong interest in producing sons.

In some patrilineal societies, there are additional norms which increase the burden to parents of raising daughters. Religious beliefs and social customs such as dowry and 'purdah' serve as further instruments of patriarchal control, and reinforce son preference. Furthermore, family honor often depends on the purity of patriarchal descent, which requires ensuring women's sexual purity. This often entails secluding women from outside activities—which implies reduced income-earning opportunities—and marrying them off at young ages, before they have had a chance to develop into independent people.

In contrast, where kinship rules and norms make fewer distinctions between males and females, women have greater voice in the household and in public spaces, and face fewer constraints in becoming independent economic and social actors. In these societies, women are not constrained from providing old-age support to parents, and are valued accordingly.

2.1 South and East Asia

Much of South Asia displays the patterns of patrilineal and patrilocal kinship systems. This is especially

strong in the northern part of the subcontinent. The widespread practice of dowry payments constitutes a drain on family resources, further raising the relative cost of a daughter (Miller 1981). In Bangladesh, as in much of the rest of the subcontinent, daughters are perceived as burdens due to the custom of dowry and their limited income opportunities, while sons are preferred due to their higher earning potential (Muhuri and Preston 1991).

Much of East Asia is also characterized by rigidly patrilineal kinship organization, although in recent decades women have become educated along with men and have been active in the industrial labor force. Patrilineality is most pronounced in South Korea and amongst the majority Han population of China (Goodkind 1996). Moreover, it is believed that only sons can give their parents the crucial support they need in their afterlife, to prevent their being hungry spirits for eternity. This puts additional pressure on people to have sons even if pension plans and other mechanisms for providing for one's old age are in place, as they cannot provide for the afterlife—this can only be done by sons. (Japan is an exception to the East Asian pattern, practicing a very flexible form of patrilineal kinship organization which permits daughters under many circumstances effectively to substitute for sons; Nakane 1967.)

However, in Kerala and Sri Lanka, kinship systems allow women greater flexibility to interact with their parents after marriage than in the north (Dyson and Moore 1983). Village endogamy and cross-cousin marriages are favored, allowing women to live closer to their parents. Women sometimes transfer and inherit property as well. Thus there is considerable scope for mutually supportive relations between a married woman and her parental household.

2.2 Southeast Asia

Most Southeast Asian societies have bilateral kinship systems, which allow people to maintain strong links with both their family of birth and their family of marriage. Reciprocal obligations and expectations are more flexible than in strongly patrilineal societies. In much of Southeast Asia—as in Thailand, Indonesia, and the Philippines, for example—women have traditionally had considerable social and economic autonomy, playing a significant role in marketing and other independent income-generating activities. This makes it possible for them to provide old-age support for their parents should it be needed. In some cases, daughters are the main source of support—in Northern Thailand, for example, the norm is for newly married couples to live with the wife's parents for some years, and for the last daughter to stay on and inherit the land and house after the parents' death (Limanonda 1995). Although Islam has been actively practiced in several of these societies for some centuries, this had relatively little impact on traditional kinship systems, as documented for Malaysia (Kling 1995).

On the other hand, the Chinese populations of Southeast Asia—including those in Malaysia and Singapore—display patrilineal and patriarchal kinship systems (Wongboonsin and Ruffolo 1994). Vietnam is an interesting amalgam of kinship systems, with strongly patrilineal values stemming from the incursion of Confucian ideologies from the north, superimposed on more bilateral Southeast Asian kinship systems, resulting in moderate levels of son preference (Goodkind 1994). As in China, sons alone can perform traditional rights of ancestor worship.

2.3 Middle East and North Africa

North Africa and the Middle East are characterized by patrilineal systems in that the family line is traced through males, but people are permitted to marry within their lineage, including cousins. This means that women often live in close proximity to their own parents, and can be a source of physical and emotional support to them. Economic reasons for son preference also do not appear to be strong—in Egypt, for example, parents value the labor of both sons and daughters, marriage expenses may actually be higher for males, and daughters can remain close to their parents after marriage (Makinson 1986). There are significant groups of educated and employed women in most of these countries, especially amongst the middle classes (Rostam-Kayali 1999).

However, the extent to which daughters can provide material support is constrained by the norms of Islam and the Arab family, whereby the husbands are the dominant decision makers in the household and most women are discouraged from active participation in the world outside the home (Obermeyer 1995). Normative pressures for bearing sons is strong, and women derive social status while men continue their name and lineage through sons. Overall, women are not as well placed as in Southeast Asia to contribute to their parental households, but not nearly as constrained from doing so as in East Asia and the northern parts of South Asia—making for moderate rather than strong son preference in the region.

2.4 Sub-Saharan Africa

Sub-Saharan Africa includes societies which are patrilineal, bilateral, and matrilineal. Despite these apparent differences, they have a great deal in common in terms of gender relations. Men tend to be dominant in the household—even in several of the matrilineal societies, property is passed through the female line,

but from men to men (Kevane and Gray 1996). However, conjugal bonds in sub-Saharan Africa are weaker than in South Asia, and women have a higher degree of economic independence (Cain 1984). They have freedom to remarry upon divorce or widowhood (Bledsoe 1990). Marriage often involves bride-price from the groom to the kin of the prospective bride, rather than dowry (Oppong 1987).

Women typically function easily as independent economic actors who can operate comfortably in public arenas—whether as self-employed market women, or as managers of their own farm plots. In several West African societies, women are very active traders and are completely accustomed to being the mainstay of the household, through marital union and dissolution. There is substantial flexibility in terms of those whom people can choose to live with and support, and parents have little reason to be desperate about old-age support if they do not have sons. Hence, one is unlikely to see strong preferences for children of a particular sex on the part of parents.

2.5 *Latin America*

Central and South American societies are formed from the confluence of many different ethnic groups, each with their own customs and social organization. For all their differences, however, these societies derive from cultures which are not heavily patrilineal or gender differentiated. The dominant culture is that of those of European origin, mixed with those of sub-Saharan African origin, neither of which place rigid constraints on women's participation in public life and interaction with their parental household. The Latin American family structure allows women a high degree of female autonomy. The conjugal bond is weaker than in countries of South Asia, the Middle East, and North Africa (Cain 1984). Childbearing outside formal marriage or stable unions is common, and women may spend much of their reproductive lives in consensual unions with men. Economic dependence of women on men is inconsistent with such family patterns. The result is that there is a strong tradition of male machismo, but there are few rigid rules constraining women from participating actively in social and economic life, or from helping their parents. Thus, there is little reason to expect strong gender preferences among parents—and whether or not they receive support from daughters is a matter of choice and circumstance, rather than one of custom and ability.

3. *Quantitative Evidence of Son Preference*

Son preference is manifested in quantitative data in various ways. In this section, we focus on two measures—survey data on desire for additional children by number and sex of living children, and the sex ratios of children across developing countries—to examine the correspondence between family systems and these measures of revealed son preference.

As an indicator of sex preference among currently married nonpregnant women, the Demographic and Health Surveys (DHS) asked respondents whether or not they wanted another child, by sex distribution of their living children. The results from the latest available country studies are reproduced in Table 1, for women with two and three living children of the same sex. (This indicator is limited in that it focuses on a selected sample of women with a specific number of children rather than on the entire population of women of childbearing age, but is suggestive, nonetheless.) Son preference is indicated if, for a given number of children, a lower percentage of women wanted an additional child if all their children were sons, as compared with if they were all daughters. The number of children at which such preferences manifest themselves depends on the overall level of fertility in the country—it is only when people approach their desired family size that they are compelled to make decisions on additional childbearing to achieve their desired sex composition. Thus, where overall levels of fertility are high—as in some countries of South Asia, the Middle East, and North Africa, for example—gender preferences manifest themselves at higher numbers of children than where fertility levels are low.

Another way to measure sex preference is to examine the sex ratios of children aged below five years (Fig. 1). This reflects the extent to which the stated preference for sons actually translates into significant levels of excess female child mortality. Excess female child mortality can result from discrimination before birth (sex-selective abortion), at birth (female infanticide), and during early childhood (most commonly through neglect in providing timely health care). These are all motivated by the desire to rear more sons than daughters, and which method is used depends largely on the technological choices available.

There is no single 'expected' ratio of boys to girls, as this depends on the level of mortality. The observed sex ratio at birth is around 1.05 on average, and excess male mortality in early infancy is the biological norm (Waldron 1998). A recent study of 82 developing countries that had populations in excess of a million in 1990 confirms that mortality among male infants (aged below a year) is higher than that of females, the median sex ratio of infant mortality being 1.18—although actual figures vary around this median, with lower ratios in societies with higher life expectancy (United Nations Secretariat 1998). In the age-group one to four, the median sex ratio of child mortality is about 1. Based on this information, sex ratios among children aged zero to 4 can be expected to be below 1.05, *on average*. We have used a cut-off ratio of 1.045 (see Fig. 1), and sex ratios above this expected value

Table 1
Percentage of currently married nonpregnant women who want another child, by number and sex of living children

	Two Children			Three Children		
	No sons	*2 sons*	*2 sons/No sons*	*No sons*	*3 sons*	*3 sons/No sons*
South Asia						
Bangladesh	65	39	0.60	48	20	0.42
India	57	22	0.39	50	11	0.22
Nepal	77	25	0.32	66	10	0.15
Pakistan	77	61	0.79	79	33	0.42
Sri Lanka	47	36	0.77	18	16	0.89
Unweighted mean			0.57			0.42
Southeast Asia						
Indonesia	47	50	1.06	35	26	0.74
Philippines	44	46	1.05	24	30	1.25
Thailand	33	26	0.79	19	20	1.05
Unweighted mean			0.97			1.02
Middle East and North Africa						
Egypt	48	38	0.79	27	18	0.67
Jordan	75	66	0.88	76	55	0.72
Morocco	81	68	0.84	72	38	0.53
Tunisia	66	58	0.88	56	33	0.59
Turkey	25	13	0.52	18	4	0.22
Unweighted mean			0.78			0.55
sub-Saharan Africa						
Botswana	75	75	1.00	64	52	0.81
Burkina Faso	89	91	1.02	76	83	1.09
Burundi	92	91	0.99	89	78	0.88
Cameroon	79	83	1.05	84	73	0.87
Central African Republic	76	73	0.96	72	68	0.94
Ghana	81	80	0.99	60	52	0.87
Kenya	68	69	1.01	47	45	0.96
Liberia	85	87	1.02	95	82	0.86
Madagascar	83	86	1.04	65	70	1.08
Malawi	75	69	0.92	62	68	1.10
Mali	79	82	1.04	79	70	0.89
Namibia	52	65	1.25	46	61	1.33
Niger	88	92	1.05	75	82	1.09
Nigeria	79	83	1.05	82	76	0.93
Rwanda	91	80	0.88	72	71	0.99
Senegal	92	88	0.96	86	81	0.94
Sudan	86	86	1.00	81	77	0.95
Togo	95	91	0.96	83	85	1.02
Uganda	96	95	0.99	88	99	1.13
Zambia	90	90	1.00	87	89	1.02
Zimbabwe	82	79	0.96	63	73	1.16
Unweighted mean			1.01			1.00
Latin America						
Bolivia	29	28	0.97	19	16	0.84
Brazil	31	31	1.00	15	15	1.00
Colombia	26	36	1.38	8	13	1.63
Dominican Republic	42	49	1.17	15	13	0.87
Ecuador	39	36	0.92	26	23	0.88
Guatemala	56	58	1.04	43	41	0.95
Mexico	37	37	1.00	24	12	0.50
Paraguay	57	63	1.11	47	53	1.13
Peru	23	27	1.17	14	14	1.00
Trinidad and Tobago	36	44	1.22	24	29	1.21
Unweighted mean			1.10			1.00

Source: Arnold 1997
For countries with results from multiple years, the latest year is reported

Map 1: Juvenile Sex ratios in Africa, Asia, and Latin America, 2000

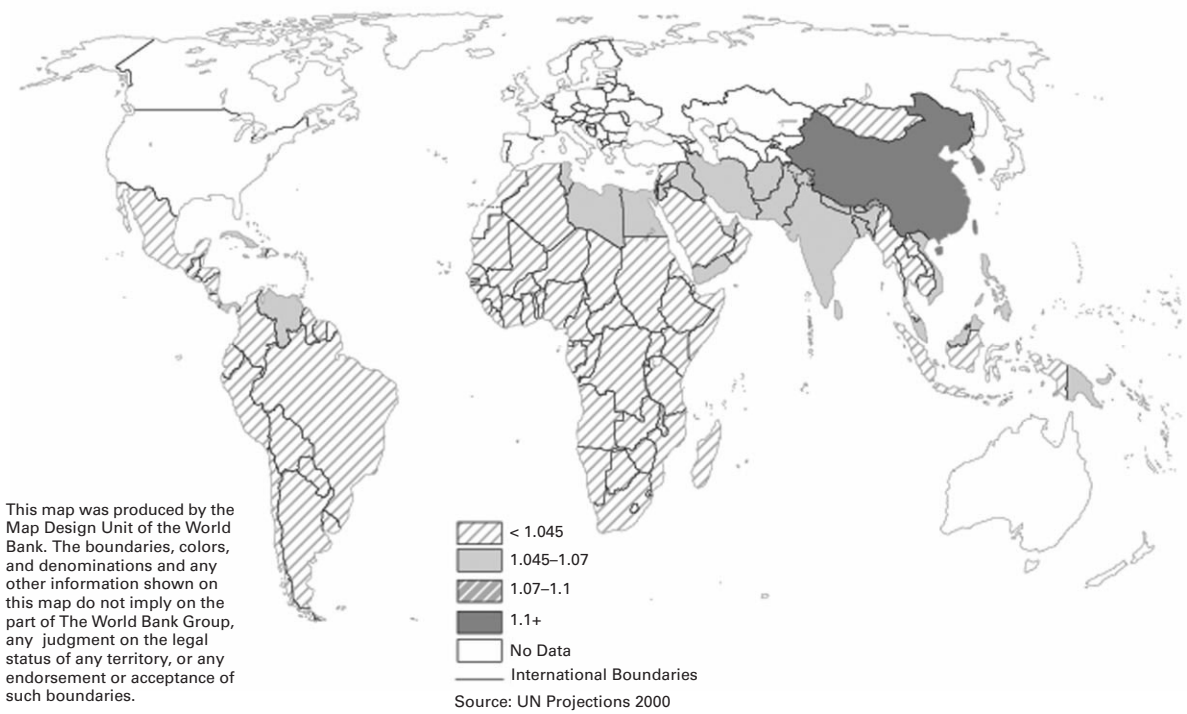

Figure 1
Juvenile sex ratios in Africa, Asia, and Latin America, 2000

can be viewed as indicative of excess female child mortality. The higher the observed sex ratio above this cut-off, the stronger the manifestation of son preference.

The data presented in Table 1 and Fig. 1 are consistent with the broad regional differences in family systems outlined above. In Latin America, sub-Saharan Africa, and Southeast Asia, where there is a higher degree of female autonomy, there is little consistent evidence of son preference in either juvenile sex ratios or desire for additional children. In Latin America, there are actually instances of a mild preference for daughters, in countries such as Colombia, Trinidad and Tobago, and Paraguay (Table 1).

On the other hand, in regions of the world where patrilineal and patrilocal family systems prevail, the data reveal a distinct preference for sons. Son preference in the desire for additional children is evident in all countries in the Middle East and North Africa, and even more strongly in South Asia, at a parity of 2 (Table 1). (Since the DHS surveys were not conducted in East Asia, these observations are missing from the table.) These preferences are strengthened at a parity of 3—reflecting that in high-fertility regions, son preferences are more evident at higher parities. (Only Sri Lanka in South Asia displays a different pattern; see Table 1.) These regions, along with China and South Korea in East Asia, have distinctly higher juvenile sex ratios than average as well—suggesting excess female mortality (Fig. 1).

Estimates indicate that in China and South Korea, around 7 percent of girls were 'missing' in 1990 on account of excess prenatal or postnatal mortality (Das Gupta et al. 1997). Globally, estimates indicate that there are between 60 and 100 million fewer women than there would be in the absence of gender discrimination—mostly in South and East Asia (Coale 1991, Klasen 1994, Sen 1992). (These estimates are based on comparing the actual sex ratio of the population against the sex ratio that would theoretically exist in the absence of gender discrimination.)

Since China and India are large countries with considerable cultural heterogeneity, it is instructive to look at province-level data (Fig. 2). This shows patterns which accord further with the description of family systems above. In India, the northwest shows sex ratios as excessively masculine as South Korea and the mainstream Han-populated provinces of China. South India shows less discrimination against girls, as do the interior provinces of China, where minority

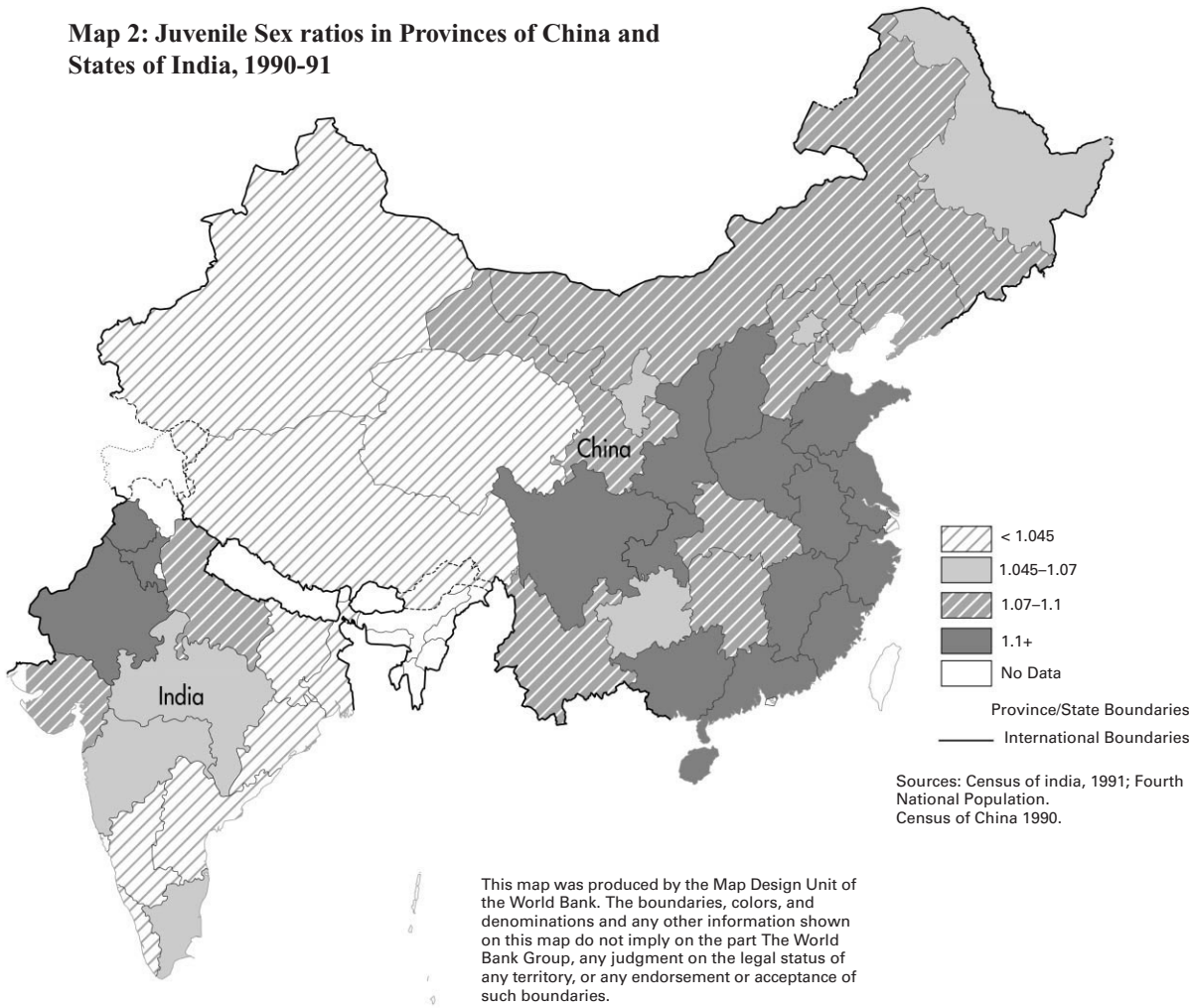

Figure 2
Juvenile sex ratios in provinces of China and states of India, 1990–1

ethnic groups are clustered. Thus the strongest evidence of excess female child mortality is from the regions of East and South Asia characterized by rigidly patrilineal kinship systems. Similar consistency between kinship systems and son preference is evident in the data from Malaysia, where the Malays with their Southeast Asian patterns of bilateral kinship show little sign of excess female child mortality, while the Chinese and Indians show it clearly (Goodkind 1996).

4. *Other Factors Affecting Son Preference*

Although son preference is culturally determined, the extent to which the preferences within a given society are translated into actual discrimination against daughters is influenced by the pressures operating on parents at a given point in time. Such pressures can take many forms. For example, in times of adverse income shocks, investment in girls may suffer more than investment in boys. (See, for example, Behrman and Knowles 1998 for Vietnam, which suggests that the income elasticity of investment in daughters is higher than that of investing in sons. This implies that if household income falls, discrimination against daughters will rise.) It is not that parents are unwilling to invest in children, but that they would prefer to invest less in daughters who will be subsequently lost to them.

External crises, such as famine and war, increase the pressures on parents to discriminate against daughters, as the difficulties of nurturing all household members

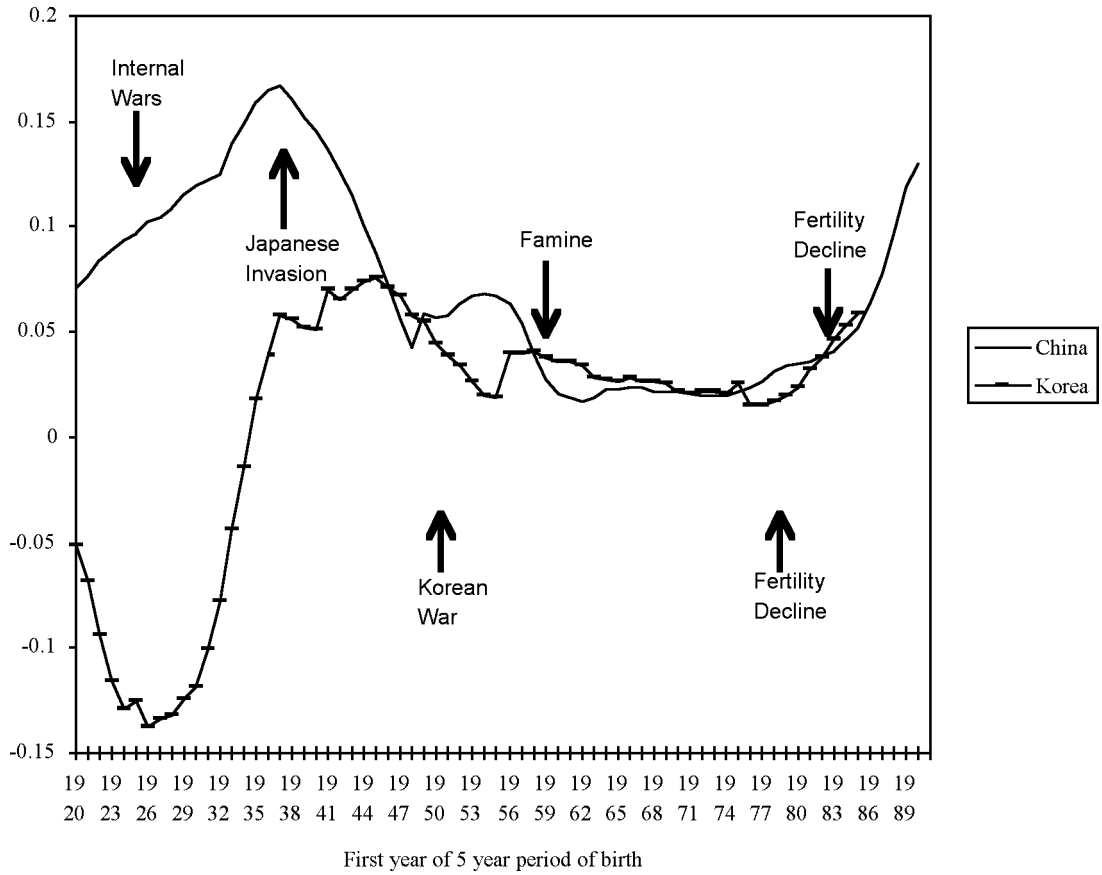

Figure 3
Excess sex ratios at census time by five-year birth cohorts in China and South Korea, 1920–95
Note: The excess ratios peak amongst cohorts born just before a war or famine, because those who were young girls at the time of the crisis experienced the maximum excess mortality
Source: Das Gupta and Li 1999

rise and harsher choices have to be made. Fig. 3 shows how the extent of actual discrimination against girls was affected by wars and famines in China and South Korea during this century (Das Gupta and Li 1999). (China experienced many severe crises during the twentieth century—during the first half of the century, civil war and invasion in China led to an enormous amount of disruption, which was followed by the famine of 1959–61.) The South Korean data show a massive deficit of males associated with the loss of young men during the Korean War in the early 1950s, followed by a rise in sex ratios as fertility declined.

Fertility decline also adds to the pressures to discriminate against daughters. When parents desire a smaller number of children, this further limits the number of daughters they would like to rear. If, for example, they want to have no more than two children, and the first is a girl, they are under severe pressure to ensure that the next is *not* a girl. Thus child sex ratios were observed to become more masculine when fertility declined significantly in India, China, and South Korea (Das Gupta et al. 1997).

5. *Conclusions*

This entry provides a broad overview of family systems in the developing world, and how they influence sex preference for children. In general, evidence for son preference is strongest where women are not in a position to contribute to the well-being of the family.

See also: Family and Gender; Family, Anthropology of; Family Size Preferences; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Fertility Control: Overview; Fertility: Proximate Determinants; Gender-

related Development; Infant and Child Mortality in the Less Developed World; Parenting: Attitudes and Beliefs; Sex Preferences in Western Societies

Bibliography

Behrman J R, Knowles J C 1998. Household income and child schooling in Vietnam. *World Bank Economic Review* (*International*) **13**: 211–56

Bledsoe C 1990 Transformation in sub-Saharan African marriage and fertility. *Annals of the American Academy of Political and Social Science* **510**: 115–25

Cain M 1984 *Women's Status and Fertility in Developing Countries: Son Preference and Economic Security*. World Bank, Washington, DC

Coale A J 1991 Excess female mortality and the balance of the sexes: An estimate of the number of 'missing females'. *Population and Development Review* **17**: 517–23

Das Gupta M, Li S 1999 Gender bias and the marriage squeeze in China, South Korea and India 1920–1990: The effects of war, famine and fertility decline. *Development and Change* **30**: 619–52

Das Gupta M, Jiang Z, Xie Z, Li B, Nam-Hoon C, Woojin C, Mari Bhat P N 1997 *Gender Bias in China, South Korea and India: Causes and Policy Implications*. Harvard University Press, Cambridge, MA

Dyson T, Moore M 1983 On Kinship structure, female autonomy and demographic behavior in India. *Population and Development Review* **9**: 35–60

Goodkind D M 1994 Sex preference for children in Vietnam. Paper presented at the International Symposium on Issues Related to Sex Preference for Children in the Rapidly Changing Demographic Dynamics in Asia, Seoul, Republic of Korea, 21–4 November

Goodkind D M 1996 On substituting sex preference strategies in East Asia: Does prenatal sex selection reduce postnatal discrimination? *Population and Development Review* **22**: 111–25

Kevane M, Gray L 1996 'A woman's field is made at night': Gendered land rights in and norms in Burkina Faso. *Feminist Economics* **5**: 1–26

Klasen S 1994 'Missing Women' reconsidered. *World Development* **22**: 106–71

Kling Z 1995 The Malay family: Beliefs and realities. *Journal of Comparative Family Studies* **26**: 43–66

Limanonda B 1995 Families in Thailand: Beliefs and realities. *Journal of Comparative Family Studies* **26**: 67–82

Makinson C 1986 *Sex Differentials in Infant and Child Mortality in Egypt*. Ph.D. thesis, Princeton University

Miller B D 1981 *The Endangered Sex: Neglect of Female Children in Rural North India*. Cornell University Press, Ithaca, NY

Muhuri P K, Preston S H 1991 Effects of family composition on mortality differentials by sex among children in Matlab, Bangladesh. *Population and Development Review* **17**: 415–34

Nakane C 1967 *Kinship and Economic Organization in Rural Japan*. Athlone Press, London

Obermeyer C M (ed.) 1995 *Family, Gender and Population in the Middle East*. American University in Cairo Press, Cairo, Egypt

Oppong C (ed.) 1987 *Sex Roles, Population and Development in West Africa: Policy-related Studies on Work and Demographic Issues*. Heinemann, Portsmouth, NH

Rostam-Koyali J 1999 The politics of women's rights in the contemporary Muslim world. *Journal of Women's History* **10**: 205–15

Sen A 1992 More than 100 million women are missing. *New York Review of Books* **December**: 61–6

United Nations Secretariat 1998 Levels and trends of sex differentials in infant, child, and under-five mortality. In: *Too Young to Die: Genes or Gender*. United Nations, New York, chap. 4, pp. 84–108

Waldron I 1998 Sex differences in infant and early childhood mortality: Major causes of death and possible biological causes. In: *Too Young to Die: Genes or Gender*. United Nations, New York, chap. 3, pp. 64–83

Wongboonsin K, Ruffolo V P 1994 The future of Thailand's population policy: Potential directions. Paper presented at the International Symposium on Issues Related to Sex Preference for Children in the Rapidly Changing Demographic Dynamics in Asia, Seoul, Republic of Korea, 21–4 November

A. Basu and M. Das Gupta

Family Systems in Europe

Contributing to a long tradition of scholarship on family life in Europe, a group of scholars in the 1960s redirected research by arguing that a unique system characterized the traditional European family. They criticized what they took to be the received wisdom on European family life in the past, which portrayed the nuclearization of the coresidential family as a recent product of industrialization and urbanization. The final decades of the twentieth century saw increasing criticism of this revisionist perspective, leading to a more nuanced and theoretically sophisticated understanding of the nature of coresidential processes in the European past.

1. *A European Family System*

The debate over the existence of a European family system has revolved around two related elements—a marriage system and a household system—emphasizing the demographic aspects of each. The former focuses on the age at which men and women marry, as well as the proportions who never marry. The latter concerns the principles by which coresidential units form and dissolve. The two are linked via the principles governing postmarital residence, that is, where newlyweds live following marriage.

These are only two aspects of what could be seen as the larger matrix of a family system. The anthropologist Goody (1983) has called attention to a variety of features that together historically differentiated European family dynamics from those found elsewhere in the world—including a prohibition on poly-

gamy and child adoption, and various restrictions on marriage to kin—differences which he attributes in good part to the influence of the Church. An understanding of family systems more generally would also include consideration of kin ties beyond the household; however, relatively little empirical attention has been focused on such issues historically.

1.1 From Le Play to Laslett

Most prominent among the nineteenth-century progenitors of the social scientific study of European family systems was the Frenchman, Le Play (1895), who combined a series of empirical studies of families across much of Europe with pronouncements about the evolution of family life in the west, warning of an alarming erosion in traditional family forms that was eating at the moral fabric of European society. He contrasted what he took to be the traditional family, consisting of a large, patrilaterally extended group of kin under the senior male's autocratic control, with the unstable family that was rapidly spreading throughout the industrializing population of a rapidly changing Europe.

Between the two extremes was the household form that Le Play thought best, the stem family, which mixed a high degree of respect for authority and concern for other kin with an element of independence and room for innovation. The farmstead lay at the center of stem family life, and was passed on, intact, from generation to generation, to a single child, preventing land fragmentation.

Le Play believed that the large patriarchal family had already receded in Europe, remaining primarily among Eastern nomads, Russian peasants, and the Slavs of Central Europe. The epic struggle that gripped nineteenth-century Europe, in Le Play's view, pitted the stem family against the unstable family.

Such evolutionary perspectives as Le Play's were taken up by American sociologists in the middle decades of the twentieth century. Many assumed that large, complex family households had characterized the European past, and were being ineluctibly replaced by small, nuclear family households as a result of modernization, however defined (Goode 1963). Challenging these assumptions, Laslett (1965), using British historical evidence, argued that the view of the large, complex household as typical of the European past was a 'myth.' According to Laslett, western Europe was long characterized by neolocality (i.e., couples established their own independent households at marriage) and it was rare for people in the past to live in large, complex family households. The Cambridge Group for the History of Population and Social Structure, which Laslett and Wrigley cofounded at this time, became an influential center for historical household studies, and the 1972 volume edited by Laslett and Wall, *Household and Family in Past Time*, set the terms of the debate for years to come and spawned hundreds of local-level studies throughout Europe.

1.2 The European Marriage Pattern

Just as Laslett was attacking the myth of the European extended family, Hajnal (1965) advanced the claim that a unique marriage system had characterized the whole of Europe except for the east and southeast for at least two centuries preceding 1940. It had two key features: a high age of marriage and high proportions of people who never married. A line could be drawn, he argued, stretching from St. Petersburg (then Leningrad) southwest to Trieste. To its east the system was not found.

Two decades later, Hajnal (1983) partially retreated from his by then famous 'line,' portraying the European marriage pattern as characterizing northwestern Europe and excluding not only eastern Europe, but the south as well. In this reformulation, he emphasized a third characteristic of the European family system, in addition to high marriage age (men 26, women 23) and high proportions (10 per cent or more) never marrying: young people in large numbers left their parental home and spent the years before marriage circulating between houses of nonkin in the capacity of servants. It was this circulation of servants that gave flexibility to the nuclear family system, facilitating the redistribution of a household-based rural labor force in response to each household's labor needs.

1.3 Geographic Diversity

As Hajnal's retreat from his own 'line' suggests, the proponents of a traditional European family system soon found themselves under heavy criticism from scholars who determined that some form of a more complex household system had indeed been typical of the areas they were studying. Both Hajnal and Laslett attempted to salvage their model of the European family system by restricting its territorial reach. It is revealing that French family historians never participated in these efforts to find a single European family system, for they were all too aware of the fact that even within a single country—in this case, their own—more than one family system had existed. In northern France a system much like that described by Laslett and Hajnal appeared, but in the central and southern regions larger, complex family households were common in past centuries. Similarly, work on Spain (Reher 1997), Portugal (Kertzer and Brettell 1987), and Italy (Barbagli 1984) found clear internal divisions: in each country nuclear family systems predominated in the south but more complex household systems were found in parts of the north. Nuclear family systems and neolocality were linked to systems

of partible inheritance or contexts in which there was little of productive value to inherit (as in the case of the proletariat).

Laslett (1983) responded to this new evidence by proposing a new division of traditional Europe into four areas—the West, Middle, Mediterranean, and East—each having a distinctive family system. It turned out that the 'myth' that Laslett had punctured, if it was a myth at all, pertained only to one of the four areas of Europe, the West. In this new model, not only did Eastern Europe have low female marriage age and high proportions of joint-family households, but so did Mediterranean Europe.

Laslett continued to hold to Hajnal's notion of a family system in which neolocality, high female age at marriage, and the prevalence of service as a life stage go together. Yet the Italian case, among others, shows that this is not necessarily true: in Italy in past centuries, it was the central and northern areas having high proportions of joint family households that had high proportions of servants, while in the south, where nuclear family residence prevailed, life-cycle service was rare (Da Molin 1990). Moreover, age of female marriage was typically higher in the center and north than in the south, again going against the Hajnal/Laslett model.

2. *Rethinking European Family Systems*

Laslett's later work on European family systems rejected crude generalizations about 'the' European family. A more fully satisfactory approach should build on these insights, recognizing the role played by political economic, demographic, and cultural forces in local contexts. It should also abandon the notion that northwestern Europe is the kernel of Europe and the lingering attempts to rescue a quintessential European family system.

2.1 *Political Economy*

Family historians frequently have pointed to two major, related economic pressures shaping family systems: the household as production unit, and inheritance norms. The traditional model of rural life in preindustrial times depicts a peasant society in which the household was the unit of production. Where this is the case, the composition of the household, and the timing of marriages, are shaped by labor needs and relations between peasants and the land they farm (or ownership of other productive resources, such as livestock). A dramatic case comes from central Italy, where sharecroppers typically lived in large joint family households. Landowners sought to maximize the number of adult laborers on each plot of land, to maximize their income (classically one-half of the produce). In the Alpine area in the north of Italy, likewise, Viazzo (1989) describes a system in which complex family households prevailed, allowing peasants to diversify their holdings and their risks, combining herding with farming scattered plots of land at different altitudes.

Political forces entered into these dynamics both indirectly, via political support for various land-holding systems, and directly, as when systems of household taxation or military conscription influenced peasants to live in more extended or more nuclear households. In the case of serf systems of Eastern Europe, controls on movement out of estates had a similar influence (Czap 1982).

Inheritance systems are tied closely to family systems. In Spain, France, and elsewhere, areas practicing impartible inheritance tended to have stem family systems. On the other hand, where peasants did not own their own land, the link did not operate in this way. Italian sharecropping households practiced partible inheritance, but the pressures to have large numbers of coresiding adults resulted in joint family households.

2.2 *Demographic Constraints*

A continuing theme in debates about European family systems concerns the constraints placed on family dynamics by the brute facts of demography. In 1965, Levy argued that the great majority of people in the past lived in relatively small nuclear family households regardless of family formation norms, because high mortality rates interfered with the formation of extended family households. In critiquing Laslett's original thesis of the prevalence of the nuclear household system in Europe, Berkner (1972) took up this idea, arguing that census data showing that the great majority of households at any one time are nuclear does not in itself provide evidence of a nuclear family system. Rather, he argued, under preindustrial demographic conditions, people following extended family norms would spend most of their lives living in nuclear households.

Given the workings of such demographic constraints, inheritance was typically a much more messy process than a focus on norms would suggest. A study of a nineteenth-century Austrian community, in which impartible inheritance was the rule, for example, found that large numbers of households heads died before they had a son old enough to take over. When their widows remarried, tensions between stepfather and sons resulted (Sieder and Mitterauer 1983). Yet work in sharecropping Italy and on the Russian serf estates shows that such demographic constraints were much less evident in joint family systems: where all sons are supposed to bring their brides into the household at marriage a nuclear phase is much less common, as many other combinations of coresiding kin are possible (Kertzer 1989, 1991, Czap 1982). Indeed, there is

evidence that in the nineteenth century the proportion of people living in large, complex family households was growing in several areas as a result of a combination of demographic and economic factors (Ruggles 1987). In central Italy, with a largely fixed number of sharecropper farms available, a growing rural population, and the primary alternative—agricultural wage labor—providing a much poorer living, there was increasing pressure on the children of sharecropper families not to strike off on their own at marriage, and in this area the nineteenth century saw the historical zenith in the proportion of people living in joint family households (Barbagli 1984).

3. Culture

Understanding European family systems requires not only attention to historically changing political economic and demographic forces, but to cultural norms regarding proper family life as well. For example, each society has norms about the proper timing for a woman's marriage. While in some settings marriage age was affected by changing economic conditions (i.e., marriage being postponed during hard economic times) (Wrigley and Schofield 1981), there is also much evidence of the resistance of cultural norms to economic pressures. One manifestation of this is the number of studies showing a very similar female marriage age across different economic segments of local stratified populations (Knodel 1988, pp. 130–6).

The great bulk of the work on European household systems—following Laslett and Hajnal—focuses on the cultural norm of postmarital residence as the key (identifying neolocality with nuclear family household systems). However, it is becoming clear that another cultural norm must be entered into the model: the fate of older people as they become frail or widowed (Kertzer 1995). Even where newlyweds routinely establish their own household, a pure nuclear family system does not necessarily result. Recent studies show that in the very heartland of Hajnal and Laslett's northwestern European family system, large numbers of older people did live in extended family households. Wall (1995), for example, in a sample of English communities for 1599–1796, found that half of the men aged over 65 and over a third of the women lived with at least one child. While some of these adult children were not married, many of them were.

Finally, while most work on European family systems has focused on household units and on marriage, any satisfactory understanding of the role of family must consider relations that extend beyond single households. Reher (1997, p. 114), for example, found that in the Spanish community of Cuenca in the middle of the nineteenth century, while neolocality prevailed about three-quarters of the newlyweds set up their households in the same immediate vicinity of one of their parental households.

In short, no single family system—in the sense intended by Hajnal or Laslett—characterized 'traditional' Europe, nor even western Europe. While broad general differences among family systems in different parts of Europe can be distinguished, attempts to divide the continent into distinct family zones are misguided. In those areas where complex family households were commonly found, significant sections of the population lived in simpler households, while even in the heart of northwestern Europe, the epicenter of the presumed nuclear family system, more complex household dynamics were often at work.

See also: Families and Households, Behavioral Demography of; Families and Households, Formal Demography of; Family and Kinship, History of; Historical Demography; Western European Studies: Gender and History; Western European Studies: Society

Bibliography

Barbagli M 1984 *Sotto lo stesso tetto*. Il Mulino, Bologna, Italy

Berkner L 1972 The stem family and the developmental cycle of the peasant household. *American Historical Review* **77**: 398–418

Czap P 1982 The perennial multiple family household, Mishino, Russia. *Journal of Family History* **7**: 5–26

Da Molin G 1990 *La famiglia nel passato*. Caccuci Editore, Bari

Goode W J 1963 Industrialization and family change. In: Hoselitz B F, Moore W E (eds.) *Industrialization and Society*: UNESCO, New York

Goody J 1983 *The Development of the Family and Marriage in Europe*. Cambridge University Press, Cambridge, UK

Hajnal J 1965 European marriage patterns in perspective. In: Glass D V, Eversley D (eds.) *Population in History*. Edward Arnold, London

Hajnal J 1983 Two kinds of pre-industrial household formation system. In: Wall R, Robin R, Laslett P (eds.) *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK

Kertzer D I 1989 The joint family household revisited: Demographic constraints and household complexity in the European past. *Journal of Family History* **14**: 1–15

Kertzer D I 1991 Household history and sociological theory. *Annual Review of Sociology* **17**: 155–79

Kertzer D I 1995 Toward a historical demography of aging. In: Kertzer D I, Laslett P (eds.) *Aging in the Past*. University of California Press, Berkeley, CA

Kertzer D I, Brettell C B 1987 Advances in Italian and Iberian family history. *Journal of Family History* **12**: 87–120

Knodel J 1988 *Demographic Behavior in the Past*. Cambridge University Press, Cambridge, UK

Laslett P 1965 *The World We Have Lost*. Methuen, London

Laslett P 1983 Family and household as work group and kin group: Areas of traditional Europe compared. In: Wall R, Robin R, Laslett P (eds.) *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK

Le Play F 1895 *L'Organisation de la famille selon le vrai modèle signalé par l'historie*, 4th edn. Alfred Mame, Paris

Marion L Jr 1965 *Aspects of the Analysis of Family Structure*. Princeton University Press, Princeton, NJ

Reher D S 1997 *Perspectives on the Family in Spain Past and Present*. Clarendon Press, Oxford, UK
Ruggles S 1987 *Prolonged Connections: The Rise of the Extended Family in Nineteenth-century England and America*. University of Wisconsin Press, Madison, WI
Sieder R, Mitterauer M 1983 The reconstruction of the family life course. In: Wall R, Robin J, Laslett P (eds.) *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK
Viazzo P P 1989 *Upland Communities: Environment, Population and Social Structure in the Alps since the Sixteenth Century*. Cambridge University Press, Cambridge, UK
Wall R 1995 Elderly persons and members of their households in England and Wales from preindustrial times to the present. In: Kertzer D, Laslett P (eds.) *Aging in the Past*. University of California Press, Berkeley, CA
Wrigley E A, Schofield R S 1981 *The Population History of England 1541–1871*. Edward Arnold, London

D. I. Kertzer

Family Theory and the Realities of Childbearing Behavior

Childbearing takes place within the socially defined context of family life, but there have been substantial changes in family life in recent decades and it differs between societies and population subgroups. While childbearing patterns have also changed substantially in developed societies since the 1950s, there has been a tendency for trends to be broadly similar (Fig. 1(a)). The annual level of fertility in the post-1945 period rose to a peak around the late 1950 to mid-1960s, followed by a sharp decline in the 1970s and a slight rise and relative stabilization in the recent period, although some countries such as those in Southern Europe have continued to decline to levels of fertility just above one child per woman in the late 1990s (Coleman 1996). The average age at childbearing fell until the late 1960s or early 1980s and then rose substantially (Fig. 1(b)).

Annual levels and mean age at childbearing are closely related since a move towards earlier childbearing will increase the annual fertility rate even if the women having their children in that period do not ultimately have any more children by the end of their reproductive lifetime than other cohorts of women. Most of the large-scale movements in fertility in the postwar period are attributable to shifts in the age at childbearing rather that to changes in the overall propensity to have a given number of children: 'most of the baby boom would have occurred without any change in the numbers of births per woman, and most of the decline since the baby boom as well' (Ryder 1980, pp. 39–40). While the level of childbearing in a given year may reflect contemporary factors such as income or employment, some women may be having children then because they have postponed births from an earlier period or wish to bring their births forward for particular reasons. The fact that married women are much more likely to have a child than those not in a partnership, with cohabiting women generally in an intermediate position, means that earlier experiences of entry into and exit out of partnerships also need to be taken into account.

The main alternative theories that have been put forward to explain childbearing patterns will be reviewed briefly, before considering the extent to which they can be empirically validated.

1. Explanations of Childbearing

1.1 Proximate Determinants of Fertility

Fertility is often usefully analyzed in terms of a proximate determinants framework (Bongaarts and Potter 1983) which for developed countries means concentrating on three main factors; the proportion of women in a sexual union, their use of contraception, and the level of induced abortion. These factors biologically determine childbearing, but women's choices about partnership and family limitation may also be consequences of their fertility decisions. One reason why women may enter a partnership is in order to have children, but it is unlikely to be the only one. On the other hand, use of contraception and abortion on 'social' grounds are determined by fertility preferences. Over time, there has been a change from a period when methods of family limitation were generally ineffective, difficult to obtain, and/or disapproved of, to one when methods are safe, easy to use, relatively inexpensive, reliable, and acceptable. This change occurred with the introduction of the contraceptive pill and IUD around 1960 in many developed countries, and the subsequent introduction of laws permitting abortion on wider grounds than previously. Studies in the US showed that the level of 'unwanted' fertility in the 1960s was about 0.5 of a child per woman (Westoff and Ryder 1977), similar to the difference in cohort fertility between women completing their childbearing at that time and today. The availability and acceptability of efficient contraceptive methods has had a major contribution to the change in timing which is the major determinant of overall period change. The effect of access to improved contraception not only affects childbearing, but permits couples to have a childless relationship (and, in particular, to facilitate the emergence of cohabiting relationships that have low rates of childbearing) and to decide on the pattern of childbearing with greater precision, since, for example, the possibility of an unwanted birth can be eliminated by contraceptive sterilisation. In turn, this has permitted the increase in mean age at childbearing and hence led to changing attitudes about women's roles since a full satisfying

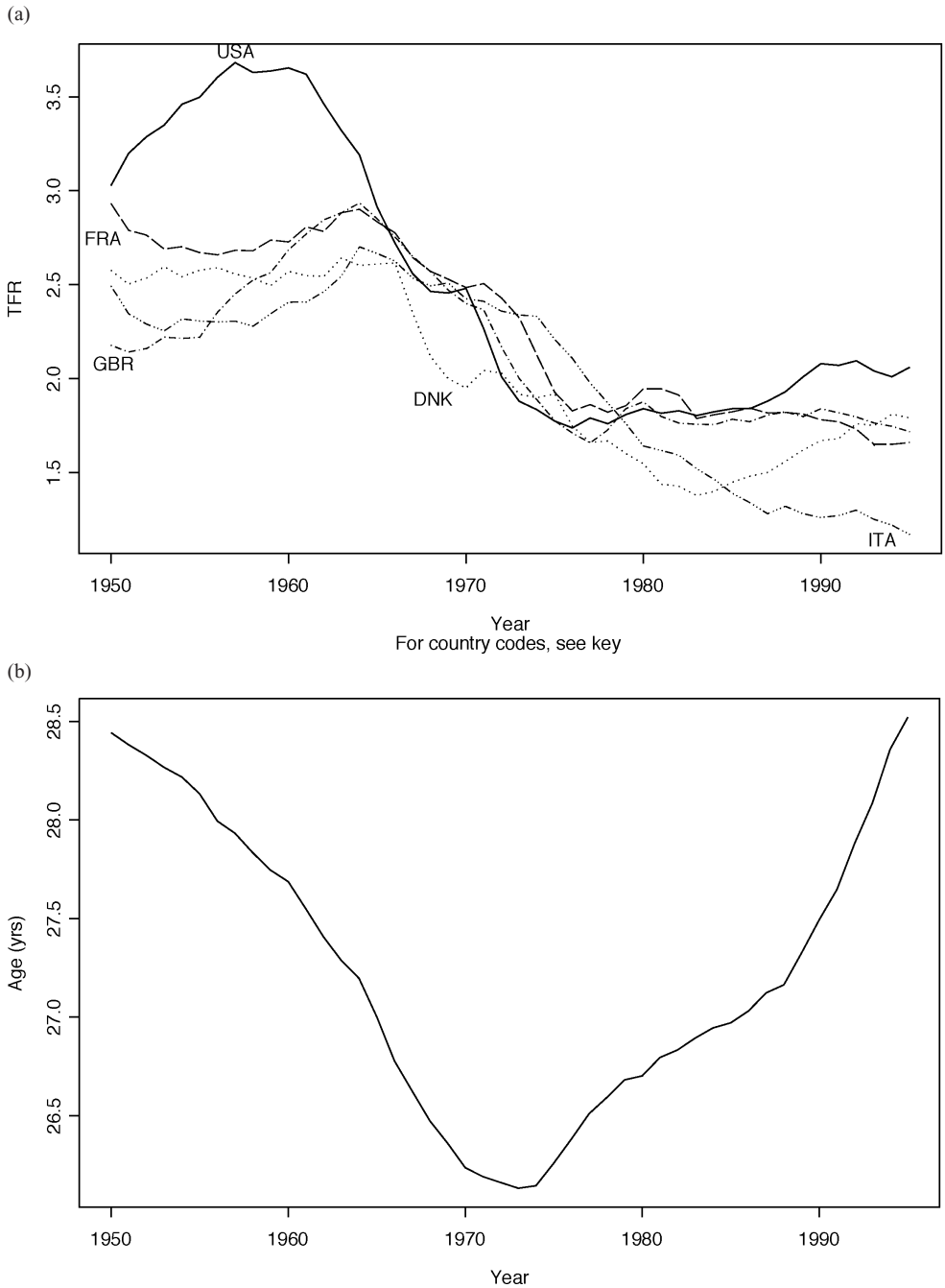

Figure 1
(a) Total fertility rate, selected OECD countries, 1950–95. (b) Mean age at childbearing, England and Wales, 1950–95. Key for charts: AUS, Australia; AUT, Austria; BEL, Belgium; CAN, Canada; DNK, Denmark; FIN, Finland; FRA, France; DEU, Germany; GRC, Greece; ISL, Iceland; IRL, Ireland; ITA, Italy; JPN, Japan; KOR, Rep. of Korea; LUX, Luxembourg; NLD, Netherlands; NZL, New Zealand; NOR, Norway; PRT, Portugal; ESP, Spain; SWE, Sweden; CHE, Switzerland; GBR, England and Wales; USA, United States (Source: World Bank and National Statistical Offices)

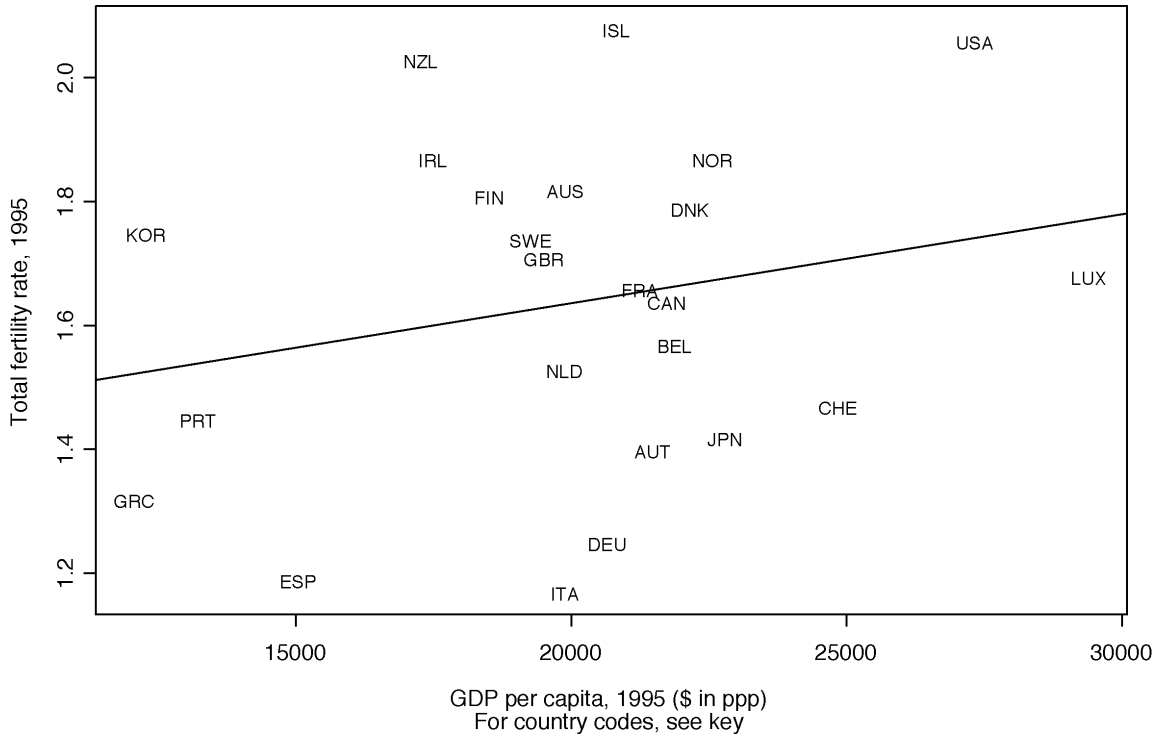

Figure 2
Total fertility rate vs. GDP per capita for a range of countries in 1995

sex life is now compatible with career progression by married women that would have been difficult in the past.

As improved access to methods of family limitation have increased across and within nations, variability in fertility has declined, so that the maximum difference in TFR levels between the OECD member countries in 1995 was 0.9 of a child, compared with a value of 1.9 in 1965 and 1.8 in 1980.

1.2 Economic Approaches

In the past, a positive relationship between fertility and income (in cross-section) and with the business cycle (in time trend) had been accepted. However, in the period since 1945, fertility and real incomes generally have been negatively rather than positively related in developed countries, since fertility is now at an all time low, whereas average real incomes have risen. The current widely used microeconomic model of fertility is the 'new home economics' one particularly associated with the work of Becker (1981) that has been extended to both marriage and marital breakdown.

The model brought together two features within the framework of microeconomic analysis to account for the postwar baby boom and bust: the first was the increasing importance of women's work outside the home and its competition with childrearing for the wife's time. Since women undertake the greater burden of childrearing, the costs of childbearing, especially in terms of earnings foregone while they are looking after children and future labor market prospects, fall most heavily on them. Therefore, the number of children is expected to be strongly negatively related to her earnings, but weakly positively related to her husband's earnings, and the opportunity costs of childbearing are greatest for women in highly paid jobs.

The second element, the demand for children, is determined in part by economic factors and in part by 'tastes' which may differ by religion, race, age, and time period, etc. Decisions concerning childbearing have two major components, those related to quantity and 'quality': 'quality' is simply measured by the amount spent on a child on items such as clothing, education, etc. ('higher quality' does not mean superior, just more expensive). Thus, parents can choose higher 'quality' of a lower quantity of children for a given level of overall expenditure. It is assumed that increased income will be diverted more towards

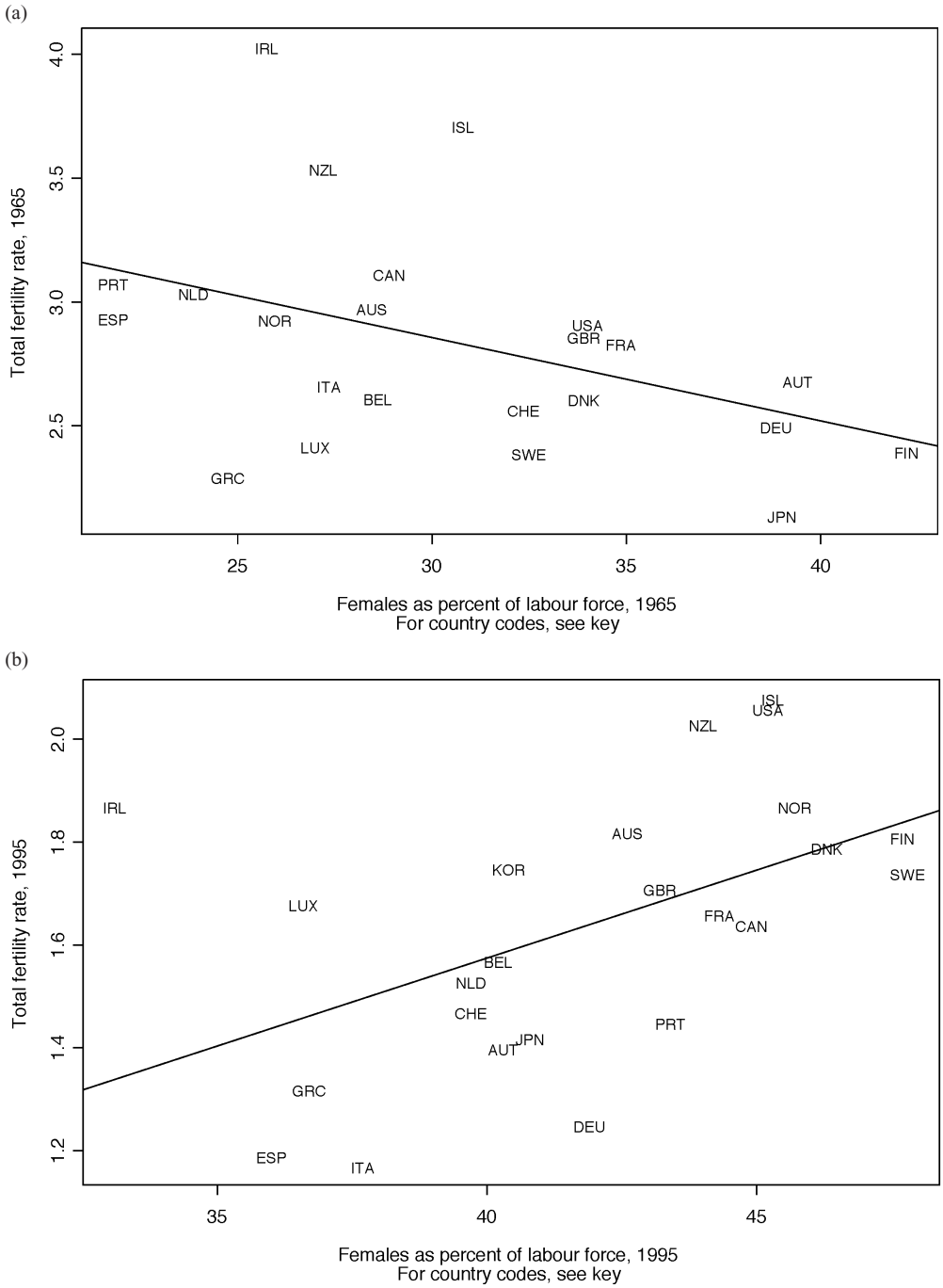

Figure 3
Total fertility rate vs. females as a percentage of the labor force

increases in quality rather than quantity, and it is possible that if the shift towards higher quality is so substantial, increased income may result in fewer children. At present, the cross-national relationship between average income and fertility is positive but weak in developed countries (Fig. 2).

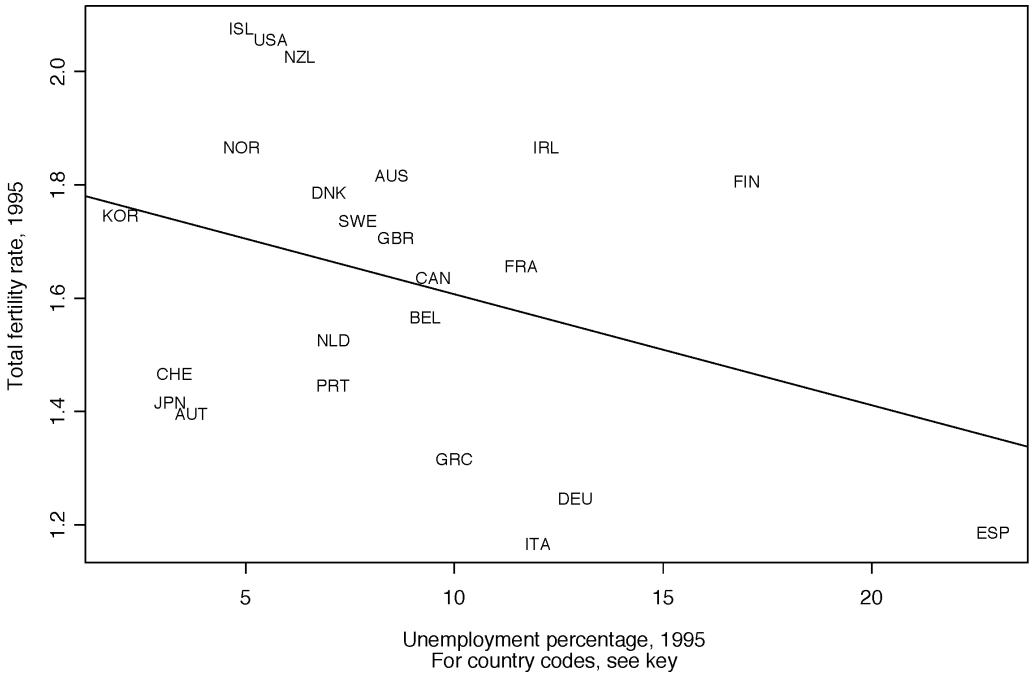

Figure 4
Total fertility rate vs. unemployment percentage, 1995

The relationship between female employment and fertility is expected to be more clearly negative since a basic incompatibility between work outside the home and childrearing is assumed as they are competing demands for the mother's fixed amount of available time. (This approach has been criticized on the grounds that it emphasises the differentiation of gender roles: however, since it reflects how things are, rather than how many think that things ought to be, such criticisms are unfair.) While in the 1960s, there was the expected negative relationship between female employment and fertility across countries, by the 1990s, the relationship had become positive (Fig. 3). In the 1990s, Scandinavian countries have relatively high fertility and female employment, while Mediterranean ones score low on both counts, whereas although the female employment rankings have remained relatively constant since the 1960s, the fertility ones have reversed. Of course, such simple correlations do not establish that the nature of the relationship has switched over time, but it is clear that the other factors that cause the correlation to be in the unexpected direction in the 1990s must also have a very substantial, if not dominant, effect on childbearing.

An alternative economic-demographic model is that of Easterlin (1980). In simplified form, it assumes that individuals' (or couples') expectations are set by the standard of living in the home that they had as children. Those born into large birth cohorts will have many direct competitors for education, employment, housing, etc., and are likely to be relatively disadvantaged compared with those from smaller cohorts and they will react by marrying later and having fewer children. There is little empirical evidence that large cohorts are disadvantaged and earnings are more likely to be associated with other period and structural factors. This model emphasis men's as well as women's employment situations, but because it appeared to have little explanatory power, the role of male employment has tended to be neglected compared to that of women. However, the declining position of men relative to women in the labor market, and, in particular, the rise in unemployment that particularly affected young men is likely to inhibit leaving home and partnership formation, as well as making young men less attractive lifetime partners to young women (Oppenheimer 1994). Such considerations may be especially important among disadvantaged groups such as young black males in developed countries. Overall, there is a negative relationship between unemployment and fertility in the OECD countries (Fig. 4). Although the unemployed are disproportionately drawn from groups that would in any case be expected to start childbearing early and for this reason, among others, may have higher fertility than those in employment, a high level of unemployment is likely to have the effect of inhibiting partnership formation and childbearing of those in employment.

1.3 'Cultural' Explanations

In a study of the determinants of fertility decline in European historical and contemporary Third World countries, Cleland and Wilson (1987) concluded that 'economic' theories are much less successful than 'cultural' explanations. For example, fertility change was more closely associated with factors such as language and voting patterns than with income or employment (Coale and Watkins 1986).

Attitudes to increased sex role egalitarianism, and permissiveness to divorce and cohabitation have become more positive over time, whereas attitudes to religion, traditional family orientation, and deference to authority have declined. These changes are associated with a move towards greater individualism that has been associated with a range of emerging patterns of family formation and dissolution often referred to as a 'Second Demographic Transition,' based on individual autonomy, female emancipation, and increased consumerism (van de Kaa 1987, Lesthaeghe 1995). Such approaches integrate the interaction of different life events such as partnership, childbearing, and employment within the historical context, both national and individual, including cohort effects.

2. Assessing the Empirical Validity of Alternative Explanations

While swings in annual levels of childbearing have attracted much comment, especially the very sharp and largely unanticipated change between the late 1950s and 1970s, from a long-term perspective, it is the baby-boom period that stands out as the anomaly as fertility has shown a generally declining trend for over a century in many developed countries. In the 1950s the main alternative to remaining at home with one's parents was early marriage (sometimes precipitated by pregnancy), followed by a relatively quick start to childbearing, both because this was expected and the possibilities of delaying births were limited. The relatively high level of fertility at that period reflects the selection by women of a package of husband/home/children rather than solitude/parental residence/work. In addition, couples expected that the marriage would last and that employment would be permanent, so increasing the attractiveness of marriage. More recently, people have had the opportunity of childbearing without partnership and vice versa, and other options such as living alone have become more feasible. Therefore, the nature of the choices made with respect to childbearing have changed substantially over time, which complicates the specification of appropriate empirical models.

There are many levels at which analysis and interpretation may be undertaken: international comparisons; subpopulations within nations; families/households; individuals; and, most generally, combinations of data from different levels. Childbearing is clearly sensitive to short-term factors such as pill scares, which may therefore provide good short-term explanations of fertility change (Murphy 1993). However, establishing the reasons for the long-term changes shown in Fig. 1(a) is more difficult. Some of the most widely cited studies for empirical underpinning of the theory have been based on the macrolevel time-series studies of the effects of income on fertility in developed societies. The problem in using aggregate-level data to draw conclusions about individual-level behavior is widely recognized.

A rational choice model underpins most discussion of childbearing, although this is often less helpful than might be thought since the number of factors that enter into the choice are very broad in both time and space. Childbearing is likely to have substantial effects on the emotional, lifestyle, and financial circumstance of the mother for at least two decades, and for most fathers too. Over time, the balance of advantage will have changed considerably. Some of these aspects may be much easier to measure than others. For example, higher social benefits for unmarried mothers which are relatively easy to measure, would be expected to lead to more births outside marriage, but a more important factor could be a reduction in public disapproval for such childbearing, and the availability of benefits a consequence of changing public attitudes, which may be difficult or impossible to quantify especially if no contemporary data exist.

It may be useful to make a distinction between those factors that have an unambiguous effect on childbearing and those for which the effect is likely to be context-specific. For example, it has been assumed that the effect of women's work outside the home is to reduce their fertility, as seen in the 1960s in Fig. 3(b). However, by the 1990s, the reverse was the case in that those countries with high rates of female employment had higher fertility also. The range of reasons for this could include:

(a) formerly it was considered inappropriate for women with young children to be in paid employment, whereas now it is not considered harmful to the child (social influence);

(b) to bring up a child, two incomes are now required whereas one would suffice in the past (reversed female income effect);

(c) improved childcare facilities and technology makes it more feasible to combine childrearing and work than in the past (institutional and technological change);

(d) withdrawal from the labor force for childrearing could have harmful long-term economic and career prospects for women, especially in societies with high rates of marital breakdown (external demographic effect);

(e) societies that have low female employment have high unemployment and poor job security that inhibits family formation (historical macroeconomic factor);

(f) societies that have exhibited substantial fertility declines recently may reflect reduced fertility by older women who have sufficient children and postponement by younger women (transitory effect).

This list is not exhaustive but serves to emphasise the difficulty in establishing empirically the nature of the relationship between childbearing and the single variable of female employment. The statement that if increases in paid work outside them by women of childbearing age makes bringing up children less attractive as an option, with other things being constant, then childbearing will fall is uncontentious. However, these other relevant factors that are not constant may be moving and be the dominant determinants of change. The dependent variable in the new home economics model is not number of children born, but it is the benefits obtained from their children, yet this cannot be clearly specified and so number of children is often used as the variable to be explained. The biological, cultural, institutional, economic, and historical factors that influence childbearing decisions are complex and it is impossible to obtain comparable data on these different domains, so in practice, it is almost impossible to specify a comprehensive formal model to 'explain' contemporary patterns. It is, therefore, unsurprising that after many decades of investigation, although a number of valuable insights have been made about the determinants of childbearing trends and differentials between subpopulations, full explanations remain elusive.

See also: Demographic Transition, Second; Families and Households, Behavioral Demography of; Families and Households, Formal Demography of; Family as Institution; Family Processes; Family Size Preferences; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Childbearing; Fertility Control: Overview; Fertility: Proximate Determinants; Lone Mothers in Affluent Nations

Bibliography

Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Bongaarts J, Potter R G 1983 *Fertility, Biology and Behavior*. Academic Press, New York

Cleland J, Wilson C 1987 Demand theories of the fertility transition: An iconoclastic view. *Population Studies – J. Demog*. **41**: 5–30

Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ

Coleman D (ed.) 1996 *Europe's Population in the 1990s*. Oxford University Press, Oxford, UK

Easterlin R A 1980 *Birth and Fortune: The Impact of Numbers on Personal Welfare*. Basic Books, New York

van de Kaa D J 1987 *Europe's Second Demographic Transition*. *Population Bulletin* **42**: 1

Lesthaeghe R 1995 The second demographic transition in western countries: An interpretation. In: Oppenheimer Mason K, Jensen A-M (eds.) *Gender and Family Change in Developed Societies*. Clarendon Press, Oxford, UK, pp. 17–62

Murphy M 1993 The contraceptive pill and women's employment as factors in fertility change in Britain 1963–1980: A challenge to the conventional view. *Population Studies – J. Demog*. **47**: 221–43

Oppenheimer V K 1994 Women's rising employment and the future of the family in industrial societies. *Population and Development Review* **20**: 293–342

Ryder N B 1980 Components of temporal variations in American fertility. In: Hiorns R W (ed.) *Demographic Patterns in Developed Societies*. Taylor & Francis, London, pp. 15–54

Westoff C F, Ryder N B 1977 *The Contraceptive Revolution*. Princeton University Press, Princeton, NJ

M. Murphy

Family Theory: Competing Perspectives in Social Demography

Two competing perspectives on the effect of postwar changes in men and women's economic behavior on the family have characterized recent demographic research. Each primarily focuses on one of two fundamental theoretical pillars which underlay most demographic research on families in Western societies: (a) the economic implications for family behavior of an independent nuclear family system; and (b) the importance of sex-role specialization in marriage. This article will undertake an examination of each of these perspectives in terms of these theoretical underpinnings, how they are affected by changing economic behavior, and the implications of this for the emerging nature of the family.

1. Overview

A number of major changes in marriage and family behavior have occurred in the West since the early 1970s. Age at marriage and marital instability have risen considerably while fertility has declined substantially (Blossfeld 1995, Mason and Jensen 1995, Oppenheimer 1994). The magnitude of these changes has fostered considerable debate about their determinants and consequences and what these imply for the future of the family. However, in most cases, the debate is not so much about major theoretical differences on the nature of the family but differences in which aspects of two long and commonly accepted underpinnings should be emphasized the most. One is that marriage as an institution, as well as marriages as

functioning social forms, is based on sex-role specialization. Although there are variations in how this specialization argument is conceptualized, the general idea is that the husband's role is largely economic while the wife's lies in the care of the home and in the reproduction and socialization of the children.

The second theoretical underpinning to most discussions is that an independent nuclear family system is a long-standing feature of populations of Western European origin. Such a family system is characterized by husband–wife families which are economically independent and live separately in their own household, although a single aged parent might also have sometimes been included in the past. This is opposed to an extended family system which is usually characterized by joint family households consisting of several nuclear families. In general, historical demographic research has indicated that marriage and fertility behavior have often been quite different in populations with independent nuclear as opposed to joint family systems. In the former case marriage is contingent on the couple's ability to be self-supporting and able to set up a separate household; in the latter, the economic ability of the household as a whole to absorb new members has tended to be more important. Since young people's economic position is often precarious, the requirement of economic independence has historically led to somewhat later marriages and also to periodic fluctuations in marriage timing, depending on variations in economic circumstances. This has often resulted in substantial short- and long-term effects on fertility, especially in periods before the deliberate control of marital fertility became common.

This article focuses on two opposing positions regarding the effect of shifts in the economic behavior of men and women. One emphasizes sex-role specialization, arguing that the rising employment of women has been the major factor leading to the observed changes in marriage and family behavior because it has reduced specialization and hence permitted women to become economically independent of marriage, thereby reducing the gain to marriage. Assuming women's employment is unlikely to decrease, the result will be a major and irreversible transformation of the family in modern industrialized societies.

The opposing position is that the family has not been transformed as much as the women's independence argument maintains nor has women's rising employment had a negative effect on marriage formation, although it undoubtedly mitigates against high fertility. On the other hand, there is some evidence that a deterioration in young *men's* economic position has had an important effect on marriage and family behavior because it has seriously affected the ability of young couples to set up an independent household. If so, then the assumed irreversibility of present trends is open to question, although it is extremely unlikely that a return to postwar patterns of the 1950s to 1970s will occur.

These two alternatives do not exhaust the many possible reasons that have been proposed for changing family behavior, although they cover the major debates related to the effect of changes in economic behavior. Nevertheless, it is often the case that different explanations may be implicitly, if not explicitly, dependent on one or both of these positions. For example it has been argued that modernization and the rise of individualism is the major force behind the recent demographic changes (see *Demographic Transition, Second*). However, neither men nor women can be autonomous individualists if they are still economically dependent on traditional marital and kinship ties.

2. *Independent Nuclear Family System*

Given the assumption of sex-role specialization, demographic research on marriage and family behavior in independent nuclear family systems in preindustrial and early industrial populations has tended to focus primarily on the economic position of men, although there has long been a general recognition that women could help facilitate the formation of a marriage via earnings before marriage. In farming communities, marriage was often considerably delayed as the heirs had to wait for the father to die or retire before they inherited the property and with it the means of marrying. Even where leasing rather than owning land was the rule, as in England, there were often extensive delays until a man could save enough to lease a farm and set up a household. The effect of this pattern on marriage timing—and hence on fertility in historically noncontracepting populations—was often substantial with high proportions not marrying until their thirties and relatively substantial proportions never marrying (see *Family Systems in Europe*; Arensberg and Kimball 1968, Goldstone 1986).

With the growth of wage labor in the cities and on the farms as industrialization progressed, young men could much more easily become economically independent at a relatively young age. This promoted a reduction in the age at marriage, although offsetting factors were the need to save up to establish an independent household and prior obligations to contribute to the parental household, especially if one or both parents had died or were incapacitated in some way. In general, economic development also increased the sensitivity of young men's economic position to business cycles and structural changes in the economy. Recessions or depressions reduced marriage propensities, especially among the young, and recoveries led to sharp rises in rates. Moreover, the structural changes in the economy associated with economic development changed the skill levels and types of skills demanded of workers. This has placed an increasing premium on education and technical training in modern industrial

societies which has, in turn, postponed the full-time entry of young people into adult-level employment and, with it, entry into marriage and the achievement of the degree of economic stability usually considered necessary for starting to raise a family.

2.1 Role of Relative Cohort Size

Postwar analyses of the influence of men's economic position on marriage formation and fertility was initially dominated by the work of Richard Easterlin who hypothesized that fluctuations in relative cohort size had a major impact on young men's labor market position, thereby affecting marriage timing and marital fertility (see *Baby Booms and Baby Busts in the Twentieth Century*; Easterlin 1987). Thus, Easterlin argued that the postwar baby boom, such as the US experienced in the early postwar period, was largely caused by the favorable economic position of the small depression cohorts entering the labor market at that time. This encouraged them to marry earlier and have more children than previous cohorts. However, the relatively large cohorts they produced crowded the labor market, negatively affecting their lifetime earnings. This led to postponed marriage and childbearing, lower completed fertility, and greater marital instability. In addition, more wives would be working to compensate for their husband's poorer economic position. The result was smaller birth cohorts in the next generation—the baby bust cohorts—who would be in a much more favorable relative labor-market position. According to the theory, this, in turn, should lead once again to earlier marriage and childbearing, somewhat higher fertility, increasing marital stability, and the slowing of the growth in married women's employment. In short, the theory posited a self-perpetuating cycle.

The relative cohort size argument ran into empirical difficulties as research in the 1980s showed that postwar marriage and fertility fluctuations were largely *period* rather than *cohort* driven, i.e., fluctuations in demographic rates occurred simultaneously for numerous birth cohorts, regardless of cohort size. Moreover, the theory essentially accepted the premise that marriage is based on sex-role specialization. Women's labor-market behavior was considered to vary primarily in response to men's earnings position rather than in response to women's own employment opportunities, independent career goals, or family cycle variations in economic needs. However, women's employment continued to rise rapidly throughout the postwar period, regardless of variations in cohort size. As a result, the focus of research interest began to shift to the familial consequences of the transformation of women's economic roles. Moreover, since the concern with the economic position of young men had become so closely identified with the relative cohort size argument, interest in men's changing economic position flagged for a number of years while the research focus shifted to analyzing the effect of women's rising employment.

3. Sex-role Specialization

At the same time that relative cohort size and hence male-oriented explanations of changes in marriage and family behavior began to lose favor, the continuing rapid rise of married women's employment has seemed to provide a ready explanation for the ongoing changes in marriage and family behavior. Hence, the theoretical emphasis shifted from a concern with the ability to set up and maintain an economically independent household to the critical importance of sex-role specialization for family formation and cohesion.

While it has only recently taken center stage as an explanation of changes in marriage and family behavior, a long-standing and influential tradition in the social science literature is the importance of differentiated sex roles for a stable marriage system. In the early postwar period this idea was emphasized by Talcott Parsons (1949) who argued that sex-role segregation was a functional necessity for marital stability and even for the viability of society itself. More recently, this same theme has been elaborated somewhat differently in the economic theory of marriage by Gary Becker (see *Family Theory: Economics of Marriage and Divorce*). In an argument very similar to one Durkheim made over a century ago, Becker maintains that the major gain to marriage lies in the mutual dependence of spouses, each of whom specializes in certain functions—the woman in domestic production (and reproduction), the man in market work. Marriage then involves trading the fruits of these different skills. However, in response to economic growth and the rising wages it produces, women's market work also rises. The result is that women become less specialized and, as their earnings approach those of potential mates, more economically independent, leading, in turn, to a decline in the desirability of marrying or of staying married as well as the desirability of high fertility.

Not all scholars necessarily agree with Becker's argument in its entirety. Nevertheless, an economic independence argument of one sort or another has had wide appeal among sociologists as well as economists and is currently one of the major contenders in trying to explain recent marriage and family trends in the United States (Cherlin 1992, Oppenheimer 1997). One reason is that the model seems to have tremendous face validity. Not only does the notion of specialized sex roles fit in with long-held views of how the traditional family historically functioned, but the rapid changes in marital behavior appeared to have followed very closely upon the rapid rise in married women's employment, especially that of young

women. Moreover, the theory has the elegance of simplicity, yet it can apparently explain a wide variety of complex changes and differentials in marriage and family behavior—from delayed marriage to nonmarriage, marital instability, nonmarital cohabitation, female-headed households, declining fertility, and so forth. Nevertheless, the independence hypothesis has also run into a number of serious theoretical and empirical problems.

3.1 Mismatches

One problem the theory has encountered is that there are number of mismatches between the phenomena it claims to explain and the behavioral changes that are actually occurring. First, the independence hypothesis is essentially a theory of *non*marriage for it argues that if, by their own endeavors, women can achieve approximately the same income as a prospective spouse there is not much to gain by marrying and specializing in home production. However, so far, while there has been a considerable increase in *delayed* marriage, as yet there is little evidence of a substantial rise in *non*marriage (Oppenheimer 1994, Blossfeld 1995). For example, in the United States although white women reaching age 20–24 in 1980 were more likely to be single than earlier cohorts at that age, by the time they were 35–39 in 1995 just under 10 percent were still never married, below that for late nineteenth century cohorts who were presumably living in a 'traditional' family system. Hence one reason for the apparent support for the hypothesis is due to a serious confounding of *delayed* marriage with *nonmarriage* in the statistics. Moreover, to date, there has also been little effort to articulate how the independence argument could lead to *delayed* instead of *non*marriage.

A major difficulty in assessing the outcome of recent marriage trends lies in the recency of the trends and the necessity of waiting 20 years or more before one can accurately determine whether there actually will be a significant rise in nonmarriage. Efforts to estimate the future proportions never marrying are often misleading because they typically use the current behavior of older people as a guide. However, during a period of rapid timing shifts, the *current* behavior of somewhat older people, who are the relatively sparse and atypical remains of much earlier marrying cohorts, is often a particularly poor guide to the future behavior of the much larger and less select group of younger people who have been delaying marriage in more recent years.

A second kind of mismatch lies in using the family behavior of the 1950s and 1960s in the US (and somewhat later in Europe) as typifying the traditional family patterns when sex-role specialization prevailed; compared to this, the large subsequent changes in marriage and family behavior certainly appear to represent a major retreat from marriage and the family. However, this was the baby boom era when fertility was unusually high compared to the previous 50 years; age at marriage was unusually low—far lower than at the turn of the century—and, in the US at least, even marital disruption was somewhat less likely than in previous years. In short, this benchmark was highly atypical historically and unrepresentative of the supposedly 'traditional' family. Hence using it as a benchmark has led to a considerable exaggeration of the extensiveness of the subsequent break with historically traditional patterns and of the historical origins of many recent trends.

There is still another type of mismatch characteristic of a theory which posits that the stability of the family is founded on women specializing in home production. The proposed model seems strangely out of tune with the needs and characteristics of modern societies. Much of the specialized home production of women in the past was devoted to bearing and rearing children who never survived to adulthood. For women to be equally occupied in contemporary low-mortality societies would result in large families. However, even moderate family sizes in a low-mortality society lead to rapid population growth. Moreover, couples do not just want to produce children *per se*, they want to produce children like themselves—that is, they are interested in social, not just biological reproduction. But the cost of social reproduction is high in a society where increasingly substantial and lengthy investments in human capital for each child are required. Hence, high fertility is usually not a realistic option for most people in modern industrial societies. With low fertility, however, the female specialist would not be doing anything highly productive for much of her life. In short, specialization in home production appears to be a rather anachronistic institutional role for women in industrial societies.

3.2 Empirical Status of the Independence Hypothesis

Much of the empirical work on the women's independence hypothesis has not supported it. At any given time, empirical analyses show that indicators of women's economic independence have not generally had a negative effect on marriage formation. For example, under an independence argument, better-educated women should be more economically independent of marriage; however, microlevel regression analyses show that, once school enrollment is taken into account, they have a *higher* rather than a *lower* propensity to marry.

The empirical work on marital dissolution is more mixed. Some studies have found evidence of an independence effect, whereas others have not. There are a number of possible reasons for these conflicting findings. A major one is that the measurement of women's economic independence is difficult and inconsistent across studies. Some studies use absolute

income while others use some measure of relative income. While the hypothesis implies that the *relative* earnings of wives should be the important factor and hence should be directly modeled, using the ratio of wives' to husbands' income alone can lead to ambiguous results. The wife's earnings in some families may be higher than in others, reflecting a better labor-market position and, consequently, indicating a greater financial independence of her husband. However, the ratio could be higher because the husband's earnings are low compared to other couples, while the wife herself may also be in a very weak labor-market position.

A second problem that is more likely to occur in analyses of marital dissolution than marital formation is the greater difficulty of determining the appropriate causal direction. Wives who are more extensively employed in one year may be more likely to separate the following year—but not necessarily because they are more economically independent. Instead, their marriage may have already been in trouble, causing them to increase their labor-market involvement in anticipation of a marital breakup—the reverse causal direction than that hypothesized.

3.3 Coping with Risk and Change

Although rarely considered in explications of the theory, historical research on independent nuclear families indicates that extreme sex-role specialization in marriage is essentially a high-risk and inflexible family strategy unless accompanied by supplementary support mechanisms (Oppenheimer 1994). Even with such supports, specialization often entailed considerable individual and social costs. An inherent problem is that the temporary or permanent loss of one specialist in a family can mean that functions vital to the well-being of the complementary specialist and children are not being performed. Husbands/fathers can die or become ill or disabled; they can lose their jobs and have difficulty finding another one; they could desert the family for a variety of reasons or become an alcoholic; and so on. The result is that the family is left without its major source of income. Except for employment-related shifts, there are similar problems involving the wife-mother specialist. In that case, there could be no one to take care of the children or the home. Specialization may be a feasible strategy in a large extended family household where no particular individual is indispensable because of the redundancy in personnel that can characterize such a system. However, for independent nuclear families and their individual members, specialization entails considerable risks.

Extreme sex-role specialization is also not a very flexible way of dealing with the varying needs of nuclear families over their developmental cycle. Since individuals' consumption needs and productive capabilities vary markedly by age, a basic feature of nuclear families is that the ratio of consumers to producers, and hence the family's level of living, can vary substantially over the family's developmental cycle. Hence, specialization involves a potentially serious inflexibility in dealing with changes in both a family's internal composition as well as the stresses posed by its environment.

The large literature on family history in Western societies indicates that a variety of strategies were developed to maintain economic stability over the family's developmental cycle and in the event of the temporary or permanent loss of specializing parents (Goldin 1981, Haines 1979, Rowntree 1922, see Oppenheimer 1994 for a review). In the nineteenth and early twentieth centuries these strategies usually involved utilizing the productive labor of children, daughters as well as sons. The evidence also indicates that utilizing the labor of one's children could exact substantial costs. A sufficient number of children old enough to make an economic difference was generally not available until the middle or later stages of the family cycle. Families who temporarily or permanently lost the contribution of the father early in the family cycle were not greatly helped by such a strategy. If the mother was lost when the children were young, the family might break up with children being parceled out among relatives or even going to orphanages. In general, this economic reliance on one's offspring often led to a pattern of 'life cycle poverty' where periods of poverty and comparative plenty alternated over the life cycle of workers. Another well-known disadvantage to the extensive employment of children to supplement their family's income was that it tended to discourage schooling and hence had a negative effect on children's adult socioeconomic status.

Even aside from the drawbacks of using children's work, the economic advantages of this strategy were eventually bound to decline during industrialization as the adult labor force became more skilled and hence the relative economic contribution unskilled children could make declined. This suggests that other equilibrating mechanisms were likely to develop. One mechanism may be governmental social support systems. At the family level, however, from a historical perspective, the rise in married women's employment might be viewed as a functional substitute for the work of their children, facilitating the more extensive schooling of the next generation and thereby fostering upward intergenerational social mobility. In addition, wives' employment is not limited to the later stages of the family's developmental cycle and wives, being adults, and usually having educational attainments roughly similar to their husbands, can command a much higher wage than unskilled children could. Moreover, recent research indicates that wives' earnings are currently playing an important role in offsetting the less favorable earning position of their husbands (Levy 1998). It seems likely, then, that wives'

employment could be a source of cohesion in marriages; in difficult times, a helpmate may be of much greater value than a complete economic dependent.

Even aside from wives employment during episodic periods of an increased need for additional income, wives who work regularly may also increase the interdependence of spouses. Two-earner families where the couple pool their resources are families whose level of living is based on their *combined* income. Even when the income contribution of each spouse is equal, given economies of scale, neither partner could live as well on his/her own or save and invest as much. Hence, instead of the asymmetric dependence of the wife which is characteristic of complete specialization, there is a symmetric dependence where each spouse depends on the other to maintain living standards. What this may also indicate is that the combined income of the two-earner family has come to form the social standard rather than the husband's income alone. To the extent this is the case, it becomes increasingly difficult for single earners and married couples with a more traditional division of labor to achieve the same level of living as the two-earner family. Hence, the mutual dependence of the two-earner family may not only contribute to their own gain to marriage but may also reduce the relative gain to being single and to marriages characterized by a specialized division of labor.

In sum, it is unclear that specialization is the only, or even the optimum, method of creating cohesion in marital relationships in modern industrial societies. The specialization model itself seems highly problematic and women's rising employment may actually be promoting a more collaborative model of marriage.

4. *Rising Economic Inequality*

Given the empirical and theoretical limitations of the specialization model of marriage, research attention has been returning once again to the role of economic difficulties in setting up and maintaining an independent nuclear family. While the effect of relative cohort size earlier dominated research in this area, the more recent focus is on such issues as changing economic structure, chronically high unemployment rates (particularly among young adults), and rising earnings inequality (again especially among the young but also among the moderately-to-less educated as well). These analyses have also returned to the question of men's changing economic position which had been neglected for quite some time. Research in this area is still in its early stages but some of the major questions to be addressed are: (a) whether, how, and why young men's labor-market position affects cohabitation and marriage and what difference socioeconomic differentials make; (b) and whether economic changes in a society are likely to have had an important impact on young people's economic position and have thereby influenced changes in their marital behavior.

Virtually all recent research has shown that men who are more likely to be working full-time/year-round, had higher current earnings, and/or greater future career prospects were more likely to marry in any given short time interval (Clarkberg 1999, Huinink 1995, Oppenheimer 1997, Oppenheimer et al. 1997). Some research in this area has also started to look at the role of economic factors in the formation of cohabiting unions and this indicates that here too, some level of economic stability is important, although apparently not as much as for marriage formation (Bracher and Santow 1998, Clarkberg 1999). Oppenheimer (1988) has theoretically and empirically elaborated on these issues in a variety of ways. First, she has conceptualized the transition to work as a career-entry *process* rather than a *step*, which was more commonly the approach in previous life-course analyses. This change in approach emphasizes that embarking on a relatively stable occupational career takes time, during which a young man is in a relatively 'immature' career state. One aspect of career immaturity directly impeding the formation of an independent household is relatively low income; however, another is uncertainties regarding what kind of stable employment may eventually be achieved, when this may occur, what lifestyle the individual may ultimately expect or desire, and so forth. Such uncertainties can impede mate selection and affect the individual's own desirability as a prospective marriage partner. Hence, career immaturity is not only likely to foster delayed marriage for economic reasons but for social-psychological ones as well. Second, viewing the transition to work as a *process* expands the emphasis beyond the effect of career immaturity at any single point in time to the length of time it takes to achieve some career stability and the cumulative impact this will have on marriage timing. The delaying effect of only a modest effect of career instability in any given year, for example, may nevertheless have a large cumulative impact if the instability persists over many years. Moreover, this approach is well suited for examining how the length and difficulty of career development could vary substantially among different groups and over time as well, thereby facilitating analyses of socioeconomic differentials in marriage timing and trends in marriage behavior.

Using 12 years of an American longitudinal survey of young people, Oppenheimer and her colleagues (1997) found that not only did a variety of indicators of career immaturity have a substantial effect in any given year on marriage formation, but that the cumulative impact was substantial when the career-development process was followed over time. Moreover, there were large race–schooling differences in the length and difficulty of the career-entry process which, in turn, produced major differences in marriage timing.

Because the analysis was limited to the young adult life course of young men born between 1957 and 1964, how much of the rise in delayed marriage over the past 25 years can be accounted for by career-development problems has yet to be directly examined. However, there has been a substantial and well-documented rise in inequality in the US since the 1970s, and this has had a particularly severe effect on moderately-to-less educated younger workers, suggesting that career-entry difficulties may be playing an important role in these trends. In Western Europe there has been a substantial increase in persistent unemployment, especially among the young which also indicates that career-entry difficulties may be important in changing European marital behavior (Levy 1998).

See also: Families and Households, Behavioral Demography of; Family Theory and the Realities of Childbearing Behavior; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Marriage and Divorce; Family Theory: Role of Changing Values; Marriage and the Dual-career Family: Cultural Concerns; Reproductive Rights in Affluent Nations

Bibliography

Arensberg C M, Kimball S T 1968 *Family and Community in Ireland*. Harvard University Press, Cambridge, MA

Blossfeld H-P (ed.) 1995 *The New Role of Women: Family Formation in Modern Societies*. Westview Press, Boulder, CO

Bracher M, Santow G 1998 Economic independence and union formation in Sweden. *Population Studies – Journal of Demography* **52**: 275–94

Cherlin A J 1992 *Marriage, Divorce, and Remarriage*. Harvard University Press, Cambridge, MA

Clarkberg M 1999 The price of partnering: the role of economic well-being in young adults' first union experiences. *Social Forces* **77**: 945–68

Easterlin R A 1987 *Birth and Fortune*. University of Chicago Press, Chicago

Goldin C 1981 Family strategies and the family economy in the late nineteenth century: the role of secondary workers. In: Hershberg T (ed.) *Work, Space, Family and Group Experience in the Nineteenth Century*. Oxford University Press, New York, pp. 277–310

Goldstone J A 1986 The demographic revolution in England: a re-examination. *Population Studies – Journal of Demography* **49**: 5–33

Haines M 1979 Industrial work and the family life cycle, 1889–1890. *Research in Economic History* **4**: 449–95

Huinink J 1995 Education, work, and family patterns of men: the case of West German. In: Blossfeld H-P (ed.) *The New Role of Women: Family Formation in Modern Societies*. Westview Press, Boulder, CO, pp. 247–62

Levy F 1998 *The New Dollars and Dreams: American Incomes and Economic Change*. Russell Sage, New York

Mason K O, Jensen A-M 1995 *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK

Oppenheimer V K 1988 A theory of marriage timing. *American Journal of Sociology* **94**: 563–91

Oppenheimer V K 1994 Women's rising employment and the future of the family in industrial societies. *Population and Development Review* **20**: 293–342

Oppenheimer V K 1997 Women's employment and the gain to marriage: the specialization and trading model. *Annual Review of Sociology* **23**: 431–53

Oppenheimer V K, Kalmijn M, Lim N 1997 Men's career development and marriage timing during a period of rising inequality. *Demography* **34**: 311–30

Parsons T 1949 The social structure of the family. In: Anshen R (ed.) *The Family: Its Function and Destiny*. Harper, New York, pp. 173–201

Rowntree B S 1922 *Poverty: A Study of Town Life*. Longmans, Green, London

V. K. Oppenheimer

Family Theory: Complementarity of Economic and Social Explanations

1. The Problem

With the growing specialization in the social sciences the various disciplines have often developed their own traditions and paradigms in explaining social phenomena. There is, however, a risk that this may lead to segregated 'communities of belief' that advance only 'positive proof' in support of their own positions. This has, on occasion, led to the presentation of methodologically flawed research that is supposed to enhance the supremacy of one discipline over the other (see Burch 1996).

As far as family and population research is concerned, contributions are made regularly from a wide variety of disciplinary backgrounds. The phenomena that needed to be explained, such as the fertility or mortality transitions or the change in demographic behavior and household structures, have drawn the attention of sociologists, historians, economists, anthropologists, geographers, biologists, epidemiologists, and statisticians alike. As a consequence, the field of population studies—or demography in a broad sense—has always had firm interdisciplinary roots. The leading journals, such as *Population Studies* (UK), *Demography* (US), *Population and Development Review* (US), and *Population* (France) have remained loyal to their interdisciplinary tradition, and they are still firmly in the lead over their more 'mono-disciplinary' competitors.

Yet, there have been skirmishes and frustrations. The former have often been similar to the nature–nurture debate in psychology. But in the field of family and population research, they emerge in the form of an economics vs. culture debate. The demand theory used by economists in explaining declining fertility or differential fertility between social groups is matched by a diffusion theory, advanced by economists and noneconomists alike. In the latter, the roles of social

learning and the worldwide spreading of new ideas and attitudes are being accentuated. Today, there is still a small cluster of neoclassic economists that hold the belief that nothing more is needed on the explanatory side than changing constraints and opportunity structures (see Becker 1991, Stigler and Becker 1977), but at the other extreme, there is also a hard core of diffusionists who claim that historical and contemporary fertility transitions are occurring under such highly contrasting economic conditions that only the cultural diffusion of new ideas—now also followed by family planning programs—can account for them (e.g., Cleland and Wilson 1987). Obviously such discussions have major policy implications, and the debate has had a lasting life (see the contrasting results on the effects of family planning in Schultz 1997 and in Bongaarts et al. 1990).

In contrast, the frustrations among the mainstream participants in family and population studies stem from the observation that the different explanations live side by side as distinct narratives, as separate 'good stories.' Van de Kaa (1996) describes an anchored narrative by the following two characteristics: (a) there must be an easily identifiable central action (such as a detailed statistical account of changing trends or of newly emerging differentials) and (b) there must be a setting or theory that allows for an easy interpretation of that action. A well-anchored narrative consists of subnarratives that can be ordered from the general to the highly detailed. The question is whether the various social science disciplines can achieve a sufficiently high degree of sophistication that lifts them beyond anchored, but still compartmentalized narratives.

A classic recipe for theory integration emerged from the great debate in philosophy of science on the evolution of scientific knowledge as it took place in the late 1960s and much of the 1970s with Karl Popper and Thomas Kuhn as major protagonists (see Lakatos and Musgrave 1970). The gist of Lakatos's (1970) strategy of progressive problem shift is that a scientific theory H_1 can be regarded as falsified only if another theory H_2 has been proposed such that:

(a) The counter-evidence to H_1 is corroborating evidence for H_2;

(b) H_2 satisfactorily explains all the empirical successes hitherto explained by H_1 (i.e., partial inclusion of H_1 into H_2); and

(c) H_2 is also capable of explaining or predicting new facts or facts that were unlikely or impossible according to H_1 (i.e. enlarged content of H_2).

If these three principles are applied to the theories and anchored narratives in the broader field of demography, one comes to the following observations (see Lesthaeghe 1998):

(a) Many theories, or substantial parts of them, are not mutually exclusive but merely complementary: H_1 and H_2 explain different aspects or account for different subnarratives;

(b) Mechanisms described in H_1 are often connected to mechanisms identified by H_2. Theories escape from these connections through *ceteris paribus* clauses with respect to different but nonetheless highly relevant omitted variables. If put together, however, the respective mechanisms in H_1 and H_2 often produce mutually reinforcing (or neutralizing) consequences or lead to important interaction effects; and

(c) Nonoverlapping parts of H_1 and H_2 that produce opposite predictions have rarely been falsified. Rather, their respective predictions have often come true in different contexts. Hence, a substantial degree of historical path dependency or context specificity emerges. This points in the direction of conditionality on factors not specified in either H_1 or H_2 but to be identified in a broader theory H_3.

These points will now be illustrated with respect to the changes in family formation and dissolution in industrialized countries since the 1950s.

2. *A Comparison of Three Theories of the 'Second Demographic Transition'*

The term 'second demographic transition' was introduced by Lesthaeghe and van de Kaa (1986) to describe the changes in family formation, union dissolution, and patterns of family reconstruction in Western nations during the latter part of the twentieth century. By now, these patterns are also emerging in the former Eastern bloc countries of Europe including Russia. The changes in family formation operate through the postponement of marriage, a rise in single living, prolonged residence in the parental household, increases in incidence and duration of cohabitation (both pre- and postmarital), a postponement of parenthood leading to overall subreplacement fertility, and an increase in procreation within consensual unions. Patterns of union dissolution are characterized by rising divorce rates occuring at shorter durations of marriage and by high separation rates among cohabitants. These phenomena lead to more single person households and to more female-headed single parent households. More children grow up in nonconventional families. Although countries differ with respect to incidence and timing (e.g. Lesthaeghe 1995), most changes have consistent international trends.

These features of the narrative have been explained in three ways:

(a) by the theory of increased female economic autonomy combined with other ingredients of the neoclassic demand theory in economics (Becker et al. 1977, Becker 1991, Hotz et al. 1997, Bergstrom 1997);

(b) by the theory of relative economic deprivation (Easterlin 1976, Easterlin et al. 1990);

(c) by the theory of ideational shift (Preston 1986, Lesthaeghe and Meekers 1986, Thornton and Camburn 1987, Lesthaeghe and Surkyn 1988, Bumpass 1990).

The neoclassic economic theory of the family posits that (a) increasing investment in the quality (e.g., education) of children rather than in their numbers is being made as individuals and societies become wealthier, and (b) that increased female education and the opening up of employment opportunities for women has led to increased female economic autonomy and to increased opportunity costs associated with household tasks and child rearing. Services provided by the market are substituted for such tasks. Economic autonomy also means that women are less dependent on husbands, so that entry costs into marriage have become higher and exit costs lower. From these core propositions one can easily derive the changing position of men in households (sharing of tasks and child rearing activities, the quest for gender symmetry, weakening of male authority, etc.). By the same token, the rules of the game are being altered, especially with respect to intrafamilial exchange: less trust from family members in the family head is exchanged for less altruism from the family head toward the members. All of this leads to an economic disinstitutionalization of the family, and the narrative of the 'second demographic transition' describes the demographic outcomes produced by these forces.

The theory of relative economic deprivation provides a number of other key ingredients. The outcomes of this narrative essentially are produced by the tension between the consumption aspirations and employment opportunities of successive male generations. Easterlin originally used the theory to account for the baby boom of the 1960s and the baby bust of the 1970s and thereafter. The generations that produced the marriage and baby boom were raised in periods of less economic growth and had not developed high consumption aspirations during their childhood. These cohorts, however, benefited from expanding economic opportunities during the 1960s, and as a result they could more easily satisfy both consumption aspirations and family formation desires. Subsequent cohorts experienced high levels of consumption in their families of origin and developed high consumption aspirations themselves, but could not as easily fulfill these as a result of less advantageous labor market conditions. All the crucial steps in family formation were, therefore, postponed and new arrangements with a more temporary character, such as prolonged residence in the parental household or cohabitation, emerged as a necessity while waiting for better times. Furthermore, high consumption aspirations in tandem with scarcer opportunities for males fostered the need for an additional partial or full income to be provided by the female partner. From there onwards, the story could easily be completed according to the lines provided by the 'female economic autonomy' version of the neoclassic school.

Yet, both the starting point and the mechanism of the narratives are different. Becker et al. (1977) start from the female side, whereas Easterlin et al. (1990) start from the male point of view. Moreover, the Easterlin version has the built-in mechanism of cohort succession and a link to values (concerning consumption) fostered during childhood and adolescence. In the original version, a demographic factor was also given greater prominence: cohort size would acerbate the consumption aspirations–income tension since larger cohorts would face more competition and hence also tougher employment conditions. Higher female education and a larger female labor force supply (see the neoclassic version) would of course result in similar effects.

From the above narratives it is clear that the two economic theories H_1 and H_2 are not mutually exclusive and that their respective mechanisms have, to put it metaphorically, several cog wheels that grip into each others.

The theory of ideational shifts links the features of the second demographic transition to long-term trends, often starting in the nineteenth century, toward greater individual autonomy in the ethical, religious, and political domains. Typical expressions of this trend are the steady rise of secularism in the West over the last two centuries, the growth of emancipation movements operating first in the domain of social stratification, and then in the area of gender relations, the steady rise of postmaterialist political aspirations (see Inglehart 1990) in each successive birth cohort since 1920, the historical alteration of priorities in child socialization values (see Alwin 1990); or more recently the growth of scepticism toward institutions and the weakening of civil morality. In this respect, the baby boom in the original six member countries of the EU corresponded with the arrival at childbearing age of a cohort of parents that had hardly progressed on the postmaterialism scale with respect to political aspirations, whereas the baby bust corresponds with the childbearing of the cohort that made the largest leap of all on the same scale (Lesthaeghe and Surkyn 1988). Very much the same result emerges with respect to the relative pace of secularization in Europe, and also among American youths, a significant shift away from community orientations and toward a self-centered preoccupation with material riches was observed (Easterlin and Crimmins 1991). On the whole, the rise of secular individualism and the concommittantly diminishing acceptability of institutional regulation in the sphere of the family are part of the same trend. This also implies that individuals are allowed more freedom to opt out of any arrangement if returns are judged to be inadequate and hence, that the cost–benefit calculus, which features so prominently in the neoclassic economic version, has gained legitimacy in most spheres of life.

A distinct feature of the ideational shifts theory is its preoccupation with intergenerational values transmission, the predictive power of value orientations on subsequent choices with respect to patterns of family formation and dissolution (selection effects), and the

feed-back effects in the form of value adjustments (affirmation or negation) in function of certain life cycle events. This line of research has made extensive use of panel data in assessing the strength of these effects net of other predictors commonly used in microeconomic empirical analyses (see Thornton and Camburn 1987, Axinn and Thornton 1993, Clarkberg et al. 1995, Barber 1999, Moors 1999). These meticulously designed studies all show that value orientations, even when measured at young ages (including those related to consumption aspirations and labor force commitment), have highly significant independent effects on future life course choices such as cohabitation vs. marriage, parenthood within and outside wedlock, timing and quantum of fertility, separation and divorce. While not refuting predictions derived from economic theory, these studies document the extra predictive power of ideational dimensions.

Finally, it is clear that several ingredients of the ideational theory can be linked directly to those featured by the economic views. For instance, the neoclassic central focus on female economic autonomy leads directly to the changes in gender relations, and these indeed matter in the unfolding of alternative forms of family formation. In addition, both the Easterlin version and the ideational theory pay ample attention to the role of socialization in successive generations and both are essentially cohort theories of social change. Ideational theory, however, traces the trends in causal factors much further back in time, and it does not of course restrict itself to consumption aspirations alone. But in doing so, the ideational shifts theory provides support for the neoclassic economic position by showing how increased legitimacy of individual economic rationality is taking the lead over the forces of institutional normative regulation.

3. *Predictions and Contexts: The Example of Cohabitation*

The three theories just discussed offer different mechanisms for explaining the rise of premarital cohabitation in the West since the 1960s. According to the neoclassic theory, the phenomenon results from increased economic independence of women, and expects, therefore, that better educated women with better employment opportunities and higher incomes take the lead. In the relative economic deprivation version, cohabitation is a preliminary stage typically emerging prior to the full establishment of economic self-sufficiency of a new household. Cohabitation is, therefore, more typical for lower income groups with less employment stability, or for subgroups that have not yet matched their incomes to their consumption aspirations. In the ideational version, cohabitation is not so much predicted on the basis of economic constraints or opportunities, but it stems from a set of value orientations already emerging during the 'formative years.' As these values are often linked to more advanced education, a positive correlation between education and cohabitation is expected. Furthermore, antecedents in the family of origin (such as parental divorce) also contribute to earlier home leaving and cohabitation (see, for example, Cherlin et al. 1995, Kiernan and Hobcraft 1997), *inter alia*, via the accentuation of values of ethical autonomy.

The neoclassic economic theory and the ideational theory both draw support from the initial positive correlation between education and cohabitation in a number of countries. At the microlevel, attitudes associated with career commitment and gender symmetry, also tend to be positively related to premarital cohabitation. However, in several contexts (e.g., Scandinavia, Austria), historical patterns of premarital cohabitation had survived in specific subgroups, and the diffusion of premarital cohabitation did not originate from the college educated or in large urban areas, but rather in the working class and more remote areas (e.g., northern Sweden, Alpine valleys). In addition, in most other Western countries, younger generations from working-class backgrounds and with less education adopted the pattern of cohabitation very quickly, so that the initial positive correlation with education vanished or even reversed. The predictive power of specific value orientations is also weakening as cohabitation has ceased to be an expression of protest among the highly unconventional (for these trends see Villeneuve-Gokalp 1990, Oppenheimer and Lew 1995, Boulanger et al. 1997, Bracher and Santow 1998, Kravdal 1999).

To sum up, cohabitation has become widespread among very different socioeconomic strata, suggesting (a) the existence of particular diffusion patterns in various national contexts, and (b) a rapid differentiation of motivations behind the choice. This form of family formation can be initiated for reasons suggested by all three theories. As a result, a particular theory may have higher or lower predictive power depending on the phase of the process of diffusion and on national or historical contexts.

4. *Conclusion*

The scientific traditions in the different social science disciplines leave an important imprint on both the theoretical conceptualization and the research agenda. Neoclassical economists have a strongly deductive approach and typically start with theory construction and the specification of the corresponding equations. Often they stop short of adequate operationalizations of basic concepts or fail to incorporate noneconomic variables in tandem with economic proxies (Robinson 1997). Demographers, sociologists, and statisticians use both deductive and inductive strategies, and they are engaged continuously in concept operationalizations and statistical testing or exploration. The latter

group is, therefore, more likely to report 'mixed results' than the former. Empirical scrutinity of associations derived from theory invariably points in the direction of theoretical incompleteness. As shown by the example above, the separate narratives are seldom mutually exclusive, they complement each other in more than one way, and they produce predictions that are restricted to context, time, or social group. All three theories are, therefore, candidates for inclusion into a larger theoretical framework.

Finally, there are two particularly fruitful ways of advancing empirical research. First, the use of panel data, with measurements of both socioeconomic and attitudinal variables at each wave, is ideally suited to check predictions deduced from various theories. And second, much can be learned from comparisons across contexts if attempts are made to specify and to test the presumed contextual causes at work.

See also: Demographic Transition, Second; Families and Households, Behavioral Demography of; Family Size Preferences; Family Systems and the Preferred Sex of Children; Family Theory and the Realities of Childbearing Behavior; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Family Theory: Role of Changing Values

Bibliography

Alwin D 1990 Historical changes in parental orientations to children. *Sociological Studies of Child Development* **3**: 65–86

Axinn W G, Thornton A 1993 Mothers, children, and cohabitation: The intergenerational effects of attitudes and behavior. *American Sociological Review* **58**: 233–46

Barber J S 1999 Ideational influences on the transition to parenthood: Attitudes toward childbearing and competing alternatives. Paper presented at the Annual Meetings of the Population Association of America, New York, March 25–27

Becker G S 1991 *A Treatise on the Family* (enlarged edn.). Harvard University Press, Cambridge, MA

Becker G S, Landes E M, Michael R T 1977 An economic analysis of marital instability. *Journal of Political Economy* **85**(6): 1141–87

Bergstrom T C 1997 A survey of theories of the family. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 21–76

Berrington A, Diamond I 1999 Marital dissolution among the 1958 British birth cohort: The role of cohabitation. *Population Studies* **53**(1): 19–38

Bongaarts J, Mauldin W P, Phillips J F 1990 The demographic impact of family planning programs. *Studies in Family Planning* **21**: 299–310

Boulanger P-M, Deboosere P, Lambert A, Lesthaeghe R, Surkyn J 1997 *Ménages et Familles*, Census Monograph 4. National Institute of Statistics, Brussels

Bracher M, Santow G 1998 Economic independence and union formation in Sweden. *Population Studies* **52**(3): 275–94

Bumpass L L 1990 What's happening to the family? Interactions between demographic and institutional changes. *Demography* **27**(4): 483–98

Burch T K 1996 Icons, strawmen, and precision: Reflections on demographic theories of fertility decline. *The Sociological Quarterly* **37**(1): 59–81

Cherlin A J, Kiernan K E, Chase-Lansdale P L 1995 Parental divorce in childhood and demographic outcomes in younger adulthood. *Demography* **32**(3): 299–318

Clarkberg M, Stolzenberg R M, Waite L J 1995 Attitudes, values, and entry into cohabitational versus marital unions. *Social Forces* **74**: 609–32

Cleland J, Wilson C 1987 Demand theories of the fertility transition: An iconoclastic view. *Population Studies* **41**(1): 5–30

Easterlin R A 1976 The conflict between aspirations and resources. *Population and Development Review* **2**(3): 417–25

Easterlin R A 1980 *Birth and Fortune*. Basic Books, New York

Easterlin R, Crimmins E 1991 Private materialism, personal self-fulfillment, family life, and public interest: The nature, effects and causes of recent changes in the values of American youth. *Pública Opinion Quarterly* **55**: 499–533

Easterlin R A, Macdonald C, Macunovich D J 1990 How have American baby boomers fared? Earnings and economic well-being of young adults 1964–1987. *Journal of Population Economics* **3**(4): 277–90

England P, Farkas G 1986 *Households, Employment, and Gender: A Social, Economic, and Demographic View*. Aldine De Gruyter, New York

Folbre N 1996 *The Economics of the Family*. Edward Elgar, Cheltenham, UK

Hotz V J, Klerman J A, Willis R J 1997 The economics of fertility in developed countries. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 276–347

Inglehart R 1990 *Culture Shift in Advanced Industrial Society*. Princeton University Press, Princeton, NJ

van de Kaa D J 1996 Anchored narratives: The story and findings of half a century of research in the determinants of fertility. *Population Studies* **50**: 389–432

Kiernan K E, Hobcraft J 1997 Parental divorce during childhood: Age at first intercourse, partnership, and parenthood. *Population Studies* **51**(1): 41–56

Kravdal O 1999 Does marriage require a stronger economic underpinning than informal cohabitation? *Population Studies* **53**(1): 63–80

Lakatos I 1970 Falsification and the methodology of scientific research programmes. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK, pp. 91–196

Lakatos I, Musgrave A (eds.) 1970 *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK

Lesthaeghe R 1995 The second demographic transition—An interpretation. In: Mason K O, Jensen A-M (eds.) *Gender and Family change in Industrialized Countries*. Clarendon Press, Oxford, UK, pp. 17–62

Lesthaeghe R 1998 On theory development: Applications to the study of family formation. *Population and Development Review* **24**(1): 1–14

Lesthaeghe R, van de Kaa D J 1986 Twee demografische transities. In: Lesthaeghe R, van de Kaa D J (eds.) *Bevolking—Groei en Krimp*. Van Loghum-Slaterus, Deventer, pp. 19–68

Lesthaeghe R, Meekers D 1986 Value changes and the dimensions of familism in the European Community. *European Journal of Population* **2**: 225–68

Lesthaeghe R, Surkyn J 1988 Cultural dynamics and economic theories of fertility change. *Population and Development Review* **14**(1): 1–45
Moors G 1999 Gender role attitudes, values, and family formation. Paper presented at the Annual Meetings of the Population Association of America, New York, March 25–27
Oppenheimer V K 1988 A theory of marriage timing. *American Journal of Sociology* **94**: 563–91
Oppenheimer V K, Lew V 1995 American marriage formation in the 1980s. How important was women's economic independence? In: Mason K O, Jensen A.-M (eds.) *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK, pp. 105–38
Pollak R A, Watkins S C 1993 Cultural and economic approaches to fertility: Proper marriage or mésalliance? *Population and Development Review* **19**(3): 467–96
Preston S H 1986 Changing values and falling birth rates. *Population and Development Review* **12**(suppl.): 176–95
Robinson W C 1997 The economic theory of fertility over three decades. *Population Studies* **51**(1): 63–74
Schultz T P 1997 Demand for children in low-income countries. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 349–430
Stigler J, Becker G S 1977 De gustibus non est disputandum. *American Economic Review* **67**(2): 76–90
Thornton A, Camburn D 1987 The influence of the family on premarital sexual attitudes and behavior. *Demography* **24**(3): 323–40
Villeneuve-Gokalp C 1990 [From marriage to cohabitation—a recent history of changes in sensual unions.] Du marriage aux unions sans papiers: Histoire récente des transformations conjugales. *Population* **45**(2): 265–98

R. Lesthaeghe

Family Theory: Economics of Childbearing

Economic theories of childbearing relate decisions about family size and the timing of births over people's lives to elements of the constraints on choice that people face, particularly prices and incomes. This article presents the basic concepts and approaches used in these theories, discusses their main lines of argument and summarizes some of their empirical implications. It first examines 'static' theories of completed family size and then discusses dynamic theories of the timing of childbearing. While mainly concerned with theories of marital fertility, it concludes with a discussion of nonmarital fertility.

1. *'Child Quality' and Fertility*

Economic theories of fertility have been around for over 200 years. The ones advanced by Thomas Malthus are the best known. In a few words, Malthus' view was that fertility was determined by the age at marriage and frequency of coition during marriage. He argued that an increase in people's income would encourage them to marry earlier and have sexual intercourse more often. Modern economic theories of fertility date from a paper by Becker (1960), which generalized and developed the Malthusian theory. Until Sect. 4, it is assumed that fertility decisions are made in a stable marriage.

Becker's point of departure is that children are usually a source of psychic satisfaction for parents. An important aspect of his seminal contribution was to point out that this satisfaction is likely to depend on the amount that parents spend on children as well as the number of children that they have. For instance, parents are likely to receive higher satisfaction if their children obtain a university degree than if they leave school as teenagers. For short, he calls children who have more spent on them 'higher quality' children, hastening to add 'that "higher quality" does not mean morally better' (Becker 1960, p. 211). The basic idea is that if parents voluntarily spend more on a child, it is because they obtain additional satisfaction from the additional expenditure. It is this additional satisfaction that is called 'higher quality.'

The concept of 'child quality' continues to be an important element in the economic analysis of fertility today. The characterization of 'child quality' has moved in the direction of identifying child quality with the lifetime well-being of the child, which can be increased by investing more in the child's 'human capital' (i.e., his or her ability to earn income as an adult) or by the direct transfer of wealth to the child. That is, the parents receive more satisfaction from having children who are better off throughout their lives, and they make monetary transfers and human capital investments to influence their child's lifetime standard of living. Thus, child quality could be thought of as the child's future 'quality of life.'

Having incorporated the quality dimension of reproduction decisions, the well-developed model of consumer choice is applied to them. Parents are assumed to choose the best combination of the number of children, their quality and the parents' own standard of living subject to parents' lifetime income and the prices that they face in markets. An increase in parents' income is expected to increase both the quantity and quality of children. In light of the tendency for people to increase the quality of most consumer durable goods proportionately much more than the quantity of such goods when their income increases, Becker argued that this is also likely to be the case for children. In economics, the 'income elasticity' of a good is defined as the percentage increase in the demand for that good in response to a 1 percent increase in income. Thus, using this terminology, Becker's contention is that the income elasticity of the number of children ('quantity') is small compared to the income elasticity of child quality. Put somewhat differently, an increase in parents' income

may increase the amount spent on children substantially, but this would mainly take the form of higher quality rather than more children. The introduction of the child quality concept generalizes the Malthusian theory, which only allows responses to income along the quantity dimension.

1.1 Quantity–Quality Interaction

A new twist to the quantity–quality model was added in the influential papers by Willis (1973) and Becker and Lewis (1973). They assumed that parents view child quantity and quality as substitutes and that parents treat all their children equally, in the sense that child quality is the same for each child. The parents' lifetime budget constraint is that the sum of expenditures on their own consumption, expenditures on their children's human capital acquisition, and monetary transfers to their children equals the parents' lifetime income. The latter two elements of spending represent expenditure on child quality, which is the product of the number of children, quality per child and the full cost per unit of child quality (which is discussed later). This model can be generalized to incorporate costs of the number of children that do not depend on child quality and costs of quality that do not depend on the number of children. As an example of the former, a decline in the cost of averting births, say because of the introduction of the oral contraceptive pill, would increase the marginal cost of a birth without affecting the marginal cost of child quality.

This product in the budget constraint implies that the cost (or 'shadow price') of an additional child is proportional to the level of child quality, and the cost ('shadow price') of raising child quality is proportional to the number of children the parents have. As a consequence, there is an important interaction between family size and child quality. Suppose, for example, that there is a *small* exogenous decrease in fertility, say because of lower contraception costs. This lowers the shadow price of child quality, which in turn raises child quality, which raises the shadow price of children, which lowers family size further, and so on. Thus, a cumulative process favoring child quality and reducing fertility ensues. Similarly, a *small* exogenous increase in desired child quality could produce a *large* decrease in family size and a much larger increase in child quality.

This interaction between child quality and fertility does not derive from assuming that child quality and quantity are *close* substitutes for one another. Indeed, just the opposite must be the case, because if both fertility and child quality are to be nonzero, then they cannot be close substitutes. Thus, family size can be highly responsive to changes in prices and incomes, even though children have no close substitutes.

Another consequence of this interaction is that it may appear that the income elasticity of fertility is negative, even though children are 'normal goods,' in the sense that parents want more of them when parental income increases. The reason is that the true income elasticity is defined with prices constant, but, because of the interaction, it is not possible to hold the shadow prices of an additional child and of child quality constant when measuring the elasticity. The measured income elasticity of family size is more likely to be negative (even though the true elasticity is positive), the more responsive is child quality to changes in income relative to the responsiveness of fertility (i.e., the higher is the quality elasticity relative to the quantity elasticity), and the better substitute is family size for child quality or parental consumption.

1.2 Child Mortality and Fertility

The ultimate manifestation of low child quality is a child not surviving to adulthood. In light of the 'demographic transition' (i.e., the change from a high fertility–high child mortality environment to a low fertility–low mortality one), an interesting question is how fertility responds to changes in the 'risk' of child mortality. Because parents may know the risk, but not exactly how many of their children will die for any given number of births, this is a problem of choice under uncertainty. Sah (1991) assumed that parents choose the number of births to maximize their 'expected utility.' This is the weighted sum of the utilities of having particular numbers of surviving children (inclusive of all benefits and costs), with the weights being the probabilities of obtaining that number of surviving children from given numbers of births.

Suppose first that only surviving children bring costs and benefits. Sah (1991) showed that the number of births cannot increase with an increase in the probability that the child survives to adulthood. While the parents may wish to have more births when survival chances are low, because some may die before adulthood and they will not have enough time to replace them in their fecund period, their incentive to practice such 'hoarding' is less when survival chances are better. Thus, their fertility will not be higher when child survival is more likely.

A cost may, however, be incurred by a birth regardless of whether or not the child survives. In this case, a higher probability of child survival reduces the effective price of a surviving birth, thereby encouraging higher fertility. Sah finds that this effect will be smaller than the 'hoarding effect' above if parents are sufficiently 'risk averse,' or there is a 'target fertility level,' in the sense that the marginal utility of the last child is nonpositive if all the children from an optimally chosen number of births were to survive. If either of these conditions hold, better survival chances for children tend to reduce fertility.

In the analysis by Cigno (1998), parents can influence the chances that their child survives (an element

of 'child quality') by spending more on him/her. In his model, an exogenous reduction in child mortality may either raise or lower fertility. When the level of child mortality is high, reductions in it are likely to raise both fertility and survival-enhancing expenditures on children, because it lowers the price of a surviving child. When, however, the level of child mortality is already low, further reductions in it are likely to reduce both fertility and survival-enhancing expenditures on children.

2. *The Cost of Children*

Consumer theory tells us that, like any other price, a higher full cost of child quality per unit ought to reduce both fertility and child quality. Elucidation of the factors affecting this cost was another important contribution to the literature on the economics of fertility. In particular, it acknowledged the key role of parental time, especially the mother's, in the rearing of and investment in children. The foundations for this analysis were laid in a paper by Mincer (1963) and in Becker's (1965) theory of time allocation, and Willis (1973) built on them in his analysis.

The concept of 'household production' is important in understanding what influences the cost of child quality. Following Becker (1965), the parents' standard of living, the number of children and quality per child yield satisfaction to the parents, but these cannot be purchased directly in the market. Rather, the parents' standard of living and child quality are both produced within the household by combining parents' own time and purchased goods and services. It is the costs of these inputs that determine the cost of child quality relative to the cost of the parents' living standard. Parental time in the production of child quality is primarily the mother's time, and the rearing of children is assumed to be *time-intensive* relative to other home production activities. That is, the proportion of the total cost of producing children represented by the value of the mother's time input to children is larger than the proportion of the total cost of producing the parents' standard of living represented by the value of her time input to that production. Thus, the unit cost of child quality relative to the cost of the parents' living standard is directly related to the mother's cost of time. If she has ever been in paid employment, her cost of time is the wage she could earn in employment (i.e., her forgone earnings). The higher her wage, the higher the cost of an additional child and of additional quality per child.

This analysis suggests, therefore, that men's and women's wages should have different effects on fertility. Higher wages mean higher income for the parents, encouraging them to have more children and to invest more in the human capital of each child or to make larger monetary transfers to them (i.e., higher quality). But higher wages also mean more income lost by those caring for children and investing in their human capital. Because most child-rearing is done by the mother, higher women's wages raise the opportunity cost of a child as well as increasing family income, while higher men's wages mainly affect childbearing through their effect on the couple's income. If the opportunity cost effect of women's wages is larger than their effect on fertility through higher family income, then higher women's wages reduce family size and child quality. It then follows that an increase in the ratio of women's to men's wages would reduce fertility.

In this model, the fraction of the woman's time supplied to the labor market is also a choice variable, which depends on the wage that the woman can earn and her partner's wage. Women who can earn high wages, or whose partners have low earnings, tend to devote a high proportion of their time to the labor market and remain childless. At the other extreme, women who could only earn low wages, or who have partners whose earnings are high, do not work in the labor market at all and have large families. For this latter group, women's wages have no impact on their fertility and higher men's wages increase the opportunity cost of children, because home production time is scarce for these women (they cannot increase it by reducing time in paid employment) and the production of child quality is assumed to be time-intensive relative to other household production activities. Thus, higher husband's wages have a smaller effect on fertility among this group of couples than among couples in which the woman spends some time in paid employment.

In between these two extremes are the bulk of the population of women, who combine childbearing with paid employment. Within this group, there would tend to be a negative correlation between family size and the fraction of their life they spend in paid employment. This is a correlation between two choice variables. Variations in each partner's wages and in tastes drive both variables.

Purchased childcare can weaken the link between a woman's wage and the cost of an additional child. In Ermisch (1989) it is treated as an increasingly imperfect substitute for the mother's time in child-rearing as the amount of it increases. Mothers with high wages tend to purchase a much larger proportion of childcare time. For them, higher wages have little effect on the cost of children, making it more likely that they increase fertility by raising family income. Similarly, in countries with heavily subsidized childcare, mothers contribute much less of childcare themselves, making it more likely that women earning higher pay have larger families. At low to moderate levels of wages, a higher mother's wage reduces fertility, but its negative impact attenuates as her wage rises, or the price of childcare falls, because mothers purchase a larger proportion of childcare time under these conditions. The impact of the price of childcare on fertility displays a similar interaction, becoming more negative as the mother's wage rises.

Nevertheless, when examining changes in fertility over time, the cost of purchased childcare and women's wages tend to move together, because women's labor is such an important input to the provision of childcare services. Thus, over time women's pay relative to men's and fertility should be negatively related, because higher women's pay raises the cost of children.

3. *Dynamic Fertility Models*

The theories discussed so far are 'static' in nature, the relevant time unit being the parents' lifetime. In order to consider decisions about the timing of births, imperfect fertility control, or the consequences of unexpected outcomes like birth-control failure or child mortality, a dynamic model is needed. First, some features of dynamic life-cycle theories are discussed, and then a model of fertility dynamics across generations is introduced.

3.1 *Life-cycle Dynamics*

Dynamic models incorporating a sequence of decisions and imperfect, costly fertility control are more realistic than the static theories discussed above. Their predictions often depend on what is assumed about parents' ability to borrow and save. One of two polar assumptions has been made in the literature: either parents can borrow and save as they choose ('perfect capital markets') or they cannot do so at all (i.e., they must consume all of their current income).

In one of the earliest dynamic fertility models (Heckman and Willis 1976), a couple's fertility is controllable by the contraceptive strategies they adopt, but control is not perfect. Because it is assumed that couples start obtaining satisfaction from children as soon as they are born, they wish to have their children early. Couples may, however, find it best to adopt a 'precautionary' contraceptive strategy early in their partnership in order to reduce the risk of having more children than they would have chosen to have if fertility control were perfect and free. Such a strategy would result in postponing the age at first birth. If people cannot borrow or save, and their incomes rise with age, they would also wish to practice contraception in order to try to postpone childbearing to a time when the marginal benefit of income for parental consumption is lower (when they can 'better afford' children). Subsequent spacing of births (rather than having children as quickly as possible) would also be a response to this tension between their desire to have children early and the economic incentive to have them later, when parents' income is higher. The more rapid the increase in income with age, the greater the incentive to postpone the start of childbearing and to space births. In other words, births are timed in order to 'smooth' parental consumption over time.

Another common (but not universal) feature of dynamic models is allowance for the possibility that a woman's current participation in paid employment improves her future earning capacity, and that absence from employment may reduce her pay when she returns to employment. Thus, a mother's absence from paid employment because of a birth has two costs: a loss of current earnings and a loss of future earning potential. An early example of such a model is that of Moffitt (1984).

Even when there are perfect capital markets, couples may choose to postpone childbearing because these opportunity costs of childbearing are high when the woman is younger and on the steep part of her career-earnings profile. Women in jobs in which work experience has little effect on earnings have less incentive to postpone. In addition, if the care of young children is particularly intensive in the mother's time, then a birth produces a temporary increase in the value of her time in the home to a level above the value of her time in the labor market, causing her to leave paid employment. It also increases the cost of an additional child, encouraging her to postpone the next birth until the value of her time declines sufficiently through the aging of the child. Thus, even with perfect capital markets, there is an economic incentive to space births. An unexpected birth, because of contraceptive failure, would generate similar incentives for the mother to leave paid employment temporarily and postpone having another child.

The death of a child encourages its replacement, because the marginal benefit of an additional child is likely to increase substantially when a child dies. Replacement makes a woman's number of births increase with the experience of child mortality. Unless child mortality is expected when the mother is older, replacement is a better response to child mortality risk than 'hoarding,' because the latter involves a larger deviation from what the parents would have done in the absence of child mortality. Wolpin (1984) was the first to incorporate uncertainty about child survival into a sequential, dynamic fertility model, which permits analysis of replacement in response to child deaths and of the effect of infant mortality risk on fertility.

Recent detailed surveys by Hotz et al. (1997) and Schultz (1997) discuss the empirical strategies for applying the theories discussed above and the results of many empirical studies.

3.2 *Easterlin's Relative Income Hypothesis*

In Easterlin's (1980) view, an important factor affecting a young couple's willingness to marry and have children is their outlook for supporting their material aspirations. If they expect their earning potential to be high relative to their aspirations, they will be more optimistic and will feel freer to marry and have children. If they expect that they will find it difficult to

achieve the standard of living to which they aspire, they will hesitate to marry and have children. The ratio of the couple's potential earning power to their material aspirations is called their 'relative income.'

A person's material aspirations are formed through experience; for a young person, the primary relevant experience is the standard of living experienced in his/her parents' home. Young people's expectations concerning their earning potential are largely shaped by their experience of working.

A person's relative income may be related to the size of his/her generation relative to the size of previous generations. There are important differences between the characteristics of new labor force entrants and experienced workers. The former have not yet acquired a high level of skills and are engaged in a considerable amount of job search with consequent high job turnover. Older workers are more skilled and occupy career jobs with relatively low job turnover. As a consequence of these differences, the ability to substitute between young and older workers is limited. Thus, when young workers are scarce relative to older workers, their wages are higher relative to older workers and their chances of upward mobility improve. When they are in relative abundance, their relative wage declines and their promotion chances deteriorate. If there is a steady growth in labor demand, the earnings of young people (say, aged 15–29) relative to older people (say, aged 30–64) tends to be inversely related to the number of young people relative to older people ('relative generation size'), which suggests that their relative income also moves inversely with relative generation size.

The relative income theory of fertility would then lead us to expect that people from smaller birth cohorts would have higher fertility than people from larger cohorts. If couples feel constrained to choose family size within certain limits, say two to four children, then 'scarce' generations, who start childbearing earlier because of their favorable economic prospects relative to their material aspirations, would cut back later as they reach their upper limits, and people from 'surplus' generations would catch up later to attain their lower limits. Such behavior tends to make the fertility rates of women of different ages (different generations) move together. The period total fertility rate would then also move inversely with relative generation size, and it would fluctuate much more than completed family size.

4. *Non-marital Fertility*

Since 1990 there has been a shift away from the paradigm of the married couple making fertility decisions as a single entity. This reflects in part the fact that births outside marriage now represent a substantial proportion of births in many countries, which suggests that men and women should be treated as individual agents in fertility decisions. Higher divorce rates would reinforce such a treatment.

This is clearest considering births outside a live-in partnership, drawing on the analysis of Willis (1999). Child quality is a 'collective good' to the parents (i.e., both the mother's and father's satisfaction depend on child quality). As a consequence, men fathering children outside a live-in partnership may choose to make transfers to the mother. Assume that the amount of transfer is determined in a noncooperative way: the mother has custody and chooses child quality and her own consumption taking transfer payments from the father as given. Fathers choose transfers and their own consumption, taking into account how the mother reacts to such a transfer. Then a father would voluntarily make a transfer if his income is high relative to the mother's income and he cares sufficiently for the well-being of his child. The higher is his income relative to hers, the larger is the transfer. But the mother, father, and child could be even better off if the couple cooperated in their decisions. Because of difficulties in monitoring expenditures on children, cooperation is facilitated by the parents forming a live-in partnership. Thus, these relatively affluent fathers would usually not want to father children outside a live-in partnership.

In a frictionless 'marriage market,' the collective-good nature of child quality encourages the highest income man to live with and have a child with the highest income woman, the next highest income couple would live together and have a child, and so on. If there are more men than women, then the men with the lowest incomes remain without a partner and all childbearing would be within live-in partnerships. If the number of women exceeds the number of men and if women at the lower end of the income distribution have incomes which are sufficient for them to want to raise children with their own resources, then fatherhood outside a live-in partnership would be free. If low-income men can father children by more than one woman, out-of-partnership fatherhood may be more advantageous for them than forming a union with a low-income woman, even though the mother and child would be better off in the union. This is more likely if the man's income is not much higher than the woman's, in which case he would not want to make transfers to her. The single low-income mothers gain by becoming mothers rather than remaining childless.

Thus, there may be what Willis (1999) calls an 'out-of-wedlock equilibrium' in which men with low incomes seek to father children outside live-in partnerships and women oblige them in preference to remaining childless. This outcome is more likely when there is an excess of unmarried women relative to unmarried men, men's incomes are not much higher than women's in the lower portion of the income distribution, and the combination of women's earnings and government transfer programs are adequate to raise children without the father's help. Thus, this

theory explains fertility outside live-in partnerships as a rational choice by low-income men and women, while at the same time higher income couples choose to have children within live-in partnerships.

In the world described by this model, people either form live-in partnerships or remain single forever. A simple model of time-consuming search for a partner, in which a woman may have a birth before entering a live-in partnership but still partner subsequently, yields a few similar predictions. Pre-partnership childbearing is more likely when the incomes of single mothers are higher relative to the incomes of single women (perhaps because of welfare benefits to single mothers), both because more partnership offers are rejected and because the value of becoming a single mother increases relative to the value of remaining single and childless. It is also more likely when single motherhood does not substantially reduce the chances of subsequent live-in partnering. The fathers of these children, whose partnership offers have been rejected, tend to have low incomes.

There has been little economic analysis of childbearing within cohabiting unions, which has grown in importance in many countries. In unions that are expected to have a short life, it may be akin to the analysis of out-of-partnership childbearing discussed above. More generally, it shares some features of the analysis of marital fertility, but there are also unique attributes of cohabiting unions that are probably important for fertility decisions. One is their high risk of dissolution. In light of it, it is essential to treat the fertility decision as problem of choice under uncertainty in which each partner is an individual agent whose preferences, expectations, and resources influence the decision. This presents new and difficult analytical problems, which are also relevant to marital fertility in high-divorce countries. These are new challenges in the economic theory of fertility.

See also: Family Size Preferences; Family Theory and the Realities of Childbearing Behavior; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Feminist–Economist Critique; Fertility: Institutional and Political Approaches; Fertility of Single and Cohabiting Women; Fertility: Political and Political–Economic Perspectives; Fertility Theory: Caldwell's Theory of Intergenerational Wealth Flows; Fertility Theory: Embodied-capital Theory of Life History Evolution; Mortality, Biodemography of; Motherhood: Economic Aspects

Bibliography

Becker G 1960 An economic analysis of fertility. In: *Demographic and Economic Change in Developed Countries*. National Bureau of Economic Research, Princeton University Press, Princeton, NJ, pp. 209–31

Becker G 1965 A theory of the allocation of time. *Economic Journal* **75**: 493–517

Becker G, Lewis H G 1973 On the interaction between the quantity and quality of children. *Journal of Political Economy* **81**: S279–S288

Cigno A 1998 Fertility decisions when infant survival is endogenous. *Journal of Population Economics* **11**: 21–8

Easterlin R 1980 *Birth and Fortune*. Grant-McIntyre, London

Ermisch J 1989 Purchased child care, optimal family size and mother's employment. *Journal of Population Economics* **2**: 79–102

Heckman J, Willis R 1976 Estimation of a stochastic model of reproduction: An econometric approach. In: Terleckyj N (ed.) *Household Production and Consumption*. Columbia University Press, New York

Hotz V J, Klerman J, Willis R 1997 The economics of fertility in developed countries. In: Rosenzweig M, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 241–347

Mincer J 1963 Market prices, opportunity costs and income effects. In: *Measurement in Economics: Studies in Mathematical Economics in Honor of Yehuda Grunfeld*. Stanford University Press, Stanford, CA, pp. 67–82

Moffitt R 1984 Optimal life-cycle profiles of fertility and labor supply. *Research in Population Economics* **5**: 29–50

Sah R 1991 The effect of child mortality changes on fertility choice and parental welfare. *Journal of Political Economy* **99**: 582–606

Schultz T P 1997 Demand for children in low income countries. In: Rosenzweig M, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 349–430

Willis R 1973 A new approach to the economic theory of fertility behavior. *Journal of Political Economy* **81**: S14–S64

Willis R 1999 A theory of out-of-wedlock childbearing. *Journal of Political Economy* **107**: S33–S64

Wolpin K 1984 An estimable dynamic stochastic model of fertility and child mortality. *Journal of Political Economy* **92**: 852–74

J. Ermisch

Family Theory: Economics of Intergenerational Relations

In economics, the term 'transfer' is used to refer to the delivery of money or other assets, without counterpart. We do not talk of transfers, therefore, in the context of a commercial transaction, where money is the counterpart of goods or services. Public transfers, by contrast, are the direct subsidies that governments pay to households and firms. Although such subsidies are financed out of tax receipts, there is in fact no direct link between what any particular firm or household pays in taxes, and what it gets back in subsidies. A gray area is represented by certain types of old-age pensions, where the size of the benefit is related in some way to contributions made, but the rules on the basis of which the calculation is made are subject to change by political decision.

Transfers between generations may be voluntary, or forced by the tax system. Here, we are concerned with

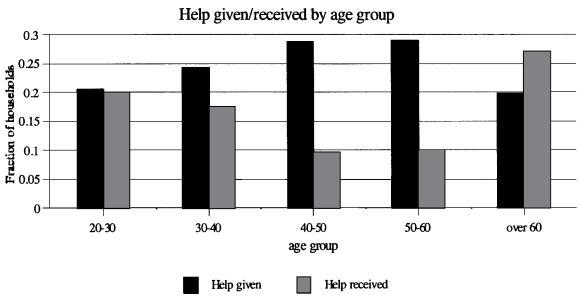

Figure 1
Help given/received by age group

voluntary transfers between different generations of the same family, but we are also interested in how these respond to public transfers. Typically, the latter redistribute income away from working-age cohorts, who pay more taxes and receive less subsidies than younger and older cohorts. A similar age pattern applies to voluntary *inter vivos* transfers: Fig. 1, drawn from Cigno and Rosati (2000), shows the distribution, by age of the household head, of monetary help to and from non-coresident 'friends or relatives' (overwhelmingly relatives) by Italian households during the late 1980s. It transpires that the probability of making a transfer is highest, and that of receiving one lowest, when the household head is in the 40–59 age range. Similar pictures can be constructed for other countries. What can be the reason for such a pattern?

Economic explanations fall into two broad categories. One follows what is called the altruistic approach, whereby individuals are assumed to derive direct utility (satisfaction) from the wellbeing of other family members. The other assumes that individuals are ultimately self-interested. Transfers are straight, no-strings-attached gifts according to the first of these explanations, part of a strategy that is expected to bring, directly or indirectly, benefits to the donor according to the second.

1. *Altruistic Explanations*

Simple altruistic explanations assume that sentiment is unidirectional: either from parents to children (*descending altruism*) or from children to parents (*ascending altruism*). More complicated versions allow for altruism in both directions.

Descending altruism underlies Richard Barro's and Gary Becker's theories of voluntary transfers. This assumption allows them to model transfers within the family as the outcome of a single optimization, whereby parents decide how much each family member will consume, subject to two constraints. One is that the sum of the consumption expenditures of all family members cannot exceed the sum of their incomes. The other is that transfers received (the difference between consumption expenditure and income) by any family member other than the decision-maker(s) must be positive or zero, otherwise they would be rejected. Under standard conditions, the model predicts that transfers will not decrease (increase) as the income of the donor (beneficiary) increases. This is consistent with the observation (Fig. 1) that the probability of making a transfer is highest, and that of receiving one lowest, in middle age, when incomes are generally at their peak. It is difficult to reconcile, however, with the observation that people continue to receive *inter vivos* transfers after the age of 60 (or even 75), when few of them have living parents.

If it is assumed that parents make transfers in the form of bequests, rather than *inter vivos*, the descending-altruism model predicts that household saving is increased, dollar for dollar, by any current government deficit (Barro 1974). That is because, conscious of the government intertemporal budget constraint, altruistic taxpayers perceive any current deficit as a tax on their descendants (this is the so-called *Ricardian equivalence*, from the name of David Ricardo, who first detected it). According to the model, taxpayers thus respond to a government deficit by saving more, in order to be able to leave their descendants bequests large enough to pay the extra taxes. One of the mechanisms that might generate such a deficit is a pay-as-you-go public pension system, where current pensions are paid out of current contributions. Whichever the mechanism, however, any net public transfer will be matched by private transfers of opposite sign.

Ricardian equivalence does not hold if fertility is endogenous i.e., if parents can choose how many children to have (or, more realistically, condition the probability distribution of births), and parents derive satisfaction from the number of children, as well as from how much each of them consumes (Becker and Barro 1988). Suppose, then, that the government were to take a dollar off the social security contributions due by each of the present workers, without any corresponding reduction in their pension entitlements. Each member of the next generation would then have to pay (r/n) more dollars in taxes, where r is the intergenerational interest factor, and n the ratio of future to current taxpayers (roughly, the population growth factor). Knowing that, present workers would then want to compensate their own children by making larger bequests to each of them. But this would make it more expensive to have an extra child relative to increasing the wellbeing of existing children. Under standard conditions, the present generation would then choose to have fewer children, and to *more* than compensate each child for the heavier tax burden he or she will have to shoulder. Of course, the change in fertility behavior would result in a lower n, and thus raise the burden on each future taxpayer, but the effect

of each couple's actions on the aggregate fertility rate is infinitesimal. Couples, therefore, will not take this effect into account when they decide how many children to have. Hence, fertility would fall. Saving could rise or fall, depending on whether the rise in transfers per child is proportionally larger or smaller than the reduction in the number of children. By contrast, if the government were to promise higher benefits to current taxpayers already covered by old-age security, or extend coverage to new categories, and simultaneously raise taxes or contributions so as not to force a transfer in favor of present workers, there would be no effect on either saving or fertility behaviour. There is evidence (Cigno and Rosati 1996, among others) that old-age security *per se* (i.e., without intergenerational spillovers) encourages household saving and discourages fertility, while deficits discourage saving. This empirical finding is consistent with the Becker–Barro prediction that a deficit *could* result in lower savings, but contradicts that model's prediction that, in the absence of intergenerational transfers, saving and fertility would remain the same.

The implications of ascending altruism (Nishimura and Zhang 1992) are somewhat different. The direction of transfers is not from working-age parents to young children, as in the case of descending altruism, but from working-age children to retirement-age parents. This is consistent with the observation that old people receive *inter vivos* transfers, but difficult to reconcile with the one that they also make such transfers. As children are a source of income, parents treat them as an asset. If fertility is endogenous, working-age individuals (or couples) then choose the composition of their portfolios so as to equate the marginal return to having children to that of holding conventional assets (capital). According to this model, a gift to current tax payers at the expense of future ones would induce parents to save more, not in order to leave larger bequests, but in order to put some of the windfall aside for old age. By contrast, a fully funded expansion in pension coverage (a form of forced saving) would reduce voluntary saving. These predictions clash, however, with the empirical finding, already mentioned, that deficits discourage voluntary household saving, while social security coverage *per se* has the opposite effect.

Whichever the direction, unilateral altruism is thus an inadequate explanation of what we observe. Can bilateral altruism (parents love children, children love parents) do better? If family members are unanimous in their preferences, intrafamily transfers can again be described as the outcome of the maximization of a single preference function (this time agreed upon by all family members), subject only to the constraint that the sum of the expenditures cannot be greater than the sum of the incomes of all family members. There is now no restriction on the direction of money flows: depending on life-cycle patterns, transfers from middle-aged children to elderly parents may be positive in some families and negative in others. This is consistent with evidence of two-way flows between generations, but the model also has a stronger behavioral implication: since the amount consumed by each family member depends on *total* family income, and not on his or her own individual income, changing the intrafamily distribution of income should not change the intrafamily distribution of consumption. Statistical tests based on this property of the model (Cox 1987, Altonji et al. 1992) reject the hypothesis of altruism. That property also has the strong implication that redistributive policies are ineffective: taking a dollar off the rich (the parents) and giving it to the poor (the children) would reduce the amount voluntarily transferred from the former to the latter by exactly one dollar.

In general, however, different family members are likely to have different ideas about the most desirable pattern of transfers. Even if all family members were altruistic towards one another, each of them might in fact appreciate his or her own consumption a little more (or, for that matter, a little less) than another family member. If that is the case, transfers cannot be viewed as the solution to a single optimization. They must rather be seen as the outcome of a game (in the sense of formal game theory) in which family members bargain with one another, or respond strategically to one another's moves, pretty much as if they were pursuing purely selfish aims (Manser and Brown 1980, Stark 1993). In such a situation, transfers depend on individual fall-back positions (what the player would get, if he or she dropped out of the game), as well as on preferences and characteristics. It is then difficult to distinguish empirically (McElroy 1990) bilateral altruism from self-interest, to which we now turn.

2. *Non-altruistic Explanations*

Non-altruistic explanations are based on the assumption that individuals are interested only in their own lifetime consumption stream. If a person voluntarily surrenders money to another, it will be for profit, or in order to improve the allocation of consumption over the life-cycle, or across states of the world. It must also be true that the person in question could not get a better deal from the market. There does not seem to be any difference of substance, therefore, between this and giving money to the grocer in exchange for a dozen eggs. Nonetheless, it is customary, maybe due to lack of information, to classify as transfers any moneys changing hands within the family whenever there is no explicit *quid pro quo* (not, however, in the case of a father paying his child to wash his car).

Mutual insurance ('I help you if things go well for me and badly for you, you help me if it is the other way round') is a typical form of intrafamily arrangement. But why should people want to enter into such informal, maybe even tacit, agreements, if it is possible to buy insurance from the market with all the safeguards

afforded by the law? The insurance market has the advantage over the family that individual risks are less likely to be correlated in a large heterogeneous population than in a small homogeneous group. Because of adverse selection and moral hazard problems, however, the market might not offer insurance, or offer it on very unfavorable terms, to certain individuals or for certain types of risk. Such problems are generally less acute within the family, the members of which have privileged information on one another's characteristics, and lower costs of monitoring one another's behavior. In some circumstances, it may thus turn out that intrafamily arrangements are more advantageous than market insurance, or that there is no alternative.

Mutual insurance may or may not result in intergenerational transfers (not if the agreement is among contemporaries). Intergenerational transfers, by contrast, are essentially another kind of insurance arrangement whereby a person (parent) transfers wealth or pre-commits to leave a bequest to a younger relative (child) in exchange for a promise of indefinite old-age support, effectively in exchange for an annuity (Kotlikoff and Spivak 1981). That way, the parent buys insurance from the child against the risk of running out of money after too long a period of retirement.

Another type of mutually advantageous intrafamily deal, also giving rise to intergenerational transfers, stems from life-cycle income variations and capital market imperfections. Standard microeconomic theory predicts that individuals smooth consumption over the life-cycle by borrowing from the market when they are young, and lending (buying assets) when they are middle-aged. In reality, however, minors cannot sign contracts, and young adults have difficulty in obtaining credit against future earnings. The very fact that people survive to adulthood is thus evidence that the young receive support in other ways, but from whom? We have just seen that nobody makes genuine presents. Self-interested relatives, on the other hand, will make loans, not presents, and only if the risk-adjusted return is at least as high as that offered by the formal capital market. And what is there to induce a selfish adult to honor an informal debt incurred 20 or so years earlier?

Imagine a set of family rules (family constitution) prescribing the amount that a parent must pay a young child, and the amount that a grown-up child must pay an elderly parent, subject to the clause that nothing is due to a parent who disobeyed the rules. Such a constitution is self-enforcing, in the sense that it is in everybody's interest to comply, if the rules are so designed that complying with them is at least as advantageous, for a middle-aged person, as the alternative strategy of having no children, giving nothing to parents, and providing for old age by buying capital (Cigno 1993). This line of reasoning explains transfers to the young as implicit loans, and transfers to the old as implicit loan repayments. In such a situation, paying back one's own parents is a condition for getting the same treatment from one's own children. Now suppose that the level of old-age security is raised, or coverage extended to new categories. The present middle-aged affected by the policy will then want to shift less of their current spending capacity to old age. Since the amount they are supposed to pay to their elderly parents has not changed, however, some of these middle-aged persons will find it advantageous to disobey the constitution, and consequently give up the idea of being supported by their own children in old age. They will then default on their implicit debt towards their own parents, have no children, and possibly top up their expected pensions by buying assets. Conversely, if the government were to cut current taxes or social security contributions without reducing pension entitlements, some middle-aged individuals, who would otherwise have found it advantageous to buy assets, would have children instead. This is consistent with the already mentioned empirical evidence that aggregate household saving responds positively to old-age security, and negatively to deficits. It also explains why fertility responds negatively to old-age security.

See also: Age: Anthropological Aspects; Family and Gender; Family and Kinship, History of; Family, Anthropology of; Family as Institution; Family Systems in Europe; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Childbearing; Family Theory: Feminist–Economist Critique; Family Theory: Role of Changing Values; Kinship in Anthropology; Parenting: Attitudes and Beliefs

Bibliography

Altonji J G, Hayashi F, Kotlikoff L J 1992 Is the extended family altruistically linked? Direct evidence using micro data. *American Economic Review* **82**: 1177–98

Barro R J 1974 Are government bonds net wealth? *Journal of Political Economy* **82**: 1095–1118

Becker G S 1974 A theory of social interactions. *Journal of Political Economy* **82**: 1063–93

Becker G S, Barro R J 1988 A reformulation of the economic theory of fertility. *Quarterly Journal of Economics* **103**: 1–26

Cigno A 1993 Intergenerational transfers without altruism: Family, market and state. *European Journal of Political Economy* **7**: 505–18

Cigno A, Rosati F C 1996 Jointly determined saving and fertility behaviour: Theory, and evidence for Germany, Italy, UK and USA. *European Economic Review* **40**: 1561–89

Cigno A, Rosati F C 2000 Mutual interest, self-enforcing constitutions and apparent generosity. In: Gérard-Veret L A, Kolm S C, Mercier-Ythier J (eds.) *The Economics of Reciprocity, Giving and Altruism*. Macmillan, London

Cox D 1987 Motives for private income transfers. *Journal of Political Economy* **95**: 508–46

Kotlikoff L J, Spivak A 1981 The family as an incomplete annuities market. *Journal of Political Economy* **89**: 372–91

Manser M, Brown M 1980 Marriage and household decision making. *International Economic Review* **21**: 31–44
McElroy M B 1990 The empirical content of Nash bargained household behavior. *Journal of Human Resources* **25**: 559–83
Nishimura K, Zhang J 1992 Pay-as-you-go public pensions with endogenous fertility. *Journal of Public Economics* **48**: 239–58
Stark O 1993 Nonmarket transfers and altruism. *European Economic Review* **37**: 1413–24

A. Cigno

Family Theory: Economics of Marriage and Divorce

From an economic point of view, marriage is a voluntary partnership for the purpose of joint production and joint consumption. As such, it is comparable to other economic organizations that aim to maximize some private gains, but are subject to market discipline. This article describes the working of the marriage market, as a feedback system in which the aggregate divorce and marriage rates influence and are influenced by individual choices, and explains the features of this market that have led to the sharp increase in marital turnover.

1. *The Gains from Marriage*

The economic analysis of the family naturally begins with the gains from marriage. These gains are a major determinant of the decision to form a new union or to break an existing relationship. The main economic advantages of being married, rather than being single, are as follows.

(a) *Children*. Although children can be produced and raised outside the family, the family has a substantial advantage in carrying out these activities. There are two inter-related factors that cause this advantage: by nature, parents care about their own children and, because of this mutual interest, it is more efficient that the parents themselves determine the expenditure on their children. Because children are viewed as a public good by their parents, an efficient allocation of family resources between public and private uses requires some coordination between the parents. If the parents live separately, as either single or remarried, the noncustodian parent loses control of child expenditures and is less likely to contribute. In addition, lack of contact may reduce the parent's benefits from the children's 'quality' (i.e., their well being and success).

(b) *Division of labor*. Family members coordinate their work activities to exploit comparative advantage and increasing returns. For instance, one partner works in the market while the other works at home. This pattern is economically efficient if the partners have different market wages or different productivity at home. Such differences may be inherent or, more likely, acquired through investment in skills. Investment in human capital is more profitable the higher its rate of utilization. This type of increasing returns leads to specialization that can, in principle, create mutual gains through redistribution.

(c) *Sharing*. There are collective (nonrival) goods that can be jointly consumed by both partners: for example, child quality, common leisure activities, housing expenditures, and shared information. The share of such public goods in family expenditures can be substantial. Lazear and Michael 1980 estimate that two single individuals can almost double their purchasing power by forming a union.

(d) *Credit and investment*. Married partners often engage in implicit loans that facilitate investment activities. For example, one partner works when the other is in school, expecting to share in the future gains from the investment. Similarly, parents often finance the schooling of their children, in part in the anticipation of old-age support. Such internal transfers can support investments that are mutually beneficial and would not be carried out because of market imperfections.

(e) *Risk pooling*. A family can spread the risks of idiosyncratic shocks to individual members. For example, one partner can work when the other is sick or unemployed. The family jointly accumulates precautionary savings that can serve *either* member in case of need.

There is plenty of evidence for division of labor within families. Typically, husbands work more in the market than do their wives. In addition, married men work longer hours in the market and have substantially higher wages than unmarried men. Similarly, married women have lower wages and work more at home than unmarried women. This division of labor is influenced by the higher market wages of males, which are reinforced by their higher participation in the labor market. Somewhat surprisingly, the evidence for mutual insurance is not easy to come by. In particular, an employed person is generally *more* likely to have an employed spouse than an unemployed person, suggesting that common factors affect the unemployment of both spouses.

The family is not the only organization that fulfills these functions. Markets and government institutions compete with and complement the family in this regard. If all goods and work activities are marketable, there is no need to form marriages to enjoy increasing returns or to pool risks. Similarly, with good medical or unemployment insurance one does not need to rely on one's spouse. Generally, the role of the family varies depending on market conditions and government policies. In addition, human partnerships need not be confined to couples of opposite sexes. One also observes 'extended families' of varying structures that

coordinate the activities of their members and provide self-insurance. The prevalence of male–female partnerships has to do with sexual attraction, which triggers some initial amount of blind trust and the production of children. These emotional and biological considerations are sufficient to bring into the family domain some activities that could be purchased in the market. Then, the accumulation of specific 'marital capital,' such as children and shared experience, increases the costs of separation and creates incentives for a lasting relationship. In this sense, there is an accumulative effect in which economic considerations and investments reinforce the natural attachment.

The potential gains from marriage generally depend on the personal traits of the two partners. Some individual traits are complements, and the gains from marriage rise if the partners have similar endowments. For instance, similarity in tastes facilitates decision on collective goods, such as the education of the children. In other cases, the traits are substitutes, and gains from marriage rise if the partners have different traits. For instance, the gains from division of labor and from credit are higher when the wages or incomes of the partners differ.

2. *Family Decision Making*

The existence of potential gains from marriage is not sufficient to motivate marriage and to sustain it. Prospective mates need to form some notion as to whether families realize the potential gains and how they are divided. Because families consist of several members with potentially conflicting interests, there is a question as to how families reach decisions. The old notion that consensus leads to joint maximization of a common objective appears too narrow. However, because of their proximity and repeated interaction, it is plausible to assume that married partners reach some sort of agreement that specifies an *efficient* allocation of resources and a *stable division rule*. These two principles replace the principle of rationality in individual choices.

In a special case, referred to as *transferable utility*, it is possible to separate the issues of efficiency and distribution. This situation arises if there exists a commodity (say money) that, upon changing hands, shifts utilities between the partners at a fixed rate of exchange. In this case, the family decision process can be broken into two steps: first choose an action that maximizes a weighted sum of the individual utilities, then use money to divide the jointly maximal outcome. The basic idea is that the partners can agree on most aspects of their choice and restrict their bargaining to transfers of money, which do not detract from total resources. In general, the problems of efficiency and distribution are intertwined. The family can still be described as maximizing a weighted sum of the individual utilities, but the weights depend on the individual bargaining powers and any shift in the weights will affect the family choice. The bargaining power may depend on individual attributes, such as earning capacity, subjective factors such as impatience and risk aversion, and on market conditions, such as the sex ratio and availability of alternative mates.

Realization of the potential gains from marriage requires coordination between the two partners and durable relationship. Durable relationships are motivated by the long-term investment in children and the accumulation of marital capital, which is lost or diminished in value if separation occurs. However, to induce the actions that generate the gains from marriage, it is necessary for the partners to exchange *commitments*. For instance, a young wife is unlikely to support her husband by working while he is in medical school, if she expects him to leave her and marry a young nurse when he gets his degree. To some extent, emotional factors such as love and guilt replace formal commitments and facilitate efficient investments and effort levels. As an example, consider the role of an altruistic family head. Suppose that the head is given control over family resources and can make transfers as the head sees fit. The only requirement for becoming a head is that they should care about *all* family members, willing to raise their income, whenever the total resources at the head's disposal rise. Under these conditions, selfish family members voluntarily act in the interest of the group. The reason is that any productive (destructive) action which increases (decreases) total family resources is rewarded by an increased (decreased) transfer from the family head (Becker 1991).

It is empirically testable whether families maximize a joint utility function, whether they act efficiently, and it is also possible to recover the sharing rule from observed family choices. Recent evidence shows that, holding total family income constant, an increase in the income of one member shifts the allocation of consumption goods in their favor. These findings reject the hypotheses that the family maximizes a joint objective, which would imply that within family, allocations are invariant under such redistribution of total resources, but are consistent with both cooperation and noncooperation. However, only cooperation implies efficiency, which yields testable cross-equation restrictions on family demand. Based on data on consumption and work patterns within families, efficiency is not rejected, suggesting that cooperation is the relevant mode (Browning et al. 1994). In the context of uncertainty, however, efficiency has some further implications that are rejected. Efficient risk sharing among linked individuals implies that, holding aggregate consumption constant, the consumption of each family member is independent of idiosyncratic shocks such as fall into unemployment or bad health. Stated differently, all individuals in the household are affected by a random shock to any individual income

and their consumption levels move together. Because of lack of data on individual consumption, the tests of this hypothesis involve larger units, such as villages or extended families, and, at this level of aggregation, efficiency is rejected.

3. *Divorce and its Economic Consequences*

Divorce is motivated by two general considerations. First, because it takes time to find a suitable match, people may enter a relationship which they intend to break in the future if a better match is met. Second, because of changing economic and emotional circumstances, the gains from marriage may dissipate. As time passes, new information on match quality and outside options is accumulated, and each partner decides whether to dissolve the partnership or to continue the marriage. In making this choice, partners must consider the expected value of each alternative, where the value of remaining married includes the option of later divorce, and the value of divorcing includes the option of later remarriage. Divorce occurs endogenously whenever one partner has an alternative option that the current spouse cannot, or is unwilling to, match by a redistribution of the gains from marriage.

The model outlined above yields several testable implications:

(a) It is the *unanticipated* changes in the characteristics of the partners or the quality of match which trigger divorce. It is clear that falling out of love can destroy the gains from the current marriage and cause divorce. It is less obvious how unanticipated changes in personal attributes, such as earning capacity, influence divorce, because they affect both the gains from the current marriage and the outside options. However the partners were matched, based on their traits as observed at the time of marriage, any unanticipated change reduces the gains from interaction. Thus, if the husband is unexpectedly wealthy, he will seek a better wife, and if he becomes unexpectedly poor, she will seek a better husband. Note that both positive and negative 'surprises' can disrupt the marriage.

(b) If the gains from marriage are substantial, small shocks will not lead to divorce. Therefore, the probability of divorce will be lower amongst couples who are well matched. Anticipating that, couples sort into marriage according to characteristics that are likely to enhance the stability of the marriage. In this respect, the decisions to marry and divorce are linked.

(c) By the same logic, costs of divorce mitigate the impact of unanticipated shocks on marital dissolution. The costs arise from loss of specific capital, such as information about the current spouse, and loss of coordination, especially with regard to child expenditures. There are also emotional and legal costs associated with the break up.

(d) Somewhat more controversial is the role of divorce laws, in particular whether the legal possibility unilaterally to walk away from a marriage influences the divorce rate. With transferable utility, such a change should only affect the shares in the gains from marriage but not the decision to separate, because the partner who wishes to continue the marriage can compensate the one who wishes to leave by giving up part of the gains from the marriage. Conversely, the person who wants to leave can pay damages to the one left behind. In general, because the unexpected shock differs between partners and because the options for compensations are limited, divorce is less likely if mutual consent is required.

These implications have been tested empirically. There is evidence that surprises matter. Specifically, an unexpected increase in the husband's earnings capacity reduces the divorce hazard, while an unexpected increase in the wife's earning capacity raises the divorce hazard. The important roles of search and costs of divorce are indicated by the findings that higher age at marriage and the presence of children stabilize the marriage. There is only weak evidence that divorce rates are higher in states where 'fault' is not a prerequisite for divorce. There is clear evidence that individuals sort into marriage based on their anticipation of divorce. Thus, couples with similar schooling attainments at the time of marriage are less likely to divorce, and individuals are more likely to marry if they have a similar amount of schooling. The same patterns hold for religion and ethnicity (Weiss 1997).

The costs of divorce are influenced by the post-divorce transfer of resources in the form of alimony and child support. It has been observed that divorced husbands, even if relatively well to do, fail to support their ex-wives and their children at the standard to which they were accustomed during marriage. Consequently, divorced women and children in their custody suffer a large decline in income. The possible explanations for this phenomenon are: lack of binding marriage contracts, the inability of noncustodial parents to monitor expenditures by the custodian, and loss of interest in the child by the noncustodian parent. In most cases, the wife obtains custody and controls the expenditure on children. This situation creates an *agency problem*, because out of every dollar transferred to the custodial wife with the intention of raising the welfare of the child, part is used for her own consumption. Because of this 'tax,' noncustodial fathers reduce their transfer, and consequently child quality is reduced. The assignment of custody to the wife is usually motivated by her comparative advantage in child care, but the agency problem implies that such a practice has shortcomings that are not easy to overcome, because the courts cannot verify within-household allocations.

The adequacy of transfers must be judged not only on the basis of the efficiency of allocation of family

resources in the aftermath of divorce, but also on the basis of partners' options and expectations at the time of marriage. From the latter point of view, efficiency requires that the consumption levels during marriage and divorce are tied together. Courts do take into account these two broad considerations in awarding divorce settlements. These are usually based on post-divorce incomes and the accustomed consumption of the wife and children during marriage. However, the evidence suggests that actual transfers fall short of the efficient level from both the *ex post* and *ex ante* points of view (Weiss 1997).

With deficient transfer mechanisms, the partners must better prepare for the event of divorce. One important instrument is the allocation of time within marriage. By investing in human capital, each partner can be less dependent on transfers in the event of divorce. However, such investments may detract from marital output. For instance, a wife who works is better defended against divorce but has less time to spend with children. Indeed, there is evidence that women tend to increase their investment in market work in anticipation of divorce. Thus, if the risk of divorce rises, children may suffer even before the marriage breaks.

4. The Marriage Market

Individuals in society have many potential partners. This situation creates competition over the potential gains from marriage. In modern societies, explicit price mechanisms are not observed. Nevertheless, the assignment of partners and the sharing of the gains from marriage can be analyzed within a market framework, because an undesired marriage can be avoided or replaced by a better one.

Any market solution is a particular assignment of males to females. The major questions of interest are: what types of matches are likely to form in equilibrium; for instance, would a class structure emerge where the rich marry the rich? What determines the rates of marital turnover and the proportions of married and single individuals in society? Does the marriage market operate efficiently and, if not, what type of intervention is required?

Matching models provide a starting point for such analysis. These models investigate the mapping from preferences over prospective matches into a *stable assignment* (Roth and Sotomayor 1990). Reflecting the assumption that marriage is voluntary, an assignment is said to be stable if no married person would rather be single and no two (married or unmarried) persons prefer to form a new union. To illustrate, assume that each male is endowed with a single trait, m, and each female is endowed with a single trait, f, which positively affects the gains from marriage, denoted by $z = g(m, f)$. Suppose, first, that z is a public good that *both* partners can consume jointly. Then, the only stable assignment is that in which males with high m marry females with high f, and, if there are more (less) eligible men than women, the men (women) with the lowest endowments remain unmarried, a positive *assortative mating*. However, if one assumes, instead, that z can be divided between the two partners, a positive (negative) assortative mating occurs only if the two traits are complements (substitutes). Matching models refer to these two situations as nontransferable and transferable utility, respectively. With transferable utility, a man with low m may obtain women with high f by giving up part of his private share in the gains from marriage. Complementarity (substitution) means that the two traits interact in such a way that the benefits from a woman with high f are higher (lower) for a male with high m than for a male with low m. The type of interaction in the gains from marriage determines the willingness to pay for the different attributes. An important lesson is that in a marriage market, with sufficient scope for compensation within marriage, the best man is not necessarily the one married to the best women, because, with negative interaction, either one of them can be bid away by the second best of the opposite sex (Becker 1991).

The process of matching in real life is characterized by scarcity of information about potential matches. The participants in the process must spend time and money to locate their best options. The realized distribution of matches and the division of the gains from each marriage are therefore determined in an equilibrium which is influenced by the costs of *search* and the search policies of other participants. The main ingredients of the search model are as follows. There is a random process that creates meetings between members of society of the opposite sex. When a meeting occurs, the partners adopt a reservation policy where the match is accepted only if it provides expected gains above some prespecified value. Otherwise, they depart and wait for the next meeting to occur (Mortensen 1988). Because meetings are random and sparse in time, those who actually meet and choose to marry enjoy a positive rent. An important issue is the division of these rents between the partners.

The literature mentions two considerations that determine the division of the gains from marriage: outside options, reflected in the value of continued search, and the self-enforcing allocation that would emerge if the marriage continued without agreement. Combining these two considerations, the sharing rule will be influenced by both the value of search as singles (outside the current match) and the value of continued search during the bargaining process, including the option of leaving when an outside offer (whose value exceeds the value of potential agreement) arrives. In this way, a link is created between the division of gains and the market conditions. For instance, if there is excess supply of women in a particular marriage market, the value of continued search by females is

reduced and consequently their share in the gains from marriage declines.

With friction, there is still a tendency to positive (negative) assortative mating, based on the type of the interaction in traits. If the trits are complements, individuals of either sex with higher endowment will adopt a more selective reservation policy and will be matched, on the average, with a highly endowed person of the opposite sex. However, with sufficient friction, it is possible also to have negative assortative mating under complementarity. The reason for this result is that, because of the low frequency of meetings, males with low m expect women with high f to accept them, and if the gains from such a match are large enough, they will reject all women with low f and wait until a high f arrives (Burdett and Coles 1999).

Generally, one would expect negative sorting on wages and positive sorting on nonwage income, because the gains from division of labor are larger when wages differ, while the gains from sharing goods are larger when incomes are similar (Becker 1991). Empirical findings suggest positive sorting on both wage and nonwage income. In particular, there is a substantial correlation in the schooling achievements of partners to marriage. Similarity in schooling has opposing effects on the gains in marriage; it reduces the gaps in wages and thus the gains from the division of labor, but it also induces similarity in tastes, which facilitates the allocation of public goods. It seems that the latter effect is empirically more important. We should further note that meetings are not really random, and that unattached individuals select jobs, schools, and leisure activities so as to affect the chances of meeting a qualified person of the opposite sex.

Frictions and search explain why, in any given moment in time, part of the population is single, even though there are positive gains from marriage. The aggregate stock of unattached individuals is determined by the rates of entry and exit into marriage implied by the arrival rates of offers and the optimal reservation policies chosen by individuals. These decisions, in turn, depend on the aggregate proportion of singles, because if one meets an unattached person it is more likely that a new union will be formed than upon meeting an attached person, who already has some positive rents. Therefore, the higher the proportion of singles, including divorcees, the higher the private expected gains from divorce. This type of reinforcement can lead to multiple equilibria, whereby both high and low aggregate divorce rates can be sustained through the induced change in the individual incentives to marry and divorce.

The time pattern of aggregate divorces is consistent with multiple equilibria. For instance, the divorce rate in the USA doubled during the decade between 1965 and 1975, standing at roughly 10 percent from 1940 to 1965 and at roughly 20 percent from 1975 to 1995, suggesting a switch from a low to a high equilibrium. Explanations for the timing of the change include the appearance of the contraceptive pill, break up of norms, and legal changes (Michael 1988). The main conceptual point, however, is that any change in exogenous factors, even a small one, can have marked impact on the aggregate divorce rate, because of the inherent reinforcements in marriage markets.

The marriage market determines not only the assignment of partners but also the division of resources and activities within the family. As market conditions change, a shortage of suitable partners of a particular kind leads to an increase in their gains from marriage. For instance, there is evidence that an increase in the demand for men, created by faster population growth combined with the tendency of men to marry younger women (a marriage squeeze), has led to an increase of dowries in rural India. In modern societies, up-front payments are rare and market forces are mostly revealed by the division of labor within families. The trend of rising female participation in the labor force among married people is probably associated with a larger share that women can extract from marriage, following the rise in female wages. Additional information on the (expected) gains from marriage is contained in the decisions to enter marriage and to stay married. For instance, it has been observed that black women in the USA delay their marriage and have children out of wedlock, because of a shortage of eligible black men (Willis 1999).

5. *Some Policy Issues*

Despite its firm roots in nature and its antiquity in human society, the future of the family institution has been recently put into question. This was motivated by recent trends, common to all Western societies, of declining marriage rates, declining fertility, higher divorce rates, and a rise in alternative arrangements such as cohabitation, single-parent households, and single-mother families. Social observers view these trends with considerable alarm, especially because of the potential harm to children (more than one million children in the USA are involved in a divorce every year).

Although there is ample evidence that divorce reduces the welfare of single wives and of children with single or step-parents, this is only one part of the picture. Continuation of marriage under adverse conditions can have equally harmful results, although these are harder to identify. Broadly viewed, divorce is a corrective mechanism that enables the replacement of bad matches by better ones. There is a risk that a better match will not be found, in which case the person who has divorced and the children are worse off. But rational agents take this consideration into account and can make financial arrangements to ensure that separations occur only if they improve the welfare of all parties. The issue, though, is whether the courts can enforce such binding contracts and the extent to which they should intervene.

Because meetings are random and uncoordinated, the search behavior and the decision whether to marry or divorce of each individual influence the marriage (and remarriage) prospects of all other members of society who are potential matches. Therefore, legal intervention must take into consideration the impact of the law not only on particular families but also on the marriage market as a whole. These two considerations may be in conflict. As an illustration, consider an increase in the size, or the enforcement, of child-support payments. Holding the aggregate divorce and remarriage rates fixed, each family in isolation may be better off as a result of the additional insurance. However, the reduced propensity to divorce can have a negative effect on the remarriage prospects of those who have suffered a large negative shock to their quality of match and who wish to separate. Another externality that must be considered, and may operate in the opposite direction, is that the availability of alternative mates can reduce investments that are specific to the current match and detract from its quality. This is indeed the logic behind the laws and religious norms aimed at restricting divorce.

The observed changes in family arrangements are closely related to the dramatic changes in labor-market participation of women, and the associated technological advances in the workplace and in the home. Whatever the causal relationships, it is clear that these processes reinforce each other and generate complex dynamics. While the reduction in marriage has been fairly smooth, the changes in divorce appear discontinuous, suggesting search externalities and perhaps the breaking of norms in the processes governing divorce and remarriage. These abrupt changes cause a special adjustment problem that requires social and legal restructuring. However, the expected outcome for the early twenty-first century is that the marriage market will settle at a new equilibrium with higher marital turnover, reflecting the lower gains from marriage as the disparity in the market productivity of men and women gradually declines. Although reduced, the gains from marriage are unlikely to disappear and the family, in one form or another, will continue to have a major role.

See also: Divorce and Gender; Divorce, Sociology of; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Intergenerational Relations; Marriage

Bibliography

Becker G 1991 *A Treatise on the Family*, enl. edn. Harvard University Press, Cambridge, MA

Browning M, Bourginion F, Chiappori P A, Lechene V 1994 Incomes and outcomes: A structural model of intra-household allocation. *Journal of Political Economy* **102**: 1067–96

Burdett K, Coles M G 1999 Long-term partnership formation: Marriage and employment. *Economic Journal, Series F* **109**: 307–34

Lazear E P, Michael R T 1980 Family size and the distribution of real per capita income. *American Economic Review* **70**: 91–107

Michael R 1988 Why did the US divorce rate double within a decade? *Research in Popular Economics* **6**: 367–99

Mortensen D T 1988 Matching: Finding a partner for life or otherwise. *American Journal of Sociology* **94** (suppl.): 215–40

Roth A, Sotomayor M 1990 *Two Sided Matching: A Study in Game-theoretic Modeling and Analysis*. Cambridge University Press, Cambridge, UK

Weiss Y 1997 The formation and dissolution of families: Why marry? Who marries whom? and What happens upon divorce? In: Stark O, Rosenzweig M (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam

Willis R 1999 A theory of out-of-wedlock childbearing. *Journal of Political Economy* **107** (Part 2 suppl.): 33–64

Y. Weiss

Family Theory: Feminist–Economist Critique

1. Feminist Economics

The critique from feminist economics of economic family theory largely has been directed at the version developed in neoclassical economics, often referred to as the 'new home economics,' though Marxist economic approaches to the family have also come in for their share of the criticism. The latter approaches largely were explicated in the 'domestic labor debates' of the 1980s, with articles appearing on both sides of the Atlantic showing how Marxist concepts could be used to study household production. Both these bodies of family theory were developed mostly in the 1970s, though their origins go back to the nineteenth century and even earlier.

Feminist economics can also be traced back to earlier centuries, but it emerged as a discipline in its own right in the 1970s and has grown most strongly in the 1990s. It is the branch of economics that seeks to understand the economic aspects of the relations between women and men and the gendered aspects of economic activity. With this dual goal of inquiry, feminist economics substantially broadens the scope of human activity to which economics can be applied. While economics is most commonly defined as the study of the allocation of scarce resources to meet human needs, in practice that study has been limited to primarily market transactions. Scarcity implies choice and neoclassical economics generally posits that 'economic man' acts rationally to maximize his own well-being, choosing options subject to constraints. Feminist economics suggests broadening the scope to a comprehensive study of provisioning and distribution, or to how society meets all its needs for everyone (Nelson 1993).

This would include housework and home production as well as the labor of caring for others, much

of which is done by women within the family. Caring necessarily intertwines affection and work, and maximization of an individual's self-interest may not be the guiding principle of this type of production, requiring that, at a minimum, the simple neoclassical model be made more complex. Since a monetary measure is absent, it is difficult to determine precisely how much production occurs in the household sphere, but worldwide it has been estimated that one-third of all human production results from the unpaid work of women (UNDP 1995, p. 97). Feminist economists argue that if these aspects of economic activity are not considered, one has a very incomplete view of the economy.

As Peterson and Lewis point out in the introduction to their 800-page compendium on feminist economics (Peterson and Lewis 1999), this is a branch of economics that welcomes diversity and heterodoxy. Feminist economists borrow from many traditions of economic thought as well as insights from other social science disciplines (Kuiper et al. 1995). While some feminist economists work largely within the neoclassical tradition and seek to use its methods to answer feminist questions, others work more within the Marxist framework or institutional economic tradition, while yet others use philosophical, literary, or postmodern methods. Most agree, however, that traditional economics of all schools has failed to explicate the economic aspects of gender relations, specifically the differences in economic power between women and men or the subordination of women to men. Why do women earn less than men, on average, just about everywhere, why do women do most of the family care in virtually all human societies, why do men hold most of the positions of economic leadership as owners, top managers, financiers, and so on?

2. *The Neoclassical Theory of the Family*

Neoclassical economists did not turn their attention to the question of how economic factors affect decision-making within the family and the allocation of the time of family members among various pursuits until the market started to impinge on the family and to change family dynamics in an obvious way when labor force participation by married women began to grow in the United States in the 1960s. Before this period, the family was thought of mainly as a consumption unit spending income earned by the man of the family, who went outside the family to produce goods and services for the marketplace. As Albelda (1997) notes, this was the typical family arrangement of the white, middle-class men who founded the separate discipline of economics and increasingly professionalized it in the late nineteenth and early twentieth centuries in the United States and Europe.

Becker (1965, 1973) and Mincer (1962) developed a model of the family unit in which men who specialize in market work (at which they are assumed to be relatively more efficient) and women who specialize in domestic production (at which they are assumed to be relatively more efficient) marry each other and maximize their joint satisfaction in a sort of trade exchange based on their comparative advantages. Each human needs both market goods and home-produced goods to survive (since the goods are substitutable but not perfectly so). If relative prices were to change, women and men would decide to trade different amounts with each other. The division of labor within the family (and the supply of family members' labor time to the market) can change with the male–female wage differential, changes in male–female productivity in housework, and the availability and quality of substitutes, all of which can be affected by technological change. Rising real wages, especially for women, would tend to lead women to work more outside the home, substituting market-based goods and services for their housework time. Fertility will decline as women will want to spend more time at market work and less at home production. Moreover, as technological change occurs and society becomes more complex, the quality of children (their education and socialization) becomes more important than their quantity. More market-based goods (more education and better food and clothing) can be combined with fewer children to yield higher quality output (children with more human capital). In this way the productivity of home production keeps pace with the growing productivity in market-based goods. In the several decades since its origins, the neoclassical theory of the family has been developed by the work of many economists carefully examining such issues as the allocation of time, the supply of labor, the development of human capital, population growth, and economic development.

The neoclassical approach views the household's role in the larger economy and society as the provision of labor to the market and the consumption of final products provided by the market. The household center (or factory) which performs these two functions combines market work, nonmarket work, and consumer products in an economically rational way: the mainspring of the system is the choice of individuals—who, faced with market options and given their own preferences and relative productivities, allocate their time and money accordingly. Thus, household production is the creation and maximization of utility to consumers in the household. It is assumed that the members of the household have harmonious interests which can be maximized by joint effort and decision-making (Hartmann 1974).

3. *The Feminist Critique*

Feminist economics challenges nearly every aspect of neoclassical economics. First, feminist economics challenges the limited role accorded the family or household, noting that caring labor is a large part of social reproduction and involves not only creating new

laborers by raising children and maintaining adult workers through daily care but also the care of those unable to work, not only the young, but also the sick, disabled, and elderly. Women's unpaid family work also includes maintaining kin networks and relationships, which provide a type of social insurance for the family. Women also contribute volunteer time to school, neighborhood, religious and community groups, which provide necessary services to family members, especially in countries without extensive welfare state services. Elson (1994) refers to all this crucial social reproductive work as the 'caring economy,' a term which brings its importance to the fore. The quantity and quality of the actual labor that takes place in the family is seldom examined carefully by neoclassical economists and thus the family portrait they paint is simplistic and incomplete. Too often women's caring labor is not seen as skilled work that must be learned but rather as an innate ability that women have, or as the manifestation of their natural affection.

Moreover, since in feminist economics caring labor combines affection and work, economic rationality and love, or its withering away, may dictate different 'choices.' A disabled child almost always requires more care and attention than an able-bodied one, yet investing more in such a child who most likely will be less productive hardly seems economically rational; such a 'choice' may be more a matter of affection and love than one of economic rationality. Economic rationality implies 'free' choice, choice that is not fundamentally coerced even though it is faced with constraints. If love or affection has waned, even otherwise enjoyable activities like sex could become oppressive. Women sometimes stay on even in violent relationships because their alternatives seem so limited to them—can such behavior be deemed a free or rational choice?

Second, the neoclassical economics model assumes choice by individuals of different relative productivities exercising their individual preferences, but it fails to explain why domestic labor is most often performed by women. It offers no convincing explanation of why this would be so (other than to assert that women's and men's abilities and preferences differ markedly on average, an assertion that has not been proved). Moreover, one cannot just assume that men's and women's preferences are determined individually. Instead, social roles may rather be collectively determined or politically enforced to such an extent that it is difficult for an individual to express a different preference. Even in the present-day United States, where it is common knowledge that there is increased flexibility and diversity in family structures, a typical married mother may seem to make an individual choice to stay home and care for the children because her wage rate is lower than her husband's, yet the wages women and men face in the labor market are affected by societal norms, including discrimination against women. As much as she might like to work, given her wage and the cost of childcare, she may have no individual choice to make. Also, the neoclassical model assumes that on the margin men and women can make individual decisions about time allocation, and that an individual could choose to work 33 hours per week, or 17 hours, or 51. In practice, the 'choice' of the number of hours of work is also largely socially determined by the convention of the 40-hour workweek as full-time work. And in some countries, for the most part only full-time jobs come with all the fringe benefits (health insurance, unemployment insurance, and pensions) that families need. Such benefits are difficult to get elsewhere.

Third, it is hard to see whose utility is being maximized, since nothing in economic theory suggests that two adults often have identical utility functions. Sometimes neoclassical economics seems to assume a joint utility function; at other times the possibility of altruism enters, in which, for example, the male head makes altruistic decisions within the family but acts in a completely self-interested way in the market place. In reality it is most logical and quite common that men and women in families often have different interests, preferences, and utility functions, as do parents and children (Hartmann 1981, Folbre and Hartmann 1988, Woolley 1999). Neoclassical economics does not deal well with the possibility of serious conflict in the family.

Fundamentally, feminist economists challenge this model because it justifies the status quo of the gendered division of work, labor and care within the family. It assumes that women are disadvantaged socially, have less access to the economic resources of good wages and property, and are often oppressed by individual men. It, therefore, hides what feminists are most interested in explaining. As Galbraith (1973, p. 25) has put it, neoclassical economists have made the concept of the household 'a disguise for the exercise of male authority.' Albelda (1997, p. 119) writes: 'At the most fundamental level, neoclassical economics argues that women receive lower wages than men and perform more unpaid labor than men largely out of women's and men's own rational choices and desires.'

4. *Marxist Domestic Labor Debates and the Feminist Critique*

A debate about the unique aspects of production in the home took place within the Marxist economic school of thought in the 1970s, but it was never resolved. Does housework, like wage work, produce surplus value that is appropriated both by capitalists (who can pay workers a lower wage than they otherwise could) and/or by husbands (who appropriate personal services and who have a higher standard of living than they would otherwise)? Does home production constitute a separate mode of production, a feudal mode,

directly appropriated by the lord of the house, rather than a capitalist one based on wage labor?

As Himmelweit (1999) notes in a useful review, this debate was part of a feminist attempt to use Marxist economics to identify the material base of women's oppression. Since the Marxist framework gives prime attention to production, housework (i.e., the unique work of women) was scrutinized. Thus, earlier than in neoclassical economics, the actual content of housework was studied and theorized, but just as in the neoclassical domain, the role of housework had to be discovered by women. Just as the first neoclassical 'new home economics' treated the family as an unexamined black box with a unitary utility function, early Marxist analysis of the family generally suggested that the reproduction of the working class could safely be left to itself (i.e., to the invisible labor of women in the home). Since then the concept of housework as production has been firmly established and the importance of the changing nature of the boundaries between the spheres of domestic production and market production is understood everywhere. But the feminist critique also established that what is unique in women's work in the home to reproduce the next generation is 'its caring and relational aspects' (Himmelweit 1999, p. 133).

5. *The Development of a Feminist Model*

Using aspects from three major streams of economic thought, namely the neoclassical, institutional, and Marxist approaches, Table 1 describes the attributes of the 'traditional neoclassical; model of the family (or economic decision-making more generally) in column 1. The table also shows how they are modified when institutional perspectives and feminist critiques informed by Marxist and institutional economics are added. (The table has been adapted from Nancy Folbre's (1994) much cited *Who Cares for the Kids?*) In column 2 it shows that neoclassical institutionalists modify the simple model and emphasize that rules, norms, and preferences are not all exogenously given but are at least partially endogenous. As Folbre notes, it is often easier to act based on norms and on what everyone else is doing than to calculate each decision anew. Individuals also sometimes act collectively. Folbre argues that these modifications come largely from institutional economics, which rejected both individual choice and Marxist class analysis and instead emphasized group behavior by voluntary interest groups, such as labor unions (Figart and Mutari 1999). This model notes that a transaction can involve bargaining and cooperation (based on love and affection) as well as simply buying and selling at fixed prices, particularly within the family, where there are no fixed prices at which a wife can sell her home-produced services to a new husband. Maximization of self-interest exists side by side with altruism in real life. One fruitful area of feminist research within this modified neoclassical framework is the use of game theory and models of cooperative and noncooperative bargaining to elucidate gendered outcomes in the family. (See Seiz 1999, for an overview, critique, and suggested extensions of this work.)

As the third column in Table 1 shows (the 'stylized feminist' description), not only are rules, norms, and preferences structural, but so are the assets with which the different parties begin the transaction. Using the concepts of material base and unequal ownership of assets from Marxist economics, feminist economists note that the genders do not come to the playing field as equals. Feminist economists also note that individuals do not choose all their group memberships, as the institutionalists often assume; many such memberships are assigned at birth. And, as in Marxist economics, important economic transactions are likely to involve coercion and production as well as cooperation and exchange. The relational aspect of caring is captured in the process of cooperation. Thus, in this example of a feminist model, the traditions of neoclassical, institutional, and Marxist economics are all challenged and incorporated in changed forms.

6. *Impact of the Critique from Feminist Economics*

The emergence of common elements of feminist approaches, from whichever economic framework they start, has contributed to a considerable policy impact of the feminist economic critique. The rec-

Table 1
Models of choice/action/collective identity

	Traditional neoclassical	Neoclassical institutionalist	Stylized feminist
Factors	Rules, norms, preferences (exogenously given)	Rules, norms, preferences (partially endogenous)	Assets, rules, norms, preferences (structural)
Agents	Individuals	Individuals, voluntary groups	Individuals, chosen groups, given groups
Processes	Exchange	Exchange, bargaining, coordination	Coercion, production, exchange, coordination
Sites	Markets	Markets, social institutions	Firms, states, markets, families

Source: Folbre N 1994, pp. 24, 49

ognition of the different work, interests, and utilities of wives and husbands, mothers and fathers, and the durability of unequal gender relations within the family, has informed the policies of economic development and family planning around the world. The success of these policies requires an understanding of women's roles in the family and nonmarket production. Empirical research has shown that mothers use economic resources more for the welfare of children than do fathers both in more and less advanced countries.

Hence empowering women through education and employment is thought to be the surest way to improve the health and education of children and to raise a country's productive capacity in the long run. More economically conservative structural adjustment policies, often fostered by international financial institutions to curb inflation and public deficits, seek to shrink the public sector and lessen wage demands. These policies have been shown to have disproportionately negative impacts on women, who try to make up for the lack of market wages and government aid through more intensive unpaid work. The limits of women's ability to absorb the costs of adjustment necessarily limit the success of these and other macroeconomic policies. Feminist economic analysis has led to the understanding that integrating women into paid work or into production for market exchange is not sufficient to overcome gender roles that remain unequal in the family, nor is it enough to call for more grassroots activism to allow affected women to participate in policy discussions (Bakker 1999). It calls instead for 'gender mainstreaming' or putting gender issues at the center of policy changes everywhere, including in international agencies such as the FAO, ILO, and UNDP.

National and international statistical efforts have also become informed by the feminist economic critique of the family model. For the first time, the UN Platform for Action from the Fourth World Conference on Women held in Beijing in 1995 endorses the measurement of all unpaid labor activities, including those for own consumption in the household (i.e., housework). It also endorses their inclusion as satellite accounts in the System of National Accounts for each country (Vanek 1996). In its 1995 *Human Development Report* the UNDP for the first time included measures of work-time, covering both paid and unpaid time, for women and men. Its data for 31 countries show that everywhere women work more hours than men. Beneria (1996) has called for the continuous collection and analysis of such data from all countries to better measure the effects of policy changes, such as structural adjustment policies. In the United States, the Bureau of Labor Statistics is considering a regular, periodic survey of time use (Committee on National Statistics 2000).

Feminist economists call for new measures of family poverty that challenge the assumption of income pooling and equal sharing within the family (Shaw 1999). In the United States feminist analysis of poverty, single mothers, and welfare reform calls attention to the similarity of women's income packaging strategies across class—women package income from men or other family members, their own labor market participation, and government programs (such as welfare, unemployment compensation, or social security). Just as most middle-class married women now work both inside the home for no pay and outside the home for some pay, most mothers who receive welfare or welfare-state benefits do not rely on it alone but combine income from paid work and from other family members with their government benefits. But we still know too little about who controls income within the family and how resource allocation decisions are made (Spalter-Roth et al. 1995).

7. *The Growth of Feminist Economics*

The establishment of a list serve for feminist economists (Femecon-L) in 1991; the formation of the International Association for Feminist Economics (IAFFE) in 1992; its annual conferences held in diverse locations around the world; its regular presence as the sponsor and co-sponsor of sessions at the annual meetings of the Allied Social Science Association in the United States and in many other associations in many countries; and the founding of its journal, *Feminist Economics*, first published in 1995, have all contributed to the growing visibility of feminist economists and their ideas (Shackelford 1999). As is characteristic of so much other growth in social science in the past several decades, the development of feminist economics is distinctly diverse in method and approach, distinctly international in scholars, and distinctly global in its impact.

See also: Economic development and women; Family Planning Programs: Feminist Perspectives; Family Theory: Complementarity of Economic and Social Explanations; Feminist Economics; Feminist Theory; Rationality and Feminist Thought; Rational Choice Theory: Cultural Concern

Bibliography

Albelda R 1997 *Economics and Feminism: Disturbances in the Field*. Twayne Publishers, New York

Bakker I 1999 Development policies. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK and Northampton, MA, pp. 83–95

Becker G S 1965 A theory of the allocation of time. *Economic Journal* **75**: 493–517

Becker G S 1973 A theory of marriage, Part I. *Journal of Political Economy* **81**: 813–46

Beneria L 1995 Toward a greater integration of gender in economics. *World Development* **23**(11): 1839–50
Beneria L 1996 Thou shalt not live by statistics alone, but it might help. *Feminist Economics* **2**(3): 139–42
Committee on National Statistics 2000 *Time-Use Measurement and Research: Report of a Workshop*. National Academy Press, Washington, DC
Elson D 1994 Micro, meso and macro: Gender and economic analysis in the context of policy reform. In: Bakker I (ed.) *The Strategic Silence: Gender and Economic Policy*. Zed Press and the North-South Institute, London
Figart D M, Mutari E 1999 Feminist political economy: Paradigms. In: O'Hara P A (ed.) *Encyclopedia of Political Economy*. Routledge, London and New York, pp. 335–7
Folbre N 1994 *Who Pays for the Kids? Gender and the Structures of Constraint*. Routledge, London and New York
Folbre N, Hartmann H I 1988 The rhetoric of self-interest: Ideology and gender in economic theory. In: *The Consequences of Rhetoric*. Klamer A, McCloskey D, Solow R (eds.) Cambridge University Press. Cambridge, UK, pp. 182–206
Galbraith J K 1973 *Economics and the Public Purpose*. Houghton Mifflin, Boston
Hartmann H I 1974 *Capitalism and Women's Work in the Home, 1900–1930*. Ph.D. Thesis, Yale University
Hartmann H I 1981 The family as the locus of gender, class, and political struggle: The example of housework. *Signs* **6**(3): 366–94
Himmelweit S 1999 Domestic labour. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 126–35
Kuiper E, Sap J, Feiner S, Ott N, Tzannatos Z (eds.) 1995 *Out of the Margin. Feminist Perspectives on Economics*. Routledge, London and New York
Mincer J 1962 The labor force participation of married women. In: Gregg H (ed.) *Aspects of Labor Economics*. National Bureau of Economic Research, New York
Nelson J A 1993 The study of choice or the study of provisioning? Gender and the definition of economics. In: Ferber M A, Nelson J A (eds.) *Beyond Economic Man: Feminist Theory and Economics*. University of Chicago Press, Chicago, pp. 23–36
Peterson J, Lewis M (eds.) 1999 *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA
Shackelford J 1999 International Association for Feminist Economics (IAFFE). In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 486–9
Seiz J A 1999 Game theory and bargaining models. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 379–90
Shaw L B 1999 Poverty, measurement and analysis of. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 634–9
Spalter-Roth R, Burr B, Hartmann H I, Shaw L 1995 *Welfare That Works: The Working Lives of AFDC Recipients*. Institute for Women's Policy Research, Washington, DC
UNDP (United Nations Development Program) 1995 *Human Development Report*. Oxford University Press, Oxford, UK
Vanek J 1996 Generate and disseminate! The UN platform for action. *Feminist Economics* **2**(3): 123–4
Woolley F 1999 Economics of family. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 328–36

H. Hartmann

Family Theory: Role of Changing Values

1. Introducing the Issue

This article is about the significance of changing values in understanding demographic behavior. Questioning the significance of values may seem odd since in everyday life few people would doubt that values are important. Among social scientists, however, the issue of values still evokes feelings of ambivalence and controversy. Three issues that we think are pertinent to a sociological understanding of the significance of values in family research will be broached. First, the significance of values as such is dealt with, including a discussion of the concept of values. Second, we argue that if values theory is to play a role in family research, it should be able to complement other types of explanations as well as to have an intrinsic quality to explain demographic changes both at the individual and the societal level. Finally, two types of values theories, that is socialization and role theory are merged into one framework that highlights the dynamics of the values-behavior relationship.

A note on terminology: although logically one should write about 'value theory' (without the -s) the plural form 'values' is preferred because 'value' and 'values' are distinct concepts similar to 'medium' and 'media.' In writing about 'values' the totality of values is referred to as ideational factor, and not to 'value' in the sense of the worth of an object.

2. The Significance of Values

In general three types of attitudes toward the concept of values can be found among social scientists: denial, trivialization, and emphasis. Probably the most radical type of denial of the 'significance' of values is to refuse to give any role to 'values' altogether. In most economic theories for instance, 'value' (without the 's) enters the picture as a synonym of 'price.' 'Values' as a cultural factor are rarely considered. This is especially true for neo-classic economists (cf. G. Becker 1981). If ideational factors are allowed in economic models, it is often in terms of tastes, preferences, or needs, and not in terms of values in a sociological sense of the word (cf. infra). But even in sociology some approaches (e.g., structural theories) are notorious for denying the significance of culture (Thompson et al. 1990). Among approaches that explicitly give credit to

a cultural argument, several also raise doubt whether the concept of values is needed in the study of culture and social change (Adler 1956, Blake and Davis 1968; see also J. F. Scott's phenomenological approach (Cancian 1975). They question the usefulness of the concept of values in causal explanations because values can only be inferred from behavior itself or from specific norms and goals. Hence, in this view a direct explanation of human behavior by reference to values is a sociological tautology. Although these critics differ in opinion regarding the type of cultural factors that are significant, they all deny the significance of values.

Trivialization of values may manifest itself in several ways. Values may be trivialized by 'explaining it away,' that is by arguing that values are merely spuriously related to behavior. Such a view is congruent with the former critics regarding the tautological nature of any values explanation of behavior. Values may also be trivialized by making them a 'constant' or 'exogenous' factor in explaining behavior. Within rational choice theory the exogenous character of values is explicit: assuming that there is a rational actor is implying that 'rationality' is a natural characteristic rather than a cultural construct. Defining values as an exogenous factor or arguing that values are merely spuriously related to behavior is minimizing their role. However, maximizing their importance can also lead to trivialization. This is illustrated in the work of Parsons (1937) who—perhaps ironically enough—protested strongly against the aforementioned 'natural science' kind of attitude toward values. Parsons' own functionalist approach to values, however, can also be criticized for having an intrinsically trivial nature (Spates 1983). Part of the trivialization is due to the teleological conceptualization of values within the Parsonian tradition: 'a value is a conception…of the desirable which influences the selection from available modes, means and ends of action' (C. Kluckhohn, in: Parsons and Shils 1951, p. 395). Values determine behavior by definition. By consequence, the question whether they do so is irrelevant. Spates (1983) raises a second caveat, that is that Parsons' values theory faces the problem of deductive imposition by arguing that values can only be deduced by a careful theorist. Hence, observable reality is forced into accord with a preconceived model. Although theoretically values are emphasized, empirically they are trivialized. Dropping the element of action probably leads to a less trivial conceptualization.

The empirical tradition of studying values by definition emphasizes the significance of values. Conceptualization of values, however, has not proved to be its strongest point. Many researchers do not hesitate to use the concept of values implicitly in their work, assuming a mutually intuitive understanding of the concept. Others have claimed that the empirical tradition merely added to the conceptual confusion rather than anything else (Lautmann 1971). Although there is a grain of truth in these claims, it is felt that they tend to exaggerate divergence. Within the ideational approaches of the concept of values, there is a tendency to fit in with two traditions. ('Fitting in' does not mean that scholars fully subscribe to the conceptualization within a tradition, but rather that they share its point of departure.) The Parsonian tradition has already been discussed in the preceding section. Researchers working within this framework seem to agree with Kluckhohn's idea that values are 'conceptions of the desirable.' The second tradition was initialized by the social psychologist Rokeach (1968, 1973), who explicitly opposed the Parsonian concept of values. Instead he argued that 'a value is an enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence' (1973, p. 5). The major advantage of this definition is that it tries to avoid letting issues open to empirical questioning—such as reference to action—constitute elements of that definition. Presumably, this is one of the reasons why the two values theories that will be discussed in the last section adopt a similar concept of values. Nevertheless, one could interpret Rokeach's concept of values as being a conservative force since they are defined as enduring beliefs. Although socialization theories (cf. infra) tend to agree with the latter statement, Rokeach himself has pointed out that complete stability is not at issue. To avoid misinterpretation it is suggested the adjective 'enduring' is dropped from the nominal definition.

In spite of these arguments several researchers still reject Rokeach's definitions and prefer Kluckhohn's conceptualization (Mohler 1987 and Miceli and Castelfranchi 1989). This merely illustrates that both traditions are competing.

It is obviously important to distinguish 'values' from other ideational factors (Rokeach 1973, Cancian 1975, Fishbein and Ajzen 1975, and Miceli and Castelfranchi 1989). In family research, three such factors seem important, namely, norms, attitudes, and intentions. Especially the latter two concepts have received much more attention than values in studies of demographic behavior. Since there is little agreement on the meaning of these concepts a view is presented on the major differences. 'Values' and 'attitudes' differ because attitudes are more 'object' and 'situation specific' and because values are of a more general nature. Intentions are even more object-related. Consequently, an 'organization of attitudes' transcends the specificity of each attitude and may be interpreted as a 'values orientation.' Analyses of the European and World Values Studies often adopt such a strategy (Ester et al. 1993) by collapsing scores on different attitudes into one or several values dimensions. Hence, attitudes and values are strongly related. Intentions, on the other hand, are more closely related to behavior. For that reason measured intentions are often modeled as dependent variables as an alternative to indicators

of behavior in values research (Inglehart 1977, 1990). At least Fishbein and Ajzen's theory of reasoned action (1975) demurs at such a point of view. Instead they argue that the best predictor of behavior is the intention to do so. A second difference between values on the one hand and attitudes or intentions on the other is that the latter are individual predispositions, whereas values may also be characteristics of depersonalized entities, as in 'the values of the working class.' In such a connection values bridge the gap between attitudes and norms. Norms prevail at the society or macro level as products of cultural inheritance. They refer to prescriptions of how to behave. In fact they include sets of (normative) rules. Internalization of these rules may define values orientations.

3. *The Role of Values in Family Theory and Research*

The values argument belongs in a broader context concerned with the role of culture in social demography. As far as family research is concerned, one cannot 'neglect' noncultural types of explanation. One should not ignore the impact of G. Becker's New Home Economics (1981) and R. Easterlin's Social Deprivation Theory (1976, 1980) in explanations of demographic change in recent decades. Pollak and Watkins (1993) recently summarized the various perspectives on how cultural approaches relate to economic perspectives and argued that cultural and economic approaches are compatible when they are consistent or equivalent to one another or when cultural explanations attempt to specify the scope of economic rational choice explanations. Preston (1986) and Lesthaeghe and Surkyn (1998) are referred to in particular on this issue. Similarly, Lesthaeghe and Moors (1994) and Lesthaeghe (1998) have demonstrated the usefulness of a frame of reference that includes the complementarity of theoretical perspectives in studying family formation (see *Family Theory: Complementarity of Economic and Social Explanations*). They argue that our understanding of demographic transitions may improve if we avoid rigidly adopting a single grand theory or aiming at enhancing the supremacy of one particular paradigm. Treating theories as distinct but not mutually exclusive views gives rise to a multicausal theoretical framework that can be used to 'explain' historical and geographical variations (Preston 1986, Lesthaeghe 1995) as well as individual differences (Lesthaeghe and Meekers 1986, Lesthaeghe and Moors 1994). Irrespective of these notions, Oppenheimer (1994) remains skeptical of researchers who rely exclusively on the idea that values influence behavior. In her opinion, they face the problem that there is little empirical research that actually tests the presumed causal effect. However, the little evidence that does exist supports the theory of ideational change. Using panel data, Moors (1997)—on values—and Thornton and coworkers (1992, see also Barber 1999)—on attitudes—clearly demonstrated that ideational factors influence demographic behavior autonomously. Hence, the ability of values theory to complement other types of explanations, its intrinsic quality in explaining demographic changes, and its empirical merits lead us to conclude that values have a role to play.

4. *Values Theory?*

In empirical research on values two theories developed in the 1970s drew much attention, namely, Inglehart's socialization theory (1977, 1990) and Kohn's role theory (1977, Kohn and Slomczynski 1993). They produced an independent body of research and stirred controversy, but have gained prominence in later decades. In demography Inglehart's theory has been used by the researchers focusing on values (cf. infra) because the theory is founded on the notion of ideational change. There have been some modifications of Inglehart's argument, but its core has remained intact. It is founded on two basic premises. First, Inglehart argues that an individual's values priorities reflect his or her socio-economic environment, people attach relatively more importance to relatively scarce objects. This scarcity argument is complemented by a socialization argument that stresses the importance of experiences in the so-called 'formative years.' Inglehart assumes that—in reaching adulthood—values tend to crystallize in the personality. By consequence he 'predicts' a generational pattern of change in values orientations toward increasing preference for autonomy and consequent behavior. The idea that a process of social metabolism (i.e., the succession of generations with distinctive profiles) may generate social change was nothing new to demographers familiar with the work of N. Ryder (1965). By consequence, it has been a small step to link a generational pattern of values change with changes in behavior, for example, rising levels of cohabitation, independent living and divorce, an increase in single parenthood, etc. Inglehart's values theory primarily emphasizes that values influence choices people make. The possibility of reverse causation is not explicitly discussed.

Whereas Inglehart focuses on historical conditions during the formative years, Kohn claims that values are directly influenced by 'here-and-now experiences' at the workplace. In Kohn's theory, values play a mediating role in relating social structure to behavior. As such he primarily focuses on 'contemporary conditions' rather than historical circumstances. The key argument of the Kohn thesis is that the normative requirements of a job generate values orientations which are appropriate to the actual circumstances of

the job. Hence, the relationship between social class and values can be explained by referring to occupational conditions, that is the level of occupational self-direction. Generalizing this point of view one could argue that people learn to value characteristics that are appropriate to their conditions of life. In contrast to Inglehart, Kohn highlights the adaptation of values to changing conditions. At first glance, the theories yield contradictory propositions. On consideration, however, these theories can be reconciled because they bring together different mechanisms that generate values orientations, that is, values may be socialized (Inglehart) as well as learned by taking up social roles (Kohn). Furthermore, even their difference in locus regarding the issue of values stability and values adaptation need not imply contradiction. To Inglehart values influence behavior because values are relatively enduring characteristics of individuals. Although Rokeach (1973) holds a highly similar position, he also demonstrated within experimental designs that values may change due to strong 'incentives,' for example, dramatic or intense experiences (Ball-Rokeach et al. 1984). Kohn's theory suggests that—given the centrality of work in the life of people—the everyday experience in the workplace may induce such change. Consequently, Inglehart's notion of relatively enduring values orientations after the formative years is compatible with Kohn's argument about the significance of temporal occupational conditions. After all, important changes in the life-course may re-affirm values rather than altering them. Young people who value traditional family life may decide to marry for that reason, and may find affirmation of their opinion by the very experience of marriage itself. Hence, even stability forms part of a dynamic process. As a consequence, the key question is no longer whether values change after the formative years, but when (e.g., under what conditions) they do so. By the same token the dynamics of the values-behavior relationship becomes manifest. Values influence the choices people make and the consequences of these choices—in turn—re-affirm or alter the values these people hold. It is important to reconcile both these theoretical perspectives in demography since there is little doubt that changes in family life can be intense experiences. As the research of Thornton et al. (1992, see also Axinn and Thornton 1993) has demonstrated, both historical and contemporary experiences affect demographic transitions. Hence, a life-course approach focusing on the sequential and recursive relationship of values and demographic transitions is most appropriate.

See also: Families and Households, Behavioral Demography of; Families and Households, Formal Demography of; Family and Gender; Family and Kinship, History of; Family, Anthropology of; Family as Institution; Family Processes; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Intergenerational Relations; Family Theory: Economics of Childbearing; Family Theory: Economics of Marriage and Divorce; Fertility: Institutional and Political Approaches; Fertility: Political and Political–Economic Perspectives; Values, Sociology of

Bibliography

Adler F 1956 The value concept in sociology. *American Journal of Sociology* **65**: 272–9

Axinn W G, Thornton A 1993 Mothers, children and cohabitation: The intergenerational effect of attitudes and behavior. *American Sociological Review* **58**: 233–46

Ball-Rokeach S J, Rokeach M, Grube J W 1984 *The Great American Values Test. Influencing Behavior and Belief Through Television*. Free Press, New York

Barber J S 1999 Ideational influences on the transition to Parenthood: Attitudes toward childbearing and competing alternatives. Paper presented at the P.A.A. annual meeting, New York, March 1999

Becker G 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Blake J, Davis K 1968 Norms, values and sanctions. In: Faris R E L (ed.) *Handbook of Modern Sociology*. Rand McNally and Company, Chicago, pp. 456–84

Cancian F M 1975 *What are Norms? A Study of Beliefs and Action in a Maya Community*. Cambridge University Press, Cambridge, UK

Easterlin R 1976 The conflict between aspirations and resources. *Population and Development Review* **2**: 417–26

Easterlin R 1980 *Birth and Fortune. The Impact of Numbers on Personal Welfare*. Basic Books, New York (2nd edn. 1987)

Ester P, Halman L, de Moor R 1993 *The Individualizing Society. Value Change in Europe and North America*. Tilburg University Press, Tilburg, The Netherlands

Fishbein M, Ajzen I 1975 *Belief, Attitude, Intention and Behavior. An Introduction to Theory and Research*. Addison-Wesley, Reading, MA

Inglehart R 1977 *The Silent Revolution*. Princeton University Press, Princeton, NJ

Inglehart R 1990 *Culture Shift in Advanced Industrial Society*. Princeton University Press, Princeton, NJ

Kohn M L 1977 (1st edn. 1969) *Class and Conformity. A Study in Values*. The University of Chicago Press, Chicago

Kohn M L, Slomczynski K M (eds.) 1993 *Social Structure and Self-Direction: A Comparative Analysis of the United States and Poland*. Blackwell, Oxford, UK

Lautmann R 1971 *Wert und Norm: Begriffsnanalyse für die Soziologie*. Westdeutscher Verlag. Opladen, Germany

Lesthaeghe R 1995 The second demographic transition: An interpretation. In: Mason K O, Jensen A M (eds.) *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK, pp. 17–62

Lesthaeghe R 1998 On theory development: Applications to the study of family formation. *Population and Development Review* **24**(1): 1–14

Lesthaeghe R, Meekers D 1986 Value changes and the dimensions of familism in the European Community. *European Journal of Population* **2**: 225–68

Lesthaeghe R, Moors G 1994 Living arrangements, socioeconomic position and values among young adults: A pattern description for France, Germany, Belgium and the Netherlands 1990. In: van den Brekel J C, Deven F (eds.) *Population and Family in the Low Countries 1994*. Kluwer, Amsterdam

Lesthaeghe R, Surkyn J 1988 Cultural dynamics and economic theories of fertility change. *Population and Development Review* **14**: 1–46
Miceli M, Castelfranchi C 1989 A cognitive approach to values. *Journal of the Theory of Social Behaviour* **19**: 169–93
Mohler P P 1987 Cycles of value change. *European Journal of Political Research* **15**: 155–65
Moors G 1997 The dynamics of values-based selection and values adaptation. With an application to the process of family formation. Ph.D. thesis Vrije Universiteit, Brussels, Belgium
Oppenheimer V K 1994 Women's rising employment and the future of the family in industrial societies. *Population and Development Review* **20**: 293–342
Parsons T 1968 (1st edn. 1937) *The Structure of Social Action*. Free Press, New York
Parsons T, Shils E A (eds.) 1951 *Toward a General Theory of Action*. Harper, New York
Pollak R A, Watkins S C 1993 Cultural and economic approaches to fertility: Proper marriage or mésalliance? *Population and Development Review* **19**: 467–95
Preston S H 1986 Changing values and falling birth rates. *Population and Development Review* **12** (suppl.): 176–95
Rokeach M 1973 *The Nature of Human Values*. Free Press, New York
Rokeach M 1976 (1st edn. 1968) *Beliefs, Attitudes and Values. A Theory of Organization and Change*. Jossey-Bass, San Francisco
Ryder N B 1965 The cohort as a concept in the study of social change. *American Sociological Review* **30**: 843–61
Spates J L 1983 The sociology of values. *Annual Review of Sociology* **9**: 27–49
Thompson M, Ellis R, Wildavsky A 1990 *Cultural Theory*. Westview Press, Boulder, Co
Thornton A 1992 The influence of the parental family on the attitudes and behavior of children. In: South S J, Tolnay S E (eds.) *The Changing American Family: Sociological and Demographic Perspectives*. Westview Press, Boulder, CO, pp. 247–66
Thornton A, Axinn W G, Hill D H 1992 Reciprocal effects of religiosity, cohabitation and marriage. *American Journal of Sociology* **98**: 628–51

G. Moors

Family Therapy, Clinical Psychology of

The origins of family therapy can be traced to the 1950s, although its historical roots go back further. Several largely independent movements that deal with relationships among family members contributed to the development of family therapy. The oldest of these is the marriage counseling movement while, later, a separate movement toward conjoint marital therapy developed within the psychiatric community. Family therapy emerged from a union of several seminal thinkers from a variety of backgrounds. The main factors influencing its development were the expansion of psychoanalytic psychotherapy to include a broader range of emotional disorders and the application of psychoanalytic principles to the entire family; the investigation of the role of the family in the development of schizophrenia and other mental and physical disorders; and the growth of the child guidance movement.

1. Historical Perspective of Family Therapy

1.1 Evolution of the Field

Social work, psychiatry, and psychoanalysis in both Great Britain and the United States all made enormous contributions to the development of family therapy. In her influential book *Social Diagnosis* (1917), Mary Richmond was one of the first proponents of studying the whole family and its needs. At the same time, the dominance of the psychoanalytic model in the mental health field focused on the mother–child dyad as the source of psychopathology. This model, with its linear–causal attributions, blamed maternal deficiencies for any childhood disturbance. Adler became the first of Freud's pupils to openly challenge such an explanation of human behavior. He offered an alternative and more socially rooted theory of psychodynamics, organized child guidance clinics in Vienna, and counseled children, parents, and teachers.

By the 1930s, a strong community of analysts had developed in the United States. Among the most influential to emerge was Harry Stack Sullivan. He focused on interpersonal relations in his research into family dynamics of schizophrenia. By the time family therapy pioneers began to experiment in the 1950s, there was a well-established bias toward social explanations among American analysts. This trend was underscored by a paradigm shift that occurred with the development of *General Systems Theory* (GST). GST implies that family therapy is grounded in a set of assumptions about the interchange between individual, family and social processes, which operate according to certain principles that apply to all human systems (Bateson 1971). Family members are interrelated such that every individual affects each other, while the group as a whole affects the first member in a circular chain of influence, making every action in a sequence also a reaction.

1.2 Pioneers of the Family Therapy Movement

John Bowlby's work at the Tavistock Child Guidance Clinic in London during the late 1940s exemplifies the transition from an individual to a family approach. These beginnings of family therapy were tentative in that Bowlby felt he had to 'reassure' readers that family interviews were 'rarely employed.' However, family interviews prior to the initiation of treatment

became routine, and served as a major stimulus to the development of this new approach in the United States. By the end of the 1950s family therapy emerged as a connected movement whose members exchanged correspondence and visits, and began to cite one another.

Nathan Ackerman came to family therapy from the field of child psychiatry. His involvement in a study of mental health problems among the unemployed, along with other experiences, convinced him that both the environment and the psyche could influence emotional problems. Although family interviews began as an experiment for Bowlby, Ackerman saw this as the major form of treatment in child guidance clinics. He saw the family as the unit of diagnosis and treatment, and began sending his staff on home visits to study families.

In 1958 Ackerman published the first book-length treatment of family relationships, *The Psychodynamics of Family Life*. By 1961 family therapists met to prepare the way for the first joint handbook and to found a common journal, *Family Process*. In addition, John Bell's respected monograph, *Family Group Therapy* (1961), constituted one of the founding documents of the profession. Carl Whitaker, who is noted as the most brazen of the founders, is generally credited with having called the first formal meeting in the family therapy movement in 1953. As family therapy developed, he played an integral role, as evident from his membership on the first board of editors for *Family Process*.

Among the pioneers, Lyman Wynne was particularly well prepared by formal training to become a family researcher and therapist. After his medical training in 1948, he embarked on doctoral studies in the Department of Social Relations. As a graduate student, he came into contact with many of the leaders in the field of sociology, social psychology, and social anthropology. Initially, he saw whole families only when individual treatment or joint interviews with the mother and patient were not effective. Gradually, however, he worked out his own theory of the family structure of schizophrenic patients. Murray Bowen, a psychiatrist who specialized in schizophrenia, felt that the parents—in particular, mothers, should be required to live in the hospital with their disturbed child. In 1954 he joined Lyman Wynne at the National Institute of Mental Health, where he established a program for treating families together.

The Palo Alto Group has probably one of the strongest claims to have been at the forefront of the development of family therapy. This group consisted of five richly individual persons, viz. Gregory Bateson, Jay Haley, John Weakland, Don Jackson and Virginia Satir. Bateson, an anthropologist and philosopher by training, and considered by many as the founder of family therapy, wrote important works such as *Steps to an Ecology of Mind* (1971) and *Mind and Nature* (1979). Bateson and his group hypothesized that family stability is achieved by a feedback system, i.e., whenever the family system is disturbed, it will quickly move toward reestablishing balance/homeostasis. In 1956 this group generated one of the most discussed papers in the history of psychiatry. In '*Toward a Theory of Schizophrenia*,' they introduced the concept of double-bind as the crucial familial determinant of schizophrenia in children. Jackson published a seminal paper on conjoint family therapy, arguing that it was more effective than seeing family members individually (Jackson 1959). At the same time he founded the Mental Research Institute (MRI), and invited Satir to join him. The MRI, compared to the original Bateson group, was more focused on the treatment of families as opposed to research.

The influence of behavioral family therapy can also not be overlooked as it was launched as a specialist area in the late 1960s with the seminal work of Stuart (1969) and Liberman (1970). Behavioral approaches to family therapy can be characterized as goal oriented, a general reliance on principles of learning, directedness in interventions, and preference for short-term, problem-solving involvement.

While the pioneers dominated the 1960s and 1970s, they also saw new innovations. One innovator who has taken his place among the founders as one of the most influential of all family therapists is Salvador Minuchin. In the late 1960s he became the director of the Philadelphia Child Guidance Clinic and convinced Jay Haley, among others, to join him. The Minuchin–Haley team made substantial clinical and methodological contributions. 'Structural therapy' emerged from their interaction, and will be discussed below. In addition to carrying over the communications and systems elements developed at MRI, Minuchin and Haley gave added emphasis to the realignment of counterproductive family coalitions, and tied their theory to a family developmental framework.

2. *The First Family Therapy Models*

The field of family therapy encompasses many modalities, each with distinctly different ways of conceptualizing and treating families.

2.1 *Psychodynamically Oriented Family Therapy*

Nathan Ackerman was among the first to integrate psychoanalytic principles with systems theory. He viewed individual functioning as a reflection of constant interactions between the individual, the family, and society. For Ackerman, understanding a person requires an understanding of their intrapsychic processes and interpersonal interactions. Achievement of this goal involves helping family members adopt social roles that are flexible so as to permit role complementarity.

Murray Bowen's *Extended Family Systems Therapy* expands GST beyond the nuclear family. Behavioral disorders are the result of a multigenerational transmission process in which progressively lower levels of differentiation, i.e., independence of self from others, occur as the family's lack of differentiation is transmitted from one generation to the next. The primary goal of therapy is to increase the differentiation of family members. Bowen regards family relationships as involving 'triangles,' and therapy typically includes only two family members (usually the spouses) while the therapist becomes the third member in a 'therapeutic triangle.' As long as the therapist remains objective and neutral (does not become emotionally triangled), his/her presence helps family members resolve the fusion between them and achieve higher levels of differentiation.

Contemporary proponents of psychodynamically oriented family therapy include Christopher Dare, John Byng-Hall, and James Framo.

2.2 *Problem-solving Approaches to Family Therapy*

2.2.1 Communication interaction approach. This approach to family therapy grew primarily out of research conducted by Gregory Bateson and his colleagues. They recognized the importance of communication in family and individual dysfunction (e.g., the role of 'double-bind' communication in the development of schizophrenia), and developed a school of family therapy based on the principles of communication. This approach incorporates principles derived from information theory, cybernetics, and GST. Family therapists adopting this approach accept a circular model of causality that regards a symptom as both a cause and an effect of dysfunctional communication patterns. Although therapists do not deny that individual intrapsychic factors underlie family problems, the observable interactions between family members are viewed as appropriate targets of therapy. The primary goal of therapy is to alter interactional patterns that maintain the presenting symptom(s).

2.2.2 Structural family therapy. Salvador Minuchin (1974) originally developed structural family therapy through his work with poor, inner-city populations. In his seminal work, *Families and Family Therapy*, he argued that symptoms are the result of family structural imbalances. Structural family therapy focuses on concepts that describe space configurations: for example, malfunctioning hierarchy and boundaries, and maladaptive reactions to developmental and environmental changes. Structuralists use the key notion of 'complimentarity' to denote a fit among matching parts of a whole. Symptomatic behavior is seen as one such part that fits into a dysfunctional whole/organization. For instance, an adolescent's anorexia nervosa may be related to a mutual invasion of the patient's and their parents' territories. In a dysfunctional family, development has been replaced by inaction. When a family is stuck in a rigid arrangement, they cannot solve their problems and continue to grow.

Breaking away from such an organizational impasse requires the mobilization of resources that the family already latently possesses and which are often apparent in a different context. Systemic change, in the structural view, equals an increase in the complexity of the structure—an increment in the availability of alternative ways of transacting. The function of the therapist is to create a context for the family to experience these alternative patterns as accessible, possible, and necessary. What the structural therapist is trying to build through his or her restructuring efforts is more important than what he or she is trying to uncover. The therapist does not emphasize individual change or the prescription of specific solutions. Instead, he/she attempts to reorganize the family structure, or increase its flexibility. Through parental leadership and the creation of clear subsystems and boundaries, more adaptive coping is promoted.

2.2.3 Strategic or systemic family therapy. In addition to the strategic and systems models of the Palo Alto group, other innovative early approaches were Jay Haley (1976) and Cloe Madanas' (1981) problem-solving approach, and the Milan team approach (Selvini-Palazzoli et al. 1978). Haley and Madanas are now most associated with strategic family therapy, which combines the communication/interaction approach with Minuchin's structural family therapy. Haley coined the term 'strategic therapy' when describing the techniques used by the psychiatrist Milton Erickson. Like communication/interaction therapists, strategic therapists emphasize the role of communication in maladaptive behavior, but their focus is on how communication is used to increase one party's control in a relationship. Whereas struggles for control are inherent in any relationship, they become pathological when one or both parties denies his/her intent to control the other person, and when this results in symptomatic behavior. Symptoms are thus seen as interpersonal phenomena rather than intrapsychic events. The focus in strategic therapy is not on understanding the source of problems, but on alleviating current symptoms through altering family organization, hierarchy, and generational boundaries.

The Milan team consisted of Mara Selvini-Palazzoli, Luigi Boscolo, Gianfranco Cecchin, and Giuiana Prata. Their book, *Paradox and Counterparadox* (1978), presented the conceptual framework

and methodology of their systemic family therapy. Stimulated by Bateson's writings (Bateson 1971), the Milan group radically revised their analysis of systems. Until this point, they had seen the therapeutic process as one wherein the therapy team observed the family system in order to understand it. Working as a team, utilizing a one-way mirror, circular interviewing, and employing systemic analysis to describe the family's interactional map, interventions were designed to ensure that the therapist remains separate from the family. This analysis and approach was labeled 'first-order cybernetics.'

Through revisiting Bateson's work, the Milan clinicians began to take into account the context of observation itself, and hence the observational system as well. In first-order cybernetics a distinction is arbitrarily drawn between the family and the therapist, as though they were two distinct systems. The Milan clinicians began to see the family in the context of the therapist's interaction with them. This new awareness included a second level of observation that parallels the first. This second level of observation is that of the system created by the interaction between the therapist and family within the therapeutic encounter. This second level observation was termed 'second-order cybernetics.' Instead of only studying the interactional patterns of family members, the feedback loops that develop between interviewer and interviewees were also considered.

3. *Recent Family Therapies*

The 1980s ushered in a new stage in the development of family therapy, in part inspired by the Milan group. First, an era of skepticism and a reexamination of the old, under the banner of postmodernism, emerged. Postmodern psychologies concern themselves with how people make meaning in their lives; how they construct reality. Constructivism (how an individual creates his/her own reality) and social constructionism (how social interaction creates reality) took hold of family therapy in the early 1980s and have since exerted a powerful impact on the field (see *Constructivist Psychotherapies*). Second, a movement to deconstruct established knowledge expanded in the form of feminist critique. This exposed the patriarchal bias in the assumptions and practices of many family therapies (Hare-Mustin 1986). Third, postmodern family therapy became more collaborative, which led to a cross-fertilization among therapists, as well as between therapists and families. Fourth, the mechanistic aspects of systems thinking have become less popular. Instead the focus has shifted to the stories that govern a family's life rather than its structure or interactional sequences of behavior. Fifth, postmodernists have become less wedded to universal patterns, and emphasize diversity and pluralism.

The result of this review has been the realization that a favored therapeutic approach may only be one of many useful approaches. The field of family therapy has also shifted its view of the family as the source of symptoms to a biopsychosocial view. From this perspective, assessment and intervention attend to the impact of biologically based conditions as well as larger cultural influences, such as race, class, gender, ethnicity, religion, and sexual orientation on the family. Family therapy today is characterized as having a resource perspective as opposed to concentrating on deficits within the family. It has also shifted from a model of the therapist as expert to one of respectful collaboration with family members. Assessment and intervention have shifted their emphasis on problems and how they are maintained to the attainment of solutions (Walsh 1993).

3.1 *Postmodern Approaches*

Growing out of the strategic/systemic models, the post-Milan or 'second-order' family therapists invented techniques to engage families in 'conversation' about their problems; these are known as narrative (White and Epstein 1990) or conversational approaches (Anderson and Goolishian 1988). These therapies are based in constructivist and social constructionist views of reality. Normality is viewed as socially constructed. In narrative therapy, for instance, the therapeutic conversation and process of 'restorying' a problematic experience is emphasized. The goal is to envision optimistic life stories that are more empowering and satisfying. Another constructivist approach that has gained popularity is Steve de Shazer's solution-focused therapy (Berg and de Shazer 1993). This is similar to the narrative approach in that it uses leading questions to separate clients from their problem-focused stories. The therapist is affirmative in that he/she deconstructs the problem story by having clients focus on exceptions, i.e., times when their problem did not occur. No judgments are made about the solutions a family may discover. Goals in these therapies are achieved by externalizing problems, focusing on future-oriented potential, and the search for exceptions and unique outcomes.

3.2 *Psychoeducational Approaches*

The psychoeducational model was developed for family intervention with schizophrenia and other persistent mental and physical illnesses (Falloon et al. 1984). In schizophrenia, for instance, family management of the illness has been demonstrated as effective in reducing levels of 'Expressed Emotion' and lowering rates of relapse. Family therapy, combined with drug maintenance and social skills training, has produced the best results by reducing relapse rates and improving functioning. The process of change is

facilitated through information and management guidelines, social support, and a respectful collaboration between therapist and family.

4. Research in Family Therapy

The roots of family therapy are in research, as many founders came to this field through studying interactional patterns associated with problem families. Initially, no distinctions were drawn between researchers and therapists, as research had direct clinical relevance, and hypotheses were developed and tested in clinical settings. The growth in research depended disproportionately on intuitive appeal as opposed to evidence. This research was mainly qualitative and impressionistic in nature.

Since the early investigations, research in family therapy has grown dramatically and undergone at least two major transformations (Sprenkle and Moon 1996). First, the field has evolved from an impressionistic to a quantitative emphasis. This shift forced family therapists to operationalize less well-defined concepts, and to develop more reliable and valid measures. The early decades were characterized by a strong push for outcome research and, although the scientific rigor of some of the earlier studies may be questionable, it did give credence to an emerging discipline. Second and more recent, a transformation that is characterized by a shift from quantitative methods to the incorporation of qualitative methods can be discerned. This transformation was facilitated by the advent of the new epistemology constructivism, or social constructionism. Linear causality, assumed by traditional quantitative research, has been replaced by more systemic thinking. Proponents of the second transformation argued that family therapy concepts were operationalized before being truly understood and, consequently, seemed removed from clinical reality. Therefore, the second transformation emphasizes the 'context of discovery,' i.e., the intricate subtleties of the therapeutic process, as opposed to the 'context of verification.' The integration of outcome and process research that has characterized the first and second transformations, respectively, has gained the support of family therapy researchers. The Maudsley group in London exemplifies this integration with their family treatment studies for eating disorders (Eisler et al. 1997).

This 'third transformation' in family therapy research is called pluralism. A pluralistic research community values and appreciates different inquiry paradigms, methodologies and designs, and encourages an atmosphere of understanding, openness, and tolerance. Methodological pluralism encourages research from a wide variety of perspectives, and increases the tools at our disposal for understanding family therapy processes and outcomes (Gurman et al. 1986).

5. Future Directions

Family therapy is generally effective, and family treatments that focus on specific problem areas have been shown to have positive and enduring effects. If treated systemically, problems with clear family relations implications show a greater likelihood of a good outcome. The growing *diversity* of families poses new challenges for the family therapy field. Culture, class, race, ethnicity, sexual orientation, and religion contribute greatly to family beliefs and practices (Falicov 1995). A broader definition of normalcy has to be considered if family therapy is to remain responsive to this pluralism. Recent shifts from action to narrative, and from challenging to collaborating, have opened new possibilities for doing family therapy. Family therapy has developed greater humility about the influences of existing theoretical models, diminished the need to prescribe to and change people, and increased trust in a family's own resources.

See also: Child Abuse; Child and Adolescent Psychiatry, Principles of; Counseling Psychology; Divorce and Children's Social Development; Family and Psychiatric Illness/Family Therapy; Family as Institution; Family Health; Family Processes; Family Theory: Role of Changing Values; Marital Interaction: Effects on Child Development; Parenting: Attitudes and Beliefs; Parents and Teachers as Partners in Education

Bibliography

Anderson H, Goolishian H A 1988 Human systems as linguistic systems: Preliminary and evolving ideas about the implications for clinical theory. *Family Process* **27**: 371–93

Bateson G 1971 *Steps to an Ecology of Mind*. Ballatine, New York

Berg I K, de Shazer S 1993 Making numbers talk: Language in therapy. In: Friedman S (ed.) *The New Language of Change*. Guilford Press, New York

Eisler I, Dare C, Russell G F M, Szmukler G I, Le Grange D, Dodge E 1997 Family and individual therapy in anorexia nervosa: A five-year follow-up. *Archives of General Psychiatry* **54**: 1025–30

Falicov C J 1995 Training to think culturally: A multidimensional comparative framework. *Family Process* **34**: 373–88

Fallon I R H, Boyd J L, McGill C W 1984 *Family Care of Schizophrenia: A Problem-Solving Approach to the Treatment of Mental Illness*. Guilford Press, New York

Gurman A S, Kniskern D P, Pinsof W M 1986 Research on the process and outcome of marital and family therapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York

Haley J 1976 *Problem-Solving Therapy*, 1st edn. Jossey-Bass, San Francisco, USA

Hare-Mustin R T 1986 The problem of gender in family therapy theory. *Family Process* **26**: 15–27

Jackson D D 1959 Family interaction, family homeostasis, and some implications for conjoint family therapy. In: Maserman

J (ed.) *Individual and Family Dynamics*. Grune and Stratton, New York
Liberman R 1970 Behavioral approaches in family and couple therapy. *Journal of Orthopsychiatry* **40**: 106
Madanas C 1981 *Strategic Family Therapy*. Jossey-Bass, San Francisco
Minuchin S 1974 *Families and Family Therapy*. Harvard University Press, Cambridge, MA
Selvini-Palazzoli M, Boscolo L, Cecchin G, Prata G 1978 *Paradox and Counterparadox*. Aronson, New York
Sprenkle D H, Moon S M 1996 Toward pluralism in family therapy research. In: Sprenkle D H, Moon S M (eds.) *Research Methods in Family Therapy*. Guilford Press, New York
Walsh F 1993 Conceptualization of normal family processes. In: Walsh F (ed.) *Normal Family Processes*, 2nd edn. Guilford Press, New York
White M, Epstein D 1990 *Narrative Means to Therapeutic Ends*, 1st edn. Norton, New York

D. le Grange

Farmland Preservation

One might think that the escalation of the world's population and widespread famine would supply an ample reason for the preservation of agricultural land in the better-off countries where agricultural production is also the most know-how-intensive and efficient. But in fact, wherever conversion to urban development is more lucrative than the income from farming, agricultural land is in danger of losing out, unless special public policies are installed. This almost axiomatic conflict means that farmland preservation is a concern of every industrialized country. The relentless expansive energy of cities often affects precisely those areas that are most attractive to farming.

But despite the importance of preserving agricultural land from a worldwide perspective, the stimulus for farmland preservation in advanced economy countries is increasingly tilting away from concerns about agricultural production *per se*, towards environmental, landscape heritage, and quality of life concerns for the urban residents. Bruce (1998) shows how the interests and needs of the farmers themselves have receded almost to disappearance in the discourse about farmland preservation, while environmental and quality of life goals have risen. Moreover: the very goal of preserving farmland is not beyond dispute. It is sometimes challenged by competing goals, some stemming from the changing economics of particular types of agriculture, others from social equity policies or overriding national priorities, such as the supply to affordable housing. In the countries across the Atlantic, such goal conflicts are more intense than in North America.

Beyond the shared characteristics among advanced economy countries, there are significant differences in the specific goals and means for farmland preservation and in the degree of success achieved. While focusing mostly on the USA, this article takes a cross-national view of farmland preservation goals, means, and outcomes in selected advanced economy countries (see Table 1). The author bases this article on research about farmland preservation laws, policies, contexts, and degrees of success in the USA, Canada, the UK, France, The Netherlands, and Israel, and in more general terms, also Germany and Sweden. For a detailed exposition see Alterman (1997).

1. *Comparing the Contexts*

The conflict over land use has much to do with degree of urbanization, economic well-being, population

Table 1
Comparative background statistics of selected advanced-economy countries, 1999

Country	Population (in millions)	Average annual growth 1990–2000 (%)	Percent of population Urban	GNP per capita (US $ current prices)	Fertility rate per woman	Surface area (sq. km)*	Density (residents per sq. km)
USA	272.0	0.91	76	29,080	2.0	9,363,520	29
Canada	30.3	1.14	77	19,640	1.6	9,970,610	3
France	58.5	0.41	75	26,300	2.2	551,500	105
UK	58.5	0.22	89	20,870	1.7	244,100	240
Germany	82.1	0.41	87	28,280	1.3	356,733	229
Netherlands	15.6	0.55	89	25,830	1.5	40,844	378
Sweden	8.9	0.41	83	26,210	1.6	449,964	20
Israel	6.0	2.88	91	16,180	2.7	21,056	286

Sources: The Economist World in Figures 2000, The Europa World Yearbook 1999, UN Statistical Yearbook 1999, Statistical Abstracts of the United States 1999, Israel Statistical Yearbook 1999.

growth, and population density. In the advanced economy countries, where the Gross Domestic Product per person is relatively high, growing food for subsistence is not a major problem because there is enough wealth to supplement what is locally grown with imported food. Local food self-sufficiency is not yet an item on the agenda of most local land-use planning agencies, though some have argued that it should be (see Pothukuchi and Kaufman 2000). As a group, these countries contrast sharply with less developed countries or East European countries, where the figures for annual GDP per person would typically run in the hundreds or very few thousands. The advanced economy countries also differ significantly among themselves (see Table 1).

The USA and Canada are large countries and less densely inhabited than most European countries. In view of their sophisticated farming technology and know-how, the USA and Canada not only supply most of their internal needs for food and fiber, they are major food exporters on which many other countries depend. While these two countries share some aspects of farmland policies, they differ on others. Canada has by far the lowest population density among the countries surveyed. The US population density of 27 per square km, and Canada's 3 (only Australia and Iceland are Western countries with similar densities), are low compared to most countries across the Atlantic. Although most of Canada's land area is quite inhospitable climatically and its 8 percent agricultural land is concentrated along a thin strip along the US border, the area of farmland per Canadian is considerably higher than in the US.

Crossing the Atlantic brings us to a different world of generally smaller farms than in the USA or Canada, much higher population densities, and distinctly more compact cities. France, with a lower population density than most continental European countries, has 105 persons per square km, while at the other extreme The Netherlands has 378 persons. Israel, though currently less densely inhabited than the Netherlands, is in the long term the densest because it has the highest population growth and fertility rates among advanced economy countries.

Even though it has less farmland per person than North America, Western Europe faces a reverse problem: agricultural surpluses endanger the maintenance of price levels and have led the European Union to adopt policies of compulsory 'set asides' of farmland. These pose a clear challenge to the very notions of farmland preservation.

There is no agreement cross-nationally about the size of farming units that should be the target for preservation. In the USA, hobby farming by wealthy ex-urbanites is regarded as a threat to commercial farming. Europeans, who are concerned more with preserving the visual and environmental character of the traditional countryside, tend to focus more on encouraging 'family farms.' Although there is also a concern in the USA with supporting the family farm, the status of such farms is probably more secure there than across the Atlantic in view of the much-larger average size of American family farms.

2. *How Public Awareness and Planning Policies Emerged*

In both the USA and Canada, farmland preservation gained public support only in the 1980s; later than in most Western European countries. The landmark US National Agricultural Lands Study published in 1981 arose out of this growing concern and propelled it further. That study found that farmland conversion increased almost three-fold between the 1960 and the 1970s, rising from 1.1 million acres a year in the 1960s to 3 million acres a year in the 1970s. In Canada, policymakers' awareness about the depletion of farmland came approximately at the same time as the USA, and the major provincial programs, Ontario, British Columbia, and later Quebec, ensued (Reid and Yeates 1991).

In the USA and Canada, policies and methods for farmland preservation were a relative novice to the land-use planning and implementation kit of tools. No prescriptive formulae with proven results existed in the professional literature or in practice. Farmland preservation was not part of the routine work of local or state governments; someone had to 'invent' them. In the USA, the process whereby these tools emerged and tested (which, incidentally, is quite typical of other areas of land-use planning and implementation) may be called a 'survival of the fittest' process whereby more successful local policies survive and gain national recognition. While such policies were being developed and disseminated sporadically across local and state authorities in the 1980s, several 'how to' publications appeared that collated the emerging 'kit of tools' for planners and decision makers (for a listing of these see Alterman 1997). Scholarly evaluation of the degrees of success and impacts of alternative techniques has taken longer, and the jury is still out about some of them. In the Western European and Israeli context, where regional and national land-use planning is generally better entrenched in national legislation and tradition (for a comparative analysis of the broader land-use planning contexts in these countries see Alterman 2001), farmland preservation methods may therefore have been more 'natural.' That is not to say that they were more successful. Along with the general de-legitimization of direct government intervention, gradual changes occurred in the type of tools used for farmland preservation.

The other side of the coin of farmland preservation is containment of urban development and prevention of sprawl. This concern arrived in the USA and Canada much later than across the Atlantic where

compact cities are the tradition. American planners had to invent a special term, 'growth management' (and since the mid- 1990s, also 'smart growth') to discuss what is a routine part of land-use planning in much of Europe. 'Growth management' describes an open-ended set of policies and methods to phase and control the growth of cities and suburbs and to rationalize infrastructure investments. This approach often also prescribes a more coordinated institutional structure for land-use and fiscal planning (several of many titles on growth management include De Grove 1992, Stein 1993, Daniels 1999. A full review of growth management and 'smart growth' law and policy is presented by I Kushner 2000). However, only a minority of US states has seriously adopted growth management policies and tools.

3. *Farmland Preservation Policies*

With 50 states and innumerable localities, generalization about US policies is difficult. When compared with the other countries surveyed, American farmland preservation policies are characterized by a rich variety, innovativeness, and capacity for experimentation, but usually with only lukewarm effects.

A distinctive attribute of American farmland preservation policies is that planners and policymakers usually view them as a package of tools that cut across policy areas. By contrast, in some of the countries across the Atlantic, there is or was a dominant method of preservation usually based on top-down control that restricted the authority of lower-level governments to grant permits for conversion of farmland. The American-style package of tools offers not only land-use controls directly regulating the conversion of farmland to development, but also a variety of other tools such as taxation and developer exactions.

Direct control of land-use conversion from farming to non-farming use seems to be the most intuitively obvious tool. However, it does not hold center stage in the USA except in a few states, foremost Hawaii and Oregon. The most stringent version, called 'exclusive zoning,' restricts all uses of land except for farming or farming-related construction, but it is the least common of farmland preservation tool in the USA (Coughlin 1991, American Farmland Trust 1997). It is more prevalent in the countries cross the Atlantic, and ostensibly the most restrictive in Israel. In all the countries studied, exclusive zoning has been receding in importance and effectiveness. Yet in Oregon, where it is a major tool and is applied on a statewide level through a state institution, it has been found to be reasonably successful (Howe 1993, Nelson 1992). Much more common across the USA is a more flexible form of farmland conversion control, called 'nonexclusive zoning.' It permits many other land uses and puts few hurdles before conversion. Usually applied on the local and county levels only, its effect is often quite lukewarm (Coughlin 1991).

Another direct control method that is so characteristics of British and some other Western European land-use policies is an effectively regulated 'greenbelt' around urban areas (an in-depth analysis of British greenbelt policies and their effectiveness is by Elson et al. This approach is not as common or effectively used in the USA. For example, the city of Madison, a mid-size Midwest city, discussed such a concept but interpreted it as the designation of several narrow one to three-mile buffers in a few spots along the circumference to be used as golf courses.

The dominant tools for farmland preservation in the USA are those that focus on the economics of conversion rather than seeking to limit the permission to build directly. These tools hope to defer the time when farming will no longer out-power real estate, but not to prevent it altogether. Among the most common are property or other tax relief programs such as existing use appraisal rules (where farming land is assessed for its value for agriculture, even if its real estate value has risen steeply). Also common are tax deferral programs, sometimes associated, as in Wisconsin, with agreements drawn with developers to continue farming for a preset number of years. Most widely used by all states are the so-called 'right to farm' laws that provide farmers immunity from lawsuits for causing a nuisance related to agricultural work. However, fiscal means are only mildly effective because once farmers feel that it is lucrative to sell, they do not deter conversions (see Nelson 1992).

Another economic tool, minimum lot size ordinances, aims to discourage suburban housing and thus defer conversion by making the land too costly for most population groups. However, research findings have shown that it is difficult to determine what is a viable cutoff line (Nelson 1992). Florida's infrastructure concurrency requirements are intended to render development in outlying areas less lucrative, but in fact at times work in the reverse.

Property-rights based tools (for which the USA is probably the world leader) are a promising set that attempt to prevent development by buying up the very right to develop, or in effect, by compensating for the farmer's willingness to forgo the opportunity for development. These include 'land trusts' by public agencies or, most commonly, voluntary sector organizations that purchase the development rights or easements from farmers while farmers retain the right to continue farming (Endicott 1993, American Farmland Trust 1997). Transfer of development rights is potentially a superb 'made in America' tool. In an area targeted for preservation, development rights are 'purchased' by landowners in another area, tagged a 'receiving area,' who wish to increase their development rights. However, research has shown that this tool can be used only in very particular market and land-use situations, otherwise, it is ineffective (see

Pruetz 1997). A simpler tool that can be effective in some situations is cluster zoning whereby development rights are concentrated in one part of a larger land area (see Daniels 1998). Significant in their absence from the typical USA package of tools (but used very selectively in the other countries studied) are those whereby governments directly intervene in land tenure. They do this either through expropriation ('condemnation') of land, consolidation of lots that are too small for viable agriculture (prevalent in France), redistribution of lots that are too large, or most ambitiously, 'land readjustment' that redistributes development rights more justly while preserving targeted land (used in Israel in spot cases). Also rare in the USA but prevalent in The Netherlands is land banking as an instrument for timing the conversion of farmland.

4. *Levels of Government and Legal Constraints*

An important variable in the effectiveness of farmland preservation efforts is the level of government involved. From an international point of view, the USA stands out in the extreme degree of freedom held by local governments in most US states. Despite the 'smart growth' movement that has called for greater state action on land management, in most states, especially those in the heartland, the fate of farmland is still largely in the hands of local government or local civic groups. An important factor is the extensive self-incorporation and annexation powers held by local residents and local governments in many states. A majority of landowners in a particular tract of land who wish to annex themselves to an urban jurisdiction can initiate such a procedure. Although an urban municipality is empowered to refrain from rezoning agricultural land, local interests often achieve both annexations and rezoning.

In a few states, often called the 'growth management' states, the state government does play a major role in farmland preservation. Oregon, recognized as the flagship state for preservation, has developed an especially rich package of policies. Compared with the national rate of farmland conversion, Oregon has been quite successful through a combination of exclusive agricultural zones implemented by a special state-level commission, alongside policies for containing urban growth. But even this comparative success is limited to only a few farm-size categories, not all (as shown by Nelson 1992, Howe 1993). Hawaii is a second flagship state. It is unique in that it has full statewide zoning, including agricultural, that covers approximately half the state. But Hawaii has been only mildly successful in preventing conversions to urban use and, indeed, the conversion rate has doubled as the farm economy faltered (Callies 1994 describes the Hawaiian system of land-use controls. Empirical findings can be found in Ferguson and Khan 1992). Most states do not have state-level agricultural zoning. The Wisconsin tax relief and agreements program, while less illustrious than Oregon and Hawaii's, is a more typical degree of state involvement. Research into its implementation in Dane County shows little effectiveness and rampant circumvention.

What of the US federal level? National-level involvement in farmland conservation in the USA has been strikingly weak when compared with other environmental areas such as air, water, wildlife, coastlines, and other environmental areas. An attempt to enact a Federal Land Use Act was aborted in the 1970s and never reattempted (see Kayden 2001). The almost nonexistent federal role in farmland preservation (on the books is the US Farmland Protection Policy Act of 1981 oriented to minimizing the extent to which projects initiated by the federal government contribute to farmland conversion, but it has been called 'another case of benign neglect' because of its failure in implementation; see Ward 1991) contrasts with the UK, The Netherlands, and Israel (but interestingly, not France), where legislation has given central government a major role (however, the contents of the preservation goals themselves are changing across the Atlantic as well; see Alterman 1997).

A major legal impediment to farmland preservation would be if a designation of land for farming use were to be legal grounds to claim compensation from government. In none of the countries studied across the Atlantic would denial of an application to convert farmland currently designated as agriculture constitute a good ground for claiming compensation. The USA is a possible exception, but even there, the likelihood of a successful claim is very small. The inherent fear of all American planners is the 'taking issue': the constitutional limitation on excessive regulation that might be ruled by the courts to be tantamount to a taking of property without just compensation. Since the law on this issue in the USA is perpetually inconclusive and farmland preservation has not yet come up directly before the US Supreme Court (Myren 1993), there is still uncertainty about whether total denial of marketable development rights on farmland could be ruled a taking. In the past, several scholars have expressed optimism that restrictions on development would not likely be regarded as a 'taking' so long as some economically viable use is left. However, the 1992 US Supreme Court decision regarding 'Lucas' (1992) may have made the situation somewhat more uncertain; see Mandelker 1993, Lazarus 1992. The court applies a severe *per se* rule to any situation where no 'economically viable use' whatsoever is left, but leaves the law vague regarding situations of near total takings. So long as agriculture is economically lucrative, past optimism has no reason to recede. But it might prove to be less secure if farming economics were to render the 'beneficial use'

assumption tenuous, especially since farmland preservation often involves additional strata of regulations regarding runoff water, soil erosion, fertilizers, etc.

5. *The Relationships between Methods and Degrees of Success*

Achieving effective preservation is apparently much more complex and elusive than simply selecting from among the set of tools outlined above. Research shows that there is little direct relationship between the degree of success in preservation and the ostensible stringency of the tool applied (the research findings are presented in Alterman 1997). The two greatest success stories in farmland preservation among the countries studied—The Netherlands and Britain—do not primarily rely on direct control of farmland conversion. Furthermore, success in preservation does not depend on the characteristics of the broader planning system. The UK and The Netherlands have planning systems that, within the European vintage, are very different from each other (for a comparative analysis of national land-use planning in 10 advanced economy countries, including those analyzed in this article, see Alterman 2001), yet both are very successful. Nor is success intimately tied with the question of whether compensation for 'taking' is due. Most of the countries surveyed—indeed, all but possibly the USA—do not grant landowners compensation rights for denial of development permission on farmland, yet they exhibit widely varying degrees of success in preservation. In the USA there are significant variations between states in degrees of success in preservation, even though takings law in its constitutional aspects, is basically similar.

What, then, are the secrets of success in farmland preservation? First, the laudable achievements of the UK and The Netherlands could not have come about had the doctrine in these countries continued to focus on the preservation of agricultural land *per se*. In both The Netherlands and the UK, the concept of farmland preservation today draws much less attention than in the USA and Canada, despite the former countries' higher population densities and smaller area of farmland per capita. Rather, in the UK and The Netherlands (and belatedly in Israel) there has been an overt redefinition of farmland preservation as countryside preservation. Second, both countries have achieved this outcome through national level policies that are planning driven, shared and applied effectively by local and regional planning authorities, and are less reactive to incremental developer initiatives than in the USA. Third, the UK and The Netherlands focus on the containment of growth of urban areas through infill and higher densities, many times higher than those prevalent in most US cities and suburbs, even where 'smart growth' is adopted. Fourth, and perhaps most crucially, in both countries, preservation policies enjoy wide public support among electoral constituencies. Finally, planning policies in both countries are intolerant of sporadic exurban initiatives. They centrally determine the number and size of local governments, and notions such as 'bottom up' annexation initiatives are unthinkable. By contrast, since the 1980s, France, though relatively successful in the use of higher densities, is slack on local government management capacity and on national and regional planning. Now that the old tenure related preservation mechanisms are becoming economically obsolete, France is likely to achieve only lukewarm success in preservation.

Our most distinctive success story, The Netherlands, demonstrates that reliance on agricultural land conservation can no longer assure countryside conservation. The two are separate goals: economics may at times unite them, but will increasingly dissociate them, and may even place them at direct odds with each other by making industrialized farming in buildings lucrative. Countryside preservation must be considered in its own right.

See also: Agriculture, Economics of; Environmental and Resource Management; Environmentalism: Preservation and Conservation; Food Security; Green Revolution; Greening of Technology and Ecotechnology; Nature, Concepts of: Environmental and Ecological; Rural Planning: General

Bibliography

Alterman R 1997 The challenge of farmland preservation: Lessons from a six-nation comparison. *Journal of the American Planning Association* **63**(2): 220–43

Alterman R 2001 National-level planning in democratic countries: A comparative perspective. In: Alterman R (ed.) *National-Level Planning in Democratic Countries: An International Comparison of City and Regional Policy-making*. Liverpool University Press, Liverpool, UK, pp. 1–42

American Farmland Trust 1997 *Saving American Farmland: What Works*. American Farmland Trust Publication

Bruce M 1998 Thirty years of farmland preservation in North America: Discourses and ideologies of a movement. *Journal of Rural Studies* **14**(2): 233–47

Burchell R W, Listokin D, Galley C C 2000 Smart growth: More than a ghost of urban policy past, less than a bold new horizon. *Housing Policy Debate* **11**(4): 821–79

Callies D L 1994 *Preserving Paradise: Why Regulation Won't Work*. University of Hawaii Press, Hawaii

Coughlin R E 1991 Formulating and evaluating agricultural zoning programs. *Journal of the American Planning Association* **57**(2 Spring): 183–92

Daniels T L 1998 *When City and Country Collide: Managing Growth in the Metropolitan Fringe*. Island Press

DeGrove J M, Miness D A 1992 *The New Frontier for Land Policy: Planning and Growth Management in the States*. Lincoln Institute of Land Policy, Cambridge, MA

Endicott E (ed.) 1993 *Land Conservation through Public/Private Partnerships*. Lincoln Institute of Land Policy, Cambridge, MA
Elson M J, Walker S, Macdonald R 1993 *The Effectiveness of Greenbelts*. The Stationery Office Books, HMSO, London
Ferguson C A, Akram K 1992 Protecting farm land near cities: Trade-offs with affordable housing in Hawaii. *Land Use Policy* (**Oct**): 259–71
Gale D E 1999 Land use planning, environmental protection and growth management: The Florida experience. *Journal of the American Planning Association* **65**(3): 344–5
Howe D A 1993 Growth management in Oregon. In: Stein J M (ed.) *Growth Management: The Planning Challenge of the 1990s*. Sage Publications, Newbury Park, CA, pp. 61–75
Kayden 2001 National land use planning and regulation in America: Something whose time has never come. In: Alterman R (ed.) *National-Level Planning in Democratic Countries: An International Comparison of City and Regional Policymaking*. Liverpool University Press, Liverpool, UK, pp. 44–65
Kushner J A 2000 Smart growth: Urban growth management and land-use regulation law in America. *Urban Lawyer* **32**(2): 211–38
Lowry R W 1998 *Preserving Public Lands for the Future: The Politics of Intergenerational Goods*. Georgetown University Press, Washington
Mandelker D R, Payne J M 2001 *Planning and Control of Land Development: Cases and Materials,* 5th edn. Mathew Bender and Co
Myren R S 1993 Growth control as a taking. *The Urban Lawyer* **25**(2, Spring): 385–405
Nelson A C 1992 Preserving prime farmland in the face of urbanization: Lessons from Oregon. *Journal of the American Planning Association* **58**(4, Autumn): 467–88
Pothukuchi K, Kaufman J L 2000 The food system: A stranger to the planning field. *Journal of the American Planning Association* **66**(2), Spring: 34–45
Pruetz R 1997 *Saved by Development*. Arje Press, Burbank, California
Reid E P, Yeates M 1991 Bill 90—An act to protect agricultural land: An assessment of its success in Laprairie County, Quebec. *Urban Geography* **July-August**: 295–309
Sax L J 1994 *Property Rights and The Economy of Nature: Understanding Lucas v. South Carolina Coastal Council*. Land Use and Environment
Stein J M (ed.) 1993 *Growth Management: The Planning Challenge of the 1990s*. Sage Publications, Newbury Park, CA
US Department of Agriculture and the Council on Environmental Quality 1981 National Agricultural Land Study
Ward R M 1991 The US farmland protection policy act: Another case of benign neglect. *Land Use Policy* **Jan**: 63–8

R. Alterman

Fashion, Sociology of

While the word 'fashion' in English comes originally from the Latin *facere*, to make, in many European languages it derives from *modus*, to measure. Fashion can be said to be that which is understood as being right in a specific time and context. Its underlying principal is revealed in the link between the *modus*-derivatives and the term 'modern,' with its original meaning of 'now' or 'for today.'

1. The Scope of Fashion

Knowing what is 'right for today' appears to be fundamental to all periods and societies. It was only at the turn of the twentieth century, however, that fashion became the subject of serious sociological study. Its universal nature taken as read, much has been made of its growing importance. People's lives have been increasingly affected in new ways, while more of the general population has the chance to follow fashion. Sapir (1931) contrasts fashion's functional irrelevance with its tremendous symbolic significance. This explains in part why fashion can affect so much. While usually expressed outwardly in clothing and adornment, there is nothing to prevent modes of thought, morals, art, and the like from being the equivalent of clothing fashions.

More recently, attention has been drawn to fashion's growing importance to the general population. Fashion was once a limited phenomenon, with very few in a society able to participate. In modern societies, however, increasing numbers have become conscious followers of fashion. Economic well-being has made it possible for the vast majority to participate in the 'fashion race.' Wealth is not the only factor, however; people also enjoy far greater psychological mobility, and the new form of communication it generates makes fashion more important today. In *Psychologie économique* (1902), Tarde coined the term 'inter-psycholgie.' Society is the sum of people's mental condition in constant interaction. Free time and consumption become increasingly significant vehicles for such communication. New needs are born out of this mental interaction.

2. Fashion in Different Societies

Fashion's universal nature does not preclude strong variations in its social significance from society to society. Thus, Tarde (1903) differentiated between customary and fashionable societies. In the former, people see *custom* as governing their lives; they are more passionate about their country than about their period because custom prizes the past above all. In periods where *fashion* predominates, people are instead prouder of their era than their country.

Variations between societies are thrown up by different conditions. In economically stagnant periods, conventional stratification variables such as class and income are more decisive in people's lifestyles. In economically expansive periods, increasing importance is laid on fashion instead. Modern society is often contrasted with earlier epochs, and it is argued that

fashion's role has markedly increased, although many reasons are given for the change. An influential argument is that the differences in lifestyle of different groups have increased in number and scope, making it possible for people to exhibit their group's distinctive nature.

Simmel sees the 'need for distinction' and the 'need for union' as preconditions for fashion. Certain societies therefore lack the necessary motivation. Simmel (1904) here mentions that in fourteenth-century Florence, male dress was devoid of fashion. The 'need for union,' the need to express how certain groups were distinct from others with the help of fashion, was absent, and everyone could dress according to personal preference. Another of Simmel's examples was Venice at the same time. Again we find that fashion is absent from upper-class dress. Venetian aristocrats were ordered to dress in black so that the lower classes would not know how few they were. They did *not* want to differentiate themselves, and thus the 'need for distinction' that fashion requires was missing.

Variations in fashion's importance have also been tied to women's standing in society. According to Veblen (1899), it was woman's role to be a vicarious consumer; her consumption symbolized the man's wealth, and thus she became the vehicle of fashion. In Simmel (1904), however, we find an early formulation of compensation theory. Fashion acts as a safety valve. Deny women self-expression in other areas, and the only thing left for them is fashion. In fourteenth- and fifteenth-century Germany, the chance of personal development had increased, which, however, was denied to women. Rarely have more 'hypertrophic' modes of female dress been seen than in this period.

Simmel (1904) further explains fashion differences between societies. Fashion has two fundamental, symbolic functions: an ability to mark difference and an ability to express community. In many non-industrialized societies, the need for union is acute, while the need for distinction is weak. Simmel mentions certain African tribes who lacked both class differentiation and the concomitant need to signal differences. Where people in these societies want to note differences, it is often a matter of open hostility, which in turn is sterile ground for fashion. Given that fashion functions best where the distance between groups is otherwise only vaguely established, in instances where open hostility already exists, fashion as a sociological form becomes meaningless.

A similar argument can be found in König (1974), who identifies three different stages in society, each with a different type of fashion. The first significant historical change was the emergence of a fixed class hierarchy; at this stage, fashion is the preserve of an exclusive, feudal nobility. In the second stage, rigid differences between upper and lower classes are weakened; lower classes adopt the fashion of the elite, ultimately forcing it to alter fashion constantly to retain its individuality. The third stage sees the development of egalitarian democracies and industrial technology. This modern form of society brings with it the spread of fashion through mass consumption.

3. *Fashion as a Process of Change*

Fashion prescribes what is 'the right thing' for 'now.' The question of what drives fashion's continuous changes has kept researchers busy, and a series of different theories has been given in answer. In Simmel's analysis of fashion we find a foretaste of many of the theories that were to follow.

The theory of *social differentiation* argues that fashion is based on class, the upper classes abandoning a fashion as soon as it is adopted by the lower classes (see Sect. 3.1). The theory of fashion's *internal dynamic* contends that changes are powered by an *Eigendynamik* in fashion (see Sect. 3.2). The theory of fashion as a manifestation of the *expression for other social change* is a variant of *spirit of the age* theory (see Sect. 3.3). Additionally, different *system theories* of fashion have been proposed, pointing to various factors, for example fashion producers, that accelerate the speed with which fashion changes (see Sect. 3.4). Theories of fashion as a *violation of convention* argue that fashion always breaks with the past, for example with prevailing conventions and fashions (see Sect. 3.5). Finally, there are theories of fashion as a channel of psychological and social *needs for expression* (see Sect. 3.6).

3.1 *Social Differentiation*

Many have stressed fashion's origins in class. The analysis by Veblen (1899) of patterns of fashion dispersal has become a classic. A fashion begins with the 'leisure class,' the group with most money. It is attractive to this class because fashion items cost a great deal more. By degrees the fashion spreads to the upwardly aspiring middle classes. Simmel (1904) also describes how fashion must constantly be supplied with new content if it is to give social expression both to similarities within groups and differences between them. To describe fashion's downward spread through society, the *trickle-down* model was adopted by Barber and Lobel (1953), who dismissed the argument that fashion is irrational as obscuring its connections with the economic system. Meanwhile, named for the imitation of the elite that brings in turn an alteration in elite habits, we have the *chase and flight* model and the *paper chase* model.

Bell (1976) follows Veblen in seeing economic competition as a force behind changes in fashion. Dress for the middle and upper classes is an expression of wealth. This wealth can take the form of *conspicuous consumption, conspicuous leisure*, and *conspicuous*

waste. To these, Bell adds his own category of *conspicuous outrage* to convey the fashion-setting class' conscious choice of clothing that does *not* conform with prevailing notions of good taste.

The theory that fashion is the means of social distinction for the upper classes in a hierarchical society has had great influence. Mimicry of social superiors and an urge to be different from 'inferiors' are often noted as the reasons for fashion's changes, while an individual's choice of style is seen as a significant means of securing his or her position in society. Such choices often establish their distance from those below them in the social pecking order, and express subtle distinctions of degree within one and the same class. Researchers who present other models often start from the theory of social differentiation, even if they distance themselves from its application to today's society. Some reason that the idea that fashion was the distinguishing mark of the elite was probably valid in the eighteenth and nineteenth centuries, but not today. It has been argued that the traditional upper class has lost its fashion primacy to be replaced, for example, by today's powerless elite. This new elite, termed *Divi* by Alberoni (1967), uses the mass media to present new lifestyles and new consumption.

3.2 Fashion's Inner Dynamic

Simmel's analysis included a further suggestion of what drives fashion. Fashion is made up of internal, conflicting forces that necessarily lead to change; it is distinguished by both differentiation *and* union, superficiality *and* profundity, freedom *and* dependence, individuality *and* imitation, destruction *and* construction, leadership *and* submission. Antagonistic forces are united in fashion in a way that ensures continual change. Nedelmann (1990) develops the idea of the inherently contradictory nature of fashion, describing its *Eigendynamik*. Fashion, by its nature, contains a stimulus for change derived from polar opposites. People like to imitate, thus spreading a certain fashion, but are equally drawn to differentiation. Imitation leads to the spread of fashion, while innovation and differentiation lead to its contraction. Nedelmann invokes Simmel, writing that the greater the degree of concord within the upper classes, the more frantic the search for imitation from beneath, and the more constant the search for new fashions by the upper classes. On the other hand, the more the members of the lower classes use style to differentiate between themselves, the less the upper classes are inclined to come up with new fashions. Fashion in this society thus has less significance.

Davis (1992) has also noted the ambivalence of fashion, and identifies its source in sexuality. Davis writes of a continuous dialectic in clothing between the erotic and the modest. Sellerberg (1994) similarly stresses the significance of antagonism as a driving force of change. Fashion reduces *and* creates complexity. It is a continuous break with authority, but its break is authoritative: 'this is what matters, for now.' In this contradiction one finds the motor of change.

3.3 Fashion as an Expression of the Spirit of the Age

A given fashion draws a line between the present and the past, expressing what is right 'for the day': fashion represents the present. This view is the basis for theories that fashion is the spirit of the age. Fashion not only grows out of the designers' world view, but also simply reflects a time. Laver (1937), who worked mainly on dress, drew up a fashion timetable for when an item of clothing is ahead of, of, or after its time.

Indecent	10 years before its time
Shameless	5 years before its time
Daring	1 year before its time
Smart	
Dowdy	1 year after its time
Hideous	10 years after its time
Ridiculous	20 years after its time
Amusing	30 years after its time
Quaint	50 years after its time
Charming	70 years after its time
Romantic	100 years after its time
Beautiful	150 years after its time.

Blumer (1969) approached fashion as an expression of the present. In interviews with designers intended to identify the stimuli that they felt influenced their creations, they mentioned trying to 'feel' what was current within art, literature, and public debate. World events left their trace in fashion. It has to express what happened at the time, it must be a 'Zeitgeist.'

3.4 System Theories

Davis (1992) differentiates between two types of sociological model used to explain the dissemination of clothing styles. The first, the *populist* model, centers on the consumer's role and the laymen's innovations in dress. Researchers who use the populist model study general consumption, for example, how groups such as teenagers, surfers, gays, skateboarders, or feminists set their own stamp on fashion. The other model is a *system* model. Some system models are predicated on the existence of established social centers for innovation, for example Paris or Milan, with their fashion shows, designers, editors, wholesalers, stylists, and so on. The fashion system is here seen as a closed circle of producers, distributors, and consumers.

Trickle-down theory has been criticized by system theorists because it does not focus on the complex

structures of organizations and marketing that channel and mediate the fashion process. Those who use a system model aim to analyze the influence of the different elements of the fashion system. One subject of controversy has been whether designers are creative interpreters or merely passive intermediaries in the commercially vital process of heeding consumers' wishes and reconciling them with the producers' requirements.

3.5 Fashion as a Break with Convention

The destruction of what exists and the construction of something new in the process are the elements that unite in fashion, according to Simmel. Researchers who today see fashion as breaking current rules usually identify groups other than the socioeconomic upper class as leading fashion. Campbell (1992) has criticized the class differentiation theorist's explanations and identifies different types of 'new': the new, the innovative, and the novel. The novel—new as different—is central to fashion as it breaks with what already exists. The necessary conditions for creating novelty exist in peripheral groups such as bohemians and artists. It is these groups who set fashion in motion, *not* an upper class. To argue that fashion is unconventional, however, is also to argue that fashion continuously reestablishes social conventions and social order. Blumer (1969) sees both similarities and differences between fashion and custom. In both cases it is a question of social ordering. Custom, however, is static or changes only slowly, while fashion is a constant creator of social order. In Blumer's argument, every fashion emerges from an intensive process of 'collective selection.' The alternatives that are ultimately selected become fashion in different areas. Fashion conjures order out of the vast jumble of possible styles where no objective norm can determine 'right' or 'wrong.'

3.6 Fashion as a Psychological and Social Expression

Psychologically oriented theories of fashion frequently center on general human needs such as the wish to be appreciated. It is often pointed out that people's need to display and strengthen their ego is particularly strong. Fashion has great shock value in the hunt for self-expression; it provides a tangible 'look at me.'

In a discussion of fashion as an expression of personality we can return to Simmel. Fashion makes it possible to express opposites. While conveying personality (paradoxically in the extreme imitation of the modern), it also offers the chance to *avoid* conveying individuality, and thus functions as the ego's 'iron mask.'

Fashion offers a means of psychological expression, but also a social expression of protest. People express in fashion the impulses and wishes that are *not* sanctioned by normal social conventions. There is thus a constant tension between the more established 'domain of culture' and the marginal 'domain of fashion.' In fashion, young people express their stance through wearing what König (1973) calls a 'Gegen-uniform.'

The idea of fashion as a vehicle of social protest has been criticized, however. Protest styles are usually incorporated into fashion and exploited economically (Emberley 1987). Others argue that fashion's break with convention is actually a pseudodeviation. Although certainly defying convention, each break loses its original edge as soon as it is expressed in fashion (Wilson 1985).

Fashion has been interpreted as an expression of the search for a personal identity in an insecure world, and 'fads and fashions' as an expression of the worry that typifies our time. This has sometimes been interpreted as a negative phenomenon. Blumer (1969) thinks that the operation of fashion becomes natural in a changing world where people must continuously cut their ties with the past. Individuals are liberated from earlier models, and orient themselves towards the future.

4. Fashion's Cycles

Some researchers have been less interested in fashion's social dynamic than in an in-depth descriptive study of its cycles. Richardson and Kroeber (1940) use empirical data to identify six different formal aspects in women's dress taking, for example, the changes in length and width of skirts over 300 years. A number of researchers have followed Kroeber's empirical tradition, in which specific changes in fashion are seen to occur within a continuous cycle of longer periods. For example, Robinson studied beard fashions between 1842 and 1972 as time series analyses. To understand the elapse between the reappearance of the same fashion, Robinson argues that certain ground rules apply for fashion's cyclical changes, one being that as long as a number of people still follow a certain fashion, it prevents younger people from adopting it.

Fashion in this perspective is not born, but is rediscovered. Cross-cyclical changes receive particular attention as demonstrating sharp reactions to extreme styles. For example, the streamlined form's soft curves are contrasted with the ensuing period's particularly angular design.

5. Ambivalence Towards Fashion: The Vitally Unimportant

Fashion's symbolic significance has been contrasted with its lack of practicality. This has led many today to express ambivalence towards fashion, and there is a

tendency to legitimize it using functional arguments; the rational and practical is acceptable where the impractical is not. Function and symbolism can sometimes be reconciled in fashion, however. Aerodynamics has been described as a discovery of great significance for industrial design, and for a time all household appliances were streamlined. 'Streamlining' symbolized America's enlightened practice of throwing over old conventions to create a new and dynamic civilization. Those who chose, for example, a streamlined kettle, demonstrated that they also chose the rational, the future, the practical.

6. *Fashion as Unpredictable*

Increasingly, fashion's analysts have come to emphasize the inner dynamic in fashion. Individual theories are run together to reflect fashion's inherent contradictions: it is coercion *and* individual choice; it is social order *and* a break with social order; it is union *and* differentiation. The nature of fashion is seen as ever more complex.

There is a more simplistic perspective, however, common to those who base their interpretation of fashion on industrial products. Here one finds phrases such as 'the most thorough logic of planned aging,' and 'industry is based on extolling a perpetual neology.' Fashion, however, is to be found everywhere, not only in areas that are subject to commercial planning, while the aging of fashion is impossible to predict precisely, however much industry demands precise guidelines. The problems stem from fashion's most central characteristic: because it symbolizes 'now,' it can never be foreseen with any certainty. Fashion pinpoints what is right in the present, but the moment it can be predicted accurately, it is no longer a matter of fashion. The transient symbols of what is right for now can never be nailed down in advance.

See also: Action, Collective; Art and Culture, Economics of; Art, Sociology of; Collective Behavior, Sociology of; Consumer Psychology; Consumption, Sociology of; Cultural Expression and Action; Culture, Sociology of; Differentiation: Social; Dress and Fashion; Expressive Forms as Generators, Transmitters, and Transformers of Social Power; Hegemony: Cultural; Identity: Social; Sapir, Edward (1884–1939); Simmel, Georg (1858–1918); Taste and Taste Culture; Values, Sociology of

Bibliography

Alberoni F 1967 *Consumi e società*. Società editrice il Mulino, Bologna, Italy

Barber B, Lobel L S 1953 Fashion in women's clothes and the American social system. In: Bendix R, Lipset S M (eds.) *Class, Status and Power*. Free Press, Glencoe, IL, pp. 323–32

Bell Q 1976 *On Human Finery*. Hogarth Press, London

Blumer H 1969 Fashion: from class differentiation to collective selection. *Sociological Quarterly* **10**: 275

Campbell C 1992 The desire for the new. Its nature and social location as presented in theories of fashion and modern consumerism. In: Silverstone R, Hirsch E (eds.) *Consuming Technologies: Media and Information in Domestic Spaces*. Routledge, London

Davis F 1992 *Fashion, Culture and Identity*. University of Chicago Press, Chicago, IL

Emberley J 1987 The fashion apparatus and the deconstruction of post-modern subjectivity. *Canadian Journal of Political Social and Theory* **11**: 38

König R 1973 *The Restless Image: A Sociology of Fashion*. Allen & Unwin, London

König R 1974 *A la Mode: On the Social Psychology of Fashion*. Seabury Press, New York

Laver J 1937 *Tastes and Fashion: From the French Revolution Until Today*. G C Harrap, London

Levine D N 1971 *On Individuality and Social Forms: Selected Writings*. University of Chicago Press, Chicago, IL

Nedelmann B 1990 Georg Simmel as an analyst of autonomous processes: the merry-go-round of fashion. In: Kaern M, Phillips B S, Cohen R S (eds.) *Georg Simmel and Contemporary Sociology*. Kluwer, Dordrecht, The Netherlands

Richardson J, Kroeber A L 1940 *Three Centuries of Women's Dress Fashions: A Quantitative Analysis*. University of California, Berkeley, CA

Robinson D E 1976 Fashions in shaving and trimming of the beard: the men of the illustrated London news, 1842–1972. *American Journal of Sociology* **81**: 1133–41

Sapir E 1931 *Fashion. Encyclopedia of the Social Sciences*. The Macmillan Company, New York, Vol. VI, pp. 139–44

Sellerberg A-M 1994 *A Blend of Contradictions: Georg Simmel in Theory and Practice*. Transaction Publishers, New Brunswick, NJ

Simmel G 1904 Fashion. *Fashion International Quarterly* **10**: 130–55

Tarde G 1902 *Psychologie Economique*. F Alcan, Paris

Tarde G 1903 *Laws of Imitation* [trans. from the 2nd French edition, Parsons E C] *Custom and Fashion*. Holt, New York, Chap. vii

Veblen T B 1899 *The Theory of the Leisure Class: An Economic Study in the Evolution Of Institutions*. MacMillan, New York

Wilson E 1985 *Adorned in Dreams*. Virago, London

A.-M. Sellerberg

Fast Food Enterprises

Fast food enterprises are companies that supply catering services on a fast turnover basis. The fast food sector comprises hamburger, chicken, pizza and other food outlets, coffee, pastry and icecream shops, ethnic food outlets as well as home delivery services for some of these. Many are franchised to international food chains. The rapid growth of this sector across cultural, national, and geographical frontiers has been taken, first, to signify fundamental changes in consumer

tastes and lifestyles, associated with changing family structures and employment patterns. Second, it is viewed as one of the most tangible features of cultural and economic globalization, marking the colonization by transnational organizations of one of the most important cultural terrains, namely what, how, when, and where people eat. Third, fast food enterprises seem to offer a prototype of the transition of the service industries from skill-based and people-based technologies to Taylorist and Fordist approaches which revolutionized manufacturing in the earlier part of the century. It is further argued that the fast food industry has become a paradigm for profound changes in the nature of contemporary societies. The term McDonaldization (Ritzer 1993–6) has been used to indicate the increasing dominance of the core principles embodied by the fast food restaurant in virtually every sector of society. Fourth, employment in fast food has become the first 'real' work exposure for increasing numbers of young people, affecting subsequent work attitudes and experiences. Lying at the meeting point of work and leisure, fast food enterprises represent a vital area on which to develop and test theories regarding the formation of contemporary identities as well as the interface between customer and employee.

1. *Rationalization*

Numerous factors account for the spectacular growth of this sector since the late 1950s. These include new franchising systems, pioneered by McDonald's Ray Kroc, as well as broader social changes such as increasing participation of married women in the workforce, devaluation of housework, decreasing size of households, ever more hectic lifestyles, the greater importance of children as consumers, increasingly personalized and differentiated tastes in food, and ever greater consumer exposure to advertising images. All of these factors can all be seen as contributing to the success of 'convenience eating.' Above all, however, the success of fast food enterprises is often seen as the outcome of the belated rationalization of production, delivery, and marketing of a consistent, reliable, and inexpensive product. Writing in 1972, Levitt argued that fast food marked the industrialization of service, the application in catering of 'technocratic thinking which … replace[s] the high-cost and erratic elegance of the artisan with the low-cost, predictable munificence of the manufacturer' (Levitt 1972, pp. 43–4). Taylorist and Fordist principles of management are clearly observable in the organization of the fast food service. These include a highly standardized and uniform product, routinized and fragmented food preparation procedures, highly rationalized technologies, and the employment of staff with virtually no catering skills and minimal training. Thus, Levitt views a McDonald's retail outlet as 'a machine that produces, with the help of totally unskilled machine tenders, a highly polished product. Through painstaking attention to total design and facilities planning, everything is built integrally into the machines itself, into the technology of the system. The only choice available to the attendant is to operate it exactly as the designers intended' (Levitt 1972, p. 46).

Rationalization in fast food enterprises extends into several other areas—the glamorization of their brands through relentless advertising, the systematic control of the customer through numerous discursive and nondiscursive practices (the location of outlets, menus, products, pricing, queuing, seating and waiting arrangements, background music, lighting, and outlet décor), and the organized attempt to control the emotional displays of the employees through the emblematic 'service smile' (Sturdy 1998). In all of these respects, fast food enterprises have moved rationalization far beyond Fordist and Taylorist regimes. Ritzer (1999) has argued along Weberian lines that excessive rationalization brings about the dissolution of magic and disenchantment. This leads to attempts on the part of producers to re-enchant consumption, by transforming their outlets into 'cathedrals' where fantasies are enacted—thus museums, cruise-ships, holiday resorts, theme parks, shopping malls, and even solid utilitarian institutions, like hospitals and universities, are gradually becoming arenas where consumers can fantasize, explore, experiment, and spend, after the model of fast food outlets. In a paradoxical way, then, attempts to re-enchant consumption are organized around the same rationalizing principles which account for disenchantment.

2. *McJobs*

Employment in fast food enterprises ('McJobs' in Ritzer's expression) is characterized by extensive use of young workers, part-time employment, casualization, and high levels of staff turnover. Very few employees, including managers, view their work in fast food as anything other than short-term. Levels of remuneration are generally poor, as many employees live with parents or other relatives. Union representation is low, partly due to the casual nature of employment and partly because well-organized resistance of unions by employers. Job satisfaction is generally low, even by the low standards common in catering (Gabriel 1988). Discretion and individuality at work are very limited. While there is evidence that employees frequently break the rules even with the collusion of managers, this is most often in the interest of meeting demands from customers rather than in the form of resistance or self-expression. Gossip, joking, story-telling, day-dreaming, and some game-playing are all used as survival mechanisms.

Deskilling and control in fast food industries go beyond their counterparts in the manufacturing sector.

On the basis of fieldwork, Leidner (1993) and Reiter (1996) have argued that counter work entails extensive emotional labor (Hochschild 1983, Fineman 1993), a systematic attempt to control the employees' total emotional outlook towards the customers rather simply their displays of emotion. While this argument has been contested, there is little doubt that much of the work performed in fast food enterprises combines low levels of skill with constant exposure to the demanding and critical gaze of the customer. Regular visits incognito by company inspectors add to the sense of constant surveillance. Interactions with the customers may not be scripted, but they are constantly monitored and assessed.

3. *The Consumers*

The high visibility of fast food enterprises, their restaurants occupying conspicuous positions in virtually every city and town, their logos, signs, and advertisements instantly recognizable, has made them an almost archetypal symbol of globalization. Some have argued that Americanization is a more appropriate term than globalization. Virtually all major fast food chains (including those associated with ethnic cuisine, like Mexican or Italian) originated in the United States, their products firmly developed on American food tastes. More importantly, fast food is symbolically linked to an American lifestyle, recognizable the world over, fast, informal, cool, fun, strong, reliable, open, modern, and democratic. The democratic quality of fast food is embodied in the orderly queues, the equal treatment of all customers, as well as well-publicized visits to outlets by celebrities, such as Prime Ministers, film stars, and tycoons, including Mr. Bill Gates, the world's richest man. The fun quality is highlighted by the ambience, the music, the smiling faces. The lifestyle symbolized by fast food outlets is especially attractive to children, who in many countries have been the consumer vanguard adopting fast food restaurants as their preferred venues for family entertainment. Fast food restaurants are not merely children-friendly places; they are places where the authority of the paying parent is diminished and replaced by an informal atmosphere of egalitarianism. The food has been prepaid, the tables have no heads to be occupied by the head of the family, the décor and music devised to mark the space as a space for children, where adults are welcome as long as they relinquish any claim to authority and discipline.

Based on a series of ethnographic studies of the growth of McDonald's in five East Asian countries, Watson claims that 'more than any other factor, ..., McDonald's success is attributable to the revolution in family values that has transformed East Asia' (Watson 1997, p. 19). The decline in family size, the weakening of extended kinship systems, and the erosion of parental authority have all contributed to the considerable success of fast food in many developing countries. This has given rise to criticism of fast food as a type of American cultural imperialism, contributing, among other things, to the corruption of local traditions, the increasing uniformity of tastes, and the unstoppable growth of commercialization and consumerism. Watson and his colleagues observe unmistakable signs of cultural change brought about by the growth of transnational enterprises. An interesting example is the adoption of Western birthday parties in countries which neither celebrated birthdays nor even viewed them as significant dates. Moreover, they document evidence of direct resistance to the expansion of fast food restaurants ranging from rumor-mongering in Japan ('hamburgers contain cat-meat') to overt charges of antipatriotism leveled against fast food patrons in Korea. They conclude, however, that the impact of fast food on local cultures is neither wholly negative nor wholly one way. They observe that many local traditional restaurants have been forced to improve the standards of cleanliness, hygiene, and service in order to remain competitive against Western fast food chains. They also point at evidence of 'localization,' arguing that when an international institution is adopted by a new culture it is modified (occasionally as a token gesture) to suit local cultural and social conditions. While initially accepted as an exotic import, such an institution eventually becomes acclimatized to local culture. Some elements of fast food, like queuing, self-provisioning and self-seating, have been accepted, others have been modified. In Asian countries, McDonald's outlets are often turned into de facto leisure centers, coffee houses, and after-school clubs.

The charges leveled against the fast food business, however, go well beyond those of cultural imperialism. Some of these charges surfaced in 1994 during a well-publicized case in which McDonald's sued for libel two British activists who had published a pamphlet strongly critical of the company. This accused McDonald's of (a) causing widespread environmental damage, (b) exploiting young consumers through invasive advertising, (c) maltreating their employees, (d) causing unnecessary cruelty to animals, (e) promoting unhealthy eating habits, and (f) causing hunger in the Third World. The 'McLibel' trial became the longest trial in English history and was a public relations disaster for the company, spawning a much visited Website and becoming a focal point for antifast food activism. The verdict, handed down in 1997, amounted to a Pyrrhic victory for the company—the judge ruled that while some of the activists' allegations had been excessive, many were based on fact (Vidal 1998). In addition to giving the original pamphlet unprecedented publicity, the trial also had the effect of revealing the extent to which McDonald's resorted to litigation to silence its critics, resulting in further negative publicity for the company.

In spite of such incidents which indicate resistance to the fast food business on political, environmental, and cultural grounds, their growth appears unstoppable. The demographic, cultural, and employment factors which account for their early successes, if anything, appear to accelerate and spread to the newly industrialized countries. It appears that the major challenges to fast food enterprises in the future will come, first, from an increasing difficulty in recruiting casual, low-paid, young workers, and, second, from customer fatigue and the transience of fashions. It is quite possible that already some of the major fast food chains are losing customers to more up-market, fashionable, and at least temporarily alluring competitors.

See also: Consumer Culture; Consumer Psychology; Consumption, Sociology of; Customer Satisfaction; Globalization and World Culture; Leisure and Cultural Consumption; Leisure, Psychology of; Popular Culture; Retail Trade; Service Economy, Geography of; Services Marketing

Bibliography

Fineman S 1993 *Emotion in Organizations*. Sage, London
Gabriel Y 1988 *Working Lives in Catering*. Routledge, London
Hochschild A R 1983 *The Managed Heart: Commercialization of Human Feeling*. University of California Press, Berkeley, CA
Leidner R 1993 *Fast Food, Fast Talk*. University of California Press, Berkeley, CA
Levitt T 1972 Production-line approach to service. *Harvard Business Review* **50**: 41–52
Reiter E 1996 *Making Fast Food: From the Frying Pan into the Fryer*. McGill-Queen's University Press, Montreal, PQ
Ritzer G 1993–6 *The McDonaldization of Society*. Sage, London
Ritzer G 1999 *Enchanting a Disenchanted World: Revolutionizing the Means of Consumption*. Pine Forge Press, Thousand Oaks, CA
Sturdy A 1998 Customer care in a consumer society: Smiling and sometimes meaning it? *Organization* **5**(1): 27–53
Vidal J 1998 *McLibel*. London, Macmillan
Watson J L 1997 Introduction: Transnationalism, localization, and fast food in East Asia. In: Watson J L (ed.) *Golden Arches East: McDonald's in East Asia*. Stanford University Press, Stanford, CA, pp. 1–38

Y. Gabriel

Fatherhood

Fatherhood defines a biological and social relationship between a male parent and his offspring. In common usage, 'to father' means to impregnate a woman and beget a child, thus describing a kinship connection and allowing for the intergenerational transfer of wealth and authority in a patrilineal descent system. Fatherhood also refers to a patterned set of parenting behaviors and reflects a society's ideals about the rights and obligations of men in families. In general, fathers (including biological fathers, step-fathers, divorced fathers, and father-figures) are expected to love, support, protect, nurture, teach, discipline, and control their children, though many men do not fulfill all aspects of this idealized role. The concept of fatherhood also generalizes to other social and symbolic relationships, as when Christians refer to 'God The Father,' Catholics call priests 'Father,' Germans speak of their native country as 'The Fatherland,' and Americans label George Washington 'the father' of their country. Fatherhood thus reflects a normative set of social practices and expectations that become institutionalized within the family, religion, law, and culture.

1. Cross-cultural Variation in Fathering

Although folk beliefs suggest that fathering and mothering entail behaviors that are fixed by reproductive biology, humans must learn how to parent much like they learn other social behaviors. In every culture and historical period, men's family roles are shaped by social, economic, and cultural pressures, resulting in unique fatherhood ideals and practices. Although women have been the primary caretakers of young children in all cultures, the father role has varied from virtually no direct involvement to active participation in all aspects of children's routine care, feeding, and protection.

Anthropologists have identified two general patterns of male involvement in domestic life across the world's societies—one intimate and the other aloof. In the intimate pattern, men eat and sleep with their wives and children, talk with them during evening meals, attend births, and participate actively in infant care. In the contrasting aloof pattern, men often eat and sleep apart from women, spend their leisure time with men, stay away during births, and seldom help with child care (Whiting and Whiting 1975).

Cross-cultural research shows that about half of the world's known societies have exhibited close father–child relationships, with contact typically increasing as children grow older. Compared to societies with distant father–child relationships, those with involved fathers are more likely to be peaceful and include nurturing deities of both sexes in their creation stories (Sanday 1981). Distant-father societies are more likely to have stern male gods, with groups of men routinely excluding women from community decision-making, demanding deference from wives, and engaging in competitive displays of manliness (Coltrane 1996). Higher levels of father involvement with children are thus associated with more intimate husband–wife relationships and more gender equality

in the society, whereas lower levels of father involvement are associated with more distant husband–wife relationships and more overt male dominance.

2. Theories of Fatherhood and Families

To understand cross-cultural variation in fatherhood and identify the essence of family, Malinowski (1929) developed a theory based on the 'principle of legitimacy.' Malinowski observed that in Western societies, families were organized around the authority of fathers, with patrilineal descent and patrilocal residence assuring the position of the male household head. In matrilineal Trobriand Island society, in contrast, Malinowski observed that fathers were virtual strangers living among their wives' relatives, making few contributions to the family household and exercising little power. The important link was not between husband and wife as in Western societies, but between sister and brother, with the latter being the symbolic and actual head of the household. These observations led Malinowski to question the universality of the father-centered household pattern and the importance of biological fatherhood, suggesting an alternative focus on the legitimacy of children. To be legitimate, a child must be socially approved through a formal marriage that has taken place between his or her parents. Malinowski thus argued that the important focus was social fatherhood, (not biological fatherhood). The purpose of social fatherhood was attaching a child to a particular man (whether father or uncle), thereby securing a place for that child in the social structure.

Subsequent theories of family and fatherhood suggest that Malinowski's formulation may have been too narrow, as well as containing some gender-biased assumptions (Coltrane and Collins 2001). Societies have existed where the majority of children are not born to mothers married to husbands (e.g., in the Caribbean Islands), and similar patterns are becoming increasingly common in urban centers of modern industrial societies. Malinowski's theory focuses on connecting children to social fathers because this brings them legal and social status, but it ignores the possibility that children could gain that status from mothers. Since males usually control more wealth and exercise more political power than females, it has been more important for a child's economic fate to be connected to a man than to a woman, but it is important to see how this pattern can vary under different social and economic conditions.

According to Collins' (1975) conflict theory, a strong emphasis on legitimacy is most common in societies with high levels of male dominance. Nonmarital birth ('illegitimacy') is punished most severely in societies where men have the most extreme monopoly on power. In medieval Arab societies, for example, if a woman of the harem had an illegitimate child, she was punished by death, with the sentence carried out by the woman's husband, father, or brother. On the other hand, if the father of the illegitimate child was powerful enough, the mother could have a reasonably honorable status as a concubine or mistress, and the child could grow up with a recognized social position. This pattern has been repeated in male-dominated societies throughout the world. In medieval France, for instance, one duke's son was known as William the Bastard; later, after he founded the English throne, he was called William the Conqueror. Considering variations in male dominance, Collins (1975) argues that illegitimacy rules do not so much say *no child without an official father* as they say there should be *no childbearing woman who does not officially belong to a man*. Collins extends Malinowski's legitimacy theory by conceiving of the family as a property system regulating three kinds of property: rights of sexual possession, economic property rights, and intergenerational property rights. Such approaches help explain how fatherhood ideals and practices maintain men's dominance over women in both premodern and modern societies.

3. Historical Development of Fatherhood

Fatherhood has been linked to the exercise of family authority at least since Ancient Greece and Rome, but it has been fathers' public lives—their work, political exploits, and heroic battles—that have been chronicled in most histories. Fathers have also been care providers and teachers in different cultures and eras, and recent scholarship documents such practices. Before the modern era, childrearing was a collective enterprise and most family matters, including father–child relationships, were ruled primarily by duty and obligation (Shorter 1975). Because men's work as farmers, artisans, and tradesmen occurred in the family household, most fathers were a visible presence in their children's lives. Men introduced sons to farming or craft work within the household economy, oversaw the work of others, and were responsible for maintaining harmonious household relations. The family home was thus a system of control, as well as a center of production, and both functions tended to reinforce the father's authority (Griswold 1993). Though mothers provided most direct care for infants and young children, men tended to be very active in the training and tutoring of children, and most parental advice was addressed to fathers. Because they were moral teachers and family heads, fathers were thought to have far greater responsibility for, and influence on, their children than mothers (Pleck 1987).

As market economies took over from home-based production, the father's position as head and master of the household and moral instructor of his children was transformed. Men were increasingly called upon to seek employment outside the home and their direct

contact with family members declined. Although men's physical presence in the home was diminished, most did not stop participating in family activities. In fact, the first 'new' fatherhood movement emerged in the nineteenth century as middle-class fathers helped their wives through childbirth, romped with their offspring, forged close emotional bonds with daughters, and worried about sons' vocations and education (Griswold 1993). As the wage labor economy developed, however, men's occupational achievement outside the household took on stronger moral overtones and men came to be seen as fulfilling their family and civic duty not by teaching and interacting with their children as before, but by supporting the family financially (i.e., being a 'good provider'). The middle-class home, which was previously the normal site of production, consumption, and virtually everything else in life, was transformed slowly into a nurturant child-centered haven set apart from the impersonal world of work, politics, and other public pursuits. The ideal of separate spheres for men and women—work for him and home for her—became one of the defining features of the late nineteenth and early twentieth centuries.

4. *Recent Trends*

By the middle of the twentieth century, mothers had become almost solely responsible for home and children. Separate spheres ideology and the popularity of Freud's ideas about mother–infant bonding led to widespread acceptance of concepts like 'maternal deprivation.' Though few researchers bothered to ask who besides mothers took care of children, some researchers focused on 'father absence,' worrying that single-mother families could not set firm limits nor provide adequate masculine role models for sons. These studies have been criticized for confusing father absence with poverty, and for adopting the narrow view that aggressive and unemotional masculinity was both normative and desirable.

Starting in the 1970s, psychologists and sociologists began to report that a new fatherhood ideal was gaining in popularity. Films and television shows celebrated men's love for children, even if the men they pictured were initially shown as comically inept. Most social scientists had assumed that men were not capable of, nor interested in, actual hands-on parenting, but research showed that, at least in laboratory settings, fathers could care for infants and young children (Parke 1996). Following an individual differences research tradition in psychology, studies tended to focus on 'sex differences' in parenting style. Fathers typically were found to be more likely than mothers to engage in rough and tumble play, to be more directive, to give more attention to sons, and to treat boys and girls in sex-stereotyped fashion. Fathers' unique play styles were found to enhance children's emotional and social competence, in part because they forced children to self-regulate (Parke 1996). Other studies by sociologists and psychologists showed that men and women performed child-care in different settings and in conjunction with different activities. Women most often did child-care in the home and frequently were engaged in multiple household tasks while simultaneously attending to children's needs. Fathers, in contrast, were more likely to watch children in public settings (e.g., parks and playgrounds) and to take care of them as a sole activity. The conventional pattern was thus for child care to be an ongoing and taken-for-granted task for mothers, but a novel and fun distraction for fathers (Coltrane 1998).

In response to the women's movement and maternal labor force participation, many observers predicted that fathers would begin assuming responsibility for a larger share of parenting and other family work. Several decades of research, however, have shown that while men are doing more, women continue to do most of the housework, remain the child care experts, and serve as emotional managers for their families. In general, family divisions of labor become more gender-segregated upon having children, with men increasing, and women decreasing, their paid work hours, even though both tend to increase their hourly contributions to family labor after a birth. Mothers are still more likely than fathers to take time off from their jobs to provide continuous child-care, spending significantly more time than fathers feeding, dressing, cleaning, and watching young children. Studies find that mothers spend double the time of fathers in these activities, even though many men did increase the time they spent with preschool and school-aged children during the 1980s and 1990s (Parke 1996, Pleck 1997). Two-parent, dual-earner families tend to share the most, with increasing numbers working different shifts and alternating routine child care between them. Research shows that couples share more family work when they are both employed full time, their incomes are similar, they believe in gender equality, and there are fewer, older, and male children (Coltrane 2000).

5. *Measuring Father Involvement*

Most researchers now agree that what fathers do with and for children is more important than simple co-residence or frequency of father-child contact (Parke 1996). The most influential approach distinguishing among various types of father involvement was offered by Lamb and his colleagues (Lamb et al. 1985) who suggested three principal components of fathering: interaction, availability, and responsibility. Interaction refers to the father's direct contact with his child through care giving and shared activities. Recent measurement strategies in this domain pay attention to the form and content of interaction, distinguishing between play, instruction, and care-taking, and at-

tempting to assess the quality of interaction, as well as the time spent interacting. Availability is a related concept focusing on the father's potential availability for interaction, by virtue of being present or accessible to the child. Responsibility refers to the role the father takes in ascertaining that the child is taken care of and arranging for resources to be available for the child both inside and outside the home. This type of parental management is one of the least studied aspects of fathering, but is one of the most important (Parke 1996). Mothers remain child-care managers in the vast majority of households, but evidence indicates that at least some fathers are taking a more active role in this domain (Coltrane 1996, Pleck 1997).

6. *What Difference Do Fathers Make?*

As scholars pay more attention to fathers, we are beginning to understand what influence their involvement might have on child development and gender relations. We know that fathers' resources improve children's life chances. When fathers also share child care and housework with their wives, employed women escape total responsibility for family work, evaluate the division of labor as more fair, are less depressed, and enjoy higher levels of marital satisfaction (Coltrane 2000). When men care for young children on a regular basis, they emphasize verbal interaction, notice and use more subtle cues, and treat sons and daughters similarly (rather than focusing on play, giving orders, and sex-typing children). These differences between everyday and occasional fathering deserve further study, for they portend differential social impacts. Most studies show that children with actively involved fathers demonstrate enhanced intellectual and socio-emotional adjustment, as well as developing more balanced ego structures and gender expectations (Coltrane 1996, Parke 1996).

At the same time that many fathers are spending an increasing amount of time with their children, high levels of divorce and increasing rates of nonmarital childbirth are keeping other men away from children. In many industrialized countries, growing numbers of men rarely see their children and do not support them financially. In the US, for example, single-parent households are increasing rapidly, but only about half of children eligible for child support from nonresident parents obtain court orders for it, and only about half of those receive full payments. Both trends in fatherhood—toward more direct involvement and toward less contact and financial support—are responses to the same underlying social developments, including women's rising labor force participation and the increasingly optional nature of marriage.

Longitudinal research shows that children from one-parent households (usually mother-headed) are at greater risk for negative adult outcomes (e.g., lower educational and occupational achievement, earlier childbirth, relationship problems) than those from two-parent families (Cherlin et al. 1995, McLanahan and Sandefur 1994). Debates continue over whether these outcomes are specifically attributable to the actions or inactions of fathers, and whether state support and community programs might ameliorate negative effects. Some commentators call for a return to tougher fault-based divorce laws and gender-segregated models of fatherhood, but others label such responses a reactionary backlash against women's increasing autonomy. Reinstating fathers as household heads and expecting only minimal direct domestic contributions from them is likely to maintain hierarchical structures both inside and outside families. Whether absent or physically present and doing little, the ability of fathers to stay aloof from daily family life is linked to patriarchal family authority and the reproduction of systemic male dominance. Fathers' performance of everyday family work, in contrast, is associated with more intimate husband–wife relationships, more democratic family decision-making, and more gender equality in the society. The twenty-first century will undoubtedly see a continuing struggle between various intimate and aloof styles of fatherhood, with profound implications for men, women, and children.

Bibliography

Cherlin A, Kiernan K E, Chase-Landsdale P L 1995 Parental divorce in childhood and demographic outcomes in young adulthood. *Demography* **32**: 299–318

Collins R 1975 *Conflict Sociology*. Academic Press, New York

Coltrane S, Collins R 2001 *Sociology of Marriage and the Family*. Wadsworth, Belmont, CA

Coltrane S 1996 *Family Man*. Oxford University Press, New York

Coltrane S 1998 *Gender and Families*. Pine Forge Press, Thousand Oaks, CA

Coltrane S 2000 Research on household labor. *Journal of Marriage and the Family* **62**: 363–89

Griswold R L 1993 *Fatherhood in America*. Basic Books, New York

Lamb M E, Pleck J H, Charnov E L, Levine J 1985 Paternal behavior in humans. *American Zoologist* **25**: 883–94

McLanahan S, Sandefur G 1994 *Growing Up with a Single Parent* Harvard University Press, Cambridge, MA

Malinowski B 1929 *The Sexual Life of Savages in North-Western Melanesia*. H. Liveright, New York

Parke R D 1996 *Fatherhood*. Harvard University Press, Cambridge, MA

Pleck J H 1987 American fathering in historical perspective. In: Kimmel M S (ed.) *Changing Men*. 3rd edn. Sage, Newbury Park, CA, pp. 83–97

Pleck J H 1997 Paternal involvement. In: Lamb M E (ed.) *The Role of the Father in Child Development*, Wiley, New York, pp. 66–103

Sanday P R 1981 *Female Power and Male Dominance*. Cambridge University Press, Cambridge, UK

Shorter E 1975 *The Making of the Modern Family*. Basic Books, New York

Whiting J W M, Whiting B 1975 Aloofness and intimacy of husbands and wives. *Ethos* **3**: 183–207

S. Coltrane

Fear Conditioning

1. Basic Terms

The pairing of an initially neutral stimulus (the conditioned stimulus—CS) with a biologically relevant stimulus (the unconditioned stimulus—US) comes to elicit a response (conditioned response—CR) that is usually but not always similar to the response previously associated with the unconditioned stimulus (the unconditioned response—UR). In fear conditioning, the US is an aversive fear-eliciting stimulus such as painful electric shock or loud noise, the CS is a neutral tone or light stimulus. The unconditioned and the conditioned response consist of changes on the subjective, the behavioral and the physiological level and include (in humans) enhanced subjective fear and responses such as freezing, changes in heart rate and skin conductance, the release of stress hormones, reduced pain sensitivity and startle reflex potentiation.

The development of the CR is based on the formation of an association between a neutral stimulus and a stimulus with innate biological significance (Rescorla 1988). Most studies involving fear conditioning have used cue rather than context conditioning, i.e., discrete CSs were presented rather than using the environment of the animal (e.g., the cage) as CS. In addition, delay conditioning where the CS terminates with the US rather than trace conditioning where the CS and US are separated in time were used in most studies. Fear can be viewed as a specific reaction to threatening stimuli. It can turn into an anxiety disorder when the fear becomes disproportionate to the stimulus that elicits it or when fear is experienced in inappropriate situations.

2. Mechanisms of Fear Conditioning in Animals

The psychobiological basis of fear conditioning in animals, especially rodents, has been well established (LeDoux 2000, Davis 1992, Maren and Fanselow 1993). When a tone is used as CS and electric shock as US, the auditory CS reaches the lateral nucleus of the amygdala from the thalamus or the cortex, probably depending on the complexity of the CS. When contextual stimuli are employed as CS, the hippocampus is also involved and transmits the signal to the basolateral and the accessory basolateral nucleus of the amygdala. The nociceptive US also reaches the

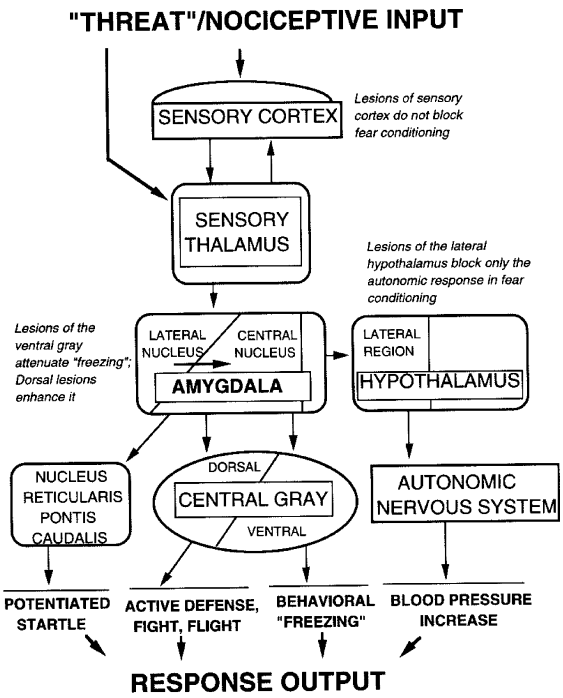

Figure 1
Fear stimuli proceed from sense receptor systems to the sensory cortex and/or sensory thalamus and then to the amygdala. There are three important connections efferent to the amygdala: (a) a projection from the central amygdala to the lateral hypothalamic area that mediates the autonomic components of the emotional response; (b) projections to the midbrain central gray region mediates defensive freezing and escape behavior and (d) a direct projection to the nucleus reticularis pontis caudalis modulates the startle reflex. (Based on Bradley Lang 2000 with kind permission)

lateral nucleus of the amygdala via thalamic and cortical pathways. Conditioning seems to require an intact lateral nucleus of the amygdala when cue conditioning and an intact basolateral and accessory basolateral nucleus when context conditioning is involved. The central nucleus of the amygdala receives input from the lateral and the basolateral nuclei and it projects to brainstem nuclei that initiate the expression of peripheral fear responses. Whereas the lateral or basolateral nucleus are thus important for the acquisition of the CR, the central nucleus is essential for the expression of the CR (see Figure 1).

The initial site of plasticity seems to be the lateral nucleus of the amygdala; however, thalamic and cortical reorganization have also been shown during fear conditioning (cf. Weinberger 1998). It is to date not clear if the memory trace is stored in the amygdala or if the amygdala only modulates memories that are stored elsewhere, e.g., in the cerebral cortex. If the

amygdala is inactivated during fear conditioning, no CR is formed. If the inactivation occurs after conditioning, then the memory trace is not abolished (LeDoux 2000). In addition, cognitive aspects of fear may involve other pathways.

3. Fear Conditioning in Humans

In humans, the role of the amygdala is less clear. In contrast to rodents, humans possess a much larger frontal cortex and it is likely that it takes over some of the functions that are related to the amygdala in rodents. Damage to the amygdala in humans, has, however, also been related to deficits in fear conditioning whereas lesions in the hippocampus have been shown to interfere with declarative learning and contextual fear conditioning (cf. Bechara et al. 1995). Most studies using positron emission tomography (PET) were unable to replicate the findings from animal experiments that assigned an important role to the amygdala in fear conditioning (for a review see Büchel and Dolan 2000). They instead reported activations in areas such as the hypothalamus, the thalamus, the cingulate gyrus, the premotor and the parietal cortex and, recently, also in the cerebellum (Fischer et al. 2000). Later studies, using differential conditioning paradigms, found some amygdala activation that was positively correlated with activations in the auditory cortex, the frontal cortex and the thalamus. An additional PET study used masked CSs and found significant amygdala activation with the right amygdala showing more activation with unaware (masked) and the left amygdala showing more activation with aware (unmasked) CS presentation. The notion that a direct pathway from the thalamus to the amygdala also exists in humans was recently supported by Morris et al. (1999) who showed that activation of the right amygdala in a masked conditioning paradigm was associated with activity in the superior colliculus and pulvinar but not with activity at cortical sites.

Several recent studies used functional magnetic resonance imaging (fMRI) to study fear conditioning in humans (for a review see Büchel and Dolan 2000). Probably due to the higher spatial and temporal resolution of this method compared to PET, amygdala activation was reported in all studies. In addition to the amygdala, the acquisition of a fear response created activation in the cingulate gyrus and the premotor cortex. Activation of the insulae has also been reported when painful stimuli were used as US. When trace conditioning was employed, activation in the hippocampus in addition to activation in the amygdala and cortical areas was noted.

Based on the observation that not all stimuli are equally easy to condition Seligman (1971) suggested that phylogenetically relevant stimuli such as snakes or spiders might serve as 'prepared stimuli' and be more conditionable than others. Subsequently, Öhman (for a summary see Öhman et al. 2000) has shown that conditioned skin conductance responses to a prepared CS such as an angry face or a picture of a snake persist for a longer time (even when the presentation is masked) than when a neutral stimulus such as a picture of a mushroom is used as CS and combined with electric shock as US. Moreover, masked presentations of these stimuli also lead to the acquisition of conditioned responses suggesting that unconscious learning of these stimulus associations is possible.

More recent accounts of fear conditioning have emphasized the role of cognitive variables as well as genetic predisposition in the acquisition of fear (e.g., Davey 1995, Kendler et al. 1992). Part of this may be a cognitive bias termed covariation bias that suggests that aversive outcomes are overestimated in the face of fear-relevant stimuli (e.g., Tomarken et al. 1989). However, although these cognitive variables have recently been discussed extensively as important mediations of the acquisition of fear, the biases themselves can be acquired through conditioning processes and it is not clear to what extent altered cognitive processing precedes or follows the acquisition of fear.

4. Genetic Factors

Genetic factors seem to play a role in the development in anxiety disorders in humans. In familial studies, twin studies and adoption studies show genetic variances of 20 to 40 percent. The same was found for questionnaire variables such as neuroticism. Fear conditioning ability and predisposition to learn fear responses were extensively studied in animals. In rodents, fear processes or anxiety measured in the open field test or avoidance conditioning have revealed a strong genetic basis explaining most of the interindividual variance. However, the genes responsible for that trait have not yet been identified.

The situation is different for antisocial personality disorder and criminal behavior. Assuming that a deficit in anticipatory fear conditioning is the central symptom from which all consequent social problems follow, all evidence ranging from twin to adoption studies and gene sequencing suggests a moderate to strong genetic influence. If sensitivity to alcoholism and novelty-seeking are included as comorbidity factors the genetic variance increases further. Whether this is related to genetic fear conditioning remains to be seen (Plomin et al. 2001).

5. Significance of Fear Conditioning for the Understanding of Anxiety Disorders

Although classical fear conditioning has been implicated as an important mechanism in the development of most anxiety disorders, studies that support

the etiological significance of fear conditioning have been rare. Retrospective data from either adult or child patients or the parents of children with specific phobias revealed that about 40–75 percent of the subjects or the parents could recall a traumatic event (e.g., Menzies and Clarke 1993), a substantial number of the remaining subjects experienced vicarious conditioning or had been verbally informed about potential dangers related to the feared object. In evaluating these data it must be kept in mind that all these reports are retrospective and that fear conditioning may be nondeclarative and therefore not be accessible to conscious recollection. Of greater relevance are therefore studies that examine fear conditioning in clinical groups or analogue samples.

In social phobics, several studies used differential conditioning with faces as CS and aversive odor or shock as US. In two studies, a conditioned response was already present prior to the acquisition phase with the social phobics showing enhanced amygdala and (in one study) orbitofrontal responses to the face CS suggesting that they displayed a fear response to neutral faces that never elicits amygdala activation in healthy humans (e.g., Birbaumer et al. 2000). In another study stronger amygdala activation was present to the CS+ than the CS− only in social phobics whereas the healthy controls showed deactivation. However, the scanning was performed late in the acquisition phase and it has been shown that the amygdala is only responsive early in the acquisition process. Additional studies that used conditioning paradigms or phobia-relevant cues during imaging also found activation increases in the anterior cingulate and the insulae in accordance with the findings of fear conditioning in healthy subjects.

In persons with psychopathy and antisocial personality disorder disturbed fear conditioning seems to be present with a concomitant complete absence of orbitofrontal, insular and cingulate activation during fear conditioning (Birbaumer et al. 2000). This absence of associative learning with respect to the aversive consequences of a stimulus might be the basis of the antisocial and often criminal behavior of these persons.

In posttraumatic stress disorder, higher conditionability as evident in enhanced and more difficult to extinguish fear responses has been observed. It is, however, not clear to what extent this condition predated the disorder or if it must be viewed as a consequence of the disorder and to what extent conditioning of trauma-related stimuli actually occurred because none of the studies involved trauma-relevant cues (e.g., Orr et al. 2000).

6. *Future Perspectives*

Future research must focus on the elucidation of the neuronal correlates of fear conditioning in healthy humans and patients. It is not clear to what extent the amygdala plays the same role in fear conditioning in humans as it does in rodents. Furthermore, the role of awareness of the CS and the CS–US relationship has not been determined by neuroimaging methods. The importance of fear conditioning for the understanding of anxiety disorders needs to be explored using paradigms that involve fear-relevant and prepared stimuli as well as longitudinal measurements. Finally, the role of cognitive and genetic variables and their relationship to fear conditioning needs to be further explored.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Classical Conditioning and Clinical Psychology; Classical Conditioning, Neural Basis of; Eyelid Classical Conditioning; Fear: Potentiation of Startle; Fear: Psychological and Neural Aspects

Bibliography

Büchel C, Dolan R J 2000 Classical fear conditioning in functional neuroimaging. *Current Opinion in Neurology* **10**: 219–23

Bechara A, Tranel D, Damasio H, Adophs R, Rockland C, Damasio A R 1995 Double dissociation of conditioning and declarative knowledge relative to the amygdala and the hippocampus in humans. *Science* **269**: 1115–18

Birbaumer N, Veit R, Flor H, Patrick C 2000 Excess and lack of fear: an fMRI study with social phobias and psychopaths. *Society for Neuroscience Abstracts* **26**: S1072

Bradley M M, Lang P J 2000 Measuring emotion: behavior, feeling and physiology. In: Lane R D, Nadel L (eds.) *Cognitive Neuroscience of Emotion*. Oxford University Press, New York, pp. 242–76

Davey G C 1995 UCS revaluation and conditioning models of acquired fears. *Behavior Research and Therapy* **27**: 521–8

Davis M 1992 The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience* **15**: 353–75

Fischer H, Andersson J L, Furmark T, Fredrikson M 2000 Fear conditioning and brain activity: a positron emission tomography study in humans. *Behavioral Neuroscience* **114**: 671–80

Kendler K S, Neale M C, Kessler R C, Heath A C, Eaves L J 1992 The genetic epidemiology of phobias in women: the interrelationship of agoraphobia, social phobia, situational phobia and simple phobia. *Archives of General Psychiatry* **49**: 273–81

LeDoux J E 2000 Emotion circuits in the brain. *Annual Review of Neuroscience* **23**: 155–84

Maren S, Fanselow M S 1993 The amygdala and fear conditioning: has the nut been cracked? *Neuron* **16**: 237–40

Menzies R G, Clarke J C 1993 The etiology of childhood water phobia. *Behavior Research and Therapy* **31**: 499–501

Morris J S, Öhman A, Dolan R J 1999 A subcortical pathway to the right amygdala mediating 'unseen' fear. *Proceedings of the National Academy of Science USA* **96**: 1680–5

Öhman A, Flykt A, Lundqvist D 2000 Unconscious emotion: evolutionary perspectives, psychophysiological data and neuropsychological mechanisms. In: Lane R D, Nadel L (eds.) *Cognitive Neuroscience of Emotion*. Oxford University Press, Oxford, pp. 296–327

Orr S P, Metzger L J, Lasko N B, Macklin M L, Peri T, Pitman R K 2000 De novo conditioning in trauma-exposed individuals with and without posttraumatic stress disorder. *Journal of Abnormal Psychology* **109**: 290–8
Plomin R, DeFries J, McClearen G, Roth M 2001 *Behavioral Genetics* 4th Edn. Freeman, San Francisco, CA
Rescorla R A 1988 Pavlovian conditioning: It's not what you think it is. *American Psychologist* **43**: 151–60
Seligman M E P 1971 Phobias and preparedness. *Behavior Therapy* **2**: 307–20
Tomarken A J, Mineka S, Cook M 1989 Fear-relevant selective associations and covariation bias. *Journal of Abnormal Psychology* **98**: 381–94
Weinberger N M 1998 Physiological memory in primary auditory cortex: characteristics and mechanisms. *Neurobiology of Learning and Memory* **70**: 226–51

H. Flor and N. Birbaumer

Fear: Potentiation of Startle

When people are afraid they tend to show a bigger startle response to a loud sound. Rats show a similar effect. This phenomenon, called 'fear-potentiated startle,' has been used to study the neural mechanisms involved in fear and anxiety in both rats and humans.

1. *Fear-potentiated Startle in Rats*

Capitalizing on anecdotal evidence that people startle more when they are afraid, Brown et al. (1951) demonstrated that the amplitude of the acoustic startle reflex in the rat can be augmented by presenting the eliciting auditory startle stimulus shortly after onset (e.g., 3s) of a cue (e.g., a light) previously paired with a footshock. This fear-potentiated startle occurs with an auditory, visual, or tactile conditioned stimulus (CS), and when startle is elicited by either a loud sound or an airpuff in rats (Davis et al. 1993). In this test a central state of fear is considered to be the conditioned response. Conditioned fear is defined operationally by elevated startle amplitude in the presence of a cue previously paired with a shock. Thus, the CS does not elicit startle. Furthermore, the startle-eliciting stimulus is never paired with a shock. Instead, the CS is paired with a shock, and startle is elicited by another stimulus in the presence or absence of the CS. Fear-potentiated startle only occurs following paired rather than unpaired or random presentations of the CS and the shock, indicating that it is a measure of conditioned fear. Discriminations between visual and auditory conditioned stimuli, and between auditory cues and visual cues that differ in duration, have also been demonstrated with potentiated startle. Pairing one cue with shock does not lead to generalization of fear-potentiated startle in testing with other cues. Fear-potentiated startle still occurs reliably at least one month after original training, making it appropriate for the study of long-term memory as well.

It has been suggested, however, that potentiated startle may not reflect increased fear in the presence of a CS, but instead results from the animal making a postural adjustment (e.g., crouching) which is especially conducive to startle, in anticipation of the impending footshock. However, spinally transected rats that were held rigidly in a modified stereotaxic instrument, which prevented obvious postural adjustments, still showed fear-potentiated startle, measured with the pinna component of startle. Potentiation of startle measured electromyographically in neck muscles also occurs in the absence of any obvious postural adjustment. In addition, the magnitude of potentiated startle correlates highly with the degree of freezing, a very common measure of fear.

1.1 *Effects of Different Drugs*

Drugs that reduce fear or anxiety in humans decrease potentiated startle in rats (Davis et al. 1993). Benzodiazepines, opiates, β-noradrenergic antagonists, drugs that decrease norepinephrine release, certain serotonin agonists, nicotine, and perhaps alcohol decrease or fully block fear-potentiated startle. In many cases these treatments do not depress startle levels in noise-alone trials, although this is not always the case. Conversely, drugs that increase norepinephrine release or decrease GABA (γ-aminobutyric acid) mediated neurotransmission and exaggerate anxiety in anxious people increase the magnitude of potentiated startle in rats.

1.2 *Neural Systems Involved in Fear-potentiated Startle*

A major advantage of the fear-potentiated startle paradigm is that fear is measured by a change in a simple reflex. Hence, fear is expressed through some neural circuit that is activated by the CS and ultimately impinges on the startle circuit. Figure 1 shows a schematic summary diagram of the neural pathways we believe are required for both the acquisition and expression of fear-potentiated startle in rats using a visual CS. These pathways involve convergence of the CS and the unconditioned shock stimulus at the lateral and basolateral amygdala nuclei, which project to the central nucleus of the amygdala, and this in turn projects directly to the startle pathway.

1.2.1 The primary acoustic startle pathway. Because the acoustic startle reflex has such a short latency (e.g., 8 ms measured electromyographically in the hindlegs), it must be mediated by a simple neural pathway. Using a variety of techniques, we have de-

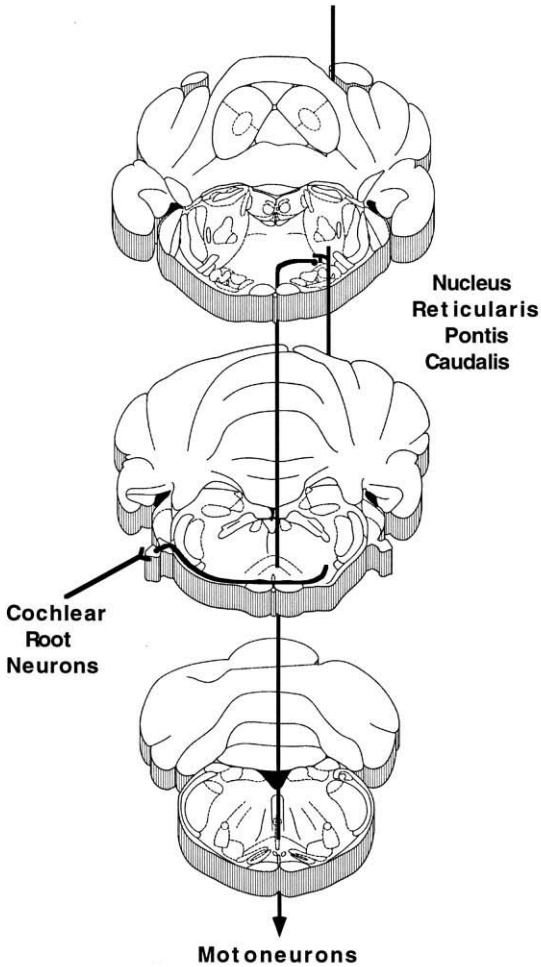

Figure 1
Schematic summary of the pathways believed to be involved in fear-potentiated startle in rats using schematic drawings of coronal brain slides in the rat (reproduced by permission of the Society for Neuroscience from *Journal of Neuroscience* 1996, **16**: 3775–89)

termined that the primary acoustic startle reflex in rats is probably mediated by three central synapses (Lee et al. 1996). The first involves terminals of the auditory nerve that synapse with large neurons embedded in the auditory nerve, called cochlear root neurons. These send exceedingly thick axons (sometimes as large as 7 μm in diameter) to neurons in the nucleus reticularis pontis caudalis (PnC). These cells project directly to motoneurons in the facial motor nucleus (pinna reflex and perhaps the eyeblink component of startle) and spinal cord (whole-body startle). Chemical lesions of these neurons or local infusion of glutamate antagonists in the vicinity of these synapses eliminated acoustic startle. Although other investigators have proposed that additional connections between the cochlear nucleus and the PnC may be important for startle (Koch and Schnitzler 1997), we believe that this simple, three-synapse pathway is the main acoustic startle pathway. For example, although the dorsal cochlear nucleus also projects directly to the PnC, lesions of the dorsal cochlear nucleus only decreased startle amplitude when high-intensity startle stimuli were used.

1.2.2 The role of the thalamus. Visual information uses parallel pathways that involve either (a) direct retinal inputs to the lateral posterior nucleus of the thalamus, which projects directly to the perirhinal cortex, or (b) retinal inputs to the dorsal lateral geniculate nucleus, which projects indirectly to the perirhinal cortex via the visual cortex. Lesions of both of these areas of the visual thalamus together, but not either one alone, blocked fear-potentiated startle using a visual CS (Shi and Davis 1996).

If an auditory CS is used, this involves parallel inputs from the auditory thalamus to the perirhinal cortex, either directly or indirectly via the auditory cortex (Campeau and Davis 1995a). Pre- or post-training lesions of the entire auditory thalamus completely blocked fear-potentiated startle to an auditory CS but not to a visual CS. Post-training lesions restricted to the main body of the medial geniculate, which projects to the perirhinal cortex via the auditory cortex, also specifically blocked fear-potentiated startle to the auditory CS.

Pain information reaches the amygdala via parallel pathways that include the caudal granular/dysgranular insular cortex and the posterior intralaminar nuclei of the thalamus. Pretraining lesions of both the insular cortex and the posterior intralaminar nuclei of the thalamus, but not lesions of either structure alone, blocked the acquisition of fear-potentiated startle (Campeau and Davis 1995a, Shi and Davis 1999). However, post-training lesions of both of these areas together did not prevent expression of conditioned fear. These results suggest that parallel cortical and subcortical pathways are involved in relaying shock information during fear conditioning.

1.2.3 The role of the anterior perirhinal and insular cortex. The perirhinal cortex, which receives either visual or auditory CS information, projects directly to the lateral and basolateral amygdala. Post-training lesions of the anterior perirhinal cortex completely blocked the expression of fear-potentiated startle using a visual CS (Rosen et al. 1992), provided the lesion destroyed both the dysgranular and agranular

portions of the perirhinal cortex. Post-training lesions of the perirhinal area (including secondary auditory cortices) blocked fear-potentiated startle to both auditory and visual CS (Campeau and Davis 1995a). However, reliable potentiated startle was observed after retraining in animals that had sustained main geniculate body lesions (which would destroy cortical connections between the thalamus and perirhinal cortex), or following pretraining lesions of the perirhinal area. These data suggest that cortical areas are normally used for the expression of fear conditioning, but that subcortical areas can take over if the cortex is damaged. Finally, as mentioned before, shock information seems to require the insular cortex, which in turn projects directly to the lateral nucleus of the amygdala.

1.2.4 The role of the amygdala. The perirhinal cortex projects directly to the lateral and basolateral amygdala. Selective destruction of cell bodies in the lateral and basolateral nuclei by local infusion of neurotoxins caused a complete blockade of fear-potentiated startle when the lesions were made, whether before or after training (Sananes and Davis 1992). This blockade did not seem to result from a disruption of vision, and other studies found that neurotoxic lesions of these amygdaloid nuclei also blocked fear-potentiated startle using an auditory CS. In contrast, lesions of a variety of other subcortical areas, such as the hippocampus, the septal area, the cerebellum, or the bed nucleus of the stria terminalis, had no consistent effect. The lateral, and especially the basolateral, nuclei of the amygdala project directly to the central nucleus of the amygdala, lesions of which blocked the expression of fear-potentiated startle using either a visual (Hitchcock and Davis 1986) or an auditory CS (Campeau and Davis 1995b).

Because CS and shock information converge neurally at the lateral and basolateral amygdala nuclei, this could be the site of plasticity for fear-potentiated startle. In fact, local infusion of NMDA (*N*-methyl-D-aspartate) antagonists into the amygdala blocked the acquisition, but not the expression, of fear-potentiated startle using either a visual or an auditory CS (Davis et al. 1993). This was probably not a result of shock information failing to reach the amygdala, because this treatment also blocked the acquisition of second-order fear-potentiated startle, which does not involve shock during second-order training.

By eliciting startle-like responses electrically from various points along the startle pathway before and after presentation of a light previously paired with a shock, we concluded that fear ultimately alters transmission at the PnC. Consistent with this, local infusion into the PnC of several compounds, such as glutamate antagonists and corticotropin-releasing hormone antagonists, blocked fear-potentiated startle at doses that have no effect on baseline startle (Koch and Schnitzler 1997).

The central nucleus of the amygdala projects directly to the PnC, and lesions at several points along this pathway blocked the expression of fear-potentiated startle. Low-level electrical stimulation of the amygdala markedly increased the acoustic startle amplitude (Rosen and Davis 1988) at currents and durations that did not produce any other signs of behavioral activation. Transit times from the amygdala to the startle circuit appear to be about 4–5 ms or even less. Electrical stimulation of the amygdala in anesthetized rats facilitated tone-evoked activation of cells in the PnC (Koch and Ebert 1993).

However, it is now clear that there is at least one critical synapse between the amygdala and the startle pathway, located in an area in or very close to the mesencephalic reticular formation, dorsal periaqueductal gray or deep white layers of the superior colliculus (Koch and Schnitzler 1997). Neurotoxic lesions of this area or local infusion of muscimol blocked the expression, but not the acquisition, of fear-potentiated startle.

2. *Fear-potentiated Startle in Humans*

Fear-potentiated startle can also be measured in humans by using the eyeblink component of the startle reflex. In one test, people are told that when a certain colored light comes on they might get a shock on the wrist, whereas they will not get a shock when a different colored light comes on. Startle is elicited with bursts of noise through earphones, and the eyeblink component of startle is measured electromyographically from the orbicularis oculi muscles. Startle amplitude was consistently higher in the presence of the light that signaled shock (Grillon et al. 1991). The size of this increase depended on the time when the subject expected the shock on the basis of verbal instructions, and also occurred when conditioning procedures were used.

Thus far, there is a close correspondence between results gathered in rats and humans. Brain imaging studies show that the amygdala is activated during this verbally mediated fear-potentiated startle test, and that patients with lesions of the amygdala fail to display fear-potentiated startle. People also show an increase in startle when they see scary pictures—for example, of a gun in their face, a dog about to bite them, or mutilated bodies (Lang et al. 1990). The size of the increase in startle was related directly to how the subjects rated the pictures in terms of negative valence and arousal.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Autonomic Classical and

Operant Conditioning; Cardiovascular Conditioning: Neural Substrates; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Eyelid Classical Conditioning; Fear Conditioning; Fear: Psychological and Neural Aspects

Bibliography

Brown J S, Kalish H I, Farber I E 1951 Conditional fear as revealed by magnitude of startle response to an auditory stimulus. *Journal of Experimental Psychology* **41**: 317–28

Campeau S, Davis M 1995a Involvement of subcortical and cortical afferents to the lateral nucleus of the amygdala in fear conditioning measured with fear-potentiated startle in rats trained concurrently with auditory and visual conditioned stimuli. *Journal of Neuroscience* **15**: 2312–27

Campeau S, Davis M 1995b Involvement of the central nucleus and basolateral complex of the amygdala in fear conditioning measured with fear-potentiated startle in rats trained concurrently with auditory and visual conditioned stimuli. *Journal of Neuroscience* **15**: 2301–11

Davis M, Falls W A, Campeau S, Kim M 1993 Fear-potentiated startle: A neural and pharmacological analysis. *Behavioural Brain Research* **58**: 175–98

Grillon C, Ameli R, Woods S W, Merikangas K, Davis M 1991 Fear-potentiated startle in humans: Effects of anticipatory anxiety on the acoustic blink reflex. *Psychophysiology* **28**: 588–95

Hitchcock J M, Davis M 1986 Lesions of the amygdala, but not of the cerebellum or red nucleus, block conditioned fear as measured with the potentiated startle paradigm. *Behavioral Neuroscience* **100**: 11–22

Koch M, Ebert U 1993 Enhancement of the acoustic startle response by stimulation of an excitatory pathway from the central amygdala/basal nucleus of Meynert to the pontine reticular formation. *Experimental Brain Research* **93**: 231–41

Koch M, Schnitzler H-U 1997 The acoustic startle response in rats—circuits mediating evocation, inhibition and potentiation. *Behavioural Brain Research* **89**: 35–49

Lang P J, Bradley M M, Cuthbert B N 1990 Emotion, attention, and the startle reflex. *Psychological Review* **97**: 377–95

Lee Y L, Lopez D E, Meloni E G, Davis M 1996 A primary acoustic startle pathway: Obligatory role of cochlear root neurons and the nucleus reticularis pontis caudalis. *Journal of Neuroscience* **16**: 3775–89

Rosen J B, Davis M 1988 Enhancement of acoustic startle by electrical stimulation of the amygdala. *Behavioral Neuroscience* **102**: 195–202

Rosen J B, Hitchcock J M, Miserendino M J D, Falls W A, Campeau S, Davis M 1992 Lesions of the perirhinal cortex but not of the frontal, medial prefrontal, visual, or insular cortex block fear-potentiated startle using a visual conditioned stimulus. *Journal of Neuroscience* **12**: 4624–33

Sananes C B, Davis M 1992 *N*-Methyl-D-aspartate lesions of the lateral and basolateral nuclei of the amygdala block fear-potentiated startle and shock sensitization of startle. *Behavioral Neuroscience* **106**: 72–80

Shi C, Davis M 1996 Anatomical tracing and lesion studies of visual pathways involved in fear conditioning measured with fear potentiated startle. *Society for Neuroscience Abstracts* **22**: 1115

Shi C-J, Davis M 1999 Pain pathways involved in fear conditioning measured with fear-potentiated startle: Lesion studies. *Journal of Neuroscience* **19**: 420–30

M. Davis

Fear: Psychological and Neural Aspects

Fear is an innate behavioral–cognitive–emotional system that evolved because of its evolutionary success in protecting animals (including humans) from danger. While fear to certain kinds of stimuli is unlearned, fear can also be acquired rapidly and lastingly to other kinds of stimuli, allowing animals to respond adaptively to new environmental threats. This entry will present some of the behavioral and physiological aspects of fear that make it a useful model system for understanding the psychological and neurobiological mechanisms of emotion, learning, and memory, as well as the development of a variety of human fear-related disorders.

1. Fear—Definition and Background

All organisms are subject to threats and dangers from their environments, such as predation, dominance, and territoriality. Through the process of natural selection, evolution has endowed all organisms with defensive mechanisms (e.g., chemical toxins in plants; fight-flight responses in animals) for self-preservation. Fear serves this same purpose. In humans and other higher functioning animals, 'fear' as a scientific term describes a behavioral–cognitive–emotional state in which a set of adaptive responses is activated in the presence of danger. As the term is more commonly used, 'fear' refers to a sensation (or feeling) of strong unpleasant emotion aroused by stimuli or events that are perceived as threatening or dangerous.

While this more anthropomorphic (psychological) sense of the word 'fear' may be useful in describing a person who is experiencing an environmental threat (e.g., electric shock) and exhibiting certain stereotyped behaviors (such as a fearful facial expression, autonomic arousal, behavioral readiness, and other defensive responses), it is probably inaccurate to use the term to describe a goldfish (or any other lower-functioning animal) that is exhibiting similarly stereotyped behaviors in response to electric shock (such as avoidance or escape). In this entry, fear will be defined as a by-product (or an emergent property) of a coordinated pattern of brain activities and bodily changes that appear in response to threats and dangers. Thus, fear serves an adaptive function, but evolved relatively recently among the continuum (or phylogeny) of defensive mechanisms.

Fear as an emotion has been a topic of interest to many subdisciplines of psychology, including psychoanalysis, behaviorism, cognitive psychology, and

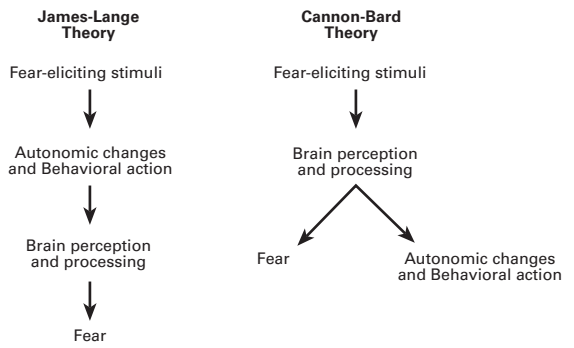

Figure 1
Two influential theories of emotion propose different sequences by which threats and danger trigger fear. The main tenet of the James–Lange theory is that the animal feels fear after its body reacts (emphasizing bodily feedback in the brain representation of fear). In contrast, the Cannon–Bard theory states that threats and danger simultaneously activate the brain representation of fear and the bodily responses (bodily feedback is not a precursor of the brain representation of fear)

biological psychology (for detailed treatment and further references, see LeDoux 1996). Historically, two opposing theories of emotion, the James–Lange theory and the Cannon–Bard theory, have been especially influential (Fig. 1). According to the James–Lange theory, threatening or dangerous stimuli trigger physiological arousal and behavioral responses (via the autonomic nervous system), which are perceived and then processed as fearful by the brain (emphasizing the importance of bodily feedback to the brain). In contrast, the Cannon–Bard theory posits that threats and dangers simultaneously trigger brain activities representing fear and initiate physiological arousal and behavioral readiness (emphasizing the importance of the brain activity). Thus, the critical distinction between the two theories is whether the appearance of immediate danger (e.g., a bear) is postulated to activate the central fear system in parallel with fleeing and autonomic responses (Cannon–Bard theory), or to first trigger fleeing and autonomic responses, which are then detected and interpreted as fear by the brain (James–Lange theory). There is evidence in support of both theories, suggesting that central processes and bodily feedback (as well as other, cognitive processes) both contribute to the emergence of fear as an emotion.

2. *Neuronal Substrates of Fear*

Because humans share many biological and behavioral features with other animals, experimental studies with nonhuman animals provide valuable information

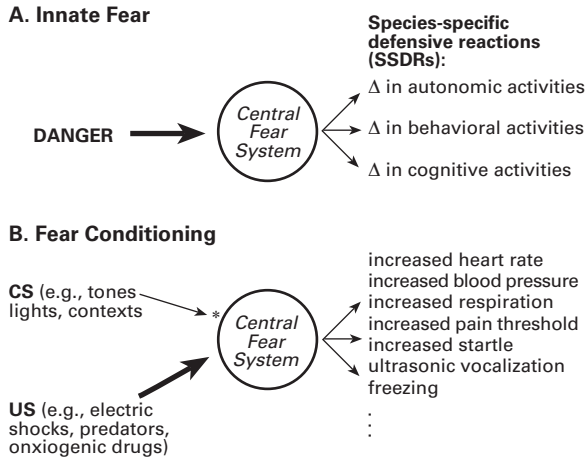

Figure 2
Simple models of fear system and fear conditioning. A. Innate fear: presence of danger triggers a coordinated set of programmed behavioral physiological and cognitive responses, which rapidly prepares the animal to contend with the immediate threat. B. Fear conditioning: fear of novel stimuli can be acquired through the process of classical conditioning. By virtue of association with the US, the CS comes to activate fear
* The putative locus of neural plasticity involved in fear conditioning

toward understanding the physiological as well as the psychological aspects of fear. Darwin postulated in *The Expression of Emotions in Man and Animals* (1872) that emotions such as fear are innate, highly specific to situations, and the product of coordinated brain activities, and that different animal species with diverse environmental demands (or histories) have evolved dissimilar specialized sets of fear (or defensive) responses to maximize survival. It follows from this line of reasoning that more related species will exhibit more similar fear responses than will less related species, whose nervous systems share fewer characteristics. Thus, it is highly likely that the neuronal substrates underlying fear (or other defensive mechanisms) in humans are more similar to those of other mammals (e.g., chimpanzees and rats) than those of nonmammalian animals (e.g., fish and birds).

A repertoire of innate fear responses that is characteristic of a particular animal species is often referred to as species-specific defensive reactions (SSDRs) (Bolles 1970) (Fig. 2, A). For instance, the universal expressions of fear in humans include particular facial expressions (seen in members of all cultures), changes in autonomic nervous system activities (e.g., increased heart rate, increased blood pressure, rapid respiration, perspiration, elevated pain threshold, pupil enlargement, and decreased digestive activity) and beha-

vioral readiness (e.g., freezing, increased propensity to startle, and sharpening of cognition) (see LeDoux 1996). These fear responses in humans are 'hardwired' (innate) as indicated by the observation that newborn infants, when exposed to aversive stimuli (e.g., a loud noise) for the very first time, exhibit responses, such as sudden eye closure, rapid breathing, clutching of hands, and crying, which are indicative of a general state of fear (e.g., Watson 1924). The innateness of the fear response is also demonstrated in animals; for example, laboratory-reared rats confronted with danger (e.g., a cat) for the very first time exhibit changes in autonomic nervous system activities and behavioral readiness (responses that are remarkably similar to those seen in humans), and ultrasonic (22 kHz) distress vocalization (a response that is an example of an SSDR) (e.g., Blanchard and Blanchard 1971, Blanchard et al. 1991, Bolles 1970).

Several lines of evidence point to temporal lobe structures (specifically the amygdala, an almond-shaped group of nuclei buried deep within the temporal lobes) as the key neural system underlying the central state of fear (see LeDoux 1996). The classic studies of Klüver and Bucy (e.g., 1937), who examined the effect of temporal lobe ablations in rhesus monkeys, were pioneering in this respect. Klüver and Bucy noted dramatic postlesion behavioral changes, including hyperactivity, hypersexuality, and a so-called 'psychic blindness' that was manifested as an apparent indifference to the significance of previously frightening and/or arousing stimuli. While the monkeys were wild and aggressive prior to surgery, they were remarkably tame and agreeable to human contact afterwards, suggesting that a crucial emotive brain center had been removed or significantly affected by the lesions. Subsequent work, such as that of MacLean and Delgado (1953) using focal stimulation of temporal lobe structures in awake, behaving monkeys, pointed to the amygdala as the critical site. The importance of the amygdala in fear has been reconfirmed in numerous studies employing diverse species of animals (e.g., rats, mice, cats, rabbits, monkeys) and various experimental techniques (e.g., lesions, stimulation, drug infusions, recordings), and most current theories of amygdalar function reflect the importance of the structure in the processing of emotional stimuli (see LeDoux 1996). Consistent with this view, recent human neuropsychological and brain-imaging studies have found that patients with amygdalar damage are selectively impaired in the recognition of facial expressions of fear (Adolphs et al. 1994), and that there is a significantly increased blood flow to the amygdala (as measured by functional magnetic resonance imaging, fMRI) when normal subjects are presented with pictures of fearful faces (Morris et al. 1996).

The amygdala, as one of the principal structures of the limbic system, has access to sensory input from diverse areas of the brain (e.g., olfactory cortex, hippocampus, thalamus, hypothalamus, neocortex), and projects to various autonomic and somatomotor structures that mediate specific fear responses (e.g., bed nucleus of stria terminalis for activating stress hormones, periaqueductal gray matter for defensive behavior, lateral hypothalamus for sympathetic activation) (see LeDoux 1996).

Given this pattern of connectivity, it is not surprising that damage to the amygdala eliminates stereotypical fear responses, such as defensive freezing (characterized as a motionless, crouching posture that functions to reduce the probability of detection by predators) exhibited by rodents in the presence of predators (Blanchard and Blanchard 1971). Nevertheless, while the amygdala is apparently an important component of the central fear-processing system, it does not necessarily follow that the amygdala is the seat of fear as an emotion (or a feeling). Given its intricate reciprocal connections with the neocortex, a structure implicated in higher cognitive functioning, it is very likely that the interactions between amygdala and neocortex (particularly sensory and limbic association cortices) are responsible for the emotion of fear. Thus, although the amygdala is essential in activating various fear responses (or SSDRs) in both humans and rats, the fact that there are great differences in the neocortices of humans and rats suggest that perhaps there might be important differences in fear as experienced by the two species.

3. *Acquired Fear*

While humans and other animals are genetically predisposed to fear certain stimuli/events, it is also adaptive for animals to have the capacity to learn about new dangers in their environments. For instance, although newborn infants innately exhibit fear to certain stimuli (e.g., loud noises), they do not show innate fear to flame or height (two stimuli that most children learn and adults remember to avoid). Thus, fear responses to many stimuli and events in the environment appear to be acquired.

Classical or Pavlovian fear conditioning has been widely employed for studying the mechanism(s) by which fear is acquired. Fear conditioning occurs when initially neutral conditioned stimuli (CSs) such as tones, lights, or distinctive environmental settings, are contingently paired with aversive unconditioned stimuli (USs) such as electric shocks, which reflexively activate unconditioned fear responses (URs). Through CS–US association formation, the CS comes to elicit various conditioned fear responses (CRs) that are similar to innate fear responses (Fig. 2, B). Perhaps the best known example of fear conditioning is the Little Albert experiment by Watson and Rayner (1920). Little Albert was an 11-month-old infant who initially exhibited curiosity (no fear) to a white rat by touching and playing with it. Watson and Rayner attempted to

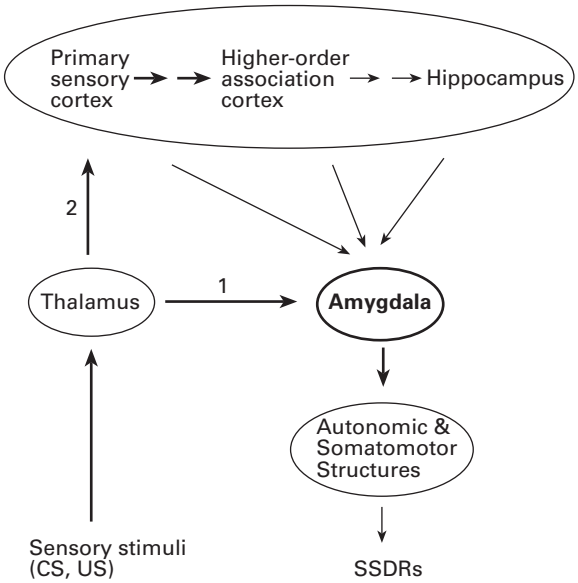

Figure 3
A simplified putative fear-conditioning circuit. Conditioned stimuli (e.g., tones) are processed via two separate pathways to the amygdala (where fear conditioning is thought to take place). One pathway is via the direct (1) thalamo-amygdalar projections. The second pathway is via the indirect (2) cortico-amygdalar projections
Source: Adapted from Armony and LeDoux 1997

condition Albert to fear the rat by striking a steel bar with a hammer behind his head (US) as his hand touched the rat, causing him to startle, fall forward, and cry (UR). A week later, when the rat (CS) was placed near Little Albert's hand, he withdrew his hand and began to cry (CR). This fear of the rat was also said to be observed in the presence of other white furry animals and objects (e.g., rabbits, dogs, fur muffs).

Modern investigations of fear conditioning typically employ rats as subjects and use a tone as a CS and a mild footshock as an US. Under these circumstances a single CS–US pairing results in robust (rapid and enduring) fear learning, as evidenced by a variety of conditioned fear responses exhibited upon subsequent presentations of the tone. Thus, the biological significance of fear conditioning is reflected by the quick and permanent nature of changes in the behavior toward the initially neutral stimulus that has been paired with an aversive experience.

Fear can also be rapidly acquired through instrumental or operant conditioning, in which the presentation of an aversive stimulus is contingent upon the behavior of the animal. A widely employed procedure with rodents is the passive (or inhibitory) avoidance task (see McGaugh et al. 1996). In this paradigm, the animal's response (e.g., entering a dark compartment of a box when placed in an adjacent lighted compartment, or stepping down from a platform onto a grid floor) is paired with an aversive experience (e.g., a footshock). As a function of this response–stimulus pairing, the animal learns to avoid making the response that was followed by the aversive experience.

Similar to innate fear, evidence from lesion, pharmacological, recording, and stimulation studies indicates that the amygdala is essential not only to innate fear, but also to the acquisition of fear, as with fear conditioning (see Lavond et al. 1993, LeDoux 1996). Thus, amygdalar lesions (particularly lesions of the central nucleus) impair the acquisition (learning) and retention (performance) of various fear-related behaviors in rats, rabbits, and cats; neurons in the central nucleus of the amygdala are responsive to both conditioned and innate fear stimuli and undergo plastic changes during fear conditioning; and restricted stimulation of the amygdala elicits specific fear responses (e.g., freezing). Recent findings from human studies that patients with amygdalar damage are impaired in fear conditioning, and that blood flow to the amygdala is increased during fear conditioning in normal subjects, further add to the evidence that the amygdala is crucially involved in fear learning.

One hypothesis of amygdalar function in fear conditioning is presented schematically in Fig. 3. In this model, an emotional stimulus (such as a tone that has been paired with a footshock) is presented, and its sensory characteristics are processed by way of two parallel pathways that ultimately converge in the amygdala. One of these pathways is 'direct' and provides relatively 'raw,' unprocessed information about the stimulus, but does so very quickly. If the neocortex is the seat of consciousness, this thalamic pathway to the amygdala can be considered the subconscious (nondeclarative or implicit) fear route. The second pathway is 'indirect' and provides highly processed information about the stimulus, but does so at the cost of an increased processing time (LeDoux 1996).

This cortical (conscious, declarative, or explicit) pathway consists of projections through primary auditory cortex to higher-order auditory association areas and the hippocampal formation, which then project to the amygdala. Upon reaching the amygdala (specifically the lateral/basolateral nuclei), these two converging representations of the stimulus are subjected to additional processing within the amygdala itself. Most current theories assume that the essential plasticity underlying this form of learning occurs in the amygdala and corresponds to a neural instantiation of the association between the CS (the tone) and the US (the footshock) long assumed by psychologists to lie at the heart of classically conditioned behavior. The precise functions of the inputs to and outputs from the amygdala, as well as the intraamygdaloid processing of CS- and US-related information, remain a topic of intensive research.

4. Fear Conditioning and Clinical Implications

Fear conditioning serves as a model system for understanding not only the neural basis of acquired fear in nonhuman animals, but also the etiology of a variety of fear-related disorders in humans (e.g., anxiety, phobias, post-traumatic stress disorder, and panic disorder). Since the characteristic behavioral and physiological symptoms of human anxiety disorders closely resemble many of the fear responses observed in fear conditioning paradigms, it is possible that anxiety and fear are closely related (Table 1). (It is important to note, however, that the fact that the symptomatic manifestation of anxiety disorders is similar to the fear responses observed in fear conditioning does not necessarily mean that the underlying neuronal substrates of fear and anxiety are the same (*post hoc*, *ergo propter hoc*).) While anxiety disorders can arise unpredictably, it is known that most patients experience their first symptoms in early adult life, indicating that maladaptive anxiety may be at least partly acquired. Indeed, clinicians from the time of Freud have speculated that anxiety patients have acquired their disorder as a result of traumatic learning experiences. Since the paradigm *par excellence* of traumatic learning is fear conditioning, fear conditioning appears to be useful in understanding the genesis of anxiety disorders (Fig. 3).

The question of whether anxiety is analogous to fear conditioning might be one of semantics; the difference is between normal, adaptive fear and abnormal, maladaptive fear (anxiety, phobias). However, there is an important distinction between anxiety disorders and fear conditioning (e.g., Seligman 1971) that might be taken to suggest that the two have very different bases. That is, while conditioned fear responses may be extinguished by, for example, actively preventing an animal from making escape or avoidance responses to a danger-signaling stimulus, fear-related disorders in humans (e.g., phobias) are quite resistant to extinction. In fact, anxiety disorders are characterized by fears that the patients recognize to be irrational but are unable to control.

Based on Fig. 3, it is interesting to speculate that fear-related disorders (irrational fears) might be due to

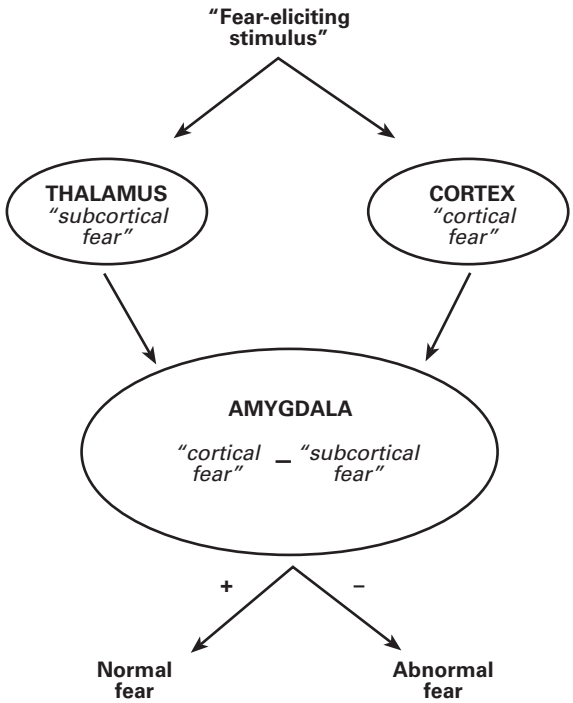

Figure 4
A simple model of normal and abnormal fear based on the putative circuitry of cortical and subcortical fear-conditioning pathways. The subcortical fear route provides a quick but imprecise representation of sensory information for activating fear (a nonconscious pathway), whereas the cortical fear route provides relatively slower but more detailed information for activating fear (a conscious pathway). In normal fear expression there is a balance between the cortical fear and subcortical fear pathways, such that the cortical component is stronger than or equal to the subcortical component. Abnormal fear might arise when this balance is disrupted such that the activity of the subcortical fear system overwhelms the activity of the cortical fear system

Table 1
Behavioral and physiological responses associated with anxiety disorders and fear conditioning

Anxiety disorders	Fear conditioning
Palpitations; pounding heart, or accelerated heart rate	Increased blood pressure, tachycardia
Sweating	Increase in Galvanic skin response measures
Sensations of shortness of breath or smothering	Increase in respiration, quick, shallow breathing
Nausea or abdominal distress	Ulcer
Paresthesia (numbness or tingling sensations)	Hypoalgesia or analgesia
Exaggerated startle responses	Potentiated startle

unbalanced activities in the cortical and thalamic pathways to the amygdala. For example, in most people a snake seen behind a glass window will elicit a different response than will a snake encountered in the woods. Both cortical and subcortical routes will process information about the snake, but the cortical (conscious) route will integrate the contextual information of the glass window and assess the safety of the situation (possibly acting as a gate). However, if the cortical pathway is somehow disrupted, the subcortical activity will be left unchecked to activate fear, even in the inappropriate context. It is possible that this type of unregulated subconscious activation of fear contributes to the etiology of fear-related disorders.

Figure 4 illustrates a hypothetical model of normal, adaptive fear and abnormal maladaptive fear. It is conceivable that genetic predisposition and environmental factors (e.g., stress) interact in such a way as to disrupt the normal balance between cortically activated fear and subcortically activated fear, leading to fear-related disorders. If it is then supposed, as some have suggested (LeDoux 1996), that the direct thalamic inputs to the amygdala are relatively resistant to extinction, whereas the cortical inputs are not, it follows that fear disorders based on overactivity in the subcortical pathway should be 'irrational,' in the sense that they do not subside even when the patient recognizes them as such. A more comprehensive understanding of the neuronal substrates and, subsequently, the molecular mechanisms of the fear system will be useful in understanding and developing treatments for fear-related disorders.

5. *Conclusion*

The study of fear, a form of emotion that evolved relatively recently among the phylogeny of successful defensive mechanisms, is but one means of understanding emotion, but it is the most fruitful one to have been employed to date. This is likely due to the primacy of fear: unlike other emotions, fear is apparently shared by a wide variety of species. This ubiquity speaks to its importance as a mechanism for self-preservation.

See also: Amygdala (Amygdaloid Complex); Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Emotion, Neural Basis of; Emotions, Evolution of; Fear Conditioning; Fear: Potentiation of Startle

Bibliography

Adolphs R, Tranel D, Damasio H, Damasio A 1994 Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372**: 669–72

Armony J L, LeDoux J E 1997 How the brain processes emotional information. *Annals of the New York Academy of Science* **821**: 259–70

Blanchard R J, Blanchard D C 1971 Defensive reactions in the albino rat. *Learning and Motivation* **2**: 351–62

Blanchard R J, Blanchard D C, Agullana R, Weiss S M 1991 Twenty-two kHz alarm cries to presentation of a predator, by laboratory rats living in visible burrow systems. *Physiology and Behavior* **50**: 967–72

Bolles R C 1970 Species-specific defensive reactions and avoidance learning. *Psychological Review* **71**: 32–48

Darwin C 1872 *The Expression of the Emotions in Man and Animals*. Murray, London

Klüver H, Bucy P C 1937 'Psychic blindness' and other symptoms following bilateral temporal lobectomy in rhesus monkeys. *American Journal of Physiology—Regulatory Integrative and Comparative Physiology* **119**: 352–3

Lavond D G, Kim J J, Thompson R F 1993 Mammalian brain substrates of aversive classical conditioning. *Annual Review of Psychology* **44**: 317–42

LeDoux J 1996 *The Emotional Brain: the Mysterious Underpinning of Emotional Life*. Simon and Schuster, New York

MacLean P D, Delgado J M R 1953 Electrical and chemical stimulation of frontotemporal portion of limbic system in the waking animal. *EEG Clinical Neurophysiology* **5**: 91–100

McGaugh J L, Cahill L, Roozendaal B 1996 Involvement of the amygdala in memory storage: Interaction with other brain systems. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 13508–14

Morris J S, Frith C D, Perrett D I, Rowland D, Young A W, Calder A J, Dolan R J 1996 A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* **383**: 812–15

Seligman M E P 1971 Phobias and preparedness. *Behavioral Therapy* **2**: 307–20

Watson J B 1924 *Behaviorism*. Norton, New York

Watson J B, Rayner R 1920 Conditioned emotional reactions. *Journal of Experimental Psychology—Human Perception and Performance* **3**: 1–14

J. J. Kim and K. M. Myers

Feature Representations in Cognitive Psychology

One of the most basic assumptions of cognitive psychology is that entities in the world can be described in terms of simpler components called features. For example, when asked to list the features that tend to characterize a bird, the typical undergraduate would mention that a bird has wings, flies, lays eggs, builds nests, has a beak, has feathers, and so on. This article discusses psychological research on how people represent features, and examines a number of challenges to developing a theory of feature representation.

1. *Introduction*

Cognitive psychologists have often attempted to characterize the entities of a domain in terms of a

relatively small set of such features, with different entities having different combinations of these features. To take a simple example, consider capital letters. Each capital letter can be described uniquely by some combination of seven basic features: vertical, horizontal, or oblique lines; continuous or discontinuous curves; and right or acute angles between the junctions of lines. For instance, the letter R has a vertical line, two horizontal lines, an oblique line, a discontinuous curve, and three right angles. In contrast, the letter O has a single feature—a continuous curve.

Being able to characterize the entities of a domain in this manner would be an important step in specifying how entities are identified. For example, suppose that in a person's long-term memory, capital letters were represented as combinations of basic features like those described above. Then by using processes that identified such features and matched them to ones represented in long-term memory, a person could identify a letter. For example, upon detecting an oblique line and a continuous curve, a person would determine that these features matched those of the representation of Q and identify the letter as a Q. Evidence suggests that the human visual system has feature detectors (i.e., neurons). For example, some neurons respond to lines of different orientations and to junctions of lines (i.e., corners) (Hubel and Wiesel 1965).

Identifying entities is the starting point for many important, everyday cognitive tasks. In this example, identifying the letter would be an important step in identifying a word, which in turn would be an important step in reading. As another example, by identifying an object, we gain access to a large amount of stored information that is useful in interacting with that object. For instance, by recognizing something as a bear you can then access stored knowledge which lets you anticipate things about the bear that may not be immediately present (e.g., that the bear may approach and attack you).

In many domains, the representation (and subsequent processing) of features is considerably more complex than implied by the example of capital letters. In contrast to capital letters, it is not always clear what features characterize an entity. In addition, multiple factors often determine how a feature is identified. Further, relations between features can be important for identifying entities. Finally, people not only identify features but must also learn new ones. In the remainder of this article, I discuss these and other issues related to feature representation and processing.

2. Context-dependent Feature Representations

The features that we use to represent entities in the world are often influenced by the surrounding context (what other information is present at encoding, our expectations, etc.). A classic example is illustrated by a study involving the rat-man figure (see Fig. 1). Bugelski and Alampay (1961) found that when preceded by several pictures of human faces people tend to see the rat-man figure as a man, but when preceded by several pictures of animals they tend to see the rat-man figure as a rat. In these cases, people represent the same perceptual input in terms of different features. For example, the small circles in the figure are seen as ears or as eyeglasses depending on the preceding pictures. As another example, whether a white area on an X-ray indicates a lung tumor, a bone, or an artifact of the procedure can depend on a variety of contextual factors, including the radiologist's knowledge of the patient's case history, the location of bones, whether the X-ray is a chest X-ray, and the likelihood that the X-ray was underexposed. These and many other examples suggest that the representation of a feature is not just based on perceptual input but also on what is in the head of the perceiver. As result, the features associated with a perceptual input can vary depending on the perceiver's prior knowledge.

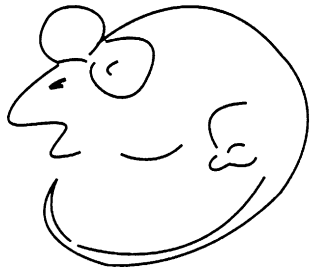

Figure 1
The rat-man figure used by Bugelski and Alampay (1961)

3. Relations between Features

Feature representations often capture relations between other features. These relations are necessary in order to identify an entity accurately. For example, a bucket and a coffee cup can both be described by the same two basic features (an open cylinder and a handle). Thus, it would not be possible to distinguish these objects if they were only represented by such features. Note, though, that the spatial relationship between the two features differs in the two objects. That is, the handle of the bucket is attached to the rim of the open cylinder, whereas the handle of the coffee cup is attached to the side of the open cylinder. Thus, people could distinguish these objects if their featural representations included such spatial relationships. Biederman (1987) describes a model of object recognition that represents objects in terms of part-like features (called geons), and the relations between them. As another example, the words 'team' and 'meat' are made up of the same basic features, called

phonemes (i.e., the smallest significant sound units in a language). However, the temporal relationship between these phonemes is different. For example, the 't' sound is pronounced first in 'team,' but last in 'meat.' Thus, people could distinguish these spoken words if their featural representations included such temporal relationships.

4. Construction of New Features

In a number of areas of cognitive psychology, the researcher and subject agree implicitly on the features which characterize the cognitive task being investigated. For example, in studying how people learn about new categories of entities, researchers often present subjects with stimuli whose features are well-specified, familiar, and unambiguous. The question of interest is how people determine which features are important for distinguishing one category from another. However, learning does not just involve selecting relevant features. In any domain, people must sometimes figure out the features themselves (i.e., acquire new features; see Biederman and Shiffrar 1987, Lesgold et al. 1988, Schyns et al. 1998, Wisniewski and Medin 1994, for examples).

To illustrate this point, consider a study by Schyns and Rodet (1997) in which subjects learned about categories of novel stimuli called 'Martian cells' (see Fig. 2). One group of subjects first learned a category whose members had *x* blobs followed by a category with *y* blobs, and then a category with *xy* blobs (a fusion of the separate *x* and *y* blobs; see Fig. 2a). A second group learned the categories in the reverse order (see Fig. 2b). These different orders affected how subjects determined the features of the category with *xy* blobs. The first group perceived the *xy* blob as a conjunction of the *x* and *y* features. They had previously learned that *x* was diagnostic of the *x* blobs category, and that *y* was diagnostic of the *y* blobs category. Thus, previously learning the *x* and *y* features biased the first group to segment the *xy* blob into these two features. In contrast, the second group perceived the *xy* blob as a single, unitary feature *xy*. Because the initial category that they learned was the one with *xy* blobs, their perception of the *xy* blob was not influenced by having learned features *x* and *y*.

5. Local vs. Global Feature Representations

In many areas of cognitive psychology, researchers use symbols to represent features that correspond closely to language, and which tend to bottom out at the level at which the language bottoms out (Solomon and Barsalou in press). For example, in the English language, people typically use the word 'stripes' to refer to the stripes of zebras, barber poles, the American flag, and so on, rather than a more detailed term. In turn, researchers represent features using similar, language-like terms. Often, these representations take the form of attribute-value pairs. For

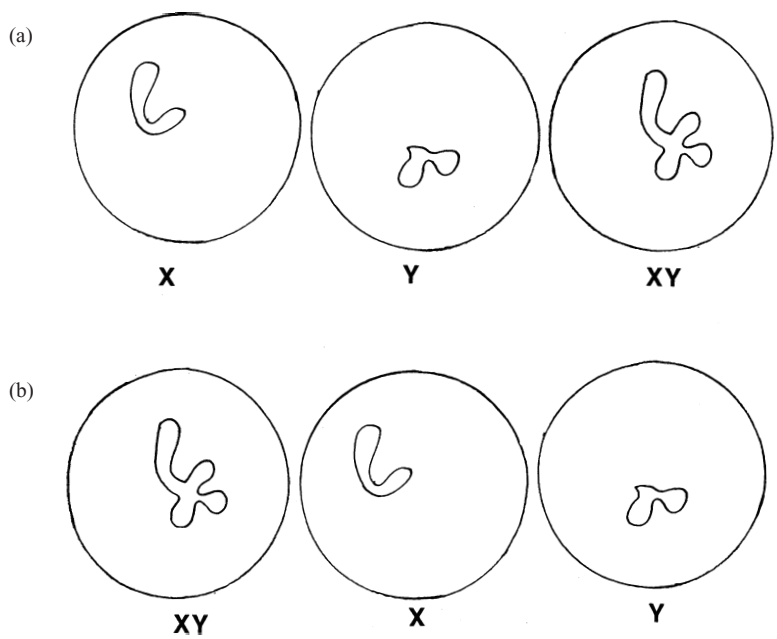

Figure 2
Schematic drawing of two different orders of novel categories learned by subjects in a study by Schyns and Rodet (1997)

example, stripes could be represented as the attribute texture. Do researchers believe that such representations have no further specification of meaning? They probably do not, as intuition strongly suggests that there is something more to such representations. Instead, such a feature representation may be a shorthand term for the commonalities shared by its referents (which researchers usually leave unspecified). For example, one possible commonality among the stripes of entities is that they have a color which contrasts with the background color of the entity. Thus, researchers generally assume that a feature has a similar representation across the entities to which it applies (i.e., a global representation).

However, evidence suggests that feature representations are also more fine-grained and specific or local to their associated entities. For example, Solomon and Barsalou (in press) tested the local representation view in a series of property verification tasks. Subjects saw an object (e.g., a bus) followed a property name (e.g., seat) and had to verify whether the property was true of the object. On a subsequent trial, verifying that property for a different object was facilitated if the property had the same form in the second object. For example, having previously verified that seat is true of bus, subjects more quickly verified that seat was true of truck. However, this benefit was not obtained when verifying that seat was true of bicycle. This finding implies that the representation of seat in truck and bus is similar, but differs from that in bicycle. As another example, many properties are relative. For instance, a large mouse and a large house are not very similar in size. A large mouse is large compared to other mice, but a large house is large compared to other houses. Thus, relative properties also have local representations.

6. Determining the Appropriate Features

As much of the discussion above implies, a central issue in feature representation is determining the appropriate features of an entity. In principle, one can describe any entity by an infinite number of features (e.g., a German shepherd dog has 4 legs, but it also has fewer than 6 legs; it has a tail, but also a tail of a particular length; it is found on the planet Earth, but not at the bottom of the ocean; it has a nose, but it also has a nose of a particular color and shape, and so on). Clearly, there must be constraints on what counts as a feature (Murphy and Medin 1985). Many of these constraints are provided by the perceptual system (see Barsalou 1999, Schyns et al. 1998, for discussion of perceptual mechanisms that extract features). For example, the perception of an object part is influenced by discontinuities and points of maximum curvature on its surface (e.g., Hoffman and Richards 1984).

Other constraints are related to the categories to which an entity belongs. For example, a small, umbrella-shaped entity growing in a forest might be categorized as a mushroom. The representation of mushroom might contain the features stem and umbrella-shaped cap so as to distinguish it from other plants. On the other hand, a mushroom belongs to a more specific category such as Death Cap—the most poisonous mushroom. Its representation might contain the features green-tinged cap and thick stem, so as to distinguish it from other mushrooms. Thus, the features associated with a category may be partially constrained by its contrasting categories. Further, the same aspect of an entity may have multiple feature representations (e.g., cap, green-tinged cap).

7. Final Comment

Ultimately, any theory of how features are represented must address the symbol grounding problem (Harnard 1990). That is, the representations of features that are in our heads must make contact with what they represent—what is out in the world. Much of the utility of representations is in allowing us to interact successfully with the world. For example, the representation of the green-tinged cap must allow mushroom pickers to identify actual green-tinged caps of mushrooms so that they can avoid eating them. As another example, a relatively abstract feature such as 'plays music' for a CD player is not particularly useful unless it indicates more specifically how one can use a CD player to play music.

See also: Concept Learning and Representation: Models; Face Recognition Models; Letter and Character Recognition, Cognitive Psychology of; Neural Representations of Objects

Bibliography

Barsalou L W 1999 Perceptual symbol systems. *Brain and Behavioral Sciences* **22**: 577–609

Biederman I 1987 Recognition-by-components: A theory of image understanding. *Psychological Review* **94**: 115–47

Biederman I, Shiffrar M M 1987 Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory and Cognition* **13**: 640–45

Bugelski B R, Alampay D A 1961 The role of frequency in developing perceptual sets. *Canadian Journal of Psychology* **15**: 205–11

Harnad S 1990 The symbol grounding problem. *Physica D* **42**: 335–46

Hoffman D D, Richards W 1984 Parts of recognition. *Cognition* **18**: 65–96

Hubel D H, Wiesel T N 1965 Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neuropsychology* **28**: 229–89

Lesgold A M, Rubinson H, Feltovich P, Glaser R, Klopfer D, Wang Y 1988 Expertise in complex skill: Diagnosing X-ray pictures. In: Chi M T H, R Glaser Farr M (eds.) *The Nature of Expertise*. Erlbaum, Hillsdale, NJ, pp. 322–51

Murphy G L, Medin D L 1985 The role of theories in conceptual coherence. *Psychological Review* **92**: 289–316

Schyns P G, Rodet L 1997 Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition* **23**(3): 681–96
Schyns P G, Goldstone R L, Thibaut J P 1998 The development of features in object concepts. *Behavioral and Brain Sciences* **21**: 1–54
Solomon K O Barsalou L W (in press) Representing properties locally: Evidence for the perceptual grounding of concepts. Manuscript submitted for publication
Wisniewski E J, Medin D L 1994 On the interaction of theory and data in concept learning. *Cognitive Science* **18**: 221–81

E. Wisniewski

Fechnerian Psychophysics

Fechnerian psychophysics is based on the idea that an appropriately chosen measure of local discriminability in a continuous stimulus space (the degree with which an observer can discriminate a stimulus from its very close neighbors) can be used to compute 'subjective' distances between very close stimuli; and that by integrating these small distances appropriatly along paths connecting different stimuli one can compute 'subjective' (*Fechnerian*) distances among all stimuli comprising the space. This idea, together with experimental procedures for measuring local discriminability, was originally proposed by Fechner (1860, 1877) for unidimensional stimulus continua (such as the space of fixed-frequency tones varying in intensity, or visually presented line segments varying in length). Historians often date the beginning of scientific psychology from Fechner's work.

1. *Classical Theory* (*Refined*)

If a stimulus x changes along a single physical dimension (an interval of reals), and $F(x)$ is the *local discriminability measure* at x (positive and continuous), then the *Fechnerian distance* between any two stimuli a, b is defined as

$$G(a, b) = \left| \int_a^b F(x)\,dx \right| \tag{1}$$

The Fechnerian distance clearly satisfies the defining properties of a metric: $G(a, a) = 0$, $G(a, b) > 0$ for $a \neq b$, $G(a, b) = G(b, a)$, and $G(a, b) \leqslant G(a, x) + G(x, b)$.

Assuming that x is much greater than the *lower absolute threshold* x_{inf} of the stimulus continuum, so that the random variability in x_{inf} can be ignored, the distance $G(x, x_{\text{inf}})$ can be called the 'Fechnerian magnitude' of x (traditionally, 'sensation magnitude').

Let $\xi_x(y)$ be the probability with which y is perceived to be greater than x, and assume that $\xi_x(x) = 1/2$ (the latter can be ensured by an appropriate recalibration of the reference stimuli x). Then the *local discriminability measure* $F(x)$ is defined as

$$F(x) = C \left. \frac{d\xi_x(y)}{dy} \right|_{y=x+} \tag{2}$$

where certain regularity assumptions ensure that the derivative exists and continuously changes with x; $C > 0$ is a proportionality coefficient allowed to be different for different stimulus continua.

For $1/2 \leqslant p < 1$, let $W_p(x)$ denote the value of y at which $\xi_x(y) = p$; then $W_{1/2}(x) = x$. The difference $W_p(x) - x$ is called the 'just noticeable difference' (JND) at x on level p, or p-JND, for short. Putting $y = W_p(x)$ and using the identity $\xi_x\,[W_p(x)] = p$, Eqn. (2) can be written as

$$F(x) = C \left. \frac{dp}{dW_p(x)} \right|_{p=1/2+} = C \lim_{p \to 1/2+} \frac{p-(1/2)}{W_p(x)-x} \tag{3}$$

Since, by Eqn. (1)

$$\lim_{p \to 1/2+} \frac{G[a, W_p(a)]}{p-(1/2)} = \left. \frac{dG[a, W_p(a)]}{dp} \right|_{p=1/2+}$$
$$= \left. \frac{d \int_a^{W_p(a)} F(x)\,dx}{dp} \right|_{p=1/2+}$$

one can use the chain rule and apply Eqn. (3) to $F[W_p(a)]|_{p=1/2+} = F(a)$ to show that

$$\lim_{p \to 1/2+} \frac{G[a, W_p(a)]}{p-(1/2)} = C \tag{4}$$

By fixing p at a value sufficiently close to $1/2$ one concludes that $G[a, W_p(a)]$ is approximately constant for all a, because of which $G(a, b)$ in Eqn. (1) is roughly proportional to the number of chained p-JNDs, $W_p(a) - a$, $W_p[W_p(a)] - W_p(a)$, etc., that fit between a and b (see Dzhafarov and Colonius 1999 for details). The p-JNDs can be estimated from empirical data by a variety of techniques (see, e. g., Falmagne 1985). The use of the methods of *limits* and *adjustments* in Fechnerian psychophysics, as these provide JNDs on unknown probability levels, is predicated on the assumption that these JNDs are roughly proportional to p-JNDs on a constant level p sufficiently close to $1/2$.

2. *Psychophysical Laws*

If, with some choice of the physical measure for x, $d\xi_x(y)/dy$ at $y = x$ is proportional to $1/x$ (*Weber's law*), then Eqns. (1) and (2) yield $G(a, b) = C \log(a/b)$

(*Fechner's psychophysical law*). In spite of the enormous importance attached to Weber's and Fechner's laws in the history of psychology, the logic of Fechnerian psychophysics by no means implies them and is not contingent on their validity. Fechnerian psychophysics is equally consistent, for example, with such often found empirical approximations as $F(x) \propto 1/(x+\delta)$ (*modified Weber's law*) or $F(x) \propto 1/x^{\alpha}$ (called the *near miss to Weber's law* if α is just below 1, and the *Fullerton–Cattell law* if $\alpha \approx 0.5$). These two forms of $F(x)$ imply, respectively, $G(a, b) = C\log[(a+\delta)/(b+\delta)]$ and $G(a, b) = C'(b^{1-\alpha} - a^{1-\alpha})$ as substitutes for Fechner's law.

3. Criticisms of Classical Theory

3.1 Infinitesimals vs. Finite Differences

The account in Sect. 1 is a refined interpretation of Fechner's original theory. It is supported by Fechner's writings, but being voluminous and rather lacking in mathematical rigor, these writings lend themselves to alternative interpretations. According to one of them, Fechner believed in 'true' irreducible JNDs, $W_*(x) - x$, and simultaneously maintained that $G[x, W_*(x)] =$ const ('Fechner's postulate'), and that Eqn. (1) holds with $F(x) = C/[W_*(x) - x]$ compare with Eqn. (3). These propositions, however, are mutually inconsistent (Luce and Edwards 1958). Fechner knew of this criticism, and his 1877 rejoinder indicates that the interpretation it is based on is not correct. Whatever the historical truth, it is a sound scientific strategy to reserve the term 'Fechnerian psychophysics' for the interpretation that preserves the essence of Fechner's approach while freeing it from logical contradictions.

3.2 Fechnerian Distances and Subjective Matching

Consider four stimuli, say, tones with amplitudes and frequencies (A_1, f), (A_2, f), (A_1', f'), (A_2', f'), and let (A_1, f) and (A_1', f') be matched in loudness, and the same be true for (A_2, f) and (A_2', f'). Then the 'loudness-wise' Fechnerian distance between (A_1, f) and (A_2, f) should equal that between (A_1', f') and (A_2', f'), and one could expect that the numbers of chained amplitude JNDs fitting between the first two and the last two tones are roughly equal (under the assumption stipulated in Sect. 1). Riesz (1933) found this prediction wrong: for instance, the number of amplitude JNDs between two 4000-Hz tones may exceed that between the corresponding 200-Hz tones by a factor of four. This may be thought to invalidate the Fechnerian theory, as the difference seems too large to be attributed to the approximation error associated with counting finite JNDs (on unknown probability levels). The counter-argument is as follows. If the amplitudes coupled with frequency f and those coupled with frequency f' can be treated as *separate* stimulus continua, then the values of C in Eqn. (3) (and hence the approximate proportionality coefficients between Fechnerian distances and JND counts) may be different for f and f'. If, on the other hand, the amplitude-frequency combinations should be treated as comprising a single two-dimensional stimulus space (see Sect. 5), then it is no longer justifiable to measure the Fechnerian distance between (A_1, f) and (A_2, f) along the straight line connecting them. For a detailed treatment of subjectively unidimensional (e.g., 'loudness-wise') discriminations among multidimensional stimuli see Dzhafarov and Colonius (1999).

3.3 Fechnerian Magnitudes vs. 'Directly' Estimated Magnitudes

For certain stimulus continua, called *prothetic* (intensity, length, duration, etc.), the Fechnerian magnitudes $G(x, x_{\text{inf}})$ computed from discrimination data typically are found nonlinearly related to magnitudes obtained by 'direct scaling' methods (Stevens 1975). For instance, if Weber's law holds on a continuum of x, implying $G(x, x_{\text{inf}}) = C\log(x, x_{\text{inf}})$, the observer's numerical estimates of x are often found closely adhering to $S(x) = Kx^{\alpha} = \exp[\lambda G(x, x_{\text{inf}})]$ (*Stevens' psychophysical law*). Many have agreed with Stevens that $S(x)$, and not $G(x, x_{\text{inf}})$, is the 'true' measure of the 'sensory magnitude' of x. As no independent definition of sensory magnitude is available, however, the argument is largely semantic. The classical Fechnerian theory does not predict that all judgments of stimuli interpretable in terms of 'how large,' ' how high,' etc., must be proportional to Fechnerian magnitudes (although this happens to be the case for some *metathetic* continua, such as frequencies of constant-amplitude tones). A meaningful question to ask within the framework of Fechnerian psychophysics is whether one can propose a plausible mechanism, process, or computation by which an observer arrives at the $S(x)$-measures of prothetic stimuli based on their Fechnerian magnitudes $G(x, x_{\text{inf}})$. A naive but refinable example of such a proposal can be found in Ekman (1964).

4. Probability-distance Hypothesis

According to Eqns. (1) and (2), Fechnerian distances $G(a, b)$ are computed from discrimination probabilities $\xi_x(y)$ taken in arbitrarily small vicinities of $y = x$. A natural question to ask, therefore, is whether $\xi_a(b)$ (provided it is not zero or one) is determined uniquely by $G(a, b)$ for all stimulus pairs a, b. More generally: can one find a continuous *internal metric* $D(a, b)$ and a continuous function f such that

$$\xi_a(b) = f[D(a, b)] \tag{5}$$

and if so, what is the relationship between $D(a, b)$ and the Fechnerian metric $G(a, b)$? Metric $D(a, b)$ is called *internal* (also, *inner* or *intrinsic*) to the given unidimensional continuum if $D(a, b) + D(b, c) = D(a, c)$ for any $a < b < c$ on this continuum. Thus, the Fechnerian metric G is internal. The assumption that Eqn. (5) holds true can be called the *probability-distance hypothesis*, and the problem of finding D and f satisfying this hypothesis is referred to traditionally as 'Fechner's problem' (Luce and Galanter 1963).

Primarily due to the criticism mentioned in Sect. 3.1, Fechnerian psychophysics since the 1960s has been reduced essentially to the probability-distance hypothesis. One should note, however, that the classical theory presented in Sect. 1 does not imply the truth of this hypothesis. At the same time, the implication holds in the reverse direction: Pfanzagl (1962) proved that (subject to some regularity conditions) Eqn. (5) may hold for an internal metric D only if $D(a, b) = G(a, b)$. That is, discrimination probabilities may not depend on any internal metric other than the Fechnerian one.

Falmagne (1985) proved that Eqn. (5) is equivalent to either of the following two conditions: (weak bicancellation) $\xi_a(c)$ is determined uniquely by $\xi_a(b)$ and $\xi_a(c)$, for any a, b, c; (weak quadruple condition) $\xi_a(b) = \xi_{a'}(b')$ if and only if $\xi_a(a') = \xi_b(b')$.

5. *Multidimensional Fechnerian Psychophysics*

A generalization of Fechnerian psychophysics to continuous stimulus spaces of arbitrary dimensionality (such as the CIE color space, or space of tones varying in both frequency and amplitude) was proposed by Dzhafarov and Colonius (1999). A special form of this generalization, however, can be traced back to the tradition, originated by Helmholtz and Schrödinger, of computing colors metrics from color-discrimination data (Wyszecki and Stiles 1982). The discrimination of multidimensional stimuli $\mathbf{x} = (x_1, \ldots, x_n)$ cannot be defined in terms of 'greater–less,' and the function $\xi_a(y)$ of the classical theory should be replaced with $\psi_{\mathbf{x}}(\mathbf{y})$, the probability of $\mathbf{y}$ being perceived has *different* from $\mathbf{x}$. Let $\psi_{\mathbf{x}}(\mathbf{y})$ attains its minimum at $\mathbf{y} = \mathbf{x}$ (if necessary, after an appropriate recalibration of reference stimuli $\mathbf{x}$). The local discriminability measure $F(\mathbf{x}, \mathbf{u})$ *at stimulus* $\mathbf{x} = (x_1, \ldots x_n)$ *in direction* $\mathbf{u} = (u_1, \ldots, u_n)$ is defined as

$$F(\mathbf{x}, \mathbf{u}) = C \lim_{s \to 0+} \frac{\Phi[\Psi_{\mathbf{x}}(\mathbf{x} + \mathbf{u}s) - \psi_{\mathbf{x}}(\mathbf{x})]}{s} \tag{6}$$

where the transformation Φ is one and the same for all $\mathbf{x}$ and $\mathbf{u}$. The assumptions about $\psi_{\mathbf{x}}(\mathbf{y})$ underlying the theory guarantee that $F(\mathbf{x}, \mathbf{u})$ is determined by $\psi_{\mathbf{x}}(\mathbf{y})$ uniquely, and is continuous and positive, while Φ is determined asymptotically uniquely (put roughly, uniquely in a small vicinity of zero). The local discriminability $F(\mathbf{x}, \mathbf{u})$ can be used to compute the *psychometric length* L of any path $\mathbf{x}(t)|_{\mathbf{a}}^{\mathbf{b}}$, $a \leqslant t \leqslant b$, lying within the stimulus space and connecting stimuli $\mathbf{a} = \mathbf{x}(a)$ and $\mathbf{b} = \mathbf{x}(b)$:

$$L[\mathbf{x}(t)|_{\mathbf{a}}^{\mathbf{b}}] = \int_{\mathbf{a}}^{\mathbf{b}} F[\mathbf{x}(t), \dot{\mathbf{x}}(t)]\, dt \tag{7}$$

The Fechnerian distance between $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$G(\mathbf{a}, \mathbf{b}) = \inf L[\mathbf{x}(t)|_{\mathbf{a}}^{\mathbf{b}}] \tag{8}$$

where the infimum is taken over all paths connecting $\mathbf{a}$ with $\mathbf{b}$. Under certain conditions, this infimum may equal the psychometric length of a certain ('shortest') curve connecting $\mathbf{a}$ with $\mathbf{b}$, a *Fechnerian geodesic*. Fechnerian distances are determined uniquely and are invariant with respect to all smooth one-to-one physical reparametrizations of the stimulus space. If $\psi_{\mathbf{x}}(\mathbf{x}) = \text{const}$ (*constant self-similarity condition*), then Fechnerian distances are also invariant with respect to smooth monotone transformations $T[\psi_{\mathbf{x}}(\mathbf{y})]$ (e.g., due to *response bias* changes).

The construction just described identifies the Fechnerian metric as *internal to the stimulus space*. (In general, a distance function is internal to a space if the distance between any two points in the space is the infimum of the lengths of all paths connecting these points and lying entirely within the space.) A multidimensional generalization of Pfanzagl's theorem on the probability-distance hypothesis holds: $\psi_{\mathbf{a}}(\mathbf{b})$ (different from zero and one) may be determined uniquely by an internal metric $D(\mathbf{a}, \mathbf{b})$ only if $D(\mathbf{a}, \mathbf{b}) = G(\mathbf{a}, \mathbf{b})$.

The multidimensional Fechnerian theory, when specialized to unidimensional continua, sheds additional light on some aspects of the classical theory. While the concept of a unidimensional Fechnerian distance readily generalizes to spaces of arbitrary dimensionality, no natural generalizations exist for the notions of a Fechnerian ('sensation') magnitude and psychophysical law, such as Fechner's or Stevens'. This suggests that the importance traditionally attached to these notions may be unwarranted. By the same criterion (multidimensional generalizability), it is desirable to reformulate the classical theory in terms of the 'same–different' discrimination probabilities $\psi_{\mathbf{x}}(y)$, instead of the traditional 'greater–less' probabilities $\xi_x(y)$, and accordingly to revise the empirical procedures for computing the local discriminability measure $F(x)$.

6. *Concluding Remark*

With discrimination among stimuli being arguably the most basic cognitive function, Fechnerian psychophysics is motivated by an expectation that distances

computed from discrimination probabilities should have a fundamental status among behavioral measurements, and that in the final analysis various kinds of perceptual judgments could be shown to depend prominently on Fechnerian distances among stimuli involved. The eventual status of Fechnerian psychophysics will depend on the extent to which this expectation will be confirmed by future experimental and theoretical developments.

See also: Memory Psychophysics; Psychophysics; Signal Detection Theory; Signal Detection Theory, History of

Bibliography

Dzhafarov E D, Colonius H 1999 Fechnerian metrics in unidimensional and multidimensional stimulus spaces. *Psychological Bulletin and Review* **6**: 239–68

Ekman G 1964 Is the power law a special case of Fechner's law? *Perceptual and Motor Skills* **19**: 730

Falmagne J C 1985 *Elements of Psychophysical Theory*. Oxford University Press, Oxford, UK

Fechner G T 1860 *Elemente der Psychophysik* [Elements of Psychophysics]. Breitkopf and Härtel, Leipzig

Fechner G T 1877 *In Sachen der Psychophysik* [In the Matter of Psychophysics]. Breitkopf and Härtel, Leipzig

Luce R D, Edwards W 1958 The derivation of subjective scales from just noticeable differences. *Psychological Review* **65**: 222–37

Luce R D, Galanter E 1963 Discrimination. In: Luce R D, Bush R R, Galanter E (eds.) *Handbook of Mathematical Psychology*. Wiley, New York, Vol. 1, pp. 191–244

Pfanzagl J 1962 Über die stochastische Fundierung des psychophysischen Gesetzes [On the Stochastic Foundations of Psychophysical Law]. *Biometrische Zeitschrift* **4**: 1–14

Riesz R R 1933 The relationship between loudness and the minimum perceptible increment of intensity. *Journal of the Acoustical Society of America* **4**: 211–16

Stevens S S 1975 *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. Wiley, New York

Wyszecki G, Stiles W S 1982 *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley, New York

E. N. Dzhafarov

Federalism

'Federalism' refers to both an arrangement of political institutions and a philosophy of government. Federalism as a set of institutional arrangements denotes a political system in which power is shared between a central government and regional governments. Federalism as a philosophy of government holds that such federal structures are a desirable governmental form.

1. Federalism and Related Terms

The distinction between 'federalism' and related terms should be clarified. A government displaying the institutional characteristics of federalism is referred to as a 'federal system,' and the institutional structures themselves are described with the adjective 'federal.' A federal system is contrasted with a 'unitary system' of government in which all major decisions are made by the central government; some degree of administrative decentralization, however, may be present in a unitary system (see *Unitary State*). The terms 'federalism' and 'federal system' have been used as overarching terms which encompass a variety of more specific types of political arrangements, including federations, confederacies, and leagues (Riker 1975). The origins of the word 'federal' can be traced to the Latin *foedus*, denoting a league, treaty, or covenant. The term 'intergovernmental relations' refers to interactions between different governmental units within a political system. This term includes but is not limited to the study of interactions between central and regional governments within federal systems.

2. The Normative Study of Federalism

The study of federalism can be usefully divided into normative and empirical branches. The normative branch, which dates back centuries and remains very important today, includes the arguments in favor of the US constitution put forth in *The Federalist*. This normative literature consists of systematic consideration of federalism as a philosophy of government, including the enumeration of possible advantages and disadvantages of federal systems.

Proponents of federalism argue that federal systems are desirable because they may produce a number of advantageous consequences. In particular, advocates of federalism suggest that federal systems protect citizens against tyranny by dividing power among central and regional governments; expand possibilities for political participation by increasing the number of governments within a political system; allow for regional variations in public policy that are reflective of geographic differences in citizen preferences; and provide multiple governmental 'laboratories' in which innovative policy responses to public problems can be tested (Shapiro 1995).

Other scholars have argued that federal systems may produce a variety of negative consequences. For example, federal systems may encourage regional inequalities in the distribution of the costs and benefits of public policies; allow for the oppression of local minorities by local majorities within the subnational jurisdictions; and produce an undesirable 'race to the bottom,' in which regional governments attempt to attract economic enterprises by lowering regulatory standards and reducing governmental taxation and expenditure (Shapiro 1995).

3. The Empirical Study of Federalism

The empirical branch of the study of federalism has focused on the description and measurement of federal systems, and also on the establishment of relevant causal relationships.

3.1 Descriptive Approaches

In the descriptive realm, scholars have been concerned above all with identifying the core characteristics of federalism, characterizing variation in federal systems, and describing the mechanisms through which federal systems operate. At the heart of the federal idea is the division of powers and responsibilities between an overarching central government exercising authority over a broad territory, on the one hand, and a number of constituent governments each exercising authority over a particular subset of that territory, on the other. This division of power is generally enshrined in a constitution that guarantees the existence of all of the governments and provides at least a rough outline of the division of power among them.

Within these broad parameters, scholars have offered a range of more specific definitions of federalism, and no universal agreement about the term's precise definition has been reached. Different definitions produce somewhat differing classifications of political systems into federal and nonfederal categories. The result is that the term 'federal' can be applied to a variety of political systems in which power is shared between general and constituent governments. The political institutions of ancient Israel and the Greek city-states displayed federal characteristics. The US federal system, established in 1789, has served as the prototype of modern federalism (Elazar 1968). At the turn of the twenty-first century, significant federal elements existed in political systems across the globe, including Argentina, Australia, Austria, Belgium, Brazil, Canada, the European Union, Germany, India, Malaysia, Mexico, Nigeria, Pakistan, the Russian Federation, South Africa, Spain, Switzerland, the USA, and Venezuela (Elazar 1994).

Among federal systems, there is variation in the degree and character of the distribution of power between national and regional governments. This variation can be observed both in formal constitutional arrangements and in the working relationships among different levels of government within a political system. One component of the variation between federal systems, and within individual systems over time, is the extent to which powers and responsibilities rest with the central government or with the regional governments. Federal systems also vary in the extent to which these powers are either strictly divided between levels of government or shared between those levels of government.

With respect to the first of these dimensions, scholars have not yet reached a consensus about a precise way to measure the relative centralization and decentralization of federal systems (Watts 1998). The formal distribution of powers set forth in constitutions may provide some insight into the comparative degree of centralization between systems. Such a constitutional focus has limitations, however, and is unlikely to be as useful in measuring changes over time within a federal system. Here, close attention must be paid to legislative and judicial decisions altering the balance of power between the central government and the constituent governments. One summary measure that is commonly examined is the proportion of governmental expenditures undertaken by the central government, either in total or in particular policy areas (Peterson 1995). Efforts to characterize variation in the distribution of powers between general and constituent governments in federal systems have also led to the development of a wide range of metaphors and conceptual models (e.g., 'dual federalism,' 'cooperative federalism'). These metaphors and models attempt to characterize either the degree of centralization in a federal system or the extent to which powers are strictly divided or cooperatively shared between levels of government, or both. In all, Stewart (1982) found more than 300 such models and metaphors that had been proposed in the literature.

Descriptive studies have also examined, both statically and dynamically, the mechanisms through which federal systems operate. Core areas of concern here include the constitutional, legislative, and judicial processes through which powers are allocated between levels of government, and the fiscal (e.g., grants-in-aid), regulatory, and political instruments through which governments within a federal system influence each other's actions (Wright 1988).

3.2 Causal Approaches

The causal branch of the empirical study of federalism addresses two major areas of concern: the causes and consequences of the establishment and dissolution of federal systems, and the causes and consequences of changes in how powers are divided between governments within federal systems.

Federal systems can be established either by joining together previously independent governments, or by decentralizing power in a previously unitary state. The formation of federal systems involves somewhat different processes and conditions in each of these two cases. In all cases, however, federal systems emerge under political conditions that contain strong pressures for both political unity (in the form of a national government) and political autonomy (in the form of regional governments with meaningful powers).

The pressures for unity most commonly result from a desire for military or economic advantage that constituent governments feel cannot be achieved independently. When federal systems are formed from previously independent governments, the reluctance

of these governments to cede too much power to a newly formed central government creates pressure for regional autonomy. More generally, pressures for autonomy most commonly involve protecting constituent governments' diverse interests, whether based on nationality, language, ethnicity, religion, economy, or other variables. Against the backdrop of these general dynamics, scholars have debated what conditions are necessary and sufficient to produce a federal system, with particular discussion of cultural, economic, political, social, and historical variables (Riker 1975).

Scholars have sought to understand not only why federal systems are formed, but also what the consequences of federal systems are. For example, federal systems can have important consequences for the management of ethnic conflict, either mitigating or exacerbating ethnic tensions, depending on the particular institutional arrangements in place.

The late twentieth century witnessed the disintegration of some political systems that were federal in principle or practice, including the Soviet Union, Yugoslavia, and Czechoslovakia. The survival of federal systems is imperiled when institutional arrangements reinforce rather than crosscut existing lines of division within the system, or when one group or constituent government representing a supermajority dominates the overall system (Watts 1998, Elazar 1968).

At a less macroscopic level, the causal study of federalism focuses to a significant extent on the evolution of federal systems: why the distribution of powers changes in federal systems, and with what consequences. There are numerous studies of the proximate impact of legislative, judicial, and executive decisions on the distribution of powers in federal systems. Other analysts have focused on the role of underlying forces such as changes in technology, culture, ideology, society, and economy (Anton 1984).

In all, there is significant dissatisfaction among scholars with the current state of our ability to construct robust theories of why federal systems evolve in the ways that they do (Beam et al. 1983). Explanations focusing on proximate causes have been criticized for being atheoretical and for too often relying on single case studies. Analyses focusing on underlying causes have not produced a clear consensus (Anton 1984). Understanding the causes of centralization and decentralization in federal systems will remain an important area of investigation in the decades ahead. Scholars may increasingly attempt to link explanations focusing on proximate and underlying causes, constructing theories that specify how underlying forces such as culture or economy might in turn produce centralizing or decentralizing actions by proximate institutions such as legislatures or courts.

Scholars have also discussed the consequences of centralization and decentralization within federal systems. For example, there is evidence that decentralized systems make decentralized political party structures more likely (Riker 1975). In another realm, some scholars have argued that decentralization of redistributive and regulatory functions may lead to a competitive spiral in which subnational governments reduce their interventions in these areas (Peterson 1995).

4. *Future Directions*

Several areas of the study of federalism will likely be of particular importance in the first decades of the twenty-first century, as the literature continues to respond to important political developments. Scholars will be particularly interested in: the interaction between economic globalization and federal structures (see *Globalization: Political Aspects*); the emergence and evolution of supranational federal arrangements, including the European Union (see *European Integration*; *Regional Integration*); and the relationship between federal structures and the management of conflict in multiethnic political systems.

See also: Electoral Systems; Globalization: Political Aspects; Political Representation; Political Thought, History of; Regional Government; Representation: History of the Problem

Bibliography

Anton T J 1984 Intergovernmental change in the United States: An assessment of the literature. In: Miller T (ed.) *Public Sector Performance: A Conceptual Turning Point*. Johns Hopkins University, Baltimore, MD, pp. 15–64

Beam D R, Conlan T J, Walker D B 1983 Federalism: The challenge of conflicting theories and contemporary practice. In: Finifter A (ed.) *Political Science: The State of the Discipline*. American Political Science Association, Washington, DC, pp. 247–79

Elazar D J 1968 Federalism. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 5, pp. 353–67

Elazar D J (ed.) 1994 *Federal Systems of the World*, 2nd edn. Longman, Harlow, UK

Peterson P E 1995 *The Price of Federalism*. The Twentieth Century Fund, New York

Riker W H 1975 Federalism. In: Polsby N W, Greenstein F I (eds.) *Handbook of Political Science*. Addison-Wesley, Reading, MA, Vol. 5, pp. 93–172

Shapiro D L 1995 *Federalism: A Dialogue*. Northwestern University, Evanston, IL

Stewart W H 1982 Metaphors, models, and the development of federal theory. *Publius: The Journal of Federalism* **12**: 5–24

Watts R L 1998 Federalism, federal political systems, and federations. *Annual Review of Political Science* **1**: 117–37

Wright D S 1988 *Understanding Intergovernmental Relations*, 3rd edn. Brooks/Cole, Pacific Grove, CA

A. Burris

Female Excision: Cultural Concerns

1. Overview

Female excision, also known as female circumcision, is a euphemism for a variety of genital surgical procedures. It is a time-honored cultural tradition that has recently come under intense international scrutiny. Practices of genital alteration have existed in recent times in Australia, Asia, Latin America, America, and Europe. In fact, some Western surgeons claimed to have invented similar procedures as treatment for sexual dysfunction. Presently, excision is most common in Africa. Thus, this article focuses on African experiences. Given the discrepancies in existing studies, it is difficult to obtain accurate information about the exact prevalence of the practice. However, it is roughly estimated that about 80 to 110 million African women and children are affected.

There are at least four major forms of excision. In the mildest form, the clitoris is barely nicked or pricked to shed a few drops of blood. The second kind which is often called *sunna* involves the removal of the clitoral prepuce, hood, or outer skin. The third gradation, clitoridectomy, more accurately describes the process of excision. Here, the clitoral glans and some of the *nympha* or *labia minora*—the narrow lip-like enclosures of the genital anatomy, are severed. Infibulation is the least popular and most extreme procedure; it entails scraping the *labia majora*—the two rounded folds of tissue that contour the external boundaries of the vulva, and stitching the remaining raw edges together in a manner that reduces the vulva opening after the surgery heals. Different forms of the operation affect the genitalia differently.

2. Rationales

Nobody can unequivocally identify where, when, or why excision originated. The practice can mean different things to different actors in different milieus. Embedded in an intricate web of habits, attitudes, and values, excision has both functional and symbolic connotations. It is validated and undergirded by a spectrum of temporal and spiritual beliefs. While some of the justifications are at odds with ample medical knowledge, efforts to counter the practice by illuminating the medical contradictions have not been very successful in eroding commitment to it. This is mainly because, for adherents, the attraction and imperative of the practice has little to do with medical science. Recurring themes such as sexuality and fertility signify and affirm paramount indigenous values like solidarity, public recognition of lifecycle change, and procreation for social continuity. That it deals with the genetalia and is suffused with gender constructs does not denote unequivocally exclusive or primary concern with gender and sexuality.

Among the Sande for example, it has been reported that initiates spend months in seclusion being instructed on topics as broad as the cultural secrets of successful living and traditional medicine. With the Chagga, excision is said to test courage by simulating birth labor pains. To the Mende and Sherbro, it fosters cooperation among women. The Bambara and Mossi assume it reduces mortality and morbidity. Some Bantu and Sudanese use it for purification. In other areas, excision is employed as a means of self-definition, self-assertion, and expression of personhood and identity; it has also been characterized as the equivalent of an ornamental tatoo, or a kind of cosmetic surgery akin to ear-piercing. On a continuum with these worldviews are a competing set of beliefs which tend to emphasize the dimensions of sexuality and gender. But even these beliefs are tempered and nuanced in the manner in which they inform everyday realities.

The Mandingo, Massai, and Swahili believe that excision renders the body fertile and fit for the social order. Among some Yoruba it is considered a contraceptive device. In some other groups, excision reflects a fear of a primal hermaphroditic human nature and an attempt to differentiate the sexes. Still others justify it on grounds of aesthetics; to these, unaltered female genitalia are unsightly, and occluded or attenuated body orifices are more preferable. Some communities subscribe to the belief that a girl is ripe for sexual relations only after being ushered into womanhood by circumcision. Finally, there are communities that, to enforce moral values and ethical standards, discourage premarital sex. These groups believe that circumcision attenuates sexual desire and this mindset underscores its incorporation as a rite of passage. Regardless of the prevalence of this mindset in select groups, excision does not necessarily encode absolute prenuptial chastity nor does it *prima facie* signal a distortion of women's sexual experiences. From all indications, the practice is not homogenous; it has divergent contexts, meanings, and consequences. However, some opponents of the practice insistently explain it monolithically to privilege connotations of sexual politics and patriarchal control.

3. Emerging Trends

For various reasons, many communities where excision was once an aspect of elaborate rites of passage celebrations with minimum age requirements now excise children who are too young to comprehend the ritual implications of the experience. A corollary of this development is that some persons undergo the operation independent of accompanying ceremonies. In contradistinction to the evolving tendency to cut the genitalia without performing complementary ritual ceremonies, a few communities are embracing alternative rites that replace physical cutting with

symbolic enactments. Here initiation modules only entail health education, transmission of traditional knowledge, and festivity.

Patterns of change are also manifesting in more decidedly oppositional guises. Often, fierce opposition is a catalyst for other forms of transformation. Accordingly, it is not unusual to find the circumcision of the under-age and/or unceremonious cutting as forms of resistance to intense expressions of anti-circumcision impulses. These conflicts mutually reinforce each other.

An abiding bone of contention in the controversy about excision is the question of its health effects. Increasingly, the veracity of extant claims about health hazards and the premium that is placed on it as an arsenal for the anti-circumcision campaign, are being attacked. Searching analyses of available research suggest that advocates of abolition are quicker abstractly to chronicle an array of hazards than they are to produce objective evidence to back up their claims about the severity and frequency of health complications. Where these claims cannot be substantiated empirically, they operate as mere propaganda that remain unrepresentative of the lived experiences of circumcised persons. Although existing data are deficient, there is sufficient evidence to demonstrate that the health complications that are attributable to excision vary with the extent and circumstances of the surgery. The health hazards of the operation typically are compounded by the crude circumstances under which it usually occurs, including the lack of anesthesia and the use of unsterile instruments.

Drastic forms of excision such as clitordectomy are associated with an increased risk of bleeding and infection; the complications of infibulation include serious bleeding, infections, urinary problems, infertility, labor and delivery difficulties, and adhesions and obstructions. There is no objective evidence that directly links excision with mortality. The World Health Organization categorizes a range of possible negative health consequences as follows: short-term complications which include pain, injury to adjacent tissues, potentially fatal hemorrhage and shock, urinary retention, and acute and chronic infections; long-term complications that range from difficulty in passing urine, urinary tract infections, pelvic infections, infertility, keloid scars, abscesses and cysts, to menstrual difficulties, dyspareunia and sexual dysfunctions, and problems in pregnancy and childbirth as well as sexual, mental, and social consequences.

This list is more concise and less exaggerated than the inventories of deleterious sequalae in other literature; nevertheless, it is subject to caution as it derives from cross-sectional studies that are not always able to establish direct causation between the operation and the putative complication. This list is also comparable to other studies in the sense that it offers little guidance for inferring the rates of complications. Recent systematic reviews of relevant materials show that consistent and conclusive evidence of far-reaching complications are rare. These observations call for the qualification of the claims at the heart of the anti-circumcision campaign strategy. Notwithstanding the observations, the anti-circumcision alliance has done a remarkable job of mobilizing international consensus against the practice by emphasizing its harmful effects.

3.1 Eradication

Systematic efforts to abolish excision date back to the colonial era. In 1946, the British colonial administration in Sudan amended its Penal Code to make illegal all forms of genital surgeries, except *sunna*. In Kenya, interference by missionaries condemned the practice to secrecy; as a clandestine custom, it was particularly difficult to assess the incidence of casualties. In material respects, the intervention also politicized the significance of the operation as some began to appropriate it as a symbol of nationalism. According to Kenyatta (1938), the controversy came to a head in 1929–30 when families who interpreted the colonist measures as an affront to their freedom, social customs, and cultural beliefs hastened to circumcise their daughters, even at the hands of unskilled persons with inevitable tragic consequences.

The backlash recurred in 1989 when the incumbent President of Kenya, Arap Moi, denounced the custom and scores of women and girls volunteered for the procedure in protest of the official sanction. More recently, some grassroots communities are reinforcing, instead of resisting heightening global pressure against circumcision. To this end, several villages have adopted renunciation oaths and declarations that they complement with active learning and socioeconomic empowerment programs. Some eradication strategies train former excisors in grassroots management skills and facilitate their involvement in business enterprises. In lieu of material benefits, certain competing models—especially those that reconfigure the rite to eliminate the ritual of cutting—incorporate creative ways of enlisting the support of traditional excisors.

Several sovereign powers have prescribed an assortment of steps to stem excision. These range from economic sanctions and extradition to prosecution and punishment. In 1991, France became the first country to establish that the threat of 'genital mutilation' is sufficient grounds to grant women refugee status. Subsequently, Canada pioneered guidelines expanding the basis of refugee claims to include gender-related persecution which was construed to encompass circumcision. More recently in the United States, a few African women have succeeded in obtaining political asylum or staying deportation by conceptualizing circumcision as persecution.

A series of human rights instruments and provisions can also be read as prohibiting excision. These range

from regulations concerning equal protection and nondiscrimination to those pertaining to the rights of children, the right of sexual and corporal integrity, the right against torture, the right to life, and the right to health. The World Health Organization officially condemned surgical alterations of female genitalia in 1982. The World Medical Association, the International Federation of Gynaecology and Obstetrics, and the Population Fund issued warnings about the risks of the practice and oppose the participation of health professionals in it. In 1990, the Committee responsible for the implementation of the Convention for the Elimination of Discrimination against Women recommended the eradication of the practice.

Although a few African countries have instituted various initiatives to address excision, it appears that the force of culture and the poverty of resources inhibit meaningful implementation of these initiatives. A wide range of official declarations against excision have been made in Benin, Nigeria, Ghana, Burkina Faso, and Senegal. Some European countries have also promulgated relevant laws and policies. Several interventions promote prevention through information, education, and communication campaigns. However, the predominant trend proscribes the practice as a crime that is subject to prosecution. As of the summer of 1994, France was the only country in which the practice of circumcision has actually resulted in criminal trials. In 1982, the Royal College of Obstetricians and Gynecologists found that excision was a very infrequent occurrence in Britain. By 1985 however, a scare of waves of occurrences prompted the promulgation of the British Prohibition of Circumcision Act. Section 1 of the Act makes it an 'offense for any person (a) to excise, infibulate or otherwise mutilate the whole or any part of the labia majora or labia minora or clitoris of another person; or (b) to aid, abet, counsel or procure the performance by another person of any of those acts on that other person's own body.' Section 2 of the Act indicates that genital surgeries are not absolutely prohibited; its precise wording seems to accommodate exceptions under certain circumstances. This section provides that 'Subsection 1(a) of Section 1 shall not render unlawful the performance of a surgical operation if that operation (a) is necessary for the physical or mental health of the person on whom it is performed and is performed by a registered medical practitioner.' This provision allows the continuation of 'trimming' surgeries that typically are performed on women who are dissatisfied and psychologically disturbed by the shape or size of their external genitalia. Plausibly, the mental health exemption can be construed to support psychologically-grounded claims for circumcision. However, section 2 (2) of the Act precludes the consideration of psychological effects induced by any belief 'that the operation is required as a matter of custom or ritual.'

In 1996, the US Congress made the excision of a person who has not attained the age of 18 a felony punishable by up to five years imprisonment. This statute equally exempts a surgical procedure performed by a licensed medical practitioner, which is either necessary for the health of the woman on whom it is performed, or is performed prenatally or postnatally on a woman. The US Congress further authorized the imposition of economic sanctions against any countries that have a cultural custom of female circumcision and have not implemented educational programs to counteract the practice. Resorting to this course of action as an antidote to circumcision is not without limitations. Withholding or conditioning aid as a mechanism for controlling the practice subordinates humanitarian considerations to political expedience. While there obviously are no simple solutions, it is important to recognize that when deeply embedded cultural attitudes are reinforced by complex economic, social, and political variables, a more nuanced strategy may be more effective than essentially top-down or carrot-and-stick approaches.

3.2 Medicalization

Empirical evidence of the range and frequency of adverse health outcomes posit a strong correlation between the probability of harm and the degree of cutting, the nonhygienic and anesthetic conditions under which it is performed, and the limited skills of the excisor. Drawing on these findings, some authorities recommend a spectrum of medical interventions as a way of containing potential harm. In one initiative, nurses dispense prophylactic antibiotics, anti-tetanus injections, and sterile razors to girls who are later cut by traditional circumcisors. In another model, traditional birth attendants are trained in septic and precautionary procedures for genital cutting. In urban parts of Sudan and in Egypt, the governments have attempted to curtail the adverse health consequences that can be traced to excision by clinicalizing mild forms of the procedure. In the Netherlands, a Welfare, Health and Cultural Ministry report and officials at the Harborview Medical Center in Seattle in the US recommended a similar course of action.

Apparently some African immigrants in Holland circumcised themselves under circumstances that compound the potential harms of the procedure. Their dilemma arose partly because, while they could not afford to be alienated for failure to comply with the cultural mandate for circumcision, the Dutch medical and social workers denied them assistance in executing the operation. To address the needs of these women, the Ministry, on the basis of the findings of a study it sponsored among 500 Somali refugee women, recommended a distinction between mutilating or tissue-impairing circumcision and nonmutilating ritual incisions. The report further argued for official stipulation of conditions under which doctors might

be allowed to perform simple incisions or ritual perforations of the clitoral covering as an alternative to infibulation. The report drew a storm of protests and eventually was superseded in 1992.

In 1996, a similar controversy erupted in Seattle, Washington, when in the interest of some Somalian immigrants, pediatricians at Harborview Medical Center took under advisement the possibility of performing 'symbolic blood-letting' procedures on girls. The proposed procedure would only involve a negligible nicking of the prepuce far less invasive than what is allowed in the circumcision of infant males. Beseiged by outraged opponents, Harborview decided to abandon the prospective accommodation.

The Dutch and Seattle experiences illustrate the extent to which sworn opponents of the practice favor prohibition and are vehemently opposed to medicalizing any variant of the practice—even as an interim measure to arrest potentially dire consequences.

One reason proffered for this position is that the surgical compromise ultimately signals complicity in patriarchal control and diminution of the right of women to sexual and bodily integrity. Some ridicule operations performed under medical conditions and/or supervision as an empty caricature devoid of traditional connotations. Others who envisage a scenario where traditional midwives are displaced by forum change seek to enhance their fundamental opposition to the practice by coupling it with an ancillary aversion to the implicit loss of income for the midwives. Another concern is that medicalization may unduly burden and deplete public funds which could be deployed for less controversial purposes. It is also contended that the compromise creates a slippery slope that can be abused to camouflage severe forms of the operation.

Whatever the validity of these concerns, it is worth ruminating over the merits of changing the forum and form of the practice. The options may be problematic, but as already suggested, there are no exact solutions. Criminalization, which is overwhelmingly championed by the anticircumcision coalition, is not a panacea. Indeed, where not being excized is perceived as a transgression of social mores, attempting an undifferentiated assault may jeopardize the prospects for enduring reform. The rhetoric of the opposition is premised largely on claims about the health hazards of excision. However, a recent examination of how much available research supports widely publicized allegations about the prevalence and harmful effects of female genital surgeries found that most of what are purported to be universal 'facts' vis-à-vis the deleterious consequences are drawn from studies of the Sudan where most women are infibulated. The review confirmed that obtaining irrefutable epidemiological data is limited by logistical and ethical constraints involved in monitoring over time the experience of individuals who undergo the surgeries.

The fact that adverse complications are exacerbated by the poor skills, anatomical knowledge, sanitary conditions, and surgical equipment that characterize excision in traditional settings reinforces arguments for medicalization. By the same token, instituting a change of form such as symbolic pricking is inconsequential from a health standpoint. Without a probable risk of harm, it is difficult to make a compelling case against the accommodation of consensual genital marking. In this light, medicalization reconciles the need to protect women with a respect for embodied sociocultural identities. Some would even argue that the virtue of tolerance and the principle of self-determination allow accommodation of more extensive genital alterations, as long as the procedure is demonstrably benign. Given the current wave of anticircumcision sentiments, this view is subject to dispute and far from endorsement. Nevertheless, it raises issues that deserve serious attention.

3.3 Education

Several authorities hail education as the single most important factor in combating the practice. For one, law alone seldom changes behavior. Effective legal engineering can hardly occur without favorable public disposition, which in turn can be stimulated by fostering meaningful interactions between legal and educational strategies. In this vein, it is noteworthy that gains achieved through the medium of education are often contingent and enhanced by concurrent poverty alleviation. The emphasis on education and connection with poverty, however, can only go so far. The substantial prevalence rate of the genital operations among very educated and privileged women in many ethnic groups suggests that advances in literacy and socioeconomic status do not necessarily decrease allegiance to circumcision. Undoubtedly, good education is an effective antidote for challenging erroneous beliefs. However, many women undergo circumcision on the bases of rationales that cannot be readily refuted by objective evidence. Ritual thought and action belong to specific symbolic frames of reference and are best understood within the totality of their natural context. Attempting to falsify indigenous explanations of circumcision by regarding them as assertions of facts that can be disconfirmed through formally logical discursive categories obscure valid psychosocial processes.

See also: Female Genital Mutilation; Gender, Feminism, and Sexuality in Archaeological Studies; Gender Ideology: Cross-cultural Aspects; Gynecological Health: Psychosocial Aspects; Health: Anthropological Aspects; Health in Developing Countries: Cultural Concerns; Human Rights, Anthropology of; Human Rights, History of; Ritual; Ritual and Symbolism, Archaeology of; Sexuality and Gender

Bibliography

Association of African Women for Research and Development 1983 A statement on genital mutilation. In: Davies M (ed.) *Third World: Second Sex*. Zed Books, London

Boddy J P 1989 *Wombs and Alien Spirits: Women, Men and the Zar Cult in Northern Sudany*. University of Wisconsin Press, Madison, WI

Case Western Law Review 1996 Colloquium: Bridging society, culture, and law: The issue of female circumcision. *Case Western Law Review* **47**: 263

Coleman D L 1998 The Seattle compromise: Multicultural sensitivity and Americanization. *Duke Law Journal* **47**: 717

Daly M 1978 African genital mutilation: The unspeakable atrocities. In: *Gyn/ecology: The Metaethics of Radical Feminism*. Beacon Press, Boston

Dorkenoo E, Elworthy S 1992 *Female Genital Mutilation: Proposals for Change*. Minority Rights Group, London

Giorgis B W 1981 *Female Circumcision in Africa*. United Nations Economic Commission for Africa, Addis Ababa

Kenyatta J 1938 *Facing Mount Kenya; The Tribal Life of Gikuyu*. Secker and Warburg, London

Lightfoot-Klein H 1989 *Prisoners of Ritual: An Odyssey into Female Genital Circumcision in Africa*. Harrington Park Press, New York

Lionnet F 1992 Women's rights, bodies and identities: The limits of universalism and the legal debate around excision in France. *Inscriptions* **6**: 98

Obermeyer C M 1999 Female genital surgeries: The known, the unknown, and the unknowable. *Medical Anthropology Quarterly* **13**(1): 79–106

Saadawi N 1980 *The Hidden Faces of Eve: Women in the Arab World*.

Shell-Duncan B, Herlund Y in press *Female Circumcision in Africa: Culture, Change and Controversy*. Lynne Rienner, Boulder, CO

Shweder R in press What about FGM? And why understanding culture matters in the first place. In: Shweder R, Markus H, Minow M (eds.) *The Free Exercise of Culture: How Free Is It? How Free Ought It To Be?* Russel Sage, New York

Toubia N 1994 Female circumcision as a public health issue. *New England Journal of Medecine* **331**: 712–16

Van Gennep A 1960 *The Rites of Passage* (trans. Vizedom M, Caffee G). University of Chicago Press, Chicago

Walker A 1992 *Possessing the Secret of Joy*. Vintage Books, London

Wilson B R 1970 *Rationality*. Blackwell, Oxford

World Health Organization 1982 *Seminar on Traditional Practices Affecting the Health of Women and Children*.

World Health Organization 1996 *Female Genital Mutilation*. WHO, Geneva

L. A. Obiora

Female Genital Mutilation

Female genital mutilation is a collective term for various surgeries performed on young girls, often done as part of initiation rituals into adulthood. Other terms commonly used are female genital cutting, female circumcision, clitoridectomy, infibulation or Pharaonic circumcision, excision, and *sunna*. Except for female genital cutting and female circumcision which also are general terms and used interchangeably with female genital mutilation, the other terms are more specific, indicating the particular type of operations being performed.

1. Extent and Distribution

Sources estimate that perhaps as many as 130 million women in the world today have been mutilated genitally (Toubia 1995). Since relatively little extensive statistical research has been conducted on the topic, prevailing figures quoted tend to vary. Figures usually are derived from approximate population censuses from areas where mutilation is known to be practiced.

As a cultural practice, female genital surgeries are particularly widespread on the African continent and to some extent on the Saudi Arabian peninsula, but are only sporadically documented from the rest of the world. Instances of genital mutilation are reported from selected communities in the Middle East and from Muslim Malaysia. In Europe and the USA, clitoridectomy was until the 1930s medically administered in diagnosis of excessive masturbation, nymphomania, and hysteria (cf., e.g., Roheim 1932, Bonaparte 1953). Cosmetic surgery of enlarged clitorises and labia minora has been reported recently both from the USA and Europe (cf. Walley 1997). In Europe and the USA, female genital mutilation was primarily a single instance phenomenon and not a social convention. During the last few decades of the twentieth century, however, large-scale immigration to Europe, the USA, and elsewhere from African countries where genital operations are mandatory, has extended the geographical location of female genital mutilation (cf. Lionnet 1992)

In Africa, the practice of genital surgeries is contained to certain regions only, most notably to the northeastern and eastern parts of the continent, from Egypt in the north to Tanzania in the south and across the continent along the Sahelian countries from Ethiopia to Senegal. It is practiced in some 28 countries and by a wide variety and number of ethnic groups. The extent and type of surgical intervention as well as its cultural meaning and social context vary considerably making it difficult to generalize across the societies in which female genital surgeries are the norm. Female genital mutilation occurs among Muslims, Christians (particularly among Copts), as well as adherents of traditional systems of belief, among agricultural and nomadic peoples alike and in rural as well as urban communities. The highest regional concentration, that is, where a majority of women in a particular area are mutilated, is in Northeast Africa; in the Horn (Somalia, Djibouti, and Eritrea), across Ethiopia, and into northern Sudan

and Egypt. In Somalia and northern Sudan female genital mutilation is almost universal (Gruenbaum 1996). Elsewhere in Africa, genital surgeries as a cultural practice are scattered and alternate with ethnic affiliation, religious faith, economic class, and level of education. Within the same community, practices may even vary between subgroups and individual families, causing a situation where mutilated women live side by side with nonmutilated females.

As a general rule, where female genital mutilation is performed, male circumcision is also the norm. People themselves often draw a symbolic and social parallel between boys' and girls' circumcision. This is, among other things, reflected in a common term for the interventions in the local vernacular. While a number of societies in Africa and beyond practice male circumcision without cutting their women, the opposite, however, is never the case.

2. *Operation Practices*

Mutilation practices vary greatly. In some areas, such as Amharic Ethiopia, girls are operated upon within 40 days after birth, while in neighboring Somalia the operation usually is performed when the girls are between 6 and 8 years old (Hicks 1993, Talle 1993). In Kenya and Tanzania the girls are cut routinely just before puberty and often in association with elaborate initiation ceremonies (cf. Murray 1974). In a few, relatively exceptional cases reported among others from West Africa, the surgery is performed after puberty. It also has been reported that married women who have defied traditions and remained unmutilated, in cases of barrenness have undergone the surgery after being married in order to promote fertility. Not only the girl's age, but also the surgical intervention, the performance context, and ritual elaboration exhibit substantial variations. In places in Egypt and among devoted Muslims in towns in the Sudan, for instance, only the prepuce or the tip of the clitoris is excised. This operation is referred to as *sunna* ('duty' in Arabic, i.e., to follow tradition). Surgically, *sunna* is the mildest form of mutilation and physiologically comparable to the removal of the foreskin at male circumcision. Although encouraged by religious leaders, health personnel, and opinion makers as a commendable mutilation form, it is still uncommon. What people label *sunna*, however, in practice is often really a clitoridectomy.

The most widespread form of female genital mutilation in fact is clitoridectomy, an intervention that implies removal of the whole or part of the clitoris, sometimes together with the *labia minora*. In the latter case the operation is also referred to as excision. Various forms of clitoridectomy-cum-excision are practiced from Senegal and Mauritania in West Africa, across Sahel to Egypt in the northeast, and southwards to Tanzania in East Africa. It is carried out across the whole area where female genital surgeries occur.

The surgically most severe form of operation is called infibulation (from Latin *fibulae*, clip) or Pharaonic circumcision. This is a form of surgery where the *labia minora* and parts of *labia majora* are excised wholly or partly, and often the clitoris; the sides of the vulva are then stitched together. A tiny orifice is left open serving as an outlet for urine and menstrual blood. The surgery transforms the genitals of a woman into a flattened, smoothed, and closed vulva with an oblong scar and a miniscule opening at the lower part. The practice of infibulation is particularly prevalent in the Horn of Africa, in the Sudan (Northern), and in southern (Nubian) parts of Egypt. It also occurs widely in Mali, in northern Nigeria, and among ethnic Somalis in other East African countries. The infibulated women are opened (defibulated) at marriage, either surgically—often by the practitioner who did the closing—or by the husband at penetration.

In medical and anthropological literature it is common to differentiate between four types of genital surgeries: *sunna*, clitoridectomy, excision, and infibulation. The types are classified by a gradient of surgical severity. Variations and modifications of the different types may occur over time as well in individual cases within the same geographical location. In the Sudan, for instance, an intermediate type, a compromise between excision and infibulation, sprung up during the colonial area as a local response to legislation against the practice imposed from outside (El Dareer 1982). The skills and techniques of the practitioners as well as the preferences of individual families may contribute to variation and differences in mutilation styles.

Genital operations usually are performed in the home of the girl and, as a rule, undertaken by an elderly woman skilled in the task. Often these women are local birth attendants as well. In Somalia, the practitioners belonged traditionally to a ritually ambiguous and despised subclan within the community. In other places, and these days even in Somalia, any adult woman with the necessary courage and skills may engage in the practice.

The operation normally is performed without any extensive form of anaesthetic or antiseptic precautions. Different herbs, animal fat, dung, or other substances believed to have homeostatic and cicatrizing properties are applied to the wound after the operation to prevent excessive bleeding and infection. Depending upon the routine, skills, and preference of the practitioner, the instruments used for the surgery are either a small knife prepared for that kind of work, or a razorblade. In the case of infibulation, the severed parts are stitched with thorns of the acacia species, or recently, with catgut. In urban areas, medically trained personnel or paramedics commonly perform 'circumcision' on both boys and girls. In addition to giving prophylactic antibiotics and tetanus injections, the

urban practitioners normally apply local anesthetics when operating. Many parents prefer these modern operations as they tend to reduce the pain and minimize adverse effects, such as hemorrhaging or infection from the operation. With the medicalization of the genital surgeries, men have also become involved in operating on girls. Traditionally, men operated only on boys, while women operated on girls. The same-sex pair of circumcizer and circumcized reflected the thinking of male circumcision as an analogy of female circumcision.

Among peoples in Kenya and Tanzania and in large parts of West Africa, the surgical intervention is associated with elaborate rituals of initiation into adulthood (e.g., Murray 1974). In northeastern Africa, where infibulation is widespread, the operation, however, is not normally followed by any extensive ceremonial celebration. The operated girl is usually given a few presents and the family mark the day with some extra food and drinks, but the event is not a ceremony of great societal worth. In cases of physical cutting being part of some more significant rites of passage into adulthood, the waning of such traditions has not necessarily led to a decrease in the surgery itself. Instead of being part of a larger event, genital mutilation is now often a single act performed on girls at almost any age.

3. *Historical Roots*

The existence of female genital mutilation dates back to pre-Islamic and pre-Christian time. Although the practice as such does not stem from religious affiliation, many people make an association between circumcision and religious faith. This is particularly the case in West Africa where genital surgeries both in men and women are more common in Muslim than Christian communities (Koso-Thomas 1987).

The term Pharaonic circumcision originates from the folk legend that the pharaohs of ancient Egypt performed similar operations on their women. Whether the pharaohs actually infibulated their women or not will probably never be substantiated, but the term nonetheless indicates that we are dealing with a cultural phenomenon of great antiquity.

Greek merchants and travelers who crossed the Mediterranean in classical time mention that female genital mutilation occurred among Egyptians as well as among the peoples along the Red Sea. The sparse descriptions emanating from these travelers, however, do not allow us to ascertain whether it was infibulation or clitoridectomy, or both, people practiced in earlier times.

European travelers visiting Egypt much later reported on the occurrence of clitoridectomy (Browne 1799, Seligmann 1913). However, female slaves from the south (perhaps Nubians from the Sudan or southern Egypt), whom the wealthy Egyptians favored as mistresses, were infibulated (Widstrand 1964). Some authors held that these girls were mutilated upon arriving in the north by the slave merchants or by their masters to avoid loose living or unwanted pregnancies. This explanation is disputed as early reports suggest that infibulation occurred for some time among the peoples of the upper Nile in northern Sudan. Burckhardt who visited these areas early in the nineteenth century claimed that infibulated girls ('virgins') received a higher price on the Egyptian slave market than noninfibulated ones (Burckhardt 1819).

The beginning of female genital mutilation remains obscure. It has not been possible to determine whether mutilation originated in one area and from there spread to other areas or whether it originated independently in various places. With reference to the African continent, however, it is not unlikely that the practice arose in a culturally, commercially, central area and from there reached and influenced other parts of the continent. This assumption is based upon the fact that the practice of female genital mutilation appears to have had varying historical origins depending when people came into contact with each other (cf., e.g., Murray 1974). Historical and linguistic sources point to the Chusitic and Semitic-speaking peoples living along the Red Sea Coast as a core area of female genital mutilation (Seligmann 1913, Hicks 1993).

Recently, the practice of female genital mutilation is reported to have spread from one ethnic group to another in countries such as Chad, the Sudan, and Tanzania (cf. Leonard 2000, Gruenbaum 1996). In the Sudan, lower hierarchy groups often assimilate the practice of mutilation in order to be able to marry those of higher rank. Young girls in Chad are reported to perform clitoridectomy as a mark of modernity; since around 1980 the operation has become a kind of fashion in at least one reported area (Leonard 2000). Studies from places in West Africa suggest that the expansion of Islam is conducive to the spread of female genital mutilation. Among the ethnic Meru in Northern Tanzania, clitoridectomy has been revitalized as a sign of cultural pride and resistance against rapid modernization. These sketchy, contemporary examples show that the meaning of female genital surgeries may assume various values following shifts in social reality. To know its origin does not explain its present existence (Gruenbaum 1996). One distinguishable trait in the spread of female genital mutilation, however, appears to be the copying of cultural practices of higher status groups by people of lower rank (cf. Mackie 1995).

4. *The Cultural Explanation*

The practice of female genital mutilation is embedded deeply in cultural and moral preferences, gender identity and person forming, perceptions of body

aesthetics, and ethnic marking. The unsuccessful attempts made by the British colonial administration during the 1930s and 1940s to abolish or modify the most severe types of female genital mutilation in the Sudan and in Kenya testify to the social importance of the practice (cf. Murray 1974, El Dareer 1982). Anthropologists and others often associate the tradition of female genital mutilation with a patriarchal social structure, practically and symbolically privileging female reproductive powers over female sexuality. (cf. Boddy 1982, Talle 1994).

People themselves give many explanations to defend the practice. For instance, the pastoral Maasai of Kenya and Tanzania claim that a girl can neither be married nor give birth to healthy children if she is not clitoridectomized. Should she deliver before she is operated upon, the child will die or be born disabled (Talle 1994). In the Sudan and Egypt, uncut female genitals are considered ugly and dirty. Men are reluctant to marry such girls because by not being 'purified' by genital excision, they are held to be immoral in the widest sense. For the Bantu-speaking Kikuyu in Kenya a girl must have her clitoris removed to be considered an adult and responsible person. And so on. People present a plethora of meanings and motives for upholding the practice. Although seldom elaborated on explicitly, the control of female sexuality is at stake. References to moral looseness, unmarriagability, and fertility failures speak strongly to the societal value of female chastity and of male control of female fertility and sexuality (cf., e.g., El Sadaawi 1980, Gordon 1991, Talle 1993).

In Somalia, where the practice of infibulation is particularly widespread, the operation is performed in two stages: the excision and the stitching. First, the clitoris is excised. This is considered an act of purification (*xalaalays*; from the Arabic *xalaal*, allowed by religion). The clitoris (*kintir*) is a part of the body which is associated with dirt and childhood. A girl who has not been operated on is referred to as *kintirleey* (childish, i.e., having clitoris) and teased by peers. To adult Somali women the word *kinterleey* has a very negative and somewhat embarrassing ring. Another Somali synonym for *kintir* is *awle* which means approximately 'with (or having) father.' This term suggest that the Somalis, as a number of other peoples performing genital surgeries, see the clitoris as a male part in the female body and thus held to be unclean. When excised, women become more feminine (Talle 1993).

Where infibulation is practiced, virginity is not held to be a natural condition in women but has to be forcibly implanted in girls while they are still young. The hymen of a woman, which attracts so much interest and attention in many other Muslim societies and evokes strong sanctions if broken before marriage, has little social or cultural significance among infibulation practicing people. To them a chaste girl means a 'sewn' girl. A tight infibulation is a sign of distinction and moral excellence. Symbolically, the closing is more important than the excision as the cutting of the flesh is irrevocable while the closing of the vulva can, in principle, be executed over and over again. Although never as tight as the first time, later surgeries (after childbirth and at divorce) return the woman's body to its previous state by remaking her virginity (El Dareer 1987). Virginity under such circumstances is not an anatomical facticity but a social category (Hayes 1975).

The closing of her vulva is looked upon as a protection for the girl against indecency and immorality. A reason often given, for instance by Somalis, as to why originally they began to infibulate their girls was to prevent sexual assault of young girls while out herding alone. The infibulation, or the creation of the artificial virginity, is seen as a safeguard against unwanted pregnancies and unidentified paternity. Both in northern Sudan and in Somalia a nubile, 'virgin' daughter upholds and marks the moral standing of the partrilineal group (Hayes 1975, Kennedy 1980, Talle 1993). To leave a daughter or sister 'uncut' is held to be unnatural; the vision of such a woman is beyond comprehension, something which cannot even be expressed in words.

5. *A Global Concern*

During the feminist wave of the 1970s in Europe and the USA, female genital surgeries became a politicized issue. At the Copenhagen conference in 1980 during the United Nations Decade of Women (1975–85) the issue was brought to widespread international attention. Opponents of the practice asserted that female genital mutilation was a severe abuse of women's and children's rights and demanded legislation against it. The term 'female circumcision' was substituted by 'female genital mutilation' (Hosken 1982). This change of name from circumcision which is used for genital operations on men, was meant to draw attention to the severity of intervention, primarily in terms of health risks, on girls and women as compared with that of men (cf., e.g., Toubia 1995).

Legislation against female genital mutilation is on the agenda of international organizations such as Save the Children, Amnesty International, WHO, UNICEF, and other UN agencies. To prevent the spread of such surgeries into their countries, European governments are passing laws criminalizing all forms of genital mutilation (cf. Lionnet 1992). African countries, Egypt, the Sudan, and Kenya among others, have also passed legislation against female genital surgeries.

The global discourse on female genital mutilation has often been moralizing, racially prejudiced, and without sufficient understanding of the social and cultural context of the practice. Many African women have reacted to the paternalistic attitude that often

permeates the global debate. The term female genital mutilation, which has received widespread acclaim, but which holds the implicit assumption that parents and kin deliberately intend to harm children, is contested (cf. Walley 1997). A more neutral term proposed recently is 'female genital cutting' (Shell-Duncan and Hernlund 2000). African women—themselves frequently victimized by genital surgery—have, however, also been among the strongest opponents to the practice (e.g., Abdalla 1982). The latest voice in a string of personal testimonies of an excruciatingly painful intervention with lifelong suffering comes from an internationally celebrated Somali model, now a UN ambassador against female genital mutilation (Dirie and Miller1998).

Although opposition is growing everywhere, women continue to favor the practice. They do it not necessarily because they like it, but because they see no other alternative. Africans in exile, exposed to a majority society where mutilation is neither a norm nor an ideal and, additionally, relieved of the social pressure to perform it, may spearhead the eventual eradication of female genital mutilation.

See also: Cultural Relativism, Anthropology of; Female Excision: Cultural Concerns; Gender and Feminist Studies in Anthropology; Gender Ideology: Cross-cultural Aspects; Group Decision Making, Social Psychology of; Islam and Gender; Male Dominance; Near Middle East/North African Studies: Gender; Reproductive Rights in Developing Nations; Ritual; Sexuality and Gender

Bibliography

Abdalla R M D 1982 *Sisters in Affliction: Circumcision and Infibulation of Women in Africa*. Zed Press, London

Boddy J 1982 Womb as oasis, the symbolic context of Pharaonic circumcision in rural northern Sudan. *American Ethnologist* **9**(4): 682–98

Bonepart M 1953 *Female Sexuality*. International Universities Press, New York

Burckhardt J L 1819 *Travels in Nubia*. Darf Publishers, London

Dirie W, Miller C 1998 *Desert Flower: The Extraordinary Journey of a Desert Nomad*. William Morrow, Location

Dareer A 1982 *Women, Why Do You Weep? Circumcision and its Consequences*. Zed Press, London

Sadaawi N 1980 *The Hidden Face of Eve: Women in the Arab World*. Zed Press, London

Gordon D 1991 Female circumcision and genital operations in Egypt and the Sudan: A dilemma for medical anthropology. *Medical Anthropology Quarterly* **5**(1): 3–14

Gruenbaum E 1996 The cultural debate over female circumcision: The Sudanese are arguing this one out for themselves. *Medical Anthropology Quarterly* **10**(4): 455–75

Hansen H H 1972/73 Clitoridectomy: Female circumcision in Egypt. *FOLK* **14/15**: 15–26

Hayes R O 1975 Female genital mutilation, fertility control, women's roles, and the patrilineage in modern Sudan: A functional analysis. *American Ethnologist* **2**(4): 627–37

Hicks E K 1993 *Infibulation: Female Mutilation in Islamic Northeastern Africa*. Transaction Publishers, New Brunswick and London

Hosken J P 1982 *The Hosken Report: Genital and Sexual Mutilation of Females*. Women's International News Network, Lexington, MA

Koso-Thomas O 1987 *The Circumcision of Women: A Strategy for Eradication*. Zed Press, London

Kennedy J G 1970 Circumcision and excision in Egyptian Nubia. *Man* **5**: 175–91

Leonard L 2000 'We did it for pleasure only': Hearing alternative tales of female circumcision. *Qualitative Inquiry* **6**(2): 212–28

Lionnet F 1992 Identity, sexuality, and criminality: 'Universal rights' and the debate around the practice of female excision in France. *Contemporary French Civilization* **16**(2): 294–307

Mackie G 1996 Ending footbinding and infibulation: A convention account. *American Sociological Review* **61**(6): 999–1017

Murray J M 1974 *The Kikuyu Female Circumcision Controversy with Special Reference to the Church Missionary Society's 'Sphere of Influence'*. Ph.D., University of California, Los Angeles. Available from Silverprint-Order Entry

Roheim G 1932 Psycho-analysis of primitive cultural types. *International Journal of Psycho-analysis* **XIII**: 1–221

Seligmann C G 1913 Some aspects of the Hamitic problem in the Anglo-Egyptian Sudan. *Journal of the Anthropological Institute* **XLIII**: 593–705

Shell-Duncan B, Hernlund Y 2000 Female 'circumcision' in Africa: Dimensions of the practice and the debates. In: Shell-Duncan B, Hernlund Y (eds.) *Female 'Circumcision' in Africa: Culture, Controversy, and Change*. Lynne Rienner Publishers, Boulder, CO

Talle A 1993 Transforming women into 'pure' agnates: Aspects of female infibulation in Somalia. In: Broch-Due V, Rudie I, Bleie T (eds.) *Carved Flesh/Cast Selves: Gendered Symbols and Social Practices*. Berg, Oxford/Providence

Talle A 1994 The making of female fertility: Anthropological perspective on a bodily issue. *Acta Obstetricia et Gynecologica Scandinavica* **73**: 280–3

Toubia N 1995 *Female Genital Mutilation. A Call for Global Action*. Women Inc., New York

Walley C J 1997 Searching for 'voices': Feminism, anthropology, and the global debate over female genital operations. *Cultural Anthropology* **12**(3): 405–38

Widstrand C G 1964 Female infibulation. *Studia Etnografica Upsaliensia* **XX**: 95–124

A. Talle

Feminist Economics

This article analyzes the consequences of the underrepresentation of women in the foundation, developments, and applications of economic science.

The use of the term 'feminist economics' appeared in the 1990s with the emergence of a new, extensive, and evolving literature challenging economic theory and research by applying a feminist perspective in all fields of economics. This research, focusing on the relationship between gender and economics, and exploring the extent to which gender bias has influenced economics in all its fields, is based on the idea

that conventional economics is neither universal nor impartial, but that its concepts are built on masculine concerns.

Pluralism and diversity are the key characteristics of feminist economics and they appear in every aspect of the subject. The three main approaches to research in feminist economics are as follows:

(a) one group of studies is devoted to the analysis of the under-representation of women in the economics profession and applies discrimination theories to the profession of economics, stressing the consequences of the absence of women;

(b) a second approach, sometimes described as the 'add women and stir' approach, consists of studies in which a gender component is simply added into classical economics in order to make it more realistic, but without calling into question the fundamental hypothesis and the tools of analysis; and

(c) a third approach comprises studies that fundamentally re-examine the bases and hypotheses of so-called 'universal' and 'gender neutral' mainstream economic theories.

1. *Feminism in Economics*

The contemporary literature on women in economics developed during the course of the 1960s. Its development was linked to a reassessment of the microeconomic analysis of the behavior of households to adapt to the considerable change in family structures that took place and was due partly to the sustained and continued growth of female labor supply. Within couples, two-earner households have in many countries become the majority. An increase in the number of lone-parent families, mainly women raising their children by themselves, has also been noted.

The use of the label 'feminist economics' appeared more recently with the emergence of a new school of economic thought still marginal within the traditional discipline of economics.

Over the last decade, there has been an increase in research focusing on the relation between gender and economics, which is due largely to the dynamism of the International Association for Feminist Economics (IAFFE), established in 1992 in the USA to provide a feminist perspective in the economics profession. IAFFE annual summer conferences, held since 1992, and the publication of a scholarly journal, *Feminist Economics*, first published in 1995, have contributed considerably to the development of a fruitful challenge of economic theory and research, applying a feminist perspective in all areas of economics. The Association now has over 600 members based in 38 countries.

At the same time, and in a similar vein, the first international conference on feminist economics, entitled 'Out of the margin: Feminist perspectives on economic theory,' was held in Amsterdam in 1993 with over 300 participants from more than 20 nations. This proved to be a major event, bringing together a series of groundbreaking papers focusing on the relation of gender to economic science. In 1995, a research network working on issues pertaining to gender and work, called 'MAGE,' was created in France, fostering comparative work and cross-disciplinary discussions.

The 1990s also saw the publication of the first anthologies of feminist economics, by Ferber and Nelson (1993), Humphries (1995), Kuiper and Sap (1995), Hewiston (1999) and Peterson and Lewis (1999).

The establishment of the IAFFE in the USA and of MAGE in France and their success testify to the inability of gender-neutral associations to reflect the questions and interests of feminist economists: in these associations, the feminist perspective is anything but mainstream; its sessions generally have a limited audience. The scant interest shown by economists in gender aspects is reflected in the findings of Randy Albelda's (1995) survey of 400 US-based members of the American Economic Association. The salient results (reported by Hewiston (1999, p. 29)) showed that the surveyed economists believed that feminism had made virtually no impact on the methodology of economics or on economic theory (exceptions: labor market analysis 12.3 percent, and household behavior and family economics 11.1 percent); less than one-fifth believed that more space should be devoted in economic journals to 'feminist perspectives on economic analysis'; more than 70 percent either strongly disagreed with or had no opinion on the statement that 'mainstream economics would be enriched if it incorporated more feminist analysis.'

Practitioners of the male sex heavily dominate economic science, and this dominance has heavily influenced the working hypotheses, theories, and methods developed and used. Whereas the discourse associated with research of a high standard is that of objectivity and universalism, the choice of themes covered and methods of analysis would appear to derive inevitably from social constructs which, in the case of economics, are a reflection of tradition: 'Victorian ideology is at the root of the treatment of women in the neoclassical paradigm' (Pujol 1995, p. 17). Economic science is not neutral, it is not a universal science, and it is permeated with sexist hypotheses. The pivotal hypothesis of the rational economic man is not an abstract, sexless concept: the rational economic man appeared to be a male, white, and Western person.

The androcentrism typical of theoretical and applied economic research has engendered bias, particularly in the analysis and design of economic policies, by ignoring the fact that assumptions about gender influence virtually all aspects of economic reasoning. It is this bias, evident in all fields of economic science, which is nowadays being highlighted in publications by feminist economists.

2. Under-representation of Women in Economics

The founding fathers of economics generally regarded women as separate beings, confined to household tasks and devoid of rational behavior. By the same token, the contributions of female economists to the history of economic thought have often been played down or ignored. Even today, female economists are under-represented in positions of responsibility, despite the sizeable growth in the number of female economics students, whose examination results often surpass those of their male peers.

2.1 Women in Economic History

In a forceful book, Pujol (1992) conducted an extensive historical study of the treatment of women by the founding fathers of classical and early neoclassical economics. In a later essay (Pujol 1995, p. 18) she analyzed five elements that characterize early neoclassical views of women based on the writings of the founding fathers, Marshall, Pigou, Edgeworth and Jevons:

(a) All women are married, or if not yet, they will be. Similarly, all women have or will have children.

(b) All women are (and ought to be) economically dependent on a male relative: father or husband.

(c) Women are (and ought to be) housewives; their reproductive capacities specialize for that function.

(d) Women are unproductive (whether absolutely or relative to men is not always clear) in the industrial workforce.

(e) Women are irrational, they are unfit as economic agents, and they cannot be trusted to make the right economic decisions.

These elements, which fell within the unstated and undiscussed assumptions, justified the poor position of women in the labor market or their exclusion from it. Pujol (1995, p. 29) argued that 'approaches to women in the neoclassical paradigm have not changed much since the founding of the paradigm in the last decade of the nineteenth century.'

Pujol (1992) also underlined another important aspect of the under-representation of women in economics, namely the underestimation of the contribution by female economists to the history of economic thought. She showed the influence of Harriet Taylor on the work of John Stuart Mill as well as her personal contribution, and presented various works on the condition of women by Bodichon, Millicent Garret Fawced, Adav Heather-Bigg, Eleanor Rathbone and Beatrice Potter Webb that were ignored by historians of economics.

One example, among many others, of the underestimation of women's contributions to economics is the scant attention paid to the writings of Margaret Reid (1896–1991), who worked on households, and of Charlotte Perkins Gilman (1860–1935), who focused specifically on gender in economic relationships. 'Reid's colleagues at the University of Chicago, Theodore Schultz and Gary Becker, incorporated Reid's idea of household production into their 'new home economics' (Woolley 1999, p. 335). 'Reid's contribution has been largely ignored in the histories of the 'new home economics,' which Becker is credited with "fathering"' (Lewis 1999, p. 437).

2.2 Women in the Profession

As in most professions in Europe and the USA, there is vertical segregation among economists: women are under-represented in positions of responsibility, in all sectors of activity.

In academic circles they do not occupy the most senior positions and are under-represented in the most prestigious institutions. They do not win prizes: not one woman has so far received the Nobel Prize for Economics; only exceptionally do they chair associations of economists, both in the USA and in Europe; likewise, their representation on the boards of journals is poor.

In the publicity leaflet for *Who is Who in Economics* (Blaug 1999), only two out of 19 photographs are of women: Joan Robinson and the newcomer Deirdre McCloskey.

This segregation is also to be found in the field of public administration as well as private companies. The glass ceiling, that invisible barrier hindering women's access to top jobs, is operating efficiently in economics.

Whereas there is everywhere a growing number of female students and graduates (first degrees) in economics, whose examination results are in no way inferior to those of their male peers, fewer women proceed to Ph.D.s and pursue academic careers. The factors put forward in an attempt to explain these behavioral differences include the following:

(a) The absence of women in the classical economics textbooks and the choice of male-oriented or even phallocratic examples to illustrate theories: the white Western male is the norm in economics textbooks, but he is concealed behind a gender-neutral, universal façade. Similarly, few pages deal with gender and inequality: women are excluded as objects of investigation.

(b) A classroom environment unfriendly to women, the shortage of female lecturers to act as role models, the examination process, the intensive use of mathematical tools and abstract modeling have also been mentioned; however, the superior examination results of young women in mathematics do not corroborate this last explanation.

Moreover, in the same way as women occupy a traditional role in civil society, if female economists are to be acknowledged and valued by their male colleagues they must confine themselves to narrowly circumscribed fields: home production, voluntary

work, discrimination. Women economists have often written about so-called 'women's issues such as women's labor force activity and wages, which seemed of little interest to male economists and benefit from a lower level of reward than other fields.'

The Massachusetts Institute of Technology established a committee on women faculty to understand and ameliorate the long-term effects of discrimination in academia: their key conclusion is that (Massachusetts Institute of Technology 1999, p. 3) 'gender discrimination is subtle but pervasive and stems largely from unconscious ways of thinking that have been socialized into all of us, men and women alike. … Our first instinct is to deny that a problem exists, or to blame it on the pipeline or the circumstances and choices of individual women.' These conclusions could probably be extended to most faculties and universities in the world.

3. *Consequences of Women's Under-representation in Economics*

Women's under-representation engenders bias and error in the foundations, development and applications of economic science. The androcentric bias in basic neoclassical hypotheses is substantial: historically the entire theory of utility is based on the rational behavior of a *Homo economicus* who may choose to allocate his time to paid work or leisure. Only in recent economic research have male economists realized that life does not boil down to only a choice between paid work and leisure.

Women and the family are invisible in the economic analysis of human behavior: even when gender is central to the subject under investigation it is often ignored, as evidenced by studies of poverty.

3.1 *Sex Discrimination*

The question of inequality between the sexes has not been the subject of any key works of normative economics; instead, discrimination between men and women is either included in the broader issue of social inequality or ignored.

Discrimination has mainly been the subject of writings in applied and comparative economics, which have quantified it and measured its development but have left extensive shady areas in respect of its causes and dynamic. Two approaches have been followed in formulating the models appraised: human capital theory, which explains the differences evident between men and women on the labor market through differences in the investment in human capital—education, training, experience, health—and discrimination theories, which analyze the differences between individuals whose production capacity is identical—discrimination exists when persons of equal capacity and with the same skills are treated differently because of their sex (Blau and Ferber 1992).

The factors explaining the segregation of men and women are to be found in the theories of labor supply and demand, institutional theories, and transaction cost theories (Jonung 1998, pp. 36–71). Labor-supply theories highlight the differences between the preferences and aptitudes of men and women, the socialization of roles and the domestic division of labor, all of which affect decision making; demand theories explain segregation in terms of the differing investment in human capital and the notion of discrimination that colors the attitude of employers towards women. Transaction-cost theories point to the incomplete information available to employers, which leads them to adopt discriminatory attitudes based on observations of average behavior; and finally, institutional theories stress the legal and social constraints, the traditions that influence individual decision making.

3.2 *Household Economics*

For feminist economists, existing economic theory and analysis are unsatisfactory when it comes to studying intrafamily relationships and the distribution of resources within the family unit. According to the neoclassicists, the household or family is a black box: it matters little who produces or who consumes resources; this share-out is a technical matter or one of preference, something of no concern to economic analysis.

The situation of households is characterized by global earnings or global consumption managed by an altruistic head of household who looks after the interests of his loved ones and behaves as a rational economic agent.

This simplistic view can be found in most European-level comparative studies on poverty or earnings inequality. The hypothesis is that resources are shared equally among the members of a household: poor individuals belong to poor households. Households are classified according to the age, occupation, and sector of activity and educational level of the head of the household. Rarely does the composition of the household feature, other than through the simplistic application of a scale of equivalence.

However, certain studies have underlined the importance of power in the distribution of earnings and consumption within households, not merely of preferences and abilities, the sole elements to impinge on the reasoning of economists. Recent empirical research has confirmed that the allocation of resources among family members depends on age and gender. Studies conducted in Asia and Africa have emphasized the high rates of female mortality due to the systematic deprivation of resources suffered by women within households (not necessarily poor ones; see Sen (1989)).

In one of the few studies on distribution within American households, Lazear and Michael (1988) reveal that the division of consumption between parents and children alters radically with the number of adults in the household: as the number of adults rises, the children's share declines proportionately. Furthermore, access to paid work considerably changes the share-out of power within the household. On average, women earning an income have more decision-making power within the family, and this share of the power increases in step with their earnings (England and Kilbourne 1990).

Findlay and Wright (1996, pp. 335–351) have assessed, for Italy and the USA, the way in which an unequal distribution of resources within households affected the measurement of poverty. They conclude that the conventional methods of gauging poverty within households lead to a systematic underestimation of female poverty and an overestimation of male poverty.

The criticisms of feminist economists have addressed the lack of analysis of the household as an economic entity, but they have also addressed the problematic analysis of the household in the new home economics: 'For feminists, the minimum requirement is a model of the household that allows for the conflict of interests and patriarchal power relations within and outside the family' (MacDonald 1995, p. 185). New home economics, an example of the 'add women and stir' type of studies, is based on an implicit hypothesis of women's natural superiority in nursing and housekeeping which justifies, from the point of view of efficiency, the division of labor within the family. It used to treat the sexual division of labor in the family as if the differences were biologically predetermined and not socially constructed. The neoclassical model or child support model disadvantages women by obscuring and justifying men's neglect of their parental responsibilities.

Bargaining theories and game theory are also becoming increasingly popular in dealing with intrahousehold allocations: the game theory model of household decision making is an interesting approach to overcome the restrictive assumptions of household intrafamily consent and formally develop a more realistic family behavior pattern. One application is the treatment of 'fertility as a prisoner's dilemma' (Ott 1995, p. 89). However, formal modeling limits the reality of analysis and the 'cooperative conflict' approach developed by Sen (1990) and Agarwal (1994) which incorporate qualitative, historical, and cultural variables provides an interesting alternative to game-theoretic models.

3.3 Labor Economics

Labor economics, along with the new home economics, is the first field to incorporate gender into the neoclassical framework. Gender has been included in the models without challenging the theoretical foundations; it was merely a matter of arriving at a more accurate analysis of observable reality.

The mainstream understanding of women's employment in labor economics is 'men as breadwinners and women as second earners.' The concentration of women in low-wage occupations and atypical employment appears to be a result of their own choice.

In fact, there is circular reasoning at work here: women choose less well-paid posts which enable them to take on the domestic chores for which they have a preference and are made in the knowledge that the performance of domestic chores reduces the earnings which they could command from employers.

Numerous econometric studies (Silvera and Sonnac 1996) have attempted to explain the persisting wage gap between women and men: even when controlling for differences in working time, educational attainment, training, tenure, occupations, sectors, etc., an unexplained gap amounting to 11–14 percent of women's wages remains. This gap, a measure of pure discrimination, exists even in the countries where protective laws concerning wage equality are in application. However, and contrary to neoclassical principles, women's lower wages are not associated with lower unemployment rates: in nearly all European Union countries women's unemployment rates are considerably higher than men's.

3.4 Data and Methods in Empirical Economics

The same bias is evident in the development of statistical tools: to include a feminist perspective in empirical economic analysis, new kinds of data and indicators have to be constructed, and there is a need for more qualitative data and for panel data sets.

The available data concerning the labor market, for example, are constructed around standard male practice: full-time paid work without interruptions. In most cases these are individual data with no information as to the individuals' household or lifestyle.

The data available from national statistical agencies to study women on the labor market are inadequate in many ways (MacDonald 1995, p. 188):

(a) there is a preponderance of cross-sectional data, and very few panel data;

(b) the level of aggregation makes it impossible to study issues of segmentation within occupations;

(c) labor force data are collected on the individual, with information on the family context either missing or inadequate;

(d) the use of categories such as 'personal or family reasons' versus 'economic reasons' in labor force surveys, to classify reasons for not looking for work or for working part-time, embed gender biases in the data;

(e) skill measures have been shown to be gender biased (Boyd 1990); and

(f) it is difficult to measure changes in the quality of jobs.

National accounting ignores all non-monetary aspects: women's traditional contribution to growth—domestic work and child rearing—are never quantified. To develop alternative indicators of economic and social welfare is an important component of feminist economic research.

Bias in measuring economic growth, ignoring domestic work, leads to wrong political decisions insofar as no account is taken of cost when policies negatively affect home production. The fact that no value is attributed to home production, and that it is statistically overlooked, leads to an inferior allocation of public expenditure.

3.5 Policies

One example of the adverse effects of mainstream economics on economic policy is the structure of social protection systems. These often refer back to the neoclassical patriarchal model: a single main income provider per household. This reference engenders discriminatory practices, which can adversely affect the female supply of labor and can keep women in a situation of dependence. The existence of discriminatory measures, the trend in family structures, and the contradictions between the labor market and social security systems are three elements that lie behind the issue of individualizing social security entitlements.

As they operate at present, Europe's social security systems exert a discriminatory effect on women by two means. First, these systems do not ensure adequate protection for workers in atypical circumstances, among whom women are over-represented: such persons are often deprived of proper protection, mainly with respect to unemployment and pensions. Second, the absence of individual entitlements and the proliferation of derived rights forces women into relationships of dependence tinged with insecurity, which discourage them from entering the official labor market and results in inequalities between the benefits received according to family circumstances. Europe's taxation and social security systems contain various 'inactivity traps' which were deliberately set and still exist: derived rights in social protection systems are an incentive for women to remain at home or to engage in undeclared work, since marriage automatically confers on a spouse the healthcare and pension benefits enjoyed by the partner. The advantages granted to the home-based spouse by the tax system, namely due to dependents' allowances, as well as all the allowances linked to inactivity, such as career breaks, constitute inactivity traps that affect a far larger number of persons of working age than the unintentional traps that are currently being singled out for corrective action (Jepsen et al. 1997, pp. 81–98).

Whereas the Welfare State gave men their independence in the market, it also reinforced the economic dependence of women on men by ignoring the sexual division of labor.

4. The Future: a Reconstruction of Concepts

Ferber and Nelson (1993, p. 8) identified five, not necessarily mutually exclusive, responses to perceived inadequacies in the academic discipline:

(a) Affirmative action: based on the idea that the central problem is the under-representation of women and acting to correct this by monitoring but not criticizing the discipline itself.

(b) Feminist empiricism: not calling into question the tools but rather the way in which they are used and the manner in which the issues are addressed: 'It is not the theory that is patriarchic, but the questions male economists have asked and the conclusions they have drawn and particularly the policy implications based on the research' Gustafsson (1990, p. 6); Ferber and Nelson (1993, p. 8) noted that 'Casual conversation suggests that most feminist economists currently adhere to this view.'

(c) Feminist difference: the replacement of male bias by a feminist bias, not a quest for objectivity but subjectivity, emotion, holistic approach (less well represented).

(d) Feminist postmodernism: to deconstruct traditional conceptions, the key point: is gender a meaningful categorization?

(e) Feminist constructionism: analyzing the social production of gender identity.

Whatever the classification, the reconstruction of economic concepts underpins the work of feminist economists: the concepts used for analysis and the way concepts are linked together in the theory may be better suited to explain the activities of some groups than others: they are culturally biased and show a tendency to treat male values and practices as universal. A gendered view applied to all parts of economics questions the hypothesis, methods, and data used.

The aim is also to take greater account of complex social constructs that have been ignored or inadequately incorporated into economic modeling. This means taking account of societal variables often neglected by economists, and implies a reconstruction of traditional databases that are unsatisfactory for gender mainstreaming.

Feminist economics also considers gender as a central category for analysis: a feminist reconstruction of economic theory can improve all the different fields; this new economics will not just address feminist issues, but will be better economics.

See also: Economic Development and Women; Economic Globalization and Gender; Economics, History of; Economics: Overview; Feminist Epistemology;

Feminist Political Ecology; Feminist Theory; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Sociology

Bibliography

Agarwal B 1994 *A Field of One's Own: Gender and Land Rights in South Asia*. Cambridge University Press, Cambridge, UK

Albelda R 1995 The impact of feminism in economics—beyond the pale? A discussion and survey results. *Journal of Economic Education* **26**: 253–73

Barker D K 1999 Gender. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK, pp. 393

Blau F D, Ferber M A 1992 *The Economics of Women, Men, and Work*. Prentice Hall, Englewood Cliffs, NJ

Blaug M (ed.) 1999 *Who is Who in Economics*. Edward Elgar, Cheltenham, UK

Boyd M. 1990 Sex differences in occupational skill: Canada 1961–1986. *Canadian Review of Sociology and Anthropology* **27**: 287–315

England P, Kilbourne B S 1990 Marriages, markets, and other mates: The problem of power. In: Friedland R, Robertson A F (eds.) *Beyond the Market Place: Rethinking Economy and Society*. Walter de Gruyter, New York, pp. 163–8

Ferber M A, Nelson J A (eds.) 1993 *Beyond Economic Man—Feminist Theory and Economics*. University of Chicago Press, Chicago

Findlay J Wright R E 1996 Gender, poverty and the intra household distribution of resources. *Review of Income and Wealth*, Series 42, No. 3, September, pp. 335–351

Gustafsson S 1990 Half the power, half the incomes and half the glory: The use of microeconomic theory in women's emancipation research. Inaugural Lecture at the Official Assumption of the Chair of Labor Market Issues with Special Attention to Women's Emancipation. University of Amsterdam

Hewiston G J 1999 *Feminist Economics. Interrogating the Masculinity of Rational Economic Man*. Edward Elgar, Cheltenham, UK

Humphries J (ed.) 1995 *Gender and Economics*. Edward Elgar, Aldershot, UK

Jepsen M, Meulders D, Plasman O 1997 Protection sociale: le rôle des droits dérivés. *Les Cahiers du Mage* No. 3–4, pp. 81–98

Jonung C 1998 Occupational segregation by sex and change over time. In: Persson I, Jonung C (eds.) *Women's Work and Wages*. Routledge, London, pp. 36–71

Kuiper E, Sap J (eds.) 1995 *Out of the Margin*. Routledge, London

Lazear E P, Michael R T 1988 *Allocation of Income Within the Household*. University of Chicago Press, Chicago

Lewis M 1999 History of economic thought. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK, pp. 433–43

MacDonald M 1995 The empirical challenges of feminist economics. In: Kuiper E, Sap J (eds.) *Out of the Margin*. Routledge, London, pp. 175–97

Massachusetts Institute of Technology 1999 *A Study on the Status of Women Faculty in Science at MIT*. Massachusetts Institute of Technology, Boston, mimeo

Ott N 1995 Fertility and division of work in the family. In: Kuiper E, Sap J (eds.) *Out of the Margin*. Routledge, London, pp. 80–99

Persson I, Jonung C (eds.) 1998 *Women's Work and Wages*. Routledge, London

Peterson J, Lewis M (eds.) 1999 *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK

Pujol M A 1992 *Feminism and Anti-feminism in Early Economic Thought*. Edward Elgar, Cheltenham, UK

Pujol M A 1995 Into the margin. In: Kuiper E, Sap J (eds.) *Out of the Margin*. Routledge, London, pp. 17–34

Sen A 1989 Women's survival as a development problem. *Bulletin of the American Academy of Sciences* **43**: November

Sen A 1990 Gender and cooperative conflicts. In: Tincker I (ed.) *Persistent Inequalities: Women and World Development*. Oxford University Press, New York, pp. 123–49

Silvera R, Sonnac N 1996 *Le Salaire des Femmes: Toutes Choses Inégales par Ailleurs*. La Documentation Française, Paris

Woolley F 1999 Family, economics of. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK, pp. 328–35

D. Meulders

Feminist Epistemology

Reflecting on feminist epistemology, namely the claim that there are distinctively feminist perspectives on the theory of knowledge, initially proposed at the end of the 1970s, is a little like recalling the English Chartist movement with its utopian program of radical political reform. Within what historians would regard as rather little time, what was seen as an absurd and impossible political project, not least the demand for universal suffrage, was to become an everyday reality. The utopian dreamers turned out to be more practical than the dismissive establishment. The equally utopian proposal for a distinctively feminist epistemology was placed firmly on the theoretical agenda of the second wave of feminism by the end of the 1970s. Within just two further decades, substantial aspects of the feminist epistemology debate began to find acceptance, not just among the diversity of feminists, but much more widely within culture and society.

That hot debate among the scholars and activists of the women's movement, which gave birth to the claims for feminist epistemology, took rather little notice of disciplinary boundaries. The new field they were in the process of creating, 'Women's Studies,' was self-consciously not only highly innovative but was also transgressive. Having found that the academic gaze was blind to the bodies and lives of women, to say nothing of sex–gender relations, there seemed little purpose in deferring to the canon of the old exclusionary disciplines. Thus the new epistemology drew in theorists whose disciplinary formation lay in philosophy, psychoanalysis, biology, history, physics, and sociology, but whose cultural products often showed little trace of their formation. These feminist epistemology debates, without quite ever settling down and forging an unequivocal consensus, have entered

and helped foster profound changes in many academic disciplines, particularly where the mode of production of knowledge requires rather modest capitalization. The so-called 'Little Sciences' have proved much more susceptible to change than the 'Big Sciences.' Thus the most power charged techno-sciences of molecular biology and informatics are proving to be remarkably resistant to the kinds of internal changes that primatology or sociology have experienced.

1. *The Birth of Academic Feminism*

Despite the power of late twentieth century feminism to touch women worldwide, the development of academic feminist theorizing required financial support. During this key early period support given to Women's Studies by both the Ford and Rockefeller foundations enabled US academic feminism to develop more strongly than its counterpart within Europe, not least in the development of feminist approaches to the theory of knowledge. By contrast, although Scandinavian research policy was swift to support Women's Studies, it did so as part of policy-oriented, rather than as basic research, while in Western Europe, with the exception of The Netherlands, support from either the state or from foundations was slow to arrive. Thus, while there was a rich outpouring of articles and books published by feminists from many countries criticizing existing theories of knowledge and proposing alternatives, the confident assertion that there were distinctively feminist perspectives on epistemology had to wait until the publication of *Discovering Reality* (Harding and Hintikka 1983). The editors, the US philosophers of science, Sandra Harding and Merrill Hintikka, described their theoretical project.

During the 1990s, feminist thinkers provided brilliant critiques of the political and social beliefs and practices of patriarchal cultures. But less attention has been given to the underlying theories of knowledge, and to the metaphysics which mirror and support patriarchal belief and practice. Are there—can there be 'distinctive feminist perspectives on epistemology, metaphysic, methodology, and philosophy of science' (p. 9)?

The US contributors, divided between deconstructionist accounts of the values embedded within organized knowledge or scientia, and those which try to set out the conditions for rebuilding these as feminist knowledges. A key target was the natural sciences, which had, with immense success, claimed to be above culture. In setting the natural sciences as their target, the feminist critics of science shared more than a little in common with the new, post-Mertonian, sociology of science, although there was little visible connection between the two. However, the very title of the book indicates that these were daughters, albeit dissident daughters, of the enlightenment. As feminist critics of science, they sought to replace what they saw as flawed accounts of reality with more reliable accounts free from ideologies and relations of domination. For this generation of theorists, the fact that reality was both real and knowable was not significantly at issue; the nominalism which was to sweep the 1980s and 1990s, was not, at that moment, a powerful current. Recognizing the power of foundationalist thought, the authors set out to analyze the patriarchal constructions of those cultural giants who had deeply shaped Western culture from antiquity, to the early theorists of modernity to the late modern figures of Darwin, Marx, and Freud. Within the natural sciences it was not by chance that the focus was Darwin and the life sciences, and not, say Einstein and physics, for biology was the discourse which claimed cultural authority to define women's nature.

The definition of women's nature was an old battleground, with pioneer nineteenth century feminists arguing that in Darwinian theory, only men evolved. Unquestionably, scientific sexism along with scientific racism demanded critical attention; clearing away the culture of biology-as-destiny would help create a political space in which new human subjects could claim agency. Where in the past such detailed critiques had been mounted predominantly by progressive men biologists, as women entered the laboratories, their consciousness raised by the second women's movement, they too had the skills to enter the lists in their own right. The pioneers among these feminist biologists include Ruth Hubbard, Marian Lowe, Ethel Tobach, Betty Roscoff, Anne Fausto Stirling, Ruth Bleier, and Lynda Birke. These insisted that the androcentricity which produced and reflected a patriarchal culture also produced poor biology, and that a more faithful, more reliable account of women's nature could only be produced through feminist science resisting the relations of domination. It was the pioneering work of Carolyn Merchant (1980) as a historian of science that linked the problem of gender domination with the domination of nature. While this link by the feminist historians opened a continuing connection with the burgeoning environmentalist movement, the feminist life scientists made a powerful connection with the women's health movement. Both had, and have, epistemological and political implications.

Their deconstructionist critiques (without Derrida) constituted an immense methodological challenge to contemporary biomedical research. They charged successfully that biomedicine had taken the body of the male of European bio-geographic ancestry, or 'race,' and fetishized it as the universal human body. They demonstrated forcefully the ways through which women's bodies, apart from their problematic reproductive systems and their troublesome psyches, along with the bodies and minds of other Others, were subsumed within a demonstrably false universalism. Under the impact of this epistemological questioning,

together with the increasing presence of historically excluded groups within the laboratories, the discourse of biomedical research has gradually but dramatically changed, above all in the study of the pre-eminent lethal diseases of the West. Thus, heart research now routinely explores difference without hierarchy, where previously the study of the heart of the pale male was equated with some universal human organ. Researching cancer also underwent a sea change. It was not simply the immense pressure to address breast cancer seriously but, almost ironically, this attention to sex–gender difference has led to greater attention being paid to the specific problems of men, not least prostate cancer. Even mental health began to address gender and 'race' difference without automatic notions of domination. This was feminist epistemology in action.

2. *The New Feminist Epistemology*

This successful critique of biomedical science as skewed and inadequate, bad for women and bad for science, fostered an initial comfortable solidarity between the pioneer feminist science critics, embracing both the feminist biologists and those, typically from the social sciences and humanities, engaged in setting out the theoretical project of a distinctively feminist epistemology. One of the most influential theorists of the new feminist epistemology was political scientist Nancy Hartsock (1983), who formulated and named feminist standpoint theory. Her socialist feminist project revisioned Marx's theory of knowledge. Taking the project of a proletarian science, she runs it through the sieve of feminist analyses of the specificities of women's reproductive labor. She argues that it is a feminist consciousness that gives rise to the distinctively feminist standpoint. This approach was shared by Harding in the same volume, and by sociologist Hilary Rose in her *Signs* article (1978). In the important earlier natural science issue of *Signs*, the historian Donna Haraway (1978) had already spelled out the need to rebuild the life sciences outside the relations of dominance. In this project of a successor science these feminists reflected the strong Hegelian and Marxist influences in their work. The heretical claim that nature and not just society could be more reliably 'seen from below' challenged the hegemony of the claims of the natural sciences to be a culture of no culture.

Feminist standpoint theory, particularly when it was directed to the study of nature, represented a theoretical development from the 'two sciences thesis,' articulated initially as the struggle between proletarian and bourgeois science by the social relations of the science movement of the 1930s and 1940s, lost in the Stalinist period, but subsequently re-articulated as a rainbow politics of 'science for the people' by the radical science movement of the 1960s and 1970s. Other influential feminist science theorists came at the epistemology question from very different theoretical and political trajectories. Thus, philosopher Helen Longino, while concerned to define a feminist science, did so by specifically repudiating the Hegelian master/slave argument; for her the view from below carried no epistemic privilege. Evelyn Fox Keller, by contrast, took a psycho-historical approach, focusing on the gendered origin stories of science. She saw masculinity as embedded in that mechanical philosophy which had triumphed in those founding years of both modern science and capitalism.

Not only did early feminist epistemology theorizing show little deference to the existing canon, initially it paid little formal attention to the paradigm-breakers emerging from the crisis of French social thought in the wake of the defeat of 1968. Few of these early feminist epistemological texts paid attention to the theorists of post-structuralism and post-modernism. As noted above, self-proclaimed feminist deconstructionists paid little attention to Derrida or to the discourse analysis of Foucault, or to Lyotard's post-modernistic project, even though debates were raging around these in the mainstream disciplinary discourses. Despite the parallels which have led to some over facile assertions such as 'feminism is part of post-modernism,' these new currents were relatively slowly taken into feminist theoretical production. They first entered the humanities and then, with continuing resistance, the social sciences.

3. *Disciplining Feminism?*

Feminism's successful arrival within the academy by the end of the 1980s brought with it substantial epistemological implications, not least the restoration of disciplinarity within feminist discourse. Feminists began to address the specific canon and current debates of their own disciplines; the ethical and political commitment to address the feminist movement both in and out of the academy increasingly weakened. The language of feminist theorizing soon changed. The early commitment to transparency and accessibility ceded to a new complexity, even obscurantism. There was a sudden interest in post-modern theorists, whom an early, more politicized, generation of feminists had set aside for their anti-feminism; earlier theorists such as Nietzsche became widely discussed despite an evident anti-semitism and misogyny. While disciplinary boundaries appeared, the claims of theorists became grandiose. Feminists in the humanities often spoke of 'feminist theory,' by which they meant more precisely feminist humanities theory, but tacitly let the term carry with it the aggrandizing claim that this had universal power over the diversity of fields. Even feminist sociology, something of a bastion of critical realism, entered a period of self-questioning. When the differences between women, whether of class, ethnicity, and bio-

geographic ancestry were so very visible, how did it make sense to speak of a single feminist standpoint? Theoretical developments responding to these challenges by feminist sociologists such as Canadian Dorothy Smith, German Frigga Haug, and African American Patricia Hill Collins gave standpoint theory a new and enriched life.

4. Feminist Science Wars?

Among feminist science studies scholars it was the historian Donna Haraway who embraced post-structuralism decisively, simultaneously throwing into question the possibility of universal reliable knowledge. In a key paper (which echoed silently philosopher David Bloor's methodological stance of refusing to privilege the truth claims of the natural sciences), Haraway, despite her respect for Hubbard as a pioneering feminist biologist, pointed to her uneven treatment of truth claims. She argues that for Hubbard, good science remained privileged; only bad science was to be deconstructed. This epistemological difference between the feminist scientist and the feminist historian foreshadowed, with some strained courtesy, the divisions among the feminist approaches to science theory which were to develop.

However, the tensions between feminists science theorists were never characterized by the savagery of those within mainstream epistemological fights, above all those of the Science Wars. Science theory feminists, despite their theoretical differences, frequently shared a desire to build solidarity between those in science studies and those in science. Thus it was matter of common recognition that natural scientists, particularly biologists, had to be realists. It was equally clear that, for science theorists, more relativistic positions were entirely possible, whether they were historicized as by materialism or full-blown as by nominalism. Feminist science theorists routinely drew back from stark 'either/or' choices, not least those dichotomously opposing relativism to realism. Thus, where the early Harding had shared Rose's concern at the hyper-reflexivity of the strong program of science, and the reduction of science just to social relations proposed by Robert Young and the *Radical Science Journal*, her subsequent publication suggested a new theoretical ambiguity. Certainly reviewers who were outside science theory of *The Science Question in Feminism* (Harding 1986) diversely claimed the book both for standpoint theory and as support for the more relativistic approach of post-structuralism. The ambiguity was not accidental but typical of the feminist science theorists as they fought to avoid an unwelcome dichotomy.

Later, Haraway made a sophisticated attempt to spring the 'either/or' trap: unwilling to lose all claim to objectivity she proposes a feminist objectivity built from 'partial critical knowledge located in multiple embodied sites sharing a web of connections called solidarity in politics.' Harding too made an attempt to go beyond the 'either/or' dichotomy in her advocacy of 'strong objectivity,' an argument which has much in common with the earlier externalist analysis of Marxism.

5. Global Technoscience and Global Feminisms

In a quite extraordinary way, the feminist epistemology debates have paralleled and offered an entirely different face to the widespread demands for equal opportunities within the labor market, including that of science. This, whether claimed for reasons of justice or for economic efficiency, has become subtly linked to arguments about changing the character of cultural production, typically side-stepping whether the claim entails an appeal to social, cultural, or essentialist difference. Today, innovative corporations concerned for their global markets sound remarkably like dedicated multiculturalists in their arguments that products aimed at a socially diverse market are better if designed and produced by a matching diversity of producers. That the former tend to equate better with 'sells more' and the latter equate better with 'is more democratic' and 'will produce a more sustainable science and technology' is almost incidental. Both are responding to the feminist epistemological challenge that who gets to produce knowledge, and who is excluded, matters.

But it has been the advent of what social theorists have termed the risk society, that is, where the development of science and technology have become associated intrinsically with major environmental risks, which has also fostered an immense turn towards democratic control to protect the entire socio-ecosystem. This immense cultural and political move has paralleled the struggle within the sciences. Working from outside science, a mixture of NGOs and mass popular movements from both the North and the South have confronted the technosciences and the corporations. As we have seen with the fate of the Kyoto agreement, individual governments and global corporations continue to mobilize the old constructions of science as outside culture and capable of producing certainty, in order defend themselves against the need to take socio-ecological risk seriously. Nonetheless, although it would be a mistake to announce the environmentalist cause won, it has become less and less easy for the corporations and supportive governments to pursue their commercial and technological objectives without any sense of global responsibility and concern for either human or green nature. Where they display such indifference they find themselves confronted by new alliances capable of mobilizing immense popular criticism. The more sophisticated recognize the need for sustainable development and are more open to new more socially

inclusive forms of governance. How far these are serious or tokenistic is too soon to judge.

It is only possible in a brief article to do little other than hint at the links between the feminist epistemology debates and these huge cultural and political developments. Sometimes, as with the mass people's science movement in India and the work of the feminist science theorist and activist Vandana Shiva, the links are more evident. But it will be historians looking back at this extraordinary shift in the cultural status of science and technology at the cusp of the twentieth and twenty-first centuries, who will have the pleasure of disentangling the cultural and political contribution of the feminist epistemology debate to these sea changes.

See also: Feminist Ethics; Feminist Political Theory and Political Science; Feminist Theory; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Gender and Feminist Studies; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Rationality and Feminist Thought; Situated Knowledge: Feminist and Science and Technology Studies Perspectives

Bibliography

Bleier R 1986 *Feminist Approaches to Science*. Pergamon, Oxford, UK

Haraway D 1990 *Primate Visions: Gender, Race and Nature in the World of Modern Science*. Routledge, New York

Harding S, Hintikka M (eds.) 1983 *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology and Philosophy of Science*. Reidel, Dordrecht, The Netherlands

Harding S 1986 *The Science Question in Feminism*. Open University Press, Milton Keynes, UK

Keller E F 1985 *Reflections on Gender and Science*. Yale University Press, New Haven, CT

Longino H 1990 *Science as Social Knowledge, Values, and Objectivity in Scientific Enquiry*. Princeton University Press, Princeton, NJ

Rose H 1978 Signs: Special issue on women, science and society. *Journal of Women in Culture and Society* **4**(1)

Rose H 1989 *Staying Alive: Women, Ecology and Development*. Zed Press, London

Rose H 1994 *Love, Power and Knowledge: Towards a Feminist Transformation of the Sciences*. Indiana University Press, Bloomington, IN

H. Rose

Feminist Ethics

Feminist ethics has its roots in the more general feminist theory and politics, going back to Mary Wollstonecraft's Declaration of the Rights of Women, and on through the work of Jane Addams and Charlotte Perkins Gilman early in the twentieth century, work that took women's moral visions seriously. As an intellectual discipline in Anglophone nations, it was born in the 'second wave' of the Women's movement in the mid- to late-1960s, taking a few clues from de Beauvoir (1948). The early feminist theories in this 'wave' were written by activists in the movement as position papers to provide conceptual and theoretical bases for charting strategy. The basic concerns and projects of feminist ethics today have grown from the political roots.

As time passed, the feminist activists entered professions across the disciplines. They established journals and courses and eventually the discipline of women's studies itself was legitimized and institutionalized, with its own subspecialties that include feminist ethics. Although feminist intellectual work has always been multidisciplinary, its institutionalization has taken place in the traditional context of departments, course assignments, disciplinary specialties, etc. One outcome has been that feminist ethics, like traditional ethics, has tended to fall within the territory of the humanities, despite the fact that the social sciences also have essential contributions to make.

Intellectually, feminist ethics was shaped by criticisms of the traditional moral theories that were developed to suit life in secular, Western capitalist democracies. Moral subjects and agents were defined as autonomous individuals who were equal under morality and the law—a stereotypically male subject. The traditional theories incorporate, for the most part, a variety of dualisms in which one side is valued over the other, underwriting an ethics that supports dominance and subordination of many sorts. The dualisms fundamentally value reason over emotion, with men identified with reason and mind and women (and the racially subordinated) identified with emotion, body, and nature. Out of this soil a variety of feminist moral theories have grown, ranging from modified, liberal theories of equal opportunity and rights; to theories of mothering, care, and relationship; to lesbian ethics; liberatory ethics; standpoint approaches; third world and minority approaches; and the feminist ecological ethics that is now being developed.

From the beginning, feminists worked in 'applied ethics,' for example, on moral issues of abortion, violence, prostitution, AIDS, clitoral cutting, the environment, peace, immigration, and many others. All of these theorists insist on taking women's experience seriously. In addition, some feminists focus on issues of power and dominance, insisting that feminist ethics must be 'liberatory' and provide guides for action. The liberatory focus can be seen as asking, 'What would an ethics be like that takes patriarchy, or racism, or anthropocentrism seriously as a social and political structure?'

The dominant public ethics in the Western democracies is an abstract moral theory of impartial reasoning, under which autonomous individuals freely chose principles of rights and justice and make their decisions accordingly. It has been used to good effect in some places as women argue for equal opportunity and rights. The impartial ethics is implicitly radical because it is supposed to apply to all individuals regardless of gender, race, age, ethnicity, relationship, etc. In practice, the radical extension to all individuals was politically won during nineteenth and twentieth century struggles to extend the suffrage, change property laws, and so on. In the process, the ethics itself was modified so as to include women (and others) among the autonomous individuals—leading to the intellectual discovery that the original moral agents of the theory were conceived to suit a stereotypically Western, white, higher class, male model. A family of feminist moral theories sprang up to remedy the problem.

The dominant Western ethics of justice, rights, and autonomy relies heavily on the reason–emotion dualism. One feminist alternative has been to remedy the imbalance by emphasizing women's experience of care and responsibility. Long ago, Addams (1907) called for a 'social ethic' that would offer a guide to women's action by expanding women's morality in home and family into the public sphere—the Settlement House movement and the women's peace movement were outcomes. This alternative puts the focus on care and relationship rather than on autonomous agents acting individually according to their interests, self-chosen principles, etc., and it has provided an important direction for feminist ethics. Diana T. Meyers, for example, developed an account of moral autonomy in terms of competency that could be exercised either under an ethics of impartial reason or under an ethics of responsibility and care—using studies in social psychology to support her work. Gilligan (1982) distinguished two 'voices' in moral reasoning, and two developmental modes, one of them charted by Lawrence Kohlberg's scale, the other by a scale she developed through interviewing women about their abortion decisions.

Any moral theory presupposes a social theory and a theory of human nature—often implicitly. Kohlberg (1981) was explicit: his view on human rationality relied on Kant, and so did his social theory, traced back to Kant through Piaget and Durkheim. Like the latter, Kohlberg supposed there is an intimate connection between social organization and development of moral reasoning. Differences in the rate of moral development, he said, are due to differences in opportunities for social role taking—and for the highest stage of development, opportunities in the 'public sphere' institutions of government, law, and economy are required (historically 'men's sphere'), in contrast to the 'domestic or personal sphere' of family, etc. Women in general tested out at the lower stages, and they were diagnosed as confusing moral problems with issues of interpersonal relations.

Kohlberg measured development in terms of subject responses to abstract moral dilemmas. Gilligan developed a different scale for measuring development by asking women about their actual abortion decisions. The basic concepts of her theory were those of care and responsibility in relationships—in contrast to individual rights and justice. She constructed the ethics in the mid-1970s, using interviews with women, but her later work indicates that both men and women use care and responsibility reasoning and develop along her scale.

Gilligan was one of the early theorists of the care and relationship approach, but there were many others, particularly those who focused on mothering as an alternative paradigm to the 'public man' that operated in a society of autonomous moral equals. These 'maternal ethics' concern care and responsibility among those who are not equal in power or ability—those who are cared for and those who give care, those who are dependent and those on whom they depend. Ruddick's (1980) work on mothering was widely influential.

To a degree, these 'care' ethics accept the dualisms of reason/emotion and public/personal, but they argue for the distinctive importance of women's experience in developing an ethics for both men and women. However, there are political implications of accepting the dualisms. Tronto (1993) connects the traditional separation of ethics and politics with the division into public and domestic (or personal) spheres, marking it as a boundary that has excluded women from power. She develops a positive alternative of an 'ethico-politics' of care.

The ethics of care and responsibility makes a major contribution to philosophical ethics by analyzing moral relations among persons who are not equals. For example, Baier's (1995) work on trust considers relations among unequals in a way that is applicable in the 'public sphere'—investigating trust between the governed and the governors. Baier also mined the history of philosophy to find alternatives to the 'rights and justice' approach, taking David Hume as a candidate for 'the women's theorist' because he took sentiment and relationship as basic. Other feminists have broadened the ethics to other personal relationships, such as friendship or love, while some have developed an ethics of relationship within lesbian communities (Hoagland 1988, Mohin 1996).

The 'care and responsibility' approach has sometimes been labeled 'feminine,' rather than 'feminist,' because it emphasized women's traditional place and feminine virtues. Proponents of what is labeled 'feminist ethics' claim that any feminist theory must be liberatory, or at least must provide a basis for arguing issues of injustice and domination in the public sphere. For this reason, many feminists have worked to inte-

grate the care and justice approaches (see Held 1995). They insist not only that women's experience be taken seriously, but that patriarchy, racism, class, and other oppressive social structures be taken seriously. Jaggar (1995) has offered an agenda for any ethics that is called feminist: the ethics must underwrite critiques of actions and practices that perpetuate women's subordination; it should prescribe morally justified ways of resisting those practices; and it should envision morally desirable alternative promises of women's emancipation. Feminist ethics should offer a guide to feminist action.

Traditional ethics was criticized because the ideal of a moral agent was modeled on certain men's experience, excluding the experience of women. Both feminine and feminist ethics have themselves been criticized for using an ideal of Western white, middle-class women's experience as the experience of all women—excluding the experiences of many women worldwide. They have been accused of 'essentialism'—assuming women everywhere share certain essential experiences of reproduction and of subordination under patriarchy. This is a very serious accusation, because it questions whether women share a common experience across race and class, in rich and poor nations, in radically different personal and public relations. Because feminists claim to take women's experience seriously, a 'common experience' is needed to make a general 'ethico-politics' that works for all women—whether it is classified as feminine or feminist.

Many critics argue that 'women's experience' is heterogeneous, and it varies along the significant lines of dominance that mark racial, economic, and political oppression. Women of color, lesbians, and working-class women have raised objections—with the result that the varieties of women's experience have been brought into the light. Collins's (1991) work on black feminist thought has offered both criticism and positive alternatives. Narayan (1995) has written on colonialism and the ethics of rights and care, and the place of 'the other' in feminist theory. The early and influential feminist theorist, Flax (1993), now calls for 'an ethics of multiplicity.' Those working in both feminine and feminist ethics agree to the need, but it is by no means clear how the ethics might offer either a general theory of care and responsibility or a general guide to feminist action. As Sandoval (2000) writes, 'US third-world women' have their own strategies of action which suit their own needs in their own situations.

Feminists are committed to respecting the experiences of all women—and avoiding the imperialism of traditional Western approaches. This commitment may be impossible to fulfill, given the multiplicity of ways of life in all the many places of the earth. Globally, women's situations and their moral and political agency differ radically. It is not clear how a secular, Western feminist ethics would apply in a situation where Islam is important to the lives, identities, and politics of women, for example. In other communities, women have objected to the ethics of women's individual autonomy that is grounded in the ethics. Western feminists have themselves been accused of using Western standards and meanings to define third-world subjects (Ong 1988). 'Postmodern' feminists have argued that feminist ethico-politics presupposes women as the subject of action, and as moral agents, when in fact gender and agency are not givens but are political constructs, enacted in particular times and places (Butler 1990).

Other interesting developments in feminist ethics have been made by the 'ecological feminists.' These feminists offer general criticisms of hierarchies involved in dominance and subordination. They insist that the dualisms of the traditional ethics are incorporated into the social order so as to support dominance in many forms—with domination by men over women and man over nature being central. Ecological feminists share some of their ethics with nonfeminists in the environmental movement, but they insist that understanding patriarchy is fundamental to understanding the hierarchies involved in domination of nature (Warren 1995). One of the earlier books that had an impact was Merchant's *The Death of Nature* (1980), that presents a history of the change in the relation of man and nature from one of respect for a living being (Mother Earth) to a resource to be exploited. Feminist spirituality and 'third world' women have also been important influences in ecological feminism (Ruether 1996).

Within feminist ethics proper, there is criticism of the androcentrism or male centeredness of traditional ethics, but ecological feminists also criticize its anthropocentrism—putting humans at the center of the value universe in a way that permits exploitation of nonhumans. Human interests and human life are the sole moral issues, not only for traditional ethics but for most feminist ethics as well—adding women's interest to men's interest doesn't change that fact. The most interesting of the ecological feminist theories require respect for nonhumans, and they insist on including them within the range of ethics.

Over the years, feminist ethics has come to have a wide influence in other fields. It has made major contributions to biomedical ethics, to business, engineering, and other professional ethics (Leidka 1996, Tong 1997, Whitbeck 1997). The work on care in relationship and the discussions of trust have been particularly important. In this way, feminist ethics has offered practical, new 'guides to action' in some significant arenas of the public world.

See also: Cultural Critique: Anthropological; Cultural Relativism, Anthropology of; Ethics and Values; Feminist Epistemology; Feminist Political Ecology; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory:

Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Theory: Sociological History of

Bibliography

Addams J 1907 *Democracy and Social Ethics*. MacMillan, New York

Baier A C 1995 *Moral Prejudices: Essays on Ethics*. Harvard University Press, Cambridge, MA

de Beauvoir S 1948 *The Ethics of Ambiguity* (trans. Frechtman B). Philosophical Library, New York

Collins P H 1991 *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, New York

Flax J 1993 *Disputed Subjects*. Routledge, New York

Gilligan C 1982 *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA

Gilman C P 1966 *Women and Economics*. Harper & Row, New York

Gilman C P 1979 *Herland*. Pantheon, New York

Held V (ed.) 1995 *Justice and Care*. Westview, Boulder, CO

Hoagland S 1988 *Lesbian Ethics*. Institute of Lesbian Ethics, Palo Alto, CA

Jaggar A 1995 Caring as a feminist practice of moral reason. In: Held V (ed.) *Justice and Care*. Westview, Boulder, CO

Jakobsen J 1998 *Working Alliances and the Politics of Difference: Diversity and Feminist Ethics*. Indiana University Press, Bloomington, IN

Kohlberg L 1981 *The Philosophy of Moral Development*, 1st edn. Harper & Row, San Francisco, CA

Leidka J 1996 Feminism, morality and competitive reality: A role for an ethic of care? *Business Ethics Quarterly* **6**(2): 179–200

Merchant C 1980 *The Death of Nature: Women, Ecology, and the Scientific Revolution*. Harper & Row, San Francisco, CA

Meyers D T 1989 *Self, Society, and Personal Choice*. Columbia University Press, New York

Mohin L (ed.) 1996 *An Intimacy of Equals: Lesbian Feminist Ethics*. Harrington Park Press, New York

Ong A 1988 Colonialism and modernity: Feminist representations of women in non-western societies. *Inscriptions* **3**(4): 79–93

Ruddick S 1989 *Maternal Thinking: Towards a Politics of Peace*. Beacon Press, Boston

Ruether R R 1996 *Women Healing Earth: Third World Women on Ecology, Feminism, and Religion*. Orbis Books, New York

Sandoval C 2000 *Methodology of the Oppressed*. University of Minnesota Press, Minnesota

Tong R 1997 *Feminist Approaches to Bioethics*. Westview, Boulder, CO

Tronto J C 1993 *Moral Boundaries: A Political Argument for an Ethics of Care*. Routledge, New York

Walker M U 1998 *Moral Understandings: A Feminist Study in Ethics*. Routledge, New York

Warren K J (ed.) 1996 *Ecological Feminist Philosophies*. Indiana University Press, Bloomington, IN

Whitbeck C 1997 *Understanding Moral Problems in Engineering*. Cambridge University Press, Cambridge, UK

K. Addelson

Feminist Legal Theory

1. Background

Although it was only as recently as the 1970s that feminist legal theory or feminist jurisprudence were first heard as names for the diversity of gender-focused currents in legal scholarship and argumentation, these strands of contemporary thought have their roots in the 'rights revolution' that took place in the political thought of seventeenth and eighteenth century Europe and the Americas. The foundations of early feminist thought can be traced to explicit efforts to appropriate the emancipatory liberal ideals for women.

The extension of the right to vote was central to the nineteenth century feminist legal agenda. Informed by the social contract theory, early feminists believed that women's political enfranchisement would facilitate their emancipation from historical oppression. The women's rights movement was closely linked to the struggle for the abolition of slavery. In the United States, the campaign for women's vote is associated with the 1848 Seneca Falls convention and the Declaration of Sentiments that emerged from the meeting. It was at Seneca Falls that the former slave, Sojourner Truth, gave her celebrated 'Ain't I a woman?' speech, which questioned the parallel that is often drawn between the legal status of white women and that of slaves in the rhetoric of early nineteenth century feminism.

Despite the significance of voting as the emblematic act of citizenship and of full participation in the *polis*, gender-based discrimination was no less pervasive in other spheres of life. Under the common law doctrine of coverture, a woman lived under the tutelage of her father or other male guardian, until through the sacramental covenant of marriage, her civil legal identity was subsumed by that of her husband under whose cover she derived protection. A woman was typically incapable of making contracts or testamentary dispositions, owning and alienating real property, testifying as a witness in court against her husband, suing or being sued in her own name, or controlling her earnings. She could not even protect her own physical integrity—her husband had the right to chastise her, although only with a switch no thicker than his thumb (see *Gender and the Law*).

The legal fiction of the unity of spouses as well as the rise of separate spheres ideology exempted the private

sphere of the family from state intervention, and rendered women vulnerable to an array of abuses, including rape by their husbands. Beginning in the middle of the nineteenth century, a series of statutes—generically called the Married Women's Property Acts—removed some of the most severe civil disabilities pertaining to women, and acknowledged their personhood for the purposes of the law. Ultimately, struggles for the right to vote, for property rights, for protective labor legislation, for access to education and professional careers, and for related entitlements helped transform the legal landscape. Over the past few decades, feminist strategies have continued to effect change in consciousness and in laws. Legal efforts have focused, *inter alia*, on expanding reproductive freedom, deterring sexual violence, and eliminating sex-based discrimination in employment, education, family, welfare, and related contexts.

2. *Definitions*

Feminist legal theory is a legacy both of the nineteenth century campaigns for women's emancipation and of the renaissance of feminist activism in the United States and elsewhere since the 1960s. Although there is much debate about its definition, feminist legal theory can be understood as the manifestation in the legal academy of a range of efforts to understand the nature of women's subordination—especially the ways in which it is rationalized and naturalized—and to propose remedies for this subordinate condition. Drawing upon cross-cutting academic and activist traditions, feminist inquiry approaches law both as an ensemble of potential tactics for reform and, more generally, as a site of struggle for a more general political transformation. Most strands of contemporary feminist legal theory have conceptualized law as a historically and socially constructed artifact embedded in particular social relations. Feminist legal theory elucidates the ways in which law facilitates patriarchal dominance and control. Explicating how the doctrines, discourses, institutions, and culture of law perpetuate gender hierarchy, feminist legal theory has provided the impetus for reconceptualizing tightly bonded concepts and premises of legal thought.

Distinctions are often made among various strands of contemporary feminist legal theories. Liberal feminists emphasize the similarities between men and women as rights-bearing autonomous human beings (see e.g., Williams 1982, Kay, 1985, Littleton 1987). Radical feminists argue that society is premised on relations of male domination and female subordination, and that the very epistemological preconditions of law and politics are profoundly gendered (MacKinnon 1982). Postmodern legal feminists show how identities are constituted through multiple structures and discourses that in various ways overlap, intersect, and contradict each other (Frug 1992, Williams 1991). Cultural feminists emphasize the differences in experience and perspective between men and women, challenging the reification of instrumental capacities of men and denigration of the affective capacities of women which prefer the connection of the individual self to others (see e.g., Menkel-Meadow 1985, West 1987).

The various strands identified above are not an exhaustive inventory of the competing perspectives in feminist legal thoughts, and each has been criticized for its limitation. For example, the cultural feminist perspective is sharply countered by critics who view it as simply an atavistic, throwback to nineteenth century stereotypes, which reified and naturalized the separate-spheres moralizing that entrenched gender asymmetry (Dubois et al. 1985). By the same token, the radical view is criticized for presupposing a unique female point-of-view that does not allow for social differentiation (Cornell 1990). Whatever their drawbacks, it is important to note that these strands are not at all mutually exclusive; their overlap is particularly evident in their commitment to make the world better for women by reconstituting legal practices that have excluded, devalued, or undermined women's concerns.

3. *Legal Education*

The emergence of the feminist legal theory critique is no doubt linked to the unprecedented entry of women into the legal academy over the past few decades. It was not until 1972 that all bar-accredited law schools in the United States admitted female students. In other countries, the participation of women in legal education has also not been without resistance (see e.g., Dahl 1987, Sachs and Wilson 1978). Feminist theorists have argued that legal education is a microcosm that reflects, reinforces, and reproduces the asymmetric gender relations of the society at large (Obiora 1996). The entire structure of law—its hierarchical organization, its adversarial orientation, and its privileging of formal logic—is considered evidence of its gendered historical and ideological foundations. Studies suggest that the standard law school curriculum marginalizes, distorts, and omits issues of particular concern and interest to women (see e.g., Charlesworth et al. 1991). A recurrent criticism is that the behavioral norms elicited and rewarded by dominant pedagogy are at fundamental variance with socio-cultural norms of female behavior and cause high levels of alienation and role conflict for women in law school (O'Brien and McIntyre 1986).

The idea that gender differences matter is by no means new in the legal terrain. Western philosophical and legal thought is informed and characterized by dichotomous paradigms that may be traced to ancient assumptions of male superiority. Such assumptions found particular expression in systematic sex-role

constraints that allocated public power and authority to men and increasingly relegated the activities of women to the private sphere. One of the theories justifying the exclusion of women from the public sphere posited that there was a discrepancy between womanhood and adulthood; the qualities deemed necessary for adulthood—the capacity for autonomous thinking, rational decision making, and logical action are associated with masculinity, but considered undesirable attributes of the feminine self. The opinion of Justice Bradley in the 1873 United States Supreme Court decision in *Bradwell* vs. *Illinois* (83 U.S. 130), which affirmed the Illinois bar's refusal to admit a woman, epitomized the operation of the ideological generalization about women's intellectual deficiency and providential ordination for domestic concerns. As he put it, '[t]he paramount destiny and nature of woman is to fulfill the noble and benign offices of wife and mother.'

4. *Reform Initiatives*

Feminist legal theory has had diverse targets of inquiry, and has sought to answer many difficult, sometimes intractable, questions about law and society. At the methodological level, it has foregrounded gender as an analytic category. Consonant with techniques of consciousness raising which made visible the personal consequences of gender inequities and the commonalities amongst women's life stories, feminist legal method draws and builds upon analyses of women's experiences. Through the critical scrutiny that it has brought to bear on the form and content of orthodox legal theory and practice, feminist methodology has exposed the ambiguities, contradictions, deficiencies, and politics of law. Interrogating the assumptions, arguments, and conclusions which underlie the patterns of constraint that pervade and define the everyday lives of women, it has illuminated how gender hierarchies have been constructed and legitimated in the legal arena.

Contemporary feminist legal theory has provided a framework for the questioning of numerous gendered social practices, such as prostitution and pornography, and made possible the recognition of sexual harassment and various forms of rape as legal categories (see e.g., MacKinnon 1989, Estrich 1986, Smart 1995, Strossen 1993, Shrage 1994). Feminist theorizing has also shown how the private sphere, as exemplified by the family, which had been treated in law as an autonomous realm deserving of protection from state intervention, incorporates myriad forms of gender subordination. Insisting that the 'personal is political,' feminist scholars have exposed the gendered consequences of domestic violence and of the allocation and valuation of 'invisible' domestic labor, such as child care and housekeeping (see e.g., Schneider 1991, Pateman 1988, Olsen 1983). They have also reinvigorated debates concerning family law issues, ranging from the consequences for women of traditional marriage, divorce, custody, and child support to issues relating to the control of reproduction (see e.g., Fineman 1991). Exposing how women's unpaid domestic labor and socialization patterns determine the extent to which they can enter the marketplace of paid employment and the terms on which they do so, feminist inquiry has illuminated the connection between domestic arrangements and the structure of the workplace, the regulation of motherhood, the availability of welfare, etc. (Finley 1986, Rose 1992, Radin 1987, Roberts 1991, Abrams 1995, Nedelsky 1990).

Just as beliefs about differences between men and women have rationalized gender inequality, other notions of sexual difference have inspired feminist theorizing. Given the centrality of the principle of equality from the Declaration of Independence to present day political discourse, it is not surprising that feminist legal theory began with an attack on discrimination. Criticisms of restrictive governmental categories were framed in terms of equality of treatment and opportunity. Within this framework, feminists have successfully challenged the assumptions and policies underlying many laws, such as those of female dependency and unsuitability for certain jobs, both in courts and in legislatures. Condemning policies or practices whereby women are denied opportunities simply because of their gender, the liberal feminist enterprise has stressed the similarities between men and women and sought to minimize their differences. Especially during the 1960s and 1970s, feminist jurists relied on a formal equality model which required sameness of treatment for similarly-situated persons and disparate treatments for persons who are not alike as a strategy to trigger the judicial invalidation of many sex-based discriminatory classifications and practices (see e.g., Franke 1995).

To survive judicial scrutiny in the United States, sex-based classifications must bear a substantial relationship to an important governmental purpose. Employing this standard of review, the Supreme Court has insisted on gender-neutral military, social security, welfare, and workers compensation benefits. Under the rule of gender neutrality, in many jurisdictions, the law of custody and divorce has been transformed, giving men (in principle) equal rights to the custody of their children and to spousal support. By and large, the objective of challenging unequal opportunities and the ideology that legitimated them has resulted in a certain undeniable progress. Women have increasingly attained equality in formal rights, but they still experience substantial inequity in social, economic, and political status. The persistence of such disparities is in part attributable to the continuing devaluation of women's capabilities and women's work. The ultimate defeat of the proposed Constitutional Equal Rights Amendment (ERA) in the

USA during the 1980s underscored the extent to which 'traditional' views of gender roles have continued to frame debates. Even in places such as Nigeria, where the mechanism for challenging gender-based discrimination is enshrined in the Constitution, efforts to implement the law remain constrained by deeply embedded customs and tradition.

Faith in formal legal equality obscured its political and practical implications (see *Discrimination*). Equality-based arguments were necessarily value-laden to the extent that they affirmed the status quo in general and accepted of the view of law as rational, objective, neutral, and apolitical. Such equality for women came to mean equality with men—in the United States, with white, middle-class men. In this fashion, the social attributes of men set the norm against which women often appeared deficient. Where there was no man to set the norm (as in pregnancy), it was assumed that there was no discrimination and differential treatment was not condemned. Discourse regarding pregnancy discrimination, for example, ignored women's reproductive health needs. The achievement of formal equality in some settings has aggravated the real world inequality of women in others.

The substantive approach to equality modifies the formalist argument by focusing on the outcome of actual differences between men and women. Proponents of substantive equality demand that rules address disparate outcomes resulting from sex differences. A variant of this model attempts to remedy past discrimination through measures such as affirmative action and comparable worth schemes. An alternative model of substantive equality invokes sex-based biological differences as a rationale for sex-specific protections. A more comprehensive strategy advocates special accommodations to counteract the costs of social norms and practices that disadvantage women (see *Affirmative Action, Sociology of*).

The debate over differential treatment strategies caused a rift in the women's movement. For some, emphasizing women's unique characteristics has exacerbated their economic disadvantage and magnified underlying stereotypes of their frailty. With time, equality has been reconceptualized as an anti-subordination principle to counter discrimination, not just because the classical focus on formal equality implied unfair treatment of specific individuals, but also because it perpetuated gender hierarchy. This perspective, also known as dominance theory, focuses attention on the imbalance of power between men and women (MacKinnon 1989). Increasingly, feminist legal theory has been influenced by a view of the individual as being constituted of multiple, intersecting, and contradictory institutional and ideological forces. This emerging view has been directed not only outward, against conventional legal principles and values, but also inward against feminist legal theory itself (see e.g., Minow 1988).

5. *Critique of Identity*

Feminist legal theory has been faulted for its failure to condemn the other dimensions of injustice that also mediate gender identity. The preoccupation with differences between men and women has obscured the differences among women and the way that other factors such as race, class, age, disability and sexuality mediate these differences (see *Critical Race Theory*; *Sexual Orientation and the Law*; *Law and People with Disabilities*). At the same time that women's choices have been limited, the role of some women—white women of the privileged classes—has been idealized and unreflectively generalized to the situations of other less fortunate women. Consistent with the recognition that the perspective from which one speaks affects the meaning of what is said, some feminists have criticized the liberal mainstream tendency to homogenize and universalize a tacitly normative concept of 'woman'—typically a white, heterosexual, bourgeois, able-bodied, female citizen of an advanced industrial democracy (see e.g., Harris 1990, Crenshaw 1989, Kline 1989, Williams 1991).

Influenced by the post-modern acknowledgment of the plurality of situated realities, these feminists have insisted that the analytic focus of the feminist paradigm take account of social context, structure, and history. Arguing that the category 'woman' has no universal explanatory force, they reject any foundationalist view of identity based on gender distinction alone. They challenge the very notion that there could exist a prototypical woman who can be described in ways that reflect and have meaning for the lives of many different women living in very different geographical, economic, political, and social settings (see e.g., Obiora 1996).

6. *Global Dimensions*

Gender relations in many non-Western societies are complicated by an intermingling of custom and religion in the law of personal status that governs gender relations. In nineteenth century British debates around numerous gendered practices in colonized regions, the agenda of rescuing non-white women from the barbarity of the culture into which they had the misfortune of being born played an important role in justifying the imperial project. With the intensification of global cosmopolitanism from the 1980s onward, local gender relations have increasingly been required to justify themselves in dialog with alternative modes of being. Gender equity norms symbolize the advance of the human rights agenda.

The most eloquent expression of the prohibition of gender discrimination is codified in the United Nations Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW). The imperatives that it articulates have been consolidated in subsequent documents and conferences such as

the 1993 Vienna Conference on Human Rights. Notwithstanding the force of the appeal of a global norm of gender parity, there is reason to question how well it has traveled across radically different histories, cultures, and socioeconomic structures. Amid the enthusiasm for spreading human rights ideas across the globe, it is easy to lose sight of the fact that abstract formal rights are seldom a panacea. Gender reform initiatives, laudable as they may be, show clearly that meaningful sociolegal acts of transformation must be sensitive to context and culture.

See also: Fatherhood; Feminist Epistemology; Feminist Movements; Feminist Political Theory and Political Science; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Political Science; Gender and the Law; Rationality and Feminist Thought

Bibliography

Abrams K 1995 Sex wars redux: Agency and coercion in feminist legal theory. *Columbia Law Review* **95**: 304

Babcock, Freedman, Ross, Williams, Copelon, Rhode, Taub 1996 *Sex Discrimination and the Law: History, Practice and Theory*. Little Brown, Boston

Bartlett K T 1994 'Gender law'. *Duke Journal of Gender Law and Policy* **1**: 1

Bartlett K, Harris A 1998 *Gender and Law: Theory, Doctrine, Commentary*. Aspen Law

Barnett H 1998 *Introduction to Feminist Jurisprudence*. Cavendish, London

Beauvoir S de 1949 *The Second Sex*. Picador

Becker, Bowman, Torrey 1994 *Feminist Jurisprudence: Taking Women Seriously*. West

Charlesworth H, Chinkin C, Wright S 1991 Feminist approaches to international law. *American Journal of International Law* **85**: 613

Cornell D 1990 Sexual difference, the feminine and equivalency: a critique of MacKinnon's. *Toward a Feminist Theory of the State, Yale Law Journal* **100**: 2257

Crenshaw K 1989 'Demarginalizing the intersection or Race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. *The University of Chicago Legal Forum*: 139–67

Dahl T S 1987 *Women's Law: An Introduction to Feminist Jurisprudence*

Dubois E C, Dunlap M C, Gilligan C J, MacKinnon C A, Menkel-Meadow C 1985 Feminist discourse, moral values and the law. *Buffalo Law Review* **34**: 11

Estrich S 1986 *Rape*

Fineman M 1991 *The Illusion of Equality: The Rhetoric and Reality of Divorce Reform*

Fineman M, Thomadsen N 1985 *At the Boundaries of Law: Feminism and Legal Theory*. Routledge, London

Finley L 1986 Transcending the equality theory: A way out of the meternity and workplace debate. *Columbia Law Review* **86**: 1118

Fiss O What is Feminism? *Arizona State Law Journal* **26**: 413

Franke K 1995 The central mistake of sex discrimination law: the disaggregation of sex from gender. *U. Pa. L. Rev.* **144**: 1

Friedan B 1963 *The Feminine Mystique*

Frug M J 1992 *Postmodern Legal Feminism*

Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA

Harris A 1990 Race and Essentialism in feminist legal theory. *Stanford Law Review* **42**: 581

Kay H H 1985 Equality and difference: the case of pregnancy. *Berkeley Women's Law Journal* **1**: 1

Kay H, West M 1996 *Sex-Based Discrimination*. West

Kline M 1989 Race, racism and feminist legal theory. *Harvard Women's Law Journal* **12**: 115

Larson J 1993 Can feminists use the law to effect social change in the 1990s? *Northwestern Law Review* 1252

Littleton C 1987 Reconstructing sexual equality. *California Law Review* **75**: 1279

MacKinnon C 1982 Feminism, Marxism, method and the state: an agenda for theory. *Signs: Journal of Women in Culture and Society* **1**(3): 515–44

MacKinnon C 1989 *Toward a Feminist Theory of the State*. Harvard University Press, Cambridge, MA

MacPherson C A B 1962 *The Political Theory of Possessive Individualism: Hobbes to Locke*. Oxford University Press, Oxford, UK

Menkel-Meadow C 1985 Portia in a different voice: speculations on a lawyering process. *Berkeley Women's Law Journal* **1**: 39

Mill J S 1869 *The Subjection of Women*. Cambridge

Minow M 1988 Feminist reason getting it and losing it. *Journal of Legal Education* **38**: 47

Nedelsky J 1990 *Law Boundaries and the Bounded Self*

Obiora L A 1996 Neither here nor there: of the female in American legal education. *Law and Social Inquiry* **21**: 355

Obiora L 1997 Feminism, globalization and culture: after Beijing. *Journal of Global Legal Studies* **4**: 355

O'Brien M, McIntyre S 1986 Patriarchal hegemony and legal education. *Canadian Journal of Women and the Law* **2**: 69

Olsen F 1983 The family and the market: a study of ideology and legal reform. *Harvard Law Review* **6**: 9

Olsen F 1995 *Feminist Legal Theory*. New York University, New York

Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, NJ

Radin M 1987 Market-inalienability. *Harvard* **100**: 1849

Rhode D 1989 *Justice and Gender*. Harvard University Press, Cambridge, MA

Rhode D 1991 The 'no-problem' problem: feminist challenges and cultural changes. *Yale Law Review* **100**: 1731

Roberts D 1991 Punishing drug addicts who have babies: women of color, equality and the right of privacy. *Harvard Law Review* **104**: 1419

Rose C M 1992 Women and property: gaining and losing ground. *Va L. R.* **78**: 421

Rossi A (ed.) 1989 *Essays on Sex Equality*. University of Chicago, Chicago

Sachs A, Wilson J 1978 *Sexism and the Law*

Schneider E The violence of privacy. *Connecticut Law Review* **23**: 973

Shrage L 1994 *Moral Dilemmas of Feminism: Prostitution, Adultery and Abortion*

Smart C 1995 *Law, Crime, and Sexuality: Essays in Feminism*. Sage, London

Smith P 1993 *Feminist Jurisprudence*. Oxford University Press
Strossen N 1993 A feminist critique of 'the' feminist critique of pornography. *Val. L. Rev.* **79**: 1099
United Nations 1980 *Convention on the Elimination of All Forms of Discrimination Against Women*. U.N. Document A/RES/34/46
Weisberg D K 1993 *Feminist Legal Theory*. Temple University Press, Philadelphia, PA
West R 1987 The difference in women's hedonic lives: A phenomenological critique of feminist legal theory. *Wisconsin Women's Law Journal* **3**: 71–145
William J 1991 Dissolving the sameness/difference debate: a critical post-modern path beyond essentialism in feminist and critical race theory. *Duke Law Journal* 296
Williams P 1991 *Alchemy of Race and Rights: Diary of a Law Professor*
Williams W 1982 The Equality Crisis: Some Reflections on Culture, Courts and Feminism. *Women's Rights Law Report* **7**: 175
Wollstonecraft M 1967 *Vindication of the Rights of Women*. Norton

L. Obiora and R. Perry

Feminist Movements

Feminist movements are, in the broadest sense, collective efforts to improve the situation of women. Yet much scholarly debate has focused on the question of what movements qualify as 'feminist.' The problem is one of both time and place. 'Feminism' is a term of relatively recent origin, and it was first used in the countries of Europe and the 'neo-Europes,' places such as the United States, Canada, and Australia where European settler colonies flourished. Furthermore, the meaning of feminism has changed over time. Scholars have proposed different definitions of feminism and different ways of categorizing feminist movements. This article considers these various approaches, explores the origins and development of feminist movements in different times and places, and ends with the coming together of women in international feminist movements.

1. *Defining Feminism*

The term feminism originated as the French word *féminisme* at the end of the nineteenth century and then spread to other languages and parts of the world (Offen 1988). It referred in general to women's emancipation but quickly developed multiple and even conflicting meanings. In the international women's movement in the years before World War I, the term feminism had come to connote, quite specifically, support for legal equality and equal opportunity in the labor market for women and men, particularly opposition to special—what was known as 'protective'—labor legislation for women (Wikander 1992). Since then a raft of definitions and associations have developed, along with a multitude of qualifying adjectives preceding the word feminism. For some scholars, a feminist movement must specifically embrace that identity, while for others, any activism by women can qualify as part of a feminist movement. The former position means that there can be no feminist movements before the late nineteenth century, while the latter can claim such actions as food riots as feminist. Lerner (1993) locates the origins of the creation of feminist consciousness in medieval Europe, defining feminist consciousness as:

> the awareness of women that they belong to a subordinate group; that they have suffered wrongs as a group; that their condition of subordination is not natural, but is societally determined; that they must join with other women to remedy these wrongs; and finally, that they must and can provide an alternate vision of societal organization in which women as well as men will enjoy autonomy and self-determination.

Cott (1989) comes down on the opposite side of the question, arguing that 'not even all activities undertaken by women who claim to have "women's interests" or "women's needs" at heart' are feminist. She proposes, instead, that we distinguish feminist consciousness from 'female consciousness,' that stems from women's shared sense of obligation to preserve and nurture life, and 'communal consciousness,' a sense of solidarity between women and men of any class, ethnic, racial, or other group.

One way scholars have attempted to deal with the complexity of feminism is by distinguishing among different varieties.

2. *Categorizing Feminist Movements*

The most influential scheme for categorizing feminism and feminist movements has been Offen's (1988) distinction between individualist and relational feminism. Arguing that the anomalous Anglo-American brand of individualist feminism that emphasizes human rights and personal autonomy has been taken as the yardstick to determine what was and was not feminist, Offen asserts that relational feminism, which advocates equity for women in their traditional familial roles as wives and mothers, has been more characteristic throughout time and in different locations. Critics have questioned whether relational feminism is really feminist.

Other schemes for categorizing feminist movements in the past include 'social feminism' and 'hard-core feminism' (Cott 1989) or 'social feminism' and 'equity feminism' (Black 1989). These pairs essentially mirror Offen's distinction. At bottom, the basic divide in these categories is between those who emphasize the sameness of women and men (individualist, hard-core, and equity feminism) and those who emphasize dif-

ference (relational and social, as well as 'domestic,' 'maternal,' and 'cultural' feminism). This ideological distinction has its counterpart in the goals of movements, some of which work to win identical treatment for women and men and some of which advocate different treatment as the right route to equality. In fact, there is no neat fit between an ideological tendency to sameness and the goal of identical treatment, since some feminists have believed that women and men should have identical rights precisely so that women can bring their different (and better) qualities into public life.

For more contemporary movements, scholars have distinguished among liberal, socialist, and radical feminism. Liberal feminism seeks equality for women within the existing social structure, holding that women lack power because as a group they are denied equal opportunity to compete and succeed in the male-dominated economic and political arenas. Thus, liberal feminist movements advocate equal access and equal opportunity. Socialist feminism identifies both class and gender as equally important in creating inequality and calls for a transformation of both capitalism and patriarchy. Radical feminism, in contrast, sees women as a 'sex class and views gender as the primary foundation for the unequal distribution of rewards and privileges. Radical feminist movements, which arose in the industrialized West in the late 1960s and 1970s, like socialist feminist movements seek more than access to existing privileges. Radical feminists call for a fundamental transformation of society based on the superiority of the 'female values' of caring, nurturing, egalitarianism, pacifism, and cooperation.

The categories of feminism have proliferated as research on different kinds of feminist activism has expanded. Scholars have identified Nazi feminists, Islamic feminists, lesbian feminists, and imperial feminists, to name just a few. Increasingly the plural, feminisms,' is used to capture the diversity of these varieties of feminism. From a social movement perspective, Rupp and Taylor (1999) have proposed to shift focus away from the ideology of feminist movements to the collective identities that are created and sustained within movements. Such identities exist at three levels: organizational, movement, and solidary. Women develop a sense of themselves as members of discrete organizations, as part of a movement, and, at the broadest level, as women. By exploring the construction of these overlapping identities within specific contexts, it is possible to see how women with very different understandings and strategies and goals might all claim the label 'feminist.'

3. *The Origins of Feminist Movements*

Social movement scholars point out that everywhere and at every time, women have had sufficient grievances to stimulate the rise of feminist movements. Yet collective activity on the part of women aimed at improving their own status in society has blossomed primarily during times of generalized social upheaval. In addition, scholars have identified structural conditions that underlie the emergence of feminist protest, at least in the West: changes in patterns of female labor force participation, increased access to education, and shifts in fertility rates and reproductive roles. Chafetz and Dworkin (1986) argue that industrialization and urbanization play a critical role in the emergence of women's movements. In industrializing and urbanizing societies, a large middle class develops, women gain access to education, married women enter the labor force in larger numbers, and the resulting role expansion and conflict results in the development of gender-based and ultimately feminist consciousness. The levels of industrialization and urbanization, the size of the middle class, the extent of women's access to education, and the numbers of married women in the labor force, in combination with the attitude of governments to women's demands, will determine the size of women's movements and whether they qualify as feminist. Despite Chafetz's and Dworkin's attention to movements throughout the globe, their model privileges the western pattern of movement emergence. Feminist movements, scholars have argued, have emerged in other parts of the world in conjunction with nationalist movements, not industrialization and urbanization (Jayawardena 1986).

Although the conditions that lead to the emergence of feminist movements may vary in different parts of the world, there do seem to be broad commonalities in the ways such movements originate and develop. A first stage often begins with the writings of elite women (and men) and demands for greater access to education. In Egypt, e.g., educated, elite women in urban harems began to call for educational opportunity in the context of Islamic modernism at the end of the nineteenth century (Badran 1995). In Argentina, Uruguay, Chile, Brazil, Mexico, and Cuba, women schoolteachers took the lead in the late nineteenth century in calling for expanded education for women (Miller 1991). Chinese women, many educated in Japan, began in the early twentieth century to relate China's weakness to the lack of education for women (Beahan 1975). Typically, women who have gained some educational advantage begin to spread their ideas, often through their own publications. Throughout the Middle East, Asia, and Latin America, women's journals played a central role in introducing feminist ideas.

A next stage was often the formation of organizations devoted to various kinds of social reform. In the United States in the mid-nineteenth century, these included radical projects such as abolitionist groups as well as more societally acceptable activities such as literary societies and church groups. In many cases, demands that originally sprang from women's interest in other causes—that women be allowed to speak in

public or participate in the political process in order to bring about some change in society—eventually came to be voiced as a basic human right. During the national revolution of 1919–22, Egyptian feminists who had earlier formed philanthropic groups worked actively to free Egypt from British domination but experienced disappointment when they found themselves deprived of the vote in the new Egyptian state. This was when the word feminist (in French) first came into use, claimed by the Egyptian Feminist Union, founded in 1923 (Badran 1995). The move from participation in nationalist or anti-imperialist movements to demands for basic political rights and for social freedoms was a common one throughout the colonized and dependent countries.

As feminist movements developed in different places, they often split along ideological, political, religious, or other lines. In countries such as Germany, e.g., the bourgeois women's movement and socialist women's movement had little to do with each other. In England, and to a lesser extent in the United States, the major divide was between moderate and militant groups. The French women's movement splintered along political lines into Republican, Catholic, and socialist organizations. Because feminism as an ideology and movement emerged earlier in the United States and Europe, feminist movements in the rest of the world often give rise to charges of imperialist influence. Despite feminism's origins as an indigenous movement in Egypt, the fact that the upper and middle-class urban women who claimed the label tended to wear western dress and speak French associated feminism with foreign and imperialist forces (Badran 1995). Korean New Women in the 1920s made use of Japanese, British, and Swedish feminist discourses for their own purposes in developing a critique of traditional Korean marriage and sexuality, giving rise to the notion that they were simply dupes of the West (Kwon 1998).

4. *The Two-wave Model of Feminist Movements*

Most scholarship on the development of women's movements conceptualizes two waves of activism, the first in the mid-nineteenth century to the early twentieth century and the second since the 1960s. Evans (1977) explores a variety of feminist movements in Europe, the United States, and Australasia during the first wave, charting the shifting aims of the movements from economic and educational to moral and then political. Evans concludes that the feminist movements of the first wave declined along with the liberalism with which they were so closely associated by the early twentieth century. But there is more continuity in the existence of women's movements than this model suggests. In fact, movements grew outside the Euroamerican arena precisely when scholars think of the decline of feminism in the aftermath of World War I. It may still be useful to think of two peaks or waves of feminist activism, as long as we remember that not all movements followed the same trajectory and that feminist movements did not die out between the two waves (Rupp and Taylor 1987). Nevertheless, it is clear that feminist movements burst forth with renewed vigor in the 1960s and 1970s in association with civil rights, student and left-wing movements, and movements of national liberation. Often divided on questions of ideology, strategy, goals, and membership, these late twentieth-century movements mobilized around a broad range of issues, from reproductive rights to sexual and economic exploitation to violence against women. Feminist movements both conceptualized women as sharing basic experiences of oppression, asserting that 'sisterhood is global,' and increasingly came to emphasize differences among women on the basis of race, class, ethnicity, nation, and so on. This was true both within national feminist movements and even more so on the international stage.

In the United States, scholars identify two wings of the second-wave women's movement: the women's rights branch, consisting of older professional women organized in bureaucratic national groups working for change within the existing system, and the women's liberation branch, made up of younger women associated with the civil rights and New Left movements who organized small collective groups on the local level and worked for a radical transformation of society (Freeman 1975; Ferree and Hess 1994). The women's rights branch, dominated by the National Organization for Women, founded in 1966, originally worked for legal and public policy reform, while the women's liberation groups pioneered consciousnesss-raising as a strategy for change and engaged in dramatic 'zap' actions to protest such things as the institution of marriage and beauty pageants and the stock market. Yet over time the distinctions between the two wings dissipated.

The political contexts of different countries in the late twentieth century have been critical in shaping feminist movements and in facilitating or constraining success in moving toward feminist goals. In Western Europe and the United States, feminist movements have had the most success in policymaking where the Left is either very strong, as in Sweden, or very weak, as in the United States (Katzenstein and Mueller 1987). Likewise, in India, the women's movement in Calcutta, which is dominated by the communist party and has a strong left culture, functions more as a political party, while the movement in Bombay, where there is more political competition, takes on more autonomous forms of organizing, including the formation of groups receptive to Western ideas (Ray 1999).

5. *International Feminist Movements*

Since the 1970s, regional and international organizing among women has blossomed, although the origins of

international women's organizations date back to the closing decades of the nineteenth century. In fact, women from the long-established international groups played a vital role in bringing the issue of women's equality to the founding conferences of the United Nations in the aftermath of World War II (Rupp 1997). It was the United Nations-sponsored International Women's Year in 1975, the Decade for Women conferences in 1975, 1980, and 1985, and the Beijing conference of 1995, with their attendant meetings of non-governmental organizations, that nurtured global feminist dialogue (Miles 1996). Although the official conferences and the affiliated forums have proven extremely contentious, struggling over questions of politics, religion, sexuality, and social custom, they have grappled with such basic issues as women's worldwide economic disadvantage, women's lack of reproductive freedom, and different forms of violence against women. Despite variations in feminist movements around the world, international conferences have diffused feminist ideology and practice throughout the globe.

See also: Civil Rights Movement, The; Feminist Theory; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Radical Lesbian; Gay/Lesbian Movements; Right-wing Movements in the United States: Women and Gender; Social Movements, History of: General

Bibliography

Badran M 1995 *Feminists, Islam, and Nation: Gender and the Making of Modern Egypt*. Princeton University Press, Princeton, NJ

Beahan C 1975 Feminism and nationalism in the Chinese women's press, 1902–1911. *Modern China* **1**: 379–416

Black N 1989 *Social Feminism*. Cornell University Press, Ithaca, NY

Chafetz J S, Dworkin A G 1986 *Female Revolt*. Rowman & Allenheld, Totowa, NJ

Cott N F 1989 What's in a name? The limits of 'social feminism': Or, expanding the vocabulary of women's history. *Journal of American History* **76**: 809–29

Evans R 1977 *The Feminists*. Croom Helm, London

Ferree M M, Hess B B 1994 *Controversy and Coalition: The New Feminist Movement across Three Decades of Change*, rev. edn. Twayne, New York

Freeman J 1975 *The Politics of Women's Liberation*. McKay, New York

Jayawardena K 1986 *Feminism and Nationalism in the Third World*. Zed Books, London

Kwon I 1998 The new women's movement in 1920s Korea: Rethinking the relationship between imperialism and women. *Gender & History* **10**: 381–405

Lerner G 1993 *The Creation of Feminist Consciousness: From the Middle Ages to Eighteen-seventy*. Oxford University Press, New York

Miles A R 1996 *Integrative Feminisms: Building Global Visions, 1960s–1990s*. Routledge, New York

Miller F 1991 *Latin American Women and the Search for Social Justice*. University Press of New England, Hanover, NH

Offen K 1988 Defining feminism: A comparative historical approach. *Signs* **14**: 119–57

Ray R 1999 *Fields of Protest: Women's Movements in India*. University of Minnesota Press, Minneapolis

Rupp L J 1997 *Worlds of Women: The Making of an International Women's Movement*. Princeton University Press, Princeton

Rupp L J, Taylor V 1987 *Survival in the Doldrums: The American Women's Rights Movement, 1945 to the 1960s*. Oxford University Press, New York

Wikander U 1992 International women's congresses, 1878–1914: The controversy over equality and special labour legislation. In: Eduards M L et al (eds.) *Rethinking Change: Current Swedish Feminist Research*. Humanistisk-samhällsvetenskapliga forskningsradet, Uppsala, Sweden, pp. 11–36

L. J. Rupp

Feminist Political Ecology

Feminist political ecology offers explanations and analyses of human–environment relations by uniting feminism with political ecology, a subfield prominent within geography and anthropology. Political ecology emphasizes how power relations and politics shape the dynamics of economic development, environmental transformation, and social change across geographic scales of analysis from the local to the global. This approach lies at the interdisciplinary intersection of critical development studies, feminism, and environmental studies. Feminist political ecology analyzes identity, difference, and meanings in their relation to sites of environmental change, degradation, and struggle (Rocheleau et al. 1996). This subfield emphasizes policy relevant natural resource management issues and environmental and social justice concerns in both urban and rural places around the world. Feminist political ecology explains how gender identities and social relations shape, and are shaped by, power relations and social inequality, and how gendered inequalities are linked to questions of liberation struggles, environmental change, and degradation. This article examines the changing dimensions of these cultural, political and ecological intersections through an overview of the emergence of feminist political ecology since the 1980s.

1. Feminism and Environmentalism

Feminism's link with environmental issues and concerns embraces both social and ecological theory and political activism that is rooted in the rise of environmental and feminist activism and intellectual activity of the late 1960s and early 1970s. Ecofeminism, environmental feminism(s), and feminist political ecology all refer to the linkages of people-gender-culture-

environment-nature with varying emphases and frameworks (Merchant 1980, Plumwood 1993, Rocheleau et al. 1996).

Since the 1980s, activists' concerns focused upon nuclear warfare and militarism, toxic waste pollution, and deforestation among others. Ecofeminist work, which spanned academic, popular, and literary texts, emphasized that women were both biologically and socially linked to nature. Women's status and subordination was thus linked to environmental degradation and change so that the oppression and degradation of nature and the environment paralleled the oppression and degradation of women (Merchant 1980). However, when viewed from an international perspective, gendered human perceptions of nature as well as gendered material and sociocultural links to the environment varied over place and time and among cultures. In numerous societies around the world, social scientists have documented and explained how changing cosmologies and geographic histories construct and reproduce key binaries: man/woman; culture/nature; self/other. Structuralism and semiotics influenced much of this initial work. Women are commonly, but not necessarily, associated with nature, and landscape, the fertility of the earth, the waxing and waning of the moon, with passivity, and the hearth (Merchant 1980, Kolodny 1984). These cultural metaphors shaped individual lives as well as public discourse through images of Mother Nature/Mother Earth as contrasted with Spaceship Earth, for example. Early ecofeminist work provided insight as to the ways in which nature and the environment are gendered, and then employed this insight to analyze environmental transformation or celebrate the connection in its spiritual dimension. Ecofeminists defined masculinity and femininity as socially constructed categories which are detrimental to people and the environment and are linked to larger social processes such as colonialism and development. Subsequent critiques pointed to the monolithic and essentialist foundations of earlier ecofeminist work. These critiques argued against a monolithic view of woman that was unmediated by other social categories such as race, ethnicity, gender, age, sexual orientation, and citizenship and pointed to the necessity of site specific empirical case studies (Agarwal 1992).

2. *Critical Ecofeminism*

Critical ecological feminism involves a conscious decision to align human and nature in a move beyond the binary dualism embedded epistemology and ideology in the West. In feminist postructural critiques, masculinity and femininity as modes of identity are examined in relation to our store of knowledge about the human link to nature and the ways in which nature and environment are gendered and socially constructed. Feminist political ecology rejects the dualisms of man/nature and emphasizes multiplicity, diversity, and the complex interconnections among ecologies, societies, and politics in questions of environmental change of both short and long-term duration. This insight is linked to the critical examination of development policies and practices and increasing social inequality. A key question is how identity and difference shape resource control, access, and distribution in multiple historical and ecological contexts initially in Africa, Asia, and Latin America. Research examined how the marginalization and exploitation of local communities and ecologies is linked to larger social and economic networks, institutions, and processes.

3. *Gender, Environment, and Development*

Critiques of mainstream economic development and modernization theory at the international scale revealed how women were invisible in modernization initiatives and within development theory and policy more generally. For example, the majority of women in sub-Saharan Africa are food producers who are responsible for provisioning their households. However, throughout Africa and elsewhere, women were, and still are, unable to gain access to the social, financial, technical, and natural resources that would enable them to provide adequately for their households. As women were incorporated slowly and unevenly into the modernist project through economic development programs aimed at their inclusion, subsequent research revealed how gender bias in development and nesting patriarchal ideologies could exacerbate as well as alleviate their particular vulnerabilities. In addition, this research demonstrated how technology transfer deskilled or did away with working class women's work, and how modernization and development policy and practices in Africa, Asia, and Latin America transformed rural environments and indigenous social and cultural networks, often reproducing or reinforcing gender stereotypes and inequalities. These critiques of mainstream development theory and practice drew from structuralism and Marxian political economy. Studies of the nexus of capitalism, development, populations, and ecologies stressed how oppression and injustice are linked to environmental transformation and change, the industrial geographies of resource extraction and capitalism's creations of surplus (Di Leonardo 1991, Seager 1993).

Examining diverse dimensions of women's work and access to natural resources was critical in revealing the ways in which women were largely disadvantaged by mainstream modernization efforts and policies. Intersections of imperial and indigenous forms of patriarchal power and control could constrict access to land, credit, technology, and education and shape human reproduction. Environmental justice move-

ments and feminist research linked poverty and environmental degradation such as tropical deforestation, irrigation, and pesticide poisoning to gendered relations within the colonialism and to development-induced scarcity. Household vulnerability, changing individual, household, regional and national entitlements, and a lack of democracy are linked to the persistence of poverty and to environmental degradation. Gendered identities and relations crosscut racism, class, age, marital status, ethnicity, and nationality.

Alternatives to mainstream development explore indigenous agroecologies and ecological knowledge, the organization and endurance of social movements linked to environmental concerns, and examine popular struggles and resistance by women and by marginalized and indigenous peoples. The ecofeminist project is one of inclusion and interconnectedness. Feminist political ecology refuses to accept binary dualisms linking gender relations and activities to nature and valuing women and/or nature as inferior or unequal (Plumwood 1993). Thus, ecofeminism and feminist political ecology concern struggles for liberation.

4. *Human Ecology Politicized*

Political ecology emerged from the subfields of human/cultural ecology prominent in geography, anthropology, and sociology. The next section examines the emergence of political ecology and feminist political ecology from these subfields.

Cultural ecology examines society–environment relations and uses the concepts of adaptation and equilibrium and organic metaphors to explain these relations over time and place. Cultural ecologists focused upon case studies involving extractive economies and peasant societies in Africa, Asia, and Latin America. Critics of cultural ecology pointed out its tendencies to explanations, which were functionalist, teleological and devoid of considerations of politics and power relations. Subsequent critiques questioned whether organic metaphors were appropriate for defining human societies and also indicated that research in ecology had moved beyond conceptualizing environments in terms of equilibrium states and were defining them in terms of dynamism, impermanence, and constant change. Women's access and use of natural resources within a cultural and ecological frame progressed in this period as research focused upon issues of women as resource managers and farmers.

5. *Gendering Cultural/Social/Political Ecologies*

Since the 1980s, political ecology has examined the political and social dimensions of environmental transformation and change primarily in rural areas of Africa, Asia, and Latin America. It emerged as a critique of cultural ecology's inattention to power relations and considerations of the politics of resource distribution, access, and transformation. Political ecology combines a broadly defined political economy of resource development and change centering upon the role of social relations and processes to environmental change and degradation, resource distribution, access, and control and the social constructions of nature. In its critique of adaptation and equilibrium, political ecology emphasized agrarian and environmental transformation and change throughout the colonial period up to the present day by examining the ways in which capitalism transformed societies, environments, and the relation between them.

Poverty and environmental degradation are proximate outcomes of the workings of capitalism and world market integration. Peet and Watts (1996) define four contemporary directions in political ecology. First, there are attempts to connect explicitly the dynamics of capitalist growth and environmental outcomes at varying levels and scales. Second, is the integration of power relations and politics into examinations of social and environmental change at varying levels of scale from the global to the micro-politics of the household. Third is the focus examining the roles of ideology and social institutions and organizations. This involves probing the roles of knowledge construction and legitimization as well as state–society relations in terms of the politics of environmental development and change. Fourth, concerns the myriad perceptions of environmental and resource problems and the ways in which these problems are struggled over.

Notably, political ecology is also prominent in its critique of neo-Malthusian explanations of environmental degradation and change. These explanations avoid considerations of social structures and processes and gender identities and power relations within specific historical and geographic contexts as ways to understand the linkage between environment and population growth.

Bryant and Bailey (1997) outline four major research themes in political ecology. The first centers on a particular environmental problem or set of problems such as soil erosion, deforestation, or water. The second studies concepts that are linked to political ecology such as critiques of sustainable development. A third examines political and ecological problems by means of a specific regional case study. The final approach which is explicitly linked to feminism emphasizes individual and group social characteristics such as class, age, ethnicity, gender, and race.

Over roughly the same period, feminist studies in international political economy had taken up considerations of gender and environments, gendered resource access and control, and activism and social movements as they involve the environmental change and degradation ranging across geographic scale.

6. Research Directions in Feminist Political Ecology

Contemporary research directions in feminist political ecology concern how knowledge about society, environments, spaces, and places are gendered. Studies take up questions of local, gendered knowledge about ecologies and agrarian, rural, and urban landscapes. Gendered knowledges refer to particular bodies of knowledge such as that of plant species, agricultural techniques, and food preparation and processing which are frequently the provinces of women or indigenous groups. A second strand of research examines changing environmental rights and responsibilities as they pertain to resource access, distribution, and control. Gendered rights and responsibilities may involve differential access to natural resources such as education, technology, land and trees, and the impacts upon female headed households, the poor, or ethnic minorities. Third, feminist political ecology examines the social struggles that fuse social and environmental concerns and contribute to progressive social change.

Environmental degradation is linked to survival and the quality of life. Survival pivots upon questions of entitlement within rural and urban spaces and is linked in by a diversity of production systems and subject to collective struggles. These microlevel questions and struggles are linked necessarily to broader processes of globalization and ecological dynamics across geographic scales.

See also: Cultural Geography; Ecology, Cultural; Ecology, Deep; Ecology, Political; Feminist Epistemology; Feminist Theory: Ecofeminist and Cultural Feminist; Gender and Environment; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender Ideology: Cross-cultural Aspects; Human–Environment Relationships

Bibliography

Agarwal B 1992 The gender and environment debate: lessons from India. *Feminist Studies* **128**(1): 119–55

Bryant R L, Bailey S 1997 *Third World Political Ecology*. Routledge, London

Di Leonardo M 1991 *Gender at the Crossroads of Knowledge*. University of California Press, Berkeley, CA

Kolodny A 1984 *The Land Before Her*. University of North Carolina Press, Chapel Hill, NC

Merchant C 1980 *The Death of Nature*. Harper & Row, San Francisco

Peet R, Watts M 1996 *Liberation Ecologies*. Routledge, London

Plumwood V 1993 *Feminism and the Mastery of Nature*. Routledge, London

Rocheleau D, Thomas-Slayter B, Wangari E 1996 *Feminist Political Ecology*. Routledge, London

Seager J 1993 *Earth Follies: Coming to Feminist Terms With the Global Environmental Crisis*. Routledge, New York

L. Jarosz

Feminist Political Theory and Political Science

In popular discourse the term feminism is normally used to mean support for women's rights, while a feminist is someone whose central concern is with the position of women and their struggle for emancipation (Delmar 1986). The definitions are contested by feminist scholars and subject to change. In the 1970s feminism was frequently defined as opposition to the dominance or subjugation of women. By the end of the 1990s it was impossible to define feminism without reference to the concept of gender and it was unusual to refer to feminism in the singular. Feminisms are therefore standpoints. They are also political practices (most often social movements), and sets of ideas and theories. To complete the definition it is necessary to add the concept of gender. Gender is a category that was developed to explore what counts as 'woman' and what counts as 'man' (Squires 1999, p. 54). Thus, one useful definition of feminisms declares them to be oppositions to gender hierarchy, capturing the different situations and interests of different women in the pluralization of 'feminism,' while the social construction of masculinity and femininity is signaled by the term 'gender.'

1. Gender Studies and Feminist Scholarship

Challenging the approaches and preoccupations of mainstream scholarship has proved to be a continuing focus of feminist politics. Originally mainly concerned with 'women's studies,' the attention of feminist scholarship gradually shifted to the study of gender and especially to gender relations and the nature of masculinity and femininity. Feminist writings include critiques of most areas of contemporary scholarship calling into question their sensitivity to issues of gender. The shift had important implications for the social sciences including the study of politics. In the social sciences an important feminist concern is to establish the category of gender as a fundamental social division. The feminist critiques of political science are part of this tendency, addressing the standard themes of research on politics, questioning narrow definitions of the 'political,' and also drawing upon research and thinking about gender in other disciplines.

Feminist political science is therefore interdisciplinary and radical, offering a broad definition of the

field. Feminist political studies take three main forms: (a) feminist political theory which is predicated on a critique of liberalism and challenges the conventions of political theory, (b) feminist political science which is a critique of political institutions and the ways they have been understood, and (c) feminist accounts of gender and politics which combine explorations of gender as a category, as a hierarchy, and as performance in studies of how gender relations are implicated in power. Each area is work in progress, complicated and contested. Similar patterns of development have characterized each of the three kinds of discussion.

2. *Feminist Political Theory*

The women's liberation movement of the 1960s and 1970s gave a new impetus to feminist politics. This was at first new social movement politics, suspicious of formal hierarchical organizations which were regarded not only as undemocratic but also as sexist. However, feminism was, and is, profoundly political. The early slogan 'the personal is political,' questioned the nature of power and challenged the conventional restriction of politics to the 'public sphere.' Initially, feminist political theory was a matter of unmasking sexist assumptions and the neglect of women by Western Political Theory. Carol Pateman argued that the subjugation of women was a central underpinning of modern political thought and that gender difference made a public sphere possible at the expense of women. She contended that the distinction between state and civil society is only visible if the family is ignored. If civil society is labeled private then family is forgotten in political discussion (Pateman 1988). Hence the addition of women to the concerns of political theory brought the central assumptions of liberalism into question, notably the separation between public and private life.

Recognizing that writing gender differences into political theory required a major reconsideration of political ideas, feminist theorists addressed issues of equality, justice and representation, initiating debates that grew increasingly sophisticated. Originally theoretical arguments were built around binary oppositions or dualisms such as equality/difference, nature/nurture, justice/care, inclusion/exclusion, all thought to parallel male/female dichotomies. Arguments in the equality/difference debates considered whether men and women should be treated the same or whether differences should be taken into account in order to treat both fairly. The justice/care debates addressed a perceived difference between an ethic of justice (in the liberal distributive sense) and an ethic of care (in a maternalist sense), argued to be male and female, respectively.

Representation arguments considered whether women need women to represent them, reasoning that the identity of the speaker matters to the interpretation of what is said (Phillips 1995, Young 1990). Discussions of justice, equality, and representation tended to converge into arguments about citizenship in which feminists seek to redefine politics in terms that integrate public and private concerns, breaking through the public/private divide (Dietz 1985, Mouffe 1995, Lister 1997). Since the 1980s a tendency to contest the oppositions has become more prevalent as theorists seek to transcend or displace the hierarchies, often by redefining issues and concepts (Squires 1999, pp. 226–33).

3. *Feminist Political Science*

Although there were very few links with feminist political theory, feminist political science has followed a similar trajectory, developing from the inclusion of information about women to the subject matter of politics to an intellectual concern to establish a gendered discipline. Feminist political science first sought to eliminate the 'bad science' of exaggerating and misconstruing differences between women and men, assuming a male political universe and using data without their original accompanying qualifications. They were troubled by the distortions present in the mainstream of behavioral political science arising, they believed, from an epistemology that separated facts and value and privileged so-called value-free factual accounts. The effect of concern only with putative facts was to allow behavioral political science to factor women out of consideration as political subjects. An overemphasis on measurable political behavior prevented recognition of '…not only the… exclusion of women from what is traditionally political, but also the inclusion of politics in what women have traditionally done' (Nelson 1989, quoted in Carroll and Zerilli 1993). Families, communities, and voluntary groups all went largely unresearched as constitutive of political life. Inevitably, feminist political scientists sought to expand definitions of politics, lending strength to other challenges to overconcern with electoral behavior and elite participation, at the same time as examining differences in men's and women's political behavior. Discussion of the significance of the public/private split and the confinement of women to the private realm were an important challenge to narrow constructions of political power. So too were studies of the crucial part played by the state in regulating sexuality (see Carroll and Zerilli 1993, Lovenduski 1998).

Once it was necessary (and possible) to study the effects of women on previously male-dominated political institutions and organizations, the need for an adequately theorized notion of gender was accepted. As increasing numbers of women appeared in legislatures, councils, and other representative bodies as well as in executive organizations and public

administrations of various kinds, a considerable debate developed about the causes, stability, and results of gender differences in political activity. An especially powerful line of argument for both feminist activists and their opponents was the contention that women should not seek equal presence to men in powerful positions because they would be 'incorporated' and changed by it, they would become unable to represent women because they had been turned by male created institutions into political men. The counterclaim was that once a significant number and diversity of women occupy decision making positions in previously 'masculine' institutions they would be transformed, their gendering altered.

This argument, so important in feminist debate, has echoes in the conventions of the study of political processes (Randall 1991). Once arrangements, practices, and rules are institutionalized, actors learn them, invest in them, and make decisions accordingly. Particular institutions become associated with particular behavior and outputs that may be classified as masculine or feminine (Duerst-Lahti and Kelly 1995). The practices that insulate the positions of typical members and are notoriously difficult to change are gendered.

4. *Feminism and Gender Studies*

Feminist intellectual concern about women and power generated research in the arts, humanities, and social sciences. In the fields of history, sociology, social psychology, literature, and cultural studies, a substantial body of work about different aspects of gender and power was produced. In addition a significant field of 'new' men's studies examined masculinity as a category both of privilege and oppression (Carver 1996). Analysis of masculinities has also received considerable attention from feminist theorists who also claimed that differences among women were at least as important as differences between women and men, that race, class, age, ethnicity, sexuality, and physical ability were, like sex, constitutive of political identity and therefore of political behavior and power.

5. *Feminism, Gender Studies, and Political Science*

When feminist political scientists contested narrow definitions of politics, it was because they wanted to expand their research to aspects of 'private' life and to examine the role of government and the state in constructing and enforcing certain kinds of relationships between women and men. As it became more widely understood that both women and men have gender, debates across disciplinary boundaries became a common feature of feminist research on politics. Once perceived, the effects of gender cannot be left to one side or held constant for the sake of other experiments. As it has come to be defined, gender is embedded in individuals, relationships, institutions, and organizations. The gender debates are part of a continuing process of definition refinement and theory production that enhances understanding of political institutions and processes. A 'gendered' political science looks 'not just at particular theories and assumptions as they have been applied to women, but as they are constructed and apply in any case' (Sapiro 1991). Institutional research traditions in political science require the close description of contexts, norms, processes, and cultures to which a gendered vocabulary may be applied. It follows that the gendered interests present in political arrangements should be acknowledged and analyzed.

In summary, feminist scholars have shown that it is difficult to think of an aspect of politics that is unaffected by gender. However, gender is not a neat and tidy concept to be slotted into unchanging research strategies. Using the concept of gender involves rejection of the definitions of politics implied by the conventions of the public/private split, reexamination of the productions of conventional political science for gender dimensions and, on that foundation, the construction of a political science that is gender informed.

See also: Feminist Movements; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Gender and Feminist Studies in Sociology; Rules in the Legal Process; Women's Suffrage

Bibliography

Carroll S J, Zerilli M G 1993 Feminist challenges to political science. In: Finifter A W (ed.) *Political Science the State of the Discipline, II*. American Political Science Association, Washington, DC

Carver T 1996 *Gender is not a Synonym for Women*. L. Reiner, Boulder, CO

Dahlerup D 1988 From a small to a large minority: Women in Scandinavian politics. *Scandinavian Political Studies* **11**: 275–98

Delmar R 1986 What is feminism? In: Mitchell J, Oakley A (eds.) *What is Feminism?* Blackwell, Oxford, UK, pp. 8–33

Dietz M 1985 Citizenship with a feminist face: The problem of maternal thinking. *Political Theory* **13**(1)

Duerst-Lahti G, Kelly R M (eds.) 1995 *Gender Power, Leadership and Governance*. University of Michigan Press, Ann Arbor, MI

Lister R 1997 *Citizenship: Feminist Perspectives*. Macmillan, London

Lovenduski J 1998 Gendering research in political science. *Annual Review of Political Science* **1**: 333–56

Mouffe C 1995 Democracy, power and the political. In: Benhabib S (ed.) *Democracy and Difference: Contesting the Boundaries of the Political*. Princeton University Press, Princeton, NJ, pp. 245–57
Pateman C 1988 *The Sexual Contract*. Polity Press, Cambridge, UK
Phillips A 1995 *The Politics of Presence*. Clarendon Press, Oxford, UK
Randall V 1991 Feminism and political analysis. *Political Studies* **39**: 513–32
Sapiro V 1991 Gender politics, gendered politics: The state of the field. In: Crotty W (ed.) *Political Science: Looking to the Future*. Northwestern University Press, Evanston, IL, pp. 165–87
Squires J 1999 *Gender in Political Theory*. Polity Press, Cambridge, UK
Young I M 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

J. Lovenduski

Feminist Psychotherapy

1. *What is Feminist Therapy?*

Feminism is defined in Webster's New Word Dictionary (1978) as (a) the principle that women should have political, economic, and social rights equal to those of men, and (b) the movement to win such rights for women. While this is a relatively benign definition, the range of reactions and beliefs attached to the word feminism is vast and often emotionally laden.

These diverse views of feminism have resulted in conflicting definitions in the field of psychotherapy and have led to questions of whether one should more aptly refer to *feminisms* and psychotherapies rather than a unified theory of feminist therapy. In fact in many ways 'feminist therapy' is a moniker for an approach or value system that can be applied to all mental distress. Neither monolithic nor static feminist perspectives can be applied to all therapeutic orientations (e.g., *Interpersonal Psychotherapy*; *Feminist Theory: Psychoanalytic*).

1.1 *Core Principles*

Although they may vary in their application, at the core of a feminist approach are the following assumptions (Worell and Remer 1992).

(a) Women's problems cannot be solved in isolation from gender politics which often result in women's lower social status and oppression in most societies.

(b) Equal status and empowerment are vital not only for women but for all oppressed groups.

(c) Values enter all human enterprises: neither science nor clinical cases can be value free.

(d) Women's experience and knowledge should be appreciated and studied.

(e) Few individual women can achieve parity alone, individual and collective action is necessary to achieve the social and political change that underpins many stated 'mental' problems.

Brown (1994) captures it well when she says 'what makes practice feminist is not who the clients are or the specific techniques used, but how the therapist thinks about what she does' (p. 2).

2. *The Birth of Feminist Therapy*

The women's movement in the late 1960s prompted a reconsideration of roles for women and men in western society. Consciousness raising groups provided a forum for women to make the link between their personal experience and the political context in which these experiences were constructed. In spite of their turning away from the accepted expertise of the 'establishment' one of the outcomes of consciousness rising groups was the development of feminist therapists. Through setting up 'women's' groups and 'feminist' supervision, services run by women for women began to develop.

As emergent ideologies challenged traditional views, the field of psychology was also pushed in new directions. This produced revised theories of psychological development and a demand for new explanations to prevent and remediate human problems. In 1975 a new discipline, The Psychology of Women, was established providing a foundation or (at least one academic home) for an applied science dedicated to counseling women.

While the consideration of gender in the prevalence, etiology, diagnosis, and treatment of human problems had long been neglected (*Age, Race, and Gender in Organizations*), there is now increasing awareness that the labeling of illness, the election to seek care, the dialogue between the professional and the 'helper,' and the 'cure' offered are all transactions that are influenced and bound by social organizations, as well as socially constructed themselves.

2.1 *Data Supporting the Disproportionate Number of 'Mentally Ill Women'*

Studies of clinical and community samples report that a high proportion of individuals with signs of depression, anxiety, panic, and eating disorders are women. Community health and mental health utilization rates are higher for women and a disproportionate share of psychoactive drugs are prescribed to women—many with deleterious or unknown side effects. In addition, it has been suggested that continuous exposure to disempowered positions may be

associated with women's depression (Feree et al. 1999). In addition physical and sexual abuse has been recorded in significant numbers of women seeking mental health care (*Childhood Sexual Abuse and Risk for Adult Psychopathology*).

2.2 Dissatisfaction with Traditional Treatment

Worell and Remer (1992) in a classic text delineate numerous sources of discontent with prevailing treatment options.

(a) Dissatisfaction with traditional theories that depicted male traits as the norm and women as deficient by comparison.

(b) Omission of women from the knowledge base of psychology.

(c) Failure to recognize that psychopathology may reflect power imbalances rather than intrapersonal difficulty.

(d) Uncovering sex bias in psychotherapy.

(e) Disregard by many health professionals for women's self reported experience.

(f) The practice of attributing blame and responsibility to women for their experience of sexual and physical violence.

(g) 'Mother blaming' such that women's interdependence and involvement is pathologized and men's responsibility for their abuses of power remains unchallenged.

(h) Increasing medicalization of women's psychological problems, including misuse of diagnoses and prescriptive drugs.

3. From The Western Individual to The Global Issue: Changing Focus Over Time

The initial focus of feminist therapy was to name the biases indicated above and to create a validated discipline to address them. Consciousness raising groups increased personal insights and improved relationships between women. Although these groups helped women see how the personal could be political, the wider political impact of these groups was limited and the focus was on the individual (Seu and Heenan 1998). The movement was primarily contained to western societies with the experience of the 'other' women of the east infrequently noted.

The next step was to describe and implement clinical practice. Alongside grass roots movements, scholars began to focus on issues specific to women's needs. An important contribution was Gilligan's (1982) *In a Different Voice* in which she argued that women's relationship-oriented concern for care and connectedness is devalued by the dominant androcentric culture's emphasis on logic and reasoning. She called for men and women to embrace a wider range of emotional qualities.

The 1980s and 1990s also saw a proliferation of writing, primarily in the USA, on ways to teach 'empowering behaviors' such as assertiveness training and psychoeducation.

In addition, psychoanalytic scholars in America and Europe explored ways of incorporating feminist object relations theory and its emphasis on connectedness into its framework, while Lacan (1977) provided direction for many professionals working on cultural theory.

Although the field became more adept at questioning, the answers were not forthcoming and indeed the inherent tension between political theory and therapeutic practice has not been resolved. There has, however, been greater attention paid to the global condition of women.

As a result, the third stage within feminist therapy will be the naming of the biases within the profession (Seu and Heenan 1998). Influenced by postmodern deconstructionist thought, there have been continuing challenges to the notion of 'feminist therapy' as a singular idiom. Eurocentric and heterosexual biases in teaching and treatment have been criticised and the esteemed position of gender bias as the 'ultimate' social oppression has been questioned.

Feminist analyses have become increasingly cross-cultural and cross-disciplinary as a greater effort has been made to understand not only the complexity of the social issues 'masquerading' as individual psychopathology but the near universal struggle to achieve and maintain power in a complex, modernizing, global environment.

Studies exploring the social construction of illness—borrowing from sociology, anthropology, and history as well as medicine and psychology have focused as much on gender issues as on power issues. By widening the gender lens to survey hierarchal imbalances more generally, feminist theories have found a way to answer male and female needs in an array of areas from self-definition, personal agency, and respectful care delivery to the negotiation of improved working environments and rejection of limiting fashion norms.

The tension between the individual and the societal approach along with connotations attached to 'feminisms' has led to several interesting questions for the future. For example (a) should a 'true' feminist approach to care even involve treating the individual? or (b) do we even need the term feminist at all if in fact the goal is parity for all humans?

4. Research Questions and Concerns

To amply answer the questions generated by a feminist perspective one must first gain support from societal (governmental, academic, and personal) agencies that manage funding. That questions of female mental health have been omitted in the past reflect that this is

not all easy. However, all who study the mental health of women need not be feminists or utilize feminist research methods, in fact they could be men!

Striegel-Moore (1994) has delineated research guidelines for feminist scholarship in this field: affirmation of a positive view of women (challenges the deficit model of women vs. men), adoption of a 'contextual' approach, utilization of a broad spectrum of research methods (applies qualitative and quantitative methods and establishes a collaborative relationship with research participants), and consideration of the implications of research findings for social change (studies diverse women in natural settings).

The challenge for the future will be to apply these methods across disciplines and across countries to help answer questions such as 'how the outside (environmental influences) get inside (and become individual pathology)?,' and when problems are mostly female specific, Why women? Why now? (at this point in history). It may be that by looking across specific diagnoses to syndromes of female distress that new approaches to treatment and prevention will be found (see *Syndromal Diagnosis versus Dimensional Assessment, Clinical Psychology of*). It may also be possible to better understand societal dilemmas by decoding these personal struggles.

5. *Future Directions: The World as Living Laboratory*

Changing societies and gender roles may provide *in vivo* opportunities to explore the impact of modifying political and social structures on human adjustment. Cross- disciplinary input will be the key. In addition, access to technology may enhance efforts to distribute knowledge and care to a broader audience (see *Technological Change in a Multicultural Context, Applied Psychology of*). The impact of globalization and technology on women's freedoms and their applications in the therapy environment will be critical areas for study in the twenty-first century.

See also: Feminist Theory: Psychoanalytic; Gender and Feminist Studies in Psychology; Women and Psychiatry

Bibliography

Brown L S 1994 *Subversive Dialogues: Theory in Feminist Therapy*. Basic Books, New York

Feree M M, Lorber J, Hess B B 1999 *Revisioning Gender*. Sage Publications, London

Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA

Hare-Mustin R T, Marecek J (eds.) 1990 *Making a Difference: Psychology and the Construction of Gender*. Yale University Press, New Haven, CT

Lacan J 1977 *Ecrits: A Selection*. Tavistock, London

Seu I B, Heenan M C 1998 *Feminism and Psychotherapy: Reflections on Contemporary Theories and Practices*. Sage Publications, London

Striegal-Moore R 1994 A feminist agenda for psychological research on eating disorders. In: Fallon P, Katzman M A, Wooley S (eds.) *Feminist Perspectives on Eating Disorders*. Guilford Press, New York

Worell J, Remer P 1992 *Feminist Perspectives in Therapy—An Empowerment Model for Women*. Wiley, New York

M. A. Katzman

Feminist Theology

Feminist theology presumes what many world religions have for centuries precluded—namely, women's full humanity, the immediacy of women's relation to the divine, and the cultural import of women's reflections on the sacred. Committed to supporting cultural transformation through the politics of spiritual practice (Spretnak 1994), feminist theology has attempted to dismantle dominative power and hierarchy in all their guises—from the private, familial and sexual to the institutional, ecclesiastical and political.

In the interest of transforming power relations, feminist theology works both to deconstruct the androcentric religious imagination, its philosophical underpinnings and its socio-political incarnations, and to construct or to create livable, religious imaginals that shelter the cultural becoming of women and that generate the formation of open, just and egalitarian communities. The term feminist theology, broadly defined, may include women's re/invention of goddess and wiccan traditions, critical reconstruction of the world religions, i.e., Judaism, Christianity, Islam, and Buddhism, as well as women's philosophical, non-theistic, or spiritualist departures from them.

1. *Philosophical Background*

Philosophically speaking, feminist theology developed in response to the 'deracination' (Daly 1984) of women's lives within the scope of patriarchal, monotheistic religions. Deracination names the effect on women of those religions whose engagement with 'the transcendent' subliminally assumes an abjection of matter—from the Latin 'mater' meaning mother (Kristeva 1982). Inasmuch as Platonic philosophy—which radically split spirit from matter, while aligning 'man' with enspirited mindfulness and 'woman' with (fallen or corrupted) nature—has served as infrastructure for the western monotheistic religions, women have been engendered theologically as devoid of spirit. As a sexual and reproductive body, woman

was subjugated to male authority, e.g., wife to husband, daughter to father, penitent to male priest, etc. Additionally inlaid with Aristotle's hierarchical 'chain of being,' western religions granted men greater ontological status and the privilege, therefore, of ruling over women, animals and earth—all of which, given their assumed passive and receptive contours, awaited man's invigorating agential influence. Built upon the story of man's 'second birth' into the transcendent realms, a birth without resort to woman, these religions idealized male selfhood and generated essentially homosocial cultures (Irigaray 1985). These religions libidinally installed an androcentric socio-political and economic matrix which feminists named 'patriarchy,' because all 'goods' (in terms of values and political rights as well as economics) were circulated through the generations of the fathers and sons. Consequently, women within the purview of these religions have been, as philosopher Daly put it, 'physically, mentally, and emotionally separated from our Original, elemental Race…' (1984). That is, women have been religiously precluded from exercising historical agency on behalf of ourselves as women and on behalf of the culture at large.

Androcentric religions have in various ways instilled in women a sense of self-loathing, existential shame and fear of autonomy. In Christianity, e.g., women have been blamed for all frustrations with the flesh, rationalized as the consequence of Eve's eating of the Tree of Knowledge of Good and Evil (Genesis 3). According to the normative reading of this story, one woman's exercise of free will ontologically fractured and debased the world, bringing death, disease, suffering, and transience. Since the fleshy body has been viewed as the inheritance of the mother; and since woman has been viewed as a derivative of man (Genesis 2, 1 Corinthians 11:8), Christianity disavowed the flesh and imagined salvation as a resurrection of 'the (ideal or imaginal) body' freed from the 'bondage' of development and decay, death, disease, and sexual desire (Romans 8). Not only do women thus serve as social scapegoats for Christianity, women have been the physical buffers between the projects generated by 'man's self-infinitizing spirit' (Ruether 1983) and the material impingements of finitude. That is, by socially domesticating women, men withhold from the purview of culture the needs and delimitations of physical bodies, including 'bodies' of land, water, etc.

2. *Theoretical Contours*

Insomuch as the women's movement initially understood itself to be a 'spiritual' movement (Christ and Plaskow 1979, Spretnak 1994), feminist theology could be said to have been inherent to feminism. Women's exclusion from societal power structures was, feminists surmised, fundamentally linked to women's exclusion from socio-religious symbol systems. As Daly observed wryly, where 'God is male, the male is God' (1985b)—a sentiment seconded by Spretnak, who observed that '(t)he long-standing justification in Western culture for regarding the male as the norm… has been the implicit acknowledgement that men are the same sex as God' (1994). This insight into the culturally constitutive nature of the religious imagination, a critique supported by Feuerbach's definition of religion as projection (*The Essence of Christianity* 1957), led feminist theologians to attempt to interrupt the cultural absolute of male socio-political power by displacing male religious images and by interrupting their sacred storylines. (Consequent to this insight and owing to feminists' own interest in 'changing the divine subject' so as to authorize women's power thereby, women's theological enterprise has sometimes been named 'feminist thealogy'—'thea-' being the feminine of the Greek masculine 'theo-,' meaning 'God.' See Goldenberg (1979).) Given the incarnational relation between religious symbol and cultural power relations, among feminist theologians' first tasks was the critique of the masculine images by which religions have referred to the sacred as God, Father, Lord, King, Son, etc., and to humanity as 'sons' or 'men of God.' Not only do such images exclude women as religious and political subjects, these images—feminist theologians were convinced—led to the abuse and violation of women and children. 'Women's experience,' feminist theology's epistemological locus, referred to women's shared experiences of living under the purview of this exclusive male engendering of power. Specifically, such images religiously legitimated the practice of dominative power in relation to all that the system conversely objectified. Hence, women experienced this 'dominion of God' as 'domination'—specifically as battering, rape, poverty, the dislocation and domestication of our lives, etc.

Citing instead texts such as Genesis 1:27 ('So God created humankind in (God's) image,… male and female…'), Galatians 3:28 (In Christ, 'there is no longer male and female…'), and the Pentecostal outpouring of the Spirit upon 'all flesh' (Joel 2, Acts 2), feminists asserted women's unmediated relationship to the sacred, claiming thereby women's theological authority, the validity of women's spiritual experience, and the theological subject matter of women's daily lives. Authorized by Spirit, feminist theologians began a critique of male clerical and ecclesiastical leadership, of the male authorship and interpretation of the canon, of the functional closure of revelation vis-a-vis the sacred canon. Feminist theology not only supported, but was identified actively with the Women–Church movement and the movement for women's ordination—accomplished in Reformed Judaism as well as within many Protestant (e.g., Presbyterian, Lutheran, American Baptist, United Methodist) and Anglican–Episcopal Christian

circles. Yet the problem for women in the religions of 'the sacred book' remained: sacred texts have been written by and for men and have circulated for centuries in closed circuits of male interpreters. Where females did appear in these texts, feminists soon surmised that men had 'used (these) women to think with' (Karen King). Further, the religious sacraments of these communities have often worked to delegitimate even that which was prescribed as 'women's work,' e.g., in the church birthing becomes baptism, feeding becomes eucharist. Faced with the question of whether sacred texts and existing religious institutions had the symbolic resources to promote women's well-being, some women, beginning in the early 1970s and onward, have left and invested themselves in the recovery of goddess and wiccan spiritualities, traditions which have supported human embodiment (Eller 1993). Among those women who (also) stayed related to the existing monotheistic religions, a 'hermeneutic of suspicion' was employed to generate a critical distance from canonical texts. By invoking 'suspicion,' Biblical scholar Schüssler-Fiorenza discouraged women from expecting to discover truth 'in' a text conceived by men and encouraged feminists instead to deduce the 'truth' of scripture by a text's ability to raze systems of domination and exploitation (1984). Another feminist strategy for creating an authority base has been the recovery of the hidden histories of women, both in canonical texts, e.g., Wire's rhetorical positioning of Paul's otherwise anonymous discursants as *The Corinthian Women Prophets* (1990), and in Christian communities throughout the centuries, e.g., Ruether and McLaughlin's *Women of Spirit: Female Leadership in the Jewish and Christian Traditions* (1979). Nevertheless, Trible's *Texts of Terror* (1984) reminded feminists that, given the patterns of 'inferiority, subordination and abuse of the female in ancient Israel and the early church' (3), e.g., Hagar, Tamar, and the daughter of Jephthah, women's tragic stories do not always yield to a positivist reconstructive resolution. Along with the recuperative labor of feminist Biblical and historical theologians, women have attempted to re-imagine the sacred, a project that has outraged some denominational leaders and led to heresy charges against some women. First, recovering feminine images of the divine from scripture, e.g., Sophia, Wisdom, Spirit, feminists have then applied themselves to rethinking ritual and theology based upon these images. Among academic theologians, McFague's (1987) *Models of God* developed the metaphors of God as Mother, Lover, and Friend so as to promote a change of consciousness—from a consciousness founded upon the theological notion of transcendence as separation from and ascendance over women and nature, towards a consciousness of belonging to the 'evolutionary ecosystem of our cosmos' (9). In *She Who Is* Johnson (1992), proposed a Christian trinitarian model of Spirit–Sophia, Jesus–Sophia and Mother–Sophia—a constructive project aimed at thinking women into the economy of divine love, while also thinking divine sociality in terms of the lives of women.

As feminist theology developed, feminists, observing that traditional theology reflected only men's spiritual experience, undertook a reconstruction of the central doctrines of their respective communities. Whereas Christian theology had, following Augustine, psychically repressed the corporeal body, senses, emotions, and relations, feminists made the body—women's minds, wills, desires, and relationships—the central site of revelatory encounter and salvific experience. Whereas God had been conceived as absolutely transcendent, feminists have insisted on divine immanence—hence, the transition from invoking the divine as 'God' to 'Spirit.' Similarly, if according to the androcentric imagination creation was said to have transpired by divine fiat or to have come about through artisanal handicraft, i.e., God as 'Creator' or 'Maker,' feminists have spoken of the divine 'birthing' the world, of the world as 'God's Body' (McFague 1993, Jantzen 1984) and of humans as cocreators of cosmic process.

Finally, whereas for Christians Jesus' crucifixion has been seen historically as the occasion for sacrificial grace entering in to right a recalcitrant and wayward world, feminists have questioned positioning a human injustice as an ontological necessity and have gravitated toward two other views. First, feminists, rather than memorializing a unique individual from the historical past, reconstrued christology as the prophetic practice of justice and egalitarian relations within the on-going Jesus-community. Second, some feminists have shifted the christological focus away from crucifixion and towards incarnation. The use of the incarnational metaphor, pervasive in contemporary feminist philosophical theory, resonates with the histories of Christian women who have asserted their subjective agency in and through this analog, e.g., Prous Boneta and Guglielma of Milan (Ruether and McLaughlin 1979).

3. *Context and History of Development*

While prefigured by the late nineteenth century feminists (e.g., Stanton's *The Woman's Bible* 1895), 'feminist theology' refers to the theological critique and reconstruction of religious symbols and systems over the past four decades of the women's movement (1960s–1990s). Initial publications by Schüssler-Fiorenza and Daly took place in Europe (1968), where they had worked with other early European feminists, including Catherina Halkes. By the early 1970s, Schüssler-Fiorenza and Daly re/turned to the United States where they joined the scene that included Rosemary Ruether, Letty Russell, and Judith Plaskow. Women's theological writing swelled during the 1970s and 1980s with publication's such as Daly's

Beyond God the Father (1985a) and *Pure Lust* (1984), Ruether's *Sexism and God-Talk* (1983) and Schüssler-Fiorenza's *In Memory of Her* (1983).

As feminist theory became conscious of the contextual situatedness of knowledge, feminist theology was refracted increasingly through the differences of all who had been forced together under the universalization of history's anonymous 'other,' 'woman.' Inasmuch as 'women's experience' was invoked as the epicenter of feminist theology in the 1970s, and inasmuch as this category was—women soon learned—dependent upon contextual, ethnic, class, and race factors, feminist theology in the 1980s became pluralized. Written now as feminist theologies, the term recognized the distinctive liberative agendas of women of color.

African–American women, protesting that feminist calls for 'sisterhood' glossed over black–white racism and the double oppression of race and gender, gleaned the name 'womanist' from Alice Walker's *The Color Purple*. If feminists read with suspicion, womanists—Delores Williams suggested—read with an eye toward survival; she found the epitome of the black woman survivor in Hagar (1993). Womanists' theological reflection often incorporates black women's literature, especially that of Toni Morrison, Alice Walker, and Paule Marshall, womanists having found in this literary canon models more relevant to the goal of improving black women's quality of life.

As differences continued to refract, Hispanic women like Ada Maria Isasi-Diaz wrote *Mujerista Theology* (1996). The indigenous voice of Paula Gunn Allen in *The Sacred Hoop* (1986) interrupted any too facile linkage of North America with the Christian continent. Jewish women resisted the supercessionist imperialism of Christian women (Plaskow 1990). From the beginning, lesbian voices have checked Christianity and Judaism's too facile heterosexist perspectives (e.g., Carter Heyward 1989 *Touching Our Strength*).

As with women of color in North America, women's theology in the densely populated and economically poorer nations has had to address survival needs and has often been undertaken without standing against men, who may consider the gender question secondary to the community's goal of liberation from neocolonism. Significant voices from feminist theology in global perspective include Mercy Amba Oduyoye from Ghana, Kwok Pui-Lan of Hong Kong, Chung Hyun-Kyung from South Korea, Mary John Mananzan of the Philippines, Ivone Gebarra from Brazil and Elsa Tamez from Costa Rica. Women from the nations evangelized by the western religions have come more recently to celebrate the hybridity of their religious subjectivity. The recovery of indigenous traditions—some of which, like Korean shamanism, claim stronger spiritual leadership roles for women—have become significant factors in creating out of the received religious tradition a vital and practical religious syncretism (Chung, Hyun Kyung 1990 *Struggle to be the Sun Again*).

In summary, the pluralization of 'woman' lead to distinct feminist theologies. Questions remain as to whether 'feminist' can serve as a canopy or whether such a move asks women of color to submit to the universalism of white, western experience. Though European theologians, e.g., Dorothee Soelle and Luise Schottroff, have—given greater impetus by their proximity to socialist analysis—included more economic reflections, feminist theology in general has not done as well at critiquing class as race issues. If feminist theology initially was configured by uniting women around victimization, increasingly feminist theologies have attempted to build constructively around women's shared goals. Parallel to this, a significant shift in the energies of feminist theologies can be felt—a move away from the ensconcing of opposition, especially 'resistance,' a position still within the logic of dualism; a move towards a 'nonapocalyptic' feminism, employing a 'methodological ambivalence' to sift through the 'power lines' so as to transform cultures (Catherine Keller 1996 *Apocalypse Now and Then*).

4. *Probable Directions*

In her 1975 book *New Woman New Earth*, Ruether noted that '(s)exism and ecological destructiveness are related in the symbolic patterns of patriarchal consciousness…' The disclosure of this symbolic connection between women's oppression and the domination of the earth led to the ecofeminist theology of the 1990s. If the collections of ecofeminist essays are indicative (e.g., Adams), ecological concern may convene an even greater diversity of feminist theologians—including women activists from Buddhism, Hinduism, and indigenous religions worldwide. Since women's bodies have served as buffer from life's material impingements for those who believe in the extraterrestrial transcendence of androcentric religions; and, inasmuch as ecosystemic collapses from industrial, agricultural, and multinational industries appear to be increasing, women—who, according to UN statistics, perform two-thirds of the human work hours, receive 10 percent of the pay, and own less than 1 percent of the world's habitable land—will be increasingly adversely impacted (*Progress for Half the World's People* in issues of the 1980s by the UN Association of the US 1980). Insofar as feminist theologies remain orthopraxic, i.e., committed to good practice, rather than correct belief ('orthodoxy'), a burgeoning of ecofeminist theologies might be expected.

Further, a new interest in the sciences could revitalize a certain stream of feminist theology. In the 1980s, the articulation of a relational ontology, undergirded with the process metaphysics of Alfred North

Whitehead, lead to the process-oriented theologies of Rita Nakashima Brock, Catherine Keller, Marjorie Suchocki, and Sheila Devaney, among others. More recently, feminist philosophers of science, e.g., Carolyn Merchant, Donna Haraway, and Vandana Shiva, have been critiquing Christianity's influence upon western scientific paradigms. Given this alliance of interest, feminist theologians might be expected to further the deconstruction of Western Christian mythemes in hopes of interrupting the planetary destructiveness of current and envisioned biotechnologies. At the same time, feminist theologians could be expected—given the fertile overlap between the new physics and theology—to offer new cosmological articulations.

In the 1990s, feminist theology took on a decidedly philosophical hue. To be sure, women like Daphne Hampson (England) and Daly, convinced that Christianity was and would remain inherently patriarchal, dated themselves 'post-Christian.' Consequent to that move, Daly insisted that she wrote philosophy, not theology. However, for those who continue to work within the sacred texts and traditions of Christianity and Judaism, postmodern discourse—heralding the end of totalizing worldviews and the disappearance of the subject—has caused feminist theology—like feminist theory, upon which it is dependent—to rethink its critical terms. While both a totalizing worldview, i.e., 'patriarchy,' and the evocation of female subjectivity had been used to achieve feminism's early liberative goals, this new philosophically reflective stage should not be seen necessarily as a departure from feminist theology's commitment to political transformation. Increasingly, postcolonial theory—especially in the field of Biblical studies—and deconstruction are being used to revitalize or reshape liberative theologies. Sharon Welch modeled the viability of employing post-structuralism on behalf of emancipatory transformation in her 1985 *Communities of Resistance and Solidarity* as did Rebecca Chopp in her 1989 *The Power to Speak*. More recently, the French psychoanalytic and poststructuralist insights of Kristeva and Irigaray have been used as hermeneutical tools to address topics such as eucharist, belief, the incarnation, spirit, etc. (Kim et al. 1993). Nevertheless, whether or not this philosophical tone might— amidst the transformation of other critical terms, e.g., the divine, transcendence, belief, etc.—represent a reconsideration of what counts as theology remains an unarticulated and therefore unanswered question for the field.

See also: Feminist Epistemology; Feminist Ethics; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Radical Lesbian; Religion and Gender; Religion: Evolution and Development

Bibliography

Adams C A (ed.) 1993 *Ecofeminism and the Sacred*. Continuum, New York

Christ C P, Plaskow J (eds.) 1979 *Womanspirit Rising: A Feminist Reader in Religion*. Harper & Row, San Francisco

Daly M 1984 *Pure Lust: Elemental Feminist Philosophy*. Beacon Press, Boston

Daly M 1985a *Beyond God the Father: Toward a Philosophy of Women's Liberation*. Beacon Press, Boston

Daly M 1985b *The Church and the Second Sex*. Beacon Press, Boston

Eller C 1993 *Living in the Lap of the Goddess: The Feminist Spirituality Movement in America*. Crossroad, New York

Goldenberg N R 1979 *Changing of the Gods: Feminism and the End of Traditional Religions*. Beacon Press, Boston

Irigaray L 1985 *Speculum of the Other Woman*. Cornell University Press, Ithaca, NY

Irigaray L 1993 *Sexes and Genealogies*. Columbia University Press, New York

Jantzen G 1984 *God's World, God's Body*. Westminster, Philadelphia, PA

Kim C W M et al. 1993 *Transfigurations: Theology and the French Feminists*. Fortress, Minneapolis, MN

Kristeva J 1982 *Powers of Horror: An Essay on Abjection*. Columbia University Press, New York

McFague S 1993 *The Body of God: An Ecological Theology*. Fortress Press, Minneapolis, MN

Plaskow J 1990 *Standing Again at Sinai: Judaism From a Feminist Perspective*. Harper & Row, San Francisco

Ruether R R 1983 *Sexism and God-Talk: Toward a Feminist Theology*. Beacon Press, Boston

Ruether R, McLaughlin E 1979 *Women of Spirit: Female Leadership in the Jewish and Christian Traditions*. Simon and Schuster, New York

Russell L M (ed.) 1988 *Inheriting Our Mother's Gardens: Feminist Theology in Third World Perspective*, 1st edn.Westminster Press, Philadelphia

Schüssler-Fiorenza El 1983 *In Memory of Her: A Feminist Theological Reconstruction of Christian Origins*. Crossroad, New York

Schüssler-Fiorenza El 1984 *Bread Not Stone: The Challenge of Feminist Biblical Interpretation*. Beacon Press, Boston

Spretnak C (ed.) 1994 *The Politics of Women's Spirituality: Essays by Founding Mothers of the Movement*, 2nd edn. Anchor/Doubleday, New York

Stanton E C 1895 *The Woman's Bible*. European Publishing Co., New York

Trible P 1989 *Texts of Terror*. Fortress, Philadelphia, PA

Williams D S 1993 *Sisters in the Wilderness: The Challenge of Womanist God-Talk*. Orbis Books, Maryknoll, NY

Wire A C 1990 *The Corinthian Women Prophets*. Fortress, Minneapolis, MN

S. V. Betcher

Feminist Theory

Feminist theory encompasses a range of diverse ideas, all of which originate with the following beliefs: (a) society is patriarchal, structured by and favoring men;

(b) traditional ways of thinking support the subordination of women and the neglect or trivialization of issues particularly affecting women; and (c) this patriarchal order should be overthrown and replaced with a system that stresses equality for both sexes. Feminist theory impacts all institutions—medical, legal, academic, and social, for example—and can be used to illumine all issues affecting humans. The diversity of thought within feminist theory lies in the fact that women across the world differ from each other in many ways—including race, ethnicity, nationality, class, sexual orientation, or educational background, for example—and that these differences result in varying views of feminism and gender equality as expressed in liberal feminist, Marxist–socialist, radical 'libertarian' and radical 'cultural,' postmodern, and global feminist theories. Yet for all the diversity within feminist theory, there remains the belief that despite women's many differences, women everywhere share some basic 'sameness.'

In understanding feminist theory or feminist philosophy, it is necessary to understand theory and why humans develop theories. Philosophy helps us understand what it means to be rational beings, to have knowledge, and to express virtues such as benevolence or justice. Humans develop theories to explain elements of reality, how we interact with one another, and who we are as persons.

However, the history of philosophy and those humans who have developed the theories has centered around male humans. Western philosophy is dominated by the writings and thinking of men and, therefore, it often reflects stereotypes or biases against women, or simply omits any point of view that might be particular to women and their lives. For example, basic concepts of political theory—autonomy, freedom, individualism, power, competition, justice—have been designed to fit the experiences of males in male-dominated cultures (Kourany 1998). Particularly reflecting the daily lives of men who do not birth and raise children or provide the major, mundane caretaking duties for a family, these ideas represented for many centuries the heights to which human nature should aspire. Conversely, traits associated with women for many centuries—emotionality, passivity, cooperation, nurturing, compassion—were thought of as inferior and weak. Clearly, here are philosophical ideas that favor the experiences of men over women, relegating women's experiences to the inferior realm.

Feminist theory, usually described as beginning in the late eighteenth and early nineteenth centuries along with women's movements and suffrage movements, reconceives these historical theories and includes the experiences or general characteristics of women in addition to men. Women are, after all, the sex with the capacities of gestation, lactation, and menstruation. There are specific differences in the bodily functions of women and men. Traditionally, women have been associated with these bodily functions, associated with the 'lesser' of many hierarchical oppositions: mind/body, sacred/profane, reason/emotion, active/passive, spiritual/material. This duality—particularly where it privileges so-called 'masculine' virtues and values over 'feminine' ones—is rejected in feminist theory.

Feminist theory, as a 'new' way of thinking, is sometimes difficult to understand because it does not represent one uniform view. The basic premise of feminist theory is that women have been excluded from traditional philosophy; moreover, not only have women been excluded, but also people of 'different' economic classes, races, ethnicities, or nationalities. Feminist theory recognizes many flaws in traditional (Western, Judeo–Christian, European, male) philosophy and offers a new way of thinking about issues affecting humans and our world. Additionally, feminist theory operates with the understanding that our current society is patriarchal, structured by and favoring men, and that this patriarchal order should be overthrown and replaced with a system that emphasizes equal rights and just and fair distribution of resources for both sexes. Because this element of feminist theory particularly impacts political and moral theory and the basic functions of our society, there are, as mentioned above, different 'political' interpretations or understandings of feminist theory, ranging from so-called liberal perspectives to so-called global perspectives. So great is the diversity of feminist theory that some critics fail to see that no matter how they label their particular way of thinking, virtually all feminist theorists agree that in order to be 'feminist' a theory must (a) proceed on the assumption that women and men do not share precisely the same situation in life; (b) offer action guides 'that will tend to subvert rather than reinforce the present systematic subordination of women'; (c) provide strategies for dealing with issues that arise in private or domestic life; and (d) 'take the moral experience of all women seriously, though not, of course, uncritically' (Jaggar 1992).

1. *Different Feminist Theories*

Liberal feminism receives its classic formulation in Mary Wollstonecraft's *A Vindication of the Rights of Woman* and in John Stuart Mill's *The Subjection of Women*, two nineteenth-century works. Liberal feminists find the roots of women's oppression in the cultural constraints that hinder women from competing in the public world, including the worlds of politics, medicine, law, and finance. Because traditional society holds the false belief that women are, by nature, less mentally and physically capable than men, society often excludes women from the arenas mentioned above. Insisting this is discriminatory, liberal feminists demand that society give women the same educational and occupational opportunities that men

have. Liberal feminism is clearly reflected in the work of the National Organization of Women, current laws against sexual harassment or discrimination in the workplace, and past lawsuits against same-sex schools that receive funding from the state, including female tax-payers.

In contrast to liberal feminism, radical feminism holds that the patriarchal system that oppresses women is so flawed it cannot be reformed, but must be completely eliminated. In overthrowing patriarchy's institutions and assumptions, radical feminism singles out for attack social and cultural institutions such as the family and church, as well as traditional ideologies about the sexual and reproductive rights, responsibilities, needs, and preferences of men and women.

Initially, radical feminists—sometimes referred to as 'radical libertarian feminists' (Tong 1998)—aimed to explore what they saw as the pleasures of sex: consensual sex between men and women, lesbian sex, sex with both men and women, autoeroticism, sadomasochistic sex, and even intergenerational sex. They sought to free women from the beliefs that 'good' sex could be experienced only in a 'love relationship,' and that sex for sex's sake was somehow 'bad' or promiscuous. In addition, radical libertarian feminists wished to help women avoid the burdens of human reproduction, going so far as to recommend that natural reproduction be replaced by technological reproduction. They agreed with thinkers such as Shulamith Firestone that no matter how much educational, political, and economic equality women achieve, nothing fundamental will change for women so long as their reproductive roles remain the same. Natural reproduction, said Firestone, is not in women's best interests. Pregnancy is 'barbaric,' and natural childbirth is 'at best necessary and tolerable' and at worst 'like shitting a pumpkin' (Firestone 1970) and had best be replaced by *in vitro* fertilization and *ex utero* gestation. What is more, insisted Firestone, natural reproduction is not in men's or children's interests either, for as soon as men and women are truly free to engage in polymorphous, perverse sex, it will no longer be necessary for men to display only masculine identities and behaviors and for women to display only feminine ones. Freed from their gender roles at the level of biology (i.e., reproduction), said Firestone, women would no longer have to be passive, receptive, and vulnerable, sending out 'signals' to men to dominate, possess, and penetrate them in order to keep the wheels of human procreation spinning. Instead, men and women would be encouraged to become either equally masculine and feminine (mono-androgynous) or as differently masculine and/or feminine as they wished (polyandrogynous). As a result, not only would men and women become androgynous persons, all of culture would become androgynous. Furthermore, in this newly-evolved androgynous culture, the categories of the technological and the aesthetic, together with the categories of the masculine and the feminine, would disappear through what Firestone termed 'a mutual cancellation—a matter–antimatter explosion, ending with a poof!' (Firestone 1970).

Firestone's 'poof' proved too much not only for the general public and most liberal feminists, but also for some radical feminists, particularly those who were beginning to wonder whether women would really gain true liberty by engaging in permissive sex, refusing to bear children, and becoming androgynous persons, that is, persons who are free to develop both 'female' qualities—nurturance, compassion, tenderness, sensitivity, cooperativeness—and 'male' qualities—aggressiveness, leadership, initiative, competitiveness. This concerned group of radical feminists—sometimes called 'radical cultural feminists' (Tong 1998) or 'essentialists' (Alcoff 1988)—began to caution that sex, usually understood as heterosexual sex, is more dangerous than pleasurable for most women. They urged women to extricate themselves from the institution of so-called compulsory heterosexuality (Rich 1994), which they viewed as 'characterized by an ideology of sexual objectification (men as subjects/masters; women as objects/slaves) that supports male sexual violence against women' (Ferguson 1984). Radical cultural feminists insisted that, as it has been experienced so far, heterosexuality is men's sexuality. It is about men seeking to control women's sexuality: representatively in pornography, and actually through the use of prostitutes or the selective harassment, rape, and physical abuse of women in their power. Only if women can free themselves from sex as men want it, can women discover what sex as women want it might be.

In addition to stressing the dangers of heterosexual relations and the pleasures of lesbian relations, radical cultural feminists emphasized that artificial reproduction would more likely disempower than empower women. They urged women to see artificial insemination by donor, *in vitro* fertilization, and plans for an artificial womb not as new procreative options for women but as means for men to exercise complete control over women's procreative powers—their ability to determine whether the human species will continue or not (Rich 1976, Ruddick 1984).

Finally, radical cultural feminists rejected the idea of androgyny as a desirable goal for feminists, replacing it with proposals to affirm women's essential 'femaleness' (Echols 1983). Far from believing the liberated woman must exhibit both masculine and feminine traits and behaviors (traits or behaviors culturally or stereotypically associated with a particular gender), radical cultural feminists expressed the view that it is better to be female/feminine than it is to be male/masculine. Women should not try to be like men, they said. On the contrary, they should try to be more like women, emphasizing the values and virtues culturally associated with women including 'interdependence, community, connection, sharing,

emotion, body, trust, absence of hierarchy, nature, immanence, process, joy, peace, and life' (Jaggar 1992)), and de-emphasizing the values and virtues culturally associated with men including 'independence, autonomy, intellect, will, wariness, hierarchy, domination, culture, transcendence, product, asceticism, war, and death' (Jaggar 1992). Women's liberation, according to radical cultural feminists, exists in women's conviction that it is better to be women than men.

Somewhat unconvinced by the liberal and radical feminist agendas for women's liberation, Marxist–socialist feminists claim it is impossible for anyone, especially women, to achieve true freedom in a class-based society where the wealth produced by the powerless many ends up in the hands of the powerful few. With Engels, Marxist–socialist feminists insist women's oppression originated in the introduction of private property, an institution that obliterated whatever equality of community humans had previously enjoyed. Private ownership of the means of production by relatively few persons, originally all male, inaugurated a class system whose contemporary manifestations are corporate capitalism and imperialism. Reflection on this state of affairs suggests that capitalism itself, not just the larger social rules that privilege men over women, is a primary cause of women's oppression. If all women—rather than the 'exceptional' ones alone—are ever to be liberated, the two-headed beast of capitalist patriarchy or patriarchal capitalism (take your pick) must be killed. Agreeing with this thought, Marxist–socialist feminist Mitchell (1971) stressed that women's condition is determined by the structures of production (as Marxist–socialists believe), reproduction and sexuality (as radical feminists believe), and the socialization of children (as liberal feminists believe). Women's status and functions in all of these structures must change if women are to be men's equals.

Common to liberal, radical, and Marxist–socialist feminist thought is a desire to view women as somehow the same, even if this sameness falls short of the extreme 'essentialism' of radical–cultural feminism. Women's sisterhood and solidarity are the dominant notes of feminist theory as discussed so far. The problem with such a view of women is, however, that women are not only different from men, but also from each other depending on their class, race, ethnicity, nationality, age, religion, marital status, sexual orientation, and so forth. This point about women's difference, about the tendency of some feminist theorists to confuse 'Women' with a particular kind of woman—perhaps white, Western, middle-class—is the core conviction of multicultural feminism. In the 1990s, the buzzword 'cultural diversity' affected social, legal, educational, and religious institutions in the US, and multicultural feminism emerged as part of this diversity. Multicultural feminists stress that although it is hard to be a woman, it is harder to be a woman of color than a white woman; a poor woman than a rich woman; a lesbian woman than a heterosexual woman; an old or average-looking woman than a young or beautiful woman. Repeatedly, multicultural feminists explain how the idea of 'sameness'—the seemingly benevolent notion that 'down deep' we are all the same—could counter-intuitively be used as an instrument of oppression rather than liberation. In this connection, multicultural feminists reflect in particular upon the works of feminist theorists such as Spelman (1988) who pleaded with feminists not to make the mistake historian Kenneth Stampp made by asserting 'that innately Negroes are, after all, only white men with black skins, nothing more, nothing less.' Why, asked Spelman, is it that black men are really white men 'down deep' and not that white men are really black men 'down deep?' Could it be that white people still think 'white' is definitely the best way to be—the 'gold standard' for all people? Fearing the presence of some well-intentioned 'Kenneth Stampps' within the circle of feminist theorists—most of whom are white, relatively-privileged, and heterosexual—Spelman warned that 'If, like Stampp, I believe that the woman in every woman is a woman just like me, and if I also assume that there is no difference between being white and being a woman, then seeing another woman "as a woman" will involve seeing her as fundamentally like the woman I am. In other words, the womanness underneath the Black woman's skin is a white woman's, and down inside the Latino woman is an Anglo woman waiting to burst through a cultural shroud' (Spelman 1988). If feminist theorists really value equality, insisted Spelman, then they must take care not to use themselves as the measure of 'womanhood.'

In addition to multicultural feminists, so-called postmodern feminists emphasize women's differences. For example, Julia Kristeva, a postmodern feminist, notes that feminism has neglected multicultural diversity. Oliver, author of an article entitled *Julia Kristeva's Feminist Revolutions*, explains:

> Feminists in the United States are struggling with this very issue. The feminist movement has had to realize that it is a white middle class movement that has worked to exclude women whose interests and needs are somehow different. Paradoxically as soon as feminism defines 'woman' it excludes all sorts of women (Oliver 1993).

Inspired by a variety of antifoundationalists, especially the deconstructionist Jacques Derrida and the psychoanalyst Jacques Lacan, postmodern feminists resolve the old philosophical debates about the nature of reality—is it 'one' or 'many'?—decidedly in the direction of 'many-ness.' Along the way postmodern feminists such as Hélène Cixous challenge Western dualistic thinking itself (reason/emotion, mind/body, male/female) as the primary cause of human oppression, seeking to liberate human beings from a rigid conceptual grid that privileges some of society's

members over others (Cixous and Clement 1986). Insisting that not being one of society's favored members (white, rich, heterosexual, male) is an advantage, postmodern feminists stress the advantages of being rejected, unwanted, shunned, frozen out, disadvantaged, unprivileged, abandoned, dislocated, and otherwise marginalized. They claim that precisely because women have not been socially favored, women can lead the revolution to let human differences flourish.

Reconsidering their emphasis on the idea of differences as the 'essential' description of the human condition, some multicultural and postmodern feminists now wonder whether they might have gone too far in their rejection of the idea of sameness, thereby endangering feminism's ability to speak on behalf of the best interests of women in general. After all, if women have nothing more in common than two X chromosomes, it is conceivable that women share few, if any, interests in common. If women do not share common interests, then it makes little, if any, sense to insist, for example, that pornography degrades women or that female genital mutilation harms women or that confining women to the domestic sphere limits their human development. If women are essentially different, if they are more different from each other than similar to each other, then 'issues' affecting women in general do not exist.

Increasingly convinced that, like the idea of sameness, the idea of difference can become destructive, many multicultural and postmodern feminists have cast aside or at least modified their perspectives in order to join hands with so-called global feminists and ecofeminists. Although global feminists are well aware that women in the US are very different from women in Kenya or Thailand, for example, they are nevertheless inclined to believe that the women in these countries share enough common interests to become each other's political allies and moral supports. In a recent article, feminist theorist Okin (1998) reminds 'First World' feminist theorists that feminist activists, particularly those in the 'Third World,' are finding that women across the world do have a lot in common. Okin stresses that at several international meetings, women attending from countries throughout the world acknowledged 'that women (everywhere) are greatly affected by laws and customs having to do with sexuality, marriages, divorce, child custody, and family life as a whole;' that they 'are much more likely to be rendered sexually vulnerable than men and boys;' and that their 'work tends to be valued considerable less highly than…men's work' (Okin 1998). First World women, insists Okin, should work with Third World women to achieve for all women the kind of freedom and well-being some women already have. To do so is not an exercise in cultural imperialism, but a simple response to a call for assistance.

Okin's case for cooperation between First World women and Third World women is a cogent one. But feminists who are sincere about eliminating all forms of human oppression, beginning with gender oppression, will need to do more than talk about the need to help all women achieve all that is rightfully theirs as women. Privileged feminists must, as many global feminists have suggested, be prepared to give up some of their material luxuries so that women far more oppressed than they can attain all of their necessities (Mies 1993a). Stressing that there is only so much of any one material good to be distributed among the world's population, global feminists claim that feminists must take the lead in living more simply so that life on earth can continue through the next millennium and more. Bluntly put, global feminists maintain that if feminists are really serious about ending all oppression, beginning with gender oppression, they must stop being oppressors themselves.

Of course, it is not easy to stop being an oppressor, particularly if one intends to actively carry out the theories of ecofeminists. According to ecofeminists, human beings are connected not only to each other, but also to the nonhuman world: animal and even vegetative. Unfortunately, we do not always acknowledge our responsibilities to each other, let alone to the nonhuman world. As a result, we deplete the world's natural resources with our machines, pollute the environment with our toxic fumes, and stockpile weapons of total destruction. In so doing, we delude ourselves that we are controlling nature successfully to benefit ourselves. In point of fact, says ecofeminist King (1995), nature is already rebelling, and each day the human self is impoverished as yet another forest is 'detreed' and yet another animal species is murdered to extinction. The only way not to destroy ourselves, insist ecofeminists, is to strengthen our relationships to each other and the nonhuman world by refusing to engage in acts of violence, particularly warfare; refusing to eat animal flesh; and refusing to lead luxurious lifestyles. The question remains, of course, whether all feminists are required to lead lives of self-restraint so that not only gender equity but human equity is achieved; or whether, instead, ecofeminist are feminist 'saints' and 'heroes' that all feminists need not follow.

2. *Feminist Theory and Other Disciplines*

Many of the feminist theories described above center on political or social sciences and systems. Although most work has been accomplished in the social sciences, and much of the feminist movement in the past 200 years has emphasized these areas, feminist theory impacts every discipline and particularly impacts the increasingly important areas of science, medicine, technology, religion, ethics, education, as well as philosophy.

Since it is impossible to describe in a few pages the ways in which feminist theory has transformed the

entirety of human knowledge, I focus here only on the way it has affected two of philosophy's many subfields: epistemology and ontology. Feminist theory challenges the traditional ontological assumption that the more separate the 'self' is from others, the more autonomous, unique, successful, and superior that self is. Traditionally, philosophers have portrayed 'autonomous man' as a biological male—an independent self geared toward maximizing his self-interest effectively and efficiently. Such a self is always on guard against the ominous 'other' who many interfere with his life projects. Even when the 'other' convinces him to establish cooperative relations in a community, autonomous man proceeds cautiously, always seeing every individual he encounters as a potential threat and competitor for resources. Feminist theory, however, offers a different interpretation of how people come to understand their 'selves.' According to feminist ontologists, including Whitbeck (1989), people come to understand their selves through others, not against them. We are historical creatures, shaped by our relationships with other people, our parents, siblings, friends, and colleagues. Although we differ from one another, we relate to each other reasonably well because we also have much in common. This different ontology would result in a different view of the world—one in which we work together for a common good instead of competing against each other in order to secure the best only for ourselves.

Similarly, feminist theory impacts the field of epistemology. As described by Code (1998), 'the issue is less of doing philosophy "in a feminist voice", or even "in a different voice"—as some feminists propose—than of discerning whose voices have been audible, and whose muffled, in the articulation of prevailing theories …' As with the impact of feminist theory on other disciplines such as history, literature, or art, we find that the story has been told from the male point of view, using a male's voice, and using the experiences of males. At least in the Western world, philosophers and other theoreticians (mostly men until relatively recently) have claimed that the only way to achieve true knowledge is by distancing one's self from the object of one's inquiry and subjecting it to the powers of dispassionate reason. This style of knowing is in estimation of many feminists 'masculine' or 'male.' Because men tend to conceive of themselves as separate, autonomous selves, they favor ways of thinking that permit them to keep their study subjects at arm's length. Keller (1983, 1985) notes that in science, for example, objectivity has been defined in terms of the separation of the subject (the scientist) from the object of study. She believes it is more than coincidence that the men developing science conceived their methodology in terms of what was emphasized as 'masculine'—separative autonomy. Emotional connections with one's subject matter were seen as contaminating knowledge—as subjective, biased, 'feminine.' Keller insists, however, that emotional connections sometimes yield useful insights. Indeed, some of our deepest insights come from the ability to empathize and to connect with those whose behavior we study; and because women typically understand themselves as intimately and intricately connected to others, it stands to reason that women might be able to see, or first see dimensions of reality to which many men are initially oblivious (Keller 1983, 1985).

3. *First, Second, and Third Waves of Feminism*

Feminist theory, in sum, is an evolving set of ideas. Feminist theory—or 'academic feminism' (Nussbaum 1999) in America—historically has been 'allied to the practical struggle to achieve justice and equality for women. Feminist theory has been understood by theorists as not just fancy words on paper; theory is connected to proposals for social change' (Nussbaum 1999). As a discipline or philosophical approach to other disciplines, feminist theory may not be separated from the very real struggles of the feminist movement.

At this point, the beginning of the twenty-first century, the feminist movement has accomplished great equality in many areas, including politics, the workplace, and the legal system. However, there is much more work needed for true equality in all spheres of life. Feminist theory—understood here as the scholarship or thinking behind such practical feminist actions such as the reform of rape, domestic violence, or sexual harassment laws, for example—changes according to the position of women in society and the world. Just as the feminist movement is often described as having three waves—we are in the midst of the third wave now—feminist theory has evolved similarly. The first wave of US feminism is rooted in eighteenth- and nineteenth-century liberal thought as articulated by Mary Wollstonecraft, John Stuart Mill, and Harriet Taylor Mill, in particular. As described above, liberal feminism emphasized the equal rational and moral abilities of men and women, and the fact that women must be able to leave the domestic realm and enter the public realm if they are to use these equal abilities. One of the major issues of first-wave feminism in the US was securing for women the right to vote, that is, the right to participate as equal members in our political society.

Second-wave feminism centered on the fact that suffrage would never make women equal to men unless women had the same educational, occupational, and professional opportunities that men had. In addition to participating in the public realm, women must be freed from obligations in the domestic world, including sexual and reproductive obligations. Major issues in second-wave feminism included birth control and abortion, and attacks on the images portraying women as sex-objects or baby machines.

Third-wave feminism is very similar to the global and ecofeminist theories described above. In par-

ticular, third-wave feminism emphasizes that the feminist movement in the US does not represent feminism all over the world. Third-wave feminism is aware of women's differences, including race, class, and national origin; but it is also aware that women across the world have much in common, and that women can work together as equals to achieve the particular forms of freedom and well-being that mean the most to them. Like first- and second-wave feminists, third-wave feminists hunger for justice. However, their vision of justice is exceptionally large and demanding for it is based on the assumption that oppression, including gender oppression, will continue to define the human condition unless feminists and other like-minded theorists and activists take it upon themselves to convince humankind as a whole that in order for each person to have enough, no person can 'have it all.' According to ecofeminist Mies, this thought was best expressed by Kamla Bhasin, an Indian feminist. Bhasin stated that:

> sustainable development … is not compatible with the existing profit-and-growth oriented development paradigm. And this means that the standard of living of the North's (First World) affluent societies cannot be generalized. This was already clear to Mahatma Gandhi sixty years ago, who, when asked by a British journalist whether he would like India to have the same standard of living as Britain, replied: 'To have its standard of living a tiny country like Britain had to exploit half the globe. How many globes will India need to exploit to have the same standard of living?' From an ecological and feminist perspective, moreover, even if there were more globes to be exploited, it is not even desirable that this development paradigm and standard of living was generalized, because it has failed to fulfill its promises of happiness, freedom, dignity and peace, even for those who have profited from it' (see quotation in Mies 1993b).

Emerging between the lines of Bhasin's words is, it would seem, the profile of the ideal third-wave feminist, a feminist for a truly new millennium.

4. Conclusion

Feminist theory includes many ideas, and may be frustrating for someone seeking a one-sentence answer to the question 'what makes a view feminist?' However, this characteristic of feminism is not new or particular to contemporary, third-wave, global feminism or ecofeminism. In 1913, a woman wrote: 'I myself have never been able to find out precisely what feminism is … I only know that people call me a feminist whenever I express sentiments that differentiate me from a doormat, or a prostitute' (Gibbs 1992). In essence, feminist theory is a set of ideas originating with the belief that women are not subordinate to men or only valuable in relationship to men (servant, caretaker, mother, or prostitute), and that the disciplines, systems, and structures in place in our world today may be changed for the better if infused with a feminist point of view. But it is more than this. Feminist theory sets an agenda for action, the aim of which is justice and equality for women everywhere and, of course, also for the men and children to whom they are inextricably linked.

See also: Beauvoir, Simone de (1908–86); Critical Race Theory; Family Theory: Feminist–Economist Critique; Feminist Epistemology; Feminist Ethics; Feminist Legal Theory; Feminist Movements; Feminist Political Ecology; Feminist Theology; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gay/Lesbian Movements; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Geography; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender, Feminism, and Sexuality in Archaeological Studies; Multicultural Feminism: Cultural Concerns; Political Thought, History of; Queer Theory; Social Movements and Gender; Theory: Sociological

Bibliography

Alcoff L 1988 Cultural feminism versus poststructuralism: The identity crisis in feminist theory. *Signs: Journal of Women in Culture and Society* **13**(3): 406–22

Cixous H, Clement C 1986 Sorties. In: Wing B (trans.) *The Newly Born Woman*. University of Minnesota Press, Minneapolis, MN

Code L 1998 Voice and voicelessness: A modest proposal? In: Kourany J (ed.) *Philosophy in a Feminist Voice: Critiques and Reconstructions*. Princeton University Press, Princeton, NJ

Echols A 1983 The new feminism of Yin and Yang. In: Snitow A, Stansell C, Thompson S (eds.) *Powers of Desire: The Politics of Sexuality*. Monthly Review Press, New York

Engels F 1972 *The Origin of the Family, Private Property and the State*. Pathfinder Press, New York

Ferguson A 1984 Sex wars: The debate between radical and liberation feminists. *Signs: Journal of Women in Culture and Society* **10**(1): 106–35

Firestone S 1970 *The Dialectic of Sex*. Morrow, New York

Gibbs N 1992 The war against feminism. *Time* **March 9**: 50–4

Harding S, Hintikka M (eds.) 1983 *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. Reidel, Dordrecht, Germany

Jaggar A 1992 Feminist ethics. In: Becker L, Becker C (eds.) *Encyclopedia of Ethics*. Garland, New York

Keller E F 1983 *A Feeling for the Organism: The Life and Work of Barbara McClintock*. Freeman, San Francisco, CA

Keller E F 1985 *Reflections on Gender and Science*. Yale University Press, New Haven, CT

King Y 1995 Healing the wounds: Feminism, ecology, and nature/culture dualism. In: Tuana N, Tong R (eds.) *Feminism and Philosophy*. Westview Press, Boulder, CO

Kourany J 1998 Introduction: Philosophy in a feminist voice? In: Kourany J (ed.) *Philosophy in a Feminist Voice: Critiques and Reconstructions*. Princeton University Press, Princeton, NJ
Mies M 1993a The myths of catching-up development. In: Mies M, Shiva V (eds.) *Ecofeminism*. Zed, London
Mies M 1993b The need for a new vision: The subsistence perspective. In: Mies M, Shiva V (eds.) *Ecofeminism*. Zed, London
Mitchell J 1971 *Women's Estate*. Pantheon, New York
Nussbaum M C 1999 The professor of parody. *The New Republic* **February 22**: 37–46
Okin S M 1998 Feminism, women's human rights, and cultural differences. *Hypatia* **13**(2): 32–52
Oliver K 1993 Julia Kristeva's Feminist Revolutions. *Hypatia* **8**(3): 94–114
Rich A 1976 *Of Woman Born*. Norton, New York
Rich A 1994 Compulsory heterosexuality and lesbian existence. In: Jaggar A (ed.) *Living with Contradictions: Controversies in Feminist Social Ethics*. Westview Press, Boulder, CO
Ruddick S 1984 Maternal thinking. In: Trebilcot J (ed.) *Mothering: Essays in Feminist Theory*. Rowman and Allanheld, Totowa, NJ
Spelman E V 1988 *Inessential Woman: Problems of Exclusion in Feminist Thought*. Beacon, Boston
Tong R 1993 *Feminine and Feminist Ethics*. Wadsworth, Belmont, CA
Tong R 1998 *Feminist Thought: A More Comprehensive Introduction*, 2nd edn. Westview Press, Boulder, CO
Whitbeck C 1989 A different reality: Feminist ontology. In: Garry A, Pearsall M (eds.) *Women, Knowledge, and Reality: Explorations in Feminist Philosophy*. Unwin Hyman, Boston

R. Tong

Feminist Theory and Women of Color

Feminist theory has been at the forefront of new directions in political, social, and cultural theory. These developments are inherently indebted to the internal critique within feminism made by 'women of color' who have been pivotal in raising questions of 'difference' around such social axes as class, racism, ethnicity, sexuality, and the problematic of global inequities. The point is that the experience of 'being a woman' may vary greatly according to class background, color, sexual orientation, religion, and so on. The critique consists of debates that emerged through political contestation both within and outside the women's movements, drawing attention to the centrality of slavery, colonialism, imperialism, and postcoloniality in understanding contemporary gender relations and global predicaments.

1. Emergence of 'Women of Color' as a Political Subject

The terms 'women of color' and 'white women' throw into relief the political nature of discourses and practices in and through which these terms emerged as political subjects and became conceptual components of social, political, and cultural theory. They show how seemingly neutral words such as 'color' may assume specific meanings in different contexts so that, as in this instance, the color of 'whiteness' is placed into question. The political subject of 'women of color' decenters 'whiteness' as a modality of power constituted in varying forms in different times and places through historical processes of slavery, colonialism, and imperialism. It challenges the ways in which white European-descent woman came to be seen as superior to other categories of women. Although black, brown, and other categories of 'nonwhite women' have been engaged in feminist practices for centuries, the concept of 'women of color' first emerged in post-World War II Northern America and Western Europe, especially Britain. Hence, the primary focus here is upon the anglophone debate in these locations.

Political tensions in Western feminism surrounding the interrelationship between 'race' and other factors such as class and gender in the USA date back to the antislavery campaigns. During the decade of the1830s, for example, American women became increasingly active in the abolitionist movement where they learnt to champion their own right to engage in political work and where their experience of relative marginalisation compelled them to form separate women's antislavery societies. The first female antislavery society was formed in 1832 by black women in Salem, Massachusetts, followed by similar societies established by white women in other locations. Paradoxically, when the motion for women's suffrage was first introduced amid immense controversy at the Seneca Falls Anti-Slavery Convention of 1848, black women were conspicuous by their absence. This omission was incredible, not least because black women already had brought into the arena of public debate issues such as women's education which the Convention was only just beginning to address. In May of 1866, when women decided to establish an Equal Rights Association incorporating struggles for Black emancipation and woman's suffrage into a single campaign, a number of eminent speakers, including the white feminist Elizabeth Cady Stanton, argued against it. In class terms too, this was predominantly a middle-class women's movement. While a few individual black women, as for instance Sojourner Truth, were able to participate in specific events, nonetheless, the movement overall did not take on board the contradictory relationship between racism and class, or the question of sexuality in any significant way. Nor was the plight of Native American peoples, or non-European immigrants an identifiable feature of these debates and activities (Davis 1981).

As a form of self-conscious protest, feminism in Britain arose in the seventeenth century. Here, as in the USA, the early women's rights movement and later the Suffragists failed to give sufficient priority to the needs of working-class women or the issue of 'race.'

This is not to deny that there were some women as, for example, Annie Besant who was active on the anti-colonial front as well as on gender issues, and Sylvia Pankhurst who placed considerable emphasis on the conditions of the working class. Nevertheless, the effects of racism and class inequality did not become a major feminist concern at this stage. Such amnesia about issues of 'race' cannot be attributed to the lack of presence of Irish, Jewish, African, and South Asian descent people—the primary target of racisms of the period—because they were far from absent in Britain from the sixteenth to the early twentieth century. Moreover, the first usage of the term 'race' in Britain also dates back to the sixteenth century and the subsequent history of the discourse of 'race' is interlinked with slavery, colonialism, and the Holocaust of Jews, gays, gypsies, and others. Racism therefore can be said to be one of the key factors in the formation of Western societies.

The last five decades of the twentieth century witnessed the emergence of a variety of social movements: anticolonial movements for independence, Civil Rights and the Black Power movement, the Peace movement, student protest, and the Workers' movements. The Women's Liberation Movement was formed during the late 1960s and the 1970s as a part of this intellectual and political ferment. Women of color were active in all these groups. Yet, a significant number of early publications by 'Second Wave' white feminists seemed to display a certain disregard of racism as an *internal* feature of Western patriarchal relations. This neglect drew critical scrutiny from women of color. One of the first critiques was launched by the Combahee River Collective, a black, lesbian feminist organization from Boston, USA. In 1977, they produced a document that demonstrates the complexity of theorizing women's subordination when analyzing experiences based on simultaneous intersection of diverse forms of injustice: 'We know that there is such a thing as racial-sexual oppression which is neither solely racial nor solely sexual, e.g., the history of rape of Black women by white men as a weapon of political repression (Smith 1983, p. 275). The Collective saw itself 'actively committed to struggling against racial, sexual, heterosexual, and class oppression' and advocates 'the development of integrated analysis and practice based upon the fact that the major systems of oppression are interlocking (Smith 1983, p. 272). Pointing to the global dimensions of gender, the text speaks of the impact of 'political-economic systems of capitalism and imperialism' and emphasizes the question of institutional racism as well as 'racism in the white women's movement.' It repudiates those political positions which accept that the end justifies the means: '… as feminists we do not want to mess over people in the name of politics' (Smith 1983, p. 282). In taking a stance against biological determinism, this feminist discourse articulates a certain non-essentialism even as its notion of identity politics would seem to exemplify what Spivak-Chakravorty (1999) was later to call 'strategic essentialism.' The idea of 'simultaneously interlocking oppressions' is one of the most productive insights of this period.

While the composition of 'women of color' in terms of ethnicity may vary from place to place, the generalized category itself remains wide-ranging and global. According to the anthology *Home Girls*, for example, 'the terms Third World Women and women of color … designate Native American, Asian American, Latina and Afro-American women in the USA and the indigenous peoples of Third World countries wherever they may live' (Smith 1983, p. 1iv). Taken together, women of color comprise the great majority of women of the world and a very high proportion live in conditions of material disadvantage. Hence, the centrality of a concern with social inequality in women of color feminisms.

Writings by women of color have a history of being treated as 'personal testaments' rather than 'theory.' The question of what counts as theory and whose intellectual work gets treated as being theoretical has been the subject of much debate (cf. Collins 1990). A related point concerns the *theoretical* subject of 'women of color.' In her analysis of the book *The Bridge Called My Back*—a seminal American text of the early 1980s—Norma Alarcon argues that the theoretical subject of *Bridge* is a figure of multiplicity, representing consciousness as a 'site of multiple voicings' seen 'not as necessarily originating with the subject but as discourses that traverse consciousness and which the subject must struggle with constantly.' It is associated with subjectivity that is deeply marked by 'psychic and material violence' and it demands a thorough 'reconfiguration of feminist theory' (Alarcon in Anzaldúa 1990, pp. 359–65). In other words, the theoretical subject of 'women of color,' in its very constitution, decenters the Eurocentric discourse of the rationalist, masculinist subject of modernist philosophy. It challenges essentialist readings of skin tone or physical appearance as inherent 'difference' and disrupts any notion of 'woman' as a unitary category.

2. *Black British Feminism*

A similar yet different process occurred in1970s Britain when 'women of color' came to be figured as 'black.' This was a consequence of coalition politics among African, Caribbean, and South Asian descent groups who borrowed the Black Power vocabulary but re-signified 'black' to embrace all 'nonwhite' people. The concept of 'black' was designed to substitute the colonial term 'colored people' and interrogate 'chromatism'—the hierarchy based on differences of skin tone—that could potentially threaten unity between these groups. Black British feminism was forged

through the work of local women's organizations active around such issues as wages and conditions of work, immigration law, fascist violence, reproductive rights, and domestic violence. By 1978, local groups had combined to form a national body called the Organisation of Women of Asian and African Descent (OWAAD). This network held annual conferences, published a newsletter, and served as an active conduit for information, intellectual conversations, and political mobilization. The ensuing dialog entailed sustained analysis of racism, class, and gender with much debate as to the best means of confronting their outcomes while remaining alive to cultural specificities:

> Our group organises on the basis of Afro-Asian unity, and although that principle is maintained, we don't deal with it by avoiding the problems this might present, but by having on-going discussions…Obviously, we have to take into account our cultural differences, and that has affected the way we are able to organise (OWAAD cited in Mirza 1997, p. 43).

Like its 'women of color' counterpart in the USA, emerging black feminism in Britain came to prefigure later theories of 'difference.' The internal conflicts within OWAAD, especially around homophobia, proved salutary so that, even as British 'black feminism' assumed a distinctive political identity separate from 'white feminism,' engaging the latter in heated debate, it did not become oblivious to the contradictions of its internal heterogeneity. The careful reflection and analysis that went into the deliberations resulted in a very significant corpus of scholarship.

Hazel Carby's essay *White woman listen* is symptomatic of conflicts and tensions that characterized feminist debate during the 1980s. Carby develops a trenchant critique of key feminist concepts—'patriarchy,' 'the family,' and 'reproduction'—drawing out their interconnection with racism and colonial history. Taking issue with the Euro-centric approach of certain white feminist discourses, she stresses that racism is not simply about exclusion but equally about the manner of inclusion itself. As Chandra Talpade Mohanty has argued, Western feminisms during this period tended to reproduce the very categories through which the 'West' constructs and represents itself as superior to others. In 1984, a group of British feminists of color collectively edited a special issue of the journal *Feminist Review* (Issue 17). This text, together with the edited collection *Charting the Journey* (Grewal et al. 1988), represents intellectual intervention through work that was the outcome of collective thinking. Some of the main points addressed in these publications may be summarized as follows:

(a) As a consequence of capitalist colonial relations, labor migrants from former colonies are constituted as *a racialized class fraction* in the metropolitan society.

(b) The institutionalization of racism and gender inequality within global sociopolitical and cultural relations means that 'women of color' may face some similar injustices all over the world, but they will also share experiences with the metropolitan white working class.

(c) Women of color in Britain are subjected to 'double patriarchy': a generalized, all pervasive, racialized form of British patriarchy; and, specific configurations of patriarchal norms current among particular groups of people of color.

(d) Lesbians of color have to contend with: heterosexism in British society; homophobia among people of color; and, eurocentricity and racism among white lesbians and gay men. These aspects must be addressed as part of feminist theoretical and political practice.

3. *Postmodernism and the Theory of Difference*

Feminist theories of the 1970s and 1980 were informed by conceptual repertoires drawn largely from 'modernist' theoretical and philosophical traditions of European Enlightenment such as liberalism and Marxism. The 'postmodernist' critique of these perspectives, including their claims to universal applicability, had precursors within anticolonial, antiracist, and feminist critical practice. Postmodern theoretical approaches found sporadic expression in Anglophone feminist works from the late 1970s. But, during the 1990s they became quite a significant influence, in particular their poststructuralist variant. Some of the main points emerging from the encounter between women of color feminism and postmodern/poststructuralist theories are that:

(a) It is no longer tenable to conceptualize 'white feminism' and 'feminism of color' as if they are mutually exclusive entities, each carrying some unchanging transhistorical 'essence.'

(b) Contrary to analysis where process is reified and understood as personified in the bodies of individuals, these two distinctive yet overlapping sets of feminisms are now viewed as representing historically contingent relationships, contesting fields of discourses, or 'reading strategies.'

(c) The concept of 'agency' has been reconfigured through poststructuralist appropriations of psychoanalysis in order to take account of psychic and emotional life. The concept of subjectivity now covers both conscious agency and unconscious psychic dynamics, and it represents a site of multiplicity where effects of racism and ethnicity intersect with other structural and experiential dimensions.

(d) Reassessment of the notion of experiential authenticity has highlighted the limitations of 'identity politics.' Experience is now understood primarily as a 'signifying practice' referring to the process of making sense of the world symbolically and narratively.

(e) Poststructuralist insistence that meaning is relational; that identity and subjectivity are not 'products' but on-going processes; that power can be

both coercive and productive; that subordination can occur through modes of inclusion no less than exclusion means that poststructural paradigm has much in common with theoretical interventions made by women of color over the years.

Overall, critical but productive conversations with theories of poststructuralism have resulted in new theoretical frameworks for refashioning the analysis of 'difference' (cf. Hooks 1990, Grewal and Kaplan 1994).

4. *Postcoloniality, Transnationality, and Diasporic Space*

One outstanding feature of the work of feminists of color during the 1990s is a concern with the potential of combining strengths of modern theory with postmodern insights. This interest in interdisciplinary 'creolized theory' would seem to stem from the following preoccupations:

(a) an interest in analyzing the effects of racism without recourse to the discourse of 'race';

(b) the need for theorizing ethnicity and cultural difference outside the imperatives of racism;

(c) the importance of understanding political identities as contingent forms of resistance;

(d) the necessity of distinguishing between political identity as conscious agency and identity conceptualised in terms of subjectivity;

(e) the importance of analyzing locality and globality as relational categories;

(f) sustaining the previous emphasis on colonial histories and 'material conditions' at the same time as analyzing cultural issues; and

(g) the need to attend to the problematic of the category 'mixed race.'

This work has taken several forms. Some developments, especially in the field of literary studies, are known as 'postcolonial' theory. When used in the sense of a reading strategy, postcolonial theory emphasizes the point that both the 'metropolis' and the'colony' were altered deeply by the colonial process and that these articulating histories have a mutually constitutive role in the present. Postcolonial feminist studies foreground 'Othering' processes underlying colonial and postcolonial discourses of gender. Frequently, such work uses poststructuralist frameworks, especially Derridean deconstruction and Foucauldian discourse analysis. Some scholars have attempted to combine poststructualist approaches with neo-Marxist or psychoanalytic theories. Other feminists of color have transformed 'border theory' (Anzaldua 1990, Mani 1998, Spivak 1999)

A related development is associated with valorization of the term diaspora. The concept of diaspora increasingly is used in analyzing the mobility of peoples, commodities, capital, and cultures in the context of globalizaton and transnationalism. The concept is designed to analyze configurations of power—both productive and coercive—in 'local' and 'global' encounters in specific spaces and historical moments. Brah (1996) addresses the concept of diaspora alongside that of Anzaldua's theorization of 'border' and the widely debated feminist concept of 'politics of home.' The intersection of these three terms is understood through the concept of 'diaspora space' which covers the entanglements of geneaologies of dispersal with those of 'staying put' so that the 'native' becomes as much a 'diasporian' as the diasporic subject becomes the native. The term 'homing desire' is used to think through the question of home and belonging; and, both power and time are viewed as multidimensional processes. Importantly, diaspora space embraces the intersection of 'difference' in its variable forms, placing emphasis upon emotional and psychic dynamics as much as socioeconomic, political, and cultural differences. Difference is thus conceptualized as social relation, experience, subjectivity, and identity. Home and belonging is also a theme of emerging literature on 'mixed-race' identities which interrogates the concept of 'race' as an essentialist discourse with racist effects (Tizzard and Phoenix 1993, Zack 1993, Ifekwunige 1999). Accordingly, the idea that you are mixed-race if you have black and white parents is jettisoned. Instead the analytical focus is upon varying and variable subjectivites, identities, and the specific meanings attached to 'differences.'

Theoretical perspectives developed by women of color are central features of innovations in feminist theory. They herald a new phase in feminist alliances and coalitions. They promise new ways of formulating democratic futures (Alexander and Mohanty-Talpade 1997).

See also: Critical Race Theory; Cultural Assimilation; Cultural Critique: Anthropological; Cultural Relativism, Anthropology of; Evolutionism, Including Social Darwinism; Feminist Epistemology; Feminist Ethics; Feminist Theology; Feminist Theory; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Multicultural Feminism: Cultural Concerns; Race and Gender Intersections; Theory: Sociological

Bibliography

Alexander J, Mohanty-Talpade C (eds.) 1997 *Feminist Genealogies, Colonial legacies, Democratic Futures*. Routledge, London and New York

Anzaldúa G (ed.) 1990 *Making Face, Making Soul*. Aunt Lute Foundation Books, San Francisco
Bhavnani K K (ed.) 2000 *A Reader*. Oxford University Press, Oxford, UK
Brah A 1996 *Cartographies of Diaspora, Contesting Identities*. Routledge, London and New York
Collins P H 1990 *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Unwin Hyman, Boston
Davis A Y 1981 *Women, Race and Class*. Women's Press, London
Essed P 1991 *Understanding Everyday Racism: An Interdisciplinary Theory*. Sage, Newbury Park, CA
Grewal I, Kaplan C (eds.) 1994 *Scattered Hegemonies: Postmodernity and Transnational Feminist Practices*. University of Minnesota Press, Minneapolis, MN
Grewal S et al. (eds.) 1988 *Charting the Journey: Writings by Black and Third World Women*. Sheba, London
Hooks B 1990 *Yearning: Race, Gender and Cultural Politics*. South End Press, Boston
Ifekwunigwe J O 1999 *Scattered Belongings: Cultural Paradoxes of 'Race', Nation, and Gender*. Routledge, London and New York
Jarrett-Macauley D (ed.) 1996 *Reconstructing Womanhood, Reconstructing Feminism: Writings on Black Women*. Routledge, London
Lewis G 2000 *Race, Gender, Social Welfare: Encounters in a Postcolonial Society*. Polity, Cambridge, MA
Mani L 1998 *Contentious Traditions: The Debate on Sati in Colonial India*. University of California Press, Berkeley
Mirza H S (ed.) 1997 *Black British Feminism: A Reader*. Routledge, London and New York
Smith B (ed.) 1983 *Home Girls: A Black Feminist Anthology*. Kitchen Table Press, New York
Spivak G Chakravorty 1999 *A Critique of Postcolonial Reason: Towards History of the Vanishing Present*. Harvard University Press, Cambridge, MA
Sudbury J 1998 *Other Kinds of Dreams: Black Women's Organizations and the Politics of Transformation*. Routledge, New York and London
Tizzard B, Pheonix A 1993 *Black, White or Mixed Race?: Race and Racism in the Lives of Young People of Mixed Parentage*. Routledge, London and New York
Trinh T Min-Ha 1989 *Woman, Native, Other: Writing Postcoloniality and Feminism*. Indiana University Press, Bloomington and Minneapolis
Zack N 1993 *Race and Mixed Race*. Temple University Press, Philadelphia.

A. Brah

Feminist Theory: Ecofeminist and Cultural Feminist

Just as there are a variety of feminisms and feminist theories, there are a variety of ecofeminisms and ecofeminist theories. What distinguishes them as 'ecofeminist' is a twofold commitment to analyzing and undoing historical connections between the unjustified dominations of women (and other human subordinated groups) and nonhuman nature. This article explores the varieties of ecofeminist theories. It reveals similarities and differences among ecofeminist theories, and suggests directions for future ecofeminist theory and research.

1. *The Terms 'Ecofeminist' and 'Cultural Feminist'*

The term '*ecofeminisme*' was coined in 1974 by French feminist Francoise d'Eaubonne (1974) to call attention to women's potential to bring about an ecological revolution. Ecofeminism has emerged around the globe as a powerful political movement and theory. Despite important differences among ecofeminist theories, they share a common concern for making visible historical associations of 'women' with 'nature,' and, at least in Western societies, the attendant devaluation of both as inferior to men and male-gender identified 'culture' (Merchant 1980).

2. *Three Types of Ecofeminist Positions*

Historically, the first strand of ecofeminism was *cultural feminism* (sometimes referred to as 'radical' or 'spiritual' feminism). Developed in the early 1970s, cultural feminism reclaims women-nature connections as liberating and empowering expressions of women's capabilities to care for nature. Some cultural ecofeminists argue that women's reproductive capacities provide a biological tie with nature, making women 'closer to nature' than men. Other cultural ecofeminists claim that 'women's closeness to nature' is embedded in deep social and psychological structures, making women's ways of knowing and moral reasoning better suited to solving environmental problems. Some argue for a resurrection of pre-patriarchal religions and spiritual practices (e.g., Goddess worship or pagan rituals) which honored women's bodies and procreative powers. But all cultural ecofeminists agree that women have a special and useful relationship to the physical world (bodies, nature) which can help end the unjustified dominations of both women and nature.

Cultural ecofeminism has been criticized by *social ecofeminists* for making essentialist, universalist, and ahistorical claims about both women and nature. Social ecofeminists claim that there is no essential (biological, natural, innate) nature of 'women' (*contra* cultural ecofeminism's essentialism), no homogenous 'women's experience' or 'women's way of knowing' (*contra* cultural ecofeminism's universalism), and no ahistorical concept of 'women' (*contra* cultural ecofeminism's ahistoricism). Rather, women's identities are socially constructed, historically fashioned, and materially reinforced through the interplay of a diversity of race/ethnicity, class, sexual orientation, age,

ability, marital status, and geographic factors ('social constructivism').

A middle ground is suggested by *materialist* (or *socialist*) *ecofeminism*. Socialist/materialist ecofeminism claims that women–nature connections are both socially constructed *and* biologically predisposed: women's biology (bodies, reproductive capacities) has played a key role historically in the oppression of women and the identification of 'women' with 'nature,' but women's biology is *not* destiny. Rather, it is the social, material, and political relationships between women and nature which are important.

What all three strands of ecofeminism add to feminism is the insistence on the relevance of ecology and environmentalism to feminism. While contemporary feminism has made progress in making visible the interconnections among various forms of human oppression (e.g., sexism, racism, classism, heterosexism, ageism, ableism, ethnocentrism), all ecofeminists insist that feminist analyses be extended to show how the unjustified domination of nonhuman nature (or 'naturism') is connected to sexism and other human 'isms of domination.'

3. *The Intellectual Context of Ecofeminist Theory and Practice*

Ecofeminism has always been a grassroots political movement motivated by pressing pragmatic concerns: women's and environmental health; deforestation, desertification, and soil erosion; the treatment of animals; science, development and technology; peace, and antinuclear activism. The Chipko Movement in India, Mothers of East Los Angeles, and the Women's Pentagon Actions are examples of the sort of political organizing actions characteristic of ecofeminism (Warren 1997b).

Ecofeminist theory grew out of such grassroots activism. It provides perspectives about gender from a combined environmental *and* feminist perspective, by showing ways human relationships to both 'nature' and 'culture' are gendered, raced, and classed, and how the dominations of women and nature are embedded in patriarchal (and other) structures and practices. Ecofeminist theory offers ecological notions of the self which challenge dominant Western models of rationality, knowledge, and ethics, and which establish connections between biodiversity and cultural diversity.

Ecofeminist scholarship is interdisciplinary. In philosophy, it focuses on conceptual issues by challenging patriarchal world views characterized by value-hierarchical and dualistic thinking, coupled with a 'logic of domination' which 'justifies' relationships of dominance and subordination (Warren 1996, p. 21). Ecofeminist philosophy challenges the 'culture/nature' dualism, as well as related 'reason/emotion,' 'mind/body,' 'mental/physical,' 'objective/subjective,' and 'human/nonhuman nature' dualisms. That is because these value dualisms historically have functioned to sanction as superior the male-identified realm of culture (mind, reason, objectivity) and as inferior the female-identified realm of nature (the physical, the body, emotion, subjectivity). Ecofeminist theorists argue that such hierarchical value dualisms are either false or conceptually incoherent. They incorrectly presuppose a clear separation of culture from nature (mind from body, reason from emotion, human from nonhuman), an ecologically unsound view of humans as discontinuous with nature, and an untenable ethical view which justifies the control, conquest, and mastery of the nonhuman environment in order to serve human needs.

Ecofeminists have offered different historical accounts of the connections between patriarchy, androcentric (male-centered) thinking, and the destruction of nonhuman nature. For example, environmental historian Carolyn Merchant argues in *The Death of Nature* that the change in gender-laden metaphors about nature in the history of science contributed to 'the death of nature.' Merchant argues that the classical Greek metaphor of nature as 'organic,' kindly, benevolent, nurturing female metaphor was replaced in the sixteenth and seventeenth centuries by the metaphor of nature as 'mechanistic,' disorderly, chaotic, instrumental, passive, and inert. While both metaphors feminize 'nature,' only the mechanistic metaphor permitted the removal of ethical constraints against the 'rape of Mother nature,' thereby permitting the 'mining of her womb' for the sole benefit of man.

By contrast, Riane Eisler argues that patriarchal patterns of domination began about 4500 BC E with the invasion of Indo-European societies by nomadic tribes from Eurasia. Ecofeminist philosopher Val Plumwood locates the historical roots of naturism in patriarchal dualisms which originated in classical Greek philosophy and the rationalist tradition (whereby rationality is taken to be both the hallmark of humanness and what elevates humans over nonhuman animals and nature).

Ecofeminist theorists often focus on ways sexist-naturist language maintain, sanction, and perpetuates harmful value hierarchical dualisms. Women are described in animal (nature) terms as pets, dogs, cats, pigs, cows, sows, foxes, chicks, serpents, bitches, beavers, mares, old mares, old bats, old hens, mother hens, pussies, pussycats, cheetahs, bird-brains, hare-brains, elephants, and whales. Animalizing or naturalizing women *in a patriarchal culture* where animals are seen as inferior to humans (men) thereby reinforces and authorizes women's inferior status. Similarly, language which feminizes nature *in a patriarchal culture* where women are seen as subordinate or inferior reinforces and authorizes the domination of nature. 'Mother Nature' is raped, mastered, conquered, mined, controlled, poked, prodded, and pried into; her secrets are penetrated and her womb is put

into service by the 'man of science.' Virgin timber is felled, cut down; fertile (not virile) soil is tilled and land that lies fallow is barren (not impotent) and useless, like a woman who cannot conceive a child. The exploitation of nature and animals is justified by language which is feminizing them; the exploitation of women is justified by language which is naturalizing women (Warren 1997b).

Ecofeminist theory has spawned a new area called 'ecofeminist literary criticism' (Murphy 1995). Focusing on contemporary writing by women (particularly women of color), ecofeminist literary criticism uses ecofeminist insights to develop a body of literature which honors biotic and human (especially gender) differences, while challenging mainstream notions of humans, culture, nature, and wilderness in the traditional Western literary canon (including traditional nature writing by men). Paula Gunn Allen, Margaret Atwood, Rachel Bagby, Annie Dillard, Sally Gearhart, Susan Griffin, Linda Hogan, Ursula Le Guin, Marge Piercy, Leslie Marmon Silko, Alice Walker, and Terry Tempest Williams are just a few of the authors who explore these connections through literature.

One of the first ecofeminist literary texts was cultural ecofeminist Susan Griffin's *Woman and Nature* (1978). In the prologue to her book Griffin writes,

> He says that woman speaks with nature … He says he is not part of this world, that he was set on this world as a stranger. He sets himself apart from woman and nature.
>
> And so it is Goldilocks who goes to the home of the three bears, Little Red Riding Hood who converses with the wolf, Dorothy who befriends a lion, Snow White who talks to the birds, Cinderella with mice as her allies, the Mermaid who is half fish, Thumbelina courted by a mole. (And when we hear in the Navajo chant of the mountain that a grown man sits and smokes with bears and follows directions given to him by squirrels, we are surprised. We had thought only little girls spoke with animals.)
>
> We are the bird's eggs. Bird's eggs, flowers, butterflies, rabbits, cows, sheep; we are caterpillars; we are leaves of ivy and sprigs of wallflower. We are women. We rise from the wave. We are gazelle and doe, elephant and whale, lilies and roses and peach, we are air, we are flame, we are oyster and pearl, we are girls. We are women and nature. And he says he cannot hear us speak.
>
> But we hear (Griffin 1978, p. 1).

Griffin's writing has been controversial because of its essentialist and universalist underpinnings. Still, it is one of the earliest and most powerful pieces of ecofeminist literature to reveal the deeply-seated symbolic and philosophical associations between women and nature in the Western intellectual tradition.

An ecofeminist concern about language, images, stories, rituals, and symbol-systems about 'women' and 'nature' is at the heart of cultural ecofeminism's spiritual claims. Cultural ecofeminists defend earth-based, feminist spiritual symbols (such as Gaia and the Goddess) as a necessary part of ecofeminism. Charlene Spretnak writes:

> Ecofeminism grew out of radical, or cultural, feminism (rather than from liberal feminist or socialist feminism) … In the mid-1970s many radical/cultural feminists experienced the exhilarating discovery, through historic and archaeological sources, of a religion that honored the female and seemed to have as its "good book" nature itself … We would not have been interested in "Yahweh in a skirt," a distant, detached, domineering godhead who happened to be female. What was cosmologically wholesome and healing was the discovery of the Divine as immanent in and around us (Spretnak 1990, p. 5).

Similarly, cultural ecofeminist Starhawk defends a nonpatriarchal, earth-based spirituality as crucial to ecofeminist politics:

> Earth-based spirituality is rooted in three basic concepts that I call immanence, interconnection, and community. The first—immanence—names our primary understanding that the Earth is alive, part of a living cosmos. What that means is that spirit, sacred, Goddess, God—whatever you want to call it—is not found outside the world somewhere—it's in the world: it *is* the world, and it is us … When we understand that the Earth itself embodies spirit and that the cosmos is alive, the we also understand that everything is interconnected … The kind of spirituality and the kinds of politics we're called upon to practice are rooted in community … The goal is the creation of a community that becomes a place in which we can be empowered and in which we can be connected to the Earth and take action together to heal the Earth (Starhawk 1990, pp. 73–4).

On the other hand, ecofeminist theologian Elizabeth Dodson Gray was among the first ecofeminists to examine the roles that sexist-naturist imagery plays in the patriarchal heritage of the Judeo-Christian tradition *without* entirely rejecting those traditions. In her book, *Green Paradise Lost*, Gray claims that a destructive hierarchy of beings is at the heart of biblical accounts of creation:

> In this biblical view of the nature of things woman comes after and also below man. Woman was created (according to this chronologically earliest account of the creation of the world in Gen. 2) out of man's body (rather than from a woman's body as happens naturally) … Then come children, so derivative that they are not even in the Creation story … Then come animals, who do not have the unique human spirit at all … Thus animals are below. Further down are plants, which do not even move about. Below them is the ground of nature itself—the hills and mountains, streams and valleys—which is the bottom of everything just as the heavens, the moon and the stars are close to God at the top of everything (Gray 1979, p. 5).

In this hierarchical order of being, the higher up one goes, the closer one is to all that is spiritual and superior:

> What is clearly articulated here is a hierarchical order *of being* in which the lower orders—whether female or child or animal or plant—can be treated, mistreated, violated, sold, sacrificed or killed at the convenience of the higher states of spiritual begins found in males and in God … Nature … is

> obviously a prize candidate for the most ruthless "mastering" of all (Gray 1979, pp. 5–6).

Much of ecofeminist theory involves ethical considerations. Minimally, ecofeminist ethics attempts to develop theories and practices concerning humans, nonhuman animals, and the natural environment which are not male-biased and which provide a guide to action in the prefeminist present. Typically, ecofeminist ethics reconceives humans as both ecological selves and moral agents, and reframes human–nonhuman nature relationships in ways which centralize values of care, appropriate reciprocity, kinship, integrity, and well-being—values typically underplayed in traditional Western ethics.

Ecofeminism also challenges the conception and practice of mainstream Western science and epistemology. As philosopher Lori Gruen states,

> Following the arguments made by feminist philosophers of science, Marxists, cultural critics, and others, ecofeminists believe that facts are theory-laden, theories are value-laden, and values are molded by historical and philosophical ideologies, social norms, and individual processes of categorization (Gruen 1994, p. 124).

According to Gruen, ecofeminist theory always grows out of and examines the social contexts in which moral and epistemic claims are generated. Recognition of the interdependence of science and society, facts and values, reason and emotion 'is the first step towards any legitimate knowledge' (Gruen 1994, p. 134).

Biologist Donna Haraway concurs. Her arguments for 'situated knowledges' and 'cyborg feminism' challenge scientific assumptions of impartiality, objectivity, fact, and truth by revealing their embeddedness in structures of power and dominance. Haraway locates the promise of ecofeminism in disrupting the nature–culture dualism by seeing nature as both active, unpredictable, unstable, ungendered subject with agency (not simply some 'resource to be mapped and appropriated') and also as constructed artifact (Haraway 1991, p. 199).

Ecofeminism's deep roots in political activism have led to the emergence of ecofeminist political theory based on coalition-building, direct action, and new conceptions of democracy. Political scientist Noël Sturgeon argues for a conception of ecofeminism as an 'oppositional political discourse and set of practices imbedded in particular historical, material, and political contexts' (Sturgeon 1997, p. 3). Sturgeon characterizes ecofeminism as a social movement through which change is produced by numerous kinds of 'action,'

> including that of the deployment of symbolic resources, shifts in identity construction, and the production of both popular and scholarly knowledge—as well as direct action, civil disobedience, strikes, boycotts, demonstrations, lobbying, and other more traditionally recognized forms of political action (Sturgeon 1997, p. 4).

Sturgeon argues that, as a social movement, ecofeminism is and must be involved in theorizing about concepts (e.g., of nature, women, race).

The power of ecological feminist theory lies in its ability to provide insights into the interconnections among the dominations of women and other human 'others' and the domination of nonhuman nature. The future of ecofeminist theory and research is to continue to show why, where, and how to accomplish the dismantling of interlocking human systems of unjustified domination.

See also: Feminist Ethics; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Environment; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology

Bibliography

d'Eaubonne F 1974 *Le Feminisme ou La Mort*. Pierre Horay, Paris

Eisler R 1987 *The Chalice and the Blade*, 1st edn. Harper and Row, Cambridge, MA

Gray E D 1979 *Green Paradise Lost*. Roundtable Press, Wellesley, MA

Griffin S 1978 *Woman and Nature: The Roaring Inside Her*, 1st edn. Harper and Row, New York

Gruen L 1994 Toward an ecofeminist moral epistemology. In: Warren K J (ed.) *Ecological Feminism*. Routledge, London, pp. 120–38

Haraway D J 1991 *Simians, Cyborgs and Women: The Reinvention of Nature*. Routledge, New York

Merchant C 1980 *The Death of Nature: Women, Ecology, and the Scientific Revolution*, 1st edn. Harper and Row, San Francisco

Murphy P D 1995 *Literature, Nature, and Other: Ecofeminist Critiques*. State University of New York Press, Albany, NY

Plumwood V 1993 *Feminism and The Mastery of Nature*. Routledge, London

Spretnak C 1990 Ecofeminism: our roots and flowering. In: Diamond I, Orenstein G F (eds.) *Reweaving the World: The Emergence of Ecofeminism*. Sierra Club Books, San Francisco, pp. 3–14

Starhawk 1990 Power, authority, and mystery. In: Diamond I, Orenstein G F (eds.) *Reweaving the World: The Emergence of Ecofeminism*. Sierra Club, San Francisco, pp. 73–86

Sturgeon N 1997 *Ecofeminist Natures: Race, Gender, Feminist Theory and Political Action*. Routledge, New York

Warren K J (ed.) 1997a *Ecofeminism: Women, Culture, Nature*. Indiana University Press, Bloomington, IN

Warren K J (ed.) 1997b Taking empirical data seriously. *Ecofeminism: Women, Culture, Nature*. Indiana University Press, Bloomington, IN, pp. 3–20

Warren K J 1996 The power and the promise of ecological feminism. In: Warren K J (ed.) *Ecofeminist Philosophy*. Rowman and Littlefield, Lantham, MD, pp. 19–41

K. J. Warren

Feminist Theory: Liberal

Liberal feminist theory is liberal political theory and feminist political theory—meaning that it advances the theoretical tradition of liberalism while it contributes to a political theory based on the foundation that men and women are of equal human worth. It is also a critical practice of assessing the laws and social practices sustained by liberal theory which might inhibit the realization of men's and women's equality.

Classical liberalism describes a limited government intended to secure the individual freedom and the political equality of (some) men (Locke 1998). Rational, free, and equal men contract in the public sphere to organize limited government that will protect their autonomy. Contemporary welfare liberals respect the rational autonomy and freedom of individuals but interpret equality more broadly and accept a role for the state in fostering political equality.

Because liberal feminists respond to theoretical traditions, laws, and social practices that have changed over time, liberal feminism is not defined quickly or narrowly. Some liberal feminists, particularly early liberal feminists, work within the framework of classical liberalism and seek to extend equality to women based on the assertion that women like men are rational individuals. More recently others challenge the public/private dichotomy that is sustained by the theory of the state as a contractual arrangement among autonomous rational individuals (Pateman 1988). Thus, some liberal feminists work within liberal theory to change its implications for women while others stretch the bounds of liberal theory to adapt it to feminist demands (Phillips 1993). In addition, due to their concern about inclusion, many liberal feminists engage in criticism of some aspects of liberal feminist thinking itself. Thus, the boundaries between liberal feminist inquiry and the broader feminist project of making theory, law, and social practice and the criticism thereof more inclusive social processes are not tidily delineated.

1. *The Historical Tradition of Liberal Feminism*

Liberal feminists of the eighteenth, nineteenth, and early twentieth centuries asserted women's ability and right to participate in public life at least through the suffrage, examined the institution of marriage and the legal framework of divorce and property laws, and challenged the practices of denying women access to the same quality education as men and to the professions. Liberal feminists sought to end the treatment of women as legal dependents on their husbands or fathers. Through examination of laws and practices, liberal feminists including Mary Astell (1666–1731), Mary Wollstonecraft (1759–99), Harriet Taylor (1807–58), John Stuart Mill (1806–73), Elizabeth Cady Stanton (1815–1902), and Virginia Woolf (1882–1941) drew on the liberal tradition's value of equality and individual freedom to argue that, just as social status at birth was no longer a legitimate basis on which to discriminate among men as liberals argued, so also sex at birth was no longer a legitimate basis on which to discriminate against women.

Among liberal feminists, justifications varied. To defend their claim, some feminists argued that men and women were of equal human worth; some argued that some women were superior to some men and thus all women should be given opportunity to demonstrate their worth; and some argued that women's virtues were superior to men's and thus warranted influence on public economic and political activity. Mill drew on each of these when it was strategically prudent. However, in general, the liberal argument took the form of asserting that individual development required the ability to educate and expand one's faculties such that to deny women access to education and the professions was to deny the development of their full human potential. Further, liberal feminists argued that society as a whole would benefit from the public contributions of women.

While liberal feminists asserted that society in general would benefit from educated women's contributions, generally these early liberal feminists did not pay adequate attention to the circumstances of working class women. The male model of life experience to which they wanted women to gain access was that of a white professional class man. However, more activist liberal feminists such as Maria Stewart (1803–79) and Susan B. Anthony (1820–1906) did pay attention to the circumstances of Black women and working class women.

In the latter twentieth century the arguments for women's equal access to education and the professions and for treatment as autonomous individuals (rather than individuals assumed to be dependent on a father or husband) by the law were heard from feminists again. However, although some feminists focussed their attention on the plight of the white, middle class, heterosexual, married Western woman (Friedan 1974), in general the theoretical and practical attention of liberal feminists switched from equality to inclusion. While demanding that women get the same opportunities as advantaged men, feminists also interrogated the very desirability of a western male model of life experience where professional and public activities were seen as more important and fulfilling than community and family activities. Instead, these

feminists sought to change the gendered practices of men and women at home and in public such that domestic caregiving was given social, legal, and economic recognition commensurate with its social value (Okin 1989; Steinem 1983) (see *Gender and Feminist Studies*).

Further, women of color, lesbians, and women from around the world encouraged western feminists to think beyond the bounds of white, middle class, Western experience (Lorde 1979, Mohanty 1984). Consequently, during the twentieth century, the political agenda of liberal feminism moved beyond a focus on married women's property rights, divorce laws, and the suffrage to include labor laws, social issues including child-care, welfare, health-care and education, access to safe legal abortion, international development, the distribution of economic resources, and human rights (NOW 1967, Sen and Grown 1987, Bunch 1990).

Even as the liberal feminist agenda expanded, theorists and activists wondered if liberal feminism was theoretically adequate for challenging norms of colonialism, global inequality, racism, patronage politics, and other forms of hierarchy that complicate analysis of gender inequality. For example, at the dawn of the twenty-first century, women's human rights activism is a global movement of national and international feminists acting in locally specific ways and sharing their experience globally. This movement ostensibly promotes the liberal values of freedom and equality, but the scope of freedom and equality extends well beyond classical liberal bounds to include global provision of basic needs, education, and healthcare and access to development resources (Tinker 1990, Kabeer and Subrahmanian 1996). Thus, the social interests of contemporary liberal feminists extend far beyond those of early liberal feminists. Most early twenty-first century liberal feminists seek to challenge fundamental values, practices, and norms within and across societies (Charlesworth 1994, Afkhami 1995, Okin 1999).

2. *Feminism within Liberal Theory*

In the tradition of liberal feminist theory, theorists have not only promoted liberal values of freedom and equality for women but also re-examined those same values in light of women's experiences. While working within the liberal tradition, liberal feminists work to change liberal theory by examining its accepted arguments and underlying assumptions. Thus, liberal feminists advance feminism and liberalism. For example, Carole Pateman (1988) shows that liberal theory's use of contract doctrine assumes freedom and equality that do not exist among the contracting parties. In a private contract between the sexes, women lose their ability to contract as autonomous individuals with others in public. Thus, women are not equal parties to the original political contract on which classical liberal social contract theorists' base liberal democracy. For women the 'sexual contract' is not an exercise of freedom. Rather than wanting to rewrite the sexual contract and the political contract as some feminist theorists have argued, and as some political theorists have done by sanitizing their texts of sex-specific language, Pateman criticizes the use of classical liberal theorists' fundamental tool, consent theory or social contract theory. Pateman undermines the oppressive roots of contract theory and reveals its use in the validation and regulation of slavery, employment contracts, marriage and divorce, prostitution, and surrogacy. She reveals that the liberal premise of natural individual freedom and equality is a fallacy and as such is an inappropriate basis for political theory.

MacKinnon (1993) argues that liberal notions of equality are based on an historical tradition dating back to Aristotle according to whom treating people equally means treating likes alike, and unlikes differently. MacKinnon argues that contemporary liberal theory sustains this perspective. Consequently, although equality is an important foundation of liberal theory, as defined it is not an adequate basis for challenging exploitative inequality where the inequality is the result of a difference that is perceived to be natural or ordained by God. MacKinnon proposes an alternative conception of equality as the absence of hierarchy. Such a definition enables the 'unlikes' to demand equality.

In addition to challenging the underpinnings of liberalism, liberal feminists also challenge the scope of liberal inquiry. For example, Okin (1989) challenges the liberal assumption that justice can be assumed to exist within families and between family members and instead argues that the social institutions of marriage and childrearing practices create a vulnerability of women in marriage that is unjust and leads to harm to women and children. Okin proposes either changing the traditional gendered division of private and public work and related gender differentiation of men's and women's roles in families and in public or compensating those women and their children who are made vulnerable by the existing gendered system. Specifically, until society does not perpetuate gender difference in economic means and opportunities, wives are entitled to and should receive half of their husbands' incomes. Such an arrangement should continue even in divorce for as long as the gendered arrangement existed during the marriage. With these and related proposals, Okin challenges society to take responsibility for the vulnerability of women and children that is created by the institution of gendered division of work in marriage and its impact on other social institutions.

Because they challenge the root norms of liberalism—freedom and equality—these feminists might be classified as 'radical'; yet their contributions

to liberal theory's cornerstones of freedom, equality, and justice make them liberal feminists as well. By re-examining the meanings of freedom and equality and the institutions that liberals rely on to promote them, liberal feminists advance liberal theory while advancing the feminist agenda of promoting theory, law, and social practices that respect men and women as having the equal worth of all human beings (see *Feminist Theory: Radical Lesbian*).

3. Feminist Critiques of Liberal Feminism

While engaged in a critical project of its own, liberal feminist theory has also been the target of criticism. However, liberal feminists generally respond to criticism by expanding their collective agenda and self-understanding such that liberal feminism continues to evolve.

Marxist feminists have criticized capitalism as causing the sexual division of labor in the family and have criticized liberal feminists' focus on equal opportunity (see *Feminist Theory: Marxist and Socialist*). Liberal feminists likewise are concerned about the sexual division of labor in the family (see *Family and Gender*) but they seek to change social norms regarding male and female roles in family maintenance—increasing women's role in providing the economic resources for the family and promoting equal opportunity for women in economics and politics while increasing men's role in providing for the care of the family (Okin 1989). In addition liberal feminists working in development have challenged the global distribution of resources and economic power without objecting to a capitalist organization of economics (Sen and Grown 1987).

Radical and lesbian feminists accuse liberal feminists of accepting patriarchy and the heterosexual model of family such that any challenge to gender norms within the family is not radical enough. In other words, according to their radical critics, liberal feminism does not get at the root causes of gender inequality (see *Feminist Theory: Radical Lesbian*). Liberal feminists respond by proposing family related policies that incorporate the experiences and needs of single and lesbian women and do not assume that families are headed by heterosexual married couples.

Although generally responding to its critics in ways which expand the scope of its own agenda, liberal feminist theory is less responsive to postmodern feminist theory. Postmodern feminists interrogate categories and challenge analyses that rely on norms of 'woman' and 'gender' as being exclusionary and limiting even in their attempt to be critical and analytical (see *Feminist Theory: Postmodern*). Liberal feminists respond by being self-conscious about their own ability to exclude through their assumptions (as liberal theorists exclude through their assumptions), but they are unwilling to abandon the use of categories of analysis as a tool for their critical endeavor (Benhabib 1995). Moreover, they accuse postmodern feminists of abandoning a genuine concern for the excluded by undermining the tools necessary to analyze their exclusion (Okin 1995, Nussbaum 1992).

In sum, because it responds to theoretical traditions, laws, and social practices that have changed over time, liberal feminism is not narrowly definable. Moreover, due to their concern about inclusion, many liberal feminists not only engage in a critique of liberalism but also incorporate into their analyses criticism that has been directed toward certain aspects of liberal feminist thinking. Thus, liberal feminism is part of the broader feminist project of making theory, law, and social practice and the criticism of these a more inclusive social process.

See also: Evolutionism, Including Social Darwinism; Feminist Epistemology; Feminist Ethics; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Mobility: Social; Women's Suffrage

Bibliography

Afkhami M (ed.) 1995 *Faith and Freedom: Women's Human Rights in the Muslim World*. Syracuse University Press, Syracuse, NY

Anthony S B 1972 Woman wants bread, not the ballot! In: Schneir M (ed.) *Feminism: The Essential Historical Writings*, 1st edn. Random House, New York, pp. 137–42

Astell M 1970 *Some Reflections upon Marriage*. Source Book Press, New York

Basu A (ed.) 1995 *The Challenge of Local Feminisms: Women's Movements in Global Perspective*. Westview Press, Boulder, CO

Benhabib S 1995 Feminism and postmodernism: An uneasy alliance. In: Benhabib S, Butler J, Fraser N, Cornell D (eds.) *Feminist Contentions: A Philosophical Exchange*. Routledge, New York, pp. 17–34

Bunch C 1990 Women's rights as human rights: Toward a revision of human rights. *Human Rights Quarterly* **12**(4): 486–98

Charlesworth H 1994 What are 'Women's International Human Rights?' In: Cook R J (ed.) *Human Rights of Women: National and International Perspectives*. University of Pennsylvania Press, Philadelphia, pp. 58–84

Flax J 1995 Race/gender and the ethics of difference: A reply to Okin's 'gender inequality and cultural differences.' *Political Theory* **23**(3): 500–10

Friedan B 1974 *The Feminine Mystique*, 10th aniversary edn. Norton, New York

Kabeer N, Subrahmanian R 1996 Institutions, relations and outcomes: A framework and case studies for gender-aware Planning. IDS discussion paper, Institute for Development Studies, Brighton, UK
Locke J 1988 *Two Treatises of Government*. Cambridge University Press, Cambridge, UK
Lorde A 1979 The master's tools will never dismantle the master's house. In: Moraga C, Anzaldua G (eds.) *This Bridge Called My Back: Writings by Radical Women of Color*, Kitchen Table, Women of Color Press, New York, pp. 98–101
Lorde A 1981 An open letter to Mary Daly. In: Moraga C, Anzaldua G (eds.) *This Bridge Called My Back: Writings by Radical Women of Color*. Kitchen Table, Women of Color Press, New York, pp. 94–7
MacKinnon C 1993 Crimes of war, crimes of peace. In: Shute S, Hurley S (eds.) *On Human Rights: The Oxford Amnesty Lectures 1993*. Basic Books, New York, pp. 83–109
Mill J S 1980 *The Subjection of Women*. Attar, Arlington Heights, IL
Mohanty C T 1984 Under Western eyes: Feminist scholarship and colonial discourses. In: Talpade Mohanty C, Russo A, Torres L (eds.) *Third World Women and the Politics of Feminism*. University of Indiana Press, Bloomington, IN, pp. 51–80
National Organization for Women 1967 NOW Bill of Rights. In: Jaggar A M, Rothenberg P S (eds.) *Feminist Frameworks: Alternative Theoretical Accounts of the Relations between Women and Men*. McGraw-Hill, New York, p. 159
Nussbaum M 1992 Human functioning and social justice: In defense of Aristotelian essentialism. *Political Theory* **20**(2): 202–46
Okin S M 1989 *Gender, Justice and the Family*. Basic Books, New York
Okin S M 1994 Gender inequality and cultural differences. *Political Theory* **22**(1): 5–24
Okin S M 1995 Race gender and the ethics of difference—Response. *Political Theory* **23**(3): 511–16
Okin S M 1999 *Is Multiculturalism Bad for Women*? Princeton University Press, Princeton, NJ
Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA
Phillips A 1993 *Democracy and Difference*. The Pennsylvania State University Press, University Park, PA
Sen G, Grown C 1987 *Development, Crisis, and Alternative Visions: Third World Women's Perspectives*. Monthly Review Press, New York
Stanton E C 1972 Seneca Falls Declaration. In: Schneir M (ed.) *Feminism: The Essential Historical Writings*, 1st edn. Random House, New York, pp. 76–82
Steinem G 1983 *Outrageous Acts and Everyday Rebellions*. New American Library, New York
Stewart M 1987 *America's First Black Woman Political Writer*. Indiana University Press, Bloomington, IN
Taylor H 1970 Enfranchisement of women. In: Rossi A S (ed.) *The Subjection of Women in Essays on Sex Equality*. University of Chicago Press, Chicago, pp. 91–121
Tinker I 1990 A context for the field and for the book. In: Tinker I (ed.) *Persistent Inequalities: Women and World Development*. Oxford University Press, New York, pp. 3–13
Wollstonecraft M 1975 [1792] *A Vindication of the Rights of Women: An Authoritative Text, Backgrounds, Criticism*, Poston C (ed.). Norton, New York

B. A. Ackerly

Feminist Theory: Marxist and Socialist

1. *Definition of the Field*

Socialist feminist and Marxist feminist theory has been a major form of both feminist theory and of socialist theory. Key elements include: the analysis of gender in its intersection with class; a focus on the material, economic, and political dimensions of social relations; a prioritization of the analysis of inequality rather than difference; a focus on social transformation rather than either the recognition of difference or of equal opportunities; and location within the Enlightenment tradition with its commitment to reason, rather than postmodernist scepticism about the possibility of objective explanation.

It was conventional in the 1980s to distinguish three major types of feminist theory in which socialist and feminist theory was one alongside liberal feminism (with a focus on individuals and equal opportunities) and radical feminism (with a focus on men's oppression of women). Since then the major division in feminist theory has become that between modernist theories such as socialist feminism and postmodernist theories (described below). There is a fuzzy rather than clear-cut boundary around socialist and Marxist feminist theory. For instance, there is an issue of how much autonomy can be granted to gender relations from class relations before the theory is better classified in materialist feminism or radical feminism rather than socialist feminism. Or again, it is unclear how much significance can be accorded to culture before a different categorization becomes appropriate.

There are various sources of variation within socialist feminism including among others: the extent of the centrality of class relations as against the centrality of men's direct exploitation of women and the relative importance of economic and cultural dimensions.

Marxist feminist and socialist feminist theory overlap significantly, though the extent of this varies in different times and places. Sometimes the terms are used interchangeably, while in other contexts there are specific demarcations. When there are demarcations, this most often involves Marxist feminism being more focused on class analysis and socialist feminism on the specificity and autonomy of gender dynamics. Further, materialist feminist theory overlaps with socialist feminism, in that it has an emphasis on the material exploitation of women by men, though the extent of the priority given to gender is such that it is sometimes distinguishable from socialist feminism.

2. *Historical Development*

Socialist feminist theory first developed in Western capitalist societies in the eighteenth century, initially in the context of radical communitarian forms of social-

ism which embraced women's liberation as a core part of their practice, such as that of Owenism. Marxist feminist thought developed a little later, in the nineteenth century, in the context of Marxist socialist practice and theory and of both class struggle and first-wave feminism.

While Marx himself had little to say about feminism, Engels (1941) did, writing of the changing forms of the family using a historical materialist method. He drew attention to reproduction as well as to production, and the way in which the different social organization of the form of reproduction in the family affected many other dimensions of society. The merit and precise meaning of the distinction between reproduction and production has been the subject of extensive later analysis. Gilman (1896), a materialist feminist and a major theorist of first-wave feminism, focused the analysis of women's labor onto the way that men were the beneficiaries of the exploitation of women. This theoretical focus is echoed in the work of contemporary theorists such as Delphy (1984), who likewise see the expropriation of women's labor in the household as key to the explanation of women's subordination. Materialist feminism shares with socialist and Marxist feminism a concern with the centrality of labor, but pulls away from it as a result of the assumption of the independence of gender dynamics from those of class. In the revitalization of feminism in the West from the 1960s, there was a flowering of many hues of socialist feminist and Marxist feminist thought, both within and without the academy. This led to a major rethinking of the significance of gender for social theory and the social and behavioural sciences more broadly. Specific attention was paid to the analyses of women's labor both at home and in the labor market. During the 1970s and 1980s there was a significant extension in the range of substantive concerns which were subjected to a socialist feminist analysis, including the state, sexuality, culture, and male violence.

3. *Domestic Labor Debate*

The domestic labor debate of the 1970s set out to theorize the relationship between housework and capitalism. Housework was understood broadly in the context of domestic labor. This was intended to include childcare as well as housework, and indeed other forms of care work and servicing that women did at home without a wage. The protagonists debated the possible benefits of domestic labor for capitalism, and for husbands and fathers. Capitalists were seen to benefit from the nurturance and care of male workers who were more productive as a consequence, and also from the rearing of the next generation of well-socialized workers. The debate explored the extent to which it was possible to utilize traditional Marxist terminology and apply it to gender inequality, and the extent to which a new set of concepts was needed. In particular, they asked whether domestic labor was productive, unproductive, or nonproductive, of value, and surplus value. This was important to Marxist theory of capitalism, since labor that was productive of value and surplus value was considered to be more central to the workings of capitalism than work that was unproductive or nonproductive. These feminist writers thus challenged successfully the traditional Marxist assumption that women's domestic labor was marginal to capitalism, as part of their argument that Marxist theory needed a better appreciation of the significance of gender relations, and of women as a potential political force.

The debate was a highly sophisticated exploration of these theoretical issues, which eventually left far behind its starting point in the everyday oppression of women—too far behind for some critics of the entire debate. The whole field was criticized for economism and its use of arcane technical language. With the fall of the kind of socialism practiced in Eastern Europe Marxist conceptualization has become regarded as less significant. Today, Marxist concepts are utilized rarely in the understanding of housework. Nevertheless, the question of the nature of the relationship of domestic labor to the market economy remains an important one, and one pursued by socialist feminists using a less overtly Marxist terminology. Today the focus is on domestic labor as care-work (Gardiner 1997), and on the variations in the extent to which the state provides for the collective and social performance of this labor, rather than its being left as a private matter for individual households (Folbre 1994).

4. *Women's Waged Labor*

The inequality between women and men in the labor market was theorized as resulting from capitalism, with major debates as to the mechanisms by which the connection was to be understood. The inequalities include: lower wages; lesser job security; fewer fringe benefits such as employer-based pension schemes; occupational segregation. The nature of the causes of the greater vulnerability to exploitation of women's labor than men's was the subject of much debate. This greater vulnerability was seen sometimes as stemming from women's position in the family that ill-prepared them with the skills needed for the labor market. Further, divisions and lack of solidarity between workers as a result of gender and racialized divisions were seen to facilitate such exploitative strategies by employers (Cockburn 1983).

A variant of socialist feminism, sometimes referred to separately as dual-systems theory, theorized the subordinate position of women in the labor market and family as a result of the interlocking of the system of capitalism with a system of patriarchy (Hartmann 1976). Here the subordination of women in the system of patriarchy interacted with that of women's subordination within capitalism, and was, in certain

periods and locations, fuelled by the organization of men on their own behalf, such as in the male-only or male-dominated trade unions of the late nineteenth century. Dual-systems theory played an important role in the development of analyses of occupational segregation, since it was hard to see that this form of labor market rigidity was in a simple sense a result of employers' interests, or indeed women's interests, but rather that to be understood it was necessary to include a notion of male workers' interests. The relationship between patriarchy and capitalism was sometimes seen as harmonious, in that the subordination of women in one system supported their subordination in the other, and sometimes in tension, as when the demand for labor by expanding capitalist economies pulled women out of patriarchal households (Walby 1986).

Recent socialist feminist analysis on women's labor has included analysis of the strategy for delivering comparable worth, that is, equal pay for work of equal value using a legally-based intervention. This focuses on the extent to which patriarchal work systems are resilient to feminist campaigns rooted in equality politics. While most of the analysis of comparable worth has taken place in the US, with largely pessimistic conclusions as to the tenacity, depth, and flexibility of patriarchal practices in the workplace, the strongest legislation is actually within the EU, where analysis is still in its early stages.

Economics as a discipline has integrated a feminist perspective to a lesser extent than some other disciplines, for instance, sociology. Yet, there has been a sustained critique of and engagement with mainstream economics, in which socialist feminist theoretical leadership has played a significant role (England 1992, Cook et al. 2000).

5. *Class Divisions*

In socialist and Marxist feminist theory class divisions are important in a number of ways. First, the class divisions between capital and labor structures women's lives, both their domestic labor and their participation in waged labor. Second, the class divisions between women mean that it is not possible to generalize about women as a whole, since women in different class locations will have different interests. For instance, a woman who works as a cleaner in a private household will not have the same economic interests as the woman who employs her.

6. *Culture*

Most socialist and Marxist feminist theory has conceptualized culture as a secondary dimension of women's lives. While patriarchal ideologies that support women's oppression are noted, these are not usually seen as the most important part of the explanation.

However, since the late 1980s the perceived importance of culture has grown within much Western social science, although the extent of this change is uneven and is more acute in the UK than some other countries. This 'cultural turn' has had an impact upon socialist and Marxist feminist theory in two ways. First, the turn to culture was incorporated within a self-defined Marxist feminism, as part of a new and more sophisticated interpretation of Marxism. For instance, there were attempts to address questions of sexuality and to integrate Marxism with Freud and with Foucault. Second, there was a turn away from socialist and Marxist feminism, and a critique of this school of thought as economistic. The overall impact was to displace the concern with women's paid and unpaid labor from its central location in theoretical questions within feminist theory, even though empirical studies continued on women's labor.

7. *'Race,' Ethnicity, and Gender*

Since the 1980s there have been criticisms of feminist (and other) academic writing for insufficient attention to issues of 'race' and ethnicity. This was not only a criticism of the scarcity of studies focused on issues of race and ethnicity. More importantly, it was a criticism of the lack of theoretical centrality given to race and ethnicity in feminist theory. The strongest criticism was that socialist and Marxist feminist theory was inherently unable to deal with the complexity needed to theoretically engage with 'race' and gender simultaneously.

However, attempts at analyzing the relationship of gender and ethnicity by the critics have themselves been subject to problems, not least of which have been tendencies to, first, empirical depth at the expense of explanatory power; and second, a concern with difference at the expense of inequality. These issues are re-articulated in the debate on postmodernist developments in feminist theory.

The integration of issues of analyses of ethnicity, gender, and class remains an important theoretical debate today, and re-emerge in different forms in the topics discussed below, especially those of: postmodernisn/modernism; recognition/transformation; agency/structure; development and globalization.

7.1 *Postmodernism/Modernism*

Socialist and Marxist feminist theory is modernist rather than postmodernist and draws strongly on an Enlightenment heritage. This involves several elements of which two will be noted here. First, there is an assumption that explanation, while fallible, is both possible and desirable, and that merely describing the world, albeit in complex and interesting ways, is too modest an ambition for social science. The search for an explanation of the world is necessary in order to understand how to change the world. Without an

explanation of the way the world works, attempts at change, such as a feminist political project, are directionless. Second, there is a search for simplicity in the explanation of large and complex phenomena, in terms of powerful underlying motors of change, and of underlying large-scale social structures. This is in contrast with the postmodernist fascination with complexity that has a tendency to merely descriptive accounts of processes.

7.2 Recognition/Transformation

Socialist and Marxist feminist theory seeks to analyze processes of transformation rather than simply recognizing existing social relations or limiting demands for social justice to that of equal opportunities (Fraser 1995). There has been a vigorous debate in feminist (and other social) theory as to the merits of the politics of recognition and the politics of equality, as to which is more effective in achieving social justice. Equality politics has been criticized because, in its focus on the unequal distribution of material privileges, it neglects the significance of cultural differences that are alleged to be more important. Recognition politics focuses on cultural dimensions of social life and on the positive valuation of difference in the political arena. Two examples of recognition politics are the demand for respect for religious customs and for quotas for the representation of subordinate social groups in political bodies, such as parliaments. An example of equality politics is a drive for equal pay through equal opportunities laws. The socialist feminist critique of this debate holds that the strategy of recognition politics can merely embed existing inequalities even as it recognizes the cultural specificity of subordinate groups, while equal opportunity politics, by seeking redistribution within an existing hierarchical system, does not fundamentally alter that system. By contrast, the socialist feminist argument is that a focus on transformation is needed since it is the overarching system of social inequality which needs to be changed, not just which groups, such as women or racialized minorities, are disadvantaged. This entails a less hierarchical division of labor and less unequal distribution of income. It is argued that the goal of transformation demands reasoned argument to build a new social project, rather than embracing existing differences and the projects of existing groups, or seeking the redistribution of existing privileges. This debate has echoes in many branches of contemporary social and political theory, as it grapples with integrating theoretical concern with difference into traditional social theory.

7.3 Agency/Structure

Socialist and Marxist feminist analysis uses concepts of social structure. This is in contrast to some forms of feminist theory that focus primarily on agency and performativity. This conceptualization of the social world is needed in order to capture the deep rootedness of social processes, their persistence over time, and to enable an analysis which connects large-scale processes of different kinds into an overarching system. This does not deny the relevance of agency, despite many criticisms of structural analysis to this effect, since collective political agency is regularly theorized by socialist feminists as a key route to social change. But rather it involves rooting agency in structure, so as to understand that while people choose, they do not do so under conditions of their own making.

7.4 Globalization and Development

While within some branches of Western feminist theory there was a move away from structural accounts of gender inequality during the 1980s and 1990s, this change has not been universal. In particular, within Development Studies the socialist feminist tradition has remained strong (Moghadam 1996). This has been partly because of the agreed significance of the issue of economic poverty and of the role of international capital in the shaping of the process of development.

Since the late 1990s the interest in the issue of globalization is leading to a resurgence in the types of analysis which traditionally have been at the core of socialist feminist analysis, such as the transformation of work and collective political agency. This is because of the desire to understand the interconnectedness of large-scale global processes, and the role of international capital in the changes associated with globalization.

8. The Mainstreaming of Socialist Feminist Theory?

It would be conventional to suggest that socialist and Marxist feminist theory has lost most of its vibrancy during the 1980s and 1990s and that the ascendancy in feminist theory has passed to theories which are more cultural and literary in substantive and theoretical focus. Indeed since the fall of the Wall between East and West Europe Marxist theory has become unfashionable, being regarded as disproven by historic events. Certainly hypotheses about the revolutionary potential of the proletariat need to be rejected.

However, this should not lead to the underestimation of the impact of socialist feminist theory and the continuing significance of its core theoretical concerns. Several of the initial positions of socialist feminist theory which were once regarded as unusual and original have become accepted by traditions which might once have been hostile opponents of such theory. These include, for example, arguments as the importance of labor to understanding gender inequality, and of the centrality of gender to the understanding of the labor market, which are now mainstream assumptions. Today, it is rare for main-

stream social science to deny the relevance of a gender perspective to their subject matter, unlike in the 1970s. Once gender issues were confined to family studies, but today the relevance of gender to a wider set of issues, such as the labor market, is much more readily accepted. To a significant extent then, socialist feminist theory has become mainstreamed and in so doing has lost some of its distinctive identity. In parallel the broadening of the substantive and theoretical concerns within socialist feminist theory has meant the loss of the specificity of some dimensions of a distinctive socialist feminist project (Walby 1997).

However, the move to postmodernism in much feminist theory in the 1980s and 1990s, left socialist feminism as an important carrier of the commitment to modernist forms of theorizing, with a focus on explanation, a search for deep structures and their interconnections, and an interest in the conditions under which gender relations might be transformed. With the development of analysis of globalization the core theoretical concerns of social feminism are again central to current debates, even if they are no longer labeled as socialist feminist.

See also: Feminist Epistemology; Feminist Ethics; Feminist Political Ecology; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Marxism in Contemporary Sociology; Marxist Geography; Marxist Social Thought, History of; Theory: Sociological

Bibliography

Cockburn C 1983 *Brothers: Male Dominance and Technological Change*. Pluto Press, London

Cook J, Roberts J, Waylen G (eds.) 2000 *Towards a Gendered Political Economy*. Macmillan, London

Delphy C 1984 *Close to Home: A Materialist Analysis of Women's Oppression*. Hutchinson, London

Engels F 1941 *The Origin of the Family, Private Property and the State*. Lawrence and Wishart, London

England P 1992 *Comparable Worth: Theories and Evidence*. Aldine de Gruyter, New York

Folbre N 1994 *Who Pays for the Kids? Gender, the Structures of Consent*. Routledge, London

Fraser N 1995 From redistribution to recognition? Dilemmas of justice in a 'post-socialist' age. *New Left Review* 212 (July/August): 68–93

Gardiner J 1997 *Women, Care and Economics*. Macmillan, Basingstoke, UK

Gilman C P (1896/1966) *Women and Economics: A Study of the Economic Relation between Men and Women as a Factor in Social Evolution*. Harper & Row, New York

Hartmann H 1976 Capitalism, patriarchy and job segregation by sex. *Signs* **13**: 137–69

Moghadam V (ed.) 1996 *Trajectories of Patriarchy and Development*. Clarendon Press, Oxford, UK

Walby S 1986 *Patriarchy at Work: Patriarchal and Capitalist Relations in Employment*. Polity Press, Cambridge, UK

Walby S 1997 *Gender Transformations*. Routledge, London

S. Walby

Feminist Theory: Postmodern

The work of postmodern theorists has been a major influence in feminist theory since the 1980s. Building on the work of theorists such as Michel Foucault, Jacques Derrida, and Jean-Francois Lyotard, postmodern feminist theory emphasizes the constituting function of language and discourse in the construction of sexuality and the structure of gender relations. Breaking from modernist feminist theories such as liberal feminism and Marxist/socialist feminism, postmodern feminists reject the founding assumptions of modernist/Enlightenment thought: the attempt to find universal truth, to comprehend the totality of social reality and to define the essential nature of 'woman.' Through the postmodern method of deconstruction they attempt to displace these certainties. They emphasize the discursive nature of social reality and analyze the mechanisms by which gender identity is constituted discursively.

1. Postmodern Theories

Theories termed postmodern have their roots in Nietzsche's challenge to the fundamental values of the Western tradition, his 'transvaluation of values.' French structuralism and psychoanalysis also played a major role in the development of postmodernism. Levi-Strauss's structuralism focused on the way in which the overall structures of cultures create meaning for social actors. Jacques Lacan's psychoanalytic theory asserted that language constitutes the psyche, specifically the definitions of masculinity and femininity; he defined the possibility of meaning in language as deriving from the phallus. These and related theories produced a generation of thinkers, particularly in France, who looked to language and discourse as the constituent elements of social reality. They turned away from the quest for universal truth, what they called 'metanarratives,' and turned instead to the specific definition of 'truth' established by discourses within particular societies.

Lyotard's influential *The Postmodern Condition* (1984) defined the parameters of the postmodern

project. Lyotard argued that 'knowledge' in the contemporary world was necessarily plural and diverse, that the metanarratives that defined modernism no longer hold sway. Identifying postmodernity as discontinuous with modernism, Lyotard argued for new definitions of truth and knowledge appropriate to the postmodern condition. Michel Foucault and Jacques Derrida further challenged the modernist definition of a single, absolute Truth. Their distinctive contribution to postmodern thought was the theory of the death of 'man' or the subject. Foucault argued that the subject of modernism, the autonomous agent who constitutes meaning, is neither universal nor absolute, but, rather, the product of specific discourses at a particular juncture in Western thought. Derrida's work attacked the opposition between the rational, autonomous subject of the Western tradition and the object of knowledge that this subject constitutes. His critique called into question a particular aspect of this subject: the author who creates the meaning of the text. Both Foucault and Derrida argued that this subject, the centerpiece of modernism and humanism, is doomed to extinction.

One of the principal aspects of postmodern theories is an attack on the binaries—dichotomous, hierarchical oppositions—that are at the center of modernist thought. Postmoderns argue that modernist thought is structured by a series of binaries: rational/irrational, nature/culture, subject/object, reason/emotion, absolute/relative, male/female. A major aspect of the postmodern project is the attempt to displace these binaries. Postmoderns argue, first, that they are the result of historically situated discourses and, second, that they distort and limit understanding. The focus of postmodern method is the attempt to 'deconstruct' these binaries, to reveal the unstated presuppositions informing them and the strictures they impose on thought.

2. *The Convergence of Feminism and Postmodernism*

There is widespread agreement among intellectuals that Western thought is witnessing a major shift at the end of the twentieth century: the turn away from the absolutes of modernist thought and its central figure, the autonomous subject, and the embrace of situational, perspectival knowledge and a relational, constituted subject. Both postmodernism and feminism are at the forefront of this intellectual shift. Both engage in a radical critique of modernist thought, attacking the roots of its understanding of knowledge and the subject. They agree that the binarisms that inform modernist thought are necessarily gendered, rooted in the central opposition between male and female and the subordination of the feminine entailed by this opposition. Postmoderns attack this binarism on the grounds that it distorts knowledge and presupposes a false opposition between the essentially feminine and the essentially masculine. At the heart of the modern feminist movement is an attack on the male/female binarism on the grounds that it presupposes the inferiority of women. The postmodern and feminist convergence on this central issue was the initial basis of the attraction of postmodern theories for feminists. The postmodern project of deconstructing 'man' converged in many ways with the feminist project of challenging patriarchy.

Many feminists first encountered postmodern themes in the work of what are now called the 'French Feminists.' Luce Irigaray, Julia Kristeva, and Helene Cixous theorized about the construction of the feminine within language, particularly the construction of female sexuality. Building on the work of Jacques Lacan who defined the position of women in language as a 'lack,' they argued that the oppression of women in Western societies is rooted in language and what they identified as its 'phallocratic' assumptions. They emphasized that language creates subjects, particularly the subject 'woman,' and that this creation both defines and distorts female sexuality. The French Feminists accepted Lacan's thesis that the basis for meaning in language is the phallus, but, instead of accepting his conclusion that women are, therefore, erased by language, they attempted to fashion a 'feminine writing' that would give women a voice.

The work of the French Feminists, although not strictly postmodern, presaged many of the themes of postmodern feminism. The French Feminists identified language as the source of meaning, subjectivity, and sexuality. They questioned the binarisms that structure Western thought and the inferior status for women that they entail. Their attempt to define a 'feminine writing,' however, appears to contradict the postmodern impulse to reject essentialism. This aspect of the French Feminists' thought continues to be controversial for postmodern feminism.

3. *Postmodern Feminist Theories*

The clearest statement of the major issues surrounding postmodern feminism is found in an influential article by Fraser and Nicholson (1990). They argued that feminism and postmodernism are the most important political–cultural currents of the 1980s and that the two approaches can both profit from a closer alliance. Their central thesis is that the strengths of each movement could and should correct the weaknesses of the other. The strength of postmodernism, its rejection of metanarratives, can counter feminism's tendency to construct such narratives. Conversely, feminism's orientation toward social criticism can rectify postmodernism's avoidance of social critique. Calling for an integration of the two approaches, Fraser and Nicholson argued that a postmodern feminism would

be an approach that avoids the metanarratives of modernism but could nevertheless engage in large historical narratives and analyses of societal macrostructures.

The late 1980s and 1990s witnessed an explosion of discussions of postmodernism and feminism. Two books outlined the issues facing a postmodern feminism (Hekman 1990, Weedon 1987). Butler's (1990) book defined the distinctive approach to the subject that would characterize postmodern feminism. Relying on the postmodern critique of the modernist subject as well as psychoanalytic theory, Butler argued that there is no essence to gender identity but, rather, that gender is constituted through the performance of gender roles. Butler's thesis was that gender identity is a regulative ideal that grounds the norm of heterosexuality. She asserted that it is precisely the belief in gender identity, a belief that has been central to feminist theory and political practice since the advent of the feminist movement, that prevents us from identifying the political processes that produce that coherent identity. She thus concluded that gender identity, far from being the means of women's liberation, was the ground of women's oppression.

Butler's book served as a kind of lightening rod for both defenders and critics of postmodern feminism. A series of books and articles in the 1990s further established the parameters of the approach. A notable element of this formulation was the connection to Marxism/socialism. Several prominent postmodern feminists argued for a redefinition of socialist feminism along postmodern lines. Haraway (1990), in her influential article, argued that women need a new identity for the postmodern age that can replace the constructed revolutionary identity of socialist feminism. Like Butler, Haraway attempted to demonstrate how the postmodern critique of identity could be used to formulate a feminist politics appropriate to the postmodern era. In addition to extending the discussion of gender identity, her work also brought postmodern themes into another area of feminist concern: the critique of Western science. Another argument for the connection between postmodern feminism and Marxism/socialism came from Gayatri Spivak. Proclaiming herself a Marxist feminist deconstructionist with an overtly political agenda, Spivak argued for the political usefulness of postmodern feminism. Arguing against critics of the approach who defined it as exclusively a white, male theory, she also demonstrated how deconstructive methods could be used to critique issues involving race and colonialism.

Psychoanalytic theories became important in the emerging definition of postmodern feminism as well. Flax (1990) argued for the convergence of postmodernism, feminism, and psychoanalysis. Identifying feminist theory as a type of postmodern philosophy, Flax argued that, together with psychoanalysis, postmodernism and feminism contribute to the fundamental questioning of the certainties of the Western tradition. Together, she asserts, these three movements force a redefinition of those certainties and encourage questioning of the need for imposing order on social reality and defining its universal structure.

Another influential postmodern feminist theorist, Drucilla Cornell, applied postmodern theories to feminist legal studies. Although legal theory would seem to be an inappropriate subject for postmodern deconstruction, Cornell (1991) argued that legal theory needs a new choreography of sexual difference that can only be supplied by postmodern deconstruction. In a series of books on deconstruction and the law, Cornell developed the ethical dimensions of deconstruction and argued for its usefulness in liberal legal theory.

Postmodern themes have appeared in nearly every aspect of feminist thought. The strongest proponents of a postmodernism feminism are found in English literary criticism where the influence of Derrida and Foucault is pervasive. But postmodern themes also appear in women's history, political theory, philosophy, cultural studies and feminist critiques of science.

4. *Critiques of Postmodern Feminist Theories*

Since it was first introduced in feminist theory, postmodernism has had a mixed reception in the feminist community. The well-known collection, *Feminism/Postmodernism* (Nicholson 1990), is representative of this reception: the articles are almost evenly divided between critics and defenders of a post-modern feminism. The feminist critiques of postmodern feminism that have emerged in the 1980s and 1990s have focused on the fundamental aspects of the postmodern approach, questioning its appropriateness for feminist theory and practice.

The most pervasive critique of a postmodern feminism is that it precludes the possibility of a feminist politics. Di Stefano, for example, argued that feminism is 'firmly, if ambivalently, located in the modernist ethos' (1990). The basic concepts of the contemporary women's liberation movement, 'woman' and 'liberation,' are, she asserts, thoroughly modernist concepts. These concepts are deconstructed and rejected by postmodernism, thus destroying the foundation of feminist theory and politics. Di Stefano also questions the possibility of founding feminist politics on an approach that ignores gender altogether.

Ebert (1996) elaborates on this critique in one of the most comprehensive attacks on postmodern feminism. Ebert decries what she defines as the dominance of 'ludic feminism.' She argues that a feminism founded on postmodern assumptions about linguistic play, difference, and the priority of discourse will result in the substitution of a politics of representation for

radical social transformation. Ebert laments the fact that many socialist feminists have embraced postmodernism and thus abandoned the revolutionary resistance politics that defines Marxism. Ebert and other critics of postmodern feminism claim that by reducing everything to discourse and rejecting the notion of the truth of social reality, postmodernism precludes the possibility of a feminist politics. Far from fostering a politics of resistance, these critics claim, postmodernism entails either nihilism or an ultimately conservative defense of the *status quo*.

Defenders of postmodern feminism argue that politics in general and feminist politics in particular do not require the metanarratives of modernist politics. They also point to the work of postmodernists such as Foucault who clearly place their work in a tradition of political resistance. But, although Ebert and other critics of a postmodern feminism concede that some postmodern thinkers are interested in resistance, they assert that without a view of the truth of social reality, feminists cannot proclaim the truth of women's oppression, a necessary element of any feminist politics.

This critique is related closely to another pervasive criticism of postmodern feminism: its failure to produce systematic knowledge of social reality. Although Fraser and Nicholson argued that systemic social critiques are not precluded by postmodernism this argument has not been persuasive for many feminists. On the contrary, many feminists have argued that postmodernism's approach to knowledge produces nothing but meaningless textual analysis of the infinite differences of social reality (Hartsock 1990). Bordo (1990) argued that this aspect of postmodernism leads to 'gender scepticism' and the infinite play of differences without distinction. These critics conclude that postmodernism's resistance to comprehensive theories precludes a coherent understanding of social reality without which social change is impossible.

Another major focus of critiques of a postmodern feminism is the concept of the subject. Butler and other postmodern feminists' deconstruction of the modernist subject is central to their theoretical project. Yet their critics have claimed that this deconstruction leaves feminism with nothing to replace this subject. The modernist subject, they claim, had agency and autonomy, the postmodern subject has neither. Worse, postmodern theories of the subject produce what its critics call a 'social dupe,' a subject which is wholly constituted by discourse and thus incapable of agentic action.This not only produces passivity but fails to explain the possibility of resistance, including feminist resistance politics. The defenders of postmodern feminism claim that this critique is rooted in modernist assumptions that falsely link agency to the modernist subject. But the question of agency and the 'social dupe' remain central to critiques of postmodern feminism.

Postmodernism's emphasis on texts and discourse is at the root of another critique of postmodern feminism: the charge that deconstructing materiality denies the reality of women's pain. Critics such as Bordo (1993) argue that a central concern of feminism must be an account of the very real, material pain inflicted on women's bodies by the structures of patriarchy and that postmodern feminism's rejection of materiality precludes this. The postmodern feminist rejoinder formulated by Butler (1993) is that focusing on discourses does not ignore the materiality of pain, but, rather, offers the possibility of transforming it. Butler claims that materiality must be the object of feminist inquiry, not its ground.

The origins of postmodern theories have also been the object of feminist critiques. Postmodern theorists, critics charge, are elite white men who ignore gender issues and write jargon-filled discourse that excludes the uninitiated. Feminists writing on race and ethnicity have been particularly vehement critics of this aspect of postmodern feminism. They claim that what Barbara Christian called the 'race for theory' marginalizes women of color. Specifically, these critics assert that the unique perspectives and theorizing formulated by women of color are ignored and negated by postmodern feminist theory.

One consequence of these critiques of postmodern feminism has been a return to the universals of modernism on the part of some feminists. Okin has argued that the anti-essentialism and hyperconcern with differences among women that characterizes contemporary feminism is a dangerous trend. She claims that the result of these emphases is a paralyzing degree of cultural relativism and the impossibility of making any generalizations about women's condition. A similar concern motivates the recent work of Martha Nussbaum. Nussbaum asserts that without a concept of universal human functioning feminists will be unable to address the wrongs done to women throughout the world. Like Okin, she is concerned about the extremes of what she sees as cultural relativism fostered by postmodern feminism and other approaches that emphasize differences among women.

In summary, the jury is still out on postmodern feminist theories. They have had a pervasive influence in contemporary feminism and have generated vehement defenders and critics. On the positive side postmodern theories have aligned feminism with the intellectual shift that characterizes contemporary Western thought and brought it into the era of differences. On the negative side postmodernism has raised questions about the foundations of any feminist politics and the status of the truth claims feminists advance.

See also: Cultural Critique: Anthropological; Cultural Relativism, Anthropology of; Deconstruction: Cultural Concerns; Feminist Epistemology; Feminist Ethics; Feminist Political Ecology; Feminist Theology; Feminist Theory; Feminist Theory and Women

of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Political Thought, History of; Postmodernism in Geography; Postmodernism in Sociology; Postmodernism: Methodology; Postmodernism: Philosophical Aspects; Theory: Sociological

Bibliography

Bordo S 1990 Feminism, postmodernism, and gender scepticism. In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 133–56

Bordo S 1993 *Unbearable Weight: Feminism, Western Culture, and the Body*. University of California Press, Berkeley, CA

Butler J 1990 *Gender Trouble*. Routledge, New York

Butler J 1993 *Bodies that Matter*. Routledge, New York

Cornell D 1991 *Beyond Accommodation: Ethical Feminism, Deconstruction and the Law*. Routledge, New York

Di Stefano C 1990 Dilemmas of difference: Feminism, modernity, and postmodernism. In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 63–82

Ebert T 1996 *Ludic Feminism and After: Postmodernism, Desire, and Labor in Late Capitalism*. University of Michigan Press, Ann Arbor, MI

Flax J 1990 *Thinking Fragments: Psychoanalysis, Feminism, and Postmodernism in the Contemporary West*. University of California Press, Berkeley, CA

Fraser N, Nicholson L 1990 Social criticism without philosophy: an encounter between feminism and postmodernism. In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 19–38

Haraway D 1990 A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s. In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 190–223

Hartsock N 1990 Foucault on power: A theory for women? In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 157–75

Hekman S 1990 *Gender and Knowledge: Elements of a Postmodern Feminism*. Polity Press, Cambridge, UK

Lyotard J-F 1984 *The Postmodern Condition: A Report on Knowledge* (trans. Bennington G, Massumi B). University of Minneapolis Press, Minneapolis, MN

Nicholson L (ed.) 1990 *Feminism/Postmodernism*. Routledge, New York

Weedon C 1987 *Feminist Practice and Poststructuralist Theory*. Blackwell, New York

S. Hekman

Feminist Theory: Psychoanalytic

Psychoanalytic theory, originated by Sigmund Freud and elaborated by numerous competing camps of disciples, has been a major influence on feminist theory throughout the twentieth century. Feminists were among the earliest proponents of psychoanalysis because it acknowledged sexual desires in all people including women, who at the time were considered asexual. But over the course of the century, the theory fell out of favor among feminists, many of whom maintained that it contributed to women's oppression instead of their liberation. In the 1970s interest in psychoanalysis among feminists enjoyed a renaissance, and by the end of the century it formed part of the canon of feminist theory in the humanities. However, the theory still faces a great deal of resistance from feminists in the social sciences.

The most important aspect of psychoanalysis for feminists is its theory of sexual development. In the Freudian account, all infants are polymorph perverse, meaning that humans are not innately predisposed to heterosexual genital activity, but also capable of homosexuality, fetishism, perversions, and all other sexual variations. Freud argued that boys and girls begin life with similar psychological dispositions, but they follow different paths to adult heterosexuality owing to their anatomical differences and the expectations placed upon them by their families and society in general. Boys' development diverges from girls' around age three, during the oedipus complex, when they must relinquish their erotic longing for their mother in the face of castration threats from their father. To maintain their physical integrity, boys will typically project their aggressive fantasies onto their fathers, identify with their fathers, and then internalize the aggression imputed to the father in the form of a punitive superego. Freud wrote relatively little about girls' development, which is surprising since he treated many women patients, some of whom were the subjects of his published case studies. He did note that as beings without penises, the threat of castration was useless against girls' incestuous desires. But he maintained that the inherent superiority of the penis as a satisfying sexual organ inspired such envy among women that they are (unconsciously) compelled to seek a substitute penis of their own, typically by subordinating themselves to men in the hopes of obtaining a male offspring.

Freud's description of the childhood stages of sexual development recognized that individuals could and often did stray from the typical patterns of resolution, due to idiosyncracies of their physiology and the peculiar configurations of their family lives. When individuals did not successfully progress through the stages of development, mental illness could result. Freud maintained that unresolved early childhood traumas associated with these stages of sexual development become lodged in the unconscious and reappear in neurotic symptoms. These symptoms can be relieved through psychoanalytic therapy, a technique that involves free association, recovery of repressed traumatic memories and affects, and trans-

ference, all in the context of a uniquely structured relationship between analyst and patient. Thus, his theory provided both an account of typical sexual development, and a method for curing adult neurosis produced by deviations from those typical patterns.

This theory of sexual development appealed to many feminists who were contemporaries of Freud. Although several of Freud's women proteges, including Karen Horney, rejected his notion of penis envy in women, she and other feminist theorists applauded his emphasis on the psychological similarities between men and women. At a time when most of his notable contemporaries emphasized essential differences in men's and women's natures, Freud observed that individuals contain both masculine and feminine characteristics, and that women are as passionate as men, ideas that resonated with feminists Emma Goldman and Margaret Sanger. Feminists also took from Freud an indictment of women's proscribed roles in society. In Freud's view, marriage and family arrangements demanded an enormous and inordinate amount of repression and sacrifice on the part of women which made them hostile to civilization and often generated mental disorders. But where Freud saw few options or alternatives to these conventional practices, some feminist theorists, including Elsie Clews Parsons, saw this as a critique of Western marriage and an inspiration to experiment with alternative forms of sexual relationships.

Feminism was dormant in mid-twentieth century, but psychoanalysis at the time was experiencing a period of growth and popularity. When feminism reemerged in the 1960s and 1970s, psychoanalysis was targeted as an enemy of women's liberation. Instead of being associated with sexual modernism, which it had represented in the 1920s, psychoanalysis had become alloyed with social conservatism. Feminist activists and theorists, including Betty Friedan, charged that psychoanalysts typically blamed strong and assertive mothers for causing mental illness in their families and a whole host of other social problems. The concept of 'penis envy' was again criticized for defining women as innately inferior to men. Therapy was seen as a tool of patriarchy for enforcing submission to sex roles and brainwashing women into accepting compulsory heterosexuality by diagnosing lesbians as sick and deranged; Mary Daly noted that 'therapist' could be read 'the/rapist,' symbolizing the attitude of psychoanalysis to women's oppression and suffering (Buhle 1998, Gardiner 1992).

In this climate of feminist antagonism to psychoanalysis, two books appeared that eventually inspired a reassessment of Freudian theory by feminists: Juliet Mitchell's *Psychoanalysis and Feminism*, published in Great Britain in 1974, and Nancy Chodorow's *Reproduction of Mothering*, published in the US in 1978. Both writers were active in the women's liberation movement, and both identified themselves as socialist feminists. They turned to psychoanalysis because it offered a compelling explanation of the origin and the intractability of gender domination. Psychoanalysis was unique, they argued, in its ability to explain how patriarchal attitudes and practices are inscribed and internalized in both men and women. They read Freud as showing how, through the normal course of sexual development, both men and women learn to unconsciously desire and reproduce male domination and devaluation of women. In other words, they understood the theory as a description of the inner workings of a patriarchal society—not a proscription for one. Freud had in fact elided this distinction in his work, but Mitchell and Chodorow insisted that there were critical moments in psychoanalytic theory that could be emphasized to enhance its usefulness for feminists intent on changing society to benefit women.

For example, Chodorow demonstrated that the psychoanalytic theory of sexual development could account for why women typically conform to society's expectation that they bear primary responsibility for rearing children. She agreed with radical feminists who argued that this primary mothering is ordered by a sexist society that provides few options for emotionally fulfilling productive activity for women. But she used psychoanalytic theory to uncover the unconscious dynamics that caused women to willingly embrace this family role. She argued that when women do all the mothering, this sets up asymmetries in boys' and girls' sexual development that leads to gender differences in adult men's and women's ability and unconscious desire to render adequate care to infants. Girls develop their primary identity through connectedness to their mothers, whereas boys develop a sense of self as separate from their primary caretaker and all the nurturing qualities she typically embodies. Boys learn that to be 'masculine' requires that they deny and denigrate feminine qualities in themselves and in others. Thus, Chodorow demonstrated that psychoanalysis provides a compelling account of the origin of gender differentiation, helps to explain the desires and proclivities of women to nurture infants, and offers an explanation of men's general devaluation of women, without succumbing to either biological or social determinism, pitfalls that feminist critics of Freud seldom avoided. Psychoanalysis also suggests a strategy for changing these patterns—co-parenting. Chodorow argued that if men were equally involved in childcare, women would be freed to perform other creative and productive work. Also, children would learn to associate nurturing activities with both men and women, so boys would no longer denigrate women (and qualitites currently associated with femininity) in the course of developing a masculine sense of self.

Chodorow's approach draws heavily from the 'objects relations' branch of psychoanalysis, which emphasizes the importance of social relationships and interactions with other people (unfortunately termed objects in the development of the self). This tradition

is the most popular one among the sociologists who advocate psychoanalysis, including Chodorow, Jessica Benjamin, and Miriam Johnson. Miriam Johnson, for example, analyzes the function of the peer group in male psychological development, focusing on its importance in directing men's aggression against women. Political theorist Jane Flax also uses objects-relations theory to challenge notions of universal reason and justice, which she regards as undergirding male domination and other forms of political inequality. The goal of these and other feminist works in the objects relations tradition is to explore the gendered psychological underpinnings and consequences of various forms of social organization with the ultimate aim of criticizing those forms and advocating social and political change.

The other major branch of psychoanalysis that feminists have appropriated is associated with the work of French psychoanalyst Jacques Lacan. This tradition is strongest in the humanities. Juliet Mitchell, who began her career as a literary critic, endorsed a Lacanian approach in her appropriation of psychoanalysis for feminism, arguing that language is an important mechanism through which socialization occurs and unconscious conflicts are expressed. Because of its emphasis on fantasy and imagination, this perspective lends itself more readily than objects-relations theory to the analysis of literature and culture. Other feminist theorists who are associated closely with the Lacanian tradition include Julia Kristeva, Jane Gallop, and Elizabeth Grosz. While all criticize Lacan as sexist, they share his view that acquisition of the culture's dominant symbolic language is the means through which male subjectivity is privileged and male domination is enforced. They also use Lacan's methods of textual analysis to uncover feminine resistance to this symbolic order through hidden expressions of women's unique sexual desire. These theorists reject the notion of a stable self or gender identity; masculinity and femininity are considered the products of discourse, not the psychological attributes of individuals (e.g., Grosz 1990).

The popularity of this perspective is reflected in the fact that familiarity with this branch of feminist psychoanalytic theory is now expected in most graduate gender programs in English, literature, and cultural studies. Psychoanalysis has been less successful at integrating the canon of gender scholarship in the social sciences, however. The perspective is virtually ignored, if not vilified in most psychology departments, and sociologists have generally been wary about embracing the theory.

Feminist social scientists have been reluctant to endorse psychoanalytic theory for a variety of reasons, some of which reflect resistances common among social scientists in general. First, the methodology of psychoanalysis—the in-depth, case study of a neurotic individual typically compiled through prolonged one-on-one discussions with a therapist—is considered incompatible with accepted research methodologies in the social sciences. Psychoanalysis violates principles of random selection, and its emphasis on the abnormal or pathological to explain the normal patterns of development, goes against the social science logic of focusing on central tendency.

A related problem concerns the different levels of analysis in psychoanalysis and the social sciences. It is very difficult to detect patterns of group behavior based on personal, idiosyncratic information about an individual's unconscious motivations, or to infer anything about individuals' experiences based on broad-based, impersonal social forces. Feminists are not alone among social scientists in detecting this incompatibility, but it has special consequences for theorizing about gender. Since feminism begins with the premise that women as a group are disadvantaged compared to men, generalizations about gender differences are central to the project of feminist theory. But psychoanalysis focuses on individual uniqueness; the assumption is that no two people share identical unconscious needs, desires, or motives. Consequently, gender identity itself can and does have different meanings for different individuals, a point that has been emphasized by Chodorow in her later work, aptly titled *Femininities, Masculinities, Sexualities* (1994). Not coincidentally, Chodorow's movement away from large-scale sociological generalizations about gender identity followed her completion of training as a professional psychoanalyst.

The psychoanalytic focus on the unconscious and the irrational is a second obstacle to its incorporation into feminist social science. In general, social scientists trained in positivistic and quantitative methodologies have raised questions about the reliability of psychoanalytic data. In addition to this generic concern, feminist social scientists have voiced some reluctance to attribute nonrational motivations to women's behavior. For some, this theoretical approach reproduces sexist beliefs about women's reputed emotionalism and inability to act rationally. Interestingly, sociologists specializing in the relatively new field of masculinity studies have been more open to psychoanalytic theory, perhaps because the emphasis on the irrational motives of men is less worrisome to feminists (e.g., Connell 1995, Segal 1990, Williams 1995).

A third set of criticisms made of psychoanalytic theory by a number of feminist social scientists concerns the possibilities for social and political change. Many object to psychoanalytical emphasis on treating the individual neurosis without addressing the social problem that causes the illness. To the extent that feminist psychoanalytic theory has concerned itself with social change, the focus has been exclusively on changing the family, and not the economy, the political system, or other public institutions that contribute to discrimination against women. Even the changes that psychoanalytic feminists endorse have

been criticized: some feminists reject the co-parenting strategy because it ignores, if not demonizes, family patterns that deviate from the conventional heterosexual nuclear form, such as gay and lesbian families and single-parent families. Although theoretically these family types are not incompatible with feminist psychoanalysis, critics are correct that little attention has been paid to the psychodynamic consequences of being raised in these alternative family settings.

Incorporating diversity is one of the major challenges facing feminist psychoanalytic theory in the future. For the most part, feminist psychoanalysis has focused on white middle class heterosexuals, leading some feminists to reject the theory as irrelevant for understanding the lives of those suffering from the effects of racism, poverty, and heterosexism. Some efforts are being made to broaden the scope of feminist psychoanalytic theory to incorporate these experiences, but more work is needed (e.g., Abel et al. 1997).

A second challenge for feminist psychoanalytic theory is to move beyond the critique of the family as the mainstay of male domination and privilege in society. Other institutions, including schools, the popular media (especially television), and day-care centers, are increasingly important in the development of children's gender identity. Our understanding of childhood socialization in these settings could be enriched with a psychoanalytic account (e.g., Martin 1996).

Finally, feminist psychoanalytic theorists must develop the interpretive tools needed to demonstrate the operation of psychodynamic forces outside the therapeutic environment. Lacanian feminists have been successful at applying psychoanalytic concepts to the interpretation of literary texts. What is needed is a similar effort directed to understanding the unconscious motives that underlie lived engagement in the social world. Sexual harassment, workplace segregation, domestic violence, and obsession with body image are all examples of seemingly intractable social problems that call out for the kind of deeper analysis of male domination that psychoanalytic theory can provide. But this will require new psychoanalytic methodologies that will permit generalization, and a greater willingness on the part of social scientists to incorporate unconscious and irrational motives in their explanations of social behavior.

See also: Feminist Epistemology; Feminist Ethics; Feminist Political Ecology; Feminist Psychotherapy; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender Differences in Personality and Social Behavior; Gender Ideology: Cross-cultural Aspects; Klein, Melanie (1882–1960); Masculinities and Femininities; Psychoanalysis in Sociology; Psychological Anthropology; Theory: Sociological

Bibliography

Abel E, Christian B, Moglen H 1997 *Female Subjects in Black and White: Race, Psychoanalysis, Feminism*. University of California Press, Berkeley, CA

Buhle M J 1998 *Feminism and Its Discontents: A Century of Struggle with Psychoanalysis*. Harvard University Press, Cambridge, MA

Chodorow N J 1978 *The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. University of California Press, Berkeley, CA

Chodorow N J 1994 *Femininities, Masculinities, Sexualities: Freud and Beyond*. University Press of Kentucky, Lexington, KY

Connell R W 1995 *Masculinities*. University of California Press, Berkeley, CA

Daly M 1978 *Gyn/Ecology: The Metaethics of Radical Feminism*. Beacon Press, Boston

Flax J 1993 *Disputed Subjects: Essays on Psychoanalysis, Politics and Philosophy*. Routledge, New York

Gardiner J K 1992 Psychoanalysis and feminism: A humanist's view. *Signs* **17**(2): 437–54

Grosz E 1990 *Jacques Lacan: A Feminist Introduction*. Routledge, New York

Johnson M M 1988 *Strong Mothers, Weak Wives: The Search for Gender Equality*. University of California Press, Berkeley, CA

Martin K A 1996 *Puberty, Sexuality, and the Self: Girls and Boys at Adolescence*. Routledge, New York

Mitchell J 1974 *Psychoanalysis and Feminism*, 1st edn. Pantheon, New York

Segal L 1990 *Slow Motion: Changing Masculinities, Changing Men*. Rutgers University Press, New Brunswick, NJ

Williams C L 1995 *Still a Man's World: Men Who Do 'Women's Work*.' University of California Press, Berkeley CA

C. Williams

Feminist Theory: Radical Lesbian

Perhaps what most obviously differentiates mainstream liberal feminism from her more radical sisters is that while the former continues to hope that the master's tools can be deployed to dismantle and then rebuild a better version of the master's house, the latter has largely quit this project and jettisoned its tools in favor of building not merely a house but an entire conceptual village whose collective mission it is to embody the experiences, values, and desires of

women. That is, while the primary focus of mainstream liberal feminist theory has been the critical evaluation of patriarchal or male-dominated institutions such as government, medicine, family, and religion in the interest of retooling these institutions to affect greater equality or inclusion for women within them, radical feminism takes its point of departure to be the position that any thoroughgoing critique of patriarchy must include an analysis of the tools themselves.

How, in other words, do concepts like 'woman,' 'man,' 'body,' and 'mind' inform the maintenance of these institutions? How have we defined notions like 'the rational,' 'the desirable,' 'the good,' 'the moral,' and 'the beautiful,' such that being born male virtually guarantees greater access to opportunities and resources? What are the conceptual and ideological anatomies of patriarchies as we find them across culture and history? How do patriarchal ideologies help to maintain patriarchal institutions?

Perhaps, however, the central issue which gave birth to contemporary radical feminism in the late 1960s and early 1970s was the emerging recognition that it could never be enough to agitate for an equality of inclusion within a conceptual framework which defines women's value only in reference to men. The liberal feminist project cannot fail to risk the self-defeat that accrues to retaining a conceptual framework—the master's tools—which systematically inferiorizes women. The merely cosmetic improvement produced in some women's lives (mostly Caucasian, middle-class, and educated like their male counterparts) inevitably fails to address the very real life and death issues which confront many women both locally and globally. Equipped with this insight, radical feminists undertook the conceptual as well as political quest to discover, for example, how the inferiorizing of women is accomplished, and how the domination of women varies with respect to class, ethnicity, culture, ability, age, sexual orientation, and institution. As late feminist philosopher Audra Lorde put it, the master's tools cannot be used to dismantle the master's house. The revolutionary change necessary to produce lasting improvement in all women's lives requires that we fashion our own tools.

Some radical feminists argue, for example, that the notion of equality is itself tethered to a view of human nature which so privileges attributes traditionally associated with men such as aggression, hierarchical forms of organization, rugged individualism, competition, and ethical systems that strongly favor impersonal rule-following over relationality, that women are only able to be included within such conceptual frameworks as honorary men. It is little wonder, they argue, that examples offered to young women by liberal feminists as role models and heroines are modeled after male heroes; the accomplishments of women like Harriet Tubman, or Marie Curie (or movie heroines like those found in *Terminator II*, *Aliens*, or *Femme Nikita*) are valued according to a standard that rewards domination over domesticity, conquest over caring, and militarism over mothering, in short masculinity over femininity. This is not to imply that these are not examples of very real valor, but that what is rewarded in them goes further to reinforce a male-centered or masculinist vision of the good than it does to foster a vision centered around values identified as feminine such as cooperation, compromise, collective decision-making, or compassion. It is thus simply not enough, argue radical feminists, for women to settle for reformist programs that aim at equality. For given the conceptual constraints that determine value, such programs will either fail or, even if they succeed, will do so only at the cost of continuing to devalue qualities associated with femininity, and hence with women.

Liberal feminism does not, in other words, necessarily represent the desires or interests of all women, whether we take these desires to be culturally constructed or endowed by nature. Here, however, we arrive at another important juncture in radical feminist theory: if what conceptual analysis shows is that the ways in which we conceive of gender operate consistently to privilege men, ought we to consider these ways a reflection of natural propensity, and then aim to elevate the value of feminine characteristics like nurturing, emotional responsiveness, and the tendency to focus on relationships over abstract autonomy? Or, alternatively, should we eschew traditional femininity as the construction of patriarchal culture, arguing that there are no *essential* qualities that importantly distinguish men from women? If so, at least two other possibilities present themselves: One meaning of *androgyny* favors the promotion of the masculine and the feminine in all of us, acknowledging that while gender may be constructed, such qualities may still be salvageable reinterpreted in a nonsexist context. Another meaning of androgyny, however, suggests that we ought to abandon notions of femininity and masculinity altogether as unworkable in favor some other notion of personhood. Here the question that arises concerns what such a notion would look like, and whether it is possible to envision a conception of human being that can comprehend biological differences without merely reproducing an atomistic individual so divorced from the bodily that it becomes subject to the same criticism radical feminists level at liberal feminism. That is, can we envision a concept of 'embodied person' which, irreducible to the determinations of either biology or culture, is also rich enough to account for differences relevant to and affected by both?

Whatever the resolution to the essentialism/anti-essentialism debate among radical feminists, their main critique of liberal feminism remains unaffected, namely, that what the history of patriarchy demonstrates is that the oppression of women is sanctioned through the denigration of whatever qualities are

identified with femininity. Whether we seek liberation through the elevation or the abolition of femininity what remains clear is that reformist programs for inclusion can never by themselves be adequate. We have not, however, exhausted our avenues of critique, for there remains another possibility: Marxist/socialist feminism whose primary aim it is to show how, given the marriage of patriarchy to *capitalism*, it is illuminating to conceive of women not merely as a group but as a *class* whose oppression is maintained first and foremost through economic means. Combined with a radical feminist analysis of patriarchal concepts of motherhood and domestic labor, a Marxist/socialist account shows how divisions of labor are themselves typically sexed in at least two ways: first, because women's *labor qua* procreation (the reproduction of human life itself) is assigned to the 'natural' and hence is not counted as labor at all, and second, because the domestic labor performed almost exclusively by women, (the reproduction of the conditions of life) has itself become naturalized in a capitalist economy that identifies all labor as *paid* or *wage* labor; that is, if it is not paid, it is not labor. Cast as the vehicle for the reproduction of both life and the means to life, 'woman' appears a paradox. 'She' is unexploitable in the sense that her labor is not counted as labor within a capitalist conceptual framework at all, yet 'she' also epitomizes the very prototype of exploitability in that 'her' labors are systematically and maximally undercompensated.

Again, the valuable point can be made without entering the essentialism/anti-essentialism debate: regardless what qualities we assign to nature or nurture, *that* women's oppression is intimately linked to the relationship between her economic status and her status as provider of hearth, home, and progeny *for men* well supports the radical feminist claim that there exist no social institutions in which men do not have a controlling interest, including economy and family. What recommends this approach over liberal feminism, then, is the emphasis on the role of economic class, for whereas the best that liberal feminists can hope for is an equality of competition between men and women within capitalism, the radical Marxist/socialist feminists are free to imagine a world within which neither sex nor class oppression are operative; radical feminist theory is thus utopian in a way that liberal feminism cannot be. However, though no doubt radical Marxist/socialist feminism goes further than liberal feminism or than radical feminism alone, even it ultimately fails the project of genuine emancipation. For while such a critique of capitalism shows something important about the specifically sexed parameters of economic oppression, there remains considerable work to be done to show the relationship between economic and other forms of oppression such as sex, sexual orientation, race, age, ability, and so on.

Given the need for this work, many radical feminists distinguish themselves from both mainstream liberal feminism and Marxist/socialist feminism by taking a more skeptical, self-reflective approach to theorizing, arguing that it is *at least* an open question whether any patriarchal institution is salvageable in any form. There is little more revolutionary a declaration than that capitalism must be overthrown, but if this is not accompanied by the overthrow of other patriarchal institutions, the status of women could remain unchanged or worsened as men's conditions improve. What revolutionary changes would institutions like the family be required to undergo to achieve equality not only with respect to gender, but with respect to any number of other axes of oppression, including race, ability, economic class, indigenous culture status, age, sexual orientation, and gender expression?

Some radical feminists such as Mary Daly argue that patriarchy must be overthrown or abandoned in favor of *separatist* or women-only cultures and societies. In her 'radical elemental feminist manifesto,' *Quintessence*, Daly describes the fundamental ideology of patriarchal institutions as *rapist*, 'characterized by invasion, violation, degradation, objectification, and destruction of women and nature; the fundamental paradigm of sexism, racism, classism, speciesism, and all other oppressive-isms (1998, p. 8). Not unlike the radical Marxist/socialist feminist's view that the sexed division of labor lies at the root of economic exploitation, Daly argues that at the root of all forms of oppression lies that sexual domination—the male prerogative to rape—on the basis of which degradation and objectification, in short, *enslavement*, becomes the naturalized condition of virtually all nonmale, nonwhite, nonheterosexual, non-Western, nonhuman, nonaffluent, persons.

If Daly is correct, (a) all forms of oppression and exploitation take sex as their essential archetype. Hence, (b) feminism's emancipatory objectives cannot be achieved without serious attention paid to the ways in which sexism informs the other 'isms' which sustain a patriarchy that (c) is malignant in all of its past and present forms, and whose beneficiaries, primarily Caucasian, affluent, heterosexual males continue to have a stake. Finally, (d) to whatever extent sex oppression is archetypal of all oppression so too must heterosexuality become *compulsory*, that is, a characteristic of the *heteropatriarchal* way we define sex and hence sexuality such that other sexualities are either completely occluded or, when acknowledged at all, identified with the perverse and unnatural. Within compulsory heterosexuality, an action only counts as sex if (a) it involves a penis-penetrating-a-vagina, (b) it is compulsory in the sense that no other option counts as an acceptable choice, and (c) it negatively associates abstinence with frigidity, mental illness, or homosexuality. On this view, then, the domination of sex cannot be accomplished without the domination of sexuality. Given the combined compulsion of (a)–(c), moreover, it is little wonder that Daly characterizes heteropatriarchy as rapist urging the view that re-

sistance to patriarchy must be accompanied by resistance to heterosexuality. Radical feminism thus becomes radical *lesbian* feminism as both the rejection of male sexual access and as the radical embrace of the value, lives, and bodies of *women*. Such is the meaning of dismantling the master's house with *new tools*. That is, radical lesbian feminists (or radical feminist lesbians) argue for two distinct but intimately related claims: first that compulsory heterosexuality must be repudiated, and second, that lesbianism must itself be theorized as a positive, life-affirming, morally responsible choice.

One potentially compelling objection to this view is that it is surely possible to embrace the political goals of lesbian feminism without jettisoning sexual relationships with sympathetic men. However, because even the best of men are still disproportionately empowered in heterosexual relationships such intercourse cannot fail to remain compulsory, and hence self-defeating. This is not to say that heterosexual sex either is or is not *inherently* harmful to women, but rather that because male sexual access can be identified as one of the primary sources of women's oppression, cauterizing this source of harm is instrumental to ending heteropatriarchal power. Indeed, whether heterosexual sex is inherently harmful or not is largely beside the point given that from a radical lesbian feminist point of view it is hard to imagine conditions under which the potential for harm could be sufficiently deflated, given men's history of sexual and domestic violence, to make the risk worth taking. Nothing, moreover, requires us to theorize men's 'essential' natures as aggressive or women's as passive; it is enough to simply recognize that history and culture are so saturated with the glorification of male sexual conquest and violence, that the prospects for change, and the likelihood of that change being initiated by men is less than compelling. Not unlike Marx's conception of the permanent revolution of the working class, radical lesbian sexual separatism constitutes much more than a political strategy; it constitutes part of the bedrock of a burgeoning *lesbian culture* in the sense that what propels this revolution is a *positive choice to live one's life among women*, not to live among women until men learn to behave better. To opt merely for the latter is tantamount to abandoning in advance the worth of having chosen the former at all.

Philosophers such as Sarah Hoagland, for example, argue that a constitutive feature of a coherent lesbian ethic includes the repudiation of all forms of domination and subordination. To the extent, then, that heterosexuality is entrenched in and defined by these values, it can find few comfortable bedfellows—literally or figuratively—among radical lesbian feminists. Is a nonoppressive, nondominating heterosexuality possible? In theory, no doubt it is. But, as feminist philosopher Marilyn Frye points out, the practical likelihood of this given the stakes men have and are likely to continue to have in heteropatriarchy seems illusive at best. Lesbians, she argues, are simply impossible within the heteropatriarchal conceptual framework; indeed, they are like 'arm wrestling ducks.' That is, no arms, no wrestling, no penis, no sex. Hence while gay men can be condemned for having perverse sex, lesbians cannot have sex at all, and without the possibility of sex, there can be no lesbians. As a clever male college student was once heard to put the point: 'There can be no arm-wrestling ducks, but a good arm-wrestling fuck, now that's sex.' Frye goes on to argue that given the biological facts of reproduction as well as the social and political expectations attached to child-rearing, it is hardly a wonder that exercising control over the conditions under which pregnancy, birth, and lactation take place becomes imperative for men who are likely to experience the termination of privilege as the unjust withdrawal of something owed. As Frederick Engels had shown more than a century earlier, men's investment in sexual domination is rooted first and foremost in their investment in the products of their own (hetero)sexuality, namely, the offspring who represent a paradigmatic form of property, a reflection of self-interest, and the potential for future inheritance. To deny men sexual access, then, is in effect to deny them the opportunity to control the conditions of women's reproductive lives, and hence of women's lives in general, particularly in nations or regions where access to contraceptives and abortion is heavily regulated if not banned by heteropatriarchal laws, mores, and religion.

The affirmation of women's lives, bodies, values, and choices forms the second critical axis of any radical lesbian feminist theory. Yet, while this affirmation has often been theorized as necessitating some variety of separation from men—*separatism*—it is less clear what exactly this might mean, for whom it is realistically practicable, and to what extent it requires physical and material independence from both men and patriarchal institutions. This much, however, seems clear: the choice to no longer engage with men sexually constitutes a minimal criterion of separatism whereas the choice to embrace a life composed primarily or even exclusively of 'women-loving-women' goes much further—and is also far more difficult to realize in practice. There are several reasons why this is the case: first, because we live in a capitalist economy, regionally, nationally, and globally, wherewithal *is* economic wherewithal. Hence, the creation of any women-only space, permanently (a communal farm, for instance) or temporarily (such as women's music festivals) is an economic undertaking which, in virtue of this, is always at risk of excluding some in favor of those who can afford the price of separatism. As some lesbian feminist women-of-color point out, separatism continues to be a luxury of whiteness to whatever extent economic affluence does. Second, because every institution, economic, social, cultural, medical, to which we have access is imbued with heteropatriarchal values, it remains somewhat obscure

what separatism could mean short of wholesale abandonment. Third, short of the invention of parthenogenesis, a contribution of male-produced sperm is still a requisite of human reproduction.

Such a characterization of separatism may, however, be guilty of a straw fallacy, that is, perhaps separatism appears impracticable only because its parameters have been defined too narrowly or too simplistically. Whether this is so, however, depends precisely on what kind of conceptual work such a notion is supposed to be able to do, for whom, and for what duration. The range of other potentially separatist choices is very wide and includes at least the following: the formation of one's primary affective relationships, sexual or otherwise, with women and girls only, choosing a woman-centered spirituality, adopting and rearing girls within a lesbian culture, creating woman-centered works of art, dance, music, poetry, comedy, film, philosophy, and literature, rejecting other male-centered practices such as nonhuman animal flesh consumption, hunting, or other hierarchical forms of organization, choosing political affiliations and activities that are consistent with a woman-centered life. The difficulty here, of course, is twofold: first, how ought we to define what counts as woman-centered or lesbian (and do these mean the same thing?)? How do we avoid reproducing the same forms of oppression which so disable us within heteropatriarchy? (Should preoperative transsexual men-to-women be allowed into women-only spaces? Can a woman-centered life include rearing sons? Should radical lesbian feminists form political alliances with gay men? If so, should we sign on to their political causes such as the promotion of legalized marriage for gays and lesbians?)

Second, if 'separatist' is itself subject to evolving cultural, ethnic, and historical interpretation, surely it is as dangerous to treat it ahistorically as it is to treat 'woman,' 'man,' and so on. What, then, does separatism require beyond the refusal of sexual access to men? Does anyone's refusal matter more than others? Can lifelong lesbians who do not identify as feminists be separatists? Can women who have left long-standing heterosexual relationships and now identify as lesbian qualify as separatists in mid-life? How separate is separate? The point, of course, is that many questions attend the endeavor to create new tools, and much more attention needs to be paid to the ways in which oppressions related not only to sex, but to race, ability, age, gender expression, and class intersect with compulsory heterosexuality. There is much to worry and much to celebrate. Nonetheless, if the radical lesbian feminist critique of capitalist heteropatriarchy is correct, this endeavor is more than just important; it is necessary to the emancipation of women.

See also: Evolutionism, Including Social Darwinism; Feminist Ethics; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology

Bibliography

Allen J (ed.) 1986 *Lesbian Philosophy: Explorations*. Institute for Lesbian Studies, Palo Alto, CA

Card C (ed.) 1991 *Feminist Ethics*. University of Kansas Press, Lawrence, KS

Daly M 1998 *Quintessence: A Radical Feminist Manifesto*. Beacon Press, Boston

Frye M 1983 *The Politics of Reality: Essays in Feminist Theory*. Crossing Press, Freedom, CA

Hartmann H 1990 Capitalism, patriarchy, and job segregation by sex. In: Hansen K V, Phillipson I J (eds.) *Women, Class, and The Feminist Imagination*. Temple University Press, Philadelphia, PA

Hoagland S 1988 *Lesbian Ethics: Toward New Value*. Institute of Lesbian Studies, Palo Alto, CA

hooks b 2000 *Feminist Theory From Margin to Center*. South End Press, Cambridge, MA

Lee-Lampshire W 1999 Spilling all over the wide fields of our passions. *Hypatia* **14**(3): 1–16

Lorde A 1984 *Uses of the Erotic: The Erotic as Power*. In: Lorde A (ed.) *Sister Outsider*, Crossing Press, Trumansburg, NY

Lugones M 1987 Playfulness, world-traveling, and loving perception. *Hypatia* **2**(2): 3–19

Rich A 1994 Compulsory heterosexuality and lesbian existence. In: Jaggar A (ed.) *Living with Contradictions*. Westview Press, Boulder, CO

W. L. Lee

Fertility and Culture: Anthropological Insights

In its narrowest meaning, fertility refers to the production of a living child that adds one person to a population. Between this biological event and the level of culture, which defines and shapes the events that produce a child, lies an engaging set of issues. The mainstream of anthropology has never been interested in the question of numbers of children in the way that demography has, in its focus on the stocks and flows of population. Conversely, the anthropological penchant for rich, contextualized cultural meaning draws limited interest from demographers seeking parsimonious ways to explain fertility levels. Still, the

two disciplines are finding increasing common ground in the domain of reproduction. This article describes anthropological attempts to derive cultural insight in studies of fertility.

Demography seeks to identify and measure key variables related to population and to demonstrate relations among them. But to frame its questions and explain the results, it has tended in the past to draw on other disciplines, particularly on sociology, economics, and history. In recent years, demographic interests have turned as well to anthropology and to its methods. There have been two principal reasons for this growth of interest. The first was the impact of the Princeton European Fertility Project (Coale and Watkins 1986). The most important finding of this project was that the decline in marital fertility in Europe was explained less by the expected economic factors of income and level of development than by factors such as religion, dialect, ethnicity, and region—factors the project termed 'cultural,' although the awkwardness of the nomenclature was recognized. In Belgium, for example, villages experienced fertility decline in consonance with linguistic groupings. Walloon villages tended to experience the transition to lower fertility earlier than did Flemish villages, irrespective of their locations or economic characteristics (Lesthaeghe 1977). The implication drawn was that information diffused more readily within culturally or linguistically similar regions. The second reason why demography has turned to anthropology has been an increasing interest in patterns of fertility and mortality in developing countries. Questions were emerging particularly about the social supports for high fertility and about such tremendous variation in reproductive practices that researchers trained largely in the Western quantitative disciplines felt poorly equipped to handle. Because anthropology is the discipline that has so routinely attempted to study non-Western societies, anthropological works have drawn increasing demographic attention.

1. *What is Anthropology, and How Has it Been Applied to Demographic Studies of Fertility?*

Anthropology seeks to understand why people behave the way they do and to understand their beliefs and values within particular socioeconomic contexts. Anthropological work begins with the premise that the cultural frames from which our understandings of human motivation and action spring are likely partial, if not altogether wrong. They are therefore misleading. This is true, no matter how familiar the culture appears to be to the researcher. In fact, the more familiar the culture appears to be, the greater the risks of being misled and the more wary the analyst should be. To a field whose practitioners start with the expectation of being wrong, most valued are data that embody local people's own ideas about the world, in contrast to those based only on the researcher's predetermined constructs. To this end, open-ended approaches are critical. A key goal is to achieve a subjective understanding: that is, to comprehend how local people's experience the world and the logic that may underlie their choices. Doing so, it is felt, should lead not simply to better insights into the cultural beliefs of people living in remote places, but to a more encompassing view of reality and hence to better theory.

Whereas anthropology was an exotic oddity 15 or 20 years ago in demographic circles, demographic studies increasingly integrate methods and concepts conventionally associated with anthropology into their enterprise. Many demographic projects have begun to supplement their survey analyses with field projects, which include an anthologist or individuals trained in ethnographic methods and qualitative methods are increasingly common in demographic project designs. Even the international Demographic and Health Surveys (the DHS) now have a 'focus group' component in which statements and explanations, rather than simply numeric and short-answer responses, can be elicited from groups of people.

1.1 *Operationalizing Cultural Traits in Statistical Taxonomies*

What anthropology has had to offer the demographic enterprise has been largely a function of the state of its own field. When significant demographic interest first turned to anthropology in the middle of the twentieth century, a particular methodological approach, sometimes called the cross-cultural method, dominated American anthropology. This was a kind of statistical taxonomy in which societies were described by traits that were typed and 'operationalized'—that is, cast in the form of discrete variables—and coded in order to test quantitative hypotheses and look for cross-cultural patterns. The primary data used in this endeavor was the Human Relations Area Files (Murdock 1975), which contained detailed codes of the organizational features of hundreds of ethnic groups. Prevalent as well in American anthropology at this time wave questions about the relationship of culture to personality from Freudian perspectives. A number of anthropologists sought to combine this theoretical perspective with the cross-cultural methodology. Whiting et al. (1958), for example argued from HRAF data that life under harsh environmental conditions required parents to protect a child by a long period of breastfeeding and a ban on sexual relations. Such practices, because they led to strong psychological attachments of the child to the mother, were said to necessitate dramatic manhood initiation rituals for boys, to separate them from their mothers and to foster the development of their masculinity. While personality development drew little demographic interest, there was considerable interest in the potentials of such studies to shed light on facts such as the

duration of breastfeeding and postpartum sexual abstinence, which had more direct connections to fertility. Applying this kind of analysis to data on kinship and economic subsistence, demographer Frank Lorimer (1954) argued that populations like those on small islands, with highly constrained and intensively utilized resources, developed types of social organization and customs that checked family size. In Polynesia, he maintained, such customs included coitus interruptus, abortion, infanticide banning further reproduction to the mother of a married son, and encouraging sea voyaging for young men.

1.2 Culture as Ideas and Mental Schemes

The taxonomic approach drew considerable interest in demography. Yet within anthropology itself, it encountered increasing antipathy. Critics charged that the approach, in extracting practices from the contexts in which they drew meaning and treating them as a checklist of discrete, modular traits, effectively froze what was in reality a multifaceted and fluid entity. Particularly with the impact of Clifford Geertz (1973), anthropology's conception of culture shifted decisively away from such approaches and moved toward one of shared mental schemes or symbols that signaled meanings and interpretations of the world. By contrast to efforts to operationalize cultural data in order to break them down into pieces and assimilate them to our own analytical categories, most anthropologists now have the opposite purpose. They stress that culture cannot be broken down into separable pieces. Such pieces would in any case have no meaning apart from a context. Anthropologists also take cultural oddities as possible signs of some facet of a reality that the analyst has not seen and hence has not coded into his or her own cultural ideas.

Anthropologist's efforts to reclaim culture from the domain of operationalizeable concepts have not always been easy to explain to other disciplines. The work of demographer John Caldwell, however, took a notable stand in its efforts to make culture figure more explicitly as ideas or attitudes. The most notable example is his description of the proliferation of Westernization: the rise of new attitudes about work, family relations, and fertility practices, particularly through schooling. By comparison to traditional values of authoritarian family structures, a distant and severe father, and marriage as an arrangement between elders, Caldwell (1980) describes infant and primary school textbooks in three African countries as attempting to convey messages of parental closeness to children: kindly, understanding fathers and love-match marriages. So powerful are such ideas that with the shift to mass education, all members of the population, including those who have not attended school, may be affected by the changes generated by a new cultural climate of an educated populace. These ideational changes, he argues, have facilitated a shift to nuclearization of families and to smaller family size goals.

The evolutionary implications associated with traditional/modern frameworks have fallen out of favor with sociocultural anthropology. Contemporary anthropologists reject any necessary assignation of labels of traditionalism or modernity, and they recognize all societies as inherently fluid and changing. Nonetheless, the move toward seeing culture as a set of shared mental schemes that can shed light on ideas and behaviors has opened up significant possibilities for understanding demographic phenomena.

An engaging example that attempts to link forms of belief to fertility behavior is *Mussallam's* (1983) account of the allowable uses of contraception for men in classical Islam. Islamic practice has often been cast by analysts as opposing the use of contraceptives. *Mussallam*, however, describes a more complex picture. Although contraception was seen as an act that fell short of religious perfection for a man, Islamic practice allowed a wide range of exceptions. Contraception was allowed to preserve health and well-being. A man could contracept to protect his wife from the dangers of childbirth, especially if she were young or sickly, or simply to protect her health and beauty from the wearing effects of childbearing. He could contracept as well to protect a young nursing child from the nutritional competition posed by a new pregnancy. Another salient domain of concern was economic welfare. A man who found himself tempted to engage in theft to meet his economic responsibilities to his children could practice contraception to limit the number of his offspring to a manageable size. Analogous concerns emerged in the desire to safeguard human property. Because a man could not sell a slave after she became the mother to his child, using contraception was seen as a lawful measure to retain liquidity of slave assets. A man could use contraception as well when having sex with another man's slave to prevent his own children from being born into slavery. Related to this was the need to safeguard the welfare of the Islamic community under conditions of warfare. If a man had sex with a woman behind enemy lines, he risked not only losing his own children by procreating but swelling the numbers of the infidels unless he took contraceptive precautions.

1.3 Culture in Flux; Culture in Contestation

Despite the intrinsic appeal of pursuing cultural pathways of logic like this, it is risky to assume that behavior reflects cultural beliefs in straightforward ways. Most contemporary anthropologists hold that people have multiple cultural templates, and that culture is a fluid set of interpretive resources that people can draw on in particular political and historical contexts. Even a body of religious canon, as powerful as its injunctures may seem, has the potential for numerous interpretations.

Among the most lively examples of new research in anthropology that demands a fluid, resourceful vision of culture is the topic of male fertility. Fertility research is by far the largest sub-discipline in contemporary demography. Yet until a very few years ago, women comprised virtually its sole objects of study. Men, if they appeared at all, were shadows, partners-by-implication of those engaged in childbearing. Even if genetic tests could verify the relationships among household members in a survey, this would hardly begin to solve the challenges surrounding the analysis of male fertility. While some men do not know how many children they have or deny connection to offspring born outside of wedlock, other men claim to have fathered children they have never seen. The fact that men so often redefine parenthood can work to the disadvantage of children, but also to their advantage. For example, Townsend (2000) shows that in rural Botswana, where the demands of migrant labor are so taxing, a young man who lives full time with his children is an oddity. At the same time, men provide considerable support for children, even though it is often not their own children but those of other men, whether these are uncles, grandfathers, or distant cousins. Thus, despite the widespread perception in the current international population policy world of massive male disengagement from parenthood responsibilities, Western cultural preoccupations with a biological definition of paternity have masked some critical variations in the cultural assignment of parenthood.

One of the most important points of debate in current socio-cultural theory that is now being brought to demography is the question of societal forces of politics and economics versus individual agency. That is, to what extent are people's choices limited by wider powers beyond their control? Can people, on the other hand, make their own decisions about reproductive actions by resisting or evading state powers? The increasing answer to this question is that both forces are invariably at work, and that cultural ideologies are often at the center of the contestation. Greenhalgh and Li's account (1995) of the heavy-handed Chinese state policy of one child per couple that was implemented in 1979 reveals a dynamic of mutual influence. Greenhalgh and Li show that women began to resist the formal one-child policy, strategizing to have two children, especially if the first was a girl. Women attempted to adopt out daughters, move to another city to give birth to a second child, or remove state-imposed intrauterine devices. They might also attempt to conceal unauthorized pregnancies until it was too late for a state-required abortion, or (as ultra-sound technology spread) engage in sex-selective abortions. By 1991, the state, faced with such resistance, was forced to rewrite policy. It 'encouraged' one child for all and prohibited third children, yet allowed two children if the first was a girl. In sum, individual women may have been liberating themselves individually from unwanted daughters, yet their strategies of resistance effectively reinforced the state's gender hierarchy.

The fact that culture is increasingly seen in anthropology as a resource that people may put to use is best examined in a context in which reproductive volition has been seen as low. For this reason, some of the best examples are found in topics related to high fertility: a pattern that has been seen in demographic convention as an outcome of biology and custom and not as the calculus of conscious choice (Coale 1973). In populations said to have what has been called natural fertility (Henry 1961), the demand for children is high, and the number of children women produce is assumed to be limited only by how quickly they can be supplied, given the biological rhythms resulting from specific customary regimes. In sub-Saharan Africa, for example, most people continue to report on surveys that they want 'all the children God gives them,' and typical reactions to suggestions to reduce fertility range from polite interest to outrage. By contrast, in societies where desires for large numbers of children have abated, couples are said to engage in parity-specific behaviors, especially the use of modern contraceptives, in order to control completed family size.

However, ethnographic accounts of practices such as infanticide, child fosterage, child abandonment, adoption, and abortion make it clear that people in all societies exert some form of control over reproduction. Whereas Western models assume that reducing fertility is the best way to address economic troubles, families in many African societies see the diverse skills and social ties that they manage to cultivate through children as the key to their future capacity to cope with economic and political adversity. Adults treat the original biological event of parenthood not as a secure hold on a child's future support, but as the beginning of a long, continuously negotiated relationship in which they try to influence children's obligations. Young children are commonly fostered to 'grannies.' Later, when they are older, they may be sent to urban households that offer training in specialized apprenticeships or a chance to attend a respected school. In short, rather than trying to reduce child numbers, many African families would find this a drastic solution to economic problems. They prefer to have a number of children, cultivating differentiated opportunities for them (Bledsoe 1994). Yet although the effects of certain practices on cutting reproductive costs may be analytically discernable, they do not necessarily constitute people's motivations for taking actions that have effects on fertility or household composition. Most anthropologists insist that it is wrong to reduce family structure to economic or demographic terms alone.

Another example of people's efforts to exert control over childbearing appears in the cultural reinterpretation of contraceptive technologies in sub-Saharan Africa (Bledsoe 1994). In Western cultural per-

ceptions, contraceptives are devices that limit fertility, and people who do not wish to limit births would be unlikely users of devices that would do so. In West Africa, however, people see contraceptives more as devices to ensure the survival of many children by spacing their births carefully than to limit births. Some of the most common contraceptive users are thus breastfeeding women who want to avoid overlapping children, one in the womb and the other nursing, but who have resumed their menses before the breastfeeding child can be weaned. As this suggests, families exert prodigious efforts to control the timing and circumstances of children's births, though, again, not necessarily to limit numbers of children.

1.4 Reconciling Local Cultural Knowledge and Western Science

Pursuing threads of counterintuitive cultural logic that are opened up by scrutinizing demographic patterns can yield not simply observations about culture but empirical truths that have been lost to contemporary scientific view. An example comes again from the use of contraception in West Africa (Bledsoe 2001).

Contemporary Western writings, both scientific and humanistic, depict the ability to bear children as one limited by time. That is, Western culture posits a time-dependent ability to conceive as the key to achieving fertility. Accordingly, the number of children a woman may bear is calculated by dividing the total amount of time she can conceive by the interval required for each pregnancy. The use of contraception plays directly into this presumed calculus. To the extent that the use of contraception or abstinence blocks off time in this potential reproductive span, a woman's fertility will remain low. In the African cultural view, however, the ability to reproduce is limited not by time but by bodily resources. Among the forces that drain bodily resources and erode the ability to reproduce, childbirth is by far the most powerful. Each birth takes a mounting toll on a woman's physical reserves, cumulatively 'aging' her body (or wearing it out, in many local translations), and eventually making further childbearing dangerous, irrespective of her chronological age. In this vision, aging is not the same thing as age. Although aging tends to be equated with the passage of time in Western analyses, its pace, in the African view, is contingent on the physical ordeals the body suffers. Difficult and/or closely spaced births, combined with a poor diet, illness, and strenuous work responsibilities, can exact a disproportionate toll. In such situations, a woman who is trying to have all the children God gives her will almost certainly run up against the anatomical and energy limits of the body before she runs into any temporal boundaries on the ability to conceive. Because contraceptives allow women more opportunity to recuperate between births, and because reproductive capacity is seen as governed by taxing events rather than by time, it is clear why contraceptives can be used to produce more, rather than fewer, children and even to stem the pace of aging.

The African vision of contingent wear and cumulative physical toll over a reproductive lifetime might appear to be 'just common sense.' It is more accurate, however, to say that our increasing personal distance from the high-and-risky end of the fertility horizon, combined with our unprecedented levels of preventive care and copious backup in case of emergencies, has sharply truncated our vision of the tolls that reproduction can take under conditions that we no longer experience or fully understand.

2. Future Possibilities for Anthropology and Demography in Studies of Reproduction

In the evolving relationship between anthropology and demography, the spectrum of topics continues to broaden. Common interests can be found in the politics of fertility control, gender relations, the negotiation of fertility decisions among spouses, and the re-opening to scrutiny of taken-for-granted associations between fertility and variables such as educational achievement. Anthropology's current interests in the new reproductive technologies holds the potential for studies of the timing and circumstances of entry into parenthood. Finally, it can be noted that although anthropology made itself as a discipline on the study of exotic people living in distant places, increasing numbers of practitioners are coming 'back home,' studying social and cultural phenomena in the urban United States and Europe. This trend may facilitate fresh collaborations with demographers on studies of reproduction in Western societies. Some of the most productive areas of common ground in studies of reproduction for the two disciplines may thus lie in studies of migration. New anthropological interests in global flows of people, ideas, and technologies are coming at the time when demographers are taking up the challenges of studying international population movements and refugees. The opening of the European Union and, in the United States, the removal of national-origin quotas for immigrants has spawned migration across national borders on an unprecedented scale. Such a mix of national ideologies and policies of social support that must underpin these dynamics as people envision a reproductive career invoke are virtually uncharted areas. It will no longer be possible to rest on conventional notions either of a population boundary or of local cultural practices. Clearly, in any case, anthropology's style of engagement with demography and the insights to be gained from such mergers will continue to change as the fields themselves change.

See also: Family Theory and the Realities of Childbearing Behavior; Fertility: Institutional and

Political Approaches; Fertility: Political and Political–Economic Perspectives; Fertility: Proximate Determinants

Bibliography

Bledsoe C 1994 Children are like young bamboo trees: Potentiality and reproduction in sub-Saharan Africa. In: Kiessling K L, Landberg H (eds.) *Population, Economic Development, and the Environment*. Oxford University Press, New York, pp. 105–38

Bledsoe C 2001 The Contingent Life Course: African Challenges to the Culture of Western Science. University of Chicago Press

Caldwell J C 1980 Mass education as a determinant of the timing of fertility decline. *Population and Development Review* **6**(2): 225–55

Coale A 1973 The demographic transition reconsidered. In: *International Union for the Scientific Study of Population* (ed.). *International Population Conference, Liege* IUSSP, Liege, Vol. 1, 53–72

Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ

Geertz C 1973 *The Interpretation of Cultures: Selected Essays by Clifford Geertz*. Basic Books, New York

Greenhalgh S, Li J L 1995 Engendering reproductive policy and practice in peasant China: For a feminist demography of reproduction. *Signs: Journal of Women in Culture and Society* **20**(3): 601–41

Henry L 1961 Some data on natural fertility. *Eugenics Quarterly* **8**: 81–91

Lesthaeghe R J 1977 *The Decline of Belgian Fertility, 1800–1970*. Princeton University Press, Princeton, NJ

Lorimer F 1954 *Culture and Human Fertility*. UNESCO, Paris

Malthus T R 1798 *An Essay on the Principle of Population*. J. Johnson, London

Murdock G P 1975 *Outline of World Cultures*, 5th edn. Human Relations Area Files, New Haven, CT

Mussallam B F 1983 *Sex and Society in Islam: Birth Control Before the Nineteenth Century*. Cambridge University Press, Cambridge, UK

Townsend N W 2000 Male fertility as a life time of relationships: Contextualizing men's biological reproduction in Botswana. In: Bledsoe C, Lerner S, Guyer J I (eds.) *Fertility and the Male Life Cycle in the Era of Fertility Decline*. Oxford University Press, Oxford, UK

Whiting J W M, Kluckhohn R, Anthony A S 1958 The function of male initiation ceremonies at puberty. In: Maccoby E E, Newcomb T, Hartley E (eds.) *Readings in Social Psychology*. Holt, New York, pp. 359–70

C. Bledsoe

Fertility Change: Quantum and Tempo

In the measurement of fertility the term 'quantum' usually refers to the average number of births occurring over the lifetime of a cohort of women (Ryder 1980, Pressat 1985). This number is called the cohort total fertility rate. More generally, quantum can refer to the average numbers of births that occur in a group of women under certain defined circumstances, such as marriage, or its measurement can be limited to births of a certain order, such as first births. The term 'tempo' usually refers to the average age at which births occur in a cohort of women who have reached the end of the childbearing years (Ryder 1980, Pressat 1985). More generally, tempo can refer to the average time elapsed between a birth and a specific preceding event (e.g., marriage) or its measurement can be limited to births of a certain order.

Estimation of the quantum and tempo of childbearing for a cohort of women is straightforward provided the required data are available and accurate. These variables quantify key dimensions of a cohort's reproductive history but they have the considerable disadvantage of representing past experience: women currently aged 50 did most of their childbearing two to three decades ago when they were in their 20s and early 30s. Trends in cohort measures therefore do not reflect current conditions, and they are not often used to monitor recent developments because fertility can change substantially in a short period of time. Instead, an alternative measure, the period total fertility rate (PTFR), is used far more widely as an indicator of current fertility. It is defined as the average number of births a woman would have if she were to live through her reproductive years (ages 15–49) and bear children at each age at the rates observed in a particular year or period. It is a hypothetical measure because no real group of women has experienced or will necessarily experience these particular rates. The advantage of the PTFR is that it gives up-to-date information on levels and trends in fertility. The PTFR is a quantum indicator of period fertility, and the mean of the age-specific fertility schedule is a tempo measure.

The simplicity and wide availability of the PTFR have contributed to a neglect of some deficiencies in this fertility indicator. The demographic literature on the measurement of fertility includes a number of criticisms of the PTFR (Ní Bhrolcháin 1992). The most important of these is the distortion of the PTFR that occurs when the timing of childbearing changes. During years in which women delay childbearing, period fertility rates are depressed relative to cohort fertility; in years when childbearing is accelerated, period fertility is raised (Hajnal 1947, Ryder 1956, 1964, 1980, 1983). The quantum and tempo dimensions of period fertility interact.

To demonstrate how these tempo distortions operate, it is helpful to examine a set of simple illustrations as presented in Fig. 1. In these hypothetical illustrations, every woman is assumed to have exactly one birth so that the cohort total fertility rate equals one for all cohorts, and these births are assumed to take place at exactly the same age within each cohort. Figure 1(a) depicts the reference situation in which this age is set at exactly 20 years. Since there are no tempo effects,

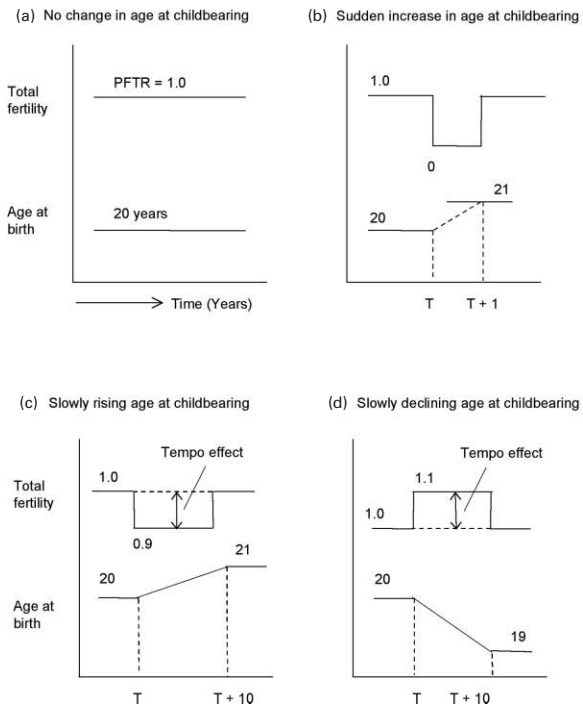

Figure 1
Hypothetical patterns of childbearing, with and without tempo effects

period fertility as measured by the PTFR is also constant and equal to one birth per woman. Assume next that the age at childbirth is raised by one year starting at time T, as shown in Fig. 1(b). In this case, women who would have given birth at age 20 immediately following T have to delay their childbearing for a year until they reach their 21st birthday. As a consequence, no births occur during the year between T and $T+1$ and the PTFR then equals 0. In other words, one year's worth of births is lost because the mean age at childbearing increased by one year. In reality, the mean age at childbearing rarely changes suddenly and rapidly. A more typical situation is summarized in Fig. 1(c), which illustrates the effect of a slow rise in the age at childbearing. Starting at time T, successive cohorts delay childbearing so that the mean age at first birth rises linearly from 20 to 21 years over a 10-year period from T to $T+10$. In this case the PTFR drops from 1 to 0.9 at time T, remains at 0.9 from T to $T+10$, and finally returns to 1 after time T. The PTFR is distorted by 10 percent or 0.1 birth per woman during the interval T to $T+10$. Over the 10-year period, one year's worth of fertility is lost from the PTFR as a result of a one-year increase in the mean age at childbearing. Conversely, as shown in Fig. 1(d), if the mean age at first birth declines linearly from 20 to 19 years over the same period of 10 years, the PTFR rises from 1 to 1.1 at time T, remains at 1.1 from T to $T+10$, and returns to 1 birth per woman after T. Bongaarts and Feeney (1998) present a fuller discussion of these effects.

The existence of timing distortions is readily documented when the age at childbearing is declining significantly. In that case, implausible results are often obtained for the birth-order components of the PTFR, computed as the sum over age of the period-specific fpjertility rates of the second kind by age and birth order (see *Demographic Techniques: Rates of the First and Second Kind*). For example, the component for births of order one ($PTFR_1$) equals the average number of first births women would have by age 50 if they were to bear first births at the age-specific rates of the second kind observed in a given year or period. In most years during the late 1940s and early 1950s, the $PTFR_1$ in the US exceeded 1 (Whelpton 1954), a fact which seems to imply that women had more than one first birth on average. This is impossible and these $PTFR_1$ estimates must therefore be reinterpreted. The main explanation of why $PTFR_1$ exceeded 1 during the baby-boom years is that the age at childbearing declined, resulting in a temporary inflation of the PTFR and its additive birth-order components:

$$PTFR = \sum_n PTFR_n \tag{1}$$

An extensive analysis of the effects of variations in the childbearing tempo on period fertility indicators was undertaken by Ryder in a series of influential papers (Ryder 1956, 1964, 1980, 1983). Ryder proposed 'translation' equations to calculate the period fertility measures from corresponding cohort measures when the timing of cohort fertility is changing. These procedures have not found wide acceptance for two main reasons. First, in his work on the translation issue, Ryder assumes that the tempo and quantum of cohort fertility are the determinants of the PTFR and other period fertility measures. However, extensive empirical analysis of this issue has demonstrated that this is not the case (Ní Bhrolcháin 1992). Period effects are clearly the prime source of variation in fertility rates, and the completed fertility of cohorts is largely determined by the period fertility rates during the cohort's reproductive years, rather than the other way around. Second, the two dimensions of aggregate cohort fertility—quantum and tempo—are not independent in practice. When cohorts reduce their fertility, they do this primarily by reducing childbearing at higher birth orders. As a result, the mean age at childbearing for all births to the cohort declines even if the timing of individual births does not change. In other words, a decline in the cohort quantum can lead to changes in the cohort's mean age at childbearing that do not cause tempo distortions. Such distortions only occur if the mean ages at births of specific orders rise or fall. Ignoring this effect—as

Ryder does in much of his writing—therefore gives biased results except when cohort fertility is constant. Fortunately, this second problem can be solved by analyzing fertility trends separately for each birth order rather than for overall fertility.

In a recent study Bongaarts and Feeney (1998) propose a procedure for removing tempo effects from the PTFR. The approach is an outgrowth of Ryder's original translation equation. The above objections to Ryder's approach are addressed by assuming that fertility is only affected by period effects, and by applying adjustments separately to the birth-order components ($PTFR_n$) of the total fertility rate rather than to the total fertility rate itself. Under these conditions, it is possible to estimate a tempo-free total fertility rate in any given year (or period) from the conventional $PTFR_n$ at each birth order n with the formula

$$PTFR'_n = PTFR_n/(1-m_n) \qquad (2)$$

where $PTFR_n$ is the observed total fertility rate component for birth order n, $PTFR'_n$ is the tempo-free total fertility rate component for birth order n, and m_n is the annual *change* in the mean age of the age-specific fertility schedule for birth order n. By dividing the observed total fertility rate by (1-m_n) at any given birth order n, we obtain an estimate of the total fertility rate that would have been observed if there had been no change in the timing of childbearing. Adding up results for the various birth orders gives the overall tempo-free total fertility rate

$$PTFR' = \sum_n PTFR'_n \qquad (3)$$

The PTFR′ equals the distortion-free quantum component of the PTFR, and the difference PTFR′–PTFR equals the tempo distortion.

According to this argument, an annual increase of one-tenth of a year in the mean age at childbearing ($m_n = 0.1$) reduces the $PTFR_n$ by 10 percent below its tempo-free level, because $PTFR_n = 0.9PTFR'_n$. Similarly, an annual decline in the mean age at a rate of just 0.1 years per year ($m_n = -0.1$) inflates the $PTFR_n$ by 10 percent. Clearly, rather modest changes in the timing of childbearing at any birth order can produce substantial changes in observed fertility.

It should be emphasized that the above simple formula for removing tempo distortions assumes that the shape of the age-specific period fertility rates at each birth order does not change over time. Kohler and Philipov (2001) have suggested a more general equation that incorporates variance effects. Their formula reduces to the simpler version presented here when the shape of the fertility schedule is invariant.

The existence of tempo distortions of period fertility measures has recently become an important issue in contemporary developed countries. In most of these societies, fertility has dropped below the replacement level. If future fertility remains at these low levels, populations will decline in size and age rapidly. The fact that ages at childbearing have been rising quite rapidly since the 1970s implies that the PTFRs of these countries contain a downward tempo distortion. The size of this distortion can be estimated with the above equation, but the required data—age-specific fertility rates by single age and birth order for successive years—are available only for a restricted number of countries. Bongaarts (1999) estimates the tempo effects for seven post-transitional societies in the late 1980s and finds downward distortions ranging up to 0.4 births per woman. This effect may appear modest, but in some cases it is the primary reason why the PTFR is below the replacement level.

Assessment of fertility levels and trends should include an analysis of the tempo effects because the conventional PTFR gives a misleading estimate of the actual rate of childbearing when the age at childbearing is changing. This is the case in many post-transitional populations. Childbearing is being delayed and women are actually bearing more children than is indicated by the observed PTFRs. This distortion in the PTFR will continue for as long as the mean age at childbearing rises, but once the deferment ends the distortion is removed, and the PTFR will then rise unless there are offsetting declines in the quantum component of period fertility.

See also: Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Reproductive Rights in Developing Nations

Bibliography

Bongaarts J 1999 Fertility decline in the developed world: Where will it end? *American Economic Review* **89**: 256–60

Bongaarts J, Feeney G 1998 On the tempo and quantum of fertility. *Population and Development Review* **24**: 271–91

Hajnal J 1947 The analysis of birth statistics in the light of the recent international recovery of the birth-rate. *Population Studies* **1**: 137–64

Kohler H-P, Philipov D 2001 Variance effects in the Bongaarts-Feeney formula. *Demography* **38**: 1–16

Ní Bhrolcháin M 1992 Period paramount? A critique of the cohort approach to fertility. *Population and Development Review* **18**: 599–629

Pressat R 1985 *The Dictionary of Demography*. Basil Blackwell, Oxford, UK

Ryder N B 1956 Problems of trend determination during a transition in fertility. *Milbank Memorial Fund Quarterly* **34**: 5–21

Ryder N B 1964 The process of demographic translation. *Demography* **1**: 74–82

Ryder N B 1980 Components of temporal variations in American fertility. In: Hiorns R W (ed.) *Demographic Patterns in Developed Societies*. Taylor and Francis, London, pp. 15–54

Ryder N B 1983 Cohort and period measures of changing fertility. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries*. Academic Press, New York, pp. 737–56

Whelpton P K 1954 *Cohort Fertility: Native White Women in the United States*. Princeton University Press, Princeton, NJ [Reissued 1973 Kennikat Press, Port Washington, NY]

J. Bongaarts

Fertility Control: Eastern Europe

This article reviews the decline of fertility in the countries of Eastern Europe, broadly defined to include all regions of the former Soviet Union. Direct data on contraceptive utilization which have recently become available are examined, together with indirect measures of fertility limitation.

1. *Total Fertility*

Fertility levels in Eastern Europe and the former Soviet Union span most of the range of variation observed around the world (see Fig. 1). As recently as the mid-1970s, Uzbekistan, Turkmenistan, and Tajikistan registered total fertility rates (TFRs) from six to seven, well within the range of high-fertility Third World countries. At earlier dates (in the 1950s) this was also true of Albania and Azerbaijan. However, the vast majority of the countries in question qualify as low-fertility countries, having experienced TFRs on the order of two to three for most of the period since World War II. In general, all of the countries in question have experienced fertility decline during the postwar period. This includes the period after the dissolution of the Warsaw Pact and the breakup of the former Soviet Union.

It is possible to group the countries under consideration according to their phase of the 'Demographic Transition' from high, natural fertility to low, controlled fertility (Anichkin and Vishnevskiy 1994, Vishnevskiy and Volkov 1983). For the purposes of the present discussion, 'low fertility' is taken to refer to fertility at or below the 'replacement level,' which is the level of fertility that would be consistent with a stationary population in the long run; a TFR of 2.1 is considered conventionally to be equivalent to the replacement level. The vast majority of East European countries, together with the European republics of the former USSR, were in the intermediate stages of the transition prior to World War II, and completed the transition during the postwar period. However, certain of these countries, such as Estonia, had actually

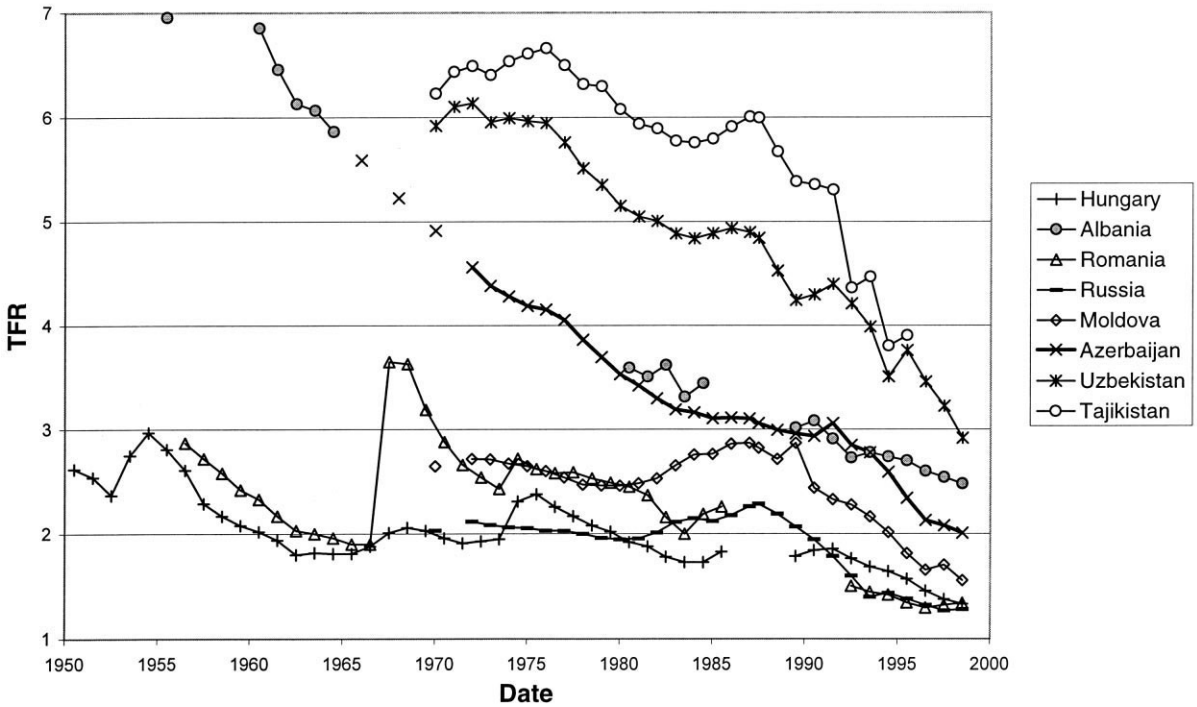

Figure 1
Total fertility rates in selected countries of Eastern Europe and the former Soviet Union

achieved below-replacement fertility prior to World War II. Azerbaijan has completed the transition from high to replacement-level fertility since 1959. The trend in Albania's fertility level resembles closely that of Azerbaijan, although Albania's fertility has not fallen quite to the replacement level. At the opposite extreme, in the indigenous population of former Soviet Central Asia, fertility decline did not take hold until the mid-1970s. These populations remain in the early phases of the demographic transition, exhibiting sustained fertility decline.

There is little evidence of any fundamental difference between the fertility levels and trends of the countries which comprised the former Soviet Union, and other East European countries (see *Fertility Trends in the Formerly Socialist Countries of Europe*). Fertility levels in the European former Soviet republics have fallen below replacement. Russia and Latvia, whose respective TFRs were 1.28 and 1.09 as of 1998, figure among the lowest-fertility countries in Europe. The only major difference between these regions (former USSR and East Europe) at the time of writing is that the former Soviet Central Asian populations, in which fertility decline is a comparatively recent phenomenon, have not yet attained the low fertility levels which typify the European populations. The fertility levels of most of the former Soviet republics discussed here are adjusted for underregistration of births according to the procedures of the International Programs Center of the US Bureau of the Census.

Both the low-fertility formerly Socialist countries of Eastern Europe and the republics of the former USSR, whose governments were pronatalist in orientation, experimented with policy measures to stimulate childbearing (David and McIntyre 1981, David and Skilogianis 1999). The East European countries were the first to introduce these measures, and did so in the late 1960s and the 1970s. In most cases (e.g., Bulgaria, Hungary, Poland), these measures involved financial incentives such as extended maternity leave, cash awards, and housing subsidies. Typically, the introduction of such measures was followed by a modest rise in fertility of relatively brief duration, and a longer-term slackening or cessation of fertility decline. These pronatalist policies were generally unsuccessful in bringing about increases in fertility to levels substantially above replacement. Only Romania achieved a major rise in TFR, which rose spectacularly from below replacement to over 3.5, then stabilized in the neighborhood of 2.5 for more than ten years. The Romanian achievement owed much to the absence of modern methods of family planning coupled with a draconian ban on abortions, enforced to the extent of subjecting women to regular gynecological examinations to ensure that they were not employing any modern means of fertility limitation. Once these restrictions were lifted, Romania's fertility declined to a level close to that of Russia, one of Europe's lowest.

The Soviet Union introduced a major set of pronatalist incentives in 1981. Shortly afterwards, fertility began to rise in the European republics of the USSR, together with Armenia and Georgia. Intriguingly, the fertility levels of these countries converged to approximately the replacement level in the mid-1980s. In Azerbaijan, fertility decline was halted temporarily in the mid-1980s. Concurrently, levels of fertility rose somewhat in Tajikistan, and to a lesser extent in Turkmenistan and Uzbekistan. Although various legislative articles were implemented to enhance the effect of the pronatalist measures throughout the 1980s, they succumbed ultimately to inflation and to the collapse of the Soviet economy. By 1990, the TFRs of the European Soviet republics had returned to approximately their 1980 levels, while fertility decline had resumed in Soviet Central Asia. After the breakup of the USSR, fertility levels continued to decline in all former Soviet republics (Vishnevskiy 1999).

2. *Contraceptive Utilization*

Direct data on utilization of contraceptive methods in the countries of Eastern Europe and the former Soviet Union is available in many cases from surveys conducted at various points since World War II, especially recently. Table 1 summarizes the data available from the International Demographic Database maintained by the US Bureau of the Census. Two aggregate categories of contraceptive methods are distinguished. The category of 'All' methods of contraception includes traditional methods such as abstinence and withdrawal, whereas 'Modern' methods include the contraceptive pill, IUDs, condoms, sterilization, and a variety of other chemical and mechanical interventions. According to the figures, in all of the East European and former Soviet countries with available data on both forms of contraception, a majority of the population of reproductive-age women were practicing contraception in one form or another at all dates of observation. However, the same cannot be said with regard to modern means of contraception: In most instances, less than half the population were using modern methods.

The data in Table 1 depart from conventional expectations in that the East European countries register levels of modern contraceptive utilization that are usually no greater than those of the former Soviet countries. In several instances, the East European levels are substantially lower than those of former Soviet countries. The levels of modern contraceptive utilization exhibited by Bulgaria and Romania in the 1970s are comparable to that observed in Zaire in 1991. Although this is to some extent a function of time, as access to modern contraceptives improved after the dissolution of the Warsaw Pact and the former USSR, the level of modern contraceptive utilization registered by Romania in 1993 remained well below what is typical for industrialized countries.

Table 1
Percentage of reproductive-age women using contraception

Country	Year(s)	All methods	Modern methods
Bulgaria	1976	76.0	8.0
Czech Republic	1993	68.9	44.8
Hungary	1958	58.0	18.0
Hungary	1974	74.0	44.0
Hungary	1993	84.4	74.6
Poland	1972	60.0	12.0
Poland	1977	75.0	26.0
Romania	1978	58.0	5.0
Romania	1993	57.3	13.9
Slovakia	1991	74.0	42.0
Slovenia	1989	91.6	56.8
Russia	1994	66.8	48.6
Belarus	1995	50.4	42.1
Estonia	1994	70.0	56.0
Latvia	1995	67.8	55.5
Lithuania	1994–5	65.9	22.1
Moldova	1997	73.7	50.1
Kazakhstan	1995	59.1	46.1
Kazakhstan	1999	66.1	54.4
Kyrgyzstan	1997	59.5	48.8
Uzbekistan	1996	55.6	51.3
Congo (Kinshasa)	1991	8.0	2.0
France	1972	64.0	21.0
France	1998	79.5	73.8
Germany	1992	75.0	72.0
United States	1955	58.8	47.8
United States	1988	74.3	69.0

Source: US Bureau of the Census, International Programs Center.

With regard to the former Soviet countries, the data on contraceptive utilization also depart from standard expectations in that the high-fertility Central Asian countries exhibit levels of modern contraceptive utilization as high as, or higher than, a number of European former Soviet republics with much lower fertility levels. Perhaps this reflects the emphasis placed on fertility reduction in these republics in the population policies of the former Soviet government, as well as those of several international organizations that have been active in promoting contraception in Central Asia after the breakup of the USSR.

The above discussion raises the question of the extent to which contraceptive utilization as measured in available surveys relates to realized fertility levels. The answer to this question is positive with respect to modern means of contraception: use of modern contraception has a decidedly negative and statistically significant correlation ($r = -0.534$) with the total fertility rates registered by the countries for which data are available at the appropriate dates. However, when traditional means of contraception are included, the relationship deteriorates into a minor positive correlation ($r = 0.153$) which is not statistically significant.

3. *Overall Fertility Limitation*

The limited coverage of, and unexpected results obtained with, the available direct data on contraceptive utilization in the East European and former Soviet countries justifies the use of indirect estimation methods to assess the extent of fertility reduction in a fuller range of countries and dates. Table 2 contains estimates of Coale's *m* (Coale and Trussell 1974) as of 1995 for those countries for which there were sufficient data available to permit calculation of this measure. The results leave an overall impression of greater conformity with general expectations than obtains for direct measures of contraception. In particular, the Czech Republic exhibits a higher level of fertility limitation, as reflected by a higher value of *m* than any of the former Soviet countries, while Hungary and Slovakia register levels comparable to the Slavic former Soviet countries (Russia, Ukraine, Belarus), along with Moldova. These former Soviet republics, in turn, register higher levels of fertility limitation than any of the Transcaucasian or Central Asian former Soviet republics. The principal departure from standard expectations is that the Baltic countries of the former USSR (Estonia, Latvia, Lithuania) register lower values of *m* than one would anticipate. This may partly reflect the fact that these countries had achieved

Table 2
Index of fertility limitation (*m*) for former Soviet and certain East European countries, 1995

Country	Index
Hungary	2.294
Czech Republic	2.575
Slovak Republic	2.252
Russia	2.365
Ukraine	2.379
Belarus	2.402
Moldova	2.175
Estonia	1.625
Latvia	1.918
Lithuania	2.026
Georgia	1.872
Armenia	2.026
Azerbaijan	1.921
Kyrgyzstan	1.300
Kazakhstan	1.772
Uzbekistan	1.977
Turkmenistan	1.384
Tajikistan	0.769

Note: Figures calculated from data maintained by the US Census Bureau, International Programs Center

high levels of fertility control earlier during the former Soviet period, and that the recent further reductions experienced in these countries have caused the resultant marital fertility schedules to deviate from the model age pattern of fertility regulation.

See also: Eastern European Studies: History; Eastern European Studies: Society; Fertility Control: Overview; Fertility: Proximate Determinants; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Reproductive Rights in Affluent Nations; Revolutions of 1989–90 in Eastern Central Europe; Soviet Studies: Society

Bibliography

Anichkin A A, Vishnevskiy A G 1994 Three types of fertility behavior in the USSR. In: Lutz W, Sherbov S, Volkov A (eds.) *Demographic Trends and Patterns in the Soviet Union before 1991*. Routledge, London

Coale A J, Trussell J T 1974 Model fertility schedules: Variations in the age structure of childbearing in human populations. *Population Index* **40**(2): 185–258

David H P, McIntyre R J 1981 *Reproductive Behavior: Central and East European Experience*. Springer, New York

David H P, Skilogianis J 1999 *From Abortion To Contraception: A Resource to Public Policies and Reproductive Behavior in Central and Eastern Europe from 1917 to the Present*. Greenwood Press, Westport, CT

Vishnevskiy A G 1999 Demographic processes in the post-Soviet states *Revue Baltique*, No. 13: 23–37

Vishnevskiy A G, Volkov A G 1983 *Воспроизводство населения СССР* (Reproduction of the Population of the USSR). Nauka, Moscow

W. W. Kingkade

Fertility Control: Overview

The primary immediate, or proximate, biological and behavioral determinants of the levels and patterns of human fertility are (a) the extent and effectiveness of contraceptive use; (b) the degree to which abortion is practiced; (c) the intensity and duration of breastfeeding behavior; and (d) the degree of exposure to sexual intercourse, often proxied by various measures of the extent of marriage. The levels of the various proximate determinants are, in turn, functions of social structure, which differs across populations and over time (Davis and Blake 1956).

The term 'fertility control' can be taken to refer to any or all of the proximate determinants. However, while marriage patterns clearly affect fertility, it is often thought that marriage behavior is motivated and explained by factors that are not related directly to a desire to control fertility. The same could be considered true regarding breastfeeding behavior. Thus, some demographers would define 'fertility control' only in terms of those behaviors known to be motivated by a desire to control fertility: contraception and abortion. The primary focus of this article is to provide an overview of these two determinants of fertility.

1. A Variety of Reasons Humans Have Desired to Control Fertility

Judging from what is known about nineteenth- and twentieth-century preliterate populations, as well as from a variety of written evidence left behind by ancient through early modern societies, it appears that men and women have been motivated to control fertility—so that there would not be too few, nor too many children—for at least the last several thousand years (Himes 1970). Additional evidence of the (postnatal) desire to control fertility is seen in the resort to infanticide, child negligence, and child abandonment in the ancient world and throughout much of the Middle Ages and beyond (see *Infanticide and Induced Abortion*). The focus here is on the various reasons humans have desired to limit fertility.

Health concerns have always been an issue; pregnancy and childbearing still present serious medical risks to women in much of the developing world, as they have until fairly recently in the currently industrialized world. Even if continuous pregnancy and childbirth did not prove fatal, the sickness, pain, and suffering involved were motivation enough for wanting to postpone, avoid, or terminate the next pregnancy. A rapid pace of childbearing presented risks to surviving children too, particularly the youngest, since it taxes parents' ability to care for their family. For example, birth spacing was seen as desirable in order to prolong breastfeeding, often considered important for a child's health and development.

In addition, there has often been the motivation to prevent or terminate pregnancy that resulted from extramarital or premarital sexual relations, either because such relations were socially sanctioned or because the consequent childbearing was deemed undesirable. Offspring born as a consequence of such sexual relations could prove extremely inconvenient to the fathers (and their legitimate children) and devastating to the mothers.

Within marriage, there has also been the desire to space and limit the number of births, for various social and economic reasons (see e.g., McLaren 1990). In terms of the timing of fertility, there was a desire to avoid childbearing during the busy agricultural seasons, since women would be working hard in the fields. In other contexts, parents who wished to avoid dividing their estates into too many parts may have wanted to limit the number of their descendants. Later on, upwardly mobile middle-class and working-class families may have chosen to have smaller families

because of increasing costs of raising children or in order to invest more in the education and health care of each child. While there have been many reasons for men and women to want to control their fertility in different periods and in different cultures, that desire is found in all societies and is very old indeed.

2. *A Variety of Fertility Control Methods*

Over the thousands of years that humans have desired to exercise control over their fertility, various methods have been developed, many through trial and error, some partially effectively, some highly effective, and some not effective at all. Some methods have been medically harmful to the women who used them, while others harmless. In this section, descriptions of the types of methods used, their historical significance, mechanisms of action, degrees of efficacy, and where possible, extents of use in the past, are presented.

Regarding efficacy, it is important to keep in mind that a contraceptive method which does not guarantee that any one woman will not conceive while the method is applied, can nonetheless have a profound effect on the overall probability of conceiving in a population. To see this, consider the following: It is estimated that if 100 couples exposed to vaginal intercourse do not use any form of contraception, approximately 85 of the women will become pregnant within one year. A contraceptive method such as *coitus interruptus*—one which is not considered by modern standards to be among the most effective—dramatically reduces the annual risk of pregnancy from 85 percent to approximately 19 percent. In this article, references will be made, when possible, to contraceptive failure rates which indicate the proportion of women who would become pregnant during the first year of typical use among couples in the United States. Typical use refers to method use which may be imperfect or inconsistent. The failure rates cited should always be compared with the probability of pregnancy when no method is used—that is, approximately 85 percent. The reader should also keep in mind that 'perfect-use' failure rates—the proportion of women who would become pregnant within one year of correct and consistent use of a specific method of contraception— are often substantially lower than typical-use failure rates (Hatcher et al. 1998).

2.1 *Methods Which are Completely Ineffective*

The application of ineffective methods provide evidence for the widespread desire and hope for fertility control, rather than for human ability to translate those hopes into effective action. A long list of amulets, symbolic magic, and rituals that have no contraceptive or abortive effect have nonetheless been known and practiced in ancient and contemporary populations. For example, written evidence from the Middle Ages in Europe describe a belief that lying or sitting down upon fingers will prevent conception for a number of years that is equal to the number of fingers. In addition, it has long been believed that the passivity of women during sexual intercourse, or their 'holding back' an orgasm will prevent conception. A wide variety of herbal preparations have also been used with the intent of fertility control; doubt remains as to whether any of these potions do indeed have contraceptive or abortive effects in humans (Riddle 1997, Santow 1998).

2.2 *Behavioral Methods of Contraception*

This section reviews methods of preventing contraception which do not require the use of any type of physical or chemical substance, but which rather rely on the modification of sexual and reproductive behavior. In this category fall some of the oldest effective methods known. Reduced coital frequency, even if practiced without regard to the timing of sexual intercourse during the menstrual cycle, will affect the monthly probability of conception to an extent which is roughly proportional to the percentage decline in coital frequency (Trussell 1979). Clearly, sexual abstinence, or continence—that is, refraining from vaginal intercourse—reduces fertility. David and Sanderson (1986) argue that sexual restraint of these kinds were important factors in bringing about the nineteenth-century decline in fertility in the United States.

Fertility awareness methods of periodic abstinence, by which the fertile days during the cycle are correctly identified and avoided, will be more effective than a simple overall reduction in coital frequency. Various types of methods are in use today (e.g., Calendar Rhythm Method, Basal Body Temperature method, Cervical Mucus charting), but historically, these are fairly new developments. It was not until the early 1930s that the timing of ovulation within the menstrual cycle was correctly identified and a chart was worked out for use by women desiring to identify the days it was 'safe' to have sexual intercourse. Typical-use failure rates of periodic abstinence hover around 25 percent. Prior to the second quarter of the twentieth century, attempts at periodic abstinence, based on incorrect beliefs about the timing of the fertile period, while often discussed, were generally ineffective.

Prolonged and intensive lactation have the effect of delaying the return of ovulation after a live birth. There is evidence that certain preliterate societies, as well as the ancient Egyptians, were aware of the fertility-inhibiting function of breastfeeding, and practiced prolonged lactation for the purpose of increasing the spacing between births. Such practices were com-

mon in the currently developed nations in the 1900s, and they are still in use there among various religious groups. In addition, throughout history, it has often been believed that semen would spoil the milk of lactating women. Customs of post-partum abstinence during lactation, more common in polygynous societies, enhance the fertility-inhibiting effect of breastfeeding, whether or not they are practiced for the purpose of fertility control (see *Fertility Control: Prevalence and Consequences of Breastfeeding*).

Modifications of sexual practices, such as oral sex, anal sex and *coitus reservatus*—vaginal penetration without ejaculation—may have been used for the purposes of preventing conception. Probably the ancient Peruvians, Greeks, and Hebrews, among others, performed such practices. However, the behavioral modification invoked most universally, both in terms of time and space, and most certainly for the purposes of fertility control, is *coitus interruptus*, or withdrawal, whereby male ejaculation occurs outside the vaginal cavity, following vaginal penetration.

Evidence indicates that preliterate populations, the ancient Hebrews, ancient Islamic societies, and others used withdrawal. The Bible mentions *coitus interruptus* in *Genesis* 38:7–10 in the story of Er and Onan. Numerous references to the practice can be found afterwards, some in favor, many denouncing it. Santow (1995) provides a compelling argument for the demographic importance of *coitus interruptus*, even centuries before the modern transition to low fertility. There is no doubt that its use was very important in bringing about the decline in fertility in many parts of Europe during the last part of the nineteenth century and early part of the twentieth century. Even in countries like the USA, where its prevalence declined relatively early, *coitus interruptus* was among the most frequently cited, if not the most cited, method of birth control reported in studies undertaken in the 1930s. In France and Belgium, significant reliance on withdrawal continued into the 1970s. Today, *coitus interruptus* remains a frequently-used method of birth control only in some parts of Southern and Eastern Europe, as well as in parts of the developing world, such as West Asia (Santow 1993).

2.3 Physical, Mechanical, and Chemical Means of Contraception

Various techniques have been developed over time to remove semen from the vagina after ejaculation. Physical methods of sperm removal cited in early literature include such means as the woman squatting and straining, or jumping following intercourse (Himes 1970). Probably the most widely known and used technique for sperm removal in past times was douching, referred to, for example, in Hindu writing of around 1,500 BC. Some primitive douches have an additional spermicidal effect, such as those which use solutions of vinegar, salt, lemon juice, or tannic acid. Charles Knowlton, an early American birth control advocate, recommended postcoital douching with a solution of zinc and alum in his 1832 book *The Fruits of Philosophy*. David and Sanderson (1986) suggest that douching was a method frequently used by urban American birth controllers in the early nineteenth century, perhaps in conjunction with reduced coital frequency and abstinence. Woycke (1988) also writes that douching was popular in nineteenth-century Germany, partly because it could be seen as a having a legitimate hygienic function, even by those ambivalent about the moral rectitude of pregnancy prevention. Typical-use failure rates of douching are probably close to 40 percent (Kestelman and Trussell 1991).

Other physical and mechanical means of contraception include those that attempt to prevent entry of sperm into the cervix. Barrier methods described by ancient Greek writers include smearing of oils on the cervix. Similar strategies are discussed in ancient Indian writings. A description of the use of crocodile dung as a pessary appears as early as 1850 BC in the ancient Egyptian Petri papyrus. In other references, alternative sticky substances, such as honey or melted beeswax were suggested for use as pessaries (Himes 1970).

A tuft of wool or cotton, frequently referred to in some Jewish theological writings (Talmud) as a *mokh*, was probably placed in the vagina before coitus or after coitus, in order to absorb the semen and prevent it from entering the cervix (Feldman 1995). In certain contexts, the wool or cotton was used as a vehicle for some type of sticky substance or primitive spermicide. Modern-day sponges are small and pillow-shaped, made of polyurethane, imbued with spermicide, and have typical-use failure rates of approximately 20 percent among nulliparous women, 40 percent among parous women.

Another important barrier method of birth control, the male condom, also has a long history. In use in various forms by preliterate peoples as well as early Egyptians for noncontraceptive purposes (e.g., for decoration), condoms became associated with prostitution during the sixteenth century, when they were used primarily for protection against sexually transmitted diseases. Even today in some contexts, condoms carry a negative stigma associated with promiscuity or prostitution. At least from the second century AD, condoms were often made from animal membranes. It was not until the middle of the nineteenth century that the development of the vulcanization of rubber allowed for production of less expensive condoms. However, even rubber condoms were probably too expensive in the mid-nineteenth century for widespread use by the lower classes. In the 1930s, the improved technology that allowed for latex condoms made condoms even cheaper and more appealing for use. Their popularity spread, so that

surveys of the 1930s show condoms competing with withdrawal for popularity in the USA. At the beginning of the twenty-first century, unlike in the past, condoms are high quality and reliable, so that breakage is rare. However, typical-use failure rates approximate 14 percent, largely because of inconsistent use.

Various other barrier methods, forerunners of the modern-day cervical caps and diaphragms, included the hollowed-out rinds of lemons, discussed, for example, by Casanova in the eighteenth century. The vaginal barriers cover the cervix in order to prevent the entrance of sperm. The rubber cervical cap and diaphragms were developed and perfected by German physicians and scientists in the nineteenth century—the cap by F. A. Wilde in 1838 and the diaphragm by W. Mensinga in 1880. The use of the latter method spread from Germany to Holland, and then to England, where it became known as the 'Dutch Cap.' Use of the diaphragm and cap have been supplemented with various spermicidal elements, such as quinine compounds, and later, modern spermicidal creams, introduced into the interior cavity of the objects before being inserted into the vagina or placed on the cervix. Promoted by birth control propagandists both in England and in the USA in the 1920s and 1930s—Dr. Marie Stopes in England advocated use of the cervical cap, Margaret Sanger the diaphragm in the USA—use of these vaginal barrier methods was limited largely to the middle classes, but apparently well accepted in this sector of the population. Modern-day diaphragms, used in conjunction with spermicidal cream or jelly, have typical-use failure rates in the range of 20 percent. Failure rates for the cervical cap are generally higher, especially among parous women.

2.4 Hormonal Methods

The search for an oral contraceptive has a long history in herbal formulations. However, only in the mid-1920s did scientists discover the antifertility effects of estrogenic compounds, and it was not until 1960 that the United States Food and Drug Administration approved the first hormonal oral contraceptive. Modern-day combined oral contraceptives, among the most extensively studied medications ever prescribed, are safe; in many countries, but not in the USA, oral contraceptives are available without a doctor's prescription. The pills are highly effective and prevent pregnancy primarily by suppressing ovulation through the combined actions of estrogen and progestin. Typical-use failure rates are approximately 5 percent.

Following its introduction in 1960s, there was rapid, widespread acceptance of the pill in the US, especially among young women. By 1965, 24 percent of all married contraceptors were relying on the pill, and by 1970, the pill was the single most popular method used by married American women. Increasing reliance on the pill was accompanied by a reduction in use of barrier methods, rhythm, and withdrawal (Westoff and Ryder 1977). In many other currently industrialized nations as well, pill use has eventually overtaken older methods of contraception at various times, depending on when it was introduced and made easily available. Today, the pill is the most widely-used reversible method of contraception in the Western developed countries. It is important to recall, however, that the major declines in fertility in the currently industrialized nations began in the late nineteenth and early twentieth centuries, well before the development of the oral contraceptive, and even before the mass production and distribution of barrier methods such as condoms and diaphragms.

For a variety of reasons, some women want to avoid hormonal contraceptives which contain estrogen. In these cases, progestin-only methods of contraception, such as injectibles, implants, and pills can be safe, highly effective alternatives. In the USA, Depo-Provera injections and Norplant implants have typical-use failure rates of less than 1 percent. The progestin-only pill, or minipill, is probably only slightly less effective than combined oral contraceptives (Hatcher et al. 1998).

Emergency contraceptive pills (ECPs)—consisting of progestin only or a combination of estrogen and progestin—can be used within 72 hours following unprotected intercourse. Emergency contraception acts to prevent pregnancy, and is not an abortifacient. If 100 women had unprotected intercourse once during the second or third week of their menstrual cycle, about eight would become pregnant. Following treatment with ECPs, only one or two would become pregnant. Thus, there is a reduction of about 75 percent–88 percent in the probability of pregnancy resulting from use of ECPs (http://ec.princeton.edu/).

2.5 Intrauterine Devices (IUDs)

Physicians began experimenting with intrauterine pessaries for women in the middle of the nineteenth century. However, use of early IUDs, made of silk and silver wire, often led to serious pelvic infections, and their use was thus considered by most physicians to be unwise up through the middle of the twentieth century. The use of plastic instead of wire, begun in the late 1950s, improved the safety of IUD use.

The modern-day IUD is inserted by a physician into the uterus, with a string that should protrude from the uterus to aid in removal. The device can remain in the uterus and provides contraceptive protection for a number of years. The IUD is thought to work primarily by preventing fertilization; it is not currently thought to be an abortifacient. Highly effective, the method has typical-use failure rates that generally lie below the 2 percent level. Some types of IUDs can also

be used as emergency contraception, whereby they are inserted into the uterus within a few days of unprotected intercourse.

2.6 *Contraceptive Sterilization*

Primitive attempts at producing sterility in men have taken the form of drastic actions such as the castration of male servants chosen to safeguard female residents of imperial palaces in ancient China circa 1,100 BC. In an attempt to make women infertile, surgical removal of the ovaries was used by the ancient Egyptians, and retroflexion of the uterus to block the cervical canal was used by certain tribes of Java, as reported in the late nineteenth century (Finch and Green 1963).

During the nineteenth and twentieth centuries, more effective and safer methods of sterilization were developed with improvements in asepsis and anesthesia. Today, female sterilization is accomplished through various methods of mechanically blocking the fallopian tubes to prevent the sperm and egg from uniting. Male sterilization (vasectomy), medically even simpler, safer, and less expensive than female sterilization, involves blocking the *vas deferens* in order to prevent passage of sperm into the ejaculated seminal fluid. Although it is sometimes possible to reverse the procedures, neither male nor female sterilization is considered reversible. Highly effective, both methods have typical-use failure rates of less than 1 percent.

In the first half of the twentieth century in the US and in Europe, sterilization was thought to be appropriate primarily in the eugenic context of those thought mentally unfit to be parents. However, since the 1960s, there has been widespread acceptance of voluntary sterilization, so that now roughly 50 percent of married contraceptors of reproductive age in the USA rely on either female or male sterilization. Worldwide, female sterilization is the dominant form of contraception used today.

2.7 *Induced Abortion*

A popular notion of the difference between contraception—the prevention of pregnancy—and abortion—the termination of pregnancy—is a modern-day distinction which did not usually exist in the past. At least during the early modern period, if not before, and to the end of the nineteenth century, abortion performed prior to quickening was viewed traditionally by European women as just another point on the continuum of fertility-control practices, and not generally less legitimate or acceptable than the rather indelicate means of contraception available at the time (McLaren 1990).

Abortion was probably more commonly used than was contraception before modern times (Himes 1970). Various traditional and folk methods were employed throughout the ages, and some are still in use today, where legal or safe abortion is not widely available. Examples include jumping off tables or down stairs, engaging in vigorous exercise, and performing abdominal massages. Purportedly emmenagogic or abortificient potions and pills made of lead, arsenic, phosphorus, gin, quinine, or different herbal extracts, to name a few, were often either poisonous and potentially lethal to the pregnant woman or, at most partially effective (Santow 1998, Woycke 1988). Insertion into the uterus of objects such as hat pins, knitting needles, and infamously, coat hangers, is dangerous, primarily because of risk of infection and hemorrhage. Sometimes these mechanical abortions are self-induced, sometimes they are performed by abortionists of varying reputation.

Over time, improvements in techniques and asepsis allowed for the development of safe and effective surgical abortion. By the 1930s, the dilation and curettage method was recognized as being as safe as full-term delivery. Today, legally induced abortion is significantly safer medically than continuing a pregnancy to term. Introduced to the USA in 1967, vacuum aspiration is currently the most commonly used method of abortion there. Recent developments in medical abortion have the potential of reducing the need for these types of surgical abortion. For example, mifepristone (formerly known as RU486), in combination with prostoglandins, is used widely to abort early pregnancies in several European countries, and has been found to be safe and effective.

The contribution of abortion to the decline in fertility in currently industrialized countries cannot be quantified, but there is strong reason to believe that it is significant. Probably no country has been able to attain replacement-level fertility without the recourse to induced abortion. Generally, use of abortion increases during the early stages of fertility decline, and then decreases as more effective contraception is adopted by a larger proportion of the population (Ross and Frankenberg 1993; see also *Infanticide and Induced Abortion*).

3. *Recent Trends in Contraceptive Use*

Accurate information on the use of abortion is not available in many countries, even when abortion is legal, but especially when it is not (Henshaw 1990). Tracking trends in this sort of data is even more difficult. Therefore, this section focuses only on levels and trends in contraceptive use. One important measure of contraceptive use is contraceptive prevalence, defined as the percentage of couples (with the woman of reproductive age) who are currently using contraception.

Systematic surveys of contraceptive prevalence and method use began to be conducted only in the 1960s in developing countries, as well as in most developed

Table 1
Contraceptive use[a] by method type and by region of the world, circa 1993

World and regional averages	Contraceptive prevalence[b]	Percentage distribution of couples, by type of method used						
		Sterilization					Other supply methods[c]	Nonsupply methods[d]
		Female	Male	Pill	IUD	Condom		
World	**58**	**19**	**4**	**8**	**13**	**4**	**3**	**8**
Less developed regions	**55**	**21**	**4**	**6**	**14**	**2**	**2**	**5**
Africa	20	2	< 1	7	4	1	2	4
Asia[e]	60	24	5	5	17	3	2	4
Latin America and the Caribbean	66	28	1	14	7	4	2	9
Oceania	29	9	< 1	5	1	1	6	7
More developed regions	**70**	**9**	**5**	**17**	**6**	**14**	**2**	**19**
Japan	59	3	1	< 1	2	46	1	6
Europe	72	3	2	20	8	10	2	26
North America	71	24	14	15	1	10	4	3
Australia/New Zealand	76	26	12	23	5	6	1	4

Source: United Nations 1999.
[a] Contraceptive use among currently married women of reproductive age, including, where possible, those in consensual unions. [b] Contraceptive prevalence is the percentage currently using contraception among couples with the woman of reproductive age. [c] Other methods requiring supplies or medical services, including injectibles, diaphragms, cervical caps, and spermicides. [d] Including periodic abstinence or rhythm, withdrawal, douche, total sexual abstinence if practiced for contraceptive reasons, folk methods and other methods not separately reported. [e] Excluding Japan.

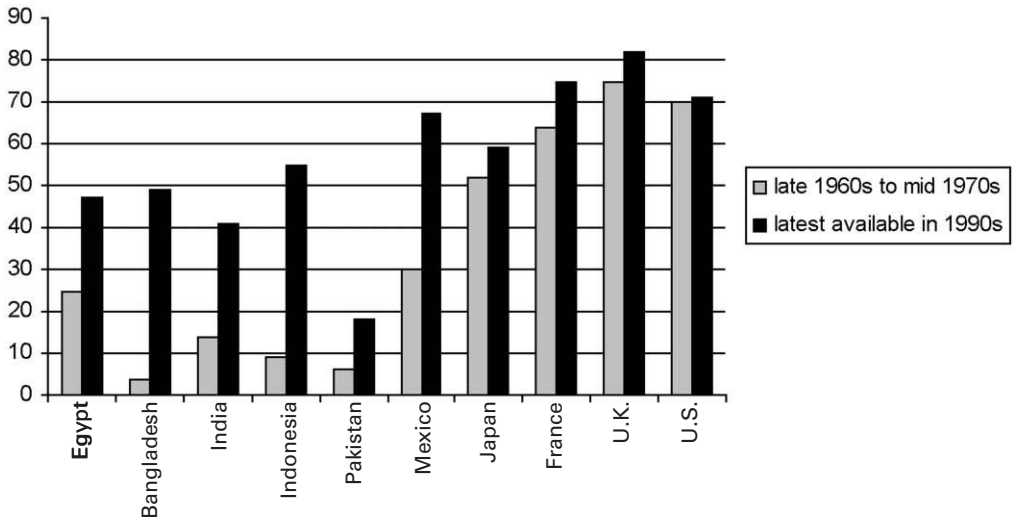

Figure 1
Trends in contraceptive prevalence, selected developing and developed countries (Sources: United Nations (1996, 1999))

countries. Increased concern over rapid population growth and high fertility levels in post-World War II developing countries led to the employment of large-scale surveys of contraceptive use that allow for documentation of prevalence and trends in those regions that is more comprehensive and up-to-date than in currently developed regions. Table 1 presents contraceptive prevalence rates in different regions of the world, based on the most recent data that is currently available. Data in the table refer roughly to the period around 1993.

Contraceptive prevalence was approximately 58 percent for the world as a whole in the years around 1993, but there are important differences between regions. Overall, less developed regions have somewhat lower proportions who use some method of

contraception (55 percent) than do more developed regions (70 percent). The very high level of prevalence in China, currently estimated at 83 percent, raises substantially the average for the less developed regions because of China's large population size. It is estimated that in less developed regions, excluding China, contraceptive prevalence is only 42 percent. In Africa, particularly Western Africa, the prevalence level is lowest in the world (8 percent). Most countries in Latin America and the Caribbean have moderate to high levels of contraceptive use, with nearly all countries reporting prevalence rates of at least 40 percent. Among the more developed regions, the range in prevalence levels is much smaller than among the less developed regions.

Eighty-seven percent of couples who currently use contraception worldwide rely on supply methods—that is, methods which require supplies or medical services. The major supply methods used by couples today are female and male sterilization, the pill, the IUD, and condoms. Supply methods actually comprise a larger percentage of total contraceptive use in developing regions (91 percent) than in developed regions (74 percent). This is probably a function of the role in developing countries of family planning programs, which began massive supply work just when modern methods became available in the 1960s (see *Family Planning Programs: Development and Outcomes*). Only in certain areas such as Southern and Eastern Europe, and parts of West Asia, are nonsupply methods, primarily *coitus interruptus* and rhythm, still widely used among married couples.

Worldwide, and in less developed regions overall, female sterilization is the method most often used. In China and in India, between a quarter and a third of all married women of reproductive age are sterilized. In developed regions overall, the pill is the most commonly used supply method among married couples, but sterilization is the most dominant method in North America and Australia/New Zealand. Condom use is more common in the more developed regions, particularly in Japan, than in the less developed regions, while IUD use is more common in the less developed regions, especially in China. In the more developed regions, IUD use is relatively common in Northern and Eastern Europe. Use of vaginal barrier methods is negligible everywhere.

Nearly all developing countries for which trend data is available have experienced increases in contraceptive prevalence between the years 1970–2000. It has been estimated that contraceptive prevalence for less developed countries has risen by approximately 41 percentage points since the period 1960–5, when it stood at only 14 percent. Increases in prevalence rates among developed countries have been much more modest than among developing countries, in part because the latter started from lower levels (see *Fertility Control: Eastern Europe*). Figure 1 provides a few examples of changes in prevalence rates for a selection of less and more developed countries. In terms of method mix, rising use of supply methods, particularly female sterilization, is the dominant trend in more developed as well as in less developed regions, but the experiences of individual countries is varied. In many more developed countries, increasing use of supply methods has been accompanied by decreasing use of non-supply methods.

See also: Fertility: Proximate Determinants; Reproductive Rights in Developing Nations

Bibliography

Bongaarts J, Potter R G 1983 *Fertility, Biology, and Behavior: An Analysis of Proximate Determinants*. Academic Press, New York

David P A, Sanderson W C 1986 Rudimentary contraceptive methods and the American transition to marital fertility control, 1855–1915. In: Engerman S L, Gallman R E (eds.) *Long-Term Factors in American Economic Growth*. University of Chicago Press, Chicago, pp. 307–90

Davis K, Blake J 1956 Social structure and fertility: An analytic framework. *Economic Development and Cultural Change* **4**: 221–35

Draper E 1972 *Birth Control in the Modern World: The Role of the Individual in Population Control*, 2nd edn. Penguin Books, Harmondsworth, Middlesex, UK

Feldman D M 1995 *Birth Control in Jewish Law*. New York University Press, New York

Finch B E, Green H 1963 *Contraception Through the Ages*. Peter Owen, London

Hatcher R A, Trussell J, Stewart F, Cates W Jr, Stewart G K, Guest F, Kowal D 1998 *Contraceptive Technology*, 17th edn. Ardent Media, New York

Henshaw S K 1990 Induced abortion: A world review, 1990. *Family Planning Perspectives* **22**: 76–89

Himes N E 1970 *Medical History of Contraception*. Schocken Books, New York

Kestelman P, Trussell J 1991 Efficacy of the simultaneous use of condoms and spermicide. *Family Planning Perspectives* **23**: 226–32

McLaren A 1990 *A History of Contraception from Antiquity to the Present Day*. Blackwell, Oxford, UK

Ranum O, Ranum P 1972 *Popular Attitudes toward Birth Control in Pre-Industrial France and England*. Harper & Row, New York

Reed J 1983 *The Birth Control Movement and American Society, From Private Vice to Public Virtue*. Princeton University Press, Princeton, NJ

Riddle J M 1997 *Eve's Herbs: A History of Contraception and Abortion in the West*. Harvard University Press, Cambridge, MA

Ross J A, Frankenberg E 1993 *Findings from Two Decades of Family Planning Research*. The Population Council, New York

Ross J, Stover J, Willard A 1999 *Profiles for Family Planning and Reproductive Health Programs 116 Countries*. The Futures Group International, Glastonbury, CT

Santow G 1993 Coitus interruptus in the twentieth century. *Population and Development Review* **19**: 767–92

Santow G 1995 Coitus Interruptus and the control of natural fertility. *Population Studies* **49**: 19–43

Santow G 1998 Eve's Herbs: A history of contraception and abortion in the West. *Population and Development Review* **24**: 869–75

Trussell J 1979 Natural fertility: Measurement and use in fertility models. In: Leridon H, Menken J (eds.) *Natural Fertility*. Ordina Editions, Liege, Belgium

Trussell J, Kost K 1987 Contraceptive failure in the United States: A critical review of the literature. *Studies in Family Planning* **18**: 237–83

United Nations 1996 *Levels and Trends Contraceptive Use as Assessed in 1994*. United Nations, New York

United Nations 1999 *World Contraceptive Use 1998 Wall Chart*. United Nations, New York

Westoff C F, Ryder N B 1977 *The Contraceptive Revolution*. Princeton University Press, Princeton, NJ

Woycke J 1988 *Birth Control in Germany 1871–1933*. Routledge, London

B. S. Okun

Fertility Control: Prevalence and Consequences of Breastfeeding

The fertility inhibiting consequences of breastfeeding have been known for centuries. Among the ancient Greeks, Aristotle noted that women who nursed their infants for extended periods of time were able to delay the onset of postpartum menstruation and conception. Yet only recently have demographers begun fully to appreciate breastfeeding's impact on global fertility levels. Since the 1960s a large number of scientific studies from many parts of the world have consistently demonstrated that breastfeeding is closely associated with prolonged periods of postpartum amenorrhea, long birth intervals, low completed fertility, and optimal health and wellbeing for the mother and infant. Breastfeeding is now recognized as an important determinant of fertility, especially in the developing world, where modern contraceptive usage is low and breastfeeding rates are high. In this article we review the recent transition in thinking regarding the contraceptive role of breastfeeding.

1. Early Demographic Research

Beginning in the 1950s researchers hypothesized that variation in breastfeeding behavior contributed to fertility differences among populations. Studies showed that the survival of one child is positively associated with the mother's succeeding birth or pregnancy interval. In Senegal, researchers noted that the interval between one birth and the next increased by at least nine months when the age at weaning of the first of those children increased by one year. Also, average birth intervals were 25 months when Senegalese mothers breastfed their children for 12 months or less, but were as long as 41 months when they breastfed for up to 35 months. Still other studies demonstrated that in the absence of any breastfeeding the average duration of postpartum amenorrhea is one or two months with a corresponding birth interval of just over one year. By the 1970s the weight of the accumulated demographic evidence led one researcher to conclude that, in developing countries, breastfeeding prevents more pregnancies than all modern forms of contraception combined (for details see Rosa 1975).

By the mid-1970s other researchers using linear regression techniques found that when regressing the average duration of postpartum amenorrhea on the average duration of breastfeeding, each additional month of breastfeeding increases amenorrhea by somewhat less than one month. These types of empirical findings helped form the basis for what has become known in demographic circles as the 'proximate determinants of fertility' framework. This modeling approach tries to account for variation in fertility between populations by quantifying a limited and exhaustive list of variables having a direct bearing on fertility. The period of time in which nursing mothers are not menstruating or ovulating and are unable to conceive following a pregnancy is referred to as lactational infecundability. As the name implies, breastfeeding is assumed to be the leading or sole cause of the delayed return to the fecundable state in nursing mothers. Used widely as an analytical tool, the proximate determinants of fertility framework has demonstrated, worldwide, that lactational infertility is an important determinant of the total fertility rate between populations. For example, in an analysis of World Fertility Survey data (1974–84) using the proximate determinants framework, demographers estimated the relative contributions of breastfeeding and artificial methods of contraception to the restriction of fertility in 29 developing countries in Africa, the Americas, and Asia. The analysis showed that in Africa the amount of potential fertility reduced by modern contraceptive use is relatively low, whereas breastfeeding inhibits an average of four births per women. In Asia, breastfeeding remains important for fertility control, but overall, modern contraceptive use is more important than in Africa. In Mexico, Central America, and South America, modern contraceptive use plays the predominant role with only one-fifth of the total fecundity reduction due to breastfeeding. In poorer developing countries like Bangladesh, Peru, and Haiti, however, breastfeeding still plays the dominant role in reducing fecundity. Similar results have been found using more recent data from the Demographic and Health Surveys (DHS) for currently developing countries. Results using the proximate determinants framework in 25 DHS countries showed that for each additional month increase in the duration of breastfeeding there is a one-half to three-fourths month increase in the duration of postpartum amen-

orrhea (for detailed treatment see Goldman et al. 1987, Guz and Hobcraft 1991, Thapa et al. 1992).

Others have developed models incorporating the underlying physiological processes of postpartum infertility. In doing so they have pointed out an important limitation of the proximate determinants approach. The proximate determinants framework uses the total duration of breastfeeding to assess postpartum lactational infecundability, when in fact amenorrhea typically ends well before breastfeeding is terminated. Thus simply showing empirically that the duration of breastfeeding is highly correlated with the duration of postpartum amenorrhea in a regression model reveals little about the underlying causal relationship between the two. Furthermore, physiological models provide a means of assessing the contribution of full, unsupplemented breastfeeding, not simply the total duration of breastfeeding used in linear regression approaches. Results suggest that for each month of exclusive breastfeeding, anovulation and/or amenorrhea would on average be postponed by *more* than one month. Nonetheless, the strong statistical relationship between the duration of breastfeeding and the duration of postpartum amenorrhea, whether measured at aggregate or individual levels, is remarkably robust, and indicates a strong underlying biological link between the two (see details in Habicht et al. 1985, Santow 1987).

2. *Neuroendocrine Control of Postpartum Lactational Infertility*

Thanks to a series of clinical studies carried out in a small homogeneous group of breastfeeding mothers in Edinburgh, Scotland, the neuroendocrine regulation of lactational infertility is now fairly well understood. When breastfeeding is initiated, afferent nerve impulses pass directly along nerve axes leading from the mother's nipple to the hypothalamus in the brain, directly disrupting the release of the hormones responsible for ovulation. Disruption continues until the neural suckling stimulus declines sufficiently to allow for normal hormonal release to be restored. When suckling is reduced or stopped, ovulation, fertilization, implantation, embryogenesis, and pregnancy are once again biologically possible. Therefore, it is the suckling stimulus itself that plays the key role in the neuroendocrine control of postpartum lactational infertility, not the direct effect of delayed weaning. Only in cases where weaning or the death of a child terminates breastfeeding prior to the resumption of menses can the total duration of breastfeeding, *per se*, be considered causally related to the return of postpartum menstruation.

Apart from elucidating the underlying neuroendocrine mechanisms of postpartum infertility, the Scottish studies also revealed significant new insights into the relationship between postpartum infertility and specific aspect of infant feeding behavior. The research showed, for example, that both the number and the duration of suckling episodes played key roles in suppressing ovarian function. None of the Scottish women in the study ovulated if they breastfed six or more times and more than 60 minutes a day. The introduction of supplemental foods prompted a decline in both suckling frequency and duration with suckling duration declining somewhat earlier than frequency.

The more abrupt the introduction of supplements the more rapid the decline in duration compared to frequency and the more rapid the resumption of ovarian function. Women who maintained regular night-time feeds tended to breastfeed with a greater frequency and longer duration and to remain anovulatory for a longer period of time than those who did not. Overall, supplementation undermined and reduced the intensity of the infant's suckling stimulus (however measured) and led to an earlier resumption of ovarian activity and menstruation than if exclusive breastfeeding were maintained (for details see Tay et al. 1996).

3. *Hazard Model and Life Table Studies of the Fertility Inhibiting Consequences of Breastfeeding*

To date only a handful of large-scale, longitudinal studies, including those in India, Taiwan, Bangladesh, Guatemala, Indonesia, and the Philippines, have been undertaken to investigate, either directly or indirectly, the determinants of postpartum lactational infertility in nonclinical settings. The strength of these studies derives, in part, from their prospective data collection frameworks and their measurement of nursing behavior. The investigations lasted for periods of one to three years or longer in some cases. They prospectively collected information on infant nursing behaviors, supplemental feeding practices, maternal and infant health, the duration of postpartum amenorrhea and birth interval, as well as a rich array of other sociodemographic and control variables.

An additional strength of these prospective studies arises from their use of an alternative body of statistical techniques known as event history or hazard model analysis. These statistical procedures have been developed since the 1980s specifically to handle duration or event variables, censoring (when a subject, for whatever reason, leaves a prospective study before it has ended or the event of interest does not occur before the study ends), and other problems inherent in longitudinal study designs. Hazard model analysis provides the ideal statistical framework for unambiguously quantifying the effects of nursing behavior on the duration of postpartum amenorrhea in nonclinical settings while simultaneously controlling for the joint effects of confounding variables (see *Event-history Analysis in Continuous Time*).

To cite one example, women living in rural Indonesia were interviewed monthly about their current menstrual status and their average daily breastfeeding pattern. The women were classified into four groups based on nursing intensity. Univariate life-table analysis revealed that low-intensity nursing mothers resumed menses at 11.5 months, while high-intensity nursers resumed menses at 20.4 months. Medium-low and medium-high intensity nursing groups resumed menses at 14.0 and 17.7 months, respectively. A series of multivariate proportional hazard models showed that, when controlling for confounding factors, the higher the intensity of suckling the slower the rate of return of menses. Further multivariate analysis also showed that partial breastfeeding (with supplements) significantly increased the risk of resumption of postpartum menstruation when controlling simultaneously for nursing intensity and other confounding factors.

In the Philippines, research also demonstrated that partial (supplemented) breastfeeding significantly increases the risk of early return of postpartum menstruation. Hazard model analysis showed that factors such as active suckling, the use of two breasts versus one, and breastfeeding on demand vs. a fixed nursing schedule significantly decreased the risk of early return of postpartum menstruation. By contrast, the feeding of milk other than breast milk and nonnutritive or low caloric nonmilk liquids increased the risk of early resumption of postpartum menstruation.

In the final analysis, these longitudinal studies and the clinical research from Edinburgh share an emergent consensus concerning breastfeeding and fertility. The consensus is that, in spite of the wide range of variability in infant feeding patterns and practices across populations and individuals, it is the sum total of all direct and indirect breastfeeding activity affecting the intensity of the suckling stimulus, and thus the strength of the neural input to the brain, that is the most important determinant/mediator of postpartum infertility. (For references and details see Jones 1989, 1990, Popkin et al. 1993, Zohoori and Popkin 1996.)

4. *Breastfeeding and Family Planning*

The emergent scientific consensus regarding the reproductive consequences of breastfeeding represents a marked departure in thought and sophistication from that of just a few decades earlier. After carefully reviewing the evidence at hand, a group of researchers meeting in Italy in the late 1980s issued a statement regarding the contraceptive role of breastfeeding. Known as the Bellagio Consensus Statement, the announcement noted that, if three conditions prevail, postpartum mothers could expect a pregnancy rate of less than 2 percent. These three conditions are: a woman must remain amenorrheic since delivery, she must be fully or nearly fully breastfeeding (little or no supplements), and she must be within six months of delivery. This is referred to as the lactational amenorrhea method (LAM) of family planning (for details see Kennedy et al. 1989, 1990).

5. *Conclusion*

The accumulation of scientific evidence since the 1960s has confirmed what many, especially women, have known intuitively over vast stretches of human history and prehistory: that breastfeeding acts as nature's way of keeping human fertility rates low and of guaranteeing enough time between births for the child to be adequately nourished and developed and for the mother to sufficiently recover from the rigors of child bearing and rearing. Prospective research undertaken in both clinical and real-life settings has confirmed the contraceptive effects of breastfeeding. Knowledge gained from these studies has lead to the development of guidelines to enable breastfeeding mothers to use lactation as a means of fertility control, as least during the first six months postpartum, or possibly up to 12 months postpartum, if mothers are actively pursuing breastfeeding and are receiving appropriate family planning counseling.

Worldwide it is estimated that more than 90 percent of mothers with infants still breastfeed, at least for a short period of time. If women were to be counseled on LAM use, this could potentially provide a natural way to prevent pregnancy immediately after birth, much as it has for most of humankind's evolutionary history. Unfortunately, however, the initiation, intensity, and duration of breastfeeding are declining due to urban lifestyles and other changes in the postindustrial years. The recent Demographic and Health Surveys have shown that today in many developing countries young women who are educated, urban, and affluent are abandoning breastfeeding in favor of bottle-feeding. Often these women are the trend setters in their communities, and their influence is spreading to the other urban and rural poor women who are the least likely to have access to or use modern contraceptives and whose infants have the most to gain from breast milk.

In recent years, however, the advent of the HIV pandemic presents a potentially more dangerous public health problem than declining rates of breastfeeding. Even though HIV, the virus that causes AIDS, can be transmitted by breastfeeding, WHO and UNICEF have advised that where other infectious diseases and malnutrition are the primary cause of infant deaths, breastfeeding should remain the standard advice to pregnant women, including those who are known to be HIV-infected, because their baby's risk of becoming infected through breast milk is likely to be lower than its risk of dying of other causes if deprived of breastfeeding Whether or not this advice can or should be implemented is currently a matter of strong debate. Only time and more research will be

able to resolve these important social, medical and ethical issues (see details in Kennedy et al. 1990, Family Health International 1992, World Health Organization Global Programme on AIDS 1992, Trusssell et al. 1992; see also *HIV and Fertility*).

See also: Family Theory and the Realities of Childbearing Behavior; Family Theory: Complementarity of Economic and Social Explanations; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants

Bibliography

Bongaarts J, Potter R G 1983 *Fertility, Biology and Behavior*. Academic Press, New York

Family Health International 1992 Breastfeeding. *Network* **13**

Goldman N, Westoff C F, Paul L E 1987 Variations in natural fertility: The effect of lactation and other determinants. *Population Studies* **41**: 127–46

Gray R, Leridon H, Spira A (eds.) 1993 *Biomedical and Demographic Determinants of Reproduction*. Clarendon Press, Oxford, UK

Guz D, Hobcraft J 1991 Breastfeeding and fertility: A comparative analysis. *Population Studies* **45**: 91–108

Habicht J-P, Davanzo J, Butz W P, Meyers L 1985 The contraceptive role of breastfeeding. *Population Studies* **39**: 213–32

Jones R E 1989 Breast-feeding and post-partum amenorrhoea in Indonesia. *Journal of Biosocial Science* **21**: 83–100

Jones R E 1990 The effect of initiation of child supplementation on resumption of post-partum menstruation. *Journal of Biosocial Science* **22**: 173–89

Kennedy K I, Fortney J A, Bonhomme M G, Potts M, Lamptey P, Carswell W 1990 Do the benefits of breastfeeding outweight the risk of postnatal transmission of HIV via breastmilk? *Tropical Doctor* **20**: 25–9

Kennedy K I, Rivera R, McNeilly A S 1989 Consensus statement on the use of breastfeeding as a family planning method. *Contraception* **39**: 477–96

Knobil E, Neill J D, Ewing L L, Greenwald S, Markert C L, Pfaff D W (eds.) 1994 *The Physiology of Reproduction, Volume 2*, 2nd edn. Raven Press, New York

Leridon H, Menken J (eds.) 1977 *Natural Fertility*. Ordina Editions, Liege, Belgium

Leridon H 1977 *Human Fertility: The Basic Components*. University of Chicago Press, Chicago

Popkin B M, Guilkey D K, Akin J S, Adair L S, Udry J R, Flieger W 1993 Nutrition, lactation, and birth spacing in Filipino women. *Demography* **30**: 333–52

Rosa F W 1975 The role of breastfeeding in family planning. *Protein Advisory Group Bulletin* **5**: 5–10

Tay C C K, Glasier A F, McNeilly A S 1996 Twenty-four hour patterns of prolactin secretion during lactation and the relationship to suckling and the resumption of fertility in breast-feeding women. *Human Reproduction* **11**: 950–5

Thapa S, Kumar S, Cushing J, Kennedy K 1992 Contraceptive use among postpartum women: Recent patterns and programmatic implications. *International Family Planning Perspectives* **18**: 83–92

Thapa S, Short R V, Potts M 1988 Breast feeding, birth spacing and their effects on child survival. *Nature* **335**: 679–82

Trussell J, Grummer-Strawn L, Rodriquez G, VanLandingham M 1992 Trends and differentials in breastfeeding behavior: evidence from the WFS and DHS. *Population Studies* **46**: 285–307

Wood J W 1994 *Dynamics of Human Reproduction: Biology, Biometry, Demography*. Aldine De Gruyter, New York

World Health Organization Global Programme on AIDS (WHO/GPA) 1992 Current and Future Dimensions of the HIV/AIDS Pandemic: A Capsule Summary. WHO, Geneva

Zohoori N, Popkin B M 1996 Longitudinal analysis of the effects of infant-feeding practices on postpartum amenorrhea. *Demography* **33**: 167–80

R. E. Jones

Fertility: Institutional and Political Approaches

Much of fertility analysis is based on the model of an autonomous actor like the one assumed in microeconomic theory, characterized by a preference set against which responses to price and other signals are determined. In fertility studies, the actor is a woman (rarely a man) of reproductive age or a composite 'couple.' Fertility surveys relating individual childbearing patterns to individual characteristics, family-size desires, and practice of birth control help to maintain this focus. Fertility, however, is also a societal process and outcome. It is affected by the powerful close-range forces at work within the microsociety of the family. Moreover, both individuals and families are enmeshed in larger social groups—clans, villages, and other communities, castes, social classes, religious groups, nations—which may affect their fertility preferences and are the locus of institutional constraints within which their fertility-related decisions are taken. The ways in which fertility is influenced by these broader social organizational and institutional settings, and by the patterns of social control that operate within them, are the subject of this article.

1. Institutional Determinants of Fertility

Institutions are enduring clusters of behavioral rules, known but often unwritten, governing human actions and relationships in recurrent situations. They are not rigid constraints requiring compliance, though there may be external or self-imposed sanctions for violation. They are not unchanging: they are constantly being made and remade by those coming into contact with them—and sometimes by deliberate efforts at policy design. They have to do with 'structure,' but not only in the sense of tangible reality. Cultural values and ideational systems may also be institutionalized.

Marriage and the family are familiar examples of institutions—more precisely, the sets of behavioral

rules and expectations characterizing marriage and the family. Clearly, both are closely linked to fertility outcomes. Other institutions that potentially influence fertility, some more distantly, include the body of family and property law, whether customary or codified; the formal authority structure in the community and local government; the stratification system, and the mobility paths (class- and gender-specific) that it supports; public systems of interage transfers, such as child allowance and pension schemes; and the labor market, particularly as it affects employment of women. The forms such institutions take, emerging from each society's unique history, support a 'fertility regime'—a distinctive pattern of family formation and array of fertility preferences, behaviors, and outcomes. This can be illustrated by some examples of different regimes.

1.1 Social Regulation of Fertility in Preindustrial Agrarian Societies

Many societies facing static land resources and productivity, and recurrent high mortality found institutional means of keeping population growth low, with fertility well below its maximum level and able to respond promptly to mortality crises. In English parishes, for example, prior accumulation of property was required before a person could marry and set up an independent household; in periods of economic hardship, saving took longer and marriage was delayed. A similar kind of community-level social control of household formation existed in preindustrial Japan. In China, social pressures, particularly from the family or lineage, applied not to marriage (which for women was near to universal) but to fertility within marriage, through expectations of lengthy interbirth intervals. Other institutions—'service' in England (sending children to work for a period as servants in other households), adoption in China—helped to smooth family consumption and production imbalances (see Smith 1981, Lee and Wang 1999; *Family Systems in Europe*).

In most traditional societies it is possible to discern distinctive patterns of institutional regulation of fertility that yield somewhat comparable demographic outcomes. The patterns vary with the nature of the family system, with the strength or weakness of territorially-based communities, with the kind of economy, and with the structure and roles of the larger political and administrative systems in which the local society is set. Very high fertility is not a feature of such societies: where it occurs it usually indicates the disruption of customary, institutionalized behavior.

1.2 Institutional Factors in Demographic Transition

Early accounts of and explanations for demographic transition—the secular declines in mortality and fertility from high to low levels—were largely cast in institutional terms. Social institutions like those just discussed constituted a normative order that sustained high fertility in the face of the prevalent high-mortality risks. As those risks lessened and as technological advances created new economic opportunities, that normative order was slow to change. Hence the pressures and incentives facing individuals tended for a time to maintain the earlier patterns of fertility. Ultimately, however, families discerned their economic interest in having fewer children and acted accordingly. Institutional control was supplanted by rational individual or family-level decision making (see Szreter 1993).

More recent explanations of the transition do not emphasize this supposed shift from institutional to personal and conscious control. Fertility decisions are seen as rational (or boundedly rational) throughout—in the sense of being related to the actor's knowledge and perceived constraints and options at the time. Fertility is also institutionally conditioned throughout: those constraints and options are set by the society's institutional configuration as well as by its members' individual backgrounds and circumstances. This is the case no less in periods of rapid social and economic change than in the 'stable' preindustrial situation. Indeed, such change is in large part a change in that institutional setting (see North 1990). For example, industrial development creates new avenues of social mobility, but ones that require education and often relocation to urban centers; the social capital of kinship and other ties within the community is supplanted by human capital, transforming the economics of the family. High fertility is increasingly seen as inconsistent with the new conditions. (For various accounts of demographic transition as institutional change see Davis 1963, Caldwell 1982, McNicoll and Cain 1990, *Fertility: Political and Political–Economic Perspectives*.)

Demographic transition in turn induces institutional change. The faster population growth occurring with sustained high (or even rising) fertility but falling mortality early in the transition is a factor promoting change to more intensive agricultural systems (Boserup 1965). A rapid decline from high fertility levels, such as took place in many countries since the 1960s, substantially lowers child dependency rates (its effect in increasing old-age dependency is felt only later) and provides a boost to development performance.

1.3 Institutional Factors Perpetuating Low Fertility

The social pressures that support high fertility do not long survive in the conditions of urban industrialized societies. Other institutional influences come to bear on fertility with the opposite effect. The labor market offers new career opportunities to women, but ones difficult to combine with childbearing. With women's

greater economic independence the institution of marriage erodes; the individual rather than the family becomes the basic unit of society in matters such as public finance. The state intrudes further in the education of children, lessening parental roles in socialization; more generally, peer influence becomes more significant, intergenerational influence less so. At the same time, the cost of childraising—both the direct cost of education and the opportunity cost of forgone income and leisure time—increases markedly in the new knowledge-based economy, much of it having to be borne by the parents.

There is little agreement on how low fertility will become in modern affluent societies, and whether there are inherent forces that might cause it to rebound (see *Demographic Transition, Second*). Some observers foresee an emerging retreat from individualism and restoration of family and community roles as a natural cyclical phenomenon (Fukuyama 1999). Others point to innate biological drives that may sustain fertility. But many see little likelihood of a return to near-replacement levels. Once the lingering effects of earlier high fertility have worked their way through the age distribution, however, the prospect of natural population decrease of 20 percent or more per generation is likely to provide a strong rationale for pronatalist interventions by the state in the design of social and economic arrangements (see Chesnais 1996, Sect. 3.3).

1.4 Linking Institutions to Fertility

Fertility at its simplest is a summary measure of reproductive outcome. Institutional arrangements in a society sometimes deal directly with that outcome, as in systems rewarding or penalizing births in a family beyond a certain prescribed number (China's one-child policy, for instance). But fertility is also a process—a series of individual decisions and biological events—providing many opportunities for indirect institutional influence on outcomes.

The biosocial process leading to a birth can be decomposed into its elements, the so-called intermediate variables or proximate fertility determinants such as age at marriage, duration of breastfeeding, and use of contraception. Each such element is potentially subject to separate institutional influence. Where an intermediate variable is distant from the fertility outcome, the effect on fertility may be merely incidental. For instance, the institutional shift from arranged marriage to love marriage is likely to be accompanied by a rise in average ages of marriage and first birth, tending to depress fertility. Labor market changes affecting women's employment may lead to shortened breastfeeding and hence (through shortened amenorrhea) to higher fertility—to be offset eventually by greater use of contraception (see Davis and Blake 1956).

If the intermediate variables are strongly interdependent, with individuals or couples having some given family size goal in view, subject to economic circumstances, then the main institutional determinants of fertility are those that affect that goal or its attainment. The intermediate variables in this case lose much of their behavioral interest. Where the preferences of one spouse differ from the other, changes in power relations within the family, or in the broader gender system in the society, may be a significant source of fertility change. Greater gender equality is widely believed to promote lower fertility. Other institutional effects on fertility goals would be those that change the perceived costs and benefits of children to parents or, more fundamentally, alter people's concept of the family and its role in the society. Effects on the attainment of fertility goals would include any overt social or political pressures bearing on reproductive behavior (see McNicoll 1994, Potter 1983).

2. Origins and Maintenance of Fertility Regimes

The theory of institutions investigates the origins of conventions and other institutionalized behavior, drawing particularly on transaction-cost economics and game theory. Institutional change bears an analogy to technological change, in being characterized by path dependence and possibilities of multiple equilibria. Some of this theory illuminates fertility regimes.

2.1 Transaction Cost Approaches to the Family

Transactions among members of a society entail 'transaction costs' associated with advertising or acquiring information, negotiating terms, and enforcing contracts. Social institutions, such as a banking system, may lower such costs and thus make market transactions more efficient. It is also possible to economize on transaction costs by internalizing some transactions within an enduring group or organization, members of which come to know and trust each other and thus do not need to fully specify the contractual terms. The firm is such an organization; so is the family (see Ben Porath 1980). The kinds of transactions that are best dealt with in such organizations are those where the identity of the parties matters. Examples are exchange of labor services in situations where effort cannot readily be measured, and parental investment in a child's education in expectation of a much later return.

The family, from this perspective, is the most important institution for nonmarket transactions. Its form is influenced by the emergence of market institutions, such as savings institutions, that lessen its transaction-cost advantages. The implicit contract between parent and child is one such intrafamily

transaction subject to change over the broader course of institutional development in the society, with strong implications for fertility. (See Pollak 1985.)

2.2 *Path Dependent Institutional Change*

Fertility, however, is not sufficiently described as an output of the family treated as a miniature firm-like organization. Family and gender systems require a fuller institutional specification to accommodate their distinctive patterns and fertility outcomes. The persistence of such systems over many generations in preindustrial societies suggests that they were stable solutions to fundamental problems of organizing reproduction and dealing with child and old-age dependency. Such problems include how children are to be protected and socialized, how sexuality is regulated and parenthood affirmed, how claims to property are maintained and transmitted, and how the dependent elderly are supported. Each admits of many different solutions. Institutionalized, some combinations of those solutions dominate over time, partly through characteristics of the resulting society—among them, its reproductive success—and partly perhaps by fortuity. Belief systems develop alongside such structures, generating explicit rules of practice and sanctions for violation. Ties beyond the immediate family are an essential means of risk-spreading in high mortality conditions, and the reciprocal relationships thus established acquire cultural validity as a 'moral economy.' Social groups or statuses advantaged by a given structure have evident interests in its maintenance, entrenching a system of social control.

Such an account suggests that variant family and gender systems are stable institutional equilibria, and their resulting fertility regimes are to some degree fortuitous. Different regimes subjected to the same pressures for fertility change (new resources, new technologies, larger surviving birth cohorts) would respond differently, some more resistant to change than others. Variant paths of fertility transition are in fact observed.

The insight is quite general. A society's institutional structure is historically contingent and thus in some measure accidental. Institutional change, like change in technology, is path dependent: the direction of change is constrained not merely by its existing state but by its history. Institutions, in part the residues of past social action, limit the possibilities of subsequent social action (see David 1993). But change is also influenced by hopes and expectations about the future, which are not just reflections of immediately perceived options but are formed out of information about the world and the opportunities it seems to present. A role for expectations introduces the possibility of bandwagon effects in social change, as individuals, in making their own decisions, take account of how they anticipate others will decide. Fertility outcomes derive from both history and expectations.

Fertility policy, of course, is also path dependent. Governments do not have a blank slate on which to make policy changes in accordance with the best advice from experience elsewhere; they deal with the politically feasible, meaning a situation inherited from the past.

2.3 *Institutions and Culture*

One strand of economic theorizing on fertility assumes a largely 'institution-free' world, in which, in Becker's (1976) words: 'all human behavior can be viewed as involving participants who maximize their utility from a stable set of preferences and accumulate an optimal amount of information and other inputs in a variety of markets.' The need to specify preferences, or tastes, introduces a domain of values, hence presumably a cultural system, but one kept in the background. A second strand, institutional economics in the manner of Ben Porath, or North, derives from the analysis of transactions and iterated games. It is structurally more sophisticated but at its core assumes the same model of a utility-maximizing agent. Institutions, by implication, are economic institutions—or at least institutions amenable to economic analysis.

A consequence of this situation is that an institutional perspective on fertility is often contrasted with a cultural or ideational perspective, as if the two were entirely distinct. The antithesis is false. Even in the study of economic development, where resources and technologies are dominant realities, economic incentives and constraints are necessarily interpreted through a cognitive lens. Controversies over the role of the Protestant ethic or the Confucian ethic in promoting development are not just about abstract values, but about rules of conduct. For fertility, the relevant systems of social control typically combine tangible rewards and sanctions with the symbolic. Ideational systems provide a structure of meaning and a framework against which to evaluate life-course opportunities and options, to determine the scope of tradeoffs in economic calculation, and (internally) to assess gratification or guilt. Where those ideational structures are widely shared, behavior that accords with them can be described as institutionalized. (For discussions of the role of cultural systems and ideational change in fertility determination see Kreager 1986, Kertzer and Fricke 1997. See also *Fertility Transition: Cultural Explanations*; *Family Theory: Complementarity of Economic and Social Explanations*.)

3. *Governance and Fertility Change*

Since societies, and nations in particular, have an evident interest in their demographic futures, fertility (along with migration) is an object of policy attention.

For most societies most of the time it is not high on the policy agenda, but it typically becomes so in situations of very high or very low fertility. However, governments are routinely engaged in maintaining and modifying systems of law and administration which, though having no deliberate demographic intent, have a significant influence on fertility.

3.1 Incidental Effects of Government Action

There is a wide consensus on the kinds of political and administrative setting that are conducive to economic growth. In establishing such a setting a government is likely also, but incidentally and perhaps unintentionally, to be promoting fertility change along the path of demographic transition. Individuals in seeking economic opportunity are likely to see low fertility as advantageous. A striking illustration of this is seen in those Western countries, including the United States, which attained close to replacement-level fertility early in the twentieth century in the face of strong government disapproval of birth control (see Watkins 1991).

In modern welfare states a range of other government measures bears unintentionally on fertility. Elaborate social policy objectives are pursued through family law, employment regulation, and tax expenditures. Recognition of individual rights and entitlements, translated into social policy, sometimes works against the interests of families. The general social interest in reproduction—the issue of bearing and raising the next generation—is left to the private actions of individuals and families, with no assurance that the latter will aggregate to the social interest (see Demeny 1986).

3.2 Political and Administrative Pressures to Lower Fertility

Many governments pursuing economic development have sought to lower fertility by offering economic incentives to participants (or to recruiters of participants) enrolling in birth control programs. The effects, when discernible at all, have been small. More forceful measures, however, have sometimes attained significant results. Examples are China's campaign of the 1970s for later marriage, longer birth intervals, and fewer births per woman, and China's subsequent one-child policy; and India's sterilization program during the period of rule under the declaration of national emergency in 1975–7. In these campaigns the government's political authority and administrative apparatus were drawn on to implement central policy directives that might otherwise have lain within the duties of a relatively powerless line ministry. In some other countries more routine family planning programs have also sometimes made use of the local civil administration in securing clients (see McNicoll 1997).

The personal costs of coercive measures are high. The political costs are also potentially large: they were, for example, a factor in the change of government in India in 1977. Even authoritarian governments have limited political capital to allocate over the policy spectrum, and may be loath to expend much of it on birth control programs. A program's unpopularity at the grass-roots level can provide a source of solidarity that spills over into other kinds of opposition to government authority.

3.3 Institutional Innovation in Very Low Fertility Settings

Under a continued rate of fertility well below two children per woman, a country will experience declining population size and rapid population aging. It is likely to seek ways of raising fertility closer to replacement level. Direct state subsidies in the form of child endowments are expensive and have not proven very effective. The amounts offered typically are dwarfed by the actual costs of childraising, especially for more affluent parents; moreover, large windfall gains to those recipients who would not have required the subsidy are inescapable. Symbolic rewards have fared even worse. More promising are measures that increase the compatibility of women's (and more generally, parents') employment and childrearing. Examples are flexible working hours, provision of childcare, job-sharing, and telecommuting. These may sometimes be in the employer's interest; in other cases their cost would have to be allocated between employer and government. A third category would cover efforts to reaffirm parental interest in childrearing by providing parents with a greater tangible return to them. (For proposals of this kind see Coleman 1993 and Demeny 1987.)

The diverse efforts of the many societies likely to be seeking to raise their fertility if very low birth rates persist will provide a natural experiment that will help to determine the relative efficacy and costs of alternative measures.

See also: Demographic Transition, Second; Fertility: Political and Political–Economic Perspectives; Fertility: Proximate Determinants

Bibliography

Becker G S 1976 *The Economic Approach to Human Behavior*. University of Chicago Press, Chicago

Ben Porath Y 1980 The F-connection: Families, friends, and firms and the organization of exchange. *Population and Development Review* **6**: 1–30

Boserup E 1965 *The Conditions of Agricultural Growth*. Aldine, Chicago

Caldwell J C 1982 *Theory of Fertility Decline*. Academic Press, London

Chesnais J C 1996 Fertility, family, and social policy in contemporary Western Europe. *Population and Development Review* **22**: 729–39

Coleman J S 1993 The rational reconstruction of society—1992 presidential-address. *American Sociological Review* **58**: 1–15
David P A 1993 Historical economics in the long run: Some implications of path dependence. In: Snooks G D (ed.) *Historical Analysis in Economics*. Routledge, London
Davis K 1963 The theory of change and response in modern demographic history. *Population Index* **29**: 345–66
Davis K, Blake J 1956 Social structure and fertility: An analytic framework. *Economic Development and Cultural Change* **4**: 211–35
Demeny P 1986 Population and the invisible hand. *Demography* **23**: 473–87
Demeny P 1987 Pronatalist policies in low fertility countries: Patterns, performance, and prospects. In: Davis K, Bernstam M S, Ricardo-Campbell R (eds.) *Below-Replacement Fertility in Industrial Societies: Causes, Consequences, Policies*. Cambridge University Press, Cambridge, UK
Fukuyama F 1999 *The Great Disruption: Human Nature and the Reconstitution of Social Order*. Free Press, New York
Kertzer D I, Fricke T 1997 *Anthropological Demography: Toward a New Synthesis*. University of Chicago Press, Chicago
Kreager P 1986 Demographic regimes as cultural systems. In: Coleman D, Schofield R (eds.) *The State of Population Theory: Forward from Malthus*. Blackwell, Oxford, UK
Lee J, Wang F 1999 *One Quarter of Humanity: Malthusian Mythology and Chinese Realities, 1700–2000*. Harvard University Press, Cambridge, MA
McNicoll G 1994 Institutional analysis of fertility. In: Lindahl-Kiessling K, Landberg H (eds.) *Population, Economic Development, and the Environment*. Oxford University Press, Oxford, UK
McNicoll G 1997 Governance of fertility transition: Reflections on the Asian experience. In: Jones G W, Caldwell J C, D'Souza R M (eds.) *The Continuing Demographic Transition*. Clarendon Press, Oxford, UK
McNicoll G, Cain M (eds.) 1990 *Rural Development and Population: Institutions and Policy*. Oxford University Press, Oxford, UK
North D C 1990 *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, UK
Pollak R A 1985 A transaction cost approach to families and households. *Journal of Economic Literature* **23**: 581–608
Potter J E 1983 Effects of societal and community institutions on fertility. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries*. Academic Press, New York
Smith R M 1981 Fertility, economy, and household formation in England over three centuries. *Population and Development Review* **7**: 595–622
Szreter S 1993 The idea of demographic-transition and the study of fertility change: A critical intellectual history. *Population and Development Review* **19**: 659–701
Watkins S C 1991 *From Provinces into Nations: Demographic Integration in Western Europe, 1870–1960*. Princeton University Press, Princeton, New Jersey

G. McNicoll

Fertility of Single and Cohabiting Women

Historically, fertility outside formal marriage has been referred to variously as bastardy, illegitimacy, or out-of-wedlock childbearing, with such terms typically connoting the inferior status of children born in such circumstances. More recently, the use of these terms has declined, in part because they presume an answer to the question of whether 'inferior status' is a necessary consequence of being born to an unmarried woman or whether some of this disadvantage might stem from other causes. Instead, terms such as nonmarital fertility or childbearing by single or cohabiting women, which carry more neutral connotations, have seen increasing usage.

1. *Measurement*

Measuring nonmarital fertility requires knowledge of at least two demographic statuses—fertility and marriage—and ideally a third—whether a birth occurs within a cohabiting union. By and large, identifying births as occurring within or outside formal marriage is a straightforward matter given sufficiently detailed data. However, it is useful to identify potential inconsistencies across alternative definitions of nonmarital fertility; for example, 'the fertility of single or cohabiting women' vs. 'childbearing to unwed parents.' Consider a birth conceived within a marital union in which the father dies before the child is born. Viewed one way, the child was conceived to married parents; viewed another way, the mother was single at birth by virtue of the death of her spouse. Similar issues arise for legitimated births or births to separated parents. In the latter case, because the birth occurs prior to the dissolution of a formal marital union, one might view it as a marital birth. However, the child may have been conceived by the woman and a man other than her husband, either before or after marital separation; thus, another definition might classify such a birth as nonmarital. Furthermore, note that the period between separation and divorce can be lengthy, with the timing of a formal divorce decree often a response to various factors, including the outcome (e.g., childbearing) under study. Consequently, some researchers have adopted an operational definition that shifts emphasis from formal to *de facto* marital status, thus including as nonmarital those births that occur between separation—marking the informal but behavioral dissolution of a union—and divorce—marking the formal dissolution of a union.

Additional measurement issues arise when attempting to distinguish between nonmarital births to cohabiting women and to women in neither a marital nor cohabiting union. It is sometimes difficult to determine when or whether a cohabiting union has begun, a problem similar to that posed by common law marriage in historical studies. These issues have led some to argue that cohabitation might be better conceptualized not as a discrete status but rather in terms of a continuum capturing the extent to which a couple might be said to have formed a union. Nevertheless, estimates from alternative data sources are typically in substantial agreement, providing evidence, albeit of an

indirect nature, suggesting that female respondents can report accurately on their union status at birth, even retrospectively.

As a general accounting principle, one might suppose that the number of births to males and females should, in the aggregate, be identical as when data are available from official birth registers. Still, this principle need not hold under some definitions of non marital fertility (e.g., when both parents are married, but not to each other) given the types of data typically available to the analyst. Moreover, comparisons of survey and registration data often reveal that marriage and fertility data supplied by male respondents are markedly inferior in quality to those from female respondents. One consequence has been that nearly all research on nonmarital fertility to date has focused on the nonmarital fertility of women. Although such an emphasis is often sensible, in that childrearing responsibilities continue to fall most heavily on women, it has nevertheless produced an asymmetry in research findings, with little being known about the men who have fathered children outside marriage.

2. Demographic Trends

Nonmarital fertility has increased substantially in both Europe and the USA. Official statistics for the USA suggest that, as of 1998, roughly one in three births (32.8 percent) were to unmarried women, a level about three times higher than that in 1970, at which time one in ten births (10.7 percent) were to unmarried women. These levels are mirrored in much of Europe. More than every third birth occurs outside marriage in France, the UK, Finland, and the former East Germany, and roughly every second birth is nonmarital in the Scandinavian countries of Iceland, Sweden, Norway, and Denmark (Kiernan 2001).

In the USA, nonmarital first births have been and continue to be prevalent among teenage women; however, the composition of nonmarital births (counting all parities) increasingly consists of those at higher birth orders to nonteenage women. As of the early 1990s, second and higher-order nonmarital births accounted for 48 percent of births to unmarried white women and fully 60 percent of births to unmarried black women. Nonmarital fertility in the USA has also increased substantially among women over age 19. In 1970, one-half of all nonmarital births were to teenage women, but by the early 1990s, nonmarital births to women aged 20 and older accounted for more than two-thirds of all nonmarital births. These period trends have been accompanied by marked declines in the likelihood of marriage following a nonmarital first birth, with some attributing this decline in marriage to changes in men's economic circumstances and others to period change in both men's and women's economic circumstances.

Nonmarital fertility in the USA continues to exhibit striking differentials by race and ethnicity. In 1998, slightly more than one in four white births (26.3 percent) were nonmarital, while nearly seven in ten black births (69.0 percent) were nonmarital. These differences also extend to births in cohabiting unions. Births to cohabiting white women rose sharply in the last two decades of the twentieth century, accounting for virtually all of the observed increases in white nonmarital fertility (Bumpass and Lu 2000). By contrast, births to cohabiting black women accounted for a far smaller proportion of nonmarital black births. Fewer than one out of five of recent nonmarital births occurred to cohabiting black women and, over time, births within cohabiting unions have constituted a declining fraction of nonmarital births to black women (Wu et al. 2001).

Nonmarital fertility in the USA and Europe differs in several important respects. The overlap between teenage and nonmarital fertility, characteristic of the USA, is largely absent in the European context. In addition, births within cohabiting unions constitute a far larger fraction of European than of US nonmarital births. Fewer than half of US nonmarital births are to parents cohabiting at the time of the birth (Wu et al. 2001), whereas the overwhelming proportion of nonmarital births in Europe are to cohabiting couples (Kiernan 2001). As a result, recent cohorts of US children are likely to spend more years, on average, growing up without a father present in the child's household than are their European counterparts.

One might expect that the decoupling of both sexual activity and childbearing from marriage and the accompanying erosion of norms proscribing nonmarital fertility would imply that nonmarital fertility has increasingly diffused beyond poor and disadvantaged populations to the non-poor and more highly educated segments of society. The available empirical evidence is in fact quite mixed on this point. On the one hand, the prevalence of cohabiting births in Europe, particularly Scandinavia, has been interpreted by some as consistent with this view. On the other hand, the expected diffusion is difficult to detect in the USA and Great Britain, where nonmarital fertility continues to occur predominantly among those with the least education and where there is accumulating evidence of diverging trends in the sociodemographic characteristics of married and unmarried mothers (see, e.g., Wu et al. 2001, Haveman et al. 2001).

3. Functionalist, Historical, and Policy Literatures

The distinction between social and biological parenthood implicit in the discussion of measurement of nonmarital fertility was an explicit emphasis of early functionalist authors in anthropology and sociology. A classic statement is Malinowski's:

> The most important moral and legal rule concerning the physiological state of kinship is that no child should be

> brought into the world without a man—and one man at that—assuming the role of sociological father, that is guardian and protector, the male link between the child and the rest of the community. I think that this generalization amounts to a universal sociological law, and as such I have called it … *The Principle of Legitimacy*. (Malinowski 1930, pp. 137–8)

Accumulating empirical evidence of violations of this principle led Goode (1960) to argue that it was more likely to be followed or enforced when fathers were high in status, as when a father possessed substantial income or assets, or when his stake in the child was high.

By contrast, historians have typically regarded nonmarital fertility as a rough empirical proxy for premarital or illicit sexual activity, although many historians have also emphasized the ebbs and flows in both the incidence of 'bastardy' in different historical periods and the tenuous connection between levels of nonmarital fertility and the intensity of social or moral disapproval. Moreover, while public and moral authorities often reacted to nonmarital childbearing as concrete evidence of fornication and adultery, real concern appears to have often centered on the financial burdens posed to the community by children born out of wedlock. Wells (1979), in particular, concludes that local authorities in colonial America were by and large 'unwilling to punish children for the sins of their parents,' noting, for example, that 'both William Franklin, the last royal governor of New Jersey, and Alexander Hamilton were able to achieve political eminence … despite the fact that their parents were not legally married.'

The question of whether the sins of parents should be visited on children remains highly relevant to contemporary policy debates. A continuing policy dilemma is that social and economic resources directed to children born outside marriage also in effect provide assistance to parents—it is often difficult to target social policies and provisions in ways that benefit children but not parents. This concern has been less evident among policy-makers in Europe, with many European nations having had a longer and more generous tradition of social welfare provisions. It has, however, played a central role in US policy debates, where concern that welfare programs such as Aid to Families with Dependent Children (AFDC) may have had the unintended consequence of increasing nonmarital births was a key factor in the dismantling of these programs in favor of those intended to discourage nonmarital childbearing.

The specific issue of whether more generous AFDC benefits have produced increases in nonmarital births in the USA has spawned an extensive research literature. The available evidence, constructed largely from cross-sectional US data (see Moffitt (1998) for a review), finds support for such an association, with most researchers finding that this association is small in magnitude. However, a persistent puzzle generated by these cross-sectional findings is that they appear inconsistent with period trends in nonmarital births and welfare generosity, with nonmarital births increasing and the inflation-adjusted value of AFDC and other benefits falling between 1975 and 1995. One possibility is that there is in fact no real inconsistency between the time-series and cross-sectional evidence, in that falling AFDC benefits may have acted to reduce nonmarital births, but that other factors may have operated in the opposite direction to increase nonmarital births. Surprisingly little work has been done with respect to this issue, save for Moffitt (2001), who presents evidence in support of this possibility.

4. *Socioeconomic Determinants*

A consistent finding in the US literature (see, e.g., McLanahan and Sandefur 1994) is that the risk of a nonmarital first birth is associated with family structure, with the standard finding being that these risks are higher for women who resided in a single-mother family at some point during adolescence. This empirical association appears remarkably robust, for example, to adjustments for unmeasured factors shared by siblings or the tendency for women with nonmarital first births to have initiated sexual activity slightly earlier than other women.

A difficulty with many of these studies is their reliance on a 'snapshot' of family structure, typically measured at age 14 or 16, which makes problematic the interpretation of what exactly it is about family structure that may account for the higher observed risk of a nonmarital first birth. Somewhat surprisingly, evidence to date suggests no support for widespread expectations concerning the intergenerational transmission of nonmarital childbearing—that being born outside marriage raises the risk that a young woman will herself bear a child outside marriage (Wu and Martinson 1993, Wu 1996, Haveman et al. 2001). Instead, what appears to be transmitted across generations is the propensity to bear a child at an early age (Haveman et al. 2001). Another consistent finding is that low income and low educational attainments of parents appear substantially associated with premarital first birth risks (Haveman and Wolfe 1994, McLanahan and Sandefur 1994, Wu 1996, Ermisch 2001). Three studies (Wu and Martinson 1993, Haveman and Wolfe 1994, Wu 1996) found that turbulent family environments, in the form of repeated changes in family structure, are associated with higher risks of a premarital first birth.

5. *Consequences for Women and Children*

Nonmarital childbearing is commonly thought to carry adverse consequences, both for the women bearing the child and for the children born into such families. An early view capturing this sentiment (Campbell 1968) held that:

> The girl who has an illegitimate child at the age of 16 suddenly has 90 percent of her life's script written for her. She will drop out of school; even if someone else in her family helps to take care of her baby, she will probably not be able to find a steady job that pays enough to provide for herself and her child; she may feel impelled to marry someone she might not otherwise have chosen. Her life choices are few, and most of them are bad.

Although this view remains highly influential in many respects, the research community has increasingly regarded it with some skepticism. For example, Furstenberg et al. (1987) reported substantial variability in the life-course outcomes for women who began childbearing while in their teenage years, many of whom had births while unmarried. These findings have caused many to question whether a woman's life script is in fact '90 percent written' as a consequence of teenage (and nonmarital) childbearing. Wu et al. (2001), using nationally representative US data, have documented similar variability for women with nonmarital births, suggesting that the results of Furstenberg et al. may also hold for more recent cohorts of women. Another basis for skepticism is the realization by researchers that previous studies may have overstated the magnitude of the negative effects of teenage childbearing. Indeed, a considerable controversy has arisen over whether teenage childbearing might be held to have any causal relationship to the later social and economic disadvantages observed for teenage mothers (Geronimus and Korenman 1993, Hoffman et al. 1993a, 1993b).

Although a large empirical literature has examined outcomes for children who resided at some point with a single parent, this literature rarely distinguishes between whether such families were formed through divorce or nonmarital fertility. Thus, surprisingly little is known concerning outcomes for children born to unmarried mothers, aside from the fact that these children are substantially more likely to grow up in poverty. As noted above, several studies have found no empirical evidence in support of the transmission of nonmarital childbearing across generations. Korenman et al. (2001) found that the higher prevalence of unintended births among unmarried mothers relative to married mothers appears to account for little of the observed differential in parenting behaviors, infant health, or child development between marital and nonmarital births. Haveman et al. (2001) have found that children born to unmarried mothers are significantly less likely to complete high school than children born to married parents.

Even less is known about the consequences of being born to cohabiting parents. Empirical studies of Great Britain and the USA (Ermisch 2001, Wu et al. 2001) have found that children born within a cohabiting union are more than twice as likely to experience the loss of a parent—most typically the father—through the dissolution of their parent's cohabiting or marital union as children born within marital unions, but this instability appears far less true in the continental European context (Kiernan 2001). As a result, it is difficult to know whether outcomes for children born within a cohabiting union would more closely resemble those of children born to single mothers or married parents. It is also worth emphasizing that what is observed—the presence or absence of a marital or cohabiting union—is but a crude proxy for the economic, social, and emotional commitments of the parents to the child, commitments which may, in turn, vary substantially over time.

6. *Conceptual and Theoretical Challenges*

Unlike empirical research, theoretical work to date has taken seriously the 'two-sex' nature of nonmarital fertility—for example, how nonmarital fertility might respond to both men's and women's characteristics. Notable examples include Wilson's (1987) argument concerning a dearth of 'marriageable' males in the US black community and Willis' economic models of nonmarital fertility (Willis and Haaga 1996, Willis 1999), in which different distributions in the relative wages of men and women may produce behavioral equilibria in which men move from a pattern of high to low paternal investment in the children they father. Neal (2000) has proposed a model in which the prevalence of nonmarital fertility can vary substantially if the choice set for women is expanded to consist not only of giving birth within or outside marriage, as assumed by both Wilson and Willis, but also of the option of foregoing a birth entirely.

The study of nonmarital fertility poses other unusual challenges for both theoretical and empirical researchers. One is that nonmarital fertility lies at the intersection of several joint (and potentially endogenous) processes by which families are formed—childbearing, marriage, and, increasingly, the formation of cohabiting unions. Many researchers have only begun to grapple with the additional analytic issues that arise when including cohabitation in their behavioral and statistical models. A second challenge concerns the likely life-course dynamics underlying nonmarital fertility. Advancing knowledge is complicated by the fact that many key aspects relevant to this process are typically unobserved by the researcher, including the timing of initial (and ensuing frequency) of sexual activity and how the resolution of a nonmarital pregnancy is weighed by the woman and by other relevant actors. A final challenge is that recent trends in nonmarital fertility may well reflect an evolving response by individuals to fundamental changes in the long-term contract represented by marriage, a contract that has itself changed in response to other social forces, including the nearly universal levels of sexual activity prior to marriage, the increasingly widespread diffusion of cohabitation, changes in the nature and levels of labor force participation of men and women,

and continuing high levels of divorce and union instability.

See also: Families and Households, Behavioral Demography of; Family Theory and the Realities of Childbearing Behavior; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants; Sexual Behavior: Sociological Perspective; Teen Sexuality; Teenage Fertility

Bibliography

Bumpass L L, Lu H-H 2000 Trends in cohabitation and implications for children's family contexts. *Population Studies* **54**: 29–41

Campbell A A 1968 The role of family planning in the reduction of poverty. *Journal of Marriage and the Family* **30**: 236–45

Ermisch J 2001 Cohabitation and childbearing outside marriage in Britain. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 109–39

Furstenberg F F Jr., Brooks-Gunn J, Morgan S P 1987 *Adolescent Mothers in Later Life*. Cambridge University Press, Cambridge, UK

Geronimus A, Korenman S 1993 The socioeconomic costs of teenage childbearing: Evidence and interpretation. *Demography* **30**: 281–96

Goode W J 1960 Illegitimacy in the Caribbean social structure. *American Sociological Review* **25**(1): 21–30

Haveman R, Wolfe B 1994 *Succeeding Generations: On the Effects of Investments in Children*. Russell Sage Foundation, New York

Haveman R, Wolfe B, Pence K 2001 Intergenerational effects of nonmarital and early childbearing. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 287–316

Hoffman S D, Foster E M, Furstenberg F F Jr. 1993a Reevaluating the costs of teenage childbearing. *Demography* **30**: 1–13

Hoffman S D, Foster E M, Furstenberg F F Jr. 1993b Reevaluating the costs of teenage childbearing: Response to Geronimus and Korenman. *Demography* **30**: 291–6

Kiernan K 2001 European perspectives on nonmarital childbearing. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 77–108

Korenman S, Kaestner R, Joyce T 2001 Unintended pregnancy and the consequences of nonmarital childbearing. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 259–86

Malinowski B 1930 Parenthood, the basis of social structure. In: Calverton V F, Schmalhausen S D (eds.) *The New Generation: The Intimate Problems of Modern Parents and Children*. Macaulay, New York, pp. 113–60

McLanahan S S, Sandefur G 1994 *Growing Up with a Single Parent*. Harvard University Press, Cambridge, MA

Moffitt R A 1998 The effect of welfare on marriage and fertility. In: Moffitt R A (ed.) *Welfare, The Family, and Reproductive Behavior: Research Perspectives*. National Academy Press, Washington, DC, pp. 50–97

Moffitt R A 2001 Welfare benefits and female headship in U.S. time series. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 143–72

Neal D 2000 The economics of family structure. Paper presented at the Workshop on Problems of the Low-Income Population. Institute for Research on Poverty, University of Wisconsin-Madison, WI

Wells R V 1979 Illegitimacy and bridal pregnancy in colonial America. In: Laslett P, Oosterveen K, Smith R M (eds.) *Bastardy and its Comparative History*. Harvard University Press, Cambridge, MA, pp. 349–61

Willis R J 1999 A theory of out-of-wedlock childbearing. *Journal of Political Economy* **106**: S33–64

Willis R J, Haaga J G 1996 Economic approaches to understanding nonmarital fertility. In: Casterline J B, Lee R D, Foote K A (eds.) *Fertility in the United States: New Patterns, New Theories*, pp. 67–86

Wilson W J 1987 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago

Wu L L 1996 Effects of family instability, income, and income instability on the risk of a premarital birth. *American Sociological Review* **61**: 386–406

Wu L L, Bumpass L L, Musick K 2001 Historical and life course trajectories of nonmarital childbearing. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 3–48

Wu L L, Martinson B C 1993 Family structure and the risk of a premarital birth. *American Sociological Review* **58**: 210–32

L. L. Wu

Fertility: Political and Political–Economic Perspectives

Since the rise of modern demography in the early twentieth century, fertility has largely been the province of demographic research. In the 1980s other fields—in particular anthropology, but also gender studies, social history, and cultural studies—began to take an active interest in childbearing as sociocultural and political–economic process. Concerned less with the numbers of children born than with the macro- and microlevel dynamics surrounding the reproduction of families, societies, and polities, these perspectives have flourished along the borders between sociocultural anthropology, population studies, gender studies, and cultural studies of science and technology. This article traces the development of three of the most vigorous streams of such research: the culture and political economy of fertility, the politics of reproduction, and Foucauldian perspectives on 'population.' Although these fields of research are multidisciplinary, because anthropologists have been at the forefront of these intellectual developments, and because of limitations of space and knowledge, this

article emphasizes anthropological work, especially that of US-based scholars.

These three perspectives are rooted in different sets of substantive concerns and theoretical preoccupations. This article describes each analytic framework in turn, tracing its intellectual and political sources, its analytic agenda and research concerns, and its status at the end of the twentieth century. It then turns to some of the methodological, ethical, and political dilemmas faced by scholars conducting research on reproductive politics and political economy. Whereas conventional research places social and economic forces at the heart of fertility change, the perspectives reviewed here emphasize instead the political, political–economic, cultural, and linguistic or discursive dimensions of reproductive transformation. With their political and cultural interests, qualitative research methodologies, and concern about researcher reflexivity, these perspectives do not so much supplant as supplement the more conventional quantitative research of demography. Their emergence on the intellectual scene in the 1980s and 1990s marks a significant broadening of interest in reproduction across the human sciences. As anthropologists and other students of social life have turned their attention to reproduction, the study of fertility has grown in theoretical sophistication. At the same time, human reproduction, long peripheral to theory in the human sciences, has assumed a more central place in social, cultural, and political theory. The appearance of these fresh ways of understanding human fertility may turn out to be one of the more exciting developments in social theory in the late twentieth century.

1. *The Culture and Political Economy of Fertility*

The first of these broadly political perspectives, the culture and political economy of fertility, grew out of the dissatisfaction of anthropologists in the mid- to late 1980s with conventional demographic approaches to fertility. A highly quantitative social science, demography has relied on relatively simple theoretical formulations to guide its research. The field's main contribution to theorizing about fertility, demographic transition theory, posits a universal fertility 'transition,' or decline from high to low levels of childbearing, that follows urbanization, industrialization, the spread of education, and the fall in infant mortality associated with the worldwide process of 'modernization.' These changes are thought to alter the economics of childbearing, lowering desired family size. Developed in the mid-1940s, demographic transition theory was part of the larger body of modernization theory that dominated social science research on Asia, Africa, and Latin America for roughly the first 25 years after World War II. In the 1970s, the consensus surrounding transition theory began to erode; by the early 1990s it was no longer an area of active theory building. What undermined faith in transition theory was the massive European Fertility Project, which failed to show the expected connections between the theory's socioeconomic variables and the historic decline in European fertility. The only 'variable' consistently related to fertility was culture, defined operationally as language, ethnicity, or geographical region. Accordingly, in the 1980s and 1990s, demographers increasingly turned their attention to 'culture,' understood primarily as ideational change, more specifically, diffusion of new ideas about the acceptability of modern contraception.

As the disciplinary experts on 'culture,' anthropologists were invited to join the search for the forces underlying fertility change. Yet anthropological notions of culture were embedded in disciplinary traditions that rooted individual behaviors such as childbearing in larger-scale structures and processes ranging from family economy and community, kin and class relations all the way to global processes such as the historic spread of capitalism and, today, the transnationalization of cultural, social, and economic life. Out of one such tradition, anthropological political economy, known now as the anthropology of globalism, emerged a culture and political economy of fertility that sought to broaden, politicize, and historicize research on childbearing across space and time. This body of work traces its intellectual lineage to Marxian models of global capitalism that emerged as critiques of modernization theory in the 1960s and 1970s, namely, dependency theory, world-systems theory, and the anthropological political economy of Eric R. Wolf. More recently, this perspective has drawn inspiration from theories of disorganized capitalism and flexible accumulation advanced by critical geographers such as David Harvey, as well as work on transnationalism by anthropologists such as Ulf Hannerz and Arjun Appadurai.

From the vantage point of these larger global political economies of social life, demographic research on fertility carves out too narrow and too micro a slice of social organization. It also underestimates the importance of power and politics, employs a relatively static notion of culture, has difficulty capturing human agency, and takes too unilinear, evolutionary, and progressivist a view of history. Demographic work also tends to adopt a Eurocentric stance according to which fertility transition is caused by and in turn causes further Westernization and reproductive Westernization is beneficial for all. Underlying these latter notions is the problematic assumption that Europe and its offshoots are superior to the rest of the world and the source of all significant change.

A culture and political economy of fertility takes as its object the location of reproductive subjects at the conjuncture of historically developed local, regional, national, transnational, and global processes. It directs attention to the embeddedness of community inst-

itutions shaping fertility in structures and processes operating at supracommunity levels, and to the historical roots of those macro–micro linkages. Students of the culture and political economy of fertility read demographic history as saying there are many reproductive transformations, each driven by a combination of forces that are to some extent institutionally, culturally, and temporally specific. Accordingly, they focus on specific transformations in different parts of the world, using generalizations derived from case studies as the building blocks of general theory. Aiming to create 'whole demographies' of particular sites, they attend to political and economic as well as social and cultural forces underlying reproductive change, stressing the interrelations among these. Emphasizing the political negotiation of all aspects of social life, the culture and political economy of reproduction places gender, class, race, and other forms of inequality at the heart of reproductive dynamics. It views culture as a cluster of negotiated meanings and motivations that, taken together, form larger cultural logics that shape reproductive aspirations and practices. In this perspective people are seen as active agents of their own reproduction, who ongoingly and often strategically tinker with biology in what Caroline Bledsoe (1990) calls the 'social construction of demographic reality.'

The social constructedness of reproductive outcomes illuminates the limitations of the biological notion, 'fertility,' though students of reproductive political economy generally continue to use the term because it facilitates communication with other population specialists. In the culture and political economy of fertility, history is indeterminate and contingent because it is actively constructed by situated actors. People everywhere are seen as part of an interconnected world in which non-Euro-Americans actively shape not only their own childbearing, but also the reproductive dynamics of the world at large. Clearly, the implied research agenda of such a perspective is very large. It includes such topics as the reproductive impact of the extension of capitalist relations of production, the rise of states interested in counting, surveiling, and policing their families, and the spread of family planning institutions and ideologies to the most remote corners of the third world.

While a great many authors incorporate political-economic insights into their work, the most explicit theoretical statements of the political–economic embedding of reproductive practices can be found in Greenhalgh (1990), contributions in Greenhalgh (1995), and Schneider and Schneider (1996). Exemplary book-length case studies include Kligman (1998) on political demography and the medicalization of reproductive repression in Ceausescu's Romania; Schneider and Schneider (1996) on the capitalist-driven, class-specific reproductive transitions of Sicily; Jeffery and Jeffery (1997) on the classed and gendered political economy of reproduction in rural India; Inhorn (1994) on the devastating consequences of infertility for Egyptian women; and Kertzer and Hogan (1989), and Kertzer (1993) on the economics and church and state politics of infant abandonment and family change in Italy.

With its global reach and historical stretch, the culture and political economy of fertility presents a vast array of important research problems, few of which have yet been resolved for any time and place. In the early 1990s, however, as anthropologists working on demographic topics began to link up with those working on the political dynamics of reproduction, this field of research began to lose its distinct character and merge with other, more political, feminist, and poststructuralist (or Foucauldian) perspectives on population.

2. *The Politics of Reproduction*

A second stream of innovative research, the politics of reproduction, coalesced into an identifiable body of work in the early 1990s. A product of the encounter between the social sciences and post-1960s feminism, with its critical views of the reproductive status quo, this work takes issue with the biomedicalization of reproduction in the industrialized world, and with the ideologies and practices of population control in the third world, both of which treat women as objects of control rather than subjects or active agents of their own reproductive lives. The politics of reproduction is not only an intellectual field of activity, although many of its participants are university based. It is better described as a scholar-activist movement aimed at transforming the politics of reproduction in a way that places control over whether, when, and how to have children in the hands of women themselves. The development and internationalization of this feminist project on reproduction were greatly stimulated by the events surrounding the United Nations International Conference on Population and Development (ICPD) held in Cairo in 1994. At that meeting women's health advocates gained significant influence over international population policy, transforming its agenda from one based on demographic concerns with population numbers to one focused on health and equity concerns with women's sexual and reproductive well-being and rights. Work on reproductive politics, then, is rooted in both intellectual and political developments.

Two relatively distinct clusters of research on reproductive politics can be distinguished: multidisciplinary critiques of international population control activities and broadly anthropological analyses of the local/global dynamics of reproductive practices, policies, and politics. The multidisciplinary critique of international population policy combines broadly neo-Marxian concerns with contemporary capitalism and class relations with feminist concerns

with the consequences of large-scale political and economic developments for poor women. Perceiving shared political and economic interests linking transnational pharmaceutical corporations, the international population control establishment, and third world population policymakers, this view holds that the top-down, technocratic, too often high-pressured interventions of third world family planning programs sacrifice women's health and safety in the name of a neo-Malthusian ideology that wrongly blames third world poverty on population growth. The ideology of population control, it suggests, obscures the true cause of growing poverty and inequality in the world. The real source of these problems is not population growth, but the post-1973 global spread of neoliberal capitalism, with the attendant privatization of state sectors of the economy and retreat of the state from support for social programs.

Some writers believe that new, post-Cairo policies have appropriated the feminist language of ICPD to advance a newly antifeminist population agenda. Whatever their position on current policy, these writers agree that the only effective solutions to these problems are to empower women economically, install popular-based health care systems that give women a voice in their own health care, and guarantee women's sexual and reproductive rights as human rights. Important statements of these critiques and proposals are, from a southern feminist perspective, Correa (1994); US-based feminist angles, Dixon-Mueller (1993) and Hartmann (1995); and, a neo-Marxist feminist approach, Bandarage (1997). Petchesky and Judd (1998) illuminate the diverse understandings of sexual and reproductive rights held by grassroots women around the world, crucial information to guide the transformation of international population policy into policies guaranteeing women's sexual and reproductive rights.

The anthropological work steps back from these charged policy debates to reflect on the complex intersections of global-to-local politics of reproduction in different cultural settings. Scholars working in this vein are interested less in the number of children women have than in the dynamics by which childbearing is negotiated by actors situated differently in hierarchies of gender, class, race, sexuality, and so forth. Starting with Friedrich Engels' insight about the mutual determination of reproduction and production, this work draws on the social anthropology of kinship, marriage, and parenting, adding to it a new interest in women's agency in negotiating the contradictory forces shaping their reproductive lives. Viewing reproduction as an aspect of the distribution of power in interconnected local and global arenas, this perspective pays particular attention to ways people use local cultural logics and social relations to incorporate, revise, resist, and sometimes even reshape larger-scale processes. Its aim is to produce accounts of reproductive politics that are culturally grounded, ethnographically nuanced, and theoretically multilayered. Key problems for research include the stratified nature of reproduction, in which certain ethnic, class, gender, sexual, and national categories of people are empowerto reproduce and nurture while others are not; the micropolitics attending the spread of Western biomedicine and its reproductive technologies worldwide; and new cultural understandings of kinship and social life that are produced as people imagine and enact the creation of the next generation. Important theoretical statements and broad-based collections on reproductive politics include Ginsburg and Rapp (1991, 1995) and Franklin and Ragone (1998). Recent case studies include Rapp's (1999) study of prenatal testing in the US, Franklin's (1997) account of assisted conception technologies in Britain, and Strathern's (1992) reflections on the implications for English kinship of the new reproductive technologies.

At the end of the twentieth century, the politics of reproduction was a burgeoning field of research activity. As new technologies for 'perfecting' reproduction continue to proliferate and spread around the world, and the politics of Cairo work themselves out in diverse transnational and southern settings, scholarly-activist interest in the gendered, raced, classed, and north–south politics of reproduction can only be expected to grow.

3. *Foucauldian Perspectives on 'Population'*

Emerging in the mid- to late 1990s, a third stream of political research reflects the extraordinary influence of the French 'poststructuralist' philosopher and social critic, Michel Foucault, on the human sciences in general and anthropology in particular. Foucault's writing have left their mark on both of the above perspectives, but in the work discussed here his voice is more pervasive, shaping not only some of the terms of discourse, but the central problems addressed. Given limitations of space, this article covers only a few of his most provocative ideas on the interlacing of power and population.

In Discipline and Punish, *History of Sexuality*, and other works, Foucault (1978, 1979) identified a new, modern form of power, biopower, which takes as its object life itself. Biopower operates by constituting its objects of knowledge, to which it then responds with the development of new knowledges, technologies, and arts of government. Far from being 'neutral' and power-free, the biological, medical, and social sciences are seen as parts of the apparatus of modern power. Through their scientific discourses, analytic categories, and intellectual practices, the natural and human sciences help to construct 'fertility,' 'the body,' 'population,' and 'the social,' in the process subjecting people to new, distinctly modern forms of power that affect how they constitute themselves as subjects.

Given the sciences' close relationships with bureaucratic states in most settings, scientific knowledges of reproduction have often served the political objectives of regimes in promoting such goals as socioeconomic modernity, eugenic purity, and national power. In Foucault's view, modern power is embodied in discourses and practices that are not concentrated in any institution, such as the state or medical establishment, but widely dispersed throughout the social system. Yet hegemony is never complete, for 'where there is power, there is resistance,' leading to unstable configurations of power that are always open to contestation. Even this brief review of some Foucauldian concepts makes clear that Foucault's ideas unsettle the received terms of debates over reproduction. Among other things, they make language a central focus of analysis, redefine power, problematize the categories in which reproduction is understood, and turn the sciences, long thought to be beyond politics, into objects of political study in their own right.

Anthropologists and other students of reproduction have taken Foucault's broad, Eurocentric ideas, grounded them historically and culturally, and given the human subject of biopower more agency as a political actor. In the 1990s, two interrelated programs of research on reproduction have emerged, one focusing on the individual body, the other on the social body, or population as a whole. These are the two poles around which biopower is said to operate. Foucauldian work on the individual reproductive body focuses largely on the discourses and practices of biomedicine, which 'medicalizes' ever more phases of the female reproductive life cycle, turning what were once 'normal' aspects of procreation into 'abnormal' medical conditions, which it then proceeds to treat. Viewing the female body as a site of biomedical domination and resistance, researchers have explored the medicalization of pregnancy, childbirth, and menopause, and the accommodations and subversions of women for whom medical interventions are simultaneously seductive and dangerous. Key collections of anthropological work on such issues are Lock and Kaufert (1998), which examines the 'microphysics of power' and its operation in everyday life, and Davis-Floyd and Sargent (1997), which traces how authoritative knowledge about childbirth is produced, displayed, resisted, and challenged in different social, political, and clinical contexts.

While the anthropological research tends to work in the present and from the bottom up, focusing on the daily encounters between ordinary subjects and the agents of reproductive power, it exists in productive conversation with research in cultural studies of technoscience. The technoscience studies more often work historically and from the top down, tracing the development of the biological life sciences and their construction of the discourses, practices, and technologies over which contemporary reproductive contests are waged. Two exemplary studies of this sort are Oudshoorn's (1994) account of endocrinology and the hormonally constructed, sex-differentiated body, and Clarke's (1998) history of the scientific study of reproduction in biology, medicine, and agriculture, which resulted in the technoscientific production of modern means of contraception.

Foucauldian work on the population as a whole focuses on the role of demography, sociology, social work, social hygiene, urban planning, and other fields in constituting inappropriate ('too high' or 'too low') fertility as a 'social problem.' Working in and with state bureaucracies, these fields first identify fertility as an object of knowledge and management, and then foster the creation of new social technologies—the census, the social survey, the urban plan, the family planning program—designed to manage it at the level of the reproductive body, the home, the family, the city, and the nation-state. A relatively new research topic, the disciplinary and bureaucratic production of 'population' has so far yielded one major monograph, Horn's (1994) study of modernity and social reproduction in 1920s and 1930s Italy. Yet the Foucauldian notion of 'population' is an exciting and important construct that is certain to produce more such studies in the years ahead.

While these three political and political–economic perspectives grew out of different sets of intellectual and political concerns, by the late 1990s, as population had moved closer to the center of social theory, the lines between them, always thinly drawn, had begun to fade and disappear. This blurring of boundaries can be expected to continue into the future, giving rise to a broad politics of reproduction and population that is multilevel in scope, multidisciplinary in approach, and politically engaged in the policy implications of the research.

4. *Methodological Issues; Ethical and Political Dilemmas*

With their interests in the political and cultural dimensions of reproduction, these perspectives require research methods that are more akin to those of anthropology than of demography. Conventional demographic research relies on the social and demographic survey, aimed at reducing sociodemographic behavior to a series of discrete, measurable variables whose relationships can be mathematically modeled. Such research produces knowledge that, although empirically useful, is thin and undercontextualized, individual in level, and collected in categories developed by the analyst. Led by John Caldwell, since the 1970s a small but growing number of demographers have recognized the shortcomings of survey methods for gathering nuanced data on multiple levels of social reality. In place of large-scale surveys, they have turned to 'microdemographic' research in the field. While microdemographic fieldwork marks an impor-

tant departure from the survey method, such fieldwork is usually short in duration, often conducted through translaters, and rarely informed by cultural and political theories of the sort discussed above.

In contrast to these demographic approaches, anthropological work on political and cultural process entails what Tom Fricke (1997) calls 'thick, cultural demography.' In thick demography, or reproductively thick ethnography, the ethnographic ear of the anthropologist is coupled with the textual eye of the cultural analyst and the documentary eye of the historian, and all are informed by in-depth knowledge of the language, culture, history, and social relations of the area studied. Thick reproductive ethnography is based on long-term ethnographic fieldwork in one or a small number of locales. Such fieldwork generally involves the use of a variety of specific methods: participant observation; small-scale censuses and/or surveys; structured and semistructured interviews; the gathering of life histories, textual materials, and so forth. The forces of interest are interconnected processes that exist on multiple levels of social reality: social, institutional, cultural. These processes are embedded in local systems of meaning and are negotiated by local actors differently placed with regard to class, clan, and other systems of social inequality. Concerned with gathering local knowledge in culturally relevant categories, the research aims to 'situate fertility,' that is, to show how it makes sense given the historically developed sociocultural and political-economic context in which it is embedded (Greenhalgh 1995). Situating fertility is an interpretive task well accomplished through narrative modes of explanation. Such forms of explanation offer 'readings' of cultural texts, oral expressions, or social situations that give coherence to something that was poorly understood.

Students of reproductive politics and culture are especially concerned with promoting self-reflexivity in their work. By self-reflexivity is meant explicit attention to the politics of one's own research practices. Attention to one's own politics entails recognition by the researcher that, despite every effort to be objective, the social scientist inevitably is a political actor whose scholarly representations have the power to create a reproductive 'other' (often 'traditional,' non-Western, high fertility) who is different from, and somehow less than, one's (scientific, Western, low fertility) self. This recognition is based not only on critiques of objectivity in science and social science generally, but also on critical studies of actual reproductive research. Two such studies have been particularly influential. Mahmood Mamdani's (1972) early account of the Khanna birth control study in northern India suggests that the scientific apparatus underlying the large-scale promotion of family planning may be based on village research that is marked by profound misunderstanding of village life on the part of foreign researchers and deliberate fabrication of lies on the part of bemused yet polite villagers. Mamdani found that demographic research of the sort conducted in Khanna strips facts of their social context, producing analytic results and policy suggestions that contradict the vital interests of local people. Agnes Riedmann's (1993) more recent and theoretically motivated study of demographic fieldwork among the Yoruba of Nigeria reveals the exercise of power at all stages of demographic research, from gaining entree to the study population to deciding on the questions to be asked, negotiating the interview process, analyzing the data gathered, and writing up the research results. Funded by first-world donors, Reidmann argues, 'world-system demography' serves first-world interests of prediction and control, while neglecting the concerns of third-world study populations and willfully disrupting traditional cultures and societies. In short, population research serves as a tool for cultural imperialism in which Western values of family and fertility are aggressively imposed on third-world study populations, legitimizing the purposeful limitation of family size. While these arguments misread the motivations of first-world researchers and understate the agency and resistance of third-world study populations, what cannot be gainsaid is that the stakes involved in fertility research are very high, and that research *sans* reflexivity involves a risk that the research itself may exacerbate inequality in the name of enhancing welfare.

Research like this generates suspicions about the power relations embodied in the numbers demographers use. In response to such concerns, some have called for a 'demography beyond, or even without, numbers.' In place of number collection, Nancy Scheper-Hughes (1997) proposes a critical interpretive methodology that ongoingly questions the neutrality of the research process, ponders the multiple and contradictory meanings that attach to the 'facts,' and recognizes the partiality of all truths—biomedical, demographic, and anthropological. Other students of reproductive culture and politics believe that numbers are useful and important—after all, they point out, without numbers one has no measure of the levels and trends of fertility, the subject of primary interest. Yet even the most vocal advocates of quantification insist that the numbers be gathered in a culturally sensitive manner and used in ways that are sensitive to the political implications of population research.

Closely related to these political problems are ethical dilemmas that attend research on such intimate topics as reproductive politics and culture. Ironically, in some ways the ethnographic methods that are preferred by scholars in these fields are more exploitative than the positivistic methods of survey research, for the ethnographer intrudes deep into his or her informants' private lives, leaving them vulnerable first to exposure and later to misinterpretation or even distortion of their viewpoints. The dilemma is that the more sensitive the method, the greater the ethical risk

of exploitation of researched by researcher. While there are no easy solutions to this dilemma, researchers can at least strive for ethical accountability by acknowledging that their own knowledge is both partial and positioned—that is, influenced by their social locations of class, gender, race, sexuality, nationality, and so forth. In research that spans difference—national, racial, class, and so forth—researchers can employ collective decision making about procedures, and commit themselves to give something back to the community in which they work. Petchesky and Judd (1998) describe an exemplary project of this sort. Finally, at the point that research becomes the basis for policymaking or other forms of political action, they can engage in a politics of coalition and affinity, in which apparently incompatible elements join together to promote a shared cause. Given the policy and programmatic consequences of reproductive research, ongoing reflexivity about its politics and ethics is of great importance. These issues have been worked and reworked with particular care in Brazil, where scholar-activists attentive to the political and cultural dimensions of reproduction have been unusually active in the formulation of population policy (Pitanguy 1994, Barroso and Correa 1995). Students of reproduction everywhere have much to learn from them.

See also: Family Theory and the Realities of Childbearing Behavior; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Childbearing; Fertility: Institutional and Political Approaches; Fertility of Single and Cohabiting Women; Fertility: Proximate Determinants; Reproductive Rights in Developing Nations

Bibliography

Bandarage A 1997 *Women, Population and Global Crisis: A Political-Economic Analysis*. Zed, London

Barroso C, Correa S 1995 Public servants, professionals, and feminists: The politics of contraceptive research in Brazil. In: Ginsburg F D, Rapp R (eds.) *Conceiving the New World Order*. University of California Press, Berkeley, CA

Bledsoe C 1990 The politics of children: Fosterage and the social management of fertility among the Mende of Sierra Leone. In: Handwerker W P (ed.) *Births and Power: Social Change and the Politics of Reproduction*. Westview Press, Boulder, CO

Clarke A E 1998 *Disciplining Reproduction: Modernity, American Life Sciences, and the Problems of Sex*. University of California Press, Berkeley, CA

Correa S 1994 *Population and Reproductive Rights: Feminist Perspectives from the South*. Zed, London

Davis-Floyd R E, Sargent C F (eds.) 1997 *Childbirth and Authoritative Knowledge: Cross-Cultural Perspectives*. University of California Press, Berkeley, CA

Dixon-Mueller R 1993 *Population Policy and Women's Rights: Transforming Reproductive Choice*. Praeger, Westport, CT

Foucault M 1978 *The History of Sexuality, Vol. 1: An Introduction*. Pantheon Books, New York

Foucault M 1979 *Discipline and Punish: the Birth of the Prison*. Vintage Books, New York

Franklin S 1997 *Embodied Progress: A Cultural Account of Assisted Conception*. Routledge, London

Franklin S, Ragone H (eds.) 1998 *Reproducing Reproduction: Kinship, Power, and Technological Innovation*. University of Pennsylvania Press, Philadelphia

Fricke T 1997 Culture theory and demographic process: Toward a thicker demography. In: Kertzer D I, Fricke T (eds.) *Anthropological Demography*. University of Chicago Press, Chicago

Ginsburg F, Rapp R 1991 The politics of reproduction. *Annual Review of Anthropology* **20**: 311–43

Ginsburg, F D, Rapp R (eds.) 1995 *Conceiving the New World Order: The Global Politics of Reproduction*. University of California Press, Berkeley, CA

Greenhalgh S 1990 Toward a political economy of fertility: Anthropological contributions. *Population and Development Review* **16**: 85–106

Greenhalgh S (ed.) 1995 *Situating Fertility: Anthropology and Demographic Inquiry*. Cambridge University Press, Cambridge, UK

Hartmann B 1995 *Reproductive Rights and Wrongs: The Global Politics of Population Control*, rev. edn. South End Press, Boston

Horn D G 1994 *Social Bodies: Science, Reproduction, and Italian Modernity*. Princeton University Press, Princeton, NJ

Inhorn M C 1994 *Quest for Conception: Gender, Infertility, and Egyptian Medical Traditions*. University of Pennsylvania Press, Philadelphia

Jeffery R, Jeffery P 1997 *Population, Gender and Politics: Demographic Change in Rural North India*. Cambridge University Press, Cambridge, UK

Kertzer D I 1993 *Sacrificed for Honor: Italian Infant Abandonment and the Politics of Reproductive Control*. Beacon, Boston

Kertzer D I, Hogan D P 1989 *Family, Political Economy, and Demographic Change: The Transformation of Life in Casalecchio, Italy, 1861–1921*. University of Wisconsin Press, Madison, WI

Kertzer D I, Fricke T (eds.) 1997 *Anthropological Demography: Toward a New Synthesis*. University of Chicago Press, Chicago

Kligman G 1998 *The Politics of Duplicity: Controlling Reproduction in Ceausescu's Romania*. University of California Press, Berkeley, CA

Lock M, Kaufert P A (eds.) 1998 *Pragmatic Women and Body Politics*. Cambridge University Press, Cambridge, UK

Mamdani M 1972 *The Myth of Population Control: Family, Caste, and Class in an Indian Village*. Monthly Review Press, New York

Oudshoorn N 1994 *Beyond the Natural Body: An Archaeology of Sex Hormones*. Routledge, London

Petchesky R P, Judd K (eds.) 1998 *Negotiating Reproductive Rights: Women's Perspectives Across Countries and Cultures*. Zed, London

Pitanguy J 1994 Feminist politics and reproductive rights: The case of Brazil. In: Sen G, Snow R C (eds.) *Power and Decision: The Social Control of Reproduction*. Harvard Center for Population and Development Studies, Cambridge, MA

Rapp R 1999 *Moral Pioneers: Fetuses, Families, and Amniocentesis*. Routledge, New York

Riedmann A 1993 *Science that Colonizes: A Critique of Fertility Studies in Africa*. Temple University Press, Philadelphia

Scheper-Hughes N 1997 Demography without numbers. In: Kertzer D I, Fricke T (eds.) *Anthropological Demography*. University of Chicago Press, Chicago
Schneider J C, Schneider P T 1996 *Festival of the Poor: Fertility Decline and the Ideology of Class in Sicily, 1860–1980*. University of Arizona Press, Tucson, AZ
Strathern M 1992 *Reproducing the Future: Anthropology, Kinship and the New Reproductive Technologies*. Routledge, New York

S. Greenhalgh

Fertility: Proximate Determinants

1. Background

Fertility in human populations is determined by a broad variety of factors. Specifying the relative importance among them as well as understanding the mechanisms through which they affect fertility long eluded scientists. Davis and Blake (1956) brought about an invaluable advancement by creating a conceptual analytical framework which distinguishes 'intermediate' (direct) from social, economic, political, environmental, and other (indirect) fertility determinants (see *Fertility: Political and Political—Economic Perspectives*; *Demographic Measurement: General Issues and Measures of Fertility*; *Fertility Transition: Cultural Explanations*; *Fertility Transition: Economic Explanations*; *Fertility Transition: China*; *Fertility Transition: East Asia*; *Fertility Transition: Southeast Asia*). The principal revelation was twofold. The distinction between the two types of determinants and the penetrating insight that societal factors affecting fertility can operate only indirectly through the intermediate or proximate determinants, which directly modify fertility by affecting conception, gestation, and successful parturition. In other words, social, economic, and cultural circumstances affect human behavior regarding patterns of entering into and remaining in or exiting from stable sexual unions, and practicing breast-feeding, contraception, or induced abortion. It is this behavior which produces the level of fertility and its changes generate fertility trends.

Further crucial advancements were achieved two decades later. Bongaarts (1978, 1982)

(a) identified a simplified set of proximate fertility determinants (seven rather than eleven);

(b) established which of these are central in creating fertility differentials and trends (four out of the seven); and

(c) devised formulas to measure these factors from generally available data, and how to quantify their respective impact on fertility levels and trends in a comprehensive reasonably simple model.

The direct relationship between the proximate determinants and fertility is embodied in the fact that the direction of change of fertility is given by the direction of change of the respective proximate determinant. If, for instance, breast-feeding declines, fertility increases. An increase in contraceptive prevalence or the practice of induced abortion will engender a fertility decline. To generalize, if a proximate determinant changes then fertility necessarily changes also, and the change is in a predictable direction, provided the other proximate determinants remain constant. This is not true for an indirect determinant. An increase in educational attainment may lead to a decline in breast-feeding which in turn signifies increased fertility. A similar rise in education in another population may cause contraceptive prevalence to increase which in turn signifies a fertility decline. Also, both of these mechanisms may be operating simultaneously in which case they cancel each other out and fertility will not change. Detailed analyses are required to determine how changes in the social, economic, cultural, religious, and environmental circumstances affect proximate determinants so that their effect on fertility trends can be understood.

2. The Proximate Determinants

The proximate determinants can best be identified by examining the events that most immediately influence the duration of the reproductive period and the rate of childbearing within it. A more detailed discussion follows a brief overview.

The beginning and end of the reproductive period are biologically determined. The potential reproductive years start at menarche, the woman's first ovulation which reveals itself in the first menstruation. The childbearing years end with the cessation of ovulation, i.e., the onset of permanent sterility, which occurs before or at menopause. The effective reproductive period is equal to the time women spend in stable sexual unions, marriages, and consensual unions. Consequently, the level of fertility is dependent directly on the length of time women spend in stable sexual unions.

The next important circumstance shaping the level of fertility is the average number of children borne by women during their reproductive life span. It is inversely related to the length of the periods between births. Short periods between births imply high fertility, i.e., a relatively large number of children born per woman, and vice versa.

The period between births has several components. A birth is followed by postpartum infecundability, a fecundable or ovulatory period, and a subsequent pregnancy. The length of the postpartum infecundability interval after a birth during which a woman cannot conceive is determined mainly by the length and intensity of breast-feeding. This period lasts through the first postpartum ovulation which is at the same time the beginning of the fecundable interval,

also called the ovulatory interval or the waiting time to conception. The length of this period is determined by fecundability, the monthly probability of conception, by the monthly risk of conception in the absence of contraception, and by the effectiveness of any contraception practiced. Once conception takes place it can be followed by a full-term pregnancy of nine months, or by an incomplete pregnancy terminated by a spontaneous or induced intrauterine death. Postpartum infecundability after an intrauterine death tends to be much shorter than following a birth.

The overview of the reproductive process identifies the following proximate determinants:

(a) The beginning and discontinuation of stable sexual unions, marriages, and consensual unions.
(b) The onset of sterility.
(c) Postpartum infecundability.
(d) Fecundability.
(e) Contraception.
(f) Spontaneous intrauterine mortality.
(g) Induced abortion.

2.1 The Beginning and Discontinuation of Stable Sexual Unions

Fertility is related directly to the duration of time a woman is exposed to conception, which basically is the period she is in a stable sexual union. A widespread form of such unions is formal marriage. In Latin America and the Caribbean, in parts of Africa and in the Nordic countries of Europe consensual unions have traditionally been common. In the second half of the twentieth century informal cohabitation is increasingly practiced throughout Europe and other developed countries. Occasional sexual encounters can also have an effect on fertility, but this tends to be small. The average age of union formation varies considerably. Traditionally it has been relatively early with the exception of most European countries. There the mean age of marriage was often about 25, whereas elsewhere it was frequently around 17. In recent decades of rapid social change the mean age of marriage has been increasing almost everywhere, including the developing countries. The considerable postponement of marriage taking place in many developing countries is an important factor in lowering fertility. Trends in the formation of consensual unions may have been similar, however, data to confirm this assumption worldwide are lacking.

Proportions of women who never marry differ among populations. Such women were rare in traditional societies, whereas in the developed countries their proportions have at times been 10 percent or more.

Widowhood is an important factor modifying fertility as long as mortality is high and women bear children late in their reproductive period. With declining mortality only small proportions of women loose their spouses before the end of the childbearing period.

In recent decades divorce rates have been increasing, mainly in many developed countries. The impact of this trend on fertility has been insignificant as remarriage is frequent.

2.2 The Onset of Sterility

The average age of menopause is around 50. Effective sterility starts several years earlier due to irregular ovulation among many aging women and the higher risk of spontaneous intrauterine death with increasing age. A certain increased sterility of spouses has an additional effect of lowering the age of couple sterility. An average age of last birth around 40 in societies not practicing contraception confirms the assumption that combined sterility sets in when women are in their lower forties.

2.3 Postpartum Infecundability

Breast-feeding is nature's contraceptive. The mother provides the infant with food and protection and the process of suckling acts as a contraceptive for the mother. In some known contemporary hunter-gatherer societies, infants are constantly with their mothers and can suckle up to several times an hour. Here births tend to be spaced four years apart without the use of any contraception. Suckling on the nipple produces nervous impulses which pass from the breast to the brain thus inhibiting the release of hormones responsible for ovulation from the pituitary gland (see Potts and Short 1999 and *Fertility Control: Prevalence and Consequences of Breastfeeding*). In the absence of breast-feeding menses return shortly after birth, on average after 1.5–2 months. In developed countries periods of exclusive breast-feeding have been reduced to only a few months. In developing countries profound changes are also underway, but the postpartum infecundability period may still be considerable. The longest average duration documented by a national survey was in Bangladesh in 1975, namely 18.6 months.

2.4 Fecundability

The monthly probability of conception, fecundability, is at most 0.15–0.25 among newlyweds and declines with increasing age and duration of marriage. It is this low because fertilization can occur only during about two days around the time of ovulation in the middle of the menstrual cycle. In addition some cycles are anovulatory and a substantial proportion of fertilized eggs fail to implant or are aborted spontaneously during the first two weeks after fertilization. Typical

delays to conception range from five to ten months. It is assumed that there are no significant differences between populations in this respect. The overall effect of involuntary sterility tends to be minor and significant variations between populations are the exception.

2.5 Contraception

Contraception is arguably the main factor modifying fertility in contemporary societies. There have been exceptions in eastern Europe where for some periods of time the practice of induced abortion was extremely high and thus more important in inhibiting fertility. In a few developing countries the use of contraception is still low, but practically everywhere it has been growing. During the second half of the twentieth century the unprecedented increase in the use of contraceptives and the availability of new modern more effective contraceptives have led to the perception that a 'contraceptive revolution' has taken place (see Westoff and Ryder 1977).

The effectiveness of contraception is measured as the percentage reduction in fecundability. For instance, a contraceptive with a 95 percent effectiveness used by women with a 0.2 fecundability will yield a 0.01 monthly probability of conception. Actual contraceptive effectiveness of a method depends primarily on the extent to which it is used as prescribed. Improper use results in contraceptive failure. The effectiveness of modern methods, such as oral contraceptives and intrauterine devices, is more than 95 percent in developed countries and that of conventional methods, such as the condom, diaphragm, or spermicides, about 90 percent. Depending on the knowledge and motivation of the average user, those levels tend to be lower in developing countries. In a number of countries female and/or male sterilization is popular and is 100 percent effective. The mix of methods used differs very much from one country to another.

2.6 Spontaneous Intrauterine Mortality

While there is no doubt that many conceptions are spontaneously aborted, it has proven difficult to estimate the magnitude of this factor. Not only retrospectively but even for women in prospective studies it is very difficult to accurately determine and report intrauterine deaths in the early stages of pregnancy. Carefully designed studies have provided estimates of 20 percent of conceptions not resulting in live births with half of these occurring before the third month of pregnancy. It is known that the rate of intrauterine mortality is the lowest in the middle of the reproductive period. Differences between societies appear to be insignificant, except with regard to stillbirths, defined as intrauterine deaths after the twenty-eighth week of pregnancy. These occur more frequently in some poor countries, 4 percent of conceptions compared with 1 percent in developed countries, presumably due to health and environmental factors.

2.7 Induced Abortions

Induced abortions have been practiced in every society and usually play a meaningful role in some stage of a country's fertility transition. The incidence depends mainly on motivation to restrict childbearing, the availability and use of contraception, and to some extent on the legal status of induced abortion (see *Demographic Techniques: Family Reconstitution*). The practice of induced abortion was exceptionally widespread in some countries of eastern Europe in the second half of the twentieth century where they were legalized earlier than elsewhere and where access to contraceptives, especially modern ones, was restricted. Around 1990 in Romania, for instance, three-quarters of all pregnancies were intentionally terminated. In contrast, the incidence was low in west European countries with easy access to modern contraceptives.

Table 1
Rating of proximate determinants with respect to sensitivity and variability among populations

Proximate determinants	Sensitivity of fertility to determinant	Variability among populations	Overall rating
1. The beginning and discontinuation of stable sexual unions, marriages, and consensual unions	+ + +	+ + +	+ + +
2. The onset of sterility	+ +	+	+
3. Postpartum infecundability	+ +	+ + +	+ + +
4. Fecundability	+ +	+ +	+ +
5. Contraception	+ + +	+ + +	+ + +
6. Spontaneous intrauterine mortality	+	+	+
7. Induced abortion	+ +	+ + +	+ + +

Source: Bongaarts (1982)
+ + + = high; + + = medium; + = low or absent.

During the second half of the twentieth century the number of countries with liberal induced abortion legislation was increasing steadily. By the end of the twentieth century over 60 percent of the world's population was living in such countries. In contrast, 25 percent resided in countries where induced abortions were generally prohibited. Almost all of the latter are developing countries, where induced abortions are performed frequently by unqualified personnel under unsanitary conditions and thus women's health is often seriously impaired.

3. *The Relative Importance of Proximate Determinants*

The joint effect of all the proximate variables determines fertility levels and trends. For the analysis of fertility differentials and trends only those determinants are meaningful whose effect differs in time and space. If large variations of a determinant have only a small effect on fertility trends and differentials, or if a determinant hardly varies among populations, its analysis will not yield useful scientific or policy relevant insights. Thus the selection of the analytically important factors depends on how sensitive fertility is with respect to a particular determinant and the variability of the determinant among populations. In Table 1 the proximate determinants are rated according to these criteria.

Abundant evidence confirms that fertility is the least sensitive to the risk of spontaneous intrauterine mortality and the most sensitive to time spent in sexual unions and contraception. Variability is the lowest for onset of sterility and the risk of spontaneous intrauterine mortality. Altogether, for the analysis of fertility levels and trends, the four most important proximate determinants are the time spent in stable sexual unions, postpartum infecundability, contraception, and induced abortion.

See also: Adolescent Behavior: Demographic; Family Theory: Economics of Childbearing; Fertility and Culture: Anthropological Insights; Fertility Control: Overview; Fertility Control: Prevalence and Consequences of Breastfeeding; Fertility: Institutional and Political Approaches; Fertility of Single and Cohabiting Women; Fertility: Political and Political–Economic Perspectives; Teenage Fertility

Bibliography

Bongaarts J 1978 Framework for analyzing proximate determinants of fertility. *Population and Development Review*. **4 (1)**: 105–32

Bongaarts J 1982 The fertility-inhibiting effects of the intermediate fertility variables. *Studies in Family Planning*. **13 (6–7)**: 179–89

Bongaarts J, Potter R G 1983 *Fertility, Biology, and Behavior: An Analysis of the Proximate Determinants*. Academic Press, New York and London

Davis K, Blake J 1956 Social structure and fertility: An analytic framework. *Economic Development and Cultural Change*. **4** (3): 211–235

Leridon H 1977 *Human Fertility: The Basic Components*. University of Chicago Press, Chicago, IL

Potts M, Short R 1999 *Ever Since Adam and Eve: The Evolution of Human Sexuality*. Cambridge University Press, Cambridge, UK

Westoff C F, Ryder N B 1977 *The Contraceptive Revolution*. Princeton University Press, Princeton, NJ

T. Frejka

Fertility Theory: Caldwell's Theory of Intergenerational Wealth Flows

1. *Introduction*

John Caldwell's wealth flows theory proposes a direct link between family structure and fertility (Caldwell 1976b). According to the theory, there are only two major forms of family structure, differing principally in the direction of wealth flows among generations. In 'primitive' and 'traditional' societies, net wealth flows are primarily upward from younger to older generations, and individual interests are subjugated to corporate interests. In developed nations, family structure is organized in terms of downward wealth flows where parents are expected to provide for children's economic well-being.

The theory proposes that fertility decisions in all societies are economically rational responses to familial wealth flows. In societies with net upward wealth flows, the economically rational decision is to have as many surviving children as possible (within the constraints imposed by biology), because each additional child adds positively to a parent's wealth, security in old age, and social and political well-being. In societies with net downward wealth flows, the economically rational decision is to have no children or the minimum number allowed by a psychological disposition that derives pleasure from children and parenting. The worldwide transition from high to low fertility is the result of a change in family structures from upward to downward wealth flows.

This change in family structure was due to the spread of new values that placed a premium on individual satisfaction and achievement (Caldwell 1980). Those values emanated from the educated, middle-class in the west and are now being exported to the developing world through mass formal education. Implicit in the educational materials and expectations of schools is the individualistic value system that

produces downward wealth flows. Labor markets make the adoption of these individualistic values, which are adversative to the family or group production characteristic of pretransition societies, economically feasible. The transition from 'traditional' to 'modern' family structure occurs when a critical mass of individuals adopt the new values, and respond with low fertility. The attainment of mass education in a country should therefore precipitate and predict the fertility transition.

The most serious challenge to the wealth flows theory has come from evolutionary biologists. The argument that prior to modernization, upward wealth flows characterized human family structures is inherently antithetical to theory in evolutionary biology (Turke 1989). Biologists expect net wealth flows to be downward in all organisms, including humans. This research generated by the evolutionary biological critique of wealth flows theory has stimulated the development of a new theoretical framework for understanding human demographic patterns (see *Fertility Theory: Embodied-capital Theory of Life History Evolution*).

2. *Historical Context*

Wealth flows theory is not unique in proposing a link between costs of children and fertility. Notestein (1945, see *Fertility Transition: Economic Explanations*) proposed that the cost of children and their economic value are major determinants of fertility. Becker (1960) also developed formal models of the demand for children based on the household production function, focusing on the tradeoff between quantity and quality of children. Leibenstein (1957), also building on Notestein, argued that two forms of utility derived from children in pretransition (agricultural) societies were the proceeds of child labor and old age security. Evidence regarding the value of children's labor in agricultural societies was offered by Boserup (1965), Clark (1967), and Mamdani (1972). During the 1970s Cain (1978), Caldwell (1976b), Mueller (1976), and Nag et al. (1978) all developed theories concerning the economic contribution of children and the importance of child labor in peasant societies and how these factors impact family size and structure.

Caldwell's 'wealth flows' theory differs from the others by expanding the definition of intergenerational transfers across the life course and by directly linking changing value systems regarding intergenerational transfers of wealth to fertility transition. Defining wealth as 'all the money, goods, services, and guarantees that one person provides to another' (Caldwell 1982, p. 333), the theory directly incorporates effects on status and political position as well as on material wealth. Together, those material and nonmaterial transfers across the life course determine the net flow of wealth between generations and also affect societal norms regarding the emotional value of children. As with economic models, Caldwell's theory recognizes technological impacts on wealth flows, such as the changing economic value of education. While there is unique pleasure derived from having children, pleasure will be derived from fewer children as their economic value decreases or the costs of educating them increase (Caldwell 1982, p. 338). However, unlike the more strictly economic models (Becker 1981), value systems regarding support in old age and the relationship between the individual and the larger family are the primary determinants of intrafamilial wealth flows in Caldwell's view.

3. *The Economic Contribution of Minor Children*

Wealth flows theory has been tested in several ways. Some tests are designed to determine if changes in the economic roles of children are associated with changes in fertility. Several researchers have found evidence of such effects in sub-Saharan Africa. In addition to Caldwell's (1976a) original microlevel study in southwest Nigeria, DeLancey (1990) and Makinwa-Adebusoye (1994) both found through meta-analysis of sub-Saharan African national level survey data that parents perceived upward wealth flows in high fertility contexts.

Similar results have been reported in several Asian countries with national level data. Chang (1990) reports that the adoption of socialist economic principles in China may reduce the economic contribution of children and result in fertility decline, but Moore (1998) suggests recent economic policy initiatives encouraging family economic independence may promote higher fertility. Some studies, using micro-level data, do not find evidence of upward intergenerational wealth transfers in high fertility contexts in India, (Das Gupta 1994 (in Punjab), Vlassoff 1982).

Willis (1982) has criticized Caldwell's hypothesis strongly on theoretical grounds, arguing that members of an altruistic older generation will see offspring success as a substitute for their own, and will therefore be willing to invest in offspring human capital at a higher rate than would be predicted if parents were acting solely in their own self-interests.

In general, the tests discussed above are weak because they do not measure the absolute flow of wealth but rather indicators of changes in the economic contributions of children. Strong tests of the theory require that children provide *net* economic benefits (i.e., amount provided minus amount received) to parents in high fertility societies and net economic costs in low fertility societies. Studies that have actually measured the net flow of wealth in high fertility societies do not support a strict interpretation of the theory. On the basis of data on net production and consumption through the life course, Kaplan (1994) has shown that in three lowland South Amer-

ican indigenous populations engaged in traditional economic pursuits of foraging and swidden horticulture, offspring do not become net producers until their twenties and never compensate their parents in terms of economic investment. Mueller (1976) in a study of 27 peasant agricultural societies found that while children were able to contribute to the household economy at a far higher rate than those in Kaplan's study, they still do not become net producers until their late teens and never repay the initial parental investment. In a review of all published studies, Lee (1996) concludes there are no data showing a net wealth flow from children to parents for any society—ironically, net upward wealth flows between generations are found in *low* fertility societies when public transfers, such as Medicare and social security, are included (Lee 1996).

4. *Old Age Security*

According to wealth flows theory, a second period in the life course for parents to receive economic benefit from their offspring is during old age. Those benefits motivate high fertility in traditional societies where adult children are expected to support their aging parents. In modern society, the ethic of individualism and public supports for the aged remove this motivation. Empirical evidence bearing on this hypothesis is also mixed. Hugo (1997), using national level data in Indonesia, found high fertility to be associated positively with perceptions regarding the likelihood of receiving old age support. Using historical data from England and Wales, Schellekens (1993) also found that old age security was correlated with high fertility. Cain has argued that children provide risk insurance for the elderly in south Asia (Cain 1981). Using microlevel data from Bangladesh, he also suggests that women desire high fertility to receive support from sons in old age as insurance against the loss of support from husbands in this patriarchical society. In contrast to Caldwell, however, Cain also argues that the breakdown of extended kin networks makes women even more dependent on sons for security. Vlassoff and Vlassoff (1980), using microlevel data from India, found no evidence that old age security motivated high fertility levels. (See Dharmalingam (1994) for a review of the debate between Cain and the Vlassoffs.)

Several studies using national level data have found that there are high expectations or levels of old age support in peasant agricultural societies in sub-Saharan Africa (De Lancey 1990), Mexico (regional level data) (Nugent and Gillaspy 1983), Malaysia (Lillard and Willis 1997), and Taiwan and the Philippines (Lee et al. 1994). Many studies have found extremely low levels of support or expectations of support of the elderly by adult children in the United States and other industrialized countries (Eggebeen 1992). None of these studies in the developing or developed world were explicitly concerned with the relationship of fertility to old age security.

Critiques of wealth flows theory concerned with the economic contribution to the household of minor children are equally appropriate with regards to a net upward flow of wealth realized as support of the elderly. The evidence in support of wealth flows theory is based upon associations of high fertility with expectations of support. There is no solid evidence about whether high fertility actually increases the likelihood of support at the individual level, nor about the net flow of goods and services between older people and their descendent kin in high fertility societies. In fact, there is some evidence that older men and women make contributions to household economics in peasant agricultural societies through the skills and knowledge they bring to crisis situations. Researchers working with hunter-gatherers and forager-horticulturalists also observe that elderly individuals are often net producers and engage in other forms of work to support members of younger generations (Hawkes and Jones 1997).

5. *Mass Education*

Education plays a prominent role in many theories of fertility transition. In the wealth flows theory, education affects fertility through three routes, all affecting the net costs of rearing children. First, education subverts traditional communally-oriented value systems and trains children to pursue individual goals at the expense of supporting their parents (see also Cochrane 1979); second, it reduces children's ability to contribute to the household economy through diversion of time and loss of traditional skills and knowledge (Bock 2001, Caldwell 1980); and third, it increases the direct and indirect costs of children by increasing their demand for consumer goods and by establishing childhood as a period for receiving investment rather than contributing to the household economy (Caldwell 1980).

While there is a great deal of empirical evidence regarding the relationship between education and fertility, less is known about the relative contributions of these alternative causal routes, and particularly about education's impact on fertility through changing values regarding familial wealth flows. The empirical evidence regarding the impact of education on values is mixed. A meta-analysis of national level data from 37 sub-Saharan African countries found that primary school enrollment for both males and females in 1960 and 1980 had no significant effects on total fertility rate 15 and 30 years later. Secondary school enrollment for males and females in 1960 had a weak negative effect on total fertility rate, while only male secondary school enrollment in 1980 had a strong negative effect on actual and estimated total fertility rates 10 to 25 years in the future (Adamchak and Ntseane 1992).

Several studies show that children's school attendance affects parental fertility, (see, for example, Axinn (1993)). However, the causal process underlying this association is unclear, especially since parents who actively seek formal education for their children may differ from parents who do not with respect to characteristics that are themselves determinants of lower fertility. Lloyd et al. (2000) also found that children's school attendance predicts the timing of fertility transitions in a cross-national sample of African countries. Many studies report a negative impact of women's education and fertility, but again several alternative causal hypotheses are consistent with the data. At present, there is no strong evidence for or against the strict interpretation of fertility transition as a cultural watershed generated by mass education.

6. *Summary of Wealth Flows Theory*

Consistent with the logic of wealth flows theory, there is a great deal of suggestive evidence that changes in the costs of rearing children and in parental perceptions of those costs are associated with fertility transition. People in many high fertility societies also report higher expectations regarding economic support during old age than do people in low fertility societies. Also consistent with the theory is that education and fertility are negatively related in virtually every study measuring their association.

There are two major deficiencies in the theory, however. The first is that the data do not support a strict interpretation of the wealth flows hypothesis. There is no direct quantitative evidence of net-upward wealth flows in any traditional high-fertility society. Virtually all studies supporting the theory have used national-level, cross-sectional data, and have relied on proxy measures and informant reports. Actual measures of wealth flows in micro-level, longitudinal, ethnographic studies, using behavioral observation and measurement of intrahousehold productivity and exchange, all show *downward* wealth flows in traditional high fertility societies. Caldwell recognized that adequate tests of theory would require longitudinal, microlevel, ethnographically oriented research due to the subtleties and complexity of intrafamilial wealth flows (Caldwell 1982).

The second deficiency is that the theoretical foundations for the determination of familial wealth flows are not well specified. This is evident in the treatment of education. In the theory, education is at once a vehicle of values transmission, a response to micro- and macroeconomic forces, and a direct cost to parents. The theory does not specify why certain value systems governing wealth flows arise, nor the relationships between exogenous economic factors, human psychology, and cultural evolution.

7. *Conclusion*

According to wealth flows theory, cultural transmission of new family values is the principal driving force in fertility transition. However, without a theory which can link ecological conditions to cultural features and their distribution, the concept of culture risks being vague, both cause and effect, the unexplained explanator. Culture is many things: a body of accumulated information, a set of customs and rituals, group beliefs, and a mindset about the organization of the social and physical environments. Each of these components may be related to fertility; but a generalized culture concept cannot be successfully used to explain patterning and variation in fertility without risking causal circularity (Bock 1999).

The wealth flows theory of fertility transition is a major contribution to demography. The theory is general in that it explains both high and low fertility regimes. A second strength is that it addresses both exogenous variables and the psychological and social processes that translated those exogenous variables into individual behavior, emphasizing both economic and social factors in understanding fertility transitions. The theory helped to lay the groundwork for theories of the family to become a major perspective within the field of demography, and broadened the scope of demography by directing attention to culture. It also inspired and influenced a large number of empirical studies. However, in the strict sense of predicting two fertility regimes to be associated with upward and downward wealth flows, empirical evidence does not support the theory. In the weak sense of predicting a negative relationship between net wealth flows to children and fertility, there is a good deal of empirical support.

Bibliography

Adamchak D J, Ntseane P G 1992 Gender, education, and fertility: A cross-national analysis of sub-Saharan African nations. *Sociological Spectrum* **12**: 167–82

Axinn W G 1993 The effects of children's schooling on fertility limitation. *Population Studies – Journal of Demography* **47**: 481–93

Becker G S 1960 An economic analysis of fertility. In: National Bureau of Economic Research (ed.) *Demographic and Economic Change in Developed Countries*. Princeton University Press, Princeton, NJ, pp. 209–31

Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Bock J 1999 Evolutionary approaches to population: Implications for research and policy. *Population and Environment* **21**: 193–222

Bock J 2001 Learning, life history, and productivity: Children's lives in the Okavango Delta, Botswana. *Human Nature*, in press

Boserup E 1965 *The Conditions of Agricultural Growth*. Aldine, New York

Cain M T 1978 The household life cycle and economic mobility in rural Bangladesh. *Population and Development Review* **4**: 421–38
Cain M 1981 Risk and insurance: Perspectives on fertility and agrarian change in India and Bangladesh. *Population and Development Review* **7**: 435–74
Caldwell J C 1976a Fertility and the household economy in Nigeria. *Journal of Comparative Family Studies* **7**: 193–253
Caldwell J C 1976b Toward a restatement of demographic transition theory. *Population and Development Review* **2**: 321–66
Caldwell J C 1980 Mass education as a determinant of the timing of fertility decline. *Population and Development Review* **6**: 225–55
Caldwell J C 1982 *Theory of Fertility Decline*. Academic Press, London
Chang K-S 1990 Socialist institutions and family wealth flows reversal: An assessment of post-revolutionary Chinese rural fertility. *Journal of Family History* **15**: 179–200
Clark C 1967 *Population Growth and Land Use*. St. Martin's Press, New York
Cochrane S H 1979 *Fertility and Education*: *What Do We Really Know*? Johns Hopkins University Press, Baltimore, MD
Das Gupta M 1994 What motivates fertility decline? A case study from Punjab, India. In: Egero B, Hammarskjold M (eds.) *Understanding Reproductive Change:Kenya*, *Tamil Nadu*, *Punjab*, *Costa Rica*. Lund University Press, Lund, Sweden, pp. 101–33
DeLancey V 1990 Socioeconomic consequences of high fertility for the family. In: Acsadi G T F, Acsadi G J, Bulatao R A (eds.) *Population Growth and Reproduction in Sub Saharan Africa: Technical Analyses of Fertility and its Consequences*. The World Bank, Washington, DC, pp. 115–30
Dharmalingam A 1994 Old age support: Expectations and experiences in a south Indian village. *Population Studies – Journal of Demography* **48**: 5–19
Eggebeen D J 1992 Family structure and intergenerational exchanges. *Research on Aging* **14**: 427–47
Hawkes K, Jones N G B 1997 Hadza women's time allocation; offspring provisioning; and the evolution of long postmenopausal life spans. *Current Anthropology* **38**: 551–77
Hugo G 1997 Intergenerational wealth flows and the elderly in Indonesia. In: Jones G W, Douglas R M, Caldwell J C, D'Souza R M (eds.) *The Continuing Demographic Transition*. Oxford University Press, New York, pp. 111–33
Kaplan H 1994 Evolutionary and wealth flows theories of fertility: Empirical tests and new models. *Population and Development Review* **20**: 753–91
Lee R 1996 A cross-cultural perspective on intergenerational transfers and the economic life cycle. *Seminar on intergenerational economic relations and demographic change*: *papers*. International Union for the Scientific Study of Population [IUSSP], Committee on Economic Demography, Liege, Belgium, pp. 1–25
Lee Y J, Parish W L, Willis R J 1994 Sons, daughters, and intergenerational support in Taiwan. *American Journal of Sociology* **99**: 1010–41
Liebenstein H M 1957 *Economic Backwardness and Economic Growth*. Wiley, New York
Lillard L A, Willis R J 1997 Motives for intergenerational transfers: Evidence from Malaysia. *Demography* **34**: 115–34
Lloyd C B, Kaufman C E, Hewett P 1999 *The Spread of Primary Schooling in Sub-Saharan Africa: Implications for Fertility Change*. Population Council, New York
Makinwa-Adebusoye P 1994 Changes in the costs and benefits of children to their parents. In: Locoh T, Hertrich V (eds.) *The Onset of Fertility Transition in Sub Saharan Africa*. International Union for the Scientific Study of Population [IUSSP], Liege, Belgium, pp. 175–192
Mamdani M 1972 *The Myth of Population Control*: *Family*, *Caste*, *and Class in an Indian Village*. Monthly Review Press, New York
Moore T W 1998 Fertility in China 1982–1990: gender equality as a complement to wealth flows theory. *Population Research and Policy Review* **17**: 197–222
Mueller E 1976 The economic value of children in peasant agriculture. In: Ridker R (ed.) *Population Development: A Search for Selective Interventions*. John Hopkins University Press, Baltimore, MD, pp. 98–153
Nag M, White B N F, Peet R C 1978 An anthropological approach to the study of the economic value of children in Java and Nepal. *Current Anthropology* **19**: 293–306
Notestein F 1945 Population: The long view. In: Schultz T (ed.) *Food for the World*. University of Chicago Press, Chicago, pp. 36–57
Nugent J B, Gillaspy R T 1983 Old age pensions and fertility in rural areas of less developed countries: Some evidence from Mexico. *Economic Development and Cultural Change* **31**: 809–29
Schellekens J 1993 Wages, secondary workers, and fertility: A working-class perspective of the fertility transition in England and Wales. *Journal of Family History* **18**: 1–17
Turke P W 1989 Evolution and the Demand for Children. *Population and Development Review* **15**: 61–90
Vlassoff M 1982 Economic utility of children and fertility in rural India. *Population Studies – Journal of Demography* **36**: 45–59
Vlassoff M, Vlassoff C 1980 Old age security and the utility of children in rural India. *Population Studies – Journal of Demography* **34**: 487–99
Willis R J 1982 The direction of intergenerational transfers and demographic transition: The Caldwell hypothesis reexamined. *Population and Development Review* **8**: 207–34

H. S. Kaplan and J. Bock

Fertility Theory: Embodied-capital Theory of Life History Evolution

The goal of this article is to provide an introduction to the embodied-capital theory of life history evolution and the empirical research upon which it is based. The embodied-capital theory generalizes existing life-history theory in biology by treating the processes of growth, development, and maintenance as investments in stocks of somatic or embodied-capital (see Kaplan (1996, 1997) and Kaplan et al. (2000) for more complete treatments). In a physical sense, embodied capital is organized somatic tissue—muscles, digestive organs, brains, etc. In a functional sense, embodied capital includes strength, immune function, coordination, skill, knowledge, and other abilities. Since such stocks tend to depreciate in response to metabolic

activity (e.g., the production of free radicals and mutations associated with mitosis) and aging, allocations to maintenance can also be seen as investments in embodied capital. The general theory is introduced and then applied to human hunting and gathering societies, which characterized most of our species' evolutionary history. This framework is then used to develop models of modern fertility behavior as the result of an interaction between historically novel conditions and the evolved psychological, cultural, and physiological mechanisms governing human parental investment and fertility.

1. *The Embodied-capital Theory of Life History Evolution*

Hominids lived as hunter–gatherers for the vast majority of their evolutionary history (more than 2 million years). Agriculture originated only 10,000 years ago and has been practiced by the majority of the world's population for just two or three millennia, a relatively brief period of time for selection to act. The shift to a fully mercantile, highly urbanized economy with competitive wage–labor markets and a vast array of consumer goods since the mid-nineteenth century is completely novel. Since natural selection is a historical process, humans should be well adapted to the foraging lifestyle, but there is no reason to expect that they will respond adaptively (in the sense of maximizing reproductive fitness) to today's novel circumstances (see Kaplan et al. (1995a, 1995b) for evidence that modern fertility behavior does not maximize fitness). Therefore, an evolutionary explanation of human fertility and parental investment has three requirements: it must be consistent with our general understanding of evolution by natural selection; it must explain the behavior of humans living under traditional hunting and gathering conditions characteristic of our evolutionary history with adaptive models; and it must be able to predict the pattern of modern responses to novel conditions. Our approach to these requirements has been to follow a four-step procedure (Kaplan 1996, Kaplan and Lancaster 2000). The first step is to develop a model of natural selection on fertility and parental investment for organisms in general. The second step is to apply that model to the specific selection pressures acting on human hunter–gatherers. The third step is develop a model of the proximate physiological, psychological and cultural mechanisms produced by selection and how they would interact to determine fertility and parental investment behavior. The fourth step is to generate predictive models of how those mechanisms would respond to conditions currently prevailing in the world. The theory is both ecological and psychological, integrating socioeconomic influences with a theory of mind to identify the processes by which salient features of the social and physical environment relate to people's reproductive and parenting behavior across diverse social, cultural, economic, and ecological contexts.

2. *The Action of Natural Selection on Fertility and Parental Investment*

Natural selection acts on variability in the traits of individual organisms within populations. Traits (and the genes which code for those traits) whose *average* effects on the individuals possessing those traits act to maximize their long-term production of descendents through time increase in frequency relative to other traits. (Technically, selection acts on the 'inclusive fitness' of genes coding for traits. Inclusive fitness includes effects both on the reproductive success of the individual bearing the gene and on other individuals, related by common descent, who also bear the gene. For example, selection on genes affecting alarm calls in response to predators depends both on their effects on the reproductive fitness of the caller (who may risk a greater threat of predation) and on relatives bearing those genes (whose lives may be saved by the call).) Fertility is the most direct contributor to an organism's fitness (i.e., the number of descendents it produces). In fact, all other fitness components, such as mortality, only affect fitness through their effects on fertility (for example, mortality rates affect fitness by affecting the probability of living to the next reproductive event). All else constant, any increase in fertility increases an organism's fitness. However, there are two trade-offs affecting natural selection on fertility.

The first is the trade-off between present and future reproduction. By growing, an organism can increase its energy capture rates in the future and thus increase its future fertility. For this reason, organisms typically have a juvenile phase in which fertility is zero until they reach a size at which some allocation to reproduction increases fitness more than growth. Similarly, among organisms that engage in repeated bouts of reproduction (humans included), some energy during the reproductive phase is diverted away from reproduction and allocated to maintenance so that it can live to reproduce again. Natural selection on age of first reproduction and on the adult reproductive rate is expected to maximize total allocations of energy to reproduction over the life course, and it depends on ecological factors, such as mortality rates (Charnov 1993) and the effects of body size on both energy capture (Charnov 1993) and vulnerability to predation (Werner 1986).

The second trade-off is between quantity and quality of offspring, where quality is a function of parental investment in offspring and reflects its ability to survive and reproduce. In general, natural selection on offspring number and investment per offspring is expected to maximize the number of offspring that survive to reproduce themselves during an organism's lifetime

(Smith and Fretwell 1974). Ecological factors, such as food supply, disease and predation rates, also affect the optimal fertility rates and optimal expenditures on offspring.

Virtually all complex organisms exhibit flexibility in both age at first reproduction and fertility rates. Natural selection has resulted in physiological and psychological mechanisms by which individuals adjust fertility onset and fertility rates in relation to changing environmental conditions. For example, plants often exhibit many thousand-fold variations in fertility rates in relation to yearly and seasonal variations in rainfall and other environmental factors. Extensive research on many bird species has shown that this phenotypic plasticity tracks fitness fairly well (see Godfray et al. (1991) for a review). Birds under variable conditions adjust clutch sizes in ways that tend to maximize the number of surviving young produced during the life course. The pervasiveness of adaptive phenotypic plasticity in fertility among nonhumans suggests that humans are also likely to possess adaptations to adjust fertility onset and fertility rates in relation to changing environmental conditions.

3. The Embodied-capital Theory and its Application to Fertility and Parental Investment Among Traditional Humans

The embodied-capital theory treats the present–future reproduction trade-off in terms of optimal investments in own embodied capital vs. reproduction, and the quantity–quality trade-off in terms of investments in the embodied-capital of offspring. Models based on the theory allow us to address problems that have not been addressed with standard life history models. First, an exclusive focus on physical growth *per se* is an impoverished way of understanding human development. The large human brain, a stock of embodied capital, supports a great deal of learning and knowledge acquisition during both the juvenile and adult periods. Second, parental investment among humans affects not only survival to adulthood, but also the adult socioeconomic status of offspring. This is true not only in modern societies, but also among traditional hunter–gatherers. These models of investment in embodied capital have produced some fundamental results. Of central interest here, the models show that investments in embodied capital affecting adult income or energy capture co-evolve with investments affecting mortality and longevity. The longer the time spent growing and learning prior to reproducing, the more natural selection favors investments in staying alive to reap the benefits of those investments. Similarly, any investments that produce increased energy-capture rates later in life select for additional investments to reach those older ages. The converse is also true. Ecological features or investments that increase the probability of survival to older ages also produce selection for greater investments in income-related embodied capital. These co-evolutionary effects have been particularly important in human life-history evolution.

The human adaptation is broad and flexible, in one sense, and very narrow and specialized, in another. It is broad in the sense that as hunter–gatherers, humans have existed successfully in virtually all of the world's major habitats. This has entailed eating a very wide variety of foods, both plant and animal, both within and among environments. It also has entailed a great deal of flexibility in the contributions of different age and sex classes of individuals. The relative contributions of men and women to food production vary from group to group, as do production profiles by age.

The human adaptation is narrow and specialized in that it is based on extremely high investments in brain tissue and learning. In every environment, human foragers consume the largest, highest-quality, and most difficult-to-acquire foods, using techniques that take years to learn (Kaplan et al. 2000). This adaptation has profound effects on fertility, mortality, and wealth flows. Compared with other primates and mammals, there are at least four distinctive characteristics of human life histories: (a) an exceptionally long lifespan, (b) an extended period of juvenile dependence, resulting in families with multiple dependent children of different ages, (c) support of reproduction by older postreproductive individuals, and (d) male support of reproduction through the provisioning of females and their offspring.

The co-evolution of brains, learning and life histories is evident in Fig. 1. This figure plots net expected cumulative productivity by age (cumulating the probability of being alive at each age times the net productivity at that age) for human hunter–gatherers and wild-living chimpanzees, averaging over all available data from different groups (see Kaplan et al. (2000) for details). The long human training period is evident when the troughs in the human and chimpanzee curves are compared. Humans are not net producers until about age 18–20 years, but are very productive as adults, whereas chimpanzees begin to produce as much as they consume by about age 5 years and are much less productive as adults. In fact, peak production rates for both human males and females do not occur until the mid-thirties for the most difficult-to-acquire resources. Humans also have an expected adult lifespan of about 40 years when they become net producers, whereas chimpanzees have an expected adult lifespan (beginning at about age 13) of 14 years. The dashed line is a hypothetical cross of human production profiles with chimpanzee survival rates. It shows that the human production profile would not be viable with chimpanzee survival rates, because expected lifetime net production would be negative. The human adaptation depends on about 20 years of parental support and a long life span to recoup the investments in learning.

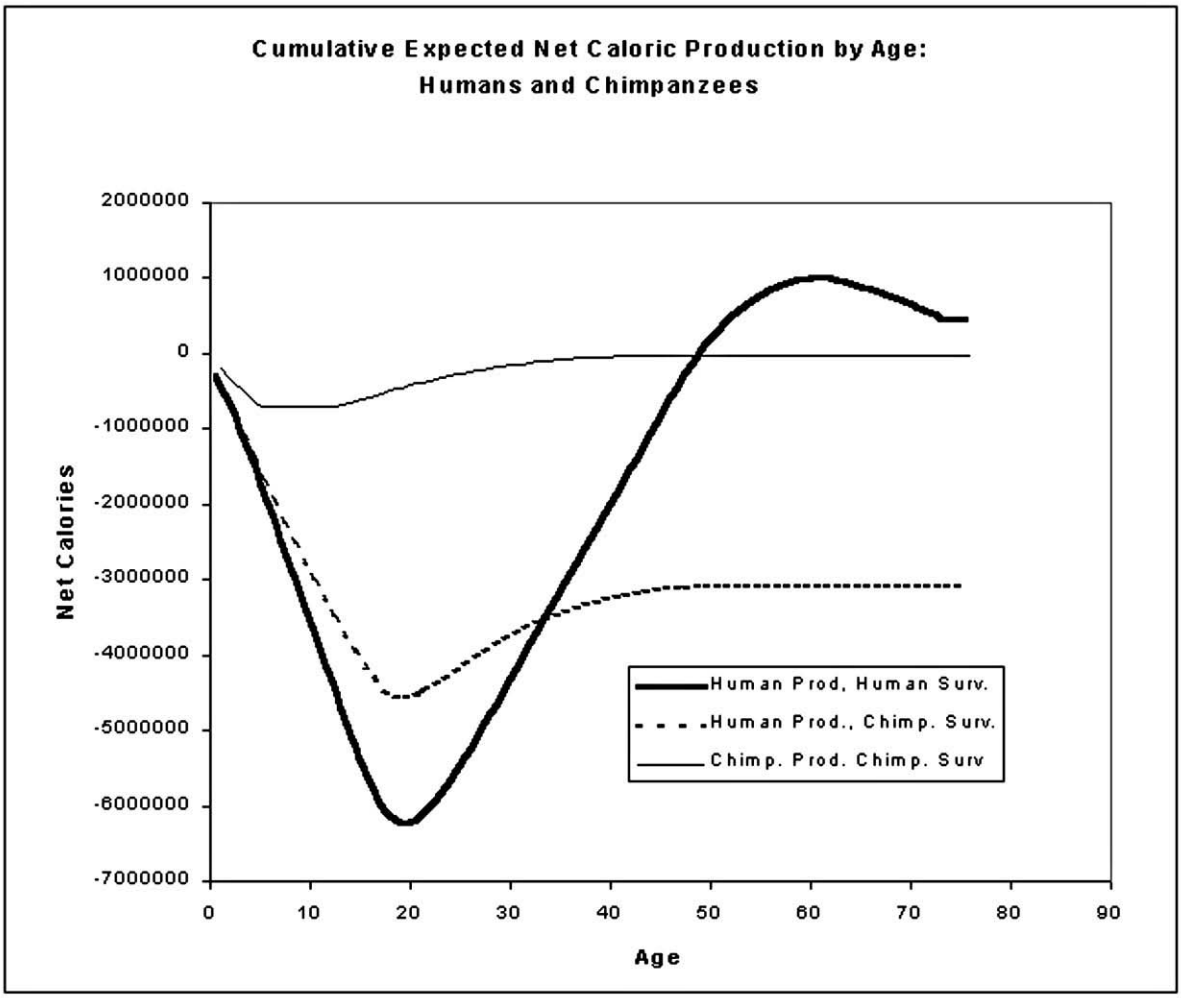

Figure 1
Cumulative expected net caloric production by age: humans and chimpanzees

The first major demographic transition separating our species from apes is a shift to low mortality, an increased lifespan, a long dependent development period, a large commitment to learning and intelligence, and an age profile of production shifted towards older ages with downward wealth flows across generations.

4. *Ecology and the Proximate Mechanisms Governing Fertility and Parental Investment*

The mortality and production profiles discussed above have direct implications for fertility regulation. Since human parents and grandparents provision children, it is likely that natural selection would have produced mechanisms by which fertility could respond to the number of children parents could afford to raise at any given time. Figure 2 presents a conceptual model of the evolved mechanisms of fertility regulation in humans and its responsiveness to physical, biological and social conditions (*socioecology*). The figure depicts two classes of socioecological factors, one affecting survival and the other embodied capital, and three domains of evolved responses, psychology and culture, behavior, and reproductive physiology. The plain text and bold arrows refer to the traditional hunting and gathering lifeway, which we consider first. The italicized text and lighter arrows include additional factors and causal pathways existing under modern conditions.

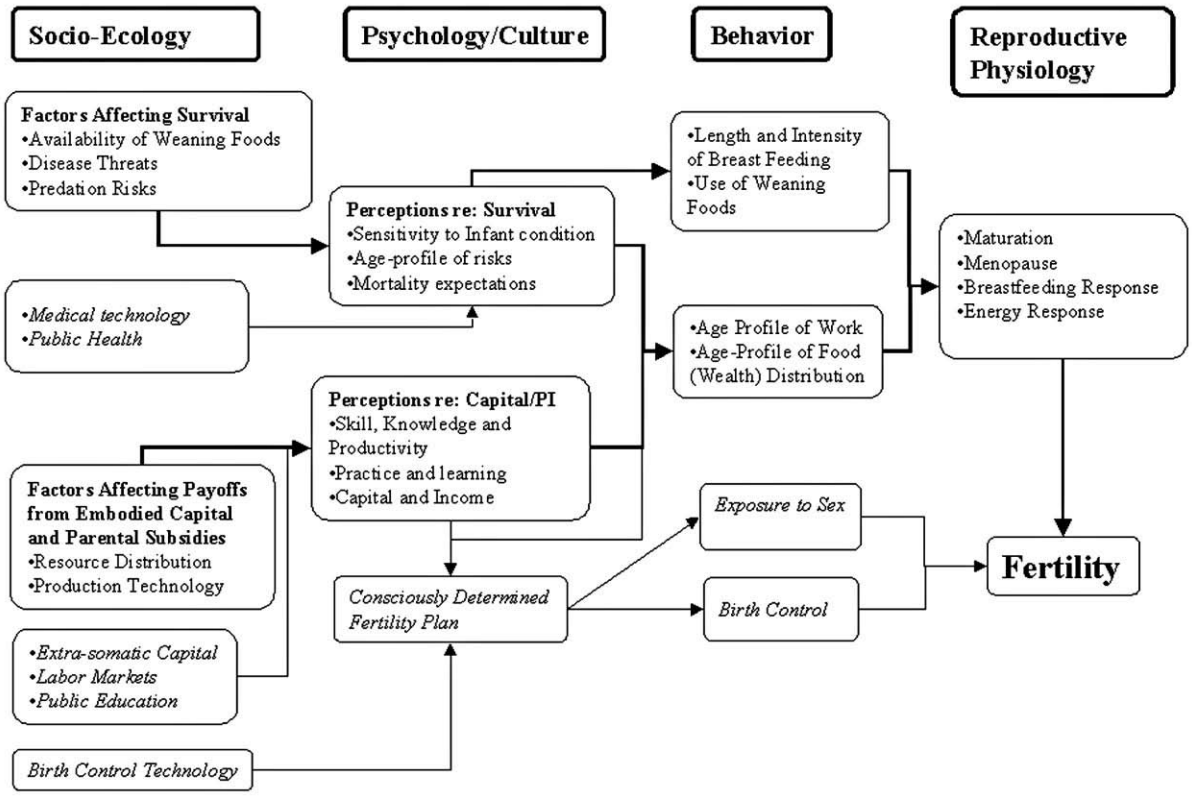

Figure 2
A mode of fertility determinants in traditional and modern societies

4.1 Traditional Foraging Societies

There are two main causal pathways regulating fertility that are both related to the energy constraints in provisioning offspring, one involving breastfeeding and the other involving the age profiles of work and food distribution. Natural selection on the timing and rate of weaning should reflect a balance between quantity effects (favoring earlier weaning) and quality effects on survival and growth of offspring (favoring later weaning). Socioecological conditions, such as the availability of easy-to-digest weaning foods and disease threats, are likely to impact the relationship between age at weaning and survival. Similarly, infant condition is likely to be individually variable and differentially affected by weaning. We expect that natural selection has acted on both infant and maternal psychology to make them responsive to those variable conditions. For example, sickly infants may be more demanding of the breast, and larger and healthier babies more interested in weaning foods. In the human case, maternal perceptions regarding breastfeeding are not only guided by her individual experience and sensitivity to her infant's condition, but also by the cumulative experience of her group, as reflected in cultural conventions and ideas about proper breastfeeding practices. Maternal and infant psychology interact in determining the length and timing of breastfeeding. The evolved reproductive physiology translates this pattern of breastfeeding into differing degrees of fecundity.

The second causal pathway is due to the additional energy constraints involved in provisioning children. The age/sex profile of work and productivity determine the total energy available for consumption. Given some age profile of food distribution and of mortality, the fertility rate must be constrained so that energy consumed does not exceed energy produced. The longer children are provisioned and the greater proportion of their food needs subsidized by parents and grandparents, the lower the reproductive rate that can be supported with a given adult income. Natural selection should therefore act both on psychological abilities to assess the short- and long-term costs and benefits of different activity profiles and on fertility regulation so that energy flows are balanced. One fundamental trade-off with respect to activity profiles

is between present and future production (Kaplan 1996, Bock 2002). If children engage in easy-to-perform foraging activities, such as fruit collection, this will provide immediate caloric benefits and lower the parental subsidy required. However, time allocation to those activities detracts from time spent learning more difficult activities, such as hunting and extractive foraging. During the learning phase, those activities produce very little food, but generate high rates of production in the future. Another fundamental trade-off is between production and mortality risks, since foraging exposes people to predation, accidents, and getting lost. This suggests that parental psychology should be responsive to age-specific mortality rates and how they are affected by alternative activity profiles, and on their short- and long-term consequences for production/productivity.

There is a growing body of evidence supporting the view that people in foraging societies are sensitive to these trade-offs and adjust behavior accordingly. In a seminal series of papers, Blurton Jones and co-workers (Blurton Jones et al. 1994a, 1994b, Hawkes et al. 1995) show that Kung hunter–gatherer parents consciously assess foraging risks to children and that differences between Kung and Hadza children's productivity are due to features of the local ecology related to ease of resource extraction and dangers associated with productive tasks. Bock (1995, 2002), in a study of forager/agropastoralists in Botswana, shows that the age profiles of children's activity budgets reflect immediate skill and strength requirements as well as opportunities to learn. In all foraging societies for which data are available, people are aware of the many years of learning required to become a competent hunter and young men are provisioned while they learn those skills (Kaplan et al. 2000).

These time-allocation decisions affect fertility through maternal energetics. In addition to breastfeeding effects, fecundity responds to women's net energy flows. Seasonal changes on food intake (Bailey et al. 1992), net energy expenditure (Hurtado and Hill 1987) and weight (Bailey et al. 1992, Tracer 1996) all have been shown to predict fecundity as do individual differences among women (Hill and Hurtado 1996).

To summarize, in this model fertility is determined by the interaction of individual psychology and cumulative cultural knowledge, beliefs and practices, behavior, and reproductive physiology. Each of these domains co-evolved by natural selection to respond to ecological variability in mortality regimes and payoffs to parental subsidization of skill acquisition. The connection between psychology and fertility is indirect in that psychology adjusts parental investment (through productive behavior, wealth flows, and breastfeeding) and reproductive physiology translates parental investment decisions into fertility. The key to this system is that maximizing lifetime expected resource production through the optimal allocation of activities and wealth flows will tend also to maximize fitness when all wealth is in the form of food and extra food translates into higher fertility.

4.2 Fertility Under Modernizing Conditions

In Fig. 2, the additional socioecological factors and causal pathways proposed to result in fertility transition are shown with italicized type and thin arrows. Our proposal is that this fertility regime results from the interaction of new socioecological conditions with the psychological processes governing parental investment and wealth flows, evolved in response to the hunting and gathering lifeway. The two critical new factors are the existence of extrasomatic capital and changes in the payoffs to embodied capital, especially in the form of education.

Physical capital, such as land and money, is stored and invested in children in the form of inheritance and parental support for education. Unlike the traditional case where all wealth is embodied, there is no guarantee that a wealth-maximizing psychology will maximize fitness, because much wealth may never be converted into offspring and remain stored in other physical forms.

Changes in the payoffs to embodied capital occur for two reasons. Changes in the technology of production to education-based labor markets have led to very high returns on parental investments in children's education. There are multiple inputs into the production of education-based embodied capital. Parental time inputs prepare children for school and affect their progress in school. Variation in the quality of those inputs, perhaps based on parental education, may predict both levels of parental investment and fertility. Teachers, educational materials, and the quality of student preparation all affect rates of learning (Card and Krueger 1992).

Changing medical technology and public health efforts greatly reduced mortality rates for all age groups. Increased survival rates during the period of parental investment increase the expected costs per child born, favoring further increases in offspring quality (Kaplan 1996). Increased survival during the adult period increases the expected years of return of educational investments, further increasing the incentive to invest in children's education. It will be especially productive to investigate the public and private dynamics involved in this process. Increased private interest in education may have inspired more public investment in education. The quality of public education, in turn, increased the rate of return on private investments in education. Similarly, increased private interest in health (perhaps inspired, in part, by increased investment in education) also stimulated increased public investment in health. This, in turn, increased payoffs to educational investments.

These factors, in combination with parental psychology, result in fertility being regulated by a consciously determined childbearing plan realized

through birth-control technology and/or controlled exposure to sex. In response to the increased payoffs to investments in education and the expected costs of those investments, parents determine the number of children they can afford to raise, given their wealth. The low mortality rates also allow parents to plan reproduction at the outset, because the number of children born accurately predicts the number of children that will be raised to adulthood. Reproductive physiology affects fertility in the present in that it sets the broad limits within which a childbearing plan can be realized, and may actually lower fertility when plans include significant delays to first reproduction (Kaplan et al. 2001).

A great deal of further research is necessary, however, before we can understand why these changes in socioecology have resulted in such low levels of fertility and high levels of parental investment and wealth consumption. It does not appear that people are simply maximizing family wealth because net wealth of families would be maximized by higher fertility than is currently observed. It is also clear that people are not maximizing personal consumption; in that case, they would have no children. One possible hypothesis is that the social dynamics of small groups in hunting and gathering economies resulted in greater fitness for those of higher social standing and selected for a psychology in which relative social position of self and offspring is valued highly. Such a psychology would also emerge if people judged the well-being of themselves and their children based upon the wealth and consumption of others. This psychology may have been fitness maximizing under traditional conditions. If relative, as opposed to absolute, wealth and social standing guide human decisions regarding wealth flows, parental investment, and fertility, it is possible that 'run-away' consumption and investment in children's education result from the interaction of this psychology and modern education-based labor markets and consumption possibilities.

5. *Implications of Embodied-capital Theory for Understanding Current Fertility Transitions*

Embodied-capital theory links several existing theories of demographic transition and, in so doing, provides a new perspective on each one. It rationalizes the shift from natural fertility to parity-specific fertility in terms of a changing quality/quantity trade-off, as a transition from high fertility/low parental investment to low fertility/high parental investment. It also shows why the shift to lower mortality rates can lead to lower fertility, but for reasons that are very different from what traditional demographic-transition theory proposes. According to the latter, lower mortality precedes and drives lower fertility, because fewer children are required to maintain the population equilibrium and cultural ideologies evolve to maintain population equilibrium. According to the present view, mortality rates are but one important factor in the overall rate of return on investments in embodied capital.

This model can further serve to organize research on fertility transitions, with the fundamental expectation that exogenous factors affecting the rate of return on parental investment in the embodied capital of offspring will predict the timing of fertility decline and its distribution within populations. Understanding these dynamics may provide insights into the course of fertility reduction and its variation across countries and among individuals within countries. For example, most studies of predemographic transition societies show that there is a positive association between wealth and fertility (Borgerhoff Mulder 1992). There is some evidence for a negative relationship during the demographic transition. (Kaplan et al. 1995a, Vining 1986) and no relationship after the transition is complete (Kaplan et al. 1995a). It may be that only the wealthy can take advantage of high levels of investment in children's embodied capital when there are few schools and limited public expenditures on health and education. The process of demographic transition may progress through different sectors of the society as more and more people can take advantage of increased rates of return due to changes in the economy and changes in public investments. These same dynamics, in combination with international transfers of educational and health technology, may also explain the national differences in the timing of fertility reduction.

Understanding these dynamics also requires an understanding of how people process the mix of information derived from direct assessment, observation of others, and cultural guidelines. Theoretical work modeling the action of natural selection on information processing suggests that the relative reliance on direct experience and on cultural sources should vary according to environmental conditions (Bock 1999). Cultural learning should be a better guide for behavior when environments are relatively stable than when they are changing (Bock 2001). The rapid change in economics of production, public education and public health may lead people to reject existing cultural models, look for new models, and assess costs and benefits of alternative courses of action more consciously on the basis of direct experience.

International family planning programs present one set of new models, promoting lower fertility and greater investment in children's education. When those models are consistent with individual assessments of costs and benefits, the direction of change is clear. It is less clear, however, how people will respond to family-planning messages in the short and long run if the payoffs to investment in embodied capital are low, either because schools are poor or because employment opportunities for educated people are limited (as appears to be the case in many sub-Saharan African countries).

6. Conclusion

According to embodied-capital theory, fertility and parental investment behavior result from the interaction of exogenous conditions with an evolved response system shaped by natural selection. In addition, a distinctive feature of the learning-intensive nature of human psychology is that information about the environment and guidelines for behavior are obtained through cultural sources as well as through direct personal assessment (Bock 1999, 2001). Given the very limited capacities for cultural transmission displayed by nonhuman primates (Lefebvre 1995), it is evident that the capacity for culture itself evolved by natural selection. We still know very little about the mix of physiological, psychological, and cultural processes underlying decisions regarding wealth flows, educational investments and consumption patterns, or how those processes relate to fertility decisions. Understanding the complex interplay of exogenous conditions, cultural forces, and evolved response systems in the determination of family structure and fertility presents a formidable challenge, but one that promises great rewards.

See also: Fertility: Institutional and Political Approaches; Fertility: Political and Political–Economic Perspectives; Life Course: Sociological Aspects

Bibliography

Bailey R C, Jenike M R, Ellison P T, Bentley G R, Harrigan A M, Peacock N R 1992 The ecology of birth seasonality among agriculturalists in Central Africa. *Journal of Biosocial Science* **24**: 393–412

Blurton Jones N G, Hawkes K, Draper P 1994a Differences between Hadza and Kung children's work: Original affluence or practical reason. In: Burch E S (ed.) *Key Issues in Hunter–Gatherer Research*. Berg, Oxford, UK, pp. 189–215

Blurton Jones N G, Hawkes K, Draper P 1994b Foraging returns of Kung adults and children: Why didn't Kung children forage? *Journal of Anthropological Research* **50**: 217–48

Bock J A 1995 The determinants of variation in children's activities in a Southern African community. PhD thesis, University of New Mexico

Bock J 1999 Evolutionary approaches to population: Implications for research and policy. *Population and Environment* **21**: 193–222

Bock J 2001 Evolutionary theory and the search for a unified theory of fertility. *American Journal of Human Biology* (in press)

Bock J 2002 Learning, life history and productivity: children's lives in the Okavango Delta, Botswana. *Human Nature* (in press)

Borgerhoff Mulder M 1992 Reproductive decisions. In: Smith E A, Winterhalder B (eds.) *Evolutionary Ecology and Human Behavior*. Aldine, Hawthorne, NY, pp. 339–74

Card D, Krueger A B 1992 Does school quality matter: Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy* **100**: 1–40

Charnov E L 1993 *Life History Invariants*. Oxford University Press, Oxford, UK

Godfray H C J, Partridge L, Harvey P H 1991 Clutch size. *Annual Review of Ecology and Systematics* **22**: 409–29

Hawkes K, O'Connell J F, Jones N G B 1995 Hadza children's foraging: Juvenile dependency, social arrangements, and mobility among hunter–gatherers. *Current Anthropology* **36**: 688–700

Hill K, Hurtado A M 1996 *Ache Life History: The Ecology and Demography of a Foraging People*. Aldine, Hawthorne, NY

Hurtado A M, Hill K R 1987 Early dry season subsistence ecology of the Cuiva foragers of Venezuela. *Human Ecology* **15**: 163–87

Kaplan H S 1996 A theory of fertility and parental investment in traditional and modern human societies. *Yearbook of Physical Anthropology* **39**: 91–135

Kaplan H S 1997 The evolution of the human life course. In: Wachter K, Finch C (eds.) *Between Zeus and Salmon: The Biodemography of Aging*. National Academy of Sciences, Washington, DC, pp. 175–211

Kaplan H S, Lancaster J B, Anderson K G, Tucker W T 2001 An evolutionary approach to below replacement fertility. *American Journal of Human Biology* (in press)

Kaplan H S, Lancaster J B 2000 Skills-based competitive labor markets, the demographic transition, and the interaction of fertility and parental human capital in the determination of child outcomes. In: Cronk L, Irons W, Chagnon N (eds.) *Human Behavior and Adaptation: An Anthropological Perspective*. Aldine, New York, pp. 283–322

Kaplan H S, Lancaster J B, Bock J A, Johnson S E 1995a Does observed fertility maximize fitness among New Mexican men? A test of an optimality model and a new theory of parental investment in the embodied capital of offspring. *Human Nature* **6**: 325–60

Kaplan H S, Lancaster J B, Bock J A, Johnson S E 1995b Fertility and fitness among Albuquerque men: A competitive labor market theory. In: Dunbar R I M (eds.) *Human Reproductive Decisions: Biological and Social Perspectives*. St. Martin's Press, New York, pp. 96–136

Kaplan H S, Lancaster J B, Hill K R, Hurtado A M 2000 A theory of human life history evolution: Diet, intelligence, and longevity. *Evolutionary Anthropology* **9**: 156–85

Lefebvre L 1995 Culturally-transmitted feeding-behavior in primates: Evidence for accelerating learning rates. *Primates* **36**: 227–39

Smith C C, Fretwell S D 1974 The optimal balance between size and number of offspring. *American Naturalist* **108**: 499–506

Tracer D P 1996 Lactation, nutrition, and postpartum amenorrhea in Lowland Papua, New Guinea. *Human Biology* **68**: 277–92

Vining D R 1986 Social versus reproductive success: The central theoretical problem of human sociobiology. *Behavioral and Brain Sciences* **9**: 167–216

Werner E E 1986 Amphibian metamorphosis: Growth-rate; predation risk; and the optimal size at transformation. *American Naturalist* **128**: 319–41

H. S. Kaplan and J. Bock

Fertility Transition: China

Between 1970 and 1980, in one decade, China's national fertility level was cut in half. This is his-

torically the most drastic fertility decline among all human populations of significant size, aside from wars, epidemics, and famines. As the largest population in the world, China's fertility transition has also had the greatest impact on the slowing down of the world's population growth. China's fertility decline alone reduced the world's population growth by at least 250 million individuals during the second half of the twentieth century. This entry reviews the process, its historical background, the roles of the government programs and other factors, and the consequences of this transition.

1. The Process of Fertility Transition

At the national level, China's fertility transition began in the early 1970s. In 1970, the total fertility rate was 5.8, close to the highest national level ever recorded for China. In the following decade, a most drastic reduction in fertility occurred. China's total fertility rate declined to 3.6 by 1975, and to 2.3 by 1980. The fertility level fluctuated between 2 and 3 throughout the 1980s. By the early 1990s, it reached and thereafter stayed at the replacement level of 2.1. China, in other words, completed its fertility transition in two decades, from 1970 to 1990.

In China's urban areas and among the more educated subgroups of the population, the fertility transition began much earlier, dating back to the early 1950s (Lavely and Freedman 1990). Since 1957, urban Chinese fertility declined almost consistently, with only one exception, when fertility rebounded in 1963 after a plunge caused by the Great Leap Forward famine. By 1966, total urban fertility dropped to 3.1, and since 1974, it has never been higher than 2. In Shanghai, China's most cosmopolitan city, the total fertility rate dropped to below the replacement level (of 2.1 births per woman) for the first time in 1964, and to below 1 in 1975. By 1970, on the eve of China's first national family planning program, use of contraception and abortion among young women had already risen to 35 and 20 percent, respectively, in cities. Even in the countryside, over 15 percent of women at age 35 had already used a modern contraceptive method, and 7 percent had had at least one induced abortion.

In China's rural areas, where close to 80 percent of the population resided in the 1970s, fertility transition did not begin until much later–around 1970. In fact, during most years between 1950 and 1970, rural fertility showed signs of increase under the collective commune system. Fertility generally stayed at a level of above 6, with the exception of the Great Leap Famine years of 1959 to 1961, when rural fertility was nearly halved. Consistent fertility decline in rural China began after 1970. The total fertility rate dropped to 5 by 1973, to 4 by 1975, and to 3 by 1979. After 1980, the total fertility rate has fluctuated mostly between 2 and 3.

There were also regional variations in the fertility transition in rural China. In provinces with a more advanced economy and a stronger tradition of population control, such as the Lower Yangtze provinces of Jiangsu and Zhejiang, rural fertility declined earlier. In 1973, when the national level was 5, the total fertility rate was only 2.8 and 3.5 in these two provinces. By comparison, rural fertility was much above the national average in economically less developed provinces: it was 7.4 in Guizhou and 6.5 in Gansu (Coale and Chen 1987).

2. Historical Background

China's large population size along with a strong cultural preference for sons and a practice of early and universal marriage among Chinese women have often led many to believe that the Chinese did not attempt to control fertility before China's contemporary fertility transition. Malthus, for instance, stated that 'To account for this population, it will not be necessary to recur to the supposition of Montesquieu, that the climate of China is in any peculiar manner favourable to the production of children, and that the women are more prolific than in any other part of the world.' Among the causes he identified is 'The extraordinary encouragements that have been given to marriage, which have caused the immense produce of the country to be divided into very small shares, and have, consequently, rendered China more populous, in proportion to its means of subsistence, than perhaps any other country in the world' (Malthus 1971, pp. 101, 103).

Contrary to such earlier observations, many Chinese controlled their demographic outcomes in the past by relying on multiple mechanisms. The best-documented and most widely cited examples pertain to infanticide, especially female infanticide (Lee 1981, Lee et al. 1994, Lee and Campbell 1997). Chinese families killed some of their newborns in order to achieve their desired goals for the number and sex composition of their offspring. Poor harvests and economic difficulties also exacerbated the practice of infanticide.

Moreover, studies of Chinese historical population data also reveal that Chinese fertility within marriages was also significantly lower than their pretransition European counterparts (Barclay et al. 1976, Lee and Wang 1999). This lower marital fertility level was achieved by a much longer interval between marriage and first birth, longer interbirth intervals, and an earlier ending of childbearing. The same rationales that motivated infanticide also operated for fertility control. Chinese parents sped up or slowed down reproduction according to the number and sex composition of their existing children, and according to their economic familial circumstances (Wang et al. 1995, Lee and Campbell 1997, Zhao 1997). Fertility control in the past relied mainly on sexual abstinence,

prolonged breast-feeding, and, to a lesser extent, abortion and traditional contraceptives (Hsiung 1994, Lee and Wang 1999).

These behaviors suggest that controlling births is not a novel idea to the Chinese population. The modern fertility transition did not require new ideas for regulating population behavior, but rather a new environment in which fertility control could be seen as necessary and feasible. In contrast to traditional birth control that relied on both prenatal and postnatal methods, the modern fertility transition is mostly prenatal, facilitated by new contraceptive and abortive technologies. Historical traditions of demographic control not only lay the foundation for contemporary fertility transition in mainland China, but they also help to explain the similarities in the pace and patterns of fertility transitions observed in several populations. In spite of having different political and economic systems, mainland China, Taiwan, Hong Kong, and Singapore all exhibited highly similar patterns of fertility transition, partly due to a shared cultural tradition (Coale and Freedman 1993). The Chinese fertility transition process is therefore fundamentally a continuous process, extending a long tradition of demographic control.

3. *Government Programs and Other Factors*

The most prominent feature of China's fertility transition is the role played by government family planning programs. The most drastic fertility decline coincided with the launching of the most forceful government policies and programs of population control. Development of government policy can be traced to the early 1950s. Early concerns about rapid population growth and the need for a government family planning program emerged in the early 1950s, when China's 1953 census confirmed its large population size. Abortion was legalized and contraceptives began to be provided on a limited scale. Unlike later programs, which were motivated by government birth control goals, these early provisions were largely to meet an increasing demand among the population for birth control. These early efforts, however, soon became entangled in ideological debates about whether China needed to control its population growth and were interrupted by a series of political campaigns. As a result, it was not until the 1960s that the government seriously promoted birth control in urban and densely populated rural areas. Overall, government family planning programs at this early stage may have facilitated the early decline, but they were not the main driving force.

Government intervention in reproduction reached a new phase in the early 1970s when the policy of *wan* (later marriage), *xi* (longer birth intervals), and *shao* (fewer births) was formed. Although largely accepted by most people on a voluntary basis, the implementation of the policy became increasingly intensified and regularized. The government utilized its control over the media to promote family planning as a national policy, and mobilized a whole system of vertically controlled work and political organizations to implement the policy. Special government organizations in charge of family planning were also organized and expanded during this period. The official marriage age was stipulated to be 25 and 23 for urban males and females, and 22 and 20 for rural youths. Two children were considered the proper family size for an urban couple and three for most rural couples. Under this program, Chinese ages at marriage rose sharply and fertility declined swiftly. Nationally, the female mean age at first marriage rose from 21 in 1970 to 23 in 1979, and the fertility level was halved (Banister 1987).

In spite of this rapid fertility decline of the 1970s, the Chinese government was not content. Instead, motivated by a keen desire to increase quickly the per capita living standard within a short period of time, it launched an even more ambitious and much more controversial family planning program in 1979. Known as the one-child policy, it required each couple to have only one child, and was an emergency measure to further reduce China's population growth rate. Under the pressure of such an unprecedented and extreme policy, officials and especially family planning cadres resorted to all methods, including physical abuse in some locales, to meet the goals established by various levels of government. Women with two children, sometimes even only one, were required to have an intrauterine contraceptive device (IUD) inserted, and women with three or more births were required to undergo sterilization. Pregnancies not meeting the requirement of family planning had to be aborted. Family planning campaigns were also launched in the 1980s, resulting in sharp increases in the numbers of abortions and sterilization (Hardee-Cleaveland and Banister 1988).

This extreme policy met strong resistance from the Chinese population from the very beginning, especially from the vast population in the countryside. Whereas the earlier policies of family planning were by and large acceptable to Chinese peasants, this new extreme measure required them not only to forgo their traditional perceptions of family and kinship, but also to reduce or even remove their reliance on children for labor and support in old age. Such resistance led to a relaxation in implementation in the Chinese countryside and to changes in the policy. As a result, Chinese couples in cities (accounting for about 20 percent of the population in the late 1970s) obeyed the one-child policy, whereas most rural couples had two, and some more than two, children (Wang 1996, Feeney and Yuan 1994).

The draconian government policy may have masked the roles of many other factors underlying China's fertility transition process. In addition to the historical tradition of demographic regulation, many contem-

porary social and economic changes contributed to China's fertility transition, especially before the 1980s. These changes included a drastic decline in mortality, especially infant mortality, a rise in nonagricultural and nonfamilial employment, a rapid reduction in female illiteracy, and increased provision of social welfare, especially to the urban population (Peng 1989). A changing social and economic context thus produced rapid adaptation to changes in childbearing in the Chinese mainland, just as it did elsewhere in East Asia (Greenhalgh 1988). At both the regional and the individual level, fertility differentials associated with employment type, residence, education, and income existed from as early as the mid-1950s (Lavely and Freedman 1990).

4. Consequences

In its speed and intensity, the Chinese fertility transition has produced several distinctive consequences. The first is a whole generation of single children; and a whole generation of parents with only one child. These families are almost all in urban China, where the one-child policy was effectively implemented. Children who grew up in these families have no siblings, and in the future their children will not have any uncles, aunts, or cousins, all of whom are important members of the traditional Chinese kinship network. Moreover, there will be a severe shortage in familial support in their old age. At least half of all parents of these single children will not be able to co-reside with their married children in their future old age, when emotional and functional support from children may be needed.

Second, rapid fertility decline also means an accelerated process of population aging. This is again especially true for China's urban population, in which the ratio of the elderly aged 60 and above to those at the working age of between 20 and 60 will reach an unprecedentedly high level, the highest in the world, around the year 2040. Nationally, the rate of population aging will also be high, but not as high as in urban areas due to a higher fertility rate in the vast rural areas. Rural to urban labor migration, therefore, can alleviate the shortage of labor in urban areas.

Lastly, China's rapid fertility transition has also led to an acutely unbalanced sex ratio between male and female births. During the 1980s, the sex ratio at birth rose steadily, from 107 boys to 100 girls in 1981 to 115 boys to 100 girls in 1991. The sex ratio among births of higher orders rose even more. Among second births, the recorded sex ratio rose from 107 in 1981 to 116 in 1985, and to 130 in 1993 (Gu and Roy 1995). Much of this increased sex ratio is due to the widespread use of sex-selective abortion. Whereas the increase in the sex ratio may be attributable in part to its government program, the fact that similar increases in sex ratio are also observed in several other East Asian populations suggests this may at least be a temporary by-product of the fertility transition process in these societies. In China, such a lopsided sex ratio will have long-lasting effect on marriage patterns and on population growth in the long run.

See also: East Asian Studies: Gender; Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility Control: Overview; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Infant and Child Mortality in the Less Developed World; Land Rights and Gender

Bibliography

Banister J 1987 *China's Changing Population*. Stanford University Press, Stanford, CA

Barclay G W, Coale A J, Stoto M A, Trussell J 1976 A reassessment of the demography of traditional rural China. *Population Index* **42**: 606–35

Coale A J, Chen S 1987 *Basic Data on Fertility in the Provinces of China, 1942–1982*. East-West Population Institute, Honolulu, HI

Coale A J, Freedman R 1993 Similarities in the fertility transition in China and three other East Asian populations. In: Leete R, Alam I (eds.) *The Revolution in Asian Fertility: Dimensions, Causes, and Implications*. Clarendon Press, Oxford, UK, pp. 208–38

Feeney G, Yuan J 1994 Below replacement fertility in China? A close look at recent evidence. *Population Studies* **48**: 381–94

Greenhalgh S 1988 Fertility as mobility: Sinic transitions. *Population and Development Review* **14**: 629–74

Gu B, Roy K 1995 Sex ratio at birth in China, with reference to other Areas in East Asia: What we know. *Asia–Pacific Population Journal* **10**: 17–42

Hardee-Cleaveland K, Banister J 1988 Fertility policy and implementation in China, 1986–88. *Population and Development Review* **14**: 245–86

Hsiung P 1994 More or less: Cultural and medical factors behind marital fertility in late imperial China. Paper presented at the International Seminar on Abortion, Infanticide and Reproductive Culture in Asia, Kyoto, Japan

Lavely W, Freedman R 1990 The origins of Chinese fertility decline. *Demography* **27**: 89–116

Lee B 1981 Infanticide in China. In: Guisso R, Johannson S (eds.) *Women in China*. Philo Press, Youngstown, NJ, pp. 163–77

Lee J, Campbell C 1997 *Fate and Fortune in Rural China: Social Organization and Population Behavior in Liaoning, 1774–1873*. Cambridge University Press, Cambridge, UK

Lee J, Wang F 1999 *One Quarter of Humanity, Malthusian Mythology and Chinese Realities 1700–2000*. Harvard University Press, Cambridge, MA

Malthus T R 1971 *An Essay on the Principle of Population, Seventh Edition*. Augustus M. Kelley, New York

Peng X 1989 Major determinants of China's fertility transition. *The China Quarterly* **March: 117**: 1–37

Wang F 1996 A decade of the one-child policy: Achievement and implications. In: Goldstein A, Wang F (eds.) *China: The Many Facets of Demographic Change*. Westview, Boulder, CO, pp. 97–120

Wang F, Lee J, Campbell C 1995 Marital fertility control among the late imperial Chinese nobility: Implications for two types of preventive check. *Population Studies* **49**: 383–400
Zhao Z 1997 Deliberate birth control under a high-fertility regime: Reproductive behavior in China before 1970. *Population and Development Review* **23**: 729–68

F. Wang

Fertility Transition: Cultural Explanations

1. Introduction

Demographers have used two approaches to culture in their attempts to understand fertility transitions, both borrowed from anthropology and, to a lesser extent, sociology. The first sees culture as a set of largely inflexible shared beliefs that constrain action, and thus as a barrier to fertility transitions. The second sees culture as shared interpretations that are subject to negotiation by actors; whether these hinder or facilitate fertility transitions is contingent. Demographers have made considerably more use of the former than the latter, although the more flexible and contingent view of culture is evident in some recent work. This article describes these two approaches to culture, giving examples of their uses in the demographic literature and their links to anthropology and sociology. Although the following begins with the introduction of culture in analyses of fertility transitions in Europe in the nineteenth century, the focus is on fertility transitions in third world countries. Much of the literature cited concerns Africa, but the approaches to culture in understanding fertility transitions are similar in other parts of the world.

2. The Princeton European Fertility Project

The interpretations of culture by scholars participating in the Princeton European Fertility Project set a persistent pattern. An influential paper summarized the project's findings with respect to culture: 'cultural setting influenced the onset and spread of fertility decline independently of socioeconomic conditions. Proximate areas with similar socioeconomic conditions entered the transition at different times, whereas areas differing in the level of socioeconomic development but with similar cultures entered the transition at similar times' (Knodel and van de Walle 1986, p. 412; see also Coale 1973). By cultural setting, the project meant region, taken to be a proxy for ethnicity and language (which were considered synonymous); other potential markers of regional cultures, such as gender, were not explored (Watkins 1993). Regional cultures were assumed to be exogenous and inflexible. Culture was also assumed to hinder rational action. Thus, the Princeton Project's model of culture was one that has its roots deep in the European Enlightenment: individuals are assumed to be fundamentally rational, but their vision could be clouded by superstitions and myths.

3. Culture as Barrier

As demographers began studies in the third world, interest in culture increased. This interest had several sources. First, researchers saw third world couples as profoundly different from Western couples—even more different than, say, the differences among regions in Europe. They emphasized aspects of culture that appeared quite exotic from a Western perspective—polygamy, strong son preference, long post-partum abstinence—and were seen to be associated with ethnicity (sometimes a specific ethnicity, sometimes a broader category of 'African' or 'Asian' culture). Second, those promoting family planning programs intended to alter the reproductive patterns of third world couples interpreted their initial failures as due to the 'cultural beliefs' of those who should want fewer children and who should use family planning but who were resistant to reproductive change.

The literature on third world fertility declines is replete with views of culture as barriers to change. These views that are expressed by analysts from third world countries as well as from the West, suggest that they inhabit the same cultural space. The solutions both propose are also similar: education to dispel ignorance and 'myths and rumors,' and cultural change. The latter is either presented as the inevitable accompaniment of modernization, or as something that can be promoted through appropriate policies.

The definition of culture implicit in much of the literature on cultural barriers draws on the view of culture that dominated cultural and social anthropology of the 1920s to the 1940s (Hammel 1990). Because of their far-flung empire, British social anthropologists were particularly influential. For the British social anthropologists who worked in Africa, this was structural functionalism, which emphasized the integrated social wholeness, indigenous morality, and the logical coherence of societies; particular attention was paid to kinship as a central aspect of culture (S. F. Moore 1994, p. 23). Culture was viewed as external to the individual, a set of beliefs shared by a local group, with little attention to differences within a culture or to change. Moreover, it was 'traditional' societies that had culture, whereas 'modern' societies did not.

Despite the views of anthropologist that there was a logical coherence to culture or that reproductive behavior might be the outcome of motivations that were far from reproduction (Kreager 1985), demographers focused on customs or practices thought to be closely related to reproduction, particularly those believed to support high fertility such as polygamy and social control by clan or lineage elders, but also local

practices that reduced fertility such as extended postpartum abstinence (Lesthaeghe 1989, Caldwell and Caldwell 1987). Again, other aspects of local cultures were ignored. The analytic models of demographers sometimes included ethnicity or region, but no other indicators of local culture, and much of the attention was focused on aspects of modernization that were expected by sociologists and economists to diminish the hold of culture.

4. *Theoretical Shifts in Anthropology and Sociology*

Between the 1940s and the 1960s, the interpretation of culture in anthropology—and in sociology—changed substantially. Anthropologists working in third world countries began to focus on the break up of local cultures associated with modernization, especially migration to urban areas (S. F. Moore 1994), and even ethnicity—the stalwart of demographers' views of culture—has come to be seen as fluid and socially constructed (Ekeh 1990). Beginning roughly in the 1970s, there was an even more significant reorientation of theory towards social actors and their strategies for everyday life. A particularly influential figure on this reorientation is the interpretative school associated with Clifford Geertz, who turned attention to meaning in terms of the actors rather than the models of analysts. Geertz proposed attempting to understand the perceived world of actors and their motivations: 'the trick is to figure out what the devil they think they are up to' (Geertz 1983).

Geertz's insights were extended and modified, particularly in the direction of recognizing more within-culture diversity and strategic agency. For example, although the cultural anthropologists of an earlier era, and demographers, saw a close correspondence between beliefs and behavior, in the new interpretations of culture, beliefs and norms were diverse and often contradictory, and which ones were chosen to guide (or perhaps rationalize) behavior was negotiated (Holy and Stuchlik 1983). Pierre Bourdieu (1977) and others emphasized the extent to which individuals know the rules of the game, and use them strategically.

In this approach, culture remains supraindividual, as it was for the structural-functionalists, but it is clearly endogenous, altered, and adapted to specific circumstances. Although local actors may invoke 'our culture since time immemorial' and they themselves may speak of the customs that appear to outside analysts to be exotic, culture is flexible, and concerns rules for daily living. Anthropologists and sociologists began to speak of culture as a 'spice rack' or a 'tool kit' from which various items may be drawn as needed (Greenhalgh 1988, Swidler 1986). Thus, it is argued, analysts need to pay attention to the process by which norms and behavior are evaluated, i.e., what people can expect regarding appropriate behavior, how they assess motivation, behavior, and the social worth of others (H. Moore 1986, p. 40).

In this interpretive view, then, culture changes through communication. An important part of this communication takes place in 'microcultures,' where 'shared meanings [are] directly tied to specific, likewise shared, experiences of people, settings and events' (Hannerz 1992, p. 77, Watkins 2000). Some of it occurs through gossip, a mechanism by which people not only interpret the behavior of others but also discover other people's interpretations; they can thus learn and change cultural rules at a distance. Culture may provide ideal rules to apply to particular circumstances, but the contingencies of life themselves restructure the rules, even change them over time. Thus, in gossip people may justify different conclusions in the same way: for example, in a community in Malawi, one respondent, when asked how many children she wanted, replied 'Only a few, I am too poor to have more,' whereas another said 'Many, I am poor so I need many children to help me.' Much of this communication is local, but it may also be global, following the 'cultural flows' that appear to be increasingly linking societies and are collectively interpreted not only in villages but in global agencies and international meetings and among functionaries of national agencies such as the Ministry of Health (Appadurai 1990, Meyer et al. 1997, Bongaarts and Watkins 1996, Renne 1996).

5. *What Role Do Interpretive Views of Culture Play in Demography?*

The interpretive views of culture have been a basis of vigorous critiques of demographers' interpretations of culture (Greenhalgh 1995, Kerzer 1995, Kertzer and Fricke 1997). Although many demographers have continued to draw on 'fossilized' views of culture drawn from an older era of anthropology (Hammel 1990), the newer views of culture have made some inroads, primarily through anthropologists with an interest in reproduction (many of the younger ones of whom received training in demography through a program established by the Mellon Foundation). Three articles illustrate the sorts of insights that derive from analyses informed by the newer views of culture.

Alaka Basu and Sajeda Amin (2000) emphasize the role of culture in the region of Greater Bengal, but interpret region very differently from the Princeton Project model. Their region is characterized by a common language and a high degree of identity with that language. But the region crosses the border between Bangladesh and West Bengal in India, such that the Bengali language is shared by groups that do not share political or religious boundaries. Nonetheless, they say, cross-border interaction has long been sustained and intense among the elites, as well as between urban and rural areas. More significantly—

and in stark contrast to the common view among demographers that the culture shared through the common language is a barrier to fertility change—they argue that in this specific context culture has been characterized for several centuries by its openness to innovation, evident in receptivity to Western contraception as well as to social reform in other areas of social life.

Tom Fricke's (1997) analysis of fertility control among the Tamang of Nepal shows that the same cultural value can lead to precisely opposite reproductive action when the frame shifts. Between his fieldwork in 1981–2 in the village of Timling and a subsequent visit in 1987, he found that some men had undergone a vasectomy. These men were not characterized by more education than others, nor by more contact with the world outside their village. In listening to their explanations of why they had made such a move, he found that they referred constantly to an underlying cultural value: that of reciprocity and exchange, which 'provides Timling actors with an interpretive model for their own actions and the actions of others' (p. 262). The ethos of reciprocity and exchange is a core cultural value among the Tamang and has associated secondary values, including having many children in order to reproduce broader exchange and social networks and providing their children with a share of inheritance.

Over time, these secondary values had come in conflict. Thus, the men justified their decision to undergo a vasectomy as necessary to protect the inheritance shares of the children they had already borne. The conflict was provoked by changes in the material context and individual experience, but it was the actors' commitment to preserving a core cultural value that led them to vasectomy.

An even more radical step is taken by Caroline Bledsoe, Fatoumatta Banja, and Allan Hill (1998), who contrast the cultural model of high fertility that population scientists use to guide their analyses with the cultural model of reproduction that rural Gambians use to guide their behavior. The former is familiar to demographers: age and number of children lead a woman to change her fertility behavior, those who use contraceptives intend to limit live births, and early contraceptive users are expected to be educated and relatively autonomous from family pressure. Some users of contraception in the Gambia conform to this Western model of reproduction. For others, however, their behavior is inconsistent with this model: they are motivated to use Western contraception because they want to *protect* the woman's reproductive potential so that she may bear more children, either in the current marriage or, should this marriage fail and another be necessary, in a future marriage.

Moreover, reproductive potential is measured not terms of age, as in Western models, but in terms of 'body resource expenditure,' the loss of 'muscles,' strength, and blood that accompany childbearing. Indeed, the authors found that the highest rates of use of Western contraceptives to be among women who had experienced reproductive mishaps—abortion or miscarriage, or child death—and who wanted to 'rest' in order that they might bear more children. Fertility is highly important in this society, but it is less the number of children that matter to the women than marriage.

There are other examples of the way that a more interpretive and activist view of culture has informed demographic analyses of fertility transitions, not only in the work of these authors but in the work of others (e.g. Szreter 1993, Townsend 2000, Kaler 2000).

6. *Conclusions*

Both the older and the newer views of culture can be useful in understanding fertility transitions. Culture may at times be a barrier to fertility change, but it may also facilitate fertility change, and culture itself may change as people alter old reproductive models or adopt new ones. The critics of the use of culture by demographers are fundamentally correct in arguing that the perception that culture is a barrier to rational action—a view that has been adopted by those writing on the spread of HIV/AIDS—is based on a view of third world people as implausibly passive. Indeed, the transformation in reproductive behavior in much of the third world between 1960 and 1990 suggests that new cultural models of reproduction can supercede old ones, or old models can lead to different behavior in new circumstances.

What are the implications of this newer view of culture for demographic analysis? Most demographers will undoubtedly continue to do the kind of statistical analyses they have always done, and are not likely to abandon these for the methods of cultural anthropologists or sociologists of culture. Thus, they will continue to require labels for aspects of culture, labels that can be attached to individuals or groups. The emphasis of the new approaches on communication as a source of cultural flexibility and change suggests that linguistic labels may still be useful, supplemented by attempts to open the 'black box' of culture in order to understand better differences in perceptions and behaviors across (and within) groups. Considering other ways of categorizing individuals and groups may also be fruitful for statistical analysis.

The newer view of culture suggests, however, that what these categories mean may change as the perceptions and motivations of actors change, as well as how that change might occur and the implications for demographic behavior. Most importantly, the recent empirical work guided by the newer understanding of culture suggests that demographic analysis would benefit were our analytic models closer to the models that actors who produce fertility transitions use to guide their behavior.

When demographers have taken culture seriously, it has typically been local cultures in third world countries. A few, however, have suggested new directions. One is the interpretation of what might be called governmental and bureaucratic cultures, both national and global. The second is that demographers examine their own culture, one that has a particular model of reproduction and views culture as a barrier to reproductive change. The newer approaches to culture as explanations for fertility change suggest that the current interpretations of culture in anthropology and sociology may change, and that the culture of demographers may also be flexible.

Bibliography

Appadurai A 1990 Difference and disjunction in the global cultural economy. *Public Culture* **2**: 1–24

Basu A M, Amin S A 2000 Conditioning factors for fertility decline in Bengal: History, language identity and an openness to innovations. *Population and Development Review* **26**: 761–94

Bledsoe C, Banja F, Hill A G 1998 Reproductive mishaps and Western contraception: An African challenge to fertility theory. *Population and Development Review* **24**: 15–57

Bongaarts J, Watkins S C 1996 Social interactions and contemporary fertility transitions. *Population and Development Review* **22**: 639–82

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK

Caldwell J C, Caldwell P 1987 The cultural context of high fertility in sub-Saharan Africa. *Population and Development Review* **13**: 409–37

Coale A J 1973 The demographic transition reconsidered. In: *International Population Conference, Liege, 1973*. International Union for the Scientific Study of Populations, Liege, Vol. 1

Ekeh P P 1990 Social anthropology and two contrasting uses of tribalism in Africa. *Comparative Studies in Society and History* **32**: 660–700

Fricke T 1997 Culture theory and demographic process: Toward a thicker demography. In: Kertzer D, Fricke T (eds.) *Anthropological Demography: Toward a New Synthesis*. University of Chicago Press, Chicago

Geertz C 1983 *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York

Greenhalgh S 1995 Anthropology theorizes reproduction: Integrating practice, political economic, and feminist perspectives. In: Greenhalgh S (ed.) *Situating Fertility*. Cambridge University Press, Cambridge, UK

Greenhalgh S 1988 Fertility as mobility: Sinic transitions. *Population and Development Review* **14**: 629–74

Hammel E A 1990 A theory of culture for demography. *Population and Development Review* **16**: 455–86

Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York

Holy L, Stuchlik M 1983 *Actions, Norms and Representations*. Cambridge University Press, Cambridge, UK

Kaler A 2000 Who has told you to do this thing?: Toward a feminist interpretation of contraceptive diffusion in Rhodesia, 1970–1980. *Signs: Journal of Women in Culture and Society* **25**: 677–708

Kerzer D I 1995 Political-economic and cultural explanations of demographic behavior. In: Greenhalgh S (ed.) *Situating Fertility*. Cambridge University Press, Cambridge, UK

Kerzer D I 1997 The proper role of culture in demographic explanation. In: Jones G W, Douglas R M, Caldwell J C, D'Souza R M (eds.) *The Continuing Demographic Transition*. Clarendon Press, Oxford, UK

Kertzer D I, Fricke T 1997 Toward an anthropological demography. In: Kertzer D, Fricke T (eds.) *Anthropological Demography: Toward a New Synthesis*. University of Chicago Press, Chicago

Knodel J, van de Walle E 1986 Lessons from the past: Policy implications of historical fertility studies. In: Coale A J, Watkins S C (eds.) *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ

Kreager P 1985 Demographic regimes as cultural systems. In: David C, Schofield R (eds.) *The State of Population Theory*. Basil Blackwell, Oxford, UK

Lesthaeghe R 1989 Production and reproduction in sub-Saharan Africa: An overview of organizing principles. In: Lesthaeghe R J (ed.) *Reproduction and Social Organization in sub-Saharan Africa*. University of California Press, Berkeley, CA

Meyer J W, Boli J, Thomas G M, Ramirez F 1997 World society and the nation-state. *American Journal of Sociology* **1**: 144–81

Moore H L 1986 *Space, Text and Gender: An Anthropological Study of the Marakwet of Kenya*. Cambridge University Press, Cambridge, UK

Moore S F 1994 *Anthropology and Africa: Changing Perspectives on a Changing Scene*. University of Virginia Press, Charlottesville, VA and London

Renne E P 1996 Perceptions of population policy, development, and family planning programs in Northern Nigeria. *Studies in Family Planning* **27**: 127–36

Swidler A 1986 Culture in action: Symbols and strategies. *American Sociological Review* **51**: 273–86

Szreter S 1993 The idea of demographic transition and the study of fertility change: A critical intellectual history. *Population and Development Review* **19**: 659–701

Townsend N 1999 Male fertility as a lifetime of relationships: Contextualizing men's biological reproduction in Botswana. In: Bledsoe C, Guyer J, Lerner S (eds.) *Fertility and the Male Life Cycle in the Era of Fertility Decline*. Oxford University Press, New York

Watkins S C 1993 If all we knew about women was what we read in demography, what would we know? *Demography* **30**: 551–77

Watkins S C 2000 Local and foreign models of reproduction in Nyanza Province, Kenya. *Population and Development Review*. **26**: 725–59

S. Cotts Watkins

Fertility Transition: East Asia

East Asia experienced extremely rapid declines in fertility after World War II. In the last decades of the twentieth century, fertility declined further to below replacement levels in some countries of the region, notably in Japan, the Republic of Korea (called 'Korea' hereafter for linguistic simplicity), and the

Republic of China (called 'Taiwan' hereafter for the same reason). Focusing on these three countries, this entry looks at the patterns and factors associated with the postwar fertility transition. After an overview of trends in fertility levels, the entry discusses changes in the two major demographic factors that influence fertility: the timing of women's marriage and fertility within marriage. The article then discusses socioeconomic factors that influence fertility including women's education and employment.

1. Fertility Trends

Although there were differences in timing and magnitude, Japan, Korea, and Taiwan all experienced a dramatic downturn in fertility after World War II (see Fig. 1). In Japan the total fertility rate (TFR) declined from 4.5 children per woman in 1947 to 2.0 in 1957. After this rapid decline, fertility stabilized at about the replacement level (a TFR of 2.0 to 2.1 children per woman) until it started to decline again in the mid-1970s, reaching an all-time low of 1.3 children per woman in 1999. High fertility persisted in Korea until rapid economic development began in the early 1960s. Then the TFR dropped precipitously from 5.9 children per woman in 1963 to 1.6 in 1987. This remarkable decline was followed by a modest upturn in the early 1990s and subsequent stabilization at a level of 1.6 to 1.8 per woman. Taiwan's fertility transition was equally dramatic. Starting in the early 1950s, the TFR declined steadily from 7.4 children per woman in 1951 to 1.9 in 1985. Since the mid-1980s, TFR has stabilized at 1.7 to 1.8 children per woman.

In summary, Japan experienced two fertility transitions, the first a rapid decline from historically high to replacement levels of childbearing, which occurred shortly after World War II, and the second an ongoing decline from replacement to very low levels of fertility. Korea and Taiwan both experienced one continuous fertility transition from very high to below replacement levels.

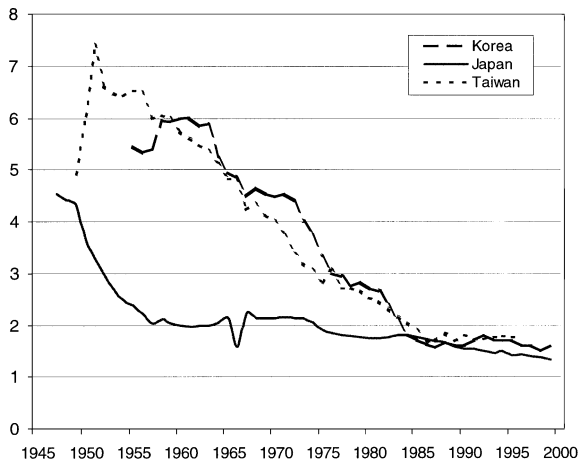

Figure 1
Total fertility rates in Japan, Korea, and Taiwan after World War II

2. Demographic Factors in Fertility Decline

In all three of these countries the fertility transition is related to two demographic factors, namely delayed marriage among women and declining fertility within marriage. Because out-of-wedlock childbearing is minimal, these two factors determine the overall levels of fertility. Demographic analysis indicates that roughly 90 percent of the decline in Japan's TFR between 1950 and 1975 was related to a decline in marital fertility, and the remaining 10 percent was related to delayed marriage among women (Tsuya and Mason 1995). Similarly, approximately 80 percent of the TFR decline in Korea from 1960 to 1990 was due to declining fertility within marriage, and the remaining 20 percent was due to delayed marriage. In Taiwan, around 75 percent of the decline in fertility between 1960 and 1985 was due to a decline in marital fertility, and the remaining 25 percent was due to delayed marriage. By contrast, the fertility decline in Japan *after* 1975 was due solely to delayed marriage among women, while marital fertility showed moderate increases. In a corresponding manner, around 40 percent of the fertility decline from 1980 to 1990 can be attributed to delayed marriage in Korea, and around 85 percent in Taiwan.

2.1 Decline in Marital Fertility

The dramatic fertility declines in these three East Asian countries were caused primarily by a reduction in fertility among married women, who gained control of their family size through contraception and induced abortion. The rate of contraceptive use among married women under the age of 50 increased steadily in Japan from around 20 percent in 1950 to 62 percent in 1979 (Kobayashi 1990). After 1979, the prevalence of contraception stabilized. With the expansion of government family-planning programs, contraceptive practice spread even more rapidly in Korea and Taiwan. In Korea contraceptive use among married women aged 15 to 44 increased from 20 percent in 1966 to 70 percent in 1985 (Palmore and Bulatao 1989). In Taiwan contraceptive use among married women aged 20 to 39 was 23 percent in 1965; among married women aged 15 to 49 it was 78 percent in 1985 (Freedman et al. 1994).

Induced abortion is a traditional means of fertility control long practiced in East Asia. Along with the rapid increase in contraceptive use, rates of induced abortion also increased in the three countries. In Japan the proportion of married women of repro-

ductive age reporting having experienced an abortion increased from 15 percent in 1952 to 41 percent in 1961 (Muramatsu 1990). Since the early 1960s, this proportion has remained in the range of 30–40 percent. In Korea the proportion of married women of reproductive age reporting having experienced an abortion increased from around 10 percent in the mid-1960s to nearly 40 percent in the late 1970s (Cho et al. 1982). Although the rate declined somewhat in the 1980s, abortion remains an important option for terminating unwanted pregnancies. The reported use of abortion among married women of reproductive age also rose steadily in Taiwan from the 1960s to the 1980s (Freedman et al. 1994).

2.2 Delayed Marriage

As we have noted already, being older at the time of marriage has also played an important role in the fertility transitions. The postponement of marriage for women in the three countries is remarkable in light of the East Asian tradition of universal marriage and the cultural expectation that women should marry by their mid-twenties. Between 1975 and 1995 the proportion single rose from 21 percent to 48 percent among women aged 25 to 29, and from 8 percent to 20 percent among women aged 30 to 34 in Japan. Korean women are also marrying later, although almost all marry by their mid-30s. Between 1955 and 1995, the proportion never married increased from 21 percent to 83 percent among women aged 20 to 24, and from 3 percent to 29 percent among women aged 25 to 29. Taiwan has experienced an equally dramatic delay in women's first marriage. From 1956 to 1995, the proportion single rose from 29 percent to 79 percent among women aged 20 to 24, and from 5 percent to 40 percent among women aged 25 to 29. Since the mid-1970s, the percentage single has also risen among women in their early thirties—from 3 percent in 1975 to 16 percent in 1995.

3. Socioeconomic Factors in Fertility Decline

3.1 Women's Educational Attainment

Dramatic socioeconomic changes after World War II brought about important changes in women's status; educational attainment and employment grew, and these changes have also affected fertility. Women's education is a key factor affecting fertility through the timing of marriage, marital fertility, employment, and earnings. In Japan most women were already receiving primary education before World War II; the most notable postwar change has been the steady, rapid increase in the proportion of women with higher education. The proportion of female high-school graduates advancing to higher education increased from 6 percent in 1960 to 48 percent in 1995. The proportion of women aged 25 to 29 who had higher education increased from 10 percent in 1970 to 43 percent in 1990, reaching a par with men in the same age group.

In contrast with Japan, the educational attainment of Korean women was very low in the early postwar years and a sizable proportion of older women had no formal education. This situation has changed dramatically. Only 3 percent of women born in 1936–40 had higher education, compared with 41 percent of women born in 1971–5. The corresponding proportion for men increased from 16 percent to 44 percent; thus the once vast gender gap has almost disappeared. Taiwan has also made impressive improvements in women's education. Educational attainment among Taiwanese women was very low in the early postwar years, with a sizable proportion of older women illiterate. Only 20 percent of women aged 20 to 24 had a high-school education in 1974, increasing to 34 percent in 1995. Since the mid-1970s, about 11 percent of women in this age group have experienced higher education.

3.2 Economic Changes

Japan, Korea, and Taiwan all experienced extremely rapid economic growth after World War II. Since the mid-1950s, Japan's per capita GNP in constant prices has quadrupled every ten to fifteen years, so that Japan is now the second largest economy in the world after the USA. Between the mid-1960s and the mid-1980s, Korea's per capita GNP also quadrupled every 10 to 15 years. Over the same period, Taiwan's per capita GNP doubled every 10 years.

In the course of rapid economic growth, Japan, Korea, and Taiwan experienced rapid urbanization. In Japan, the proportion of the population living in urban areas increased from 37 percent in 1950 to 63 percent in 1960 and 78 percent in 1995. Korea has experienced even more rapid urbanization, with the urban population rising from 21 percent in 1950 to 81 percent in 1995. In Taiwan the proportion of the population living in cities of 50,000 or more increased from 27 percent in 1952 to 73 percent in 1988.

Along with rapid economic growth and urbanization, these countries experienced dramatic changes in the structure of their economies. In Japan the proportion of the workers employed in primary industries fell from 49 percent in 1950 to 6 percent in 1995. In Taiwan it dropped from 56 percent in 1952 to 11 percent in 1995. In Korea the proportion of workers engaged in agriculture fell from 66 percent in 1960 to 12 percent in 1995.

3.3 Women's Employment

With economies shifting rapidly from primary to secondary and tertiary industries, female employment also underwent notable changes. In Japan, although the overall female labor force participation rate has

remained roughly 50 percent over the past 40 years, the type of female employment has changed dramatically. The proportion of employed women who were family workers declined steadily from 43 percent in 1960 to 12 percent in 1995, while the proportion of employed women in paid employment increased from 41 percent to 78 percent. Although Japanese women are still more likely to work before and after rather than during their peak childbearing and child-rearing years, they are increasingly likely also to work between the ages of 25 and 34.

In Korea, female labor force participation increased from one-quarter in 1960 to about one-half in 1995, but women in their peak childbearing and child-rearing years are still much less likely to work than younger or older women. Whereas labor force participation among women aged 20 to 24 increased considerably in the last decades of the twentieth century, participation rates for women aged 25 to 34 have remained stable. This suggests that while employment of young (mostly single) women has increased, paid employment among married women has remained limited. In Taiwan, overall female labor force participation rose from 19 percent to 45 percent between 1956 and 1995, and rates for women at peak childbearing and child-rearing ages increased especially dramatically—from 17 percent to 65 percent among women aged 25 to 29, and from 15 percent to 58 percent among women aged 30 to 34.

See also: East Asian Studies: Economics; East Asian Studies: Gender; East Asian Studies: Society; Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Land Rights and Gender

Bibliography

Cho L, Arnold F, Kwon T H 1982 *The Determinants of Fertility in the Republic of Korea*. National Academy Press, Washington, DC

Freedman R, Chang M C, Sun T H, Weinstein M 1994 The fertility transition in Taiwan. In: Thornton A, Lin H-S (eds.) *Social Change and the Family in Taiwan*. The University of Chicago Press, Chicago, pp. 264–304

Kobayashi K 1990 General remarks on contraception. In: Population Problems Research Council (ed.) *Summary of Twentieth National Survey on Family Planning*. Mainichi Shimbun, Tokyo, pp. 39–52

Muramatsu M 1990 Induced abortion. In: Population Problems Research Council (ed.) *Summary of Twentieth National Survey on Family Planning*. Mainichi Shimbun, Tokyo, pp. 83–97

Palmore J A, Bulatao R A 1989 The contraceptive method mix: An overview. In: Bulatao R, Palmore J A, Ward S (eds.) *Choosing A Contraceptive: Method Choice in Asia and the United States*. Westview, Boulder, CO, pp. 3–24

Tsuya N O, Mason K O 1995 Changing gender roles and below-replacement fertility in Japan. In: Mason K O, Jensen A (eds.) *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK, pp. 137–67

N. O. Tsuya

Fertility Transition: Economic Explanations

1. Approaches of Economics and Demography to the Fertility Transition

With the establishment of modern economic growth, fertility has tended to decline, first across high income countries starting largely in the last quarter of the nineteenth century, and then occurring even more rapidly across most low income countries in the last quarter of the twentieth century (Kuznets 1966, Schultz 1981). Economic models of the allocation of time and wealth of families among production and consumption activities suggests several hypotheses linking changes in the economic environment in which people live to their reproductive demands and fertility (Becker 1960, Mincer 1963). Empirical counterparts for the price, income, and environmental constraints on people's fertility choices have been used to explain aspects of fertility and other related outcomes of the family, strengthening the view that this household demand model is a source of insights into the determinants of the historic fertility transition (Schultz 1974, Becker 1981).

This economic approach to fertility determinants has been criticized by those who believe that fertility is not a rational response to changing economic conditions (Cleland and Wilson 1987). The proposed alternative to the economic approach, called 'ideational change' emphasizes social interactions, but has not yet evolved to a form that is empirically distinguishable. However, research is trying to quantify the role of groups external to the family such as peers and neighborhoods, and to interactions between family members as postulated in cooperative and noncooperative game strategies within families. But because these groups are generally self selected, it is complex to specify the factors that affect the formation of the group and do not also affect the behavior of individuals who enter such groups (Heckman 1979, Manski 1995). There is active debate on how to apply the economic framework and to exploit statistically cross sectional and time series variation in fertility to improve our understanding of the determinants of fertility (e.g., Schultz 1997).

There is a fundamental methodological divide between those who want to describe a 'structural model' that involves interactions between choices made at the household level, including those choices spanning an adult's lifetime, such as between a woman's fertility, marriage, and labor force history, and those who

investigate a 'reduced form model' where only the fixed constraints on household choice are considered as potential explanatory variables—market prices, income sources, technologies, and the environment. It is widely recognized that the associations between lifetime choices represent more (or less) than a simple causal effect of one choice on the other. These covariations between family choices embody possible effects transmitted in both directions, effects of unobserved factors which impact both outcomes, and heterogeneity due to unobserved differences, perhaps in individual preferences (Schultz 1981). Strong identification restrictions are generally required to estimate from nonexperimental data the critical parameters to such 'structural models,' and therefore the empirical conclusions drawn from them are often controversial, because they depend on a host of untested and debatable modeling choices (Manski 1995).

The alternative approach is to estimate fertility as a function of explanatory variables that are not family choice variables or are not likely to be impacted by unobserved factors that could also affect fertility. By delimiting the potential list of explanatory variables, these 'reduced form' relationships represent a summation of the direct effect of the environmentally fixed constraints on the couple's fertility, plus the possible indirect effect of the same constraints on fertility that operate through the modification of other lifetime choice variables, such as the woman's age at marriage, extent of labor force activity outside of her home, contraceptive and sexual behavior, etc. The limitation of reduced form estimates is that they do not necessarily clarify the mechanism by which the constraints impact fertility, for example through the timing of marriage, or the adoption within marriage of various so-called 'natural' or 'unnatural' forms of birth control. The strength of reduced form estimates is that they claim to provide an unbiased approximation for the total effect of the constraint on fertility, which is often a central concern of the social scientist, forecaster, or policy maker.

Some historical demographers assign special importance to the moment when married women begin to restrict their birth rates in response to their prior number of births, or surviving children (Coale and Watkins 1986). Although this may be a well-defined and interesting threshold that predicts the onset of the long term decline in fertility in certain European populations, other populations may adopt other mechanisms to reach their changing fertility goals, such as a change in median age at marriage (i.e., Malthus' preventive check), an increase in celibacy or the proportion who never marry in a birth cohort, an increase in the spacing of births within marriage, as appears to be evident in Africa today (Caldwell et al. 1992, Shapiro and Tabashe 2000), or a change in nonmarital fertility. Even in the region of Northern Europe, standard measures of the emergence of parity-specific fertility control within marriages are not precisely defined by existing data (Guinnane et al. 1994). Thus, economist generally consider comprehensive measures of cohort fertility, and do not focus on one component, such as marital fertility.

2. *Stages of the Demographic Transition*

The 'demographic transition' refers to a sequence of three periods. In the first period, fertility and mortality rates are high and mortality tends to be highly variable, with population growth fluctuating widely about a moderate long run trend of growth or decline. Short run responses in fertility (and mortality) can be attributed to cycles in weather or harvests or possibly to other exogenous variables using time series statistical methods, such as vector-auto-regression.

In the second period, age-specific mortality rates decrease gradually, raising life expectancy at birth from 30–35 years in pre-industrial Europe, to 70–75 today in the high income countries, and from 25–30 years in the low income countries in the 1920s, to levels which today range from about 45–73, excluding countries ravaged by war or the AIDS epidemic. This second period is one of accelerating population growth, which some observers such as Malthus attributed to improved food supplies, and more specifically to nutrition, often initiated by technical progress in agriculture, followed by improved industrial productivity, and advances in transportation and communication (Fogel 1999). A free press capable of publicizing food shortfalls in the twentieth century may have further alleviated excess mortality in the wake of periodic famines. Improvements in private and public health technologies are assigned an important role in reducing mortality, but not much before the start of the twentieth century.

In the third period, the secular decline in fertility begins, after which the number of births per woman falls by more than half, from five or six to about two, in the high income countries, and six to eight to about two or three in most low income countries. When historically high fertility levels have declined by more than 10 percent, they are not expected to rise again on a sustained basis (National Research Council 2000). Medium-term swings in fertility are nonetheless distinguished from time series, occurring in response to business cycles (e.g., the post-World War II baby boom), wars, and economic shocks, as during the transition from centra lly planned to market oriented economies in the 1990s in Eastern Europe and Russia, or recently due to reversals in sub-Saharan Africa.

3. *Economic Determinants of Fertility*

Notestein (1945) described the early dimensions of the demographic transition, and speculated on the economic pressures which he expected to reduce fertility in

the low income world: rising cost of children in urban life, individualistic promotion of child health, and education. Economic models of a lifetime fertility choice emphasize different facets of this decision process, but the trade-off between quantity of children and their quality (i.e., education and health) has become a central feature (Becker 1960, Mincer 1963, Schultz 1981). Because women's time in child-care is a large fraction of the modern opportunity cost of fertility, her allocation of time is treated as jointly determined with fertility. Many cross-sectional and a few time-series studies have proceeded to relate statistically these fixed constraints to fertility outcomes.

A couple's lifetime wealth (endowments of human capital and nonhuman capital) is expected to increase with modern economic growth, and to encourage parents to demand more children, if the relative costs and functions of children did not change (Becker 1960, Schultz 1981). However, for reasons discussed below, the opportunity cost of children increased along with the cost of other goods and activities that complement the rearing of children, whereas the private price of some substitutes for children, such as old-age pensions and healthcare, have been socialized and subsidized by governments. Marked improvements in the technology of birth control were developed in the 1950s, translating into the rapid diffusion of the pill (oral steroid contraceptive) and the IUD (inter-uterine device) to most regions of the world. The monetary and psychic cost and inconvenience (independent of coitus) of avoiding too many births was reduced, and should thereby have reduced the number of births. Family planning programs are not, however, a necessary condition for the occurrence of a fertility transition, because birth rates declined in high income countries without the benefit of new control technologies, or of family planning extension activities designed to foster the adoption of birth control. Finally, improvements in child health technology increased the proportion of births that survived to adulthood. These health developments put downward pressure on birth rates, under two assumptions: parent reproductive goals are defined primarily in terms of their number of surviving children, and these demands for survivors are highly price inelastic (Schultz 1981).

The actual expenditures of a couple on each of their children is not a suitable indicator of the 'market price' of a child on which a couple would base their demand for children. Expenditures per child are also a family choice, which Becker called the demand for 'child quality.' The market-determined prices of the inputs that households combine to rear children are the relevant prices that condition fertility. There is no consensus on which inputs to children are most important and how their unit prices are measured.

If one accepts Becker's (1960) hypothesis that quality and quantity of children are viewed by parents as substitutes, and much of the variation in child quality is embodied in the education parents give their child, then the net price of schooling would affect negatively the demand for quality and affect positively the demand for quantity of children. There is growing evidence in the twentieth century that wage returns to schooling increased, and this can be viewed as a decline in the price of schooling, encouraging parents and societies to invest more resources in the education of children. If parents view schooling per child as a substitute for numbers of children, as is now often empirically confirmed (e.g., Rosenzweig and Wolpin 1980), the increasing wage return to schooling among workers should encourage parents to substitute child schooling for number of children. The increase in wage returns to schooling can be attributed to a technological bias favoring skilled labor, or to a complementarity in production between the schooling of workers and the accumulating stock of technical knowledge available in the global economy.

Becker (1981) also assumed that the income effect on the demand for child quality was a larger positive magnitude than the income effect on the demand for child quantity. This might also have encouraged parents to substitute child quality for child quantity as their incomes increased from one generation to the next. Here was a second possible explanation for the anomalous fall in fertility in a period when *per capita* income was increasing.

4. *The Increasing Value of Parents' Time as an Opportunity Cost of Children*

In the twentieth century, the fraction of national income paid out to labor has risen, despite the increase in capital employed per worker and the decrease in hours worked/year per worker (Kuznets 1966, Fogel 1999). Wage rates have grown relative to *per capita* income, implying that consumption activities that use more time of the consumer have become relatively more expensive over time. Children are thought to be a more 'time-intensive commodity' than the average of all consumption commodities, and thus the opportunity costs for household's to produce and consume a child have risen as parent wages have risen, offsetting some part of the increased demand for children that might have been expected in a period when *per capita* income is rising (Mincer 1963, Becker 1965).

The household production-consumption model drew attention to gender specialization in social production, and to the 'female time-intensity' of children, thereby suggesting the woman's value of time could be a key component in the price of children. In environments where the opportunity cost of female (i.e., mother's or child-care giver's) time is higher, fertility is expected to be lower, other things equal (Schultz 1981). The market wage of women might thus be an 'opportunity price' variable, and it has notably increased in many countries in the twentieth century at a time when fertility has fallen. But just as 'parental

expenditures per child' does not represent a fixed market-determined price of a child, the woman's market wage does not represent a fixed price of a woman's time input to child rearing. First, the wage a woman receives in the labor force is affected by her prior job experience, which accumulates at different rates depending on her preferences among types of work and the allocation of her time. Consequently, a woman's wage is an endogenous choice variable closely linked to her lifetime allocation of time, which in turn is intimately related to her fertility decisions (Mincer 1963).

A second problem with the woman's wage variable is that it is observed only for a self-selected fraction of women who work for a wage in the survey period, and this fraction has tended to rise over time in many populations, and to differ by socioeconomic strata within a population. It is possible that women who work for wages are those who receive an unusually high market wage offer, suggesting that the observed average wage of women overstates the value of time of an average woman. Conversely, women who are unusually productive in nonwage work, and thus are less likely to participate in the wage labor force, may still have an above average market-wage offer, in which case the observed average wage of women would understate the value of time of an average woman. Sample selection bias could distort the wage as a measure of the value of women's time, and thus bias estimates of the woman's wage effect on fertility, probably overstating the causal connection in time series (Heckman 1979, Schultz 1981).

5. *Determinants of the Value of Women's Time and the Cost of Children*

The variation in the opportunity price of children that arises from the variation in women's opportunity value of their time remains a key economic determinant of fertility, and econometric procedures are required to estimate the importance of this connection for the fertility transition. In other words, because the observed wage of women cannot be treated as a fixed determinant of fertility, it cannot be included in a reduced form equation for fertility. How then, does one approximate in a more structural model the effect of employment and productive opportunities for women on the opportunity cost of a woman's time, which in turn affects fertility?

One approach to this problem is to employ the woman's education as a proxy for her wage; schooling tends to be the best variable available for predicting in the cross-section women's wages, just as it is for men's wages. For most women their human capital investment in schooling is completed before childbearing starts, and thus the schooling decision can be viewed as predetermined with respect to fertility. If the schooling decision is made by the child's parents, it may also be argued that the woman's schooling is unaffected by her adult preferences between labor market and fertility choices, and then schooling may be a satisfactory 'instrumental variable' for predicting the adult lifetime wage opportunities for women, which constrain the choice of fertility.

A second approach is to treat the structure of wages in the local labor market as fixed and thus independent of the woman's preferences. This appears to be a reasonable strategy, proceeding from the individual wage to the community averaged wage, and thus eliminating the idiosyncratic individual variation in preferences between working and fertility. Analyses of fertility can then include all women based on the regional wage, and not be restricted to a selected sample of women who are currently reported working for a wage. But another problem could still bias the estimated effect of women's regional wage opportunities on regional or individual fertility. Because an increase in the labor supplied by women in a region will tend to depress the market-clearing wage women receive locally, any omitted regional variables that contribute to greater female labor force participation are expected to also be correlated with lower wages, and hence be correlated with higher fertility. Consequently, these regional omitted labor supply variables would introduce a positive bias in the estimated effect of woman's regional wage on fertility, masking the theoretically expected negative impact of aggregate demand-induced variation in women's wages on fertility.

One estimation strategy to deal with this regional omitted variable problem is to specify a regional variable that shifts only the aggregate regional economy's demand for female labor (relative to the demand for male labor), and thereby affects the local wage available to women (relative to men). Variables of this form which shift the aggregate demand schedule for women's labor are candidates for being an 'instrumental variable' that can account for demand-induced variation in women wages, and can thus be used to obtain an unbiased estimate of the effect of women's wages on fertility. One such demand variable is the regional location of natural resources or the composition of industries, which influence the gender composition of employment opportunities and wages.

Coal mines, for example, employ mostly men, and communities dependent on coal mines for their livelihood have relatively low wages for women. Historically, these coal communities in the UK and the United States reported relatively high fertility levels, other things equal (Haines 1979). In the Northern provinces of Sweden where the price and quantity of timber increased at the end of the nineteenth century due to exports, men's wages increased relative to women's wages, and fertility remained high in these regions well into the twentieth century. By contrast, fertility in the Southern provinces of Sweden declined after the 1860s, as women's wages increased relative to

those of men, in response to higher output prices for exports of livestock, dairy, and food manufactured products, which employed relatively more women (Schultz 1985). Natural resource exports strengthen government finances by providing an inelastic source of taxation, notably in the case of oil exports, with a tendency to sustain customary social institutions that have been associated with relatively low levels of female education, low female labor force participation, and high levels of fertility, holding constant for other factors (Schultz 1994).

Labor economists have noted since the 1960s that in regions where the share of employment in certain industries, for example the services, is relatively large, female labor force participation rates tend to be higher. The importance of part-time, less-skilled, and physically less-demanding jobs may thus be associated with a larger share of women in the labor force, and a rise in the economic status of women relative to men. This form of 'female bias' in the industrial composition of local employment is found to predict increased women's wages relative to men's and lower fertility.

Changes in fertility over time can then be explained by changes in factors external to the household sector which govern the expansion of employment opportunities in industries that hire a disproportionate share of women (or men). The price of the industry's output when it is determined in international markets is one such factor. Terms of trade indexes linked to gender-specific employment and wages opportunities can predict women's wages relative to men's and provide a basis for estimating how demand-driven variation in women's employment opportunities influence fertility. However, these changes in relative prices and employment prospects must be perceived by couples as sufficiently long term in nature that they feel the need to adapt accordingly their reproductive goals.

6. *Evidence from European Fertility Transitions*

The decline in the price of food grains in Europe in the nineteenth century led to such a profound reallocation of labor and allied resources. With the opening of the Erie canal in 1825 and the reduction in ocean transportation costs, exports from the USA, Russia, and other world grain exporters depressed grain markets across Europe. Although restrictions of trade attempted to protect domestic agriculture, European farmers gradually adapted to their changing comparative advantage and left agriculture for urban jobs or emigrated, whereas those who remained in agriculture reallocated their effort from grain to livestock production and to the expansion of food and dairy processing industries. Because grain production traditionally employed more male labor, and livestock and dairy production used more female labor, these changing output prices boosted women's wages relative to men's. This reallocation of resources occurred more rapidly where economies were more open to international trade in agricultural commodities. This transformation of composition of European agriculture or its consequences on the productive roles of women and men is not quantitatively studied in the Princeton European Fertility Project, although these changes could have influenced the diffusion of the fertility transition (Coale and Watkins 1986).

In Sweden, as noted earlier, these agricultural relative price changes were closely associated with the improvement in women's wages relative to men's, and with declines in fertility, by province over time, starting after 1860. The level of real wages for men was related to younger childbearing, but not to a change in total fertility rates. About one-third of the 25 percent decline in total fertility rates for Sweden in the period 1860 to 1910 can be attributed to these externally driven increases of women's wage opportunities, whereas the rest of the decline in Swedish fertility in this half century is associated with urbanization and the increase in child survival, under the assumption that these are exogenous developments (Schultz 1985). A similar demographic pattern unfolded in Denmark, which was also relatively open to international agricultural trade, while its demographic transition was only slightly less rapid than in Sweden.

The German fertility transitions has also been analyzed from 1880 to 1910. When regional persistent or fixed-effects are introduced in the analysis of times series of cross sections, the evidence of strong partial effects of industrialization and urbanization emerge (Richards 1977). Smaller regional units were examined in Prussia from 1875 to 1910, with the same regional fixed-effect specification for fertility (Galloway et al. 1994). However, the female labor force participation rates are used to explain fertility in a reduced form specification, as though the labor force participation of women was a fixed constraint on fertility decisions. Not surprisingly, this choice variable reflecting women's time allocation accounts for the largest fraction of the fertility decline, and if it is jointly determined by women with their fertility, as seems likely, the reported estimates are subject to simultaneous equation bias and are difficult to interpret as evidence for any causal model.

7. *Advancement in Women's Schooling*

Another widespread development in the world, which has contributed to increasing the cost of children relative to most other forms of consumption, is the advance in women's schooling. In the USA, by the start of the twentieth century, the gender gap in schooling was modest, except for college education. But this gap remained substantial in most countries in Southern and Eastern Europe until after the World War I. Moreover, women's schooling attainment began to gain relative to that of men in low income countries only in the last few decades. This progress has been particularly sluggish in South Asia where the

fertility transition has also been most gradual. Given the empirical tendency for wages to increase for women by at least the same percentage as they do for men when they obtain an additional year of schooling, the closure in the gender gap in schooling implies that women's wages are likely to continue to increase in most parts of the low income world for some time, and indeed increase at a more rapid rate than will the wages of men (Schultz 1995). The factors that have triggered these differential trends in the schooling of women and men have not been thoroughly studied, but the expansion of women's schooling closely parallels the decline in fertility across high and low income countries since 1960. Indeed, changes in women's schooling accounts for the largest share of the changes in total fertility rates observed in low income countries since 1960, and in Africa since 1980 (Schultz 1994). Even in an environment where economic progress has been minimal since the 1960s—Kinshasa City in Congo—the increasing rate of female education is the key factor explaining different levels of fertility across the population, and accounting for the decline in fertility in the 1980s and 1990s (Shapiro and Tambashe 2000).

However, educated women could have lower fertility for many reasons, in addition to the greater opportunity cost of their time in child rearing. The social and intellectual advantages that educated women enjoy help them in deciphering, adopting, and using effectively, new and old forms of birth control, and thereby avoiding unwanted births.

Female schooling may also affect fertility through its impact on the risks of child mortality. Censuses and surveys after the World War II have documented for dozens of countries that child mortality rates tend to be 5 to 10 percent lower for mothers who have completed an additional year of schooling (Schultz 1981, 1997). However, the proportionate decline in fertility associated with women's education is larger than the proportionate increase in their child survival rates. Therefore, better educated women have smaller numbers of surviving children, which implies they contribute less to the overall rate of population growth than their less educated peers (Schultz 1994).

Education of women is also closely related to their labor force participation outside of their family, holding constant for their husband's income or education (Schultz 1981). Household models of production and consumption generally assume that family labor supply choices are jointly determined, and women's education is expected to affect not only lifetime fertility, but also the timing of marriage and the matching of the productive characteristics between husband and wife in the marriage 'market.' Within such a framework of family coordinated decision making, it is clearly difficult to assess how one of these jointly determined life-cycle choices made by women affects another of their choices, as stressed earlier. To decompose the variation in fertility into that which is due to the timing of marriage, marital fertility rates, and non-marital fertility rates is a suggestive arithmetic exercise, but one that can misguide statistical analysis because the three components are likely to be stochastically interdependent. Consequently, when analyses of fertility differentials during the demographic transition focus only on marital fertility differentials, correlations can be causally understood only if marriage behavior and fertility demands are behaviorally independent of each other, a working assumption that few social scientists would adopt (Coale and Watkins 1986, Cleland and Wilson 1987).

8. *Conclusions*

Economic explanations for the fertility transition recognize the catalytic contribution of increasing returns to schooling, especially for women, which has encouraged women's education relative to men's, and fostered the rise in women's wages. These attributes of development are not equally evident in all countries. Declines in fertility have therefore proceeded quite unevenly, first across the high income countries, and more recently across the low income countries, but following a pattern of economic developments that has been widely quantified in household surveys and suggestively followed in cross-country comparisons. At an aggregate level, states have simultaneously legislated socialized support systems to provide for the consumption requirements of the elderly, and for the school and health investment requirements of the young, which may have also on balance, reduced the private demands of parents for children.

See also: Children, Value of; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Political and Political–Economic Perspectives; Fertility: Proximate Determinants; Fertility Theory: Caldwell's Theory of Intergenerational Wealth Flows; Fertility Theory: Embodied-capital Theory of Life History Evolution; Fertility Transition: Cultural Explanations; Gender and Feminist Studies in Economics; Land Rights and Gender

Bibliography

Becker G S 1960 An economic analysis of fertility. *Demographic and Economic Change in Developed Countries*. Princeton University Press, Princeton, NJ

Becker G S 1965 A theory of the allocation of time. *Economic Journal* **70**: 493–517

Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Caldwell J C, Orubuloye I O, Caldwell P 1992 Fertility decline in Africa: A new type of transition. *Population and Development Review* **18**(2): 211–42

Cleland J, Wilson C 1987 Demand theories of the fertility transition: An iconoclastic view. *Population Studies* **41**(1): 5–30

Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ
Fogel R W 1999 Catching up with the economy. *American Economic Review* **89**(1): 1–21
Galloway P R, Hammel E A, Lee R D 1994 Fertility decline in Russia 1875 to 1910: Pooled cross section time series. *Population Studies* **48**(1): 135–58
Guinnane T W, Okun B S, Trussell J 1994 What do we know about the timing of the European fertility transition. *Demography* **31**(1): 1–20
Haines M R 1979 *Fertility and Occupation: Population Patterns in Industrialization*. Academic Press, New York
Heckman J J 1979 Sample selection bias as a specification error. *Econometrica*, **47**(1): 153–61
Kuznets S 1966 *Modern Economic Growth: Rate Structure and Spread*. Yale University Press, New Haven, CT
Manski C F 1995 *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA
Mincer J 1963 Opportunity costs and income effects. In: Christ C, Friedman M, Goodman L, Griliches Z, Harberger A, Liviatan N, Mincer J, Mudlak Y, Nerlore M, Patinkin D, Telser L, Theil H (eds.) *Measurement in Economics*. Stanford University Press, Stanford, CA
National Research Council 2000 *Beyond Eight Billion*. National Academy Press, Washington, DC
Notestein F W 1945 Population—the long view. In: Schultz T W (ed.) *Food for the World*. University of Chicago Press, Chicago
Richards T 1977 Fertility decline in Germany: An econometric appraisal. *Population Studies* **31**(3): 537–53
Rosenzweig M R, Wolpin K I 1980 Testing the quantity-quality fertility model. *Econometrica* **48**: 227–40
Schultz T P 1981 *Economics of Population*. Addison-Wesley, Reading, MA
Schultz T P 1985 Changing world prices, women's wages and the fertility transition: Sweden, 1860–1910. *Journal of Political Economy* **93**(6): 1126–54
Schultz T P 1994 Human capital, family planning and their effects on population control. *American Economic Review* **83**: 255–60
Schultz T P 1995 *Investment in Women's Human Capital*. University of Chicago Press, Chicago
Schultz T P 1997 The demand for children in low-income countries. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. North Holland, Amsterdam, The Netherlands, Vol. 1A, Chap. 8
Schultz T W (ed.) 1974 *Economics of the Family*. University of Chicago Press, Chicago
Shapiro D, Tombashe B O 2000 *Women's Education, Employment and Fertility: Kinshasa in Transition*. Book manuscript, Pennsylvania State University, PA

T. P. Schultz

Fertility Transition: Latin America and the Caribbean

In the second half of the twentieth century Latin America experienced a remarkable demographic transformation, with fertility declining from a regional average of more than six children per woman in 1950 to less than three children per woman by the year 2000. In this article we describe the nature of the decline, assess the role played by the proximate determinants of fertility, and discuss the ultimate causes and consequences of this profound transformation.

1. Fertility Levels and Trends

CELADE (1999) has published estimates and projections of the total fertility rate (TFR) for 20 countries in Latin America and the Caribbean for the period 1950–2000. These estimates update Chackiel and Schkolnik's (1996) excellent overview of the transition. The estimates are based on population censuses and vital statistics analyzed using direct and indirect techniques, and are generally consistent with alternative estimates based on survey data, such as the World Fertility Survey (WFS) and the Demographic and Health Surveys (DHS). Figure 1 plots the TFRs for a subset of 10 countries that represent the types of transitions observed in the region.

Back in 1950 most countries in Latin America and the Caribbean had TFRs in excess of six children per woman. Notable exceptions (the last four solid lines in Fig. 1) were Argentina and Uruguay, with rates around three; and Cuba and Chile, which hovered around four and five, respectively, but dropped below three in the mid- to late 1970s. Cuba then went on to have the lowest fertility in the region, 1.6 children per woman by the year 2000. A second group of 10 countries experienced declines from much higher fertility to cross the four-child level in the 1970s and 1980s. These include Panama, Brazil, El Salvador, and the Dominican Republic (the dashed lines in Fig. 1), as well as Colombia, Costa Rica, Ecuador, Mexico, Peru, and Venezuela (not shown, but their trajectories are bracketed at both ends by the countries shown). By the end of the twentieth century only six countries still had TFRs above three, but they all had experienced substantial fertility declines for 20 years or longer. This third group includes Guatemala and Paraguay (the top two solid lines in Fig. 1, representing the highest and lowest TFRs in this group in 2000), as well as Bolivia, Haiti, Honduras, and Nicaragua (not shown).

While there are substantial differences among countries, age-specific fertility has fallen more rapidly at older than at younger ages, as one would expect of a transition driven primarily by a desire to limit fertility rather than space births (Chackiel and Schkolnik 1996). This notion is confirmed by birth-interval analyses that show a substantial decline in parity progression ratios (the proportion of women proceeding from one parity to the next), particularly at high parities, with little if any change in average birth intervals (Juárez and Llera 1996).

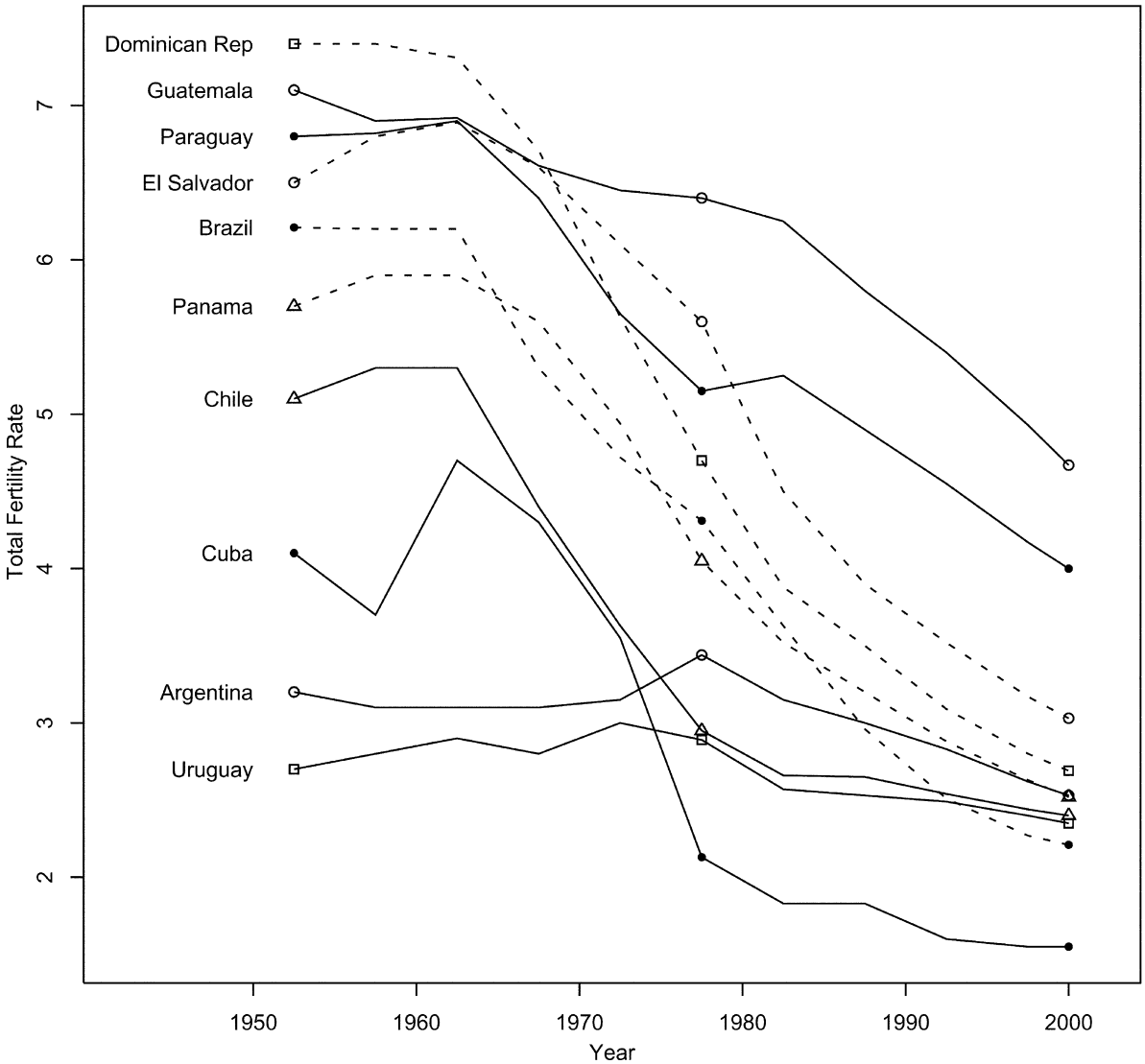

Figure 1
Total fertility rates, Latin America and the Caribbean, 1950–2000

The national averages conceal substantial variation across social groups in each country. Not surprisingly, fertility in Latin America has been substantially lower in urban than in rural areas; it decreases with increasing education of women, and with increasing status in occupation of the husband as one moves from agricultural workers through manual laborers to white-collar strata. The fertility decline started in the higher strata, but it soon spread to other groups, eventually reaching rural areas and women with no education married to agricultural workers. As a result, differentials increased at the start of the transition but became narrower as fertility levels declined (Rodríguez 1996, Chackiel and Schkolnik 1996).

2. *The Proximate Determinants*

Following Davis and Blake's classic framework, changes in fertility can be traced to changes in intermediate variables affecting exposure to the risk of pregnancy, to conception, and to gestation. Most analyses focus on three key factors: marriage, contraception, and postpartum infecundity. A fourth important factor, abortion, is often omitted for lack of

reliable data. Moreno and Singh (1996), in a thorough analysis of these variables using WFS and DHS data, apply Bongaarts' model as well as an alternative model by Moreno to show that the bulk of the fertility decline in Latin America can be attributed to increases in contraceptive use, with changes in nuptiality and breastfeeding making only modest contributions.

Rosero-Bixby (1996) has examined the role of nuptiality in more detail. Using census and survey data he finds a slight increase in the time spent in union between 1950 and 1960 (consistent with isolated fertility increases in this period) followed by a flat trend. The proportion of consensual unions exceeds 50 percent in several countries, particularly in Central America and the Caribbean, and has increased over time. More remarkably, he finds no systematic trend in age at marriage. These results suggest that nuptiality trends could hardly contribute to the reduction in fertility in the region after 1960, and a more formal analysis shows only modest effects in selected countries.

Weinberger (1996) analyzed the role of contraception. She estimates a prevalence rate for the region during 1990 of 56 percent of couples where the wife is in the reproductive ages, with country-specific rates ranging from 7 percent in Haiti to around 70 percent in Costa Rica. The earliest national-level surveys measuring contraceptive practice go back to the 1970s and document a rapid and continuing spread of contraceptive use that is consistent with the observed decline in fertility. A distinctive characteristic of contraceptive use in the region is the high prevalence of female sterilization (20 percent of couples) and oral contraception (16 percent), which together account for about two-thirds of all contraceptive use. Other methods used by at least 5 percent of couples are periodic abstinence or rhythm, and intra-uterine devices. In contrast, male sterilization is rare (around 1 percent of users). In general, changes in the method mix over time have tended towards increased effectiveness, a trend largely due to increased reliance on female sterilization.

The role of induced abortion is hard to determine because of a dearth of reliable data. Frejka and Atkin (1996) argue convincingly that induced abortion is used widely in the region, but can provide only indirect evidence in most cases. Exceptions include Cuba, where abortion has been available on request since 1960, and Chile, where abortion is illegal but reasonable estimates are available from health statistics and surveys; both countries had total abortion rates in excess of two in the 1960s, coinciding with periods of steep fertility decline. (Interestingly, the early surveys in Chile helped motivate the medical profession to provide family planning services as an alternative to abortion.) Singh and Sedgh (1997) use indirect estimates based on hospitalizations for the treatment of complications related to induced abortions to show that, in addition to helping determine fertility levels, abortion has played a significant role in the fertility decline in Brazil (where fertility in the 1990s would have been about 13 percent higher had abortion not increased from its level in the late 1970s), and more moderate roles in Colombia and Mexico.

Finally, postpartum infecundity appears to be the least important of the proximate determinants in terms of its role in the fertility transition in the region. The duration of breastfeeding in Latin America and the Caribbean is relatively short (with the exception of Haiti and Peru) and has not changed much over time.

3. *Ultimate Causes and Consequences of the Decline*

Traditional explanations of fertility decline have focused on socioeconomic changes that reduce the demand for children, but more recently there has been a resurgence of interest in theories that emphasize ideational changes and the diffusion of new ideas and forms of behavior. Several features of the Latin American transition suggest that the process represents more an innovation than adaptation to changing economic conditions.

Starting in the 1960s, Latin America underwent a process of modernization that resulted in remarkable improvements in educational levels, a shift away from agricultural work, greater participation of women in the labor force, and a doubling of per capita income between 1955 and 1980. Impressive as these achievements are, the changing socioeconomic composition of the population explains only a small part of the overall decline in fertility, which was driven largely by changes within social strata (Rodríguez 1996). Moreover, the economic situation changed drastically in the 1980s, as the region experienced a profound crisis (the 'lost decade') that eroded standards of living and reduced upward mobility; yet had little if any perceptible effect on the pace of the decline.

More direct evidence on the demand for children comes from data on fertility preferences. Starting in the 1960s, a number of surveys found preferences for small families in high-fertility regimes where women were not using contraception—the so-called 'KAP gap'—that evolved later into more sophisticated measures of unmet need for contraception and decompositions of fertility into its wanted and unwanted components (Westoff and Moreno 1996.) In countries like Costa Rica the transition seems to have occurred, at least initially, without a corresponding reduction in ideal family size (Rosero-Bixby 1999). Once the decline is established, however, preferences may well be reduced further. In Peru, for example, the TFR in 1977 to 1978 was 5.3 and its wanted component was only three. Ten years later the TFR had declined to four, but by then wanted fertility had further declined to two children per woman (Bongaarts and Lightbourne

1996). Still, back in the 1960s demand was already low and there was ample room for fertility reduction by better implementing existing preferences.

The gap persisted because the means to reduce fertility were not widely available at the time, but the situation was soon to change. Family planning programs were introduced in Latin America starting in the 1960s, largely as the result of private initiatives organized with help from the International Planned Parenthood Federation (IPPF) and, in some cases, the United States Agency for International Development (USAID). A leading private program is Profamilia, which was established in Colombia in 1965 and was a pioneer in the development of distribution networks and the use of radio to promote family planning. In the 1970s a number of governments started to provide services as well: a leader in this area was Mexico, where government clinics opened in 1974. Today most countries have a mixture of private- and public-sector programs, the latter often integrated with maternal and child health services. While estimates of the demographic impact of these programs vary, there is little doubt that they facilitated the transition by increasing the acceptability and availability of contraception (Mundigo 1996, Bongaarts et al 1990).

Once modern contraceptive methods became generally available, all that remained was for women to adopt them. It is here that diffusion or social contagion enters the picture, with friends and relatives playing a key role in terms of information flow and demonstration effects, at least in the early stages of the transition. A multivariate analysis of the spacing and limiting components of fertility shows that once the transition starts in a social stratum, it follows a broad but well-defined path that is consistent with a basic diffusion process (Rodríguez 1996). In a detailed case study of Costa Rica, quantitative spatial analysis combined with qualitative focus group research provides strong evidence of diffusion effects (Rosero-Bixby 1999). There is also a growing body of evidence that the mass media, particularly radio and television, can play a significant role. Perhaps the best documented example comes from northeast Brazil, where the fertility decline tracks the spread of television and the extremely popular *telenovelas*, nightly soap operas that often deal with life-styles, family values, sexuality, and reproductive behavior, and which can have strong if unintended influences on attitudes and values (Faria and Potter 1999).

Assessment of the consequences of the fertility transition is difficult because these changes occurred in times of economic crisis and turmoil (Potter 1996). There is, however, clear evidence that the decline resulted in lower infant mortality rates (Taucher 1996). Perhaps of more enduring consequence, the status of women has changed as a result of improved education and reduced fertility. In the words of a *New York Times* editorial on the subject, 'For all the machismo in Latin America, women there have gained substantially more rights and better treatment than women in the rest of the developing world.'

See also: Economic Development and Women; Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Land Rights and Gender; Latin American Studies: Education; Latin American Studies: Gender; Latin American Studies: Religion; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender

Bibliography

Bongaarts J, Lightbourne R 1996 Wanted fertility in Latin America: Trends and differentials in seven countries. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Bongaarts J, Mauldin W P, Phillips J F 1990 The demographic impact of family planning programs. *Studies in Family Planning* **21**: 299–310

Centro Latinoamericano de Demografía (CELADE) 1999 Latin America: Urban and rural population projections 1970–2025. *Boletín Demográfico/Demographic Bulletin*, No. 63. CELADE, Santiago, Chile

Chackiel J, Schkolnik S 1996 Latin America: Overview of the fertility transition, 1950–1990. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Faria V E, Potter J E 1999 Television, telenovelas and fertility change in northeast Brazil. In: Leete R (ed.) *Dynamics of Values in Fertility Change*. Oxford University Press, Oxford, UK

Frejka T, Atkin L C 1996 The role of induced abortion in the fertility transition of Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Juárez F, Llera S 1996 The process of family formation during the fertility transition. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Leete R (ed.) 1999 *Dynamics of Values in Fertility Change*. Oxford University Press, Oxford, UK

Moreno L, Singh S 1996 Fertility decline and changes in the proximate determinants in the Latin America and Caribbean regions. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Mundigo A I 1996 The role of family planning programmes in the fertility transition of Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

New York Times 1999 *Latin America's Birth Surprise*. Editorial, June 13, New York Times, New York

Potter J E 1996 The social consequences of rapid fertility decline during a period of economic crisis. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Rodríguez G 1996 The spacing and limiting components of the fertility transition in Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Rosero-Bixby L 1996 Nuptiality trends and fertility transition in Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Rosero-Bixby L 1999 Interaction, diffusion, and fertility transition in Costa Rica: Quantitative and qualitative evidence. In: Leete R (ed.) *Dynamics of Values in Fertility Change*. Oxford University Press, Oxford, UK

Singh S, Sedgh G 1997 The relationship of abortion to trends in contraception and fertility in Brazil, Colombia and Mexico. *International Family Planning Perspectives* **23**: 4–14

Taucher E 1996 The impact of fertility decline on levels of infant mortality. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Weinberger M B 1996 Changes in the mix of contraceptive methods during fertility decline: Latin America and the Caribbean. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

Westoff C F, Moreno L 1996 Reproductive intentions and fertility in Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

G. Rodríguez

Fertility Transition: Middle East and North Africa

1. Introduction

Countries in the Middle East and North Africa (MENA) region provide a complex and rich experience on fertility transition. During the second half of the twentieth century, fertility trends took more divergent courses than ever before. Several forces influenced fertility in conflicting directions. On the one side the forces of modernization, development, modern family planning technology and women's education, among others, have worked toward rapid fertility decline. On the other side, political, religious, cultural and institutional forces have supported early marriage and high procreation. The balance of these two sets of conflicting forces has, over time, influenced the onset and magnitude of fertility transition in the region.

2. Fertility Transition During the Period 1950–70

Throughout this period fertility was high in a group of 22 predominantly Arab and Muslim MENA countries (henceforth these countries will be referred to by the acronym AMENA). The total fertility rates (TFR) were five or more live births per woman (see Table 1 and Fig. 1) and the crude birth rates (CBR) were above 40 births per 1,000 persons. Some of these countries experienced a slight fertility increase. For instance, fertility increased from 5.74 to 6.36 in Lebanon and from 6.93 to 7.25 in Tunisia. Similar increases occurred in Egypt, Bahrain, Jordan and Libya. An odd country among this group is the former Soviet Union of Azerbaijan where fertility was relatively lower (TFR = 5.49; see Fig. 1).

A second group of four MENA countries experienced low and declining fertility ($3 \leqslant \text{TFR} \leqslant 4.9$; see Table 1 and Fig. 1). Between 1950 and 1970, fertility declined, albeit slowly, from 4.49 to 3.45 and from 3.0 to 2.61 in the Soviet republics of Armenia and Georgia, respectively. Cyprus experienced a similar slow decline from 3.71 to 2.79 during the same period. Also, fertility declined from 4.16 to 3.79 in Israel, which represents a special case because of its large immigrant population from Europe, the Arab States and Africa (Okun 1997, Goldscheider 1992). Henceforth, this group of countries will be referred to by the acronym (NAMENA), as a separate group in the MENA region.

Table 1
Distribution of MENA countries by categories of total fertility rates (TFR) 1950–2000

TFR	1950–55	1960–65	1970–75	1980–85	1990–95	1995–2000
$\leqslant 7.0$	11	15	9	6	2	2
6.0–6.9	9	6	9	4	2	0
5.0–5.9	2	1	2	5	3	3
4.0–4.9	2	1	2	4	6	3
3.0–3.9	2	2	2	3	7	7
2.2–2.9	0	1	2	3	4	7
$\geqslant 2.1$	0	0	0	0	2	4
Total	26	26	26	26	26	26

Source: UN Population Division. World Population 1950–2050 The 1998 Revision (Medium Variant)

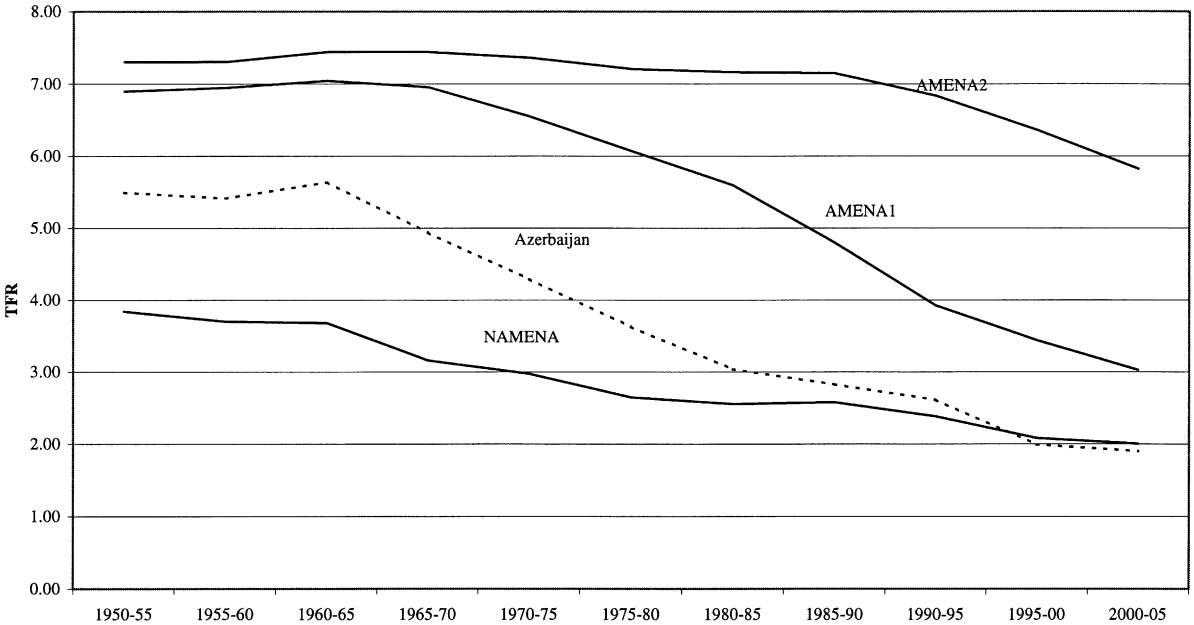

Source of Data: UN Population Division World Population 1950-2050 (The 1998 Revision Medium Variant)
AMENA2=Yemen,Oman,Saudi Arabia, Iraq & Gaza Strip. AMENA1=Remaining Arab and Muslim Countries
NAMENA = Israel, Cyprus, Georgia & Armenia

Figure 1
Average total fertility rates in the MENA countries

Fertility trends in the above mentioned two groups of MENA countries (AMENA and NAMENA) reflect considerable differences in reproductive behavior during this period. High fertility in the AMENA countries was due to political, religious, cultural, and institutional factors, and to the supremacy of patriarchal and extended family norms (Yousif et al. 1996, Courbage 1999). Moreover, public policies were against family planning interventions to regulate fertility (Obermeyer 1992). By contrast, culture and public policies in the NAMENA countries were in support of small family size, women's education and the use of modern contraceptive methods (Okun 1997).

3. Fertility Transition During the Period 1970–2000

The onset and magnitude of the fertility transition during this period varied considerably (see Table 1 and Fig. 1). The slow fertility decline mentioned before in Cyprus, Georgia, and Armenia continued unabated and reached replacement level (TFR = 2.1) during this period. However, the traditional high fertility in the AMENA countries gradually declined. Fertility continued to decline in Azerbaijan and it reached below replacement level by the year 2000 (see Fig. 1). It also declined to 2.2–2.9 live births per woman in six countries (Lebanon, Tunisia, Turkey, Iran, Kuwait, and Bahrain) and to 3.0–4.9 in ten others (Jordan, Sudan, Syria, Algeria, Libya, Qatar, United Arab Emirates, Egypt, Morocco, and Western Sahara; see Fig. 1). Fertility remained high (TFR $\geqslant$ 5.0) in only five AMENA countries (Yemen, Saudi Arabia, Oman, Iraq, and the Gaza Strip).

Tunisia, a pioneer country in the AMENA group, provides a good example of a sharp fertility decline from 6.21 in 1970 to 2.55 in 2000—a drop of 3.66 children in 30 years. The total fertility rate declined in Egypt from 5.6 in 1976 to 4.1 in 1991and to 3.9 in 1992; a drop of 30 percent in 17 years. Kuwait provides a third example with a TFR decline from 5.9 in 1980 to 2.89 in 2000; a drop of three children in 20 years. Fertility declined in the Maghrib countries too, by 2.8 children in Algeria (1977–92) and 2.9 children in Morocco (1971–92).

By contrast, fertility remained high in Iraq, Saudi Arabia, Oman, and the Gaza Strip where the TFR ranged between 5.3 to 8 children per woman (see Fig. 1) and the CBR stayed above 40 throughout the second half of the twentieth century. A slight fertility decline finally occurred in a couple of these countries during the 1990s. Fertility declined from 6.8 to 5.8 in Saudi Arabia, and from 6.2 to 5.2 in Iraq. By contrast, fertility in Yemen and the Gaza Strip remained at a

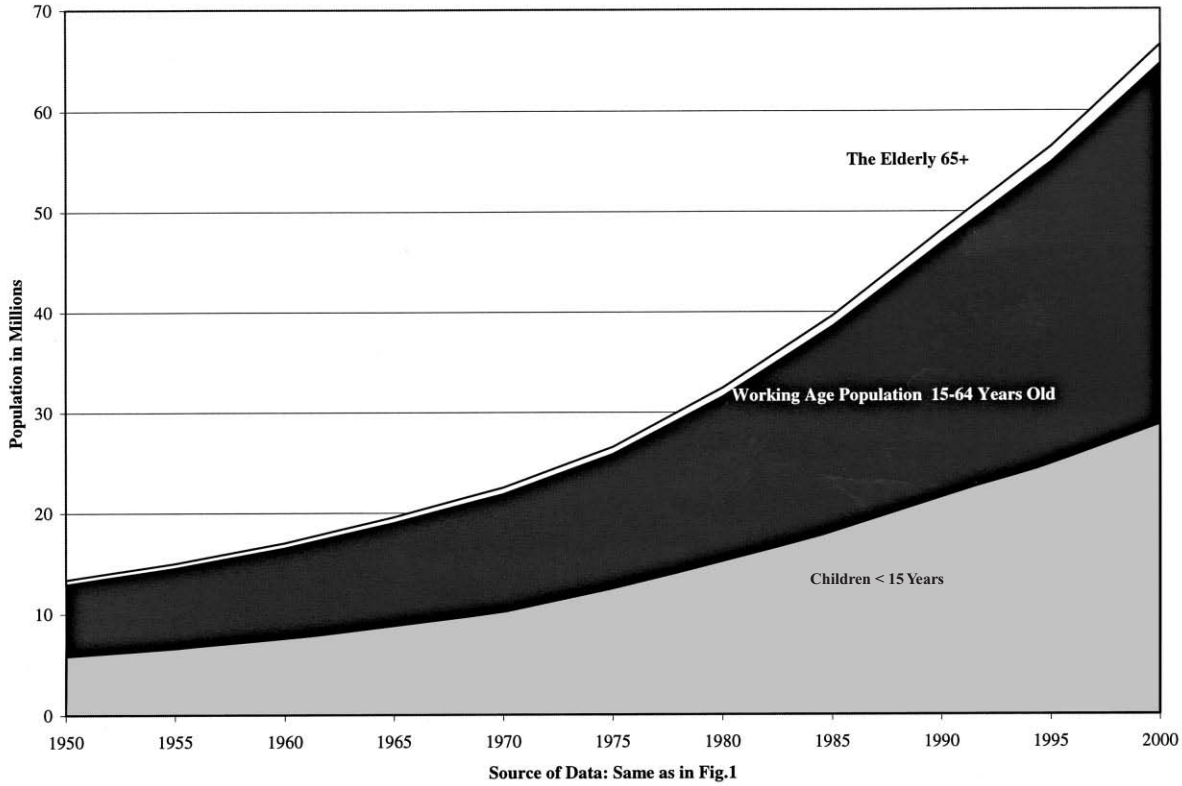

Figure 2
Population distribution by age in the AMENA 2 countries

high level of seven or more children per woman throughout the 1950–2000 period.

4. *Urban vs. Rural Fertility*

Place of residence (urban–rural) has a pervasive influence on fertility transition in the AMENA countries. It encapsulates many differences in opportunities available to women. Urban women have better access to the media and information technology, education, health, modern jobs, and family planning methods than their rural counterparts, most of whom are still under the grip of traditional beliefs and values (Yousif et al. 1996). The fertility decline is considerably higher and its onset earlier in urban than in rural areas. For instance the total fertility rate had dropped to below replacement level in Beirut by 1999, while in North Lebanon it averaged four live births (Kulczycki and Saxena 1999). In some AMENA countries a fertility decline occurred in the urban areas only. In Sudan, for example, fertility has declined to 3.7 in Khartoum and has remained high at above six children per woman in rural areas (Yousif et al. 1996). The TFR declined to 3.6 in urban areas and it remained high at a level of 6.3 children in the rural areas of Egypt.

5. *Women's Education*

Women's educational attainment has contributed significantly to the fertility transition in the MENA countries. However, countries have differed from each other with regard to the level of education (or years of schooling) at which fertility declined. In several countries fertility declined in response to a few-years increase in women's education (Yousif et al. 1996). The rapid spread of women's education during the period 1970–2000 reduced the influence of traditional values, attitudes and norms and strengthened women's social status within households and communities. It facilitated their employment in modern activities outside the home as well. Consequently, the mean age at marriage for girls has increased to 25 or more years in several AMENA countries. Moreover, generations of educated girls have acquired modern aspirations and family desires.

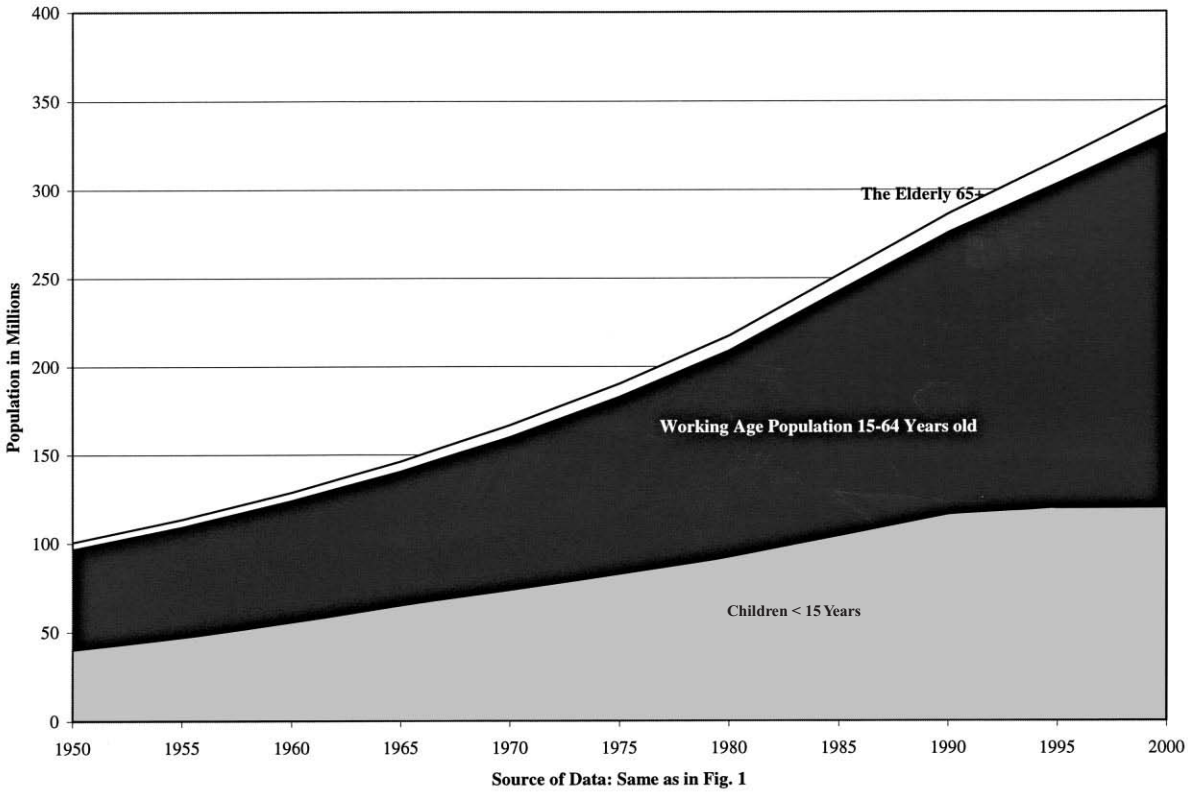

Figure 3
Population distribution by age in the AMENA 1 countries

6. *Impact of the Fertility Transition on Age Composition*

The impact of past high fertility, which often is associated with a high population momentum, on the age composition of the population is universal in the AMENA countries. This is reflected in increasing absolute and relative numbers of children below the age of 15 and in a precipitous increase in the working age population (15 to 64 years of age). The high population momentum is so pervasive that the impact of the recent fertility decline in these countries has not yet taken its full swing (see Figs. 2 and 3).

However, the impact is clear in the AMENA countries that have experienced fertility decline earlier than the others (see Fig. 3). In Azerbaijan and Tunisia, for example, the percentage of children below age 15 is declining, and it reached 30.3 percent and 28.6 percent, respectively, in 2000. In Lebanon and Kuwait this proportion is also declining (to 33.7 percent and 32.7 percent by 2000). In countries like Yemen, the Gaza Strip, Saudi Arabia, Iraq, Jordan, and Syria, children still make up more than 40 percent of the total population. However, in all AMENA countries, the percentage of the population at working age is increasing, which implies a high demand for social services, particularly family planning, health, and education, as well as jobs.

By contrast, continuous fertility decline in the NAMENA has led to a relatively stable absolute and relative number of children below age 15 (see Fig. 4). The working age population is growing rapidly and its proportion of the total population exceeded 62 percent in all of these countries. The relative and absolute numbers of elderly have increased too and it has reached 12.6 percent in Georgia and 9.8 percent in Israel. This process will continue in the future, with profound implications for health and old age security schemes in these countries. These changes reflect an older age structure compared with the AMENA countries.

7. *Impacts of Public Policies on the Fertility Transition*

Public policies in the MENA countries provided the ideological and operational context for the fertility

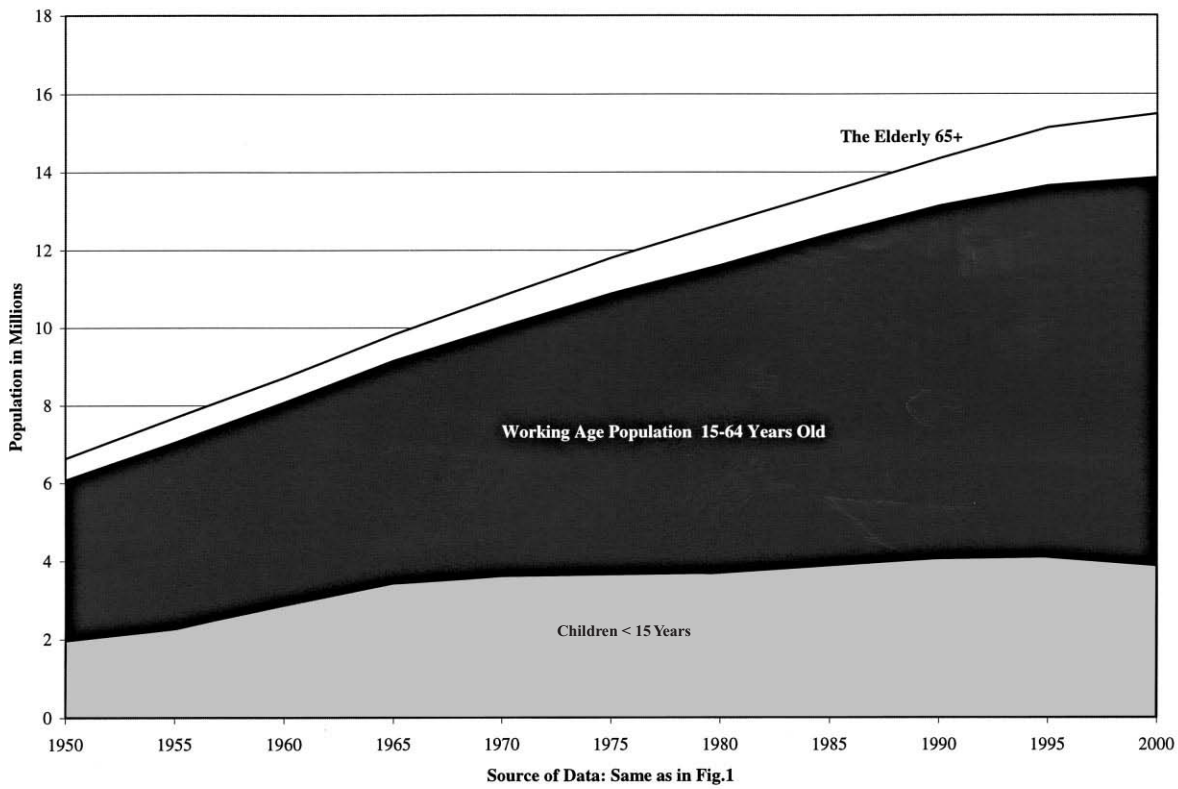

Figure 4
Population distribution by age in the NAMENA countries

transition. During the period 1950–70, population issues in the AMENA countries were seen as long-term problems of secondary importance, neither pressing nor urgent (Yousif et al. 1996). The reproductive role of women was regarded more important than their education and participation in modern employment. This public policy stand was evident during the International Population Conference in Bucharest in 1974. Delegates from the AMENA countries invariably argued in support of development, as opposed to family planning, as the best method for resolving population problems. Ten years later, by 1984, the political stand of these governments had changed, and in 1994 several AMENA countries reached an advanced stage on issues of population policy and development. Also, government support to population policies and programs gained importance, albeit at different times, in the AMENA countries.

Public policy in some countries changed direction several times. In Egypt, for instance, the government was against the use of modern contraceptive methods during the 1960s and 1970s. Starting in 1980, the government has changed its stand to support family planning to regulate fertility and address population issues in the country. Its main challenge is in the rural areas, particularly in Upper Egypt, where fertility is highest and the potential for decline greatest.

Sudan and Iran provide other examples of countries with contradictory or vacillating public policies on population issues. The Sudan family planning program, which was progressing well during the 1970s and 1980s, has recently come to a standstill due to lack of government support. In Iran, the government's opposition to family planning after the Islamic revolution in 1979 has now changed completely to the opposite (Courbage 1999). In the Gulf States public policies on population issues and family planning are in their infancy.

8. *Conclusions*

Fertility transition mirrors divergences in culture and development of the MENA countries. Family values and traditions, social institutions, norms, ethnic divisions, religious beliefs, and all the various forces of

modernization and development are reflected in the fertility trends of the populations in this region. The downward trend in fertility in many countries and in urban areas is irreversible in view of ongoing social development, improving educational attainment, and increasing use of contraceptive methods. The potential for further fertility decline in the future is high, particularly in rural areas and in the AMENA countries where fertility is currently high.

See also: Fertility and Culture: Anthropological Insights; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Islam and Gender; Islam: Middle East; Middle East and North Africa: Sociocultural Aspects; Near Middle East/North African Studies: Economics; Near Middle East/North African Studies: Gender; Near Middle East/North African Studies: Religion; Near Middle East/North African Studies: Society and History

Bibliography

Courbage Y 1999 Economic and political issues of fertility transition in the Arab world—answers and open questions. *Population and Environment* **20**: 353–80

Goldscheider C 1992 *The Embeddedness of the Arab–Jewish conflict in the State of Israel: Demographic and Sociological Perspectives*. PSTC Reprint Series, No. 92–02. Brown University, Population Studies and Training Center Providence, Rhode Island

Kulczycki A, Saxena P C 1999 New evidence on fertility transition through wartime in Lebanon. *Genus* **LV**: 131–52

Obermeyer C M 1992 Islam, women, and politics: The demography of Arab countries. *Population and Development Review* **18**: 33–60, 205–7

Okun B S 1997 Innovation and adaptation in fertility transition: Jewish immigrants to Israel from Muslim North Africa and the Middle East. *Population Studies* **51**: 317–35

Omran A R, Roudi F 1993 The Middle East population puzzle. *Population Bulletin* **48**

Yousif H M, Goujon A, Lutz W 1996 *Future Population and Education Trends in the Countries of North Africa, RR-96-11*. International Institute for Applied Systems Analysis, Laxenburg, Austria

H. M. Yousif

Fertility Transition: South Asia

South Asia is defined here as the countries of former British India (India, Pakistan, and Bangladesh) plus Nepal and Sri Lanka. Their combined population was estimated to be 1.34 billion in the year 2000, equivalent to 22 percent of the global total. Fertility trends in this region are thus a crucially important determinant of future global population size. In this article the course of fertility over the past 50 years will be described, attempts to explain changes in reproductive behavior will be reviewed, and prospects for future fertility decline assessed.

1. Data Sources

Though only Sri Lanka possesses a vital registration system with high coverage of births, there is nevertheless a relative abundance of demographic data sources in South Asia, including decennial censuses and large-scale single-round demographic surveys. In terms of regular demographic monitoring, the outstanding achievement of the region is India's Sample Registration System (SRS). The SRS has yielded annual demographic estimates for most states in India since the early 1970s.

The collection of accurate demographic data in South Asia is extremely difficult and all national datasets are flawed to some extent. This has generated considerable controversy over the timing and magnitude of reproductive change in South Asia. Genuine uncertainty still exists over the details, particularly in Pakistan, where the quality of data is especially poor, but the broad contours of demographic change are well established.

2. Fertility Trends

The historic level of fertility in South Asia was probably similar to that of Europe and China, namely an average of 5.5 to 6.0 live births per woman, with fluctuations in the range of 4.5 to 6.5. Natality rose during the 1950s and 1960s in some parts of the region, a reflection of declines in widowhood and changes in breastfeeding patterns and sexual behavior within marriage. Fertility peaked at around seven births per woman in Bangladesh and Pakistan in the 1970s but remained about one birth lower in the other three countries.

Sri Lanka was the forerunner of fertility transition (see Fig. 1). The onset of decline dates back to the 1950s, and began largely in response to a rising age at marriage (Fernando 1975). The first appreciable fall in national fertility in India occurred in the 1970s, in Pakistan and Bangladesh in the 1980s, and in Nepal in the early 1990s. Comparing the highest post-1950 fertility level with estimated levels between 1995 and 2000, childbearing has fallen most steeply in Sri Lanka (63 percent), moderately in India and Bangladesh (about 50 percent), and least in Pakistan and Nepal (about 28 percent). Undoubtedly the most remarkable development has been the steep fall in fertility in Bangladesh. Between 1980 and 1995 fertility fell from 6.5 to 3.2 births per woman. This is among the most

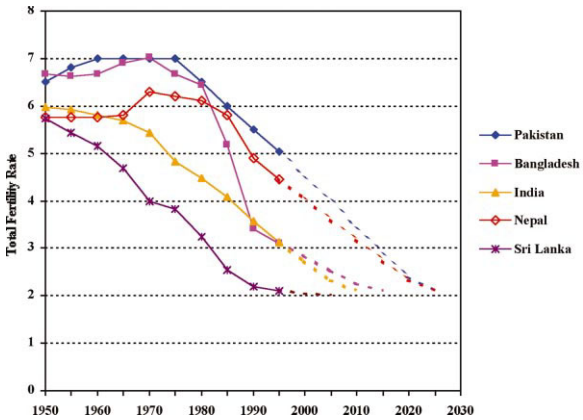

Figure 1
Past and projected trends in total fertility rates in South Asian countries
Source: United Nations 1999 *World Population Prospects: The 1998 Revision*. Population Division, New York

abrupt and pronounced changes in reproduction ever recorded in a developing nation.

These national trends mask geographical, ethnic, and socioeconomic variations in the timing of the onset and speed of reproductive change. In Bangladesh, fertility is lowest in the Rajshahi Division (in the north), characterized by relative poverty and landlessness, and highest in the Chittagong Division, generally considered to be the most conservative part of the country. In Nepal, fertility transition is more advanced in the Hill Zone than in the Terai (adjacent to India) or in the high-altitude Mountain Zone. The highest fertility in Pakistan is recorded in the least developed province, Baluchistan.

The diversity of reproductive trends in India is much greater (see Fig. 2). In the four northern states of Bihar, Madhya Pradesh, Rajastan, and Uttar Pradesh, which comprise nearly 40 percent of the total population, demographic modernization has lagged far behind the rest of the country. Women's literacy, age at marriage, and life expectancy are all low; fertility has declined moderately but the level of childbearing was still above four births per woman in the mid-

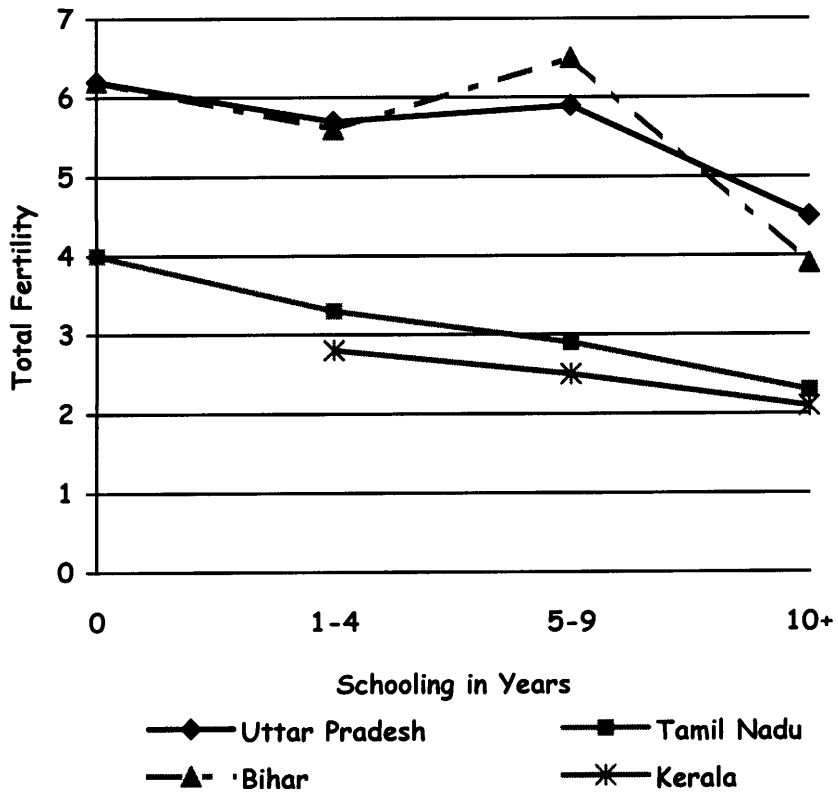

Figure 2
Rural total fertility rates by educational level of women, selected states of India, 1984
Note: Kerala estimate for 0 years of schooling not shown, because too few cases exist
Source: Office of the Registar-General 1990 *Fertility Differentials in India*. Vital Statistics Division, New Delhi

1990s. By contrast, in the southern states of Kerala and Tamil Nadu fertility transition began early, probably in the late 1950s (Zachariah and Rajan 1997). According to the SRS the level of childbearing had already fallen to about four births per woman by the early 1970s. Kerala reached replacement level fertility (i.e., the level required to maintain stability of population size in the long term) of a little above two births by 1988, eventually stabilizing at 1.8 births per woman. Replacement level was reached in Tamil Nadu in 1993.

3. *Determinants of Transition*

3.1 *Direct Causes*

South Asian transitions to lower fertility have been driven by two direct causes: a rise in marriage ages, thus reducing teenage childbearing; and increased family size limitation by contraception and abortion. There is little evidence of postponement of live births following marriage, or of changes in birth spacing. Contraception and abortion have been used to limit family size but not to delay first births, nor for birth spacing. Reproduction is increasingly compressed within the age span of 20 to 30 years.

Throughout most of South Asia, early and universal marriage for women—with a wide gap of five to eight years in the ages of brides and grooms—has been the historic norm. Apart from Sri Lanka and parts of India (e.g., Kerala), the singulate mean age at marriage for women at the start of the twentieth century was about 13 years. (The singulate mean age is calculated from census or survey data on proportions single by current age: for details see Hajnal 1953 and also Ní Bhrolcháin.) Marriage has remained universal but women's age at marriage has steadily increased. By the 1980s, the singulate mean ages were: Bangladesh 16.6, India 18.3, Nepal 18.9, Pakistan 20.7, Sri Lanka 24.4.

Marriage postponement probably accounts for 20 to 30 percent of the decline in total fertility in Pakistan and Sri Lanka but rather less in the other three countries. However, delayed marriage is not consciously motivated by a desire for smaller family sizes. The fundamental cause is a change in the marriage market. In a rapidly growing population, each age group is larger than the preceding older one and, in societies where females traditionally marry older men, this creates a surplus of marriageable women. It thus becomes more difficult and expensive for parents to find suitable bridegrooms for their daughters. Increasing secondary schooling for girls accentuates this upward pressure on marriage ages.

The major direct cause of the South Asia fertility transition is increased use of contraception and abortion. Abortion is believed to be commonly practiced in all five countries but its precise contribution to fertility decline is impossible to gauge with any confidence because in Nepal, Sri Lanka, and Pakistan it is illegal, and in Bangladesh and India most terminations are unrecorded.

In contrast, trends in contraceptive practice have been monitored reliably. A very pronounced link exists between reported contraception and fertility. In the 1990s, contraceptive prevalence (i.e., the percentage of married women in the reproductive age span who report current use of any contraceptive method) was highest in Sri Lanka (66 percent), and lowest in Pakistan (24 percent). Highly effective modern methods are dominant. In India and Nepal there is a heavy reliance on sterilization, mainly of women. In 1992 and 1993, sterilization accounted for 76 percent of contraceptive protection in India, and the median age of sterilization of women was 27 years. In contrast, the great surge in contraceptive practice in Bangladesh in the late 1980s and 1990s was dominated by increased use of hormonal methods.

3.2 *Socioeconomic, Cultural, and Policy Influences*

During the past 50 years, a fertility transition has spread across South Asia despite the variation in material, social, and cultural circumstances. Presumably some common underlying factor, or interrelated cluster of factors, accounts for the onset of transition over the same relatively short timescale. Such factors might include preceding mortality declines that have increased numbers of surviving children, or monetization of economies and growing aspirations for educating children and for material goods, all of which increase the cost of childbearing. The variations in the timing and speed of transition demand more context-specific explanations.

Population scientists disagree about the causes of change (see *Fertility Transition: Cultural Explanations* and *Fertility Transition: Economic Explanations*), and rather few widely-accepted generalizations can be made. One of the few is that the fertility transition in the region is not primarily a response to rising living standards. Bangladesh remains one of the poorest countries in the world and the southern states of India have income levels that are below the national average. Indeed, some commentators have proposed that reproductive change has been driven by poverty rather than prosperity, though the micro-level evidence does not support this view: poorer families have more children than richer families.

Progressive social policies, resulting in universal schooling, high life expectancy, relative equality of incomes, and state welfare provision, offer a compelling explanation for the early transitions in Sri Lanka and Kerala. But this thesis applies less obviously to Tamil Nadu, where literacy and life expectancy are not exceptionally high, and is implausible in the case of Bangladesh.

Gender inequality has been a major explanatory theme (see *Reproductive Rights in Developing Nations*).

In an influential paper, Dyson and Moore (1983) interpreted the north–south gradient in Indian fertility in terms of the kinship structure. North India, Pakistan, and the Indo-Aryan majority in Nepal are part of a wider West Asian kinship system dominated by patrilineage. Its key characteristics include early exogamous marriage for women, a joint family structure, and an emphasis on purdah, which result in very limited autonomy for young wives and considerable pressure to produce sons.

In contrast, southern India and Sri Lanka are part of a South and East Asian kinship system, where endogenous marriage and nuclear family are more prevalent, women's lives are less rigidly controlled, and preference for sons over daughters is less pronounced. According to Dyson and Moore this system, with its greater autonomy for women, is more amenable to demographic innovation, including the adoption of modern birth control. Bhat (1996) concluded that these elements of social structure and kinship account for some 30 to 50 percent of the north–south fertility difference in India. Possibly even more important is exposure to mass media, a result interpreted by Bhat as testimony to the powerful influence of Western ideas on reproductive behavior.

Considerable stress has also been laid on the role of education, particularly for women, as a facilitator of fertility transition. At the individual level the reproductive behavior of unschooled women and those who have received primary schooling differs little, whereas those with secondary schooling invariably have fewer children. Yet the contextual influence of schooling may be more important than the impact of individual attainment. Women in Kerala and Tamil Nadu have lower fertility than those in Bihar and Uttar Pradesh, regardless of their length of schooling (see Fig. 2). While it is widely accepted that high adult literacy has been conducive to reproductive change, it is also clearly not a precondition. In 1991, when Bangladesh was in the midst of rapid fertility transition, only 23 percent of adult women and 36 percent of men were literate.

The contribution of state family planning policies to the fertility transition is also controversial (see *Family Planning Programs: Development and Outcomes*). All five countries have population control policies with explicit demographic goals. All have offered financial incentives (or compensation payments) to individuals who agree to sterilization, and all have been criticized for coercive elements . In Sri Lanka, India, and Nepal, vigorous government promotion of family planning has probably accelerated fertility decline (see, for example, Caldwell et al. 1982). In Bangladesh it may have been even more important. A well-known experiment in the Matlab district demonstrated that the appropriate provision of services could reduce fertility (Phillips et al. 1988) and it has been claimed that, while a preceding mortality decline was the underlying cause of the fertility transition, the family planning program was largely responsible for the timing and speed of change (Cleland et al. 1994). In Pakistan, the program has been largely ineffective because of weak political support.

4. *Prospects for the Future*

India's fertility is projected to reach replacement level by 2015, Bangladesh by 2020, and Nepal and Pakistan by 2030 (see Fig. 1). The prospects for the continuation of past downward trends are good. Recent surveys indicate that at least 20 percent of births are unwanted in Pakistan, Nepal, and the high-fertility northern states of India. However, a desire for two sons is still prevalent in these parts of South Asia. Acting on this preference implies a fertility rate of over four births per woman unless sex-selective abortion becomes widespread. Similar strong preferences for sons exist in East Asia, but the desire for two sons has subsided into a desire for one (see *Fertility Transition: China*; *Fertility Transition: East Asia*). The most likely prognosis is that the remaining high-fertility parts of South Asia will follow this example.

See also: Family Size Preferences; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Reproductive Rights in Developing Nations

Bibliography

Bhat P N M 1996 Contours of fertility decline in India: A district level study based on the 1991 Census. In: Srinivasan K (ed.) *Population Policy and Reproductive Health*. Hindustan Publishing Corporation (India), New Delhi, India, pp. 97–177

Bose A, Premi A K (eds.) 1992 *Population Transition in South Asia*. B. R. Publishing Corporation, New Delhi, India

Cain M T 1983 Fertility as an adjustment to risk. *Population and Development Review* **9**: 688–702

Caldwell J C, Reddy P H, Caldwell P 1982 The causes of demographic change in rural south India. *Population and Development Review* **8**: 689–727

Cleland J, Phillips J F, Amin S, Kamal G M 1994 *The Determinants of Reproductive Change in Bangladesh: Success in a Challenging Environment*. World Bank Regional and Sectional Studies, Washington, DC

Dyson T, Moore M 1983 On kinship structure, female autonomy, and demographic behavior in India. *Population and Development Review* **9**: 35–60

Fernando D F S 1975 Changing nuptiality patterns in Sri Lanka 1901–1971. *Population Studies* **29**: 179–90

Hajnal J 1953 Age at marriage and proportion marrying. *Population Studies* **7**: 110–36

Jeffery R, Basu A M (eds.) 1996 *Girls Schooling, Women's Autonomy and Fertility Change in South Asia*. Sage, New Delhi, India

Leete R, Alam I (eds.) 1993 *The Revolution in Asian Fertility*. Clarendon Press, Oxford, UK

Phillips J F, Simmons R, Koenig M A, Chakaborty J 1988 The determinants of reproductive change in a traditional society: Evidence from Matlab, Bangladesh. *Studies in Family Planning* **19**: 313–34

Vlassoff M, Vlassoff C 1980 Old age security and the utility of children in rural India. *Population Studies* **34**: 487–99

Zachariah K C, Rajan S I (eds.) 1997 *Kerala's Demographic Transition: Determinants and Consequences*. Sage, New Delhi, India

J. Cleland

Fertility Transition: Southeast Asia

The population of Southeast Asia was only one third the size of Europe's in 1950 (182 million compared to 547 million in Europe), but by 2050, the population of Southeast Asia is projected to be 25 percent larger than Europe's (786 million vs. 628 million, United Nations 1999, pp. 442–3). This dramatic reversal is the product of a lag, of about a century, in the demographic transitions of the two regions. Fertility began to decline in several northwestern Europe countries during the last quarter of the nineteenth century and spread to all of Europe over the first half of the twentieth century. Fertility in Southeast Asia first showed signs of decline in a few countries in the late 1960s, and then moved downward, at varying speeds, throughout the region from the 1970s to the 1990s. Although fertility in Southeast Asia is likely to approach the replacement level (around two births per woman) early in the twenty-first century, the force of demographic momentum and the power of compound rates of growth will continue to expand the population of Southeast Asia relative to Europe (and North America).

Southeast Asia consists of the 11 countries that lie between the Indian subcontinent and China (see Fig. 1). While there are common geographical, historical,

Figure 1
Countries of Southeast Asia

Table 1
Estimates of population, life expectancy, and fertility in Southeast Asia

Southeast Asian Countries	POP (in 000) 2000 estimate	Life Expectancy		Total Fertility Rate	
		1965–70	1995–00	1965–70	1995–00
Brunei Darussalam	328	64.9	75.5	5.9	2.8
Cambodia	11,168	45.4	53.4	6.2	4.6
East Timor	885	37.5	47.5	6.2	4.4
Indonesia	212,107	46.0	65.1	5.6	2.6
Laos	5,433	40.4	53.2	6.2	5.8
Malaysia	22,244	59.4	72.0	5.9	3.2
Myanmar	45,611	47.4	60.1	6.0	2.4
Philippines	75,967	56.2	68.3	6.0	3.6
Singapore	3,567	67.9	77.1	3.5	1.7
Thailand	61,399	56.7	68.8	6.1	1.7
Vietnam	79,832	47.9	67.4	6.0	2.6
Total	518,540	47.8	65.7	5.8	2.7

Source: United Nations. 1999. *World Population Prospects: The 1998 Revision*: Volume 1: Comprehensive Tables. United Nations, New York

and cultural features of Southeast Asia, diversity is the hallmark of the region (Reid 1988, 1993, Wolters 1999). Indonesia, one of the world's most populous countries, comprises about 40 percent of population of the region as a whole (see Table 1). The other largest countries are the Philippines and Vietnam, each with estimated populations of more than 75 million in 2000, and Thailand at approximately 60 million. At the other extreme are the microstates of Brunei and East Timor, each with less than a million people. Singapore, a city-state, has a population of only 3.5 million, and Laos is only slightly larger at 5 million (Table 1). Myanmar (Burma) at 45 million, Malaysia with 22 million, and Cambodia with a population of 11 million are the other countries in the region.

The major geographical and cultural divisions are between mainland and insular Southeast Asia and between the lowlands and uplands within each region (and country). The populations of the mainland countries generally practice Buddhism while Islam is the major religion of Indonesia, and Christianity is the predominant religion of the Philippines. While part of Malaysia (Peninsular Malaysia) is on the mainland, it is usually considered as belonging to insular Southeast Asia because the Malay population of Malaysia shares a common language, religion, and cultural heritage with much of the Indonesian population. The lowland countryside of both mainland and insular Southeast Asia consists of villages knitted together by small and medium size market towns. Wet (irrigated) rice agriculture is the predominant feature of the countryside, but rubber, oil palm, coconut, and sugar plantations (and small holdings) as well as mining remain as legacies of the colonial export economies that once dominated the region. Every country has remote highland and mountainous regions, which are often populated by cultural and religious minorities (a minority in one country is sometimes the dominant cultural group in neighboring countries).

Located along the major sea route between the great civilizations of East Asia and South Asia, the 'plural society' has been one of the defining features of Southeast Asia, which has been relatively open and absorptive of peoples, ideas, and cultural practices from elsewhere. In the twentieth century, assimilation into Southeast Asian societies and interethnic relations became more difficult with the creation of 'national' states and the translation of cultural differences by religion, language, and region into political categories with differential status and rights. Throughout the region, the 'overseas Chinese' continue to be over-represented in commercial and professional occupations. Over the centuries, a good share of Chinese-origin peoples have intermarried and blended with local populations in many Southeast Asian counties, but there remains a segment of the Chinese population that has been residentially and socially segregated even after three or four generations of local residence (Cushman and Wang 1988, Reid 1996, Chaps. 1–3).

The new Southeast Asian economies are dominated by the major metropolitan areas, including Jakarta, Surabaya, Bangkok, Singapore, Manila, Rangoon, Kuala Lumpur, and Ho Chi Minh City. These major cities typically originated as trading and administrative areas along major rivers or coastal ports. At the present time, these cities are bustling metropolises with large-scale manufacturing, commercial, service, and governmental functions. Contemporary indicators of development in Southeast Asia, including very low levels of mortality and almost universal secondary schooling, are approaching the prevailing standards of developed countries. The economic crisis that hit the region in 1997 notwithstanding, the

prevailing trajectory of Southeast Asia appears to be heading along the lines of the modernizing East Asian countries of Japan, Korea, and Taiwan.

1. A Demographic Overview

The major demographic attributes and trends of Southeast Asia, based on the latest UN population estimates and projections, are reported in Table 1. The extraordinary pace of demographic and social change in the region is registered in the measures of life expectancy and the total fertility rate (a summary measure of annual fertility expressed in terms of the average number of births per woman) from the late 1960s to the late 1990s. Over this 30-year period, the average life expectancy in Southeast Asia increased from 48 years to 66 years and fertility declined from almost six births per woman to less than three. These rates of change are unprecedented in the historical record and have only been exceeded in the modern era by several East Asian countries. In the late 1960s, the only countries in the region with life expectancy above 60 were Singapore and Brunei, although Malaysia was not far behind. Thirty years later, every country had life expectancies above 60, except for Cambodia, East Timor, and Laos (the poorest countries in the region). The declines in fertility are no less pervasive. Even in the relatively poor countries of Myanmar (Burma) and Vietnam, fertility has dropped below three children per couple. Fertility is below the replacement level of two births per couple in Singapore and Thailand.

The broad picture of rapid fertility decline conceals significant variations from country to country. For example, there has been very little socioeconomic development and only modest demographic change in Laos and Cambodia, countries that were victims of the Indochina wars of the 1960s and 1970s. After the 'victory' of the Khmer Rouge in 1974, the mortality rate may have doubled as a result of executions and starvation during the Pol Pot regime (United Nations 1999, p. 124, Heuveline 1998). There are signs of fertility declines in the UN data for Laos, Cambodia, and East Timor, but the absolute levels of fertility remain very high in these countries. Much further along the path of fertility decline are the large countries of Vietnam and Myanmar, although economic conditions in these counties remain far below the rest of the region (Haugton 1997, Myint 1991).

Fertility dropped below the replacement level in Singapore in the late 1970s (Saw 1999). Perhaps the most celebrated case of rapid fertility decline in the region is Thailand, where fertility declined from more than six births per woman in the 1960s to less than two births per woman in less than 30 years (Knodel et al. 1996, Hirschman et al. 1994). Fertility declines have also been registered in other Southeast Asian countries with estimates of total fertility rates ranging between three and four births in Malaysia, the Philippines, and Indonesia (Hirschman and Guest 1990a, Palmore et al. 1995). The case of falling fertility in Indonesia is particularly salient, given the size of the country and its major influence on the region as a whole.

2. Explanations of Fertility Decline in Southeast Asian Countries

The postponement of marriage has contributed modestly to the temporal decline in fertility in several Southeast Asian countries, but the major component has been declines in marital fertility (Retherford and Cho 1973, Hirschman and Guest 1990a). Declines in marital fertility are due primarily to increased contraceptive use. Indeed other factors, such as decreases in spouse mortality and reduced durations in breastfeeding, would probably have led to rising marital fertility in the absence of increased contraceptive practice. The underlying question is what has caused the increase in contraceptive use.

There are two widespread interpretations for the widespread adoption of contraception and rapid fertility declines in the region: the efficacy of family planning programs and the rapid pace of socioeconomic change in the region. These interpretations are not incompatible, and in their classic study, John Knodel and colleagues conclude that Thailand's reproductive revolution is due to the combination of four factors: rapid social and economic development, a favorable cultural setting (female autonomy and weak intergenerational influences in reproductive decisions), a latent demand for fertility control, and an effective national family planning program (Knodel et al. 1987). The challenge for the field has been to test the relative significance of these explanatory factors in different times and places. At present, the research record points to a broad mosaic of varied patterns and causes without a simple overarching interpretation.

The fundamental prerequisite for fertility transitions is a mortality transition, especially declines in infant and child mortality. In Southeast Asia, mortality declines began in the decade after World War II, as antibiotics, other aspects of modern medicine, and preventive health programs were introduced and spread in many Southeast Asian countries. The continuing pace of mortality declines in the region is shown in Table 1. According to demographic transition theory, declines in mortality and other aspects of socioeconomic change are predicted to increase pressures on families and married couples to lower their desired number of children and to increase their 'demand' for family planning.

In almost every Southeast Asian country, there have been empirical studies that find the predicated associations between modern characteristics and lower levels of fertility. Social change, or modernization, is usually indexed by a variety of individual level characteristics,

such as female education and employment, urban residence and nonagricultural employment, exposure to the mass media, and ownership of modern goods. Such associations, however, do not explain how macrochanges (in society as a whole) lead to individual-level modernization nor whether cross-sectional associations reflect actual historical (over time) patterns of social change. Hirschman and colleagues (Hirschman and Guest 1990b, Hirschman et al. 1998) have attempted to explain fertility declines in four Southeast Asian countries (Indonesia, Philippines, Malaysia, and Thailand) from 1970 to 1990 in terms of changes in provincial level measures of social change in the status of women, the economic roles of children, and infant mortality. Although based on fertility transitions still in process, the findings are consistent with the logic of demographic transition theory, albeit with wide variations in the significance of specific independent variables over time and across countries.

Impressionistic evidence would suggest that governmentally sponsored family planning programs have been the major reason for rising contraceptive use in Thailand, Indonesia, and other Southeast Asian countries with declining fertility. Family planning programs were instituted in many Southeast Asian countries in the 1970s, just about the same time that fertility began to edge downward in the region. Several authors have attributed the decline of fertility in Indonesia to a very successful family planning program there (Freedman et al. 1981, Hull et al. 1977, Ross and Poedjastoerti 1983), and some observers attribute the slower declines in fertility in Malaysia and the Philippines to relatively less energetic family planning programs in these countries. These conjectures, however, are very difficult to evaluate empirically because the emergence and characteristics of family planning programs are correlated with other determinants of fertility decline, most notably socioeconomic development.

One analytical approach to this question is to measure the spatial pattern of fertility decline relative to the geographical location of family planning clinics and aspects of the program that may provide family planning information or contraceptives. The problem with this approach is that the location of family planning clinics may be endogenous, i.e., program resources such as clinics are directed to populations with very high fertility (just as police stations are often located in high crime areas). Moreover, contraceptives are available from nongovernmental outlets such as private clinics and drugstores. These conditions mean that empirical tests of the impact of family planning programs on fertility decline are extraordinarily difficult to estimate empirically.

Using econometric techniques to control for the endogenous role of family planning inputs, Gertler and Molyneaux (1994) find that almost all of the fertility decline in Indonesia from 1982 to 1987 was due to rising incomes and education and that very little of the fertility decline could be attributed directly to family planning efforts. These findings do not rule out the interpretation that the Indonesian family planning programs played an important role through the provision of readily available means for fertility control (and at a modest cost) for women who were already motivated to limit their number of births, which is what Robinson and Rachapaetayakom (1993) argue was the major role of the Thai family planning program. Family planning programs may play a more influential role in the origins of a fertility transition than on its continuation.

Govindasamy and DaVanzo (1992) conclude that public policies, other than family planning programs, have been more influential in promoting more rapid fertility declines among the Chinese and Indian populations than among the Malay population of Malaysia during the 1970s and 1980s. The Malaysian government provided educational and economic subsidies (as part of an affirmative action program) for Malays that reduced the costs of children in a rapidly modernizing society. Leete (1996) adds the additional interpretation that Malay fertility has remained high because of the resurgence of Islam and pronatalist cultural values. There are, however, quite wide variations in levels and trends of Malay fertility between Singapore and Malaysia and across states in Malaysia that raise questions about a simple cultural or religious preference for high fertility (Hirschman 1986, Jones 1990).

Perhaps the single most important cultural characteristic of Southeast Asia is the relatively high status of women, especially when compared with East Asia and South Asia (Reid 1988, pp. 162–72, Van Esterik 1982). While women still face many socioeconomic obstacles in Southeast Asia, the situation is quite different from the patriarchal societies of East and South Asia. Southeast Asian kinship systems are typically bilateral with equal importance attached to the husband's and wife's families. For the most part, there is no strong sex preference for children in Southeast Asia (Wongboonsin and Ruffolo 1995) nor are there strong prescriptions on residence with the groom's family after marriage. Cultural variables are difficult to quantify, especially at the microlevel, which makes it difficult to estimate their relative role in shaping fertility levels and change, but it seems likely that greater freedom for women to pursue nonfamilial roles and to influence household decision making have been important factors in the region's rapid fertility decline.

3. *The End of the Fertility Transition in Southeast Asia*

If the current pace of decline continues, replacement-level fertility (two children per woman) should be reached early in the twenty-first century in most

countries and regions of Southeast Asia. Indeed, if the recent history of East Asia, and Singapore, is any guide, modern fertility levels may not stop at the replacement level, but continue to decline below the level of fertility needed to replace the parental generation. The explanations for these rapid fertility declines, although still hotly debated, are likely to be the complex interplay of rapid socioeconomic change, organized family planning programs, and spatial and social diffusion processes that vary in significance and intensity over time.

An additional factor that is likely to accelerate the decline of fertility is the postponement and perhaps decline in marriage. The most celebrated case is Singapore, where the 'the flight from marriage' and declining fertility among the highly educated population prompted the government to organize social activities to encourage unmarried persons in their twenties and thirties to marry and have babies (Leete 1994, p. 815, Yap 1995). The trend toward marital postponement throughout the region has been underway for 30 years, but the 1990 round of censuses revealed how far the trend had progressed. In Malaysia, Thailand, and the Philippines, from 13 to 15 percent of women in their early thirties remained unmarried (Jones 1995, 1997).

The figures for marital postponement are even higher in cities and among those with higher education. In several countries of Southeast Asia, the average age at marriage of women and men is older than in many Western countries. Since the historical pattern in Southeast Asia was of young and universal marriage, this is a remarkable transformation in custom. Perhaps, there may be a lag in the adjustment of social institutions to replace traditional matchmaking arrangements. Another hypothesis is that some highly educated men in Southeast Asia may prefer to marry women with lower educational attainment who are considered to more likely to follow traditional gender roles than highly educated women.

Although there is likely to be an adjustment in marriage patterns before high levels of celibacy are reached, the current pattern of marital postponement is certain to likely to accelerate the downward trend in fertility in most Southeast Asian countries. Indeed, the fertility transitions in Southeast Asia are among the most rapid in the developing world, and the trend may accelerate. Nonetheless, rates of rapid population growth of 2 percent per annum during much of the second half of the twentieth century and subsequent demographic momentum will lead to continued rapid growth of the population of Southeast Asia during the first half of the twenty-first century.

See also: Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility Control: Overview; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Land Rights and Gender; Southeast Asia: Sociocultural Aspects; Southeast Asian Studies: Gender; Southeast Asian Studies: Health

Bibliography

Cushman J, Wang G (eds.) 1988 *Changing Identities of the Southeast Asian Chinese Since World War II*. Hong Kong University Press, Hong Kong

Freedman R, Khoo S-E, Supraptilah B 1981 Use of modern contraceptives in Indonesia: a challenge to the conventional wisdom. *International Family Planning Perspectives* **7** (March): 3–15

Gertler P J, Molyneaux J W 1994 How economic development and family planning programs combined to reduce Indonesian fertility. *Demography* **31**: 33–63

Govindasamy P, DaVanzo J 1992 Ethnicity and fertility differentials in peninsular Malaysia: do policies matter? *Population and Development Review* **18**: 243–67

Haughton J 1997 Falling fertility in Vietnam. *Population Studies* **5**: 203–11

Heuveline P 1998 Between one and three million: toward the demographic reconstruction of a decade of Cambodian history (1970–1979). *Population Studies* **52**: 49–65

Hirschman C 1986 The recent rise in Malay Fertility: a new trend or a temporary lull in a fertility transition?. *Demography* **23** (May): 161–84

Hirschman C, Guest P 1990a The Emerging Demographic Transitions of Southeast Asia. *Population and Development Review* **16**: 121–52

Hirschman C, Guest P 1990b Multilevel models of fertility determination in four Southeast Asian countries: 1970 and 1980. *Demography* **27**: 369–96

Hirschman C, Tan J-E, Chamratrithirong A, Guest P 1994 The path to below replacement-level fertility in Thailand. *International Family Planning Perspectives* **20**: 82–7, 107

Hirschman C, Tan J-E, Chamratrithirong A, Guest P 1998 Explaining the rapid fertility decline in Thailand. Unpublished paper, Center for Studies in Demography and Ecology, University of Washington, Seattle, WA

Hull T H, Hull V J, Singarimbun M 1977 Indonesia's family planning story: success and challenge. *Population Bulletin* **32(6)**: 1–51

Jones G W 1990 Fertility transitions among Malay populations of Southeast Asia: puzzles of interpretation. *Population and Development Review* **16**: 507–37

Jones G W 1995 Population and the family in Southeast Asia. *Journal of Southeast Asian Studies* **26**: 184–95

Jones G W 1997 The demise of universal marriage in East and South-East Asia. In: Jones G W, Douglas R M, Caldwell J C, D'Souza R M (eds.) *The Continuing Demographic Transition*. Clarendon Press, Oxford, UK pp. 51–79

Knodel J, Chamratrithirong A, Debavalya N 1987 *Thailand's Reproductive Revolution: Rapid Fertility Decline in a Third-World Setting*. University of Wisconsin Press, Madison, WI

Knodel J, Ruffolo V P, Ratanalangkarn P, Wongboonsin K 1996 Reproductive preferences and fertility trends in post-transition Thailand. *Studies in Family Planning* **27**: 307–18

Leete R 1994 The continuing flight from marriage and parenthood among the overseas Chinese in East and Southeast Asia. *Population and Development Review* **20**: 811–29

Leete R 1996 *Malaysia's Demographic Transition: Rapid Development, Culture, and Politics*. Oxford University Press, Kuala Lumpur

Myint N 1991 Recent levels and trends of fertility and mortality in Myanmar. *Asia-Pacific Population Journal* **6**: 3–20
Palmore J A, Made Mamas S G, Arifiyanto Y 1995 Fertility decline in Indonesia, 1971–1991. *Journal of Population* (*Jakarta*, *Indonesia*) **1**: 45–66
Reid A 1988 *Southeast Asia in the Age of Commerce 1450–1680. Vol. 1. The Lands Below the Winds*. Yale University Press, New Haven, CT
Reid A 1993 *Southeast Asia in the Age of Commerce, 1450–1680. Vol. 2. Expansion and Crisis*. Yale University Press, New Haven, CT
Reid A (ed.) 1996 *Sojourners and Settlers*: *Histories of Southeast Asia and the Chinese*. Allen and Unwin, St. Leonards, Australia
Retherford R, Cho L-J 1973 Comparative analysis of recent fertility trends in East Asia. In: *International Population Conference*. International Union for the Scientific Study of Population, Liege, Vol. 2, pp. 163–81
Robinson W C, Rachapaetayakom J 1993 The role of government planning in Thailand's fertility decline. In: Leete R, Alam I (eds.) *The Revolution in Asian Fertility: Dimensions, Causes, and Implications*. Clarendon Press, Oxford, UK pp. 54–66
Ross J, Poedjastoerti S 1983 Contraceptive use and program development: new information from Indonesia. *International Family Planning Perspectives* **9**: 68–77
Saw S H 1999 *The Population of Singapore*. Singapore Institute of Southeast Asian Studies, Singapore
United Nations 1999 *World Population Prospects: The 1998 Revision: Volume 1: Comprehensive Tables*. United Nations, New York
Van Esterik P 1982 *Women of Southeast Asia*. Dekalb: Center for Southeast Asian Studies, Northern Illinois University, Urbana, IL
Wolters O W 1999 *History*, *Culture*, *and Region in Southeast Asian Perspectives*, rev. edn. Southeast Asia Program, Cornell University, Ithaca, NY
Wongboonsin K, Ruffolo V P 1995 Sex preference for children in Thailand and other Southeast Asian countries. *Asia-Pacific Population Journal* **10**: 43–62
Yap M-T 1995 Singapore's 'three or more' policy: the first five years. *Asia-Pacific Population Journal*. **10**: 39–52

C. Hirschman

Fertility Trends in the Formerly Socialist Countries of Europe

The Formerly Socialist Countries of Europe (FSCE) include Albania, Bulgaria, Hungary, Poland, and Romania, as well as the newly independent states, namely the Czech Republic and the Slovak Republic from the former Czechoslovakia; Belarus, Estonia, Lithuania, Latvia, Moldova, Russia, and Ukraine from the former USSR; Croatia, Bosnia and Herzegovina, Slovenia, Macedonia, and Yugoslavia from the Socialist Federated Republic of Yugoslavia; and finally the eastern *Länder* of Germany, which from May, 1949 to October 3, 1990 formed an independent state—the German Democratic Republic (GDR).

Relatively heterogeneous demographically, the FSCE have *grosso modo* followed a common pattern during the twentieth century, namely a late onset and high speed of the fertility transition as compared to Western and Northern Europe; a specific environment of family economics and other family policies applied under the socialist system during the four post-war decades; and a rapid transition from a centralized to a market economy since the beginning of the 1990s.

1. *The Fertility Transition*

The fertility transition started during the last quarter of the nineteenth century in the more industrialized and urbanized Czech Republic (which was part of the Austro-Hungarian Empire until 1918), in the former GDR (Western Prussia), and in Hungary (the Kingdom of Hungary in the Austro-Hungarian Empire). Elsewhere (Albania excluded), the beginning of a stable fertility decline may be dated to about the first two decades of the twentieth century, though exact estimation is difficult because of social troubles and wars in 1912–20 (Coale 1969, Pavlik 1991). The completed fertility (CF) of women born in 1910, a central cohort whose reproductive behavior reflected fertility dynamics during the first half of the twentieth century, was below 2.5 births per woman in the Czech Republic, the GDR, and Hungary (as it was in England and Wales), about 3.0 in Bulgaria and Slovakia (as in The Netherlands), and between 3.5 and 4.0 in Poland, Rumania, Yugoslavia, and Russia (Festy 1991). In general, the FSCE followed the West European fertility transition pattern between the two World Wars, and on the eve of World War II, Albania was the only country not to have begun its fertility transition.

2. *Trends after World War II*

Fertility in the FSCE in the second half of the twentieth century was marked by a notable deviation from the Western and Northern European pattern. While fertility rose throughout the West after the 1950s (see *Baby Booms and Baby Busts in the Twentieth Century*), in the FSCE it continued to decline without interruption, reaching the then European low in the early 1960s (Sardon 1998). In Hungary, the TFR fell below 2 in 1961, in Romania in 1964, in the Czech Republic in 1967, in the Ukraine in 1965, and in Russia in 1968. The TFR level necessary for generation replacement (at least 2.1) was still maintained in Poland, the GDR, Belarus, Bulgaria, Moldova, and all of Yugoslavia,

but this was mainly caused by a lowering of childbearing age, which fell by two years between 1950 and 1965 (see *Fertility Change: Quantum and Tempo*), while cohort fertility continued to fall.

Furthermore, in Yugoslavia the relatively high national fertility level concealed strong regional and ethnic variance: while the TFR was about 3.5 in Bosnia and Herzegovina in 1965, 3.7 in Macedonia, and 3.1 in Montenegro, it was approaching replacement level in the other Yugoslav republics: 2.02 in Croatia, 2.4 in Slovenia and Serbia. Within Serbia, again an enormous difference existed between Kosovo, where the TFR exceeded six births per woman in the mid-1960s (as it did in Albania), and the other two autonomous provinces, Central Serbia and Vojvodina, where it was 2.03 and 2.13, respectively.

From the mid-1960s, when the fertility decline took new speed in Western Europe (see *Demographic Transition, Second*), the TFR stabilized for almost two decades at values of 2.0 to 2.5 in Central and Eastern Europe. Thus in the 1970s and 1980s the FSCE once again had the highest fertility levels in Europe. The stabilization of fertility in the FSCE might be explained by a particular 'socialist' type of family formation resulting from policies of social paternalism towards the family, especially toward the young family, applied throughout the FSCE since the beginning of the 1950s. Under the conditions generated by these policies, based on the principles of social egalitarianism and state guarantees of equal and universal access to education and health services, family formation (marriage and childbearing) became one of the first stages of an adult's social adaptation. It often was a condition for independent housing arrangements and it preceded a definitive choice of profession. This resulted in younger childbearing and in a strong compression of fertility around the modal age at childbearing. Furthermore, this policy stimulated the development of a two-child family system. Throughout the FSCE, the proportion of women who were childless at age 50 was very close to the natural level of sterility, i.e., about 10 percent. At the same time, families with four and more children were rare, especially in the countries that had a high level of industrialization in the pre-socialist period (the GDR, the Czech Republic, Hungary, Slovenia).

In the countries where economic modernization had not been achieved by World War II (those east of the 'Hajnal line'), the traditions of early and universal marriage as well as early childbearing persisted, because the 'gentle' socialist form of industrialization avoided the explosion of unemployment, a drop in living standards, and mass rural–urban exodus.

Another factor explaining the stabilization of (or even slight increase in) fertility was the special pronatalist policies implemented between the mid-1960s and the mid-1980s throughout the FSCE, again with the exception of Albania (Henry and McIntyre 1981). Based on an array of incentives such as family allowances, birth payments, paid maternity and post-maternity leave, and subsidized credit, these policies aimed at stimulating second-, third- and fourth-order births (Klinger 1991). The most powerful pronatalist incentives were introduced in Hungary, Czechoslovakia, Bulgaria, and the GDR, where they substantially influenced period fertility indicators, and less strongly the fertility of female cohorts born around 1950 (± 5 years). The strongest growth in cohort fertility was observed in Hungary, where CF rose from 1.9 for women born in 1942–45 to 2.03 in the cohort of 1961. In Czechoslovakia, CF gained less, from 2.03 in the cohort of 1943 to 2.1 in the cohort of 1952–58, while in the GDR and Bulgaria, the pronatalist measures only stopped the decline and stabilized cohort fertility (at around 1.8 in the former for the cohorts born in 1952–58 and 2.05 in the latter in the cohorts born in 1945–57. Thus, all these incentives were unable to raise cohort fertility to replacement level. In 1983–86, the USSR repeated this pronatalist policy experiment, but without any restrictions on abortion, and obtained the same result in its European republics, namely a short-term increase of the period TFR to 2.2 in 1984–87, followed by a rapid decline after the fall of the Soviet Union in 1988. In Romania and Poland, the pronatalist measures were more modest. Finally, in Yugoslavia, child allowance payments were decentralized and increased for lower-income families, which suggests a social policy orientation rather than a purely pronatalist policy (Henry and McIntyre 1981).

The availability of induced abortion was another key element of the family policies of a number of FSCE countries. Following the example of a Soviet abortion law passed in November 1955, all the FSCE except the GDR and Albania liberalized induced abortion during the late 1950s. Later, the abortion law became more restrictive again in Romania (1966), Bulgaria (1968), Czechoslovakia (1973), and Hungary (1974). These restrictions had little long-term effect on fertility because they were accompanied by an extension in contraceptive use, encouraged by the governments to protect women's reproductive health (see *Fertility Control: Eastern Europe*). Only in Romania, where the level of modern contraceptive use was very low, did the strongest ban on abortion ever (implemented on 1 November, 1966 by Decree 770) result in the most spectacular rise in fertility in post-war Europe: the TFR increased from 1.92 in 1966 to 3.67 in 1967. Fertility fell quickly again, after 1969 however, and later amendments reinforcing Decree 770 (in 1974, in 1984, and in 1986) had only a very weak influence on trends in the TFR.

The main results of these fertility-stimulating and restrictive incentives were a strong shift in fertility to younger ages (the cohort MAC—mean age at childbearing—decreased to 25 ± 1 years) and a greater compression of fertility spacing toward its natural limits within the frame of the two-child system, with a

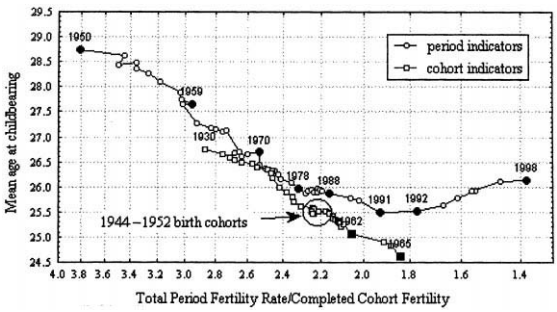

Figure 1
Evolution of period and cohort fertility indicators in the formerly socialist countries of Europe
Source: Data Base of European Demographic Observatory communicated by G Calot and J P Sardon; the estimation of the TFR in the USSR republics for the period until 1975 was performed by Alain Blum. (Personal communication)

CF in the region of 1.9 births per woman (see Fig. 1). In the GDR, where fertility compression was relatively moderate by FSCE standards, about 60 percent of fertility was concentrated in the age range of 19–25 years in the mid-1970s. By the beginning of the 1980s, this pattern of reproductive behavior had spread throughout the FSCE except for Albania, the autonomous province of Kosovo and, probably, Macedonia.

In the early 1980s, when the crisis of the socialist system emerged throughout the FSCE and was absolutely obvious in Poland and the former Yugoslavia, the first signs of a corrosion of the 'socialist' pattern of family formation also became visible. This showed up in an increase in the mean age at childbearing (Hungary, the GDR, Czech Republic, Croatia, Slovenia), followed by a decrease in TFR since the mid-1980s. In Central Serbia and the autonomous province of Vojvodina, where the economic crisis began earlier, the MAC started to rise in the early 1970s while fertility was below replacement level. Only in the European republics of the USSR, Poland, Bosnia and Herzegovina, Macedonia, and Kosovo was the fertility decrease accompanied by a reduction in the MAC until the mid-1990s.

With the collapse of the socialist system in Europe (1989–1991) and the disappearance of the social and economic support of the 'socialist' pattern of family formation, fertility plummeted in all FSCE countries. In the ex-GDR, the Czech Republic, the Slovak Republic, Romania, and the newly independent states of the former USSR, the first reaction of fertility to the social changes has been described as a paralyzing shock that led to an abrupt but brief fall in the mean age at childbearing due to a decrease in fertility rates at all ages. Subsequently MAC began to increase while the TFR decreased. The most spectacular fertility drop occurred in the former GDR, where the TFR fell from 1.56 in 1989 to 0.77 in 1993–1994. In the rest of the FSCE, fertility followed the trends of the late 1980s, except perhaps a little faster. It seems that the economic troubles definitively brought the fertility of women born in 1962–65 down below the level that would have been reached otherwise. The fertility of younger women has been checked too, but this may perhaps be made up later (Philipov and Kohler 1999). Moreover, a rapid rise in out-of-wedlock births is observed all over the FSCE and is especially remarkable in Estonia, where the proportion of births outside marriage rose from 20.7 percent in 1985 to 51.6 percent in 1997. In the former GDR, the increase was from 33.8 percent to 44 percent, in Slovenia from 19.1 percent to 32.7 percent, in Latvia from 14.4 percent to 34.8 percent and in Bulgaria from 11.7 percent to 30.1 percent.

In summary, it seems likely that a new pattern of family formation is now emerging in the FSCE, characterized by later marriage and by late and diversified fertility similar to that observed in Western Europe. This observation is supported by the stabilization of (or even a slight increase in) period fertility observed in the former GDR, Moldova, Romania, and the Czech Republic in the second half of the 1990s, in contrast to the fertility decline still going on elsewhere in the FSCE. At the present time it is, however, impossible to draw any definitive conclusions about the near future of fertility in the FSCE. What the future will bring depends in large part on the political stability and the economic situation in these countries.

See also: Eastern European Studies: Culture; Eastern European Studies: Economics; Eastern European Studies: History; Eastern European Studies: Politics; Eastern European Studies: Society; Fertility Control: Overview; Reproductive Rights in Developing Nations

Bibliography

Coale A 1969 The decline of fertility in Europe from the French Revolution to World War II. In: Berhman S J, Corsa L, Freedman R (eds.) *Fertility and Family Planning: A World View*. Ann Arbor: University of Michigan Press, pp. 3–34

Festy P 1991 La fécondité en Europe de l'Est depuis 1950. *Population* **46**(3): 479–510

Henry P D, McIntyre R J 1981 *Reproductive Behavior: Central and Eastern European Experience*. Springer, New York

Klinger A 1991 Les politiques familiales en Europe de l'Est depuis 1950. [Population policies in Eastern Europe since 1950]. *Population*. **46**: 511–26

Macura M 1997 Fertility and nuptiality changes in Central and Eastern Europe: 1982–1993. *Stanovnistvo* **3–4**: 11–43

Pavlik Z 1991 Les tendances démographique longues en Europe de l'Est. [Long-term demographic trends in Eastern Europe]. *Population* **46**(3): 463–78

Philipov D, Kohler H-P 1999 *Tempo Effects in the Fertility Timing and Quantum Effects on Fertility in Bulgaria. The Czech Republic and Russia*. Max Planck Institute for Demographic Research. Working Paper 1999–008

Sardon J-P 1998 Fécondité, bouleversements politiques et transition vers l'économie de marché en Europe de l'Est. *Population, Espace, Sociétés* **3**: 339–60

A. Avdeev

Feud and Internal War: Legal Aspects

Feuding and internal war are two types of violent collective conflict. Long of interest to anthropologists, feuding and internal war may be most common in simple, kin-based societies, but, as lethal violence between modern urban gangs attests, are not confined to those settings. A central issue for scholars has been the relationship between feuding and internal war and the state's system of social control or law: are feud and internal war the antithesis of law or the functional equivalent of law? In fact, they are neither. The emergence over the past quarter-century or so of a new field of sociological inquiry known as conflict management provides a new conception of the relationship, one in which law and violence are similar, though not identical, ways of handling disputes triggered by similar, though not identical, social conditions (Black 1976, 1993).

1. *The Terms 'Feud' and 'Internal War'*

Scholars use two strategies to define the terms 'feud' and 'internal war,' strategies that might be labeled qualitative and quantitative, respectively. The qualitative strategy treats feud and war as distinct subtypes of violence; the quantitative strategy locates feud and war at different points on the same continuum of violence.

The work of Keith Otterbein exemplifies the qualitative strategy, the more common of the two. In an influential paper, Otterbein and Otterbein (1965) isolate three types of violence: feuding, internal and external war. Feuding the Otterbeins define as violence within a political community; internal war as violence between political communities within a society; and external war as violence between societies. Useful though these distinctions may be for some purposes, they do not appear to be necessary to answer the central theoretical question: what explains violent conflict? Otterbein and Otterbein found that the best predictor of feuding in a society is not the complexity of political organization but the presence of co-resident related males organized for mutual self-protection or 'fraternal interest groups.' In a later paper, Otterbein (1968) reports that fraternal interest groups are also strongly associated with internal war (compare Ericksen and Horton 1992). If the same variable predicts both types of violence, much of the rationale for treating them as qualitatively distinct disappears.

The quantitative strategy is used by Donald Black. Black has developed a field of sociological inquiry known as social control or conflict management that addresses the myriad ways in which people and groups handle disputes. Diverse though conflict management behavior is, Black argues that it can be reduced to five 'elementary forms': negotiation, avoidance, toleration, and the two most material to the present discussion, self-help (or aggression), and third-party settlement (such as mediation, arbitration, or adjudication). Each elementary form can be found at different structural levels: among individuals, organizations, nations. For Black, there is therefore no sharp discontinuity between interpersonal, group, and international self-help. However, self-help does vary in the amount of violence it entails. Measures of the violence of self-help include the presence and lethality of weapons, the number of people attacked, the reciprocity of the violence, and whether the assault is accompanied by torture or mutilation. On this line of thought, feud and war (as well as other forms of self-help, such as assault, homicide, and retaliatory killing) differ not in kind but in degree: war is more violent. (Consistent with it, the remainder of this entry will refer to both feuding and internal war as feuding, except where it is important to distinguish the two.) In practice, there is often an additional contrast: the parties to war are usually large groups, such as societies (external war) or tribes (internal war) while feuds are contested by smaller groups, such as families. In the quintessential war, then, armies trade multiple rounds of lethal violence, seek to slaughter as many of the enemy as possible, and exempt nobody from attack. In the quintessential feud, antagonistic families kill at least once, abide by the principle of 'a life for a life,' and avoid targeting women, children, and non-kin.

2. *Feuding and Its Alternatives*

Although feuding has been practiced in many societies, there is no such thing as a feuding society. Only some conflicts give rise to feuds, even among those people well known for feuding, such as medieval Icelanders (Miller 1990), highland Albanians (Hasluck 1954), and Australian aborigines (see, e.g., Warner 1958). For instance, while homicides between families may precipitate prolonged and bitter feuds, those within families are likely to attract nothing more than some criticism and avoidance of the killer (see, e.g., Peters 1967).

Even when homicides, insults to honor, or other offenses occur within relationships that can result in a feud, a feud is rarely inevitable. In many societies, the victims's kin have the option of accepting compensation from the offender's kin in lieu of retaliation (see, e.g., Boehm 1984). Whether they agree to compensation or launch an attack (which may then trigger a counterattack) depends on a variety of factors, including their social status relative to that of the other side, and the nature and number of cross links between them (see Black 1993, Chap. 3). If they decide to seek compensation, they can arrange payment either by directly negotiating with the other side or, more frequently, by calling upon the services of a third-party skilled in the resolution of conflict, such as a mediator. Among the Nuer (of Sudan), for instance, a third-party known as the Leopard-Skin Chief traditionally intervened in disputes, such as homicides, with a view to having the parties agree to settle the case non-violently through a transfer of cattle from the killer's kin to the victim's kin (Evans-Pritchard 1940, pp. 150–76). The Leopard-Skin Chief could not force the parties to resolve the matter peaceably; nor did he referee the merits of the dispute; the only tools available to him were his powers of persuasion and his threat of supernatural sanctions if the parties did not settle. The Leopard-Skin Chief's mediation was most likely to succeed in disputes between neighbors and acquaintances: if the parties were intimates his intervention was likely to be unnecessary; if they were relative strangers his efforts were likely to be disregarded.

3. *The Relationship Between Law and Feud*

The noted anthropologist, Paul Bohannan (1967, p. xiii) once wrote that 'there are basically two forms of conflict resolution: administered rules and fighting. Law and war.' Later developments in the sociology of conflict management reveal that Bohannan's dichotomy is too simple. Recall Black's analysis of the five elementary forms, a scheme that goes beyond law and warfare to include negotiation, avoidance, toleration, and non-legal forms of third-party settlement. However, even though Bohannan's dichotomy neglects these additional forms, it may be correct in pitting law and war as opposites, as radically contrasting ways of handling conflict.

3.1 *Law and Feud as Opposites*

The idea that law and violence are fundamentally opposed is a popular one. Where law represents order and reason and civilization, feud represents disorder, passion, and tribalism. Pospisil (1968, p. 392) articulates a central tenet of this position when he states:

> That law stands in opposition to feud and that the latter is the antithesis of the former … is well documented in those societies in which feuds are stopped by legal decisions of the over-all authority who either has enough power to enforce his will or possesses the skill to persuade the quarrelling parties to accept his solutions.

In reality, however, law does not invariably strive to eliminate violence. Early states appear to tolerate feuding (see, e.g., Otterbein and Otterbein 1965). Moreover, even when law does seek to bring fighting to an end it often must do so with tactics and procedures similar to those used by the protagonists. If the parties persist in pursuing their conflict violently, the only way to stop them is to use overwhelming force. However, in employing violence to end violence, the law becomes not so much a party outside the feud as a more powerful party that wins the feud. That law's authority ultimately rests on physical force blurs the distinction between law and feud.

3.2 *Law and Feud as Complementary*

A second view is that law and feud are complementary institutions, at least where law is unavailable. Several eminent anthropologists, including Evans-Pritchard (1940) and Max Gluckman (1956) view feuding and the threat of feuding as a mechanism by which social order is attained and maintained in stateless societies. Evans-Pritchard, for instance, studied the Nuer who have a segmentary lineage system under which opposing kinship subunits unite to support their kinship unit in a conflict with another unit, and those opposing kinship units themselves unite when conflict occurs between larger kinship units. Feuds are least likely to occur within subunits where the parties are bound by ties of kinship and the necessity of daily cooperation; feuds are more common, and more difficult to settle, as the parties become more structurally distant and cooperate less in matters of economics or conflict. Evans-Pritchard (1940, p. 159) proposes that the 'function of the feud … is, therefore, to maintain the structural equilibrium between opposed tribal segments which are, nevertheless, politically fused in relation to larger units.'

The idea that the feud is the functional equivalent of law in stateless societies qualifies but does not negate the idea that law and feud are opposed. Functional equivalence only applies to stateless societies but feuds and internal wars involving gangs, guerrillas, and other groups may be found in societies with well-established legal systems.

Moreover, only in some stateless societies can feud be the functional equivalent of law. There are stateless societies that have neither law nor feuding, societies in which people employ entirely peaceful methods of settling disputes (see, e.g., Howell and Willis 1989).

Finally, the idea that violence is functional sits better with feuding than with internal war: the indiscriminate nature of war can result in the political and even physical destruction of a society, rather than its continued existence.

3.3 Law and Feud as Conflict Management

What, then, is the relationship between law and feuding? The young discipline of conflict management sheds new light on this issue, revealing both the contrasts and continuities between law and violence. As modes of handling conflict, law and feuding, it is true, are opposed in certain ways. In law, the parties use words; in feuding, they use weapons. Law tends to be hostile to feuding such that in developed or strong states, feuding is met with forceful criminal punishment. Moreover, in law, it is not the adversaries who typically determine the outcome, but neutral third parties, often applying general rules. However, consider some of the ways in which law and feuding are similar:

(a) In both law and feuding the parties confront one another. They dispute through hostile encounters. In so doing, they behave quite differently from disputants who simply avoid their adversary or tolerate the grievance (see Baumgartner 1988).

(b) In both law and feuding the parties seek victory. They do not want or expect to compromise. By contrast, when people negotiate their conflicts or invoke a mediator they normally anticipate that any settlement will require them to make concessions to the other side.

(c) In both law and feuding third-party supporters typically get involved. Legal conflicts draw in lawyers and witnesses; feuds draw in warriors. In either case, supporters are often crucial to the outcome of the case (Black 1993, Chap. 7).

Black's work not only allows the close similarities and the distinct dissimilarities to be described but to be explained as well. Black argues that the explanation of why violence or law or any other type of conflict management is found lies not in the substance of the conflict but in its social structure. The social structure of the conflict consists of the social characteristics of the participants, including their wealth, social integration, reputations for deviance, and relational distance. Black (1995, p. 855, note 130) summarizes the social structure of the classic feud:

> (1) intermediate relational distance between the parties (neither strangers nor close associates), (2) relational segmentation between the parties (divided by a social chasm, without cross links), (3) functional independence of the parties (lack of division of labor), (4) solidarity of each party (intimate and homogeneous), (5) homogeneity between the parties (same ethnicity), and (6) equality between the parties (similar size and resources).

Black adds:

> Narrow the distances in the model by reducing the relational separation and independence of the parties, and the reciprocity and continuity of the violence declines. Increase the distances …, and the violence becomes more indiscriminate and warlike.

Now consider litigation. According to Black's (1976) theory of law (a theory supported by a substantial amount of evidence), the first three characteristics of the social structure of the feud above would also be conducive to extensive litigation. However, the fourth, fifth, and sixth would not: litigation is attracted more to conflicts between unequals than equals, and to people of different ethnicities than the same ethnicity. In other words, conflicts in which the parties fight and those in which they invoke law have social structures that, while not identical, are similar.

3.4 The Fate of Feud-conflicts under Law

The affinity between the social structures of law and of violence provides an answer to another puzzle: how do disputants handle the underlying conflicts when feuding becomes increasingly difficult? What happens to the feud-conflicts that legal authorities successfully prohibit being fought? Cross-cultural evidence suggests that, in the absence of violence, the contending parties generally turn to forms of conflict management similar to violence. Rarely, for example, do the parties simply tolerate the grievance. Nor do they avoid one another or even engage in extensive negotiations. Most commonly, they resort to one of two courses of action. The first is to replace fighting with litigation. Parties who once traded blows now trade legal claims in a kind of sublimated violence. For instance, before the arrival of state authority, the Mae Enga of New Guinea had a substantial amount of interclan fighting, resulting in a comparatively high rate of lethal violence. The colonial administration's efforts to end the fighting were initially unsuccessful, but eventually the Enga began to turn to litigation:

> This litigation the Central Enga quickly defined as 'fighting in the courts,' and they regarded the outcomes as equivalent to military victories or losses affecting whole phratries. Thus the men of a growing clan who believed their group required more land would carefully concoct a case, often spurious, against a weaker non-fraternal neighbor. After patiently rehearsing his followers in the statements they were to make, the Big Man [i.e., Enga leader] … took the matter to the kiap with the realistic expectation that the sheer weight of testimony would persuade the overworked officer, who had neither the time nor the facilities to sift all the evidence. The indignant defendants of course argued strenuously in rebuttal, but, given the absence of written land records and the intrinsic difficulty of determining 'historical facts' … it is

> understandable that resolute lying often carried the day. However the case was decided, the disgruntled losers would soon fabricate a counterclaim (or polish up a valid complaint) against an apparently vulnerable fraternal clan of the victors (Meggitt 1977, p. 155).

The second course of action is to make increased use of indigenous forms of third-party settlement. If, previously, feuding and mediated compensation payments were alternative ways of handling, say, homicides, mediation payments now come to be the dominant method. For instance, after coming under colonial rule, the Nuer began, with the encouragement of the authorities, to bring all homicides to the Leopard-Skin Chief to be resolved through compensation (though the killer was also prosecuted and punished separately). Thus, by settling homicide cases between lineages that once might have resulted in reprisals, the role of feud contracted while that of mediation expanded (Howell 1954, p. 61).

Whichever option—using judges or mediators—is pursued, both are possible because of the similarity in the underlying social structure of violence and litigation and, more generally, third-party settlement. The arrival of law into a setting previously stateless provides a new outlet for disputes but it does not, at least in the short term, fundamentally alter the underlying relationship between the parties. Hence, when law forbids feuding, former feudists are usually attracted more to litigation, mediation, or arbitration than to forms of conflict management bearing fewer resemblances to violence. Law and feuding, then, are neither opposites nor complements, neither strangers nor siblings. Perhaps they are best thought of as cousins.

See also: Conflict and Conflict Resolution, Social Psychology of; Domestic Violence: Sociological Perspectives; Internal Warfare: Civil War, Insurgency, and Regional Conflict; Violence, History of; Violence in Anthropology; War, Sociology of; Youth Gangs

Bibliography

Baumgartner M P 1988 *The Moral Order of a Suburb*. Oxford University Press, New York

Black D 1976 *The Behavior of Law*. Academic Press, New York

Black D 1993 *The Social Structure of Right and Wrong*. Academic Press, San Diego, CA

Black D 1995 The epistemology of pure sociology. *Law and Social Inquiry* **20**: 829–70

Boehm C 1984 *Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies*. University of Pennsylvania Press, Philadelphia, PA

Bohannan P 1967 Introduction. In: Bohannan P (ed.) *Law and Warfare: Studies in the Anthropology of Conflict*, 1st edn. Natural History Press, Garden City, NY

Ericksen K P, Horton H 1992 Blood feuds: cross-cultural variations in kin group vengeance. *Behavior Science Research* **26**: 57–85

Evans-Pritchard E E 1940 *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People*. Oxford University Press, London

Gluckman M 1956 *Custom and Conflict in Africa*. Barnes and Noble, New York

Hasluck M 1954 *The Unwritten Law in Albania*. Cambridge University Press, Cambridge, UK

Howell P P 1954 *A Manual of Nuer Law*. Oxford University Press, London

Howell S, Willis R (eds.) 1989 *Societies at Peace: Anthropological Perspectives*. Routledge, London

Meggitt M 1977 *Blood is their Argument: Warfare among the Mae Enga Tribesmen of the New Guinea Highlands*, 1st edn. Mayfield, Palo Alto, CA

Miller W I 1990 *Bloodtaking and Peacemaking: Feud, Law, and Society in Saga Iceland*. University of Chicago Press, Chicago

Otterbein K F 1968 Internal war: a cross-cultural study. *American Anthropologist* **70**: 277–89

Otterbein K E, Otterbein C S 1965 An eye for an eye, a tooth for a tooth: a cross-cultural study of feuding. *American Anthropologist* **67**: 1470–82

Peters E L 1967 Some structural aspects of the feud among the camel-herding Bedouin of Cyrenaica. *Africa* **37**: 261–82

Pospisil L 1968 Feud. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York

Warner W L 1958 *A Black Civilization: A Social Study of An Australian Tribe*. Harper and Brothers, New York

M. Cooney

Feudalism

Definitions of feudalism 'vary widely in their degree of comprehensiveness' (Max Weber). Historians and social scientists of the nineteenth and twentieth centuries thus do not agree on a standard definition of the term. There is a great deal of controversy about the scope and the explanatory value of the term feudalism. Characteristic of this debate is a plurality of concepts of feudalism. The diversity of definitions and concepts is attributable to differences in both the epistemological interests and in the historical and social theories that inform the approaches of each particular scholar, school of thought, and orientation.

1. Etymology

Etymologically, the adjective 'feudal' and the noun 'feudalism' are derived from the Germanic *feod* and the Latin *feudum*, which originally meant 'cattle, money, property.' During the course of the eleventh century, *feudum* became a technical legal term for the fief with which the lord invests a vassal under specific conditions. Being provided with this fief obliged the vassal to serve his lord militarily. Feudal law, the sum of all recorded feudal legal norms, was known as *ius*

feudale in medieval and early modern legal terminology and as 'féodalité' beginning in the seventeenth century. During the course of the eighteenth century, structural changes of a social and political nature brought about a fundamental shift in the meaning of the term 'féodalité.' Since then, the words 'féodal' and 'féodalité' have been used to designate not only the legally vested relationship of dependence between a lord and his vassal, but also an entire societal condition: the political and social condition of the *ancien régime*. A legal term was thus transformed into a kind of political catchword in order to call into question the legitimacy of a traditional political and social order. Thus associated with socio-critical connotations, 'feudalism' entered into the languages of Europe as both a word and a concept.

2. *Historical Controversies about Feudalism*

In his *Esprit des Lois* (1748) Montesquieu equates the '*lois féodales*' with the breakdown of public power and the disintegration of the central authority of the French state into territorial centres of power. Voltaire (1694–1778) denounces the '*system féodale*' as a system of sovereignty in which public power was divided among 'countless little tyrants.' In *Qu'est-ce le tiers état*, a text appearing in 1789, Abbé Sieyès attacks the nobility and the church as 'agents of feudality' whose privileges were incompatible with the ideal of equality under the law. In his view, freeing France from the 'ruins of feudality' was the most important task at hand. Influenced by persistent and widespread unrest among the peasants, the national assembly decreed the 'abolition of feudalism' (*abolition de la féodalité/du régime féodal*) on August 11, 1789. In the writings of Karl Theodor Welcker (1790–1869), one of the leading thinkers of Southern German liberalism, the feudal system is a glaring counterpart to the liberal order of the early Germanic period with its public cooperations. In keeping with the polemical use of the word '*féodalité*' in revolutionary France, Welcker employs phrases like 'feudal despotism' (*tyrannie féodale*) and 'feudal anarchy' (*anarchie féodale*) in order to characterize feudalism as a form of domination which, based on the violent usurpation of power, was blatantly inconsistent with the early Germanic ideal of popular self-determination ('*Volksfreiheit*').

3. *'Feudal Rule' and 'Feudal Mode of Production' in the Work of Georg Wilhelm Friedrich Hegel and Karl Marx*

In Hegel's philosophy of history, the 'feudal system' constitutes a step in the historical process which brought about the 'progress from feudal government to monarchy.' Hegel defines feudal government as a 'system of private dependence and private commitment' in which 'the general injustice' and 'the general rightlessness' assumed an institutionalized form. Under the conditions of feudal rule, the lawful power of the state was superseded by a more personalized form of power which individuals held on the basis of their 'private property.' According to Hegel, 'the wilful brutality of particularistic law' prevented the formation of an administrative and controlling authority that could act for the common good to protect the freedom and equality under the law of its civil subjects. In Hegel's view, only the legal government of a monarch can adequately guarantee freedom. 'Feudal rule is a polyarchy; it consists of nothing but masters and slaves. In a monarchy, on the other hand, there is only one master and no slaves because under a monarch, the bonds of slavery are broken and justice and law prevail. True freedom issues from a monarchy.'

In Marx's view of history, 'feudal society' with its typical 'feudal relations of property' and its 'feudal mode of production' constituted a 'progressive economic form of society that produced a new class, the bourgeoisie, which broke down the 'feudal barriers to production.' Marx does not, however, work out an explicit theory of the origins, structure, and transformation of feudalism in detail; he never systematically treats the dynamics of the feudal mode of production. For him, terms like the Middle Ages, feudalism, and feudality are interchangeable. Marx links the Middle Ages and feudalism to the idea of a premodern society that created the necessary conditions for the rise of a capitalistic society which eventually overcame and dissolved feudalism itself. In the 'Communist Manifesto,' he writes that 'the means of production and exchange on whose foundation the bourgeoisie built itself, were generated in feudal society.' 'Modern bourgeois society that has sprouted from the ruins of feudal society has not done away with class antagonisms. It has but established new classes, new conditions of oppression, new forms of struggle in place of the old ones.' Marx is interested in the Middle Ages as the epoch that preceeded the 'epoch of the bourgeoisie.' He is concerned more with the political state of medieval society than with an analysis of the feudal mode of production. Like the liberal critics of prerevolutionary corporate society, Marx assumes that in the feudal society of the Middle Ages, 'the supreme command in war and justice is based on landed property.' In his opinion, past and present share common structures with regard to this connection between power and property, for in both feudal society and a bourgeois society, the form of ownership determines the form of rule. The 'supreme command in industry is based on the ownership of capital,' while rule in feudal society was based on the control of land. However, the difference between feudal and bourgeois power relations was that in the Middle Ages, control of landed property determined

both social standing and legal rights to rule; the 'constitution of private property' was also the 'political constitution.' In feudal society, 'property or membership to a privileged family or the nature of one's occupation … were essential elements of political life in the form of landlordism, class, and corporation.' Control over land formed the basis for rule over people. Noble birth was a kind of guarantee for the privileges of a lord. Private corporations such as guilds administered justice over their members, a right which the modern constitutional state reserves exclusively for itself. Narrow 'personalized relations of dependence' characterized the 'social conditions of material production.' Bond men were personally dependent on a small class of landholding nobles; as serfs, they were obliged to pay dues to their feudal lord and provide him with certain services.

Interpreting history from a Eurocentric perspective, Marx equates classical antiquity with a slave-holding society, the Middle Ages with feudalism, and the present times with capitalism. This classification of history reflects Enlightenment philosophy's belief in progress and combines it with the view of history held by postrevolutionary historians and social scientists. In their view, '*féodalité*' is a term to describe both an epoch and a society: in the linear, evolutionary course of world history, '*féodalité*' or more exactly the '*système théologique et féodal*' was a link between the society of ancient slaveholders and the '*système industriel et scientifique*' that produced the French Revolution.

4. *Feudalism as Conceived by Historians and Social Scientists: Max Weber, Otto Hintze, and Marc Bloch*

Weber's concept of feudalism has nothing in common with revolutionary and liberal authors' view of feudalism as a system of anarchy, injustice, and violence. Weber integrates socio-political, legal, economic, and spiritual-religious structural elements into a single model. Weber defines feudalism as a traditional system of governance that combines economic, social, and cultural factors. The idea that a feudal society is determined exclusively by socioeconomic characteristics—its 'feudal mode of production'—is quite alien to him. In his view, feudalism is not a type of economic management, but rather a type of domination that requires an economic basis. Weber argues that as a form of organizing and exercising domination, feudalism is a political phenomenon. The structural principle of feudal power relations is the fief. Feudal relations determine the structure and form of the political order, a hierarchically graded feudal pyramid that consists of feudal lords, vassals, and their subvassals. Weber defines *enfeoffment* as 'any investment of rights, especially the right to use land or to exercise territorial domination in exchange for military or administrative services.' The polity was not held together by—to his mind—the integrative forces of a bureaucratically controlled state, but drew on the personal relationship of loyalty between lord and vassal for stability and duration. During investiture this relationship took on the character of a legally binding contract in a symbolic act which included commendation (*homagium*) and an oath of fealty (*fides*). In the absence of a centralized bureaucracy, 'the personal feudal bond of allegiance and the feudal code of honor serve as the glue of political cohesion.' Charlemagne's reform of the army led to a feudalization of military affairs. Knights were provided with economically useful fiefs which obliged them to provide military service. A cavalry composed of heavily armed feudatories replaced an army that once consisted of freeborn peasants. The feudalization of administrative structures transformed the early and high medieval realm into a system of ruled territories (duchies, ecclesiastical and secular principalities, margravates, earldoms) whose bearers were bound to the king by feudal law as tenants in chief. The feudal system was of fundamental significance for the political-legal organization of the kingdom. As a political principle of organization that divided the power of lords, feudalism laid the foundations for a decentralization of political power. Contractually stipulated division of power caused rivalries and conflicts between the central authority of the king and local vassals. By banding together to form corporations, feudatories transformed the feudal union into a corporative state.

Weber also points out that the manor constituted the economic basis of feudalism. The manor was 'the standard object of feudal tenure.' The power of nobles and princes was founded on ownership of land, human beings, and livestock. The conditions under which feudalism emerged and its function as a system of rule lent it a social character. The social differentiation between peasants and trained warriors grew out of the emergence of an armored cavalry. Membership to the noble class of lords was necessary in order to exercise power in the feudal system. Ownership of a manor which yielded rent entitled one to a 'lordly mode of existence' and enabled one to live according to one's social standing. Honor was the most important norm influencing conduct. Work-free rental income was indispensable for a life 'from which the social forms of prince and nobleman originate.' 'Feudalism, especially in the form of free vassalage and, more importantly, in the feudal system appeals to "honour" and personal, freely rendered and sustained "allegiance" as constitutive motives of conduct.' Honor and allegiance, the embodiment of the 'feudal ethics' which are characteristic of Western feudalism, constitute the 'extreme opposite of all objective business relations.' Feudalism produced and cultivated an ethics of the nobility in the form of heroic asceticism, heroic honor, and hero-worship; these ethics contradict the objectiveness of business relations and operations. The idea of the fief

determined the various forms of 'feudal incorporation,' 'all sorts of social relations: from one's relationship to the Saviour to that to his lover.'

To continue with Weber's concept of feudalism, in the long term, 'feudal structures of power with their clearly defined rights and duties' contributed to the 'stability of the economy.' The 'constancy of the legal system' contributed to the 'emergence of a rational capitalistic system and fostered its penetration into the gaps and seams of the feudal system.' According to Weber, the feudal system also had the rationalizing effect of molding society into 'an at least relatively constitutional formation.' The feudal relationship between lord and vassal created a 'constitutional state' which did not arise on the basis of a legal system, but rather on the basis of subjective rights agreed upon by free partners. By strengthening the personal rights of the individual feudatory, a system of social relations emerged in which contract and consensus constituted the predominant forms of societal integration.

In *Character and Spread of Feudalism* (1929) Hintze follows in the steps of Weber's methodology to 'describe the ideal type that underlies the concept of feudalism.' In his effort to precisely define feudalism as a socio-historic category, Hintze follows the rules of modeling ideal types to go beyond the hermeneutical interpretation of historical texts and concepts. Hintze argues that feudalism is more than 'a self-contained and coherent system of institutions' that constitute the 'feudal state.' His concept of feudalism is thus intended to understand and include other phenomena that are not a part of the feudal system in a strict sense. He therefore tries to link political, military, and socio-economic structural elements in his definition of feudalism. He understands feudalism not only as a principle of political organization, but also as 'a principle of military, economic, and social constitution.' According to Hintze, feudal structures had an effect on the development of sovereignty and social stratification in three ways: first, in the 'selection of a highly trained and loyal class of warriors who, bound by private contract, assumed a privileged position; second, 'in the development of a manorial, rural economic form which provided for the privileged class of warriors'; and third, in the 'elevation of this military nobility to the status of local lords who exercised considerable influence or who, because of their arbitrary way of ruling, assumed a privileged position' within the loosely structured feudal state.

Hintze distinguishes between three main phases in the development of feudalism in the history of the Romance-Germanic peoples: 'first, the era of early feudalism in which the military factor predominated—it extends roughly until the end of the twelfth century; second, the era of high feudalism in which the military nobility attained the height of its political influence, either by attaining relative independence from their feudal overlord as was the case under the arbitrary rule of princes in Germany, or in the form of a corporative federation within a more solid feudal monarchy as it occurred in England when barons joined together in parliament against the crown—it extends roughly up to the sixteenth or seventeenth century; and third, the era of late feudalism in which the main interest of the nobility was in preserving and exploiting its socio-economic position as lords of manors and estate owners—it extends up until the era of the French revolution and up until the dissolution of traditional relationships in rural areas in the nineteenth century.' Hintze sets up guidelines for applying the term feudalism to non-European states, societies, and cultures. In his view, the term feudalism is, strictly speaking, only applicable to a society in which 'military, socio-economic and political feudalism are fully developed.' Working under this premise, Hintze argues that 'feudalism in its fullest sense can only be found—albeit with considerable deviations and particularities—in Russia, the Islamic nations and Japan.' Egypt, the ancient Orient, the Roman Empire, India, and China only exhibit initial signs of a feudalism that extended equally over the political order, the manorial, agricultural economic form and a military nobility which, supported by work-free income, was bound in allegiance to the ruler. In view of the poorly developed possibilities for communication, large kingdoms required a decentralized administration and court system which were meant to function on the basis of the feudal bond of allegiance between ruler and noble vassals.

Like Otto Hintze, Marc Bloch also views feudalism as a social type. His book *La société féodale* is considered to be 'a model of analysis of social structure,' a 'paradigm of a *histoire totale*.' Bloch does not limit himself to describing an ideal type which other scholars can use as a starting point for their own research. He also analyzes and describes European social history from the mid-ninth to the mid-thirteenth centuries, never losing sight of feudal society as the main connecting thread throughout this era. He sees vassalage and the manorial system as 'leading characteristics' of feudalism as a type of society. He designates the manor as 'the fundamental unit of the feudal regime.' According to Bloch's understanding, feudalism encompasses more than just those forms of social interaction that result from vassalage and the feudal and manorial system. He sees feudalism as a society that should be reconstructed as a complete interactive system. He therefore examines the basic structural conditions—including transportation, trade, population, the system of communication, and the monetary system—that constitute the environment of society. He is concerned with describing the '*atmosphère mentale*,' the 'culture of the feudal period' that influenced the social behavior and societal life of the people by means of its 'forms of feeling and thinking' including its literature (heroic song and epic), its education and scholarship, its attitudes toward nature, time and history, and its matters of

faith and religious practices. Bloch identifies 'fear of punishment in hell' as 'one of the strongest social powers of this time.' The 'kingdom of the feudal period' had a 'holy character.'

Personal ties and relationships of dependence had a great influence on the societal interactions which Bloch designates as feudalism. 'A network of bonds of dependence that ran through all levels of human society from top to bottom gave the culture and society of European feudalism its distinctive character.' Forms of personal dependence that existed within families and clans were based on blood ties. Under vassalage and the feudal system, personal bonds were forged through freely made agreements, thus initiating a relationship in which services and duties were interchanged. Legal tradition, voluntary devotion, or involuntary coercion produced legal, economic, and social inequality on manorial estates between noble lords, serfs, and tenant farmers. Central elements of Bloch's concept of feudalism include vassalic dependence, which bound vassals to perform military services for a lord, and manorial dependence, under which villeins and serfs (*coloni*, *villani*, *mancipia*, *servi*) were made to pay their landlords dues and tributes and perform statute labor. Like Max Weber, Bloch sees vassalage as patterns of behavior and relations which shaped relationships within the family, between lovers, and between the believer and God.

Because the members of societies generally designated as 'feudal' were subdivided into groups which were stratified according to profession, prestige, wealth, and power, Bloch examines social divisions as well as legal and economic inequality between the nobility, burghers and peasants, high and low clerics. He views bearers of 'public power' in the 'feudal epoch' (emperors and kings, princes and prince bishops, dukes and earls) not as representatives of the so-called 'feudal state,' a system of public rule distinct from society, but rather as bearers of 'traditional powers' in terms of their social bonds, limitations, and functions. It was, after all, their duty in a legally and socially 'divided society to maintain certain principles of order and unity.'

5. *History and the Concept of Feudalism: Marxist, Historical, and Social Science Perspectives*

Marx and Lenin are only interested in feudalism as a societal formation that constituted a preliminary stage of capitalism. Influenced by Marx, Lenin, and Stalin, socialist historians and social scientists have developed a theory of feudal society that not only explains its origins and characteristics, but also takes the change in conditions of production as a starting point to develop a criteria for the division of history into epochs. According to the prevailing theory of Marxism, feudal conditions of productions and feudal social conditions gradually evolved between the end of the fifth and the middle of the ninth centuries. The 'epoch of developed feudalism' lasted from the middle of the eleventh until the end of the fifteenth century. 'It was a time in which feudal conditions of production and exploitation in Western and Central Europe prevailed and influenced the economy and society, politics and class struggle, culture and ideology in spite of critical disturbances during the late medieval centuries and variations in the degree reached by capitalistic conditions of production in the most advanced European regions.' Feudalism was able to endure over a span of several epochs because the 'early bourgeois revolution' (peasants' war, Reformation) miscarried. As a 'late feudal order' characterized by 'unrestricted rule of the nobility,' 'absolute feudal monarchy,' 'fragmentation of the feudal state,' the continuation of feudal conditions of production, and the 'development of capitalistic forms of production,' feudalism continued until the end of the eighteenth century, when it was overcome and dissolved during the French Revolution by 'bourgeois emancipation.'

The question of the 'essence' and 'genesis of feudalism' grew out of an interest in the structures of the precapitalistic form of society. The following characteristics of premodern feudal society arise out of a comparison between the two systems feudalism and capitalism: low level of technique, dominance of a barter economy, and feudal and hence divided ownership of land. Strictly speaking, the owner of the land was the feudal lord. He had the absolute ownership (*dominium directum*) of the parcels of land which were lent to the peasant. The peasant was granted the right to use the land by way of a loan (*dominium utile*). The instruments of production necessary to work the land belonged to him. The villein and serf owed their lord 'surplus work' (*corvée*) and 'surplus products' (payment in kind, dues) for the use of the land. The 'extra-economic compulsion' which issued from the lord lent the exploitative feudal conditions duration and stability. Extra-economic compulsion refers to the connection between property and sovereign rights. Because the lord of the manor also had the right of jurisdiction over the peasants to whom he granted use of his land, control of jurisdiction was a constitutive element of the exploitative conditions of production.

The majority of non-Marxist historians exclude forms of manorial dependence from their concept of feudalism. They use feudalism and feudality as synonyms for the feudal system and feudal law. In their view, a definition of feudalism that includes all forms of dependence including that of the peasant on the manor is an imprecise and thus unsuitable term for ordering and explaining this historical phenomenon. They reject a definition of feudalism that is more broadly conceived to include all areas of political and social life because they say that it cannot be verified by the sources themselves. They argue that the term feudalism loses its heuristic value if the clear-cut

reference to the fief of a vassal who is obliged to perform military service is lost; in their view, extending the term further is linguistically and objectively ungrounded. Faced with the dilemma of defining feudalism either as a form of society or as the feudal system, François Louis Ganshof chooses the latter. His justification for this choice is symptomatic for all historians who, by means of an exact interpretation of the sources, claim to be able to determine 'how it actually was.'

Ganshof defines 'the feudal system as a collection of institutions,' 'that create and regulate two kinds of bonds between one free-born man, called vassal and another, called lord or seignior: the 'vassal' is obliged to obey and serve—chiefly militarily—the 'lord', and the 'lord' is obliged to provide the 'vassal' with protection and a livelihood. The lord normally fulfilled his obligation to provide the latter by granting the use of land known as the 'fief.'

It has often been pointed out that conceiving of the feudal state as a counterpart to a bureaucracy leads to a 'one-sided exaggeration of domination' and does not adequately take 'the communal, co-operative, and corporative elements in the village, the city, and in the country' into account (Otto Brunner). Research in social and constitutional history that focuses on the role of the community in towns and in the country confirms this skepticism. It is argued that there is 'no logical connection' between 'feudalism,' which overemphasizes the rule of the nobility, and 'communalism,' which sees the community as an organ of civic and peasant self-determination (Peter Blickle). Hence by seeing the peasant as a mere object of domination and exploitation, the 'feudal' point of view ignores historical records and documents to overlook the fact that the peasants themselves were also active subjects in the long development from serfdom into freedom.

Historians who examine changes in social structures in order to explain the development of modern capitalism (Perry Anderson, Rodney Hilton, Maurice Dobb, Eric Hobsbawm, Hans Ulrich Wehler) use feudalism as a term of demarcation. Free of dogmatic fixation, the work of Ludolf Kuchenbuch, Guy Bois, and Rodney Hilton examines the sources closely and focuses directly on its subject in order to empirically substantiate and better clarify the term 'feudal mode of production.' *Feudalism and History* (1956), an anthology edited by Rushton Coulborn, attempts to turn the definition of feudalism as a 'territorial system of governance' into a category of world history that could be used for international comparisons. Drawing on this definition of feudalism, the volume comes to the conclusion that only Western Europe, Japan, and early Chou-China display evidence for the existence of a feudal social system. English and French historians make little attempt to justify their use of the terms feudalism, '*féodalité*,' and '*féodalism*' to describe structures of medieval and early modern agrarian societies. This usage is part of the established terminology of the field. The intense debate about feudalism in the former GDR ended after the unification of Germany. The *Jahrbuch für die Geschichte des Feudalismus* (*Annual for the History of Feudalism*) ceased publication in 1990. German historians avoid the term feudalism as an ideologically loaded construct. They consider the term manor ('*Grundherrschaft*') on the other hand, as indispensable. However, the term manor also originated as a politically tendentious term; at the end of the eighteenth and beginning of the nineteenth centuries, enlightened and liberal authors who saw the manor as an illegitimate form of aristocratic despotism used the term to stigmatize the connection between domination and ownership of property which made the peasant dependent on noble lords and removed land from circulation on the free market.

6. *Feudalism as a Type of Society*

Adherents of a liberal concept of the town which originated in the nineteenth and early twentieth centuries object to usage of the term 'feudalism' or 'feudality' to label medieval society as a whole because, they argue, the political order and the socioeconomic life of the town—municipalized since the twelfth century—can hardly be characterized as 'feudal.' As Otto von Gierke argues in *Rechtsgeschichte der deutschen Genossenschaft* (1868), 'free union,' a new form of incorporation that formed the basis for the medieval town, was destined to 'break the feudal system.' Max Weber recognizes the medieval town as one of the factors that 'broke the restraints of the outer and inner authorities of the feudal era.' According to Marc Bloch, social relations within the town were inconsistent with the 'feudal type.' Because of its 'liberties,' its 'economic expansion,' and 'collective independence,' medieval towns remained 'foreign bodies in feudal society.' As a community in which personally free citizens with equal rights vowed their mutual solidarity by oath, the upward bound aspiring citizens became a 'destructive element for the feudal system.' Experts in the fields of law and the social sciences have recently emphasized that assuming the form of a commune, the medieval town developed a 'principle contrary to the hierarchically structured forms of domination in feudal society' (Gerhard Dilcher). Further, the medieval town proved itself to be 'a point of crystallisation for an alternative principle of legitimacy,' for a 'legitimacy by virtue of statute, not tradition and piety,' and for a 'republican, not a feudal or estate-based model of governance' (Wolfgang Schluchter). Otto Brunner, on the other hand, contends that the medieval municipality developed into 'a special configuration in the world of European feudalism.' He argues that the bond of fealty that was characteristic for the European system of governance,

'so-called feudalism,' also determined the relationship between the town lord and the municipality. In the Marxist interpretation of history, which defines the Middle Ages as 'age of feudalism,' medieval and early modern towns are referred to as 'economic and social units of feudal society.'

In order to compare and distinguish societies with different formations and structures, it is useful to have a relatively wide range of general definitions. The term feudalism is just this kind of general definition: it is a tool of historical study in the attempt to understand and describe premodern society in its entirety. This conclusion cannot be called into question or disproved by the fact that the medieval town had its own special political, social, and economic character, and therefore contributed to the transformation of the so-called feudal society into a civil society. The towns of Old Europe were not sovereign polities; they were bound up in the constitution of kingdoms and principalities. Max Weber refers to the 'feudal town lords' to whom the town council had to pledge loyalty and obedience each time the crown changed hands or new lords came to power. Weber also points out the 'feudal character' that marked the lives of urban patricians and oligarchs and makes reference to the existence of 'feudal unions' within the towns of Northern Italy and the development of a clientele at the hands of the nobility in the towns of Southern Europe. Patrician families claimed to possess the qualities of the nobility and therefore wanted to be regarded as nobility and receive the same status. The councils of late medieval and early modern towns often saw themselves as an authority appointed by God and worthy of the loyalty and obedience of the citizens. The council saw its rule as a system of command and obedience and derived its position not from a mandate of the citizens but rather from its *summa potestas*. When towns tried to increase their power by setting up territories of their own, they acted as feudal lords themselves by exercising dominion over territory and people. The different quality of the medieval and early modern town did not preclude structural elements of a feudal provenance and character. Because empirical evidence indicates that the borders between systems are permeable, it is probably justified to include the town as a special configuration in a more open concept of feudalism.

The concept of feudalism is useful as a kind of comprehensive model which draws connections between the various constituent areas of reality in a given society (politics, economics, law, social behavior, religion, and culture) in order to show and explain how these areas function independently as well as how they interplay with each other as elements in a complete social system. If the concept of feudalism is to be used in this way, then it is necessary to expand on the meaning it has acquired during its genesis with new contents. Only by formulating the concept of feudalism anew is it possible to free it from the ideological burdens of the eighteenth and nineteenth centuries and make it useful for our current understanding of the past.

The following features can be considered as typical structural characteristics of a feudal society: (a) Understood as political feudalism, the feudal system—in view of the limited communication possibilities—provides regions in the process of expansion which are difficult to organize politically: access to the dispensation of justice, administrative services, and military protection. Organized according to feudal law, land, administrative posts, and legal rights can be equally invested and bequeathed; the result is a system of governance determined by the interplay between manorial and cooperative powers at the level of both the kingdom and the regions. (b) Political feudalism also refers to fragmentation of executive and judicial powers. During the medieval and early modern eras, a monopoly on state power in the hands of a king was—in most cases—unknown. The monarchy and the nobility divided public power among themselves. Noble lords ruled over land and people by virtue of both their own right and royal investment. Competition for power, a consequence of divided and fragmented executive power, allowed for the development of a cooperative social structure that, in the long run, contributed to the disintegration of feudal power and social conditions. The medieval town emerged only because feudalism did not represent a fully realized structural principle with regard to politics and society. Processes of social differentiation could therefore take hold in the seams and gaps of the system, thus forming a basis for the development of the town as a new kind of settlement and social form. (c) In the chiefly agrarian structured societies of early and high medieval Central and Western Europe, the territorial sovereignty practiced by kings, noble lords, and municipal cooperatives was not one of many principles structuring society, but rather the fundamental way of ordering life and was of importance for the whole of society. (d) Decisive for the organization of social life were relations of personal dependence: the vassal was dependent on his feudal lord, the client on his patron, the serf and tenant on their ecclesiastical and secular landlord, the apprentice on this master, the craftsman, homeworker, and wage earner on his manufacturer. (e) The exercise of dominion as well as the extent and range of political participation were closely connected with one's social position. The distribution of political power and social organization were closely linked. This relationship was not only an expression of legal and social inequality, but also of the absence of the division between public and private law. The medieval *societas civilis* was constituted neither as a system of social and economic needs which was separate from the state nor as a society of citizens with equal rights. (f) Medieval and early modern societies were devoid of the inherent laws of secular orders and institutions. Christianity constituted an interpretative and formative power which encom-

passed the whole of society and pervaded and regulated all realms of life. The connective and binding power of the political order was rooted in ideas and principles of the Christian religion which attribute justice and power to God. Not only did manifestations of social inequality require legitimacy in religion, but also justice, the economy, science and art, war, and peace. The social character of religion encompassed more than just works of Christian *caritas*; it also constituted a *vinculum societatis*, a kind of bond that integrated groups and unions through common convictions concerning right and wrong.

Defining feudalism involves the difficult task of comprehending and describing both the unifying and the differentiating aspects of the premodern *societas civilis*, which became thinkable as a social unity through the medieval reception of Aristotle. However, because the urban and rural societies of the Middle Ages and the Early Modern underwent various structural changes, it is crucial from a methodological standpoint to use the term feudalism in a way that makes the difference between ideal-typical model and concrete historical reality clear.

See also: Bloch, Mark Léopold Benjamin (1886–1944); Hintze, Otto (1861–1940); Honor and Shame; Marx, Karl (1818–89); Middle Ages, The; Weber, Max (1864–1920)

Bibliography

Anderson P 1974 *Passages from Antiquity to Feudalism*. Atlantic Highlands: Humanities Press, London

Bader K S, Dilcher G 1999 *Deutsche Rechtsgeschichte. Land und Stadt-Bürger und Bauer im Alten Europa*. Springer, Berlin

Bartel H (ed.) 1982 *Deutsche Geschichte, Vol. 1: Von den Anfängen bis zur Ausbildung des Feudalismus Mitte des 11.* Jahrhunderts, Köln

Bartel H 1986 *Deutsche Geschichte, Vol. 2: Die entfaltete Feudalgesellschaft von der Mitte des 11. bis zu den siebziger Jahren des 15*. Jahrhunderts Deutscher Verlag der Wissenchaffer, Berlin

Blickle P 2000 *Kommunalismus. Skizzen einer gesellschaftlichen Organisationsform*. R. Oldenbourg, München, Vols. 1–2

Bloch M1930 Artikel feudalism. In: Seligman E R A, Johnson A (eds) *Encyclopaedia of the Social Sciences*. New York (Reprint 1950), Vol. 6, pp. 203–10

Bloch M 1939 *La société féodale. La formation de liens de dépendance*. Albin Michel, Paris

Bloch M 1940 *La société féodale. Les classes et le gouvernement des hommes*. Albin Michel, Paris

Bloch M 1999 *Die Feudalgesellschaft*. Klett-Gotta, Stuttgart, Germany

Breuer St 1988 Der okzidentale Feudalismus in Max Webers Gesellschaftsgeschichte. In: Schluchter W (ed.) *Max Webers Sicht des okzidentalen Christentums. Interpretation und Kritik*. Suhrkamp, Frankfurt a. M. pp. 437–75

Brown R A 1973 *Origins of English Feudalism*. Allen & Unwin, London

Brown E A R 1974 The tyranny of a construct: Feudalism and historians of Medieval Europe. *The American Historical Review* **79**: 1063–88

Brunner O 1968 'Feudalismus'. Ein Beitrag zur Begriffsgeschichte. In: Brunner O (ed.) *Neue Wege der Verfassungs- und Sozialgeschichte*. Vandenhoeck & Ruprecht, Göttingen, pp. 128–59

Brunner O 1975 Artikel 'Feudalismus'. In: Brunner O, Conze W, Koselleck R (eds.) *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*. Klett-Gotta, Stuttgart, Germany Vol. 2, pp. 337–50

Byres T J, Mukhia H (eds.) 1985 *Feudalism and Non-European Societies*. F. Cass, London

Demski R 1996 *Adel und Lübeck. Studien zum Verhältnis zwischen adliger und bürgerlicher Kultur im 13. und 14.* Jahrhundert, Lang Frankfurt a.M., Germany

Dilcher G 1993 Kommune und Bürgerschaft als politische Idee der mittelalterlichen Stadt. In: Fetscher I, Münkler H (eds.) *Pipers Handbuch der politischen Ideen*. Piper, München, Germany Vol. 2

Duby G 1978 *Les trois ordres ou l'imaginaire du féodalisme*. Gallimard, Paris

Duus P 1969 *Feudalism in Japan*. Knopf, New York

Fourquin G 1976 *Lordship and Feudalism in the Middle Ages*. Publisher, New York

Heuvel G v d 1988 Artikel 'Féodalité, Féodal'. In: Reichardt R, Schmitt E (eds.) *Handbuch politisch-sozialer Grundbegriffe in Frankreich 1680–1820*. Oldenbourg, München, H.10, pp. 1–48

Hilton R H 1985 *Class Conflict and the Crisis of Feudalism. Essays in Medieval History*. Hambledon Press, London

Holton R H 1985 *The Transition from Feudalism to Capitalism*. St Martin's Press, New York

Hintze O 1929 Wesen und Verbreitung des Feudalismus. In: Oestreich G (ed.) *Otto Hintze, Staat und Verfassung. Gesammelte Abhandlungen zur allgemeinen Verfassungsgeschichte*. Vandenhoeck & Ruprecht, Göttingen, Germany pp. 84–119

Krieser H 1984 *Die Abschaffung des 'Feudalismus' in der Französischen Revolution. Revolutionärer Begriff und begriffene Realität in der Geschichtsschreibung Frankreichs (1815–1914)*. Ulstein, Frankfurt a.M., Germany

Kuchenbuch L 1977 *Feudalismus-Materialien zur Theorie und Geschichte*. Ulstein, Frankfurt a.M., Germany

Kuczynski J 1992 *Asche für Phönix. Aufstieg, Untergang und Wiederkehr neuer Gesellschaftsordnungen. Eine vergleichende Studie zu Feudalismus, Kapitalismus und 'Realem Sozialismus'*. Papy Rossa-Verlag, Köln, Germany

Martin J E 1983 *Feudalism to Capitalism: Peasant and Landlord in English Agrarian Development*. MacMillan, London

Milson St F Ch 1976 *The Legal Frame Work of English Feudalism*. S. F. G. Milson, Cambridge, UK

Mitteis H 1958 *Lehnrecht und Staatsgewalt. Untersuchungen zur mittelalterlichen Verfassungsgeschichte*. Böhlaus NachFolger, Weimar, Germany

Mitteis H 1944 *Der Staat des hohen Mittelalters. Grundlinien einer vergleichenden Verfassungsgeschichte des Lehnzeitalters*. Böhlaus NachFolger, Weimar, Germany

Müller-Mertens E (ed.) 1985 *Feudalismus—Entstehung und Wesen*. Akademie-Verlag, Berlin, Germany

Neubauer H 1968 Artikel 'Feudalismus'. In: Sowjetsystem und demokratische Gesellschaft. *Eine vergleichende Enzyklopädie*. Herder Freiburg, Germany, pp. 477–90

Poggi G 1988 Max Webers Begriff des okzidentalen Feudalismus. In: Schluchter W (ed.) *Max Webers Sicht des okzidentalen Christentums. Interpretation und Kritik*. Suhrkamp, Frankfurt a.M., Germany pp. 476–97

Rader T 1971 *The Economics of Feudalism*. Gordon & Breach, New York
Roth P 1863 *Feudalität und Unterthanverband*. Böhlau, Weimar, Germany
Sander P 1906 *Feudalstaat und Bürgerliche Verfassung. Ein Versuch über das Grundproblem der deutschen Verfassungsgeschichte*. Bath, Berlin, Germany
Schluchter W 1979 *Die Entwicklung des okzidentalen Rationalismus. Eine Analyse von Max Webers Gesellschaftsgeschichte*. Mohr (Paul Siebeck), Tübingen, Germany
Schreiner K 1983 'Grundherrschaft'. Entstehung und Bedeutungswandel eines geschichtswissenschaftlichen Ordnungs- und Erklärungsbegriffs. In: Patze H (ed.) *Die Grundherrschaft im Mittelalter*. Thorbecke, Sigmaringen, Germany pp. 11–74
Sidney P 1961 *Feudalism and Liberty: Articles and Addresses*. Johans Hopkins Press, Baltimore
Stephenson C 1948 *Medieval Feudalism*. Cornell University Press, Ithaca, NY
Strayer J R 1965 *Feudalism*. Van Nostrand, Princeton
Sweezy P u. a. 1976 *The Transition from Feudalism to Capitalism*. Verso, London
Sydney H 1921/1969 *The Fall of Feudalism in France*. Barnes & Noble, New York
van Horst Bartel 1969 *Sachwörterbuch der Geschichte Deutschlands und der deutschen Arbeiterbewegung* Dietz, Berlin, Germany Vol. 1, pp. 582–6
Von Gierke O 1868/1954 *Das Deutsche Genossenschaftsrecht*. Graz Akademische Druck-und Verlangs-Anstadt, Graz, Austria Vol. 1
Ward J O 1985 *Feudalism: Comparative Studies*. Sydney Association for Social Studies in Society & Culture, Sydney
Weber M 1921/1972 *Wirtschaft und Gesellschaft. Grundriß der verstehenden Soziologie*, 5th edn. Mohr (Paul Siebeck), Tübingen
Wehler H-U 1987 *Deutsche Gesellschaftsgeschichte, Vol.1: Vom Feudalismus des Alten Reichs bis zur Defensiven Modernisierung der Reformära 1700–1815*. Beck, München, Germany
Wunder H 1974 Der Feudalismus-Begriff. Überlegungen zu Möglichkeiten der historischen Begriffsbildung. In: Wunder H (ed.) *Feudalismus*. Nymphenburger Verlag, München, Germany
Wunder H 1989 Artikel 'Feudalismus'. In: Sautier R H (ed.) Lexikon des Mittelalters. Artemis, München, Germany, Vol. 4, pp. 411–15

K. Schreiner

Fiducial and Structural Statistical Inference

Fiducial inference is a statistical approach to interval estimation first advocated by R. A. Fisher as an alternative to the then dominant method of inverse probability, i.e., using Bayes' Theorem. Considerable effort has gone into formalizing Fisher's notions using such concepts as statistical invariance and pivotal quantities. This entry describes elements of the fiducial approach and relates them to other currently more widely used statistical approaches to inference. Section 1 introduces some basic inferential ideas via a simple example.

1. A Simple Example

Consider a variable y that is directly available or has arisen by some preliminary reduction process and suppose that y measures θ in an unbiased manner and has error that is normal with known variance σ_0^2. Then we can say for example that $P(y \leqslant \theta + 1.64\sigma_0; \theta) = 95$ percent.

Parenthetically, we note that the preliminary reduction could have occurred as part of the underlying investigation or as part of some subsequent simplification of the statistical model by one of the common statistical reduction methods, sufficiency or conditionality. If the reduction is to a sufficient statistic, then the conditional distribution describing possible antecedent data has no dependence on the parameter and the model for the sufficient statistic is used in place of the original model. If the reduction is by conditionality then there is typically an ancillary variable with a distribution free of the parameter and the given model for the possible original data is replaced by the conditional model given the ancillary (supportive) variable; see also *Statistical Sufficiency*.

With a data value y^0 the fiducial methodology would take the above probability expression and substitute the value y^0 and then treat θ as the variable for the probability statement, thus giving a 95 percent fiducial probability statement $P(\theta \geqslant y^0 - 1.64\sigma_0; y^0) = 95$ percent. The structural approach would consider the normal $(0; \sigma_0^2)$ distribution for the error $e = y - \theta$ and might record for example the probability $P(e \leqslant 1.64\sigma_0) = 95$ percent; then with data value y^0 the probability statement would be applied to the error $e = y^0 - \theta$ giving the structural probability $P(y^0 - \theta \leqslant 1.64\sigma_0;\ y^0) = 95$ percent or equivalently $P(\theta \geqslant y^0 - 1.64\sigma_0; y^0) = 95$ percent. In a somewhat related manner the Bayesian methodology might use a uniform prior $cd\theta$ and obtain a posterior distribution for θ that would have $P(\theta \geqslant y^0 - 1.64\sigma_0; y^0) = 95$ percent. For this simple example the three methods give the same result at the 95 percent level and also at other levels, thus saying essentially that with data value y^0 the probability distribution describing the unknown θ is normal $(y^0; \sigma_0^2)$.

With more complicated models the results from the three methodologies can differ and philosophical arguments concerning substance and relative merits arise. However, for one straightforward generalization the methods remain in agreement: the normal distribution can be replaced by some alternative distribution form; this is discussed in some detail in Sect. 2.

2. Fiducial Probability

Fisher (1922, 1925; see also *Fisher, Ronald A (1890–1962)*) had already introduced most of the

fundamental concepts of statistical theory, such as sufficiency, likelihood, efficiency, exhaustiveness (minimal sufficiency), when he chose (Fisher 1930) to address directly the aspiration mentioned above. He took Laplace and Gauss to task for 'fall(ing) into error on a question of prime theoretical importance' by adopting the Bayesian approach that 'Bayes (had) tentatively wished to postulate in a special case' and which was published posthumously (Bayes, ibid). Be then proposed in a restricted context the fiducial method, as discussed.

Neyman and Pearson (1933) then gave a mathematical formulation of fiducial probability that became known as confidence intervals. Fisher (1956) however treated Neyman and Pearson's formulation as a 'misconception having some troubling consequences …'; logical and philosophical arguments between the two sides were intense for many years. In particular the slight to Laplace and Gauss may well have affected the views of the more mathematical participants.

Fisher (1930) entitled his paper 'Inverse Probability' and examined a statistic $t(y)$ whose distribution depended on a single parameter θ. Let $P = F(t; \theta)$ be the distribution function of t, and let P itself be what we might now call a p-value for assessing θ; of course in the usual continuous case P has the uniform distribution on (0, 1). 'If now we give P any particular value such as 0.95, we have … the perfectly objective fact that in 5 percent of samples' t 'will exceed the 95 percent value corresponding to the actual value of θ ….' Then to 'any value of' t 'there will moreover be usually a particular value of θ to which it bears this relationship; we may call this the 'fiducial 5 percent value of θ' corresponding to' the given t. This led to Neyman and Pearson's (ibid) confidence methodology but Fisher treated this as a misconception and he followed different directions and interpretations for the fiducial methodology; for a view on related approaches see *Estimation: Point and Interval*. For this present simple case with scalar t and scalar θ, there seems little difference between the fiducial and the confidence approaches and interpretations.

This discussion effectively ascribes a distribution to θ based on an observed t; this is called the fiducial distribution for θ. Just as the density of t for given θ is obtained as $(\partial/\partial t)F(t; \theta)$ so also the fiducial density is obtained as $(-\partial/\partial\theta)F(t;\theta)$; the negative sign is inconsequential and is merely the result of $F(t; \theta)$ being examined typically for the case that is increasing with θ. The fiducial distribution for cases like the present can also be called the confidence distribution, which recently has a variety of uses in statistical inference. This alternative name emphasizes its more conventional role and attempts to avoid the inappropriate stigma that surrounds the fiducial concept.

For the fiducial approach in more general contexts, Fisher (1935) recommended a maximum reduction to a statistic $t(y)$ by the use of sufficiency. For example, with a sample $y_1,\ldots, y_n$ from the normal $(\mu; \sigma_0^2)$ distribution, the reduction would be to $t(y) = (\bar{y}, s_y^2)$. The methodology then suggests the use of a pivotal quantity $p = p(t; \theta)$ with a fixed distribution and a one-one relationship between any two of p, t, θ; recall more generally that a pivotal quantity is a function of the variable and parameter that provides a measure of departure of variable value from parameter value, and has a fixed distribution which allows an assessment of an observed departure. Thus for the example a natural pivotal is

$$\left\{ z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}, \quad \chi^2_{n-1} = \frac{(n-1)\,s_y^2}{\sigma^2} \right\}$$

which has independent components, normal (0, 1) and chi-square with $n-1$ degrees of freedom. Fisher (1956, p. 172), however, rather deviously rejected this as a legitimate part of the fiducial methodology, but he was somewhat less explicit about what would be legitimate. This pivotal reduction procedure is now however a rather familiar component of standard inference theory and in particular of confidence theory.

The final step is to invert the pivotal quantity, that is, to insert the observed values for the variables and then transform the distribution of the pivotal quantity to the parameter. For the example this gives

$$\mu = \bar{y} - \{z/(n-1)^{-1/2}\chi_{n-1}\}s_y/\sqrt{n}$$
$$\sigma^2 = (n-1)\,s_y^2/\chi^2_{n-1}$$

where the $\bar{y}$, s_y^2 have their observed values. We can then write

$$\mu = \bar{y} - ts_y/\sqrt{n}$$

where t is Student with $n-1$ degrees of freedom. This fiducial calculation closely parallels that for the confidence approach, except that the limits here are calculated from the Student distribution for μ rather than from the Student distribution of the pivotal t. In quite wide generality fiducial regions can correspond to confidence regions; it is just a matter of whether the limits are calculated before or after the data are observed, a non-issue from the Bayesian viewpoint.

There are however cases where routinely obtained confidence and fiducial regions can differ. For example, consider a sample $(y_{11}, y_{12}),\ldots, (y_{n1}, y_{n2})$ from a bivariate normal distribution with means μ_1, μ_2, variances σ_1^2, σ_2^2, and correlation ρ, and suppose primary interest centers on the regression coefficient $\beta = \rho\sigma_2/\sigma_1$ of the second variable on the first variable. Standard regression analysis of the second variable on the first variable can produce an interval that differs

from that obtained from the joint fiducial or structural distribution for all the parameters by integrating out the unwanted parameters; see for example, Fraser (1979, pp. 189, 204, 280, and 293). Also there are other familiar cases where fiducial, structural and Bayesian methods can differ; see for example, the Behrens-Fisher and the Fieller-Creasy problems, and for some discussion see Wallace (1980).

Now suppose we generalize and examine a sample $y_1,\ldots, y_n$ from some distribution with location μ and scale σ. For example the basic form $f(z)$ of the distribution might be, say, the longer-tailed logistic, or the Student with 6 degrees of freedom often cited as a realistic error pattern. The distribution for the underlying errors is then $f(z_1)\cdots f(z_n)$ and for the sample is

$$\sigma^{-n}\prod f\{\sigma^{-1}(y_i-\mu)\}$$

With a typical general error shape we do not have a simple minimal sufficient statistic as with the normal error pattern above. Fisher (1934) then recommended conditioning on an ancillary statistic and suggested the statistic

$$d_1=\frac{y_1-\hat{\mu}}{\hat{\sigma}},\ldots, d_n=\frac{y_n-\hat{\mu}}{\hat{\sigma}}$$

called the configuration statistic, whose coordinates are standardized residuals, standardized here with respect to maximum likelihood values. The conditional distribution of $(\hat{\mu}, \hat{\sigma})$ is fairly easily expressed in terms of likelihood. Let

$$L^0(\mu, \sigma) = c\sigma^{-n}\prod f\{\sigma^{-1}(y_i^0-\mu)\}$$

be the observed likelihood function from data $(y_1^0,\ldots, y_n^0)$, and $(\hat{\mu}^0, \hat{\sigma}^0)$ be the observed maximum likelihood values. Then the conditional distribution given the observed configuration statistic $(d_1^0,\ldots, d_n^0)$ is

$$g(\hat{\mu}, \hat{\sigma}\,|\,\mathbf{d}^0; \mu, \sigma)\frac{d\hat{\mu}d\hat{\sigma}}{\hat{\sigma}^2} =$$

$$cL^0(\hat{\mu}^0+(\mu-\hat{\mu})\hat{\sigma}^0/\hat{\sigma}, \sigma\hat{\sigma}^0/\hat{\sigma})\frac{d\hat{\mu}d\hat{\sigma}}{\hat{\sigma}^2}$$

which can be used to give confidence intervals; also a straightforward fiducial argument following that for the normal case gives

$$L^0(\mu, \sigma)\frac{d\mu d\sigma}{\sigma}$$

for the fiduicial distribution. If f is replaced by the standard normal we get the result discussed earlier and we get it without any reference to sufficiency. This with some other results suggests (Fraser and Reid 2000) that the primary concept sufficiency can be replaced quite widely by appropriate conditioning, and with a major gain in generality of viewpoint.

Fisher (1935) did not use maximum likelihood notation for this location scale model. The notation here adapts to more general contexts (Fraser 1968, 1979) but needs to be modified if the error distribution involves a shape parameter which would then make the maximum likelihood estimates parameter-dependent.

For more general problems the fiducial can have non-uniqueness difficulties, and these tend to arise from the choice of what pivotal to use. For example with the bivariate normal discussed, different pivotals can produce different fiducial distributions (Fisher 1956, p. 172, where he elusively denigrates the pivotal approach).

There are also issues connected with finding a marginal fiducial for a component parameter. For a general discussion of marginalization paradoxes connected with methodologies concerning distributions for parameters, including the Bayesian methodology, see David et al. (1973).

For a mathematical analysis that avoids some of the difficulties with the fiducial methodology, see the transformation group framework in Fraser (1961a, 1961b); this mathematical analysis led to the development of the structural approach described below. The marginalization issues still remain, however, and this is related to the common assessment technique of examining a procedure with many repetitions from the same distribution with fixed parameter value. Alternatives to this are suggested in Fraser and Reid (2000).

3. *Structural Probability*

In many applied problems it is possible to ascribe an underlying error distribution to the model. For example with measurements that are normal with unknown error scaling we might write

$$y_i=\mu+\sigma z_i \quad i=1,\ldots, n$$

where the $z_1,\ldots, z_n$, are independent standard normal variables. If we then examine (Fraser 1966) these expressions with observed data we can note rather easily that many characteristics of the underlying realized errors can be numerically evaluated. Specifically, let

$$d_1=\frac{y_1-\bar{y}}{s_y},\ldots, d_n=\frac{y_n-\bar{y}}{s_y}$$

be standardized residuals; then the observed $y_1,\ldots, y_n$ and the underlying realized $(z_1,\ldots, z_n)$ have the same standardized residuals. In fact, $n-2$ characteristics as presented by the residuals are available concerning the

underlying errors. Most theory concerning applications of probability would then say the analysis should be conditional on these observed characteristics, that is, on what you know concerning the realized error (Fraser 1976, Chap. 4, pp. 161–2 and Chap. 11, pp. 456–66). This for quite general problems automatically gives the conditioning used by Fisher and avoids the need to invoke a conditioning principle and seek an ancillary such as the configuration statistic.

This present type of analysis carries through equally (Fraser 1966) if the normal distribution is replaced by some other error pattern. The conditional error distribution can be viewed from an invariance argument as being valid after the full data are observed. This distribution then gives the structural distribution for μ and σ. The mechanics follow closely what is presented for the fiducial case and the resulting structural distribution agrees with the fiducial distribution, $L^0(\mu, \sigma)\, d\mu d\sigma/\sigma$. The difference here is that a rather strict invariance structure is assumed as part of the argument. This is summarized in general notation.

The structural approach is possible in cases where the underlying error can be directly modeled and the observed response is obtained as some transformation of this. Let Z designate a general error variable with known density $f(Z)$ on say R^n. Let Y designate the response as obtained by a transformation or re-expression θ in some space Ω. Then we have

$$Y = \theta Z, \quad f(Z)\, dZ,$$

For the invariance we assume that a transformation θ carries R^n into R^n and that the collection Ω of such transformations is closed under composition and inversion. Also for details here we assume that the transformations are smooth and at most one transformation carries a point Z into a point Y. For some general background on transformations and invariance, see *Causal Inference and Statistical Fallacies*.

If we consider all the transformation Ω and apply them to a data point Y^0 we obtain a set ΩY^0 of images, of possible antecedent error values. If Z^0 is the unknown realization of error that produced Y^0 then we have $\Omega Z^0 = \Omega Y^0$ or equivalently we have that Z^0 lies in the set ΩY_0. This means that a probability assessment should be made conditional on the observed set ΩY^0. Let $D(Z)$ be a reference point in the set ΩZ and let $[Z]$ in Ω be the transformation in Ω that carries $D(Z)$ to Z. Then we have that the model $Y = \theta Z$ with data Y^0 can be rewritten as

$$[Y] = \theta[Z], \quad D(Y) = D^0,$$

giving the conditioning $D(Y) = D(Y^0) = D^0$ that appears as an observed ancillary in the fiducial framework. The conditional distribution of $[Z]$ given $D(Z) = D^0$ then describes the reduced model $[Y] = \theta[Z]$ with data $[Y^0]$. The structural distribution of θ is then given as $\theta = [Y^0][Z]^{-1}$ using the conditional distribution $[Z]$ just described.

In most contexts the support measure dZ can be replaced by an invariant measure $dM(Z)$

$$dM(Z) = \frac{dZ}{J(Z)}$$

where $J(Z)$ is an appropriate Jacobian (Fraser 1979). Expressions are simpler if we use density $\bar{f}(Z)$ with respect to the measure $dM(Z)$: $\bar{f}(Z) = f(Z)J(Z)$.

The distributions of $[Y]$ and $[Z]$ conditional on $D(Y) = D(Z) = D^0$ are

$$c\bar{f}(\theta^{-1}[Y]D^0)d\mu[Y], \quad c\bar{f}([Z]D^0)\, d\mu[Z]$$

where $d\mu[Z]$ is the corresponding invariant measure on the group Ω. The inverted distribution, called the structural distribution for θ, is

$$c\bar{f}(\theta^{-1}[Y^0]D^0)\, dv(\theta) = cL^0(\theta)\, dv(\theta)$$

using the right invariant measure v on the group Ω. In the Bayesian context the typically preferred noninformative prior is the right invariant measure $dv(\theta)$. Thus the structural distribution coincides with this preferred Bayesian posterior.

This group type model covers a wide range of regression multivariate and spherical distribution models but not in the simplified notation used here; for further details see Fraser (1968, 1979).

Recent likelihood asymptotics has been able to obtain many aspects of these results in a general asymptotic context; see Fraser et al. (1999) and Fraser et al. (1999).

4. *Comment*

Fiducial inference led to confidence intervals, with both having evolved from the Bayesian procedures initiated in the eighteenth century. In terms of freedom of usage, we have Bayesian inference as the more liberal. The foundational issues center on validity and interpretation of the probabilities for the parameters. Here the conditions are strictest for the structural and minimal for the Bayesian. In the nice cases they agree; and in the less standard cases they pressure the proponents to clarify their assumptions.

For some recent discussion of confidence and fiducial interconnections plus references see Barnard (1995), with complementing views in Fraser (1996). Also see Fraser and Reid (2000) for a discussion of the interplay with objective Bayesian methods.

Bibliography

Barnard G A 1995 Pivotal models and the fiducial argument. *International Statistical Review* **63**: 309–24

Bayes T 1763 An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society* **53**: 370–418 [Reprinted (1958) in *Biometrika* **45**: 293–315]

Dawid A P, Stone M, Zidek J V 1973 *Journal of the Royal Statistical Society* **35**: 189–233

Fisher R A 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* **A222**: 309–68

Fisher R A 1925 Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**: 700–25

Fisher R A 1930 Inverse probability. *Proceedings of the Cambridge Philosophical Society* **26**: 528–35

Fisher R A 1934 Two new properties of maximum likelihood. *Proceedings of the Royal Society of London* **A144**: 285–307

Fisher R A 1935 The fiducial argument in statistical inference. *Annals Eugenics* **6**(4): 391–8

Fisher R A 1956 *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, UK

Fraser D A S 1961a The fiducial method and invariance. *Biometrika* **48**: 261–80

Fraser D A S 1961b On fiducial inference. *Ann. Math. Statist.* **32**: 661–76

Fraser D A S 1966 Structural probabilities and a generalization. *Biometrica* **53**: 1–9

Fraser D A S 1968 *The Structure of Inference*. Wiley, New York

Fraser D A S 1976 *Probability and Statistics, Theory and Applications*. ITS, Toronto

Fraser D A S 1979 *Inference and Linear Models*. McGraw Hill, New York

Fraser D A S 1996 Some remarks on pivotal models and the financial argument in relation to structural models. *International Statistical Review* **64**: 23–5

Fraser D A S, Reid N 2000 Strong matching of frequentist and Bayesian inference. *Journal of Statistical Planning and inference*, to appear

Fraser D A S, Reid N, Wu J 1999 A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrica* **86**: 249–64

Fraser D A S, Wong A, Wu J 1999 Regression analysis, nonlinear or nonnormal simple and accurate *p* values for likelihood analysis. *Journal of the American Statistical Association* **94**: 1286–95

Neyman J, Pearson E S 1933 The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society* **29**: 492–10

Wallace D L 1980 The Behrers–Fisher and Creasy–Fieller problems. In: Fienberg S E, Henlele D V (eds.) *R. A. Fisher: An Appreciation*. Springer Verlag, New York, pp. 139–45

D. A. S. Fraser

Field Observational Research in Anthropology and Sociology

Field observational research in anthropology and sociology—or the process of doing ethnography—brings the social science investigator directly to the scene of the human behavior she or he describes. It involves 'being there,' or 'I-witnessing.' So also, of course, do other forms of documentation: journalism, film, travel writing, explorer or missionary's report, even background research by a novelist. What makes ethnography distinctive as a research practice is that it forms one point of a triangular process of knowledge construction which also includes comparative social theory and documentary contextualization.

The fieldnotes that result from ethnographic discovery procedures are not merely 'written up,' but are filtered and interpreted against theoretical propositions and the comparative record of other field observational studies; they are also grounded and enhanced by historical, ecological, demographic, economic, and other documentary sources that provide background and context. As the products of this mode of research are read—monographs and articles also referred to as ethnography—they stimulate new comparative theoretical thinking, which in turn suggests further problems and interpretations to be resolved through more field observational research. Ethnographies also regularly lead to new demands and rising standards for documentary contextualization—more history, more ecological or demographic backgrounding, more attention to state policy, economic trends or the world system. This triangle of ethnography, comparison, and contextualization is, in essence, the process by which field observational research in anthropology and sociology is utilized to explain and interpret human cultures and social life.

1. History and Scope in Anthropology and Sociology

In 1851 Morgan published his *League of the Ho-de-no-sau-nee, or Iroquois*; combining fieldwork and comparative and theoretical interests in political organization, this book was the first anthropological ethnography. It depicted the structure and operation of Iroquois society, detailing matrilineal kinship, political and ceremonial life, material culture and religion. After formation of the US government Bureau of American Ethnology in 1879, a stream of ethnographic accounts of native American societies began, including landmark field observational studies of the Zuni by Cushing, the Baffin Island Inuit by Boas, and, after 1900, ethnographies of Plains, Eastern Woodland, and California groups by Boas' students trained at Columbia University. During the 1930s the scope broadened to include rural and urban studies in Latin America, the Caribbean, Africa, and Europe, and also in contemporary US settings.

In Great Britain's colonies, local government official or missionary fieldworkers followed the topical guidebook *Notes and Queries on Anthropology*, drafted by Oxford University theorist Tylor and others, and first published in 1874. This armchair scholar/man-on-the-

spot division of labor ended two decades later when Oxford-trained Spencer collaborated in field observational study with local expert Gillen, producing their *Central Tribes of Native Australia* in 1899, and providing detailed coverage of cosmology, ritual, and social organization which revealed unheralded complexity, and which sparked theoretical work by Durkheim and Freud. In 1898 a team from Cambridge University conducted fieldwork in Torres Strait north of Australia, and member Rivers used genealogical data to probe community history, migration trajectories, marriage patterns, demography, and inheritance. Rivers and his students conducted further ethnographic studies in India and the Pacific, but their work was eclipsed by the more thorough Trobriand Islands field observational research of Malinowski, resulting in *Argonauts of the Western Pacific* in 1922 and its famous first chapter on method. Malinowskian ethnography dominated British social anthropology thereafter, and by the 1940s reshaped American cultural anthropology as well.

In sociology, there is less agreement about the disciplinary origin of field observational research methods. Conventionally, the monographs of the University of Chicago department during the 1920s and 1930s are cited (Madge 1962): studies of itinerant 'hobo' workers, youth gangs, dancehall employees and patrons; Italians, Poles, Jews, French Canadians, Mexicans, and Chinese; and central city 'gold coast' apartment, artist colony, and rooming house dwellers. Two decades before this research, however, Du Bois used similar methods (both he and the Chicago school were inspired by British and American urban survey and settlement house movements) to produce *The Philadelphia Negro* (1899), which depicted life in an inner city neighborhood comprised of southern migrant and northern-born African-Americans. Even earlier, Engels in 1844 studied new classes, Irish labor migration, urban spatial arrangements, symbolic attachments, and modes of observable behavior in Manchester and other rapidly industrializing British cities, although his *Condition of the Working Class in England* was not published in English until 1892.

Since the 1920s there has been much cross-fertilization between anthropological and sociological ethnography. A few of the Chicago studies were conducted by researchers identifiable as anthropologists, and the Lynds, authors of *Middletown* (1929) which credits anthropologists Rivers and Wissler for methodological inspiration, and Whyte, author of *Street Corner Society* (1955 [1943]) and influenced by Harvard anthropologists, are presented as major figures in the development of American sociology (Madge 1962). Gluckman, a Manchester social anthropologist whose fieldwork was conducted in Africa, was influential in the development of ethnographic fieldwork by British sociologists working in schools, gangs, and factories during the 1950s and later (Burgess 1984). In recent decades, both anthropologists and sociologists have undertaken field observational research in a wide range of overlapping settings, at home and abroad; so increasingly do scholars from folklore, education, environmental psychology, political science, and other disciplines. Ethnography, however, remains central to credentialing, teaching, publishing, and theory-construction in anthropology, while it occupies a more marginal position in contemporary sociology, or in other fields.

2. *Selection, Access, Rapport*

The selection of a particular population or site for field observational research is ordinarily related to some unanswered question or outstanding problem in social theory. Personal connections or predilections of researchers also shape this selection, but the fieldworker must justify his or her choice in terms of some significant theory to which the project is addressed. Usually this is made explicit in a written proposal for funds to underwrite the fieldwork.

While ethnography is thus lodged from conception in social theory, a comparative viewpoint also molds the fieldwork process. First, investigators acquire a cross-cultural or cross-site perspective by training and reading; when addressing any aspect of social life—marriage, leadership, ethnicity, etc.—field observational researchers recall examples of similarities and differences elsewhere. Second, the comparative perspective focuses attention on trends and transitions, not just on similarities and differences at random (which are infinite). Rather than treating each fieldwork case as unique, ethnographers place the social phenomena they study within comparative frames (foraging, horticultural, agricultural, pastoral, industrial, colon ial, neocolonial economies; cooperative, competitive, individualistic societies; gender subordination, complementarity or equality; etc.) Ethnographic results in turn provoke debate, revision, and innovation in such theorizing.

While significant theories bring fieldworkers to particular locations, actors, and activities, once they arrive they begin to observe and listen. Often they must learn to listen—by learning the language, the local vocabulary, and current verbal conventions. Field observational research now turns away from theoretical discourse to the viewpoints and concepts of the people (informants, subjects, actors, consultants) themselves. Ethnographers aim to document how the people see and talk about their everyday social activities and groupings, and about the wider worlds they live in. It is their normal scenes of activity, topics of conversation, and standards of evaluation that are the objects of field observational research.

This does not begin by announcing: 'I'm your ethnographer; when can I interview you?' Investigators must be honest about their role and sponsorship (covert field observational research remains contro-

versial and rare), but their paramount aim is to listen and to move as quickly as possible into the natural settings of social life, the places people would be, doing what they would be doing, if the ethnographer was not there. This involves a continuing process of requesting and obtaining access to the many locations, activities, and groups that comprise the overall fieldwork arena. As ethnographers observe and listen in a wide-ranging manner, within parameters set by the significant theories that bring them there, they begin to understand culturally meaningful behavioral conventions, and to formulate culturally appropriate utterances. This indexes a second continuing process of building interpersonal rapport with informants, a topic that preoccupies many autobiographical accounts of field observational research—one of the most instructive being Whyte's 1955 [1943] Appendix to *Street Corner Society*.

As the ethnographic process unfolds, the fieldworker must constantly make decisions about where to be, whom to listen to, what events to follow, and what to safely ignore and leave out. These decisions are guided both by the significant theories prefiguring fieldwork, and by new theories of significance that arise in the field. These latter theories—hunches, hypotheses, ideas about connections and relationships—emerge as listening and observation proceed. They suggest which persons and activities to focus upon, what places and events to attend to, and which objects and their circulation to follow. As this occurs, the fieldwork 'funnel' narrows, to use Agar's (1980) apt metaphor. The early phase of securing access and building rapport is wide, open, and nearly all-encompassing; as theories of significance emerge, pan out, or are discarded, the funnel of informants, events, and activities narrows. Goals crystallize. Research design sharpens.

3. *Discovery Procedures: Listening*

Ethnography's greatest strength is situated listening. Interviews become useful at later stages of fieldwork, but field observational research begins by listening to what British anthropologist Richards (1939) called 'speech in action.' Here the actors control topicality—talking to each other about what they usually do—and the researcher secures access to their turf—the locations they usually occupy. Early on, as rapport begins to be established, fieldworkers deliberately place themselves in a wide sampling of such situations. Then, as the research funnel narrows and other methods are deployed, the ethnographer becomes more selective about where to listen.

The importance of situated listening was emphasized memorably by Whyte. After he asked an intrusive question during illicit gambling activities, and stopping the ongoing flow of behavior, Whyte's key informant Doc told him, 'Go easy on that "who," "what," "why," "when," "where" stuff, Bill. You ask those questions and people will clam up on you. If people accept you, you can just hang around, and you'll learn the answers in the long run without even having to ask the questions.' 'I found that this was true. As I sat and listened, I learned the answers to question that I would not even have had the sense to ask if I had been getting my information solely on an interview basis' (1955 [1943], p. 303).

Situated listening becomes the basis for understanding both what can be talked about in later conversations and interviews, and what categories and evaluations the actors use to observe and interpret ongoing behavior.

4. *Discovery Procedures: Observing*

Many veteran field observational researchers bristle when they are told that their methods are not quantitative. Using both emic categories articulated by the actors, and etic categories formulated by the community of social science observers, ethnographers count, sample, and calculate ratios, proportions, and probabilities as part of their fieldwork activities (Johnson 1978). These countable quantities include numbers of people in events, their positions, their comings and goings; objects, inventories; and exchanges, movements, orderings, sequences, associations, assemblages, and arrangements of many sorts.

Malinowski's classic *Argonauts of the Western Pacific* (1922) incorporated his numerous observations of the interisland *kula* exchange of seashell valuables—necklaces traded for armlets—as canoes from one community arrived at another, exchanges were conducted, and then additional exchanges of items including foodstuffs, timber, greenstone, and pottery followed. Who exchanged what, the order, etiquette, volume, frequency, and other variables were all counted and analyzed. In Whyte's *Street Corner Society* (1955 [1943]) the size, composition, duration of activity, and initiation of interaction were recorded and analyzed for numerous group events; Whyte even demonstrated quantitatively a relationship between hierarchical group leadership structure and bowling scores.

Harris (1964) furnishes field observational researchers with an etic model of 'the great stream of uninterrupted behavior' that extends backward in time and unfolds continuously. He provides a technical vocabulary to operationalize recording of isolatable events and linked segments of the behavior stream—from individual body movements and their environmental effects at the most basic level, through sequences of multiactor scenes connected by actor, place, or object. Sanjek (1978) applied this approach in a multiethnic Ghanaian urban neighborhood to

collect (through interviews) sequences of multiactor scenes for 40 adults, each over four-day periods, in order to test whether their social networks were structured more by common ethnicity or by class position, and in which behavioral settings ethnicity and class were stronger or weaker.

Johnson (1975, 1978) developed an observational fieldwork approach to formulate statistically valid time allocation patterns in his research among the Machiguenga of Peru. By randomly selecting actors and times for observation at their most likely location, and recording their ongoing behavior at the time of each random visit, a powerful quantitative record of how a given population allocates its activities among production, reproduction, leisure, and other categories resulted. This and similar methods have been employed in a wide range of fieldwork settings (Gross 1984).

5. *Discovery Procedures: Interviewing*

Barth recounts of his highland New Guinea research, 'My field procedure was ... to rely very heavily on observation of spontaneous, unelicited word and act. I paid particular attention to this in Baktaman questions and explanations *to each other*—in men's houses, in the fellowship of hunting and working, during initiations and other rites. Only secondarily, and with much reticence, did I ask my own questions, and then only when they seemed cast in a native pattern and received easy and natural response' (1975, p. 226).

Interviewing is problematic (Briggs 1986). Human beings are apt to reinterpret or reformulate the past to make it conform with their ongoing sense of the present. Moreover, the ever present danger of an observer effect in fieldwork—that the presence of the observer affects and reshapes an actor's behavior and response—is especially difficult to control in an artificial one-to-one interview situation where normal social activity is suspended. Still, interviews are essential to field observational research. They are best left for later stages of the ethnographic encounter when, to use Agar's funnel metaphor, the speech events of fieldwork move from wide and open to narrow and more focused.

After the investigator has become at ease with the language and local codes and references, and has successfully gained access to natural conversations and amassed a stock of cultural knowledge via situated listening, deliberate intervention into speech events may begin. These steps may be graded from least to most intrusive and controlled.

(a) Still on the informants' turf, and still in the accustomed activities of daily life, the field observational researcher begins to enter natural conversations as a speaker, and begins to shift topicality to her or his own interests. This process starts gently, by moving appropriately into rounds of chatting, gossiping, and ordinary comment. As cultural competence increases, and as theories of significance emerge, the fieldworker attempts to direct conversations by introducing questions and suggesting topics for responses from informants.

(b) After an initial period of fieldwork—a few months perhaps—arranged interviews may begin. This class of speech events is disruptive—the informant is removed from her or his turf, either to the ethnographer's household or office, or by transforming an everyday location into a scene of ethnographer–informant dialogue—an activity that would otherwise not be occurring there. Typically the earliest of these deliberate breaks in time–place behavioral flow reserve topicality for the actor. In such open-ended or discovery interviews, the informant is encouraged to move the conversation according to his or her own interests.

(c) In later and more productive interviews, the ethnographer begins to assert control. General topics are introduced, allowing informants to expand freely upon their own points of view and knowledge. In more structured ethnographic interviews, topicality is more firmly shaped and directed by the fieldworker; informant responses now move away from orations and free commentaries and to more specific responses to questions.

(d) In the most focused forms of interviewing, the ethnographer controls both turf and topicality as fully as possible. Questionnaires and interview schedules may be used, and the objective is to obtain particular types and pieces of data. These may include household interviews, psychological tests, or reports of disputes, or may encompass repeated interview sessions to secure lengthy life histories, with the fieldworker now guiding the subject according to preset standards of scope and comprehensiveness.

(e) Though not a major part of ethnographic practice, in some instances, and when still on the informants' turf, the fieldworker may ask direct and pointed questions, and attempt to secure precise pieces of data for his or her records. Interventions of this sort are especially dangerous—the inappropriateness of such seizures of topicality in everyday settings may be jarring to the actors. Typically speech events of this sort occur in the final days of fieldwork, when local rapport is at its peak, research goals are most pressing, and the fieldwork funnel approaches its vanishing point.

6. *Notes and Records*

The production of notes and records (Sanjek 1990) begins to move the field observational research process toward its ultimate written products. Focused in-

terview sessions with seated informants often permit direct transcription of verbal statements, but in open-ended and ethnographic interviews, brief written notes—what Ottenberg (in Sanjek 1990) terms scratch notes—are taken during the session, and these form the basis for the construction later of fuller written fieldnotes. Fieldworkers often go through this two-step process even when interviews are tape recorded, both as a backup to and index of the taped session, and because of the analytic gains many ethnographers note in transforming their scratch notes into fuller descriptive fieldnotes.

In field observational research in natural behavior stream settings, similar brief jottings may be inscribed, but major attention is directed to the event in progress. Often it is not even possible to record scratch notes, and both they and fuller fieldnote description occur later. Mead (1977) wrote about the nagging pressure to type up fieldnotes from scratch notes, and about the danger of scratch notes growing 'cold' when this was delayed, even by one day. But she also wrote of the satisfaction of being caught up with this work, and of the importance for later ethnographic writing of the insights gained in moving from scratch notes to descriptive fieldnotes.

Ottenberg sees this process as the interaction of scratch notes and headnotes—the stored memories and interpretations that arise from direct participant observation as filtered by the ethnographer's overall theoretical stance. Headnotes form an essential complement to fieldnotes (and to more formal fieldwork datasets or records). Headnotes are employed to make sense of one's fieldnotes when they are re-read later for ethnographic writing projects. The importance of headnotes is particularly evident when one ethnographer attempts to use another's fieldnotes, and quickly realizes how difficult it is to understand them without headnotes of one's own (Sanjek 1990).

Fieldnotes typically are kept in running chronological order, and may minimally be indexed separately by actor and topic. Additionally, field observational researchers often generate records, or well-defined sets of data organized by topic, person, or other category. These may include records of household composition, land holdings, ritual performances, life histories, folktales, etc. Sometimes decisions to collect sets of records are part of the initial research design, and preset forms to enter such data may be brought with the investigator to the field setting. In other cases, a decision to collect a systematic body of records develops during fieldwork, often emerging within general running fieldnotes and then being separated out in new record files.

Fieldnotes and records present ethnographers with great masses of information—hundreds, even thousands of pages—that while being accumulated may be arranged only by chronological order or by topic. Malinowski (1922) urged that ethnographers constantly re-read and begin to organize and analyze their notes while still conducting field research, but commonly the more focused work of indexing them occurs when fieldwork is over. As ethnographers then turn to ethnographic writing, they readdress the theoretical discourse they turned away from in fieldwork. Fieldnotes and records must now be related directly to the comparative and contextual points of the knowledge-construction triangle.

7. *Canons of Validity*

Field observational research aims to maximize validity—'the degree to which scientific observations actually measure or record what they purport to measure'—rather than reliability—'the repeatability, including interpersonal replicability, of scientific observations' (Pelto and Pelto 1978, p. 33). Reliability is valued in survey, medical, and product safety research. In ethnography, restudies to assess change may be conducted, but independent investigation of the same locale or population is unlikely or impossible.

Validity depends on the intensity and depth of field observational research, but it always remains a goal rather than a property that can be independently measured (Wolcott 1995, pp. 167–70). A quantitative, and highly reliable, randomized sample study that investigates the gender, ethnicity, class, and other characteristics of the five best friends of members of a given population would be less valid than one in which cultural meanings and categories of friendship were first discovered through listening to natural conversations or eliciting such information in interviews. In the case of Indian immigrants in New York City, for example, the 'five best friends' study might miss entirely the category of *rakhi* ritual friendships between adult males and females which are affirmed in an annual festival where the man ties a flower bracelet on the woman's wrist.

If validity remains important, but elusive, how can it be assessed in field observational research results? Sanjek (1990, pp. 393–404) proposed three canons of validity by which ethnographic writing may be evaluated. The first is theoretical candor, the openness with which the ethnographer addresses the significant theories, and even more, the local theories of significance that structured the fieldwork process. A second canon calls for explicit depiction of the ethnographer's fieldwork path—the number of informants from whom information was obtained, in what ways, and their relationship both to the wider population the field observational study concerns and to each other. A third canon concerns information about the fieldnote evidence itself: not simply 'how much' and its basis in listening, observing, or interviewing, but more significantly the precise relationship of notes and records to the written ethnography. Some ethnographies utilize fieldnotes directly, even masses of them; others, for rhetorical or narrative purposes, do

not, and need not. What matters in the end is that readers of field observational research have a clear picture of what the ethnographer did and why, who they talked to and learned from, and what they brought back to document it.

Field observational research produces results that can be obtained in no other way. This ethnographic tradition entails that 'the description of people's activities, their interactions with each other, [and] their verbal behavior' be 'copious and detailed'; 'cumulative development … consists above all in the fact that such data, provided they are adequate, never become obsolete and can be used by analysts … in researching questions not visualized by the researchers who collected them' (Anglin 1979, p. 49).

See also: Fieldwork in Social and Cultural Anthropology; Observational Studies: Overview; Participant Observation

Bibliography

Agar M 1980 *The Professional Stranger: An Informal Introduction to Ethnography*. Academic Press, New York

Anglin A 1979 Analytic and folk models: The Tallensi case. In: Holy L (ed.) *Segmentary Lineage Systems Reconsidered*. Department of Social Anthropology, The Queen's University of Belfast, Ireland

Barth F 1975 *Ritual and Knowledge Among the Baktaman of New Guinea*. Yale University Press, New Haven, CT

Briggs C L 1986 *Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research*. Cambridge University Press, New York

Burgess R G 1984 *In the Field: An Introduction to Field Research*. George Allen and Unwin, London

Emerson R M, Fretz R I, Shaw L L 1995 *Writing Ethnographic Fieldnotes*. University of Chicago Press, Chicago

Gross D R 1984 Time allocation: A tool for the study of cultural behavior. *Annual Review of Anthropology* **13**: 519–58

Harris M 1964 *The Nature of Cultural Things*. Random House, New York

Johnson A 1975 Time allocation in a Machiguenga community. *Ethnology* **14**: 301–10

Johnson A W 1978 *Quantification in Cultural Anthropology: An Introduction to Research Design*. Stanford University Press, Stanford, CA

Madge J 1962 *The Origins of Scientific Sociology*. Tavistock, London

Malinowski B 1922 *Argonauts of the Western Pacific*. Routledge, London

Mead M 1977 *Letters From the Field, 1925–1975*. Harper and Row, New York

Pelto P J, Pelto G H 1978 *Anthropological Research: The Structure of Inquiry*. Cambridge University Press, New York

Richards A 1939 The development of field work methods in social anthropology. In: Barlett F C et al. (eds.) *The Study of Society*. Routledge and Kegan Paul, London

Sanjek R 1978 A network method and its uses in urban ethnography. *Human Organization* **37**: 257–68

Sanjek R (ed.) 1990 *Fieldnotes: The Makings of Anthropology*. Cornell University Press, Ithaca, NY

Whyte W F 1955 [1943] *Street Corner Society: The Social Structure of an Italian Slum*, 2nd edn. University of Chicago Press, Chicago

Wolcott H F 1995 *The Art of Fieldwork*. AltaMira, Walnut Creek, CA

R. Sanjek

Fieldwork: Ethical Aspects

Fieldwork involves a range of methodologies including ethnographic participant observation, in-depth interviews, direct observation of individual or group behavior, or a combination of both qualitative and quantitative approaches (Denzin and Lincoln 1998). Social scientists conducting fieldwork confront unique ethical dilemmas in the course of implementing their investigations, particularly when researchers are actively engaged with participants in their natural environments. Ethical dilemmas occur when there are conflicts about what is considered to be an appropriate response to a situation involving the investigators, research participants, or the production and use of research materials. A full treatment of these issues is beyond the scope of this paper. This article briefly explores ethical issues associated with informed consent, the protection of confidentiality, the development of personal relationships with research participants, involvement in illegal activities, and intervention in on-going social activities. Finally, a framework for resolving ethical dilemmas in fieldwork is outlined.

1. Informed Consent

National and international guidelines for ethical conduct in research with human subjects identify specific requirements for informed consent: the provision of information, comprehension of information, and voluntary participation (Faden and Beauchamp 1986, Weisstub 1998). Voluntary participation in research depends upon the individual's ability to understand not just the meaning of the research, but the impact it may have on his or her life. Excessive financial compensation or unrealistic promises may constitute coercion, especially if someone is vulnerable because of social factors such as poverty or ethnic background. Voluntary participation may be inhibited when there are social status and social class differences between fieldworkers and research participants—differences based on gender, age, education, income level, religious affiliation, and ethnic identity. The institutional affiliation of the field researcher may also pose a threat to participants. Additionally, the fluid and unbounded nature of many social science field studies complicates

the straightforward approach to informed consent associated with other forms of social science methodologies.

A number of factors influence informed consent in fieldwork—the specific goals of the research, the cultural context of the research project, and communication issues that influence comprehension of information. In international settings, the challenges associated with implementing informed consent in fieldwork are heightened because of language barriers that may inhibit effective communication, particularly regarding the translation of scientific concepts. Misunderstandings and miscommunication about social science research are more likely to occur when fieldworkers and participants speak different languages. Language barriers may be reduced through the use of an interpreter. However, translators must negotiate not only language, but also cultural and contextual factors. Some of the problems associated with interpretation in social science fieldwork include: the inability to translate equivalent expressions easily across languages; paraphrasing that results in omissions or erroneous substitutions of terms; varying levels of comprehension; and the influence of conflicting cultural beliefs and values concerning the meaning and purpose of research.

Differences in beliefs about who has the authority to give permission to participate in research also influence the consent process. Beliefs about personhood, individual autonomy, and decisional capacity are embedded within the social and cultural patterns of family ties and community obligations. In Western industrialized nations, where personal autonomy is emphasized, research participants are expected to make decisions for themselves or through designated surrogates. However, in many non-Western settings, religious or tribal leaders or a person's extended family may play a significant role in decisions concerning research participation. Cultural differences in beliefs about the nature of personhood and the location of decisional authority for consent have been problematic for social science investigators conducting fieldwork in international settings (Levine 1991). Negotiating informed consent with designated authorities in non-Western populations requires fieldworkers to move beyond narrow definitions of personhood, autonomy and self-determination. While respect for persons demands cultural sensitivity to local customs, social scientists conducting fieldwork are never authorized to conduct research without regard to potential harm and without attempts to be informative throughout the duration of the study (Levine 1996). Moreover, although a tribal chief, village elder, or community leader may express approval of a research agenda, sensitivity to cultural customs is secondary to honoring individual choice regarding participation (World Health Organization and the Council for International Organizations of Medical Sciences 1993).

1.1 Informed Consent in Observations of Group Interactions

In fieldwork involving direct observation of group activities, arrangements usually are made prior to initiating the study to inform group members that the fieldworker will be present in the course of routine activities. In closed systems such as a hospital unit, a community center, or other institutional setting, informed consent should be obtained from all those who are at the facility on a regular basis. Individuals who are not present on a regular basis but whose behavior may be observed in public group activities should be alerted to the presence of the fieldworker if it is feasible to do so. In some group observations conducted in the course of fieldwork, it may not be possible or necessary to obtain informed consent from every person present. For example, at informal gatherings such as family or community events, or in public places where individuals interact informally, it would be intrusive for the fieldworker to explain the study to everyone present.

2. Protection of Confidentiality

Protection of confidentiality and privacy for research participants is an important aspect of fieldwork (Cassell and Jacobs 1987). Fieldwork in social science investigations often involves the collection of sensitive information concerning an individual's personal life, including information on physical conditions, mental illness, or behavior regarding illegal activities. Individuals or communities could be harmed if information was inadvertently or purposely disclosed. In some cases, information could be used to ostracize or marginalize research participants, particularly if research participants suffer from a stigmatizing illness such as HIV/AIDS, or if they are involved in activities that may be viewed as socially deviant.

Data collection strategies associated with fieldwork require careful attention to the representation of individuals and communities in written transcripts derived from interviews and observations (Fluehr-Lobban 1991). Descriptive data recorded in field notes, daily logs, or diaries should be entered using a coding system that protects the confidentiality of the research participants. Additionally, field data should also be secured outside of the research setting in order to safeguard the identity of participants. Pseudonyms, rather than the names of participants, must be used in published manuscripts that include case narratives. This gesture, however, will not necessarily protect confidentiality because it may be possible to identify the individuals involved based on the details of the case and the research setting. Nevertheless, at each stage of the research—data collection, data analysis, and data reporting—every effort should be made to protect the privacy of the study participants. In

obtaining informed consent, investigators are obliged to advise participants in fieldwork about the methods designed to protect their privacy.

3. Relationships with Research Participants

Social scientists involved in fieldwork have the opportunity to form close personal relationships with research participants because of the collaborative nature of ethnographic approaches such as participant observation. The development of trust between fieldworkers and the individuals involved in the research enhances the potential for collecting important information relevant to the study. However, in some circumstances, investigators may find themselves in compromising situations because of their personal relationships in the field.

3.1 Intimate Relationships

Although it is rarely addressed in the literature (for an exception, see Kulick and Willson 1995), fieldworkers may become intimately involved with research participants. Sexual relationships with field participants are more likely to occur in the context of extended fieldwork. Traditional anthropological fieldwork—where the investigator may spend months or years in an environment far removed from familiar social networks (often in a different country)—represents the type of setting that contributes to the development of intimate personal relationships with research participants. Despite what some might view as the natural inclination to form a sexual attachment with a close companion, the prevailing opinion among social and behavioral scientists is that sex with participants in the field should be avoided. Sexual relationships between researchers and study participants may diminish the relative objectivity of the investigator.

More problematic is the potential for coercion of socially vulnerable individuals. Moreover, attitudes concerning sexual behavior are socially constructed and the meanings attached to sexual intimacy may vary significantly between the fieldworker's home community and the research setting (Singer et al. 1999). The social complications that can result from a sexual relationship (e.g., jealousy, accusations of infidelity, gossip) jeopardize the integrity of the research and place the investigator at risk for problematic interactions with the broader community.

3.2 Illegal Activities

Investigators conducting fieldwork with populations such as illegal immigrants, or with individuals involved in activities such as drug use or the sex trade, may be asked to participate in behavior that is legally questionable. For example, fieldworkers investigating the use of illegal drugs may be asked to share drugs with study participants or to assist in procuring or holding drugs or drug paraphernalia. While some researchers may believe that participation in these types of activities strengthens the bonds of trust between a fieldworker and research participants and thus increases the likelihood of securing information that is meaningful and necessary to achieve the study goals, there are strong reasons for avoiding participation in illicit or illegal behavior. As with sexual involvement with study participants, engaging in illegal activities seriously threatens the integrity of study findings and places the researcher at risk socially, professionally, and in some cases, physically. Fieldworkers may find themselves targeted for blackmail or other forms of coercion. Additionally, fieldworkers who participate in illegal behavior diminish the potential to secure funding for future research. A full consideration of the ethical dimensions of involvement in illegal activities cannot be explored in this essay.

4. Fieldworker Intervention in Ongoing Social Activities

Questions concerning the intervention of a fieldworker in problematic situations that occur in the course of conducting research are challenging. The fieldworker's decision should be a judgment based on the purpose of the study, the context, and the specific event that occurs. When investigators continually intervene in field situations that are directly related to the issue being investigated, the possibility that their research will have an impact is jeopardized since they would be unable to document the events being studied. For example, fieldwork with injection drug users to investigate HIV risk behavior would be inconclusive if researchers attempted to stop practices known to facilitate the transmission of the virus. The fundamental problem is one of advocacy—when to be an advocate for research participants and when to maintain the more circumscribed role of the investigator (Bosk 1985). In general, non-intervention is the accepted mode for fieldwork.

5. Recommendations for Resolving Ethical Dilemmas

Investigators who confront moral dilemmas in the course of conducting fieldwork must consider carefully the full range of issues involved in the problem. A systematic approach to ethical decision-making begins with a robust description of the research dilemma.

This would include details about the purpose of the study, the research design, the sponsors of the study, and the individual participants and community involved in the research. The researcher then must consider the cultural and social values represented in the ethical problem. The relevant values of the study participants, the study community, the professional community, and the study sponsors should be outlined. The researcher must determine the principal value conflict. At this point in the process of decision-making, the fieldworker should consider whose values are threatened and who are the most vulnerable to potential harms. This process of reflection facilitates the identification of the key issues involved in the ethical dilemma. When the primary ethical issues are determined, the investigator should outline a full range of strategies and consider the potential risks and benefits associated with each solution. The decision regarding a course of action should maximize respect for the individual and group values identified. The vulnerability of research participants and the communities within which they live should be of paramount importance in resolving the ethical dilemma encountered.

See also: Experimenter and Subject Artifacts: Methodology; Fieldwork in Social and Cultural Anthropology; Linguistic Fieldwork; Research Conduct: Ethical Codes; Research Ethics, Cross-cultural Dimensions of; Research Ethics: Research; Research Subjects, Informed and Implied Consent of

Bibliography

Bosk C 1985 The fieldworker as watcher and witness. *Hastings Center Report* **15**: 10–14

Cassell J, Jacobs S (eds.) 1987 *Handbook on Ethical Issues in Anthropology*. Special Publication of the American Anthropological Association, Number 23, Washington, DC

Denzin N K, Lincoln Y S (eds.) 1998 *Collecting and Interpreting Qualitative Materials*. Sage Publications, Thousand Oaks, CA

Faden R R, Beauchamp T L 1986 *A History and Theory of Informed Consent*. Oxford University Press, New York

Fluehr-Lobban C (ed.) 1991 *Ethics and The Profession of Anthropology*. University of Pennsylvania Press, Philadelphia PA

Kulick D, Willson M 1995 *Taboo: Sex, Identity and Erotic Subjectivity in Anthropological Fieldwork*. Routledge, New York

Levine R J 1991 Informed consent: Some challenges to the universal validity of the Western model. *Law, Medicine, and Health Care* **19**: 3–4

Levine R J 1996 International codes and guidelines for research ethics: A critical appraisal. In: Vanderpool H (ed.) *The Ethics of Research Involving Human Subjects: Facing the 21st Century*. University Publishing Group, Frederick, MD

Singer M, Marshall P, Trotter R T, Schensul J J, Weeks M R, Simmons J E, Radda K E 1999 Ethics, ethnography, drug use, and AIDS: Dilemmas in federally funded research. In: Singer M, Marshall P, Clatts M (eds.) *Integrating Cultural, Observational, and Epidemiological Approaches in The Prevention of Drug Abuse and HIV/AIDS*. US Department of Health and Human Services, National Institutes of Health Publication 99–4565, Washington, DC

Weisstub D N (ed.) 1998 *Research on Human Subjects: Ethics, Law and Social Policy*. Pergamon, Oxford, UK

World Health Organization and Council for International Organizations of Medical Sciences (WHO-CIOMS) 1993 *International Ethical Guidelines for Biomedical Research Involving Human Subjects*.

P. Marshall

Fieldwork in Social and Cultural Anthropology

1. Origins of a Method

Fieldwork is a term which has been employed for nearly a century by social/cultural anthropologists, for their major methodological tool and a profound professional experience that leaves its mark on their lives throughout their careers. It designates a complete concentration on observations in their chosen site of research, for a considerable period of time, usually of no less than six months and up to a few years. Originally, that engagement involved a total separation from the researcher's ordinary life, and his/her immersion into a remote and alien social environment. Anthropological fieldwork is known also as the 'ethnographic method.'

The idea of fieldwork owes its origin to the interests of scientists at the turn of the nineteenth century concerning the evolution of the human species, as much as to the expansion of European colonialism that brought under its rule vast territories inhabited by unknown 'exotic' peoples. But no less, that movement coincided with a growing awareness in the United States about the vanishing cultures of native Americans. The new science of anthropology that intended to satisfy these interests was at its first stages mostly speculative, based on secondary sources of data, collected by missioners, traders, colonial administrators, travelers, adventurers, etc. The practice of pre-fieldwork anthropologists was nicknamed 'arm-chair anthropology.' Its most celebrated, as well as discredited representative, remains James Frazer whose monumental work was all based on library material.

The invention of fieldwork as the characteristic method of anthropology, as well as of its corollary aspect of 'participant observations,' has often been attributed to Bronislaw Malinowski, the British–Polish scholar, who is also considered among the founders of the school of functionalism in anthropology. His claim to fame as fieldworker is based on his long stay, of more than two years, during World

War I, among the Trobriand Islanders. Although he was not the first to employ the method engaging participant observation, nevertheless, his linguistic skills, the meticulous ethnographies he produced, and the introductory chapter on his method of inquiry that opened his most famous work (*Argonauts of the Western Pacific* 1922) have become the hallmark and quintessential model of fieldwork for almost half a century.

The fieldwork projects carried out by Malinowski, his colleagues, and the later generations of younger anthropologists entailed observations of the minutiae of everyday life in the various domains of social behavior and cultural traditions. In sum, it was supposed to record the totality of the social experience in the studied society. The essence of fieldwork was encapsulated, for example, in the following statements by representatives of three generations in British anthropology:

> This goal is, briefly, to grasp the native's point of view, his relation to life, to realize *his* vision of *his* world (Malinowski 1922, p. 25).
> First, in my estimation, is the point that fieldwork in the empirical mode remains the *sine qua non* both for the testing of theory and, what is more important, for making new discoveries (Fortes 1978, p. 24).
> 'Good' fieldworkers are those who are prepared to have good 'trips,' that is, suspend as far as possible their own social conditioning in order to have sensory and mental knowledge of what is really happening around and to them (Turner 1985, pp. 205–6).

In the United States, the method of ethnographic fieldwork was initiated by Franz Boas and his students most of whom had until World War II concentrated their work in the not too far away reservations of Indian tribes. Their trips to the fields of research had rarely involved a long departure from their ordinary life and work. They were also far more dependent on key informants who provided them with data on specific subjects. Consequently, they were not as deeply engaged in the routine of everyday life of the people they studied. The ethnographies by Ruth Benedict are good examples of the achievements and deficiencies of the tradition of short fieldwork trips and reliance on informants. Margaret Mead, probably the most celebrated among Boas' students, who went to Samoa in a style of research closer to the British tradition (1928), was later seriously criticized for her poor work as fieldworker. It was mostly since the mid-1940s that American anthropologists have been equally engaged in intensive fieldwork projects away from home. Compared with the British tradition, American anthropologists, since the early stages, have developed a distinctive orientation in their fieldwork projects, more broadly directed toward inquiries of culture and personality (represented in later years, for example, in the works by Clifford Geertz and his colleagues).

The method of fieldwork was adopted in most Western European countries, though in France the interests of leading anthropologists (Lévi-Strauss in particular) were far more concentrated in oral and textual traditions such as tribal myths of creation. They attributed little importance to the ethnographer's experiences *in vivo*. The ethnographic tradition was also adopted outside Europe and the United States.

2. *The Practice of Fieldwork*

As a first step in their project, fieldworkers in the Malinowskian tradition have always been acutely aware they must develop *rapport* and friendly relationships with the people in their chosen field. They are expected to listen attentively, and witness whatever happens around them in the field, without disturbing the natural flow of events. As they consolidate their position with locals, they usually add information through more active modes of communication (such as initiating conversations, interviews, collecting statistical data, taking photographs, etc.), but they are careful not to instigate a flow of information into preconceived patterns of their own making. They invest much of their time recording this material in their notebooks or computers. The bulk of that store accumulates into their most treasured fieldwork product—'fieldnotes' (to be later transformed into an ethnography). The hallmark of anthropology, fieldwork had also come to differentiate anthropology from sociology. That major difference became more prominent a few decades later when anthropologists were no longer confined to the study of remote Third World societies.

The practice of fieldwork has raised many, as yet not fully resolved, problems of methodology and ethics. Considered a 'method,' it is assumed to maintain a standardized guide for professional behavior, as well as technical rules concerning the collection and recording of data. But throughout the history of anthropology, the craft of fieldwork was practiced, mostly by individuals, who went on their own to explore new research sites where they confronted conditions and constraints not envisioned by their teachers and colleagues. For many years, the teaching of anthropology in many universities did not include specific courses in methodology. The apprenticeship in anthropology was gained through more informal methods such as a close familiarity with the relevant corpus of ethnographies (and particularly those associated with the candidates' departments, for example, 'The Manchester School'), attending presentations and seminars by colleagues who returned from the field, communication with those who had already gone through the experience, and finally, through the good or bad fortunes of the novice on his arrival in the field and his success in establishing rapport with his native hosts. No surprise, fieldwork had been often described

as the major *rite de passage* for students of anthropology. Many anthropologists have been engaged in one major fieldwork project only. But it seems that more ethnographers have gradually become involved in a few fieldwork ventures throughout their career.

Ethnographic work in the various sites of fieldwork, in the early days and today, has been based on an ethos of trust. Anthropologists, mostly engaged in a lone journey, have never been able to implement a procedure of testing the accuracy and truthfulness of fieldnotes reported by colleagues and students. Rarely did fieldworkers choose to return to a field studied already by another ethnographer. The few cases of contested reports have gained much attention, in particular, the controversy between Oscar Lewis and Robert Redfield who studied the same community of Tepotzlan, and Derek Freeman's refutation of Margaret Mead's observations in Samoa. But long-term research, carried out by anthropologists whose continuing interest and commitment to a particular field has been supported by sufficient funds or easy access, also offered a strategy of verification of observations made at earlier trips.

Although anthropologists have been expected to master the language of the society they studied, that prerequisite was rarely fully achieved. Constraints of time and linguistic skills often impelled them to rely on 'informants' as interpreters until they gained a better command of communication in the native languages. The length of stay in the field was usually determined by budgetary as much as by personal circumstances. But an informal norm of one year seems to have dominated fieldwork projects.

How fieldworkers present their role, and to what extent their subjects comprehend the purpose and consequences of their work, remains an unsolved ethical problem. However, some norms that pertain to a universal code of behavior have been institutionalized at an early stage, such as the disguise of the community studied. More rigorously maintained is the tradition of disguising the identity of individual participants. But no regulations were designed in regard to the rights of privacy in recording and disclosing of intimate information collected during fieldwork. Anthropologists were left to make their own decisions concerning ethical issues under the assumption of a deeply felt responsibility toward 'his/her people.' This dilemma was addressed, for example, by Geertz (1968, p. 157): 'A professional commitment to view human affairs analytically is not in opposition to a personal commitment to view them in terms of a particular moral perspective.' In any case, fieldworkers were expected to record in their fieldnotes, without censoring, as much as they could remember from their daily observations and conversations.

During many years of practice, however, anthropologists informed their colleagues, also in a more public fashion, about the circumstances they encountered during fieldwork. The early publications had mostly revealed the personal tribulations of individuals in their respective fields (e.g., Powdermaker 1966, Freilich 1970). But certain specific subjects of methodology had engaged a more rigorous professional treatment. For example, 'the extended case-method' (Gluckman 1958), though mostly a method of ethnographic presentation, suggested a follow-up strategy that stimulated the fieldworker to concentrate his/her observations, thereby revealing the hidden connection between disjoined events. Another theme that gained much attention concerned the various types of informants one might meet at entrance to a new field, and the strategies and precautions one might consider working with them (e.g., Casagrande 1960, Shokeid 1988). Detailed manuals and reference materials for the various aspects of fieldwork have gradually developed into books of methodology so long absent (e.g., Epstein 1967, Wax 1971, Agar 1996, Ellen 1984, Shaffir and Stebbins 1991). A more meticulous approach came to the fore examining the methods of recording and safekeeping of fieldnotes, the most valuable product of fieldwork (Sanjek 1990). Also, conferences dedicated to specific methodological issues have become frequent events (such as 'fieldwork at home').

3. *Transformations in Fieldwork Research*

The idea of fieldwork had gone through immense changes since the days of Malinowski and the generations that established the ethnographic tradition of participant observations in remote, apparently homogeneous societies, clearly identified within territorial and cultural boundaries. The consequences of World War II, the decline of colonialism on the one hand, but the expansion of anthropology in academia on the other, made both the entry and the budgeting of long trips to remote continents an impractical goal for the growing number of students. Instead, many candidates turned their interest to various attractive fields within the borders of their own societies. No doubt, there had been previous successful projects in the United States and Europe of ethnographic studies in fields 'at home' (for example the most celebrated work in an Italian lower class Boston neighborhood by Whyte 1970). But these few research endeavors were unusual cases, in a tradition that was directed to confront 'other cultures' away from home.

The new search for fields close to home seemed at first to involve only a transfer of sites, such as the study of the apparently clearly defined urban ethnic neighborhoods in the United States (e.g., Lewis 1966, Hannerz 1969). However, that novelty demanded a change in the style and intensity of participant observation (Plotnicov 1973). But of more profound consequence, the next change of spatial sites of fieldwork came to also include institutions that offered

specific services to permanent or transient populations (for example, Myerhoff's (1978) work in a community center for aged people). That transformation opened the door to a choice of fields that, hitherto, have not been considered suitable for observation (for example, fieldwork in the scene of gay life, such as by Newton 1972, Read 1980, Shokeid 1995). In these fields, the fieldworker is often unable to observe the participants throughout a wide spectrum of daily activities. Moreover, in the fields of anonymous participants, the observer's ability to record behavior and communicate with his subjects of research had been very limited in scope. These new fields were instrumental in the development of the method and theory of *symbolic interaction*.

The transformation of methods could also be identified in the choice of problem-oriented ethnographic projects. Even when the relevant site was actually a bounded territorial unit, it engaged a modified strategy of fieldwork (e.g., Ginsburg 1989). But another major change of sites entailed a complete eradication of the bounded spatial characteristics of fieldwork. These unbounded sites included observations of participants in particular occupations (e.g., Agar 1986, Hannerz 1998), observing cultural performances (e.g., Heilman 1983), as well as observing the members of dispersed social categories (e.g., Ortner 1997). These fields, no longer designed to record the totality of personal and communal life experiences, have been variously defined: *delocalized*, *multisited*, *postcommunity*, *focus groups*, and *transnational* spaces.

The ongoing change of the spatial research entities have not necessarily transformed some of the major characteristics of fieldwork. Anthropologists in the new fields continue to initiate their observations mostly as individuals, men or women, engaged in their lone projects. It seems, however, there are less couples involved in these projects, compared with the tendency of married anthropologists to take along their families to the more remote field sites.

In recent years, however, some issues not considered before, concerning the fieldworker's feelings and demeanor, have been openly raised in conferences and edited volumes. Anthropologists preparing for a fieldwork project have never been requested to consult about their personality traits and social capabilities to accommodate what might become stressful field circumstances. But since Malinowski's widow published his personal diaries (1967), revealing the prejudices and the emotional stress experienced in the field by the mythical founder of fieldwork, a more relaxed attitude was gaining place in the professional discourse about fieldworkers' personal problems. This attitude has been closely related to the genre of reflexivity that since the late 1970s gained popularity in ethnographic text (e.g., Rabinow 1977). The lone fieldworker was no longer perceived through the image of the infallible man or woman, expected to withstand successfully all obstacles in a valiant test of personality that proves he/she are capable of penetrating the social codes of alien cultures in far away territories.

The new fieldwork sites in Western urban societies contributed to that growing awareness and doubts about some long established taboos that separated the fieldworker from the subjects of his observations (e.g., Kulick and Willson 1995, Lewin and Leap 1996). The border between the observer and the observed that was naturally upheld in the old fields, where the alien researcher had been conspicuously distinguished (by race and other visible marks of personal characteristics), are no longer easily maintained in many of the new sites of fieldwork. The photographs displayed in many ethnographies of the founding generation, showing them in a landscape of exotic trees and simple huts, posing in the company of minimally clothed natives, have become a symbolic emblem for the changing circumstances of fieldwork.

It is probably an irony that by the end of the twentieth century, a century that gave rise to modern anthropology, fieldwork, its most inventive method and major signifier of professional identity, seems to have lost its incontestable paramount position. Critical views of the validity of positivist methodology, as well as the growing influence of postmodernist currents in cultural studies, are challenging the authority of ethnographic texts based on participant observations (e.g., Marcus and Fisher 1986). These intellectual movements, popular in recent years, inspire and legitimize anthropological projects that are based on other types of data (such as textual and mass media creations) which can be collected without the need for fieldwork. It is too early to predict whether the new genre of arm-chair anthropology will have a more lasting impact on the role of fieldwork and its practitioners.

See also: Anthropology; Anthropology and History; Anthropology, History of; Cultural Critique: Anthropological; Cultural Relativism, Anthropology of; Ethnography; Ethnology; Field Observational Research in Anthropology and Sociology; Psychological Anthropology; Qualitative Methods, History of; Symbolic Interaction: Methodology; Thick Description: Methodology

Bibliography

Agar M H 1996[1980] *The Professional Stranger*, 2nd edn. Academic Press, San Diego, CA

Agar M H 1986 *Independents Declared: The Dilemmas of Independent Trucking*. Smithsonian Institute Press, Washington, DC

Casagrande J B (ed.) 1960 *In the Company of Man: Twenty Portraits by Anthropologists*. Harper, New York

Ellen R F 1984 *Ethnographic Research: A Guide to General Conduct*. Academic Press, London

Epstein A L (ed.) 1967 *The Craft of Social Anthropology*. Tavistock, London

Fortes M 1978 An anthropologist's apprenticeship. *Annual Review of Anthropology* **7**: 1–30
Freilich M (ed.) 1970 *Marginal Natives: Anthropologists at Work*. Harper & Row, New York
Geertz C 1968 Thinking as a moral act: ethical dimensions of anthropological fieldwork in the new states. *The Antioch Review* **28**: 139–58
Ginsburg F D 1989 *Contested Lives: The Abortion Debate in an American Community*. University of California Press, Berkeley, CA
Gluckman M 1958[1940] *Analysis of a Social Situation in Modern Zululand*, Rhodes–Livingstone Paper No. 28. Manchester University Press, Manchester, UK
Hannerz U 1969 *Soulside: Inquiries into Ghetto Culture and Community*. Columbia University Press, New York
Hannerz U 1998 Reporting from Jerusalem. *Cultural Anthropology* **13**: 548–74
Heilman S C 1983 *The People of the Book: Drama, Fellowship and Religion*. University of Chicago Press, Chicago
Kulick D, Willson M (eds.) 1995 *Taboo: Sex, Identity and Erotic Subjectivity in Anthropological Fieldwork*. Routledge, London
Lewin E, Leap W L (eds.) 1996 *Out in the Field: Reflections of Lesbian and Gay Anthropologists*. University of Illinois Press, Urbana, IL and Chicago
Lewis O 1966 *La Vida: A Puerto Rican Family in the Culture of Poverty—San Juan and New York*. Martin Secker and Warburg, London
Malinowski B 1922 *Argonauts of the Western Pacific*. Routledge & Sons, London
Malinowski B 1967 *A Diary in the Strict Sense of the Term*. Routledge & Kegan Paul, London
Marcus G E, Fischer M M J (eds.) 1986 *Anthropology as Cultural Critique*. University of Chicago Press, Chicago
Mead M 1928 *Coming of Age in Samoa*. Morrow, New York
Myerhoff B 1978 *Number Our Days*, 1st edn. Dutton, New York
Newton E 1972 *Mother Camp: Female Impersonators in America*. Prentice-Hall, Englewood Cliffs, NJ
Ortner S 1997 Fieldwork in the postcommunity. *Anthropology and Humanism Quarterly* **22**: 61–80
Plotnicov L 1973 Anthropological fieldwork in modern and local urban contexts. *Urban Anthropology* **2**: 248–64
Powdermaker H 1966 *Stranger and Friend: The Way of an Anthropologist*, 1st edn. W.W. Norton, New York
Rabinow P 1977 *Reflections on Fieldwork in Morocco*. University of California Press, Berkeley, CA
Read K E 1980 *Other Voices: The Style of a Male Homosexual Tavern*. Chandler & Sharp, Novato, CA
Sanjek R (ed.) 1990 *Fieldnotes: The Makings of Anthropology*. Cornell University Press, Ithaca, NY
Shaffir W B, Stebbins R A (eds.) 1991 *Experiencing Fieldwork: An Inside View of Qualitative Research*. Sage, London
Shokeid M 1988 Anthropologists and their informants: marginality reconsidered. *Archives Européennes De Sociologie* **29**: 31–50
Shokeid M 1995 *A Gay Synagogue in New York*. Columbia University Press, New York
Turner V W 1985 *On the Edge of the Bush: Anthropology as Experience*. The University of Arizona Press, Tucson, AZ
Wax R H 1971 *Doing Fieldwork: Warnings and Advice*. University of Chicago Press, Chicago
Whyte W F 1970[1943] *Street Corner Society*. University of Chicago Press, Chicago

M. Shokeid

Figurative Thought and Figurative Language, Cognitive Psychology of

1. *A Definition*

Different types of linguistic expressions (metaphors, idioms, proverbs, etc.) form figurative language. Since the seminal work of Lakoff and Johnson (1980 but see Black 1979), the term 'figurative' is also used to denote conceptual entities that contribute to shaping our understanding of the world. Figurative language in general, and metaphor in particular, plays a central role not only in everyday discourse and in linguistic change but also in reflecting and shaping how people think in a broad range of domains (Cacciari 1998).

1.1 *Literal vs. Figurative Language*

When a speaker/reader encounters a sentence, she/he is generally able to judge whether it is intended literally or not. The ease with which people deal with this distinction in everyday discourse belies the difficulty and complexity of the distinction between literal and figurative language in the linguistic and psycholinguistic literature. No universally accepted criteria for discriminating between literal and nonliteral expressions have yet been devised. Many researchers expressed serious doubt as to whether a sharp distinction can (or should) be drawn between these two varieties of language use and proposed to substitute it with the notion of levels of conventionalization of meanings and sentence uses.

Despite the pervasiveness of figurative language, many language researchers still assume that it represents a peripheral issue that language comprehension and production models can easily ignore. For instance, current psycholinguistic models of production are almost silent as to why a speaker would select a figurative expression instead of a 'corresponding' literal expression (if any). Likewise, standard semantic theories still assign a rather peripheral and theoretically uninteresting role to figurative language (with the important exception of cognitive linguistics). Two long-standing assumptions contribute to this error.

(a) Figurative expressions can always be paraphrased literally since they only represent a 'stroke of style' reflecting, as Black (1979, p. 27) ironically stated, the 'incidental pleasure of stating figuratively what might just as well have been said literally.'

(b) The opposition between literal language as the realm of clear shared meanings, and figurative language as the realm of idiosyncratic, obscure meanings.

The cognitive psychology of figurative language does not consider figurative language as an ornament added to everyday straightforward literal language but it views it as a powerful communicative and

conceptual tool. This shift in the view of figurative language, and particularly of metaphor, from a strictly linguistic entity to a conceptual structure has opened a new wave of studies on the mental representation and the comprehension of figurative language. The main stance of this new viewpoint consists of treating metaphor as a bridge that forces the human mind to construct higher-order linkages between the entities referred to in metaphorical assertions. This bridging function is very widespread since metaphors can be used to connect a variety of different domains: abstract entities; abstract concepts to sensory-perceptual experiences and to concrete objects; perceptual experiences belonging to different sensory modalities (as in synaesthetic metaphors).

2. *The Comprehension of Figurative Language*

Contemporary cognitive research on figurative language comprehension began in the 1970s with a predominant focus on metaphor only recently extended to other linguistic expressions, particularly to idioms. Despite the fact that a vast family of expressions form figurative language, not all tropes have been investigated with a theoretical and experimental effort similar to that employed for metaphors and idioms. For instance, proverb comprehension has been scarcely investigated (see Honeck 1997); analogies have almost always been treated as a separate problem generally studied independently from metaphor, even though the two tropes are often confused and sometimes actually difficult to differentiate (see Gentner and Jerziorski 1993). The privilege accorded to metaphor is not surprising if one considers that, since Aristotle, it has been considered the trope *par excellence* often even serving as a collective label for any and all rhetorical figures.

3. *Metaphors as Comparisons*

Metaphor, according to the still most influential definition that comes from Aristotle, is said to be a form of transfer of a name from one object to another. The 'ground' of a metaphor (i.e., its content) can be the set of properties that two concepts share (e.g., in 'Dew is a veil' both are transparent, covering, shimmering). In some cases, it can be a superordinate category that subsumes both concepts (e.g., as in 'the tooth of the mountain' where the mountain peak and the tooth partake of the category 'sharpened forms'). In other cases, it may be necessary to reconstruct the underlying analogy involving both concepts (e.g., as in Virginia Woolf's 'She allowed life to waste as a tap left running').

The comparison view of metaphor has generated a class of comprehension models that can be characterized as feature- or attribute-matching models (for a review, see Cacciari and Glucksberg 1994). Matching models adopt the assumption that in attributive metaphor, of the form 'An A is a B,' the A and B concepts are represented either as sets of features or as positions in an *n*-dimensional semantic space. Regardless of the specific representational assumptions, matching models assume that metaphors are first transformed into comparison statements, and then the features of the vehicle are compared to, or mapped onto, those of the topic.

The two basic assumptions – that metaphors are implicit comparison and that the comparison is derived from a sharing of features – are each problematic, for different reasons, and have been extensively criticized insofar as they do not contribute to definitively clarify what metaphors are and how they are understood. These assumptions are in fact unable to account for several factors: first, that metaphorical comparisons, unlike literal ones, are nonreversible or asymmetrical, that is, they change their meaning if the order of the topic and the vehicle is modified. Second, they do not provide any explicit mechanism for distinguishing between relevant and irrelevant features or properties nor for introducing new features into the mental representation of metaphor's topic. In processing terms, then, they posit that in order to understand a metaphorical assertion, one should first transform it into an explicit comparison. Hence this process should take longer than with a comparable literal expression. On the contrary, a consistent body of evidence suggests that attributive metaphors are understood as quickly as literal sentences when embedded in appropriate contexts, a result which confirms that identical comprehension processes are involved irrespective of language use. More importantly, the feature-matching process may be too limited to capture the full range and impact of metaphors.

Metaphors may have more to do, as Levinson (1983, p. 150) noted, with 'the contingent, factual (real-world) attributes of the referents of the metaphorical focus than with the semantic features that can be claimed to express its meaning.' Metaphor does something more than simply calling our attention to pre-existing similarities between the semantic features of the topic and the vehicle. Whether similarity is pre-existing or is created by the metaphor depends of course on the nature and quality of the metaphorical assertion. In many cases, good metaphors certainly made us perceive resemblances previously unperceived.

4. *Metaphors as Categorizations*

A metaphor can imply a comparison, but it is not in itself merely a comparison. Provided that it is not an implicit comparison, how can we characterize a metaphorical statement? Two theoretical answers have contributed to radically change the cognitive studies on metaphor representation and comprehension: the

Conceptual model (Lakoff and Johnson 1980, Lakoff 1993, Gibbs 1994) and the Property Attribution model (Glucksberg and Keysar 1990, 1993).

These models share the idea that metaphor's use and understanding involves not only linguistic or discourse processes but also reasoning and conceptual processes. Metaphor is based at the same time on the use of acquired categories and conceptual schema and on the creation of new categories and conceptual schemas. These models differ in important respects: first of which the locus of metaphoricity. According to the Property Attribution model proposed by Glucksberg and Keysar, metaphors belong to language use since they are linguistic entities expressing assertions of categorization similar to those expressed with literal language. They are used not simply to call the reader's attention to superficial similarity between concepts but to label categories that have no conventional names of their own. Metaphors create new categories employing the predicate to name the class of things that it exemplifies. When one says, 'My job is a jail,' she/he wants to suggest that his/her job is part of the metaphorical category of situations that are involuntary, unpleasant, and so forth, and to attribute specific properties to the metaphor topic 'job.'

In nominal metaphors, the metaphor topic and the vehicle play distinctive roles in that the topic provides constraints on what is likely to be attributed, while the vehicle provides properties to be attributed to the topic. Metaphorical vehicles are located at two different levels of category abstraction: the literal concrete level (the subordinate level), where they are members of a class of entities, and the more general, superordinate, attributive category level. A metaphor vehicle such as 'jail' can refer simultaneously to a concrete physical structure, and to the metaphorical category of situations that are involuntary, unpleasant, confining, and so forth. The literal 'jail' contains all the properties of the superordinate category 'jail' plus the properties that are specific to a jail *per se* (e.g., having painted walls, being made of concrete, etc.).

According to the Conceptual model proposed by Lakoff and colleagues, the locus of metaphor is not in language, but in the way in which people conceptualize one mental domain in terms of another. 'The general theory of metaphor is given by characterizing such cross-domain mappings' (Lakoff 1993, p. 203). The linguistic expression of metaphor is a simple epiphenomenon of a more general feature, i.e., that we think in metaphorical ways. Lakoff contends that everyday metaphor is characterized by a huge system of thousands of cross-domain mappings that motivates the idea that metaphor is primarily conceptual. It is not just 'a matter of language, but of thought and reason. The language is secondary. The mapping is primary, in that it sanctions the use of source domain language and inference pattern for target domain concepts. The mapping is conventional, that is, it is a fixed part of our conceptual system' (Lakoff 1993, p. 208). People use figures of speech in everyday conversation because they conceptualize much of their experience through the figurative schemes of metaphor, idioms, metonymy, and so forth.

These two models also differ in their theoretical scope: whereas Glucksberg and Keysar want to provide a psycholinguistic model of how people use and comprehend metaphors in discourse, Lakoff's aim is to single out a general account of human thought independently from any processing validity (but see Gibbs 1994).

5. *Idioms: What They are and How They are Comprehended*

The task of defining what an idiom is, how it is acquired and understood is still rather controversial although enough evidence has been cumulated to formulate satisfying models of the comprehension processes (for a review see Cacciari and Tabossi 1993, Gibbs 1994). One of the problems in defining idioms is that idiomatization is a process: a given structure is not idiomatic once and for all, but gradually acquires its idiomaticity. A very general definition is that idioms are composed of freely occurring units whose global meaning cannot be reduced simply to the meanings of these units.

Idioms behave at the same time as words and as larger-than-words chunks. From a processing viewpoint, this has been taken to imply that they were multiword lexical units accessed as such from the mental lexicon. The 'Lexical look up' models of idiom comprehension share the idea that idiom meanings are apprehended by direct memory retrieval, not by linguistic processing. They differ in terms of representational assumptions: according to Bobrow and Bell (1973), idioms are stored in a separate mental idiom list that is accessed and scanned when the literal analysis of a given string fails. This assumption has been rejected on the basis of the robust finding that idioms are understood as quickly as (if not more quickly than) comparable literal expressions. According to Swinney and Cutler (1979), idioms are represented as long words together with all ordinary words of the mental lexicon.

The received view is that idioms are noncompositional strings of words. Despite this, it has been noted that even in the most frozen of idioms (e.g., 'kick the bucket'), the semantic of the single constituent words is still available even though these idioms behave much as words, with their meaning acquired and stored in memory, as for other memorized strings of words (lines of poetry, snatches of songs, etc.) one cannot 'slowly kick the bucket' due to the semantics of 'to kick' that implies an abrupt action (Nunberg 1978, Glucksberg 1993). On more general grounds, this reflects the fact that language comprehension is nonoptional: in spoken language, the

meanings of the words are activated and discarded all the time since understanding occurs automatically without conscious control by the subject.

The recent psycholinguistic literature has shown that during the comprehension of idioms there is activation of the literal meaning of the words forming them, despite what is predicted by the direct look up models. According to the Configuration model (Cacciari and Tabossi 1988, Tabossi and Zardon 1993), idioms are no different from any other type of familiar, memorized string of words. When during a conversation a listener encounters one of such strings (e.g. 'being in seventh heaven'), the word meanings are activated up to the point (and even sometime later) in which the string itself is recognized as an idiomatic configuration. A configuration is made up of the same lexical units that need to be activated during the comprehension of literal discourse. This implies that there is only one processing of an idiomatic string up to the point in which, after the 'keyword' of the idiom is activated, the configuration can emerge. In some idioms, the keyword can appear very early, for instance, immediately after the main verb (e.g., as in 'spilling the beans' or 'breaking the ice'); in other cases it can be located at the end of the idiomatic configuration (as in 'being in seventh heaven' or 'looking for a needle in a haystack'), changing the time course of the comprehension process.

6. *The Role of Semantic Transparency*

The role of syntactic and semantic factors in idioms' interpretation has emerged in the last few decades corresponding to a new wave of studies that has concerned two main areas. The first one concerns the interplay between lexical and syntactic information, that is, the syntactic flexibility of idioms. A consistent body of evidence suggested that the number of syntactic variations an idiom can undergo still retaining its idiomatic meaning may be predicted on semantic grounds: the more the idiomatic meaning can be analyzed, the more flexible the idiom is. The second area has to do with the interplay between lexical and semantic/conceptual systems, that is, with the problem of semantic transparency of idiom meanings. The semantic transparency of an idiom's meaning has been defined as the extent to which the meanings of the words forming an idiom contribute to its overall figurative interpretation. Enough evidence already exists to reject the received view that idioms are semantically empty strings, suggesting that semantically transparent idioms (e.g., 'spill the beans', 'skate on thin ice') are easier to comprehend and to acquire for both children and adults.

These studies argued that adults and children could reliably differentiate among types of idioms (for a review, see Cacciari and Tabossi 1993), for instance among:

(a) Analyzable-opaque: the relation between an idiom's elements and the idiom meaning may be opaque, but the meaning of the individual words can nevertheless constrain both interpretation and use (e.g., 'kick the bucket,' 'by and large,' 'speak one's mind').

(b) Analyzable-transparent: there is a clear semantic relation between the elements of the idiom and the components of the idiom's meaning usually because of metaphorical correspondence between the idiom's elements and the components of the idiom's meaning. The elements of an idiom map onto the components of the idiom's meaning (e.g., in 'spill the beans,' 'spill' corresponds to the act of revealing or letting out, and 'beans' to the materials that had been heretofore concealed or otherwise unknown).

(c) Quasi-metaphorical idioms: here the literal referent of the idiom is itself an instance of the idiomatic meaning. A particularly relevant case or instance is used to name all possible other situations that one might want to describe using the semantic characteristics epitomized by the idiom referent. For instance, 'carrying coals to Newcastle' is simultaneously an ideal or prototypical instance of a useless transfer of something and also a phrase that can refer to any instance of bringing something to a place that has a surfeit of that something (as in 'Selling rice to Japan is like carrying coals to Newcastle'). The literal referents of this type of idiom are relevant and appropriate to their idiomatic meanings.

Quasi-metaphorical idioms may serve the same communicative function of metaphor vehicles in expressions such as 'My lawyer is a snake' or 'My job is a jail.' In these metaphors, vehicles such as 'snake' or 'jail' serve as exemplars of their metaphoric categories—vicious people and confining, unpleasant situations, respectively—and simultaneously as names for those categories (Glucksberg and Keysar 1990). Consider a metaphorically intended term such as 'snake.' It provides a set of candidate properties (many and different) that can be plausibly attributed to a person. Usually, mutual knowledge among the speakers and/or the discourse context provide the relevant knowledge about the person to which such a vehicle ('snake') refers so that only a subset of properties will be attributed to him/her. Metaphorically speaking, a person can be a 'snake' in several respects: she/he can be vicious, dangerous, hit you unexpectedly, and so on. Topics such a lawyer or a supposedly good friend would require, for instance, the instantiation of a different set of attributive properties (or values) insofar as each of these two topics put constraints on what is more likely to be attributed to them.

Linguistic processing and idiom look-up can occur in parallel. Ordinarily, people understand idioms following the basic principles of language comprehension to court. Language is full of noncompositional units that are not idiomatic at all. Consider, for

instance, nominal compounds such as 'carpet sweeper' or 'vacuum cleaner': a carpet sweeper sweeps carpets but a vacuum cleaner does not clean vacuums. As a processing default, people retrieve an idiom as a configuration from the lexicon. Still, the meanings of the words forming it and its underlying structure (be it metaphorical, analogical, metonymical, etc.) are available and might play a role if necessary for discourse purposes. In these cases, people can use their intuitions on the structure of an idiom as well as on the conceptual templates underlying it to elaborate on an idiom's meaning.

7. *Sense Creation and Extension*

The notion that semantic properties of the constituent words of an idiom play an important role in the comprehension and production of idioms has been extended to also account for the problem of meaning extension, i.e., polysemy (Glucksberg 1993). The words forming familiar idioms might become polysemous in that they have at last two meanings: their default context-free literal meanings, and the meanings that are induced by the idiom context. With repeated use, familiar idioms might become able to induce polysemy adding the idiomatic meaning to each word's set of possible meanings. The verb 'spilling,' for instance, might acquire as a secondary meaning that of 'revealing.'

Figurative language is arguably the most powerful source for meaning creation and sense extension. Words evolve additional meanings through it. When such an evolution is consistent and stable enough within a speech community, then a new lexical entry can be added to the list of possible meanings of a word. At that point, lexicographers will officially confirm a new sense by adding it to the dictionary.

Words as well as idioms do not acquire new senses randomly. Most often idioms derive their meaning from metaphors that for some reason emerge in language. According to many authors, this process is not at all arbitrary (for a review, see Cacciari 1998). As Quinn (1991, p. 79) noted, 'the invention of new metaphorical expressions is constrained…by the necessity of finding metaphors that make sense in terms of the cultural models and by preference for metaphors that do a particular good job of this.' Hence, metaphors that become conventionalized (the first step toward idiomatization) presumably represent 'satisfying instantiations' of culturally shared models.

Language comprehension may always involve a mixture of sense selection and sense creation. As Clark and Gerrig (1983) noted, in general speakers consider the information they share with the addressee, evaluate the salience and coherence of potential interpretation against this common ground, and select or create specific interpretations, 'all with deceptive ease' (1983, p. 607). This 'deceptive ease' reflects human's ability to simultaneously process information coming from different domains: lexical, syntactic, semantic, pragmatic, and so forth. Figurative language, as an old tree, extends its roots across these domains, taking its nourishment where it springs. That's why it is such a crucial hinge among different language systems.

See also: Creativity and Cognition; Cognitive Development in Childhood and Adolescence; Comprehension, Cognitive Psychology of; Language Acquisition; Language and Thought: The Modern Whorfian Hypothesis; Scientific Reasoning and Discovery, Cognitive Psychology of; Semantic Similarity, Cognitive Psychology of; Sentence Comprehension, Psychology of; Word Meaning: Psychological Aspects

Bibliography

Black M 1979 More about metaphor. In: Ortony A (ed.) *Metaphor and Thought*. Cambridge University Press, Cambridge, UK, pp. 19–43

Bobrow S, Bell B 1973 On catching on to idiomatic expressions. *Memory and Cognition* **1**: 343–6

Cacciari C 1998 Why do we speak metaphorically? Reflections on the functions of metaphor in discourse and reasoning. In: Katz A, Cacciari C, Gibbs Jr R W, Turner M (eds.) *Figurative Language and Thought*. Oxford University Press, Oxford, UK, pp. 119–57

Cacciari C, Glucksberg S 1994 Understanding figurative language. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, New York, pp. 447–77

Cacciari C, Tabossi P 1988 The comprehension of idioms. *Journal of Memory and Language* **27**(6): 668–84

Cacciari C, Tabossi P (eds.) 1993 *Idioms. Processing, Structure and Interpretation*. Laurence Erlbaum, Hillsdale, NJ

Clark H H, Gerrig R J 1983 Understanding old words with new meanings. *Journal of Verbal Learning and Verbal Behavior* **22**: 591–608

Gentner D, Jeziorski D 1993 The shift from metaphor to analogy in Western science. In: Ortony A (ed.) *Metaphor and Thought*, 2nd edn. Cambridge University Press, Cambridge, UK, pp. 447–80

Gibbs W R 1994 *The Poetics of Mind. Figurative Thought, Language and Understanding*. Cambridge University Press, Cambridge, UK

Glucksberg S 1993 Idiom meaning and allusional content. In: Cacciari C, Tabossi P (eds.) *Idioms. Processing, Structure and Interpretation*. Erlbaum, Hillsdale, NJ, pp. 3–26

Glucksberg S, Keysar B 1990 Understanding metaphorical comparisons: beyond similarity. *Psychological Review* **97**(1): 3–18

Glucksberg S, Keysar B 1993 How metaphor works. In: Ortony A (ed.) *Metaphor and Thought*, 2nd edn. Cambridge University Press, Cambridge, UK, pp. 401–24

Honeck R P 1997 *A Proverb in Mind: The Cognitive Science of Proverbial Wit and Wisdom*. Erlbaum, Mahwah, NJ

Lakoff G 1993 The contemporary theory of metaphor. In: Ortony A (ed.) *Metaphor and Thought*, 2nd edn. Cambridge University Press, Cambridge, UK, pp. 202–51

Lakoff G, Johnson M 1980 *Metaphors We Live By*. The University of Chicago Press, Chicago

Levinson S C 1983 *Pragmatics*. Cambridge University Press, Cambridge, UK

Nunberg G 1978 *The Pragmatics of Reference*. Indiana Linguistic Club, Bloomington, IN
Quinn N 1991 The cultural basis of metaphor. In: Fernandez J W (ed.) *Beyond Metaphor. The Theory of Tropes in Anthropology*. Stanford University Press, Stanford, CA, pp. 56–93
Swinney D A, Cutler A 1979 The access and processing of idioms. *Journal of Verbal Learning and Verbal Behavior* **18**: 523–34
Tabossi P, Zardon F 1993 The activation of idiomatic meaning in spoken language. In: Cacciari C, Tabossi P (eds.) *Idioms. Processing, Structure and Interpretation*. Laurence Erlbaum, Hillsdale, NJ, pp. 145–62
Wasow T, Sag I, Nunberg G 1983 Idioms: An interim Report. In: Hattori S, Inoue K (eds.) *Proceedings of the XIIIth International Congress of Linguistics*. CIPL, Tokyo, pp. 102–5

C. Cacciari

Film and Video Industry

The production, and distribution of cinema occurs throughout the world but since the 1920s has been dominated by eight or less giant Hollywood studios. These vast multinational enterprises create films that are presented in virtually every country in the world, frequently capturing a major share of the business despite competition from locally made and distributed motion pictures. Almost continuous technological change—led by the Hollywood oligopolists—did not lessen Hollywood's industrial power, and so all other film industries sought to ape Hollywood's continuous success. The coming of sound and color, and the innovations of pay-TV and home video (played on VCRs) simply extended Hollywood's influence, and increased its profits. Analysis of the Hollywood studio system has led to praise or damnation of this omnipresent, inescapable, and highly influential industry.

1. *The Film and Video Industry*

A handful of companies have defined the Hollywood film industry, dominating the US and world markets. They have weathered a world war, and a Great Depression, innovated wide screen and color technologies, made peace with television, learned to exploit home video, and were more powerful at the end of the twentieth century than ever before.

The coming of sound solidified control by five studios over the world cinema market: Paramount Pictures, MGM, Twentieth Century Fox, Warner Bros., and RKO. During the 1930s and 1940s by controlling the key cinemas in the downtowns of nearly all major US cities, the 'Big Five' took in nearly all box-office take. Only after they granted their own theaters first-runs, did they then permit smaller, independently owned theaters to scramble for the remaining bookings months, or even years, after a film's premiere.

Paramount represented the most profitable, powerful, and traditional business-like Hollywood company. More than any other member of the Big Five, Paramount relied on its chain of more than 1,000 theaters to maintain its corporate might. Not surprisingly, a former theater operator, Barney Balaban, stood at the top of this corporate colossus, hiring more lawyers and MBAs than movie stars.

Metro-Goldwyn-Mayer (MGM) technically functioned as a successful unit of Loew's, a theater chain concentrated in the Eastern USA. MGM, more than any other studio, relied on the star system, producing films from its massive Culver City, California-based 'factory' of 27 sound stages (for interior shots), and a 168-acre back lot (for outdoor scenes).

Twentieth Century Fox ranked behind Paramount and MGM with neither as large a theater chain or studio facility. Studio boss Darryl F. Zanuck used what he had to authoritatively produce films for world distribution. His decisions were final, frequently bellowing to all subordinates: 'Don't say yes until I've finished talking!'

Warner Bros. represented the only family-run operation among the major movie studios, with eldest brother Harry as President, middle brother Abe supervising worldwide distribution, and baby brother Jack heading studio production. The innovation of sound pushed Warners into the Big Five, yet financially the company struggled through the 1930s and 1940s.

Radio-Keith-Orpheum (RKO) was formed so that radio's RCA could sell its sound equipment to Keith–Orpheum theaters. While the studio created many memorable films—such as *Citizen Kane* and the Astaire–Rogers musicals—RKO's only regular profitability came from the distribution of Disney animation. TV killed RKO.

The Big Five tolerated some competition—from Universal Pictures, Columbia Pictures, and United Artists—as the 'Little Three' depended on the Big Five for access to the top theaters in the USA. The suburbanization of the US, and the coming of television allowed Universal, Columbia, and United Artists to prosper because the advantage of owning theaters disappeared because of antitrust action by the US government. The former Big Five, save RKO, tried to re-invent themselves.

Warner Bros. began its transition into what became Time Warner when in 1956 the founding brothers sold out, and a series of new owners embraced television production. In 1969, Kinney National Services, a New York conglomerate, purchased Warners, and Steven Ross, son-in-law of Kinney's founder, began to fashion the modern Hollywood-based prototypical multinational media conglomerate. In 1989 Ross merged Warner with Time, and the largest media

company in the world was created. A decade later AOL took control.

Paramount stubbornly hung on until 1966 when Charles Bluhdorn's Gulf+Western Industries conglomerate, after a 2-year battle, took over. Through the 1970s and 1980s the 'new' Paramount thrived as part of a classic conglomerate, but in another 2-year struggle, was taken over in 1993 by another media conglomerate, Sumner Redstone's Viacom, Inc.

Twentieth Century Fox began its reinvention in 1956 when longtime studio boss Darryl F. Zanuck resigned. His replacements struggled, but did lead the way to licensing feature films to major television networks. NBC's 'Saturday Night at the Movies' began TV's showcasing of feature films. The surprise hit *Star Wars* (1977) moved Fox to the top of the Hollywood studio hierarchy, and within a decade led to the studio's purchase by Australian media mogul Rupert Murdoch. From the Fox studio base, Murdoch created the Fox TV network.

MGM never was able to reinvent itself, even with its merger with United Artists. The two other members of the Little Three did far better. Columbia went through several ownership changes (including during the 1980s when Coca-Cola owned it), but it was not until 1989 that stability came when Japanese electronics giant Sony purchased Columbia. During the 1990s Sony never was able to exploit its Hollywood studio to sell more VCRs, but did learn to create vast profits.

Universal also struggled until 1962 when the MCA talent agency acquired it. Thereafter, under Lew Wasserman, Universal ascended to the top of Hollywood movie and television studio hierarchy as Wasserman innovated the made-for-TV movie, and the modern film blockbuster—*Jaws* (1975). Wasserman retired in the early 1990s, and in 1995 Canadian liquor giant Seagram took control.

Disney joined Warners, Paramount, Fox, Sony, and Universal to form the 'Big Six' of the 1990s. Beginning in 1984, Michael Eisner leveraged Disney's weak position into major studio status, a significant change in an industry with so much long-term stability. Indeed, how had this small number of companies survived and thrived for so long—with no real newcomers joining their ranks?

First, even with all the social technological and industrial changes, the major companies alone continued to be able to distribute films around the world. At considerable expense, all maintained offices not only in the USA, but in every region of the world. In most countries around the world, owners of theater chains knew the Hollywood majors alone could continuously deliver popular films, and so even theater owners outside the USA preferred Hollywood films to native productions. The Hollywood studios' record at the box office satisfied theater owners—a conservative lot with most of their assets invested in real estate. Economies of scale prevented rivals from building and maintaining equal world distribution networks.

Second, modern Hollywood fashioned the blockbuster strategy to turn television into a friend, not a foe. Through the final quarter of the twentieth century, the Hollywood oligopoly learned to generate profits from creations that premiered with blitz TV advertising, which in turn was married to a multirevenue generating machine to sell musical sound tracks, novelizations of the stories, 'action figures' from movie characters, and endless sales on video. Each movie aspired theatrically to become a smash hit because if it succeeded then the feature turned into a 'product line,' designed to fill all entertainment needs and desires from toys to theme park rides to campaigns to sell more McDonald hamburgers.

The blockbuster strategy in turn led to ever escalating costs of production so it became more difficult for rivals to join the exclusive Hollywood club. By the close of the twentieth century the average Hollywood feature cost in excess of $50 million to produce, and half again that amount to publicize. *Titanic* doubled those average cost figures, but then earned in excess of $2 billion in revenues. Potential filmmakers knew that if they ever aspired to have a world audience see their creation and to attain undreamed of wealth, they needed to be 'green lighted' by one of the Hollywood major studios.

Alone, Hollywood could keep the revenues flowing year after year—particularly from home video. While theaters still premiered new films, it was 'downstream' on video where the bulk of the monies were generated. The theater functioned as the 'voting booth' where hits were made, and necessary publicity generated. But once the film had fully milked the box-office it went into release on pay-TV and home video that created more than two-thirds—in the long run—of the average film's revenues. Hollywood executives carefully planned—in classic price discrimination fashion—to drain all possible revenues from each of these release venues, only releasing it to home video when they reasoned all the revenues from theatrical and then pay-TV had been procured. Each window in this sequence was an exclusive; a new window opened only when all value of the previous window had been captured. This exploitation of video represented Hollywood's newest advantage in the film industry.

2. *The Rest of the World*

While the Hollywood majors maintained their economic power and influence, filmmakers outside Hollywood long struggled to find their place in their national cinemas, often requiring the assistance of their governments. Only sporadically have large nation-states—France, Italy, Japan, and India in particular—been able to mount a studio system as an alternative and serious rival to Hollywood in their own countries, let alone elsewhere.

Globally, only in the rare nation does Hollywood not capture more than half the business. The Hollywood majors maintain offices around the globe, where their representatives maintain constant contact with the heads of the dominant theater chains, pay-TV outlets, and home video retailers. Hollywood's regular production of hit films provides a strong incentive for foreign theater, pay TV, and home video companies to consistently deal with them rather than take a flyer with some new company—even a native company.

But other nations did try. During the 1920s European nation-states—led by the United Kingdom, France, and Germany—sought to build their own studio systems, and were successful for short periods of time, until their leading filmmakers left for Hollywood. During the 1920s, the new Soviet Union tried to build an alternative national cinema industry under Communism. The 1930s ended these experiments. At the close of World War II, the focus shifted to state sponsored television, but still in Japan and India studio systems did succeed, even while conceding Hollywood a vast share of the monies paid for cinema attendance.

After the advent of the blockbuster, it became harder for foreign governments to subsidize a native cinema industry. Most chose to focus on television instead. But Germany and Australia, for example, did for a time in the 1980s create enough different types of popular productions long enough to be discovered as the latest 'new wave.' Japan had long had a studio system, but the coming of television and the takeover of Hollywood studios by Sony, and temporarily, Matsushita, ended the importance of that native industry. But India, for example, was still able to maintain native popular cinema, complete with a star system and defined genres like Hollywood. But these films were intended for domestic consumption, with minimal export possibilities. Community viewing in cities and poor areas alike still flourished.

Hollywood has long maintained an influential trade association—the Motion Picture Association of America noted for its ratings of films, but far more influential around the world for lobbying to smooth the way for international distribution. The MPPA lobbied the US State Department to keep cinema trade 'free' to permit the Hollywood oligopoly to continue pulling billions in revenue from the foreign [read: non-US] market. Foreign governments, on the other hand, resented US pressure, and formally and informally sought to protect the national film industries, and their cultural integrity and identity.

3. *Industrial Analysis*

The common thread of Hollywood power and influence, and its growing presence around the world, have led to much analysis of the Hollywood film industry. Three different interpretations have dominated.

3.1 *Marxist Industry Analysis*

Marxist critics have long seen this influential cultural industry as a classic example of monopoly capitalism, focusing on how these long lived corporations colluded to devise ways to maintain their power and cultural imperialism. While Wall Street celebrated Twentieth Century Fox and Paramount's co-financing of *Titanic*, Marxist critics would see this as yet another example of studios working together—not competing—to maximize profit. Since there were at most eight players, they did not have to worry about going out of business, but simply seeking how much their exploit and keep in profits.

The long-term propensity toward concentration of ownership, endemic to monopoly capitalism, was simply exemplified by the Hollywood motion picture industry. That is what history should teach us, and that is how Hollywood has behaved even in the face of new technologies. Marxists focus on the expansion of giant corporations and the spread of their influence to television, music, toys, theme parks, and other entertainments. Market concentration and anticompetitive behavior describe the operation and ownership of the Hollywood movie industry.

Marxist industry analysis usually focuses on international distribution, seeing the trend toward globalization as evidence of cultural and economic imperialism. While Hollywood publicity focuses on the production of films—its stars, stories, and special effects—international distribution has always been a key to corporate longevity. No other national film industry has been ever been so far-reaching, endangering indigenous, nonprofit seeking culture. They correctly emphasize that Hollywood taught the mass culture businesses to focus on and exploit the advantages of globalizing as seen in the work of Wasko (1994) and Guback (1969).

3.2 *Neoclassical Industry Analysis*

Neoclassical economists emphasize that movies face competition from all forms of mass culture. While there are and have been a few very powerful studios, consumer sovereignty means that no Hollywood company can ever force a customer to pay to see a film—even *Titanic*. These analysts focus on the forces and interplay of supply and demand, seeing the movie theater as a classic voting booth, where top grossers win and unpopular films are pulled almost immediately. They stress that there is a great deal of product diversity that permeates the entire landscape of the mass media. This is why genres—from science fiction to westerns, from dramas to comedies—ebb and flow in relative popularity. With the rise of the VCR, for example, the choice of titles to rent or buy needs to measured in the thousands. Like the Marxists who come to their analysis with a predetermined

anticapitalist assumption, these neoclassical analysts enter the debate with a procapitalism assumption as seen in the work of Litman (1998).

3.3 Industrial Organization Analysis

Industrial organization economists start with neither pro- or anticapitalistic assumptions, and instead seek first to define 'who owns the media,' then analyze what economic conduct emerges from that industrial structure, then detail how the industry performs given specified criteria, and finally recommend possible public policy corrections to make the industry do 'better.' Industrial organization economists have long recognized the host of advantages that enabled them to maintain their considerable economic power and keep out the competition.

For example, cross-subsidization enabled a Hollywood corporation—with interests in a number of media markets—to extract profits from a thriving area to prop up another less financially successful area. Single-line corporations do not have this luxury and so aspiring single-line Hollywood operations invariably fail as Gomery (1986) discusses.

Or, to take another example, reciprocity enables a diversified member of the Hollywood major studios to choose to whom they will sell and then only deal with those companies that cooperate with other units of the media conglomerate. For example, Universal Pictures might not sell movies to Time Warner's HBO unless Time Warner's cable franchises book Universal's (partially owned) USA television network.

Horizontally, year after year, the Hollywood oligopoly has generated considerable profits from a wide spectrum of mass-media enterprises, including theme parks, music, and television production. There is no such thing as a single-line major movie company. The major Hollywood companies also control the process of production and distribution—as they long have—but also more and more of the increasing avenues of presentation, from pay-per-view channels to home video to digital home viewings.

Bibliography

Alexander A, Owers J, Carveth R 1968 *Media Economics*. Erlbaum, Mehwah, NJ

Balio T 1985 *The American Film Industry*. University of Wisconsin Press, Madison, WI

Bruck C 1992 *Master of the Game: Steve Ross and the Creation of Time Warner*. Simon & Schuster, New York

Conant M 1960 *Antitrust and the Motion Picture Industry*. University of California Press, Berkeley, CA

Gomery D 1986 *The Hollywood Studio System*. St. Martin's, New York

Gomery D 1992 *Shared Pleasures: A History of Movie Presentation in the United States*. University of Wisconsin Press, Madison, WI

Guback T 1969 *The International Film Industry*. Indiana University Press, Bloomington, IN

Kindem G 1982 *The American Movie Industry*. Southern Illinois University Press, Carbondale, IL

Lardner J 1987 *Fast Forward: Hollywood, the Japanese, and the Onslaught of the VCR*. Norton, New York

Litman B 1998 *The Motion Picture Mega-industry*. Allyn and Bacon, New York

Morin A 1996 *Film Policy: International, National, and Regional Perspectives*. Routledge, London

Neale S, Smith M 1998 *Contemporary Hollywood Cinema*. Routledge, London

Wasko J 1994 *Hollywood in the Information Age*. University of Texas Press, Austin, TX

Wyatt J 1994 *High Concept: Movies and Marketing in Hollywood*. University of Texas Press, Austin, TX

D. Gomery

Film: Genres and Genre Theory

Genre, a term notoriously difficult to pin down, refers to a type of film or other art object where instances of the type share similarities in form and style, theme and content, as well as communicative function. A film genre is thus based on a set of conventions that influence both the production of individual works within that genre and audience expectations and experiences.

1. The Study of Film Genres

The concept of genre has been important since the beginning of the history of film, as a category for the industry and thus for the production, marketing, and distribution of films; for film critics and the academic, esthetic, historical, and theoretical analysis of film; and for audiences as a framework for the selection and experience of films. The study of film genres is therefore placed in a triangular structure relating to the institutional, social, and cultural context of films; to the esthetic, formal, thematic, and stylistic dimensions; and to audiences and reception understood both in an empirical social and psychological sense.

The development of studies of film genre from classical to modern film theory have shifted between focus on different aspects of this triangular structure. Studies of particular film genres, e.g., the Western or the family melodrama, may stress the social and cultural aspects of genre, as ideological or ritual expressions of American history, values, and mythology, or as part of the Hollywood studio system and the institutional development of production norms. Or studies may stress the formal, narrative, visual, iconographic, and thematic structure of particular groups of films in order to find similarities and variations and historical changes and developments.

But studies of genre have also focused on psychological aspects of genre such as gender and the imaginary, using Freudian, Lacanian, and feminist terminology or, as it has been the case in recent thinking about genre, from a cognitive and emotional theory of basic prototypes of film genres. But many theories of genre do not stress just one part of the triangle but clearly indicate that genres are unstable and fuzzy categories. Rather than defining film genres as sharp, distinctive, and essential categories, modern genre theory tends to define genres as prototypes, and to see the establishing of genres as processes related to institution, text, and audience, and historical changes.

1.1 Auteur and Genre

Although film genres are found as categories from the beginning of film history, used by both production companies and film critics, the systematic theoretical and historical reflections on film genres were developed in the 1950s as European auteur theory met US genre films. The early theories of film were often very strongly tied to a more individualized and esthetic understanding of film as art, and therefore often viewed mainstream cinema and genres as commercially dominated formulas, that did not allow artistic freedom and innovation for the auteur. But although this discussion is often repeated even today, film genre theory in general has explicitly questioned this normative division between an industrialized product and mainstream cinema on the one hand and film art and the auteur concept on the other.

In the 1950s French critics centered around *Cahiers du Cinéma* and the New Wave stressed the variations and auteurism of classical genre films in their writings on classical Hollywood cinema. Bazin, despite his central role in the formulating of a realist film theory and an auteur theory, was also engaged in describing the special qualities of classical US genre-films, such as the Western. In his two seminal articles (Bazin 1971a, 1971b) published in French in 1953 and 1954, he discussed the Western, and especially John Ford's Westerns as US film art *par excellence*. In these articles, Bazin saw the Western as a visual and narrative embodiment of American history and myths, and in his definition of what constitutes a genre, he focused on both thematic elements and elements of cinematic technique and style, narrative structure, and *mise-en-scène*. He also initiated later discussions on the historical transformation of the Western, defining both its pure classical forms and newer forms, which he called 'superwesterns' and 'novelistic Westerns.'

Bazin's articles mark a transition from a normative esthetic notion of genre films as just degraded, commercialized, and industrialized art. He combined auteur perspectives and genre perspectives into a more general esthetic and cultural perspective on all forms of film. This was also the intention behind a number of later US contributions to the general understanding and re-evaluation of US film genres. Braudy (1976) not only explicitly criticized the division between film as art and film genres, but also claimed that mainstream genres can express broader cultural, social, and esthetic values than classical works of art can. Buscombe ([1970] 1995) takes a broad discussion of literary genre theory as his starting point, and specifically uses Wellek and Warren's distinction in *Theory of Literature* ([1949] 1956) between 'outer form' and 'inner form' and their cinematic equivalents, thus opening a more structural and esthetic specification of genre definition.

1.2 Semiotics and the Ritual and Ideological Turn

The move from auteur to genre from the late 1960s was sustained and further developed through the influence of semiotics, structural anthropology, and literary genre theory to a broader ritual and ideological approach to the study of film and film genres. The ritual approach involved both a more systematic analysis of cinematic texts and a form of institutionalized practice expressing and negotiating fundamental cultural values. The ideological perspective involved a critical study of cinema and cinematic genres and the reproduction of dominant forms of ideology from sociological, Marxist, Freudian, and feminist perspectives. The two approaches do not necessarily exclude each other but in the ritual approach there is a tendency to see the relation between genre and audience as one of negotiation and exchange, whereas the ideological perspective more often stresses the principle of dominance.

In Cawelti's broad ritual study of popular literature and film (Cawelti 1976) and in Wright's classical ritual and structural analysis of the Western (Wright 1975), genres are seen as fundamental story types that represent and embody cultural images and myths with both universal and historical elements, and which are recognized by the audience. Genres and the act of genre reading balance between mimesis and escapism. Cawelti's idea is that genre formulas work because they are able to express and negotiate historical conflicts and mental and cultural oppositions and contradictions in a way that is accepted, understood, and appreciated by a large audience, not just dominated by ideology. He rejects deterministic theories in favor of symbolic or reflective theories, where assimilation, negotiation, and change take place but without denying that formula stories reflect and affirm dominant forms of thought and attitudes.

Cawelti defines five universal story types that can be used and combined in different ways: adventure, romance, mystery, melodrama, and stories of alien beings and states. In Schatz's (1981) study of classical Hollywood genres he talks about genres in a fundamental semiotic way in which he sees 'film genre as

a specific grammar or system of rules of expression and construction and the individual genre film as a manifestation of these rules' (Shatz 1981, p.19). Basic film genres form the *langue* and individual genre films form the *parole* of genre filmmaking and film language, and he defines the grammar in terms of a genre's iconography, characters, settings, and plot structure. Shatz's analysis is a study of some of the most stable film genres in the classical Hollywood system, the individual films, and the directors, but he also points to two very basic types of supergenres: genres of order (Western, gangster, detective) and genres of integration (musical, screwball comedy, family melodrama) with two very different narrative strategies, settings, types of hero, and solution models of conflicts that either express violent elimination and male dominance or emotional, familial integration and couple or female orientation.

Shatz's approach points to problematics also addressed in the more ideological and feminist genre theories developed at the same time, but with a background in French film semiotics or the influential UK journal *Screen*. One of the seminal articles articulating the ideological approach is Comolli and Narboni's ([1969] 1992) article, printed originally in *Cahiers du Cinema* and in English in *Screen* (1972). Every film is political and ideological, they state, in so far as no film or filmmaker can escape the commodity form and the capitalist economy. According to Comolli and Narboni, filmmaking at large is nothing but an expression of the prevailing ideology, which is working behind the back of both the filmmaker and the audience. But the article also defines a kind of ideological genre taxonomy, in which films are defined in relation to whether and how they try to oppose the dominant ideology. Among the seven genres defined are: the majority of films, which are just unconscious instruments of ideology; films that have an explicit critical content and or/form; films that have an ambiguous secret message and form; and films like documentary films, that are seen as only illusionary breaks with the dominant ideology.

Ideological genre analysis found dominance of ideology disguised behind classical realism or inscribed in the apparatus of cinema, and therefore viewed the audience as rather passive. However, ideological analysis did look for both a countercinema and counterstrategies to the dominance of ideology in genre film and mainstream cinema. But whereas the ritual approach saw qualities in the mainstream film and its relation to the audience on a more direct level, the ideological approach often investigated ambiguities and contradictions in genre films dominated by ideology. Writers belonging to the ideological and semiotic project (e.g., Wood [1977] 1995 and Klinger [1984] 1995) therefore also found elements of artistic personality still visible in genre films and often described subversion of the dominant tendencies in mainstream film.

1.3 Gender and Genre

The story of genre theory and gender starts with the feminist project based on the ideological and Freudian-Lacanian-inspired genre theory that sought to find the dominant. Mulvey's ([1975] 1992) much quoted essay combined elements from all aspects of the ideological genre theory, but with the aim of demonstrating the male dominance of the feminine in mainstream Hollywood. She pointed to the male gaze and the feminine characters as the looked-upon object, and she also described the asymmetric relation of active/passive roles. But Mulvey's article did not just open the feminist critique of male dominance, but also the demand for another esthetic, a feminist countercinema. Eventually it also led to the investigation of subversive female readings of classical genre films, most notably 'women's films,' e.g., melodrama or film noir as it is very broadly demonstrated in Gledhill's (1987) anthology on melodrama and women's films and Kaplan's ([1978] 1998) on women and film noir, in which prominent feminist film genre critics, e.g., Mary Ann Doane, Tania Modleski, and Annette Kuhn, figure. Later feminist writers have directly entered more action-oriented and excessive 'body genres,' such as Williams studying pornography, horror, and melodrama (Williams 1989) or Clover's (1992) thorough study of the horror genre. And just as feminist criticism has opened towards seemingly more 'masculine genres', gender studies of film genres have also included studies of gay and lesbian films or gay and lesbian aspects of mainstream genre films.

1.4 Genre, Reception, and History

In literary genre theory (Todorov [1970] 1975) the distinction between theoretical genres is defined in more general terms by criticism and theory, and historical genres are defined as genres actively used and recognized by critics and readers. This distinction in a way is a starting point for Neale's (1980, [1990] 1995) and Altman's ([1984] 1995, 1999) contributions to a genre theory. Neale (1980) represents screen theory as combining ideological and semiotic-esthetic aspects. But he expands the concept of genre to a very multifaceted process. Genres are both social and textual institutions and, according to Neale, they work to create a coherence and narrative closure, which positions the viewer safely inside a well-known structure: 'a genre is, simultaneously, a coherent and systematic body of film texts, and a coherent and systematic set of expectations' (Neale 1980, p.37). Altman ([1984] 1995) also clearly continues the semiotic project of genre definition, but at the same time seeks a merging of the ritual and ideological approach and the more general description of genres as a grammar and language. His suggestion in the article is

thus to combine the semantic approach to genres, defining its 'building blocks', and a syntactic approach defining the actual structures in which the semantic elements are placed. But at the same time he argues for a stronger historical complexity in the study of genres, and for a stronger emphasis on the study of genres in historical contexts and to actual interpretative communities.

This line of argument focusing on reception and the historical process of 'genrification,' and genre as a kind of 'game' between critics, producers, and audiences in a concrete historical process, is strongly underlined in Neale ([1990] 1995) and Altman (1999). Taking some historical examples of how different early film genres were defined in industrial and critical terms and how long it took for well-established genres to stabilize as both institutional and audience categories, Neale argues for genre studies with more real historical integrity, actually studying the discourses on genre used in a historical context, rather than only constituting general, theoretical genre concepts. Along the same lines, Altman focuses very much on the unstable and fuzzy nature of genres and the differences in audiences' readings of individual genres and individual films. He adds a pragmatic dimension to his earlier semantic-syntactic dimensions to include the user-contexts aspect. So key terms in this reception-oriented historicizing of film genre theory are different readings, multiple users, and conflicting understandings of genres, instead of genres understood as relatively stable categories based on both general textual similarities and a social and cultural contract or common understanding.

1.5 Cognition, Emotion, and Genre

A move in the opposite direction back to a more general definition of genres can be found in the strong new trend toward cognitive and pragmatic studies of genres in the 1990s. The earlier semiotic and ritual sociocultural approach to genres defined dominant structures based on more general patterns of mental, generic procedures, that partly explain why mainstream genres are so strong and dominant. This point of view was based on structural arguments that were not very firmly anchored in a psychological dimension. Cognitive-pragmatic genre studies also define very basic patterns and schemata in our active construction of narrative forms and basic generic formulas. But the cognitive-emotional theories focus on more internal psychological dimensions, but do not exclude a pragmatic-historical dimension, since all such mental aspects of communication are situated and embodied in a specific historical and communicative situation. The cognitive-emotional genre theory also opposes the dominant Freudian, Lacanian, or Jungian-inspired psychology in the screen tradition of genre studies. Important contributions to this new cognitive-emotional genre theory is Bordwell's (1985) description of different narrative prototypes and the way they cue the spectator and Carroll's studies of horror (1990) and his later article on emotion and genre (Carroll 1999), in which melodrama, horror, and suspense are described from a cognitive-emotional perspective.

But the most ambitious attempt to formulate a cognitive and emotional theory of genre is that of Grodal (1997). Grodal defines a new typology of very basic genres based on the viewer's active–passive relation with the film, the kind of distantiation or identification involved, and the typical moods and emotional responses activated during viewing: associative lyricism, canonical narratives of action, obsessional paratelic fictions, melodramas, horror, schizoid fictions, comedy, and metafiction. Behind these basic forms we then find a number of more concrete, historical subgenres. The basic assumption of the book is that genres of visual fiction are mental structures that to a large degree are modeled on the basis of a simulated reality of actions and emotions to which we react with much the same experiences and capacities as in real life experiences. In his book Grodal describes both the typical embodied flow, during the experience of visual fiction with reference to cognitive-emotional psychology, and the reason for the strength and popularity of the prototypical genres of visual fiction and their relation to different mental and emotional activities and experiences. Grodal's book is, of course, not strongly oriented toward the cultural and historical aspects of moving images, but there is no essential opposition between a cultural, historical, and stylistic approach to visual fictions and Grodal's position, as long as one realizes that certain aspects of the experience and processing of moving images cannot be described as culturally constructed in any short-term sense.

2. Film and Genres

How many film genres are there, and how do we define them and distinguish one genre from another? This is the basic question raised by film genre theory, and this is also the question raised in more concrete, historical analysis of particular film genres and their development. Theoretically there are at one end theories that try to define genres as relatively stable and accepted narrative, stylistic, and thematic structures that transcend individual films and clearly influence and determine both production and audience response; at the other there are theories that stress the unstable and processual nature of genres and the fuzziness of genre. Some theories argue for few and very basic generic prototypes with a variety of possible subgenres, that may vary and change historically, whereas practical criticism and the history of film genre categories often show that you can construct genres on the basis of many criteria, for instance on content and dominant

type of narrative action (war films, porno films), on specific stylistic features (film noir, musicals, animated films), on the basis of the expected audience (women's films, children's films), on the basis of emotional response (comedy, melodrama), etc.

Although genres are often related very clearly to the dominance of US cinema worldwide, national variations of genres are also very important and often have more universal influence, e.g., the Asian martial arts films or the British heritage film or costume drama. But although historical, national, and other variations are very common and the number of actual genres defined in some critical context is probably very large, then some basic narrative modes and generic fomulas are stronger and more permanent and universal than others.

2.1 Basic Film Forms and Universal Formulas

Kracauer (1960) defined two basic types of film: narrative films (story films), and nonnarrative films (nonstory films) which he divided into 'films of fact' and experimental/avant garde films. He considered narrative film to be the most important and basic form. In a similar way Williams (1984) has suggested that before we talk about specific subgenres, we should perhaps talk about only three, very basic or principal film forms: narrative film, avantgarde/experimental film, and documentary film. This division is also supported by Bordwell's (1985) division of films belonging to the more canonical, classical narration which dominates both Hollywood mainstream formats and national European mainstream films; art cinema narration with its more fragmented and less linear models of narration found in, for instance, European neorealism, New Wave cinema, and modernist and postmodern films; and finally nonnarrative forms such as documentary. All three forms have very distinct institutional backgrounds and are normally very easily recognized by the general audience, because distribution and indexing of art cinema, for example, is linked to special cinemas, festivals, and a special critical discourse, and documentary films have a totally separate distribution and are indexed much more for their special reference to reality.

2.2 Classical Hollywood and Beyond

Very often the classical Hollywood genres form the basis of what is generally understood as the fundamental genres, and these would most typically include melodrama, comedy, action-adventure, science fiction, horror, Western, film noir, crime genres, and musical. But each of these categories has subcategories, such as the crime genre, which is often further divided into gangster films, thrillers, hardboiled detective films, spy films, police films, etc. A popular Internet site like the Internet movie database lists 11 general genre categories: action and adventure, art house and international, classics, comedy, drama, horror and suspense, kids and family, musicals, science fiction and fantasy, special interests and studio specials, and documentary. But each of these categories again has a large number of subcategories. Action and adventure, for instance, includes war films, Westerns, and disaster films. Neither the historical, critical, nor the popular taxonomy of genres are defined in very rigid and stable forms, although some basic forms are clearly present. As indicated in the ritual approach, the cognitive approach, and in most studies of the classical Hollywood system, it is however possible to define a number of more general generic forms and prototypes, which then often have different subforms and furthermore can change and vary historically.

Several models of typical historical develoments of genres have been suggested in film genre theory. Neale ([1990] 1995), for instance, discusses the evolutionary model, used by Schatz (1981). In this model genres move from an experimental stage to a classical stage and to stages of more opaque, baroque, ironic, or even metatextual modes in which the basic formula elements are so familiar that directors and audiences have a more playful and reflexive relation with them. Neale points to another model in which genres occupy shifting positions as dominant, and run through phases of canonization, automation, and reshuffling. In recent theories on genre mixing and intertextuality it is furthermore often indicated that it is an increasingly common strategy by both the producer and the director to try to mix genres, to have maximum audience effect, and that furthermore the general audience today is much more used to genre mixing and intertextual and metatextual elements even in mainstream films.

2.3 Nonfiction Films and Genre

Many theories and discussions of film genres relate mainly to fictional film; however, both nonfiction film genres and the distinction between fiction and nonfiction are very important for the understanding of the fundamental film genres. Historically the term documentary was defined by John Grierson in 1926 as 'creative treatment of actuality.' This definition points to the fact that documentary films, like all nonfiction genres, have a special and more direct reference to reality than even the most realistic fictional film, but that on the other hand all documentary films are also a product of rhetorical and esthetic choices. Plantinga (1997) gives a pragmatic definition of the distinction between nonfiction and fiction: Nonfiction is related to those discourses that make explicit claims about reality and fiction to discourses that do not, a distinction between an assertive and a fictional mode, in which the

spectator treats the reference from text to reality in two fundamentally different ways.

Historically, esthetically, and rhetorically the documentary film has developed many different ways of representing and talking about reality. Barnouw (1993) uses different communicative roles to define different rhetorical purposes in documentary: explorer, reporter, advocate, poet, promoter, and observer. Renov (1993) in his theorizing about documentary forms makes a distinction between four fundamental functions of documentary films: to record, reveal, and preserve; to persuade or promote; to analyze and interrogate; and to express. Combinations of each of these communicative roles or purposes of course result in different subgenres of documentary.

Documentary genre theory offers several ways of defining the basic genres, but one of the most widely cited is Nichols's (1991) definition of four fundamental modes of representation: *the expository documentary* (with direct address to the viewer, often through authoritative voice-over and based on explicit arguments); *the observational documentary* (often based on simply edited visual observation and people expressing views in their observed reality, but without explicit commentary); *the interactive documentary* (based on a more active intervention and interviewing into reality, but still without an authoritative voice and comment); and *the reflexive documentary* (with a dominance of more self-reflexive and poetic strategies foregrounding the film's own strategy and relation to reality). If these modes can be taken as basic prototypes, then of course different subgenres can mix these modes and rhetorical strategies and it should also be added that mixing of fictional forms and documentary forms also has a long history. Docudrama and drama documentary or even more more complicated fact–fiction formats have developed strongly in recent film and TV culture adding to the esthetic and rhetorical variations of documentary forms or 'faked' documentary forms (Rosenthal 1999).

See also: Art and Culture, Economics of; Art, Sociology of; Culture, Sociology of; Film and Video Industry; Television: General; Television: Genres; Television: History

Bibliography

Altman R [1984] 1995 A semantic/syntactic approach to film genre. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 26–41

Altman R 1999 *Film/Genre*. British Film Institute, London

Barnouw E 1993 *Documentary. A History of the Non-fiction Film*. Oxford University Press, New York

Bazin A 1971a The evolution of the Western. In: Bazin A (ed.) *What is Cinema? II*. University of California Press, Berkeley, CA, pp. 149–58

Bazin A 1971b The Western: Or the American film par excellence. In: Bazin A (ed.) *What is Cinema? II*. University of California Press, Berkeley, CA, pp. 140–49

Bordwell D 1985 *Narration in the Fiction Film*. University of Wisconsin Press, Madison, WI

Braudy L 1976 *The World in a Frame: What We See in Films*. Anchor, New York

Buscombe E [1970] 1995 The idea of genre in the American cinema. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 11–26

Carroll N 1990 *The Philosophy of Horror, or Paradoxes of the Heart*. Routledge, New York

Carroll N 1999 Film, emotion and genre. In: Plantinga C, Smith G M (eds.) *Passionate Views. Film, Cognition, and Emotion*. Johns Hopkins University Press, Baltimore, pp. 21–48

Cawelti J 1976 *Adventure, Mystery and Romance*. University of Chicago Press, Chicago

Clover C 1992 *Men, Women and Chain Saws: Gender in the Modern Horror Film*. Princeton University Press, Princeton, NJ

Comolli J-L, Narboni J [1969] 1992 Cinema/ideology/criticism. In: Mast A, Cohen M, Braudy L (eds.) *Film Theory and Criticism*. Oxford University Press, New York, pp. 682–9

Cook P, Bernink M (eds.) 1999 *The Cinema Book*, 2nd edn. British Film Institute, London

Gledhill C 1987 *Home is Where the Heart Is. Studies in Melodrama and The Woman's Film*. British Film Institute, London

Grodal T 1997 *Moving Pictures. A New Theory of Film Genres, Feelings and Cognition*. Oxford University Press, Oxford, UK

Kaplan A [1978] 1998 *Women in Film Noir*, rev. edn. British Film Institute, London

Klinger B [1984] 1995 Cinema/ideology/criticism revisited. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 74–91

Kracauer S 1960 *Theory of Film*. Oxford University Press, London

Mulvey L [1975] 1992 Visual pleasure and narrative cinema. In: Mast A, Cohen M, Braudy L (eds.) *Film Theory and Criticism*. Oxford University Press, New York, pp. 746–58

Nichols B 1991 *Representing Reality*. Indiana University Press, Bloomington, IN

Neale S 1980 *Genre*. British Film Institute, London

Neale S [1990] 1995 Questions of genre. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 159–87

Plantinga C 1997 *Rhetoric and Representation in Nonfiction Film*. Cambridge University Press, Cambridge, UK

Renov M (ed.) 1993 *Theorizing Documentary*. Routledge, New York

Rosenthal A (ed.) 1999 *Why Docudrama? Fact-Fiction on Film and TV*. Southern Illinois University Press, Carbondale, IL

Schatz T 1981 *Hollywood Genres: Formulas, Filmmaking and the Studio System*. Random House, New York

Todorov T [1970] 1975 *The Fantastic*. Cornell University Press, Ithaca, NY

Wellek R, Warren R [1949] 1956 *Theory of Literature*. Harcourt, Brace & World, New York

Williams A 1984 Is a radical genre criticism possible? *Quarterly Review of Film Studies* **9**(2): 121–5

Williams L 1989 *Hard Core: Power, Pleasure, and the 'Frenzy or the Visible.'* University of Berkeley Press, Berkeley, CA

Wright W 1975 *Sixguns and Society: A Structural Study of the Western*. University of California Press, Berkeley, CA

Wood R [1977] 1995 Ideology, genre, auteur. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 59–74

I. Bondebjerg

Film: History

As the art of film enters its second century, film historians not only encounter a longer and more heterogenous film-historical past, but do so from within contexts unimaginable during the 1930s, when an important wave of book-length film histories appeared. Most notably, film history is now a field of study within higher education, with specialized journals, professional organizations, and conferences. An indication of film history's evolved condition is the extent to which film-historical research today proceeds through explicit revision of the findings and assumptions of established film histories.

Contemporary film historiography's self-critical orientation, together with the discovery and restoration of lost films, has enabled a fundamental rethinking of key periods of film history. Examples include recent studies of silent-era cinemas that had preceded or outsurvived the formation of Hollywood's so-called classical cinema. In attending to silent cinema's differences from later Hollywood practice, the new historiography has complicated established interpretations of film history as a progressive movement toward worldwide adoption of classicism. New interpretations of other periods in film history are certain to occur as films continue to be found, identified, reconstructed, and made available to scholars, especially films produced outside of film history's familiar research domains of Western Europe and the USA. Indeed, as new case studies accumulate, it is questionable whether world cinema can be said to have a history, if the term history is understood to imply a single narrative that might encompass the heterogeneity of the world's films and film practices.

1. *Film as Art and Industry*

Critics have often distinguished film from traditional arts, such as painting, music, and literature, by insisting that film is simultaneously an art and an industry. This familiar claim points to differences between the industrialized processes of filmmaking and the craft-based traditional arts that have differentiated the historiography of film from that of the other arts. Here three basic points can be made. First, as products of an industrial system, films lend themselves to a variety of forms of analysis, ranging from traditional art-historical approaches, centered on issues of esthetics and filmmakers' biographies, to diverse economic, technological, sociopolitical, and quasi-anthropological investigations (see Allen and Gomery 1985, Lagny 1992). Indeed, film history's object of study is rarely defined exclusively in esthetic terms, as is evident in the use of nonesthetic events, such as the end of World War II, to demarcate periods in film history. Even film histories concerned mainly with esthetic questions typically also situate films in economic, technological, and social contexts. In fact, certain film histories conceptualize film primarily as a technological, economic, and/or social phenomenon, to the point of excluding esthetic considerations altogether. Such has been the case, for instance, in studies of the economics of national film industries.

Second, the evidently industrial nature of film production has encouraged modernist conceptions of historical change, according to which film history moves forward, in a single direction, with new films necessarily counting as a technical advance over old films. The histories of the established fine arts traditionally have been interpreted in terms of cyclical patterns of change, according to which artistic movements rise, fall, and emerge again, in the manner of the life cycle of a natural organism. In contrast, filmmaking, so evidently a function of technological, industrial, and commercial conditions and forces, also invites conceptualization as an ongoing, forward-moving process, forever open to additional improvement. This conception of historical change as a modernizing process underlies both the film industry's celebration of the latest advances in film technique and critics' nostalgia for the lost artistry of the filmmaking of earlier periods.

Third, film history's progressivist outlook has been supported by the relatively short duration of the cinema's history. The histories of painting, poetry, and music have typically been studied irrespective of present-day artistic developments. Film history, however, has been concerned with an art whose past coincides with a recent and relatively short span of historical time. Thus, film historians have often examined the cinema's past in light of its latest trends. Indeed, on occasion, current developments in film practice have enabled new directions in film-historical inquiry. For instance, during the 1930s, the conversion to sound cinema had made it seem as if 'silent' films belonged to the historical past, thereby providing justification for the creation of film archives in Germany, the UK, the USA, France, and elsewhere.

2. *Beyond Classical Cinema*

Film history often displays a retrospective orientation in which the cinema's past comes up as a prelude to the cinema of the historian's present. Examples can be found in film histories that understand world cinema as having evolved toward the adoption of the so-called classical film style associated with studio-era Holly-

wood. Such adoption is generally taken to have occurred by the late 1930s, when filmmakers, in film-producing countries throughout the world, began making sound movies in imitation of Hollywood's films and filmmaking methods. Popular throughout the world, Hollywood films appear to have met transcultural expectations concerning story structure; moreover, it has been argued that the style of Hollywood films conforms to universal perceptual dispositions (see Bordwell et al. 1985). For instance, in a classical film, the duration of a shot is never too short or too long, but instead just long enough for the viewers, regardless of their cultural experience, to register the shot's significance to the unfolding of the film's narration.

Although Hollywood's studio system had collapsed by 1960, classicism continues to provide the context for inquiries into recent narrative cinema, as if classicism's apparent conformity to naturalized perceptual and cognitive habits has made it difficult to see how mainstream film practice might evolve beyond it. Yet since the 1970s, such evolution appears to have been happening, as the contemporary commercial cinema, with its emphasis on spectacular effects and tendency toward episodic narration, increasingly exhibits nonclassical attributes. Thus, following the pattern whereby today's film trends prompt reconsideration of the cinema's past, the nonclassical salience of current mainstream cinema has served to cast new light on established areas of film-historical inquiry. Such areas have begun to include studio-era Hollywood itself, as is evident in recent critical studies of the nonclassical aspects of Hollywood genres such as the musical, film noir, melodrama, and comedy (see Altman 1999).

3. *A Revised Chronology*

Since the 1970s, film historians, working in various research fields, have attempted to understand past films and film movements as if classicism's status as the pregiven *telos* or goal for world cinema can no longer be taken for granted. An indication of this development can be seen in a basic shift in historical method. Film historiography has traditionally displayed what might be called a 'vertical' interest in the position of a film, or a film-historical period, within a developmental sequence. In contrast, the new historiography can be said to be defined by 'horizontal' approaches, in which films are analyzed as embedded within contemporaneous contexts of production, exhibition, and reception.

3.1 *Early Cinema, 1895–1906*

Recent research on the early cinema of the period from 1895 to about 1906 provides a prototypical example of contemporary film historiography's horizontal orientation. The period of early cinema begins in the mid-1890s with the first screenings of films by inventors of motion picture cameras and projectors. Canonical examples include screenings by the Lumière brothers in France and the Skladonowsky brothers in Germany in late 1895, and, in early 1896, by Birt Acres in London and Thomas Edison in the USA. The period continues until roughly 1906, when the sites of film exhibition were no longer mainly fairgrounds, vaudeville houses, and rented theatres, but came also to include nickelodeons and other permanent venues intended specifically for film screenings.

Accounts of early cinema typically center on the development of techniques of fiction filmmaking, especially devices of film editing. The films of inventors such as Edison and the Lumières were made up of only a single shot, usually lasting the length of a standard reel of film. But by 1899, in response to a greatly increased demand for films, film companies in France, the UK, and the USA began producing relatively lengthy narrative films made up of multiple shots. Historians have analyzed these films as attempts to solve the problem of how separate shots might be staged and connected together so as to create the impression of a continuous story. In established accounts, early multiple-shot films such as *The Great Train Robbery* (made by Edwin S. Porter in 1903) mark important moments in the development of the system of stylistic conventions associated with studio-era Hollywood (see Jacobs 1939).

Since the late 1970s, however, a very different history of the cinema of 1895–1906 has been written (see Elsaesser and Barker 1990). Instead of examining early cinema as a forerunner of the commercial narrative cinema of the late 1910s, the new historiography has analyzed early films in the context of their function within vaudeville routines, illustrated song slides, public lectures, and other forms of popular spectacle and fairground-like attraction. Examining how live music, spoken narration, and other exhibition practices affected the viewing experience, recent studies of early cinema bring forth dimensions of early films occluded within traditional developmental accounts. An example is the theatrical or presentational aspect of early cinema, which became a focus of historical study in the 1980s. Investigating early cinema's practices of direct address, adapted from vaudeville, the fairground, and other forms of popular theater, historians have come to understand early films less as primitive attempts at classical storytelling and more as performances within the context of live, spectacle entertainment at the turn of the century.

3.2 *The Transition to Classicism*

The cinema of the period 1906–1917 has traditionally been conceptualized as a transition from early cinema

to classicism. Key industrial developments include: a shift in editorial control away from exhibitors and toward producers; the creation of exchanges which rented film prints to exhibitors rather than sold them outright; the organization of film production according to a hierarchical division of labor that gave producers authority over directors and camera operators; the establishment of Hollywood as a production center; and the cinema's further expansion worldwide, marked, for instance, by the emergence of Denmark and Italy as major film-producing countries. Basic stylistic changes include: the dominance of the fiction film over the travelogs and documentary 'actualities' that had been an important part of early cinema; a new emphasis on celebrity performers or 'stars'; and a continuing increase in film length, leading, by 1914, to the establishment of the feature-length film as the norm for commercial movie-making.

As with other silent-era periods, established histories of the 1910s have emphasized the development of devices of film editing. Since the 1910s, critics noted that editing enabled forms of acting unique to cinema, in which actors' performances were constructed through a planned assembly of numerous shots rather than captured in a lengthy, uninterrupted run of the camera. During the 1920s, film theorists in France and the USSR argued that editing, in differentiating film from theater, had served to establish film as a singular art.

Accounts of the cinema of the 1910s have long centered on the work of the director D. W. Griffith. Dubbed the father of narrative cinema, Griffith has been credited with having conceived of the scene as a composite of shots, staged, filmed, and assembled so as to construct a unified dramatic world. In some 500 short one- and two-reel films, directed by Griffith between 1908 and 1912, devices such as the closeup, cross-cutting, and the eyeline match are employed as elements within what film theorists have described as a storytelling 'language' unique to cinema. Films such as *The Lonely Villa*, made by Griffith in 1909, feature elaborate last-minute rescues, made up of dozens of shots. Such sequences include events occurring in as many as three separate locales, cut together in rhythms that accelerate or slow down in relation to the intensity of the dramatic action. Innovations associated with Griffith became essential to US film practice, thus providing the basis for the classical style that would take form in Hollywood circa 1917.

During the 1990s, some historians have challenged the traditional emphasis on editing practices, arguing that proper appreciation of the cinema of the 1910s requires attention to techniques of *mise-en-scène*, i.e., the staging of dramatic action, through arrangement of actors, settings, and props within the camera's visual field (see Brewster and Jacobs 1997). Such attention seems especially warranted in the case of films made outside the USA, where innovations in film technique often were prominent in areas other than editing. For instance, recent archival discoveries have shown that the film industries of Italy, Russia, France, Denmark, Sweden, and Germany produced hundreds of narrative films whose stories, in crucial respects, were told through techniques of staging. With scenes that play out in extended long takes, European films of the 1910s commonly feature far fewer shots than comparable US films, instead presenting story events through careful choreography of multiple actors' movements relative to their positions within the film's frame. The ensemble performances, in which several dramatic actions may occur simultaneously within a single framing, serve to situate character interactions within a shared environment. Thus, the narration of European films of the 1910s might be said to exhibit a psychological or social dimension unavailable to the contemporaneous US cinema's stress on linear, physical action.

3.3 The Decade of National Cinemas

By 1918, with the operation of Europe's film industries having been interrupted by World War I, Hollywood had established itself as the world's foremost film exporter, with distribution branches in South America, Australia, the Far East, and Europe. In response to Hollywood competition, Europe's film industries imposed quotas and other trade restrictions on Hollywood films, and drew inspiration from modernist trends in the arts to invent distinctive national film styles. The major art cinemas of the 1920s have long been a focus of film historiography. Such cinemas include expressionism and the *Neue Sachlichkeit* ('new objectivity') in Germany, impressionism and surrealism in France, and constructivism in the USSR. In France, where there was no clear industrial division between commercial cinema and art cinema, avant-garde techniques were integral to many mainstream films.

The experimental film movements evolved within a new international film culture, organized around specialized journals, film clubs, and art-cinema theaters. An unprecedented intellectual interest in film became manifest in a proliferation of theoretical writings and manifestos. Theorists stressed the need to overcome the apparent theatricality of established film styles, while numerous filmmakers, in Europe and elsewhere, undertook radical explorations of film's artistic potential. Thus, the films of the major European film movements of the 1920s frequently exhibit far more editing technique than had earlier European films. The new art films also often featured experimental cinematographic devices, from filtered lenses to superimposed images, double-exposures, and split-screen compositions.

Certain European film movements explored film's potential for radical alternatives to conventional scenography. Prominent examples include: the mon-

tage films of the USSR, such as Sergei Eisenstein's *The Battleship Potemkin* (1925) and Dziga Vertov's *Man with a Movie Camera* (1929), famous for their radical experiments in film editing; and French Impressionist films, from Louis Delluc's *Fièvre* (1921) to diverse films by directors such as Germaine Dulac, Jean Epstein, and Abel Gance, who employed multiple superimpositions and other optical effects to simulate diverse moods and affective states. Also important were German films such as Robert Wiene's *The Cabinet of Dr. Caligari* (1920) and Fritz Lang's *Metropolis* (1927), acclaimed internationally for their spectacular, studio-created settings. In addition, German films such as F. W. Murnau's *The Last Laugh* (1924) were known for the use of optical effects and elaborate moving-camera shots to suggest a character's embodied or psychological 'point of view.' By the mid-1920s, with increasing numbers of films and filmmakers crossing national borders, an international avant garde had taken form, encompassing filmmakers in Europe, North America, and Japan. 'International' films, such as Teinosuke Kinugasa's *Page of Madness* (1926) and Carl Dreyer's *La passion de Jeanne d'Arc* (1928), confounded national labels with their unusual mix of styles.

Questions remain concerning how new research into national cinemas besides those of France, Germany, the USSR, and the USA might alter the historiography of world cinema during the 1920s. For instance, during this period, Japan and India were among the world's largest film-producing countries. Yet, as with many other national cinemas during the years between the world wars, the films were rarely distributed beyond the national and regional film markets, and few of them have been preserved by distributors and archives. Thus, historical assessments of world cinema during the interwar years remain incomplete and provisional.

3.4 The Conversion from Silent to Sound Cinema

The introduction of recorded sound to cinema has attracted much historical analysis and commentary. Beginning in 1927, the unexpected success of Alan Crosland's *The Jazz Singer* (1927) provoked debate worldwide concerning the future of cinema. By the summer of 1928, as the film industry in the USA began its conversion from silent to sound cinema, observers described it as nothing less than an artistic, technological, and economic revolution. Historical accounts of the conversion period have traditionally emphasized various traumas, such as: the sudden collapse of established actors' careers, the flood of new stage and technical personnel into the film industry, the technical difficulties of sound filmmaking, the esthetic limitations of early sound films, the refinancing of the film industry by large banks, and the curtailment of avant-garde experimentation brought about by the high costs of sound-film production. The impression of crisis encouraged partisan assessments, whether in the form of Hollywood's celebration of sound cinema as a great technological advance or critics' nostalgia for the vanished art of silent film. In either case, film history had seemed to start over again with the introduction of recorded sound.

Concerning film esthetics, sound conversion is said to have produced two main consequences. First, early sound films appeared to signal an artistic regression from the sophisticated art films of the 1920s. The available technologies had restricted the style of many early sound films to relatively long takes, stationary framings, and immobile actors who recited dialog in styles familiar to the popular stage. For many modernist critics, sound cinema amounted to a return to the allegedly theatrical cinema of the years prior to World War I. Established histories of the early sound years have focused on exceptional films such as Rouben Mamoulian's *Applause* (1929), René Clair's *Sous les toits de Paris* (1930), and Fritz Lang's *M* (1931), whose innovations in multitrack recording, postsynchronization, playback, and other techniques are credited with having restored movement to cinema.

Second, the technical demands of sound filmmaking are said to have pushed filmmakers everywhere toward adoption of film classicism, thereby producing a worldwide homogenization of film style. For critics familiar with the distinctive national film styles and transnational film movements of the 1920s, sound films often seemed to belong to a common family, regardless of where they were made. Even the USSR, famous during the 1920s for its counter-Hollywood film culture, was, by 1934, producing films that looked and sounded much like commercial narrative films made elsewhere.

In assessing sound's impact on film style, it is important to distinguish between filmmaking in Hollywood, which had undergone full industrialization prior to sound conversion, and filmmaking in other countries. One issue concerns the industrial conditions for film style. Whatever the personal, technical, and administrative upheaval, sound conversion did not fundamentally change Hollywood's processes of serial manufacture. Thus, sound conversion occurred in Hollywood under circumstances that enabled a continuation of silent-era practice. Most notably, Hollywood's sound-era editing style remained essentially consistent with established methods, involving many of the same techniques for constructing a unified, linear narrative. After a brief period of experimentation, even the new sound techniques were brought into line with classical precedent (see Crafton 1997). For instance, in 'backstage musicals' such as Harry Beaumont's *The Broadway Melody* (1929), song-and-dance sequences no longer came forward strictly as recordings of self-contained stage performances but instead functioned within narratives concerning the offstage lives of show-business personalities. By the summer of 1931, with

the industry-wide adoption of multitrack recording and mixing, the foregrounding of sound technique as an artistic novelty had been definitively abandoned in favor of the stylistic balance characteristic of Hollywood films of the late silent period.

In many other countries, however, sound conversion had entailed fundamental industrial change, and hence radical consequences for film esthetics. Small countries such as Canada were unable to sustain the industrial production of feature films, and instead came to produce mainly short films, such as documentaries and animated films. In countries such as India and Mexico, where filmmaking had been minimally industrialized, complete restructuring of the industry occurred, and new, sound-era genres, such as musicals that drew on indigenous entertainment traditions, came to define the national film output.

Research into the worldwide conversion to sound remains in its early stages. Relatively little is known about the output of many important film industries of the period. Recent research on Japanese cinema provides a case in point. Although Japan's large film industry dates from the 1920s, research on the history of Japanese cinema began to flourish only in the 1970s, with the rediscovery of the prewar films of Yasujiro Ozu, Kenji Mizoguchi, and other prominent art-cinema directors. Assessments of world cinema during the 1930s are certain to undergo revision pending further research into other important film industries of the period. In this event, the familiar notion that world cinema became homogenized during the 1930s may prove to have been a function of the restricted corpus of films then available to historians.

3.5 Postwar Internationalism

The decade after World War II is known for the emergence of a revived international art cinema. Prominent institutional developments include: the creation of international film festivals in cities such as Cannes, Venice, Berlin, Edinburgh, and Tokyo; increased government funding of film archives, thereby enabling expansion of the International Federation of Film Archives; new film journals devoted to international coverage; and the opening of special art-house movie theaters in cities throughout the world. These developments had prewar antecedents, and observers noted parallels between the new art cinema and the international avant-garde of the 1920s. Nonetheless, the new film culture differed in the extent to which its global network functioned independently from nation-state institutions.

An example can be found in the worldwide reception of Italian neorealism, whose status as the pre-eminent postwar film movement depended crucially on critical acclaim from outside Italy. Films such as Roberto Rossellini's *Rome, Open City* (1945) and Vittorio De Sica's *Bicycle Thieves* (1948) drew admiration from filmmakers from every continent. Most notably, in evident contrast to the commercial cinema's linear and formulaic plots, the neorealist films emphasized the acts and events of ordinary life, including chance occurrences, routine, nondramatic actions, and inconclusive chains of events. Neorealist films were also acclaimed for the documentary-like 'realism' manifest in the use of nonprofessional actors, improvised performances, actual locations, and natural lighting.

Neorealism has commanded attention from historians as a beginning for prominent subsequent movements and tendencies in world cinema. Indeed, neorealism is said to have affected filmmaking in Asia, Africa, and Latin America more so than any other film movement of the sound era. For many filmmakers, neorealism offered a coherent, low-budget alternative to the style of commercial cinema, with great potential both for naturalistic social representation and for sociopolitical allegory. Filmmakers who have cited Italian neorealism as a personal inspiration include: Fernando Birra and Tomás Gutiérrez Alea, leading directors during the 1960s in the film industries of Argentina and Cuba, respectively; Satyajit Ray, the famous Bengali director from the Indian subcontinent; and Nelson Pereira dos Santos, founder of the *Cinema Novo* movement in Brazil.

The emergence of innovative film movements in Latin America, Asia, and, by the late 1960s, in Africa, with the Senegalese director Ousmane Sembene, had been conditioned by two basic developments. First, in 1947, movie attendance in North America had begun to fall, and would continue to do so through much of the 1960s. The decline became apparent just as Hollywood faced an important 1948 decision by the US Supreme Court, which required that the five major studios sell off their theater chains, and that both major and minor studios cease a variety of other monopolistic practices. In response to the economic and political adversity, the major Hollywood studios produced fewer films throughout the 1950s, targeted now toward specific markets. Second, during this same period, many countries experienced booms in film attendance, especially in the so-called developing world, where television had not yet emerged as an entertainment alternative. Responding to the demand, major film industries emerged in Egypt and Hong Kong, while the new nations of Burma, Pakistan, South Korea, and the Philippines also began producing films on an industrial scale. Also during the 1950s, India became one of the world's largest film producers, making hundreds of films exclusively for the South Asian film market. By the mid-1960s, some 60 percent of all of the world's feature films were made in Asia, Latin America, and Africa.

Coinciding with the worldwide boom in film production, 'new waves' and 'young cinemas' took form in diverse countries, beginning in the late 1950s and continuing through the 1960s. Young filmmakers

worked to distinguish themselves from the established filmmakers of their native countries by emulating new currents in world cinema. Most prominent was the French New Wave, known for films such as François Truffaut's *Les Quatre Cent Coups* (1959) and Jean-Luc Godard's *Breathless* (1960), which featured inventive adaptations of conventions drawn both from Hollywood and European genre films and also from the international art cinema. Important new waves also emerged in Japan, Czechoslovakia, and, indeed, in virtually all countries, in one form or another, during the 1960s. A famous example is the Young German Cinema, whose self-conscious opposition to Germany's established film culture opened the way for the international prominence during the 1970s of directors such as R. W. Fassbinder, Werner Herzog, and Wim Wenders.

3.6 Film Historiography Since the 1970s

The new film culture of the 1960s would have a profound impact on the writing of film history. An example can be found in the notion of the film director as artist or *auteur*. This notion had been formulated in Paris in the 1950s by a group of young film critics, a remarkable number of whom would go on to become key directors of the French New Wave. These critics argued that film directors should be seen as performing the kind of creative role traditionally accorded to novelists, playwrights, painters, and composers. A precedent could be found in certain currents in film critical writing during the 1920s, which had recognized filmmakers such as Charlie Chaplin, F. W. Murnau, and Erich von Stroheim as artists. But the auteur critics extended the sphere of art to encompass a remarkable diversity of films, including, in some cases, even low-budget sound-era genre films. Provocatively comparing movie directors to Mozart and Shakespeare, the auteur critics expanded the canon of great film artists to include current commercial directors such as John Ford, Orson Welles, Alfred Hitchcock, and Howard Hawks.

The 1970s might be said to have opened a new era in film historiography. During the 1970s, the integration of film study into university curricula in Western countries opened up new contexts for film-historical research. For instance, the notion of the director as artist came to inform the academic study of film, leading to the publication of numerous monographs on a wide variety of filmmakers. Moreover, professional approaches to film study, involving the critical re-examination of established film history, also took form. A case in point is the new historiography of silent cinema. Whereas previous film historians had often relied on other historians' accounts, together with personal memories of films, the new generation of archivists and professional scholars studied large collections of film prints, comprised of recognized classics as well as many lesser known or forgotten works. Additional historical sources included a wide variety of print documents, including film reviews, trade publications, film companies' catalogs, and paper prints of films from the US Library of Congress. The result of this research was a thorough reconceptualization of the early silent period. This work, in turn, is likely to serve as an exemplar for historians working on later periods, such as the period of sound conversion, which, like early cinema, demands both empirical research and historiographical reflection.

As the cinema now enters its second century, new technological changes can be expected to enable new ways of seeing the cinema's past. For instance, as computer-generated digital images come to replace the cinema's photographic image, cartoons and other nonphotographic animated films may come to appear central to film history. In any case, as old films continue to be restored and made available to scholars, significant changes in film historiography continue to occur. At annual silent-film festivals, such as those held in Bologna and Pordenone, large bodies of films, representing the output of diverse directors, actors, studios, film movements, and national cinemas, are made accessible to critics and historians. Each year, new discoveries stimulate revision of established scholarship, while at the same time underscoring the incompleteness of the historical record, and hence the certainty that film history will remain a work in progress.

See also: Art and Culture, Economics of; Art, Sociology of; Culture, Sociology of; Film: Genres and Genre Theory; Television: History; Television: Industry

Bibliography

Allen R, Gomery D 1985 *Film History: Theory and Practice*. Knopf, New York

Altman R 1999 *Film/Genre*. British Film Institute, London

Bardèche M, Brasillach R 1935 *Histoire du cinéma*. Denoël et Steele, Paris

Bazin A 1958 *Qu'est-ce que le cinéma?* Editions du Cerf, Paris

Bordwell D, Staiger J, Thompson K 1985 *The Classical Hollywood Cinema: Film Style and Mode of Production to 1960*. Columbia University Press, New York

Brewster B, Jacobs L 1997 *Theatre to Cinema: Stage Pictorialism and the Early Feature Film*. Oxford University Press, Oxford, UK

Cherchai-Usai P 1995 *Burning Passions: An Introduction to the Study of Silent Cinema*. British Film Institute, London

Crafton D 1997 *The Talkies: America's Transition to Sound, 1926–1931*. Scribner, New York

Elsaesser T, Barker A (eds.) 1990 *Early Cinema: Space, Frame, Narrative*. British Film Institute, London

Gunning T 1991 *D. W. Griffith and the Origins of American Narrative Film*. University of Illinois Press, Urbana, IL

Hill J, Church Gibson P 2000 *World Cinema: Critical Approaches*. Oxford University Press, Oxford, UK

Houston P 1994 *Keepers of the Frame: The Film Archives*. British Film Institute, London

Jacobs L 1939 *The Rise of the American Film*. Harcourt Brace, New York
Lagny M 1992 *De l'histoire du cinéma: Méthode historique et histoire du cinéma*. A. Colin, Paris
Salt B 1992 *Film Style and Technology: History and Analysis*, 2nd expanded edn. Starword, London
Sklar R 1993 *Film: An International History of the Medium*. Prentice Hall and Harry N. Abrams, New York
Thompson K, Bordwell D 1994 *Film History: An Introduction*. McGraw Hill, New York
Vasey R 1997 *The World According to Hollywood, 1918–1939*. University of Wisconsin Press, Madison, WI

C. O'Brien

Finalization in Science

The finalization model addresses the question of how internal (cognitive) and external (social) regulations interact in the dynamics of modern science. It assumes that paradigm building in scientific disciplines will eventually lead to stable theories that solve the fundamental problems in the field and are unlikely to be repealed by a Kuhnian scientific revolution (Kuhn 1962). With the advent of paradigmatic theory, research programs in disciplines become responsive to social and political goals on the level of theory formation. Therefore, cognitive autonomy in the sense of a compelling internal logic of the growth of knowledge is a transient feature and hence more characteristic of an immature science than a mature one. (The finalization model is explicated in Böhme et al. 1983; it has been reassessed in contributions to a symposium for *Social Science Information* Vol. 36, see Leeming 1997.)

1. *The Internalism–Externalism Debate*

The internalism–externalism debate—a debate that has intrigued historians and sociologists of science since the 1960s—fell into disrepute in the 1980s when mainstream studies of science became preoccupied with the analysis of the social construction and cultural embeddedness of knowledge (see *Laboratory Studies: Historical Perspectives*). These studies have tended to discard any distinction between the cognitive and the social aspects of science (Shapin 1992). However, a (relative) distinction must be upheld if one is to explain not only how science can be shaped by social context, but also how society can be transformed through science (see *Knowledge Societies*).

Modern science emerged from the struggles over cognitive order that waged throughout the seventeenth century, eventually establishing the notion of, and a niche for, the pursuit of natural knowledge which sets empirical facts categorically apart from religious belief, social choice, or political preference (see Webster 1976). The 'edge of objectivity' (Gillespie 1960) is culturally biased: modern science applies an operational (experimental) conception of truth that reflects a social interest in prediction and control. Despite that bias, however, it yields universally valid knowledge. Accordingly, when observations in the laboratory fail to identify any interesting epistemological difference between the pursuit of knowledge and the pursuit of power (Knorr Cetina 1995, p. 151), this tells more about sociology than about science. The finalization model acknowledges the premise of social construction, but still emphasizes the institutional and cognitive differentiation and specificity of modern science as a system of knowledge production. It considers the internal logic of the system—albeit as an emergent property and evolving process.

2. *The Rise of Scientific Internalism and its Transition*

2.1 *Paradigmatic Science*

Historically, scientific fields develop through phases in which the mechanisms of internal problem generation prevail. The *locus classicus* of such growth, both in cognitive and institutional terms, are academic disciplines or specialties (e.g., mechanics, electrodynamics, physical chemistry, genetics, biochemistry) that pursue research programs in which the fundamental laws governing the phenomena in their respective object areas are elaborated and confirmed. In this phase of theory formation or paradigm building (Kuhn 1962), frontiers of research are largely determined through explanatory ideals and strategies within science. The flow of problems in understanding the nature and mechanisms of light, electricity, matter, heredity, or cell differentiation regulate the dynamics of the disciplines. The full logic of this internalism is not well understood and remains a matter of ongoing debate among philosophers and sociologists of science (see Hacking 1983, Pickering 1995). The search for more fundamental and more comprehensive theories to explain natural phenomena of increasing complexity seems to be an element of that logic.

2.2 *Postparadigmatic Science*

The finalization theory adopts Thomas Kuhn's phase model of disciplinary development (preparadigmatic, paradigmatic, normal science) with one major modification: it assumes that the dynamics of paradigmatic science can lead to mature theories which provide a fundamental and, in a certain sense, conclusive understanding of the phenomena in their domain. To assume that all theories will eventually be falsified or superseded through scientific revolutions, and completely replaced by new paradigms, fails to account for the extraordinary stability of modern theories. For at least

100 years, consolidated theories of motion, heat, light, electricity, matter, or heredity have not been rendered obsolete, despite the Kuhnian revolutions that established new paradigms. Werner Heisenberg (1971) proposed the concept of 'closed theory' to capture the stability of classical mechanics and electrodynamics in the transition to relativity and quantum mechanics (see also Böhme 1980).

To the extent that disciplines achieve theoretical maturity, the core of their dynamics will shift from research that strives for the discovery of new fundamental laws to research that applies these laws to the wealth of phenomena in their domain. Science is a truly 'endless frontier' (Bush 1945) not so much because fundamental problems are multiplied beyond limits and paradigm change is always imminent, but rather because existing paradigms can be extended to an indefinite number of applications. Such applications are termed 'postparadigmatic' science in the finalization model.

2.3 *The Social Direction of Theory Development*

Applying paradigmatic theories is not a simple exercise in deduction. The general principles of hydrodynamics or quantum mechanics or organic chemistry are valid for fluids or solid matter or the processes of plant nutrition, respectively, but they cannot be applied to them without a series of steps of theory specialization through which the principles are extended and adapted to the complexity of real objects. These steps involve such fundamental achievements as the boundary layer concepts in hydrodynamics, or the band concept in solid-state physics or the metabolic cycles in biochemistry. Postparadigmatic science (Kuhnian 'normal science') is not puzzle solving in a trivial sense, it is theory development at the frontiers of basic research. The point the finalization model raises is that internal regulation of science becomes weak at these frontiers. While problem generation may flow from the natural complexity of real phenomena in the domain of general theories, it can also flow from the task of coping with the boundary conditions defined by technical and social tasks. For example, fluid mechanics had been developed into a theory of flow around airplane wings by Prandtl early in the twentieth century; more recent efforts in fusion research extended the theories of magneto-hydrodynamics to the special conditions required for controlled fusion in a reactor: stability and containment of a plasma in toroidal geometry.

3. *Limitations of the Model*

The finalization model is biased on physics as the exemplary scientific discipline and on theory building as the prevalent pattern of scientific internalism. While the model may cope with the fact that there are other internal frontiers in physics, such as the exploitation of instruments for discovery (Hanson 1965), a critical point is that the meaning of theory may vary among disciplines. Such variation will be reflected in what sociologists describe as the 'epistemic culture' of a field (see Knorr Cetina 1999). Further analysis is needed to determine whether models of scientific growth derived from the case of physics are transferable to genetics or physiology or ecology. The case of agricultural chemistry suggests that theoretical maturity may not be a necessary condition for goal oriented theory development (Krohn and Schäfer 1983).

Finalization is not the only option for the development of disciplines that have reached the paradigmatic stage. 'Normal science' can also pursue the extension of available theories to objects of purely academic interest. In addition, there is also the possibility that the paradigmatic dynamics in the discipline is renewed; non equilibrium theories that go beyond classical mechanics are one example. Such renewal does not challenge the assumptions of finalization as long as new theories do not supersede established paradigms, but indeed provide a more precise understanding of their domain of validity.

Applied science must not wait for finalization to become feasible. To a great extent the mobilization of science for external objectives, today as in the past, does not occur on the level of targeted basic research and theory development; rather, less ambitious strategies are pursued that aim at the quick exploitation of findings and methodologies available in science. Application strategies are pervasive in science. They even intrude disciplines in the paradigmatic phase. However, it seems that the priority and the logic of paradigm building remain unaffected in this case. A study of the relation of early molecular biology to medicine indicates that external orientation can operate as long as it converges with the internal frontiers of the discipline. In addition it is possible to transfer the findings and techniques of basic research to contexts of application (Hohlfeld 1983).

Finalization is a pattern of goal orientation in science that becomes feasible with the growth of knowledge in scientific disciplines. However, critical social and political demands for scientific knowledge often transcend the boundaries and resources of existing disciplines. Ecological prognosis of climate change resulting from the burning of fossil fuels is a case in point. While such demands may be rejected as 'transscientific' from the perspective of traditional disciplines (Weinberg 1972), they do find resonance in what has been called 'post-normal science' (Funtovicz and Ravetz 1993). This is basically a recursive effort to combine scientific analysis with political judgement of how one should responsibly proceed in view of the fact that hypotheses about imminent danger cannot at present be tested by scientific means. See also van den Daele and Krohn (1998) as to how the methodology of

research is translated into a policy design for dealing with the inherent risks of the implementation of new technology.

4. The Science for Society Perspective

Ever since Bacon and Descartes, modern science has been hailed as a ready tool for social and political goals. Historically, however, it took time until the sweeping promises began to materialize. Now, at the beginning of the twenty-first century, the applied mode of science seems to dominate (Gibbons et al. 1994). Pressure to cooperate with various consumers of science, mainly industry, is extended even to academic research. Such pressure does not just reflect a change of political climate or the emergence of a new narrative of the science–society relationship (Elzinga 1997): science itself has become more susceptible to external direction. The finalization model provides an understanding of this susceptibility by analyzing the supply-side of science, so to speak. With theoretical maturity, social and political objectives can be pursued through targeted research at the core of scientific disciplines. Postparadigmatic sciences flourish in many fields of physics (National Academy of Sciences 1972, p. 404); the same would hold for chemistry and considerable parts of molecular genetics. The advancement of science is bound to bring more disciplines to a postparadigmatic stage.

Finalization affects the social relations of scientific disciplines. New issues of responsiveness and accountability arise. Social goals for basic research may be defined by scientists sensitive to social issues, but must ultimately be justified in political, not scientific, terms. The generation of problems in postparadigmatic research can be easily converted to a political process that entangles scientists in negotiations with stakeholders in society over the future development of their fields. Finalization, however, implies a commitment to theory formation. While bound to social objectives, it thus operates at some distance to problem definitions derived from the immediate instrumental needs of companies, governments, or regulatory agencies. Such distance may be needed if the true potential of science for society is to unfold. Theory building provides broader horizons of both explanation and technology, in light of which present definitions of problems and goals may have to be reconsidered.

The finalization model suggests that the distinction between social context and cognitive contents is indispensable for the analysis of the social relations of science. Within a conceptual framework that reduces knowledge to social construction and discards any reference to scientific internalism, one cannot account for the changing patterns of scientific growth and their impact on society. These points deserve more consideration in contemporary sociology of science. To that end, it is not enough to provide a new language that can be applied equally to natural events, experimental data, and theoretical arguments on the one hand, and to social actors and processes on the other (see Latour 1993). The notion that knowledge is categorically separate from power, interest, and social relations must be rehabilitated. This will not jeopardize the social studies of science: it will only eliminate some of the misleading epistemological claims from sociologists who observe science, and perhaps spare us more 'science wars' (Gross and Levitt 1997) from researchers who do science.

See also: Innovation, Theory of; Kuhn, Thomas S (1922–96); Science and Social Movements; Scientific Culture; Truth and Credibility: Science and the Social Study of Science; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Böhme G 1980 On the possibility of closed theories. *Studies in the History and Philosophy of Science* **11**: 163–72

Böhme G, van den Daele W, Hohlfeld R, Krohn W, Schäfer W 1983 Finalization in science. The social orientation of scientific progress. In: Schäfer W (ed.) *Boston Studies in the Philosophy of Science*, Vol. 77. Reidel, Dordrecht, The Netherlands

Bush V 1945 *Science, the Endless Frontier*. US Government Printing Office, Washington, DC

van den Daele W, Krohn W 1998 Experimental implementation as a linking mechanism in the process of innovation. *Research Policy* **27**: 853–68

Elzinga A 1997 The science-society contract in historical transformation: With special reference to 'Epistemic Drift'. *Social Science Information* **36**: 411–45

Funtovicz S O, Ravetz J R 1993 Science for the post-normal age. *Futures* **25**: 739–56

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Schott P, Trow M 1994 *The New Production of Knowledge*. Sage, London

Gillespie C C 1960 *The Edge of Objectivity: An Essay in the History of Scientific Ideas*. Princeton University Press, Princeton, NJ

Gross P R, Levitt N 1997 *Higher Superstition: The Academic Left and its Quarrels With Science*. Johns Hopkins University Press, Baltimore, MD

Hacking I 1983 *Representing and Intervening: Introductory Topics in the Philosophy of Nature Science*. Cambridge University Press, Cambridge, UK

Hanson N R 1965 *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press, Cambridge, UK

Heisenberg W 1971 Der Begriff 'abgeschlossene Theorie' in der modernen Naturwissenschaft. In: Heisenberg W (ed.) *Schritte über Grenzen*. Piper, Munich, Germany, pp. 87–94

Hohlfeld R 1983 Cancer research. A study of praxis-related theoretical developments in chemistry, the biosciences and medicine. In: Böhme G, van den Daele W, Howfeld R, Krohn W, Schäfer W (eds.) *Finalization in Science*. Reidel, Dordrecht, The Netherlands, pp. 93–126

Knorr Cetina K 1995 Laboratory studies. The cultural approach to the study of science. In: Jasanoff S, Markle G E, Petersen J C, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA, pp. 140–66
Knorr Cetina K 1999 *Epistemic Cultures*. Harvard University Press, Cambridge, MA
Krohn W, Schäfer W 1983 Agricultural chemistry. The origin and structure of a finalized science. In: Böhme G, van den Daele W, Howfeld R, Krohn W, Schäfer W (eds.) *Finalization in Science. The Social Orientation of Scientific Progress*. Reidel, Dordrecht, The Netherlands, pp. 17–25
Kuhn T 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
Latour B 1993 *We Have Never Been Modern*. Harvester Wheatsheaf, New York
Leeming W 1997 Revisiting finalization. *Social Science Information* **36**: 387–410
National Academy of Sciences 1972 *Physics in Perspective*, Vol. I. National Academy of Sciences, Washington, DC
Pickering A 1995 *The Mangle of Practice*. University Press of Chicago, Chicago
Rheinberger H-J 1997 *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford, CA
Shapin S 1992 Discipline and bounding: The history and sociology of science as seen through the externalism-internalism debate. *History of Science* **30**: 333–69
Webster C 1976 *The Great Instauration, Science, Medicine and Reform 1626–1660*. Holmes and Meyer, New York
Weinberg A 1972 Science and transscience. *Minerva* **10**: 209–22

W. van den Daele

Finance, Geography of

1. Money and Finance

The geography of money and finance is an important part of what money and finance actually are: the spatial pattern of automated teller machines and bank branches, the passions aroused by currency zones and credit ratings, the flow of financial information through fiber-optic cable networks and satellites, the hierarchy of international and offshore financial centers—all attest to this fact. Yet it is remarkable how little understanding of these different geographies existed until comparatively recently. With a few honorable exceptions, the geography of money and finance was neglected until the 1980s. Since then, however, geography has become a major element of the study of money and finance, for three main reasons.

First, the 1980s were a period of major changes in cultural perceptions of money and finance. Whereas previously finance had been seen as the preserve of stolid accountants, solid bank managers, and squalid dealers, in the 1980s, through the offices of film portrayals, television series, and print journalism, it became recast as a dynamic, even sexy, area of the economy.

Second, money and finance had become such a central element of the geography of modern economies that it was impossible to ignore. The 'financialization' of the economy, especially the tendency of economies to run to essentially short-term financial criteria such as shareholder value, had become so noticeable that it was no longer possible to envisage money and finance as a subsidiary activity whose role was simply to oil the wheels of production.

Third, geographers produced a series of works on money and finance which has proved the value of a distinctive geographic perspective that emphasizes how space creates money, as well as how money creates space. Foremost among these works were David Harvey's *The Limits to Capital* (1982), a Marxian analysis of capitalism which included a seminal chapter on money; subsequent books by Leyshon and Thrift (1997a), McDowell (1997), and Clark (2000); and edited collections from Corbridge et al. (1994) and Martin (1999). In turn, this literature has been paralleled by work in politics, sociology, and even economics. In this article, each of four different but related geographies of money and finance is examined in turn.

2. International Finance

Geography first approached money and finance through the lens of *international finance*. Events like the developing-country debt crisis of the 1980s, the developed-country debt crisis of the early 1990s, and the establishment of the euro zone in the late 1990s demonstrated both the power of international finance and its essentially geographical nature. Geographers gravitated to four main areas of work. Not surprisingly, the first of these was the link between international finance and urbanization implicit in the hierarchy of world cities. Work taking place at a number of levels attempted to show the mechanics of this relationship; for example, specific financial markets were related to the success of particular cities. A further, natural extension of this work consisted of investigations into the global hierarchy of offshore financial centers, which, though smaller in scope, are also key features of the international financial system.

The second area of work concerned new financial instruments. In particular, the geography of new forms of options have proved of interest, not least because they have had direct impacts on the urban and regional fabric through various market imperfections and financial scandals (Tickell 1996, 2000, Pryke and Allen 2000).

One more area of research has been international debt. The debt crises of the 1980s galvanized geographers, and led to a continuing stream of work on finance and development. Then, finally, wrapping all of these areas of work together has been the issue of

governance. The shifting balance of power between the national state and financial markets, the growth of supranational regulatory institutions, and the status of world cities as centers of global governance in their own right are topics that underline just what an important area of work this has become (Thrift and Leyshon 1994, Gilbert and Helleiner 1999), one which overlaps, in particular, with the concerns of international political economy.

3. Retail Finance

Geographers gradually became interested in the national- and regional-scale financial systems through the issue of *consumer finance*. That interest has taken a number of different forms. To begin with, there was a general interest in the changing geographies of retail finance: for example, variations in employment in the industry between states in the USA, the hierarchy of domestic financial centers in the UK, the variable gender and ethnic composition of financial services labor forces, and so on. This work has produced a base from which it has been possible to inquire into the dramatic restructuring of financial services, covering subjects as diverse as the rise of call centers (and their urban and regional consequences) and the diffusion of new financial management ideas. In turn, it has also become possible to intervene in a number of policy domains. The most dramatic intervention has been in policies towards financial exclusion, through careful study of the large numbers of poorer people who do not have access to the formal domestic financial system and who are often further disadvantaged as a result (Leyshon and Thrift 1995, 1997b). While the number of those excluded has fallen in many Western economies, the fact remains that this is a pressing problem only made worse by some of the effects of restructuring, such as the closure of bank branches in poorer areas. The result has been both an interest in changing government policy to force the financial system to acknowledge its civic role and a widespread interest in financial institutions able to be used to construct alternatives: credit unions, microcredit, alternative currencies (e.g., time dollars), small-firm financial schemes, local exchange trading schemes (LETS), and the like.

4. Electronic Finance

The impact of electronic telecommunications on money and finance has been critical. Historically, finance has always been one of the industries most likely to feel the impact of changes in telecommunications, given its emphasis on speed. However, of late, this impact seems to be so great that it is actually redefining what money is by producing new spatial and temporal frames, in rather the same way as in the changeover from coin to paper money (Leyshon and Thrift 1997a, Pryke and Allen 2000). International finance has been in a process of change since at least the 1960s as a result of continual advances in telecommunications which are leading to nearly instantaneous 24-hour financial markets, trading globally. In turn, these developments are producing changes to the current hierarchy of world cities (e.g., Thrift 2000). An important finding made by geographers has been that the increasing amount of electronic information is not necessarily leading to a decline in face-to-face interaction. Especially in the larger international financial centers, which receive very large amounts of electronic information, face-to-face interaction in tightly defined social networks is vitally necessary to lessen the interpretive load in many markets, as well as to allow new stories about what is happening to be created and put into general circulation. Retail finance is also being changed by electronic telecommunications, and, in the long run, the changes here may be even more dramatic than in international finance. Currently a transition is under way from intermediate technologies such as call centers—which still include human elements—to much more heavily automated systems (for example, employing automated voice recognition to replace call center agents). In turn, this transition may well produce a new geography of web-based banking in which consumers will use intelligent agents to undertake much of their financial business (Leyshon and Thrift 1999, Thrift 2000).

Perhaps the most important outcome of electronic money and finance will turn out to be the growth of a combination of software and databases. The effect of the rise of automated decision systems, combined with large synthetic data sets, on the economic circumstances of most individuals has serious connotations, which are now being explored. For example, the concentration of data in the hands of a few private-sector firms raises obvious issues concerning civil liberties and regulation. Geographers specializing in geographical information systems have been intimately involved in the construction of many of these systems and must shoulder some of the responsibilities for the outcome.

In turn, the advent of this new electronic monetary space has contributed to financial exclusion. It seems likely that cities will 'splinter' even further, with serious consequences for social equality. In particular, a large minority of people will not only be excluded electronically but, given the changing status of money, financially as well.

5. Social and Cultural Issues

Finally, the social and cultural elements of the geography of money and finance have increasingly become visible. As an electronic site of exchange, money has increasingly become an evacuated signifier,

a kind of cultural degree zero. In such a situation, social and cultural aspects become more rather than less important, since they become the crucial identifiers of what counts as money in the absence of money as a tangible object. So, in certain senses, money has proliferated; all kinds of different monies, labeled according to their social and cultural usage, have sprung up.

Such developments have produced considerable geographical interest. To begin with, there has been a large amount of work on the historical iconography of money and finance, and especially the role of money in establishing national identity (see Gilbert and Helleiner 1999). Then there has been work on how money is read through a growing discursive apparatus of television programs, newspaper articles, and the like, which demands 'financial literacy' as a ticket of entry (Leyshon et al. 1998). And there are the social and cultural fault-lines created by money, both within its labor force (as, for example, in the matter of gender and sexuality; see McDowell 1997) and in its ability to produce social and cultural capital for certain elite segments of that labor force, which has enormous spread effects in cities like London and New York (Leyshon and Thrift 1997a).

6. *Further Directions of Research*

Further directions of research on money and finance are clear. The first of these is the increasing privatization of welfare, with welfare functions that were often produced or guaranteed by the state being run down, leading to a situation in which individuals increasingly have to make their own private arrangements for health and old age. Thus we are seeing the birth of what Clark (2000) has called '*pension fund capitalism*,' which, especially given the general aging of populations in many countries of the world, is becoming a pressing economic issue with profound consequences for the distribution of income, and for life chances. The second direction is *software*. Increasingly, the financial system will be run by machine codes, yet remarkably little is known about the agency of code or how to regulate its production. Here is a pressing theoretical and practical issue. The third direction is *speed*. Increasingly, electronic money is producing instantaneous exchange. There is much hyperbole about this state of affairs but precious little research informing us on how social skills and practices are adjusting and thereby changing what money is. Then, finally, there is the matter of financial exclusion. As already noted, this is a pressing issue which demands the design of different kinds of financial institutions. Geographers have been very active in this area and have been able to demonstrate the full range of competencies that they now have concerning money and finance, competencies that as recently as the 1980s would perhaps have been regarded as surprising.

See also: Consumer Economics; Financial Institutions in Economic Development; Retail Trade

Bibliography

Allen J, Pryke M 1994 The production of service space. *Environment and Planning, D. Society and Space* **12**: 453–75

Clark G L 2000 *Pension Fund Capitalism*. Oxford University Press, Oxford, UK

Corbridge S, Thrift N J, Martin R L (eds.) 1994 *Money, Power and Space*. Blackwell, Oxford, UK

Gilbert E, Helleiner E (eds.) 1999 *Nation-States and Money*. Routledge, London

Harvey D 1982 *The Limits to Capital*. Blackwell, Oxford, UK

Leyshon A, Thrift N J 1995 Geographies of financial exclusion: Financial abandonment in Britain and the United States. *Transactions, Institute of British Geographers, NS* **20**: 312–41

Leyshon A, Thrift N J 1997a *Money/Space: Geographies of Monetary Transformation*. Routledge, London

Leyshon A, Thrift NJ (eds.) 1997b Special issue on financial exclusion. *Environment and Planning, A* **29**(3)

Leyshon A, Thrift N J 1999 Lists come alive: Electronic systems of knowledge and the rise of credit-scoring in retail banking. *Economy and Society* **28**: 434–64

Leyshon A, Thrift N J, Pratt A 1998 Reading financial services: Texts, consumers and financial literacy. *Environment and Planning, D. Society and Space* **16**: 29–55

Martin R L (ed.) 1999 *Money and the Space Economy*. Wiley, New York

McDowell L 1997 *Capital Culture: Gender at Work in the City*. Blackwell, Oxford, UK

Pryke M, Allen T 2000 Monetized time–space: derivatives: Money's new imaginary. *Economy and Society* **29**: 264–84

Sidaway J, Pryke M 2000 The strange geographies of emerging markets. *Transactions of the Institute of British Geographers, NS* **25**: 187–201

Thrift N J 2000 Less mystery, more imagination: The future of the City of London. *Environment and Planning, A* **32**: 381–90

Tickell A 1996 Making a melodrama out of a crisis: Reinterpreting the collapse of Barings Bank. *Environment and Planning, D. Society and Space* **14**: 5–33

Tickell A 2000 Dangerous derivatives: Controlling and creating risks in international money. *Geoforum* **31**: 87–99

N. Thrift

Financial Accounting and Auditing

After situating accounting and audit as social practices, this article examines research objectives and methodology in these fields.

Accounting and audit are closely linked social practices. By means of the information it produces, accounting assumes two principal functions: the ren-

dering of accounts and the taking of financial decisions. Standardization and audit then come into play, since accounting information must necessarily be a reliable and believable reflection of the events and conditions it portrays. The initial role of auditing is to assess the reliability and credibility of the accounting portrayal. Taken together and in their complementary roles, accounting and audit are major elements of social regulation.

The resulting importance of accounting and audit in contemporary societies, for better or for worse, has made them the object of active research. Depending on the research methodology employed, research may be conducted from a historical, sociological, or economic perspective.

1. Accounting and Auditing as Social Practices

The origins of accounting have been lost in the mists of time, predating even figures themselves. Indeed, we have been counting for a very long time (and continue, moreover, to do so) without the help of figures. However, generally it is agreed that modern accounting was born at the end of the Middle Ages in Northern Italy with the 'invention' of double-entry bookkeeping. A chapter of the Franciscan monk Luca Pacioli's work on mathematics (the *Summa de Arithmetica*, *Geometria*, *Proportion e Proportionalita*), published for the first time in 1494, introduces this invention which was to enable accounting to fulfill its present-day functions.

1.1 The Functions of Accounting

In short, let us say that accounting fulfills, and has probably always fulfilled, two main functions: on the one hand, a rendering and control function, and on the other, an aid to decision-making function.

Its first function may be related to the old master–servant relationship. When a servant is given any sort of task or assignment by his master, he must, sooner or later, give a narrative or numerical account of the manner in which he has carried it out. Accounting is precisely one of the instruments which enables a servant to give an account to his master, and the master to check his work.

In the light of this, it is not surprising that modern accounting was 'invented' at the end of the Middle Ages in the big towns of northern Italy: Florence, Genoa, and Venice. Their bankers and merchants, at the head of veritable multinationals, operated all over the Mediterranean and northern Europe. To carry out business activities that were often far away from their home-towns successfully, they were obliged inevitably to rely on agents and representatives that their invention of double-entry records enabled them to control more effectively.

Similarly, it is easy to understand the importance accounting assumed in the nineteenth century when joint-stock companies mushroomed and the split between partner–capital contributors and directors appeared. Partners removed themselves, or were removed, from the running of companies, and directors were obliged legally to provide them with information on their administration at regular intervals by the production of balance sheets and income statements. What financial theory terms an 'agency relationship,' therefore, became a legal requirement. To use agency theory vocabulary, one might say that the partners played the role of principals, and the directors that of agents. Accounting is, therefore, the instrument that enables the principal to control what his agent has done, and the agent to give an account of what he has done. In this relationship, the accountant occupies a somewhat unusual position since he draws up the accounts—destined for the principal—of the agent who must give an account. English expresses this triangle very well since it has three words, to designate respectively he who has an account to give (the accountor), he to whom the account must be given (the accountee), and finally, he who does the accounts (accountant).

This 'accounting game' is obviously a complex social game to the extent that the person giving the account possesses more information than the person to whom the account must be given, and to the extent also that he may be tempted to manipulate the accounts and present them in a way favorable to himself. The accountant, who draws up the accounts, occupies, despite himself, the position of mediator, and so it is a good idea for him to belong to an organized profession with a strong sense of ethics.

Accounting's aid to decision-making function is as old as the preceding one and linked closely to it, moreover. The bankers and merchants of northern Italy, mentioned above, had to join forces to amass the capital needed to finance their businesses. They obviously had to share the profits, and to do so, they needed an instrument which would enable them to calculate what these were. The new double-entry accounting was also capable of responding to this need and naturally still exists in today's big companies. A company's annual accounts allow shareholders not only to keep track of the activities of their directors, but also to take the decision which is, as it were, the high point of their annual general meeting, the decision concerning the allocation of profits.

The two main functions of accounting have taken on a dimension that matches the responsibilities of today's companies. Today, companies must give an account not only to those who have entrusted them with their capital, but also to many other partners such as the Government, their suppliers, customers, employees, and in more general terms, the public. So their accounts (and in particular their balance sheet and income statement) have to satisfy various needs for

information which correspond to very different models and control and decision-making rationales. More specifically, they should, for example, allow:

(a) investors, as previously mentioned, to assess the way the directors administer the business and to take into account their decisions regarding purchasing, and the retention or sale of stocks and shares;

(b) banks and financial establishments to judge the financial solidarity and reimbursement capacity of the company;

(c) employees to analyze its social effectiveness and its durability;

(d) suppliers to understand its capacity to honor its debts;

(e) customers to measure its capacity to fulfill the orders it receives; and

(f) the Government to control its taxable profit.

At this stage, one might say that this type of accounting is the system of information through which a company, or any other organization (including nonprofit-making organizations), gives an account to its economic and social partners, and, in doing so, submits itself to their control and allows them to take decisions. It is because the information it produces is, for the most part, of a financial nature that it is described in this way. Moreover, financial accounting which responds to the need of third parties for information can be distinguished from the type of accounting known as management accounting which responds more specifically to a company director's need for information. If it is to assume its vocation of providing information for third parties effectively, financial accounting must obey recognized rules and the application of these rules must be controlled. This is the role of the audit.

1.2 Standardization and Auditing

In all major developed countries, accounting obeys standards fixed by the accounting profession and/or public authorities. In addition to purely national standards, there are now international standards produced by the *International Accounting Standards Committee* (IASC). The aim of standardizing corporate accounting is to make the financial information it produces intelligible and credible. However, to the extent that it is the companies themselves who do their own accounting, standardization is not an adequate condition for the credibility of their accounts. The accounts must be audited.

Auditing a company's accounts consists of checking that they have been drawn up in accordance with current procedures and standards and that they give a true and fair view of the company's financial situation. Obviously, the audit cannot consist of redoing the company's accounts. The auditor proceeds essentially by means of investigation and cross-checking in order to identify anomalies likely to call into question the reliability and credibility of the accounts on which they must give an opinion. This opinion, which is the result of these investigations, serves to certify the accounts. Certification can be positive or negative, and accompanied by observations and reservations of varying degrees of importance. It should allow the recipients of the accounts, and in particular investors, to assess their validity.

In the course of the twentieth century, the accounting audit received legal consecration in very many countries. Joint-stock companies must have their accounts audited by independent professionals. In Anglo-Saxon countries, the audit assignment is undertaken by certified (US) or chartered (GB) accountants. These providers of credibility form what the sociologists call a 'profession' in the sense that they take on a task which is in the general interest and are, for this reason, subject to requirements regarding their competence and professional code of ethics. Auditors' codes of ethics place great emphasis on their independence.

The audit has expanded considerably its initial accounting vocation and in the second half of the twentieth century has undergone considerable, protean development, generating a major economic sector dominated at world level by several large consultancies such as Arthur Andersen, Price Waterhouse Coopers, Deloitte Touche Tohmatsu, Ernst and Young, KPMG, (known as the 'Big Five'). In the widest sense of the term, the audit now covers all 'listening' activities ('audit' comes from the Latin 'audire,' to hear), evaluation, supervision, and external control of organizations. To the accounting audit previously mentioned has been added the management audit, the quality audit, the environmental audit, etc. Moreover, all organizations, and not simply companies, make use of the audit.

Together, therefore, accounting and auditing form one of the mechanisms which contribute to maintaining links of collaboration and confidence between the various economic and social players, whether these be individuals or organizations. The multiplication and increasing complexity of the relationships between social players has created an enormous demand for accounting and auditing. One might even say that modern society is an 'audit society' (Power 1997).

2. Accounting and Auditing as Research Subjects

Accounting has never been a fixed, immobile technique. For five centuries at least, if one traces its modern origins back to the Middle Ages, it has adapted constantly to developments in companies and their environment. This adaptation is, in a sense, the expression of a form of applied research that one might qualify as spontaneous, carried out by the

practicians themselves, heads of companies, and accountants. But it was not until the twentieth century that accounting and audit became objects of research for the social sciences, of interest in particular to historians, sociologists, and economists.

2.1 Accounting History

During the twentieth century, accounting has been the subject of very many historic investigations. All the major academic accounting reviews (*Abacus*, *Accounting and Business Research*, *Accounting-Organization and Society*, *The Accounting Review*) publish articles on the history of accounting. Entire reviews are even dedicated to it such as *The Accounting Historian's Journal* and *Accounting Business and Financial History*. Through these many publications, it would seem that the history of accounting can be approached in several different ways.

Certain historians are interested in accounting as a recording technique and pay particular attention to the genesis and evolution of the double-entry system that still is the most striking aspect of accounting. Other historians wish to see accounting as something more than pure technique and consider it to be an instrument of social and economic regulation. From this point of view, its history is not an autonomous history and is part of several histories: the history of the economic system, the history of companies, the history of techniques, and the general history of ideas.

One of the first to adopt this point of view was the economist Werner Sombart who postulated that there was a very strong link between the development of the capitalist system and that of accounting, going so far as to claim that one was not conceivable without the other, that double-entry accounting allowed the company to be set up as a place of economic calculation, to give a numerical expression to the profit objective, and thence to rationalize the allocation of resources. Although contested, Sombart's thesis has contributed to reviving the history of accounting.

2.2 Accounting Sociology

Since the past 1970s, and above all in Great Britain, the sociology of accounting has undergone major developments. These developments are sufficiently differentiated for two distinct schools of thought to be identifiable within this field of research: an interpretative, and a radical school.

The interpretative school, represented by authors such as Giddens, Morgan and Foucault, aims to interpret accounting practices, in particular national standardization practices, through reference to their technical, economic, social, cultural, and political context. This type of research tries to highlight the various conditions which influence accounting and which mean that it cannot be considered as an isolated technique. Although it relies mainly on sociological methods, in particular the monograph, it also uses the historical method. Its main organ is the review *Accounting Organizations and Society*.

The radical school, which is Marxist or neo-Marxist in inspiration (with references to Habermas and Marcuse) considers that accounting and auditing are mechanisms designed and implemented by the dominant social groups to defend their interests and strengthen their domination. It is, therefore, a question of dispelling the mystery surrounding the neutrality and objectivity often attributed to them and of highlighting their ideological function. Researchers who position themselves within this trend willingly support the oppressed (poor countries, women, and in the United States, blacks) and lend them their expertise. From a methodological point of view, they make highly eclectic use of social science approaches. Intentionally provocative, and with a leaning towards intellectualism, they go so far as to call upon poetry to illustrate and defend their theses. Their main organ is *Critical Perspectives on Accounting*.

2.3 Accounting Economics

Economic research has given birth to what is commonly known as positive accounting theory. According to Jensen (1983), this is a theory capable of explaining '…why accounting is what it is, and what consequences accounting figures have on people and on the allocation of resources.' At the end of the 1970s, the theory was promoted by two professors from the University of Rochester, Ross L. Watts and Jerold L. Zimmerman.

The research conducted in the wake of these two pioneers was inspired principally by the agency theory and aimed to highlight the determinants of choices of accounting methods made by company directors, where standardization authorizes such choices. In agency theory, the company is a 'nexus' of contracts, formal and informal, concluded mainly between the directors and the stockholders. To the extent that accounting produces information that serves to draw up and control the execution of these contracts, company directors are said to pursue their personal interests and choose methods that give them an advantage.

Within the context of this problem, it is possible to formulate many hypotheses on the accounting behavior of company directors. One might suppose, for example, that directors linked to their company by a profit-sharing contract would choose accounting methods that tend to increase present profits at the expense of future profits. One might also suppose that

company directors whose debt ratios are bordering on the thresholds tolerated by the lenders would once again choose accounting methods that increase the profits and diminish the debt ratios. Such hypotheses may be subjected to the statistical tests common in economic research.

Finally, positive accounting explains the choices of accounting methods made by company directors by reference to various economic and financial characteristics of the companies: their size, level of debt, Stock Exchange listing, etc.

This research trend has been the subject of very many methodological and epistemological criticisms (see, e.g., Christenson 1983), but it remains the mainstream of accounting research in the United States. The resulting research papers are published mainly in the *Journal of Accounting Research* and in the review specially created by Watts and Zimmerman, the *Journal of Accounting and Economics*.

To these various types of research (historic, sociological, and economic), that are cognitive in nature, may be added applied research, the object of which is to define accounting concepts, principles, and procedures. Research aims to define a theoretical framework for accounting practice and is, to a greater or lesser extent, linked to the work of the standardization bodies. In contrast to the preceding types of research, it is described as normative.

See also: Competitive Strategies: Organizational; Conflict: Organizational; Corporate Culture; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; Information and Knowledge: Organizational; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; International Business; Monetary Policy; Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Strategy: Organizational; Technology and Organization

Bibliography

Burchell S, Clubb C, Hopwood A, Hughes A, Nahapict J 1980 The roles of accounting in organisations and society. *Accounting Organizations and Society* **5**(l): 5–27

Christenson C 1983 The methodology of positive accounting. *Accounting Review* **58**(l): 1–22

Dunn J 1996 *Auditing*: *Theory and Practice*. Prentice Hall, New York

Edwards J R 1989 *A History of Financial Accounting*. Routledge, London and New York

Flint D 1988 *Philosophy and Principles of Auditing*. Macmillan, London

Gray R, Owen D, Adams C 1996 *Accounting and Accountability*. Prentice Hall Europe, London

Hopwood A G, Miller P (eds.) 1994 *Accounting as Social and Institutional Practice*. Cambridge University Press, Cambridge, UK

Jensen M C 1976 Reflections on the state of accounting research and the regulation of accounting. In: *Stanford Lectures in Accounting*, Stanford University Press, Stanford, CA, pp. 11–19

Jones T C 1995 *Accounting and the Enterprise: A Social Analysis*. Routledge, London and New York

Mathew M R, Perera M H 1991 *Accounting Theory and Development*. Chapman and Hall, London

Power M 1997 *The Audit Society: Rituals of Verification*. Oxford University Press, New York

Roslender R 1992 *Sociological Perspectives on Modern Accountancy*. Routledge, London

Ryan B, Scapens W, Theobald M 1992 *Research Method and Methodology in Finance and Accounting*. Academic Press, London

Watts R L, Zimmerman J L 1978 Towards a positive theory of the determination of accounting standards. *Accounting Review* **53**(January): 112–34

Watts R L, Zimmerman J L 1986 *Positive Accounting Theory*. Prentice Hall, Englewoods Cliffs, NJ

B. Colasse

Financial Institutions in Economic Development

The financial system can be considered as a composite of financial activities and the institutions that implement and regulate these activities. Financing of firms is one activity of the system, and capital (securities) markets and banks are their principal financing sources. The sociology of financing economic development examines the institutions that shape the development of firms and nations—particularly in emerging markets.

There have been several sociological analyses of development, including firm, industry, and country studies; but what remains largely unexplored is the role of banks and other financial institutions in this process, and how this relation is influenced by national and international contexts and events, to say nothing of local populations and their initiatives (cf. Evans 1995, Gereffi and Korzeniewicz 1993). After reviewing some theoretical approaches to financing growth and economic development (development) this entry focuses on banks, the principal institution in financing development in most emerging countries. It will identify studies that are rooted in social structure and that focus on such questions as how banks emerge, the sociopolitical reasons for their involvement, which forms are prevalent, and how development is financed. To conclude, some policy-related research challenges will be posed.

1. The Financial Context

Financing of households, businesses, and the government are the three main financial activities in any country. Financing is mainly achieved through capital

markets and/or banks as well as own savings by economic units. A firm can draw upon its savings to finance its production, it can issue stocks, bonds, or it can borrow from banks. Households usually borrow from banks, while governments issue bonds and/or borrow from banks. Cross-nationally and cross-temporally, there are significant differences in these activities. Businesses in the United States and the United Kingdom use capital markets (securities issues), while those in France, Germany, and Japan mainly use bank loans. Among these countries there are differences in bank–industry relations. German banks are inclined to intervene in industry because, in addition to being important shareholders, they control access to all sources of finance. In France, on the other hand, businesses have access to a wider array of public and private institutions whose lending terms are shaped by the government (Zysman 1983).

There is little consensus over which system contributes most to the overall development of an economy. Evidence indicates that having a robust banking and capital market is correlated with economic development (Demirguc-Kunt and Levine 1999). Levine (1997) reminds us of one classic controversy, i.e., whether Germany's development in the latter half of the nineteenth century was more rapid than Great Britain's because its economic structure was based on banks with close links to industry, whereas Great Britain's was more securities market-based and ties between banks and industry were less intimate. As countries mature and evolve into service-oriented economies, there is a growing belief that capital markets are increasingly more effective at intermediation than banks, as Ness (1974) and Haber (1997) illustrate for Brazilian development. Nonetheless, banks are essential to small and medium sized firms as they are less able to tap capital markets for funding. Similarly in emerging countries where there tends to be a lack of a tradition of investing, where capital markets are absent, or are only beginning to mature, banks are the principal financial intermediaries and have been the primary engines for financing economic development. They are heavily linked through equity and management to industry. In these countries the state may own a good portion of the banks, and share the market with foreign or large domestically owned banks. In China nearly 100 percent of the banks belong to the state; in India 70 percent are state owned. In contrast, in the United Kingdom more than 50 percent of the banks are foreign owned (*Economist* 1997). For an open-banking country in which foreign banks are treated the same as domestic banks, the United States has relatively little foreign ownership (less than 10 percent) of separately chartered banks (Barth et al. 2000). And as the twenty-first century begins, throughout the global economy banks are growing in size and becoming more multifaceted in their operations, often as a result of huge mergers and acquisitions which themselves are fueled by the advent of new regional markets (Makler 1999). Bank growth is in response to an increasing and rapid demand for huge, often multilateral, financing coupled with banks' desire to provide all or one-stop financial services.

1.1 Banks in Emerging Economies: Some Anomalies

Economists view banks as the principal financial intermediary between savers and investors. Banks obtain deposits from the public and invest what they receive in loans and securities. But their role is broader. They are important centers in the formation of wealth, the distribution of economic power, and they influence the way societies change. Political scientists see banks as a powerful interest group attempting to influence public policy or elections. Anthropologists see banks as 'unorganized,' informal institutions that provide capital to small rural or urban businesses.

Sociologists have focused mainly on bank power and have elaborated Berle and Means' analysis of corporate ownership and control. Few have studied banks as institutions or have explored the meaning of their wide array of relations. Few have related form to performance. Even fewer have focused on emerging countries where banks have played a significant role in change and development, as Mizruchi and Stearns (1994) recognized. This would require comparisons among a diverse selection of countries in addition to those that have a similar standard of living such as Germany, Japan, the United Kingdom, and the United States (Levine 1997).

2. Development and Banks in the Social Sciences

While few studies of the role of banks in growth and development are found in sociology, a number of distinctive theoretical approaches have emerged in the general study of financial institutions that illuminate the conditions under which one or another type of bank will appear and explain what differences this makes for economic development.

In economics, one debate has to do with supply and demand. Some have characterized banks as demand-following, or as passive, permissive intermediaries that merely provide services or loans to creditworthy borrowers. Economists such as Goldsmith (1969), Gurley and Shaw (1967), and Patrick (1966) saw banks as supplying capital, encouraging efficient credit allocation, increasing accumulation, and sometimes assisting in the founding of new enterprises or in promoting mergers, but only upon the demands of their clients. Others viewed banks as supply leading. Among these was Schumpeter (1961), who brought credit creation by banks to the fore, identifying them as supply-leading and growth-inducing agents of development. In creating new sources of credit, banking systems were ranked slightly behind those of

entrepreneurship and innovation. For Schumpeter, well-managed banks fuel technological innovation by identifying and funding entrepreneurs who were identified as most likely to introduce and implement new products and production processes sucessfully. But in his discussion of commercial banks he simply mentioned the ability of banks to create purchasing power, without analyzing their method of capital accumulation or their relations with entrepreneurs.

It was Gerschenkron (1962) who introduced the importance of context and institutional form. In an often-cited explanation of the role of banks in development, he observed that the historical timing of a nation's industrialization determines which economic actors and social institutions will be involved. One of his principal hypotheses was that the more backward an economy relative to others in the same period, the greater the likelihood that financial institutions would be created to supply capital in advance of demand. In the transformation of once backward countries to industrialization, the role of the state, and the emergence of specialized private and state banks designed to increase the supply of capital and provide better-informed entrepreneurial guidance to nascent industries were also emphasized as important to development. As for form, he saw the pressure for industrialization in latecomers resulting in larger industries and banks. And because of the magnitude of bank lending to industry Gerschenkron predicted that banks would acquire 'a formidable degree of ascendancy over industrial enterprises which extended beyond the sphere of financial control and into that of entrepreneurial and managerial decisions' (Gerschenkron 1962, p. 14). Thus institutional dominance was introduced into his analysis.

Classical Marxist theorists also viewed banks as dominant. Lenin saw banking as concentrating into a 'single collective capitalism,' as predatory, and as controlling industry, causing capital concentration and conglomeration. To Hilferding (1981), finance capital (or that capital supplied to industrialists and controlled by financial intermediaries) was the means by which big banks forged concentration among non-financial corporations and created a new class 'under the direction of high finance.' And through investment in more and more basic sectors, banks propelled the state as their agent into the world market (Hilferding 1981, Harvey 1982). In applying parts of this theory to the United States, sociologists Mintz and Schwartz (1985) and Fennema (1982) found in some periods extensive interlocking between the directorates of major US corporations and banks, that industry was dependent on banks because of bank control over financial resources, capital flows and knowledge of money markets, and that American banks interlocked with important foreign banks to form a discrete consortium in world finance. But like Gerschenkron, neither Hilferding, Lenin, nor more recent scholars have demonstrated successfully that such a strategy will result in finance capital's economic or political hegemony. To the contrary, in studying bank transformation in Brazil, Garman and his colleagues (1998) illustrate that while banking concentration resulted during privatization there was little evidence of finance capital's domination of the process.

The typical institutional form that emerged in Europe in the nineteenth century was the financial conglomerate or multi-sector company, transacting in different markets owned and controlled by a family-owned bank. In their quest for growth and development, governments have been bankers, have recruited domestic and foreign banks, and encouraged financial conglomeration and financial concentration. The classical cases in financing development are Germany and Japan where, in preparation for World War II, military governments nurtured large banks and encouraged their spread into and control of non-financial sectors in order to develop these sectors (Hadley 1970, Roe 1994). In Japan, the conglomerates were known as *zaibatsus* and today they are referred to as *keiretsus*. In South Korea they are *chaebols* and in Latin America they are usually known as *grupos* (Leff 1978, Granovetter 1994). In most emerging countries where the state is a formidable economic actor, state-owned banks share the market with large family-owned and foreign banks. These banks were recruited by the state, which provided subsidies to encourage the development of basic, agro-industrial, and export industries. As a result, the families achieved economic security and upward mobility, and the foreign investors became more embedded in the economy. The bank, usually at the core of the conglomerate, served primarily as a direct source of capital for member firms that were faced with increased capital requirements posed by industrialization and export markets. It also served as a source of employment for family members and close friends (cf. Gerlach 1992, Lamoreaux 1994). In exchange for the promise of investment, foreign banks also received 'benefices' such as authorization to open branches or to purchase domestic banks. During Brazil's industrialization in the 1970s foreign capital formed alliances with prominent family-owned firms in order to access local markets. In exchange these firms and their controllers were provided with technology, capital, and access to international markets (Evans 1979, Cardoso and Faletto 1979). What soon became known as Brazil's economic miracle was largely fostered by the banking sector, its conglomerates, and its alliances (Makler 2000). The state–banking sector relation in Brazil and other emerging countries is thus more symbiotic than in some advanced economies (cf. Hall 1991, Weber 1968).

Interlocking boards of directors and the resulting exclusivity favor large business growth over small. But this pattern seems to be tempering. Davis and Mizruchi (1999) found that the interlocking between the boards of American banks and major industrial

corporations declined significantly from 1982 to 1994, reflecting a strategic shift by banks away from lending to major corporations. While corporations are turning more to capital markets for their financing, interlocking and, as a consequence, lending is likely to be increasing between banks and smaller firms that are less able to secure funds from capital markets (cf. Barth et al. 2000). In relations between banks and smaller firms, commercial transactions are more likely to be embedded in social attachments that improve loan acquisition, offer favorable credit terms, and motivate smaller firms to achieve, as Uzzi (1999) has shown.

Another topic that has intrigued sociologists is the impact of foreign capital on social structure in emerging nations. Some nations, in their quest for rapid economic development, have encouraged foreign direct investment, including banks, because of the promise of technological and industrial growth that they represent. Foreign banks have been cast as stimulating competition in their sector, better able to withstand macroeconomic shocks because of their geographic diversity, and as relieving scarce regulatory resources because they are regulated by their home country (World Bank 2000). Yet some development economists as well as some sociologists have been dubious about foreign capital's contribution, while others have viewed it more favorably. In studying late nineteenth-century Chile, Zeitlin (1984) found that important family groups that had merged with foreign capital to develop the nitrate industry at first flourished but later were economically and politically weakened when the government, in an attempt to curb an export crisis, threatened nationalization and limitations on further foreign capital participation. However, other studies suggest that associations with foreign capital can foster development, local entrepreneurship, and provide mobility opportunities for local businesses. In comparing Brazil, India, and South Korea, Evans (1995) demonstrates that foreign capital and national governments can perform a nurturing or a midwifery role in the development of certain economies. In an earlier study of Brazil, Evans (1979) found that foreign capital formed alliances with prominent family-owned firms in order to access local markets. In exchange, these firms were promised technology and access to international markets.

There are few sociological studies of foreign banks in emerging countries, and those that exist are inconclusive as to whether banks contribute to development. In the 1990s foreign banks proliferated in Latin America. Guillén and Tschoegl (2000) have documented the dramatic rise of Spanish banks but advocate further research to understand their investment strategies. In Brazil, foreign banks have spread but it is unknown whether their lending, like many of the leading domestic banks, is mostly directed toward larger foreign or state-owned corporations, or whether it is becoming more widespread to include small and medium sized businesses. In that country, banking secrecy provisions preclude detailed analysis of size and regional distribution of loans, and for their part banks resist studies of their lending, presumably fearful that such information will be used by their competitors (Ness 2000).

Values, politics, and laws underlie this proclivity to conglomerate, to concentrate finance capital, and to allow foreign or outside banks to practice in national markets. In some societies, centralization and concentration of economic and political power are tolerated while in others they are not. Because of their feudal heritage, a populist, anti-big-bank ethos has historically been absent in Japan and South Korea while instead large enterprises are celebrated (cf. Roe 1994, Orru et al. 1991). In several Latin American countries centralized power and large private and public institutions are also preferred. In Brazil, because of its patrimonial heritage, large financial conglomerates flourish, and Brazilian federalism leads development-prone administrations to decentralize financial power to them in exchange for their economic and political tribute (Makler 2000). In contrast, American ideology has distrusted large monopolies and banks and has favored smaller, encapsulated, locally accessible, and locally embedded institutions. Since the early nineteenth century the American banking system has been fragmented, consisting of numerous small banks without extensive branching and a few large banks that have been hampered legally from capitalizing in other financial services (Roe 1994). The mixing of banking and commerce is prohibited, and although banks can now engage in securities and insurance their spread into real estate activities is restricted. The state 'recruits' banks to invest in nonfinancial sectors such as community housing and urban development programs, not through continuing subsidies and dispensations as in patrimonial systems but through regulation.

This ideology is rooted in American federalism that fostered fragmented, decentralized, and 'separate but equal' banking where each state chartered and protected its own banks while excluding those from other states or foreign countries. National legislation echoes this ideology. The National Bank Acts of 1863 and 1864, and the Glass-Steagall Act of 1933 restricted banks from spreading geographically and sectorally into other financial and non-financial sectors, thus leading Roe (1994) to predict that the United States might have achieved greater growth earlier with stronger, less fragmented, money centered banks. Yet eleven Southeastern states closed their markets to outside banks. This gave resident banks a monopoly within the region (known as the Southeast Compact). As a result these banks registered greater profitability, made proportionately more property loans, and for a decade (1980–90) more economic growth occurred in this region than elsewhere in the country (Barth and Jahera 1994). Southeastern growth subsided with the

debut of national banking deregulation and banking liberalization in the 1990s. This legislation repealed six decades of constraining legislation and resulted in nationwide banking, mergers, acquisitions, and the spread of banks into securities, insurance, and other sectors. Concentration is increasing and the multi-sector financial conglomerate is likely to become the dominant organizational form, as in many European countries, Latin America, and other emerging countries. With the advent of electronic banking and nationwide conglomeration, the relation between local deposits and local financing will become increasingly tenuous.

3. *Challenges*

A growing apprehension with globalization and its effects on the development of national markets is casting an important challenge for sociologists as it did for economists who were concerned with post-World War II recovery half a century ago. Moving from studies of financial structure and power, sociological research at the beginning of the twenty-first century should focus on outcomes of financial intermediation, and these should be analyzed in ways that could inform public policy.

First, the ways that economic units, be they regional markets, nations, or large or small businesses, finance their starts and their growth should be clarified. For mature, more service-oriented economies, capital markets (securities) are viewed as more efficient but for emerging economies and for new businesses banks (loans) are the principal form. Is this the case, or will there be increasing multi-sourcing as access to financial markets is widened through deregulation, technological innovation, and new financial institutions?

Second, monitoring whether major changes in local and national economies are accompanied by changes in financing enables identification of new forms of financial intermediation that finance development. Granovetter (2001) and other scholars (cf. Kaplan 1999, Kenney and Florida 2000), for example, found that the rapid growth of California's Silicon Valley information technology industry depended upon the creation of a new form of financial intermediation: venture capital. Early venture capitalists were mainly successful engineers who invested their profits and often took equity and managerial positions, unlike most financiers who hardly associated with their investments.

Third, in studying emerging countries, recognition of the embeddedness of market segmentation or the duality of the financial market is challenging. Paralleling the two forms of financial intermediation (banks and securities) is informal financing. This is particularly prevalent in emerging markets among poorer populations that have little, if any, access to (or knowledge of) formal financial institutions that serve large businesses (World Bank 1989, Carstens 1995). Biggart (2001) studied micro credit programs such as rotating savings and credit associations that lend modest amounts of money for entrepreneurial business activities to poor women who are organized in solidarity groups. By starting with the demand side, with credit seekers, pertinent financial institutions can best be identified and from there related to performance. Answering these two questions is a promising start: What is the relevance of one financial institution compared to another for development; and how, in turn, does development change and evolve to affect that institution and its organization? But more challenging is understanding the relation between financial development and social organization. Capital markets create dispersed savers (investors) who are interested primarily in financial return and risk, while banks have been more locally embedded and attentive to the special needs of their borrowers. Capital market development might well result in a depersonalized form of capitalism. The banking industry is becoming more concentrated in nationwide conglomerates rather than in locally oriented institutions, and globalization has brought foreign banks to many nations. For emerging economies these banks represent financial efficiency and economic stability, yet for all nations the sovereignty of their regulatory institutions and their economic development has been challenged.

See also: Development, Economics of; Development: Rural Development Strategies; Development: Socioeconomic Aspects; Economic Sociology; Finance, Geography of; Foreign Investment: Direct; Spatial Pattern, Analysis of

Bibliography

Barth J R, Jahera J 1994 Barriers to the integration of banking markets: The case of the United States. *CEPR Working Paper 440*. Center for Economic Policy Research, Stanford, CA

Barth J R, Nolle D E, Root H L, Yago G 2000 *Choosing the Right Financial System for Growth*. Milken Institute, Santa Monica, CA

Biggart N W 2001 Banking on each other: The situational logic of rotating savings and credit associations. In: Wagner J A III (ed.) *Advances in Qualitative Organizational Research*, Vol. 3, pp. 129–53

Cardoso F H, Faletto E 1979 *Dependency and Development in Latin America*. University of California Press, Berkeley, CA

Carstens C M 1995 *Las Finanzas Populares en México: El Redescubrimiento de un Sistema Financiero* [*Popular finance in Mexico: The Rediscovery of a Financial System*]. Centro de Estudios Monetarios Latinoamericanos, Editorial Milenio, Instituto Tecnológico Autónomo de México, Mexico City

Davis G, Mizruchi M 1999 The money center cannot hold: Commercial banks in the US system of corporate governance. *Administrative Science Quarterly* **44**: 215–39

Demirguc-Kunt A, Levine R 1999 Bank-based and market-based financial systems: Cross-country comparisons. *World Bank Policy Research Working Paper 2143*. World Bank, Washington, DC
Economist 1997 Banking in emerging markets. *The Economist*, April 12
Evans P 1979 *Dependent Development: The Alliance of Multinational, State and Local Capital in Brazil*. Princeton University Press, Princeton, NJ
Evans P 1995 *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press, Princeton, NJ
Fennema M 1982 *International Networks of Banks and Industry*. Nijhoff, The Hague, The Netherlands
Garman C, Silva Leite C, Silva Marques M 1998 *Impactos das Relacoes Banco Central x Bancos Estaduais no Arranjo Federativo Pos-1994* [*The Impacts of Central Bank-State Bank Relations within Brazil's Federalist Arrangements after 1994*]. Paper presented at the XXII Encontro Anual da ANPOCS (Associacao Nacional de Pos-Graduacao e Pesquisa em Ciencias Sociais), October
Gereffi G, Korzeniewicz M 1993 *Commodity Chains and Global Capitalism*. Greenwood, Westport, CT
Gerlach M 1992 *Alliance Capitalism: The Social Organization of Japanese Business*. University of California Press, Berkeley, CA
Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Harvard University Press, Cambridge, MA
Goldsmith R 1969 *Financial Structure and Development*. Yale University Press, New Haven, CT
Granovetter M 1994 Business groups. In: Smelser N, Swedberg R (eds.) *The Handbook of Economic Sociology*. Russell Sage Foundation and Princeton University Press, Princeton, NJ
Granovetter M 2001 A theoretical agenda for economic sociology. In: Guillén M, Randall Collins R, England P, Meyer M (eds.) *Economic Sociology at the Millenium*. Russell Sage Foundation, New York
Guillén M, Tschoegl A 2000 At last the internationalization of retail banking? The case of Spanish banks in Latin America. Working paper, Wharton School, University of Pennsylvania, PA
Gurley J, Shaw E 1967 Financial structure and economic development. *Economic Development and Cultural Change* **15**(3): 257–68
Haber S 1997 Financial markets and industrial development: A comparative study of governmental regulation, financial innovation, and industrial structure in Brazil and Mexico, 1840–1930. In: Haber S (ed.) *How Latin America Fell Behind: Essays on the Economic Histories of Brazil and Mexico, 1800–1914*. Stanford University Press, Stanford, CA
Hadley E 1970 *Antitrust in Japan*. Princeton University Press, Princeton, NJ
Hall J R 1991 The patrimonial dynamic in colonial Brazil. In: Graham R (ed.) *Brazil and the World System*. University of Texas Press, Austin, TX
Harvey D 1982 *The Limits of Capital*. Blackwell, Oxford, UK
Hilferding R 1981 *Finance Capital: A Study of the Latest Phase of Capitalist Development*. Routledge & Kegan Paul, Boston
Kaplan D A 1999 *The Silicon Boys and Their Valley of Dreams*. William Morrow, New York
Kenney M, Florida R 2000 Venture capital in Silicon Valley: Fueling new firm formation. In: Kenney M (ed.) *Anatomy of Silicon Valley: Understanding an Entrepreneurial Region*. Stanford University Press, Stanford, CA
Lamoreaux N 1994 *Insider Lending: Banks, Personal Connections, and Economic Development in Industrial New England*. Cambridge University Press, New York
Leff N 1978 Industrial organization and entrepreneurship in developing countries: The economic groups. *Economic Development and Cultural Change* **26**(4): 661–75
Levine R 1997 Financial development and economic growth: Views and agenda. *Journal of Economic Literature* **35**: 688–726
Makler H 1999 Regional integration and trends in financial services. In: Haar J, Dandapani K (eds.) *Banking in North America: NAFTA and Beyond*. Pergamon, Oxford, UK
Makler H 2000 Bank transformation and privatization in Brazil: Financial federalism and some lessons about privatization. *Quarterly Review of Economics and Finance* **40**(1): 45–69
Mintz B, Schwartz M 1985 *The Power Structure of American Business*. University of Chicago Press, Chicago
Mizruchi M, Stearns L 1994 Money, banking, and financial markets. In: Smelser N, Swedberg R (eds.) *The Handbook of Economic Sociology*. Russell Sage Foundation and Princeton University Press, Princeton, NJ
Ness W L 1974 Financial markets innovation as a development strategy: Initial results from the Brazilian experience. *Economic Development and Cultural Change* **22**(3): 452–73
Ness W L 2000 Foreign banks in Brazil: What do they do differently? Unpublished essay, Business School, Pontifical Catholic University of Rio de Janeiro, October
Orru M, Biggart N W, Hamilton G 1991 Organizational isomorphism in East Asia. In: Powell W, DiMaggio P (eds.) *The New Institutionalism in Organizational Analysis*. University of Chicago Press, Chicago
Patrick H 1966 Financial development and economic growth in underdeveloped countries. *Economic Development and Cultural Change* **14**(2): 174–89
Roe M 1994 *Strong Managers, Weak Owners*. Princeton University Press, Princeton, NJ
Schumpeter J 1961 *Theory of Economic Development*. Oxford University Press, New York
Uzzi B 1999 Social relations and networks in the making of financial capital. *American Sociological Review* **64**(4): 481–505
Weber M 1968 Patriarchalism and patrimonialism. In: Roth G, Wittich C (eds.) *Economy and Society*. Bedminster Press, New York, Vol. 3, pp. 1031–77
World Bank 1989 *Financial Systems and Development. World Development Report*. Oxford University Press, New York
World Bank 2000 *Entering the 21st Century. World Development Report 1999/2000*. Oxford University Press, New York
Zeitlin M 1984 *The Civil Wars in Chile or the Bourgeois Revolutions that Never Were*. Princeton University Press, Princeton, NJ
Zysman J 1983 *Government, Markets and Growth: Financial Systems and the Politics of Industrial Change*. Cornell University Press, Ithaca, NY

H. M. Makler

Fine Arts

Whereas the arts and other forms of creation may be the results of qualities inherent to the human species, the fine arts constitute a particular subset of this broad domain. For the most part a specifically European phenomenon, a social construction embedded in his-

torical, socioeconomic, and political developments in the context of the late Middle Ages, they reached their apogee with the rise of the nation state. Sustained by the development of a discourse that singled out certain art forms as unique and superior to other forms of creation, their high standing was achieved through association with prestigious supporters, penetrating esteemed milieus, and the establishment of selective procedures that became institutionalized over several centuries. Having become a well-entrenched cultural structure, even with the decline of the conditions and institutions that had sustained them, they continue to exert influence over other genres, some of which came into existence only much later.

1. Introduction

Neither timeless nor abstract, like all categories of thought, the fine arts came into being over time through a process of construction and subsequent reconstructions (Zolberg 1990). The term as used today cannot be assumed to possess the same connotations as when it first appeared. Moreover, although functional equivalents of the fine arts have been found in several geopolitical regions and at different historical moments (Alsop 1982), the fine arts represent a complex of meanings which were until relatively recently associated with European civilization. From a social scientific standpoint the fine arts need to be understood in relation to the arts and cultural creation more generally. The first part of this article deals, therefore, with the standing of the fine arts as a subset of aesthetic creation in relation to other forms.

Three developments supported the emergence and subsequent durability of the cultural structure of the fine arts:

(a) The constitution of its intellectual roots as an integral part of European discourse within the changing socioeconomic and political context of the Middle Ages and the Renaissance.

(b) Somewhat independently, professionalization of artistic creation in which individual achievements came to be favored over the communalism of the previously pre-eminent institution of creation, the guild system.

(c) Intersecting these trends, in the place of the previously dominant forms of patronage by the church, the nobility, and fairly autonomous cities, the emergence of new support structures in whose constitution creative artists played an active part: first, official patronage through academies; next, the dealer–critic system.

In the final section, this article analyzes processes of cultural change that, moving beyond anti-academicism, itself a part of the modernist movement, takes the far more radical stance of postmodernism. Whereas modernism accepted the position of fine art as a hierarchically superior domain of creation, postmodernism contests the very idea of such a hierarchy. It disregards barriers between the fine arts and commercial art forms, professional artists and self-taught creators, art world members and outsiders to it. It rejects the separation of art from politics, science, nature, and even life itself. In trying to clarify the conditions under which certain forms, old and new, gain legitimacy as fine art, this article explores the possible consequences that expanding this construct to encompass existing fine art worlds, cultural politics, and commercial entertainments may have for the idea of fine art in the future.

2. The Making of the Fine Arts Core

2.1 Structuring Aesthetic Culture

Distinguished by a set of artistic categories that made up what has been authoritatively designated 'the modern system of the arts,' the fine arts owe their uniqueness to a particular cultural development. It encompasses the transformation of a body of ideas, largely derived from classical thought, thence reformulated in late medieval scholasticism and filtered through Renaissance conceptions. Culminating in the intellectual innovations of the Enlightenment, the fine arts became the established aesthetic structure with which all later arts and scholarship had to contend. Fully constituted by the eighteenth century, as Paul Kristeller noted in his classic formulation, it consisted of a core of five fine arts—painting, sculpture, architecture, music, along with the *belles lettres*–poetry and eloquence (Kristeller 1951, p. 497). While other forms migrated into and out of this cultural structure—dance and garden design, for example—these five became the core of what was meant by Art.

Kant, the first major philosopher to integrate aesthetics into his intellectual system, posited the enduring qualities of art, its universalism, the pleasure it gave, yet insisted that art should not provide too much ease of enjoyment. In this way Kant's analysis attributed to fine art a seriousness equal to his theory of truth (epistemology or metaphysics) and his theory of goodness (ethics) (Kristeller 1952, p. 42). As later critics have noted, an important criterion in Kant's formulation was to be art's gratuitousness, its freedom from social necessity and usefulness. A striking consequence of this construction is that it can be seen as a bulwark for ideologies justifying social inequality (Bourdieu 1984, pp. 494–5).

Although it served as a foundation for philosophical writing throughout the nineteenth century, the system of the arts has been challenged by, among other things, ideas associated with the Romantic movement and by

new technologies. As Kristeller noted, aesthetic analysis based on the modern system of the arts is hard pressed to be appropriate to new mediums and art forms such as the modern concert grand piano and motion pictures (Kristeller 1952, p. 46). Thus, knowing the intellectual roots of the concept is insufficient for understanding the social significance of the fine arts. Without reducing one to the other, the orientations of intellectual and art history with their focus on autonomous genres and their structure, and influences emanating from within their particular domain need to be situated within the social, political, and economic changes with which they intersect. With aesthetic analysis and history of art as the intellectual foundation, the social history of the arts has provided an important perspective by emphasizing the changing social context within which the arts developed and with which they interacted (Hauser 1951)

2.2 *Patronage of the Fine Arts: A Contribution of Social History*

In their studies of artistic transformations, social historians of the arts stress the effects of changing patterns of patronage: shifts in support structures from the patronage; of the classical city states, imperial courts, churches, secular cities, and states to individual aristocrats, royal courts, or magnates (Hauser 1951, Castelnuovo and Guinzburg 1981, Kempers 1992). They have observed that the social standing of art forms and artists is frequently commensurate with the status of their patronage source. Most social historians contend that patrons—those who 'pay the piper'—call the tune. Further careful examination suggests that the nature of their influence is highly complex. The idea that art is a mere 'reflection' of its support system has been shown to be oversimplified (Balfe 1993, p. 4). Instead, it is important to recognize the diversity of intermediate institutions of culture and the expansion of varieties of possible sources of patronage, especially since the middle of the nineteenth century. While social historians provide insight into the impact of major social structural changes on the arts, much social history neglects the agency of artists, individually or in groups, to act on their own behalf. Recent studies indicate that artists played an active role, for example, in establishing the Royal Academy of Fine Arts in France (Heinich 1993) and, in the process, constituting what for over a century was considered 'the canon.'

For painters who sought to gain the support of royalty and the nobility, the uncertainty of their social position appears to have been the main incentive underlying their strategies. With the establishment of the Academy in the seventeenth century (Heinich 1993), recognition by the highest status groups in society went a long way to providing that standing, at least for a lucky few. In conjunction with subsequent developments, the fine arts (*beaux-arts*) came to be treated as inhabiting a separate sphere with its own intellectual legitimacy and, under certain conditions, a degree of autonomy from mundane social constraints. This new autonomy is implied in Max Weber's observation that 'art becomes a cosmos of more and more consciously grasped independent values which exist in their own right. Art takes over the function of a this-worldly salvation.' Although Weber goes on to caution that 'the refusal of modern men to assume responsibility for moral judgments tends to transform judgments of moral intent into judgments of taste ('in poor taste' instead of 'reprehensible')' (Gerth and Mills 1946, pp. 941–2), he correctly notes that many Western artists since the late Middle Ages aspired to that autonomy. In that regard, the most durable supporting institution became the academic system (Pevsner 1940).

2.3 *The French Academy: Institutional Support for Official Creativity*

As an administrative structure encompassing educational institutions, competition, and rewards for talent, the French Academy became the principal gatekeeper in commissioning art for official public purposes. Following a rationale derived from the prestigious humanistic classical scholarship, with the purpose of symbolizing the glory of the monarch and (not very differentiated from him) the State, academicians selected aspirants, rigorously educated them in a certain kind of creativity, established and maintained standards of quality, and legitimated art forms. With the ruler's accord, this entailed institutionalizing certain art forms and genres based on subject matter, materials, and style, elevating some hierarchically over others. In the process they established the 'great culture' that redefined Art (Corvisier 1978, Heinich 1993). The French academic system was widely admired, envied, and imitated in many nations. With variations in accord with local conditions, traditions, and diverse contingent factors (Pelles 1963), it became the successful rival to the previous leading patronage structure, the Church, and overwhelmed the provincial nobility.

Academic dominance of art careers had both symbolic and material consequences. Sustained by an aura of the sacred derived from the King himself, enjoying the acclaim of scholars and amateurs of high standing, fine art was at times treated as a quasi-religion of high culture. But the very fact of attaining high standing contributed to the ambiguity of the position of fine art. Intricately connected to leading institutions of religion, politics, education, and finance, its sacrality was constantly threatened by contamination from the worldliness associated with those domains. The penalty for artists who too obviously were concerned with financial success rather

than high-minded disinterestedness might be condemnation to separate development as part of a provincial 'little culture' to a weakened, marginalized position as a client of the Church (Corvisier 1978), or—worse yet—to the mercies of market forces.

Artists whose works were refused in official *Salons* where they could be viewed by high-status private collectors or clients and public officials with authority to commission and purchase art for public buildings, cathedrals, or churches with elite parishioners, had little recourse but to sell their paintings from their own studios or, worse, at public markets and open air fairs, to an indiscriminate clientele of lesser standing and with considerably lower financial rewards.

At the height of their pre-eminence as gatekeepers, academies and the choices they made were confronted by challenges emanating from artists and other art world participants, including dealers, critics, scholars, and collectors, who were outsiders to the official art world. In the context of eighteenth and nineteenth century shifts in political power away from an aristocracy that had been allied with a narrow stratum of wealthy financiers, an expanded and heterogeneous middle class came to prominence. It provided a diverse market that was no longer coterminous with the taste promoted by the official milieu. Instead, its members seemed to share an elective affinity with the entrepreneurship of the new dealer and critic system. This combination created conditions favorable to aesthetic innovations that were to challenge the officially recognized fine arts even more profoundly (White and White 1965, Boime 1971, Zolberg 1983).

2.4 From the Academy to Anti-academic Movements

Whereas under the Old Régime, fine art was what the monarch preferred, through much of the nineteenth century it was largely economic elites whose tastes, whether conventional or audacious, came to dominate (Boime 1971). Its demise as a result of the Great Revolution had left a void in the aesthetics that it had imposed on French cultural life. Remade after its virtual abolition and modified several times, the academic system flourished for the better part of the nineteenth century. Its eventual re-establishment as a set of state institutions was welcomed by many aspiring artists, who found that career avenues reopened to them (Boime 1971). But its success was soon diluted, overwhelmed by increasing numbers of aspirants (White and White 1965).

Through administrative control of admissions, the assignment by their juries to better or worse locations for art works on its walls, the prestigious biennial *Salons* might become opportunities for artists to be rewarded with prizes, government or private commissions, and future sales or, on the contrary, be quite unnoticed. Alienated from the traditional or official artistic values of their time, convinced of its favoritism toward the most timid conformity, many artists became disenchanted with the academic institution that earlier artists had helped to launch. Highly innovative at its beginnings, the official academy, as is the case of many formal organizations, was accused of becoming increasingly rigid and oligarchic for excluding innovative ideas and creators. By the middle of the nineteenth century, academic administrators were being charged with behaving like crude bureaucrats, unreflectively enforcing old ideas, favoring the predictable and the safe over the new and audacious.

These accusations were sometimes exaggerated by critics and disgruntled aspirants. In fact, its officials were not resistant to change across the board. The hierarchical dominance of history painting, portraits of monarchs, nobles, and very wealthy officials and businessmen with which the century had begun was being overshadowed by the end of the nineteenth century with the elevation of what had previously been classed as lesser genres. They included landscape paintings and nominally allegorical or mythical themes, often in a style accentuating narrative as opposed to timelessness, sweetness rather than nobility (Bouguereau) (Boime 1971). But the ethos of tameness imposed on art students and members conflicted with the increasingly prevalent conception of the artist as Romantic genius with a singular vision (Bourdieu 1995). The academy became synonymous with the overbearing institutions that trammeled their imaginations.

In trying to account for the eventual demise of academic authority, Harrison and Cynthia White have argued that the very success of the institution made it the goal of too many aspiring artists. Under the onslaught of swelling numbers of applicants, the academy was obliged to introduce increasingly selective mechanisms (White and White 1965). From an ethnographic perspective, the analytic categories developed by Howard Becker provide further insight into the social roles of artists in complex institutional settings or 'Art Worlds.' For Becker, the world of art is made up of several different types of artists. Certain artists are 'certified'—as 'integrated professionals,' or they may be nonconformists—'mavericks' (Becker 1982). In the exceedingly centralized French system especially, mavericks were viewed and often presented themselves as outsiders and innovators who knew the rules and *deliberately* violated them.

As Becker suggests, most were trained in the academic institutional system that rejected them, or studied with academically trained artists. But many were willing to risk exclusion from its sales institution, the official *Salon*, in order to be true to their ideas. Although, as later research indicates, in fact, relatively few of them were very innovative or audacious. This was even the case of the most celebrated event of the Second Empire, the *Salon des Refusés*, when Impressionists' works were displayed, as well as various

later dealer-organized exhibitions. But even the more conventional ones formed alliances and exhibited with independent groups of real mavericks (Huston 1989).

One of the most important results of their refusal to conform is the seemingly unending succession of new style after style. Associated with groups of artists who provided moral support for their efforts, each new stylistic form became a contender for official and public recognition. Its leaders were: Courbet and Naturalism; Corot and Millet and other Barbizon painters who led the movement for revalorizing landscape painting; Manet and numerous painters influenced by him who made up the Impressionist group, rejecting the foundation of classical drawing they had acquired in favor of color experiments that provided a direct relationship to nature; Neoimpressionists such as Signac and Seurat who undertook quasiscientific experimentation with the optics of color in their *pointillisme*; Van Gogh and his deliberate distortions of Expressionism; Gauguin and his combination of gothic line and form with eroticized exoticism; Cézanne and his uniquely difficult-to-characterize works (Rewald 1988). Most of these 'independent' artists found themselves barred from a place in official exhibitions for most of the nineteenth century.

For artists of modest means, the dealer–critic system that expanded and flowered in the nineteenth century was essential for permitting them to survive honorably as painters. As fine artists, they relished the informality of contractual arrangements that made agreements with dealers rest on a gentlemanly handshake (Vollard 1978, Moulin 1987). Even artists who were independently wealthy, such as Manet and Degas, sought to build support among new middle class collectors through the growing dealer system. Unexpected as it may have been to their critics, eventually a number of their aesthetic innovations came to be more highly prized than the more conventional art of their contemporaries. As controversial and willing to risk failure as they were, however, for all their innovations, they continued to work in the basic mediums and genres of art, rather than create entirely new forms of art.

3. *How the Core Collapsed—On Artistic Change*

3.1 *The Tradition of the New*

From approximately the middle of the nineteenth through nearly half of the twentieth century it became customary to speak of artistic change as a revolutionary succession of art movements or *avant-gardes* (Poggioli 1971). Most of them challenged stylistic strictures governing artistic form, some challenged the social structures of art worlds, and certain of them used their art to attack political and social arrangements beyond art worlds. Through these independent artists of *avant-garde* or secession movements that challenged the academy's legitimacy, the academic fine arts fell into disrepute. Several of the new styles gained ascendancy, but even as anti-academic artists, they did not eradicate the idea of fine art itself, nor their aspiration to aesthetic autonomy.

The story of the apparently inexorable march of new styles encompassed elements that contributed to the 'tradition of the new,' as Harold Rosenberg termed it (1965). Its pattern of innovation, rejection, and acceptance has become familiar and conventional in the fine arts, and is acknowledged as part and parcel of the flow of modernism. In the post-World War II period, the United States produced what was then considered the most innovative artistic movement. Referred to variously as Abstract Expressionism, Action Painting, or simply the New York School, its members aspired to replace the figurative content and narrative of art, largely hegemonic at the time, by abstract representations of the artist's feelings and impulses. Viewed by many as incomprehensible or self indulgent, as it gained acceptance among a few collectors and a place in adventurous museums, in line with the century-long pattern of the challenge by the new, it too was faced by rivals: even without an institution such as the academic system, avant-gardism itself had become institutionalized. Soon, they too found themselves and the aesthetics they had constructed challenged (Crane 1987).

As unexpected as they were, the new genres and styles—Pop Art, Minimalism, Pattern, Photo-Realism—could claim as much legitimacy as had previous newcomers. There was little basis for holding them off except through assertions of bad faith on the part of artists, decline of standards, lack of discrimination. On the one hand, these resemble the reproaches that have been invoked for a century and a half. On the other, the Pop Artists had crossed the barrier separating naked commercialism from disinterested art. Their popularity assailed what might be called the dominant canon of abstractionist modernity, leading the way to what has been termed the postmodern condition.

3.2 *Postmodernism and Challenge from the Margins*

Despite their confrontational stance, few artists of anti-academic movements actually repudiated 'the modern system of the fine arts' until well into the twentieth century (Danto 1986). Instead, the idea of 'fine art' has until very recently retained its lofty status. What has changed is that it may be freely attached to new genres, many of them nonexistent at the height of academic dominance, and most unrecognizable to nineteenth century observers. Whereas modernism accepted the position of fine art as a hierarchically superior domain of creation, postmodernism contests the very idea of such a hierarchy. It disregards barriers between the fine arts and commercial art forms,

professional artists and self-taught creators, art world members and outsiders to it. It rejects the separation of art from politics, science, nature, entertainment, and even life itself.

Maverick professionals have attempted to overthrow the existing canon that has been the foundation of their training (Becker 1982). Excluded groups on the fringes of established art worlds are a constantly renewed source of aesthetic change, whether through challenging aesthetic ideas or simply to gain entry without changing prevailing aesthetic forms or hierarchies. Most prominent are women, racial or ethnic minorities, or disadvantaged people more generally, who attempt, either on their own behalf or through the mediation of influential gatekeepers, to penetrate to the center (Lang and Lang 1990).

But over the years the margins of the art world have come to defy easy categorization. As has become startlingly evident, the barriers between high and low art, art and politics, art and religious rite, art and emotional expression have been repeatedly blurred and breached. One outcome of these assaults on conventions, in which the periphery has come to be valorized beyond all predictions, is that it is virtually impossible to find an aesthetic center. These developments have had their critics, many of them simply conservative either politically or aesthetically, or sometimes both.

How certain artists and art forms gained ascendancy has often involved political and social contexts that are intertwined with aesthetics. They may involve the construction of institutional support favoring particular stylistic forms while excluding others. Under authoritarian political regimes in modern times this has led to fierce repression of artists and their works (Nazi Germany, Stalinist Soviet Union, Maoist China, among others). Even in liberal regimes, artists may be restricted in the free expression of their creation. Whereas in the past the most important sources of support came from absolutist monarchs who were attempting to centralize their nation states at the expense of competing institutions and what they constructed as peripheral cultures, today the ground is shifting. This is indicated by growing interest in the construction of social identities that are not as co-extensive with a nation's boundaries as was the ideal expectation of the past. The survival or revival of cultural conflict around language, ethnicity, and religion provide a fertile ground for jockeying for pre-eminence by winners at the expense of dominated groups.

A familiar cliché has become that of imagining the artist facing a philistine world that misunderstands him, struggling for inspiration, and gaining recognition only belatedly. Applied to nineteenth-century composers such as Beethoven and Schubert; painters such as Gauguin and Van Gogh, in the twentieth century this model was moved to the lives of Jazz 'greats'—Scott Joplin, Billy Holliday, Thelonius Monk, and many others. The insertion of Jazz and other popular music into the domain of the fine arts is far removed from academies and official legitimacy. But by now almost the entire body of popular commercialized culture and the constitution of stardom could be included in this analysis of forms that are outside of fine art canons.

Supposedly born as low-life brothel music, Jazz grew in the late 1910s and 1920s into a wickedly exciting, commercially powerful form of popular music. Taken up as an art form by European intellectuals, who incorporated it into ideas of an essentialist primitivism that were popular among artists and writers, and appealed to many amateurs, their attention provided the cachet that ordinary commercial musical forms lacked. Eventually, certain American intellectuals were drawn into defining Jazz as an expression of revolt against the conventional lives they were expected to live. By the 1940s, with the advent of be-bop and the regular forays of composers such as Duke Ellington into concert halls, Jazz began to take on the trappings and pretentions of Art, moves supported by patrons, official government bodies, foundations, while commercial entrepreneurs became less important. Its legitimacy functioned at two levels, one aesthetic and the other political. Rather than being content with the entry of the new musical forms into common usage through appropriation by industry agents, even at increased financial returns, African Americans became active in asserting the importance of their identity through its innovations.

The rise of Jazz to respectability and fine art standing is based, by now, on a long history reconstructed (or constructed). It includes a cataloged repertory and a body of admirers and practitioners. It has a closely perceived linkage with black cultural identity and hence has the support of related interest groups who advocate funding for minority cultural and political projects. Jazz (or 'black classical music' as some call it) is firmly embedded in the curricula of institutions of higher education and conservatories. It is deemed worthy of support by governmental agencies such as the National Endowment for the Arts and State Arts Councils. By now, Jazz has attained a level of respectability that would have been difficult to predict less than a century ago.

The very nature of art in the late twentieth century entails a constant process of absorbing, sampling, appropriating, amalgamating new and unusual forms, and incorporating a changing body of personnel. Currently, the field is much larger and the number of contenders has multiplied, as has the core itself. The line between classical and commercial entertainment has been crossed repeatedly. Beyond professionals and mavericks, creative work by folk and naïve artists has penetrated the museums and galleries that had been host to creators formed in the attachment to autonomous fine art of whatever form. No longer a single canon, the centralized academy and its in-

stitutional descendants have given way to competing groups who promote a broad range of genres or forms (Huyssen 1986, p. 218). It may be that, as Andreas Huyssen has suggested, the previous domination of the fringes by the West may be replaced by a healthy resistance of the dominated in the form of a productive tension between the political and the aesthetic (Huyssen 1986, p. 221).

See also: Avant-garde Art and Artists; Collective Identity and Expressive Forms; Culture, Production of; Nationalism and Expressive Forms; Popular Culture

Bibliography

Alsop J 1982 *The Rare Art Traditions: The History of Collecting and Its Linked Phenomena Wherever They Have Appeared*. Harper and Row, New York

Balfe J H (ed.) 1993 *Paying the Piper: Causes and Consequences of Art Patronage*. University of Illinois Press, Urbana, IL and Chicago

Becker H S 1982 *Art Worlds*. University of California Press, Berkeley, CA

Boime A 1971 *The Academy and French Painting in the 19th Century*. Phaidon Press, London

Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste. Trans. by Richard Nice*. Harvard University Press, Cambridge, MA

Bourdieu P 1995 *The Rules of Art: Genesis and Structure of the Literary Field* [trans. Emanuel S]. Stanford University Press, Stanford, CA [Orig. Éditions du Seuil 1992]

Castelnuovo E, Guinzburg C 1981 Domination symbolique et géographie artistique dans l'histoire de l'art italien. *Actes de la Recherche en Sciences Sociales* **40**(November): 51–72

Corvisier A 1978 *Arts et sociétés dans l'Europe du XVIII° Siècle*. Presses Universitaires de France, Paris

Crane D 1987 *The Transformation of the Avant-Garde: The New York Art World 1940–1985*. University of Chicago Press, Chicago

Danto A 1986 *The Philosophical Disenfranchisement of Art*. Columbia University Press, New York

Gerth H, Mills C W 1946 *From Max Weber*. Oxford University Press, New York

Graña C 1964 *Bohemian versus Bourgeois: French Society and the French Man of Letters in the 19th Century*. Basic Books, New York

Hauser A 1951 *The Social History of Art*. Knopf, New York, 4 Vols.

Heinich N 1993 *Du peintre à l'artiste: Artisans et Académiciens à l'âge classique*. Éditions de Minuit, Paris

Huston L 1989 *L'Evolution des expositions d'art à Paris, et le pouvoir dans la vie artistique (1864–1914)*. Ph.D. thesis. Concordia University, Montréal, Canada

Huyssen A 1986 *After the Great Divide: Modernism, Mass Culture, Postmodernism*. Indiana University Press, Bloomington, IN

Kempers B 1992 *Painting, Power, and Patronage: The Rise of the Professional Artist in the Italian Renaissance* [Trans. from Dutch Jackson B]. Allen Lane/Penguin Press, London

Kristeller P O 1951 The modern system of the arts. *Journal of the History of Ideas* **12**: 496–527

Kristeller P O 1952 The modern system of the arts II. *Journal of the History of Ideas* **13**: 17–46

Lang G E, Lang K 1990 *Etched in Memory: The Building and Survival of Artistic Reputation*. University of North Carolina Press, Chapel Hill, NC

Moulin R 1987 *The French Art Market, A Sociological View* [Trans. Goldhammer A]. Rutgers University Press, New Brunswick, NJ [abridged from orig. Éditions de Minuit, 1967]

Pelles G 1963 *Art, Artists and Society: Origins of a Modern Dilemma, Painting in England and France, 1750–1850*. Prentice-Hall, Englewood Cliffs, NJ

Pevsner N 1940 *Academies of Art: Past and Present*. Cambridge University Press, Cambridge, UK

Poggioli R 1971 *The Theory of the Avant-Garde*. Harvard University Press, Cambridge, MA

Rewald J 1988 *Cézanne and America: Dealers, Collectors, Artists and Critics, 1891–1921*. Princeton University Press, Princeton, NJ

Rosenberg H 1965 *The Tradition of the New*. McGraw-Hill, New York

Vollard A 1978 *Recollections of a Picture Dealer* [Trans. MacDonald V M]. Hacker Art Books, New York [Enlarged from orig. Albin Michel, 1937]

White H C, White C 1965 *Canvases and Careers, Institutional Change in the French Painting World*. Wiley, New York

Zolberg V L 1983 Changing patterns of patronage in the arts. In: Kamerman J B, Martorella R (eds.) *Performers and Performances: The Sociology of Artistic Work*. Praeger, New York

Zolberg V L 1990 *Constructing a Sociology of the Arts*. Cambridge University Press, Cambridge, UK

V. L. Zolberg

Fire, Human Use, and Consequences

1. Introduction

The ability to handle fire is, like language and the use of tools, a universal human attainment: no human society in historical times is known to have lived without it. The ability to handle fire is also exclusively human; while other animals have developed rudimentary signals and tools, only humans have learned to handle fire.

Today fire is continuously present in all human societies. It is used in many different guises, some highly visible, others largely hidden from public view and consciousness. Its control has played a far more important part in human history and social evolution than is generally acknowledged in the literature of the social sciences.

2. The Original Domestication of Fire

Fire is, according to dictionary definitions, a process of combustion, manifested by heat and light. It occurs relatively rarely in nature; when it does, it does not last long, with a few exceptions such as the burning coalfields in Northern China which have been alight for

many centuries. Most natural fires have rapid and drastic effects. They cause a quick disintegration of organic matter, reducing it to ash and smoke in a process of transformation that is almost the prototype of 'destruction.'

During the Paleolithic, human groups have learned to turn this destructive and essentially purposeless natural force into a regularly available asset that could be used for a variety of 'productive' purposes. They 'domesticated' fire, they 'tamed' it and 'incorporated' it into their own societies. They were able to do so thanks to their general capacity for culture—for learning new forms of behavior, sharing those with others, and transmitting them to following generations. Adjusting their behavior in such a way that a fire could be kept burning continuously, they developed a fire regime which enabled them to expand their control over their natural environment, while at the same time affecting their own social arrangements and their personal habits and feelings (cf. Goudsblom 1992, pp. 12–41).

The uses of fire were diverse, ranging from the enjoyment of heat and light to clearing land for hunting and gathering and to cooking. Their overall tendency was to make life in human groups more secure and comfortable, but also more vulnerable and dangerous. *Fire increased at once people's destructive and productive potential.* Its control thus reinforced a basic trend in human history and social evolution: the increasing differentiation in behavior and power between humans and all biologically related species.

3. *Fire and Agrarianization*

The formation of the fire regime may be regarded as a paradigm for other regimes that were developed at later stages in human history. A paradigm in a double sense: first of all, substantively, the regime by which humans learned to extend their care and control over fire could serve as a model for subsequent forms of care and control over other forces in nonhuman nature such as plants and animals. Second, the domestication of fire may be seen as a model case in a methodological sense, since it exhibits some basic principles which may be observed again and again in human history and social evolution.

Toward the end of the Paleolithic, between 15,000 and 10,000 years ago, some human groups began to extend their care and control over selected plants and animals. This marked the beginning of what is often called 'the First Revolution' but is better viewed as the second great ecological transformation brought about by humans. That transformation could not have taken place without the firmly established control of fire. Once it was under way, it in turn deeply affected the further development of the part played by fire in human society.

The domestication of fire was a precondition for agrarianization first of all because in most areas using fire ('slash and burn') was the most efficient technique of opening up fields for cultivation by removing the existing vegetation. Moreover, most of the crops selected for cultivation needed cooking before consumption as a staple food. As a further, less tangible precondition, the solidly established 'species monopoly' over fire gave humans a dominance over other animals which enabled them not only to bring certain species under direct control but also kept most of the remaining 'wild' animals at a distance from their crops and herds. Finally, the long familiarity with a fire regime probably helped to prepare people for the strains of an agrarian regime demanding self-imposed 'deferred gratification.'

With agriculture and livestock, fire was no longer the only nonhuman source of energy under human control. Gradually it ceased to be the prominent focus of group life that it had been for many thousands of generations, becoming increasingly more dispersed over various sorts of specific 'containers,' such as hearths, ovens, and lamps, beyond which it was not allowed to spread. Its use was subjected to stricter regulation. Whereas during the initial stages of domestication, the main concern had been to keep the one fire burning, the main concern now became to prevent the many fires from causing a conflagration.

As the more advanced agrarian societies grew more populous, the number of fires in use increased proportionally. Those fires were increasingly concentrated within the 'fire-protected zones' of towns and villages, in special containers. The uses of fire became more specialized, with various crafts and professions such as smiths and potters developing their own pyrotechnic skills. With proliferation, concentration, and specialization came new forms of organization, if only to meet the growing need for fuel. Unavoidably, the use of fire also reflected the process of stratification, as some privileged people came to command huge supplies of fuel, and could use fire for impressive displays of power, whereas many others might never call any fire their own.

A common problem in all advanced agrarian societies was the prevention of uncontrolled fire in cities. For the growing urban populations the proliferation of fires increased the risks of blazes, while the accumulation of property meant that people had more to lose.

Of course, it was fire as a natural force with its aimless and self-generating destructiveness that was feared. In case of a military raid people had to reckon with the common practice of 'sack and burn.' But ordinarily they encountered fire only in its domesticated settings where the greatest danger lay in carelessness and where they had to rely on their own and their fellow citizens' incessant caution.

Once a fire did break out in a preindustrial town, there was little one could do. Usually the only way of

stopping it was by creating a 'fire break' and tearing down the buildings towards which the fire was advancing; deliberate destruction by human hands had to put an end to the blind destruction by fire. By far the most effective way of restricting the risk of conflagrations was then, as it still is, prevention.

There is a tendency today to conceive of problems of fire prevention primarily in terms of technology: of building materials and technical equipment. However, the problems have always been at least as much civilizational. The crucial issue was human behavior in its social context of informal and formal controls.

Little is known about the organization of fire brigades in preindustrial cities. The available evidence suggests that their effectiveness, and their very presence or absence, reflected important aspects of the social structure. In ancient Rome, fire brigades at first operated as private bands, and were later brought under imperial military command. A glimpse of the situation in the Roman provinces is shown in a letter from the emperor Trajan, written in 112 CE, in which he refused permission to the governor of Nicomedia to establish a fire brigade of citizens on the grounds that such an organization could easily 'turn into a political club' (Goudsblom 1992, p. 118). One of the social innovations in the towns of medieval Europe was that they had civil fire brigades.

4. *Fire in Industrial Society*

After the original domestication of fire and the rise of agriculture and animal husbandry, industrialization was the third major ecological transformation set in motion by humans. Again, natural forces originally outside the human realm were brought under human control. These natural forces consisted primarily of large, untapped sources of energy stored up in fossil fuels. Tapping the sources of coal, and later also of oil and gas, enabled people to exploit a great variety of other mineral resources as well, from iron to plutonium.

From the start, industrialization involved an enormous increase in the number, variety, and intensity of man-made fires. Just as the first stages of agrarianization were marked by extensive and conspicuous use of fire in the clearing of land by slash and burn, early industrialization dotted the landscape with smoke stacks and steam locomotives. In agriculture, the extensive use of fire gradually made way to more intensive forms of tillage with less conspicuous use of fire. Similarly, fire and smoke have become far less prominent in the industrial scene, as production and transport were powered increasingly by electricity and other forms of technologically processed energy (cf. Hughes 1983).

Yet modern industry as well as agriculture continue to be highly fuel-intensive. Most of the energy used is still derived from fossil fuels. Combustion processes continue to play a central role, but they are relegated to special containers so that most people are not confronted with any of the features that can make the presence of a fire annoying and dangerous. The caprices of the flames are fully controlled. Soot, smoke, and risks of burning are reduced to a minimum. The furnaces and combustion chambers in which great heat is concentrated remain cool on the outside.

Typical products of modern fuel-intensive industry are cars, with engines designed to be propelled by finely tuned and minutely controllable combustion processes. Cars are made of steel, plastic, and glass—materials produced and processed at high temperatures. Yet when people get into their vehicle and turn on the electrical ignition to start the engine they are unlikely to be consciously aware of using fire and the products of fire. Nor are they used to stopping and thinking of the unintended contribution they are making to CO_2 emissions and, probably, to global warming (cf. Fleming 1998).

The highly fuel-intensive character of modern technology is matched by the advanced standards of fire prevention. Every branch of industrial production, from the building trade to the manufacture of clothes and furniture, is subject to strict measures with respect to fire. In such branches as the aircraft or the hotel industry, regulation has become international.

Along with productivity the social potential to destroy, or 'destructivity,' has increased greatly. Not only have the means of preventing and fighting fires become more effective, but so have the means of starting fires, while cities as targets for arson have also increased in number and size. Consequently, the hostilities of World War II have caused some of the largest urban fires in history, as in Hamburg, Dresden, Tokyo, Hiroshima, and Nagasaki. Forest fires caused directly or indirectly by human action have also been more numerous and larger than in any previous century (cf. Pyne 1995).

Forms of fire use that were common and relatively inconsequential before the industrial age have become far more consequential under today's socioecological regimes. Thus, in many parts of Asia and Africa the need for firewood is now significantly contributing to deforestation. The high concentrations of wood stoves and ovens in the crowded metropoles in China and elsewhere have detrimental effects on human health.

5. *Symbolic and Religious Aspects*

From all over the world myths have been collected relating how humanity first came to master fire, often in a feat of cunning and deceit (Frazer 1930). Those myths convey a sense of the preciousness of fire which is represented as a unique possession, separating humanity from the animal world. Many people told that neighboring groups living further away in the

bush did not have fire and were unable to cook their food; those stories have all been proven slanderous.

Fire is generally believed to have purifying effects. The domain of culture is seen as coinciding with the range of domesticated fire. Thus, cooked meat is regarded as edible for humans, raw meat as only fit for beasts. Similarly, land cleared by burning is trusted as belonging to the human domain, whereas uncleared bush and forest are feared as a dangerous hideout for predators, serpents, insects, and evil spirits.

The sense of preciousness has also found expression in rites. The communal fire would be the center of group life, and people would treat it as sacred, surrounding it with ceremony and taking care that it would never cease burning. The Roman cults of the hearth and of Vesta, the fire goddess, show how such ancient rites could persist in an urban setting (cf. Pyne 1997).

Yet, with urbanization, the sense of the preciousness of fire was often superseded by a general anxiety of the dangers it posed. Thus while Hindu and Zoroaster fire cults kept alive the age-old worship of fire, they were surrounded by ceremonial care. In Christianity and Islam, some old fire traditions persist, as in the burning of candles and incense. Fire no longer plays any part in beliefs about the origin of human society, however. Instead, it is strongly associated with evil. Sinners and unbelievers are threatened with a life after death in which they will be tormented by eternal burning.

Fire festivals, such as those held all over Europe on Saint John's day, when things considered impure would be committed to the flames, were occasions of combined joy and horror. Modern literature and art abound with evocations of the terrors of fire. Most novelists who have described a blaze attribute the disaster, quite realistically, not to natural forces such as lightning but to human action—either to personal negligence, madness, or malice, or to collective aggression, in war.

A more dispassionate way of regarding fire, not in terms of its good or bad effects upon human beings but as a natural force in its own right, was first developed by the philosophers of ancient Greece. Out of various antecedents a generally accepted doctrine emerged according to which the universe consisted of four elements: earth, water, air, and fire. In China and India similar systems of thought developed. With some variations the view in which fire constituted one of the major elements remained the dominant cosmology for natural scientists and physicians well into the modern era.

Alchemists and chemists, as the direct successors to the metallurgists and 'fire masters' of a previous age, made the study of fire into their central concern. Just as modern mechanical industry in its initial stages owed a great deal to the steam engine, the rise of modern physical science would have been inconceivable without the laborious efforts of earlier investigators experimenting with, and theorizing about, fire.

6. *Conclusions*

The control over fire has been a human species monopoly since the Paleolithic. It constitutes an integral part of the apparatus with which humans have established and continue to maintain their dominance over other species.

Ever since the earliest stages of its domestication, the same sequence of events has recurred again and again in human relations with fire: deliberately sought advances in control are followed by unintended increases in dependency, and as the dependency sinks deeper into the social and economic infrastructure, it tends to become increasingly less clearly perceptible. Like agrarianization, industrialization began with conspicuously huge applications of fire. As industrial production became more specialized and highly organized, so did the use of fire. In everyday life in highly industrialized contemporary societies, flames are visibly present only in such highly domesticated guises as cigarette lighters, candles for ceremonial use, or wood fires intended to create a sphere of comfort and relaxation. When, on the other hand, a fire is shown on television or in the newspapers, it almost always spells war and disaster. The regularly controlled combustion processes upon which industrial production largely rests are now mostly relegated away from public view.

In accordance with this apparent eclipse of fire, it has also ceased to be regarded in the natural sciences and philosophy as one of the elements out of which the world is composed. The very concept of fire has disappeared from scientific discourse, to be replaced by the more abstract concept of 'energy'—something that cannot be directly seen, heard, smelled, or felt. Similarly, many a day may pass in the life of the average citizen of a modern industrial state during which he or she does not witness any physically observable fire. In the social sciences, too, fire has become almost completely unnoticed. It can be concluded from the foregoing that this is unwarranted.

See also: Agricultural Sciences and Technology; Climate, History of; Environmental and Resource Management; Environmental Challenges in Organizations; Food Production, Origins of; History of Technology; Human–Environment Relationships; Land Use and Cover Change

Bibliography

Bachelard G 1938 *La psychanalyse du feu*. Librairie Gallimard, Paris [1964 *The Psychoanalysis of Fire*. Beacon Press, Boston]

Canter D (ed.) 1980 *Fires and Human Behaviour*. Wiley, New York

Fleming J R 1998 *Historical Perspectives on Climate Change*. Oxford University Press, Oxford, UK

Frazer J G 1930 *Myths of the Origin of Fire*. Macmillan, London

Goudsblom J 1992 *Fire and Civilization*. Penguin, London
Hughes T P 1983 *Networks of Power. Electrification in Western Society, 1880–1930*. Johns Hopkins University Press, Baltimore, MD
Jones E L, Porter S, Turner M 1984 A Gazetteer of English urban fire disasters. *Historical Geography Research Series 13*. Geo Books, Norwich, UK
Pyne S J 1982 *Fire in America. A Cultural History of Wildland and Rural Fire*, 1st edn. Princeton University Press, Princeton, NJ
Pyne S J 1995 *World Fire. The Culture of Fire on Earth*. Henry Holt and Company, New York
Pyne S J 1997 *Vestal Fire. An Environmental History, Told through Fire, of Europe and Europe's Encounter with the World*. University of Washington Press, Seattle, WA

J. Goudsblom

Firm Behavior

At the heart of the supply side in the theory of how markets function is the theory of the firm. This article surveys the received economic wisdom on how firms make their decisions as to the types and quantities of products they will supply, the prices at which they sell, and the investments they make toward future operations.

That economists have adopted the convention of talking and writing about the behavior of 'the firm' is semantically curious. Supply decisions are made by entities bearing a host of names—corporations, companies, enterprises, businesses, ventures, farms, factories, establishments, and, of course, firms. The word 'firm' is derived from the Latin verb *firmare*, which evolved over time to describe the confirmation of a transaction by signature. The notion of a 'firm' metamorphosed during the nineteenth century to define houses of business organized through written agreements, notably, partnerships, and in modern lay parlance, it is used most frequently to describe partnerships in such fields as law and accounting.

1. Complex Organizational Forms

Meanwhile, the forms of business organization were changing in legal structure and scope. Corporations replaced partnerships, and multi-unit business enterprises operating both within a nation and multinationally came to supply the lion's share of services and (especially) products. Already in the sixteenth century, the Venetian merchant Antonio was said by Shakespeare to have 'an argosy bound to Tripolis, another to the Indies … a third at Mexico, a fourth for England, and other ventures he hath squandered abroad.' In 1995, US business corporations reported sales of roughly US$14 trillion, partnerships US$0.85 trillion, and non-farm individual proprietorships US$0.81 trillion. In 1982, 7,030 companies with 500 or more employees surveyed by the US Bureau of the Census (1986) generated 41.3 percent of all business employment. On average, those relatively large enterprises operated 64 separate domestic business establishments or plants each, not including foreign branch units. Thus, much of the industrialized world's business is done by organizations of considerable geographic scope and complexity.

Within this array of organizational forms, the theory of the firm is ambiguous, or one might say flexible, in isolating the locus of supply decision-making. The decisions encompassed under the theory may be made at the individual branch or factory level, at the headquarters level, at some intermediate management level, or across a composite of levels.

Few 'firms,' however they may be bounded, carry out internally all the activities needed to provide the goods or services they ultimately supply. Many intermediate inputs, supplies, machines, and specialized production and marketing services are procured through arms-length market transactions from other entities. The characteristic defining the bounds of a 'firm,' Ronald Coase argued in a seminal article (1937), is 'suppression of the price mechanism.' According to Coase, activities are collected within what we call a firm when transaction costs incurred in using the price mechanism for external purchases or sales exceed the cost of organizing those activities through direct managerial controls. Transaction costs may be high when work specifications are complex or rapidly changing, contract negotiations and renegotiations are burdensome, and /or there is a substantial risk of opportunistic deviations by one or another party once contracts have been fixed. The higher the transaction costs, the more a firm is likely to be vertically integrated to carry out internally the activities needed to implement its ultimate supply objectives.

2. The Goals of the Firm

If an organization is to act purposefully, it must have more or less clearly defined goals. In the standard static theory of the firm, the goal is simple: the firm maximizes profits. On the output side, this entails expanding production until marginal cost rises into equality with marginal revenue (which, for a purely competitive firm, is the product's parametric price). On the input side, profits are maximized by hiring the diverse inputs until the value of their marginal product, or in monopoly cases their marginal revenue product, is equal (when inputs are competitively supplied) to the inputs' prices. In addition, the firm must operate on its production function rather than off it; that is, it must avoid the manifold risks of X-

inefficiency (see, for example, Leibenstein 1966). This will be examined further later in this document.

Matters become more complicated when time and uncertainty enter the picture. On the time dimension, two general problems can be distinguished. For one, firms are faced with a continuing menu of opportunities to invest in longlived production facilities and new products. The standard theory says that they should accept all mutually independent investment projects whose internal rate of return exceeds the firm's marginal cost of investible funds. (Interdependence or mutual exclusivity of projects requires the satisfaction of additional conditions.) Additionally, when a firm has monopoly power and hence a choice of alternative output price levels, setting a price that maximizes current profits encourages the entry and expansion of competitors who erode future profit-earning opportunities. The firm must then make a tradeoff in its pricing between present and future profits. In the most elegant control-theoretic formulation of this problem (Gaskins 1971), the firm holds its price below the short-run profit-maximizing level so as to slow the rate of competitors' expansion and hence the erosion of future profits, appropriately discounted to present value, by the optimal amount.

One crucial variable in these calculations is the marginal cost of investible funds, also known as the firm's discount rate. Standard treatises on corporate finance suggest that the firm's discount rate be measured as the weighted average cost, taking into account income tax implications, of the various external sources of funds, i.e., common stock, long-term debt, and short-term loans; and the extent to which external financing, as compared to the reinvestment of current earnings rather than paying dividends, is utilized. Two complications arise here. First, the optimal mix or portfolio of financing methods must be chosen. There is disagreement in the literature over the extent to which external finance costs should and do vary with those portfolio choices. Second, debt costs vary over time with changes in macroeconomic money market conditions; and common stock financing costs fluctuate with both general macroeconomic conditions and with investors' volatile 'animal spirits,' to use Keynes' (1936, p. 161) metaphor. There is disagreement among scholars over whether firms should accept current market signals unequivocally in setting their investment discount rates or whether, recognizing that investments are for the long term, they should try to see beyond current market trends and 'smooth' anticipated external finance market costs in computing discount rates. If they fail to heed short-run equity market trends, they may be subject to hostile takeover by another enterprise more closely attuned to maximizing current shareholder value. Compare Marris (1963), Jensen (1988), and Scherer (1988).

A related question is how a firm's decision makers should treat the intrinsic uncertainty of investment payoffs. In principle, external financial investors evaluate the overall risks of a firm's investments and, by their reactions, introduce appropriate risk premiums into the firm's cost of capital. Using these, internal decision makers should presumably choose investments so as to maximize the discounted expected value of their investments. Often, however, risk aversion creeps into investment evaluations when decision makers or their staff adjust discount rates upward, shade estimated payoffs downward, or avoid through sensitivity analyses projects with substantial downside risk. Again, the firm that internally introduces more risk aversion than external investors prefer thereby increases its vulnerability to a hostile takeover. A further perplexity has been largely ignored in the relevant literature. Under standard practice, firms array their prospective investment projects in descending order of internal rates of return (IRRs), accepting all projects with IRRs in excess of the cost of external capital. This descending array of projects is the microeconomic analogue of what Keynes (1936, pp. 135–6) called 'the schedule of the marginal efficiency of capital.' It is illustrated by the line ME_1ME_1 in Fig. 1.

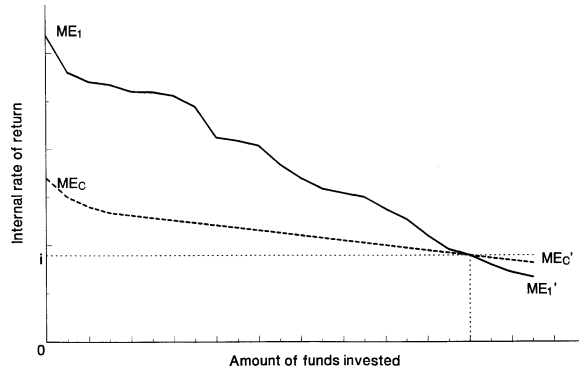

Figure 1
Marginal efficiency of investment schedule for the firm

Assuming this procedure to be unbiased, all projects but the marginal one (at point Z in Fig. 1) yield a return in excess of the rate Oi required to attract outside capital; that is, they yield a supranormal return. Received theory states that in long-run equilibrium, competitive firms do not realize supranormal returns, although monopoly firms may do so. But if the representative competitive firm faces a schedule of investment opportunities such as ME_1ME_1, how can it not enjoy supranormal returns, contradicting the predictions of the theory? There are two possible answers.

(a) Competitive firms, unlike monopoly firms, may systematically overestimate the profitability of their investments because they fail to take into account the likelihood that product prices and hence returns on investment will be reduced when many other firms,

each an atom relative to the market, make similar investments. This answer can be defended theoretically, but seems inconsistent with the premise that investors try to factor into their decisions all the expected environmental changes that could affect the returns on their investments. Compare Richardson (1959).

(b) The difference between competitive and monopoly firms may lie as much in the shapes of their marginal efficiency functions as in the expected output and price behavior of rival firms. In particular, the monopoly firm's MEI function may resemble ME_1ME_1 in Fig. 1, whereas the typical competitive firm, faced with competition in exploiting investment opportunities as well as in pricing its output, confronts a function more like the dash–dash schedule ME_cME_c, with much less (and in the limit no) possibility of rationally expecting supranormal returns.

3. *Getting Things Done*

Even when it is agreed that business firms should maximize either their current profits or the discounted present value of their long-term profits, there is many a slip between cup and lip. Difficulties intrude because complex and difficult decisions must be taken in the midst of what von Clausewitz called the fog of war and because decisions must be implemented within complex organizational frameworks. Organizational complexity in turn engenders communication breakdowns, coordination failures, goal conflicts, and motivational problems.

3.1 *Bounded Rationality*

Running anything but the most pedestrian business enterprise is difficult. Firms' decision makers seldom have a clear view of how consumers will respond to price changes outside the range of those with which experience has been accumulated; that is, demand curves are imperfectly visualized. In oligopolistic markets, the reactions of rivals are crucial to profitability but difficult to predict. Whether technological innovations will function as hoped is uncertain. Even less certain is how consumers will respond to product design changes, minor or radical. Decisions that might turn out brilliantly if rival firms remain passive could fail if rivals respond with aggressive countervailing strategies.

In a debate that raged during the mid twentieth century, it was asserted that business decision makers were not only subjected to great uncertainty about the environment in which they operated, but didn't understand the basic logic of profit maximization as taught by economists. In a famous reply, Machlup (1946) insisted that the variables used in firm decision making are only 'subjective estimates, guesses, and hunches' and that, just as automobile drivers can solve the problem of passing another car on a two-lane, two-way road without comprehending the differential equations of acceleration and convergence, so also business decision makers (or at least, the successful ones) intuitively solve the equations of profit maximization. Since that time, hundreds of thousands of young people trained in graduate business administration programs have entered the real world of business. It is surely true, even if it once was not, that business decision makers at least understand what marginal analysis means.

The more serious problem is undoubtedly what Herbert Simon (1959) called 'bounded rationality'—that is, the impossibility of having all the information about present and future conditions needed to formulate and solve decision problems with the rigor economic theorists might desire. Seeking to explain how business decision makers function under these conditions, scholars have proposed and extended a 'behavioral theory of the firm' that recognizes real-world information availability and processing limitations (see, for example, Cyert and March 1963). Rather than assuming profit maximization as a well-focused objective, the behavioral theory of the firm postulates that decision makers adopt more limited proximate goals—e.g., setting plausible short-term sales and profit targets—and then 'satisfice' in striving to attain them. The targets may then be ratcheted upward over time to converge, at least under favorable conditions, toward a true profit maximum. Similarly, decision-making processes are simplified to emphasize 'rules of thumb' and 'routines' whose pursuit leads in the general direction of profit maximization and whose modification, under the stress of adverse outcomes or the recognition that environmental variables have changed, tends in a rough way to emulate the outcomes that would emerge from a perfectly rational but informationally unattainable profit maximization exercise (see, for example, Nelson and Winter 1982).

3.2 *Decentralization in Complex Organizations*

Some suppliers of goods and services such as traditional family farms and 'Mom and Pop' retail outlets have only primitive organizational structures. However, much of the world's business is done by geographically dispersed multi-unit, multi-product enterprises. For them, complex organizational problems cannot be avoided. Hardly anything could be achieved by a large organization, whether it be a business enterprise, a church, or an army, without resort to decentralization. In the simplest models, goals and means are chosen by the organization's head, called the 'principal,' and the work is carried out by lower-level employees, called the 'agents' in the large body of

literature on principal–agent relationships. In more realistic multi-unit enterprises, strategic decisions on the markets to be served, the units responsible for serving them, and the financial and physical means at those units' disposal are set by top management; more detailed decisions on product specifications, prices, outputs, and marketing methods are made by unit managers; and the work is done by individuals within the various subordinate organizational layers. The paradigm for such decentralization, surveyed historically by Chandler (1962) and said by Williamson (1970, p. 175) to be 'American capitalism's most important single innovation in the twentieth century' is the so-called multidivisional or M-form organization. Under it, corporate leaders supported by a small central staff provide broad strategic direction and assign goals to virtually autonomous operating division heads, whose performance in meeting their goals is monitored by the corporate staff, with rewards and punishments being dealt out by the corporate leadership in proportion to the divisions' goal attainment. Or in still more complex variants, the corporate office sets overall policy and strategy; group executives are responsible for achieving broad product line or geographic region goals; and individual product division heads must operationalize the efforts needed to meet the goals (or bear the consequences if they fail). Although no one-to-one correspondence is possible, it is the individual product division within complex corporate organizations that is most like the 'firm' as it has been traditionally construed in microeconomic theory.

3.3 *Information Flow Occlusions*

With decentralization come several problems. One is the decentralization of knowledge. Central office executives usually know more than individuals lower in the corporate hierarchy about overall corporate goals (entailing *inter alia* understanding what stockholders want) and sources of finance. Production foremen have the most detailed knowledge of what can and cannot be done in cranking out products. Field sales representatives often know best what kinds of products their customers want and what rival firms are doing. R&D staff members have expertise in determining what product features can be achieved physically—insights that are sometimes sharpened by interaction with customers, but that for more radical product changes lie outside the range of customers' perception. Division managers serve as information collection points and are best positioned to see the 'big picture' in their lines of responsibility. Information on customer needs and production possibilities must flow upward from the shop or field to division managers and corporate officials; information on policies and decisions must flow downward for implementation. But information flows are commonly attenuated and distorted. Lower-level employees are loath to admit their known shortcomings, problems they anticipate but hope to avoid, and the organizational 'slack' they could turn into higher output if pressed. Through hubris, superiors often speak without listening or hear only what they want to hear. Revolutions in information transmission, processing, and storage technology have mitigated some of these problems, but they have by no means eliminated them. As Andrew Grove, one of the most thoughtful information technology business leaders of the twentieth century, observed (Byrne 2000):

> Brains don't speed up. The exchange of ideas doesn't really speed up, only the overhead that slowed down the exchange … You can reach people around the clock, but they won't think any better or any faster just because you've reached them faster.

3.4 *Suboptimization Failures*

Decentralization implies pushing decisions down to operating levels at which hands-on knowledge and responsibility for implementation are present. But when this is done, suboptimization failures can occur. Division A may supply a key input to Division B. But to meet its financial performance targets, Division A needs a generous markup of prices above costs which could adversely affect the financial performance of Division B or lead it to substitute a less suitable input for the high-priced input. Timely intervention by higher management, with a possibly adverse impact on the motivation of the sacrificing entity, is required to solve such 'transfer pricing' problems. Or functional group X may have the special skills needed to help functional group Y solve a problem, but the effort required from group X may come at the expense of meeting X's direct performance targets, so the desired cooperation fails. Again, subtle intervention by higher-level management can solve the problem, but only by deviating from the decentralization principle.

3.5 *Goal Conflicts*

Perhaps the greatest challenge is ensuring that all members of a business organization conform to higher-level goals. The problem begins at the top. In principle, it was said earlier, firms maximize profits, long-run or short-run, to serve the interests of their owners—in the case of corporations, the equity shareholders. But there are many reasons why this may not happen. A pathbreaking book by Berle and Means (1932) observed that as the ownership of modern business corporations became dispersed away from founder–owners toward increasing numbers of shareholders, each with too small a stake to warrant intervening actively in their investment target's management, the ownership function became separated

from the control function. Top managers, Berle and Means continued, exercised effective control over their organizations and enjoyed within broad limits discretion as to the objectives they would seek to advance. They might opt to maximize the value of shareholders' ownership interests, but they might equally well pursue alternative and conflicting goals—for example, maximizing the size of the business empires they control by acquiring other companies on terms that reduce original shareholders' stock values; paying themselves lavish salaries; enjoying perquisites such as richly furnished offices and trips in corporate jets to posh golf and skiing locales; contributing enthusiastically to local community cultural establishments; or simply seeking a 'quiet life' devoid of entrepreneurial risk-taking.

Corporate boards of directors are—in principle—supposed to represent stockholder interests. Reality is more complex. During much of the twentieth century, directors of American business corporations were chosen and appointed by company managers, owed their continuing presence on the board to the managers, and adopted codes of behavior that discouraged sharp questioning and opposition to management positions except in times of crisis (see Mace 1971). As a result, they provided at best a weak check against managerial actions that reduced shareholder value. In other nations, widely differing board of directors structures exist, but again, board members have often served as less than diligent stewards of common shareholder interests.

Gradually, this has changed. During the last two decades of the twentieth century there were pressures from institutional investors (such as pension funds), government regulators, and the stock exchanges throughout much of the world to improve 'corporate governance,' rendering corporate directors less beholden to inside management and more dedicated to defending stockholder interests. In the USA, for example, corporations have been induced to restructure their boards so that 'outside' members comprise a majority, to create board member nomination committees dominated by 'outside' directors, and to form compensation committees in which only outsiders determine the salaries and bonuses of inside managers. Much more emphasis has been placed on bonuses and stock options rather than straight pay as means of motivating top managers to perform well on behalf of stockholders. As a result, boards have come to intervene more aggressively in company decision-making and, when performance has been less than satisfactory, to fire the responsible top executives.

Nevertheless, problems remain. The compensation of top corporate managers in the USA during the 1990s rose to unprecedentedly high multiples of average company salaries. One reason was a booming stock market that enhanced the value of stock options without requiring extraordinary performance on the part of managers. Also, with superior information on company prospects, inside managers could 'game' the timing of stock option awards and bonus target-setting to maximize the chances of a favorable outcome. And to help them in the task of top manager compensation-setting, board committees hired outside management compensation consultants who instigated emulatory competition among companies to reward their executives—few of them leadership superstars—at least as liberally as peers and in the process raised the average for all.

At the end of the day, leadership and the ability to create effectively functioning teams are crucial to organizational success. Even among those who have reached the top ranks in business organizations, the quality of leadership varies widely. Graduate schools of management have found it much harder to teach leadership than any of the other requisite arts and skills (but see Heifetz 1994). Economists have few qualifications for saying anything sensible about how superior leadership emerges, so the subject is left to other disciplines contributing to this *Encyclopedia*.

The goal conflicts observed between shareholders and top managers are replicated manyfold at the lower levels of corporate organizations. Operating unit managers 'game' against their corporate superiors to set performance targets and bonus schemes that are biased on the side of yielding generous compensation. Knowing that personnel turnover is high and corporate staff memories short, operating unit managers take actions that sacrifice durable profitability for the sake of higher short-run profits and bonuses. Production and sales force supervisors recognize that their tasks could be carried out with smaller staffs, but refrain from profit-enhancing cutbacks to hold reserves against future contingencies and to maintain a 'nice guy' reputation. These incentive conflicts might be reduced in companies with 'superincentive' compensation schemes; for example, in high-technology startup companies that award potentially lucrative stock options to all employees in responsible positions. The phenomenon is a relatively new one on which clarifying research is needed.

4. *Natural Selection*

Achieving good economic performance at the operating levels of complex organizations is not easy. Communication and incentive failures abound. There is considerable X-inefficiency. But goods and services do get delivered at levels that on average are nothing short of remarkable. One reason is that in national economies with firmly implanted capitalist institutions, most business firms must strive to improve their internal incentive systems and hew to something approximating a profit maximization line lest they be displaced by competitors who do a better job of satisfying consumer wants. In vigorously competitive

industries, firms that stray too far from the path of profit maximization are unlikely to withstand the forces of natural selection (see Alchian 1950, Nelson and Winter 1982). When competitive pressures are weak, however, product market forces may be insufficient to jeopardize non-maximizers' survival. In such cases—more the exception than the rule in well-functioning economies—the disciplinary job is left to the still-imperfect instruments of corporate governance.

See also: Bounded Rationality; Business History; Corporate Culture; Corporate Governance; Information and Knowledge: Organizational; Information, Economics of; Organizational Decision Making; Organizations, Sociology of; Population Ecology

Bibliography

Alchian A A 1950 Uncertainty, evolution, and economic theory. *Journal of Political Economy* **58**: 211–21

Berle A A, Means G 1932 *The Modern Corporation and Private Property*. Macmillan, New York

Byrne J A 2000 Back to the future: Visionary vs. visionary. *Business Week*. August 28: 210–14

Chandler A D Jr 1962 *Strategy and Structure: Chapters in the History of American Industrial Enterprise*. MIT Press, Cambridge, MA

Coase R H 1937 The nature of the firm. *Economica* **4**: 386–405

Cyert R M, March J G 1963 *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ

Gaskins D W Jr 1971 Dynamic limit pricing: Optimal pricing under threat of entry. *Journal of Economic Theory* **3**: 306–22

Heifetz R A 1994 *Leadership Without Easy Answers*. Harvard University Press, Cambridge, MA

Jensen M C 1988 Takeovers: Their causes and consequences. *Journal of Economic Perspectives* **2**: 21–48

Keynes J M 1936 *The General Theory of Employment Interest and Money*. Harcourt, New York

Leibenstein H 1966 Allocative efficiency vs. "x-efficiency". *American Economic Review* **56**: 392–415

Mace M L 1971 *Directors: Myth and Reality*. Harvard Business School Division of Research, Boston

Machlup F 1946 Marginal analysis and empirical research. *American Economic Review* **36**: 519–54

Marris R 1963 A model of the managerial enterprise. *Quarterly Journal of Economics* **77**: 185–209

Nelson R R, Winter S G 1982 *An Evolutionary Theory of Economic Change*. Harvard University Press, Cambridge, MA

Richardson G B 1959 Equilibrium, expectations, and information. *Economic Journal* **69**: 223–37

Scherer F M 1988 Corporate takeovers: The efficiency arguments. *Journal of Economic Perspectives* **2**: 69–82

Simon H 1959 Theories of decision-making in economics and behavioral science. *American Economic Review* **49**: 253–83

Simon H 1979 Rational decision-making in business organizations. *American Economic Review* **69**: 493–513

US Bureau of the Census 1986 *1982 Enterprise Statistics: General Report on Industrial Organization*. ES-82-1, Washington, DC

Williamson O E 1970 *Corporate Control and Business Behavior*. Prentice-Hall, Englewood Cliffs, NJ

F. M. Scherer

First Language Acquisition: Cross-linguistic

In this article, motivations for cross-linguistic research into children's acquisition of their first language(s) are noted with a brief historical outline of the field (Sect. 1). This is followed by a review of trends which appear to be universally shared across children acquiring different languages (Sect. 2), compared with the impact of typological properties of particular target languages on processes of acquisition (Sect. 3). The article concludes with a note on desiderata (Sect. 4).

1. Approaches to the Field

The major motivation for the cross-linguistic study of language acquisition, as of cross-linguistic research in general, is to compare what is universally shared across children irrespective of the specific target language or speech community to which they belong, on the one hand, with interlanguage variation and the impact of properties peculiar to a particular language or language type on the other (Slobin 1997b). A related, psycholinguistically motivated question is how the specifics of target-language typology interact with acquisitional principles and strategies deriving from more general extralinguistic factors of perception, cognition, and social interaction (Slobin 1985b).

The first studies of children's acquisition of their native language(s) which can be characterized as cross-linguistic in orientation date from the nineteenth and early twentieth centuries. These were generally diary studies of childhood bilingualism, among children raised in more than one language, mainly in Europe. Such studies provided potentially rich data for cross-linguistic comparisons, as did a range of studies of children acquiring languages other than English (Ferguson and Slobin 1973). However, these were by and large not systematically directed to the testing of psycholinguistically motivated hypotheses, and they tended to focus on phonology, inflectional morphology, and the lexicon, with little concern for syntax or semantics.

The rise of developmental psycholinguistics in the USA in the 1960s created a field of inquiry which aimed to integrate insights from linguistic theory and description with developmental patternings in first language acquisition. Such studies tended to be anglocentric, but provided a basis for research in other

languages. For example, the general developmental stages delineated in Brown (1973) for the early years of grammatical development appear to apply across languages (see Sect. 2 below).

This anglocentricism was countered by the Berkeley project on four typologically distinct languages (Slobin 1982). Since then, data have been collected from a wide range of languages, both across and outside Europe. These include surveys of several dozen typologically distinct languages in the five volumes edited by Slobin (1985–1997). Other collections focus on particular domains including: early clause structure (Bowerman and Perdue 1990), the lexicon–syntax interface in verb acquisition (Lieven 1998), and narrative construction (Berman and Slobin 1994, Hickmann et al. 1996). Generative grammars gave rise to a parameter-based analysis of acquisition in different languages and domains, for example, Hyams (1986) and Weissenborn (1992) on null subjects, Meisel (1992) on functional categories and the ordering of constituents (see *Generative Grammar*).

Research in child language since the 1980s has thus yielded a large body of data on acquisition of numerous languages, including ones with non-European typology. Many of the latter studies are anchored in ethnographic-type fieldwork, for example, for Inuktitut by Shanley Allen, for Mayan Quiché by Clifton Pye, and for Sesotho by Katherine Demuth; or in research conducted by native speakers, as in Soonja Choi's work on Korean. This work combines to demonstrate that (a) children early on adapt to the particular constraints of their native-language typology, as verb-initial or verb-final, accusative or ergative, as synthetic or agglutinating; (b) they do so on the basis of the particular cultural norms for caretaker–child interaction in their speech community; (c) they do so with equal rapidity and ease across languages; and (d) within languages, the acquisition of certain subsystems may be delayed for language particular reasons of relative structural and/or semantic complexity, on the one hand, and because they belong to restricted registers of usage not accessible to young children, on the other.

2. *Universal Trends in Language Development*

Different points of departure underlie claims about universals in language acquisition. A nativist, Chomskyan-based view will specify these in terms of the formal principles characterizing a generative model of Universal Grammar (UG) with cross-linguistic variation accounted for by differential setting of structural parameters according to the target language. A Piagetian orientation will explain them in terms of general cognitive principles and shared conceptual underpinnings which provide a universal epistemological basis for acquisition of linguistic knowledge (Slobin 1985). The view espoused by this author is that children start out the task of acquisition equipped with categorial and semantic distinctions of the type termed by Chomsky 'substantive' as against formal universals, akin to Keenan's (1979) idea of 'naive' or primitive universals (Berman 1993b). For example, children do not need to learn that there are two main kinds of sound segments, consonants and vowels; that propositions distinguish between predicates and arguments; or that there are intransitive and transitive clauses, with one-, two-, or three-place predicates.

Whatever the explanation they propose, researchers agree that there is a clear biological underpinning to the overall developmental stages manifested by children in command of language production irrespective of mother tongue. Just as children sit before they stand, and walk before they run, so children across the world babble before they produce single words, rely on single words before they start combining words, and produce isolated clauses before they join these together by increasingly specific lexical means and complex syntactic devices.

Overall progression thus involves an evolution in construction types. Language production typically starts out with utterances, as behavioral elements which provide the initial basis for expressing relations between semantic content and phonetic form. These typically take the form of isolated lexical items or unanalyzed strings of words, starting around the end of the first year of life. Structure dependence emerges when these lexical units are combined in simple-clause structure, realized by morphosyntactic markings of verb-argument relations. Next, adjacent clauses are combined, first by coordination, subsequently by various kinds of subordinate constructions—initially mainly finite and subsequently nonfinite—leading up to complex clause-linkage by derived nominalizations (Berman 1993a). This needs to be qualified for a verb-final language like Turkish, where the unmarked means of clause linkage is by nonfinite, nominalized constructions. These emerge relatively early in Turkish compared with European or Semitic languages.

With respect to acquisition of morphology, the domain in which languages tend to differ most, children generally acquire the bulk of the inflectional system of their language before they master its derivational morphology. This is understandable since inflection is more regular, productive, and grammatically obligatory than derivation. Besides, derivational processes depend on a level of vocabulary expansion not achieved by young preschool age. In fact, in some instances, like derived nominals or the high-register Latinate lexicon of English, children will master derivational processes only well into school age. In contrast, they by and large command the inflectional systems of their language by the age of three, irrespective of how morphologically opaque or synthetic these happen to be. Cross-linguistic differences affect

acquisition of inflection along two main dimensions. Psychological processing principles like perceptual saliency coupled with regularity and transparency of form-meaning mappings account for the fact that, for example, Turkish-speaking children gain command of the bulk of their rich system of agglutinating, inflectional suffixes by a much earlier age than children acquiring languages with more synthetic or syncretic morphologies like Hebrew or Icelandic. And experiential factors of frequency of use in the everyday colloquial speech that constitutes the linguistic input and output of young children explain why certain grammatical categories are acquired relatively late in languages where they constitute high-register or literate constructions.

In the domain of syntax, children typically gain command of simple clause structure in their language between ages two to three years. Basic word order emerges very early indeed, first between phrasal constituents, only later within phrases. Constituent ordering is followed by grammatical marking of verb-argument relations, around the same time or somewhat before inflectional marking of grammatical agreement. The overall progression is thus much as Brown (1973) defined for English: from single words to early word combinations, thence to simple clause structure, followed by a range of devices for clause linkage.

Another commonality is the general development observed across children from rote-learning of unanalyzed items to partial productivity followed by command of structure-dependent rules and generalizations, which are subsequently constrained by conventions of lexical and discourse usage. This has been demonstrated, for example, for how children acquire the semantics, morphology, and syntax of causative verbs in three typologically distinct languages: English at the more analytic or syntactic end of the scale, Hebrew with relatively synthetic bound morphology, and Inuktitut as a polysynthetic language which marks causativity by both morphological and lexical means (Allen, in Lieven 1998).

General language development thus proceeds across different phases, from pregrammatical to structure-dependent to appropriate use. And it becomes both more elaborated and varied in structure and use, on the one hand, and more restricted or constrained by target-language specifics, on the other. Thus, in acquiring the sound system, babies proceed from universal crying to uttering a range of speech sounds, followed by babbling, which is increasingly restricted to elements of the native language to a point where eventually the system will be so constrained as to impose a non-native pronunciation in other languages. Similarly, in morphology, across languages children proceed from nonalternation between forms to immature strategies for initial markings of grammatical categories, thence to 'creative errors' of over-regularization, until eventually the system becomes fully constrained by lexical convention. And in the acquisition of the lexicon, again across languages, children move from a restricted vocabulary of nursery terms and over or underextensions of conventional items to a period of creative innovation; eventually the conventional lexicon takes over, while at the same time vocabulary growth is both constrained and elaborated by socially and culturally appropriate distinctions of register and usage, lasting into adolescence.

3. *Impact of Target-languageProperties*

As against these cross-language commonalities and claims for universal constraints, whether formal-linguistic or perceptual-cognitive, current research reveals a marked effect of target-language typology from a very early age. Such more language-specific claims have been explained in cross-linguistic perspective as reflecting a modified, neo-Whorfian characterization of the relation between language and thought (Slobin 1996) and as demonstrating the critical impact of caretaker input from the very start (Bowerman 1995). This very early sensitivity to target-language properties has been observed in a variety of linguistic domains. For example, language-specific effects exert an influence on speech perception and babbling during the first year of life (Jucszyk 1997). And well before they have command of grammatical inflection, children will adjust their speech output to the prosodic character of their language, as intonation or tone-based (Demuth 1993), and they will show rapid command of vowel-harmony where required, as in Turkish (Aksu-Koç and Slobin 1985).

Children's encoding of the semantics of space and which distinctions they mark linguistically (e.g., in shape or location of figure and ground respectively) is also markedly influenced by target-language specifics. Thus, the universality of the early preference among children acquiring European languages (and also Hebrew) for the prepositions *in*, *on*, and then *under*, which was convincingly explained in terms of perceptual properties of space and shared cognitive preferences, has now been queried, since it turns out that quite different factors govern the encoding of spatial distinctions in languages like Korean or Tzeltal compared with English and Dutch (Bowerman 1995). Children appear to relate to the categories relevant to their own native language without initially manifesting universal patternings in the expression of spatial distinctions.

Null subjects are another domain in which universalist claims have been made. Thus, Hyams (1986) suggested that children first set the prodrop parameter to the unmarked value of plus, irrespective of target language, since English-speaking children in the early stage of grammar acquisition produce numerous

subjectless sentences, as do their Italian peers. Weissenborn (1992) showed, however, that target language input has a marked influence from very early on in this respect, since children acquiring French produce fewer subjectless sentences than those acquiring German, and the latter do so less than children acquiring strongly pro-drop languages like Italian and Spanish.

The very early impact of target language is evident not only in which categories are distinguished, but also how. This is clearly shown in Clark's (1993) cross-linguistic comparisons in the domain of word formation. When young children express the notion of 'undoing' or reversal of actions, they do so from early on in ways peculiar to their native tongue: by the prefixes *un-* in English and French *dé-*, and by particles like *ab* and *aus* in German, *af* and *uit* in Dutch. Hebrew-speaking children, in contrast, must do this lexically or periphrastically, since their language lacks the morphological means for this semantic category. Similarly, in order to express causativity, children will start by using syntactic means with verbs like *make* in English and *faire* in French, a construction type used only rarely by Hebrew-speaking children for this purpose. Subsequently, English-speaking children will overextend verbs like *fall* or *sing* to transitive contexts; in Inuktitut, children will first use morphological and then lexical means of causativity while in Hebrew, they will first neutralize and then alternate morphological patterns to mark verb transitivity (Allen in Lieven 1998, Berman 1993b).

This research converges to suggest that apparently universal acquisitional principles of structural simplicity and semantic transparency in form-meaning mappings are affected and perhaps even superseded by typologically pervasive properties of the target language. For example, it has been argued that children will find it easier to juxtapose words than to affix bound elements to a stem or root, and that these processes are perceptually more salient and so more accessible to young children than stem-internal vowel alternation. However, in a language like Hebrew, which uses vowel alternation to express basic grammatical distinctions such as verb-tense, children mark past vs. present tense at much the same time as their counterparts acquiring languages which use more linear morphological means. Moreover, although their language allows both noun compounding and external, linear affixation to a stem for new word formation, Hebrew-speaking children typically innovate nouns, verbs, and adjectives by the more typically Semitic device of attaching syllabic affixes to a consonantal root: and they rarely use the option favored by English-speaking children for creating novel verbs by zero derivation or syntactic conversion (Berman 1999).

Children's acquisition of word classes and the composition of their early lexicon are other areas which suggest that claims for universal underpinnings of acquisition may need to be modified. Prior research on children's lexical repertoire at the one-word stage in several languages showed that more nouns were acquired earlier than verbs, while the vocabulary spurt in the period before early word combinations also consisted mainly of noun like terms. Gentner (1982) explained this in terms of general cognitive principles, since verbs as relational terms are holistic rather than coherently packaged entities, while nouns (at least those in the input and output of one- to two-year-olds) tend to encode semantically 'natural' categories that refer to concrete objects which are perceptually identifiable. Recent research has shown, however, that this noun bias may be an artifact of the nature of the lexicon and lexical input in the target languages, and that in some cases, verbs may be acquired earlier and in greater quantities than nouns (for example, the studies by Brown and by Choi in Lieven 1998).

At more advanced stages, developments in form-meaning mappings in extended discourse provide further evidence for the impact of 'typological imperatives.' A five-language cross-linguistic sample of picturebook narratives produced by children aged three to five and nine-year-olds compared with adults (Berman and Slobin 1994) showed that the expression of textual temporality was critically affected by native-language tense/aspect systems even among our youngest subjects. In Hebrew, children marked only grammatical tense, in Spanish they distinguished both simple from progressive and perfective from imperfective aspect, while in Turkish they gave grammatical expression to modality. Moreover, there was no evidence for 'compensation' in the sense that children attempted to give lexical expression to notions not grammatically encoded in their language (e.g., by using periphrastic adverbs to mark aspectual distinctions in languages like German or Hebrew). On the contrary, not only 'do speakers tend to limit their attention to those semantic distinctions marked in the grammar, they may even overattend to such distinctions in early phases of development' (Berman and Slobin 1994, p. 619). Slobin (1996a, 1996b, 1985–97) has since extended these observations to the expression of locative trajectories. Proceeding from Talmy's (1985) distinction between verb-framed languages (Hebrew, Spanish, and Turkish in our sample) and satellite-framed languages like English and German that rely heavily on locative particles, he shows that the way young children—like adult speakers of the same language—talk about motion, paths, and directionality depends critically on how these notions are grammatically and lexically encoded in their language. Here, too, the underpinnings of such distinctions are evident very early: at the one-word stage, Germanic-speaking children use particles like *off* or *aus* to express direction of motion in ways encoded by verb-stems by children acquiring French, Spanish, or Hebrew. For example, at the age of 18 months, my Hebrew-English bilingual daughter would say either English *down* or a

truncated stem form of the Hebrew verb *la-redet* 'to get/go down,' when she wanted to be taken out of her high chair.

In another cross-linguistic project comparing children's narrative texts produced by children aged from five to 15 years on the basis of two picture-series, Hickmann et al. (1996, p. 391) focused on the devices for introducing, maintaining, and shifting reference to characters in a story. They found that the English, French, German, and Chinese 'rely differentially on local vs. global devices to mark newness: postverbal position is obligatory in Chinese …, indefinite determines in the other languages; across languages: newness markings emerge late (age seven), local markings emerge before global, and local and global markings are strongly related.' Their findings point to an important interaction between shared developmental trends and linguistic specificity. Relatedly, in the Berman and Slobin (1994) study, across languages, three-year-olds tended to relate to each picture in isolation, four- and five-year-olds were able to encode linearly expressed local relations between events in terms of local temporal relations of sequentiality and subsequently of causality, school-age children constructed a globally organized hierarchical narrative, and adults tended to adopt individual rhetorical styles with rich interpretation and a personal narrative stance.

Ongoing cross-linguistic research on the production of narrative and expository texts by schoolchildren and adults reaffirms such interaction between generally shared developmental trends and typologically determined expression of semantic distinctions and selection of rhetorical options. Nine-year-olds can all construct wellformed personal experience narratives, but are unable to produce coherent expository texts. Further, only from high-school age will one find consistent, deliberate, and appropriate differentiation between language usage across spoken and written texts, marking highly language-particular register distinctions. In contrast, overall genre distinctions appear similar across languages. For example, narrative texts are couched in past tense and mainly perfective aspect, and they have light subjects (mainly pronominal in English and French or null subjects in Hebrew and Spanish), while expository texts rely more on generic or extended present tense and have more lexical and heavier, more noun-phrase subjects and objects. This, too, interacts with target-language options and preferences. For example, in order to express a detached, impersonal perspective in English, passive is quite widely used even by grade-school children, increasingly by adults; and passive is also quite common in the Icelandic texts. In contrast, passive is used far less in other languages, even though it occurs as a structural option: Swedish subjects rely heavily on the impersonal pronominal *man* corresponding to the very common use of *on* in French; Spanish and Hebrew achieve the same rhetorical effect by means of subjectless or second person impersonal constructions. Similar patterns emerged for differential use of pronominal compared with null subjects and also for relative clauses as devices for syntactic packaging in the picturebook narratives.

Discourse-based cross-linguistic comparisons thus suggest that reliance on certain linguistic forms may depend not only on structural factors of relative morpho-syntactic complexity or productivity, but also on the extent to which a given rhetorical option is favored for specific discourse purposes in current language usage.

4. *Conclusion*

In sum, on the basis of shared cognitive and linguistic universals, even very young children are attentive to key typological properties of their language. From an early age, they recognize 'where the action is at' in the input language, not only which categories are formally distinguished, but also how. *When* this type of sensitivity finds expression will depend on general developmental, largely extralinguistic factors. For example, the spatial distinctions studied by Bowerman (1995) and her colleagues will precede command of derivational marking of linguistic subcategories, and these will emerge earlier than rhetorical mastery of linguistic forms in the context of extended narratives. But in each case, the specific ways in which children encode form-meaning relations accord with how this is done by adult speakers of the same target language rather than by children of the same age in other languages.

As noted in the introduction, a large body of data is currently available for acquisition of different domains in a rich range of languages from Europe and elsewhere. What now seems necessary is to extend systematic and deliberately comparable research designs across a range of languages and linguistic domains—along lines conducted for morpho-syntax by Slobin (1982), for word formation and derivational morphology by the structured elicitation and diary studies reported in Clark (1993), for semantic encoding of space by Bowerman (1995), for temporality by Weist and co-workers (Weist et al. 1993), as well as for narratives by Hickmann (1996) and by Berman and Slobin (1994). Such studies might profitably focus on two contrasting directions of cross-linguistic comparison: (a) languages which are typologically very distinct, for example, ergative or accusative, verb-initial or verb-final, analytic versus synthetic or agglutinating; and (b) languages which are typologically close, like Dutch and German, or Arabic and Hebrew, but which differ along certain restricted dimensions. Such work should aid in specifying the relative weight of the various factors touched on here. It would help to delineate the interaction between language-universal and language-particular, between intralinguistic,

cross-linguistic, and extralinguistic factors in acquisition. And it could provide psycholinguistic motivation for the distinction between language-specific, typologically shared, and cross-linguistic or universal features of language structure and use in developmental perspective.

See also: Language Acquisition; Language Development, Neural Basis of; Linguistic Typology; Second Language Acquisition; Vocabulary Acquisition, Psychology of

Bibliography

Aksu-Koç A A, Slobin D I 1985 The acquisition of Turkish. In: Slobin D I (ed.) *The Cross-linguistic Study of Language Acquisition: Vol. 1*. Erlbaum, Hillsdale, NJ, pp. 839–78

Berman R A 1993a Crosslinguistic perspectives on native language acquisition. In: K Hyltenstam, A Viberg (eds.) *Progression and Regression in Language*. Cambridge University Press, Cambridge, UK, pp. 245–66

Berman R A 1993b Developmental perspectives on transitivity: A confluence of cues. In: Levy Y (ed.) *Other Children, Other Languages: Issues in the Theory of Language Acquisition*. Erlbaum, Hillsdale, NJ, pp. 189–241

Berman R A 1999 Children's innovative verbs versus nouns. In: Menn L, Bernstein-Ratner N (eds.) *Methods for Studying Language Production*. Erlbaum, Mahwah, NJ, pp. 69–93

Berman R A, Slobin D I 1994 *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Erlbaum, Hillsdale, NJ

Bowerman M 1995 Learning how to structure space for language: A crosslinguistic perspective. In: Bloom P, Peterson M, Nadel L, Garrett M (eds.) *Language and Space*. MIT Press, Cambridge, MA, pp. 385–486

Bowerman M, Perdue C (eds.) 1990 *Linguistics Special Issue: The Structure of the Simple Clause in Acquisition*, pp. 28–36

Brown R 1973 *A First Language: The Early Stages*. Harvard University Press, Cambridge, MA

Brown P 1998 Children's first verbs in Tzeltal: Evidence for an early verb category. In: Lieven E (ed.) *Linguistics Special Issue: Developing a Verb-category: Cross-Linguistic Perspectives*, pp. 713–54

Clark E V 1993 *The Lexicon in Acquisition*. Cambridge University Press, Cambridge, UK

Demuth K 1993 Issues in the acquisition of the Sesotho tonal system. *Journal of Child Language* **20**: 275–301

Ferguson C A, Slobin D I (eds.) 1973 *Studies of Child Language Development*. Holt, Rinehart and Winston, New York

Gentner D 1982 Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In: Kuczaj S A II (ed.) *Language Development, Volume II: Language, Thought, and Culture*. Erlbaum, Hillsdale, NJ, pp. 301–34

Hickmann M 1998 Form, function, and context in narrative development. *Journal of Pragmatics* **29**: 35–56

Hickmann M, Hendriks H, Roland F, Liang J 1996 The marking of new information in children's narratives: A comparison of English, French, German, and Mandarin Chinese. *Journal of Child Language* **23**: 591–620

Hyams N M 1986 *Language Acquisition and the Theory of Parameters*. Reidel, Dordrecht, The Netherlands

Jusczyk P W 1997 *The Discovery of Spoken Language*. Bradford Books, Cambridge, MA

Keenan E L 1979 On surface form and logical form. *Studies in Linguistic Sciences* **8**: 2

Lieven E 1998 (ed.) *Linguistics Special Issue: Developing a Verb-category: Cross-Linguistic Perspectives*, pp. 36–4

Meisel J M (ed.) 1992 *The Acquisition of Verb Placement*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Pye C 1994 A crosslinguistic approach to the causative alternation. In: Levy Y (ed.) *Other Children, Other Languages: Issues in the Theory of Language Acquisition*. Erlbaum, Hillsdale, NJ, pp. 243–64

Slobin D I 1982 Universal and particular in the acquisition of language. In: Wanner E, Gleitman L R (eds.) *Language Acquisition: The State of the Art*. Cambridge University Press, Cambridge, UK, pp. 128–70

Slobin D I 1985–1997 (ed.) *The Crosslinguistic Study of Language Acquisition, Volume 1* (1985a), *Volume 2* (1985b), *Volume 3* (1993), *Volume 4* (1997a), *Volume 5* (1997b). Erlbaum, Hillsdale, NJ

Slobin D I 1996 From 'thought and language' to 'thinking for speaking'. In: Gumperz J J, Levinson S C (eds.) *Rethinking Linguistic Relativity*. Cambridge University Press, Cambridge, UK, pp. 70–96

Talmy L 1985 Lexicalization patterns: Semantic structure in lexical forms. In: Shopen T (ed.) *Language Typology and Syntactic Description, Vol. 3: Grammatical Categories and the Lexicon*. Cambridge University Press, Cambridge, UK, pp. 36–149

Weissenborn J 1992 Null subjects in early grammars: Implications for parameter-setting theory. In: Weissenborn J, Goodluck H, Roeper T (eds.) *Theoretical Issues in Language Acquisition: Continuity and Change in Development*. Erlbaum, Hillsdale, NJ, pp. 269–99

Weist R M, Kaczmark A, Wysocka J 1993 The function of aspectual configurations in the conversational and narrative discourse of Finnish, Polish, and American children. *Papers and Studies in Contrastive Linguistics* **27**: 76–109

R. A. Berman

First Past the Post, in Politics

1. *First Past the Post, in Politics*

First past the post (FPTP) is the name given to the simplest possible system of electing persons to office. Also known as *single member plurality* system, FPTP refers to contests where the candidate with the most votes, and only that candidate, is elected from a given electoral district. FPTP is used in legislative elections in 68 of the 211 independent states and related territories identified in the *International IDEA Handbook of Electoral System Design* (1997, p. 20). Even though FPTP is used in about a third of the states and territories, countries with FPTP constitute nearly half of the populations included in these units, because

they include populous countries such as Bangladesh, India, Nigeria, Pakistan, the UK, and the USA. Many of the countries with FPTP are former British colonies.

Two features distinguish FPTP from other electoral systems: the number of elected seats per district (one) and the rules by which those seats are allocated among candidates (simple plurality). Because both of these factors contribute to the unique character of FPTP, it is worth considering them in sequence.

1.1 Single-member Districts

Under FPTP, only one person is elected per district. In the US House of Representatives, for instance, although there are 435 districts, each district is represented by a single person (hence the term 'single-member district'). The same is true for the British House of Commons. In contrast to FPTP, in many national legislatures, the top two, three, or *n* vote-getters are elected to represent the district. In the extreme case, the entire country is a single multi-member district, and all legislators can be said to represent the same district. The Israeli Knesset, for example, has a solitary district comprising the entire country, and all 120 members of Knesset represent it. The same is true of the Dutch legislature, which has 150 members from a single national district.

In the systematic comparison of types of electoral systems, the number of elected seats within a district is termed the *district magnitude*. In FPTP, where the district magnitude is one, voters usually vote directly for candidates. If a party's candidate wins in a district, that party has won essentially 100 percent of the representation for that district, despite receiving only a portion, and not necessarily more than half of the total votes. For example, suppose that there are three parties evenly distributed over a hundred districts. If the vote in each district is 40 percent for party A and 30 percent each for parties B and C, party A will have 100 percent of the seats with 40 percent of the votes. This pattern (with more realistic numbers) often gives the largest single party a proportion of seats in the legislature that is higher than its proportion of votes in the electorate. Thus, leading or majority parties in FPTP tend to be over represented.

This exaggeration of the seat shares of the parties that are largest in votes is not unique to FPTP. Virtually all electoral systems do so, but FPTP does so more than proportional representation systems (Rae 1971). However, it is also possible in FPTP systems for a party to receive a majority of votes, but only a minority of seats, for example, when such a party's supporters are heavily concentrated in a few districts.

In systems where the district magnitude is high (and therefore, by definition, not FPTP), voters typically cast their ballots for parties who are then allotted legislative seats more or less in proportion to their share of the electorate's vote. In general, higher district magnitudes imply some kind of proportional representation (see *Proportional Representation*).

1.2 Plurality Rule

Within single-member district systems, a winner can be picked via one of several alternative rules. Plurality, the simplest rule, implies that the one with the most votes wins. Other options include absolute (more than 50 percent) majority requirements, with runoff elections between the top two finishers if necessary; approval voting, wherein voters give one vote to each candidate they approve of; the Condorcet procedure, in which the winner of a series of paired comparisons (if there is one) is designated the winner; and the Borda count, wherein points are assigned by voters to candidates in order of their preference.

When there are only two candidates, simple plurality rule is equivalent to majority rule and has four desirable properties: decisiveness, anonymity, neutrality, and positive responsiveness (May 1952). When there are more than two candidates, neither plurality rule nor any other system fulfils another (weaker) handful of simple desiderata: universal admissibility of preference orderings, the Pareto criterion, independence of irrelevant alternatives, and nondictatorship (Arrow 1963).

1.3 Consequences of FPTP

The great virtue of FPTP is its simplicity. Voters make a single choice and the candidate with the most votes wins. However, FPTP has fewer desirable properties than many other methods in multicandidate elections. Compared to other rules, plurality is the least likely to pick a Condorcet winner, i.e. a candidate who is able to beat all other candidates in pairwise contests. Also, simulations show that the plurality rule results in average social utilities that are lower than those under other rules (Merrill 1988). Such drawbacks may not be very consequential when there are only two candidates. Furthermore, when there are only two options, as in a two-party system, there is no need for more complex electoral systems.

When three or more candidates compete in a FPTP election, some of the candidates may divide the votes of some segment that might otherwise agree on a winning candidate. This could result in a victorious candidate who is less preferred by the electorate than one of the losers. If voters are aware of this eventuality they may be able to coordinate in order to elect a more preferred candidate. FPTP encourages voters to cast their ballots strategically and consider the behavior of other voters. If a voter's first-choice candidate has little chance of getting a plurality of votes, the voter

may find it advantageous to vote for a less preferred but more viable candidate instead of 'wasting' his or her vote on a candidate with little chance. In contrast, multi-member districts and even single-member districts with decision rules other than plurality may reduce the incentives for voters to deviate from their first choice.

Strategic behavior and an aversion to wasting one's vote means that FPTP elections are typified by contests between two major candidates, whereas alternative systems often exhibit several viable candidates. Extending this result to parties, FPTP systems routinely have only two major parties, a phenomenon that has been coined *Duverger's law*, named after the French scholar who extensively described the relationship (Duverger 1954). Proportional representation systems, however, usually have more than two parties, a result that has come to be known as *Duverger's hypothesis* (Riker 1982, Cox 1997).

When there are only two viable candidates in an FPTP election, no abstentions, and the candidates are choosing their positions to maximize their electoral support, the candidate ideologically closest to the median voter is advantaged (Downs 1957). FPTP systems thus tend to produce more moderate outcomes than alternative systems. The incentive for candidates to take moderate positions can produce candidates who are only minimally differentiated, leading some to conclude that the choices in FPTP are often of little consequence. One potential advantage to such a system, however, is that the outcomes from one election to the next tend to be more ideologically consistent, resulting in small, incremental policy changes over time.

See also: Electoral Systems; Political Representation; Voting, Sociology of; Voting: Tactical

Bibliography

Arrow K J 1963 *Social Choice and Individual Values*, 2nd edn. Wiley, New York

Cox G W 1997 *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge University Press, Cambridge, UK

Downs A 1957 *An Economic Theory of Democracy*. Harper, New York

Duverger M 1954 *Political Parties*. Wiley, New York

International IDEA Handbook of Electoral System Design 1997. International Institute for Democracy and Electoral Assistance, Stockholm

May K O 1952 A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica* **20**: 680–4

Merrill S III 1988 *Making Multicandidate Elections More Democratic*. Princeton University Press, Princeton, NJ

Rae D W 1971 *The Political Consequences of Electoral Laws*, rev. edn. Yale University Press, New Haven, CT

Riker W H 1982 The two-party system and Duverger's law: an essay on the history of political science. *American Political Science Review* **76**: 753–66

G. D. Adams and W. R. Keech

First World War, The

1. Causes

The outbreak of the First World War in the summer of 1914 has been attributed to accident, design and confusion. The first position suggests that no one intended that armed conflict would break out in 1914; the second, that German and Austrian military elites planned the conversion of a Balkan diplomatic crisis into an armed confrontation; the third, that political and military leaders throughout Europe began to maneuver with some notion that war was possible, but found that they could not resist the momentum of confrontation (Joll 1992).

What was the sequence of events subject to these interpretations, and which has commanded the support of most historians who have addressed the issue of the descent into war in 1914? On 28 June 1914, the Archduke Franz-Ferdinand of Austria–Hungary, heir to the throne, was assassinated in Sarajevo. The crime was carried out by a group of Serb student patriots in neighboring Bosnia, then a province of the Austro–Hungarian empire. Their aim was to demonstrate the force of Slavic nationalism, and to challenge the domination of these south Slavic lands by imperialists in Vienna and Budapest.

In this aim, they succeeded completely. The government in Vienna decided to make this crime a matter of state, and to lay the blame for this political murder on the independent Kingdom of Serbia, and its intelligence services. On 5–6 July, Austrian officials went to Germany to seek support for their policy of making Serbia 'pay' for the assassination. German officials gave their approval, offering a 'blank check' to their Austrian allies. With this backing, on 23 July the Austrians presented Serbia with a list of demands which constituted a challenge to the status quo in the Balkans.

This is where Russia entered the conflict. A humiliation for Serbia would humiliate her Slavic ally, Russia. With Russian backing, Serbia met most of the Austrian demands, but refused to capitulate completely. In support of Serbia, Russia ordered partial mobilization of her armies on 29 July. In support of Austria, and in response to the numerical superiority of Russian forces, the German army mobilized too.

Once German mobilization was in progress, then the crucial link appears between a conflict in the Balkans and an all-European war. German military

planning had for years established a set of priorities to safeguard the security of the German empire. The primary point here was the need to avoid fighting a two-front war. Thus, it was necessary for Germany to eliminate a western threat on the part of Russia's ally, France, before facing the huge armies Russia could put in the field. To do so, a military plan—the Schlieffen plan—was developed through which German forces could move in a huge arc from Cologne in southwestern Germany westward through Belgium and then southward through France. The destination, after 42 days, was in the vicinity of Paris, where the French army would be destroyed, just as had occurred in 1870.

The Schlieffen plan was thus a German response to the threat of a two-front war. That threat had materialized out of the confrontation of Germany's ally Austria–Hungary with Russia's ally Serbia. But the German plan to destroy the French army by an arcing move through Belgium threatened to bring in the UK, guarantor of Belgian independence (Ritter 1958).

And that is precisely what happened. The German invasion of Belgium precipitated a state of war among the five great powers: Germany and Austria–Hungary on one side, and France, Russia and the UK on the other. By 4 August 1914, the First World War had begun.

Had anyone intended this moment to occur? Probably not. While the chief catalyst of the war crisis was the decision by Austria and Germany to press Serbia to pay a political price for the assassination, the outbreak of war was less a conspiracy than a complex mixture of arrogance, a sense of national 'honor,' ignorance, and confusion. All the actors in this drama misjudged the other side. Each move forward in the crisis was followed by a further heightening of tension. First Austrian honor was at stake; then Russian honor; then German fears of encirclement led her to respond to a perceived threat from Russia by threatening Russia's ally, France; and when confronted by a likely invasion of France, the UK hesitated, and then in defense of her honor, she declared war on Germany.

Two elements in the crisis were crucial in the prevention of a diplomatic solution. The first was temporal; the second, structural. First, timing. The war crisis lasted for one month. But a closer view of these events presents us with an even shorter timetable in the slide towards war. One week separated the Austrian ultimatum to Serbia—on 23 July—from Russian and German mobilization—29–31 July. In that time period, it was simply impossible for alternatives to be explored, or for the weight of antiwar and pacifist opinion, by no means negligible, to be brought to bear on political leaders. The crisis moved too fast to be stopped.

The second determinant of a general European war was, broadly speaking, structural. That is, the chronic instability of southeastern Europe was insufficient in and of itself to bring the Great Powers to war. What turned a minor conflict into a major crisis was the linkage between the Balkan conflict and the longer-term balance of power between Germany and the UK. Once the UK and Germany were aligned on opposite sides, then the crisis of 1914 became a contest for control over northwestern Europe. Should Germany defeat France in 1914, as she had done in 1870, the German navy would occupy the Channel ports. British shipping lanes, necessary to feed her home population, would be open at the pleasure of the German navy. No British government could accept that prospect (Kennedy 1980).

It is true that the British Foreign Secretary, Lord Grey, did not make this issue clear in his diplomatic handling of the war crisis. The German Chancellor, Bethmann-Hollweg hoped that the UK would stay out of the war. But geographical and strategic imperatives overshadowed diplomatic miscalculations. An invasion of Belgium and France posed a real and present danger to British vital interests. And when the UK entered the war, so did the British empire. What had started in Sarajevo, echoed in Sydney, Capetown and Calcutta. By 4 August 1914, the world was indeed at war.

2. *Conduct*

The 1914–18 war spanned the globe, but its outcome was determined by what happened on the battlefields of Western Europe. This article therefore offers an outline of the major military encounters of the war, encounters which left lasting traces both on the landscape of Europe and on the contours of the rest of the twentieth century.

The battlefield in 1914 was Victorian in character. Artillery provided more fire power than ever before, but most of it was horse-drawn. Four years later, the battlefield was transformed. By the end of the war, it took on contours recognizable to most soldiers who would come to fight in later conflicts. Infantry, artillery, tanks, and air power were all coordinated through communications systems and supply systems of great complexity and sophistication. Supplying animal power meant relatively little; supplying machines meant everything (van Creveld 1977). The 'storm of steel' in the words of the German infantryman Ernst Jünger, had come to stay (Jünger 1929).

2.1 *1914*

The two major military encounters of 1914 produced very different outcomes. In Belgium and France, the German army, one million strong, pushed forward in a swing westward and then southward. It met stiff resistance from Belgian forts, and then was harried by stubborn French and British defensive action. As the

German army moved west into Belgium, the French army moved east towards Alsace and Lorraine, provinces lost to Germany in 1871. German defenses held and inflicted massive casualties on the French. On one day, 24 August 1914, over 24,000 French soldiers died in action. This 'Battle of the Frontiers' was a massive defeat for the French. But the strength of their east–west rail links enabled the bulk of French forces to move west to meet the real threat of the German advance towards Paris. There on the river Marne, the German thrust was stopped. The German army retreated northward and dug in on the river Aisne. When French and British forces pursuing the German army reached the Aisne, they were unable to move the invaders. There the 'Western Front' was born. It soon stretched from Belgium to Switzerland.

There is little doubt that the Battle of the Marne was a German catastrophe. The whole purpose of German strategy had failed. There would be a two-front war. By the end of 1914 the chief of staff of the German army, von Moltke, had been replaced by Erich von Falkenhayn. But what redressed the strategic balance was a massive German military victory on the Eastern front. Two columns of Russian troops advanced into East Prussia at the same time as the German army moved through Belgium and France. These columns were separated by a string of lakes. A German staff officer, Max Hoffmann, took advantage of the geography and of the existence of a north–south rail link. Under the cover of darkness, the German army in the north moved south. The Russian column in the north had no idea they were facing an empty landscape; the Russian column in the south had no idea they were facing a numerically superior force, that is, until they were annihilated at the Battle of Tannenberg. The commander of Russian forces in the south, Samsonov, committed suicide. The two commanders of the German forces, who had not planned the operation, but who reaped its rewards, became national heroes. They were Paul von Hindenburg and Eric Ludendorff (Showalter 1991).

2.2 1915

The war spread in 1915, but the rough stalemate of 1914 remained unbroken. One major attempt to break it was in Turkey, then an ally of Germany. To 'stop our troops from chewing on barbed wire' on the Western front, as Winston Churchill, Britain's First Lord of the Admiralty, put it, a plan was hatched to knock Turkey out of the war by a naval operation. The aim was to use British naval power to force the straits that connected the Mediterranean Sea to the Black Sea. Once done, then the Turkish capital of Constantinople would be indefensible. The defeat of Germany's ally Turkey would serve many purposes. It would open supply lines to Russia; it would convince all and sundry, especially in the Islamic world, of the error of siding with Germany in the world war; it might even convince Germany to withdraw from France and Belgium.

Nothing of the sort happened. Here too, a mixture of arrogance, ignorance and confusion ruled. The enemy was Oriental, and all the distortions attending Western attitudes towards the Orient flourished. The enemy was underestimated. Attempts to push a naval force through the Dardanelles failed in March 1915. The Allies decide to mount a joint land–sea operation, but no one had a clear idea of the terrain on which soldiers would land. When they went ashore on 24–25 April, they faced steep cliffs and stiff Turkish resistance. They never got off the beaches. A six-month stalemate ensued, after which Allied forces withdrew.

2.3 1916

Two further attempts were made to shift the balance of power on the Western front. The first was at Verdun, in eastern France. There on 21 February 1916, more than one million shells fell on French positions in the hills north of the garrison city of Verdun. The idea was to force the French into total defense of this relatively unimportant position. The French, Falkenhayn reasoned, valued Verdun as a symbol of French freedom and power (Afflerbach 1994). He was right. The French committed 259 out of its full complement of 330 infantry regiments to the struggle for Verdun. They moved into Verdun along a vast conveyor belt, which the French called *la voie sacrée*, the sacred way. It certainly was a field of sacrifice. Falkenhayn succeeded in bleeding the French army white, but he did so at a massive cost to his own troops. Perhaps half a million men died at Verdun, where, by November, the French army had recaptured all the ground lost in the early days of the battle. Lasting ten months, the Battle of Verdun was the longest battle in history. The front remained about where it had been before the bloodbath (Horne 1962).

The same failure to break the stalemate occurred further to the north, where the British and French armies opened a major offensive on 1 July 1916. Here the fault lay in an underestimate of the enemy's fortifications and an overestimate of the power of artillery to destroy them. A one-week barrage by British guns was supposed to obliterate German defenses between Amiens and St Quentin in the lowlands of the Somme. But the deep dugouts under the German lines withstood the barrage, and when British forces moved out early in the morning of 1 July, they were mown down by German machine guns. Of the 100,000 men who went over the top, 60,000 were casualties, of whom 20,000 died in that single day. Thereafter, British and German forces regrouped. A further advance on 14 July was somewhat more successful, but by September, it was apparent that no breakthrough would take place. German lines held, but at a high price. German casualties numbered approximately 450,000; together

French and British casualties surpassed 600,000. One million men fell, and the lines remained roughly where they had been at the beginning of 1916 (Keegan 1976).

Stalemate was also the rule at sea, where the British fleet kept the German fleet penned up to the east of Denmark. One attempt to shift the war to the west at Jutland resulted in mixed fortunes for both navies. But the ultimate result was that the German high seas fleet was kept away from Britain.

2.4 1917

In 1917, the balance of power remained unchanged on the Western front. But in the east and in Italy, Germany and Austria were in the ascendant. First came the crisis in Russia. After the February revolution which toppled the Czar and put a provisional government in power, a provisional government vowed to carry on the war. This was its worst mistake, made evident by a catastrophic offensive launched on 1 July 1917. When the Russian offensive failed, a counter-offensive produced a massive Russianretreat which effectively destroyed the Russian army as a fighting force. When the Bolsheviks came to power in October, they saw what had to be done. They sued for peace, and signed an armistice on 3 December 1917. Germany and her allies had won the war on the Eastern front.

In Italy, the Allied cause suffered a series of alarming reverses. At Caporetto, the Italian army collapsed, and fled westward 100 km where a new defensive line was hastily formed. In France, a spring attempt to pierce German lines at the Chemin des Dames failed, and when the French commander Nivelle persisted in trying to do the 'un-doable,' his army mutinied (Smith 1994). Nivelle was replaced by Pétain; the offensive was called off; the anger of frontline troops abated. British attempts to shift the balance of power in Flanders were also unsuccessful, in part through the unhappy coincidence of their attack at Ypres occurring in the wettest summer on record. The British offensive, which ended in the rubble of the village of Passchendaele on 10 November 1917, literally drowned in mud (Prior and Wilson 1996).

The German navy attempted to shift the balance of power by launching unrestricted submarine warfare in February 1917. This was a colossal miscalculation. It did not starve the UK into negotiating an armistice. Despite massive losses to British shipping, domestic food supplies remained adequate (Offer 1991). What was worse from the German point of view, was the provocation posed by naval attacks to the USA, which entered the war in April 1917.

2.5 1918

In January 1918, it would have been difficult for a neutral observer to predict Allied victory by the end of that year. How did it happen? First, the political and economic balance of power shifted with US entry into the war in 1917. Given the time needed to raise and then move an army of two million men across the Atlantic, it was apparent that US troops and US supplies would directly enter the balance of power only in 1918. This gave the Allies a new and deep reservoir of men and materiel which the Central Powers could not match. Thus, it was not US firepower that turned the tide, but US reserves.

This became apparent after German army launched its last major military gamble of the war. On 21 March, 47 divisions of the German army broke through at the point where British and French forces met on the Western front. Finally, the Western front moved. Within a week, German forces had moved 40 miles (65 km) to the west. In April they attacked at Ypres; in May and June to the south, reaching the Marne. But there the offensive broke down. The gamble had failed. The German army had lost one million men in a last, desperate and futile effort to win the war (Keegan 1999).

When German forces saw that they could not win, the balance of power shifted radically towards the Allies. Starting in early August, German troops began to surrender en masse. High Command was getting reports of low morale among troops, who were convinced that further sacrifices were in vain. At this point, unrest on the home front merged with despair over the military situation. The High Command saw that the game was up (Deist 1991). At the same time, Austrian and Bulgarian positions crumbled. Ludendorff urged the Kaiser to turn to the US President, Woodrow Wilson, to broker an Armistice. Ludendorff resigned. The Kaiser fled to the Netherlands, leaving it to others to face a disaster he and his military had engineered. The Armistice came on 11 November 1918. The Great War was over.

3. Consequences

3.1 Demographic

Total casualties and losses as a proportion of those who served passed a threshold beyond previous experience. Wherever the threshold is, the total of roughly nine million dead soldiers (according to varying estimates) is beyond it: this constitutes roughly one in eight of the men who served. Adding statistics on other casualties, it is apparent that roughly 50 percent of the men who served were either prisoners of war, wounded or killed.

The most murderous theatre of operations was the Eastern front, where disease and enemy action described the course of a nineteenth century war waged with twentieth century weapons. Of all Serbs who served in the war, 37 percent were killed; roughly one in four Rumanians, Turks and Bulgarians also perished. On the Western Front, where the war was won and

lost, combat was about half as lethal: German and French losses were about one in six of those who served; British losses were one in eight.

Initially casualties among social elites were higher than among the rest of the population. The longer the war lasted, the greater was the democratization of loss. The reason is that officer casualties were higher than those in the ranks, and the social selection of the officer corps mirrored inequalities in prewar life. Consequently in its initial phases, the higher up in the social scale a man was, the greater were his chances of becoming a casualty of war. By 1917, elites were sufficiently decimated to require the armies to draw junior officers from wider social groups which in their turn suffered disproportionately higher casualties in the last two years of the war.

Among the poor and the underprivileged, the story is different. Prewar deprivation saved the lives of millions of working class men and poor peasants, whose stunted stature and diseases made it impossible for them to pass even the rudimentary standards of medical fitness for military service during the war. In the British case, roughly 35 percent of the men examined for military service were either unfit for combat or unfit to wear a uniform at all (Winter 1985).

Numerically, workers and peasants formed the vast majority of the army of the dead. The same was true of disabled men, visible reminders of the war on street corners and in farming villages throughout the world.

More lethal than combat was a visitation of epidemic influenza, colloquially known as 'the Spanish flu.' In 1918 and 1919 this mutant virus killed young, healthy adults throughout the world. The war may have helped spread the disease, though it did not cause it.

3.2 Political

The Treaty of Paris, the peace settlement ending the Great War, was signed on 28 June 1919, five years to the day after the assassination of Franz Ferdinand in Sarajevo. Its major effects were to ratify the break-up of the Austro–Hungarian and Turkish empires, and to establish a new state system in Europe, wherein Germany was stripped of Alsace and Lorraine in the west and much of East Prussia in the east. Primary

Table 1
Some estimates of military losses among combatant countries in the 1914–18 war

Country	Total killed or died	Total mobilized	Prewar male population (aged 15–49)	Total prewar population	Total killed		
	(in thousands)				per 1,000 mobilized	per 1,000 males (aged 15–49)	per 1,000 people
Britain and Ireland	723	6,147	11,540	45,221	118	63	16
Canada	61	629	2,320	8,100	97	26	8
Australia	60	413	1,370	4,900	145	44	12
New Zealand	16	129	320	1,100	124	50	15
South Africa	7	136	1,700	6,300	51	4	1
India	54	953	82,600	321,800	57	1	0
France	1,327	7,891	9,981	39,600	168	133	34
French colonies	71	449	13,200	52,700	158	5	1
Belgium	38	365	1,924	7,600	104	20	5
Italy	578	5,615	7,767	35,900	103	75	16
Portugal	7	100	1,315	6,100	70	5	1
Greece	26	353	1,235	4,900	73	21	5
Serbia	278	750	1,225	4,900	371	227	57
Rumania	250	1,000	1,900	7,600	250	132	33
Russia	1,811	15,798	40,080	167,000	115	45	11
USA	114	4,273	25,541	98,800	27	4	1
Allied total	5,421	45,001	204,018	812,521	120	27	7
Germany	2,037	13,200	16,316	67,800	154	125	30
Australia–Hungary	1,100	9,000	12,176	58,600	122	90	19
Turkey	804	2,998	5,425	21,700	268	148	37
Bulgaria	88	400	1,100	4,700	220	80	19
Central Powers' total	4,029	25,598	35,017	152,800	157	115	26
Grand total	9,450	70,599	239,035	965,321	134	40	10

Source: Winter (1985).

responsibility for the war, and for making material reparations for it, were laid at the feet of the new German Republic.

Through a League of Nations, international conflict would be regulated, if not resolved. An International Labour Organization was established to help address problems of working class life, a clear response to the Bolshevik revolution. The new Soviet Union was not a party to the peace negotiations or to the settlement. Instead a kind of quarantine was applied to the Soviet Union, and tacit support was given by the victorious powers to modest military intervention in Russia on the side of the anti-Bolshevik forces. This intervention was a total failure.

While the principle of self-determination for all was established, colonial and imperial issues were resolved through bilateral talks among imperial powers. Thus spheres of influence in the Middle East were divided between France and the UK, and contradictory promises were given to Arabs and Jews as to the future postcolonial disposition of the area.

In short, the Great War set the terms of the international order for the rest of the twentieth century (Boemeke et al. 1998). Domestically, the political consequences of war were mixed. Socialist and labor parties had participated in government. Right wing nationalist groups also grew. Among them were fascist parties, whose elevation of the virtues of armed struggle was mixed with a hatred of the political left. The center of politics weakened, evidenced by the waning of support for Liberal parties. The polarized left–right divide took on violent form in Italy, Germany, and throughout Eastern Europe, mixing ideological and territorial conflicts in such a way as to ensure the outbreak of a new and more lethal world war in just 20 years.

The war opened a phase of domestic strife in which the upheaval of war was mixed with revolutionary struggles. Civil war in Russia lasted for three years after the Armistice; armed conflict continued in the new states of Poland and Hungary. The victorious Allies quarantined the new Soviet Union, and began a phase of cold warfare which lasted for 70 years. For this reason alone, some historians see the war as inaugurating the 'short twentieth century,' from 1914–91 (Hobsbawm 1995).

3.3 Economic and Social

The Great War was both the apogee and the beginning of the end of European domination of the world economy. The big loser was the UK, whose export economy suffered from import substitution in countries throughout the world who developed finished goods during the conflict, when British exports were unavailable. Chronic financial instability weakened London as a banking center. The era of free trade was over (Milward 1984).

The losers paid for the war in a host of ways. Reparations were forced on Germany, though she managed to avoid paying most of the bill. Worse was the continuation and acceleration of wartime inflation into the postwar years, wiping out savings and deepening domestic social and political conflict in Germany and Austria (Feldman 1993). This drastic reduction in the value of money had serious effects on the social structure and political outlook of middle class groups, threatened by proletarianization (Kocka 1984).

Domestically, economies of scale were introduced during the war which proved important after it. Hence military conflict helped accelerate the movement away from the small, family firm towards corporate economic life. State interest in and intervention in scientific research and education also expanded during and after war, in the light of evident strategic considerations (Hardach 1987).

After the war, the state continued to play a greater role in economic life than ever before. This was not through the ownership of industry, but through the management of debt. In most industrialized countries, taxation levels doubled over the war decade, and much of it was spent servicing war debt. The USA was the exception here, turning because of the war from a debtor to a creditor nation, though the chronic weakness of countries indebted to it after the war, like the UK, helped turn the Wall Street crash of 1929 into the world depression.

3.4 Cultural

Every combatant country mourned its war dead in public. The outcome is a vast array of war cemeteries near the fields of battle and of war memorials dotted in villages and towns throughout Europe, North America, Africa, Asia and the Antipodes. Annual commemorative ceremonies were (and still are) held on 11 November (Winter 1995) (Table 1).

The monumental character of the war has indelibly marked the literature, poetry, and visual arts of the twentieth century (Cork 1994, Fussell 1977, Hynes 1991). The Great War is the point of reference for much subsequent writing about catastrophe, and has introduced terminology—like shell shock—which has entered colloquial language. In many countries, though not in the former Soviet Union or the USA, its shadow is longer than that of the Second World War.

See also: Contemporary History; Imperialism, History of; Military History; Russian Revolution, The; Second World War, The; War: Anthropological Aspects; War: Causes and Patterns; Warfare in History

Bibliography

Afflerbach H 1994 *Falkenhayn: Politisches Denken und Handeln im Kaiserreich*. Oldenbourg, Munich, Germany

Boemeke M F, Feldman G D, Glaser E (eds.) 1998 *The Treaty of Versailles: a Reassessment after 75 Years*. Cambridge University Press, Cambridge, UK
Cork R 1994 *A Bitter Truth: Avant Garde Art and The Great War*. Yale University Press, New Haven, CT
Deist W 1991 *Militär, Staat und Gesellschaft: Studien zur Preussisch—Deutschen Militärgeschichte*. R. Oldenbourg, Munich, Germany
Feldman G D 1993 *The Great Disorder: Politics, Economics, and Society in the German Inflation, 1914–1924*. Oxford University Press, New York
Fussell P 1977 *The Great War and Modern Memory*. Oxford University Press, London
Hardach G 1987 *The First World War, 1914–1918*. Penguin, Harmondsworth, UK
Hobsbawm E J 1995 *Age of Extremes: The Short Twentieth Century, 1914–91*. Abacus, London
Horne A 1962 *The Price of Glory: Verdun 1916*. St. Martin's Press, New York
Hynes S 1990 *A War Imagined: the First World War and English Culture*. Bodley Head, London
Joll J 1992 *The Origins of the First World War*, 2nd edn. Longman, New York
Jünger E (ed.) 1929 *Storm of Steel: From the Diary of a Storm Troop Officer on the Western Front*. Chatto and Windus, London
Keegan J 1976 *The Face of Battle*. Cape, London
Keegan J 1999 *The First World War*. A Knopf, New York
Kennedy P M 1980 *The Rise of the Anglo–German Antagonism 1860–1914*. Allen and Unwin, London
Kocka J 1984 *Facing Total War: German Society 1914–1918*. Berg, Leamington Spa, UK
Milward A S 1984 *The Economic Effects of the Two World Wars on Britain*, 2nd edn. Macmillan, London
Offer A 1991 *The First World War: An Agrarian Interpretation*. Clarendon Press, Oxford, UK
Prior R, Wilson T 1992 *Command on the Western Front: the Military Career of Sir Henry Rawlinson, 1914–18*. Blackwell, Oxford, UK
Prior R, Wilson T 1996 *Passchendaele: The Untold Story*. Yale University Press, New Haven, CT
Ritter G 1958 *The Schlieffen Plan: Critique of a Myth* [Wilson A, Wilson E trans.]. Praeger, New York
Showalter D E 1991 *Tannenberg, Clash of Empires*. Archon Books, Hamden, CT
Smith L V 1994 *Between Mutiny and Obedience: The Case of the French Fifth Infantry Division during World War I*. Princeton University Press, Princeton, NJ
van Creveld M 1977 *Supplying War*. Cambridge University Press, Cambridge, UK
Winter J M 1985 *The Great War and the British People*. Macmillan, Basingstoke, UK
Winter J M 1995 *Sites of Memory, Sites of Mourning: The Great War in European Cultural History*. Cambridge University Press, Cambridge, UK

J. Winter

Fisher, Irving (1867–1947)

Irving Fisher is widely recognized as one of the greatest American scientific economists. He has been credited with being a primary architect of the 'pillars and arches' of modern economics—to borrow a phrase from J. A. Schumpeter—through his contributions to the theory of capital and interest, to monetary theory, and to the construction of index numbers. He was also an activist in the promotion of economic and social reforms. His reputation suffered in the 1930s when his approach to the discipline was eclipsed by Keynesianism. His accomplishments were then neglected and had to be rediscovered.

1. *Beginnings*

Irving Fisher was born on February 27, 1867 in Saugerties, New York and was the elder son of a Congregational clergyman. His was a home grown talent, nurtured in a single American institution: Yale University. Thus, unlike most American economists reaching professional maturity in the last decades of the nineteenth century, he had no direct exposure to the teachings of the German historical school. His intellectual development was shaped by two forceful members of Yale's faculty: Willard Gibbs, a mathematical physicist, and William Graham Sumner, a political economist who was also a champion of Social Darwinism. Their joint influence was observable in his choice of topic for a Ph.D. dissertation—*Mathematical Investigations in the Theory of Value and Prices*. With the publication of this exercise in general equilibrium theorizing in 1892, Fisher was hailed as a path breaker in mathematical economics. Fisher joined the Yale faculty in 1892 as an instructor in mathematics. In the following year, he married Margaret Hazard, a daughter of a wealthy Rhode Island textile manufacturer; their union was to produce two daughters and a son. In 1895, his appointment at Yale was shifted to the Department of Political Economy, an affiliation he retained until retirement in 1935. In 1898, he was promoted to full professor, but was soon diagnosed thereafter with a potentially fatal case of tuberculosis that idled him for the next three years. Upon recovery, he decided that the life of an ivory-tower academician was no longer sufficient to satisfy him. Indeed he was convinced that scientific professionals had an obligation to engage with the world by promoting economic and social uplift.

2. *Contributions to the Theory of Capital, Income, and Interest*

Fisher had undertaken preliminary inquiries into the nature of capital before the interruption occasioned by illness. He returned to this theme in 1904 when his energies approached full restoration. Results emerged quickly in two major treatises.

In *The Nature of Capital and Income* (1906), Fisher set out a rigorous formulation of key concepts that had formerly been clouded by confusion. Capital (including human capital) was identified as the stock of assets that yield services, whereas income referred to the flow of services that generated satisfactions. He further demonstrated that the value of capital was determined by discounting the future income stream. This finding marked a distinct analytic advance. Though now generally accepted, Fisher's insistence that asset values were oriented toward the future, rather than to historical cost, was controversial when first presented. One feature of the book's argument has continued to be controversial, i.e., his claim that the correct concept of income should identify it with consumption spending and should exclude saving. *The Rate of Interest* (1907) was a logical sequel to *The Nature of Capital and Income*.

The earlier book had established the pivotal role of the interest rate in capitalizing the value of an anticipated income stream. It thus remained to demonstrate how the rate of interest was itself determined. When Fisher first addressed this issue, debate was polarized between two rival positions: one explanation of interest focused on increments in output arising from the use of capital (the technological school); a second emphasized the necessity to reward savers for deferring consumption (the psychological school). Fisher's initial analysis was centered on 'impatience'—or an unwillingness to forgo today's consumption without the prospect of greater consumption possibilities tomorrow—as the principal explanatory factor. When he returned to this topic in *The Theory of Interest* (1930), he argued that the technological and the psychological arguments were complementary, that is, that the interest rate was determined by the interaction between 'impatience to spend income and the opportunity to invest it.' The real interest rate was thus unambiguously a real phenomenon, not a monetary one. This line of analysis would stand in sharp contrast to the monetary theory of interest that Keynes was to offer in his *General Theory of Employment, Interest and Money* (1936). For Fisher, money still mattered because of its impact on the price level. It could thus affect the nominal interest rate, which Fisher treated as the real interest rate plus the inflation rate.

3. *Early Work in Monetary Theory*

The Purchasing Power of Money (1911) was conceived as an exercise in establishing the validity and usefulness of the quantity theory of money, a doctrine that had been politically contaminated in the polemics over 'free silver' in the 1890s. In Fisher's formulation, 'the equation of exchange' was written as $MV + M'V' = PT$. (In this expression, M represented the quantity of cash and V the velocity of its circulation; M' stood for total deposits subject to check and V' for the average velocity of their circulation; P was defined as an average price and T as the volume of trade.) This book was a vigorously monetarist document in which Fisher maintained that changes in the general price level were linked to proportionate changes in the money supply. This was true in the 'normal period,' defined as a state of equilibrium in which the values of T and the V's were constant. His choice of terminology on this point bred misunderstandings: he recognized readily that economic reality was typified by 'transition periods'—when adjustments to disturbances were being worked out—in which the constancies did not hold.

Fisher was not content to offer solely an analytic explanation for variations in the purchasing power of money. Convinced as he was that monetary instability produced distributive injustices by distorting the relative positions of creditors and debtors, he felt obliged to suggest a corrective. His proposed remedy prescribed that changes in the general price level should be offset by variations in the gold content of the dollar (i.e., the 'compensated dollar' plan). But this policy recommendation did not mesh with the analysis presented in *The Purchasing Power of Money*. The 'compensated dollar' scheme rested essentially on a 'commodity theory of money' in which the real value of money was determined by the value of its gold equivalent. This conclusion did not follow from the 'equation of exchange' that Fisher had set out as the basis for his 'quantity theory of money.' Schumpeter (1948) and Patinkin (1993) have suggested that this oddity can be explained because Fisher, the reformer, got the better of Fisher, the scholar.

4. *Analytic Innovations in the 1920s*

Given Fisher's concern for stabilization of the general price level, his interest in perfecting the instrument for measuring its behavior—namely, the index number—was readily comprehensible. He addressed this in *The Making of Index Numbers* (1922), a work which has continued to be influential. Shortly thereafter, he created an Index Number Institute (INI) that provided weekly reports on leading economic indicators, along with commentary on current events. These analyses were distributed nationally via a syndicated news service. At its peak, Fisher's INI column had a potential to reach about seven million readers.

By the mid-1920s, Fisher's studies had convinced him that the stakes in price level instability were higher than he had earlier thought them to be. When writing *The Purchasing Power of Money*, he regarded distributive injustices as the major evil associated with unstable prices. His later analyses indicated that instability of the general price level was responsible for another major problem—fluctuation in the volume of

trade and, with it, variations in aggregate income and employment. This conclusion, he maintained, was supported by statistical evidence indicating that the rate of change in the general price level was highly correlated with variation in the volume of trade. This relationship, however, was not instantaneous.

His empirical investigations suggested a lag of approximately seven months between the turning point in the rate of change in prices and change in aggregate economic activity. In this research, Fisher introduced a novel statistical technique—the distributed lag—which has become a standard tool in econometrics. With its aid, his results appeared to show conclusively that a rising price level stimulated trade, while a falling price level depressed it. It thus followed that the way to eliminate macroeconomic fluctuations was to maintain a stable price level. In that spirit, the Fisher of the mid-1920s wrote that the 'so-called business cycle' was really 'a dance of the dollar' and that there was no reason why such disturbances should be tolerated.

5. *Crusading for Causes*

Fisher, the recovered TB patient, was an indefatigable champion of causes. This was evident in his tireless promotion of monetary reforms: in the first instance, in lobbying for a 'compensated dollar' that promised to stabilize the dollar's purchasing power. In the 1920s, he lobbied vigorously for legislation that would compel the Federal Reserve to maintain price level stability. His views by then had changed on how this should be accomplished. In 1923, with the discovery of the potency of 'open market operations'—a technique for altering commercial bank reserves through the purchase or sale of government securities by the Federal Reserve system—Fisher readily embraced this procedure as the major instrument for central bank control of the money supply. He was a frequent visitor to Washington as a star witness before Congressional committees with jurisdiction over monetary affairs. He also poured considerable energy into mobilizing public opinion behind reform proposals through the creation of organizations such as the Stable Money League.

Fisher's reformist agenda was not restricted to matters directly linked to his expertise as an economist. He was an enthusiast for measures to improve the health and longevity of the population. On his initiative, a group of leading citizens was organized in 1907 to lobby—unsuccessfully as it turned out—for the creation of a Federal Department of Health that would establish national standards for public health practices and monitor their enforcement. He enjoyed greater success in the promotion of a Life Extension Institute to persuade life insurance companies that it would be profitable to require and pay for annual physical examinations for policy holders. To reinforce its activities, he co-authored a first-of-its-kind manual of good health practices, entitled *How to Live*, that went through some 21 editions.

Fisher was one of the nation's staunchest advocates of national prohibition of the consumption and sale of alcoholic beverages. Once the 18th amendment to the Constitution (1920) had been enacted for that purpose, he was a leading voice in arguing against its repeal. Even though enforcement of prohibition left much to be desired, he maintained that this legal ban had still contributed positively to American economic expansion in the 'new era' of the 1920s through its effect on improved worker efficiency and on reduced workplace absenteeism. This line of thinking was responsible, in part, for his disastrous misreading of economic conditions in the autumn of 1929 when he predicted that the American economy was on the threshold of permanent prosperity. In the 1920s, he was also an inveterate campaigner for an international organization to promote world peace and for membership of the US in the League of Nations. He was a prominent figure as well in the eugenics movement that he hoped would encourage state governments to enact legislation permitting sterilization of the mentally defective and the criminally insane.

Fisher contributed significantly from his own pocket to support these causes. His ability to do so was enhanced by the sale of one of his inventions—an Index Visible filing system—to Rand–Kardex (shortly to become Remington Rand) for a handsome block of stock and for a position on the board of directors. When informing his son about this turn of fortune, he wrote: 'This added income will enable me to further the four chief causes which we have at heart, the abolition of war, disease, degeneracy, and instability of money (Fisher to Irving Norton Fisher, June 17, 1925).

6. *Coping with the Disaster of Depression in the 1930s*

Fisher was totally caught off-guard by the collapse of the stock market in the autumn of 1929 and by the events that followed. His personal financial circumstances were left in shambles: he managed to avert personal bankruptcy only by borrowing heavily from his wife's unmarried sister. The Great Depression was also a professional embarrassment. It did stimulate him to rethink where his earlier line of analysis had failed him. The result—reported in *Booms and Depressions* (1932)—was the debt-deflation theory of Great Depressions. The problem, as he then saw it, stemmed from a situation of overindebtedness in the late 1920s. Once a break in the price level had occurred, the real burden of debt increased. This led to distress selling, to contraction in deposit currency, and to slackening in

the velocity of monetary circulation. Deflation—far from being self-correcting—simply bred more deflation. The remedy was implicit in the diagnosis: 'reflation' of the price level to its elevation at the time when most outstanding debts were contracted. When that point had been reached, the general price level should be stabilized.

To accomplish this, Fisher urged a number of proposals on the Roosevelt administration. For example, he endorsed 'stamped scrip'—a scheme which called for a special currency issue that would lose 2 percent of its value each week unless a two-cent stamp were affixed to it. This device (which he had borrowed from Gesell) was intended to accelerate the velocity of monetary circulation. Though the idea was not original with him, he later put his full weight behind advocacy of a plan that would require commercial banks to hold '100 percent reserves' against demand deposits, thereby eliminating their capacity to create (or destroy) check-book money by extending (or calling) loans. The money supply would instead be controlled by a public authority (e.g., a Currency Commission) that would expand the money supply as needed to achieve a prescribed 'reflation' target and thereafter to stabilize. Fisher's conception of a 'monetary rule' was tied to outcomes, as opposed to the 'rule' recommended in Milton Friedman's brand of monetarism that stipulates a fixed rate of growth in the money supply.

Fisher's thinking about recovery from the Great Depression always had a monetary orientation. He was unsympathetic toward fiscal stimulants—such as Keynesian-style programs of public works—and he made no secret about that. He was also unsympathetic to the analytic message of Keynes's *General Theory*. Even so, he and Keynes came close to convergence in their recommendations for tax policy in wartime: both championed a tax targeting consumption. The analytic props to their conclusions, however, had nothing in common. Fisher's views—which he spelled out in his final book, *Constructive Income Taxation* (1942)—rested on the analysis he had worked out in 1906 in *The Nature of Capital and Income*. This argument maintained that income should properly be defined as expenditure on consumption: thus, a sound tax system, in all circumstances, should take aim at consumption spending. Keynes's position, on the other hand, was informed by his conception of the management of the components of aggregate demand, as formulated in *The General Theory*.

Despite the fact that the Roosevelt administration took little heed of his unsolicited advice, Fisher still believed that something hopeful could be salvaged from the wreckage of depression in the form of lasting improvement in scientific investigation in economics. It had long been his aspiration to build an international organization to bring economists interested in linking economic theory with mathematics and statistics into closer collaboration. Circumstances of the early 1930s permitted this dream to become a reality. The presence at Yale in 1930 of Ragnar Frisch of the University of Oslo made it easier for three who shared the dream—Fisher, Frisch, and C. F. Roos—to lay the groundwork for the Econometric Society. The obvious shortcomings of economic forecasting in the early depression years—and an apparent need for improvement—convinced a well-to-do businessman to underwrite this new organization.

7. *A Word about the Legacy*

Fisher was accorded numerous honors by professional colleagues: among them, election to the presidencies of the American Economic Association, the American Statistical Association, the Econometric Society, the American Association for Labor Legislation. But he lacked a committed cadre of disciples: there was no 'Yale circus' to match Keynes's 'Cambridge circus.' In the last decade of his life, his approach was largely overshadowed by Keynesianism. To younger members of the profession, he appeared to be out-of-step with the times, with the consequence that his work tended to be ignored.

With the waning of the Keynesian ascendancy, Fisher's achievements have been revisited. His pioneering analyses of capital and interest have now approached the status of an orthodoxy in the graduate training of prospective professionals. A 'rediscovery' in the 1970s documented that an article he wrote in 1926—entitled *A Statistical Relation between Unemployment and Price Changes*—anticipated the essential insight of the Phillips Curve by more than three decades. It is interesting to speculate on how the principals in an intense controversy that divided the profession in the 1960s and 1970s would have behaved had they known that Fisher had identified the 'reswitching problem' in 1907.

In the 1990s, some of Fisher's practical proposals have enjoyed a renewed lease on life. In 1997, the US Treasury issued a 'stabilized bond,' the return on which would be adjusted to maintain constancy in its purchasing power. Fisher had advocated the creation of this type of financial instrument some seven decades earlier. In 1995, the US Department of Commerce resurrected another Fisher conception, dating from 1922, when adopting his formula for an 'ideal index number' to calculate the GNP deflator.

Fisher died on April 29, 1947 in New York City.

See also: Economics, History of; Health Economics; Keynes, John Maynard (1883–1946); Monetary Policy

Bibliography

Allen R L 1993 *Irving Fisher: A Biography*. Blackwell, Cambridge, MA

Fisher I 1997 *The Works of Irving Fisher*, 14 Vols. Pickering and Chatto, London
Patinkin D 1993 Irving Fisher and his compensated dollar plan. *Federal Reserve Bank of Richmond Economic Quarterly* **79**: 1–33
Schumpeter J A 1948 Irving Fisher's econometrics. *Econometrica* **16**: 219–31

W. J. Barber

Fisher, Ronald A (1890–1962)

Sir Ronald Fisher, the father of modern statistics, was for most of his working life a professor of genetics, first in London and then at Cambridge. He made lasting contributions to mathematical and evolutionary genetics as well as to statistical theory. In statistics his contributions stand alongside those of his great predecessors Gauss and Laplace, while in evolutionary theory he is acknowledged as the greatest of Darwin's successors.

Fisher was born in London on February 17, 1890, the son of George Fisher, a fine-art auctioneer, and his wife Katie. His twin brother was stillborn. At Harrow School, which he entered in 1904 as a scholar, he distinguished himself in mathematics despite being handicapped by poor eyesight which prevented him working by artificial light. His teachers used to instruct him by ear, and Fisher developed a remarkable capacity for pursuing complex mathematical arguments in his head. This manifested itself later in life in an ability to reach a conclusion whilst forgetting the argument, to handle complex geometrical trains of thought, and to develop and report essentially mathematical arguments in English (only for students to have to reconstruct the mathematics later). Fisher's early interest in natural history was reflected in the books chosen for special school prizes at Harrow, culminating in his last year in the choice of the complete works of Charles Darwin in 13 volumes. Fisher entered Gonville and Caius College, Cambridge, as a scholar in 1909, graduating BA in mathematics in 1912. At college he instigated the formation of a Cambridge University Eugenics Society through which he met Major Leonard Darwin, Charles's fourth son and president of the Eugenics Education Society of London, who was to become his mentor and friend. He spent a postgraduate year in the Cavendish Laboratory, Cambridge, studying the theory of errors under F. J. M. Stratton and statistical mechanics and quantum theory under J. H. Jeans (later Sir James Jeans).

Prevented from entering war service in 1914 by his poor eyesight, Fisher taught physics and mathematics in schools for the duration of the war and in 1919 was appointed Statistician to Rothamsted Experimental Station, an agricultural station at Harpenden north of London. In 1933 he was elected to succeed Karl Pearson as Galton Professor of Eugenics (i.e., of Human Genetics, as it later became) at University College, London, and in 1943 he was elected Arthur Balfour Professor of Genetics at Cambridge and a Fellow of Gonville and Caius College, succeeding R. C. Punnett. After he retired in 1957, Fisher traveled widely, spending his last few years in Adelaide, Australia, as an honorary research fellow of the C. S. I. R. O. Division of Mathematical Statistics. He died there of a postoperative embolism on July 29, 1962. His ashes lie under a plaque in a side aisle of Adelaide Cathedral.

Fisher married Ruth Eileen Guinness in 1917 and they had two sons and six daughters, and a baby girl who died young. He was elected a Fellow of the Royal Society of London in 1929 (as a mathematician) and was created Knight Bachelor by Queen Elizabeth II in 1952 for services to science. He was the founding President of the Biometric Society (now the International Biometric Society) in 1947, and served as President of the Royal Statistical Society (of the UK), of the International Statistical Institute, of the Genetical Society of Great Britain, and of his Cambridge college, Gonville and Caius. He received many honorary degrees and accepted the honorary membership of many academies at home and abroad, and was awarded all the principal medals of the Royal Society, the Royal (1938), Darwin (1948), and Copley (1956).

Fisher's first publication, whilst still an undergraduate, proposed the method of maximum likelihood for statistical estimation (though not in those words) and before graduating he had already given a paper remarking on the surprisingly good fit of Mendel's data. In 1916, without any academic appointment but encouraged by Leonard Darwin, he completed the founding paper of biometrical genetics 'The correlation between relatives on the supposition of Mendelian inheritance' (1918), eventually published in 1918 after it had been rejected for the *Philosophical Transactions of the Royal Society* on the advice of the referees, Karl Pearson and R. C. Punnett (both of whom he later succeeded, as he was fond of remarking). In this paper, Fisher introduced the word 'variance' and established the basis of the analysis of variance.

From his post of statistician at Rothamsted, Fisher made advances which revolutionized statistics, whilst his advances in genetics and evolution were hardly less pathbreaking. In a single publication in 1922 he proved that heterozygotic advantage gives rise to a stable gene-frequency equilibrium, introduced the first stochastic model into genetics (a branching process), and initiated the study of gene-frequency distributions by means of the diffusion approximation, and in another paper he applied the method of maximum likelihood for the first time, to the estimation of linkage.

Also in 1922, building on his method of maximum likelihood, Fisher published 'On the mathematical

foundations of theoretical statistics' in the *Philosophical Transactions* (1922a), a revolution in statistical thought. This time the Royal Society's referees, A. S. Eddington and G. U. Yule, were more complimentary. Amongst the new words and phrases Fisher introduced in the paper in order to clarify the basis of his approaches to statistical inference are likelihood, consistency, degrees of freedom, efficiency, sufficiency, information, significance level, test of significance, sampling distribution, percentage point, statistic, and others of a more technical nature still. Information in Fisher's statistical sense (1925) antedated Shannon's (1948) use of the word for a somewhat different purpose: Fisher information measures the content of a message, whereas Shannon's measures the ability of a channel to transmit a message, independently of its meaning. In 1930, Fisher published the first satisfactory account of what J. Jerey Neyman was later to call the 'confidence' property of interval estimation.

In distribution theory Fisher built on his 1915 paper deriving the exact distribution of the correlation coefficient (for whose solution he introduced the concept of a sample-space) and used his considerable mathematical ability to solve many further distributional problems, including that of the multiple correlation coefficient and the variance ratio (which G. W. Snedecor of Iowa State University, where Fisher was visiting professor in 1931 and 1936, called 'F' in his honor). His work at Rothamsted on agricultural statistics soon led to the creation of the science of experimental design with the appreciation of the fundamental role of randomization, the development of the principles of blocking and confounding, and the introduction of Latin squares and factorial analysis and similar designs chosen for their statistical efficiency. The methods which Fisher invented or improved were brought together in 1925 in his book *Statistical Methods for Research Workers* (1925b), which he updated continually over the next 23 years until the thirteenth edition was published in 1958 (a posthumous fourteenth edition was prepared from notes written before his death).

In addition to the statistical work associated with his employment at Rothamsted, Fisher pursued his interest in evolutionary theory, guided and encouraged by Leonard Darwin. Putting his deep knowledge of Mendelian theory and Charles Darwin's writings to good use he built on the success of his earlier investigations of 1918 and 1922 to write *The Genetical Theory of Natural Selection* (1930), published in 1930. This book is now seen as the chief document of the 'Neo-Darwinian revolution,' that revival of Darwinism in the 1930s and 1940s from the neglect it had suffered in the early years of the century following the rediscovery of Mendel's paper in 1900. The book contains a wealth of new evolutionary arguments, from the fundamental theorem of natural selection to ideas about sexual selection, inclusive fitness and parental expenditure. More than any other work *The Genetical Theory* established a firm basis for the modern view that evolution by natural selection is primarily a within-species phenomenon.

Taking up his appointment as Galton Professor of Eugenics at University College London in 1933, Fisher's pace did not slacken. Now able to pursue experimental work, which previously he had only undertaken in a private capacity, he demonstrated his ability as a natural historian. Experimental organisms included mice, poultry, and the purple loosestrife, and even dogs (under the auspices of the Genetical Society). In a long collaboration with E. B. Ford of Oxford he studied natural populations for the purpose of estimating the power of natural selection, thus contributing to the foundation of a school of ecological genetics. But it is in human genetics that he made his most lasting contribution of this period. In 1935 he secured funds from the Rockefeller Foundation to establish a Blood-Group Serum Unit at the Galton Laboratory with the express purpose of initiating the construction of a linkage map of man, for he had already seen the connection between 'Linkage studies and the prognosis of hereditary ailments' (to use the title of his lecture to the International Congress on Life Assurance Medicine in that year). Here is to be found the intellectual origin of the Human Genome Project. At the same time Fisher, with J. B. S. Haldane and L. S. Penrose, was advancing the special statistical theory required in the estimation of human linkage.

Fisher (1935) also saw the publication of *The Design of Experiments* in which his pioneering work in that subject at Rothamsted was expounded fully for the benefit of workers in agriculture and many other fields. Like *Statistical Methods for Research Workers*, the book was an outstanding success, running to seven editions by 1960 (an eighth was later prepared from notes left by Fisher). It opened with a famous description, often quoted, of an experimental design for the problem of testing the assertion of a lady who maintained that she could tell by tasting a cup of tea whether the milk or the tea had been the first to be poured into the cup. In the course of this discussion Fisher coined the phrase 'null hypothesis' for the hypothesis under test.

In 1936, Fisher subjected Mendel's published data on segregation in the garden pea to close statistical scrutiny, concluding that in some respects the data seemed to follow too closely the expectations on Mendel's genetic hypothesis, a discovery that worried him greatly, and left him speculating that perhaps Mendel's assistant, knowing full well what his master hoped to find, was tempted to adjust some of the counts accordingly.

In 1943, Fisher moved to Cambridge as Arthur Balfour Professor of Genetics, where he was reunited with his colleagues from the Blood-Group Unit which had been evacuated there during the war. An immediate consequence was his brilliant solution of the Rhesus blood-group puzzle, involving three closely

linked loci which between them explained the array of serological reactions which to everyone else had appeared chaotic: Fisher did for Rhesus what Mendel did for round and wrinkled.

After the second world war ended in 1945 Fisher attempted to establish bacterial genetics in his Cambridge department and to retain for Cambridge the Blood-Group Unit, but without success. Work in his small department revolved around linkage in the mouse and studies on purple loosestrife, wood sorrel, and primroses, always with a strong background of mathematics and statistics. His *Theory of Inbreeding* (1949) was published in 1949, and in 1950 he published the first paper applying a computer to a biological problem.

In the subject of statistics, developments in the USA during and after the war had tended to obscure the great advances which Fisher had made in the subject, and he found himself having to defend many of his views against critics brought up in the more mathematical and operational climate then dominant. His repeated insistence that statistical inference was a mode of reasoning separate from, though related to, approaches based on decision theory and other behavioral considerations won him few friends, and his continued (if somewhat puzzled) defense of his troublesome 1930 concept of fiducial probability was cited as evidence that his judgment was fallible. He defended his position in a fourth book, *Statistical Methods and Scientific Inference* (1956), published in 1956 with a second edition in 1959 and a posthumous edition with many additions and clarifications. In this work he particularly emphasized the direct role of likelihood in statistical inference, in addition to the method of maximum likelihood, and continued his life-long criticism of the improper use, as he saw it, of prior probabilities and Bayes's theorem in statistical inference, a methodology previously called 'inverse probability' for which he had himself coined the word 'Bayesian' in 1950.

As well as writing his books, Fisher collaborated with Frank Yates, also of Rothamsted, to produce the most widely-used compilation of statistical tables (before computers rendered such volumes redundant). First published in 1938, *Statistical Tables for Biological, Agricultural and Medical Research* (1938) went through many editions.

In retirement Fisher continued his academic work, but also became embroiled in a controversy as to the extent to which the published data supported the hypothesis, then novel, that cigarette-smoking was a cause of lung cancer. Perhaps not fully appreciating the strong feelings which the issue was to generate, he focused on the finer points of the statistical analysis which interested him, on the need always to disentangle causation and correlation, and on the fact that the evidence did not seem to him to disallow the possibility that genetical variability might be a contributory factor in the observed association.

Fisher's papers number nearly 300, and he also wrote many reviews, particularly in *The Eugenics Review* between 1915 and 1935, and letters to journals. He was an accomplished formal lecturer as may be seen from his many presidential and similar addresses, and an occasional broadcaster on scientific topics. His interests were wide, as may be instanced from two examples from his publications: in 1934 he discovered and published the idea of a randomized or 'mixed' strategy independently of von Neumann, and in 1958 he published the first suggestion that the theory of games might be applied profitably to problems in evolutionary biology, the possibility having been drawn to his attention by Dr. Cavalli (later Professor L. L. Cavalli-Sforza).

Small of stature, with thick glasses and a beard, Fisher did not suffer fools gladly. He was a skilled controversialist in conversation, but his quick temper sometimes rendered further discussion impossible. His contemporaries divided rather cleanly into those who regarded him with awe and affection and gratitude for the generosity with which he offered his ideas to them, and those who found him tetchy, difficult, and remote. Especially as a professor at Cambridge, he showed great interest in the few students who passed through his small department, and he always enjoyed the company of young people. Of the many thumb-nail sketches which his greatness has inspired, perhaps the following comment on Fisher by the Cambridge cosmologist Sir Fred Hoyle contains the closest likeness in the smallest span: 'I am genuinely sorry for scientists of the younger generation who never knew Fisher personally. So long as you avoided a handful of subjects like inverse probability that would turn Fisher in the briefest possible moment from extreme urbanity into a boiling cauldron of wrath, you got by with little worse than a thick head from the port which he, like the Cambridge mathematician J. E. Littlewood, loved to drink in the evening. And on the credit side you gained a cherished memory of English spoken in a Shakespearean style and delivered in the manner of a Spanish grandee' (Hoyle 1999).

Fisher was one of the great intellects of the twentieth century. His contributions to evolutionary theory and mathematical genetics ensure him a place in the Darwinian pantheon whilst he is the undisputed father of modern statistical theory.

The statistical methodology he developed has had a far-reaching impact in every field of science and social science. Whether one considers the mathematical theory of distributions, the design and analysis of experiments, the theory of estimation, the theory of significance testing, or the logic of statistical inference, Fisher's is the primary influence. As a teacher of research workers through his statistical books he had no peer.

In genetics he was the principal architect of the application of statistics, much of his theory having been inspired by practical problems, especially in

human genetics. With J. B. S. Haldane and others he initiated the study of linkage in man specifically for the purpose of throwing light on hereditary disease risks, thus foreshadowing the human genome project, and he encouraged the study of associations between blood groups and disease. His studies on natural and artificial populations of both plants and animals would alone have given him an honored place in the development of genetics in the twentieth century. His '1918 paper' is accepted as the starting point for the entire development of biometrical genetics, with its ramifications in plant and animal breeding.

In evolutionary biology, his book *The Genetical Theory of Natural Selection* (1930) hides a brilliant series of evolutionary insights behind a density of writing that requires more concentration than most readers have been prepared to give it, but is now acclaimed as the cornerstone of the neo-Darwinian revival of the mid-twentieth century.

See also: Distributions, Statistical: Approximations; Estimation: Point and Interval; Hotelling, Harold (1895–1973); Natural Selection; Significance, Tests of; Statistical Methods, History of: Post-1900

Bibliography

Bennett J H (ed.) 1971–74 *Collected Papers of R. A. Fisher*, 5 Vols. University of Adelaide, Adelaide, Australia

Bennett J H (ed.) 1983 *Natural Selection, Heredity, and Eugenics, Including Selected Correspondence of R. A. Fisher with Leonard Darwin and Others*. Clarendon Press, Oxford, UK

Bennett J H (ed.) 1990 *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Clarendon Press, Oxford, UK

Box J F 1978 *R. A. Fisher: The Life of a Scientist*. Wiley, New York

Fisher R A 1918 The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**: 399–433

Fisher R A 1922a On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A* **222**: 309–68

Fisher R A 1922b On the dominance ratio. *Proceedings of the Royal Society of Edinburgh* **42**: 321–41

Fisher R A 1925a Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**: 700–25

Fisher R A 1925b *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh

Fisher R A 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK

Fisher R A 1935 *The Design of Experiments*. Oliver and Boyd, Edinburgh

Fisher R A 1949 *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh

Fisher R A 1950 *Contributions to Mathematical Statistics*. Wiley, New York

Fisher R A 1956 *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh

Fisher R A, Yates F 1938 *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver and Boyd, Edinburgh

Hoyle F 1999 *Mathematics of Evolution*. Acorn Enterprises, Memphis, TN

A. W. F. Edwards

Fishing Societies

The theme of this article is the place of fishing activities and coastal economies in social scientific thought and some of the major issues developed in the relevant literature; the category of fishing, natural and social models of fishing activities and coastal communities, gender relations, ecological knowledge, and resource management.

1. The Category of 'Fishing'

According to some of the evolutionary theories of the nineteenth century, fishing represented a separate and early stage in the history of humans. Such an idea was underlined in many contemporary accounts of particular groups of people largely dependent on fish. The evidence on the role of early fishing, however, is fairly limited. Some scholars suggest that, from the point of view of early humans, aquatic environments must have been both hostile and an inadequate source of food. Others suggest that coastal zones must have been quite attractive, providing a worthwhile challenge, and that people may have exploited them for a long time without leaving modern groundlings any evidence. While the archaeological evidence does not indicate that fishing predated the hunting of terrestrial animals, nonetheless among hunter-gatherers fishing may have played quite an important role. The abundance of resources in coastal zones may, indeed, have provided an opportunity for the development of complex civilizations. There is evidence for a transformation of hunter-gatherer social relations in coastal regimes although many of the important issues involved are far from settled (see *Hunting and Gathering Societies in Anthropology*). Rather than seeing aquatic resources themselves as determinants of complexity, however, one should regard coastal niches as just one possible avenue for intensification (Kelly 1992). With the advent of fieldwork and participant observation, evolutionary speculations tended to be replaced by descriptive accounts. The category of 'fishing' became much less visible than before—sometimes being subsumed under the label 'hunting and gathering.'

While the category of fishing may seem straightforward, it need not be taken for granted. The distinctiveness of fishing activities, it has been argued (Hewes 1948), has two aspects. First, objects behave in a particular manner while in an aquatic substance, due to special conditions of buoyancy, turbulence, solubility, and refraction of light. Second, fishers and their prey occupy different media. An emphasis on differ-

ences in habitat yields a definition of fishing as 'that category of human activity which is connected with the capture or gathering, of animals (or plants) which regularly dwell in the water' (Hewes 1948). The concept of fishing as a particular kind of hunting which happens to yield fish is one element of a common anthropological scheme for classifying types of subsistence: gathering, collecting, hunting (including trapping), husbandry (including fish farming), and plant cultivation. Much like medieval Europeans often distinguished between fishing, fowling, and hunting, modern anthropology tends to operate with three concepts of foraging—fishing, gathering, and hunting. For some scholars, fishing is a broad category that includes shell fishing and the pursuit of large aquatic animals. Others have argued to the contrary that the pursuit of large aquatic animals more properly is classified as hunting and that shell fishing should be classified as gathering.

While some linguistic evidence suggests that fishing and hunting are perceived as similar strategies, differing only with respect to the focus of subsistence efforts, indigenous terminologies do not distinguish necessarily between hunting and other subsistence activities, including trapping, collecting, and gathering. Also, the classification of aquatic organisms, including 'fish,' varies from one society to another. Although in many societies, fish are classified according to an elaborate taxonomy, depending on their shape and usefulness to humans, the category of 'fish' represents one of the largest and most diverse categories found in most environments; at times fish terms are extended to all creatures inhabiting aquatic environments, including turtles, frogs, crocodiles, crustacea, and fish-shaped mammals. In some languages, there is no term for fish; fish are included in a heterogeneous category along with snakes, worms, lizards, and other elongated creatures.

2. *Natural and Social Models*

Many approaches to fisheries, both of the present and the past, suggest a 'natural' model of fishing, emphasizing material context and ecological relations. Social scientists have often drawn attention to the fact that fishing takes place in a relatively uncertain environment (see Pálsson 1991). People who adapt to earning a living by exploiting marine resources, it is argued, often seem to manage their lives in similar ways. Thus, crew organization tends to be flexible and based more on voluntary ties than kinship obligations, to ensure cooperation and the right combination of skills, to reduce uncertainty, and spread the risks of production. Fishing societies, as a result, seem to have a range of characteristics in common due to the fact that their members have to adapt to corresponding environments and cope with identical problems. Such approaches are reminiscent of the method of cultural ecology, according to which social life is rather mechanistically adapted to the material world.

In line with the assumptions of the natural model of fishing, there has been a strong tendency to look for parallels between trawling ('industrial hunting') and small-scale fishing on the basis of a common link to water. While focusing on the boundary between land and water is helpful for drawing contrasts between economic and social systems which are organized in similar ways—between, say, aquatic and terrestrial industrial systems—much recent research has emphasized that a unitary category of fishing diverts attention from important issues, in particular the social relations in which production is necessarily to be found. Although there are real empirical differences between fishing and other modes of subsistence, the use of such differences gives misplaced importance to ecology.

One way to understand similarities and differences among fishing systems, to emphasize the social context of production, is to distinguish between societies in terms of mode of circulation—the motivation of the producers and the destination of the products (see *Mode of Production*). Production may be primarily for exchange or primarily for use. Many fisheries conflicts derive from the different rationales of production of the household economy of local small-scale fishing and the market economy of industrial fleets (McGoodwin 1990). Another way to distinguish fishing systems is to underline different forms of managing access to fish. While knowledge of ecology and fishing techniques is important for the understanding of different forms of managing access, one should not ignore the social space in which they occur. The informal exclusion of outsiders by means of secrecy and formal ownership of local territories should not be regarded as functionally equivalent devices. Differences among fishing economies in terms of modes of circulation of products and access to resources insufficiently represent the variety of production systems there is. On the other hand, they help to illustrate the fundamental point that fisheries are embedded in social life. Any study of what happens on board fishing vessels must consider variations in fishing systems and the larger contexts of fisheries and fishing communities.

3. *Gender Roles and Identities*

In removing 'fishing' from the context of social relations, social scientists have often failed to recognize the importance of the relations of men and women and the significant economic role of women in fishing economies. The issue of gender tends to be suppressed or distorted. Nevertheless, in recent years an extensive literature on the role of gender in fishing and coastal communities has been developed (see, for instance, Nadel-Klein and Davis 1988). In fishing communities, it has sometimes been suggested, one can expect a

sharp division of labor along gender lines in that men do the fishing while women are responsible for the processing of the catch. A rigid sexual division of labor should not, however, be regarded as inevitable. In many cases women's participation in extraction is no less important than that of men. The ethnography of fishing is biased with respect to gender. This observer's bias is partly a reflection of the popular Western image of fishing as a particularly 'manly' activity. Thus, the pursuit of small aquatic animals has sometimes been regarded as mere 'gathering' or 'collecting' if performed by women, but as 'fishing' if performed by men.

Many indigenous accounts of fishing economies are no less gender-specific than observers' accounts, treating women's labor as not only different but also somehow secondary to that of men. While the devaluation of women's work is not restricted to market economies, the change from a subsistence economy to markets often has important implications for the social construction of gender. The 'domestic' labor of women tends to be presented as private labor, on the grounds that it takes place internal to the household and that its consumption products are 'perishable' use values, requiring immediate use. Conversely, only the nondomestic work of men is regarded as socialized since it alone involves cooperation between households, the production of exchange values and durable commodities. Domestic laborers are presented as natural beings outside the social life of the market economy. And women only become social or visible insiders to the extent that they engage in nondomestic production, outside the sphere of reproduction. Women, then, tend to be presented as unproductive (see Cole 1991). But just as extraction is a social activity whether or not it takes place in a group or by isolated individuals, labor is necessarily a social activity whether or not it takes place in a domestic or non-domestic context. The household is not a private place devoid of social content but a social institution constituted by a particular set of relations—the social relations of domestic production involved in the creation of use values, particularly cooking and child care (see *Gender and Feminist Studies in Anthropology*).

4. *Ecological Knowledge*

Since fish are relatively incapable of coping with human predators, their chances of using evasive tactics are small. The difference in medium, however, has contradictory implications and some observers have emphasized the disadvantages to fishers, pointing out that if humans are invisible to the fish so are fish to humans. Accordingly, it is argued, knowledge of marine organisms is likely to be relatively restricted. Nevertheless, as several ethnographic accounts have shown, in fishing communities people accumulate a vast amount of knowledge of fish and aquatic habitat. The social distribution and nature of ecological knowledge and its importance for resource management, on the other hand, are matters of some debate.

In fisheries generally, catches vary from one boat to another. Accounts of and interest in such differences vary from one context to another, depending on the nature of the social system involved and the nature of the fishery. Some explanations emphasize differences in luck, others underline personal differences in fishing skill. Some anthropologists have attempted to statistically measure the personal contribution of the leaders of fishing operations, the 'skipper effect' (see Pálsson and Durrenberger 1990). It is difficult, however, to separate the factors involved and recent studies of fishing success have drawn attention to the continuity of the social world to which the skipper belongs, focusing on the importance of the crew for decision-making and, by extension, relative fishing success. In a very real sense, the practical skills of fishing skippers are the result of mutual enskillment, the collective property of a community of practice (Pálsson 1994). Not only has there been a growing theoretical interest in the practical learning of fishers and the general divide between indigenous and scientific knowledge, but fisheries science itself is increasingly seen as an embedded enterprise (Smith 1996). Some scholars argue that learning, whether in science or every-day life, primarily takes place with the internalization of a cognitive scheme (see *Cognitive Anthropology*). Others have emphasized that learning involves the whole person interacting with the social and natural environment (Lave 1990; see *Apprenticeship: Anthropological Aspects*).

In many fishing societies there is little attempt to draw upon the knowledge that practitioners have achieved in the course of their work, for the purpose of marine ecological research. Analyses of human-environmental relations are the privileged task of people formally trained in public institutions. There are important signs of change, however. A greater reliance on practical knowledge, it is sometimes argued, is a practical necessity, given the randomness and complexity of many marine ecosystems. Multispecies marine ecosystems, in particular, are highly unpredictable, with constant fluctuations in interactions among species and between species and their habitat (Acheson and Wilson 1996). This does not mean that governance is impossible; it suggests, however, increasing reliance on a finer spatial and temporal scale of observation and engagement—a scale that only the skilful practitioner is able to apply. There are good grounds, therefore, for exploring how fishers acquire knowledge about the ecosystem within which they operate, how their knowledge differs from that of professional biologists, and to what extent the former could be brought more systematically into the process of management for the purpose of ensuring sustainability.

5. Resource Management

Many fisheries are currently in a state of crisis. Increasingly social scientists are addressing the problems involved, both internationally and in particular ethnographic contexts (McCay and Acheson 1987). Two main positions can be identified, the market approach which gained momentum in the early 1970s and the approach of co-management which has received increased attention since the 1980s. Comparative studies of the implications and consequences of different management regimes represent an important and growing forum for social research.

The market approach assumes that the current crisis in fishing is due to the open-access nature of fisheries; in the absence of both government control and the institution of private property, it is assumed, people will inevitably drive the ecosystem beyond its sustainable limits. The paradigm further suggests that only the 'hidden' forces of the market can ensure efficiency and sustainable use of resources. Finally, the issues of ethics and equity typically are presented as irrelevant theoretical distractions. In several fisheries in different parts of the world, fishing stocks are being turned into private property. First, the resource is appropriated by regional or national authorities and later on the total allowable catch (TAC) for a season is divided into quota holdings and allocated to producers, often the owners of boats. At a still later stage, the temporary privileges of quota holders are turned into a marketable commodity, individual transferable quotas (ITQs).

In some fisheries the crew's contribution and its relevance for quota allocation has been a pressing political issue. Since fishing quotas are often allocated with reference to 'fishing history,' fishers, boat owners, and managers debate on what exactly makes fishing history, focusing on the relative importance of the vessel itself and its technology (favoring the owners of capital) and the crew (favoring labor power, in particular that of the skipper). A related contested issue involves the practical and theoretical merits of effort quotas (number of fishing trips) vis-à-vis catch quotas (size of catch).

The other main approach to fisheries management, co-management (Jentoft 1989), emphasizes democratic decision making and the importance of practical knowledge. A number of anthropological studies have shown that in many fishing societies people have developed indigenous means of regulating access to fishing grounds. In some cases, local groups of users successfully control the reproduction of renewable resources without external intervention, effectively co-managing local resources. In some parts of the world, especially Asia and the Pacific, co-management has a long tradition (Ruddle and Johannes 1985). Recently, such examples have attracted much attention among both researchers and managers, informed by renewed interest in 'indigenous' or 'traditional' ecological knowledge on the international scene (Dyer and McGoodwin 1994) as well as growing doubts about the modernist concept of 'resource management,' with its assumptions of human mastery, control, and predictability.

Traditionally, anthropologists have focused on isolated communities. Modern anthropology, however, provides useful perspectives on cultural flows and the link between micro and macro levels. Recently, anthropological students of fishing societies have increasingly moved from the narrow context of face-to-face encounters and local communities to the nation state and beyond (LiPuma and Meltzoff 1994). Thus, several anthropological studies have drawn attention to the ways in which local and professional identities are constructed within larger contexts.

Bibliography

Acheson J M, Wilson J A 1996 Order out of chaos: the case for parametric fisheries management. *American Anthropologist* **98**: 579–94

Cole S 1991 *Women of the Praia: Work and Lives in a Portuguese Coastal Community*. Princeton University Press, Princeton, NJ

Dyer C L, McGoodwin J R (eds.) 1994 *Folk Management in the World's Fisheries: Lessons for Modern Fisheries Management*. University Press of Colorado, Niwot, CO

Hewes G 1948 The rubric 'fishing and fisheries.' *American Anthropologist* **50**: 238–46

Jentoft S 1989 Fisheries co-management: delegating government responsibility to fishermen's organizations. *Marine Policy* **13**(2): 137–54

Kelly R L 1992 Mobility/sedentism: concepts, archaeological measures, and effects. *Annual Reviews of Anthropology* **21**: 43–66

Lave J 1990 The culture of acquisition and the practice of understanding. In: Stigler J W, Shweder R A, Herdt G (eds.) *Cultural Psychology: Essays on Comparative Human Development*. Cambridge University Press, Cambridge, UK

LiPuma E, Meltzoff S K 1994 Economic mediation and the power of associations: toward a concept of encompassment. *American Anthropologist* **96**(1): 31–51

McCay B M, Acheson J M (eds.) 1987 *The Question of the Commons: The Culture and Ecology of Communal Resources*. University of Arizona Press, Tucson, AZ

McGoodwin J R 1990 *Crisis in the World's Fisheries: People, Problems, and Policies*. Stanford University Press, Stanford, CA

Nadel-Klein J, Davis D L (eds.) 1988 *To Work and to Weep: Women in Fishing Economies*. Institute of Social and Economic Research, St. John's, Newfoundland

Pálsson G 1991 *Coastal Economies, Cultural Accounts: Human Ecology and Icelandic Discourse*. Manchester University Press, Manchester, UK

Pálsson G 1994 Enskilment at sea. *Man: The Journal of the Royal Anthropological Institute* **29**(4): 901–27

Pálsson G, Durrenberger E P 1990 Systems of production and social discourse: the skipper effect revisited. *American Anthropologist* **92**: 130–41

Ruddle K, Johannes R E 1985 *The Traditional Ecological Knowledge and Management of Coastal Systems in Asia and the Pacific*. UNESCO, Jakarta, Indonesia

Smith M E 1996 Public policy, sciencing, and managing the future. In: Nader L (ed.) *Naked Science: Anthropological Inquiry Into Boundaries, Power, and Knowledge*. Routledge, London

G. Pálsson

Folk, Indigenous, and Customary Law

The term 'customary law' in its most general meaning refers to unwritten legal systems. Though its roots are much older, the theoretical foundations were laid in the eighteenth century by the German historical school of legal theory, which was based on the assumption that all law originated from customs. There, it obtained the meaning of law that does not emanate from the state. The term became fashionable and of eminent practical importance in the nineteenth century, when colonial powers had to determine their relationship to existing legal systems in their colonies. These systems were in the majority of cases unwritten and at best partly written down. With the political demands for self-determination of indigenous peoples, customary law, as one of the fundamental features of their culture and claims for sovereignty, has received renewed interest and political importance. This article will discuss the various terms used for unwritten law and the advantages and disadvantages of each of them. It will then present an overview of the political controversies around the recognition of customary law. Then follows a discussion of the historical developments in the research themes concerning customary law. Finally the debate about the question whether customary law is an invention by colonial powers will be presented.

1. *Customary Law, Traditional Law, Indigenous Law, Folk Law, Local Law*

The terms 'traditional law,' 'customary law,' 'indigenous law,' 'folk law,' and—for Indonesia—'*adat* law' are often used interchangeably. Each one carries its own connotations and has its specific advantages and disadvantages (van den Bergh 1986). The problem with terms like customary and traditional is that they refer to a presumed unchanged past, though in actual fact each kind of law is subject to continuous change. Sometimes and in some places changes occur rapidly, and at other times law changes slowly and gradually. But there is no legal system that does not change at all. To presume too much continuity poses analytical as well as political problems. It leads too easily to a fixed and traditionalist representation of local law which has little to do with the social reality of such law. However, when people do not want to comply with proposals or orders from the government, often one of the few legitimate and accepted arguments to oppose government interference is to say government plans go against their traditional law. The real reasons for opposition may be quite different and have to do with distrust of corrupt government officials or differences as to the economic soundness of the proposed plans. This very argument strengthens the perception of government officials that traditional law is indeed an obstacle for development.

The struggle for self-determination of indigenous peoples also suffers from traditionalization of indigenous law. Because reference to long traditions strengthens the claim to self-determination, attention is called away from the flexibility and adjustability of such legal systems. The problem is complicated by the fact that usually there is more than one version of 'traditional' law; the law as it is used and maintained by the local communities themselves, and traditional law as it is interpreted by government institutions.

The Dutch equivalent for indigenous, *inheems*, has had strong colonial connotations and for that reason went out of use. However, in connection with the new discussions about the position of indigenous peoples, it has come into use again. The term 'folk law' (Dutch and German: *Volksrecht*) has been suggested, and in several European languages is a more neutral alternative. In English, it resembles the term 'folklore' too much and is therefore not a good alternative. For these reasons it has been proposed to use the neutral term 'local law' as a generic term for law that is being used and maintained at a local level, from whatever source it is derived.

2. *Political Controversies Surrounding the Recognition of Customary Law*

Until the nineteenth century colonial governments throughout the world had been mainly interested in controlling trade centers. However, with the change from mercantile to production colonies, colonial administrations had to deal increasingly with local populations at large. In their attempts to gain control they had to determine their relationship towards the existing political and legal systems. In general, they refrained from much interference with family law, but other sectors of law—such as criminal law and land law—pertaining to the political and economic position of colonists and their economic relationships towards the local populations could not simply be left to customary law. Different colonial powers developed very different policies. One extreme was France, with its strong centralistic tradition; it reduced recognition

of existing legal systems as much as it could. England represented the other extreme, with its indirect rule and its concomitant wide recognition of customary law. The Netherlands took a middle position. As in all other colonial countries heated debates took place in The Netherlands about the status and scope of influence of what was called '*adat*-law' of the Dutch East Indies. Issues of both public administration, and material, economic interests were involved in these debates, which touched upon control over large tracts of land. At an abstract level, the question concerned whether colonial law was the exclusive guarantee for economic development, or whether customary law could also support economic growth and development. Ultimately, it was also a debate about the legitimacy of the colonial legal order, the colonial administration, and its monopoly on the legitimate use of force. In a more concrete way, the struggle was about the extent and legitimacy of colonial control over the most important economic assets, i.e., land. See Hooker (1975) for an overview of the legal policies of various colonial powers.

After independence, new states had to reconsider the form and constitution of their legal order. Most countries opted for continuation of the policy of their former colonial motherland and simply took over the former colonial law, including its policy concerning recognition of customary law. Some, among them Malawi, tried to establish a more 'indigenous' legal order. But usually they reverted to the old system after a while. Indonesia, although it claimed to indigenize its legal system, by and large continued the development towards unification that had been initiated by the government of the Dutch East Indies. In most former colonies the issue of recognition is hardly debated any more; there are no serious attempts in countries of the Third World to found a national legal system upon customary law.

However, in countries with indigenous populations —such as Canada, the USA, Latin America, New Zealand, and Australia—the discussion has presented a high political profile since the 1980s. Human rights, in particular the right of self-determination and the protection of cultural rights of indigenous peoples, form the focus of the debate. Many countries negotiate forms of (internal) self-determination and a certain degree of autonomy for indigenous peoples. Debates about minorities in Eastern Europe and the former Soviet Union, and discussions about ethnic minorities in Western Europe, show parallel features with the debates about indigenous peoples (Finkler and K. von Benda-Beckmann 1999). Notions of cultural identity are increasingly used as an argument for autonomy and recognition of group rights. They are also used as an argument to explain and defend 'delinquent' behavior (i.e., the 'cultural defense').

The debate recurs in the context of natural resource management and environmental protection. Indigenous groups and their representatives are trying to develop 'indigenous' forms of management of tropical forest, land, and water, and call for recognition of local communal rights. Participatory development programmes for natural resource management are also based on the—contested—idea that customary forms of regulation are more adequate than other types of regulation. Notions of participation, self-government, 'good governance,' and sustainability, developed in international law and administration, play an important role in these discussions.

3. *Studying Customary Law*

The empirical study of customary legal systems began at the end of the nineteenth century. During that century missionaries, travellers, and civil servants had provided a host of widely diverse descriptions of the customs that they encountered. The colonial governments had become particularly interested in land tenure systems, family and inheritance laws. Scholars in the then emerging social sciences developed an interest in customary law as a means to study the evolution of modern society. Thus it was a combination of practical concerns of the colonial government and scholarly interest that elicited a long and rich tradition of the study of customary law.

One problem was how to study unwritten legal systems. The ordinary approach of lawyers to look at legislative and court decisions was inappropriate, because there were virtually no documents. Over the years, different approaches to solve this problem have been developed with different aims. The German legal scholars Post and Kohler were the first to set up a systematic worldwide collection of unwritten legal systems to be used for comparative study. Unlike researchers of the twentieth century, who went out to do extensive field research to learn the ins and outs of particular legal systems, Post and Kohler sent out questionnaires with detailed questions on all aspects of a legal system, to missionaries, travellers, and civil servants. Another German scholar, Bachofen, and his contemporary, the British scholar Sir Henry Maine, each published a comparative study of legal systems in 1861—*Das Mutterrecht* and *Ancient Law* respectively —works firmly rooted in an evolutionary perspective.

From the beginning of the twentieth century a large number of empirical studies began to appear on customary legal systems. Until then the focus had been on a description of the rules of a particular society. But due to scholars like Ehrlich, Malinowski, and Van Vollenhoven (see Holleman 1981), the focus shifted towards a study of the working of law in society. It was their contention that the law could be distilled from ordinary social behaviour. Malinowski's *Crime and Custom in Savage Society* (1926) became one of the most famous examples in this tradition.

From the 1930s on, British, South African, and American scholars doing research in Africa developed

a rich research tradition in the study of customary law with a strong focus on disputes. Standing firmly in the evolutionary tradition, they sought to find the original, authentic customary law, unaffected by modernity and the law of the colonial state. Gluckmann (1965, 1967), Gulliver (1963), and Llewellyn and Hoebel (1967) carefully—sometimes even purposefully—disregarded the colonial state and its representatives. As a consequence, they did not consider it necessary to deal with complex normative structures. Pospišil (1958, 1967) and Bohannan (1965) did write about legal complexity, although they did not deal with the law of the colonial state. The disregard of the state came under vehement attack in the 1970s from critics such as Moore (1978), Spittler, and Von Trotha, who argued that, contrary to what these authors suggested, the state and its law were not at all absent.

Since the 1970s, several important developments have taken place in the study of customary law.

(a) While the earlier studies concentrated on disputes as a source for learning about law, later on the disputing process itself moved into the centre of attention. The 'Berkeley village law project' under Laura Nader led to the classic *Law in Ten Societies* (Nader and Todd 1978). Schott (1978) wrote *Das Recht gegen das Gesetz*. The main objective was no longer the study of customary rules and regulation as such, but the study of modes of dispute management outside courts or court-like institutions (Morse and Woodman 1987); Gulliver's work on negotiation and mediation (1979) is a famous example. These studies have been influential in the movement towards alternative dispute resolution (ADR) in the USA and Europe. ADR is becoming an important export article of the USA to developing countries. For a critical examination of these modern forms of customary law, see Nader (1996); see also K. von Benda-Beckmann (2000).

(b) Customary law is no longer exclusively studied in the context of disputes. Based on the theoretical insight as developed by Ehrlich, Malinowski, and Van Vollenhoven (see Holleman 1981), that law operates primarily in ordinary social life, the role of law in ordinary social relations and social behavior has become a major object of research (F von Benda-Beckmann 1979).

(c) Customary law is nowadays always seen as a part of complex legal systems. It is studied in its relation not only to the law of the state but also to religious law, with which it merged in some societies, while in others the two are considered to be different bodies of law. It embraces, besides the more classical forms of customary law, all kinds of self-regulation within companies and in business branches, ethnic communities, the Mafia, local communities, and organizations of various kinds. In many instances mixed forms of law develop, or unnamed new forms of law (De Sousa Santos 1977). International law, including Human Rights of Indigenous Peoples, has provided a new stimulus in the study of customary law in its relation to state law and international law (Merry 1997).

4. *The Construction of Customary Law*

Many authors have warned that what is often thought to be authentic customary law, in the sense of untouched by Western influence, has in fact been fundamentally influenced by colonial rule. The terms 'invented' or 'fabricated' customary law are proposed to indicate the changed character of customary law as it is usually found today. Some authors maintain that it does not make sense to even speak of customary law as some form of changed or transformed indigenous law, because what has been constructed by the state institutions is an entirely new type of law, invented by the state (Moore 1986, Wiber 1993).

This is not a problem that belongs to the past. In former colonies most courts still apply customary law to many disputes in which the rural population is involved. Parts of customary law are thus incorporated into the law of the state. This is a case of constructed legal pluralism. However, application of customary law regards substantive rules only; procedures of state courts are always designed by the state. Thus 'traditional' substantive law is applied under procedures of state law, as a result of which the substantive rules have undergone important changes. In fact, two kinds of customary law coexist: local law as it is developed and applied by the local population itself; and 'customary law,' or 'traditional law,' as it is applied by government institutions. These two types of law are certainly not fully independent, but stand in complex relationships towards each other and towards the law of the state. Dutch authors writing about colonial law, such as Van Vollenhoven were well aware of what was happening (Holleman 1981). The term *adat* law was designed precisely to indicate the difference from *adat*—the Malay term for local customs in the widest sense of the word, the way of life. Today the term *adat* law is generally used for both the state defined and the local version of customary law.

Courts play a central role in the creation of customary law of the state. Other government institutions (notably land registration offices and extension services) as well as banks adopt the courts' interpretations, enlarging the realm of state-defined customary law. Sometimes these institutions even develop their own version which is slightly different from the courts' version. Thus the influence of international institutions like the World Bank and also internationally operating NGOs and donor agencies has not only renewed the interest in customary law, it also has made the study of customary law more complex.

5. Conclusion

For a long time, customary law was regarded as a vestige of the past that was bound to disappear, but it is clearly far more resilient than was expected. It is subject to change, and it always develops in relationship with, and in response to, other normative orders —of which most important are the state legal system and religious law—but increasingly so in response to international law. In addition, self-regulation and other forms of regulation not emanating from the state or state-like institutions are new twigs on the branch of customary law, giving the concept a far broader scope than it used to have.

See also: African Legal Systems; Common Law; Conventions and Norms: Philosophical Aspects; Law: Change and Evolution; Legal Systems, Classification of; Norms; Tradition, Anthropology of; Traditions in Sociology; Traditions: Social

Bibliography

Benda-Beckmann F von 1979 *Property in Social Continuity: Continuity and Change in the Maintenance of Property Relationships Through Time in Minangkabau, West Sumatra*. Martinus Nijhoff, The Hague, The Netherlands

Benda-Beckmann K von 2000 Transnational dimensions of legal pluralism. In: Fikentscher W (ed.) *Begegnung und Konflikt—eine kulturanthropologische Bestandaufnahme*. C H Beck, Munich, Germany

Bergh G van den 1986 The concept of folk law in historical context. In: Benda-Beckmann K von, Strijbosch F (eds.) *Anthropology of Law in The Netherlands*. Foris, Verhandelingen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde 116, Dordrecht, The Netherlands, pp. 67–89

Bohannan P 1965 The differing realms of the law. In: Nader L (ed.) The ethnography of law. *American Anthropologist* **67**: 33–42

Finkler H, Benda-Beckmann K von (eds.) 1999 *Folk Law and Legal Pluralism: Societies in Transformation*. Papers of the XIth International Congres. Department of Circumpolar Affairs, Ottawa, Canada

Gluckman M 1965 *The Ideas of Barotse Jurisprudence*. Manchester University Press, Manchester, UK

Gluckman M 1967 *The Judicial Process Among the Barotse of Northern Rhodesia*, 2nd edn. Manchester University Press, Manchester, UK

Gulliver P H 1963 *Social Control in an African Society*. Boston University Press, Boston

Gulliver P H 1979 *Disputes and Negotiations: A Cross-cultural Perspective*. Academic Press, New York

Holleman J F (ed.) 1981 *Van Vollenhoven on Indonesian Adat Law*. Martinus Nijhoff, The Hague, The Netherlands

Hooker M B 1975 *Legal Pluralism: An Introduction to Colonial and Neo-colonial Laws*. Clarendon Press, Oxford, UK

Llewellyn L, Hoebel E A 1967 *The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence*, 2nd edn., 4th print. University of Oklahoma Press, Norman, OK

Malinowski B 1926 *Crime and Custom in Savage Society*. Routledge, London

Merry S 1997 Legal pluralism and transnational culture: The Ka Hoókolokolonui Kanaka Maoli Tribunal, Hawaï, 1993. In: Wilson R A (ed.) *Human Rights, Culture and Context: Anthropological Perspectives*. Pluto Press, London, pp. 28–48

Moore S F 1978 Archaic law and modern times on the Zambezi: Some thoughts on Max Gluckman's interpretation of Barotse law. In: Gulliver P H (ed.) *Cross-examinations: Essays in Memory of Max Gluckman*. E. J. Brill, Leiden, The Netherlands, pp. 53–77

Moore S F 1986 *Social Facts and Fabrications: 'Customary' Law on Kilimanjaro 1880–1980*. Cambridge University Press, Cambridge, UK

Morse B W, Woodman G R (eds.) 1987 *Indigenous Law and the State*. Walter de Gruyter, Dordrecht, The Netherlands

Nader L 1996 Civilization and its negotiations. In: Caplan P (ed.) *Understanding Disputes: The Politics of Argument*. Berg, Oxford, UK, pp. 39–63

Nader L, Todd H F (eds.) 1978 *The Disputing Process: Law in Ten Societies*. Columbia University Press, New York

Pospišil L 1958 *Kapauku Papuans and their Law*. Yale University Publications in Anthropology, New Haven, CT

Pospišil L 1967 The attributes of law. In: Bohannan P (ed.) *Law and Warfare: Studies in the Anthropology of Conflict*. Natural History Press, Garden City, NY, pp. 25–41

Schott R 1978 Das Recht gegen das Gesetz. In: Kuhlbach F, Krawitz W (eds.) *Recht und Gesellschaft. Festschrift für Helmut Schelsky zum 65. Geburtstag*. Duncker & Humblot Berlin, pp. 605–38

Sousa Santos B de 1977 The law of the oppressed: The construction and reproduction of legality in Pasagarda. *Law and Society Review* **12**(1): 5–126

Wiber M 1993 *Politics, Property and Law in the Philippine Uplands*. Wildrid Laurier University Press, Waterloo, Ontario, Canada

K. von Benda-Beckmann

Folk Religion

The term 'folk religion' is used rarely today, as it is loosely defined, with the word folk bearing several overtones. It is associated primarily with the English word 'folk-lore,' first used in 1846, and referring to traditional popular beliefs, stories, customs, and superstitions. It relates to the nineteenth century belief that research into 'survivals,' or ancient customs and traditions, as practiced by, e.g., English peasants, and/or 'primitive' peoples, could provide reliable clues to the origins, evolution, or 'progress' of mankind. Folk religion, as part of folklore, is therefore linked to certain European Renaissance, Enlightenment, and post-Enlightenment concepts within psychology and the philosophy of history, concerning the nature of mankind and modes of cultural reasoning. These include human evolution and diversity, comparative religion, and the social functions of religion, including its cultural adaptations.

1. Folk Religion in England

In England and in many other Christian countries, folk religion referred to the survival of pre-Christian and medieval practices. The latter were rooted in the Roman Catholic traditions of pre-Reformation England, when peasants learned through performative ritual rather than intellectual adherence to theological beliefs. The Latin prayers of the medieval Church blurred into charms and superstitions when offered by an illiterate and uncomprehending laity to ward off disaster and disease. Fertility and purification rites, ranging from corn dollies to the churching of women, were customary in rural areas. They became part of folklore, as did continuing beliefs in fairies, portents, ghosts, and witches. Such beliefs often continued well into the twentieth century. In the sixteenth century, the Protestant Reformation brought about the religious use of the vernacular, advances in printing, and increasing literacy. Nevertheless, in the sixteenth and seventeenth centuries, the boundaries of magic, science, and technology overlapped. Dreams were often accepted as supernaturally inspired, on a par with divine judgments and the direct intervention of God in the physical world. From the eighteenth century onwards, such beliefs continued to bridge the divisions between folk and conventional religion.

2. Explorations and Learning in Europe

Throughout Europe, the geographical explorations and new learning of the Renaissance was followed, in the eighteenth and nineteenth centuries, by increasing interest in man's origins and in the natural world. Great interest was shown in the Americas, Asia, the Pacific islands, and in their inhabitants. Writers ranging from Michel de Montaigne to Jean-Jacques Rousseau drew numerous comparisons between these newly discovered peoples and the ancient founders of European civilization, the Greeks and the Romans. At first this took the form of social speculation, of antiquarianism, and of archaeological excavations, such as the eighteenth century excavations of Pompeii and Herculaneum by Johann Winckelmann, which attempted to prove the historical reality of places and events described by the ancient writers. In the nineteenth century, with the expansion of the French and British empires, the interest in antiquity began to include the Middle East. Efforts were made to verify sites mentioned in the Bible. This was partly in response to devastating scientific attacks upon the literal creationism of the Book of Genesis, notably those made by the concepts of geological time, and by evolutionism (as expressed in Charles Darwin's *The Origin of Species on the Basis of Natural Selection* 1859). As well as those interested in establishing the origins of mankind through Biblical truths, others sought the missionisation and Christian conversion of technologically simple peoples recently encountered in Africa, Australia, and other newly colonized areas. Others saw possible parallels between such people's behavior, customs, and beliefs, and those of early humans.

3. Comparative Religion and Human Origins: Totemism and Evolution

Such beliefs included religion. Under this came polytheism, monotheism, and the concept of a high god; the after-life; theodicy; and divine agency. Opinions were divided as to whether there existed any peoples so 'primitive' as to recognize no god at all. Some commentators, such as John McLennan, and William Robertson-Smith, were preoccupied by sacrifice and cannibalism, sometimes relating these to the origins of Christianity. In *Lectures on the Religion of the Semites* (1889), Robertson-Smith argued that totemism was the earliest form of religion, and that the periodic eating of the totemic animal was both a form of sacrifice and a form of cannibalism, an idea later adopted by Sigmund Freud. Others, such as James Frazer (*The Golden Bough* 1900–22), believed that the answer lay in careful comparative study of all known religious and other rites. Frazer's attempt to elucidate the meaning of divine sacrifice assumed that the comparison of similar customs throughout the world might help to understand the behavior of early humans. This inferred the racial and psychic unity of all mankind. Frazer believed in evolution and the idea of progress, as did his contemporary, Edward Tylor. Tylor, the founder of what is now seen as modern British anthropology, was influenced by Auguste Comte, and became particularly interested in animism and the definition of culture. Frazer and Tylor theorized that the intellectual development of mankind could be described as a movement from magic to religion, and from religion to scientific thinking, an idea later taken up by Bronislaw Malinowski (1925).

4. Folk Religion: Social Functions, Evolution, and Diversity

Folk religion, like folklore, implies the fusion of ancient and modern religious practices, customs, and beliefs. As such it refers to what would today be called multiculturalism, cultural adaptation, and syncretism. These are studied with reference to their social functions, rather than to the theological and philosophical questions which beset man's search for meaning in the universe. As an idea, folk religion belongs both to the Enlightenment and counter-Enlightenment traditions within anthropology. Its most notable representatives include Tylor and Frazer, who tried to present religious beliefs in evolutionary

terms, and Johannes Herder and Max Müller, who insisted on the irreducible diversity of human cultures. Herder argued that cycles of growth and decay affected both man and nature, and that the art and literature of each nation, the 'Volksgeist,' had a unique cultural value that was incompatible with developmental hierarchies. Belief in the Volksgeist was to fuel nineteenth-century nationalism and claims that particular ethnic groups were entitled to their own homelands or nation-states. Müller, a folklorist and noted religious scholar, held that ethnic identity sprang from two human universals, religion and language.

5. *History and Psychology: Society's Impact on the Individual*

The belief that mankind consisted of one species, rather than several racial subspecies, meant that concepts of psychic unity coincided with questions about society's impact on the individual. This was already a well-established theme in the philosophy of history, with the idea that the course of history was ruled by something greater than the influence of individuals who participated in it. The Renaissance historian Giovanni Vico, who developed a cyclical theory of civilisation, attributed this to divine intervention. Georg Hegel saw competing ideas, 'thesis' and 'antithesis,' as determining mankind's behavior. Karl Marx, who had read both Vico and Hegel, claimed ideas were not autonomous but firmly rooted in material conditions which produced an ideological superstructure, a 'false consciousness,' which affected human action. In psychology, William James emphasized the social functions of religious practice. Freud, reworking Christian concepts of original sin, promulgated belief in the unconscious: Carl Jung, that of the collective unconscious. Emile Durkheim's *Les Formes élémentaires de la Vie Religieuse* (1912) (translation *The Elementary Forms of the Religious Life*) held that religious rites were mechanisms for affirming social solidarity, involving collective representations of the group. Durkheim's idea of collective representations influenced Lucien Lévy-Bruhl's concept of prelogical mentalities and discussion of the ways in which the mental capacities and/or, reasoning devices of so-called primitive people were thought to differ from those seen as more civilized and rationalistic. As Evans-Pritchard (1965) pointed out, those who were engaged in the sociological and anthropological study of religion generally regarded it from an Enlightenment, rationalist viewpoint, as a series of superstitions which could only be explained in terms of cognitive variation or social function.

Religion became linked to the study of emotions, to the nature of the self, and to psychological interpretations of social functions, but also to theories of economic exchange, social control, and political legitimization: links which continue today. One effect of this was that religion tended to be seen less as a holistic phenomenon than as the sum of several parts, any of which could be used to support arguments relating to the functions and meanings of other social institutions.

6. *Cultural Groups and Modes of Reasoning*

The ways in which modes of reasoning appear to differ between cultural groups is a perennial theme in anthropology, even though the arguments have long since moved from questions of innate capacities, and from the relationship between magic, science, and religion, to the ways in which particular cultural premises favor different forms of interpretation. Malinowski and Arthur Radcliffe-Brown were the first two European anthropologists to base their arguments upon personal fieldwork rather than armchair speculation. Malinowski held that the religious practices of 'primitive' peoples displayed a rational response to their surroundings, rather than one infused by the mysticism of Lévy-Bruhl. Radcliffe-Brown, influenced by Durkheim, interested himself in the expressive and affective aspects of tribal ceremonies, rather than in their possible origins.

Evans-Pritchard, like his student Mary Douglas, emphasized the logic underlying belief in witchcraft and other quasireligious practices, showing that indigenous interpretations depended on analogies, rational deduction, and the influence of emotions upon a particular line of reasoning.

7. *Folk Religion: Tradition and Adaptation*

Even though the phrase is obsolete, folk religion usually denotes either the traditional beliefs and practices of a particular people, or, more often, remnants from an earlier age. In *Folkways* (1906), the political scientist William Sumner argued that folkways, or the belief in and enactment of folklore, were the result of illiteracy combined with an intellectual closure produced by overwhelming deference to tradition. He pointed out that all beliefs were constantly adapting themselves to the exigencies of everyday life, so that 'pure' customs did not exist. Today, folk religious beliefs and practices may refer to those which have survived attempts to destroy an indigenous culture by colonialism or other means. They are then often described in terms of ethnicity and cultural resistance to forms of oppression. Such oppression may range from the exclusion of certain ethnic groups from formal learning systems to the class exclusion of the illiterate or poorly educated, including those living in rural areas and traditionally defined as peasants. Cultural resistance may encourage the oppressed to cling to beliefs and practices discarded by other social

or ethnic groups, or to create particular meanings and interpretations from their experiences, which are not countered by the interpretations given in formal meaning systems. The result is cultural adaptation, as elements of both the dominant and traditional cultures fuse together. This may give rise to religious syncretism, which may be both formal and informal, encompassing both distinctly new forms of cultural expression (e.g., Brazilian spiritist movements such as Umbanda and Candomblé) and the reworking of the previously discarded beliefs and practices of formal religious systems. Such peasant remnants and reworkings may be ignored by rationalistic and educated believers with higher social status (for a description of this, see Christian 1972). The former may continue to exist without reworking, with few substantial changes, so that they effectively consist of religious survivals, the folk religion of many years ago.

See also: Anthropology; Anthropology, History of; Belief, Anthropology of; Evans-Pritchard, Sir Edward E (1902–73); Folklore; Freud, Sigmund (1856–1939); Healing; Magic, Anthropology of; Malinowski, Bronislaw (1884–1942); Millennialism; Myth in Religion; Popular Culture; Religion, Sociology of; Shamanism; Tradition, Anthropology of; Witchcraft

Bibliography

Christian Jr. W A 1972 *Person and God in a Spanish Valley*. Seminar Press, New York
Evans-Pritchard E E 1965 *Theories of Primitive Religion*. Clarendon Press, Oxford, UK
Malinowski B 1974 *Magic, Science and Religion, and Other Essays* (1925). Souvenir Press, London
Morris B 1987 *Anthropological Studies of Religion: An Introductory Text*. Cambridge University Press, Cambridge, UK
Stocking Jr. G W 1987 *Victorian Anthropology*. Free Press, New York
Thomas K 1971 *Religion and the Decline of Magic*. Weidenfeld & Nicolson, London

M. A. Taylor

Folklore

'Folklore' has four basic meanings. First, it denotes oral narration, rituals, crafts, and other forms of vernacular expressive culture. Second, folklore, or 'folkloristics,' names an academic discipline devoted to the study of such phenomena. Third, in everyday usage, folklore sometimes describes colorful 'folkloric' phenomena linked to the music, tourist, and fashion industries. Fourth, like myth, folklore can mean falsehood. While in this text, focus is placed on the first two meanings, the other two are often present, implicitly or explicitly.

1. History

1.1 A Term is Invented

The word folklore was coined by British antiquarian William John Thoms who, in 1846, proposed that this 'good Saxon compound' be used to describe 'the manners, customs, observances, superstitions, ballads, proverbs' and other materials 'of the olden time.' Thoms was inspired by the works of the Brothers Grimm and 'folklore' is apparently a translation of the German 'Volkskunde' introduced in 1806.

Although the word was new, the interest in the phenomena themselves was not. European rulers had long ordered the collecting of traditions that would glorify them, and religious leaders had gathered dangerous beliefs in order to weed them out. However, during the national romantic currents of the late eighteenth and early nineteenth centuries, intellectuals added something new, when they insisted that the customs and arts of illiterate peasants were remnants from a remote human past and should be collected before they were bound to disappear as civilization advanced. Like Johann Gottfried von Herder, many intellectuals also saw a reflection of the true spirit of nations in the traditions of the anonymous folk living close to the national soil.

1.2 A Discipline is Founded

During the nineteenth century, many people were fascinated with folklore, among them scholars with interests in literature, philology, religion, psychology, history, archeology, anthropology, and other fields. In some countries enthusiasts formed societies: the British Folklore Society was founded in 1878, and the American equivalent 10 years later. In other countries, primarily in northern and central Europe, the idea of folklore was directly linked to nation formation. In Finland, which was in the process of freeing itself, first from Swedish and then from Russian domination, Elias Lönnrot pieced together the *Kalevala Epic* from songs collected in oral tradition. Published in 1835, the *Kalevala* became the first example of literature written in Finnish and a national symbol. In addition, the songs formed the core of a national folklore archive and, in 1898, the world's first professorship in folklore was established: the chair in 'Finnish and Comparative Folklore' at the University of Helsinki. Other departments were to follow in other European countries, often in conjunction with the founding of national museums and archives.

However, in many countries, folklore continued to be studied outside formal institutions. In some cases, such as England, separate university departments were never established, whereas in others, notably the United States, they were founded as late as after World War II (see Sect. 3).

1.3 A History of Approaches

On the whole, the ideas employed to explain and interpret folklore coincide with the major schools of thought in anthropology, beginning with evolutionism and diffusionism and moving on toward functionalism, structuralism, and later tendencies. Yet even more than in anthropology as a whole, the study of folklore was long focused on questions of origin; an array of explanations were proposed to account for the origin of such ancient and widely disseminated tales as '*The Dragon Slayer*' and other traditions. The Indianist school championed by Theodor Benfey sought to establish that all European folktales had originated in India and spread from there. The mythological school headed by Max Müller found that all folktales were survivals of ancient nature-myths, while ritualists such as Pierre Saintyves tried to demonstrate that all folktales had originated in rituals.

While these explanations now have mostly antiquarian interest, other early trends of thought remain important. Among these is, first, the diffusionist 'historical-geographical' method developed by Finnish scholars at the end of the nineteenth century. In this cumbersome procedure all recorded versions and variants of a tale type (or a ballad or a riddle) are studied so that the location of the earliest traits (and theoretically also the tale's place of origin) can be determined. The gathering of comparative evidence resulted in massive indexes and catalogues that are still often consulted (Aarne and Thompson 1961). A second intellectual movement that remains important to the understanding of folklore is psychoanalysis. Siegmund Freud, C. G. Jung, Otto Rank, and others came to regard folklore as essential expressions of the individual or collective human unconscious. A third essential strand of thought is the insight that forms of folklore are not only widely spread but also firmly localized and therefore must be studied in their immediate social contexts. Franz Boas' ethnographies of the stories of native peoples on the north west coast of the United States initiated an influential format for the study of folklore in local contexts.

Functionalism, not least Bronislaw Malinowski's analyses of story telling among the Trobrianders, is a fourth development that has had a lasting impact on the study of folklore. Particularly influential is the psycho-functionalist premise that folklore is 'socially sanctioned' fantasy; as such it permits action that is otherwise not approved (Dundes 1965, p. 277). Finally, structuralist thought has also endured among students of folklore. One structuralist approach derives from Russian formalism and emphasizes features of plot sequence; in an influential study Vladimir Propp demonstrated that all Russian fairytales share the same plot structure. A second approach derives from the work of Claude Lévi-Strauss and points to the ways in which logical relationships within myths and tales (such as binary oppositions) reflect universal processes of the human mind. Lévi-Strauss' thinking has been central to the analysis of a wide range of traditional forms, not least to Henry Glassie's monumental investigations of vernacular material culture.

2. Key Words

Despite the many decades of inquiry, it remains difficult to define folklore. In one recent attempt it is suggested that folklore embraces 'expressive forms, processes, and behaviors' that are learned, taught, and utilized 'during face-to-face interactions.' It is also suggested that people judge these phenomena to be traditional 'because they serve as evidence of continuities … through time and space in human … thought, belief, and feeling' (Georges and Jones 1995, p. 1).

However, many scholars avoid generalized statements and prefer to look at folklore as perspectives and processes that might be illuminated through such keywords as 'folk,' 'tradition,' 'genre,' 'performance,' 'text,' and 'art' (cf. Feintuch 1995).

2.1 Folk

Influenced by the typologies of early social thought and by such conceptualizations as Robert Redfield's 'ideal folk society,' students of folklore long took it for granted that the folk (i.e., the carriers and transmitters of folklore) were illiterate peasants or, to some extent, tribal peoples. However, during the last few decades, the word folk is increasingly regarded as a cumbersome inheritance from earlier thinking and most scholars agree that all human beings in some way or another are involved in folklore processes, as performers, participants, watchers or audiences—sometimes even as indifferent or disdainful audiences (see Sects. 2.4–2.6). But it has been difficult to abolish the word folk. For one thing, it remains important in studies of folklore in feudal systems or other class hierarchies. Second, it remains alive in such compounds as folk art, folk music, and folk dance; attempts to replace 'folk' with 'people,' 'vernacular,' or other words have not been successful (see Sect. 3). Furthermore, despite their conviction that all human beings partake in folklore processes, contemporary students also concentrate on marginal or subaltern peoples. But their reasons for doing so differ from those of nineteenth century predecessors; many students now argue that the role of folklorists today is to assist peoples who lack political power in their quest to gain a public voice.

2.2 Tradition

Used both to name a process of transmission and to name specific cultural elements (i.e., traditions), the concept of tradition is basic to the understanding of folklore and folkloristics. It has long been recognized that tradition not only involves historical continuity but also change. Because folklore forms are often transmitted orally or in other informal ways, they are extremely adaptable to fleeting moments of social interaction. One could even say that it is because of their adaptability that simple string-figures or guessing games have remained in circulation through millenia.

The materials deposited in archives and listed in comparative indexes demonstrate both the adaptability of some themes and their continuity through time and space. For example, comic book hero Superman travels through space in a womblike vessel only to be found and raised by a poor couple in the middle of North America. His story has parallels in the stories of Moses, Cyros, and other heroes who are similarly found as children floating around in small containers. Furthermore, identity changes are not only characteristic of Superman but also of countless other heroes and heroines in world traditions. The transformation of an orphaned young girl into a rich and beautiful queen is reenacted over and over in oral tales, films, and real life. Folklore themes have a psychological force that makes people recreate them century after century in different communicative modes and media.

During recent decades, students of folklore increasingly utilize the verb 'traditionalize' to name processes in which the past is interpreted or reinvented in a given present. Many students of folklore now emphasize that they do not investigate given sets of traditions but all sorts of practices of traditionalization in social life. Furthermore, by focusing on active traditionalization processes, scholars avoid cementing an opposition between tradition and creativity. They see tradition as a source of creativity, not as a constraint to it (Briggs 1988).

2.3 Genre

Genre means 'kind' and a genre may be seen as a bounded distinguishable form characterized by a specific combination of stylistic traits and elements of content. The classification into genres was for a long time so central that scholars thought that enumerating genres was the easiest way to describe what folklore is (Dundes 1965).

When analyzing genres, students of folklore often find themselves caught in dilemmas familiar to all students of culture. On the one hand, all peoples make local, culture specific, or 'emic' distinctions between genres in their repertoires: coyote tales, superstitions, dirty jokes, little girls' games, gossip, songs while making baskets, and so on. On the other hand, scholars may observe more overriding features in the same materials and may note how elements of content, style, or structure are shared across cultural boundaries. The extent to which local classifications are emphasized and the extent to which broader ones come into play have to vary with different analytic tasks.

One distinction that was long taken for granted but has proven problematic is that between 'fixed' genres, such as riddles and proverbs, and 'unfixed' genres, such as many narratives. While some genres, magic formulas for example, are fixed in the sense that they must be recited exactly as learned in order to have power, other genres that were long called fixed change too. For example, people often do not cite proverbs in full but merely hint at them through tone of voice. Thanks to performance-centered approaches (see Sect. 2.4), scholars have recently gained new insights into the ways in which oral genres are malleable social phenomena. In social interaction people mix or blur materials so that artful hybrids emerge. At the same time, genres are needed as 'social contracts' that aid memory and establish structure.

2.4 Performance

Since the 1960s, the concept of performance has transformed the study of folklore. Thanks to increasing refinements in recording techniques, scholars have been able to analyze folklore as part of communicative processes in ways that had not been possible before. Two performance-centered approaches will be recognized here; in both, careful ethnographic fieldwork is essential.

One approach is rooted in the ethnography of speaking and other sociolinguistic methods. A basic idea is that verbal art and other forms of folklore emerge in social interaction as aesthetically marked forms. They are recognized as something special thanks to cues made through tone of voice mimickry, gestures, and other devices. Arguing that performance is 'constitutive' of verbal art, Dell Hymes, Richard Bauman, and others have developed ways of studying folklore as artistic communication in everyday life, communication in which, as Hymes says, people shape 'deeply felt values into meaningful, apposite form.' If preachers, story-tellers, weavers, carvers, and carnival participants in the past were seen as transmitters of handed down traditions, they were now regarded as active artists with power to transform social life.

A second performance-centered approach is oral formulaic theory. This was first developed by Albert Lord and others who in studies of South Slavic oral epic poetry show how songs that are many thousands of lines long can live in tradition, although they have not been written down. They are remembered because singers learn to draw on a store of formulae and formulaic expressions and learn to adjust their singing

to different social events. Several scholars have continued along these paths and oral formulaic studies are now enriching the understanding of both ancient (in particular Homeric) epic singing and of contemporary epic performances in India, China, Africa, and Eastern Europe (Finnegan 1992, Honko 1998).

2.5 Text

It has long been recognized that the distinction between oral and written texts is far from firm and that all kinds of texts can be of interest to students of folklore. Some 'folklore texts' have always existed primarily in writing, for instance court records, religious documents, newspaper articles, and xeroxlore. Others have circulated widely as visual representations, for example fables painted inside medieval churches. Furthermore, literary authors have continuously made use of traditional themes and, in a sense, folklore texts can be found in the works of many celebrated writers. For example, contemporary African authors, such as Ayi Kwei Armah, acknowledge their profound debt to oral narrative art and Shakespeare often builds his plays on traditional materials. Actually, a great many art forms can profitably be studied as folklore texts. Classical ballet, for example, shares some of its stylized and repetitive character with the European folktale.

Recently, performance-centered approaches have opened up oral texts in unprecedented ways in that scholars have invented new methods to translate oral performances into writing. In procedures that have sometimes been named 'laying out texts ethnopoetically' students transcribe not only words but also pauses, cadences, gestures, mimicry, and other oral features (Briggs 1988). These procedures have led to many insights. One is that oral narration in many cultures tends to be closer to drama than to prose literature. Another is that repetitions and other features of oral narration (or visual folk art) are not to be dismissed as poor artistry as was often the case in the past. Rather, repetitions can be seen as ways to control complex materials. A repetitive structure simultaneously offers restrictions and creative freedom; it aids memory at the same time as it frees fantasy and improvization.

During the last few years, the term 'textualization' is increasingly used to describe modes of translating oral performances onto the written page (Honko 1998). Long regarded as routine exercises, textualization processes are now seen as central analytic tasks.

2.6 Art

On some level, scholars always seem to have agreed that folklore has to do with art, creativity, and imagination. Even at times when the study of folklore was reduced to enumerations of genres and tale types, there was a sense that folklore touches upon the aesthetic dimensions of life. This sense has become heightened during the last few decades, in part due to the interest in folklore as performance. This is particularly the case in the United States, where all kinds of vernacular expressions—from singing to cooking and quilting—are often called the 'the folk arts.'

The focus on art is linked to several perspectives. First, students make a distinction between less articulate folklore performers and artists who are celebrated in their communities and have the power to move others to action. Second, students of folklore have made important discoveries about social life by focusing, not on finished products, but on moments when material objects or community events are being planned or created. Third, the interest in art has led to new analyses of the languages of folklore. Whether verbal, visual, musical, kinesthetic, or tactile, these languages are bold in their use of intense and multivalent metaphors and other tropes.

It is because of this expressive intensity that folklore, like other forms of art, can be a vehicle through which people can examine or critique the lives they lead. For example, to people who have fled from one part of the world to another, a few songs, stories, or dishes can become embodiments of a precious inheritance which must be preserved and reconstituted in new lands. Conversely folklore can incarnate some of the worst aspects of life: detested symbols of racism and humiliating jokes that outsiders direct at one's own group. Because of its power of presence, folklore can be invoked to hold communities together or to separate them from one another.

3. Folklore and the Politics of Culture

Among contemporary scholars there is both confidence and unease about the future of folklore and folklore study. There is confidence, for example, about the supply of material. The idea that folklore is 'dying out' is no more true today than it was 200 years ago. People shape rituals and tell stories, perhaps more than ever before, but they do so not only in face-to-face interactions but also on the electronic media. There is also confidence in the sense that a great deal of scholarship is being conducted. Indeed, some scholars are convinced that, regardless of discipline, those who study folklore today play a special role as advocates of the artistic expressions and traditions of all those groups and individuals who are caught in the cracks of modern life.

But, on the other hand, scholars are uncertain about the directions of folklore study. There are renewed critical reflections on the discipline's long and eurocentric history and on its often naive, and sometimes disastrous, ideological and political entanglements (Briggs and Shuman 1993). One indication of the

uncertainties are the current controversies surrounding the designation of the field. In Germany, the term 'Volkskunde' is now often avoided, because it is deeply implicated in fascist scholarship and in the creation of symbols for the Third Reich. Here and in some other countries, 'ethnology' is preferred to compounds containing 'folk' (see Sect. 2.1). In other countries, the word folklore is perceived as an inheritance from a colonial era and is being replaced by new coinages, such as the South Indian 'janapada.'

Many critical reflections concern the paradoxical situation that folkloristics regards itself an international field at the same time as it is almost everywhere a construction in the service of nation building and related local endeavors. This was true in the nineteenth century and it is true in the twenty-first. For example, in some newly formed or recently reconstituted nations, folklore and folklore archives play a critical part in efforts to construct visions of homogeneous and exclusive national heritages. In other situations, folklore and folklorists are linked to nation building in other ways. In the United States, for example, they have become important in affirming a vision of the country in which a multitude of diverse cultures can co-exist. This is particularly true of so-called public folklorists. Employed in museums, at the American Folklife Center in Washington DC, or as state folk arts coordinators, these professionals have special skills in collecting, preserving, and presenting the folklore of an immense diversity of groups. But also on a global level folklorists participate in cultural politics. Many work with UNESCO in efforts to safeguard folk traditions around the globe from the onslaughts of modernity.

Whether nationally or globally based, all these endeavors to collect, preserve, and present folklore raise many questions. In whose interest are these efforts being carried out? Whose traditions should be safeguarded and whose should be deleted? What happens to all those traditions that are not beautiful or presentable but offensive, frightening, or incomprehensible? In the view of many scholars, it is the many different kinds of heritage politics and ideological mazes that ought to become an important focus of attention for folklore scholarship in the years to come.

Bibliography

Aarne A, Thompson S 1961 *The Types of the Folktale. A Classification and Bibliography*. FF Communications 184, Suomalainen Tiedeakatemia, Helsinki

Bausinger H 1961 *Volkskultur in der teknischen Welt*. Kohlhammer, Stuttgart (1990 *Folk Culture in a World of Technology*. Indiana University Press, Bloomington, IN)

Briggs C L 1988 *Competence in Performance: the Creativity of Tradition in Mexicano Verbal Art*. University of Pennsylvania Press, Philadelphia, PA

Briggs C L, Shuman A (eds.) 1993 Theorizing folklore: toward new perspectives on the politics of culture. *Western Folklore* **52**(April): 109–400

Cocchiara G 1952 *Storia del Folklore in Europa*. Editore Boringhieri, Torino (1981 *The History of Folklore in Europe*. Institute for the Study of Human Issues Press, Philadelphia, PA)

Dundes A (ed.) 1965 *The Study of Folklore*. Prentice-Hall, Englewood Cliffs, NJ

Feintuch B (ed.) 1995 Common ground. Keywords for the study of expressive culture. *Journal of American Folklore* **108**(430): 391–550

Finnegan, R 1992 *Oral Traditions and the Verbal Arts. A Guide to Research Practices*. Routledge, London

Georges R A, Jones M O 1995 *Folkloristics: An Introduction*. Indiana University Press, Bloomington, IN

Glassie H H 1982 *Passing the Time in Ballymenone. Culture and History of an Ulster Community*. University of Pennsylvania Press, Philadelphia, PA

Honko L 1998 *Textualizing the Siri Epic*. FF Communications 264, Helsinki

Toelken B 1996 *The Dynamics of Folklore*. Utah State University Press, Logan, UT

B. Klein

Food in Anthropology

The study of food in anthropology involves simultaneous consideration of both culture and biology. Work spans a wide array of emphases from largely biological to largely cultural, and comprehends a range of topics, methodologies, and theoretical approaches. Food and culture studies emphasize the social and symbolic dimensions of food to gather insights about gender, class, political economy, power, cultural ideology, and cultural identity (Counihan and Van Esterik 1997). Nutritional anthropology gives more weight to biological considerations and examines the interplay of food supply and hunger, nutritional status, breast-feeding, disease, household dynamics, and human evolution (Goodman et al. 1999). The two different approaches overlap and complement each other and reveal the power of anthropological holism to explain the human condition.

1. History and Scope of the Field

Many early ethnographers paid attention to food habits because food is a central focus of the economy in most cultures. In noncapitalist societies, the production, distribution, and consumption of food take up a great deal of people's time, energy, and attention. Marcel Mauss (1923) laid an important foundation of foodways research in anthropology with his pioneering study of exchange in *The Gift* (see also *Exchange in Anthropology*). Mauss directed our attention to the importance of giving, receiving, and repaying in initiating, maintaining, or breaking social and political relationships, and reminded us of how often food is a

key substance in exchanges. Gifts, Mauss pointed out, are 'total social phenomena,' a concept that applies beautifully to food. Mauss used ethnographic and historical literature to explore gifts in a variety of cultural forms, for example the Northwest Coast Indian potlatch and the Maori 'taonga.' He described the meaning of the ancient Hindu food offerings thus: 'The thing given brings a return in this life and the other … Food given away means that food will return to the donor in this world; it also means food for him in the other world and in his series of reincarnations' (1967, pp. 54–5). Food gifts establish obligations of reciprocity in recipients whether they are gods or humans.

The centrality of food in gifts and economic transactions has ensured its central role in expressing religious beliefs, communicating with others, establishing and breaking relationships, defining social groups and families, and expressing cultural values. Yet few early anthropologists focused explicitly on food, the exception being Audrey Richards who conducted pioneering work on land, labor, diet, and hunger in Southern Africa (1932, 1939). She took a functionalist approach to food and examined the interconnectedness of food-centered production, distribution, consumption, and beliefs (see also *Functionalism in Anthropology*). She consistently examined not only food's important economic and nutritional dimensions, but also its role in constructing social relationships, showing for example how gifts of food declared and enacted the new obligations between families at marriage. Richards was the first anthropologist to do a full fieldwork study centered on food. Although her work was well regarded, it did not inspire followers for several decades.

The current burgeoning of studies in food anthropology began in the 1960s with the stunning work of Claude Lévi-Strauss (1969 [1964]) in France and Mary Douglas (1966) in England. Lévi-Strauss developed his structural theory by applying it to a symbolic analysis of food and other important symbols in hundreds of myths collected from Amazonian Indians. He introduced the idea of the 'gustatory code' (1969, p. 164) which is the most important of the five basic codes (stemming from the five senses) that humans use to transmit cultural messages. The gustatory code refers to the predominance of food symbols and metaphors in human expression which is due to 'the truly essential place of cooking in native thought: not only does cooking mark the transition from nature to culture, but through it and by means of it, the human state can be defined with all its attributes' (1969, p. 164). The study of myths and the consistent patterns that they display on diverse levels of meaning reveals that 'myths signify the mind that evolves them … there is simultaneous production of myths themselves, by the mind that generates them and, by the myths, of an image of the world which is already inherent in the structure of the mind' (1969, p. 341).

Mary Douglas took a different approach to the symbolic study of foods by focusing on their ability to reveal not the mind's structure but social structure. She demonstrated how foodways revealed concepts of purity and pollution and constructed social boundaries in diverse cultural settings, most prominently in her study of Jewish dietary law. She argued that Jewish rules about food consumption continually reinforced concepts of holiness and renewed the community of the faithful. Together, the work of Douglas and Lévi-Strauss gave an influential demonstration of the role of food as 'good to think' as well as 'good to eat' and overcame the anthropological tendency to take food for granted.

Food studies in anthropology proliferated in the 1970s, 1980s, and 1990s. Researchers have used ethnographic, archaeological, and biocultural fieldwork methods to study foodways—the beliefs and behaviors surrounding the production, distribution, and consumption of food. Douglas and Lévi-Strauss' emphasis on the symbolic role of food did not go unchallenged. Marvin Harris (1985) responded to their work by arguing explicitly that food habits exist for utilitarian purposes. 'Preferred foods (good to eat) are foods that have a more favorable balance of practical benefits over costs than foods that are avoided (bad to eat)' (1985, p. 15). Harris gave priority to material over symbolic concerns in explaining Jewish dietary law. He argued that the pig was an economic competitor with humans in the arid lands of the Middle East where it made much better economic sense to raise ruminants who do not need high-protein foods to flourish as pigs do. Thus, he argued food aversions and preferences evolve primarily to support utilitarian practices based on nutritional and ecological concerns rather than to support religious belief systems.

Douglas and Harris's different explanations for Jewish dietary law show how the study of food provides diverse and complementary insights into human behavior. Anthropologists have examined foodways to gain insights into power in politics, family, gender, class, and race. They have studied human evolution, health, and disease; famine, hunger, and malnutrition; and the relationships between cultural ideology and behavior.

2. *Food and Power*

Because food is the most insistent and basic daily survival need, it is a central pawn in power struggles in many different settings. In Papua New Guinea, for example, men use food to acquire and demonstrate political leadership. In Kalauna, men 'fight with food' by shaming others through excessive gifts of comestibles (Young 1971). In Wamira, Papua New Guinea (Kahn 1986), men demonstrate their leadership and public renown by accumulating food and giving it away to followers and enemies alike at public feasts.

At the same time, they refrain from eating and thus demonstrate their self-control and prestige through 'the potency of [their] hunger-assuaging magic' and their ability 'to tame the most frightening of uncontrollable behaviors—namely ravenous, rapacious desire—both gustatory and sexual' (Kahn 1986, p. 86).

The attainment of power through giving food away typical of Melanesian and other subsistence cultures contrasts dramatically with the goal of capitalists, who attempt to withhold food and sell it for profit and power, as Mintz (1985) capably demonstrated in his anthrohistorical study of the worldwide commodification of sugar. Sugar production involved Europeans in conquest of Third world countries, seizure of raw materials, and enslavement of Africans to provide land and labor to produce sugar cheaply enough to fuel a vast expansion of sugar consumption and exorbitant profits. The story of sugar provides a case study in how foodways research can uncover the interworkings of political, economic, and cultural systems of power.

Goody (1982) focused explicitly on the role of foodways in constituting class relations by examining the conditions for the emergence of high and low cuisine among some African, European, and Asian cultures. He examined how 'industrial food' can be a form of cultural domination of Third World cultures and explored the socioeconomic structures that enable them to resist. Meillassoux (1975) used a Marxist perspective and economic focus on food production and distribution to examine the persistence of the household economy under the corrosive conditions of colonization and capitalist labor exploitation. Weismantel (1988) showed how specific food behaviors identify class and race in Highland Ecuador. Indians regularly eat potatoes but only on special occasions do they consume the white rice that forms the staple food and symbol of class and race privilege of the whites. In the USA, distinctions between rich and poor, male and female, and white and 'ethnic' Americans are all enacted through foodways (Counihan 1999).

3. *Food and Gender*

Food is intimately connected with the definition and assertion of gender identity in many cultures (Counihan 1999). Food work ordinarily falls to women and whether this work is revered or denigrated is an important constituent in women's status. Among the Zumbagua of the Ecuadorian Highlands, women use their cooking to express approval or disapproval of their husbands, and they establish hierarchy in the family by their allocation of scarce and valued meat (Weismantel 1988). Among Wamirans of Papua New Guinea, women's roles in producing food and children are likened to each other and valued highly (Kahn 1986). To emulate women's important role in procreation and establish symbolic gender parity, Wamiran men cultivate taro, which they call their 'children' and which represents their 'personal status and virility' (Kahn 1986, p. 91). Among the Hua of highland Papua New Guinea, men's elaborate system of food rules reveals how they construct gender and overcome their own fears of women's power of procreation (Meigs 1984). They define foods in two categories, koroko and hakeri'a, female and male. Female foods are wet, cold, fertile, soft, and fast growing; male foods are dry, hot, infertile, hard, and slow growing. Men publicly aver that women and the foods associated with them are polluting and dangerous, yet in rituals they secretly consume the female foods to gain the power, fertility, and vitality of women.

In all cultures, gender relations are clearly expressed through the attitudes of men and women towards eating, feeding, and the body. In contrast to the great majority of cultures, where women aspire to plumpness and pleasure in eating, in the USA, women reproduce their subordination by restricting their eating, reducing their bodies, and loathing their own appetites. As capitalism transforms women from producers into consumers and commodities, they lose social value and power, their bodies become objects, and they lose subjecthood and agency (Counihan 1999).

Anthropologists have shown how cultural practices and attitudes towards breast-feeding vs. formula feeding reveal much about women's power and status. Van Esterik (1989) used fieldwork data and documentary sources on cases in Thailand, Indonesia, Kenya, and Colombia to argue that the adoption of formula-feeding robs women of an important bodily ability, costs them money, and generates profits for multinational corporations at the expense of infant health, especially among the poor in the USA and the world. Empowerment of women depends on creating conditions where they can breast-feed if they choose with dignity and respect so as to ensure the best possible health of children and validation of women's corporeal abilities.

4. *Nutritional Anthropology*

The relationship between cultural practices, food consumption, health, and disease past and present has been a central focus of nutritional anthropology. Harris and Ross (1987) developed their edited volume *Food and Evolution* to demonstrate the important role in human development of the interplay of biological and behavioral 'repertoires' surrounding food. Goodman et al. (1999, p.1) defined nutritional anthropology as the study of 'the interrelationships of biological and social forces in shaping human food use and the nutritional status of individuals and populations.' Their volume explored the evolution of human foodways with attention to both fossil and primate studies and the importance of plant and animal domestication

on diet and health. It looked at case studies on lactose tolerance, maize processing technology, and bitter cassava consumption to explore the development and impact of specific diets and food processing techniques and their ecological and health implications. They examined the biocultural causes and effects of hunger, malnutrition, and overeating.

5. *Hunger and Malnutrition*

Few anthropologists have written books on the causes and effects of hunger, the exception again being the pioneering work of Audrey Richards (1932). Recently there have been a growing number of good articles, some of which have been, anthologized (see Counihan and Van Esterik1997, Goodman et al. 1999, Harris and Ross 1987). Dettwyler (1994) is one of few anthropologists to produce a book-length study of hunger. She focused on feeding practices, health, and disease in Mali by combining ethnographic observations with nutritional methods. She used dietary nutrient assessment, height and weight measurements of children and adults, collection of health beliefs, observations of feeding and breast-feeding practices, and blood, urine, and stool sampling. She was able to assess health status and parasitic invasions and show their relationship to cultural practices and attitudes. For example, Malian beliefs about children's normal appearance being thin and pot-bellied obscured their ability to recognize or end malnutrition in children. Dettwyler's goal was to use her study to develop 'culturally appropriate nutrition education programs, aimed at improving people's understanding of the links between diet and health, and the importance of making sure young children get enough food, of sufficient quality, to meet their needs for growth and health' (1994, p.7). Nutritional anthropology demonstrates the analytic power of combining sensitive social observations with scientific data-gathering to understand the causes and potential cures of hunger, malnutrition, and the diseases of poverty.

6. *Future Directions*

Future studies of food in anthropology will continue to address links between cultural systems, social inequality, food, and hunger. Studies can do this by examining food in popular culture—particularly the use of food images in advertising, television, film and popular music—where the messages transmitted often penetrate public consciousness with little critical reflection. The study of McDonalds in East Asia by Watson (1997) and his collaborators links the anthropology of food to the theme of globalization. Anthropology's symbolic and biocultural approach is particularly useful in studying the growing epidemic of eating disorders, including anorexia nervosa, bulimia, and compulsive eating. These disorders strike women in much greater numbers than men and seem to have psychocultural causes associated directly with women's subordinate status and socially constructed role as those who should feed others but not themselves. The cross-cultural study of the spread of, and resistance to, eating disorders in Third World countries where they were previously unknown will provide important insights into their causes, consequences, and cures.

As multinational corporations increasingly dominate food production and distribution, anthropologists will need to keep a close eye on the effects of food profit-making on access to and quality of food. Issues surrounding the bioengineering, genetic manipulation, loss of genetic diversity, and environmental pollution and protection in agricultural production will be paramount. Because of the centrality of food to human physical well being and cultural expression, anthropologists must continue to document and promote efforts to end hunger and to establish food as a human right.

See also: Consumption, History of; Fishing Societies; Food Production, Origins of; Functionalism in Anthropology; Horticultural Societies; Hunter–Gatherer Societies, Archaeology of; Hunting and Gathering Societies in Anthropology

Bibliography

Counihan C M 1999 *The Anthropology of Food and Body: Gender, Meaning and Power*. Routledge, New York

Counihan C, Van Esterik P (eds.) 1997 *Food and Culture: A Reader*. Routledge, New York

Dettwyler K A 1994 *Dancing Skeletons: Life and Death in West Africa*. Waveland, Prospect Heights, IL

Douglas M 1966 *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*. Praeger, New York

Goodman A H, Dufour D L, Pelto G H (eds.) 1999 *Nutritional Anthropology: Biocultural Perspectives on Food and Nutrition*. Mayfield, Mountain View, CA

Goody J 1982 *Cooking, Cuisine and Class: A Study in Comparative Sociology*. Cambridge University Press, Cambridge, UK

Harris M 1985 *Good to Eat: Riddles of Food and Culture*. Simon and Schuster, New York

Harris M, Ross E B (eds.) 1987 *Food and Evolution: Toward a Theory of Human Food Habits*. Temple University Press, Philadelphia, PA

Kahn M 1986 *Always Hungry, Never Greedy: Food and the Expression of Gender in a Melanesian Society*. Cambridge University Press, Cambridge, UK

Lévi-Strauss C 1964 *Le Cru e Le Cuit*. Librairie Plon, Paris [1969 *The Raw and the Cooked: Introduction to a Science of Mythology*. Harper and Row, New York]

Mauss M 1923 Essai sur le don: Forme et raison de l'échange dans les sociétés archaïques. *L'Année Sociologique* **1**: 30–186 [1967 *The Gift: Forms and Functions of Exchange in Archaic Societies*. Norton, New York]

Meigs A S 1984 *Food, Sex, and Pollution: A New Guinea Religion*. Rutgers University Press, New Brunswick, NJ

Meillassoux C 1975 *Femmes, greniers et capitaux*. F Maspero, Paris [1981 *Maidens, Meals and Money*. Cambridge University Press, Cambridge, UK]
Mintz S W 1985 *Sweetness and Power: The Place of Sugar in Modern History*. Viking, New York
Richards A I 1932 *Hunger and Work in a Savage Tribe*. Routledge, London
Richards A I 1939 *Land, Labor and Diet in Northern Rhodesia*. Oxford University Press, London
Van Esterik P 1989 *Beyond the Breast-Bottle Controversy*. Rutgers University Press, New Brunswick, NJ
Watson J L (ed.) 1997 *Golden Arches East: McDonalds in East Asia*. Stanford University Press, Stanford, CA
Weismantel M J 1988 *Food, Gender and Poverty in the Ecuadorian Andes*. University of Pennsylvania Press, Philadelphia, PA
Young M W 1971 *Fighting with Food: Leadership, Values and Social Control in a Massim Society*. Cambridge University Press, Cambridge, UK

C. M. Counihan

Food Preference

Food and food choice constitute a major part of waking human activity, and, in the Third World, the major expense. Furthermore, food choice may be the single most important selective force in animal evolution, as suggested by the fact that many animal groups are named for their food habits (e.g., carnivores, insectivores). In human evolution, adaptations to enable the exploitation of large animals as foods played a significant role in the evolution of the human brain and human social organization. As argued persuasively by Jared Diamond (1999), early advances in food procurement, agriculture, and domestication set the stage for major advances in technology, the development of urban centers, and warfare. As a fundamental and frequent part of human life, food has served as a foundation for the cultural evolution of non-food systems; thus, food plays a major role in religion, ritual, and social exchange. It is also a source of metaphors, with words such as 'bitter,' 'meat,' and 'swallow' used in many non-food contexts. Thus, an understanding of human relationships to food promises to enlighten a major area of human concern, as well as other areas. The full range of food in human life is beautifully described in Leon Kass' book, *The Hungry Soul* (1994).

1. *The Basic Relationship Between Humans and Their Food*

For any person, the world can be divided into the self and everything else. Eating involves taking matter from outside the self and putting it inside the self. This is a very intimate act. It is not surprising that people feel strongly about what they eat. The costs in terms of toxins, microorganisms, or imbalanced nutrients are high, but the benefits are at least as high: survival is at stake.

Humans (and rats and cockroaches) are food generalists. They eat a wide range of foods; virtually anything that can fit into the mouth is potential food. Generalists have few innate determinants of food choice, simply because it is not easy to predict the nutritive and toxic properties of a potential food on sensory grounds. There are only a few documented innate biases in human (and rat) food choice. First, there are innate predispositions to like or dislike certain tastes. There is an innate preference for sweet tastes, which, in nature, are predictive of calorie sources. The long history of sweetness in human culture, from fruit preferences, to cultivation of fruits, to extraction of sugar from fruits, to colonization of the Americas partly to get a source of sugar, to the development of sugar substitutes, is all driven by the innate liking for sugar.

There is also an innate tendency to dislike bitter tastes, and there are probably innate predispositions to reject very strong tastes including oral irritation (as from peppers) and to like fatty textures. Second, there is a suspicion about trying new foods (on account of potential toxicity), but also a conflicting interest in them (on account of their potential as a new nutrient source). This has been described as the generalist's dilemma. Third, there is a special learning mechanism that allows learning about the consequences of ingestion, even when these may occur hours after ingestion. Fourth, there are a few specific internal states that seem to signal the need for specific nutrients. The best example is thirst. Hunger, of course, signals a general need for energy, and there may be a specific internal signal associated with sodium deficiency.

Humans are mammals, and hence have a unique first food, milk. Humans and other mammalian generalists make a transition from this single food to a very wide range of foods. This transition is made under the guidance of parents and other caretakers, and usually consists of a scheduled and graduated introduction of new foods. For all mammal species except a subset of modern humans, milk is a food limited to infancy. The development of dairying afforded the opportunity for humans to exploit milk and its products as adult foods.

2. *Determinants of Human Food Choice*

For humans, food choice is accounted for by some combination of biological, psychological, and cultural factors. The role of biological factors is limited for two reasons. One, as discussed above, is the human generalist orientation to food. Second, the modern

developed-culture world presents to its members a food world that is vastly different from the ancestral world in which the biological predispositions evolved. A shortage of food has been replaced by a surplus of food. A modest range of natural, local foods has been replaced by an enormous range of foods, representing most of the world's cuisines at a neighborhood supermarket. A major risk of food poisoning from toxins and microorganisms has been sharply reduced; the longer lifetime that has resulted from this and other medical advances establishes a new and more subtle link between degenerative diseases and dietary patterns over decades. Limited knowledge about foods and their consequences is replaced by exposure through the media to massive amounts of evidence relating dietary pattern to wellness and disease. And finally, the linkage between energy expenditure and finding food (or finding anything else) has been almost eliminated. Thus, while the ancestral food environment focused on shortage and short-term consequences of eating, the modern environment stresses overabundance and long-term consequences.

Culture is the most powerful determinant of human food choice. More can be learned about an individual's food attitudes and preferences from his or her cultural identity than from any other single piece of information. Food choice is highly constrained by availability and cost, and both of these are primarily determined by culture. In addition, since availability and cost control exposure to food, they indirectly affect the developing of liking for foods, since that is substantially determined by exposure.

Cost and availability aside, human preferences and food attitudes can be framed by a psychological (as opposed to nutritional) taxonomy of foods, developed by Paul Rozin and April Fallon. A basic distinction can be drawn between preferences based on liking, that is on the intrinsic, sensory properties of foods, and instrumental preferences, based on the consequences of eating particular foods. There appear to be three types of reasons for rejecting or accepting a food: sensory–affective, that is, how pleasant a food tastes and smells; anticipated consequences, what the expected consequences are of eating a food; and ideational, what is known about a food (for example, where it comes from, what the nature of it is).

Food rejections can be understood in terms of the selection and interplay of these reasons. One category of rejections is called 'distaste.' These are entities rejected because of negative sensory–affective properties, such as lima beans, broccoli, beer, or chili pepper, for those who find these foods distasteful. A second category is 'danger.' These are things rejected primarily because they are believed to be harmful, because of acute or long-term consequences. The emotion of fear is often associated with their consumption. A third category of rejected things—the largest category—is called 'inappropriate.' These are things that the culture labels as inedible, such as pencils, grass, paper, or cloth. They might taste good, and might be harmless, but they are rejected for ideational reasons. The fourth category is 'disgust.' Rejection of food as disgusting is, like the previous category, based on ideational, culturally transmitted information but, unlike inappropriates, there is a strong belief that disgusts taste bad and are harmful. Unlike the affectively neutral response to inappropriates, the response to disgusts is strongly negative and emotional. Disgust is the most powerful reaction people have to food. Disgusting entities are so powerful that if they touch an otherwise acceptable food, they render it undesirable, disgusting, and inedible (the principle of contamination or contagion; see *Disgust, Psychology of*).

On the positive side, there are four comparable categories. Good taste (acceptance principally because of sensory properties), beneficial (acceptance largely because of consequences), appropriate (acceptance because it is culturally designated as food, or food for a particular occasion), and transvalued (food enhanced because of its prior history). The transvalued category is much weaker and smaller than the disgust category, in most cultures. In Hindu India, food that has been 'shared' with the gods (via donation to the priests in the temple, and then returned, in part), called *prasad*, is an example of transvaluation.

3. *The Acquisition of Food Preferences*

For newborn infants, the only functioning categories are good taste (e.g., sweet) and distaste (e.g., bitter). Generally, infants will place in their mouth anything that might fit in, including feces and potentially toxic foods. They reject only on the basis of sensory properties. Gradually, they acquire distastes or good tastes, and learn about danger and beneficial foods. It is much later that the full manifestations of disgust appear. The full adult categorization is in place by roughly five to eight years of age.

Most is known about the distinction between distaste and danger. When ingestion of a food is followed by nausea, it tends to become disliked, that is, a distaste. However, if ingestion of a food is followed by most other negative symptoms (e.g., lower gut pain, skin rash, respiratory distress), the food typically becomes a danger. That is, people reject a food that has caused such symptoms, but it does not usually become a disliked taste. This distinction has also been demonstrated in the laboratory with rats. With respect to affect and emotion, it is notable that dislikes (distastes) and dangers have very different properties, although the outcome (rejection) is the same. The nausea-based acquired distaste (often called a conditioned taste aversion), unlike dangers, is not based on a legitimate sense of danger. Even if a person knows that the nausea/upper gastrointestinal illness was not produced by the food, the aversion remains.

Thus, people who get nauseous and often vomit after a meal usually develop an aversion to some food in the meal, even if they know that the illness was simply the onset of influenza.

The acquisition of good tastes is more complex and less understood. Mere exposure to a food, in itself, often seems sufficient to produce an acquired like. In addition, the pairing of a food with an already positive event (an already-liked food mixed with it, positive regard by a respected person, a pleasant environment), by a process called evaluative conditioning, can lead to acquired likes (or acquired dislikes, if the paired events are negative, as in conditioned taste aversions). Leann Birch has demonstrated that indications of liking by a significant other (peer, older child, teacher, parent) may cause acquisition of liking. The process at work here is not understood. It could be a form of evaluative conditioning. But it also may involve an important instance of communication of affect or emotion. The expressed pleasure by a significant other, on consuming a food, may directly induce a pleasant state in an observer, or it may induce a mimicked positive facial expression. Either of these responses may cause enhancement of liking. Birch has demonstrated, in the laboratory, that efforts by adults to promote liking by emphasizing the beneficial consequences (better health, a specific reward for eating) seem to block the acquisition of liking for the food. It seems that when a child observes respected others enjoying a food, this promotes liking; when she is rewarded for consuming it, this seems to block the acquisition of liking.

The acquisition of knowledge of appropriate and inappropriate foods seems to be largely cognitive, and affect free. The affect-laden acquisition of disgust contrasts with this. Some communication of affect (facial and other) is almost certainly involved, but there are also important cognitive aspects; it is the nature of the foods and their history that is central to this category. Disgusting foods have contamination properties. This property requires a realization that appearance does not equal reality: a contaminated food looks like any other food, but is distinguished by its history. Young children have difficulty making the appearance–reality distinction, and it is only by age four to seven that we see signs of the ability to appreciate contamination, and hence disgust.

Individuals in specific cultures develop likings (the good taste category) for some foods that are innately unpalatable (e.g., bitter, very strong tasting, irritating). Innately unpalatable foods are typically among the favorite foods: chili pepper, black pepper, ginger and other irritant spices, coffee, bitter chocolate, tobacco, alcohol, burnt foods, and highly salted foods. Such reversals of innate aversions are common in humans, and rare in animals. We do not know how these preference reversals occur. In part, it may be through the same processes that produce normal likes: mere exposure, evaluative conditioning, and social approval. But there may also be special mechanisms that are involved in preference reversals. One is that innately negative foods may come to produce the secretion of endorphins, chemicals secreted in the brain in response to pain and irritation. These may modulate the displeasure, and like their pharmacological analog, morphine, can produce pleasure. Normally, one ceases to interact with a bad-tasting food. However, cultural forces continually reintroduce innately unpalatable foods to the child. It may be that with repeated exposures, the body endorphin response becomes stronger, and eventually cancels and then overwhelms the displeasure. Another account is that it is the very displeasure that is the source of the pleasure. This places liking for innately unpalatable foods in the same category as thrill-seeking activities such as roller coaster riding. That is, humans get pleasure out of situations in which innate aversions or fears are stimulated, but in which there is no real danger. Roller coasters are safe, and so is chili pepper. This may be case of pleasure derived from 'mind-over-matter.'

4. *Family Influences and the Family Paradox*

There are strong arguments for high parent–child resemblance in food habits. There is the common genetic heritage, the fact that parents control access (and hence exposure) to foods, and that the principal affective signals about foods in early life come from the parents. Yet parent–child correlations in liking for foods are very low, in the range of 0 to 0.3, in comparison to parent–child correlations in values (such as attitudes to abortion), temperament, or abilities. These low parent–child correlations appear whether the children studied are four-year-olds or college students. This family paradox cannot presently be explained. However, it suggests that the second six years of life, when children are much more heavily influenced by peers and teachers, as opposed to parents, may be more critical for establishing adult preferences than the first six years.

5. *Preferences and Values: Moralization and Vegetarianism*

Selecting and consuming a food has social and moral implications, as well as physiological consequences. The source of a food (e.g., who prepared it, animal or vegetable origin) can powerfully influence acceptance. Particular food or food product boycotts in recent decades testify to this fact. Hindu Indian concerns about food fall largely within the moral domain. In the United States, there is a long history linking particular foods with the good life or immorality. Through the process of moralization, choice of a particular food may move from the domain of preferences to the domain of values. This has clearly happened with cigarette smoking in the United States, and it fre-

quently happens in the course of developing vegetarianism. For some vegetarians, eating of animals is an immoral act. When ingestion of a food becomes moralized, censure of consumers becomes licensed and governments and institutions may take limiting or prohibiting actions. Such foods often become offensive (disgusting), and the rejection becomes intrinsic as opposed to instrumental.

6. *Food, Fear, and Pleasure: Cultural Perspectives*

In the non-human animal world, and among traditional humans, food is basically a source of pleasure. Even in the developed world, cost and availability aside, the principal factor influencing food choice is taste, that is to say, pleasure. Convenience and health issues are also important determinants. However, with the major changes in the human environment in the First World, obesity and the consequences of long-term exposure to particular diets have become more salient. As a result, the earlier simple pleasure of eating has given way, for many people, to ambivalence, or even outright fear. This is most well developed in American women, who often fear food as much as they savor it. In some, there is now embarrassment and guilt about consuming high fat foods, great concern about thinness, and great concern about the healthfulness of every bite of food. This is not a necessary consequence of the modern environment since, in France, the pleasures of eating still dominate.

See also: Disgust, Psychology of; Eating Disorders, Determinants of: Genetic Aspects; Food in Anthropology; Food Production, Origins of; Hunger and Eating, Neural Basis of; Taste and Smell, Psychology of; Taste and Taste Culture; Taste, Neural Basis of

Bibliography

Barker L M (ed.) *The Psychobiology of Human Food Selection*. AVI, Bridgeport, CT

Birch L L, Fisher J O, Grimm-Thomas K 1996 The development of children's eating habits. In: MacFie H, Meiselman H (eds.) *Food Choice, Acceptance and Consumption*. Blackie, Glasgow, UK, pp. 161–206

Booth D A 1994 *Psychology of Nutrition*. Taylor & Francis, London

Diamond J 1999 *Guns, Germs, and Steel*. Norton, New York

Fischler C 1990 *L'homnivore*. Editions Odile Jacob, Paris

Kass L 1994 *The Hungry Soul*. Free Press, New York

Logue A 1991 *The Psychology of Eating and Drinking*, 2nd edn. Freeman, New York

MacFie H, Meiselman H (eds.) 1996 *Food Choice, Acceptance and Consumption*. Blackie, Glasgow, UK

Rozin P 1982 Human food selection: The interaction of biology, culture and individual experience. In: Barker L M (ed.) *The Psychobiology of Human Food Selection*. AVI, Bridgeport, CT, pp. 225–54

Rozin P 1990 The acquisition of stable food preferences. *Nutrition Reviews* **48**: 106–13

Rozin P 1998 *Towards a Psychology of Food Choice*. Danone Chair Monograph. Institut Danone, Brussels, Belgium

Rozin P 1999 Food is fundamental, fun, frightening, and far-reaching. *Social Research* **66**: 9–30

Rozin P, Fallon A E 1987 A perspective on disgust. *Psychological Review* **94**: 23–41

Shepherd R, Raats M M 1996 Attitudes and beliefs in food habits. In: MacFie H, Meiselman H (eds.) *Food Choice, Acceptance and Consumption*. Blackie, Glasgow, UK, pp. 346–64

P. Rozin

Food Production, Origins of

A bird's eye view of the human career from inception to now reveals four major landmarks: origins of hominids; origins of *Homo sapiens sapiens*; origins of food production; origins of urbanized, state-based civilizations. Like the other categories of origins research, the emergence of the first farmers and herders has long attracted attention from archaeologists, anthropologists, historians, and other scholars. Early in the twentieth century, a prominent archaeologist, Vere Gordon Childe, discussed the origins of agriculture in arid western Asia at the end of the Ice Age (Childe 1936, Chap. V 'The Neolithic Revolution,' and 1942, Chap. III 'Neolithic Barbarism'). It was not Childe, however, but another archaeologist who set himself the task of actually digging up evidence for the food producing revolution in western Asia.

1. *The Origins of Food Production in Western Asia*

Shortly after the end of World War II, Robert J. Braidwood (1960) organized the Iraq-Jarmo Project aimed at excavating the oldest agropastoral community known anywhere in the world at that time: Jarmo in Iraqi Kurdistan. Braidwood's expeditions to Jarmo between 1948 and 1955 are famous for their interdisciplinary nature, and for his suggestion that food production in western Asia began in environments like that of Jarmo, a situation he described as 'the hilly flanks of the Fertile Crescent.' He was referring to the foothill and piedmont zone, extending from the Levantine coast along the southern slopes of the Taurus and down the westerly face of the Zagros Mountains, where wild species ancestral to the earliest plant and animal domesticates were distributed.

Following Braidwood's example, interdisciplinary, international archaeological teams undertook intensive research on agricultural origins in the hilly flanks and other regions of western Asia. The work is ongoing today wherever political conditions permit; hence the evidential base is continually changing. At present, however, it appears that the first recognizably domestic food plants are small grains (rye, wheat, barley) and legumes (lentils, peas, vetches) from sites in the Levant, and in the hilly flanks zones of northwest Syria, southeastern Anatolia, and northern Iraq. Archaeobotanist Daniel Zohary (1996) thinks that most of the founder crops (wheat, barley, and legumes) were domesticated only once or twice, and then spread both east and west. The first domesticated food animals are sheep and goats from sites in the Zagros mountain portion of the hilly flanks, pigs and cattle appearing later. Although the archaeological and paleoenvironmental data are by no means unambiguous, most scholars invoke climatic factors prominently (especially fluctuations between warmer and colder intervals in the aftermath of the last Ice Age) to explain how and why food production arose in western Asia (for further details see Bar-Yosef and Meadow 1995, Miller 1992, Harris 1998, Hillman 2000).

2. *The Origins of Food Production in other World Areas*

Systematic research on agropastoral origins is now widespread around the globe. Interdisciplinary archaeological projects are recovering evidence for early food production in Africa, South and East Asia, New Guinea, and the Americas. At present, it appears that there may have been no more than seven independent trajectories to agriculture, or agropastoralism during the Early and Middle Holocene (i.e., between 10,000 and 5,000 years ago). A brief global survey follows.

2.1 *Africa*

The lower and middle Nile Valley portion of northeastern North Africa is linked to agropastoral developments in western Asia. Domestic species of wheat, barley, sheep, and goat were derived from the Levant by about 7,000 years ago, and were combined with the earlier gathering, hunting, fishing economy. It is likely, however, that cattle were domesticated in the eastern Sahara from local populations of wild *Bos primigenius* by 9,000 years ago. Many of the early herders of cattle, sheep, and goats were probably also systematically harvesting wild species such as sorghum, *Cenchrus* (krem krem), and the oil plant *Canarium*, but in ways (e.g., beating into trays or baskets) that did not result in morphologically identifiable domesticates.

On the present scanty evidence, there are no domesticated plants in Africa outside the Nile Valley prior to about 4,000/3,500 years ago. Domestic pearl millet (*Pennisetum*) of that date is reported from archaeological sites across central Africa, probably introduced by Saharan herders. Other plants believed to have been first domesticated in Africa include African rice, sorghum, finger millet, tef, enset, and noog. Domestic sorghum and rice are present at the site of Jenne-Jeno in Mali about 2,000 years ago, whereas finger millet, tef, noog, and enset are East African plants first domesticated in the Ethiopian/Eritrean region (by ca. 1,200 years ago for finger millet; no data as yet for tef, enset, and noog).

There are several other plants, including a number of roots and tubers, that were used intensively, extensively, or both in various parts of sub-Saharan Africa, with little or nothing in the way of archaeological histories as yet. In South Africa, agriculture and pastoralism are thought to have arrived in the early Iron Age about 2,000 years ago with the spread of food producers into this part of the continent (for further details see Harlan 1992, Marshall 1998, Wetterstrom 1993, 1998).

2.2 *South Asia*

The origin and spread of agriculture and pastoralism in South Asia are complex, long-term processes beginning about 9,500 years ago. Most experts think domesticated wheat and barley as well as sheep and goats were introduced from western Asia, but humped cattle (*Bos indicus*) are thought to have been domesticated indigenously by 7,000 years ago, as perhaps were water buffalo.

Agriculture in the northerly portion of South Asia was altered significantly beginning about 4,500 years ago when a series of summer crops were introduced and grown in various combinations with the winter wheat and barley: finger millet, pearl millet, and sorghum, all from Africa; broomcorn millet, foxtail millet, and rice, all probably from East Asia. One other millet was apparently indigenously domesticated by about this same time: little millet (for further details see Meadow 1998).

2.3 *East Asia*

A great deal of archaeological work has been carried out at early village sites in China since World War II, but only very recently have explicit, systematic efforts been made to obtain archaeobotanical and zooarchaeological materials from them. Most of the present evidence comes from South China and centers upon rice. Of the 14 direct (AMS) radiocarbon dates on rice remains reported by Crawford and Shen (1998), the oldest is 9,000 years bp ('bp' = before present; this is a calibrated date) at the site of Hemudu (on the east coast of China south of the Yangtze delta) with 9,000 to 8,000 bp determinations on rice grains at Pengtoushan (in Hunan Province), and the northerly site of

Jiahu (Henan Province). Crawford and Shen say that the people who lived in these three early sites were harvesting rice, but whether it had been planted, or was being reaped from wild stands is unknown. In any case, it appears that rice was probably first domesticated somewhere in the general region of the middle Yangtze River drainage, and domestic rice may have spread to Korea and Japan between 6,000 and 3,000 years ago.

The other indigenously domesticated Chinese grain crops are foxtail millet (believed to have been domesticated by ca. 7,500 (calibrated) years ago at one site in Hebei Province, and at three sites in Henan Province); and broomcorn millet, domesticated prior to 8,500 years ago in the Yellow River drainage. On the present evidence, rice use (and domestication) in the central Yangtze drainage of southern China seems to predate the appearance of domestic millet in the Yellow River drainage.

Even less is known about the origins of domestic food animals in China than about domestic plants, but dogs, pigs, and probably domestic chickens seem to be present in North China sites at about the same time as the millets. Although the data are not very detailed, the origin of food production in China appears to have entailed processes somewhat more like those of western Asia (i.e., a suite of founder species, which then spread widely) than like the patchier, mosaic pattern of Africa and South Asia (for further details see Cohen 1998, Crawford 1992, Crawford and Shen 1998).

2.4 *Southeast Asia and the Islands of the Southwest Pacific*

Although much of the older literature on the development of Southeast Asian agriculture focuses heavily on rice (now thought to have diffused there from China's central Yangtze River region), there was a complex and important array of early tuber, root, and tree crops in Southeast Asia, New Guinea, and the islands of the southwestern Pacific. Evidence of these is difficult to retrieve archaeologically, and much of the discussion about them revolves around ethnographic, ethnobotanic, and linguistic data as well as delineation of prehistoric agricultural terraces, ditches, and other landscape features. A strong case has been made for landscape modification in New Guinea, presumably in the interests of growing taro and native yam species, long preceding the ca. 3,500 bp expansion of people from Southeast Asia onto the Pacific islands, accompanied by domestic animals (probably dogs, chickens, and pigs) and an array of domesticated tuber and root species (see *Pacific Islands, Archaeology of*). There is, however, considerable disagreement about the exact nature of pre-3,500 bp highland New Guinea agriculture, and how significant it was prior to the arrival of Southeast Asian derived agricultural systems. (Hence, although some scholars regard New Guinea as one of the world areas that witnessed an independent transition to agriculture in prehistory, that claim is still being debated. For further details see Bellwood 1996, Golson 1989, Hather 1996, Spriggs 1996).

2.5 *The Americas: Mexico*

Maize is the most famous crop of the Americas, and by far the most complex of the world-class grains domesticated in prehistory. Although its wild ancestor (teosinte) and its origin place (the Balsas River drainage in western Mexico) are generally agreed upon, the exact details and the chronology of initial domestication are still under debate. AMS dates on early maize from excavations directed by Richard MacNeish in the 1960s at two rockshelters in the Tehuacan Valley, Puebla, and from two other rockshelters in the Ocampo region of Tamaulipas have established that none of these specimens is more than about 5,000 years old, previous estimates having ranged from 7,000 to 5,000 bp or even older. Some researchers, persuaded by seemingly early pollen and phytolith data, continue to advocate very early estimates for domestic maize in spite of concerns by other scholars about the dating of these microremains. Similarly, there is lack of consensus on how, and how rapidly, domesticated maize spread north and south from its place of origin.

Archaeological work at the Mexican sites noted above, plus excavations by Kent Flannery in 1966 at another dry rockshelter—Guila Naquitz in Oaxaca—have resulted in information about several other cultigens: squashes, bottle gourd, beans, and chili peppers. Also present at some of these sites are *Amaranthus*, *Chenopodium*, *Agave*, and several root crops and tree fruits thought to have been encouraged and harvested if not actually domesticated in early prehistory. Recent AMS dates on squashes from MacNeish's excavations in Tehuacan and Tamaulipas (*Cucurbita pepo*, *C. argyrosperma*, *C. moshata*) and from Flannery's work at Guila Naquitz (*C. pepo* gourds) have returned 9,000 bp determinations for *C. pepo* there, and 4,000 to 5,000 bp for the other two cucurbits. No fieldwork on early Mexican agriculture comparable to MacNeish's and Flannery's has taken place since the 1960s (for further details see Fritz 1994, Piperno 1994, Smith 1997a, 1997b).

2.6 *The Americas: Central and South America*

Origins research on early agriculture in the American lowland neotropics is being actively pursued, with new information emerging at a rapid rate. The foci of attention in the lowlands are tubers and root crops (including potatoes, sweet potatoes, yams, and manioc), as well as tree fruits, pineapple, chili peppers, avocados, maize, squashes, beans, cotton, and bottle gourd. Because of preservation problems, macro-

botanical remains are less likely to be recovered from humid regions than from arid locales like those of the Mexican rockshelters. Hence, several researchers have turned to pollen, phytoliths, and starch grain analyses. This is important, innovative work, but does pose serious problems for secure dating, and—because these trajectories are so new—for consensus in identifications and interpretations.

The Andean highlands of South America are very different from the tropical lowlands, but similar in that the origins of food production there are archaeobotanically documented only preliminarily). The crops centrally involved are *Chenopodium quinoa*, squashes and gourds, beans, cotton, maize, and several tuber species including the potato, *Solanum tuberosum*. Early animal domesticates are llamas, alpacas, and guinea pigs. Chronologies and other details are not precise, but quinoa may have been domesticated by 5,000 to 4,000 years ago, gourd and squash perhaps somewhat earlier, common beans by 4,400 bp (calibrated). There are no direct AMS dates for prehistoric finds of possibly domestic potatoes, and the chronology for llama/alpaca and guinea pig domestication is also vague. Estimates for domestication of all three animal species range around the same time as for quinoa, and the potato may have been domesticated about the same period or even earlier.

The arid west coast of South America in Peru and Chile has its own food-producing story to tell, related to that of the highlands in some ways, but with significant differences. Preservation of even very delicate plant parts as well as textiles and human remains makes the archaeology attractive to professionals and to looters. Evidence for plant use begins early at some sites, perhaps as far back as the Early Holocene for gourds, and includes many of the same species referred to above for other areas of South America. The coastal valleys of Peru currently provide the oldest directly dated evidence for beans (one lima bean pod) at ca. 5,600 bp (calibrated); Mexican common and tepary beans are no older than 2,500 bp (calibrated) (for further details see Pearsall 1992, 1995, Piperno et al. 2000, Piperno and Pearsall 1998).

2.7 The Americas: Eastern North America, with Comparative Reference to the Southwestern United States

Since 1980, archaeological and paleoethnobotanical research has shown that an indigenous agricultural system was created in Eastern North America beginning about 4,500 years ago. Sunflower, sump weed, goosefoot, may grass, bottle gourd, and pepo gourd are the initial members of the complex at 3,000 bp, with knotweed, little barley, and tobacco being added several hundred years later. Maize first appears in Eastern North America 2,000 years ago, but is a very minor component of the archaeobotanical record until ca. 1,200 bp when it became a major staple in communities of the Ohio and Mississippi River drainages. Common beans enter the picture even later (ca. 700 bp). The two gourds are present in a few places 7,000 years ago—well before the rest of the complex appears—but it is not clear whether they were domesticated that early, or were simply being harvested from wild plants. In any case, they were probably domesticated separately from parallel processes in Mesoamerica and South America, just as *Chenopodium* seems to have been domesticated separately in Eastern North America, Mexico, and South America. Fritz (1999) has argued persuasively that the early solo presence of cultivated or domesticated gourds of whatever kind does not mean the gourd cultivators were farmers. Full-scale agriculture may appear much later, as seems to be the case in several places within the Americas.

The archaeobotanical database in Eastern North America—arguably the best in the world—demonstrates use of wild plant and animal foods throughout the entire sequence from early Middle Holocene to the time of European contact. Domesticates appear in differing combinations and at different levels of intensity across space and time in an intricate mosaic.

Formerly, it was thought that the Mexican agricultural triad of maize, squash, and beans diffused north into the southwestern US, then much later moved across the Plains to Eastern North America. It is now abundantly clear, however, that the Southwest and the East were on separate trajectories, each area experiencing its own complex set of developments. Maize did arrive in the Southwest earlier than it appeared in the East (pre-3,000 bp vs. 2,000 bp) with squash and beans arriving perhaps about the same time, but these plants were adopted into very different kinds of pre-existing Archaic societies across the region: some small, mobile, and staying that way, others already somewhat sedentary and becoming more so. In all cases, considerable use was made of indigenous southwestern plants, some domesticated, others semi-domesticated or encouraged, and others harvested from wild stands. The only domesticated mammal was the dog, but turkeys were kept in some places. Hunting of deer, antelope, rabbits, and other small game was very important everywhere throughout the entire prehistoric sequence (for further details see Fritz 1990, Minnis 1992, Smith 1992, Yarnell 1994).

3. Discussion and Conclusions

There is a vast literature on agricultural origins and the development of prehistoric food-producing economies. The actual corpus of empirical, archaeological, archaeobotanical, zooarchaeological evidence about early food production is much, much smaller. Only in Western Asia and Eastern North America is there sufficient data to constrain conjecture in any signifi-

cant way. Nevertheless, it is clear that humans, plants, and animals everywhere were always interdependent in multiple ways, and that there were many specific pathways to food production. The areal sketches provided in the previous section indicate how diverse those pathways could be. The present evidence, viewed from a high level of abstraction, suggests that there were seven world areas where indigenous and distinctive food-producing economies were developed: Western Asia, Saharan and sub-Saharan Africa (perhaps with West African and East African foci), East Asia (perhaps with separate foci in North China and South China), Eastern North America, Mexico, Central and South America (with neotropical and highland foci), and—intriguing but enigmatic—highland New Guinea.

Discussion of early crops in all these places centers heavily upon those that dominate modern world economies (wheat, rice, maize), sometimes to the detriment of understanding the nature and functioning of ancient subsistence systems, whether food producing or not. Given present concerns about the grave dangers of monocropping on an areal, let alone a global, scale, this tendency is doubly unfortunate. Lost crops, lost breeds, and lost varieties of current crops and domestic animals are vital resources for modern agropastoralism, but can be retrieved only through the interdisciplinary work of archaeologists, paleoethnobotanists, zooarchaeologists, on the one hand, and molecular biologists and plant and animal breeders on the other.

Two more issues central to any consideration of ancient food-producing economies are chronology and explanation. Was the origin of food production a nearly simultaneous event around the world triggered by Late Pleistocene/Early Holocene climatic change? Were the causes and processes of all six or seven independent, indigenous transitions to food production essentially the same? How one answers these questions obviously depends almost entirely on how detailed and how secure areal chronologies and areal databases may be. There is, at best, only semi-security for two of the pristine-origins candidates. Nevertheless, one can argue a negative response to both questions. On the evidence of present chronologies, created and continuously refined via heavy dependence on AMS radiocarbon dating, it appears that only in West Asia and East Asia was there significant impact by end Pleistocene/Early Holocene climatic changes to impel intensification of the food-producing sort. In both places, food production economies seem to be phenomena of the earliest Holocene, taking shape in the first two millennia after the end of the last ice age, and spreading relatively rapidly, almost in package form. Elsewhere, when sufficient information is available, chronologies for the origins of food production are longer, with food-producing systems being more fluid and patchy or mosaic-like, varying considerably through time and across space. David Harris (1996, pp. 557, 568) contrasts the temperate West Asian package (small grains, legumes, sheep, goat) with the vegecultural systems (roots, tubers, rhizomes) nutritionally (the latter is heavy on carbohydrates, the former also includes fat and protein) and territorially (the former is expansive, the latter is not). Somewhat similarly but from a global perspective, Jared Diamond (1997) points out the great contrasts among geography, flora, and fauna that contextualize subsistence systems in different parts of the world.

Prime movers used to aid explanation of transitions to food production range from the older favorites, climatic or environmental change (see above) and population pressure (e.g., Cohen 1977) to newer models drawn from economics and evolutionary ecology, to competition among lineage or clan heads, or to the necessity of intercommunity trade for forging alliances guaranteeing mutual assistance in hard times (e.g., Bender 1978, Hayden 1996, Price and Brown 1985). Some of these models are quite ingenious, and plausible for specific origins cases, but they do not fit others at all, making them less than satisfactory as general explanations for *the* origin of food production in the human past. That abstraction is very useful for certain purposes, but self-destructs when placed in contact with all the evidential bases now available, meager though many of them are. Although we do not know nearly enough, we do know too much, and what we know is changing too rapidly, to sustain current general explanations for why food production happened in prehistory. We know, for example, that gatherer–hunter–fisher–foragers always manipulated and managed their physical and biological environments, that they often cultivated and/or harvested wild plants, tamed wild animals, and restructured the topography as well as the native plant and animal distributions in their home territories; and that they were quite capable of intensifying any or all of these management procedures if they perceived intensification to be necessary or desirable. We know that in some places and times the kind of intensification we call food production was implemented, but that mostly it was not; that gatherer–hunters could and did coexist for long periods (centuries, millennia) with food producers; and that food producers in some times and places became gatherer–hunters as well as vice versa.

In spite of this chronic mismatch between theory and data, inevitable in a field as broad and lively as agricultural origins research, it is essential that high-level, theoretical formulations be devised and tested against independent relevant data, explicitly sought if not already available, because that is the most efficient means to expand knowledge about what, in the abstract, is indeed the most important transition in human prehistory.

See also: Agricultural Change Theory; Cultural Evolution: Overview; Food in Anthropology; Food Security; Hunter–Gatherer Societies, Archaeology of;

Hunting and Gathering Societies in Anthropology; Pacific Islands, Archaeology of; Pastoralism in Anthropology

Bibliography

Bar-Yosef O, Meadow R H 1995 The origins of agriculture in the Near East. In: Price T D, Gebauer A B (eds.) *Last Hunters, First Farmers: New Perspectives on the Prehistoric Transition to Agriculture*. SAR Press, Santa Fe, NM

Bellwood P 1996 The origins and spread of agriculture in the Indo-Pacific region: Gradualism and diffusion or revolution and colonization? In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London

Bender B 1978 From gatherer–hunter to farmer: A social perspective. *World Archaeology* **10**: 204–22

Braidwood R J 1960 The agricultural revolution. *Scientific American* **203**: 130–48

Childe V G 1936 *Man Makes Himself*. Watts, London

Childe V G 1942 *What Happened in History*. Penguin, Harmondsworth, UK

Cohen D 1998 The origins of domesticated cereals and the Pleistocene–Holocene transition in East Asia. In: Bar-Yosef O (ed.) The Transition to Agriculture in the Old World. *The Review of Archaeology* **19**(2): 22–9

Cohen M 1977 *The Food Crisis in Prehistory*. Yale University Press, New Haven, CT

Crawford G W 1992 Prehistoric plant domestication in East Asia. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture, an International Perspective*. Smithsonian Institution Press, Washington, DC

Crawford G W, Shen C 1998 The origins of rice agriculture: Recent progress in East Asia. *Antiquity* **72**: 856–66

Diamond J M 1997 *Guns, Germs and Steel: The Fates of Human Societies*. W W Norton, New York

Fritz G J 1990 Multiple pathways to farming in precontact Eastern North America. *Journal of World Prehistory* **4**: 387–435

Fritz G J 1994 Are the first American farmers getting younger? *Current Anthropology* **35**: 305–9, 639–43

Fritz G J 1999 Gender and the early cultivation of gourds in Eastern North America. *American Antiquity* 417–29

Golson J 1989 The origins and development of New Guinea agriculture. In: Harris D R, Hillman G C (eds.) *Foraging and Farming: The Evolution of Plant Exploitation*. Unwin Hyman, London

Harlan J R 1992 Indigenous African agriculture. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture, an International Perspective*. Smithsonian Institution Press, Washington, DC

Harris D R 1996 The origins and spread of agriculture and pastoralism in Eurasia: An overview. In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London

Harris D R 1998 The origins of agriculture in Southwest Asia. In: Bar-Yosef O (ed.) The Transition to Agriculture in the Old World. *The Review of Archaeology* **19**(2): 5–11

Hather J G 1996 The origins of tropical vegeculture: Zingiberaceae, Araceae and Dioscoreaceae in Southeast Asia. In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London

Hayden B 1996 Pathways to power: Principles of creating socioeconomic inequalities. In: Price T D, Feinman G M (eds.) *Foundations of Social Inequality*. Plenum, New York

Hillman G C 2000 Overview: The plant-based components of subsistence at Abu Hureyra 1 and 2. In: Moore A M, Hillman G C, Legge A J (eds.) *Abu Hureyra and the Advent of Agriculture*. Oxford University Press, Oxford, UK

Marshall F 1998 Early food production in Africa. In: Bar-Yosef O (ed.) The Transition to Agriculture in the Old World. *The Review of Archaeology* **19**(2): 47–58

Meadow R H 1998 Pre- and proto-historic agricultural and pastoral transformations in northwestern South Asia. In: Bar-Yosef O (ed.) The Transition to Agriculture in the Old World. *The Review of Archaeology* **19**(2): 12–21

Miller N F 1992 The origins of plant cultivation in the Near East. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture, an International Perspective*. Smithsonian Institution Press, Washington, DC

Minnis P E 1992 Earliest plant cultivation in the desert borderlands of North America. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture, an International Perspective*. Smithsonian Institution Press, Washington, DC

Pearsall D M 1992 The origins of plant cultivation in South America. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture: An International Perspective*. Smithsonian Institution Press, Washington, DC

Pearsall D M 1995 Domestication and agriculture in the New World tropics. In: Price T D, Gebauer A T (eds.) *Last Hunters, First Farmers: New Perspectives on the Prehistoric Transition to Agriculture*. SAR Press, Santa Fe, NM

Piperno D R 1994 On the emergence of agriculture in the New World. *Current Anthropology* **5**: 637–9

Piperno D R, Pearsall D M 1998 *The Origins of Agriculture in the Lowland Neotropics*. Academic Press, San Diego, CA

Piperno D R, Ranere A J, Holst I, Hansell P 2000 Starch grains reveal early root crop horticulture in the Panamanian tropical forest. *Nature* **407**: 894–7

Price T D, Brown J A 1985 *Prehistoric Hunter-Gatherers: The Emergence of Cultural Compexity*. Academic Press, Orlando, FL

Smith B D 1992 *Rivers of Change: Essays on Early Agriculture in Eastern North America*. Smithsonian Institution Press, Washington, DC

Smith B D 1997a The initial domestication of *Curcurbita pepo* in the Americas 10,000 years ago. *Science* **276**: 932–4

Smith B D 1997b Reconsidering the Ocampo caves and the era of incipient cultivation in Mesoamerica. *Latin American Antiquity* 342–83

Spriggs M 1996 Early agriculture and what went before in Island Melanesia: Continuity or intrusion? In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London

Wetterstrom W 1993 Foraging and farming in Egypt: The transition from hunting and gathering to horticulture in the Nile Valley. In: Shaw T, Sinclair P, Andah B, Akpoko A (eds.) *The Archaeology of Africa: Food, Metals and Towns*. Routledge, London

Wetterstrom 1998 The origins of agriculture in Africa, with particular reference to sorghum and pearl millet. *The Review of Archaeology* **2**: 30–46

Yarnell R A 1994 Investigations relevant to the native development of plant husbandry in eastern North America: A brief and reasonably true account. In: Green W, Arzigian C M

(eds.) *Agricultural Origins and Development in the Mid-continent*. Office of the State Archaeologist, University of Iowa, Report 19, Iowa City, IA
Zohary D 1996 The mode of domestication of the founder crops of southwestern Asian agriculture. In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London

P. J. Watson

Food Security

Food is undoubtedly one of the most basic human needs. Access to enough food to lead an active and healthy life is also one of the most fundamental human rights. Although the global food supply is generally satisfactory, the basic food needs of hundreds of millions of people are not satisfied, and the right to food is widely neglected. This article defines food security, assesses the current and future world food situation, and presents theoretical and conceptual approaches to analyze food security from the perspective of the social and behavioral sciences.

1. *Definition of Food Security*

Although nutritional scientists distinguish between 'food security' (availability of food on global, national, local, and household levels), on the one hand, and 'nutrition security' (satisfactory nutritional status of individuals), on the other (Oltersdorf and Weingärtner 1996, p. 23), social and behavioral scientists tend to consider 'food security' as a more comprehensive term that incorporates both concepts. This view is evident from most of the definitions available on food security, not only in social science, but also in official GO and NGO documents (see M. Smith et al. 1992). The following World Food Council (WFC) definition may serve as an example:

> Food security implies two things. First, it implies that food is available, accessible, affordable—when and where needed—in sufficient quantity and quality. Second, it implies an assurance that this state of affairs can reasonably be expected to continue; or in other words, that it can be sustained. To put it simply, food security exists when adequate food is available to all people on a regular basis (World Food Council 1988, p. 2).

This definition will serve as the base for the following discussion on social and behavioral aspects of food security.

Along the line of this definition, food insecurity implies problems with the availability, accessibility, and affordability of food. This links the concept of 'food insecurity' to the notion of the 'food system' which is widely used in the discussion of food (in)security (Cannon 1991). Food systems are defined as integrated systems of the four subsystems of production, exchange, delivery, and consumption (Cannon 1991, p. 298).

It is important to distinguish between two kinds of food insecurity, chronic and transitory. The difference between the two is emphasized by a World Bank definition (1986, p. 1):

> Chronic food insecurity is a continuously inadequate diet caused by the inability to acquire food. It affects households that persistently lack the ability either to buy enough food or to produce their own. Transitory food insecurity is a temporary decline in a household's access to enough food. It results from instability in food prices, food production, or household incomes—and in its worst form it produces famine.

2. *The World Food Situation*

A comprehensive assessment of the world food situation by experts from the International Food Policy Research Institute (Pinstrup-Andersen et al. 1997) addresses recent developments, emerging issues, and long-term prospects, and also distinguishes between production, distribution, and consumption issues. On the one hand, in the past couple of years, developments in global food supply, demand, and trade have raised concerns about the world's future food supply. Prices of grain rose rapidly, and at the same time global cereal stocks fell sharply. Many countries of Eastern Europe and the former Soviet Union failed to make significant advances in economic transition and agricultural development. Moreover, flows of food aid and official development finance to developing countries declined. In sub-Saharan Africa, the population growth rate has exceeded the rate of growth in food production considerably. Weather fluctuations and climate change could lead to sizeable food production shortfalls and deterioration in food security in many parts of the world. Declining natural base, first of all growing water scarcity and declining soil fertility, put severe threats on future agricultural development. Stagnating income levels of the poor, high rates of underemployment, and severe cuts in the social security budgets put further stress on the food security of the poor and vulnerable sections of society. South Asia alone is home to around 50 percent of the world's poor and food insecure (ACC/SCN 1997). Moreover, there are escalating concerns about genetic engineering and food quality.

On the other hand, other developments offer glimmers of hope. The United Nations, for example, has once again revised its population projections downward, thus reducing expected pressures on future food supplies. Progress has also been made on international trade liberalization, although developing countries do not equally benefit as developed countries. The World

Food Summit of November 1996 has raised awareness for world food security problems and is expected to stimulate commitments for action.

Projections of food production, distribution, and consumption to the year 2020 offer some signs of progress, but prospects of a food-secure world, a world in which each and every person is assured of access at all times to the food required to lead a healthy and productive life, remain bleak if the global community continues with business as usual (Pinstrup-Andersen et al. 1997, p. 8). Projections for the future world food situation under several scenarios show that, under the most likely scenario, 150 million children under the age of six years will be malnourished in 2020, just 20 percent fewer than in 1993. Projections by the Food and Agricultural Organization (FAO) on the number of food insecure people paint a similarly mixed picture. FAO projects that 680 million people could be food insecure in 2010, down from 840 million in 1990–2. Worldwide, per capita availability of food is projected to increase around 7 percent between 1993 and 2020. However, the future world food situation will be significantly influenced by issues such as rapid urbanization, declining resource base, policy-induced draw down of cereal stocks, and greater variability in agricultural production caused by climate change. Widespread conflict and political instability are increasingly adding to food insecurity in many countries of the world.

To conclude with the IFPRI study (Pinstrup-Andersen et al. 1997, p. 7):

> All of these issues suggest potentially larger fluctuations in food production and prices, and higher associated risks of food insecurity for the world's most vulnerable countries and people. The challenge for policy makers, researchers, and others is how best to minimize these risks to achieve food security for all people.

3. *Food Security Research and Theoretical Approaches*

The scientific debate on the causes of food insecurity was opened by Thomas Malthus (1798) more than 200 years ago. Malthus's basic argument was that the world's natural resources could not assure expansions in food supply that would match population growth. This argument was turned around by Boserup (1965) who postulated that population growth is one of the most powerful driving forces for agricultural development and increased food production.

Experience has proved that Malthus certainly underestimated both technical innovations in agricultural productions and checks in population growth. Although the Malthusian arguments are still prevalent, first of all in the context of the current situation in sub-Saharan Africa, the debate has now shifted away from perspectives of food availability to concerns with food entitlements. This paradigmatic change can be traced back to Amartya Sen's seminal work on *Poverty and Famine* (1981) which won him the Nobel Prize in Economics in 1998. The entitlement concept has primarily addressed the conditions under which chronic food insecurity collapses into mass starvation and in this sense, it endeavors to provide a theory of food crisis and famine causation. A person must necessarily starve if his/her entitlement set does not include a commodity bundle with enough food; a person is reduced to starvation if his/her endowment or exchange entitlement makes it no longer possible to acquire a commodity bundle with sufficient food (Sen 1990, p. 37).

Sen's entitlement theory can be regarded as a milestone in social science risk research, in general, and in the field of food vulnerability analysis, in particular. However, when addressing issues of vulnerability to food insecurity and famine, recent debates have clearly shown that there can be no single theory of food (in)security, because the problems are simply too complex. Consequently, Watts and Bohle (1993) and Downing et al. (1994) have tried to integrate three theoretical vantage points, namely human ecology, entitlement theory, and political economy, into a comprehensive concept of vulnerability analysis which combines issues of demography, ecology, poverty, empowerment, and long-term structural crisis proneness into one causal model.

A new strand of research is to systematically perceive food systems from two different perspectives, namely an 'external' and an 'internal' side. As Chambers had already pointed out in 1989, vulnerability is:

> the exposure to contingencies and stress, and difficulty coping with them. Vulnerability has thus two sides: an external side of risks, shocks and stress to which an individual or household is subject; and an internal side which is defenselessness, meaning a lack of means to cope without damaging loss (Chambers 1989, p. 1).

Contrary to the 'external' side of vulnerability, the 'internal' side of coping is still awaiting a firm theoretical base. Recent empirical research has revealed the enormous complexity in the scope, mechanisms, and efficiency of people's coping strategies for food security (for example, Spittler 1989, Bohle and Adhikari 1998). Thus, concepts are being developed currently that address the structural and behavioral context of coping strategies in more general terms, for example the 'access' model (Blaikie et al. 1994) or the notion of 'assets' as propagated by UNDP (HDR 1997/1999).

4. *Future Directions*

Although there is a wealth of empirical social science and behavioral studies on food security, their theoretical foundation is still rather weak, especially in the

fields of risk analysis and coping behavior. Moreover, the link between scientific approaches and practical needs is still underdeveloped. The theoretical frameworks are generally not operational for practitioners, although vulnerability analysis as developed by social sciences has become a standard procedure in food security projects and programs. Social and behavioral sciences should systematically provide insights into processes that allow goal-oriented structural changes for enhanced food security (e.g., promoting access and providing relevant assets). Finally, food security concerns have to be integrated more systematically into broader issues of 'human security' (HDR 1997/1999). In the field of transitory food insecurity (food crisis and famines), social and behavioral science has started to develop concepts for early warning and mitigation, such as risk mapping, to contribute to more precise social and geographical targeting procedures for food security. There is still broad scope for social and behavioral science in these fields.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Ecological Economics; Ecology, Cultural; Feminist Political Ecology; Food Production, Origins of; Human–Environment Relationship: Carrying Capacity; Population and Technological Change in Agriculture

Bibliography

ACC/SCN United Nations Administrative Committee on Coordination. Sub-Committee on Nutrition 1997 Third Report on the World Nutrition Situation. WHO, Geneva

Blaikie P, Cannon T, Davis I, Wisner B 1994 *At Risk. Natural Hazards, People's Vulnerability, and Disasters*. Routledge, London and New York

Bohle H G, Adhikari J 1998 Rural livelihoods at risk. How Nepalese mountain farmers cope with food insecurity. In: *Mountain Research and Development*, Vol. 18.4, pp. 321–32

Boserup E 1965 *The Conditions of Agricultural Growth: The Economics of Agrarian Change Under Population Pressure*. Allen & Unwin, London

Cannon T 1991 Hunger and famine: Using a food systems model to analyse vulnerability. In: Bohle H G, Cannon T, Hugo G, Ibrahim F N (eds.) *Famine and Food Security in Africa and Asia*. Naturwissenschaftliche Gesellschaft, Bayreuth

Chambers R 1989 Vulnerability, coping and policy. *IDS Bulletin* **20**(2): 1–7

Downing T E, Bohle H G, Watts M J 1994 Climate change and social vulnerability: Toward a sociology and geography of food insecurity. In: *Global Environmental Change*, Vol. 4, pp. 37–48

Kasperson J X, Kasperson R E, Turner II B L (eds.) 1995 *Regions at Risk: Comparisons of Threatened Environments*. United Nations University, Tokyo

Malthus T 1798 *An Essay on the Principle of Population*. London

Oltersdorf U, Weingärtner L 1996 *Handbuch der Welternährung. Die zwei Gesichter der globalen Nahrungssituation*. Deutsche Welthungerhilfe, Bonn, Germany

Pinstrup-Andersen P, Pandya-Loren R, Rosegrant M W 1997 *The World Food Situation: Recent developments, Emerging Issues, and Long-term Prospects*. Food Policy Report, IFPRI, Washington, DC

Sen A K 1981 *Poverty and Famines: An Essay on Entitlement and Deprivation*. Clarendon Press, Oxford, UK

Sen A K 1990 Food, economics, and entitlements. In: Drèze J, Sen A K (eds.) *The Political Economy of Hunger*. Clarendon Press, London

Smith M, Pointing J, Maxwell S 1992 *Household Food Security: Concepts and Definitions. An Annotated Bibliography*. The Institute of Development Studies, Brighton, UK

Spittler G 1989 *Handeln in einer Hungerkrise: Tuaregnomaden und die Große Dürre von 1984*. Opladen, Germany

UNDP 1997/1999 *Human Development Report*. United Nations Development Program, New York

Watts M J, Bohle H G 1993 Hunger, famine and the space of vulnerability. *Geojournal* **30**(2): 117–25

World Bank 1986 *Poverty and Hunger*. World Bank, Washington, DC

World Food Council 1988 *Towards Sustainable Food Security: Critical Issues*. Report by the Secretariat, 14th Ministerial Session, Nicosia, Cyprus, 23–26 May

World Food Council 1999 *Current World Food Situation*, 116th Session, Food and Agricultural Organization, Rome, 14–19 June

H.-G. Bohle

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

For

Foreign Investment: Direct

Direct Foreign Investment or Foreign Direct Investment (FDI) is defined as the ownership (partial or full) and control of assets in one country by foreign residents. FDI in the USA is defined as the foreign ownership or control of, directly or indirectly, at least 10 percent of the voting securities of an incorporated US business enterprise or an equivalent interest in an unincorporated US business enterprise. Two major forms of FDI are the acquisition of an existing firm or real property or the establishment of new operations. The OECD countries account for most FDI flows in the world. Furthermore, large-scale mergers and acquisitions, particularly between British firms, and US firms have been an important development in the 1990s. The USA is the top ranking country in terms of both FDI inflows and outflows followed by the UK. Other dominant sources of FDI are Canada, Japan and the West European countries. With the exception of Japan, both Canada and West European countries are also the top ranking recipients of FDI. Notable trends in the 1990s include the flow of FDI in East European economies and China in response to national level policy changes towards foreign investments. Another development has been the growth of FDI in East and Southeast Asia in the 1990s by non-Asian firms in response to deregulation, as well as opportunities for low-priced acquisition following the Asian financial crisis. Prior to the Asian crisis, intra-Asian investments dominated FDI flows in East and Southeast Asia.

FDI studies can be classified into two broad categories: (a) analyses of factors affecting the flow of FDI from one country to another country; and (b) the location characteristics of FDI within a host economy. Other important topics include the analysis of the impact of FDI (e.g., employment and technology transfer) in source and host countries. However, due to data limitations, impact analyses are limited to corporate surveys.

Most explanations of FDI have been based on the trans-nationalization process of large USA based multinationals in manufacturing industries (Caves 1971, Hymer 1976, Vernon 1966). In the 1970s and 1980s, the upsurge of FDI from Western Europe, Japan, and the newly industrializing nations, as well as FDI in service industries called for a critique of the existing explanations and further theorization of the determinants of FDI (Dunning 1989, Lall 1983). In the 1980s, the USA became the top recipient of FDI and several studies focused on the spatial and temporal dynamics of FDI in the USA with a special emphasis on FDI from Japan (Bagchi-Sen and Wheeler 1989, McConnell 1980, O'Huallachain 1985). The worldwide growth of FDI also led to the analysis of multinational activity in other destinations (e.g., the UK, Australia, Eastern Europe, and East and Southeast Asia) (Dicken 1998, Edgington 1990).

1. Determinants of FDI

Initially, the principle of comparative advantage was presumed to govern foreign investment flows. However, the neoclassical explanation did not account for the continuation of the bilateral flow of investment among capital-rich nations such as the USA and the countries of Western Europe. The neoclassical explanation of trade also failed to recognize the importance of technology, know-how, and managerial or marketing expertise, as significant determinants of FDI

1.1 The Industrial Organization Approach

Hymer's analysis of the trans-nationalization process of firms based on the theory of market imperfections suggested that foreign direct investors must possess specific advantages over competing host country firms (Hymer 1976). These competitive advantages can be derived from the factor market (e.g., discriminatory access to capital, technology, and managerial expertise), product market (through product differentiation and pricing policy), internal and external economies of scale, and government regulations (e.g., limitations on output or entry). These advantages will allow the direct investor to overcome barriers (e.g., the lack of information about local markets, legal issues, infrastructure and logistics, and culture among others) encountered in the host country. Caves (1971) suggested that in order to compete with domestic firms, foreign firms must internalize certain firm-specific advantages, such as patented technology, managerial know-how, marketing strategy, product differentiation techniques and economies of scale. Internalization refers to an extension of firm's activities rather than the externalization of such activities through contracts with independent host country firms. Alternatives to internalization, such as licensing, man-

agement contracts, franchises, and subcontracts are not feasible methods of appropriating firm-specific advantages.

1.2 The Product Cycle Theory

One of the major explanations of the internationalization process of firms as a sequential development is the product cycle theory (Vernon 1966). The product cycle theory explains why market-oriented firms replaced their export markets with direct investment in overseas facilities. The relevance of this theory becomes questionable when a firm's investment approach involves risk diversification through the integration of worldwide operations rather than the dependence on one market alone. Furthermore, some firms do not follow the sequential process and may skip the export phase altogether and undertake a direct investment. In the case of Japanese FDI, some products (e.g., color television) were developed exclusively for the US market. Europeans on the other hand reacted to US investor's pricing strategy in European markets by undertaking FDI in the USA (exchange of threat). The analysis of multinationals from developing countries shows that innovative power and wealth are not necessary preconditions of FDI—these firms often undertake FDI following defensive, imitative, dependent, and opportunist strategies.

1.3 The Eclectic Approach

In recent years, Dunning's eclectic paradigm has been used to explain FDI in both manufacturing and services from all possible sources (Bagchi-Sen and Sen 1997). Ownership, location, and internalization advantages are the three determinants of trans-nationalizaton (Dunning 1977). Once a firm possesses ownership-specific advantages (knowledge, management-marketing-financial skills, control of resources and markets, and the ability to diversify risks) and successfully internalizes its proprietary know-how, remaining variables that further influence the choice of a destination are location or country-specific characteristics (e.g., market demand, trade barriers). The analysis of firm structure and behavior of US-based, German and Japanese firms by Pauly and Reich (1997) highlights source-country based differences in corporate governance, financing, research and development, direct investment activities and intra-firm trade. The findings emphasize that a global model of internationalization is not the norm.

2. Spatial and Temporal Variations in FDI

Studies on location preferences of foreign investors offer a variety of explanations based on samples drawn from different regions, time periods, and groups of firms (Hayter 1997). Selected findings from these location analyses over the past three decades are discussed. The most prevalent pattern is the location of foreign firms in the core regions of host countries (Blackbourn 1974, McConnell 1980). These core regions offer economies of agglomeration and lower search cost as well as uncertainty for newcomers. In the USA, for example, the major mode of entry is acquisitions—therefore, core regions are preferred because these regions offer an array of possibilities for mergers and acquisitions. In both developed and developing countries, the large metropolitan centers attract FDI, especially FDI in service sectors (Bagchi-Sen and Sen1997, Edgington 1990). Latecomers in manufacturing have shown the tendency to locate in emerging industrial spaces (e.g., the preference of foreign manufacturers to locate in the southern and western part of the USA as opposed to the northeastern and New England states) or sometimes foreign investors create their own industrial corridors in a host country (e.g., the location of Japanese automakers and suppliers in southern Ontario, Michigan, Indiana, Ohio, Illinois, Kentucky and Tennessee in North America). A preference for peripheral areas for new investments is often influenced by the incentive packages offered by state and local governments. Investments by state/local governments to attract Japanese autoproducers ranged between US$33 million to US$149 million in the USA. Similar investments have been incurred by local governments in other countries such as Setubal (Portugal), West Midlands and northeast England (UK), and Lorraine (France), among others, (Dicken 1998) to attract automakers from the USA, Korea, and Germany. A further observation is that foreign investors often prefer particular regions based on sector characteristics of FDI, physical proximity of the investor to the host, and personal choice/bias of the decision-maker. For example, Japanese automakers avoided places with existing US auto plants and States where another Japanese autoproducer had located their manufacturing operations.

3. Conclusion

The growth of FDI in the 1990s surpassed the growth of exports in goods and services, world output and gross domestic capital formation. The importance of FDI in host economies cannot be ignored because in several countries FDI's share of gross domestic product is well over 10 percent. Furthermore, the impact of FDI is far reaching. A nation's capital and financial markets, employment pattern, industrial structure and competitiveness, technology, as well as trade and development patterns, are affected by FDI. The nature of the impact depends on the characteristics of FDI (e.g., mode of entry, function, and attributes), the timing of FDI and the bargaining relationship between the foreign entity and the nation state (Cantwell 1989 Dicken 1998).

See also: Development and the State; Development, Economics of; Economics: Overview; Financial Institutions in Economic Development; Industrialization; International Business; International Organization; International Trade: Economic Integration; Multinational Corporations

Bibliography

Bagchi-Sen S, Wheeler J O 1989 A spatial and temporal model of foreign direct investment in the United States. *Economic Geography* **65**: 113–29

Bagchi-Sen S, Sen J 1997 The current state of knowledge in international business in producer services. *Environment and Planning A* **29**: 1153–74

Blackbourn A 1974 The spatial behavior of American firms in Western Europe. In: Hamilton, F E I (ed.) *Spatial Perspectives on Industrial Organization and Decision-Making*. John Wiley, London, pp. 245–64

Cantwell J 1989 *Technological Innovation and the Multinational Corporations*. Blackwell, London

Caves R E 1971 International corporations: the industrial economics of foreign investment. *Economica* (New Series) **38**: 1–27

Dicken P 1998 *Global Shift. Transforming the World Economy* 3rd edn. Guilford, London

Dunning J H 1977 Trade, location of economic activity and the MNE: a search for an eclectic approach. In: Ohlin B, Hesselborn P O, Wijkman P M (eds.) *The International Allocation of Economic Activity*. Macmillan, London, Chap. 12

Dunning J H 1989 *Transnational Corporations and the Growth of Services: Some Conceptual and Theoretical Issues*. United Nations, New York

Edgington D W 1990 *Japanese Business Down Under: Patterns of Japanese Investment in Australia*. Routledge, London

Hayter R 1997 *The Dynamics of Industrial Location. The Factory, the Firm and the Production System*. Wiley, New York

Hymer S 1976 *The International Operation of National Firms: A Study of Direct Foreign Invertment*. MIT Press Cambridge, MA (PL.D. thesis 1960, MIT)

Lall S 1983 *The New Multinationals: The Spread of Third World Enterprises*. Wiley, Chichester, UK

McConnell J E 1980 Foreign direct investment in the United States. *Annals of the Association of American Geographers* **70**: 259–70

O'Huallachain B 1985 Spatial patterns of foreign investment in the United States. *Professional Geographer* **37**: 155–63

Pauly L W, Reich S 1997 National structures and multinational corporate behavior: enduring differences in the age of globalization. *International Organization* **51**: 1–30

Vernon R J 1966 International investment and international trade in product cycle. *Quarterly Journal of Economics* **80**: 190–207

S. Bagchi-Sen

Foreign Language Teaching and Learning

Foreign language learning and teaching refer to the teaching or learning of a non-native language outside of the environment where it is commonly spoken. A distinction is often made between 'foreign' and 'second' language learning. A second language implies that the learner resides in an environment where the acquired language is spoken. In the area of research, however, the term second language acquisition is a general term that embraces foreign language learning. Research into the acquisition of a non-native language includes the disciplines of psychology, linguistics, education, neurobiology, sociology, and anthropology. Inquiries of learning and teaching innovations have provided new insights into successful language learning strategies and environments designed to increase language achievement and proficiency.

1. Definition

A language is considered foreign if it is learned largely in the classroom and is not spoken in the society where the teaching occurs. Study of another language allows the individual to communicate effectively and creatively and to participate in real-life situations through the language of the authentic culture itself. Learning another language provides access into a perspective other than one's own, increases the ability to see connections across content areas, and promotes an interdisciplinary perspective while gaining intercultural understandings. Language is the vehicle required for effective human-to-human interactions and yields a better understanding of one's own language and culture. Studying a language provides the learner with the opportunity to gain linguistic and social knowledge and to know when, how, and why to say what to whom (National Standards in Foreign Language Education Project 1996).

Language scholars distinguish between the terms acquisition and learning: 'acquisition' refers to the process of learning first and second languages naturally, without formal instruction, whereas 'learning' is reserved for the formal study of second or foreign languages in classroom settings. One usually distinguishes between the relatively effortless process of second language acquisition by children and the more formal and difficult second language acquisition by adults. Foreign language education refers to the teaching of a modern language that is neither an official language nor the mother tongue of a significant part of the population.

2. Theories of Language Learning

Foreign language learning and teaching have undergone a significant paradigm shift as a result of the research and experiences that have expanded the scientific and theoretical knowledge base on how students learn and acquire a foreign language. Traditionally, learning a foreign language was thought to be a 'mimetic' activity, a process that involved students repeating, or imitating new information. Grounded in

behaviorist theories of learning and structural linguistics, the quality and quantity of language and feedback were regarded as the major determinants of language learning success. A popular method of teaching in the 1950s, called the audio-lingual approach (ALM), promoted an imitation and practice approach to language development. The major figure in the ALM classroom was the instructor who was cast into the role of drill sergeant, expert, and authority figure. Students were relegated to practicing and imitating patterns to a point of automatic response in the belief that the learner would then merely have to slot in lexical items appropriate to the conversational situation. It was believed that the first language interfered with the acquisition of the second language and that a transfer would take place from the first to the second language, resulting in errors. In 1959, Noam Chomsky's review (Chomsky 1959) of B. F. Skinner's (1957) *Verbal Behavior* dramatically changed the way of looking at language by arguing that language was a rule-governed activity, not a set of habits. Chomsky argued that stimulus–response psychology could not adequately account for creativity involved in generating novel utterances using internalized rules. The creative aspect of language behavior implies that the human mind is involved in deep processing of meaning rather than in memorized responses to environmental stimuli. Chomsky's view of language and cognitive psychology, dubbed generative transformational grammar, regarded language acquisition as an internal thinking–learning process. Chomsky claimed that children are biologically programmed for language and have an innate ability to discover for themselves the underlying rules of a language system. Chomsky's ideas led to the demise of structural linguistics, behaviorist psychology, and the ALM approach to language learning.

An alternative theoretical position emerged centered on the role of the linguistic environment in combination with the child's innate capacities in acquiring language. This position (interactionist) viewed language development as the result of a complex interplay between innate language capacities of the learner and the learner's environment. Unlike the innatist position (e.g., Chomsky 1959), the interactionists claimed that language had to be modified to the ability of the learner. According to Long (1985), language input was made comprehensible by simplifying the input, by using linguistic and extralinguistic cues, and by modifying the interactional structure of the conversation. Long maintained that speakers adjust their language as they interact or negotiate meaning with others. Through negotiation of meaning, interactions are changed and redirected, leading to enhanced comprehensibility. Long proposed that learners, in order to acquire language, cannot simply listen to input, rather they must be active co-constructive participants who interact and negotiate the type of input they receive.

Each of these theories of language acquisition addresses a different aspect of a learner's ability to acquire a language. Behaviorist explanations explain systematic aspects, whereas innatist explanations explain the acquisition of complex grammar. Interactionist explanations assist in understanding how learners relate form and meaning in language, how they interact in conversation, and how they use language appropriately.

3. *Emphases in Second Language Research and Teaching*

Research has revealed that knowledge of language structures demonstrated on discrete-point tests does not ensure communicative ability when the measure of language knowledge is one of more spontaneous language use. Further studies have shown that there is little correlation between the rules learners are taught and their developing knowledge of the second language. Language scholars have demonstrated that certain aspects of second language learning cannot be altered through instruction and that intermediate, non-native-like second language competencies, known as stages of interlanguage, characterize the progression of second language acquisition. Selinker (1974) viewed interlanguage as an intermediate system located on a continuum stretching from the native language to the target language. Corder (1978) stated that, in the interlanguage process, the learner constantly and progressively adjusts the native language system to approximate the target language system more closely (restructuring continuum). Corder noted that not all learners showed evidence of transfer from native language to target language and suggested that there was a uniformity about the way second language learners progress and that they follow approximately the same sequence of development regardless of their native language (developmental continuum). According to this view of second language acquisition, the controlling factor is the innate ability for learning language that all human beings possess. Pica (1983) determined that all language learners progressed through a fixed series of stages, known as developmental sequences, in learning particular linguistic subsystems, such as word order, negation, or relative clauses. In English negation, for example, when communicative samples were examined, it was revealed that both foreign language and second language learners progressed through the same four-stage sequence, defined in terms of placement of negation. Ellis (1986) reviewed several studies that involved Japanese, Spanish, German, and Norwegian children, adolescents, and adult learners. He concluded that all English-as-a-second language learners pass through the following prescribed set of stages: (a) no + phrase, e.g., 'No drink'; (b) negator moves inside the phrase, e.g., 'I no can swim'; (c) negator is attached

to modals, e.g., 'I can't play this one'; and (d) auxiliary system is developed and learner acquires correct use of not and contractions, e.g., 'He doesn't know anything.' This suggests that learners make particular kinds of errors at particular stages in the acquisition of a structure. Each stage marks some kind of restructuring in the mind of the learner regarding that particular structure. Structure evolves over time. These studies had significant implications for how foreign languages should be taught.

4. *Learner-centered Instruction*

Two communicative approaches, the input model and the input interaction model represent two models of foreign language theory and teaching that investigate the language acquisition process from the perspective of the learner. Krashen (1982) is the principal advocate of the input model of foreign language teaching. His theory is grounded in (a) Chomsky's generative linguistics, (b) research on the effectiveness of different second/foreign teaching methods, and (c) research on affective factors (motivation, anxiety, personality, etc.). Krashen posited that second language acquisition occurs when the learner comprehends the language input in a low-anxiety, high-motivation situation, and proposed that the teacher's role is to create such a learning environment. Krashen further claimed that conscious grammar teaching/learning is effective only in a monitoring capacity to check for grammatical accuracy, not in the acquisition of the second language itself.

Because classrooms remained a major setting for language learning, the pursuit to determine those elements that enhanced classroom language achievement became particularly important. Why do two learners who seemingly have the same instructional opportunity achieve varying levels of language proficiency? Investigations focused on individual skills or abilities and environmental factors that may impact foreign language achievement and proficiency.

Individual cognitive (e.g., intelligence, aptitude, or ability) and affective (e.g., attitude and personality variables) factors were analyzed. Skehan (1986) noted a fairly strong relationship between cognitive variables such as aptitude, intelligence, and language achievement for learners in foreign language classrooms. The predictive power of such traits, however, decreased as the criteria for language proficiency became more communicative and the learning setting became more natural (versus formal and instructional). The most avid pursuit in research occurred in investigations of the role of motivation in learning language and the learner's attitude toward the target language and culture. Using Gardner and Lambert's (1972) differentiation between integrative and instrumental motivation, researchers reported no significant advantage for an integrative (intrinsic) motive and others reported an advantage when the learner was driven by instrumental (extrinsic) motives. Integrative motivation was defined as one in which the target language was being learned by an individual in order to be accepted by the native speaker community. Instrumental motivation was one in which the language was being learned for external benefits, such as securing a better job.

Results of studies investigating environmental factors reported on the effect on achievement scores. Carroll (1975) conducted a survey of French instruction in eight countries and noted effects on achievement by gender, school type, and teacher gender, and mixed effects according to parental interest. Social factors outside the school were determined to have a significant impact on the development of language proficiency. Both cognitive and affective factors were investigated to explain the variance in foreign language achievement. Motivation, attitudes, anxiety, self-esteem, tolerance of ambiguity, risk-taking, cooperation, and competition proved to be key variables that explained individual differences in foreign language learning (Ellis 1994). Successful language learning was determined to be largely dependent on who was learning the language, under what circumstances, and for what purposes. Foreign language acquisition was revealed to be a complex, multidimensional process influenced by both learner and environment variables. The questions generated by these theories and research studies began to focus on significant new responsibilities on the part of the teacher in the design and support of individual and personalized learning tasks.

5. *Learning and Measurement*

Language teaching throughout the 1990s has experienced numerous curricular innovations in response to the importance of providing students with opportunities to acquire and practice the foreign language in contextualized and meaningful language communicative tasks at all stages of the second or foreign language acquisition process. Communicative language teaching (CLT), the term most associated with current discussion of method, emerged as a significant approach that found universal resonance and support in theory and application in many contexts and across disciplines (linguists, methodologists, curriculum developers). Central to the rise of CLT was the realization that linguistic competence does not on its own achieve communicative competence (Canale and Swain 1980) and that language used in meaningful, authentic contexts is more readily acquired.

Pair work, group work, cooperative/collaborative learning settings, authentic materials, culturally integrated lesson content and interactive tasks focused on the cognitive and affective domains were integrated into foreign language classrooms. Constructivist

teaching practices, influenced by Vygotsky's emphasis on social interaction in learning and development, helped learners to internalize and reshape new information. The theoretical underpinnings of Vygotsky's (1978) view of language learning that maintained contextualized input in cooperative, meaningful interactions with others enhanced language acquisition took hold in classrooms around the globe. The interaction between an expert (teacher) and novice (learner) in a problem-solving task (scaffolding) in which the expert's role was to provide the novice with scaffolded help became the model for communicative tasks in the foreign language classrooms. The expert's/teacher's role was to gain the learner's interest in the task, simplify the task, keep the learner motivated, point out important features, reduce anxiety and frustration during problem solving, and model appropriate form.

In accordance with the new responsibilities, the role of the classroom teacher shifted to that of an architect, creating meaningful, interactive, cooperative learning tasks designed to engage the learner actively in negotiating language meaning in authentic contexts.

The focus on student language proficiency as measured through performance-based tasks made itself felt both in language learning research and in teaching. Questions emerged regarding how language proficiency could be enhanced and how best to measure the level of language proficiency.

As the proficiency movement gained momentum in the 1980s, most especially in the USA, consensus was sought about describing and measuring language abilities. The development of the Provisional Proficiency Guidelines by the American Council on the Teaching of Foreign Languages (ACTFL) defined what language users are able to do with language in speaking, listening, reading, and writing at various levels of performance. These Guidelines marked a major shift in language pedagogy from methodology to measurement and a focus on learner outcomes. In 1996, content standards were published (National Standards in Foreign Language Education Project 1996) that delineated what learners should know and be able to do. More recently, the ACTFL Performance Guidelines for K-12 Learners (ACTFL 1999) described language performance within three modes of communication (interpersonal, interpretive, presentational) to assist teachers in understanding how well students demonstrate language ability at various points along the language-learning continuum. A similar effort by the International Association for the Evaluation of Educational Achievement's Language Education Study is seeking a comparison and evaluation of the outcomes of different educational systems across Europe. The Standards Movement, seeking to promote the establishment of guidelines for the teaching of foreign languages for all learners, indicates the growing concern with learner outcomes and accountability. In a standards-driven environment the shift to student performance will require that teachers have a repertoire of approaches that target specific goal areas or standards.

6. *Future Directions*

Future directions in research should investigate the potential use and effect of computer technology on foreign language learning. As classroom tasks become more focused on real-world issues, texts, or events, and problem-solving-based tasks, computers introduce a new dimension to the teaching and learning process. Technology allows students to manipulate learning materials/language at their own pace and according to individual needs. Students examine reports, authentic documents and web pages to find information that can be synthesized and discussed later and can collaborate electronically with youth from around the world. In such a learning environment the role of the teacher changes from one of authority figure/expert who delivers knowledge to one who facilitates guides and supports student learning. The teacher assumes greater responsibilities in designing and supporting individual and personalized learning tasks. This has tremendous implications for teacher educators and teacher trainers to act as agents of change.

One of the most effective research methodologies that emerged in the last few years has been action research. Inquiring into one's own instructional practices through classroom-based investigations, teachers actively contribute to the research endeavor and change practices based on findings. Such research promises to improve teaching practices that are of interest to both researchers and teachers.

Methodologically classroom-oriented research has been largely conducted within the framework of correlational approaches, case studies, survey research, ethnographic research, experiments, and discourse analysis (Johnson 1992). While the choice of research method is largely determined by the nature of the research question to be investigated, or by the hypothesis to be tested, thoughtful combinations of qualitative and quantitative research on foreign/second language learning conditions will provide valuable insights into language-acquisition processes. Greater use of qualitative methods investigating students in their classrooms with special attention to cultural, situational, and longitudinal contexts is needed and recommended. As foreign language research draws on related disciplines (psychology, psycholinguistics, neurobiology, neurolinguistics, sociology, linguistics) to better explain conditions that lead to greater language proficiency and differential success among foreign language learners, a deeper understanding of how languages are acquired and consequently how they should be taught will be gained. As learning and teaching innovations are

tested and researched, new insights will be gained that will influence teaching practices globally.

See also: Communication and Social Psychology; Language Acquisition; Second Language Acquisition; Verbal Communication, Psychology of

Bibliography

ACTFL 1999 *ACTFL Performance Guidelines for K-12 Learners*. American Council on the Teaching of Foreign Languages, Yonkers, NY

Canale M, Swain M 1980 Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* **1**: 1–47

Carroll J B 1975 *The Teaching of French as a Foreign Language in Eight Countries*. Wiley, New York

Chomsky N 1959 Review of *Verbal Behavior* by B. F. Skinner. *Language* **35**: 26–58

Corder S P 1978 Language-learning language. In: Richards J C (ed.) *Understanding Second and Foreign Language Learning: Issues and Approaches*. Newbury House, Rowley, MA

Ellis R 1986 *Understanding Second Language Acquisition*. Oxford University Press, Oxford, UK

Ellis R 1994 *The Study of Second Language Acquisition*. Oxford University Press, Oxford, UK

Gardner R C, Lambert W E 1972 *Attitude and Motivation in Second Language Learning*. Newbury House, Rowley, MA

Jonson D M 1992 *Approaches to Research in Second Language Learning*. Longman, White Plains, NY

Krashen S 1982 *Principles and Practice in Second Language Acquisition*. Pergamon, Oxford, UK

Long M H 1985 Input and second language acquisition theory. In: Gass S, Madden C (eds.) *Input in Second Language Acquisition*. Newbury House, Rowley, MA

National Standards in Foreign Language Education Project 1996 *Standards for Foreign Language Learning: Preparing for the 21st Century*. Allen Press, Lawrence, KS

Pica T 1983 Adult acquisition of English as a second language under different conditions of exposure. *Language Learning* **33**: 465–97

Selinker L 1974 Interlanguage. In: Schumann J H, Stenson N (eds.) *New Frontiers in Second-language Learning*. Newbury House, Rowley, MA

Skehan P 1986 The role of foreign language aptitude in a model of school learning. *Language Testing* **3**: 188–221

Skinner B F 1957 *Verbal Behavior*. Methuen, London

Vygotsky L S 1978 *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

A. K. Moeller and E. Koubek

Foreign Policy Analysis

What causes states to adopt certain foreign policies? Why do some nations seek to extend their influence abroad while others define their interests in narrower, more inwardly focused terms? Statesmen, historians, and political scientists have long pondered these questions. Though there is no shortage of answers, most observers have followed one of two distinct paths: the first emphasizes states' external or international situation; the second stresses their internal or domestic makeup. This article examines each of these traditions or approaches in turn.

1. *Realpolitik*

'Realism,' which emphasizes a state's relative position in the international distribution of power, is usually regarded as the dominant approach to the study of foreign policy. Certainly, it is the oldest. Ever since Thucydides (1972) wrote about the Greek city-states' struggle for power and security, many have embraced realism's basic postulates as the best guide to explaining state behavior: anarchy is the basic, irreducible fact of international life (Waltz 1959). In the absence of a central authority to settle disputes over territory and wealth, states must provide for their own security and welfare. Exactly how they go about this, and the conditions that make success most likely, are the central issues animating realism.

1.1 *The Classical School*

Although scholars classify realism into different variants ('offensive' and 'defensive' realism and 'classical' and 'structural' realism), in the foreign policy field, the term is most often associated with classical realists, scholars like E. H. Carr (1939), Nicholas Spykman (1942), and Hans J. Morgenthau (1978). Writing during the years between World Wars I and II, these self-proclaimed realists sought to catalogue the 'do's' and 'don'ts' of international power politics. Their work continues to be a touchstone of realist thinking about foreign policy.

Classical realists start from the assumption that states are influence-maximizers: they seek, first and foremost, to control and shape their external environment and, all things being equal, will seek more rather than less influence. How ambitious and single-minded their leaders are in this regard depends on many things, but for the realist none is more essential than a nation's power, or more precisely, its relative power, where power refers to its share of material resources such as military strength, world trade, and population.

For realists, power is indispensable. Knowing how much a nation has, they argue, tells much about the foreign policies it will likely pursue. In general, as states' international power or capabilities increase, they will seek to extend their political interests abroad; as their power ebbs, their actions and aspirations will be scaled back accordingly. In short, power

shapes intentions: the stronger the state's international position, the more expansive its foreign policy interests will be.

1.2 Realism's Strengths and Weaknesses

One of realism's great virtues is that it reminds us of the importance of power in understanding state behavior. Big changes in a nation's international position—such as Britain's rise to great powerdom in the late eighteenth century, Germany's ascent under Otto von Bismark after 1870, and American dominance after World War II—can lead to big changes in its foreign policies. The theory that a country's power determines its intentions is crisp and parsimonious, and much of realism's appeal stems from this fact.

A major weakness of realism, however, is that it is underdetermining. A nation's international position may well create incentives for action. But rarely is a state's 'relative power' vis-à-vis its allies, competitors, and neighbors clear and unambiguous, and rarely is there a consensus among its leaders over the appropriate response to foreign policy opportunity or challenge. For every rising state that has acted as realism predicts, there is one that has not. The same is true about states in decline (Kennedy 1987). Declining powers often continue to expand abroad, failing to make the adjustments realism predicts. Sometimes this results in disaster as realism would predict, but sometimes they get away with it.

Realists try to address this shortcoming in one of two ways. The first involves substituting 'elite perception' about international strength for objective reality. By dispensing with objective measures of power in favor of policy makers' perceptions of power, it is possible, realists contend, to explain gaps or lags in a country's response to increases (or decreases) in its power as assessment failures (Friedberg 1988, Jervis 1976). Strategic miscalculations or misjudgments arise because of difficulties inherent in judging military strength, commercial prowess, and so on accurately.

A second solution to the problem of indeterminacy involves the introduction of domestic decision-making structures into the model. Even if leaders assess their nation's power correctly, they may not be able to respond appropriately because of domestic institutional constraints (Krasner 1978, Zakaria 1998). The weaker (i.e., the more fragmented and decentralized) a state's political institutions, the harder for its leaders to mobilize internal resources and act with dispatch. Here, state strength functions as an 'intervening variable,' mediating the transmission of capabilities (realism's independent variable) into behavior (its dependent variable).

Such efforts to bring theory and reality into line are useful. But introducing variables such as elite perception and state strength also carries a methodological price, most notably a loss of parsimony. One needs to know a great deal about the country in question, as well as its decision makers, to make the theory work. Even more serious is the inevitable loss of distinctiveness that accompanies such broadening of the theory. The more realists turn to domestic variables, the harder it becomes to distinguish *Realpolitik* from its competitor: *Innenpolitik*.

2. Innenpolitik

If relative power in the international arena is realism's core idea, *Innenpolitik*'s is domestic political structure. Like realism, this approach claims that the context in which foreign policy decisions are made shapes policy. Where they differ is how they understand 'context' or, more accurately, in their judgment about which setting, the international or the domestic, should receive pride of place in the analysis. For realists, a state's external situation is determinant; for *Innenpolitiker*s a country's internal makeup is decisive.

2.1 Many Theories, One Approach

Like realism, *Innenpolitik*'s roots are European. The approach traces its origins back to early-twentieth-century accounts of British imperialism and German *Weltpolitik*, though at least one variant of the general approach, the so-called 'democratic peace theory,' can be traced to Immanuel Kant's discourses. Vladimir I. Lenin's (1916) theory of colonialism, and economist Joseph A. Schumpeter's (1955) theory of imperialism are classic examples of the *Innenpolitik* genre. In the North American context, the earliest example is historian Charles Beard's (1934) revisionist account of America's entry into World War I.

As these examples suggest, *Innenpolitik* is less a single theory than a family of theories, each stressing a different, specific domestic variable in the explanation of foreign policy outcomes. Some argue for the importance of a nation's culture, others its socioeconomic structure, and still others its political institutions. What ties these theories together is the common assumption that foreign policy is best understood from the 'inside out,' as a product of domestic conditions and circumstances.

In its strongest forms, *Innenpolitik* makes predictions about foreign policy based solely on information about a state's domestic makeup. The notion that democracies behave differently from nondemocracies is a popular, if overdrawn, example. A more subtle version of the argument is democratic peace theory (Doyle 1983). Liberal institutions and norms, it is claimed, make democracies less inclined to fight each other, even if they show no such reluctance when dealing with authoritarian regimes.

The key distinction in democratic peace theory is regime type—democratic or authoritarian. How war-

prone states are depends upon their domestic political institutions. The implication is clear: change a state's domestic structure, and its foreign policy behavior will change accordingly. International conditions, while not completely ignored, are secondary. They neither constrain nor dictate foreign policy; they provide opportunities. Whether states seek to exploit international opportunities or not depends on how constrained or motivated their leaders are by 'checks and balances,' interests of domestic constituencies, and pressures to gain office in a competitive political environment.

A focus on states' domestic makeup also informs many theories of imperialism, the most well-known of which, Marxism, claims that capitalist states, due to problems of overproduction and underconsumption, are more prone to expansionism than non-capitalist ones. Whether or not they pursue an aggressive imperialist foreign policy, and how likely they are to succeed if they do, may depend on the international balance of power, but their expansionist impulses do not. They are internally generated. Again, domestic structure is determinative.

2.2 *Choice and Decision-making*

Theories of democratic peace and imperialism are structural theories, and in this regard are like realism. Strictly speaking, information about the beliefs, perceptions, and prejudices of decision makers is unnecessary. For *Innenpolitiker*s, the implicit claim is that states with similar domestic structures will pursue similar foreign policies. Conversely, dissimilar states will differ in their foreign policies.

The problem is that there are many examples of states with similar domestic structures that have acted differently in foreign policy. The reverse is true, too: dissimilar states sometimes act similarly. It turns out that just as a state's external position rarely precludes a range of responses, its internal makeup leaves room for meaningful choice. To understand how those choices are made, domestic-level theorists, like realists, have sought to incorporate decision-making into their models.

The decision-making theory that has had the biggest impact on the *Innenpolitik* approach to studying foreign policy is 'bureaucratic politics.' A form of interest-group theory, its central claim is that foreign-policy decisions are the result of bargaining among government officials and agencies (Allison 1971). Foreign-policy choices are thus seen as a product of the distribution of power and initiative among governmental actors, which varies by policy issue.

That government officials' choices matter in the making of foreign policy is clear. Yet how useful bureaucratic politics is in explaining those choices is less obvious. One problem is that the decision-making process inside government is not as autonomous and insulated from larger political processes as bureaucratic politics theorists would have it: 'palace politics' often reflects larger social interests and powers. Moreover, the theory seems to work best when it matters least: on issues that are not seen as crucial to the state's security, and that are therefore handled mostly inside the foreign-policy-making bureaucracy.

3. *The State of the Art*

As structuralist theories, realism and *Innenpolitik* both seem to flatten the sense of choice and politics that characterizes statecraft. The future of foreign policy studies lies in finding systematic ways to transcend this problem: that is, to introduce choice and politics into models that recognize that foreign-policy making is constrained by domestic institutions and international balances of power.

Some possible strategies have been proposed. One is to introduce politics by analyzing the uneven impact of international forces on the different constituencies, sectors, or interest groups within one country (Gourevitch 1986, Trubowitz 1998). Groups compete within the framework of domestic political institutions to promote what is, in the final analysis, an inherently partisan vision of 'the national interest.'

Another is Robert Putnam's (1988) solution, which is to envision statesmen as rational maximizers who are engaged in 'two-level games.' Foreign-policy makers must strategize in domestic political arenas, as 'democratic peace' theorists suggest, and at the same time seek stability or advantage in the competitive international settings that are the centerpiece of realist analysis.

Finally, there are constructive theories that dissolve the distinction between 'international' and 'domestic' (Katzenstein 1996). Constructivist writers see 'structural constraints' themselves—understandings of international power balances, state interests, and domestic priorities—as the product of ideologies, cultural bias, and discourses about state and nation.

The main empirical challenge for foreign policy analysts is to figure out ways to conceptualize 'international context' and 'national interest' at a time when international relations is fluid and dominated increasingly by forces not controlled directly by governments (such as international capital flows), and when traditional definitions of security must be broadened to include environmental, economic, and other non-state sources of challenge.

See also: Balance of Power: Political; Conflict/Consensus; Decision-making Systems: Personal and Collective; Diplomacy; Dispute Resolution in Economics; Imperialism: Political Aspects; International Law and Treaties; Military and Politics; National Security Studies and War Potential of Nations; Peace; Power: Political; Risk: Theories of Decision and Choice; War, Sociology of

Bibliography

Allison G T 1971 *Essence of Decision: Explaining the Cuban Missile Crisis*. Little Brown, Boston

Beard C A 1934 *The Idea of National Interest: An Analytical Study in American Foreign Policy*. Macmillan, New York

Carr E H 1939 *The Twenty Years' Crisis, 1919–1939: An Introduction to the Study of International Relations*. Macmillan, London

Doyle M W 1983 Kant, liberal legacies, and foreign affairs. *Philosophy and Public Affairs* **12**: 205–35

Friedberg A L 1988 *The Weary Titan: Britain and the Experience of Relative Decline, 1895–1905*. Princeton University Press, Princeton, NJ

Gourevitch P 1986 *Politics in Hard Time: Comparative Responses to International Economic Crisis*. Cornell University Press, Ithaca, NY

Jervis R 1976 *Perception and Misperception in International Politics*. Princeton University Press, Princeton, NJ

Katzenstein P (ed.) 1996 *The Culture of National Security: Norms and Identity in World Politics*. Columbia University Press, New York

Kennedy P 1987 *The Rise and Fall of the Great Powers: Economic Change and Military Conflict from 1500 to 2000*. Random House, New York

Krasner S 1978 *Defending the National Interest: Raw Materials Investments and US Foreign Policy*. Princeton University Press, Princeton, NJ

Lenin V I 1916 (1939) *Imperialism, the Highest Stage of Capitalism*. International Publishers, New York

Morgenthau H J 1978 *Politics Among Nations: The Struggle for Power and Peace*, 5th edn. Knopf, New York

Putnam R 1988 Diplomacy and domestic politics: The logic of two-level games. *International Organization* **42**(3): 427–60

Schumpeter J A 1955 *Imperialism and Social Class*. Trans. H Norden. Meridian, New York

Spykman N J 1942 *America's Strategy in World Politics: The United States and Balance of Power*. Harcourt, Brace, New York

Thucydides 1972 *History of the Peloponnesian War*. Trans. R. Warner. Penguin, New York

Trubowitz P 1998 *Defining the National Interest: Conflict and Change in American Foreign Policy*. University of Chicago Press, Chicago

Waltz k 1959 *Man, the State, and War*. Columbia University Press, New York

Zakaria F 1998 *From Wealth to Power: The Unusual Origins of America's World Role*. Princeton University Press, Princeton, NJ

P. Trubowitz

Forensic Psychiatry

A number of definitions of forensic psychiatry must be considered. The American Academy of Psychiatry and the Law (AAPL) (1985) endorses the definition of forensic psychiatry adopted by the American Board of Forensic Psychiatry Incorporated which is:

> Forensic Psychiatry is a subspecialty of psychiatry in which scientific and clinical expertise is applied to legal issues in legal contexts embracing civil, criminal, correctional or legislative matters: forensic psychiatry should be practiced in accordance with the guidelines and ethical principles enunciated by the profession of psychiatry.

In the preface to the British textbook of forensic psychiatry, its field of operation is defined as

> the overlap, interface, and interaction of psychiatry and the law in all its aspects; criminal behaviour, civil litigation, family law, the diagnosis, care and treatment of psychiatric patients where the disorder is associated with abnormalities of behaviour, legislation and numerous other problems such as the management of violence in the study of sexual deviation.

It adds that

> forensic psychiatry continues to be concerned primarily with mentally disordered offenders, but now encompasses a wide range of offences and gives much more prominence to diagnosis, management and treatment in prison, hospital or the community using the skills of psychologists, nurses, social workers, and probation officers in addition to psychiatrists.

The Canadian approach to forensic psychiatry mirrors the British definition as is seen in the recent application for subspecialty status by forensic psychiatry under the auspices of the Canadian Academy of Psychiatry and the Law. The goal of forensic psychiatry is defined as

> promoting the mental health of Canadians involved in legal matters. We aim to provide a functional interface between psychiatry and the law and to improve Canadian society as a whole through early intervention programs, better diagnosis and treatment and ultimately a reduction in offending.

In twentieth-century Europe, forensic psychiatry has generally centered on advances in the fields of neurology, psychiatry, and sciences related to these disciplines. In general terms, although there is no single definition of forensic psychiatry in Europe, it would tend to follow the British and Canadian approaches. In summary therefore, as we move into the twenty-first century there appears to be a convergence of the American definition of forensic psychiatry with the British, Canadian, and European definitions which relate to the evaluation and treatment of the mentally abnormal offender while at the same time being involved in the psychiatric assessment (for legal purposes) of fitness to stand trial, criminal responsibility in forensic psychiatry dealing with criminal matters, and similar assessments involving the civil aspects of the law. As we approach the twenty-first century, forensic psychiatry in all parts of the world appears to be moving and establishing its roots as a medical specialty and a subspecialty of psychiatry.

Even though the medical roots are emphasized, it is quite clear that the forensic psychiatrists must be familiar with basic issues in law.

1. The Concept

The forensic psychiatrist needs to be an expert clinician as well as being well versed in law and psychiatry as it applies in the jurisdiction where he or she lives. Law is understood to encompass criminal, civil, and mental health law. In medicine as well as the social sciences there is an increasing number of attempts to complete research studies involving large samples in order better to define the characteristics of disease and aberrant as well as normal behavior. This has led to progress in many areas including forensic psychiatry. This allows the individual to be classified according to whether they have certain clinical and other characteristics in common with a certain reference group. An estimate of the individual's personality characteristics can be measured against a reference sample. The difficulty in forensic psychiatry is that the legal system looks at unique characteristics of a given individual before the courts, rather than any theoretical concept of how he or she may fit a given reference group. This can potentially create a conflict where the scientific aspect of forensic psychiatry could be seen as more prejudicial than probative in the context of the legal system.

The development of forensic psychiatry in Britain, Europe, and the United States was based on historical events in the eighteenth and nineteenth centuries. In Britain the writings of Sir Mathew Hale and the trials of Arnold, Ferris, Hadfield, Oxford, and McNaughton were major landmarks in the recognition of forensic psychiatry. The development of the Bethlem Hospital and later Broadmoor as special facilities for the criminally insane were also fundamental to the growth and development of forensic psychiatry. This was extended to other facilities such as the special psychiatric facilities at Rampton and Parkhurst in the early part of the twentieth century.

The development of forensic psychiatry in Europe followed a similar path. Forensic psychiatry emerged from events in the eighteenth and nineteenth centuries. The French Revolution brought to light documents about prisons, workhouses, and civil hospitals being used to detain members of the public. This was directed specifically to the way and rationale for the detention of persons who had been found insane. Concerns over the method of detention led to pressure for scientists of the day to define insanity more precisely. The influence of religion declined and this allowed the scientific roots of modern clinical psychiatry to develop. Philippe Pinel had an enormous influence in his *Medico-Philosophical Treatise on Mental Alienation or Mania*, which was rapidly translated into English and German. This led to a system of classifying mental disorders that was soon adopted in Europe and England. This helped in the development of forensic psychiatry as it brought more consistency to the testimony in insanity hearings before the courts. Esquirol, Morel, Lombroso, Krafft-Ebing, and others contributed to these important advances. As in Britain, special institutions for the criminally insane were developed and further research occurred in these institutions.

The United States saw a similar development with Benjamin Rush delivering the first lecture on forensic psychiatry under the guise of medical jurisprudence. The lecture covered topics ranging from civil incompetence caused by psychiatric disorder, and financial incompetence, to mental disorder and criminal responsibility. Rush was followed by Beck who in 1823 published *Elements of Medical Jurisprudence*. He referred to British and French psychiatry as well as Benjamin Rush in his book. Isaac Ray followed Rush and Beck and published the *A Treatise on the Medical Jurisprudence of Insanity* in 1838. This was used to defend Daniel McNaughton. Ray's approach in which he carefully studied the law and the judicial opinions of the day—in particular, anything in the law that obstructed the psychiatrically ill (the insane) being able to receive treatment. He became the superintendent of the Maine Insane Asylum and later the Butler Hospital. Perhaps most importantly he was one of the original founders of the American Psychiatric Association, and was its President from 1855 to 1859. In 1869 Ray went on to write a paper *Confinement of the Insane* that has remained fundamental to forensic psychiatry throughout the twentieth century.

> In the first place, the law should put no hindrance in the way of the prompt use of those instrumentalities which are regarded as the most effectual in promoting the comfort and restoration of the patient. Secondly, it should spare all unnecessary exposure of private troubles and all unnecessary conflict with popular prejudices. Thirdly, it should protect individuals from wrongful imprisonment. It would be objection enough to any legal provision, that it failed to secure these objects, in the completest manner. (Ray 1869)

Highly significantly, Quen writing in *The Principles and Practices of Forensic Psychiatry* (1990) reports 'These elements remain necessary, valid, and unachieved in America today.'

As outlined above forensic psychiatry was fundamental to the development of psychiatry in the eighteenth and nineteenth centuries and was adopted by the founders of British, European, and American psychiatry as part of their role to promote psychiatry as a medical specialty.

2. Criminal Law and Forensic Psychiatry

This involves fitness-to-stand-trial evaluations and assessments for criminal responsibility. It involves the forensic psychiatric assessment of mentally abnormal

offenders. The forensic psychiatrist is, by virtue of his or her professional training, an expert in psychiatry and specifically in mentally abnormal offenders. This expertise is applied to legal matters such as fitness to stand trial and criminal responsibility. The challenge to the forensic psychiatrist is that the traditional doctor–patient relationship does not exist under these circumstances. The traditional aims of psychiatry of treating or helping the patient are also not obvious and may even be contradicted. The definition of forensic psychiatry by the well-known American forensic psychiatrist Dr Seymour Pollack (1974) addresses these contradictions. Pollack defined forensic psychiatry thus:

> forensic psychiatry is limited to the application of psychiatry to evaluations for legal purposes. Psychiatric evaluation of the patient is directed primarily to the legal issues in which he is involved, and consultation is concerned primarily with the ends of the legal system, rather than the therapeutic objectives of the medical system.

Pollack is correct in terms of evaluations for fitness to stand trial and criminal responsibility. In the former, the initial role of psychiatric evaluation is to establish the diagnosis, from the standpoint of either a medical illness or a psychiatric disorder affecting the person's mental functioning. Once this has been established the next task for the forensic psychiatrist is to assess whether the degree of disorder present affects the person's capacity to proceed to trial. The legal basis may vary slightly between jurisdictions but all are similar and would be analogous to Section 2 of the Criminal Code of Canada, which defines 'unfit to stand trial' as follows:

> unfit to stand trial means unable on account of mental disorder to conduct a defence at any stage of the proceedings before a verdict is rendered or to instruct counsel to do so, and, in particular, unable on account of mental disorder to
> (a) understand the nature and object of the proceedings
> (b) understand the possible consequences of the proceedings, or
> (c) communicate with counsel (Section 2 Criminal Code of Canada 2000).

The process is clearly conducted for a legal purpose and, arguably, not necessarily in the person's best interest, as it may mean finding them fit for trial which may in turn result in a guilty verdict and possibly lengthy incarceration in a correctional setting. Alternatively, if they were found unfit and remained unfit, while they would still be incarcerated, they would probably be detained in a hospital setting. In Canada and most other jurisdictions, a finding of unfitness means that a judicial hearing would follow as to whether psychiatric treatment would render the individual fit to stand trial. In Canada this could result in a treatment order issued by a judge, ordering psychiatric treatment (most commonly antipsychotic medication) to render the person fit to stand trial. This can be done independently of the person's capacity to consent or refuse psychiatric treatment.

An assessment of criminal responsibility involves judging whether the person's mental status at the time the crime took place rendered him or her incapable of appreciating the nature of their actions as well as knowing whether their actions were wrong. It is beyond the scope of this article to go into criminal responsibility in detail as the legal basis varies considerably in different jurisdictions. This particularly true in the United States where the criminal law is under state jurisdiction and therefore each state varies in terms of what it uses as the legal basis for a finding of not criminally responsible on the basis of mental disorder. The most commonly used standard in Western countries is the McNaughton Test or some modification of the test. In recent years a number of American states have either abolished or severely limited the application of 'insanity' (mental disorder) to reduce the *mens rea* component of criminal responsibility. In order to be criminally responsible for a crime the state must prove the accused committed the unlawful act (*actus reus*) and that they also had the intent to commit the crime (*mens rea*). Mental disorder if present at the time the crime was committed can negate *mens rea*, leading to a finding of 'not guilty by reason of insanity,' not criminally responsible on the basis of mental disorder, or guilty but mentally ill. Such a verdict usually results in confinement in a forensic psychiatric facility until the acquitee is regarded as having recovered and no longer being dangerous to the community at large. In other cases, a mental disorder may result in a diminishing of criminal responsibility, allowing the accused to be found guilty of a lesser offence carrying a less severe penalty. This is seen in situations where a homicide has been committed and the individual suffered from a mental disorder at the time of the homicide. The degree of mental disorder is regarded as not sufficient to negate *mens rea* but to reduce it. This could result in a finding of manslaughter as opposed to murder where the intent to kill was reduced although the accused was still criminally responsible for the death of the victim. The original McNaughton Test was strictly a cognitive test. In its original form it was:

> To establish a defence on the grounds of insanity, it must be clearly proved that at the time of committing the act, the party accused was labouring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing, or if he did know it, that he did not know he was doing what was wrong. (Quoted in Hunter and MacAlpine 1963.)

Various modifications of the original McNaughton Test have occurred in the United States and elsewhere. The most important of these has been the adoption, in the United States, of the American Law Institute Test (ALI Test), which includes both cognitive and volitional aspects. To complicate things even further the

actual legal interpretation of the meanings of the various tests of criminal responsibility is subject to case law. This means the interpretation may change, depending on a specific interpretation by a particular judge in a specific court case that is upheld by various appeal courts. Forensic psychiatrists who act as expert witnesses in cases of criminal responsibility based on mental disorder are therefore becoming involved in a very complicated process where they have to use their psychiatric skills initially to establish a psychiatric or medical diagnosis. Drawing on the observations of others as well as the medical history of the accused, they must establish, by a process of reconstruction, whether the disorder was present at the time of the offence. If it was present, they must assess its degree or severity and how that would affect the operating mind of the accused both cognitively and possibly volitionally. Then they must form an opinion with reference to the legal test for criminal responsibility in the particular jurisdiction where the trial is being held. He or she will need a working knowledge of the case law interpretations as they apply in the specific jurisdiction and in the specific case. Finally, as expert witness he or she must understand the facts of the case as they emerge at the time of the trial and be able to incorporate them into the opinion being rendered.

In addition to criminal responsibility and fitness to proceed with trial, forensic psychiatrists are also involved in the sentencing of mentally abnormal offenders. Here the role of an expert is to address the third principle of sentencing, which is rehabilitation (the first two being retribution and deterrence). In certain jurisdictions this may be the most prominent role for the forensic psychiatrist as it is used to mitigate the sentence while at the same time allowing the judge to direct treatment in a correctional facility for a psychiatric disorder. It therefore provides a link between forensic psychiatry and correctional psychiatry. This linkage varies in different jurisdictions and countries according the degree of development of forensic psychiatry and the availability of treatment resources in the correctional system. A highly contentious issue related to sentencing is the participation of forensic psychiatrists in the death penalty process in the United States, including pretrial assessment, the sentencing phase, the determination of competence to be executed, and most controversially the assessment of restoration of competence to be executed as a result of psychiatric treatment.

3. *Civil Law and Forensic Psychiatry*

This branch of forensic psychiatry involves psychiatric malpractice; psychiatric disability determinations; personal injury evaluations; testamentary capacity and guardianship; and family law matters, including parental competence and child custody evaluations. Included here is the specialty of the child and adolescent forensic psychiatrist. These individuals must combine the skills of two psychiatric subspecialties, namely, child and adolescent psychiatry, and forensic psychiatry. In some university divisions of forensic psychiatry these skills are available through a family court clinic. This clinic would receive consultations from the family court system to evaluate parental competence in child abuse and neglect cases and also assist in determining the best interests of the child in divorce custody cases. Family court clinics have also traditionally provided evaluation of young offenders. The evaluation of young offenders, principally adolescents, requires the clinical skills and experience necessary for evaluating children and adolescents. The importance of these dual skills cannot be underestimated and at times has called for the recognition of a sub-subspecialty of child and adolescent forensic psychiatry in some countries.

Testamentary capacity may be a relatively small component of forensic psychiatry, but there is no doubt as to its complexity and the special skills necessary for these retrospective evaluations and being able to communicate the evaluation to the legal authorities.

4. *The Legal Regulation of Psychiatry and the Forensic Psychiatrist*

The legal regulation of psychiatry involves informed consent and competence; the right to refuse psychiatric treatment and the right to psychiatric treatment; psychiatric hospitalization and involuntary civil commitment; outpatient civil commitment; confidentiality; the duty to warn or protect against professional misconduct, including the impaired physician, competence to practice, and professional sexual misconduct. These are areas of psychiatry and the law where the average psychiatrist also needs to develop skills and competence and are, therefore, not an exclusive area of practice for the forensic psychiatrist. The forensic psychiatrist, however, contributes special skills and training in the legal system, and the skill to be a competent expert witness. This means their role is to assist the non-forensic psychiatrist by consultation and advice and, in the more complex cases, forensic psychiatric expertise may be desirable.

5. *Special Clinical Skills of the Forensic Psychiatrist*

A wide range of special clinical skills is unique to the forensic psychiatrist. Special skills in evaluating and treating violent psychiatric patients are a fundamental requirement for all forensic psychiatrists. This is also expanded to include the assessment of the danger posed by an offender and risk assessment, as well as the risk management of violent mentally abnormal

offenders. Associated with these clinical skills are skills in assessing antisocial personality disorder as well as impulse control disorders. Because of the association between organic brain dysfunction and impulsivity and violence, forensic psychiatrists need to develop skills in assessing the neuropsychiatric aspects of psychiatric conditions that can result in violence or can affect competence. They also need to be able to evaluate and treat the most severe and violent major mental disorders. Forensic psychiatrists must develop special skills in evaluating malingering and the use of hypnosis. The forensic psychiatric aspects of certain psychiatric disorders (e.g., multiple personality disorder) are also within the skill set of the average forensic psychiatrist. The assessment and treatment of sexual deviation is also a required skill.

6. *Correctional Psychiatry*

The strict definition of correctional psychiatry is 'psychiatry practiced within a correctional system.' Broader definitions would include any agency of social control that is involved in the treatment and rehabilitation of adult or juvenile offenders. These populations have significant psychiatric morbidity and clearly would benefit from psychiatric intervention and this had led to a growth of correctional psychiatry in the last few decades. There is a close relationship between correctional psychiatrists and forensic psychiatrists with correctional psychiatric programs often being outreach programs of university departments of forensic psychiatry (as, for example, in Canada).

7. *Organization and Future of Forensic Psychiatry*

This is a subject of debate in North America at the beginning of the twenty-first century. There is a consensus that the future growth and development of forensic psychiatry must lie within the field of medicine as a subspecialty of psychiatry. This is already well established in general terms in Canada and Europe. If we accept this consensus, some comment as to the ideal forensic program is warranted. This author strongly endorses the subspecialty status of forensic psychiatry as a critical component of its future development. This implies that the ideal forensic program would be an academic division within a university department of psychiatry. In turn, this would mean that education, clinical training, and research would be fundamental components of the program. General psychiatric residents should be required to do a rotation as part of their residency requirement. As the largest interface between medicine and the law is between psychiatry and the law, it should be part of the basic training of all psychiatric residents. It is almost impossible for the average psychiatrist to go through his career without some interaction with law and psychiatry tribunals, the criminal justice system, or the civil legal system.

An ideal program would provide fellowship training in forensic psychiatry in an accredited program for a minimum of one year. The forensic clinical program should have a medium or maximum inpatient assessment unit to provide pretrial evaluations of fitness and criminal responsibility. In addition, inpatient and outpatient facilities are required for the treatment and rehabilitation of individuals who have been found unfit to stand trial or not criminally responsible on the basis of mental disorder. The length of stay should allow for both acute and chronic care so a cautious approach to treatment and rehabilitation would prevail. There also need to be outreach programs to correctional facilities and an active participation in pretrial diversion programs. In an ideal situation, specialized clinics for the assessment and treatment of sexual deviation, and aggression and impulse control disorders should also be present. A family court clinic with child and adolescent forensic psychiatrists on the staff should also be included. A multidisciplinary team approach to assessment and treatment is essential. Research into all aspects of forensic psychiatry is critical for its future growth and the development of its profile as a medical subspecialty.

See also: Child and Adolescent Psychiatry, Principles of; Civil Law; Criminal Defenses; Criminal Justice, Ethics of: Aspects of Human Dignity; Expert Witness and the Legal System: Psychological Aspects; Law: Defense of Insanity

Bibliography

American Academy of Psychiatry and the Law (AAPL) 1985 Definition of Forensic Psychiatry.

Barras V, Bernheim J 1990 The history of law and psychiatry in Europe. In: Blueglass R, Bowden P (eds.) *Principles and Practice of Forensic Psychiatry*. Churchill Livingstone, London, pp. 103–9

Blueglass R, Bowden P (eds.) 1990 *Principles and Practice of Forensic Psychiatry*. Churchill Livingstone, London

Bradford J M W, Glancy G 2000 Commentary on 'Ensuring that forensic psychiatry thrives as a medical specialty in the 21st century.' *The Journal of the American Academy of Psychiatry and the Law* **28**: 20–2

Forshaw D, Rollin H 1990 The history of forensic psychiatry in England. In: Blueglas R, Bowden P (eds.) *Principles and Practice of Forensic Psychiatry*. Churchill Livingstone, London, pp. 61–101

Glancy G, Bradford J, Gagne P 2000 Psychiatry and the law in Canada: A proposal for subspecialization: Part I. *Bulletin of the Canadian Psychiatric Association*, pp. 20–6

Hunter R, MacAlpine I 1963 *Three Hundred Years of Psychiatry 1535–1860. A History Presented in Selected English Texts*. Oxford University Press, London

Quen J M 1990 The history of law and psychiatry in America. In: Blueglass R, Bowden P (eds.) *Principles and Practice of Forensic Psychiatry*. Churchill Livingstone, London, pp. 111–16

Pollack S 1974 *Psychiatry in Criminal Law*. University of Southern California, Los Angeles
Ray I 1869 Confinement of the insane. *American Law Review* **3**: 193–217
Rosner R (ed.) 1994 *Principles and Practice of Forensic Psychiatry*. Chapman and Hall, New York
Weinstock R, Leong D B, Silva A J 1994 Defining forensic psychiatry: Roles and responsibilities. In: Rosner R (ed.) *Principles and Practice of Forensic Psychiatry*. Chapman and Hall, New York
Watt D, Fuerst M 1998 *The 1998 Annotated Tremeear's Criminal Code*. Carswell, Toronto, ON

J. Bradford and G. Glancy

Forensic Psychology

Forensic psychology is the branch of psychology concerned with the production and application of psychological knowledge and principles within the legal process. (The term 'forensic,' meaning 'legal' or 'judicial,' is derived from the Latin word *forensis*, which literally means 'relating to the forum or market.' In ancient Rome, a forum was an open square, which served as a market, as well as a place where legal matters were disposed of.)

1. Historical Background

Since both psychology and law are concerned with human behavior, they have been intertwined since the dawn of recorded history (Gudjonsson and Haward 1998). Indications of interest shown in the law by philosophers, from whose ranks psychology would later develop, are found in the earliest Asian, Egyptian, Greek, and Latin writings. Later Greek philosophers continued to express themselves on practical questions of law and topics that today constitute the subject matter of psychology. For example, Plato (ca. 427–347 BC) in his *Republic* considered the question whether mentally disordered people were responsible for their behavior (Allan 1995). However, it took more than 2000 years for psychological knowledge to mature to a level where it was officially allowed in the courtroom for the first time. This happened in 1896 when Albert Von Schrenk-Notzing, student of Wilhelm Wundt who founded the first psychology laboratory in Germany in 1879, testified in a murder trial concerning the effect of pretrial publicity on memory and suggestibility (Blau 1998). Since then, forensic psychology has made great strides. As we enter the twenty-first century, there is hardly an area of the law where psychological expert testimony is not implemented. Ewing (1985) believes that the modern justice system could not function without significant participation by behavioral scientists such as psychologists.

2. Roles of the Forensic Psychologist

As both psychology and law essentially cover all aspects of human behavior, it is understandable that the forensic psychologist serves as an expert on a variety of matters including those considered in the following sections.

2.1 Competency to Stand Trial

Also known as fitness to plead, competency to stand trial refers to the defendant's mental capacity to make a proper defense, i.e., to understand the legal proceedings and the charges against him or her.

Contrary to popular opinion, the mere existence of a mental disorder *per se* does not render an individual incompetent to stand trial. Many mental disorders are of such a nature and intensity that the defendant's competency is not significantly influenced. The crucial question is whether the mental disorder affects the defendant's current functional abilities within the context of relevant legal criteria for determining competency (Gudjonsson and Haward 1998).

As laws and 'legal cultures' differ from country to country, it is understandable that statistics concerning competence to stand trial will vary considerably. For example, proportionally speaking, about 100 times more defendants in the USA are found incompetent than in England and Wales. Even within one country, the percentage of defendants found to be incompetent often varies between 1 percent and 80 percent across jurisdictions.

It does seem that individuals adjudicated incompetent to stand trial often have certain characteristics in common, such as a history of mental health treatment (especially for schizophrenia and depression), lower intellectual functioning, and a present charge of violent crime.

The evaluation of competency is usually very reliable and the agreement rate between psychologists often exceeds 80 percent. Next to the traditional assessment techniques, particularly the clinical interview, several competency assessment instruments are also used.

2.2 Criminal Responsibility

Criminal responsibility (liability/accountability) should not be confused with competency to stand trial. Competency to stand trial is specifically concerned with the defendant's mental capacity during the legal proceedings. Criminal responsibility, also known as the insanity defense, refers to the time of the alleged offense and more specifically asks the question whether the defendant had the mental capacity to distinguish between right and wrong at the time of the alleged offense and could control his or her conduct accordingly. Should the defendant be found not criminally

responsible, he or she rarely goes free. The person is usually referred to a mental health facility where he or she will remain as long as the commitment criteria are met.

The insanity defense is one of the most controversial issues in criminal law. One reason for this are the many misconceptions surrounding this plea. For example, although the public believes (probably as a result of the publicity given to such cases) that the insanity plea is very often entered, such cases are rare: often fewer than 1 in every 1,000 criminal cases. The success rate of the defense varies from country to country and from jurisdiction to jurisdiction, with 25–50 percent as a broad average.

Most defendants whose insanity pleas are successful are diagnosed with major psychotic disorders and have extensive mental health histories, often accompanied by prior civil commitments or findings of incompetence (Golding et al. 1999).

Insanity evaluations are usually regarded as more complex than competency evaluations. It is, after all, exceedingly difficult to determine what another person was thinking, feeling, and doing at the time of an offense that usually happened weeks, months, or even years earlier (Bartol and Bartol 1994). However, it seems that specialized training and experience can lead to an agreement rate between mental health professionals of more than 80 percent. Several measuring instruments have been developed to assist mental health professionals in their assessment of the insanity plea.

2.3 Other Defenses

Next to competency to stand trial and criminal responsibility, the expertise of psychologists is employed in several other defenses in both criminal and civil cases.

(a) *Intoxication* by alcohol or other drugs, because it is self-induced and temporary, is seldom given complete exculpatory effect (Melton et al. 1997). The major exception is when the substance abuse has been prolonged to a point where it has produced a serious mental disorder. However, in most countries voluntary intoxication can be used as a diminished capacity defense claim. In such cases mental health professionals have to produce evidence that the intoxication did indeed affect the defendant's mental capacity significantly.

(b) *Amnesia* (the total or partial loss of memory) at the time of the offense is one of the most common defenses heard in the criminal court, especially concerning violent crimes (see *Amnesia*). It is also one of the most difficult defenses to assess. It is not only difficult to determine the weight that should be given to the claimed amnesia, but also to distinguish between genuine and simulated amnesia. An even more complex task is to determine whether the defendant was really amnesic when committing the crime, or whether he or she only developed the amnesia afterwards. Although psychologists frequently give expert testimony on amnesia, their ability to assist the court with scientific evidence is questioned. Courts are therefore often skeptical about accepting their testimony as a viable defense element.

(c) *Dissociative identity disorder*, formerly called *multiple personality disorder*, is primarily characterized by the presence of two or more distinct personalities that in turn take control of the person's behavior (American Psychiatric Association 1994). However, the disorder is highly controversial and sometimes called the 'UFO of mental health.' Many mental health professionals believe that the disorder does not exist and amounts to no more than role playing. Although screening instruments and structural interviews have been developed and hypnosis is used to diagnose dissociative identity disorder, experts admit that it is very difficult to distinguish between real cases and malingering (Alloy et al. 1999). It is therefore understandable that legal professionals tend to be very cautious about making use of this defense. However, in rare cases in several countries, dissociative identity disorder has been accepted as a defense, ranging from an insanity plea to extenuating circumstances.

(d) *Post-traumatic stress disorder* (PTSD) is characterized by intensely distressing symptoms (e.g., irritability and sleeplessness) caused by exposure to a traumatic event such as rape, death, combat, accident, and even harassment and difficult work conditions (see *Post-traumatic Stress Disorder*). PTSD can therefore be a factor in both criminal and civil cases. However, establishing the existence of PTSD poses several difficulties for the forensic psychologist. For example, the validity of the diagnoses is often questioned. Although psychological measuring instruments and techniques to assess PTSD are available, they are usually based on self-report. The result is that malingering cannot be ruled out easily.

2.4 Prediction of Violent Behavior

As one of the primary duties of the court is to protect society, assessing the offender's potential for future violence often forms an integral part of the judicial process. This is especially true if the offender is charged with a violent crime. Determining the propensity for future violence is therefore an important part of the forensic psychologist's functions. However, the role of the mental health professional in violence prediction is controversial (see *Clinical Psychology: Validity of Judgment*). The main reason for this is that past research showed that mental health professionals have a bad track record when it comes to predicting the probability of a person committing future violence, and even more so when it comes to predicting the type and severity of such potential violence. As such,

professionals can pay a heavy social, professional, and legal price for failing to predict violent behavior, they tend to 'play it safe' and therefore often overpredict. The result is that research has shown that more than 50 percent of persons predicted likely to commit future violence, do not do so.

Ongoing research in this field is fortunately beginning to bear fruit. For example, several variables have been identified which can be regarded as risk factors for violence, e.g., a history of violence, substance abuse, symptoms of a psychosis or antisocial personality disorder (psychopathy), weapon availability, early onset of violent behavior, and absence of social support. Measuring instruments to assist psychologists have also been developed.

2.5 *Eyewitness Testimony*

'It is … inevitable that what a witness says in court will be a subjective and incomplete version of 'the truth' interspersed with small portions of fiction.' Although most psychologists will agree with this statement by Ainsworth (1998, p. 3), the testimony of an eyewitness still remains a very important part of evidence presented in both criminal and civil cases. However, as a result of a rapidly growing data bank in this widely researched area, especially in the 1980s and 1990s, the courts are currently examining some of their assumptions about the reliability and thus the value of eyewitness testimony. It is therefore understandable that courts increasingly rely on psychologists to provide expert testimony in this regard (see *Expert Witness and the Legal System: Psychological Aspects*).

The following are some of the more important research findings that have been used in court by psychologists: (a) eyewitnesses tend to be less accurate in stressful situations; (b) people experience difficulty making cross-racial identifications; (c) eyewitnesses focus on weapons rather than faces; (d) memory decay starts immediately after a perception; (e) memory gaps are often and easily replaced by unconscious fabrication or by external suggestions; (f) there is not necessarily a correlation between the degree of certainty evinced by the eyewitness and accuracy (Melton et al. 1997); and (g) although younger children are often more suggestible and remember less detail than older children and adults, they do not necessarily have poorer memories (Saywitz and Goodman 1996).

2.6 *Personal Injury*

The world is progressively becoming a litigious society in which monetary compensation is sought for almost any maloccurrence (Tsushima and Anderson 1996). As many such claims involve 'psychological pain and suffering,' psychologists are increasingly becoming involved in so-called mental injury cases. Although determining the presence and cause of mental injury constitutes the bulk of psychologists' work in this field, they are also involved in injury claims of a more physical nature, e.g., to assess brain damage or determine the psychological effect of losing a limb.

To assess whether the plaintiff does indeed suffer from a mental injury is not always easy. First, it has to be proved that there is a causal relationship between the specific incident and the plaintiffs symptoms, i.e., that the symptoms did not already exist before the incident or were not caused by other factors after the accident. Second, as possible financial remuneration is involved, plaintiffs often tend to exaggerate or even fake their symptoms. No foolproof method to determine such relationships and distinctions is available. However, by using the most objective, valid, and reliable tests appropriate to the complaints, conducting thorough interviews, and carefully recording case notes and test reports, psychologists can make a valuable contribution towards clarifying the extent to which psychological factors may be involved in personal injury cases (Butcher and Miller 1999).

2.7 *Sentencing*

If a defendant is found guilty of the charge against him or her, the court has to impose an appropriate punishment. In most countries the aim of such punishment consists of a combination or all of the following: deterrence, retribution, prevention, and rehabilitation. Since these goals have a strong psychological element and psychologists are supposed to be well equipped to provide the court with a psychological profile of the defendant, they often testify in this regard. For example, the psychologist can provide the court with information concerning the defendant's potential for future violence, the sincerity of the repentance shown, the possibility of rehabilitation, a variety of mitigating factors, and the effect of different sentences on the defendant as a unique individual.

Although there seems to be a high correlation between psychologists' recommendations and the final sentencing decision of the courts, it remains an open question as to what extent the courts do indeed take the testimony of the psychologist into account or merely use it to 'scientifically' justify decisions they have already made (Melton et al 1997).

2.8 *Profiling*

Criminal profiling refers to the process in which the nature of a crime is used to make inferences about the personality and other characteristics of the likely offender.

The sensational and dramatic elements of profiling portrayed in various well-known movies, television

series, and books (usually those depicting a serial killer) seem to be in sharp contrast with reality. Profiling is not widely accepted in the psychological and legal community, and some courts have even ruled profiling testimony inadmissible. There are two main reasons for this (Gudjonsson and Haward 1998). First, a criminal profile only gives a broad indication of the type of person who may have committed the crime. It does not indicate a specific individual who happens to fit the profile. The profiler is therefore unable to say whether it is more probable than not that a specific offender did, in fact, commit the crime. Second, there is no scientific evidence to support the reliability and validity of criminal profiling in solving crimes. Indeed, it seems that when profiling does assist the police in solving a case or in opening up new lines of inquiry, it is the exception rather than the rule.

On the other hand, it cannot be denied that criminal profiling has proven helpful in some, albeit exceptional, cases. However, much more research needs to be done before criminal profiling will earn its place as a valuable forensic tool.

A variation of profiling that has evoked some interest in the news media is psychological autopsy, which comprises the compilation of a psychological profile of especially well-known deceased individuals. It is also used in suicide cases, for example, to determine whether the deceased could indeed have committed suicide. However, as in the case of criminal profiling, its reliability and validity are also questioned.

2.9 Child Custody

With the worldwide increase in divorce rates and the resulting child custody battles between divorcing parents, mental health professionals have come to play an increasingly important role in such disputes.

Two principles play a major role in courts' decisions in child custody cases: the tender-years principle and the best-interest-of-the-child principle. The tender-years principle reflects the traditional belief that the mother is inherently better equipped to care for children of a 'tender age' (which particularly refers to children around or under the age of seven). However, several factors have led to the questioning and even abolishment of this principle in some courts. For example, research has shown that fathers and mothers are equally capable of fulfilling the psychological and other needs of children in their 'tender years.' The principle is also seen as a violation of equal gender rights. Most courts base their decisions in child custody cases on the best-interest-of-the-child principle, which involves criteria such as (a) the love, affection, and other emotional ties between parent and child, as well as between the child and his or her siblings or any other significant person; (b) the ability of the parent to provide for the child's psychological, educational, cultural, and environmental development; (c) the ability of the parent to provide the basic physical needs of the child, e.g., food, clothing, and housing; (d) the mental and physical health and moral fitness of the parent; and (e) the child's preference as to his or her custodian. However, this principle has also come under fire from critics who, for example, claim that many of the criteria are vague, difficult to assess, and open to subjective interpretation.

Assessment in child custody disputes usually includes interviewing the parents and children, as well as any other third-party source that may provide valuable information, observation of the interaction between especially parents and child, and the use of traditional psychological and specialized tests. However, the interview is usually recommended as the best method. The reasons are the following. The observation technique is regarded as artificial and thus unreliable (the individuals know that they are being observed). The traditional psychological tests are often irrelevant (e.g., there is no correlation between good parenthood and IQ). The specialized tests are also not very reliable and valid. It is generally accepted that a psychologist, acting according to professional guidelines, can make a valuable and often decisive contribution in child custody cases.

3. Forensic Assessment

Forensic assessment refers to the procedures, methods, and techniques that psychologists use to gather information in order to assist the court in coming to a final decision. The nature of the assessment will depend on the theoretical orientation, training, experience, and preference of the psychologist, as well as the nature of the specific case. Usually a combination of procedures, methods, and techniques are used.

The forensic interview can be considered the central element of the assessment process. The interview can be structured (the questions being determined and systematized in advance), unstructured (the interview being allowed to develop spontaneously), or a combination of the two. The forensic interview is much more than just an 'ordinary' interview. It is a scientific method for gathering information and requires specialized training and experience. As mentioned, it is usually necessary to interview not only the specific individual, but also family members, friends, and other third-party sources. Such collateral information may be helpful not only in acquiring a broader picture, but also in determining the reliability of the information provided by the different sources.

In addition to this basic assessment method, several others are employed, and examples are given below.

3.1 Psychological Testing

Virtually all forensic psychologists incorporate psychological tests in their assessment armament. Psycho-

logical test results can be powerfully persuasive instruments because of society's long-standing high regard and respect for these measuring instruments (Tsushima and Anderson 1996). Many courts also seem to expect that such tests will be used.

However, it has also become clear that psychological tests have not always conformed to the high qualities attributed to them. Not only are the reliability and validity of many of the tests used by psychologists generally disputed, but also they have not been developed specifically for forensic issues. The result is that their usefulness in forensic settings is questioned. Attempts to develop specialized forensic assessment instruments have also not yet been very successful, mainly owing to psychometric flaws. Another problem especially encountered in underdeveloped and developing countries is that tests are used on groups for which they have not been developed. An example is the worldwide use of various American tests developed for the US population on groups for which they were not standardized. Cultural differences cast doubt on the reliability of such results.

On the other hand, can there be little doubt that, in the hands of a well-trained and experienced examiner who is well aware of the pitfalls, psychological tests can be very valuable tools and can provide valuable information.

3.2 Forensic Hypnosis

There are few topics in psychology more shrouded in mystery and misconception than hypnosis. Mainly responsible for this is the sensational representation of this technique by the movie industry, news media, and stage hypnotists. Even in the scientific world controversy surrounds the hypnosis phenomenon, with some regarding it as a valuable tool in many areas, while others do not believe in the existence of such a state, and mainly contribute it to suggestion and role-playing (see *Hypnosis, Psychiatry of*).

The primary use of forensic hypnosis centers on enhancing the memory of a witness, victim, or even defendant, especially with a view to helping him or her remember forgotten or vague details of a crime. Most research shows that, compared with people who have not been hypnotized, people under hypnosis tend to recall more detail about an event. They also usually feel confident that their recollections are correct. However, their stories contain a great deal of inaccurate information, even more so than in the case of nonhypnotized subjects. It is generally accepted that forensic hypnosis is seldom more efficient in improving memory than any other standard psychoforensic procedure. Courts therefore hesitate to accept testimony acquired through hypnosis. In some countries such testimony is even inadmissible. Others do allow it, but usually only if certain safeguards and guidelines are applied. It is therefore important than forensic psychologists meet these preconditions if they wish to employ this controversial forensic tool.

3.3 Polygraph

The polygraph, also called 'lie detector,' is an apparatus that measures physiological variables such as heart rate, blood pressure, and respiration rate. Polygraphers believe that when people are asked threatening questions to which they respond by lying, physiological changes will occur which will show up on the polygraph record. This makes the polygraph a potentially valuable forensic tool.

Unfortunately, the validity of the polygraph leaves much to be desired. Polygraph results seem to be wrong in at least 30 percent of all cases. This concerns innocent people who are wrongly accused, as well as guilty persons who plead innocence. For this reason, most courts do not admit evidence based on polygraph results, or admit it only if it adheres to certain stipulations (e.g., that all parties agree to the test before it is administered).

Ironically, it seems that one of the polygraph's biggest strengths lies in people's belief that it is reliable. It seems that about 30 percent of suspects confess after they have failed a polygraph test (Memon et al. 1998). However, it does not necessarily mean that these suspects had been lying. It could also be that innocent suspects tend to confess because, having failed the test, they see little possibility of convincing the court of their innocence thereafter. By confessing they hope to get a lighter sentence.

4. The Future

Only a few decades ago the relationship between psychology and law was compared to an unhappy marriage. In the twenty-first century, it is rather regarded as a 'normal' marriage with its ups and downs, but definitely not heading for a divorce. There can be no doubt that both professions find the interaction and collaboration enriching.

However, in order to make full use of this interdependent relationship, much more needs to be done. The key lies in research, aimed not only at developing new procedures, methods, and techniques, but also at determining weaknesses in the general professional approach of forensic psychologists. Forensic psychologists should especially try to move away from the 'hired guns' stigma attached to them. This can only be achieved by telling 'the truth and nothing but the truth.' This principle can be best served by presenting objective and critical both-sides-of-the-story evidence without subjectively favoring one party.

The saying, 'It is not a question whether forensic psychology is good, but what it is good at,' should also serve as a principle guideline for acknowledging that there are areas where little or no contribution can made and sticking to those areas where it is clear that psychologists are held in high esteem by courts. With such an approach and quality research, there can be no doubt that forensic psychology will flourish in the future.

See also: Confessions: Psychological and Forensic Aspects; Criminal Defenses; Expert Testimony; Expert Witness and the Legal System: Psychological Aspects; Forensic Psychiatry; Liability: Legal

Bibliography

Ainsworth P B 1998 *Psychology, Law and Eyewitness Testimony*. Wiley, New York

Allan A 1995 The role of the psychologist in South African courts. Ph.D. dissertation, University of the Orange Free State, Bloemfontein

Alloy L B, Jacobson N S, Acocella J 1999 *Abnormal Psychology. Current Perspectives*, 8th edn. McGraw-Hill, New York

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*, 4th edn. American Psychiatric Association, Washington, DC

Bartol C R, Bartol A M 1994 *Psychology and Law. Research and Application*, 2nd edn. Brooks/Cole, Pacific Grove, CA

Blau T H 1998 *The Psychologist as Expert Witness*, 2nd edn. Wiley, New York

Butcher J N, Miller K B 1999 Personality assessment in personal injury litigation. In: Hess A K, Weiner I B (eds.) *The Handbook of Forensic Psychology*, 2nd edn. New York, Wiley, pp. 104–26

Ewing C P 1985 *Psychology Psychiatry and the Law: a Clinical and Forensic Handbook*. Professional Resource Exchange, Sarasota, FL

Golding S L, Skeen J L, Roesch R, Zapf P A 1999 The assessment of criminal responsibility: current controversies. In: Hess A K, Weiner I B (eds.) *The Handbook of Forensic Psychology*, 2nd edn. Wiley, New York, pp. 379–408

Gudjonsson G H, Haward L C R 1998 *Forensic Psychology. A Guide to Practice*. Routledge, London

Melton G B, Petrila J, Poythress N G, Slobogin C 1997 *Psychological Evaluations for the Courts*, 2nd edn. Guilford Press, New York

Memon A, Vrij A, Bull R 1998 *Psychology and Law. Truthfulness, Accuracy and Credibility*. McGraw-Hill, New York

Saywitz K J, Goodman G S 1996 Interviewing children in and out of court. In: Briere J, Berliner L, Bulkley J A, Jenny C (eds.) *The APSAC Handbook of Child Maltreatment*. Sage, Thousand Oaks, CA

Tsushima W T, Anderson R M Jr 1996 *Mastering Expert Testimony*. Erlbaum, Mahwah, NJ

D. Louw

Foucault, Michel (1926–84)

Michel Foucault is one of the most influential, and also most controversial, social thinkers of the century. His work represents a sustained effort to reconstruct the way in which the specifically modern Western form of subjectivity has become what it is through institutionalized practices of truth-telling and the search for knowledge.

1. Foucault's Intellectual Trajectory

Foucault was born on October 15, 1926 in Poitiers. His upbringing was organized around knowledge since his early childhood (he started to attend school at the age of three), while major political events, especially wars (the Ethiopian war, the Spanish Civil War, and World War II) also had a major impact on his formation. In September 1945 he moved to Paris where he listened to the lectures of Jean Hyppolite during a preparatory high-school year, and was then admitted to the prestigious *Ecole Normale Supérieur*. These were troubled years for Foucault, intellectually, personally, and politically alike. He had a short spell with the Communist Party and had difficulties in managing his homosexuality.

After graduating in philosophy in 1951, he received further qualifications in 1952 and 1953 in psychopathology and experimental psychology. He then started teaching, working on his thesis and also attending the St Anne psychiatric hospital. However, under the impact of an intensive reading of Nietzsche, his interests gained focus, and he decided to leave the standard academic career, even the country.

In between 1955 and 1960 Foucault worked as a cultural diplomat, first in Uppsala (at the invitation of Georges Dumézil), then in Warsaw and Hamburg. He thus experienced personally the different economic and political regimes of the period, noticing their latent similarities, and also worked on his doctoral thesis. This was defended in 1961, under the supervision of Georges Canguilhem, and published as *Folie et déraison* (*Madness and Unreason*) with the support of Philippe Ariès.

After completing the book, Foucault continued his work on the reconstruction of the joint emergence of modern forms of scientific (especially medical) knowledge and its institutional context, and published in 1963 *The Birth of the Clinic*. At the same time, he also got increasingly involved in literary criticism, publishing various essays and a book on Raymond Roussel. As a result, he came to 'bracket' institutional practices temporarily in his work and turned towards language and discourse—a shift he would later much regret. The result was *The Order of Things*. Published in 1966, at the height of structuralism, the book immediately became a huge success, identifying Foucault as a main protagonist of the field.

Becoming increasingly uneasy about this label, Foucault soon embarked on a major self-reflexive exercise, attempting to clarify his project and its underlying methods for others and for himself. This eventually led to the *Archaeology of Knowledge* and a series of major methodological papers. At the same time, he moved from Paris to Tunis, and this is where he lived during the events of May 1968. However, instead of a systematic and final clarification, the reflexive exercises rather brought about a major reorganization of Foucault's entire work. Appointed as Chair of the History of Systems of Thought at the prestigious *Collège de France*, a few months after his inaugural lecture, Foucault suddenly shifted his intellectual interest from knowledge to power, and at the same time he became engaged in political struggles as the founder of GIP (group for information on prisons).

His new interests led to the publication of *Discipline and Punish*, easily his most widely read and influential work. However, around the time of its publication, Foucault was again already reflecting on his work. This yielded a book-length study published as the first of six planned volumes on the *History of Sexuality* and a series of significant and quite influential interviews and shorter pieces (see Foucault 1980). But perhaps most importantly, it led to another major reorientation of his project. When trying to specify the main characteristics of his approach to the study of power, he eventually came to realize that the question of power was not central to his work. This change—often interpreted as Foucault's 'crisis'—is visible in successive displacements of the time horizon of the work, first to the fifteenth–sixteenth centuries, then to early Christianity, and finally to Greco-Roman Antiquity; and also in a reorganization at the level of central concepts: from power-knowledge to government by truth; from the genealogy of power to the hermeneutic of the subject; from technologies of power to techniques of self. It was strongly influenced by Foucault's encounter with the works of major scholars of antiquity such as Pierre Hadot and Peter Brown. At the same time he also increasingly recognized affinities with the works of social scientists like Norbert Elias and especially Max Weber. These encounters, together with a systematic reflection on the trajectory of his past work, helped Foucault to identify the underlying problem of his work as being concerned neither with power, nor with discourse, rather with the link between truth and subjectivity (Foucault 1997, p. 281).

The new orientation is visible in the published Volumes Two and Three of the *History of Sexuality* project that discuss the way practices related to pleasure were problematized in ancient Greece and Rome, respectively. Volume Four, covering early Christianity, only exists in the form of an uncorrected manuscript, as Foucault was hospitalized, terminally ill, just days after he started to revise it for publication.

The best guides for the last, arguably most significant period of Foucault's work are the courses he gave at the Collège de France in between 1978 and 1984, especially the last two courses on *parrhesia* or frankness of speech and truth-telling. Foucault died in Paris on June 25, 1984.

2. *The Work and its Context*

Throughout his writings, covering over three decades, Foucault's work focused around the reasons and manner in which the individual subject, especially its 'true being,' has become the center of attention in the pursuit of knowledge and the search for social order, considered as an evident starting point and foundation. Thus proceeding, Foucault posed question marks to what has been taken for granted by the main available approaches in philosophy and the social sciences, without dismissing them as mere ideology or illusion. This explains much of both the dynamics of Foucault's intellectual trajectory (discussed in this section) and its reception (covered in the next).

The trajectory has three major peculiarities. First of all, far from following a straightforward path of development, it was quite bumpy and curvy, progressing through crises (see Szakolczai 1998). This was probably unavoidable as Foucault's work did not simply represent an addition to the existing stock of knowledge, neither a general attack on it, rather a problematization of the taken for granted. Furthermore, and for much the same reasons, the trajectory cannot be separated from the substance of the ideas. Given its detours, it would seem particularly desirable to get over such details and focus on the main results. However, this would undermine the effectivity of Foucault's undertaking, as doing so one would fail to problematize one's own hidden related preconceptions. Finally, Foucault was quite aware of the pitfalls of his project, and throughout his career kept reflecting on its directions and stakes. These reflections had a significant effect on the trajectory and contain a concise introduction to Foucault by Foucault, a starting point for any interpretative effort.

For these reasons the exact mapping of Foucault's early steps is of crucial relevance, as this not simply defines the starting point of the investigation that would later be assumed as given, rather indicates the way in which Foucault plunged into the questioning and mapping of the aspect of the taken for granted he singled out for attention as problematic. The first anchorage point to be questioned was academic philosophy that assumed, from Descartes through Kant to Husserl and Merleau-Ponty that the 'subject of knowledge' or the 'transcendental ego' was the starting point of any philosophical undertaking. Foucault recognized that behind the epistemological assumptions of this approach there lay an entire anthropology, and was particularly interested in Kant's related work, which he translated and introduced as his minor thesis.

Given his problems with academic philosophy, he turned to psychology, or an empirical approach to the Kantian question 'what is man.' At the same time, he experimented with the two main alternative theoretical frameworks to academic philosophy, Marxism and psychoanalysis, finding them unsatisfactory. It is in this combination that the profound significance and originality of Foucault's work lies. His background incorporates both theoretical and empirical work: the philosophy of the subject in its various kinds and positivism on the one hand, and their main critiques, the works of Marx and Freud on the other, while being orthogonal to this entire epistemological field.

In his attempts to go beyond the available approaches, Foucault found inspiration in the works of Nietzsche. Instead of starting from the experiences *of* the subject, gained through a combination of sense perception and the application of 'universally' valid categories of mind, Foucault rather posed the question of how this subject has been actually historically formed *through* experience.

In his last writings, Foucault would argue that there are three main axes of experience: knowledge, power, and subjectivity. This conceptualization, however, was only the fruit of late reflection on his own work. In his first major book, *Folie et déraison*, all three axes were dealt with, and their separation was not at all clear-cut. Furthermore, instead of directly reconstructing the subject of knowledge, he rather started at the opposite end, as if following a reflexive strategy, by reconstructing the opposite or the mirror, the subject of nonknowledge or the 'nonsubject,' the mad, the carrier of 'unreason.' The aim was to show that 'mental illness' was not simply an objective aspect of the human condition that was there, waiting to be discovered finally by the methods of true science, but as a complex form of behavior was itself shaped by these very methods and their institutional context.

In this book and the follow-up work, *The Birth of the Clinic*, Foucault tied closely together the reconstruction of the views concerning the aim and methods of collecting knowledge about individual human beings and its context of closed or total institutions, or the knowledge and the power axis. This changed with his third main book, *The Order of Things*, where the institutional dimension or actual scientific and social practice was left out of the picture. On the one hand, this enabled Foucault to sharpen his tools here and in the following, reflexive works, defining his archaeological method as 'an attempt to introduce the diversity of systems and the play of discontinuities into the history of discourses' (Foucault 1991). On the other hand, however, this also meant a certain one-sidedness, the suspension of the study of the link between displacements in thought and broader historical processes.

As a result, in the 1970s, Foucault reversed his approach and turned to the power axis, developing conceptual tools like power-knowledge (which never meant a simple identification between knowledge and power, only the assertion of the impossibility of a complete separation) or the microphysics of power. This meant a shift of disciplinary allegiances from philosophy and the life sciences to sociology and the social sciences, and a reorientation of focus from the 'subject of knowledge' of phenomenology to the liberal individual as a subject of rights, the taken for granted starting point of modern political and social theory. Here as before, Foucault chose an indirect approach, focusing his reconstructive historical work on the formation of the subject deprived of rights: the criminal.

Discipline and Punish not only returned Foucault's interest to institutional practices but sharpened his analytical tools concerning the power axis. However, if his work in the 1960s was slightly affected by flirtations with literary criticism and structuralism, his new work showed too much affinity with the contemporary leftist, neo-Marxist vogue, most visibly at the level of rhetorics.

Foucault's next series of works on sexuality were intended to continue and complement the studies of power-knowledge as started in *Discipline and Punish*. However, when finishing the first, reflexive and methodological volume of the series, *La volonté de savoir*, Foucault again stepped back from his approach and undertook another major reorientation. This meant that while previously, whether focusing on knowledge or institutional practice, Foucault studied how certain human beings shaped other human beings as objects of power and knowledge, now he set out to analyze directly how individual human beings where driven to reflect on their own being as subjects.

This was accompanied by a marked change in style, even a refocusing of his entire project. While Foucault's aim so far was to render visible the singularity of the idea that the individual subject should be the foundation of knowledge and social order, now he recognized the seriousness of the problem to which this came to be the answer. The shifting of the focus of interest to the individual was not simply an epistemological mutation, nor a particularly cunning turn in the exercise of power, but was an answer to the manner in which being has become problematized at a certain point; the way in which human beings came to pose question marks to 'what they are, what they do, and the world in which they live' (Foucault 1984a). While this turn on the one hand rendered visible the Heideggerian inspiration behind his work, on the other it also brought to the surface the close affinities between Foucault and Max Weber.

3. *The Foucault 'Effect'*

Foucault has been one of the most influential social theorists of the past decades. His works are widely read in all branches of the humanities and the social

sciences, including philosophy, sociology, history, anthropology, cultural studies, and literary criticism. However, they have always been controversial and most probably will remain so in the future.

Both characteristics can be traced to the same source. Foucault's work developed not simply as a dissent from the then ruling paradigms in philosophy and psychology, phenomenology and positivism, but was similarly critical, after some initial flirtation, of the undertakings of Marx and Freud. Based on his reading of Kierkegaard, Heidegger, and especially Nietzsche, Foucault went into directions that did not easily fit on the contemporary intellectual landscape. This led to a persistent problem of understanding, reception, classification, and (mis)labeling. It was only deepened by some strategic concessions Foucault made when trying to establish bridges between his lonely undertaking and the work of his contemporaries.

Foucault's work first gained wide currency in the mid-1960s when, after the publication of *The Order of Things*, he was considered as a prime exponent of structuralism, together with Barthès, Lacan, and Lévi-Strauss. His archaeological method, aiming at the establishment of discontinuities at the level of discourses and considering texts written in between two breakpoints as belonging to a discursive formation, governed by certain autonomous rules, is still influential in wide circles within philosophy and literary criticism. The work largely followed the footsteps of Foucault's earlier works, relying more on Dumézil, Canguilhem, and Bachelard than on Lacan or Lévi-Strauss. It was also different from the structuralist project and mentality, as it made no attempt at developing formal schemes with claims of general validity. Still, the label 'structuralist' stuck due to the gestures Foucault made in its last chapter to the then rising structuralist tide.

Though Foucault's work of the 1970s on power-knowledge came to be inserted in a different debate and exerted its impact on a different field, the modalities were remarkably similar. As a conscious break with his earlier work, *Discipline and Punish* aimed at direct political relevance and impact. While *The Order of Things* was critical of Marxism and took up no side in the political landscape, the new work was clearly aiming at making contact with the far left and started by using a language heavily indebted to Marxism. As the work progressed, Foucault steadily retrieved his central interest in the formation of the self through 'true' discourse, in this case the shaping of the criminal. However, in its style and rhetoric, the ballast was still present, and the work came to be read and regarded as a prime example of the 'poststructuralist' and 'postmodern' alternative to old-fashioned Marxism. While in the 1960s Sartre considered Foucault's work as the 'last rampart' of the bourgeoisie, his work of the 1970s served as a lifebuoy for many on the sinking boat of the radical left.

Soon Foucault became just as deeply disturbed by this reception of his work as with the structuralist label in the 1960s. His self-reflexive exercises of the late 1970s and the early 1980s not only helped him to regain focus in his own work but also to find at the end, after some flirtation with the thought of the Frankfurt school, the most fruitful intellectual connections with the works of philosophers such as Gadamer and Hadot, and social scientists like Elias and Goffman. Most important of all, however, was his late in-depth encounter with the work of Max Weber, in whom he recognized a kindred spirit and a selected predecessor.

Though this was not lost in some segments of the Foucault reception (Dreyfus and Rabinow 1982, Gordon 1991, Owen 1994), by then the damage had already been done. Foucault thus came to be regarded as a prime exponent of the very intellectual currents whose presuppositions he wanted to question and map in his work. In this, his fate can again be best compared to Max Weber's. As a result, in the burgeoning literature for or against Foucault, the problem he started to pose almost half a century ago concerning the taken for granted, allegedly universal but quite peculiar and questionable character of the way modern Western subjects form their subjectivity (selfhood or identity)—a problem very close to the Weberian interest in inner-worldly asceticism and in the methodical conduct of life, or the interest of Elias in the formation of *habitus* through the civilizing process—became lost. This is demonstrated most paradoxically in the way the new literature on sexual, gender, race, etc., identities takes Foucault as one of its sources, although his work should rather be seen as a most effective way in which this 'politics of identity' could be criticized.

While Foucault, just like Weber, is one of the most widely read social theorists of the century, his work is similarly still more of an unfulfilled promise than a closed, exhausted chapter in the history of thought.

See also: Control: Social; Crime: Sociological Aspects; Crime, Sociology of; Gender Ideology: Cross-cultural Aspects; Heterosexism and Homophobia; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Medicine, History of; Postmodernism in Geography; Postmodernism in Sociology; Postmodernism: Philosophical Aspects; Psychiatry, History of; Rationality and Feminist Thought; Transsexuality, Transvestism, and Transgender

Bibliography

Dreyfus H L, Rabinow P 1982 *Michel Foucault: Beyond Structuralism and Hermeneutics*. University of Chicago Press, Chicago

Eribon D 1991 *Michel Foucault*. Harvard University Press, Cambridge, MA

Foucault M 1961 *Folie et déraison: Histoire de la folie à l'âge classique*. Plon, Paris [1965 *Madness and Civilization*. Harper, New York] {partial version}
Foucault M 1963 *Naissance de la clinique: Une archéologie du regard médical*. Presses Universitaires de France, Paris [1975 *The Birth of the Clinic*. Vintage, New York]
Foucault M 1966 *Les mots et les choses: Une archéologie des sciences humaines*. Gallimard, Paris [1973 *The Order of Things*. Vintage, New York]
Foucault M 1969 *L'archéologie du savoir*. Gallimard, Paris [1976 *The Archaeology of Knowledge*. Harper, New York]
Foucault M 1975 *Surveiller et punir: Naissance de la prison*. Gallimard, Paris [1979 *Discipline and Punish*. Vintage, New York]
Foucault M 1976 *La volonté de savoir: Histoire de la sexualité 1*. Gallimard, Paris [1980 *The History of Sexuality, Volume 1: An Introduction*. Vintage, New York]
Foucault M 1980 *Power/Knowledge: Selected Interviews and Other Writings by Michel Foucault, 1972–1977*, Gordon C (ed.). Harvester Press, Brighton, UK
Foucault M 1984a *L'usage des plaisirs: Histoire de la sexualité 2*. Gallimard, Paris [1986 *The Use of Pleasure*. Vintage, New York]
Foucault M 1984b *Le souci de soi: Histoire de la sexualité 3*. Gallimard, Paris [1987 *The Care of the Self*. Vintage, New York]
Foucault M 1984c *The Foucault Reader*. Rabinow P (ed.). Pantheon, New York
Foucault M 1991 Politics and the study of discourse. In: Burchell G, Gordon C, Miller P (eds.) *The Foucault Effect: Studies in Governmentality*. Harvester Wheatsheaf, London
Foucault M 1994 *Dits et écrits*. Defert D, Eward F (eds.). Gallimard, Paris, 4 Vols
Foucault M 1997 The ethics of the concern of the self as a practice of freedom. In: Rabinow P (ed.) *Ethics: Subjectivity and Truth, The Essential Works of Michel Foucault*. The New Press, New York, Vol. 1
Gordon C 1991 Introduction. In: Burchell G, Gordon C, Miller P (eds.) *The Foucault Effect: Studies in Governmentality*. Harvester Wheatsheaf, London
Owen D 1994 *Maturity and Modernity: Nietzsche, Weber, Foucault and the Ambivalence of Reason*. Routledge, London
Szakolczai A 1998 *Max Weber and Michel Foucault: Parallel Life-Works*. Routledge, London

A. Szakolczai

Fourier Analysis

1. Fourier Series

1.1 Trigonometric Functions

The term Fourier analysis originally refers to the possibility to approximate functions (up to scaling) in the open unit interval $I :=]-1/2, +1/2[$ by linear combinations of trigonometric functions

$$t_0(x) := 1, t_{2n}(x) := \sqrt{2}\cos(2\pi nx),$$
$$t_{2n-1}(x) := \sqrt{2}\sin(2\pi nx) \quad \text{for} \quad n = 1, 2, 3, \ldots \quad (1)$$

and commemorates the role that the French baron Jean-Baptiste-Joseph Fourier played in furthering the development of this method. (Here, the colon marks that side of an equation which is defined by the other side; thus, the symbol '$:=$' could be read as 'defined as.') Fourier's influential work has its precursors in early studies of the vibrating string by the Swiss mathematicians Daniel Bernoulli and Leonhard Euler. The formulas for what are nowadays called Fourier coefficients date back as early as 1728 and are due to Bernoulli. These formulas are best understood by invoking the geometric metaphors of length and orthogonality, which can be described in terms of a so-called inner product $\langle \cdot | \cdot \rangle$, in our current context specifically given by

$$\langle f|g\rangle := \int f(x)^* g(x)\,\mathrm{d}x \quad (2)$$

for f, g in the set $BC(I)$ of bounded, continuous, complex-valued functions on I (here z^* denotes the complex conjugate of the complex number z; for real numbers r we have $r^* = r$). As a matter of convenience and simplicity of notation we may consider the integral in Eqn. (2) to extend over the whole real line $\mathbb{R}$ by considering f, g as well as the functions t_k, $k = 0, 1, 2, \ldots$, as defined on $\mathbb{R}$ but vanishing outside the interval I.

The inner product induces the concept of a distance between two functions $f, g \in BC(I)$ defined by $\|f-g\| := \sqrt{\langle f-g|f-g\rangle}$. Here, '$\ldots \in \ldots$' should be read as '… element(s) of …' In particular, the distance of such a function f to the constant function 0 (with value zero for all arguments) $\|f\| = \|f-0\| = \sqrt{\langle f|f\rangle}$ is called the norm (or length) of f. Functions f with norm $\|f\| = 1$ are called normalized. Two functions f and g are called orthogonal when their inner product vanishes, that is, when $\langle f|g\rangle = 0$. In fact, the specific functions of the family $(t_n)_{n\in\mathbb{N}}$, where $\mathbb{N} := \{0, 1, 2, \ldots\}$ denotes the set of natural numbers, are all normalized and pairwise orthogonal if we interpret t_n to be zero outside of the unit interval I. A family of functions such as $(t_n)_{n\in\mathbb{N}}$ is, therefore, also called an orthonormal system. The interest in such orthonormal systems is based on the simplicity with which 'best approximations' to given elements can be found in terms of linear combinations of elements of such orthonormal systems. Indeed, let $f \in BC(I)$ be given, then

$$\|f - \sum_{k=0}^{N} \langle t_k|f\rangle t_k\| \quad (3)$$

is the smallest possible distance between f and any linear combination of the t_k for $k = 0, 1, 2, \ldots, N$. The so-called Fourier coefficients $\langle t_k|f\rangle$, $k = 0, \ldots, N$, are obviously independent of $N \in \mathbb{N}$ as N grows. It should be emphasized that this feature is indeed shared by any orthonormal system. The term Fourier coefficients is also used in the general situation. For sake of brevity,

however, we will have to focus here and in the following predominantly on the issue of Fourier analysis in a narrower sense.

1.2 Fourier Series

Letting $N \to \infty$ one finds that the quantity in Eqn. (3) tends indeed to zero. This property is known as the completeness of the orthonormal system $(t_n)_{n\in\mathbb{N}}$. According to the Fischer–Riesz theorem the convergence of $\sum_{k=0}^{\infty} c_k t_k$ (with respect to the norm $\|\cdot\|$) is determined completely by the convergence of the numerial series $\sum_{k=0}^{\infty}|c_k|^2$. It turns out that there are more functions than just functions in $BC(I)$ which can be approximated by Fourier series, that is, by series of the form $\sum_{k=0}^{\infty} c_k t_k$ (assuming only that the Fourier coefficients are such that $\sum_{k=0}^{\infty}|c_k|^2$ exists). The set of all such functions is denoted by $L_2(I)$, which is the set of measurable, square-integrable functions. Identifying functions $f, g \in L_2(I)$ if $\|f-g\| = 0$, we obtain a 1-to-1 and onto correspondence between elements of $L_2(I)$ and complex number sequences in $\ell_2(\mathbb{N}) := \left\{(c_k)_{k\in\mathbb{N}} \middle| \sum_{j=0}^{\infty}|c_j|^2 \text{ convergent}\right\}$. Indeed, one finds

$$\|f\| = \left(\sum_{j=0}^{\infty}|\langle t_j|f\rangle|^2\right)^{1/2} \tag{4}$$

for all $f \in L_2(I)$ and

$$f = \sum_{j=0}^{\infty}\langle t_j|f\rangle t_j \tag{5}$$

Thus, the mapping $\mathscr{T}$—defined as the association of $f \in L_2(I)$ with the sequence $(\langle t_k|f\rangle)_{k\in\mathbb{N}}$—is norm-preserving if we define

$$\|(c_k)_{k\in\mathbb{N}}\|_{\#} := \left(\sum_{j=0}^{\infty}|c_j|^2\right)^{1/2} \tag{6}$$

for $(c_k)_{k\in\mathbb{N}} \in \ell_2(\mathbb{N})$ as norm in $\ell_2(\mathbb{N})$, that is

$$\|\mathscr{T}f\|_{\#} = \|f\| \tag{7}$$

This means in particular that all the information of f is preserved just in the sequence of the Fourier coefficients.

The advantage of using the functions in Eqn. (1) as an orthonormal system is that real-valued functions result in real Fourier coefficients. However, using the equivalent complex orthonormal system $(\eta_k)_{k\in\mathbb{Z}}$

$$\begin{aligned}\eta_k(x) &:= \exp(2\pi \mathrm{i} kx)\\ &= \cos(2\pi kx) + \mathrm{i}\sin(2\pi kx) \quad \text{for } x \in I\end{aligned} \tag{8}$$

for $k \in \mathbb{Z} := \{\ldots, -3, -2, -1, 0, +1, +2, +3, \ldots\}$, yields a simpler, more unified description $(\eta_k)_{k\in\mathbb{Z}} = (\exp(2\pi \mathrm{i} k\cdot))_{k\in\mathbb{Z}}$. The relation between the new Fourier coefficient $(\langle\exp(2\pi \mathrm{i} j\cdot)|f\rangle)_{j\in\mathbb{Z}}$ and the previous ones is (due to Eqn. (8)) easily established. Similarly, as before we obtain

$$f = \sum_{j=-\infty}^{\infty}\langle\exp(2\pi \mathrm{i} j\cdot)|f\rangle \exp(2\pi \mathrm{i} j\cdot) \text{ in } L_2(I) \tag{9}$$

and the mapping $\mathscr{F}_{\#}$ associating the sequence of Fourier coefficients

$$\begin{aligned}&(\langle\exp(2\pi \mathrm{i} k\cdot)|f\rangle)_{k\in\mathbb{Z}} \in \ell_2(\mathbb{Z})\\ &\qquad := \left\{(c_k)_{k\in\mathbb{Z}} \middle| \sum_{j=-\infty}^{\infty}|c_j|^2 \text{ convergent}\right\}\end{aligned} \tag{10}$$

with the element $f \in L_2(I)$ is also norm-preserving, that is

$$\|\mathscr{F}_{\#}f\|_{\#} = \|f\| \tag{11}$$

2. The Fourier Transform

The Fourier coefficients $(\langle\exp(2\pi \mathrm{i} j\cdot)|f\rangle)_{j\in\mathbb{Z}}$ suggest the following definition

$$\hat{f}(p) := \langle\exp(2\pi \mathrm{i} p\cdot)|f\rangle \quad \text{for} \quad p \in \mathbb{R} \tag{12}$$

where f should now vanish outside a bounded interval (rather than just the unit interval I). The mapping $\mathscr{F}$ that associates with f the function $\hat{f}$ is called the Fourier transform (see Gaskill 1978, Goodman 1968 as general references). The function $\hat{f} = \mathscr{F}f$ is referred to as the Fourier transform (or the spectrum) of f. In view of the Fourier coefficients, we note that any such function can be forced into $L_2(I)$ by rescaling with a suitably large scaling factor $L \in \mathbb{R}: \sigma_L f := \sqrt{L} f(L\cdot)$. The Fourier transform interacts with rescaling in the following way

$$\begin{aligned}\mathscr{F}(\sigma_L f) &= \sqrt{|L|}\mathscr{F}(f(L\cdot))\\ &= \frac{1}{\sqrt{|L|}}(\mathscr{F}f)(\cdot/L)\\ &= \sigma_{1/L}(\mathscr{F}f) \quad \text{for } L \in \mathbb{R},\ L \neq 0\end{aligned} \tag{13}$$

Any such function may also be considered as an element of $L_2(\mathbb{R})$ (the set of measurable, square-integrable functions on $\mathbb{R}$).

Using rescaling with -1, that is, reflection at the origin, we get $\mathscr{F}^* := \sigma_{-1}\mathscr{F}$ as the inverse Fourier transform and so we are led to the so-called Fourier inversion formula

$$f = \int \exp(2\pi \mathrm{i} p\cdot)\hat{f}(p)\,\mathrm{d}p =: \mathscr{F}^*\hat{f} \tag{14}$$

From Eqn. (14) it can be concluded that $\mathscr{F}$ is norm-preserving (and so also preserves the inner product $\langle\cdot|\cdot\rangle$)

$$\|\mathscr{F}f\| = \|f\| \tag{15}$$

which in turn allows the extension of the Fourier transform (as well as Eqns. (14), (15), (13)) by continuity to all $f \in L_2(\mathbb{R})$ (although then the original integral Eqns. (13) and (14) do not in general make sense any more). Analogously to Eqn. (15), the inverse Fourier transform $\mathscr{F}^*$ also preserves the inner product. We note the following important relation between $\mathscr{F}$ and its inverse $\mathscr{F}^{-1} := \mathscr{F}^*$

$$\langle \mathscr{F}f|g\rangle = \langle f|\mathscr{F}^*g\rangle \tag{16}$$

for all $f, g \in L_2(\mathbb{R})$. Equation (15) shows again that by transforming with $\mathscr{F}$ no information is lost. This information is, however, encoded in such a form that local information in $\hat{f}$ represents global information in the original f. This feature is utilized (actually for the two-dimensional case; see the brief discussion in Sect. 5) in image processing where local information is removed from or added to the Fourier transform $\hat{f}$ of an image f (modeled as an illumination distribution function f) to actually improve the visual appearance of the original f by, for example, removing grid structures or sharpening edges. The visual system appears to perform image processing, at least approximately, in this manner (see Cornsweet et al. 1984, Georgeson 1979, Graham 1981). Another application of the particular properties of Fourier transform encoding can be found in holography (see Stroke 1969). Holographic encoding has been utilized as a model for particular features in memory and perception (Pribram et al. 1974). The resistance of the original signal f against 'noise' afflicting the Fourier transformed signal $\hat{f}$ is also utilized in data transmission security.

In concrete applications one has to deal with real quantities (the complex notation just being a matter of formal convenience). In this context the so-called power spectrum $|\hat{f}(\cdot)|^2$ is of particular interest (the inverse Fourier transform of the power spectrum is known as the autocorrelation function). Also note that if f is even, that is, $f(-x) = f(x)$, then $(\mathscr{F}_{\cos} f)(x) := \langle \cos(2\pi x\cdot)|f\rangle = (\frac{1}{2}(\mathscr{F} + \mathscr{F}^*)(f))(x)$ (the Fourier cosine transform) and if f is odd, that is, $f(-x) = -f(x)$, then $(\mathscr{F}_{\sin} f)(x) := \langle \sin(2\pi x\cdot)|f\rangle = (\frac{1}{2i}(\mathscr{F} - \mathscr{F}^*)(f))(x)$ (the Fourier sine transform) all of which are real quantities. The Fourier cosine transform would recover the even part $E(f)(x) := \frac{f(x)+f(-x)}{2}$ of f from $\hat{f}$ and the Fourier sine transform its odd part $O(f)(x) := \frac{f(x)-f(-x)}{2}$. Readers who are uncomfortable with complex numbers may restrict their attention to odd and even functions separately (the less mathematically inclined reader may also find in Weisstein 1980 an interesting and entertaining survey on Fourier transforms). Their information is then fully preserved by the (real) Fourier cosine and the (real) Fourier sine transform, respectively.

To give at least a flavor of how Fourier transforms act, we give two interesting examples (for a visual approach to Fourier transforms see Harburn et al. 1975). The Gauss distribution function γ—well-known from statistics—is given by $\gamma(x) := \exp(\pi x^2)$ for $x \in \mathbb{R}$. It has the remarkable property that

$$\mathscr{F}_{\cos}\gamma = \mathscr{F}\gamma = \gamma \tag{17}$$

Let χ_M denote the characteristic function of a set M (given by

$$\chi_M(x) := \begin{cases} 1 & \text{for } x \in M \\ 0 & \text{for } x \notin M \end{cases} \tag{18}$$

for $x \in \mathbb{R}$). The characteristic function $\Pi := \chi_I$ of the above unit interval I has the so-called sinc function as Fourier transform

$$\mathscr{F}_{\cos}\Pi = \mathscr{F}\Pi = \text{sinc} \tag{19}$$

where

$$\text{sinc}(x) := \begin{cases} \frac{\sin(\pi x)}{\pi x} & \text{for } x \neq 0 \\ 1 & \text{for } x = 0 \end{cases} \tag{20}$$

for all real x. The sinc function is important in connection with the Shannon sampling theorem.

A discretization of Eqn. (14) yields the finite sum

$$\exp(-2\pi i(N/(2L))\cdot)\sum_{j=0}^{N-1}\hat{f}((j-N/2)/L)\times \exp(2\pi i(j/L)\cdot)\frac{1}{L} \tag{21}$$

as an approximation for f as N and L become large. Representing f by the samples of

$$f((\cdot - N/2)/L) \tag{22}$$

at the points $j = 0, 1, 2, \ldots, N-1, N$ and letting $L = \sqrt{N}$ we may consider Eqn. (21) as a product of a column vector by the matrix $\frac{1}{\sqrt{N}}[\exp(2\pi i(jk/N)]_{j,k=0,1,\ldots,N-1}$ approximating f. This matrix is unitary and known as the inverse discrete Fourier transform (IDFT). Consequently, its inverse, the unitary matrix

$$\frac{1}{\sqrt{N}}[\exp(-2\pi i jk/N)]_{j,k=0,1,\ldots,N-1} \tag{23}$$

is called the discrete Fourier transform (DFT). For practical purposes it is worth noting that the number

of numerical operations required to calculate the discrete Fourier transform (or its inverse) can be dramatically reduced by a factorization technique known as the fast Fourier transform (FFT).

The connection to the Fourier series transform $\mathscr{F}_{\#}$ is apparent. However, since the transform $\mathscr{F}_{\#}$ can be recovered from the Fourier transform $\mathscr{F}$ (see Eqn. (45) below), we will here and from now on focus on the latter.

3. The Fourier Transform as a Spectral Representation

3.1 Derivatives

The significance of representing a function in terms of its Fourier transform becomes apparent if we consider how that transform interacts with differentiation ∂. Introducing for convenience the symbol $D := \frac{1}{2\pi i}\partial$ we find that

$$D = \mathscr{F}^{-1}\mathbf{m}\mathscr{F} \text{ in } L_2(\mathbb{R}) \tag{24}$$

where $\mathbf{m}$ stands for the multiplication-by-argument operation, that is

$$(\mathbf{m}f)(p) := pf(p) \quad \text{for all } f \in L_2(\mathbb{R}),\ p \in \mathbb{R} \tag{25}$$

It should be noted that being in the domain of the differential operator D (as naturally characterized by the well-definedness of the right-hand side of Eqn. (24)) implies a generalized concept of differentiability. The elements in the domain turn out to be even in $BC(\mathbb{R})$. Equation (24)—it is called a spectral representation of D—shows the equivalence of differentiation $\frac{1}{2\pi i}\partial$ with multiplication-by-argument $\mathbf{m}$. The obvious symmetry of $\mathbf{m}$, that is

$$\langle \mathbf{m}f|g\rangle = \langle f|\mathbf{m}g\rangle \tag{26}$$

implies the symmetry of D, that is

$$\langle Df|g\rangle = \langle f|Dg\rangle \text{ in } L_2(\mathbb{R}) \tag{27}$$

3.2 Functions of D

The Fourier transform spectral representation in Eqn. (24) gives rise to an extension to polynomials in D and $\mathbf{m}$, respectively. Equation (24) thus suggests to generally define functions of D by letting

$$S(D) := \mathscr{F}^{-1}S(\mathbf{m})\mathscr{F} \text{ in } L_2(\mathbb{R}) \tag{28}$$

where S can be fairly general (and even matrix-valued, in which case the Fourier transform is to be carried out component-wise). This concept is extremely powerful, since a large class of models discussed in so-called linear systems theory (see Gaskill 1978, Norman 1981) can be understood as such functions $S(D)$ acting on an input u resulting in an output $S(D)u$. Characterizing properties of such systems are that their input–output relations are linear and translation invariant, that is, sums, multiples, and shifts of input result in sums, in (up to scaling) the same multiple, and equally shifted output. Indeed, translation τ_h defined by its action on functions f as a shift of the argument

$$\tau_h f := f(\cdot + h) \tag{29}$$

is itself a function of D in the above sense

$$\tau_h = \exp(2\pi i h D) = \mathscr{F}^{-1}\exp(2\pi i h\mathbf{m})\mathscr{F} \tag{30}$$

It may be interesting to note that by substituting the series expansion for the exponential function, this formula (if applied to suitable functions) is recognized as nothing but the Taylor expansion. Such translation-invariant, linear systems occur (usually in their higher-dimensional form, compare Sect. 5) as good approximations of many aspects of visual and auditory perception.

3.3 Convolution

The action of a function $\hat{S}(D)$, where $\hat{S}$ is the Fourier transform of S defines what is called convolution (of S with g)

$$S * g := \hat{S}(D)g \tag{31}$$

which is defined for any g in the domain of $\hat{S}(D)$. In many cases, $S * g$ is actually given as an integral expression, the convolution integral

$$(S * g)(x) = \int S(x-y)g(y)\,\mathrm{d}y \tag{32}$$

In accordance with Eqn. (31) for such convolution integrals the so-called convolution theorem holds

$$\mathscr{F}(S * g) = \hat{S}\hat{g} = \hat{S}(\mathbf{m})\hat{g} \tag{33}$$

Particularly, the above-mentioned relationship between the autocorrelation and the power spectrum reveals that the autocorrelation is a convolution.

4. Generalized Fourier Transform and Some Applications

Generalizing the meaning of $\langle\cdot|\cdot\rangle$ we introduce generalized functions f (also called distributions) as linear functionals by *prescribing* the value of

$$f(\varphi) := \langle f|\varphi\rangle \tag{34}$$

for any smooth φ vanishing outside a bounded interval in $\mathbb{R}$—we shall write $\varphi \in C_0^\infty(\mathbb{R})$—such that linearity holds, that is

$$\langle f|\varphi + \alpha\psi\rangle = \langle f|\varphi\rangle + \alpha\langle f|\psi\rangle \tag{35}$$

for all complex numbers α and all φ, $\psi \in C_0^\infty(\mathbb{R})$. Clearly, any $f \in L_2(\mathbb{R})$ is also such a generalized function, but also $\exp(2\pi i p \cdot)$, although not an element of $L_2(\mathbb{R})$, is such a generalized function, as Eqn. (12) of the Fourier transform shows. A generalized function, which is indeed *not* given by a function, is featured by the distribution $\delta_{\{t\}}$ defined by

$$\delta_{\{t\}}(\varphi) = \langle \delta_{\{t\}} | \varphi \rangle := \varphi(t) \quad \text{for all } \varphi \in C_0^\infty(\mathbb{R}) \tag{36}$$

(this distribution apparently takes a sample of the 'test function' φ at point $t \in \mathbb{R}$!). The special choice $\delta := \delta_{\{0\}}$ is the so-called Dirac δ-distribution. Since it can only be distinguished from 0 by testing with $\varphi \in C_0^\infty(\mathbb{R})$ satisfying $\varphi(0) \neq 0$ we say $\delta = 0$ on $\mathbb{R} \setminus \{0\}$. However, δ is (by definition) not identically zero as a functional. It is particularly useful in many contexts to have $\delta_{\{t\}}$ $(= \tau_{-t}\delta)$ as a model for something very small but nevertheless non-negligable such as a mass point, an elementary charge, a light point, a short impulse, or a molecule at location t (see (Sect. 5) for the particularly interesting higher-dimensional case). With sufficient care a large class of such distributions f (e.g., the so-called tempered distributions) can be established for which the concept of derivative, multiplication by the argument, and Fourier transform can (guided by Eqns. (27) and (16)) be generalized by defining

$$\begin{aligned} \langle Df | \varphi \rangle &:= \langle f | D\varphi \rangle \\ \langle \mathscr{F} f | \varphi \rangle &:= \langle f | \mathscr{F}^* \varphi \rangle \\ \langle \mathscr{F}^* f | \varphi \rangle &:= \langle f | \mathscr{F} \varphi \rangle \end{aligned} \tag{37}$$

for all $\varphi \in C_0^\infty(\mathbb{R})$. As an example for this generalized differentiation we find for the characteristic function $\chi_{]0,\infty[}$ of the interval $]0, \infty[$ of all positive real numbers

$$D\chi_{]0,\infty[} = \frac{1}{2\pi i}\delta \tag{38}$$

The Fourier transform of the distribution $\delta_{\{t\}}$ is according to Eqn. (37)

$$\mathscr{F}\delta_{\{t\}} = \exp(-2\pi i t \cdot) \tag{39}$$

and conversely we have

$$\mathscr{F}\exp(2\pi i t \cdot) = \delta_{\{t\}} \tag{40}$$

The latter fact can be used to detect oscillatory behavior of frequency t which may be hiding in seemingly random data. The oscillation would show up in the Fourier transformed data as a 'peak' at point t.

The concept of functions of D and in particular the concept of convolution can also be carried over to tempered distributions. This will be illustrated by a few applications.

4.1 Impulse Response

The importance of the Dirac δ-distribution is that a translation-invariant, linear system $S(D)$ is completely described by its impulse response $S(D)\delta$. Indeed, the general response $S(D)f$ is given by convolution with the impulse response

$$S(D)f = S(D)(\delta * f) = (S(D)\delta) * f \tag{41}$$

In the case that $S(D)$ is the solution operator $S(D) = P(D)^{-1}$ of the differential equation $P(D)u = f$, *P being a* polynomial, the impulse response is also known as a fundamental solution (or Green's function).

4.2 Shannon's Sampling Theorem

Another distribution of particular interest is the (sampling or) comb distribution $\text{III} := \sum_{n=-\infty}^{+\infty} \delta_{\{n\}}$ which takes samples at every integer point. This distribution is reproduced by the Fourier transform, that is

$$\mathscr{F}\text{III} = \text{III} \tag{42}$$

(which is another way of stating the Poisson summation formula). If $f \in L_2(I)$ then $\text{III} * f$ is simply the periodic extension of f to all of $\mathbb{R}$. Therefore, by 'cutting-off' with the characteristic function Π we recover

$$f = \Pi(\text{III} * f) \tag{43}$$

As we now demonstrate, the so-called Shannon sampling theorem follows from Eqn. (42). In fact, applying Eqn. (43) to $\hat{f} \in L_2(I)$ we get after the inverse Fourier transformation has been applied that

$$\begin{aligned} f &= \mathscr{F}^*(\Pi(\text{III} * \hat{f})) \\ &= \text{sinc} * \left(\sum_{n=-\infty}^{+\infty} f(n)\delta_{\{n\}} \right) \\ &= \sum_{n=-\infty}^{+\infty} f(n)\,\text{sinc}(\cdot - n) \end{aligned} \tag{44}$$

Any function f with $\hat{f}$ vanishing outside a bounded interval is called band limited. Since any such function can be easily rescaled to have $\hat{f} \in L_2(I)$, Eqn. (44) shows that any band-limited function f can be completely recovered from equidistant sampling. It is worth noting that by interchanging the role of f and $\hat{f}$ in the Shannon sampling theorem we obtain

$$\hat{f} = \mathscr{F}(\Pi(\text{III} * f)) = \sum_{n=-\infty}^{+\infty} \hat{f}(n)\,\text{sinc}(\cdot - n) \tag{45}$$

and after applying the inverse Fourier transform the Fourier series Eqn. (9) is recovered. The sampling theorem has many applications, in particular in the analysis of auditory and visual perception (see, e.g., Yellot 1990).

4.3 Wavelets

It is interesting to note that $(\tau_{-n}\mathrm{sinc})_{n\in\mathbb{Z}}$ is also an orthonormal system in $L_2(\mathbb{R})$. By virtue of $(\tau_{-n}\mathrm{sinc})_{n\in\mathbb{Z}}$ being an orthonormal system we get (as for the trigonometric series)

$$\sum_{n=-\infty}^{+\infty} \langle(\tau_{-n}\mathrm{sinc})|g\rangle\tau_{-n}\mathrm{sinc} \tag{46}$$

as the best approximation of g in these terms. According to Eqn. (44) we have that for functions with bandwidth less than or equal to 1 (i.e., $\hat{f}\in L_2(I)$) g actually is represented by this sum. To cover all band-limited functions one would have to allow for sufficiently large scaling factors L. Letting, for example, $L = 2^p$, $p\in\mathbb{Z}$, then $(\sigma_{2^p}(\tau_{-n}\mathrm{sinc}))_{p,n\in\mathbb{Z}}$ can be used to approximate any element $f\in L_2(\mathbb{R})$. All elements of $(\sigma_{2^p}\tau_{-n}\mathrm{sinc})_{p,n\in\mathbb{Z}}$ are generated by the combination $\sigma_{2^p}\tau_{-n}$ of scaling and translation from the single function sinc, the so-called scaling function of this particular construction. Unfortunately, $(\sigma_{2^p}\tau_{-n}\mathrm{sinc})_{n\in\mathbb{Z}}$ is only orthonormal for fixed $p\in\mathbb{Z}$. By a suitable orthonormalization procedure (multiresolution analysis) one can obtain, however, that $(\sigma_{2^p}\tau_{-n}(\mathrm{sinc}-\sqrt{2}\sigma_2(\mathrm{sinc})))_{p,n\in\mathbb{Z}}$ is an—even complete—orthonormal system in $L_2(\mathbb{R})$. Any orthonormal system generated by $\sigma_{2^p}\tau_{-n}, p,n\in\mathbb{Z}$, from a so-called mother wavelet—here $\mathrm{sinc}-\sqrt{2}\sigma_2(\mathrm{sinc})$—is called an orthonormal wavelet basis. Wavelets derive their utility largely from the fact that they do not have a fixed base scale like, for example, the Fourier system $(\exp(2\pi i n\cdot)\Pi)_{n\in\mathbb{Z}}$, which by allowing for translations would also yield an orthonormal system $(\tau_r(\exp(2\pi i n\cdot)\Pi))_{n,r\in\mathbb{Z}}$ for $L_2(\mathbb{R})$. Wavelets naturally allow for the 'right' scale in $L = 2^p, p\in\mathbb{Z}$, to emerge for any given f according to their weight in its orthonormal expansion. As the Fourier transform may be considered to be a continuously scaled Fourier series, there are also continuously scaled versions of wavelet expansions known as wavelet transforms.

5. Higher-dimensional Fourier Transform

Much of what has been explored above can be generalized to higher dimensions. This is due to the particular product structure of the higher-dimensional exponential

$$\begin{aligned}&\exp(2\pi i(x_1p_1+x_2p_2+\cdots+x_np_n))\\&\quad=\exp(2\pi i x_1p_1)\exp(2\pi i x_2p_2)\ldots\exp(2\pi i x_np_n)\end{aligned} \tag{47}$$

for $x_i, p_i\in\mathbb{R}$, $i = 1, 2, \ldots, n$, that is, $(x_1, x_2, \ldots, x_n)$, $(p_1, p_2, \ldots, p_n)\in\mathbb{R}^n$. Thus, the n-dimensional Fourier transform, $n\in\mathbb{N}$, $n\geqslant 2$, can be understood as a repeated one-dimensional Fourier transform, that is

$$\begin{aligned}(\mathscr{F}f)(p) &= \iint\ldots\int\exp(-2\pi i xp)f(x)\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n\\&=\int\exp(2\pi i x_1p_1)\int\exp(-2\pi i x_2p_2)\ldots\\&\quad\ldots\int\exp(-2\pi i x_np_n)f(x)\mathrm{d}x_n\ldots\mathrm{d}x_2\mathrm{d}x_1\end{aligned} \tag{48}$$

where $x = (x_1, x_2, \ldots, x_n)$, $p = (p_1, p_2, \ldots, p_n)$, and $xp = \sum_{i=1}^n x_ip_i$. As a consequence, many of the above considerations can be carried over to the higher-dimensional case. For example, with $x^2 := xx$, $x\in\mathbb{R}^n$, the definition of the Gauss distribution function carries over to the n-dimensional case and Eqn. (17) remains true. The n-dimensional Dirac δ-distribution $\delta_{\{x\}}$ simply takes a single sample at $x\in\mathbb{R}^n$ and still has Fourier transform $\exp(-2\pi i x\cdot)$. Also, $\mathrm{III} := \sum_{x\in\mathbb{Z}^n}\delta_{\{x\}}$, which takes samples at every point of $\mathbb{R}^n$ with integer coordinates, still satisfies

$$\mathscr{F}\mathrm{III} = \mathrm{III} \tag{49}$$

Thus, in particular, corresponding variants of the Shannon sampling theorem and of the Fourier series expansion hold. The scaling, however, has to be generalized to $\sigma_L f := |L|^{n/2}f(L\cdot)$, but then Eqn. (13) also carries over to the n-dimensional case. The higher-dimensional Fourier transform finds its applications in higher-dimensional, translation-invariant linear systems (Norman 1981). A particularly prominent application is in optics (see, e.g., Gaskill 1978, Goodman 1968, Taylor 1978). As pointed out earlier, the visual system may be considered at least approximately to be a two- or three-dimensional example and may be analyzed with Fourier techniques (see e.g., Cornsweet et al. 1984, Georgeson 1979, Graham 1981, Julesz 1980, Yellot 1990).

See also: Bernoulli, Jacob I (1654–1705); Computational Imagery; Pattern Matching: Methodology; Spatial Pattern, Analysis of; Spectral Analysis; Statistical Pattern Recognition

Bibliography

Cornsweet T N, Wandell B, Yellot J I 1984 The beginnings of visual perception: The retinal image and its initial encoding. In: Darian-Smith I (ed.) *Handbook of Physiology, Sect. 1. The Nervous System. Vol. 3: Sensory Processes.* American Physiological Society, Bethesda, MD pp. 257–316

Gaskill J D 1978 *Linear Systems, Fourier Transforms and Optics.* Wiley, New York

Georgeson M 1979 Spatial Fourier analysis and human vision. In: Sutherland N S (ed.) *Tutorial Essays in Psychology.* Erlbaum, Hillsdale, NJ, Vol. 2, pp. 39–88

Goodman J W 1968 *Introduction to Fourier Optics.* McGraw-Hill, San Francisco

Graham N 1981 The visual system does a crude Fourier analysis of patterns. SIAM-AMS Proceedings, 13, pp. 1–16

Harburn G, Taylor C A, Welberry T R 1975 *Atlas of Optical Transforms.* Bell & Sons, London

Julesz B 1980 Spatial-frequency channels in one-, two-, and three-dimensional visions: Variations on an auditory theme by Bekesy. In: Harris C S (ed.) *Visual Coding and Adaptibility*. Lawrence Erlbaum Associates, Hillsdale, NJ

Norman F M 1981 Lectures on linear systems theory. *Journal of Mathematical Psychology* **23**: 1–89

Pribram K, Nuwer M, Baron R 1974 The holographic hypothesis of memory structure in brain function and perception. In: Krantz D H, Atkinson R C, Luce R D, Suppes P (eds.) *Contemporary Developments in Mathematical Psychology*. W. H. Freeman, San Francisco, Vol. 2, pp. 46–57

Stroke G W 1969 *An Introduction to Coherent Optics and Holography*. Academic Press, New York

Taylor C A 1978 *Images: A Unified View of Diffraction and Image Formation with All Kinds of Radiation*. Wykenam Publications, London

Weisstein N 1980 The joy of Fourier analysis. In: Harris C S (ed.) *Visual Coding and Adaptibility*. Lawrence Erlbaum Associates, Hillsdale, NJ

Yellot J I 1990 The photoreceptor mosaic as an image sampling device. In: *Advances in Perception. Proc. Symp. on Frontiers of Visual Science. National Research Council*. National Academy Press, Washington, DC

R. Picard

Fourth World

The Fourth World is both an imagined community and a political coalition of indigenous peoples, transcending the borders of nation states. Indigenous identity is grounded in a people's claim to be descended from the original inhabitants of a country, vis-à-vis ethnic groups which came later as conquerors and colonists; it also implies a radical difference from other life ways. Thus, the Fourth World aligns indigenous peoples with the Third World (see *Third World*) in opposition to the neocolonialism of the First (capitalist) World, but opposes them to both and to the Second (communist) World, in rejecting modernity in favor of spirituality and closeness to nature. Notwithstanding their weakness in terms of conventional political resources, and their often minuscule numbers, Fourth World peoples were able to achieve substantial recognition in the second half of the twentieth century, by linking local with global mobilization.

1. *Recognizing Indigenous Peoples*

The current use of the term 'indigenous' has its origins in the European colonization of the New World, and the settlers' attempt to make their homes in places that were already inhabited. It can be dated conveniently from Columbus's landfall in the Americas in 1492, though events subsequently followed a similar course in Australasia, parts of Africa, and some Pacific islands. Settler societies commonly identified indigenous populations by the use of generic terms such as Indians, natives, aborigines, and indigenes. 'Indigenous' has become the preferred term, particularly in Fourth World discourse.

People identifying as indigenous live alongside the descendants of settlers and also recent immigrants in the former colonies of the Americas and Australasia, and most of the governments of these states recognize their presence. Indigenous groups also exist in Arctic Scandinavia and Russia, and in Japan, following the extension of central government into marginal areas, and incursions by the majority populations. These peoples survive either as internal colonies, or integrated with lower strata of the majority society. Some minorities in the recently decolonised nations of Africa and Asia have also identified themselves as indigenous, in resisting the encroachments and intrusions of a dominant population and government. There are other people who might be regarded as indigenous but who have been unable to participate in the Fourth World coalition, or who prefer some other kind of identification.

2. *Indigenous Peoples and National Governments*

Official criteria of indigenous status vary; thus, some Latin American countries recognize only those who speak an indigenous language, regarding others as assimilated; Australia accepts a combination of self identification and community recognition, while in the US, status derives from treaties made between the government and certain tribes. Taking official estimates from a selection of countries, the indigenous population exceeds 40 percent in the Andean republics and Guatemala; ranges between five percent and 20 percent in New Zealand, Chile, Mexico, Panama, and Nicaragua; between one and four percent in Australia, the United States, Canada, Colombia, Norway, Malaysia, and the Philippines. The percentages are even smaller in Brazil, Sweden, Russia, Japan, and Botswana. However, the size of an indigenous population is no guide to its political visibility on the global arena. From the metropolitan perspective, the position of indigenous peoples is more or less marginal. From their own perspective, keeping the majority at arm's length enables them to preserve an indigenous space, perhaps grounded in a distinctive mode of livelihood, such as fishing or herding or subsistence cultivation. However, the history of indigenous peoples has been characterized by encroachments on this space, and if they can no longer disengage, their lives become hostage to the schemes of the majority. In the second half of the twentieth century indigenous spaces all around the world were subjected to a new round of incursions, in such forms as logging, agribusiness, mining, commercial fishing, and tourism. Resistance to previous incursions had been localized, and often

violently suppressed, but from the 1960s, changes in international relations provided indigenous peoples with new freedoms and opportunities.

3. International Politics and Transnational Coalitions

In the context of the Cold War and the replacement of the old imperial domains by nonaligned 'new nations', colonialism and racial discrimination became issues in international politics and reproaches that most governments preferred to avoid. The new medium of television transformed domestic issues such as the African–American civil rights movement in the US and the Bantu struggle against *apartheid* in South Africa, into global dramas, on which people elsewhere could make judgments and relate to their own situations. The eruption of the Cold War in Latin America and Southeast Asia created a climate of dissent in the metropolitan countries involved, as well as others that were not directly involved, for example, in Western Europe.

From the 1970s the world became more receptive to indigenous demands, and indigenous mobilization took on an international dimension, both through coalitions of indigenous peoples across national borders, and the involvement of nonindigenous organizations. Indigenous Americans were the first to mobilize across national borders, with the formation of the International Indian Treaty Council 1974 (Ortiz 1984). In 1975 indigenous Americans formed the World Council of Indigenous Peoples (WCIP), presently opening their ranks to Aboriginal Australians, New Zealand Maori, and Norwegian Saami; others followed. Sympathetic governments facilitated the travel of indigenous spokespersons and several hosted the world conferences that have been held annually ever since. The United Nations recognized the WCIP and its participant organizations in 1977.

As early as the 1950s, the International Labor Organization had drafted a resolution on discrimination against indigenous peoples (Ortiz 1984); but it was not until 1972 that the UN, in response to representations from various nongovernmental bodies, set up the United Nations Sub-Committee on the Prevention of Discrimination and the Protection of Minorities and commissioned a study of the situation of indigenous peoples (Van Cott 1994, Pritchard 1998). In 1977, after further lobbying from nongovernment organizations, the UN convened a conference in Geneva on discrimination against indigenous populations in the Americas, in which the main participants were indigenous. In 1982, it set up the Working Group on Indigenous Populations, bringing to Geneva indigenous peoples not only from the Americas but Australasia and Europe.

Following the Columbus quincentennial, the UN proclaimed an International Year and Decade of Indigenous Peoples. Since 1995, its Commission on Human Rights has been discussing a draft declaration on the Rights of Indigenous Peoples, to be ratified by the participating governments. These developments at the international level have coincided with the emergence, in Europe as well as North America, of nongovernment organizations specifically concerned with indigenous peoples. The World Council of Churches has also maintained a continuous interest since 1972 and organizations with broader concerns, such as the environmentalists, have discovered interests in common with indigenous groups. As a result of these developments, governments must expect not only to have their own indigenous minorities subject to international scrutiny, but also to take a position on the practices of other governments.

4. Political Process and Indigenous Mobilization

Indigenous groups have finally to settle matters with the governments of the countries in which they live, but appealing to international opinion now offers a means of applying a pressure that is not otherwise available. Through the Internet, indigenous organizations are able to disseminate information instantly to sympathisers around the world, and through them to national media and political bodies; where indigenous activists are at risk of reprisals, international observers can be on the scene at short notice. In less critical circumstances, indigenous protesters can take advantage of a national festivity, when the international media are watching, to publicize their grievances through what has become the politics of embarrassment. Otherwise powerless indigenous groups have become skilled in the performance of protest, deploying the signs of their identity in colorful ethnodramas which draw the television cameras (Paine 1982).

The outcome of indigenous mobilization depends on the relative strength of the opposing interests, and—since indigenous peoples live under a variety of regimes—the forms of political process available to them. Even in democracies, a small, low status, or inaccessible minority may have difficulty in attracting the attention of voters. International pressure has induced governments to be more responsive to indigenous claims than their electorates. In some countries, legal process has proved more productive than political process and there has been a substantial development in the law of indigenous rights, with the sharing of experience and insights across national boundaries (Pritchard 1998).

5. The Politics of Indigenous Difference

It took some time for a distinctively indigenous politics to take shape. At first indigenous people might mobilize within a framework of agrarian struggle or

civil rights campaigns, with class or race as the basis for alliance. With the emergence of difference politics in the 1970s, the indigenous movement began to articulate its own special needs and aspirations. Thus, a Bolivian Aymara leader argued the need to progress from 'democratic and political reductionism, to an ethnic democracy' (Van Cott 1994).

Despite the diversity among indigenous peoples, certain issues are now regarded as characteristic of all Fourth World peoples, being woven into a transnational discourse of indigeneity. Definitive is the trope of survival, referring to a past of resistance to domination and assimilation, and a future in which indigenous peoples will regain the space, physical and cultural, in which to practice their unique kind of difference. This difference often is grounded in subsistence economic practices that use resources in ways quite different from those of the mainstream economy. Land is not only an economic resource but also the site of the group's mythology and history. Where the land has been expropriated, there is a demand for its return to the traditional owners, and the legislation of some kind of native title. Where recuperation is not a possibility, there may be a call for compensation, to acquire other lands, or for royalties that can be used to underwrite some other basis for indigenous revival.

Indigenous peoples also demand the rehabilitation of their cultural heritage, including language. However, a return to a precolonial existence is rarely an option. Revival of indigenous religion raises particular difficulties, since the groups in question may not only have been practicing Christians for generations, but be indebted to the church for local support and international linkages in its political struggles. Adherents of New Age religions may welcome the return of the old religions, but more common is a negotiation with the churches over the indigenization of clergy and forms of worship. Whether in Christian or other form, the Fourth World proclaims its spirituality in the face of the materialism of developed and developing worlds.

The demand of Fourth World peoples for the restoration of sovereignty constitutes the most direct challenge to the state, just as the claim to be 'first nations' challenges the self image of the nations which encompass them. Many indigenous peoples are too small, too scattered, or too dependent economically, for sovereignty to be realizable. A number have been able to negotiate a degree of self determination within the framework of the state, in respect of a particular territory; the Nunavut agreement in Northern Canada is the most far reaching arrangement of this kind. Peoples who live interspersed with the majority population are harder to accommodate, but the Saami Parliaments in Scandinavia and the Aboriginal and Torres Strait Islander Commission in Australia have assumed responsibility for some of the special needs of their constituents. As with land ownership, it may not be possible or appropriate to try to recreate pre-existing institutions of government, so that new forms have to be devised. As with the Aboriginal and Torres Strait Islander Commission, power is more centralized than formerly. In the US, by contrast, power continues to lie mainly with the tribes, constituted as a result of treaties drawn up in the nineteenth century.

The Fourth World can be described as an imagined community (Anderson 1991) but it is not one that aspires to nationhood; the aim of coalitions such as the WCIP is to enhance rather than assume the sovereignty of its constituent peoples and to enable them to proclaim their particular kinds of difference to their world. It has, nevertheless, provided the milieu in which a Fourth World discourse could be articulated.

6. *The Fourth World and its Constituents*

Fourth World discourse represents indigenous peoples as the custodians of a unique wisdom which, if lost, will be lost to all humankind. Cultural survival, then, is not just a matter of minority rights but of global concern. The qualities that make up indigeneity are more or less explicitly alternative to modernity including a spiritual rather than exploitative attitude towards the environment and a focus upon essential human needs over material accumulation and economic development. It is because of these shared qualities that indigenous people are able to talk to one another and imagine people like themselves, even though they have not met them.

Fourth World discourse necessarily presents the qualities of indigeneity as generalities, but this is on the understanding that their value is only realized through their embeddedness in particular local life ways. While it celebrates the special qualities that indigenous peoples share, any tendency to standardize or homogenize them runs up against the countervailing emphasis on the cultural diversity of the constituency. Consistent with this, Fourth World gatherings are celebrations of particularity, in which the various groups perform their cultures in such accessible forms as music and dance, and cuisine. Removed from their sociocultural settings, these practices cannot be apprehended in their full complexity; they are rather the signs of identity, cultural capital which is deployed to achieve a sense of equivalence among Fourth World peoples, and to display difference to others.

While Fourth World leaders speak as members of the particular peoples from which they come, participation in the transnational arena takes them away from their communities for long periods; some may be exiled from their homelands. Thus, paradoxically it is possible to speak of cosmopolitan indigenes, whose time is spent working with other indigenous peoples, and representing the Fourth World to nonindigenous people, leaving little time to keep contact with their grass roots. For the present, however, the Fourth

World is a culture of negotiation and representation; it is not an identity in itself.

7. Scholarly Studies of the Fourth World

The discipline of anthropology emerged in the wake of colonization as a study of the peoples whose descendants make up the Fourth World. The original aim of this research was to reconstruct the early history of humanity. However, from early in the twentieth century anthropologists along with ethnohistorians studied the effects of European settlement and colonization on indigenous peoples and the ways they responded to these influences (Wolf 1982).

Since the 1960s, anthropologists and ethnohistorians have also been involved in the indigenous rights movements, coming forward as experts in court cases, and advising community leaders. Anthropologists tend to work from the perspective of the local community but there have been comparisons of conditions within both North and South America (Kroeber 1994; Urban and Sherzer 1991), and between continents (Dyck 1985). Studies at the level of the nation state have been rather the province of historians, political scientists, and lawyers, some of whom have also worked comparatively, and on the relation between indigenous peoples and the UN (Ortiz 1984, Van Cott 1994, Pritchard 1998). Nongovernmental organizations, such as Cultural Survival, Survival International, and the International Work Group for Indigenous Affairs, are concerned specifically with indigenous peoples and also sponsor and publish research in these areas. What remains unresearched is the actual working of Fourth World conferences and delegations and the development of a meta-indigenous culture among those indigenous people who work at this level.

See also: Colonization and Colonialism, History of; Development, Economics of; Development: Social-anthropological Aspects; Development: Socioeconomic Aspects; Globalization: Geographical Aspects; Globalization: Political Aspects; Industrialization; Inequality

Bibliography

Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, New York

Dyck N (ed.) 1985 *Indigenous Peoples and the Nation-State: 'Fourth World' Politics in Canada, Australia and Norway*. Institute of Social and Economic Research, Memorial University of Newfoundland, St Johns, Canada

Kroeber K (ed.) 1994 *American Indian Persistence and Resurgence*. Duke University Press, Durham and London

Ortiz R D 1984 *Indians of the Americas: Human Rights and Self-Determination*. Zed Books, London

Paine R 1982 *Dam a River, Damn a People? Saami (Lapp) Livelihood and the Alta/Kautokeino Hydro-electric Project and the Norwegian Parliament*. International Work Group for Indigenous Affairs, Copenhagen, Denmark

Pritchard S (ed.) 1998 *Indigenous Peoples, the United Nations and Human Rights*. Zed Books, London

Urban G, Sherzer J (eds.) 1991 *Nation States and Indians in Latin America*, 1st edn. University of Texas Press, Austin, TX

Van Cott D L (ed.) 1994 *Indigenous Peoples and Democracy in Latin America*. St Martin's Press, New York

Wolf E R 1982 *Europe and the People without History*. University of California Press, Berkeley, CA

J. Beckett

Free Will and Action

According to the standard account of action, an agent's acts are those of his bodily movements that spring in an appropriate way from his will. Bodily movements that are caused in some other way—the bodily movement of someone who gets blown over by the wind, say—are not among the actions agents perform. More precisely, an agent's actions are those of his bodily movements that occur because he wills them to, where an agent's will is, in turn, identified with his desires and means-end beliefs: actions are bodily movements that occur because an agent desires some outcome and believes that the outcome can be produced by his moving his body in the relevant way. Bodily movements that trace their causal history to other states or events in an agent's psychology, or to nonpsychological states or events, are not actions (Davidson 1963, 1980).

The standard account of action suggests that all animals that are capable of desire and belief, whether those animals are human or nonhuman, are capable of action. The bodily movement of a mouse that occurs because it wants some cheese and believes that it can get some by running in the direction of the crumbs on the floor is an action of the mouse, just as much as the intricate performance of a gymnast or a ballet dancer. By contrast, the movement that the mouse makes when it is picked up by a cat and flung across the room is not an action, any more than the movement of a ballet dancer who is unceremoniously dropped to the floor by her partner.

Not all of an agent's actions, whether the agent is human or nonhuman, are free, however, notwithstanding the fact that they all spring from his will. When a mouse is caught in a trap by its tail, it may give up the struggle and sit down near the trap. Its body remains under its control, not under the control of some alien force, but staying near the trap is not something that the mouse does freely. Similarly, when

a man is locked in a prison cell he may give up the struggle and sit on his chair behind bars. But staying behind bars is not something that he does freely. What, then, within the class of actions, is the mark of those of an agent's actions that are free, as opposed to those that are unfree?

Some hold that free actions are all and only those that are done intentionally: that is, they are those of an agent's actions which, suitably described, can be seen to have been rationalized by his desires and beliefs (Davidson 1973, 1980). This idea is supposedly supported by paradigm cases of unfree action, such as that of the man and the mouse. For though the mouse stays near the trap and the man stays behind bars, neither does what they do intentionally under these descriptions. The man, for example, intentionally sits on a chair, as opposed to on the ground—that is something that is rationalized by his desires and beliefs—and one consequence of his sitting on the chair is that he stays behind bars. But he does not have any desires and beliefs that rationalize his staying behind bars. That is merely a consequence of what he does intentionally, not itself something that he does intentionally.

But though this certainly captures one sense in which we might say of actions that they are under an agent's control, it does not capture the idea of a free action. It entails, for example, that whenever someone who acts in a certain way knowingly, but not intentionally, he does not act freely, and this is surely quite wrong. Suppose that someone, A, desires to convey to another, B, the low opinion he has of a third person, C. But let's suppose further that he goes on to tell B in full knowledge of the fact that C is standing within earshot, and so causes C to become embarrassed. He could have waited until C left, but he didn't because he didn't care. A then knowingly causes C to become embarrassed, but he does not do so intentionally, as there is no desire of A's such that he believes he can satisfy that desire by embarrassing C. C's embarrassment is merely a foreseen consequence of what he does do intentionally, which is to tell B his low opinion of C. It would, however, be extremely implausible to suppose that the fact that C's embarrassment was merely foreseen, rather than intentionally produced, makes it the case that A didn't cause C's embarrassment freely. It was, after all, entirely up to A whether or not C would be embarrassed by what he said. He wasn't forced to embarrass C.

The more standard idea has thus been that an agent's acts are free just in case he could have done otherwise than he in fact did when he acted (Wiggins 1973). The question 'Did the agent act freely?' is in this way supposed to reduce to the question 'Did the agent have any alternative option?' A's causing C to become embarrassed is therefore indeed a free action, on this way of seeing things, because A had the option of acting in some other way instead. He could have waited until C was out of earshot. This is the relevant sense in which free actions are under an agent's control. The reason that a mouse caught in a trap does not stay in the vicinity of the trap freely is thus that it has no other option: it cannot do otherwise but stay near the trap. The reason a man who sits on a chair in a locked prison cell does not stay behind bars freely is that he, too, has no other option: he cannot do otherwise than stay behind bars.

However, this more standard account of free action immediately raises a difficult question, for without a special explanation of what the 'could' means, it seems to imply that freedom of action is impossible in a deterministic world. Every action performed by an agent in a deterministic world is fixed by the way the world was long before the agent was born together with the operation of causal laws. In a very familiar sense of 'can,' no agent can ever do otherwise than he in fact does. Agents are therefore always forced to do what they do by their past and the forces of nature.

Incompatibilists—those who think that freedom is impossible in a deterministic world—happily accept this conclusion (van Inwagen 1975). But they in turn divide over the substantive question whether the world is or is not deterministic. Those who think that it is—the so-called 'hard' determinists—draw the conclusion that there are no free actions. Those who think that there are free actions—the 'libertarians'—draw the conclusion that the world is not deterministic. But not everyone is an incompatibilist. Compatibilists find it difficult to see how the mere fact of causal determinism is supposed to undermine freedom of action. If anything, they insist, things are the other way around. Events that cannot be explained causally are completely random: this is what it means to say that an event eludes causal explanation. But free actions can hardly be randomly produced. Free actions, like all actions, are produced in a very reliable fashion by the desires and beliefs of their agents. They therefore conclude that free actions, if there are any, are caused by desires and beliefs, and hence that the sense of 'can' at issue must be different from the one that so impresses the incompatibilists (Ayer 1954, 1982).

Compatibilists have therefore coupled the suggestion that an agent's acts are free just in case he could have done otherwise with a special explanation of what this 'could' means. To say of an agent that he could have acted otherwise, in the sense relevant to establish freedom of action, is, they suggest—or, at any rate, this is the way the suggestion was originally put—to say of him that he would have done otherwise if he had chosen to (Moore 1911). In other words, freedom of action turns on whether or not, in the nearby possible worlds in which agents choose to do otherwise than they do in actuality—these are possible worlds in which the past is ever so slightly different—they succeed in doing what they choose to do. The truth of falsity of determinism is irrelevant.

The intuitive idea here should be clear enough, for the suggestion amounts to the claim that agents enjoy freedom of action if the way the external world presents no barrier to their efforts to act as they will: freedom is a matter of the absence of such external barriers or constraints. To see how this allows us to sort cases of free from cases of unfree action, consider the difference between two men who stay in a prison cell, one of whom is locked in, and one of whom isn't. The difference is that the one, but not the other, would have walked out of his prison cell if he had chosen to do so: in the nearby possible worlds in which the agent who is locked in his prison cell chooses to walk out, he fails to do so, as he is still locked in, whereas in the nearby possible worlds in which the agent who is not locked in his prison cell chooses to walk out, he succeeds in doing so, as his cell presents no barrier to him. The one has the option to walk out, whereas the other doesn't. Moreover this account is, apparently, fully general. Of two mice that stay nearby a trap, the one whose tail isn't caught in the trap, and so moves away in the nearby possible worlds in which it chooses to do so, stays nearby the trap freely, whereas the one whose tail is caught in the trap, and so doesn't move away from the trap in the nearby possible worlds in which it chooses to do so, does not stay nearby the trap freely. The external world presents a barrier to the movements of the latter, it serves as a constraint, but it presents no such barrier or constraint to those of the former.

Intuitive though this idea might be, however, the suggestion in the form in which it was originally made is problematic. The problem is that, since choosing is an action, it is something that people can do freely or unfreely (Chisholm 1964, 1982). But now suppose we ask whether a particular action is free. This turns on whether the agent could have done otherwise than he in fact did, and, according to the suggested analysis of 'could,' this turns on what happens in those nearby possible worlds in which the agent chooses to act otherwise. But now note that, whatever answer we give to this question, we have been forced to admit that the status of the original action as free or unfree turns on what happens when another merely possible act is performed, an act whose status as free or unfree in turn needs to be determined. Moreover, the status of this merely possible act turns on whether or not, in those nearby possible worlds in which the agent chooses otherwise than to choose to act otherwise, he succeeds in so choosing. Whatever answer we give to this question, it follows that we have therefore been forced to concede the possibility of yet another act whose status as free or unfree needs to be determined. And so we must go on.

The problem here is the need to posit further acts of choosing to act otherwise, and choosing otherwise than to choose to act otherwise, and so on. The problem is not just that questions about freedom of action are never answered, once and for all—every question answered leaves another posed but unanswered, and so on *ad infinitum*—the problem is rather that it is implausible to suppose that every creature capable of free action is capable of such ever higher and higher order acts of choosing. In order to raise the mundane question whether a mouse, caught in a trap, sits nearby the trap freely or unfreely, it is surely quite incredible to suppose that we must thereby concede the possibility of its choosing to do otherwise than to sit near the trap, and choosing otherwise than to choose to do otherwise than to sit near the trap, and so on. Mice simply aren't cognitively sophisticated enough to make such choices.

How are we to solve this problem? The suggested analysis of 'could have done otherwise' clearly needs to be rephrased so that the possibility of the freedom of an agent's actions does not entail the possibility of his making ever higher and higher order acts of choice. We should therefore suppose that for it to be true that an agent could have done otherwise it has to be the case that he would have done otherwise, not if he had chosen, but rather *if he had wanted*. This not only removes the problem just encountered—it is indeed possible for agents whose acts are free, men and mice alike, to want other than they actually want, and that possibility does not commit us to the possibility of their making higher order acts of choice—it also better captures the intuitive idea presented at the outset, the idea that an agent's actions are free to the extent that they are dependent on his will. For, as we saw, an agent's will is to be identified with his wants and means-end beliefs. To say of an agent, be he man or mouse, that he acts freely is thus, according to the rephrased suggestion, to say of him that, in the nearby possible worlds in which he wants something different, he acts in that different way. This is what it is for him to have alternative options.

Note two consequences of this account of free action. The first is that certain acts which we would ordinarily suppose to be unfree aren't really unfree. When someone says 'Your money or your life!,' and I give him my money, we would ordinarily say that I don't act freely because I have been forced to hand over my money. But, according to the account on offer, I do act freely. I have another option: I refuse in those nearby possible worlds in which I want not to give him my money more than I want him not to kill me. Many are unhappy about this. They think that there must still be a sense in which my freedom is compromised by such a threat. They point out that though, in the nearby possible worlds in which I want to refuse, I do refuse, it isn't the case that wanting to refuse would be at all reasonable, in those circumstances. Indeed, not only would it not be reasonable, it would be completely crazy. There is, they suggest, thus the following sense in which the person who threatens me forces me to act: he puts me in circumstances in which there is no other reasonable option but to comply with his threat. But if this right—that is, if this

is the sense in which threats compromise my freedom—then it seems to follow that people who make me irresistible offers equally compromise my freedom, for they too put me in circumstances in which the only reasonable option is to accept their offer. Whether it is acceptable to suppose that the only sense in which threats compromise my freedom is a sense in which offers also compromise my freedom clearly thus needs much more thought (Frankfurt 1973, 1988).

The second consequence concerns the relationship between free action and moral responsibility. According to this account of free action, freedom is not sufficient for an allocation of moral responsibility. Mice often act freely, but they are not morally responsible for any of their acts. Something else is therefore required. Other more complicated cases tell in favor of the same conclusion. Consider, for example, the plight of a reluctant alcoholic who wants to have a drink, and so has one, knowing he shouldn't. He presumably drinks freely, in the sense provided by the above account of free action, as in the nearby possible worlds in which he wants not to have a drink, he does not drink. Yet it is surely an open question whether or not he is morally responsible for his drinking. If, for example, his desire to drink is literally irresistible, then he is not responsible. Indeed, we would even think that he isn't free, in the important sense of being unable, in the circumstances, to want to do anything else except to drink. This more complicated case thus suggests that in order to understand the conditions of moral responsibility we must look not so much to an account of free action, but rather to an account of freedom of the will. The idea is that, though the alcoholic's actions may be free, his will is not, and that this is therefore the source of our judgements about moral responsibility (Frankfurt 1971, 1988).

Is the performance of a free action a necessary condition of moral responsibility? The answer seems to be that we can and do hold agents responsible for acting when they have no alternative, and that acting freely would thus seem not to be a necessary condition either. Those who act in the false belief that their acts are free are often held morally responsible for what they do, notwithstanding the fact that they don't act freely. Imagine, for example, that when a family visits an old prison they disperse so that they can explore the place individually, but that they agree to meet at a certain time to go home. When the agreed time to leave arrives suppose that one of the children fails to turn up. His parents look for him and find him in a cell. Because he doesn't want to leave, he has hidden from them. But now suppose that the cell door is closed, and that, unbeknown to the child, it is not just closed but also locked. It then follows that the child had no alternative but to be in the prison cell when it came time to leave. He couldn't have met the rest of his family. Is the child off the hook in terms of responsibility? Hardly. His parents would surely quite rightly continue to hold him morally responsible for failing to turn up. The fact that he couldn't have done otherwise is irrelevant because it doesn't figure, in the right kind of way, in the explanation of the child's failure to meet them. In order to see what is relevant, though, we must once again look at what does figure. This case, too, suggests that we need to look more carefully to an account of freedom of the will (Frankfurt 1969, 1988).

Now that we have seen what it is for an agent's acts to be free, we must ask what it is for his will to be free. It might be thought that we should model our account of freedom of the will on the account we give of freedom of action. Thus, the suggestion might be, just as freedom of action is a matter of being in circumstances in which one can do whatever one wants to do, so freedom of the will is a matter of being in circumstances in which one can want whatever one wants to (Frankfurt 1971, 1988). Where freedom of action is a matter of the absence of external constraints on action, so freedom of the will is a matter of the absence of internal constraints on wanting. An agent's will is free just in case, in those nearby possible worlds in which he wants to want to act in a certain way, he wants to act in that way.

There is clearly something right about this idea. Those who enjoy freedom of the will do seem to have the capacity to control their will, where this includes the capacity to control their first-order desires. The problem with the reluctant alcoholic seems to be that he has no such ability: his desire to drink is beyond his control. What is less clear, however, is that this idea is well captured by the suggestion just made. The desires an agent has about his first-order desires are, after all, just more desires he has, and are hence just another part of his will, a part which may or may not be under the agent's control (Watson 1975, 1982). Thus, even supposing it to be true that in those nearby possible worlds in which an agent wants to want to act otherwise, we can always imagine that the agent is unable to have wanted to want to act otherwise. If the agent were unable to want to want to act otherwise then we must surely ask in what sense he enjoys freedom of the will. Certain of the desires such an agent has in certain possible worlds are controlled by certain others in those worlds, but the agent himself doesn't seem to be in control of any desires at all.

Some theorists have argued that in order to solve this problem we need to postulate a special kind of higher order desire: a desire which captures the agent's own sense of identity (Frankfurt 1987, 1988). When an agent has this sort of higher order desire, they suggest, he identifies himself with the lower order desires that are this higher order desire's objects. Freedom of an agent's will is, the suggestion goes, a matter of his being able to have desires that match this sort of higher order desire: in other words, it is a matter of having desires with which he identifies.

The difficulty with this suggestion, however, is to come up with a suitably independent account of

identification. There certainly is a sense in which an agent who acts of his own free will identifies with the desires on which he acts, as he is himself the originating source of such desires. But this is not an independent sense of identification. It is a sense that we define in terms of the idea of desires originating in a free will. There is also at least the following independently characterizable sense of identification. Some people who take drugs against their will, and who struggle to overcome their addiction, ultimately give in, and so come to identify with their addiction in the sense of building all of their expectations around it. In other words, they come to see themselves as hopeless drug addicts: this becomes their self-conception. But this is hardly a sense of identification that will help us understand the idea of desires originating in a free will.

There is something right about connecting freedom of the will with the possession of desires with which an agent identifies, however. Since agents are the originating source of such desires it does seem that, when they act of their own free will they express their true nature. The desires on which they act are, in some sense, essential to who they are. But though this seems right, it in fact suggests a quite different account of free will, one which does not depend crucially on our having an independent idea of what it is for an agent to have desires with which he identifies.

Since the Enlightenment we have come to conceive of people as being essentially truth-oriented reasoners. They are creatures who have the capacity to recognize the truth and to respond appropriately to it. Insofar as we are deliberators, what this suggests is that we are essentially creatures who have the capacity to recognize and respond to the facts about what we have reason to do. Thus, according to this account, an agent whose will is free is someone who asks himself what there is reason to do and, in answering this question, he is sensitive to the available evidence. This is not to say that he will inevitably acquire true beliefs about what there is reason to do, but it is to say that when he fails to do so, and when the evidence was available to him, his failure will be traceable to a failure to exercise his belief-forming capacities appropriately: he could and should have formed the correct belief, but he failed to do so. Moreover, an agent whose will is free is someone who, in the formation of desires, is sensitive to the beliefs he has about what he has reason to do. Again, this is not to say that he will inevitably form desires that accord with his beliefs about his reasons, but it is to say that when he fails to do so, his failure will be traceable to a failure to properly exercise self-control: he could and should have formed the correct desire, but he failed to do so (Wolf 1990).

As is perhaps clear, this conception of free will differs crucially from the one suggested earlier. What is important for freedom is indeed a capacity to desire otherwise, but the possession of this capacity is not to be analyzed in terms of what happens in nearby possible worlds in which the agent wants to want otherwise. Instead it is to be analyzed in the same way, whatever that way is, that we analyze the ability rational thinkers possess to respond appropriately to evidence in the formation of their beliefs. Just as a rational thinker who fails to adjust his beliefs to the available evidence makes a mistake he shouldn't have made, and so, we suppose, must have had the capacity to make the adjustment and believe rightly instead, so a rational deliberator who fails to form beliefs about what he has reason to do in accordance with the available evidence, or who fails to desire in accordance with his beliefs about what he has reason to do, makes a mistake he shouldn't have made, and so is supposed to have had the capacity to control himself and acquire the correct belief or desire instead (Pettit and Smith 1996).

Note two consequences of this account of freedom of the will. The first is that it looks tailormade to provide an account of the conditions of moral responsibility. Those who have no capacity to believe that they have reason to do anything other than they in fact believe they have reason to do, or to desire to do anything other than they in fact desire to do, can hardly be praised for having the right beliefs and desires, if it so happens that they're right, or blamed for having the wrong ones, if it so happens that they're wrong. This explains why we don't hold the alcoholic responsible. The irresistible nature of his desire to drink makes him incapable of an exercise of self-control, despite the fact that he believes, correctly, that he has reason to abstain. It also explains why we don't hold mice morally responsible for what they do. Their lack of cognitive sophistication deprives them of the capacity even to form beliefs about their reasons for doing what they do, let alone the capacity to desire accordingly.

Moreover, nothing other than the capacity to form the right beliefs and desires would seem to be required for people to be responsible either. This is why we hold the child described earlier morally responsible for failing to meet his parents at the appropriate time, notwithstanding the fact that he was locked in a prison cell and so couldn't have met them. We hold him responsible because we assume that his behavior was in fact the product of a desire he had which accorded with his beliefs about what he had reason to do, a belief that we take to be culpably mistaken. It is this that attracts our censure.

The second consequence concerns the possibility of freedom of the will in a deterministic world. To my knowledge, no one yet has a fully worked out theory of what it is for someone to have a rational capacity. It therefore follows that, to this extent at least, a proof that freedom of the will is possible within a deterministic world is yet to be provided. Provision of such a proof would, after all, require a demonstration that individuals in a deterministic world have the wherewithal to possess just such a rational capacity. On the face of it, however, it must be said that there doesn't

look to be any particular difficulty in principle involved in providing such a demonstration. We regularly discriminate between people who do and people who don't have rational capacities, and, in so doing, we seem to pay no regard whatsoever to issues of determinism. This is because even if determinism is true, that only shows that people will make the mistakes that they make. It doesn't show that they couldn't help making those mistakes, in the sense of having no rational capacity to get things right.

But, whatever we ultimately come to think about the possibility of rational capacities in a deterministic world, the crucial lesson of this account of freedom of the will is that, if there do turn out to be such difficulties, then it isn't just the idea of freedom of the will that becomes problematic, it is the very idea of a rational agent as someone who has the capacity to recognize and respond appropriately to the truth. Whether freedom of the will is possible in a deterministic world or not, its possibility goes hand in hand with the possibility of agents who have the capacity to form beliefs and desires in a rational manner.

Bibliography

Ayer A J 1954/1982 Freedom and necessity. In: Watson G (ed.) *Free Will*. Oxford University Press, Oxford, UK

Chisholm R M 1964/1982 Human freedom and the self. In: Watson G (ed.) *Free Will*. Oxford University Press, Oxford, UK

Davidson D 1963/1980 Actions, reasons and causes. In: Davidson D (ed.) *Essays on Actions and Events*. Clarendon Press, Oxford, UK

Davidson D 1973/1980 Freedom to act. In: Davidson D (ed.) *Essays on Actions and Events*. Clarendon Press, Oxford, UK

Davidson D (ed.) 1980 *Essays on Actions and Events*. Clarendon Press, Oxford, UK

Frankfurt H G 1969/1988 Alternate possibilities and moral responsibility. In: Frankfurt H G (ed.) *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press, Cambridge, UK

Frankfurt H G 1971/1988 Freedom of the will and the concept of a person. In: Frankfurt H G (ed.) *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press, Cambridge, UK

Frankfurt H G 1973/1988 Coercion and moral responsibility. In: Frankfurt H G (ed.) *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press, Cambridge, UK

Frankfurt H G 1987/1988 Identification and wholeheartedness. In: Frankfurt H G (ed.) *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press, Cambridge, UK

Honderich T (ed.) 1973 *Essays on Freedom of Action*. Routledge & Kegan Paul, Boston, MA

Moore G E 1912/1944 *Ethics*. Oxford University Press, Oxford

Pettit P, Smith S 1996 Freedom in belief and desire. *Journal of Philosophy* **93**(9): 429–49

van Inwagen P 1975 The incompatibility of free will and determinism. In: Watson G (ed.) *Essays on Actions and Events*. Clarendon Press, Oxford, UK

Watson G (ed.) 1982 *Free Will*. Oxford University Press, Oxford, UK

Watson G 1975/1982 Free agency. In: Watson G (ed.) *Free Will*. Oxford University Press, Oxford, UK

Wiggins D 1973 Towards a reasonable libertarianism. In: Honderich T (ed.) *Essays on Freedom of Action*. Routledge & Kegan Paul, London

Wolf S 1990 *Freedom within Reason*. Oxford University Press, New York

M. A. Smith

Freedom/Liberty: Impact on the Social Sciences

The concept of freedom, or in another rendering in the English language: liberty, is one of the most basic concepts in social and political thought. In many respects, it is the basis of all thought about politics and society. The assumption, in some form, of free will in human beings, namely, raises the question of action as making a difference in the world, and the relation of that difference to the course of the world had no action occurred (see *Free Will and Action*). More particularly, the term political freedom describes the possibility of a multitude of human beings deliberating about what they want in common and acting in common (see *Freedom: Political*).

The tension between the freedom of the singular human being and political freedom is at the center of all modern political thought, that is, all political thought that assumes the human ability of giving themselves their own laws. (Sometimes, this difference finds expression in the terminological choice between liberty, more consistently used for individual freedom, and freedom, for political freedom. Most authors, however, do not make any distinction between these two terms, and none will be made here.) It is also in the background of the theoretical *problématique* of agency and structure in the social sciences.

Comparative research on the history of the idea of freedom is but little advanced, partly due to prevailing preconceptions about the Western origins of freedom, partly due to the intrinsic difficulties of historical comparisons of the most basic linguistic concepts. Refraining from any explicit comparison, however, it is possible to state that the 'Western' tradition of political thought is marked by the confluence of both ideas, individual and political freedom (Patterson 1991). The latter originated in ancient Greece with the idea of self-government of the city states, and led into the Republican tradition, first of the Roman Republic, and then refounded by the Italian city republics, emerging from under the imperium, in the course of

what then became called the Renaissance (Skinner 1978). The former had important roots in early Christianity, as a religion that established more direct links between the singular human beings and their conscience and God, not least under conditions in which the declaration of Christian belief could lead to persecution by the political rulers.

The tension between the two concepts of individual freedom and political freedom is widely acknowledged in political philosophy (see, for instance, Berlin 1969, Miller 1991). This is not to say, however, that a resolution of that tension, or a reconciliation between the concerns that stand behind the two ideas, is not possible. Many authors argue, for instance, that republicanism is a political theory that succeeds in combining both exigencies (see *Republicanism: Impact on Social Thought*), or that romanticism is a social philosophy that underpins self-realization while also developing an idea of a viable polity (see *Romanticism: Impact on Social Thought*). Historically, though, it has been the case that a political theory emerged as central for the self-understanding of the European nation–states as well as for the USA that was based on the concept of individual liberty. Increasingly so during the nineteenth and twentieth centuries, republican views of political freedom were marginalized, and romanticist views of the embeddedness of the singular human beings were assigned a merely supportive, even though not insignificant, role in sustaining polities.

Recent research in the history of political concepts has emphasized that this shift in balances between various political languages took place between the middle of the eighteenth and the middle of the nineteenth century. This period is at the same time the period of the emergence of the early social sciences. It is also the period of the 'democratic revolutions' (R. R. Palmer), of the American and the French Revolution, which are often seen as aiming at giving institutional expression to the principle of liberty. It is from this historical constellation that an assessment of the impact of the concept of liberty on the social sciences has to start out.

1. *Liberty and the Moment of the Revolutions*

The pervasiveness of political issues, and in particular of the issue of liberty, in the early social sciences, that is, in authors like Smith, Condorcet, or Hegel, is evident. In political terms their modes of reasoning were, however, marked by a curious paradox. These authors were working in the shadow of liberal thinking (in a very broad sense), namely accepting the right of human beings to self-determine their individual and social lives, even though not always without important qualifications (see *Liberalism: Impact on Social Science*). At the same time, though, they were holding that the free action of human beings would not actually generate, as one might innocently assume, a very wide range of possible individual and social forms and orientations of behavior. Given complete autonomy, human beings would reveal themselves not to be free in a radical sense, but driven by a limited number of well intelligible inclinations. One might even infer that it was exactly this predictability and order that gave grounds and reasons—to these authors themselves and to more hesitant others—not to reject the normative project of liberty. And this linkage of freedom and predictability became particularly important at the historical moment when the externally imposed barriers to free deliberation were threatened to be removed, the moment of the American and French Revolutions.

These revolutions gave institutional expression to the political aspect of a broader culture of individual autonomy that is a key element of modernity. In this sense, much of this era can be seen as a liberation of human beings from imposed ties, but this liberation was far from unproblematic. As Lefort (1986, pp. 214–15) once described this feature of modernity: 'When he is defined as independent, the individual does not exchange … one certainty against another one. … The new mode of existence of the individual within the horizons of democracy does not merely emerge as the promise to control one's own destiny, but also and not less as the dispossession of the assurance as to one's identity—of the assurance which once appeared to be provided by one's place, by one's social condition, or by the possibility of attaching oneself to a legitimate authority.' Liberation is here interpreted as an increase of contingency and uncertainty in the lives of human beings.

If this view were unequivocally valid, one should expect a philosophy of contingency—in Rorty's style (1989), for instance—linked to a liberal-individualist political theory to dominate the intellectual scene forever after the successful revolutions. However, historically this was not at all the case. In contrast, 'the historical moment, about which we speak, emerges in such a way that the real rising of the political instance entails its theoretical abatement' (Manent 1994, p. 123). Instead, the historical moment of liberty coincided with the rise of the social sciences. 'Society' as the object of the social sciences has been a 'post-revolutionary discovery'; or, to put it even more succinctly, 'the sociological point of view constitutes itself in the moment when the notion of liberty becomes the principal articulation of the human world' (Manent 1994, p. 75, 113, Therborn 1976, for an earlier analysis of the emergence of sociology 'between two revolutions'). Such apparent paradox reveals the aporia of political thought after liberation. Very generally speaking, the social sciences are exactly a part of the response human beings gave to their new condition of—self-incurred, one might say—contingency and principled uncertainty. Being unable to rely any longer on externally defined certainties,

political thinkers started searching for regularities and continuities which exist without being commanded. The social sciences have been a means to decrease contingency.

2. Concepts of Liberty before the Revolutions

Earlier political thought had already recognized the inevitable circularity of a reasoning in which unity should rise from diversity as well as, at the same time, impose itself on diversity. At that time, however, there was room to try to approach the issue by mere conceptual construction and theoretical determination. As long as there were barriers to the practice of liberty, every conceivable view on its outcomes and consequences for the polity could be held. The mode of reasoning was bound to change when those barriers were removed and experiences were made and when practical issues, such as the founding of political institutions, had to be tackled.

Political philosophy, in its prerevolutionary versions from Machiavelli and Hobbes onwards, gradually had developed an understanding of that which people have to regulate in common on the basis of the assumption that they do this on their own, without any external demands. The line of thinking that leads from Hobbes to Locke and Rousseau is often seen as the source of Western liberalism. It is most often presented as the theory of the social contract in which free individuals relinquish some of their rights in exchange for security and the regulation of the common realm provided by the state. In recent years, however, it has become common to distinguish the Hobbesian from the Machiavellian tradition. In this view, liberals define liberty as non-interference. The Hobbes–Lockean state, founded by free contract, dominates over the individuals, but it interferes with their liberties only to the degree in which the maintenance of order makes this necessary. The liberal tradition needs to draw a strong boundary between the public and the private, and since non-interference of the public into the private is the supreme principle, it can only have a 'thin' concept of membership in a polity. Significantly, the concepts of the state and of membership can be developed in the abstract, since they are based on a reasoning each rational individual is seen to arrive at, regardless of the situation.

In contrast, republicans define liberty as nondomination. Drawing via Machiavelli on Roman political thought, nondomination is conceptualized in stronger terms than noninterference; it requires security against interference. Such security stems at least in part from the ways in which citizens relate to each other, so that there is a less sharp divide between the private and the public and a 'thicker' concept of membership than in liberalism (see Pettit 1997, Skinner 1998 for recent accounts). As a consequence, the shape and substance of each republican polity is specific, since it depends on its members and their deliberations. Most contributors to this recent debate emphasize the distinction between liberalism and republicanism to underline that individualist liberalism was not the only alternative available to provide the discursive underpinning of political modernity at its onset. However, they also agree that republicanism was by and large abandoned around the turn of the eighteenth century, and liberalism, though it neither appeared particularly powerful nor coherent before, very soon emerged as the pivotal political theory in postrevolutionary polities.

3. The Impact of the Revolutions on the Concept of Liberty

At a closer look, it is not too difficult to understand such intellectual shift through a look at the deep shock the revolutions meant to political thinking (Wagner 1998). The revolutions were certainly inspired by republican thinking. Looking at them in the somewhat stylized way that an emphasis on the French experience has long suggested, one recognizes, however, that they aimed at combining two objectives that proved to be practically impossible to hold together. On the one hand, they aimed at transforming state sovereignty in the hands of the monarch into popular sovereignty, that is, they worked with extended notions of citizenship and liberty. On the other hand, they held such a transformation of the polity to be conceivable only in the form of the existing territorial state and within its dimensions.

Such double transformation entailed, first, that the existing relations between the people, tainted with the suspicion of domination and privilege characteristic of feudal society, had to be weakened or abolished. Thus, however, a major available resource for a substantive, 'thick' grounding of a modern republic was rejected. Second, the idea of extending political rights widely cast doubts about the viability of a demanding concept of liberty such as the one upheld in the republican tradition. Caution seemed to demand, not least for some more conservative observers, to limit the substance of the concept at the moment at which its reach was extended. As a consequence, the public realm, the state, was robbed of most of its substance and the formal process through which common deliberations were reached was emphasized instead. About such processes, though, there is much less to reason in politico-philosophical terms. The liberal conception that things regulate themselves as soon as atomistic individuals are endowed with the right and the capacity to follow their own interests and passions makes political thought almost superfluous. It suggests that nothing beyond some basic assumptions and rules as to how to set up such processes is needed (though

such rules are much less innocuous than it may have appeared to some of the revolutionaries).

This change is most visible in the context of the foundation of the United States of America, at least in a reading of it, inspired by Gordon Wood, that tries to go beyond the debate over the republican or individualist beginnings of the American polity. In their reaction against obtrusive and illegitimate government, the Americans consciously tried to build institutions so that 'power may not actualize itself anywhere.' The 'disembodiment of government … goes along with a disentangling of power, of the law, and of the knowledge of the ultimate ends of a society' (Lefort 1991, p. 27). Many of the most important themes of earlier political philosophy were simply ejected as illegitimate from reasoning about political institutions. 'For the first time, the place of power is easily recognized as being empty' (Lefort 1991, p. 27). The Americans of the revolutionary generation had 'broken through the conceptions of political theory that had imprisoned men's minds for centuries' through the renunciation of any substantive foundation of the polity and 'a total grounding of government in self-interest and consent' (Wood 1998, p. 614, 612).

Even in the American context, this deliberate political construction could not exorcise all concerns about the substantive foundations of the public realm. Beyond the calls for republican virtue during the early period, the search for a preordained unity emerged in the nineteenth century, pursued often by Americans with some German intellectual background (Gunnell 1991). However, such political science could not establish itself as a dominant paradigm and was mostly abandoned after the end of the nineteenth century. In Europe with its long-established statist institutions, the idea of unity was not abandoned, but it was transformed rather radically. As in the US, the observation seemed undeniable that, once the reasonable will of the human beings had been cast into institutions, the political order must be seen as intrinsically satisfactory (Manent 1994, pp. 228–9). In this organized context, however, the fear that such new order based on human liberty might turn out unsustainable was much greater than in North America. 'The effect of liberty to individuals is, that they may do what they please: We ought to see what it will please them to do, before we risque congratulations,' as Burke (1993 [1790], pp. 8–9) put it in his reflections on the French Revolution.

4. *Liberty and the Rise of the Social Sciences*

In this situation, individualist liberalism provides the less demanding concept of ('negative') liberty as noninterference that may live up to Burke's requirements under the proviso that the state is capable of maintaining order for and above the individuals. However, there is another way out of the aporia of liberty, namely the attempt to arrive at knowing 'what it will please them to do' by other means. In Europe, indeed, the idea of an emptiness of the political space did not fully emerge. In the Enlightenment combination of freedom and reason, the state was rather reinterpreted as the incarnation of reason once the expression of human freedom fed into its construction. This is one of the central themes of Hegel's writings, but it can be found in many other contemporary works as well (Wuthnow 1989, p. 315). This linkage of knowledge and politics, the idea of 'legislative reason' (Bauman 1987), was characteristic of much of the early social science thinking during the revolutionary period in Europe. Reason, rather than merely serving as an ideal point of philosophical orientation (which it also continued to be), was to be made practical in the elaboration and application of the rules of politics. The movement to transform political reason into a science of politics found its most ardent spokespersons in France where politics had most radically been put onto new foundations. Beyond the voluntary commitment to construct a new social order, there was also a feeling of dire necessity to do so on well-informed grounds, not least motivated by the shock of revolutionary terror. The writings of Saint-Simon and Comte mark one major line of such thinking.

The American and French revolutions, thus, strongly suggested to study what held human beings together, how they would actually organize their lives—individually, in 'associations' or 'social movements,' and in the polity and the 'nation'—and what kinds of regularities and orders could be expected, if people were permitted to do so on their own, without imposed restrictions. This is the search for social ties that is one major root of the social sciences, and it is in this sense a politically motivated search.

The hope and aspiration was that the moral and political sciences should and could now achieve 'the same certainty' as the physical sciences (Baker 1975, p. 197). Certainty was a requirement of some urgency, since the new political order needed assurances of its viability. But it was also regarded as an historically new possibility, since political action was liberated of the arbitrariness of decisions by rulers of doubtful legitimacy and given into the hands of the multitude of reason-endowed human beings. The 'blend of liberalism and rationalism' (Baker 1975, p. 385, about Condorcet) can thus be explained as stemming from the same source, the Enlightenment linkage of freedom and reason—though possibly in its most optimistic version.

This linkage, if it could be sustained, made a science possible, where according to the classical view it would have been unthinkable, in the realm of the political. Condorcet subscribed to the new concept of free expression of political wills, but for him there was no contradiction between that view and the possibility of scientific analysis. 'Societal choice' could well be

transformed 'into the rational decision-making of the idealized republic of science' (Baker 1975, p. IX, 193).

5. *Varieties of Social Science Responses to Liberty*

The rights-endowed individual became thus, in such views, the only conceivable ontological as well as the methodological foundation of a science of political matters after the revolutions. Once the rights of man had been generally accepted as self-evident and unalienable, it seemed obvious, to Turgot and Condorcet for instance, that they were also 'the logical foundation of the science of society' (Baker 1975, p. 218). In rights-based liberalism, the individual is the only category that need not, often in fact: cannot, be debated. The individual is simply there, whereas everything else—for instance, what criteria of justification are to be applied when determining the collective good—needs to be argued about. Substantive aspects of human interaction are subject to communication and consensus. And, to make the issue even more complicated, with whom one should enter into communication, that is, the boundaries of the community, is itself not given, but subject to agreement.

Once this assumption was accepted, basically two avenues of constructing a science of the political had opened. Both these forms of social theorizing connect modernist political philosophy, that is, individualist liberalism, to a science of the political. One possibility was to try to identify by theoretical reasoning the basic features of this unit of analysis, the individual human being, and its actions. Since this unit was conceived as an ontological starting-point, devoid of all specific, historical and concrete, ties to the world, its characterization was to proceed from some inherent features. From earlier debates, those features had often been conceived as twofold, as passions and as interests. In the late Enlightenment context, the rational side of this dichotomy was regarded as the one amenable to systematic reasoning. It thus allowed the building of a scientific approach to the study of at least one aspect of human interaction with the world, namely the production and distribution of material wealth. This approach inaugurated the tradition of political economy, later to be transformed into neoclassical economics and, still later, into rational choice theory. The moral and political philosophy of the so-called early modern period split into a political theory based on the idea of the social contract and a rationalized moral theory based on the idea of exchange. In both cases, the individual is the starting-point and unit of the analysis.

While political economy was based on a highly abstract, but for the same reasons extremely powerful, assumption of human rationality, the other conclusion from the individualist foundational principle was arguably even more reductionist but much more cautious. Avoiding any substantive assumptions on the driving forces in human beings at all, the statistical approach, often under the label of political arithmetic, resorted to the collection of numerically treatable information about human behavior. The space of substantive presuppositions was radically emptied in this thinking, but the methodological confidence in mathematics increased in inverse proportion (Brian 1994, Desrosières 1993).

Thus, two strands of political thought that had been proposed and elaborated for some time rose to new and greater prominence, political economy and political arithmetic. The denominations these approaches were known by in the late eighteenth century referred explicitly to political matters. Both were to lose these attributes in the nineteenth century when they had consolidated their ways of proceeding and when the application of these cognitive forms had established predominance over political deliberation in decision on common matters, at least in the view of many economists and statisticians. Mostly, this terminological change has been interpreted as an autonomization of cognitive approaches and as a differentiation of the sciences into disciplines. However, it is not exactly appropriate to say that economics and statistics separated from politics. Once the approaches of the former two are accepted as valid, there is nothing political left to study.

The acceptance of the economic and the statistic ways of conceiving of the social world did not go without criticism; and they were never accepted as the only possible ways anywhere. However, the critiques and alternatives that were proposed most often accepted the fundamental change in political reasoning after the construction of a polity based on the assumption of free individuals. (Since nineteenth-century polities all worked with restrictions that are incompatible with a full-fledged individualist liberalism, one should say more precisely that such a polity was put on the horizon of political debate through the revolutions, rather than made actual in institutional form.) After such a polity had begun to come into existence, new problems were identified. These were now essentially liberal problems; they resulted, one might say, from the observation that not everything that was needed for organizing a liberal polity could indeed be derived from an 'original position' (John Rawls). Two main types of problems may be distinguished by reference to the hypothetical original position in which individuals meet under a veil of ignorance.

On the one hand, the range of conclusions that could be drawn from the assumption of free and equal individuals was too limited. These individuals' relations were structured by the existence of politically important 'prepolitical' social facts, of orientations and links between human beings that were seen to

already exist before individuals entered into political communication and deliberation. On the other hand, the working of the liberal rules would themselves produce new kinds of social relations, 'postpolitical' relations, which would have a structuring impact on the polity in turn.

Attempts to theorize prepolitical relations all start out from the critical observation that the human being who enters into political relations is not such a kind of individual as liberal political theory described it, and that the hypothesis of any original position would lead to serious flaws in the conclusions. This thinking emphasizes the rootedness of the singular human beings in contexts from which their ways of giving meaning to the world stem. The broadest intellectual movement of this kind has been the cultural-linguistic theory of the boundaries of the polity, which inaugurated culturalist thinking in the social sciences and also became one source of later nationalism.

The other main line of 'postrevolutionary' social thought started out from the insights, first, that the basic liberal assumptions, once they were cast into effective rules, would have durable and important effects on what social scientists would soon call the 'structure' of social relations. In this sense we can refer to those relations as being conceived as 'postpolitical.' And second, the question of such relations was forced onto the agenda of social and political thought by the fact that the liberal assumptions on their own did not suffice to create and justify a political order. The observation of structures of representation served the interest of enhancing some stability and certainty in political procedures that could appear to be opened to all contingencies by the abolition of any legitimacy of preordained orders (see *Representation: History of the Problem*). We find here again the two main strategies for rediscovering certainties, systematic observation and reflective conceptualization. These two intellectual responses to the political *problématique* inaugurate two further modes of social theorizing, the behavioralist and the structural-functional one. Unlike economics and statistics, they do not make individualistic assumptions but aim at grounding sociopolitical life in purely social forms.

After the claim for a social science had been made by authors like Saint-Simon or Comte, the themes of the emerging social sciences formed around the observation of this internal structuration of societies. In Germany, Lorenz von Stein reported about the 'social movements' in France that announced a major change in the social order. At mid-century, Robert von Mohl diagnosed a transitory situation after the *Ancien Régime* had disassembled and no liberal order had been able to assert itself in the German states, while beginning industrialization and urbanization placed new demands on the political orders (Maier 1980, Wagner 2000). Karl Marx gave a central place in his social theory (and philosophy of history) to the newly forming social phenomenon, the 'working class.' And Émile Durkheim provided a representation of society in which the elements of the social order were defined according to their position in the division of social labor and their relations regarded as interlocking in the form of 'organic solidarity.' Steps like these marked the construction of a sociological view-point which was clearly identifiable around the turn of the twentieth century.

6. *Liberty after the Social Sciences*

All these approaches to the social sciences can, under one aspect, be regarded as ways of dealing with the problem of contingency after the assertion of human freedom. It is on those grounds that sciences of the social and the political are proposed. If, in contrast, politics is seen as that human activity that by its nature is open, plural, and diverse (Arendt 1958), then any cognitive linkage of free action and predictable outcome is inconceivable. Orderly outcomes can only result from planned or routine activities, work, and labor in Arendt's terminology, over which certainty can be established before they are started. In contrast, political action in a context of liberty must go along with contingency of outcomes. From an Arendtian viewpoint, thus, those early social scientists established an impossible connection. If they were heading for a 'social science' trying to identify laws and regularities of human action and societal development, then they necessarily abandoned the heritage of political philosophy, the emphasis on creative agency, irreducible diversity, and the permanent possibility of unpredictable beginnings. It is in the light of such considerations that the closing decades of the twentieth century have witnessed a revival of political philosophies of freedom, often going beyond concepts of liberty as held in liberal political theory. These works, by authors such as Claude Lefort, Pierre Manent or, more historically oriented, Quentin Skinner, are not merely contributions to political philosophy or its history. Rather, they challenge the very separation of the social sciences from political philosophy.

See also: Citizenship: Political; Civil Liberties and Human Rights; Civil Rights; Civil Society, Concept and History of; Freedom: Political; Human Rights, History of; Methodological Individualism: Philosophical Aspects; Political Economy, History of; Rational Choice Theory: Cultural Concerns; Revolutions, History of

Bibliography

Arendt H 1958 *The Human Condition*. The University of Chicago Press, Chicago

Baker K M 1975 *Condorcet: From Natural Philosophy to Social Mathematics*. The University of Chicago Press, Chicago

Bauman Z 1987 *Legislators and Interpreters*. Polity, Cambridge, MA
Berlin I 1969 Two concepts of liberty. In: Berlin I (ed.) *Four Essays on Liberty*. Oxford University Press, Oxford, UK
Brian E 1994 *La mesure de l'Etat. Administrateurs et géomètres au XVIIIe siècle*. Michel, Paris
Burke E 1993 [1790] *Reflections on the Revolution in France*. Oxford University Press, Oxford, UK
Desrosières A 1993 *La politique des grands nombres. Histoire de la raison statistique*. La découverte, Paris
Gunnell J G 1991 In search of the state: Political science as an emerging discipline in the US. In: Wagner P, Wittrock B, Whitley R (eds.) *Discourses on Society. The Shaping of the Social Science Disciplines*. Kluwer, Dordrecht, The Netherlands, pp. 123–61
Lefort C 1986 *Réversibilité: Liberté politique et liberté de l'individu. Essais sur le politique. XIXe–XXe siècles*. Seuil, Paris, pp. 197–216
Lefort C 1991 Introduction. In: Wood G S (ed.) *La création de la république américaine 1776–1787*. Belin, Paris
Maier H 1980 [1965] *Die ältere deutsche Staats- und Verwaltungslehre*. Beck, Munich
Manent P 1994 *La cité de l'homme*. Fayard, Paris
Miller D (ed.) 1991 *Liberty*. Oxford University Press, Oxford, UK
Patterson O 1991 *Freedom in the Making of Western Culture*. Tauris, London
Pettit P 1997 *Republicanism. A Theory of Freedom and Government*. Clarendon, Oxford, UK
Rorty R 1989 *Contingency, Irony, Solidarity*. Cambridge University Press, Cambridge, UK
Skinner Q 1978 *The Foundations of Modern Political Thought, Vol. 1, The Renaissance*. Cambridge University Press, Cambridge, UK
Skinner Q 1998 *Liberty before Liberalism*. Cambridge University Press, Cambridge, UK
Therborn G 1976 *Science, Class and Society. On the Formation of Sociology and Historical Materialism*. New Left Books, London
Wagner P 1998 Certainty and order, liberty and contingency; the birth of social science as empirical political philosophy. In: Heilbron J, Magnusson L, Wittrock B (eds.) *The Rise of the Social Sciences and the Formation of Modernity*. Kluwer, Dordrecht, The Netherlands, pp. 241–63
Wagner P 2000 An entirely new object of consciousness, of volition, of thought. The coming into being and (almost) passing away of 'society' as an object of sociology. In: Daston L (ed.) *Biographies of Scientific Objects*. The University of Chicago Press, Chicago, pp. 132–57
Wood G S 1998 [1969] *The Creation of the American Republic 1776–1787*. University of North Carolina Press, Durham, NC
Wuthnow R 1989 *Communities of Discourse. Ideology and Social Structure in the Reformation, the Enlightenment, and European Socialism*. Harvard University Press, Cambridge, MA

P. Wagner

Freedom of the Press

No press system is absolutely free from various social, political, economic, or legal controls of its society. Most discussions of freedom of the press center on the relationship of government and the press, and here the basic question is not whether government controls the press, but how and to what extent. Freedom of the press refers to negative freedom—freedom from governmental constraint, or it may refer to positive freedom—the capacity to do or achieve certain ends. Most issues concerning press freedom have revolved around negative freedom, understood as the independence of the news media from governmental restraint.

1. The Term 'Press'

The 'press' in 'freedom of the press' is elusive. The general public considers the word press as the conventional institutional news media including newspapers, magazines, and radio and television broadcasters. But the popular image of the term press is too limiting to be accepted in US constitutional law. The press goes beyond mainstream newspapers and magazines to 'every sort of publication which affords a vehicle of information and opinion' (Lovell v. Griffin 1938).

Can the term press be distinguished from 'speech' in 'freedom of speech?' Thus far, the answer to this press–speech dichotomy question is 'No' in US law. The now well-established US rejection of the distinction between press and nonpress communicators presents a striking contrast to the clear-cut statutory recognition of the press in continental Europe and some Asian countries. It should be emphasized, however, that the presence (or absence) of a direct press law does not necessarily ensure a privileged (or unprivileged) press in a country: 'The theory that a special press law in itself constitutes a threat to press freedom has merit only if a special press regime is identified with an authoritarian press regime' (Lahav 1985).

2. Theories of Press Freedom

The key values sought by society in protecting freedom of expression are that freedom of expression assures individual self-fulfillment, helps to discover the truth, secures participation by citizens in social and political and economic decision-making, and promotes a stable change in society (Emerson 1966). The press is protected as the surrogate of the public when performing its informing and checking functions as a Fourth Branch, supplementing the executive, legislative, and judicial branches of government.

The foremost value of press freedom comes from the premise that a press free from government control is essential to a political system based on the will of the people. The 'checking value' of a free press posits that the press should be guaranteed freedom to report

abuses of power by public officials. This compels a logical conclusion that the government should not prohibit the people from receiving news critical of the government. *New York Times* columnist Anthony Lewis succinctly discusses the relationship between the news media's checking function and the media's quest for information:

> If big government is to be effectively criticized and controlled, it will take more than the exhortations of an orator. It will take the countervailing force of big newspapers and broadcast networks and public-interest groups and lobbying organizations of all kinds. And they cannot succeed without information. In our society information is power. Officials struggle to control it and, in that struggle, the *citizen-critic* needs constitutional support (Lewis 1980).

The libertarian concept of press freedom, most fully developed and most ardently supported in the US, focuses exclusively on negative freedom. Since the 1940s, it has been challenged by views that emphasize 'freedom for' as much as 'freedom from.' Concerned about the inexorably growing influence of the mass media on society, so-called 'social responsibility' theories of the press have emerged as a significant counterpoise to the *laissez-faire* libertarian view. The increasing significance of broadcasting, regulated by government 'in the public interest' from early on, has no doubt contributed to this emphasis.

The criticism of a libertarian concept of press freedom stems from an innovative proposition that freedom of the press carries little meaningful value unless it is supported by an independence of editorial decision-making from both internal and external pressures. In this context, 'journalistic freedom' is distinguished from press freedom. Journalistic freedom is concerned with a relationship between the working journalists of a news medium and the managing executives and editors of the news medium. It is an individualized, not institutional, type of press freedom (Merrill 1989, Baker 1989).

In the US, press freedom is to a large degree guaranteed and practiced institutionally, but journalistic freedom is rarely accepted. But several free-press countries in Europe and Asia try to accommodate the conflict between the institutional press freedom and the internal journalistic freedom. A case in point is South Korea. To ensure separation between the editorial department and management, the *Hankyoreh Shinmun* (One Nation Newspaper) and several Korean newspapers have reporters directly involved in deciding on the overall editorial policy of their newspapers, with managing editors elected by reporters or selected from those recommended by them. The stronger voice of the editorial departments of an increasing number of Korean newspapers has resulted from reactivated labor unions of journalists since 1987, when the Korean government embraced a series of sweeping sociopolitical democratic reforms.

3. *Censorship of the Press*

In connection with press freedom, both institutional and journalistic, censorship is an unending concern to the print and broadcast media around the world. Censorship is 'the systematic control of the content of any communications medium, or of several or all of the media, by means of constitutional, judicial, administrative, financial or purely physical measures imposed directly by, or with the connivance of, the ruling power or a ruling elite. It may or may not be accompanied by violence, it may or may not be total, it may or may not include propaganda. But the concept of a *systematic* imposition of controls is...essential to determining whether or not censorship is being practised in a given country, just as is the concept of intent' (Scammell 1988).

Censorship is intertwined with power so inextricably that it is referred to as the 'handmaiden of power.' Censorship in the sense of prior restraint on the press is more effective than subsequent punishment because of its broader scope and lack of countervailing mechanism such as public appraisal of its rationale. And it has a propensity to decide against publication. 'The function of the censor is to censor. He has a professional interest in finding things to suppress' (Emerson 1955).

Censorship can be 'total' when all the media are systematically controlled, 'partial or selective' when the censorship is selectively enforced, or 'random' when it is capricious or arbitrary. Further, the methods of censorship are multifarious, ranging from prior restraint to fines and prison sentences for violations to suspensions of a publication to closing down a publication to the many varieties of self-censorship journalists exercise to stay out of trouble (Scammell 1988).

Does freedom of the press start to flourish when control of the media through state censorship ends? It is not as simple as that. An enterprising press is not the rule but the exception in societies in transition to democracy. Where journalists have been accustomed to being told what to report and what not to report, they have been trained and tamed by censorship and do not necessarily take advantage of their new-found freedom easily (Williams 2000).

3.1 *Censorship Checklist*

Censorship is a crucial criterion against which press freedom is measured. The World Press Freedom Center (WPFC) in the US asks 20 questions to determine whether the news media in a particular country are free:

(a) Are there restrictive press laws?

(b) Who owns the media, print, and broadcast? Private, government, or both?

(c) Are journalists prosecuted for what they write?

(d) Are journalists in jail? Reasons.
(e) Is libel a civil or criminal offense?
(f) Are journalists required to have a government enforced license to work?
(g) Do journalists have unrestrictive access to government proceedings?
(h) Are journalists harassed while covering the news?
(i) Does the government pay journalists?
(j) Are newspapers or broadcasters subsidized?
(k) Is there government-supported censorship? Is there self-censorship?
(l) Are there restrictions on the means of production, such as government allocations of paper, control of distribution systems, and ownership of printing facilities?
(m) Is government advertising allocated fairly?
(n) Is there a legally mandated right of reply, which includes government officials?
(o) Are insult laws routinely used to shield officials' conduct from public scrutiny?
(p) Are courts able to judge news media cases independently?
(q) To what extend are media outlets owned by political, government-linked entities, or others desiring to control content?
(r) Are crimes against journalists prosecuted by authorities?
(s) Are the activities of government—courts, legislature, officials, records—open to the press?
(t) Do journalists themselves consider themselves free to write or broadcast the news as they find it? (Greene 2000).

3.1.1 Right of reply. The WPFC checklist reflects a US-centric approach to evaluating press freedom against the exercise of censorship in that it emphasizes the negative-freedom view on the press rather than positive freedom. The right of reply is a good illustration. In the US, the mandatory right of reply statutes governing the news media as a whole and the print press in particular are seen as violations of freedom of the press.

The US Supreme Court in 1974 struck down a right of reply statute on the ground that it would infringe upon the editorial rights of the press: 'A newspaper is more than a passive receptacle or conduit for news, comment, and advertising. The choice of material to go into a newspaper, and the decisions made as to limitations on the size and content of the paper, and treatment of public issues and public officials—whether fair or unfair—constitute the exercise of editorial control and judgment. It has yet to be demonstrated how governmental regulation of this crucial process can be exercised consistent with First Amendment guarantees of a free press as they have evolved to this time' (Miami Herald Publishing Co. v. Tornillo 1974)

At least 13 countries in Europe, including Germany and France, however, recognize the right of reply. The right of reply in German law is derived from French law. The policy justifications for the right of reply in Germany are not so much to provide the public with access to the media as to protect individuals from false defamation. This differentiates German law from the South Korean law on the right of reply. The right of reply in Korea is justified partly because it can contribute to the public's discovering truth and forming diverse public opinions by creating a more open marketplace of ideas by enhancing the objectivity of news information. In this vein, the right of reply offers access for the Korean public to the news media in a positive way. Especially noteworthy is the international acceptance of the right of reply. The International Convention on the Right of Correction and the American Convention on Human Rights provide the right of reply.

The explicit recognition of the right of reply in several civil-law countries in Europe and Asia and under international law signifies that there are alternative visions of press freedom, which stand in sharp contrast with the US approach to the negative freedom of the press, that is, access to the press is not entirely antithetical to freedom of the press, but it can be viewed as a plausible 'positive' notion of press freedom in ensuring more opportunity for freedom of expression.

3.1.2 Press subsidies. Do government subsidies for the press create a cause of concern for censorship? Evidence from the Scandinavian press subsidies disproves an assumption underlying the adversarial press–state relations in America that newspapers receiving direct public subsidies will lose their editorial independence and autonomy to the government. The Swedish subsidy program 'has had little effect on the willingness of subsidized publications to criticize government' (Gothberg 1983, Gustafsson 1980)

3.1.3 Licensing of journalists. Licensing of journalists is incompatible with freedom of the press as a basic human right. As the Inter-American Court of Human Rights (IACHR) ruled in 1985: 'Such a law (on licensing of journalists) would contain restrictions to freedom of expression...and would consequently be in violation of not only the right of each individual to seek and impart information and ideas through any means of his choice, but also the right of the public at large to receive information without any interference' (IACHR 1985).

3.1.4 Administrative vs. judicial censorship. The legality and scope of press censorship is sometimes

examined by focusing on who primarily dictates the censorship mechanism. Is censorship entirely in the hands of executive officers with no judicial oversight? On this issue, the US differs from Japan. The US law does not allow even court-issued prior restraint under the First Amendment to the US Constitution. In Japan, however, the anticensorship principle does not extend to judicial restrictions because censorship is limited to administrative imposition of prior restraint on the press.

4. *Self-censorship of the Press*

Self-censorship is viewed as a widespread and insidious problem among reporters and editors. Journalists resort to self-censorship when they decide to avoid legitimate stories not because of their professional judgments but on nonprofessional grounds. In some cases, self-censorship can be intimately intertwined with state censorship. This is the case in Israel: 'When acting as concerned members of society, editors and journalists are of one mind with the censors concerning the prohibition of public exposure of issues considered crucial to the physical welfare of Israel or to its ideological *raison d'être* (even though the public may not share that view)' (Liebes 1997).

Intrusion of commercial interests into editorial decision-making is another major cause of self-censorship. 'Market forces are seen as the primary reason why worthwhile stories are not pursued, and this factor is especially prevalent in the broadcast arena,' reports a recent survey of American journalists on self-censorship (Kohut 2000). Self-censorship is related closely to threat of lawsuit for defamation, invasion of privacy, breach of confidentiality agreements, and related issues.

4.1 *Libel Law*

Libel was one of the earliest legal actions available against the press, and it is still the most common legal danger to the media. Libel is an occupational hazard for media professionals. Seventy percent of all libel actions are filed against the mass media; at least two-thirds of those actions are brought against newspapers in the US. Libel law protects the reputational interests of individuals because it involves the expression of opinions about them by others. It also attempts to compensate for economic and emotional injury and promotes human dignity by providing a civilized forum in which a dispute is settled. Most importantly, however, libel law acts as a deterrent on the publication of false and injurious speech through the award of damages. In addition, libel law serves a vital social interest by providing a check on media power. It opens the newsgathering and decision-making process of the media to public scrutiny and accountability.

4.1.1 Transnational libel litigation. The globalizing media confront a new type of libel threats as offshore litigation against US and other transnational news media is growing. One much-publicized case involved Ariel Sharon, the former Defense Minister of Israel, who sued against Time, Inc. for a defamatory article published in *Time* magazine. Sharon filed his lawsuits in a US federal court and in an Israeli court in 1983.

Most recently, former Indonesian President Suharto filed a $27 billion lawsuit against the US-based *Time* magazine in a Jakarta court. Suharto claimed that *Time* published an article alleging that he had transferred billions of dollars from Switzerland to an account in Austria in an effort to stash his multibillion-dollar fortune abroad. The Indonesian court ruled in November 1999 that it had the authority to try Suharto's civil libel suit against *Time*.

Suing American media in foreign courts is one thing; enforcing foreign court judgments in the US is another. Thus far, American courts have consistently refused to recognize and enforce foreign libel judgments, holding that the judgments were based on libel law repugnant to the US Constitution on press freedom.

4.1.2 Internet libel. The 'information superhighway' epitomized by the Internet exerts an enormous impact on libel law all over the world. This is especially the case with jurisdiction and enforcement of libel law. While other countries are still in the process of sorting out a vast array of jurisdictional and other issues, the US Congress in 1996 passed a law immunizing Internet service providers (ISPs) from liability for third-party postings of libelous material. One leading English libel lawyer suggested that the US law would pressure England and other European countries to adopt similar standards for Internet libel 'to reflect the transnational nature of such publications' (Hooper 2000).

5. *Preventive Counseling for News Media*

Because freedom of the press is litigated in increasing frequency in the US, England, and other countries, media organizations involve lawyers in prebroadcasting and prepublication reviews. Lawyers in newsrooms have now become a fact of life for many, if not all, reporters and editors in dealing with a host of legal problems ranging from libel to access to in-

formation. Some journalists are concerned about the chilling effect of lawyers' presence in and outside the American newsroom. CBS News anchor Dan Rather claimed: 'I'd rather take my chances in front of a jury than take my chances in front of corporate lawyers' (Grossman 1996). Nonetheless, the journalist–lawyer relationship is not necessarily harmful to freedom of the press insofar as lawyers are sufficiently 'sensitive to the murky line between legal advice and journalistic or artistic judgment' (Smolla 2000). No less important is the fact that journalists, the integral part of the media system, are more often responsible than lawyers for determining what to broadcast or publish or what not to broadcast or publish in the US. Likewise, a 1997 study of the impact of libel law on the media both in England and in Scotland concluded, '(A) solicitor's *advice* is just that, and ... on occasion editors do decide to ignore it' (Barendt et al. 1997).

6. *Access to Information and Free Press vs. Fair Trial*

The press's role of serving as a 'public watchdog' on the government and of imparting information on matters of public interest is difficult, if not impossible, to fulfill without access to public records. Without some protection for seeking out the news, freedom of the press could be a farce. Now freedom of information (FOI) is accepted more widely around the world than ever. This is a profound development in freedom of the press as a positive concept. In this connection, it should come as no surprise that the Scandinavian countries that embrace a 'proactive' concept of press freedom has the most comprehensive FOI laws. For example, Sweden, whose government information statute is considered the oldest in the world, has adopted the policy on access to official records as a fundamental foundation of its democracy since 1766. The presumption of openness in government documents in Sweden is unambiguous when the Swedish FOI act states that all official records are public if there is no specific statute restricting the access to the records.

Traditionally, freedom of the press has been confined to the right of news reporters and editors to disseminate any information in their possession, which does not necessarily extend to the legal right for the media to gather information. But newsgathering has now emerged as a vital issue for most journalists and news media organizations. In the US, how to deal with access to government records, meetings, institutions, courtrooms, and the like constitutes by far the largest body of mass media law.

Freedom of the press is often in conflict with the accused's right to a fair trial in that court proceedings and judicial records are an important source of news information, and trials are government proceedings for the public to know. Some countries such as Australia, Sweden, and the US favor open proceedings. Other countries including the UK are more restrictive of the press coverage of trial proceedings.

7. *Journalist's Privilege to Shield Confidential Sources*

Equally important is the issue of whether journalists should be privileged to protect their sources' confidentiality from forced disclosure. In the US, a journalist's privilege is qualified and lags far behind Australia, France, Germany, and Sweden, where the strongest legal protection is provided to the confidentiality of sources. In the European countries, the protection of confidential sources is premised on the conclusion that 'society is better served by encouraging people to disclose matters of public interest to the press, even when the information may include evidence of their own wrongdoing, than by identifying and possibly convicting a particular wrongdoer,' and 'journalists rarely divulge information that would compromise sources, courts rarely compel disclosure, and even more rarely do courts enforce a disclosure order through imprisonment' (Coliver 1993).

8. *Freedom of the US Press in the Globalizing Era*

Freedom of speech and the press has been the international buzzword since the end of the Cold War in the late 1980s. And the positive concept of press freedom has crystallized the undisputed paradigm throughout the world and it inspires the standard for quality journalism. The growing impact of the now modified libertarian press theory, with an emphasis on the affirmative role of the state in expanding press freedom, on many newly independent countries and emerging democracies in Asia, former eastern Europe, and Latin Africa is credited partly to the American successful experiment with a free press over the past 200 years. Professor Owen M. Fiss of Yale Law School offers a cogent note: 'In building a free press, the reformers [in formerly Soviet bloc nations] should look to the American experience, but only selectively. They must create for the press a measure of autonomy from the state without delivering the press totally and completely to the vicissitudes of the market' (Fiss 1996).

See also: Censorship and Secrecy: Legal Perspectives; Censorship in Research and Scholarship; Control: Social; Cultural Expression and Action; Freedom/Liberty: Impact on the Social Sciences; Freedom: Political; Media and History: Cultural Concerns; Media, Uses of; Political Communication; Political Discourse

Bibliography

Baker C E 1989 *Human Liberty and Freedom of Speech*. Oxford University Press, New York

Barendt E, Lustgarten L, Norrie K, Stephenson H 1997 *Libel and the Media*. Clarendon Press, Oxford, UK

Coliver S 1993 Comparative analysis of press law in European and other democracies. In: Coliver S (ed.) *Press Law and Practice*. Article 19, London

Emerson T I 1955 The doctrine of prior restraint. *Law & Contemporary Problems* **20**: 648–71

Emerson T I 1966 *Toward a General Theory of the First Amendment*. Random House, New York

Fiss O M 1996 *Liberalism Divided*. Westview Press, Boulder, CO

Gothberg J A 1983 Newspaper subsidies in Sweden pose no dangers. *Journalism Quarterly* **60**: 629–34

Greene M J 2000 *New Code Words for Censorship*. World Press Freedom Committee, Reston, VA

Grossman L 1996 Lessons of the 60-minute Cave-In. *Columbia Journalism Review* **Jan/Feb**: 39–51

Gustafsson K 1980 The press subsidies of Sweden: A decade of experiment. In: Smith A (ed.) *Newspapers and Democracy*. MIT Press, Cambridge, MA

Hooper D 2000 *Reputations under Fire*. Little, Brown, London

Inter-American Court of Human Rights (IACHR), San Jose, Advisory Opinion of November 13, 1985, No. OC-5/85—Compulsory membership in an association prescribed by law for the practice of journalism (Arts. 13 and 29 American Convention on Human rights)

Kohut A 2000 Self-censorship: Counting the ways. *Columbia Journalism Review* **May/June**: 42

Lahav P 1985 Conclusion: An outline for a general theory of press law in a democracy. In: Lahav P (ed.) *Press Law in Modern Democracies*. Longman, New York

Lewis A 1980 A public right to know about public institutions: The first amendment as a sword. *Supreme Court Review*, pp. 1–25

Liebes T 1997 *Reporting the Arab–Israeli Conflict*. Routledge, New York

Lovell v. Griffith, 1938 303 U.S. 444

Miami Herald Publishing Co. v. Tornillo 1974 418 U.S. 241

Merrill J C 1989 *The Dialectic in Journalism*. Louisiana State University Press, Baton Rouge, LA

Scammell M 1988 Censorship and its history: A personal view. In: Boyle K (ed.) *Article 19 World Report 1988*. Times Books, New York

Smolla R 1999 *Law of Defamation*, 2nd edn. West Group, St. Paul, MN

Williams L 2000 Censors at work, censors out of work. In: Williams L, Rich R (eds.) *Losing Control: Freedom of the Press in Asia*. Asia Pacific Press, Canberra, Australia

K. H. Youm

Freedom: Political

1. Introduction: Liberty, Equality, Fraternity; Two Concepts of Liberty

Since the bourgeois revolutions, modern states have been committed, at least formally, to the ideals of liberty, equality, and fraternity. From the start there have been tensions between those three aspirations. Attempts to resolve these tensions have taken different forms, from rampant *laissez-faire* libertarianism to interventionist, redistributivist socialism. The theories examined in this article focus mainly on the relation between liberty and equality, with an occasional nod towards fraternity. Since political freedom or liberty is an essentially contested concept, it is impossible to start from a shared definition of it. One can only trace the sources of a number of incompatible conceptions and establish an elementary taxonomy of them.

An important moment in the modern debate about political freedom is Berlin's *Two Concepts of Liberty* (1958). It is an unsystematic mixture of different elements: conceptual analysis, political engagement in the postwar antitotalitarian struggle, and a rich, if eclectic, essay in intellectual history. But for all its shortcomings, it made the opposition between positive liberty and negative liberty central to much subsequent discussion.

Berlin's political task was to unmask the perversion of language by Stalinist ideologues, who presented the worst forms of oppression as 'real (socialist)' freedom, as opposed to 'mere bourgeois (formal)' freedoms. Berlin aligned the former with 'positive,' and the latter with 'negative,' liberty, and reasserted the value of the latter against its critics.

Historically, Berlin identified negative liberty with liberalism, the tradition that defended the individual against 'interference,' particularly when exercised by the state. But the idea of 'interference' is hard to tie down. Berlin himself was ambivalent as to whether the interference must be deliberate or not, but he was adamant that an inability caused by the social structure (e.g., the inability of the worker to control the means of production in traditional Marxist theory) is not a case of unfreedom. He agreed that the 'classical English political theorists' were 'prepared to curtail freedom in the interests of other values and indeed of freedom itself.' They held that without law, 'freedom for the pike is death for the minnow.' A natural interpretation of that insight is that the law positively creates freedom for the minnow, allowing it to do things it would otherwise be unable to do. In the words of Locke (a quintessential negative libertarian for Berlin): 'the end of Law is not to abolish or restrain, but to preserve and enlarge freedom.' Berlin could agree that the pike's freedom is restricted for the sake of the minnow's freedom ('curtailed in the interests of liberty itself'), but his polemical drive was against partisans of positive liberty who would maintain that the pike's freedom too is (really) enhanced by the restriction.

Another ambivalence within negative liberty concerns the connexion between freedom and wishes. Berlin began by tying the two together directly, defining oppression as 'the part I believe to be played by other human beings … in frustrating my wishes.' But later he rejected the definition of freedom as 'the

ability to do what one wants' because it is compatible with a situation in which all one's wants are restricted by brainwashing or the removal of desirable objects. At the heart of negative liberty lies 'the right to be let alone,' but that right turns out to be remarkably elusive.

If negative liberty (NL) is hard to tie down, positive liberty (PL) is even more so. PL contains both (a) political elements (PL-pol) and (b) moral ones (PL-mor). (a) is essentially the freedom to participate in the political process; (b) includes the ability to be the autonomous master of one's own drives, which is associated with the romantic notion of the divided self. Rousseau says that both (a) and (b) are the outcome of entering civil society, but he is careful not to identify them. When Constant (1988) introduced the distinction between the 'liberty of the ancients' and the 'liberty of the moderns,' he was concerned with (a). His polemical claim was that the 'liberty of the ancients' nowadays meant the unbridled exercise of popular sovereignty (PL-pol) to the detriment of NL. For Berlin, the danger of modern totalitarian ideology is that it harnesses (a) to (b). It suggests that, through laws endorsed by 'the people,' irrational empirical selves should be forced to coincide with rational ends (their 'real selves').

Five broad approaches to political freedom can now be distinguished. In each of these, the three fundamental values of liberty, equality, and fraternity are addressed. Writers inspired by Berlin focus on incompatibilities between those values. Others seek to show that when they are more clearly understood, it will be seen either that they are compatible or that the very idea of balancing one against the other must be abandoned. In their analyses, the terminology of PL and NL frequently recurs, but the distinction made above between dimensions (a) and (b) of PL frequently is ignored.

The approaches to be considered are: (a) Welfare liberalism; (b) Libertarianism; (c) Freedom, economics and power; (d) Freedom and community; (e) Civic republicanism.

1.1 Welfare Liberalism

For theorists of welfare liberalism, political freedom is exercised in a context of representative democracy constrained by constitutional checks to protect individual rights and by a fluctuating sense of obligation to redress the inequalities thrown up by the operation of the market. The two outstanding proponents of this approach are John Rawls and Ronald Dworkin.

1.1.1 Rawls. Rawls's *A Theory of Justice* (1971) is the most comprehensive, systematic work of modern political philosophy. In it the author expounds the moral basis of pluralist welfarism, combining the three elements of liberty, equality, and fraternity into a coherent whole.

At the heart of the book lie the Two Principles of Justice. The First Principle states that 'Each person is to have an equal right to the most extensive total system of basic liberties compatible with a similar system of liberty for all.' 'Basic liberties' are specified as 'roughly speaking, political liberty (the right to vote and to be eligible for public office) together with freedom of speech and assembly; liberty of conscience and freedom of thought; freedom of the person along with the right to hold (personal) property; and freedom from arbitrary arrest and seizure as defined by the concept of the rule of law.' These are what Marxists called 'bourgeois liberties.' The list is not equivalent to NL-pol, since they include certain 'positive' political elements, even if the 'negative' liberties of privacy may be more important than 'positive' ones to modern citizens.

In subsequent work, Rawls (1993) has responded to Hart's criticism that the idea of 'maximizing freedom' is incoherent. He has rewritten the First Principle to read: 'Each person has a claim to a fully adequate scheme of basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value.' He has also added the 'integrity' (physical and psychological security) of the person to the list of basic liberties to be protected by the First Principle. But he has not modified the central claims of *A Theory of Justice* that the modern liberal state must (a) guarantee 'basic liberties,' and (b) strive to maximize the welfare of the worst-off members of society by redistributive mechanisms, and that the First Principle must not be sacrificed to the Second, since 'liberty can be restricted only for the sake of liberty.'

Most of the other important revisions have concerned structure and methodology rather than content. Rawls has anchored the Two Principles more explicitly in a conception of citizens of a modern pluralist state, as it has developed since the Reformation. They are 'regarded as free and equal persons,' endowed with 'two moral powers … namely a capacity for a sense of justice and a capacity for a conception of the good.' Such citizens display their freedom in the following ways. They have 'the moral power to have a conception of the good' and are 'capable of revising and changing this conception.' They 'regard themselves as self-authenticating sources of moral claims.' They are capable of taking responsibility for their 'ends.'

The Two Principles of justice remain unaltered and are still subject to attack from both left and right. Left-wing critics charge that First Principle 'bourgeois' liberties are hollow and worthless for the mass of the population. To them Rawls had already responded in *A Theory of Justice* by distinguishing equal liberty

from equal worth of liberty, acknowledging that it is the latter that is important, and predicting, perhaps optimistically, that it would be enhanced by the application of the Difference Principle. Right-wing 'libertarian' critics reject all redistributive mechanisms, because they infringe the individual's unrestricted right to accumulate property. Rawls' response to them is also contained in *A Theory of Justice*, namely that individuals have a basic right to own *personal* property, but that right does not entail any particular property system, whether capitalist or socialist. In each particular case the choice of system will be determined by its ability to maximize the welfare of the worst-off.

1.1.2 Dworkin. Dworkin's political philosophy is an attempt to 'fine-tune' both the methodology and the political recommendations of Rawls's welfare liberalism. Methodologically, Dworkin denies that liberalism must balance liberty against equality. That image wrongly presupposes that liberty can be quantified, so that 'fundamental' liberties weigh more heavily than trivial ones. Instead, 'the fundamental liberties are important because we value something else that they protect.' And that 'something else' is, unexpectedly, a form of equality, the principle that a government should treat all its citizens with 'equal concern and respect.' The 'fundamental' liberties, then, are entailed by the equality principle itself, according to which 'the government (should) treat all its citizens as free, or as independent, or with equal dignity.' Dworkin accepts the Rawlsian requirement that 'political decisions must be … independent of any particular conception of the good life …' Thus governments are bound by the negative requirement to be neutral between lifestyles, and also by the positive requirement to 'respect' the 'dignity and independence' of its citizens.

In terms of political recommendations, Dworkin proceeds from a tension at the heart of Rawls's Two Principles. While the First Principle defends the conditions of basic freedoms, and addresses people as autonomous agents, the Second Principle might undermine people's motivation to exercise their freedom to 'better themselves,' addressing them as passive recipients of benefits. In response to this tension, Dworkin introduces a key modification. He follows Rawls in compensating people for 'irrelevant differences, like differences in talent.' But he argues that people should be 'responsible for the costs of their choices.' The system, in short, must be both 'ambition-sensitive' and 'endowment-insensitive.' But Dworkin's modification may have introduced a new tension of its own, since however finely tuned it is, its two requirements must work against one another: 'the more we try to make the distribution sensitive to people's ambitions, the more likely it is that some people disadvantaged by circumstances will be undeservedly penalized, and vice versa' (Guest 1997). Everything now hinges on the possibility of making a realistic distinction between ambition and endowment, between 'personality' and 'circumstances.' The fact that this is so difficult suggests that the tension between freedom and equality may still be unresolved.

2. *Libertarianism, Freedom as the Unrestricted Right to Own Private Property: Nozick*

Nozick's *Anarchy, State and Utopia* (1974) was a polemical response to Rawls' *A Theory of Justice*. At the time of its publication, there was a backlash against welfarism and state intervention in the economy. That backlash was driven by the economic theory that managed markets are inefficient, since redistribution through taxation takes away incentives. But although Nozick's book was congenial to that theory, his philosophical position is independent of it. Nozick would defend his position on the grounds that it was the correct one, even if it were inefficient. He would in fact reject appeals to economic efficiency as improperly utilitarian.

Though labeled a 'libertarian,' Nozick's starting point is not the idea of 'liberty' or 'freedom,' but rather the Lockean one of 'self-ownership.' This has polemical value because it allows Nozick to proceed against Rawls' (and other welfarists') programs of redistribution by appealing to a hatred of slavery shared by all across the political spectrum since the 'bourgeois revolutions,' and suggesting, by a rhetorical sleight of hand, that 'taxation of earnings from labor is on a par with forced labor.' The slide has some source in Locke's 'labour theory of property,' the idea that you have a property right to something if you have 'mixed your labour with it.' But Locke did not suggest that that right is absolute in the same sense that your right to your bodily integrity is.

The sleight of hand at the heart of libertarianism is this. We all agree that if A has the right to make use of B's body, of which B is the rightful owner, without B's consent, then B is A's slave. By parity of reason, it is suggested, if A has the right to make use of B's property, of which B is the rightful owner, without B's consent, by taxing B, then B is A's slave. The sleight of hand depends on accepting the extension of self-ownership from ownership of one's own body to ownership of whatever one has acquired legally, an extension which is not evidently justified. Nonetheless, libertarians may be able to deploy a 'slippery slope' argument against welfarism without making this particular sleight of hand. Rawls's Two Principles of justice are intended 'to balance the twin demands of (a) respecting choices and (b) rectifying circumstances.' Rawls intends (a) to be applied to individuals, (b) to groups. But, according to the libertarian, there is

nothing within Rawls's theory to prevent the radical welfarist from extending rectificatory treatment to individuals too. How is one to define what intervention is legitimate in order to rectify circumstances? If the line is to be drawn around the body, it would seem to mean that any nonbodily interference is at least in principle legitimate in order to rectify circumstances. And that sounds sinister even to those who are not extreme 'libertarians.'

3. *Freedom, Economics, and Power*

In the past, egalitarians, Marxists and non-Marxists alike, derided 'bourgeois freedoms' for masking from the workers their real enslavement. The modern egalitarians considered in this section all take seriously the demand for real equality of freedom. But unlike their predecessors, they all also take seriously the importance of the 'basic' freedoms that lie at the heart of liberalism. The three positions considered in this section are linked by the insight that freedom and equality are linked intimately, and that the conditions of equality and inequality are zero-sum relations of economic power. In each of the three, real freedom is 'positive' in the sense that it requires more than just the 'right to be let alone.' On the other hand, it is not the 'liberty of the ancients,' since it recognizes the importance of private life in the modern world.

3.1 *Neo-Marxism: Cohen*

In the most sophisticated version of neo-Marxism, that of Cohen, 'there is … a conflict between social equality and the liberty of some people,' but that should not stand in the way of 'the pursuit of social equality, since a humane concern for liberty must first of all direct itself to the condition of those who enjoy hardly any of it.'

Neo-Marxists thus agree with liberals that freedom of the individual is the most important political value, and that modern capitalism delivers it to all members of society, including proletarians, in several important ways. First, all members of a capitalist society (worker and capitalist alike) are legally free. Second, any particular worker is always legally free, and often economically free, to leave the proletariat and become a petty bourgeois or even a capitalist. But, within a capitalist society, 'although most proletarians are free to escape the proletariat, indeed even if all are, the proletariat is an imprisoned class.' Third, capitalism has delivered 'important freedoms beyond that of buying and selling.' These include 'freedom of speech, assembly, worship, publication, movement, political participation … .' The neo-Marxist Cohen is agnostic about 'how accidental the connexion between capitalism and those freedoms has been and is,' but is committed to the view that only 'freedom to buy and sell belongs to capitalism's inmost nature.' But Marxists must still acknowledge that 'bourgeois freedoms' really are freedoms: '… when socialists suggest that there is no real liberty under capitalism, or that socialism promises liberty of a higher and unprecedented kind … their line is theoretically incorrect and politically disastrous. For liberty under capitalism is, where it exists, just that, liberty; and if socialism will not give us plenty of it, we shall rightly be disappointed.'

In recent writings, Cohen (1995) has begun to investigate the idea of real freedom, understood as 'autonomy, the circumstances of genuine control over one's own life.' With this turn, neo-Marxism returns to the old site of conflict, since we must 'ask what kind and degree of control over external things a person must have to enjoy autonomy, and then to ask whether such control is compatible with socialist equality.' Cohen's project now is to translate the idea of real freedom into terms that are applicable to the real world, and, where possible, quantifiable. His model of 'equal access to advantage' now stands alongside van Parijs's (1995) and Sen's (1992) models. It thus forms part of a rich developing research program. Whether it should still be called Marxist is another question.

3.2 *Post-Marxism: van Parijs*

In the post-Marxist program of van Parijs, the normative goal of equalizing access to freedom is maintained. At the same time, the Marxian *theory* of exploitation is abandoned, and the capitalist market is now seen as the only vehicle that can deliver 'real freedom for all' in *practice*. The goal is to 'empower' the greatest number of people. The method is to guarantee them an unconditional 'basic income' that will prevent personal dependence and allow individuals to make meaningful choices between life-styles. The 'basic income' project has now entered the real political agenda in western Europe.

The goal of the post-Marxists is to reconcile freedom and equality by the principle to 'maximization of an unconditional income, subject to formal freedom being protected and to no-one's endowment being unanimously found worse than someone else's.' That would guarantee the widest, and fairest, distribution of 'real' freedom that could also be delivered by a 'sustainable' economic system. Van Parijs maintains that that system would be a highly regulated, interventionist form of capitalism. His reasons are fourfold. First, it is (probably) more efficient than socialism. Second, it can (probably) guarantee a higher basic income than socialism can. Third, it can (certainly) guarantee formal freedoms, whereas socialist régimes have still to prove that they can. Last, and this is a decisive consideration, even if there are reforms of international economic affairs, it will be impossible to promote an 'autarkic' socialist régime in one country

without imposing unacceptable restrictions on the movement of its people to other (higher paying) countries.

3.3 Neo-Aristotelian Freedom as Capability: Sen

The third egalitarian model combines sophisticated economic theory with a deep philosophical anthropology derived from Aristotle. The leading proponent of this model is Sen.

The aim here is to construct a model of freedom which is sufficiently objective to be used in cross-cultural comparisons, but which is also sensitive to the different weightings given to the elements of a flourishing life in different cultures. Sen (1992) draws up a list of 'functionings' that include 'being adequately nourished' and 'having self-respect,' and defines 'well-being freedom' as the capability to achieve those functionings. But there is also 'agency-freedom,' defined as the capability to achieve whatever one's objectives are. Since many people have objectives other than their own well being, these two freedoms may conflict.

It is agency freedom that has been at the center of most liberal discussion. Sen acknowledges its importance, but weighs it alongside well-being freedom in evaluating the totality of a person's life. Even within agency freedom, Sen advises liberals not to exaggerate the importance of the agent's direct control of an outcome. Instead, he proposes a counterfactual formula, 'I am free when control is exercised in line with what I would choose.' That formula can be generalized to cover all forms of freedom, even when the term 'freedom' is extended to cover 'freedom from hunger,' 'freedom from malaria,' etc.: 'One values living without malaria, desires such a life, and would have chosen it, *given the choice*.'

Society has different obligations to its members concerning their well being and their agency. It has an evident 'responsibility to a person's well-being, especially when it is in some danger of being particularly low' which it does not have for that person's particular agency objectives. In advanced liberal societies at least,'in which the focus is on the *freedom* that people have,' no patterned outcome can be imposed. For 'even when complete equality of well-being freedom is … entirely realized, this need not lead to an equality of well-being *achieved*, since different individuals may give different priority to the pursuit of their own well-being.' In rejecting patterned outcomes, Sen has shown that the capability model does not conflict with the central values of liberalism expounded by Rawls and Dworkin.

The capability model inherits from Aristotle the idea that human beings are endowed with specific capacities that can be more or less fully realized in a flourishing life. In his presentation of the model, Sen is concerned that Aristotle's particular 'view of human nature (with a unique list of functionings for a good human life) may be tremendously overspecified.' He prefers to leave open 'both … the exact grounds underlying the determination of relative weights, and the actual relative weights chosen.' This gives Sen's version of the 'capability approach' a degree of incompleteness relative to Aristotle's, but also, he thinks, a 'considerable "cutting power"' for use in the real world of cultural diversity.

3.4 Freedom as Attaining Autonomy Through Community: Raz

Communitarian critics including Taylor, MacIntyre, and Sandel brought two charges against liberalism. At the level of theory, they claimed that it is based on a methodological individualism which in turn derives from an impoverished, atomistic conception of human nature. At the level of practice, they argued that because the liberal order fails to provide a satisfactory sense of identity for its members, 'intolerance flourishes most where forms of life are dislocated, roots unsettled, traditions undone.'

Along with the communitarians, Raz (1986) rejects methodological individualism, but against them he argues that this does not entail a rejection of liberal values. Raz's liberalism, 'the doctrine of limited government,' is rooted in a 'non-individualist conception of morality.' In opposition to Rawls's antifoundationalism, Raz propounds a 'doctrine of political authority which rests on a perfectionist political defence and promotion of liberty and autonomy.' But this is not a brutish libertarianism, since freedom is 'a distinct value, but one which is intimately intertwined with others, and cannot exist by itself.' Raz in short is a foundationalist, who seeks to establish a 'continuity' between a moral base and a political–legal superstructure.

For someone to attain 'the ideal of autonomy' and live an autonomous life, certain conditions must be satisfied. These are the possession of 'appropriate mental abilities,' of an 'adequate range of options,' and of 'independence.' The ideal is positive, comprehensive. It constitutes the moral base of the system, but is not 'moralistic.' In particular, it does not involve a teleological conception of self-realization, since, as an autonomous person, one is free to choose whether or not to realize oneself. But without teleology it is difficult to make autonomy foundational. If autonomy involves integrity ('identification with one's life'), self-awareness and commitment to a project, it can still be asked why it is valuable. It is insufficient to answer that 'many people desire to be autonomous,' since people's desires are many, conflicting, and mutable. Rather, says Raz,'the value of autonomy does not depend on choice … the conditions of autonomy concern a central aspect of a whole system of values of a society …' The kind of society in question

is a pluralistic society, which both supports and is supported by this central value 'valuing autonomy leads to the endorsement of moral pluralism.'

Other modern liberals, notably Rawls, are cautious about seeking to found political legitimacy on a particular understanding of freedom. They hold that particular individuals and groups can value freedom for different reasons, some welcoming it because it is a purely negative guarantee against interference, others because it embodies their highest aspirations. Against them Raz holds that most people in a liberal state must subscribe to the positive value of freedom as autonomy, otherwise its tolerant pluralism will perish.

4. *Civic Republicanism*

In the fifth model of political freedom, the 'liberty of the ancients' re-enters the modern debate. In both versions of 'civic republicanism' considered below, political participation plays a key role, a purely instrumental role in the first, and a constitutive role in the second.

4.1 The Instrumental Notion of Civic Republicanism: Skinner

Skinner's model, known variously as 'civic republicanism,' 'classical republicanism,' and 'neo-romanism,' is rooted in the Renaissance, in particular in the idea derived from Machiavelli that a state which is 'free from all external slavery' guarantees 'personal liberty, understood in the ordinary sense … each citizen remains free from any elements of constraint, especially those which arise from personal dependence and servitude, free to pursue his own chosen ends … To be free … is simply to be unconstrained from pursuing whatever goals we may happen to set ourselves.' The independence of the state is linked to the freedom of the individual insofar as 'a self-governing republic is the only type of régime under which a community can hope to attain greatness at the same time as guaranteeing its citizens their individual liberty.' Such a state requires two capacities or virtues from it citizens, courage and determination to defend their community against external attack, and practical reason which qualifies them to play an active part in public life. That means that the state can, legitimately, force the citizen to be free. Despite appearances, this 'strenuous view of citizenship' is, according to Skinner (1990), consistent with negative liberty in two respects. First, it does not presuppose any specific teleological theory of human nature. Second, when we are 'forced to be free,' we are being forced not to 'reason about ends,' but to 'recognize the range of actions we have good reason to perform in order to bring about the ends we actually desire.' Skinner illustrates the latter point by a contrast between 'liberals' like Hobbes and Locke on the one hand, and Machiavelli on the other. The 'liberals' hold only (a) that the law preserves A's freedom by coercing B (the potential law-breaker). Machiavelli holds both (a) and (b) that the law also preserves A's freedom by coercing A 'into acting in a particular way … forcing us out of our habitual patterns of self-interested behaviour … into discharging the full range of our civic duties and thereby ensuring that the free state on which our own liberty depends is itself maintained free of servitude.' In short, because 'freedom depends on service and so on our willingness to cultivate the civic virtues … we may have to be coerced into virtue and thereby constrained into upholding a liberty which, left to ourselves, we would have undermined.' Skinner comments that 'contemporary liberalism … is in danger of sweeping the public arena bare of any concepts save self-interest and individual rights.' Ideologies suspicious of the whole modern agenda of individual freedom have stepped in to fill that vacuum. Skinner (1998) thinks that 'republican freedom' provides a robust defence against such reactionary positions.

4.2 The Constitutive Notion of Civic Republicanism: Pettit

In the second version of civic republicanism, espoused by Pettit (1997), Berlin's dichotomy of the positive freedom of self-mastery and the negative freedom of noninterference is not exhaustive. Beyond those two lies republican freedom, which means absence of domination, not absence of interference. Domination differs from interference in important respects. Domination is a function of the relationship of unequal power between persons or groups of persons. It relates to the *capacity* of one person or group of persons to interfere in the choices of other persons or groups, rather than to actual *instances* of interference. But that would not be enough to pick out the undesirable character of domination, since any non-anarchist social order will empower certain persons or groups, giving them authority over others in given circumstances. So domination is not just the possession of power, but the possession of *arbitrary* power. The goal of republican freedom, then, is that 'no-one is able to interfere on an arbitrary basis—at their pleasure—in the choices of the free person.' With this formula Pettit seeks to extract the rational kernel of both negative and positive freedom, while showing that neither of them is sufficient in itself. On the one hand, freedom as nondomination 'will facilitate the achievement of autonomy,' as positive freedom requires. On the other hand, it will respect people's empirical choices, as negative freedom requires, but it will go beyond it by showing that dominated people are not really free even if they have the good fortune to suffer no interference when their dominators choose to refrain from exercising their arbitrary power.

Insofar as republican freedom is tied to power, it is essentially egalitarian. It is held to protect each

individual against arbitrary power, and also to be a 'communitarian good,' allowing people to identify with a state that protects their freedom. This version of republican freedom is heavily influenced by Rousseau, purged of totalitarian accretions and updated to the advanced capitalist societies of the late twentieth century. They are now explicitly inclusive, bestowing their benefits on all members of society, and also multicultural, displaying liberal neutrality towards different substantive conceptions of the good. How far such societies can provide a stable balance between the participatory core of republican freedom and the centrifugal drives of modern pluralism remains to be seen.

See also: Contractarianism; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Liberalism; Liberalism: Historical Aspects; Libertarianism; Republicanism: Philosophical Aspects; Solidarity: History of the Concept

Bibliography

Berlin I 1958 *Four Essays on Liberty*. Oxford University Press, London

Cohen G A 1995 *Self-ownership, Freedom and Equality*. Cambridge University Press, Cambridge, UK

Constant B 1988 The liberty of the ancients compared with that of the moderns. In: *Political Writings*, (trans. and ed.) Fontana B. Cambridge University Press, Cambridge, UK

Dworkin R M 1977 *Taking Rights Seriously*. Duckworth, London

Dworkin R 1981 What is equality Parts 1 and 2. *Philosophy and Public Affairs* **10**: 185–246, 283–345

Guest R 1997 *Ronald Dworkin*. Edinburgh University Press, Edinburgh, UK

Nozick R 1974 *Anarchy, State, and Utopia*. Blackwell, Oxford, UK

Pettit P 1997 *Republicanism: A Theory of Freedom and Government*. Clarendon Press, Oxford, UK

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Rawls J 1993 *Political Liberalism*. Columbia University Press, New York

Raz J 1986 *The Morality of Freedom*. Clarendon Press, Oxford, UK

Sandel M J 1982 *Liberalism and the Limits of Justice*. Cambridge University Press, Cambridge, UK

Sen A 1992 *Inequality Reexamined*. Clarendon Press, Oxford, UK

Skinner Q 1990 The republican idea of political liberty. In: Bock G, Skinner Q, Viroli M (eds.) *Machiavelli and Republicanism*. Cambridge University Press, Cambridge, UK

Skinner Q 1998 *Liberty before Liberalism*. Cambridge University Press, Cambridge, UK

Taylor C 1979 What's wrong with negative liberty? In: Ryan A (ed.) *Essays in Honour of Isaiah Berlin*. Oxford University Press, Oxford, UK

Van Parijs P 1995 *Real Freedom for All*. Clarendon Press, Oxford, UK

T. O'Hagan

French Revolution, The

The French Revolution was a momentous historical event that set enduring patterns for modern revolutionary movements and for much of modern politics in general. Contemporaries experienced the French Revolution as a set of interlocking changes or stages that seemed driven by some kind of mechanism or impetus. Analysis of this mechanism has preoccupied revolutionaries and scholars ever since. The conflicting political reactions to the French Revolution fed into the development of ideologies from conservatism to communism. Scholarly controversies about the significance and interpretation of the French Revolution have been animated in particular by Karl Marx and Friedrich Engels' use of it as the prime example of a 'bourgeois revolution' that prepared the way for the triumph of capitalism. Although the controversies over the Marxist interpretation continue, recent scholarship has turned in new directions: gender relations; the notion of citizenship and its connection to religion, ethnicity, and race; and the Revolution's impact on the colonies.

1. The French Revolution as Event

The year 1789 marks the clear beginning of the French Revolution, but opinion is much more divided about its ending: did it end in 1794 with the fall of Maximilien Robespierre, in 1799 with the rise of Napoleon Bonaparte, in 1804 when Napoleon established his empire, or in 1815 with the final defeat of Napoleon at the Battle of Waterloo? Debate on the question of endings cannot be resolved definitively because it is so closely entwined with interpretive questions. The choice of ending depends on the determination of the central significance of the French Revolution.

What most struck contemporaries and has fascinated commentators ever since was the unpredictability and yet seeming inexorability of the unfolding of events. Even in the midst of this unfolding, political actors and witnesses spoke of 'the system' of 'the French Revolution,' as if revolutionary events had an internal logic or mechanism pushing in a consistent direction. After spending years studying this logic, and having published his enduringly influential study of the French Revolution, Alexis de Tocqueville still complained of his difficulty in grasping 'this malady of the French Revolution,' this '*virus* of a new and unknown kind' (Tocqueville 1977).

1.1 Revolutionary Stages

As a series of political events, the French Revolution is usually divided into five periods or stages: the constitutional monarchy or liberal revolution of 1789–91; the second revolution and the Terror of 1792–4; the

Directorial regime or liberal republic of 1795–9; Napoleon Bonaparte's consulate of 1799–1804; and the Napoleonic empire of 1804–15. This vertiginous succession of regimes established many of the prototypes of modern political authority and even much of modern political language, including the key distinction between Left and Right (determined by the location of deputies vis-à-vis the speaker in the National Assembly). The very possibility of ascertaining periods or stages—already apparent to contemporaries—raised questions about the natural history or morphology of revolution: was it like a sickness or fever from which a polity eventually recovered? Or was it rather a kind of speeding up of historical development, a bootstrap up to the next stage of capitalism or modernity?

1.2 Models of Political Authority

As the French Revolution moved from stage to stage, the revolutionaries generated at least three different models for modern (secular, nonmonarchical) political authority. By abolishing 'feudalism' (eliminating serfdom and seigneurial dues), suppressing noble titles, declaring the 'rights of man,' and founding the first French republic with universal manhood suffrage (1789–92), the French revolutionaries laid the foundations for democratic, representative government. By arresting dissidents, executing opponents, and carrying out a cultural revolution in the form of de-Christianization and republican political reeducation (the Terror of 1793–4), the revolutionaries set the mold for 'totalitarian' government with its repressive insistence on political conformity in the name of a future social good. Napoleon Bonaparte developed a third model of paternalist, authoritarian government by a state with a strong military caste. In other words, the French Revolution adumbrated virtually all the possibilities of modern political authority.

2. The Origins of Ideology

The French Revolution gave birth both to the term 'ideology' and to many different ideologies that have continued to organize political discussion right up to the present day. Under the impact of revolutionary events, tradition lost its givenness, hierarchy no longer seemed natural, and monarchy no longer went without justification. The revolutionaries showed that society could be reorganized by an act of self-conscious political will (the definition of revolution itself). The attempt to reshape hearts and minds through festivals, paintings, engravings, operas, plays, and even ordinary items such as snuffboxes and playing cards drew dramatic attention to revolution as a mode of action. This attempt did not produce unanimity, however; rather it produced social conflict. As a consequence, revolution brought 'the social' as a category into sharper relief. As early as 1792, Antoine Barnave wrote his *Introduction à la Révolution française*, arguing that new forms of wealth produced ineluctably new forms of political authority (Barnave 1988). As the social determinants of politics became more visible, i.e., more of a problem to be investigated, ideologies took shape as formal doctrines about the proper relationship between politics and society.

2.1 From Conservatism to Communism

The first ideology to emerge from the French Revolution was conservatism. Its founding text, Edmund Burke's *Reflections on the Revolution in France* (1790), was written in response to the events of 1789. The National Assembly's efforts to remake the French constitution led Burke to denounce what he termed a new 'political metaphysics' based on natural rights and abstract reason. Unlike many reactionary royalists, however, Burke did not just excoriate the French Revolution as the incarnation of evil and impiousness. He understood that the revolutionaries were in the process of erecting an entirely new foundation for politics, and he undertook to counter them point by point. By arguing for a conservatism that accommodated change through practical experience and gradual amelioration, he admitted the necessity of confronting demands for change and the parallel demand for an elaborated doctrine to support the status quo. Even for conservatives, then, the French Revolution changed the rules of the game. The status quo now required explicit defense.

Although liberalism had its origins in the social contract theory of John Locke and the political economy of Adam Smith, after 1789 liberals too had to announce their position vis-à-vis the French Revolution. Most liberals supported the constitutional revolution of 1789 and rejected the Terror of 1793–4. This distinction was not always easy to uphold, for conservatives routinely accused liberals of supporting policies and positions that might open the way to more radical experiments, just as those of the revolutionaries of 1789 had opened the way to the Terror of 1793–4. The unfolding of stages in the revolutionary process thus shaped ideological debate, either explicitly or implicitly, for many generations after the French Revolution.

Socialists, and especially Marxists, traced their lineage directly back to the French Revolution. The early socialists—Claude Henri de Saint-Simon and Charles Fourier, for example—concluded from the French Revolution that society could be reorganized according to a technocratic project (thus 'utopian socialism' according to Marx). Marxists believed that only class struggle could produce a new social order. According to Marx and Engels, the French Revolution marked

the overthrow of the aristocracy by the bourgeoisie. The bourgeoisie would develop capitalism on the ruins of the feudal order it dismantled, thus preparing the way for a subsequent revolution in which the proletariat formed by capitalist development would overturn the bourgeoisie and replace capitalism with communism. The French Revolution was therefore a crucial stage in the class struggle and an essential forerunner to the future communist order.

Marx's emphasis on the need for self-conscious revolutionary organization had been at least partially inspired by Gracchus Babeuf and his 'Conspiracy of Equals' of 1796. Although Babeuf had few followers and his plans were foiled by police infiltration, his dream of a revolutionary seizure of political power by a secretly organized conspiratorial group lived on through the efforts of the Italian revolutionary Filippo (or Philippe) Buonarroti. Buonarroti, a member of Babeuf's circle in 1796, published a history of the Conspiracy of Equals in 1828. It brought the message of revolutionary seizure of power to secret societies throughout Europe. One of Buonarroti's followers, the French revolutionary socialist Auguste Blanqui, combined Buonarroti's appeal for a secret revolutionary brotherhood with the idea of a 'dictatorship of the proletariat,' a term coined by Blanqui to define the political goal for socialism. Marx and Engels incorporated both Babeuf's and Blanqui's notions into their program for proletarian revolution.

2.2 The Foundations of Modern Social Science

Because the French Revolution brought the relationship between society and politics into sharper focus, it also accelerated the development of the social sciences, first in France and then more broadly in Europe. The Directory regime created the first research institute for the social sciences with its department of political and moral sciences in the Institut. Bonaparte, himself a member of the Institut, took more than 150 scientific experts with him on his expedition to Egypt in 1798. They set up an Institut d'Egypt in Cairo and helped inaugurate 'oriental studies' in Europe. They also discovered the Rosetta Stone which enabled experts to decipher Egyptian hieroglyphics.

The French Revolution had an especially dramatic impact on historical studies in France. The first full-length histories of the Revolution were published in the 1820s by François-Auguste Mignet and Adolphe Thiers, journalists and editors in the political opposition. They used the history of the French Revolution to defend the notion of a constitutional monarchy and to criticize the restored Bourbons. In the 1840s, Louis Blanc, Alphonse Lamartine, and Jules Michelet wrote histories of the Revolution as part of their campaign for a republic. Although Marx did not write a history of the French Revolution of 1789, he did read the French historians and made constant reference in his work to the French Revolution of 1789.

Marx was not alone in considering the French Revolution an important stage in historical development. As the notion of historical evolution spread in the nineteenth century, the French Revolution almost inevitably became an important benchmark. Auguste Comte, one of the founders of sociology, took as his point of departure the last writings of the Marquis de Condorcet, a French nobleman and revolutionary who committed suicide rather than face the guillotine in 1794. Comte proposed three stages of historical development: the theological, the metaphysical, and the positive or scientific. The French Revolution helped bring about the ongoing transition to the third, positive stage by liberating modern social elements from the dead hand of the Old Regime. The task of social science was to complete this transition, which could only be accomplished finally in ideas and opinions, according to Comte, not by revolutionary action. The ultimate aim of his system was the reorganization of society by means of scientific sociology.

Alexis de Tocqueville, who helped lay the foundations for political sociology and political science with his *L'ancien régime et la Révolution* (1856), considered the French Revolution not as a stage so much as an accelerator of the general modern trend toward democracy and equality. Tocqueville had read the previous historians of the French Revolution, though he rarely cited them, and absorbed the view that social changes had inevitable political repercussions. He wove his analysis of social changes into a subtle examination of French political culture and especially the dominance of the central state. In so doing he created an alternative to the Marxist emphasis on class struggle. Tocqueville considered the French Revolution pivotal in the advance of modernity but largely negative in its effects. Tocqueville saw in the French Revolution the tendency of democratic revolution to reproduce the despotism of monarchical authority by leveling all institutions that might have provided a check on state power. Although the French made their revolution for liberty and equality, they gave up liberty in order to be assured of equality and ended up with an equality of servitude to the Napoleonic state. This line of analysis had a long posterity—Hannah Arendt's analysis of revolution has clear affiliations to Tocqueville's—and continues to influence political and social theory right down to the present.

3. Scholarly Controversies

Since analysis of the French Revolution has played such a prominent role in the development of competing political ideologies, it is hardly surprising that it has generated disputes among scholars as well. In the nineteenth century writers used the French Revolution

to argue for and against monarchy, empire, or republicanism as a form of government. After the eruption of the Russian Revolution in 1917, however, debate shifted to the terrain defined by Marxism and communism, and in particular, the French Revolution as the exemplar of bourgeois revolution.

3.1 The Question of Marxism

Within France, Marxist historians of the French Revolution enjoyed the advantage of having captured the chair of the history of the French Revolution at the Sorbonne (University of Paris). For most of the twentieth century, the chair was occupied by a succession of talented socialists and communists. They pioneered the study of social tensions and social classes, of 'history from below.' The most remarkable of them was Georges Lefebvre, who wrote wide-ranging studies of the peasantry during the Revolution. Lefebvre's combination of narrative skills and analytical studies gave him an unrivaled reputation; even those who objected to his Marxist views regarded him as the most accomplished academic historian of the Revolution.

The Right did not have the same initial foothold within the French university system. Hippolyte Taine, for example, wrote his four volumes on the Old Regime and the French Revolution as a general literary and cultural critic, not as a professional historian. Taine's views remained fundamental to the Right-wing criticism of the Revolution. He had nothing but contempt for 'the people' so beloved by Michelet. He frequently likened the lower classes to beasts or insects; at worst, they were a crazed, frenzied rabble. Augustin Cochin toned down this virulence, arguing that the lower classes were manipulated by professional politicians who came to power through the activities of secret societies such as Masonic lodges. The Right took up some of the claims of the Left—that the French Revolution foreshadowed the Russian Revolution of 1917—but gave them a negative rather than a positive spin.

Before the 1970s, the Marxist Left enjoyed the dominant position in French academic debates, in part because of its strength within the university, and in part because the Right was associated with anti-democratic politics. The balance began to shift, however, as critics of the Revolution finally accepted the republican form of government. A more moderate republican Right got support from a growing body of work produced in the Anglophone countries, especially the United States and the United Kingdom, which attacked the central Marxist thesis that the French Revolution was a bourgeois revolution. Critics argued that the capitalist bourgeoisie had not made the Revolution (first aristocrats and then middle-class officials and lawyers had led it, not manufacturers); that the Revolution had not overthrown feudalism (because feudalism no longer existed); and that the Revolution did not pave the way for capitalism (since industrialization did not take place until the 1840s).

François Furet, a former communist, brought together the new Anglophone work in a frontal assault on the Marxist interpretation. This 'revolutionary catechism,' this 'Lenino-populist vulgate,' he argued, fundamentally distorted the history of the French Revolution by depicting it as a class struggle that anticipated communism. Furet claimed as his ancestors Tocqueville and Cochin; he did not reject democracy, but rather wanted to understand why democracy in the French Revolution developed into Terror or totalitarianism. The social underpinnings of the French Revolution could never explain this skidding out of control into Terror, he maintained; revolution was an autonomous political and ideological movement, not the expression of a social class (Furet 1978). By 1989, the year of the bicentennial of the French Revolution, Furet's version of the French Revolution had became the new orthodoxy. The political balance had shifted, not only in France but almost everywhere else as well. The French Revolution no longer seemed a communist preserve.

3.2 Current and Future Directions

Since 1989, as the polemics over Marxism have subsided, the work of historians from other countries has attracted more attention. American, British, German, and Italian historians now offer alternative views of their own. No one school, whether Left-wing or Right-wing, French or Anglophone, currently dominates the historiography of the French Revolution. The shift of political interest away from Marxism has opened up new areas of research, including gender relations, definitions of citizenship, and the role of the colonies. These new perspectives reflect concerns with human rights and globalization.

Since the French Revolution began by enshrining the notion of 'the rights of man,' it has been a prime site for work on the dilemmas of universalism. Feminists have pinpointed the Enlightenment's masculinist bias as the chief cause of the failure of the revolutionaries to include women in their promise of citizenship; Enlightenment writers, especially Montesquieu and Rousseau, used nature and reason to justify women's exclusion from the public sphere, and the revolutionaries simply drew the logical conclusion from these assumptions. Other feminists, however, have insisted that the Revolution opened up new opportunities for women, legalizing divorce, assuring equal inheritance to girls, and affording women the opportunity to establish political clubs. Unlike women, men who were Protestants, Jews, free blacks, mulattos, or even slaves gained political rights during

the French Revolution. Yet even these enfranchisements revealed telling fault lines in the revolutionary conception of citizenship; the National Assembly, for instance, insisted that the Jews must give up their particular identity and corporate privileges as Jews in order to become French citizens. Since discussions of free black and slave rights inevitably involved the slave colonies, and since France's colony of Saint Domingue (San Domingo) was home to the first successful slave revolt in history beginning in 1791, the interest in citizenship has also stimulated study of France's slave colonies. Because the French Revolution put both revolution and modernity on the agenda, it continually offers rich possibilities for historical and theoretical debate.

See also: Bourgeoisie/Middle Classes, History of; Citizenship, Historical Development of; Democracy; Democratic Theory; Enlightenment; Freedom/Liberty: Impact on the Social Sciences; Ideology: History of the Concept; Nationalism: General; Political Representation; Representation: History of the Problem; Rousseau, Jean-Jacques (1712–78); Social Science, the Idea of; Socialism: Historical Aspects; Tocqueville, Alexis de (1805–59)

Bibliography

Arendt H 1963 *On Revolution*. Viking Press, New York

Barnave A 1988 *Introduction à la Révolution française. De la Révolution et de la Constitution*. Presses Universitaires de Grenoble, Grenoble, France

Buonarotti P 1828 *Conspiration pour l'égalité dite de Babeuf suivie du procès auquel elle donna lieu, et des pièces justificatives*. A la Libraire Romantique, Brussels, 2 Vols.

Burke E 1790 *Reflections on the Revolution in France, and on the Proceedings in Certain Societies in London Relative to that Event*, 3rd edn. J. Dodsley, London

Furet F 1978 *Penser la Révolution française*. Gallimard, Paris

Furet F 1986 *Marx et la Révolution française*. Flammarion, Paris

Godineau D 1988 *Citoyennes tricoteuses: les femmes du peuple à Paris pendant la Révolution française*. Alinéa, Aix-en-Provence, France

Hunt L A 1984 *Politics, Culture, and Class in the French Revolution*. University of California Press, Berkeley, CA

James C L R 1938 *The Black Jacobins: Toussaint L'Ouverture and the San Domingo Revolution*. Secker & Warburg, London

Landes J B 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca, NY

Taine H 1885–88 *La Révolution*. Hachette, Paris, 3 Vols.

Tocqueville A de 1856 *L'ancien régime et la Révolution*, 2nd edn. Michel Lévy frères, Paris

Tocqueville A de 1977 Oeuvres complètes. In: Jardin A, Lesourd J-A (eds.) *Correspondance Tocqueville-Louis de Kergolay*. Gallimard, Paris, Vol. 13

L. Hunt

Frequentist Inference

1. Data, Models, and Inference

The starting point of a statistical analysis is a set of data, for example, of counts or measurements. One aim may be simply to study what these data have to tell us. If they consist of a set of real numbers we might want to see, for example, whether they are small or large; tightly concentrated or spread out, whether they are stable or tend to increase with time, etc. If they are points in the plane we can get an idea of the shape of this set of points, for example, whether they cluster about a line. The branch of statistics dealing with this kind of investigation used to be called descriptive statistics, but now goes by the name data analysis or, more precisely, exploratory data analysis (EDA), a term introduced by Tukey (see *Exploratory Data Analysis: Univariate Methods*. Instead we consider how we can quantify the conclusions or decisions drawn from an analysis. Frequentist inference requires that any quantifying measure be interpretable in terms of frequentist probability of events (see *Frequentist Interpretation of Probability*) . That is, we assume the data are random quantities produced by some probability distribution and that something is known about this distribution. For example, if we have a set of n measurements of some quantity, we may assume that these measurements are independently and identically distributed. We may stop there or may go further and make some assumptions about this common distribution.

Statistical analyses based on such assumptions were common in the nineteenth century, and isolated instances can be found in the eighteenth century; but the first general framework was proposed by R. A. Fisher (1922) (see *Fisher, Ronald A (1890–1962)*). He states:

> The object of statistical methods is the reduction of data. A quantity of data which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole.
>
> This object is accomplished by constructing a hypothetical infinite population of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all quantities under discussion.

In modern terminology we might paraphrase this proposal by saying that we construct a mathematical model, according to which the data are produced by a probability distribution assumed to belong to some specified parametric family of distributions.

What are the statistical methods based on these models supposed to achieve? Fisher called the aim 'inductive inference' or 'inductive reasoning,' and described it variously as 'learning by experience' and

as 'drawing inferences from the particular to the general, from consequences to causes' or—using statistical language—'from the sample to the population.' 'The purpose of inductive reasoning based on empirical observations,' he wrote 'is to improve our understanding of the systems from which these observations are drawn.'

This interpretation was criticized by Neyman (1961) who objected that 'After a conscientious effort to find the exact meaning of this term [inductive reasoning] I came to the conclusion that, at least in the sense of Fisher, the term is empty.' Neyman proposed instead that the purpose of statistical analysis is to serve as guide to appropriate action; and, in contrast to Fisher's inductive reasoning, named it inductive behavior (see also *Neyman, Jerzy (1894–1981)*).

Neyman's point of view was taken up by Wald who used it to construct a comprehensive framework for statistical decision making (Wald 1950). It involved three principal elements

(a) A family of probability distributions representing the various possible true situations.

(b) A set of actions from which it is the statistician's task to choose the most appropriate one.

(c) A loss function which measures the loss resulting from any action taken for any of the possible true situations.

Wald's formulation was castigated by Fisher as completely inappropriate for scientific work. 'It is important,' he wrote (Fisher 1973, p. 106) 'that the scientific worker introduces no cost functions for faulty decisions. To do so would imply that the purposes to which new knowledge was to be put were known and capable of evaluation. As workers in Science we aim at methods of inference which shall be equally convincing to all freely reasoning minds, entirely independently of any intentions that might be furthered by utilizing the knowledge inferred.'

While it is difficult not to sympathize with Fisher's stress on scientific aims as an alternative to immediate utilitarian use, his statements are distressingly vague. An effort to be more specific, which still seems rather vague, was made by Tukey (1960). For the outcome of a nonaction-oriented statistical analysis he proposed the term 'conclusion.' Concerning the difference he writes: 'Conclusions are established with careful regard to evidence but without regard to consequences of specific actions in specific circumstances. Conclusions are withheld until adequate evidence has been accumulated.'

The distinction between Fisher's point of view and that of Neyman and Wald, between conclusions and decisions, concerns the interpretation of the results of a statistical analysis. How those results are reached from the given model by means of frequentist inference f orms the subject of this article. (For the frequentist meaning of probability see *Frequentist Interpretation of Probability*.) It includes the derivation of appropriate procedures, the comparison of different procedures, investigating the sensitivity of a procedure to departures from the assumed model, and the checking of the suitability of the model. Another topic of interest, considered in Sect. 3.6, is the selection of an appropriate model.

There is an alternative approach to inference which can embody both the decision theory aspects of the Neyman–Wald approach and Fisher's views on learning by experience. This is the Bayes approach which requires not only the frequentist's specification of a family of possible distributions for the observation, but also that the distribution of the observation has itself been drawn from the specified model according to a known probability mechanism. This is coupled with the subjective point of view of probability in which the chance of an event represents a quantitative measure of the observers' odds on the occurrence or nonoccurrence of the event. This point of view, though attractive, puts a burden on the observer to specify his or her state of mind with possibly unreasonable precision and, of course, leaves open the issue of the generalizability of the conclusions by scientists at large (see also *Bayesian Statistics*).

The problems of inference have been the subject of much discourse among philosophers as well as scientists whose primary interest was not statistics. We refer to Keynes (1921), von Mises (1928) and Jeffreys (1939) among others.

A problem in discussing frequentist inference is that, while distinctions such as those between Fisher, Neyman, and Wald are required in order to formulate the goals of theory, they are artificial in practice. Even in scientific inquiry the cost of initiating extensive further investigation on the basis of weak evidence has to be kept in mind. On the other hand, the most interesting consequences may not be foreseen and the costs of actions difficult to quantify. Is foisting a slightly better-than-average drug with serious side effects on most patients better than doing nothing? What if the drug turns out to have major benefits for a small group of patients not represented in the original clinical trial of the drug? A chemist, E. B. Wilson (1952), after considering the work of the thinkers we mention, pleads eloquently, 'There is a great need for further work on the subject of scientific inference. To be fruitful it should be carried out by critical original minds who are not only well-versed in philosophy but also familiar with the way scientists actually work (and not just with the way some of them say they work).'

Wilson concludes pessimistically: 'Unfortunately the practical nonexistence of such people almost suggests that the qualities of mind required by a good philosopher and those needed by a working scientist are incompatible.'

In what follows we

(a) begin, for illustrative purposes, by describing a few simple models that have arisen from a frequentist point of view (Sect. 2), and

(b) discuss by example the principal frequentist inference procedures (hypothesis testing, point estimation, confidence regions, prediction, and model selection) (Sect. 3).

We use the term frequentist inference if all probabilities involved (such as significance level, power, and coefficient) are interpreted as frequentist probabilities in the sense of the article (*Frequentist Interpretation of Probability*). In particular, the assertion that the value of such a probability is p then means that in a large number of independent cases with probability p of some outcome, we expect the outcome to occur with a frequency close to p. In this connection it is important to realize that there is no need for these situations to be at all like each other, as long as the probability p is the same in each (see Neyman 1937).

2. *Some Simple Models*

Frequentist models are characterized by two principal features.

(a) The variability of the data. If the observations were repeated, different values would be obtained. This variability is represented in the model by postulating that the data are generated by a probability distribution.

(b) The unknown aspect of the situation which the statistical investigation is to elucidate. This unknown feature is represented in the model by the fact that the probability distribution of the observations is assumed to be only partially known.

These two features are seen clearly in the following model, which was mentioned briefly in the preceding section.

Example 1. (Error-measurement model) If $X_1, \ldots, X_n$ denote n measurements of an unknown quantity θ, this model can be written as

$$X_i = \theta + \varepsilon_i. \tag{1}$$

Here θ represents the unknown quantity we wish to estimate or test. The errors ε_i are the source of the variability of the observations. This frequency model assumes that, in a replication of the situation, the unknown value of θ would be the same but that the ε's would take on different values. The most common assumption concerning their variability is that they are independent random variables with a common distribution with mean zero. This distribution may be assumed to be known. More usually it too embodies some unknown features; for example, its form may be known but it may contain an unknown scale parameter.

Simpson and others toward the middle of the eighteenth century considered various possible forms for this distribution; but by the early nineteenth century it was agreed generally that the most suitable distribution of such measurement errors, in most cases, was the normal. The basis for this belief was the hypothesis of 'elementary errors,' which assumes that observational errors are the sum of a large number of small, elementary errors; and, by the Central Limit Theorem, are therefore approximately normally distributed.

If the ε's in (1) are assumed to be independently normally distributed with the mean 0 and variance σ^2, the model (1) is equivalent to assuming that

$$X_1, \ldots, X_n \text{ are independent } N(\theta, \sigma^2). \tag{2}$$

Example 2 (Normal one-sample model) Model (2) arises also in contexts quite different from the measurement situation described above. Suppose X is some numerical characteristic of a person such as height, weight, blood pressure, intelligence, etc., or of an animal, plant or manufactured product. However, this time we are not taking several observations X_i on the same subject but one observation each on different subjects drawn at random from a population which, in the model, we shall assume to be infinitely large. In this setting θ is the mean value of X in the population, and (2) is called the normal one-sample model. (The measurement situation can be viewed as a special case if one considers the n measurements actually taken as a sample from an essentially infinite number of measurements that could be taken.)

In these applications the assumption of normality is frequently made even though it is often not suitable and each situation ought to be considered in its own right. Often normality can be brought closer by applying it not to the observations X_i themselves but to some transformation $T(X_i)$. In other cases the normal model (2) is replaced by the more general model

$$X_1, \ldots, X_n \text{ are independently distributed according to } F(x-\theta) \tag{3}$$

where F has mean zero and otherwise is arbitrary, or where it is assumed to be an arbitrary distribution that is symmetric about zero. Some smoothness conditions may also be imposed; for example, that F has a density f.

Example 3. (Linear Model) Large areas of statistical methodology are based on the following extension of (1), due to Gauss (1809) and known as the linear model

$$X_i = \sum_{j=1}^{p} \beta_j z_{ij} + \varepsilon_i \quad (i = 1, \ldots, n) \tag{4}$$

where the z's are known constants (sometimes called explanatory variables) and the β's unknown parameters. The ε's are the observational errors as before and are assumed to have expectation zero.

Examples of (4) abound. Here are a few.

(a) The measurement model (1) is the special case of (4) with $p = 1$, $z_{i2} = 1$ and $\beta_1 = \theta$.

(b) The two-sample problem with $n_1 + n_2$ independent variables $X_1, \ldots, X_{n_1}: N(\theta_1, \sigma^2)$ and $X_{n_1+1}, \ldots, X_{n_1+n_2}: N(\theta_2, \sigma^2)$ corresponds to $p = 2$,

$$z_{11} = \cdots = z_{1_{n1}} = 1, \quad z_{1n_1+1} = \cdots = z_{1n_1+n_2} = 0, \quad \beta_1 = \theta_1 \tag{5}$$

and

$$z_{21} = \cdots = z_{2n_1} = 0, \quad z_{2n_1+1} = \cdots = z_{2n_1+n_2} = 1, \quad \beta_2 = \theta_2 \tag{6}$$

(c) The k-sample model is defined analogously. The two-way layout without interactions assumes that the variables X_{ijk} $(k = 1, \ldots, n_{ij}; i = 1, \ldots, a; j = 1, \ldots, b)$ are independent normal with common variance σ^2 and with means

$$E(X_{ijk}) = \mu + \alpha_i + \beta_j (\Sigma\alpha_i = \Sigma\beta_j = 0). \tag{7}$$

The k-sample model and the two-way layout are the simplest cases of analysis of variance models.

Another important class of examples of (4) are models for regression. The simplest case is

(d) Simple linear regression, given by

$$X_i = \alpha + \beta t_i + \varepsilon_i \tag{8}$$

where the t's are known constants, and α and β unknown parameters.

This is a special case of

(e) Polynomial regression, with

$$X_i = f(t_i) + \varepsilon_i \tag{9}$$

where f is a polynomial (for example, quadratic) of known degree and with unknown coefficients.

Other functions f, of course, are also possible.

A modification of (4) that is sometimes suitable is to have some of the z's random. For example (7) in some situations is replaced by

$$X_{ijk} = \mu + A_i + \beta_j + \varepsilon_{ijk} (\Sigma\beta_j = 0) \tag{10}$$

where the A's are unobservable random variables assumed to be independent $N(0, \sigma_A^2)$ and independent of ε. For more, see *Linear Hypothesis*.

When the distribution of the observational errors the ε's in (1) and (4) is specified, the models considered so far are parametric, that is, they can be smoothly parameterized by Euclidean labels. Models such as (3) with the distributional form of errors unspecified up to symmetry are semiparametric. Nonparametric models are ones in which as little is assumed as possible.

Example 4. (Nonparametric regression) This model is suitable in situations where, on each member of a sample of size n from some population, we observe not only a response Y but also observe characteristics (covariates) $\mathbf{Z} = (Z_1, \ldots, Z_k)$. For example, Y might be the income of a randomly sampled individual, Z_1 might be age, Z_2 educational level, etc. Of principal interest here is the average relation between Y and $\mathbf{Z}$

$$E(Y|Z_1 = z_1, \ldots, Z_k = z_k) \equiv m(z_1, \ldots, z_k). \tag{11}$$

If we assume nothing about the form of m, then subject to the assumptions that $E|Y| < \infty$ and that $(Y_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{Z}_n)$ are independent and identically distributed, this model is completely general since we can always write

$$Y_i = m(Z_{i1}, \ldots, Z_{ik}) + \varepsilon_i \tag{12}$$

where $E(\varepsilon_i | Z_i) = 0$. Note that if we assume

$$m(z_1, \ldots, z_k) = \sum_{j=1}^{k} \beta_j z_j,$$

then, given $\mathbf{Z}_i = z_i$ $(i = 1, \ldots, n)$ we are back in the linear model (4).

In order to identify m in (12) we need to make some assumptions, for instance that m is continuous.

An important method for constructing more complex models from simpler ones is that of hierarchical models. As exemplified in (10), such models often contain unobservable random variables which are themselves of interest.

From this point of view Bayes models can be considered as two-stage hierarchical models in which the first stage is a frequentist model involving unobservable parameters for which the Bayes model specifies a distribution as a second stage. If the frequentist model is itself hierarchical, a last stage is added when the model is Bayesianized. To this difference—that the unknown parameters of the frequentist model become unobservable random variables in the corresponding Bayes model of course, is added the difference in interpretation of the probabilities involved as frequentist or subjective.

3. *Inference Methods*

As indicated in Sect. 1 linked closely with the frequentist paradigm is that of decision theory. The principal types of problems considered in that framework are

(a) hypothesis testing,

(b) point estimation,

(c) confidence regions,

(d) prediction,
(e) model selection.
We shall discuss these in the above order.

3.1 Hypothesis Testing

Wald's decision theory grew out of the Neyman–Pearson theory of hypothesis testing (Neyman and Pearson 1933) which the authors formulated as choosing between two decisions, whether a prespecified statement (the hypothesis H) about the distribution P generating the data is or is not correct. If $\theta \in \Theta$ parameterizes the model, this is equivalent to determining whether $H: \theta \in \Theta_0$ or the alternative $K: \theta \in \Theta_1$ is true.

This formulation is satisfactory in many practical contexts such as occur, for example, in medicine, agriculture, or industry. For instance, in Example 3(b) the two samples may correspond to control and treatment groups of patients. In Example 3(b) the assumption of normality is made, and implicitly that of an additive treatment effect measured by the difference $\theta_2 - \theta_1$. The hypothesis to be tested is usually formulated as $H: \theta_2 \leqslant \theta_1$; i.e., that the treatment has no beneficial effect. The alternative $K: \theta_2 > \theta_1$ claims the existence of a beneficial effect.

The test for this problem is a test function δ which takes on the values $\delta(x) = 0$ or $\delta(x) = 1$ for the sample points x for which we decide to accept or reject H. The performance of such a test is measured by the probabilities of

(a) Type I errors—rejecting H when H is true, and
(b) Type II errors—accepting H when H is false,
that is, by the probabilities

$$P_\theta[\delta(x) = 1] \text{ when } \theta \in \Theta_0 \quad \text{and}$$
$$P_\theta[\delta(x) = 0] \text{ when } \theta \in \Theta_1. \tag{13}$$

The test δ is required to satisfy the condition

$$P_\theta(\text{Type I error}) \leqslant \alpha \quad \text{for all } \theta \in \Theta_0, \tag{14}$$

where the level α of the test is preassigned (usually as 0.05 or 0.01) as the maximum probability of false rejection to be tolerated. Subject to (14) we look for tests δ which make P_θ (Type II error) as small as possible.

This formulation is meaningful if it is possible to distinguish sharply between hypothesis and alternatives, between those θ-values for which rejection of H is clearly desirable or undesirable. In Example 3(b), for instance, we can specify that only beneficial effects exceeding $\varepsilon\sigma$ (for some given $\varepsilon > 0$) are of interest. Such a distinction is often possible in the contexts we have mentioned.

The Neyman–Pearson formulation of hypothesis testing in terms of the two types of error works well when we have clear definitions of both the hypothesis and the alternatives. In scientific contexts it is often the case that the hypothesis is clear but the alternatives are vague. It is then the unlikelihood of observations that would be surprising if the hypothesis were true (but less so if the vaguely entertained alternatives were valid) which takes on particular significance. It is still necessary to define a test statistic T and to suppose the hypothesis is rejected when its values are sufficiently large, since such values would be surprising if H were true but not if H were false. Then the probability of observing values greater than or equal to the observed value of T is the p-value or significance probability of the hypothesis. If it is sufficiently small, it draws attention to anomalies not expected under H. The calculation of such probabilities is central to science. However, whether such attention should be followed by action as assumed in the Neyman–Pearson theory depends on extraneous considerations.

The Bayesian approach to testing assigns prior probabilities to hypothesis and alternative and calculates the ratio of posterior probabilities, the Bayes factor, as a measure of evidence in favor of or against the hypothesis rather than the significant probability. This approach is very attractive in providing exactly what everyone wants: the probability of the hypothesis being true, given the data. However, this probability depends on the prior probability distribution assigned to the parameters; and these priors can overwhelm strong opposite data evidence. Disparate priors can result in conflicting conclusions based on the same data by analysts holding different prior opinions (see also *Hypothesis Testing in Statistics*).

3.2 Point Estimation

The decision-oriented point of view of Neyman and Pearson led to the full-fledged decision theory of Wald, based on assigning losses to the consequences of inappropriate decisions. Perhaps the most successful example of this approach is point estimation (for which the idea of loss functions had already been discussed by Gauss and Laplace).

The prototypical estimation problem is that of estimating an unknown physical constant subject to measurement error discussed in Example 1. The loss function most commonly considered is squared error leading to its expected value, the mean squared error (MSE), as its risk. The standard assumption of independent normal errors with mean zero and common variance implies that the sample mean $\bar{X}$ is the appropriate estimator according to a number of different criteria.

(a) Minimax. This is a worst case analysis in which the maximum risk is minimized.

(b) Uniform minimum variance unbiased. An estimator of θ is said to be unbiased if its expectation is equal to θ. For unbiased estimators the MSE is equal to the variance; and, in the present case, $\bar{X}$ minimizes the variance among all unbiased estimators.

(c) Maximum likelihood (ML). This is a non-decision-theoretic approach based on the likelihood the density of the data viewed as a function of the parameter. Fisher considered the likelihood of an observation as a measure of support for the different θ-values that might have generated it and proposed the θ-value maximizing it as an appropriate estimator of θ. When a uniform prior distribution can be placed on the parameter, the ML estimator is just the mode of the posterior distribution. However, when θ ranges over the infinite line (as is the case in the measurement model and many others), no uniform probability distribution for θ exists; and only an 'improper Bayes' interpretation is then possible.

The ML principle is popular for a number of reasons

(a) general applicability,

(b) invariance under reparameterization,

(c) in cases such as Example 2, where meaning can be attached to 'large sample sizes,' the method is approximately optimal,

(d) in many standard cases it leads to explicit solutions.

However, there are examples in which the resulting estimators are completely misleading, (LeCam 1990, Ferguson 1996). In particular, ML is either not applicable or can lead to very poor results in many non- and semiparametric models such as Example 4. For a general discussion of different approaches to estimation see, for example, Bickel and Doksum (2000), Lechmann and Casella (1998).

The assumption of joint normality of the data, for instance in Examples 1–3, is often seriously in doubt. In such cases robustness criteria are invoked and lead to alternatives to the linear procedures appropriate for normally distributed data. For example, the median rather than the mean (or some intermediate trimmed mean) might be used in Example 1, and minimum L_1 (least absolute values) rather than L_2 (least squares) estimators in Example 3. For an extended discussion of robustness see Hampel et al. (1986) and Staudte and Sheather (1990).

3.3 Bayes Estimation

Still another principle is that of Bayes estimation, i.e., minimizing the expected risk under a prior distribution π of the parameter $\boldsymbol{\theta}$. Since this is the same as minimizing the average risk, averaged with respect to the weight-function π, such estimators are of interest both from a Bayesian and a frequentist point of view. For estimating a real-valued parameter $g(\boldsymbol{\theta})$ with data X and squared error loss, the Bayes estimator is

$$E[g(\boldsymbol{\theta})|X] \tag{15}$$

the mean of the posterior distribution of $g(\boldsymbol{\theta})$ given X.

Bayes estimators are essentially never unbiased (Girshick and Savage 1951) since they naturally 'pull' the estimator toward one's prior opinion, the prior mean of $g(\boldsymbol{\theta})$. For example, in the one-sample model of Example 2. If π is normal (μ, τ^2), the Bayes estimator is

$$\left(\frac{n\bar{X}}{\sigma^2}+\frac{\mu}{\tau^2}\right)\Big/\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right). \tag{16}$$

Frequentist and Bayes inference generally merge for large sample sizes since the data then wash out the influence of the prior distribution (provided the prior does not exclude parts of the sample space); see, for example, Blackwell and Dubins (1962).

In estimation even more can be said. If the prior distribution is locally uniform, Bayes estimators are asymptotically optimal in a frequentist sense by the Bernstein–von Mises theorem (see, for example, Lehmann and Casella 1998, Bickel and Doksum 2000). This is, however, a limit result; in practice, the effect of a strong prior often persists even for substantial sample sizes.

Both aspects can be illustrated by the Bayes estimator (16). For fixed μ, the difference of this estimator from $\bar{X}$ is of the order $1/n\tau^2$. For fixed τ^2, it is accordingly of the order $1/n$, while the difference of both estimators from θ is of the order $1/\sqrt{n}$. On the other hand, the convergence of (16) can become arbitrarily slow if $\tau^2 \to 0$ as $n \to \infty$. The basic message is that optimal frequentist estimators are appropriate for Bayesian problems and vice versa.

For the sake of simplicity, we have here focused attention to the estimation of real-valued parameters. Of course, estimation problems arise also for multivariate parameters (for instance, in Example 3) and for function-valued parameters such as the function m in Example 4 or in the estimation of an unknown density. Squared error loss is then replaced by summed or integrated MSEs. However, the univariate and multivariate situations present unexpected differences. For example, the typically unique minimax estimator of the univariate case is replaced by an infinite set of minimax estimators; and among these the natural extension of the classical univariate estimator is no longer optimal (Stein 1956).

Point estimates by themselves are hardly ever sufficient. Typically one also requires an idea of the error committed in using the estimator; that is, an estimate of the MSE. This is again the problem of estimating a parameter, and there is little to add. However, interpretation of the estimated measure of error is often difficult. It is clarified by the concept of confidence estimation, which we take up next.

3.4 Confidence Regions

When estimating θ in the measurement error model of Example 1, it is customary to indicate the reliability of

the estimate $\bar{X}$ by attaching to it the error estimate $\pm S/\sqrt{n}$ with $S^2 = \Sigma(X_i - \bar{X})^2/(n-1)$ as a proxy for $\pm\sigma/\sqrt{n}$. In the case of normal errors with mean zero.

$$P_\theta\left[\bar{X} - \frac{\sigma}{\sqrt{n}} \leqslant \theta \leqslant \bar{X} + \frac{\sigma}{\sqrt{n}}\right] = 0.67 \quad \text{for all } \theta. \tag{17}$$

If σ is known, (17) provides intervals of fixed length (and with random midpoints) which, in many repetitions of the experiment of taking n measurements, will 'cover' the true θ 67 percent of the time whatever are the true values of θ and σ. The intervals (17) are confidence intervals for θ with confidence coefficient 0.67.

In practice, σ is usually unknown. We can then obtain confidence intervals for θ by replacing σ by its estimator S. If t_α is the $100(1-\alpha/2)$ percent point of the t-distribution with $n-1$ degrees of freedom, the intervals

$$\bar{X} - t_\alpha\frac{S}{\sqrt{n}} \leqslant \theta \leqslant \bar{X} + t_\alpha\frac{S}{\sqrt{n}} \tag{18}$$

cover the true θ with probability $1-\alpha$ so that the statement (18) is correct about $100(1-\alpha)$ percent of the time. The length of the intervals is no longer fixed but random, its length tending to increase as the accuracy of the observations (the inverse of which is measured by σ and estimated by S) decreases.

In general, in estimating a parameter θ taking values in an arbitrary space Θ, a 100 $(1-\alpha)$ percent confidence region for θ is a random subset $C(X)$ of Θ, depending on the data with the property that

$$P[\theta \in C(X)] \geqslant 1-\alpha \tag{19}$$

for all probability distribution P in the model. Here the region $C(X)$ can, for example, be an interval as above, an ellipse as in the Scheffé regions for the parameters of the linear model in Example 3, or a 'band' about an unknown distribution function.

Confidence regions can be viewed as simultaneously making statements about the acceptability of a family of hypotheses (see Lehmann 1986). In particular, they provide a measure of the acceptability of the alternatives. A 95 percent confidence interval for the treatment effect of Example 3(b), for instance, which includes zero, tells us not only that the null hypothesis is accepted at the 5 percent level but also specifies all alternatives to the null hypothesis which are accepted at that level. An interval which does not contain zero but to which zero is close suggests that although we have seen something that is surprising under the null hypothesis, the data are consistent with alternatives that do not differ materially from zero.

The interpretation of confidence regions is conceptually difficult because the probability statements refer not to any random variation of the parameter, which, although unknown, is considered fixed, but to that of the data-dependent region. In contrast, a Bayesian analysis views θ as random and constructs credible regions $C^*(X)$ such that the posterior probability of θ falling into $C^*(X)$ is $\geqslant 1-\alpha$. This is just the kind of conclusion one would like to make; but, as in the case of testing, tends to be strongly influenced by assumptions concerning the prior.

Again, as in point estimation, Bayes regions derived on the basis of optimality criteria of size (length, volume, etc.) for large sample sizes agree approximately with frequentist confidence regions satisfying minimax size properties subject to fixed probability of coverage (see also *Estimation: Point and Interval*).

3.5 Prediction

This is the part of frequentist inference that is closest to the Bayesian approach and that fits least well into the decision-theoretic framework. Typically we are given a sample from a population of the form $(\mathbf{Z}_i, Y_i)$, $i = 1, \ldots, n$ where $\mathbf{Z}_i$ is a vector of predictors. Using this sample (which provides information concerning the relationship of $\mathbf{Z}$ and Y) and a new $\mathbf{Z}_{n+1}$ we wish to predict the unobserved Y_{n+1}. The predictor of Y_{n+1} is a function

$$\delta(\mathbf{Z}; \mathbf{Z}_1, Y_1, \ldots, \mathbf{Z}_n, Y_n) = \hat{\delta}(\mathbf{Z})$$

such that $\hat{\delta}(\mathbf{Z}_{n+1})$ is used to predict Y_{n+1}. In the continuous case (regression), the classical measure of loss resulting from incorrect prediction is squared error,

$$[Y_{n+1} - \hat{\delta}(Z_{n+1})]^2.$$

On the other hand, in the case of categorical variables taking on values $\{1, \ldots, k\}$ (classification), the standard loss function takes on the value 0 if $Y_{n+1} = \hat{\delta}(\mathbf{Z}_{n+1})$ and 1 otherwise. Prediction has the attractive feature that it allows nonparametric estimates of error since for any procedure we can compute its performance by applying it to the training sample on which it was built, for instance, $(1/n)\Sigma_{i=1}^n(\hat{\delta}(Z_i) - Y_i)^2$. This measure typically underestimates error, but techniques such as cross validation can be used to adjust it.

3.6 Model Selection

An important aspect of a frequentist analysis is the specification of a model. As an illustration, consider Example 3(e) where the regression function is assumed to be well approximated by a polynomial. Suppose the errors are normally distributed with zero mean and common variance. To complete the specification of the model, it is then necessary to decide on the degree d of the polynomial. With classical (least squares) prediction methods, the variance of the prediction error increases and its bias decreases with d. On a frequentist

basis the data can then be used to select d so as to obtain a desirable balance between these two aspects. Criteria for selecting d (and making analogous choices within a family of proposed models in other problems) have been developed by Akaike, Mallows, Schwarz, and others (for more details, see, for example, Linhart and Zucchini 1986):

Bayesian solutions to this problem differ in a number of ways, reflecting whether we assume that the true model belongs to the hypothesized class of models (e.g., is really a polynomial) or can merely be approximated arbitrarily closely by such models. For more on this topic see Shao (1997).

See also: Estimation: Point and Interval; Robustness in Statistics.

Bibliography

Bickel P, Doksum K 2000 *Mathematical Statistics*, 2nd edn. Prentice Hall, New York, Vol. 1

Blackwell D, Dubins L 1962 Merging of opinions with increasing information. *Annals of Mathematical Statistics* **33**: 882–6

Ferguson T S 1996 *A Course in Large Sample Theory*. Chapman & Hall, London

Fisher R A 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* **222**: 309–68

Fisher R A 1973 *Statistical Methods and Scientific Inference*, 3rd edn. Hafner Press, New York

Gauss C F 1809 *Theoria Motus Corporum Celestium*. Perthes, Hamburg, Germany

Girshick M A, Savage L J 1951 *Bayes and Minimax Estimates for Quadratic Loss Functions. Proceedings of the 2nd Berkeley Symposium of Mathematics, Statistics, and Probability*. University of California Press, Berkeley, CA

Hampel F, Ronchetti E, Rousseeuw P, Stahel W 1986 *Robust Statistics*. Wiley, New York

Jeffreys H 1939 *Theory of Probability*. Clarendon Press, Oxford, UK

Keynes J M 1921 *A Treatise on Probability*. Macmillan, London

LeCam L 1990 Maximum likelihood: An introduction. *International Statistical Review* **58**: 153–71

Lehmann E L 1986 *Testing Statistical Hypotheses*, 2nd edn. Springer, New York

Lehmann E L, Casella G 1998 *Theory of Point Estimation*, 2nd edn. Springer, New York

Linhart H, Zucchini W 1986 *Model Selection*. Wiley, New York

Neyman J 1937 Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society Series A* **236**: 333–80

Neyman J 1961 Silver Jubilee of my dispute with Fisher. *Journal of the Operations Research Society of Japan* **3**: 145–54

Neyman J, Pearson E S 1933 On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society Series A* **231**: 289–337

Shao J 1997 An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* **7**: 221–66

Staudte R G, Sheather S J 1990 *Robust Estimation and Testing*. Wiley, New York

Stein C 1956 Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proceedings of the 3rd Berkeley Symposium of Mathematics, Statistics, and Probability*. University of California Press, Vol. 1, pp. 187–95

Tukey J W 1960 Conclusions versus decisions. *Technology* **2**: 423–33

von Mises R 1928 *Wahrscheinlichkeit, Statistik and Wahrheit*. Springer, Wien, Austria

Wald A 1950 *Statistical Decision Functions*. Wiley, New York

Wilson E B 1952 *An Introduction to Scientific Research*. McGraw Hill, New York

P. J. Bickel and E. L. Lehmann

Frequentist Interpretation of Probability

If the outcome of an event is observed in a large number of independent repetitions of the event under roughly the same conditions, then it is a fact of the real world that the frequency of the outcome stabilizes as the number of repetitions increases. If another long sequence of such events is observed, the frequency of the outcome will typically be approximately the same as it was in the first sequence.

Unfortunately, the real world is not very tidy. For this reason it was necessary in the above statement to insert several weasel words. The use of 'roughly the same,' 'typically,' 'approximately,' and 'long sequence' make it clear that the stability phenomenon being described cannot be stated very precisely. A much clearer statement is possible within a mathematical model of this phenomenon. This discovery is due to Jacob Bernoulli who raised the following question.

It is well known, Bernoulli says in Part IV of his book *Ars Conjectandi* (published posthumously in 1713), that the degree to which the frequency of an observed event varies about the probability of the event decreases as the number of events increases. He goes on to say that an important question that has never been asked before concerns the behavior of this variability as the number n of events increases indefinitely. He envisages two possibilities.

(a) As n gets larger and larger, the variability eventually shrinks to zero, so that for sufficiently large n the frequency will essentially pinpoint the probability p of the outcome.

(b) Alternatively, it is conceivable that there is a positive lower bound below which the vari ability can never fall so that p will always be surrounded by a cloud of uncertainty, no matter how large a number of events we observe.

Bernoulli then proceeds to prove the law of large numbers, which shows that it is (a) rather than (b) that pertains. More precisely, he proves that for any $a > 0$

$$\Pr\left(\left|\frac{X}{n}-p\right| < a\right) \to 1 \quad \text{as} \quad n \to \infty \tag{1}$$

where X/n is the frequency under consideration.

It is easy to be misled into the belief that this theorem proves something about the behavior of frequencies in the real world. It does not. The result is only concerned with properties of the mathematical model. What it does show is that the behavior of frequencies in the model is mirrored in a way that is much neater and more precise, the very imprecise stability phenomenon stated in the first paragraph of this article.

In fact, a result for the model much more precise than (1) was obtained by De Moivre (1733). It gives the normal approximation

$$\Pr\left(\left|\frac{X}{n}-p\right| < c/\sqrt{n}\right) \to \int_{-c/\sqrt{pq}}^{c/\sqrt{pq}} \varphi(x)\,\mathrm{d}x \qquad (2)$$

where φ denotes the standard normal density. This is again a theorem in the model. The approximate corresponding real-world phenomenon can be seen, for example, by observing a quincunx, a mechanical device using balls falling through 'random' paths to generate a histogram.

De Moivre's result was given a far reaching generalization by Laplace (1810) in the central limit theorem (CLT) concerning the behavior of the average $\bar{X}$ of n identically, independently distributed random variable $X_1,\ldots, X_n$ with mean ξ and finite variance σ^2. It shows that

$$\Pr\left(\ |\bar{X}-\xi| < c/\sqrt{n}\ \right) \to \int_{-c/\sigma}^{c/\sigma} \varphi(x)\,\mathrm{d}x. \qquad (3)$$

This reduces to (2) when X takes on the values of 1 and 0 with probabilities p and q, respectively. The CLT formed the basis of most frequentist inference throughout the nineteenth century.

The first systematic discussion of the frequentist approach was given by Venn (1866), and an axiomatization based on frequencies in infinite random sequences (*Kollectives*) was attempted by von Mises (1928). Because of technical difficulties his concept of a random sequence was modified by Solomonoff (1964), Martin-Löf (1966), and Kolmogorov (1968), with the introduction of computational complexity. (An entirely different axiomatization based on events and their probabilities rather than random sequences was put forward by Kolmogorov in 1933, and has successfully served as a basis for both frequentist and subjective interpretations of probability). For more details on these different approaches see Barnett (1982).

The frequentist concept of probability described so far has met considerable criticism. One of the main objections is that it is not applicable to many situations to which one might want to apply probability assessments. To see this, consider the following three possibilities.

(a) An actual sequence of repetitions may be available; for example, a sequence of coin tosses or a sequence of independent measurements of the same quantity.

(b) A sequence of repetitions may be available in principle but not likely to be carried out in practice; for example, the polio experiment of 1954 involving a sample of over a million children.

(c) A unique event which by its very nature can never be replicated, such as the outcome of a particular historical event; for example, whether a particular president will survive an impeachment trial. The conditions of this experiment cannot be duplicated.

The frequentist concept of probability can be applied in cases (a) and (b) but not in the third situation. An alternative approach to probability which is applicable in all cases is the notion of probability as degree of belief; i.e., of a state of mind (for a discussion of this approach, see Robert (1994)). The inference methods based on these two interpretations of the meaning of probability are called frequentist and Bayesian, respectively.

Although frequentist probability is considered objective, it has the following subjective feature. Its impact on a particular person will differ from one person to another. One patient facing a surgical procedure with a 1 percent mortality rate will consider this a dire prospect and emphasize the possibility of a fatal outcome. Another will shrug it off as so rare as not to be worth worrying about.

There exists a class of situations in which both approaches will lead to the same probability assessment. Suppose there is complete symmetry between the various outcomes; for example, in random sampling which is performed so that the drawing favors no sample over another. Then we expect the frequencies of the various outcomes to be roughly the same and will also, in our beliefs, assign the same probability to each of them.

Let us now turn to a second criticism of frequentist probability. This concerns the difficulty of specifying what is meant by a repetition in the first sentence of this section. Consider once more the surgical procedure with 1 percent fatalities. This figure may represent the experience of thousands of cases, with the operation performed by different surgeons in different hospitals and—of course—on different patients. The rate of fatalities may vary from one hospital or surgeon to another and may, in particular, vary drastically with the condition, for example, the age and general health, of the patient.

Suppose a young woman requires this operation although her general health is very good. The frequency of a fatal outcome with patients sharing these characteristics may be much lower, and the 1 percent figure in that sense would be quite misleading for her. And yet she might be considered to have been obtained under 'roughly the same conditions,' namely to be drawn at random from the total population of persons

requiring this surgery. To obtain the most useful figures one should identify the most important variables, classify the cases accordingly (for example, young, middle-aged, old; male, female; etc.) and then provide the frequency for each class. They will, of course, be meaningful only for the classes which contain a reasonable number of cases.

1. A Terminology Note

A source of much confusion in the discussion of probability is the fact that 'probability' is used both as a mathematical term, i.e., as a concept in the mathematical model and also in everyday language when talking about real events. When reading about probability, it is important to be aware of these two meanings and to keep them distinct.

See also: Frequentist Inference; Probability: Formal; Probability: Interpretations; Statistical Methods, History of: Post-1900; Statistical Methods, History of: Pre-1900

Bibliography

Barnett V 1982 *Comparative Statistical Inference*, 2nd edn. Wiley, New York

De Moivre A 1733 1738, 1756. In: *Doctrine of Chance*, Millar, London, Reprinted by Chelsea Publishers, New York (1967), pp. 235–43 (1738), 243–54 (1754) 1878–1912

Kolmogorov A 1933 *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Ergebnisse der Mathematik, Vol. 2. Springer, Berlin

Kolmogorov A 1968 Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory* **14**: 662–4

Laplace P S *Oevres complètes de Laplace*. Gauthier-Villars, Paris

Martin-Löf P 1966 The definition of random sequences. *Information and Control* **9**: 602–19

Robert C P 1994 *The Bayesian Choice*. Springer, New York

Solomonoff R J 1964 The formal theory of inductive inference. *Information and Control* **7**: 224–54

Venn J 1866 *The Logic of Change*. Reprinted by Chelsea Publishers, New York, (1967)

von Mises R 1957 *Probability, Statistics and Truth*. Macmillan, New York

P. J. Bickel and E. L. Lehmann

Freud, Sigmund (1856–1939)

1. Family Background and Early Life

Sigmund Freud, neuroscientist turned founder of psychoanalysis, was born on May 6, 1856, in Freiberg, Moravia, now part of the Czech Republic. His father, Jakob (1815–96), a wool merchant, came from Galicia, and married Amalie (1835–1930, neé Nathansohn), 20 years younger, when he was 40. Jakob had two sons by a previous marriage, Emanuel and Philipp. Jakob was warm and affectionate, with a strong sense of humor, and Freud evidently admired and respected him. Amalie was a lively and attractive woman who was proud of Sigmund, her first-born.

The family multiplied quickly: five daughters and two more sons were born within 10 years. The first of these, Julius, died when 8 months old and Sigmund was aged 19 months. Both parents were Jews, with no religious affiliations, though a devout Catholic Nanny took Sigmund to Church. She left when he was two and a half. A nephew, John, a year older than Freud was very close to him: the two fought and played together. That childhood ambivalence, Freud reflected years later, must have had a profound effect on his character and on his ability to defend himself (Freud 1900).

Financially, times were hard. In 1859 the family moved to Leipzig, and in the following year to Vienna. Freud missed the beauty of the countryside round Freiberg. Although the children were well looked after, poverty was not assuaged by rapid family growth. By 1875 family fortunes improved: Amelie's family were helping, making possible a move to a larger house. Freud had a room of his own that served as a study.

Freud's memories of his childhood as revealed in his writings are without early detail, but one unforgettable recollection was of urinating in his parents' bedroom while they were present. Reprimanding him, his father said bluntly: 'The boy will come to nothing!' Freud's wounded ambitions were reflected recurrently in dreams in which he listed his achievements as if to say: 'You see! I have come to something!' (Freud 1900; see also Shengold 1993).

2. Education and Early Interests

Freud's schooling began with his mother, until his father took over before sending him to a private school. Freud learned rapidly, and was reading Shakespeare (whom he loved all his life) from the age of eight years. When nine years of age, he won a place at the Sperl Gymnasium, where he was a distinguished pupil, passing out, at 17, with the distinction '*summa cum laude*.' He became an accomplished linguist, with a firm foundation in Latin and Greek, a sound knowledge of Hebrew, fluency in English and French, and good Italian. He corresponded with an adolescent friend in Spanish. He became one of the greatest stylists in the German language (Kaufmann 1980). He studied at every opportunity. He loved the Arts, but took an early dislike to music, which he never felt able to appreciate. He was fond of walking, swimming, and skating. He liked travel, and made his first visit to England at the age of 19.

3. *Early Career and Scientific Outlook*

Freud (1925) said that curiosity and human concerns motivated his interest in medicine, but it was 'hearing Goethe's beautiful essay on Nature read aloud…that decided me to become a medical student.' He was attracted strongly by Darwin—a necessary step towards his own evolutionary approach to the history of the individual's mental development. His neurological work was affected profoundly by the evolutionary approach of John Hughlings Jackson, the founder of modern neurology. Furthermore, Freud's admiration of Goethe as scientist as well as writer was life long. Goethe's belief in the importance of observation and reflection was profound, and his meticulous examination of a single sheep's skull in 1790 led to his deduction that the vertebrate's skull developed from vertebrae. This type of detailed study—applied by Freud to the individual case—is the essence of poetic (or romantic) science that stands in contrast with (though not in opposition to) Newtonian (or classical) science. (Goethe's unfortunate opposition to Newton's theory of color does not vitiate this point.) In psychology, the corresponding distinction is between the ideographic (romantic) approach and the nomothetic (classical). Both methods have drawbacks and advantages; for Freud's psychology, in particular, the nomothetic approach has serious limitations. The distinctions have been discussed extensively by Kaufmann (1980) and by the founder of modern neuropsychology, Luria (1979). It is impossible fully to appreciate Freud's work outside this context.

Freud began his medical training in October 1873 and qualified MD in March 1881. The years were crowded: Freud dealt with the tedium of a year's compulsory military service by translating a book by John Stewart Mill (Jones 1953). He was influenced by many illustrious figures, none more important than Ernst Brücke (1819–92) head of the Institute of Physiology. Freud, as a research scholar, studied the histology of nerve cells, and his work led to a number of significant conclusions. After qualification, Freud continued working at the Institute, where his love for research had full scope. He had already found 'rest and full satisfaction' there (Freud 1925).

4. *Neuroscience*

But in 1882 Freud was still poor and, reluctantly, accepted Brücke's advice to practice medicine. He left the laboratory to start work in the General Hospital. He spent 6 months under Theodor Meynert (1833–1892), the Professor of Psychiatry, whose work and personality Freud greatly admired. His research at the Institute of Cerebral Anatomy led to important publications. He began to study nervous diseases, and quickly established a reputation as an outstanding diagnostician. He was still urgently short of money: he was by now engaged to Martha Bernays (1861–1951), and the need to earn enough to keep a wife and family was compelling.

In the spring of 1885 Freud was appointed Lecturer (*Dozent*) in Neuropathology. Brücke's influence resulted in the award of a valuable Traveling Bursary, and Freud fulfilled an old ambition to visit Paris and learn from Charcot at the Salpêtrière. His offer to translate a new volume of Charcot's lectures into German was accepted by the great man, and Freud was soon admitted to Charcot's inner circle. Charcot's lectures and demonstrations were outstanding, but Freud was astonished by his investigations into hysteria. Charcot demonstrated that hysterical symptoms were genuine; that hysteria was common in men as well as women; and that hysterical paralyses produced by hypnosis were identical with spontaneous attacks. Charcot met skepticism with patience and firmness: once, faced with an objection made on theoretical grounds, he replied: 'Theory is all very well, but it doesn't stop facts from existing.' Freud never forgot that remark.

The Paris visit changed the direction of Freud's career. Charcot had broadened his grasp of neurology, but the neuroses had captured his imagination. He left Paris at the end of February 1886, intent on further investigations of hysteria. But before he returned to Vienna he stayed for some weeks in Berlin to study nervous disorders in children, and this important detour led in time to the publication of a number of monographs. Freud's report on his visits, to his sponsors in Vienna, is a model of clarity and concise exposition (Freud 1956).

Indeed, Freud's interest in neuropathology by no means stopped abruptly. Important papers lay ahead: observations on hemianopsia in two very young children (Freud 1888) and his monograph on *Aphasia* (Freud 1891b) occupy a permanent place in the history of aphasiology. Freud had already abandoned the strict localisation theory of brain functioning—to which Meynert strongly adhered—in favor of the hierarchical model, with its concomitant theory of evolution and dissolution of the nervous system, advanced by Hughlings Jackson (e.g., 1884). Freud recognized the importance of these major advances, which were otherwise scarcely appreciated for another quarter of a century. With his young friend Oscar Rie he wrote a work on the unilateral cerebral palsies of children—a book still of repute in neuroscience (Freud 1891a). Freud's overall contributions to the field were massive, and he published around 200 papers.

On his return to Vienna, however, Freud was trying to build up a practice to support a family: he married Martha in September 1886, starting a life long partnership of unusual harmony and mutual support. Both were devoted to their six children: in order, Mathilde, Martin, Oliver, Ernst, Sophie, and Anna—the only one to follow Freud into psychoanalysis. The first 5 years of marriage covered the closing phase of

neurological work, and saw the growth of a friendship with the distinguished physician Joseph Breuer (1842–1925), with profound effects on Freud's career.

5. *The Birth of Psychoanalysis*

Between 1880 and 1882 Breuer had treated the now famous Frau Anna O. for severe hysterical symptoms, including multiple limb paralyses, disturbances of sight and speech, contractures, an intense cough, and anorexia. Accounts of the illness, its treatment, outcome, and significance for the foundations of psychoanalysis, and its effect on the Breuer–Freud friendship vary (e.g., Breuer 1882, Breuer and Freud 1895, Freud 1914, 1925, 1932, Jones 1953, Gay 1988, Ellenberger 1972, Hirschmüller 1989), but, at the risk of distortion by brevity, the present summary seems safe. Anna O. suffered from states of autohypnosis alternating with near-normality. When, under self-hypnosis, she recounted the history of one of her symptoms, recalling the details of its first appearance, Breuer was astonished to find that it disappeared. Under induced hypnosis other symptoms were treated in similar fashion, and the method was christened the 'talking cure.' Breuer told Freud about it in 1982, and during the next two years they discussed it repeatedly. But, in Paris, Charcot showed little interest in the story, and Freud allowed the matter to lapse.

Back in Vienna, Freud treated a number of hysterical patients by hypnotic suggestion. But 'recovery' was too often transient, and Freud's interest in Breuer's treatment of Anna O. revived. When her first symptom appeared, the patient was nursing her dying father, and her recall resulted in 'abreaction' (the release of pent-up emotion attached to the event). Adopting this 'cathartic method' over several years, Freud repeatedly confirmed Breuer's findings. The links with traumata were hitherto unconscious.

A preliminary communication by Breuer and Freud (1893) preceded the *Studies on Hysteria* (Freud 1895d). Breuer contributed a theoretical chapter and an account of Anna O; Freud, four case histories. Differences of opinion became a strain in 1894, and the collaboration ceased in 1896. Freud valued Breuer's achievement, but was thinking on different lines. Breuer believed that pathological processes arose during 'hypnoid states' (e.g., Anna O's autohypnosis). For Freud, this begged the question of how 'hypnoid states' arose. He considered that psychoneurosis involved psychic defences against unacceptable internal states of 'emotional excitation,' habitually sexual in kind. The actual neuroses, however, were states of anxiety arising from unrelieved sexual excitation linked with abstinence or *coitus interruptus*. Breuer could not accept these views; he had failed to find any sexual component in Anna O's disturbance. But Breuer had never recognized Anna O's strong attachment to him ('transference') as sexual; alarmed by its intensity, he brought the treatment to an end. Anna O. responded to this loss with a phantom pregnancy (Freud 1932). Breuer continued to see mentally ill patients, but never again used the cathartic method.

In time, Freud abandoned the use of hypnosis altogether. The power of transference (from the patient's unrecognized re-experience, in the treatment process, of infantile relationships) meant that excellent clinical progress could still be vitiated by disturbances in the doctor/patient relationship. That recognition profoundly affected the future development of psychoanalysis.

For a time, Freud used encouragement, persuasion, and similar measures as aids to treatment, but increasingly found them unnecessary. They were replaced slowly by the developing technique of free association, which took over completely by 1897–8. The patient lay on a couch (as before) with the analyst sitting behind him, but was now expected to try to say everything that entered his mind, irrespective of apparent irrelevance, distastefulness, or indiscretion. The analyst listened impartially, without giving more weight to any one utterance than he did to another. This major shift in clinical method had surprising results. Conflicting mental forces, undetected by hypnosis and related devices, were now revealed. Memories of striking internal and external mental experiences in early life had been denied access to consciousness because they had been 'alarming or painful or shameful by the standards of the subject's personality' (Freud 1925). Memory was actively opposed by the force of repression, experienced by the clinician as resistance to unwelcome material. Forces striving to break through this barrier stemmed from the drives, founded substantially in infantile sexuality.

6. *Psychoanalytic Career*

Until the spring of 1897 (Jones 1953) Freud believed his female adult patients whose associations pointed to seduction in childhood, usually by the father, as the traumatic root of their illness. But what at first presented as memory was found increasingly to be fantasy. Seduction in childhood undoubtedly occurred, but could no longer be invoked as the root of all neurosis. As it stood, Freud's theory was wrong. This devastating conclusion nonetheless pointed to the importance of fantasies in mental life, and to the realization that fantasy that is dynamically unconscious, that is, inaccessible to introspection, had all the force of reality since it could not be tested in the light of objective experience. A second, fateful, event, did more to foster psychoanalytic thinking than anything else: Freud courageously (and painfully) undertook his self-analysis, beginning in 1897 and continuing systematically through and beyond the writing of *The Interpretation of Dreams* (1900). With interruptions, it was life long.

Freud's letters to his friend Wilhelm Fliess (1858–1928), an ear, nose, and throat specialist living in

Berlin, provide a unique source of information about these events. Never intended for publication, they were rescued by Princess Marie Bonaparte of Greece (the letters from Fliess have not survived). Starting in 1887, they set out in astonishing detail Freud's reflections, doubts, and formulations as they arose, and can be followed until after the turn of the nineteenth century. An edited version (Freud 1954) was followed by the complete letters (Masson 1985). The demise of the seduction theory and Freud's comments on his self-analysis are described in both.

The self-analysis was crucial for the development of basic psychoanalytic concepts. The richness and complexity of infantile life, including its sexuality, was not fully appreciated until well after *The Interpretation of Dreams* saw the light of day. The Oedipus complex, for example, is described in that book, but its fuller significance was underlined only in footnotes added much later. But Freud regarded the dream book as his single most important work. With copious illustrations from his self-analysis and clinical experience with patients, the study goes beyond the meaning of dreams and approaches the baffling problem of how the mind works.

The theory is involved, but a few key points may be mentioned. The function of dreaming is the preservation of sleep. Stimuli arising from outside the organism (noises, the pressure of bedclothes) or from within (pain, excitation), could often be contained when switched into nonmotor channels (dreaming). The formation of the dream is more complex. In normal waking life, accessible mental content is either conscious at any given moment, or is preconscious—capable of reaching awareness with varying degrees of difficulty. But some mental content is dynamically unconscious—barred from reaching the preconscious by repression. During sleep, this barrier is relaxed, and a partial regression to earlier forms of thinking allows normally unacceptable childhood wishes to find expression, though only in disguise. Even in sleep a censorship is partially operative, and the result is a distorted representation of infantile wishes that gives the dream its seemingly bizarre character. Any semblance of coherence is imposed during waking and recollection: the repression barrier is restored and the dream is (unwittingly) clothed in a measure of rationality.

The recollected (manifest) dream content arises from several sources. Stimuli that threaten sleep are the most obvious, but events and experiences of the day before (day residues) are also used as building blocks. These have no intrinsic importance: they simply provide convenient representations for forbidden wishes. But emerging wishes are themselves distorted by primitive thinking before they are disguised further by the censorship. For Freud, the dream was a hallucinatory wish fulfillment, though generally far removed from the wishes of everyday appearance.

Do dreams provoking anxiety or portraying punishment also conceal wishes? Freud argued that, even in sleep, a wish can be so strong that its forbidden character can still be recognized and occasion distressing anxiety. In punishment dreams, the transgression portrayed brings with it a wish for punishment. This became clearer still when Freud, later, had conceptualized an unconscious sense of guilt.

The model of the mind outlined in the dream book was the foundation for all Freud's later theoretical formulations. Since it was written during his self-analysis, his pride in it also reflected a new self-confidence. Jones (1955) claims that, at the age of 45, Freud had reached full maturity: intellectual development, always precocious, had not hitherto been matched emotionally. For many years there had been 'continual periods of restlessness, uncertainty, instability and even more definitely neurotic disturbance.' A deep self-assurance had been hidden behind 'strange feelings of inferiority, even in the intellectual sphere.' He had over-idealized men who were, certainly, important to him; of these, Breuer had already withdrawn his support, and the relationship with Fliess was loosening. Self-analysis had revealed the neurotic nature of the tie.

Intellectual loneliness did not diminish productivity. *The Psychopathology of Everyday Life* (Freud 1901) dealt with unconscious motivations in forgetting, slips of the tongue and pen, as well as some 'accidents.' The *Fragment* of Freud's analysis of 'Dora' (1905) described her treatment in 1900. *Three Essays on the Theory of Sexuality* (Freud 1905) met with ferocity hard to imagine today. Freud had widened the concept of sexuality: all existing definitions failed to fit the facts and could not, for example, account for the sexual perversions. Sexuality was now understood within a much wider, developmental perspective. It was considered diphasic: development from infancy to roughly the age of five, centered around bodily erotogenic zones, was followed by a period of comparative quiescence (latency) before a second stage of sexual efflorescence started around puberty. *Jokes and their Relation to the Unconscious* (Freud 1905) arose out of the work on the dream book.

7. *The Growth of the Discipline*

By the time these works appeared Freud was no longer intellectually isolated. Already, in 1902, two young Viennese doctors, Kahane and Reitler, and, more famously, Stekel and Adler were intrigued by psychoanalysis. Freud invited them to meet him for discussions; they gathered together on Wednesdays as 'The Psychological Wednesday Society.' This was the nucleus of an expanding group renamed, in 1908, the Vienna Psychoanalytical Society. By then much had happened. In 1904 Eugene Bleuler, Professor of Psychiatry in Zurich, wrote to inform Freud that he and his staff, most notably his Chief Assistant, C. G.

Jung, had been making use of psychoanalytic ideas for the past two years. (In particular, Jung had applied them in his word-association tests.) By 1906 Freud and Jung were corresponding, and in 1907 Jung visited Freud in Vienna. The friendship grew rapidly and when, in 1910, the International Psychoanalytic Association was formed, Freud asked Jung to be President. Jung's bearing, intellect, qualities of leadership, psychiatric background, and dedication seemed suitable qualifications. Freud was attracted by his personality and regarded him as his 'son and heir.' Jung was greatly impressed by Freud's penetrating intellect. The rise and fall of their relationship is chronicled in their correspondence (McGuire 1974).

Otto Rank had joined the group in 1906; and Karl Abraham, a major figure in the psychoanalytic movement, had long been fascinated by Freud's work when he visited him in 1907. Again, the correspondence between the two gives a vivid picture of their relationship; it continues until Abraham's premature death in 1925 (Abraham and Freud E 1965). In 1907, too, Sandor Ferenczi visited from Budapest, establishing himself the following year as a major figure in the discipline. With the visit of Ernest Jones in 1908, and Hans Sachs in 1910, Freud had a close circle of friends and collaborators.

In 1908, Jung organized an international conference in Salzburg. Freud's work was now discussed widely, in Europe, the USA, and Australia. In 1909 Freud lectured at Clark University, Massachusetts, accompanied by Ferenczi, Jones, and Jung. In 1910 a second international congress was held at Nuremberg, organized by Jung. But internal dissent soon surfaced. In 1911 Adler defected: to Freud's dismay, he refused to try to resolve their conceptual differences. Stekel, gifted, charming, but scientifically unreliable (Jones 1955) left about the same time. Although Freud disliked leadership, he was a fatherly figure for his young colleagues, and bickering and rivalry, especially among the Viennese (cf. Jones 1955) threatened stability. Freud's obvious preference for Jung was resented. But by 1912 Jung was showing signs of disaffection. Many Swiss psychiatrists denounced Freud's theory of sexuality, and, during a course of lectures in the United States Jung omitted any reference to it. Personal relations between the two men cooled, but it was not until 1914 that Jung left the movement on the grounds of incompatible views. The story of the rupture is complex: the Freud/Jung letters and the account by Jones (1955) repay study.

Freud's continuing productivity defied these dissensions. He addressed the problem of the sexual enlightenment of children (Freud 1907), discussed children's sexual theories (Freud 1908), and considered the relationship of the character traits of orderliness, parsimony, and obstinacy to anal eroticism—a paper greeted with indignation when it appeared (Freud 1908). He tried to show that psychoanalysis could contribute to the understanding of art in discussions of Jensen's *Gradiva* (Freud 1907) and of the relationship between creative writing and day dreaming (Freud 1908). He wrote many other papers in these industrious days.

Two extensive clinical studies appeared in 1909. The first (Freud 1909) describes the very first analysis of a child, though one undertaken by proxy. The successful treatment of 'Little Hans,' who had developed a phobia of horses, was carried out by his father, under Freud's supervision. Hitherto, the childhood roots of neurosis in adults had been understood through a process of reconstruction; with Little Hans those sources could be identified at the time they were operative. Freud believed that the analysis could not be conducted by a stranger, and that the affection and authority of the father (in this case, a physician, too) made treatment possible. Child analysis as a speciality was yet unborn. It was many years before the possibility of directly analysing young children, under favorable conditions, was recognized.

The second patient suffered from an obsessional neurosis (Freud 1909). The symptoms included fears that harm would befall his father and a woman he admired; compelling impulses, one of which was to cut his throat with a razor; and self-imposed irrational prohibitions. He was horrified but attracted by an account of a punishment in which rats ate their way up the victim's rectum. The unconscious sadism underlying his 'fascination' with the torture was among the features on which the analysis cast light. A deeper psychoanalytic formulation of obsessional neurosis emerged from the treatment. Later critics of this case report, and that of Dora, have suggested that Freud neglected the role of the mother in the genesis of the disorders. But Dora's treatment was brief, unfinished, and took place when clinical analysis was still in its infancy. In the case of the Rat Man, the criticism, when sound, is largely based on hindsight.

Another case study (Freud 1911) concerned a subject whom Freud had never seen. Dr. Schreber suffered from a paranoid state, and was under care for 6 years. He later wrote his *Memoirs of My Nervous Illness*, though Freud read them only in 1910. The illness fell into two phases: distressing delusions of persecution preceded delusions of grandiosity. Freud made an ingenious and fascinating reconstruction of the course of the illness. His work stimulated much clinical research. Psychoanalytic theories of psychotic illness fall into two broad groups: either it is part of a neurosis–psychosis continuum, or, unlike neurosis, it involves the loss of higher psychic functions and the re-emergence of more primitive ones, hitherto kept in check by the lost capacities. Many claim that Freud supported the first (unitary) theory; others the second (discontinuity theory). The controversy deserves close attention and is summarized elsewhere (Yorke et al. 1989).

In 1910 Freud (1910) published a study of a childhood memory of Leonardo da Vinci, and the first

of three essays on the psychology of love. In 1911, the first of a series on psychoanalytic technique, and from 1910 to 1912 some short papers on theory that paved the way for later and larger works. An excursion into social anthropology, *Totem and Taboo*, was written between 1912 and 1913. Before war broke out in the summer of 1914, Freud's works included a history of the psychoanalytic movement (Freud 1914), giving his version of the disagreements with Adler and Jung; a fascinating paper on the *Moses* of Michelangelo (Freud 1914); a discussion of the scientific status of psychoanalysis (Freud 1913); and a wide ranging paper on narcissism from which so many ideas were later developed (Freud 1914). These were productive years indeed.

During the war Freud was anxious about two of his sons involved in it. Martin fought in Galicia and Russia; Ernst against Italy after that country entered the conflict. Martin was decorated for gallantry. Medical reasons prevented Oliver's acceptance by the army. Optimism about the outcome diminished after 1915. In 1916 the food shortage began to bite and got worse as the war continued. Freud had few analytic patients. Contact between analysts faced obstacles, some severe, even in the central powers.

For all the difficulties, Freud continued to write. He wrote 12 papers on the theoretical basis of psychoanalysis between 1914 and 1917: only five were published, and a sixth manuscript was found in 1983. There is no trace of the others. *Thoughts for the Times on War and Death* (Freud 1915) anticipated the correspondence with Einstein (Freud 1933) *Why War*? At the University, in the winter terms of 1915–6, and 1916–7, Freud gave a total of 28 lectures on psychoanalysis, and wrote them down, from memory, later (Freud 1916–7). Important changes were made in the *New Introductory Lectures* (Freud 1933), but, taken together, the two series offer an excellent entry into Freud's psychological thinking. The most extensive of Freud's case histories (the *Wolf Man*), written in 1914, was published much later (Freud 1918). In 1919 the baffling problem of masochism (sexual pleasure in pain), and a psychoanalytic look at 'the uncanny' (Freud 1919) were among the many subjects Freud addressed.

The return of many old colleagues to Vienna in 1919 ended wartime isolation. For some time life was hard, but psychoanalysis continued to grow worldwide. In 1919 Martha Freud contracted influenza leading to pneumonia and was ill for many months; the following year Sophie Freud died from a similar cause. Freud grieved deeply.

In 1923 Freud's heavy cigar smoking seemed responsible for a growth affecting the jaw and palate, removed surgically. This was the first of many operations for cancer that Freud survived over the next 16 years, bearing discomfort and pain, often severe, with fortitude. But he continued to work. In that same year, he set out major changes in psychoanalytic theory (Freud 1923). He had already felt obliged to give aggression (more accurately, destructiveness) the same status as libido. For clinical and theoretical reasons the structure of the mental apparatus was redefined. The ego was now the executive mental agency, holding a balance between the conflicting demands of reality, the id, and the superego. The id strove for immediate drive gratification; its mentation followed the primitive primary process; and its content was dynamically unconscious. But it intruded into an unconscious part of the ego where it encountered checks and controls against inappropriate discharge. The superego was thought of as a greatly extended conscience, based on internalized but modified parental aims and proscriptions, but a substantial part of its operations were unconscious. Its functions included that of the earlier censor. With the aggressive drives of the id partly at its disposal, it put pressure on the ego to withstand those inner urges that ran counter to its prohibitions, aims, and ideals. It was responsible for the sense of guilt, conscious, and unconscious. It must be added that higher mental functions such as conscious thinking and reflection (secondary process), reality testing, and judgment were functions of the ego.

Freud continued to elaborate and modify this model (it should not be reified). He made radical changes in his theory of anxiety (Freud 1926); extended his work on masochism (Freud 1924); and in the field of social psychology added to his earlier study of the psychology of groups (Freud 1921) the far reaching *Civilization and its Discontents* (Freud 1930). The work of his female colleagues and his own clinical experience led him to re-examine the many problems of female sexuality (Freud 1931, 1933). This work aroused opposition, particularly among feminists: Eissler (1977) countered some of these criticisms, but there have been many subsequent developments. It is worth remembering that Freud's close friendships included many women (see, e.g., Freud and Andreas-Salome 1972); and many found his work did much to reduce guilt about sexuality. In *Analysis Terminable and Interminable* Freud (1937) emphasized the therapeutic limitations of psychoanalysis, but he was apprehensive about the possible reception of *Moses and Monotheism* (Freud 1939), which extended his inquiries into early social organization (*Totem and Taboo* 1912–3, *Group Psychology* 1921).

8. *Exile: England*

In 1938 the Nazis entered Vienna. Freud's books had already been burned in Germany (Freud thought this more advanced than burning him). Through the good offices of Princess Marie and Ernest Jones Freud, together with Martha, their daughter Anna, and others, were allowed to leave for London. Freud was quite ill and tired by the journey, but was pleased by the warm reception that awaited him. The British

Psychoanalytical Society gave him a house at 21 Maresfield Gardens in Hampstead. Anna Freud was his devoted nurse. His remarkable *Outline of Psycho-Analysis* (Freud 1940) was written there. It was there, too, that that death ended his considerable pain and suffering on September 23, 1939. His ashes remain at Golder's Green Crematorium.

Of his many honors, he particularly valued the award of the Goethe Prize for Literature (1930) and his Corresponding Membership of the Royal Society (1936).

9. *Freud's Impact on Twentieth-century Thought and Culture*

There is still a widespread view that Freud's impact on twentieth-century thought, on the arts, sciences, social attitudes, and beliefs, has been far-reaching, even immense. Almost every intellectual discipline, it is said, has been influenced by his ideas. The importance of unconscious motivation is acknowledged widely. Furthermore, many of his concepts—such as the 'Freudian slip' of pen or tongue—have passed into the common language of everyday social intercourse; and his views on sexuality have had a liberating impact on social attitudes and broadened sexual mores.

This view is by no means universal. Freud is undeniably a household name, but more than any other major figure, is the subject of widely varying and conflicting claims. These differences may arise in part from the nature of his subject, for it may be said, with a certain truth, that every man is his own psychologist. They may also arise from an investigative method difficult to understand without personal acquaintance. Freud's neuroscience aside, his principle field of inquiry was subjective and internal, and not readily open to methods appropriate to objective and external research. Many academic psychiatrists and psychologists consider that both Freud's clinical method and his theoretical constructs lack empirical support: in this, they generally equate empiricism with mensuration.

This fact stands in the way of dispassionate assessment even for those actively and intelligently involved in debate. The casual bystander, faced with controversy on a substantial scale, may be simply bemused or follow an opinion in line with personal appeal. And if the intelligent and industrious layman turns to Freud's own writings, he will still encounter significant obstacles. Freud's ideas cannot be picked up piecemeal, but comprehensive reading is difficult without informed guidance. Moreover, reliable introductory works are hard to come by. So it seems unlikely that the forces behind the 'Freudian slip,' and the distinctions between the preconscious and the unconscious, are understood widely. (The term 'subconscious' is more likely to be heard—a word Freud never used except by way of criticism.) Finally, while the liberalization of sexual attitudes may owe something to Freud, social, economic, and political forces are surely of greater importance. These include the availability of reliable contraception.

Indeed, Freud's views on sexuality, when they are not bowdlerized, are misunderstood widely. But the wider point to be made is this. Most people, in all sections of society, who have any interest in Freud, do not know what he really said. They only know what they think he said. And that is just as true for most critics and commentators who seemingly write with authority, including psychologists, 'hard' scientists, sociologists and art critics, whether the views they express are positive or negative. The viewpoint expressed here may at first seem novel. It was presented in extended form at a meeting of the *Psychoanalytic Heritage Foundation* on October 6, 1992, when it became evident that many Freud scholars were thinking on similar lines. At the recent Conference in London (see below), Douglas Watts stated from the platform that psychoanalysts needed to do more to make the general public more familiar with Freud. 'At present all they have is a caricature.'

For reasons that are not dissimilar, though highly complex, even psychoanalysts have widely diverse approaches to both theory and clinical practice and to the weight they attach to basic Freudian principles (Yorke 1995). Psychoanalysis is avowedly pluralistic (Wallerstein 1988, 1990). And although psychotherapy is often said to derive from the 'talking cure,' similar considerations apply.

In a few instances, Freud's influence, directly or indirectly, on other disciplines is evident, though localized. Disciplines not already mentioned include: education (e.g., many members of the Montessori movement); criminology (The Institute for the Study and Treatment of Delinquency in London); family law in relation to children (e.g., liaison between the Child Study Center and the Law School at Yale University). Properly understood, Freud's work and its true derivatives is capable of fostering advances in many intellectual and sociological fields.

It is pleasing to end on a positive note—in relation to Freud's first love, neuroscience. Freud always hoped that the underlying basis for mental activity in terms of the functioning brain would, one day, yield to investigation. For the greater part of a century that hope seemed slim indeed. But the last few years have brought startling developments. Solms (1997), for example, in a fascinating series of experiments, showed convincingly that the neuropsychology of dreaming was very much in line with the argument set out in *The Interpretation of Dreams*. Work on the neuropsychology of affects, for example, attracted interest from an increasing number of psychoanalysts, and in 1999 a new scholarly journal, *Neuro-psychoanalysis*, was founded. In the first issue, Solms and Nersessian (1999) demonstrated, in a wide-ranging survey, the extensive nature of Freud's contributions to affect

theory. On July 21–3, 2000, the first annual Neuropsychoanalysis Conference was held in London entitled *Neuropsychology and Psychoanalytic Persepectives on Emotions*, addressed by the world authorities Jaak Panksepp, Antonio Damasio, and Mark Solms, and introduced by Oliver Sacks. The meeting was a heartening success, and the compatibility between the findings of the two disciplines was seen to be remarkable.

All references to Freud, other than those in the bibliography, are to the *Standard Edition of the Complete Psychological Works of Sigmund Freud* (1953–74) Strachey J (ed.) 24 Vols. Hogarth, London. For specific volumes and page numbers see Vol. 24.

See also: Adorno, Theodor W (1903–69); Allport, Gordon W (1897–1967); Bleuler, Eugen (1857–1939); Bowlby, John (1907–90); Defense Mechanisms; Dreaming, Neural Basis of; Ego Psychology and Psychoanalysis; Erikson, Erik Homburger (1902–94); Foucault, Michel (1926–84); Hypnosis, Psychiatry of; Hysteria; Janet, Pierre (1859–1947); Jung, Carl Gustav (1875–1961); Mead, George Herbert (1863–1931); Mental Health and Normality; Nosology in Psychiatry; Pavlov, Ivan Petrovich (1849–1936); Personality Development in Childhood; Personality Theories; Psychiatry, History of; Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis: Current Status; Psychoanalysis, History of; Psychoanalysis in Clinical Psychology; Psychoanalysis in Sociology; Psychoanalysis: Overview; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychology: Historical and Cultural Perspectives; Sexual Perversions (Paraphilias)

Bibliography

Abraham H C, Freud E L (eds.) 1965 *A Psycho-analytic Dialogue: The Letters of Sigmund Freud and Karl Abraham*. Hogarth, London

Breuer J 1882 Report to Bellevue Sanatorium, Kreuzlingen, on Bertha Pappenheim. In: Hirschmüller A (ed.) (1989) *The Life and Work of Joseph Breuer*. New York University Press, New York

Breuer J, Freud S 1893 *On the Psychical Mechanism of Hysterical Phenomena: A Preliminary Communication*. Hogarth, London

Breuer J, Freud S 1895 *Studies on Hysteria*. Hogarth, London

Eissler K R 1977 Comments on penis envy and orgasm in women. *Psychoanalytic Study of the Child* **32**: 29–83

Ellenberger H F 1972 The story of 'Anna O': A critical review with new data. *Journal of the History of the Behavioral Sciences* **8**: 267–79

Freud S 1891b *Zur Auffassung der Aphasien* [trans. Stengel E as *On Aphasia* 1953]. Imago, London

Freud S 1932 Letter to Stefan Zweig, June 2, 1932. In: Freud E L (ed.) *Letters of Sigmund Freud, 1873–1939*. Hogarth, London

Freud S 1954 The Origins of Psycho-analysis: Letters to Wilhelm Fliess, drafts and notes: 1887–1902. (Bonaparte M, Freud A, Kris E (eds.) Imago, London

Freud S 1987 A Phylogenetic Fantasy. (Grubrich-Simitis I (ed.) Belknapp, Cambridge, MA

Freud S, Andreas-Salome L 1972 *Letters* [trans. Robson-Scott W, Robson-Scott E]. Hogarth, London

Gay P 1988 *Freud: A Life for Our Time*. Dent, London and Melbourne

Hirschmuller A 1989 *The Life and Work of Joseph Breuer: Physiology and Psychoanalysis*. New York University Press, New York

Jackson J H 1884 Evolution and dissolution of the nervous system. In: Taylor J (ed.) *Selected Writings of John Hughlings Jackson*. Basic Books, New York

Jones E 1953 *Sigmund Freud: Life and Work*. Hogarth, London, Vol. 1

Jones E 1955 *Sigmund Freud: Life and Work*. Hogarth, London, Vol. 2

Jones E 1957 *Sigmund Freud: Life and Work*. Hogarth, London, Vol. 3

Kaufmann W 1980 *Discovering the Mind*, 3 Vols. McGraw-Hill, New York

Luria A R 1979 *The Making of Mind: A Personal Account of Soviet Psychology*. Harvard University Press, Cambridge, MA

McGuire W (ed.) 1974 *The Freud/Jung Letters* [trans. Manheim R, Hull R E C]. Hogarth & Routledge, London

Masson J M (ed.) 1985 *The Complete Letters of Sigmund Freud to Wilhelm Fliess: 1887–1904*. Belknap Press of Harvard University Press, Cambridge, MA

Shengold L 1993 *'The Boy Will Come to Nothing!' Freud's Ego Ideal and Freud as Ego Ideal*. Yale University Press, New Haven, CT

Solms M 1997 *The Neuropsychology of Dreaming: A Clinico-Anatomical Study*. Erlbaum, Mahwah, NJ

Solms M, Nercessian E 1999 Freud's theory of affect: Questions for neuroscience. *Neuro-psychoanalysis* **1**: 5–14

Wallerstein R S 1988 One psychoanalysis or many? *International Journal of Psycho-analysis* **69**: 5–21

Wallerstein R S 1990 Psychoanalysis: The common ground. *International Journal of Psycho-analysis* **71**: 3–20

Yorke C 1992 *Freud's impact on the twentieth century*. Paper read to the Freud Literary Heritage Foundation, New York, October 6

Yorke C 1995 Freud's psychology: Can it survive? *Psychoanalytic Study of the Child* **50**: 3–31

Yorke C, Wiseberg S, Freeman T 1989 *Development and Psychopathology: Studies in Psychoanalytic Psychiatry*. Yale University Press, New Haven, CT

C. Yorke

Friendship, Anthropology of

Friendship is an informal social relationship. Contrary to kinship, it is based on choice and voluntariness: friends are sought and must be won. Friendship is an acquired not an ascribed status. The relationship is often viewed as affective. However, the 'emotional content' of friendship can vary strongly (Paine 1969, p. 507). Mutuality, concerning interest in the relationship, appears to be one important requirement.

In fact, in some societies there exist asymmetrical friendships, sometimes being difficult to distinguish from patronage. Friendship is based on sharing: besides matters of material value friends share their time, their problems, plans, hopes, and thoughts. Therefore, trust and loyalty are basic requirements for friendship (Kennedy 1986, 128 ff.). But this aspect also involves dangers: if friends share secrets, and the friendship is eventually broken off, it can lead to gossip and enmity. Some authors note that generally no fixed role expectations are attached to friendship. But in fact many societies do have well-defined expectations.

From the perspective of comparative social and cultural anthropology it is obvious that besides common core aspects, notions of friendship are highly variable. To this day there is a lack of data concerning the different meanings friends and friendship can have in different societies, and even within one society. Most languages have many terms for friends (in English among others: *pal*, *chum*, *buddy*, *bosom friend*, *old friend/chap*, *acquaintance*, etc.), which differentiate between kinds and meanings of social relationships called 'friendship.' One also finds different concepts of friendship within one society relating to sex, age, social status, as well as the different places or contexts in which friendship takes place, e.g., neighborhood, school, or place of work. The 'growth' of friendship has also been an underestimated aspect in research. A friendship does not only change in the course of time, it also changes due to the period of its existence: after 20 years a friendship, like all other emotional relationships, is no longer the same as it was in the beginning. As Sarah Uhl (1991, p. 90) states, friendship is a type of 'noninstitutionalized institution' distinguished from blood-brotherhood and fictive kinship—overlooking the fact that with a new stage in a life cycle, friendships may be transformed to other more institutionalized relationships like co-parenthood.

1. *Anthropological Interest in Friendship*

Uhl (1991) states that research on friendship is difficult due to the fact that despite its being a public phenomenon, friendship is viewed as private. But the same can be said of marriage, and marriage has been one of the most important topics in anthropology since its beginnings. More probably anthropologists have focused too strongly on kinship (see *Kinship in Anthropology*), thus becoming unaware of sometimes even overlapping concepts of friendship.

Until now most empirical research has been conducted in industrialized societies, and basically there are only few studies by cultural anthropologists. If anthropologists have studied friendship at all, they have concentrated on formal relationships with well-defined mutual obligations such as ritual kinship, exchange, trade, or working relationships. These anthropological studies—most of them conducted in Southern Europe—overemphasized male friendships and neglected friendships between females (Boissevain 1974, Gilmore 1975). For a long time it was supposed that adult women in particular form their most important social relationships inside the family and with relatives. Different forms of friendship between men are on the other hand an important part of public life. These factors have contributed to the lack of research concerning friendship between women, with the exception of some more recent studies (e.g., Kennedy 1986).

One finds very little information on meanings of friendship in nonindustrial societies. One exception is Thomas Kiefer's work, who in 1968 published a major article on institutionalized friendship and warfare among the Tausug of Jolo, in the Philippines. The Tausug perceive strangers and all people not related by kinship as potential enemies. In this context friendship—formally sworn to by oath on the Koran—is very important for forming alliances between kin groups, as well as, on a higher regional level, between local leaders. These can be former or—should the oath be broken—future enemies. This formal friendship in Tausug society derives its meaning from enmity—friends are allies against common enemies (Kiefer 1968). However, Kiefer's study once again focuses on highly formalized male friendship patterns. Institutionalized and formalized friendship patterns, trade and exchange partnerships, like the bond friendship' (*tau soa*) in Tikopia described by Firth (1967), are still today given priority in studies by cultural anthropologists. On the level of intercultural comparison (including more traditional societies), it would be of importance to investigate the assumed declining significance of kinship and the possibly growing importance of friendship.

2. *Themes and Methods*

Within the realm of anthropological research different dimensions of friendship have been stressed. Eric R. Wolf differentiates between 'emotional' and 'instrumental relations,' which both satisfy a deficit of some kind in each participant of the dyad (1966, 10 ff). The labels emotional/instrumental are somewhat misleading, because both types have a quite instrumental quality. Wolf's approach links friendship with wider societal forms: societies produce different deficits, which different types of friendship respond to. His conception is a useful starting point for further research, embedding friendship in a wider theoretical frame. In that way it differs from Robert Paine's (1969) above-mentioned conception of an emotional, more inward-turning modern middle-class friendship.

Jacobson's approach to research on friendship emphasizes the situational aspect. He is not only concerned with constant friendship and its character-

istics but 'with the labelling process itself, that is, with the situations in which a person gives and takes away the label of "friend".' (Jacobson 1975, p. 225). Both friendship and affiliation with an ethnic group can change depending on the situation. In combination with an examination of group identities, research on friendship could lead to important results in the investigation of interethnic relationships. Friends can be made across ethnic boundaries; besides binding themselves by marriage, people can form alliances between different ethnic groups. An example of the discussion of interethnic relationships in folklore about friendship is given in a story about a Kipsigis and a Masai, who exchange honey and arrows, cheating each other in the process (Dundes 1971, pp. 176–7). Dundes explains that the topic of making and breaking friendship in African societies is a common motif in narratives. In a story about a crippled and a blind person, for example, who join forces during an attack by enemies, the latter takes the first on his back and runs, while the crippled person directs him. No matter whether the failing of a friendship or its advantages are the central theme, exchange is always the constituent core of the described relationship. The different motifs in African narratives show the conflict between social norms and actual behavior. This discrepancy between a value system and behavior is another aspect of friendship that has not yet been thoroughly studied.

Reina (1959) discusses two concepts of friendship, existing parallel to one another among Indians and Ladinos in Guatemala. The Ladino term *cuello* means taking up many relationships and alliances, which can change under different circumstances. The Indian concept of friendship implies having only very few friends, who are made during adolescence. This example shows that divergent notions of friendship can prevent close relationships between members of different ethnic groups. But without more detailed descriptions of intraethnic concepts of friendship, studies about cultural differences and interethnic friendship are impossible.

In early network analysis the issue of friendship played an important role (Bott 1957). Network analyses seemed to be appropriate for research about weak institutionalized social relationships in complex societies. Friendship is one possible content of such social relations, but until now research has been restricted mostly to Western industrialized societies (e.g., Allan 1989, Fischer 1982). Analysis of networks of friends in more 'traditional' societies must be preceded by studies which describe who is referred to as a friend, and what kind of different categories of friends exist within a given cultural context. As shown above, it cannot be taken for granted that concepts of anthropologists and informants are congruent. With a better knowledge of the meanings and terms attached to the notion of friendship, network analysis and other quantitative methods might also be applied in interethnic situations. From the interplay between participation in everyday life, an evaluation of one's own experiences, and a systematic investigation, a more differentiated picture of the ambiguous term 'friendship' could emerge.

3. *Fieldwork and Friendship*

Friendship is not only an important research topic, it is also an important relationship during fieldwork (see *Fieldwork in Social and Cultural Anthropology*). The situation regarding friendship between anthropologists and people with whom they live, often for long periods of time, is hardly documented, even though it is quite probable that these are very relevant personal experiences during fieldwork and in some cases they are even the precondition of fieldwork. Exceptions being the articles by Friedman Hansen (1976), Hendry (1992), and Reina (1959). Friendship formed in the course of fieldwork has neither been treated very honestly, nor has it been dealt with systematically in anthropology to this day. Most anthropologists note having made many friends during their fieldwork. In publications it is obligatory to thank all 'friends,' without whom research would have been impossible. In most cases the nature of this friendship is not described, nor is the significance it had for the fieldwork. Therefore, the question of how anthropologists cope with unfamiliar concepts of friendship is left open. Most studies do not make explicit what is meant by friendship, and few authors relate whether friendship with informants is called so from the perspective of the fieldworker or from the emic view. Other intercultural forms of friendship—those between anthropologists and informants are only one special case—have gone unstudied.

In the literature on fieldwork, friendship is sometimes depicted from a very one-sided angle: 'Informants become friends. This creates another dilemma, for it leads to a very instrumental, and often dishonest, approach to friendship. The interest of science demands that you milk your informants. You do this by making them your friends, by exchanging confidences, giving presents, talking for hours about subjects that bore you. Are these friends or scientific objects? The subjects themselves often realize that they are being used. For some fieldworkers the scientific ends justify the means. They have no moral problem. For many others, however, the dilemma remains.' (Boissevain 1985, p. 273)

Other authors warn against making friends with informants on the grounds of possible role conflicts. Spradley, for example, advises anthropologists, repeating the same questions over and over again on their endless quest for explanations, to refrain from bothering their friends with this boring task. Apart from this, however, he has no objection to making friends during fieldwork (Spradley 1979, p. 28).

An exceptionally good account of friendship in the course of fieldwork has been published by Joy Hendry (1992). In detail she describes the changes within the relationship to her Japanese friend that spanned nearly two decades. She also compares, at least marginally, to what extent concepts of who is related to as a friend, as well as expectations regarding how friends should relate to one another, correspond in Great Britain and Japan. Hendry mentions that in Japan it is not possible to bring inequality in line with friendship. As long as she did research in rural Japan, similarities and equality dominated the relationship to her friend Sachiko, who also did not come from a rural area. Later, influenced by their respective cultures, their lives developed differently from one another. Apart from individual changes due to the respective transformations in their lives, i.e., from being students to married mothers of school children, the actual research project nearly caused a breaking off of their friendship. Conflicts arose when Hendry began to study manners of communicating in Sachikos *nakama* (reference group). She integrated her into the project, which caused Sachiko to have role conflicts within her own *nakama*. However, the author gathered more insight out of this conflict than out of any other contacts with Japanese women. She evaluated her fieldwork experience as one aspect of participant observation—it having been, on the one hand, scientifically successful but personally nearly having led to a loss. The situation became more relaxed when Hendry left Japan, after enough time had elapsed, and she had analyzed and discussed the experience with Sachiko in a number of letters.

The example concerning Guatemaltecan Indians equally shows that friendship in the context of fieldwork may cause problems due to an initial unawareness of cultural differences between notions of friendship. Guatemaltecan Indians have clearly defined expectations regarding friendship, exclusiveness being one criteria for friendship. Ruben E. Reina states that not only his informants, but also those of his wife, terminated relationships, because both anthropologists had also formed contacts to other informants, thus not meeting the expectations of exclusiveness. Only through detailed explanations of why anthropologists carrying out fieldwork must speak to more than one person was the problem solved (Reina 1959, p. 48).

Role conflicts are one of the major topics discussed in writings by anthropologists on friendships formed in the context of fieldwork. In an article by Judith Friedman Hansen (1976), role conflicts are mentioned as the central problem. This is hardly astonishing, since not only during fieldwork but also in other work-related situations friendship can cause conflicts regarding loyalty, possessiveness, and role expectations. At home anthropologists have to play different roles within their families, partnerships, and within academia as well—thus, cross-cutting and intersecting friendship patterns are not typical of fieldwork alone. The problem is only more obvious and perceived more consciously during fieldwork. Coping with these conflicts requires time, much talking, patience, and a deeper understanding of the rules of the other culture.

See also: Friendship: Development in Childhood and Adolescence; Interpersonal Attraction, Psychology of; Interpersonal Trust across the Lifespan; Patron–Client Relationships, Anthropology of

Bibliography

Allan G A 1989 *Friendship: Developing a Sociological Perspective*. Westview Press, Boulder, CO

Boissevain J 1974 *Friends of Friends: Networks, Manipulators and Coalitions*. Basil Blackwell, Oxford, UK

Boissevain J 1985 Ethnographic fieldwork. In: Kuper A, Kuper J (eds.) *The Social Science Encyclopedia*. Routledge & Kegan Paul, London, pp. 272–4

Bott E 1957 *Family and Social Network*. Tavistock, London

Dundes A 1971 The making and breaking of friendship as a standard frame in African folktales. In: Maranda P, Köngäs Maranda E (eds.) *Structural Analysis of Oral Tradition*. University of Pennsylvania Press, Philadelphia, pp. 171–85

Firth R 1967 [1936] Bond friendship. In: Firth R (ed.) *Tikopia Ritual and Belief*. Allen & Unwin, London, pp. 108–15

Fischer C S 1982 *To Dwell Among Friends: Personal Networks in Town and City*. University of Chicago Press, Chicago

Friedman Hansen J 1976 The anthropologist in the field: Scientist, friend, and voyeur. In: Rynkiewich M A, Spradley J P (eds.) *Ethics and Anthropology. Dilemmas in Fieldwork*. Wiley, New York, pp. 123–34

Gilmore D C 1975 Friendship in Fuenmayor: Patterns of integration in an atomistic society. *Ethnology* **14**: 311–24

Hendry J 1992 The paradox of friendship in the field: Analysis of a long-term Anglo–Japanese relationship. In: Okely J, Callaway H (eds.) *Anthropology and Autobiography*. Routledge, London

Jacobson D 1975 Fair-weather friend: Label and context in middle-class friendships. *Journal of Anthropological Research* **31**: 225–34

Kennedy R 1986 Women's friendships on Crete: A psychological perspective. In: Dubisch J (ed.) *Gender and Power in Rural Greece*. Princeton University Press, Princeton, NJ

Kiefer T 1968 Institutionalized friendship and warfare among the Tausug of Jolo. *Ethnology* **7**: 225–44

Leyton E (ed.) 1974 *The Compact. Selected Dimensions of Friendship*. Newfoundland Social and Economic Papers, No. 3. Memorial University of Newfoundland, Newfoundland

Paine R 1969 In search of friendship. *Man* **4**: 505–24

Reina R E 1959 Two patterns of friendship in a Guatemalan community. *American Anthropologist* **61**: 44–50

Spradley J P 1979 *The Ethnographic Interview*. Holt, Rinehart and Winston, New York

Uhl S 1991 Forbidden friends: Cultural veils of female friendship in Andalusia. *American Ethnologist* **18**: 90–105

Wolf E R 1966 Kinship, friendship, and patron–client relations in complex societies. In: Banton M (ed.) *The Social Anthropology of Complex Societies*. Tavistock, London

B. Beer

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Friendship: Development in Childhood and Adolescence

Most children and adolescents have friends. Friendships are characterized by mutual liking and expectations that friends spend more time with one another than acquaintances, that cost–benefit balances in the social exchanges between friends are favorable, and, among older children and adolescents, that friends are loyal, trusting, and intimate (Bukowski et al. 1996). This article deals with friendships during the first two decades of life and their developmental significance.

1. *Incidence*

Infants and toddlers sometimes show preferences for one another although it is not always easy to compare these relationships to the friendships of older children. The word 'friend' begins to be used by children during the third and fourth years and, by that time, about three-fourths of children are involved in reciprocated friendships. Proportions rise through middle childhood (when about 85 percent of children have a best friend and usually several good friends) and decline slightly during adolescence. Only about 10 percent of older children and adolescents have no friends (Hartup and Stevens 1997).

Friendship networks are small among young children, averaging about 1.7 friends for boys and 0.9 friends for girls during the preschool years. These numbers increase to 3.0 to 5.0 best friends during middle childhood and adolescence, depending on whether one includes unreciprocated choices as well as reciprocated ones. Time spent with friends changes with age too, being greater in middle childhood and adolescence than at any other time (29 percent of time awake among teenagers).

2. *Friendship Expectations*

Friendships are understood by children and adolescents (especially) to provide the individual with companionship and intimacy more frequently than parent–child relationships, but less frequently with experiences of compliance and control. Most dramatically, friendships are understood to differ from parent–child relationships in power sharing (Furman and Buhrmester 1985). Overt affection is provided less frequently to children by friends than by parents but nurturance in general is understood to be provided by both.

Differences between children and adolescents in the ways that they describe their friends occur mainly in the number, complexity, and organization of information and ideas about these relationships (Rubin et al. 1997). Among young children, for example, friendship expectations are most likely to emphasize common activities and concrete reciprocities as well as the 'here and now' ('We play'). Among older children, friends are described as sharing values and rules about loyalty and trust; friends also expect to spend time together and engage in constructive conflict resolution. Among adolescents, friends are expected to share interests, be understanding, and empathetic; similarity between oneself and one's friends is especially important among teenagers. These developmental changes have sometimes been interpreted as reflecting changes in the child's understanding of reciprocity or mutuality (Youniss 1980). Indeed, the emphasis on mutuality in these expectations at all ages suggests that the meaning structure underlying these relationships is 'symmetrical reciprocity.'

3. *Behavior with Friends*

Children spend more time with their friends than with nonfriends. Some researchers, in fact, have used time-sharing as a means of identifying friendships (Hinde et al. 1985), although most investigators prefer validation by the children themselves or an outside observer (a parent or a teacher).

The social exchanges of friends and nonfriends are different in four major respects (Newcomb and Bagwell 1995): *positive engagement* (friends talk, smile, and laugh together more frequently than nonfriends); *task activity* (friends are more task-oriented and spend more time on-task than nonfriends); *relationship mutuality* (friends are more affirmative with one another, more mutually oriented in their interaction, and more likely to emphasize equality in their exchanges than nonfriends); and *conflict management* (friends use disengagement and negotiation as opposed to power assertion proportionally more often than nonfriends). Mutuality and symmetry thus mark the social exchanges of children and their friends as well as adolescents; these results are another reason for suggesting that reciprocity and symmetry are the behavioral hallmarks of friendship during the first two decades of life.

4. *Similarity between Friends*

Children and adolescents are more similar to their friends than to nonfriends. Greatest similarity is evident between friends in gender, race, and age—extending from early childhood through adolescence. School-aged children are more similar to friends than to nonfriends in prosocial behavior, antisocial behavior, shyness dependency, depression, and social acceptance/rejection. Friends also share biases in social perception, although it should be recognized that these similarities do not mean that friends are carbon copies of one another (Hartup 1996). By

adolescence, friendship similarities vary according to attribute, with greatest similarities occurring in school-related attitudes, aspirations, and achievement along with life-style behaviors that include smoking, drinking, drug use, and sexual activity. Generally, friends are not very similar to one another in personality traits measured with paper and pencil tests—either in childhood or adolescence. Similarities between friends are multidetermined. Owing to the manner in which schools and neighborhoods are organized in most industrialized countries, children are more likely to come together with children who are like themselves than who are unlike them. Once in the same school or neighborhood, they are more likely to be exposed to the same socialization agents (e.g., teachers). Beyond these sociodemographic conditions, children are also more attracted to children who are like themselves rather than different; children and adolescents also become more like one another as an outcome of mutual socialization. The relative extent to which mutual selection and socialization contribute to the similarity between friends depends on many conditions, including characteristics of the children themselves, the nature of their interaction, and the behavioral attributes being measured (Urberg 1999).

5. *Gender Differences*

One of the most notable characteristics of friendships in childhood and adolescence is their gender concordance. Opposite-sex friendships are relatively rare, even among preschoolers. Boys and girls do not differ in the proportion who have friends, although friendship networks are likely to be somewhat smaller among girls than among boys (Rubin et al. 1997).

Everyone knows that the activities of boys and their friends differ from the activities of girls and their friends. Certain interpersonal provisions also differ. Girls, for example, expect to be provided with greater affection, intimacy, and instrumental help by their friends than do boys. Intimacy concerns are more common in girls' talk about their friendships than in boys' talk; self-ratings of girls' friendships are more intimate; and self-disclosure between girls and their friends is more common than between boys and their friends. These gender differences are recognized and accepted by both boys and girls. Not well studied, however, are gender differences in intimacy when the construct is defined in terms of camaraderie and shared mastery rather than in terms of empathy and self-disclosure.

6. *Children Who Have Friends vs. Those Who Do Not*

Children who have friends are more likely to evince better adjustment over time than those who do not. Research supports this conclusion in that: (a) children and adolescents who have friends are better adjusted than those who do not; (b) children referred for clinical treatment are more likely *not* to have friends than children who have never been referred for treatment; and (c) youngsters who have friends feel better about themselves and are less lonely than youngsters who do not (Hartup 1996). These results have been widely reported, but they are difficult to interpret. First, 'having friends' usually means having 'good friends' in the sense that the relationships in question are harmonious and mutually supportive. This confounding is important since research with adults shows that *not* having problematic friendships is more closely related to good adjustment than actually having supportive and intimate ones (Hartup and Stevens 1997). Second, correlational results cannot be used to support causal interpretations: having friends may contribute to individual well-being but the reverse may also be true. Third, existing data may not be adequately refined: having friends may be associated with certain behavioral adaptations (e.g., self-esteem) but not others (conduct disorder).

Longitudinal studies assist in sorting out these issues when adequate assessment of early status is included in the research design and multiple measures are used to assess outcome. Accordingly, certain studies show that having a friend increases self-esteem and decreases psychosocial difficulties among preadolescents during changes from lower to middle to high school. Also, having friends in middle childhood predicts young adult feelings of self-worth, family attitudes, and the absence of depression, but not sociability, school performance, educational aspirations, and job performance (Bagwell et al. 1998). These results indicate that merely having a friend contributes selectively rather than generally to developmental outcome.

One other consequence of having friends during childhood should be mentioned—success in early romantic relationships. Having same-gender friends during middle childhood forecasts having romantic relationships during adolescence. Having friends in middle childhood also enhances success in early romantic relationships, even when the quality of early parent–child relationships is factored out (Sroufe et al. 1999). These results conform to theoretical predictions that childhood opportunities for intimacy in an asexual context are necessary precursors to heterosexual adaptation.

7. *Characteristics of One's Friends*

Considerable weight in determining developmental outcome is carried by characteristics of the child's best friends. In many instances, more can be learned by knowing who children's friends are than by knowing merely that they have a best friend. For example, following divorce or remarriage (both stressful events), having socially skilled friends who have few behavior problems promotes resilience whereas having imma-

ture friends who have many behavior problems does not (Hetherington 1999). Conversely, considerable evidence shows that friendships between antisocial children increase the risk of antisocial behavior among both children (Coie and Dodge 1997).

What socialization experiences between friends contribute to these outcomes? Although observational learning (modeling) may make a contribution, it is also the case that the interaction between friends varies according to the characteristics of the children involved. Overtly aggressive boys, for example, are more contentious and conflict-instigating with their friends than nonaggressive boys are, and engage in more deviant talk (Dishion et al. 1995). 'Relationally aggressive' children, who hurt others by exclusion or doing things that damage the victim's relationships with other children, have friendships that are intimate and exclusive but, at the same time, marked by frequent displays of relational aggression (Grotpeter and Crick 1996). One can easily guess how this admixture of intimacy and relationship-oriented aggression incites conflict and uncertainty, as well as provides instruction in how to draw others into dysfunctional relationships.

8. *Friendship Stability*

Children who have friends at one time are likely to have friends at other ages. Changes in partners are common, although friendships are more stable than is sometimes appreciated (Hartup and Stevens 1997). Preschool-aged children often maintain these relationships for many months. Friendship stability, however, increases with age until about 70 percent of adolescents report that their friendships last for a year or more; the majority of 13- and 14-year olds who have friends report that these relationships have lasted between one and five years.

Stability varies according to child characteristics: friendships between aggressive, antisocial children are less stable than friendships among nonaggressive children (Dishion et al. 1995). Disabilities and psychosocial difficulties are also associated with reduced friendship stability. In these instances, stability is reduced as a consequence of limitations on the child's ability to regulate emotion and other deficits in social skills.

Friendship stability has considerable developmental significance. Kindergarten children, for example, who maintain friendships across the school year show improvement in attitudes toward school and other school-related behaviors, whereas children who do not maintain their friendships show poorer adaptations. In other instances, friendship stability is correlated with outcome according to characteristics of the child's friends: stable friendships with children who have conduct problems, for example, increase children's own behavior problems, whereas stable relationships with children who are socially withdrawn do not increase the children's social withdrawal. Friendship stability thus affects children differently according to the behavior in question and other aspects of their relationships (Berndt et al. 1999).

9. *Friendship Quality*

Friendships among children and adolescents vary in their social and emotional qualities. Some relationships are supportive and intimate, some conflict-ridden. One can differentiate among these relationships in terms of reciprocity, self-disclosure, overt hostility, and covert hostility. Friendship structures have been most commonly described, however, simply as 'positive' or 'negative' (Bukowski et al 1996).

Supportiveness between children and their friends is positively correlated with school involvement and achievement while being negatively correlated with school-based problems. Supportiveness between school-aged friends is positively correlated with self-esteem in early adulthood, as well as negatively correlated with later identity problems (Bagwell et al. 1998). Friendship quality is associated with most developmental outcomes, however, in combination with child characteristics and friendship stability. For example, boys with very low-quality friendships and who are delinquent at age 13 to 14 escalate more over the next year in delinquent activity than either delinquent or nondelinquent boys in better relationships (Poulin et al. 1999). Other results show that sociability increases most and social withdrawal increases least across the transition from elementary school to junior high school when friendships are *both* high quality and stable (Berndt et al. 1999). Although causal direction has not been established in these data, the weight of the evidence suggests that good developmental outcomes are most likely when a child has friends, those friends are socially skilled, and these friendships are supportive and stable. Friendships between children who are not well socialized are mixed blessings.

10. *Conclusion*

Friendships are developmentally significant from early childhood through adolescence and beyond. Friendships are cognitive and affective resources, fostering social problem solving, self-esteem, and well-being. Friends socialize one another (sometimes for better, sometimes for worse). Supportive and intimate friendships generally are developmental advantages, whereas contentious relationships between individuals at risk generally are disadvantages.

Even so, friendships are not all alike. Children differ in the company they keep and in the social and affective qualities of their relationships. Developmental assessment is more complete when one knows who

a child's friends are and what these relationships are like. Developmental outcomes need to be better understood, however, especially in terms of children's involvement in relationships that can place them at risk. Particularly needed are studies of the manner in which friendship experience combines with temperament and early experience, social skills, family relationships, and the social context (especially the child's encounters with stress) to affect developmental outcome. One must also consider that childhood friendships may be more important in accounting for developmental outcome in some individuals than others.

See also: Divorce and Children's Social Development; Friendship, Anthropology of; Infancy and Childhood: Emotional Development; Interpersonal Trust across the Lifespan; Trust: Philosophical Aspects; Trust, Sociology of

Bibliography

Bagwell C L, Newcomb A F, Bukowski W M 1998 Preadolescent friendship and peer rejection as predictors of adult adjustment. *Child Development* **69**: 140–53

Berndt T J, Hawkins J A, Jiao Z Y 1999 Influences of friends and friendships on adjustment to junior high school. *Merrill-Palmer Quarterly-Journal of Developmental Psychology* **45**: 13–41

Bukowski W M, Newcomb A F, Hartup W W (eds.) 1996 *The Company They Keep: Friendship in Childhood and Adolescence*.Cambridge University Press, Cambridge, UK

Coie J D, Dodge K A 1997 Aggression and antisocial behavior. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*. Wiley, New York, Vol. 3, pp. 779–862

Dishion T J, Andrews D W, Crosby L 1995 Antisocial boys and their friends in early adolescence: Relationship characteristics, quality, and interactional process. *Child Development* **66**: 139–51

Furman W, Buhrmester D 1985 Children's perceptions of the personal relationships in their social networks. *Developmental Psychology* **21**: 1016–24

Grotpeter J K, Crick N R 1996 Relational aggression, overt aggression, and friendship. *Child Development* **67**: 2328–38

Hartup W W 1996 The company they keep: Friendships and their developmental significance. *Child Development* **67**: 1–13

Hartup W W, Stevens N 1997 Friendships and adaptation in the life course. *Psychological Bulletin* **121**: 355–70

Hetherington E M 1999 Social capital and the development of youth from nondivorced, divorced, and remarried families. In: Collins W A, Laursen B (eds.) *Relationships as Developmental Contexts, Minnesota Symposia on Child Psychology, Vol. 30*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 177–209

Hinde R A, Titmus G, Easton D, Tamplin A 1985 Incidence of 'friendship' and behavior with strong associates versus non-associates in preschoolers. *Child Development* **56**: 234–45

Newcomb A F, Bagwell C L 1995 Children's friendship relations: A meta-analytic review. *Psychogical Bulletin* **117**: 306–47

Poulin F, Dishion T J, Haas E 1999 The peer influence paradox: Friendship quality and deviancy training within male adolescent friendships. *Merrill-Palmer Quarterly-Journal of Developmental Psychology* **45**: 42–61

Rubin K H, Bukowski W M, Parker J G 1997 Peer interactions, relationships, and groups. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*. Wiley, New York, Vol. 3, pp. 619–700

Sroufe L A, Egeland B, Carlson E 1999 One social world: The integrated development of parent-child and peer relationships. In: Collins W A, Laursen B (eds.) *Relationships as Developmental Contexts*. L. Erlbaum Associates, Mahwah, NJ, Vol. 30, pp. 241–61

Urberg K A 1999 Introduction: Some thoughts about studying the influence of peers on children and adolescents. *Merrill-Palmer Quarterly-Journal of Developmental Psychology* **45**: 1–12

Youniss J 1980 *Parents and Peers in Social Development: A Sullivan-Piaget Perspective*. University of Chicago Press, Chicago

W. W. Hartup

Frontiers in History

Frontiers are as ancient as the state and the history of frontiers begins as far back as Herodotus, but the frontier as a myth is a recent invention. In Europe, the word 'frontière' first appears in French in the fourteenth century to indicate a facade in architecture or a military order of battle and only gradually came to mean the limits of state control. It passed into English and was widely used in the sixteenth century to mean a barrier against attack. In the thirteenth century the Slavic word 'greniz' or 'granitsa' was applied, particularly by the Teutonic Order, to the area of contact between Germans and Slavs, and replaced the Gothic term 'mark'. In the scholarly literature, definitions of frontiers have shifted from line and place to process, symbol and mythology as the disciplines involved—anthropology, geography and history—interacted creatively with one another under the general influence of the contextual approach. This has not been a linear evolution; nor has it been uniformly accepted, least of all by advocates of geopolitics, who still perceive frontiers as constituting a power grid, even when they have been determined by cultural or ideological factors.

1. Frontiers and the Historical Sciences

It is just over 100 years since Frederick Jackson Turner published his famous essay that still serves as the main frame in a triptych of historical approaches to frontiers in history (Turner [1893] 1921). Given the large number of American historians and their methodological restlessness, the comparative historiography of fron-

tiers has been dominated by their revisions of Turner. Flanking Turner's frontier on one side is a different spatial concept tied to the rise and consolidation of the centralized state that developed out of the French experience, as interpreted by the Annales school. On the other side a third panel represents the symbolic geographies, that is, the construction of imaginary borders between civilizations, as exemplified by the classical dualism of Europe and Asia, and the trialism of the Cold War: the West, the Communist Bloc and the Third World.

Anthropologists came later than historians to frontier studies, but then arrived in considerable force. In the 1950s they began to engage in extensive fieldwork on the margins of sedentary and nomadic societies (Khazanov 1994). Their insights were rapidly absorbed by historians, who had already been seriously exploring frontiers for over half a century. Finally, geographers based in local history and influenced by cultural studies began to redefine the ideas of space and frontiers in terms of linguistic and social contexts (Godlewka and Smith 1994). As a result the current state of research on frontiers is multidisciplinary and comparative.

2. *Types of Frontiers*

Frontiers may be separated into three basic types: consolidated state frontiers; dynamic frontiers of advancing settlements; and symbolic frontiers. On occasion, features of two or even three of these types coincide with one another. Each has a variety of subtypes. All are characterized by the existence of zones of influence and exchange on both sides of the real or imaginary line that presumes to divide them.

2.1 *State Frontiers in Antiquity*

The earliest historical evidence of frontiers expresses the needs of the state to define the limits of public authority, as in the empires of ancient Mesopotamia and Egypt, or later of private property in the Greek city states. In both cases, boundary markers were endowed with a sacred character and their removal signified the victory of a conqueror (Foucher 1991). The Athenian and Roman Empires and, independently, the Chinese Empire introduced two additional features: the construction of walls or defensive lines, and, second, the definition of peoples beyond the frontiers as culturally inferior, that is, barbarian. The attempts of the Roman and Chinese emperors to establish fixed frontiers repeatedly broke down under the pressure of migration, raiding and the development of extensive cultural and economic exchange. In the case of the Chinese-steppe frontier, the process of acculturation, transculturation and assimilation, sometimes simplified to 'trade or raid,' worked in both directions long before the Great Wall was constructed and continued into the twentieth century. Relations between the Romans and peoples on the periphery of their control varied greatly according to location over centuries. Even the walls (limes), especially in the UK and Tunisia, were designed less to exclude, and more to control the movement of people and goods. The collapse in the West is increasingly regarded as a process of mutual accommodation of Romans and barbarian tribes through the emergence of a new military elite of the warlord type (Whittaker 1994).

2.2 *Islamic State Frontiers*

The Ottoman Empire represents a subtype of state frontier, intermediate between the ancient empires and the modern European state system. Embodying the precepts of Islam, it did not recognize until the late eighteenth century the concept of state frontiers in the European sense, but rather a division of the world into two cultural spheres: the Dar ul-Islam, the abode of Islam, and the Dar ul-Harb, the abode of war. But even this rigid division could not be strictly maintained. For example a contested space occupied by nomads opened up all along the Islamic–Byzantine frontiers during the relative stability of the tenth century. By the thirteenth century, there had arisen a culture of holy warriors (gazhis) with a corresponding type on the Byzantine side (Akritai), at first composed of Greek, and then of Turkomen tribesmen. This intermediate zone where war and trade often alternated facilitated the penetration and conquest of the Byzantine Empire by the Ottomans (Wittek 1938). At the western end of Ottoman power, an even more complex military frontier emerged where three empires, the Ottoman, Habsburg and Venetian (triplex confinium) confronted and co-existed with one another for three centuries (ca. 1500–1800). Small wars, population transfers, natural disasters and deeply rooted uncertainties shaped a border society that invented a variety of survival strategies, from cross-border trade and blood brotherhood to religious conversions, and gave rise to a variation of the ghazi warrior culture in the form of the Christian Uskoks (Kaser 1997, Roksandic 1998). Along the northern Ottoman frontier, similar conditions existed at the meeting ground of both the Poles, Russians and Ottomans in the Ukraine, and of the Russians, Ottomans and Iranians in the Transcaucasus, which also produced a warrior culture of Cossacks. Curiously, Ottoman specialists have yet to make the frontier experience the central theme in the history of the empire.

By the nineteenth century, the Ottoman frontier societies became more sharply divided by fundamental cultural conflicts. As Ottoman power receded in southwestern Europe and along the Black Sea, littoral

new state boundaries were drawn by international negotiation (first with Russia in the Treaty of Kuchuk-Kainardji, 1774) and great power intervention. Later, both the idea of rigid linear frontiers along national lines and the power to enforce them came from western Europe.

2.3 West European State Frontiers

A very different kind of frontier emerged in Western Europe during the transition from the feudal fragmentation of political power to the centralized absolutist monarchy. France, which set the pace frontiers as the limit of state power, evolved from a military to an ideological, financial and linguistic line accompanied by a more scientific cartography; in 1815, the French government produced the first map in the world drawn on a geodesic triangulation. Propagandists of the ancient regime and men of the revolution both claimed for France its 'natural frontiers', first in the form of the four rivers (Rhone, Saone, Meuse, and Scheldt), and then invoking Strabo and Caesar, the quadrilateral limits of ancient Gaul, the Rhine, Alps and the Ocean. To these myths the revolutionaries added another, the exact correspondence between 'natural' and 'national' frontiers before it was entirely clear who or what constituted the nation. It took another century before the administrative, customs and juridical instruments were sufficiently firmly established to enforce the claim (Bergeron and Revel 1989). In the process, the national territorial boundary was shaped as much by the local, frontier societies as by the central state, and as a result, the inhabitants retained their sense of place, even as they accepted their identity with the nation (Sahlins 1989).

The French (or Jacobin) model that the frontiers of a state ought to correspond to the natural and the national with one language and one culture was internationalized, at least for France, in the treaties of Vienna of 1815. Subsequently, it was widely diffused throughout Europe, Latin America and—with the participation of the British—Africa and the Middle East with mixed results.

One subtype of the French model emerged in the late nineteenth and early twentieth centuries in the small successor states of central, southeastern and eastern Europe. But there the prolonged existence of multicultural societies (Habsburg, Ottoman, and Russian) shaped by dynastic interests delayed into the twentieth century the shift from historic to linguistic (national) frontiers with dire consequences. By this time, internal migration, colonization, forced resettlement, and expulsions had created a kaleidoscope of peoples which could not be neatly compartmentalized by drawing linear frontiers.

Redrawing the map of southeastern Europe as the Ottoman power gradually receded was a cartographic nightmare, the result of a series of unhappy compromises between the conflicting interests of the great powers and the aspirations of the subjugated peoples striving not only for independence but also for frontiers based on historicist claims—Greater Serbia, Greater Croatia, Greater Romania, Greater Hungary, Greater Bulgaria, Greater Greece, etc.—claims that were mutually irreconcilable.

German frontiers represent another subtype, sharing elements of the consolidated state and the dynamic moving frontier. In the north, west and south, the German frontiers evolved along the lines of the French model. But in the vast eastern space of the German–Slavic encounter, the discrepancy between sovereignty and settlement remained unresolved for 1000 years. For German statesmen, publicists and intellectuals from the eighteenth to the mid twentieth century, the task of consolidating a fixed state frontier with the Slavic east became a profoundly disturbing socio-psychological as well as political problem producing its share of historical myths (Demandt 1990).

3. The Dynamic Type of Frontiers

The second type of frontier may be characterized as an advancing line of settlements engaged primarily in agricultural or mixed economy, confronting a semi-nomadic and/or technologically less developed culture. In the modern period, three subtypes may be distinguished: the United States, the former British dominions and the Imperial Russian and Chinese Empires. The United States was the only former British colony where, once it had achieved independence, the frontier experience became both a national myth and a dominant historiographic theme. The other former British dominions constitute another subtype, where geographical and political factors placed the idea of the frontier in a different context. The continental expansion of Russia and China constitute a third subtype where the advance of the two societies brought them into direct contact and competition with one another.

3.1 The American Frontier as a Subtype

Neither the revisionists in American historiography nor the comparativists on an international scale have been able to free themselves entirely from the Turnerian concept of the frontier, even as they demolished its tenets one by one. What Turner did was to establish the frontier as a historical problem and to ask what it meant to be an American. The complexity of his approach guaranteed its longevity even as a foil. He combined elements of economic geography, social structure and ideology even though he was not always clear or consistent in drawing the relationship between them. His brilliant essays added up to something less than a full-blown theory and left the way open for

multiple interpretation, extensive revision and critical refutation. Like many great theses in historiography, its importance lies more in the literature it generated than in its original propositions. The debate over his legacy took place in two arenas, the national (American) and the comparative (global).

In American historiography, the revisions and reformulations of the Turner thesis radiate around two conceptual poles: new methodologies, particularly anthropology and environmental studies, and fresh interpretive perspectives reflecting the shifting concerns of a restless society in search of itself. At present, as in so many other areas, the task of charting the currents of the debate has itself become a minor cottage industry (Nash 1991). Turner's frontier as an advancing line of settlement has given way, in the work of the regionalists, to a multiplicity of locations. But Turner's tendency to conflate the frontier as a line and 'the Great West' as a region continues to be a source of confusion. There has also been a trend away from spatial to cultural definitions of the frontier, that is, from place to process. Turner's concept of the frontier as dividing 'savagery from civilization' or, alternatively, as a forward movement into 'empty lands' has, under the influence of anthropologists, yielded to the idea of the frontier as a zone of interaction involving two or more previously distinct societies that engendered a variety of cultural and commercial 'frontier exchanges' between Native Americans and colonists. Turner's concept of the West as an environment that bred rugged individualism and privileged self-reliance was challenged by historians and sociologists who gave equal billing to social collectivities, and uncovered a darker side of the frontiersman in the persistence of a gun culture and the cult of violence. Environmentalist historians exploded Turner's arcadian picture of natural harmony on the frontier by exposing the depredations, manipulations and 'species shifting' impact of the western advance (Cronon et al. 1992). Turner's insistence on the defining role of the frontier experience in building American democratic institutions was counterbalanced by the contributions of urban life and the antecedent values the pioneers brought with them. At one end of the historigraphical spectrum, Turner's thesis lies in ruins and even the term 'frontier' is rejected or excoriated. At the other end, attempts have been made to synthesize the newer ideas on race, class gender and region, in order to rescue once again the importance of the frontier in American history. The persistence of interest in Turner, marked more often by invoking his 'ghost' or 'spirit' rather than the body of his work, lies in large part with the power of his explanatory theme to explain American national identity. Pluralistic interpretations have never conveyed the same emotional force. Turner's thesis has passed into the realm of historical myth and nowhere is this more evident than in the realm of mass culture.

The commercialization of American history found the perfect vehicle in the frontier. The transformation of the frontier into a popular myth and its subsequent permutations was a function of changes in both the medium and the message. New social and environmental concerns left their mark on Turnerian stereotypes. In film and television, the heroic cowboy in the white hat, the intrepid farmer fighting off the Indians, the cavalry arriving at the last minute became more ambiguous figures. Black cowboys and troopers, environmentally sensitive Native Americans, even women gun-slingers challenged the dominant white male image. The Marlboro Man made his last stand on American billboards.

3.2 The British Empire as a Subtype

An analysis of subtypes of the dynamic frontier cannot be treated without reference to the comparative approach largely inspired by, if not always accepting, the Turner thesis and its revisions. The influence of the debate over the Turner thesis has been felt most keenly in the historiographies of the former dominions of the British Empire, and the Russian and Chinese Empires. Historians of North America and South Africa have produced a particularly well-designed comparative analysis of their respective frontiers (Lamar and Thompson 1981). It sought to achieve theoretical clarity and consistency by establishing a universal definition rather than adopting the USA experience or Turner thesis as a model, even though the original inspiration remained clear. Three elements in defining frontier societies are considered essential: first, the geographical–territorial, conceived as a zone rather than a line with different carrying capacities, attractiveness and resources; second, the cultural, in terms of the interpenetration between two previously distinct societies, one indigenous and the other intrusive; and third, the operational, the process of opening and closing, the former dated by the arrival of representatives of the intrusive society and the latter, more-complex stage because it was variable and potentially reversible, as it indeed proved to be in South Africa, by the establishment of a single political hegemony through extermination (Tazmania), expulsion (Trans-Mississippi), subjugation (South Africa) or incorporation (Brazil). Specific comparisons of the USA and South Africa reveal more differences than similarities, although the white communities in both mythologized the frontier experience as formative. Before the recent reversal in South Africa, the critical difference appeared to be the different relationship of the indigenous people in the USA and South Africa to capitalism.

Comparative studies, with Turnerian overtones, of the frontier experience in Canada, Australia and New Zealand suggest that despite superficial similarities with the USA, geographical factors and the prolonged tutorship of the metropolitan center in the UK

prevented the rise of a mythology of frontier exceptionalism (Nugent 1994). In Canada there were close parallels between the farming frontier in Ontario and Saskatchewan, and in North Dakota and Montana. But the expansion west was delayed by the physical barrier of the Canadian shield. When it came, it was promoted by Royal Governors established national police (the Royal Mounties) before mass immigration and negotiated settlements with the Blackfoot tribes thus heading off the local violence and large-scale white-native American wars that characterized the American west. To an even greater extent, the geography of Australia, dominated by a vast arid region in the center, 'a hinterland but no heartland,' relegated the bulk of the population to the cities on the east and south coast while royal land policy did not favor the small farmer in competition with the large sheep ranges. Yet even here there are regional similarities between the semi-arid fringes of California and Victoria. And Australia had its frontier folk hero in the form of the bush-ranger, although he was connected to the convict life rather than the freebooter born in an open society. The Northwest Frontier of India was the only zone of encounter between the British colonial expansion and a seminomadic population where the odds were rated even. The character of this military frontier resembled more that of the great older empires, the Roman, Chinese and Ottoman, confronting a persistent threat to security along one or several of their peripheries. The British inherited the problem from their predecessors the Moghuls, who, like the founders of several Chinese dynasties, were originally 'eastern barbarians' before they came to settle in the fertile river valleys of the south. Although the Northwest Frontier could hardly be said to have dominated British colonial politics, there was a constant tension between the proponents of a forward policy to anchor the frontier on the Hindu Kush and the advocates of 'masterly inactivity' (Fraser-Tytler 1950). And there was concern and debate over the Russian advance behind the tribes. The British crossed the frontier in force three times to fight the Afghans without settling much of anything. It was a classical military frontier zone exhibiting sporadic warfare, trade, smuggling, movement of peoples and a mythology of heroism immortalized by Kipling. Today, the Northwest frontier between Afghanistan and Pakistan has lost nothing of this except its glamour.

3.3 The Imperial Russian and Chinese Empires as a Subtype

The long, porous and shifting Eurasian frontiers of Russia and China have also attracted comparativists who employ Turner's thesis and its revisions as a point of departure in their analysis. However, even before Turner, nineteenth-century Russian historians, like S. M. Solov'ev and V. O. Kliuchevskii, stressed the central role of colonization and frontier in shaping the history of their own country, and drew from many of the same formative influences as their American counterpart (Bassin 1993). By contrast, though, their interpretations emphasized the negative effects: the draining of resources from the center, the unfavorable conditions of climate and soil, the threat of nomadic incursions. Historians in the early Soviet period added to this list the exploitation of the conquered people. A strong reversal set in only after World War II when the new state school of Soviet historians extolled the Russian peasant colonizer as a heroic figure and employed the specialized term 'merging' (sblizhenie), a form of assimilation, to signify the civilizing mission of the Russian people. A third turn has been taken only recently by the emerging regional (Siberian) school of historians who, in readjusting their focus to the specific features and values of the periphery at odds with those of the center have, nonetheless, acknowledged their debt to Turner.

Meanwhile, Western scholars, mainly Americans and British, assumed the mantle of the pre-revolutionary Russian historians and invoked Turner's thesis, in order to draw parallels and contrasts between the Russian conquest and colonization of Siberia and the American western expansion. The earliest attempts to apply the Turner thesis to the Russian experience lagged behind the extensive revisionist literature in the USA that had already forced a re-evaluation of its central tenets. However, since the 1980s a new generation of American and British students of Russian colonization have incorporated new ideas borrowed from the historiography of the USA, rather than from direct comparison with the American experience. Multiple locus and process have replaced conquest and resistance (Barrett 1995).

On the Chinese side, after more than six decades the dominant figure in frontier studies remains Owen Lattimore. His earliest work combined the keen first-hand observations of the linguistically competent, experienced traveler with historical and anthropological training (Lattimore [1940] 1951). Studded with comparative insights (Roman Empire, American West, Northwest Frontier Province), its central thesis was to establish that frontiers, as distinct from boundaries, were the outer limit of zones on the margins of a socioeconomic systems that represented their 'optimal limit of growth.' While admitting that Turner was 'an acute observer' he added that 'what he saw so clearly he saw standing on his head. In large measure, when he thought he saw what the frontier had done to society, he was really seeing what society did to the frontier' (Lattimore 1962). By comparing the social and economic character of Chinese and Russian empires, he was able to distinguish between their different 'margins of expansion.' The study of Chinese frontiers has increasingly focused on the long-term sinefication of the frontier provinces: Manchuria, Inner Mongolia, Sinkiang, and Tibet.

4. Symbolic Frontiers

The frontiers of symbolic geographies focus on the meeting ground of distinctive civilizations. The attempt to define and defend these frontiers was a feature of both the Greco-Roman and Chinese empires, the first to identify the boundary between civilized and barbarian. It was continued by the Islamic and Christian (European) traditions which were contemporaneous and contiguous hence more conflictual. But religious fissures (Sunni-Sh'ia on the one hand and Eastern Orthodox and Latin on the other) created intercivilizational frontiers that proved just as difficult to cross. Within Europe, a significant shift in symbolic geography occurred in the eighteenth century when the philosophers realigned the main cultural divide between the enlightened and the backward from a south–north to an west–east axis (Wolff 1994). But the exact demarcation line was much disputed, especially when confused with the symbolic frontier between Europe and Asia as Metternich's famous quip, 'Asia begins at the Landstrasse,' well illustrates.

In the twentieth century, the Soviet rhetoric of frontiers infused the ancient 'sacred' concept with an exclusionist ideological meaning enabling another rhetorician, Churchill, to coin the phrase 'Iron Curtain,' which retained its symbolic power long after events demonstrated its imprecision and inadequacy. The collapse of the Soviet Union and the Communist bloc overturned the theory of stable state frontiers, which had dominated American political science from 1950–90 (Anderson 1997). It also spurred a recurrent debate about the cultural line dividing western and eastern Europe and the West from other civilizations. Both during and after the Cold War there were divergent attempts to locate the historic origins of the dividing lines. Some held that Latin and Orthodox Christianity constituted a single civilization for 1,000 years until the communists rang down the curtain and arbitrarily split it along the Oder–Neisse line (Halecki 1950). Others, like Arnold Toynbee, drew the frontiers of civilizations along religious lines and argued that totalitarian Russia had its roots in Eastern Orthodoxy. After 1989, several émigré and former dissident intellectuals in the region resurrected the idea of a different more westerly and long-established cultural frontier between 'central Europe,' which was equated with liberal democracy and the spirit of capitalism, from 'eastern Europe' associated with the authoritarian state and cultural backwardness. Only in this context is it possible to measure the enormous psychological impact on the people of the region of being accepted or rejected for membership in NATO or the European Union, real institutions which, nevertheless, constitute the new symbolic frontiers of civilization.

See also: Area and International Studies in the United States: Intellectual Trends; Borders, Anthropology of; Civilization, Concept and History of; Colonization and Colonialism, History of; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: East Asia; Historiography and Historical Thought: Islamic Tradition; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Nations and Nation-states in History; State Formation; State, History of

Bibliography

Anderson M 1997 *Frontiers: Territory and State Formation in the Modern World*. Polity Press, Malden, MA

Barrett T 1995 Lines of uncertainty. The frontiers of the north Caucasus. *Slavic Review* **54**(3): 578–601

Bassin M 1993 Turner, Solov'ev and the 'Frontier Hypothesis': The nationalist significance of open spaces. *Journal of Modern History* **3**: 473–511

Bergeron L, Revel J (eds.) 1989 *L'Espace Français*. Seuil, Paris

Cronon W, Miles G, Gitlin J 1992 *Under an Open Sky: Rethinking America's Western Past*. W. W. Norton, New York

Demandt A (ed.) 1990 *Die Grenzen in der Geschichte Deutschlands*. Verlag C H Beck, Munich, Germany

Foucher M 1991 *Fronts et Frontières. Un Tour du Monde Géopolitque*. New edn. Fayard, Paris

Fraser-Tytler W K 1950 *Afghanistan. A Study of Political Developments in Central Asia*. Oxford University Press, London

Godlewka A, Smith N 1994 *Geography and Empire*. Blackwell, Oxford

Halecki O 1950 *The Limits and Divisions of European History*. Sheed and Ward, London

Kaser K 1997 *Freier Bauer und Soldat. Die Militarisierung der agrarischen Gesellschaft an der kroatisch–slawonischen Militargrenze (1535–1881)*. Verlag, Vienna

Khazanov A M 1994 *Nomads and the Outside World*, 2nd ed. University of Wisconsin Press, Madison, WI

Lamar H, Thompson L (eds.) 1981 *The Frontier in History: North America and Southern Africa Compared*. Yale University Press, New Haven, CT

Lattimore O [1940] 1951 *Inner Asian Frontiers of China*. American Geographical Society, New York

Lattimore O 1962 *Studies in Frontier History. Collected Papers, 1928–1958*. Oxford University Press, London

Nash G D 1991 *Creating the West: Historical Interpretations, 1880–1990*, 1st edn. University of New Mexico Press, Albuquerque, NM

Nugent W 1994 Comparing wests and frontiers. In: Milner C A II et al. (eds.) *Oxford History of the American West*. Oxford University Press, New York pp. 803–33

Roksandic D (ed.) 1998 *Microhistory of the Triplex Confinium*. Institute on Southeastern Europe, Budapest, Hungary

Sahlins P 1989 *Boundaries. The Making of France and Spain in the Pyrenees*. University of California Press, Berkeley, CA

Turner F J [1893] 1921 *The Frontier in American History*. Henry Holt, New York

Whittaker C R 1994 *Frontiers of the Roman Empire. A Social and Economic Study*. Johns Hopkins University Press, Baltimore, MD

Wittek P 1938 *The Rise of the Ottoman Empire*. Royal Asiatic Society, London
Wolff L 1994 *Inventing Eastern Europe. The Map of Civilization on the Mind of the Enlightenment*. Stanford University Press, Stanford, CA

A. J. Rieber

Functional Brain Imaging

Since the 1990s the field of cognitive neuroscience has emerged as a very important growth area in neuroscience. Cognitive neuroscience combines the experimental strategies of cognitive psychology with various techniques to actually examine how brain function supports mental activities. Leading this research in normal humans are the new techniques of functional brain imaging: positron emission tomography (PET) and magnetic resonance imaging (MRI) along with event related potentials (ERP's) obtained from electroencephalography (EEG) or magnetoencephalography (MEG). This article focuses on PET and MRI.

1. *Antecedents of Functional Brain Imaging*

1.1 *Neuropsychology*

Early discussions of the 'mind-brain problem' treated the brain largely as a 'black box.' Then, in 1861, French surgeon and anthropologist Pierre Paul Broca described a clear relationship between a patient's difficulty speaking and an injury to a specific part of the patient's brain due to a stroke. Since this seminal observation, a vast body of scientific literature has accumulated implicating various parts of the human brain in specific aspects of human behavior, including language. This approach to the study of human brain function has achieved a remarkable level of sophistication and represents an important tool in neuroscience.

The features of brain organization arising from the study of patients with brain injury nevertheless raise some questions of interpretation. The size and location of brain injury varies greatly from patient to patient, making a precise correlation between damage to a particular area of the brain and the function normally served by that area sometimes difficult to determine. Furthermore, each patient may be assumed to have some features of brain organization that are unique to him or her. Finally, it remains uncertain whether one can simply attribute a lost or disrupted function to a particular area of injury. Because of the interconnected nature of areas of the brain, injury in one area is likely to have effects on other areas that cannot necessarily be predicted from the location and size of the injury. Thus, as valuable as our insights concerning the organization of the human brain have been, from the study of patients with stroke and other types of brain injury, it has remained an open question exactly how this information relates to the normal organization of the human brain.

Only recently have scientists interested in this question had the opportunity to explore it analytically—to peer inside the black box as it functions normally. This ability stems from the developments in imaging technology since the1980s, most notably PET and MRI. These techniques can now capture localized changes in blood flow in the normal human brain associated with behaviorally induced changes in brain cellular activity (Raichle 1998).

1.2 *Blood Flow and Brain Function*

The quest for an understanding of the functional organization of the normal human brain, using techniques to assess changes in brain circulation, has occupied mankind for more than a century (for a historical review see Raichle (2000)). An enduring theme throughout this work has been that changes in the cellular activity of the normal brain of humans and laboratory animals is invariably accompanied by changes in the blood flow to the region of the brain affected.

1.2.1 Early quantitative blood flow measurements in humans. At the close of World War II, Seymour Kety and co-worker opened the first chapter in quantitative studies of brain circulation and metabolism in humans and laboratory animals. Working with Lou Sokoloff and others, Kety developed the first quantitative methods for measuring *whole* brain blood flow and metabolism in humans. The introduction of an in vivo tissue autoradiographic measurement of regional blood flow in laboratory animals by Kety's group (Kety 1960, Landau et al. 1955) provided the first glimpse of quantitative changes in blood flow in the brain related directly to brain function.

Soon after Kety introduced quantitative methods for measuring whole brain blood flow and metabolism in humans, David Ingvar, Neils Lassen, and their Scandinavian colleagues introduced methods applicable to man that permitted regional blood flow measurements to be made using scintillation detectors arrayed like a helmet over the head (Lassen et al. 1963). They demonstrated directly in normal human subjects that blood flow changed regionally during changes in brain functional activity. The first study of functionally-induced regional changes in blood flow using these techniques in normal humans was actually reported by Ingvar and Risberg at an early meeting on brain blood and metabolism (Ingvar and Risberg 1965) and was greeted with cautious enthusiasm and a clear sense of its potential importance for studies of

human brain function by Kety (Kety 1965). However, despite many studies of functionally induced changes in regional cerebral blood, which followed, this approach was not embraced by most neuroscientists or cognitive scientists at the time. It is interesting to note that this indifference was to disappear almost completely in the 1980s, a subject to which the article will return.

2. Introduction of Modern Imaging Techniques

The modern era of medical imaging began in the early 1970s when the world was introduced to a remarkable technique called X-ray computed tomography, now known as X-ray CT or just CT (for review see Kevles 1997). South African physicist Allan M. Cormack and British engineer Sir Godfrey Hounsfield independently developed its principles but it was Hounsfield who constructed the first scanner successfully.

2.1 X-ray Computed Tomography (CT)

CT takes advantage of the fact that different tissues absorb varying amounts of X-ray energy. The denser the tissue, the more it absorbs. A highly focused beam of X-rays traversing the body will exit at a reduced energy level depending on the tissues and organs through which it passes. A beam of X-rays passed through the body at many different angles through a plane provides sufficient information to reconstruct a picture of that body section. It was crucial to the development of CT that clever computing and mathematical techniques emerged for processing the vast amount of information necessary to create the images themselves. Without the availability of sophisticated computers, the task would have been impossible to accomplish.

CT had two consequences. First, it changed the practice of medicine forever, because it was much superior to standard X-ray techniques. For the first time, physicians could safely and effectively view living human tissue such as the brain with no discomfort or risk to the patient. Second, CT immediately stimulated scientists and engineers to consider alternative ways of creating images of the body's interior using similar mathematical and computer strategies for image construction. These efforts went beyond the picture of human anatomy to focus on function.

2.2 Positron Emission Tomography (PET)

One of the first groups to be intrigued by the possibilities opened by CT consisted of experts in tissue autoradiography, a method used for many years in animal studies to investigate organ metabolism, biochemistry and blood flow. In tissue autoradiography, a radioactively labeled compound is injected into a vein. After the compound has accumulated in the organ under investigation, the animal is sacrificed and the organ (e.g., the brain) removed for study. The organ then is sectioned carefully, and the individual slices are laid on a piece of film sensitive to radioactivity. Much as the film in a camera records a scene as it was viewed originally, this X-ray film records the distribution of radioactively labeled compound in each slice of tissue.

When the X-ray film is developed, scientists have a picture of the distribution of radioactivity within the organ and hence can deduce the organ's specific functions. The radioactive compound injected determines the type of information. A radioactively labeled form of glucose, for example, measures brain metabolism, because glucose is the primary source of energy for the cells of the brain. Central to functional brain imaging with PET is the measurement of brain blood flow, which is accomplished by the injection of radioactively labeled water.

Investigators adept at tissue autoradiography were fascinated when CT was introduced. They suddenly realized that if they could reconstruct the anatomy of an organ by passing an X-ray beam through it, as CT did, they could also reconstruct safely the distribution of a previously administered radioisotope. One had simply to measure the emission of radioactivity from the body section. This realization was the birth of autoradiography of living human subjects.

One crucial element in the evolution of human autoradiography was the choice of radioisotope. Workers in the field selected a class of radioisotopes that emit positrons, which resemble electrons except that they carry a positive charge. A positron produced with the tissue almost immediately combines with a nearby electron. The positron and the electron annihilate one another in this interaction, emitting two high-energy gamma rays in the process. Since the gamma rays travel in nearly opposite directions, radiation detection devices arrayed in a circle around the organ of interest can detect the pairs of gamma rays and, with the aid of computers, locate their origin with remarkable precision. The crucial role of positrons in human autoradiography gave rise to the name positron emission tomography or PET (Phelps et al. 1975, Ter-Pogossian et al. 1975).

2.3 Magnetic Resonance Imaging (MRI)

Finally, another technology emerged contemporaneously with PET and CT. This was MRI. MRI is based upon yet another set of physical principles that have to do with the behavior of hydrogen atoms or protons in a magnetic field. These principles were discovered independently by Felix Block and Edward Purcell and colleagues in 1946 and expanded to imaging by Paul Lauterbur in 1973 (for review see Kevles 1997)).

Initially MRI provided superb anatomical information but inherent in the data also was important metabolic and physiological information. An opening for MRI in the area of functional brain imaging emerged when it was discovered that during changes in neuronal activity there are local changes in the amount of oxygen in the tissue (Raichle 1998). By combining this observation with a much earlier observation by Pauling and Coryell (1936) that changing the amount of oxygen carried by hemoglobin changes the degree to which hemoglobin disturbs a magnetic field, Ogawa et al. (1990) were able to demonstrate that *in vivo* changes in blood oxygenation could be detected with MRI. The MRI signal (technically known as T2* or 'tee-two-star') arising from this unique combination of brain physiology (Raichle 1998) and nuclear magnetic resonance physics (Pauling and Coryell 1936, Thulborn et al. 1982) became known as the blood oxygen level dependent or BOLD signal (Ogawa et al. 1990). There followed quickly several demonstrations of BOLD signal changes in normal humans during functional brain activation (Kwong et al. 1992, Ogawa et al. 1992) giving birth to the rapidly developing field of functional MRI or fMRI.

It is important to keep in mind that when a BOLD signal is detected blood flow to a region of brain has changed out of proportion to the change in oxygen consumption. When blood flow changes more than oxygen consumption, in either direction, there is a reciprocal change in the amount of deoxyhemoglobin present locally in the tissue changing the local magnetic field properties. As you will see both increases and decreases occur in the BOLD signal in the normal human brain.

3. Functional Brain Imaging

3.1 The Strategy of Functional Brain Imaging

As a result of collaboration among neuroscientists, imaging scientists, and cognitive psychologists, a distinct behavioral strategy for the functional mapping of neuronal activity emerged. This combination of disciplines and the strategies resulting from it probably account for the remarkable success and the resulting acceptance of functional brain imaging in cognitive neuroscience. The general strategy of functional brain imaging is based on a concept introduced by the Dutch physiologist Franciscus C. Donders in 1868 (reprinted in Donders (1969)). Donders proposed a general method to measure thought processes based on a simple logic. He subtracted the time needed to respond to a light (say, by pressing a key) from the time needed to respond to a particular color of light. He found that discriminating color required about 50 msec. In this way, Donders isolated and measured a mental process for the first time by subtracting a *control state* (i.e., responding to a light) from a *task state* (i.e., discriminating the color of the light).

One criticism of this approach has been that the time necessary to press a key after a decision to do so has been made is affected by the nature of the decision process itself. By implication, the nature of the processes underlying key press, in this example, may have been altered. Although this issue (known in cognitive science jargon as the *assumption of pure insertion*) has been the subject of continuing discussion in cognitive psychology, it finds its resolution in functional brain imaging, where changes in any process are signaled directly by changes in observable brain states. Events occurring in the brain are not hidden from the investigator as in the purely cognitive experiments. Careful analysis of the changes in the functional images reveals whether processes (e.g., specific cognitive decisions) can be added or removed without affecting ongoing processes (e.g., motor processes). Clearly, some areas of the brain active at one stage in a hierarchically designed paradigm can become inactive as task complexity is increased. Even more surprising, areas of the brain active in the resting state can also become inactive during certain cognitive states (Raichle et al. 2001). While changes of this sort are hidden from the view of the cognitive scientist they become obvious when brain imaging is employed.

A final caveat with regard to certain cognitive paradigms is that the brain systems involved do not necessarily remain constant through many repetitions of the task. While simple habituation might be suspected when a task is tedious, this is not the issue referred to here. Rather, when a task is *novel* and, more importantly, conflicts with a more habitual response to the presented stimulus major changes can occur in the systems allocated to the task. Such changes have both practical and theoretical implications when it comes to the design and interpretation of cognitive activation experiments. Functional brain imaging obviously provides a unique perspective that is unavailable in the purely cognitive experiment.

3.2 The Assumptions of Functional Brain Imaging

It is important to point out that the underlying assumptions of current brain-mapping studies with PET and fMRI do not embrace a modern version of phrenology. The phrenologists of the nineteenth century posited that single areas of the brain, often identified by bumps on the skull, uniquely represented specific thought processes and emotions. In contrast, modern thinking posits that single areas of the brain each contribute quite simple mental operations that form the elementary components of the observable behaviors. Observable behaviors and thought processes emerge through the cooperative interactions of many such areas. So, just as diverse instruments of a large orchestra are played in a coordinated fashion to

produce a symphony, a group of diverse brain areas, each performing unique elementary and unique mental operations, work together in a cooperative fashion to produce human behavior. The prerequisite for such analyses is the conviction that complex behaviors can be broken down into sets of constituent mental operations. This is not a new view. In 1909, the great neuroanatomist Korbinian Brodmann commented that 'With regard to the infinite variety and richness of the psyche one should rather envisage the situation that in each particular case supposed "elementary functional loci" are active in differing numbers, in differing degrees and in differing combinations. One must further accept that specific complex processes occur mainly in one locality and others mainly in another. One must therefore also assume a certain regional preference for higher activities, sometimes more in occipital and temporal areas, sometimes more in frontal. *Such activities are, however, always the result (and not merely the sum) of the function of a large number of suborgans distributed more or less widely over the cortical surface; they can never be the product of a morphologically or physiologically independent "center"*' (Brodmann 1909, 1994).

3.3 The Future of Functional Brain Imaging

Functional brain imaging took a major step forward with the introduction of fMRI. This resulted not only from the improved spatial resolution of the technique and the wide availability of MRI scanners, many now appearing outside of hospitals in departments of psychology, but also the ability to obtain information on the response of the brain to individual stimuli (Buckner et al. 1996). Earlier studies with PET required that stimulus types be 'blocked' in order to accommodate the time necessary to make a single measurement of blood flow (usually about 1 m). With fMRI stimulus types can be randomized in ways that are identical to those used traditionally by cognitive psychologists. Thus, following the completion of the study, image data can be sorted according to stimulus type or response characteristics (Brewer et al. 1998, Wagner et al. 1998), providing a much more sophisticated analysis of brain function in relation to behavior or stimulus type.

Even with the introduction of fMRI we still do not have an adequate description of the temporal dimension of responses occurring within the brain. We know from direct measurements of the responses of neurons in the brains of laboratory animals that information is processed among areas of the brain with time differences in the range of tens of milliseconds. This is far too fast to be captured with fMRI. Electrical techniques such as the EEG and MEG do capture this temporal dimension but do so with very poor spatial resolution. The hope, therefore, is to combine eventually the spatial dimension of responses captured with fMRI with the temporal dimension provided by EEG or MEG. This combination of techniques remains a technically challenging problem but progress has been made (Liu et al. 1998, Rosen et al. 1998).

Those readers interested in a general survey of the current state of functional brain imaging work might wish to consult a variety of new books that are appearing regularly (Moonen and Bandettini 1999; Toga and Mazziotta 1996, 2000).

See also: Cortical Activity: Differential Optical Imaging; Magnetoencephalography; MRI (Magnetic Resonance Imaging) in Psychiatry; Neuroanatomical Methods.

Bibliography

Brewer J B, Zhao Z, Desmond J E, Glover G H, Gabrieli 1998 Making memories: Brain activity that predicts how well visual experience will be remembered. *Science* **281**: 1185–7

Brodmann K 1909 *Vergleichende Lokalisationlehre der Grosshirnrinde*. Barth, Leipzig, Germany

Brodmann K 1994 *Localisation in the Cerebral Cortex*. Smith-Gordon, London

Buckner R L, Bandettini P A, O'Craven K M, Savoy R L, Petersen S E, Raichle M E, Rosen B R 1996 Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 14878–83

Donders F C 1969 On the speed of mental processes. *Acta Psychologia* **30**: 412–31

Ingvar G H, Risberg J 1965 Influence of mental activity upon regional cerebral blood flow in man. *Acta Neurological Scandinavica* **Suppl 14**: 183–6

Kety S 1960 Measurement of local blood flow by the exchange on an inert diffusible substance. *Methods Medical Research* **8**: 228–36

Kety S 1965 Closing comments. *Acta Neurological Scandinavica* **Suppl 14**: 197

Kevles B H 1997 *Naked To The Bone. Medical Imaging in the Twentieth Century*. Rutgers University Press, New Brunswick, NJ

Kwong K K, Belliveau J W, Chesler D A, Goldberg I E, Weiskoff R M, Poncelet B P, Kennedy D N, Hoppel B E, Cohen M S, Turner R, Cheng H M, Brady T J, Rosen B R 1992 Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 5675–9

Landau W M, Freygang W H J, Roland L P, Sokoloff L, Kety S S 1955 The local circulation of the living brain: values in the unanesthetized and anesthetized cat. *Transactions of the American Neurology Association* **80**: 125–9

Lassen N A, Hoedt-Rasmussen K, Sorensen S C, Skinhoj E, Cronquist B, Bodforss E, Ingvar D H 1963 Regional cerebral blood flow in man determined by Krypton-85. *Neurology* **13**: 719–27

Liu A K, Belliveau J W, Dale A M 1998 Spatiotemporal imaging of human brain activity using functional MRI constrained magnetoencephalography data: Monte Carlo simulations. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 8945–50

Moonen C T, Bandettini P A 1999 *Functional MRI*. Springer, New York

Ogawa S, Lee T M, Kay A R, Tank D W 1990 Brain magnetic resonance imaging with contrast depedent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America* **87**: 9868–72

Ogawa S, Tank D W, Menon R, Ellermann J M, Kim S-G, Merkle H, Ugurbil K 1992 Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 5951–5

Pauling L, Coryell C D 1936 The magnetic properties and structure of hemoglobin, oxyghemoglobin and caronmonoxyhemoglobin. *Proceedings of the National Academy of Sciences of the United States of America* **22**: 210–6

Petersen S E, Fox P T, Posner M I, Mintum M A, Raichle M E 1988 Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* **331**: 585–9

Petersen S E, Fox P T, Posner M I, Mintun M A, Raichle M E 1989 Positron emission tomographic studies of the processing of single words. *Journal of Cognitive Neuroscience* **1**: 153–70

Phelps M E, Hoffman E J, Mullani N A, Ter-Pogossian M M 1975 Application of annihilation coincidence detection to transaxial reconstruction tomography. *Journal of Nuclear Medicine* **16**: 210–24

Raichle M E 1998 Behind the scenes of functional brain imaging: A historical and physiological perspective. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 765–72

Raichle M E 2000 A brief history of human functional brain mapping. In: Toga A W, Mazziotta J C (eds.) *Brain Mapping. The Systems*. Academic Press, San Diego, pp. 33–75

Raichle M E, Fiez J A, Videen T O, MacLeod A M, Pardo J V, Fox P T, Petersen S E 1994 Practice-related changes in human brain functional anatomy during nonmotor learning. *Cerebral Cortex* **4**: 8–26

Raichle M E, MacLeod A M, Snyder A Z, Powers W J, Gusnard D A, Shulman G L 2001 A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 676–82

Rosen B R, Buckner R L, Dale A M 1998 Event-related function MRI: Past, present and future. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 773–80

Talairach J, Tournoux P 1988 *Coplanar Stereotaxic Atlas of the Human Brain*. Thieme Medical Publishers, New York

Ter-Pogossian M M, Phelps M E, Hoffman E J, Mullani N A 1975 A positron-emission tomograph for nuclear imaging (PETT). *Radiology* **114**: 89–98

Thulborn K R, Waterton J C, Matthews P M, Radda G K 1982 Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field. *Biochimica et Biophysica Acta* **714**: 265–70

Toga A W, Mazziotta J C 1996 *Brain Mapping. The Methods*. Academic Press, New York

Toga A W, Mazziotta J C 2000 *Brain Mapping. The Systems*. Academic Press, New York

Wagner A D, Schacter D L, Rotte M, Koutstaal W, Maril A, Dale A M, Rosen B R, Buckner R L 1998 Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science* **281**: 1188–91

M. E. Raichle

Functional Data Analysis

Most statistical analyses involve one or more observations taken on each of a number of individuals in a sample, with the aim of making inferences about the general population from which the sample is drawn. In an increasing number of fields, these observations are curves or images. Curves and images are examples of functions, since an observed intensity is available at each point on a line segment, a portion of a plane, or a volume. For this reason, we call observed curves and images 'functional data,' and statistical methods for analyzing such data are described by the term 'functional data analysis' (FDA), coined by Ramsay and Dalzell (1991). Though the individual methods and techniques are important, functional data analysis is also a general way of thinking, where the basic unit of information is the entire observed function rather than a string of numbers. A few illustrations will be offered in this article, but Ramsay and Silverman (1997) may be consulted for many more examples and methods, for consideration of the underlying philosophy, and for further references to the literature.

The goals of functional data analysis are essentially the same as for other branches of statistics, and include the following:

(a) to represent and transform the data in ways that aid further analysis,

(b) to display the data so as to highlight various characteristics,

(c) to study important sources of pattern and variation among the data, and

(d) to explain variation in an outcome or dependent variable by using input or independent variable information.

We shall illustrate the nature of functional data, these goals, and FDA tools available or under development, through a series of examples.

1. Human Growth Data: Looking at Velocity and Acceleration

Figure 1 shows the growth patterns of 10 Swiss boys involved in the Zurich Longitudinal Growth Study (Falkner 1960). Each boy's height is measured at 29 separate ages, but it is more instructive to think of this as a sample of 10 growth curves or functions than as a set of 290 numerical observations. Of particular interest in the study of growth is the way that growth speeds up and slows down at various stages of

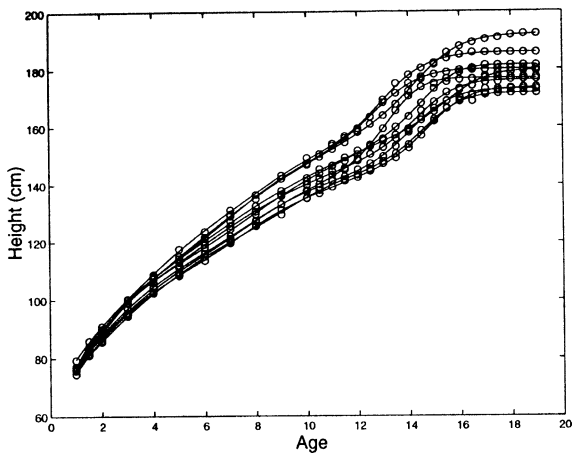

Figure 1
The heights of 10 Swiss boys measured at 29 ages
Notes: Points indicate unequally spaced ages of measurement; solid line is a function fitted to these data by monotone smoothing.

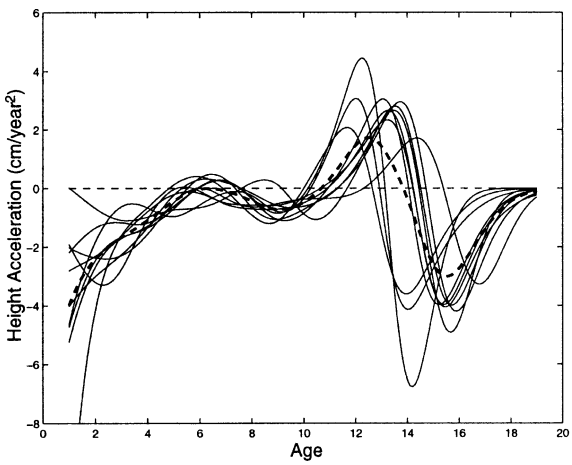

Figure 2
The estimated accelerations of height for 10 Swiss boys
Notes: Heavy dashed line is cross-sectional mean. This is a rather poor summary of the curves because of the presence of the timing of the pubertal growth spurt, called phase variation.

development, and this is much better illustrated by studying the velocity and acceleration of height. Figure 2 gives estimates of the acceleration or second derivative of these 10 height functions, and shows that around puberty there is a strong pulse of acceleration in growth, followed by sharp deceleration. This is the pubertal growth spurt (PGS), and there is also often a smaller spurt, hardly visible in Fig. 1 but obvious in Fig. 2, at around age 6.

2. The Mean Function and the Registration Problem

How would one estimate an overall acceleration curve to quantify the size of these spurts? The dashed curve in Fig.2 is the average acceleration over the sample at each time point, but the PGS for this average is smaller in size and longer in duration than that of any of the individual spurts. Averages are supposed to look like typical observations; so what went wrong?

The problem is that the curves vary in two ways, termed amplitude variation and phase variation. Alternatively, these are called vertical and horizontal variation, respectively, since amplitude concerns the height of a function at a given point, and phase the location in time of a feature such as an acceleration peak. In short, the actual acceleration peaks in the PGS vary in both size and timing. An artificial example demonstrating these two sorts of variation is presented in Fig. 3. The phase shift displayed in this figure is constant across the timescale, but many data sets require more complicated and nonlinear transformations of the time axis in order to align curve features like the peaks and valleys in Fig. 3. The presence of phase variation arises in FDA because many processes involve a system time, or in this case biological time, that does not unfold at the same rate as clock time. In this case, the biological clock of each boy marks off time at its own unique rate.

A key part of the FDA methodology is time-warping or registration of the data curves to align specific features or to minimize variability, since estimating the mean and most other analyses are thrown off by the presence of phase variation. The time-warping functions may be simple constants, linear functions, or more complicated, nonlinear

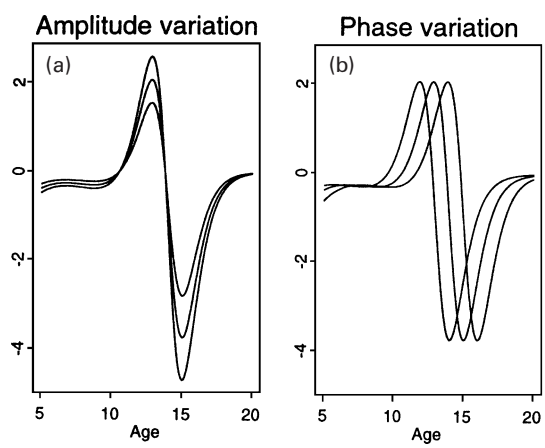

Figure 3
Panel (a): three height acceleration curves varying only in amplitude; Panel (b): three curves varying only in phase

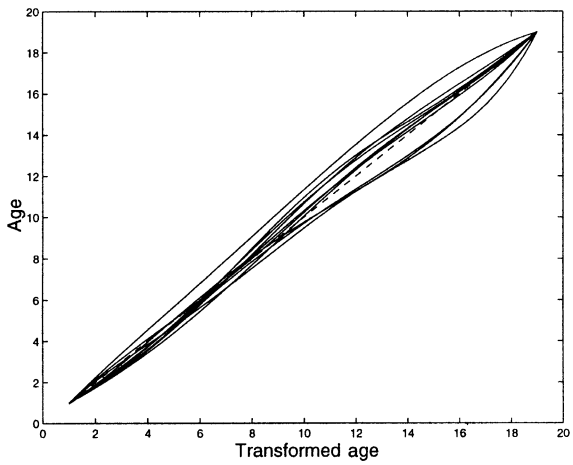

Figure 4
Time-warping functions taking chronological age on the horizontal axis into biological age on the vertical axis

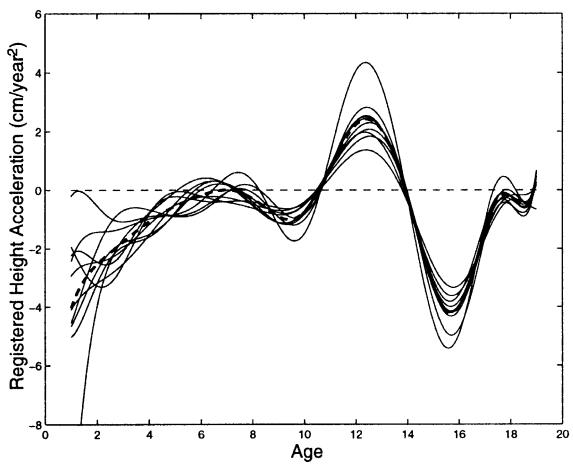

Figure 5
The height acceleration curves in Fig. 2 aligned or registered by measuring height with respect to the biological time shown in Fig. 4
Notes: Heavy dashed line is the mean acceleration. This provides a much better summary of the actual curves than Fig. 2

functions. The amount or shape of warp may be of great interest. See Ramsay and Silverman (1997, Chaps. 5 and 8). Figure 4 shows time-warping functions and registered height acceleration curves for the growth curves of Fig. 1 estimated using a method developed by Ramsay and Li (1998). The time-warping curve for each individual shows his maturation pattern. A warping function consistently above the diagonal dashed line implies that clock time lags behind development, and therefore development is late in clock terms. If height is measured relative to biological age, the curves in Fig. 2 are aligned as shown in Fig. 5. Differences in height acceleration after registration correspond to pure amplitude variation in growth patterns. The mean of the registered curves in Fig. 5 is a much better summary than the mean of the unregistered data in Fig. 1.

3. *The Nondurable Goods Index and More Derivatives*

Valuable information may be gained by plotting one derivative of a functional observation against another, a technique called phase/plane plotting. The US index of nondurable goods production is an important economic indicator, and considerable insight may be gained by plotting the acceleration of this index against its rate of increase. If the index had a smooth trend plus a sinusoidal seasonal variation, this plot would be approximately circular with suitable choice of scales. The typical year 1970 in Fig. 6 has three main cycles, of which the spring cycle is rather smaller than the summer and fall/winter cycles. Comparisons of plots for different years is instructive. During the stable period 1964–6, there is little year-to-year change, but the period 1973–5 in Fig. 6 shows considerable instability, due primarily to the end of the Vietnam War. A typical plot for a more recent year, 1994, shows that the whole pattern has changed, indicating structural changes in the methods and economics of production. For further details, see Ramsay and Ramsey (2001).

4. *Functional Principal Components Analysis*

Many FDA methods are adaptations of classical multivariate methods such as principal components analysis (PCA), linear modeling, and analysis of variance (see *Linear Hypothesis* and *Multivariate Analysis: Overview*). Functional PCA demonstrates the way in which a set of functional data varies from its mean, and, in terms of these modes of variability, quantifies the discrepancy from the mean of each individual functional datum. Figure 7 (fig. 6.2 in Ramsay and Silverman (1997)) shows the first four principal component weight curves for temperature records for 35 Canadian weather stations. The first mode of variability is overall temperature, but with larger effects in the winter; the second corresponds to variation in annual range of temperature; the third partly to a time-shift of the overall cycle, and the fourth to short, hot summers and long, mild winters, or vice versa. Individual weather stations can then be

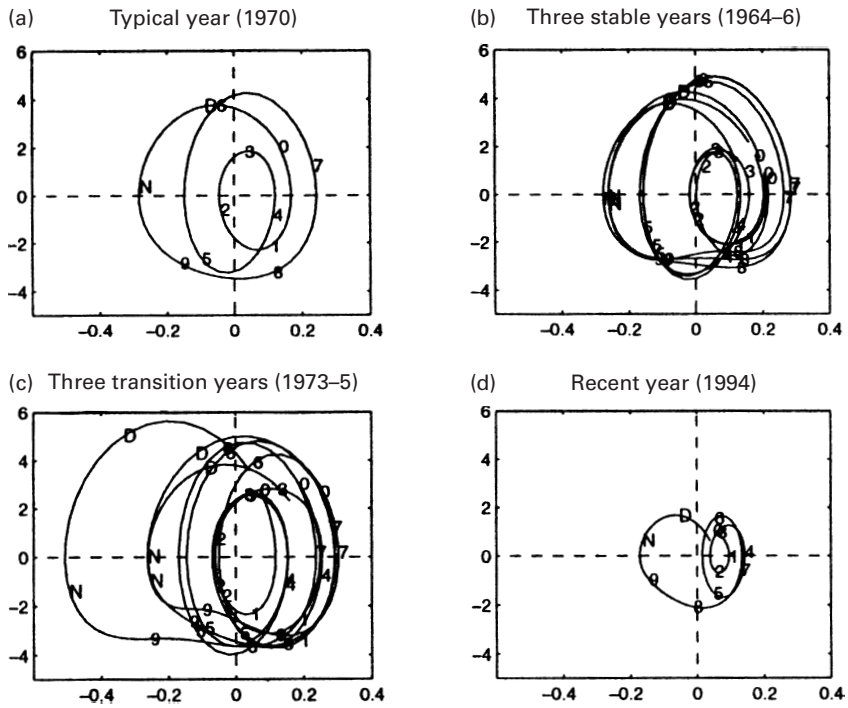

Figure 6
A phase/plane plot of *US index of nondurable goods production*. Each panel plots second derivative on vertical axis against first derivative on horizontal axis. Plots show structure of seasonal variation in index, and how structure varies with time

characterized according to their scores on these four criteria. The interpretability of the components can often be improved by a suitable rotation of the original principal components; see Ramsay and Silverman (1997, Chap. 6).

5. *A Functional Linear Model and Regularization*

An example of functional linear modeling is provided by the study of the dependence of the acceleration of the lower lip (lipacc) in speech on neural activity, as measured by electromyographical (EMG) recording. It is reasonable to suppose that the neural activity at time s continues to affect lip movement for only a short period of time, which we denote by δ and this leads to a model of the form

$$lipacc(t) = \alpha(t) + \int_{t-\delta}^{t} EMG(s)\beta(s,t)\, dt + \varepsilon\,(t),$$

developed by Malfait et al. (2000). The bivariate function $\beta(s, t)$ plays the same role as the regression coefficient in standard linear regression, and quantifies the relationship between the independent functional variable EMG, and the dependent functional variable lip. One particular issue is the appropriate choice of the time lag δ the length of time within the model that neural activity can continue to affect movement. Models for increasing values of δ correspond to nested linear models, and in this case a study of the amount of variation explained demonstrates that there is little effect beyond a lag of about half a second.

An important theoretical and practical issue is the necessity or otherwise of incorporating some form of smoothing or regularization into a functional analog of a classical multivariate procedure. In principal components analysis, the principal components, such as those displayed in Fig. 7, are smoothed or regularized and in some procedures, such as canonical correlation analysis (see *Multivariate Analysis: Overview*) or certain types of linear modeling for the lip/EMG data, regularization is necessary to give any meaningful results at all; see Leurgans et al. (1993), and Ramsay and Silverman (1997, Chaps. 11, 12).

6. *Modeling with Derivatives: A Central Theme*

It is the smoothness of the processes generating functional data that differentiates this type of data

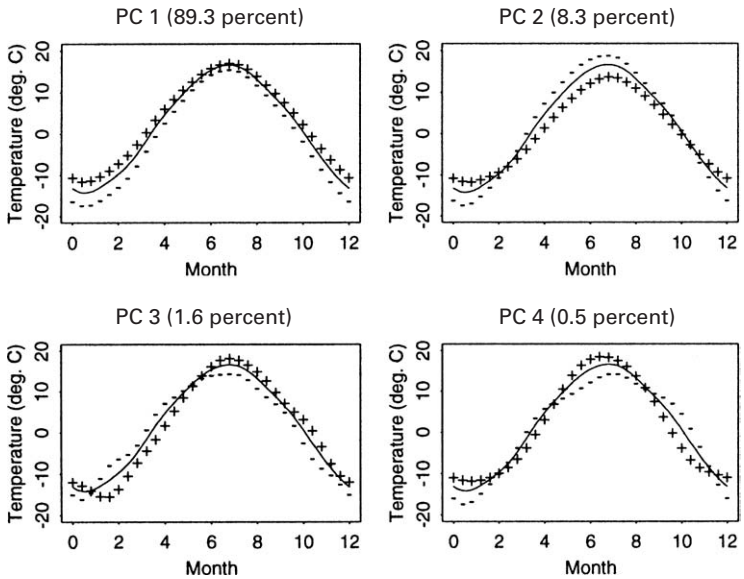

Figure 7
A principal components analysis of mean temperature curves for 35 Canadian weather stations
Notes: Solid line is the mean curve taken across all stations. The effects of adding (+) and subtracting (−) a suitable multiple of each component are also shown.

from more classical multivariate observations. This smoothness means that we can work with the information in the derivatives of functions or images. We looked at acceleration in the growth data, found something interesting in the joint variation of the velocity and acceleration of nondurable goods production, and used derivatives to define roughness penalties; see Ramsay and Silverman (1997, Chap. 4) for the weather data.

We may also use derivatives to construct models for data. Such models are called differential equations, and are part of the standard model-building toolkit in

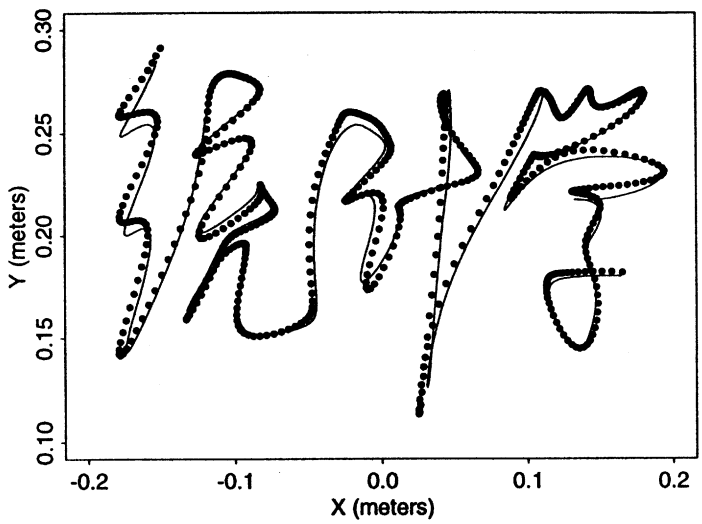

Figure 8
Observations of the tip of the pen while writing the word 'statistics' in Mandarin
Notes: Dots indicate 2400 observations; solid line is an approximation based on a second-order linear differential equation.

the natural sciences and engineering. Differential equation models are often called dynamic models, since they are especially effective at representing processes unfolding in time. Ramsay (2000) developed a linear differential equation for handwriting involving velocity, acceleration, and the third derivative of the tip of the pen. Figure 8 shows a sequence of pen positions during the writing of the word 'statistics' in Mandarin, and also shows as a solid line its approximation on the basis of the dynamic equation estimated from 50 such samples of Chinese script. The equation not only captured the essential features of this particular sample, but was also able to model a fair amount of the sample-to-sample variation as well. And, finally, it also provided an excellent account of the velocity, acceleration, and jerk functions at the same time. Ramsay and Silverman (1997, Chaps. 13, 14) take up this topic of dynamic modeling with functional data, including an account of a technique called 'principal differential analysis,' developed by Ramsay (1995).

7. *Relationships with Other Branches of Statistics*

Of course, statisticians have been exploring curves for a very long time, and there are many older statistical methodologies that can continue to inform and enrich an FDA. First of all, FDA usually begins with some kind of nonparametric regression (e.g., see *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)*) in order to replace discrete and possibly noisy curve values by smooth, continuous functions. Density estimation also involves data smoothing. Typically, time series analysis (Brillinger 1981, Box and Jenkins 1994) considers long sequences of curve values corresponding to equally spaced time values that are more or less stationary; that is, the data show little overall trend, and the covariance among neighboring values does not change much from one time region to another. However, methods for nonstationary processes are also under active development, and there are strong structural links between times series methods such as state–space modeling and FDA methods such as smoothing. Longitudinal data analysis or multilevel modeling (Diggle et al. 1994, Searle et al. 1992) usually looks at shorter sequences of curve values sampled over many cases, and studies the structure of between- and within-curve variation. It seems likely that this field and FDA will merge in a number of useful ways. Multivariate statistics has many methods developed specifically for functional data that have preceded FDA, but which have not explicitly exploited the smoothness of functional data or made other uses of derivatives (see *Probability Density Estimation*; *Longitudinal Data*; *Multivariate Analysis: Overview*).

8. *Conclusion*

By the smoothness of a functional observation we have meant the possibility that one or more derivatives, such as velocity or acceleration, may be examined or used in the modeling process. Otherwise, FDA makes only minimal assumptions about the structure of variation within and between curves. We have seen, however, that teasing apart amplitude and phase variation is a central issue, and one that has largely been ignored until recently. Certain types of models considered to be fairly routine and well understood in multivariate statistics, such as linear regression, have many more variants in the functional domain, and are being researched actively. Software for FDA is also being developed, and it is likely that FDA will soon be considered as a routine item in a data analyst's black bag of instruments and nostrums. Further information on FDA and related issues may be obtained in Pinheiro and Bates (2000), Ramsay and Silverman (1997), and Simonoff (1996).

Bibliography

Box G E P, Jenkins G M 1994 *Times Series Analysis: Forecasting and Control*, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ

Brillinger D R 1981 *Time Series: Data Analysis and Theory*. Holden Day, San Francisco

Diggle P J, Liang K-Y, Zeger S L 1994 *Analysis of Longitudinal Data*. Clarendon Press, Oxford, UK

Falkner F (ed.) 1960 *Child Development: An International Method of Study*. Karger, Basle, Switzerland

Leurgans S E, Moyeed R A, Silverman B W 1993 Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society, Series B* **55**: 725–40

Malfait N, Ramsay J O, Froda S 2000 The historical functional linear model. Unpublished manuscript. McGill University, Montreal, PQ

Pinheiro J C, Bates D M 2000 *Mixed-effects Models in S and S-PLUS*. Springer, New York

Ramsay J O 1995 Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society* **58**: 495–508

Ramsay J O 1998 Estimating smooth monotone functions. *Journal of the Royal Statistical Society, Series B* **60**: 365–75

Ramsay J O 2000 Functional components of variation in handwriting. *Journal of the American Statistical Association* **95**: 9–15

Ramsay J O, Dalzell C 1991 Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society* **53**: 539–72

Ramsay J O, Li X 1998 Curve registration. *Journal of the Royal Statistical Society, Series B* **60**: 351–63

Ramsay J O, Ramsey J B 2001 Functional data analysis of the dynamics of the monthly index of non-durable goods production. *Journal of Econometrics* (in press)

Ramsay J O, Silverman B W 1997 *Functional Data Analysis*. Springer, New York

Searle S R, Casella G, McCulloch C E 1992 *Variance Components*. Wiley, New York
Simonoff J S 1996 *Smoothing Methods in Statistics*. Springer, New York

J. O. Ramsay and B. W. Silverman

Functional Equations in Behavioral and Social Sciences

Theories in the behavioral, social, and natural sciences are often formalized by equations involving unknown functions, i.e., by functional equations. For instance, a theorist may be reluctant to make specific assumptions regarding the form of functions involved in a mathematical model, but the qualitative formulation of the model itself may impose constraints on the initially unknown functions. Such constraints often reduce the possibilities and occasionally are so severe that they restrict the possible forms to a few. A recent general reference to functional equations is Aczél and Dhombres (1989) where references to earlier and still useful surveys are given.

In the simplest cases there is one unknown, real-valued, function φ. An example, named after the famous French mathematician A. L. Cauchy, is the Cauchy equation $\varphi(x+y) = \varphi(x)+\varphi(y)$, where x, y are in the set of real numbers $\mathbb{R}$. More restrictive domains are sometimes studied. With no further restrictions, the solutions to the Cauchy equation can be wild. But assuming that φ is monotonic, or continuous, or bounded over a finite interval, the solution reduces to $\varphi(x) = cx$ for some constant c. By taking $\psi = \exp\varphi$, we get $\psi(x+y) = \psi(x)\psi(y)$, which is the so-called 'lack of memory property' and it is easy to see from the previous case that the strictly monotonic solutions are $\psi(x) = e^{cx}$, $c \neq 0$.

A celebrated psychophysical example is the connection between empirical just noticeable differences and G. T. Fechner's hypothesis that sensation corresponds to subjective differences being equal. By one interpretation of what he meant, one is led to the family of functional equations, named Abel equations after the famous Norwegian mathematician N. H. Abel, $\eta[x+g(k,x)]-\eta(x) = h(k)$, where the variable x and the parameter k are in the set of nonnegative real numbers $\mathbb{R}_+$. For the special case of Weber's law where $g(k, x) = (k-1)x$, this equation reduces to $\eta(kx) - \eta(x) = h(k) = \eta(k)-\eta(1)$, i.e., with $\varphi(z) = \eta(\exp z) - \eta(1)$, $k' = \log k$, $x' = \log x$, we have $\varphi(k'+x') = \varphi(k')+\varphi(x')$, which if k' is treated as a variable is Cauchy's equation. The restriction of strict monotonicity yields $\eta(x) = a\ln x+b$, $a \neq 0$, b constants, as the solution. For references and a general discussion of closely related issue see Falmagne (1985, Chap. 4).

Sometimes a functional equation has multiple, qualitatively different solutions. Yet the scientist arriving at the functional equation has the strong intuition that only one of these solutions is really appropriate for the scientific problem in question. Whenever this happens, the challenge is to discover additional behavioral properties that seem to be empirically correct and that serve to isolate the desired solution. A reason for fully determining the several solutions is that the same functional equation may arise in an entirely different empirical context and, for that context, one of the previously unacceptable solutions may be appropriate. So the complete characterization is clearly of interest.

As will be seen, functional equations arise in the social sciences in at least three main ways. One occurs when one knows how to measure numerically the same attribute in two different ways, which is the case more often than not. Then the two measures are related by an unknown strictly increasing function. An empirical law linking the two underlying measurement structures manifests itself as an equation restricting that unknown function.

Functional equations also arise when some invariance condition holds. The Fechner hypothesis above is an example. Others are given later.

Economic aggregation problems are a third source of functional equations. Consistent aggregation both rules out some *ad hoc* aggregation functions and leads to families of functions that are indeed consistent. Despite the intuitive appeal of additive aggregation, that assumption is inconsistent with the most common production functions used in economics. Examples are given of production functions and of permissible aggregation rules.

1. Independent Measures of an Attribute

1.1 A Physical Example: Mass Measures

Consider mass measurement where one decides which of two objects x and y is more massive by observing which arm of an equal-arm beam balance drops (in a vacuum). This provides a mass ordering. One can construct a numerical measure m of mass, unique up to multiplication by a positive constant (ratio scale), by placing pairs of objects, denoted by $x \circ y$, on the pans. Under reasonable, empirically-testable assumptions m is additive over the operation of combining, i.e., $m(x \circ y) = m(x)+m(y)$. One can also construct a measure unique up to power transformations (log-interval scale) by varying the velocity of objects and determining which has greater momentum. Under empirically reasonable assumptions, the measure of momentum is multiplicative in powers of a mass and a velocity measure, i.e., $\alpha m'(x)^{\beta}v'(x)^{\beta}$, where $\alpha, \beta > 0$ but are otherwise unspecified. It can be shown that the two

measures m and m' preserve the same mass order, and so they must be related by a strictly increasing function ξ. A natural question is: What is the form of ξ?

The following qualitative distribution property is empirically justified. Suppose (x, v) denotes the object x moving at velocity v and that $(x, v) \sim (x', v')$ means the two objects have the same momentum. Suppose also that $(y, v) \sim (y', v')$. Then, distribution asserts $(x \circ y, v) \sim (x' \circ y', v')$. This condition forces ξ to be multiplicative in the sense of satisfying $\xi(rs) = \xi(r)\,\xi(s)$, where the variables are real numbers with $r > 0$, $s > 0$. By setting $\xi = \exp\varphi(\ln)$ we see from the strictly increasing solutions of the Cauchy equation that $\xi(r) = r^{\beta}$, $\beta > 0$. Thus, one mass measure is a power function of the other one. See Luce et al. (1990, p. 125).

1.2 Preferences Among Gambles

Something quite analogous occurs in studying people's preferences among uncertain alternatives, often called gambles. A binary gamble $(x,C;\, y,\bar{C})$ means that x is the consequence if event C occurs and y otherwise when a chance 'experiment' with possible outcomes $E = C \cup \bar{C}$ is conducted. The preference order over gambles, which is assumed to be connected and transitive, is denoted by $\succcurlyeq$. We assume there is a status quo e, and a consequence x is said to be a gain if $x \succcurlyeq e$ and a loss if $e \succcurlyeq x$. A well developed theory of preferences over gambles, which generalizes subjective expected utility (Savage 1954), was worked out during the 1980s (see Quiggin 1993). It establishes that under certain conditions a utility function U over gambles and a weighting function W over chance events exists such that for gains x, y, and $g = (x,C;\, y,\bar{C})$

$$U(g) = \begin{cases} U(x)W(C) + U(y)[1 - W(C)], & \text{if } x \succcurlyeq y \\ U(x)[1 - W(\bar{C})] + U(y)W(\bar{C}), & \text{if } x \preccurlyeq y \end{cases} \tag{1}$$

$U(g) = U[(x,C;\, y,\bar{C})]$ is usually written just as $U(x,C;\, y,\bar{C})$, which is a minor abuse of notation. Because we assume $x \sim (x,C;\, x,\bar{C})$, U on gambles is also on consequences as well. This representation is called rank-dependent utility because the weight used in the average depends upon the ordering of the consequences. The model can be extended to cover losses and mixed gains and losses, but we do not go into that here (see Luce 2000).

One can also consider having or receiving pairs of valued consequences and gambles. Suppose g and h are gambles (including, as special cases, pure consequences) and let $g \oplus h$ denote having or receiving both. This binary operation $\oplus$ is called joint receipt. Assume that the preference order $\succcurlyeq$ extends to the domain of joint receipts. It is plausible that, for gains, $\oplus$ and $\succcurlyeq$ satisfy the same formal properties as does mass where no change from the status quo, e, plays the role of an identity. Thus, there is a real-valued 'value measure' V that preserves the order $\succcurlyeq$ and is additive over $\oplus$, i.e., $V(x \oplus y) = V(x) + V(y)$.

A natural question is: How do U and V relate? Luce and Fishburn (1991) considered the following linking hypothesis (or putative 'law') called segregation: For all gains x, y and events C

$$(x,C;\, e,\bar{C}) \oplus y \sim (x \oplus y, C;\, y,\bar{C}) \tag{2}$$

Note that both sides yield the same thing under the same conditions, so this is a highly rational property. If we introduce the notation $G(u, v) = U[U^{-1}(u) \oplus U^{-1}(v)]$ one can show that segregation implies that for $u = U(x)$, $v = U(y)$, $w = W(C)$ and $u \geqslant v$

$$G(uw, v) = G(u, v)w + v(1 - w) \tag{3}$$

For commutative the solution to this functional equation (Luce and Fishburn 1991) is that for some real δ

$$G(u, v) = u + v - \delta uw \tag{4}$$

From this equation one can show that U and V are related as follows. If $\delta = 0$, then $U = \alpha V$. If $\delta \neq 0$, then either $U_+ = 1 - e^{-\kappa V}$ or $U_+ = e^{\kappa V} - 1$, where $U_+(x) = |\delta| U(x)$, $x \gtrsim e$. Thus, if one knows how V depends on the consequences, then one knows up to a constant how U_+ does also. Conditions are given that determine the form of V over money and W over probability. [In like manner, a theory for losses is possible with constant δ' and $U_-(x) = |\delta'| U(x)$, $x \prec e$.]

From these expressions and the additivity of V over $\oplus$ one can deduce formulas for the utility U over $\oplus$ and for gambles when the consequences are a gain and a loss. Some interesting non-bilinear expressions arise (Luce 2000).

Recently, Luce and A. A. J. Marley (summarized in Luce 2000) have suggested a generalized representation of binary gambles over gains, namely, for $x \gtrsim y \gtrsim e$,

$$\varphi_0[(x,C;\, y,\bar{C})] = \varphi_1[u_1(x)w_1(C)] + \varphi_2[u_2(y)w_2(\bar{C})] \tag{5}$$

where $\varphi_1(0) = \varphi_2(0) = 0$. The ranked-additive decomposition in this form has been axiomatized by Wakker (1991, 1993) and the separability of $u_1 w_1$ can be defended axiomatically for the first component (Luce 2000). Assume, as seems reasonable,

$$x \sim (x,E;\, y,\varnothing) \tag{6}$$

$$y \sim (x,\varnothing;\, y,E) \tag{7}$$

$$x \sim (x,C;\, x,\bar{C}) \tag{8}$$

where $\varnothing$ denotes the null event and E the universal one underlying the gamble. Then noting $w_1(E) = 1$, $w_2(\varnothing) = 0$, the first indifference yields $\varphi_0(x) = \varphi_1[u_1(x)]$. If we set $y = x$, $v = u_1(x)$, $w = w_1(C)$, $w_2(\bar{C})$

$= Q[w_1(C)] = Q(w)$, then the second of the above conditions implies $x \sim (x,\varnothing; x,E)$, and so $\varphi_1(v) = \varphi_0(x) = \varphi_2[u_2(x)]$ or rewriting $u_2(x) = \varphi_2^{-1}[\varphi_1(v)] = \psi(v)$. Thus, setting $\varphi = \varphi_1$ the third yields the functional equation

$$\varphi(v) = \varphi(vw) + \varphi(\psi^{-1}[\psi(v)Q(w)]),\ v \in [0,k[,\ w \in [0,1] \tag{9}$$

where the unknown functions φ, ψ are strictly increasing, and Q is strictly decreasing. Aczél et al. (2001) have determined all solutions to this equation. From these results, Luce and Marley (see Luce 2000) showed the following utility representation: For $x \succcurlyeq y \succcurlyeq e$, there exist constants $c > 0$ and $\mu > -1/k^c$ such that

$$U(x,C; y,\bar{C}) = \frac{U(x)W(C) + U(y)[1 - W(C)] + \mu U(x)U(y)W(C)}{1 + \mu U(y)W(C)} \tag{10}$$

where $U = (u_1)^c$ and $W = (w_1)^c$. If, however, one supposes that segregation holds, then one can show that $\mu = 0$, which of course is just the rank-dependent model.

2. Invariance Principles

2.1 Value Function for Money

Section 1.2 left open the question of how V depends upon money. The most obvious assumption is that if x, y are sums of money, then $x \oplus y = x + y$. Thus, $V(x) + V(y) = V(x \oplus y) = V(x + y)$, which is Cauchy's equation. Because V is strictly increasing, V is thus proportional to money. Some empirical work suggests that, in general, this hypothesis may be too strong (Thaler 1985). An alternative assumption is that $\oplus$ is invariant under proportional changes in money, i.e., for any $\lambda > 0$,

$$x \oplus y \sim z \Rightarrow \lambda x \oplus \lambda y \sim \lambda z \tag{11}$$

Using the additivity of V,

$$V^{-1}[V(\lambda x) + V(\lambda y)] = \lambda z = \lambda V^{-1}[V(x) + V(y)] \tag{12}$$

For $x,y \geqslant 0$, functional equation arguments lead, through multiplicative functions, to $V(x) = \alpha x^\beta$ and $x \oplus y = (x^\beta + y^\beta)^{1/\beta}$, for some $\alpha > 0$, $\beta > 0$. This result, combined with those of Sect. 1.2, gives U to within one additional free parameter β (note that κ of U_+ and α are not separately identifiable).

2.2 Prelec's Weighting Function

Section 1.2 also left open the question of the form of the weighting function W. When the gambles are characterized in terms of probabilities $p = \Pr(C)$, much empirical work shows that in general for some p_0 in]0, 1[

$$W(p) \begin{cases} > p, & \text{if } 0 < p < p_0 \\ = p, & \text{if } p = p_0 \\ < p, & \text{if } p_0 < p < 1 \end{cases} \tag{13}$$

with W strictly increasing initially concave, and then convex. Such weighting functions are described as inverse S-shaped. Until Prelec (1998), suggestions of the mathematical form have not been derived from behavioral principles. He provided a somewhat complex behavioral invariance condition leading to $W(p) = \exp[-\gamma(-\ln p)^\alpha]$. Gonzalez and Wu (1999) showed that the Prelec form fits empirical data remarkably well. Not only is this function inverse S-shaped for $\alpha < 1$, but for $\alpha = 1$ it includes the 'rational' weighting function that arises from the conditions that gambles of the form $(x,p; 0)$ have a separable representation $U(x,p; 0) = U(x)W(p)$ and that $((x,p; 0),q; 0) \sim (x,pq; 0)$ holds.

Luce (cited in Luce 2000) gave a simpler condition, called reduction invariance, equivalent to Prelec's function which generalizes a property of the rational condition to the non-rational case: For all gains x and p, q, r in [0, 1],

$$((x,p; 0), q; 0) \sim (x,r; 0) \text{ implies, for } N = 2,3, \text{ that } ((x,p^N; 0),q^N; 0) \sim (x,r^N; 0) \tag{14}$$

From this, one first shows that the right side holds for N replaced by any real exponent $\lambda > 0$. Then, using separability, one arrives at the functional equation

$$(W^{-1}[W(p) + W(q)])^\lambda = r^\lambda = W^{-1}[W(p^\lambda) + W(q^\lambda)] \tag{15}$$

Note that with $V = \ln W(\exp)$, $x = \ln p$, $y = \ln q$ this reduces to the same functional equation as that for the value of money.

2.3 Additive Conjoint Weber Laws

Consider stimulus pairs (a, x) such as a tone of intensity a in a background noise of intensity x. Denote by $P(a,x; b,y)$ the probability that a subject perceives a in x more easily than b in y. Theoretical psychophysicists wish to understand the mathematical structure of such families of probabilities.

To do so, they explore various regularity or invariance conditions. For example, a homogeneity condition called a 'conjoint Weber law' assumes that for all real $\lambda > 0$

$$P(\lambda a, \lambda x; \lambda b, \lambda y) = P(a, x; b, y) \tag{16}$$

That is, the probability is invariant under multiplication of all four intensities by any positive number λ. A second important hypothesis is that of 'additive probability' in the following sense: There are con-

tinuous functions f strictly increasing, r strictly decreasing, and H strictly increasing in the first and strictly decreasing in the second variable such that

$$P(a,x;\, b,y) = H[f(a)+r(x), f(b)+r(y)] \tag{17}$$

In place of the last equation, Falmagne and G. Iverson (see Falmagne 1985, pp. 289–91) assumed the weaker condition

$$P(a,x;\, b,y) = H\,[q(a,\, x),\, g(b,\, y)] \tag{18}$$

of simple scalability where H is as above and g is strictly increasing in the first and strictly decreasing in the second variable, and the variable pairs (c, z) lie in subsets closed under multiplication by any $\lambda > 0$. They showed that then P must be of one of two forms:

$$P(a,x;\, b,y) = M\left[\frac{T(a/x)x}{T(b/y)y}\right] \tag{19}$$

or

$$P(a,x;\, b,y) = Q\left(\frac{a}{x}, \frac{b}{y}\right) \tag{20}$$

where $T > 0$, M, and Q are continuous, T and M are strictly increasing, and Q is strictly increasing in the first and strictly decreasing in the second variable.

In particular, they determined the form of all representations that are both additive and satisfy a conjoint Weber law as follows. The last equation when rewritten as

$$P(a,x;\, b,y) = Q(e^{\ln a - \ln x},\, e^{\ln b - \ln y}) \tag{21}$$

is additive with

$$H(u,\, v) = Q(e^u, e^v),\, f(a) = \ln a,\, r(x) = -\ln x \tag{22}$$

The first homogeneous, simple scalable form is additive if and only if

$$M\left[\frac{T(a/x)x}{T(b/y)y}\right] = H\,[f(a)+r(x), f(b)+r(y)] \tag{23}$$

Holding $b = b_0$, $y = y_0$ constant and letting

$$K(t) = T\left(\frac{b_0}{y_0}\right) y_0 M^{-1}(H[t, f(b_0)+r(y_0)]) \tag{24}$$

we get

$$K[f(a)+r(x)] = T\left(\frac{a}{x}\right) x \tag{25}$$

Thus, for all $\lambda > 0$,

$$K[f(\lambda a)+r(\lambda x)] = \lambda K[f(a)+r(x)] \tag{26}$$

This functional equation and its method of solution is similar to but somewhat more complicated than those in Sects. 2.1 and 2.2. Moreover, it has two distinct sets of solutions:

$$f(a) = Aa^{\beta}+B, \quad r(x) = Cx^{\beta}+D \tag{27}$$

and

$$f(a) = A\ln a+B, \quad r(x) = C\ln x+D \tag{28}$$

Accordingly, for all homogeneous, additive, conjoint Weber laws, the probabilities are of one of the following three forms:

$$P(a,x;\, b,y) = Q\left(\frac{a}{x}, \frac{b}{y}\right) \tag{29}$$

$$P(a,x;\, b,y) = G\left(\frac{a^{\beta}+\gamma x^{\beta}}{b^{\beta}+\gamma y^{\beta}}\right) \tag{30}$$

$$P(a,x;\, b,y) = G\left(\frac{ax^{\gamma}}{by^{\gamma}}\right) \tag{31}$$

where Q is as above, G is continuous and strictly increasing, and $\beta > 0$, $\gamma \neq 0$ are constants.

Results of this character are useful both in understanding the implications of general invariance principles and in suggesting to experimentalists how to plot and analyse their data.

3. *Aggregation*

Consider an industry with m producers using n inputs. Denote by x_{jk} the amount of inputs k used by producer j. Let $y_j = F_j(x_{j1}, \ldots, x_{jn})$ denote the output of producer j and $z_k = G_k(x_{1k}, \ldots, x_{mk})$ denote the total consumption of input k. Consistent aggregation holds if there is a macro production function F and macro input function G such that

$$\begin{aligned} F(z_1, \ldots, z_n) &= F\,[G_1(x_{11}, \ldots, x_{m1}), \ldots, G_n(x_{1n}, \ldots, x_{mn})] \\ &= G[F_1(x_{11}, \ldots, x_{1n}), \ldots, F_m(x_{m1}, \ldots, x_{mn})] \\ &= G(y_1, \ldots, y_m) \end{aligned} \tag{32}$$

This equation is called $m \times n$ generalized bisymmetry. Although a full treatment deals with general domains (see, e.g., Aczél and Maksa 1996, Aczél 1997), here it is assumed that all of the variables and functions lie in the set of non-negative real numbers $\mathbb{R}_+$. Assuming

that all functions are strictly monotonic and continuous, Maksa (1999) has proved that there are continuous functions α_k, β_{jk}, γ_j, and φ, $j = 1,\ldots, m$, $k = 1, \ldots, n$ such that

$$F(z_1, \ldots, z_n) = \varphi^{-1}[\alpha_1(z_1) + \cdots + \alpha_n(z_n)] \quad (33)$$

$$F_j(x_{j1}, \ldots, x_{jn}) = \gamma_j^{-1}[\beta_{j1}(x_{j1}) + \cdots + \beta_{jn}(x_{jn})] \quad (34)$$

$$G(y_1, \ldots, y_m) = \varphi^{-1}[\gamma_1(y_1) + \cdots + \gamma_m(y_m)] \quad (35)$$

$$G_k(x_{1k}, \ldots, x_{mk}) = \alpha_k^{-1}[\beta_{1k}(x_{1k}) + \cdots + \beta_{mk}(x_{mk})] \quad (36)$$

Aggregation functions of this quasi-additive type had been obtained earlier, but under stronger regularity conditions (for a survey see Aczél 1997). Maksa's achievement was to get the same results under appreciably weaker conditions.

One way to interpret this result is that if the variables and functions are 'properly' measured, i.e., as $\varphi(F)$, $\varphi(G)$, $\gamma_j(F_j)$, $\alpha_k(G_k)$, $\alpha_k(z_k)$, $\beta_{jk}(x_{jk})$ and $\gamma_j(y_j)$, then aggregation is additive. This is, however, very different from assuming that the 'raw' inputs and outputs are related additively. Consider the following examples. Two commonly used production functions are the Cobb-Douglas (CD) function

$$F(z_1, \ldots, z_n) = a z_1^{c_1} z_2^{c_2} \ldots z_n^{c_n}, \quad a, c_k > 0 \quad (37)$$

and the Constant Elasticity of Substitution (CES) Function

$$F(z_1, \ldots, z_n) = (c_1 z_1^b + \cdots + c_n z_n^b)^{1/b}, \quad c_k > 0,\ b \neq 0 \quad (38)$$

Aczél and Maksa (1996) have shown that if any one of the aggregation functions is simple addition at the level of the 'raw' inputs or outputs, then it is impossible for any production function to be either CD or CES with $b \neq 1$. This does not say that inputs (or outputs) cannot actually be additively aggregated. Rather, it says that such additive aggregation is not consistent with common and realistic production functions. (One can prove that it is consistent only with linear production functions.)

Another question addressed is: If the F_j functions are in some sense of the same form, does that dictate that the macro F is also of that form? This is called the representativeness problem. The answer depends upon exactly what one means by 'the same form.' In a trivial sense, the answer is 'Yes' since the F and F_j's in the above quasi-additive solution have roughly the same structure. A somewhat more sophisticated answer is that it need not be so. For example, if all of the F_j are CD functions, then

$$\gamma_j(y) = \ln\left(\frac{y}{a_j}\right), \quad \beta_{jk}(x) = c_k \ln x \quad (39)$$

Or if they are all CES functions, then

$$\gamma_j(y) = y^b, \quad \beta_{jk}(x) = c_k x^b \quad (40)$$

Neither conclusion restricts φ and α_k in any way. Thus one may choose F to be CD or CES or neither.

So the general conclusion from solving the aggregation problem is that consistent and even representative aggregation is possible only for appropriately chosen functions. In general neither is possible if the aggregating functions are preselected in the 'wrong way.' The implications for the possibility of macro production models are considerable.

Related ideas and results concerning aggregation of probabilistic models of choice, with results for Luce's choice model as a special case, appear in Aczél et al. (1997).

4. Conclusions

For many static questions in the behavioral and social sciences, functional equations serve as a powerful tool in understanding theoretical constraints. Three types of constraint have been illustrated: (a) the linking of measurement structures that have a common qualitative attribute; (b) the imposition of invariance principles of various types of which three were illustrated; and (c) the question of whether one can aggregate over inputs and separately over industries and arrive at the same macro conclusion.

See also: Algorithms; Axiomatic Theories; Mathematical Psychology; Measurement, Representational Theory of; Measurement Theory: Conjoint; Ordered Relational Structures; Psychophysical Theory and Laws, History of; Utility and Subjective Probability: Contemporary Theories

Bibliography

Aczél J 1997 Bisymmetry and consistent aggregation: Historical review and recent results. In: Marley A A J (ed.) *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*. Erlbaum, Mahwah, NJ

Aczél J, Dhombres J 1989 *Functional Equations in Several Variables with Applications to Mathematics, Information Theory and to the Natural and Social Sciences*. Cambridge University Press, Cambridge, New York

Aczél J, Maksa G 1996 Solution of the rectangular $m \times n$ generalized bisymmetry equation and of the problem of consistent aggregation. *Journal of Mathematical Analysis and Applications* **203**: 104–26

Aczél J, Maksa G, Marley A A J, Moszner Z 1997 Consistent aggregation of scale families of selection probabilities. *Mathematical Social Sciences* **33**: 227–50

Aczél J, Maksa G, Ng C T, Páles Z 2001 A functional equation arising from ranked additive and separable utility. *Proceedings of the American Mathematical Society* **129**: 989–98

Falmagne J-C 1985 *Elements of Psychophysical Theory*. Oxford University Press, New York
Gonzalez R, Wu G 1999 On the form of the probability weighting function. *Cognitive Psychology* **38**: 129–66
Luce R D 2000 *Utility of Gains and Losses: Measurement–Theoretical and Experimental Approaches*. Erlbaum, Mahwah, NJ
Luce R D, Fishburn P C 1991 Rank- and sign-dependent linear utility models for finite first-order gambles. *Journal of Risk and Uncertainty* **4**: 25–59
Luce R D, Krantz D H, Suppes P, Tversky A 1990 *Foundations of Measurement*. Academic Press, New York Vol. III
Maksa G 1999 Solution of generalized bisymmetry type equations without surjectivity assumptions. *Aequationes Mathematicae* **57**: 1–25
Prelec D 1998 The probability weighting function. *Econometrica* **66**: 497–527
Quiggin J 1993 *Generalized Expected Utility Theory: The Rank-dependent Model*. Kluwer Academic Publishers, Boston
Savage L J 1954 *The Foundations of Statistics*. Wiley, New York
Thaler R H 1985 Mental accounting and consumer choice. *Marketing Science* **36**: 199–214
Wakker P P 1991 Additive representations on rank-ordered sets. I. *Journal of Mathematical Psychology* **35**: 501–31
Wakker P P 1993 Additive representations on rank-ordered sets. II. The topological approach. *Journal of Mathematical Economics* **22**: 1–26

J. Aczél and R. D. Luce

Functional Explanation: Philosophical Aspects

Functional explanation has received a great deal of philosophical attention in recent years, partly as a consequence of a surge of interest in the philosophy of biology, and partly as a consequence of attempts to make functions and functional explanation central to issues in the philosophy of psychology. In both of these areas there has been some controversy over what the correct account of functional explanation is, with the contending accounts splitting into two main camps. Both camps agree that functional explanation is dependent on an account of functions, but disagree on what the correct account is. Within evolutionary biology the favored view is that functions emerge over time as a consequence of evolutionary processes, so a certain feature of an organism will only have a function if it has a suitable history. The opposed view is that the operation of a system can be analyzed into the way in which parts of the system function to make the system behave in that manner. This version has many names, but is perhaps best known as a 'systems' or 'Cummins' account of functions and functional explanation, but it has also been called 'causal-role' functionalism. In what follows both of these views will be examined; in addition the controversy over what it is that functional explanations explain will be outlined.

1. Functions and Systems

The 'systems' account of functions takes as its starting point the fact of (biological) complexity and proposes to understand the complex functioning of a biological entity (or system) in terms of the working of its parts. The fact of complexity suggests that the systems are goal directed, and if this is taken for granted one can recover what is distinctive of many explanations in biology, that they invariably cite the effects of causes as explaining why the cause is there. It does this by specifying how a part of the system operates in the production of the aimed-at goal.

The best known account of how such an explanation operates is given by Cummins (1975, 1983). He distinguishes explanation of changes, which is done by 'transition' theories, from explanation of properties. Transition theories are suitable for explaining 'What caused S to acquire P?' whereas a property theory explains what it is for S to instantiate P. This is best done by constructing an analysis of S, such an analysis adverting to the properties of S's components and their organization. So the kinetic theory of gases explains what it is for a gas to have temperature via an account of the properties of the molecules contained within the gas. Cummins calls the analysis of such a system compositional analysis: the analysis of a dispositional property being called functional analysis. The general claim is that in science the ascription of a function to something is to specify a capacity, such a capacity being identified by its role in enabling a system to operate in a certain manner. Thus, in biology the capacities of an organism are explained by analyzing them into a number of systems (circulatory, digestive, etc.), such systems having a particular task to perform. How these tasks get performed is in turn explained by functional analysis, such an analysis citing the capacities of parts of these systems. Such capacities are the functions of the parts, so we get the Cummins-style definition of function:

> X functions as an F in s (or: the function of X in s is to F) relative to an analytical account A of s's capacity to G just in case X is capable of F-ing in s and A appropriately and adequately accounts for s's capacity to G by, in part, appealing to the capacity of X to F in s (Cummins 1975, p. 762).

This account clearly does not restrict the attribution of functions to biological organisms; one can recognize the attribution of this type of function to parts of societies within sociological and anthropological functionalism. Here the system was seen as a society or culture, and the parts were the subsocietal structures or the recurrent patterns of behavior, the function of which was to maintain the structural continuity of the whole. The fact that this account of functionality is applicable to any system has given rise to two criticisms of this approach, one seeing a problem in how to specify nonarbitrarily which behavior or capacity of the containing system is relevant for the

functional analysis to analyze, the second seeing a difficulty in producing a sufficiently determinate notion of a 'system' for which a functional analysis is appropriate. The problem is that the choice of both a capacity and a system may seem to depend upon our decisions in a way which could make the functional analysis of any capacity or system seem justifiable. The pertinent question here concerns the explanatory gain made by the appeal to functional analysis. There seems to be no explanatory pay-off in some cases where functional analysis may be applied. Take the 'system' which is my study, which includes as parts various items of furniture, books, and me. The property of the room that we want explained is its temperature. Say my presence in the study raises the temperature of the room by 1/2 a degree. One does not gain any insight into the working of this system by attributing to me the function of enabling the room (the system) attain its present temperature.

Cummins was aware of this trivialization of functional analysis, and suggested constraints on the choice of a system and capacity that could be analyzed illuminatingly. One problem with the above analysis is that the capacity to be explained via functional analysis, the temperature of the room, is similar in kind and complexity to the property doing the explaining, my body temperature. One gets an explanatory pay-off when the analyzing property is both less complicated than, and different in kind from, the analysandum. This suggests the third constraint imposed by Cummins: one is not going to get a less sophisticated property doing all the explanatory work in showing how a complex system operates. The analysis will have to appeal to the operation of component parts and their properties, their combined action producing the complex behavior of the system. It is through the organization of interdependent parts that simplicity generates complexity, and functional analysis succeeds when it shows how this can be achieved.

The choice of system and capacity to be analyzed, then, is constrained: functional analysis will be applied fruitfully only to a complex system whose behavior is produced by the action of simpler but well organized parts. It is unsurprising, then, that the systems most often favored as exemplars of systems suitable for functional analysis are those that have been designed. The interdependence of the parts of such systems is often attained by careful planning. The second theory of functions that will now be examined focuses on this 'design' aspect of the systems to whose parts functions can be aptly attributed.

2. *Functions and Selection*

It is the improbability of complex systems existing by chance that suggests that such systems are generated by design, and so have parts which function so as to produce the intended behavior of the system. With artifacts it is easy to see how functionality can be ascribed: the complex system is designed intentionally to have the capacity to produce certain effects, and the parts are chosen for their ability to contribute towards that goal. They are selected for that task. This type of functionality is derived from intentionality. The achievement of Darwin and Wallace was to provide a nonintentional explanation of how such complexity could originate, and they did this via the theory of natural selection. The major strength of the historical account of functions is that it ties attribution of functions to 'design' by natural (nonintentional) selection. The essential features of this account are dependent upon the idea that to attribute a function to an item is to say that the item has that feature or behavior because it has a certain history: in the past that feature produced certain consequences, those consequences themselves having the effect of reproducing items with that feature. It is this latter aspect which makes the account teleological. One can say that the item has the relevant property in order to produce those consequences. It is because of (some of) its effects that the item with this capacity is selected for. What the theory of natural selection does is make respectable the apparent purposefulness of the 'in order to.' It does this by showing how environmental pressures on variable items can produce varied reproduction of those items; those items with certain features will become more prevalent because those features have typical effects on the reproductive capacity of the items which bear them.

The long neck of the giraffe, to take a standard example, is an adaptation and has the function of enabling giraffes to eat the leaves at the top of tall trees. On the historical account, this capacity is the function of the long neck because previous giraffes whose genes produced longer necks than their contemporaries were better producers of offspring, and were so just because of the advantage conferred on those giraffes in the food wars. This results, in the long run, in more long-necked giraffes, until the giraffe is a long-necked species. Put as briefly and generally as possible, the etiological account says that functionality arises because some variants within a group (a mutant gene, for example) have novel features with capacities which are favourable to their possessors' ability to reproduce (i.e., more favourable than the nonpossession of those features by the other members of the group). Such features are transmitted to their descendants, proliferating within the group in the process. The feature of the organism will then, and only then, have as its function the exercise of the favorable capacity.

As with Cummins-style functions, this account of functionality is very general. It applies not just to biological entities, but to all items that have the required characteristics of:

(a) variability in the items to be selected,

(b) selection of some of those items, and
(c) variable reproduction of the items selected for.

Selection cannot take place if there is no initial variation; the same selection forces operating on a homogeneous population will have no effect. For a cause to be there in order to produce certain effects it must also be the case that instances of properties with those causal powers have been selected in the past because they produced those effects. The selection process requires that the causal property interact with salient features of the environment in such a way that other items without that property are seen to 'suffer' some loss as a consequence of their non-possession of the property. Finally, it is because the favored properties are transmitted to descendants of the items having those features that the present items have these properties. Then it can be said that the property is, or was functional, its function being to produce those types of effects which lead to its ancestors reproducing items with the property in question.

As the giraffe example indicates, it is in its forging a connection between function and adaptation that the strength of this account resides. Adaptations are produced historically via the operation of natural selection, so they are functional. A creature may have a feature which renders it 'adapted' to its environment because it so happens that that feature will help it flourish in that environment, but such a feature, lacking the relevant selectionist history, will be adaptive for the creature, but not an adaptation. The historical (etiological) theory of functions claims that it is because adaptations are selected, not accidental, features of organisms that we are entitled to the use of functional explanation.

The historical and systems account of functions are not necessarily competitors. It has been seen that the systems for which functional analysis is appropriate are complex, with the analysis proceeding by showing how the operation of simpler, interdependent parts can produce the relevant capacity of the system. Such systems, it can be claimed, come into being either by intention or by selection, so functional analysis will be appropriately applied in just those cases where some design-mechanism is operating. On this view, functional analysis is an accessory to the historical theory of functions. There are, however, two aspects of the historical conception that render it significantly different from the systems version. The first is a consequence of what has been suggested above: it is probable that complex systems are generated by some design process, but it is not impossible for such a system to emerge by chance. For such a system, a functional analysis would still be appropriate, but the historical version of functions would disallow the attribution of any function to the system or its parts, these not having the appropriate history. Second, the historical theory has the ability to include malfunctions within its purview. On the systems account, if a part 'functions' so as to produce a system with a different capacity from that expected, that part cannot be said to be malfunctioning. A functional analysis will proceed as usual, with the new capacity being analyzed, and the analysis showing just how it is produced given the nature of the contained parts, and their causal powers. On the historical conception, given a suitable account of reproduction, one can allow for mistakes to be made in the reproducing process, so that the items which are considered to be descendants of some common ancestor need not be perfect copies of that ancestor, and in particular may not have the causal powers for which items of that type have been selected previously. The item may then not do what it is supposed to do—it may malfunction. To appreciate this better, and to make manifest the power of the explanations which the historical theory makes available, that theory needs to be fleshed out in more detail.

3. *Elaborations: Malfunction and Novelty*

An essential aspect of the historical account is that a function results from selection operating over a period of time on items that are descended from each other. What is it to be descended in the relevant sense? Initially one may think in terms of copying, one item being a descendant of another if it is an exact copy of the former, and is so as a result of some law-like process that ensures that such a replica is forthcoming. This is too strict a definition: for one thing, it would be impossible to provide a decent conception of a malfunctioning device if all descendants had to be exact copies of their anestors. A malfunction involves some mistake, so a device that malfunctions will not be an exact copy of a properly functioning device. One has to allow for an item being a descendant of another even though the copying process is not perfect. One way of making what is involved more precise is due to Millikan (1984, 1993), and it goes like this: individual A is a reproduction of individual B if and only if B has some determinate properties in common with A, and this correlation of properties can be explained by a natural law operating in that environment. These common properties are the reproductively established properties of B, and the devices sharing these properties form reproductively established families (refs.). (Only first order refs. reproduce in this manner. Higher order refs (horefs.) are produced by first-order refs. My and my ancestors' two-leggedness is the result of common genes, such genes being first-order refs. They have certain properties in common, this being ensured by the laws of genetic inheritance.) Direct proper functions are functions of devices that are members of refs. or horefs.

So how do we get malfunctioning members of refs. or horefs? Millikan suggests that a malfunctioning member of a ref. can be produced provided it has been produced by a device the proper function of which is to produce members of the ref. in question, that it is in

some aspects similar to other members of the ref., and that its production has an explanation which approximates (in some undefined degree) to a normal explanation for the production of members of the ref. Because a particular reproductive process need only approximate the 'perfect' reproductive process it can produce imperfect copies, ones which malfunction.

The historical theory's emphasis on the role of reproduction and selection has given rise to the thought that functional explanation, construed in this way, cannot explain novel behavior. On the historical account it is inheritance from past (selected for) patterns of behavior that gives rise to the present functional behavior. However, to suppose that novel behavior cannot be produced by something whose function it is to produce that behavior is to underestimate the variety of functions. Millikan notes that there can be a relational proper function, which is a function to do something bearing a certain relation to something else. The chameleon's skin-color mechanism has the relational function of making the chameleon's skin match that of its environment (whatever that color may be). Given a specific color to adapt to, the mechanism then acquires an adapted proper function. So when it's sitting on a brown plant the adapted proper function of the device is to produce brown skin. Crucial here is that the brown skin may have never been produced before, so it is not a member of a ref., and the production of it is not a direct proper function of the device. That to which an adapted device adapts is its adaptor (the brown plant being the adaptor in this case). One can extend this picture to include the functions of those who may be consumers of the output of an adapted function. Consider the dance of the bees which acts as a signal to other bees, 'telling' them to go southeast, where the nectar is to be found. This dance is adapted to the location of the nectar, so has an adapted function. The response it calls forth, the flight of the bees to the southeast, is adapted to this adapted function. It is not a short step to extend this approach to the communicative linguistic and other cultural capacities of human beings. Complications abound, but the fundamentals for doing this are in place. For the extension to be seen as a realistic prospect, two important features of these extensions of the historical approach need to be noted. The first is that functional explanation includes a normative aspect: a functional property is there in order to produce an effect, the effect being one it ought to cause, given that the property is instantiated here because past instantiations have produced that effect. Malfunctioning items are not doing what they ought to do. The extent to which functional explanation can account for psychological and sociocultural properties is partly dependent upon whether the normativity intrinsic to these latter properties can be explicated by the norms of functional explanation.

The second is that the above extensions indicate the explanatory reach of that theory. Any items (including patterns of behavior (e.g., rituals), social structures, ideas, 'memes'), can be the bearer of functions, provided that the basic conditions for this are met: variety, selection, varied reproduction. The reproduction need not be gene-centered, nor even broadly 'biological.' The transmission of bodies of knowledge and ways of behaving across generations is a form of reproduction, and so these items are possible bearers of functional properties. Whether they are or not depends on whether selection operates on these reproduced items, and here it is important to note that selection can take many forms. As behaviorists emphasized, behavior can be molded by painful experience, so pain is one selector. Emotional experience can be another. These have been called 'vicarious' or 'secondary' selectors (Campbell 1960), indicating that they themselves have been selected for by natural selection, their purpose being to be local, less severe eradicators of those types of behavior that are inimical to the flourishing of the organisms which produce the behavior. Given that reproduction and selection can be ubiquitous, it would appear that functional explanation has a much wider reach than once thought.

There have, though, been suggestions both that the explanatory scope of functional explanation has become too broad, that it is being abused by those who tend to see functionality everywhere, and that it is too narrow, not encompassing explanations of items which have no past history of selection, but are nevertheless beneficial to the organism. In the next section these criticisms of the historical theory are reviewed.

4. *Difficulties and Additions*

4.1 *Panglossianism*

The etiological theory is supposed to explain the persistence and spread of a trait in terms of its adaptive history, such explanations involving historical claims as to why the feature being explained was adaptive, and then tying this adaptation to its present function. At times it is extremely difficult to trace the relevant history, and the temptation is to ascribe functionality anyway, with the hope that the historical detail will be forthcoming. This has led to the accusation that such explanations are 'Panglossian,' presuming that this world is the 'biological best' world, one in which all features of organisms have some function or other (see Gould and Lewontin 1978). This accusation is particularly relevant to the sociobiological explanations of psychological and social behaviors, as the relevant history supporting these explanations is often unknown. The complaint is a welcome caution against epistemologically unconstrained attributions of function, but does not attack the heart of the historical theory. It simply warns against its abuse.

4.2 Exaptations

A related worry about attributing functions incorrectly, also voiced by Gould (Gould and Vrba 1982), concerns a trait that may be adaptive for one reason, but then change function in mid-stream, either doubling up in its functions or (because of environmental change maybe) sloughing off the old function in favor of the new, without any change of structure. Now the present persistence of the trait cannot be explained via its previous adaptive history, as that would give a misleading account of its present function. Such a change in function has been called an exaptation. An example of such an exaptation is the wing feathers of birds, the function of which we now think is to enable birds to fly. A selectionist story, however, would have to say that their function was as a thermal regulator, as that is what they were selected for originally. Clearly some complication of the original selectionist theory is called for, but the complications are hardly unexpected. Causal explanation and accounts of causality in general have difficulties dealing with causal overdetermination, and what we have in the above examples is functional overdetermination. What we need to say, in the functional case, is that the functions of an item are those fitness enhancing effects which advantage (or advantaged) the bearer of the item relative to rivals, and which have been reproduced differentially in the relevant population. With the change of function of feathers from thermal regulation to flight, the stage at which some feathers produced a flight-response would be critical in noting the acquisition of the new function. As soon as the feathered and flighted birds became preponderant, then one can say that the feathers had acquired a new function.

4.3 Dispositional and Virtual Selection

Once one notes that selection can be recent, a question arises as to how recent selection can be for functionality to be correctly attributed. To some extent the answer will be vague; there seems no exact stage at which one can say: before now there was no function, but now there is. But given that selection can be recent, a natural suggestion will be that present selection suffices for functionality. The radical proposal is that history can be irrelevant, that the essential component in the historical theory is the contribution of a feature to the reproductive success of the item possessing that feature, and that a trait without any history of selection can be seen to be useful for its possessor in the precise sense of making a (future, or possible) positive contribution to the survival or reproductive capacity of the organism. This 'dispositional' or 'virtual' selectionism involves the recognition that selection does play some role in the attribution of functionality, but it need not be actual selection. Differential reproduction need not have taken place. On this account, a trait is functional because of its potential contribution to the fitness of the organism possessing the trait, given the environment in which the organism operates.

A weak version of this view can be developed by requiring the environment to engage actively with the trait, so that the trait can be seen to contribute positively to the organism's fitness. On this view, suitably developed, the trait is seen as being currently under a selection pressure. That is, it is seen as potentially contributing to the differential reproduction of those organisms with the trait, where 'potentially' only means 'barring accidents, it will contribute to the differential reproduction of those organisms.' So the weak view agrees with the historical view in all respects except that it allows for functionality prior to differential reproduction. The stronger interpretation does not require actual engagement with the selection-relevant environment; it only requires that there are factors present in the environment that could influence the organism's chances of reproduction and survival.

The crucial question facing these amendments to the historical theory is: what is the advantage in using the notion of function here? What does it add to a simple nonfunctional causal explanation of what is happening? On the historical view, part of the distinctive force of functional explanations derives from the normative elements implicated in attributions of function, the idea that the functional trait is there in order to do something. The selectionist story, cashed out in terms of differential reproduction, makes this normativity naturalistically respectable. Without the selection occurring, this distinctive feature of functional attribution is threatened.

It is clear that these accounts cannot give any functional explanation for the existence, proliferation, or shape of the trait. Take the first generation of long-necked giraffes. The existence of the long-necks (their proliferation) cannot be given an explanation relating to the reproductively beneficial effect which the long necks have had in the past. The only explanation seems to be a straightforward causal explanation in terms of mutations. A suggestion is that one change the explanandum, which is what Pettit (1996) does in his advocacy of virtual selection. The idea is that there can be an explanation of a fact that is not an explanation of its existence or actual persistence, but an explanation of why it would persist in certain nonactual conditions, an explanation of its resilience. This would be an explanation of the robustness of the fact under various hypothesized circumstances. An example given is that of equilibrium explanation, where the explanation of, say, a certain ratio of prey to predator in a given environment need not explain why this ratio (which is in fact an equilibrium) has come about—why it exists—but why, if disturbing influences were brought to bear upon it, it would return eventually to this ratio. It is in this sense that the

resilience of the pattern is being explained. With functional explanation, the thought is that the absence of the operation of the function, the absence of the 'good' effect, may itself initiate certain mechanisms or processes that will reinstate the function's effect. So independently of how the trait performing the function emerged or proliferated, so long as it has this modal property of being re-instatable (in certain conditions) when disturbed, then the attribution of function will be in order. There will be something for that attribution to explain: namely the resilience of the pattern.

The problem for the advocate of functions here is that a normal causal story (perhaps one with feedback built into it) seems adequate to the task of explaining what is going on. The notion of function adds nothing, or at least, the plausibility it has derives only from the systems idea of functions. There is a system here which works so as to maintain an item, or a pattern of behavior, and any subsystemic processes which work so as to maintain the item or behavior 'functions' to that end. There is nothing more that functional explanation does here, and this notion of functional explanation does not have the distinctive normative character that the historical account imparts to functionality.

4.4 The Explanandum

There is some debate about what it is that a functional explanation explains. One suggestion is that, given selection only 'chooses' among the traits which are in existence, favoring some at the expense of others, all it can explain is why there are (so many) items with that feature. The proposal is that selection can not explain how individuals come to have the traits that they do have. The explanation for the traits possessed by any individual would be simply the causal story instantiated in the genetic processes that produced the trait in that individual. Sober (1984) called this latter type of explanation developmental explanation, and the good point he was making was that selection operates on traits already in existence.

One can accept this point, though, and still deny that all that functional explanation can explain is the very general fact concerning the proliferation rates of the causally explained traits. Neander has pointed out that that each such trait will have an evolutionary history, the elaboration of which will show how the trait came to have the form it now has (Neander 1995). Given the falsity of saltationist doctrines, traits such as opposable thumbs or long necks did not emerge fully fledged; there were many mutations, and those which increased, say, opposable thumbness were favored. A restriction to looking at 'selection of pre-existing traits' has almost an air of saltationism about it. It is as though the traits came into existence formed as they now are, and were then selected for on that basis. Ignoring the more gradual, at times almost imperceptible, changes that took place in order to arrive at the present trait is to ignore a part of the explanation as to why the trait is what it is. It is because functions can be performed by different mechanisms that attending to the historical detail is important in showing how and why this mechanism evolved into its present form. The central point is that although selection does only distribute characteristics, the history of successive distributions shows selection as also molding those characteristics. It is then possible to explain not only why people with thumbs were selected, but also the shape of the thumb, by citing facts about the shapes of its predecessors.

See also: Adaptation, Fitness, and Evolution; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Evolutionary Theory, Structure of; Explanation: Conceptions in the Social Sciences; Functionalism, History of; Functionalism in Sociology; Methodological Individualism in Sociology; Natural Selection

Bibliography

Campbell D T 1960 Blind variation and selective retention in creative thought and in other knowledge processes. *Psychological Review* **67**: 380–400

Cummins R 1975 Functional analysis. *Journal of Philosophy* **72**: 741–65

Cummins R 1983 *The Nature of Psychological Explanation*. MIT Press, Cambridge, MA

Gould S Lewontin R 1978 The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London, series B: Biological Sciences* **205**: 581–98

Gould S J, Vrba E S 1982 Exaptation—a missing term in the science of form. *Paleobiology* **8**(1): 4–15

Hull D, Ruse M (eds.) 1998 *The Philosophy of Biology*, Part IV. Oxford University Press, Oxford, UK

Millikan R G 1984 *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT, Cambridge, MA

Millikan R G 1993 *White Queen Psychology and Other Essays for Alice*. MIT, Cambridge, MA

Neander K 1995 Pruning the tree of life. *British Journal for the Philosophy of Science* **46**: 59–80

Pettit P 1996 Functional explanation and virtual selection. *British Journal for the Philosophy of Science* **47**: 291–302

Sober E 1984 *The Nature of Selection: Evolution Theory in Philosophical Focus*. MIT, Cambridge, MA

G. Macdonald

Functionalism, History of

Functionalism is a method of sociological analysis which has its roots in nineteenth-century social thought. It originated from the application of methods from biology, the rising science at that time, to the analysis of society. Three features are characteristic of

this orientation toward biology: society is more or less strictly conceived as an organism; its development is understood as an evolutionary process; special attention is given to the relationship between the parts and the whole constituted by them, particularly to functions which have to be fulfilled by the parts for the working of the whole. The parts are institutions such as forms of family life, economic production, government and religion as well as the organizational embodiments of such institutions: families, enterprises, governmental bodies, and churches. The institutions are binding rules which regulate the activities and relationships between people. The functions are, for example, biological reproduction for the family, generating resources and allocating them to needs for economic production, binding decisions for the collectivity for the government or the production of meaning for religion. It is assumed that, for its simple continuation, society needs the subserving of such functions by institutions and their organizational embodiments.

1. Origins of Functionalism: Auguste Comte and Herbert Spencer

It was the fundamental change from traditional agrarian to modern industrial society which constituted the background for the rise of functionalism in nineteenth century social thought, particularly in the work of Auguste Comte and Herbert Spencer. Comte (1830–42/1969) carried on the system developed by his former teacher, Saint-Simon, from whom he separated after having served him in his younger years. In his *Social Statics*, Comte conceived of society as a whole which, in its ideal state, is kept in perfect harmony by the balanced contribution of its constituent parts to its working in the process of fulfilling their specific functions. In his *Social Dynamics*, he conceptualized social change as an evolutionary process from lower to higher forms of societal organization. Because he expected science to bring about the perfect organization of social life in its final stage in modernity, he couched the evolutionary process in terms of the evolution of thinking from the theological to the metaphysical to the positive stage shaped by 'positive' science.

The other major source of functionalism in nineteenth century social thought next to Comte is Herbert Spencer (1909). Though he shared the adoption of biological and evolutionary thought with Comte, he demarcated his approach explicitly from Comte's by conceiving society and social development not as a matter of ideas, but as a matter of the objective things, namely the behavior of human individuals. The major process in the development of society is differentiation; functions become more numerous and the tendency is a function that is fulfilled by one specific institution specialized in it. The more society grows in size, the more differentiated will be its functions and corresponding institutions. The first differentiation occurs between regulative, sustaining, and distributive functions. The corresponding institutions are government, economy, and the infrastructure of channels for transport and communication. With the differentiation of functions and institutions, society works like higher organisms, it becomes more heterogeneous, definite in the demarcation of institutions, and integrated; group life is transformed from the coexistence of groups with the same structure without coherence between them (families, clans) to greater heterogeneity but also greater coherence between groups. The evolution of social life goes from incoherent homogeneity to coherent heterogeneity in the process of specialization and growing division of labor (see *Evolutionism, Including Social Darwinism*).

2. The Classical Framing of Functionalism: Emile Durkheim

Among the great classical founders of sociology at the turn of the nineteenth to the twentieth century, it is Emile Durkheim (1893/1973a, 1897/1973b, 1912/1968) who was most influenced by Comte and Spencer and who made a distinctive contribution to functionalism as a method of sociological analysis. From Comte he borrowed the holistic view of society, from Spencer the idea that the growing division of labor is a core feature of social evolution. In his *Rules of the Sociological Method*, Durkheim made a clear distinction between functional and causal analysis and a clear statement that every functional analysis needs to be complemented by causal analysis. He argues that the utility of an institution in serving a need for individuals or in performing a function for society does not suffice to explain its existence.

In his study on *The Division of Labor in Society*, Durkheim points out that specialization and division of labor are tools for the self-preservation of the individual and the society and for integrating society. They help to overcome the struggle for scarce resources. This struggle is caused by the increasing material and dynamic density that results from population growth and the extending as well as denser weaving of ways and means of transport and communication. Nevertheless, serving self-preservation and social integration do not produce specialization and the division of labor by themselves. In Durkheim's eyes, this functional explanation of specialization and division of labor needs a complementation by causal analysis. In doing so, he looks for forces which lead the selection of behavior and individuals in the direction of increasing labor division. He says that emigration, suicide, and crime would have been other possible solutions of the problem besides labor division. He explains the fact that they have not been selected to the same extent by the greater resistance called for by

them. The bonds to one's home country work against emigration, the bonds to one's life against suicide, the sympathy for our fellows prevents us from committing a crime. They are on average stronger and more resistant feelings than the habits which could prevent us from further specialization. The latter had to give way to the pressure toward specialization. In doing so, Durkheim points out the 'functional equivalents' which would serve self-preservation and then gives answers to the question as to the forces which lead human behavior in the direction of one of them. This is his complementation of functional analysis by causal explanation in his study on the division of labor.

In his later work on *The Elementary Forms of Religious Life*, Durkheim draws on anthropological studies on Australian aborigines. Here he makes use of his idea of the circular link between cause and functional effect in the study of ritual practices in which the sacred is strengthened in distinction to the profane and in which the community is regularly becoming conscious of itself. We have another circular link here. People gather because they want to influence the demonic powers positively for practical reasons of a good harvest or hunt, their willingness to come together is, however, supported by their feeling of belongingness and solidarity, which is again strengthened by the ritual itself.

3. A Paradigm of Functional Analysis: Robert K. Merton

Durkheim's analysis of tribal societies was continued in the anthropological study of tribal societies, particularly by the leading British anthropologists, Alfred R. Radcliffe-Brown and Bronislaw Malinowski. According to Malinowski (1948), every social phenomenon has to be explained by an indispensable function it fulfills not only for the whole of society, but also for the human individual: biological needs for food, shelter, and reproduction, psychological needs for support, needs for maintaining culture and social organization. Institutions of exchange, decision-making, social control, and education should be explained by such functions. Radcliffe-Brown (1952) criticized Malinowski's emphatic functionalism and preferred to call his approach structuralism. He did not accept the presumption that every social phenomenon exists because of a necessary function it fulfills. In his view, the unity of society was not a postulate claimed for every society, but a hypothesis which allowed a functional explanation of a social phenomenon by its contribution to the persistence of society. One of his examples is the explanation of lineage systems by the fact that they help to avoid conflicts about property rights after the death of property holders and thus contribute to the maintenance of solidarity. It is obvious that such an explanation is nevertheless insufficient in the light of Durkheim's methodological objections against purely functional explanations.

At this point, it is quite reasonable to turn to the detailed methodological analysis of functionalism contributed by Robert K. Merton (1949/1968) in his paradigm of functional analysis. Merton points out three postulates of functionalism, which all have to be doubted in his eyes. (a) Functional unity postulates that society is one consistent whole and every phenomenon contributes to its working and that there is no division of society into parts with conflicting needs. (b) Universal functionalism postulates that every social phenomenon performs a positive function for the whole of society. (c) Functional indispensability postulates that every social phenomenon performs an indispensable function for the whole of society and cannot be replaced by some other phenomenon. Merton says that all three postulates are unfulfilled by concrete societies, particularly by modern societies, so that functional explanations of social phenomena cannot be based on them.

Instead, Merton introduced a more realistic paradigm of functional analysis with a set of 12 requirements. (a) Only standardized social phenomena like institutions should be analyzed with regard to their functions. (b) Subjective dispositions of actors and objective consequences of their action in social practices should be outlined and distinguished. (c) Positive, negative (dysfunctional) and neutral effects of standardized practices should be discerned. (d) The units for which an item fulfills a function have to be exactly determined. (e) The functional requirements for maintaining a system have to be denoted. (f) The mechanism of their fulfillment must be specified. (g) The functional alternatives have to be enumerated. (h) The structural context needs description regarding the narrowing down of space for functional alternatives. (i) Dysfunctions have to be detected and analyzed with regard to effecting stress and strain with resulting pressures for change. (j) The functional explanations should be validated in comparative research. (k) Functional analysis must be neutral in ideological terms. (l) Functional analysis of a social practice needs to include; (i) the location of participants in the social structure; (ii) alternative modes of behavior excluded by the pattern; (iii) the emotive and cognitive meaning of the pattern for the participants; (iv) a distinction between motivation and objective behavior; (v) behavioral regularities that are part of the pattern but are not recognized by the participants.

A good example of Merton's idea is his functional explanation of the political machine by the various needs it fulfills for the whole of society and for a variety of groups. As an informal network with a boss at its head, it concentrates power and thus makes political decisions possible in the American political system of checks and balances; it also serves as a vehicle for helping the poor and weak, legitimate and illegitimate business, and provides channels for

upward mobility. Thus the political machine becomes stabilized because—according to Merton's requirement—the positive functions outweigh the dysfunctions by far.

4. *Functional Prerequisites and Social Stratification: Aberle et al., Davis and Moore*

In the 1940s and 1950s, functionalism achieved the dominant paradigm of sociological analysis under the name of 'structural functionalism.' The investigative strategy of structural functionalism was the discovery of the functions which have to be fulfilled in order to secure the survival of society and the discovery of the corresponding structures serving these functions (Aberle et al. 1950/1967). Another representative contribution to structural functionalism is the functional theory of social stratification outlined by Kingsley Davis and Wilbert E. Moore (1945) in their much debated article 'Some Principles of Social Stratification' published in the *American Sociological Review* in 1945. Davis and Moore argue that every society needs a system of social stratification for the following reasons. (a) The members of the society share common views on the importance of a certain position needed for the provision of public and private goods and services. (b) To perform the corresponding roles, more or less talent, training and/or effort is required. (c) Thus, positions can be more or less important, the personnel needed for filling the positions more or less scarce. In order to fill the positions there must be a differentiation of rewards in terms of money or prestige according to the importance of the position and the scarcity of the personnel available. The greater the importance and scarcity, the greater the reward. The result is that positions are always stratified. The representative criticism of this functional theory of stratification was formulated by Melvin M. Tumin (1953) in an article published in the American Sociological Review eight years later, in 1953. It is couched in terms of conflict theory: the class structure is based on the power of the dominant class, which defines the importance of positions; its major effect is the differentiation of access to the higher positions according to socialization advantages and training of the offspring originating from the dominant class. Rather than contributing to the working and integration of society, the stratification system has disruptive effects so that society is not stabilized by value consensus and stratification, but by the application of power.

5. *Functionalism as an Attempt at Synthesis: from Talcott Parsons' Structural Functionalism to Neofunctionalism*

Five years earlier than Davis and Moore, in 1940, Talcott Parsons published another functional theory of social stratification, which he revised twice, in 1954 and 1975 (Parsons 1954a, 1954b, 1977). In the perspective developed by Parsons, social stratification has to be explained by the interaction of dynamic and controlling forces of action, interests, and power on the one hand, and solidarities and cultural values on the other. Interests and power determine the access to positions and the enforcement of positional differences and outcomes; particularistic solidarities of family and lineage explain the inertia of class structures because of the tendency of families to secure their offspring remaining in the same class position. The *legitimacy* of the stratification system in distinction to its factual production and reproduction is, however, an effect of the consensus on a grading scale according to which prestige is attributed to positions. Only in as much as such a consensus on a grading scale is achieved will the system of stratification be accepted as legitimate. Otherwise, the definition of the importance of positions and the access to them will be a matter of the war of all against all. In general terms, this is the problem of double contingency.

Other than Merton's deliberate plea for empirical middle range theories, Parsons' program was analytical grand theory. Merton stood for empirical functionalism, Parsons for analytical functionalism (Parsons 1951). He introduced his four-function paradigm as an analytical instrument of analyzing structures as well as dynamic processes with regard to abstract functions (Parsons 1966, 1967, 1978, Parsons and Smelser 1956). By cross-tabulating the instrumental/consummatory and the internal/external orientation of action, he defined four abstract functions: instrumental/external: adaptation (A); consummatory/external: goal attainment (G); consummatory/internal: integration (I); instrumental/internal: latent pattern maintenance (L). For further differentiation, Parsons distinguishes three levels of applying the four-function scheme: social system: economic (A), political (G), community (I), fiduciary (L) system; general action system: behavioral (A), personality (G), social (I), and cultural (L) system; human condition: physico-chemical (A), organic (G), action (I), and telic (L) system. Further differentiation results from the application of the scheme to one of the subsystems. After having developed an analytical action theory in his first work of 1937, *The Structure of Social Action* (Parsons 1937/1968), Parsons established structural functionalism as a link between analytical functionalism and systems theory in *The Social System* in the early 1950s, and extended that link with the introduction of his theory of generalized media of interchange (money (A), political power (G), influence (I) and value-commitments (L) for the social system) and his theory of evolution as instruments for the analysis of dynamic processes of inflation and deflation and social change in the 1960s. With his turn to functionalism, systems, media, and evolutionary theory he did, however, not give up the core position

of his early action theory: the idea that social order emerges from the interaction between controlling and dynamizing forces. In his systems theory, he couched this idea in the double cybernetic hierarchy of conditions and controls. From L to A energy and dynamics in the system increase, while information and control decrease, and vice versa.

Looking at one analytical level, the social system, this system is composed of subsystems which fulfill their specific functions (A G I L) with specific structures (markets A, authority G, association I, and communication L) and specific media of interchange. They build an inter-related set of systems with in- and output processes of factors and products processed by the circulation of the media from one system to the other (interpenetration). Crises in the working of the systems can be analyzed as inflation–deflation spirals of the media involved.

In his theory of evolution, Parsons tries to point out the evolutionary advantage of specific institutions. Religion, communication by language, social organization through kinship, and technology are the first conditions of building human societies. A breakthrough to higher forms of social organization is realized by social stratification according to prestige based on achievement instead of ascription by origin. The rigidity of kinship systems is pushed back in this way. A further step forward is the replacement of tradition by cultural legitimation through general values. The developmental stage of a modern society is realized by the establishment of bureaucratic organizations, markets and money, universalistic norms, and democratic association. In two volumes, he outlines the evolutionary process in four dimensions according to his AGIL scheme: A adaptive upgrading, G structural differentiation, I inclusion and L value generalization.

Evolutionary achievements require breakthroughs in all four dimensions. He points out such breakthroughs in his analysis of primitive societies, archaic societies (Egypt and Mesopotamia), intermediate historic empires (China, India, Islamic, and Roman empires), seedbed societies (Israel, Greece) and, finally, the system of modern societies from ancient cultural roots via the Renaissance and the Reformation to the Enlightenment and the building of modern citizenship, state, and capitalism. The major modern breakthroughs are the industrial, the democratic, the educational, and eventually the 'expressive' revolution of the 1960s and 1970s (Parsons 1966, 1967, 1971).

According to his position that social order always has to reside in some form of consensus and solidarity in order to avoid an unlimited power struggle, Parsons lays special emphasis on the development of citizenship as a core of a modern 'societal community.' It is a modern form of community which has the greatest potential for inclusion, it associates people of the most varied ethnic and cultural origins. In this respect, it is the cornerstone of modern social organization. He sees the society of the United States farther developed in this respect than the ethnically more homogeneous European welfare states. The modern community of citizens is the core which is complemented by a capitalist economy, democratic government, and universalistic institutions of education, science, professional work, and voluntary religious association. This is the abstract path of sociocultural evolution, which has to be distinguished from concrete historical development that can more or less deviate from this path. There is particularly always the danger of fundamentalist movements against modernity, which originate from the severe tensions caused by the evolution of modern forms of social organization. For Parsons, fascism as well as communism were such fundamentalist reactions to modernity, which have, however, no chance of long-enduring stabilization in his eyes. Therefore, he expected that the Soviet system would crumble some time, an expectation grounded in 'abstract' evolutionary theory which turned out to be very 'realistic' eventually.

The dominant role played by Parsonian functionalism in the 1950s provoked countermovements which concentrated mostly on some particular aspect of social organization, which was relativized within the very general Parsonian scheme. The need for revitalizing the Parsonian approach has called for a new movement, which aims at continuing with the general program of Parsons on the contemporary level of development. This is the movement of neofunctionalism which arose in the 1980s. Neofunctionalism tries to advance the Parsonian legacy by way of learning from new developments in social theory and social research, for example, from advancements provided by Clifford Geertz or Jürgen Habermas. This new movement has contributed fresh interpretations of Parsons' work along with a new reading of the classics, studies on civil society and public discourse, on the evolution of the culture of modernity, on communication society, European integration and globalization (Alexander 1982/83, 1998, Münch 1982, 1998).

6. *Functionalism and Modern Systems Theory: Niklas Luhmann*

A comprehensive renewal of functionalism has been undertaken by Niklas Luhmann (1970, 1984, 1997). He, too, combines functionalism with systems, media, and evolutionary theory. In his early phase of theory building, Luhmann introduced a turn from structural-functional to functional-structural systems theory. To justify this turn, he claims that the question of function has to be posed in a more fundamental way. Before we ask for the structures of systems which have to fulfill certain functions, we should raise the question as to which function is served by systems-building as such.

What is the problem solved by systems building? This is how he asks his question. His answer is: world complexity which needs reduction so that it can be handled. Systems just perform this function of reducing world complexity.

In his second phase of theory building, Luhmann undertook another turn of his functionalist systems theory: the conceptualization of social systems as autopoietic, self-referential systems. Because there is always the danger of too much complexity, the major problem is systems building and the continued maintenance of systems boundaries, which are boundaries of meaning. Autopoiesis is a form of systemic operation which guarantees the maintenance of boundaries and, at the same time, allows for maximum absorption of environmental complexity. An autopoietic system is operationally closed, but open to any kind of complexity in the environment. Operational closure is guaranteed by a binary code according to which any kind of environmental complexity is processed. The concrete processing is organized by a system's program. Special links to other systems in the environment are stabilized by structural coupling. Generalized media of communication allow for most flexible forms of systemic processing: money, power, law, love, for example. Social systems are built up by communication, they originate from that and continue to exist as long as communication takes place according to the established code. Communication consists of three elements: information, message, and interpretation.

An autopoietic system produces and reproduces all its elements—communication—by itself, that is by systemic communication, but it makes use of environmental complexity. It does not take inputs from the environment to transform them into outputs, but receives impulses from the environment to which it shows more or less response. The system observes events in the environment and absorbs them according to its own code and program. Problems of maladaptation to the environment can be discovered by second-order observation, that is, self-observation and self-description.

Luhmann distinguishes three levels of social systems: interaction systems, organizational systems, and societal systems. According to his theory of evolution, more complex societal systems evolve by variation, selection, and stabilization. There are three (later four) stages of societal evolution which are defined by the form of differentiation of systems: segmentary, estate (later: center/periphery), and functional differentiation.

Modern societies are functionally differentiated and no longer have a center. The economic, political, legal, science, educational, arts systems are more and more differentiated on a world scale and thus transcend their centering around a societal community and/or a state. Thus the nation-state societies are transitional stages on the way to world society. Paying vs. not paying, having power vs. having no power, right vs. wrong, true vs. false are the binary codes of the economy, polity, law, and science systems. The evolutionary process works toward the establishment of functionally differentiated systems on a world scale, which are linked by structural coupling and by mutually stabilizing co-evolution.

7. *Future Need for Functionalism*

The question finally is: what will be the future of functionalism after its renewal by neofunctionalism and by Luhmann's functionalist systems theory? Whatever the movements against functionalism have contributed to our knowledge, we cannot say that they can replace the insights of functionalism completely. They have merely emphasized some particular aspect of social reality and some particular question. But they did neither raise nor answer the questions posed by functionalism from Comte and Spencer to Parsons and Luhmann. Because their questions are still legitimate, there is still a place for functionalist theorizing, whatever form it might take in the future.

See also: Action, Theories of Social; Biology's Influence on Sociology: Human Sociobiology; Comte, Auguste (1798–1857); Differentiation: Social; Durkheim, Emile (1858–1917); Evolution, History of; Explanation: Conceptions in the Social Sciences; Functional Explanation: Philosophical Aspects; Functionalism in Anthropology; Functionalism in Sociology; Law, Autopoiesis in; Luhmann, Niklas (1927–98); Malinowski, Bronislaw (1884–1942); Parsons, Talcott (1902–79); Social Stratification; Spencer, Herbert (1820–1903); System: Social

Bibliography

Aberle D F, Cohen A K, Davis K, Levy M J Jr., Sutton F X 1950/1967 The functional prerequisites of a society. In: Demerath N, Peterson R (eds.) *System, Change and Conflict*. Free Press, New York, pp. 317–31

Alexander J C 1982–83 *Theoretical Logic in Sociology*, 4 vols. University of California Press, Berkeley, CA

Alexander J C 1998 *Neofunctionalism and After*. Blackwell, Oxford, UK

Comte A 1830–42/1969 *Cours de philosophie positive*, 6 Vols. Culture et Civilisation, Brussels (1855 *The Positive Philosophy of Auguste Comte*. Calvin Blanchard, New York)

Davis K, Moore W E 1945 Some principles of stratification. *American Sociological Review* **10**: 242–9

Durkheim E 1893/1973a *De la division du travail social*. Presses Universitaires de France, Paris (1964 *The Division of Labor in Society* [trans. G Simpson]. Free Press, New York)

Durkheim E 1895/1973b *Les règles de la méthode sociologique*. Presses Universitaires de France, Paris (1982 *The Rules of the Sociological Method* [ed. Lukes S, trans. Halls W D]. Macmillan, London)

Durkheim E 1912/1968 *Les formes élémentaires de la vie religieuse*. Presses Universitaires de France, Paris (1976 *The Elementary Forms of Religious Life* [trans. J Swain W]. Allen & Unwin, London)
Luhmann N 1970 *Soziologische Aufklärung*. Vol. 1. Westdeutscher Verlag, Opladen (1982 *The Differentiation of Society* [trans. Holmes S, Larmore C]. Columbia University Press, New York)
Luhmann N 1984 *Soziale Systeme. Grundriß einer allgemeinen Theorie*. Suhrkamp, Frankfurt am Main
Luhmann N 1997 *Die Gesellschaft der Gesellschaft*, 2 Vols. Suhrkamp, Frankfurt am Main
Malinowski B 1948 *Magic, Science and Religion*. Doubleday, New York
Merton R K 1949/1968 *Social Theory and Social Structure*. Free Press, New York, pp. 73–138
Münch R 1982 *Theorie des Handelns. Zur Rekonstruktion der Beiträge von Talcott Parsons, Emile Durkheim und Max Weber*. Suhrkamp, Frankfurt am Main
Münch R 1998 *Globale Dynamik, lokale Lebenswelten. Der schwierige Weg in die Weltgesellschaft*. Suhrkamp, Frankfurt am Main
Parsons T 1937/1968 *The Structure of Social Action*. Free Press, Glencoe, IL
Parsons T 1951 *The Social System*. Free Press, Glencoe, IL
Parsons T 1954a *Essays in Sociological Theory*. Free Press, New York, pp. 69–88
Parsons T 1954b *Essays in Sociological Theory*. Free Press, New York, pp. 386–439
Parsons T 1966 *Societies. Evolutionary and Comparative Perspectives*. Prentice Hall, Englewood Cliffs, NJ
Parsons T 1967 *Sociological Theory and Modern Society*. Free Press, New York
Parsons T 1971 *The System of Modern Societies*. Prentice Hall, Englewood Cliffs, NJ
Parsons T 1977 *Social Systems and the Evolution of Action Theory*. Free Press, New York, pp. 321–80
Parsons T 1978 *Action Theory and the Human Condition*. Free Press, New York
Parsons T, Smelser N J 1956 *Economy and Society*. Free Press, New York
Radcliffe-Brown A R 1952 *Structure and Function in Primitive Society*. Macmillan, New York
Spencer H 1909 *The Principles of Sociology*, 3 Vols. Appleton, New York
Tumin M M 1953 Some principles of stratification: a critical analysis. *American Sociological Review* **18**: 387–94

R. Münch

Functionalism in Anthropology

In its simplest common-sense form functionalism in anthropology describes approaches which purport to explain cultural or social phenomena in terms of the purpose they serve.

Since the 1960s, specialized research into the sociological and historical framing of anthropology has revealed the extent to which discussions of functionalism contribute to analysis of knowledge as culture. As yet, however, it falls short of meeting the criticism that it tends to treat the discipline of anthropology as a closed discourse, despite calls that it be viewed as part of a larger discursive field in which anthropology operates (Asad, in Stocking 1991; Trouillot, in Fox 1991).

1. *The Intellectual History of an Idea*

The term 'function' was first introduced as a mathematical concept by Gottfried Wilhelm Leibniz in 1714, to refer to a variable considered in relation to one or more other variables in terms of which it might be expressed.

Two main bundles of ideas underlay the concept of 'function' that dominated academic anthropology in the first half of the twentieth century: (a) a relation of interdependence and (b) an orientation toward given ends. The first involved reciprocity and reversibility in terms of cause and effect. The ends might be the needs of the individual, as in the utilitarian functionalism of Malinowski (1884–1942) or the needs of society, as in the structural–functionalism of Radcliffe-Brown (1881–1955). Both adopted notions of culture or society in the modern pluralized usage as totalities with a certain unity and integrative character, a notion derived from the German philosopher Johann Gottfried Herder.

In the second half of the century, the premise remained that cultures and societies were structured wholes, now more consistently referred to as cultural or social systems. At this point functionalism's orthodoxy came under challenge in the aftermath to the Second World War which saw the withdrawal of sovereignty claims over a plethora of colonial territories within the British Empire and the establishment of United States 'trusteeship' in Oceania. This spurred British functionalists to address complex 'new nations' and provided American anthropologists *de novo* with a laboratory of functioning colonial cultures (Bashkow, in Stocking 1991).

By the 1970s, functionalism (as an achieved, shared paradigm which defined problems and procedures) was seen to be outmoded. It was replaced, amidst repeated calls for anthropology to reinvent itself, by struggle among new modes of discourse, chief among them Levi–Straussian structuralism, Marxism, and symbolic and interpretive anthropology. Nevertheless, echoes of functionalism lingered on into the decentered, fragmented, pluralizing, flexible, refractive, postmodern world of the new millennium.

2. *Functionalism's Multivocality*

A distinction among (a) functionalism as doctrine (or theory); (b) functionalism as method; and (c) functionalism as interest or ideology was a critical feature in Bronislaw Malinowski's modernist legacy to functionalism (Vincent 1991). This, along with the essential dualism between natural science and the humanities (particularly science and history) that entered mod-

ernist anthropology with Franz Boas's *The Study of Geography* (1897, reprinted in Stocking 1996), goes some way towards ordering the cultural complexity of functionalist knowledge. The scientific tendencies of the physicist who tries theoretically to reduce reality to simple elements and the humanist tendencies of the cosmographer who seeks to comprehend a complex phenomenon as a whole are both contained within the functionalist moment in anthropology.

2.1 Functionalism as Doctrine or Theory

Three cross-cutting anthropological trajectories may be recognized within functionalist anthropology.

2.1.1 Utilitarian functionalism (Malinowskian). Recent research has exposed the extent to which Malinowski's utilitarian functionalism, which dominated British anthropology between the wars, derived from his engagement with the neopositivism of Ernst Mach in his doctoral dissertation—particularly Mach's view that knowledge is a response to biological human needs (Thornton 1985; Thornton and Skalnik 1993). In his last, posthumously published, position paper, *The Functional Theory*, Malinowski takes his stand on two axioms: first, that every culture must satisfy the biological system of needs, and second, that every cultural achievement that implies the use of artifacts or symbols is an instrumental enhancement of human anatomy, and refers directly or indirectly to the satisfaction of a bodily need (Malinowski 1944, p. 171).

Strictly followed, these axioms would restrict cultural anthropology to the economism of the rationalizing individual (human nature), or the ecologism of selective advantage (external nature) (Sahlins 1976), but Malinowski's practice deviated considerably from his theorizing. As with Boas, his sentient functionalism lay immanent in his ethnography; neither man was deemed 'a theorist.'

2.1.2 Hybrid cultural functionalisms (Boasian/Parsonian). This complex functionalist trajectory, triggered by Franz Boas's (1858–1942) 'vaguely functionalist ideas' (Sahlins 1976, p. 71), drew explicitly on German humanist intellectual traditions. It comprised several contradictions which manifested themselves transgenerationally. Initially this took the form of Alfred Kroeber's (1876–1960) functionalist historicism and Robert Lowie's (1883–1957) neopositivist empiricism. A second generation was further divided; it was also developing in the context of the 'British' functionalist paradigm.

Thus, in 1934, Alexander Lesser (1902–82), recognizing functionalism as the most vigorous tendency extant in anthropology, set out to distinguish the particular interests and conceptions of individual American functionalists from the broader, foundational Boasian functionalist approach. He adopted a Leibnizian mathematical definition of a functional relation, established by observation, while arguing that contemporaneous events associated by the observer might be merely coexistences. Functional historicity recognized that 'institutions, customs, beliefs, artifacts have careers in time, and that their form and character is molded more by what has happened to them in the course of that history than by what particular things they occur associated with at any one time' (Lesser in Mintz 1985, p. 60). Radcliffe-Brown's position paper 'On the concept of function in social science' (1935) was prompted by his taking Lesser's critique personally.

It was actually directed at Lesser's fellow Boasians at Columbia University in New York, Ruth Benedict (1887–1948) and Margaret Mead (1901–78). Romantic holistic sensibility figured most prominently in Benedict's cultural configurations or patterns of culture. Mead, on the other hand (after corresponding with Radcliffe-Brown), designed a framework for the comparative functional analysis of social systems. This was published in *Cooperation and Competition among Primitive Peoples* (1937).

At the same time a more scientific anthroposociological functionalism was being developed at Harvard where sociologist Talcott Parsons (1902–79) joined forces with anthropology to set up a Department of Social Relations. One of its earliest products was arguably the most thoroughgoing functionalist ethnography ever published, Conrad Arensberg and Solon Kimball's *Family and Community in Ireland* (1940). Under the direction of W. Lloyd Warner (formerly Radcliffe-Brown's student) this extended functionalist anthropology from simple/primitive to complex/modern cultures. Later critics faulted the work for presenting as eternally cultural what was no more than a well-observed historical moment in county Clare.

2.1.3 Structural-functionalism (Radcliffe-Brown). Structural-functionalism's core concepts are, in harness, structure and system. Under the auspices of Radcliffe-Brown (1881–1955) it emerged as a clear-cut alternative to hybrid Boasian functionalism and successor to Malinowski's particular brand of economism and radical individualism. It derived not from German and British precursors but mainly from Durkheimian French sociology with supplementary insights from the Russian geographer and anarchist, Peter Kropotkin (1842–1921).

Structural-functionalism emphasized the formal ordering of parts and their functional interrelations as contributing to the maintenance needs of a structured social system. The function of any institution (or 'recurrent social activity') was the part it played in the maintenance of the larger structural whole. This

assumption attributed to social systems an internal integration of parts similar to that found in organisms.

Whereas Malinowskian functionalism was seen to have outworn its usefulness by 1945, and American cultural functionalism remained factionally divided, Radcliffe-Brown's successors, particularly Meyer Fortes (1906–83), Edward Evans-Pritchard (1902–73), and Max Gluckman (1911–75), expanded structural-functionalism to address situation, selection, and history while continuing to work implicitly with equilibrium models of processes internal to the socio-cultural system. Marxisant anthropological analyses of contradictions in structure and function supplemented structural-functionalism with a compatible, historically oriented, approach to competition, conflict, and change.

2.2 Functionalism as Method

Malinowski claimed that the so-called functional method in modern anthropology constituted a specific and necessary method of knowledge production. The relationship between theory and practice was reciprocal. Practice consisted of the intensive fieldwork that had emerged in contradistinction to the scientific survey methods of the Cambridge University Torres Straits Expedition of 1898–9, and the writing up of research findings in monographic form. The 1912 edition of *Notes and Queries*, that Malinowski carried with him into the field, contained W. H. R. Rivers, (1864–1922) delineation of 'intensive' fieldwork in which 'the worker lives for a year or more among a community of perhaps four or five hundred people and studies every detail of their life and culture; in which he comes to know every member of the community personally; in which he is not content with generalized information, but studies every feature of life and custom in concrete detail and by means of the vernacular language' (Rivers 1913, p. 7, quoted in Urry 1993, p. 28). This became functionalism's ideal mode of production.

Functionalism's methods of reproduction (knowledge as culture) included grantsmanship, correspondence, seminars, public events, unpublished and published lectures, oral histories, biographies, and the writing of monographs. Commemorations and conferences held in Cracow, London, and New York to memorialize Malinowski in 1984 (Gellner 1987) and in America to memorialize Boas in 1992 exposed a more complex functionalist heritage than that commonly recognized.

Paradigmatic adjustment was essential to the production and reproduction of functionalism. Thus, the last volume of Malinowski's Trobriand ethnography, *Coral Gardens and their Magic* (1935), documents the extent to which he was beginning to question 'the wholeness of culture,' criticize his neglect of the colonial order and of the inadequacy of the concept of culture contact or acculturation. He noted the possibility of placing his own functional analyses of land tenure in the same frame as political controversy over land rights in Ireland and India (Vincent 1991). But such paradigmatic adjustment was rare. More routinely, it was the thick, detailed ethnographies functionalism engendered that empowered the reader to re-evaluate functionalism long before the textualization of theory became emblematic of paradigmatic change. The perdurance of a romantic ethnographic tradition contributed significantly to the persistence of inherent functionalist notions in anthropology even when the theoretical adequacy of functionalism remained an issue (Gellner 1987).

2.3 Functionalism as Interest: The 'Functionalist Movement'

Functionalism was both a concept and a scientific community contesting the constitution and distribution of cultural capital. Throughout the first three decades of the twentieth century in both Britain and America evolutionism, diffusionism and functionalism represented warring trajectories and interests and the paradigmatic dominance of functionalism was by no means secure (Vincent 1986). Explanations of dominance, as George Stocking has consistently demonstrated, are to be found not in the valency of ideas but in academic power, control, and institutionalization—to which observation must be added strategies of reproduction. In anthropology departments where field materials were presented in seminars (as at the London School of Economics and Manchester) 'fieldnotes were treated as facts, as concrete things, as sources of authority, as weapons in the struggle to gain the high ground of functionalist debate' (Bond in Sanjek 1990, p. 279).

3. Methodological Issues or Problems in Functionalism

As with other -isms, functionalism has been described most clearly and simply by those with most to gain from detracting interest from it. Yet anthropological functionalism has been characterized by constant immanent critique as well as external challenge. For critics it has been something of a moveable feast. What is noted depends on whose functionalism they are addressing and at what phase in that functionalist's thinking. Broad, generic charges laid against anthropological functionalism relate to positivism and pragmatism, on the one hand, and primitivism, on the other. Specific critiques have been directed against its holism along with its assumption of interdependence of parts, synchronism, homogeneity and the notion of collective consciousness, uniformity and its transgenerational replication through custom. Re-evaluations of structural-functional classics demonstrated that even so-called simple societies/cultures were internally

differentiated and heterogeneity gave rise to conflicting concerns among different sectors of the population. As the paradigm expanded further, new charges were made that functionalism could not accommodate the analysis of power.

4. Conclusion

The radical antipositivism of the 1960s put functionalism on trial. Functional anthropologists who declined to commit anthropology to the scrap heap argued ethnography's value as social history. Others argued its continued relevance for specific subfields such as ecological and applied anthropology.

It is not likely that future theoretical and research problems will follow the trends set by functionalism except, perhaps, in re-evaluations of problems in Foucault's adoption of functional terminology in his conceptualization of 'discourse'; in the work of anthropologists who holistically find instances of cultural resistance to what they perceive to be Gramscian style 'hegemony'; and in the relationship of anthropological functionalism to other modernisms (Webster, in Manganaro 1990).

The main thrust is likely to be the further undercutting of the two pluralized concepts integral to functionalism: 'cultures' and 'societies' (Wolf 1988, Barth 1992). This new emergent paradigm, 'New Critical Theory,' retains functionalism on its agenda, albeit in a literary and historical mode. It promises to further problematics addressed by anthropological historical realism (sometimes in the guise of political economy) and postinterpretive cultural critique. In both instances, functionalism, in anthropology, lingers on (just as evolutionism and diffusionism linger on) not just as situated knowledge ever in the making but as historical ethnography of the operations of culture.

See also: Anthropology, History of; Functionalism, History of; Functionalism in Sociology; Malinowski, Bronislaw (1884–1942); Parsons, Talcott (1902–79)

Bibliography

Barth F 1992 Towards greater naturalism in conceptualizing societies. In: Kuper A (ed.) *Conceptualizing Society*. Routledge, London and New York

Fox R G (ed.) 1991 *Recapturing Anthropology: Working in the Present*. School of American Research Press, Santa Fe, NM

Gellner E 1987 The Zeno of Cracow, or revolution at Nemi, or the Polish revenge. In: *Culture, Identity and Politics*. Cambridge University Press, Cambridge, UK

Malinowski B 1944 *A Scientific Theory of Culture, and Other Essays*. University of North Carolina Press, Chapel Hill, NC

Manganaro M (ed.) 1990 *Modernist Anthropology: From Fieldwork to Text*. Princeton University Press, Princeton, NJ

Mead M (ed.) 1937 *Cooperation and Competition among Primitive Peoples*. McGraw-Hill, New York

Mintz S W (ed.) 1985 *History, Evolution, and the Concept of Culture: Selected Papers by Alexander Lesser*. Cambridge University Press, Cambridge, UK

Radcliffe-Brown A R 1935 On the concept of function in social science. *American Anthropologist* **37**: 394–402 (reprinted 1952 in *Structure and Function in Primitive Society*. Free Press, Glencoe, IL, pp. 178–87)

Sahlins M D 1976 *Culture and Practical Reason*. University of Chicago Press, Chicago

Sanjek R (ed.) 1990 *Fieldnotes: The Makings of Anthropology*. Cornell University Press, Ithaca, NY

Stocking G W Jr. (ed.) 1991 *Colonial Situations: Essays on the Contextualization of Ethrographic Knowledge. History of Anthropology*. University of Wisconsin Press, Madison, WI, Vol. 7

Stocking G W Jr. 1992 *The Ethnographer's Magic and Other Essays in the History of Anthropology*. University of Wisconsin Press, Madison, WI

Stocking G W Jr. (ed.) 1996 *Volksgeist as Method and Ethic: Essays on Boasian Ethnography and the German Anthropological Tradition. History of Anthropology*, Vol. 8. University of Wisconsin Press, Madison, WI

Thornton R 1985 Imagine yourself set down …: Mach, Frazer, Conrad, Malinowski and the role of imagination in ethnography. *Anthropology Today* **1**(5): 7–14

Thornton R J, Skalnik P (eds.) 1993 *The Early Writings of Bronislaw Malinowski*. Cambridge University Press, Cambridge, UK

Urry J 1993 *Before Social Anthropology. Essays on the History of British Anthropology*. Harwood, Chur, Switzerland

Vincent J 1986 Functionalism revisited: An unsettled science. *Reviews in Anthropology* **13**: 331–9

Vincent J 1991 Engaging historicism. In: Fox R G (ed.) *Recapturing Anthropology: Working in the Present*. School of American Research Press, Santa Fe, NM

Wolf E 1988 Inventing society. *American Ethnologist* **15**(4): 752–61

J. Vincent

Functionalism in Sociology

Functionalism is a sociological research program of historical and systematic significance. Sociological explanations in terms of functions are teleological: social phenomena are to be explained by their effects and consequences—chiefly of a beneficial, morally desired or non-intended character—and/or by a circuit of equilibration (a homeostatic loop) which is maintaining a certain state of proper functioning. The notion of (mono-)functionally specialized subsystems of society can be seen as the most important precipitate of functionalist thinking in sociology.

1. Basic Extra-disciplinary Orientations of Sociological Functionalism

Beginning with the renaissance, the quest for a truly empirical analysis in the natural sciences has given the concept of function a crucial leverage in replacing the metaphysical notion of substance (Cassirer 1980). Substance has been recast as the functional interdependence of (minimally two) related variables. In

sociology functionalist thinking has spread with the works of Auguste Comte, Herbert Spencer, and Émile Durkheim, who drew on contemporary physiology and evolutionary biology as the disciplines that offered the most sophisticated notions of functional interdependence so far.

Of crucial significance for the development of sociological functionalism is the work of Harvard University's biochemist and physiologist Lawrence Joseph Henderson, who played the role of a mediator between physiology and sociology as regards equilibrating processes. In his famous seminar 'Pareto and the Methods of Scientific Investigation' (1932–42) Henderson exerted considerable influence on several generations of Harvard sociologists, among them Robert K. Merton and Talcott Parsons, who became leading figures in the movement of functionalist thinking in the USA. The basic ideas of a refined sociological functionalism originate within this context.

1.1 System, Interdependence, Equilibrium

A mixture of ice, soda water, and whisky in a tightly stoppered thermos bottle can serve as an example for an isolated system of different components: water, alcohol, and carbon dioxide are its main components. There are three phases: a solid phase (ice), a liquid phase, and a gaseous phase. All three components exist in the liquid and the gaseous phase in this system. Concentration, pressure, and temperature form a set of related variables. If one thrusts the stopper more deeply into the neck of the bottle the pressure will increase and there will be a movement of components from the gaseous phase to the liquid phase; there will also be a change in concentration and temperature. All variables that characterize the system stand in a relation of mutual dependency with one another and make up an equilibrated state (Henderson 1935). If the system exhibits a minimal complexity, its internal process cannot be conceptualized as a temporal succession of single cause–effect chains any more. The alternative is the simultaneous variation of interdependent, but relatively autonomous, variables. The two single most important sources of Henderson's analysis of systems are Claude Bernard's experimental physiology—which deeply impressed Durkheim as well—and Josiah Willard Gibbs' research *On the Equilibrium of Heterogeneous Substances* (Gibbs 1948), which decisively extends the applicability of the laws of thermodynamics. On the basis of Gibbs' insights, the biochemical equilibrium of a *milieu intérieur*—of the circulation of blood, for example—can now be given a description in terms of functional equations (Russett 1966). This 'homeostatic process,' as it has been called by Henderson's colleague Walter B. Cannon, provides the systematic starting point for functionalist thinking in sociology. The conceptual scheme of the social system involves the variation of heterogeneous, interdependent components that can be accounted for in a model of equilibrium—in the ideal case, in the form of a system of functional equations. The facts, that the variables of a social system are not easily quantifiable, and that the system cannot be experimentally isolated, do not detract from the value and applicability of this conceptual scheme.

1.2 The Problem of Teleology

Equilibrating processes are not random processes. They are not the result of deterministic causal mechanisms, and they do not issue from a deterministic process of natural selection, from a 'survival of the fittest.' The internal integration of an organism and its adaptation to external conditions are based on *reciprocal* relations with the environment: as the organism is adapted to the environment, the environment is adapted to the organism; there is something like the 'fitness of environment' which is based on the internalization of the environment's properties within the organism—water and carbon dioxide, for example, play a very important role for the organism in both milieus: the external and the internal (Henderson 1913).

At least *ex post factum* the world seems to be a world of interdependent functional systems and can be described as a process of 'negentropy'; a process of organization. This is the basic assumption on which the teleological character of functionalism is grounded. But commitment to this assumption is not a necessary condition to make use of equilibrium models in sociological functionalism.

2. Structural Functionalism

Structural functionalism marks the phase of dominance of functionalism in sociology, especially in the USA from the 1940s to the 1960s. Structural functionalism does not constitute a single and unified research program. There are simpler versions of functionalism than the refined functionalism already characterized above by its extradisciplinary grounding in equilibrium models. A paradigmatic version of the simpler forms of structural functionalism can be found in the work of British anthropologist A. R. Radcliffe-Brown (1952). The anthropological study of institutions in the context of a social totality can be seen as a blueprint for the procedure of functionalist research: the careful description of structural patterns, i.e., of non-random interrelations between components of a society or a social system that is at the core of this procedure.The *functioning* of these structural patterns secures the adaptation and/or integration, i.e., the persistence of the social unit in question. The structural

patterns are explained in terms of their consequences. Such a structural pattern would be, for example, the repressive law in archaic societies which Durkheim saw as indicative of a mechanical solidarity. Repressive law contributes to the integration of society. Retributive acts revive the feelings on which mechanical solidarity is based. All sociological research tends to exhibit the formal characteristics of functionalism in so far as it is pursued carefully (Davis 1959).

2.1 Simple Functional Explanation

A simple functional explanation follows the following sequence:

(a) A system S is adapted to its environment/is integrated, if the corresponding functional prerequisite P is fulfilled;

(b) If the functional prerequisite P is fulfilled, pattern A or its functional equivalents B, C … exist;

(c) Patterns B, C … do not exist in the system;

(d) System S is adapted to its environment/is integrated, because pattern A exists in the system. This type of explanation does not imply an equilibrium model or a system of variables in the sense of Gibbs/Henderson.

Two questions have to be asked here. First, how can one explain *historically* that pattern A contributes to the fulfillment of the functional prerequisite? The functional explanation has to be linked to a theory adequate to ground the claim that pattern A would not exist at all if it did not make a difference in the fulfillment of the functional prerequisite P or in the adaptation/integration of system S, respectively (Runciman 1990). Second, what is the *origin* of functional prerequisites? Either all consequences of action are functional prerequisites, or a limited set of functional prerequisites has to be justified theoretically and/or empirically. In the latter case a more refined version of functional explanation is called for, which will be taken up in the next step. In the first case an encompassing functional unity of society cannot be presupposed. Here, functionalism advances on an inductive basis. Such a middle-range approach to functionalism is presented by Robert K. Merton, who distinguishes between manifest functions on the basis of intended and recognized consequences of action on the one hand, and latent functions on the basis of non-intended and non-recognized consequences of action on the other (Merton 1963).

Since Merton, functionalism has become associated with a concept of (mono-)functional systems that are the product of an interweaving of non-intended consequences of action and have gained substantial autonomy as against the intentions of actors. The discrepancy between actors' intentions when acting in economic systems and the actual outcome of market processes as an aggregated result of these intentional actions may serve here as an illustration. As Merton demonstrates, the definition of functions on the basis of consequences of action does not imply any preference towards an existing social order. It is impossible to make the disruption of social order the center of a functional perspective. Merton's repudiation of the commonplace critique of conservatism proves him to be one of the liberal representatives of functionalism.

2.2 Refined Functional Explanation

The greatest danger for functional explanation lies in the arbitrary selection of consequences of action as defining a function—this holds true for both simple and refined functional explanations. A limited set of functional prerequisites could be gained inductively from the empirical comparison of singular societies. The alternative is to justify a limited set of functional prerequisites on theoretical grounds. It has to be clarified, in which sense functions such as adaptation or integration make sense, not only for the organisms of biology and physiology, but for social systems as well.

In particular, the question has to be answered, how an essential change of the system can be distinguished from mere surface variations in its operations. What is missing is a clear analogon in sociological functionalism for the concept of the 'death of an organism' (Döbert 1973).

In his version of structural functionalism Talcott Parsons has provided a theoretical, normativistic account of functional prerequisites: a cultural system of values that is shared by all members of a society constitutes the central criterion for the adaptation and integration of society. The processes of society are functional in so far as they help to realize this common system of values. Parsons cannot introduce this system of values independently from an analysis of contemporary society that makes his theoretical grounding of functional problems a mere affirmation of consequences. At the same time he has to admit, that he is not able to conceive the equilibrating process of society in terms of a system of functional variables, in terms of a Gibbs/Henderson system. In particular, it does not seem possible to explain the *change* of action systems in terms of a functionalist analysis of equilibrating processes, especially processes of a moving or even a dynamic equilibrium.

That leaves Parsons with static, normative structural categories and little difference compared with the 'no frills' strategy of simple functional explanation. The concept of function serves for him as a means for the dynamic interpretation of static categories that can eventually explain no more than a static equilibrium—the boundary maintenance of society and social systems, as they already exist. Parsons sees the concept of function as a logical, but not really satisfactory, equivalent for a system of simultaneous functional equations; structural functionalism therefore is only the 'second best type of theory,' biased for a normatively defined static equilibrium (Parsons

1954, 1959). Since an existing system of values is used here as the grounding interpretation for a limited set of functional prerequisites, the charge of conservatism that has been brought against Parsons is difficult to repudiate. However, it should be clear that the reasons for these shortcomings of Parsons' theory are not political or ideological in character, but due to theoretical deficiencies. Finally, Parsons' normative version of structural functionalism has to be put into the historical context: it is a reaffirmation of the cultural identity of American society during the dramatic developments on the eve of World War II, when the identity of American society indeed seemed to be at risk (Parsons 1973).

The guiding ideal of a Gibbs/Henderson system certainly is an admonition for Parsons to avoid simple analogies and to search thoroughly for a system of interdependent functional equations. This clearly distinguishes Parsons' effort from simple forms of functional explanation, but it is not the only difference. Formally the essential difference is marked by Parsons' use of an equilibrium model (Nagel 1957, 1961). A state of equilibrium E is dependent on the variable contributions of several functions—a case of hierarchical or vertical organization of equilibrium: E is a system's overriding concern. The range of variation of these functional contributions is large enough to endanger the stability of E. If the stability of E is endangered, certain mechanisms intervene to start a process of re-equilibration. A state of equilibrium E can be brought about by several different combinations of the values of the functional variables, but there are also combinations that characterize a disequilibrium: the loss of stability. In contrast to the hierarchical, vertical organization stands the horizontal organization of equilibrium: in a horizontally organized equilibrium functional variables are on a par with each other and constitute a mutual equilibrium.

Of crucial importance for equilibrium models and for the development of functionalist thinking are the mechanisms that regulate the processes of (re)equilibration. There are three different mechanisms:

(a) *Compensatory mechanisms*. The interdependence of the functional variables is based on direct relations between them: a decreasing contribution of the functional variable A directly causes the compensating, increased contribution of the functional variable B. The limits of compensation are given in the functional variables' range of variation. Increasing the pressure in the thermos bottle illustrates a process of re-equilibration on the basis of compensatory relations between pressure, concentration, and temperature.

(b) *Mechanisms of a central control of equilibrium*. The cybernetic self-regulation of machines can provide an example for this kind of re-equilibrating mechanism. The thermostat of a refrigerator is an instance of a central re-equilibrating control.

(c) *Mechanisms of reflexive control*. In this case the system must be able to observe its own operations and to decide on re-equilibration on the basis of a 'conscious' or self-referential act of taking these observed operations into account.

3. *Systemic Functionalism*

Parsons' systemic functionalism exemplifies an equilibrium model characterized by a central control. Systemic functionalism replaces the deficient structural functionalism and develops an equivalent to the Gibbs/Henderson system. This theory claims to be able to explain the change of action systems in terms of a model of a moving or even a dynamic equilibrium. Parsons makes temporal process a constituent of his functional variables. The standard reproach directed against this type of functionalism concerns its deductive character. It may be deductive, but the theory is nevertheless empirically grounded in the empirical studies of small groups and interaction processes by Robert F. Bales, who collaborated with Parsons in developing the so-called 'four-function scheme' (adaptation, goal attainment, integration, latent pattern maintenance: AGIL). These four basic functions of social systems are derived from the analyses of processes of problem solution in small groups (Bales 1950, Parsons et al. 1981).

Parsons conceives of society as a functionally differentiated unity consisting of four subsystems: economy, polity, societal community and the cultural, fiduciary subsystem. Parsons reconstructs the interdependence of these functional subsystems in terms of communication or interchange.

Each subsystem has a specialized language of communication at its disposal—generalized media of communication or interchange: money, power, influence and moral appeals (see Fig. 1). In this scheme of

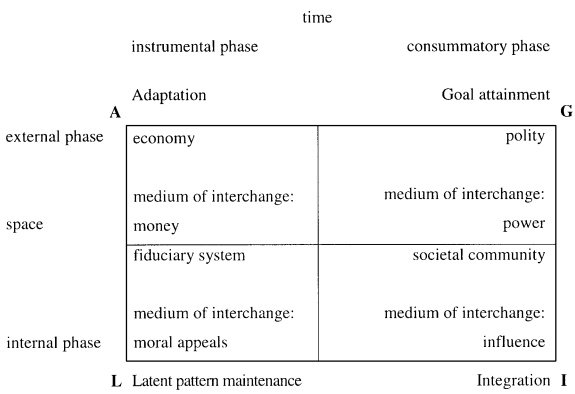

Figure 1
Parsons' four-function scheme, the functional subsystems of society, and the societal media of interchange

interchange relations between the subsystems in terms of these media, Parsons has constructed his final equivalent to a system of functional equations in the sense of Gibbs and Henderson. Moving and even dynamic equilibrium now can be explained in terms of this communicative interchange. An empirical application of this scheme can be found in the study of Parsons and Gerald Platt on the system of higher education in the USA (Parsons and Platt 1973).

All functional processes are regulated in the final instance by a *central control mechanism*: the top system of a cybernetic hierarchy which Parsons has integrated into his four-function scheme since the 1960s. This top system is culture or—on the level of societal subsystems—the fiduciary subsystem. This is a theoretical assumption that perpetuates the 'cultural determinism' of Parsonian theory.

4. *Functional Structuralism*

The theory of German sociologist Niklas Luhmann offers an outstanding example for the 'reflexive control' of the equilibrium process. Although Luhmann's theory has been referred to as a 'second encoding' of Parsons' functionalism, it is really an inversion of Parsons' structural functionalism. Therefore it may be called functional structuralism: Luhmann is not—as is Parsons—primarily interested in the function of given structures. His inverted theoretical perspective is characterized by the search for equivalent solutions of problems, i.e., for equivalent structures that fulfill given functional prerequisites. The basic assumption of his theory is that for almost all problems there are also other solutions than the ones already selected; every solution is a contingent one. This perspective brings ever-new potentialities and the continuing change of systems into view. The basic functional problem is not the boundary maintenance of systems but the functioning of systems in the face of contingency. There is no culturally defined structure, no program, no code that has to be maintained, executed, or realized to preserve the identity of the system. Luhmann shifts Parsons' problem of the boundary maintenance of systems radically into the dimension of time, shaking off all the normative connotations of Parsonian theory. This leaves only the continuity of the system over time as a problem—and as a radically purged understanding of persistence in an overly complex world of other possibilities, of other equivalent solutions (Luhmann 1970).

From the 1980s on, Luhmann began to rethink the distinction between system and environment, drawing on the works of Argentinian biologists F. J. Varela and Humberto Maturana and on the so-called 'second cybernetics' developed by researchers such as Heinrich von Foerster and Ernst von Glasersfeld. At the core of this process of rethinking stands the concept of the autopoiesis (self-creation) of systems. Even before this rearrangement of theoretical concepts, Luhmann developed the notion of a self-referential system. Self-referentiality took the strategically important place which the top system of the cybernetic hierarchy, culture, held in Parsons' theory. This was a first step to introduce a reflexive control of equilibrium process. Even more significant is the concept of autopoietic systems. Autopoietic systems repeat the difference of system and environment internally, within the system. Autopoietic systems use this difference to observe (and to regulate, to condition) their own operations. The distinction between system and environment is transformed, the difference of identity (what belongs rightfully to the system makes up its identity), and difference (what does not belong to the system, what is foreign to it) forms the basis for the system's operations.

Autopoietic systems produce the typical elements they consist of on their own by observing themselves in the process of production, by using the difference of identity and difference—in a metaphorical sense one could call this the self-consciousness of the system. Autopoietic systems are closed, and operate only under the condition of self-contact, whereas Parsons postulated open systems engaged in processes of mutual interchange. The only way for autopoietic systems to influence each other is in a manner that Luhmann calls the 'structural coupling' of systems: mutual irritations and resonance effects which the operations of one system have in the other. In a sense, autopoietic systems discover their own functionality and internalize the process of the search for equivalent solutions (Luhmann 1984). Luhmann's theory can eschew Anthony Giddens' criticism that equilibrium process, the self-regulation of systems, eludes an analysis in terms of *reflexivity* (Giddens 1977).

For Luhmann, social systems consist of communications as elements. The differentiation of society into (mono-)functionally specialized subsystems—the economic system, the political system, science, art, religion, love, etc.—is based on the development of generalized media of communication. These specialized types of communications form the elements which autopoietic systems have to reproduce endlessly. The economy, for example, is made up of money communications: payments.

It is observing itself in terms of the difference payment/non-payment (or liquid/broke). The maintenance of liquidity is the single most important functional prerequisite which has to be fulfilled to secure the continuity of the economic system's operations (Luhmann 1988).

5. *Neofunctionalism*

Neofunctionalism is a research program that has been founded on the basis of a thorough critique of Parsonian theory, in particular by Jeffrey C. Alexander (Alexander 1985, 1998). Although the name suggests a

close relationship with other forms of functionalism, the central criterion for belonging to the class of functionalist approaches is clearly not satisfied: neofunctionalism does not provide simple or refined functional explanations in terms of the teleological interpretation of consequences of action or in terms of equilibrium models. The only reason that the label 'neofunctionalism' is justified at all can be seen in its renewal of Davis' assertion, that all careful sociological analysis is (neo)functional analysis. As a movement of theory and empirical research, neofunctionalism is a reorientation of sociological analysis from systems and equilibrium to concrete actors and interpretative processes.

See also: Control: Social; Differentiation: Social; Integration: Social; Labor, Division of; Luhmann, Niklas (1927–98); Parsons, Talcott (1902–79); System: Social; Values, Sociology of

Bibliography

Alexander J C (ed.) 1985 *Neofunctionalism*. Sage Publications. Beverly Hills, CA

Alexander J C 1998 *Neofunctionalism and after*. Blackwell, Oxford, UK

Bales R F 1950 *Interaction Process Analysis: A Method for the Study of Small Groups* Addison-Wesley, Cambridge, MA

Cassirer E 1980 *Substanzbegriff und Funktionsbegriff. Untersuchungen über die Grundfragen der Erkenntniskritik*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany

Davis K 1959 The myth of functional analysis as a special method in sociology and anthropology. *American Sociological Review* **24**: 757–72

Döbert R 1973 *Systemtheorie und die Entwicklung religiöser Deutungssysteme. Zur Logik des sozialwissenschaftlichen Funktionalismus*. Suhrkamp, Frankfurt, Main, Germany

Gibbs J W 1948 *The Collected Works of J Willard Gibbs*. Yale University Press, New Haven, CT

Giddens A 1977 *Studies in Social and Political Theory*. Hutchinson, London

Henderson L J 1913 *The Fitness of Environment*. Macmillan, New York

Henderson L J 1935 *Pareto's General Sociology. A Physiologist's Interpretation*. Harvard University Press, Cambridge, MA

Luhmann N 1970 *Soziologische Aufklärung 1. Aufsätze zur Theorie der Gesellschaft*. Westdeutscher Verlag, Opladen, Germany

Luhmann N 1984 *Soziale Systeme. Grundriß einer allgemeinen Theorie*. Suhrkamp, Frankfurt, Germany

Luhmann N 1988 *Die Wirtschaft der Gesellschaft*. Suhrkamp, Frankfurt, Germany

Merton R K 1963 *Social Theory and Social Structure*, rev. edn. The Free Press, Glencoe, IL

Nagel E 1957 *Logic Without Metaphysics, and Other Essays in the Philosophy of Science*. The Free Press, Glencoe, IL

Nagel E 1961 *The Structure of Science: Problems in the Logic of Scientific Explanation*. Harcourt, New York

Parsons T 1954 *Essays in Sociological Theory*, rev. edn. The Free Press, New York

Parsons T 1959 *The Social System*. The Free Press, Glencoe, IL

Parsons T 1993 *Talcott Parsons on National Socialism*. Aldine De Gruyter, New York

Parsons T, Bales R F, Shils E A 1981 *Working Papers in the Theory of Action*. Greenwood Press, Westport, CT

Parsons T, Platt G M 1973 The American University, Harvard University Press, Cambridge, MA

Radcliffe-Brown A R 1952 *Structure and Function in Primitive Society*. Cohen & West, London

Runciman W G 1990 *A Treatise on Social Theory. Vol. I: The Methodology of Social Theory*. Cambridge University Press, Cambridge

Russett C E 1966 *The Concept of Equilibrium in American Social Thought*. Yale University Press, New Haven, CT

Sztompka P 1974 *System and Function. Toward a Theory of Society*. Academic Press, New York

Wenzel H 1991 *Die Ordnung des Handelns. Talcott Parsons' Theorie des allgemeinen Handlungssystems*. Suhrkamp, Frankfurt am Main, Germany

H. Wenzel

Fundamental Rights and Constitutional Guarantees

'Fundamental rights' are typically the protections of individual (and sometimes group) liberties in modern constitutions. International human rights law provides a baseline of minimum protections that all members of the international community ought to observe. Many constitutional orders supplement these minimum protections with additional guarantees.

1. Types of Fundamental Rights

Fundamental rights developed in three stages. Initially governments committed themselves to protect a set of basic civil and political rights, and later began to treat social welfare rights as fundamental. In the later part of the twentieth century, rights to cultural and environmental protection came to be understood as fundamental as well.

The idea that people have constitutional rights makes sense only in the context of accounting for why some exercises of power are arbitrary and unjustified. Postmedieval Western political theory abandoned accounts in which sovereign power was by definition unlimited, and replaced them with accounts of sovereign power—usually the power of governments, but sometimes the power of individuals over others—according to which there were limits on the justified exercise of power. Initially, the accounts defined a set of rights, which came to be described as 'civil' rights, that people had simply by virtue of the fact that they lived in an organized society (rather than in a state of nature). At the core of this set were the right to own property, the right to dispose of one's property

according to one's personal preferences, and the right to enter into legally enforceable contracts. Protecting these core rights were collateral procedural rights, usually associated with judicial proceedings, to ensure that one could effectively protect property rights and enforce contracts.

Successful popular and elite struggles against arbitrary royal governments led to the creation of institutions to limit government (e.g., parliaments), and to the development of another set of rights. These rights, denominated 'political' rights, arose because people lived in particular societies, each with its own political form of regulating arbitrary power. Civil rights were universal, held by all people outside the state of nature; political rights were held by the people of particular political society, and might vary from one society to the next. As parliamentary, and more generally popular, government spread, the core political right came to be the right to vote. Some rights that became widespread in this first period checked arbitrary power by protecting the institutions of civil society; examples are the press and religious institutions (other than an established church).

Accompanying the development of civil and political rights was the expansion of their distribution among the population. Conceptually, civil rights inured to people simply because they were members of an organized society, and political rights did so because they were members of a particular political order. As the idea that people were entitled to civil and political rights took hold, the denial of those rights to certain groups—women, slaves, men with little or no property—became increasingly difficult to defend. Social movements pressed for the extension of civil and political rights to those groups. Slavery was abolished in western societies by the middle of the nineteenth century. Anglo–American law recognized the rights of married women to hold property and to enter into contracts in their own names through a series of statutes enacted during the same century. Most generally, the franchise was expanded gradually, beginning with the elimination of property-based qualifications for male voting in the United States in the 1830s and 1840s, extending to universal male suffrage from the 1860s through the early years of the twentieth century in the United States and western Europe, and culminating in the elimination of gender-based qualifications for voting from the 1920s through the late 1940s in those regions.

The idea of a generalized right to equal treatment became increasingly credible as civil and political rights were extended gradually. Equal protection of the law was implicit in the theory of civil and political rights, but the manifest denial of equality made it difficult to conceptualize equality as a fundamental right. As barriers to equal protection of the law crumbled, equality as such became a fundamental right. In the contemporary United States and in discussions elsewhere influenced by the US experience, the term 'civil rights' now refers not to a category of rights, as it did earlier, but to the general right to equal treatment. Controversies over civil rights concern whether rules denying some benefit to specific groups violate their right to equal treatment, not whether access to the benefit is itself a right.

A second generation of fundamental rights, dealing with social welfare, began to emerge in the late-nineteenth century. The rise of socialist parties in western Europe provoked a response from conservatives, such as Germany's Chancellor Otto von Bismarck, and from the Roman Catholic Church. The socialist parties challenged the capitalist order for its failure to secure conditions of material well-being, such as minimal guarantees of shelter, food, and employment. According to the conservative perspective and Catholic social teaching, providing these second generation rights is a basic task of modern governments. The western commitment to social welfare rights deepened during World War II; in the War's immediate aftermath, socialist parties again played a large role in the constitutional restructuring of western Europe. The commitment was extended geographically as well, as former colonies gained independence and incorporated second generation social welfare rights in their constitutions.

The socialist arguments for social welfare rights rested to some extent on the fact that first generation civil and political rights were already viewed as fundamental. They were purely formal protections, because they guaranteed a right to hold property without guaranteeing that a person would actually hold property. In more modern terms, they protected opportunities without ensuring outcomes. Critics argued that first generation rights were empty guarantees. As they saw it, the right to own property was meaningful only to people who actually owned it. More generally, people could really have and exercise certain first generation rights only if they had second generation rights as well.

Social welfare rights were connected to first generation rights in another way. People who lacked guarantees of material well-being were likely to be dependent on others for their lives; they might, therefore, be under the command of those on whom they depended. They might be so consumed by their efforts to reach minimal levels of material well-being that they could not exercise rights to vote or participate in politics. Proponents of second generation rights argued that the fair exercise of formally guaranteed rights could occur only when people had social welfare rights as well.

First generation civil and political rights and second generation social welfare rights generally had a fundamentally individualistic character: They were rights that each person held simply because of his or her status as a human being or as an individual member of a particular society. A third generation of rights became important in the late 1960s and thereafter,

although there were hints of this new category of rights earlier. This group includes rights of national minorities to sustain their own cultures and languages within states dominated by other national groups and, relatedly, rights of aboriginal peoples to obtain support for their cultures and some degree of reparation for their displacement by settler and immigrant communities (see *Aboriginal Rights*). Rights to the protection of the environment may also be included in enumerations of third generation rights.

Although most first generation rights were individual, some had a collective orientation. Rights to religious liberty are individual only under a particular theological view of the relation between individuals and their god(s); they are collective rights when held by a group organized, for example, as a church. Third generation rights expanded the range of rights that were importantly collective in nature. A person can have a right to maintain the language of his or her origin only if, and to the extent that, other people similarly have and exercise that right. Environmental protection benefits society as a unit, rather than any particular member of society.

2. *Permissible Limitations of Fundamental Rights*

No legal system gives absolute protection to every constitutional guarantee. Doing so would sometimes substantially impair a government's ability to pursue social goals that most people regard as worthwhile. Modern formulations of the power of governments to restrict fundamental rights often resemble the one used in Canada's Charter of Rights and Freedoms (1982), which authorizes restrictions on the rights it protects when such restrictions are 'reasonable limits … [that] can be demonstrably justified in a free and democratic society.'

Historically, governments asserted the power to limit constitutional guarantees to advance the public interest in such matters as health, safety, and (more controversially) morality. These were called the 'police power' limits on constitutional guarantees. Frequently, governments found it relatively easy to assert reasonable police power interests to support restrictions on fundamental rights. This undermined the status of those rights as truly fundamental.

Constitutional systems, often through their constitutional courts, now restrict the occasions on which ordinary police power interests can justify restricting fundamental rights. In societies characterized by a wide range of views about what constitutes morality, it is now difficult to justify restrictions by reference to morality: the claim that some restriction advances a specific moral goal is met with the response that, in a morally pluralistic society, public policy ought not seek to advance that goal. Most systems require that restrictions on fundamental rights be 'proportional' to the harms the restrictions are designed to reduce. Many constitutional systems require that when efforts to promote uncontroversial goals such as health and safety impair fundamental rights, those efforts must be justified by showing, for example, that they infringe the fundamental rights no more than is absolutely necessary to advance the police power goal, and that they will be effective in doing so. Terms such as 'strict scrutiny' or 'substantial showing' are common in describing the limits on a government's ability to invoke its police powers to restrict a fundamental right.

Sometimes, constitutional guarantees conflict with each other. A defendant's right to a fair trial may conflict with the press's free expression right to publish information about a crime and with the public's right to know about the conduct of public proceedings such as trials. A person's rights of free expression, when they take the form of hate speech, may conflict with the ability of subordinated groups to achieve true equality.

Constitutional systems address these clashes in two ways. They may give one right substantial priority over others, allowing infringements on the preferred right only when such infringements appear to be the only way to protect the other rights. That is the solution in US constitutional law to the conflict between free speech and fair trial rights: the US Supreme Court allows restrictions on dissemination of information about criminal proceedings only when no other methods of protecting the defendant's right to a fair trial seem likely to succeed. However, the court must provide reasons for choosing the right that has the preferred status, and must explain the fact that such choices may produce intrusions on the subordinated rights that the alternative approach could mitigate.

The other solution is to balance the conflicting constitutional guarantees. Under this approach, a statute prohibiting some forms of hate speech might be acceptable if decision-makers reasonably conclude that the restrictions it imposes are outweighed by the increase in equality that it will produce. While common among constitutional orders, balancing solutions are problematic because they call upon decision-makers to compare interests that are not truly commensurable.

The emergence of third generation rights poses particular difficulties in the theory of rights, because those rights sometimes conflict quite dramatically with first and second generation rights. Some national minorities, for example, may have internal norms that subordinate women in ways incompatible with the exercise by their female members of the full set of first and second-generation rights. A right to individual mobility may undermine the ability of a linguistic minority to sustain its language, as individual members of the minority move away, or simply choose to conduct their affairs in the majority's language. Constitutional orders have found it difficult to address

these sorts of conflicts. The balancing approach seems even more inapt here than it does when there are conflicts between political rights, because the rights at stake so obviously are different. Further, the evolutionary development of rights means that first and second-generation rights have a stronger theoretical basis than third-generation rights, and commitments to first and second-generation rights usually are held more deeply, if only because more familiar. When third-generation rights conflict with others, then, the third-generation rights are likely to yield, making them legally and practically less fundamental and secure than other rights.

3. *Methods of Protecting Fundamental Rights*

Some fundamental rights are identified in written documents, such as constitutions, statutes, or international human rights treaties. Particularly in Anglo–American legal systems, others derive from the unwritten law. It is conventional to regard judicial enforcement as the necessary concomitant of fundamental rights. The means of enforcing fundamental rights are more varied, however. Notably, the institutional mechanisms for enforcing international human rights consist of quite weak judicial institutions and significantly more potent moral and political ones.

Most of the world's nations have signed treaties committing themselves to enforcing international human rights defined as minimum standards universally applicable. Some, but not all, of these treaties also require the nations to submit to adjudication by international human rights courts. Not all states have submitted themselves to these courts, however. An extremely limited number of human rights violations, such as genocide, can be enforced against wrong-doers in *any* court in the world willing to exercise its power, and the 1990s saw the creation of international criminal courts with authority to convict upon finding that a defendant had committed criminal violations of core human rights. Otherwise, the power of international tribunals to enforce international human rights is rather limited.

Fundamental rights can be effectively guaranteed even without judicial enforcement. International human rights, for example, are enforced in part by moral pressure generated by adverse publicity produced by some organs of the international community and by nongovernmental human rights organizations (NGOs), an increasingly potent participant in the international arena. The mere presence in a constitution of a fundamental guarantee may place political pressure on governments to honor the guarantee. Legislatures and chief executives can create committees or administrative units charged with reviewing legislative and administrative proposals to determine whether they are consistent with constitutional rights.

The question of enforcement is particularly pressing in connection with second generation rights. Fully guaranteeing social welfare rights would be extremely expensive in many countries, often far beyond the nation's fiscal capacity. Some constitutions address this problem by describing the guarantees of second-generation rights as 'directive principles of social policy,' a phrase first used in the Irish Constitution of 1937 and then adopted by the postwar Constitution of India. As constitutional 'directives,' second-generation rights are committed to the care of the legislature and sometimes, as in the Irish Constitution, specifically excluded from judicial enforcement. Other constitutional systems, such as the Italian and Japanese, treat second-generation rights as judicially enforceable, but allow legislatures a wide range of discretion in choosing how to go about actually providing those rights.

Modern constitutional systems typically provide some mechanisms for the judicial enforcement of first generation rights. Courts may develop the unwritten (or 'common') law or interpret statutes in ways expressing sensitivity to concerns about fundamental rights, even if they acknowledge that the legislature has the power to alter the law however it chooses (as is the case, at least in theory, in many parliamentary systems). They may find that when executive officials violate fundamental rights, they have acted outside the authority conferred on them by statute, a doctrine known as *ultra vires*.

Sometimes, the mechanisms for judicial enforcement of constitutional rights are so limited in scope or so cumbersome that individuals may be unable to raise claims that their personal constitutional rights have been violated in specific cases. In France, for example, individuals cannot challenge statutes once they have been promulgated; only legislators and a few other high officials have the power to invoke constitutional review, and then only before a statute has become binding law.

Proponents of judicial enforcement of fundamental rights often overlook a serious problem with judicial enforcement. Even when all agree that there is a fundamental right, described in general terms, there may be reasonable disagreement over the precise content of that right and, in particular, over whether a specific statute in fact violates the right. Some legal doctrines take account of the existence of reasonable disagreement of this sort. The European Court of Human Rights, for example, gives states a 'margin of appreciation,' allowing them to adopt interpretations of the agreed-upon rights that other states reject and that the judges themselves might reject if they had the power to specify the content of the right themselves.

Judicial enforcement means that the views of judges concerning the precise content of an agreed-upon right prevail over the views of legislators or of executive enforcement officials that no violation of fundamental rights has occurred. Where disagreement regarding

the precise content of a fundamental right is reasonable, the case for judicial enforcement is weaker than its proponents believe.

See also: Aboriginal Rights; Bill of Rights; Civil Liberties and Human Rights; Civil Rights; Civil Rights Movement, The; Constitutional Courts; Discrimination; Equality and Inequality: Legal Aspects; Human Rights, History of; Human Rights: Political Aspects; Rights: Legal Aspects

Bibliography

Alston P (ed.) 1999 *Promoting Human Rights through Bills of Rights: Comparative Perspectives*. Clarendon Press, New York

Henkin L, Neuman G L, Orentlicher D F, Leebron D W 1999 *Human Rights*. Foundation Press, New York

Jackson V C, Tushnet M 1999 *Comparative Constitutional Law*. Foundation Press, New York

Kull A 1992 *The Color-blind Constitution*. Harvard University Press, Cambridge, MA

Tushnet M 1999 *Taking the Constitution Away From the Courts*. Princeton University Press, Princeton, NJ

M. Tushnet

Fundamentalism (New Christian Right)

'New Christian Right' (NCR) is the label given to a cluster of US social movement organizations and political pressure groups formed in the late 1970s and early 1980s, of which the best known was the Moral Majority, headed by Baptist fundamentalist pastor and televangelist Jerry Falwell, of Lynchburg, VA.

Unlike such ethnic and nationalist movements as right-wing Afrikanerdom and Ulster Unionism, where religion had long played a part of identifying a people and legitimating their political struggles, the political involvement of American fundamentalism has always focused on sociomoral issues and that engagement has been sporadic. Fundamentalists have alternated between periods of pietistic retreat from the world and attempts to re-assert their values on what was seen as an increasingly secular and immoral public sphere. In the 1920s and 1930s fundamentalists campaigned for temperance and against Darwinian evolution. In the 1950s they were involved in anticommunist campaigns. They were relatively quiet during the 1960s and early 1970s.

The mobilization of the NCR in 1979 was initiated by a group of right-wingers who had previously supported third-party candidates against the liberal Republican establishment. It had two novel features. It made extensive use of direct mailing (using the lists of those who had contributed to fundamentalist radio and television programs) to raise funds, to spread information, and to create mass pressure from very small acts of commitment. For example, potentially supporters in a state with a liberal senator would be sent a pre-printed and pre-addressed card protesting about some piece of legislation and threatening to vote against the senator should he or she not vote the right way. All the recipient had to do was sign the card and mail it. The NCR also, for the first time, used white fundamentalist churches in a way that the Civil Rights movement has used black churches: to provide a network of contacts and venues for politicization. Where the law permitted a voter-registration booth would be set up in a fundamentalist church and the pastor asked to preach on the Christian's duty to vote.

Although Falwell and other NCR leaders were avowedly nonpartisan, they quickly won more friends among Republicans than Democrats and during the 1980s, the movement evolved into a pressure group within the GOP. In 1987, Marion G. 'Pat' Robertson, a religious broadcaster, Pentecostal minister, and founder of the Christian Broadcasting Network (later the Family Channel) ran for the Republican nomination for the presidency.

Specifically the NCR was opposed to an Equal Rights for Women Amendment to the constitution, gay rights, abortion, divorce, public displays of sexuality and promiscuity, and to the teaching of evolution and 'secular humanism' in schools. It was in favor of public prayer in schools and the teaching of Creationism (which it tried to recast as 'Creation Science'). More generally, it wanted public re-affirmation of the superiority of conservative Protestant culture and identity. Although this was often denied there was also a large amount of racism in the appeal of the NCR. Criticism of liberal welfare state and immigrant policies provided a respectable language for a 'nativist' reaction to Blacks and Hispanics. Support for tax tuition credits for independent schools, although presented as pro-parental choice, was also a way of opposing racial integration.

The NCR had considerable early success in gaining publicity for its cause. Falwell, Robertson, and other religious broadcasters were masters of the mass media. With the technical assistance and experience of right-wing lobbyists, they were able to create an appearance of power far beyond the reality. In this they were aided initially by the credulity of naive commentators who over-reacted to the reappearance of a large constituency which the liberal and cosmopolitan centers of the US had forgotten. Thus, the NCR's claims to have been responsible for the ousting of four leading liberal Democrat senators in 1980 was reported without the qualification that the swing against those four was the same as the swing against the Democrats in states where the NCR had not campaigned actively. The NCR also claimed the credit for the 1980 election of Ronald Reagan to the presidency, even though

Reagan's success owed far more to the appeal of his economic liberalism and his aggressive foreign policy than to his conservative rhetoric on socio-moral issues.

1. *The NCR's Record*

Despite the 12 years of Reagan and then George Bush in the White House, the NCR conspicuously failed to make much progress on its agenda. In areas where fundamentalists were strong, it won a number of court cases and managed to have state legislatures pass such measures as bills requiring that teaching of evolution be balanced by equal time for creationism, but such gains were always reversed once the issue had been appealed upwards. Appeal courts decided that Creation Science was religiously-inspired rather than evidence-driven science and hence that its teaching in public schools was unconstitutional. Attempts to have any form of instruction that did not specifically promote Christianity regarded as expressions of the 'religion' of secular humanism were rejected repeatedly by the courts. A landmark judgment ruled that teaching children about beliefs and values they did not share was not in itself a form of unconstitutional propaganda. On its most popular issue, opposition to abortion, the NCR made very little headway. Although the Supreme Court was willing to permit very slight constraints (for example, on public funding of information about abortion), it turned down opportunities to outlaw it repeatedly.

The NCR's electoral record was similarly unimpressive in the light of the unprecedented sums that were spent. Being able to deploy groups of highly committed activists allowed NCR groups to win power and promote candidates in some branches of the GOP but such advances were usually reversed by mainstream Republicans. When they were not, movement candidates usually were rejected by the electorate. In particular, the NCR alienated middle-class, women voters who saw the entire conservative agenda as an attempt to reverse their gains in the labor market and in public life.

Where the NCR had some success was in presenting itself as a cultural minority the rights of which had been infringed. It was thus able to reverse a number of overly zealous decisions of the separation of church and state. For example, a Christian student group won back funding that had been denied it while the University funded a variety of equally contentious political and cultural groups.

2. *Explaining the NCR's Failure*

Idiosyncratic causes of the NCR's failure can be identified. It certainly did Robertson's campaign no good that it coincided with two major television evangelists being exposed as frauds, perverts, and hypocrites. However the major roots of failure lie in the dynamics of motivation and in US social structure.

Despite the title of Falwell's organization, NCR supporters were never a majority. They could only achieve their goals if they formed alliances with other conservative groups but this created motivational problems because it required alternating between two sets of values. Fundamentalists were encouraged to become politically active by various threats and insults to their socio-moral values, which were derived from their religious beliefs. But to be effective politically, they had to set aside their religious preferences and cooperate with conservative Catholics, Jews, and atheists. While the NCR leadership could separate religion and politics, many of the rank-and-file could not. They either became disillusioned at the compromises involved in pressure group politics or refused to make such compromises and by their anti-Catholicism, anti-Semitism, and racism offended, respectively, Catholics, Jews, Blacks, and Hispanics, groups which had considerable sympathy for aspects of the NCR agenda.

Potential for frustration was inadvertently built into the movement from the start by its flawed explanation of what was wrong with America. Because it saw all that offended it as the deliberate work of militant 'secular humanists,' it supposed that all that was required to 'bring America back to God' was an equal counterforce of militant socio-moral conservatives. The NCR failed to appreciate that many of the changes that concerned it were the accidental by-product of two social forces that, in general, its members supported: prosperity and individual autonomy.

The one sense in which liberals were implicated directly in the failure of the NCR was in successful counter-attack. Once leading scientists realized that they could not take their cultural dominance for granted but had to persuade people, they did so extremely effectively. Organizations such as People for the American Way mobilized very effectively to produce a compelling defence of diversity and toleration. But the liberal counterattack worked because it rested on values that many fundamentalists themselves accepted. Unlike, for example, Islamic fundamentalists, NCR sympathisers were not opposed thoroughly to the division of the public and private spheres or to the principle of toleration. Hence, they were often ambivalent about their own legislative program; while polls repeatedly showed most Americans opposed to abortion, more sensitive questioning showed considerable willingness to make exceptions and some reluctance to impose morality on others. In addition, most voters put socio-moral concerns relatively low down their list of priorities and attempts by NCR-supported candidates to stress such matters reinforced the impression that they were single-issue candidates who would not be effective politicians. The result was that the NCR failed to turn the very large constituency of born-again Christians into committed supporters.

What was clear from opinion polls and voting patterns in the southern states in Pat Robertson's election campaign is that even many who shared Robertson's religion were unenthusiastic about supporting a conservative and fundamentalist pastor for the presidency.

While it is easy to assess the NCR's record on specific agenda items and conclude that it failed, it is less easy to assess its more general and long-term effects. Given that the proportion of women in full-time employment, divorce rates, and abortion rates have continued to rise, we may conclude that the movement failed in its more general objectives. The one sense it can claim to have succeeded is in making sociomoral issues extremely contentious and emotive. Although the commonly used phrase 'culture wars' is an exaggeration, the NCR has contributed to the development of an extremely acrimonious tone in American public life signaled, for example, by murderous attacks on the staff of abortion clinics.

3. *Conclusion*

Although interesting in its own right, the career of the NCR allows us to understand some basic features of modern democracies. While the more ethnically and culturally homogenous and smaller states of Europe tend to have centralized and closed systems of political power and public administration, the size and diversity of the US has resulted in an open, federal structure that allows sizable minorities considerable opportunity to initiate debate and promote their agendas, and to win some local victories. However, those issues of central importance to the operation of the state are dealt with in a centripetal structure that prevents even well organized and highly motivated minority movements carrying the day. In electoral politics and in the formulating of legislative programs, the NCR could press its case but then suffered repeated rejection at higher levels as that case was scrutinized by politicians answerable to wider constituencies. Similarly victories in lower level courts were reversed in various levels of appeal.

What is clear from the decisions of such influential elite groups as Congressional politicians and judges, and from the voting patterns of ordinary Americans, is that there is widespread implicit understanding that the protection of individual liberty and the preservation of social harmony in a culturally diverse democracy require a division of the public from the private and that, while considerable liberty can be permitted in the private sphere, the public sphere has to remain religiously neutral.

See also: American Studies: Politics; American Studies: Religion; Christian Parties: European; Creationism, Evolutionism, and Antievolutionism; Parties/Movements: Extreme Right; Protestantism and Gender; Religious Fundamentalism: Cultural Concerns; Republican Party; Televangelism; Tolerance

Bibliography

Bruce S 1998 *Conservative Protestant Politics*. Oxford University Press, New York

Moen M C 1992 *The Transformation of the Christian Right*. University of Alabama Press, Tuscaloosa, AL

Wilcox C 1992 *God's Warriors*. Johns Hopkins University Press, Baltimore, MD

S. Bruce

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

G

Gaia

In the early 1970s, the English chemist, James Lovelock, and the American microbiologist Lynn Margulis postulated the Gaia hypothesis that life on Earth is part of a self-regulating system that binds the atmosphere to the continents and oceans in a quasi-physiological process (Lovelock and Margulis 1975). The ideas behind Gaia were developed as part of NASA's efforts to find life on other planets (Lovelock 1965). As a theory of evolution, Gaia holds that life forms have co-evolved with their environment in such a way that environmental conditions are held steady in the face of a continual influx of energy and matter. Gaia posits first that life on Earth is a planetary phenomenon, and second, that living organisms have to regulate their planetary environment otherwise the forces of physical and chemical evolution would render it uninhabitable (Lovelock 1979).

The evidence put forward for Gaia draws attention to the fact that the chemical composition of the Earth's atmosphere has stayed remarkably constant and quite unlike the composition that would be expected in the abiological steady state of a lifeless planet. The geological record suggests that neither global freezing nor boiling conditions have ever prevailed. Mean surface temperature has probably never departed from the range 5–50 °C (Lovelock and Whitfield 1982). The Gaian perspective has spawned new avenues of scientific inquiry aimed at furthering the understanding of biogeochemical cycles, especially in the atmospheric sciences. For example, the importance of methane in climate change came to be taken seriously with the recognition that biological organisms play a vital role in regulating the atmospheric environment.

Lovelock (1988) coined the term geophysiology to denote an emerging systems approach to earth science that might assist in furthering understanding of the implications of human intervention into biogeochemical cycles in a way that distinguishes situations which threaten the planetary operating system from those that do not. By using the term geophysiology, Lovelock distanced himself from teleological criticisms and the mysticism that has come to surround the notion of Gaia. For Lovelock, Gaia is a quasi-living entity that has the capacity for global homeostasis operating according to cybernetic principles, which like a thermostat does not require will or sentience.

Gaia caught the public imagination in the 1980s, along with an emerging conviction in the scientific community that the global environment was being seriously affected by human action (Joseph 1990). For environmentalists, Gaia became a symbol of the need for promoting an ecocentric view of the world around us, in contrast to the dominant anthropocentric view. For many, Gaia became also an argument that cooperation rather than competitiveness was needed to unite humanity in the face of threats to the global environment and an argument for alternative ways of living.

It is important to distinguish between the Gaia hypothesis of earth science, and Gaia, the popular symbol of environmentalists, which underpins a call for an ecological egalitarianism to end all forms of human domination of nature (Serafin 1988). A considerable literature has developed exploring Gaia both as a scientific concept and as an ethical, philosophical, and sociological proposition. Debates in the literature about Gaia have been set increasingly in the historical context of an emerging science of the biosphere. Gaia is seen as an elaboration of ideas put forward by the Englishman James Hutton who first proposed the Earth to be a living system in 1785 and the Russian Vladimir Vernadsky who realized the importance of life in transforming geology in the early part of the twentieth century.

See also: Cassandra/Cornucopian Debate; Ecology, Political; Environment and Development; Environmental and Resource Management; Environmental Economics; Nature, Concepts of: Environmental and Ecological; Sustainability Transition: Human–Environment Relationship

Bibliography

Bunyard P (ed.) 1996 *Gaia in Action: Science of the Living Earth*. Floris Books, Edinburgh, UK

Joseph L E 1990 *Gaia, the Growth of an Idea*. St. Martin's, New York

Lovelock J E 1965 A physical basis for life detection experiments. *Nature* **207**: 568

Lovelock J E 1979 *Gaia: A New look at Life on Earth*. Oxford University Press, Oxford, UK

Lovelock J E 1988 *The Ages of Gaia: A Biography of Our Living Earth*. Norton, New York

Lovelock J E, Margulis L 1975 The atmosphere as circulatory system of the biosphere—the Gaia hypothesis. *CoEvolution Quarterly* **6**: 30–41

Lovelock J E, Whitfield M 1982 Life span of the biosphere. *Nature* **296**: 561

Schneider S H, Boston P (eds.) 1992 *Scientists on Gaia*. MIT Press, Cambridge, MA
Serafin R 1988 Noosphere, Gaia and the science of the biosphere. *Environmental Ethics* **10**: 121

R. Serafin

Galton, Sir Francis (1822–1911)

Francis Galton was born in Birmingham on February 16, 1822, the third son of Samuel Tertius Galton, a wealthy banker whose fortune was mainly derived from his father's musket manufactories. Francis was also the grandson of Erasmus Darwin, in whose scientific and evolutionist footsteps he was later to follow. However, he was originally destined for a career in medicine. Aged 16, at his father's behest, Francis enrolled as a medical student at the Birmingham Free Hospital, from where he moved to King's College, London to study medicine in 1839, and to Trinity College, Cambridge in 1840 to read mathematics. Having achieved only a pass degree at Cambridge, largely due to the first of the mental breakdowns that punctuated his career, Galton continued with his medical instruction despite a growing aversion to it. Coming into a substantial fortune on the death of his father in 1844, he immediately abandoned medicine and spent several indolent years 'hunting, fishing and shooting.' Then, in 1850, he led and financed an exploratory expedition to South-West Africa (modern Namibia). On his return, in 1852, Galton was lionized by the scientific establishment. And having achieved this entrée into the scientific community, gradually he established himself as one of Victorian Britain's most respected men of science. He was knighted in 1909.

Francis Galton is now remembered chiefly for his innovations in the field of statistics and for advocating the practice of eugenics. But his interest in both of these areas initially arose from an attempt to provide scientific evidence for the inheritance of human mental attributes. In fact, Galton devoted most of his professional energies to investigating varied aspects of heredity, a predilection largely inspired by reading Charles Darwin's (his half-cousin) *The Origin of Species* in 1859. Throughout his career, Galton's views of mind and morality were unequivocally materialist and evolutionary. In 1865, he published an article in two parts, under the heading 'Hereditary talent and character,' that were later expanded into his first book on the subject of mental heredity, *Hereditary Genius* (1869). In each of these publications he sought to show that the pedigrees of geniuses contain considerably more forebears of high social distinction than the lineages of those with mediocre parentage. Galton was aware that any positive results might instead suggest the importance of nepotism and unequal access to education. But on the basis of several ingenious—albeit less than convincing—arguments he insisted that for the highly intelligent neither education nor preferment are necessary for the fulfillment of their inborn intellectual potential. This permitted him to interpret any instance of relatedness among the eminent as being due to shared heredity.

Having presented detailed pedigrees for several hundred eminent men Galton calculated that 50 percent of illustrious individuals have at least one distinguished relation. On this basis he concluded that heredity is the primary determinant of intellectual ability. He also devoted a section of *Hereditary Genius* to estimating, in terms of the overall stock of hereditary intelligence, the 'comparative worth' of different nations and races. Galton reproduced the typical mid-Victorian conception of a racial hierarchy stretching from the Caucasian to the Negroid, although he argued that the ancient Athenians were appreciably more intelligent on average than any modern civilized nation.

During the 1870s, Galton continued his researches into human mental heredity. In 1874 he issued questionnaires to leading men of mid-Victorian science, hoping to shed light on the relative importance of 'nature and nurture' in their choice of vocation. Predictably, his data interpretation was prejudiced in favor of a hereditarian explanation. In 1875, although he had no conception of the modern distinction between monozygotic and dizygotic twins, Galton turned to the phenomenon of twinning as a means of delineating between heredity and environment. He collected the life histories of pairs of like-sexed twins and sought to determine if those that were similar at birth and in childhood became unlike upon leaving their similar home environments. He also asked if like-sexed twins that were dissimilar at birth subsequently became more similar in the shared home environment. In both cases Galton concluded that character had been formed prior to birth and had thereafter shown only limited flexibility in response to environmental pressures.

On the basis of his studies of heredity, Galton proposed the peremptory regulation of human breeding so as to maximize the biological quality of the nation. In 1869 he recommended the confinement of the socially undesirable in 'celibate monasteries or sisterhoods,' and the creation of incentives for the intellectual elites to reproduce more frequently. The term 'eugenics' was coined in 1883 by Galton from the Greek *eugenes* meaning 'good in stock.' Yet, eugenics was widely considered impracticable during the 1860s and it was only during Galton's last decade that he began championing the idea. By then, foreign challenges to Britain's industrial and mercantile hegemony, and fears of an increasingly volatile urban underclass had rendered eugenics an appealing notion to many *fin-de-siècle* Britons. In 1908, after initial

reluctance occasioned by his advanced age, Galton became President of the newly formed Eugenics Education Society.

Political considerations bulked large in Galton's advocacy of eugenics. Running through his work on heredity and eugenics, and becoming more pronounced with time, was a strongly elitist conviction that the vast majority of the lower classes were sexually profligate and irremediably inferior to the higher social classes. Accordingly, he expressed the fear that the lower classes were out-breeding the 'respectable' classes. Moreover, Galton's predilection for studying intelligence and his commitment to enhancing the reproductive rate of the most intelligent are symptomatic of his upbringing as a member of the Victorian professional middle class. This was a social stratum that set great store by intellectual accomplishment and produced many of the century's foremost social reformers, commentators and scientists: Galton unconsciously celebrated the core values and accomplishments of this important social group. In advocating eugenics he was also influenced by the widely held belief in the possibility of human perfectibility—of which his grandfather Erasmus Darwin had been a leading exponent—and by the potent mid-Victorian ideologies of human and social progress advanced most forcefully by his friend and colleague Herbert Spencer. Lastly, Galton's eugenic thought reflects the growing impetus among Victorian life-scientists for making their discipline practically useful to the state.

In the context of studying the inheritance of human mental and bodily traits, Galton emerged as an innovative statistician. His maxim was 'Whenever you can, count,' and his researches suggest a man with an almost obsessive compulsion to do so. In *Hereditary Talent* and *Hereditary Genius*, Galton became the first person to employ the Gaussian error curve in classifying human *mental* attributes. Although his use of statistics was at this stage rudimentary, he accomplished the important step of measuring—and focusing almost entirely upon—statistical variation. This aspect of the normal curve had not received much attention from his contemporaries (with the exception of the Belgian astronomer and statistician Adolphe Quetelet), many of his predecessors regarding variation from the mean as 'error' rather than a natural phenomenon worthy of investigation. All of Galton's major statistical insights stemmed from this concern for statistical spread as opposed to averages alone. In particular, his work was pivotal in the development of both the correlation co-efficient and regression analysis. Galton formulated a graphical method to measure correlation in 1886 by plotting the distributions of both variables, and two years later he devised an 'index of co-relation.' In fact, his computational procedures for 'co-relation' involved measuring the slope of the regression line and producing a measure of the regression coefficient. The mathematical underpinnings of correlation were developed more fully by his admirer and first biographer, Karl Pearson. From Galton's idea of regression, Pearson also developed a statistical system for the linear production of two continuous variables.

Galton's role in the development of regression analysis is more ambiguous. As a means of studying the mathematics of hereditary variation, during the late 1870s Galton began to breed sweet peas. He found that the F_1 generation peas were, on average, only a third as extreme in any given quality as the parental generation. Thus, through his breeding experiments, Galton had stumbled upon the statistical principle of regression. However, he always assumed that the phenomenon of regression related to a fundamental physiological process, and his explanation for it was biological rather than statistical. Galton asserted that each individual is composed of distinct hereditary units derived both from parents and from earlier ancestors. In each generation a random selection of the ancestral and parental hereditary material is made, such that only some of the total complement of hereditary units are actually expressed in each offspring. Galton believed that, from a purely probabilistic point of view, the entire hereditary material—stretching back several generations—would on average be close to the population mean. However, because only a random selection of hereditary units gain expression, it is possible for more extreme characteristics to be produced. Regression occurs when, as is likely, the more mediocre hereditary units regain supremacy in the following generation. This conception of the nature of heredity led to Galton's formulation of the 'ancestral law of heredity,' a law stipulating that any two parents together contribute approximately one-half of the total hereditary material of the offspring, the four grandparents one-quarter, and so on until the proportions become vanishingly small. Galton believed this law to describe the actual hereditary composition of each individual and the model formed the basis of his statistical treatment of heredity in *Natural Inheritance* (Galton 1889). During the twentieth century, 'Galton's law' would be eclipsed by the advent of Mendelian genetics.

'Galton's law' may not have long survived the rediscovery of Mendel's work, but Galton has been credited with anticipating August Weismann's concept of the continuity of germplasm. Unlike Darwin and the majority of his contemporary life-scientists, from as early as 1865 Galton was largely dismissive of the notion of the inheritance of acquired characteristics. During the early 1870s he experimentally tested Darwin's theory of 'Pangenesis'—according to which bodily tissues throw off the germ cells from which offspring are produced—by transfusing blood among different varieties of rabbit. His predictable conclusion brought no pleasure to Darwin, and Galton's 1871 rejection of Pangenesis in *Nature* elicited from him a swift rebuke. Subsequently, Galton's resistance to the inheritance of acquired characteristics hardened. In

the models of hereditary transmission he published during the first half of the 1870s, he claimed that the 'patent' hereditary units that become bodily structures only rarely return to the reproductive organs to be transmitted to progeny. In contrast, the 'latent' elements remain in the testes and ovaries ready to be transmitted to the next generation. In this rather circumscribed sense his hereditarian model did foreshadow Weismannism.

However, aside from the many technical inconsistencies between Galton's and Weismann's concepts of heredity, it is important to note that after 1875 Galton departed significantly from the position outlined above. He gradually came to appreciate that his case for the continuity of traits between parents and offspring was weakened if it were assumed that the latter inherited only the former's latent elements and very few of their patent elements. As a result, by the late 1880s Galton was writing of patent and latent elements as being equally heritable. Whilst he remained only mildly sympathetic to the notion of the inheritance of acquired characteristics, he had abandoned his earlier emphasis upon the near-absolute separation of bodily from reproductive cells.

Galton's dual fascination with heredity and quantification also generated an interest in anthropometry, the measurement of bodily morphology. During the 1870s, appalled that so little was known about the physique of the British people and fearful that a biological decline was already in progress, Galton sought to obtain anthropometric information on children from different social classes. Subsequently, he became a member and later chairman of a British Association committee investigating occupational disparities in size and weight. Throughout his involvement with this committee Galton was convinced that heredity was the chief determinant of anthropometric variation. As a means of procuring raw data from which to study hereditary transmission in humans, in 1884 he set up an elaborate anthropometric laboratory at the International Health Exhibition in London. Equipment was installed to measure the height, weight, breathing-capacity and, *inter alia*, the strength of pull of members of the public. Also measured were such qualities as keenness of sight, hearing, and color discrimination.

These last tests reflect Galton's interest in what he dubbed 'psycho-physics,' the scientific investigation of mental phenomena. Generalizing his earlier conviction that individual nurture is a subordinate force to heredity, he emphasized aspects of unconscious cerebral function in which the 'will' seems to play little if any role. In his 1883 *Inquiries into Human Faculty*, Galton described and, where feasible, quantified a range of unconscious and sensory phenomena, including number forms, visual imagery, color associations, visions, and sensory awareness among the blind. He also documented the results of his introspective investigations into the power of the individual will. In the course of these studies, Galton devised the method of 'word-association' with which to demonstrate his claim that very few thoughts or actions are ever the spontaneous product of the will but relate to desires, ideas and associations of which we have little conscious awareness.

In addition, Galton was a pioneer in the design and employment of composite photography. He hoped that through superimposing numerous photographic portraits on top of each other he would be able to discern the common hereditary features shared by families or groups. His main work in this area involved an attempt to discover the physiognomical indices of different forms of criminality from a series of photographs supplied by the Inspector of Prisons. To his disappointment, Galton found no evidence of distinctive criminal types. However, this interest in hereditary variation, measurement and their relevance to criminal investigations also culminated in his extensive study of fingerprints. Galton was not an innovator in this field but his several publications on the subject drew attention to the enormous value of fingerprints as a means of personal identification.

Although it is clear that most of Galton's scientific work stemmed from his early fascination with heredity, his scientific curiosity was nonetheless polymathic in breadth. Over the course of half-a-century's involvement with the Royal Geographical Society, he was able to make significant contributions to the expeditions of such explorers as R. Burton, J. Speke, and J. Grant, and he worked hard to place geography on a more scientific footing. As an active meteorologist he was instrumental in the construction of weather maps and was responsible for discovering the 'anticyclone' in 1862. Finally, he was an enthusiastic advocate of establishing scientific activity as an occupation of dedicated, professional and, if necessary, salaried, researchers. Alongside T. H. Huxley, John Tyndall, Joseph Hooker, and others, he campaigned vigorously to free science from the limitations imposed by religious orthodoxy. To this end, many of Galton's works carried thinly veiled attacks on theology and the Church, most strikingly his 1872 attempt to prove statistically that prayer is inefficacious.

In retrospect, there can be little doubt that Galton gave an important impetus to the scientific study of heredity and the development of psychometric testing. Despite the shortcomings later revealed in his own work, he was the first to subject human intellectual attributes to systematic numerical analysis. With the aid of W. F. R. Weldon and Karl Pearson, he was also intimately involved with establishing a statistical approach to the study of evolution. Perhaps Galton's most lasting contributions to science arose from his statistical breakthroughs of the 1870s and 1880s. Statistical procedures derived from his initial insights are now indispensable to both the natural and the social sciences. More practically, his private wealth enabled him to set up a Eugenics Research Office at

University College London in 1904. He honored and extended this commitment upon his death (on January 17, 1911) when he created the Galton Professorship of Eugenics and named Karl Pearson as the Chair's first holder.

Any summary of Galton's legacy must recognize that eugenic ideas were implemented after his demise with horrific consequences, and that the hereditarian ideologies he promoted have often been used as means of social, sexual and racial oppression. In the wake of the Nazi atrocities of the 1930s and 1940s, and of the civil rights movements of the 1960s and 1970s, Galton's stock as an important scientific figure has declined rapidly. However, in the interests of a proper understanding of the past, it should be pointed out that to implicate Galton in the rise of twentieth-century eugenic practices is to over-personalize a complex series of events and set of national circumstances. Eugenic measures were linked to a wide range of social, economic and political factors against which Galton's advocacy is scarcely more than trivial. Furthermore, even in 1865 the notion of selective human breeding was far from novel: Galton's originality lay only in buttressing the idea with apparently rigorous scientific research.

Bibliography

Forrest D W 1974 *Francis Galton: The Life and Work of a Victorian Genius*. Paul Elek, London

Galton F 1863 *Meteorgraphica, or Methods of Mapping the Weather*. Macmillan, London and Cambridge

Galton F 1869 *Hereditary Genius*. Macmillan, London

Galton F 1874 *English Men of Science: Their Nature and Nurture*. Macmillan, London

Galton F 1883 *Inquiries into Human Faculty and its Development*. Macmillan, London

Galton F 1889 *Natural Inheritance*. Macmillan, London

Galton F 1908 *Memories of My Life*. 2nd edn. Methuen, London

Kevles D J 1986 *In the Name of Eugenics: Genetics and the Uses of Human Heredity*. Penguin, Harmondsworth, UK

Keynes M (ed.) 1993 *Sir Francis Galton, FRS: The Legacy of His Ideas*. Macmillan, Basingstoke and London

Pearson K 1914/1930 *The Life, Letters and Labours of Francis Galton*. Cambridge University Press, Cambridge, UK

Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press, Cambridge, MA

J. C. W. Waller

Game Theory

1. Introduction

Game theory is the mathematical study of situations of conflict of interest. It is applicable to military situations, but has also been applied to economics and other social sciences, and even to biology.

We think of game theory as the fourth in a sequence of paradigms, explaining how people attempt to achieve goals; call these type I, II, III, and IV thinking: in type I thinking, the question is whether a goal can be met. In type II thinking, the decision maker, satisfied that the goal can be met, tries to meet this goal in the most efficient way possible (at minimal cost); alternatively, having decided that the goal cannot be met, he tries to come as close as possible to meeting the goal. In type III thinking, the decision maker realizes he is faced, not by nature, but by other decision makers whom he can influence. He therefore tries to influence them so that they (acting according to type II thinking) will carry out actions helpful to him. In type IV thinking, the decision maker realizes that he is faced by other individuals whose actions can affect him, and who in turn are trying to influence his actions. Thus he will generally try either to outsmart the others, or to cooperate with some or all of them. It is this last type which represents game theory.

A game is any situation in which two or more agents (players), not having identical interests, can affect the outcome. We assume that all players seek to maximize their individual utility. There may or may not be random moves.

Games can be noncooperative or cooperative. In the noncooperative case, we are mainly interested in the strategies chosen by each player. In the cooperative case, we are also interested in the process of bargaining and coalition formation. Games may also have complete or incomplete information. In case of incomplete information, there is interest in the process of learning, as players elicit information during the play of a game.

Essentially, given a game situation, game theory tries to answer the following questions:

(a) What strategies should each player use?

(b) Where there are gains to be made from cooperation, how should the players bargain?

(c) Where there are many players, what coalitions should form?

(d) What information can be deduced from other players' actions?

2. Extensive and Normal Forms

There are several ways of representing a game. The simplest perhaps is to draw a graph, with nodes representing the several positions in the games, and arcs representing moves. This is the extensive form of the game. For example, consider the simple game of matching pennies:

Example 1. Player I puts a coin on the table, with either heads (H) or tails (T) up. Player II, not seeing the coin, calls either H or T. If II has called correctly, she wins $1 from I. If the call is incorrect, I wins $1 from II.

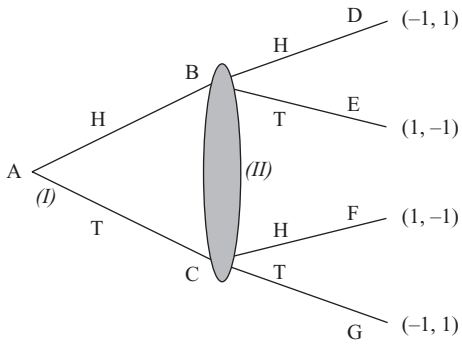

Figure 1
The game of Matching Pennies

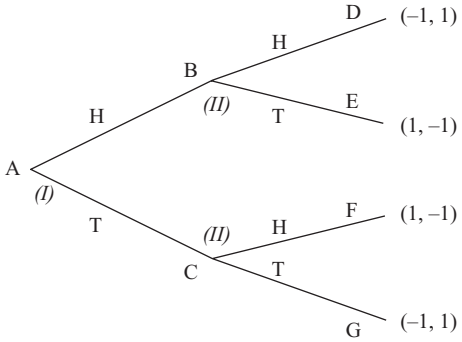

Figure 2
The game of Matching Pennies, modified

This game is represented by the graph in Fig. 1. Play starts at node A, which 'belongs' to player I (it is his turn to play there). Depending on I's choice of H or T, play then moves to either node B or node C, both of which belong to II. After II makes her choice, play terminates at one of the nodes D, E, F, or G, all of which are terminal nodes of the game. Numbers are assigned to the terminal nodes: they are payoffs for the game. Thus, at D, the payoff (-1, 1) means that, since II guessed H correctly, I must pay a dollar to II. Similarly at the other terminal nodes.

Note the connection (shade) between nodes B and C. This is meant to represent II's lack of information when she moves here: she knows the position is either B or C, but cannot distinguish between them. Thus B and C are said to belong to the same information set.

Example 2. Suppose, in the previous example, II were allowed to see how I had placed the coin. Otherwise, assume the game is unchanged. In this case, we obtain the graph shown in Fig. 2. The difference between the two figures is that nodes B and C are now in different information sets. II, when moving here, can distinguish between these two positions.

Alternatively, it is frequently more convenient to represent a game by means of its strategies. A strategy is a rule that tells a player what to do in each conceivable situation, i.e., at each of his information sets. Thus, in example 1, each player has two strategies, namely H and T. In example 2, Player I still has only two strategies, H and T. Player II, however, has two information sets (vertices B and C) at each of which she has two alternatives. This gives her a total of four strategies, which we might call S(HH), S(HT), S(TH), and S(TT) (where HT means 'call H in position B, T in position C,' etc.).

For a given choice of strategies (one strategy for each player), the outcome of the game is determined except for the effect of random moves. Assuming that the probability distributions at chance moves are well defined, and known to the players, it is quite straightforward to obtain expected payoffs for each choice of strategies. The function which assigns to each n-tuple of strategies these expected payoffs is known as the normal form of the game.

For the game of example 1, above, the payoff is given in the form of a table or, more precisely, a 2×2 matrix, in which the rows correspond to Player I's choice of strategy, and the columns to II's choice. The entries are pairs of numbers, representing the payoffs:

	H	T
H	$(-1, 1)$	$(1, -1)$
T	$(1, -1)$	$(-1, 1)$

For the game of example 2, the normal form is given as a 2×4 matrix, much as in example 1:

	S(HH)	S(HT)	S(TH)	S(TT)
H	$(-1, 1)$	$(-1, 1)$	$(-1, 1)$	$(-1, 1)$
T	$(1, -1$	$(1, -1$	$(1, -1$	$(1, -1$

3. *Two-person Zero-sum Games*

3.1 *The Game Matrix*

In the simplest of cases, a game situation will have only two players, with directly opposed interests. This can be thought of as a closed system in which, whatever one player wins, the other loses. Thus the sum of the two players' utilities will be zero at all times. It follows that it is not necessary to state explicitly Player II's payoff: this is always the negative of that to Player I. The outcome of such a game can be represented, then, by a single number: I's payoff. The normal form of such a game becomes a matrix in which, as before, rows represent I's strategies, and columns, II's stra-

tegies. The entries are payoffs to I, which I wishes to maximize, and II, to minimize. Thus, for the game of example 1, the normal form becomes the matrix:

$$\begin{matrix} -1 & 1 \\ 1 & -1 \end{matrix}$$

and, for example 2, it becomes:

$$\begin{matrix} -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{matrix}$$

3.2 Equilibria and Saddle Points

In a desire to 'solve' games, we look for equilibria, i.e., choices of strategies such that no one of the players can gain by a unilateral change of strategy. In a zero-sum matrix game, a pair of strategies will be in equilibrium if the payoff they define is simultaneously the largest in its column and smallest in its row. This is frequently called a *saddle point*, by analogy with the surface of a (riding) saddle.

Example 3. In the matrix game:

$$\begin{matrix} 6 & 1 & 2 \\ -4 & 0 & 9 \\ 8 & -1 & -4 \end{matrix}$$

the entry in the first row, second column is a saddle point. Similarly, in example 2, above, both entries in the second column are saddle points.

When a matrix game has a saddle point, game theory recommends that the players choose the strategies giving rise to that point. Unfortunately, as in example 1 above, most matrix games will not have saddle points. Another example of a game without a saddle point is:

Example 4:

$$\begin{matrix} 5 & 1 & 0 \\ -1 & 2 & 1 \\ 1 & -3 & 3 \end{matrix}$$

3.3 Mixed Strategies

If a game does not have a saddle point, it is important to keep one's opponents from learning one's plans. Game theory then recommends the use of a mixed strategy. In a mixed strategy, a player chooses among his strategies (the rows and columns of the matrix, hereafter called pure strategies) by a randomization scheme. The mixed strategy is then an assignment of probabilities to the several pure strategies. Thus it may be possible for an opponent to compute one's mixed strategy (since this particular assignment of probabilities is chosen in a rational manner), but it is impossible to know the actual pure strategy used in any play of the game. We generally represent a mixed strategy (for a given player) as a vector with m components (m being the number of pure strategies that player has), non-negative and adding to 1.

Suppose, then, that in matrix game A, Player I has chosen the mixed strategy x, while II has chosen y. It is not too difficult to see that, assuming these two randomization schemes are independent, there will be an expected payoff equal to the matrix product $x\mathrm{A}y$. Once again, I wishes to maximize, while II wishes to minimize, this quantity.

The question now is how the mixed strategies should be chosen. A very important theorem, the minimax theorem, discovered independently by E. Borel and J. von Neumann during the 1920s, tells us that, for any matrix game, there exist mixed strategies, x^* and y^* (optimal strategies) and a number v (the value of the game) such that, if I uses x^*, he will always have an expected payoff at least equal to v, no matter what II does. Similarly, if II uses y^*, the expected payoff will never be greater than v, no matter what I chooses. The triple (x^*; y^*; v), consisting of two optimal strategies and the value, is the solution of the game.

As an example, note that, in example 1 above, the optimal strategies are $x^* = (1/2, 1/2)$ and $y^* = (1/2, 1/2)$, guaranteeing the value $v = 0$. In example 4, the optimal strategies are $x^* = (2/7, 4/7, 1/7)$ and $y^* = (1/7, 2/7, 4/7)$, guaranteeing the value $v = 1$.

4. Two-person General-sum Games

In other games, the players' interests will not be directly opposed. Thus cooperation between the two players may well be possible, unless the rules of the game (e.g., antitrust laws) specifically forbid this. We therefore consider two cases: the noncooperative and the cooperative game.

4.1 Noncooperative Games

In the noncooperative case, the emphasis has generally been on equilibrium points, (still defined as pairs of strategies such that neither player can gain by a unilateral change). Much as in the case of matrix games, it can be shown that, when mixed strategies are allowed, every finite n-person game has at least one equilibrium n-tuple of strategies. Unfortunately, equilibrium points are not generally as satisfactory here as in the case of matrix games. In fact, a game might have

several different equilibrium points; moreover, even if a game has a unique equilibrium point, it is not clear that it will be best for players to play in equilibrium.

Example 5. In the two-person (non-zero-sum) game:

(4, 1)	(0, 0)
(1, 0)	(2, 5)

the strategy pairs (row 1, column 1) and (row 2, column 2) are both in equilibrium. Note, however, that Player I clearly prefers the first of these, while II prefers the second. It is not clear how the game should then be played.

Example 6. In the game:

(5, 5)	(0, 10)
(10, 0)	(1, 1)

(usually known as the prisoner's dilemma), the pair (row 2, column 2) is the unique equilibrium point. Note, however, that both players will receive a higher payoff if (row 1, column 1) is used.

4.2 Cooperative Games

In case cooperation is allowed, it is usually possible for both players to gain by profitable deals. Since, however, many such deals may be possible, the question of bargaining comes into play. Essentially, the search in such games has been for a 'fair' arbitration rule, one that takes into account what each player can do for himself (without the other player's help) and what the two of them can accomplish together. J. Nash in 1950 developed an axiomatic model for this. Assuming a set S of possible bargains (represented as pairs of payoffs), Nash looks for an outcome (bargain) which will be Pareto-optimal (it is not possible for one player to gain more without taking something from the other) and individually rational (each player gains at least as much as he could by himself) and which moreover satisfies certain technical conditions. Nash shows that the only arbitration rule which satisfies these axioms is the rule that chooses the bargain which maximizes the product of the two players' increments in utility (over what they could get by themselves).

Example 7. Two players are to divide the sum of \$100; if they do not agree on a division of the \$100, each gets \$0. Assume player I's utility is proportional to the amount of money he receives; for Player II, however, suppose that \$$y$ represents a utility equal to the logarithm of $100+y$.

In this case, any division (x, y) with $y \leqslant 100-x$ is permissible. This will give rise to any utility payoffs of the form (x, w), where $w \leqslant \log(200-x)$. Nash's solution is the point which maximizes the product $x(w-\log(100))$, subject to the constraint above. Using differential calculus, it is not difficult to see that this maximum is attained at $x = 54.4$, so that I should receive \$54.40 while II receives \$45.60.

Nash's system of axioms is not the only one possible. Several other systems of axioms, as well as nonaxiomatic approaches, have been suggested. There is, unfortunately, no generally accepted solution to the bargaining problem.

5. n-*Person Games*

5.1 The Characteristic Function

In case there are three or more players, the cooperative game is further complicated by the fact that each player might choose among several possible coalitions. J. von Neumann and O. Morgenstern (1947) approached these games through the *Characteristic Function*. Let $N = \{1, 2, ..., n\}$ be the set of all players in a game. For a given coalition S (nonempty subset of N), $v(S)$ is the amount of utility which the members of S can obtain jointly. It is assumed that this utility can be divided among the members of S in any way, if they agree. Thus v, the characteristic function, assigns a number to each coalition. The problem then is to look for a division of the total amount $v(N)$ which is available in the game, in such a way that each player, i, receives at least the amount $v(\{i\})$. Any payoff vector that does this is an imputation of the game v.

Von Neumann and Morgenstern then introduce the notion of domination. Given two imputations, x and y, we say that x dominates y if there is at least one coalition S such that (a) $x_i > y_i$ for every $i \varepsilon S$ (so that all members of S prefer $\mathbf{x}$), and (b) $\Sigma_{i\varepsilon S}\, x_i \leqslant v(S)$ (so that S is strong enough to obtain x).

5.2 The Core and Stable Sets

One possibility, now, is to look for the undominated imputations. The set of these is known as the core, and is closely allied to the economic concept of market equilibrium.

Example 8. In a three-person game, player 1 owns a horse for which she has no use. Players 2 and 3 would both like to buy the horse; player 2 values it at \$200 while 3 values it at \$100. We will equate money with utility.

It may be seen that any single player can create no utility in this game. Thus $v(\{i\}) = 0$ for $i = 1, 2, 3$. Similarly, $v(\{2, 3\}) = 0$ since the two buyers together can do nothing.

Now, if 1 sells the horse to 2 for \$$x$, then 1 gains x (she has exchanged a useless horse for \$$x$), while 2 gains $200-x$ (he has obtained a horse worth \$200 for \$$x$). Together, then, they gain 200, and we set $v(\{1, 2\}) = 200$. Similarly, $v(\{1, 3)\} = 100$ and $v(\{1, 2, 3\}) = 200$.

For this game, an imputation is any vector (x_1, x_2, x_3) with non-negative components adding to 200. To be in the core, the imputation must additionally satisfy the conditions

$x_1+x_2 \geqslant 200$ and $x_1+x_3 \geqslant 100$. This can only happen if $100 \leqslant x_1 \leqslant 200$, $x_2 = 200-x_1$, and $x_3 = 0$. In effect, it seems that 1 will sell the horse to 2 for a price between \$100 and \$200. Player 3 receives nothing, but helps to drive the price up. (Note that, if player 3 were removed from the game, the price could be anything between \$0 and \$200.) We should point out that, in classical economics, such a situation would reach equilibrium with precisely such a price.

In many cases, unfortunately, the core is empty, and von Neumann and Morgenstern thus develop the concept of a solution (now better known as a stable set). A stable set (for a given game) is a set V of imputations such that (a) no imputation in V dominates another in V and (b) every imputation not in V is dominated by at least one in V. For some two decades, it was conjectured that every game in characteristic function form had at least one stable set, but W. F. Lucas (1967) gave an example of a ten-person game with no stable sets.

Example 9. In a three-person game, there is a trunk full of valuable objects, worth 100 units of utility. This trunk is so heavy that no one of the three players is strong enough to lift it; alternatively, any two acting together can carry the trunk and collect the utility.

We represent this game by the characteristic function u, where $u(\{i\}) = 0$ for $i = 1, 2, 3$, and $u(S) = 100$ for any coalition S with two or three members. For this game, we see that the imputations are vectors with three nonnegative components adding to 100. Now, for an imputation to be in the core, each two-person coalition must receive at least 100 units. But this means that the players must receive on average at least 50 units; i.e., the three together must receive at least 150 units. This is impossible since there are only 100 units available, and we conclude that, for this game, the core is empty.

However, it is not difficult to check that the set V, consisting of the three imputations (50, 50, 0), (50, 0, 50), and (0, 50, 50) is a stable set. (This is not the only stable set for this game.)

5.3 The Shapley Value

Both the core and the stable set solution concepts suffer from the fact that (a) the first may be empty, the latter might not exist, and (b) in many cases they contain an enormous number of imputations. Game theorists have therefore looked for other solution concepts. Foremost among these is L. S. Shapley's value.

First developed in 1953, the Shapley value or power index is based on the three axioms of efficiency (if a game is played entirely by a set of players, these players will share the total available utility), anonymity (the value depends only on the players' possibilities, and not on the names given them) and additivity (the value of the sum of two games is equal to the sum of the two games' values). Shapley proved that there is a unique value scheme satisfying these three axioms over the set of all n-person games, and this scheme assigns to player i, in game v, the quantity $\phi_i[v]$ given by the formula

$$\phi_i[v] = \sum_{S \subset N} \gamma_{n,s} \, [v(S \cup \{i\}) - v(S)]$$

where the sum is taken over all coalitions, S, and for these, s is the number of players in S, while $\gamma_{n,s} = s!(n–s–1)!/n!$.

As an example, the horse game (example 8) has value $\phi[v] = (116.7, 66.7, 16.7)$. The game of example 9, being totally symmetric, has value $\phi[u] = (33.3, 33.3, 33.3)$.

The Shapley value is also very useful as a measure of voting power in political situations. A voting structure is a game v whose players are the voters, N, and such that, for any coalition S, $v(S) = 1$ if S has enough votes to win an election (or approve a motion), and $v(S) = 0$ otherwise.

Example 10. In a committee with four members, assume member 1 (the chairman) has the power to break ties. In this case, $v(S) = 1$ if S has three or four members, or if S has two members, including the chairman. For this game, the value is $\phi[v] = (0.5, 0.167, 0.167, 0.167)$.

5.4 The Bargaining Set(s)

Of other solution concepts for n-person games, we mention only the bargaining sets, developed by R. Aumann and M. Maschler (1995). Given a coalition structure, $T = \{T_1, T_2, ..., T_m\}$ (a partition of the set of players) we consider only payoff vectors $\mathbf{x}$ which give to each of the sets T_j (the coalitions which are assumed to form) their total possibility $v(T_j)$ and to each player, i, at least $v(\{i\})$. Such pairs $\langle \mathbf{T}; \mathbf{x} \rangle$ are individually rational payoff configurations (i.r.p.c.'s). Now, given an i.r.p.c., a player, i, can object against another player, k, if i can receive greater utility in another i.r.p.c. $\langle \mathbf{U}; \mathbf{y} \rangle$ in which she does not have player k as a coalition partner. Of course, k can counterobject if there is yet another i.r.p.c. $\langle \mathbf{V}; \mathbf{z} \rangle$ in which he does not need i as a partner, and such that $z_k \geqslant x_k$. In this case, the objection is unjustified; otherwise (i.e., if there is no counter-objection) the objection is justified.

Now, an i.r.p.c. is said to be stable, or belong to the bargaining set $\mathbf{M}$, if there are no justified objections. Because of different possible relations allowed between the objecting and objected players, i and k, there are several somewhat different bargaining sets. Aumann and Maschler prove that in one case, namely when i and k have to be partners in the original structure $\mathbf{T}$, the bargaining set—here denoted $\mathbf{M}_1^{(i)}$, is nonempty in the sense that, for any game v, and any coalition structure $\mathbf{T}$, there exists at least one $\mathbf{x}$ such that $\langle \mathbf{T}; \mathbf{x} \rangle$ belongs to $\mathbf{M}_1^{(i)}$.

Example 11. In the three-person game of example 9, if the coalition structure is $T = \{\{1, 2\}, \{3\}\}$, the only stable payoff vector is (50, 50, 0). If, instead, the grand coalition ($\mathbf{T} = \{N\}$) forms, then the only stable payoff vector is (33.3, 33.3, 33.3).

In the four-person game of example 10, if the coalition structure $\mathbf{T} = \{\{1, 2\}, \{3\}, \{4\}\}$ forms, then the stable payoffs are of the form (t, 1–t, 0, 0), where $0.5 \leqslant t \leqslant 0.67$.

6. *Games with Incomplete Information*

In some cases, players are unaware of some of the available strategies, or might be uncertain as to the payoffs which correspond to certain choices. In particular, they may be unaware of the utilities which other players assign to certain outcomes (e.g., we frequently do not know whether we are dealing with an honest or dishonest person—an honest person being one who assigns high utility to keeping an agreement). In such situations—known as games with incomplete information—repeated play will often give us further information—a person who does not keep an agreement may simply have made an honest mistake, but if this happens repeatedly, we generally conclude he is dishonest. Of course, a dishonest person, realizing this, will frequently keep some less important agreements so as to 'lull' others in preparation for a big scam. In some important work dating back to the 1960s, J. Harsanyi (1977) on the one hand, and Aumann and Maschler on the other, address this problem of learning and deception.

7. *Conclusion*

Game theory represents an important attempt to give a mathematical basis to behavior in conflict situations. Admittedly, there are elements of human activity which are not easily modeled, and this leads to certain gaps in the theory. Nevertheless, it has served to give a mathematical structure to several well-known ideas (bluffing, bargaining) and to develop new insights into human behavior. As of this writing, it is considered an important mathematical tool for economists and other social scientists.

See also: Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision Theory: Bayesian; Game Theory and its Relation to Bayesian Theory; Game Theory: Noncooperative Games; Luce's Choice Axiom; Mathematical Psychology

Bibliography

Aumann R J, Maschler M B, Stearns R E 1995 *Values of Repeated Games with Incomplete Information*. MIT Press, Cambridge, MA

Damme E van 1987 *Stability and Perfection of Nash Equilibria*. Springer, Berlin

Harsanyi J C 1977 *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge, UK

Luce R D, Raiffa H 1957 *Games and Decisions*. Wiley, New York

Owen G 1995 *Game Theory*, 3rd edn. Academic Press, San Diego, CA

von Neumann J, Morgenstern O 1947 *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

G. Owen

Game Theory and its Relation to Bayesian Theory

This entry explores some open controversies relating to a contrast between rival decision paradigms. Alas, it cannot be a report on a unified theory of games: such a thing does not exist. The controversies between Bayesian decision theory and traditional Game Theory have a common source, however, which may be identified informally, as follows.

Bayesian decision theory comes in many varieties, Good (1971). Common to all is one rule: the principle of maximizing (subjective) conditional expected utility. Generally, an option in a decision problem is depicted as a (partial) function from possible states of affairs to outcomes, each of which has a value represented by a (cardinal) utility. Where Bayesian decision makers are uncertain about which state of affairs obtains, they assign conditional probabilities to those states, conditional upon the choice of the option. The Bayesian decision rule is to choose an option from the set of available options that maximizes subjective conditional expected utility—assuming a maximum exists, as will be done here. (Some Bayesian theories, e.g., Savage (1954), use a restricted formulation where acts and states are probabilistically independent, and dominance considerations apply. Then the Bayesian decision rule is simplified: maximize subjective expected utility.) Thus, the primary Bayesian decision rule assesses the value of an option by a weighted average of the utilities of potential outcomes, weighted according to the decision maker's (conditional) personal probability over states (see *Decision Theory: Bayesian*).

Traditional Game Theory, by contrast, requires that rational players choose strategies that form an equilibrium—though which variety of equilibrium they adopt is also a matter of serious debate among game theorists, as noted below. However, under each sense of 'equilibrium' that is proposed, a player is required to assess the value of a strategy in terms of a best response that an opponent might make against it. That is, by game theoretic standards, against each available strategy that the player might select, typically

there is one state of affairs, namely the *best reply* that might be chosen by the (collective) opponent(s). That *best reply* is what determines the value of the strategy for the player. When all the players' choices jointly meet the criterion of being best replies, a *Nash equilibrium* results.

At the core of the debate, then, is this conflict. On the one hand is the Bayesian decision rule as that might be applied to games. It is a rule that specifies the value of a player's choice of a strategy in a game as a weighted average over possible replies that an opponent might make. Here, the set of potential strategies available to the (collective) opponent(s) constitute the states of affairs over which the Bayesian player is uncertain, and the weights assigned to these are the player's subjective probability distribution for what the opponent might do. On the other hand, there is the game theoretic rule that uses an opponent's best reply as the primary criterion to determine the value of a player's choice of strategy. Of course, since a distribution with all its weight concentrated on one point—corresponding to the state of affairs that is the opponent's *best reply*—is not precluded by Bayesian theory; sometimes, but usually not, the two approaches agree. But Bayesian theory generally does not privilege the Nash equilibrium solution. Before we examine games where the two approaches give different recommendations, and some recent attempts to mitigate these differences, it is worthwhile to review a few points of shared agreement.

1. Illustrations Where Expected Utility Considerations Are Welcome Within Traditional Game Theory

1.1 Decisions Under Risk

First, in their *magnum opus*, *The Theory of Games and Economic Behavior*, von Neumann and Morgenstern (1944), formalize the assumption that each player's preferences for outcomes in a game can be quantified by a (cardinal) utility. That is, they assume that a player's preferences over outcomes satisfy the expected utility rule applied to decisions under *risk*. Specifically, when options are presented as *lotteries* over outcomes, i.e., when a (simple) probability distribution for outcomes is well defined, then von Neumann and Morgenstern assume that the decision maker's preferences accord with the Bayesian rule of maximizing expected utility. This assumption is critical to the received game theoretic interpretation of *mixed strategies*, whose values are given by expectations over the numerical outcomes of pure, i.e., non-mixed strategies.

That is, the numerical values used to describe outcomes in games depend for their meaning on the application of expected utility in decisions under *risk*. This much of von Neumann and Morgenstern's theory is in common with Bayesian decision theory, though it involves a game with one player. Of course, Bayesian decision theory, but not Game Theory, extends the rule to maximize expected utility also to decisions under *uncertainty*—where well-defined probabilities over states of affairs are not specified by the assumptions of the decision problem. In particular, Game Theory does not extend the Bayesian decision principle to problems under uncertainty with more than one player where the relevant states of affairs are decisions by an opponent!

1.2 deFinetti's Game of Coherent Previsions

In a seminal work, deFinetti (1937) proposed a game theoretic basis for using subjective expected utility. Here is a simplified version of that game-simplified to avoid important technical considerations involving deFinetti's use of merely finitely additive probability that arise with infinite state spaces.

Let $\Omega = \{\omega_1, \ldots \omega_n\}$ be a finite set of pairwise exclusive and mutually exhaustive states, i.e., Ω is a finite partition of the sure event. Call the elements of Ω the *states*. Let $\{X_i\colon i = 1,\ldots\}$ be a set of random variables defined on Ω. That is, each $X_i\colon \Omega \to \mathscr{R}$ is a simple, real-valued function defined on the partition Ω. *Player*$_1$, whom we call the *bookie* in the game, is obliged to play first and to reveal a set of real-valued quantities, called *previsions* $\{p_i\colon i = 1,\ldots\}$. A set of previsions constitutes a strategy for the *bookie*. (The game requires that the bookie reveals a pure strategy.) *Player*$_2$, whom we call the *gambler* in the game, plays second and, having seen the strategy chosen by the *bookie* is then allowed to choose a set of real-valued terms $\{\alpha_i\colon i = 1,\ldots\}$, subject to the constraint that only *finitely* many of these terms may differ from zero. (For added simplicity in this presentation, assume that $\sum_i |\alpha_i| \leqslant 1$.) A set of terms constitutes a strategy for the *gambler*.

Next, consider the (simple) random variable, $Y = \sum_i \alpha_i |X_i - p_i|$, whose value $Y(\omega_i) = y_i$ is fixed by the state. Given a strategy choice by each player, the payoff to the *bookie* in deFinetti's game is the maximum $_\Omega\{yi\} = y^*$, while $-y^*$ is the corresponding payoff to the *gambler*. Hence, this is a 2-person, zero-sum game in which the *bookie* plays first.

Since the *gambler* may choose all the terms equal to 0, in which case $Y = 0$ for sure, the *gambler* can ensure that the value of the game to the *bookie* is bounded above by 0. Now, either the *bookie*'s *previsions* $\{p_i\colon i = 1,\ldots\}$ form a set of expected values for the random variables $\{X_i\colon i = 1,\ldots\}$, based on a single probability distribution P over Ω, or they do not. deFinetti showed that one of two circumstances obtain, respectively: If there exists a probability P such that $\mathbf{E}_P[X_i] = p_i$, then the *gambler* can do no better than securing the value 0 for the game, and the *bookie* is said to have *coherent previsions*.

Aside: This can be seen as follows. When the bookie's previsions are coherent, then $\mathbf{E}_P[Y] = 0$ regardless of the gambler's choice of strategy. But a weighted average of terms (y_i) cannot exceed its maximum term (y^*); $0 = \mathbf{E}_P[Y] \leqslant y^*$. Thus, the coherent bookie is assured a value of at least 0.

If, however, the *bookie*'s chosen strategy does not agree with the expectations from a probability P, then the gambler may find a reply that insures a strictly negative outcome to the *bookie*: then the *bookie* is said to have *incoherent previsions*. That is, if and only if the *bookie* does not play as though the *previsions* are (subjective) expected values of the random variables, calculated with some subjective probability P, the *bookie* is in jeopardy to suffer a sure loss from a clever choice of strategy by the *gambler*. Thus, though the *bookie*'s prudent, equilibrium strategy is to play a coherent set of previsions, just as though maximizing expected utility, which equilibrium from the continuum of coherent sets of previsions is chosen remains unspecified.

Aside: When the random variables are indicator functions for events, i.e., when they are indicators for subsets of Ω, then this result reduces to the familiar Dutch Book Theorem (see *Probability: Interpretations*).

For an original application of deFinetti's prevision game to problems of cooperative decision making among several agents, see Nau's interesting work (1999).

1.3 Admissible Statistical Decision Functions

Wald (1950) proposed a very general theory of statistical decisions where, in effect, the statistician plays a zero-sum game against Nature. The statistician chooses a statistical decision rule and Nature fixes the unknown state of affairs. In outline and simplified for presentation here, Wald's influential program for framing statistical problems uses the following structure. Let θ be a real-valued statistical parameter with parameter space Θ, a partition of the relevant states of affairs. Let X be a random value whose distribution $\boldsymbol{P}_\theta(X)$ depends upon the parameter θ. The statistician anticipates observing $X = x$ and in advance of the observation chooses a statistical *decision rule*, denoted $d(\cdot)$, that may depend upon the value of X to be observed. The result to the statistician of the *terminal decision* $d(x)$ depends upon the unknown parameter θ. The value of that result is specified by a loss function, $\boldsymbol{L}(d(x), \theta)$, understood as the negative (cardinal) utility of the consequence to the statistician of $d(x)$ when θ obtains. In Wald's presentation, typically the loss function is nonnegative. When, for each value of θ the minimum (over the set of terminal decision) of the loss function is 0, the utility for payoffs is then said to be formulated as a *regret* function.

Aside: In Bayesian decision theory, when acts and states are probabilistically independent and utility is bounded—as in Savage's theory—then, without loss of generality, the decision problem may be reformulated with utility represented by a regret function.

Here are key definitions in Wald's theory.

The *risk function* $\boldsymbol{R}$ for a decision rule $d(\cdot)$ is the expected loss of adopting $d(\cdot)$ as a function of the unknown parameter θ.

$$\boldsymbol{R}(d(\cdot), \theta) = \mathbf{E}_\theta[\boldsymbol{L}(d(\cdot), \theta)].$$

A decision rule $d'(\cdot)$ is *inadmissible* if there exists another decision rule $d(\cdot)$ such that, for each value θ the risk for d is never greater than the risk for d', and for some value of θ the risk for d is less. In that case, say that rule d *dominates* rule d'.

That is, for each θ $\boldsymbol{R}(d(\cdot), \theta) \leqslant \boldsymbol{R}(d'(\cdot), \theta)$, and for some θ $\boldsymbol{R}(d(\cdot), \theta) < \boldsymbol{R}(d'(\cdot), \theta)$.

A decision rule that is not dominated is *admissible*.

Given a probability distribution $\boldsymbol{P}(\theta)$ for the parameter, the Bayes-risk $\boldsymbol{r}$ for a decision rule $d(\cdot)$ is the expected risk with respect to $\boldsymbol{P}$ of adopting $d(\cdot)$.

$$\boldsymbol{r}(d(\cdot), \boldsymbol{P}(\theta)) = \mathbf{E}_P[\boldsymbol{R}(d(\cdot), \theta)] = \mathbf{E}_P[\boldsymbol{E}_\theta[\boldsymbol{L}(d(\cdot), \theta)]].$$

A decision rule $d(\cdot)$ is *Bayes-admissible with respect to the distribution* $\boldsymbol{P}(\theta)$ if, for each dominating rule $d'(\cdot)$, the set of parameter values where $d'(\cdot)$ has strictly lower risk has $\boldsymbol{P}$-probability 0.

A decision rule $d(\cdot)$ is minimax for the statistician if it minimizes the maximum risk over the parameter space (see *Decision Theory: Classical*).

The following result (Berger (1985, p. 254) or Schervish (1995 p. 155) helps to explain the relevance of Bayesian expected utility considerations in identifying admissible decision rules.

Theorem: Suppose that the Bayesian statistician's 'prior' distribution $\boldsymbol{P}(\theta)$ gives positive probability to each open interval. Assume that the risk function $\boldsymbol{R}(d(\cdot), \theta)$ is continuous in θ for each decision rule $d(\cdot)$. Then if $d^*(\cdot)$ is Bayes-admissible with respect to $\boldsymbol{P}$, it is admissible.

The resulting connection of Bayesian decision rules to Game Theory, as first developed by Wald, is indicated by the following (see Berger (1985, p. 353) or Schervish (1995, p. 167)).

Theorem: If decision rule $d(\cdot)$ has constant risk and if it is admissible, then it is minimax.

Thus, if $d^*(\cdot)$ is Bayes-admissible with respect to the distribution $\boldsymbol{P}$ and has constant risk, it is minimax. Also, in that case $\boldsymbol{P}$ is said to be the *worst-case* prior, since it is Nature's (mixed) strategy for achieving a maximin solution in the statistical game. Though the

details are beyond the presentation here, similar results apply to certain limits of Bayes-admissible solutions, obtained by considering a sequence of (proper) priors. (See Wald (1950 Theorem 3.17) for a result showing that every admissible decision rule either is Bayes-admissible for some distribution $\boldsymbol{P}$ or is the limit of a sequence of Bayes-admissible solutions.) The following illustrates the more general case.

Example: Given the parameter θ, let X have a Normal $\mathbf{N}[\theta, 1]$ distribution with $\Theta = \mathscr{R}$. The statistician is required to provide a point-estimate for θ under a squared-error loss function, $\boldsymbol{L}(d(x), \theta) = (d(x)-\theta)^2$. The intuitive rule $d(x) = x$ is admissible with constant risk $\mathbf{E}_\theta[(x-\theta)^2] = \text{Variance}[X] = 1$. For the (conjugate) Normal prior $\boldsymbol{P}$, with $\theta \sim \mathbf{N}[0, \sigma^2]$, a Bayes-admissible decision rule is $d(x) = x\sigma^2/(\sigma^2+1)$, with risk function $\boldsymbol{R}(d(\cdot), \theta) = (\sigma^4+\theta^2)/(\sigma^2+1)^2$ and Bayes-risk $\boldsymbol{r}(d(\cdot), \boldsymbol{P}(\theta)) = \sigma^2/(\sigma^2+1)$. Consider a sequence of such conjugate priors, $\boldsymbol{P}_n$, where $\theta \sim \mathbf{N}[0, \sigma_n^2]$, with $\lim_n \sigma_n^2 \to \infty$. The corresponding Bayes-admissible rules are just, $d_n(x) = x\sigma_n^2/(\sigma_n^2+1)$. The limiting Bayes-rule is the intuitive rule, $d(x) = x$, which is admissible and has constant risk; hence, it is minimax. From the statistician's point of view, the *worst case prior* is the 'improper' uniform distribution for θ: that is Nature's maximin strategy. (See Schervish and Seidenfeld (1996) for a minimax theorem that applies when the worst case prior is 'improper.') Thus, the intuitive decision rule in this problem, $d(x) = x$, illustrates a case where the game theoretic equilibrium and extended Bayesian expected utility theory (using limits of Bayes-admissible strategies) coincide.

In the preceding example, each of the decision rules $d_n(x)$ is Bayes-admissible with respect to the prior $\boldsymbol{P}_n$, and is an admissible decision rule as well. Thus, even in this heuristic example we see the tension between the two decision paradigms. The Bayesian statistician may have prior beliefs more accurately depicted by one of the (proper) conjugate priors $\boldsymbol{P}_n$, rather than by the limiting, worst-case prior. In such a situation, the Bayesian statistician will not be motivated by equilibrium considerations to adopt the game theoretic solution instead of maximizing expected utility with the decision rule $d_n(x)$. This is the tension explored in the next section.

2. *Equilibrium Refinements and Maximizing Expected Utility: Rationalizable Strategies, Correlated Equilibrium, and Trembling Hands*

What if each player in a game is a subjective expected utility maximizer, models the opponents that way, and these assumptions are common knowledge among the players? That is, though each player is uncertain about what choices opponents will make, nonetheless each player behaves as though he or she holds a (subjective) probability over those unknown states of affairs and maximizes expected utility against that probability distribution. But, insofar as players think that their opponents also maximize expected utility, their probabilities must be supported by the set of their strategies that maximize their expected utilities. That all the players share these assumptions leads to a modest but interesting restriction on the set of Bayesian equilibria that result.

Bernheim (1984) and Pearce (1984) consider the important cases where the Bayesian players have probability models for each other's play where, regarding simultaneous play, acts and states are probabilistically independent. That is, assume that when the game requires players to make moves in ignorance of (some of) the moves of their opponents, then from the perspective of each player, the opponents' moves are probabilistically independent of choices at that time. The result is the set of *rationalizable* strategies, which carries the following elegant description. A strategy is *rationalizable* if and only if it survives the recursive operation of deleting successively options that are *strictly* dominated by another option not already deleted.

> Aside: The reader is alerted to the fact that the recursion mentioned above iteratively changes the set of states of uncertainty for a player, i.e., the set of surviving options available to the opponents, that are relevant to a player when determining whether strict dominance obtains between two available strategies.

For an examination of how rationalizability differs from the traditional criterion of Nash equilibrium, particularly for sequential games in extensive form, see the articles by DeGroot and Kadane (1983) and Kadane and Seidenfeld (1992).

The set of rationalizable strategies in a game, as judged by the common standards of traditional Game Theory, typically makes an uncomfortably large collection of permissible moves, offering too little guidance about how the game should be played. An even more radical departure from the traditional point of view is proposed by Kadane and Larkey (1982) who examine the question of what considerations become salient to a Bayesian game-player who does *not* assume Bayesian-rationality on the part of opponents. They argue that the distinctions that are most important to traditional Game Theory are of only passing relevance under their radical reformulation of the same decision problems. For the practical side of this matter, see *Negotiations: Statistical Aspects*.

In the other direction, Aumann (1974, 1987) proposes a narrowing of the rationalizable strategies by introducing additional components of common knowledge among the players. Specifically, Aumann argues for a theory of rationality in which the Bayesian players share a common 'prior' over all the states of uncertainty that are relevant to choices made in the game. The upshot is a powerful criterion of *correlated equilibrium*, where players' *mixed strategies* form a best reply against each other, while relying on what are

in effect probabilistically *dependent* events. This allows for Pareto gains, where all players do better with a correlated equilibrium compared to their expected gains from a Nash equilibrium, when the mixed strategies of the players use jointly *independent* events only.

The relevant states of affairs over which the players have a common prior in Aumann's theory include the choices of all the players. This leads to the following controversy. Can Bayesian players assign nontrivial personal probabilities to their own (current) choices? Levi 1997 examines this intricacy and argues against the coherence of Aumann's condition.

Another interesting approach to reconciling equilibrium theory with expected utility analysis is found in the important works by Harsanyi (1977), and Harsanyi and Selten (1988). Harsanyi's doctrine is that Bayesian agents ought to use special distributions—oftentimes uniform probabilities—to reflect their ignorance about whatever states of uncertainty are relevant to their choices. (Note the connection with the *worst case* prior of the *Example* given in Section 1.3.) In the joint work with Selten, cited above, their novel proposal relies on the following idea: Though Bayesian players have definite preferences over strategy choices, they can achieve only a *trembling-hand* selection, where some residual probability is invariably assigned to other options besides the one most desired. In short, the players' options are modeled as nontrivial mixed strategies with some residual probability attached to even undesirable choices, and this is common knowledge. Trembling-hand equilibrium analysis examines limits of Nash equilibrium strategies as the trembles diminish. The results underwrite the criterion that choices in extensive form games should obey the principle of *subgame perfection*, i.e., that self-defeating threats are only idle threats. This criterion is particularly significant for application of the rule of *backward induction* in the theory of sequential choices, e.g., extensive form games when these are viewed from the Bayesian perspective.

3. Additional Sources

Luce and Raiffa's (1957) work remains a first-rate source of basic ideas and results contrasting the rival decision theories discussed here. Blackwell and Girshick's (1957) book is an excellent introduction to the interplay of ideas between Bayesian decision theory and traditional Game Theory relating to the design of statistical experiments. Binmore's stimulating textbook (1992) is a rich source of heuristic examples and informed discussion illuminating rival accounts of equilibrium refinements, going well beyond the few varieties alluded to here. Binmore is not shy, commenting openly on some of the more radical attempts at applying Bayesian decision theory to games. Even a casual reading of Binmore's textbook will convince the reader of the truth of the opening sentences in this article!

See also: Game Theory; Game Theory: Noncooperative Games

Bibliography

Aumann R J 1974 Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* **1**: 67–96

Aumann R J 1987 Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* **55**: 1–18

Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Springer-Verlag, New York

Bernheim B D 1984 Rationalizable strategic behavior. *Econometrica* **52**: 1007–28

Binmore K 1992 *Fun and Games: A Text on Game Theory*. D. C. Heath and Company, Lexington, MA

Blackwell D, Girshick M A 1954 *Theory of Games and Statistical Decisions*. John Wiley, New York

deFinetti B 1937 La prévision, ses lois logiques, ses sources subjectives. *Annals de l'Institute Henri Poincaré* **7**: 1–68

DeGroot M H, Kadane J B 1983 Optimal sequential decisions in problems involving more than one decision maker. In: Rustagi R H, Siegmund D (eds.) *Recent Advances in Statistics*. Academic Press, New York, pp. 197–210

Good I J 1971 46,656 varieties of Bayesians. *American Statistician* **25**: 62–3

Harsanyi J C 1977 *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge, UK

Harsanyi J C, Selten R 1988 *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA

Kadane J B, Larkey P D 1982 Subjective probability and the theory of games. *Management Science* **28**: 113–20

Kadane J B, Seidenfeld T 1992 Equilibrium, common knowledge, and optimal sequential decisions. In: Bicchieri C, dalla Chiara M L (eds.) *Knowledge, Belief, and Strategic Interactions*. Cambridge University Press, Cambridge, UK, pp. 27–45

Levi I 1997 Prediction, deliberation, and correlated equilibrium. In: Levi I (ed.) *The Covenant of Reason: Rationality and the Commitments of Thought*. Cambridge University Press, Cambridge, UK, Chap. 5

Luce R D, Raiffa H 1957 *Games and Decisions*. Dover, New York

Nau R 1999 *Arbitrage, Incomplete Models, and Interactive Rationality*. Technical Report. Fuqua School of Business, Duke University, Durham, NC

Pearce D G 1984 Rationalizable strategic behavior and the problem of perfection. *Econometrica* **52**: 1029–50

Savage L J 1954 *The Foundations of Statistics*. John Wiley, New York

Schervish M J 1995 *Theory of Statistics*. Springer-Verlag, New York

Schervish M J, Seidenfeld T 1996 A fair minimax theorem for two-person (zero sum) games involving finitely additive strategies. In: Berry D, Chaloner C, Geweke J (eds.) *Bayesian Analysis in Statistics and Econometrics*. John Wiley, New York, pp. 557–68

von Neumann J, Morgenstern O 1944 *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

Wald A 1950 *Statistical Decision Functions*. John Wiley, New York

T. Seidenfeld

Game Theory: Noncooperative Games

Games are mathematical models of interactive strategic decision situations: there are various actors (players) involved that jointly determine the outcome and each tries to obtain that outcome that is most favorable to him. A game is said to be noncooperative if there are no possibilities for commitment (unilateral or multilateral) outside the rules of the game. In contrast, in cooperative games, players can form coalitions with the possibilities for doing so not being explicitly modeled within the rules of the game. Nash (1953) argued that the two approaches are complementary and he proposed building noncooperative models of cooperative games, an idea that is referred to as the Nash program. The article describes noncooperative game models and solution concepts and some applications are indicated (see Aumann 1987 for a more extended overview).

1. The Problem

Imagine an interactive decision situation in which various individuals (players) are involved, whose decisions affect each other. Depending on the decisions that are taken, a certain outcome will result; players have preferences over these outcomes and typically there is a conflict of interest as different players prefer different outcomes. Each player now faces the question what decision is best. As the decisions are interdependent, the answer to the question will depend not only on the player's own preferences, but also on the decisions that the other players take. The player, hence, has to make a prediction about what the other players will do. For an outsider there is the related question of predicting the overall outcome of the game. Based on the assumption of rational behavior of all players involved, game theory offers a set of tools and concepts that provide answers to these questions.

The above mentioned problem arises in a variety of circumstances, ranging from parlor games (chess, poker, bridge) to various economic, political, military, or biological situations. Game theory offers a collection of formal models and solution concepts to analyze these situations. While the history of the field can be traced back to the analysis of the two-person card game *le Her* by James Waldegrave in the early eighteenth century, with important contributions by the French scientists Cournot and Borel in the nineteenth and twentieth centuries (Baumol and Goldfeld (1968), Weintraub (1992)), John von Neumann generally is regarded as the founder of the field. In von Neumann (1928), he showed that the above mentioned problem allows a solution in the case of two players with strictly opposed interests (zero-sum games). The book *Theory of Games and Economic Behavior* that von Neumann wrote together with the economist Morgenstern (von Neumann and Morgenstern 1944) demonstrated convincingly that the tools developed initially for parlor games could be applied successfully to a variety of social conflict situations.

2. Noncooperative Game Models

The simplest type of model is one in which there is no dynamics: each player has to make only one decision and decisions are effectively made simultaneously as no player has knowledge about the decision that has been made by another player. As an example, one might think of a sealed bid procurement auction: without knowing the bids of the competitors, each player submits their bid in a sealed envelope; the player making the lowest bid is awarded the contract and in turn receives a payment (say) equal to their bid. Such a game is called a game in normal form or also a game in strategic form.

The formal mathematical definition of an n-player normal form game is as follows. Let S_i be the (finite) set of possible decisions (also called actions or strategies) that player i might take. Furthermore, let $S = X_{i=1}^{n} S_i$ be the set of strategy profiles, hence, $s \in S$ specifies a strategy for each and every player in the game. Assume that each player's preferences over the set of outcomes of the game can be described by a (von Neumann and Morgenstern 1953) utility function, hence, each player wants to maximize their utility and only cares about expected utility. Von Neumann and Morgenstern (1953) give conditions under which such a utility function can be found; some theory has been developed also without this assumption, but it is convenient to maintain it here. Each strategy profile $s \in S$ produces a certain outcome and right $u_i(s)$ for the utility of player i associated with this outcome. The normal form game is then completely specified by the strategy sets S_i and by the utility functions u_i with u_i: $S \to \mathbb{R}$, hence $G = \langle S_i, \ldots, S_n, u_i, \ldots, u_n \rangle$.

Games in which players move more than once, and/or in which moves of different players are sequential can be represented by means of a tree. Formally, such a game is said to be in extensive form. The nodes of the tree correspond with decision points of players, with the arcs at a node representing the decisions that are possible at that node. Each endpoint of the tree represents an outcome of the game and again it is associated with a utility for each of the players. A special type of extensive form game is one in which there is perfect information: when a player has to move, they have full information about where they are in the game and what decisions have been taken

before. Chess can be viewed as an extensive form game with perfect information. So-called information sets are used to represent limited information in a tree: when a player cannot distinguish two nodes in the tree, these are in the same information set. When a player has to move, they can condition their action only on the information set that is revealed to them, not on the actual node. The theory has been developed mainly for games with perfect recall, that is, each player is assumed to fully remember their previous actions and information. Kuhn (1953) provides a formal development. For some recent contributions dealing with imperfect recall, see Rubinstein (1998).

Von Neumann (1928) introduced the fundamental concept of strategy by means of which an extensive form game can be reduced to one in normal form. Formally, a strategy for a player is a full plan of action that specifies for each information set of this player the action that this player intends to take if and when that information set is reached. Clearly, once a strategy has been determined for each player, the outcome is determined as well. Von Neumann argued that, for rational players, there is no loss of generality in forcing them to think through the game in advance, hence, to force them to choose a strategy. Consequently, the extensive form can be reduced to the simpler normal form.

The above description abstracted from randomization. A player, however, might randomize before the start of the game which (pure) strategy they might use during the game. Alternatively, they might randomize to determine which action they will choose once a certain information set is reached. The concept of mixed strategy is used to refer to the former case, the concept of behavior strategy refers to the latter case of local randomization. Kuhn (1953) showed that both concepts are equivalent for games with perfect recall, that is, whatever a player can do with a strategy of one type they can also do with a strategy of the other, and vice versa. Note that since players are assumed to have von Neumann–Morgenstern utilities, replacing 'payoff' (utility) by 'expected payoff' creates no special difficulties.

The traditional theory has been developed for games of complete information, that is, each player is assumed to know the utility functions of all the players in the game. Frequently, a player will not have all information about their competitors. For example, in the procurement context referred to above, a player may not know how many orders a competitor has in stock, hence, they may not know how much value an opponent assigns to winning the present contract. Harsanyi (1968) showed how to incorporate incomplete information into the model. To represent asymmetric information, Harsanyi introduces an artificial chance move at the beginning of the game that determines which piece of private information each player will have. This piece of private information is also called the player's type. It is assumed that the set of possible types and the probability distribution over this set is common knowledge, however, which type realization results for player i is typically only known to player i himself. A normal form game with incomplete information, hence, does not only specify the players and their utility functions, but also their types and the joint probability distribution on players' types. Formally, such a game is given by a tuple $G = \langle S_1,\ldots, S_n, T_1,\ldots, T_n, u_1,\ldots, u_n, p\rangle$ where S_i is player i's strategy set, T_i is the set of possible types of player, i, $u_i: SxT \rightarrow \mathbb{R}$ is player i's payoff function (where $T = X_{i=1}^{n} T_i$), and p is the probability distribution on T. The play of the game proceeds as follows: A type profile $t \in T$ is determined according to p and player i is informed about their type t_i. Based on this information, player i updates information about the types of the others, computing the posterior probability $p_i(t/t_i)$ and chooses an action to maximize the associated expected payoff. Harsanyi's modeling technique has shown to be very powerful (see *Auctions*; *Information*, *Economics of*).

3. Solution Concepts

3.1 Individual Rationality

Concepts of 'individual rationality' from decision theory can also be used in a game context. If one strategy s_i of player i is (strictly) dominated by another strategy s_I of this player (i.e., it yields a lower payoff no matter what strategy profile s_{-1} the other players use, formally $u_i(s_i, s_{-i}) < u_i(s'_i, s_{-i})$ for all s_{-i}) then a rational player i will not use s_i and it should be possible to eliminate s_i without changing the solution. Furthermore, dominated strategies can be iteratively eliminated. Alternatively, a 'Bayesian' player will represent their uncertainty about their opponent's strategies by a probability distribution σ_{-i} on $S_{-i}(= X_{j \neq i} S_j)$ and they will thus only choose strategies that are best responses against some such correlated strategy σ_{-i}. By using duality theory from linear programming it can be shown that s_i is a best response against some correlated strategy σ_{-i} of the opponents if and only if s_i is not strictly dominated, hence, the two concepts are equivalent. In most of the theory it has been assumed that the fact that players decide independently implies that a player will represent their uncertainty by a profile of mixed strategies of the opponents, that is, the components of σ_{-i} are independent. The set of strategy profiles that remain after all strategies of all players that are not best responses against independent mixtures of opponents have been iteratively eliminated is called the set of rationalizable profiles (Bernheim 1984, Pearce 1984). The set of rationalizable strategies may be smaller than the set of iteratively undominated strategies.

In his theory of the 2-person zero-sum game, von Neumann also used an individualistic rationality concept. If player i uses mixed strategy σ_i then the worst that can happen to is to receive the expected payoff $\min_{\sigma_{-i}} u_i(\sigma_i,\sigma_{-i})$, hence, player i can guarantee the payoff $\max_{\sigma_i}\min_{\sigma_{-i}} u_i(\sigma_i,\sigma_{-i})$ by choosing their mixed strategy appropriately. Note that a mixed strategy may 'guarantee' a higher expected payoff than a pure strategy. For example, in 'matching pennies' (two players simultaneously show the side of their penny with player 1 (resp. player 2) receiving both pennies if the sides are the same (resp. different) it is obviously better to keep one's side hidden, hence, to use a mixed strategy. Von Neumann showed that, for finite 2-person zero-sum games, there exist number v_1 and v_2 with $v_1+v_2=0$ such that each player i can guarantee v_i. We say that is the security level of player i: against a 'rational' opponent, player i cannot expect to get more than this level. Strategies that guarantee the security level are called optimal strategies and v_1 is also termed the value of the game. Von Neumann's theorem is also called the minmax theorem, it thus shows that there is indeed an optimal way to play a finite 2-person zero-sum game.

3.2 Equilibrium

Nash (1950, 1951) was the first to break away from the individualistic rationality concepts. Rather than focusing on the direct question 'what should a player do' (or 'what should a game theorist advise a player to do'), he approached the problem indirectly and asked 'if the game has a solution, what properties would it have.' In fact, a suggestion to proceed in this way was already contained in von Neumann and Morgenstern (1953), but the authors of that book had not followed it through. Nash assumes that a satisfactory solution of a game exists. He further assumes that the solution is unique. Hence, the solution takes the form of a (mixed) strategy profile σ, recommending to each player i a course of behavior σ_i. Rational players will know this solution and a player will be willing to behave according to it only if it is in their interest to do so, that is, if the strategy σ_i is a best response against the strategy profile of the opponents, $u_i(\sigma) = \max_{\tau i} u_i\ (\tau_i, \sigma_{-i})$. A Nash equilibrium is a profile of strategies σ such that this condition is satisfied for each player i. To summarize the argument, a necessary condition for σ to be a satisfactory solution of the game is that σ be a Nash equilibrium.

By relying on fixed point theorems (either the one of Brouwer or the one of Kakutani can be used), Nash proved that every finite game admits at least one Nash equilibrium, albeit possibly in mixed strategies. Indeed, in 'matching pennies' each player should choose each side of the coin with 50 percent probability in the unique equilibrium. In a 2-person zero-sum game, σ is a Nash equilibrium if and only if it is a pair of optimal strategies, hence, Nash's concept generalizes the solution provided by von Neumann and Morgenstern.

Pursuing the above rationale leading to Nash's solution concept, three questions remain: (a) why does a player represent their uncertainty about the behavior of their opponents by an independent mixed strategy profile? (b) do the requirements determine a unique solution? and (c) is being a Nash equilibrium sufficient to qualify as a satisfactory solution? The first issue is taken up in Aumann (1974) where an alternative concept of correlated equilibrium is developed that allows for more general beliefs. One interpretation is that players discuss before the game and construct a common randomization device, a correlated strategy σ. Upon hearing what pure strategy the device selected for player i, this player updates their beliefs on their opponent's actions; σ is a correlated equilibrium if each player is always willing to follow the recommendation of σ. The answer to the second question is a clear no, it is easy to construct games with multiple Nash equilibria and many games, also some with practical relevance, do have multiple equilibria. Hence, the rationale that has been given for the equilibrium concept appears incomplete, and the question is, whether, in a game with multiple equilibria there is any argument for focusing on any of these. Alternatively, should one look in an entirely different direction for a satisfactory solution, for example, by giving up the assumption that the solution be unique? We discuss these questions below after first having taken up the third question and having shown that that one also has a negative answer: not every Nash equilibrium can be considered as a satisfactory solution.

Consider the following extensive form ultimatum game. Player 1 divides \$10 in any (integer) way that they want; player 2 observes the division and decides whether or not to accept it; if player 2 accepts, each player receives the amount that was proposed, otherwise neither player receives anything. In addition, assume that each player cares only about the amount of money that they receive and that each prefers more money to less, say $u_i(x) = x_i$. Player 1 then knows that player 2 is sure to accept as long as they are offered at least \$1, hence, there seem only two possible solutions of the game—10; accept all—and—9, accept iff $x_2 \geqslant 1$ (the first number is the amount that player 1 asks for himself). Indeed these two strategy profiles are Nash equilibria, however, there are other equilibria as well, for example, 3, accept iff $x_2 \geqslant 7$ is a Nash equilibrium. (If player 1 demands 3 for themself, then player 2's strategy prescribes to accept and that is a best response; on the other hand, if player 2 indeed accepts only if offered at least 7, then it is optimal for player 1 to offer exactly 7.) In this equilibrium player 2 'threatens' to reject a positive amount that is offered to, such as \$5, even though strictly prefering to accept in case \$5 would be offered. The reason the profile is in

equilibrium is that an amount such as \$5 is not offered and, hence, the threat is not called. Player 1 behaves as if player 2 is committed to this strategy, in the extensive form game, however, such commitments are not possible and facing the *fait accompli* that \$5 is offered, player 2 will accept. Player 1 can neglect the incredible threat.

Starting with Selten (1965) a literature developed dealing with the question of how to eliminate equilibria involving incredible threats. Selten (1965) proposed the concept of subgame perfect equilibrium that is based on the idea of 'persistent rationality': everywhere in the game tree, no matter what happened in the past, the player that moves will play a strategy that, from that point on, is a best response against the strategies of the opponents. Formally, a subgame perfect equilibrium is a strategy profile that constitutes a Nash equilibrium in every subgame of the game, a subgame being a part of the game that constitutes a game in itself. For games with perfect information, subgame perfect equilibria can be found by the backward induction (dynamic programming) procedure that was used already in Zermelo (1913) to show that the game of chess is determined completely: starting at the end of the game one work backwards each time reducing the game by substituting a decision set of a player by an optimal decision and it's associated payoff. It is worthwhile to remark that this assumption of 'persistent rationality' was criticized in von Neumann and Morgenstern (1944).

Selten (1975) noted that requiring subgame perfection was not sufficient to eliminate all equilibria involving incredible threats and proposed a different solution. He reasoned that, if the problem is caused by nonmaximizing behavior at unreached information sets, it can be eliminated by ensuring that all information sets are reached, if only with small probability. Formally, Selten (1975) defines an equilibrium to be (trembling hand) perfect if it satisfies a robustness test: each player should be still willing to play their equilibrium strategy if each player with some small probability makes mistakes and, as a consequence, every information set is reached with positive probability. In the ultimatum game above, if player 1 makes mistakes, player 2 is forced to accept any positive amount and player 1 will not offer more than \$1 to player 2 in a perfect equilibrium. The robustness test proposed is weak (the strategies are only required to be optimal for some sequence of small mistakes that converges to zero) and this guarantees that each game admits at least one perfect equilibrium. However, at the same time this implies that some equilibria are perfect only because of the fact that the mistake sequence can be artificially chosen, hence, the perfectness concept does not eliminate all equilibria that are considered as 'counterintuitive.' The same remark applies to the closely related concept of sequential equilibria that has been proposed in Kreps and Wilson (1982). This concept requires that, at each information set, a player constructs beliefs over where they are in the set and optimizes their payoff, given these beliefs and taking the strategies of the opponents as fixed. Kreps and Wilson require that the beliefs be explainable in terms of small mistakes, hence, the close connection between their concept and Selten's.

Intuitively, one would like an equilibrium not to be robust against just some trembles but rather against 'plausible' ones: if there is a good explanation for why the information set is reached after all, then preference should be given to such an explanation. Plausibility is hard to define formally, but, if an equilibrium is robust against all trembles, then clearly it will also be robust against the plausible ones. Unfortunately, equilibria satisfying this strict robustness test need not exist. However, Kohlberg and Mertens (1984) have shown that sets of equilibria with this desirable property exist always. They, furthermore, argue that it is natural to look at sets of 'equivalent' equilibria. For example, if a player is fully indifferent between two of their strategies, there is no reason to make them choose a specific one. Similarly, if the choice only makes a difference if and after another player has played a strictly dominated strategy, one might be satisfied to leave the choice undetermined.

Sets of stable equilibria satisfy several desirable properties, such as being robust against elimination of strategies that are dominated or that are not a best response against any element of the set. The latter property is also called 'forward induction' and it has proved powerful to reduce the number of equilibrium outcomes in signaling games (see *Information, Economics of*). The original stability concept proposed in Kohlberg and Mertens (1984) was of a preliminary nature and was not completely satisfactory. The ideas have been further developed in Mertens (1989) but that concept has not seen too many applications yet as it is difficult to handle. In addition, part of the profession considers this a too demanding rationality concept (see van Damme 1991 for further discussion).

The normative, 'rationalistic' interpretation of Nash equilibrium that has been considered thus far relies on the assumption that the solution to a game is unique, yet many games admit multiple equilibria, even multiple stable ones. The question thus arises about equilibrium selection: how will players coordinate on an equilibrium and on which one? Consider the following simple normal form stag hunt game as an example. Two players simultaneously choose between the numbers 1 and 2. If both choose the same number, they each receive a payoff equal to the number chosen, if they choose different numbers the player choosing the lower number receives 1 from the other. The game has two equilibria in pure strategies: (1,1) and (2,2). The latter yields both players higher payoffs than the former, however, choosing 2 is also more risky than choosing 1: while the latter guarantees the payoff 1, the former might result in a loss of 1. How to resolve the problem?

Harsanyi and Selten (1988) provide a general theory of equilibrium selection in noncooperative games that can be used in these circumstances. The theory can be seen as an extension of Nash's (1953) equilibrium selection theory for bargaining games. The authors start by formulating general properties that a theory should satisfy (such as symmetry, ordinality, efficiency, monotonicity, consistency) but they discover quickly that not all of these can be satisfied at the same time. Indeed, the stag hunt game given in the above paragraph has the same best reply structure as the game in which payoffs are 2 (resp. 1) if each player chooses 1 (resp. 2) and where payoffs are 0 if different numbers are chosen, and in this game the efficient equilibrium involves both players choosing 1. Hence, choices have to be made and, as Harsanyi and Selten admit, different selection theories are possible. One of the major innovations of Harsanyi and Selten (that in some form will probably play a role in these alternative theories as well) is the concept of risk dominance. Intuitively, one equilibrium is said to risk dominate another if, in a situation where attention is confined to this pair, players eventually come to coordinate on the first as they consider the second to be more risky. The formal definition makes use of the tracing procedure, an algorithm (homotopy) that transforms any mixed strategy profile (which represents the players' initial assessment about how the game will be played) into an equilibrium of the game. This tracing procedure is supposed to capture the reasoning process of rational players in the game. If, in a situation of uncertainty about equilibria σ and σ, the tracing procedure produces equilibrium σ, then σ is said to risk dominate σ. The concept allows a simple characterization for 2-player 2×2 games with 2 strict equilibria. Let $p_i(\sigma,\sigma')$ be the probability that player i has to assign to σ_i in order to make their opponent j indifferent between σ_j and $\sigma'_j =$. Intuitively, the larger $p_i(\sigma,\sigma')$, the more attractive $\sigma_j =$ is for player j. In this case, σ risk dominates $\sigma =$ if and only if $p_1(\sigma, \sigma') + p_2(\sigma, \sigma') < 1$. In the above example, the equilibrium (1,1) risk dominates the equilibrium (2,2).

3.3 Evolution and Learning

The above discussion of Nash equilibrium (and its variants) has been based entirely on the rationalistic interpretation of this concept, viz. that a self-enforcing theory advising players how to play should prescribe a Nash equilibrium in every game. Nash (1950), however, provided also a second interpretation of his concept. When the game is played repeatedly by myopic players who best respond against the current situation, then, if a stable state is reached, that state has to be a Nash equilibrium. In this second interpretation it is thus unnecessary to assume that players know the full structure of the game, or are able to perform complex reasoning processes. On the contrary, the interpretation applies in a context of limited information where players behave in a boundedly rational way. After Shapley (1964) had published an example showing that a certain dynamic adjustment process need not converge at all, interest in this second approach dwindled for some time, but it was revived in the 1990s. Since then a systematic study has begun under which circumstances a variety of processes will converge and, if they converge, to which equilibria. It is too early to draw broad conclusions, instead the reader is referred to Fudenberg and Levine (1998) for a partial overview. Interestingly, some of the concepts that were introduced in the 'rationalistic' branch reappear in this branch of the literature, in particular this holds for the concept of risk dominance.

Consider the stag hunt coordination game described above and imagine that it is played repeatedly. When a player is called upon to play they 'ask around' to find out what other players in the population that have already played have been confronted with and then plays a best response against the resulting sample. Young (1998) shows that the process will converge to one of the two equilibria. Now imagine that sometimes players make mistakes in implementing their desired action and choose the other one instead. In this case the process may move from one equilibrium to the other through a sequence of consecutive mistakes. Young (1998), however, shows that, if the probability of making mistakes tends to zero, the limit distribution will put all mass on the risk dominant equilibrium as that equilibrium is more robust to mistakes than the other one. Hence, in the long run when mistake probabilities are negligible the process will end up in the risk dominant equilibrium, irrespective where it starts out. The process not only produces an equilibrium, it leads to a very specific one. Research effort is under way to investigate the robustness of this result with respect to both the underlying game and the dynamic process under consideration.

Related to the above is the application of the Nash equilibrium concept in biology. As Maynard Smith (1982, p. vii) writes 'Paradoxically, it has turned out that game theory is more readily applied to biology than to the field of economic behavior for which it was originally designed.' In biological applications the concept of utility is replaced by that of Darwinian fitness (expected number of offspring), human rationality is replaced by evolutionary stability and, according to Maynard Smith, both contribute to making the theory more applicable. Note that in this context it is assumed that there is no conscious choice of strategy at all; strategies exist and individuals are 'programmed' to play certain strategies, an evolutionary process selects those strategies that have highest fitness and in the end an equilibrium results.

To illustrate, one may consider the so called Hawk-Dove game in which two animals are contesting a certain valuable resource. There is a (finite) repertoire of possible behaviors, S, and when an individual

programmed to play strategy s meets another individual playing strategy s' the gain in fitness to the first is given by $u(s, s')$. The situation may then be represented by a game in which there are 2 players, and each has strategy set S and payoff function u, that is, the game is symmetric. The question is which strategy will be selected, that is, which one will survive in the long run. Obviously, if $u(s', s) > u(s, s)$ then s' can displace s so that a necessary condition for survival of s is that (s, s) constitutes a Nash equilibrium of the game. A second necessary condition for selection of s is that, once s has been established, random mutations that enter into the population are selected against, that is, these are driven out again. This will be the case if they cannot spread, that is, if the entrant s' is an alternative best response against s, then the incumbent strategy obtains a higher payoff against the mutant than the mutant gets against itself (if $u(s', s) = u(s, s)$, then $u(s, s') > u(s', s')$). A strategy s satisfying both conditions is called an evolutionarily stable strategy or ESS. One sees that an ESS corresponds to a Nash equilibrium satisfying an additional stability requirement (see Hammerstein and Selten 1994 for further discussion).

4. Behavior

Game theory, at least its traditional variant, relies on the assumption of rational behavior, that is, behavior that is consistent and is motivated solely by the desire to obtain well-defined goals, with, furthermore, no cognitive constraints standing in the way. In contrast, the biological branch of the theory relies on no rationality at all, but assumes long time periods for selection to work. The question, hence, is how the theory is relevant for humans that are boundedly rational and have relatively little time to learn and adjust behavior. Real players have only limited rationality: they face cognitive limitations which may make it impossible to get a full overview of the situation and to consistently evaluate and rank all alternatives, they may decide not to think and instead rely on automated procedures that have proved satisfactory in the past or in other related contexts, and even if rational deliberation suggests a decision, emotional factors may override and produce an alternative decision.

The question, hence, is to what extent a theory that is based on strong rationality assumptions is relevant for actual interactive decision making. For sure, a normative solution can serve as a benchmark to better understand actual behavior, it might point to systematic differences and it might even be that the differences are negligible in certain contexts, such as when players have had sufficient time to familiarize themselves with the situation. Unfortunately, a discussion of the issue of relevance is made difficult by the fact that alternative theories are still undeveloped and since relatively little is known yet about the actual structure of human decision making (see Selten 1991). Nevertheless, some systematic deviations from the theory that are found in the experimental laboratory may be described (see Camerer 1997 for more details).

First of all, a remark that does not apply that much to the solution concepts, but rather to the modeling aspect. Game theory analyses the consequences of selfish behavior, where selfishness is interpreted as the individual following their preferences. However, it is not assumed necessarily that individuals are selfish in the more narrow sense of being only interested in their own material well-being. The latter assumption would indeed be too narrow. Experiments have shown that, in the ultimatum game referred to above, responding players generally dislike being treated 'unfairly,' hence, they are willing to reject proposals in which they get positive amounts but the distribution is uneven. A relevant game model of the situation should take this into account and as a consequence it may well advise to player 1 to allocate a substantial proportion of the cake to player 2 (see Kagel and Roth 1995 for further discussion on this issue). Suffice it to note here that tests of game theoretic predictions are always combined tests of the underlying game and the solution concept applied, that it is difficult to control the players' preferences, hence, that one should be careful in drawing strong inferences from results observed in the experimental laboratory.

Second, when confronted with an interactive decision situation, real-life players typically construct a simplified model of the situation. The 'mental model' may be an oversimplification and, as a consequence, important strategic aspects may be missed. Furthermore, a player needs to take into account the models constructed by other players: if these do not incorporate certain elements, it does not make sense for the player to signal along these dimensions. Selten (1998) argues that the reliance on superficial analysis may explain presentation or framing effects, that is, the way in which a game is presented to the players may have a strong influence on actual behavior. The basic reason is that the presentation may trigger a certain reasoning process or may make some outcomes more focal than others. Related is the fact that human players do not analyze each game in isolation but rather make use of analogies; a principle that has proved useful in a certain context may also be used in another one. In experiments, for example, it is observed that 2-player games with sequential moves in which the second mover receives no information about the first move may be played as if there was such information, where according to the theory the game is strategically equivalent to the game with simultaneous moves. The discussion of the presentation effect in strategic contexts originates with Schelling (1963), but his concept of 'focal points' still largely awaits formalization.

Where standard game theory assumes that all players are equally rational and, hence, homogeneous, experiments have revealed considerable player het-

erogeneity. Some players are more of the adaptive type and do not approach a game analytically. Instead, in repeated contexts, they make use of *ex post* rationality considerations and move in the direction of better responses against the previous outcome. Players who rely on analytic approaches construct simplified models, as discussed above, and they may limit the 'depth' of their analysis to a couple of steps and, hence, may not reach a game theoretic equilibrium. For example, in the 'beauty contest game' in which a set of players choose numbers between 1 and 100 and in which the winner is the one that chooses the number closest to half of the average of all numbers, the equilibrium is to choose 1 and that equilibrium is reached through iterative elimination of dominated strategies. However, most players choose larger numbers and those that choose the equilibrium number do not win the game. Relatedly, in extensive form games players do not necessarily 'look ahead and reason backwards,' but instead may use a myopic somewhat forward looking process and, as a consequence, the outcome obtained may be different from the subgame perfect equilibrium one.

At present, research is under way to construct an empirically-based behavioral game theory that steers a middle ground between over-rational equilibrium analyses and under-rational adaptive analyses, but it is too early to survey this field (see *Behavioral Economics*; *Experimental Economics*).

5. Conclusion

The space allotted has not allowed a discussion of any applications in depth, however, as Aumann (1987) has argued forcefully, the models and solution concepts from game theory should be judged by the insights that they yield. In addition to the entries of this Encyclopedia already referred to above, we may point to the areas of finance (market microstructure), industrial organization, and antitrust and regulation, in which ideas from noncooperative game theory have been influential and are used widely.

Bibliography

Aumann R J 1974 Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* **1**: 67–96

Aumann R J 1987 Game theory. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave Dictionary of Economics*. Macmillan, London, pp. 460–82

Aumann R J 1985 What is game theory trying to accomplish? In: Arrow K, Honkapohja S (eds.) *Frontiers of Economics*. Blacksburg, VA, pp. 28–100

Baumol W J, Goldfeld S 1968 *Precursors in Mathematical Economics: An Anthology*. London School of Economics and Political Science, London

Bernheim B D 1984 Rationalizable strategic behavior. *Econometrica* **52**: 1007–28

Camerer C 1997 Progress in behavioral game theory. *Journal of Economic Perspectives* **11**: 167–88

Damme E E C van 1991 *Stability and Perfection of Nash Equilibria*, 2nd edn. Springer Verlag, Berlin

Fudenberg D, Levine D 1998 *The Theory of Learning in Games*. MIT, Cambridge, MA

Hammerstein P, Selten R 1994 Game theory and evolutionary biology. In: Aumann R J, Hart S (eds.) *Handbook of Game Theory*, Vol. II. Elsevier, Amsterdam

Harsanyi J 1967–1968 Games with incomplete information played by 'Bayesian' players, parts I, II and III. *Management Science* **14**: 159–82, 320–34, 486–502

Harsanyi J, Selten R 1988 *A General Theory of Equilibrium Selection in Games*. MIT, Cambridge, MA

Kagel J, Roth A 1995 *Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ

Kohlberg E, Mertens J F 1986 On the strategic stability of equilibria. *Econometrica* **54**: 1003–39

Kreps D, Wilson R 1982 Sequential equilibria. *Econometrica* **50**: 863–94

Kuhn H 1953 Extensive games and the problem of information. In: Kuhn H, Tucker A W (eds.) *Contributions to the Theory of Games II*. Princeton University Press, Princeton, NJ, pp. 193–216

Maynard Smith J 1982 *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK

Mertens J F 1989 Stable equilibrium: A reformation. *Mathematics of Operations Research* **14**: 575–625

Nash J (1950) Non-cooperative games. Ph.D. thesis, Princeton University

Nash J 1951 Non-cooperative games. *Annals of Mathematics* **54**: 298–95

Nash J 1953 Two-person cooperative games. *Econometrica* **21**: 128–40

Neumann J von 1928 Zur theorie der gesellschaftspielen. *Mathematische Annalen* **100**: 295–320

Neumann J von, Morgenstern O 1944/1953 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

Pearce D G 1984 Rationalizable strategic behavior and the problem of perfection. *Econometrica* **52**: 1029–50

Rubinstein A 1998 *Modeling Bounded Rationality*. MIT, Cambridge, MA

Schelling T 1963 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA

Selten R 1965 Spieltheoretische behandlung eines oligopolmodells mit nachfragetragheit. *Zeitschrift für die gesamte Staatswissenschaft* **121**: 301–24, 667–89

Selten R 1975 Re-examination of the perfectness concept for extensive form games. *International Journal of Game Theory* **4**: 25–55

Selten R 1991 Evolution, learning and economic behavior. *Games and Economic Behavior* **3**: 3–24

Selten R 1998 Features of experimentally observed bounded rationality. *European Economic Review* **42**: 413–36

Shapley L 1953 Some topics in 2-person games. In: Kuhn H, Tucker A (eds.) *Contributions to the Theory of Games II*. Princeton University Press, Princeton, NJ, pp. 305–17

Weintraub E R 1992 *Toward a History of Game Theory*. Duke University Press, Durham, NC

Young P 1998 *Individual Strategy and Social Structure*. Princeton University Press, Princeton, NJ

Zermelo E 1913 Über eine anwendung der mengenlehre auf die theorie des schachspiel. In: *Proc. 5th Int. Congress of Mathematicians.* Vol. II, pp. 501–4

E. van Damme

Gangs, Sociology of

The term, 'gang' is both a theoretical construct and an object of varying definitions in legal statutes, legal and social agency policy, and common discourse. Although the term defies precise measurement, it is the subject of a large scholarly and popular literature.

1. Historical Background of Scholarly Research on Gangs

'Gang' and other terms (mob, syndicate, outfit, etc.) have been applied to many types of groups, including those associated with organized and professional crime and incarcerated felons. Scholarly attention, however, has focused primarily on youth gangs. Frederic Thrasher's *The Gang: A Study of 1,313 Gangs in Chicago* (1927) was the first attempt to survey the extent of youth gang activity in a major city, and perhaps the first in any jurisdiction. The project involved census and court records, personal observations, and personal documents collected from gang boys and from persons who had observed gangs in many contexts. Studies by Clifford R. Shaw and his collaborators provided even more information about the collective nature of youthful delinquency, offering consistent documentation that most boys who were brought before the juvenile court committed their delinquent acts in the company of others. Case studies documented patterns of friendship, the association of younger with older offenders, and the influence of organized crime and other forms of adult criminality in communities with high rates of juvenile delinquency (Shaw et al. 1929, Shaw 1930, Shaw and McKay 1931, Shaw and Moore 1931). Ecological studies located delinquency in space and in relationship to urban development, documenting economic and institutional contexts within which urban lives were lived, and identifying forces that shaped the ability of communities to aid in the socialization of children and exercise control over misbehavior.

Theoretical insights from these studies continue to be important in thinking about youth crime. Foreshadowing 'labeling theory,' for example, Shaw's Jack-Roller—reflecting on his confinement in a reformatory—noted that he was no longer 'just a mischievous lad, a poor city waif, a petty thief, a habitual runaway,' but 'a criminal' (Shaw 1930, p. 103). Conversely, both Shaw and Thrasher noted that incarceration and criminal notoriety often had positive value among gang members, as it does today: '… residents in the vicinity south of the stock yards were startled one morning by a number of placards bearing the inscription 'The Murderers, 10,000 Strong, 48th and Ada.' In this way attention was attracted to a gang of thirty Polish boys, who hang out in a district known as the Bush' (Thrasher 1927, pp. 62–3).

Thrasher's Murderers were involved in a good deal of criminal activity, but their primary pastimes 'were loafing, smoking, chewing, crap-shooting, card-playing, pool, and bowling.' Their 'rudeness and thievery' were 'an awful nuisance' to local shop keepers and neighbors (Thrasher 1927, pp. 62–3). The arcane language notwithstanding, these excerpts highlight similar patterns among contemporary and earlier gangs involving 'hanging out,' minor and more serious criminal behavior.

The most important impact of earlier gang research was its contribution to social disorganization theory. Youth gangs were found overwhelmingly in communities with high rates of crime and delinquency, poverty, and population heterogeneity and turnover. Social disorganization theory, hypothesizing that such communities lack both effective institutions and local informal means of control, continues to be refined by recent research. Thus, Sampson and his colleagues (2000) find that intergenerational closure (the linkage between adults and children), reciprocal local exchange (interfamily and adult interaction with respect to children), and expectations that community residents will share child control responsibilities, are associated with the ability of local communities to exert effective control over violent and other forms of criminal behavior.

Midway between the publication of Thrasher's classic work and the turn of the century, the book was abridged and reissued. Social changes had vastly altered the gang landscape, and gang research, heretofore largely descriptive, was changing in response to seminal theoretical proposals by Cohen (1955), Cloward and Ohlin (1960), and Miller (1958). Competing explanations for the origins of the delinquent subculture (Cohen), variations in delinquent subcultures (Cloward and Ohlin), and the role of lower-class culture in producing gang delinquency (Miller) stimulated a large body of empirical research and subsequent theorizing.

By the last decade of the twentieth century, however, events that could hardly have been anticipated by earlier researchers had overtaken both theory and research. Whereas midcentury theories had outstripped available data, modest theoretical advances and related empirical research were overwhelmed by the rapid proliferation of gangs in the USA and their spread to many other countries (Klein 1995, Klein et al. 2000, Moore and Terrett 1999). In the USA sophisticated firearms became more readily available to young people (Blumstein 1995), often turning what previously had been 'non-zero-sum' contests between gangs into lethal confrontations (Short and

Strodtbeck 1965, Anderson 1999). Weakened conventional economic, political, and family institutions lost legitimacy, contributing to rising crime rates (LaFree 1998); and the technological revolution in electronic communication and entertainment media created heretofore unknown problems for institutions of social control.

2. *Change and Continuity*

It is possible, nevertheless, to note both changes and continuities in knowledge concerning the collective forms taken by the delinquent and criminal behavior engaged in by the young.

(a) Youth gangs are more prevalent in many countries than previously. Law enforcement agencies in all large US cities (over 250,000) reported the presence of youth gangs, as did many suburban counties, small cities and rural counties, where gang prevalence is quite recent (National Youth Gang Center 1998). European and UK researchers have identified variations in youth gang organization and behavior, some of them patterned after US gangs (Klein et al. 2000).

(b) Gangs in the US, Continental Europe, and the UK are comprised predominantly of youth who are marginalized in mainstream society by virtue of race, ethnicity, or economic status. In the USA these are predominantly African-American and Hispanic, but with increasing numbers of new Asian immigrant children and children of immigrant parents. In Continental Europe and the UK they tend to be the children of refugees from other countries or migrant workers. Gangs comprised of majority population youth, nearly always white, sometimes form in opposition to these marginalized groups, thus becoming marginalized themselves.

Youth gangs are comprised overwhelmingly of young males, and the experience of females who associate with male-dominated gangs, or in all-female gangs, is quite different from the gang experience of males (Chesney-Lind and Hagedorn 1999).

(c) The distribution and patterning of collective youth crime reflect the ecology of the places where they are located. In the USA and many other cities, these are inner-city areas, much as they were in Thrasher's and Shaw's day. Where public housing projects are located outside central cities, gangs of loosely structured, ethnically mixed groups often 'commute' to central city areas where they engage in carousing, 'smash-and-grab' hits on stores, rolling drunks, and robberies.

Large aggregations of gang 'nations' emerged in a few large US cities, most notably Chicago and Los Angeles, during the 1960s. Local gangs—wherever they are found—are primarily 'home grown,' but the influence of gang nations is widespread through mimicry and migration of gang members. Many—probably most, systematic data are lacking—gangs begin as nondelinquent, friendship groups, 'hanging' groups, or groups formed around particular nondelinquent activities (sports, singing, dancing, etc.) (see, e.g., Schwendinger and Schwendinger 1985, Hagedorn 1988, Sullivan 1989, Schwartz 1987, Short 1997). The primary distinguishing characteristic of all youth gangs is that they are self-determining with respect to membership criteria, group structure and behavior, and—most importantly—they are unsupervised by adults.

(d) A variety of studies suggest that such unsupervised youth groups become violent, and otherwise delinquent, as a result of one or more of the following processes:

(i) escalation of the natural rough-and-tumble punching and wrestling that occurs among most male groups, and the association of status with fighting prowess;

(ii) competition with rival groups involving status-enhancing activities or economic motivations;

(iii) conflicts based on racial/ethnic differences;

(iv) definitions of groups as gangs, by others, and behavior toward them that pushes a violent identity on the group;

(v) group processes that create or reinforce group cohesion based on violent or other delinquent behavior.

Street gangs are diverse in behavior, organization, continuity across generations, and in their relationships with other gangs and with their communities. They change, as members age and move into adulthood. The age range of many gangs has expanded in many countries, in response to the lack of attractive legitimate jobs. This change often is accompanied by increased violence, as gang members have turned to selling drugs and other illegal activities in order to support themselves, and often their families (see Fagan 1996, Fagan and Wilkinson 1998, Hagedorn 1998).

These changes tend to blur distinctions between street gangs and a variety of types of youth and young adult collectivities, such as 'wilding' groups, 'skinheads,' and perhaps the mass shootings that have occurred on school properties. Although differing among themselves, these groups have in common their hate of 'outsiders,' based on racial, ethnic, or other exclusionary criteria, as, also, do crowds that gather to attack community outsiders, usually members of minority groups. Violence and other delinquent activities often occur, also, in 'staging' areas in which young people gather with friends and others simply in order to escape adult supervision (Schwartz 1987, Anderson 1999).

Drug and other 'working' gangs are organized primarily around revenue-producing activities, such as drug selling, stealing, fencing stolen goods—all activities requiring greater organizational skills and structure than is characteristic of traditional street gangs. Also unlike street gangs, 'bikers,' 'punkers,'

and sports fanatics and 'hooligans' are organized around their specialized interests. Except for the latter (most of whom appear to be disadvantaged socially, rather than economically), these groups are comprised primarily of youth who have been marginalized by minority ethnic and/or economic status, and who have poor prospects for integration or social mobility within dominant majority societies. The excessive behavior of bikers and sports hooligans also appears to be motivated in some measure by resentment of minorities.

3. *Youth Culture*

Virtually all youth are heavily influenced by the development, rapid growth, and diffusion of youth culture. Peddled by media advertising, and augmented by large social and economic forces that fail to provide meaningful roles for young people but cater to their distinctive fads, appetites, and currency, youth culture and gang culture together underly the excesses of consumption and violence perpetrated by gangs.

Youth culture developed rapidly during the 1960s, especially in the USA. Young people born during the 'baby boom' following World War II shared in the general level of affluence in the country, thereby creating a youth market with great economic power; the period of education prior to entering the labor market was extended; increasing numbers of women joined the workforce, further separating mothers from youth in the home and in the neighborhood; increasingly, adults worked in large organizations away from young people; and the mass media greatly expanded and focused their attention on the youth market (see Coleman et al. 1974, pp. 114–19). These changes accelerated as the century wore on, expanding their influence throughout the world.

Field researchers document the pernicious seductions of media-advertised products among less affluent young people, leading often to thefts and assaults. Anderson (1999) describes the 'zero-sum quality' that pervades the 'code of the streets,' in the search for respect—often associated with items of clothing such as shoes and jackets—among young African-American men in the ghetto. Sullivan identifies 'mutually valorizing cultural symbols' (clothing, drugs, alcohol, and recreational artifacts and activities) that cut across racial and ethnic boundaries, often resulting in invidious comparison, confrontation, theft, and assault (Sullivan 1989, pp. 248–9).

Observers of youth groups everywhere note that status differences within and between them are extremely variable and highly refined. Status criteria within a group, or between one group and another, may be based on race or ethnicity, relative economic affluence, skills in valued activities, public appearance, school performance, or most importantly, on lifestyle differences. These bases of stratification typically also become criteria of inclusiveness and exclusiveness which, in turn, create opportunities for both friendship and rejection.

Adolescence is a period of especially intense identity formation, intense relationships, and shared feelings of friendship, acceptance, and respect. Conversely, feelings of rejection and disrespect also are especially intense, and often the basis for group and subcultural formation.

Next to families, schools are perhaps the most important contexts for adolescent friendship, achievement, and recognition. Because of this, school contexts (including the journey to and from schools) are the settings for much adolescent behavior, including delinquent and criminal behavior. On occasion they have also been the setting for the most extreme forms of violence (e.g., mass killings by students or others alienated from their fellows or from mainstream institutions in general). Although the specific causes of such extreme alienation are complex, it seems clear that schools have been especially targeted precisely because of their importance in the lives of adolescents, as symbols of rejection by both peers and an adult world that seems far removed from adolescent concerns. The ready availability of guns at times transforms normal adolescent turmoil and conflict into deadly confrontation.

More than ever before in history, young people, targeted for commercial exploitation and isolated from mainstream adult roles and institutions, confront economic conditions beyond their control. Economic decline, severe unemployment, and the unavailability of 'good jobs,' are associated not only with street gangs, but with their transformation into 'economic gangs' (including drug gangs), and with ethnic, racial, and class-related identities and antagonisms that lead to other types of collective violence (Hagedorn 1998 Pitts 2000). These same forces alter both intergang relationships and relationships between gangs and their communities. Relationships between adults and young people, shaped by generational backgrounds and community cultures, 'translate' these and other macrolevel forces at the local community level.

4. *Research Needs*

Research and theory concerning gangs, and effective means of controlling the violence and crime associated with gangs, are hampered by the lack of a viable typology of youth groups and groups comprised of young adults that would situate the many observed variations of gangs and gang-like groups within even broader collective forms of behavior. Comparative research in many countries suggests that distinctions among and between such groups vary a great deal, and that the meaning and the function of behavioral variations lack universality.

Varying definitions, meanings, and behaviors have consequences. To the extent that gangs and gang membership confer 'master status' on groups and individuals—among law enforcement personnel, in media portrayals, and among the general public—they contribute to the deterioration of relationships between economic, racial, and ethnic groupings and between the generations. Reification and demonization of gangs, by media, law enforcement, the public, and occasionally by scholars, for example, may contribute to gang problems rather than to their solution.

See also: Control: Social; Crime, Sociology of; Delinquency, Sociology of; Norms; Violence: Public; Youth Gangs

Bibliography

Anderson E 1999 *Code of the Streets: Decency, Violence, and the Moral Life of the Inner City*. W.W. Norton, New York

Blumstein A 1995 Violence by young people: Why the deadly nexus? *National Institute of Justice Journal* **August**: 2–9

Chesney-Lind M, Hagedorn M (eds.) 1999 *Female Gangs in America*. Lake View Press, Chicago

Cloward R A, Ohlin L E 1960 *Delinquency and Opportunity*. Free Press, New York

Cohen A K 1955 *Delinquent Boys*. Free Press, New York

Coleman J S, Bremner R H, Clark B R, Davis J B, Eichorn D H, Griliches Z, Kett J F, Ryder N B, Doering Z B, Mays J M 1974 *Youth: Transition to Adulthood. Report of the Panel of the President's Science Advisory Committee*. University of Chicago Press, Chicago

Fagan J 1996 Gangs, drugs, and community change. In: Huff C R (ed.) *Gangs in America*, 2nd edn. Sage, Thousand Oaks, CA, pp. 39–74

Fagan J, Wilkinson D L 1998 Guns, youth violence, and social identity in inner cities. In: Tonry M, Moore M (eds.) *Youth Violence. Crime and Justice*. University of Chicago Press, Chicago, Vol. 24, pp. 105–88

Hagedorn J M 1988 *People and Folks: Gangs, Crime, and the Underclass in a Rustbelt City*. Lake View Press, Chicago

Hagedorn J M 1998 Gang violence in the postindustrial era. In: Tonry M, Moore M (eds.) *Youth Violence. Crime and Justice*. University of Chicago Press, Chicago, Vol. 24, pp. 365–419

Klein M 1995 *The American Street Gang*. Oxford University Press, UK

Klein M, Kerner H-J, Maxson C L, Weitekamp E G M (eds.) 2000 *The Eurogang Paradox: Street Gangs and Youth Groups in the US and Europe*. Kluwer-Plenum, Amsterdam

LaFree G 1998 *Losing Legitimacy: Street Crime and the Decline of Social Institutions in America*. Westview Press, Boulder, CO

Miller W B 1958 Lower class culture as a generating milieu of gang delinquency. *Journal of Social Issues* **23**(3): 5–19

Moore J P, Terrett C P 1999 *Highlights of the 1997 National Youth Gang Survey*. OJJDP Fact Sheet, No. 97. Office of Juvenile Justice and Delinquency Prevention, Washington, DC

National Youth Gang Center 1998 *1996 National Youth Gang Survey*. Office of Juvenile Justice and Delinquency Prevention, Washington, DC

Pitts J 2000 *Discipline or Solidarity: The New Politics of Youth Crime*. Macmillan, Basingstoke, UK

Sampson R J, Morenoff J D, Earls F 2000 Beyond social capital: Structural sources and spatial embeddedness of collective efficacy for children. *American Sociological Review* **64**: 633–60

Schwartz G 1987 *Beyond Conformity or Rebellion: Youth and Authority in America*. University of Chicago Press, Chicago

Schwendinger H, Schwendinger J 1985 *Adolescent Subcultures and Delinquency*. Praeger, New York

Shaw C R 1930 *The Jack-Roller: A Delinquent Boy's Own Story*. University of Chicago Press, Chicago

Shaw C R, McKay H D 1931 *Report on the Causes of Crime*. National Commission on Law Observance and Enforcement Report, Vol. 2. USGPO, Washington, DC

Shaw C R, McKay H D [1942] 1969 *Juvenile Delinquency and Urban Areas*. University of Chicago Press, Chicago

Shaw C R, Moore M E 1931 *The Natural History of a Delinquent Career*. University of Chicago Press, Chicago

Shaw C R, Zorbaugh F, McKay H D, Cottrell L S 1929 *Delinquency Areas*. University of Chicago Press, Chicago

Short J F Jr 1997 *Poverty, Ethnicity, and Violent Crime*. Westview Press, Boulder, CO

Short J F Jr, Strodtbeck F L 1965 *Group Process and Gang Delinquency*. University of Chicago Press, Chicago

Sullivan M 1989 *'Getting Paid': Youth Crime and Work in the Inner City*. Cornell University Press, Ithaca, NY

Thrasher F M 1927 *The Gang: A Study of 1,313 Gangs in Chicago*. University of Chicago Press, Chicago

J. F. Short Jr.

Garbage Can Model of Behavior

The garbage can (GC) model is a framework for analyzing decision making in 'organized anarchies'—organizations characterized by problematic preferences, unclear technologies, and fluid participation (Cohen et al. 1972). The model conceives of organizations as conglomerates of semiautonomous decision arenas or 'garbage cans' through which problems and solutions flow as independent streams. Decision outcomes are sensitive to the precise mix of problems and solutions represented in a garbage can at the moment of decision. This mix, in turn, depends on the number of decision arenas, the structure of access to them, the overall organizational load of problems and solutions, and the allocation of energy and attention across arenas.

1. A Challenge to the Rational Model of Organization

The GC model may be understood as an important step in a developing critique of the rational model of organization by the so-called 'Carnegie School.' This school builds on Herbert Simon's treatment of organizations as information-processing systems and his argument about cognitive limits on individual

information-processing (see *Bounded Rationality*). Simon (1985) argued that just as individuals could partially overcome their cognitive limits by decomposing complex decisions into simpler choices and routines, so too could organizations decompose complex collective decision making by means of nested hierarchical structures ('nearly decomposable systems'). The subsystems of these nested hierarchies represented relatively autonomous bundles of routines. While Simon's analysis offers a critique of the rational model, he ultimately seeks to demonstrate that problem decomposition and routinization permit cognitively limited individuals and organizations to approximate, if only imperfectly, rational organization. The GC model, in contrast, poses a more fundamental challenge to the rational model.

The GC model goes beyond Simon's analysis by focusing on the ambiguity and loosely coupled nature of individual and organizational decision making. Whereas Simon's model of bounded rationality emphasizes cognitive limits on evaluating alternative courses of action, the GC model argues that the ambiguity of decision making stems from the ill-defined and inconsistent nature of individual and organizational goals (problematic preferences), and the interdeterminacy of knowledge and of techniques for achieving goals (unclear technologies). The GC model also problematizes Simon's notion of organizations as bundles of semiautonomous subsystems. The metaphor of the organization as a collection of 'garbage cans' implies a set of semiautonomous decision-making arenas where problems are neither so neatly decomposed nor so well integrated as conceptualized by Simon. Furthermore, the GC model argues that participation in these decision arenas is unstable rather than routine (fluid participation). In making these arguments, the GC model incorporates and extends many of the earlier arguments of the Carnegie school, including insights into the political and symbolic character of decision making, the coalitional character of organization, the sequential nature of decision making, the role of organizational slack, and the tendency of subunits to develop narrow and parochial decision premises.

In seeking to turn the rational model 'on its head,' the authors of the garbage can model sometimes state their alternative assumptions in fanciful ways. For example, they argue that organized anarchies can be understood as 'collections of choices looking for problems, issues and feelings looking for decision situations in which they might be aired, solutions looking for issues to which they might answer, and decisionmakers looking for work' (Cohen et al. 1972, p. 1). This ironic tone may be read as either usefully provocative or as a hindrance to understanding. Stated in the extreme, the assumptions of the model often appear as unrealistic as the assumptions of the rational model. The argument that problems are decoupled from solutions and flow through the organization as independent streams is difficult to grasp, because it is intuitive to think that the people who formulate problems are typically the ones best positioned to devise solutions to them. Yet if we were to dismiss the GC model because of its implausibility *in extremis*, we would also miss its powerful insights. Shorn of its irony, it is easy to recognize situations in which people become advocates for certain types of solutions and promote them whenever and wherever they can, often without clearly understanding the problems which are being addressed. It is easy to see that people often bring issues and concerns to a decision arena extraneous to the ostensible decision at hand. And it is obvious that decision outcomes are highly sensitive to the mix of issues brought to bear in any decision arena, that this mix is often highly dependent on who participates, and on the range of other decisions being made at any given time. One value of the GC model is that it treats these very intuitive and commonplace circumstances seriously.

2. *The Simulation of Chaotic Decision Making*

Part of the purpose of stating their assumptions in exaggerated form is that it allows the authors (brilliantly) to transform their analysis of the essentially *ad hoc* nature organizational behavior into a systemic analysis of organizational decision making. They do this by linking together the following ideas. As already mentioned, the model emphasizes the *loose coupling* or *decoupling* of intentions, actions, and constraints. It argues that individual intentions are not tightly coupled to individual action, that collective goals are not tightly coupled to collective action, and that organizations are not tightly coupled to their environments. By conceptualizing problems and solutions as flowing independently through organizations, the GC model imagines organizations to be *demographic systems* in which problems and solutions (via their attachment to individuals) enter and exit particular decision-making arenas. Decisions or nondecisions can then be understood as dependent upon the composition of problems and solutions in a decision arena at any point in time. At the systemic level, this composition is expected to be very sensitive to the issues to which actors happen to be paying attention at any given moment (attention is treated as a scarce resource) and to the structure of access into decision-making arenas. The model is *processual* in that it suggests that outcomes are not easily predicted from starting conditions, but depend rather heavily on the precise interactions between elements of the system. In general, processual theories suggest that interactions are complex, nonlinear, and characterized by the prevalence of unintended consequences. Taken together, this characterization of organized anarchies as loosely coupled, demographic, and processual allows the authors to treat decision making as a *stochastic*

process, introducing elements of probability and randomness into the model. Beginning with the original paper of Cohen et al. (1972), the GC model has encouraged a strategy of simulation modeling in order to capture the dynamics of these stochastic, demographic processes (e.g., Padgett 1980, Carley in March and Weissinger-Baylon 1986).

3. *The Garbage Can Model and Political Institutionalism*

If organizational decision making is often disorderly and fraught with ambiguity, we rarely experience the extreme versions of chaos the GC model may conjure. Yet the model also suggests how disorder and ambiguity may be transformed into relative order and certainty. It is very possible, for instance, that disorder and ambiguity may be tidied up on the output side of decision-making processes: order and certainty then represent *post hoc* rationalizations of essentially *ad hoc* decisions. Yet order and certainty may also be produced during decision making through a process known in current parlance as *social construction*. Ambiguity is transformed into relative certainty by social processes in which individuals develop and institutionalize shared understandings and judgments. March and Olsen elaborate on this constructivist mechanism in their book, *Rediscovering Institutions* (1989), where they situate garbage can theory in a larger theoretical framework of political institutionalism (see *Institutionalism*). Arguing that politics can be understood as the collective interpretation and enactment of meaning, they suggest that institutionalization leads individuals and organizations to act according to a 'logic of appropriateness.' March and Olsen (1989, pp. 69–116) provide a nice example of how garbage can processes, political institutionalism, and social constructivism can complement one another in their historical analysis of the failure of comprehensive administrative reforms.

The GC literature also suggests at least two other mechanisms through which anarchies organize: leadership and learning. With respect to leadership, for instance, this literature suggests that unobtrusive leadership which embraces rather than rejects the disorder and ambiguity of decision making may be successful in steering organized anarchies toward collective purposes (Cohen and March 1974, Kingdon 1995, Padgett 1980). Learning in organized anarchies is problematic because the lessons we should draw from experience are themselves ambiguous. While learning under conditions of ambiguity may lead to superstitious beliefs, the GC literature suggests that organized anarchies may adapt and learn successfully (Cohen and March 1974, March and Olsen 1979). In fact, it is currently fashionable to see a bit of chaos as necessary for producing creativity and innovation. Hood (1999) notes the interesting parallels between the garbage can model and Mintzberg's model of 'adhocracy,' which is widely promoted as an organizational design for promoting creativity and flexibility.

The GC model has been applied in numerous contexts. As March and Olsen have noted, the theory probably applies to many organizations some of the time and some organizations most of the time. It is often seen as a good fit for educational institutions, especially universities. Sproull et al. (1978) argue that the characteristics of organized anarchies characterize 'most new government organizations, irrespective of their mission, and most agencies operating in social policy domains, irrespective of their age' (Sproull et al. 1978, p. 5). In addition to their book on the National Institute of Education, other studies of political organization inspired by the GC model include analyses of presidential management (Padgett 1980), military organizations and operations (March and Weissinger-Baylon 1986), public-health management in the context of an 'interorganizational garbage can' (Clarke 1989), policy making and agenda setting in the federal government (Kingdon 1995), and foreign policy decision making (Newmann 1998). While these differ in their details, the GC model should also be seen as similar in spirit to Graham Allison's model of bureaucratic politics, to Charles Lindbloom's model of disjointed incrementalism (see *Decision-making Systems: Personal and Collective*), to Norton Long's notion of politics as an 'ecology of games,' and to Jeffrey Pressman and Aaron Wildavsky's theory of implementation (see *Implementation: Political*). Taken together, these models have made a major contribution to our understanding of the limitations of the rational model of decision making and to the possibility of alternative rationalities.

See also: Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Organizational Behavior, Psychology of; Organizations, Sociology of; Rational Choice and Organization Theory; Rational Choice in Politics; Rational Choice Theory in Sociology

Bibliography

Clarke L 1989 *Acceptable Risk? Making Decisions in a Toxic Environment*. University of California Press, Berkeley, CA

Cohen M D, March J G 1974 *Leadership and Ambiguity: The American College President*. McGraw-Hill, New York

Cohen M D, March J G, Olsen J P 1972 A garbage can model of organizational choice. *Administrative Science Quarterly* **17**: 1–25

Hood C 1999 The garbage can model of organization: Describing a condition or a prescriptive design principle. In: Egeberg M, Lægreid P (eds.) *Organizing Political Institutions: Essays for Johan P. Olsen*. Scandinavian University Press, Oslo, Norway, pp. 59–78

Kingdon J W 1995 *Agendas, Alternatives, and Public Policies*, 2nd edn. HarperCollins, New York

March J G, Olsen J P 1979 *Ambiguity and Choice in Organizations*, 2nd edn. Universitetsforlaget, Bergen, Norway
March J G, Olsen J P 1989 *Rediscovering Institutions: The Organizational Basis of Politics*. Free Press, New York
March J G, Weissinger-Baylon R 1986 *Ambiguity and Command: Organizational Perspectives on Military Decisionmaking*. Pitman, Boston
Newmann W W 1998 Foreign policy decision making, garbage cans, and policy shifts: The Eisenhower administration and the 'Chances for Peace' speech. *American Review of Public Administration* **28**: 187–212
Padgett J F 1980 Managing garbage can hierarchies. *Administrative Science Quarterly* **25**: 583–604
Simon H 1985 *The Sciences of the Artificial*. MIT Press, Cambridge, MA
Sproull L, Weiner S, Wolf D 1978 *Organizing an Anarchy: Belief, Bureaucracy, and Politics in the National Institute of Education*. University of Chicago Press, Chicago

C. K. Ansell

Gastrointestinal Diseases: Psychosocial Aspects

Research in the area of gastrointestinal diseases has focused primarily on biological and medical aspects of these disorders. Little consideration has been given to psychosocial aspects of these medical conditions. However, there has been one general area, namely irritable bowel syndrome (IBS) that has received increased attention in the area of psychosocial aspects over the past few decades. More specifically, IBS has been increasingly conceptualized using a biopsychosocial perspective (Drossman et al. 2000). This perspective provides a framework for understanding the biological and psychosocial components of IBS, and can provide a model for other areas within the digestive system.

The purpose of this article is to offer an understanding of the psychosocial aspects of gastrointestinal illness. Because it is beyond the scope of this article to supply an extensive and detailed discussion of all the specific diseases of the digestive system, we have limited our discussion to psychosocial aspects of IBS since this is where most of the theoretical and empirical literature has focused. The first section of this article will furnish an overview of IBS. The second section will address how a biopsychosocial perspective relates to IBS. The third section will summarize current psychosocial treatments for IBS. The final section will briefly address future directions.

1. *Definition and Symptoms of IBS*

Functional gastrointestinal disorders (FGID) are characterized by recurrent or current gastrointestinal symptoms that have no identifiable structural or biochemical basis. The most common functional gastrointestinal disorder is IBS. Recent evidence supports the belief that dysregulation of the brain–gut neuroenteric systems, rather than the presence of structural disease, promotes the development and persistence of IBS (Drossman et al. 2000). IBS is associated with abdominal pain or discomfort for at least 12 weeks or more, that has two or three of the following features: (a) relieved with defecation; and/or (b) associated with a change in frequency of stool; and/or (c) associated with a change in form (appearance) of stool. Symptoms that are associated with IBS include: abnormal stool frequency (more than three bowel movements per day and less than three bowel movements per week), abnormal stool form (hard or loose/watery stool), abnormal stool passage (straining or urgency, feeling of incomplete evacuation), passage of mucus, bloating, or feeling of abdominal distension (Thompson et al. 2000).

2. *Epidemiology*

IBS is a very common disorder, estimated to affect 9–22 percent of the Western population. However, of those affected, very few seek medical consultation or treatment for their gastrointestinal symptoms (Whitehead et al. 1988). According to Drossman et al. (1997), of those who seek treatment, IBS accounts for 28 percent of gastroenterological practice in Western societies and 12 percent of primary care. IBS is often associated with unnecessary procedures and surgeries. IBS has considerable economic impact since IBS results in decreased work capacity. Next to the common cold, it ranks as the second most common cause of work absenteeism.

IBS is a disorder that is diagnosed mostly in women (Toner et al. 2000). To date, most of our information about IBS has been drawn from female participants. Accordingly, to date, we know little about IBS in men. A review of the literature by Toner and Akman (2000) noted that the majority of studies investigating IBS use only women in their samples. Moreover, of those studies which do include men, few perform a gender difference analysis. Since a major percentage of IBS patients are women, the issue of gender must be integrated into the conceptualization and treatment of these disorders (Toner 1994).

3. *Conceptualization*

IBS is best conceptualized using a biopsychosocial model. As summarized by Drossman et al. (1999), the biopsychosocial model provides a useful framework to help us begin to understand the connections among psychosocial factors, gastrointestinal function, disease susceptibility, and clinical outcome. Symptoms may be generated from physiological disturbances (enhanced motility and visceral sensation), which are closely

connected to central nervous system activity (via the central nervous system–enteric nervous system axis). The clinical expression of these symptoms (e.g., the decision to take medication or seek health care) is strongly influenced by psychosocial factors. For this reason, the high frequency of psychosocial factors (e.g., psychiatric diagnoses, sexual abuse history) in the absence of modulating factors (e.g., social support, coping strategies) reported among IBS patients may, in part, relate to their self-selection into referral practices. Accordingly, while psychosocial factors are not part of the diagnosis of IBS, they have important effects on gut motility, symptom experience, and clinical outcome.

In light of this information, it is no longer logical that clinicians and researchers try to assess whether IBS symptoms are caused by physiological or psychosocial factors since both are operative and interactive. Rather, the goal is to determine to what extent these factors contribute to IBS and are remediable (Drossman et al. 1999).

4. *Psychosocial Aspects Discussed in the Context of Myths Associated with IBS*

It is difficult to discuss psychosocial aspects without acknowledging the stigma and myths that have been associated with IBS. Functional gastrointestinal disorders such as IBS are associated with morally pejorative connotations. Patients who have functional gastrointestinal disorders often feel that their problems are treated as 'not real' and due to a moral and/or psychological weakness or defect. Thus, when referred to health care providers, persons with IBS may believe that caregivers think that their symptoms are 'all in their heads' and not 'real' or 'serious.' It is important that health care professionals validate the reality of the person's symptoms. In doing so, health care professionals can challenge the artificial dualism between functional and organic components of illness in general and pain in particular imposed by our Western society (Toner et al. 2000). Dualism leads to a prominent myth which suggests that if pain is severe, there must be an organic cause. This is a common myth that persists as a function of our Western society's false conceptualization of pain (i.e., if the pain is severe, there must be a structural cause). However, the experience of pain is a consequence of a complex interaction among physical, cognitive, emotional, and behavioral components.

A related myth is that IBS is caused by stress. We now know that IBS is not caused by stress, but that stress may further aggravate bowel symptoms. We also know that stress is associated with bowel symptoms in most people. In fact, over 70 percent of people under stress will exhibit some GI symptoms. However, studies have found that people with IBS are more reactive since they already have a hypersensitive gut and therefore are more likely to experience more symptoms under any kind of stress including psychological, physical, dietary, or hormonal.

An additional myth is that IBS is a psychiatric disorder or that it is masked depression. Several studies have found that a substantial percentage of individuals presenting with IBS, particularly in tertiary care settings, also have an associated depressive or anxiety disorder. It is unclear why such an association exists, but the following explanations have been proposed. (a) People who also have an associated anxiety/depressive disorder may have more difficulty coping with gastrointestinal problems and may seek specialized help for their pain at higher rates than people with IBS who do not have an associated anxiety or depressive disorder. (b) Anxiety and depression are common in the general population, and their co-occurrence may simply be due to high frequency. (c) Depression/anxiety may be a consequence of living with a chronic, incapacitating disorder such as IBS (Toner et al. 2000).

A final myth is that people with IBS are sometimes described as difficult patients. Physicians and other health care providers are not adequately trained in the conceptualization and/or treatment of IBS. The lack of information coupled with the shame and trivialization associated with having a so-called functional or psychosomatic disorder leads to frustration, further distressing patients. For physicians, such patients may illicit uncomfortable feelings that might lead some health care professionals to feel powerless and frustrated in the situation. Rather than conceptualizing IBS patients as difficult, it is more helpful to conceptualize the disorder as difficult, especially in light of the fact that there is little information and a great deal of stigma associated with IBS (Toner et al. 2000). An open and honest discussion acknowledging the frustration associated with this chronic and debilitating disorder is beneficial. To date, there are no effective medication or cure for IBS and thus the focus of treatment is symptom management and improving coping strategies. As such, it is important to establish a collaborative environment whereby both the person with a diagnosis of IBS and health care provider work together in attempting to understand the most effective strategies in managing these chronic, painful symptoms.

5. *Empirical Support for Psychosocial Treatments*

Psychosocial treatments for IBS have included dynamic psychotherapy, cognitive-behavioral therapy, relaxation, biofeedback, and hypnosis. In a recent overview of this area, a multinational team including Drossman et al. (2000) reviewed 13 controlled studies to compare psychosocial treatment with conventional medical treatment. Ten of 13 studies showed signifi-

cant superiority of psychosocial over conventional medical treatment in terms of reducing bowel symptoms. Follow-up data were available for only nine studies and of these, eight showed psychosocial treatment to be superior to medical treatment.

However, it is difficult to determine the efficacy of a specific theoretical approach because most of the controlled studies have used multicomponent treatment packages (e.g., cognitive-behavioral, psychodynamic, biofeedback, and relaxation approaches). As such, it is difficult to assess which aspects of the treatments were beneficial and to what extent. Moreover, therapeutic techniques and procedures are not generally described adequately and in specific detail. This makes replications of treatment protocols difficult. Nevertheless, the data thus far seem to indicate that when compared to conventional medical treatment, psychosocial interventions are consistently more effective.

6. *Summary and Future Directions*

This article has provided a brief overview of IBS with a focus on associated psychosocial aspects. While psychosocial factors are not part of the diagnosis of IBS, clearly, psychosocial factors play an important role in symptom expression and clinical outcome of IBS. Given the significance of these factors, it is important that clinicians respond to psychosocial issues with the same attention and legitimacy that is given to the physical and biological factors associated with gastrointestinal disorders. Conceptualizing IBS from a biopsychosocial perspective allows clinicians to better understand and treat individuals who are diagnosed with IBS in a more comprehensive manner. Future research must examine how biopsychosocial factors influence other diseases of the digestive system.

Research has suggested that psychosocial treatments are effective interventions for IBS. Further research is needed to determine the usefulness of psychosocial interventions for other diseases of the digestive system. Moreover, future studies are required to determine which aspects of psychosocial treatment packages (e.g., relaxation, cognitive restructuring, etc.) account for their effectiveness and determine the patient characteristics which predict response to specific psychosocial treatments. Since this is a chronic disorder, long-term studies must also be conducted. The role of sociocultural factors and gender on IBS must be further explored in both our understanding and treatment of these chronic debilitating disorders.

See also: Behavioral Medicine; Biopsychology and Health; Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Chronic Pain: Models and Treatment Approaches; Emotions and Health; Illness Behavior and Care Seeking; Pain, Health Psychology of; Pain, Management of; Pain, Neural Basis of; Psychosomatic Medicine; Stress and Coping Theories; Stress and Health Research; Stress Management Programs; Stress, Neural Basis of; Stress: Psychological Perspectives; Symptom Awareness and Interpretation

Bibliography

Drossman D A, Whitehead W E, Camilleri M 1997 Medical position statement: Irritable Bowel Syndrome. *Gastroenterology* **112**: 2118–9

Drossman D A, Creed F H, Olden K W, Svedlund J, Toner B B, Whitehead W E 1999 Psychosocial aspects of the functional gastrointestinal disorders. *Gut* **45**(Suppl. 2): 1125–30

Drossman D A, Creed F H, Olden K W, Toner B B, Whitehead W E 2000 Psychosocial aspects of the functional gastrointestinal disorders. In: Drossman D A, Corazziari E, Talley N J, Thompson W G, Whitehead W (eds.) *The Functional Gastrointestinal Disorders*, 2nd edn. Allen Press, Lawrence, KS, pp. 157–245

Thompson W G, Longstreth G, Drossman D A, Heaton K, Irvine E J, Muller-Lissner S 2000 Functional bowel disorders and functional abdominal pain. In: Drossman D A, Corazziari E, Talley N J, Thompson W G, Whitehead W (eds.) *The Functional Gastrointestinal Disorders*, 2nd edn. Allen Press, Lawrence, KS, pp. 360–1

Toner B B 1994 Cognitive-behavioral treatment of functional somatic syndromes: Integrating gender issues. *Cognitive and Behavioral Practice* **1**: 157–78

Toner B B, Akman D 2000 Gender role in Irritable Bowel Syndrome: Literature review and hypothesis. *American Journal of Gastroenterology* **95**(1): 11–6

Toner B B, Segal Z V, Emmott S D, Myran D 2000 *Cognitive-Behavioral Treatment of Irritable Bowel Syndrome: The Brain–Gut Connection*. Guilford, New York

Whitehead W E, Bosmajian L, Zonderman A B, Costa P T Jr, Schuster M M 1988 Symptoms of psychologic distress associated with Irritable Bowel Syndrome: Comparison of community and medical clinic samples. *Gastroenterology* **95**: 709–14

B. Toner

Gauss, Carl Friedrich (1777–1855)

1. *Introduction*

The name of Gauss will already be known to many readers for its use in such phrases as Gaussian distribution, Gauss linear model, and the Gauss-Markov theorem, and for its use in the titles of the Gaussian elimination, Gauss–Jordan, Gauss–Doolittle, Gauss–Newton, and Gauss–Seidel procedures, if not for the *Système International* (SI) unit of magnetic induction and the associated terms gaussage and degauss.

Gauss contributed to all areas of science known in his day, notably to astronomy, physics, and geodesy,

but in this article we shall concentrate on his contributions to statistics and numerical methods. For more wide-ranging surveys of his work, see Eisenhart (1968) and May (1972).

2. Life of Gauss

Carl Friedrich Gauss (originally Johann Friedrich Carl Gauss) was born on 30 April, 1777 in Brunswick (Germany). He was the son of a bricklayer and gardener who also kept the accounts of a local insurance company. Gauss received his elementary education at St. Katherine's School in Brunswick. However, his humble origins were not allowed to hamper his intellectual development, for, in 1791, his talent for mathematics was brought to the attention of the Duke of Brunswick, who undertook to support his continuing education, at first at the *Collegium Carolinum* in Brunswick and later at the University of Göttingen. Gauss did not graduate from the University of Göttingen but was awarded a doctoral degree *in absentia* by the University of Helmstedt in 1799 for his work on the fundamental theorem of algebra.

In October 1806, the French land forces consolidated the command of German territory that they had established in 1805 in the battle of Austerlitz by defeating a Prussian army under the command of the Duke of Brunswick at Auerstedt (and mortally wounding the Duke). As a consequence, when in 1807 Gauss submitted the manuscript of his future *Theoria Motus* for publication in German, the political situation was such that his publisher felt obliged to insist on its translation into Latin. Consequently, the manuscript which Gauss had completed in 1806 was not published until 1809. (Fortunately, Gauss had an excellent command of Latin, so that this imposition may justly be compared with requiring a modern German scholar to translate his work into English before publication.)

In 1807 Gauss was appointed director of the University of Göttingen observatory and professor of mathematics at the University of Göttingen. He continued to hold both of these positions until his death in Göttingen on 23 February, 1855.

3. Elementary Fitting Procedures

Given any number of observations on a lesser number of variables, the linear fitting problem is concerned with establishing the relationship between one of these variables and the remaining variables. Before the middle of the eighteenth century, practitioners were obliged to employ a variety of more or less arbitrary procedures to determine the parameters of such relationships. But between 1750 and 1800 four algebraic procedures were proposed for the linear fitting problem, namely the method of averages, the method of selected points, the method of least absolute deviations, and the minimax absolute deviation procedure.

In 1805 Adrien-Marie Legendre (1752–1833) published an alternative algebraic procedure which he named the method of least squares. This procedure chooses the parameters of the relationship to minimize the sum of the squared deviations between the observed values and the fitted relationship. For a detailed history of these five algebraic procedures, see Farebrother (1999). Also note that the relevant pages of Legendre's original statement have been reproduced by Stigler (1986, p. 56).

Legendre's approach to the fitting problem was entirely algebraic and had no statistical content. Further, he did not prove many of the optimality properties he claimed for the method of least squares. Indeed, he had not even shown that the values determined by his algebraic procedure actually minimize the sum of the squared deviations function (except in the unlikely case when the fitted relationship is exact). Such a proof was subsequently supplied by Gauss (1809).

4. Most Probable Value Criterion

Gauss' first statistical derivation of the method of least squares was based on a combination of the normal or Gaussian distribution with a Bayesian variant of the maximum likelihood or maximum probability optimality criterion. The normal distribution itself was obtained from a preliminary analysis in which the arithmetic mean of a set of direct observations is associated with the maximum value of this optimality criterion.

Gauss was working in a Bayesian context, so there was no need for him to estimate the disturbance variance, but he employed a simple algebraic procedure that we would now associate with the LDL' decomposition of a symmetric matrix to obtain measures of relative precision of each of the parameter estimate.

5. Priority Disputes

In his 1809 book, Gauss referred to the method of least squares as 'our principle, which we have made use of since the year 1795' and thus sparked off a serious priority dispute with Legendre that ran on for several years. Although Legendre had certainly been the first to publish a description of the method, it is now generally agreed that Gauss had employed a similar procedure in his unpublished work, notably in his 1801 analysis of the geodetic measurements made in France, see Plackett (1972) and Stigler (1999, Chap. 17).

In this context, we should also mention that the distribution named for Gauss had previously been discussed by Abraham de Moivre (1667–1754) and Pierre Simon Laplace (1749–1827); see *Husserl, Edmund (1859–1938)*, Stigler (1999, Chaps. 14 and 22).

6. *Asymptotic Results*

In his *Théorie analytique des probabilités* of 1812, Laplace had made use of a nonrigorous asymptotic argument to establish the optimality of the method of least squares. Gauss (1816) employed a similar technique to establish the relative efficiency of the scaled sums of powers of the absolute disturbance terms when employed as estimates of the variance of their common distribution function. In particular, he found that the scaled sum of the squared disturbances is a more efficient estimator of the variance parameter than is the scaled sum of their absolute values and significantly better than the scaled sums of their third and higher powers.

7. *The Most Plausible Value Criterion*

In 1818–19 Gauss abandoned his earlier Bayesian derivation of the method of least squares and adopted an alternative approach which he later published as Gauss (1823–28). Farebrother (1999) has suggested that the principal features of Gauss' second approach to the fitting problem may be traced to a careful recombination of certain elements in Laplace's earlier work. Gauss adopted Laplace's decision theoretic approach to the fitting problem, and his restriction to what we would call a class of linear unbiased estimators, but he employed finite sample theory rather than asymptotic analysis, and chose a value for the parameter to minimize the expected value of the squared error of estimation where Laplace had chosen to minimize its expected absolute value.

The principle result to this new theory, now known as the Gauss–Markov theorem, identifies the least squares estimator as the member of the class of linear unbiased estimators with minimum mean squared error.

In this context, it is interesting to note that Gauss' restriction to the class of linear unbiased estimators was an accidental by-product of Laplace's earlier concern with estimators defined by the method of averages, see Stigler (1986) or Farebrother (1999). However, Gauss also showed that the scaled sum of squared deviations is an unbiased estimator of the disturbance variance, but without making any strong claims for the optimality of the result.

Gauss (1823–8) gave a very full account of the subject. In particular, he extended his earlier analysis by considering the effect on the fitting problem of adding a further observation, or of imposing a linear constraint on the parameters of the model. The resulting updating formulas now form the basis of the Gauss–Doolittle fitting procedure and the Kalman filter of times series analysis. They also define the set of recursive residuals employed in some statistical tests.

8. *Solution of Linear Equations*

The solution technique now known as Gaussian elimination is a development of the procedures known to Tobias Mayer (1723–62) and the ancient Chinese for solving systems of linear equations with the same number of unknowns. In his 1809 worked example, Gauss employed a solution technique which may possibly be associated with the ancient method, see Farebrother (1999), but in 1811 he outlined the steps of a solution technique that identified the elimination procedure which now goes by his name. This procedure has two phases: in the forward elimination phase the first equation is used to eliminate one of the unknown constants from the other equations; then the second equation is used to eliminate a second unknown constant from the remaining equations; and so on. This yields an upper triangular system which is then solved for the unknown constants in reverse order.

Gauss illustrated this computational technique by a worked example concerned with the least squares determination of the orbit of the newly-discovered asteroid Pallas.

The so called Gauss–Jordan procedure differs from the original procedure in that the array of numbers is diagonalized in the forward elimination phase. This minor modification (which is familiar from textbook descriptions of the simplex procedure of linear programming) leads to a considerable simplification of the second phase, but usually at the cost of some loss of numerical accuracy, see Farebrother (1988).

9. *Matrix Algebra*

As a by-product of his work in number theory and the calculus of observations, Gauss (1801, 1809, 1823–8) made use of several concepts now recognized as belonging to the theory of matrices. These include linear dependence; linear transformations; the Lagrange LDL' and Cholesky LL' decompositions of a symmetric matrix; the rank-one adjustment of a symmetric matrix; the inverse of a triangular matrix; the inverse of a symmetric matrix; and, possibly, the generalized inverse of a matrix of full rank. He also explicitly proved that the inverse of a symmetric matrix is itself symmetric.

However, despite these early beginnings, the familiar matrix formulation of the least squares problem did not materialize until 1935, and it did not become a

common feature of statistical textbooks until the second half of the twentieth century, see Farebrother (1999).

10. Iterative Fitting Procedures

Gauss (1809) also explained how nonlinear relationships may be fitted by means of an iterative variant of the method of least squares, now known as the Gauss–Newton procedure, in which the fitted relationship is linearized by a Taylor series expansion and fitted by the method of least squares. This process being repeated until a satisfactory level of convergence is obtained.

Despite its early origin, and the publication by Gauss of detailed worked examples in astronomy and geodesy, the Gauss–Newton procedure was not widely adopted for practical work in other areas until the necessary computing equipment became readily available in the second half of the twentieth century. Instead, practitioners avoided the computational burden imposed by the iterative procedure by transforming their nonlinear relationships to a more linear form before estimating the parameters of the resulting equations by a *single* application of the method of selected points, the method of averages, or the method of least squares.

A second iterative procedure usually named for Gauss is the Gauss–Seidel procedure, although Gauss and Seidel were actually not concerned with the procedure that usually goes by this name, but with a distinct fitting procedure known as the method of *relaxation*; see Gauss (1823) for details.

11. Concluding Remarks

The method of least squares and the related statistical procedures known as linear regression analysis and the analysis of variance are widely used in all areas of science. Thus there can be few numerate practitioners of the social and behavioral sciences who have not read research reports that employ these methods to fit linear or curvilinear relationships to data on the variables of interest.

The method of least squares was developed by Legendre, Gauss, and Laplace in the first quarter of the nineteenth century in the context of important practical problems in astronomy and geodesy. But it was not until later in the century that Quetelet (see *Quetelet, Adolphe (1796–1874)*), Galton (see *Galton, Sir Francis (1822–1911)*), and Edgeworth (see *Edgeworth, Francis Ysidro (1845–1926)* among others adjusted this technique to the needs of the social and behavioral sciences. There was no great shortage of practitioners trained in the use of advanced statistical techniques, but such individuals as were active in this area were not able to treat their observations with the necessary level of detachment, so that the original source of the observations remained a significant feature of their subsequent analyses.

The method of least squares was absorbed relatively rapidly into psychology, but its rate of infusion into the economic, social, and behavioral sciences (other than psychology) was more gradual. Indeed, applications of this technique were not at all common in these areas until well into the twentieth century. See Stigler (1986, 1999) for a detailed account of the gradual transformation of these subject areas (see also *Quetelet, Adolphe (1796–1874)*; *Edgeworth, Francis Ysidro (1845–1926)*; *Thorndike, Edward Lee (1874–1949)*).

The history of the method of least squares clearly demonstrates that practical scientists with little or no interest in the social and behavioral sciences can make significant contributions to this subject area by developing general analytical procedures that can be adjusted to the particular needs of social and behavioral scientists.

Bibliography

Bertrand J L F 1855 *Méthode des moindres carrés: mémoires sur la combinaison des observations par Ch.-Fr. Gauss*. Mallet-Bachelier, Paris [1957 Trotter H F Gauss's work (1803–1826) on the theory of least squares, *Technical Report 5*, Statistical Techniques Research Group, Princeton University]

Börsch A, Simon P (1887) 1964 *Abhandlungen zur Methode der kleinsten Quadrate von Carl Friedrich Gauss*, Physica-Verlag, Würzburg, Germany

Eisenhart C 1968 Carl Friedrich Gauss (1777–1855). In: Sills D (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 6, pp. 75–81

Farebrother R W 1988 *Linear Least Squares Computations*. Marcel Dekker, New York

Farebrother R W 1999 *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900*. Springer-Verlag, New York

Gauss C F (1801) 1870 *Disquisitiones Arithmeticae*. Vol. 1 of *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göttingen, Germany [1966 English trans. Clarke A A *Disquisitiones Arithmeticae*. Yale University Press, New Haven, CT

Gauss C F (1809) 1871 *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Vol. 7 of *Carl Friedrich Gauss Werke*. F. Perthes, Gotha [(1857) 1963 Davis C H *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections*. Dover, New York]

Gauss C F (1811) 1874 Disquisitio de elementis ellipticis Palladis ex oppositionibus annorum 1803, 1804, 1805, 1807, 1808, 1809. In: *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göttingen, Vol. 6, pp. 1–24

Gauss C F (1816) 1880 Bestimmung der Genauigkeit der Beobachtungen. *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göttingen, Vol. 4, pp. 109–217

Gauss C F (1823–8) 1880 Theoria combinationis observationum erroribus minimis obnoxiae (in three parts.) *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göt-

tingen, Vol. 4, pp. 3–93 [1995 Stewart G W. *Theory of the Combination of Observations Least Subject to Errors*. Society for Industrial and Applied Mathematics, Philadelphia, PA]
Gauss C F (1823) 1903 Letter to Gerling, Göttingen 26 December 1823. *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göttingen, Vol. 9, pp. 278–81 [1951 Forsythe G E Gauss to Gerling on relaxation, *Mathematical Tables and Other Aids to Computation* **5**: 255–8]
May K O 1972 Carl Friedrich Gauss. In: Gillispie C C (ed.) *Dictionary of Scientific Biography*. Scribners, New York, Vol. 5, pp. 298–315
Plackett R L 1972 The discovery of the method of least squares. *Biometrika* **59**: 239–51
Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1990*. Belknap Press of Harvard University Press, Cambridge, MA
Stigler S M 1999 *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA

R. W. Farebrother

Gay, Lesbian, and Bisexual Youth

1. Introduction

Youths who identify as gay, lesbian, or bisexual have become increasingly visible during the 1990s. Although constituting only about 2 percent of the adolescent population, these self-identified youths are, in turn, a subset of the much larger population of sexual-minority youths with same-sex attractions who do not adopt a gay, lesbian, or bisexual label. It is this larger group of youths that has been neglected by social and behavioral scientists. By focusing only on 'out and visible' youths, researchers, educators, and mental health professionals may have unwittingly created and perpetuated myths about the behavior, attitudes, and mental health of gay youths. Indeed, more recent developmental research that includes the larger category of sexual-minority youths indicate that these youths are often indistinguishable from their heterosexual peers, except in the nature of their sexual desires.

2. Definitions and Prevalence

The term *sexual minority* characterizes a broader class of individuals than those who identify as lesbian, bisexual, or gay. This is a particularly important consideration because during adolescence, youths frequently explore issues related to their sexuality that extend beyond traditional identity labels. Their sexual attractions, fantasies, and behavior may have no easy categorization. Thus, for researchers and lay people alike, the issue of *sexual identity* labels is a difficult one. Most people are familiar with the labels 'heterosexual,' 'gay,' 'lesbian,' and 'bisexual,' but not 'unlabeled,' 'bi-lesbian,' and 'not straight'—terms that current cohorts of youth, particularly young women, are likely to assume (Savin-Williams and Diamond 2000).

Included under the rubric of sexual minority are individuals who reject, for personal or political reasons, cultural definitions of any sexual category. Perhaps they believe that no term can adequately describe their attractions ('unlabeled'), are uncertain as to the nature of their sexual attractions ('questioning'), have a mixture of bisexual and lesbian attractions ('bi-lesbian'), are without sexual attractions ('asexuals'), use markers other than sex of person as the basis for their sexual attractions (e.g., personality, age, sexual behavior), or do not consider themselves to be traditionally heterosexual ('not straight').

Another contemporary sexual-minority group increasing in visibility is comprised of transgender individuals, who may be of any sexual orientation or identity. *Transgender* is an umbrella term for those who believe that they do not meet traditional definitions of masculinity or femininity consistent with their biological sex.

One simple resolution to defining who is 'gay' would be to eschew sexual labels altogether and rely on descriptions of behaviors, desires, or attractions. Rather than focusing on a particular sexual identity label, researchers should be asking questions about the existence, number, and proportion of sexual behaviors, attractions, or romantic relationships that an individual has with males and females.

Given cultural prohibitions against sexual minorities claiming their identity, it is not known with certainty how many youths experience same-sex attractions. Yet this number is likely to be far greater than the number of youths who identify as gay, lesbian, or bisexual. Three recent surveys illustrate this point. In a national, representative sample of adults, Laumann et al. (1994) reported that 1.4 percent of women and 2.8 percent of men identified as lesbian, bisexual, or gay. Far more, however, reported having thought about having sex with someone of their own gender—5.5 percent of women and 6 percent of men—or having experienced same-sex attractions at least once—7.5 percent of women and 7.7 percent of men. Second, a Youth Risk Behavior Survey of Massachusetts high schools (Goodenow and Hack 1998) reported that 2 percent of students self-identified as gay, lesbian, or bisexual; however, an additional 2 percent reported having engaged in same-sex behavior despite eschewing a sexual-minority label. Finally, in the National Longitudinal Study of Adolescent Health, a comprehensive, representative study of adolescents, 8 percent of males and 6.6 percent of females reported that they had same-sex romantic attractions or had been in a same-sex romantic relationship (Russell and Joyner 1998).

These findings indicate that far more youths experience same-sex attractions, romantic desires, sexual behavior, and romantic relationships than report that they are lesbian, gay, or bisexual. Thus, many 'potential' sexual-minority youths do not identify as such during adolescence and therefore fail to be included in research studies of sexual minorities. Hence, the overwhelming body of research on youths with same-sex attractions samples a select, likely nonrepresentative minority of this population. This conclusion is consistent with McConaghy's (1999) recent assertion about the most important 'unresolved issues in scientific sexology.' Too often researchers assume that sexual attractions are distributed categorically rather than dimensionally, rendering invisible the 'nongay majority with homosexual feelings.'

Although the reasons many youths with same-sex attractions decline to adopt a lesbian, gay, or bisexual label have not been investigated empirically, it is likely linked to negative cultural attitudes and beliefs about same-sex attracted individuals. The associated stigma causes many youths to forego altogether or to delay identification as a sexual minority. It is not known the ways in which those who adopt a sexual identity label differ from the much larger population of same-sex attracted individuals; however, some evidence suggests that they may be psychologically less healthy (see below).

Clearly, many youths who know that they are not heterosexual are unprepared, either developmentally or politically, to attach themselves to a culturally defined sexual category that, as currently constructed, carries profound psychological and social ramifications. To be gay implies belonging to a class of individuals who are subject to hate crimes, prejudice, and stereotypes. Unfortunately, youth often do not realize that embracing this label also opens the door to a community of like-minded individuals who can provide understanding, acceptance, and support for one's central sense of self. Given the perceived consequences of a gay identification, it is to be expected that some adolescents prefer to describe their sexual desires or attractions (e.g., 'I'm attracted to women.') rather than embrace a specific, stigmatized identity label.

3. *Myths and Counter Evidence*

Relatively little is known about this class of individuals who experience significant and persistent sexual attractions and fantasies for same-sex others. Until recently, researchers have ignored such individuals and behaviors, in part due to *heterocentrism* (the assumption that the universe is heterosexual) and *homonegativity* or *homophobia* (negative attitudes, beliefs, and behaviors toward those with same-sex attractions). Other reasons for this inattention include the belief that because youths are fickle with yet unformed sexuality, as evidenced by the frequency of adolescent sexual exploration, it is not worthwhile investigating the correlates and consequences of adolescent same-sex experiences. Finally, there is the common belief that researchers should not ask sexual questions because sexuality is highly private and that inquires about sexuality would not be permitted by parents, research institutions, or funding agencies. Traditional social and religious organizations sometimes fear that by raising the issue of homosexuality, they increase the 'risk' of youths engaging in same-sex behavior and identifying as gay. As a result of this negligence, many myths about sexual-minority youths have taken root. Three prevalent myths are reviewed briefly.

3.1 *Youths Develop Their Sexual Identity Through a Series of Coming-out Stages*

During the 1970s and 1980s, social scientists and clinicians made sense of these young lives by proposing a number of ideal or typical 'coming-out' models (e.g., Cass 1979). These structural models were cast in a somewhat 'one-size-fits-all' mold that outlined an orderly series of stages based on a particular theoretical understanding of human development. The life histories of sexual-minority youths were proposed to follow a series of stages, from first awareness of being different, sexual exploration, and self-labeling; to first disclosures, romantic relationships, and self-acceptance; and to political/social activism and integration of a sexual identity with the rest of one's life. With few exceptions, the empirical basis for these models remains extraordinarily weak, largely because researchers have not investigated the complexities of these models, studied youths over time, developed appropriate research tools to asses youths' 'fit' with the models, or included diverse, under-represented populations of youths. Given these limitations, it should come as no surprise that recent studies found that only 10 percent of young men followed these developmental stages and that the number of young women adhering to this hierarchy is even less (Savin-Williams 1998, Savin-Williams and Diamond 2000).

3.2 *Parents React With Horror and Reject Their Sexual-minority Child*

Popular advice books for parents often portray the child's act of disclosure as necessarily creating a crisis within the family that may never be fully resolved. Lesions in family relationships are inflicted, wounds that may not withstand later attempts to amend or overcome, and parents encounter unique developmental issues. This literature promotes the view that the disclosure of same-sex desires by children to their parents inevitably and uniformly places youths at psychological and physical risk. Disclosure may result

in youths being disowned, thrown out of the home, or emotionally or physically harassed. Because parents may feel that their child has rejected them and everything the family stands for, family relationships may be irrevocably damaged.

Empirical research, however, does not support these dire predictions. Around 5 percent of youths suffer physical harm from parents or are thrown out of the home once they are open about their sexuality (Savin-Williams 2000). For example, in a sample of Chicago support group youths, only 3 percent reported that they had been thrown out of the home and were living in shelters or with friends (Herdt and Boxer 1993). In another study, parental rejection and anger were the two least common reactions to the youth's disclosure of same-sex sexuality, both one week and six months after the event (Ben-Ari 1995)

3.3 Sexual-minority Youths Experience Mental Health Problems, Exemplified by High Rates of Attempted Suicide

The prevalent cultural image of gay youths is that they are suicide attempting, prostituting, drug-abusing runaways who are HIV positive. Encouraging outcomes of this depiction include sympathy and support from mainstream society, financing of community support groups and research, and conversations about homosexuality in the public school system (usually in health classes). However, the long-term effect may be disastrous—convincing some youths that the only way to be young and gay is to be disordered, unhappy, and problematic. The pervasiveness of this erroneous, degrading representation has re-pathologized gay youths, almost three decades after mental health professionals removed homosexuality from the annals of psychiatric illness.

Empirically, it is clear that some youths, especially young men, who identify as a sexual-minority early in life are at risk for mental health problems. This is particularly the case for those who seek the assistance of support groups (e.g., Hershberger and D'Augelli 1995, Remafedi et al. 1998). However, these researchers have noted that support group attending youths are also the individuals most likely to be raised in dysfunctional families, subjected to barrages of peer ridicule, and to be sex-atypical in their behavior and interests. More importantly, it is often overlooked that even among these high-risk populations, the vast majority of youths who identify as gay, lesbian, or bisexual did not attempt suicide, abuse drugs, pierce their bodies, or have unsafe sex.

Here the distinction between claiming a sexual identity label and experiencing same-sex sexual attractions or behavior becomes particularly important. Although many reports in the literature document that gay youths are three to 10 times more likely to attempt suicide and that 30 percent or more of all adolescent suicides are by gay youths (Fergusson et al. 1999, Garofalo et al. 1998, Goodenow and Hack 1998, Hershberger and D'Augelli 1995, Remafedi et al. 1998), there are reasons to suspect that these numbers overstate the prevalence of suicidality among the larger population of same-sex attracted individuals. For example, Russell and Joyner (1998) found that youths who reported romantic attractions and relationships with same-sex others—a much more inclusive category of sexual minorities than those who identify as gay—were twice, rather than three or 10, times more likely to report suicide attempts as heterosexual youths, and that these sexual-minority youths constituted 15 percent, rather than 30–40 percent, of reported suicide attempts. A study of college students indicated that the proportion of young women (10 percent) and men (11 percent) with same-sex attractions who attempted suicide was no greater than a matched sample of heterosexual women (8 percent) and men (2 percent) (Savin-Williams 2000). This sample is likely to be more representative of the general population of same-sex attracted individuals because it included individuals who identified after high school and because recruitment was based on sexual attractions, not sexual identity.

4. Conclusion

Social and behavioral scientists should consider two facts about the lives of sexual-minority youths. First, they are essentially similar to all youths, independent of sexual desires, in their biology, behavior, beliefs, attitudes, socialization, and mental health. This should be a basic assumption in empirical investigations, educational programs, and clinical interventions. For example, youths of all sexualities want parental acceptance and support, are preoccupied with peer status, desire love and sex, and wonder about their future. Although these concerns can be flavored by sexual orientation—it is far more difficult to date a same- than an opposite-sex peer in most high schools—they are nevertheless adolescent tasks, ultimately to be negotiated by all youth.

A second point to be recognized—and expected given basic developmental theory—is that sexual-minority youths also vary among themselves based on particular distinguishing characteristics. Indeed, on some domains a subset of sexual-minority youths mirror heterosexual youths more than they do other sexual-minority youths. For example, young women with same-sex attractions are often more akin to heterosexual young women than they are to gay male youths. They are less genitally oriented and more likely to value and pursue romantic relationships and to use these occasions to define their sexual attractions (Diamond 1998, Savin-Williams and Diamond 2000).

A dual focus on the ways in which sexual-minority youths are similar to all other adolescents and how they vary among themselves as the result of their

sexuality, personality characteristics, biological heritage, and other developmental processes should infuse all considerations of their lives. This more comprehensive perspective is now possible because a larger and more diverse sample of contemporary youths is willing to be counted as 'not heterosexual.' As these youths become increasingly visible it will become easier for scientists to include them in research that documents their diverse and similar developmental trajectories. Although investigators of sexual-minority youths have not yet adequately sampled from this newly visible population of youths with same-sex attractions, this will most likely occur when definitions of what constitutes a sexual minority extend beyond sexual identity labels to include all youths with same-sex attractions and behavior.

This dual focus can also be seen in the lives of sexual-minority youths. From the perspective of this contemporary generation of youths, a new world has emerged regarding what it means to be a sexual minority. Consequently, it is no longer rare to disclose the nature of one's sexuality earlier than in any previous generation—while still living at home and attending the local high school. Neither is it mandatory to partake in the 'counter culture'—the 'misfits' and disenfranchised youths—of the local school or community in order to receive support and a sense of community. Whereas 'gay liberation' once characterized the only available social or political organizations for teens, Gay/Straight Alliances, school-based clubs consisting of sexual-minority youths and their allies, now populate hundreds of high schools throughout the US. These allow a youth opportunities to discover others with same-sex attractions; a place to 'hide' until ready to self-disclose; and a means to obtain support, respect, and acceptance. Even though these youths tend to blend in with many others in their junior and senior high schools and in their communities, their issues must not be neglected.

See also: Feminist Theory: Radical Lesbian; Homosexuality and Psychiatry; Identity in Childhood and Adolescence; Sex-role Development and Education; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexual Preference: Genetic Aspects; Sexual Risk Behaviors; Teen Sexuality; Youth Culture, Sociology of

Bibliography

Ben-Ari A 1995 The discovery that an offspring is gay: Parents', gay men's, and lesbians' perspectives. *Journal of Homosexuality* **30**: 89–112

Cass V C 1979 Homosexual identity formation: A theoretical model. *Journal of Homosexuality* **4**: 219–35

Diamond L M 1998 Development of sexual orientation among adolescent and young adult women. *Developmental Psychology* **34**: 1085–95

Fergusson D M, Horwood L J, Beautrais A L 1999 Is sexual orientation related to mental health problems and suicidality in young people? *Archives of General Psychiatry* **56**: 876–80

Garofalo R, Wolf R C, Kessel S, Palfrey J, DuRant R H 1998 The association between health risk behaviors and sexual orientation among a school-based sample of adolescents. *Pediatrics* **101**: 895–902

Goodenow C, Hack T 1998 Risks Facing Gay, Lesbian, and Bisexual High School Adolescents: The Massachusetts Youth Risk Behavior Survey. Paper presented at the American Psychological Association Convention, San Francisco

Herdt G, Boxer A 1993 *Children of Horizons: How Gay and Lesbian Teens Are Leading a New Way Out of the Closet*. Beacon Press, Boston, MA

Hershberger S L, D'Augelli A R 1995 The impact of victimization on the mental health and suicidality of lesbians, gay and bisexual youth. *Developmental Psychology* **31**: 65–74

Laumann E O, Gagnon J H, Michael R T, Michaels S 1994 *The Social Organization of Sexuality: Sexual Practices in the United States*. University of Chicago Press, Chicago

McConaghy N 1999 Unresolved issues in scientific sexology. *Archives of Sexual Behavior* **28**: 285–318

Remafedi G, French S, Story M, Resnick M D, Blum R 1998 The relationship between suicide risk and sexual orientation: Results of a population-based study. *American Journal of Public Health* **88**: 57–60

Russell S T, Joyner K 1998 August Adolescent Sexual Orientation and Suicide Risk: Evidence from a National Study. Paper presented at the American Sociological Association, San Francisco

Savin-Williams R C 1998 *And Then I Became Gay: Young Men's Stories*. Routledge, New York

Savin-Williams R C 2000 Suicide attempts among sexual-minority youths: Controversy re-ignited. Unpublished manuscript available from the author, Cornell University Press, Ithaca, NY

Savin-Williams R C, Diamond L M 2000 Sexual identity trajectories among sexual-minority youths: Gender comparisons. *Archives of Sexual Behavior* **29**: 419–40

Savin-Williams R C 2001 *"Mom, Dad. I'm Gay". How families Negotiate Coming Out*, 1st edn. American Psychological Association Press, Washington, DC

R. C. Savin-Williams and L. M. Diamond

Gay/Lesbian Movements

For a movement to occur there needs to be a sense of common identity or purpose and sufficient political and economic space to organize. It is not surprizing that gay/lesbian movements are very much the products of modern liberal democratic and capitalist societies, nor that the globalization of these concepts had led to a rapid expansion of gay/lesbian identities across the world.

1. The Emergence of 'the Modern Homosexual'

While homosexual behavior has existed in most societies at most times in history, the creation of a

specific identity based upon homosexual behavior is a far more recent and limited phenomenon. The term 'homosexual' was coined in 1869 by the Hungarian doctor Karoly Benkert, but there is some evidence that people defined by a shared sexual attraction to the same gender had already formed loose social groupings in several European cities. More common however was a range of understandings of sexuality and gender which saw homosexual behavior as part of an undifferentiated range of possible sexual behaviors, and an equal range of acceptance and condemnation of such behaviors. Such understandings were always heavily gendered, and often ignored the potential for female homosexuality.

By the end of the nineteenth century the idea of the homosexual existed to a sufficient extent that it became possible to imagine the creation of both social and political organizations for those people who identified with the term. The first overtly political organization, the Scientific-Humanitarian Committee, was founded by the German Magnus Hirschfeld in 1897 and campaigned for the decriminalization of homosexuality.

Hirschfeld believed that homosexuals constituted a 'third sex,' and therefore were inherently different, and no more deserving of condemnation than, say, someone born left-handed. However, against the idea of the 'third sex' Freud postulated a universal 'polymorphous perversity,' so that childhood experiences determined whether or not one developed a primarily heterosexual or homosexual orientation. While many psychoanalysts developed a strong antipathy to homosexuality, seeing it as a pathology to be cured through therapy, another tradition, which would emerge in the 1960s, used Freud's works to argue against the idea of a heterosexual norm.

After World War I both Europe and the USA saw the development of small homosexual social groups in major cities, most famously Paris and Berlin. Hirschfeld's organization was ended brutally by the Nazis, and it was not until the aftermath of World War II that new organizations emerged—COC (originally the Shakespeare Club) in the Netherlands; *Arcadie* in France; *der Kreis* in Switzerland. COC remains, by far the longest lived gay/lesbian organization in the world. In the USA the first groups were established in the 1950s (the Mattachine Society in 1951, Daughters of Bilitis in 1955) but these groups remained largely underground, although some overt political protest began in the early 1960s (d'Emilio 1983).

In Europe and North America the dislocations of World War II meant that millions of people came in contact with underground gay and lesbian worlds, and although the post-war organizations remained very small, homosexual communities developed in most major western cities through the 1950s and 1960s (d'Emilio 1983, Kennedy and Davis 1993). It was the presence of both social and commercial networks which allowed for the rapid emergence of a new sort of gay/lesbian movement in the aftermath of the major social and political upheavals in most liberal democracies from the end of the 1960s.

2. *The Development of the Gay/Lesbian Movement*

The contemporary movement was in a sense born out of the political, social, and cultural changes of 'the sixties,' symbolized by the growing freedom which surrounded discussion of sexuality, the rebirth of feminism, and the student movements of 1968. In France and Italy it was the events of May 1968 which led to the emergence of a radical gay movement; in the USA the turning point is usually assumed to be the riots which followed a raid on the Stonewall Inn in Greenwich Village, out of which the New York Gay Liberation Front was born. The gay/lesbian movement of this period saw itself as part of a larger radical assault on hegemonic cultural and political ideas and institutions, and was closely linked to various elements of New Left theory (Altman 1971). Small gay liberation groups also appeared in Canada, the UK (Weeks 1990), Scandinavia, Australasia, and some Latin American countries: Argentina in 1969; Mexico in 1971; Puerto Rico in 1974. The first Israeli gay/lesbian organization was established in 1976.

The post-1968 movement differed from its predecessors in its willingness to assert a new sense of gay/lesbian identity which demanded complete equality, at both an individual and a communal level. Although the original use of the term 'gay' included both women and men, tensions quickly emerged along gender lines, in part because of the strong interconnection between radical lesbianism and feminism, which had no equivalent for gay men (Faderman 1981, Rich 1980). During the 1970s there was increasing divergence between lesbian and gay male organizations, as a growing commercial gay world opened up new ways of living as a homosexual, particularly for men. As one novelist wrote: 'Almost through an act of will, I had made myself embrace this new identity of mine and never look back. I had gay friends. I ate at gay restaurants. I went to gay bars. I had my apartment near DuPont Circle' (Russell 1994, p. 231).

While 'gay ghettoes' such as that described here only developed in a few cities outside the USA—or, as in the case of Paris's Marais, developed a decade later—the growth of gay and lesbian commercial space has been a feature of almost all western countries from the 1970s on. In many capitalist societies the 'pink dollar' has become seen as an important niche market, and this has made possible a flourishing gay press and an expansion of businesses and professional services aimed at a specifically gay and lesbian market. By the 1990s government authorities in a number of countries

were promoting services for lesbian and gay travelers, and Sydney's Gay and Lesbian *Mardi Gras* was recognized as one of Australia's most successful tourist attractions. The gay/lesbian press, originally an expression of the political movement, has become increasingly preoccupied by 'lifestyle' issues and the need to attract 'mainstream' advertizers.

Meanwhile women developed a parallel if sometimes overlapping world, and developed a set of social institutions which allowed room for a 'women-centered' politics, sometimes extended to a rejection of anything that could be seen as male-dominated (Stein 1997). In some countries links were maintained between women and men through common political organisations such as the National Gay (later Gay and Lesbian) Task Force in the USA. The Task Force's commitment to working through mainstream institutions underlines the various splits around gender, political style, and class which became apparent in the movement during the 1970s and 1980s. As the radical energies of the early 1970s dissipated, so too did organized lesbian and gay politics, though the trajectory of the movements is rather different in different countries (Adam et al. 1998). Even in the western world there were significant differences between the English-speaking democracies, where decriminalization of homosexual behavior seemed the top priority; the countries of northern Europe, especially the Netherlands and Denmark, where social acceptance seemed most assured; and those of southern Europe whose laws were less repressive and where a language of universal rights was often used to argue against any sort of identity politics.

The 1970s and 1980s saw a rapid shift in most western countries in attitudes towards homosexuality, with growing recognition of gay men and lesbians as constituting legitimate communities with a certain political influence. Openly lesbian and gay legislators were elected in Norway, Canada, and the USA, and almost all western countries abolished remaining laws which decriminalized homosexual activity. (The great exception was the USA, where homosexual acts remain illegal in a number of states, in part because of a decision by the Supreme Court in 1986 which upheld Georgia's sodomy laws.) The rate of change was particularly rapid in cases such as Spain which democratized during this period, a shift which would be echoed in Latin America in the late 1980s and some parts of Eastern Europe in the 1990s (Long 1998).

The rapidity of changes in attitudes towards homosexuality meant a corresponding backlash from moral conservatives, which was most pronounced in the USA. The mobilization of religious conservatives by groups such as the Moral Majority led to bitter attacks on homosexuality, both in the political arena and, not infrequently, in direct violence. While there were examples of similar reactions elsewhere, such as the Thatcher government's prohibitions on 'the promotion of homosexuality,' it is probably true that the gains for gay and lesbian rights was rapid and irreversible in most western democracies.

3. *The Impact of HIV/AIDS*

The first reports of what was to be named 'AIDS' (acquired immune deficiency syndrome) came with severe illnesses among young homosexual men on both American coasts, and for a short time the new disease was known as 'gay-related immune deficiency syndrome.' Within a few years two major discoveries were made: the syndrome was transmitted through sexual and blood contact, and in many parts of the world the bulk of sexual transmission was heterosexual. Nonetheless the epidemiological link to homosexuals in the rich world has a continuing impact on how the epidemic is perceived, and has had a major impact upon gay organization. The first responses to the new disease came from gay communities in major western cities, with the Gay Men's Health Crisis in New York (founded in 1982) becoming a model for groups such as the various Australian state AIDS Councils, the Terrence Higgins Trust in Britain and the *Deutsche AIDS Hilfe*.

During the 1980s the need to lobby governments, to provide information and preventive education, and to develop home-care and emotional support for those with HIV/AIDS, dominated gay movements in most western countries, and sometimes involved considerable numbers of lesbians as well. As the epidemic was discovered to be growing very rapidly in Africa and parts of Latin America, AIDS was to be the focus for a new wave of homosexual organizing in a number of 'developing' countries. Groups such as Pink Triangle in Malaysia or Triangulo Rosa in Costa Rica came into existence as *de facto* homosexual organizations because of the epidemic (Altman 1994). The development of People With AIDS groups built on the earlier gay movement concept of 'coming out.'

The impact of AIDS on the movement varied from country to country, but overall it meant a much closer relationship between the state and gay organizations, particularly in countries such as Denmark, Switzerland, Canada, and Australia where the epidemic remained largely due to homosexual transmission and national strategies incorporated community responses. At a global level the development of international responses through the Global Program on AIDS and then UNAIDS saw some institutional support for community organizations, including gay ones.

4. *The 'Queer Movement'*

One of the successes of the movement was the development of 'gay and lesbian studies,' particularly in the Netherlands and the USA. In the case of the latter the development of post-modern theory led to

new ways of conceptualizing homosexuality which sought to move beyond fixed notions of identity and to destabilize assumptions about the relations between sexuality and gender. At an academic level this was expressed in the notion of 'queer theory,' which was played out politically in two major ways: the development of a new radicalism and the extension of the movement to include issues of bisexuality and transgenderism. One might note these were both elements of the gay liberation movement of the early 1970s, which had tended to vanish in the mainstreaming of the movement in the 1980s. (See *Masculinities and Femininities*; *Queer Theory*.)

In the USA, and to a lesser extent in the UK, the movement has extended itself to recognize 'bisexual and trans-gendered' people as well as 'gay and lesbian.' Elsewhere this move has been more hotly contested, as in battles around membership criteria for Sydney's Gay and Lesbian *Mardi Gras*. Yet despite the rhetoric, the gay and lesbian movement has continued to function as do most pressure groups in western societies, using a judicious mix of popular protest, lobbying, media propaganda and in some cases direct participation in the electoral process.

In the USA two issues dominated mainstream gay and lesbian politics in the 1990s: the place of gay men and lesbians in the military and arguments about same-sex marriage. The first became a major embarrassment for the Clinton Administration in 1993, when the President's pledge to remove discrimination from the military ran into considerable Congressional opposition. The resulting 'don't ask, don't tell' policy was far weaker than the policies in a number of other western countries.

The first national recognition of same-sex partnerships came in Denmark in 1989 and arguments about state recognition of homosexual relationships have now extended to almost all of Europe. A law proposed in France in 1998 would have recognized cohabiting couples (both hetero- and homosexual) 'who cannot or do not wish to marry,' and was bitterly opposed by the right. The following year the Senate amended the proposal to cover only heterosexual couples. The idea of 'gay marriage' became a major issue in the USA in 1996, following a case in the Hawaii Supreme Court which seemed likely to recognize same-sex marriage as constitutional, following which Congress passed—and President Clinton signed—the Defense of Marriage Act, which would have refused recognition in other states of such a ruling in Hawaii.

5. *The Development of an International Gay/Lesbian Movement*

In 1978 the International Gay (later Gay and Lesbian) Association (ILGA) was formed at a conference in Coventry, England. While ILGA has largely been driven by northern Europeans, it now has member groups from over 70 countries and has organized meetings in several southern cities. Other networks, often linked to feminist and AIDS organizing, have been created in the past two decades, and emerging lesbian and gay movements are increasingly likely to be in constant contact with groups across the world. Such networks argue for a certain universality of homosexual identity, and their major strategy has involved appeals to universal norms of human rights to oppose persecution and discrimination against homosexuals across the world. It is difficult to avoid a certain amount of neo-colonialism in such universalist arguments, as Ann Ferguson has pointed out (Ferguson 1990, pp. 82–4). What is perhaps striking is that while the dominant image of gay and lesbian identity and culture is undoubtedly American, the development of international networks has owed comparatively little to American leadership. On the other hand, the two largest international gay/lesbian 'networks' are probably those based around the Metropolitan Community Church and the Gay Games, both American in origin.

It may be that a political movement is the least likely part of western concepts of homosexual identity to be adopted in many parts of the world, even as they enthusiastically embrace the mores and imagery of western queerdom. The particular form of identity politics which allowed for the mobilization of a lesbigay electoral pressure in countries like the USA, the Netherlands or even France may not be appropriate elsewhere, even if western-style liberal democracy triumphs. The need of western lesbigays to engage in identity politics as a means of enhancing self-esteem may not be felt in other societies. Nonetheless there has been a rapid growth of lesbian and gay organizing during the 1990s in countries as dissimilar as Nicaragua, Slovakia, India, and the Philippines.

The western lesbigay movement emerged in conditions of affluence and liberal democracy, where despite other large social issues it was possible to develop a politics around sexuality which is more difficult in countries where the structures of political life are constantly contested. Writing of contemporary South Africa, Mark Gevisser notes: 'Race-identification overpowers everything else—class, gender and sexuality' (Gevisser 1999). In the same way basic questions of political economy and democratization will impact on the future development of gay/lesbian movements in much of Asia and Africa. Yet in Latin America and Eastern Europe gay/lesbian movements have grown considerably in the past decade, and there are now signs of their emergence in some parts of Africa, for example in Kenya and Zimbabwe where President Mugabe has consistently attacked homosexuality as the product of colonialism. The twin impact of globalization and the discourses of international human rights are likely to see a strengthening of both gay/lesbian identities and rights in the future.

See also: Aboriginal Rights; Feminist Theory: Radical Lesbian; Gay, Lesbian, and Bisexual Youth; Gender and Feminist Studies; Identity Movements; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Masculinities and Femininities; Queer Theory; Rationality and Feminist Thought; Regulation: Sexual Behavior; Sexual Orientation: Biological Influences; Sexuality and Gender; Social Movements and Gender; Social Movements, History of: General; Social Movements: Psychological Perspectives; Social Movements, Sociology of; Transsexuality, Transvestism, and Transgender

Bibliography

Adam B, Duyvendak J W, Krouwel A (eds.) 1998 *The Global Emergence of Gay and Lesbian Politics*. Temple University Press, Philadelphia, PA

Altman D 1971 *Homosexual: Oppression and Liberation*. NY. Outerbridge and Dienstfrey (new edition New York University Press 1996)

Altman D 1994 *Power and Community*. Taylor & Francis, London

Blasius M, Phelan S (eds.) 1997 *We Are Everywhere: A Historical Sourcebook of Gay and Lesbian Politics*. Routledge, NY

d'Emilio J 1983 *Sexual Politics. Sexual Communities*. University of Chicago Press, Chicago

Faderman L 1981 *Surpassing the Love of Men*. Morrow, NY

Ferguson A 1990 Is there a lesbian culture? In: Allen J (ed.) *Lesbian Philosophies and Culture*. State University of New York Press, NY

Gevisser M 1999 Gay life in South Africa. In: Drucker P (ed.) *Different Rainbows*. GMP, London

Hendriks A, Tielman R, van der Veen E 1993 *The Third Pink Book*. Prometheus Books, Buffalo, NY

Kennedy E, Davis M 1993 *Boots of Leather. Slippers of Gold*. Routledge, New York

Long S 1998 Gay and lesbian movements in Eastern Europe. In: Adam B, Duyvendak J W, Krouwel A (eds.) *The Global Emergence of Gay and Lesbian Politics*. Temple University Press, Philadelphia, PA

Plummer K (ed.) 1992 *Modern Homosexualities*. Routledge, London

Rich A 1980 Compulsory heterosexuality and lesbian existence. *Signs* **5**: 4

Russell P 1994 *Sea of Tranquillity*. Dutton, NY

Stein A 1997 *Sex and Sensibility*. University of California Press, CA

Weeks J 1990 *Coming Out: Homosexual Politics in Britain from the Nineteenth Century to the Present*. Rev. edn. Quartet, London

D. Altman

Gellner, Ernest (1925–95)

Ernest Gellner was born in Paris on December 9, 1925 and spent his childhood in Prague. The family moved to England in 1939 ahead of the Nazi invasion. At Oxford, where his studies were interrupted by military service in the Czech Brigade, he specialized in philosophy. After teaching philosophy for two years at Edinburgh, he joined the Department of Sociology at the London School of Economics (LSE) in 1954. Here, stimulated also by his discovery of social anthropology in the department established by Malinowski, he formulated in *Words and Things* (1959) a withering and controversial critique of the linguistic philosophy that was dominant in Oxford at the time. In 1962 he obtained a doctorate in anthropology, based on fieldwork in Morocco. *Thought and Change* (1965) outlined many of the themes that he followed up in later years, including his theory of nationalism. In 1984 Gellner succeeded Jack Goody in the Chair of Social Anthropology at Cambridge, where he was elected to a Fellowship at King's College. In 1991 he returned to Prague to launch the Centre for the Study of Nationalism at the newly established Central European University. He died in Prague on November 5, 1995.

Gellner's interests ranged well beyond the conventional boundaries of the subjects in which he held academic posts. His prolific and diverse writing was securely anchored in his early philosophical studies, particularly in Hume and Kant, augmented later by his reading of Weber, Durkheim, and Malinowski. In short, he drew on classical sources in several disciplines. Negative influences were equally important. Marxism was rejected because of its economic determinism, though much of Gellner's own work emphasized a materialist standpoint and he maintained an enthusiastic curiosity for Soviet Marxist anthropology. Above all, he rejected the relativism of the later Wittgenstein.

Gellner's combination of philosophy and anthropology was far from the continental genre of philosophical anthropology. Rather, he hailed Malinowski as the innovator who made empirical observation in the course of long-term fieldwork the revolutionary basis of modern anthropological knowledge. His fascination with Malinowski led Gellner, late in his career, to explore the founder's intellectual roots in late Habsburg Cracow and to an illuminating (posthumously published) contrast with Wittgenstein (Gellner 1998).

Alongside the functionalist holism of Malinowski, Gellner was also influenced by A. R. Radcliffe-Brown's positivist insistence on the comparison of social structures. His own work prioritized pattern and system, rather than close-up ethnographic analysis of the flux of events. He was interested in all kinds of systems of ideas, and especially in linking ideologies to long-termevolutionary change and to the distribution of economic and political power. Gellner's manner of explicating societies by means of elegant structural models or ideal types was at odds with most of his contemporaries, who stressed interpretation and local knowledge. More controversially, he insisted that *one* culture, that of the modern west, based ultimately upon premises of individualism and ration-

ality, was technologically and 'cognitively' more powerful than all others. This anthropological philosophy (1975, 1985a) was complemented by a three-stage philosophy of history (1991). How to explain the dominance of the west and how best to spread the benefits of that superiority were central themes throughout his work.

Among the more specific research themes addressed by Gellner, the most significant in his lifetime was his theory of nationalism. In reaction to his LSE colleague Elie Kedourie, Gellner emphasized not the dissemination of the nationalist idea, but the functional needs of a modern industrial society, which requires a highly mobile labor force and educational institutions to promote literacy in a single 'high culture.' This model, elaborated in several variations over more than three decades, was not primarily supported with fieldwork materials but with vivid imaginary sketches, notably that of 'Ruritania' (1983). These were evidently much influenced by his own Central European roots, and perhaps by earlier scholars with a similar background, such as Hans Kohn and Karl Deutsch. Gellner's 'modernist' emphasis on the uneven spread of industrialism does not offer a fully satisfactory explanation for the actual spread of nationalism outside Europe, or even within plausible Ruritanias. It offers little help in predicting the circumstances in which national sentiment becomes 'hot.' It remains, however, as even critics have acknowledged, the best general theory of nationalism that any twentieth-century social scientist has devised (Hall 1998).

In his studies of Islam, Gellner supplemented an abstract model with data from many years of fieldwork in the High Atlas Mountains. His monograph, *Saints of the Atlas* (1969), an analysis of the role played by charismatic religious specialists in the maintenance of political order in the absence of state power, is a classic of the British school. Intellectual inspiration was drawn not only from the segmentary lineage theory of the French pioneers and from Evans-Pritchard, but also from the cyclical models of tribal–urban interaction developed centuries earlier by Ibn Khaldun. Gellner's generalization of the 'pendulum swing model' to the wider Moslem world (1981) has not been widely accepted, while even in the Moroccan case his application of segmentary theory has been criticized by later fieldworkers (see Munson, in Hall and Jarvie 1996). However, his schematic view of Islam, with its text-based high culture, as the most modern of the great religions, gave perceptive insight both into the relative absence of nationalism in these countries and the rise of 'fundamentalism.'

The Marxist–Leninist régimes of twentieth-century Eurasia were another form of social organization of abiding concern to Gellner, for reasons that connect closely with his biography. Although he visited the Soviet Union frequently and spent a year in Moscow at the height of *Perestroika*, most of his fieldwork in this case was textual. He delighted in reading between the lines of ethnographers' adaptations of Marxist theories in order to see what light these might throw on Soviet society (1988b). Though he did not predict the demise of the socialist system, his first-hand observations of socialist disenchantment helped him to get to the heart of the matter: no ideology that promised salvation here on this earth was likely to endure if the utopians could neither satisfy accelerating consumer expectations nor control corruption and widening inequalities. In his last years he found the concept of 'civil society' particularly useful in defining a blueprint for a modern democracy that would avoid both the disasters of totalitarian central planning and the tribal 'tyranny of cousins' (1994).

Testimony to Gellner's influence in his lifetime can be found in a remarkable multidisciplinary collection of essays, to which he was able to reply shortly before his death (Hall and Jarvie 1996). Part of the explanation for this fame lies in the topicality of the subjects he treated: nationalism, Islamic fundamentalism and socialism all were vital problems of the age. Another element in his appeal is the allure of a scholar who was not merely radically disrespectful of disciplinary barriers but who consistently opposed the most influential intellectual currents of the day. His critique of Oxford philosophy was followed by equally disrespectful dismissals of many other 'churches,' including Parisian structuralism, Frankfurt neo-Marxism, psychoanalysis (1985b), the 'Orientalism' paradigm of Edward Said, and every variety of 'postmodernism' (1992). Even his relationship with social anthropology, the discipline in which he came closest to finding a home, was often uneasy. He remained basically loyal to the canons of 'structural-functionalism' long after these were abandoned by the profession, yet he combined this with an even more unfashionable sympathy for evolutionist agendas. There were sharp disagreements, for example with Rodney Needham over the 'concept of kinship' (1974) and with Clifford Geertz, whom he recognized as the leading contemporary exponent of the erroneous relativist–idealistic concept of culture.

It is not therefore surprising that neither anthropology nor any of the other disciplines to which he contributed has a Gellnerian school. His diffuse impact owes much to his forthright stance in defence of 'modernist' enlightenment ideals in a century in which these foundations have seemed increasingly fragile to the great majority of the West's social scientists. His temperament and essayistic, aphoristic style were well suited to this role, which of course in the eyes of those he attacked became the role of bogeyman. This stubborn defender of Cartesian ideals of rationality, clarity and *truth* delighted in controversy and had a preference for striking images and analogies over more prosaic social science techniques; but these were an integral part of his appeal. He is best seen not as a backward-looking defender of European traditions, but as a scholar who strove to

adapt a European liberal heritage to reach a better understanding of all the problems of the modern world. Caricatures such as 'enlightenment fundamentalist' are wide of the mark (even if he himself regularly suggested such labels). Despite his criticism of Wittgenstein and of the relativist trend in anthropology, he was no more willing than Malinowski to dispense with the concept of culture. He insists on the uniqueness of modern western cognition because it is stronger than any other, but he does not claim that it is *better*. Compared to Popper and Hayek, this Central European is painfully aware that decent human communities require stronger social foundations than those which liberal individualism has been able to provide. It is this balancing act and the suggestiveness of his ideal types which make Ernest Gellner a continuing source of inspiration.

See also: Determinism: Social and Economic; Historical Explanation, Theories of: Philosophical Aspects; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: The West

Bibliography

Gellner E 1959 *Words and Things, A Critical Account of Linguistic Philosophy and a Study in Ideology*. Gollancz, London

Gellner E 1965 *Thought and Change*. Weidenfeld and Nicolson, London

Gellner E 1969 *Saints of the Atlas*. Weidenfeld and Nicolson, London

Gellner E 1974 *Contemporary Thought and Politics*. Routledge and Kegan Paul, London

Gellner E 1975 *Legitimation of Belief*. Cambridge University Press, Cambridge, UK

Gellner E 1981 *Muslim Society*. Cambridge University Press, Cambridge, UK

Gellner E 1983 *Nations and Nationalism*. Blackwell, Oxford, UK

Gellner E 1985a *Relativism and the Social Sciences*. Cambridge University Press, Cambridge, UK

Gellner E 1985b *The Psychoanalytic Movement, or The Cunning of Unreason*. Fontana Press, London

Gellner E 1988b *State and Society in Soviet Thought*. Blackwell, Oxford, UK

Gellner E 1991 *Plough, Sword and Book*. Paladin Grafton Books, London

Gellner E 1992 *Post-Modernism, Reason and Religion*. Routledge, London

Gellner E 1994 *Conditions of Liberty. Civil Society and its Rivals*. Hamish Hamilton, London

Gellner E 1998 *Language and Solitude. Wittgenstein, Malinowski and the Habsburg Dilemma*. Cambridge University Press, Cambridge

Hall J A (ed.) 1998 *The State of the Nation: Ernest Gellner and the Theory of Nationalism*. Cambridge University Press, Cambridge, UK

Hall J A, Jarvie I (eds.) 1996 *The Social Philosophy of Ernest Gellner*. Rodopi, Amsterdam

C. M. Hann

Gender and Cancer

Incidence rates for cancer, a group of diseases characterized by the rapid and uncontrolled multiplication and spread of abnormal cells, differ for men and women. Explaining this variation requires consideration of the role of both sex (biologically determined factors) and gender (socioculturally determined factors) in the onset, progression, and treatment of the disease. Although cancer etiology and treatment are complex and multidetermined, there is clear evidence that sex differences in anatomy, physiology, and hormone exposure and gender differences in health-related behavior, environmental exposure, and access to health care play central roles.

1. Cancer Incidence Rates for Women and Men

The neoplastic process can occur in any part of the body, with rates of occurrence differing substantially by anatomic organs, between women and men, and across nations. For example, breast cancer is by far the most frequently diagnosed cancer in women worldwide, due largely to high rates of breast cancer in industrialized countries. Cervical cancer, the site with the second highest incidence rate in women, is diagnosed primarily in women from developing countries, who have much lower rates of breast cancer. Lung cancer is the most common site of cancer in men, whereas it is fifth in frequency among women (Boyle and Maisonneuve 1998). Tables 1 and 2 provide the range in rates of the most common cancers for women and men, indicating where the highest and lowest incidence occurs. The prevalence of cancer in sex organs and the wide international variation within sex suggest that both sex and gender are important.

2. Problems in the Research Literature

It is difficult to ascertain the precise contributions of sex- and gender-related variables to the etiology and treatment of cancer for two sets of reasons. First, these variables have rarely been the focus of careful scientific investigation. Until very recently, subjects in research with both animals and humans have been predominantly male. Many studies included only males and, when both sexes were included, they were typically combined into one sample such that the data from the much larger proportion of males were assumed to generalize to females (Sechzer et al. 1994). The very

Table 1
Highest and lowest incidence rate for common cancer sites for women*

	Highest		Lowest	
Cancer site	Population	Rate	Population	Rate
Breast	US bay area: White	104.2	The Gambia	3.4
Cervical	Peru, Trujillo	54.6	Israel: non-Jews	2.6
Colon	Bermuda: Black	34.4	Algeria, Sétif	0.9
Liver	Thailand, Khon Kaan	38.3	Canada, Prince Edward Island	0.1
Lung	New Zealand: Maori	62.2	India, Madras	1.4
Stomach	Japan, Yamagata	42.9	India, Ahmadabad	1.5
ALL SITES	Canada, British Columbia	345.4	The Gambia	39.6

Source: Parkin et al. 1992
*Rates are per 100,000 and are age-adjusted to the world standard.

Table 2
Highest and lowest incidence rate for common cancer sites for men*

	Highest		Lowest	
Cancer site	Population	Rate	Population	Rate
Colon	US, Hawaii: Japanese	37.2	The Gambia	0.7
Liver	Thailand, Khon Kaan China, Qidong	80.0	Netherlands, Maastricht	0.8
Lung	New Zealand: Maori	119.1	The Gambia	1.0
Prostate	US, Atlanta: Black	102.0	China, Qidong	0.8
Stomach	Japan, Yamagata	93.3	India, Ahmadabad	2.1
ALL SITES	Australia, Tasmania	493.8	The Gambia	59.1

Source: Parkin et al. 1992
*Rates are per 100,000 and are age-adjusted to the world standard.

hormonal and physiological differences that appear to be central to some forms of cancer in women were considered unwanted sources of experimental error. The second set of problems is inherent in the questions of interest. Even with increasing attention being paid to sex, and to a lesser extent to gender, it is difficult to identify clear causal relations because of the complex interactions between and covariation of biological sex and sociocultural gender roles. Animal models cannot adequately explore gender, and there is little opportunity to isolate specific effects in human populations. It is difficult to make blanket statements that are accurate for all cultures and all cancers. Cultural values and constraints regarding gender roles differ substantially across societies, and cancers at different sites vary greatly in terms of incidence and mortality rates, epidemiology, risk factors and causes, and treatments (Harras et al. 1996).

3. *Etiology and Prevention*

In addition to the obvious relevance of sexual anatomy to several of the most common forms of cancer, other sex-related variables are involved in cancer etiology (Meyerowitz et al. 2000). Sex hormones often play a central role in the onset and progression of cancers of the reproductive organs and breasts, as well as other sites that have receptors for sex hormones, such as colorectal and laryngeal tissue. Also, the normal physiological variations between women and men in the functioning of virtually all systems of the body have implications for cancer (Legato 1998). For example, the slower colonic transit time and lower excretion of bile acids in women account in part for colon cancer in women, one of the few sex-neutral sites of cancer for which incidence rates for men are not higher than rates for women.

Socioculturally determined factors appear to play an even more important role in cancer etiology than hormones and physiology. It has been estimated that as many as 75 percent of cancer deaths are attributable to behavior, particularly tobacco use and diet (Boyle and Maisonneuve 1998). Women tend to lead healthier lives than men in areas that have been associated with cancer risk. They use less tobacco, drink less alcohol, eat more fruit and vegetables, consume less animal and dairy fat, and apply sunscreen more regularly than men (Meyerowitz et al. 2000). When women's health habits shift to approximate those of men, the gap

between male and female cancer rates shift as well, as seen in the dramatic increase in both smoking and lung cancer rates among non-Hispanic white women in the US over the past few decades.

There are also gender differences in cancer risk due to exposure to environmental carcinogens and, possibly, stress-inducing life circumstances. Occupational segregation in many sectors of the workforce means that women and men are exposed to different toxins and other workplace hazards (Krieger et al. 1993). For example, gender differences in exposure to secondhand smoke in hospitality jobs such as waiting on tables in restaurants or to industrial toxins in factories will vary depending on cultural norms regarding whether these are viewed as primarily jobs for women or for men. Sedentary jobs also have been associated with increased risk of some forms of cancer (Meyerowitz et al. 2000), potentially placing women in clerical jobs at higher risk than men in more physically active lines of work. Additionally, there is growing evidence that stress can influence natural killer cell activity and other aspects of immune functioning that may, in turn, effect cancer onset or progression (Cohen and Herbert 1996). Psychosocial variables that have been associated with immune suppression—inhibition of emotional expression, passive and avoidant coping, lack of social support, and multiple life stressors—can all be influenced by gender-role expectations and cultural values. For example, proscriptions against seeking social support in men and stress associated with filling multiple roles in women have been suggested as important gender differences in the US (Good et al. 2000, Sechzer et al. 1994).

In addition to their individual influences, sex and gender can interact in cancer etiology. The most salient example is cervical cancer, for which a sexually transmitted virus, Human Papilloma Virus (HPV), appears to be a primary causal agent (Boyle and Maisonneuve 1998). Women who have unprotected sex with multiple sex partners or with men who have multiple sexual contacts are at substantially increased risk of contracting the disease. Thus, social norms regarding sexual behavior for women and men and regarding women's rights to demand safe sex practices from their partners will influence cervical cancer rates.

4. *Early Diagnosis and Treatment*

Once a neoplastic process has begun, survival is dependent upon early diagnosis and effective treatment. Given that women are more likely than men to fall below the poverty line, their access to adequate care can be severely limited in countries where a fee-for-service health care model prevails. Even when finances are not a limiting factor, cultural values about the appropriateness of obtaining care can play a role, as is the case in some cultures where it is up to male relatives to determine whether women should receive medical care. Data regarding gender differences in the extent to which individuals who do have access to care adhere to medical recommendations for cancer screening and treatment are mixed, although most studies seem to indicate lower adherence rates for men (Good et al. 2000). Rigid adherence to strictly traditional gender roles for either masculinity, such as an unwillingness to accept physical vulnerability and dependence, or femininity, such as passivity in the face of medical authority or pressures to maintain a caregiver role, can impede optimal care (Good et al. 2000, Sechzer et al. 1994). Even when the individual does not hold these values himself or herself, gender roles can have an influence either through policy decision or implicit biases. For example, male physicians have been found to be less likely than female physicians to recommend breast and cervical cancer screening, whereas they are equally likely to recommend sex-neutral screenings such as blood pressure tests (see *Cancer Screening*).

In addition to the influence of gender roles on diagnosis and treatment, there is some evidence to suggest sex differences. Rates of cancer survival differ between men and women, with women having the survival advantage for most cancer sites (Legato 1998). There also is growing evidence that the pharmacokinetics and drug toxicities of some chemotherapeutic agents differ between women and men, indicating the need for sex-specific prescribing standards (Legato 1998).

5. *Quality of Life*

Psychosocial adjustment to the diagnosis and treatment of cancer is another area in which gender issues are likely to play a role in cancer outcomes. Until recently, almost all of the relevant literature has assessed quality of life among women, with the few studies with men focusing on physical symptoms and functional abilities rather than the emotional or social aspects of adjustment (Meyerowitz and Hart 1995). Therefore, it is not possible to draw clear conclusions regarding the role of gender in adjustment, and further research is required (see *Chronic Illness: Quality of Life*).

6. *Conclusions*

Clearly, values and standards regarding the appropriate behaviors and roles for men and women, both at the level of the individual and the larger society, influence cancer onset and progression. The precise nature of the influence is likely to vary by cancer site and by culture.

Relatively little research has considered the biological, behavioral, and healthcare pathways through which gender plays a role. Methodological and conceptual refinements are needed in order to allow future research to expand beyond simple description to

consider multiple causal pathways within a multidisciplinary biopsychosocial model.

See also: Biopsychology and Health; Cancer-prone Personality, Type C; Cancer: Psychosocial Aspects; Gender and Cardiovascular Health; Gender and Health Care; Gender and Physical Health; Men's Health; Women's Health

Bibliography

Boyle P, Maisonneuve P 1998 Cancer prevention in women: Clues from epidemiology. In: Kavanagh J J, Singletary S E, Einhorn N, DePetrillo A D (eds.) *Cancer in Women*. Blackwell, Malden, MA

Cohen S, Herbert T B 1996 Health psychology: Psychological factors and physical disease from the perspective of human psychoneuroimmunology. *Annual Review of Psychology* **47**: 113–42

Good G E, Sherrod N B, Dillon M G 2000 Masculine gender role stressors and men's health. In: Eisler R M, Hersen M (eds.) *Handbook of Gender, Culture, and Health*. Erlbaum, Mahwah, NJ

Harras A, Edwards B K, Blot W J, Ries L A G 1996 *Cancer Rates and Risks* (NIH Publication No. 96-961). US Department of Health and Human Services, Public Health Service, and National Institutes of Health, National Cancer Institute, Bethesda, MD

Krieger N, Avery B, Rowley D L, Phillips M T, Herman A A 1993 Racism, sexism, and social class: Implications for studies of health, disease, and well-being. *American Journal of Preventative Medicine* **9** (6 Suppl.): 82–122

Legato M J 1998 Women's health: Not for women only. *International Journal of Fertility* **43**(2): 65–72

Meyerowitz B E, Bull A A, Perez M A 2000 Cancers common in women. In: Eisler R M, Hersen M (eds.) *Handbook of Gender, Culture, and Health*. Erlbaum, Mahwah, NJ

Meyerowitz B E, Hart S 1995 Women and cancer: Have assumptions about women limited our research agenda? In: Stanton A L, Gallant S J (eds.) *The Psychology of Women's Health: Progress and Challenges in Research and Application*. American Psychological Association, Washington, DC

Parkin D M, Muir C S, Whelan S L, Gao Y T, Ferlay J, Powell J 1992 *Cancer Incidence in Five Continents*, Vol. VI. IARC Scientific Publication No. 120. World Health Organization, International Agency for Research on Cancer, Lyon, France

Sechzer J A, Rabinowitz V C, Denmark F L, McGinn M F, Weeks B M, Wilkens C L 1994 Sex and gender bias in animal research and in clinical studies of cancer, cardiovascular disease, and depression. *Annals of the New York Academy of Sciences* **736**: 21–48

B. E. Meyerowitz

Gender and Cardiovascular Health

Despite the decline in heart disease in most Western industrialized countries, the disease remains a major health problem and has reached epidemic proportions in many countries of Eastern Europe, particularly among men (Bobak and Marmot 1996, Weidner et al., in press). Typically, the category heart disease includes nonrheumatic heart disease, hypertension, and coronary heart disease (CHD). The latter comprises the major subgroup (at least 80 percent of heart disease). The focus of this article will be on this major subgroup of heart disease (see *Coronary Heart Disease (CHD): Psychosocial Aspects*).

1. The Gender Gap in CHD

Data from the Framingham Heart Study indicate that the lifetime risk of developing CHD at age 40 is one in two for men and one in three for women (Lloyd-Jones et al. 1999). Recent international CHD mortality data confirm that CHD is lower in women than in men, yielding female:male ratios ranging from 0.1 to 0.6, depending on country (Jackson et al. 1997). At first glance, this universal female advantage in countries differing in heart disease rates and lifestyles points to protective factors of a biological nature (i.e., estrogen) (see *Gender and Physical Health*). However, several studies have cast doubt on the role of sex hormones in the protection from CHD in women (Barrett-Conner 1997). Furthermore, the differences between countries are larger than the difference between the sexes, suggesting that 'male anatomy is not destiny' with regard to CHD. Also, male excess mortality from cardiovascular disease has not always been present throughout history (Nikiforov and Mamaev 1998). Finally, the recent epidemic of cardiovascular disease among young and middle-aged Eastern European men has widened the gender gap over a very brief time span. All these phenomena indicate that nongenetic factors influence the gender gap in heart disease (Weidner 1998).

2. Why Do Men Get More Heart Disease?

2.1 The Role of Traditional Risk Factors

Considerable progress has been made in the identification of physiological and behavioral risk factors for CHD. The major risks are elevated plasma cholesterol, low HDL-cholesterol, hypertension, diabetes, cigarette smoking, and abdominal obesity. Generally, men's levels of these risk factors are less favorable than those of women's (Barrett-Connor 1997). This raises the question of whether gender differences in traditional risk factors can explain the observed gender difference in CHD mortality. Analyses of data from the World Health Organization coordinated MONICA Project and the US Atherosclerosis Risk in Communities Study indicate that only 40 percent of the variation in the gender ratios of CHD mortality in 24 countries could be explained by gender differences

in five major CHD risk factors (i.e., high plasma cholesterol, low HDL-cholesterol, high blood pressure, cigarette smoking, and general obesity; Jackson et al. 1997).

In this regard, it is of interest to note that the traditional risk factors are not good candidates to account for the cardiovascular disease epidemic in Eastern European men either. For example, Kristenson et al. (1998) compared traditional and psychosocial risk factors of 50-year-old Swedish men with age-matched Lithuanian men, who are four times more likely to die from CHD than their Swedish counterparts. There were hardly any group differences in traditional risk factors. However, Lithuanian men were more likely to report greater psychosocial strain (e.g., depression, low social support, job strain) than Swedish men. This raises the possibility that psychosocial coronary risk factors do not only contribute to the recent increase in CHD among men in Eastern Europe, but may also account for some of the variance in the gender differential in heart disease.

2.2 *The Role of Psychosocial Risk Factors*

The impact of psychosocial factors on the etiology and prognosis of coronary heart disease has been summarized recently (Hemingway and Marmot 1999, Rozanski et al. 1999). These factors include coronary-prone personality attributes such as Type A and hostility, job strain, lack of social support, negative emotions (particularly depression), and low socioeconomic status. Gender-specific associations with CHD for these variables have been summarized previously (Weidner 1995, Weidner and Mueller 2000, Orth-Gomer and Chesney 1997, respectively).

Overall, men appear to be at a disadvantage when considering psychosocial risk factors. For example, they score higher on Type A behaviors and hostility than women (see *Coronary-prone Behavior, Type A*). Both of these attributes are characteristics of the male ('macho') gender role, which has been linked to behavioral risk factors, such as smoking and excessive alcohol consumption (Waldron 1997) as well as other health risks (Courtenay 2000).

Men not only report less social support than women, but also appear to have fewer sources of social support, thus increasing their dependency on a single source (see *Social Support and Stress*). The Male Aging and Women's Health Studies in Massachusetts examined sources of social support among mid-aged women and men, including support from spouse, relative, friend, and co-worker. Men were more than twice as likely as women to name their spouse or their partner as their primary provider of social support (65.5 percent vs. 26.4 percent). Furthermore, 24.2 percent of men (but only 6.1 percent of women) said this was their only source of support! These data may, in part, explain why men's health is more seriously affected by partner loss through separation, divorce, or widowhood. Finally, men were twice as likely as women to report no supporters (9.2 percent vs. 4.3 percent; New England Research Institutes 1997).

At first glance, gender differences in negative emotions appear to favor men (see *Emotions and Health*). In most studies, women report more negative emotions such as depression than men (although this is not consistently found in populations where women and men have similar roles, such as among college students). Although women may report more depression, they appear to be coping more effectively than men. Generally, men are more likely to use avoidant coping strategies, such as denial and distraction, whereas women are more likely to employ vigilant coping strategies, paying attention to the stressor and its psychological and somatic consequences (Weidner and Collins 1993). Which style is more adaptive depends largely upon the situation. Most stressful experiences consist of uncontrollable daily hassles, which are short-lived and typically of no great consequence. Here, avoidant strategies would be more adaptive ('what I cannot control and what can't hurt me, is best to be ignored'). Thus, men's strategies are likely to pay off for these types of events, contributing to their lesser experience (or report) of emotional discomfort or distress. But what if disaster hits? Avoidance of uncontrollable events—such as divorce, loss of a job or a loved one, sudden financial crisis and economic decline—may not be adaptive in the long run. Here it may be women's greater vigilance that is more adaptive: preparing for the crisis, seeking help, advice, etc. Consistent with this reasoning are recent data from the Hungarian population showing that women tend to accept their negative mood as a disorder to be treated, whereas men are more likely to engage in self-destructive behavior, such as excessive alcohol consumption (Rethelyi et al., in press). Similarly, research on how people cope with disasters (e.g., hurricanes and tornadoes) supports the notion of men's maladaptive coping: increases in alcohol consumption and depression were related to personal disaster exposure among men, whereas no such direct relationship was evident among women (Solomon et al. 1987). Thus, in addition to an unfavorable traditional coronary risk factor profile, men's psychosocial coronary risk factors seem to further enhance their susceptibility to CHD.

2.3 *Gender-specific Associations Between Risk Factors*

Another major question concerns the interrelationship among traditional, behavioral, and psychosocial risk factors. It is conceivable that gender-specific associations among risk factors may be stronger among men than among women, that is, an increase in one factor may elevate another factor in a gender-specific fashion.

There is some indication that this indeed may be the case. For example, job strain appears to be associated with increases in health-damaging behaviors (e.g., cigarette smoking, excessive alcohol and coffee consumption, lack of exercise) among men, but not among women (Weidner et al. 1997). Similarly, men respond with exaggerated cardiovascular reactions to a wider range of environmental stressors than women. A male ('macho') sex role orientation is associated with decreased motivation to learn stress management skills in men, but not in women (Sieverding 1997). Socioeconomic deprivation and depression are more closely related to mortality in men than in women (Kopp et al., in press).

In contrast, men appear to benefit more from social support (i.e., decreased cortisol response to stress) provided by their partner than do women (Kirschbaum et al. 1995, also see Orth-Gomer and Chesney 1997). Similarly, the association between physical activity and CHD appears to be stronger for men than for women. This latter finding, however, may be due to the use of questionnaires designed for men, which focus on sports and neglect physical activities associated with house work (Barrett-Connor 1997).

3. Conclusions

Women score better on traditional and psychosocial coronary risk factors when compared with men (see *Men's Health*; *Women's Health*). There is also some indication that associations between risk factors may be stronger among men than among women. While gender differences in traditional risk factors appear to contribute substantially to the gender gap in heart disease, we currently do not know to what extent psychosocial risk factors contribute to this gap. Most population studies do not include detailed assessments of psychosocial coronary risk factors, which make analyses of the sort conducted by Jackson and his colleagues impossible. It is likely that gender differences in psychosocial risk factors will explain additional variance in the gender ratios of CHD mortality. Furthermore, currently ongoing research on 'new' coronary risk factors, such as abdominal obesity, cardiovascular stress reactivity, and mechanisms related to antioxidant state in women and men promises to further illuminate factors contributing to the gender gap in heart disease (Weidner et al., in press).

See also: Coronary Heart Disease (CHD): Psychosocial Aspects; Coronary-prone Behavior, Type A; Emotions and Health; Gender and Health Care; Gender and Physical Health; Gender Role Stress and Health; Men's Health; Social Support and Health; Social Support and Stress; Stress and Coping Theories; Women's Health

Bibliography

Barrett-Connor E 1997 Sex differences in coronary heart disease: Why are women so superior? The 1995 Ancel Keys Lecture. *Circulation* **95**: 252–64

Bobak M, Marmot M 1996 East–West mortality divide and its potential explanations: Proposed research agenda. *British Medical Journal* **312**: 421–25

Courtenay W 2000 Constructions of masculinity and their influence on men's well-being: A theory of gender and health. *Social Science & Medicine* **50**: 1385–1401

Hemingway H, Marmot M 1999 Psychosocial factors in the aetiology and prognosis of coronary heart disease: Systematic review of prospective cohort studies. *British Medical Journal* **318**: 1460–7

Jackson R, Chambless L, Higgins M, Kuulasmaa K, Wijnberg L, Williams D 1997 Gender differences in ischemic heart disease and risk factors in 46 communities: An ecologic analysis. *Cardiovascular Risk Factors* **7**: 43–54

Kirschbaum C, Klauer T, Filipp S-H, Hellhammer D H 1995 Sex-specific effects of social support on cortisol and subjective responses to acute psychological stress. *Psychosomatic Medicine* **57**: 23–31

Kopp M S, Skrabski A, Szekely A 2001 Risk factors and, inequality in relation to morbidity and mortality in a changing society. In: Weidner G, Kopp M S, Kristenson M (eds.) *Heart Disease: Environment, Stress, and Gender*. IOS Press, Amsterdam

Kristenson M, Kucinskiene Z, Bergdahl B, Calkauskas H, Urmonas V, Orth-Gomer K 1998 Increased psychosocial strain in Lithuanian versus Swedish men: The LiVicordia Study. *Psychosomatic Medicine* **60**: 277–82

Lloyd-Jones D M, Larson M G, Beiser A, Levy D 1999 Lifetime risk of developing coronary heart disease. *The Lancet* **353**: 89–92

New England Research Institutes 1997 Gender differences in social supports: Data from the Massachusetts Male Aging Study and the Massachusetts Women's Health Study. *Network* **Spring/Summer**: 12

Nikiforov S V, Mamaev V B 1998 The development of sex differences in cardiovascular disease mortality: A historical perspective. *American Journal of Public Health* **88**: 1348–53

Orth-Gomer K, Chesney M A 1997 Social stress/strain and heart disease in women. In: Julian D G, Wenger N K (eds.) *Women and Heart Disease*. Martin Dunitz, London pp. 407–20

Rethelyi J M, Purebl G, Kopp M S (in press) Sociodemographic and behavioral correlates of depression in Hungarian women and men. In: Weidner G, Kopp M S, Kristenson M (eds.) *Heart Disease: Environment, Stress, and Gender*. IOS Press, Amsterdam

Rozanski A, Blumenthal J A, Kaplan J 1999 Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. *Circulation* **99**: 2192–2217

Sieverding M 1997 Die bedeutung von prototype-matching für präventives verhalten: ist die teilnahme an stressbewältigungskursen, unmännlich? ('The significance of prototype matching for preventive behavior: Is participation in stress-coping courses, unmanly?') *Zeitschrift für Gesundheitspsychologie* **5**: 272–89

Solomon S D, Smith E M, Robins L N, Fishbach R L 1987 Social involvement as a mediator of disaster-induced stress. *Journal of Applied Social Psychology* **17**: 1092–1112

Waldron I 1997 Changing gender roles and gender differences in health behavior. In: Gochman D S (ed.) *Handbook of Health*

Research I: Personal and Social Determinants. Plenum, New York
Weidner G 1995 Personality and heart disease in women: Past research and future directions. *Zeitschrift für Gesundheitspsychologie* **1**: 4–23
Weidner G 1998 Gender gap in health decline in East Europe. *Nature* **395**: 835
Weidner G, Boughal T, Connor S L, Pieper C, Mendell N R 1997 The relationship of job strain to standard coronary risk factors in women and men of the Family Heart Study. *Health Psychology* **16**: 239–47
Weidner G, Collins R L 1993 Gender, coping, and health. In: Krohne H W (ed.) *Attention and Avoidance*. Hogrefe & Huber, Seattle, WA pp. 241–65
Weidner G, Kopp M S, Kristenson M (eds.) 2001 *Heart Disease: Environment, Stress, and Gender*. IOS Press, Amsterdam
Weidner G, Mueller H 2000 Emotions and heart disease. In: Goldman M B, Hatch M C (eds.) *Women and Health*. Academic Press, San Diego, CA pp. 789–96

G. Weidner

Gender and Environment

The environment is very much a product of the interplay between societies and the natural world they inhabit. Simply by existing, human beings affect and modify this natural world and render it useful (or otherwise) to society. However, society is notoriously segmented and it follows that different sectors of society will experience the environment differently and will have different degrees of control and influence over it. Entries on environment and development, environmental justice, sustainable development, and environmental social movements explore how differences in race, ethnicity, class, and state of development shape individuals' and communities' relationships with the environment. This section extends this analysis to focus on how gender influences that relationship.

Society is powerfully structured by gender relations that conspire to render women less powerful than men. (The entry on gender defines what the term means and the importance of recognizing its social construction.) These social relations, which are explained variously by reference to class and to patriarchy, result in different, and hierarchical roles for, and opportunities available to, women and men, and structure their relationship with the environment, as well as with each other. Whilst these relationships have persisted, little has changed in structure, if not magnitude, through time. They have only been made explicit in the latter part of the twentieth century, largely due to intersecting debates on development, ecology, and feminism since the 1970s. This article will explore these intersections to show how gender affects the relationships between individuals and communities and the environment, at a variety of scales.

1. Feminist Research

Feminist scholarship has exposed inequalities between men and women which result in an uneven distribution of resources. This relationship has consigned women to the domestic (unpaid) sphere, where they enjoy little economic and political independence. This sphere is notoriously undervalued, with no recognition given to the economic and social value of unpaid work in national accounting procedures (Waring 1988). However, while women are increasingly entering the paid labor force, they do so in a multiply disadvantaged position, which has implications for their relationship with the environment. Most women are paid less, pro rata, than men, through a system of horizontal (through the gendering of different sectors, such as mining or teaching) and vertical (through the gendering of jobs within sectors, such as clerical or managers) occupational segregation, as well as through promotion inequalities. In addition, women engage in this work (full-time and part-time) in addition to their domestic responsibilities, which men have been less willing to share.

Class is often used to explain economic and occupational inequality, and Marxist analysis suggests that women have formed a reserve army of labor which is maneuvered in and out of the labor market in tune with demand. However, this explanation does not sufficiently explain inequalities in the home, nor do they explain systematic male violence towards, and control over women; factors which might better be accounted for by patriarchy. These two systems of domination (class and patriarchy), however, are increasingly being seen as interacting and mutually reinforcing (McDowell 1986).

1.1 Implications for Gender–Environment Relations

This 'double bind' of relatively low paid work outside the home, and responsibility for most domestic tasks within it, exposes women to a number of environmental problems. People on low incomes and in relative poverty are more likely to be exposed to high levels of pollution and other negative environmental effects (see *Environmental Justice*), and since women comprise a majority of adults in poverty, they are more likely to experience environmental problems. For example, DDT, banned in the West, is still used in 2001 as a pesticide in the developing world where women undertake the majority of agricultural work (80 percent worldwide—although women hold title to only 1 percent of the world's land). This chemical accumulates in body fat and is passed from mother to her breastfed children causing, amongst other things, breast cancer, endometriosis and reduced fertility in female offspring. Research in parts of Central Asia found that the Republic of Karahalpakstan (close to the heavily polluted Aral Sea) experienced the highest levels of infant mortality in the former Soviet Union.

Poor diet and hygiene, concomitant with poverty, were found to be contributory factors to mortality and infant anemia (Gabizon 1998).

DDT and other hormone disrupters provide a disturbingly clear example of how environmental manipulation can have a profound effect on the nature of sex itself. Estrogen (produced naturally, by the excretion of estrogen-based contraceptives, but most intensely by PCBs, surfactants used in cleaning materials, and phlalates used in the manufacture of flexible plastics) is being associated with a major drop in fertility and an increase in reproductive disorders. Transexuality and hermaphroditism in fish is being considered as a direct result of estrogen pollution.

The task of social reproduction accorded to women (unlike the paid tasks of production) are given no value or recognition by national economies (for example, in GNP or GDP) or by the United Nations, which specifies components of national accounting systems (Waring 1988). (Gross national product (GNP) comprises gross domestic product (GDP) plus net income from abroad. GDP is agdmonetary measure of the value, at market prices, of goods and services produced within a national economy over a given period of time, usually one year (Johnston et al. 2000)). Yet such tasks are made increasingly difficult by environmental exploitation. In the 'Global South' or Third World, in addition to undertaking most of the agricultural work, women are the main water collectors, which can involve journeys on foot of up to four hours each day. Practices that deplete water sources (such as cash crop irrigation) increase this time and energy burden on women, as does deforestation (for commercial farming, mineral extraction, and infrastructure development) for women who can spend four to five hours each day collecting fuel wood when once they would have done so every four or five days (Dankleman and Davidson 1988). In addition, women are carrying increasingly heavy loads of up to 35 kilograms, which is up to 15 kilograms over the acceptable limit laid down by the International Labor Organization Maximum Weight Convention.

In the South, women are exclusively responsible for food preparation, on which they can spend four hours each day. Cooking with firewood (the most common cooking fuel in rural areas) causes inhalation of chemicals equivalent to inhaling the smoke of twenty packs of cigarettes a day. Pollution is even greater when cooks are forced to rely on firewood substitutes, such as husks, straw, and dung, because of the commercial degradation of forests.

In the West, while women do not have the burden of collecting fuel sources and water, they are still primarily responsible for food provision and preparation and, therefore, are more directly concerned with food safety issues such as genetically modified ingredients, BSE, and pesticide use. This is illustrated by the proportion of local environmental and health campaigners who are women. (Joni Seager (1993) estimates that 60–80 percent of grass-roots environmental activists are women, although this proportion drops drastically for paid/professional appointments in environmental nongovernmental organizations [NGOs].)

In addition to social roles, cultural practices create an additional burden on women in societies in which sex can determine survival. The Chinese practice of limiting families to one child in an attempt to keep within the state's environmental carrying capacity, has resulted in a high female fetus abortion rate and the abandonment of girl children (Bandarage 1997). Likewise, in food deprivation situations Dreze and Sen (1989) have found that women are most likely to go without food. Other research has noted that women are also least likely to receive warm clothing and medical assistance and has confirmed that even pregnant and lactating mothers are last in food priorities (Bandarage 1997).

Because women have largely been confined to the domestic sphere, they have had little formal input into the way in which societies' relationship with the environment has been structured. The practice of science has been particularly masculine, both in that it is undertaken by men, and in its underlying principles aiming to control nature, although this approach is increasingly challenged with science. Even at the start of the twenty-first century, the majority of scientists and science students are men and, it could be argued, this continues to shape the subjects studied (for example, military applications) and the methodologies used (Harding 1986, Haraway 1988).

Other environmental professions, such as planning, civil engineering and architecture, political decision making fora, and business, continue to be male-dominated and masculinist and so women have few opportunities for challenging or contributing to decisions made which have profound implications for the environment itself, and for gender–environment relations, which are then perpetuated.

2. *Ecofeminism*

Feminist politics and scholarship have widely aired aspects of this debate, particularly around inequalities of income and opportunity, the uneven division of labor, and the lack of regard for unpaid work. Linking this feminist debate with environmental issues, and their differential impacts, is the ecofeminist movement that emerged in the 1970s out of the parallel social movements of feminism, ecology, and peace/nuclear disarmament. Its initiators were concerned that economic and political gender equality, without due regard for ecological concerns, would result in doubling the burden the environment would have to bear as women emulated men's behavior. The incorporation of women into 'a man's world' would be unlikely to

challenge research priorities, management practices, social values, or consumption patterns. Instead, ecofeminists argued, society needs to reassess its practices and values and particularly to value the more feminine principles of cooperation and care.

Although the term 'ecofeminism' was coined by a Frenchwoman, Francoise d'Eubonne, the thinking behind it has developed across Europe, North America, and Australasia as both an analysis of socioeconomic relations and a prescription as to how these relations can be transformed (Mary Mellor (1992) Carolyn Merchant (1996) and Val Plumwood (1993) respectively). There are two broad strands of ecofeminism: cultural or essentialist (which tend to be more enthusiastically pursued in North America) and social or constructivist (which dominates European thinking).

Cultural ecofeminism celebrates a powerful and positive link between women and nature, arguing that women are more closely in touch with nature through their reproductive potential or capacity. This connection is then thought to better place women (as opposed to men) as advocates for nature. Social ecofeminists do not believe that the essence of what is female (if such an essence actually exists) makes women inherently closer to nature than men. However, they argue that because women are subjugated by a society dominated by men (which also dominates nature), this experience of domination, and consignment to roles and jobs which are less abstracted from nature, and by their dominated status which exposes them more to environmental hazards, places them closer to nature than men who, conversely, are more likely to take on roles and jobs which distance themselves from nature.

These positions, both broadly underpinned by conceptual frameworks based on patriarchy, have led to a debate about the legitimacy of essentialism (Biehl 1991), the need to transcend dualisms of maleness versus femaleness to create a partnership ethic with nature (Merchant 1996, Plumwood 1993), the need to distinguish between different groups of women, and to challenge the unity of their experience (Jackson 1993).

3. *Gender and Development*

A third key influence on the development of gender and environment as an important relationship has emerged from the Development literature and experience, some of which has already been referred to. This is gradually replacing the arguably tokenistic 'women in development' approach to development project planning. The 'women in development' approach tended to 'bolt' women's concerns on to existing policies rather than questioning these policies and the complex social interactions between women and men which underpinned them. This inevitably affects environment/development projects. Taking these social interactions into account has resulted in a number of projects and initiatives, such as the Grameen Bank and other microcredit schemes primarily for women in poverty, which also have the potential to improve society–environment relations. Such microcredit schemes are being identified, not only as economically empowering for women, but as giving women more control over their lives in other ways, such as fertility control. Such experiences are slowly penetrating organizations which have the power to impact on society/gender relations, such as The World Bank and the United Nations, and are beginning to affect their lending policies and practices.

4. *The United Nations and Sustainable Development*

The United Nations, particularly through the *United Nations Conference on Environment and Development* in 1992 and the *World Conference on Women* in 1995, argues that there is a close relationship between the status of women and the state of the environment. It has signaled that the education of women, and the eradication of poverty (targeted at women) are both critical for improving women's status and to reduce societies' impact on the environment. *Agenda 21,* agreed at UNCED 1992, dedicated a chapter (Chap. 24) to the need to bring women into environmental decision-making. This is because women are disproportionately affected by negative environmental impacts due to their social and domestic roles and their greater tendency to be in poverty. UNCED acknowledge that these roles lead to a closer relationship with, and knowledge of the environment, whether as farmers, food and meal providers, or primary health carers. There is, however, little evidence as yet that there are any substantial changes in decision-making locally as a result of this global agreement.

5. *Conclusion*

Any discussion on gender and environment demands an understanding of the ways in which society is structured along gendered lines. In turn, this requires an understanding of how society creates, reproduces, and mobilizes gender differences. The 'gender and environment' relationship, therefore, is a contested concept which turns on various understandings of gender and gender relations and brings into question the ability to totalize the concept of gender. Whilst this entry has focused on the modernist explanatory concepts of class, and particularly, patriarchy as underpinning gender–environment relations, these relations can be further problematized by recognizing the difficulties in defining gender as a single, undifferentiated, entity. Some critics of ecofeminism draw on the literature of post-modernism to question the validity of treating all men and all women similarly, arguing that class, race, ethnicity, age, disability,

sexuality, and parenthood all cut across gender to render it inadequate as a single explanatory tool. It is important to remember the differences within as well as between genders. Research consistently shows that, while there are clearly differences in income, occupation, and opportunity depending on other factors, the relative differences between most men and most women in many societies are still significant enough to warrant concern about gender inequality and therefore the gendered impact of, and on, environmental issues.

See also: Development: Sustainable Agriculture; Ecology, Cultural; Economic Development and Women; Environment and Common Property Institutions; Environment and Development; Environmental and Resource Management; Environmental Cognition, Perception, and Attitudes; Environmental Justice; Environmental Planning; Environmental Psychology: Overview; Feminist Theory: Ecofeminist and Cultural Feminist; Gender and Place; Human–Environment Relationships; Land Rights and Gender; Social Movements: Environmental Movements; Social Stratification; Sustainable Development

Bibliography

Bandarage A 1997 *Women, Population and Global Crisis*. Zed Books, London

Biehl J 1991 *Rethinking Ecofeminist Politics*. South End Press, Boston, USA

Buckingham-Hatfield S 2000 *Gender and Environment*. Routledge, London

Dankleman I, Davidson M 1988 *Women and the Environment in the Third World, Alliance for the Future*. Earthscan in association with IUCN, London

Dreze J, Sen A 1989 *Hunger and Public Action*. Oxford University Press, Oxford, UK

Gabizon S 1998 A dying sea and a dying people. In: *Women, Health and Environment*. Women in Europe for a Common Future. Utrecht, The Netherlands

Haraway D 1988 Situated knowledges: the science question in feminism and the privilege of partial perspective. In: Keller E F, Longino H E (eds.) 1996 *Feminism and Science*. Oxford, University Press, Oxford, UK

Harding S D 1986 *The Science Question in Feminism*. Cornell University Press, Ithaca, NY

Jackson C 1993 Women/nature or gender/history? A critique of ecofeminist 'Development', *Journal of Peasant Studies* **20**(3): 389–419

Johnston R J, Gregory D, Pratt G, Watts M 2000 *The Dictionary of Human Geography*. Blackwell, Oxford, UK

McDowell L 1986 Beyond patriarchy: a class-based explanation of women's subordination. *Antipode* **18**: 311

Mellor M 1992 *Breaking the Boundaries, Towards a Feminist Green Socialism*. Virago Press, London

Merchant C 1996 *Earthcare. Women and the Environment*. Routledge, London

Plumwood V 1993 *Feminism and the Mastery of Nature*. Routledge, London

Seager J 1993 *Earth Follies: Feminism, Politics and the Environment*. Earthscan, London

United Nations Commission on Environment and Development 1992 *Agenda 21*. UNCED, Geneva, Switzerland

Waring M 1988 *Counting for Nothing, What Men Value and What Women are Worth*. Allen & Unwin, Wellington, New Zealand

S. Buckingham-Hatfield

Gender and Feminist Studies

'Gender,' as the term is used in the social and behavioral sciences, refers to all of the ways in which being a male or female, or being socially classified as such, affects one's life. The focus is on the social, political, or economic arrangements that relate to gender. Some gender differences stem from biological differences, and are studied by biologists or others focusing on evolutionary theories. Many facets of gender inequality flow from social norms, beliefs, laws, and institutional practices. Even gender differences rooted in biology will have different consequences depending on the social environment. For example, the fact that children are born from women's bodies is a biological reality, but societies and time periods vary greatly in the consequences of this fact. The consequences depend upon whether birth control is widely available, whether women can use contraception without men's permission, how much fathers participate in the care of children, how much the state collectivizes the costs of rearing children, the extent to which jobs accommodate the demands of parenting, and other factors.

The topic of gender was largely ignored in the social sciences until the 1970s. When relationships and inequalities between men and women were studied in earlier decades, many scholars 'naturalized' existing social patterns. This was true even among social scientists who advocated social explanations for most other forms of social differentiation and inequality. Since 1970 there has been a flowering of scholarship on gender. The increased interest in the topic flows in part from the increased representation of women among social scientists; although there are both women and men among prominent scholars of gender, proportionately more women than men have dedicated research careers to the topic of gender.

The term 'feminist studies' has many meanings in the social sciences and humanities. Some proponents (as well as some opponents) of feminist studies see the term to imply adopting women's point of view. Opponents believe this implies that feminist scholarship is more biased against finding the truth than conventional scholarship. In contrast, most feminist scholars believe that scholarship on gender, men, and women has been, and often still is, distorted by androcentric (male-centered) bias. Some people see the appropriate corrective as an adoption of women's

standpoint; those who view it this way doubt that there is any way the scientific method can insure objectivity. Others see feminist studies as correcting androcentric bias, not by adopting women's point of view, but by taking a less biased and broader view than conventional scholarship.

Feminist theory is a similarly broad and contested term. (See *Feminist Theory*) It generally refers to theories of women's experience and of the subordination of women by men. It is commonly taken to include both what philosophers call 'positive' and 'normative' claims. The positive claims are descriptions and explanations of how gender systems work; they are the conceptual and theoretical part of empirical gender studies. Normative claims answer philosophical questions about how gender arrangements ought to be; they deal with questions of ethics and social justice. Philosophers (including political theorists) wrote much of the normative feminist theory of recent decades.

This essay discusses three prominent themes in gender and feminist studies. Scholars have contributed positive and normative theorizing and empirical scientific research on each. They are illustrative of work in the area rather than being a summary of the field. The three themes are the exclusion of women from well-rewarded positions, the devaluation of activities and qualities associated with women, and the intersection of gender with inequalities of class, race, and nation.

1. Women's Exclusion from Positions of Power and Resources

The concentration of men in positions that bring them power and resources is an important theme in gender studies. Women and men typically hold different positions. The fact that relatively few men engage in typically female activities and the fact that relatively few women engage in typically male activities are both constitutive of gender differentiation. However, research has focused on the latter. In most societies, men have monopolized positions that give their incumbents the highest status, power, and rewards. Understanding the forces that keep women out of these positions is crucial for understanding sources of gender inequality. Social scientists from several disciplines have studied determinants of women's entrance into 'male' spheres. Anthropologists studying pre-industrial societies have concluded that cultures that are more competitive and combative are more likely to exclude women from positions of power (Sanday 1981). Economists contend that rising wage rates during the twentieth century increased female employment and lowered fertility, because higher potential wages made employment more lucrative for women (Bergmann 1986). When higher wages are available to women, their incentive to be employed is greater; the opportunity cost of staying at home is higher.

Although women's participation in paid work has increased in all industrial societies; jobs have remained somewhat sex segregated. (See *Sex Segregation at Work*.) Some economists explain the exclusion of women from male jobs in terms of 'statistical discrimination.' This term refers to using an easily observable characteristic such as sex to make inferences about harder-to-measure qualifications. The idea is that employers have some knowledge of average gender differences in the population, and, based on this information, treat all women as if they were like the average woman. For example, if the average woman doesn't have the skills to be an auto mechanic, employers may assume that this is true of every woman, and be unwilling to consider even qualified women. Social psychologists add to this that perceptions may be systematically biased. Most people exaggerate the size of gender differences. Sociologists have focused on other discriminatory barriers to women's entry into jobs typically held by men. These include harassment of women entering male jobs, as well as political maneuvering to ensure policies that will keep women out of their fields. Sociologists, anthropologists, and psychologists point out that cultural beliefs and socialization throughout the life cycle encourage both men and women to enter stereotypic jobs for their own sex. In some periods and places, laws have prohibited women's entry into many positions of power.

Liberal feminist theory focuses on this exclusion of women from well-rewarded positions. (See *Feminist Theory: Liberal*.) This focus follows from the classical liberal value placed on equality of opportunity. The exclusion of women from male spheres is a clear violation of equality of opportunity. In Western thought, philosophers had traditionally justified the exclusion of women from powerful, esteemed positions in the economy and polity by arguing that women are innately inferior to men. Liberal thought locates the basis of rights to equal treatment in rationality and autonomy, seen as the most valuable human qualities. The traditional view saw women as less rational and autonomous than men. Liberal feminists, then, argued for women's equal rights by arguing that women were as capable of rationality and autonomy as men were if given the right socialization and opportunities for education.

A key tenet of liberal thought has been a dichotomy between the public and private spheres, and a belief that, due to individual rights, the state should not intervene in what is private. In this view, justice only applies in the public sphere. But the location of the boundary between public and private is continually under dispute in both theory and real politics (Jaggar 1983). Gender relations in the family were excluded from the public sphere by the early liber6al writers (Benhabib 1987). With a few exceptions, early liberal feminists did not dispute this. Rather, they focused on women's equal rights in public matters, such as

education, politics, and careers. Exceptions to this include instances of liberal feminist support for laws against rape or violence by men against their wives, or support for laws mandating that employers give parents the job flexibility they need to combine being parents with employment (Okin 1989).

While part of the liberal tradition is *laissez-faire* and opposes state intervention in the economy, another part emphasizes the need for state action to achieve equality of opportunity by class, race, or gender. It is this second camp that is now called 'liberal' in popular discourse, but both are descendants of classical liberalism. The emphasis in egalitarian liberalism has been on equality under the law, equality of opportunity, or equality of treatment, not on equality of outcomes. The application of the egalitarian strand of liberalism to gender has led to advocacy of laws against discrimination in employment, credit, and educational opportunities. Encouraging more women in elected political positions and in the military has been another priority.

In both politics and research, the emphasis has been on what will get women into male-dominated enclaves. Less consideration has been given to what would increase men's participation in traditionally female spheres such as child rearing, paid caring labor, and other female occupations. Researchers, like the general public, think of egalitarian gender role-change largely in terms of women's movement into male spheres. But this overlooks the fact that women are unlikely to achieve equality in jobs and politics unless men share the burden of household and child-rearing work. But men seldom do; sociological research has shown that even in couples where the woman has a demanding career and earns about as much as the man, men typically perform much less household work than women. (See *Time-use and Gender*) Similarly, occupations cannot desegregate without some men taking up what have been women's jobs as well as more women moving into male fields. The failure to see that gender equality is unlikely unless gender-related change is more of a two-way street may result from an unrecognized devaluation of the traditionally female sphere. Precisely because work and qualities traditionally associated with women have been deprecated, men have had little incentive to enter them. This devaluation of everything associated with women is now a theme in its own right in feminist theorizing and in research on gender.

2. *The Devaluation of Roles and Qualities Associated with Women*

Much of feminist theory contests the claim that women are inferior to men. The liberal feminist tradition, discussed in Sect. 1, contests the claim that women are innately unqualified to fill well-rewarded public roles such as politician, soldier, scholar, manager, or worker. Liberal feminists point out that many women meet the qualifications for these positions, and many more could if provided with the same education and informal socialization men receive. Some thinkers, called radical or cultural feminists, make a different critique. (See *Feminist Theory: Ecofeminist and Cultural Feminist*; *Feminist Theory: Radical Lesbian*) In their view, the sexism in our society is not just a failure to give women the preparation to live up to the proper standards, or to judge fairly whether they meet these standards. Rather, they argue that the standards themselves are androcentric. They see sexist bias in our notions of what qualities people need to be in important positions, and the values determining which positions are rewarded most. In this view, what is seen as valuable has been socially constructed around what is seen as masculine. Examples of traditionally male characteristics highly valued in our culture include independence, abstract rationality, and repression of emotion. Contesting the valuation of these characteristics over traditionally female characteristics, these feminists ask why patriarchal and liberal views have elevated the spiritual or rational (which they associated with men) over the physical or emotional (which they associate with women)? (See *Rationality and Feminist Thought*) Where patriarchal views revere the bravery of risking one's life in hunting, sport, or war, the cultural feminist reveres nurturing for its preservation of life. Whereas patriarchal thinkers, classical liberals, and some Marxists have revered humans for their domination over nature, cultural feminism reveres harmony with nature.

The traditionally female characteristics that cultural feminists believe our culture has undervalued include nurturing, nonviolence, emotional sensitivity to the feelings of others, unselfishness, kinship with rather than domination of nature, acceptance of our physical bodies, humility, flexibility rather than rigid adherence to abstract principles, and intuition of wholes. People have always benefited from women's practice of these skills and values, but this benefit is seldom acknowledged in patriarchal societies. Rather, these virtues have been seen as weakness, lack of proper individuation, or lack of rationality; and social roles that center on these virtues (such as mothering, paid child care, nursing, teaching) are not allocated high rewards. (Examples of feminist works with this theme are Held 1993, Keller 1986, Nelson 1996, Starhawk 1987, Tuana 1993.)

When feminists argue for a valorization of women's traditional virtues and roles, this is often confused with advocating a return to traditional roles. If we see the value of the work that women have traditionally done as nurturers, this does not imply that women should stay at home and be full-time mothers. It implies that we should respect whoever does nurturing, men or women, encourage the development of these skills in all humans, and esteem and reward those who do the work of caring. Respecting traditionally female work

is entirely consistent with this work being shared by men. If the culture values and rewards traditionally female characteristics, why would it make sense to teach these things only to girls or save these roles for women? The radical and cultural feminists argue that traditionally male values are inappropriate for human behavior by either men or women in either domestic or public roles, at least without greater balance from traditionally feminine values than currently obtains.

Emotional connection and care are often involved in the activities traditionally assigned to women. By contrast, western thought has glorified a 'separative self' that is autonomous. Benhabib (197) traces the ideal of separative autonomy through the liberal political philosophy of Hobbes, Locke, Rousseau, and Kant. All presumed men to be separative and autonomous and presumed women would continue to rear children and provide emotional comfort and sexual satisfaction for men. But the value of women's work went unrecognized, as did the extent to which men were dependent upon women's work rather than autonomous. Such work was seen as 'part of nature' in a metaphysic that denigrated nature. These activities did not count as moral since only exercising autonomy in the public spheres counted as moral. This same bias in favor of individuation is seen in developmental psychology, according to Carol Gilligan (1982). She argued that Freud, Jung, Erikson, Piaget, and Kohlberg all viewed individuation as synonymous with maturation but viewed connection to others as developmentally regressive. They assumed women would do the emotional work of child rearing yet did not acknowledge learning the capacity for intimacy and nurture as part of maturation. This atomistic bias has also distorted neoclassical economics, according to many feminist thinkers (Ferber and Nelson 1993).

Paralleling the theoretical work showing a bias in western thought against qualities and activities associated with women, a body of empirical gender research has documented the low value and reward given to activities associated with women. One example is research associated with the policy issue known as comparable worth. (See *Comparable Worth in Gender Studies*.) Research in sociology, industrial psychology, and economics suggests that the sex composition of jobs exerts an effect on how much employers are willing to pay. If a job is filled largely by women, employers offer a lower wage. That is, after statistically controlling for job characteristics that affect pay, such as requirements for skill or difficult working conditions, researchers find that predominantly female jobs pay less than male jobs (England 1992). Thus, sexism affects earnings not only through keeping women out of high-paying jobs, but also through devaluing the jobs in which women are concentrated.

Sociologists and economists have recently begun to study the part of the economy that involves caring for people. Jobs that entail face-to-face care giving are especially badly paid for their educational levels. Such jobs include childcare, teaching, counseling, and social work. When children are reared well, there are broad social benefits, including a lower crime rate, a more productive economy, better informed citizens, and better care for the next generation. Many of those who receive these benefits pay nothing to those who did the paid and unpaid work of raising children. This means that the private return that markets will provide for such work will be lower than what is consistent with the broad social return. This is one rationale for government to reward the work more generously, to ensure that enough of these services are supplied, and as a matter of equity. Without this state involvement, and the sort of cultural change that cultural feminists advocate, women are likely to leave care work as other opportunities open up, or those who remain in it will pay a high price (in foregone wages) for doing this work. Whether as mothers or as workers in care work, women bear a disproportionate share of the costs of rearing children (Folbre 1994).

Many modern nations collectivize some of the costs of rearing children, through public financing of education and health care. Most European nations also provide cash family allowances to all families with children. The motivation for these programs was typically pronatalist rather than feminist, but they reduce the extent to which women bear the costs of privately rearing children, and in this sense are in women's interest. However, these subsidies to those caring for children are a relatively small part of the modern welfare state. Scholars studying gender and the welfare state have pointed out that many governments have developed two tiers to their benefits. A 'masculine' set of programs is geared to individuals who make claims as earned rights based on past employment. The 'feminine' programs support single mothers when a male breadwinner isn't present; the programs were seen to compensate for family failures. Underlying the two-tiered benefit system are gender norms that define the home as a female sphere and outside work as a male sphere. Men rely on social insurance schemes for retirement income, disability, and unemployment insurance, and women can make claims on these programs too when they are employed regularly or as wives or widows of earners. But single women with children are segregated into 'female' income-tested benefits, and these programs are more stigmatized and pay much less. (See *Social Welfare Policies and Gender*.) Women are expected to care for children, but this work does not provide as respected or rewarded an entitlement to state help in a time of need as a history of employment. More generally, feminists studying the state have argued that the seemingly gender-neutral notion of citizenship is in fact gendered; it has valorized men's typical activities such as worker and soldier, while failing to give rights and privileges for the parenting work that mothers do.

3. Intersections of Gender with Class, Race, and Nation

In the 1970s, gender scholars often wrote as if the subordination of women to men was universal. Depending on which theory one chose, patriarchy was caused by the public/private split, men's ownership of property, employment discrimination, or a combative culture. The idea was that once we found the correct explanation of patriarchy, this would reveal how women could be liberated from patriarchy. So it seemed to many of the upper-middle class white feminists in affluent nations, which were the groups predominating in the women's movement and in academic gender and feminist studies. A challenge to this view emerged from women of color in Western nations and in poorer Southern nations. (See *Feminist Theory and Women of Color*; *Race and Gender Intersections*.)

These critics argued that the analyses of gender proposed by affluent, white, Western women distorted the way gender affected the lives of many women in the world. The argument was not merely that, for example, the life of an affluent, married white professional woman in America is very different from that of an African-American welfare mother, or a Mexican immigrant to the US, or a middle-class Pakistani woman. That is obvious, since the latter three groups lack privilege on several fronts other than their status as women. The argument was deeper—that 'gender' operates differently for these different groups. The claim was that gender interacts with other axes of subordination. (See, for example, Anzaldua 1990, Collins 1990, Hooks 1989, Lorde 1984, Stoler 1996.)

For example, ideas constructed by men of a dominant racial group about womanhood in fact often create one conception of woman for members of their racial group and another for the subordinate group. In the American South, during and after slavery, white women were seen as chaste and in need of protection (especially from black men). Yet white men did not see black women as needing or deserving this chivalrous protection. They were seen as naturally sexual. Such a construction served to justify white men's interest in having a group of women who had little power to resist sexual coercion and few claims to white men's help in supporting any resulting children. Here clearly gender and race interacted. Gender oppression differed by race. White middle-class women found themselves limited to the role of asexual, married mother, while women of color were unprotected from sexual exploitation by white men. Similar patterns emerged with the European colonial expansion to North and South American, Africa, and Asia.

Hegemonic notions of manhood often impact men and women very differently depending on their class and race, even when there is agreement across class and race groups on what it means to be a man. When someone says 'Be a man!' the implied notion of masculinity revolves around power. This notion, while certainly not universal, is shared by men in most class and ethnic groups in modern societies. But men in the dominant ethnic groups have easier access to socially accepted ways of getting power, such as success in school and the labor market. This combination of cultural notions of masculinity as power with class-based or race-based lack of access to 'legitimate' sources of power creates a systemic strain making it likely that poor men will seek power through risky and illegal behavior. The gender-related nature of class and race privilege is illustrated by the fact that what men of less privileged class or race groups may envy about privileged men is that the latter often have power over women of all race and class groups. Women in subordinate class or racial groups may sympathize with the sense of powerlessness that men of their class or race feel. Yet they may also see how disadvantaged they are by the fact that they are the only actors in the social system less powerful than men of their race or class are.

Interactions of gender and nation can lead to complex patterns of identity and collective action. Consider, for example, the Islamic practice of women wearing a veil or the practice of female genital mutilation practiced in certain parts of the world. Clearly these practices are linked to male dominance. Yet, in some settings where indigenous people resent Western imperialism, which has harmed both women and men, women as well as men may see these practices to symbolize nationalist solidarity against what is seen as the decadence of Western individualism and greed. (See *Female Genital Mutilation*; *Islam and Gender*.)

The lesson from scholars of these intersections is that to understand many facets of the gender system, it needs to be studied in combination with axes such as race, ethnicity, class, sexual orientation, and nation. Yet this can sometimes lead to a rich mosaic of particulars that threaten to make gender and feminist studies descriptive rather than theoretical and explanatory. The challenge is to study gender in a way that does not falsely universalize, while looking beyond the unique aspects of each situation to search for broad patterns that help us understand what explains continuity and change in gender.

See also: Androgyny; Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Economic Development and Women; Family and Gender; Feminist Economics; Feminist Epistemology; Feminist Ethics; Feminist Legal Theory; Feminist Movements; Feminist Political Ecology; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies in Anthropology; Gender and Feminist Stu-

dies in Economics; Gender and Feminist Studies in Geography; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Economics of; Gender, Feminism, and Sexuality in Archaeological Studies; Gender History; Gender Ideology: Cross-cultural Aspects; Klein, Melanie (1882–1960); Male Dominance; Masculinities and Femininities; Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns; Population Dynamics: Two-sex Demographic Models; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Race and Gender Intersections; Sex Differences in Pay; Sexuality and Gender

Bibliography

Anzaldua G 1990 *Making Face, Making Soul*. Aunt Lute Foundation Books, San Francisco, CA

Benhabib S 1987 The generalized and the concrete other: The Kohlberg-Gilligan controversy and feminist theory. In: Benhabib S, Cornell D (eds.) *Feminism as Critique: On the Politics of Gender*. University of Minnesota Press, Minneapolis, MN, pp. 77–95

Bergmann B 1986 *The Economic Emergence of Women*. Basic Books, New York

Chodorow N 1978 *The Reproduction of Mothering*. University of California Press, Berkeley, CA

Collins P H 1990 *Black Feminist Thought: Knowledge, Consciousness, and Empowerment*. Unwin Hyman, Boston, MA

England P 1992 *Comparable Worth: Theories and Evidence*. Aldine de Gruyter, New York

Ferber M, Nelson J A 1993 *Beyond Economic Man: Feminist Theory and Economics*. University of Chicago Press, Chicago, IL

Folbre N 1994 *Who Pays for the Kids: Gender and the Structures of Constraint*. Routledge, New York

Gilligan C 1982 *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA

Held V 1993 *Feminist Morality: Transforming Culture, Society, and Politics*. University of Chicago, Chicago, IL

Hooks B 1989 *Talking Back: Thinking Feminist, Thinking Black*. South End Press, Boston, MA

Jaggar A 1983 *Feminist Politics and Human Nature*. Rowman & Allanheld, Totwa, NJ

Keller C 1986 *From a Broken Web: Separation, Sexism, and Self*. Beacon Press, Boston, MA

Lorde A 1984 *Sister Outsider: Essays and Speeches*. Crossings Press, Trumansburg, NY

Nelson, J 1996 *Feminism, Objectivity, and Economics*. Routledge, London

Okin S M 1989 *Justice, Gender, and the Family*. Basic Books, New York

Sanday P R 1981 *Female Power and Male Dominance*. Cambridge University Press, Cambridge, UK

Starhawk 1987 *Truth or Dare: Encounters with Power, Authority, and Mystery*. Harper & Row, San Francisco, CA

Stoler A 1996 Carnal knowledge and imperial power: Gender, race and morality in colonial Asia. In: Scott J W (ed.) *Feminism and History*. Oxford University Press, New York, pp. 209–66

Tuana N 1993 *The Less Noble Sex: Scientific, Religious, and Philosophical Conceptions of Woman's Nature*. Indiana University Press, Bloomington, IN

P. England

Gender and Feminist Studies in Anthropology

The development of studies of gender in anthropology received its impetus from the emergence during the late 1960s and early 1970s in Europe and the USA of what became known as the second wave feminism. This article primarily will review theoretical challenges and critical perspectives contributed by feminist anthropology to both the field of anthropology and feminist studies. The article will address main areas of feminist research and scholarly production such as reproduction, colonialism and postcoloniality, labor and transnational capital, Third World feminisms, and feminist ethnographic writing. Finally, the article will address recent work on masculinity.

1. In Search of Theoretical Frameworks

In anthropology, gender difference had played a crucial role in the theoretical development of the field providing especially during its early history a powerful metaphor for thinking about questions of cultural difference. Many characteristics attributed to 'primitive' societies in the armchair anthropology of the late nineteenth century resembled dominant definitions of feminity in Victorian society, such as nonlogical thinking and excess of emotionality. So too gender difference provided Victorian anthropology with an unquestioned 'natural' model for conceptualizing the notion of race and racial difference which was becoming increasingly central to the discipline. As feminist anthropologists came to argue, the position occupied by the notion of gender in the field of anthropology cannot be separated from the politics of colonial expansion and the development of a theory of cultural difference formulated around the notion of race. This linkage between gender, race, and colonial politics has become an important arena of contestation within contemporary feminist anthropology as this article will show. Not only as a metaphor but also as a theoretical premise, gender difference has been a cornerstone of anthropological knowledge as in, for example, Levi-Strauss' structuralist theory of culture.

While gender difference was crucial to the production of anthropological theory, gender was a

category of analysis filled with uncritical assumptions about the naturalness of inequality between the sexes. Gender was that piece of anthropological theorizing that was unproblematic, because it was conceptualized as natural and as such could provide the support for the study of other cultural variables. Such naturalization of gender difference was challenged strongly in the 1970s by a group of young feminist scholars in a series of foundational texts for contemporary feminist anthropology. Gathered predominantly in two volumes (Rosaldo and Lamphere 1974, Reiter 1975) these texts questioned the naturalness of female subordination. They departed from the premise that gender difference and gender inequality were not the product of biology but were culturally constructed, and attempted to formulate theoretical frameworks that could explain 'the universal asymmetry in cultural evaluation of the sexes' (Rosaldo 1974). These first attempts of feminist anthropology at thinking gender difference and sexual inequality argued that women's subordination was not the product of nature but the product of a number of social factors. Most important among these factors were a social organization dividing private from public domains and the cultural ideology that assigns importance to the public over the private and associates women with the latter and men with the former (Ortner 1974, Rosaldo 1974).

The cultural ideology that sustained this division of private/public social domains was aided by an economic division of labor that undervalues women's work in the reproduction of the species and labor force (Sacks 1974). It was also aided by a family structure that encouraged independence in boys and responsibility for care giving in girls (Chodorow 1974). In their search for theoretical frameworks to study gender inequality, the feminist anthropologists of the 1970s engaged the major social theories of twentieth century –structuralism, Marxism, and psychoanalysis—challenging their gender presuppositions and opening new and exciting avenues of analysis (Rubin 1975). A number of important feminist ethnographies challenged male centered accounts of culture and through the detail of ethnographic cases they showed women as social actors deploying a variety of strategies in their engagements with power, prestige, economic exchange, and other social fields.

2. *Beyond Dichotomies*

The emergent and vibrant feminist anthropology of the 1970s rested on some assumptions. Much of its theorizing was predicated on crucial dichotomies such as private/public, nature/culture, male/female. Gender was understood primarily in unitary terms as an identity all women and all men shared. At issue was the power difference between male and female. A universalist *Woman* rather than women was the object of analysis. At the turn of the 1980s the theoretical dualities characterizing feminist theorizing of gender began to be questioned and with them an essentialist conception of *Woman*. This challenge was characterized by an emphasis on history, meaning, and power in the feminist studies of the early 1980s (Ortner and Whitehead 1981, MacCormack and Strathern 1980, Bourque and Warren 1981). Some scholars studied the change in meaning of seemingly unchanging categories such as nature and culture so instrumental in the construction of gender, and argued that gender as a category needed to be examined historically. Another strand of theorizing called attention to power difference among women. New research on the impact of colonialism on women argued that the oppression of women was complicated by their position vis a vis colonial power (Leacock 1980). Thus, while European women could be oppressed as women within a gender hierarchy, they were the agents of colonial domination of native men and women. Feminist studies of colonialism also emphasized the different forms in which colonialism as system of oppression changed the realities of women in ways that were different from those experienced by men. Gender as a system of power relations was inseparable from other relations of power. To complicate things more, another strand of research noted the versatility of gender as sign and symbol (Scott 1988). These researchers affirmed that gender had been historically used to express, conceptualize, or advance other kinds of struggles, whether these were about citizenship, nationalism, or race. The point emphasized by this corpus of research was that gender could not be taken for granted as self-evident reality but had to be examined within a multiplicity of social, political, and ideological contexts as well as in historical perspective. In a path breaking essay greatly influential in the anthropology of gender Joan Scott offered a definition of gender that was both comprehensive and complex, bringing together the various foci of research. Scott (1988, pp. 42–7) defined gender as:

(a) A constitutive element of social relations based on perceived differences between the sexes which involved four interrelated elements:

(i) Culturally available symbols that evoke multiple (often contradictory) representations.

(ii) Normative concepts used for interpreting the meanings of symbols.

(iii) A notion of politics and reference to social institutions and organizations necessary to understand how the normative concepts are used.

(iv) Subjective identity which entail the question of how is gender identity constructed in relation to a range of activities, social organizations and historically specific cultural representations.

(b) Gender is a primary way of signifying relations of power. In other words, gender is a primary field within which or by means of which power is articulated. This understanding of gender as signifier of power

was very influential in feminist ethnographies of colonialism, nationalism and violence. For Scott, concepts of gender structure, not only perception of social life, but its concrete and symbolic organization. Sex related differences between bodies are summoned as legitimation of social relations and phenomena that have nothing to do with sexuality.

In anthropology the study of gender challenged the more canonic fields of social structure, power relations, and symbolic constructions. Gender was not one area of analysis but one that forced a rethinking of the thematic pillars of the discipline.

3. *New Explorations in Contentious Domains*

During the 1980s there was an explosion of very creative feminist ethnography. This scholarly production was less interested in filling the gap of knowledge about women in the discipline than in the study of women as complex social actors engaged in economic, social, and political fields. An important area of feminist production was kinship. Kinship and reproduction, which had long been a privileged area of research in anthropology, became an important domain of feminist ethnography. Feminist ethnographers challenged male centered analysis of kinship by focusing on the role of women in maintaining or transforming kinship ideologies and structures. They also looked at the ways in which kinship structures and ideologies placed women in disadvantageous positions. Some researchers argued that not only gender but also social class, ethnicity, and sexual orientation had an effect in the way kinship and family were perceived and structured. Studies of kinship among poor African-Americans questioned the theory that the high number of women headed families in this social group was a result of a culture of poverty. Instead, these studies argued that complex kinship networks, articulated around women, played a crucial role in ensuring economic and social survival of large numbers of people in a situation of chronic economic disadvantage (Stack 1974). Research on gay and lesbian families in the Bay area of San Francisco challenged the dominant thesis of an American kinship system based on the ideology of blood connections by mapping the ways in which gay and lesbian people constructed family networks and ideologies not based on blood connection (Weston 1991). In England, feminist analysis of kinship was linked to the emergence of new reproduction technologies and their effect on nationalist ideologies (Strathern 1992). These studies argued that new reproductive technologies were changing not only ideas about women and families, but also ideas about what constituted nature and culture. The association of women with nature that had its support on women's reproductive capacity was challenged with the development of new technologies of reproduction that made reproduction a highly technological affair but also allowed for a complex array of family forms not based on biological connection. Feminist anthropologists studying reproduction also focused on the ways in which women's bodies were represented in the discourse of bio-medicine and the ways in which this discourse excluded women from the control of their own bodies. Approaching science and bio-medicine as cultural discourse some researchers inquired about the gender and race assumptions underlying medical discourses.

Another important area of research during the 1980s was the study of the effects of transnational capital on women. One strand of research focused on the transformations of the role of women in the family and community triggered by the introduction of transnational manufacture in traditional societies of the Third World. Some researchers studying the new international division of labor triggered by transnational capital, argued that multinationals was producing a class of poor, working-class women worldwide. Other researchers focused on the use of traditional beliefs and rituals such as spirit possession to manage the stress and alienation of factory production and capitalist discipline (Ong 1987).

The anthropological research on transnational capital brought to the anthropological agenda new issues about the interrelation between gender and social class, the transformation of traditional societies, and the linkages between new forms of sexual violence and capitalist expansion in the Third World. During the 1990s, the growing industry of sexual tourism has become a new important area of research.

A renewed interest during these years on history and change within anthropology, as well as on the labor politics of transnational capital fueled research on the neglected areas of colonialism and nationalism. Some researchers turned to history to study the cultural assumptions about gender, sexuality, and race that colonial officers and soldiers brought with them to the colonial context (Stoler 1991). A central preoccupation of these studies has been the study of the role of tradition in colonial societies. Tradition is important because it has often been associated to gender identity and used for a variety of purposes. For example, feminist studies of the practice of Sati (the ritual burning of widows in India) show that the practice was subject to a number of contingencies and hardly generalized in precolonial India. The English, however, took this practice as an example of a barbarian tradition that had to be abolished to bring the country to the English level of civilization. In this case an argument concerning violence against women was furnished to legitimize a colonial power relation. In so doing a practice which was not tradition became formalized as such. Once Sati was understood and treated by the British as Indian tradition it began to be associated by some nationalist strands with Indian national identity as well as with a feature of gender relations (Mani 1990). Similar processes concerning

tradition and ethnic identity were at play in colonial Africa in relation to practices of female circumcision. Traditions concerning gender relations produced by colonial scholarship and colonial policy became later on an emblem of national identity. The consequence, almost everywhere, has been a greater control of women and an increase on violence against women. This charged area surrounding the definition of cultural tradition and its relation to ethnic identity has become a central part of contemporary discussions about immigration, race, and citizenship in European countries and the USA. These discussions have been dramatized in the courtrooms and mass media, triggered by charges brought against immigrants performing various forms of genital mutilation on girls. In France, for example, this discussion has confronted feminist activists against cultural specialists. Feminists have argued that female mutilation is a form of violence that violates the right of girls and women to physical integrity and, as an illegal practice, should be prosecuted by the law. Cultural specialists, in turn, have argued that genital mutilation is part of the traditional culture that constitutes the identity of those who practice it. A variety of associated meanings and situations complicate this discussion. At issue is, for example, the fact that those brought to trial are poor immigrants subjected to institutionalized forms of racism in their everyday life. At issue is, too, the forms in which the discourse of human rights has become part of French national identity. The point is that neither 'tradition' nor 'culture' are neutral or self-evident terms, but are terms used to define regulatory practices between men and women and between the state and its people. These practices of power are articulated and symbolized on the bodies of women and particularly Third World women and women of color.

A growing body of feminist anthropology has become preoccupied with the question of how women figure in relation to nationalist identity, nationalist ideologies, and postcolonial nationalist struggles. This line of research has investigated the way nationalist ideologies constrain the lives of women by making them either the repository of national tradition or the passive allegory of the nation. Part of the problem is that colonial discourse often defined colonial subjects in gender codes, attributing them a feminine incapacity for government. To counteract this definition, nationalist movements have often resorted to a reverse gender language that represents the people of the nation as masculine while representing the nation as a female object. When certain women's roles such as motherhood are defined as essential to national identity, such roles are often enforced through legislation, restricting the choice and development of women in other social areas. In postcolonial nations the pressures on women to conform to nationalist demands become complicated by the fact that the new nation represents freedom from the bonds of colonialism. Thus, to rebel against its constraints can be perceived and experienced as national disloyalty. These positions create double bind-like dilemmas for women whose aspirations go against the grain of the national project (Aretxaga 1997).

Notwithstanding its constraints, for women engaged in struggles of liberation such as Palestinian women, nationalism can open a space of social and personal liberation. Research on women in nationalist movements shows that political activism has the paradoxical effect of constraining and liberating women. Women often maneuver between contradictory demands and divided loyalties. At the same time, women's activism in struggles of liberation often leads to reformulation of gender relations but research shows that such reformulation is always precarious and shifting (Peteet 1991). The anthropological research of gender and nationalism has demonstrated that gender identity is never defined in isolation from other forms of identity such as class, ethnicity, or nationality. This research has also brought to light the differences among women showing that the very category of women even within an ethnic group is internally divided. The research on women and nationalism has counteracted a dominant ideology that has represented women in the Third World and women of color as passive victims of authoritarian nationalism or political violence.

4. *Third World Women and Women of Color*

During the late 1980s and early 1990s women of color in the USA and Europe claimed a specific epistemological and political position within the feminist movement. They criticized the feminist mainstream for silencing the voice of minority women and women of color and contended that the emphasis on women's solidarity had obscured relations of power between and among women of different class, ethnicity, and color. This critique was followed by a critique advanced by Third World women who claimed to be in a subordinated position vis-à-vis western women. Both the notion of 'Woman' and 'Western' were the object of severe criticism. If this moment had a tremendous political impact on the feminist movement worldwide, it had an impact on feminist anthropology as well. Most research on gender and women in anthropology had been produced by Western women studying women in Third World countries or minority women in Europe or the USA. The critique of imperialism launched against white mainstream feminism by Third World feminists cut right across feminist anthropology. The result among feminist anthropologists was a self-critical reflection that emphasized the awareness of the position of power the anthropologist held vis-à-vis her subject and how such position influenced the production of ethnographic knowledge. Feminist knowledge in anthropology became 'situated

knowledge' rather than unmarked universal knowledge. There was a critique of representation and an emphasis on the production of 'subjugated knowledge' (Haraway 1991). Feminist anthropologists turned their attention to questions of experience and self constitution as generating ethnographic knowledge. There was a new emphasis on the role of subjectivity in scholarship which characterized the 1990s. Theoretical production about hybridity, diversity, and partial knowledge rather than universal generalization was at the center of the feminist anthropology of the 1990s.

5. *Writing Gender*

The attempt at presenting rather than representing the voice of minority, subjugated women led feminist anthropologists to experiment with the form of ethnographic writing. Some anthropologists reflected on the difficulty of capturing the subjectivity of women in contexts where women only spoke through the conventions of male oratory. Others found that conventions of ethnographic representation were fundamentally male oriented and inadequate to narrate women's experience. Feminist ethnographers experimented with literary genres and strategies to assert a new form of writing culture suitable to convey hitherto invisible women's experience. Fiction, journal writing, first person narrative, and autobiographical narrative became part of feminist ethnographic writing during the 1990s. Some ethnographers experimented with transcribing women's narratives with minimal interference of analytic voice (Behar and Gordon 1995).

Others juxtaposed the biographical history of the ethnographer to the autobiography of the women they wrote into books, yet others resorted to mixed fictional forms to express the dilemmas of feminist ethnography in a postmodern, fragmented world divided by relations of power and increasingly hybrid. Feminist ethnographies reflected not only on the position of the women they wrote about but on the hybridity of the authors, themselves positioned oddly between the West and the Third World.

For these feminist anthropologists, ethnography in the contemporary world becomes a kind of fiction writing. Their ethnographic stories are composed of stories told in the field by informants, often informally. They freeze a fragment of social life and a fragment of history out of the ongoing unfolding of events. In ethnography, like in fiction, closure is always arbitrary and the goal for feminist anthropology is to capture the dilemmas, complexities, and uncertainties of women's experience. The kinds of questions that traverse experimental feminist ethnography gravitate towards the ambiguities of identity, community, the meaning of home, and fieldwork. In writing about other women, feminist ethnographers reflect on their problematic identity and the kind of home and community that made them feminist ethnographers.

6. *The Anthropology of Masculinity*

Although anthropology had focused traditionally on men, these were represented as universal subjects rather than men whose masculinities were culturally constituted in a variety of ways. During the late 1980s and 1990s, a growing number of ethnographic works have focused on the cultural construction of masculinity. Ethnographies of masculinity have broadened the anthropology of gender often associated with women's studies. However, as Guttmann (1997) notes there has been an awkward avoidance of feminist theory by those anthropologists studying manhood. The early anthropology of masculinity focused on how men performed their manhood in a variety of cultural contexts. More recent work analyzes masculinity in relation to a broader number of themes. Guttmann (1997) notes several topics of interest in relation to which masculinity has been discussed by anthropologists, such as nationalism, divisions of labor, family, friendship, sexuality, and struggles of power, yet there are no general theories of masculinity produced from anthropology. Part of the problem is the difficulty in defining masculinity which has been taken so much as the norm that it remains conceptually elusive in its specificity. While a number of works have explored performances of manhood in relation to other men, women, and the national identity (Herzfeld 1985), an important contribution has come from the discussion of men's subjective perceptions and feelings about other men.

7. *Conclusion*

Since the 1970s, the anthropology of gender has grown into a complex, diverse, and theoretically sophisticated field. It has gone from a focus on women's studies to include increasing study of masculinity across cultures. Much of the anthropology of gender falls into the rubric of feminist studies and this has posed political as well as epistemological challenges that have resulted in a sophisticated, mature, and rich body of scholarship.

See also: Family and Gender; Family Theory: Feminist–Economist Critique; Female Genital Mutilation; Feminist Theory; Gender and Feminist Studies; Gender and Feminist Studies in Sociology; Gender Differences in Personality and Social Behavior; Gender Ideology: Cross-cultural Aspects; Male Dominance; Masculinities and Femininities; Nation-states, Nationalism, and Gender

Bibliography

Aretxaga B 1997 *Shattering Silence: Women, Nationalism and Political Subjectivity in Northern Ireland.* Princeton University Press, Princeton, NJ

Behar R, Gordon D A (eds.) 1995 *Women Writing Culture*. University of California Press, Berkeley, CA
Etienne M, Leacock E (eds.) 1980 *Women and Colonization: Anthropological Perspectives*. Praeger, New York
Guttman M C 1997 Trafficking in men: The anthropology of masculinity. *Annual Reviews of Anthropology* **26**: 385–409
Haraway D J 1991 *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York
Herzfeld M 1985 *The Poetics of Manhood: Contest and Identity in a Cretan Mountain Village*. Princeton University Press, Princeton, NJ
MacCormack C P, Strathern M (eds.) 1980 *Nature, Culture and Gender*. Cambridge University Press, New York
Ong A 1987 *Spirits of Resistance and Capitalist Discipline: Factory Women in Malaysia*. State University of New York Press, Albany, NY
Ortner S B 1974 Is female to male as nature is to culture. In: Rosaldo M Z, Lamphere L (eds.) *Woman, Culture and Society*. Stanford University Press, Stanford, CA
Ortner S B, Whitehead H 1981 *Sexual Meanings: The Cultural Construction of Gender and Sexuality*. Cambridge University Press, New York
Peteet J M 1991 *Gender in Crisis: Women and the Palestinian Resistance Movement*. Columbia University Press, New York
Reiter R R (ed.) 1975 *Toward an Anthropology of Women*. Monthly Review Press, New York
Rosaldo M Z 1974 Woman, culture, and society: A theoretical overview. In: Rosaldo M Z, Lamphere L (eds.) *Woman, Culture and Society*. Stanford University Press, Stanford, CA
Rosaldo M Z, Lamphere L (eds.) 1974 *Woman, Culture and Society*. Stanford University Press, Stanford, CA
Rubin G 1975 The traffic in women: Notes on the 'Political Economy' of sex. In: Reiter R (ed.) *Toward an Anthropology of Women*. Monthly Review Press, New York
Sacks K 1974 Engels revisited: Women, the organization of production and private property. In: Rosaldo M Z, Lamphere L (eds.) *Woman, Culture and Society*. Stanford University Press, Stanford, CA
Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York
Stack C B 1974 *All our Kin*. Harper & Row, New York
Stoler A L 1991 Carnal knowledge and imperial power: Gender, race and morality in colonial Asia. In: di Leonardo M (ed.) *Gender at the Crossroads of Knowledge: Feminist Anthropology in the Postmodern Era*. University of California Press, Berkeley, CA
Strathern M 1992 *Reproducing the Future: Essays on Anthropology, Kinship and the New Reproductive Technologies*. Manchester University Press, Manchester, UK
Warren K B, Bourque S C 1981 *Women of the Andes: Patriarchy and Social Change in Two Peruvian Towns*. Michigan University Press, Ann Arbor, MI
Weston K 1991 *Families We Choose: Lesbians, Gays, Kinship*. Columbia University Press, New York

B. Aretxaga

Gender and Feminist Studies in Economics

A feminist approach to economics recognizes the importance of gender in structuring human experience, including economic life, and seeks economic knowledge for the purpose of combating gender-based oppression. While precursors of such feminist studies go back a century or more, feminist economics did not emerge as a body of thought among academic economists until the early 1990s. The ensuing decade brought an expansion of feminist scholarship not only about women's (and men's) economic status, but also about the gendered structure of economics as a discipline, itself. Can a discipline that seeks for reliable knowledge about economic life *not* be feminist?

1. *Defining the Terms*

For a branch of scholarship to be considered 'feminist,' it should recognize the importance of gender in structuring human experience, and should be part of a liberatory project to eliminate sexist structures that keep women in inferior and submissive positions relative to men. Work that is simply about women, or by women, but that does not challenge structures of domination, is not feminist in this sense (and hence will not be discussed here). Feminist theories include liberal, Marxist, radical, and other variants (see *Feminist Theory: Ecofeminist and Cultural Feminist*; *Feminist Theory*; *Feminist Theory: Marxist and Socialist*; *Feminist Theory: Psychoanalytic*; *Feminist Theory: Liberal*; *Feminist Theory: Postmodern*; *Feminist Theory: Radical Lesbian*).

Academic economics includes schools of neoclassical, Marxist, institutionalist, post-Keynesian, and Austrian thought. The neoclassical school, however, is clearly the 'mainstream,' as evidenced by its absolute hold on the highest-status institutions and their graduate and research programs, as well as its ubiquitous presence elsewhere. The neoclassical school is paradigmatically characterized by centrality of the question of how self-interested individual agents make rational choices when confronted with scarcity, by a strong belief in the allocative efficiency of markets, by a near-sole reliance on mathematical reasoning as the means of creating knowledge, and by the application of statistical testing to large-scale, often government-collected, data. This neoclassical definition of 'economics' is itself at issue in feminist studies, as will be discussed below. Because of the hegemony of this one school, this article will emphasize work that grows out of and/or reacts against it, although feminist economic studies can also be found within the institutionally weaker heterodox schools, the other social sciences, and popular, nonacademic, economic literature.

Some may doubt that feminist studies have a legitimate place with economics, since they may seem to mix a political agenda into an otherwise value-free, neutral scientific enterprise. Much of the accomplishment of the last decade, however, has been precisely to reveal the value-laden, masculine-biased, and evi-

dence-ignoring aspects of mainstream economic thought, and to begin to envision less biased and more satisfactory means to warranted knowledge.

2. The Rise of Feminist Studies in Economics

Feminist studies in economics, as a body of thought, formed in the early 1990s from a convergence of two streams: first, concerns about women's economic status, and second, feminist scholarship on the role of gender in the structuring of science.

Increasingly, in the 1970s and 1980s, concerned women had pointed out the neglect and/or distortion of women's experience in several areas. Marianne Ferber and Bonnie Birnbaum, for example, debated the proposition (put forth by the University of Chicago-inspired 'New Home Economics') that women's exclusive specialization in household work could be explained as their free and advantageous choice (Ferber and Birnbaum 1977). Paula England cast doubt on the empirical validity of the 'human capital' explanation of male–female wage differentials, in which women's lower pay was portrayed as the result of their own utility-maximizing choices of training and occupation (England 1982). Barbara Bergmann highlighted the importance of, and difficulties facing, the movement of women into the labor force (Bergmann 1986). Marilyn Waring pointed out that the United Nations systems of national income and product accounts excluded any consideration of unpaid work by women (Waring 1988). While some scholars took what Sandra Harding has called a 'feminist empiricist' approach, believing that 'social biases [are] correctable by stricter adherence to the existing methodological norms of scientific inquiry' (Harding 1986, p. 24), others began to question the particular forms of scientific inquiry sanctioned by the economics profession. Issues of dependence, interdependence, tradition, social beliefs, power, and oppression did not seem to fit well into mainstream forms of analysis, and feminist scholars began to search for more adequate means with which to understand the fullness of economic life.

During this same period, feminist physical scientists, philosophers, and historians (e.g., Harding) were revealing the gender biases deeply embedded in modern conceptions of science. The identification of objectivity with detachment, and the consequent model of an active, knowing subject exerting control over a passive object of knowledge, are metaphorically connected with images of male/mind activity and female/matter passivity (see *Rationality and Feminist Thought*). Within the economics profession, some attention to the rhetoric of economics (McCloskey 1985) simultaneously helped bring to light the social nature of economic discourse, allowing for the pot-ential discussion of gender influences within the discipline.

Table 1
Contrasts defining neoclassical economics.

Core	Margin
Public	Private
Market	Home
Individual	Social
Efficiency	Equity
Hard	Soft
Fact	Value
Rigor	Intuition
Precise	Vague
Objective	Subjective
Detached	Committed
Self-interest	Altruism
Autonomy	Dependence
Reason	Emotion
Masculine	Feminine

Contrasts such as those shown in Table 1 highlight the gendered construction of neoclassical economics, as those domains, methods, and assumptions chosen to be central to the discipline all have stereotypical masculine associations, while the neglected and denigrated areas are socially considered 'feminine' (see, e.g., McCloskey 1985, Nelson 1992, Jennings 1993). As such a definition of economics can be seen as one manifestation of an identification of feminine associations with inferiority, feminists have sought ways to deconstruct these gender dualisms.

The formation of the International Association for Feminist Economics in 1991 and the publication of *Beyond Economic Man: Feminist Theory and Economics* in 1993 (Ferber and Nelson 1993) signaled the beginning of feminist studies in economics as a body of thought. Continued conferences and the publication of a journal, *Feminist Economics* (started 1995) illustrate the dynamism of the field at the turn of the century.

3. Major Developments in the Field

Three major areas have absorbed the attention of many feminist scholars: first, exploration of the definition and methodology of economics; second, analysis of gendered labor market outcomes; and third, inquiries into work traditionally done by women.

3.1 Definition and Methodology of Economics

Most feminist work in economics recognizes that the discipline is socially constructed, that is, that it reflects the interests of the people who developed it and the wordviews and prejudices of its formative historical epochs, rather than simply a set of trans-

historical ideals. The dualisms in Table 1, then, reflect social, intellectual, and even psychic biases of the creators of the discipline.

One response might be to think that the point of a 'feminist' economics would be to add more women economists, and, therefore, 'women's ways of knowing.' This emphasis, however, has relatively little support among feminist economists, since it puts the causal story at the level of biology (or sex) rather than at the level of social understandings (gender). Most feminists are less interested in finding 'essential' differences between males and females, and more interested in examining the heavy cultural influences that go into creating observed differences and schemes of categorization. Since, when given the chance, female economists have proven quite adept at the types of thinking required by current disciplinary strictures, a feminist approach looks elsewhere.

The more important point is that the creators of mainstream economics lived in societies characterized by a culture-wide belief in masculine superiority. One graphic tool for working beyond the usual associations between gender and value was proposed by Nelson (1992). A vertical axis marks off the dimension of value, positive (+) or negative (−), while a horizontal axis marks off gender associations, masculine (M) or feminine (F). While sexist thinking recognizes the M+ and F− quadrants, examination of the unusual M− and F+ cells may reveal new possibilities. For example, the idea that economics should be a 'hard' science rests on an identification of hardness with (masculine-associated) strength, and softness with weakness (and femininity). Yet hardness can also mean rigidity (M−), while softness can mean flexibility (F+). Too much effort in economics has been spent on defending the masculine/feminine boundary, and too little on improving the value of economic studies. As a result, the models, methods, and pedagogy of mainstream economics tend to be characterized by rigidity, obsession with mathematical reasoning, detachment from human suffering, and narrow forms of explanation (M−). While recognizing the validity of ambitions to be strong, logical, scientific, and precise (M+), one can simultaneously recognize the value of flexibility, analogical and dialectical reasoning, concern with human life, and richness of analysis (F+).

Avoiding the hyper-rationalist mind/body split implicit in a definition of economics in terms of rational choice, and the marginalization of women's traditional activities implicit in a definition in terms of markets, feminists have suggested a definition of economics as being related to the subject matter of the provisioning of human life. Such a definition in terms of subject matter allows for the use of a larger toolbox of models and methods. Instead of models of individual autonomous in choice, agency is recognized as being limited by physical embodiment and social embeddedness. Rather than identifying knowledge solely in the forms of theorems and proofs, or statistical tests on second-hand data, the value of more qualitative and even experiential data, and of rigorous attention to the gathering and interpretation of all data, receives increased attention. Objectivity is not guaranteed by strict adherence of an individual researcher to some prescribed model or method, but rather is a characteristic obtainable only through the shared understandings of an expanding knowledge community.

Some feminist economists take this analysis a step further, and, seeing the importance of language and the partiality of positions thought previously to reflect universal truths, conclude that reliable knowledge is not possible. In the extreme deconstructionist postmodern view, the social construction of economics is taken to mean that there is no reality outside of the linguistic structure. While the feminist–postmodernist conversation (see *Feminist Theory: Postmodern*) has influenced feminist studies in the 1990s, such *relativism* is not a necessary consequence of a feminist approach, and most scholars in economics adopt at least a limited form of philosophical realism. That is, while economics is seen as socially constructed, it is also recognized that the external world puts limits on the interpretations that can be supported.

3.2 Gendered Labor Market Outcomes

The movement of women into paid labor in most industrialized countries supplied the raw material for a large literature on gendered occupational distributions and the gap between average male and female wages. Much of the work uses mainstream methods, most notably regression analysis, to determine how much of the wage gap might be attributable to sex *per se*, after taking into account the effects of observable, presumably productivity-related, factors (e.g., education, seniority, etc.). Such analyses typically find these factors to explain about half of the wage gap. Yet mainstream economists who prefer to believe that 'market forces' are stronger than 'social biases' still find ways to avoid interpreting these results as evidence of discriminatory employer behavior. The remaining gap is usually attributed, instead, to variations in some unobserved characteristics of the worker (such as 'effort,' thought to favor men) or at most to 'societal' discrimination that effects the aspirations or dedication that individual workers bring to their (presumably unbiased) employers.

Feminist economists have pointed out the limitations of using a methodology based on an *a priori* faith in markets and a narrow view of economic data (Bergmann 1989). Alternative sources of data such as that available through court cases, interviews, case studies, ethnography, and detailed-level occupational analysis, while less acceptable given mainstream methodological biases, portray a situation of continuing

sex discrimination, and the need for more integrated social and structural analysis of gendered labor market outcomes. Many feminist economic studies along such lines have been done by feminist sociologists (see *Sex Segregation at Work*) and historians.

In studying the newly industrialized countries, feminist scholars have taken a more structural approach, asking how policies of development and trade have created gendered patterns of market work. A rising question in the 1980s was whether the availability of cheap, flexible female labor had led to a pattern of 'feminization' in export-oriented sectors of many of these economies (see *Economic Globalization and Gender*).

Recent work in this area has included more attention to issues of race and sexual preference (see *Lesbians: Social and Economic Situation*) in studies of occupational segregation and wage gaps, and attention to the impact of structural change in industrialized countries.

3.3 Women's Traditional Work

While labor markets are within economists' customary domain of analysis, the historical importance of the nonmarket areas of home and family to women's economic well-being forces an expansion of the domain (Folbre 1994). Taking families as places of interesting economic production and allocation, how can these phenomena be studied? How do home and market life interact? When women move into the labor market, what happens to the tasks that they used to do? Can serious study of traditionally 'feminine' realms also change how we look at markets?

The notion that there could be something called 'home production' filtered into the discipline by the 1970s, usually in reference to constraints affecting women's labor force participation. The early models tended to perpetuate, rather than challenge, gender biases, however, by assuming that women had a comparative advantage in household work, and that the traditional allocations of home and market work reflected rational decisions given unitary 'household' preferences. A few forays into examining the inner workings of households in the early 1980s became a booming literature of bargaining models and intra-household allocation by the late 1990s (Seiz 1991, Agarwal 1997). Recognizing that households are made up of individuals who might have different and potentially conflicting interests, and different well-being outcomes, such models allow at least a limited exploration of the issues of power and perception within households (see *Family Bargaining*). The increased attention within the mainstream to intra-household phenomena may represent the most significant effect on it, so far, of feminist criticism.

In developing countries where formal markets are a smaller part of the economy, women's traditional work is likely to play an even bigger role in determining well-being (see *Economic Development and Women*). A substantial literature has developed focusing on women's 'reproductive' as well as 'productive' labor (Elson 1991, Benería 1995). In particular, macroeconomic and structural adjustment policies have come under feminist scrutiny for their differential impact on women, since the 'reproductive' sector is relied on to cushion the shocks caused by, for example, reduced public spending on health care.

Yet the division between paid and unpaid work is somewhat artificial, and the experiences of most women cannot be adequately analyzed by studying each in isolation. Sex segregation in the labor market, and the relatively low wages paid for nurturing, people-oriented activities (see *Comparable Worth in Gender Studies*), illustrate the way in which traditional expectations of, and domination of, women extends from the household into the market sphere (Hartmann 1976). Women's poverty in industrialized countries is often a combined outcome of low-paying, female-stereotyped jobs and the invisibility of needs for childcare (see *Poverty and Gender in Affluent Nations*). The need for paid childcare (and care of the elderly or ill), while most often studied as a constraint on mothers' labor force participation, is also increasingly looked at in terms of its own characteristics in creating employment—often of women less educated and more likely to be immigrants or of color compared to their employers. The division between paid and unpaid work is often also associated with a stereotypical distinction between presumably self-interested, arms-length relations in the market, and altruistic, intimate relations among family and friends. While the notion of household bargaining stretches this old dualism in the direction of recognizing individual interests within families, recent literature on 'caring labor' (see *Care and Gender*) stretches it in the direction of studying affective and socially significant relationships in the monetized sector. As family and emotional labor becomes less 'invisible' and 'natural,' and more a subject of debate and study, the myth of (male) autonomy becomes more difficult to sustain.

4. The Future of the Field

Spanning a variety of epistemological loyalties, political viewpoints, and geographical continents, feminist work promises to be a dynamic, though not unified, influence on the economics discipline for years to come. With the critique of standard economic schools relatively far along, more scholars are working towards greater development of positive feminist alternatives to the traditional paradigms. Promising routes to the development of a strong alternative to the neoclassical school include cross-fertilization from feminist work in the other social sciences, perhaps in combination with insights from institutional econo-

mics, pragmatist or postanalytical philosophy, and new views of scientific methods arising from developments in the theories of chaos and complexity. Progress is hampered, however, by resistance rooted in the traditional ideologies. While mainstream economists now tend to tolerate more study of women, or the occasional methodological article, or feminist work when it takes place in departments of cultural studies or sociology, the challenges to the core of the discipline make most economics departments and institutes reluctant to provide intellectual or institutional support for feminist studies.

See also: Gender and Feminist Studies; Gender and Feminist Studies in Geography; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender, Economics of

Bibliography

Agarwal B 1997 'Bargaining' and gender relations: Within and beyond the household. *Feminist Economics* **3**: 1–51

Benería L 1995 Toward a greater integration of gender in economics. *World Development* **23**: 1839–50

Bergmann B R 1986 *The Economic Emergence of Women*. Basic Books, New York

Bergmann B R 1989 Does the market for women's labor need fixing? *Journal of Economic Perspectives* **3**: 43–60

Elson D 1991 *Male Bias in the Development Process*. Manchester University Press, Manchester, UK

England P 1982 The failure of human capital theory to explain occupational sex segregation. *Journal of Human Resources* **17**: 358–70

Ferber M A, Birnbaum B G 1977. The 'new home economics': Retrospects and prospects. *Journal of Consumer Research* **4**: 19–28

Ferber M A, Nelson J A (eds.) 1993 *Beyond Economic Man: Feminist Theory and Economics*. University of Chicago Press, Chicago

Folbre N 1994 *Who Pays for the Kids: Gender and the Structures of Constraint*. Routledge, London

Harding S 1986. *The Science Question in Feminism*. Cornell University Press, Ithaca, NY

Hartmann H. 1976 Capitalism, patriarchy, and job segregation by sex. *Signs* **1**: 137–69

Jennings A L 1993 Public or private? Institutional economics and feminism. In: Ferber M A, Nelson J A (eds.) *Beyond Economic Man*. University of Chicago Press, Chicago, pp. 111–29

McCloskey D N 1985 *The Rhetoric of Economics*. University of Wisconsin Press, Madison, WI

Nelson J A 1992 Gender, metaphor and the definition of economics. *Economics and Philosophy* **8**: 103–25

Seiz J A 1991 The bargaining approach and feminist methodology. *Review of Radical Political Economics* **23**: 22–9

Waring M 1988 *If Women Counted: A New Feminist Economics*. Harper and Row, San Francisco

J. A. Nelson

Gender and Feminist Studies in Geography

Gender studies in geography address ways in which the material and symbolic lives of women and men are shaped by aspects of place and spatial relations and, in turn, how considerations of gender are implicated in the creation of place and space. Central within gender studies are feminist perspectives stemming from concerns about inequalities of power between women and men. In seeking points of contact between feminism and geography, Susan Hanson (1992) has identified three core analytic traditions: finding significance in everyday life; appreciating the importance of context; and thinking about difference. Bringing these traditions together, feminist geography has produced a dynamic scholarship which has introduced new themes about women's lives into the discipline, recognized the importance of space and place in understanding gender, and increasingly elaborated its focus to include the intersections between gender and other social inequalities such as those originating in 'race' and ethnicity, class, age, physical ability, religion, and nationality. The sensitivity to context that informs feminist scholarship implies not only that expressions of gender are shaped by context, but also an understanding that the creation and uses of knowledge are 'situated,' that is, they are closely tied to the cultural and political contexts in which the research is located. Given this position, interpretation of gender and feminist studies in geography has to recognize that the field is not unitary but diverse, incorporating multiple perspectives and ways of knowing.

1. The Initiation of Feminist and Gender Studies in Geography

Geographic research on gender began in the early 1970s as an academic expression of the contemporary women's rights movement and its concerns with equality, a phenomenon recognized internationally with the establishment in 1975 of the United Nations Decade for Women. Early publications came from North America, the UK, and The Netherlands, the first widely known pieces appearing in the radical journal *Antipode* in 1973. This writing reflected its political origins, addressing the unequal status of women in society and also drew attention to women's marginal positions within the geographic profession (Zelinsky et al. 1982). Studies reflected the different contexts in which they were created. Whereas Dutch initiatives emerged from student critiques of Marxist writing and relatively quickly turned to the applied and policy-related work that characterizes much Dutch geography, early research in the USA followed empirical, positivist traditions, documenting differences between women's and men's spatial behavior, for example, in daily travel patterns or migration.

From these beginnings, gender studies in geography have become global in scope (Monk 1994). Formal or informal study groups hold meetings and/or produce newsletters in Argentina, Australia, the UK, Canada, Catalonia, German-speaking countries (Austria, Germany, Switzerland), New Zealand, and perhaps other locations; within the European Union, an Erasmus program has linked feminist geographers in Denmark, Greece, The Netherlands, Spain, and England. A strong international network comes together through the Commission on Gender in the International Geographical Union. Within the latter, symposia were organized as early as 1981, a formal Study Group was approved in 1988, and the Commission was recognized in 1992. It links over 400 scholars in more than 50 countries. Feminist work is well represented in a number of mainstream geographic journals, especially in the English-language literature and in Spain (García-Ramon and Caballe 1998). A specialized journal of feminist geography, *Gender, Place and Culture* was founded in 1994. Reviews of current feminist literature also appear regularly in *Progress in Human Geography*. Nevertheless, the field remains barely visible in some parts of the world, notably eastern Europe and east Asia. The range and diversity of participation, however, means that gender studies address numerous topical themes and that research is conducted in multiple paradigms. Still, there are common strands, primarily the understanding of the concept of gender as a complex and varying social, cultural, historical, and geographic construction, not simply a binary biological distinction between women and men, that power relations are implicated in gender relations, and that space, place, and gender are interdependent.

1.1 Feminist Critiques

Important from the beginning was critique of the extent and nature of the masculine assumptions underlying the content, theories, methods, and purposes of most geographic research (Monk and Hanson 1982). Scholars pointed out that research on migration, for example, generally presented men as key actors and decision-makers and women as passive followers, despite an absence of studies confirming how decisions were made. Data also showed women as numerically dominant in migration streams at particular times and places, for example, in rural to urban migration in many Latin American countries. Urban travel models dealt with the journey to work assuming full-time employment (implicitly male), rather than attending to the more complex patterns of part-time (usually female) workers, or people combining responsibilities for household and paid work. Economic geographers usually studied formal economies in which men were the major participants (such as industrial work or export-oriented agriculture), rather than the service sector, or the informal or subsistence sectors in which women often worked. Feminist geographers advocated attention to arenas more central to women's lives, and introduced issues such as child care, housework, clerical employment, or women's fears in public spaces.

Methodologically, problems were identified with the limitations of secondary data sources such as censuses for their failure to capture women's lives, with primary data collection practices such as interviewing the male 'head' of the household, and with the use of survey methods that did not offer the subjects of research the opportunities to frame their own experiences. Feminist scholars advocated qualitative and in-depth investigations with the hope of learning how women saw their own worlds and in an attempt to modify hierarchical power relations between the researcher and those being studied.

1.2 Theorizing and Documenting Women's Material Lives

Going beyond critique, early theorizing, especially in British geography, elaborated the concept of patriarchy—social control by men originating in the authority of the father—as the source of women's oppression. Included in the arguments about patriarchy were claims that its intersections with capitalism created dual and interlocking sources of oppression. This line of theorizing was compatible both with the socialist tradition in British radical geography, with which early feminists identified, and with the substantive emphasis in research on women's material experiences, especially in urban settings. Distinctions were drawn between the city, as the site of male paid employment and public privilege, and the suburb as the site of domesticity, a private space that served as a basis for women's secondary status. This work begins to show how feminist geographers link the material world with the symbolic and ideological as they strive to understand gender-based inequalities. Though the theoretizing initially emphasized the urban scale, other scales were recognized as important. The research also highlighted the importance of place as well as space in the dynamic conjunction of capitalist and patriarchal processes. An enduring piece representing this perspective is McDowell's and Massey's (1984) analysis of ways in which different forms of economic development in four regions of Britain perpetuated male dominance, even as the specifics of economic patterns changed between the nineteenth and late twentieth centuries.

While the theoretical writing emphasized patriarchal and capitalist ideologies, a great deal of the empirical work was undergirded by related concepts useful for understanding the material aspects of labor. Numerous studies have been devoted to documenting and understanding gender divisions of labor and the

ways in which productive and reproductive work were gendered and interdependent. Attention to women's reproductive work brought new subject matter into the geographic literature, for example, on domestic employment, child care, and community-organizing. Research has been conducted in many geographic contexts, for example, on the work of Spanish rural women which demonstrates marked regional differences in its nature and extent (García-Ramon et al. 1991), and especially in the Third World where scholars have identified relationships between the gendering of work and development. This research has brought into gender studies not only scholars from North America, Europe, and Australasia, but has significantly engaged African and Asian geographers (Momsen and Kinnaird 1993). Documenting the details of daily and seasonal space-time divisions of labor in rural settings, they reveal the substantial economic contributions of women, identify relationships between household structures and labor supply and exploitation, demonstrate how the introduction of contract labor schemes, of the rearrangement of work hours, and of development schemes promoting agroforestry or irrigation have frequently increased women's burdens and fostered changes in household and community relations without improving economic outcomes for local people. Research on service and industrial employment at local scales similarly shows that from Malaysia to Massachusetts and Mexico, contemporary labor recruitment practices are implicitly and explicitly gendered with consequences for regional and international flows of migrants, daily patterns of movement, the organization of household work, and the persistence of gender-based occupational segregation (Cravey 1998, Hanson and Pratt 1995, Momsen and Kinnaird 1993).

2. *Gender and Women's Agency*

Accompanying the focus on women's oppression, disadvantages, and gender inequalities has been research on women's agency in shaping the spaces and places of their lives. Research in political geography and on environment and development provide good examples. Both exemplify ways in which feminist scholarship has contributed to revealing the linkages between not only production and reproduction but also between 'public' and 'private' spaces, questioning the validity of envisioning these as binary divisions. Studies of political struggles in Latin America (Radcliffe and Westwood 1993) show women invoking their 'private' identities as mothers to enter the public spaces to protest against state violence. These women have transgressed in public spaces, for example, seeking out clandestine cemeteries, taking and publishing photographs of the tortured and 'disappeared,' and visiting prisons. Yet they also incorporate traditional 'private' caring responsibilities of women into their public projects—organizing health and child care services and providing food and clothing for refugees. Similarly, research on Palestinian women under Israeli occupation argues that their resistance in the public political arena has been sharpened by military incursions into the secluded private spaces of the home, and by their needs to defend their children in the streets (Mayer 1994).

The motivations and tactics of environmental protests have also been interpreted in gendered terms. Women's activism may be stimulated by concerns for the health of their children and families; their public protests may draw on domestic symbols in public protests (for example, the Spanish women who nightly banged their *cacerolas* (saucepans) in protest against toxic waste dumps (Bru, in Rocheleau et al. 1996)). However, the feminist geographic literature on the environment reveals complex relationships. In this vein, Rocheleau et al. (1996) elaborated the concept of feminist political ecology. They demonstrate the importance of gender differences in access to and control over resources; how gendered knowledge is reflected in a science of survival; and the gendered nature of environmental politics and grassroots activism that is linked with women's redefinition of their identities and recognition of new environmental problems.

3. *Difference, Identity, and Embodiment*

In the 1990s, feminist geography has been strongly influenced by interdisciplinary currents, especially by the linguistic turn in the social sciences which has foregrounded discourse analysis, and by theoretical perspectives emanating from postmodernist, postcolonialist, and 'minority' feminist scholarship. Scholars have questioned grand narratives, the presumed centrality of white, Western feminisms, and the essentialism of the category 'woman.' They have focused on destabilizing dichotomous categories, such as man/woman, public/private, self/other, mind/body, and global/local (Duncan 1996). This work is particularly developed in the English-language literature, though is also attracting attention in, for example, Spanish and German-language feminist geography.

The implications of the differences among women on such bases as 'race,' ethnicity, class, sexuality, life stages, nationality, and so on are explored, as are ways in which identities based in these sources of difference are mutually constituted, situationally fluid, formed through discourse and everyday actions, and vary across specific historical and geographic contexts and spatial scales. Research explores hegemonic processes, resistances to them, and the negotiation of identities. Linda McDowell (1999) synthesizes this work, emphasizing life in post-industrial societies. Her discussion

attends to home and workplaces, the public spaces of the street and pleasure, and the nation-state. She challenges naturalized assumptions about gender and about these spaces, for example, depicting the home as simultaneously a place of safety and a trap for women with young children, a haven after work or a place where relations between gender and age have to be continually negotiated; she identifies public spaces of parks and streets as sites that can be liberatory or generative of fear and danger.

This literature recognizes spaces and places, as well as identities as simultaneously open and bounded, within and across geographic scales. How are 'hybrid' identities created, for example, by immigrants who move across international boundaries to settle in specific local places? The fluidity and contingent nature of identity in relation to place is highlighted. While some work looks to the present, past imperialisms have received significant attention as the historical discourses of women's travel narratives are re-read. Geraldine Pratt provides multiple examples of the fluid and contextual expressions of identity (Fincher and Jacobs 1998). Among the cases she presents is that of a Victorian British woman traveller in west Africa who is bound by gender at home but enabled by class and race privileges in Africa; in another example, she explores the shifting value positions on ethnicity and expressions of gender roles elaborated by Canadian couples in a multicultural neighborhood in Vancouver as they struggle with making decisions about caretakers (and care-taking locations) for their children; yet again, she speculates on the impacts on the political values of women in Worcester, Massachusetts, of moving daily between homes that are middle class (reflecting their husbands' status) and workplaces where they are side-by-side with women of working class backgrounds.

Perhaps the most notable departure in the 1990s from previous geographic work is the focus on the scale of the body and on links between sexuality and space, especially homosexuality. Research examines ways in which particular bodies—the pregnant, the homosexual, the disabled, and the non-white—are marked in space often in exclusionary ways, and also how 'performances' of gender and sexuality are place-related (Duncan 1996, McDowell 1999).

While this new work introduces exciting possibilities, cautionary notes are sounded. On the one hand, it links feminist geography with other critical scholarship concerned with challenging oppression (Jones et al. 1997). The recent collection, *Cities of Difference* (Fincher and Jacobs 1998), for example, brings together feminist work on gender (inflected by other sources of difference such as life stage) with comparable studies relating to disability, gay male identity, race/ethnicity, religious convictions, and class. On the other hand, concerns exist that decentering gender could lead to a depoliticized relativism in which a hard-won feminist political stance is silenced. So long as the research retains its focus on inequalities and power relations, however, and seeks possibilities for collaborations and coalition formation, the development may be more effective in achieving feminist goals than concentration on a unitary, and some would argue exclusionary, identity politics.

4. *Reflections on Methods and Methodologies*

Critical perspectives on the creation of knowledge have been enduring issues in gender and feminist studies. While some discussions address specific issues of methods, particularly the relative merits of qualitative and quantitative research, the fundamental concerns are with the politics involved in creating and representing understandings of the world and the partiality of all knowledge, in both senses of this word—that is, no story is considered complete or objective. Overviews (McDowell 1999), collected essays (Jones et al. 1997) and special sections of mainstream journals have addressed the theme, such as the series of articles in *The Canadian Geographer* in 1993 on aspects of ethnographic methods, in *The Professional Geographer* in 1994 on dilemmas of field research, and in the same journal in 1995 on appropriate approaches to, and the values of, quantitative work. They highlight issues of power and ethics. How does one set a research agenda/define a research question if the goals include being responsive to the people being researched and contributing to the feminist agenda for social change, while remaining constrained by the expectations and conventions of the academy? Rocheleau (Rocheleau et al. 1996) describes her involvement of rural people in the Dominican Republic in a multi-faceted research project on agro-forestry that incorporated both 'top down' and 'bottom up' generation of questions, categories of analysis, and interpretations. Gibson-Graham reports her use of an interactive, dialogic method in which women in Australian mining towns were involved in implementation of the research, engaging their domestic and community politics (Duncan 1996).

Numerous scholars reflect on power issues of their own relationships with 'subjects' especially as their interviews involve people who are of different class, 'race,' ethnic, gender, or sexual orientation from themselves and who may express views that are problematic for the researcher (for example, considering them to be sexist or racist) or who reveal knowledge that could compromise the subject or researcher (Jones et al. 1997, McDowell 1999). While such questions have been widely raised in interdisciplinary feminist literature, as Hanson (Jones et al. 1997) points out, perhaps the distinctive and most interesting contributions from geography concern the roles of location and place within the research process—in

what settings does the recruitment of 'subjects' take place, where are interviews held (homes, workplaces, community contexts)? How does the choice of context shape not only the way subjects present themselves but how they interpret the researcher's intentions and meanings? In a revealing account, Nagar identifies multiple ways that she was viewed and presented herself in research carried out in religiously, regionally, and ethnically diverse social spaces in Asian communities in Dar-es-Salaam (Jones et al. 1997). Overall, these discussions point to the importance of reflexivity that requires researchers to continually evaluate their assumptions and political positions. They also lead to a conclusion that favors combining multiple methods and exploring ways to engage the 'researched' in the validation of one's work.

Discussions about representation also figure prominently in feminist geography—not only asking whose voices are heard in the research, but also how categories are imposed on what we hear, and on where and how research is reported and read. Pulido, for example, writing on the activist group, Mothers of East L. A. (Los Angeles) indicates that while feminist work has identified women's environmental activism with their gender identities, the women involved may not see themselves and their motivations in gendered terms, but rather in relation to ethnic or community identities (Jones et al. 1997). Monk notes the irony of the hegemony of British and US scholarship in the reviews of literature by feminist geographers who simultaneously promote resistance to hegemonic discourses and listening to a plurality of voices (Jones et al. 1997) while overlooking publications written outside their own milieux. One of the most interesting examples of attempts to present research in ways that will speak to different audiences is Pulsipher's analysis of the diverse interpretive texts she has created to communicate her research in historical geography/archeology on a slave plantation in Montserrat to multiple audiences including local project workers and outside volunteers assisting in the research (and especially to engage locals in defining concepts), as well as museum curators and visitors at the Smithsonian Institution. She has also written several fictional accounts using the research material to convey the bodily feelings of life on the plantation to local schoolchildren, community members, outside visitors, and academic audiences (Jones et al. 1997).

Surprisingly rare in these methodological discussions, especially given the proliferation of cartographic metaphors in contemporary post-modern and feminist writing, are reflections on the place of mapping in feminist geography and gender studies, though several feminist atlases (for example, of women in the USA, the *Atlas of Women and Men in India* (Raju et al. 1999), and especially Seager's (1997) world atlas) present examples of the possibilities and difficulties of representing gender themes cartographically. Huffman's critique of the masculine values of traditional cartography, and his reading of feminist endeavors, opens up methodological questions, such as the ways maps and other visual representations can present fragmented data, reveal hidden perspectives, and contribute to interpretations of gendered bodies (Jones et al. 1997).

5. *Status and Emerging Directions*

In little over two decades, gender and feminist studies have become well established in geography despite their marginal origins. In particular, their critical perspectives have contributed to reshaping epistemological understandings in human geography. In addition to the substantive themes mentioned above, work is expanding on gender and the body (including sexuality and health), on gender and human rights, trans-national cultural, political, and economic processes and their intersections, and constructions of masculinity. There is every indication that gender and feminist studies will remain a vital field within geography.

See also: Feminist Theory; Nation-states, Nationalism, and Gender

Bibliography

Cravey A J 1998 *Women and Work in Mexico's Maquiladoras*. Rowman and Littlefield, Lanham, MD

Duncan N (ed.) 1996 *Bodyspace: Destabilizing Geographies of Gender and Sexuality*. Routledge, London

Fincher R, Jacobs J M 1998 *Cities of Difference*. Guilford Press, New York

García-Ramon M D, Caballe A 1998 Situating gender geographies: A bibliometric analysis. *Tijdschrift voor Economische en Sociale Geografie* **89**: 210–16

García-Ramon M D, Cruz Villalon J, Salamaña L, Valdovinos N, Vilariño M 1991 Women and farm households: Regional variations in gender roles and relations in Spain. *Iberian Studies* **20**: 81–112

Hanson S 1992 Geography and feminism: Worlds in collision? *Annals of the Association of American Geographers* **82**: 569–86

Hanson S, Pratt G 1995 *Gender, Work, and Space*. Routledge, London

Jones J P III, Nast H J, Roberts S M 1997 *Thresholds in Feminist Geography: Difference, Methodology, Representation*. Rowman and Littlefield, Lanham, MD

Mayer T 1994 *Women and the Israeli Occupation: The Politics of Change*. Routledge, London

McDowell L 1999 *Gender, Identity and Place: Understanding Feminist Geographies*. University of Minnesota Press, Minneapolis, MN

McDowell L, Massey D 1984 A woman's place? In: Massey D, Allen J (eds.) *Geography Matters*. Cambridge University Press, Cambridge, UK

Momsen J H, Kinnaird V (eds.) 1993 *Different Places, Different Voices: Gender and Development in Africa, Asia and Latin America*. Routledge, London

Monk J 1994 Place matters: Comparative international perspectives on feminist geography. *The Professional Geographer* **46**: 277–88
Monk J, Hanson S 1982 On not excluding half of the human in human geography. *The Professional Geographer* **34**: 11–23
Oberhauser A M, Rubinoff D, De Bres K, Mains S, Pope C (in press) Geographic Perspectives on Women Specialty Group. In: Gaile G, Wilmot C (eds.) *Geography in America at the Dawn of the 21st Century*. Oxford University Press, Oxford
Radcliffe S A, Westwood S (eds.) 1993 *Viva: Women and Popular Protest in Latin America*. Routledge, London
Raju S, Atkins P J, Kumar N, Townsend J 1999 *Atlas of Women and Men in India*. Kali for Women, New Delhi, India
Rocheleau D, Thomas-Slayter-B, Wangari E 1996 *Feminist Political Ecology: Global Issues and Local Experiences*. Routledge, London
Seager J 1997 *The State of Women in the World Atlas*. Penguin Reference, London
Zelinsky W, Monk J, Hanson S 1982 Women and geography: A review and prospectus. *Progress in Human Geography* **6**: 317–66

J. Monk

Gender and Feminist Studies in History

Gender history, feminist history, and women's history identify overlapping areas of inquiry whose relationships to each other have become more complex and more contested in the past decade. Emerging as an autonomous subfield by the early 1970s, women's history was a direct outgrowth of the feminist movement in the USA, Canada, Australia, and Western Europe. Certainly, earlier generations of women and of feminists had written on the historical experiences of their sex, but most did so outside the institutions and practices of professional history (Smith 1998). Only in the 1970s did women historians gain access to professional training and teaching positions in colleges and universities in significant numbers and in a sufficient array of countries to nurture this new field of study. A small number of men participated in launching women's history and a growing number employ gender analysis, but women remain the mainstay of the field.

Feminist history was initially conceived as nearly synonymous with women's history in the sense that most practitioners of the latter embraced the politics of the former, including critiques of patriarchal institutions and practices and of maldistributions of power and resources more generally. But feminist history could also define an area of investigation—the origins and development of movements for women's rights and sex equality. Gender analysis appeared as a distinct but related approach to history in the mid-1980s, though a concern with the 'social relations of the sexes' and the relative status of women and men had certainly been introduced earlier. This approach both complemented and challenged existing work in women's and feminist history by asserting the critical importance of exploring historical changes in individual and collective subjectivities and the centrality of gender in signifying forms and relationships of power (Kelly 1986, Scott 1988, *Gender & History* 1989).

The emergence of gender history as a distinct area of inquiry paralleled and converged with a number of other theoretical and professional developments that assured the transformation of feminist history. These included the rapid increase in women's history scholars, journals, books, and conferences; the resurgence of conservative and right-wing politics around the world; the rise of postmodern, postcolonial, and subaltern epistemologies; and the critique of early feminist frameworks by scholars of color and by historians of the working class, lesbians, and non-western regions and countries (Brown 1992, Meyerowitz 1992, Alexander and Mohanty 1997, Pierson and Chaudhuri 1998).

1. *The Rise of Gender History*

From the 1960s on, women's and feminist historians had incorporated some gender perspectives in their work. Not until the mid-1980s, however, did the history of gender become a substantial field in its own right. Proponents of gender analysis argued that 'the creation and reproduction of gender is a process, changing over time' and place, and operating 'at every level, and in every sphere' of historical study (*Gender & History* 1989, pp. 2–3). The ubiquity of gender—as a system of symbols and signifiers as well as a cultural construction with implications for society, the economy and politics—demanded 'a theory of meaning that assumes a multiplicity of references, a resonance beyond literal utterances, a play across topics and spheres' (Scott 1988, p. 60). In addition, gender analysis was seen as a critical weapon in transforming traditional historical narratives. Women's history could be accepted as a parallel but separate area of inquiry, but gender history—which recognized gender 'as a key axis of power in society' and provided 'a crucial understanding of how any society is organized and structured'—demanded the recasting of the past *in toto* (Hall 1992, p. 12).

Responding to earlier challenges that feminist history had long been too white, too middle-class and too heterosexual, many gender historians engaged issues of class, race, nationality, and sexuality from the beginning. They asked how class was gendered, how gender was raced, and how nationality was raced, gendered, and sexualized. Davidoff and Hall, for instance, explored the critical role of women and the family in the formation of the British middle class; Brown suggested that jazz and gumbo ya-ya were appropriate models for capturing the multivocal,

relational character of an American women's history riven with race and class differences; and Guy analyzed the ways that fears of 'white slavery,' that is, the international trade in women for sex, shaped notions of citizenship, and nationality in Argentina (Davidoff and Hall 1987, Brown 1992, Guy 1992)

There was considerable resistance, however, to the emergence of gender analysis among some feminist historians. Scholars' emerging fascination with men and masculinities led some to worry that women would quickly be subsumed within this new gender history. Other questions arose when pioneering advocates of gender analysis embraced postmodernism and deconstruction. Postmodernist critiques of grand narratives and the deconstruction of unified subjects and subjectivities posed challenges to fields like women's and feminist history that were seeking to create new narratives by building on the identity politics of the 1960s and 1970s (*International Labor and Working Class History* 1987). Nor was this concern limited to battles among white women historians. Scholars of color, committed to integrating race and nationality into studies of women and gender, were divided over the utility of postmodernism for creating new political insights and practices (Christian 1987, Alexander and Mohanty 1997). Students of sexuality, too, were alternately attracted to and repelled by theoretical perspectives that denied unitary subjects, homosexual and heterosexual alike (Meyerowitz 1992). Critics focused on the political dangers of destabilizing categories of individual and group identity at a moment when identity politics seemed to promise a potent weapon in the fight against racism, sexism, classism, homophobia, colonialism, and imperialism. In a more purely historical vein, the attack launched by postmodernists threatened not only older grand narratives, but also newer ones that celebrated the lives of women, workers, African Americans, indigenous peoples, and anticolonial activists. Yet despite such concerns, gender analysis has increasingly shaped work in women's and feminist history. And taken together, these approaches have begun to transform the study of the past in several significant ways.

2. *The Transformative Power of Feminist and Gender History*

Certain themes and interpretations provide common threads in the history that is being rewoven through analyses of women and gender. Perhaps most importantly, private life, domesticity, reproduction, and sexuality have been reclaimed as subjects. Not only do we now recognize that these topics have a history that parallels that of public life, politics, production, and war, but we also are beginning to see the ways that these two histories are deeply intertwined—private with public lives, domesticity with politics, reproduction with production, sexuality with war.

In recognizing this broader tapestry, studies of women and gender have transformed the meanings of terms central to historical analysis. Politics no longer refers only to the formal institutions of government and law, but also to mass movements, voluntary associations, maternalist campaigns, spontaneous uprisings, and alternative understandings and conceptualizations of power in workplaces, schools, cultural sites, and homes as well as courts, legislatures, parliaments, and partisan arenas. Work is no longer limited to labor performed by men in exchange for land, goods, or wages, but embraces the production of goods and services by women and men, regardless of payment or the system of exchange, and includes the labor of physical and social reproduction and the contributions of informal as well as formal economies. In a similar fashion, studies of war are no longer confined to the battlefield or the arenas of military strategy and diplomatic negotiation, but are instead open to analyses of the homefront, of low-intensity but chronic conflict, and of rape, disappearances, and torture as weapons of domination.

The history of labor—productive and reproductive—offers numerous examples of how attention to women and gender recasts our understanding of the transition to capitalism. For example, women were critical economic agents in many arenas during the medieval and early modern periods, particularly in domestic production, social reproduction, informal sectors, and petty trade. Bennett's (1996) *Ale, Beer and Brewsters in England*, illustrates the centrality of women to the production and sale of alcoholic beverages in the early medieval period, when brewing was a largely local affair. She then analyzes the process by which female brewsters gradually were pushed to the margins as larger markets, new technologies, government regulation, and gild formation turned brewing into a masculine venture. In this transition to commercial capitalism, women remained tied to local, low-paying, low-status work, while the skills they had honed in home-based shops were taken over by men once they proved sufficiently profitable. In similar fashion, Ulrich's exploration of a midwife's contributions to household sustenance and social interaction in a small town in late eighteenth- and early nineteenth-century Maine illuminate the significance of women's labor in maintaining families and communities and its contributions to larger transformations in the gendered economies in which they were embedded (Ulrich 1990). Studies of slave women across time and place and of female service, factory, and home-based workers in the industrial age add to these early portraits of women's place in both formal and informal economies (Berger 1992, Morgan 1997)

In a similar fashion, the study of politics has been transformed by the inclusion of women and of gendered perspectives. Initially feminist scholars shifted the focus away from electoral, parliamentary, and judicial activities, nearly all of which excluded

women—at least formally—until the last half century. Instead women's historians studied the means by which women and other marginalized groups shaped the political terrain through voluntary associations, petitions, food riots, mass demonstrations, and reform and revolutionary movements. Interpetrations of the most significant periods of reform in the USA—the 1830s–1850s, the 1890s–1920s, and the mid-1950s–mid-1970s—have now been rewritten to highlight the public activities of women and their larger social, political, and cultural consequences (Hewitt and Lebsock 1993). Scholars of the French Revolution and the revolutions of 1848 across Europe have highlighted both women's participation and the use of women as symbols, alternately, of liberty and degradation. And interpretations of nationalist movements and national liberation struggles in South Africa, India, Spain, Cuba, and other areas of the world are being rewritten to include women activists and gendered images (Landes 1988, Riot-Sarcey 1994, Pierson and Chaudhuri 1998).

In developing new areas of inquiry and new perspectives on traditional subjects, feminist and gender historians have been joined by historians of other groups long marginalized in historical narratives–racial and national minorities, workers, peasants, slaves, and those ostracized for sexual practices or by sexual traumas. Race, class, and sexuality were never absent from women's history, but only from the mid-1980s on did explorations of differences among women gain equal attention with investigations of differences between women and men (Brown 1992, Alexander and Mohanty 1997, Gordon 1999). In recent years, scholars have developed more and more sophisticated analyses of the interplay among race, class, nationality, and sexuality, with a growing number of feminist historians working to reveal their mutually constitutive character. Even studies of feminism and women's emancipation have been recast in terms of race, class, and nationality. In particular, the ways that many white feminists excluded women of color from their ranks and/or wielded racist or ethnocentric logic to advance their own cause has now been sharply critiqued even as historians of women's movements continue to search for moments of alliance and collaboration across these divides (Burton 1994, Pierson and Chaudhuri 1998).

Many fields have been influenced by this attention to race, class, and nationality within women's history. Research on the history of Africa and the African diaspora has been especially important in excavating the intricate links among women's multiple forms of identity. Gendered images of Africa, particularly images of African women as beastly and oblivious of pain in childbirth, justified European conquest, colonization, and enslavement. The skewed sex ratios that characterized the slave trade—with far more men than women being sent to plantations in the Caribbean and the Americas—necessitated transformations in family life and in productive and reproductive labor and over time shaped new patterns of ethnic, regional, and even religious identity in both Africa and the Americas. Moreover, while acknowledging the overwhelmingly racial character of slavery, historians have now demonstrated the powerful ways that gender shaped the experiences of bondage and of freedom. And in recent years, studies of twentieth-century Africans and African Americans have explored the ways that class, religion, and region as well as gender shaped divisions of labor, movements for social reform and social justice, and struggles for national liberation (Berger 1992, Morgan 1997, Diouf 1998).

As the mosaic of women's and gender history has expanded, a growing number of scholars have turned their attention to conservative women and women on the right. They challenge earlier feminist histories that highlighted women's engagement with progressive, emancipatory, and pacifist struggles or that treated women's engagement with conservative, right-wing, and fascist politics as solely a function of their domination by men. Rather, building on scholars' efforts to reveal women's agency in all aspects of their lives, historians of conservative and right-wing women have insisted that we recognize many women's willing embrace of antifeminist, racist, nativist, and elitist policies and projects. Beginning in the 1980s with works such as Koonz's (1987) *Mothers in the Fatherland: Women, the Family and Nazi Politics*, a number of historians have argued that women who accept traditional family and community roles often feel threatened by efforts to transform their place in society and in response support movements that celebrate motherhood, domesticity, and female subordination (Gravenhorst and Tatschmurat 1990, Sarkar and Butalia 1995). Most of these studies focus on women's engagement with the far right—facism, the Ku Klux Klan, and religious fundamentalism. Only recently, have scholars explored women's participation in more mainstream movements, demonstrating not only their support of conservative political and social agendas but also the ways that that support has undermined feminist claims for women's universal support of peace, social welfare, and other progressive causes (Delegard 1999).

These studies, though their subjects are mainly members of dominant racial and ethnic groups, reinforce claims made by historians of people of color, workers, and lesbians, for the significance of differences among women. They also hark back to earlier work on social reform that demonstrated that women and men who shared economic and racial status often held common views on social order and social change. And they fit neatly alongside more recent studies in which the social relations of the sexes take center stage in analyzing the dynamics of social, political, and cultural change (Kelly 1986, Davidoff and Hall 1987, Guy 1992, Berger 1992, Roberts 1994, Sinha 1995, Bennett 1996, Gordon 1999).

3. Globalization and Revitalizing Gender and Feminist History

The issues raised by gender history were complicated further in the 1990s by the call for more global histories. Certainly earlier work focused on diverse areas of the world. Nonetheless, the increasing globalization of the economy and of communications and the emergence of women's movements and women's histories in a growing number of nations has demanded renewed attention to international and transnational perspectives. Many nations gained their independence in the final decades of the twentieth century, even as they remained financially and culturally colonized by Western industrial powers. The emergence of women's and gender histories in newly independent nations has provided not only additional stories of women's past struggles but has also demanded a rethinking of the roles of women, and feminists, in imperial and colonial politics and culture from the seventeenth century to the present (Davis 1995). Studies of women, gender and sexuality as they shaped colonial relations in South Asia have been especially important in exploring this rich new terrain (Burton 1994, Jayawardena 1995, Sinha 1995).

Interest in comparative and transnational work and in queer theory, which extends postmodernist critiques of unitary subjects into the realm of sexuality and beyond, promise even more complex renderings of women and men, femininities and masculinities in motion around the world (Alexander and Mohanty 1997). Finally, with a profusion of work on women, feminism, and gender now appearing in a wide array of countries and regions and on a broad range of racial, ethnic, national, and sexual communities, differences within as well as between groups of women are increasingly highlighted, making generalizations more difficult even as it often makes comparisons more fruitful. Thus, for instance, differences among women of color rooted in class, color, and sexual orientation, are being explored in greater depth. Similarly, ideological differences among feminists are being analyzed at the same time that their international connections—past and present—are being highlighted and the differences in conceptualizations of women's rights and priorities are being explored in a more global context. The challenges that new and forthcoming work pose to existing interpretations combined with the rich array of research by scholars claiming identities as feminist, women's *and* gender historians suggests that these fields will continue to have profound effects on the development of history.

See also: Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Feminist Economics; Feminist Epistemology; Feminist Legal Theory; Feminist Political Ecology; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Feminism, and Sexuality in Archaeological Studies; Klein, Melanie (1882–1960); Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns; Population Dynamics: Two-sex Demographic Models

Bibliography

Alexander M J, Mohanty C T 1997 *Feminist Genealogies, Colonial Legacies, and Democratic Futures*. Routledge, New York

Bennett J 1996 *Ale, Beer, and Brewsters in England*. Oxford University Press, New York

Berger I 1992 *Threads of Solidarity*. Indiana University Press, Bloomington, IN

Brown E B 1992 What has happened here? The politics of difference in women's history and feminist politics. *Feminist Studies* **18**: 295–312

Burton A 1994 *Burdens of History*. University of North Carolina Press, Chapel Hill, NC

Christian B 1987 The race for theory. *Cultural Critique* **6**: 51–63

Davidoff L, Hall C 1987 *Family Fortunes*. Hutchinson, London and University of Chicago, Chicago

Davis N Z 1995 *Women on the Margins*. Harvard University Press, Cambridge, MA

Delegard K M 1999 Women patriots. Ph.D. Thesis, Duke University

Diouf S 1998 *Servants of Allah*. New York University Press, New York

Editors 1989 Why gender and history. *Gender & History* **1**: 1–4

Gordon L 1999 *The Great Arizona Orphan Abduction*. Harvard University Press, Cambridge, MA

Gravenhorst L, Tatschmurat C 1990 *TöchterFragen*. Kore, Verlag Traute Hensch, Frieburg, Germany

Guy D 1992 White slavery. In: Parker A, Russo M, Sommer D, Yaeger P (eds.) *Citizenship and Nationality in Argentina, Nationalisms and Sexualities*. Routledge, New York

Hall C 1992 *White, Male and Middle Class*. Routledge, New York

Hewitt N, Lebsock S 1993 *Visible Women*. University of Illnois Press, Urbana, IL

Jayawardena K 1995 *The White Woman's Other Burden*. Routledge, New York

Kelly J 1986 The social relations of the sexes. In: *Women, History, and Theory*. University of Chicago Press, Chicago

Koonz C 1987 *Mothers in the Fatherland*. St. Martin's Press, New York

Landes J 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca, NY

Meyerowitz J 1992 American women's history. *Canadian Review of American Studies* **23**: 27–52

Morgan J 1997 Some could suckle over their shoulder. *William and Mary Quarterly* **54**: 167–92

Pierson R, Chaudhuri N 1998 *Nation, Empire, Colony*. Indiana University Press, Bloomington, IN

Riot-Sarcey M 1994 *La démocratie à l'épreuve des femmes*. Michel, Paris

Roberts M L 1994 *Civilization without Sexes*. University of Chicago Press, Chicago

Sarkar T, Butalia U 1995 *Women and the Hindu Right*. Zed Press, London

Scott J 1988 *Gender and the Politics of History*. Columbia University Press, New York
Sinha M 1995 *Colonial Masculinity*. Manchester University Press, Manchester, UK
The Language of Class 1987 International labor and working class history. **31** (Special issue)
Smith B 1998 *The Gender of History*. Harvard University Press, Cambridge, MA
Ulrich L T 1990 *A Midwife's Tale*. Random House, New York

N. Hewitt

Gender and Feminist Studies in Political Science

There are two main issues which feminism continues to pose for political science. The first is the questioning of any conception of social science that sees disciplines as neutral and universal bodies of knowledge. The second, and closely related issue, is the limit of the political. In short, not only have feminists argued that traditional political science was very far from gender neutral; they have also argued that, in order to analyze politics adequately, and incorporate gender issues, it is necessary substantially to extend more traditional political science understandings of what constitutes 'politics.' In the process, feminist work has connected with wider analyses of gender, emphasizing the need to analyze gender politics in terms of the construction of both masculinities and femininities.

1. *Political Science as Neutral*

From the 1970s on, feminist academics increasingly challenged both the gender neutrality of the focus of political science and the gender neutrality of the discipline itself. As Squires has pointed out: 'until the emergence of feminist theory as a recognized academic perspective, contemporary political theory was largely assumed to be gender-neutral in focus' (Squires 1999). Not only did that neutrality come to be challenged but so did women's largely subordinate position in the political science profession. Giving her Presidential Address to the Australasian Political Studies Association in 1981, Pateman noted that: 'Women's position in the profession of political studies is very similar to our position in other professions and in public life generally; we are marginal participants in organizations that are numerically and structurally dominated by men' (Pateman 1982). While matters have improved in many respects since the time when Pateman was writing, a recent survey of the position of women in the American discipline of political science reported that women held only 21.1 percent of full-time faculty positions in 1997–8 and constituted only 11.2 percent of full professors. Improving the position of women in the discipline has been made harder by a general decline in tenure-track job opportunities (Sarkees and McGlen 1999). A survey of the impact of feminism on Canadian political science found that fewer than 640 of the nearly 17,770 pages published in the *Canadian Journal of Political Science/Revue canadienne de science politique* in the 22 years since its inception discussed women. Fortunately, the situation had improved markedly in the previous seven years (Arscott and Tremblay 1999).

2. *The Limits of 'Politics'*

Most challenging of all to conceptions of political science as 'neutral' were the challenges to the subject matter of the discipline. In particular, with the increasing impact of academic feminism, the very subject of the 'political' in political science was thrown into question. Politics, as Pateman had pointed out in her 1981 Address, tended to stop at the garden gate. Traditionally, politics was about government, the state, political parties, and other manifestations in the public sphere. Politics was not about gendered power relations in personal life, domestic labor, or women working in the volunteer welfare sector, or organizations representing sex workers in the third world, or representations of women in the media, or domestic violence or rape. Indeed, Squires has argued that it is the limited subject matter of mainstream political science that has most influenced its androcentric nature: 'The adoption of the narrow institutional or instrumental conception of the political within most dominant renderings of the discipline' is what 'accounts for the fact that the study of politics has been one of the last to take up the challenge of feminist scholarship, and most recently men's studies, and modify the cannon' (Squires 1999). Given the need to move beyond limitations of what has been defined as the political, it is not surprising that the engagement between feminists and political scientists has been continually informed by sociologists, philosophers, historians, linguists, and literary theorists rather than confined to insights generated from within the discipline (see e.g., Butler and Scott 1992).

The archetypal second-wave feminist slogan of 'the personal is political,' therefore, had a particular meaning for political science and challenged the discipline's boundaries. Initially, the tasks of feminist political science seemed relatively straightforward. Neglected areas of women's participation could be documented. The welfare state was reconceived in terms of its role in reinforcing established gender relations rather than merely analyzed in terms of the influence of Keynesian economics or postwar European social democracy. (For example, governments frequently assumed that women who were married or in *de facto* relationships should not be eligible for state

benefits. Rather, it was the job of their male breadwinner to support them.) There could be speculation regarding the possibility of undocumented feminist debates in classical Greece that may have influenced Plato and outraged Aristotle. Leading female political thinkers could be rediscovered and reassessed from fifteenth-century writers such as Christine de Pizan to nineteenth-century writers such as Flora Tristan or Harriet Taylor. Voting studies could concentrate on female voters and analyzing gender difference. The under-representation of women in parliament and the bureaucracy could be documented and critiqued. The lack of provision of services targeted for women, from childcare to women's health, could be highlighted. Feminists could analyze attempts to influence state policy and services, for example, the Australian experiment with 'femocrats,' feminist bureaucratic advisers to government (Eisenstein 1996).

Such tasks still remain a central focus of feminist political science, for example, they are amongst the aims of the *International Feminist Journal of Politics*, established in 1999. They have long been the aim of older journals such as *Women and Politics*. However, it is not a matter of just adding women on to existing analyses; feminists have argued that political science also needs to readdress some of its most fundamental theoretical and analytical underpinnings.

3. *The Reassessment of Traditional Political Theory*

Many of the issues raised by late twentieth-century feminists were not of course new. As already indicated, predecessors from fifteenth-century Christine de Pizan to better known feminists such as Mary Wollstonecraft had raised issues regarding women's relative exclusion from public life. Nonetheless, from the 1970s on, feminist political theorists along with academics from related disciplines began to reassess traditional political theory. Okin (1979) was one of the first 'second wave' feminists to argue that the fathers of modern liberal democratic thought tended to construct the citizen as a male head of household. The role of women was seen to revolve around the private sphere of the family. Women were to have only limited, if any, involvement in public, political, and economic life. Even liberal supporters of female suffrage such as John Stuart Mill assumed that, while women should be able to choose either a career or marriage, the majority of women will choose marriage. Having made the latter choice, the woman's primary commitment would be to managing the family and rearing children. Such views have led Coole (1988) to argue that, despite tendencies towards egalitarianism, liberal democrats expected women to defer to their husbands on political matters well into the nineteenth century. The issue of marriage also has a deeper significance. Pateman (1988) has argued that the marriage contract was indissolubly intertwined with the social contract. In other words, when contract theorists postulate an original social contract in which free men, living in the state of nature, joined together and agreed to give up certain individual and collective powers to government, they were telling only part of the story. Underlying the social contract was a marriage contract in which men continued to have power over their wives and their wives' property. The wage contract was similarly gendered as the employee tended to be assumed by both employers and governments to be a male wage earner supporting his family.

While there is debate about how relevant Pateman's analysis of contract is to analyzing contemporary political discourse, there is a widespread feminist view that the citizen subject is still being constituted in masculine terms. Brown (1995) paints a complex picture of how this occurs. She draws on both a conception of ideology as an obfuscating narrative and a conception of discourse as a narrative that is constitutive of a social order. Brown notes that liberalism simultaneously produces gendered subjects whose differences are seen as natural (an ideological moment in her terms), and abstract individuals who appear to be genderless (a discursive moment in her terms). The liberal individual may appear superficially to be gender neutral but is still gendered masculine in crucial ways. As numerous feminist theorists, including Brown, have pointed out, gendered dualisms operate throughout political discourse. For example, in much neoliberal discourse, categories such as self-interest are associated with the masculine, and categories such as selflessness are associated with the feminine. The citizen is meant to be an abstract, self-reliant, and entrepreneurial individual. However, far from the self-interested liberal individual being autonomous and self-reliant he is in fact dependent upon the labor and emotional support of the subordinated female 'other.' It is, therefore, hardly surprising that cutting state welfare services often results in labor being transferred to unpaid female labor in the domestic sphere. For example, as jobs and services are cut in the public sector, it is frequently women who then have to increase their level of care for sick, young, old, or disabled relatives at home. Of course, such problems are not confined to some forms of market liberalism but also appear in some of the classic works of social democracy. For example, Pateman (1996) has argued that T. H. Marshall's classic statement of the 'social' rights to which citizens were entitled was quite literally based on a conception of the citizen as male breadwinner and worker. Here, as elsewhere, women's exclusion from various 'social rights' was not addressed adequately. Not only were married women frequently denied eligibility for welfare payments but 'social' rights were conceived of according to 'male' norms. Rights to services most affecting women such as childcare were neglected. Female under-representation in parliaments and legislatures was not seen as

a major barrier to women's citizenship that needed to be urgently addressed. In short, as Phillips (1991) has pointed out, far from transcending gender, the apparently gender neutral figure is frequently based on a male norm. Feminists have therefore argued that the projects of traditional political theory were often inherently androcentric.

While the situation has improved, feminists argue that many of these problems are still apparent in contemporary western political and social theory. So, for example, Arscott and Tremblay (1999) argue that key Canadian theorists of citizenship such as Jim Tully and Will Kymlicka 'have paid next to no attention at all to gender' despite its relevance and the fact that they do deal with relevant ethnic or indigenous issues.

4. *Women and Political Representation*

The critique of traditional political science and political theory has had a strong influence on more empirical feminist analyses of women and political representation. It is not just that feminists have studied forms of politics neglected by traditional political science, but feminist analyses have also contributed new insights into mainstream political science concerns such as how politicians are selected. For example, in her analysis of Women National Leaders, Francine D'Amico reminds us that many prominent female political leaders have taken 'the widow's walk,' literally walking into the political leadership job previously occupied by their deceased husbands. Corazon Aquino, former President of the Philippines, would be a typical example. In another form of 'surrogate' politics, daughters have occupied positions occupied by their deceased fathers, for example, Benazir Bhutto, former President of Pakistan. D'Amico points out that traditional studies of political leadership tended to neglect such paths to office because of their emphasis on the 'public' rather than the 'private,' for example, on recruitment of political leaders from a background in business, the military or the professions (D'Amico 1995).

However, concentrating on a few prominent female political leaders should not blind us to the fact that, as of January 2000, only eight countries in the world, Sweden, Denmark, Germany, Finland, Norway, Iceland, the Netherlands, and South Africa had reached the UN sanctioned target of women's share of seats in national parliaments being 30 percent or above. New Zealand had reached 29.2 percent, Australia 22.4 percent, the UK 18.4 percent, and the US 12.5 percent (UNIFEM 2000). Christy has drawn attention to some of the most obvious national differences that can influence women's level and forms of political participation, even where there are not apparently huge cultural or religious differences. For example, as Christy points out in a comparison of Europe and the US, women's attitudes to politics may be different in countries with more extensive welfare states, where women can easily see the relevance of government provision in areas such as caring for children, the sick, and elderly. Electoral systems that utilize proportional representation, with party listing of candidates, tend to elect more women (Christy 1994). Multimember electorates are, therefore, a particularly important factor (Norris 1994).

The impact of women's increased political representation is still subject to debate, particularly since some prominent political leaders, such as Margaret Thatcher, have been identified with antifeminist politics. Nor can one assume that women will behave 'better' than male politicians. Nonetheless, Conway and co-workers conclude in their study of American politics that the increasing number of women in the legislatures and the political elite has already influenced policy. Female legislators are more likely to emphasize social welfare issues and support equal rights policies. They are also likely to place more emphasis on helping constituents. Their presence has 'probably led' to male members of the political elite being aware that women's issues are important (Conway et al. 1997). The claimed difference in the attitudes of female and male legislators might well reflect the persistent so-called 'gender gap' on issues such as social welfare which a number of commentators have found in longitudinal voting and opinion poll studies (see e.g., the famous study by Shapiro and Mahajan 1986).

5. *Analyzing the Politics of Masculinity*

While feminism has played a crucial role in exposing traditional political science's neglect of women, it has also drawn attention to the need for political science to analyze the role of the state and political institutions in constructing masculinity. Indeed, one could argue that political institutions are just as implicated in the construction of particular forms of masculinity as they have been in marginalizing women. If, for example, the citizen is being conceived in terms of the male citizen breadwinner, what are the implications of this for conceptions of masculinity? Do such conceptions of citizenship encourage not just a predominantly heterosexual masculinity but also forms of masculinity that emphasize financial self-reliance and paternalism? Do they encourage forms of masculine self-esteem that are dependent upon female subordination? In short, are various forms of masculinity being privileged over others here and what are the implications for men as well as women? Hearn's (1992) study analyzes constructions of masculinity within the hierarchies of patriarchy pointing out that some men oppress other men on the basis of class, race, sexuality, disability, and other categories. Work such as Yuval-Davis's (1997) has drawn attention to the ways in which nationalism is gendered in both masculine and femi-

nine terms. For example, masculinity can be mobilized for national defence while femininity can be mobilized for training children (and even resistant adults) in appropriate cultural values and national identities. However, increasingly, work on the politics of masculinity has also had to deal with the phenomenon of male organizations engaging in 'backlash' politics, as more traditional forms of masculinity are seen to be under threat from feminism (Messner 1997). There is certainly a need for more political science work on the wider normalizing political discourses that help to construct particular forms of both masculinity and femininity. However, it should be noted that mainstream political scientists sometimes have been even less keen to analyze the politics of masculinity than they have been to analyze so-called women's issues. After all, problematizing the self can be even more threatening than acknowledging the other.

6. *Analyzing the Politics of Feminism*

The analytical tools which feminist theory developed to critique gendered discourses drew on a variety of sources from sociology of gender to French theory's critique of those forms of dualistic thinking that subordinated the feminine 'other.' Increasingly, these analyses began to be turned upon the subject of feminism itself. While early feminist engagements with political science concentrated on documenting women's subordination, marginalization, and exclusion, the issues have now become far more contentious. It is not just the constitution of the 'political' which has been found to be problematic but also the very constitution of categories inherent to feminism such as 'woman.' Difference between women has been emphasized. Feminism itself has become the subject of deconstruction and critique as racial, ethnic, cultural, and class biases in particular versions of feminism have been highlighted. Sometimes, these critiques have a direct impact upon analyses of feminism and political science. For example, the authors of the study on women in Canadian political science (Arscott and Tremblay 1999) note the continued impact of Anglophone and Francophone divisions within the discipline and amongst feminists themselves—many Anglophone scholars rarely citing or teaching the work of their Francophone colleagues. Meanwhile, queer theory raises issues that earlier feminist analyses of traditional political theory tended to neglect such as the extent to which constructions of the citizen as male head of household were not only patriarchal but also constructed the citizen in heterosexual terms. The public/private conception, once so crucial to feminist critiques of traditional political theory, has been criticized for being a Western conception. Women's subordination is now as likely to be seen as being due to the continued influence of gendered binary linguistic categories as being due to more materialistic factors such as the gender division of labor.

Not surprisingly, these issues are also reflected in debates over feminist political practice. For example, there are major debates over whether feminist strategies should revolve primarily around cultural issues of 'recognition' or whether the battles should be over issues of economic 'redistribution.' There are also debates over whether feminist politics has revolved around 'states of injury' which government has then been called upon to redress, rather than upon more empowering political strategies (Brown 1995). Many 1970s' versions of feminism have been seen to result from forms of 'identity politics' which, theorists such as Butler argue, both have strategic importance and can reproduce forms of exclusion. Butler (1991) states that the constitution of the identity of the subject has wider implications since 'identity categories tend to be instruments of regulatory regimes, whether in the normalising categories of oppressive structures or as the rallying points for a liberatory contestation of that very oppression.' Consequently, political scientists such as Dean (1997) have argued for developing reflective political strategies that take issues of difference seriously while negotiating forms of feminist solidarity and political action. Dean's collection on *Feminism and the New Democracy* argues for the importance of increasing democracy not only at the level of the state but also in personal relationships.

7. *Feminism, Political Science, and Policy*

In short, contemporary feminists are still debating, and stretching, the limits of the political. In Vickers's phrase (1997), feminists need to 'reinvent' political science. However, political science also poses a number of challenges to feminist theory and practice. Key amongst these is the challenge posed by the policy-making process. For example, critiques of the universalizing category 'woman' can be useful in the policy-making process when the resulting sensitization to difference draws attention to the need to develop programs that address a range of women's problems, experiences, and lives. They can be less helpful if the differentiation goes so far that the desirability of feminists establishing women-focused programs is brought into question, or if feminists deny the insights that can be drawn across cultures from old feminist concepts such as gender stereotyping or the public/private divide (Lee 1995, Herrick and Sapieva 1997). Similarly, queer theory can be useful for drawing attention to the ways in which both minority and mainstream identities exclude those who are not seen to belong. The insights of queer theory can be less useful if the need for specifically lesbian advocacy or services is ignored. Arguing for a plurality of interpretations may well be a useful feminist political strategy in the fields of literature or film that prevents gendered meanings being fixed. Eventually, however, policy-makers will need to make definite decisions and implement specific programs. Debates over recog-

nition and redistribution usefully draw attention to the need for programs that address both aspects but are less useful if they lead to the privileging of one aspect over the other. Critiques of dual categories can be useful for drawing attention to the discursive subordination of the feminine but are less useful if dichotomous thinking is seen as the fundamental source of women's inequality that must be addressed. Here, as elsewhere, a combination of judicious eclecticism and sensitivity to difference and multiplicity can be helpful in negotiating the complex engagements between feminism and political science.

The relationship between feminism and political science has also been complicated by the times. The backlash against so-called 'political correctness' in many English-speaking countries has made it more acceptable to not merely dismiss feminist critiques of gender bias but to accuse feminists of silencing and marginalizing men. As indicated previously, the institutions and organizations of gender politics now include various aggrieved men's groups—groups mobilizing over issues of child custody, divorce settlements, and parental support—as well as women's organizations and gay and lesbian organizations. The impact of economic globalization and rapid social and technological change has contributed to an environment in which appeals to nostalgia, including gendered forms such as bringing back dependent stay-at-home Mums and reliable male breadwinners, can be electorally attractive to women as well as men. Meanwhile, the influence of neoliberalism, particularly in English-speaking countries, has undermined those feminist critiques of male dominance and female disadvantage that relied on state action to redress the balance. While there are forms of neoliberalism that attempt to incorporate some women into conceptions of the self-reliant individual, others flirt with forms of social conservatism which privilege more traditional gender relations.

All of this has made the engagement between feminism, gender studies, and political science more complex. It has also meant that analyses must be updated constantly. While appeals to nostalgia ensure that older feminist analyses are unlikely to be totally superseded, there is a need to develop analyses that can deal with the evolving political discourses and practices of the twenty-first century. Earlier feminist analyses concentrated on analyzing the politics of exclusion and marginalization. However, increasingly, feminists may need to develop sophisticated analyses of new forms of political inclusion and incorporation, (and new forms of masculinities and femininities), as changes in gender politics are played out in numerous sites and discourses. In the process, mainstream political science could usefully learn to be as self-critical and reflective as feminism has become.

See also: Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Feminist Economics; Feminist Epistemology; Feminist Legal Theory; Feminist Political Ecology; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Feminism, and Sexuality in Archaeological Studies; Klein, Melanie (1882–1960); Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns

Bibliography

Arneil B 1999 *Politics and Feminism*. Blackwell, Oxford, UK

Arscott J, Tremblay M 1999 Il reste encore des travaux à faire: Feminism and political science in Canada and Québec. *Canadian Journal of Political Science/Revue canadienne de science politique* **XXXII**: 125–51

Brown W 1995 *States of Injury: Power and Freedom in Late Modernity*. Princeton University Press, Princeton, NJ

Bryson V 1999 *Feminist Debates: Issues of Theory and Political Practice*. Macmillan, London

Butler J 1991 Imitation and gender insubordination. In: Fuss D (ed.) *Inside/Out: Lesbian Theories, Gay Theories*. Routledge, New York

Butler J, Scott J (eds.) 1992 *Feminists Theorize the Political*. Routledge, New York

Christy C 1994 Trends in sex differences in political participation: A comparative perspective. In: Githens G, Norris P, Lovenduski J (eds.) *Different Roles, Different Voices: Women and Politics in the United States and Europe*. HarperCollins, New York

Conway M, Steuernagel G, Ahern D 1997 *Women and Political Participation*. CQ Press, Washington, DC

Coole D 1988 *Women In Political Theory: From Ancient Misogyny to Contemporary Feminism*. Wheatsheaf, Brighton, UK

D'Amico F 1995 Women national leaders. In: D'Amico F, Beckman P (eds.) *Women in World Politics: An Introduction*. Bergin and Garvey, CT

Dean J (ed.) 1997 *Feminism and the New Democracy*. Sage, London

Eisenstein H 1996 *Inside Agitators: Australian Femocrats and the State*. Allen and Unwin, Sydney, Australia

Hearn J 1992 *Men in the Public Eye: The Construction and Deconstruction of Public Men and Public Patriarchies*. Routledge, London

Herrick R, Sapieva A 1997 Perceptions of women politicians in Kazakhstan. *Women and Politics* **18**(4): 27–40

Lee A-R 1995 Korean women's politicization: The effects of public/private constructions. *Women and Politics* **15**(2): 53–83

Lovenduski J, Hills J 1981 *The Politics of the Second Electorate: Women and Public Participation*. Routledge and Kegan Paul, London

Messner M 1997 *Politics of Masculinities: Men in Movement*. Sage, Thousand Oaks, CA

Nelson B, Chowdhury N (eds.) 1994 *Women and Politics Worldwide*. Yale University Press, New Haven, CT

Norris P 1994 The impact of the electoral system on election of women to national legislatures. In: Githens G, Norris P, Lovenduski J (eds.) *Different Roles, Different Voices: Women and Politics in the United States and Europe*. HarperCollins, New York

Okin S M 1979 *Women in Western Political Thought*. Princeton University Press, Princeton, NJ
Olson K 1999 *Adding Insult to Injury: Social Justice and the Politics of Recognition*. Verso, London
Pateman C 1982 Women and political studies: Presidential address to the 23rd annual meeting of the Australian Political Studies Association, Canberra, August 1981. *Politics* **May**: 1–6
Pateman C 1988 *The Sexual Contract*. Blackwell, Oxford, UK
Pateman C 1996 *Democratization and Citizenship in the 1990s: The Legacy of T H Marshall. Vilhelm Aubert Memorial Lecture 1996*. Institute for Social Research and Department of Sociology, University of Oslo, Oslo, Norway
Phelan S 1997 The shape of queer: Assimilation and articulation. *Women and Politics* **18**(2): 55–73
Phillips A 1991 *Engendering Democracy*. Polity Press, Cambridge, UK
Rae S (ed.) 2000 *International Perspectives on Gender and Democratisation*. Macmillan, Houndmills, UK
Sarkees M R, McGlen N E 1999 Misdirected backlash: The evolving nature of academia and the status of women in political science. *Political Science and Politics* **32**: 100–9
Shapiro R, Mahajan H 1986 Gender differences in policy preferment: A summary of trends from the 1960s to the 1980s. *Public Opinion Quarterly* **50**: 42–61
Squires J 1999 *Gender in Political Theory*. Polity, Cambridge, UK
Turshen M, Holcomb B 1993 *Women's Lives and Public Policy: The International Experience*. Praeger, Westport, CT
UNIFEM 2000 *Targets and Indicators: Selections from Progress of the World's Women*. UNIFEM, New York
Vickers J 1997 *Reinventing Political Science*. Fernwood, Halifax
Yuval-Davis N 1997 *Gender and Nation*. Sage, London

C. Johnson

Gender and Feminist Studies in Psychology

1. *The Emergence of a Field*

Feminist studies in psychology emerged in the late 1960s, in the context of the second wave of the women's movement (see *Feminist Movements*). A group of activists, mostly graduate students and newcomers to psychology, founded the Association for Women in Psychology in 1969. At about the same time, a group of somewhat more established psychologists worked within the American Psychological Association to create a division for the psychology of women. Divisions on ethnic minority psychology and lesbian/gay psychology were established later.

Division 35, inaugurated in 1973 and now known as the Society for the Psychology of Women, is one of the largest divisions of APA. AWP continues as an activist organization with a collective governance structure and no formal ties to APA, although Division 35 and AWP cooperate on various projects and there is considerable overlap in membership. Similar organizational development took place in Canada and the UK, where the British Psychological Society now houses a Psychology of Women Section and a Lesbian Psychology Section. For the first time in the century-long history of psychology, the structure of its major professional organizations acknowledges the presence and professional identity of diverse women. The study of women and gender, and women themselves, now have a political and organizational base within psychology.

With these developments came the founding of feminist journals devoted to research from psychology and related disciplines: *Sex Roles* (1975), *Psychology of Women Quarterly* (1977), and *Feminism and Psychology* (1991) are perhaps the most widely circulated. Their founding, along with the growth of interdisciplinary women's studies journals, provided new publication outlets for research on women and gender. In 1974, there was not a single journal article on women and achievement published, although achievement motivation was a lively research field; in 1993, there were 161. Similar growth has occurred for other topics, notably sexual violence and feminist therapy.

The growth of feminist psychology can also be seen in the number of psychology departments offering courses in the psychology of women or gender. Before 1968, there were virtually none. By 1990, 51 percent of departments that responded to an APA survey included such courses, and there were 172 offered at the graduate level. The first two textbooks appeared in 1971; today, there are many in print.

Along with facilitating the recognition of women's presence in psychology, the women's movement gave rise to a feminist critique of psychological theory, research, and practice and to the development of new approaches in all these areas. Organized activism by AWP and Division 35 led to a surge in feminist symposia and papers at APA national conferences starting around 1970. Critical books and journal articles objected to psychology's views of women. Naomi Weisstein, among the first second-wave critics, famously claimed that psychology had nothing to say about women because it knew nothing about women. Carolyn Sherif analyzed how the needs of the cold-war military had shaped psychology's research agenda. Phyllis Chesler argued that clinical psychology and psychiatry were instruments of control directed at women.

These, and the many other voices of second-wave feminist psychology, were not cautious or timid ones. The 1970s feminist critique quite openly challenged psychologists' choice of research topics and methods, their theoretical constructs, and their modes of diagnosis and therapeutic intervention, and it attacked the values underlying them. It had an assertive tone and an activist agenda that probably contributed to its exclusion from 'mainstream' journals. A similar confrontational approach had last been seen some 50 years

earlier, during the first wave of the women's movement, when some of the first American women to receive higher degrees in psychology questioned the biases of the psychology of their day, such as the ongoing search for sex differences in personality and ability. Hollingworth (1916) wrote a powerful critique of the social devices for impelling women to bear children, while Wooley (1910) characterized research on sex differences in mental abilities as 'riddled with flagrant personal bias, logic martyred in the cause of supporting a prejudice … and sentimental rot and drivel.'

A fundamental aspect of second-wave feminist critique was the claim that psychological knowledge was androcentric, or male-centered. Men were studied much more often than women, and research questions were formulated from gender-biased perspectives. For example, research on violence and aggression focused on male-male aggression ranging from animal territoriality to human warfare. Violence by men against women, particularly in the context of intimate relationships, was overlooked and under-studied.

Theories were constructed from androcentric perspectives, and women's behavior was judged against a (white, middle-class) male norm. Theories of lifespan psychological development, for example, were often derived entirely from research on middle-class men's life patterns. Women's behavior was more likely than men's to be explained in biologically determinist terms, with gender-differentiated social contexts ignored. For example, explanations of the greater incidence of depression in women focused on hormonal causes, not environmental ones like prior physical or sexual abuse. Psychological adjustment for women was articulated in terms of conformity to gender norms. However, because femininity (but not masculinity) was defined in opposition to agency, autonomy, and competence, these norms created many double binds for women. Feminist critics documented gender biases in theories of intelligence, attribution, social influence, learning, memory, thinking, identity, clinical diagnosis, and therapy, among others.

As a result of this critique, psychology began to develop new ways of thinking about women and gender, expand its research methods, and offer more gender-sensitive approaches to therapy and counseling. Contemporary feminist psychology, like most areas of contemporary psychology, is not a unified subject. Meanings are created in a community of practice: they change over time and as members of the community intersect with their other communities. Not all scholarship on the psychology of women is grounded in feminist perspectives, and no single feminist perspective unites all the work that is self-identified as feminist. Rather, an array of political stances is implicit in the scholarship in this area. Feminist psychology means different things to developmental, clinical, and social psychologists; to older second-wave activists and younger people entering psychology today; to researchers, educators, and practitioners. Most generally, feminist psychology is a form of critical psychology: it interrogates the moral, political, and scientific claims of psychology and attempts to influence the directions of the discipline. Its goal is the systematic analysis of women's psychology and of the impact of gender on human subjectivity and experience.

Subjective definitions are also important. Academic psychologists who self-identify as feminists define their identity in terms of a set of values—respecting women and their experiences, a concern with equality of power, the need for social activism, and the idea of gender as a social construct—and report that they experience feminism as a transformative personal experience.

2. *Development and Contributions*

One of the important early contributions of feminist psychology was the conceptual distinction between sex (conceived as biological differences between males and females) and gender (conceived as cultural elaborations based on biological differences). The distinction was important because it called into question the naive biological determinism and essentialism of much psychological theorizing. However, this approach later came to be seen as limiting because of its individualism. Although it distinguished social influences from biological ones, it still conceived of gender as a static set of attributes or personality traits, formed by socialization and relatively permanent within individuals. Moreover, the sex/gender distinction rested on the assumption that causation can be determined and is restricted to either nature or nurture, with little allowance for complex interactions of biological and social influences. And it foreclosed the possibility of examining the social construction of biological sex, only to be confronted with the transgender and intersex activist movements. In hindsight, the usefulness of distinguishing gender from sex was that it opened the way to conceiving of gender as a principle of social organization which structures the relations between women and men and governs access to resources. More recent formulations of gender stress its performative aspects; that is, gender is viewed as a set of agreed-upon transactions that are enacted in mundane social relationships.

Another early focus of research was replacing women in the history of psychology. The devaluation, stereotyping and invisibility of women in history had been played out in psychology; indeed, it was the recognition of the 'womanless' state of the discipline that energized the feminist critique of psychological research. Not only were women excluded from the profession, women and their characteristic experiences and activities were thought too unimportant to be a focus of inquiry. Although it originated as attempts to

restore the 'neglected foremothers' of psychology, this line of work catalyzed exploration of the structural and institutional mechanisms of exclusion. An underrepresentation of women in any scholarly discipline points to the importance of gender as a social and political category, prompting questions about how knowledge is sought, validated, and disseminated.

As more research centered on women and gender, new areas of inquiry were developed. Perhaps the most important contribution of feminist and gender studies in psychology is the wealth of new research generated. Uniquely female experiences such as menarche and childbirth have become legitimate topics of study and new feminist perspectives have been applied to them. A feminist critique of autonomy as the endpoint of psychological development grew into theories of the self-in-relation (see *Feminist Theory: Ecofeminist and Cultural Feminist*), and a feminist critique of masculinity/femininity grew into the construct of psychological androgyny (see *Androgyny*). Work on gender socialization, stereotyping, stigma, and objectification proliferated. Violence against girls and women was named and documented. Its causes, incidence, and effects were analyzed (see *Domestic Violence: Sociological Perspectives; Rape and Sexual Coercion; Sexual Harassment: Social and Psychological Issues*). The pervasiveness of gender as a cognitive category, and gender polarization as a cognitive mechanism, were explored. Developmental issues of childhood, adolescence, sexuality, motherhood, midlife, and aging were examined from women-centered perspectives (see *Motherhood: Social and Cultural Aspects; Sexuality and Gender*). New research examined the interactions of multiple roles (spouse, parent, and worker) for both sexes. Sexual identity and orientation were no longer taken for granted, but became active areas for developmental theory-building and empirical inquiry. Women's health issues, long neglected in behavioral and medical research, also became an active site of study. In clinical psychology, a new importance was attached to issues of power in social and therapeutic contexts: the power of diagnosis and labeling, power imbalances in close relationships (including the therapeutic relationship), and the effects of women's social subordination on psychological adjustment. Psychological problems that differed in incidence or characteristic manifestation according to gender (eating disorders, depression, substance abuse, and agoraphobia, among others) received new attention from feminist perspectives.

In interdisciplinary feminist theory there are many variants, including radical, socialist, psychoanalytic, postmodern, and womanist (woman of color). However, in psychology, liberal and cultural feminist approaches have predominated (see *Feminist Theory*). Both these theoretical frameworks are individualistic, leading to conceiving gender largely as a constellation of individual traits. This view of gender as rather static and internalized has resulted in prolonged debates about the extent, importance, and mutability of sex differences. Liberal feminists have tended to minimize sex differences, arguing that they are small, relatively inconsequential, and easily eliminated with corrective socialization and education. This viewpoint has predominated in debates about cognitive abilities such as mathematics performance. Cultural feminists, on the other hand, have maximized differences, arguing for essential and fundamental divergences between women and men in such areas as identity, relatedness, and moral reasoning.

As the field has matured, there has been a turn away from these debates, which seem intractable, and an increasing interest in gender processes rather than gendered traits. In other words, gender currently is conceived not only as a phenomenon inside individuals but a system of agreed-upon meanings used in social interaction. An increased interest in how gender inequity is recreated and maintained in mundane social interaction is one result, with emphasis on such topics as cognitive bases of stereotype maintenance, gender-biased attributions of others' behavior, tokenism, self-fulfilling prophecies, sex discrimination, sexual harassment, hostile and benevolent sexism, and how the illusion of equity is sustained despite inequalities of status and power (see *Small-group Interaction and Gender*). This reconceptualization has brought social constructionist theory to the forefront and underlies a turn to language and discursive processes, particularly in the UK, where discursive psychology is more developed than in the US.

Feminist psychology has brought about a notable decrease in androcentrism in psychological research, education, and practice. Tangible changes include guidelines for nonsexist research and language in APA publications, the development of courses and curriculum integration resources, and the availability of nonsexist and feminist clinical training. In the US, women now account for 60 percent of Ph.Ds awarded annually in psychology.

There is also an increased presence of women, including lesbians and women of color, in APA's structure and on its policy-making boards and committees. In over 100 years as a professional organization, APA has elected only eight women (all white) to its presidency. The impact of second-wave feminism is evident: six of the eight were elected after 1970. In the UK, the first (woman-organized) psychology of women conference took place in 1987. A decade later, feminist psychology was a theme for the annual conference of the British Psychological Society.

3. *Methodological and Theoretical Developments*

The characteristic research methods of a discipline reflect its values and the assumptions it makes about its objects of study. Both first- and second-wave

feminists criticized psychology for sexist biases in research, pointing out problems with studying women less often than men, designing research materials from masculinist perspectives, using restricted samples, failing to compare women and men under similar conditions, and exaggerating or overgeneralizing about sex differences. With the second wave, however, came a more fundamental criticism, not of flaws and biases in the use of (otherwise satisfactory) methods, but of the value of the discipline's most characteristic and central methods themselves. Psychology, it was charged, over-relied on experimental methods, which strip behavior from its social context and position the experimenter as an expert manipulator of the participant and situation. Second-wave feminists have also criticized what they see as the naive belief that psychology can discover universal laws of behavior.

There is today considerable variability in the methods used to study women and gender within psychology. The majority of articles published in the two leading US feminist journals, *Psychology of Women Quarterly* and *Sex Roles*, as well as those published in general psychological journals, use quantitative methods such as experimental and correlational designs and meta-analysis. Only the UK journal *Feminism & Psychology* routinely publishes more qualitative and discursive work. However, even outside of feminist studies, psychology has a long tradition of inquiry that transcends laboratory manipulation of isolated variables: field research, observational techniques, content analysis, participant observation, and case studies are a few examples. A significant minority of feminist researchers draws on these traditions as well as newer, postmodern influenced forms of analysis. The feminist call for methodological plurality has had some effects. Special issues of *Psychology of Women Quarterly* have focused on method and theory (1989), transforming psychology (1994), and innovative methods for feminist research (1999).

In contrast to the positivist epistemology that permeated prefeminist research, feminist methodology is characterized by at least three innovations. First, feminist methods are reflexive: they involve recognition of the social identity and involvement of the researcher. Personal reflexivity involves disciplined self-reflection on how one's identity influences one's work and, in turn, how one's work influences the self. Moreover, a broader kind of reflexivity is also apparent—the maintenance of a critical perspective on the discipline and how its dominant paradigms are sustained by powerful institutions.

Second, feminist methods are meant to serve—not drive—the inquiry. There is little methodological elitism, and a great deal of creativity, within feminist psychology. The use of multiple methods within a single research project, and adaptations of methods more characteristic of other disciplines, are emerging practices. These innovations go beyond psychology's traditional concept of convergent validity, in which different methods of measurement are applied to the same problem in order to strengthen the evidence for a hypothesis by accumulating different kinds of data. Instead, they are ways of creating dialogue between different approaches; they recognize that a method is a theory, and that different methods may lead to different—and incompatible—truths.

Finally, feminist research has a social change orientation. Within feminist psychology, advocacy and scholarship are not viewed as incompatible. Innovative work erodes the polarizations such as basic/applied and theoretical/practical that are maintained elsewhere in psychology. While the original goal of second-wave feminist critics was to eliminate gender bias and make psychology more objective, that feminist empiricist project has evolved into a more sophisticated one. Feminist psychology is no longer merely corrective or oppositional; it is transformative. Science is conceived as a social practice, and the tenets of empiricism, such as objectivity and validity, are subjects of feminist analysis. Alternative modes of research are being developed and recognized, and there is a great deal of practical research oriented toward understanding and resolving social problems.

4. *Current Constraints and Future Directions*

Despite the growth and gains of feminist psychology, the field has not developed adequately in some key areas. Theoretical initiatives have lagged behind empirical studies. The extended debate between cultural and liberal feminists over the extent and importance of fundamental sex/gender differences delayed recognition of the importance of studying gender as a set of interactional processes and a system of social meanings.

Moreover, despite the wealth of empirical research generated in the last three decades of the twentieth century, the knowledge base of feminist psychology is insufficiently integrated into psychology as a whole. Citation and content analyses indicate that the textbooks, standard courses, and major theories of psychology have been little changed by the availability of its new perspectives. Feminist work is still too easily marginalized and treated as optional.

Feminist psychology has not yet encompassed adequately the diversity of women, despite ongoing attention to this issue. Its research base and its practitioners are still largely white, middle-class, heterosexually identified, and British or North American. And its success in persuading the discipline as a whole to recognize that psychology's claims about human nature and human behavior are culturally and historically situated has been limited.

The relatively successful institutionalization of the psychology of women and gender (and of women) within psychology is not an unmitigated gain. At one extreme, some feminist critics have expressed the fear

that there can be no entity such as 'feminist psychology' without either distorting and weakening feminist commitments beyond recognition, or transforming psychology entirely. They claim that the former, and not the latter, is taking place: that the research that comes to be assimilated is a weak, individualistic 'psychology of women' and decontextualized 'sex differences,' and that there has been a loss of political activism within feminist psychology over time. However, others point to the large variety of topics, methods, and epistemological starting points represented in contemporary feminist psychology. The multiplicity of approaches allows for healthy dialogue and cross-fertilization, and the reflexivity that increasingly characterizes the field assures that the politics underlying the methods, questions, and practice of feminist scholarship will remain open to analysis.

See also: Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Feminist Economics; Feminist Epistemology; Feminist Legal Theory; Feminist Political Ecology; Feminist Psychotherapy; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Feminism, and Sexuality in Archaeological Studies; Klein, Melanie (1882–1960); Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns

Bibliography

Bem S 1993 *The Lenses of Gender*. Yale University Press, New Haven, CT

Burman E (ed.) 1998 *Deconstructing Feminist Psychology*. Sage, London

Caplan P J, Crawford M, Hyde J S, Richardson J T E 1997 *Gender Differences in Human Cognition*. Oxford University Press, Oxford, UK

Chesler P 1972 *Women and Madness*. Doubleday, New York

Crawford M 1995 *Talking Difference: On Gender and Language*. Sage, London

Crawford M, Marecek J 1989 Psychology reconstructs the female: 1968–88. *Psychology of Women Quarterly* **13**: 147–65

Crawford M, Unger R 2000 *Women and Gender: A Feminist Psychology*, 3rd edn. McGraw-Hill, New York

Fine M, Gordon L 1989 Feminist transformations of/despite psychology. In: Crawford M, Gentry M (eds.) *Gender and Thought*. Springer-Verlag, New York, pp. 146–74

Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA

Hare-Mustin R T, Marecek J (eds.) 1990 *Making a Difference: Psychology and the Construction of Gender*. Yale University Press, New Haven, CT

Hollingworth L S 1916 Social devices for impelling women to bear and rear children. *American Journal of Sociology* **22**: 19–29

Kimmel E B Crawford M in press *Innovations in Feminist Psychological Research*. Cambridge University Press, Cambridge, UK

Landrine H (ed.) 1995 *Bringing Cultural Diversity to Feminist Psychology*. American Psychological Association, Washington, DC

Morawski J 1994 *Practicing Feminisms, Reconstructing Psychology: Notes on a Liminal Science*. University of Michigan Press, Ann Arbor, MI

Parlee M 1997 Feminism and psychology. In: Gergen M M, Davis S N (eds.) *Toward a New Psychology of Gender*. Routledge, New York, pp. 65–95

Russo N F, Dumont B A 1997 A history of Division 35 (psychology of women): Origins, issues, activities, future. In: Dewsbury D A (ed.) *Unification Through Division: Histories of the Divisions of the American Psychological Association*, 2nd edn. American Psychological Association, Washington, DC

Sherif C 1979 Bias in psychology. In: Sherman J A, Beck E T (eds.) *The Prism of Sex: Essays in the Sociology of Knowledge*. University of Wisconsin Press, Madison, WI, pp. 93–133

Tiefer L 1991 A brief history of the Association for Women in Psychology: 1969–91. *Psychology of Women Quarterly* **15**: 635–49

Weisstein N 1968 *Kinder, Kirche, Kuche as Scientific Law: Psychology Constructs the Female*. New England Free Press, Boston

Wilkinson S 1988 The role of reflexivity in feminist psychology. *Women's Studies International Forum* **11**: 493–502

Wilkinson S 1997 Feminist psychology. In: Fox D, Prilleltensky I (eds.) *Critical Psychology: An Introduction*. Sage, London, pp. 247–64

Wilkinson S, Kitzinger C (eds.) 1995 *Feminism and Discourse: Psychological Perspectives*. Sage, London

Wooley H T 1910 Psychological literature: A review of the recent literature on the psychology of sex. *Psychological Bulletin* **7**: 335–42

Worell J 1996 Feminist identity in a gendered world. In: Chrisler J C, Golden C, Rozee P D (eds.) *Lectures on the Psychology of Women*. McGraw-Hill, New York

Worell J, Johnson N G 1997 *Shaping the Future of Feminist Psychology: Education, Research, and Practice*. American Psychological Association, Washington, DC

M. Crawford

Gender and Feminist Studies in Sociology

When feminists first started using the term gender in the early 1970s, the point was to make a clear distinction between how we are born (sex), from how our society shapes who we become (gender). In the process of studying how we become gendered people, feminist sociologists have changed our sense of both what gender is and of how we can best study social phenomena.

Like all sciences dominated by positivist epistemology, sociology traditionally has assumed that, appropriately conducted, it was objective, that is, had no point of view (see *Inferences, Psychology of*). As feminists began to struggle to make social science

useful for understanding the lives of women, we also began to uncover systematic biases in the tradition we inherited, for example, in the kinds of questions asked, the metaphors underlying governing theories, dominant conceptual frameworks, and the strategies used in gathering and analyzing data. These biases tended to follow a pattern. Sociology (like other sciences) was organizing knowledge from the perspective of the experiences and worldview of racially and economically privileged men (Sprague 1997).

As feminist sociologists have struggled to understand gender from the standpoint of women, we have found that it operates on at least three planes of human experience: as a personality structure, as a set of interactional practices, and as an aspect of social organization. Further, we have moved from seeing gender as something about all women and all men to discovering what gender is, how it works, and, therefore, how we might work to subvert it, varies in different intersections with race and class (see *Race and Gender Intersections*; *Social Class and Gender*). In sum, we have learned that what might appear to many people as a naturally occurring dichotomy of two opposite sexes is the complex outcome of a great deal of social effort and organization.

1. Gender as Something We Are

Most people who use the term gender probably think of it as an individual attribute, a personality structure that we are socialized into fairly early in life. There is a lot of evidence that we respond to children in gender-specific ways. Parents and other significant adults recognize different qualities in girls than in boys, for example, we tend to see girls as cute and loveable and boys as energetic and mischievous. Since we can only support the development of skills and tendencies that we can perceive, seeing girls as different from boys becomes a self-fulfilling prophecy. In attempting to please children, we give them the toys and experiences that our culture has led us to believe they will like and, eventually, that gender-specific advertising directed at children teaches them they should like. In the process, boys have been more likely to learn to like rough-housing and manipulating objects and girls have been more likely to learn to like taking care of others and dressing up and wearing makeup.

Sociologists operating in the tradition of feminist psychoanalysis (Chodorow 1989) say that the development of gendered personalities in contemporary Western society runs much deeper than developing skills and tastes that gender organizes our very sense of self and how we relate to others. Gendered selves are a consequence of gender beliefs in two ways. First, we assign primary responsibility for care-taking of the young to women and not to men. Second, we have learned to use gender as a key indicator of how we should respond and relate to one another. As a result, children form their first attachments and develop their basic sense of self and their place in the larger community in very different kinds of relationships depending on their gender.

Boys develop their first attachment in a relationship with someone who has learned to understand herself and her charge as opposite sexes. The sexual opposition propels boys into seeing themselves as separate beings earlier in life and masculinity becomes an important part of how they understand their separateness and identity. As development leads to needs for increasing autonomy, the relative absence of males from primary care means that the people boys most need to separate from are women. Thus, an important means by which boys assert their independence involves rejecting the feminine within themselves.

On the other hand, mothers are not prohibited culturally from seeing girls as extensions of themselves, making the mother–daughter attachment much stronger and keeping it in force much longer. When girls do begin to differentiate themselves from their mothers, it is not as complete as it is for boys. Girls' struggle to develop a sense of an autonomous self is not based on gender identity and does not require a complete break from psychological connection with their mothers.

These gendered psychodynamics can be exacerbated in the learning of social roles. Children spend most of their time in domestic and care-taking sites. If girls learn that their most important roles involve taking care of others, they are in places and sets of relationships in which those roles can be practiced. If boys learn that their important roles occur outside their immediate domestic environment, they must acquire a sense of their adult roles by denying their connection to their immediate context and identifying abstractly with the public roles of their fathers and other men.

Their early experiences can lead to gender differences in how people think about themselves and their orientation to the world. Men tend to develop a highly individuated sense of self, an abstract orientation to the world, and contempt for women in general. Women, in contrast, are more likely to develop a connected sense of self embedded in concrete relationships, a more concrete orientation to the world, and hostility toward their mothers in particular.

The psychoanalytic account of gendered personality structure has given feminist sociologists a good deal of insight on how gender works. It explains why gender has such emotional power; why it generates so much anger, hostility, and fear, both interpersonally and in public debate. Men's psychic need to maintain an individuated sense of self defined by a fairly abstract notion of masculinity suggests why emotional intimacy can be threatening, and why men might be more worried about appropriately displaying (proving) their masculinity. If women have a relational sense of self, founded on a relationship with a female caretaker, we can expect they will seek to form strong emotional

bonds, particularly with other women. And, if our sense of who we are depends on distancing ourselves from a powerful mother, that might explain why women in powerful positions make many of us uncomfortable.

The kind of masculine self-psychoanalysis describes is also consistent with the pattern of conceptual organization that feminists have found runs through the organization of knowledge in Western culture, from religion to science (Harding 1998). Common patterns include organizing phenomena into isolated units, arranging them in hierarchies, particularly dichotomies, privileging abstract universals over contextualized understandings, and adopting explanations that best fit the experience of the privileged.

A major shortcoming of the approach that sees gender as something we are, is that it tends to reproduce these same patterns. It tends to represent masculine and feminine as abstract universals, and people as individual bearers of gendered selves, and it generalizes from the experiences of the privileged. As a result, the picture it gives us of gender is distorted.

While all children in the West and many children elsewhere are inundated with mass mediated messages about gender that reflect the beliefs and values of affluent whites, there is reason to suspect that families and communities in other classes and ethnic communities may communicate competing notions about gender to their children. For example, poor and working-class African–American communities in the US are more likely to teach girls to be strong and self-reliant rather than the dominant view of femininity as passive and dependent. Thus, children growing up in culturally marginalized communities are very likely getting mixed messages about their gender's roles and tastes.

Similarly, the psychoanalytic approach assumes a nuclear male breadwinner/female care-taker family structure, one in which few children in the world and increasingly few in West live. This bias is partially the result of the kind of data psychoanalysis tends to rely on: clinical interviews that select for a relatively affluent group, people who have the time and financial resources to undertake psychoanalytic treatment. While it is still the case that men are rare among those who do the work of caring for the young and it seems likely that female mother-surrogates will also relate to girls as like them and boys as the opposite sex, we need to ask how a variety of gendered care-taking relationships shape the development of personality structure.

Finally, seeing gender as an individual attribute makes it seem natural, almost voluntary, downplaying the degree to which gender differences in beliefs and behavior can be outcomes of differences in power and access to resources. Standpoint epistemology prompted some feminist sociologists to ask how people understand themselves and their lives and that drew attention to the impact of situations on behavior. It is well known that people act differently than they feel in some situations, and act differently from one situation to the next.

Attempting to understand what gender means from the standpoint of actual people also made it apparent that the kinds of situations people are placed in—the constraints they face and the resources they have—differs markedly depending on what class and race/ethnicity they are. Conceiving of people as conscious actors creatively and strategically responding to variable situational demands led these scholars to uncover the situational and enacted character of gender.

2. *Gender as Something We Do*

From a social constructionist perspective, gender is not an internalized personality structure as much as it is a social fiction we work to maintain through our everyday social practices. We use gender as a lens through which to perceive and evaluate one another. We do not see one another as generic individuals but as men and women, and we have gender-specific expectations for what people are and should be like. We hold one another accountable for acting in ways our culture says are appropriate for our gender (West and Zimmerman 1987).

Knowing that others have gendered expectations of us, we take that into account in deciding how we will act. We work to construct an appearance that will make us classifiable into one of two opposed sex categories: we make choices about what we can and cannot wear, how we cut our hair, and if and where we shave. We take actions to alter the shape of our bodies: we diet, girdle, pad, exercise, and even undergo surgery. We monitor how we conduct ourselves. We understand the appropriate range of gestures and postures for our gender: more expansive and resolute for men, more constrained and passive for women. Traditionally the burden on women has been much higher than for men: the range of attractive faces and bodies, the expectations for clothes and comportment, have been much narrower, requiring more careful self-monitoring and adjustment (Kessler and McKenna 1978).

Being held accountable for performing in gender-appropriate ways in a value system that privileges a certain kind of masculinity puts women in a double-bind. At the same time women are expected to do more work on their bodies, display, and comportment, they are devalued for spending excessive time in trivialities. At the same time women are expected to do the work of facilitating conversation by asking questions, seeking to find topics that men will find interesting, they are derided for talking too much.

The nature of gendered expectations and the resources for meeting them are shaped by class and race/ethnic relations. For example, in the face of centuries of racist barriers to Black men's employment in the US, Black women often have been breadwinners

and have been criticized as matriarchs; economically discriminated against and forced into domestic work, they have been trivialized as 'mammies' (Collins 2000). Asian men immigrating to the US in the nineteenth and early twentieth centuries were 'demasculinized,' denied the opportunity to have wives by restrictions on the immigration of Asian women and laws barring interracial marriage, and forced into traditionally feminine work in domestic service (Espiritu 1997). Feeding the family, a central expectation for women in many cultures, is a constant struggle financially and time-wise for poor women, but it is constructed as trivial work assigned to lower status others by affluent women (DeVault 1991). Motherhood occupies a position of virtue and protection among the economically and racially privileged and is defined as a social problem among poor women, especially members of racial minority groups, who lack access to good paying jobs or husbands with steady employment, and often must rely on state support to raise children. Gendered interactional practices are embedded in every social institution. In schools, for example, responding to children in gender-specific ways influences how, and how much we teach them (Sadker and Sadker 1994). We are more likely to help girls by giving them the answer and to help boys by giving them clues that support their discovering the answer independently. Because girls have been trained to be more cooperative and complacent while boys are expected to be more rowdy and outspoken, engaging boys' interest becomes a higher priority for teachers. The need to keep the boys interested facilitates the development of male-centered curricula and class experiences. Boys learn they are doers, should have high expectations for performance, and develop independent problem-solving skills; girls learn they are observers rather than actors, and have lower and narrower expectations of their own ability.

Classrooms are not the only place where gender is enacted in school. Thorne (1993) has analyzed the ways that elementary school children impose gender-specific expectations on one another on school playgrounds. By fourth or fifth grade, children are imposing gender segregation. For example, boys taunt boys who like to play with girls or who do not act 'manly' enough. Boys play in large groups, evading close adult supervision while developing physical and team-building skills. Girls form shifting pairs of 'best friends' and organize themselves into emotionally charged status hierarchies based on who is friends with whom. They are verbally intimate and much of the content of conversation is about romance and boys, about 'liking' and 'going with,' that is, who can win what boy's approval.

By middle school and high school, the expectation for feminine behavior makes high academic performance a threat to displaying appropriate femininity. Many girls who previously earned self worth through school performance learn that they must switch attention to winning the recognition of an acceptable male through being sexually appealing. They must walk a narrow line, however, if they are too sexual, they are derogated as 'sluts.' Boys must demonstrate their masculinity through their ability to use force skillfully, through sports and/or technology, and through their conquest of girls. Boys who do not display some sexual power over girls have their own sexuality called into question. Again, performance resources and standards vary by class and race. Economically privileged boys may be academic achievers without undercutting their masculinity; the poorest boys may have to resort to fights to enact their masculinity. African–American girls' self-esteem does not drop as far at that of white and Latina girls (see *Education (Primary and Secondary Schools) and Gender*).

Gendered interactions permeate adult lives. We make many of our decisions about allocating resources and responsibilities in social groups and these groups use gender-specific expectations for determining who will exert influence on decision-making (Ridgeway 1993). We are much more willing to be influenced by men than women: in the absence of information, we will assume that men have competence but that women do not. Men's ability to exert influence seems independent of their interactional style or their apparent motivation. Whether and on what terms we are willing to be influenced by women depends on what gender we are. Men are more willing to be influenced by women who speak tentatively and express priorities on group interests. Thus, women interacting with men are forced to choose between having influence and being seen as competent. All-woman groups, on the other hand, are more willing to be influenced by women who speak assertively and appear to be self-interested, but women like group-oriented women more so women in female groups must choose between having influence and being liked (see *Small-group Interaction and Gender*).

Evaluation is particularly central to workplace experiences—evaluations determine hires, pay, promotions, and the distribution of opportunity—and evaluation is a thoroughly gendered process. Martin (1996) finds that managers tend to see men as having more ability and potential, thus important to groom and mentor. This assessment is shared by male workers who are more likely than women to see themselves as entitled to and thus seek out opportunities and special treatment. In evaluating performance, managers emphasize men's successes and downplay their failures while they tend to do the opposite for women. We see men as entitled to authority but find authority in women problematic, especially when women have authority over men. We are much more comfortable with openly criticizing female colleagues than male ones.

One concrete and pervasive consequence of our holding people accountable for gender-appropriate behavior is the gender gap in pay (see *Sex Differences*

in Pay). Standard job evaluation ratings systematically discount the skills and responsibilities involved in jobs primarily held by women. If those jobs involve taking care of others, they are paid even less, presumably because it is assumed that meeting human needs is somehow a natural extension of an inherent ability to nurture among women.

Knowing others hold us accountable for gender-appropriate behavior and holding ourselves accountable leads us to act in ways that may not seem to be in our individual interests. For example, moving up a career ladder often requires relocating for a promotion or a better job, but women are less willing than men to disrupt their families by changing locations for personal advancement. Also, even women who work outside the home and bring in a substantial portion of the household income will do well more than their share of household labor and feel that the arrangement is fair. Men who might want to be more involved in domestic life are reluctant to be evaluated as less serious about their work. Seeing gender as something we do exposes its malleability. If our acting in gendered ways is at least partly the outcome of gendered expectations and evaluations, we can expose those practices and make choices about their appropriateness. For example, feminist scholarship uncovering biases in job and performance ratings has facilitated the feminist political campaign for comparable worth to develop a more equitable system (England 1992). There are constraints on what we can do to change gender. A standpoint approach to methodology and situational understanding of gender as something we do have helped feminist sociologists to see that situations are not randomly and spontaneously constructed. There are some larger regularities that shape systematically the kinds of situations we find ourselves in. Over the 1980s and 1990s, feminist sociologists increasingly have been talking about gender as a social structure or institution.

3. *Gender as a Form of Social Organization*

The idea that gender is a social structure or form of social organization has not yet been developed fully. The basic picture seems to be this: gender is the outcome when a society divides labor by sex, assigns the work of meeting human needs to women, and devalues that work. While dividing labor does not lead necessarily to differences in power, this has been the pattern in what are usually thought of as the 'great civilizations,' those social formations that have dominated the world. As we divide labor, we have created a hierarchy of power and authority, largely by separating authority and power from the work of taking care of people, directly meeting their physical and emotional needs.

Historically in the West and in many other cultures, men have been assigned to different social positions than have women. During the period of industrialization, this sexual division of labor was institutionalized in a division between public and private spheres. Work directed at generating profit and rewarded with wages and other important benefits became defined as a male activity. The work of feeding, clothing, nursing, and otherwise directly meeting people's daily needs for subsistence and emotional support was defined as private, unpaid domestic work and assigned to women.

The public/private distinction has always been more fantasy than fact. The work of taking care of people is not purely a private matter: it occurs in every domain of life (see *Care and Gender*). There are whole professions built around the work of care-taking: education, medicine, and social work. Many debates about social service spending, for example, spending for childcare, healthcare, elder care—are really debates about the degree to which we should take public responsibility for paying for care-taking or allocate it to unpaid domestic labor. In most jobs, some workers make coffee, remember birthdays, listen to troubles, bring treats, and organize social events. This, too, is care-taking work and it, too, is assigned usally to women and not recognized as paid work.

The notion that women had been confined to the private sphere is also inaccurate. Even in the nineteenth century, many women, particularly poor and immigrant women, worked for wages (see *Household Production*). Also, women have done the labor of creating and sustaining communities through volunteer work for churches, charities, and social clubs. The sexual division of paid labor has been resilient over time (Reskin 1988). Even though more and more women have been entering paid labor over the last half of the twentieth century so that now a great majority of women work for wages, we still have primarily men's jobs and primarily women's jobs—there are few jobs where the proportion of women workers is even close to the proportion of men. When women have entered professions previously dominated by men, those professions frequently restratify themselves into gender hierarchies (see *Sex Segregation at Work*).

What seems to be a constant in the sexual division of labor is that directly meeting people's needs, especially basic bodily needs, is devalued and assigned to women, and, in some cases, to low-status men (Glenn 1992). The work of care-taking is unpaid or underpaid, unrecognized, or assumed to be instinctive. One indicator of privilege in a gendered form of social organization is the degree to which one can distance oneself from the 'dirty work' of care-taking. All women are held responsible for domestic work, but not all women end up doing it. At the turn of the twenty-first century, most women work for wages and still do the 'second shift' of domestic work (see *Family and Gender*). Those who are in more privileged economic situations hire poorer women to do some or all of the domestic work, either in their homes or

through the modern, commodified domestic labor of the burgeoning service sector with its restaurants, cleaning services, and services that care for and entertain children. Thus, those who are most removed from care-taking work are economically well-off men, typically white men. Those who are the most likely to be involved in it are women, particularly poor women, disproportionately women of color. Increasingly, men and women in higher status jobs are opting out of parenting altogether.

The sexual division of labor is increasingly organized internationally (Mies 1986; see *Economic Globalization and Gender*). Women in countries of the South produce much of the food, clothing, and other commodities that women in the North put to work in their domestic labor. Male-dominated multinational corporations collaborate with male-dominated governments to maintain local policies that restrict the rights of laborers. The extremely low wages and bad working conditions that result from this collaboration enable the generation of high profits. The super-exploitation of third world workers, especially young women, undercuts the situation of working-class men and women in the West who see good jobs and secure futures disappear as more and more factories move offshore.

The sexual division of labor helps support men's political and economic power: their freedom from responsibility for domestic work has allowed them the time and energy for participating in the organizations and activities that underwrite their continued domination (Connell 1987). Men control industry, central state planning, and the means of institutionalized violence, the military and paramilitary institutions. Men dominate the institutions that produce official culture: science, the academy, the mass media. Even though women do the great majority of the work in the world, wealth is concentrated in male hands.

Their command over institutional resources gives some men a base from which to form alliances in solidarity with one another and against the interests of women (see *Male Dominance*). They have used this power to limit the public resources that go to the work of raising children and caring for the frail (see *Social Welfare Policies and Gender*). They allow wages in the workforce to be determined by 'the market,' thus reproducing the existing pattern of valuing men's work over women's. They have restricted women's ability to control their own fertility via birth control and abortion (see *Reproductive Rights in Developing Nations*; *Reproductive Rights in Affluent Nations*). They have raised obstacles against women's solidarity through sanctions against lesbianism (see *Lesbians: Social and Economic Situation*). This combination of practices pushes women into dependence, either on individual men or on the collective male-dominated state (see *Poverty and Gender in Affluent Nations*; *Poverty and Gender in Developing Nations*).

Gender as a form of social organization has until recently involved two overlapping dichotomies. The social division between those who control and those who do the work of meeting human needs has until recently overlapped with a sexual dichotomizing of men and women. Perhaps what we are seeing in the West is the development of a reorganized gender order, one in which men and women are being further subdivided based on whether or not they have care-taking responsibilities.

4. Conclusion

Feminist sociologists have found that gender organizes our sense of self, is a set of expectations we encounter and must deal with in our daily interactions, and is a form of social organization shaping opportunities and constraints. These are all aspects of the same process. People's selves are the products of relationships and experiences shaped by where they are in a social order organized by gender, as well as race and class. That social order is the crystalization of myriads of daily interactions in which people evaluate and hold one another accountable for behaving in gendered ways. The nature of the expectations individuals confront and their ability to negotiate the outcomes of these interactions depend on the kinds of personal strengths, cultural advantages or disadvantages, and material resources they bring to each. Understanding how this multi-faceted process works, and how we might work to subvert it, requires looking at it from diverse perspectives, including variations in sex, race, class, and nation. The more we do, the more it seems that gender organizes us into many more than two categories.

See also: Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Feminist Economics; Feminist Epistemology; Feminist Legal Theory; Feminist Political Ecology; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Place; Gender, Feminism, and Sexuality in Archaeological Studies; Klein, Melanie (1882–1960); Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns; Population Dynamics: Two-sex Demographic Models

Bibliography

Chodorow N J 1989 *Feminism and Psychoanalytic Theory*. Yale University Press, New Haven, CT

Collins P H 2000 *Black Feminist Thought* 2nd edn. Unwin Hyman, Boston

Connell R W 1987 *Gender and Power: Society, the Person and Sexual Politics*. Polity (in association with B. Blackwell), Cambridge, UK

DeVault M L 1991 *Feeding the Family: The Social Organization of Caring as Gendered Work*. University of Chicago Press, Chicago
England P 1992 *Comparable Worth: Theories and Evidence*. Aldine de Gruyter, New York
Espiritu Y L 1997 *Asian American Women and Men: Labor, Laws, and Love*. Sage, Thousand Oaks, CA
Glenn E N 1992 From servitude to service work: Historical continuities in the racial division of paid reproductive labor. *Signs* **18**(1): 1–43
Harding S 1998 *Is Science Multicultural? Postcolonialisms, Feminisms, and Epistemologies*. Indiana University Press, Bloomington, IN
Kessler S J, McKenna W 1978 *Gender: An Ethnomethodological Approach*. Wiley, New York
Martin P Y 1996 Gendering and evaluating dynamics: Men, masculinities, and managements. In: Collinson D L, Hearn J (eds.) *Men as Managers, Managers as Men*. Sage, Thousand Oaks, CA
Mies M 1986 *Patriarchy and Accumulation on a World Scale: Women in the International Division of Labour*. Zed Books, London
Reskin B F 1988 Bringing the men back in: Sex differentiation and the devaluation of women's work. *Gender & Society* **2**: 58–81
Ridgeway C L 1993 Gender, status, and the social psychology of expectations. In: England P (ed.) *Theory on Gender/Feminism on Theory*. A. de Gruyter, New York
Sadker M, Sadker D 1994 *Failing at Fairness: How America's Schools Cheat Girls*. Scribner, New York
Sprague J 1997 Holy men and big guns: The can(n)on in social theory. *Gender & Society* **11**(1): 88–107
Thorne B 1993 *Gender Play: Girls and Boys in School*. Rutgers University Press, New Brunswick, NJ
West C, Zimmerman D H 1987 Doing gender. *Gender & Society* **1**(2): 125–51

J. Sprague

Gender and Health Care

The healthcare system is a deeply gendered social institution, often affecting men and women in dramatically different ways. In the United States and many European countries, healthcare occupations are sex segregated: men constitute a majority of physicians, while women are concentrated in occupations that are less prestigious and poorly paid. Because women are more likely than men to seek medical treatment, they are the principal consumers of healthcare. However, research has demonstrated differences in the kinds of healthcare men and women receive, differences that often place women at a disadvantage. Together with inequalities of race, socioeconomic status, and age, gender shapes both the healthcare delivery system and the quality of healthcare men and women receive (see *Socioeconomic Status and Health*; *Minority Access to Health Services: United States*). Focusing on developments in the United States, this article discusses men and women's encounters with the healthcare system, as both practitioners and recipients of healthcare.

1. Origins of Interest in Women and Healthcare

Interest in gender and healthcare developed from the engagement of the social sciences with the second wave of feminism that flourished in the 1970s. Feminists made medicine the centerpiece of their analysis, arguing that no institution so clearly epitomized women's subjugation. They developed a far-reaching critique of the healthcare system, which attacked: medicine, for its exclusion of women; the overuse and misuse of drugs, surgery, and technology; and physicians, for withholding information from their patients and trivializing women's concerns. Feminists asserted the right to be full and equal participants in medical decisions.

The women's health movement was particularly influential in the social sciences with, historians, anthropologists, and sociologists testing, applying, and extending feminist ideas. The result was a body of theory and research that reshaped how disciplines conceived of healthcare. In medical sociology, for example, theory and research on gender and healthcare has challenged fundamental assumptions of the medical model (Mishler 1981). Researchers have challenged the assumption that definitions of health and illness are unaffected by the social context by showing them to be culturally variable and historically contingent. Anthropologists, for example, have shown considerable cross-cultural variation in how menstruation and menopause are defined and treated (Lock 1993). Studies revealing gender showing that male and female patients presenting the same symptoms receive different diagnoses and treatments have challenged the assumption that medicine is scientifically neutral. Research on gender and health has been part of the move to create a social perspective on health, illness, and medicine.

2. Gender and the Medical Division of Labor

2.1 Gender and Healing: A Brief Historical Overview

Both men and women have been healers, although their roles have varied historically. By the end of the Middle Ages, the medical division of labor had assumed a form that was to continue in many European countries until the nineteenth century: a largely male stratum, consisting of guilds and dominated by university-trained physicians; and midwives and folk healers, usually women, who served the rest of the population (Ehrenreich and English 1978).

In nineteenth century America, multiple paradigms of healing existed in competition. The predominately male allopaths or 'regulars' served a wealthy clientele in Eastern cities. Regulars were known for treating disease with harsh remedies, such as purging and bleeding. Women were often rejected from regular medical schools and forced to attend women's colleges or proprietary colleges. Women (as well as some African–Americans) were also more likely to enter the schools of the sects that proliferated during that period. In contrast to regulars, sectarians avoided harsh remedies, and most served a rural or working-class clientele (Ehrenreich and English 1978).

At the beginning of the twentieth century, the leadership of the regulars, influenced by progressivism, began a campaign to make medical education more 'scientific,' culminating in the Flexner Report of 1910, a study of medical education commissioned by the Carnegie foundation. As a result of both licensing and the Flexner Report, medical education became university-based and grounded in laboratory science instruction. There were, however, additional consequences. Lacking the resources to build laboratories, proprietary colleges, including sectarian colleges, saw their enrollments dwindle. With the closure of the proprietary colleges, most women, African–Americans, and members of the working class, could no longer obtain a medical education. Medicine became a profession of a white, male elite. The process continued as many states outlawed midwives (Ehrenreich and English 1978) (see *Medical Profession, The*).

Excluded from medicine, many women chose the fledgling profession of nursing. In both the United States and England, the first professional nursing programs were established in the nineteenth-century after Florence Nightingale successfully commanded a contingent of nurses during the Crimean War. Some of Nightingale's views were to become a mixed blessing in the twentieth century. During the Crimean War, she ruled that no nurse could act without explicit orders from a doctor. Throughout the twentieth century, nursing has struggled to regain the autonomy from medicine that Nightingale had forfeited (Freidson 1970).

2.2 *The Contemporary Medical Division of Labor*

For most of the twentieth century, the gender composition of the American healthcare occupations has followed a single principle: the higher the prestige, power, and pay of the occupation, the smaller the proportion of women. Medicine, the most lucrative and prestigious occupation in the United States, remains a predominately male profession. Although the number of women physicians increased fivefold since 1970, in 1998 less than 23 percent of all physicians were women. Only three percent of all women physicians—or less than one percent of all physicians—were African–American (Pasco et al. 2000–1).

Medicine also remains internally stratified by gender. Women generally are concentrated in the medical specialties having the lowest incomes, such as pediatrics, psychiatry, and family practice—specialties involving considerable direct patient contact. Conversely, the lowest proportions of women are found in the most lucrative specialties, radiology and surgery—specialties that involve medical procedures and require less patient contact. Moreover, the medical subspecialties are themselves internally stratified by gender.Very few women enter the most lucrative surgical subspecialties, such as orthopedic surgery. Within internal medicine, the proportion of women is lowest in cardiology and gastroenterology, subspecialties similar tosurgery in their use of procedures and their salary structures (Pasco et al. 2000–1, Wassenaar and Thran 2001–2). The practice patterns of men and women physicians place women at a disadvantage. Women physicians are more likely than men to work in the least lucrative practice arrangements, that is, to be salaried employees (Wassenaar and Thran 2001–2). Less than five percent of women physicians enter administration, teaching, and research—sectors in which the leadership of the profession is found (Pasco et al. 2000–1).

Discrimination against women physicians extends to medicine's informal structure. Disturbing proportions of women physicians report experiencing sexual harassment (Olesen 1997). Women physicians are often excluded from sponsorship networks that channel medical students into prestigious specialties and enable younger physicians to establish practices (Lorber 1984).

Contrasting with medicine is a middle level consisting almost entirely of women, most of whom are white women (Olesen 1997). This level includes the allied health professions–registered nurses, pharmacists, dieticians, and therapists—as well as licensed medical technicians. These professions are also stratified according to race, with the largest percent of women of color found in the occupations with the lowest salaries.

Women allied health professionals and technicians usually advance more slowly and than men in those occupations. For example, the small number of men who have entered nursing have been promoted very quickly to administrative positions (Williams 1992). Women earn less than men who do the same work, and they also earn less than men in predominately male occupations who do different but comparable work. Pharmacy, the only predominately male allied health profession, is also the most lucrative. Pharmacists earn more than nurse supervisors even though researchers found the latter job to be more demanding (Achtenberg 1991) (see *Health Professionals, Allied*)

At the bottom of the occupational hierarchy are the health service occupations, consisting primarily of aides, orderlies, and attendants. Women represent

nearly 90 percent of this group, which also has the greatest proportion of black workers. As care of the chronically ill moved out of the hospital and into the home, the number of home health aides has increased. These jobs, typically performed by women of color and immigrants, are compensated at levels at or below the poverty line (Olesen 1997).

Women also perform most of the unpaid, informal caregiving for sick relatives in the home. Caregiving entails a number of strains, including: loss of social relationships outside the family; interference with work roles; and the strain of performing multiple caregiving roles simultaneously (Olesen 1997). Perhaps the greatest strain arises from the assumption that responsibility for providing care rests exclusively with women (see *Caregiver Burden*)

3. *Gender and Medical Encounters*

Although the number of female physicians increased dramatically, the most typical medical interview is likely to involve a male physician and a female patient. This pattern raises the question of whether the gender of either physician or patient affects the physician–patient relationship. *Are Male and Female Patients Treated Differently?* Differential treatment can assume two forms. First, physicians can interact differently with female patients by adopting a more controlling communication style or by giving them less information. Second, physicians may give a different diagnosis or recommend a different treatment to male and female patients with similar medical problems.

Some of the first studies to identify the structure of the medical interview were conducted in obstetrical and gynecological clinics. These studies found that physicians: persuaded patients to use or forego specific treatments by presenting information selectively; made recommendations that reflected cultural assumptions about gender (e.g., that only married women should have children); and gave poor women fewer choices than educated, middle-class patients (Fisher and Todd 1986).

These studies show that assumptions about gender affect the content of the medical encounter but do not demonstrate that male and female patients are treated differently. Studies comparing physicians' communication with male and female patients present contradictory findings. On the one hand, physicians spend more time with female patients and give them more explanations, possibly because women present more symptoms and ask more questions (Elkderkin-Thompson and Waitzkin 1999). On the other hand, communication between male doctors and female patients is likely to be viewed as problematic: physicians report liking male patients better, and women are more likely to report that physicians talked down to them and trivialized their concerns (Elderkin-Thompson and Waitzkin 1999).

The second question, whether gender influences medical decision making, has been studied extensively. Research undertaken in the 1970s and 1980s tested the hypothesis that primary-care physicians share the commonsense stereotype of women as more emotional than men and are more likely to interpret the women's physical complaints as psychogenic. Consequently, physicans may overlook a physical disease or incorrectly diagnose a psychological disorder. Studies of routine decisions of internists and family practitioners provide strong support for the hypothesis that physicians stereotype female patients as more emotional, and weaker, more equivocal support for the hypothesis that stereotypes result in unequal treatment (Weiss and Lonnquist 1997).

In the late 1980s and 1990s, research on gender bias focused on specialists' decisions, particularly those involving 'high technology medicine.' Even after adjusting for confounding variables, many studies have found male and female patients with heart disease to be treated differently at one or more of the following decision points: women are less likely than men to receive prompt attention in the emergency room; less likely to be hospitalized; less likely to receive noninvasive diagnostic studies, such as treadmill tests, less likely to be treated with thrombolytics ('clot busters'), and less likely to receive invasive diagnostic procedures, such as catheterization. Women, particularly black women, are less likely to undergo angioplasty or undergo coronary bypass surgery. When they do receive angioplasty or surgery, they are referred later in the course of their disease—a pattern that may contribute to women's higher surgical mortality (Beery 1995). Recently, women have also been found to be less likely than men to receive kidney transplants (Elderkin-Thompson and Waitzkin 1999).

These findings should be interpreted cautiously for two reasons. First, a substantial number of studies has failed to find evidence of gender differences in one or more of the treatments of heart disease that have been mentioned. Secondly, because these studies use information extracted from written documents, they do not reveal the actual decision-making process culminating in differential treatment. Several explanations of these findings regarding heart disease have been proposed (Beery 1995). (a) Perceived risk of procedures. Physicians may be reluctant to perform angioplasty and cardiac bypass surgery on women because they view them as more risky for women. In fact, women who undergo bypass surgery are older, have higher comorbidity, and are at a later stage in their disease. However, this pattern may also reflect physicians' reluctance to diagnose and treat women with heart disease at earlier stages. (b) Beliefs about heart disease. The misconception that heart disease is a 'man's disease' may prevent both physicians and women from recognizing the symptoms of heart disease. (c) Different symptoms. Whereas men experience chest pain, women are more likely to have

shortness of breath or epigastric pain, more subtle symptoms that might be not be correctly identified. (d) Cultural assumptions about gender. This explanation receives empirical support from a study of chest pain evaluation in an emergency room, which found men more likely to receive emergency cardiology consults, nitroglycerin, aspirin, and thrombolytic agents but women more likely to receive anti-anxiety drugs (Lehmann et al. 1996). This finding suggests that physicians may fail to diagnose heart disease in female patients because of a belief that their symptoms are psychosomatic. (e) Presentation style. Some writers suggest that women's greater tendency to introduce contextual information into the medical interview may cause their complaints to be discounted or interpreted as psychosomatic (Elderkin-Thompson and Waitzkin 1999).

3.1 Do Male and Female Physicians Differ?

Many studies have examined whether male and female physicians differ in attitudes and values, choice of medical treatment, or interactive styles. This research presents a mixed picture. On the one hand, male and female family practitioners do not differ in their evaluations of common medical problems (West 1993). On the other hand, women physicians are more likely than their male colleagues to order preventive screening; to use a more collaborative mode of interaction; to engage in discussion of psychosocial issues; to deal with emotions; and to create opportunities for their patients to participate in the medical exchange (Roter and Hall 1998). When giving instructions to patients, male physicians are more likely to use the imperative, whereas women are make requests rather than commands, thereby creating more symmetry in the doctor-patient relationship (West 1993). In most cases, then, the physician's gender is more likely to affect the *style* of communication than to affect the *content* of medical decisions.

Although women first-year medical students show greater commitment to the interpersonal dimension of healthcare, by the time they have completed medical training, women are no more patient-oriented than their male colleagues. Sex role socialization, then, is only one of many forces that shape physicians' attitudes and behavior. Its effects diminish as men and women undergo a similar training and encounter pressures arising from similar work environments (Weiss and Lonnquist 1997, Olesen 1997).

These considerations suggest that increasing the number of women in medicine is a necessary, but ultimately insufficient condition for creating a humane healthcare system responsive to the diverse needs of both men and women. It is necessary because of what women contribute to the medical encounter. It is insufficient, however, because of the absence of women of color and because very few women physicians enter teaching, research, and administration. What is needed are women who occupy positions of leadership within medicine and are able to reshape the medical curriculum, revise the textbooks, and influence funding priorities.

4. Gender and the Social Construction of Health and Illness

Research on gender and health has demonstrated that conceptions of health and illness are historically contingent, culture-bound, and ineluctably tied to the social context out of which they arise. These changing conceptions have affected the ways women's medical problems have been treated.

In the nineteenth century, for example, Victorians viewed women's psychological complaints as diseases of the uterus or ovaries caused by deviation from their social roles. Their treatments ranged from 'rest cures' to sexual surgery to remove the uterus or ovaries (Ehrenreich and English 1978). Ostensibly scientific ideas were used to make claims about women's proper place.

Although these ideas no longer hold sway, twentieth-century conceptions of illness, like those in the nineteenth, continue to be socially created. In the twentieth century, social and cultural forces have led to the increasing medicalization of women's lives—a process that proved consequential for women's healthcare.

4.1 Medicalization

A persistent trend in Europe and America is the extension of the medical model to an increasing number of deviant behaviors and natural processes. Women, particularly middle-class women, have been deeply affected by the tendency to treat the reproductive phases in the female life course—childbirth, menstruation, and menopause—as illnesses (Riessman 1983).

Since the 1800s, childbirth in the United States has been transformed from a natural event, taking place at home and assisted by midwives into a medical event managed by obstetricians. By the middle of the twentieth century, childbirth in America had been medicalized to an extent unparalled in elsewhere in the world. Women were subjected to myriad medical procedures, such as epidural anesthesia, forceps, and fetal monitors. However, US infant mortality rates continue to surpass those of other European countries in which medicalization has been less extensive (Riessman 1983).

Medicalization is also behind the controversial definition of premenstrual syndrome (PMS) as a

psychiatric disorder. Despite the lack of consensus as to how the condition should be defined and despite the lack of evidence that women's moods vary according to their menstrual cycles, PMS was included in the revised third edition of the American Psychiatric Association's *Diagnostic and Statistical Manual* and has even been proposed as a criminal defense (see Tavris 1992). Some doctors have treated women diagnosed with PMS with hormones. The medicalization of PMS can be a double-edged sword. On the one hand, women may be relieved to learn that they suffer from a real illness. On the other hand, the medicalization of PMS in the criminal law may revive the nineteenth-century view that women's mental lives are controlled by their reproductive systems.

There is reason to believe that the origins of PMS are cultural rather than hormonal. Since 1931, descriptions of this condition have varied considerably. In periods in which women were expected to work, the effects of PMS were depicted as mild. During periods in which women were expected to leave the labor market, the effects of PMS were described as debilitating. Further, although middle-class women interviewed in one study viewed themselves as suffering from PMS, working-class women did not believe in its existence.

In Europe and the United States, menopause, once viewed as a natural process has been defined as a disease requiring medical treatment. Inequalities of age and gender combine to create a highly negative view of menopause in medical texts, which depict menopause 'hormonal failure' having far-reaching consequences for the heart, the skin, and the skeleton and requiring treatment with hormone replacement therapy (Lock 1993).

These views of menopause are deeply culture-bound. In Japan, views of the aging woman are shaped by concerns with population aging, cultural views of the women's roles rooted in ancient Japanese society, and the preoccupation with work. In Japan, menopause has not been medicalized although aging women are nevertheless stigmatized as unemployed caretakers. Since women are viewed as 'natural' caretakers, government officials have suggested they should become full-time caregivers of both aging parents and aging in-laws. It is these issues, rather than their hormones, that concern Japanese women (Lock 1993).

4.2 Medicalization and Medical Technology

Medicalization is often accompanied by the growth of medical technologies. In fact, medicalization is encouraged by the drug companies, who have a stake in promulgating the ideas that pregnancy, PMS, and menopause are diseases requiring treatment.

Early studies of estrogen replacement therapy, DES, and the Dalkon shield focused on their rapid diffusion into medical practice before their consequences were known and the drug companies' delay in recalling these technologies even after their risks had become apparent. (see Riessman 1983). This pattern, however, has occurred with other medical technologies unrelated to women's health.

More recent research has focused on the social consequences of new reproductive technologies, such as genetic screening, *in vitro* fertilization, and fetal surgery healthcare. Some new reproductive technologies have diminished women's control in decision making, while augmenting the power of professionals, social movements or the state. For example, the use of implantable contraceptives by the poor or by women in developing nations raises the spectre of a powerful group or society controlling the fertility of the less powerful (see Ruzek et al. 1997). New diagnostic technologies diminish the importance of women's experience of their bodies and force them to depend on experts to interpret the new technologies. With the development of ultrasound, for example, women became less reliant on their own experience of fetal movements and more reliant upon experts to help them 'see' the fetus in an otherwise obscure sonogram (Duden 1993).

Some commentators fear the potential of some new medical technologies to devalue the mother while increasing the social value of the fetus. For example, the new specialty of fetal surgery results in the creation of an unborn patient—the fetus—while the mother becomes peripheral to the treatment process (Casper 1998). This trend has culminated in the fetal rights movement, in which African–American mothers accused of using drugs have been prosecuted for allegedly endangering fetal health or even forced to undergo Caesarian sections. In these cases, women lose control of reproduction, which has been taken over by the state.

Most critics of new medical technologies do not simply assume that all medical technologies are intrinsically harmful to women. Rather, most take the more nuanced view that the consequences of any technology depend largely on the social context in which it is deployed.

4.3 Demedicalization of Reproduction

In the United States, broadly-based social movements have attempted more recently to demedicalize childbirth. In the 1960s, some middle-class parents sought alternatives to the medical model, even to the extent of bypassing the hospital altogether and turning to lay midwives. At the same time, the women's health movement challenged professional dominance even to the extent of usurping what had been previously considered professional prerogatives (Ruzek et al. 1997).

As a result of these social movements, alternatives to labor and delivery have developed, including home

births using lay midwives or nurse midwives and delivery by nurse-midwives in free-standing birth centers (Rothman 1983, Riessman 1997). These alternatives encourage mothers and families to become actively involved in the birth process, avoid technological interventions, and rely on hospitals only when complications arise.

While organized medicine has opposed home births and attempted to regulate freestanding birth centers, hospitals have accommodated the demands of many middle-class parents for a less medicalized childbirth by creating alternative birth centers in hospitals and by offering Lamaze classes. These changes, however, do not fundamentally challenge the technological and interventionist model of conventional medicine.

Despite opposition, home births and freestanding birth centers have continued through the end of the twentieth century. However, these social movements have appealed exclusively to a small, educated, middle-class clientele—a limitation that will be discussed in the next section.

5. *Intersecting Inequalities*

In the 1980s and 1990s, research on gender and health moved away from an exclusive focus on gender and toward understanding how gender intersects with race, class, and other forms of inequality. Academic feminist influenced by postmodernism criticized treating gender as a unitary category and eliding differences among women. As scholars and activists struggled to understand why the women's health movement had appealed only to white, middle-class women, it became apparent that many women had perspectives and interests that diverged from those of white, middle-class women. For example, disability activists oppose prenatal testing and the selective abortion of 'defective' fetuses—policies that, in their view, symbolically devalue the disabled. Nonwhite women and women of low socioeconomic status, lacking access to basic healthcare, seek to the very high-technology medicine that white, middle-class women reject. Women of color, for whom forced sterilization campaigns in the 1970s are a recent memory, seek protection from the enforced use of the very technologies white women view as a social good. Because medical technologies may benefit some women to the detriment of others, it is misleading to extrapolate from the experiences of one group of women and generalize about women as a whole (Ruzek et al. 1997) (see *Minority Access to Health Services: United States*; *Socioeconomic Status and Health Care*).

The focus on differences among women has opened avenues of inquiry. One line of research examines how the divergent worldviews of middle and working-class women result in different perspectives on abortion or reproduction (Luker 1984). Another line of research explores the role of race, class, and gender in shaping healthcare policies. For example, the early birth control movement emphasized the potential of contraception to enhance reproductive freedom of all women. As control of the movement shifted to physicians and eugenicists, its focus shifted to population control, directed toward controlling the fertility of those deemed 'unfit' (Gordon 1983). Future research should also examine the differential impact of healthcare policies on different groups of women.

The mandate to study differences among women consumers of healthcare has yielded a richer and more complex picture of diverse women, whose perspectives and interests sometimes collide and sometimes coincide. Once these differences have been identified, it may be possible to draw more meaningful generalizations about the kind of healthcare that serves diverse women.

See also: Care and Gender; Caregiving in Old Age; Gender and Cardiovascular Health; Gender and Physical Health; Gender and Reproductive Health; Gender Role Stress and Health; Health Care Markets: Theory and Practice; Health Care Delivery Services; Medical Sociology; Medicine, History of; Men's Health; Minority Access to Health Services: United States; Reproductive Medicine: Ethical Aspects; Reproductive Rights in Developing Nations; Socioeconomic Status and Health Care; Women's Health

Bibliography

Achtenberg J 1991 Woman as healer: A panoramic survey of the healing activities of women from prehistoric times to the present. Shambhala, Boston, MA

Beery T 1995 Gender bias in the diagnosis and treatment of coronary artery disease. *Heart and Lung Journal of Critical Care* **24**(6): 427–35

Casper 1998 *The Making of the Unborn Patient: A Social Anatomy of Fetal Surgery*. Rutgers, New Brunswick, NJ

Duden B 1993 *Disembodying Women: Perspectives on Pregnancy and the Unborn*. Harvard University Press, Cambridge, MA

Ehrenreich B, English D 1978 *For Her Own Good: 150 Years of the Experts' Advice to Women*. Anchor, New York

Elderkin-Thompson V, Waitzkin H 1999 Differences in clinical communication by gender. *Journal of General Internal Medicine* **14**(2): 112–41

Fisher S, Todd A 1986 Friendly persuasion: The negotiation of decisions to use oral contraceptives. In: Fisher S, Todd A (eds.) *Discourse and Institutional Authority:Medicine, Education, and Law*. Ablex, Norwood, NJ

Freidson E 1970 *Profession of Medicine*. Aldine, Chicago

Gordon L 1983 The politics of birth control, 1920–1940: The impact of professionals. In: Fee E (ed.) *Women and Health: the Politics of Sex in Medicine*. Baywood, Farmingdale, NY

Lehmann J, Paulette S, Lehmann C, Savory L 1996 Gender bias in the evaluation of chest pain in the emergency department. *American Journal of Cardiology* **77**(8): 641–4

Lock M 1993 The politics of mid-life and menopause: Ideologies for the second sex in North America and Japan. In: Linden-

baum S, Lock M (eds.) *Knowledge, Power and Practice: The Anthropology of Medicine and Everyday Life*. California, Berkeley, CA
Lorber J 1984 *Women Physicians: Careers, Status, and Power*. Tavistock, New York
Luker K 1984 *Abortion and the Politics of Motherhood*. California, Berkeley, CA
Mishler E 1981 Critical perspectives on the biomedical model. In: Borwn P (ed.) *Perspectives in Medical Sociology*. Wadsworth, Belmont, CA
Olesen V 1997 Who cares? Women as informal and formal caregivers. In: Ruzek S, Olesen V, Clarke A (eds.) *Women's Health: Complexities and Differences*. Ohio State University Press, Columbus, OH
Pasco T, Silverman B, Birkhead S (eds.) 2000–1 *Physician Characteristics and Distribution in the United States, 2000–2001*. American Medical Association Center for Health Policy Research, Chicago
Riessman C 1983 Women and medicalization: A new perspective. *Social Policy* **14**: 3–18
Riessman C 1997 (1981) Improving the health experiences of low income patients. In: Conrad P (ed.) *The Sociology of Health and Illness: Critical Perspectives,* 5th edn. St Martin's, New York
Roter D, Hall 1998 Why physician gender matters in shaping the physician-patient relationship. *Journal of Women's Health* **7**(9): 1093–7
Rothman B 1983 Midwives in transition: The structure of a clinical revolution. *Social Problems* **30**(3): 262–71
Ruzek S, Clarke A, Olesen V 1997 What are the dynamics of differences? In: Ruzek S, Olesen V, Clarke A (eds.) *Women's Health: Complexities and Differences*. Ohio State University Press, Columbus, OH
Tavris C 1992 *The Mismeasure of Woman*. Simon and Schuster, New York
Wassenaar J, Thran S 2001–2 *Physicians Socioeconomic Statistics, 2002 Edition*. American Medical Association Center for Health Policy Research, Chicago
Weiss G, Lonnquist L 1997 *The Sociology of Health, Healing, and Illness*, 2nd edn. Prentice Hall, Upper Saddle River, NJ
West C 1993 Reconceptualizing gender in physician-patient relationships. *Social Science Medicine* **36**(1): 57–66
Williams C 1992 The glass escalator: Hidden advantages for men in the 'female' professions. *Social Problems* **39**: 53–67

R. R. Anspach

Gender and Language: Cultural Concerns

Research on gender and sex in sociolinguistics and discourse analysis started systematically in the early 1970s, probably as a result of women's liberation movements. Investigators examined two domains of language behavior in particular: speech behavior of men and women on the phonological level, and interactions (conversational styles) between women and men in discourse (for overviews and the history of the field see Wodak and Benke 1996, Holmes 1996, Wodak 1997, Kotthoff and Wodak 1997, Wilkinson and Kitzinger 1995, Crawford 1995, Hall and Buchholtz 1995).

1. *Setting Out the Problems*

Studies of gender-specific language behavior are often contradictory and depend on the authors' implicit assumptions about sex and gender, methodology, and samples used, etc. For example, some studies suggest that women use more standard language than men because they try to adapt to social norms (Trudgill 1974). On the other hand, other large-scale surveys have shown that young women tend to speak dialect and nonstandard language as well (Wodak and Dressler 1978), due to specific situations like family conversations and peer group talk.

The socialization of infants is often regarded as primarily the duty and work of mothers; mothers use 'motherese,' a very specific discourse adapted to the needs of infants (see Wodak and Rindler-Schjerve 1985); as a result of caretaking, children acquire the regional dialects and vernacular which their mothers use at home. Other investigations illustrated that professional women are under more pressure than their male colleagues; therefore they adopt high standard sociolects or even hypercorrect features (see Labov 1966, Moosmueller 1997). However, hypercorrectness is also clearly related to social class: Working-class and lower-middle-class men also tend to speak hypercorrectly in certain organizational contexts (Leodolter 1975). Very detailed case studies of code switching, moreover, have shown that the use of either standard or nonstandard styles as well as minority and majority languages depends highly on the context of the utterances as well as on social class (Gal 1989). Nonstandard lects can also be the manifestation of certain political opinions and attitudes as in the case of Ebonics (see Baugh 1999).

The same variation of correlations is true for studies on conversational styles: In some cultures, men clearly interrupt women more frequently and dominate topic choice, but in other cultures this need not be the case (see Shieffelin 1990). As a result, women's language has been said to reflect their conservatism, prestige consciousness, upward mobility, insecurity, deference, nurturance, emotional expressivity, connectedness, sensitivity to others, solidarity, and men's language is heard as evincing their toughness, lack of affect, competitiveness, independence, competence, hierarchy, control (Eckert and McConnell-Ginet 1992, p. 90).

Many empirical studies have thus neglected the context of language behavior and have often analyzed gender by merely looking at the speakers' biological sex (see the arguments in Nichols 1983, Cameron and Coates 1990, Henley and Kramarae 1991). By focusing

only on the biological sex, the whole group of women is contrasted to the whole group of men without taking differences in these holistic categories into account (see below).

Such a dichotomization clearly does not make sense when looking at many studies which have shown significant differences that occur with women or with men due to culture, social class, ethnicity, etc. (Wodak and Schulz 1986, Kotthoff and Wodak 1997). Some of the research has isolated the variable of sex/gender from other sociological or situational factors and has made hasty generalizations about genderlects (Chambers 1992). A context-sensitive approach, however, which regards gender as a social construct leads to more fruitful results (see Harres 1996, Wetschanow 1995). Moreover, gender is related to and overlaps with the sociocultural and ethnic background of the interlocutors, with their age, their level of education, their socioeconomic status, their emotions, and the specific power-dynamics of the discourse investigated.

The focus nowadays lies on the discursive construction of identities in interaction where gender is but one part of this attempt (Wodak et al. 1999). Thus, research on gender should take the whole interaction and context into account; of course, the biological sex will always be part of a categorization but other categories might prove more valid in a specific research which makes it possible to operationalize gender roles (the inclusion of professional roles and organizational roles with gender is of utmost importance as well as culture and ethnicity; the socialization of women who come from Islamic cultures is so different than the background of Western European women that intercultural problems clearly arise in interactions, see Martin-Rojo 2000, Wodak 2000a, Essed 1992, 1997). All these results suggest that gender has an enormous impact in many very relevant situations, like schools, hospitals, courtrooms, etc., but only in combination with other variables. However, it is not possible and is fruitless to try to decide which variable has more impact generally. Only very detailed qualitative studies can provide insight into such complex theoretical questions.

2. *Feminist Linguistics*

Analogous to the term 'racism,' the word 'sexism' was invented in the 1960s. It refers to discrimination within a social system on the basis of sexual membership. With the concept of 'sexism,' women defined themselves for the first time as a social group and as a suppressed minority expressing basic political interests of being judged and valued in equal terms as men. Of course, the history of women's fight for equal rights is much older (see Saurer 2000) but the political movements in the 1960s, together with innovations like the invention of the contraceptive pill and birth control on a larger scale started a qualitative new time for women's emancipation. As such, they sought to reveal the mechanisms of suppression, making others aware of and fighting these devices. Social groups often define themselves by means of their common language, which plays an important role in identity creation and, for subcultures, serves as a means of differentiating themselves from the outside world. This specific identity manifests itself in certain conversational styles, manifestations of emotions, etc. (see Coates 1997, Buchholtz et al. 1999).

According to Hellinger (1990), feminist linguistics (FL) is distinguished from all other disciplines by the following three aspects, which similarly analyze the relationship between language and gender:

(a) FL places female and male linguistic behavior and the linguistic phenomena connected with the designations of women and men at the center of its considerations.

(b) FL interprets person-related asymmetries in the field of language systems and language use as expressions of the linguistic discrimination of women (sexism) and links these directly to the plane of social discrimination. Traditional studies usually make do with descriptive results. Clearly, languages differ in their linguistic realizations of such inequality. In German or Polish, for example, suffixes and certain prefixes manifest gender. Kargl et al. (1997) have submitted a whole register of possibilities of equal treatment of women and men in many everyday genres in communication, starting with names and labels up to whole reports in newspapers, as well as conversations. These attempts are valid in written as well as oral speech and have led to huge debates in the media (see Pusch 1984); in German, specifically, the use of capital 'I' in written texts was at the center of such political discussions (i.e., '*LehrerInnen*' vs. '*Lehrer*' and '*Lehrerinnen*' or '*Lehrer/innen*').

(c) FL does not accept phenomena as given, but seeks alternatives in keeping with the principle of the linguistic equal treatment of women and men. It pursues explicitly political goals by criticizing ruling linguistic norms and understanding the linguistic change it advocates as part of an overall change in society (Hellinger 1990, p. 12).

3. *Sociolinguistic Studies of Sex/Gender*

In the 1970s, 'sex' was established within sociolinguistic research as a social variable next to the already existing variables of social stratum, age, nationality, ethnic affiliation, religion, class, and region. In correlative-quantitative sociolinguistic investigations on the prestige and stigma variants of languages (see Kotthoff 1992) the sex variable became a factor which significantly affected the use of language. The best-known representatives of this quantitative-correlative approach in sociolinguistics,

with its focus on urban groups of persons, are William Labov, Peter Trudgill, and Lesley Milroy. The working methods within this sociolinguistic approach diverge considerably from one another in certain aspects (unlike Trudgill and Labov, Milroy works with network analyses) but in one respect they are identical: their methodology. All of them operate quantitatively, that is, linguistic variables are defined that are realized in different forms in different varieties. Then, the different variants are correlated with sociological parameters.

Several other studies have taken different approaches and have used new and other methodologies. In particular, a lot of criticism related to the implicit chauvinistic ideologies in these first studies and also to the undifferentiated analysis of the two sexes which totally neglected the range of variation in each of the genders under observation. In a study of the language of mothers and daughters in Vienna, for example (Wodak 1984, Wodak and Schultz 1986), it could be shown that, apart from these sociological parameters, other variables, such as psychological ones, are also responsible for the variation concerning each gender. This study stressed in particular the variation that was found among the women and girls investigated, one that relativized the factor of 'sex.' It could be shown that mothers with a positive relationship to their daughters spoke differently than mothers with a conflicting relationship. The same is true for the daughters (30 mother–daughter pairs were interviewed and tested). These results illustrate that emotions and relationships as well as educational styles and the content/discontent of mothers are more important then social class in an isolated way. Moreover, this study proves that the differences in one category of sex can be statistically significant due to all these parameters.

Such results and tendencies are of utmost importance for the whole debate on gender and sex as well as social class. First, it could be shown that significant differences exist in one category (either women or men). Second, many assumptions about the differences between working-class and middle-class socialization could be contradicted—educational styles do not vary so much by social class but more by the self-assessment of mothers and their attitudes to their own life and achievements. Third, it could be shown that no significant differences appear between mothers who work at home and mothers who have professional lives outside of their homes. Therefore, many claims that children need mothers who only stay at home are wrong; the satisfaction of women with their own lives is the most important factor that affects their relationship with others, in their families or outside of their homes. Fourth, generalizations about the use of vernacular and dialect could also be contradicted: Even daughters from upper-class families used Viennese dialect to distinguish themselves from the phonological styles of their mothers and to be able to manifest their autonomy. This study illustrates the necessity of a very complex and interdisciplinary approach to gender research.

Another approach of considerable importance in recent sociolinguistics is the concept of 'communities of practice' (Eckert and McConnell-Ginet 1992, p. 92ff). Communities of practice are defined as 'An aggregate of people who come together around mutual engagement in some common endeavor. Ways of doing things, ways of talking, beliefs, values, and power relations—in short, practice—emerge in the course of their joint activity around that endeavor' (Eckert and McConnell-Ginet 1992, p. 95).

This approach relates to the proposals of context sensitivity made above: While researching gender, the emotions, attitudes, values, intergroup and intragroup relationships have to be accounted for as well. Specifically, Eckert and McConnell-Ginet studied peer groups of girls and boys and could illustrate why big differences on a phonological and conversational level occur between two groups due to all the above-mentioned factors. Eckert and McConnell-Ginet argue very convincingly and provide many examples to support their view that survey studies are too general and their level of abstractness too great because all the important sociological and psychological parameters cannot be integrated. Moreover, and this criticism applies to survey studies in general about other topics as well (e.g., research on anti-Semitism and racism, Wodak et al. 1990), surveys usually manifest self-assessments of the interviewees. A bias due to education must be considered as well: Better educated interviewees usually know what kind of answers are expected from them. They answer accordingly even though this might not be their opinion at all. This bias is very important in all studies related to stereotypes and prejudice research. Therefore, such quantitative studies must be complemented with qualitative case studies using authentic data tape-recorded in everyday interactions as well as ethnographic observations (see Titscher et al. 2000).

4. *Sexism and the Language System*

Feminists conceive language as 'a symbolical reflection of androcentric structures' (Günthner and Kotthoff 1991, p. 7), as one of the means of patriarchal society to discriminate, disregard, and incapacitate women. In their view, the language system already reflects the patriarchal structure of Western societies. Such an assumption manifests a basic consideration of sociolinguistics, namely that language and social norms are connected and that language manifests and also constructs social practices (Fairclough and Wodak 1997). If this claim is accepted (and such theories stem from Wittgenstein's 1967 concept of 'language games'), then the manifestation of women's domination must also be manifest in language.

The assumption that an individual language system has lexical elements and morphological and grammatical rules that are already sexist is based on the premise that 'due to their long history as public decision-makers, men not only determine the economic, political and social orientation of social life, but also influence the functioning and the semantic contents of each individual language' (Postl 1991, p. 89).

Once the language system has become the object of investigation for feminist linguists, it is interesting to note how the linguistic structure of an individual language is connected to the structure of society, how the structure of the language is conditioned by the structure of society, and vice versa. Unlike studies investigating divergent gender linguistic behavior, studies on the sexist use of language focus on the possibilities of reference to both genders or their practice that exist in an individual language. According to Pusch (1984, p. 7), the objects of 'feminist system linguistics' are 'patriarchalisms in diverse language systems.'

5. *The Analysis of Conversation and Discourse*

Unlike the above-mentioned research field, FL—which examines the use of language—concerns itself with gender differences in conversation and discourse (oral as well as written). Differences are investigated in the following fields of language use: voice, pronunciation, intonation, choice of words, argumentation, lexicon, syntax, interactional and conversational behavior, as well as visual features and modes and nonverbal communication. The gender-induced differences in the use of language are not treated simply as divergent variants standing side by side. Due to a patriarchally organized society, the interpretation of the differences ascertained is of major social and political significance.

The interpretation of the diverse linguistic indicators (like turn taking, indirectness, interruptions, and overlaps) varies according to the specific gender theory and ideology, which underlie the studies. (Due to the fact that there are no comparable standardized studies across cultures and social classes as well as contexts and ethnicities, no generalizations about any universal gender or sex-based differences in conversational styles are possible.) Probably the two best-known theories about the gender-induced use of language are those of 'difference' and 'dominance' (see Cameron 1997, Kendall and Tannen 1997).

Whereas investigations following the 'deficit model' have interpreted the use of such indicators as manifestations of 'powerless language of women,' research in the paradigm of the 'difference model' views the linguistic behavior of men and women as originating in opposed modes of socialization and equal in their meaning and impact. Hedging, tag questions, or indirectness can be viewed 'negatively' or 'positively,' either as signs of female insecurity or as supporting conversational work, depending on the context of the discourse and the theoretical approach adopted. This controversy led to many polemical debates which in the end shows the invalidity of such dichotomous distinctions. Thus, nowadays, other research questions have become more important, viewing gender as constructed due to context.

6. *Recent Theoretical Approaches in the Feminist Study of Discourse*

6.1 *The Construction of the 'Sexes'*

In Judith Butler's arguments against the tenability of the categories of 'woman' and 'sex,' she refers to postmodernist theories. In the manner of Foucault she ascribes normative power to the idea of gender identity and the attempt to describe it. By the mere act of defining a gender identity many bodies, practices, and discourses are excluded or devalued, whereas the constructed and hence upright character of this gender identity is simultaneously concealed.

Thus, in such feminist theories (Butler 1990), the idea of two sexes is criticized as being a construction, a prediscursive factor taking the place of reality. Following this interpretation, 'sex' cannot be separated into a biologically inherent aspect and a socially acquired one. The category of 'sex' itself, in Butler's view, is a purely cultural product of discourse. Such an opinion, of course, relates well to Foucauldian theories of discourse; 'sex,' therefore, cannot be viewed as a 'universal, suprahistorical and extrasocial entity'; it is understood to be an 'integral component of every form of life' (Butler 1990, p. 23); such a view relates well to the assumptions of context sensitivity discussed above. On the other hand, 'sex,' of course, possesses a material basis: In many relevant contexts in our societies, women and men are still evaluated, judged, promoted, or paid according to their biological sex alone. These facts are not accounted for in Butler's approach. More recently, some authors have differentiated Butler's dogmatic position: It is a question of revealing 'the reproduction mechanisms, networks and institutional compulsions that ensure that constructions become persistent and resistant and appear timeless, immovable and identical with themselves' (Hirschauer 1992, p. 333). It is a matter of exposing the arbitrary construction of this binary opposition and hence also its mutability, and not of criticising the binarily organized perception as unreal.

6.2 *Doing Gender*

Unlike a research approach that accepts sexual differences as an aggregation of qualities and deals with the qualitative behavioral tendencies of women and men,

ethnomethodologically oriented studies produced a new focus of research: 'doing gender.' Such an approach complements traditional theories which viewed gender as an aggregation of static attributes and are concerned with investigating and displaying the peculiarities of women and interpreting them as 'gender-specific' or 'gender-typical' attributes so as to reveal the asymmetry of the difference between the sexes, to criticize it and to make it politically visible. The use of a bundle of characteristics and attributes to define gender complicates or renders impossible an interactional approach since attributes are 'entities' and not processes. This concept entails that definitions of gender cannot change, transform, and are not influenced by, spontaneous interactions. A further problem raised by the concept of gender as a concept of attributes is the possibility of individualization; that is, the individual who has been seen to possess or not to possess certain attributes becomes the centre of focus, and the level of the social system is neglected.

Unlike this noninteractive approach, 'doing gender' regards membership of a sex not as a pool of attributes 'possessed' by a person, but as something a person 'does.' In this sense, membership of a gender constitutes a performative act and not a fact. Gender is continually realized in interactional form. Gender is created not only in the everyday activities that characterize 'doing gender,' but also in the asymmetry of the relationship between the sexes, the dominance of the 'male' and its normativeness. Patriarchal inequality is produced and reproduced in every interaction. This concept of 'doing gender' stresses the creative potential and the embedding of gender-typical behavior in a social context.

Female Members of the European Parliament, for example, have created very different ways of positioning themselves in interactions with male colleagues in the European Commission or in the European Parliament itself. Interviews with female MEPs (Wodak 2000a) illustrate that they are very conscious of their gender roles and that they tend to construct very different ways of self-presentation in interactions: in particular, one MEP stated the following, in talking about her first experience as a rapporteur.

when I—entered the parliament— *Orientation* (lines 1–3)
on my first report it was about Leonardo
I don't know if you know:
((laughing)) well—I said 'I'm going to speak to the commissioner'
and—I—/I knew—he only speaks very bad French
and my er my French was very bad as well.
so I said 'I want to have interpretation'
So—I went to the commissioner *Complicating Actions* (lines 4–14)
with a very good int/int/interpreter
and I/I/I/I talked more than an hour with him.
because we talked the same about it
and at the end he said—
'well: I have here the advice of my: civil servants but I—agree with you:
and this and this and this all goes through—'
so you have to be: —er—
I don't knowh/how do we call it in English in /
in the Netherlands we say [brutal]
so you have to: ((laughing)) be polite *Evaluation* (lines 16–20)
but you have to—you:/you mustn't be/
you mustn't *sit* behind your -/your *desk*—
because that doesn't help. ((laughing))
but then then you have the worse system
that I tried several times *Coda* (22–31)
then you have the Council. …

In this example, which has been marked for basic narrative structure according to Labov and Waletzky's (1967) model, we see that this MEP's story is objectively about having a successful meeting with a Commissioner while acting as rapporteur on a report about Leonardo—an EU youth and education related program established in 1995 to provide financial support for professional development and job training. In lines 4–6, the complicating actions, she shows how she went to the commissioner with an interpreter, and because she and the commissioner had the same understanding of the issues involved, he was willing to support her, despite contrary advice by his 'civil servants' on the matters involved. The main point of the story, or evaluation, from the MEP's perspective, is to show that as an MEP, to get things done, you must be active and assertive, 'not sit behind your desk.' Thus, in this narrative, she positions herself as an MEP who is proactive and who will do what it takes, including arguing directly with Commissioners, to see that her voice is heard. She also orients to being a rapporteur (line 2), which carries some responsibility in a committee, and to being from the Netherlands (line 17), although this last identity is evoked only to characterize her style of work ('brutal' in Dutch, or 'assertive'). (For further details of the analysis see Wodak 2000a). This narrative also manifests the complexity of the European Parliament, the interwovenness of languages, cultures, positions, political affiliations, and gender roles, as well as professional experiences.

Another of the interviewees is oriented to a particularly wide range of identities (left, woman, Swedish, mother, political outsider, etc.) during her interview. Most striking is the way in which she repeatedly positions herself as being an 'atypical MEP.' Here we see one such occasion.

I figure here the most common—ah civil—job—for an MEP is er to be a lawyer.
me myself *I'm far from that*
the job I had doesn't even *exist* outside *Scandinavia*.
so: it's a sort of a social teacher—so
so I'm/I'm very in/an/a very special bird in this er:
Interviewer mhm mhm so now you don't feel like you—fit into sort of a *typical* MEP er
ME *no. no: no: I'm not. I'm left I'm a woman I'm Swedish*
and I'm also everything—/everything's wrong (laughs)

In this example, the MEP contrasts herself with what she considers to be a typical profile for an MEP (lawyer by profession), emphasizing the degree to which she feels different (I'm far from that … I'm a very special bird … everything's wrong'). She also points out many of the identities that she associates with and that she perceives as marking her as different (social teacher, left, female, Swedish). This MEP is very much concerned with her self-presentation as being different, different than other female MEPs and also male MEPs. Moreover, she explains her success as a female MEP by being exotic and narrates different strategies which she applies in 'doing politics and gender.'

Even such short examples illustrate the diversity and complexity of gender roles and make clear why an interdisciplinary, qualitative, and context-sensitive approach is necessary to access the whole domain of 'gender and language.' To be able to make more general observations, however, much more cross-cultural comparable research is urgently needed.

See also: Feminist Theory; Gender and Feminist Studies; Gender and Feminist Studies in Anthropology; Gender Differences in Personality and Social Behavior; Language and Gender; Sociolinguistics

Bibliography

Baugh J 1999 *Out of the Mouth of Slaves*. University of Texas Press, Austin, TX

Buchholtz M, Liang A C, Sutton L A (eds.) 1999 *Reinventing Identities. The Gendered Self in Discourse*. Oxford University Press, Oxford, UK

Butler J 1990 *Gender Trouble*. Routledge, New York

Cameron D 1997 Theoretical debates in feminist linguistics: Questions of sex and gender. In: Wodak R (ed.) *Gender and Discourse*. Sage, London, pp. 21–36

Cameron D, Coates J 1990 Some problems in the sociolinguistic explanation of sex differences. In: Coates J, Cameron D (eds.) *Women in their Speech Communities: New Perspectives on Language and Sex*. Longman, London, pp. 13–26

Chambers J 1992 Linguistic correlates and gender in sex. *English World Wide* **2**: 173–218

Coates J 1997 *Women's Friendships, Women's Talk*. In: Wodak R (ed.) *Gender and Discourse*. Sage, London, pp. 245–62

Crawford M 1995 *Talking Difference: On Gender and Language*. Sage, London

Eckert P, McConnell-Ginet S 1992 Communications of practice: Where language, gender, and power all live. In: Hall K, Buchholtz M, Moonwomon B (eds.) *Locating Power: Proceedings of the Second Berkeley Women and Language Conference*. Berkeley Women and Language Group, University of California, Berkeley, CA, pp. 89–99

Essed P 1992 *Everyday Racism*. Sage, London

Essed P 1997 Race, gender and academic leadership. In: Kotthoff H, Wodak R (eds.) *Communicating Gender in Context*. Benjamins, Amsterdam, pp. 315–34

Fairclough N, Wodak R 1997 Critical discourse analysis. In: van Dijk T A (ed.) *Discourse as Social Interaction*. Sage, London, pp. 258–84

Gal S 1989 Between speech and silence: The problematics of research on language and gender. *IPrA Papers in Pragmatics* **3**(1): 1–38

Günthner S, Kotthoff H 1991 Von fremden Stimmen: Weibliches und männliches Sprechen im Kulturvergleich. In: Günthner S, Kotthoff H (eds.) *Von fremden Stimmen. Weibliches und männliches Sprechen im Kulturvergleich*. Suhrkamp, Frankfurt, Germany, pp. 7–52

Hall K, Buchholtz M (eds.) 1995 *Gender Articulated. Language and the Socially Constructed Self*. Routledge, London

Harres A 1996 Tag questions and gender in medical consultations. PhD thesis, Monash University, Melbourne, Vic

Hellinger M 1990 *Kontrastive Feministische Linguistik*. Hueber, Ismaning, Germany

Henley N, Kramarae C 1991 Gender, power and miscommunication. In: Coupland N, Giles H, Wiemann J (eds.) *Miscommunication and Problematic Talk*. Sage, London, pp. 60–88

Hirschauer S (ed.) 1992 *Die soziale Konstruktion der Transsexualität*. Suhrkamp, Frankfurt, Germany

Holmes J 1996 Sex and language. In: Goebel H, Nelde P H, Stary Z, Wölk W (eds.) *Kontaktlinguistik—Contact Linguistics—Linguistique de contact*. de Gruyter, Berlin, pp. 126–46

Kargl M, Wetschanow K, Wodak R, Perle N 1997 *Kreatives Formulieren. Anleitungen zum geschlechtergerechten Sprachgebrauch*. Schriftenreihe der Frauenministerin, Vol. 13, Vienna

Kendall S, Tannen D 1997 Gender and language in the workplace. In: Wodak R (ed.) *Gender and Discourse*. Sage, London, pp. 81–105

Kotthoff H 1992 Unruhe im Tabellenbild? Zur Interpretation weiblichen Sprechens in der Soziolinguistik. In: Günthner S, Kotthoff H (eds.) *Die Geschlechter im Gespräch: Kommunikation in Institutionen*. Metzler, Stuttgart, Germany, pp. 126–46

Kotthoff H, Wodak R (eds.) 1997 *Communicating Gender in Context*. Benjamins, Amsterdam

Labov W 1966 *The Social Stratification of English in New York City*. Center for Applied Linguistics, New York

Labov W, Waletzky J 1967 Narrative analysis: oral versions of personal experience. In: Helm J (ed.) *Essays on the Verbal and Visual Arts*. University of Washington Press, Seattle, pp. 12–45

Leodolter R 1975 *Das Sprachverhalten von Angeklagten bei Gericht*. Scriptor, Kronberg, Germany

Martin-Rojo L 2000 Narratives at work: When women take on the role of managers. Workshop on Theory and Interdisciplinarity in CDA 6 and 7 July 2000, University of Vienna

Milroy J, Milroy L 1978 Belfast: Change and variation in an urban vernacular. In: Trudgill P (ed.) *Sociolinguistic Patterns in British English*. Arnold, London, pp. 19–36

Moosmueller S 1997 The relevance of fundamental frequency contour for interruptions: A case study of political discussion in Austria. In: Kotthoff H, Wodak R (eds.) *Communicating Gender in Context*. Benjamins, Amsterdam, pp. 401–21

Nichols P C 1983 Linguistic options and choices for black women in the rural South. In: Thorne B, Kramarae C, Henley N (eds.) *Language, Gender and Society*. Newbury House, Cambridge, UK, pp. 54–68

Postl G 1991 *Weibliches Sprechen*. Suhrkamp, Frankfurt, Germany

Pusch L 1984 *Das Deutsche als Männersprache*. Suhrkamp, Frankfurt, Germany

Saurer E 2000 *Liebe und Arbeit: Geschlechterbeziehungen in Europa 19 und 20 Jahrhundert*. Fischer–Tuschenbuchuerlag, Frankfurt am Main
Schieffelin B 1990 *The Give and Take of Everyday Life: Language Sozialization of Kaluli Children*. Cambridge University Press, Cambridge, UK
Titscher S, Meyer M, Wodak R, Vetter E 2000 *Methods of Text and Discourse Analysis*. Sage, London
Trudgill P 1972 Sex, covert prestige, and linguistic change in the urban British English of Norwich. *Language in Society* **1**: 179–96
Trudgill P 1974 *The Social Differentiation of English in Norwich*. Cambridge University Press, Cambridge, UK
Wetschanow K 1995 Als wenns a Grammatikfehler wär: Splittingverhalten einer Gruppe mit links-alternativ feministischer Unisozialisation. MA thesis, University of Vienna
Wilkinson S, Kitzinger C 1995 *Feminism and Discourse*. Sage, London
Wittgenstein L 1967 *Philosophische Untesuchungen*. Suhrkamp, Frankfurt
Wodak R 1984 *Hilflose Nähe? Mütter und Töchter erzählen*. Bundesverlag, Vienna
Wodak R (ed.) 1997 *Gender and Discourse*. Sage, London
Wodak R 2000 National and transnational identities: European and other identities oriented to in interviews with EU officials. Workshop on European Identity, European University, Florence, June 2000
Wodak R, Benke G 1996 Gender as a sociolinguistic variable. In: Coulmas F (ed.) *The Handbook of Sociolinguistics*. Blackwell, Oxford, UK
Wodak R, de Cillia R, Reisigl M, Liebhart K 1999 *The Discursive Construction of National Identity*. Edinburgh University Press, Edinburgh, UK
Wodak R, Dressler W 1978 Phonological variation in colloquial Viennese. *Michigan Working Papers in Germanic Studies* **4**(1): 30–67
Wodak R, Pelikan J, Nowak P, Gruber H, de Cillia R, Mitten R 1990 *'Wir sind alle unschuldige Täter!' Diskurshistorische Studien zum Nachkriegsantisemitismus*. Suhrkamp, Frankfurt, Germany
Wodak R, Rindler-Schjerve R 1985 Funktionen der Mutter beim Sprachwechsel—Konsequenzen für die Primärsozialisation und Identitätsentwicklung. In: (ed.) *Linguistische Berichte*. Westdeutscher Verlag, Wiesbaden, Germany, pp. 81–82
Wodak R, Schultz M 1986 *The Language of Love and Guilt*. Benjamins, Amsterdam

R. Wodak

Gender and Physical Health

The health experiences of women and men differ, primarily reflecting gender differences in the cultural and socioeconomic circumstances of women's and men's lives. Gender is an important dimension of social life which intersects with others, especially social class and age. In this article gender differences in mortality and morbidity are examined, together with explanations. The divergent structural and role positions of women and men lead to gender differences in the nature of inequalities in health, which vary over time and among societies.

1. *The Distinction Between Sex and Gender*

Expressed simply, sex refers to the biological differences between men and women, and gender refers to cultural differences relating to appropriate behavior of men and women. Gender roles are socially constructed and therefore change over time and across societies. The cultural and religious mores of a society influence the roles, attitudes, and behavior of men and women. Gender ideology prescribes and circumscribes the social behavior of men and women, and of different age or ethnic groups. Social myths and stereotypes about appropriate or supposedly 'natural' behavior for women and men impact on their health and well-being, for example, relating to body image and weight control (Lorber 1997).

1.1 *Masculinity and Femininity*

Over recent years the simple dichotomy of biological sex has been questioned, as has the cultural gender dichotomy of male/female. The distinction between masculinity and femininity is increasingly recognized as a continuous dimension, with each individual having both masculine and feminine attributes in varying degrees. Masculinity and femininity are inherently relational concepts, arising from gender relations in particular socio-historical contexts.

In Western societies, masculinity has traditionally been identified with physical courage, stoicism, hardness, endurance, aggressiveness, and taking risks. However, notions of masculinity vary with class and ethnicity, across the life course, over time, and among societies (Watson 2000). Structural factors and power relations among men in society may influence forms of masculinity which themselves have health consequences.

The 'risk' taking behavior of men is a key concern within health promotion and public health, but here masculinity is often seen as a unitary construct. Sabo and Gordon (1995) characterize masculinity as intrinsically damaging to men's health, in terms of predisposition to drinking, smoking, drug abuse, risky sports, and so on. However, this sees maleness as problematic, fixed, and immutable (Watson 2000), and fails to examine diversity among men.

In many cultures, illness is seen as a sign of weakness, closely allied to contemporary constructions of femininity, and thereby more easily accepted by women. In particular, chronic illnesses which undermine independence and self-control may threaten masculine self-identity. Masculinity and femininity may therefore influence the way symptoms are evaluated, willingness to consult healthcare providers

and to take medical remedies. This may help explain men's lower use of available health services.

2. Women's and Men's Health

2.1 Women's Health

In the 1970s and early 1980s, the primary focus of interest within medical sociology was women's health rather than gender and health. Biological differences were emphasized, linked to women's distinctive reproductive system, with social research on the specialized services for childbirth, contraception, abortion, and infertility (Doyal 1995) (see *Gender and Reproductive Health*; *Women's Health*).

Many women in developed countries expect pregnancy and birth to be a normal event, but their experience of maternity care is often a medicalized search for pathology. Medicalization in other areas of women's lives has received less attention from sociologists. There are rising rates of hysterectomy, while the menopause provides a new opportunity for medical intervention in mid-life, especially through the promotion of hormone replacement therapy (HRT).

The impact of gender roles on reproductive health needs to be considered. For example, women's capacity to control their fertility through contraception and abortion is influenced by partners, family members, governments, religious, and other social institutions. It is also important to examine the effects on women's reproductive health of poverty, class, race and ethnicity, and sexual orientation.

2.2 Women and the Male 'Standard' Patient

The focus on women and reproductive health has led to a lack of attention to women as patients within mainstream areas of health. The biomedical or scientific model assumes that disease is a deviation from 'normal' biological functioning. However, what is regarded as normal depends on who is being compared to whom. Men are treated as the 'norm' and women compared to a standard 'male' patient (Lorber 1997). Medical norms and research have primarily been based on white, male, middle class, working age bodies. This approach also disadvantages men, since they are seen as the normal, the 'control,' to which women are compared. This inhibits exploring diversity among men's health and recognizing that maleness itself is a cultural and social phenomenon.

Epidemiological research and clinical trials often exclude women or fail to include them in sufficient numbers, for example, in research on coronary heart disease (CHD) and AIDs (Sharp 1994, Doyal 1995). This lack of medical research on women means that less is known about the effects of drug or surgical treatment regimes on them.

Differences in men's and women's reported illness rates may result from diagnostic behavior and treatment by physicians, as well as gender differences in the process of seeking care—'illness behavior.' Reported gender differences in CHD may partly reflect doctors' behavior. US women are less likely than men to be diagnosed as having heart disease or to undergo coronary angiography or bypass surgery (McKinlay 1996). Thus, rates of recorded illness may differ from the 'true' prevalence in systematic and gender-related ways.

The image of CHD as predominantly affecting men is prevalent in the media, medical advertising, and among the medical profession. There is less publicity and discussion about the risk of CHD for women than men. Yet CHD is the most common cause of death for women. Similarly, there is high public concern with female-specific cancers, such as breast cancer and cervical cancer, and less publicity and research attention on other cancers such as stomach and lung cancer among women (see *Gender and Cancer*; *Gender and Cardiovascular Health*; *Gender and Health Care*).

2.3 The Emergence of Men's Health

A marked imbalance exists between the extensive literature on women's health and the paucity of research on men's health. Men's ill-health has remained a largely neglected area, with a lack of studies examining lay ideas about health and illness among men (Watson 2000), and men are often cast as a homogeneous group (Cameron and Bernardes 1998). While some conditions, such as CHD and lung cancer, have been highlighted, other areas, such as prostate problems or mental health, have been given little priority until recently.

Empirical work on ill-health conditions specific to men reveals a variety of problems stemming from the social invisibility of men's illness. In research on prostate problems, Cameron and Bernardes (1998) found that men regarded health as women's responsibility, knew little about men's health, and tended not to talk about any prostatic health problems they had experienced. Prostate problems appeared to represent a threat to the men's sense of masculine self.

In all societies, whether symptoms are perceived as illness by individuals is influenced by cultural values, gender roles, and family interactions. Health professionals also play a key role in defining presented symptoms as 'sickness,' sanctioning entry into the 'sick role,' and making decisions about treatment, all of which may be influenced by gendered norms and assumptions (see *Men's Health*; *Body Image and Gender*).

2.4 Gender Analysis of Health

A gender analysis of health sees both women's and men's health within the wider context of social and

economic relations between women and men. It recognizes the impact of gender inequalities in power and economic resources on both men's and women's health.

Many women perform the 'double shift' of housework, childcare, and paid labor, so it is important to assess how combining paid and unpaid work affects their health and varies with socioeconomic circumstances (Doyal 1995). Men's health may benefit from the unpaid labor of their wife in providing domestic services and health care; there is substantial evidence that married men have better health and health-related behavior than never married or previously married men.

Women provide the majority of care for elderly, frail, or disabled relatives (Arber and Ginn 1991). Caregiving places a major burden on carers, with consequences for their physical and mental health. The role of unpaid carers in the community is becoming increasingly significant within Western health care provision, as hospital stays become shorter, shifting the costs of care onto family members. The policy of promoting community care and de-institutionalization therefore has a gendered effect with a greater impact on women, who form the majority of carers.

Many women experience rape and domestic violence, which have adverse physical and mental health consequences, but the extent and nature of that violence varies among and within societies (Koblinsky et al. 1993, Doyal 1995). Violence is more common where men's masculinity is defined in terms of dominance, and women are relatively powerless economically and socially. Another form of culturally-sanctioned violence in some societies is female circumcision, which results in deaths, as well as pain, chronic urinary problems, and higher levels of maternal and child mortality.

3. *Gender Differences in Mortality*

In nearly all countries women outlive men. Four broad patterns of sex differences in mortality can be identified, each related to gender roles.

3.1 *Women Outlive Men by 5–7 Years*

This pattern is found in most Western countries. For example, in the UK in 1995, expectation of life for women was 79.4 years and for men 74.1, a 5.3 year difference; in the US, it was 78.9 for women and 72.5 for men, a 6.4 year difference; in Sweden, women outlive men by 5 years (expectation of life of 81.5 and 76.5, respectively), and in Japan by 6.6 years (83.6 and 77 years, respectively) (UN 1999).

The 'norm' of women having a 5–7 year longer life expectancy than men is the product of women's biological advantage and men's greater mortality from occupational hazards and risky behaviors such as smoking, drinking, dangerous sports, fast driving, and violence. As women in Western countries have entered the workforce in larger numbers over the last 30 years and the male excess in smoking is now very small, there has been a reduction in the gender disparity of mortality (Hemstrom 1999). This trend is expected to continue as the lifestyles of women and men converge and women undertake more risky health behaviors.

3.2 *Women Outlive Men by 2–4 Years*

This pattern is exemplified by European countries in the mid-nineteenth century, and is currently typical of many developing societies in Africa and Asia. For example, in Kenya the expectation of life for women is 61.4 and men is 57.5 (a 3.9 year difference), and in China is 52.9 for women and 50.1 for men (UN 1999). This modest female mortality advantage is typical of transitional societies, where the previously high maternal mortality rate is falling, but where women occupy a disadvantageous structural position with adverse effects on their health. As countries move towards greater economic development and more education for women, the sex gap in mortality tends to increase.

3.3 *No Gender Difference or Men Outlive Women*

In some developing countries, there is little gender difference or men outlive women. In Bangladesh men outlive women, with an expectation of life of 58.6 for men and 58.2 for women. In Pakistan and India the sexes have almost equal life expectancy (in Pakistan, 59.0 for men and 59.2 for women, and in India, 59 for men and 59.7 for women) (UN 1999). This pattern occurs in countries where women's social status is very low. Women are more likely to have poor nutrition, have less access to healthcare, are subject to frequent births, and have high maternal mortality (Doyal 1995).

3.4 *Women Outlive Men by Eight or More Years*

This pattern is exemplified by countries in the former Soviet Union and among blacks in the US, partly reflecting a deterioration in men's life expectancy since 1980. In Russia, expectation of life is only 58.3 for men and 71.7 for women (a 13-year difference); the corresponding figures for Latvia are 63.9 and 75.6 (UN 1999). In some Eastern European countries and the former Soviet Union, mortality rates have risen for men over the past 20 years, but fallen or remained stable for women, showing how men and women can be affected differently by social and economic changes (Hemstrom 1999). There have been greater threats to the economic and psychosocial well-being of men in these societies, many of whom have lost a secure 'bread-winning' role. The fall in life expectancy of US

black men relates to socioeconomic disadvantage, unemployment, alienation, drug abuse, and violence (Lorber 1997).

4. Explaining Gender Differences

A range of types of explanations have been put forward to account for gender differences in mortality and in health. However, there can be no simple interpretation, since women have health advantages in some aspects while men are advantaged in others.

(a) *Biological differences* associated with genetic factors, including hormonal differences, tend to advantage women. However, the range of gender differences in mortality identified in Sect. 3 clearly demonstrates the importance of factors other than biology.

(b) *Psychosocial differences* between men and women may be related to socially-created differences in personality, coping behaviors, and self-efficacy, which influence the experience and reporting of symptoms. These include gender role orientations, associated with masculinity and femininity, discussed in Sects. 1.1 and 2.3. Emslie et al. (1999) show that high 'femininity' scores are associated with poorer reported health.

(c) *Risk behaviors*. Epidemiologists are particularly concerned with gender differences in health promoting and health damaging behaviors. In most societies, men have higher levels of smoking and drinking, including more 'binge drinking.' However, in Western countries women have a better diet and participate more in preventive health measures such as screening. The growing trend for women to engage in smoking, drinking, and more risky behaviors has been termed the 'female emancipation hypothesis' (Hemstrom 1999).

(d) *Occupational and work-related factors*. Health hazards are associated with both paid work and unpaid domestic work. Working class men have traditionally been employed in more dangerous occupations with greater industrial health hazards. Much of women's paid work is repetitive in nature and lacks autonomy, which may create stress and repetitive strain injuries (Doyal 1995). In addition, unpaid domestic or family work may involve hard physical labor, such as in developing countries carrying water long distances and heavy agricultural production.

(e) *Social roles and relationships* differ between women and men and may impact on health. Women in Western countries often have better social networks than men, and can more easily rely on contact with close friends and relatives in times of crisis and stress. Marriage may provide a major source of social support, but in a different way for men and women and across societies. In terms of social support, marriage seems to be good for men but less so for women. Men tend to rely more on their wife for social support and a confiding relationship, with divorced and widowed men reporting particularly poor self-assessed and psychosocial health.

(f) *Power and resources within the home*. Patriarchal control within the family and domestic settings may lead to exploitation of women and violence against them. Women's lack of power and influence in the home and lack of access to valued resources (of food, opportunities for leisure, income) may have adverse health consequences.

(g) *Social structural differences within society*. Women in most societies are more likely than men to be poor, have less education and live in disadvantaged material circumstances, all of which are likely to be implicated in gender differences in health.

(h) *Health service related factors*. Women may have less access to health services in developing countries, but in Western countries women generally make greater use than men, both because of using health care associated with reproduction and contraception, and the gendered meanings of illness discussed in Sect. 1.1. There may also be gender differences in diagnosis and treatment, see Sect. 2.2.

5. Gender Differences in Morbidity

It has become accepted wisdom that 'women are sicker, but men die quicker' (Lahelma et al. 1999). The higher morbidity rate but lower death rate of women was characterized by Nathanson (1975) as a 'contradiction' which required explanation. Other authors in the 1970s and 1980s repeatedly demonstrated that females had higher rates of illness than males. A number of the explanations put forward related to gender-appropriate behavior, for example, that women tend to over-report morbidity more frequently than men.

There have been recent changes; in the 1990s in European countries, relatively modest gender differences in overall measures of health and well-being were found, which are less important than the previous orthodoxy suggested (Macintyre et al. 1996, Lahelma et al. 1999). Macintyre et al. (1996) examined different measures of health using various datasets and found no female excess morbidity on many health measures.

However, major gender differences in mental health remain substantial. Women are more likely to report a range of neurotic disorders, for example, anxiety and depressive disorders, although when men's higher rates of alcohol and drug dependence are included, there is a much smaller gender difference in the prevalence of psychiatric disorders. Moreover, older women are more likely than older men to suffer from physical conditions which are nonfatal but result in chronic and disabling illnesses (Lahelma et al. 1999, Annandale and Hunt 2000). In Britain, nearly twice as many women as men over 65 suffer from functional impairments sufficient to require help on a daily basis to remain living in the community (Arber and Ginn 1991). A key concern is the extent of gender difference

in 'healthy life expectancy,' that is, the period of life without disability or chronic ill-health. Because of women's higher level of disability, the gender difference in 'healthy life expectancy' is smaller than in life expectancy.

6. Structural Inequalities in Health Among Women and Men

In the late 1970s and 1980s, many researchers began to focus on differences in health *among* women. Early work examined to what extent additional roles, such as the parental role and paid employment, had beneficial or adverse consequences for women's health (Arber 1991). While a growing body of research, particularly in Britain, studied inequalities in men's health.

More recently, researchers have stressed the importance of examining women *both* in terms of their structural position within society and their family roles. This approach examines how socioeconomic circumstances, together with marital and parental roles, influence the health of women and men. The key structural variables influencing health are occupational class, employment status, and educational qualifications. The effect of women's family roles on health depends on the material resources of her household (Arber 1991, 1997).

Changes in family structure since the 1970s, particularly the growth in lone parenthood and cohabitation, are relevant to gender differences in health inequalities. The previous orthodoxy that married women have poorer health than single women no longer holds in Britain (Annandale and Hunt 2000), possibly reflecting changes in the meaning of marriage in the late twentieth century and the opening of career opportunities for married women which previously only existed for single women. Lone mothers report the poorest health, largely due to living in disadvantaged material circumstances.

A gender analysis of health requires consideration of differences in the structural position of women and men within societies, in terms of power, occupational status, income, and lone parenthood. It is therefore essential to examine the extent to which gender differences in health may be explained by women's social, income, and political disadvantages.

7. Gender and Health Across the Life Course

During the twentieth century there have been radical changes in actual and expected gender roles, as well as in norms about age-related behavior. Gender roles and responsibilities are socially constructed in ways which vary across the life course. Thus, gender differences in health are likely to reflect gender socialization as to behavior at different stages of the life course. Thus the nature of inequalities in health for men and women may vary for different age groups (Annandale and Hunt 2000).

The health and other characteristics of women and men are influenced by their prior life course. This is vividly seen in later life, since the financial well-being of older adults is intimately tied to their previous role in the labor market and private pension acquisition. A life course approach emphasizes the linkage between phases of the life course, rather than seeing each phase in isolation.

Research often neglects gender differences in health in childhood and later life, while the dominant concern is working age adults. During the working ages, family structure has a greater effect on women's than men's health. Married men and women report the best health, irrespective of whether they have children, but lone mothers report the poorest health. Class continues to be a major factor determining the health of working age British men and women. However, other structural variables are likely to become increasingly salient in the twenty-first century. Whether or not a person is in paid employment has become a key marker of health status, while educational qualifications increasingly differentiate health, especially among women.

In later life, previous occupational class is the key determinant of health for older British men and women (Arber and Cooper 1999). The similarity in the pattern of class inequalities in health for older women and men is surprising given that the majority of the current generation of older women have spent less than half their adult life in paid employment.

The nature of health inequalities between women and men is likely to differ over time and across societies, reflecting gender roles and relationships historically and cross-nationally. Taking a life course approach provides added insights into the mechanisms underlying the creation and perpetuation of gender inequalities in health.

See also: Gender and Cardiovascular Health; Gender and Health Care; Gender and Reproductive Health; Gender, Economics of; Gender Role Stress and Health; Men's Health; Reproductive Rights in Developing Nations; Women's Health

Bibliography

Annandale E, Hunt K (eds.) 2000 *Gender Inequalities in Health*. Open University Press, Buckingham, UK

Arber S 1991 Class, paid employment and family roles: Making sense of structural disadvantage, gender and health status. *Social Science and Medicine* **32**: 425–36

Arber S 1997 Comparing inequalities in women's and men's health: Britain in the1990s. *Social Science and Medicine* **44**: 773–87

Arber S, Cooper H 1999 Gender differences in health in later life: The new paradox? *Social Science and Medicine* **48**: 63–76

Arber S, Ginn J 1991 *Gender and Later Life: A Sociological Analysis of Resources and Constraints*. Sage, London

Cameron E, Bernardes J 1998 Gender and disadvantage in health: Men's health for a change. *Sociology of Health & Illness* **20**: 673–93
Doyal L 1995 *What Makes Women Sick: Gender and the Political Economy of Health*. Rutgers University Press, New Brunswick, NJ
Emslie C, Hunt K, Macintyre S 1999 Problematizing gender, work and health. *Social Science and Medicine* **48**: 33–48
Hemstrom O 1999 Explaining differential rates of mortality decline for Swedish men and women. *Social Science and Medicine* **48**: 1759–77
Koblinsky M, Timyan J, Gay J (eds.) 1993 *The Health of Women: A Global Perspective*. Westview Press, Boulder, CO
Lahelma E, Martikainen P, Rahkonen O, Silventoinen K 1999 Gender differences in ill health in Finland. *Social Science and Medicine* **48**: 7–19
Lorber J 1997 *Gender and the Social Construction of Illness*. Sage, Thousand Oaks, CA
Macintyre S, Hunt K, Sweeting H 1996 Gender differences in health: Are things really as simple as they seem? *Social Science and Medicine* **42**: 617–24
McKinlay J 1996 Some contributions from the social system to gender inequalities in heart disease. *Journal of Health and Social Behavior* **37**: 1–26
Nathanson C A 1975 Illness and the feminine role: A theoretical review. *Social Science and Medicine* **9**: 57–62
Sabo D, Gordon D F (eds.) 1995 *Men's Health and Illness: Gender, Power and the Body*. Sage, Thousand Oaks, CA
Sharp I 1994 *Coronary Heart Disease: Are Women Special?* National Heart Forum, London
United Nations 1999 *1997 Demographic Yearbook*. United Nations, New York
Watson J 2000 *Male Bodies; Health, Culture and Identity*. Open University Press, Philadelphia, PA

S. Arber

Gender and Place

'Gender and place' as an interdisciplinary field of study emerged during the 1970s. It attempts to identify ways in which gender stereotypes and spatial arrangements reinforce each other. Over its brief history, the field has become increasingly differentiated with respect to definitions of both place and gender. Spaces that acquire meaning through use or history are 'places' consisting of cities, neighborhoods, buildings, and even the human body. 'Gender' extends beyond the dichotomy of masculine and feminine to encompass gays and lesbians in a continuum of identities.

1. *Introduction*

'Gender and place' is a nuanced term for 'sex and space.' Both phrases refer to the ways in which gender or sex-role stereotypes are built into, and reinforced by, spatial arrangements. Take the adage that 'woman's place is in the home.' When ideology dictates that women should occupy only the domestic sphere, they seldom learn skills that would prepare them to enter public spaces like colleges and the workplace. Conversely, men who spend their days in corporate boardrooms rarely are expected to excel at housework and childcare. Feminine and masculine activities both shape, and are shaped by, built environments. Homes, for example, are still more likely to evoke feminine connotations than offices and factories regardless of the number of women in the labor force.

Social scientists use 'gender' to represent socially constructed masculine and feminine characteristics, while 'sex' refers to the biologically determined categories of male and female. One's sex is determined at conception, but one's gender identity develops over the life course and can fluctuate across a wide continuum of masculine and feminine characteristics. Girls who work on the family car in the garage and boys who cook meals in the kitchen are exhibiting gender characteristics incongruent with those expected of their sex; they are also occupying spaces inconsistent with gender expectations.

Place is to space as gender is to sex. Actual spaces can be designated with commonly understood boundaries, but 'places' transcend simple attempts at measurement. Places are spaces imbued with significance, just as gender adds meaning to sex. The space of a neighborhood park may be a place to play for children, a place to sleep for the homeless, and a place to walk the dog for apartment dwellers. The park can be all of these places simultaneously or sequentially. Like gender, place is socially and subjectively created from objective elements. The park may occupy a *space* the size of a city block, but it functions as a *place* for fun, shelter, or exercise depending on one's construction of its purposes.

Space and place are quite different concepts. Space is abstract and limitless, while place is bounded and refers to a setting infused with identity and memory. Another way to think of space is as a location or site of activity. 'Location' is objective, as when real estate agents identify a house by its distance from an elementary school. Place is subjective. One neighborhood close to a school might be described as a 'good place to raise children,' while another neighborhood, equally close, fails to receive the same endorsement. Such different assessments could result from the type of housing or size of yards, but they all affect the 'feel' of the place. In short, space and location are analyzed and measured; place is experienced.

The analysis of gender and place is interdisciplinary in that the field seeks to combine the spatial orientation of geographers, architects, and urban planners with the social focus of anthropologists, sociologists, and historians to understand the ways space and gender mutually reinforce one another. A brief review of the literature reveals that its focus has grown steadily more differentiated. Places range from the geographic

scale of the city or neighborhood to the human scale of the body. Gender identities include masculine, feminine, gay, lesbian, and an ambiguous 'third gender' of queer that recognizes gender identity as variable both historically and individually over the life course (Herdt 1996, Ingram et al. 1997, Smith and Windes 1999).

2. *Origins of the Field*

One of the first social scientists to address the social construction of space was Goffman (1959), who used the theater analogy of 'front-stage' and 'back-stage' behavior to describe the spatial nature of formal and informal activities. Harvey (1973), a geographer, proposed that urban spaces and social relations are so intricately linked that a new language should be invented to capture the simultaneous occurrence of spatial form and social processes. Harvey's emphasis on the city (and the fact that the majority of American life occurs in metropolitan areas) may have directed subsequent investigations of gender and place toward its current urban focus.

Contemporary intellectual interest in the issues surrounding gendered aspects of space arose during the 1970s. When feminist political activism was at its height, women were, quite literally, entering spaces previously open only to men. Colleges, factories, and offices became increasingly integrated by sex as women's place expanded beyond the home. Young women attending, and teaching at, universities during that era were among those who challenged the boundaries of traditional scholarship. The literal and symbolic collision of place and gender soon became a conscious subject of study.

It is not surprising, then, that the first publications on gender and place were edited volumes generated from academic conferences about sex and space. Four of the most important volumes appeared almost simultaneously. *New Space for Women*, edited by Wekerle et al. (1980) emerged after the 1976 United Nations Conference on Human Settlements met in Vancouver. The editors organized a session addressing the role of women in the design and planning of urban settlements because they recognized that women's problems with housing, transportation, urban planning, and lack of access to environmental decisions transcended categories of developed and developing nations. The next year, British anthropologist Ardener published *Women and Space: Ground Rules and Social Maps* (1981) as part of a series based on lectures sponsored by the Oxford University Women's Studies Committee. *Women and Space* included ethnographies of sex differences in the use of domestic and public spaces in the Andes, Greece, Iran, Russia, Nigeria, and South Africa.

In the United States, the Lincoln Institute of Land Policy sponsored a conference that resulted in the publication of *Building for Women* (Heath 1981) edited by sociologist Keller. It dealt with housing, community, and open space from a variety of academic disciplines; one of the chapters summarized the US Department of Housing and Urban Development's Women and Mortgage Credit Project to facilitate women's ability to buy their own homes (Shalala and McGeorge 1981). HUD played a role in producing the fourth volume on women and space when it awarded a research grant to *SIGNS: Journal of Women in Culture* to commission essays for a special issue titled *Women and the American City*.The editor of *SIGNS*, Catherine Stimpson, subsequently published the volume (with Elsa Dixler, Martha Nelson, and Kathryn Yatrakis) through University of Chicago Press in 1981. Contributors constituted the pantheon of researchers on women and space: Galen Cranz, Dolores Hayden, Jacqueline Leavitt, Susan Saegert, and Gerda Wekerle.

Donna Shalala, Secretary of Health and Human Services during President Clinton's administration, worked behind the scenes to support research on gender and place. Shalala was at HUD when the Mortgage Credit Project began and when *SIGNS* received funding for its special issue. As president of Hunter College, Shalala co-chaired the Ford Foundation–Hunter College Women and Housing Seminar with urban planning professor Eugenie Ladner Birch. Members of the seminar investigated the housing needs of single mothers, the elderly, and working women from the standpoint of design and affordability. The fruits of that pedagogical venture resulted in *The Unsheltered Woman: Women and Housing in the 1980s*, edited by Eugenie Ladner Birch (Center for Urban Policy Research 1985), one of the first books to alert urban planners to the gendered nature of the housing market.

Milroy and Andrews strengthened the tradition of edited volumes with *Life Spaces: Gender, Household, Employment* (1988), a volume written entirely by Canadians about Canada. Milroy and Andrews introduced American audiences to geographer Suzanne Mackenzie, a founder of the field of feminist geography in the 1970s. Mackenzie, in turn, linked Canadian and American authors to British geographers Jo Foord, Sophie Bowlby, and Linda McDowell. *Life Spaces* departed from the model of previous edited volumes in one important respect: Wekerle's chapter on women's housing cooperatives included floor plans and photographs. Wekerle was one of the few social scientists to illustrate spatial concepts on the written page. During the first wave of scholarship, only Hayden and Birch (an architect and an urban planner, respectively) provided images of their topics. Today it seems obvious that books about space should include drawings and photographs, yet at the time authors rarely incorporated them.

Among pioneering scholars of gender and place, Dolores Hayden deserves special mention because she approaches the topic from an architectural rather than

a social science perspective. Hayden's *The Grand Domestic Revolution: A History of Feminist Designs for American Homes, Neighborhoods, and Cities* (1981) unearthed the work of nineteenth century 'material feminists' who established collective solutions to household labor and demanded pay for their work. Her next book, *Redesigning the American Dream: The Future of Housing, Work, and Family Life* (1984), addressed the disparity between the massive housing stock constructed after World War II and changing household demographics. It opened debate among architects and planners about appropriate designs for new American families. Franck and Ahrentzen's (1991) edited volume on collective, single parent, and single-room occupancy housing was a direct response to Hayden's challenge.

3. *The Place Pyramid: Boroughs, Buildings, and Bodies*

The first stage of research on gender and place addressed sex and space, as the titles of books published in the early 1980s suggest. They all included the word 'woman' and reference to an actual space such as the house or the city. Research focused on women as active participants in, and shapers of, the built environment traditionally considered men's domain. For example, Cranz (1981) chronicled women as users and suppliers of urban park services. This is what some feminist scholars call the 'add women and stir' phase; it restored women's presence for the historical record. This stage also focused on a fairly large *geographic scale*: that of cities and neighborhoods. For alliterative purposes that will become clear, I call this the *borough* stage.

During the late 1980s and early 1990s, focus expanded from the geographic scale of boroughs to include the *architectural scale* of *buildings*. Urban and neighborhood analyses were not abandoned. Rather, they were seen as the larger context in which the design of buildings reinforced stereotypical gender associations. Weisman (1992), an architect, for example, analyzed the sexual symbolism associated with specific buildings—skyscrapers, department stores, public housing—and advocated the redesign of large swaths of urban architecture to take women's needs and preferences into account. Betsky (1995) produced an architectural critique of the gap between the masculine characteristics of the exterior built environment and the feminine traits of its interiors.

A significant British book illustrating the place pyramid was Moore's *Space, Text, and Gender: An Anthropological Study of the Marakwet of Kenya* (1986). Moore drew on the theories of Foucault (1972) to analyze household space as a 'cultural text' representing gender relations in Marakwet society. The study placed the Marakwet in their geographic context with a map of Kenya and a diagram of their Sibou village, and in their building context with photographs of their dwellings. Moore's work progressed from a base of great geographical expanse to the smaller scale of the building. *Space*, *Text*, *and Gender* marked another important milestone. It was one of the first full-length ethnographies devoted to the topic of gender and place. Although edited volumes would still fill a niche, the field acquired stronger intellectual credentials with the publication of books posing theoretical arguments.

The author's own work sought to extend that tradition. Heavily influenced by previous research and theory, *Gendered Spaces* (Spain 1992) was a cross-cultural and historical study proposing that women's status is lower in societies in which men and women are separated spatially in homes, schools, and workplaces. An architectural perspective on gender and place also appeared in 1992 with *Sexuality and Space* edited by Colomina (1992). Chapters by Colomina, Lynn Spigel, and Mark Wigley dealt with the home, while one by Elizabeth Grosz examined the relationship between the body and the city. Drawing from Foucault's (1977, 1979) emphasis on bodily pleasure and pain, Grosz introduced a concept developed later by Sennett in *Flesh and Stone: The Body and the City in Western Civilization*. Sennett's (1994, p. 370) thesis, that 'urban spaces take form largely from the ways people experience their own bodies,' set the stage for subsequent emphasis on the 'body as place.'

Nowhere has the constriction of the definition of place—from the large city to the small body—been more evident than in geography. Initially, geographers worked on topics like the sexual division of space in cities and homes (McDowell 1983) and the public/private locations of production and reproduction (Mackenzie and Rose 1983, Massey 1984). More recently, however, geographers have adapted Foucault's (1977) perspective on the body as text. They write articles on *The Body and Geography* (Longhurst 1995) and on the male torso as landscape (Nash 1996). Urban spaces that were once the focus of geographic analysis have become the stages on which bodies display different degrees of fear of violence (Koskela 1997, Mehta and Bondi 1999). McDowell (1999) legitimized this move toward the body by including it among the themes currently qualifying as feminist geography (along with place, home, community, city, nation–state, public space, and travel).

4. *Current Theory and Research: The Expansion of Gender*

As the concept of place has been expanded to include a larger range of sizes, so too has the definition of gender been stretched beyond its original association with the dichotomy of biologically determined sex. This enlargement inevitably introduced homosexuality into descriptions of the relationships between

(heterosexual) men, women, and space. For example, the original four edited volumes reviewed in this essay contained not a single chapter on homosexual gender identities. In contrast, the current literature contains numerous references to gays, lesbians, queers, bisexuals, and transsexual and transgendered persons (those born one sex who identify themselves as the other sex through anatomical (transsexual) or behavioral (transgendered) modifications). Gender now reflects a continuum of identities.

Castells's mapping of San Francisco's gay male culture was one of the first explorations into the relationship between homosexual subculture and space. Castells proposed that spatial concentration is fundamental to the development of gay identity. He argued that gay men are more likely to establish urban neighborhoods than lesbian women because men's tendency toward dominance extends to territorial aspirations, while lesbians' emphasis on relationships transcends space, making lesbians 'placeless' and more politically radical (Castells 1983, p. 140, Castells and Murphy 1982; see Forsyth (1997) for a challenge to this interpretation).

Researchers have analyzed how gays and lesbians have shaped residential, commercial, and public spaces. Gay men, and to some extent lesbian women, have played significant roles in the revitalization of American urban neighborhoods (Forsyth 1997, Lauria and Knopp 1985). Commercial areas catering to gays and lesbians, such as San Francisco's Market Street and Chicago's Halsted Street, display highly visible rainbow flags to attract clientele and, increasingly, tourists. Gay and lesbian community centers, bookstores, and health clinics have become staples in many cities; gay and lesbian parades and festivals temporarily transform city streets (Forsyth 2000). Pressure is mounting to nominate prominent gay and lesbian sites as historic landmarks, such as the New York City gay bar (the Stonewall Inn) where a police raid sparked a riot (Dubrow 1998). Such places have become easier to identify as gay and lesbian histories with a spatial focus have emerged. Chauncey (1994), for example, has identified how gay men appropriated public and private spaces to provide settings for the expression of gay culture in New York City between 1890 and 1940, while Kennedy and Davis (1993) have mapped the landscape of lesbian meeting places in Buffalo, NY between the 1930s and 1960s.

What does the future hold for the study of gender and place? It will definitely involve cyberspace, as the literature already suggests (Boyer 1999, Light 1995). As the matrix of cyberspace supplements the grid of the city, women may experience the environment differently from men. Possibilities for gender ambiguities abound on the Internet in 'chat rooms' that allow more anonymity than face-to-face interactions. The digital landscape has the potential to change women's status if it changes the way knowledge is disseminated. The gendered spaces that once limited women's access to knowledge, and thus power, may dissolve.

A fitting end for this article is reference to a feminist critique of historic preservation titled *Susan B. Anthony Slept Here: A Guide to American Women's Landmarks* by Sherr and Kazickas (1994). Everyone knows that George Washington slept (and owned slaves) at his Mt. Vernon plantation, but few people know where Susan B. Anthony, leader of the movement for women's right to vote, worked or slept. The ironic title captures the trend in the literature toward greater gender and place ambiguity. The first wave of scholarship would have highlighted the gendered nature of the town where Susan B. Anthony lived. The second wave would have analyzed the home she slept in, while the third wave would ask with whom, if anyone, she was sleeping. If Susan B. Anthony were alive today, the fourth wave might explore whether it matters if her body occupies a particular physical space at all.

See also: Economic Development and Women; Economic Globalization and Gender; Gender and Environment; Gender and Feminist Studies in Sociology; Land Rights and Gender; Place in Geography; Spatial Equity; Spatial Interaction; Spatial Labor Markets; Transition, Economics of

Bibliography

Betsky A 1995 Building Sex: Men, Women, Architecture, and the Construction of Sexuality. William Morrow, New York

Boyer C 1999 Crossing cybercities: Urban regions and the cyberspace matrix. In: Beauregard R, Body-Gendrot S (eds.) *The Urban Moment: Cosmopolitan Essays on the Late-20th-Century City*. Sage, Thousand Oaks, CA

Castells M 1983 *The City and the Grassroots*. University of California Press, Berkeley, CA

Castells M, Murphy K 1982 Cultural identity and urban structure: The spatial organization of San Francisco's gay community. In: Fainstein N, Fainstein S (eds.) *Urban Policy under Capitalism*. Sage, Beverly Hills, CA

Chauncey G 1994 *Gay New York: Gender, Urban Culture, and the Making of the Gay Male World, 1890–1940*. Basic Books, New York

Cranz G 1981 Women and urban parks: Their roles as users and suppliers of park services. In: Keller S (ed.) *Building for Women*. Heath, Lexington, MA

Dubrow G 1998 Blazing trails with pink triangles and rainbow flags: New directions in the preservation and interpretation of gay and lesbian heritage. *Historic Preservation Forum* **12**: 31–44

Forsyth A 1997 'Out' in the valley. *International Journal of Urban and Regional Research* **21**: 36–60

Forsyth A 2000 Sexuality and space: Gays, lesbians, queers, and planning practice. Unpublished

Foucault M 1972 *The Archeology of Knowledge* (trans. Sheridan A). Tavistock, London

Foucault M 1977 *Discipline and Punish* (trans. Sheridan A). Allen Lane, London

Foucault M 1979 *The History of Sexuality*. Allen Lane, London

Franck K, Ahrentzen S 1991 *New Households, New Housing*. Van Nostrand Reinhold, New York
Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, New York
Harvey D 1973 *Social Justice and the City*. Johns Hopkins University Press, Baltimore, MD
Herdt G (ed.) 1996 *Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History*. Zone, New York
Ingram G B, Bouthillette A-M, Retter Y (eds.) 1997 *Queers in Space: Communities/Public Places/Sites of Resistance*. Bay Press, Seattle, WA
Kennedy E, Davis M 1993 *Boots of Leather, Slippers of Gold: The History of a Lesbian Community*. Penguin, New York
Koskela H 1997 'Bold walk and breakings': Women's spatial confidence. *Gender, Place and Culture* **4**: 301–19
Lauria M, Knopp L 1985 Toward an analysis of the role of gay communities in the urban renaissance. *Urban Geography* **6**: 152–69
Light J S 1995 The digital landscape: New space for women *Gender, Place and Culture* **2**: 133–46
Longhurst R 1995 The body and geography. *Gender, Place and Culture* **2**: 97–105
Massey D 1984 *Spatial Division of Labor: Social Structures and the Geography of Production*. Macmillan, London
McDowell L 1983 City and home: Urban housing and the sexual division of space. In: Evans M, Ungerson C (eds.) *Sexual Divisions: Patterns and Processes*. Tavistock, London
McDowell L 1999 *Gender, Identity, and Place: Understanding Feminist Geographies*. Polity Press, Cambridge, MA
MacKenzie S, Rose D 1983 Industrial change, the domestic economy, and home life. In: Anderson J, Duncan S, Hudson R (eds.) *Redundant Spaces in Cities and Regions?* Academic Press, London
Mehta A, Bondi L 1999 Embodied discourse: On gender and fear of violence. *Gender, Place and Culture* **6**: 67–84
Nash C 1996 Reclaiming vision: Looking at landscape and the body. *Gender, Place and Culture* **3**: 149–69
Smith R R, Windes R R 1999 Identity in political context: Lesbian/gay representation in the public sphere. *Journal of Homosexuality* **37**: 25–45
Weisman L K 1992 *Discrimination by Design: A Feminist Critique of the Man-made Environment*. University of Illinois Press, Urbana, IL

D. Spain

Gender and Reproductive Health

With the world's population increasing at an alarming rate, reproductive health has become the focus of growing attention in recent years. While reproductive concerns and activities also influence men's health, both scientific studies and health policies primarily focus on the female gender. Because women become pregnant and can bear children, the risks of reproduction present a greater threat to their health.

1. *Definition of Reproductive Health*

Although research has long examined how human reproduction affects overall health, the concept of 'reproductive health' is relatively new. In many industrialized countries, a clinical view continues to dominate, with reproductive health seen as a specialty of obstetrics and gynecology. Characteristic of this clinical view is its focus on negative outcomes and the separation of services dealing with reproductive issues (Graham 1998). Recent years, however, have seen the rise of a more comprehensive concept of reproductive health, formulated in 1994 at the International Conference on Population and Development (ICPD) in Cairo as follows: 'Reproductive health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity, in all matters relating to the reproductive system and to its functions and processes. Reproductive health therefore implies that people are able to have a satisfying and safe sex life and that they have the capability to reproduce and the freedom to decide if, when and how often to do so' (Program of Action 1995, p. 202).

As this definition makes clear, sexual health is considered as a central feature of reproductive health. In line with this very broad definition, the scientific literature discusses a large range of reproductive concerns, with a preponderance of reproductive problems and risks (Graham 1998). The most well-known health risks associated with reproduction are sexually transmitted diseases (STDs), including HIV/AIDS, maternal mortality, unsafe abortion, adolescent pregnancy, and infertility. Over the last several years, the comprehensive ICPD definition of reproductive health has come to include additional aspects such as female genital mutilation and violence against women.

2. *Major Reproductive Health Risks*

Reproductive heath risks or health crises are the source of high morbidity and mortality rates worldwide. Most are avoidable. The prevalence of reproductive risks or 'outcomes' varies enormously among populations throughout the world. The greatest differences in negative outcomes exist between developed and developing countries. At the same time, the clear improvement in the reproductive health of women in developed countries is a relatively recent achievement and can be traced back primarily to key medical advances (introduction of antiseptics, discovery of sulfonamide, modern obstetrics), as well as changes in gender roles. The significant reproductive risks women in Europe and the United States continued to face even until the turn of this century are described by Shorter in *A History of Women's Bodies* (1982).

The following section will outline the spread of some major reproductive health crises. Although not all concerns (such as infertility) can be presented in detail, attention will be given to the gravest dangers in terms of individual and public health (see Miller and Rosenfield 1996). Despite decreasing fertility rates in many parts of the world, it has not been possible to reduce *maternal mortality*. According to WHO statistics, every year 585,000 women die as a result of

pregnancy complications. More than 98 percent of these deaths occur in the developing world (Spicehandler 1997). In 1991, WHO calculated the incidence of maternal death for a woman in North America at 1:4,000, while her counterpart in West Africa faced a risk of 1:18. Both high overall fertility rates and the high rate of adolescent pregnancies in less developed countries contribute to the high incidence of maternal mortality. Compared with women aged 20–24, the risk of maternal death for women between the ages of 10 and 14 is five times greater, and between the ages of 15 and 19 twice as high (Miller and Rosenfield 1996).

According to WHO estimates, about every one in four pregnancies is intentionally terminated. Of an estimated 45 million abortions performed each year, almost half (ca. 20 million) are performed under unsafe and septic conditions. Most of these risky abortions, in turn, occur in developing countries. Illegal and unsafe abortions are responsible for another spectrum of morbidity (for example, cervical trauma, pelvic peritonitis, jaundice, chronic pelvic pain); if worst comes to worst, these cases end fatally (an estimated 50,000–100,000 deaths each year). Here one of the most important determinants for mortality risk is the illegality of the abortion (Miller and Rosenfield 1996). But also the high rate of legal pregnancy termination, for example in Central and Eastern European countries (until the end of the 1990s, abortion was the most common form of birth control), is problematic and has serious consequences for women's reproductive health. Short-term and longer-term repercussions of repeated abortions or abortions performed under unfavorable conditions in Eastern Europe include higher incidences of anemia, infertility, and premature delivery (Bruyniks 1994).

The prevalence of STDs has serious consequences for the health of women and their unborn children. The WHO estimates that more than 330 million cases of treatable STDs (such as chlamydia infection, gonorrhea, syphilis) occur each year, the vast majority in the developing world, particularly in sub-Saharan Africa (Drennan 1998). There are two reasons why STDs in general are understood as more serious for women than for men (Drennan 1998, Miller and Rosenfield 1996, Program of Action 1995): (a) during vaginal intercourse, the likelihood of infection from man to woman is higher than from woman to man, and (b) STDs in women are often not diagnosed and treated until the disease is in a later (and thus more serious) stage because in women the symptoms are often less conspicuous in their early stages (see Sexually transmitted diseases).

Among STDs, HIV and AIDS are particularly threatening; at the end of 2000, the HIV virus had infected almost 36.1 million people worldwide, about 18.3 million men and 16.4 million women as well as about 1.4 million children (UNAIDS/WHO 2000); the majority of these cases involve people in developing countries. In some parts of the world, the AIDS epidemic has already dramatically shortened the average life expectancy (Drennan 1998). While HIV and AIDS initially presented more of a threat to men than to women, there are now countries, including the United States and several sub-Saharan African nations, in which HIV is spreading faster among women than among men.

3. *Determinants of Reproductive Risk Behavior*

A range of factors influences reproductive health. Particularly important is the quality of medical care during pregnancy and birth. The main reason for the high reproductive morbidity and mortality rates in developing countries is found in inadequate health-care structures. Inadequate contraceptive behavior (or the lack thereof) is also responsible for a variety of risks to reproductive health (in particular unintended pregnancy and STDs); for this reason, special attention is paid to risky sexual behavior in connection with reproductive health (see *Sexual Risk Behaviors*).

Contraceptive behavior is influenced by a variety of internal and external factors, summarized in a model designed by Bruyniks (1994). Here knowledge of effective contraception and attitudes toward various methods of contraception are postulated as essential determinants of contraceptive behavior. Knowledge and attitudes are influenced by internal factors (age, intelligence, moral values) and external factors (access, social environment). It is precisely in less developed countries that access to contraceptive services is frequently blocked by financial, geographic, legal, or other barriers. Psychosocial acceptance or rejection of certain contraceptive methods in the social environment also plays a decisive role (Miller and Rosenfield 1996). Another issue is whether knowledge and positive attitudes toward effective contraception are in fact reflected in reproductive behavior, and depends on such personality variables as assertiveness or self-efficacy (Schwarzer and Fuchs 1995), as well as external barriers such as financial dependence (Wyatt 1994) or physical violence.

The social status of women is an essential factor in determining their contraceptive behavior. Social status influences not only the knowledge of efficient contraceptive methods and attitudes toward contraception, but also the possibilities for women to prevail in decisions for (or against) sexual activity and contraception, if need be against their male partner's wishes. Social status is a multidimensional concept; in terms of reproductive health, the educational level of women is the central and most frequently studied dimension. Also relevant are aspects such as political participation, employment and earnings, economic autonomy, as well as the reproductive rights of women (Kawachi et al. 1999). Williamson and Boehmer (1997) analyzed different status variables as predictors of

female life expectancy in 97 less developed countries. They found a positive correlation between the status of women and female life expectancy. Again, aspects of women's status apart from their educational level were important for determining female life expectancy. Of these other dimensions of women's status, reproductive autonomy had a particularly strong effect. This reproductive autonomy is operationalized via contraceptive prevalence as an indicator of women's control over their reproductive behavior.

The influence of social status on the (reproductive) health of women can also be found in developed countries. One recent analysis examined how the morbidity and mortality of women in 50 American states is related to different indicators of social status (Kawachi et al. 1999). Higher political participation was associated not only with lower overall female mortality but also with lower rates of death from different causes, including death due to reproductive risks such as cervical cancer or infant mortality. This analysis also examined an index for 'reproductive rights' ('reflecting reproductive well-being and autonomy') in connection with indicators of health. The index was operationalized in eight legal and political indicators reflecting the reproductive well-being and reproductive autonomy of women (for example, access to abortion services without a waiting period, public funding of infertility treatments, existence of a maternity stay law, whether gay or lesbian couples can adopt, etc.). For this indicator, however, no relationship to mortality was established, although a minor correlation to self-reported illness was found ('self-reported days of activity limitations').

Along with social status, *gender roles* have an important influence on reproductive decision-making and behavior. Gender roles refer to 'normative expectations about the division of labor between the sexes and to gender-related rules about social interactions that exist within a particular cultural–historical context' (Spence et al. 1985, p. 150). Traditional gender roles, which assign women the subordinate and dependent role in the partnership, are regarded as a particular risk for reproductive health. As Drennan (1998) shows in a detailed overview, in many developing countries men are the primary decision-makers about sexual activity, fertility, and contraceptive use. It is assumed that women who are not permitted to reject sexual intercourse or to decide about contraception have little means to influence their reproductive health. Cultural norms about femininity or masculinity might also harbor reproductive risks. In many Latin American or Caribbean countries, for example, promiscuous behavior among men is regarded as proof of their masculinity (Drennan 1998). The risky sexual behavior of their (marriage) partners frequently endangers women by increasing such risks as STD transmission. In many countries, recently married women continue to face strong pressure to 'prove' their fertility; contraception is rarely practiced, and even adolescent or very young women become pregnant with the well-known attendant risks to the health of mother and child. In India, for example, 40 percent of all women between the ages of 15 and 19 are already married, and of these only 7 percent use contraceptives (Jejeebhoy 1998).

Gender inequality is expressed in various ways, including fewer educational opportunities for women, violence against women, female genital mutilation, and economic inequities (Miller and Rosenfield 1996). Some examples of gender inequality are concentrated on less developed countries, while others, such as violence against women and economic inequalities, are also found in developed countries. Gender roles are not independent of women's social status: improvements in social status, especially better education, generally go hand in hand with changes in traditional gender roles.

4. *Promoting Reproductive Health*

The ICPD Program of Action contains wide-ranging recommendations and calls for improving the reproductive health of women, especially in developing countries. What is innovative about this program in comparison to earlier approaches is that efforts to achieve this goal do not rely solely on isolated measures such as better access to contraceptive services. Rather, the program serves as a reminder that the essential requisite to bettering reproductive health is a fundamental improvement in the social status and rights of girls and women. The ICPD Program of Action lists the following points:

(a) Access to appropriate health care services: prenatal and postnatal care, obstetrics and emergency obstetrics care, treatment of STDs, abortion by qualified medical practitioners.

(b) Availability of contraceptive services: access to safe, effective, affordable, and acceptable methods of family planning.

(c) Sex education, including education for the adolescent population.

(d) Strengthening the legal, economic, and social position of girls and women.

The recommended measures of the IPCD conference continued to direct the greatest share of attention to women as the major agents of change. In the meantime, focus has shifted to include the male partner since the critical behavior unit is the sexually active couple (Becker and Robinson 1998). Overviews of research in this area (for example Becker 1996) show that reproductive health interventions (especially family planning) targeted at couples are more effective than those targeted at one or the other sex. The pair approach to reproductive health does, however, presume a certain level of equality between man and woman. In countries where very patriarchal structures continue to predominate, it is sometimes necessary to

provide initial separate services and consultations to men and women. In their report on an intervention program in Bangladesh, for example, Becker and Robinson relate how the male villagers first had to be convinced to 'allow' their wives to use contraceptive devices (Becker and Robinson 1998).

Improving the couple's communication is an essential way to increase men's participation and to promote reproductive health (Drennan 1998). In many less developed countries, only a minority of spouses talk about issues like sexuality and family planning. Demographic and Health Surveys in Nigeria, Senegal, and Pakistan conducted in the 1990s, for example, report that only one-quarter of men currently married had discussed family planning with their wives. The percentages in Cameroon (31 percent) and Ghana (37 percent) were not much higher. Drennan (1998) discusses social and cultural factors that obstruct couple communication. It is taboo in many societies to talk about sexuality. In addition, the inferior status of women often found in traditional societies continues to aggravate communication.

Many recent approaches to promoting reproductive health and preventing unintended pregnancies and STDs/HIV share a perceptible shift from models of behavioral change centered on the individual to attempts to include the partner and to consider the social network. This development can be traced back to the increasing awareness that sexual/reproductive behavior is strongly determined by social norms (Bond et al. 1999).

See also: Family and Gender; Fatherhood; Rape and Sexual Coercion; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sexual Risk Behaviors; Sexuality and Gender; Sexually Transmitted Diseases: Psychosocial Aspects

Bibliography

Becker S 1996 Couples and reproductive health: A review of couple studies. *Studies in Family Planning* **27**: 291–306

Becker S, Robinson J C 1998 Reproductive health care: Services oriented to couples. *International Journal of Gynecology and Obstetrics* **61**: 275–81

Bond K C, Valente T W, Kendall C 1999 Social network influences on reproductive health behaviors in urban northern Thailand. *Social Science and Medicine* **49**: 1599–614

Bruyniks N P 1994 Reproductive health in Central and Eastern Europe: Priorities and needs. *Patient Education and Counseling* **23**: 203–15

Drennan M 1998 Reproductive health. New perspectives on men's participation. *Population Reports Journal* **46**: 1–35

Graham W J 1998 Outcomes and effectiveness in reproductive health. *Social Science & Medicine* **47**: 1925–936

Jejeebhoy S J 1998 Adolescent sexual and reproductive behavior: A review of the evidence from India. *Social Science & Medicine* **46**: 1275–2190

Kawachi I, Kennedy B P, Gupta V, Prothrow-Stith D 1999 Women's status and the health of women and men: A view from the States. *Social Science & Medicine* **48**(1): 21–32

Miller K, Rosenfield A 1996 Population and women's reproductive health: An international perspective. *Annual Review of Public Health* **17**: 359–82

Program of Action of the 1994 International Conference on Population and Development 1995. *Population and Development Review* **21**: 187–213

Schwarzer R, Fuchs R 1995 Changing risk behaviors and adopting health behaviors: The role of self-efficacy beliefs. In: Bandura A (ed.) *Self-Efficacy in Changing Societies*. Cambridge University Press, New York

Shorter E 1982 *A History of Women's Bodies*. Basic Books, New York

Spence J T, Deaux K, Helmreich R L 1985 Sex roles in contemporary American society. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*, Vol. II, 3rd edn. Random House, New York

Spicehandler J 1997 Issues to consider in operationalizing reproductive health. *Advances in Contraception* **13**: 345–50

UNAIDS/WHO 2000 *AIDS Epidemic Update: December 2000*. UNAIDS/00.44E-WHO/CDS/CSR/EDC/2000.9.http://www. unaids.org

Williamson J B, Boehmer U 1997 Female life expectancy, gender stratification, health status, and level of economic development: A cross-national study of less developed countries. *Social Science and Medicine* **45**: 305–17

Wyatt G E 1994 The sociocultural relevance of sex research: Challenges for the 1990s and beyond. *American Psychologist* **49**: 748–54

M. Sieverding

Gender and School Learning: Mathematics and Science

Scholarship related to gender, mathematics, and science has been extensive since the early 1970s and has accounted for a sizeable percentage of the published research in the fields of mathematics and science education. In response to concerns about the small numbers of females in upper level mathematics and science courses, the research started with the documentation of gender differences in participation and achievement, moved to the identification of related variables, and more recently has included the use of feminist and cognitive science research paradigms. A brief synthesis of the literature will set the stage for a consideration of the latest research in the field.

1. Documenting Gender Differences

1.1 Differences in Learning

It is impossible to draw firm conclusions about anything as global as overall gender differences in

mathematics and science learning. Data have been reported about a variety of populations, and measured in a variety of ways in diverse situations. The magnitude of gender differences reported varies depending on the content and/or form of the particular instrument used to assess learning. When open-ended forms of assessment are used, females tend to get higher marks while males tend to be more successful on multiple choice tests where time is a factor. For example, in Australia, gender differences in favor of males have often been reported on a multiple choice test of physics. But when the Australian state of Victoria changed the secondary school physics assessment to include extended response items and social context questions, females outperformed males (Hildebrand 1995). Females outperform males when course grades given by teachers are used as an indicator of learning (Kimball 1989).

The type of content measured also has an influence on gender differences reported. Females appear to perform as well as males do on overall measures of mathematical knowledge, but males have consistently outperformed girls on measures of complex problem solving (Hyde et al. 1990). In the Third International Mathematics and Science Study (TIMSS) study (Beaton et al. 1996), middle school females and males scored approximately the same on their test of mathematics achievement, but females scored significantly lower on the test of scientific achievement. In science, girls tend to do as well as boys in biology but far less well in physics and chemistry.

The diversity of the population being measured affects the findings. Some have suggested that gender differences are more prevalent in white populations, citing examples of populations where females outperform males at all levels. While there tended to be gender differences in favor of males in most of the participating countries of the TIMSS at the secondary level, the size of gender differences varied tremendously from country to country. It should be noted that for this portion of the study only European and American countries, predominantly composed of white citizens, participated.

To increase the complexity, the magnitude of gender differences over time may have changed. In the United States, there was some evidence of pervasive gender differences in favor of males in the 1970s (Fennema and Sherman 1977). By the 1990s, Kahle and Meece (1994) suggest that there was some evidence that gender differences in overall mathematics achievement had declined in the USA while in science they had increased. To summarize, in general males and females perform differently on measures of achievement in mathematics and science. But, while males tend to score higher on more complex measures of mathematics and science achievement, it is impossible to generalize about overall gender differences in mathematics and science learning across type of test, content, and population.

1.2 Differences in Participation

The data about gender differences in participation in mathematics and science is clearer than that regarding student learning. Females constitute diminishing proportions of the students in mathematics and science classes when these subjects become elective. In all of the countries participating in the TIMSS, with the exception of the Czech Republic and Austria, fewer females reported taking physics and advanced mathematics in secondary school than did males. The problem becomes more severe in postsecondary institutions where females constitute a small percentage of majors in engineering, physics, chemistry, and computer sciences. The differences are even more accentuated in adult life where women are dramatically underrepresented in the professional fields of engineering and information technology.

1.3 Gender Differences and Attitudes

Many studies have investigated several well-defined attitudes as they relate to mathematics and science as well as the relationship of these attitudes to learning and participation in mathematics/science. When compared to males, females have consistently reported lower levels of self-confidence in doing mathematics and science, as well as lower levels of perceived usefulness of mathematics and science. Results also have indicated consistently that males tend to report that mathematics is a male domain more than females have, while both males and females consider science to be a male domain. To illustrate this perception, students around the world typically depict a male when asked to draw a scientist. Girls tend to attribute their success to hard work rather than ability. They view their failures as evidence of their lack of ability in contrast to boys who attribute their success to high ability and their failures to bad luck or lack of effort. While a great deal of research has documented these differences in attitude and attributional style, little has been done to demonstrate if and how this affects the learning of mathematics and science (Leder 1992).

2. Understanding the Development of Gender Differences: Studies of Classroom Environments and Teachers

Although females and males sit in the same classrooms, they have different experiences in those classrooms. Researchers have observed that teachers, almost without exception, spend less time with girls than with boys. This finding has been replicated around the world in a variety of studies in both mathematics and science classrooms. Researchers have gone beyond just looking for the number of interactions with girls and boys and have studied the

types of teacher/pupil interactions. They have reported that teachers tend to ask females more routine questions for which they only have a short time to reply, while males are asked more complex questions and are given more time and encouragement to respond. Males tend to dominate the use of laboratory equipment in science classrooms and to take control of cooperative groups (Leder 1992, Kahle and Meece 1994).

While the classroom experiences of boys and girls differ, researchers have not been successful at measuring the correlation of these differences with learning. One classroom variable that has accounted for gender differences involves competition. Girls appear to learn more when classrooms emphasize cooperative activities and boys appear to learn more when classrooms emphasize competitive activities. Concern about the disparate experiences of boys and girls in coeducational settings has led some to advocate single-sex classrooms for mathematics and science. There is no evidence to indicate that girls' achievement in mathematics is significantly improved by being placed in single-sex classrooms (Leder and Forgasz 1998).

The attitudes and beliefs of teachers about their male and female students have also been studied. Teachers appear to be more aware of their male students than their female students. Teachers attribute the success of their male students to ability and the success of their female students to hard work, and hold many of the same perceptions as their students by viewing mathematics and science as masculine subjects. Based on the findings of the research on differences in attitude and classroom experiences, many educational intervention programs have been developed and implemented in various levels of schooling. While these have resulted in local change, the persistence of gender differences indicates that new perspectives on research in the field are needed to affect wide scale change.

3. *Scholarship in the 1990s*

During the 1990s, many changes have affected scholarship in the field. Feminist perspectives have become prominent. The field of cognitive science has matured, and mathematics and science pedagogy is changing. The impact of each of these developments will be explored in this section.

3.1 *Feminist Perspectives*

Many feminist scholars are critical of past research and have argued convincingly that there is a flaw in the assumption that before equity in education can be achieved, females must change. For example, many have interpreted past research to indicate that since females lack confidence in their ability, the way to decrease gender differences in mathematics is to help females develop increased confidence, assuming females are dysfunctional because they lack confidence. Feminist scholars have suggested diverse but complementary interpretations of how females in mathematics and science education can be viewed. One feminist view critiques the subject matter, pointing out that females have not been engaged in the construction of mathematics and science. These subject matters must be redefined so that they are not alienating to females (Burton 1994, Harding 1991). It follows from this, that mathematics and science must become viewed as constructed by humans rather than as fixed and objective.

Another view suggests that instruction that has been predominately organized around a male perspective should be reorganized to include a female perspective. Several specific reorganizations have been suggested. One such reorganization has been to have single-sex classrooms where females will not be competing with males, and in which instructional activities which capitalize on 'women's ways of knowing' will be used (Rossi-Becker 1995). Another reorganization includes examination of the ways in which mathematics and science contribute to reproducing social inequality and the marginalization of females. Socially relevant problems form the basis of the curriculum providing females with the knowledge of the use of mathematics and science in political arenas (Mayberry 1998).

Recently feminist scholars have recognized that gender issues affect males as well as females. Males feel pressure from their peers to misbehave in class and find that high academic achievers are treated as social outcasts. These masculine stereotypes interfere with males' learning in all subject areas and have become the focus of a great deal of concern. (See Epstein et al. 1998.)

3.2 *Perspectives from Cognitive Science*

Cognitive science is a rich, complex field that defies definition. At its simplest, it has to do with the study of covert mental activities and the construction of models of mental functioning. Within mathematics and science education, cognitive science research has focused on the study of the mental strategies that are done as an individual solves problems or develops understanding of phenomena. Most work has been done with students in a specifically defined domain such as addition/subtraction or physics. Paths of the development of understanding within these domains have been identified that reflect development in many diverse cultures. However, there has been little work related to gender because it has been assumed that these paths were universal to all humans. Recently this assumption has been called into question in mathematics.

Until a few years ago, it was assumed that although children's solution strategies for addition and subtraction matured at different rates, the developmental paths for girls and boys were similar. However a recent longitudinal study called this assumption into question (Fennema et al. 1998). Results indicated that as children moved from first grade to third grade, girls tended to use immature strategies (counting and modeling) longer while boys tended to use more abstract invented strategies earlier. A significant number of girls skipped the use of these abstract strategies almost entirely, moving to standard computational algorithms before understanding of the concepts underlying the procedures had developed.

In spite of the speed with which they learned to perform the routine procedures, this developmental sequence appeared to interfere with the girls' work on difficult problems, and by third grade, boys were better able to solve more difficult problems than the girls. The researchers suggested that the girls were failing to develop their understanding in mathematics and this might presage the gender differences that are found in complex problem solving at about adolescence. These studies have alerted researchers that gender differences in mathematics may emerge at an earlier age than previously shown. This focus on how students think about problems shows promise for uncovering one of the sources of gender differences on solving complex problems. The study does not offer any evidence about why these differences developed. However, the reporting of the study was accompanied by interpretative comments by researchers with different scholarly backgrounds. They hypothesized that the differences were due to differential learning environments that existed in the classrooms.

3.3 *Changes in the Teaching of Mathematics and Science*

Educational scholars have been encouraging schoolteachers to change the focus of their instruction to conceptual understanding rather than memorization of facts, rules, and procedures. Class discussion has become an integral part of this new form of instruction, where children have opportunities to share their problem-solving strategies and scientific observations. The curriculum includes ill-defined problems and complicated tasks which can be approached in a variety of ways.

Many educators believed that there would be fewer gender differences in classrooms with an emphasis on understanding. An examination of the research that has been done in nontraditional classrooms focusing on gender shows that this has not always been the case. It seems that when there is an emphasis on understanding, girls' performance improves but often the boys' performance improves as well, failing to ameliorate differences. New curricula do not always appeal to girls any more than old curricula. When girls actively engage with science equipment doing experiments their performance tends to improve, but girls are often denied this opportunity by their male peers. It seems that to insure gender equity is achieved in these nontraditional classrooms, teachers must pay particular attention to insuring full participation of all children.

4. *Conclusion*

Few dispute that gender differences in participation in mathematics and science careers have existed for a long time and still exist. However, there is debate about the size and importance of the differences found in learning. Some believe that gender differences in mathematics and science learning indicate a pervasive educational inequity that demand major interventions in the educational system. Others believe the differences are inconsequential, and that attending to these small differences might exacerbate the problem by contributing to an erroneous perception that mathematics and science are male domains. Still others argue that mathematics and science tend to be privileged subject matters and are accorded more respect than they deserve because of their masculine/abstract qualities. They assert that society (and scholars in the area) should not be concerned about gender differences in mathematics and science but instead should celebrate interests that are more typically feminine. But, a majority of scholars in this field (including the authors of this article) believe that the gender differences found in school learning are both important and significant. They contribute to limiting the choices available to females as they progress through school, into adult life and careers. As long as this is the case gender differences are worthy of study so that equity in mathematics and science education can be achieved.

See also: Coeducation and Single-sex Schooling; Education and Gender: Historical Perspectives; Education (Higher) and Gender; Education (Primary and Secondary Schools) and Gender; Gender and School Learning: Mathematics and Science; Mathematical and Logical Abilities, Neural Basis of; Mathematical Education; Rationality and Feminist Thought; Science Education; Sex-role Development and Education; Spatial Cognition

Bibliography

Beaton A, Mullis I, Martin M, Gonzalez E, Kelly D, Smith T 1996 *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Center for the Study of Testing and Evaluation and Educational Policy, Boston College, Boston

Burton, L 1994 Clashing epistemologies of mathematics education: Can we see the wood for the trees? *Curriculum Studies* **2**(2): 203–19
Epstein D, Elwood J, Hey V, Maw J 1998 *Failing Boys? Issues in Gender and Achievement*. Open University Press, Buckingham, UK
Fennema E, Carpenter T, Jacobs V, Franke M, Levi L 1998 A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher* **27**: 6–11
Fennema E, Sherman J 1977 Sex-related differences in mathematics achievement, spatial visualization and sociocultural factors. *American Educational Research Journal* **14**: 51–71
Harding S 1991 *Whose science? Whose knowledge? Thinking for women's lives*. Open University Press, Buckingham, UK
Hildebrand G 1995 Assessment interacts with gender: The case of girls and physics in Victoria, Australia. Paper presented at the UNESCO/Institute of Education Colloquium: Is there a Pedagogy for Girls?, University of London
Hyde J, Fennema E, Lamon S 1990 Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin* **107**(2): 139–55
Kahle J B, Meece J 1994 Research on gender issues in the classroom. In: Gabel D (ed.) *Handbook on Science Teaching and Learning*. Macmillan, New York, pp. 542–57
Kimball M M 1989 A new perspective on women's math achievement. *Psychological Bulletin* **105**(2): 198–214
Leder G 1992 Mathematics and gender: Changing perspectives. In: D. Grouws (ed.) *Handbook of Research on Mathematical Teaching and Learning*. Macmillan, New York, pp. 597–622
Leder G C, Forgasz H 1998 Single-sex groupings for mathematics: An equitable solution? In: Keitel C (ed.) *Social Justice and Mathematics Education*. Frei Universität Press, Berlin, pp. 162–79
Mayberry M 1998 Reproductive and resistant pedagogies: The comparative roles of collaborative learning and feminist pedagogy in science education. *Journal of Research in Science Teaching* **35**(4): 443–59
Rossi-Becker J 1995 Women's ways of knowing mathematics. In: Rogers P, Kaiser G (eds.) *Equity in Mathematics Education: Influences of Feminism and Culture*. The Falmer Press, London, pp. 163–74

R. Ambrose and E. Fennema

Gender and Technology

Since the 1980s an exiciting new field of study has emerged, concerned to develop a gender perspective on technology. Feminists have identified men's monopoly of technology as an important source of their power; and women's lack of technological skills as an important element in their dependence on men. There has been a tension between the view that most new technologies are oppressive and controlling of women, and the obverse view that technology would liberate women from drudgery. A key issue here is whether the problem lies in men's domination of technology, or whether the technology is in some sense inherently patriarchal. This debate resonates with, but is more recent than, the gender questions articulated in relation to science (Harding 1986).

1. Definition of Technology is Gendered

The definition of technology is cast very much in terms of male activities. Technology tends to be thought about in terms of industrial machinery and cars, for example, ignoring other technologies that affect most aspects of everyday life. The history of technology still represents the prototype inventor as male. As a result, women's contributions have by and large been left out of technological history. So, as in the history of science, an initial task of feminists has been to uncover and recover the contribution of women technologists who have been hidden from history. There is now evidence that during the industrial era, women invented or contributed to the invention of such crucial machines as the cotton gin, the sewing machine, the small electric motor, the McCormick reaper, and the Jacquard loom (Stanley 1995).

A greater emphasis on women's activities suggests immediately that women, and in particular black women, were amongst the first technologists. After all, women were the main gatherers, processors, and storers of plant food from earliest human times. It was therefore logical that they should be the ones to have invented the tools and methods involved in this work, such as the digging stick, the carrying sling, the reaping knife and sickle, pestles, and pounders. If it were not for the male orientation of most technological research, the significance of these inventions would be acknowledged. By diminishing the significance of 'women's sphere' technologies, the cultural stereotype of technology as an activity appropriate for men is reproduced.

2. Technological Change and the Sexual Division of Labor

A major concern of feminists has been the impact of technological change on women's lives, particularly in the context of paid work. The introduction of computer-based technologies into offices is a prime site of this research, mainly because the majority of clerical and secretarial workers almost everywhere are women. This research examines the effects of technological change on women's employment opportunities, their experience of work, and their skills (Hartmann et al. 1986, 1987). The exploitation of Third World women as a source of cheap labor for the manufacture of computer components has also been scrutinized (Mitter and Rowbotham 1995).

Much of this research has been influenced by a theoretical perspective, mainly Marxist in orientation, which identifies class conflict as a major factor affecting the technologies of production (Braverman 1975). In order to control the labor force and maximize profitability, capitalism continuously applies new technology designed to fragment and deskill labor, so that labor becomes cheaper and subject to greater control. Rather than technical inventions developing inexorably, machinery is understood to be the result of capitalist social relations. Building on this analysis, feminists questioned the notion that control over the labor process operates independently of the gender of the workers who are being controlled (Cockburn 1983). They pointed out that the division of labor characterizing paid jobs is a sexual hierarchy, and that its gendered nature is not incidental. Both employers as employers, and men as men, have an interest in creating and sustaining occupational sex segregation. Gender is shown time and time again to be an important factor in shaping the organization of work that results from technological change.

From this historical perspective, women's exclusion from technology may be seen as a consequence of the male domination of skilled trades that developed during the Industrial Revolution. Male craft workers actively resisted the entry of women to areas of technological work to protect their own conditions. This rigid gender division of labor within the factory meant that the machinery was designed with men in mind, either by the capitalist inventor or by skilled craftsmen. Industrial technology from its origins thus reflects male power, as well as capitalist domination. The continuing underrepresentation of women in engineering, and indeed all scientific and technical institutions, is a legacy of this history.

In the domestic sphere, by contrast, women are the main users of technology. Considerable optimism has attached to the view that technology may provide the solution to the drudgery of housework. Women's unpaid work in the home, servicing men, children, and other dependants, has for a long time been seen by feminists as a key to women's subordination. Since the 1980s, with the recognition of housework as work, research on domestic technology has grown rapidly, especially among historians (Cowan 1983).

Such research has challenged the main orientation within the sociology of technology towards paid, productive labor in the public domain. Issues that have been central here are the relationship between domestic technologies and time spent on household labor; whether technology has affected the degree of gender specialization of housework; and gender bias in the use of new technologies. Dominating the debates is the knowledge that the amount of time women spend on household tasks has not decreased with 'mechanization of the home.'

This is the central theme of Cowan's (1983) study on the failure of the 'industrial revolution in the home' to ease or eliminate household tasks. She points to the contradictions inherent in attempts to mechanize the home and standardize domestic production. Such attempts have foundered on the nature of housework—privatized, decentralized, and labor-intensive. The result is a completely 'irrational' use of technology and labor within the home because of the dominance of single-family residences and the private ownership of correspondingly small-scale amenities. Domestic technologies clearly reflect the sexual division of domestic activities and the social organization of the family.

3. *Technology Based on Women's Values?*

By the late 1980s, attention was tending to shift away from the focus on women and technology. Instead it began to examine the very processes by which technology is developed and used, and those by which gender is constituted. In other words, feminists were exploring the gendered character of technology itself. This approach has taken two broad directions: that which is influenced by radical feminism, and that which is more in tune with the social studies of technology.

According to the former view, Western technology itself embodies patriarchal values and its project is the domination and control of women and nature (Merchant 1980, Mies and Shiva 1993). This approach has been particularly influential in relation to the technologies of human biological reproduction. It is fueled by the perception that the processes of pregnancy and childbirth are directed and controlled by ever more sophisticated and intrusive technologies. Strong opposition to the development of the new reproductive technologies has been voiced by radical feminists, who see these technologies as a form of patriarchal exploitation of women's bodies (Corea et al. 1985). Central to this analysis is a concept of reproduction as a natural process, inherent in women alone, and a theory of technology as intrinsically patriarchal.

In a similar vein, ecofeminists have analyzed military technology and the ecological effects of other modern technologies as products of a violent patriarchal culture. Technology, like science, is seen as an instrument of male domination of women and nature. Just as many feminists have argued for a science based on women's values (see *Standpoint Theory, in Science*), so too there has been a call for a technology based on women's values. From this perspective, a new feminist technology would be based on a nonexploitative relationship between nature and humanity, and would embrace feminine intuition and subjectivity.

The representation of women as inherently nurturing and pacifist has been subjected to a variety of critiques (see *Feminist Epistemology*). The idea of a technology based on women's values has thus lost

much of its salience. Ecofeminism relies on an idealized femininity and overlooks the role of culture and history in shaping women's needs and priorities in different contexts. It also negates the benefits of modernization in favor of a mythical unchanging tradition.

4. Gender Relations Shape Technology

Current theory and research on gender and technology emphasize the specific characteristics of each technology and the differing positions of women in relation to them. While maintaining that male interests have structured profoundly the form of technologies that have become available, neither men nor the technologies are treated as a homogeneous group.

This approach explores the effects of gender relations on invention and innovation, the design and development processes of technologies, as a counter to the previous stress on the impact of technology on women. It developed in response to the social studies of technology (see *Technological Determinism*; *Actor Network Theory*; *Technology, Social Construction of*). Feminists have demonstrated that the almost complete exclusion of women from the technological community has had a profound influence on the design and technical content of artifacts (Wajcman 1991). Indeed, technological systems implicitly place men's experiences and men's investments at the center, without acknowledging their specificity. The corollary is the simultaneous denial of other realities, such as women's. This gender analysis of technology does not view technology as a patriarchal conspiracy. Rather, it shows how preferences for different technologies are shaped by a set of social arrangements that reflect men's power and resources in the wider society.

As few women actually design technology, we must look elsewhere for their imprint on it. The increasing recognition of the pivotal role of users in technological development has therefore been important for feminist technology studies. Although technologies are designed with specific purposes in mind, the originators of a technology cannot predict or control completely its final use. In practice, the user or consumer of an artifact interacts with it, and can challenge and renegotiate its meanings and uses. By widening the research lens from a focus on innovation to include the processes of manufacturing, marketing, and consumption, women immediately come into view as active participants in the making of technology.

A good illustration is provided by the diffusion of the telephone (Martin 1991). There was a generation-long mismatch between how customers used the telephone and how its inventors envisaged it being used. The telephone was designed by telegraph men who intended that it would be used in the same way as its parent technology, the telegraph, as a business tool. Women's use of the telephone for sociability instead eventually persuaded a reluctant telephone industry to conceive of its networks as a residential service for personal conversation. There are numerous other examples of women's capacity to subvert the intended purposes of technology and turn it to their collective advantage.

Indeed, largely as a result of actor network theory (see *Actor Network Theory*), technology is conceptualized as a fluid, relational process and the symbolic meanings attaching to it as being negotiated and reinvented continually. Contemporary feminist theory likewise stresses that gender is constructed in social interaction. Early studies of gender and technology tended to theorize gender as a fixed and unitary phenomenon, existing prior to and independently of technology, and then becoming embedded within it. Rather than construing technology and gender relations as separate spheres influencing each other, it is now understood that technologies and gender relations are mutually constitutive.

5. Technology as Culture

During the 1990s there was an explosion of feminist writing on technology, much of it being carried out at some conceptual distance from the sociology of science and technology which developed during the same period. This writing is informed explicitly by a combination of cultural studies, anthropology, and postmodern philosophy. Reflecting the postmodern emphasis on discourse, technology is conceptualized as an object of consumption, as a text, and as a communication medium. The central concern is with how technology as culture is implicated in the construction of subjective gender identities.

The most influential feminist commentator writing in this vein is Haraway (1997), who argues that we should embrace the positive potential of technoscience. Sharply critical of those who reject technology, she prefers to be a 'cyborg'—a hybrid of organism and machine parts—rather than an ecofeminist 'goddess' (see *Cyborg*). Genetic engineering, reproductive technology, and the advent of virtual reality are all seen as affecting fundamentally the basic categories of 'self' and 'gender.' Haraway and her ilk positively revel in the very difficulty of predicting what technology's effects will be, and warn against any purist rejection of the 'unnatural,' hybrid entities produced by biotechnology.

Most of the recent literature is about biomedical technologies and information technologies. The increasing preoccupation in sociological theory with the body and sexuality has been paralleled by research on biomedical technologies—technologies for the body. There are numerous studies of childbirth and contraception, in-vitro fertilization, cosmetic surgery and

genetic engineering. Since the late 1980s, feminists have begun to argue that there is now no such thing as a natural, physiological body. One consequence of this work is that the conventional distinction between sex (natural) and gender (social) has been contested and deconstructed thoroughly. Technologies, like science, are now seen as contributing to the stabilization of meanings of the body. With the rise of modern science, bodies have become objects that can be transformed with an increasing number of tools and techniques. Modern bodies are made and remade through science and technology; they too are technological artifacts. Life itself (human, plant, and animal) has been biomedicalized (Clarke 1988).

These studies illustrate the connection between the gendered discourses of the biomedical sciences and the institutionalization of medical techniques applied to women's bodies. They remind us that the conceptualization of male and female bodies as being essentially different, rather than similar, is a modern one, dating only from the eighteenth century. Women were defined in terms of sex and reproduction, and this was reflected in the establishment of gynecology as a separate branch of medicine. With the emergence of sex endocrinology in the 1920s and 1930s, the essence of femininity came to be located in hormones (Oudshoorn 1994). The development of a contraceptive pill for women, rather than for men, is the product not of biological necessity but of twentieth-century biomedical practices.

In the same way as biomedical technologies are seen as transforming the relations between the self, the body, and machines, so too are information and communication technologies. There is a profusion of studies on the Internet, cyberspace, and virtual reality. Many authors celebrate the myriad ways in which modern technology is challenging traditional notions of gender identity and creating new meanings and new entities (Turkle 1995). They argue that gender swapping, or virtual cross-dressing, on the Internet encourages people to reflect on the social construction of gender, to acquire a sense of gender as a continuum. The idea that virtual people or selves can exist in cyberspace, with no necessary link to a physical body, is sure to be taken up in future research.

New communication technologies have certainly opened up novel possibilities for sociality in cyberspace. Excitement has been generated about the capacity they offer to create new feminist communities, and the new tools they offer for global information exchange and political organizing. The significance of this technological revolution for power relations between men and women is still very much an open question.

As a result of the sociological and feminist research carried out in the last decades of the twentieth century, there is now a much more complex understanding of gender, of technology, and of the symbiotic relationship between them. Increasingly, it is accepted that masculinity, femininity, and technology are not fixed, unitary categories but that they contain multiple possibilities and are constructed in relation to each other. At the same time, there has been a rejection of the technophobia evident in earlier feminist writing in favor of a popular cyberfeminism that embraces new technology as a source of empowerment for women. Albeit reflecting diverse perspectives, the study of gender and technology is firmly established as an integral part of science and technology studies.

See also: Feminist Epistemology; Feminist Movements; Feminist Theory; Gender and Reproductive Health; Reproductive Rights in Affluent Nations; Science and Technology, Anthropology of; Women's Health

Bibliography

Braverman H 1975 *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. Monthly Review Press, New York

Clarke A E 1988 *Disciplining Reproduction: Modernity, American Life Sciences, and 'the Problems of Sex'*. University of California Press, Berkeley, CA

Cockburn C 1983 *Brothers: Male Dominance and Technological Change*. Pluto Press, London

Corea G, Klein R D, Hanmer J, Holmes H B, Hoskins B, Kishwar M, Raymond J, Rowland R, Steinbacher R 1985 *Man-Made Women: How New Reproductive Technologies Affect Women*. Hutchinson, London

Cowan R S 1983 *More Work for Mother: The Ironies of Household Technology from the Open Hearth to the Microwave*. Basic Books, New York

Haraway D 1997 *ModestWitness@SecondMillennium*. Routledge, Ithaca, NY

Harding S 1986 *The Science Question in Feminism*. Cornell University Press, New York

Hartmann H, Kraut R, Tilly L (eds.) 1986, 1987 *Computer Chips and Paper Clips: Technology and Women's Employment*. National Academy Press, Washington, DC, Vols. 1 and 2

Martin M 1991 *'Hello, Central?' Gender, Technology and Culture in the Formation of Telephone Systems*. McGill-Queens University Press, Montreal, PQ

Merchant C 1980 *The Death of Nature: Women, Ecology and the Scientific Revolution*. Harper & Row, San Francisco

Mies M, Shiva V 1993 *Ecofeminism*. Zed Books, London

Mitter S, Rowbotham S (eds.) 1995 *Women Encounter Technology: Changing Patterns of Employment in the Third World*. Routledge, London

Oudshoorn N 1994 *Beyond the Natural Body. An Archaeology of Sex Hormones*. Routledge, London

Stanley A 1995 *Mothers and Daughters of Invention: Notes for a Revised History of Technology*. Rutgers University Press, New Brunswick, NJ

Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster, New York

Wajcman J 1991 *Feminism Confronts Technology*. Pennsylvania State Press University, Philadelphia, PA

J. Wajcman

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Gender and the Law

In law, there are two different kinds of gender-related rules: those with gender-specific reference, and those with gender-specific meaning despite gender-neutral wording. These norms and their application in practice reinforce the social construction of gender. The norms depend, of course, on the historic and socio-political context in which they originated, and are subject to change. The development and extension of the equality principle as a human right has made gender inequality in law less and less justifiable, and gender discrimination has been reduced at least in the law on the books. Yet it was left to feminist theorizing to discover gender as a structural principle of modern legal orders in the Western world. This article deals with the questions why gender is used in legal norms, and how and why gender is constructed by the law.

1. *Gender and Genealogy*

All recorded legal orders contain at least some norms referring to the sex or gender of legal subjects, namely in the context of procreation where both women and men are unavoidably involved. Not everyone produces children, but everyone has a biological mother and a biological father, whether they are known or not. Yet the link between a child and its parents is established not only by biology, but also by law. In law, motherhood is in most cases linked to the mere fact of giving birth. The question of fatherhood is less easily settled, because the biological father is necessary for conception, but not for childbearing and giving birth. Therefore any sociopolitical order to which the paternal line is important has rules for establishing fatherhood legally.

The establishment of a genealogy is not necessarily a purpose in itself. However, in class societies and their respective political orders, two questions are of high importance: Who belongs to the ruling class? And who receives the property and other titles if an owner dies intestate? To avoid having to answer these questions by a legally binding act of will in every single case, a genealogy bestowing birth rights simplifies matters. Rules about linking the newborn to its elders then usually include rules about the distinction between rulers and subjects, locals and strangers, free persons and those who are not.

2. *Marriage Law and Regulation of Sexuality*

The legal mechanism of establishing fatherhood may be an act of will like recognition or adoption, but the most widespread mechanism is through marriage: all children born during an existing marriage are automatically considered to be their mother's husband's children, regardless of whether this corresponds to the biological facts. Marrying then, is an anticipating legal act, but it has preconditions. Many rules exist that regulate the question of who may marry whom, or whether he or she may marry at all. And the implications of marriage make marriage a complex legal institution which regulates such issues as the rights and/or duties of the spouses (see *Family Law*). All these rules vary widely over time and space because marriage intersects with religion (see *Hindu Law*; *Islam and Gender*; *Jewish Law*; *Judaism and Gender*), class, race (see *Race and Gender Intersections*) and state interests in population control (encouraging or discouraging procreation). Moreover, many a marriage law treats men and women differently.

Therefore, marriage law is not only the central regulatory site for the establishment of a patrilinear genealogy, but also for the construction of gender. At the end of the Middle Ages, the only type of marriage acknowledged in Europe was, due to the influence of the Catholic Church, the monogamous and principally dissoluble marriage, and it was the only accepted way to procreate, i.e., to produce legitimate heirs. Whereas previously so-called bastards were accepted by noble families when the father was an aristocrat, now children born out of wedlock were always considered 'illegitimate' and were left with their mother's status. This was safeguarded by a criminal law penalizing all forms of sexual conduct save conjugal intercourse (see *Crime and Gender*; *Sexuality and Gender*; *Sexual Orientation and the Law*). Of course, actual behavior differs from legal norms, but nevertheless the impact of such laws worked in two directions: one was the state intervention in the sexuality of its subjects, the other was the gendered discrimination of those who did not comply. When an unmarried woman became pregnant, she could less easily evade prosecution than the man, whose begetting was harder to prove. At the same time, the gendered dependency for procreation was reversed: if a child was born out of wedlock, it was no longer the man who depended on the woman to confirm his fatherhood, but now the woman depended on the man to pass on the status of legitimacy to her children.

3. *The Masculinization of Citizenship*

The marriage laws had their importance for the reproduction of the medieval classes, and for the construction of gender, but gender itself did not yet function as a predominant structural principle for inclusion in the ruling classes (see *Class and Law*). The rulers in medieval Europe were mostly, but not exclusively, men. The laws of the aristocracy (see *Aristocracy/Nobility/Gentry, History of*) usually provided for heredity in the male line, but not all of them excluded women from ruling if there was no legitimate male heir. The rural population came under the

jurisdiction of the respective land owners, whereas the cities generally did not belong to the nobility, but were self-governed by their free citizens. Free citizens were persons who were listed as such, which usually required them to swear a civic oath and to pay civic taxes. Women, whether married or not, were also eligible for full citizenship in medieval European cities (Appelt 1999). This started to change when the feudal orders of the Middle Ages and absolutism were called into question.

Generally, the American and French revolutions are seen as the beginning of the formation of the civil state in the Western world, where estates and classes are abolished. These revolutions had been influenced by the writings of political philosophers of the seventeenth and eighteenth centuries, who had conceptualized a Natural Law in opposition to the Catholic order of the Middle Ages. The Natural Law conceives all human beings as God's creatures and thus became the basis for the equality principle (see *Natural Law*). However, when this concept was made the foundation of the modern state, only men who were free owners were judged capable of being equal citizens; slaves, serfs, Jews and women were, in the beginning, explicitly excluded from the citizenry. The contradiction between the equality of all human beings in Natural Law and the exclusivity of the class of citizens was circumvented by means of the legal mechanism of qualification for legal personality (see *Legal Personality*) which required, according to the same political philosophers, an equal footing in contracting, of which neither slaves and servants nor women were capable. (The exclusion of Jews followed another mechanism linked to the emerging concept of nation.)

It is well known that French women had a major part in the revolution of 1789. Female suffrage was even discussed, but was explicitly denied in 1793. The medieval concept of citizenship was gradually converted into a concept of gendered citizenship which contained also a new concept of gender and the family. Citizens were perceived as males, equals in the public sphere, participating in public affairs by exercising their right to vote and to be elected, and as heads of family in the private sphere. Women were perceived as inherently unequal to men, their 'natural' tasks being to please the men, marry, and care for the 'legitimate' children (Rousseau 1762). As wives they were subordinated to their husbands either by 'nature' directly (Rousseau 1762), or through the marriage contract by which the woman subordinated herself (Kant 1797). The exclusion of women from politics was followed by an exclusion of women from the market during the nineteenth century, the phase of the consolidation of Western nation-states. When finally Jews, slaves, and other nonowners had been enfranchised, women belonging to these groups of the population stayed excluded like the mothers, wives, and daughters of the 'old' citizens. The marital status was the only status that was left after the abolition of statuses connected with the estates, classes, and birth. However, it had a legal meaning only for women, leaving them in some sort of inferior class, with the result that the social and state order was run exclusively along the line of gender.

4. *The Gendered Structure of Modern Laws*

The conception of the citizen as a member of the new ruling class was mirrored not only in the constitutions describing the political person who participates in ruling, but also in the civil laws describing the civil person who participates in the market. Therefore it is no accident that the nineteenth century was, in the Civil Law world (see *Legal Systems, Classification of*), also the period of 'great codifications' (see *Civil Law*). These included norms on the legal personality, on property, contract, tort, and on marriage, family, and succession. Hegel (1821) presents a concise normative foundation for the (bourgeois) civil perception of the relation between the individual, the family, society, and the state.

Marriage and family were made the ideals for both men and women, but with diametrical consequences for their legal capacity. Marriage meant for women that they fell back to the status of minor children even in countries where relief from paternal custody was granted to women of age. Wives lost their property (if they had any) to their husbands (like in England), or their right to dispose of it (like in German states). They gained in turn a claim against their husbands for maintenance, but were obliged to have sexual intercourse whenever their husbands wanted it, as can be seen from the longstanding exemption of marital rape from prosecution. The women's obligation to conceive and give birth to children was reinforced by the abortion laws. Child custody was legally given to the husband, and remained there even after divorce, if a divorce was legally attainable at all. Unmarried and widowed women were likewise judged incapable of exercising child custody, and a male custodian was appointed. (For the status of 'illegitimate' children which hardly changed between the sixteenth century and the end of the twentieth, see *Children and the Law*).

Married women were made economically dependent on their husbands not only by losing their right to dispose of the property they had brought into the marriage, but also because they needed their husband's consent for taking up paid work. However, unmarried women were not free to choose their occupation either, because of the complete exclusion of women from office holding, and an increasing exclusion from crafts and trade; they were dependent on maintenance granted by kin or the community, or on earning a living by working in agriculture, households, or the newly emerging factories—all of which was poorly paid.

This nineteenth-century segregation of the labor market along the line of gender, partly induced and

partly reinforced by legal norms, still exists at the beginning of the twenty-first century. The resulting gender inequality is extended when and where the entitlement to social security is subject to previous paid work. For further treatment of these issues, see *Sex Segregation at Work*; *Sex Differences in Pay*; *Sexual Harassment: Legal Perspectives*; *Social Welfare Policies and Gender*.

Taken all together, the laws of the nineteenth century can be characterized as founding and substantiating civil patriarchy.

5. *Law as a Means to Fight Gender Inequality*

That women were deprived of equal rights has been under criticism from the beginning of the modern constitutional nation-state, starting with the *déclaration des droits de la femme et de la citoyenne* ('declaration of the woman and citizeness') by de Gouges (1791) and the *Vindication of the Rights of Women* by Wollstonecraft (1792). The women's movements that emerged in all Western states around the middle of the nineteenth century called for political participation and civil rights, especially women's suffrage and liberation of the marital subordination of married women (see *Feminist Movements*). It has been a long uphill road to achieve most of these goals in the Western world, at least in black-letter law. The tool that was and is being used in this fight is the invocation of equal rights. So the law, in spite of its function as a means to secure heterosexual male supremacy, could also be used to fight gender inequality (see *Law as an Instrument of Social Change*).

In 1948, the United Nations included gender equality in their Universal Declaration of Human Rights, and many national constitutions followed suit (see *Civil Liberties and Human Rights*; *Civil Rights*; *Fundamental Rights and Constitutional Guarantees*). Since then, it has been much harder to legitimize discrimination against women. In most countries, at least in the Western world, much has been achieved in the name of equality and equal rights to fight women's subordination (and where not, it is still on the agenda); suffrage (see *Women's Suffrage*); access to education and professions; the right to dispose of marital property; the right of married women to work; child custody for unmarried and divorced women (see *Lone Mothers in Affluent Nations*; *Lone Mothers in Nations of the South*); access to social security; reproductive rights (see *Reproductive Rights in Affluent Nations*). Affirmative action laws in the United States (see *Affirmative Action: Comparative Policies and Controversies*) and equal opportunities legislation and adjudication in the European Union (see Hervey and O'Keeffe 1996) have their share in the promotion of gender equality. Women now have, in many countries, access to all kinds of professions and jobs, including the formerly typically male military service.

Gender inequality in law was not only related to the different treatment of men and women, but also to sexuality. In the traditional patriarchal concept of the civil state, heterosexuality was the only accepted form of sexuality until, during the last third of the twentieth century, the gay and lesbian movements in the Western world began to invoke the equality principle on their own behalf (see *Gay/Lesbian Movements*; *Sexual Orientation and the Law*). The decriminalization of homosexuality is still an ongoing process, but sexual orientation is gradually being acknowledged as a human right—e.g., in article 21 of the Charter of Fundamental Rights of the European Union (declared at the Nice European Council on December 7, 2000).

Another demand made by the gay and lesbian movements in the name of equality is directed towards access to marriage. The reason is that many rights and entitlements (e.g., in tax laws, inheritance laws, immigration laws) are restricted to married couples or spouses. Barring homosexuals from marriage according to marriage law means denying them access to these rights and entitlements. This demand has been taken up by the legislators in a number of European countries (first in Denmark in 1989) and some US States, yet without extending the term 'marriage' to such relationships, using a phrase like 'registered partnership' instead (see ILGA-Europe 1998). Only the Dutch laws allow for a 'real' same-sex marriage as of April 1, 2001

Nevertheless the effect of the laws is often limited (see, e.g., *Affirmative Action: Empirical Work on its Effectiveness*), and in no case guaranteed over time. Therefore the struggle against gendered law is still going on, both in countries with no or little achievement and in countries with comparatively high achievement.

6. *Resistance to Gender Equality*

The fight for women's equal rights has always met with considerable resistance. The resistance has three reasons.

The first is that the exclusion of women from the public sphere, including business and the market, and the ideology of marriage and family as the 'natural' profession of women led to a Catch-22 situation: women were subordinated to their husbands if they married, and ridiculed as spinsters if they stayed unmarried, with limited access to the better-paid positions in the labor market. And, most importantly, women had, at the beginning of the twentieth century, no access whatsoever to the lawmaking process—be it parliament and government, or the judiciary—because they were excluded from higher education and the professions irrespective of their marital status (see *Education (Higher) and Gender*; *Legal Education*). Although universities and law schools hesitantly began to open their doors for women from the 1860s

onwards, it was only in the 1970s and 1980s that women's share in the legal professions went beyond tokenism.

The second reason is that modern Western law is conceptualized in terms of individual rights. This is the case also for marriage and family law, although a married couple necessarily consists of two persons. The underlying image of a male head of household and breadwinner, and a female housewife and mother, proved resistant to the mere extension of rights to women—such as suffrage, and access to education and office. It was only in the last third of the twentieth century, and after the emergence of new feminist movements, that the laws were gradually adapted to the image of equal partners, even if the gendered division of tasks, namely breadwinning and family-caring, still resonates.

The third reason is that the patriarchal image of marriage and family, and a corresponding image of professions acceptable for women, had infected all other areas of law. However, the gendered character of modern law went unrecognized by mainstream legal theory until feminist lawyers and political scientists (mostly women, although the recognition of the gendered character of the law is not restricted to females) began to scrutinize the contradictions between equality as a human right, the laws reflecting the civil patriarchy, and women's lived experiences. They uncovered 'The Hidden Gender of Law' (Graycar and Morgan 1990), i.e., many contradictory and/or gender-biased traits of the law (see, e.g., Smart 1984, MacKinnon 1987, Stang Dahl 1987, Patemann 1988, Moller Okin 1989, Young 1990, Fineman 1995, Levit 1998; for further treatment see *Feminist Legal Theory*; *Feminist Theory*).

7. *Towards the End of Gender as a Legal Category? An Outlook*

One aspect of gender as a legal category still remains to be dealt with: the gender dichotomy which is not only taken up, but also created by law. This has been proven for ancient Roman law (see Thomas 1992). Hermaphrodites were even explicitly dealt with by the colified Prussian law of 1794. What is called 'the third sex' vanished, however, from the Western laws in the nineteenth and twentieth centuries when the registration of nationals and residents became increasingly regulated, and the registers attributed either the male or female sex. This official sex dimorphism made the attributed sex seem essential, and also gender became a category with seemingly two mutually exclusive expressions, fitting perfectly the gendered state and law which assigned rights and duties in a gender-specific way.

Sex was usually assigned for life until medical treatment and surgery made it possible to change the originally assigned sex/gender, and a change of sex was finally legally acknowledged (see *Transsexuality, Transvestism, and Transgender*). Yet the permission for a change of sex does not mean that the cultural dimorphism of sex and gender has been overcome. The same medical treatment is used to uphold this dimorphism in the case of hermaphrodites (intersexed, intersexuals) who, in Western countries, already as infants are treated for becoming unambiguous females or males (see Kessler 1998). It seems hardly questionable that the legally secured sex dimorphism and gender dichotomy has contributed to this branch of surgery. Up till 2000, only the Constitutional Court of Columbia (SU-337/97 of May 12, 1999, and T-551/99 of August 2, 1999) recognized the problem of genital surgery imposed on infants as a legal problem conflicting with the personal integrity and dignity of those concerned.

Altogether sex, and consequently gender, as a legal category is, at the beginning of the twenty-first century, called into question from four different, but partly related, angles:

(a) separation of 'legitimate' procreation from marriage;

(b) separation of marriage (or a functional equivalent) from heterosexuality;

(c) extension of political, civil, and economic rights to all individuals, irrespective of their sex, sexuality, and gender; and

(d) rhe (re)discovery that nature offers a greater variability than the Western dimorphic conception of sex and gender of the nineteenth and twentieth centuries.

Under such conditions, it would be superfluous to register sex/gender or refer to sex/gender in any legal provision. Legal regulation could refer to a given situation instead of generalized males or females. Even in the context of procreation (to come back to the starting point), the distinction along the sex/gender line seems disposable when motherhood and fatherhood are replaced by parenthood. The function of law for linking a newborn to someone responsible will remain, but the factual situation to which legal parenthood is knitted could be actual care instead of biological generation (which has become problematic in the light of the recent reproductive technologies; see Knibiehler 1987). Parenthood could then be assigned to one, two, or even more persons, whether the biological parents be among them or not. Likewise, laws protecting motherhood, which have a long tradition in Europe, could be linked to the actual occurrence, and, where only aimed at female reproductive capacity, be extended to everyone.

However, as long as a gendered structure can still be found in the laws, complete disregard of gender might be counterproductive to those who still have unequal access to power and the exercise of rights because the demand for equal rights, the tried and tested tool to fight discrimination by law, requires the reference to differences.

See also: ; Aboriginal Rights; Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Civil Rights Movement, The; Conflict Sociology; Critical Race Theory; Economic Globalization and Gender; Ethnic and Racial Social Movements; Ethnicity, Sociology of; Family and Gender; Family Bargaining; Family Theory: Feminist–Economist Critique; Fatherhood; Female Genital Mutilation; Feminist Economics; Feminist Epistemology; Feminist Ethics; Feminist Legal Theory; Feminist Political Ecology; Feminist Theology; Feminist Theory: Psychoanalytic; Gay, Lesbian, and Bisexual Youth; Gender and Feminist Studies in Anthropology; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Geography; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Economics of; Gender, Feminism, and Sexuality in Archaeological Studies; Gender History; Historiography and Historical Thought: Southeast Asia; Household Production; Identity Movements; Klein, Melanie (1882–1960); Labor Movements, History of; Male Dominance; Mead, Margaret (1901–78); Mortality of the Oldest Old; Motherhood: Economic Aspects; Multicultural Feminism: Cultural Concerns; Multiculturalism: Sociological Aspects; Peace Movements, History of; Plagues and Diseases in History; Queer Theory; Race Identity; Racism, Sociology of; Rape and Sexual Coercion; Rationality and Feminist Thought; Regulation: Sexual Behavior; Reproductive Rights in Affluent Nations; Sex Differences in Pay; Sex Segregation at Work; Social Movements, Sociology of; Youth Movements, History of

Bibliography

Appelt E 1999 *Geschlecht—Staatsbürgerschaft—Nation: Politische Konstruktionen des Geschlechterverhältnisses in Europa* [*Gender—Citizenship—Nation: Political Constructions of Gender Relations in Europe*]. Campus, Frankfurt-am-Main, Germany

Fineman M A 1995 *The Neutered Mother, the Sexual Family and other Twentieth Century Tragedies*. Routledge, New York

Gouges O de 1791 *Les droits de la femme*. Paris [1989 *The Rights of Woman*, trans. V Stevenson, Pythia, London]

Graycar R, Morgan J 1990 *The Hidden Gender of Law*. Federation Press, Annandale, Australia

Hegel G W F 1821 *Grundlinien der Philosophie des Rechts oder Naturrecht und Staatswissenschaften im Grundrisse*. Nicolaische Buchhandlung, Berlin [1991 *Elements of the Philosophy of Right*. A W Wood (ed.) H B Nisbet (trans.), Cambridge University Press, Cambridge, UK]

Hervey T K, O'Keeffe D (eds.) 1996 *Sex Equality Law in the European Union*. Wiley, Chichester, UK

ILGA-Europe (European Region of the International Lesbian and Gay Association) (ed.) 1998 *Equality for Lesbians and Gay Men: A Relevant Issue in the Civil and Social Dialogue*. ILGA-Europe, Brussels, Belgium

Kant I 1797 *Metaphysik der Sitten*. Nicolouius, Königsberg, Germany [1991 *The Metaphysics of Morals*, trans. M. Gregor, Cambridge University Press, Cambridge, UK]

Kessler S J 1998 *Lessons from the Intersexed*. Rutgers University Press, New Brunswick, NJ

Knibiehler Y 1987 *Les pères aussi ont une histoire*. Hachette, Paris

Levit N 1998 *The Gender Line: Men, Women, and the Law*. New York University Press, New York

MacKinnon C A 1987 *Feminism Unmodified: Discourses on Life and Law*. Harvard University Press, Cambridge, MA

Moller Okin S 1989 *Justice, Gender, and the Family*. Basic Books, New York

Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA

Rousseau J-J 1762 *Émile ou de l'Éducation*. Néaulme, La Haye, France [1979 *Emile: or On Education*, trans. A Bloom, Basic Books, New York]

Smart C 1984 *The Ties that Bind: Law, Marriage, and the Reproduction of Patriarchal Relations*. Routledge and Kegan Paul, London

Stang Dahl T 1987 *Women's Law: An Introduction to Feminist Jurisprudence*. Norwegian University Press (Universitetsforlaget AS), Oslo, Norway

Thomas Y 1992 The division of the sexes in Roman law. In: Schmitt Pantel P (ed.) *From Ancient Goddesses to Christian Saints: A History of Women in the West*, Belknap Press of Harvard University Press, Cambridge, MA, pp. 83–138

Wollstonecraft M 1792 *A Vindication of the Rights of Women: with Strictures on Political and Moral Subjects*. J. Johnson, London

Young I M 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

K. Plett

Gender, Class, Race, and Ethnicity, Social Construction of

1. Introduction

The categories gender, class, race, and ethnicity, widely used in the social sciences, history, and literature, have long been treated as distinct categories, which supposedly describe real differences among groups of people. In this article this assumption is debated and various theories about the usefulness and the interaction and interrelationship of these categories are examined.

2. Background

Since the 1970s under the impact of postmodernism, deconstruction, feminist scholarship, and challenges to it by women of color, fundamental categories of

knowledge and systems of explanation have been disputed, problematized and, in some cases overthrown. Rigidly defined categories that had long been accepted as based on 'natural,' self-evident differences, and as being disparate from one another, were now subject to critique and redefinition. This process of reevaluation took place over several decades and in many different fields. It will be discussed here category by category, although there was considerable overlap and crossover from one category to the other. Each of the categories represents some noticeable distinction from an imaginary norm. Each also is a powerful mental construct, which can be positively or negatively evaluated, according to who it is that does the defining.

3. *Challenging the Categories*

3.1 *Race*

The term 'race' first appeared in the sixteenth century; only in the nineteenth century did it become biologized on the basis of evolutionary thought. The 'races of mankind' were classified and hierarchically ordered by Europeans, with whites inevitably at the top of the ranking order. Modern science, however, gives no support to essentialist racial theories. There is no such thing as 'black blood' or 'pure blood' and the only distinctions of blood—bloodtypes—have nothing to do with race (see *Race: History of the Concept*). Biological race is, in fact, a scientifically nonexistent category. Of course, the physical characteristics (such as skin and hair color) that we associate with race are determined biologically. But which physical characteristics are seen to define race, and where lines are drawn between groups, as well as how groups are treated differently, is determined socially. For example, we note differences in eye coloring and attach no value judgments to them, but when it comes to differences in hair and skin coloring, all the contradictory and often negative value judgements accruing to race come into play.

Challenges to a naturalized concept of biological race began early in the twentieth century and found expression in the thought of African–American intellectuals. W. E. B. DuBois early in his life regarded race as a sociohistorical term, but still accepted the idea of 'a common blood.' In 1911 he wrote: 'It is not legitimate to argue from differences in physical characteristics to differences in mental characteristics' (Dubois 1911 but by 1940 he concluded: 'race is the fact that a Negro in Georgia has to ride in the Jim Crow car' DuBois 1940 p. 153.). In short, 'race' is nothing else but 'racism.'

The unscientific nature of the category race was recognized legally by the Supreme Court when it ruled miscegenation laws unconstitutional in 1967 (*Loving vs. Virginia*). Contemporary scholars, both black and white, agree with historian Barbara Fields: '… the view that race is a biological fact, a physical attribute of individuals, is no longer tenable. Race … is a purely ideological notion' (Fields 1982, p. 144; Henry Louis Gates and Kwame Anthony Appiah have expressed similar views). Many authors have searched for words which might more adequately describe the content of this complex concept, such as institutionalized racism, experiential racism, resistance formation. Literary critic Hazel Carby asked academics to stop using the term, but rather to speak of 'racial formations.' In Germany, in the aftermath of the Holocaust the term race is rarely used in academic discourse and has been replaced by the term 'ethnicity.'

However, race as a factor describing and determining people's life chances continues to exist. It is manifested in three forms: (a) as an ordering concept for dominance and discrimination, which has become institutionalized; (b) as a historic experience, a force shaping the lives and consciousness of the designated group; and (c) as a distinguishing marker for the group, transformed into a mark of pride, resistance, and a tool for liberation.

By exposing the category race as a social construct and showing its connection to power and to the unequal distribution of resources in society, scholars opened the way to a critique of the other categories. They also showed that the ability to define and categorize had developed into a powerful tool for social control. Considered from the point of view of philosophy, false universalizing (the situation where one category, such as men, is defined as the norm and a category seen as opposite, women, is defined as 'the Other' or the deviant) was found objectionable, and seen as a basic conceptual error. False universalizing was also implicated in explanatory systems pertaining to race, to sexuality, to ethnicity (Minnich 1990).

3.2 *Sex/Gender*

Simone de Beauvoir's *The Second Sex*, which appeared in the USA in 1953, explained woman's subordinate status in society as the result of her being defined as man's 'Other.' The author argued that 'one is not born, one is created a woman.' This insight led feminist scholars to new definitions, whereby Sex (the biological given) was contrasted with Gender (the socially constructed behavior, and the roles deemed appropriate to each of the sexes). She argued convincingly that the definition of woman as man's 'Other' had led to the systematic disadvantaging of women in all aspects of society, law, and culture. Gender would later be developed as a tool of analysis in many fields (Scott 1988; see *Gender and Feminist Studies*).

Comparisons between the way 'Negroes' and women were characterized, stereotyped, and discriminated against appeared in the sociological literature of the 1950s and became central to the arguments

of modern feminists (Hacker 1951). The dichotomy male–female was paralleled to white–black and other binary contrasts, like heterosexual–homosexual. Feminists criticized the binary system of thought and began to concern themselves with the parallel structures of ideas, prejudices, and social customs by which African–Americans and women were conceptualized as 'deviants' from a white, male norm. This led to other questions. How did the categories by which people were defined as deviants affect their lives? Definitions are mental constructions; discrimination in access to education, to resources, to political power is tangible and real. How did this work? Were the categories, gender, class, race, and ethnicity linked in some way and if so, how? Were they part of one or several systems? In the 1970s and 1980s the feminist debate postulated the existence of separate systems. Rubin (1975) described a single sex/gender system that linked economic oppression with gender. Marxist-feminists tried to fit a feminist analysis into Marxist categories, adding new dimensions and structures (Rowbotham 1973, Mitchell 1974, Davis 1981). O'Brien described a capitalist system of production and class oppression, which operated in conjunction with a separate system of reproduction and sex oppression. The Canadian sociologist Dorothy Smith defined gender and race in relational terms. She argued that relations of power manifest themselves in multiple, fluid structures of domination (Smith 1987).

3.3 Differences Reconsidered and Universals Challenged

In the early 1980s the concept of 'sisterhood' came under sharp attack by women of color, by post-colonialists and various ethnic groups. The emphasis of the debate now shifted to differences among women. Where previously race, class, and ethnicity had simply been treated as aspects of women's identity, which might render their situation different from one another, these differences were now seen as so great that no valid generalizations about 'women' were possible. African–American women historians challenged the notion of a homogeneous womanhood and argued for the integral connection of race, gender, and class in the thought and lived experience of black women (see, e.g., Fields 1982, Higginbotham 1992). Patricia Collins argued that the experiences of African–American women were so unique as to qualify them singularly to interpret their past from an 'Afrocentric standpoint' (Collins 1989). 'Standpoint theory' originally developed by Nancy Hartsock, has also been used in discussions of ethnicity, class, and identity (Hartsock 1983). While it is helpful in drawing attention to the faults of false universalization, it not only affirmed but seemed to increase belief in the fixity of the assigned categories.

3.4 Binary Sexual Categories

The binary distinction between heterosexuals and homosexuals mirrors the power relations in other dichotomized categories. In her incisive essay Rich (1980) showed how the marginalization of gays and lesbians derived from their being defined as 'deviant' and essentially different from the heterosexual 'norm.' She traced the intellectual constructs, the legal constraints, and the everyday cultural practices that undergird the resulting system of discrimination against yet another minority group. The pattern of construction and 'othering' is remarkably similar to that in the other categories. In recent decades gay history and theory has flourished and led to the formation of another new field, Lesbian and Gay Studies (an introduction to the growing literature can be found in Abelove et al. 1993).

4. Class

Class is a term that sorts people according to their relation to resources and power. The term is a social construct, but it is based on actual economic differences between people. There is a large literature illustrating the way race and gender affect a person's class position. Studies of a gendered labor market, of racial and gender discrimination in educational opportunities and in access to jobs, the continuing wage differential between men and women, and the income gap between black and white citizens argue for the significance of sex and race in defining class. Various ethnicities have, at different times, been subject to similar economic discrimination. Many labor historians now argue that the concept 'class' is gendered and racialized inherently (Moody and Kessler-Harris 1989). Class privileges and disadvantages vary by sex and race. For example, in nineteenth-century America the definition of male middle-class status required a man to be the sole family breadwinner. A 'white lady' by definition was someone sharply differentiated from lower-class women, from women defined as 'not respectable,' and from all nonwhite women. Another argument for the gendered nature of class is that it has been maintained historically by homogamy (the practice of marriage among persons of the same class) and by gendered inheritance laws that specified that sons were privileged over daughters in inheriting property. Both practices have given men a long-range economic advantage over women, by giving sons land and daughters dowries, when land was the most valuable property to have. Similarly, in modern times, gendered inheritance has given sons more family investment in higher education than daughters, when such education guaranteed upward mobility (Lerner 1997, pp. 146–98). There is a large literature concerning the nature of women's unpaid housework and child rearing services and its effect on class. Despite considerable analytic

differences, there is agreement that their responsibility for unpaid housework generally has debased the wages of working women and their opportunities for economic advancement (Hartman 1981, Kessler-Harris 1990, Boydston 1990).

5. *Ethnicity*

This is an affiliation of people sorted by cultural and historical characteristics. These categories are obviously imprecise and arbitrary and inadequately deal with people of mixed ethnic origin. The ethnic designations in the US census have come under constant attack by people who consider themselves of 'mixed' ethnicity or race and by others who object to the definition of categories such as 'Hispanic' (Root 1996, Anzaldua 1987, Zack 1993).

Some scholars have developed the view that ethnicity is a modern, constructed identity which is reinvented in each generation and reinforced by a variety of cultural means in the interest of nationalisms (Sollers 1989). An outgrowth of the debates on ethnicity and on race was the creation of a new racial designation and a new field of study, the creation of Whiteness. If race cannot be proven on a biological basis, neither can whiteness. White people are defined as the norm against the 'Otherness' of people of color. Whiteness bestows privileges on whites that are withheld from 'Others.' One of these privileges is that it is whites who do the defining of the social categories and who can remain unconscious of their own racial construction (Minnich 1990, Roediger 1991, Frankenberg 1993). The racialization of whiteness was used to argue both against racism and against affirmative action, by the claim that whites should be considered a group of ethnics who were disadvantaged and oppressed by race-specific legislation.

6. *The Relationship of the Categories to Each Other; The System as a Whole*

The relationship of the categories under discussion has often been described in geometric terms as 'intersection' (race and sex intersect to form class); it has been visualized as forces (sex acts upon class creating gendered class) or as a relationship approximating that of a mathematical curve. More and more, these ideas deriving from mechanical thought have given way to more holistic concepts, such as interacting, interpenetrating. The key to such an approach is to challenge the very existence of separate categories and identities and the recognition that people do not define themselves by a single identity, but by a number of identities which are interactive and mutually constitutive (Memmi 1968, Lorde 1984).

Postcolonial studies have illuminated this process and argued for its wider applicability. Stoler has shown how colonial powers, in Indonesia, Indochina, and India, constructed a concept of the 'colonizer' vs. 'the colonized' in which race did not serve as the main marker of difference. Rules of conduct and custom enforced strict segregation between the two groups, reinforced by the myth that white women needed protection from assault by colonized males. Rules against miscegenation, abandonment of 'mixed race' children by their fathers, and strictures against white single women together with the encouragement of marriage among whites, encouraged the domesticity of white women and upheld colonial rule (Stoler 1992). Mohanty described 'the ideological construction and consolidation of white masculinity as normative and the corresponding racialization and sexualization of colonized people' in the case of Africa. Other studies have strengthened this argument (Mohanty 1991).

The connections between the categories appear ever more dense and complex. There has slowly developed agreement among many scholars that the terms must be treated not as separate entities, but as interrelated. Bunch (1990) argued that various forms of oppressive systems interact. Hooks (1989) stated that: 'Sexism, racism and class exploitation constitute interlocking systems of domination.' Sacks (1989) saw 'Class, race and gender oppression as part of a unitary system.' Morgen (1990) explained 'that these social relations so interpenetrate that they are better understood as constituting...one another.' Noticing the arbitrariness of the selection of groups that are targeted for definition as 'the Other,' a number of scholars have come to view the process of forming deviant outgroups by way of categorization as part of the system of ruling (e.g., Lorde 1984, Williams 1991, Minnich 1990, Lerner 1997). Gender, race, ethnicity, and class are processes through which hierarchical relations are created and maintained in such a way as to give some men and women power and privilege over others by their control of material resources, sexual and reproductive services, education, and knowledge. Such control is maintained by a complex weave of social relations among dependent groups which offers most groups some advantages over other groups, sufficient to keep each group within the dominance system.

The categories are not just interrelated, they are also mutually constitutive. Gender is constructed racially and through class and ethnicity. Race is constructed genderically. Thus, the gender ideal of 'womanhood' was long defined for white women as embodying purity and dependency on male protection, while women of color were seen as 'impure,' open to sexual access by white men, and economically self-supporting. The gender ideal turns out to be race-specific. By a similar mental contortion middle- and upper-class mothers are today urged to stay home with their preschool children, while mothers on welfare are forced to leave their infants to go to work. This ideal turns out to be class-specific. A similar case can be

made for all other social formations. A holistic redefinition of the social categories may lead to a more accurate description of society as a functioning organism.

7. Conclusion

Scholarship on diversity and multiculturalism has gone through a process which began with the traditional acceptance of rigid, separate categories into which groups of persons were divided according to noticeable or invented differences. Challenging the categories as neither natural nor separate, denaturalizing them, led to the insight that this ordering system was connected with power—the power by which difference is turned into dominance. Underlying the construction of categories of difference was a binary system of thought, which soon came under attack from many quarters. The search for alternate modes of thinking and for models more appropriate to diversity in democratic societies led to greater concern with inclusivity, and a new scholarly interest in the problems of borderlands, mixed identities, hybridity, and holistic models.

See also: Conflict Sociology; Critical Race Theory; Ethnic and Racial Social Movements; Ethnicity, Sociology of; Feminist Theory and Women of Color; Gender and Feminist Studies; Multicultural Feminism: Cultural Concerns; Multiculturalism: Sociological Aspects; Race and Gender Intersections; Race: History of the Concept; Race Identity; Racism, Sociology of; Social Class and Gender

Bibliography

Abelove H, Barale M A, Halperin D (eds.) 1993 *The Lesbian and Gay Studies Reader*. Routledge & Kegan Paul, New York

Anzaldua G 1987 *Borderlands: La Frontera: The New Mestiza*. Spinsters/Aunt Lute, San Francisco

Berger P L, Luckmann T 1966 *The Social Construction of Reality: A Treatise in The Sociology of Knowledge*. Anchor Books/Doubleday, New York

Boydston J 1990 *Home and Work: Housework, Wages, and the Ideology of Labor in the early Republic*. Oxford University Press, New York

Bunch C 1990 Making common cause: Diversity and coalitions. In: Albrecht L, Brewer R M (eds.) *Ridges of Power: Women's Multicultural Alliances*. New Society Publishers, Philadelphia, PA

Collins P H 1989 Social construction of black feminist thought. *Signs* **14**(4): 745–74

DuBois W E B 1911 Races. *Crisis* **Aug**: 157–8

DuBois W E B 1940 *Dusk of Dawn: An Essay Toward an Autobiography of a Race Concept*. Harcourt Brace, New York

Davis A 1981 *Women, Race and Class*. Random House, New York

Fields B 1982 Ideology and race in American history. In: Kinsser J M, McPherson J M (eds.) *Region, Race and Reconstruction*. Oxford University Press, New York, pp. 143–77

Frankenberg R 1993 *White Women, Race Matters: The Social Construction of Whiteness*. University of Minnesota Press, Minneapolis, MI

Hacker H 1951 Women as a minority group. *Social Forces* **30**: 60–9

Hartman H 1981 The family as the locus of gender, class, and political struggle: The example of housework. *Signs* **6**(1): 366–94

Hartsock N 1983 The feminist standpoint: Developing the ground for a specifically feminist historical materialism. In: Harding S, Hintikka M B (eds.) *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics Methodology and Philosophy of Science*. Reidel, Boston

Higginbotham E B 1992 African–American women's history and the metalanguage of race. *Signs* **17**(2): 251–74

Hooks b 1989 *Talking Back: Thinking Feminist, Thinking Black*. South End Press, Boston

Kessler-Harris A 1990 *A Woman's Wage: Historical Meanings and Social Consequences*. University of Kentucky Press, Lexington, KY

Lerner G 1997 Differences among women and rethinking the paradigm: Race; Class. In: Lerner G (ed.) *Why History Matters: Life and Thought*. Oxford University Press, New York

Lorde A 1984 *Sister Outsider*. Crossing Press, Trumansburg, NY

Memmi A 1968 *Dominated Man*. Orion Press, New York

Minnich E K 1990 *Transforming Knowledge*. Temple University Press, Philadelphia, PA

Mitchell J 1974 *Woman's Estate*. Pantheon, New York

Mohanty C T 1991 Introduction: Cartographies of struggle, third world women and the politics of feminism. In: Mohanty C T, Russo A, Torres L (eds.) *Third World Women and the Politics of Feminism*. Indiana University Press, Bloomington, IN

Moody J C, Kessler-Harris 1989 *Perspectives on American Labor History: The Problem of Synthesis*. Northern Illinois University Press, DeKalb, IL

Morgen S 1990 Conceptualizing and changing consciousness. In: Hansen K, Philipson I (eds.) *Women, Class and the Feminist Imagination*. Temple University Press, Philadelphia, PA, p. 285

Rich A 1980 Compulsory heterosexuality and lesbian existence. *Signs* **5**: 657

Rowbotham J 1973 *Woman's Consciousness, Man's World*. Penguin, Baltimore, MD

Roediger D R 1991 *The Wages of Whiteness: Race and the Making of the American Working Class*. Verso, London

Root M P P (ed.) 1996 *The Multiracial Experience: Racial Borders as the New Frontier*. Sage, Thousand Oaks, London, New Delhi

Rubin G 1975 The traffic in women: Notes on the political economy of sex. In: Reiter R R (ed.) *Toward an Anthropology of Women*. Monthly Review Press, New York

Sacks K B 1989 Toward a unified theory of class, race, and gender. *American Ethnologist* **16**(3): 534–50

Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York

Smith D 1987 *The Everyday World as Problematic: A Feminist Sociology*. Northeastern University Press, Boston

Sollers W (ed.) 1989 *The Invention of Ethnicity*. Oxford University Press, New York

Stoler A L 1992 Sexual affronts and racial frontiers: European communities and the cultural politics of exclusion in colonial Southeast Asia. *Comparative Studies in Society and History* **34**(2): 514–51
Williams P 1991 *The Alchemy of Race and Rights*. Harvard University Press, Cambridge, MA
Winant H, Omi M 1986 *Racial Formation in the United States*. Routledge & Kegan Paul, New York
Zack N 1993 *Race and Mixed Race*. Temple University Press, Philadelphia, PA

G. Lerner

Gender Differences in Personality and Social Behavior

Research on psychological gender differences began before the beginning of the twentieth century, as scientists tried to discern differences between women and men in intelligence, reaction time, and personality, using measures that seem bizarre by today's standards, such as handwriting analysis. Sophisticated research on gender differences in personality and social behavior has flourished since the 1970s, generating hundreds of empirical studies as well as a number of highly informative meta-analyses. A meta-analysis is a quantitative review of the existing studies on a particular question; the meta-analyst extracts statistics from the existing studies and averages the results across all studies. (When reporting effect sizes from meta-analyses here, I follow the convention that a positive effect size means that men scored higher on the measure in question and a negative effect size means that women scored higher. An effect size tells how far apart the two groups—in this case, men and women—are in their average scores, standardized by dividing by the standard deviation.)

1. *Gender Differences in Personality*

Definitions of the term 'personality' vary widely. In order to sharpen the focus of this review, I will consider personality as conceptualized in the Big Five model (McCrae and Costa 1987), which specifies five basic dimensions of personality: conscientiousness, extraversion, openness to experience, agreeableness, and neuroticism. Below I review research on gender differences in each of these dimensions. A meta-analysis of gender differences in personality (Feingold 1994) is particularly helpful in discerning the overall pattern of findings.

1.1 *Conscientiousness*

Feingold located seven norming studies measuring conscientiousness using standardized personality tests. Of those, women scored higher on four, one found exactly no difference, and men scored higher on the remaining two. The unweighted mean effect size was $d = -.13$ (Feingold 1994). Therefore, the gender difference in conscientiousness appears quite inconsistently across studies and, if there is a gender difference in conscientiousness, it favors women but the difference is slight.

1.2 *Extraversion*

Most studies find no gender differences in extraversion as measured by the MMPI (Zonderman et al. 1993), or the Eysenck Personality Questionnaire (EPQ) (Martin and Kirkcaldy 1998).

In his meta-analysis, Feingold (1994) distinguished between two aspects of extraversion: gregariousness and assertiveness. Most test norming studies show women to be more gregarious, although the gender difference is small (unweighted mean $d = -.14$) and inconsistent across countries (Feingold 1994). Test norms show men scoring higher on assertiveness or dominance, ($d = .50$).

1.3 *Openness to Experience*

Openness to experience—flexibility and openness to new ideas—has been less studied. The mean effect size for gender differences was $d = .03$ (Feingold 1994), indicating no gender difference in openness to experience.

1.4 *Agreeableness*

Findings regarding gender differences in agreeableness are mixed. Feingold (1994) distinguished between two aspects of agreeableness: trust and tender-mindedness. Test norming data showed a consistent trend for women to score somewhat higher on trust ($d = -.25$) and for women to score considerably higher on tender-mindedness or nurturance ($d = -.97$). Higher scores for women on tender-mindedness or nurturance have been documented in Canada, Finland, Germany, Poland, and Russia, but not China (Feingold 1994).

1.5 *Neuroticism*

Neuroticism is the tendency to experience negative emotions such as sadness or anxiety, as well as mood swings (Costa and McCrae 1985). Those scoring high on neuroticism tend to worry or ruminate a great deal and are prone to having their feelings easily hurt. Women consistently display higher scores on neuroticism scales than men do. These differences are found on the MMPI Neuroticism scale (Zonderman et al. 1993), the Eysenck Personality Questionnaire (Martin and Kirkcaldy 1998), and the NEO (Costa

and McCrae 1985). The magnitude of the gender difference is generally around $d = -.25$ (Feingold 1994, Zonderman et al. 1993), that is, the difference is small. For comparison, the gender difference on the masculinity–femininity scale of the MMPI is approximately $d = 2.0$ (Zonderman et al. 1993).

The gender difference has been documented in a wide variety of countries including the USA, Singapore, and Bangladesh (Lynn and Martin 1997), and Northern Ireland (Martin and Kirkcaldy 1998).

The gender difference in neuroticism is a particular concern because of the association between neuroticism and depression and the well-established gender difference in depression (Roberts and Gotlib 1997). Watson and Clark (1984) have demonstrated that the personality trait labeled neuroticism is part of a broader cluster of traits including emotionality, trait anxiety, and ego strength; they label this broad personality construct 'negative affectivity.' Scales measuring traits in this cluster consistently show women scoring higher.

1.6 Self-esteem

Self-esteem, although not one of the Big Five dimensions, is an individual difference variable that has stimulated much research on gender differences. On the basis of his meta-analysis, Feingold (1994) concluded that males had slightly higher self-esteem than females, ($d = .16$).

Kling et al. (1999) conducted a more extensive meta-analysis of gender differences in self-esteem (for another meta-analysis with similar results, see Major 1999). Averaged over 216 effect sizes, ($d = .21$), indicating a small gender difference favoring males. Contrary to claims in the popular press that girls' self-esteem plummets in early adolescence, creating a large gender difference, the largest gender difference appeared in the high school years and was small to moderate in magnitude, ($d = .33$). Results based on nationally representative data sets from the National Center for Education Statistics showed smaller gender differences in the high school years, with nearly all effect sizes less than .16. When the studies were analyzed according to the ethnicity of the sample, the results indicated that a small gender difference favoring males exists for Whites ($d = .20$) whereas there is no gender differences for Blacks ($d = -.04$).

In conclusion, the gender difference in self-esteem is small at most and, in the USA, is found in White samples but not in Black samples.

1.7 Summary

On the Big Five dimensions of personality, there appear to be no gender differences in openness to experience and only a slight gender difference favoring women in conscientiousness. Within the dimension of extraversion, if one distinguishes between gregariousness and assertiveness or dominance, there is a small gender difference favoring women on the former and a moderate gender difference favoring men on the latter. For agreeableness, if one distinguishes between trust and tender-mindedness or nurturance, there is a small gender difference in the former and a large gender difference in the latter; in both cases, women score higher. A small gender difference in neuroticism, women scoring higher, appears consistently across studies and cross-nationally. Finally, the gender difference in self-esteem is small at most and, in the USA, is found in White samples but not in Black samples.

2. Gender Differences in Social Behavior

This section reviews research on gender differences in aggression, helping behavior and altruism, group behaviors, and communication and nonverbal behaviors.

2.1 Aggression

Two meta-analyses provide summaries of the existing data on gender differences in aggression. Hyde's (1986) meta-analysis used a developmental approach in reviewing 143 studies. Averaged over all studies, the average effect size d was .50; that is, the magnitude of the gender difference was moderate. Naturalistic studies of aggression yielded larger gender differences ($d = .56$) than laboratory experimental studies ($d = .29$). Studies of physical aggression produced larger gender differences ($d = .60$) than studies of verbal aggression ($d = .43$). One should note, however, that both effect sizes are positive; that is, males were more aggressive for both physical and verbal aggression. Gender differences were larger for preschoolers ($d = .58$) than for studies of college students ($d = .27$).

Eagly and Steffen (1986) took a different approach in their meta-analysis, limiting their review to social psychological research on aggression in adults, mainly college students. They obtained a mean effect size of .29. The discrepancy between this overall effect size and the value of .50 reported by Hyde (1986) doubtless lies in the fact that Eagly and Steffen reviewed only studies of adults, among whom gender differences are smaller; for studies of college students, Hyde found an effect size of .27, nearly identical to the one obtained by Eagly and Steffen.

2.2 Helping Behavior and Altruism

Eagly and Crowley (1986) conducted a meta-analysis of studies of gender differences in helping behavior. They grounded their meta-analysis in social role theory and argued that it made no sense to examine an overall effect size for gender differences in helping

behavior. Instead, whether males help more or females help more depends on whether the situation calls for helping that is consistent with the male role or the female role. The results are therefore highly dependent on the situations in which social psychologists have measured helping. Men, for example, are more likely to help in situations in which they know they are being observed ($d = .74$) whereas there is no gender difference ($d = -0.02$) when participants believe they are not being observed. Heroism is part of the male role and heroism requires onlookers. Helping in the private context of long-term relationships is part of the female role, for example, a mother helping a child. This type of helping has been neglected in social psychological research.

Eisenberg and Fabes (1998) conducted a meta-analysis of studies of gender differences in children's prosocial behavior. The overall effect size was $d = -.18$, indicating a small gender difference favoring girls. Girls scored higher on all types of prosocial behavior, including instrumental help and comforting; the largest gender difference was for being kind or considerate, $d = -.42$.

2.3 Group Behaviors

2.3.1 Instrumental and expressive roles. Research in this area has been influenced heavily by the early research and theorizing of sociologist Talcott Parsons (1955), who argued that two roles develop in all small groups including the nuclear family, the instrumental role (task-oriented, getting things done), and the expressive role (providing emotional support to group members), and that women specialize in the expressive role and men in the instrumental role. Support for the notion of instrumental and expressive roles came from a study of 14 all-male groups (Bales and Slater 1955), although Parsons apparently did not see the inconsistency between that research and his arguments about gender-based specialization. A reanalysis of a classic study that is often cited as evidence for gender differences in instrumental and expressive behaviors found that, although men exhibited significantly more task-oriented behaviors than women did, 79 percent of men's behaviors were task-oriented and 63 percent of women's were; that is, for both women and men, the majority of behavior was instrumental (Aries 1996). Several decades later, then, it is clear that, although there is evidence of gender differences in instrumental and expressive behaviors in groups, gender similarities are perhaps as striking. In addition, many features of the task and the situation influence the results (Aries 1996).

2.3.2 Leadership. In mixed-sex groups, men are more likely than women to emerge as leaders, although the magnitude of the difference is modest and the results are too complex to encompass in a single statement about gender differences. Eagly and Karau (1991) meta-analyzed studies of the emergence of leaders, most of which were laboratory studies done with college students. For task-oriented leadership, men emerged more frequently as leaders ($d = .41$). For general leadership measures, men also emerged as leaders more, ($d = .32$). Women, however, were more likely to emerge as leaders when social leadership was assessed, ($d = -.18$).

Research on leadership styles generally distinguishes between democratic and autocratic styles. Eagly and Johnson (1990) meta-analyzed research on gender differences in these styles. They found that women showed a more democratic style than men ($d = -.22$). To make matters more complex, the results were found in laboratory studies but not in field studies in actual organizations. The null findings in the real-world setting may have to do with the tendency for organizations to select leaders with particular styles.

2.4 Communication and Nonverbal Behavior

Two recent volumes have reviewed the existing research on gender differences in communication and nonverbal behavior (Aries 1996, Canary and Dindia 1998). Here I summarize the trends in findings for gender differences in language use, the content of conversations, the evaluation of speakers in relation to gender-stereotyped speech, interruptions, and nonverbal decoding.

2.4.1 Gender and language use. In 1973, Lakoff published an influential paper on gender differences in language use. In it, she argued that women use a number of speech forms—for example, tag questions, hedges, and polite forms—considerably more frequently than men, and that these forms reflect uncertainty and deference. Lakoff theorized, in turn, that this speech style represents women's adaptation to their lack of power in society. The style avoids conflict and appeases others precisely because it is not forceful. For example, 'That was a great game!' is a forceful statement, but when the tag questions is added, 'That was a great game, wasn't it?' it becomes considerably less forceful, according to Lakoff.

Lakoff's thesis contains two empirically testable hypotheses: (a) there are reliable gender differences in the use of tag questions, hedges, and polite forms, with women using more of all three; and (b) these forms have the function of signaling uncertainty and deference. I review the data for each hypothesis in turn.

In regard to tag questions, research findings are inconsistent. Some studies confirm Lakoff's belief that women use more tag questions than men (e.g.,

Carli 1990, Holmes 1984), some find that men use more tag questions (e.g., Cameron et al. 1988) some find no gender differences (Baumann 1976, cited in Aries 1996), and some find that the direction of the gender difference depends on the situation (Holmes 1984, Cameron et al. 1988, Kollock et al. 1985). In short, the evidence does not strongly support Lakoff's assertion about gender differences in the use of tag questions.

The second of Lakoff's hypotheses concerns the meaning and function of tag questions. Do they really signal uncertainty? Holmes's (1984) data indicate that tag questions can serve three functions: expression of uncertainty ('The plane is scheduled to depart at 9:35, isn't it?'), expressions of solidarity ('You're still limping a bit, aren't you?'), and politeness to soften an order or negative comment ('To get tenure, you'll have to improve your teaching, won't you?'). In her data, men actually used tags to express uncertainty more than women did, whereas women were more likely to use tags for the other two functions, which maintain and facilitate conversation. Therefore Lakoff's second hypothesis is not supported, but the data suggest that women's tag questions have different functions such as expressing solidarity and softening a difficult comment.

The pattern of findings for hedges is quite similar to that for tag questions (Aries 1996). A purported hedge such as 'I think' may be a hedge or a statement of confidence depending on several factors, including whether the emphasis is on 'think' or 'I.' For example, in the sentence 'Melbourne is the capital of New Zealand, I think,' the phrase serves as a hedge, indicating uncertainty. In contrast, 'I think Truman was the best president of the century,' with emphasis on 'I,' is a strong statement of confidence. Again, there is little support for Lakoff's assertions.

The theme that there are striking gender differences in language use was taken up again vigorously by Tannen in her books written for a popular audience, such as *You Just Don't Understand: Men and Women in Conversation* (1991). Tannen argued that men's and women's language use is so different that they essentially belong to different linguistic communities, and therefore that male–female communication is as difficult and error-prone as cross-cultural communication. Moreover, she argued that men and women use conversation for different purposes: men to express status and independence and women to establish connectedness and intimacy with the other person. She believes that these differences arise because boys and girls grow up in essentially different cultures because of the gender segregation of childhood play. Aries (1996), however, in a review of the research, found little evidence to support Tannen's propositions.

2.4.2 Self-disclosure. A large research literature exists on self-disclosure in conversations and, in particular, gender differences in self-disclosure. Dindia and Allen (1992) conducted a meta-analysis of these studies. Across 205 studies, women self-disclosed more, but the difference was small ($d = -.18$). Patterns of self-disclosure depend on the sex of the target. There is no gender difference in self-disclosure to a male target ($d = .00$), whereas the gender difference is more substantial for a female target ($d = -.35$). Patterns also depend on the relationship between the two people engaged in the conversation. The gender difference in self-disclosure is trivial when the conversation is between strangers ($d = -.07$) and is found mainly when the conversation is with a friend ($d = -.28$), a spouse ($d = -.22$), or a parent ($d = -.26$).

2.4.3 Evaluation of speakers. If college students are asked to prepare and deliver a speech and the speeches are transcribed so that they can be rated blind to the sex of the speaker, women's speeches tend to be rated higher on sociointellectual status and esthetic quality and men's speeches are rated higher on dynamism (e.g., Mulac and Lundell 1980). This has been termed the gender-linked language effect and has been found as early as fourth grade (Mulac et al. 1990).

A large body of research in social psychology has demonstrated that the stereotypes we hold influence the impression that we form of an individual. Research has shown repeatedly that people hold stereotypes about women's and men's speech (Aries 1996). For example, people expect men to interrupt more than women (Hawkins 1988). By sixth grade, children's stereotypes of male and female speech agree with adults' (Edelsky 1976).

The evidence indicates that gender stereotypes do influence raters' evaluations of men's and women's speech. In one study, participants listened to tape recordings of male and female speakers, who in fact used tag questions equally; participants perceived the women to have used more tag questions than the men (Newcombe and Arnkoff 1979). In short, people perceive women's and men's speech to confirm gender stereotypes even when the actual speech does not.

Are women and men evaluated differently, depending on whether they display male-stereotyped or female-stereotyped speech? Erickson et al. (1978) created two scripts, in the context of a courtroom trial setting; in one, the speaker used powerful or male-stereotyped speech and in the other, powerless or female-stereotyped speech. Raters were led to believe that the speaker was either male or female. The results indicated that, regardless of the speaker's sex, raters found those who used more powerful, male-stereotyped language to be more credible and more attractive. Research using such a paradigm has consistently found that, regardless of the sex of the speaker, powerless, female-stereotyped language is

viewed more negatively and powerful, male-stereotyped language more positively on most dimensions (Carli 1990).

2.4.4 Interruptions. Research in the 1970s found that, in mixed-sex conversations, men interrupted women considerably more than the reverse (Zimmerman and West 1975). This pattern was found both in natural conversations (Zimmerman and West 1975) and in controlled laboratory situations (West and Zimmerman 1983). These results stimulated much interest; they were widely interpreted as an indication of nonverbal dominance of females by males (Zimmerman and West 1975, Smith-Lovin and Brody 1989).

Recent reviews, however, have challenged both the existence of the gender difference and its interpretation (Aries 1996). First, a review of more recent research as well as a critical examination of the original studies indicates that there are no well-documented gender differences in interruptions. Second, understanding of interruptions has become more sophisticated and it is clear now that, although some interruptions are power-oriented (e.g., cutting off a speaker to change the topic), interruptions can serve many other functions, including expressing support, interest, or agreement.

2.4.5 Smiling. One nonverbal behavior that has consistently shown gender differences is smiling. In an early meta-analytic review, Hall (1984) found that females consistently smile more than males, ($r = .30$), which translates to $d = -.63$, a difference that is in the moderate-to-large range and is noticeably larger than most other gender differences reviewed here. More recent studies continue to document the same phenomenon (Hall 1998). The pattern, though, is more characteristic of White women than African-American women (Halberstadt and Saitta 1987).

Generally this difference is interpreted as indicating the extent to which smiling is a part of the female role. For example, in one study, participants were given a written description of a person, accompanied by a photograph of a man or woman who was smiling or not smiling (Deutsch et al. 1987). Women who were not smiling were given more negative evaluations; apparently people react negatively to this violation of the female role.

2.4.6 Nonverbal decoding. Nonverbal decoding (also termed nonverbal sensitivity) refers to a person's ability to read correctly the nonverbal behaviors or body language of another person. A gender difference favoring females has been documented over several decades. In an early meta-analysis, Rosenthal et al. (1979) located 133 samples that had been administered the Profile of Nonverbal Sensitivity (PONS, which bases scores on responses to short video clips); females scored higher than males in 80 percent of the samples ($d = -.41$), a moderate effect. Later meta-analyses based on a wider variety of measures obtained quite similar results; for example, Hall (1984) found $d = -.52$.

3. Future Directions

Research on gender differences in personality and social behavior must certainly move beyond questions of simple gender differences, to systematic analyses of situation *x* person and situation *x* gender interactions. Which situations call forth gender differences and which do not? How can these situational variations be drawn into a coherent pattern that is theoretically meaningful?

4. Summary

Over a wide range of personality and social variables, there is much evidence of gender similarities. For the Big Five personality dimensions, there is no gender difference in openness to experience ($d = .03$) and small differences in conscientiousness ($-.13$), extraversion or gregariousness ($-.14$), and neuroticism ($-.25$); there is, however, a large difference in one aspect of agreeableness, termed tendermindedness or nurturance ($-.97$). The gender difference in self-esteem is small (.21).

Likewise, there are many gender similarities in the realm of social behavior. The direction of the gender difference in helping behavior depends largely on the situation. In small groups, men and women are similar in their instrumental and expressive behaviors. The gender difference in democratic vs. autocratic leadership style is small ($-.22$ for democratic style). The research evidence does not support popular media claims about enormous gender differences in language use. The gender difference in self-disclosure is small ($-.18$). There are several exceptions to this pattern of gender similarities. Gender differences are moderately large for aggression (.50), smiling ($-.63$), and sensitivity to nonverbal cues ($-.52$).

See also: Gender and Feminist Studies in Psychology; Gender and Language: Cultural Concerns; Gender and School Learning: Mathematics and Science; Gender Identity Disorders; Gender-related Development; Language and Gender; Masculinities and Femininities; Small-group Interaction and Gender; Social Networks and Gender

Bibliography

Aries E 1996 *Men and Women in Interaction*. Oxford University Press, New York

Bales R F, Slater P E 1955 Role differentiation in small decision-making groups. In: Parsons T, Bales R F (eds.) *Family, Socialization, and Interaction Process*. Free Press, Glencoe, IL

Cameron D, McAlinden F, O'Leary K 1988 Lakoff in context: The social and linguistic functions of tag questions. In: Coates J, Cameron D (eds.) *Women in Their Speech Communities: New Perspectives on Language and Sex*. Longman, New York

Canary D J, Dindia K (eds.) 1998 *Sex Differences and Similarities in Communication*. Erlbaum, Mahwah, NJ

Carli L L 1990 Gender, language, and influence. *Journal of Personality and Social Psychology* **59**: 941–51

Costa P T, McCrae R R 1985 *The NEO personality inventory: Manual Form S and Form R*. Psychological Assessment Resources, Odessa, FL

Deutsch F M, LeBaron D, Fryer M M 1987 What is in a smile? *Psychology of Women Quarterly* **11**: 341–51

Dindia K, Allen M 1992 Sex differences in self-disclosure: A meta-analysis. *Psychological Bulletin* **112**: 106–24

Eagly A H, Crowley M 1986 Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin* **100**: 283–308

Eagly A H, Johnson B T 1990 Gender and leadership style: A meta-analysis. *Psychological Bulletin* **108**: 233–56

Eagly A H, Karau S J 1991 Gender and the emergence of leaders: A meta-analysis. *Journal of Personality and Social Psychology* **60**: 685–710

Eagly A H, Steffen V J 1986 Gender and aggressive behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin* **100**: 309–30

Edelsky C 1976 The acquisition of communicative competence: Recognition of linguistic correlates of sex roles. *Merrill-Palmer Quarterly* **22**: 47–59

Eisenberg N, Fabes R A 1998 Prosocial development. In: Damon W (ed.) *Handbook of Child Psychology*, 5th ed. J Wiley, New York, pp. 701–78

Erickson B, Lind E A, Johnson B C, O'Barr W M 1978 Speech style and impression formation in a court setting: The effects of 'powerful' and 'powerless' speech. *Journal of Experimental Social Psychology* **14**: 266–79

Feingold A 1994 Gender differences in personality: A meta-analysis. *Psychological Bulletin* **116**: 429–56

Halberstadt A G, Saitta M B 1987 Gender, nonverbal behavior, and perceived dominance: A test of the theory. *Journal of Personality and Social Psychology* **53**: 257–72

Hall J A 1984 *Nonverbal Sex Differences*. Johns Hopkins University Press, Baltimore, MD

Hall J A 1998 How big are nonverbal sex differences? The case of smiling and sensitivity to nonverbal cues. In: Canary D J, Dindia K (eds.) *Sex Differences and Similarities in Communication*. Erlbaum, Mahwah, NJ, pp. 155–77

Hawkins K 1988 Interruptions in task-oriented conversations: Effects of violations of expectations by males and females. *Women's Studies in Communication* **11**: 1–20

Holmes J 1984 Women's language: A functional approach. *General Linguistics* **24**: 149–78

Hyde J S 1986 Gender differences in aggression. In: Hyde J S, Linn M C (eds.) *The Psychology of Gender: Advances Through Meta-analysis*. Johns Hopkins University Press, Baltimore, MD, pp. 51–66

Kling K C, Hyde J S, Showers C J, Buswell B N 1999 Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin* **125**: 470–500

Kollock P, Blumstein P, Schwartz P 1985 Sex and power in interaction: Conversational privileges and duties. *American Sociological Review* **50**: 34–46

Lakoff R 1973 Language and woman's place. *Language in Society* **2**: 45–79

Lynn R, Martin T 1997 Gender differences in extraversion, neuroticism, and psychoticism in 37 nations. *Journal of Social Psychology* **137**: 369–73

Major B 1999 Gender and self-esteem: A meta-analysis. In: Swann Jr W B, Langlois J H, Gilbert L A (eds.) *Sexism and Stereotypes in Modern Society: The Gender Science of Janet Taylor Spence*. American Psychological Association, Washington, DC, pp. 223–53

Martin T, Kirkcaldy B 1998 Gender differences on the EPQ-R and attitudes to work. *Personality and Individual Differences* **24**: 1–5

McCrae R R, Costa P T 1987 Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology* **52**: 81–90

Mulac A, Lundell T L 1980 Differences in perceptions created by syntactic-semantic productions of male and female speakers. *Communication Monographs* **47**: 111–18

Mulac A, Studley L B, Blau S 1990 The gender-linked language effect in primary and secondary students' impromptu essays. *Sex Roles* **23**: 439–69

Newcombe N, Arnkoff D B 1979 Effects of speech style and sex of speaker on person perception. *Journal of Personality and Social Psychology* **37**: 1293–303

Parsons T 1955 The American family: Its relations to personality and to the social structure. In: Parsons T, Bales R F (eds.) *Family, Socialization, and Interaction Process*. Free Press, Glencoe, IL

Roberts J E, Gotlib I H 1997 Lifetime episodes of dysphoria: Gender, early childhood loss and personality. *British Journal of Clinical Psychology* **36**(Pt. 2): 195–208

Rosenthal R, Hall J A, DiMatteo M R, Rogers P L, Archer D 1979 *Sensitivity to nonverbal communication: The PONS test*. Johns Hopkins University Press, Baltimore, MD

Smith-Lovin L, Brody C 1989 Interruptions in group discussions: The effects of gender and group composition. *American Sociological Review* **54**: 424–35

Tannen D 1991 *You Just Don't Understand: Women and Men in Conversation*. Ballantine, New York

Watson D, Clark L A 1984 Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin* **96**: 465–90

West C, Zimmerman D H 1983 Small insults: A study of interruptions in cross-sex conversations between unacquainted person. In: Thorne B, Kramarae C, Henley N (eds.) *Language, Gender, and Society*. Newbury House, Rowley, MA

Zimmerman D H, West C 1975 Sex roles, interruptions and silences in conversation. In: Thorne B, Henley N (eds.) *Language and Sex: Differences and Dominance*. Newbury House, Rowley, MA

Zonderman A B, Siegler I C, Barefoot J C, Williams R B, Costa P T 1993 Age and gender differences in the content scales of the Minnesota Multiphasic Personality Inventory. *Experimental Aging Research* **19**: 241–57

J. S. Hyde

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Gender, Economics of

1. Gender and the Economics of Gender

The term 'gender' traditionally has referred, as has sex, to the biological differences between men and women. In the 1980s, however, increasingly it came to be used to refer instead to the distinctions which society has erected on this biological base. This article describes the relationship of this concept of gender to economics, focusing on economic analyses of gender differences in behavior and economic outcomes. This focus may be differentiated from feminist economics that is a critique of the discipline and practice of economics from a feminist perspective (see *Feminist Economics*).

2. Economists' Increasing Interest in Gender Issues

Historically, gender has not been perceived to be a central concept in economic analysis. However, as the force of current events has thrust gender-related issues to the fore, economists have responded. The outcome has been not only a better understanding of the nature and causes of gender differences, but also an enrichment of the discipline itself.

One reason for the historical neglect of gender issues by the economics discipline is that, traditionally, economists focused on market transactions and, traditionally, most married women were primarily engaged in nonmarket activities. As women have come to comprise a larger share of the paid labor force and as market work has loomed larger in the typical woman's life, interest in the determinants of gender differences in behavior and labor market outcomes has grown. At the same time, the growing interest of economists in analyzing nonmarket work and demographic outcomes such as marriage, divorce and fertility also played an important role in increasing economists' attention to gender-related issues.

While early economic analyses paid scant attention to gender issues, the nineteenth century campaign for women's rights and female suffrage did focus some attention on gender inequality. With few married women employed outside the home, observers like American feminist, Gilman (1898/1966) and Marxist, Engels (1884/1972) (see *Marxian Economic Thought*) focused on the gender division of labor itself, espousing the emancipating effects of women's participation in market work. Classical economist Mill (1869/78) eloquently argued for women's 'admissibility to all the functions and occupations hitherto retained as the monopoly of the stronger sex' and expressed the belief that women's 'disabilities elsewhere are only clung to in order to maintain their subordination in domestic life.' These views highlight the notion that understanding women's status involves an understanding of both the market and family spheres and of the interrelationship between the two. This insight remains applicable to the current day.

3. Time Allocation in the Family Context

Prompted in part by their desire to understand the causes of the rising labor force participation of married women in the post-World War II period, economists extended the traditional theory of labor supply (see *Labor Supply*) to consider household production.

3.1 The Traditional Theory of Labor Supply

The traditional theory of labor supply was a simple extension of consumer theory. Individuals maximize their utility, which is derived from market goods and leisure, subject to budget and time constraints. The labor supply decision is determined by the opportunity wage (the wage rate the individual could earn in the labor market), nonlabor income (the amount of income available to them from other sources), and the individual's tastes or preferences for work vs. leisure.

3.2 Household Production and the Allocation of Time

While this theory is sufficient for many purposes, it has substantial limitations for understanding the determinants of the gender division of labor in the family and the factors influencing women's labor force participation. The key to addressing these issues is a fuller understanding and analysis of the household production process.

An early pioneer in analyzing household production was Reid (1934), but it was the work of Becker and Mincer in the 1960s that was particularly influential in bringing these issues into the mainstream. In analyzing married women's labor force participation in the US, Mincer (1962) pointed to the importance for them of the three-way decision among market work, nonmarket work, and leisure. He attributed the growth in married women's participation during the early postwar period to women's rising real wages which induced them to substitute market work for housework. This response was large enough to more than offset the negative income effect on their labor supply associated with the growth in their husbands' real wages.

Becker (1965) advanced this process further by proposing a general theory of the allocation of time. In this and other work (summarized in Becker 1981), he laid the foundations of the 'new home economics,' including economic analyses of time allocation, marriage, divorce, and fertility. Interestingly, while Mincer opened a window on household production by distinguishing nonmarket work from leisure, Becker was

able to provide a further advance by again eliminating the distinction. However, while in the traditional labor supply model all nonmarket time is spent in leisure, in Becker's model all nonmarket time is spent in household production. Becker assumes that households derive utility from 'commodities' that are produced by inputs of market goods and nonmarket time.

These ideas may be illustrated by considering the relationship between children and women's labor force participation. Young children may be viewed as a 'time-intensive' commodity. Traditionally, it has been the mother who was the primary care giver. And, while it is possible to substitute market goods and services (e.g., babysitters, day care centers) for home time in caring for children, these alternative production techniques tend to be costly and may be difficult to implement. Thus, at any point in time, the probability that a woman will participate in the labor force is expected to be inversely related to the number of small children present. Over time, the increase in women's participation rates has been associated with decreasing birthrates, as well as increases in the availability of alternative child care arrangements.

3.3 Trends in Labor Supply

Continuing in the tradition of Mincer (1962), the labor supply model has been used to explain the trends in female labor force participation. One criticism that has been made of this type of economic modeling is that it abstracts from historical and institutional factors. Goldin (1990), however, turns this feature of economic models to her advantage in an especially innovative application. She finds that a model applied to decennial US Census data for the 1890–1980 period overpredicts married women's labor force participation for 1930, 1940, and 1950. She presents this as evidence that the process of change in participation was retarded for married women by institutional barriers like marriage bars (i.e., prohibitions against employing married women that were particularly prevalent in teaching and clerical work) and the lack of availability of part-time employment.

As wage inequality rose in many advanced countries in the last quarter of the twentieth century, increasing attention focused on the contribution of this trend to the continued rise in the labor force participation of wives and the related question of the consequences of the increased employment of wives for family income inequality. On one hand, the income effect suggests that wives of low skilled men would be drawn into the labor force to compensate for the deteriorating labor market position of their husbands. On the other hand, the substitution effect implies that the increasing rewards to labor market skills which underlie the rise in wage inequality, would draw more skilled (e.g., more highly educated) women into the labor market. Since more highly educated individuals tend to marry each other, a larger role for the income effect in explaining the trends would tend to reduce the correlation between wives' and husbands' earnings, while a larger role for the substitution effect would tend to increase the correlation.

Juhn and Murphy (1997) found for the US that employment gains were in fact largest for wives of middle and high wage men. Their findings strikingly mirror Mincer's (1962) for the immediate post-World War II period: own wage effects dominate cross-effects between husband and wife in accounting for changes in female employment. The correlation between husbands' and wives' earnings has risen in the US, both as a consequence of these participation trends and also due to a rise in the correlation between the earnings of spouses when both are employed. A similar rise has been reported for Britain, which also experienced a substantial increase in wage inequality during this time. However, while this increased correlation contributed to rising inequality in family income in both the US and Britain, the rise in family income inequality among married couples in both countries was principally due to rising inequality of husbands' earnings (Cancian and Reed 1998, Machin and Waldfogel 1994).

3.4 The Gender Division of Labor

In the discussion of the allocation of time above, it was simply assumed that women tend to bear the primary responsibility for childcare and housework. However, the gender division of labor in the family is also an issue that the new home economics addresses. According to Becker (1981), the division of labor will be dictated by comparative advantage. To the extent that women have a comparative advantage in household and men in market production, it will be efficient for women to specialize at least to some extent in the former while men specialize in the latter. In Becker's view, the increased output resulting from this arrangement maximizes family utility and constitutes one of the primary economic benefits to marriage. Thus, women's rising labor force participation is expected to reduce the gains from marriage, thereby contributing to the trend towards higher divorce and lower marriage rates.

As women have been increasingly working throughout the life cycle, more attention has focused on their ability to combine work and family responsibilities successfully. Since men tend not to face the same dilemmas as women, or not to face them to the same extent, this is a potential source of gender differences in wellbeing. Waldfogel (1998) details the prevalence of a 'family penalty' in many advanced countries, that is, lower earnings for women who are mothers, all else equal. This likely reflects the loss of their tie to the firm and associated firm-specific skills when they withdraw from the labor force upon the birth of a child.

Some evidence suggests that policy intervention could reduce the size of the family penalty. For example, Waldfogel (1998) reports evidence from analyses of microdata for the US, Britain, and Japan indicating that women who were covered by maternity leave were considerably more likely to return to their prior employer after childbirth. And women who return to their employer after a birth have higher subsequent wages. Consistent with these results, Ruhm (1998) finds, using aggregate data on 16 countries during the 1969–88 period, a positive, but not significant, effect of 1 percent on women's wages relative to men's of short leave entitlements (1–13 weeks), although longer leave entitlements (over 26 weeks) had a significantly negative effect of 3 percent. The negative effect of longer leaves could indicate that such leaves encourage women to extend their time away from the firm. It is also possible that the higher costs of long mandated leaves provide greater incentives to employers to discriminate against women in wages or job assignments.

A consideration of the difficulties women face in combining work and family raises the possibility of conflicts of interest between husband and wife which in turn makes defining 'family utility' problematic (Ferber and Birnbaum 1977). Such concerns have been addressed by the growing application of bargaining models to family decision-making, led by Manser and Brown (1980) and McElroy and Horney (1981). These models assume that the preferences of husbands and wives may diverge and that outcomes for each spouse can be modeled using game theory (see *Game Theory: Cooperative Games*; *Game Theory: Noncooperative Games*). In the most popular application of these models to the family, the cooperative bargaining model of marriage, the utility received by each spouse in the Nash bargaining solution is related positively to the utility of each partner at the 'threat point,' most commonly taken to be the utility that each would receive if the marriage irretrievably broke down. If it is assumed that, in the event of divorce, each partner maintains ownership of income received separately within marriage, a woman who works outside the home will have a higher utility at the threat point and hence a more favorable distribution within marriage. Studies reviewed by Lundberg and Pollak (1996) provide empirical support for this model in their finding that earned or unearned income received by the wife has a different effect on demand patterns than income received by the husband.

As women anticipate spending increasingly more of their adult lives in the labor market, their investments in market-oriented human capital may be expected to continue to grow and their comparative advantage in home work to decline. Moreover, as the quality of the opportunities open to women in the labor market continues to improve, the disadvantages of specialization in home work in the form of forgone earnings and possibilities for career advancement will also rise. Thus, greater sharing of household responsibilities between men and women is likely to become increasingly prevalent.

4. *Gender Differences in Labor Market Outcomes*

Women earn less than men, on average, in all advanced economies, though the extent of the gender pay gap varies considerably across countries and has been declining in many of them in recent years (e.g., Blau and Kahn 1996). Economists' understanding of gender inequality in such economic outcomes has been accommodated principally through the development of new and interesting applications of existing theoretical approaches. Such theories must explain the considerable employment segregation of men and women—by occupation, industry, and firm—as well as gender pay differentials. Segregation of men and women by occupation, for which summary data are most readily available, has been found to be prevalent in all advanced economies, though it has been declining in recent decades in most OECD countries, including the US (e.g., Blau et al. 1998, Anker 1998).

As in the case of the analysis of women's roles in the family, the catalyst for the development of analyses of pay and occupation differences was provided by external events. Some moderate degree of interest in the pay issue was generated in England by the World War I experience when there was some substitution of women into traditionally male civilian jobs, although not nearly to the degree that would later occur during World War II. Questions of the appropriate pay for women under these circumstances arose and stimulated economic analyses of the gender pay differential. These included the work of Fawcett (1918) and Edgeworth (1922), both of whom assigned a prominent causal role to occupational segregation, thus providing the antecedents for Bergmann's (1974) overcrowding model (discussed below). The analysis of gender differentials in the labor market received further impetus in the early 1960s, this time in the US, with the development of the women's liberation movement and the passage of antidiscrimination legislation.

Two broad approaches to the issue have since evolved. First is the human capital explanation which lays primary emphasis on men's and women's own choices in explaining occupation and pay differences. Second are a variety of models of labor market discrimination (see *Discrimination, Economics of*) which share the common characteristic of placing the onus for unequal outcomes on differential treatment of equally (or potentially equally) qualified men and women in the labor market. More recently, some advances have been made by considering the gender pay gap and other demographic differentials in the context of the overall structure of wages (see *Wage Differentials and Structure*).

4.1 The Human Capital Explanation

The human capital explanation for gender differences in occupations and earnings, developed by Mincer and Polachek (1974) and others, provides an explanation for gender differences in market outcomes on the basis of the traditional division of labor by gender in the family. Anticipating shorter and more discontinuous work lives as a consequence of their family responsibilities, women have lower incentives to invest in market-oriented formal education and acquire less on-the-job training than men. Their resulting smaller human capital investments will lower their earnings relative to those of men. Working in a similar direction is Becker's (1985) model in which the longer hours women spend on housework lowers the effort they put into their market jobs compared to men, controlling for hours worked, and hence reduces their wages.

These considerations are also expected to produce gender differences in occupations as women choose occupations for which human capital investments are less important and in which the skill depreciation that occurs during time spent out of the labor force is minimized. Women especially will avoid jobs requiring large investments in firm-specific skills (i.e., skills which are unique to a particular enterprise), because the returns to such investments are reaped only as long as one remains with a particular employer. At the same time, employers are expected to be reluctant to hire women for such jobs because they bear some of the costs of firm-specific training (see the discussion of statistical discrimination below).

To the extent that gender differences in outcomes are not accounted for fully by productivity differences derived from these and other sources, models of labor market discrimination offer an explanation.

4.2 Models of Labor Market Discrimination

Theoretical work in this area was initiated by Becker's (1957) model of racial discrimination. Becker conceptualized discrimination as a taste and analyzed three cases, those in which the discriminatory tastes were held by employers, co-workers, and customers or clients. Under certain circumstances, such discrimination will cause a wage differential between men and women. So discriminatory employers will only hire women at a sufficient wage discount that compensates them for the disutility of employing women. Discriminatory male workers will demand a wage premium to work with women thus raising men's relative wages, and the reluctance of discriminatory customers or clients to buy goods or services provided by women will make women less productive in terms of revenue brought in, thus depressing women's relative wages.

Models based on tastes for discrimination are also consistent with employment segregation, but do not necessarily predict it will occur. If wages are flexible, it is altogether possible that discrimination will result in lower pay for women, but produce little or no segregation. However, if discriminatory tastes against women in traditionally male pursuits are both strong and prevalent, women may tend to be excluded from these areas. If such segregation does occur, it may or may not be associated with gender pay differentials. In the presence of sufficient employment opportunities in the female sector, equally qualified women may earn no less than men.

The relationship between occupational segregation and earnings differentials is clarified in Bergmann's (1974) overcrowding model. If, for whatever reason (i.e., labor market discrimination or their own choices), potentially equally qualified men and women are segregated by occupation, the wages in male and female jobs will be determined separately by the supply and demand for labor in each sector. Workers in male jobs will enjoy a relative wage advantage if the supply of labor is more abundant relative to demand for female than for male occupations. Such 'crowding' of female occupations can also widen differentials between male and female jobs that would otherwise exist due to women's smaller human capital investments or employers' reluctance to invest in women's human capital.

One question that has been raised about the Becker model is how employer discrimination can persist in the long run in the face of competitive forces. The least discriminatory firms, which hire more lower-priced female labor, would have lower costs of production. In the long run, they should expand and drive the more discriminatory firms out of business. One answer, suggested initially by Becker himself, is that discrimination will be located in sectors of the economy that are not competitive (see *Market Structure and Performance*). While Becker emphasized monopolistic elements in the product market, a related approach targets monopsonistic power on the part of the employer in the labor market (e.g., Madden 1973).

In addition, models of statistical discrimination were later developed, in part to explain the persistence of discrimination in the long run in the face of competitive forces. Such models assume a world of uncertainty and imperfect information; thus differences between groups in the expected value of productivity or in the reliability with which productivity may be predicted are expected to result in differences in the treatment of members of each group. So, for example, employers may believe that women are more likely to quit their jobs than men with similar observed characteristics (education, experience, etc.) or they may believe that women's turnover can be less reliably predicted by their measured characteristics than men's. As a consequence, they may pay women less, exclude them from jobs requiring substantial firm-specific training, or deny them promotions.

It has been argued that such statistical discrimination, making decisions on the basis of the average characteristics of the group, is consistent with profit

maximization and can thus persist in the face of competitive forces. However, Aigner and Cain (1977) contend that such models are no more convincing in explaining the persistence of discrimination than models based on tastes for discrimination. To the extent that employers' views are accurate, the lower expected productivity of women will reduce their wages but women as a group will be paid their expected productivity. This does not constitute labor market discrimination as economists define it, that is, pay differences that are not accounted for by productivity differences. Presumably they would make a similar argument regarding hiring and promotion decisions. Moreover, they note that when employer beliefs regarding average differences are erroneous, discrimination clearly exists but in their view discrimination based on such misperceptions is even less likely to persist in the long run than discrimination based on tastes. Finally, though they acknowledge that less reliable predictions of a group's productivity combined with risk aversion by employers could produce a discriminatory differential, a perfectly elastic supply of risk neutral entrepreneurs could be expected to erode discriminatory differentials based on this factor.

However, such models may have greater usefulness in explaining the persistence of gender differentials in pay than Aigner and Cain's analysis suggest. Consider first the situation where employer perceptions are correct. In this case, the concept of statistical discrimination provides a plausible reason for the existence and persistence of pay and occupational differences between men and women with similar measured characteristics; this is in itself a useful contribution even if it does not shed light on the persistence of economic discrimination in the long run. But if the employer views are correct, is it appropriate to consider this a form of 'discrimination' in any sense? From a normative perspective, the answer may be yes, to the extent that basing employment decisions on a characteristic like sex—a characteristic that the individual cannot change—could be viewed as inequitable. Indeed, the practice of judging an individual on the basis of group characteristics rather than upon his or her own merits seems the very essence of stereotyping or discrimination. Such behavior is certainly not legal, for example, under the antidiscrimination laws and regulations.

Now consider the situation where employer perceptions are incorrect. If statistical discrimination is accompanied by feedback effects, this may be a more credible source of persistent discriminatory pay differences between men and women than Aigner and Cain suggest. For example, if employers incorrectly expect that women are more likely to quit their jobs, they may respond by giving women less firm-specific training or assigning them to dead end jobs. Faced with few incentives to remain on the job, women may respond by exhibiting exactly the higher turnover that employers expect.

This criticism of the Becker model—its inability to explain the persistence of discrimination against women in the long run—has proved to be a double-edged sword in that it has led some economists to doubt that labor market discrimination is responsible, in whole or part, for observed gender differences in labor market outcomes. Yet it is important to recognize that discrimination is an intrinsically complex process. From this perspective, it is not surprising that no easy solution has been found to the question of why discrimination may persist in the face of competitive forces. Similarly, the various models of discrimination, each emphasizing different motivations and different sources of this behavior, need not be viewed as alternatives. Rather, each may serve to illuminate different aspects of this complex reality. When it is further recognized that, as our discussion above suggests, discrimination can adversely affect women's human capital investments and labor force attachment by lowering the market rewards to this behavior (e.g., Weiss and Gronau 1981), it is clear that the potential impact of discrimination on women's labor market outcomes may be large indeed.

4.3 Evidence on Human Capital, Discrimination, and the Gender Pay Gap

The typical approach to analyzing the sources of the gender pay gap is to estimate wage regressions specifying the relationship between wages and productivity-related characteristics for men and women (see *Econometric Software*). The gender pay gap may then be decomposed statistically into two components: the first, due to gender differences in measured characteristics and the second, which is 'unexplained' and potentially due to discrimination.

Such studies provide considerable empirical support for both human capital differences and labor market discrimination in explaining the gender pay gap. For example, Blau and Kahn (1997) report that, in 1988, US women's wages were 72 percent of men's, on average. Adjusting for gender differences in human capital (i.e., education and experience) would raise the ratio to 81 percent, while additionally adjusting for gender differences in industry, occupation and unionism would further increase it to 88 percent. Thus, while measured characteristics are important, women still earn considerably less than similar men, even when all measured characteristics are taken into account.

One problem with studies of this type is that discrimination may be overestimated to the extent that unexplained pay differentials between men and women are due in whole or part to gender differences in unmeasured qualifications. On the other hand, to the extent that the regression models include controls for characteristics like industry or occupation that could themselves be due to discrimination, discrimination may be underestimated. Recent studies that take a

different approach to the question provide additional evidence of discrimination that addresses the former concern.

First, studies have applied traditional econometric techniques to especially homogeneous groups and employed extensive controls for qualifications, thus minimizing the effect of gender differences in unmeasured characteristics; and still found evidence of an unexplained pay difference between men and women (see, e.g., Wood et al. 1993) which focused on graduates of the University of Michigan Law School).

A second set of studies uses an experimental approach. Neumark (1996) analyzed the results of a hiring 'audit' in which male and female pseudo-job seekers were given similar résumés and sent to apply for waiter or waitress jobs at the same set of Philadelphia restaurants. The results provided statistically significant evidence of discrimination against women in high-priced restaurants where earnings of workers are generally higher than in the other establishments. A second study, Goldin and Rouse (2000), examined the impact of the adoption of 'blind' auditions by symphony orchestras in which a screen is used to conceal the identity of the candidate. Using data from actual auditions, it was found that the screen substantially increased the probability that a woman would advance out of preliminary rounds and be the winner in the final round.

Finally, there is some evidence that gender differentials in wages or employment are larger in less competitive industries. For example, Black and Strahan (1999) report that with the deregulation of the banking industry beginning in the mid-1970s, the gender gap in pay in banking declined. This is consistent with the expectation that discrimination will be more severe in firms that are shielded to some extent from competitive pressures and thus indirectly provides support for the view that discrimination is a factor influencing gender wage differentials.

4.4 *The Role of Wage Structure*

Both the human capital and discrimination explanations for the gender pay gap might be considered 'gender specific' in nature, that is, focused on gender differences in either qualifications or treatment. Analyses of gender differentials traditionally have tended to emphasize these types of gender-specific factors. Important work by Juhn et al. (1991) on trends in the race pay gap over time in the US, has highlighted an additional factor, overall wage structure, which is also relevant to analyzing the gender pay gap. Wage structure is the array of prices determined for labor market skills and the rewards to employment in particular sectors.

The human capital model suggests that men and women tend to have different levels of labor market qualifications (especially work experience) and to be employed in different occupations and perhaps in different industries. Discrimination models, too, suggest that women may be segregated into different sectors of the labor market. This implies that the returns to skills and the size of premiums for employment in particular sectors potentially play an important role in determining the pay gap. All else equal, the larger the returns to skills and the larger the rents received by individuals in favored sectors, the larger will be the gender pay gap.

Blau and Kahn (1996) have applied this insight to understanding international differences in the gender pay gap in order to address a paradox. The relative qualifications of American women are high compared to women in other countries, and the US has had a longer and often stronger commitment to antidiscrimination laws than most economically advanced nations. Yet the US traditionally has been among the countries with the largest gender pay gaps. The resolution to this paradox lies in the decentralized system of wage setting in the US that has resulted in a highly dispersed wage distribution and high returns to labor market skills and employment in favored sectors. In contrast, the greater role of unions and government in the labor markets of other economically advanced nations has resulted in higher wage floors and more compressed wage distributions from which employed women, as a relatively low wage group, have benefited disproportionately.

This raises the question of how, given the rising wage inequality that has occurred in many of the advanced industrialized countries, women have nonetheless been able to narrow the wage gap with men in many cases. A detailed study of the US by Blau and Kahn (1997) suggests that women have been able to compensate for the adverse effects of rising inequality and higher rewards to skills by improving their relative qualifications, particularly their average work experience and their occupational distribution. In addition, the unexplained portion of the pay gap declined, reflecting either an upgrading of women's unmeasured labor market skills or a decline in discrimination against them, or both.

Similar studies of the UK (Dolton et al. 1996) and Sweden (Edin and Richardson 1999) also find that gender specific factors were principally responsible for declines in the gender gap in those countries over the 1970s, although during a subperiod in which considerable wage compression occurred in Sweden due to its 'solidarity' wage policy (1968–71), wage structure played a major role.

5. *Conclusion*

This review suggests that women's economic roles in the family and their labor market outcomes are closely related. This has undoubtedly played a role in the

persistence of gender differences in both spheres. However, increasing opportunities for women in the labor market may be expected to reduce gender differences in family roles, while such a diminution of adherence to the traditional division of labor in the family is likely to enhance women's position in the labor market further. Thus, just as a fuller understanding of the interrelationships between women's roles in the family and their status in the labor market helps us to understand the persistence of gender inequality in economic outcomes, it also enables us to appreciate how changes in either one of these spheres, or both, can induce a mutually reinforcing process of change. Recent reductions in the pay gap in many of the economically advanced countries may well signal the operation of such a process.

See also: Comparable Worth in Gender Studies; Economic Development and Women; Feminist Economics; Gender and Feminist Studies in Economics; Gender and Technology; Sex Differences in Pay; Sex Segregation at Work; Wage Differentials and Structure

Bibliography

Aigner D J, Cain G G 1977 Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review* **30**: 175–87

Anker R 1998 *Gender and Jobs: Sex Segregation of Occupations in the World*. International Labour Office, Geneva

Becker G S 1957 *The Economics of Discrimination*, 2nd edn. University of Chicago Press, Chicago

Becker G S 1965 A theory of the allocation of time. *Economic Journal* **75**: 493–517

Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Becker G S 1985 Human capital, effort, and the sexual division of labor. *Journal of Labor Economics* **3**: 533–58

Bergmann B R 1974 Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal* **1**: 103–10

Black S E, Strahan P E 1999 Rent-sharing and discrimination: the effects of deregulation on the labor market. Working paper, Federal Reserve Bank of New York

Blau F D, Ferber M A, Winkler A E 1998 *The Economics of Women, Men, and Work*, 3rd edn. Prentice-Hall, NJ

Blau F D, Kahn L M 1996 Wage structure and gender earnings differentials: An international comparison. *Economica* **63**: S29–S62

Blau F D, Kahn L M 1997 Swimming upstream: Trends in the gender wage differential in the 1980s. *Journal of Labor Economics* **15**: 1–42

Cancian M, Reed D 1998 Assessing the effects of wives' earnings on family income inequality. *The Review of Economics and Statistics* **80**: 73–9

Dolton P, O'Neill D, Sweetman O 1996 Gender differences in the changing labor market: The role of legislation and inequality in changing the wage gap for qualified workers in the United Kingdom. *Journal of Human Resources* **31**: 549–65

Edgeworth F Y 1922 Equal pay to men and women for equal work. *Economic Journal* **32**: 431–57

Edin P-A, Richardson K 1999 Swimming with the tide: Solidarity wage policy and the gender earnings gap. Working Paper, IFAU-Office of Labour Market Policy Evaluation. Uppsala, Sweden

Engels F 1884/1972 *The Origin of the Family. Private Property and The State*. International Publishers, New York

Fawcett M G 1918 Equal pay for equal work. *Economic Journal* **28**: 1–6

Ferber M A, Birnbaum B G 1977 The new home economics: Retrospects and prospects. *Journal of Consumer Research* **4**: 19–28

Gilman C P 1898/1966 *Women and Economics: A Study of the Economic Relation Between Men and Women as a Factor of Social Evolution*. Small, Maynard, Boston

Goldin C 1990 *Understanding the Gender Gap*. Oxford University Press, New York

Goldin C, Rouse C 2000 Orchestrating impartiality: The impact of 'blind' auditions on female musicians. *American Economic Review* **90**: 715–41

Juhn C H, Murphy K M 1997 Wage inequality and family labor supply. *Journal of Labor Economics* **15**: 72–97

Juhn C H, Murphy K M, Pierce B 1991 Accounting for the slowdown in black-white wage convergence. In: Kosters M (ed.) *Workers and Their Wages*. AEI Press, Washington, DC

Lundberg S, Pollak R A 1996 Bargaining and distribution and marriage. *Journal of Economic Perspectives* **10**: 139–58

Machin S, Waldfogel J 1994 The decline of the male breadwinner: Changing shares of husbands' and wives' earnings in family income. London School of Economics Suntory-Toyota International Centre for Economics and Related Disciplines Working Paper: WSP/103

Madden J F 1973 *The Economics of Sex Discrimination*. Heath and Co., Lexington, MA

Manser M, Brown M 1980 Marriage and household decision-making. *International Economic Review* **21**: 31–44

McElroy M B, Horney J M 1981 Nash-bargained household decisions: toward a generalization of the theory of demand. *International Economic Review* **22**: 333–49

Mill J S 1869/1878 *The Subjection of Women*, 4th edn. Longmans, Green, Reader and Dyer London

Mincer J 1962 In: *Aspects of Labor Economics*. National Bureau of Economic Research, Princeton University Press, NJ

Mincer J, Polachek S 1974 Family investments in human capital: Earnings of women. *Journal of Political Economy* **82**: S76–SI08

Neumark D, Bank R J, Van Nort K O 1996 Sex discrimination in restaurant hiring: an audit study. *Quarterly Journal of Economics* **111**: 915–41

Reid M G 1934 *Economics of Household Production*. Wiley, New York

Ruhm C J 1998 The economic consequences of parental leave mandates: Lessons from Europe. *Quarterly Journal of Economics* **113**: 285–317

Waldfogel J 1998 Understanding the 'family gap' in pay for women with children. *Journal of Economic Perspectives* **12**: 137–57

Weiss Y, Gronau R 1981 Expected interruptions in labor force participation and sex-related differences in earnings growth. *Review of Economic Studies* **48**: 607–19

Wood R G, Corcoran M E, Courant P N 1993 Pay differences among the highly paid: The male-female earnings gap in lawyers' salaries. *Journal of Labor Economics* **11**: 417–41

F. D. Blau

Gender, Feminism, and Sexuality in Archaeological Studies

An important new field within archaeology, often referred to as the archaeology of gender, archaeology and gender, or gender archaeology, combines a critical feminist study of gender, sex, and sexuality with a feminist analysis and examination of central concepts within archaeology, and the ways in which archaeological knowledge is produced, presented, and represented. Although there are many different kinds of feminism, it is taken here in its broadest sense, with its major goal being a critique of the determination of women's lives by nature/biology, reducing their multifaceted lives to an essentialist category 'woman.' Rather, gender, sex, and sexuality are not constrained by essentialist and narrowly defined categories, but are seen as being historically constituted and varied categories, changing over time. The aims of feminist archaeology are to both investigate and challenge the assumed content of, and relationships between, gender categories such as women and men, femininity and masculinity; and to study gender divisions and sexuality over time and in different social and cultural contexts in the past. Archaeology is a way of knowing and investigating the past that is based on a particular androcentric understanding of objectivity that is historically constituted Euro-American. Feminist archaeology, using feminist critiques of science, challenges this traditional construction of archaeological knowledge.

1. A Slow Start

Although tied to the goals of Second Wave feminism, which not only critiqued gender relations in contemporary society, but also began very specifically to investigate how assumptions about the 'naturalness' of sex and gender affected the production of knowledge, feminism was slow in coming to archaeology. In the mid-1970s the first feminist critiques of equity issues in archaeology started in Scandinavia, particularly Norway. This concern quickly expanded to critiques of analyses and presentations of women in the past, and resulted in the founding of the first, and as yet only, feminist archaeology journal, *KAN: Kvinner i Arkeologi i Norge* (KAN: Women in Archaeology in Norway) in 1985. The previous year, Margaret Conkey and Janet Spector's groundbreaking article critiquing androcentrism in archaeology and discussing the possibilities of studying gender in the past was published in the United States. It was, however, at the beginning of the 1990s that the archaeology of gender and feminist critiques exploded on the archaeological scene.

During the course of the 1990s there were numerous publications on the archaeology of gender, particularly edited volumes of symposia and conference papers, as well as a few monographs. Few of these publications use the word 'feminist' in their titles, preferring various combinations of the words 'gender' and 'archaeology.' However, the majority are feminist in content and intention, and in their interpretations of the past.

Feminist archaeology in its beginnings was particularly an Anglo-American, Australian, and Scandinavian phenomenon, but expanded in the 1990s to other European countries, particularly Germany and Spain. The slow development of feminist archaeology in some countries can be explained partially by how archaeology is viewed within the different academic systems. Archaeology and anthropology in North America are closely linked, and feminist archaeology benefited from its close connection with feminist anthropology. In many European countries, however, archaeology is not a social science but a discipline in the humanities, and thus society and social relations were not in the forefront of many archaeological studies. At the same time archaeology is one of the few humanities that works with material culture, rather than texts, and has thus been relatively isolated from the intense feminist discussions within the humanities.

This basic disciplinary difference between the North American and the European archaeology of gender is probably one reason for the different perspectives these two archaeologies had initially. North American gender archaeology has been concerned more with social processes while European gender archaeology continued with traditional interests in cultural (now gender) identity and distinctions. However, as the archaeology of gender has developed and changed, this tendency is now minimalized as not only anthropology but social theory, post-structuralism, post-processualism, and feminist theory have become important elements of archaeological thinking in all areas.

The feminist philosopher of science Alison Wylie (1991) has considered why it took so long for feminist critiques to reach archaeology and emphasizes the essentialist assumptions in archaeological theorizing tied to macroscale analyses. Since these essentialist assumptions were based on biological determinism it was vital in the beginning to make a distinction between sex, a biological construct, and gender, a social construct. Here one must note the difference between the English language separation of sex and gender, and the fact that many other languages have but one word for both. Recent work in feminist archaeology, both in English and non-English-speaking countries, resurrects sex and sexuality from the

merely biological, but while viewing sex/sexuality as a social construction associated with gender, this work does not neglect changing biological processes.

2. Back to Basics

Much feminist work in the 1980s and 1990s concentrated on gender rather than sex and sexuality. As with feminism in general, this was necessary in order to negate the importance of previous biological determinism and essentialism in placing women in a subordinate position in the past. Androcentric archaeology considered women essentially the lesser part of the male–female, masculine–feminine, man–woman dichotomies. Women were often reduced conceptually to their biological function of giving birth. This also placed women on the lesser side of a whole series of related hierarchical dichotomies, such as public–domestic, public–private, powerful–dominated, active–passive, etc., thus giving biological reasons for social differences and social formations. For example, the origins of the nuclear family were tied closely tied to stable binary sex and gender distinctions which were seen as beginning almost 4 million years ago. This pattern was established in the archaeological literature before feminist critiques turned this set of universal hierarchical dichotomies into a problem to be investigated rather than assumed.

However, the classic feminist concern with the universal subordination of women played a peripheral role in feminist studies of the past. It was obvious that women were given a subordinated role in archaeologists' textual and visual presentations, and feminists critiqued such androcentric and essentialist assumptions. Conkey and Spector (1984) noted that despite its professed objectivity and neutrality, archaeology was 'neither objective nor inclusive on the subject of gender.' This false notion of objectivity contributed to making women invisible. The gender paradigms implicitly treated women and men differently, consistently linked certain activities to each gender/sex, and made men central and active in all societies in the past, while women were passive and 'at home.'

Feminist interpretations of the same past critiqued the subordinate role given to women in archaeological texts, exhibits, and illustrations. Feminist archaeologists have worked from the premise that universal subordination was not necessarily the case in the past, that gender ideologies and women's roles varied and were given different statuses over time and space, again emphasizing that these concepts are something to be investigated rather than assumed. In the 1980s both feminist theoretical and empirical work assumed that gender categories were not necessarily exclusive and that women were an active part of culture and society in the past. Studies were concerned with defining women's various roles, contributions, and activities. Task differentiation was developed as a framework and methodology by Janet Spector in order to study the various dimensions of women's and men's activities in past societies. Some form of this method is still often used in current analyses in the archaeology of gender, and is one of many contemporary efforts to put women back into our understanding of the past.

Following feminist critiques of science, feminist archaeologists examined critically the basic assumptions behind previous understandings of the past. The majority of articles on gender have as implicit but important goals the examination of what counts as data, what counts as interpretation of that data, and what is considered acceptable interpretation and theory in the male-dominated archaeological community. Thus they examine the selective processes which determine relevant data for interpretation, and how this over-determines interpretations and results in ignoring gendered processes and structures. Why are 'men's' artifacts simply assumed to be of primary interest? Why did important interpretational frameworks consider only men to be the prime movers of social and cultural change? In addition, they examine androcentric theoretical concepts which were previously, and often currently, considered to be the only reasonably legitimate and, indeed, the only 'natural' and 'rational,' theoretical frameworks. Engendering the past not only redefines and extends what is considered as data, interpretation and theory, but also challenges many taken-for-granted concepts and models.

3. Approaches to Gender and Sexuality

As with all feminist projects, the archaeology of gender is grounded in the analysis and understanding of women's lives. Women's lives are reflected in archaeological remains of the past, in the material culture (objects and structures) of past societies. Richly detailed empirical studies, which put women (and children) back in to our understanding of the past, abound. These are tied to differing theoretical views of gender as a structuring principle, and examine how material culture was used to express and construct gender roles, differences, ideologies and gendered identities in past societies. Foregrounding gender effects changes in traditional archaeological thinking and leads to an interest in social relations between genders and between gendered individuals of differing ages, classes, ethnicity, and race.

Engendered studies of Stone Age archaeology have gone further than a concern with ecology and technology by considering the wider perspective of social relations between individuals and groups which leads to a restructuring of traditional archaeological understandings, offering new insights into social relations,

ideologies, and sociocultural institutions. Margaret Conkey (1991) considers the wider contexts in which tools were made, used and reused at a Magdalenian site where several families lived ('aggregation sites'). By viewing tools not as single objects, but as objects representing a complex set of technological, productive, and engendered social relations, she is able to show the changing contexts of action and contexts for power in which people and objects were involved. Gender is always in process, and the sexual/gender division of labor is being changed and negotiated constantly. These changing contexts of action are associated with the renegotiation of power relations and structures. This type of analysis challenges traditional views of stable and exclusive gender categories and opens up possibilities for analyses of archaeological material that can reveal that there can be different models of gender in one society, that these models may change in relation to particular contexts and changes in the lives of individuals.

European archaeologists have made considerable use of burial material in discussions of the construction of gender. In her study of the Viking period Oseberg burial, a rich grave with the burial of two women in a ship, Elisabeth Arwill-Nordbladh (1998) uses M. Conkey's 'context of action,' here relating to textile production, farm economy, communication, and ritual performance. Artifacts in the grave were analyzed in order to see how gender could have been constructed through the various contexts of action of which these objects were a part. Gender construction is not seen as following a strict set of gender categories, but as a process that can be maintained, changed, negotiated, or resisted through practice or performative acts. In this manner gender is seen as a difference between, and as a difference within, gender categories, so that the complexity of gender dynamics comes to the forefront.

Marie Louise Stig Sørensen (1991), working with grave material from the North European Bronze Age, has done a series of studies on how objects are active in expressing gendered identity. She examines especially how gender is constructed through appearance by examining the composition and placement of grave goods in relation to the body (skeleton) and in relation to their wider sociocultural contexts of use and meaning. In this manner it is possible to analyze not only gender but also cross-cutting social categories such as age, ethnicity and status.

Concentration on gender leads to an initial neglect of sex and sexuality. However, as feminism and feminist theory developed and was increasingly concerned with difference, rather than solely with the binary distinctions of gender, and view sexuality and the body as both socially constructed and historically constituted, so too has feminist archaeology newly expanded its interest to these previously neglected areas. Sexuality as an expression of desire is not limited to the heterosexual body, but is examined in its variety of expression, including homosexuality. Much of this work is on representations of the body in art, particularly in classical Mediterranean cultures, but also in studies of prehistoric rock art.

As can be seen from this short overview, there are two threads in feminist archaeology. One is the 'archaeology of gender,' while the other is 'gendered archaeology' (Roberts 1993), what I have referred to as feminist archaeology. The archaeology of gender, while often premised on a feminist perspective and concerned with putting women and children back into our understanding of the past, reiterates traditional archaeological concepts and principles of interpretation. A gendered archaeology is explicitly feminist and tied closely to feminist critiques of science in an effort to generate new engendered ways of thinking, doing, and presenting archaeology. A gendered archaeology/feminist archaeology does not simply add women's contribution to past societies, but is concerned with interrogating and reworking our unquestioned starting premises and taken-for-granted concepts.

4. *Feminist Practice*

Although often considered to be neutral and objective, archaeological practice is highly gendered, and often implicitly so. Feminist interests include archaeological practice, particularly fieldwork and analysis, workplace issues, pedagogy, cultural heritage management, including museums, new ways of writing and presenting archaeology, and concern for the local communities in which one works.

Joan Gero (1996) has examined how fieldwork, particularly excavations and artifact analysis, is gendered. Gero has shown that observation and documentation are not neutral, objective processes but always contain an element of interpretation. Her work on excavations in Argentina show that there are gender differences in documentation, and these are often combined with differences in class and age. In addition, the gender of the person who typically does particular kinds of archaeological work varies. In North America men often lead large fieldwork projects, while women analyze extant collections. In other countries these tendencies take other forms, as in Australia, where men dominate in the universities while women dominate the work of cultural heritage management.

The status of women and equity issues in the archaeological workplace are continuing major concerns of feminists. Although there were some prominent women archaeologists in the past, these had little status and recognition in the profession. There is still a chilly climate for women in archaeology which not only slows down the advancement of female archaeologists but also seems to be preventing the full acceptance of feminist archaeology.

While an important aspect of feminism in archaeology is a concern with the institutional structure and workplace dynamics which hinder women from participating fully in the discipline, feminist archaeologists are also concerned with archaeology's relationship to the communities within which they work. In Australia, Laurajane Smith (1995) has written about the dominance of women in work with cultural heritage, or cultural resource management. Feminists have been particularly open to the importance and meaning that local communities and aboriginal peoples place on their past.

Much standard archaeological methodology is used in feminist archaeology. It is, however, reworked, because feminists ask different questions of the data. Instead of an interest in macroprocesses and population groups, typical, for example, of the systems approach, they ask questions at a microlevel, and are concerned more with social relations and processes, ideologies, and symbolism. Thus methods are combined in alternative manners. For example, emphasis is placed on entire artifact assemblages, and site assemblages, rather than tightly-defined typological categories (where the 'important' and 'interesting' types are often tools presumed to have been used by men), and comparisons over and between large regions. Ambiguous objects, often considered to be merely 'noise' in strict typological terms, are included in analyses since they are one way of analyzing the variety of social relations found at a site. In addition, interest is not placed on type of artifacts alone, but on the relationship between different kinds of artifacts and the processes involved in their production and use. This in turn emphasizes the variety of social relations involved in these processes and de-emphasizes the need to attribute the production and use of tools to an individual or group of the same sex.

5. *From Margin to Center*

Women have been invisible in archaeological interpretations of the past, but recent work grounded in feminist critiques, principles and perspectives have not only made women visible but have also shown the diversity of their lives and relations to others. Gendered archaeology/feminist archaeology has developed and changed considerably during its short existence. Feminist archaeology is exceptionally dynamic, drawing on new theoretical insights from other feminist disciplines, and expanding the scope and direction of archaeological analyses of gendered relations and ideologies in the past.

In a review article, Conkey and Gero (1997) note the lack of explicit feminist theorizing and critique of basic archaeological concepts in much of gender archaeology. There is little archaeological theorizing on gender/sexuality and there is little use of feminist theory in the archaeology of gender. Detailed empirical analyses are, of course, a necessary complement to theoretical work, but too much of the archaeology of gender continues within this arena. Feminist archaeology must move beyond this implicit essentialism. The lack of theory only marginalizes the archaeology of gender, not only in relation to mainstream archaeology but also in relation to other feminist disciplines. Future work must continue to include women in our construction of knowledge about the past, but it will also need to grapple with the fundamentals of archaeological interpretation and theorizing, and incorporate feminist theoretical insights on gender and sexuality. Feminist archaeology is dynamic and constantly expanding its areas of interest, and is well on it way to realizing its critical and reconstructive potential.

See also: Gender and Feminist Studies in Anthropology; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Heterosexism and Homophobia; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Male Dominance; Rationality and Feminist Thought

Bibliography

Arwill-Nordbladh E 1998 *Genuskonstruksjoner i Nordisk vikingtid, förr og nu*. [Gender Constructions in the Nordic Viking Period, Before and Now]. GOTARC, Gothenburg Archaeological Theses, Series B No. 9

Conkey M 1991 Contexts of action, contexts for power: Material culture and gender in the Magdalenian. In: Gero J, Conkey M 1991 (eds.) *Engendering Archaeology: Women and Prehistory*. Blackwell, Oxford, UK, pp. 57–92

Conkey M, Gero J 1997 Programme to practice: Gender and feminism in archaeology. *Annual Review of Anthropology* **4**: 11–37

Conkey M, Spector J 1984 Archaeology and the study of gender. *Advances in Archaeological Theory and Method* **7**: 1–38

Diaz-Andreu M, Sørensen, M L S 1998 *Excavating Women: A History of Women in European Archaeology*. Routledge, London, UK

Gero J 1996 The social world of prehistoric facts: Gender and power in Paleoindian research. In: Wright R (ed.) *Gender and Archaeology*. University of Pennsylvania Press, Philadelphia, PA

Gero J, Conkey M 1991 *Engendering Archaeology: Women and Prehistory*. Blackwell, Oxford, UK

Gilchrist R 1999 *Gender and Archaeology*. Routledge, London

Hager L (ed.) 1997 *Women in Human Evolution*. Routledge, New York

KAN: Kvinner i arkeologi i Norge (KAN: Women in Archaeology in Norway) 1984–present

Moore J, Scott E 1997 *Invisible People and Processes: Writing Gender and Childhood into European Archaeology*. Leicester University Press, London, UK

Roberts C 1993 A critical approach to gender as a category of analysis in archaeology. In: du Cros H, Smith L 1993 (eds.) *Women in Archaeology: A Feminist Critique*. Department of Prehistory, Occasional Papers in Prehistory No. 23, Australian National University, Canberra, Australia pp. 16–23

Smith L 1995 Cultural heritage management and feminist expression in Australian archaeology. *Norwegian Archaeological Review* **28**(1): 55–63

Sørensen M L S 1991 Gender archaeology and Scandinavian Bronze Age studies. *Norwegian Archaeological Review* **25**(1): 31–49

Wylie A 1991 Gender theory and the archaeological record: Why is there no archaeology of gender? In: Gero J, Conkey M (eds.) *Engendering Archaeology: Women and Prehistory*. Blackwell, Oxford, UK, pp. 31–54

E. Englestad

Gender History

Gender is a category of social analysis which denotes the social and cultural, as opposed to natural or biological, relations of the sexes. When this term first gained currency among historians of women in the early 1980s, gender was defined as the social relations between the sexes. As gender became an increasingly important and visible site of scholarly research and debate across the disciplines, its meaning widened to include the symbolic system or signifier of relations of power in which men and women are positioned differently. The embrace of the concept gender as a system of languages, symbols, and social practices that defined sexual difference and created sexual inequality, signaled the departure of feminist social scientists from the unchanging and universal notions of biological sex (Haraway 1991). The connotation of the English word 'gender' is a culturally formed set of social relations, distinct from sex, which is usually understood as rooted in biological or bodily differences. The word gender, however, does not exist in some languages, while in others its meaning diverges significantly from the English usage. The German term *Geschlecht*, for example, encompasses both 'sex' and 'gender' and thus blurs the core distinctions of the English term. In French the term 'sex' or *sexuel* is more commonly used than *genre*, which can refer to either grammatical or literary genre or serve as a classifying category in natural history (Riot-Sarcey 1999). The use of gender as an analytical tool does not *per se* connote a primacy of gender relative to other forms of inequality, such as race, class, or ethnicity; rather it suggests the inextricable links between gender and other social identities and categories of difference.

1. *From Women's History to Gender History*

The historical study of women, which became a vital field of scholarly inquiry during the 1960s, first sought to recover the ways in which women participated in and were excluded from processes of social transformations and political change. The empirical investigation of the private (familial, sexual) and public (political) inequality of women in past societies drove the first decade of feminist social science. From its inception, women's history had a dual character: its practitioners were historians, trained and employed in history faculties, but the arenas of debate that shaped their inquiry, paradigms and methodologies were most often housed in interdisciplinary programs, such as women's studies.

In this first phase of women's history studies, sex and class often figured as parallel forms of oppression: the female *sex* was viewed as a subordinate *class*, subjugated by a dominant *class* of men. In feminist scholarship the notion of *patriarchy* signaled a shifting understanding of sexual exploitation as the primary form of women's oppression (see, e.g., Rubin 1975, Eisenstein 1978). Taking patriarchy as a hierarchical sexual order, women's historians located its origins in 'the private family' and sought to examine the ways in which it was reproduced in social modes of production, divisions of labor, and property relations (Kelly 1976). The examination of the origins and practices of patriarchal structures and ideologies, and the recovery of the lost stories of female actors and women's agency, significantly expanded the scope of scholarship on women and family, women and labor, social policy and social movements, and elevated 'sex' to a keyword of social analysis. Yet in furnishing a more complete rendering of the contemporary and past societies, the pioneering work of women's history also revealed the limitations of the keyword 'sex.' Unitary analytical categories, such as those of 'woman'/'women' and 'man'/'men,' were based on fixed, rather than historically or socially variable, notions of biological sex, thus creating a cohesive history of the oppression of women that overlooked differences of race, class, ethnicity, and sexual preference. Further, the emphasis on the patriarchal oppression of women by men meant not only that women formed a homogeneous category, but also that they were most often studied in isolation from men, leading to a fragmented and partial understanding of the workings of sexual difference in society (Davis 1976).

As sex became a crucial category of social analysis in the 1970s, its impact on other keywords of social analysis, such as class or race, became a topic of intensive debate among social scientists. Labor historians and political theorists often attempted to fit women into prevalent notions of class, some by arguing that women formed a separate or parallel class of their own, others by examining the ways in which (female) gender presented an obstacle to the formation of cohesive classes. Still others circumvented the dilemma of 'sex and class' by ascribing to women the same class position, identity, and interests as those of their fathers, or husbands. Feminist historian Joan Gadol Kelly sought to transcend this theoretical impasse by formulating a 'doubled vision' of society,

one that emphasized the ways in which both men and women's identities of class were shaped by sex and class (Kelly 1979). In emphasizing the inextricable links between sex and class, women's history studies also indisputably widened the scope of the political to include family and household, bodies and sexualities—all once considered as belonging to the sphere of the 'private.'

Joan Kelly's formulation of this doubled vision of society marks the point at which the concept of gender came into currency as feminist scholarship turned its attention to the ways in which ideologies, norms, and symbolic systems shaped identities, social institutions, and social relations. The editors of the pioneering collection, *Sex and Class in Women's History* (Newton et al. 1983), for example, noted that they employed gender in order to understand the systematic ways in which sex differences cut through society and culture and conferred inequality upon women. This formulation makes it clear that even as 'gender' began to emerge from women's history studies in the early 1980s, 'women' and 'gender' continued to be used as virtually interchangeable terms and scholarship on women continued to provide a foundation for study of gender in most fields (e.g., Hausen and Wunder 1992).

The intellectual forces driving the shift from 'women' to 'gender' were multiple. Despite the achievements of nearly two decades of accumulated scholarship in women's history, the overarching narratives of progress, the chronological and conceptual framings of historical change, remained unshaken. While feminist historians had delivered powerful critiques of the public/private dualism, widening the scope of the political to include family and household, bodies and sexualities, and had successfully dissolved the sex/gender distinction as well, the binary pairings of home/factory, production/reproduction and production/consumption remained tenaciously in place. As feminist historians sought, empirically and theoretically, to overcome or dissolve these dualisms, they began to pose more fundamental challenges to established chronologies, categories and theories of social transformation (Kelly 1984).

The dissatisfaction with biological essentialism as an explanation of sexual inequality, for example, focused new attention on the power of language—of discourses—to construct these inequalities socially and to anchor them in social practices and institutions. The work of analytically distinguishing gender from sex also took place in a more public context, namely in the arenas of debate about 'identity politics' in the USA and Britain, in which once-cohesive ethnic, racial, national or sexual identities were increasingly understood as multiple, mutable, even contradictory. The context for a redefinition of gender became broader as race also came to be viewed as a social construction rather than a genotypic or phenotypic form of difference. At the same time, the unitary category 'woman' began to break down, as historians of race, particularly in the USA, castigated feminists for the false homogenizing of women and patriarchal oppression. Emphasizing the inextricability of gender and racial identities, feminist scholars of race argued that this category could only be dissolved by bringing race more centrally into the analysis of power (Higginbotham 1992). Feminist scholars of Western empires approached the history of white women's feminism as implicated in imperial ideologies, institutions, and practices. They called upon historians of white women's oppression to confront the racist, nationalist, and imperialist legacies of Western feminism (Amos and Parmar 1984).

2. *The Turn to Gender History*

Although a shift from women to gender was already well underway by the mid-1980s, the publication of Joan Scott's essay *Gender: A Useful Category of Historical Analysis* (Scott 1986), marks the turning point at which the two terms became distinct in the field of history studies. In the Anglo-Saxon academy the turn to gender as a cultural, linguistic, or discursive construct took place within a broader paradigm shift known as 'the linguistic turn' in the social sciences. The linguistic turn, a catch-all phrase representing a new receptivity of social scientists towards post-structuralist literary criticism, linguistic theory, and philosophy, as well as cultural and symbolic anthropology, signaled a new view of language as constituting historical events and human consciousness rather than as simply reflecting social reality (Toews 1987, Canning 1994). Scott's article, together with her essay collection, *Gender and the Politics of History* (Scott 1988), made explicit the theoretical innovation of gender, summoning 'mainstream' historians to consider gender as an essential category of historical analysis and unleashing a torrent of controversy, not least in the fields of women's studies and women's history.

At the same time Denise Riley (1988) interrogated the category 'woman,' noting its inherent and historically founded instability and pointing toward feminism as precisely the site at which that instability was systematically contested, taken apart and reconstructed. In this sense the turn to gender dissolved some of the stalwart categories of women's history, including both 'woman'/'women' and feminism, with 'experience,' 'identity,' and 'agency,' soon to follow. Furthermore, the breakdown of the category 'woman' and the more explicit formulation of gender as *relational*, as a category and process by which femininity and masculinity are mutually constituted, spurred the opening of gender history towards a more serious study of masculinity.

In the USA and Britain, feminist scholars of race also intervened at this juncture with powerful critiques of the categories and practices of women's history.

Defining race as a 'metalanguage,' as the 'ultimate trope of difference,' African–American critics explicated the powerful effects of race on the construction and representation of gender, class, and sexuality. Understanding race as a 'double-voiced discourse' revealed the ways in which it shaped both the oppression and self-representation of minorities (Higginbotham 1992).

3. *Controversies and Debated in Gender History*

Gender gained an analytic status of its own first by severing the links between sex-gender and biology and then by freeing it from the remnants of Marxism that hinged the production of gender to changing economic structures (Scott 1986). This process necessitated not only the dismantling of familiar concepts, but also the exploration of new methodologies that would illuminate the workings of language and symbols and their dissemination throughout society and culture. Feminist historians also debated the implications of extending gender historical inquiry to masculinity and men. Some of the controversy about these questions took place as an integral part of the debates among social historians on the 'linguistic turn,' while others were more specifically framed by the previous practices and boundaries of women's history. Historians of women feared, for example, that studying gender as a cultural, discursive, or linguistic construction in which masculinity and femininity are relationally and mutually constituted, would render female actors and women's agency invisible and obscure male oppression of women, ultimately producing a gender-neutral historical discourse (Bock 1991). Others suspected that the study of masculinity, while discovering the ways in which ideals of manliness were imposed upon men, might shift attention away from men's participation in those institutions and practices that oppressed women.

3.1 *Gender and the Keywords of Social Analysis*

The turn to gender, like all theoretical and methodological innovations, was an uneven and protracted process, as different fields and national scholarly arenas manifested widely differing responses to this challenge. The extent to which methodologies, concepts, and theories were redefined to meet this challenge also varies across these divergent contexts. In some cases women's history was simply renamed 'gender history,' leading to a protracted phase of interchangeability of the two without significant changes in concepts or methodologies. In other national contexts where the work of recovering the absent female subjects began later, historical research on women continues to provide the foundation for a more explicitly theoretical engagement with gender. In still other settings, where the turn to gender history coincided with the rediscovery of masculinity and men, gender history denotes a dual engagement with the histories of both masculinity and femininity.

While the study of women once sought to fit female subjects into existing categories of social analysis, the turn to gender prompted more fundamental critiques, expanded meanings, and redefinitions of the keywords of social analysis. The concept of 'class,' for example, was at the heart of debates about both gender and the linguistic turn, leading some feminist scholars to call for an emancipation from class as a 'privileged signifier of social relations and their political representations' (Alexander 1984), while others have analyzed both middle-class (Davidoff and Hall 1987) and working-class formation (Clark 1995) as processes of differentiation in which gender is always centrally implicated. Eschewing models of class formation that seek to link economic, social, cultural, and political levels or stages of development, others have analyzed class as political language, emphasizing the significance of gender in the process of assigning and contesting the contingent meanings of class (Scott 1988, Sewell 1990, Canning 1992).

Other social-scientific terms, like 'public sphere' (Fraser 1992, Ryan 1992), 'civil society' (Pateman 1988, Hull 1996), and more recently, 'citizenship' (Lister 1997) have also been the site of fruitful rethinking by scholars of gender. Similar to the location of class in socioeconomic structures, citizenship is embedded in legal and constitutional frameworks and theories of state and nation formation. Gender scholarship on citizenship has critiqued models of citizenship as a progressive process of acquisition of civil, political and social rights, as outlined, for example, in T. H. Marshall's classic, *Citizenship and Social Class* (1950). New considerations of the gender dimensions of this term have analyzed citizenship as a process, as a set of claims upon the state and public sphere rather than as a historically fixed outcome or legal status. The historical meanings of citizenship can thus be understood to encompass not only those who possess citizenship rights, but also those excluded from them.

3.2 *Gender, Sexuality, Body*

Other key terms in the vocabulary of interdisciplinary gender studies, which were redefined in the course of the paradigm shift from women to gender, are those of 'sexuality' and 'body.' In addition to critiquing and redefining keywords of social and historical analysis, such as class and citizenship, the turn to gender also generated new waves of scholarship on topics which had been submerged within the complex of women and sex and which came into their own once gender had been analytically distinguished from sexuality and body.

Influenced by the work of Michel Foucault, historians sought to analyze the historically specific

domains in which sexualities were produced, such as the systems of medical knowledge and state moral regulation that defined both normal and deviant sexuality in nineteenth-century Europe, dispersing these definitions throughout the institutions of state and social reform (Mort 1987). In this sense, it is possible to view sexuality as a more specific site of definition, intervention and experience than gender. A new emphasis on the importance of desire in the articulation of sexuality suggested that sexuality might be understood as a site of agency which was quite distinct from gender. Yet Judith Butler (1990) posited gender as lived, at least in part, through sexuality, as embodied and performed through the production and ascription of sexualities. In recent years, studies of homosexualities have raised new questions about the implications of transgressive sexualities for the study of gender, suggesting that they disrupt gender categories by breaking up the neat opposition of male/female, masculine/feminine and by unmasking 'the epistemic regime of presumptive heterosexuality' (Butler 1990, Puff 1998).

'Body,' the term in the vocabulary of gender that was most deeply embedded in biologism and essentialism, was inextricably linked to sex before the 'linguistic turn' began to disentangle these terms. During the first wave of women's studies, the body had been the foundation for unities and shared experiences of women. The 'discursivation' of the body, which took place in the course of the 'linguistic turn,' not only detached the body from unchanging notions of physical difference, but also examined bodies as sites of inscription—as instruments of changing cultural meanings. While the advent of the discursive body delivered the final blow to biological determinism, it also raised new questions about the particular embodiedness or bodily mediation of female experiences that remained central in feminist historical research (e.g., Duden 1991). In the debates of the early 1990s about the relationship between discourse and experience, for example, the body became a litmus test of this relationship (Butler 1993, Bynum 1995).

In the course of the 1990s, then, body history became a prolific offshoot of gender history. Both theoretical debate and empirical research in the field have shifted recent discussions towards exploration of the process of *embodiment*, which takes bodies as sites of mediation between constructions of the body and its materialized experiences (pain, birth, death, illness, dismemberment) and suggests new frameworks for examining the ways in which bodies are lived, experienced and acted upon by the subject and the social collectivity (Grosz 1994, Gatens 1996).

While the turn to the concept of gender spawned new waves of historical scholarship on sexualities and bodies, gender history later expanded the scope of its inquiry to two new fields of historical inquiry: the history of consumption and history of empires. The study of consumption in Western societies was invigorated in part by the collapse of communism in Russia and Eastern Europe, which made clear the ways in which consumption, as well as production, structure capitalist societies. Although consumption has historically been identified with femininity, Davidoff and Hall's *Family Fortunes* (1987) argued convincingly that domesticity, the ideology of the British middle classes, was vested in complex and elaborate practices of consumption and display. With gender as a framing concept, the historical study of consumption has expanded in recent years to encompass analysis of a social and cultural formation—consumer society—through the lens of commodification, spectatorship, commercial exchange and welfare reforms, each of which can be analyzed in conjunction with the gender identities produced through these processes (DeGrazia and Furlough 1996). Throughout the 1990s gender and sexuality also had a central place in the history of empire and figured centrally in new understandings of nations, empires, and imperial social formation.

4. *Current Issues and Future Directions*

This brief survey of gender historical scholarship indicates that the concept of gender has successfully broadened and deepened the meanings of some of the keywords of historical analysis, forged new arenas of scholarly inquiry, and helped to refurbish epistemologies and methodologies of historical analysis. Gender history has moved from the periphery once occupied by the 'herstory' of early women's history to the center in some historical fields, while in others it remains distant from the problems that define the field. The historical narrative of the French Revolution has been significantly transformed by gender scholarship, which went far beyond analyzing the exclusion of women from revolutionary citizenship during Jacobin rule to revise broader understandings of the crisis of state (e.g., Lynn Hunt's notion of 'family romance'), the emergence of the body politic, civil society, and notions of citizenship. The history of labor and labor movements, began to diminish in the late 1980s after two decades of field-defining scholarship, only to be significantly revived and recast through the wave of gender and labor scholarship of the mid- to late 1980s (Frader and Rose 1996). The histories of slavery in the antebellum South, of colonialism, empires, and post-emancipation societies have been rewritten as scholars explore the ways in which gender inequalities underpinned racial and imperial dominance. Gender has become a vital part of new scholarship on the two world wars, including military history. Yet other fields, such as the histories of the Weimar Republic or the Third Reich, for example, remain immune to the impulses of gender, despite an impressive accumulation of research on both women and gender for both

periods. Few would dispute that contests over masculinities and femininities are a crucial part of the post-World War I crisis of 'bourgeois Europe.' Yet in the historiography of Weimer Germany gender is meaningful mainly in those realms remote from high politics—realms as diverse as popular culture and sexual reform organizations. Nor has gender figured in the dramas of state, democracy, or the fragile social body of the 1920s. Recent scholarship on the Third Reich offers abundant evidence that the Nazis pursued, relentlessly and violently, a gender project, which was at the heart of the racial state they envisioned and built. But this knowledge does not significantly inform interpretations of the collapse of the Weimar Republic or the Nazis' mobilization of German society and economy for war and genocide. This uneven reception of gender history in various fields will undoubtedly shape the dynamic of gender history in years to come.

The relationship of gender history to 'mainstream' history has been of growing concern to those who formulated its emergence from women's history in the early 1980s. Joan Kelly's three-pronged project of renovating categories, chronologies, and theories of social transformation explicitly aimed at rewriting historical metanarratives, as she sought to do in her own essays on the Renaissance. Given the accomplishments of feminist history in recasting categories of historical analysis, it is now important to assess the potential of gender history to reconstruct historical narratives and to produce new syntheses of historical events and processes (e.g., Hunt 1998). Several dilemmas remain regarding gender history and the historical metanarratives it set out to subvert: should, for example, the empirical and conceptual achievements of gender history be assimilated into existing metanarratives? Should they inspire the composition of new metanarratives or prompt the abandonment of metanarratives altogether? These questions can perhaps be resolved first by breaking these metanarratives into their composite parts and examining the singular moments of subversion and incorporation along the narrative thread. Either task—that of revising existing narratives or that of composing new ones—requires a certain kind of piecework in which historical changes are rearranged on time lines and brought into new relationship with one another.

Attention to the temporalities of gender history, and to how they fit or diverge from those of established metanarratives, has only recently become possible. The genealogy of gender might be traced back to the eighteenth century, when humanist notions of the individual were synonymous with masculinity. The embodiment and masculinization of the male citizen during the French Revolution also provided the stage upon which women staked their first claims for inclusion in the terms of both equality and difference. The extensive scholarship on women and gender in nineteenth-century Europe encompasses the history of a wide range of public feminisms and domestic ideologies, the emergence of 'women' as social category, of 'women's labor' as a social question, and the dissemination of these ideologies and social categorizations, which redefined both femininity and masculinity, through the media of the emergent social and natural sciences. This suggests (a) that the history of nineteenth-century Europe could now be rewritten from the perspective of gender, and (b) that the temporality of gender history may not follow the political or social chronologies of existent metanarratives. A rethinking of twentieth-century history may uncover distinct temporalities of a different gender history, one in which gender norms and ideologies were ruptured at points of political and cultural crisis but not necessarily restored at moments of political reconstruction and stabilization. This trajectory of twentieth-century gender history may indeed help to explain its extraordinarily embodied violence. The work of aligning the temporalities of gender history with those inscribed in established historical metanarratives represents the first step towards a history that admits distinct temporalities and allows for dissonance within grand narratives.

See also: Body, History of; Family and Kinship, History of; Gender and Feminist Studies; Gender and Feminist Studies in Anthropology; Gender and Feminist Studies in History; Gender and Feminist Studies in Sociology; Gender Differences in Personality and Social Behavior; Gender Ideology: Cross-cultural Aspects; Historiography and Historical Thought: Current Trends; Household Production; Linguistic Turn and Discourse Analysis in History; Social History

Bibliography

Alexander S 1984 Women, class and sexual differences in the 1830s and 1840s: Some reflections on the writing of a feminist history. *History Workshop* **17**: 125–14

Amos V, Parmar P 1984 Challenging imperial feminism. *Feminist Review* **17**: 3–19

Bock B 1991 Challenging dichotomies: Perspectives on women's history. In: Offen K, Pierson R R, Rendall J. *Writing Women's History*. Indiana University Press, Bloomington, IN

Butler J 1990 *Gender Trouble. Feminism and the Subversion of Identity*. Routledge, New York/London

Butler J 1993 *Bodies That Matter. On the Discursive Limits of 'Sex.'* Routledge, New York/London

Bynum C 1995 Why all the fuss about the body? A medievalist's perspective. *Critical Inquiry* **22**: 1–33

Canning K 1992 Gender and the politics of class formation: Rethinking German labor history. *American History Review* **97**(3): 736–68

Canning K 1994 Feminist history after the linguistic turn: Historicizing discourse and experience. *Signs* **19/2**: 368–404

Clark A 1995 *The Struggle for the Breeches. Gender and the Making of the British Working Class*. University of California Press, Berkeley, CA

Davidoff L, Hall C 1987 *Family Fortunes: Men and Women of the English Middle Class, 1780–1950*. University of Chicago Press, Chicago
Davis N 1976 Women's history in transition: The European case. *Feminist Studies* **3**(3/4): 83–103
DeGrazia V, Furlough E (eds.) 1996 *The Sex of Things: Gender and Consumption in Historical Perspective*. University of California Press, Berkeley, CA
Duden B 1991 *Woman Beneath the Skin: A Doctor's Patients in 18th Century Germany*. [Transl. Dunlap T] Harvard University Press, Cambridge, MA
Eisenstein Z 1978 *Capitalist Patriarchy and the Case for Socialist Feminism*. Monthly Review Press, New York
Frader L, Rose S (eds.) 1996 *Gender and Class in Modern Europe*. Cornell University Press, Ithaca, New York
Fraser N 1992 Rethinking the public sphere: A contribution to the critique of actually existing democracy. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA
Gatens M 1996 *Imaginary Bodies: Ethics, Power and Corporeality*. Routledge, London
Grosz E 1994 *Volatile Bodies: Toward a Corporeal Feminism*. University of Indiana Press, Bloomington, IN
Haraway D 1991 *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York/London
Hausen K, Wunder H 1992 *Frauengeschichte/Geschlechtergeschichte*. Campus Verlag, Frankfurt/New York
Higginbotham E B 1992 African-American women's history and the metalanguage of race. *Signs* **17**(2): 251–74
Hull I 1996 *Sexuality, State, and Civil Society in Germany, 1700–1815*. Cornell University Press, Ithaca, New York
Hunt L 1998 The challenge of gender. Deconstruction of categories and reconstruction of narratives in gender history. In: Medick H, Trepp A-C (eds.) *Geschlechtergeschichte und Allgemeine Geschichte, Herausforderung und Perspektiven*. Wallstein, Göttingen, Germany
Kelly J 1976 The social relation of the sexes: Methodological implications of women's history. *Signs* **1/4**: 809–23
Kelly J 1979 The doubled vision of feminist theory. *Feminist Studies* **5**(1): 216–27
Kelly J 1984 *Women, History, and Theory: The Essays of Joan Kelly*. University of Chicago Press, Chicago
Lister R 1997 *Citizenship: Feminist Perspectives*. New York University Press, New York
Mort F 1987 *Dangerous Sexualities: Medico–Moral Politics in England Since 1830*. Routledge, New York
Newton J, Ryan M, Walkowitz J (eds.) 1983 *Sex and Class in Women's History*. Routledge, New York
Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA
Puff H 1998 *Männergeschichten/Frauengeschichten*. Über den Nutzen einer Geschichte der Homosexualitäten. In: Medick H, Trepp A-C (eds.) *Geschlechtergeschichte und Allgemeine Geschichte, Herausforderung und Perspektiven*. Wallstein, Göttingen, Germany
Riley D 1988 *Am I That Name? Feminism and the Category of 'Women' in History*. University of Minnesota, Minneapolis, MN
Riot-Sarcey M 1999 The difficulties of gender in France: Reflections on a concept. *Gender & History* **11**(3): 489–98
Ross E, Rapp R 1983 Sex and society: A research note from social history and anthropology. In: Snitow A, Starnell C, Thompson S (eds.) *Powers of Desire*. Monthly Review Press, New York
Rubin G 1975 The traffic in women: Notes on the 'political economy' of sex. In: Reiter R (ed.) *Toward an Anthropology of Women*. Monthly Review Press, New York
Ryan M 1992 Gender and public access: Women's politics in nineteenth-century America. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA
Scott J 1986 Gender: A useful category of historical analysis. *American History Review* **91/5**: 1053–75
Scott J 1988 *Gender and the Politics of History*. Columbia University Press, New York
Sewell W H Jr 1990 How classes are made: Critical reflections on E P Thompson's theory of working-class formation. In: Kaye H, McClelland K (eds.) *E. P. Thompson: Critical Perspectives*. Temple University Press, Philadelphia, PA
Toews J 1987 Intellectual history after the linguistic turn: The autonomy of meaning and the irreducibility of experience. *American History Review* **92**(4): 879–907

K. Canning

Gender Identity Disorders

1. *Definitions, Prevalence, and Sex Ratio*

The gender identity disorders are defined as disorders in which an individual exhibits marked and persistent identification with the opposite sex and persistent discomfort (dysphoria) with their own sex or sense of inappropriateness in the gender role of that sex. These characteristics may be manifest in statements such as wishing to be the other sex, dressing as, and taking on the roles of the other sex, engaging preferentially in activities usually enjoyed by the other sex, and expressing dislike for the roles and activities of their own sex, as well as a dislike for their own genitalia and other physical characteristics of their own sex. In order to distinguish the gender identity disorders from more minor cross-gender interests or role-taking these characteristics must be persistent and marked and result in distress or impairment (*Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, (*DSM-IV*), 1994; *International Classification of Diseases*, Tenth Edition, (*ICD-10*), 1992).

The word transsexual was used in the earlier literature, as was inversion, to refer to individuals with marked gender dysphoria, most of whom perceived sex reassignment as the solution to their psychological distress. It remains as a category in the *ICD-10* classification. Also in the older literature transvestite was used to designate persons who cross-dressed and may have included those who identified as the other sex as well as those in whom the cross-dressing was sexually arousing (now referred to as Transvestic Fetishism). Transgendered is a more recent term used to encompass a broad group of individuals with varying degrees of cross-gender behavior. It is not a diagnosis and is often the self-label assumed by individuals with cross-gender feelings and behaviors, particularly the more politically active. Gender role

refers to behaviors, roles, and attitudes that the specific culture defines as more appropriate to males or females. Sexual orientation refers to the sex of person to whom an individual is attracted, predominantly the opposite (heterosexual), or the same (homosexual) sex. Sexual identity connotes the adoption of an identity as a hetero- or homosexual person and involves more than the narrow notion of sexual orientation.

Gender identity disorders (GID) are relatively rare compared to other psychiatric disorders. Although there are no large scale epidemiological studies to provide true estimates recent studies suggest roughly 1:10,000 to 1:30,000. Sex ratios of adults with GID (largely based on referrals to clinics) have fluctuated with more males than females in earlier studies to a more equal ratio in many recent reports. Childhood GID is more prevalent in males, roughly 6 to 1. In adolescence the ratio is more equal (Cohen-Kettenis and Gooren 1999, provide a recent review.)

2. Historical Background

Awareness of individuals who identified strongly with the other sex has a long history. Some cultures even created a privileged role for such individuals. However, the scientific study of individuals with cross-gender identities really began in the late part of the nineteeth century. Initially many such individuals were thought to be intersex individuals, possessing the physical characteristics (both internal and external) of the other sex. It was not until the middle of the twentieth century that it became clear that most such individuals were physically normal but suffering from a psychological disorder.

The modern era of gender identity disorder began with the much publicized case of Christine Jorgensen, an American ex-soldier, who underwent sex-change surgery in Denmark in 1952, under the care of Christian Hamburger, an endocrinologist. Although demasculinizing surgery and hormonal replacement had been used earlier, typically for cases thought to be intersex, the Jorgensen case stirred the public imagination and encouraged other like-minded individuals to seek treatment for their distress (Meyerowitz 1998). Sex change surgery had been available in Europe since the early 1900s, the most prominent center being the Institute for Sexual Science in Berlin under the direction of Magnus Hirschfeld. In the USA physicians were less willing to perform such surgery, particularly on those considered to be physically normal and in some jurisdictions such body altering was considered illegal. Prior to the Jorgensen case, David Cauldwell, an American psychiatrist and editor of *Sexology*, had distinguished cases which he labeled '*psychopathia transexualis*' from intersex cases. Harry Benjamin, an endocrinologist, born and trained in Germany, had begun to advocate on behalf of individuals wishing sex change. This combination of public and professional interest set the stage for the scientific study of persons with gender identity disorders.

In 1955 John Money working with the Hampsons published his first follow-up study of individuals born with ambiguous genitalia. His conclusions that sex of rearing overrides the child's anatomy in determining gender identity was the first empirical evidence that psychological factors may be etiologically important in GID. The 1960s saw the publication of several important books and papers as clinicians working with individuals with GID began to describe their experience. Harry Benjamin, who had been diligent in his effort to bring understanding to the plight of those whose cross-gender wishes were often demeaned or dismissed, published *The Transsexual Phenomenon* in 1966. In the same year the first gender clinic in North America was established at Johns Hopkins Hospital in Baltimore. Green and Money (1969) co-edited the book *Transsexualism and Sex Reassignment*, a standard text at that time. he year earlier, Stoller (1968), a psychoanalyst, had published *Sex and Gender*, in which he articulated a theory that cross-gender identification arises from a blissful symbiosis of a son with his mother. Although Stoller's theoretical position has more recently been dismissed it sparked interest in understanding the psychodynamic origins of GID and led to more systematic study in several of the newly developed clinics for children and adolescents.

The 1970s and 1980s saw a growth of interest in treatment of adults with GID. Clinics developed in major centers in North America and Europe. Standards of Care were developed under the auspices of the Harry Benjamin International Gender Dysphoria Association in 1985 and provided some uniformity to patient management. Clinics for children were developed as clinicians realized that GID had its origins in childhood and might be prevented with early intervention. Green (1974) carried out the first systematic study of effeminate boys, published in *Sexual Identity Conflict in Children and Adults*. He compared boys displaying significant cross-gender behavior by parent report, with an equal number of age-matched masculine boys. His study provided the first systematic description of boys with GID and their parents. Green (1987) published the follow-up of these two groups of boys. The majority of the effeminate boys had a homosexual orientation when assessed in late adolescence. Only one boy continued to have significant cross-gender wishes. These findings were in marked contrast to the findings on the masculine boys who were almost entirely heterosexual in orientation and displayed no evidence of gender dysphoria.

The study of individuals with GID soon became a subject of religious, moral, and political debate. Some investigators urged treatment of GID to prevent homosexuality. In the 1990s funders of GID programs

began to debate the logic of a medical and surgical intervention for what is primarily a psychological disorder. This debate was fueled by the results of the first follow-up study of the patients from the John Hopkins clinic which suggested that surgically reassigned patients were no better off than were those who had not received surgery. Although this study was methodologically flawed, its results, together with economic pressures, led to the closure of many clinics in the USA and the removal of sex reassignment surgery as an insured service. Generally the situation in Europe, particularly in the Netherlands, remained more supportive for those with GID. In the late 1980s the gay and lesbian community began to question the pathologizing of cross-gender behavior in childhood, regarding it simply as an early manifestation of homosexuality. In the early 1990s many individuals with cross-gender identification began to organize politically, assuming the label transgendered, and arguing that society should recognize a third or intermediate position between the more traditional male and female genders.

3. *Emphases in Current Theory and Research*

3.1 *Classification and Phenomenology*

Early work focused on dividing individuals requesting sex reassignment into 'true' and 'secondary' transsexuals. This effort was abandoned as GID began to be conceptualized as being more dimensional than purely categorical. However, there was a recognition that there are two types of GID in males, one originating from early or childhood GID and being predominantly homosexual in sexual orientation, and the other, arising from Transvestic Fetishism, and being hetero- or bisexual in orientation. Individuals with the latter present most typically in their thirties in contrast to the former who present in adolescence or early adulthood. Although usually described as bi- or heterosexual, Blanchard (1991) has established that many of the latter group display what he calls autogynephilia, an erotic attraction towards oneself dressed as a female. Mature females with GID are essentially homosexual in their erotic attraction.

3.2 *Etiology*

As mentioned above, the first theorizing about etiology arose from the belief that individuals with GID are in some way intermediate in their biology between males and females. Because of the link between cross-gender identification and homosexuality and because of the relatively small numbers of individuals with GID this led to extensive investigation of the biological differences between heterosexual males and females and between these two groups and their homosexual counterparts.

Although many of these studies have provided interesting leads very few differences have been substantiated in a way that would suggest strong evidence for a significant predictor of gender pathology or of homosexuality. Rather the findings to date suggest that temperamental or other vulnerability factors may create a biological predisposition towards gender pathology which when combined with the appropriate psychodynamic factors produce gender dysphoria and cross-gender identification.

3.3 *Biological Studies*

Genetic factors are assumed to play a role in GID although there have been no large-scale studies of heritability in GID. This assumption is based on the studies of heritability in homosexuality and its putative relationship with GID. In monozygotic twins there is a greater likelihood of both twins being homosexual as compared to dizygotic twins (Bailey et al. 2000). This relationship has been shown most strongly in males. Despite a concordance rate as high as 50 percent there appears to be a strong nonshared environmental component (due to factors unique to the experience of the individual). Similarly, molecular genetics studies of homosexuality in men suggest the possibility of a marker specific to sexual orientation on the distal portion of the long arm of the X chromosome (Hamer et al. 1993) although the only other independent study failed to replicate this finding.

The prenatal hormone theory has provided the most coherent framework for an examination of factors affecting sex-dimorphic behavior in both animals and humans. This theory, now well supported by studies in animals, posits that the prenatal hormonal milieu affects the development of brain structures and organizes or shapes various sex-dimorphic behaviors, such as activity level, mounting behavior, and responsiveness to sexual partners. In humans, Congenital Adrenal Hyperplasia (CAH), a disorder in which females are exposed to high levels of circulating androgen *in utero*, provides a convenient natural experiment which gives some support to the theory. Girls with CAH have been shown to be more physically active and tomboyish than their nonaffected sisters and to report more homosexual attraction and less heterosexual involvement in adulthood. They do not, however, report more wishes to be the other sex. This latter finding may be partially a result of the small sample sizes in all studies, too small to detect a disorder such as GID, which is relatively rare.

Another approach to testing the prenatal hormone theory has been to examine hormone responses to a dose of exogenously administered estrogen. In the female such an injection produces an initial fall and then a rise in luteinizing hormone. Males do not

display the rise seen in females. Homosexual men show a pattern intermediate between that of heterosexual men and women. These studies have not been replicated consistently and have been criticized in terms of not being interpretable, as the endocrine systems of homosexual men may be compromised by such things as HIV infection.

Prenatal maternal stress is another factor thought to possibly influence the development of homosexuality (and by extension GID) through affecting the prenatal hormonal milieu and brain development. Early studies in rats showed that stressing pregnant females can produce a demasculinizing and feminizing effect on their male offspring. Although attempts to test this out on humans have been equivocal, Bailey et al. (1991) found that maternal 'stress proneness' was correlated with boyhood effeminacy. As there is an emerging literature that suggests that maternal stress may correlate with vulnerability to psychiatric disorder in offspring by affecting stress reactivity, these findings may be relevant in terms of vulnerability to psychopathology, as opposed specifically to homosexuality or GID (Bradley 2000).

Efforts to define and compare the brain areas responsible for sexually dimorphic behavior have been a part of testing the prenatal hormone theory. Once more this work has been carried out on samples of homosexual individuals, largely male. Although brain structures in rats appear more clearly sex-dimorphic, agreement about the comparable brain structures in humans has been more complicated. Swaab and Fliers (1985) reported differences between males and females in the preoptic area of the anterior hypothalamus but in a subsequent study could not define differences in volume between heterosexual and homosexual men. They did, however, find a difference in the suprachiasmatic area of the hypothalamus between homosexual and heterosexual men, an area previously not related to sex-dimorphic behavior (Swaab and Hofman 1990). Another cell group identified as possibly homologous with the sexually dimorphic area in lower mammals is the interstitial nucleus of the anterior hypothalamus, subdivided into four sections, INAH 1–4 (Allen et al. 1989). These authors found differences in size between males and females in INAH 2 and 3. LeVay (1991) subsequently reported a smaller INAH 3 in homosexual men (comparable to that of heterosexual women) in comparison to heterosexual men. This study has recently been partially replicated by Byne et al. (in press). Zhou et al. (1995) reported that the bed nucleus of the stria terminalis was smaller in six male-to-female transsexuals than in control males and was comparable to that of females.

Another area of potential difference between homosexual and heterosexual individuals is cerebral laterality. This has been examined through assessment of handedness, cognitive testing, dermatoglyphic asymmetry, and related brain volumes. Some studies have suggested that homosexual men and women have a higher likelihood of being left-handed than heterosexual men and that transsexuals (both male and female) display more left-handedness than the general population (Lalumiére et al., 2000). A preliminary study on boys with GID from the Toronto clinic examining handedness suggests a higher incidence of left-handedness (Zucker et al. 2000). Neuropsychological studies using visual hemifield dot detection and dichotic listening tasks suggest differences between homosexual and heterosexual men with the performance of homosexual men being more similar to that of heterosexual women. Studies of transsexual subjects also suggested less lateralization but again these results are preliminary. Study of fingerprint patterns, dermatoglyphy, shows sex differences with women having more ridges on the left than right hand and homosexual men showing a similar left-ward asymmetry. The anterior commissure, a fiber tract connecting the two cerebral hemispheres, is larger in women than men and has been shown to be larger in homosexual than heterosexual men. Although interpretation of these findings should remain open they suggest that early brain development and with it handedness and other indices of laterality may be affected in individuals who later become homosexual and possibly in those who develop GID.

Temperamental variables, specifically activity level and involvement in rough-and-tumble play, have been shown to differ in boys and girls. Boys with GID display a lower activity level and less interest in rough-and-tumble play than control boys. In contrast, girls with GID are reported to have a higher activity level and to be more interested in rough-and-tumble play than control girls. Although these findings suggest differences between GID children and controls, it is highly unlikely that these differences are more than predispositions that add a dimension to their general vulnerability to GID. Physical attractiveness is another factor that may act as a predisposing factor but is unlikely to be more specific to GID. Boys with GID are rated by independent raters as more attractive, beautiful, cute, handsome, and pretty than control boys. Conversely, girls with GID are rated as less attractive than control girls. Although physical attractiveness may be a social construction these findings are consistent with parental reports of their GID sons being beautiful babies and may be related to greater facial symmetry, a component of attractiveness.

Particularly feminine homosexual men and boys with GID have a greater number of brothers than controls. Being born later in the birth order also occurs more frequently in homosexual men (regardless of extent of cross-gender behavior) than in heterosexual controls. These findings have been interpreted from both biological and psychological frameworks with no evidence to suggest that either explanation is more valid.

Lastly, birthweight appears to be lower in boys with GID than in controls. Again this may reflect a neurodevelopmental factor which is related to a general vulnerability to psychopathology as low birth weight has also been found as a vulnerability factor to psychotic disorders.

3.4 Psychosocial Factors

As indicated in Sect. 2 above, Money et al. (1955) reported on their follow-up of children born with ambiguous genitalia. They concluded that sex of assignment and rearing, particularly when that was begun unambiguously prior to 24 months of age, was able to override a child's anatomy. These early studies by Money et al. led to the belief that gender identity was developed during a sensitive period and became relatively fixed at an early age. Despite challenges, this position has not been significantly refuted. Questions were raised by the finding that individuals with 5-alpha-reductase deficiency, apparently raised as girls, for the most part begin to self-identify as males at puberty when their bodies begin to masculinize. This and other similar findings appear less to have refuted Money et al.'s findings than to have suggested that in some instances gender identity may be less fixed than had originally been proposed. This same malleability of gender identity appears to be relevant in the case of males with Transvestic Fetishism who although originally identified as male begin, over a number of years, to self-identify as female. Other anomalous situations include males with accidental ablation of the penis at a young age. One such case has been used to argue that biology overrides rearing while another suggests the opposite. Generally, however, individuals with GID are physically normal and so psychosocial factors appear to be particularly relevant.

Parents have been shown to respond differently to male and female children, generally encouraging same-sex behavior more than opposite-sex behavior. Parents of children with GID tend to be neutral about if not encouraging of their child's cross-gender behavior. Further they appear to be less encouraging of same-sex behavior. Despite Stoller's premise of a 'blissful symbiosis' between a boy with GID and his mother, Green (1974) found that both parents spent significantly less time with their son with GID than did parents of controls. Further, both mothers and fathers of children with GID have elevated levels of psychopathology, which would suggest impairments in their relationships with each other and with their children. Lastly, clinical observations suggest high levels of conflict between children with GID and their parents, which these sensitive children interpret as rejection.

Etiological factors are reviewed more thoroughly in Zucker and Bradley (1995) where a complete set of references is also available. Clinical formulations of these etiological factors are available in Coates (1990) and Zucker and Bradley (1995).

4. Management

There are few formal studies assessing management of children or adults with GID. Current management of children with GID involves addressing the familial and child issues through different combinations of parent counseling, parent groups, and child psychotherapy.

Adolescents with GID may be referred for psychotherapy to rule out factors that if addressed might change the wish for sex reassignment and to ensure the stability of the cross-gender wish. Some clinics use hormonal treatment to reduce the development of secondary sex characteristics and reduce sex drive when such adolescents have been considered treatment eligible (Cohen-Kettenis and Gooren 1999). Psychological support is considered very important but often rejected by the adolescent.

Management of the adult with GID is somewhat similar to that of the adolescent but can proceed to surgical reassignment after the individual has demonstrated their capacity to surmount the hurdles of living as the other sex for a period of time (1–2 years). Once more, psychological support is considered critical but is not always welcomed by the individual with GID. Hormonal therapy (estrogens to male-to-females and androgens to female-to-males) are begun in some clinics at the beginning of the real-life test and in other clinics somewhat later. Sex reassignment surgery typically is not undertaken until sometime in early adulthood although those individuals with a history of Transvestic Fetishism may not have surgery until their thirties or forties.

5. Outcome

Children with GID can relinquish their desire to be the opposite sex. This appears to be facilitated when parents address some of the familial issues thought to contribute to the child's gender dysphoria. Children presenting at a later age (10–12) or whose families cannot address the contributing familial issues appear to do less well and continue to display significant gender dysphoria. Adults who undergo sex reassignment are generally satisfied although the capacity of surgery to produce well-functioning genitals, especially in the female-to-male, is quite limited.

6. Methodological Issues

Given the relative rarity of this condition sample sizes adequate to test hypotheses are only available in one or two clinics in the world. Evaluating treatment is similarly complicated by small sample sizes in any one

geographic location. Multisite studies are needed to address some of these issues.

7. Future Directions

Comparison of children with GID vs. contrast groups such as children with anxiety disorders are needed to address general (psychopathology) vs. specific (to GID) etiological factors. Evaluation studies of interventions in children, adolescents, and adults are needed.

See also: Body Image and Gender; Gender, Class, Race, and Ethnicity, Social Construction of; Gender Differences in Personality and Social Behavior; Gender-related Development; Identity in Childhood and Adolescence; Sexuality and Gender

Bibliography

Allen L S, Hines M, Shryne J E, Gorski R A 1989 Two sexually dimorphic cell groups in the human brain. *Journal of Neuroscience* **9**: 497–506

Bailey J M, Dunne M P, Martin N G 2000 Genetic and environmental influences on sexual orientation and its correlates in an Australian twin sample. *Journal of Personality and Social Psychology* **78**: 524–36

Bailey J M, Willerman L, Parks C 1991 A test of the maternal stress theory of human male homosexuality. *Archives of Sexual Behavior* **20**: 277–93

Benjamin H 1966 *The Transsexual Phenomenon*. Julian Press, New York

Blanchard R 1991 Clinical observations and systematic studies of autogynephilia. *Journal of Sex and Marital Therapy* **17**: 235–51

Bradley S J 2000 *Affect Regulation and the Development of Psychopathology*. Guilford, New York

Byne W, Tobet S, Mattiace L A, Lasco M S, Kemether E, Edgar M A, Morgello S, Buchsbaum M S, Jones L B (in press) The interstitial nuclei of the human anterior hypothalamus: An investigation of variation with sex, sexual orientation and HIV status. Hormones and Behavior

Coates S 1990 Ontogenesis of boyhood gender identity disorder. *Journal of the American Academy of Psychoanalysis* **18**: 414–38

Cohen-Kettenis P T, Gooren L J G 1999 Transsexualism: A review of etiology, diagnosis and treatment. *Journal of Psychosomatic Research* **46**: 315–33

Green R 1974 *Sexual Identity Conflict in Children and Adults*. Basic Books, New York

Green R 1987 *The 'Sissy Boy Syndrome' and the Development of Homosexuality*. Yale University Press, New Haven, CT

Green R, Money J (eds.) 1969 *Transsexualism and Sex Reassignment*. Johns Hopkins Press, Baltimore, MD

Hamer D H, Hu S, Magnuson V L, Hu N, Pattatucci A M L 1993 A linkage between DNA markers on the X chromosome and male sexual orientation. *Science* **262**: 321–27

Lalumiere M L, Blanchard R, Zucker K J 2000 Sexual orientation and handedness in men and women: A meta-analysis. *Psychological Bulletin* **126**: 575–92

LeVay S 1991 A difference in hypothalamic structure between heterosexual and homosexual men. *Science* **253**: 1034–7

Meyerowitz J 1998 Sex change and the popular press: Historical notes on transsexuality in the United States, 1930–1955. *Gay and Lesbian Quarterly* **4**: 159–87

Money J, Hampson J G, Hampson J L 1955 An examination of some basic sexual concepts: The evidence of human hermaphroditism. *Bulletin of the Johns Hopkins Hospital* **97**: 301–19

Stoller R 1968 *Sex and Gender*. Science House, New York

Swaab D F, Fliers E 1985 A sexually dimorphic nucleus in the human brain. *Science* **228**: 1112–5

Swaab D F, Hofman M 1990 An enlarged suprachiasmatic nucleus in homosexual men. *Brain Research* **537**: 141–8

Zhou J N, Hofman M A, Gooren L J G, Swaab D F 1995 A sex difference in the human brain and its relation to transsexuality. *Nature* **378**: 68–70

Zucker K J, Beaulieu N, Bradley S J, Grimshaw G M, Wilcox A 2000 Handedness in boys with gender identity disorder. Journal of Psychology and Psychiatry

Zucker K J, Bradley S J 1995 *Gender Identity Disorder and Psychosexual Problems in Children and Adolescents*. Guilford, New York

S. J. Bradley and K. J. Zucker

Gender Ideology: Cross-cultural Aspects

Gender ideology is concerned with normative beliefs about the proper roles for and fundamental natures of women and men in human societies. The distinction between sex and gender is central to the concept of gender ideology. When a child is born, the biological differences between males and females, or their sex differences, are used as the basis for the assignment of gender and the cultural construction of gender identities. People are assigned a gender identity in all societies, but gender systems and the gender ideologies that are thought to help sustain them are culturally variable.

1. The Historical Context for Studies of Gender Ideology

The Women's Liberation Movement of the late 1960s and 1970s, which started in the United States and then spread to Europe and other parts of the world, was an important stimulus for cross-cultural research on gender ideologies, and the politics of the movement significantly influenced this research as it emerged in the early 1970s. The most general political position of the Women's Liberation Movement that shaped the study of gender ideologies was the view that women are not equal to men in American society. They do not have the same control over their own lives and the lives of others that men have. They are dominated by men in their family life, in the workplace, and in other social domains as well, particularly religion and politics. This domination, it was argued, is bolstered by patriarchal gender ideologies that provided justification for men's domination of women. The term 'patriarchal' was used to refer to ideologies that either

assumed or asserted that the correct social order was one in which men dominate women, have authority over them, and tell them what to do. The patriarchal ideology that received the greatest attention in the Women's Liberation Movement was the American view that women are biologically inferior to men—less intelligent, physically weaker, less aggressive, and more emotional—in ways ultimately explained by differences in their biological makeup. The political theorists of this social movement actively criticized this patriarchal ideology. They acknowledged that there were biological differences between women and men, but argued that these were exaggerated by cultural processes and were in themselves not sufficient to be a basis for men's political and social domination. Instead, differences in women's and men's socialization experiences explained cultural differences between women and men that were used to justify men's authority. Academic disciplines that had contributed to the development of biological patriarchal ideology, also came under attack from the women's movement. 'Science' in general was conceptualized as a form of ideology, that is as a framework promulgated by people in power to help them maintain their power, rather than as objective factual truth. There was evidence that scientists who had studied sex differences between humans had indeed been influenced by the patriarchal ideologies of European cultures. For example, nineteenth century physical anthropologists found that women's brains were smaller than men's and argued that this was the reason women were less intelligent. Yet by the early twentieth century, physical anthropologists were arguing that intelligence was a reflection of the size of an animal's brain in proportion to its body, and women and men did not differ in this proportional relationship.

Patriarchal gender ideology was also apparent in the way women were largely absent from scholarly representations of human activity in the social sciences and humanities of the 1960s. For example, in anthropological theory about early human societies, considerable attention had been given to the importance of men's collaborative hunting for human evolution. But women's gathering of plant foods, which provided the bulk of foodstuffs eaten by early humans, was almost completely ignored. Feminist theorists argued that both biological theories which denigrated women and representations of humans that rendered women invisible conveyed gender ideologies which were psychologically harmful to girls and young women. These scholarly visions of gender undermined young women's self-confidence, encouraging them to believe that they were incapable of much in life that they were in fact quite capable of, and providing perspectives that suggested male domination was indeed an appropriate state of affairs.

This vision of the role of gender ideology in US women's subordination was projected to the world as a whole in early feminist anthropological comparative research on the topic. In the early 1970s, Rosaldo (1974) argued that the cross-cultural evidence on gender relations available at that time indicated that in all societies of the present and past men have had greater power and authority than women. In all societies, although both men and women have culturally legitimated authority in the domestic or household sphere, it is overwhelmingly men who have had culturally legitimated authority in public spheres, due to women's greater involvement in childcare. Rosaldo suggested that this arrangement was supported everywhere and reinforced by gender ideologies in which men were held to be superior to women by virtue of their greater involvement in culture and their lesser involvement in nature. Women were held to be more involved in nature than men because of their role in the natural process of giving birth to and caring for children. This argument was appealing in part because of its scope: it was a theory about women and men in all times and places. It was also appealing because it resonated with the views of feminists involved in the Women's Liberation Movement of that time.

In many ways, Rosaldo's position set the agenda for comparative research on gender ideology for the next several decades. Her position was challenged as well as embraced and the research it stimulated showed that there is far more diversity in gender ideologies than was initially envisioned. The rest of this discussion focuses on the nature of this diversity in gender ideologies. Consideration will first be given to how many gender categories there are in gender ideologies. Early work assumed two—woman and man—but this view has been questioned. Then the discussion will shift to the nature of cross-cultural similarities and differences in fundamental features of gender ideologies that are predominantly dualistic in organization. Finally, the extent to which patriarchal gender ideologies sustain male domination cross-culturally will be considered.

2. *How Many Genders Are There?*

One key source of cultural and historical variation in gender ideologies is the extent to which gender systems are thought of as fundamentally dualistic, or binary. Much contemporary Western scholarship assumes that binary gender systems are universal: sexual females become women and sexual males become men. However, research on sex and gender ideologies suggests that this is itself a culturally specific gender ideology. To begin with, binary biological sex is itself a cultural construct. Laqueur (1990) argues that until the nineteenth century, Europeans conceived of only one sex, that of male, with female as a lesser version of him. It was only in the late eighteenth century that a two-sex model of human biology began to emerge in which the differences between males and females were rhetorically elaborated. Today, although it is common to assume that all babies will be obviously biologically

male or female, sex is sometimes ambiguous. In such cases an infant will still be assigned to one category or the other and surgery may be performed to alter the body in a manner consistent with that assignment. Adults sometimes decide that they want to have the physical characteristics of the 'opposite' sex. To achieve this they engage in cultural activities that will either create this impression, as when women bind their breasts, or bring about sex changes as a physical reality through surgery and hormonal treatments.

There is no doubt that binary biological sex is partly a material reality. It is not something that we as humans have imagined completely. But it is also partly a cultural construct. Scholars agree that binary gender ideologies, as well as binary sex ideologies, are ideological constructs. They also agree that the gender ideologies articulated by people do not capture for any society the range of gender identities expressed in that society. But they disagree about how to conceptualize nonbinary models of gender. Some researchers assert the existence of societies in which there are three rather than two salient genders, where the third gender is most commonly a named category referring to a man who is womanly in dress, occupation, sexual preference, or some combination of these aspects of gender identity. For example, Europeans who first contacted North American Indian societies reported that many of them had a third gender category of men who dressed like women and carried out women's work roles, the name for which was sometimes glossed into English as 'manly hearted woman.' Other scholars argue that even in cultures with three or more named gender categories, gender as an ideological system is still and always about cultural transformations of a biological difference between males and females. In any case, gender dualism, or the idea of two genders, woman and man, is very strongly developed and salient in gender ideologies around the world and through time, and most scholars working on gender ideologies have focused on the nature, causes, and consequences of gender ideologies that posit woman and man as central categories.

3. *Cross-cultural Similarities and Differences in Dualistic Gender Ideologies*

Dualistic or two-category gender ideologies are concerned with how women and men are similar and different, with the emphasis on difference. The widespread presence of ideologies of difference indicates the importance of culture and ideology in creating women and men as unalike. Anthropologists have found that cross-culturally, people are concerned with similar aspects of human experience in their production of gender ideologies. People also engage in similar cognitive and symbolic processes as they talk about and otherwise represent women. At the same time, there are also differences from society to society and within societies in the nature of dualistic gender ideologies, so that similarities and differences across cultures are intertwined.

There are at least four aspects of human experience that regularly enter into gender ideologies. These include work, appearance, sexuality, and reproduction. People constitute themselves and are socially constituted by others as women and men through their behavior in these areas and the meanings attached to those behaviors. At the same time, societies vary in the emphasis that is given to each of these four aspects and in the extent and kind of variation tolerated in behavior relative to canonical expectations about gender.

It is very common cross-culturally for women and men to be expected to do different kinds of work, and for girls and boys to be socialized into different kinds of labor. So, for example, in hunting and gathering societies, which no longer exist in isolation from other economic types of societies, it is common for women to be expected and trained to gather plants and for men to be expected and trained to hunt for game. Among European agriculturalists, large-scale cultivation has been associated strongly with men. Consequently, where women were largely responsible for cultivation in places that have been colonized by Europeans over the last 500 years, strong pressure was often exerted to get men to replace women in these activities as, for example, with the Seneca Indians in New York state.

After about the age of 3 years, girls and boys and women and men are also commonly expected to look different from each other in ways that are specific to individual cultures. They are expected to dress differently, to wear their hair differently, to differ in the presence and nature of body decor such as jewelry and tattoos, and to differ in patterns of nonverbal behavior such as gesture and facial expression.

Sexually, binary gender ideologies entail the expectation that men's sexual object choices will be women and women's sexual object choices will be men. But whether or not and how departure from these expectations leads to the view that a male is not a man or a female is not a woman varies greatly cross-culturally.

Finally, binary gender ideologies commonly entail expectations of women and men regarding the roles they will play in the reproduction of human communities, through giving birth to, taking care of, and socializing children. It is well documented that failure to meet cultural expectations in these areas can often have major consequences in the lives of the people involved. However, a worldwide decrease in emphasis on this aspect of gender identity may be ongoing as changes in global economic processes bring about changes in family size.

Cognitively and symbolically, it is common for these substantive areas of gender ideology to be elaborated around specific gender dyads (Mathews

1992). Typically these dyads are drawn from cross-gender relationships within the family, reflecting the fundamental centrality of the family for both the maintenance of and changes in gender ideologies. Some key nuclear family dyads that are frequently ideologically elaborated cross-culturally include wife–husband, sister–brother, mother–son, and father–daughter.

However, societies differ in which gender dyads get selected for ideological elaboration. For example, in many European societies, the husband–wife relationship is the primary relationship that is ideologically elaborated, and the mother–child relationship is also the focus of a great deal of cultural production. But in the South Pacific, in Polynesian cultures such as those of Samoa, Tonga, and Tokelau, it is the sister–brother relationship that is most celebrated and discussed, and the husband–wife relationship is simply not developed explicitly or represented to the same extent in public events. So, for example, stories and plays depict the kind of work the sister should do and the kind of work the brother should do, but not what husbands and wives do, though one may infer that the division of labor will be similar for them.

While a good deal of gender ideology develops around nuclear family dyads, it also receives expression in representations of culturally idealized individual men and women, who usually represent the good, but can also represent the bad. For example, in Mexico, there are three key female icons: the Virgin of Guadalupe, La Malinche, an Indian woman who betrayed her people by consorting with the Spanish conqueror, and La Llorona, a ghostly figure who wanders riverbanks weeping over the loss of her dead children and luring men to death by drowning (Limon 1986).

The use of simple cross-gendered dyads from the family and the use of individual icons to convey gender ideology are, then, common cross-cultural cognitive and symbolic strategies. An additional such strategy is to metaphorically project gendered family dyads and individual gendered prototypes outside the family onto other more encompassing levels of social organization. For example, George Washington is referred to as the father of his country. And entire nations may be gendered, often as women, as in the expression, 'India is the mother of her people' and the term 'Mother Russia.'

The processes involved in cross-cultural similarities and differences in gender representation that have been described so far also provide resources for the production of intrasocietal as well as intersocietal similarities and differences in gender ideologies. Different segments of a society—e.g., classes, ethnic groups, or religious groups—may emphasize different aspects of gender identity in their gender ideologies. Diversity in gender ideologies within a society can consequently become the basis for criticism and conflict between groups over how women and men should be expected to behave and should treat one another. This can lead to the imposition of the gender ideology of one group on another, and to changes over time in gender ideologies.

For all of these reasons it can be problematic to talk about 'the' gender ideology of a given society or culture. It is true that some ideas about gender may be 'dominant' in a given society, in that they are represented widely or represented by those in power. But diverse views on gender are held in all societies, if only by virtue of role differentiation within the family, so that what is expected of a sister is not what is expected of a wife or mother.

4. *Gender Ideologies and Male Domination Over Women*

The final important and controversial aspect of gender ideologies to be considered here is the extent to which they provide justification for men's control of women and contribute to its perpetuation. Patriarchal gender ideologies are very widespread, but not all of them are alike. The political critique of the Women's Liberation Movement concentrated on the American and European patriarchal ideology that gender differences are rooted in biologically based sex differences which render women inferior to men. But this was and is not the only patriarchal gender ideology in the United States or elsewhere. Biological differences between women and men are not always involved. Nor is women's inferiority always asserted. Neither is necessary for a patriarchal gender ideology. What is necessary is that there be a cultural understanding that men should have power and authority over women that women should not have over themselves or men. Two important bases for patriarchal gender ideologies besides biology are tradition and religion.

Adhering to tradition means continuing to do something because it has been done in the past, often a highly valued past that itself defines who people are in the present. In religious ideologies, in contrast, patriarchal gender ideologies are often conceptualized as supernaturally ordained. This is true of gender ideology in the three world religions that originated in the Middle East and spread around the world: Judaism, Christianity, and Islam. In all these religions, gender roles and the greater authority of men is interpreted as the will of God.

Although patriarchal gender ideologies are very widespread, nonpatriarchal gender ideologies and gender ideologies that do not denigrate women but value them and represent them in positive ways are also very widespread. Quite often they co-exist with patriarchal gender ideologies. In Tonga, for example, gender ideologies reveal different attitudes toward and expectations for sisters and wives. Women are conceptualized as below their husbands, but above their brothers, so husbands can tell wives what to do and

sisters can tell brothers what to do, but brothers are not supposed to boss their sisters and wives are not supposed to boss their husbands. A good deal of comparative research on gender ideologies assumes or asserts a close relationship between the amount of power women have and the extent to which gender ideology portrays women positively in a given society. Both women's power and gender ideology about them are in turn affected by aspects of cultural systems other than the gender frameworks themselves. The idea that gender ideologies and power relations between women and men vary in ways shaped by other social forces is important. It raises the possibility that women's positions can be enhanced by human transformation of social systems through identification of processes in them that favor gender equality. Various aspects of social systems have been put forward as influencing gender ideologies and power relations. For example, many scholars have argued that women have more power and more positive images in societies where they make significant visible contribution to the economic well being of the groups they belong to. Others argue that although present day societies may all have patriarchal gender systems, this has not always been the case. Some argue that 'states' as a type of political system entail a concentration of power in men's hands that was not true of prestate political systems in the evolution of human societies. Others focus on European colonialism over the last several centuries as a cultural influence that has made many societies patriarchal which were not patriarchal before. However, these ideas have been criticized for romanticizing past societies whose gender ideologies and power relations may never be completely reliably retrieved. Since the 1990s, there have also been changes in the way the power of gender ideologies has come to be conceived. It is increasingly common for anthropologists and others to write of the 'gendering' of very large-scale social processes, particularly colonialism, the emergence of nation states, and international relationships. This work suggests that the ideological power of gender is located not so much in explicit representations of gender in myths and rituals as in a more implicit, pervasive, and diffuse semiotic gendering of all of social life. Still, the most important contribution to date from the study of gender ideologies has been the documentation of considerable cross-cultural diversity in gender ideologies that describe and explain how and why women and men are different from each other. The documentation of this diversity in itself undermines and deconstructs essentialist views that differences between women and men are natural, supernaturally ordained, or fixed for all time in any society.

See also: Belief, Anthropology of; Culture as Explanation: Cultural Concerns; Economic Development and Women; Feminist Theory; Fertility and Culture: Anthropological Insights; Folklore; Foucault, Michel (1926–84); Keynes, John Maynard (1883–1946); Male Dominance; Masculinities and Femininities; Mead, Margaret (1901–78); Witchcraft

Bibliography

Abu-Lughod L 1986 *Veiled Sentiments: Honor and Poetry in a Bedouin Society*. University of California Press, Berkeley, CA

Bozzoli B, Nkotsoe M 1991 *Women of Phokeng: Consciousness, Life Strategy, and Migrancy in South Africa 1900–1983*. Heinemann, Portsmouth, NH

Kennedy E L, Davis M D 1993 *Boots of Leather, Slippers of Gold: The Story of a Lesbian Community*. Routledge, New York

Laqueur T 1990 *Making Sex: Body and Gender from the Greeks to Freud*. Harvard University Press, Cambridge, MA

Limon J 1986 La Llorona, the third legend of greater Mexico: Cultural symbols, women and the political unconscious. *Renato Rosaldo Lecture Series Monograph* **2**: 59–93

MacCormack C P, Strathern M (eds.) 1980 *Nature, Culture and Gender*. Cambridge University Press, New York

Mani L 1998 *Contentious Traditions: The Debate on Sati in Colonial India*. University of California Press, Berkeley, CA

Mathews H 1992 The directive force of morality tales in a Mexican community. In: D'Andrade R G, Strauss C (eds.) *Human Motives and Cultural Models*. Cambridge University Press, New York

Ortner S B, Whitehead H (eds.) 1981 *Sexual Meanings: The Cultural Construction of Gender and Sexuality*. Cambridge University Press, New York

Reiter R R (ed.) 1975 *Toward an Anthropology of Women*. Monthly Review Press, New York

Rosaldo M Z 1974 Woman, culture, and society: A theoretical overview. In: Rosaldo M Z, Lamphere L (eds.) *Woman, Culture, and Society*. Stanford University Press, Stanford, CA

Sanday P R 1981 *Female Power and Male Dominance: On the Origins of Sexual Inequality*. Cambridge University Press, New York

Sanday P R, Goodenough R G (eds.) 1990 *Beyond the Second Sex: New Directions in the Anthropology of Gender*. University of Pennsylvania Press, Philadelphia

Stoler A L 1989 Rethinking colonial categories. *Comparative Studies of Social History* **31**(1): 134–61

Yanagisako S, Delaney C (eds.) 1995 *Naturalizing Power: Essays in Feminist Cultural Analysis*. Routledge, New York

S. U. Philips

Gender-related Development

Being born a girl or boy sets a wide range of forces in motion. These forces go beyond the genetic, hormonal, and genital differences inherent in one's sex. Parents provide many opportunities for learning about what it means to be a boy or girl. Teachers and peers reinforce and encourage gender-typed behaviors in children. Children also become active participants in learning about gender. Gender development is a term that

encompasses all the many gendered aspects of children's development. To understand children's gender development, the entire range of social, cognitive, and biological influences must be considered.

1. Overview of Gender Development

In recent years, gender has come to be viewed as a multidimensional construct that covers many domains including biological sex, activities and interests, personal–social attributes, gender-based social relationships, gender-based styles and symbols (such as gestures, style of dress), and gender-related values (see Ruble and Martin 1998, Huston 1983). Each domain can be examined in terms of an individual's concepts, self-perceptions, preferences, and behaviors. In the past, there was an implicit assumption that a person who was gender typed in one domain also was gender typed in other domains. The multidimensional view of gender has been useful in questioning this assumption and replacing it with the challenge to researchers that interrelations among domains need to be assessed. Also, this view has the advantage of making clear the distinctions among the different domains.

Certain aspects of gender development have been investigated more than others, including gender roles, self-perceptions of gender, gender identity, and gender-related behaviors. Gender roles have been a popular topic of study, within both psychology and sociology. Gender role is a broad term referring to the behaviors, attitudes, and traits that are associated with being male or female. Many studies have been done to identify the roles that men and women fulfill in western cultures as well as in other countries and in our history. Gender stereotypes, which are culturally defined expectations about the sexes in terms of personality, appearance, occupations, abilities, and interests, have been a very central subject for social psychologists. Also, psychologists have been involved in studying self-perceptions of gender, especially individual's perceptions of masculinity and femininity. Gender identity, the sense of oneself as being a male or female, has been studied by clinical and developmental psychologists, whereas a typical gender identity has been investigated largely by medical researchers. Gender-related behaviors and characteristics have been the subject of numerous studies, including how children develop skills and social behaviors such as their relationship preferences.

2. Age Trends in Gender Development

2.1 Gender Identity

Gender identity is the sense of oneself as a male or female. Gender identity involves more than the anatomical–it involves the inner knowledge and outer expression of one's sex. The dominant view has been that gender identity resides within the individual, although more recent views challenge this assumption by arguing that gender identity is created in the context of interactions, societal structures, and cultural expectations (Deaux and LaFrance 1998). Research on gender identity has also focused on the consequences of developing gender identity that is consistent with or different from one's genetic sex.

The predominant view has been that for typical children, the process of developing a sense of oneself as a girl or boy undergoes three stages. First, at about the age of 2–3 years, children learn to label themselves and others by sex. Second, 3–4-year-old children come to understand the stability of gender over time. Third, by the age of 4–5 years, children recognize that one's sex is a constant aspect of oneself that is not variable over situations (Slaby and Frey 1975). As children's sense of gender identity develops, they often show increased attention to people of the same sex, and are motivated to be like members of their own sex (Kohlberg 1966).

A very small group of children, about 2–5 percent of the population, develop a sense of gender identity that does not match their biological gender. These children, labeled as having gender identity disorder, express discomfort with their own sex, wish to be the other sex, and engage in behavior typical of the other sex (Zucker and Bradley 1995). The development of gender identity disorder may be related to biological factors, such as hormones and temperament, in addition to environmental factors, such as parental encouragement of cross-gender play.

An even smaller group of children are born with ambiguous genitalia or genitalia that do not match their genetic sex. These children may be assigned a sex that differs from their genetic endowment or may not be assigned a sex until later. Gender assignment, and also gender identity, become very complicated issues for these children. Although early research suggested that children who were assigned sex at an early age would develop according to their assignment regardless of their biological sex, more recent research has challenged this view. In some reports, both biological and social factors play a role in developing gender identity, whereas in others, only biological sex or only gender assignment is central in developing gender identity. Controversy continues over how to best support these children in developing their gender identities (see Ruble and Martin 1998).

2.2 Gender Differences in Interests, Behaviors, and Well-being

As early as 2 years of age, children begin to show sex-differentiated activities and interests (Ruble and Martin 1998). In the preschool years, children's preferences for stereotypic play become even more apparent. Girls and boys prefer different toys and prefer to play in different areas. The tendency to play with stereotypic

toys increases during early childhood. Both girls and boys show preferences for stereotypic activities, although boys seem to avoid other-sex activities more than girls.

Sex differences in activities persist into middle childhood and adolescence although the similarities in behavior are more striking than differences. Boys are more involved in sports, outdoor chores, and leisure activities; girls are more involved in indoor chores, shopping, and socializing. Through adolescence, girls spend increasing amounts of time in interpersonal activities, personal care, and household chores and less time in sports whereas boys' interest in sports remains stable or may increase.

Males are more active than females with the differences being more pronounced in older than in younger children. Girls and boys differ in many motor tasks, with boys outperforming girls in throwing and girls outperforming boys in tasks involving flexibility and fine-motor skills. These differences increase with age, possibly owing to differential participation of girls and boys in sports after puberty, although biological factors may also play a role in differences (see Ruble and Martin 1998).

Moderate and consistent sex differences are found in the area of physical aggression, with boys showing more aggression than girls. The exception to this is relational aggression—using behaviors to damage friendships—which girls engage in more than boys. Sex differences in aggression are larger in children than in adults, although men commit many more violent crimes than women.

The sexes differ to only a small degree in prosocial behavior and differences vary depending on the kind of behavior being assessed. Generally, girls help other people more than boys (Eisenberg and Fabes 1998). Females also are more accurate in decoding emotions from visual cues and they are more expressive and responsive than males.

Many varied indicators of well-being suggest that boys experience higher levels of well-being than girls. Beginning in adolescence, females are more likely to be clinically depressed or have depressive symptoms than males. Although the differences are small, boys report higher levels of self-esteem than girls and several studies suggest that girls experience a drop in self-esteem after the age of 9 years as compared with no change for boys. Girls have many more appearance and weight concerns than boys and these concerns increase during adolescence (see Ruble and Martin 1998). Girls' concern with their appearance may relate to their self-esteem and to eating disorders and disturbances.

2.3 Gendered Social Relationships

One of the most pervasive and consistent differences between the sexes is their choice of play partners (Maccoby 1990). Girls and boys are often found playing with members of their own sex. The phenomenon of sex segregation appears to be virtually universal across cultures. Sex segregation varies in differing situations, however. Children's tendency to interact with same-sex children is common in situations in which children have choices about playmates, such as in school, and in unstructured situations, and is less common when fewer choices are available, such as playing in their neighborhoods, or when children are in structured situations in which they need to cooperate.

Sex segregation begins very early in life. Girls tend to segregate before boys, usually starting between 2 and 3 years of age. Boys who are more socially disruptive and girls who are more socially sensitive segregate earlier than other children (Moller and Serbin 1996). Sex segregation increases as children grow older. Sex segregation continues until middle adolescence when children begin to form heterosexual dating couples. Children's friendship patterns mimic their play partners in that most often, children's friends are of the same sex as themselves.

The kinds of play typical of girls' groups and boys' groups are different. Boys tend to play more actively, encompass more space, and they often play outdoors. Girls tend to play indoors, more quietly, and use less space. Girls are more likely to encourage cooperation and interaction among play partners; boys are more often interested in issues of dominance and competition. From their experiences with same-sex partners, girls and boys learn different methods of interacting with others and may be considered to be socialized in 'different cultures' (Maccoby 1990).

The origins of sex segregation are unclear. One contributing factor to sex segregation is similar interests. Children tend to play with same-sex children regardless of activity, so similarity in interests is not the only explanation for sex segregation. Compatibility in play styles is also important. Children show higher levels of social interaction when playing with their own sex, which suggests that same-sex play is more entertaining than other-sex play. Activity level may also be a factor. Boys tend to play actively and that may be compelling to other boys but not to girls. Finally, children's expectations about the sexes also likely contribute to sex segregation (Ruble and Martin 1998).

The developmental course of social relationships in later adolescence and adulthood, especially the development of sexual orientation, has been difficult to chart. Although both sexes show very strong propensities to select other-sex romantic partners, about 5 percent of the population selects same-sex partners. The origins and developmental pathways of sexual orientation have been hotly debated. The central issues appear to be the extent to which sexual orientation is influenced by biological factors, such as genes, hormones, and brain structures, or by social factors, such as one's upbringing.

2.4 Gender Differences in Cognitive Skills

Many studies have been conducted to examine whether the sexes differ in cognitive abilities. In many areas, no sex differences are found, e.g., in overall intelligence or in analytic reasoning (Halpern 1992). When the sexes differ in a cognitive ability, the differences tend to be fairly small and hence are unlikely to account for differences in career choices or trajectories. Nonetheless, the topic of sex differences in cognitive abilities remains controversial.

Most of the research on sex differences in cognitive abilities has focused on three areas—mathematical, spatial, and verbal abilities. In the domain of mathematical abilities, the patterns of similarities and differences vary depending on many factors including the particular skill being assessed, the age and skill level of the sample, and assessment procedures. Differences tend to be larger and favor males when studies are conducted on the most mathematically skilled children and when using certain standardized tests, such as the SAT. Also, sex differences vary for different skills, with girls receiving higher grades than boys in mathematics courses, and a small difference in favor of girls for computational skills in elementary school. Males show better performance in problem-solving tasks beginning in junior high or high school. Some evidence suggests that sex differences in mathematical abilities are lessening over time.

The sexes differ in some aspects of spatial abilities but the size of the difference depends on the particular ability being assessed and the age of the participants. Few differences are found between young girls and boys. Most sex differences begin to appear in adolescence and differences increase with age. A moderate-size difference favoring adolescent and adult males is found in tasks involving mental rotation, which involves the rapid manipulation of three-dimensional objects. A moderate to small difference favoring males occurs in tasks involving recognition of the horizontal or vertical. A small difference (or no difference) is found when assessing performance on spatial visualization tasks, which involves finding hidden spatial figures (Halpern 1992).

Very small sex differences are found in verbal tasks, with females outperforming males. The differences are found most often in the areas of spelling and language and tend to be stable across age. Verbal differences may be lessening over time although not all researchers concur with this conclusion.

In a large-scale assessment of adolescents' academic abilities, researchers have found that males outperformed females in tests of science, mathematics, social studies, and vocational aptitude (mechanical reasoning and electronics), whereas females outperformed males in reading comprehension, perceptual speed, associative memory, and writing (see Ruble and Martin 1998).

2.5 Gender Stereotypes

Gender stereotypes are the beliefs that people have about the characteristics of males and females. The content of stereotypes varies over cultures and over time. These expectations are often related to the roles that the sexes fulfill in the culture.

Children learn some aspects of stereotypes at a very young age. By the age of $2\frac{1}{2}-3$-years, children show evidence of having some rudimentary knowledge of the activities and objects associated with each sex. Children's gender stereotypes of activities and occupations develop quickly during the preschool years, reaching a very high level by kindergarten. During the elementary school years, gender stereotypes broaden to include sports, school subjects, and personality traits. With age, children become increasingly knowledgeable about gender stereotypes and yet the rigidity of their stereotypes declines as they increasingly recognize the cultural relativity of these norms (Huston 1983). Some evidence suggests that boys hold more rigid gender stereotypes than girls and are held to more rigid ideals than girls. In adolescence, flexibility in stereotypes fluctuates in response to two opposing forces—increasing cognitive flexibility tends to increase adolescents' flexibility in applying stereotypes whereas increasing pressure to conform to stereotypes in preparation for sexual roles and adult status increases adherence to stereotypes (Ruble and Martin 1998).

Children use gender stereotypes to make inferences about others at a young age. When making judgments of other people, children and adults will apply their gender stereotypic expectations to them. Even more than adults, children will rely on a person's sex to make judgments and they are less likely to consider other relevant information about the person than adults are (see Ruble and Martin 1998).

3. Theories of Gender Development

3.1 Psychoanalytic Theory

The earliest and most influential theory of gender development was proposed by Sigmund Freud in the 1930s. Freud contended that boys and girls acquire gender identity by learning to negotiate desires for the opposite sex parent and by eventually identifying with the same-sex parent. Although little empirical support has been found for his theoretical framework, his ideas have been very influential in encouraging research on gender development. Over the years, additions and amendments have been made to Freud's original psychoanalytic theory. Many of these changes take into consideration the role of societal influences and include a feminist perspective.

3.2 Social Learning Theory

Social learning theorists (Bandura 1977, Mischel 1966) propose that gender development is accounted for by social factors. Girls and boys learn different roles through the effect of direct reinforcement, differential socialization, and imitation of models.

Direct reinforcement, which refers to the reaction children receive in response to their behavior, leads to children learning to discriminate between the behaviors typical of boys and girls. Children learn that there are different consequences for engaging in behaviors depending upon whether one is a girl or a boy. Behaviors considered to be stereotypic for their sex are more likely to be rewarded. Children perform activities that are consistent with their reinforcement histories (Mischel 1966).

Children also experience differential socialization of girls into feminine roles and boys into masculine roles as they are exposed to distinct environments in which gender differences are encouraged by socializing agents, such as parents, teachers, peers, and media. Parents fill boys' rooms with educational, interactive toys that promote exploration, whereas they fill girls' rooms with toys that promote nurturance and domesticity, such as doll houses and tea sets. Teachers respond differently to boys and girls: boys are more often called on to answer questions when their hands are raised than girls. Differential treatment by social agents provides boys and girls with separate sets of rules, expectations, and opportunities which, in turn, encourages traditional gender development.

Children also learn gender roles through observing others. Social learning theorists contend that individuals learn and reproduce the actions and responses of the models they encounter in their families, schools, and neighborhoods, and in books, television, and films. Gender-typed behavior develops as girls imitate the behavior of females and boys imitate the behavior of males. Children's imitation also is affected by their relationships with the model, in addition to the amount of power they perceive the model to possess (Mischel 1966, Bussey and Bandura 1984). Witnessing people's reactions to others as they display a behavior transmits information about the appropriateness of the behavior (Mischel 1966).

3.3 Cognitive Social Learning Theory

In the late 1970s and 1980s, social learning theory was adapted to emphasize the cognitive aspects involved in learning from the social environment (Bandura 1986). In 'cognitive social learning theory,' emphasis was placed on the processes involved in creating expectancies of other's behaviors and how these expectancies guide one's own actions. Specifically, it was proposed that observational learning required four processes: paying attention, being motivated to enact the behavior, having the ability to reproduce the action, and having the ability to remember the behavior (Bandura 1986).

Social learning approaches have encouraged an expanded view of the forces that influence gender development. This theory has motivated researchers to include a more diverse view of the child's social environment. In addition, it has led to closer scrutiny of the social settings provided for children and the gender-laden messages these settings contain.

3.4 Cognitive Developmental Theory

According to Kohlberg's cognitive developmental theory, children's cognitive organization of the world around them is the basis for their gender development (Kohlberg 1966). Kohlberg proposed that between the ages of 2 and 7 years, children make cognitive judgments about their gender identity (whether they are a boy or a girl) and then come to understand that gender is constant over time and situation. This understanding about gender becomes the foundation for further gender development. A child's recognition of being a male or female determines the value they place on masculine and feminine traits, initiates a child's desire to be similar to same-sex children, facilitates the development of gender stereotypes, and encourages children to hold positive views of their own sex (Kohlberg 1966).

A debate has ensued over whether Kohlberg's theory requires the acquisition of gender constancy as a necessary antecedent to learning stereotypes and attending to same-sex models (Ruble and Martin 1998). Empirical studies suggest that children have only a basic knowledge about gender before the onset of the gender-typing process, and that stereotyping can begin once children are able to label the sexes correctly. Gender constancy may help consolidate gender development, rather than initiating it (Ruble and Martin 1998). Cognitive developmental theory emphasizes the child's active role in gender development. Kohlberg's theory has been enormously influential in shaping the way researchers think about gender development.

3.5 Gender Schema Theory

Gender schema theory is a cognitively based theory that uses an information processing approach to explain how gender development occurs. The basis of this model is the cognitive representation called a schema. A schema is an organizing structure that helps simplify and categorize new information.

There are two types of gender-related schemas (Martin and Halverson 1981). The first is a general 'superordinate' schema that helps children categorize objects, characteristics, and traits into basic male and female categories. The second is a more narrow version of schema, called the 'own-sex' schema, that children

use to identify and learn in-depth information consistent with their own sex. These two schema types allow children to process information about events, objects, attitudes, behaviors, and roles and, in turn, categorize these aspects in terms of masculine or feminine, or as similar or different from the child (Martin and Halverson 1981).

Gender schema theorists (Bem 1981, Martin and Halverson 1981) proposed that gender schemas are developed and applied by children at an early age. In addition, it was proposed that the presence and use of gender schemas can explain the ways in which children and adults attend to, acquire, and remember information; how they behave; and the kinds of attitudes they develop (Martin and Halverson 1981). Gender schemas affect which kinds of information are noticed, encoded, and remembered, with information inconsistent with the existing schema being overlooked and consistent information being processed (Signorella et al. 1997). Research also shows that gender schemas can guide children's preferences, toy choices, and play partner choices.

Because schemas are used to interpret and categorize information, however, they can result in the drawing of incorrect inferences. Distortion of information has also been found to be a consequence of gender schematic processing. Altering information that does not fit into a currently held schema or placing neutral information into a gender schema leads to remembering information in a gender-consistent manner, which may not match with how it exists in reality. These processes illustrate how gender schemas can be easily maintained and how resistant they are to change.

Gender schema theory emphasizes the child as an active processor of information and that the gender schemas that guide thinking also influence children's behavior. The major strength of gender schema theory is in the understanding it provides about the maintenance and power of gender beliefs.

3.6 Biological Approaches

Much of the research using a biological approach has focused on understanding differences between the sexes. These approaches often consider, for instance, the influence of hormones and brain structure and function on sex differences.

Sex differentiation begins with the presence of an X or Y sex chromosome in the twenty-third pair of chromosomes. These sex chromosomes carry the genetic information that determines the development of fetal sex organs. The amount of a specific androgen, called testosterone, which is present in both male and female fetuses, determines further development of sex organs. Increased levels of this hormone lead to the formation of the penis and scrotum in males. Female sexual organ development continues in the absence of increased testosterone levels. Variations in the levels of hormones in these critical periods of development can lead to the formation of sex organs opposite to what would be expected from the sex chromosomes.

Hormones also influence the development of the central nervous system and the brain (Hoyenga and Hoyenga 1993). Hormones are most likely to affect brain development and behavior during two stages of life: during prenatal development and during puberty. Exposure to abnormal levels of hormones (due to genetic disorders or hormonal treatment during pregnancy) allows researchers to study the link between levels of hormones present during developmental processes and behaviors exhibited by individuals postnatally (Hoyenga and Hoyenga 1993). Studies of hormones, their effect on brain development, and the link to behavior suggest that they have an influence on gender development, although it is difficult to disentangle biological effects of hormones from social factors.

The structure and function of the brain have also been studied to compare the similarities and differences in males and females. One area of brain structure that has been studied is brain lateralization, or how much one hemisphere of the brain is responsible for carrying out specific functions. Some studies have shown that women are less lateralized in that they use both hemispheres simultaneously more often than men. It has been hypothesized that differences between men and women in a variety of abilities and behaviors could be due to a gender difference in lateralization. The findings of such studies show mixed support.

Another section of the brain that has been studied as a possible source of gender differences is the corpus callosum (CC). The CC is a group of fibers that connects the right and left hemispheres of the brain and is thought to be related to more intricate processes of thought, including language. Because the links between the structure, function, and size of the CC are not certain, research on differences in brain structure remains difficult to interpret. However, studies have shown that women have a larger CC than do men, and that enlargement of certain parts of the CC is linked to increased language skills.

Controversy surrounds the study of biological information on gender differences, especially concerning the extent of influence of biological and social factors. Most researchers of gender development accept an interactionist position, assuming that both nature and nurture are working together to produce human behavior. Furthermore, neither nature nor nurture should be viewed as steadfast or unmalleable (Hoyenga and Hoyenga 1993).

3.7 Evolutionary Perspectives

The evolutionary perspective draws on Darwinian principles of evolution and natural selection to explain

how gender differences evolve. Individuals who are fertile, live long enough to reproduce, and are able to raise their offspring will successfully pass on their characteristics to the next generation. Therefore characteristics of individuals who were successful in passing on their genes are more highly represented in future generations. According to the evolutionary viewpoint, gender differences arise because men and women play different reproductive roles (Trivers 1972).

Gender differences are hypothesized to develop because of differing parental investment levels. Evolutionary theorists contend that, because women are responsible for carrying the child during pregnancy and for producing food for the infant, their levels of parental investment are higher than men's. In contrast, men's reproductive role is limited to producing and investing their sperm. Thus, sexual selection differs for men and women. Because of a woman's increased investment level, she is more selective in her mates and searches for someone with access to resources. In contrast, men look for a woman who will care for the child, so that the child will eventually be fit to reproduce and pass on the family genes. However, because men never know if the child is actually carrying their genes, they are focused on mating more often and with shorter-term investments. Individuals who are most successful at their reproductive role are more likely to have their genes perpetuated, leading to a further divergence in men's and women's behaviors.

Evolutionary theoretical viewpoints have stirred much controversy, and empirical support for the theory's components is mixed. Critiques include the lack of ability to test the theory directly and the heavy dependence on animal studies.

4. Future Directions

The field of gender development moves quickly, making the prediction of future trends difficult. Nonetheless, some topics have garnered attention because they are important issues that have yet to be addressed. First, questions about the relation of mental health and domains of gender are worthy of additional research efforts. Why does the self-esteem of preadolescent girls drop more than that of boys? Second, research that focuses more attention on affective issues is needed to provide answers to questions about the role of self-regulation in gender development. Little is known about the affective and cognitive components that are involved in how people construct a view of their own gender and how others interpret gender identity. Third, many issues about how children are raised and about the contributions that parents, peers, and teachers can and do make to gender development are worthy of future research efforts. Finally, additional studies using multidimensional aspects of gender are important for understanding the complexities of gender development. For instance, multidimensional studies are necessary to understand how a traditional vs. nontraditional view of one's gender identity relates to a variety of outcomes such as internal psychological states, relationships, and achievement in school and careers. The picture of gender development continues to emerge but is far from complete.

See also: Education and Gender: Historical Perspectives; Family and Gender; Gender Differences in Personality and Social Behavior; Masculinities and Femininities; Psychoanalysis: Overview; Small-group Interaction and Gender; Social Learning, Cognition, and Personality Development; Social Networks and Gender

Bibliography

Bandura A 1977 *Social Learning Theory*. Prentice Hall, Englewood Cliffs, NJ

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall, Englewood Cliffs, NJ

Bem S L 1981 Gender schema theory: A cognitive account of sex typing. *Psychological Review* **88**: 354–64

Bussey K, Bandura A 1984 Influence of gender constancy and social power on sex-linked modeling. *Journal of Personality and Social Psychology* **47**: 1292–1302

Deaux K, LaFrance M 1998 Gender. In: Gilbert D, Fiske S, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York

Eisenberg N, Fabes R A 1998 Prosocial development. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*. J Wiley, New York, Vol. 3, pp. 701–78

Halpern D F 1992 *Sex Differences in Cognitive Abilities*, 2nd edn. L Erlbaum Associates, Hillsdale, NJ

Hoyenga K B, Hoyenga K T 1993 *Gender Related Differences: Origins and Outcomes*. Allyn & Bacon, Boston

Huston A 1983 Sex-typing. In: Damon W, Hetherington E M (eds.) *Handbook of Child Psychology*, J Wiley, New York, Vol. 4, pp. 388–467

Kohlberg L 1966 A cognitive-developmental analysis. In: Maccoby E E (ed.) *The Development of Sex Differences*. Stanford University Press, Stanford, CA, pp. 82–173

Maccoby E E 1990 Gender and relationships: A developmental account. *American Psychologist* **45**: 513–20

Martin C L, Halverson C F 1981 A schematic processing model of sex typing and stereotyping in children. *Child Development* **52**: 1119–34

Martin C L, Little J K 1990 The relation of gender understanding to children's sex-typed preferences and gender stereotypes. *Child Development* **61**: 1427–39

Mischel W 1966 A social learning view of sex differences in behavior. In: Maccoby E E (ed.) *The Development of Sex Differences*. Stanford University Press, Stanford, CA, p. 351

Moller L C, Serbin L A 1996 Antecedents of toddler gender segregation: Cognitive consonance, gender-typed toy preferences and behavioral compatibility. *Sex Roles* **35**: 445–60

Ruble D N, Martin C L 1998 Gender development. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*. J Wiley, New York, Vol. 3, pp. 993–1016

Signorella M L, Bigler R S, Liben L S 1997 A meta-analysis of children's memories for own-sex and other-sex information. *Journal of Applied Developmental Psychology* **18**: 429–45

Slaby R G, Frey K S 1975 Development of gender constancy and selective attention to same-sex models. *Child Development* **52**: 849–56

Trivers R 1972 Parental investment and sexual selection. In: Campbell B (ed.) *Sexual Selection and the Descent of Man*. Aldine, Chicago, pp. 1871–971

Zucker K J, Bradley S J 1995 *Gender Identity Disorder and Psychosexual Problems in Children and Adolescents*. Guilford Press, New York

C. L. Martin and L. Dinella

Gender Role Stress and Health

1. The Term 'Gender Role'

The term gender role is used in sociology and psychology to refer to the prescribed behaviors, attitudes and characteristics associated with one's gender status as a female or a male. Attributes associated with gender are the result of learning in accordance with cultural standards or prescriptions.

Gender-role socialization refers to the processes through which individuals acquire attributes appropriate to males and females. Examination of gender roles indicates differences in the content of roles seen as appropriate for men and women. Men are socialized more often than women to be active and to ignore their feelings of weakness. Men are expected to be more independent, autonomous, and strong. It is more acceptable for women to express a wide range of emotions, to pay attention to their moods, and to seek out others for help. While men may be encouraged to be independent, women are taught to be more interdependent. The term interdependence refers to the ability to express interpersonal needs, particularly in emotional relationships, and to relate meaningfully to others in relationships (Greenglass 1982).

2. The Term 'Stress'

Stress is often used to refer to the perception that the demands of the environment exceed the resources of the person, resulting in a threat to well-being. These may emanate from work, family, and daily hassles, as well as significant life events, and include time pressures, conflicts, and frustration in meeting one's goals (see *Stress and Coping Theories*).

2.1 Stress and Health

The way individuals cope with their stress is important for mental health. To the extent that individuals have resources to handle their pressures, they are less likely to suffer deleterious effects of stress. However, when individuals have difficulties in managing demands and stressors, they experience strain. Strain is a consequence of insufficient personal resources to meet demands. To the extent that individuals experience strain, they are more likely to experience deterioration in their physical and/or mental health. Some of the psychological effects of uncontrollable stress are depression, anxiety, panic attacks, hostility, and anger. High levels of stress may also result in physical symptoms including insomnia, headaches, muscular aches and pains, digestion disturbance, and difficulty in carrying out day-to-day activities (see *Stress and Health Research*).

2.2 Stress and the Feminine Gender Role

Research indicates that women are exposed to more stressful life events than are men and, as a result, have higher rates of psychological distress. Women experience depression, anxiety, and mood disorders more than men. Women may appear to be more distressed as a result of the kind of outcome measures typically used in this research. These studies tend to employ self-report measures of symptomatology. Some argue that women appear more distressed because they are more likely than men to report symptoms of all kinds. Gender-role factors contribute to the degree to which individuals report symptoms and seek help. Traditional masculine and feminine gender roles differ in the amount of vulnerability and permissibility in seeking help from others. Traditionally, men are seen as strong and invulnerable, thus restraining men from showing symptoms of mental illness or seeking help. The traditional feminine gender role allows and even encourages vulnerability to problems, thus resulting in greater reporting of symptomatology in women. It is assumed generally that if an individual is coping reasonably well, she/he is free from anxiety and depression. This conceptualization of effective coping leads to a bias in research findings given that women, more than men, often express distress (Greenglass and Noguchi 1996).

Notwithstanding women's greater 'openness' regarding their psychological symptomatology, research indicates many reasons for women's greater distress. Reproductive life events including menstruation, pregnancy, childbirth, and menopause are stressful life events for women. Within the employment sphere, women tend to be concentrated in low-paying 'female' occupational spheres, including education, healthcare, clerical, domestic, and service industries. Women have fewer promotional opportunities, make less money, and often do not have access to membership of the 'old boys' network,' all of which restricts their advancement in the employment sphere. Women professionals and managers are subject to negative stereotyped attitudes that portray women as doing a poor job

fitting into the masculine work world. Sex discrimination and harassment are ubiquitous in the workplace. Such attitudes not only harm women's salary and promotion prospects, they are also detrimental to their mental health. Women in the labor force also perform most of the domestic work, and on average spend more hours in housework and childcare than men. Employed married women manifest higher rates of distress than employed married men. Employed married women with young children experience higher rates of psychological distress than their female childless counterparts or comparable men. Given the demands on women when combining employment and familial roles, many women experience inter-role conflict and/or role overload (Barnett and Marshall 1991, Baruch et al. 1987).

Women's roles are inherently stressful in that they obligate them to respond to others, including the very young as well as the elderly. Informal care for elderly parents and relatives generally is provided primarily by women. These demands are a significant source of stress, thus playing an important role in women's mental health. Elements of the feminine gender role endanger women's mental health, since provision of care for their families and friends is often emotionally and physically draining, particularly given the relative lack of men's participation in work at home.

2.3 Stress and the Masculine Gender Role

Men have been diagnosed more than women with antisocial personality disorders in which there is a pervasive pattern of disregard for others' rights. Men are also diagnosed more as having schizoid personality disorders, characterized by detachment from social relationships and a restricted range of expression of emotions in interpersonal settings. These diagnoses may be seen as exaggerations of the traditional masculine gender role that stresses independence, separation, strength, and invulnerability (Greenglass 1982).

Research in the area of stress and burnout shows that men experience higher levels of depersonalization than women. Burnout at work commonly is measured using a scale (the Maslach Burnout Inventory) (Maslach and Jackson 1986) which yields three subscales: emotional exhaustion, depersonalization, and lack of personal accomplishment. Depersonalization refers to an attitude characterized by being impersonal, callous, and distant from others (Maslach and Jackson 1986). In this context, depersonalization may be regarded as a reflection of men's repressed emotionality and an ineffective coping form which allows men to continue their work with people yet remain untouched by the others' suffering.

The strict prohibition against emotional expressiveness in men (except perhaps anger) is one of the most powerful cognitive masculine mandates influencing men and their behavior. Male inexpressiveness is the result of gender role socialization that begins early in childhood. Men are taught that emotional expressiveness is a feminine trait and they learn to distance themselves from femininity. Masculine socialization contributes to and encourages aggressive behavior in men. Men are more involved than women in violent crime, including armed robbery, spouse and child abuse, and rape. While aggression in sports and the military is approved and rewarded socially, the social program associated with masculinity also includes competitiveness and aggression that are socially sanctioned as masculine coping styles.

While women more often express distress in internalized and depressive symptomatology (Chesney and Ozer 1995), men often cope with their feelings differently, since they are more likely to externalize their feelings such as becoming physically aggressive or by distracting themselves from their feelings by engaging in a variety of activities. The violence committed by men against women in its many forms is in part an outcome of a patriarchal society which is dominated by men who see themselves as controlling women in many different contexts. Research suggests that power, sexual control, and aggression are amplified in men in a patriarchal society. The mental health costs of male violence justify re-examination of the prescriptions of the masculine gender role (Greenglass and Noguchi 1996).

Masculine prohibitions and difficulties with emotional expressiveness make it difficult for men to maintain mutually supportive relationships. When under stress, men seem to be less flexible in their coping than women, since they have difficulty expressing positive feelings.

3. Stress, Coping, and Gender Roles

Research shows that the ways people cope with stress differ according to their gender role. Some argue that men are more problem-focused in their coping, and women are more emotion-focused (Folkman et al. 1986). For example, men have been reported to engage more in coping that alters a stressful situation. Some assert that men more often possess psychological attributes, i.e., self-esteem and mastery, which influence their coping. However, gender differences in mastery or self-efficacy are likely to be linked to differences in social experiences rather than to gender *per se*. A consistently unresponsive or negative environment has been found to affect a person's sense of self-efficacy, which in turn leads to anxiety and depression. Studies of gender differences in psychological characteristics have pointed out that models of stress and coping were developed primarily on men without consideration of women's roles and experiences, and may not encompass other aspects of coping including relationship-focused coping techniques.

In research on coping, findings indicate that when education, occupation, and/or position are controlled for, few gender differences are found in coping strategies (Greenglass et al. 1990). The type of coping an individual employs when dealing with stress is often related to the status, resources, and power associated with one's position. Many more women than men hold lower level jobs with less scope for the use of problem-solving and direct action strategies. Women are also more likely to experience high levels of demands, with low control over demands in the workplace (Greenglass 1985).

In coping research, the focus tends to be on individual behavior as the unit of analysis, and often, autonomous, agentic, and independent behavior is seen as synonymous with effective coping. But research shows that coping is highly specific to both the individual and the context. The social milieu in which a particular stressor is experienced should be incorporated into research paradigms that inquire into the relationship between gender and coping. Women may behave in a more passive way in groups with men, where stereotyped roles and social norms may be more strictly enforced. Women may appear different when acting alone or in a group of women.

3.1 *Gender, Social Support, and Mental Health*

There is a close relationship between positive health outcomes and social support. Close relationships can help a person cope with psychological distress. In such relationships people can disclose and discuss problems, share concerns, and receive advice keyed to their psychological needs. These relationships can also provide useful information, practical advice and morale boosting, all of which can assist individuals in dealing with their stressors (Cohen and Syme 1985). Both men and women benefit from supportive social interactions, especially intimate ones (see *Social Support and Health*).

Women are more emotionally involved in others' lives and serve a nurturing role for a wider network of people. In general, women give and receive more social support than men. Women have more close friends, and emphasize intimacy and disclosure in their friendships. Women's friendships are more involving and emotionally supportive. Social relationships with women are more health promoting than those with men. Wives are better providers of both instrumental and emotional support than husbands. Men tend to benefit from relationships with women; women generally benefit from relationships with their friends of the same sex (Belle 1982).

Research shows that the connection between social support and coping is stronger in women. Women employ more coping forms involving interpersonal relationships. Other research findings suggest that women use support from others through talking with one another. Women are able to make more effective use of their support networks, since they talk more as a way of coping with stress.

See also: Gender and Cardiovascular Health; Gender and Physical Health; Men's Health; Social Support and Stress; Women's Health

Bibliography

Barnett R C, Marshall N L 1991 The relationship between women's work and family roles and their subjective well-being and psychological distress. In: Frankenhaeuser M, Lunberg U, Chesney M (eds.) *Women, Work and Health: Stress and Opportunities*. Plenum Press, New York

Baruch G K, Biener L, Barnett R C 1987 Women and gender in research on work and family stress. *American Journal of Psychology* **42**: 130–6

Belle D 1982 The stress of caring: Women as providers of social support. In: Goldberger L, Bruntz S (eds.) *Handbook of Stress: Theoretical and Clinical Aspects*. Free Press, New York

Chesney M A, Ozer E M 1995 Women and health: In search of a paradigm. *Women's Health: Research on Gender, Behavior, & Policy* **1**: 3–26

Cohen S, Syme S 1985 Issues in the study and application of social support. In: Cohen S, Syme S (eds.) *Social Support and Health*. Academic Press, Orlando, FL

Folkman S, Lazarus R, Dunkel-Schetter C, De Longis A, Gruen R 1986 The dynamics of a stressful encounter: Cognitive appraisal, coping, and encounter outcomes. *Journal of Personality and Social Psychology* **50**: 992–1003

Greenglass E R 1982 *A World of Difference: Gender Roles in Perspective*. Wiley, Toronto, ON

Greenglass E R 1985 Psychological implications of sex bias in the workplace. *Academic Psychology Bulletin* **7**: 227–40

Greenglass E R, Burke R J, Ondrack M 1990 A gender-role perspective of coping and burnout. *Applied Psychology: An International Review* **39**: 5–27

Greenglass E R, Noguchi K 1996 Longevity, gender and health: A psychocultural perspective. *Japanese Health Psychology* **4**: 51–72

Maslach C, Jackson S E 1986 *Maslach Burnout Inventory Manual*, 2nd edn. Consulting Psychologists, Palo Alto, CA

E. R. Greenglass

Genealogy in Anthropology

1. *The Concept of Genealogy*

Genealogy owes its position in modern cultural and social anthropology to the genealogical method, which will be the main focus of this account, giving it an inevitable bias toward British social anthropology. The English term genealogy has three components: it refers to the descent of a person, family, or group from an ancestor or older forms; to a written or graphic account of those relationships; and to the study of

family pedigrees. English genealogy is closely associated with pedigree. Pedigree refers to the record of an ancestral line or lineage; the origin or history of something; a distinguished ancestry; and the recorded purity of breeding of an individual or strain. The overlap between genealogy and pedigree has been significant for the way genealogy has been incorporated by anthropology, especially in the study of kinship.

Anthropologists distinguished pedigree, as a vernacular account of ancestrally defined relationships, from genealogy as a systematic or scientific account of those relationships. The distinction rests upon the genealogical method, whereby the blood and marriage connections recognized by the people being studied could be systematically recorded. The genealogical framework obtained in this way provided the basis for the comparative study of terminological systems of address and reference, which were also sometimes called systems of relationship; for tracing marriage regulations (prescriptions and proscriptions); and for discerning the rules of descent and inheritance. Genealogies, or parts of them, collected in this way were (and are) often presented in graphic form to illustrate how persons or groups are related to one another. Genealogical information was the bedrock of comparative studies of kinship, descent, and marriage that were characteristic of structural-functionalist analyses of social organization during the first half of the twentieth century. Genealogy was also the basis for terminological and semantic studies of kinship, and used by alliance theorists. The use of genealogy to map the kinship system was even more pervasive, however, becoming a routine part of anthropological practice during the second half of the twentieth century.

2. *The Intellectual Context of Genealogy in Anthropology*

The genealogical method, developed by W. H. R. Rivers during the Torres Strait Expedition of 1898, was formalized in an article published in 1910 and became an important entry in the 1912 *Notes and Queries in Anthropology* and subsequent editions. The nineteenth century intellectual context of this addition to the anthropological instrumentarium lay in a rich combination of biblical, philological, heraldic, and evolutionary ideas. Darwinian theories concerning the origin of species and the descent of man had recourse both to linguistic and graphic genealogical imagery. The familiarity of biblical and historical family trees made genealogy a perfect vehicle for translating new ideas about the relationships between all forms of life, and the slow transformation (or evolution) of these forms through the immense abysses of geological time that opened up through the discovery of prehistoric sites and remains during the nineteenth century. Graphic phylogenies, such as those drawn by Haeckel (1874), expressed these evolutionary connections in the form of a tree.

Genealogies are central to the creation story in Genesis, which was a source of inspiration for late eighteenth and early nineteenth century philologists and ethnologists who sought to explain the relationship between the languages (and hence peoples) of the world in genealogical terms. Still earlier, culturally diverse objects brought from the new world to Europe from the sixteenth century onward, were accommodated by cabinets of curiosity within the category pagan, which already housed the classical world and barbarian Europe. Discussions about the origins of Incas and Aztecs, for example, revolved around their identity with lost tribes of Israel, and their being remotely descended from Noah (Shelton 1994).

Philological, biblical and Darwinian genealogical ideas influenced L. H. Morgan, author of *Systems of Consanguinity and Affinity of the Human Family* (1871), in his kinship and beaver research among the Objibwa around Marquette, which took place against a background of wilderness receding from American farms (Feeley-Harnik 1999). This was rather different from Darwin's invocation of genealogical imagery in his theory of evolution, which was filtered through the ideas of British animal breeders for whom keeping records was an important means of exercising control over procreation. The fact that the British aristocracy could trace their genealogies, just as they could trace the pedigrees of their favorite animals, such as thoroughbred horses, helped to distinguish them, in terms of descent and breeding, from middle- and working-class Britons.

Rivers borrowed and translated the English aristocratic notion of pedigree into the genealogical method. It became a means of organizing fieldwork and written ethnography, despite being later overshadowed by the long-term fieldwork practices of participant observation. The incorporation of pedigree into anthropological methodology for assimilating knowledge about others is a fascinating maneuver by a middle-class intellectual in the context of turn of the twentieth century Cambridge, which was still very much dominated by a few great families. Borrowing pedigree was simultaneously an assertion about being above this sort of hereditary class system, and yet harnessing the power of its connotations in the form of intellectual capital in the comparative study of mankind.

Genealogical thinking informed arrangements of material culture before the advent of public ethnographic museums, and continued to be influential after these collections had been systematized according to geographical provenance, or as evolutionary typologies. Museum collections played a significant role in the rise of disciplinary knowledge, including anthropology, during the nineteenth century. Rivers' early twentieth century reformulation of the genealogical method scaled down the vast abysses of nineteenth century ethnological time to the more modest tem-

poral requirements of the emerging ethnographic present. In this respect, he made a major contribution to modern ethnographic field techniques by developing an explicitly scientific method for collecting immaterial cultural data, and imbuing it with a distinctive material form.

3. *The Genealogical Method of Social Anthropological Inquiry*

Rivers' genealogical method moved from the concrete level of personal names to the abstract system of relationships. That abstract system was accessed through the terminological categories used by Ego to address and to refer to the persons named as relatives. Rivers explains the steps by which he collected a pedigree (he used the terms genealogy and pedigree interchangeably). First the descendants of the informant's 'real' (as opposed to classificatory) parents were listed, including previous or other marriages, and children listed in order of age, together with their marriages and offspring. The mother and father's respective pedigrees would be collected according to the same procedure. Rivers claimed that the most extensive pedigrees could be collected using only five terms of relationship: 'father', 'mother', 'child', 'husband', and 'wife.'

The fact that Rivers treated some categories—'brother', 'sister', 'cousin', 'uncle', and 'aunt'—as problematic, gave the entire method a lineal bias. Later commentators have argued that lineal relationships were seen as less ambiguous than either collateral or affinal ones (Barnard and Good 1984). Information concerning social condition, residence, totemic, and other group affiliation, along with other relevant facts, were to be recorded for each person on the pedigree as far as possible. An extraordinarily wide range of topics could be studied by collecting information in this way. Kinship was about the social recognition of certain genealogical relationships; and genealogy was assumed to be about biology.

Despite Malinowski's (1930) critique of the over-abstraction of kinship studies and his emphasis on the importance of long-term participant observation, the genealogical method proved to have remarkable staying power as a technique for structuring this kind of fieldwork.

4. *From Corporate Descent Groups to Critique*

One of the main uses of genealogies for Rivers and for subsequent (especially, but by no means exclusively, British) anthropologists lay in studying social organisation. This concern was sometimes connected with colonial administration, seeking to identify and use indigenous social structure for governmental purposes. Interest in descent, in studying ancestral connections reckoned through the female (matrilineal), or male (patrilineal), or both (bilateral) lines, gradually emerged as a dominant theoretical concern by the 1940s. The juridico-political importance of unilineal descent, especially in the African societies to which anthropologists turned their attention after an earlier interest in the Pacific, led Fortes to distinguish between domestic (internal) and political (external) social domains.

Fortes' (1945) 'Paradigm of the Lineage System' diagram is a particularly striking use of a genealogical-style diagram to represent the all-male lineage structure. Despite Leach's (1961) critique of the paradigm, and his insistence on the ideological rather than empirical nature of corporate lineage groups, he himself made extensive use of genealogical evidence to back up his argument. The continuing recourse to genealogy, despite the eclipse of kinship from center stage after the 1960s, helped to maintain a notion of biology as something separate (like nature) from society, until almost the end of the twentieth century. It is therefore remarkable that Barnes (1967) refers to the collection of genealogies as part of the ethnographer's stock-in-trade, and genealogical diagrams as part of the 'minimum obligation' for making fieldwork intelligible. He did, however, recommend recording the way informant's talked about their kin before embarking on formal questioning.

The ethnologists of France, who set out to study kinship at home in the same way that their colleagues did abroad, developed the technique of talking family, as a complement to genealogy, much more explicitly than their Anglo-Saxonic counterparts. Firth and his colleagues' 1960s study of middle-class kinship in north London collected and made use of genealogies only to discover that these were often part of oral knowledge—as in peasant societies.Schneider, too, constituted genealogies in his studies of American kinship (Schneider 1968), although he engaged in more open-ended questioning.

Bourdieu called for a social history of the genealogical tool, with particular attention to the functions that have produced and reproduced the need for such an instrument (Bourdieu 1977). Schneider (1984) was to criticize 'the doctrine of the genealogical unity of mankind', arguing that all comparative studies of kinship are rooted in our own beliefs about biology, and the enormous value attributed to human sexual reproduction in (what he variously called) Western/European/American cultures. Schneider's critique proved to be the catalyst for a number of new directions taken in the study of kinship. Strathern developed the argument that nature no longer serves to ground the understanding of society, which means that interventions such as the new reproductive technologies are an integral part of both English kinship and wider (political) society. Bouquet (1993) analyzed the cultural specificity of the genealogical method, arguing that it was infused by English notions of

pedigree so specific that it bears little resemblance to the Portuguese notions of relatedness, which were used to problematize English middle-class kinship. Although the genealogical diagram, with its more ramifying cultural-historical roots, simplified and facilitated the divulgation of genealogy in anthropology, it did not cancel the complexities and contradictions of genealogical ideas even within Europe (Bouquet 1996). Franklin (1997) has elucidated the effects of new reproductive technologies on the meaning of genealogy, understood as life's progression and as a way of making knowledge, among prospective parents in northern England. She coined the notion of post-modern genealogy to capture the way technology has come to stand in for and improve upon the natural process of procreation, despite all its attendant uncertainties.

5. *The Future of Genealogy*

Critical reappraisal of the assumptions implicit to genealogy in anthropology is likely to play a significant role in future studies of kinship. Genealogical thinking is certainly not universal, which renders problematic the anthropological attempt to use genealogy as a standard. Yet anthropologists have already been struck by the growth of popular genealogy in different parts of the world, and through the Internet (Segalen and Michelat 1991), both of which are subjects deserving further empirical study. Indigenous peoples' uses of historically collected genealogies in establishing land claims, constitute another highly significant development: for example, among Torres Strait and Australian Aboriginal communities. The impact of new technologies on genealogy, and new uses to which genealogies collected in the past by anthropologists may be put in the future, are but two examples suggesting that despite recent critical work, genealogy has by no means had its day in anthropology.

See also: Ancestors, Anthropology of; Clan; Kinship in Anthropology

Bibliography

Barnard A, Good A 1984 *Research Practices in the Study of Kinship*. Academic Press, London

Barnes J A 1967 Genealogies. In: Epstein A L (ed.) *The Craft of Social Anthropology*. Social Science Paperbacks in association with Tavistock Publications, London, pp. 101–27

Bouqet M 1993 *Reclaiming British Kinship. Portuguese Refractions of British Kinship Theory*. Manchester University Press, Manchester

Bouquet M 1996 Family trees and their affinities. The visual imperative of the genealogical diagram. *Journal of the Royal Anthropological Institute* (NS) **2**: 43–66

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK

Feeley-Harnik G 1999 'Communities of Blood': The natural history of kinship in nineteenth-century America. *Comparative Studies in Society and History* **41**: 215–62

Fortes M 1945 *The Dynamics of Clanship among the Tallensi*. Oxford University Press, London

Franklin S 1997 *Embodied Progress. A Cultural Account of Assisted Conception*. Routledge, London and New York

Haeckel E 1874 *Anthropogenie oder Entmicklingsgeschichte des Menschen*; *Stammesgeschichte des Menschen*. Wilhelm Engelman, Leipzig

Leach E R 1961 *Pul Eliya, A Village in Ceylon: A Study of Land Tenure and Kinship*. Cambridge University Press, Cambridge, UK

Malinowski B 1930 Kinship. *Man* **30**: 19–29

Morgan L H 1871 *System of Consanguinity and Affinity of the Human Family*. Smitnsonian Institution, Washington, DC

Rivers W H R 1910 The genealogical method of anthropological inquiry. *The Sociological Review* **3**: 1–12

Schneider D M 1968 *American Kinship, A Cultural Account*. University of Chicago Press, Chicago

Schneider D M 1984 *A Critique of the Study of Kinship*. University of Michigan Press, Ann Arbor, MI

Segalen M, Michelat C 1991 L'amour de la généalogie. In: Segalen M (ed.) *Jeux de Famille*. Presses de CNRS, Paris, pp. 193–208

Shelton A A 1994 Cabinets of transgression: renaissance collections and the incorporation of the New World. In: Elsner J, Cardinal R (eds.) *The Cultures of Collecting*. Reacktion Books, London, pp. 177–202

M. Bouquet

Generalizability Theory

1. *Overview*

Generalizability theory (G theory) provides a framework for conceptualizing, investigating, and designing reliable observations. It was originally introduced by Cronbach and colleagues (1963, 1972) in response to limitations of the popular true-score-model of classical reliability theory (Spearman 1904, 1910). Classical reliability theory centers around the notion that each observation or test score has a single true score, belongs to one family of parallel observations, and yields a single reliability coefficient (Nunnally and Bernstein 1994). While this model may be reasonable for carefully equated parallel forms of tests, it is overly restrictive and often unrealistic in situations where, for instance, raters differ in the central tendency and variance, observations depend on the context in which they occur, and constructs are patently heterogeneous. Several writers argued in the 1950s and 1960s that the same observation could conceivably belong to more than one set of parallel tests and, thus, have more than one reliability coefficient (Cronbach et al. 1963, Guttman 1953, Tryon 1957). The fact that the internal consistency reliability estimate of a multidimensional measure tends to be low even when test–retest and alternate forms reliability estimates are high illustrates

the contradictions and limitations of the classical reliability model.

It was Cronbach's vision to reinterpret classical reliability theory as a theory regarding the adequacy with which one can generalize from a sample of observations to the universe of observations that was randomly sampled. G theory acknowledges that the reliability of an observation depends on the universe about which the investigator wants to draw inferences. Because a particular measure may, conceivably, be generalized to many different universes, a measure may vary in how reliably it permits inferences about these universes and, therefore, be associated with different reliability coefficients. G theory explicitly requires investigators to specify a universe of conditions over which they wish to generalize. The extent to which different conditions are associated with different observations has implications for designing dependable observations.

G theory not only reinterprets classical reliability theory, but also shows how the traditional distinction between reliability and validity can be overcome to design dependable observations. Traditionally, construct validity is concerned with drawing inferences about a latent construct based on observable measures, whereas reliability is concerned with drawing inferences about a true score from observations across parallel measures. In G theory, a universe, its facets, and the conditions for admissible observations are defined through careful construct explication, the traditional domain of validity theory. G theory defines observations as dependable if they permit accurate inferences about the latent construct (i.e., the universe of admissible observations) they are meant to represent. In so doing, G theory blurs the traditional distinction between reliability and validity. The use of the terms 'dependability' and 'generalizability' instead of 'reliability' reflects the interest in unifying reliability and validity. How to go about designing and investigating dependable measures is the subject of G theory.

2. *Key Concepts*

2.1 *The Decision Maker*

Measurements are designed with applications in mind, and their quality has to be examined in the context of their application. This context is symbolized by the 'decision maker' who is interested in measuring a particular construct in a particular population of persons, under a particular universe of conditions, and with particular types of decisions in mind.

2.2 *Types of Decisions*

The types of decisions for which a measurement instrument is to be used must be known before an instrument that yields dependable observations can be designed. Two major types of decisions have to be distinguished. The first concerns decisions that rely on the relative ranking of individuals. Accepting the three top-scoring applicants for a position is an example of a decision that relies on the relative interpretation of scores. Measurements that emphasize the interpretation of the relative differences between scores are the domain of classical test theory with its traditional focus on interindividual differences.

The second major type of decision is based on the interpretation of the absolute level of scores. Mastery tests in education provide an example of a decision based on the absolute interpretation of a score. Similarly, pre-established admissions standards for college (e.g., minimum SAT score) or pass/fail standards on job selection tests rely on the interpretation of absolute score levels.

2.3 *The Universe of Admissible Observations*

A measurement consists of a sample from the universe of admissible observations. What defines a universe for a particular measurement is based on what the decision maker is willing to treat as interchangeable for the purposes of making a decision. Thus, a universe may consist of interchangeable three-letter syllables, symptoms of depression, or interviewers if the decision maker is willing to treat as interchangeable three-letter syllables, depressive symptoms, and interviewers. Or, a universe may consist of interchangeable time points for urine collection, survey questions about voting behavior, or situations for observing altruistic behaviors if the decision maker considers those aspects of measurement to be interchangeable.

2.4 *Estimating a Person's Universe Score*

Given a particular universe of admissible observations, a person's universe score ($\hat{\mu}_p$) can be defined as the average score based on all admissible observations (X_{pi}) of the universe of interest. The purpose of a measurement is to accurately estimate this universe score ($\hat{\mu}_p$) based on a sample of observations.

The degree to which a measurement is generalizable depends on how accurately the measurement permits us to estimate the universe score. The accuracy of an observation is captured in the variance associated with the different presumably interchangeable observations. The extent to which different items or different observers or different measurement occasions yield different observations determines how dependable a single specific observation is. In other words, if different items, observers, etc., yield similar observations, a single measurement may allow accurate inferences about the universe of observations. If different items, observers, occasions, etc., yield dis-

similar observations, inferences about the underlying universe based on a single observation may be questionable.

2.5 The Universe and its Facets

Universes can be simple or complex, homogenous or heterogeneous, and small or large depending on the construct of interest and the decision maker's interest in investigating different aspects or facets of generalizability. The terms 'facets' and 'conditions' are analogous to 'factors' and 'levels' in the literature on experimental design. A universe of observations is said to consist of one facet if the generalizability of observations regarding one source of variation in the universe, say arithmetic questions of varying difficulty, is at stake. For instance, a particular arithmetic test includes a sample of items, covering different addition, subtraction, multiplication, and division problems of one- and two-digit numbers. The decision maker is interested in general arithmetic achievement, and is indifferent to the particular questions on the test. In the one-facet design, one is interested in estimating the universe score of each person based on the sample of items included in the test. A universe is said to have two (or more) facets if the generalizability of observations regarding two (or more) sources of variation in the universe, say items and occasions, is at stake. For instance, to the extent that arithmetic achievement depends on the difficulty of the items and the test occasions (e.g., beginning or end of school year), the generalizability of a particular measure taken early in the school year on a sample of difficult items may be compromised.

2.6 Generalizability Studies

Generalizability studies (G studies) are conducted to investigate the relationship between an observed score and a universe score. They are designed to provide information regarding the sources of variability (i.e., facets of the universe) that influence the generalizability of observations. Because the dependability of a measure rests on its purpose (and thus the universe of admissible observations it is meant to represent), different G studies may be needed depending on the proposed application of the measure. Alternatively, a single G study has to anticipate the multiple uses of a measurement to provide as much information as possible about potentially important sources of variation.

In the above example, arithmetic items varying in difficulty is an obvious choice of a facet to be examined in a G study. In addition, decision makers may be interested in whether measures taken at different occasions, based on, for example, different response formats or different presentation modes yield different observations. To examine these issues, a universe has to be defined more broadly to include items of different difficulty, response formats, presentation modes, and taken at different occasions. Having defined such a multifaceted universe, a G study has to be designed to investigate the contributions of these facets and their interactions to the overall variance in observations.

2.7 Decision Studies

G studies are designed to assess the dependability of a particular measurement technique. In contrast, decision studies (D studies) are designed to gather data based on which decisions about individuals are made. D studies rely on the evidence generated by G studies to design dependable measures for a particular decision and a particular set of facets about which a decision maker would like to generalize. The goal is to design a measure that samples sufficient numbers of instances from different facets of a universe of observations to yield a sufficiently dependable estimate of the universe score which the measure is meant to represent.

3. Illustrative Generalizability Design

3.1 Generalizability Design

For illustration, the statistical models and analyses for a one-facet G study will be presented. This most simple G design requires making observations (X_{pi}) on individual subjects (1, 2, …, p, …, n_p) under different conditions of a facet (1, 2, …, i, …, n_i), where subjects and conditions represent random samples from their respective universes. In a one-facet design, the universe of admissible observations consists of observations that differ with respect to one characteristic, such as item difficulty. The generalizability question is: How dependable are observations made under different conditions to draw inferences about the universe consisting of all conditions?

3.2 The Statistical Model

A one-factorial ANOVA model can be used to describe how the observed score of person p under condition i can be partitioned into a grand mean (μ), an effect for each person (π_p), an effect for each item (α_i), and a residual (ε_{pi}).

$$X_{pi} = \mu + \pi_p + \alpha_i + \varepsilon_{pi} \tag{1}$$

In Eqn. (1), the grand mean (μ) is a constant value for all persons, positioning the score for all persons on the particular scale of measurement. The person effect (π_p) describes the difference between a person's universe score and the grand mean ($\mu_p - \mu$), indicating inter-

individual differences in universe scores. The item effect (α_i) indicates the difference between an item's difficulty and the average item difficulty ($\mu_i - \mu$), reflecting the variability associated with the facet of error being investigated. The residual combines the influence of the person-by-item interaction, systematic sources of error other than items, and random events.

According to this model, specific observations X_{pi} vary due to differences between persons (i.e., person effects), differences between items (i.e., item effect), and other unaccounted for sources (i.e., residual). Accordingly (see Eqn. (2)), the variance of a set of observed scores (σ^2_{Xpi}) can be partitioned into components due to persons (σ^2_p), items (σ^2_i), and the residual (σ^2_ε). The residual combines two sources of variance that cannot be distinguished from each other in this model. They are the person-by-item interaction (i.e., items vary in difficulty between persons; persons vary in their ranking on different items) and chance fluctuations in scores.

$$\sigma^2_{Xpi} = \sigma^2_p + \sigma^2_i + \sigma^2_{pi,\varepsilon} \tag{2}$$

3.3 G study: Estimation of Variance Components

To estimate these variance components, a G study has to be conducted in which data are collected on a representative sample of persons and items. Variance estimates can be derived from the mean squares (MS) reported as part of the standard ANOVA printout of statistical analysis programs (see Eqns. (3)–(5)), in which persons are identified as between-subjects factor and items as within-subjects factor.

$$\hat{\sigma}^2_{pi,\varepsilon} = \mathrm{MS}_{\mathrm{residual}} \tag{3}$$

$$\hat{\sigma}^2_p = (\mathrm{MS}_p - \mathrm{MS}_{\mathrm{residual}})/n_i \tag{4}$$

$$\hat{\sigma}^2_i = (\mathrm{MS}_i - \mathrm{MS}_{\mathrm{residual}})/n_p \tag{5}$$

3.4 A Numerical Example

Table 1 presents the ANOVA summary table from a G study in which 1,140 persons rated the frequency with which they experienced 20 depressive symptoms. The table presents the numerical estimates and the percentage of total variance accounted for by persons, items, and the residual.

3.5 Interpretation of Variance Components

Variance components reveal information about how different sources of variability affect the response to a single item. For instance, the variance component associated with items in the above example is 0.0230, accounting for 6 percent of the total variance. This variance component indicates how much the mean of any one randomly selected item is expected to vary from the mean of all items in the universe. Clearly, the magnitude of a variance component depends on how the items were scored (e.g., 1, 2, 3, 4 or 10, 20, 30, 40). In this example, items were scored 1, 2, 3, 4, where 1 indicates the absence of depressive symptoms and 4 indicates the symptoms were present all of the time. The average item mean (i.e., estimate of the mean of all items in the universe, $\hat{\mu}_p$) was 1.38 indicating that most respondents reported the absence of depressive symptoms. Taking the square-root of the variance component $\sqrt{0.023}$ yields 0.1517, indicating that we would expect approximately 95 percent of item means to fall between ± 0.2972 of the overall mean (i.e., CI95% = [1.08; 1.68]). Given that the possible range of item scores is 1 to 4, the expected range indicates that the item means in the universe of admissible items will mostly vary between 1 and 2. The variance component for persons is 0.0596, or approximately 15 percent of the total variance. This variance component estimates the universe score variance, that is, the between-person variability in the average item scores across all the admissible items of the universe. Based on the variance component, the standard deviation is 0.2441, indicating that 95 percent of universe scores are expected to fall between ± 0.4885 of the grand mean (1.0, 1.86).

Similar to the reliability coefficient in classical measurement theory, G theory yields a coefficient of generalizability (ρ^2) based on a particular G design. It is defined as the ratio of universe- and observed-score variances. In this example, the G coefficient for a single item is $\hat{\rho}^2 = \hat{\sigma}^2_p / \hat{\sigma}^2_{Xpi} = 0.1522$. Similarly, intraclass correlation coefficients can be determined. For one item it is $\alpha_{(1)} = (\mathrm{MS}_p - \mathrm{MS}_{res})/(\mathrm{MS}_p + (n_i - 1)\mathrm{MS}_{res} = 0.1617$, and for the set of 20 items it is $\alpha_{(20)} = (\mathrm{MS}_{p-} - \mathrm{MS}_{res})/\mathrm{MS}_p = 0.7941$. The latter is also known as the Cronbach α coefficient or KR20 formula.

By far the largest variance component is due to the residual (i.e., 0.3089), accounting for approximately 79 percent of the total variance. This suggests that there are important sources of variance not accounted

Table 1
ANOVA mean squares (MS) and estimates of variance components for a one-facet G study

Source of variation	MS	Numerical estimate	Percentage of total variance
Persons (p)	1.5005	$\hat{\sigma}^2_p = 0.0596$	15
Items (i)	26.4947	$\hat{\sigma}^2_i = 0.0230$	6
Residual (pi, ε)	0.3089	$\hat{\sigma}^2_{pi,\varepsilon} = 0.3089$	79
Total	0.3902	$\hat{\sigma}^2_{Xpi} = 0.3915$	100

for by differences between persons or differences in item difficulties (i.e., item means). Because the residual consists of the item-by-person interaction, this variance components reflects the fact that subjects indicate somewhat different levels of depressive symptoms depending on which item they are responding to. Based on the variance component, the standard deviation is 1.0893, indicating that the response to a specific item is expected to fall within ± 1.09 of a person's universe score in 95 percent of observations.

In summary, the results show that responses to a single item are mostly influenced by measurement error. Information about inter-individual differences make up only 15 percent of the total variance, and information about differences in test items make up only 6 percent of the total variance in item responses. Clearly, a single test item is of little use in estimating a person's universe score. Consequently, decision studies must sample many items to achieve more dependable estimates. Exactly how many items is the subject of D studies.

3.6 D study: Designing Generalizable Observations for Decisions

To conduct a D study, the relevant universe of observations and its facets must have been investigated in the G study. The type of decisions to be made must be known and the accuracy necessary to make the decisions specified. With this information at hand, the number of items from each facet of the universe can be determined such that dependable inferences can be drawn about the entire universe of observations at stake.

For instance, the decision maker in the previous example may want to design a measure that yields a 95 percent margin of error of ± 0.25 for the estimate of a person's universe score. That is, the decision maker wants to be 95 percent confident that a person's universe score is within ± 0.25 of the observed score. Based on this 95 percent margin of error (i.e., $0.25 = 1.96 * \sigma_{\text{error}}$), the error variance of the universe score can be determined: $\sigma^2_{\text{error}} = (0.25/1.96)^2 = 0.0163$.

For decisions based on the relative ranking of scores, the required number of items is:

$$n_{\text{i}}' = \frac{\sigma^2_{\text{pi,e}}}{\sigma^2_{\text{error|relative}}} \tag{6}$$

For decision based on the absolute ranking of scores, the required number of items is:

$$n_{\text{i}}' = \frac{\sigma^2_{\text{i}} + \sigma^2_{\text{pi,e}}}{\sigma^2_{\text{error|absolute}}} \tag{7}$$

Thus, for relative decisions, we need to sample $n_{\text{i}}' = 0.3089/0.0163 = 18.95$ or 19 items. And, for absolute decisions, we need to sample $n_{\text{i}}' = (0.0230 + 0.3089)/0.0163 = 20.36$ or 21 items. In this example, very little variance is contributed by differences between items in their difficulties (i.e., σ^2_{i}), indicating that, regardless of which items are sampled, similar absolute scores will be obtained. Consequently, only two additional items have to be sampled to yield equally dependable estimates of universe scores for relative and absolute decisions.

4. Extensions

The principles presented here for a simple generalizability scenario can be extended to more complex designs consisting of two or more factors, which may be random or fixed, and which may be crossed or nested. The interested reader is referred to the works by Cronbach et al. (1972) and Shavelson and Webb (1991). G theory can also be applied to the analysis of score profiles, composites, and difference scores. Furthermore, extensions have been developed to estimate variance components via maximum likelihood, Bayesian, and covariance structure methods, and for studying the dependability of any facet of observation. The interested reader is referred to the work by Marcoulides (1996, 1999), Shavelson et al. (1989), Wittmann (1988), and Webb et al. (1983).

See also: Classical (Psychometric) Test Theory; Content Analysis; Content Validity; Generalization: Conceptions in the Social Sciences; Indicator: Methodology; Measurement, Representational Theory of; Reliability: Measurement; Test Bias

Bibliography

Cronbach L J, Gleser G C, Nanda H, Rajaratnam N 1972 *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York

Cronbach L J, Nageswari R, Gleser G C 1963 Theory of generalizability: A liberation of reliability theory. *British Journal of Statistical Psychology* **16**: 137–63

Guttman L 1953 A special review of Harold Gulliksen's theory of mental tests. *Psychometrika* **18**: 123–30

Marcoulides G A 1996 Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling* **3**: 290–9

Marcoulides G A 1999 Generalizability theory: Picking up where the Rasch IRT model leaves off. In: Embretson S E, Hershberger S L (eds.) *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Lawrence Erlbaum, Mahwah, NJ, pp. 129–52

Nunnally J C, Bernstein I H 1994 *Psychometric Theory*, 3rd edn. McGraw Hill, New York

Shavelson R J, Webb N M 1981 Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology* **34**: 133–66

Shavelson R J, Webb N M 1991 *Generalizability Theory: A Primer*. Sage, Newbury Park, CA
Shavelson R J, Webb N M, Rowley G L 1989 Generalizability theory. *American Psychologist* **44**: 922–32
Spearman C 1904 The proof and measurement of association between two things. *American Journal of Psychology* **15**: 72–101
Spearman C 1910 Correlation calculated from faulty data. *British Journal of Psychology* **3**: 271–95
Tryon R C 1957 Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin* **54**: 229–49
Webb N M, Shavelson R J, Maddahian E 1983 Multivariate generalizability theory. *New directions in Testing and Measurement* **18**: 67–82
Wittmann W W 1988 Multivariate reliability theory: Principles of symmetry and successful validation strategies. In: Nesselroade J R, Cattell R B (eds.) *Handbook of Multivariate Experimental Psychology*, 2nd edn. Plenum Press, New York, pp. 505–60

G. E. Matt

Generalization: Conceptions in the Social Sciences

All social scientists use observed particulars to draw conclusions about abstract entities. In measurement theory, the argument that 'IQ is what an IQ test measures' is discredited, and informed debates take place to interpret what the observed scores mean in general language. Some qualitative researchers are also reluctant to generalize beyond observables, but their language is replete with summarizing statements and they invite readers to construct theoretical meanings beyond the descriptions offered. In laboratory research, scientists often study local student samples, not to construct theories about them or about students in general, but to make inferences about humans. Some laboratory researchers use animals to substitute for humans, and their case for cross-species generalization has to depend on arguments—e.g., that cats and humans share the anatomical and physiological attributes central for understanding whatever aspect of vision is under study. This entry describes and evaluates many of the procedures for generalization that social scientists actually use.

Cronbach (1982) has perhaps the most comprehensive framework for understanding generalization. We use his theory, though slightly modifying it. He claims that social scientists typically draw conclusions about four not completely independent entities—human populations, physical settings, causes, and observables (including effects and moderators of cause–effect relationships). Time, understood as historical period, might also be added. Cronbach further contends that scientists work with samples, instances, cases, exemplars, or particulars (terms he uses almost interchangeably), and that these represent study-specific populations, categories, entities, universes, classes, constructs, or categories (terms he also uses interchangeably). A major task of all scientists is to identify the class a given set of lower-order instances represents.

Cronbach also understands generalization in terms of inferences about classes that the sampling specifics manifestly do not represent. Such knowledge transfer is particularly salient, for instance, when animals are studied as proxies for humans or when laboratories substitute for other social settings. But transfer is also involved when research on American senior citizens, say, is used to extrapolate to Americans in mid-life or to seniors in Western Europe; or when findings about factories are extrapolated to schools on grounds that both are complex organizations. Extrapolation also pertains to causes and effects, as when studies of healthcare subsidies are applied to housing subsidies, or when a finding about a subsidy's effect on hospital utilization is applied to effects on mental health. Cronbach believes that methods are needed both for justifying inferences from particulars to generals (as from samples to populations or from measures to constructs) and for justifying extrapolations beyond the categories studied.

His schema also includes subcategories—e.g., the racial groups within some human population or the components constituting a construct-like stress. Researchers routinely partition data both to test a relationship's stability across the subcategories examined and to identify the conditions under which it varies. The more heterogeneous the subpopulations are over which a relationship holds, the greater its generality is usually assumed to be. In additon, the more specifically identified are the boundary conditionals limiting a relationship, the better its generality is understood. Campbell's external validity (Campbell 1957) also emphasizes generalization across entities, but without restricting it to subtypes. For him, a relationship found among Austrians should be examined not just by partitioning Austrians into regional or generational subgroups, but also by examining the same relationship across other nations. Findings that dependably replicate across very different circumstances are the foundation of all general science.

Using the Cronbach framework, we structure this exposition into four parts: (a) methods based on formal sampling theory, as they are and are not used for generalizing to clearly designated universes of people, settings, times, causes, and effects; (b) purposive sampling methods, especially as they are used in validating constructs; (c) methods for assessing the range of circumstances over which a causal connection holds; and (d) methods for extrapolating knowledge to entities with different characteristics from those on which the knowledge was originally established.

This phrasing acknowledges that many social scientists use universe (and population) differently from construct, though the later Cronbach tends to use them almost interchangeably. In statistics, populations are ostensive; their elements are real and can be pointed to. But constructs are hypothetical and more obviously theory dependent. Moreover, the formal methods statisticians prefer when sampling elements from a population cannot be used with constructs, because the necessary enumeration and sampling of elements cannot be readily achieved with measures of abstract constructs. This is why Cook and Campbell (1979) used external validity to refer to people and settings and construct validity to refer to instances of more hypothetical causes and effects. However, the distinction is partly arbitrary. Constructs have constitutive elements theoretically specified as their components, and instances of any one construct vary in which components they incorporate. Moreover, human populations are not totally ostensive *per se*. Despite official definitions, there is still room to disagree about what being an Australian means: what about someone with an Australian passport who has always lived abroad, or the illegal immigrant who has always lived in Australia without a passport? There is room for legitimate debate about the definition of any human class and about whether certain individuals fit within that class. Also, whether human populations or abstract constructs are involved, the functional task is always the same: justifying inferences from particular cases to general categories. Hence, we see no fundamental reason for treating populations and constructs differently, though we do respect social science terminology and often use the one to refer to people and settings and the other to cause and effect variables as well as causal moderators.

1. Formal Sampling Theory

The best way to represent a population is to enumerate its members before selecting a random sample from that population. When properly implemented, this guarantees that the sample will formally represent the population within known limits of sampling error. The sample and population distributions will match on all measured and unmeasured attributes. This makes random sampling the centerpiece of survey research, often complemented by other sampling details. To reduce costs, cases can be clustered within a geographic area before being randomly selected. To ensure that subpopulations are reliably represented, stratification can also occur before random selection. Formal sampling theory provides an impeccable justification for generalization to well-designated populations. And methods for implementing are now so improved that its use in describing human populations and some settings is almost routine.

The problem with random selection is its own limited generality. It rarely applies to causes and other abstract constructs (or to time). One cannot readily imagine enumerating all instances of a specific causal agent before selecting a random sample of these instances for study. In practice, causal agents are purposively chosen because of a presumed correspondence between substantive theory and operational instance(s). Likewise, it is rare for all possible measures of a construct to be enumerated before some are randomly chosen. The selection of measures to represent a construct is again purposive. Also, it is hardly practical to sample historical periods at random for generalizing, if not to eternity, then at least to some meaningful interval. Detailed measures are not always available from the past; and anyway, many theories predict temporal contingency rather than constancy. This last justifies the purposive selection of time periods for the unique light each throws on some causal process.

Even when selection is at random, unambiguous inference still depends on large samples that are not always readily forthcoming. How often can we randomly sample 50 or 100 time points for assessment? Or 50 or 100 institutions, unless the aim is to describe a particular class of institutions? What about manipulating causal agents within institutions? What proportion of those asked to participate will agree to do so? What types are these volunteer organizations likely to be—the most typical or the most proud?

Random sampling rarely applies to assessing the conditions under which cause and effect are related. We know of only six experiments where individuals or households were randomly selected before being randomly assigned to a treatment (see Shadish et al. 2001). However, survey researchers often use random sampling to test causal hypotheses, using item responses to identify variation in a possible causal agent and to provide outcome data. But this is widely acknowledged to generate inferior causal conclusions compared to experimenter-controlled intervention. So how can causal relationships be better generalized via the random selection that surveys afford if these relationships are more uncertain because a survey was used? Moreover, surveys do not routinely use random selection for other generalization-relevant features of the research. For instance, data are typically collected at one or a few time points only, and then in one type of setting (usually the home),

Random selection was not developed for extrapolation, generalizing beyond the sampling particulars. So, it is not surprising that it is not relevant to this function. Indeed, none of the extrapolation methods we later detail depends on random sampling. Other premises are involved, as they are for generalizing causal relationships and for selecting measures to represent constructs. Thanks to its logical and empirical warrant, random selection is *the* method of choice for generalizing to populations of persons

and sometimes settings. But it is rarely relevant to other tasks with which a comprehensive theory of generalization must struggle.

2. Purposive Selection and Construct Validity

Purposive samples are selected without randomization and require only that the selected cases have certain prespecified characteristics. If these characteristics coincide with those of a population, then a match is achieved between the details in the population description and the details achieved in the sampling particulars. Unfortunately, such proximal matching cannot guarantee that the cases and categories are matched on all unmeasured characteristics, let alone perfectly matched on the measured ones. Purposive selection is less well theoretically justified than random selection. By itself, it justifies claims about category membership but not category representativeness.

Yet purposive selection is used throughout the social sciences. Even passionate advocates of random sampling measure behavior by first describing the construct to be assessed and then ensuring that measures of it include all attributes in the theoretical description. Empirical patternmatching between constructs and measures is the basis for generalization in this case, and substantive theory provides the template for choosing how to measure (or manipulate). The correspondence between construct and operational representation is usually multidimensional, with the most important attributes being the prototypical ones that most clearly define the construct's uniqueness (Brunswik 1956).

Such pattern matching (also called proximal similarity) is not restricted to measurement contexts. When random selection is infeasible it also applies to selecting persons, settings, and times. Thus, experimenters might partition their data into gender subgroups. The resulting subsamples clearly correspond to the categories 'male' and 'female,' as inferred from body shape or valid self-reports. Selecting this way ensures a proximal similarity between the class 'female' and the sampled instances of that class. But do they represent their gender if all attend the University of Berlin? In such a case, the gender specifics and the construct match on some prototypical attributes; but the sample is much more internally homogeneous than the population with respect to age, intelligence, geography, and culture, among other factors. Belonging in a class does not entail formally representing that class. Proximal similarity is necessary but not sufficient for inferring the representativeness of sampled instances. Sophisticated construct validation requires more.

One auxiliary principle concerns 'the heterogeneity of irrelevancies' (Campbell 1988), ensuring that the sample is not homogeneous on attributes that are irrelevant to the inference being made—such as age, intelligence, and nation in the case of females students in Berlin. To make these conceptual irrelevancies heterogeneous requires adding samples of women of different ages, social classes, and countries. In measurement contexts, heterogeneity requires multiple measures of each construct or varying multiple instances of a treatment, provided that each instance has unique irrelevancies and that no irrelevancy is shared by all instances. It also requires testing the degree of triangulation achieved across the sources of presumed irrelevancy.

Another way to deal with homogeneous irrelevancies requires focused comparison. If female students in Berlin are compared with their male counterparts, this holds constant irrelevancies like country, language, and city of testing, while also varying gender. But such internal comparisons are not always feasible; and groups may differ in other unmeasured irrelevancies that affect responding—e.g., if female students in Berlin have higher intelligence than males on the average, or if data collection there occurs in chemistry classes that attract different kinds of females than males. Whatever the construct being targeted—gender, intelligence, or being Australian—purposive sampling requires critically examining whether the instances selected are more homogeneous than their presumed population.

Construct validation also requires 'discriminant validity' (Campbell and Fiske 1959), demonstrating that a relationship obtained with constructs labeled one way is not due to another construct that is correlated but nonetheless conceptually unique constructs. Thus, the validity of an anxiety measure requires not only showing that its measured attributes behave as hypothesized but also that they, as a whole, behave differently from conceptually related concepts like depression. Likewise, testing a hypothesis about how new computer firms are unique requires sampling more than just this type of firm. Also needed are firms making new products in other sectors. Relationships that hold across sectors can be generalized to firms in general, including computer firms. However, only relationships limited to computer firms help define what is unique about computer firms.

All sophisticated construct validity is about testing unique predictions entailed by the substantive theory of a population. We always validate interpretations of categories rather than categories themselves, even with human populations (Cronbach 1989). Lawyers might pronounce that an Australian is anyone eligible for an Australian passport. But this definition is not adamantine. It includes individuals with Australian parents who have never lived in the country; it excludes illegal immigrants who have lived there for decades; and what about persons eligible for a passport who deliberately choose not to claim one despite living in the country for decades? No specification of a construct or population is beyond reasonable criticism and so each preferred interpretation must be defended.

Usually, the interpretation can be formulated as hypotheses about some combination of components, antecedents, consequences, or correlates. Construct validation requires specifying and testing the nearest to unique of these hypotheses. Only then can we be truly confident about generalizing from purposive samples and measures to their target constructs rather than to some other entity correlated with them. However, constructing the arguments needed is more laborious and uncertainty-riddled than random sampling. Still, most social scientists sample purposively most of the time, though they vary greatly in how much auxiliary information they collect after a modicum of proximal similarity has been achieved.

3. *Generalizing Causal Connections*

Much effort goes into testing causal hypotheses—statements that variation in one construct leads to variation in another and that no alternative cause can account for their covariation. Generalizing causal statements depends on methods other than formal sampling theory, given that experiments with both random selection and random assignment are rare and surveys provide causal tests that are flawed when compared to experiments. So, how is generalization achieved when the object to be generalized is a causal claim?

One way is when a population has recently been described using measures that are also available on a purposive sample of persons or sites. Then, any causal relationships obtained at the sample level can be weighted to the population level, using the empirical relationship between sample and population for this purpose. But such weighting only approximates what the population effect size would have been if the sample had been randomly selected. Correct inference depends on the sample level effect sizes not being affected by unmeasured sample characteristics that are not fully captured by the shared, measured variables. Still, we can anticipate that weighting procedures will be used more often in the future as methods for social experiments and estimating missing data advance. After all, survey researchers now describe human populations from biased sample data using basically the same weighting technique.

Sometimes, substantive theory is specific enough to detail the conditions under which a causal relationship holds. Then, a program of research can test explicit hypotheses about specific sources of population, setting, or time variation in the strength or direction of the causal relationship. This involves deliberate attempts to falsify the hypothesis that a causal connection is general (Gadenne 1976). Rejecting the hypothesis entails identifying specific causal moderators; failure to reject implies that the hypothesis of a general cause–effect relationship remains viable. A less formal variant of this describes common practice in the laboratory sciences. A causal claim is made and then independently replicated, often under conditions that approximate the original ones. Successful replication leads to concluding that the causal relationship is provisionally true—that is, true until future failures to replicate indicate the need to invoke causal moderator variables. In both approaches, special emphasis is on identifying moderators that reduce the causal relationship to zero or that change its sign, as opposed to those that merely vary its size without affecting its sign.

Purposive sampling is also used in other less formal generalization probes. In clinical trials, hospitals recruit patients with a given disease and then adhere to the study's intervention and data collection protocols. Hospitals are basically chosen for their willingness to comply and for their heterogeneity with respect to, say, region of the country, predominant ethnic group served, or type of hospital sponsorship. Statistical power is often only adequate for cross-site analysis, suggesting that the goal is not to identify site-level main and interaction effects, but to see if the causal estimate exceeds zero despite heterogeneity in the hospitals studied. If it does, the causal effect is held to be general. If it doesn't, several alternative claims are viable and need to be tested, especially the following: there is no effect; the causal sign varies by hospital type and obscures a main effect; or the methods used cannot detect an effect of the size expected. Only increasing the sample of sites allows direct tests of which hospital types are involved in site-level interactions. Even so, relying on a purposive sample of sites that are heterogeneous on many attributes except for volunteering is only logically defensible if researchers know from other sources that no confounds are associated with all the hospitals being volunteers for the clinical trial.

Exploratory analyses often involve measuring many constructs from among a large array of possible causal modifiers in order to probe which ones affect the strength of a causal relationship. Without *a priori* hypotheses about modifiers, the approach can be time-consuming and run afoul of error rate problems due to the many tests conducted and the small sample sizes for some values on the moderator. Nonetheless, subcategory breakdowns are often used within experiments to probe how robust a relationship is across persons, settings, times, and measures. This can prevent the overgeneralization that occurs when researchers conclude that a relationship found with the entire sample holds in each subgroup within the sample. The entire sample allows one to conclude that the effect is found despite the internal heterogeneity achieved. It does not imply there is independent replication within any of the individual subgroups. The story is essentially the same when partitioning data from a single survey. With adequate sample sizes per group, one can probe the robustness of a relationship across the subgroups sampled. But only those potential conditional variables directly sampled can be assessed.

The best tests of causal conditionals come from synthesizing multiple studies on a topic rather than from subgroup breakdowns within a single study (Cooper and Hedges 1994). Experiments and surveys relevant to the same causal hypothesis accumulate and can be used in meta-analysis, the best-known form of synthesis. Meta-analysis combines data across multiple studies asking the same question, taking advantage of the between-study variation in measures, manipulations, populations of people, settings, and times. The dependent variable is each study's standardized effect size(s) and these are manipulated (a) to generate an average effect size across all the studies and all the irrelevant sources of heterogeneity they contain, and (b) to identify the specific contingency factors responsible for variation in the effect sizes.

What logic buttresses meta-analysis? It is not random sampling since individual studies are selected for their relevance to the hypothesis and the hope is to achieve a census of all published and unpublished studies. However, the more relevant target population is all studies that could be conducted on the topic, not all those done on it. In meta-analytic practice, generalization depends on the sample of achieved studies including a wide range of human populations, setting types, historical periods, and operationalizations of the cause and effect. Each study is first coded to label its sampling particulars in terms of theoretical constructs and to examine the sources of irrelevancy with which the particulars are and are not collectively imbued. Thus in a hospital experiment, hospitals might be described as being university-based, public, or private, and as being large or small. Patients might be described in terms of diagnosis, age, and gender mix, etc. The treatment might be described in terms of its components, duration, and manner of delivery; and the outcome might be described in similar terms. Then researchers typically test whether the same causal relationship holds despite these sources of sampled irrelevancy.

Sample sizes permitting, researchers also test whether the relationship holds across individually identified sources of potential irrelevancy so as to generalize to hospitals in general, people diagnosed with cancer, or whatever the target inferences might be for a particular study. Important here are the number and especially the heterogeneity of the types examined. Of course, some sources of variation might have *a priori* substantive meaning, as when university hospitals are a special target. Then one has to do a discriminant validity test and show that the effect size for such hospitals is different from that for other hospitals. Otherwise, there is nothing unique about university hospitals. So, meta-analysis depends on the very tests for proximal similarity, heterogeneous irrelevancy, and discriminant validity that are so central in construct validation. In addition, to achieve such tests multiple replication is needed across studies that may each be individually very narrow in the range of people, settings, times, and cause and effect operations it includes. Central is the heterogeneity across studies and the pattern of results achieved despite this sampled heterogeneity.

4. *Extrapolating to Nonsampled Entities*

Some social scientists probably think it silly to use samples for generalizing to populations that are not even proximally similar to the samples used. So, many animal researchers abjure all talk of generalizing to humans. But their language is often anthropomorphic, and their research is usually funded because others think it is relevant to humans. So, elaborate reasoning has evolved over time to justify generalizing from animals to humans.

In this regard, consider the use of cats in visual research (Blake 1988). The rationale for their use is in two parts. The first is entirely pragmatic and stresses the conveniences of working with cats (or *Drosophila* in genetics research), usually because of cost and ethics. The second is theoretical and is based both on cats sharing some anatomical and physiological features with humans and on behavioral research showing that cats and humans behave similarly when confronted with similar visual stimuli. However, the empirical work necessary for showing such correspondences has also revealed that cats are much less like humans in some physical ways—especially features related to night vision—and that some behaviors are not replicated across species when the same stimulus is presented. So, a contingency theory has evolved of the conditions under which cats are better or worse substitutes for humans.

But when asked, honest vision researchers will avow that convenience is the paramount rationale for using cats. Would they rather work with chimpanzees? Undoubtedly, since the chimps' visual system is much closer to humans' than the cats'. But the expense of maintaining chimps precludes using them. Would the researchers rather use humans than chimps? Undoubtedly, if ethical research procedures are possible—which they often are not. However, now that new imaging technologies are available that allow observing the human eye reacting to planned stimuli in great micro-detail and real time, many vision researchers are turning to work with humans rather than cats because the pragmatic obstacles to using humans now hardly apply. Thus within the community of vision researchers the preference for generalization to humans via within-species proximal similarity is clear. So, too, is the distaste for arguments that depend on the (only partial) proximal similarity between cat and human on some vision-related attributes given their obvious dissimilarity in others. A similar story could be told for conducting research in the laboratory rather than in the field. Convenience is paramount, and the reasons given for preferring the lab often

contain elements of rationalization. But even so, extrapolation across species and settings is surely superior when there is an empirically defensible contingency theory of the conditions under which the extrapolation is better and worse warranted.

How is extrapolation otherwise achieved? Causal explanation is often held to play a major role, explaining how or why two or more things are related (Mackie 1974). Once an explanatory process has been identified, the argument is that it can be reproduced in other settings, with other kinds of people, at future times, and with unique ways of instantiating the process. Indeed, this is the major public rationale for basic science. Once one knows why fire happens, one can replicate it with new kinds of materials (say, some newly invented polymers), in new settings (an uninhabited desert island), and with people who have never seen fire before (so long as they can understand the explanatory recipe). The problem with this view is that full explanation is hard to conceptualize and achieve. It is difficult enough to use data for discriminating among explanatory claims even at one level of explanation, let alone for simultaneously investigating factors at the biological, individual, and social levels. Even so, causal explanation is the holy grail of science, not just for its aesthetic parsimony, but also for the utilitarian causal generalization it promotes in the form of extrapolation.

Extrapolation is also achieved through careful induction. If patient education reduces the length of hospital stays and does so with patients suffering from 25 different diagnoses, then it is likely to do so with the 26th. Of course, the logical warrant for such induction is limited, and seems more reasonable the more numerous and heterogeneous are the patient types already examined. If all 25 were different forms of cancer diagnosis, or involved patients aged between 25 and 40, or were restricted to men, this limited heterogeneity would reduce our confidence in any inductive-empirical extrapolation. Meta-analysis is particularly important in this regard, especially when it permits replication across multiple different sources of irrelevancy. A relationship that has proven to be robust in whatever form it was operationalized is more likely to re-occur when other ways of operationalizing its components are attempted—more likely, but not certain. We see here the value traditionally assigned in science to discovering both general dependencies and the contingencies that describe when a given relationship does and does not hold.

The final context for extrapolation relates to formal modeling. When subgroups can be arrayed along a single continium, the functional form of a causal relationship can then be described. Is it stable? Is there a linear increase or decrease? If a pattern clearly emerges, one might want to argue that the strength of the relationship can be inferred for unobserved subpopulations with values that fall within the continuum range. Simple interpolation is involved here, while the strength of the relationship for values beyond the continuum is inferred by extrapolating from the observed function. But comfort with such extrapolation depends on how distant the nonobserved subpopulation values are from the sampled ones.

Model-based extrapolation is always tricky, but especially so when the models are more complex. Consider much of macroeconomics. Models of a given economic system are developed from theory and past findings, and changes in parameter values are introduced to see how the system reacts to this perturbation. This is basically an extrapolation exercise to discover what would happen if some details of the economic system were not as they actually are. Can we extrapolate from what we think we know, in order to learn what would happen under novel circumstances? The validity of any extrapolation depends on the model being correctly specified with respect to the variables included, the relationships postulated between them, and the strengths of these relationships. Be wrong about any of these assumptions and any conclusions about the effects of the novel perturbation will be biased to some unknown extent. The assumption that the model is complete and accurate is difficult to test, though crucial to the credibility of the results (Glymour et al. 1987). Still, the creation of valid multivariate models of causal interdependency is another major goal of science, and it is not difficult to see why. It promises the extrapolation of results beyond whatever circumstances were studied in the past.

5. *Conclusion*

In the behavioral and social sciences at the beginning of the twenty-first century, theory and practice are most developed when individuals or households are sampled to describe a human population. Here, best practice requires randomly sampling units from the population. Theory and practice are also well developed for generalizing from a measure to an abstract construct or from an experimental treatment to a more general causal agent. Here, best practice requires an explicit theory of construct validity that necessarily invokes proximal similarity, but preferably also the heterogeneity of irrelevancies, discriminant validity, and causal explanation. However, the purposive selection and theory-linked methods needed for construct validation are not as formally well supported as random selection is. Theory and practice are even less developed for generalizing to settings and historical time periods and for identifying the conditions under which a causal relationship holds. However, various deliberate sampling and replication strategies are relevant here, including meta-analysis. At their best, such strategies borrow much of the logic underpinning sophisticated construct validation. Least well developed are theory and practice for extrapolating results to nonstudied circumstances. But even here,

understanding causal explanatory processes can permit re-creation of the same causal forces in novel settings. Also, there are now decades of research and reflection about using animals to extrapolate to humans and about using the laboratory to extrapolate to other social settings. In addition, substantive modeling is sometimes used to examine what would happen to a system if one or more of its parameters were changed. Of all the methods examined, only random sampling provides an impeccable formal rationale for generalization. But its reach is, alas, limited. It is a great irony that the best formal method for generalization in the behavioral and social sciences can be used for only a small part of the generalization tasks that practicing social scientists routinely face. Given this, practicing social scientists and theorists of method reflecting on these practices have cobbled together more complex methods of generalization that depend more heavily on substantive theory, purposive sampling, heterogeneity of irrelevancies and demonstrated replication and the identification of causal contingencies.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Causation: Physical, Mental, and Social; External Validity; Generalizability Theory; Internal Validity; Sample Surveys, History of; Sample Surveys: The Field

Bibliography

Blake R 1988 Cat spatial vision. *Trends in Neuroscience*. **11**: 78–82

Brunswik E 1956 *Perception and the Representative Design of Psychological Experiments*, 2nd edn. University of California Press, Berkeley, CA

Campbell D T 1957 Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* **54**: 297–312

Campbell D T 1988 *Methodology and Epistemology for Social Science: Selected Papers* [ed. Overman E S]. University of Chicago Press, Chicago

Campbell D T, Fiske D W 1959 Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* **56**: 81–105

Cook T D 1993 A quasi-sampling theory of the generalization of causal relationships. In: Sechrest L, Scott A G (eds.) *New Directions for Program Evaluation: Understanding Causes and Generalizing about Them* (Vol. 57). Jossey-Bass, San Francisco

Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand-McNally, Chicago

Cooper H, Hedges L V (eds.) 1994 *The Handbook of Research Synthesis*. Russell Sage Foundation, New York

Cronbach L J 1982 *Designing Evaluations of Educational and Social Programs*. Jossey-Bass, San Francisco

Cronbach L J 1989 Construct validation after thirty years. In: Linn R L (ed.) *Intelligence: Measurement, Theory, and Public Policy*. University of Illinois Press, Urbana, IL, pp. 147–71

Gadenne V *Die Güttigkeit psychologischer Untersuchungen*. Kohlhammer, Stuttgart

Glymour C, Scheines R, Spirtes P, Kelly K 1987 *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press, San Diego, CA

Mackie J L 1974 *The Cement of the Universe: A Study of Causation*. Oxford University Press, Oxford, UK

Shadish W R, Cook T D, Campbell D T 2001 *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston

T. D. Cook

Generation in Anthropology

In anthropology, the term generation has two primary meanings. In one of its senses, generation refers to a group of people who have lived through a time period together and have developed some kind of shared consciousness. Generation has been used even more commonly by anthropologists in its genealogical, kinship-related sense, to refer to the relationship between parents and their children, these children and their children, and so on. Using one or both of these senses of the term, anthropological interest in generation has focused on four primary topics, pertaining to the ways people across cultures organize and envision social relationships, identities, and the processes of human social-cultural life. This article will look at each approach in turn.

1. The Cultural Construction of Intergenerational Ties

Work on the cultural shaping of intergenerational ties focuses on generation in its genealogical sense as a descent relationship. Interest in intergenerational family relationships has existed since the beginnings of anthropology as a discipline, largely due to the analysis paid by anthropologists to kinship systems. Since about the 1970s, the majority of work on intergenerational ties has come out of the subfield of anthropology of aging, where one of the central research issues has surrounded the degree to which the elderly can expect to receive various forms of support from their younger relatives. Research on intergenerational ties, however, goes well beyond the (interesting and important) problem of how the elderly are cared for, to speak to broader conceptions about the nature of the family, the relationship between individuals, families and the state, and the social-moral order.

Lamb (2000) examines such problems through a focus on West Bengal, India. She finds that Bengalis perceive intergenerational family relations to entail long-term bonds of reciprocal indebtedness extending throughout life and even after death. Juniors provide care for their elderly parents, reconstruct relations with parents as ancestors after death, and ritually nourish these ancestors, as a means of repaying the tremendous 'debts' owed for producing and caring for them in infancy and childhood. What parents once

gave their children to produce and raise them—such as a body in birth, food, material goods, money, a home, forms of love, the cleaning of urine and excrement—children are expected to provide in return for their parents years later in old age and after death. (Children, for instance, construct new spirit-bodies for their parents as ancestors, parallel to the bodies parents create for their children in birth). Other anthropologists have found that relationships of lifetime intergenerational reciprocity, conceptualized and practiced in varying ways, are not uncommon cross-culturally (see, e.g., Cattell 1990).

The shaping of intergenerational ties can bear significance beyond the family as well, to implicate the social-moral-political conditions of a broader society and nation. Indian discourses, for instance, present close, multigenerational families as signifying a 'good' and 'Indian' society; whereas disintegrating generational ties—the break-up of joint families, uncared-for old people, independent daughters-in-law—are held to be paradigmatic signs of a widely degenerate 'modern,' postcolonial, and urbanizing era (Cohen 1998, Lamb 2000).

Some anthropologists argue that the shaping of intergenerational ties can be understood through scrutinizing surrounding cultural values. Simic (1983) suggests links between an American model of intergenerational relations and a broader cultural orientation toward self-fulfillment and individualism. Children in middle-class, Anglo–American households maintain their own rooms and possessions, and by about age 18 are expected to move out and live separately from their parents. Though parents often extend financial aid to their children after they have left home, Simic observes, 'it is almost always with the overt rationale of helping them get established as independent units, rather than creating a binding reciprocity of mutual dependence between the two generations.' In later life, this same ideology prevails: among the aged economic independence is a fundamental value; dependency is 'demeaning.'

In general, research has shown that in developing countries, intergenerational family relations tend to be closer and more mutually interdependent than in capitalist, industrial settings, although many important exceptions can be found (see Albert and Cattell 1994, Sokolovsky 1990). A wealth of careful, detailed, cross-cultural data on generational ties has been produced by anthropologists, revealing a profound variety in the ways intergenerational relationships are conceived, maintained, practiced, and given value.

2. *Generation as an Element of Social Organization*

Anthropologists have also examined the ways social and cultural groups employ some concept of generation to organize social relationships beyond the family. Many cultures group people into generation-defined categories, taking some notion of genealogical succession within the family and extending it to think about broader social relations (Kertzer 1979). Examining generation in this sense, as a principle of social organization, has allowed especially social anthropologists and those concerned with social structure to address questions about the very ordering and continuity of societies (e.g., Baxter and Almagor 1978, Eisenstadt 1956, Fortes 1984).

In some societies, social generational ordering is highly formalized and elaborate. Rituals, variations in dress, forms of speech, permitted behaviors, specialized tasks, and places of residence may sharply distinguish persons of one generational group from another. The most highly institutionalized organizations of generations into social groupings, often referred to as generation-set systems, have been found among a number of East African groups (Baxter and Almagor 1978). In such groups, the generation model is extended so widely as to provide the basis for regulating social relationships in an entire society. Every male is a member of a named generation set, and the generation set a man is in is determined by the generation set his father is in.

The crucial rule in such generation-set systems is that successive genealogical generations of fathers and sons must be socially and politically separated. Some scholars suggest that this is a way of minimizing generational conflict and competition, for instance over wives, land, and political leadership, for members of different generation-sets generally cannot compete over the same resources and privileges. Generation-set systems have almost always been organized around males, and women may be tied to them peripherally as daughters, wives and mothers. Versions of such systems have also been found in Melanesia, among Australian aborigines, and among some Native American groups.

Even in societies where people are not organized into formal generation sets, some notion of generation as a principle of social organization is almost universally employed (see Kertzer 1979). For instance, the social roles or identities a person assumes may depend on whether he or she represents the junior or senior generation in a family. In many societies, a person's transition to social adulthood is linked to the birth of his or her children; and in India, the transition to a 'senior' stage in life and society comes about not so much through reaching a particular chronological age but through the marriage of one's children (Lamb 2000).

It is necessary to distinguish between generation and chronological age as principles of social organization and of cultural value. Most societies use some version of both in ordering social relations, but these categories are often confused or overlapping in social scientific analyses and among the people being studied (Kertzer 1979). Age, however, is distinct from gen-

eration: it has to do with the chronological passage of time, and rests on the formally isolated individual as the unit of reference whereas generation is family-generated. There are many situations in which relationships by generation conflict with relationships by age. For instance, among the Tallensi, no man, even if very advanced in age, can achieve jural autonomy until his father dies (Fortes 1984). In the US, social criticism may ensue if a man marries a woman 'young enough to be his daughter' (the woman as wife becomes the man's generation, but has the chronological age of his children's generation).

Several anthropologists have been intrigued by the fact that what is essentially an intrafamily principle of genealogical generations has been used so widely by human groups to think about social relations outside the family. People of diverse cultures, it seems, find families 'good to think with.'

3. *Generational Identity as an Axis of Social Difference*

A third way that anthropologists have approached the study of generation relates to what is probably the most common usage of the term in English: generation as a cohort of people who have lived through a time period together and share a sense of identification with others of their era. Sociologist Mannheim's classic essay, The Problem of Generations (1952), is often cited as the piece that popularized this sense of generational studies within the social sciences. To Mannheim, a generation is not simply a cohort—a group of people having a statistical factor, such as age or decade of birth, in common; a generation is made up of people who, born during the same time period, feel the impact of common historical events and cultural forces, and develop a shared consciousness or identity around these experiences. Anthropologists have expanded upon this sense of generation as a form of identity, using it to heed the multiple, competing perspectives, voices, and experiences that make up any culture. Along with gender, class, race, and ethnicity, generation can be a salient axis of social difference.

Newman's work (e.g., 1996) highlights such a sense of generation as an important theoretical concept for anthropology. In one study, she collected life-history data from divorced mothers in the US who had been pushed into downward mobility. The study did not begin as a generational story, yet she found that the life history texts bifurcated strongly by generation. Women who were born during the Depression were invested in different versions of American culture than those who were raised amidst the plenty of the 1950s and 1960s. 'It became clear,' Newman reflects, 'that the source of their views was a fairly tightly woven and self-conscious generational culture, or symbolic dialect,' which formed a grid though which their life circumstances were interpreted (1996).

Comaroff and Comaroff (1999) take a powerful look at generation as one of the dominant idioms through which youths in postcolonial South Africa frame their identities and place in the world. Widespread anxiety over the hardening materialities of life in an era of global capital translates frequently into bitter generational opposition. Many young black men blame their current incapacity to ensure a viable future for themselves on a perceived sinister, all-consuming aged elite, referred to as 'witches' or 'old ladies' (even when they are men). The Comaroffs cite one example of a group of youths pelting an old woman with sticks and stones. 'Why are you killing me, my grandchildren?' she wailed. Her assailants responded, 'Die, die you witch. We can't get work because of you!' The major lines of opposition in contemporary South Africa, the Comaroffs find, are not race or class, but generation, mediated by gender.

Such work reveals the potential salience of generation as a category of social experience and analysis. A society at any given moment may be perceived, by the people within it, and/or by the anthropologists studying it, as being crosscut by generations representing powerfully divergent values, beliefs, perspectives, and agendas.

4. *Generation and Social Change*

Finally, anthropologists have used generation as a means to explore social change, history, and time. Just as generation is one of the key images people around the world use to represent social change (for instance, by telling of how children are growing up in a world vastly different from that of their parents or grandparents), social theorists often think about processes of social or cultural change in generational terms. Mannheim speculated in the 1920s that cultural transformation comes about through the continuous emergence of new generations: as each generation comes into 'fresh contact' with its social and cultural heritage, it remodels what is found (1952). Anthropologist Fortes mused several decades later that such a generational model is an 'apt imagery by means of which to depict continuities and discontinuities in a community's social and cultural life over a stretch of time' (1984). It is within the framework of the succession of generations, Fortes explains, that all people acquire their cultural heritage. 'Mediated from earliest infancy by the natal family, the society readily takes on a generational form to the individual.' The concept of culture has always been defined by anthropologists as something that is 'transgenerational' at the same time that it is open to creative revision and change.

In the contemporary era of globalization and rapid social transformation, perceived 'gaps' between the lifeways of senior and junior generations may become, for many, ever more profound. Generation will be a

provocative place for anthropologists to look to apprehend the ways global processes of change take shape in particular places and lives.

5. The Importance of Generation in Future Work

Generation in anthropology has been a key topic in the study of aging, kinship and social organization, but it has otherwise remained rather undeveloped. That is, conceived as a form of subcultural identity, axis of social difference, or dimension of social change, generation has not taken center stage. The category of generation can be fruitfully exploded, however, to explore problems of current critical interest to the discipline—such as the ways processes of globalization play out in local lives, and diverse voices and identities make up particular cultural worlds.

See also: Age: Anthropological Aspects; Age, Sociology of; Age Stratification; Age Structure; Ancestors, Anthropology of; Cohort Analysis; Culture: Contemporary Views; Family, Anthropology of; Family as Institution; Family Theory: Economics of Intergenerational Relations; Generations in History; Generations, Relations Between; Generations, Sociology of; Globalization, Anthropology of; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements

Bibliography

Albert S M, Cattell M G 1994 *Old Age in Global Perspective: Cross-cultural and Cross-national Views*. Macmillan, New York

Baxter P T W, Almagor U (eds.) 1978 *Age, Generation and Time: Some Features of East African Age Organisations*. Hurst, London

Cattell M G 1990 Models of old age among the Samia of Kenya: Family support of the elderly. *Journal of Cross-cultural Gerontology* **5**: 375–94

Cohen L 1998 *No Aging in India: Alzheimer's, the Bad Family, and Other Modern Things*. University of California Press, Berkeley, CA

Comaroff J, Comaroff J L 1999 Occult economies and the violence of abstraction: Notes from the South African postcolony. *American Ethnologist* **26**: 279–303

Eisenstadt S N 1956 *From Generation to Generation: Age Groups and Social Structure*. Free Press, Glencoe, IL

Fortes M 1984 Age, generation and social structure. In: Kertzer D I, Keith J (eds.) *Age and Anthropological Theory*. Cornell University Press, Ithaca, NY

Kertzer D I 1979 Generation and age in cross-cultural perspective. In: Riley M W, Abeles R P, Teitelbaum M S (eds.) *Proc. AAAS Selected Symposium 1979. Aging from Birth to Death: Volume II: Sociotemporal Perspectives*. Westview Press, Boulder, CO

Lamb S 2000 *White Saris and Sweet Mangoes: Aging, Gender and Body in North India*. University of California Press, Berkeley, CA

Mannheim K 1952 (1927) The Problem of generations. In: Mannheim K (ed.) *Essays on the Sociology of Knowledge*. Oxford University Press, London

Newman K 1996 Ethnography, biography, and cultural history: Generational paradigms in human development. In: Jessor R, Colby A, Shweder R A (eds.) *Ethnography and Human Development: Context and Meaning in Social Inquiry*. University of Chicago Press, Chicago

Simic A 1983 Aging in the United States and Yugoslavia: Contrasting models of intergenerational relationships. In: Sokolovsky J (ed.) *Growing Old in Different Societies: Cross-cultural Perspectives*. Wadsworth, Belmont, CA

Sokolovsky J (ed.) 1990 *The Cultural Context of Aging: Worldwide Perspectives*. Bergin & Garvey, New York

S. Lamb

Generations in History

Generations—cohorts of humans born about the same time and growing up through the same slice of society and history—have played a more important role in the human past than is generally acknowledged. During the nineteenth and twentieth centuries, younger generations in particular played colorful, sometimes crucial roles in history. As awareness of generational differences and conflicts has grown, furthermore, an evolving theory of social generations has developed to explain the phenomenon. The writings of such thinkers as Karl Mannheim, José Ortega y Gasset, and a host of more recent scholars have analyzed the impact of changing sociocultural influences and traumatic historical events in forging social generations which may then themselves become powerful forces for change in history.

1. Generations: The Term and the Concept

The term generation has been commonly used in a variety of senses, some of them marginal, others of central relevance to the generational historian's task. This opening section will suggest six quite different uses of the term generation and of the more precisely focused concept of the social generation in history.

Genealogical accounts of the history of individual families commonly refer to each new wave of siblings as a new genealogical generation. Useful as it may be, however, this tight focus on an individual family evidently lacks the scope of developed generational studies, which typically seek to survey an age-defined subdivision of society as a whole.

Family history more broadly conceived may probe relations between biological generations—parents and children—across a whole society. Even this broader approach, however, still suggests that generational

relations consist exclusively of relations between biological parents and children—another unnecessary limitation on the range of generational history.

Prosopography, offering a group portrait of coevals who occupy the same social niche as politicians, writers, or business leaders, for instance, does offer a form of generational history. Emphasizing similarities of backgrounds, beliefs, careers, and achievements as well as generational contemporaneity, such analyses clearly move beyond the family into the larger sphere of society as a whole.

Another form of generational analysis divides up historical time itself into generational periods of about 20 years, the time believed to elapse between the birth of one biological generation and the next. While this approach may uncover significant differences of generational perspective, it may also impose a series of generations of unconvincingly equal length upon the alternating longueurs and crisis periods of human history.

Perhaps the most useful form of generational analysis has been that which sees social or historical generations as birth cohorts, people born about the same time. This sort of generational analysis moves beyond the family and the single professional group and recognizes that different social generations may be of varying lengths and of greater or lesser significance, depending on social or historical circumstances. Above all, cohort analysis provides a solid foundation for generational studies in a given society at a specific moment in its history.

2. *Age Groups in History*

Age groups—groups of coevals—have been recognized in most societies and have been elaborately institutionalized in some. Like other subsets of society such as race, class, and gender, age groups have sometimes been both active forces in social conflict and creative agents of social change.

2.1 *Age Groups in Human Societies*

Anthropologists have detected age groups (age-grades or age-sets) among many pre-urban peoples. Progress from one age grade to the next may be celebrated by elaborate rites of passage, and membership in a particular age group often brings both duties and privileges. Young men will be expected to fight wars, young men and women may look forward to marriage, and older people of both sexes may serve as decision-making elders or perform important ceremonial functions.

But age groups are detectable in more complex urban societies also. Apprentices in guilds, students in schools, and even conscript armies may segregate many of the young from most of their elders. At the other end of life, retirement homes and communities separate older age groups from the rest of society.

In most pre-urban societies, younger and older people apparently tend to accept the same traditions and taboos. In more complex societies, however, conflict between age groups seems to have become considerably more common.

2.2 *Conflict Between Age Groups*

There is evidence of generational conflict in earlier urban societies than ours. Aristotle saw differences between youth and age at the heart of the social wars that frequently rent the fabric of Greek city-states. Medieval observers were dismayed by the riotous behavior of guild apprentices or the town and gown violence of university students. Many of the leaders of the European Reformation were young preachers, and historians have noted that the founding fathers of the USA were actually young men.

The clearest and most numerous examples of generational conflict, however, have been detected over the last two centuries. Nineteenth-century European commentators noticed the youth of the barricade-building ideological revolutionaries who challenged authority in the 1820s, 1830s, and 1840s. Twentieth-century observers around the world recognized the youthfulness of the ideologically inspired guerrillas, urban-terrorist groups, and revolutionary cadres who played such a central role in the history of Russia, China, and the developing societies of Asia, Africa, and Latin America, as well as in the most developed nations, in the 1900s.

3. *The Generational Idea in History*

Theoretical understanding of the generational idea has evolved in tandem with this apparently increasing historical significance of generational conflict. Earlier insights into the significance of generations in human society were brief and suggestive. The nineteenth and particularly the twentieth century, however, have seen much more elaborate and sophisticated development of the concept.

3.1 *Early Awareness of Generations in Society*

Old Testament genealogies and Homeric poetry reveal clear concern with the succession of human generations. The traditional father of history, Herodotus, posited 33-year generations in his *History of the Persian War*, and the Arab world historian Ibn Khaldun proposed a basic theory of generations in his *Muqaddimah: An Introduction to History*.

Only in the nineteenth century, however, did social thinkers of the caliber of Auguste Comte, John Stuart

Mill, and Wilhelm Dilthey begin to develop a coherent and convincing theory of social generations. Julián Marías (1949), a leading historian of the generational idea, emphasizes their major contributions. Comte, he points out, clearly recognized the succession of generations as an important causal factor in social change. Mill was aware that the generational concept must be expanded beyond a familial or genealogical framework. Dilthey attempted to analyze the basic structure and significance of social generations.

3.2 Classic Formulations of the Theory of Social Generations

The first half of the twentieth century, particularly the period between the two world wars, saw classic formulations of the generational idea by the Spanish philosopher José Ortega y Gasset (1923, 1933) and the German sociologist Karl Mannheim (1928). Ortega's account of the phenomenon focuses on the internal development of the typical social generation through four 15-year stages: childhood, youth, initiation to positions of power, and domination of society. This stages-of-life perspective has influenced many generationists since.

Mannheim's analysis emphasizes the roots of generational differences in changing social and cultural circumstance. Arising in different social locations (e.g., middle or working class, rural or urban environment) and shaped by new trends in society at large, new social generations develop different points of view and patterns of behavior from those prevalent among their elders. Mannheim also analyzes the structure of the social generation, suggesting that a given birth cohort may be subdivided into a number of distinct generation units, subgroups who work up the common generational experience in very different, often ideologically definable ways.

3.3 Development and Validation of the Generational Idea

In the wake of the global pandemic of youth revolt of the 1960s and earlier 1970s, finally, sociologists, political scientists, and other social scientists researched the realities and refined the theory of generational development and interaction as never before. Such theoretical and empirical students of the generational phenomenon as Vern L. Bengtson, Richard G. Braungart, T. Allen Lambert, Robert S. Laufer, and Marvin Rintala elaborated a convincing model of the social generation in history. These scholars probed the sociopsychological roots of generational consciousness in immense detail. They documented the persistence of generational identity long after the strong orientation toward one's youthful peer group has faded. They collected masses of contemporary evidence of the existence of generational conflict and of its influence on social change.

4. The Theory of Social Generations

As the twenty-first century begins, then, one form or another of generational analysis is used routinely, not only by historians, political scientists, sociologists, anthropologists, and critics of literature and the arts, but by journalists, politicians, marketing experts, and many others. These diverse commentators explore age groups just as they do groups defined by gender, class, or ethnicity. The brief exposition of a developed, essentially Mannheimian view of the sociohistorical generation that follows will try to build on this broad awareness.

4.1 Social Circumstances Shape Social Generations

A social generation, then, consists of a birth cohort, a group of individuals born in similar circumstances during the same span of years, and thus shaped by common social and cultural influences. The critical period in the shaping of a generation comes in its earlier years, when its members are most malleable and open to the influence of their sociocultural environment.

During the childhood of a birth cohort, common family structures, child-rearing methods, and educational practices begin this socialization process. Thus a patriarchal family structure, a child-rearing philosophy of spare the rod and spoil the child, and an education built around rote memorization may tend to produce one sort of younger generation. A more egalitarian family structure, permissive child-rearing methods, and schools committed to developing the individual propensities of each child may produce a very different social generation.

During the adolescence and youth of a birth cohort, the larger society in which it develops may also exert powerful developmental influences. Hierarchical or egalitarian relations in society as a whole, traditionalist or cosmopolitan cultures, authoritarian or democratic politics, all have an impact on young people just emerging into the world beyond the family. Within the society as a whole, furthermore, the particular generational location within which youth matures will also have its influence. Young people raised in working-class ghettos will grow up with a very different view of the world from that of their coevals raised in middle-class suburbs, even if both these locations are within the same larger society. At every level, the books young people read, the films they see, the songs they hear and sing, shape the beliefs and values of a social generation in ways we are only beginning to understand.

4.2 Historical Traumas Define Rebellious Generations

The socialization process thus shapes the members of each new generation from their earliest years well into

their twenties. To these influences of home, school, and society at large, however, historical study of the generational phenomenon would add a special emphasis on the traumatic historical event in the shaping of a generation.

Popular accounts of specific social generations have tended to stress the impact of history on the young. Wars may have a much more profound impact on the lives and attitudes of the young men who fight them than on their elders. A major economic depression can be far more traumatic for young people discouraged by poverty from starting careers and families. Such historic events, furthermore, frequently challenge the view of the world the young have absorbed from other sources, producing confused, disillusioned, angry, or rebellious younger generations.

As Lewis Feuer pointed out, traumatic events may thus undermine the authority of both traditional institutions and of the older generation that imposes them (Feuer 1969). In this way the mores and morals of the young may change drastically as they shrug off the prescriptions of a de-authoritized older generation. Where anti-establishment ideas and attitudes are already present in the culture, such historic traumas may transform hitherto isolated advocates of change into leaders of large-scale generational revolts. Any historic dissonance between widely held social ideas and changing social realities may produce generational disillusionment. A traumatic historical event may multiply this effect, shaking the foundations of society and calling its deepest principles into question in the minds of the young.

4.3 Generations Shape the Course of History

It is sometimes difficult to identify the historic victories or social transformations brought about by generational revolts. Youthful rebellions like the European political revolutions of 1848 or the very different American and Chinese cultural revolutions of the 1960s may fail to achieve many of their declared goals. Yet the new commitments and the youthful activism of social generations who march to a different drummer do often make a difference.

When young black Americans, for instance, took to the streets in the early 1960s, laws and attitudes changed. As a result, doors to social advancement opened, and the lives of at least middle-class black Americans were significantly altered. Similarly, the dominantly young people who rallied to the banners of leaders like Vaclav Havel in Eastern Europe in 1989 toppled political and economic institutions, radically transforming life in such nations as Poland, the Czech Republic, East Germany, and Hungary.

The openness of younger birth cohorts to new cultural influences may also have transforming social consequences. Around 2000, the spread of capitalism, computer technology, and even such specific ideas as freedom of thought and expression, or gender equality began to change the lifestyles of the young, and through them their societies, in many parts of the world.

5. Problems and Potential of Generational Analysis

Major theoretical problems remain for the generational historian. There does, however, appear to be real potential for future development and application of the generational approach to history.

5.1 Problems of Generational Theory

Like any broad attempt to explain society and history, the theory of social generations has generated its share of challenging questions.

How long is a generation, for instance? Thirty-three years as Herodotus declared, 15 years with Ortega, or even three years, as some modern demographic studies might suggest? Historically, the length of a significant social generation seems to depend on the pace of change. The more rapidly things change, the more quickly enough changes accumulate in the shaping social matrix to produce a significantly different generational point of view.

But how important are social generations? People unsympathetic to the approach often ask this question, frequently having a rival theoretical framework to suggest, such as social class. However, generational theory makes room for class differences in the concept of generational location, which Mannheim saw as an expanded form of class identity. There would in fact seem to be room for analyses of the same population in terms of its component races, classes, genders, or social generations—all important bases for analysis, none more fundamental than the others.

Finally, are social generations historical realities at all? Some observers have noted that even when the streets are full of young demonstrators, most young people seem to share their parents' views. Certainly the socialization process does succeed more often than not, even in eras rent by generational discord. In history, however, numerical majorities are not necessary to produce significant consequences. When socialization fails with even a critical minority of the young, the resulting conflict of generations can produce historic changes in society.

5.2 Historical Research on Generational Themes

The uses to which generational analysis has been put during the twentieth century are manifold, and new topics well suited to interpretation in terms of generations emerge with every passing decade.

The generational approach has been widely utilized by historians and critics of art and literature. Students

of such increasingly important features of modern history as immigration have also found generational analysis invaluable. And there is a vast literature on younger generations mobilized in student or youth movements, countercultures, or revolutionary cadres.

During the late twentieth century, the generational approach was also applied to changing attitudes and patterns of activity in many non-Western lands. India's crisis of values has been explored in a number of such studies. Several Chinese student movements, as well as the massive Maoist Cultural Revolution, have been analyzed as revolts of younger generations. And more research is needed on the generational factor in the postwar colonial liberation movements and in the jolting changes that have accompanied urbanization, modernization, and globalization in the developing world since.

In developed societies with rapidly aging populations, finally, the other end of the spectrum of age groups would seem to demand historical study. With the progress of gerontology and our growing awareness of the persistence of generational memory as background, historians might profitably probe generational relations in past cultures in which older age groups have enjoyed both prestige and real power.

5.3 Related Concepts

Both the historic reality and the historiographical concept of the generation are ancient ones, yet we are often surprised when generational differences surface once again in history. Given the likelihood that such challenges will continue to occur, the theory of social generations would seem to merit both further development and more vigorous application to the study of society and history.

See also: Age: Anthropological Aspects; Age, Sociology of; Age Stratification; Age Structure; Diffusion: Anthropological Aspects; Education: Anthropological Aspects; Elites, Anthropology of; Generation in Anthropology; Generations: Political; Generations, Relations Between; Generations, Sociology of; Life Course in History; Life Course: Sociological Aspects; Prosopography (Collective Biography); Right-wing Movements in the United States: Women and Gender; Social Movements and Gender; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements; Youth Movements, History of

Bibliography

Bengtson V L, Laufer R S (eds.) 1974 Youth, generation, and social change. *Journal of Social Issues* **30**: issues 2 and 3

Braungart R G, Braungart M M 1986 Life course and generational politics. *Annual Review of Sociology* **12**: 205–31

Eisenstadt S N 1956 *From Generation to Generation: Age Groups and Social Structure*. Basic Books, Glencoe, IL

Esler A 1972 *Bombs Beards and Barricades: 150 Years of Youth in Revolt*. Stein and Day, New York

Feuer L S 1969 *The Conflict of Generations: The Character and Significance of Student Movements*. Basic Books, New York

Lambert T A 1972 Generations and change: Toward a theory of generations as a force in historical process. *Youth and Society* **4**: 1, 21–45

Le Vine V T 1968 Generational conflict in Africa: A paradigm. *Civilisations* **18**: 399–418

Mannheim K 1928 Das Problem der Generationen. *Kölner Vierteljahrshefte für Soziologie* **7**: 157–85, 309–29 [English translation 1952 The problem of generations. In: *Essays on the Sociology of Knowledge*, Oxford University Press, New York, pp. 276–320]

Marías J 1949 *El Método Histórico de las Generaciones*. Revista de Occidente, Madrid, Spain, [English translation 1970 *Generations: A Historical Method*. University of Alabama Press, University, AL]

Ortega y Gasset J 1923 *El tema de Nuestro Tiempo*. Revista de Occidente, Madrid, Spain [English translation 1933 *The Modern Theme*. Norton, New York]

Ortega y Gasset J 1933 *En torno a Galileo*. Revista de Occidente, Madrid, Spain [English translation 1958 *Man and Crisis*. Norton, New York]

Peyre H 1948 *Les Générations Littéraires*. Boivin, Paris

Pinder W 1926 *Das Problem der Generation in der Kunstgeschichte Europas*. Frankfurter Verlags-Anstalt, Berlin

Rintala M 1963 A generation in politics: A definition. *Review of Politics* **25**: 509–22

Sinha D 1972 *The Mughal Syndrome: A Psychological Study of Intergenerational Differences*. McGraw-Hill, Bombay, India

Spitzer A B 1973 The historical problem of generations. *American Historical Review* **78**: 1353–85

Wohl R 1979 *The Generation of 1914*. Harvard University Press, Cambridge, MA

A. Esler

Generations: Political

'Political generations' can be defined as politically distinctive age groups delimited by their years of birth or some major event, such as their year of entry into the electorate. Rintala (1979, p. 8) defines a political generation as 'a group of human beings who have undergone the same basic historical experiences during their formative years.' According to Braungart and Braungart (1989, p. 283) 'a political generation is an age group in history that mobilizes to work for political change,' although an age group could work to maintain the status quo and still be a political generation.

1. Generation and Cohort

Birth, aging, and death gradually transform the composition of any society. New generations supplant the old, but transforming the population may or may not lead to social and political change. The young may resemble the old in their political attitudes and

behaviors and, even when young people do differ from their elders, they may come to resemble them when they themselves age.

But although each generation perpetuates society, it may also change it. If the young could be isolated from their elders, it would be easier to create a new social order, and thus Plato argued that philosophers could establish the ideal polity more quickly if they expelled all citizens above the age of 10 from the republic. At times a generation may be too corrupt to establish a new political order. The Book of Numbers relates that the generation freed from Egypt, which again and again showed its faithlessness, was condemned to have its carcasses consumed in the wilderness until a new generation, born in freedom, was reared to conquer Canaan.

Although the concept of generations has biblical and Greek roots, the social scientific study of generations began with Mannheim's essay, 'The Problem of Generations' [1928] (1952).

According to Mannheim (1952), the concept of generations is both biological and sociological. Generations are produced through 'the biological rhythm in human existence,' but 'individuals who belong to the same generation, who share the same year of birth, are endowed, to that extent, with a common location in the historical dimension of the social process' (p. 290).

Mannheim (1952) maintains that historical experiences are most likely to influence persons in their late adolescence and early adulthood, postulating that 'personal experimentation with life' usually begins at about age 17. Whether young people are actually influenced by historical events depends on social and historical conditions. For example, young adults living in different cultures may experience quite different events, and, even within a single society, some persons may be relatively isolated from the mainstream of social events and thus be unaffected by major historical movements. Nor do all historical periods produce conditions that influence young adults in distinctive ways. Mannheim (1952) argues that distinctive generations are more likely to form in periods of rapid social change. But his basic point is that there is no guarantee that socially or politically distinctive generations will emerge: 'Whether a new *generation style* emerges every year, every thirty, every hundred years, or whether it emerges rythmically at all, depends entirely on the trigger action of the social and cultural process' (p. 310).

Following Mannheim's (1952) cautions, it is unwise to label a group of people a political generation unless one can establish that they are actually distinctive in their attitudes and behaviors. Therefore, social scientists need a term describing persons born during the same years, but making no other assumptions about their distinctiveness. The term 'cohort' is used for this purpose. This term derives from Roman military terminology, since the cohort was a unit within a Roman legion, but the term now has a much broader usage. Scholars usually define a cohort as a group of people born during the same years, but there are many other factors that can define a cohort: the entry class of a school, persons married during the same period, or persons who vote for the first time in the same election. But the most basic experiences are usually birth and death, and the former is usually used to define a cohort.

There is no definitive answer to how many years a cohort should include. In the 1960s, and even into the early 1970s, survey researchers often used arbitrary age categories to code years of birth, often grouping respondents into 10-year age categories in order to save one column of usage on the IBM punch cards that were used to store data. Scholars who want to work with these early data sets must employ these arbitrary age categories. With computerized storing techniques, survey age is usually coded by year of birth, and one has great flexibility in analyzing survey results For example, a major analysis by Converse (1976) divides respondents into three-year age categories, and throughout their study of the US electorate, Miller and Shanks (1996) generally divide their results into cohorts that are four years in width.

2. *Life Cycle, Generational, and Period Effects*

Whenever young adults differ from their elders there are usually two plausible interpretations for these differences. A life cycle interpretation maintains that these differences result mainly from the inexperience of the young, and postulates that the young will become more like their elders when they age. A generational interpretation maintains that these differences result from the differences between the formative experiences of young adults and those of their elders.

One of the most extensive controversies in political science research has been the well- documented finding that young US citizens tend to have weaker party loyalties than their elders. This finding was reported in *The American Voter*. Lacking longitudinal data, Campbell and his colleagues (1960) relied upon an indirect test comparing the partisan loyalties of persons of the same age who have held their partisan loyalties for different lengths of time. Even though their test was indirect, Campbell and his colleagues (1960) reached strong conclusions that favored a life cycle formulation. The weak party loyalties of young Americans, they argued, resulted from their youth and political inexperience. These conclusions were strengthened through a cross-national analysis, reanalyzing Almond and Verba's study of the USA, Britain, West Germany, Italy, and Mexico. As Converse (1969) showed, *The American Voter* formulation could explain a process through which partisan strength among the electorate is held in equilibrium.

Even though older voters with strong party ties are dying and being replaced continuously by younger voters with weak partisan loyalties, the dynamics that increase partisan strength with age could prevent the erosion of partisan strength among the electorate. If, as Campbell and his colleagues (1960) argued, party loyalties among the electorate contribute to political stability, the life cycle explanation, if valid, had major implications for the political system.

Although *The American Voter* explanation was unchallenged for a decade, in the early 1970s scholars began to question it, based upon the finding that partisan loyalties did not become stronger as cohorts were tracked across the life cycle. Converse and his supporters acknowledged that partisan loyalties after 1964 did not increase among cohorts as they aged, but maintained that this resulted from the political dislo cations of the 1960s and 1970s. In other words, the failure of partisanship to increase with age was held to result from the events of the post-1964 period. By the late 1980s at least 20 articles had focused on this controversy (for a summary, see Abramson 1989) and several more appeared in the 1990s. But the space devoted to this controversy has declined as scholars have increasingly recognized the intractable difficulties of distinguishing among life cycle, generational, and period effects. Instead, scholarship has turned to studying the impact of generational replacement upon political attitudes and behavior.

3. *The Impact of Generational Replacement*

There is no debate that the population is gradually transformed as a result of generational replacement. (Some scholars use the term 'cohort succession.') Replacement can play a major role in transforming the political attitudes and behaviors among mass publics. For example, during the 1930s and 1940s, replacement helped make the Democrats the majority party in the USA, and during the postwar years replacement was a major force eroding US party loyalties. Replacement benefited the British Labour party during the 1960s, but it contributed to the erosion of British partisan loyalties between 1964 and 1987. Between 1969 and 1988, replacement contributed to the decline of the Labor Party Alignment in Israel. During the postwar years, replacement may have contributed to a more partisan electorate in West Germany. Replacement may also help transform political values. For example, replacement helped to increase political tolerance in the USA between the mid-1950s and the mid-1970s. (For the citations to the sources documenting these trends, see Abramson and Inglehart 1995.) Women who came of age in the USA after the enactment of the Nineteenth Amendment played a major role in increasing women's turnout (Firebaugh and Chen 1995).

The most extensive analysis of the impact of generational replacement upon value change among mass publics has been conducted by Inglehart and his colleagues (Abramson and Inglehart 1995) who focus on changing levels of postmaterialism. Respondents who give high priority to postmaterialist values tend to emphasize freedom and quality of life concerns, whereas materialists emphasize economic and physical security. In all of the European societies they study, young adults are more likely to have postmaterialist values than their elders. During the two decades that Abramson and Inglehart (1995) studied, about two-fifths of the population of these societies had been replaced. As Europeans with higher levels of materialism were dying out of these societies, and as they were being replaced continuously by Europeans with postmaterialist values, population turnover had affected the overall distribution of values in these societies. Analysis of a combined sample of the European societies that can be studied between 1970 and 1992 suggests that about nine-tenths of the overall shift toward postmaterialism resulted from replacement.

4. *Methodological Issues and Future Research*

Future research on the impact of generational replacement will face methodological problems. As cohorts are tracked over time, the older cohorts die. Samples of the very old become small and nonrepresentative and, eventually, dwindle away. The standardization procedures used by Abramson and Inglehart (1995) need to assign some value to cohorts that can no longer be tracked, but imputing any such value is problematic. The best one can do is to present estimates using a variety of assumptions, some of which have been suggested by Firebaugh (1989).

See also: Age, Sociology of; Age Stratification; Cohort Analysis; Generations in History; Generations, Relations Between; Mannheim, Karl (1893–1947); Youth Culture, Sociology of; Youth Movements

Bibliography

Abramson P R 1989 Generations and political change in the United States. In: Braungart R G, Braungart M M (eds.) *Research in Political Sociology*. JAI Press, Greenwich, CT, pp. 235–80

Abramson P R 1992 Of time and partisan instability in Britain. *British Journal of Political Science*. **22**(July): 381–95

Abramson P R, Inglehart R 1995 *Value Change in Global Perspective*. University of Michigan Press, Ann Arbor, MI

Braungart R G, Braungart M M 1989 Political generations. In: Braungart R G, Braungart M M (eds.) *Research in Political Sociology*. JAI Press, Greenwich, CT, pp. 281–319

Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. Wiley, New York

Converse P E 1969 Of time and partisan stability. *Comparative Political Studies* **2**(July): 139–71

Converse P E 1976 *The Dynamics of Party Support: Cohort-Analyzing Party Identification*. Sage, Beverly Hills, CA

Firebaugh G 1989 Methods for estimating cohort replacement effects. In: Clogg C C (ed.) *Sociological Methodology, 1989*.

American Sociological Association, Washington, DC, pp. 243–62
Firebaugh G, Chen K 1995 Vote turnout of nineteenth amendment women: The enduring effects of disfranchisement. *American Journal of Sociology* **100**(January): 972–96
Mannheim K 1952 The problem of generations. In: Kecskemeti P (ed.) *Essays on the Sociology of Knowledge*. Oxford University Press, New York, pp. 276–322
Miller W E, Shanks J M 1996 *The New American Voter*. Harvard University Press, Cambridge, MA
Rintala M 1979 *The Constitution of Silence: Essays on Generational Themes*. Greenwood, Westport, CT

P. R. Abramson

Generations, Relations Between

The concept of intergenerational relations has different meanings that are not always sufficiently distinguished. One of its aspects is the relations between age groups demarcated by demographic criteria. To the extent that these relationships are regulated by the institutions of the welfare state, one speaks of a 'generational contract.' The concept of intergenerational relations also encompasses the personal relations between the members of different generations. As a major focus of theory building and empirical research, this latter aspect—intergenerational relations within the family—is the subject of this article, which also touches on intergenerational relations with friends and acquaintances. Because the modern (or now perhaps postmodern) family is usually compared with the traditional family, a brief historical review is included as well.

1. Historical Patterns of Intergenerational Relations

Historical research shows that the rise of the bourgeois family in the late eighteenth century implied a change in family relations. This change is described in terms of growth in emotional closeness, intimacy, or affective individualization (Stone 1990). Although the finding of increasing affective attachment between parents and young children has become accepted as a matter of course, historical research has been less successful at changing popular beliefs about intergenerational relations between adult children and their parents in traditional society.

Many people, including a number of sociologists and especially politicians, cling to the 'misbelief' (Laslett 1991, p. 111) that adult children in premodern times lived with their parents under the same roof. Actually, types of families in preindustrial European society were quite diverse. The stem family, in which the members of the younger generation continued to live with their parents, was quite common in south and southeast Europe. Another type of family was the zadruga, an arrangement whereby the families of brothers lived together. By contrast, the nuclear family was (and still is) the main type in north and northwest Europe (Burguiére and Lebrun 1997). In that part of Europe, the adult children did not provide housing, financial support, and care for their old parents, though the opposite claim has often been made. The relatively few persons who did survive into old age had to work as long as they could. If they were poor, frail, or sick, they received a very low level of support from the local community. People who owned property or land or who had a business decided themselves when one of their children should succeed them. Monthly contributions were regulated by contract, with the heir having to provide the parents with either money or natural products.

The nature of intergenerational relations changed at the end of the eighteenth and beginning of the nineteenth century. The new emotional quality of the early parent–child relationship became encoded in new norms that defined the conduct of a good mother. Similarly, normative ideas began to spread about obligations that adult children had toward their parents. Until the seventeenth century, there seems to have been no legal provisions that committed children to support their old parents (Grimm 1899). However, Prussian common law (*Allgemeines Preussisches Landrecht*) of 1794, the most advanced codification of its time, ruled that children had to care for their parents in case of need (Paragraph 63). It was only after life spans (longevity) increased and after the industrial economy put old people out of work that the number of three-generation households increased (Ehmer 1982).

2. Intergenerational Relations at the Turn of the Second Millennium: Parents and Their Adult Children

Contrary to the assumptions in sociological theories of modernization, premodern norms of solidarity do not seem to have been eroded by the dynamics of modernization. The norm that adult children owe their parents support and affection has been spread by modern middle-class society. Simultaneously, however, the idea has emerged that children have a right to make decisions about their own lives. Parent–child relationships have thus come to be characterized by two sometimes incompatible dimensions: attachment and the pursuit of independence.

Additional changes in intergenerational relationships have been due to demographic factors in the twentieth century. Secular decline in the birth rate, which has been particularly apparent since the 1960s, has reduced the number of siblings, uncles, aunts, cousins, nephews, nieces (a family's 'horizontal' di-

mension) but has increased the number of generations within the family (the 'vertical' dimension of a family). In the 1990s, for example, 81 percent of all children ranging from 10 through 14 years of age in Germany had a living grandfather or grandmother (Federal Ministry 1998).

Research on intergenerational relations in late phases of the family cycle began in the 1960s, when it became apparent that the rising number of old and frequently frail people was going to confront society with a problem. In response to the structural 'isolation of the conjugal unit' (Parsons 1943, p. 27), researchers have sought evidence that relationships between old parents and their adult children are characterized by solidarity (Litwak 1965). Rosenmayr and Köckeis (1965) summarized their research results with the words 'intimacy at a distance' (p. 204), a formulation that became known worldwide and that still seems to hold for the relations between old parents and their adult children, at least in Western industrialized countries. Numerous studies have meanwhile shown that parents and children provide each other instrumental, emotional, and financial support (Attias-Donfut 1995, Schütze and Lang 1996, Silverstein and Bengtson 1997). Because intergenerational financial transfer flows from the older to the younger generation, it seems plausible that some of the public transfers to the elderly are channeled back to the young through the family (Attias-Donfut 1995, Kohli 1999).

Although researchers have particularly addressed questions about the stress and feelings of gratification experienced by middle-aged children who become caregivers for their parents, care is primarily given by spouses. Often, adult children become the caregivers only if one of the parental spouses is not available. Daughters are more likely than sons to assume this responsibility, and the daughters experience much greater stress, social isolation, and disruption in their daily lives than sons do (Horowitz 1985). The number of studies on conflict and consensus in parent–child relationships, particularly on caregiving relationships, is probably due to political concerns about whether the family will continue to perform this function or whether it will fall to state institutions. The focus on sociopolitical issues has led to a relative neglect of theoretical questions, such as the transformation of the parent–child relationship during the late phases of the family cycle.

3. *Theoretical Perspectives on Intergenerational Relations*

Erikson (1959) occasionally referred to the developmental task of overcoming dependency stemming from the period of childhood. In a similar vein Blenkner (1965) developed the construct of filial maturity, arguing that a transformation occurs in the late phases of the family cycle irrespective of the quality of the early parent–child relationship. When the children cannot help but realize that their old parents are becoming frail, they experience a filial crisis because they recognize that the parents are having to rely completely on themselves. In the subsequent phase of filial maturity, the children acquire a new perspective, conceiving of their parents as individuals with their own rights, needs, and limitations. With this new perspective, there emerges a sense of responsibility that did not exist before. The concept of filial maturity may well be a way in which to integrate autonomy and attachment, moral obligation and affection, distance and closeness.

Taking Blenkner's concept of filial maturity as a model, Nydegger (1991) developed the concept of parental maturity. Parental maturity means that parents cease interference in their children's lives. Mature parents successfully cope with the fact that some hopes they had had for their children have not been met. They recognize whatever differences there may be between their values and those of their children. Mature parents also have learned to accept help and do not claim a degree of independence that is no longer realistic. Unfortunately, neither the concept of filial maturity nor that of parental maturity has been empirically researched (for an exception, see Marcoen 1995).

Another theoretical framework for the analysis of late parent–child relationships can be derived from the view of generations offered by Mannheim ([1928] 1952). Mannheim held that the members of different generations (cohorts) are located at different places in the historical process and that they therefore differ in the way they experience and cope with historical events. Mannheim applied this concept to classifications of age groups in the population, though not to family lineage. It seems legitimate, however, to ask how location in the historical process affects relations and mutual perceptions in the family. In a cross-cultural comparison, for instance, Attias-Donfut (1998) found much less convergence between value orientations and much greater emotional distance between old and younger generations (age groups) in Germany than in France. The author suggested that this generation gap in Germany may be a consequence of the Nazi period. The findings are consistent with those from the literature on family therapy, in which it is shown that Germany's Nazi past weighs heavily on relationships within the family as well (Stierlin 1981, Rosenthal 1997).

4. *Intergenerational Relations beyond the Family*

In both sociological theory and popular understanding, it is assumed that people choose friends on the basis of similarity, for friends mostly have similar social status, converge in their lifestyles, and belong to

the same sex and age group (Merton and Lazarsfeld 1954, Rosow 1967; for more recent literature see Filipp and Mayer 1999). Homogeneity of age, however, is viewed positively in relations between children or adolescents but negatively (as an indicator of discrimination) in relations that old people have. In reality, it is questionable whether the low number of friendships between older and younger people is evidence of discrimination. After all, friendship is usually between members of the same age group.

All in all, it is primarily through the family that the members of two, three, or even four generations maintain personal relations. Within these relations cultural traditions and family-specific experiences and narratives are passed on. At the same time, however, new forms of dealing with each other are created as well. There is no evidence to support suspicions of cultural pessimists that this balance between continuity and change is going to collapse.

See also: Age, Sociology of; Age Stratification; Age Structure; Cohort Analysis; Families as Educational Settings; Family as Institution; Family Processes; Generation in Anthropology; Generations in History; Generations, Sociology of; Kinship in Anthropology; Life Course: Sociological Aspects; Mannheim, Karl (1893–1947); Population Aging: Economic and Social Consequences; Tribe

Bibliography

Attias-Donfut C (ed.) 1995 *Les solidarités entre générations. Vieillesse, famille, etat*. Nathan, Paris

Attias-Donfut C 1998 Generationenverhältnis und sozialer Wandel. In: Köcher R, Schild J (eds.) *Wertewandel in Deutschland und Frankreich*. Leske & Budrich, Opladen, Germany, pp. 173–205

Blenkner M 1965 Social work and family relationships in later life with some thoughts on filial maturity. In: Shanas E, Streib G F (eds.) *Social Structure and the Family: Generational Relations*. Prentice-Hall, Englewood Cliffs, NJ, pp. 46–59

Burguière A, Lebrun F 1997 Die Vielfalt der Familienmodelle in Europa. In: Burguière A, Klapisch-Zuber C, Segalen M, Zonabend F (eds.) *Geschichte der Familie*. Campus Verlag, Frankfurt am Main, Germany, pp. 13–118

Ehmer J 1982 Zur Stellung alter Menschen in Haushalt und Familie. Thesen auf der Grundlage von quantitativen Quellen aus europäischen Städten seit dem 17. Jahrhundert. In: Konrad H (ed.) *Der alte Mensch in der Geschichte*. Verlag für Gesellschaftskritik, Vienna, pp. 62–103

Erikson E H 1959 *Identity and the Life Cycle*. International Universities Press, New York

Federal Ministry for Family, Senior Citizens, Women, and Youth 1998 *Zehnter Kinder- und Jugendbericht. Bericht über die Lebenssituation von Kindern und die Leistungen der Kinderhilfen in Deutschland*. Federal Ministry for Family, Senior Citizens, Women, and Youth, Bonn, Germany

Filipp S-H, Mayer A-K 1999 *Bilder des Alters. Alterssstereotype und die Beziehungen zwischen den Generationen*. Kohlhammer, Stuttgart, Germany

Grimm J 1899 *Deutsche Rechtsalterthümer*, 4th edn. Dietrich'sche Verlagsbuchhandlung, Leipzig, Germany

Horowitz A 1985 Sons and daughters as caregivers to older parents: Differences in role performance and consequences. *The Gerontologist* **25**: 612–17

Kohli M 1999 Private and public transfers between generations: Linking the family and the state. *European Societies* **1**: 81–104

Laslett P 1991 *A Fresh Map of Life: The Emergence of the Third Age*. Harvard University Press, Cambridge, MA

Litwak E 1965 Extended kin relations in an industrial democratic society. In: Shanas E, Streib G F (eds.) *Social Structure and the Family: Generational Relations*. Prentice-Hall, Englewood Cliffs, NJ, pp. 290–323

Mannheim K [1928] 1952 The problem of generations. In: Kecskemeti P (ed.) *Essays on the Sociology of Knowledge*. Routledge and Kegan Paul, London, pp. 276–320

Marcoen A 1995 Filial maturity of middle-aged adult children in the context of parent care: Model and measures. *Journal of Adult Development* **2**: 125–35

Merton R, Lazarsfeld P 1954 Friendship as a social process. In: Berger M, Abel T, Page C (eds.) *Freedom and Control in Modern Society*. Van Nostrand, Princeton, NJ, pp. 18–66

Nydegger C N 1991 The Development of paternal and filial maturity. In: Pillemer K, McCartney K (eds.) *Parent–Child Relations throughout Life*. Erlbaum, Hillsdale, NJ, pp. 93–111

Parsons T 1943 The kinship system of the contemporary United States. *American Anthropologist* **45**: 22–38

Rosenmayr L, Köckeis E 1965 *Umwelt und Familie alter Menschen*. Luchterhand, Neuwied, Germany

Rosenthal G (ed.) 1997 *Der Holocaust im Leben von drei Generationen. Familien von Überlebenden der Shoah und von Nazi-Tätern*. Psychosozial-Verlag, Giessen, Germany

Rosow J 1967 *Social Integration of the Aged*. Free Press, New York

Schütze Y, Lang F R 1996 Integration in family, kinship and friendship networks. In: Mollenkopf H (ed.) *Elderly People in Industrialised Societies: Social Integration in Old Age By or Despite Technology?* Edition Sigma, Berlin, pp. 25–40

Silverstein M, Bengtson V L 1997 Intergenerational solidarity and the structure of adult child–parent relationships. *American Journal of Sociology* **103**: 429–60

Stierlin H 1981 The parent's Nazi past and the dialogue between the generations. *Family Process* **26**: 379–90

Stone L 1990 *The Family, Sex and Marriage in England 1500–1800*. Penguin Books, London

Y. Schütze

Generations, Sociology of

Generations constitute—just as the closely related categories of age, youth, adulthood and old age—seemingly universal categories in all human societies signifying the different stages of human life; the same is true of tensions between generations, as well as of concepts of the age group or set—in particular the youth group—as a special social category with its own distinct self-consciousness. But it is only in very specific civilizational and historical circumstances that these features come together. Only in special circumstances does historical change become articulated in terms of generational consciousness in general, and that of youth in particular.

It was Karl Mannheim—as well as J. Marias and J. Ortega y Gassett—who emphasized the apparent universality of generational consciousness or at least posed it as a major problem of research; but this tendency, at least as reported by them, was indeed very closely connected to rather specific civilizational traditions—the European ones—and to specific historical circumstances in the development of this civilization. Accordingly a central problem in the analysis of generations is under what conditions does historical change become articulated in terms of generational consciousness in general and that of youth in particular?

1. *The Cultural Definitions of Age and of Youth*

Generational consciousness is naturally connected with that of age and age differences, and therefore it might first of all be worthwhile to specify the general conditions under which specific age groups and generational tensions and consciousness develop in human societies.

Age differences in general and youth in particular constitute a universal phenomenon. They are first of all a biological phenomenon. Although the basic biological processes of maturation (within the limits set by such factors as relative longevity) are probably more or less similar in all human societies, the definition of age differences in general and of youth in particular is always cultural. Their cultural definition varies from society to society, at least in its details.

The cultural definitions of age and age differences contain several different yet complementary elements. First, these definitions often refer to the social division of labor in a society, to one of the criteria according to which people are assigned to various social positions and roles; for instance, in many societies certain roles—especially those of married men, full citizens, independent earners—are barred to young people, while others—such as certain military roles—are specifically allocated to them. Second, the cultural definition of age is one important constituent of a person's self-identity, his self-perception in terms of his own psychological needs and aspirations, his or her place in society, and the ultimate meaning of his life.

Within any such definition, the qualities of each age are evaluated according to their relation to some basic, primordial qualities, such as vigor, physical and sexual prowess, the ability to cope with material, social and supernatural environment, wisdom, experience, or divine inspiration. Different ages are seen in different societies as the embodiments of such qualities. These various qualities seem to unfold from one age to another, each age emphasizing some out of the whole panorama of such possible qualities.

The various qualities attributed to different ages are usually interconnected in many ways, constituting a series of sorts. The characteristics of any one 'age,' therefore, cannot be fully understood except in relation to those of other ages. Only when taken together do these different 'ages' constitute the entire map of human possibilities and limitations as envisaged in different societies; and, as every individual usually must pass through them all, their complementariness and continuity (even if defined in discontinuous and contrasting terms) become strongly emphasized and articulated. The same holds true for the age definitions of the two sexes, although perhaps with a somewhat different meaning. Each age-span is often defined differently for either sex, and these definitions are usually related and complementary, as the 'sexual image' and identity always constitute basic elements of man's image in every society.

This close connection between different ages necessarily stresses the problem of transition from one point or stage in a person's life to another as a basic constituent of any cultural definition of an 'age.' The personal transition or temporal progress, or change, tends to become closely linked with the designation or perception of cosmic and societal time—with the rhythms of nature or history, the cycles of the seasons, the unfolding of some cosmic plan (whether cyclical, seasonal, or apocalyptic), or the destiny and development of society. The nature of this linkage often constitutes the focus round which an individual's personal identity becomes defined in cultural terms and through which personal experience, with its anguish, may be given some meaning in terms of cultural symbols and values.

2. *Youth as a Liminal, Transitory Phase; Kernels of Generational Consciousness*

The problems of age definition and the linkage of personal time and transition with cosmic time become very often accentuated in that age-span usually designated as youth—which constitute in most societies the most liminal of age definitions—the transition from childhood to full adult status, to full membership in the society. In this period the individual is no longer a child (especially from the physical and sexual point of view) and is ready to undertake many attributes of an adult and to fulfil adult roles. But he is not yet fully acknowledged as an adult, a full member of the society. Rather, he or she is being, as it were, 'prepared,' or is preparing himself for such adulthood.

This image of youth—the cultural definition of youth—contains all the crucial elements of any definition of age, usually in a specially articulated way, and often in confrontation with other age categories. Moreover, in this phase the problem of the linkage of the personal temporal transition with cosmic or societal time becomes extremely acute.

The fullest, the most articulate and definitive expression of these archetypal elements of youth and of their liminal potentials is best exemplified in the ritual

dramatization of the transition from adolescence to adulthood, such as the various *rites de passage* and ceremonies of initiation in primitive tribes and in ancient civilizations. In an even more diluted form, these elements may be found in the various spontaneous initiation ceremonies of the fraternities and youth groups in modern societies (Bettelheim 1954, Erikson 1963, pp. 54–93). Hence the youth period and its symbolism, and the liminal situations in which it becomes expressed, may become the focus both of continuity of the social order as well as of protest against it, and of its change.

The very emphasis on the transitory nature of this stage and on its essentially preparatory character, however, often create a somewhat paradoxical situation—namely an image of youth as the purest manifestation and repository of ultimate cultural and societal values of all the major human virtues and primordial qualities. Such an image may be reinforced by the identification of physical vigor in youth with a more general flowering of the cosmos or the society. It is this situation that contains the kernels of the development of strong generational consciousness especially around the image of youth, with strong potentialities of protest.

The social organization of any age and especially of the period of youth is constructed in all societies according to two major criteria. One is the extent to which age in general and youth in particular form a criterion for the allocation of roles in a society, whether in politics or in economic or cultural activity—aside from the family, of course, in which they always serve as such a criterion. The second is the extent to which within society there develop specific age groups, specific corporate organizations, composed of members of the same 'age,' such as youth movements or old men's clubs.

3. *The Social and Cultural Contexts of Different Patterns of Age and Generational Symbolism and Organization*

The importance of age as a criterion for allocating roles in a society is closely related to several major aspects of social organization and cultural orientation. The first aspect is the relative complexity of the division of labor. In general, the simpler the organization of the society, the more influential age will be as a criterion for allocating roles. Therefore, in parallel tribal or preliterary traditional societies, age and seniority constitute basic criteria for allocating social, economic, and political roles.

The second such aspect is that the major value orientations and symbols of a society, especially the extent to which they emphasize certain general qualities, or types of activity (such as physical vigor, the maintenance of cultural tradition, the achievement and maintenance of supernatural prowess) become expressed and symbolized in specific age-spans. Thus for instance, in those societies in which military values and orientations prevail, young adulthood is emphasized as a very important—sometimes even as the most important—age, while those in which sedentary activities prevail emphasize older age. Similarly, within some traditional societies, where older people may serve as important 'storers' of collective memory, old age may be emphasized if it is seen as the most appropriate one for expressing major cultural values and symbols—for instance, the very upholding, symbolization, and perpetuation of a given cultural tradition.

The social and cultural conditions that determine the extent to which specific age groups and youth groups develop differ from the conditions that determine the extent to which age serves as a criterion for the allocation of roles. At the same time, the two kinds of conditions may be, in some situations or societies, closely related. Age groups in general and youth groups in particular tend to arise in those societies in which the family or kinship unit cannot ensure (it may even impede) the attainment of full social and political status on the part of its members. These conditions appear especially, although not uniquely, in societies in which family or kinship groups do not constitute the basic units of the social division of labor. Membership in the total society (citizenship) is not defined in terms of belonging to any such family, kinship group, or estate, nor is it mediated by such a group.

The emergence of distinct age and generational organization and consciousness, as Meyer Fortes has shown in one of his last papers, lies in the political realm, in the search and definition of citizenship beyond the family sphere and not through the family, while generational differences are mostly elaborated in the structure of the family and kinship unit.

4. *Age Organization and Generational Consciousness in Tribal, Archaic, and Historical Societies*

The type of social division of labor and of political and cultural symbolism that generates such conditions is found in varying degrees in different societies—primitive, historical, or modern. Such a division of labor has existed, for example, in several primitive tribes: in Africa, among the chiefless (segmentary) tribes of Nandi, Masai, or Kipigis; in the village communities of Yako and Ibo; in the more centralized kingdoms of the Zulu and Swazi; and among some of the Indian tribes of the Plains, as well as among some South American and Indian tribes. Such division of labor likewise existed to some extent in several historical societies (especially in city states such as Athens or Rome), although most great historical civilizations were characterized mainly by a more hierarchical and ascriptive system of the division of

labor, in which there were greater continuity and harmony between the family and kinship groups, and the broader institutional contexts.

The fullest development of this type of social division of labor is to be found in modern, especially industrial, societies in which inclusive membership is usually based on universal criteria. In these societies the family does not constitute a basic unit of division of labor, especially not in production and distribution, and even in the sphere of consumption its functions become more limited; occupations are not transmitted through heredity. Similarly, the family or kinship group does not constitute a basic unit of political or ritual activities. Moreover, the general scope of the activities of the family has been continuously diminishing, while various specialized agencies tend to take over its functions in the fields of education and recreation.

At the same time in most modern—as well as in some historical—societies, there developed a certain dissociation between organizations of groups based on age and the symbolic distinction between different age categories. It is such dissociation that constitutes the background for the potentialities for the crystallization of specific generational consciousness formed around the image of youth and in which there develops a strong articulation of the dialectics between, on the one hand, the 'rebellious' protest-oriented aspects of youth organization and generational consciousness, and, on the other hand, continuity-oriented aspects (Bettelheim 1954, Erikson 1963, pp. 54–93).

5. *Generational and Youth Symbolism and Consciousness in Modern Societies*

5.1 *Generational Consciousness in Modern Society*

The generational consciousness, about which Mannheim, Marias, and Ortega y Gassett and many other modern writers wrote, has developed against the background analyzed above. Such consciousness is indeed connected, in a rather paradoxical way, with the fact that in modern societies the development of specific youth organizations is linked with the weakening of the importance of age in general and youth in particular as definite criteria for the allocation of roles in society become weakened. Yet this very weakening of the importance of age, as a criterion of allocation of roles, has been connected with intensive development of youth groups and movements—and lately also, perhaps paradoxically, among 'retired' people. Indeed, it is very interesting to note that in some of the contemporary industrial societies there tends to develop a distinct organization based on age as well as generational consciousness that are again coming together at the other end of the life-span—in old age—among the older people, that is, those who have retired from work.

It is because of these developments—some incipient phases of which can be also discerned in different archaic and historical societies, but which have not been studied adequately (Nash 1978)—that in modern societies there also arises the possibility of the development of very strong generational consciousness in general, and of the articulation of such consciousness in terms of youth in particular. It denotes the general development of the self-consciousness of generations, which develops in the public realm—whether the realm of politics or of culture—and it contains very strong potentialities of social, political, or cultural protest.

6. *Historical Change, Revolutionary Orientations, Generational Consciousness in Modern Youth Movements*

The development in modern societies of such a dissociation between generations, and the concomitant possibility of the development of a special type of a distinct generational consciousness, are due not only to the complex division of labor that characterizes these societies. They are also closely connected to the relatively quick tempo of historical change, and above all to the consciousness of such change. Such consciousness is often connected with changes rooted in the revolutionary orientations which constituted important components of the construction of modern civilization, cultural values, and orientations. These revolutionary orientations were most fully articulated by the major social movements that have developed as part and parcel of modern life, of the constructions of modern social and political order and of struggle around such construction. The possibility of the development of distinct generational consciousness has been enhanced in modern societies by the occurrence of rapid historical changes or ruptures, the consciousness of which is shared by members of the same age group or age cohort.

But while historical change is particularly rapid and the consciousness of such change very widespread in modern societies, it does not always give rise to fully articulated generational consciousness—although to some degree such consciousness is probably rather common in modern societies and is not necessarily confined to the symbolism and ideology of youth. It is only in rather special historical circumstances and situations—like wars, changes of regime, and the like, especially of events which entail strong ruptures—that such generational consciousness has become articulated in modern societies in terms of youth symbolism.

Among such situations in modern societies two are probably of special importance. One was developed at the end of the nineteenth and the beginning of the twentieth century, and found its fullest expression in

the free German youth movements, though it was also present in the less spectacular youth movements of Central Europe and to some extent in the youth culture of various more flexible youth groups that developed in most modern societies. In all these movements very strong attempts have been made to overcome the dislocation between personal transition and societal and cultural time. It was in them that dreams of a new life, a new society, freedom and spontaneity, a new humanity, and aspirations to social and cultural change found utterance. It was in these youth movements that the forging of youth's new social identity became closely connected with the development of new symbols of collective identity or new social-cultural symbols and meanings, which thus become foci of protest aimed at the reconstruction of society and culture.

While some of these orientations could be found in many of the youth groups that developed in most modern societies—be they in England or the United States—they developed most fully in situations where these groups became closely interconnected with major social movements oriented at the reconstruction of the centers and symbols of their respective societies. This was the case above all in the major nationalistic and revolutionary movements in Central and Eastern Europe, in Italy, Germany, Hungary, and Russia, or in the Zionist movement, where these orientations and student movements played a crucial role.

One of the major characteristics of many, probably most, of these movements was that in their ideology the symbolism of youth was very closely connected with that of the construction of new modern, especially national, collectivities that were being constructed at the period, and among the aims of these youth movements the reconstruction of such collectivities—national, social, or cultural—was central. These movements aimed at changing many aspects of the social and cultural life of their respective societies, to reconstruct the centers and premises of their respective societies in terms of some of the major revolutionary orientations of modernity. They depicted the present in a rather shabby light; they censured it for its materialism, exploitation, restrictive views, and lack of opportunity for self-fulfilment and creativity. At the same time they held out hope for the future—seemingly the not very far-off future—when both self-fulfilment and collective fulfilment could be achieved and the materialistic civilization of the adult world shaken off. They tried to appeal to youth to forge its own self-identity in terms of these new collective symbols, and this is why they have been so attractive to youth, for whom they provided a set of symbols, hopes, and aims to which to direct its activities.

The influence of such movements and orientations was very great, not only political and educational but also in many cultural arenas. In a way, education itself has tended to become a social movement. Many schools and universities, and many teachers, have been among the most important bearers of collective values. The expansion of education itself is often seen as a means by which a new epoch might be ushered in.

7. *The Routinization of Revolutionary Movements; the Demystification of the World; and Youth and Student Rebellion of the 1960s*

The second historical situation in modern societies in which generational consciousness, couched in the symbolism of youth, became very prominent in the new social movements burgeoned in the late 1960s and early 1970s, and became most fully manifest in the various student rebellions and movements of the time.

The background to these new developments was a combination of the growing institutionalization of many of the collective values and symbols of the social, national, and youth movements discussed above, and the continuous expansion of education and changes in occupational structure. These developments had several important repercussions on the structuration of the life spans of youth.

The area of social life that the specific youth or student culture encompasses has tended to expand continuously. First, it has extended over a long period of life, reaching, through the impact of the extension of higher education, to what before was seen as early adulthood. Second, it tends more and more to include areas of work, leisure time activity, and many interpersonal relations. Third, the potential and actual autonomy of these groups, and the possibility their members have for direct access to work, marriage, and family life, to political rights, and to consumption, have greatly increased, while their dependence on adults has greatly decreased.

Paradoxically enough, the growing direct access of young people to various areas of life has given rise to a growing insecurity of status and self-identity and to a growing ambiguity of adult roles. This insecurity and ambiguity has been enhanced, first, by the prolongation of the time-span between biological and social maturity and by the extension of the number of years spent in 'preparatory' (educational) institutions, and, second, by the growing dissociation between the values of these institutions and the future—especially in terms of occupational and parental roles—of those participating in them. A third factor is that for a long period many 'young' people may as yet have no clear occupational roles or responsibilities and may be dependent on their parents or public institutions for their economic needs, while at the same time they constitute an important economic force as consumers and certainly exercise political rights. In turn this situation may become intensified or aggravated by the growing demographic preponderance of the 'young' in the whole population and by the increasing possibilities of ecological mobility.

All these developments gave rise to new types of generational and youth consciousness—a new combination between the symbolism of youth and collective symbolism. The older classical movements of protest and youth movements tended to assume that the framework and centers of the nation-state constituted the major cultural and social reference points of personal identity and of the charismatic orientations to some sociocultural orders, and that the major task before them was to facilitate the access of broader strata to these centers and full participation in them and possible changes in them. In contrast, the new movements of protest have been characterized by their scepticism towards these new modern centers and by lack of commitments to them—all of which denoted important aspects of what has been called by Weber the demystification of the world.

The foci of protest promulgated by these movements shifted from demands for greater participation in national political centers or from attempts to influence their socioeconomic politics to new directions. The most important of these directions were: first, attempts to 'disrobe' these centers of their charismatic legitimacy and perhaps of any legitimacy at all; second, continuous searches for new loci of meaningful participation beyond these centers and the concomitant attempts to create new centers which would be independent of the existing center; third, attempts to formulate the patterns of participation in their centers not so much in sociopolitical or economic terms but more through symbols of primordial or of direct social participation.

These developments have often been connected with the breakdown of consciousness of historical continuity and of the conception of the future in terms of such continuity. The image of the societal future has become flattened and has been deprived of its allure. No sharp ideological discontinuity has been conceived between present and future.

Through these developments, youth has been, as it were, robbed of the full experience of the dramatic transition from adolescence to adulthood and of the dramatization of the difference between present and future. Their own changing personal future has become dissociated from any changes in their societies or in cultural activities and values. As a result of all these processes, the possibility of linking personal transition to societal and cosmic time—which was so strongly emphasized in the youth movements and observable also to some extent even in the earlier, looser youth culture—has become greatly weakened.

One of the most important results of these developments has been the development of a new configuration of youth symbolism and generational consciousness—the major characteristic of which was the development of strong antinomian intellectual tendencies as a central component of such symbolism and consciousness.

It is not just the contents of these antinomian tendencies that are new, but the convergence of these contents with the spread and location of these movements. Of special importance in this context is the fact that this type of protest was not confined only to small, closed intellectual groups but involved widespread circles of novices and aspirants to intellectual status. Given the expansion of the modern educational system and the parallel effects of the spread of mass-communication media, a very large part of the educated public have become aspirant to such status, and impinge on the centers of intellectual creativity and cultural transmission and become integral, even if transient, parts thereof.

This rather unique combination of intellectual antinomianism with youth rebellion gave rise to a new type of youth ideology that was connected to the reversal of the hitherto existing relation between the definition of different age-spans and the possibilities of social and cultural creativity. Unlike in the past—even the not-so-distant past—youth became more and more seen not only as a preparation for independent and creative participation in social and cultural life, but as the very embodiment of permissive, often unstructured creativity—to be faced later on with the constants of a relatively highly organized, constrictive, meritocratic, and bureaucratic environment. Thus youth now became, even more than before, the transmitters and symbols of a continuous antinomian confrontation between the pristine values it embodies and a demystified social reality, and the impact of this new type of generational consciousness has been of crucial importance in the crystallization of what has been designated as the postmodern trends in contemporary societies. While the intensity and the concrete manifestation of this consciousness will naturally change from one situation to another, the consciousness itself with the various antinomian tendencies analyzed above seems to have become more and more prevalent.

See also: Age: Anthropological Aspects; Age, Sociology of; Age Stratification; Age Structure; Cohort Analysis; Generation in Anthropology; Generations in History; Generations: Political; Generations, Relations Between; Life Course in History; Life Course: Sociological Aspects; Mannheim, Karl (1893–1947); Structure: Social; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements

Bibliography

Aron R 1969 *The Elusive Revolution: Anatomy of a Student Revolt*. Pall Mall Press, London

Bauer R A, Inkeles A, Kulckhohn C 1956 Generational differences. In: R A Bauer et al. (eds.) *How the Soviet System Works: Cultural, Psychological and Social Themes*. Harvard University Press, Cambridge, MA, pp. 190–8

Bettelheim B 1954 *Symbolic Wounds, Puberty Rites and the Envious Circle*. Free Press, Glencoe, IL

Bettelheim B 1963 The problem of generations. In: Erikson E H (ed.) *Youth: Change and Challenge*. Basic Books, New York, pp. 54–93
Bourges H 1968 *The French Student Revolt: The Leaders Speak*. Hill and Wang, New York
Eisenstadt S N 1956 *From Generation to Generation: Age Groups and Social Structure*. Free Press, Glencoe, IL
Eisenstadt S N 1963 Archetypal patterns of youth. In: Erikson E H (ed.) *Youth: Change and Challenge*. Basic Books, New York, pp. 24–43
Erikson E H (ed.) 1963 *Youth: Change and Challenge*. Basic Books, New York
Fortes M 1984 Age, generation and social structure. In: Kertzer D I, Keith J (eds.) *Age and Anthropological Theory*. Cornell University Press, Ithaca, NY, pp. 99–123
Glazer N 1970 *Remembering the Answers: Essays on the American Student Revolt*. Basic Books, New York
Gusfield J 1957 The problem of generations in an organizational structure. *Social Forces* **35**: 323–30
Heberle R 1951 *Social Movements: An Introduction to Political Sociology*. Appleton Century-Crofts, New York [See especially 'The Problem of Political Generations,' pp. 118–27]
Lipset S M, Altbach P G (eds.) 1969 *Students in Revolt*. Houghton Miffin, Boston
Lipset S M, Wellin S S (eds.) 1965 *The Berkeley Student Revolt. Facts and Interpretations*. Doubleday, Garden City, NY
Mannheim K 1952 The problem of generations. In: Kecskemeti P (ed.) *Essays on the Sociology of Knowledge*. Oxford University Press, New York, pp. 276–320
Marias J 1949 *El Metodo Historico de las Generaciones*. Revista de Occidente, Madrid, Spain
Nash L L 1978 Concepts of experience, Greek origins of generational thought. *Daedalus* Fall: 1–23
Ortega y Gasset 1933 *The Modern Theme*. Norton, New York
Pinder W 1926 *Kunstgeschichte nach Generationen*. Pfeiffer, Leipzig, Germany
Preston T R 1927 *My Generation of Politics, and Politicians*. Rose, Toronto
Rintala M 1963 A generation in politics: A definition. *Review of Politics* **25**: 509–22
Zeitlin M 1966 Political generations in the Cuban working class. *American Journal of Sociology* **71**: 493–508

S. N. Eisenstadt

Generative Grammar

Generative grammar is the label of the most influential research program in linguistics and related fields in the second half of the twentieth century. Initiated by a short book, Chomsky's *Syntactic Structures* (1957), it became one of the driving forces among the disciplines jointly called the cognitive sciences. The term *generative grammar* refers to an explicit, formal characterization of the (largely implicit) knowledge determining the formal aspect of all kinds of language behavior. From the start the program had a strong mentalist orientation instance, documented, for, in a fundamental critique of Skinner's *Verbal Behavior* (1957) by Chomsky (1959), arguing that behaviorist stimulus–response theories could in no way account for the complexities of ordinary language use. The 'Generative Enterprise,' as the program was called in 1982, went through a number of stages, each of which was accompanied by discussions of specific problems and consequences within the narrower domain of linguistics as well as the wider range of related fields such as ontogenetic development, psychology of language use, or biological evolution. Four stages of the Generative Enterprise can be marked off for expository purposes.

1. *Transformational Structure and Levels of Description*

The foundation of all further developments is to be found in the Chomsky's monograph *The Logical Structure of Linguistic Theory* (1975, henceforth LSLT), which was published only 20 years after the manuscript was completed. Starting from the notion of structural analysis as developed by Bloomfield and made precise by Harris (1951), LSLT introduces a radically new perspective on the nature of linguistic structure. On the one hand, LSLT takes up Harris's notion 'level of linguistic structure,' in terms of which the form of utterances, i.e., their phonetic and syntactic shape, is characterized. On the other hand, however, LSLT develops a notion of linguistic structure that differs fundamentally from the empiricist orientation of American structuralism. More specifically, Chomsky proposes a notion of grammar that provides a full and explicit account of the tacit knowledge a speaker/hearer has of the form of utterances of his or her language. Under this perspective, a grammar G is a theoretical construct, dealing with internal states of knowledge that relate only indirectly to overt behavior of the speaker. In other words, G is viewed as an empirical hypothesis about the mental structure underlying ordinary speech behavior. To make this notion precise, LSLT highlights the importance of intuitive judgments a speaker is able to make about properties of utterances, such as the grammatical well-formedness of the famous nonsense-sentence (1) as opposed to the ungrammaticalness of (2), or the rather different properties of (3) and (4), inspite of their superficial similarity:

(1) colorless green ideas sleep furiously
(2) furiously sleep ideas green colorless
(3) this picture was painted by a real artist
(4) this picture was painted by a new technique

Hence G of English has to account for the properties on which intuitions like these rely. The study of the general features of G can furthermore be assumed to reveal principles of the internal organization of mental states in general and of knowledge of language in particular. This is in fact the basic motivation underlying the research program of Generative Grammar.

More technically, LSLT gives the following account of the concepts needed for this program. First of all, a linguistic level **L** is defined as an algebraic structure of the following sort:

(5) $\mathbf{L} = [L, R_1,\ldots, R_m, \mu, \Phi, \varphi_1,\ldots, \varphi_n]$, where
L is the set of primes of **L**;
R_1 to R_m are sets of relations in **L**, such as linear ordering, inclusion, identity;
μ is a set of so-called **L**-markers, generated by the primes of **L**;
Φ is a mapping that maps μ onto the set of grammatical utterances;
φ_1 to φ_n are operations relating **L** to other levels.

Roughly speaking, an **L**-marker is the representation of the structure of an utterance on the level **L**. The particular levels assumed in LSLT include the phonetic level **Pt**, the phonemic level **Pm**, the morpheme level **M**, the word level **W**, the level of word classes **C**, the phrase structure level **P**, and the transformational level **T**. For **Pm**, the set μ consists of strings of phonemes, for **W**, the elements of μ are sequences of words, and for **P**, the set μ consists of phrase markers, roughly tree structures indicating the constituency of an utterance. The relation between the different levels defined by the operations φ_i depends on particular conditions of the various levels and is, as Chomsky argues, anything but trivial. The full structural representation of an utterance comprises the **L**-markers assigned to it at the various levels in question. The mapping Φ relating the levels of representation to the actual (or possible) grammatical utterances is correspondingly indirect and complex. The assumption in LSLT is that it may be sufficient to define Φ for phonetic representations, i.e., strings in **Pt**, provided that the relation to **Pt** is defined for all other levels. This accounts for the fact that the structure of an utterance cannot be arrived at by simple operations of segmentation and classification, as assumed in classical structuralism, but depends on highly indirect and complex relations.

Second, a grammar G is a system of operations or rules that generate the sentences of a language with their structural representations. Technically, G determines the set of **L**-markers for each level **L** and the operations by which they are related to the other levels. LSLT deals with the general form of phonological rules, defining the relation between the phonemic and phonetic level, and more extensively with phrase structure rules, generating P-markers, and with grammatical transformations. The role of transformations is a matter of central interest in LSLT (such that for quite a while the term 'transformational grammar' was taken as largely equivalent to 'generative grammar'), they are assumed to capture structural properties P-markers could not account for. For instance, the difference between (3) and (4), which are not distinct in terms of constituency, is explained by different transformational derivations, informally indicated in (3′) and (4′):

(3′) a real artist painted this picture $= = \mathrm{T}_{\mathrm{Passive}}$ $\Rightarrow$ this picture was painted by a real artist

(4′) someone painted this picture by a new technique $= = \mathrm{T}_{\mathrm{Passive}} \Rightarrow$ this picture was painted by a new technique by someone $= = \mathrm{T}_{\mathrm{Agent\text{-}deletion}} \Rightarrow$ this picture was painted by a new technique

An important point of the theory developed in LSLT is the assumption that grammars must sysbe evaluated tematically. To this effect, a simplicity-metric is introduced by which a grammar is preferred if and only if it expresses more natural generalizations than its competitors. For example, a grammar that accounts systematically for the different properties of sentences like (3) and (4) is to be preferred over one that can only assign to them the same structure. This is the reason why a grammar with a transformational component comes out as simpler than one without it. One of the factors entering the evaluation is the principle of rule ordering, which recognizes the possibility that rules might have a different structural effect, if they operate in different ordering. As a simple case in point, in (4′) the Agent-deletion can operate only after the Passive Transformation has applied. (The importance of rule ordering had already been observed in the Sanskrit grammar of Panini.)

The productivity of the proposed framework is demonstrated by an analysis of large sections of English with respect to phrase structure and transformations in LSLT and in a stimulating monograph on nominalization by Lees (1960). A seminal study of phonology is given in Halle (1959) for Russian and in Chomsky's unpublished master's thesis on the *Morphophonemics of Modern Hebrew*.

A feature that LSLT inherits from the notion of transformational analysis proposed by Harris (1957) is the notion that transformations not only relate individual sentences to their more elementary basis, as shown in (3′) and (4′), but are also used to decompose complex sentences into their constituent parts made up. Thus (6c), for instance, is composed by embedding of (6a) into (6b), replacing the bracketed pronoun:

(6) (a) this picture was painted by a new technique
(b) the director suspected [it]
(c) $= = \mathrm{T}_{\mathrm{Embedding}} \Rightarrow$ the director suspected [this picture to be painted by a new technique]

This assumption requires a distinction between simple transformations like $\mathrm{T}_{\mathrm{Passive}}$ and so-called generalized transformations like $\mathrm{T}_{\mathrm{Embedding}}$. Another distinction that must be made is that between optional and obligatory transformations. Passivization, for instance, is optional, while the fronting of relative and wh-pronouns is obligatory, as the ungrammaticalness of (7b) shows:

(7) (a) I know the man who you met last night
(b) *I know the man you met who last night

The clarification of these and a wide range of other complexities concerning the level **T** is one of the driving forces in further developments of Generative Grammar.

Another, more general feature that the Generative Enterprise inherits from American Structuralism is the assumption that all grammatical structure, from the most concrete level **Pt** of phonetics to the most abstract level **T** of transformational structure, is a matter of linguistic form which must be embedded in a wider context of a theory of semiotics that accounts for sense, reference, and use of linguistic expressions. Although there have been several proposals to include a level of semantics into the framework of generative grammar, the main emphasis has been on the knowledge of linguistic form throughout.

2. *The Standard Theory of Generative Grammer*

Immediately after the publication of *Syntactic Structures* (actually lecture notes excerpted from LSLT), the theory it proposed attracted interest from various sides, including mathematical linguistics and psychology. An overview of the formal properties of grammatical rules, especially of phrase structure grammars, in Chomsky (1963) established what became known as the Chomsky hierarchy of formal grammars as a branch of subrecursive function theory. Miller and Chomsky (1963) gave a first outline of the role of Generative Grammar for the study of language comprehension and production. Furthermore, Katz and Fodor (1963) made a first proposal to extend the theory of grammar by a semantic component, using so-called projection rules to integrate compositionally the meaning of lexical items according to their syntactic relations.

The main development, however, was due to continuous efforts to constrain systematically and strengthen the theory, increasing its explanatory power by reducing the range of possible grammars compatible with a given set of data. Such a reduction reduces at once the task assigned to the evaluation metric. Efforts to this effect led to what was later called the 'Standard Theory' of Generative Grammar. Based on preparatory ideas discussed, *inter alia*, in Katz and Postal (1964), it is formulated in Chomsky's *Aspects of the Theory of Syntax* (1965). The core point is a reassessment of the nature of transformational structure. Instead of a separate level **T** of transformations, grammatical transformations are now construed as operations that define the relation between two levels of syntactic representation, both of which are characterized by P-markers. This modification is related to one of the most influential distinctions proposed in the Standard Theory, namely that between *deep* and *surface structure*. According to this view, each sentence is assigned a deep and a surface P-marker, with transformations mapping the former on the latter. Thus except for the different choice of personal or wh-pronouns, (8)(a)–(d) are just different surface realizations of the same deep structure, where the bracketed Agent phrases might furthermore optionally be deleted by $T_{\text{Agent-deletion}}$, known from (4′)

(8) (a) John may expect her to meet him
(b) she may be expected (by John) to meet him
(c) who does John expect her to meet
(d) who may be expected (by John) to be met (by her)

On this account, the semantic interpretation of linguistic expressions could apparently be determined on the basis of their deep structure, which would be an interesting step towards an integration of grammar into an account of meaning and reference. Another important point of the Standard Theory is a more systematic distinction between (a) the lexical system, i.e., the complex system of rules that introduce lexical items with their (idiosyncratic) phonetic, syntactic, and semantic properties, and (b) the grammatical rules that integrate the lexical information into the structure of complex linguistic expressions. Finally, the distinction between levels of representation on the one hand and components of grammar generating and inter-relating these representations is made more perspicuous than in LSLT. The syntactic levels that appeared to be crucial are deep- and surface-structure, while word- and morpheme-structure do not have an independent status. The components of grammatical rules are (a) the base component, including the lexical system and the phrase structure rules, (b) the transformational component, (c) the phonological component, and, possibly (d) the semantic component. Rules and representations are inter-related according to the schema in (9), with levels of representation italized:

(9)

	Phrase Structure Rules	
Base	⇓	
Component:	Lexical Rules	
	⇓	
	Deep Structure	⇒ Semantic Component
	⇓	
	Transformations	
	⇓	
	Surface Structure	⇒ Phonological Component

The base component and the transformations together constitute the syntactic component. The semantic component raises problems to be taken up below, while the organization of the phonological component, extensively discussed by Chomsky and Halle (1968), is assumed to have a fairly clear organization, parallel to that of the syntactic component in relevant respects. The underlying phonemic structure is defined by the phonological features of lexical items and their arrangement in surface structure:

(10) *Phonemic Structure*
⇓
Phonological Component: Phonological Rules
⇓
Phonetic Structure

While there are still optional and obligatory transformations, generalized transformations are dispensed with, the deep structure of complex sentences being generated already in the base component. A particularly elegant feature that emerges from this reorganization is the principle of cyclic operation of the rules, which is added to the earlier assumption that rules are ordered. Cyclic operation apparently captures a deep property of natural languages: the same sequence of operations applies to increasingly larger parts of an utterance, defining its syntactic properties, its phonetic patterns, and the compositional integration of its semantic structure.

An important general effect of the Standard Theory is the perspicuous account it allows for the traditional notion of linguistic universals, or Universal Grammar (UG), considered as the innate, biologically fixed structure underlying the species-specific language capacity. The range of possible rules and the architecture of their interaction, as schematized in (9) and (10), could now be construed as formal universals, the potential primitive elements entering the rules and representations as substantive universals. Under this perspective, the set of possible grammars G accessible as instantiations of UG determines the diversity of possible natural languages, including the range of variation delimiting processes of linguistic change. By the same token, UG provides the innate, formal basis and predisposition that makes language acquisition possible. These accompanying considerations of the Standard Theory led to a new, principled approach to problems of language acquisition, as well as linguistic change.

3. *Towards the Principles and Parameters Theory*

Shortly after the Standard Theory was formulated, empirical analyses as well as efforts to improve the structure of the theory created modifications of the model. Controversial proposals concerned the role of semantics as indicated in (10). Lakoff (1971) and others proposed a model, according to which semantic representations are the underlying structure of linguistic utterances, from which surface structures are step by step derived by means of transformations and lexical rules. The insights and difficulties related to these proposals created a counterproposal, called the Revised Standard Theory. Based on observations discussed in Jackendoff (1972) it became obvious that meaning is not unaffected by surface structure phenomena. For example (11) and (12) are based on the same deep structure, with passivization applying to (12) but not to (11); they do not in general have the same meaning:

(11) Three students in the class wrote only one paper

(12) Only one paper was written by three students in the class

(Due to what is called the scope of quantification, in (11) the restriction of having written one paper is ascribed to three students, while in (12) the restriction of being written by three students is ascribed to a single paper.) As phenomena of this sort are by no means marginal, the Revised Standard Theory recognized a more complex relation between syntax and semantics than that indicated in (9). Only certain aspects of meaning are determined by deep structure, while scope and certain other relations depend on surface structure, such that transformations—contrary to the Standard Theory—might now affect the semantic interpretation.

A technical way to handle these problems is the so-called Trace Theory proposed in Fiengo (1977). According to this proposal, a constituent that is moved by a transformation leaves behind a trace that is phonetically empty, but may play a role in semantics, as indicated in the following example, where e_i is the trace left in the initial position of the co-indexed pronoun:

(13) who_i do you want to talk to e_i

In many cases, the trace turns out to be similar to what logicians consider a variable bound by the operator contained in the moved constituent. In this sense, invisible elements of the surface structure account for its contribution to semantic interpretation. Other empty elements of the surface structure are 'invisible' pronouns, as in (14)(a) and (b), where the actor of the infinitive clause is controlled by (i.e., coreferential with) different constituents of the main clause:

(14) (a) John_i promised Mary [e_i to take the train]
(b) John recommended Mary_i [e_i to take the train]

The structural aspect that emerges from transformations, as for instance in (13), cannot be the Deep Structure of (9), but it also differs from Surface Structure, as it contains elements like traces and empty pronouns. This more abstract surface, called S-structure, is assumed to provide the information for phonetic as well as semantic interpretation. On this account, (9) might be replaced by the general schema (15), where LF, namely the Logical Form, represents all the information G contributes to the meaning of an expression, just as PF, that is the Phonetic Form, specifies the conditions G imposes on its pronunciation:

(15) Deep Structure
|
S-structure
/ \
PF LF

The theory of traces and other 'empty' elements is not a purely technical modification, however. It is

rather part of a general orientation striving for systematic restrictions on the still much too complex possibilities to construct a grammar according to the Standard Theory. One of the reasons motivating this orientation has been called the logical problem of language acquisition. This problem consists of the requirement to specify the conditions under which an appropriate grammar can be identified on the basis of the restricted and partially defective evidence the learning child is normally exposed to. This problem cannot be approached seriously if UG is as unconstrained as it would be with all the complexities initially allowed for as transformations. On the other hand, it turned out that many of the formal complexities exploited, for instance, in operations like the standard passive transformation, can be reduced to more elementary operations if the application and interaction of these operations is subject to more systematic constraints.

Induced by the Trace theory, a number of systematic conditions have been explored which determine the conditions of empty elements of different sorts. These conditions allow transformations to be reduced ultimately to just one simple operation, called Move α, where α is any constituent. The systematic and by no means trivial conditions determining the possible choice of α, the domain and bounds of movement, the consequences of movement (such as leaving a trace), are all expressed by general principles needed on independent grounds. One of the conditions concerns relations lexical elements impose on their necessary or optional complements. This is called their Argument- or Theta-Structure. Another rather general condition entering these principles is a systematic reorganization of phrase structure rules in terms of the so-called X-bar Theory. This theory takes up an observation made already in Harris (1951) about the restricted options according to which constituents X and Y combine in order to form a complex constituent $\bar{X}$. In somewhat simplified terms, properties of $\bar{X}$ are determined by its head X, which furthermore selects Y as its complement or is modified by Y as a free adjunct.

These and a number of other developments led to a new version of the overall framework called the Principles-and-Parameters Theory, formulated in Chomsky's *Lectures on Government and Binding* (1981). The basic idea of this theory is to replace the notion of different types of highly complex rules, from which individual grammars are made up, by a system of universal principles that constrain the effect of rather basic operations like Move α and the options of X-bar theory. The principles in question determine the relation between the positions of a moved constituent, the domain of movement, the possible candidates for movement, and the conditions that require the operation. They constitute separate subsystems, called Government, Binding, Bounding, Control, Case, and Thematic Structure. The properties distinguishing different languages are no longer seen as consequences of intricate systems of different and rather complex rules. It rather seems possible to account for the relevant phenomena by the interaction of these general principles. The obvious differences between different languages is assumed to be reducible to different values of a restricted set of parameters contained in these general principles. Thus the head X of a constituent $\bar{X}$ may be in initial or final position, an embedded clause may need an introductory element or not, etc. Most (or all) of these parameters seem to be related to classes of particular lexical items so that all language particular properties are basically related to lexical information. This leads to a rather different notion of how UG determines an individual grammar G: besides lexical items making up the language particular dictionary, grammars of different languages might differ only by their choice of parameter values, while the principles these parameters rely on are given by UG and hence identical for all languages. Under this perspective, language acquisition consists essentially of identifying the lexical system of a language, thereby fixing the particular values of the parameters contained in UG. Presupposing the schema of levels given in (15), the organization of a grammar G within the Principles-and-Parameters theory can be indicated roughly as follows:

(16) (a) Lexical System, fixing the idiosyncratic phonetic, semantic, and syntactic properties of lexical items
(b) X-bar syntax, Move α
(c) Principles of Government, Binding, Bounding, Thematic Structure, Control, and Case-Assignment
(d) PF-component
(e) LF-component

Among the numerous stimulating consequences of this model is a new perspective on historical and typological variation: like language acquisition, linguistic change and typological differences are not only constrained by UG, but also closely related to choice of parameter values on the basis of universal principles.

4. *The Minimalist Program of Generative Grammar*

Research within this framework explored the principles listed in (15c) in more detail, keeping to the overall orientation of the Generative Enterprise, namely improving the explanatory power by constraining the theory as far as possible. This led in particular to the attempt to derive the principles in question from general conditions of cognitive organization plus the boundaries that are conceptually necessary to identify the domain of language. Minimal assumptions in this sense require linguistic expressions to compositionally relate patterns of the perceptual and articulatory system A-P to representations in the range of con-

ceptually and intentionally organized experience C-I. This requires the internal or I-language determined by G to provide at least the interfaces PF and LF, relating language to A-P and C-I, respectively, as indicated in (17):

(17) signal ⇔ A-P ⇔ PF ↔ LF ⇔ C-I ⇔ environment
G

Thus G must determine pairs $\langle \pi, \lambda \rangle$, where π and λ are representations in PF and LF, respectively. One must assume furthermore, that each pair $\langle \pi, \lambda \rangle$ must be based on a selection N of elements of the lexical inventory of the language. The research strategy of the Minimalist Program proposed in Chomsky (1995) pursues the aim to derive the properties of UG as far as possible from these minimal assumptions together with the hypothesis that UG is subject to conditions of structural economy. This requires first the lexical items entering the selection N to contain all and only the idiosyncratic specifications by means of which N participates in determining the pair $\langle \pi, \lambda \rangle$. This includes, besides phonetic and semantic features interpreted in P-A and C-I, respectively, certain grammatical or formal features controlling the computation of $\langle \pi, \lambda \rangle$ from N. The minimal assumption to be made about this computation is an operation *Merge* which combines two expressions X and Y into a complex expression Z, made up from the features of X and Y, adding the formal features of the head to characterize the formal properties of Z. As a matter of fact, the operation *Merge* is the minimalist version of X-bar syntax, reviving, moreover, a reduced version of the generalized transformations combining two expressions of arbitrary complexity. With this background, it is an interesting, empirical fact about natural language that the derivation from N to $\langle \pi, \lambda \rangle$ cannot in general be restricted to the simplest possibility requiring only the operation of *Merge*, but needs the effect of what was called *Move* α in order to account for facts like *Who*$_i$ *did he to talk to* e_i. As a matter of fact, the operation can be simplified to *Move*, as it is some formal feature of the moving constituent α, which triggers the operation if and only if the feature cannot be eliminated otherwise. More technically, then, the computation of linguistic expressions proceeds as indicated by (18):

(18) $N \Leftrightarrow$ *Merge*, *Move* $\Leftrightarrow$ PF, LF$\rangle$ *Spell Out* PF

A wide range of facts in different languages have been shown to follow from appropriate lexical information together with minimal assumptions about *Merge*, *Move*, plus two sorts of general conditions regulating the economy in representation and derivation of expressions of I-language. What this leads to is indicated in (19):

(19) (a) Conditions on lexical items i.e., sets of phonetic, semantic, and formal features
(b) Operations of the computational system of language: *Merge*, *Move*
(c) Economy Principles:
(i) Representational: Full Interpretation FI
(ii) Derivational: Shortest derivation from N to $\langle$PF, LF$\rangle$-pairs

The requirement of Full Interpretation presupposes that *Merge* and *Move* check and eliminate step by step the formal features, leaving only information that can appropriately be interpreted in terms of articulatory and conceptual conditions at the two interface levels. Derivational economy is a fairly abstract notion, the appropriate formulation of which is still under exploration. In any event, violation of these principles is now taken to be the source of ungrammaticalness: only optimal derivations allowing for complete interpretation yield well-formed expressions. The control of these conditions is a crucial effect of formal features, coming primarily with special lexical items called functional categories. These include grammatical words like Determiners, Auxiliaries, but also inflectional elements like Tense, Number, or Case.

From a more general perspective, the conditions in (9) are rather general, compared with the still rather special principles in the Principles-and-Parameters model (16), very much like these principles were a strong generalization compared with the rule systems of the Standard Theory in (9). The shift from (9) to (16) replaced rules of individual grammars by universal but still language-specific principles; the organization of UG indicated in (19) replaces these language-specific principles with even more general conditions of cognitive organization. Language specificity is now essentially a matter of formal features and their effect on *Merge* and *Move*, that is the operations that compositionally relate P-A to C-I, modules of cognitive organization which are largely independent of the language capacity.

From this point of view, the species-specific language capacity might consist essentially of the disposal of discrete infinity and in fact be related, as Chomsky (1982) speculates, to the computational capacity underlying arithmetic. The language capacity clearly recruits and enhances general cognitive and communicative capacities, but its evolution, which Pinker (1994) persuasively argues to be due to adaptive selection, might well be an independent step accompanying evolutionary changes in the architecture and size of the human brain.

5. *General Perspectives*

The intellectual force of generative grammar comes from the combination of two equally important factors. First, it provides a suggestive methodological and theoretical perspective for the study of language as part of the biological endowment of the human organism, supporting a central system of the overall cognitive capacity. Second, it shows how this perspective can be made precise, turned into an effective research program, and pursued with respect to a large domain of empirical facts not recognized before. It is

indeed the wide range of phenomena—from syllable structure and inflection to quantifier scope and syntactic embedding—as well as the diversity of languages—from English, German, and Chinese to Hebrew, Malayalam, and Walbiri—that gives the Generative Enterprise its unusual place in the field. The explicit formulation of the theory, its technical means, and its cognitive orientation are the reason for its strong impact on a wide range of subdisciplines from research in language comprehension and production, acquisition and aphasia to typology, historical linguistics, computer science, poetics, theory of music, and philosophy of language.

As can be seen from the above, various offshoots developed different, though related, ideas. These are, *inter alia*, Categorial Grammar, Dependency Grammar, Case Grammar, Lexical Functional Grammar (LFG), Generalized Phrase Structure Grammar (GPSG), Head-driven Phrase Structure Grammar (HPSG), Relational Grammar, to mention the more influential proposals. Another type of offspring is the Optimality Theory (OT), which develops certain ideas of the Minimalist Program in a different way, assuming that conditions on linguistic structure may be of different strength and can be violated.

A peculiar relationship must finally be noted between Generative Grammar and the various theories of semantics and pragmatics. Sticking to the initial notion that grammar can and must account for the form, but not properly for the complexities of interpretation and use of language, Chomsky did not extend his theory of language to the domain of semantics and pragmatics. It might be noted in this respect that he always considered Logical Form as a syntactic level, pertaining to the form rather than the interpretation of language. Therefore, Formal Semantics as developed by Montague (1974), Discourse Representation Theory as proposed by Kamp and Reyle (1993), Speech Act Theory as conceived by Searle (1969) and a fair number of related approaches to aspects of interpretation and use are compatible with and in part stimulated by Generative Grammar, but they have not been included into its proper domain. Even though this situation was occasionally met with disappointment, it could not in general reduce the attraction exerted by the Generative Enterprise.

See also: Constituent Structure: Syntactic; Lexical Functional Grammar; Linguistics: Theory of Principles and Parameters; Montague Grammar; Movement Theory and Constraints in Syntax; Syntax; Syntax–Phonology Interface; Syntax–Semantics Interface

Bibliography

Chomsky N 1957 *Syntactic Structures*. Mouton, S'Gravenhague, The Netherlands

Chomsky N 1959 Review of Skinner's *verbal behavior*. *Language* **35**: 26–58

Chomsky N 1963 Formal properties of grammars. In: Luce R D, Bush R R, Galanter E (eds.) *Handbook of Mathematical Psychology*. Wiley & Son, New York, Vol. II, pp. 323–418

Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Chomsky N 1975 *The Logical Structure of Linguistic Theory*. Plenum, New York

Chomsky N 1981 *Lectures on Government and Binding*. Foris Publications, Dordrecht, The Netherlands

Chomsky N 1982 *The Generative Enterprise: A Discussion with Riny Huybrigts and Henk van Rimsdijk*. Foris Publications, Dordrecht, The Netherlands

Chomsky N 1986 *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York

Chomsky N 1995 *The Minimalist Program*. MIT Press, Cambridge, MA

Chomsky N, Halle M 1968 *The Sound Pattern of English*. Harper and Row, New York

Fiengo R 1977 On trace theory. *Linguistic Inquiry* **8**: 35–62

Halle M 1959 *The Sound Pattern of Russian*. Mouton, S' Gravenhage, The Netherlands

Harris Z S 1951 *Methods in Structural Linguistics*. University of Chicago Press, Chicago

Harris Z S 1957 Co-occurrence and transformations in linguistic structure. *Language* **33**: 293–340

Jackendoff R S 1972 *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA

Kamp H, Reyle U 1993 *From Discourse to Logic*, Kluwer Dordrecht: The Netherlands

Katz J J, Fodor J A 1963 The structure of a semantic theory. *Language* **39**: 170–210

Katz J J, Postal P 1964 *An Integrated Theory of Linguistic Descriptions*. MIT Press, Cambridge, MA

Lakoff G 1971 On generative semantics. In: Steinberg D, Jakobovits L (eds.) *Semantics*. Cambridge University Press, Cambridge, UK, pp. 232–96

Lees R B 1960 *The Grammar of English Nominalizations*. Indiana University Press, Bloomington, IN

Miller G A, Chomsky N 1963 Finitary models of language users. In: Luce R D, Bush R R, Galanter E (eds.) *Handbook of Mathematical Psychology*. Wiley & Sons, New York, Vol. II, pp. 419–91

Montague R 1974 *Formal Philosophy*. Yale University Press, New Haven, CT

Pinker S 1994 *The Language Instinct*. Harper Collins, New York

Searle J R 1969 *Speech Acts*. Cambridge University Press, Cambridge, UK

Skinner B F 1957 *Verbal Behavior*. Appleton-Century-Crofts, New York

M. Bierwisch

Generative Syntax

The term generative syntax refers to research on the syntax of natural languages within the general framework of generative grammar. As in generative grammar, the term 'generative' means explicit. Thus, generative syntax differs from much other research on syntax in its emphasis on precise and explicit specifica-

tion both of general theories and of specific analyses. Most work in generative syntax like most work in generative grammar has also been seen as an investigation of I-language (internalized language), the cognitive system underlying the ordinary use of language. Like generative grammar, generative syntax has its origins in Chomsky's *Syntactic Structures* (1957). It is often seen as the central component of generative grammar and it has always been the most prominent.

1. Origins

Syntax was very much a secondary concern for pre-Chomskian linguists and their ideas about syntax were largely unformalized. Chomsky's first concern was to make these ideas explicit and to investigate their adequacy. He argued that they were unsatisfactory in various ways and proposed a richer conception of grammar; known as transformational grammar (TG).

Most pre-Chomskian work on syntax implicitly assumed that words can be classified and that they combine together into phrases which can also be classified. In other words, it assumed that sentences have a constituent structure (or more than one if they are structurally ambiguous). Constituent structures are commonly represented by tree diagrams like (10):

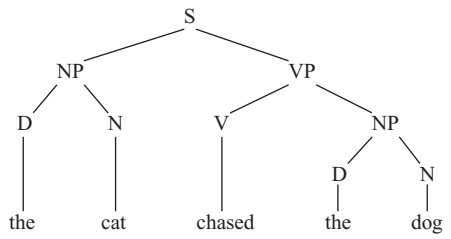

(1)

This tells us what categories the words in the sentence belong to and what phrases the sentence contains. Chomsky pointed out that constituent structures could be characterized by phrase structure (PS) rules. A PS rule has the form in (2) and the interpretation in (3).

$C_0 \rightarrow C_1\ C_2 \cdots C_n$ (2)

A C_0 can consist of a C_1, followed by a C_2, followed by … followed by a C_n (3)

Chomsky argued that grammars consisting just of PS rules could not provide an illuminating account of various syntactic phenomena, and argued for a richer conception of syntactic structure in which some constituents are introduced into the structure somewhere other than their superficial position and then moved to their superficial position. For example, he argued that the subject of a passive sentence such as (4) originates in object position.

Kim was arrested (4)

This provides an explanation for the fact that 'arrested' is acceptable here without an object although it normally requires one, as the ungrammaticality of (5) shows.

* The policeman has arrested. (5)

It also provides an explanation for the fact that 'Kim' is understood in the same way in (4) as in (6), i.e., as the person undergoing the action.

They have arrested Kim. (6)

Phenomena like these led Chomsky to propose that sentences have an underlying constituent structure characterized by PS rules and that the superficial constituent structure is derived from this by various transformational rules. Thus, for most versions of TG, the full syntactic structure of a sentence is a sequence of constituent structures. The underlying structure has been called deep structure or D-structure and the most superficial structure, surface or S-structure.

Since 1957 a variety of different theories have been developed within generative syntax and they have been applied to many different languages.

2. Data

Generative analyses make complex predictions about what is and what is not grammatical, i.e., in accord with the speaker's I-language. The most efficient way of investigating these predictions is through the judgments or intuitions of native speakers. Hence, generative syntax has always been based mainly on intuitions. Of course, intuitions have to be treated with caution. The fact that a speaker finds some sentence unacceptable does not necessarily mean that it is ungrammatical for her. A speaker might find (7) unacceptable because it is false, and might find (8) unacceptable because it is misleading. (It is natural initially to assume incorrectly that 'her' forms a constituent with 'contributions to the fund.')

Three fours are thirteen. (7)

Without her contributions to the fund would be quite inadequate. (8)

However, if a sentence is unacceptable and there is no other explanation for the unacceptability, it is reasonable to assume that it is ungrammatical.

None of this implies that other kinds of data, e.g., corpora of naturally occurring speech or writing, cannot be useful. In fact generative syntacticians have often used corpus data. Moreover, there is no alternative to corpus data if the speakers are dead. Hence, generative work on dead languages is inevitably based on a corpus supplemented by intelligent guesses about meanings. Corpus data has also played a central role in generative work on the early stages of first language acquisition.

The emphasis on intuitions has encouraged researchers to work on their own language. In the early years of generative syntax, most researchers were native speakers of English, and as a result English had more attention than any other language. However, as Newmeyer (1986, p. 40) emphasizes, generative syntax

was never limited to English. Moreover, languages other than English have been increasingly prominent since about 1980 as a result of the growing numbers of generative syntacticians who are native speakers of other languages. Work in generative syntax often focuses on a single language, but work that compares and contrasts a pair of related languages is also quite common, and some work deals with a broad range of languages.

3. *Structures*

As has been noted, most versions of TG assume that the full syntactic structure of a sentence is a sequence of constituent structures. Different versions of TG differ in how complex these constituent structures are. Typically, since the early 1980s, they have been much more complex than (1) as a result especially of the presence of various phonologically empty elements. These include traces, empty categories 'left behind' by movement processes. Thus, (4) will have a trace in object position left behind by movement of the NP 'Kim.' Empty elements also include empty functional categories. Functional categories include C (complementizer) and I (inflection), exemplified by 'that' and 'may' in (9), respectively.

I think that the cat may chase the dog. (9)

It is standardly assumed that a simple sentence contains empty members of these categories. Thus, instead of (1), one might have something like (10)

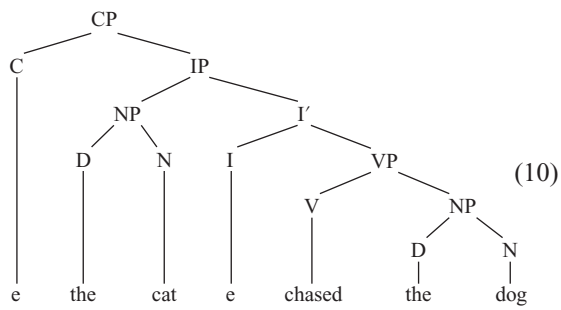

(10)

A variety of functional categories, which are commonly empty, have been assumed in recent TG work.

The assumption that the syntactic structure of a sentence is a sequence of constituent structures has been largely abandoned within the recent minimalist version of TG (Chomsky 1995, Radford 1997). Within this framework, movement processes apply within a constituent before it is combined with another constituent. In (11), for example, 'Kim' is moved to subject position in the subordinate clause before this clause is combined with the verb 'think.'

I think Kim was arrested. (11)

Hence, there is no level of structure after all processes of combination have applied and before any movement processes apply. In other words, there is no level of D-structure.

Much work in generative syntax has argued for different sorts of structures. Standard constituent structures do not incorporate labels like subject and object. It is assumed within TG that these notions can be defined in structural terms. Assuming structures like (10), one might define subject as an NP that is an immediate constituent of IP and object as an NP that is an immediate constituent of VP. Proponents of Relational Grammar (RG) argued in the early 1970s that such functional notions should be primitives and that constituent structure trees should be replaced by relational networks like (12), where 1 and 2 stand for subject and object, respectively, and P stands for predicate:

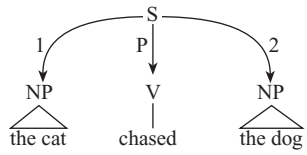

(12)

In early work in RG, the structure of a sentence is a sequence of such relational networks and the order of constituents is only significant in the most superficial network. In more recent work, the structure of a sentence is a single complex network and order is determined by its most superficial properties. (For discussion, see Blake 1990.)

The idea that certain functional notions should be primitives is also a feature of Head driven Phrase Structure Grammar (HPSG) (see Pollard and Sag 1994). Here the structure of a sentence is a single quite simple constituent structure enriched with information about whether a constituent is a head (the constituent which determines the identity of the larger expression, a noun in a noun phrase, a verb in a verb phrase, etc.), a subject, a complement, etc. (Object is not a primitive in HPSG.) A somewhat different position is found in Lexical-Functional Grammar (LFG). Here the full structure of a sentence consists of a fairly conventional constituent structure and a functional structure, in which constitutents are identified as subjects, objects, etc. Functional structures provide a way of capturing the similarities between superficially rather different languages. For example, the following sentences from English, Welsh, and Turkish, which have rather different constituent structures because subject, verb, and object are ordered differently, could have essentially the same functional structure:

The cat saw the dog. (13)

Gwelodd y gath y ci (Welsh) (14)
saw the cat the dog
'The cat saw the dog.'

Kedi köpeğ -i gör-dü (Turkish) (15)
cat dog-ACC see-Past

'The cat saw the dog.'
See Bresnan (2001) for discussion.

An important question about the various forms of generative syntax is: what sorts of categories do they assume? Chomsky (1965, 1970) argued that various syntactic phenomena require the assumption that syntactic categories are complex entities made up of smaller elements normally known as features. This position has been generally accepted since then. However, different approaches differ in how complex their categories are. Categories are particularly complex within HPSG, where, as has been noted, the structure of a sentence is a single quite simple constituent structure. In contrast, they appear to be relatively simple within recent TG, which assumes much more complex structures.

4. *Rules and Principles*

Like any scientific research, work within generative syntax is concerned to capture whatever generalizations can be found in the data. A consequence of this is that broad general statements are preferred to statements of more limited scope.

In the 1960s and early 1970s TG assumed a variety of different PS and transformational rules. More recently, the various transformations have been replaced by a single rule known as Move Alpha or in recent work just Move. This is essentially a license to move anything anywhere. It interacts with various conditions on movement and the results of movement to allow just the right movement processes. Particularly important in recent work have been certain economy conditions, which prefer simpler to more complex derivations. In much the same way, the various PS rules have been replaced by a few general principles interacting with the categorial makeup of lexical items to give the observed variety of basic, pre-movement structures. In the minimalist version of TG, it is assumed that a single rule of Merge is operative here. A similar position is found in much work in HPSG, in which it is assumed that the variety of syntactic structures are the product of a small number of rules and principles interacting with the properties of lexical items.

Within TG, the move away from large numbers of PS and transformational rules has been associated with a denial that constructions as traditionally understood are of any theoretical significance. Thus, Chomsky remarks that '[t] he notion of construction, in the traditional sense, effectively disappears; it is perhaps useful for descriptive taxonomy but has no theoretical status' (1995, p. 25). However, since the late 1980s a number of syntacticians have argued that it is impossible to characterize the full range of syntactic structures with just broad general principles and individual lexical items, and that it is necessary to recognize constructions with certain properties of their own after all. For example, Sag (1997) argues that data such as the following require the recognition of a relative clause construction with properties of its own which do not derive either from broad general principles or individual lexical items:

someone on whom I relied (16a)
someone who I relied on (16b)
someone on whom to rely (16c)
*someone who to rely on (16d)

(See Kay and Fillmore (1999) and Culicover and Jackendoff (1999) for similar arguments.)

Different approaches to syntax have taken different views about the nature of rules of syntax. Much work in TG and especially the minimalist framework has treated most rules as procedures for constructing well-formed structures. In contrast RG, HPSG, and LFG have interpreted rules as constraints to which well-formed structures must conform. Consequently these approaches are often described as constraint-based.

A further important issue is: how do grammatical constraints interact? The traditional view is that syntactic structures must conform to all relevant constraints. In contrast, proponents of what is known as Optimality Theory (OT) have argued that constraints are ranked and that a structure may violate a constraint if that is the only way to satisfy a higher ranked constraint. This view of constraints has been very influential within phonology but has also had some influence within syntax. It is combined with TG assumptions in e.g., Grimshaw (1997). Grimshaw argues, for example, that there is a constraint which rules out the movement of constituents. However, a higher ranked constraint requires movement of 'who' to sentence initial position from its underlying position as object of 'to' in the following:

Who did you talk to? (17)

The OT view of constraints has also been combined with LFG assumptions e.g. in Bresnan (2000).

5. *Syntax and Other Components of Grammar*

A central question for any theory of syntax is: how does the syntactic component of a grammar relate to other components? What is the relation between syntax on the one hand and semantics, morphology, phonology, and the lexicon on the other?

In the late 1960s and early 1970s, the relation between syntactic and semantic representations was a major topic of debate. Proponents of Generative Semantics argued that the deepest level of syntactic representation is a semantic representation. Proponents of Interpretive Semantics rejected this view and argued that semantic representation is derived by rules referring to both deep and surface structure. There are questions as to how different the two positions really were in the early stages of the debate. In the later stages they were clearly distinguished as generative semantics sought to bring more and more linguistic phenomena within the scope of grammar. Generative semantics collapsed in the mid 1970s, partly because of

a realization that some of the phenomena it had brought within the scope of grammar could be handled more satisfactorily within pragmatics. (See Newmeyer 1986, Chaps. 4 and 5 for discussion.) It is now generally accepted within generative syntax that syntactic and semantic representations are rather different, and that there is no one-to-one correspondence between them. The following examples are relevant here:

Have you seen a unicorn? (18a)
Never **have I** seen a unicorn. (18b)
Kim has seen a unicorn and so **has Lee**. (18c)
Had I seen a unicorn, I would have told you. (18d)

All these sentences exemplify what might be called the English auxiliary + subject construction. The key words are in bold. They show that this construction has a variety of different semantic interpretations. (18a) contains an interrogative, the kind of sentence that is normally used to ask a question. (18b) and (18c) contain declaratives, the type of sentence normally used to make a statement. Finally, (18d) contains a conditional clause. Here, instead of auxiliary + subject order, one might have 'if' and subject + auxiliary order.

Since the mid 1970s, most work in TG has assumed that semantic interpretation is based on a level of logical form (LF) derived from S structure by various movement processes. However, Jackendoff (1997) develops a detailed critique of this view. For HPSG, general principles provide a semantic representation for all expressions on the basis of the semantic representations of their constituents. For LFG, semantic interpretation is based on functional structure.

There are also important questions about the relation between syntax and morphology. At various times it has been proposed within TG that word stems may be combined with affixes in syntax. For example, it has sometimes been suggested that an, English verb stem combines with the past tense suffix '-ed' through a movement process. This view largely fell from favor in the 1990s and it is now widely accepted even within TG that word-formation is confined to the lexicon.

There are also questions about the relation between syntax and phonology. It has sometimes been proposed within TG that certain movement processes apply within the phonological component. For example, Chomsky (1995, p. 368) proposes that verb-second order in German is a result of a rule applying within the phonological component. The following examples illustrate this phenomenon:

Ich las schon letztes Jahr diesen Roman. (19a)
I read already last year this book
Diesen Roman las ich schon letztes Jahr. (19b)
this book read I already last year
Schon letztes Jahr las ich diesen Roman. (19c)
already last year read I this book
'I read this book already last year.'

Each example has a different constituent in initial position, but all have the verb in second position. It has also been assumed in much work in HPSG that linear order is solely a property of phonology in the sense that it is not words and phrases but only their phonological properties that are ordered.

Questions also arise about the relation between syntax and the lexicon. Within TG a central issue is where the insertion of lexical items into syntactic structures applies. It has often been assumed that it applies before any movement process. However, it has sometimes been proposed that some lexical insertion may take place after movement. Within HPSG, where lexical categories are quite complex, there are important questions about how generalizations about the categorial makeup of words can be captured. Considerable use has been made of lexical rules deriving lexical entries form lexical entries. More recently, it has been argued that most generalizations can be captured if words are classified in terms of a number of hierarchies of lexical types from which they inherit properties. Thus, 'chased' might be classified as both a transitive verb and a past tense verb and combine the properties of these two types (see Koenig 1999).

There are further issues about the scope of syntax. Should grammars include information about the contexts in which constructions are appropriate? For example, should a grammar include the information that (19) with no overt object following 'beat' is only appropriate in recipes?

Beat until stiff. (20)

It has generally been assumed that such information is something separate interacting with purely grammatical information (see, for example, Culy 1996). However, a number of researchers have argued that it is inseparable from purely grammatical information (see Hudson 1996, Paolillo 2000). Should grammars incorporate statistical information about how likely certain constructions are to occur? For example, should a grammar incorporate the information that (21a) with a clause in subject position is less likely to occur than (21b) with the clause in final position?

That Kim was there is odd. (21a)
It is odd that Kim was there. (21b)

Traditionally, statistical facts about syntactic structures have been seen as a matter of performance, but Abney (1996) argues that such information is part and parcel of grammars.

6. *Some Syntactic Phenomena*

There are many unresolved issues in generative syntax but a great deal has been learned about both the universal syntactic properties of natural languages and the ways in which they can vary in their syntax.

A major focus of research has been clause structure and a lot is known about the kinds of clause structure that a language may exhibit and about what kinds of analyses might be appropriate. This includes such

matters as the order of subject, verb, and object (see (13)–(15), the realization of tense, mood, and negation, agreement, and case-marking, and the relation between main and subordinate clauses. German, which has verb-final order in subordinate clauses, is relevant here. The following subordinate clause contrasts with the examples in (19):

dass ich schon letztes Jahr diesen Roman las (22a)
that I already last year this book read
'that I read this book last year'

There has also been a great deal of research on noun phrases and a lot is known about such matters as the realization of definiteness, the positioning of adjectives (which normally precede the noun in some languages, e.g., English, but normally follow in others, e.g., Welsh), and the positioning of relative clauses (which follow the noun in English and most European languages but precede in Turkish, Japanese, and Korean).

Perhaps the most important focus of research in generative syntax has been the various kinds of non-local dependencies which natural languages exhibit, which were largely unnoticed in pre-generative work. An important example is the dependency between a clause-initial interrogative element like 'who' and the associated gap which is a feature of many languages. A simple example is (21) in which there is a gap in object position.

Who did Kim annoy? (23)

It has been known since the 1960s and especially since Ross (1967) that these dependencies are subject to various constraints. For example, while (23) is fine, (24) with a gap following 'and' is ungrammatical.

* Who did Kim annoy Lee and? (24)

There are important differences between languages in this area. English allows a gap in prepositional object position. Hence both the following are possible:

Who did you talk to? (25a)
To whom did you talk? (25b)

However, most languages do not allow a gap in prepositional object position. The following Polish examples are typical.

* Kim rozmawiałeś z? (Polish) (26a)
whom you-talked with
Z kim rozmawiałeś? (Polish) (26b)
with whom you-talked

In contrast, English and many other languages do not allow a gap in subject position in a subordinate clause introduced by a complementizer. Hence, the following is ungrammatical:

* Who did you say that wrote this book? (27)

However, some languages do allow this, Thus, the following Italian example is grammatical:

Chi hai detto che ha scritto questo libro? (Italian) (28)
who have-you said that has written this book
'Who did you say wrote this book?'

As the translation makes clear, an English example is grammatical if there is no complementizer. See Roberts (1997, Chap. 4) for discussion of the extensive work on these matters.

Another important dependency is that between a reflexive pronoun such as English 'himself' and its antecedent, the expression from which it derives its interpretation. In English, while (29a) is acceptable, (29b) is not.

John criticized himself. (29a)
*John said Mary criticized himself (29b)

Thus, the antecedent must be quite close to the reflexive. In other languages, however, a reflexive can be further from its antecedent than in English. For example, in the following Japanese sentence, the antecedent of the reflexive 'zibun' can be either 'Ziro' or 'Taro'.

Taroo-ga Hanako-ni[Ziroo-ga zibun-o (Japanese) (30)
Taro-NOMHanako-toZiro-NOMself-ACC
hihan-sita]-to itta.
criticized-COMP said
'Taro said to Hanako that Ziro criticized self.'

Hanako is not a possible antecedent because it is not a subject. For discussion of these matters, see Roberts (1997, Chap. 3).

Non-local dependencies like these are likely to remain a major focus of research in generative syntax for the foreseeable future.

7. *Acquisition and Processing*

As has been noted, it is generally accepted within generative work that syntax is concerned with I-language, the cognitive system underlying ordinary use of language. It follows that a variety of data are relevant, for example, facts about acquisition, facts about processing, and facts about pathology.

Acquisition has been a major concern within TG. Since about 1980, it has been argued that the language faculty consists of a set of principles to which all languages conform and a set of parameters which allow a limited amount of variation between languages. For example, it is commonly assumed that there is a parameter distinguishing languages like English, in which interrogative elements appear in initial position, from those like Japanese, in which they appear in the same position as corresponding non-interrogative elements. The following examples illustrate:

John gave the book to Mary. (31a)
Who did John give the book to? (31b)

John-wa Mary-ni sonohon-o
John-TOP Mary-DATthat book-ACC
agemasi-ta-ka. (Japanese) (32a)
give-PAST-Q
'John gave the book to Mary.'

John-wa dare-ni sono hon-o
John-TOP who-DATthat book-ACC
agemasi-ta-ka? (Japanese) (32b)

give-PAST-Q
'Who did John give the book to?'
Within this framework, language acquisition is largely a matter of setting parameters. A great deal of work on the acquisition of syntax has drawn on these ideas. However, some are skeptical. There are two reasons for this. First, relatively little progress has been made towards a plausible set of parameters even for the central or 'core' phenomena of syntax, and languages have a wide range of noncore phenomena, which are also acquired (Culicover 1999). Second, it is clear that setting a parameter is a much more complex matter than it was originally thought to be (Fodor 1998). Nevertheless the idea remains a potentially important one.

Other frameworks have been more concerned with language use. It has been argued that HPSG receives some support from what is known about human language processing. HPSG is a constraint-based framework, in which all rules and principles are on a par and none takes precedence over any other. Pollard and Sag (1994, pp. 11–12) argue that this permits an explanation for some important features of human language processing. They highlight four features. First, it is highly incremental in the sense that 'speakers are able to assign partial interpretations to partial utterances.' Second, it is highly integrative in the sense that 'information about the world, the context, and the topic at hand is skillfully woven together with linguistic information whenever utterances are successfully decoded.' Third, '… there is no one order in which information is consulted which can be fixed for all language use situations.' Finally, it consists of various kinds of activity: 'comprehension, production, translation, playing language games, and the like.' They argue that these features are unproblematic for constraint-based grammars but pose serious problems for other sorts of grammars.

Facts about acquisition and processing are likely to be of continuing importance for the evaluation of proposals within generative syntax.

See also: Generative Grammar; Movement Theory and Constraints in Syntax; Syntax; Syntax–Phonology Interface; Syntax–Semantics Interface

Bibliography

Abney S 1996 Statistical methods and linguistics. In: Klavans J L, Resnick P (eds.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press, Cambridge, MA, pp. 1–26

Blake B J 1990 *Relational Grammar*. Routledge, London

Bresnan J 2000 Optimal syntax. In: Dekkers J, van der Leeuw F, van de Weijer J (eds.) *Optimality Theory: Phonology, Syntax and Acquisition*, Oxford University Press, Oxford, UK, pp. 334–85

Bresnan J 2001 *Lexical-Functional Grammar*. Blackwell, Oxford, UK

Chomsky N A 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands

Chomsky N A 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Chomsky N A 1970 Remarks on nominalization. In: Jacobs R, Rosenbaum P S (eds.) *Readings in English Transformational Grammar*. Ginn, Waltham, MA

Chomsky N A 1995 *The Minimalist Program*. MIT Press, Cambridge, MA

Culicover P W 1999 *Syntactic Nuts: Hard Cases, Syntactic Theory and Language Acquisition*. Oxford University Press, Oxford, UK

Culicover P W, Jackendoff R S 1999 The view from the periphery: The English comparative correlative. *Linguistic Inquiry* **30**: 543–71

Culy C 1996 Null objects in English recipes. *Language Variation and Change* **8**: 91–124

Fodor J D 1998 Unambiguous triggers. *Linguistic Inquiry* **36**: 1–36

Grimshaw J 1997 Projection, heads and optimality. *Linguistic Inquiry* **28**: 373–422

Hudson R A 1996 *Sociolinguistics*, 2nd edn. Cambridge University Press, Cambridge, UK

Jackendoff R S 1997 *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA

Kay P, Fillmore C 1999 Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language* **75**: 1–33

Koenig J-P 1999 *Lexical Relations*. CSLI Publications, Stanford, CA

Newmeyer F J 1986 *Linguistic Theory in America*, 2nd edn. MIT Press, Cambridge, MA

Paolillo J C 2000 Formalizing formality: An analysis of register variation in Sinhala. *Journal of Linguistics* **36**: 215–59

Pollard C, Sag I A 1994 *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago, IL

Radford A 1997 *Syntactic Theory and the Structure of English: A Minimalist Approach*. Cambridge University Press, Cambridge, UK

Roberts I 1997 *Comparative Syntax*. Arnold, London

Ross J R 1967 *Constraints on Variables in Syntax*. PhD dissertation, MIT

Sag I A 1997 English relative clauses. *Journal of Linguistics* **33**: 431–83

R. Borsley

Genes and Behavior: Animal Models

Explaining behavior in genetic terms is no small task. Behavior draws on the activities of a wide range of genes acting at many different times in an individual's life. Model organisms have provided insight into the role of genes in behavior that have major implications for humans. Many of the mutations and genetic variants (alleles) that affect behavior in *Drosophila* and mice have proved to be subtle changes in genes that, when more severely mutated, produce wider ranging effects. Studies of mutants induced in the laboratory have identified a variety of genes that affect behavior:

genes that regulate other genes (transcription factors), regulatory enzymes (protein kinases), and signaling molecules in the brain (neuropeptides and receptors). Their study has made major contributions to our understanding of biological rhythms, learning, memory, and social behavior, and has helped explain newer findings on naturally occurring variants in behavior.

1. Introduction

Unitary explanations for human behavior have a dismal history in the modern world. Whether economic, cultural, psychological, or genetic, attempts to reduce human behavior to a single influence inevitably fail to give a full account and have often produced distortions with unfortunate political consequences. Charting the relationship between genes and behavior can therefore be a perilous activity. The challenges are many: the difficulty of defining genetic and nongenetic factors, the difficulty of understanding the interactions among these many factors, the many steps that intervene between the expression of a gene and the manifestation of a behavior, and the inescapable fact that each individual is the unique product of a series of historical accidents. Historical uniqueness also applies to nonhuman biological individuals. It is a hallmark of the biological world due to the fact that no particular combination of genes and experiences is ever replicated exactly in nature.

Studies of genes and behavior in model organisms offer one way out of this conundrum. To begin with, it is far easier to control environmental conditions and genetic heterogeneity in the laboratory, and such studies are generally undertaken keeping environmental conditions as uniform as possible and genetic makeup as well defined as possible. But this would be no help if the phenomena and principles are fundamentally different from those of humans. Fortunately, they appear to be similar by an ever-increasing number of criteria. One of the most striking outcomes of modern biology and the sequencing of genomes is the similarity of genes between distantly related creatures. The fruit fly *Drosophila melanogaster* and humans share a great many of the same genes in common. Similarly, they share considerable conservation of metabolic and signaling pathways at the cellular level. Most strikingly, and most relevant for this discussion, there is growing evidence for conservation at the level of behavior and its molecular mechanisms. This has been shown most convincingly, thus far, for several behaviors in *Drosophila*: circadian rhythms (Young 1998), learning and memory (Dubnau and Tully 1998), and most recently sleep (Greenspan and Ferveur 2000). Thus, there is good reason to believe that the characteristics, interactions, and contributions of genes in the behavior of model organisms will provide important guidelines for how to think about the role of genes in human behavior.

Traditionally, there have been two separate, and often hostile, approaches to behavior genetics in model organisms: the measurement and manipulation of naturally occurring variation in laboratory or wild-caught strains, and the isolation and manipulation of newly induced mutations. The former assayed the extent of relatively mild mutational variation that survives in the world and exerts effects on behavior, without being able to know the identity of any of the genes or their mechanism of action. The latter concentrated exclusively on identifying genes and mechanisms subserving particular behaviors, but did so by producing relatively drastic genetic lesions that did not reflect the permissable variation that would have been worked on by natural selection. Both have been relevant to humans, but a synthesis between the two, made possible by recent advances in molecular biology, holds the most promise for useful insights. In this article, a few salient examples of gene effects on behavior are described as illustrations of general principles that are emerging.

2. Severe Single-gene Mutants

2.1 Circadian Rhythms

The clearest contribution single-gene mutants have made is in the realm of identifying individual genes that are central to behavioral mechanisms. The unraveling of the cellular mechanism of the circadian clock began with the discovery of the *period* gene through the isolation of long-day, short-day, and arrhythmic mutants in *Drosophila* (Fig. 1; Konopka and Benzer 1971) and proceeded with the isolation and cloning of additional mutants in flies, fungi, and mice (Young 1998). What has emerged is a picture of the circadian clock as a feedback loop in which each cell counts out its own 24 hour period by means of an oscillating cycle of gene readout in which some of the products accumulate up to a point, then begin to dampen their own synthesis through a series of time delays between the various stages of gene expression: transcription, translation, and import back into the nucleus (Fig. 1). This serves as the central regulator for specific implementation of rhythmic activities. The cellular mechanism of the circadian clock, as well as the clock's phenomenology, appears to be universal in the biological world and many of the actual genes involved, such as *period*, are well conserved between flies and mammals. The human relevance is thus transparent.

2.2 Learning and Memory

Learning and memory have many features common to flies, rodents, and humans. Behaviorally, all are

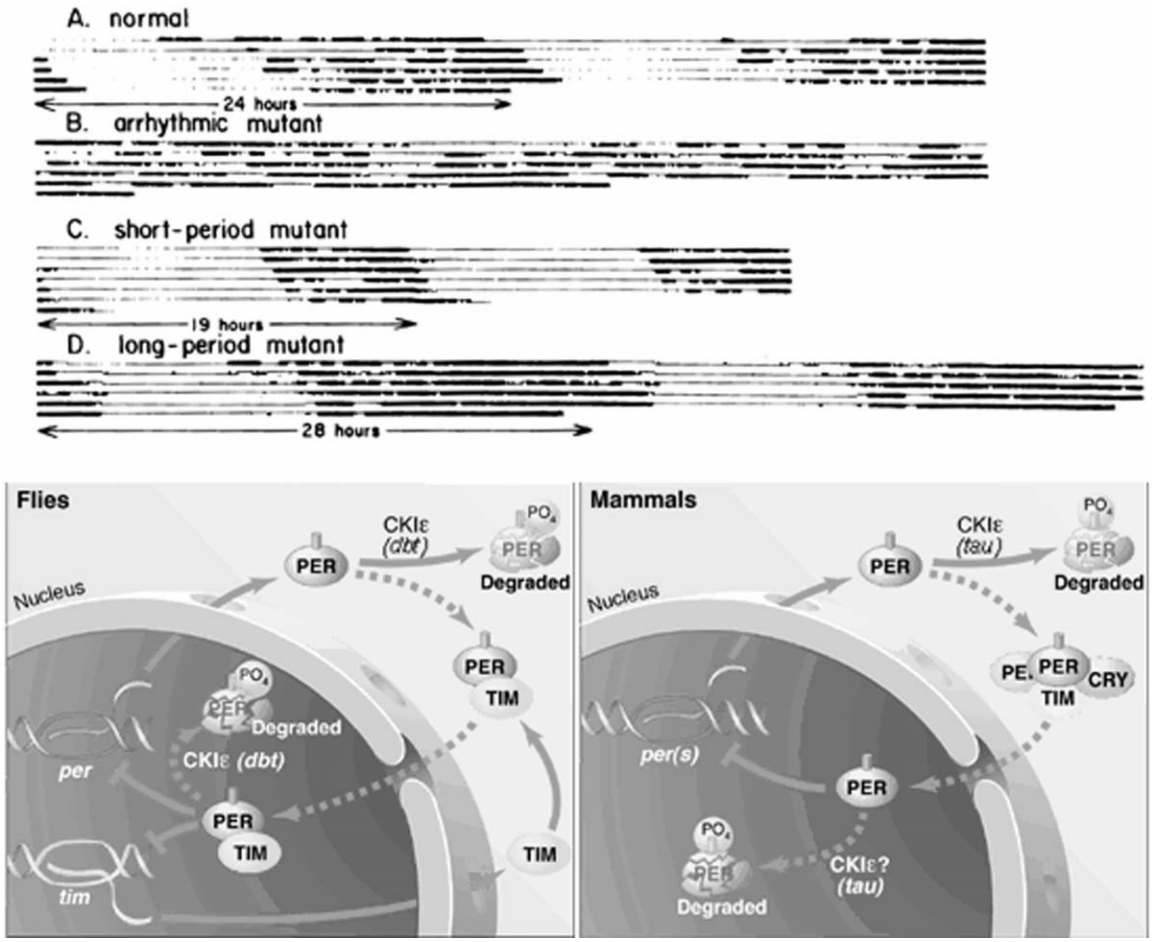

Figure 1
Top: circadian locomotor activity in normal *Drosophila* and in mutants of the *period* gene as measured in an activity monitor in the absence of a day/night cycle of light and darkness. Heavy traces represent activity, with the clock's day length indicated underneath (Source: *Proceedings of the National Academy of Sciences USA* **68**: 2112–6. Reproduced with permission). Bottom: conservation of molecular mechanisms of the circadian clock between flies and mammals. PER = *period* gene, TIM = *timeless* gene, CKIe = casein kinase Ie gene (Source: *Science* 2000 **288**: 451–3. Reproduced with permission)

capable of associative conditioning and the memory that is induced shows multiple phases that can be distinguished by pharmacology and by training regimen (Dubnau and Tully 1998). In the fly, learning takes the form of distinguishing between two odors and avoiding the one that has previously been paired with electric shock. Long-term memory is produced by repeated training trials separated by suitable time intervals. The original *Drosophila* learning mutant, *dunce* (Dudai et al. 1976), is defective in initial acquisition as well as in subsequent retention of the learned behavior (Fig. 2). Care was taken to ascertain that these mutant flies were normal in all other relevant faculties that are required for the learning task: olfaction, locomotor activity, and shock reactivity.

The product of the *dunce* gene was subsequently identified as an enzyme that is part of the signaling cascade in nerve cells that utilizes the intracellular messenger molecule cyclic AMP (Fig. 2). The gene encodes a cAMP-phosphodiesterase, the enzyme that degrades cyclic AMP. Mutations in other components of this signaling system, *rutabaga* in the synthetic enzyme (adenylate cyclase), and *Dco* and *RI* in the key enzyme regulated by cyclic AMP (cAMP-dependent protein kinase), were subsequently shown to also affect learning and memory in the fly (Dubnau and Tully

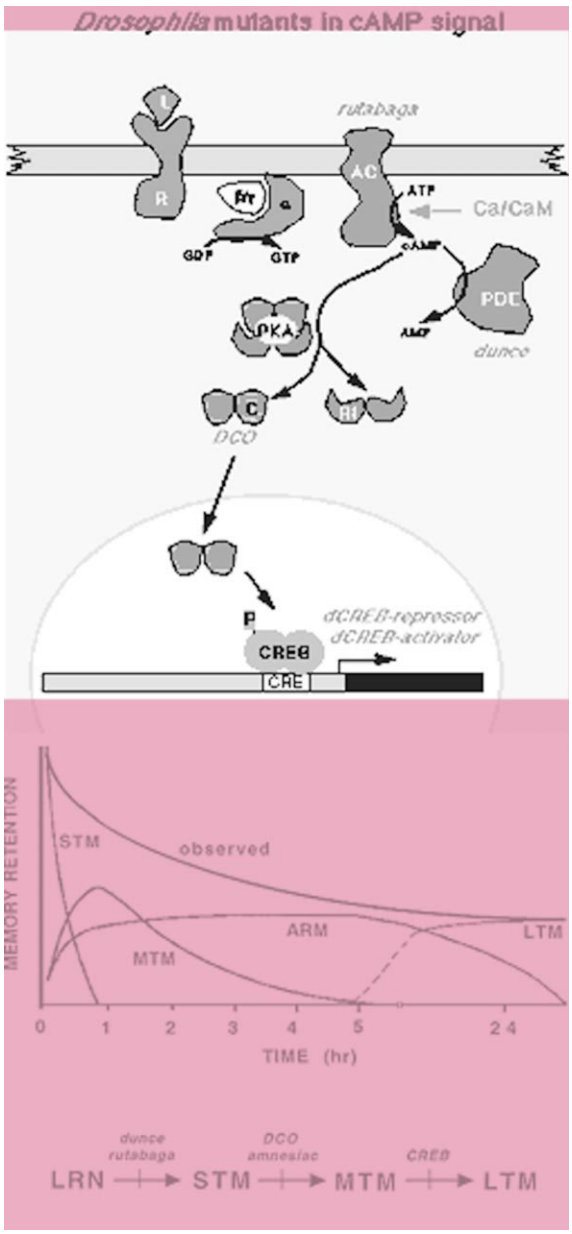

Figure 2
Top: the cAMP signaling pathway in memory formation. AC = adenylate cyclase (*rutabaga* gene), PDE = cAMP phosphodiesterase (*dunce* gene), PKA = cAMP-dependent protein kinase (*DCO* and *RI* genes), CREB = cAMP-response-element-binding protein (*dCreb* gene) (source: *Nature Neuroscience* 1998 **1**: 543–5. Reproduced with permission). Middle: memory is consolidated via distinct phases that are genetically and pharmacologically dissectable in *Drosophila* and in mammals: short-term memory (STM), middle-term memory (MTM), anesthesia-resistant memory (ARM), and long-term memory (LTM) appear sequentially, and each has a progressively slower rate of decay (source: *Trends Neuroscience* 1995 **18**: 212–7). Bottom: a genetic model of information flow showing where in the pathway single gene mutant and pharmacological disruptions have their primary effects (source: *Cell* 1994 **79**: 35–47. Reproduced with permission)

1998). The cyclic AMP signaling system had already been implicated in neuronal plasticity as central to the mechanism that modifies efficacy at synapses in the circuitry underlying learning. The isolation of mutants affecting the system from unbiased, open-ended mutant screens and the demonstration that they affected classical conditioning in the intact animal gave universality to the finding. Mouse mutants in the synthetic enzyme (adenylate cyclase) show analogous deficits in learning (Silva et al. 1998). The relevance to humans has been borne out by the fact that there are currently several memory enhancement drugs in clinical trial that target cAMP phosphodiesterase.

As additional mutants accumulated, comparison of the time course of memory decay revealed some interesting differences between them (Fig. 2), suggesting that they differentially affected memory retention. This suggestion received major support from studies of another component of the cAMP signaling system: the *Creb* (**c**AMP-**r**esponse-**e**lement-**b**inding) protein that regulates genes induced by cAMP. *Creb* exists in both repressor and activator forms. A repressor form of the gene, when engineered to be expressed in the fly just prior to training, completely blocks the formation of long-term memory while permitting normal short- and medium-term memory (Dubnau and Tully 1998). In contrast, an activator form of the gene, when engineered to be expressed just prior to training, produces long-term memory after a single trial, in contrast to the usual 10 trials with intervals that are required to induce long-term memory (Fig. 2). These findings not only placed the differential genetic effects on memory phases on firmer ground, they were also subsequently extrapolated to mice and rats (Silva et al. 1998). A *Creb* mutant in the mouse that reduces, but does not eliminate, repressor activity fails to form long-term memory in a fear-avoidance paradigm. Analogous results have been obtained in the rat after injection of a *Creb* gene blocker (anti-sense RNA) to specifically reduce levels of the protein.

A variety of other mutants disrupting learning that also affect synaptic efficacy have also been obtained in the fly and mouse, including those affecting other regulatory enzymes (calcium/calmodulin-dependent protein kinase, *fyn* kinase) and receptors (NMDA and acetylcholine). The studies of the cAMP cascade stand out, however, for the depth to which they have taken the analysis through their insight into the mechanism of differential memory phases.

2.3 Social Behavior

Genetic studies of social behavior have great relevance for humans and for considerations of human evolution. A productive avenue into this question has been taken in rodents, based on neuroendocrinological studies of the behavioral effects of oxytocin and vasopressin. Administration of these hormones produce complementary effects on a variety of behaviors including affiliative, sexual, maternal, territorial, and aggressive behavior as well as on social memory (Young et al. 1998). The detailed neural mechanisms for the hormones' actions have yet to be elaborated. A mouse mutant that lacks oxytocin altogether ($Oxt^{-/-}$) shows a distinct deficit in social memory as reflected by male olfactory investigation of a female in repeated exposures. The effect is more selective than predicted from the range of effects when oxytocin is administered. Moreover, $Oxt^{-/-}$ mutant mice are normal for olfactory responses to other, nonsocial odors, and normal for learning tasks unrelated to this social behavior.

An increase in aggression between males was also observed in the oxytocin-deficient mice. This observation falls into line with a growing number of other mouse mutant studies reporting increases in aggression. These range from mutants in receptors for serotonin or adenosine, to synthetic enzymes for the neurotransmitters GABA or nitric oxide, to regulatory enzymes (calcium/calmodulin-dependent protein kinase II) to cell adhesion molecules. In contrast to mutants of circadian rhythms or of learning and memory, there seem to be a great many genetic lesions that produce aggression in mice. When one encounters such promiscuity in genetics, it is usually diagnostic of a system that is subject to many nonspecific kinds of disruption, with all of the caveats to interpretation that accompany effects on such a system, as opposed to a sensitive indicator of key mechanistic components.

2.4 Range of Action of Behavioral Genes

The foregoing examples highlight the mechanistic insights that have been gained from behavioral studies of laboratory-induced mutants. An important criterion for these studies was the relative refractoriness of the behavior to mutational disruption: mutations in most genes do not affect these behaviors. The question then arises of the relative specificity of these genes to the behaviors they disrupt. Are these 'genes for' behaviors such as rhythms, memory, or social life, as have often been alluded to in evolutionary discussions of behavior (e.g., Wilson 1978)? For the mutants described above, as well as for the vast majority of behavioral mutants studied in both flies and mice, the answer appears to be negative. By and large, the genes that produce behavioral mutants are not limited to affecting only one behavior or only one aspect of the animal. They exert influences over many facets of the organism's biology (the genetic term for this is 'pleiotropy').

The archetypal circadian rhythms gene, *period*, regulates the song produced by a male fly during courtship as well as its 24-hour rhythms (Greenspan and Ferveur 2000). The male's song, produced by vibrating the wing, consists of a series of 'pulses' which sound like bumps occurring at the rate of about 30 per second. The rhythm lies in the slight but regular variation in the interval between pulses (Fig. 3) as the male gradually increases then decreases this interval to produce a smoothly wavering song over a period of approximately 60 seconds. Female flies respond to the male's song by becoming more receptive to their courtship advances. Mutants in the *period* gene that affect the circadian rhythm have corresponding effects on song rhythm such that long-day mutants have an 80-second rhythm, short-day mutants have a 40-second rhythm, and arrhythmic mutants have no regular rhythm (Fig. 3).

The *period* mutant also affects timing on scale greater than the 24-hour period: short-day and long-day mutants have a shorter and longer developmental time course, respectively, from egg to adult compared to normal controls. These effects of *period* occur on timescales an order of magnitude larger (10 days) or several orders of magnitude smaller (60 seconds) than the 24-hour clock. Distant as they are from the circadian timescale, they are timing mechanisms nonetheless and therefore not outside of the realm already affected by the mutants. It remains a mystery, however, as to how the relatively slow mechanism of feedback onto gene expression (Fig. 1) can also regulate a 60-second rhythm. One clue comes from a more divergent consequence of *per* and several other rhythm mutants that was revealed in pharmacological studies of cocaine responsiveness, in which they fail to become sensitized. There is no circadian or timing-dependent aspect to this phenomenon. Instead, it appears to involve a different use of their function as gene regulators: induction of the enzyme responsible for sensitization fails to occur in the mutants.

Mutants affecting learning and memory are no less wide-ranging. The *dnc* mutation produces sterility in female flies, defects in embyronic development, and shorter lifespan. The most severe *Creb* mutants are lethal in both flies and mice. Finally, $Oxt^{-/-}$ female mice are defective in lactation, as well as in infant vocalization.

The conclusion that genes producing behavioral mutants are wide-ranging in their effects (pleiotropic) is based on a large enough sampling of mutants from two different organisms to consider it likely to be the rule rather than the exception (Hall 1994, Pflugfelder 1998, Mouse Knockout & Mutation Database, http://research.bmn.com/mkmd/). It would be a mistake, therefore, to ignore it in analyses of how genes affect behavior, or in considerations of what

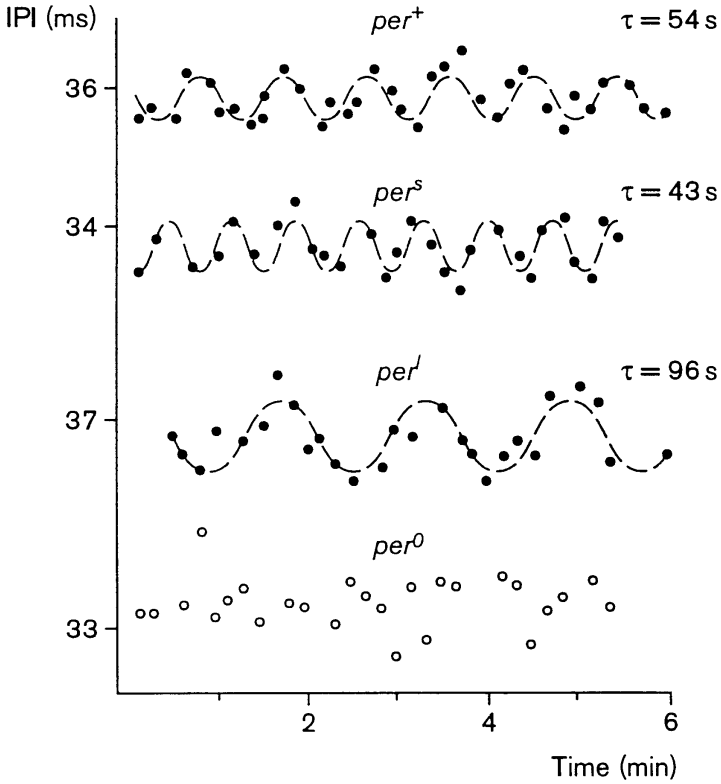

Figure 3
Rhythmic oscilliations of the coursthip song of normal (per^+) male *Drosophila melanogaster* and short-day (per^s), long-day (per^l), and arrhythmic (per^o) mutants of the *period* gene. The song consists of pulses of sound lasting 33–7 milliseconds (ms) produced by vibration of the wing. The interval between these pulses (interpulse interval, IPI) oscillates regularly with a periodicity (τ) indicated in seconds(s) (source: *Proceedings of the National Academy of Sciences USA* 1980 **77**: 6729–33. Reproduced with permission)

kinds of behavioral variation are evolutionarily permissable. A sobering conclusion that emerges from this recognition is that the path from gene to behavior is extraordinarily complex, winding its way through all possible levels of biological organization, from gene regulation through cellular biochemistry, through communication between cells, through large-scale patterns of activity in the nervous system, finally emerging as the manifest behavioral output itself.

3. *Mild Natural Variants*

Natural variation in behavior exists between and within species, and provides a major substrate for natural selection. Divergence in behavior between species is thought to be a critical component of the speciation process itself, whereas behavioral variants within a species provide it with the flexibility needed to weather the unpredictability of external conditions. Until recently, the identity and mechanistic nature of these behavioral variants has been a black box. Analyses were confined to statistical treatments of overall behavioral variation (e.g., Ehrman and Parsons 1981). A growing list of these variants has now been characterized molecularly and functionally. As foreshadowed by the experience with laboratory mutants, many have proved to be subtle variations of genes that are capable of more severe and wider ranging effects.

3.1 *Differences Between Species: Biological Rhythms*

The courtship song produced by *Drosophila* males is species-specific and plays a role in species recognition for mating (Greenspan and Ferveur 2000). The courtship song of *Drosophila melanogaster* differs from that of the closely related species *Drosophila simulans* in several parameters, the most important of which is its rhythmic oscillation (see Sect. 2.4). The *D. melano-*

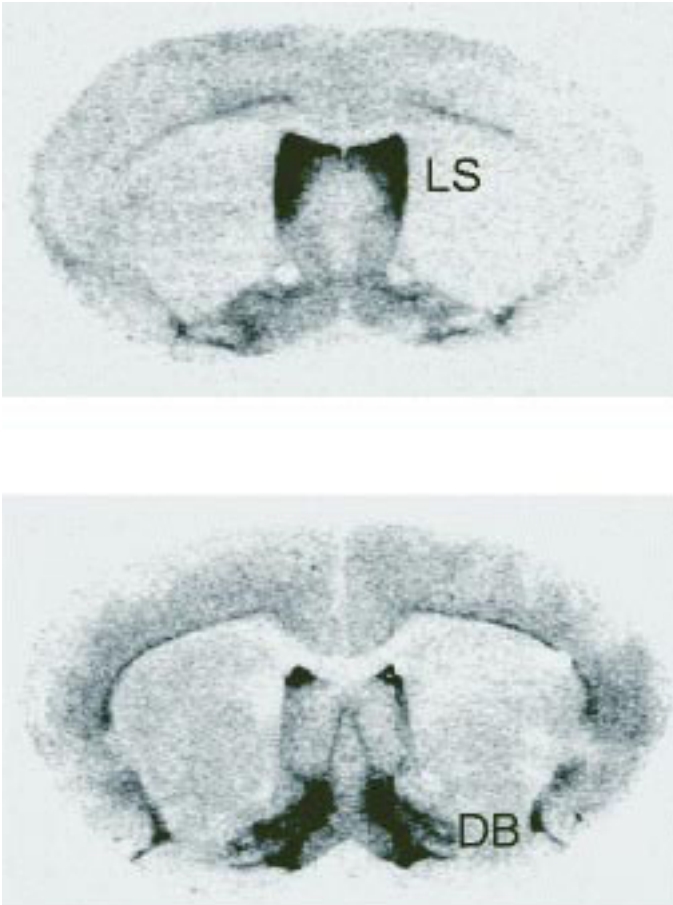

Figure 4
Anatomical distribution of the V_{1a} vasopressin receptor in the brain of the montane vole (top) and the prairie vole (bottom). Binding is more intense in the lateral septum (LS) of the montane vole but not the prairie vole, and in the diagonal band (DB) of the prairie vole but not the montane vole. Similar differences exist throughout the brain (source: *Nature* 1999 **400**: 766–8. Reproduced with permission)

gaster oscillation has a 60-second period, whereas the *D. simulans* oscillation has a 35-second period. This rhythmic difference is due to a subtle difference in the sequence of the *period* gene between *D. simulans* and *D. melanogaster*: a short segment of the protein coding region that contains repeats of the amino acid pair threonine and glycine (Fig. 3). A *D. simulans period* gene transferred into *D. melanogaster* that lack a *period* gene confers a characteristic *D. simulans* song rhythm (Wheeler et al. 1991). More specifically, a synthetic gene consisting of the entire *D. melanogaster period* gene, except for a segment of *period* from *D. simulans* consisting of a small region that is highly variable among many *Drosophila* species, also gave the same characteristic *D. simulans* song (Fig. 3). The remarkable feature of this finding is that one portion of the sequence was capable of diverging in a species-specific manner, while the rest of the gene stayed constant enough to retain the normal 24 hour circadian rhythmicity.

3.2 Differences Between Species: Social Behavior

Differences in social behavior are marked in the animal kingdom even between closely related species. Two species of North American vole illustrate this point. The prairie vole (*Microtus ochrogaster*) is monogamous, biparental, and highly social. They contrast with the montane vole (*Microtus montanus*), which is promiscuous, maternal, and minimally social (Young et al. 1998). The hormones oxytocin and vasopressin exert complementary effects on these behaviors in the

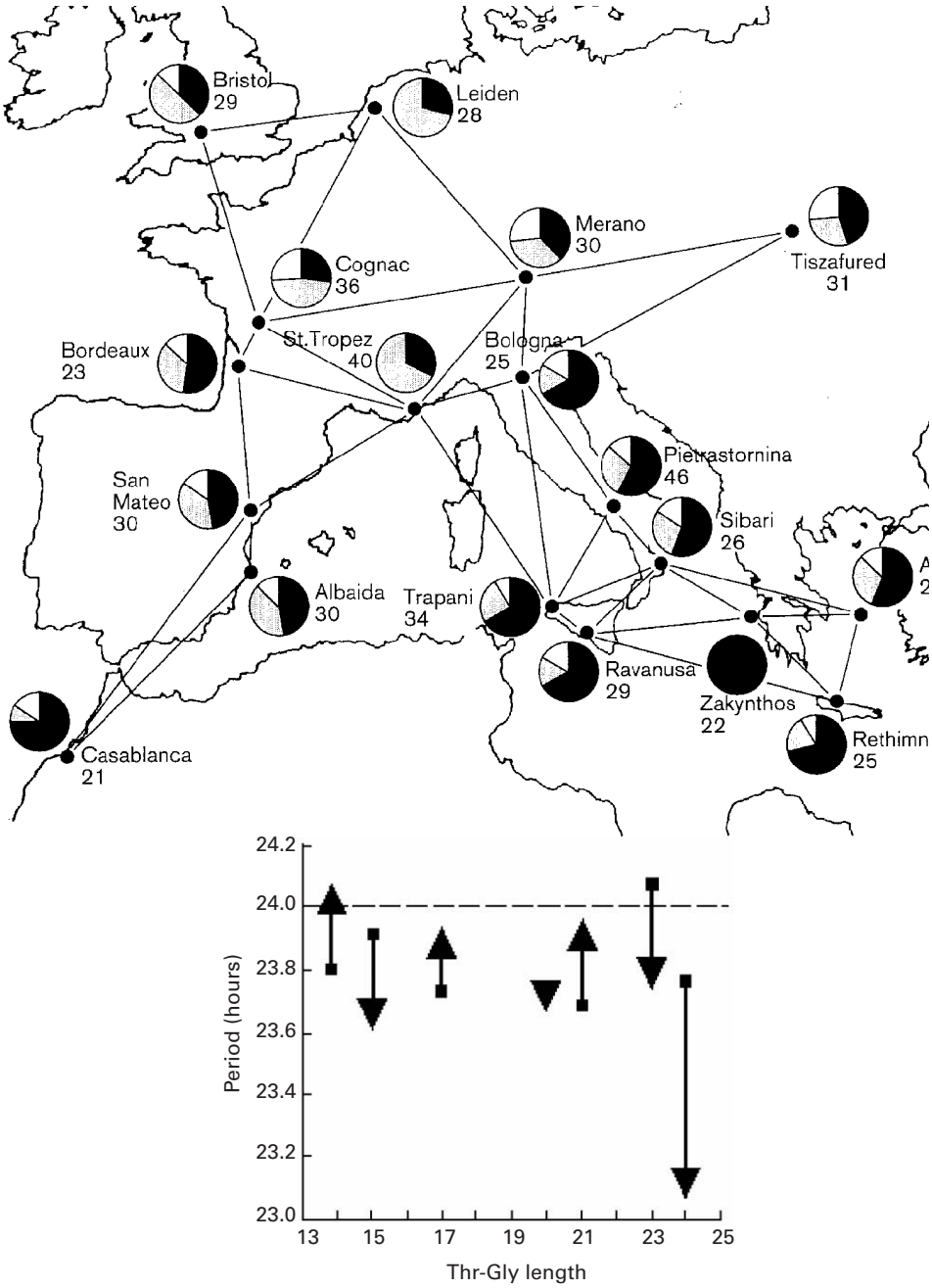

Figure 5
Top: latitudinal cline in distribution of the $(Thr\text{-}Gly)_{17}$ (dark gray) and $(Thr\text{-}Gly)_{20}$ (light gray) alleles of the *period* gene. The $(Thr\text{-}Gly)_{20}$ allele, which is more prevalent in the north, is more robust in its temperature compensation. Numbers indicate chromosomes tested (source: *Proceedings of the Royal Society of Biological Science* 1992 **250**: 43–9. Reproduced with permission). Bottom: temperature compensating ability at 18 °C (box) and 29 °C (arrowhead) of alleles of the *period* gene differing in the length of their Thr-Gly repeats. The arrow reflects the change in direction of the period at 29 °C. The $(Thr\text{-}Gly)_{20}$ variants show almost identical periods at the two temperatures (source: *Science* 1997 **278**: 2117–20. Reproduced with permission)

prairie vole, similar to their effects in rats and mice (see Sect. 2.3). Moreover, the anatomical distribution of oxytocin and vasopressin receptors in the brain differs markedly between the two vole species (Young et al. 1999).

Vasopressin administration stimulates affiliative behavior in prairie voles but not in montane voles, nor in mice. When a prairie vole vasopressin receptor gene (V_{1a}) was transferred into mice, thereby creating an anatomical distribution similar to that of the prairie vole, the mice began to exhibit affiliative behavior in response to vasopressin (Fig. 4). This result argues that changing the pattern of V_{1a} distribution is sufficient to alter this behavior. This interpretation is further supported by the fact that the V_{1a} receptor genes are virtually identical between the two vole species that differ in affiliative behavior, except for the presence in prairie voles of a small segment in the region of the gene that is likely to regulate its expression. The presence of this DNA segment correlates well with behavioral and anatomical features of two other species, the pine vole (*Microtus pinetorum*), which is like the prairie vole behaviorally as well as in its anatomical and molecular characteristics of V_{1a}, and the meadow vole (*Mictotus pennsylvanicus*), which is like the montane vole in all of these respects. Thus, it appears that subtle evolutionary tinkering with the expression pattern of the vasopressin receptor V_{1a} may account for some of the differences in social behavior between these rodent species.

3.3 Differences Within Species: Circadian Rhythms

If the genetic differences underlying behavioral variation between species are subtle, then those underlying variation within species are downright esoteric. The segment of the *period* gene that was sufficient to transfer a *D. simulans* courtship song rhythm into *D. melanogaster* also varies within the *melanogaster* species. Two versions (alleles) of the gene predominate: one containing 17 repeats of the pair of amino acids threonine and glycine $(\text{Thr-Gly})_{17}$, and one containing 20 repeats $(\text{Thr-Gly})_{20}$ (Costa and Kyriacou 1998). These two alleles are nonrandomly distributed in the wild, following a north–south gradient (known as a 'latitudinal cline') from northern Europe down to North Africa with $(\text{Thr-Gly})_{20}$ predominating in the north and $(\text{Thr-Gly})_{17}$ predominating in the south (Fig. 5). Distributions of this sort are traditionally explained as an adaptive response to climatic variation.

If these variants are undergoing selection, then the functional differences between them become relevant. A clue to such a functional difference was seen in tests of temperature compensation, one of the distinctive features of circadian rhythms which allow cold-blooded organisms to maintain their 24 hour cycle at different temperatures. Flies carrying the northerly $(\text{Thr-Gly})_{20}$ allele exhibit more robust temperature compensation than those with the more southerly $(\text{Thr-Gly})_{17}$ allele (Fig. 5). The temperature range is much greater in the north and presumably imposes greater demands on the temperature-compensating ability of the clock. The size of the effect is small, perhaps reflecting a tradeoff between the temperature compensation and conservation of the overall 24 hour rhythm.

3.4 Differences Within Species: Foraging Behavior

Most natural variation in behavior is not primarily attributable to any single gene. Where multiple genes contribute, the effect often disappears when they are separated from each other in the crosses required for genetic mapping. An exception to this trend is the *foraging* gene that affects food search behavior by *Drosophila* larvae.

Natural populations of *D. melanogaster* harbor two larval types: rovers, who search widely for food, and sitters, who do not (Sokolowski 1998). The difference is not merely locomotor because it is expressed only in the presence of food. Genetic mapping revealed it to be due primarily to a single gene, *foraging*, and molecular analysis showed it to be an enzyme involved in regulating many physiological processes, cGMP-dependent protein kinase (Fig. 6). Severe mutations of the *foraging* gene are lethal. The detailed mechanism through which the enzyme alters neuronal and behavioral activity is not known, but it appears to regulate neuronal thresholds.

A gene whose product has as wide-ranging functions as the enzyme encoded by *foraging* must have severe constraints on the kinds of variation that are viable and permissable. The explanation, in this case, appears to be that the magnitude of the natural enzyme variation is relatively small: rovers have 12 percent more activity of the enzyme cGMP-dependent protein kinase than sitters and the levels of its protein and mRNA are correspondingly higher in rovers vs. sitters (Osborne et al. 1997). This small increase thus appears to be sufficient to make a very distinctive difference in behavior, while not seriously impairing viability or obligate functions of the gene.

3.5 Genetic Basis for Selected Behaviors

Artificial selection in the laboratory has been the conventional method for detecting the presence of naturally occurring genetic variation in behavior (Ehrman and Parsons 1981). While such variation has been clearly demonstrable in virtually every behavioral selection attempted, from maze learning in rats to

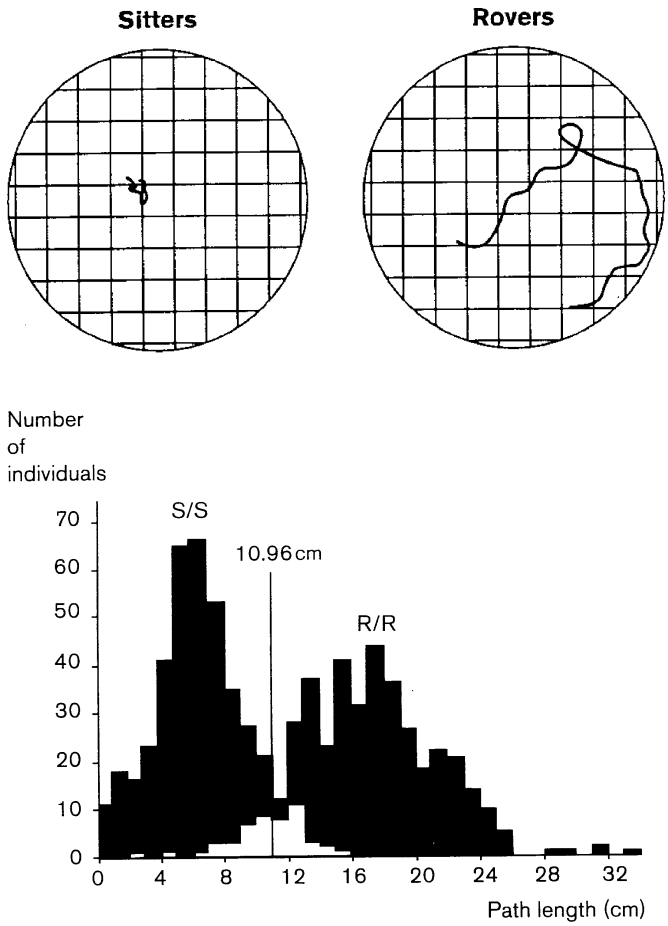

Figure 6
Top: larval foraging behavior of sitter and rover variants in *Drosophila*. Records of the trail left by larvae on a petri dish covered with yeast. Bottom: Bimodal distribution of sitter (S/S) and rover (R/R) larvae in a natural population of *Drosophila* (source: *Behavioral Genetics* 1980 **10**: 291–302. Reproduced with permission)

gravity response in *Drosophila*, there has been no way to identify any of the relevant genes. The behavioral differences are generally due to the contribution of many genes, usually of rather small effect, and thus refractory to conventional mapping techniques. Only their approximate number and putative interactions could be estimated from statistical analyses of selected lines and of F1, F2, and backcross progeny. The advent of QTL mapping (quantitative trait loci) using a widely distributed set of DNA markers has improved the prospects for mapping of such polygenic traits, but significant difficulties still attend the identification of the contributing gene itself.

The identification of natural variants in the *foraging* locus created the possibility of testing directly its response to selection under conditions likely to be relevant to its role in the wild: differences in population density. In two different experiments, strains of flies mixed for rovers and sitters were cultured at high or low larval density for over 250 generations. In both cases, the resulting high-density strains had become predominantly rover in behavior and the low-density strains predominantly sitter (Sokowloski 1998). That is, selection had eliminated the vast majority of one *foraging* allele from each population.

These findings contrast with those on intraspecies variants of the *period* gene, where the possibility of selection between alleles was inferred from their geographical distribution and made more plausible by the subtle functional differences between alleles. The conditions necessary to impose selection on the *foraging* locus are easier to imagine and implement, and the outcome leads to the conclusion that being genetically rover is advantageous under crowded

conditions, whereas being genetically sitter is advantageous under conditions of low density.

The most common sort of selection experiments have also been the most refractory to analysis. The ability to analyze the genetic basis of the strains whose behavior results from selection for variation in many genes is likely to improve with the advent of genome-wide capabilities to measure differences in gene activity (White et al. 1999). This will likely make it feasible to identify genetic differences between selected strains or natural variants based on detecting asymmetries in the activity patterns of specific genes.

4. Implications for Humans

The foregoing discussion highlights explicit and implicit aspects of the relevance to humans of gene/behavior studies in model organisms. Explicit is the conservation of function (e.g., *period* for circadian rhythms, *Creb* for long-term memory), as well as the wide-ranging nature of the genes involved. Implicit is the importance of specific context in any given gene's effect on behavior and the humility with which the question of genetic influences on behavior must be approached.

4.1 Promiscuity of Behavioral Genes

The breadth of action of the genes influencing behavior has significant implications for speculation about human behavior and its evolutionary origins. The extent to which a gene can vary is constrained by how widely its effects will radiate through the organism. For genes with wide-ranging (pleiotropic) function, this means that variants with severe effects on the gene's function are likely to have deleterious consequences. In this light, evolution looks like a balancing act, selecting for genetic variants that do the most to help the organism without doing too much harm. Sometimes this may take the form of a highly restricted alteration in a gene (e.g., anatomical distribution of V_{1a} in prairie vole vs. montane vole) and other times as a tradeoff between utility and loss of fitness (e.g., adult sluggishness in sitter). An important implication of this point is that some traits will be carried along for no reason other than their connection to some other effect of the same gene. A further corollary is that not all existing traits have been selected for themselves, and thus cannot be automatically assumed to have adaptive value.

4.2 Context

The nature/nurture dichotomy is the traditional way to describe the putative role of context in affecting gene action, but it only conveys one axis of the problem. Context is as much genetic, consisting of the particular constellation of alleles that an individual possesses, as it is nongenetic, consisting of the myriad environmental factors, both physical and social, that an individual experiences in a lifetime. Interactions between the two further complicate the matter. This means that different alleles of a given gene will not have the same effect in all contexts. A clear example of this was described earlier in the differential ability of two naturally occurring *period* alleles to compensate for temperature fluctuations (see Sect. 3.3). The $(\text{Thr-Gly})_{17}$ allele makes the individual more sensitive to its external environment than $(\text{Thr-Gly})_{20}$. This has important implications for analysis of the genetic contributions to particular behaviors (e.g., heritability).

The effect of an allelic variant on a behavior is also heavily dependent on the genetic context. This is known throughout the genetics literature as the sensitivity of a mutation to 'genetic background.' Genetic background is simply another way of describing the particular constellation of alleles at other loci present in the individual. This is exemplified in the interaction between alleles of the *foraging* locus and those of the *Chaser* locus, a gene identified based on the ability of certain of its alleles to convert sitters into rovers. In other words, mild alleles of the sort that occur naturally are even more susceptible to variations in genetic context than severe alleles.

4.3 Testability

Human genetics is primarily a correlative science. This applies as much to behavioral genetics as to any other variety. Hypotheses can be tested against observational data, but actual manipulations or 'test crosses' cannot, of course, be carried out. One of the principal virtues of model organisms, therefore, is experimental testability of gene function and interaction. Given the increasing evidence for conservation of genetic function between species as distant as *Drosophila* and mammals, they represent an important contribution to understanding and interpreting possible genetic influences on human behavior. One of these contributions should be a sense of humility before the problem. As the foregoing discussion elaborates, the road from gene to behavior is neither simple nor direct, and is difficult enough where experimental testing is possible. The difficulties are multiplied for natural variants because of the subtlety of their effects and their consequent elusive nature.

The universal importance of context in the action of genes on behavior presents further obstacles to a simple formulation. If combinations of alleles will give different effects due to interactions with each other and with environmental variables, then the problem takes on the aspect of an infinitely sliding scale. Whenever one variable is changed, the relative effect of many of the others may also change. This mallea-

bility is no doubt a source of robustness in biological systems, as well as a source for evolutionary change. Understanding it remains one of the central questions in the behavioral sciences.

See also: Adaptation, Fitness, and Evolution; Behavioral Genetics: Psychological Perspectives; Body, Evolution of; Brain, Evolution of; Comparative Method, in Evolutionary Studies; Darwin, Charles Robert (1809–82); Evolution, History of; Evolutionary Theory, Structure of; Human Evolutionary Genetics; Neurogenetics and Behavior

Bibliography

Costa R, Kyriacou C P 1998 Functional and evolutionary implications of natural variation in clock genes. *Current Opinions in Neurobiology* **8**: 659–64

Dubnau J, Tully T 1998 Gene discovery in Drosophila: new insights for learning and memory. *Annual Reviews in Neuroscience* **21**: 407–44

Dudai Y, Jan Y N, Byers D, Quinn W G, Benzer S 1976 *dunce*, a mutant of Drosophila deficient in learning. *Proceedings of the National Academy of Sciences USA* **73**: 1684–8

Ehrman L, Parsons P A 1981 *Behavior Genetics and Evolution*. McGraw-Hill Inc., New York

Greenspan R J, Ferveur J-F 2000 Courtship in *Drosophila*. *Annual Reviews in Genetics* **34**: 205–32

Greenspan R J, Tononi G, Cirelli C, Shaw P J 2001 Sleep and the fruit fly. *Trends Neuroscience* **24**: 142–5

Hall J C 1994 Pleiotropy of behavioral genes. In: Greenspan R J, Kyriacou C P (eds.) *Flexibility and Constraint in Behavioral Systems*. Wiley, New York

Konopka R J, Benzer S 1971 Clock mutants of Drosophila melanogaster. *Proceedings of the National Academy of Sciences USA* **68**: 2112–6

Osborne K A, Robichon A, Burgess E, Butland S, Shaw R A, Coulthard A, Pererira S, Greenspan R J, Sokolowski M B 1997 Natural behavior polymorphism due to a cGMP-dependent protein kinase of *Drosophila*. *Science* **277**: 834–36

Pflugfelder G O 1998 Genetic lesions in *Drosophila* behavioural mutants. *Behavioral Brain Research* **95**: 3–15

Silva A J, Kogan J H, Frankland P W, Kida S 1998 CREB and memory. *Annual Reviews in Neuroscience* **21**: 127–48

Sokolowski M 1998 Genes for normal behavioral variation: recent clues from flies and worms. *Neuron* **21**: 463–6

Wheeler D A, Kyriacou C P, Greenacre M L, Yu Q, Rutila J E, Rosbash M, Hall J C 1991 Molecular transfer of a species-specific behavior from *Drosophila simulans* to *Drosophila melanogaster*. *Science* **251**: 1082–5

White K P, Rifkin S A, Hurban P, Hogness D S 1999 Microarray analysis of *Drosophila* development during metamorphosis. *Science* **286**: 2179–84

Wilson E O 1978 *On Human Nature*. Harvard University Press, Cambridge, MA

Young L J, Nilsen R, Waymire K G, MacGregor G R, Insel T R 1999 Increased affiliative response to vasopressin in mice expressing the V1a receptor from a monogamous vole. *Nature* **400**: 766–8

Young L J, Wang Z, Insel T R 1998 Neuroendocrine bases of monogamy. *Trends Neuroscience* **21**: 71–5

Young M W 1998 The molecular control of circadian behavioral rhythms and their entrainment in *Drosophila*. *Annual Reviews in Biochemistry* **67**: 135–52

R. J. Greenspan

Genes and Culture, Coevolution of

Cultural transmission can be roughly defined as the transfer of information between individuals by social learning. Genes used to be—and for most species still are—the only means available for the accurate transfer of information across generations. In species where cultural transmission has developed, notably the human, interactions can occur between the two inheritance systems. Gene-culture coevolution refers to the evolutionary phenomena that arise from these interactions.

1. Quantifying Cultural Evolution

The principles of genetic evolution are well established. By contrast, cultural evolution is more contentious. The information that forms the basis of culture is transmitted by social learning—instruction by, or imitation of, other individuals. This information must somehow reside in the brain, but it is unclear how this is achieved. Nevertheless, just as significant insights into genetic evolution were obtained before DNA was identified as the genetic material, cultural evolution can be studied without waiting for definitive progress in cognitive science.

Central to the modern treatment of cultural evolution is the concept of a 'meme' (Dawkins 1976; also 'culturgen' of Lumsden and Wilson 1981). The meme is defined, in analogy with the gene, as a unit of information that is transmitted by social learning. Surnames are a good example of a meme that is highly congruent to a gene. Surnames are well-defined units that in many societies are transmitted from father to son, just like a gene on the Y chromosome. However, most memes cannot be delimited as precisely, so that copies of the same meme in two individuals may differ in content. Consider the ambiguity of 'how to use a personal computer.' In addition, although the meme is defined as information, it is usual to regard as a meme the more readily observable behavior that the information regularly induces.

Memes also differ from genes in that there is a multiplicity of transmission pathways. 'Vertical' transmission from parents to their children and 'horizontal' transmission between individuals of the same generation have received the most attention.

The classical, although by no means the first, quantitative study of horizontal transmission is due

two rural sociologists, Ryan and Gross (1943), who investigated the spread of hybrid seed corn. This striking technical advance in agriculture was introduced to Iowa farming communities in the late 1920s. Ryan and Gross learned through retrospective interviews when each farmer had adopted this innovation. By the time these interviews were conducted in 1941, adoption was almost complete. The cumulative number of adopters follows an S-shaped curve, which resembles the logistic curve. It is not known whether this resemblance is more than superficial.

The study of vertical transmission begins with Feldman and Cavalli-Sforza (1976). Assume two alternative memes, such as whether or not salty foods are eaten habitually. In general, denote the alternative memes by H and h. Then, the four possible combinations for husband and wife are $H \times H$, $H \times h$, $h \times H$, or $h \times h$. If children acquire their memes by vertical transmission—by social learning from their parents—the four types of families should produce different frequencies of H and h children. More specifically, if the probabilities of an H child in the four types of families are b_3, b_2, b_1, and b_0, respectively, then the inequalities $b_3 \geqslant b_2$, $b_3 \geqslant b_1$, $b_2 \geqslant b_0$, and $b_1 \geqslant b_0$ should hold. A survey of Stanford University students and their parents showed that, for the example of habitual salt use, $b_3 = 60$ percent, $b_2 = 36$ percent, $b_1 = 65$ percent, and $b_0 = 26$ percent, suggesting a fairly strong maternal influence (Cavalli-Sforza and Feldman 1981).

Cultural evolution is said to occur when the frequency (proportion) of a meme in a population changes. Memes are in principle subject to the same evolutionary forces as genes—the analog of mutation, random drift, natural and sexual selection. For example, the meme for habitual salt use may raise mortality and as a result be selected against. On the other hand, an important force acting on memes to change their frequencies, but one not normally applicable to genes, is transmission bias. The spread of hybrid seed corn in the Iowa communities is an example of cultural evolution by transmission bias, since nonadopters are more likely to be converted to adopters than vice versa.

Having briefly described cultural evolution, let us now consider in detail two well-documented instances of an interaction between genetic and cultural evolution. Many more case studies of gene-culture coevolution are discussed in books by Lumsden and Wilson (1981), Boyd and Richerson (1985), and Durham (1991).

2. *Adult Lactose Absorption and Drinking the Milk of Domesticated Animals*

Perhaps the paradigmatic example of gene-culture coevolution is the process whereby the genetic ability of adult humans to digest lactose (adult lactose absorption) became associated with their consumption of animal milk.

The fresh milk of domesticated animals, such as goats, camels, and cows, has been an important food in northern Europe and in pastoral societies for about 5,000 to 6,000 years. By contrast, animal milk is not traditionally used by hunter–gatherer and nondairying agricultural groups.

Milk contains a sugar, lactose, which must be digested by the enzyme, lactase, before it can be absorbed in the small intestine. All normal infants have high lactase activity, but during development the majority of individuals experience a drop in lactase activity, which is fully manifest by school age. Hence, adults are of two types—lactose absorbers who have high lactase activity, and lactose malabsorbers who have low lactase activity. Data from families and individuals of mixed ancestry indicate that adult lactose malabsorption is inherited as a simple Mendelian recessive.

When milk is consumed in quantity by a lactose malabsorber, the lactose passes undigested into the colon. Symptoms of intolerance such as severe diarrhea can result. Malabsorbers may also be less efficient at absorbing the calcium in milk.

Many researchers have noted the following pattern. First, populations with high frequencies of absorbers were traditionally dependent on animal milk. Second, populations not known to have used milk have low frequencies of absorbers. The highest frequencies are observed in northern Europe among the Finns, Danes, and Germans. Comparable frequencies are seen among the nomadic tribes of the north African and Arabian deserts, such as the Beja, Tuareg, and Jordanian Bedouin. On the low end are groups like the Japanese, the Pima Indians of Arizona, and the Yoruba of central Africa, none of which have a history of dairying. Although the covariation is impressive, it is not the case that all dairying populations have a high frequency of absorbers, and numerous 'exceptions' can be found around the Mediterranean. Hence, populations with a high frequency of absorbers apparently form a subset of those with a history of milk use.

Lactase activity drops at weaning in nonhuman mammals, suggesting that our ancestors were also malabsorbers. This inference and the above observations led to the proposal of the 'culture-historical' hypothesis, which posits that lactose absorbers gained a selective advantage when children and adults began to drink the milk of domesticated animals (Simoons 1970, McCracken 1971, Kretchmer 1972). Hence, the frequency of absorbers increased by natural selection in populations that had access to animal milk—or at least in some of them.

A variant of the culture–historical hypothesis is the 'calcium absorption' hypothesis (Flatz and Rotthauwe 1973). Whereas the former attributes a general nutritional advantage to the use of milk, the latter claims

that calcium was all important. Recall that high frequencies of absorbers are observed in northern Europe. Here, solar radiation is weak, and the calcium in milk may have provided protection from rickets.

A different hypothesis called the 'reverse cause' argument suggests that dairying was adopted by precisely those populations that could tolerate lactose. The genetic differentiation of human groups is assumed to have occurred before the invention of dairying by some process unrelated to milk use. In particular, drinking milk is not assumed to have conferred a selective advantage.

The two major proposals, the culture–historical hypothesis and the reverse cause argument, thus differ in the temporal priority given to cultural or to genetic change. (A third possibility is that they occur together, in which case the process would be truly coevolutionary.) However, the prevailing view in modern anthropology is that the genetic differences (if any) among human populations are not the cause of the cultural differences among those populations. In particular, the claim that a preexisting genetic contrast, such as in the prevalence of lactose absorbers, would have produced divergent cultural evolution is anathema. Hence, the reverse cause argument, in straying from dogma, has already acquired one strike against it.

More substantively, Holden and Mace (1997) have recently used phylogenetic comparative methods to show that the cultural evolution of milking probably preceded the genetic evolution of lactose absorption. In a sense, this outcome is anticipated by the fact noted above that a high frequency of absorbers is observed in only some of the dairying populations and in none of the nondairying ones. In addition, this analysis reveals that statistical evidence for the calcium absorption hypothesis is weak. Instead, the parsimonious explanation is that northern European populations are similar in having high frequencies of absorbers because they share a relatively recent common ancestral population. On the other hand, the culture–historical hypothesis is supported by a global phylogenetic analysis.

In the previous section, it was explained how a useful innovation may spread by horizontal transmission. Presumably, drinking the milk of domesticated animals can be regarded as a useful innovation, at least for absorbers. Hence, in societies where domesticated animals were available, as short a time as one generation may have sufficed for dairying to become established. If so, the claim that milk use spread first and lactose absorption evolved later may be valid.

Against this, however, the possibility must be entertained that drinking milk was an unattractive option to a society dominated by lactose malabsorbers. In other words, there may be a genetically determined transmission bias against milk use among malabsorbers. Studies conducted in modern industrial societies reveal little evidence of a clearcut difference in preference for milk between absorbers and malabsorbers. This may be because the availability of many foods, and especially the presence of lactose as an additive, makes it difficult for malabsorbers to associate the symptoms of intolerance with the ingestion of milk. By contrast, in traditional societies where food choices are limited, a difference in preference may be easier to demonstrate.

3. *Recessive Hereditary Deafness and Sign Language*

The incidence of profound childhood deafness is about one in 1,000. About one half of these cases is hereditary, with recessive genes in particular being responsible for about one third of the total incidence. Estimates of the number of recessive genes range from five to more than 30 depending on the population. Homozygosity for (having two copies of) any one of these genes results in deafness. Among the nongenetic causes are maternal rubella, neonatal jaundice due to rhesus incompatibility, and otitis media.

Sign languages are the naturally occurring forms of linguistic communication among the profoundly deaf. A sign language is not a collection of gestures. Rather, grammatical rules govern the production of words and the expression of thoughts. There are many sign languages just as there are many spoken languages. Sign languages are natural languages, whose origins cannot be determined and which have existed in the deaf communities possibly for many generations.

When deafness is recessive, an affected child may be born to normal heterozygous (having one copy of a gene for deafness) parents. In fact, about 90 percent of deaf children have hearing parents. A hearing child of hearing parents will learn the spoken language used by the parents. Vertical transmission plays an important role in the transmission of a spoken language across generations. By contrast, a deaf child of hearing parents usually cannot learn the local sign language from the parents, simply because the latter do not know it. Hence, the transmission of a sign language across generations is complicated by the fact that a significant fraction of profound childhood deafness is caused by recessive genes. Aoki and Feldman (1991), therefore, suggest that a sign language is a good example of a meme whose transmission and social distribution are subject to strong, if indirect, genetic constraints.

Another factor that interferes with the vertical transmission of a sign language is natural selection against the deaf. Selection is apparently due, not to higher mortality, but to lower marriage and fertility rates among the deaf. In the 1950s a deaf person in Northern Ireland or Japan could expect to have about 30 percent as many children as a hearing person. These observations suggest that a sign language cannot

persist under vertical transmission alone. This prediction is supported by a simple model, which is presented below to illustrate some of the principles of gene-culture coevolution.

Assume the following: (a) One gene causing deafness segregates in the population. There are two alleles (alternative states of a gene), the normal allele, and the mutant allele that causes deafness when homozygous; (b) Nongenetic causes of deafness are ignored; (c) The deaf suffer a higher mortality than the hearing, which is contrary to fact, but has approximately the same evolutionary effect as lower nuptiality or fertility. The relative viability of a deaf person is v; (d) The recessive allele causing deafness arises by mutation in each generation; (e) Deaf persons tend to marry other deaf persons. The rate of assortative mating is m; (f) Only the deaf use sign language. Sign language is transmitted vertically from signing parent(s) to deaf children. A deaf child acquires sign language with probability c, kc, or 0 depending on whether there are two, one, or no signing parents. The correspondence with the parameters of Sect. 1 is $b_3 = c$, $b_2 = b_1 = kc$, $b_0 = 0$. When the child is hearing, $b_3 = b_2 = b_1 = b_0 = 0$.

Assumptions (a)—(e) and assumption (f) determine the dynamics of the gene for deafness and the meme for sign language, respectively. Together, they constitute a model for the coevolution of genes and culture. When the strong force of natural selection and the weak force of mutation balance each other, the allele causing deafness will exist at a low equilibrium frequency. The dynamics of the frequency of signers, x, can then be approximately described by the following recursion equation, where the prime indicates the value in the next generation.

$$x' = 2kcvmx \tag{1}$$

Equation (1) is valid when the mutation rate and the frequency of signers are low, which are realistic assumptions.

Clearly, the frequency of signers will decrease and sign language will be lost if $2kcvm < 1$. On the other hand, sign language can persist if $2kcvm > 1$. Estimates of the strength of natural selection and of the rate of assortative mating are available for Japan and Northern Ireland. For example, in Japan in the 1950s, $m = 62$ percent and $v = 30$ percent. Substitution gives $2kcvm = 0.37 \times kc$. This is smaller than one, because the vertical transmission parameter, kc, cannot exceed one.

Thus, this simple coevolutionary model suggests that a sign language cannot be maintained in a deaf community by vertical transmission alone. Some linguists invoke horizontal transmission at schools for the deaf, where naive deaf children can learn from classmates who were fortunate enough to have deaf signing parents.

See also: Cultural Evolution: Theory and Models; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Memes and Cultural Viruses; Social Evolution: Overview

Bibliography

Aoki K, Feldman M W 1991 Recessive hereditary deafness, assortative mating, and persistence of a sign language. *Theoretical Population Biology* **39**: 358–72

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Cavalli-Sforza I L, Feldman M W 1981 *Cultural Transmission and Evolution*. Princeton University Press, Princeton, NJ

Dawkins R 1976 *The Selfish Gene*. Oxford University Press, New York

Durham W H 1991 *Coevolution*. Stanford University Press, Stanford, CA

Feldman M W, Cavalli-Sforza L L 1976 Cultural and biological evolutionary processes, selection for a trait under complex transmission. *Theoretical Population Biology* **9**: 238–59

Flatz G, Rotthauwe H W 1973 Lactose nutrition and natural selection. *Lancet* **2**: 76–7

Holden C, Mace R 1997 Phylogenetic analysis of the evolution of lactose digestion in adults. *Human Biology* **69**: 605–28

Kretchmer N 1972 Lactose and lactase. *Scientific American* **277**(April): 71–8

Lumsden C J, Wilson E O 1981 *Genes, Mind, and Culture*. Harvard University Press, Cambridge, MA

McCracken R D 1971 Lactase deficiency: An example of dietary evolution. *Current Anthropology* **12**: 479–517

Ryan B, Gross N C 1943 The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology* **8**: 15–24

Simoons F J 1970 Primary adult lactose intolerance and the milking habit: A problem in biologic and cultural interrelations. II. A culture historical hypothesis. *American Journal of Digestive Disease* **15**: 695–710

K. Aoki

Genes and Languages

In a study by Cavalli-Sforza et al. (1988), the spread of anatomically modern man was reconstructed on the basis of genetic and linguistic pieces of evidence: the main conclusion was that these two approaches reflect a common underlying history, the history of our past still frozen in the genes of modern populations. The expression 'genetic history' was introduced (Piazza et al. 1988) to point out that if today we find many genes showing the same geographical patterns in terms of their frequencies, this may be due to the common history of our species. A deeper exploration of the whole problem can be found in Cavalli-Sforza et al. (1994). In the following, some examples will be shown of structural analogies between linguistic and genetic geographical patterns that supply complementary information. It is important to emphasize at the outset

that evidence for coevolution of genes and languages in human populations does not suggest that some genes of our species determine the way we speak; this coevolution may simply be due to a common mode of transmission and mutation of genetic and linguistic units of information and common constraints of demographic factors.

1. *The Genetic Analysis of a Linguistic Isolate: The Basques*

The case of the Basques, a European population living in the area of the Pyrenees on the border of Spain and France who still speak a non-Indo-European language, is paradigmatic. What are the genetic relations between the Basques and their surrounding modern populations, all of whom are Indo-European speakers?

Almost half a century ago it was suggested (Bosch-Gimpera 1943) that the Basques are the descendants of the populations who lived in Western Europe during the late Paleolithic period. Their withdrawal to the area of the Pyrenees, probably caused by different waves of invasion, left the Basques untouched by the Eastern European invasions of the Iron Age. In their study of the geographic distribution of Rh blood groups, Chalmers et al. (1948) pointed out that the Rh negative allele, which is found almost exclusively in Europe, has its highest frequency among the Basques. Chalmers et al. hypothesized that modern Basques may consist of a Palaeolithic population with an extremely high Rh negative frequency, who later mixed with people from the Mediterranean area. In more recent times it was suggested (Bertranpetit and Cavalli-Sforza 1991) that the immigrants might have been the Neolithic farmers who came from the Near East and whose modern descendants have a very low frequency of the Rh negative allele.

In summary, genetic analyses have produced the following conclusions:

(a) If Basques originally settled in a large area including the Pyrenean mountains, we can postulate that the genetic differentiation of modern Basques from the neighboring French Béarn population occurred, at the latest, during the Roman period, about 2,000 years BP.

(b) On the assumption that genetic differences are proportional to the times when populations separated (which is strictly valid only when one mechanism of genetic change, technically called genetic drift, is the major cause of differentiation), it seems very plausible that genetic differentiation of Basques from European samples occurred before Celtic people settled in France and originated the French population (about 2,600 years BP). In fact recent analyses of mitocondrial DNA support the idea that the Basques are the descendants of a Paleolithic population (Richards et al. 2000).

(c) The linguistic hypothesis originally put forward by Trombetti (1926) that Basques share a common ancestry with the modern Caucasian speaking people living in the northern Caucasus (see Ruhlen 1991) is in agreement with the genetic evidence: the ancestors of modern Basques could have shared a common genetic origin with populations from the Middle East at the latest before the Neolithic introduction of agriculture in Europe (ca. 6,000–5,500 years ago).

2. *From Basque to Ligurian and Etruscan*

The population and language name 'Basque' is interesting by itself. The Basques call themselves *Euskaldunak*: the term 'Basque' derives from *Vascos*, a Spanish word. The geographical distribution of the place names ending with the suffix *asc* includes a large region where such place names are significantly represented, from the southwest of France to the Po river delta in Italy. The number of place names whose origin is documented in historical records before 1850 is about 250, and there is general agreement among linguists to associate this suffix with a pre-Indo-European *Ligurian* substratum. Place names are generally very conservative, being notoriously resistant to ethnic stratification; conquerors may impose a language on invaded populations, but usually if they modify place names, they do not change them entirely. The presence of these place names in the Alpine region between Italy and the Swiss Italian-speaking region but not in the region north to the Po River is peculiar.

The linguistic interpretation of this kind of geographical distribution—one linguistically homogeneous area surrounded by two marginal areas with the same linguistic characters—was proposed by the Italian scholar Bartoli (who together with G. I. Ascoli, Trombetti, and Terracini pioneered the new discipline called 'linguistic geography') at the beginning of the twentieth century. One of the so-called Bartoli rules says: if a central area is linguistically homogeneous and is surrounded by marginal areas which are linguistically different from the central one but equal between themselves, the language of the marginal areas is almost always older than that of the central area. In fact, the distribution of place names of Ligurian origin surrounds a region of Celtic superstratum that flooded any pre-Indo-European relic. Claimed to be among the most ancient ethnic groups of Western Europe, the Ligurians' original domain far exceeded the boundaries of the modern Italian region bearing their name.

The case of Ligurian is different from that of Basque because the latter is still spoken today while the former can possibly be detected as a linguistic relic (substratum) in the modern dialect spoken by the inhabitants of the relevant region of Italy ('Liguria'). Although the non-Indo-European nature of the Ligurian language has been debated because of the

very small number of glosses dating at least before the Celts occupied most of the Po river valley in the early sixth century BC, nobody doubts the non-Indo-European origin of the Etruscans (Pallottino 1975), and a natural question is whether we are able to recognize in today's Tuscans the fingerprints of their ancestors' genes. Apart from the number of genetic samples which have to be tested to supply enough information, the Tuscan dialect spoken today certainly cannot be defined as a non-Indo-European or a pre-Roman language: quite the reverse, it was adopted as the 'true' Italian language. What happened is that the transition from Etruscan to Latin was so difficult, the innovation so remarkable, that the process of nearly complete language substitution produced deep subconscious reactions by the speakers, and the effect of these reactions to innovation can be detected as 'fossils' in the language substratum of their descendants: the more different the languages of the conquerors and the conquered populations, the more recognizable these fossils are.

Genetics can be of great help in recognizing these fossils and possibly the population where they originated for at least two reasons. The first is that the transmission of genes is a much slower process than the transmission of a cultural trait such as a linguistic item; therefore genes can keep their identity for a longer time. Second, genetic evolution is very much dependent on population sizes. While it is perfectly legitimate to suppose that a few people can impose their language on large populations—as the Romans in fact did over the 'Romance' Europe—it is much more difficult to think of the same process in genetic terms. A few well-organized people can subdue a whole population, but it is unlikely that they would overwhelm the resident gene pool. It is reasonable to suppose that the Romans subdued the Etrurian inhabitants, but a part of the latter's probable descendants still preserve Etrurian genes. With new DNA markers, and testing of the relevant bone fossils, a better picture of the relationship between Romans and Etruscans might emerge.

3. *Co-evolution of Genes and Languages: The Origin of Indo-European*

Barbujani and Sokal (1990) found a correlation between linguistic and genetic boundaries in Europe. In the majority of cases (22 out of 33) there were also physical barriers that may have caused both genetic and linguistic boundaries. In nine cases there were only linguistic and genetic boundaries but not physical ones: three of them (northern Finland vs. Sweden, Finland vs. the Kola peninsula, Hungary vs. Austria) separate Uralic from Indo-European languages. It remains to be determined whether in these cases linguistic boundaries have generated or enhanced genetic boundaries, or if both are the consequence of political, cultural, and social boundaries that have played a role similar to that of physical barriers.

The problem of the origin of the Indo-European linguistic family and of the people speaking its languages has roused much more interest over the last years than in earlier times partly owing to the book by Renfrew (1987), who suggested that farmers, beginning to spread from Anatolia around 9,000 years ago, spoke Indo-European languages. His hypothesis was based on the suggestion originally put forward by Ammerman and Cavalli-Sforza (1984) that the spread of Neolithic farming from the Fertile Crescent was due to the spread of the farmers themselves and not only of the farming technology, and on the consideration that migrating people retain their language, if at all possible. Renfrew's hypothesis was criticized by most Indo-European linguists (for a review, see Mallory 1989, Lehmann 1993, pp. 283–8) and did not fare well when contrasted with earlier hypotheses, now identified with the name of another archaeologist, Marjia Gimbutas (1985), that Indo-Europeans migrated to Europe from the Pontic steppe area of south Russia from Dniepr to the Volga (which she called 'Kurgan' from the Russian name of mounds covering the graves), beginning with the early Bronze Age, that is, around 5,500 years ago.

Genetic data cannot give strong evidence on dates of migration, especially since the 'Kurgan' area, one of the largest pre-historic complexes in Europe, probably remained very active in generating population expansions for a long time after the Bronze Age. In that area we find at c. 6,000 years ago the Sredni-Stog culture, later (5,500–4,500 years ago) the Yamnaya cultures (formerly called pit-grave cultures) which stretched from the Southern Bug River over the Ural River and which dates from 5,600 to 4,200 years ago. From about 5,000 years ago we begin to find evidence for the presence in this culture of two and four-wheeled wagons (Anthony 1995).

Genetic data on European populations using blood typing (Piazza et al. 1995) and Y-chromosome DNA markers (Semino et al. 2000) have strongly supported a center of radiation in the Ukraine. It has been suggested (Cavalli-Sforza et al. 1994, Piazza et al. 1995) that the hypotheses of Renfrew and Gimbutas should not be treated as mutually exclusive; they may be compatible, as Schrader anticipated as long ago as 1890: 'the Indo-Europeans practiced agriculture at a site between the Dniepr and the Danube where the agricultural language of the European branch was developed' (quoted from Lehmann 1993, p. 279). The settling of the steppe by Neolithic farmers must have occurred after the beginning of their migration from Anatolia, and if the expansions began at 9,500 years ago from Anatolia and at 6,000 years ago from the Yamnaya culture region, then a 3,500-year period elapsed during their migration to the Volga–Don

region from Anatolia, probably through the Balkans. There a completely new, mostly pastoral culture developed under the stimulus of an environment unfavorable to standard agriculture, but offering new attractive possibilities. Our hypothesis is, therefore, that Indo-European languages derived from a secondary expansion from the Yamnaya culture region after the Neolithic farmers, possibly coming from Anatolia and settled there, developing pastoral nomadism.

A new analysis of the problem has been given in an unpublished analysis (Piazza et al. n.d.) of a set of lexical data (200 words) in 63 Indo-European languages published by Dyen et al. (1992). From a linguistic distance matrix whose elements are the fraction of words with the same lexical root for any pair of languages and its transformation to make the matrix elements proportional to time of differentiation, we were able to reconstruct a linguistic tree.

The root of the tree separates Albanians from the others, with a reproducibility rate (the error in reconstructing the tree) of 71 percent. The next oldest branch is Armenian. The simplest interpretation is that the language of the first migrant Anatolian farmers survives today in two direct descendants, Albanian and Armenian, which diverged from the oldest pre-Indo-European languages in different directions but remained relatively close to the point of origin.

If we give to the first split the time depth of the beginning of the expansion of the pre-Indo-European Anatolian farmers, about 9,000 years ago, we can then calculate that the origin of the European branch dates to about 6,000 years ago. The four major branches (pre-Celtic, pre-Balto-Slavic, pre-Italic, pre-Germanic) may correspond to some extent to different migratory waves, but archaeological dating is too scanty to provide unambiguous associations. It is reasonable to suggest that a first migration corresponds to the first branch, the pre-Celts (6,000 years ago, according to the tree), who settled first and went further west. Their only linguistic remnants are still alive today at the extreme of their original range. They profited from being among the first to develop an Iron Age culture, and were able to develop a wide community that spoke their language. Before Roman rule they spread to half of Europe, extending from Spain to France, most of the British Isles, northern Italy, and central Europe.

A detailed description of the other possible expansions will be found in the original paper (for a summary, see also Cavalli-Sforza 2000). In this context it is more interesting to explore how the linguistic lexical data analyzed above correlate with genetic data. The gene frequency geographical maps of the 95 blood type gene frequencies published and analyzed in Cavalli-Sforza et al. (1994) are interpolated at the geographical centers of gravity of the areas where each of the 63 languages is spoken. By taking into account the spatial autocorrelation of both—linguistic and genetic—sets of data, the following results are obtained: (a) the correlation between genetic and linguistic variables is 69 percent and is statistically significant; and (b) since there is a possible effect of the geographical distance in explaining this correlation, the partial correlation between linguistic and genetic data, given constant geographical distances, has been calculated. The resulting value of 27 percent is still statistically significant and gives evidence for a correlation between genes and languages in Europe that is not only due to the effect of the geographical distance.

4. *Towards a Global Perspective*

More than 5,000 languages are spoken today in the world, and it does not take a linguist to recognize that some languages are more closely related than others due to history. The official origin of historical linguistics can be dated to 1786, when the English judge Sir William Jones advanced the idea that Sanskrit, a classical language in India, Greek, Latin, and possibly Celtic and Gothic (the ancestor of Germanic languages) shared a common origin. These old languages were the first members of a family of languages that would become known as the 'Indo-European' family (or 'phylum'). As Indo-European is the earliest and best studied linguistic family, coevolution of genes and languages has been documented. Since the eighteenth century, however, many other linguistic families or superfamilies have been recognized. The most complete classification on a world basis was proposed by Ruhlen (1991) who lists 17 linguistic families (Khoisan, Niger-Kordofanian, Nilo-Saharian, Afro-Asiatic, Caucasian, Indo-European, Uralic-Yukaghir, Altaic, Chukchi-Kamchatkan, Dravidian, Sino-Tibetan, Austric, Indo-Pacific, Australian, Eskimo-Aleut, Na-Dene, Amerind).

The reconstruction of the relationships above the family level is hotly debated among historical linguists who have yet to agree on the existence of a single tree linking all the existing language families, that is on the possible differentiation of modern languages from a single ancestor language. Even unification at a lower level such as that of the (pre-Columbian) American languages proposed by Greenberg (1987), who grouped them into just three macro-families (Eskimo-Aleut, Na-Dene, and Amerindian), has been strongly opposed by the majority of American linguists. Interestingly, Greenberg's proposal seems to agree with the analysis of genetic markers in extant Native Americans (Cavalli-Sforza et al. 1994) and these three families seem to identify three major migrations suggested by archaeological data. Amerindian speakers appear to have come first (between 30,000 and 15,000 years ago according to genetic data), followed by Na-Dene

speakers and finally Eskimo-Aleut (both in a period between 15,000 and 10,000 years ago). It must be said, however, that at a finer level of classification contemporary Amerindian speakers show high genetic variability, and this is not easy to reconcile with linguistic taxonomy.

Even without an agreed genealogy of the linguistic families covering all tongues spoken today, it is relevant to note the impressive one-to-one correspondence of the genetic phylogeny of the world populations with the classification into the 17 large linguistic families listed above (Cavalli-Sforza et al. 1988). This correspondence is expected because there are important similarities between the evolution of genes and languages. In either case: (a) a change which first appears in a single individual can subsequently spread throughout the entire population (for genes they are called mutations; they are rare, are passed from one generation to the next and can, over many generations, eventually replace the ancestral type; linguistic innovations are much more frequent and can also pass between unrelated individuals); and (b) the dynamics of change is affected by the same demographic pressures, isolation, and migration. Two isolated populations differentiate both genetically and linguistically because isolation, which could result from geographic, ecological, or social barriers, reduces the likelihood both of marriages and cultural exchanges and, as a common result, reciprocally isolated populations will evolve independently and gradually become different. Both genes and languages will drift apart regularly over time, the former slowly, the latter much more quickly.

In principle, therefore, the linguistic tree and the genetic tree of human populations should agree since they reflect the same history of population splitting and subsequent independent evolution. The different rate of change, however, is a major source of divergence: one language can be replaced by another in a relatively short time. In Europe, for example, Hungarian is spoken in a land surrounded by Indo-European speakers but it belongs to the Finno-Ugric subdivision of Uralic. At the end of the ninth century AD, the nomadic Magyars left their land in Russia and invaded Hungary. The number of conquerors was probably less than 30 percent of the conquered population so that their genetic contribution was limited, but they imposed their language on the local Romance-speaking population. Today all Hungarians speak a Uralic language, but barely 10 percent of their genes can be attributed to the Uralic conquerors. Generally it is intuitive that the total substitution of one language for another occurs more easily under the pressure of a strong political power of the newcomers, as witnessed in the Americas. The case of Basques, on the other hand, shows that separate languages spoken in nearby countries can remain relatively unaffected for thousands of years, even when their genes experience a partial substitution.

It is remarkable that, despite the above sources of confusion, the correlation between genes and languages has been maintained through the centuries until today and is still statistically significant. The ties between biology and linguistics were already evident since the times of Darwin, who in chapter XIV of his *The Origin of Species* wrote:

> If we possessed a perfect pedigree of the mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one. Yet it might be that some ancient language had altered very little and had given rise to few new languages, whilst others had altered much owing to the spreading, isolation, and state of civilization of the several co-descended races, and had thus given rise to many new dialects and languages. The various degrees of difference between the languages of the same stock, would have to be expressed by groups subordinate to groups; but the proper or even the only possible arrangement would still be genealogical; and this would be strictly natural, as it would connect together all languages, extinct and recent, by the closest affinities, and would give the filiation and origin of each tongue.

The increasing resolving power of modern genetic data makes it possible to follow Darwin and to use the genetic phylogeny of our species to infer the earliest branches of a hypothetical linguistic tree. The most comprehensive genetic phylogeny reconstructed in Cavalli-Sforza et al. (1988) was used by Ruhlen (1994) to draw the tree of origin of human languages (some reference dates from genetic and archaeological evidence have been added). The oldest linguistic families must be African: Khoisan is probably the oldest and Afro-Asiatic the most recent, while Niger-Kordofanian and Nilo-Saharian believed by some linguists to descend from an ancestor tongue, and the Congo-Saharan, were probably spoken at an intermediate time. A more exhaustive discussion of this hypothetical tree can be found in Cavalli-Sforza (2000). As the genetic data improves with the inclusion of more representatives from those geographical areas of the world where the sampling is still scanty, the tree will be more complex but it is likely that its main features will remain unchanged.

In conclusion, our present genome keeps the record of its past evolution with an impressive richness of detail that is also reflected by our languages. Genes and languages contribute to the understanding of human history by highlighting human diversity; both are instrumental in giving some of the silent voices of our past a chance to be heard.

See also: Cultural Evolution: Overview; Evolution and Language: Overview; Languages: Genetic Classification; Linguistics: Overview

Bibliography

Ammerman A J, Cavalli-Sforza L L 1984 *Neolithic Transition and the Genetics of Populations in Europe*. Princeton University press, Princeton, NJ

Anthony D W 1995 Horse, wagon & chariot: Indo-European languages and archaeology. *Antiquity* **69**: 554–65

Barbujani G, Sokal R R 1990 Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences* **87**: 1816–9

Bertranpetit J, Cavalli-Sforza L L 1991 A genetic reconstruction of the history of the population of the Iberian peninsula. *Annals of Human Genetics* **55**: 51–67

Bosch-Gimpera A 1943 El problema de los origines vascos. *Eusko-Jakintza* **3**: 39

Cavalli-Sforza L L 2000 *Genes, Peoples, and Languages*. North Point Press, New York

Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

Cavalli-Sforza L L, Piazza A, Menozzi P, Mountain J 1988 Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences* **85**: 6002–6

Chalmers J N M, Ikin E W, Mourant A E 1948 Basque blood groups. *Nature* **162**: 27

Gimbutas M 1985 Primary and secondary homeland of the Indo-Europeans. *Journal of Indo-European Studies* **13**: 185–202

Greenberg J H 1987 *Language in the Americas*. Stanford University Press, Stanford, CA

Lehmann W P 1993 *Theoretical Bases of Indo-European Linguistics*. Routledge, London

Mallory J P 1989 *In Search of the Indo-Europeans: Language, archaeology and myth*. Thames and Hudson, London

Pallottino M 1975 *The Etruscans*. Penguin Books, New York

Piazza A, Cappello N, Olivetti E, Rendine S 1988 A genetic history of Italy. *Annals of Human Genetics* **52**: 203–13

Piazza A, Rendine S, Minch E, Menozzi P, Mountain J, Cavalli-Sforza L L 1995 Genetics and the origin of the European languages. *Proceedings of the National Academy of Sciences* **92**: 5836–40

Piazza A, Minch E, Cavalli-Sforza L L n.d. The Indo-Europeans: Linguistic tree and genetic relationships. *Antiquity*

Renfrew C 1987 *Archaeology and Language: The Puzzle of Indo-European Origins*. Cambridge University Press, New York

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Gölge M, Dimitrov D, Hill E, Bradley D, Romano V, Calì F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Oppenheim A, Norrelletto A, Nørby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozzari R, Torroni A, Bandelt H-J 2000 Tracing European founder lineages in the near eastern mtDNA pool. *The American Journal of Human Genetics* **67**: 1251–76

Ruhlen M 1991 *A Guide to the World's Languages*. Stanford University Press, Stanford, CA

Semino O, Passarino G, Oefner P J, Lin A A, Arbuzova S, Beckman L E, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti A S, Cavalli-Sforza L L, Underhill P A 2000 The genetic legacy of paleolithic *Homo sapiens sapiens* in extant Europeans: A Y chromosome perspective. *Science* **290**: 1155–9

Trombetti A 1926 *Le origini della lingua Basca*. Bologna, Italy

A. Piazza

Genetic Counseling: Historical, Ethical, and Practical Aspects

Genetic counseling is one of the most important direct applications of human genetic knowledge. Genetic counseling refers to the totality of activities that (a) establish the diagnosis of genetic disease, (b) assess the occurrence risk, (c) communicate to the patient and family the chance of recurrence, (d) provide full and unbiased information within a caring relationship regarding the many problems raised by the disease and its natural history, including the potential ethical conflicts as well as the medical, economic, psychological, and social burdens, and (e) provide information regarding the reproductive options to be taken, including prenatal diagnosis, and refer the patients to the appropriate specialists. Genetic counseling should offer guidance but will allow those being counseled to arrive at their own choices, both before and after genetic procedures (Vogel and Motulsky 1997; World Health Organisation [WHO] 1998).

1. *Paradigm Development and Concepts of Genetic Counseling*

Pre-scientific knowledge of heritability in humans dates back to the early Greek physicians and philosophers and it can be assumed that there have been practical measures, including advice to individuals and couples as to how to use such knowledge (Vogel and Motulsky 1997). The idea of an improvement of mankind by rationalized mate selection was a major aspect of ancient and Renaissance state and social utopias (Plato, *Politeia* (about 374 BC); Thomas More, *Utopia* (1516); Fra Tomaso Campanella, *The City of the Sun* (1623)). In the medical literature of the eighteenth and nineteenth centuries, reports on the familial recurrence of particular disorders begin to accumulate. An early and noteworthy example is the book of the English physician Joseph Adams (1756–1818) *A Treatise on the Supposed Hereditary Properties of Diseases* (1814), where an attempt is made to lay out the fundamentals of genetic counseling in affected families.

Towards the end of the nineteenth century, two paradigms established human genetics as a science. Gregor Mendel (1822–1884) founded the concept of the gene as the basic factor of inheritance (1866), while Francis Galton (1822–1911) introduced statistical

Table 1
Paradigm shifts in applied human genetics

Paradigm	Basis	Goal	Measure
Eugenics	Emotionally carried social utopia	Improvement of the gene pool	Directiveness, compulsion, coercion
Prevention	Rational calculation	Reduction of individual suffering, economic distribution of resources	Paternalism
Psycho(socio)-biology	Complex interaction between counselor (doctor) and client (patient)	Help in individual decision making	Empathy, nondirectiveness

methods to investigate and interpret the familial clustering of traits and thus founded the biometrical paradigm. Until the present, biometrics and Mendelian genetics have remained the basis of risk determination as the naturalistic component of genetic counseling. Galton coined the term 'eugenics' and favored measures to enhance propagation of individuals with supposedly high genetic qualities (positive eugenics). In Germany, around the turn of the nineteenth to the twentieth century, Wilhelm Schallmeyer (1857–1919), Alfred Hegar (1830–1914), and Alfred Ploetz (1860–1940) developed the idea of negative eugenics, a concept of decreasing the number of offspring of individuals with supposedly below average hereditary qualities, or of preventing congenitally diseased individuals altogether, with the aim of improving the gene pool of future generations and especially that of the presumed superior Nordic race (see Ploetz 1895).

In many countries during the first half of the twentieth century, positive and negative eugenics formed the ideological basis of state-enforced measures to exert reproductive control over individuals. Between 1911 and 1930, 24 US states proclaimed sterilization laws and eugenic thought also played a role in the enactment of restrictive immigration laws. In Europe, Denmark was the first country to proclaim a Law on Genetic Hygiene (1929), in which voluntary sterilization for genetic reasons was regulated. In Germany, the Law for the Prevention of Congenitally Diseased Progeny (1933), followed by a series of other laws and decrees, mandated and organized the sterilization of individuals with certain diseases, which were thought to be genetically determined. In Germany, eugenics, racial hygiene, and National Socialist ideology formed a gloomy alliance. The involvement of German geneticists in coercive measures and eventually the Holocaust to sections of the population that were declared to be inferior has been well documented by Müller-Hill (1997). It is beyond doubt that this historical burden strongly influences the current debate on ethics and genetics in Germany and elsewhere.

From about 1930 onwards, the eugenic movement outside Germany began to lose power, but it was not until the 1950s that the population-directed eugenic orientation was slowly replaced by a medically oriented practice with genetic disease prevention as a major focus. Nevertheless, eugenic thinking, whether acknowledged or not, has continued to play a role until the present time, especially outside the Western world (Wertz 1997). When, in 1947, Sheldon Reed coined the term 'genetic counseling,' it became clear that eugenics and what was then understood to be genetic counseling, were no longer identical. According to Reed, 'genetic counseling is a type of social work entirely for the benefit of the whole family without direct concern for its effect upon the state or politics' (1974). This definition marks a change of paradigm in applied human genetics, away from the concern about the genetic make-up of the population and towards the attitude that decision-making with regard to genetic risks should be left to the individual, even if this may create or increase social costs or worsen the gene pool. Thus, one of the fundamental ethical principles in medicine, the respect for the autonomy and self-determination of persons, is also the cornerstone of genetic counseling.

Genetic counseling has thus gone through three overlapping historical phases (Table 1). First, the eugenics phase, mainly rooted in England, the United States of America, Scandinavia, and Germany, was based on the social and biological utopia of general mental and physical health and the optimization of humankind. In order to attain these goals, depending on political circumstances, directive measures including direct and indirect coercion were applied. Eugenics was scientifically ill-founded, and the definition of inferior or superior genetic qualities a matter of ideology. Second, the preventive-medical phase laid its emphasis on the reduction of individual suffering through the prevention of genetic diseases and disabilities, in order to economize rare medical resources. It was founded on the rational calculation that any means towards that end would find acceptance as being in the immediate interest of the individual. Strategies were those of benevolence, authority, and paternalism. Third, in the psycho-(sociobio-)logical phase, the focus of genetic counseling is to help the individual or the family to manage a problematic

situation caused by genetic disease. The basis of this help is a complex interaction between the counselor (a specialized physician or a non-medically trained expert) and the client (patient) in a nondirective process of communication. Help is provided with the understanding that genetic diseases are often associated with special psycho-social problems and ethical conflicts which deserve to be managed individually. Thus, any direct or indirect coercive influence is to be avoided. Nondirectiveness has become a basic concept of genetic counseling, to which a majority of counselors worldwide feel obliged (Wertz 1997).

The history of genetic counseling shows that its purpose has always been value-determined and has never been seen outside the context of social politics in which it was embedded. In countries that acknowledge human rights as defined in the democratic state theories of the eighteenth century, the only defensible position is that of the autonomy of the individual to make decisions regarding genetic counseling and genetic diagnosis, and to draw consequences for life and family planning. The healthcare system, on the other hand, carries the ethical obligation to provide full information and the tools that may enable the individual, depending on the state of knowledge, the emotional situation, and moral positioning, to make use of that information.

1.1 Excursus: Ethics and Genetic Counseling: The Respect for Personal Autonomy

Today the term 'eugenics' is sometimes used to describe the cumulative result of individual choices after genetic counseling, especially in view of reproductive choices. The decreasing prevalence of births of children with Down Syndrome in Western Europe, especially in the UK, may be regarded as a result of individual choices to undergo prenatal diagnosis and subsequently to terminate a pregnancy after a positive test result. Whether this type of individual eugenics is socially desirable and a benign development has been discussed with some controversy by Wertz (1998). Critics argue that free, autonomous choices (after genetic counseling) not only depend on surrounding social, cultural, and other factors, but also on the full and honest presentation of information in genetic counseling, i.e., on access to valid information. In the absence of correct and unbiased information, free choices are not possible and are compromised, and information and genetic counseling becomes a means of furthering attitudes and goals of health professionals. A worldwide survey of 2,091 genetics professionals in 36 nations conducted by Wertz (Nippert and Wolff in Germany) suggests that eugenic thought underlies the perception of the goals of genetics and genetic counseling in many countries. In developing nations of Asia and Eastern Europe especially, directive counseling is practiced to influence individual decisions. Intentionally pessimistically biased information is routinely provided after positive results from prenatal diagnosis in order to avoid the birth of affected children (Wertz 1998, Cohen et al. 1997). The data from this survey suggests that the concept of nondirective counseling found in English-speaking nations and Western Europe is an aberration from the rest of the world.

2. Indications for Genetic Counseling

In medicine in general, a measure is 'indicated,' when, after weighing up the risks and benefits, it appears medically required. This definition is hardly applicable to genetic counseling, because the recognition of risks and benefits of knowledge, as well as any actions that may result from such knowledge, will emerge only in the counseling process itself, and their appraisal is often more strongly determined by individual life circumstances than by medical judgment. On the other hand, the professionalization of genetic counseling, as a specialization within medicine, and jurisdiction in regard to correctness and completeness of any medical advice, have brought about a core service that every counselee may duly expect to receive: A scientifically founded statement regarding the etiology and the (recurrence) risk of a genetic condition, comprehensive information and counseling regarding diagnostic options, and referral to preventive, therapeutic, and/or socially alleviating measures.

Generally speaking, an indication for genetic counseling is given whenever an individual regards it as a means to help solving a personal problem related to an existing or suspected genetic risk. Table 2 summarizes the most frequently encountered situations. With increasing awareness of old and novel pre- and postnatal methods of genetic diagnosis, genetic counseling is increasingly sought by individuals with genetic risks not immediately apparent to themselves. Concerns,

Table 2
Indications for genetic counseling

1. Birth of a child with congenital disease or disability
2. One or more relatives suffering from (suspectedly) genetic disease or disability
3. Affected counselee
4. Consanguineous couple
5. Recurrent miscarriage
6. Infertility
7. Mutagenic influences
7.1 Advanced maternal age
7.2 Pre- or postconceptional mutagen exposure
8. Teratogenic exposure
8.1 Infections
8.2 Drugs
8.3 Alcohol
8.4 Radiation

including moral doubts, may have been raised, for example, by offers of prenatal screening (maternal serum marker and ultrasound screening), or carrier detection, within or outside the framework of a defined screening programme, because such offers may bring up the question of termination of the pregnancy.

3. The Counseling Process

According to a statement issued by an Ad hoc Committee on Genetic Counseling of the American Society of Human Genetics, genetic counseling is a communication process which:

> deals with the human problems associated with the occurrence or risk of recurrence of a genetic disorder in a family. This process involves an attempt by one or more appropriately trained persons to help the individual or the family to 1. comprehend the medical facts, including the diagnosis, the probable course of the disorder, and the available management; 2. appreciate the way heredity contributes to the disorder and the risk of recurrence in specified relatives; 3. understand the options for dealing with the risk of recurrence; 4. choose the course of action which seems appropriate to them in view of their risk and their family goals and act in accordance with that decision; 5. make the best possible adjustment to the disorder in an affected family member and/or to the risk of recurrence of that disorder (Fraser 1974; Ad hoc Committee on Genetic Counseling 1975).

The counseling process is usually initiated by clarifying the counselee's motivation for consultation. What is the background to the problem, and what are the expectations? The counselor should explain, in general terms, what genetic counseling can achieve, and where its limits lie. This is important, because expectations are frequently unrealistic. It needs to be explained, for example, that it is technically not possible to exclude any kind of genetic risk, or to guarantee a healthy child. The next step is to record the personal and family medical history in the form of a genealogical tree, encompassing at least all first and second degree relatives. This is followed by providing information regarding the most important medical facts concerning the disease in question (natural history of the disease, the likely degree of suffering it may confer, its treatability) as well as the likelihood of its occurrence or recurrence. It is important to give unbiased, objective information. During the later phases of the counseling process, the counselor will often return to this information in order to relate it to the personal views of the counselee. The 'prior' risks for a genetic condition as deduced from its pattern of inheritance (monogenic dominant or recessive, autosomal or sex-linked, co-determined by several genes and/or the environment) may be modifiable by genetic diagnostic tests so that a 'posterior,' usually more precise, risk figure emerges. Ever more often, with the Human Genome Project nearing its completion and concomitant technical developments becoming routine practice, firm confirmation or exclusion of carrier status for any particular genetic disposition will be the rule, rather than the exception. Nevertheless, the phenotypic outcome will often be variable, due to environmental factors and complex interactions between genes. In the counseling process, the resulting uncertainty needs to be explained and worked upon with great care. Without thoroughly understanding the chances and limits of genetic testing, an informed consent to any such procedure cannot be attained. Informed consent is generally held to be the prerequisite for any genetic testing. The decision whether to make use or not of diagnostic offers can only be made sensibly in a highly personalized context, so that genetic counseling is required to be nondirective (Wolff and Jung 1995).

During the counseling process it may be necessary to make a physical examination of the counselee and/or of family members in order to search, for example, for early signs of manifestation of a late-onset disorder or for microsymptoms of a disease. The impact of such examination needs to be explained beforehand, because counselees and their family members may wish to help clarify genetic risks for others, but not necessarily for themselves. It is sometimes difficult to decide to what extent information from the family history which may confer a genetic risk to the counselee needs to be discussed, if this information has not been central to the counselee's motivation for consultation. This is particularly delicate when no therapeutic or preventive options exist. If they do, however, the counselee has a right to comprehensive information and counseling.

4. The Determination of Genetic Risks

The determination of genetic risks in any individual case is the essential naturalistic component of genetic counseling. It requires expertise in general and human genetics, including formal (mathematical) genetics, population genetics, and clinical genetics, as well as an awareness (attainable in this fast-moving field only through continuous further self-education) of novel developments, especially in the differential genetic diagnosis. In the following discussion, we provide a summary of the most important groups of conditions where a risk assessment is often sought and a brief outline of risk provision. Often, the prior genetic risks need to be modified by using information from the pedigree or genetic tests (conditional probabilities). By employing Bayesian methodology this information can be used to arrive at posterior (final) risks (Harper 1998).

If a disease is (mainly) due to the effect of a single gene, the phenotype may be inherited in a recessive or a dominant fashion. The disease may manifest itself in a sex-dependent way, if the causative gene is not

located on the autosomes (the chromosomes shared by both sexes in equal number) but rather on the sex chromosomes (X and Y). In an autosomal-recessive condition, the recurrence risk in a sibship is 1/4, but usually low (less than 1 percent) for second-degree relatives. If a parent has an autosomal-dominant condition, there is a 50 percent chance that each child will also carry the disease-causing mutation. If neither parent has a condition that is known to be autosomal-dominant, the recurrence risk in a sibship is usually negligible, because the presumption is that a spontaneous mutation has occurred in the germ-line leading to the affected individual. In X-linked recessive disease, if the mother is a carrier, half of all her sons will develop the full clinical picture, and half of her daughters will be carriers. Mothers and daughters are usually unaffected or exhibit symptoms to a lesser degree (due to non-randomness of X-inactivation during early development). In Y-linked disease (such as sub-fertility), the trait is passed on from fathers to sons. In both recessive and, most notably, dominant conditions, the inheritance patterns may be obscured by variable penetrance (proportion of carriers manifesting the phenotype at all) and expressivity (degree of disease severity). Greatly reduced penetrance and expressivity, i.e., where only a small fraction of carriers suffer from the condition, leads to the concept of genetic predisposition (susceptibility) for disease, which may be modulated by other genes or environmental factors. Polygenic and multifactorial inheritance are the etiology of most of the common diseases (such as coronary heart disease, obesity, diabetes mellitus) and the majority of cases of birth defects. In these categories, empirical risk figures apply. Numerical chromosomal aberrations are usually sporadic, with a low empirical recurrence risk in a sibship, but may also run in families, e.g., structural aberrations, such as translocations, which may be inherited as balanced or unbalanced in offspring in a more or less apparent Mendelian fashion.

5. *Genetic Counseling in the Social Context*

How have genetic counseling and its supposed high standards fared in common practice? Throughout the world and even in Europe, there is a huge variance in the professional identity and training of providers of genetic counseling services. In Germany, most counseling services are provided by physicians who are specialists in medical genetics (Nippert et al 1997) and in the UK, genetic counseling is provided primarily by nurses (Raeburn et al 1997), whereas in the US and Canada a substantial amount of genetic counseling is provided by counselors trained to the master's degree level. This implies that practice differs between countries. Surveys which compare the attitudes of genetic service and counseling providers have revealed a remarkable difference in attitudes and values. An international survey carried out in 37 nations in 1994–1995 found that, in most of the surveyed countries, genetic counseling, especially in the context of prenatal diagnosis (PD), is directive in that the counselor provides the client with intentionally slanted information in order to influence the client's decision (Wertz 1998). The survey shows that, depending on the diagnosed conditions of a fetus, on the culture's view of the condition, and on the counselor's personal view about the morality of abortion, he or she deliberately provides more optimistically or pessimistically biased information.

Although published policies and guidelines exist (Fraser 1974, German Society of Human Genetics 1996, Andrews et al 1994, WHO 1998, Fryer and Lister Cheese 1998, Holtzman and Watson 1997), the type of information and the way information is communicated vary within countries, and standard practices can hardly be found. Obviously genetic counseling practices are shaped by the characteristics of the various healthcare systems, as well as by the personal characteristics of the counselor, such as sex, professional expertise, and type of employment. The process and outcomes of genetic counseling have been documented by a few empirical studies (Sorenson 1992). The outcomes have been defined primarily on the basis of patient knowledge and patient satisfaction, but so far have hardly paid any attention to the social aspects of the information (Clarke et al 1996). Concern has been voiced that in the future more counseling will be provided by professionals not trained in genetics. Although there is insufficient empirical evidence that genetic counselors successfully achieve autonomous, informed decision-making, research from Europe and the US suggests that non-geneticists do not set a high value on providing information that allows informed decision-making before undergoing tests such as prenatal diagnosis (Dodds 1997).

With the development of more new molecular tests, implementation of these tests in various fields of medicine will drive the diversification of genetic counseling practice. Neither general healthcare providers nor non-genetics specialists have sufficient professional education and expertise in genetics, molecular testing, and genetic counseling to perform this task (Holtzman and Watson 1997). How to target the education for providers via medical curricula and continuing education remains an unsolved challenge. So far, no country has developed any strategy or has made any far-reaching efforts. Genetic specialists are arguably the most valid source for providing the needed expertise and information for their medical colleagues to facilitate their understanding of basic genetic principles and the application of molecular testing, including pre- and post-test counseling, in clinical practice. However, the more genetics becomes part of general medicine and the more genetic testing is introduced into the various specialty areas, the more the task to provide information and to interpret test

results will fall on primary care physicians and non-genetics specialists. Probably the future of genetic services will be provided in the way envisaged by Wertz:

> It is perhaps more likely that most genetic services will be provided in the same manner as other types of medicine. In managed care systems this may mean hurried, assembly-like encounters that focus on tests and treatments rather than on providing information and counseling. Most primary care providers may be ill-equipped to provide information and counseling needed in any case … Many will wonder why genetics was ever expected to produce superior provider-patient relationships, to uphold superior ethical standards, or to allow greater patient autonomy than other fields of medicine (1997, p. 343).

Most likely the future of genetic testing and counseling will be shaped by economic interests, because of the need for financial returns on heavy investments made in the field of molecular testing, by reimbursement practices, and by the extent to which national healthcare policies will encourage genetic testing. It remains unclear whether the main ethical concepts attached to genetic counseling, such as nondirectiveness, individual autonomy, and greater equality in the doctor-patient relationship, will survive or will wither away during the integration of genetic tests into routine medical practice. Hopefully genetic counseling will retain some of its best elements, namely respect for patients' rights to valid, non-slanted information and for patient autonomy in reproductive decision-making.

See also: Ethical Dilemmas: Research and Treatment Priorities; Ethical Issues in the 'New' Genetics; Ethics and Psychiatry; Eugenics, History of; Genetic Screening for Disease-related Characteristics; Genetic Testing and Counseling: Educational and Professional Aspects; Risk Screening, Testing, and Diagnosis: Ethical Aspects

Bibliography

Ad hoc Committee on Genetic Counseling of the American Society of Human Genetics 1975 Genetic counseling. *American Journal of Human Genetics* **27**: 240–2

Andrews L B, Fullarton J E, Holtzman N A, Motulsky A G (eds.) 1994 *Assessing Genetic Risks. Implications for Health and Social Policy*. National Academy Press, Washington, DC

Clarke A, Parsons E, Williams A 1996 Outcomes and process in genetic counseling. *Clinical Genetics* **50**: 462–9

Cohen P E, Wertz D C, Nippert I, Wolff G 1997 Genetic counseling practices in Germany: A comparison between East German and West German geneticists. *Journal of Genetic Counseling* **6**: 61–80

Dodds R 1997 *The Stress of Tests in Pregnancy: Summary of a National Childbirth Trust Screening Survey*. National Childbirth Trust, London

Fraser F C 1974 Genetic counseling. *American Journal of Human Genetics* **26**: 636–59

Fryer A E, Lister Cheese I A F 1998 *Clinical Genetic Services: Activity, Outcome, Effectiveness and Quality*. Report from the Clinical Genetics Committee of the Royal College of Physicians of London, The Royal College of Physicians of London, London

German Society of Human Genetics 1996 Position paper, http://www.gfhev.de

Harper P 1998 *Practical Genetic Counseling*. 5th ed. Butterworth-Heinmann, Oxford, UK

Holtzman N A, Watson M S 1997 *Promoting Safe and Effective Genetic Testing in the United States. Final Report of the Task Force on Genetic Testing*. Johns Hopkins University Press, Baltimore. http://www.med.jhu.edu/tfgtelsi/

Human Genome Project, http://www.ornl.gov/TechResources/Human_Genome/home.html

Müller-Hill B 1997 *Murderous Science: Elimination by Scientific Selection of Jews, Gypsies and Others in Germany 1933–1945*. Oxford University Press, Oxford, UK

Nippert I, Horst J, Schmidtke J 1997 Genetic services in Germany. *European Journal of Human Genetics* **5** (suppl 2): 81–8

Ploetz A 1895 Racial hygiene. In: Weingart P, Knoll J, Bayertz K (eds.) *Rasse, Blut und Gene*. Suhrkamp, Frankfurt/Main, Germany

Raeburn S, Kent A, Gillot J 1997 Genetic services in the United Kingdom. *European Journal of Human Genetics* **5** (suppl 2): 188–95

Reed S C 1974 A short history of genetic counseling. *Social Biology* **21**: 332–9

Sorenson J R 1992 What we still don't know about genetic screening and counseling. In: Annas G J, Elias S (eds.) *Gene Mapping: Using Law and Ethics as Guides*, Oxford University Press, New York, pp. 203–14

Vogel F, Motulsky A G 1997 *Human Genetics*. Springer, Berlin Heidelberg, New York

Wertz D C 1997 Society and the not-so-new genetics: What are we afraid of? Some future predictions from a social scientist. *Journal of Contemporary Health Law Policy* **13**: 299–346

Wertz D C 1998 Eugenics is alive and well: A survey of genetic professionals around the world. *Science in Context* **11**: 493–510

Wolff G, Jung C 1995 Nondirectiveness and genetic counseling. *Journal of Genetic Counseling* **4**: 3–25

World Health Organisation 1998 *Proposed International Guidelines on Ethical Issues in Medical Genetics and Genetic Services*. World Health Organisation Human Genetics Programme, Geneva. http://www.who.int/ncd/hgn/hgnethic.htm

J. Schmidtke, I. Nippert, and G. Wolff

Genetic Engineering

The term 'genetic engineering' stands for human alteration of the genetic code of an organism, so that its biosynthetic properties are changed. The major applications are for the industrial production of desired peptides or proteins, or to alter the biological capabilites of the organism. These techniques have

been used to develop crops with agronomically useful changes, such as pest resistance and ripening properties that allow for shipment. Surprising results have been obtained by silencing genes in experimental organisms, as well as the production of animal models of human disease by deriving strains of animals with mutated human genes. Much attention is being given to the production of products and methods to modify an individual's cells to treat human disease. Although the majority of activity at the present time is directed toward traits governed by a single gene, commercial and academic investigators are hopeful that complex diseases and behavior can also be modified.

1. Biological Background

The technology has developed in the context of our understanding of genetics. A vital property is the production and maintenance of variability in populations. One of the mechanisms that produces such variability is recombination. Information for a number of peptides as well as the controlling mechanisms for their synthesis are inherited together on a single chromosome. In diploid organisms, exchanges between homologous chromosomes is an important source of variation. This process occurs in the context of DNA replication, including the enzymatic unwinding of the double helix and synthesis of the new, complementary chain of nucleic acids.

Cellular properties may be altered by external mechanisms. One of the first systems that was extensively studied, and provided the tools that are currently used, is the phenomenon of infection of bacteria by bacteriophages, the bacterial viruses. The usual direction of events is followed. The message is coded in the DNA sequence. It is then transcribed into the sequence of the intermediate messenger RNA molecules, and then translated into protein. In the lytic cycle, the viral DNA instructions are used to synthesize new viral DNA and protein that can assemble into new viral particles that destroy the cell and are then released into the environment to infect other host cells. In the lysogenic cycle, the DNA of the virus integrates into the host DNA. In addition, there are a number of other particles called plasmids, which can transfer genetic information from one bacterium to another. They are small, circular DNA molecules that replicate independently of the DNA of the host bacterium. Their role is particularly prominent in the transmission of antibiotic resistance. Most animal viruses are similar to the lytic cycle of bacteriophage, however, there are a number of viruses that can become latent. There is also agents like the retroviruses, whose genome gets transcribed into the host DNA, thereby changing cellular properties. Some of these agents will also incorporate host DNA into their sequences, and then capable of transmitting this sequence into a new host.

The tools that allow us to transfer genes from one organism to another are the *restriction* endonucleases. These enzymes are found naturally in bacteria and protect them from infection by breaking down foreign DNA. They are chemically very specific, only working on specific palindromic sequences. Consequently, the kind of infection discussed in the last paragraph can only occur if the bacteriophage sequence does not contain elements that will be broken down by the host bacterial system of endonucleases. A number of different enzymes are found in different bacterial species, and the enzymes referred to by an abbreviation of the bacterial species name and a number. Similarly, there are enzymes that are used to connect fragments, the ligases (Griffiths et al. 1996).

2. Cloning Genes

All of the manipulations of genetic engineering require muliple copies of the DNA sequence or gene of interest. The original methods of getting multiple copies relied on bacteriophage or plasmid vectors to introduce the foreign DNA into bacteria to produce these copies, as each modified cell produces multiple copies, and the bacterial culture itself increases. This is done by first physically isolating the vector, opening its DNA with a restriction enzyme and binding in DNA from the organism being studied that has also been cleaved with a restriction endonuclease. A new population of bacteria is then infected with the altered vector. Given an appropriate way of selecting the population of bacteria so that it uniformly has the DNA of interest multiplying within, one can isolate a large population of vector molecules with the desired sequence, which is then freed by enzymatic cleavage once again.

Fragments of DNA are identified by physically separating them by electrical charge and molecular weight through gels. The DNA of the vectors and bacteria are generally in the range of one to ten thousand base pairs, and there are a sufficiently small number so that the fragments can be identified with a simple staining technique, usually a compound that binds to DNA and fluoresces under ultraviolet light. The larger quantity of fragments that would be isolated from more complex organisms produces a smear with such dyes, so the base-pairing property of DNA, the obligate pairing of adenine with cytosine and guanine with cytosine that allows for both recognition and synthesis of the linear sequence, is used to identify the same sequence on the gel by labeling a known fragment with an isotope or fluorescent dye. The labeled molecules are called probes. This is also the basis for identifying genetic variation in organisms, either for basic studies or identification of mutations associated with disease.

Isolation of fragments produced by digestion with several enzymes, used both singly and in combination,

allows for the construction of a physical, restriction-fragment map. Smaller fragments may be replicated, followed by the chemical analysis of the base sequence within fragments which are then assembled into the final base sequence of the gene. Once the sequence is known, production of useful amounts of a region of DNA may now be done enzymatically *in vitro* with the polymerase chain reaction (PCR). In this technique, the region between two primers, one from each strand of the final DNA molecule is copied in a logarithmic fashion by a heat-resistant DNA polymerase from a small amount of genomic DNA (it has been done with single cells), using multiple heating and cooling cycles. This technique is also used in diagnostic work (Strachan and Read 1996).

3. *Modifying Cells to Produce Proteins*

A number of strategies have been used to produce large amounts of a peptide or protein. Multiple techniques have been necessary because of the different kinds of cells that have been used. The production of human insulin, one of the earlier products produced commercially, is a useful model that is an example of expression cloning. The sequences for the A and B chains were introduced into separate vectors in sequence with a bacterial promoter (a signal to turn a gene on) and the enzyme β-galactosidase that produces colored bacterial colonies. This arrangement is necessary because bacteria will not just produce the human protein if only the structural gene is introduced. The bacteria are tricked into producing a fusion protein: a bacterial β-galactosidase with a human insulin chain sequence at the end. After selecting the colonies by color, mass cultures are produced from which the fusion proteins can be purified. The normal insulin dimer is then produced chemically by removing the β-galactosidase, purifying the chains, combining the chains and allowing them to refold and form the sulfide bridges between the two chains.

Expression cloning in bacteria works well if the peptide or protein is composed entirely of the amino acid sequence. Many functional proteins contain specific carbohydrate groups at precise locations on the chain that are attached after translation. These modifications are different in different classes of organisms, so techniques using yeast, plant, cultured insect, other mammalian, or human cells have been used to make the complete molecule. Another problem that is addressed in this approach is the elimination of bacterial toxins that cannot be removed from the desired protein product mixture. The engineered DNA can be introduced into these cells by a variety of methods. The physical techniques like electroporation, which uses an current to move the DNA inside, or coating various kinds of particles, like gold, with the DNA and then bombarding the cells to get the DNA inside has a chemical parallel in the use of calcium salts and heat. Artifical yeast chromosomes have been produced to move gene constructs. For the cultured insect cells, modified Bacloviruses are used (these are viruses that have invertebrate hosts). In mammalian systems, several viruses have been used, including retroviruses (the class of agents involved in acquired immunodeficiency and tumor transformation in experimental organisms), adenoviruses (agents that typically produce respiratory disease, however, some strains are capable of producing more severe systemic disease), and herpes simplex.

The goal of expression cloning is to produce large amounts of protein that can be purified more easily, and in much larger amounts than is possible using chemical techniques starting with cells from the normal tissue. On the research laboratory scale, the three-dimensional structure of the whole protein or the active region can be determined by a combination of nuclear magnetic resonance and x-ray crystallography. This information is useful for understanding function or designing small molecules that are expected to have pharmacological actions. This is then one element of the strategy in rational drug design. If the protein is to be used as a product, the growth of the modified cells has to be done on an industrial scale that is an engineering problem similar to that used in the production of a number of antibiotics (Glick and Pasternak 1998).

4. *Modifying Organisms*

An organism that has been engineered so that it has a gene from another species integrated into its genome is called a transgenic organism. These techniques have been used to produce plants and animals expressing mutated genes of that species to study function, with mutated or genes introduced from another species with desired agronomic properties, as models of human diseases, and to produce therapeutic products. There is a large and continuously growing literature of studies in the mouse.

Two procedures have been widely performed in mice: the interruption of a mouse gene to eliminate its function, for which the term knockout has been used, and introduction of a transgene. The 'standard' procedure employs classical, cellular, and molecular genetic techniques. The gene to be studied is first cloned, and then introduced into a targeting vector in such a way that the coding sequence of the gene is interrupted by the neomycin resistance gene. The vector also has the thymidine kinase gene from herpes. Embryonic stem cells are obtained from a brown mouse and placed in culture. The cells are then exposed to the vector. There are three possible outcomes. The vector does not insert at all, the whole construct inserts at random in a chromosome, or only the desired coding sequence and the neomycin resistance gene insert into the normal location on the chromosome.

Treating the cells with a neomycin analogue and the antiviral agent ganciclovir will kill all of the cells except the ones with the integrated sequence. The surviving stem cells are injected into the blastocyst stage embryo of a black mouse and placed in a surrogate mother. If a chimera is formed, it will have a coat of two colors. The chimeras are mated with black mice, and their brown offspring tested for the modified gene. Those that are heterozygous are mated, and one of four will be homozygous for the modification.

In addition to the introduction of genes into stem cells by vectors, injection of the desired sequence has been done by some of the physical methods discussed earlier such as electroporation, coated gold particles, and salt precipitation. Some transgenics have been derived from direct injection of the plasmid into the egg.

The production of genetically identical organisms, cloning in the original sense, was first done in amphibians by nuclear transplantation into fertilized ova from which the original nucleus was removed. This was done widely in mammals by using embryonic cells as the nuclear donor. The production of the sheep named Dolly was noteworthy because the donor nuclei were obtained from adult mammary cells that were maintained in cell culture. This precipitated a round of discussion about the feasibility and the desirability of doing this in human beings. The claim was made that this would be an alternative for infertile couples, and therefore acceptable. It is necessary to point out to people who would propagate themselves for narcissistic reasons that the derived individual would not be absolutely the same because of the developmental environment, both before and after birth (Hubbard and Wald 1997).

Another possible use of this technology that has been proposed is the therapeutic cloning of replacement tissues or organs. Cells would be biopsied from the eventual recipient and placed in culture. Nuclei from these cells would be transplanted into eggs from which the nuclei have been removed; these eggs will be allowed to develop to the blastocyst embryonic stage for the isolotion of embryonic stem cells that are treated to produce the desired cell types. This avoids the problem of using fertilized eggs, but still involves the use of human eggs and early embryonic stages (Gurdon and Colman 1999).

5. *Gene Therapy*

The goal of gene therapy is to provide a functional protein to an individual whose genotype leads to a disease because the protein is missing or modified so that it is not functional. There are two potential modes. In germ-line therapy, the modification would be heritable, and the individual's offspring potentially would not be affected. In somatic cell therapy, only the treated individual would have the modification. There have been many arguments against germ-line therapy including our inability to know whether potentially harmful changes in the sequence have been made inadvertantly. All of the trials to date have been in somatic cells for the treatment of inherited disease and cancer by control of cell death.

Two strategies are being attempted. For those disorders for which bone marrow transplantation has been effective, the gene is introduced into marrow cells, the population of which is then expanded and infused into the patient. The techniques for gene introduction are similar to the ones previously discusssed: viral vectors and chemical or physical techniques. An immune deficiency disorder, adenosine deaminase deficiency, was the first attempt reported. Although the modified cells express the enzyme, repeated infusions of treated cells have been necessary, and the patients have continued to receive chemically modified enzyme, so it is not clear whether or not gene therapy alone is sufficient.

Other trials have used viral vectors to treat target organs, such as the lung in cystic fibrosis. The therapeutic trail of an adenovirus-mediated transfer of a gene to the liver has received general media attention because of the death of a participant. Evidently, other participants in gene therapy trials have died, although it is not clear how many of those deaths are a result of the procedure (Strachan and Read 1996).

6. *Applications to Behavior*

The application of these techniques in behavior is loaded with scientific, ethical, and legal complications. Part of the problem stems from the very human desire to have a simple answer to a complex problem. Genetic, environmental, and developmental factors interact to produce behavior, and it is difficult to distinguish the role of each component. One should ask if it is even desirable to do this, because of the mixture of potential benefit and risk. It is clear that each behavioral trait is heterogeneous. There may be families where a single major gene is responsible, and a particular drug, expressed protein or even gene therapy could work, but in the more usual instance of several genes and environmental factors, it is unlikely that such a unitary approach would work. The promise of a solution to disease problems by definition of molecular mechanisms has been difficult to achieve. We still do not have an answer in the case of sickle cell anemia, the first disorder to be described in this manner. Common behavioral traits would be more difficult because of the interactions.

Because of the Human Genome Project, it is likely that a number of genes will be discovered that contribute to the development of a number of behavioral characteristics. The technology therefore will create the possibility of diagnostic tests for the presence of variants at these genetic loci that might be involved in diagnostic confirmation, or even asympto-

matic screening without necessarily having the ability to safely and ethically treat those in whom the genetic variant is identified. Furthermore, since it is well known that the genotype does not always predict the phenotype even for such things as emphysema associated with a variant α-1-antitrypsin, one should question this approach with respect to behavioral characteristics (Hubbard and Wald 1997).

Using genetic information to design pharmacologically active agents will probably be useful for drugs interacting with receptors influencing neural function. A similar problem arises here. Will the intervention be beneficial to someone who is suffering or will it be used to control individuals who are designated by those in authority to be deviant? Modifying the environment to enhance development is effective and poses less of a biological risk than drugs or gene therapy (Duyme et al. 1999). The relative economic costs are unknown, but the human ramifications are immense.

See also: Ethical Issues in the 'New' Genetics; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Testing and Counseling: Educational and Professional Aspects; Genetics and Development; Genetics and the Media; Genetics of Complex Traits Through the Life Cycle; Human Genome Project: German Perspective; Human Genome Project: Japanese Perspective; 'New' Eugenics, The: Role in Society

Bibliography

Duyme M, Dumaret A-C, Tomkiewicz S 1999 How can we boost IQs of 'dull children'?: A late adoption study. *Proceedings of the National Academy of Science USA* **96**: 8790–4

Glick B R, Pasternak J J 1998 *Molecular Biotechnology: Principles and Applications of Recombinant DNA*. 2nd ed. ASM Press, Washington, DC

Griffiths A J F, Miller J H, Suzuki D T, Lewontin R C, Gelbart W M 1996 *An Introduction to Genetic Analysis*. 6th ed. W H Freeman, New York

Gurdon J B, Colman A 1999 The future of cloning. *Nature* **402**: 743–6

Hubbard R, Wald E 1997 *Exploding the Gene Myth: How Genetic Information is Produced and Manipulated by Scientists, Physicians, Employers, Insurance Companies, Educators and Law Enforcers*. Beacon Press, Boston

Strachan T, Read A P 1996 *Human Molecular Genetics*. Wiley-Liss, New York

D. J. Harris

Genetic Factors in Cognition/Intelligence

Research has not only established the importance of both genetic and environmental influences on human intelligence, but has also begun to examine how genes shape intellectual development and the relationship among different cognitive skills (e.g., verbal and performance ability). Additionally, researchers have begun to search for DNA markers associated with human cognitive ability. This article will not only examine these new research directions, but will also discuss the implications of these findings in the context of the environment.

1. Behavioral Genetics, Genes, and Intelligence

One of the major misconceptions about behavioral genetics is that the purpose of this research is to examine whether genes *or* the environment is operating. Instead, behavioral genetics assumes that differences in intelligence, for example, are due to differences in genes, the shared family environment (the environment that family members share, such as socioeconomic status), the nonshared environment (the environment that makes family members different, such as differential parenting and different experiences), and error. Genetic influences on intelligence are statistically significant, averaging around 50 percent when collapsing across all available twin and adoption studies (see Brody 1992 for a review). In other words, genes account for roughly half of the measured variability in intelligence.

2. Beyond Heritability

While a heritability of 50 percent is important in terms of the nature/nurture debate, finding that intelligence 'is both' genetic and environmental (the other 50 percent has to be due to nongenetic influences) in and of itself does not elucidate *how* genes influence the development of human intelligence. Behavioral genetics has begun to move beyond heritability and, as a result, has reported findings that fundamentally change how we think about the development of intelligence.

2.1 Genes Become More, Not Less Important With Age

A common assumption is that genes become less important with age, as environmental influences 'add up' to shape cognitive performance. Behavioral genetic results suggest that the opposite is the case. Although genes account for 50 percent of individual difference in intelligence collapsed across all studies, heritability appears to vary with age, starting at around 40 percent in early childhood, rising to 60 percent in early adulthood, and to 80 percent in later life (see Plomin et al. 1997). More recent evidence suggests that the genetic influences on intelligence may decline to around 60 percent in old age (Finkel et al. 1998, McClearn et al. 1997).

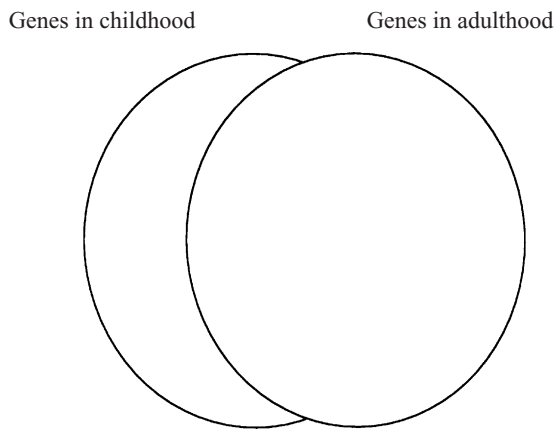

Figure 1
Genes are largely shared across development, but there is also genetic specificity across different ages

2.2 Genes Drive Stability, But Are Not Invariant Across Development

Another new research question addressed by genetic methods is the extent to which the genes that influence cognitive abilities at one age are the same genes that influence cognitive abilities at later ages. Are genes relating to intelligence immutably 'on' from birth, or do different genes turn on and off across development? Results (see Plomin et al. 1997) suggest that genes are largely stable, with some genetic change across development (see Fig. 1). In contrast, the environment is largely responsible for the changes in intelligence across development.

2.3 Genes Also Drive The Correlation Among Different Aspects of Cognitive Ability

One of the most interesting 'new directions' in behavioral genetics research involves the examination of how genes influence different aspects of cognitive ability. Are the genes that influence verbal skills the same as those that influence spatial ability or are verbal and spatial skill each influenced by different genetic processes? This debate, often called the molarity vs. modularity debate, has important implications for cognitive neuroscience. A molar system is one in which different aspects of cognitive ability, such as verbal ability, spatial ability, perceptual speed, and memory, are all influenced by a common set of genes. A modular system, in contrast, holds that genes function independently across different cognitive skills. In multivariate behavioral genetics, this degree of overlap is often measured by a genetic correlation. A high genetic correlation suggests almost complete overlap while a low genetic correlation suggests genetic independence. Behavioral genetic studies suggest that the genetic correlation approaches 1.0, suggesting almost complete genetic overlap (see Petrill 1997). The genes that influence spatial ability, for example, are largely the same genes that influence verbal ability (see Fig. 2). Genes are the 'tie that binds' different cognitive skills while the discrepancy between different cognitive skills (e.g., certain people may be better at verbal as opposed to spatial skills) is due largely to the environment.

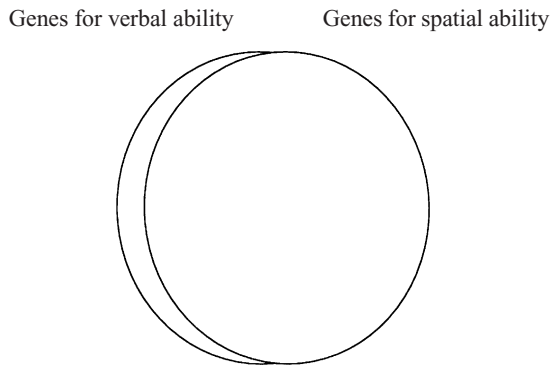

Figure 2
Genes are largely shared across different measures of cognitive ability

2.4 Molecular Genetics and Intelligence/Cognition

One of the most quickly developing areas of behavioral genetics involves the search for DNA markers associated with human cognitive ability. For example, the APOE-4 marker is associated with a four-fold increase in risk for developing late-onset Alzheimer's disease (Breitner et al. 1998). Another notable example is the continuing research on the association between DNA markers on Chromosome 6 and reading disability (see Gayan et al. 1999).

While these markers are important in that they show that differences in the human cognitive system are associated with specific DNA markers, these studies do not necessarily show that individual differences in healthy, normally distributed cognitive functioning are influenced by these same DNA markers. A study called the IQ-QTL Project is attempting to search for DNA markers associated with intelligence. In 1998 this study yielded the first replicated DNA marker associated with intelligence (Chorney et al. 1998) and is expected to complete a full genome scan, likely yielding other markers for intelligence.

3. Implications

Taken together, these data not only suggest that genes influence the development of intelligence, but that intelligence develops differently than once thought.

Genes become more, not less important as the age of the sample increases. Genes influence the correlation among measures of intelligence, both across time and across different measures of cognitive performance. Additionally these behavioral genetic results, although indirectly measured through twin and adoption studies, are beginning to be tied to specific DNA markers. What are the implications of these data?

3.1 DNA Markers Will Likely Be General In Effect

Although the search for DNA markers is in its infancy, it is likely that the first decade of the twenty-first century will yield a set of DNA markers that account for a significant proportion of the variance in intelligence. Given the fact that twin and adoption research suggests a great deal of genetic overlap, it should be expected that DNA markers will cut across different cognitive skills, and will be stable over time. This 'molar' focus will become increasingly important as researchers turn from simply trying to find markers associated with cognitive ability to attempting to delineate the mechanism by which genes influence complex cognitive skills.

3.2 The Environment Is Also Important

A significant heritability estimate, genetic correlation, or DNA marker associated with intelligence does not mean that intelligence cannot be modified by the environment. For one, the environment is an important determinant of individual differences in intelligence, especially in childhood (the heritability of intelligence is roughly 40 percent, which means that the other 60 percent has to be due to nongenetic influences). More importantly, although heritability describes why individuals differ within a group, it does not examine average increases in ability or average differences in ability between groups. Although genetic influences account for 90 percent of the individual differences in height, for example, the height of the average person has increased, presumably due to an average improvement in the environment. Similarly, although the heritability of IQ has remained stable across the 80+years that twin studies have been conducted, the measures used to measure IQ have been restandardized numerous times. The average IQ, although set to be 100 on a particular test, increases over time. Put another way, if a random sample of children in 1999 were given an intelligence test standardized in 1920, the average IQ would be significantly greater than 100 (Flynn 1998). Thus, behavioral genetic methods tell us 'what is' for a specific population at a specific time, not 'what will be' or 'what should be' (Plomin and Petrill 1997).

In fact, some of the most important discoveries about the relationship between the environment and intelligence have come through behavioral genetic research. As stated earlier, behavioral genetics separates the environment into that which is shared by family members and that which is not shared by family members. Thus, behavioral genetic methods allow the examination of environmental effects that operate between families (shared environment), and environmental effects that operate within families (nonshared environment). The same behavioral genetic studies that suggest increasing heritability and high genetic correlations also suggest that the most pervasive environmental influences across the lifespan are nonshared. In other words, the most important environmental effects are those that make family members different from one another. Behavioral genetic studies are currently attempting to identify these nonshared environmental influences.

4. Summary and Conclusions

Genetic influences *are* important to understanding intelligence. However, the environment is also important. Intelligence tests are not a veridical window into genetic potential nor are they indices of environmentally driven learning. Intelligence is defined by measures of behavior and is thus subject to both genetic and environmental effects. The key issue is understanding how genes and environment work together to yield cognitive development.

Bibliography

Breitner J C S, Jarvik G P, Plassman B L, Saunders A M, Welsh K 1998 A risk of Alzheimer disease with the epsilon 4 allele for apolipoprotein E in a population-based study of men aged 62–73 years. *Alzheimer Disease and Associated Disorders* **12**: 40–4

Brody N 1992 *Intelligence*, 2nd edn. Academic Press, San Diego

Chorney M J, Chorney K, Seese N, Owen M J, Daniels J, McGuffin P, Thompson L A, Detterman D K, Benbow C, Lubinski D, Eley T, Plomin R 1998 A quantitative trait locus associated with cognitive ability in children. *Psychological Science* **9**: 159–66

Finkel D, Pedersen N L, Plomin R, McClearn G E 1998 Longitudinal and cross-sectional twin data on cognitive abilities in adulthood: The Swedish adoption/twin study of aging. *Developmental Psychology* **34**: 1400–13

Flynn J R 1998. IQ gains over time: Toward finding the causes. In: Neisser U (ed.) *The Rising Curve*. American Psychological Association, Washington, DC

Gayan J, Smith S D, Cherny S S, Cardon L R, Fulker D W, Brower A M, Olson R K, Pennington BF, DeFries J C 1999 Quantitative-trait locus for specific language and reading deficits on chromosome 6p. *American Journal of Human Genetics* **64**: 157–64

McClearn G E, Johansson B, Berg S, Pedersen N L, Ahern F, Petrill S A, Plomin R 1997 Substantial genetic influence on cognitive abilities in twins 80 or more years old. *Science* **276**: 1560–3

Petrill S A 1997 Molarity versus modularity of cognitive functioning? A behavioral genetic perspective. *Current Directions in Psychological Science* **6**: 96–9
Plomin R, DeFries J C, McClearn G E, Rutter M 1997 *Behavioral Genetics*, 3rd edn. W. H. Freeman, New York
Plomin R, Petrill S A 1997 Genetics and intelligence: What's new? *Intelligence* **24**: 53–78

S. A. Petrill

Genetic Screening for Disease-related Characteristics

1. Introduction

Although one could imagine genetic screening for human characteristics that are not disease-related, in what follows we will restrict ourselves to disease-related screening. Furthermore, although disease-related screening could serve other purposes than the benefit of the participants, e.g., scientific or economic motives, we consider here only screening that has as its main goal the interest of people. So we place genetic screening in the medical perspective, where it is a component of community genetics, the application of medical genetics to the benefit of large numbers of people.

2. Definition

Genetic screening is the application of a test on people for the systematic early detection or exclusion of a hereditary disease, a genetic predisposition to a disease, or to determine whether a person carries a predisposition which may produce a hereditary disease in their offspring (Health Council of the Netherlands 1994). Testing is not restricted to the study of DNA or chromosomes, but includes the use of any test which is designed to detect and exclude hereditary disease, predisposition, or carrier state. Early detection means that the testing is applied to those who have not (yet) sought medical aid because of physical signs, symptoms, or anxiety. So it is typically the care system that takes the initiative. The word 'systematic' in the definition implies that screening is offered to all people in a predefined target group, e.g., neonates or pregnant women.

2.1 Terminology

Genetic screening shares its terminology with epidemiology and genetics. A test result is called positive when it points to the presence of the feature sought after. Similarly, a result is said to be negative when it indicates the absence of the feature. Both results, however, may be true or false. A test result, for instance, is called false positive if it indicates that the feature looked for is present while in reality it is absent. The probability that a certain test result is true or false depends on the prevalence of the feature and two test characteristics: sensitivity and specificity. Sensitivity refers to the proportion of cases in which the presence of the feature is identified by the test. Specificity is the proportion of cases in which the absence of the feature is correctly identified by the test. The predictive value of a test result, either positive or negative, indicates how often that test result is correct.

As to the terminology of genetics the reader is referred to other articles. An exception is made for the term 'penetrance.' Penetrance is the frequency of cases in which a particular genotype comes to expression.

3. Characteristics of Genetic Screening

As with other types of medical screening, diagnostic confirmation of a positive screening test is an integral part of the screening procedure. Unlike other types of screening, however, genetic screening looks for stable characteristics of the participants. So in general, once an individual is tested, repetition of the test is not needed. In screening for infectious diseases or cancer, the majority of participants, namely the true negatives, the false positives, and the cured, have to return for subsequent testing after a set time period.

The stability of the genetic characteristics also makes it possible to predict disease long before it would become manifest without intervention. In principle this lead time (time until disease becomes manifest) can be lifelong, or with prenatal diagnosis even longer. It is clear that this is an unwanted situation unless interventive measures are more effective when they are started early.

A further characteristic of genetic screening that differentiates it from other types of medical screening is that a positive test result may frequently have consequences for the family members. This does not apply to all types of genetic screening. For instance, finding a fetus with a regular trisomy 21 Down syndrome does not increase the risk for sisters of the pregnant woman, since this is a spontaneous mutation. But finding through screening that someone carries, for example, two mutations of the gene involved in hemochromatosis puts his or her siblings immediately at a 25 percent risk of having the same mutations (Worwood 2000). They may be grateful to know this since manifestations of this iron overloading disease can be prevented effectively by regular bleeding, but the point here is that one's decision to participate in a genetic screening program is to some extent a decision involving family members who might have preferred to make such choices themselves.

4. Target Groups

The target groups for genetic screening may consist of individuals or couples. Screening newborns for phenylketonuria (PKU) or young adults for hemochromatosis are examples where the individual is the target. In both cases the goal of identifying these persons is to offer timely intervention (dietary measures prevent the development of mental retardation in PKU, and bleeding prevents the iron overloading in hemochromatosis). Other examples are screening for other inborn errors of metabolism, congenital hypothyroidism (CHT), and familial hypercholesterolemia.

Screening for carrier status of autosomal recessive diseases such as cystic fibrosis, thalassemia, and Tay-Sachs disease can be offered to individuals, but because being a carrier is only significant when the partner is a carrier too, screening for carrier status can also wait until a partner has been found and reproductive issues arise. In this case the couple rather than the individual is the target. When both partners of a couple are carriers (for the same disease) their children are at 25 percent risk for the disorder.

Screening during pregnancy may be targeted at characteristics of one to three individuals. First, the fetus may be studied directly, for instance when ultrasound is used for the detection of congenital anomalies in general or nuchal translucency (an indication for the existence of Down's syndrome) in particular. Second, the screening may involve the pregnant woman first, such as is the case with maternal serum screening, e.g., triple test. In case of a positive test result the fetus becomes involved in a definitive diagnosis. Third, both the pregnant woman, her partner, and the fetus may become involved. This is the case with couple screening in pregnancy.

4.1 Timing

As hinted at above, apart from determining who should be screened, it is important to define the best moment for screening within the lifetime of the individual or couple. The earliest possibility is pre-implantation genetic screening (PGS) in case of *in vitro* fertilization (IVF). With pre-implantation genetic diagnosis (PGD), one is looking for a disorder for which this particular embryo has a particular high risk, while in PGS the embryo at stake has a not much higher risk than other pre-implantation or IVF embryos. PGS for chromosomal anomalies, for instance, would be very helpful to select the embryos with the highest chance of survival.

Prenatal screening, newborn screening, and adult screening are discussed above. Some people advocate screening in high school. Couple screening can be performed premaritally, preconceptionaly, and prenatally. Premarital screening is perhaps the best choice in cultures where selection of marriage partners is done by third parties, without involving the partners at all. Compulsory premarital screening, after the partners have matched, as is the case for thalassemia in Cyprus, infringes personal autonomy. Preconceptional screening is characterized by difficulties of reaching the target group in time. So the majority of couple-screening programs target pregnant women (and their partners).

4.2 Taking Advantage of Family Structure

Cascade screening is a special type of genetic screening. Starting from an individual with the characteristic of interest (the proband), one offers screening to his or her first-degree relatives. Once the relatives with the same characteristic have been identified, their first-degree relatives are also offered screening, and so on. This type of screening is suitable for dominant disorders, such as familial hypercholesterolemia, and for X-linked disorders with high penetrance and low mutation rate. It has, however, also been proposed for the detection of couples who are carriers of an autosomal recessive disease such as cystic fibrosis. The disadvantage there is that only a small proportion of at-risk couples can be identified in this way. So, screening without taking the family structure into account is a better option. A combination of non-cascade and cascade screening may, however, be very efficient.

4.3 Other Technicalities

In many cases it is not possible or feasible to have a test which is 100 percent sensitive or specific. The lack of perfect specificity can be countered by subsequent diagnostic workup. When the screening involves a DNA test, specificity is usually very high. If in the cascade approach family members are tested for the same mutation as has been found in the proband, sensitivity is also very high. In case of a test with imperfect sensitivity (e.g., 90 percent), the imperfection is multiplied in couple screening (81 percent).

Couple screening has many modalities (ten Kate et al. 1996). One may test one partner first, and only if this first partner tests positive, test the other partner (single entry or step-wise screening), or one may test both partners always (double entry). One may also decide to screen partners of carriers for more mutations than is feasible in the initial screen. One may sample both partners simultaneously or stepwise. One may give complete results, or only tell the couple whether they are both carriers or not.

5. Psychosocial Impact of Genetic Screening

The purpose of genetic screening is to provide people with information about their risks and the options available to them to avoid the birth of an affected child or to take preventive measures in case of a serious

disease. Genetic screening seems especially advantageous when intervention measures are available. For instance, for women with a high risk of hereditary breast cancer, confirmation that they are carriers makes it possible for them to take preventive measures such as participating in an intensive screening program for early detection of the disease. Because the potential impact of genetic screening is considerable, the advantages as well as the disadvantages of genetic screening should be considered carefully before deciding to offer a specific screening program routinely.

The availability of genetic screening may also have its drawbacks. Although the available choices would appear to increase people's control over their own destiny, they are actually experienced by many as an additional burden. An indication of the perceived difficulty of decisions is the finding that although most people at risk for a genetic disease say they intended to be tested in the future, many are not (e.g., Lerman et al. 1996). Some people might prefer the uncertainty of not knowing above the certainty that, for instance, they will develop a serious disease for which there is no cure, such as Huntington's disease. In addition, with the increasing possibilities and availability of genetic screening, more and more people, some with a very low risk for chromosomal abnormalities who first did not consider screening, are now confronted with the choice to undergo screening. In many countries young pregnant women with a low risk of having a child with Down's syndrome are offered prenatal screening. If they choose screening, they must cope with the possibility of a positive screening result and the implication that confirmation is needed with invasive diagnostics such as amniocentesis that itself carries a risk of miscarriage. When they decide against screening, they must deal with the feelings that they could have done otherwise in case things turn out badly and the child is affected.

The decision for or against screening is not only complicated emotionally but also cognitively. Genetic risk information is complex and replete with uncertainties, associated not only with the hereditary nature of the disease but also with (un)informativeness of the test results, the effectiveness of preventive measures, and the variability of expression of the disease (incomplete penetrance). For example, when a woman is diagnosed as 'at risk' for breast cancer based on family history, there is only a 30 percent chance that DNA testing will result in a conclusive test result. If a woman is a carrier of the breast cancer gene, she has an increased risk of getting breast cancer: 7 out of 10 women who are carriers will get breast cancer before the age of 75, and 3 of them even before 40. There is a 3 in 10 chance, however, that a woman who is a carrier will not get breast cancer. This is complicated information and difficult to integrate to make the decision about testing (e.g., Bottorff et al. 1998). Providing information about genetic risks of, e.g., breast cancer is considered an important component of genetic education and genetic counseling. The importance attributed to risk information is illustrated by the fact that recall of risk information and understanding of probabilities are considered central measures of the effectiveness of genetic counseling. The underlying assumption that accurate perception of risk could help people to make informed and individual decisions seems to be based on the principle of the autonomous and rational decision maker. From this perspective the most important goal of genetic education and counseling is to provide people with all relevant information in an understandable way. It is therefore important to find the best format to communicate risks and to stimulate a rational tradeoff of pros and cons of several options. Research has shown, however, that people find it difficult to understand risks and to make rational tradeoffs. It has been shown that knowledge of actual population risks of, e.g., breast cancer is far from adequate. People either underestimate or overestimate the risk and often fail to make a distinction between the population risk and their own risk. Genetic counseling has been shown to improve accurate risk perception, although the majority of women still have an inaccurate perception after counseling (e.g., Cull et al. 1999).

In the case of screening for a hereditary disease, the decision to screen is complicated by the potential impact that test results could have on family members. As has been pointed out, genetic screening may have implications for family members and may put familial relationships under pressure. For example, one sister may want to know more about the hereditariness of breast cancer while another sister might prefer not to know. The same goes for couples when there is no agreement as to whether to test for carrier status for, e.g., cystic fibrosis.

If the decision is made and test results become available, these need to be interpreted. Mostly, test results also involve uncertainties. A positive test result after prenatal screening for Down's syndrome means that the risk of having a child with Down's syndrome is higher than 1 in 250. A decision then must be made whether or not to confirm this chance estimate with invasive diagnostics whereby the extra information must be traded off against the risk of a miscarriage. Screening and waiting for the results have an emotional impact on people. Receiving test results, whether they are positive or negative, is also emotionally stressful. Many low-risk people are likely to undergo testing for reassurance and hope for a good result. Most of them will receive a negative test result. For those who perceive themselves as being at risk this is reassuring, but the news will perhaps be learned at the expense of knowing that they were at risk in the first place. It has been shown that anxiety exists even after negative results of prenatal screening (Green and Statham 1996). Learning what could go wrong without the possibility of ruling out all possible defects makes women more aware of birth defects and affects their

sense of vulnerability. Testing for carrier status of cystic fibrosis involves a large number of mutations but it is practicable to test only for the most common. If tested negative, the carrier status of cystic fibrosis, therefore, cannot be ruled out completely and no complete reassurance can be given that one is not a carrier. Even if testing implies that the person is not at risk, some members of risk families for, e.g., Huntington's disease find it difficult to cope with a future that no longer includes the expectation of getting a serious disease. An explanation for these negative feelings after being tested negative is that being at risk for this serious disease was part of one's identity which then needs to be changed (Tibben et al. 1997).

A positive screening test is obviously upsetting even if it turns out to be false. Women having diagnostic tests as a result of a positive screening test have been shown to be especially anxious compared to women who chose amniocentesis directly. Most of these women will be shown not to carry a child with Down's syndrome, and their anxiety can be considered to be unnecessary. Testing positive as a carrier of a recessive disease such as cystic fibrosis without actually having the disease can affect one's view of and one's optimism about the future. Being identified as a carrier for a genetic disease may affect how an individual perceives him- or herself and how he or she is perceived by others. Being tested positive for a disease for which there is no cure (such as Huntington's disease) also affects one's view on the future. Because nothing can be done to prevent the disease, this knowledge will not be considered as useful by all people (Keenen 1996). As more genetic conditions become diagnosed before any cure is in sight, more and more individuals will be added to the at-risk pool many years before a cure is possible. Some people may even be at risk their whole life for a condition they will never develop. It is therefore important to determine for which conditions and for whom genetic testing should be available.

It could be argued that at the basis of many of the problems or possible negative effects of genetic screening described above is a poor understanding of genetics, the tests, and the meaning of test results. It is therefore very important that the public as well as people with an at-risk status learn more about genetics, what it means to have some abnormal genes, how best to make a decision, what it means to have a positive screening result, etc. Professionals should learn the best way to provide genetic information in such a way that people can make a balanced decision for or against genetic screening and be able to cope with the test results so that the advantages of screening will be maximized and the disadvantages minimized (Schwartz et al. 1999).

6. *Criteria*

There are many genetic, medically relevant characteristics for which, in theory, screening could be started now. There will be many, many more in the future. It is quite clear that it is not desirable to screen for all these characteristics, and careful selection is needed. Apart from potential benefit, potential harm should be considered. Harm could be physical, psychological, social, or otherwise. Not only the effects of screening on the participants should be taken into account, but also the effects on nonparticipants.

Many individuals and groups have worked on the development of criteria that should be fulfilled for screening in general (EUROGAPPP PROJECT 1999–2000, Holtzman and Watson 1997, Nuffield Council on Bioethics 1993). Some countries have also developed specific legislation. It is widely recognized that every individual screening program should be tested to see if it reaches appropriate standards.

Basically, the information that is needed for this test can be grouped into three sets. The first set concerns what is known about the disease or characteristic and about the screening program (state of knowledge). The second set concerns the advantages and disadvantages arising from the screening, and the third set is about the guarantee of autonomous decision.

The disease should be well delineated. Its general course should be clear but also the amount of variability in its course. The effects of intervention should be known. A reliable test should be available with known sensitivity, specificity, and predictive value. The screening program should be described in detail, including the availability of intervention and follow-up. Privacy protection, and permanent quality control of all aspects of the program should be ensured.

Practical courses of action should be available for the participants, such as preventive measures or reproductive choices. The psychological and social effects should be known for the participants and nonparticipants in general, but also and in particular for those among the participants who are faced with a positive test and those among the nonparticipants who are confronted with the disease in themselves or in a child. Also, more remote effects on society should be considered, e.g., on the amount of support for handicapped people. Finally, the financial costs of the program should be taken into account and alternative strategies to combat the health problem should be considered. All this information should make it possible to compare and weigh the benefits and drawbacks of the program. The final balance should be clearly biased towards the benefits.

Participation in the screening program should be voluntary and based on good-quality information and informed consent. People should be given enough time to reach a balanced decision. This requirement may be difficult to realize in prenatal and neonatal screening, but every effort should be made to inform pregnant women about the decision to come as early in pregnancy as possible. People should have ample opportunities for pre-test consultation. People should

not be taken by surprise (as in opportunistic screening).

It is desirable to have an authority to supervise all (genetic) screening programs in a country or larger region. Criteria should be the same for screening programs offered by nonprofit organizations and commercial firms. As more and more tests become available, problems inherent in multiple testing arise. Multiple testing increases the frequency of false negative and false positive results and augment the difficulty of giving sufficient information to reach a balanced decision.

The complexity of genetic screening makes it a multidisciplinary area, involving medical doctors, laboratory specialists, epidemiologists, psychologists, sociologists, jurists, ethicists, patient organizations, and representatives of the community.

See also: Alcoholism: Genetic Aspects; Cancer Screening; Community Genetics; Dyslexia, Genetics of; Eating Disorders, Determinants of: Genetic Aspects; Ethical Issues in the 'New' Genetics; Familial Studies: Genetic Inferences; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Engineering; Genetic Testing and Counseling: Educational and Professional Aspects; Mental Illness, Genetics of; 'New' Eugenics, The: Role in Society; Obesity, Determinants of: Genetic Aspects; Risk Screening, Testing, and Diagnosis: Ethical Aspects; Schizophrenia and Bipolar Disorder: Genetic Aspects

Bibliography

Aymc S, Bobrow M, Daviello D 2000 Population genetic screening programmes: Proposed recommendations of the European Society of Human Genetics. *European Journal of Human Genetics* **8**: 998–1000

Bottorff J L, Ratner P A, Johnson J L, Lovato C Y, Joab S A 1998 Communicating cancer risk information: The challenges of uncertainty. *Patient Education Counselling* **33**: 67–81

Cull A, Anderson E D C, Campbell S, Mackay J, Smyth E, Steel M 1999 The impact of genetic counselling about breast cancer risk on women's risk perceptions and levels of distress. *British Journal of Cancer* **79**: 501–8

Green J, Statham H 1996 Psychosocial aspects of prenatal screening and diagnosis. In: Marteau T, Richards M (eds.) *The Troubled Helix: Social and Psychological Implications of the New Human Genetics*. Cambridge University Press, Cambridge, UK, pp. 140–64

Health Council of the Netherlands: Committee Genetic Screening 1994 *Genetic Screening*. Health Council, The Hague, publication no. 1994/22[E]

Holtzman N A, Watson M S (eds.) 1998 *Promoting Safe and Effective Genetic Testing in the United States*. Final report of the task force on genetic testing. Johns Hopkins University Press, Baltimore, MD

Keenen R H 1996 The at-risk health status and technology: a diagnostic invitation and the 'gift' of knowing. *Social Science & Medicine* **42**: 1545–53

Lerman C, Narod S, Schulman K, Hughes C, Gomez-Caminero A, Bonney G, Gold K, Trock B, Main D, Lynch J, Fulmore C, Snyder C, Lemon S J, Conway T, Tonin P, Lenoir G, Lynch H 1996 BRCA1 testing in families with hereditary breast-ovarian cancer. A prospective study of patient decision making and outcomes. *Journal of the American Medical Association* **275**: 1885–92

Nuffield Council on Bioethics 1993 *Genetic Screening. Ethical Issues*. Nuffield Council on Bioethics, London

Schwartz L M, Woloshin S, Welch H G 1999 Risk communication in clinical practice: Putting cancer in context. *Journal of the National Cancer Institute of Monographs* **25**: 124–33

ten Kate L P, Verheij J B G M, Wildhagen M F, Hildering H B M, Kooij L, Verzijl J G, Habbema J D F 1996 Comparison of single entry and double entry two-step couple screening for cystic fibrosis carriers. *Human Heredity* **46**: 20–25

Tibben A, Timman R, Bannink E C, Duivenvoorden H J 1997 Three-year follow up after presymptomatic testing for Huntington's disease in tested individuals and partners. *Health Psychology* **16**: 20–35

Worwood M 2000 Early detection of genetic hemochromatosis: should all young adults be offered the genetic test? *Genetic Testing* **4**: 219–28

L. P. ten Kate and D. R. M. Timmermans

Genetic Studies of Behavior: Methodology

The phrase 'methodology in genetic studies of behavior' is intended to cover any scientific analysis of behavioral phenomena that includes biometric genetic information. Human beings have multiple thousands of genes—specific DNA sequences on specific areas of the chromosomes—and each gene represents a specific sequence of amino acids that creates RNA for the formation or coding of specific proteins. Different versions of a gene appear as a pair of *alleles* at a specific position or *locus*, and a specific collection of alleles are labeled a *genotype*. While over 99 percent of all genes in humans are identical to the corresponding genes in all other humans, the other 1 percent are important because they are thought to lead to individual differences in both genotypes. Human beings exhibit multiple thousands of observed behaviors and any well-defined subset of these behaviors is labeled a *phenotype*. In these terms, the methods of genetic analysis of behaviors are used to clarify and improve our understanding of the relationships among particular variations or combinations of genotypes and particular variations or combinations of phenotypes.

Genetic methods and analyses have roots in the nineteenth century, specifically in Charles Darwin's well-known ideas on evolution and Gregor Mendel's work on the mechanisms of heredity in plants. Many techniques used in genetic analyses of behavior were initiated in the agricultural sciences to better control the size and health of plants and animals, further developed in the medical sciences to understand and control sources of disease, and are now routinely used in large-scale family studies to better understand many

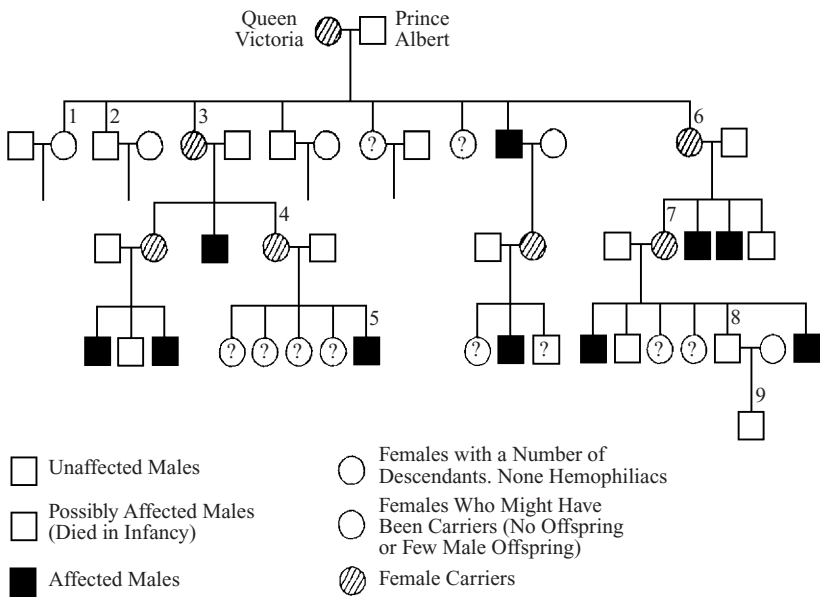

Figure 1
A pedigree chart for the inheritance of hemophilia in the family of the English Queen Victoria, Prince Albert, and their descendants (from Dixon and Johnson 1980, p. 24)

behavioral phenomena. Most of the latter methods described here stem from the classical nineteenth century work of Sir Francis Galton on the measurement of individual differences in physical and mental characteristics (Johnson et al. 1985).

In this article, we discuss some historical background, we focus on contemporary methodological issues, and we emphasize the methodological unity of most current genetic approaches. We do not deal with important contemporary issues about the use of genetic information (i.e., genetic testing and screening, ethical use of genetic information, human cloning, genetically altered food, etc.).

1. Classical Methods in Genetics and Behavioral Studies

From a historical point of view, human genotypes have been very difficult to measure and have often been treated as unobserved or latent attributes. Analyses of family trees or 'pedigrees' were among the first scientific evidence provided to clearly establish that aspects of human behavior are influenced by genetic variation. Figure 1 is a chart of a family tree or 'pedigree' from Galton of the inheritance of hemophilia in the family of the English Queen Victoria, Prince Albert, and their descendants. The pedigree method has been used to establish whether given disorders (e.g., hemophilia) run in families and, if so, what kind of inheritance (dominant or recessive, autosomal or sex-linked) best fits the observed pattern, for example hemophilia was found to be a sex-linked characteristic carried on the X chromosome. In another well-known example, clinical research on mental retardation, especially research showing the effects of inbreeding and the genetic cause of Down syndrome, has established the role for genetic regulation of impaired intellectual functioning (for references see Dixon and Johnson 1980, McArdle and Prescott 1996).

Findings from clinical disorders led researchers to investigate whether genetic factors might regulate aspects of behavior within the normal range of variation. Early evidence for genetic effects on normal

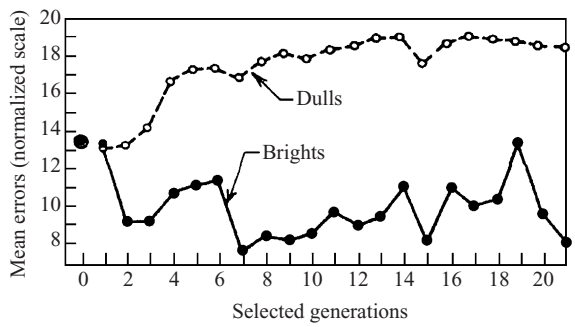

Figure 2
The results from Robert Tyron's (1940) selective breeding for maze learning over 20 generations of rats (from McGill 1965, p. 75)

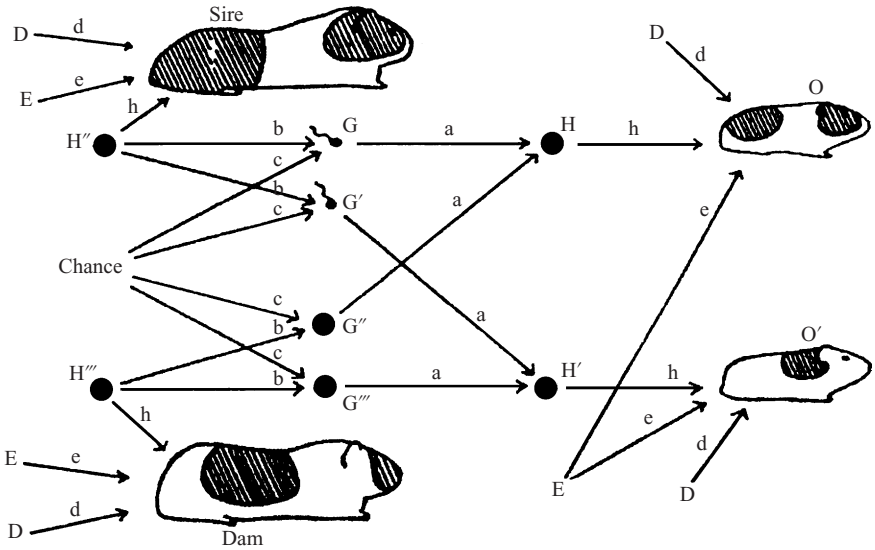

Figure 3
A path analysis model of inheritance of coat-pattern in guinea-pigs (from Sewell Wright 1920)

behaviors came from selective-breeding studies with non-human animals. Figure 2 is a plot of results from the well-known selective-breeding studies of Tolman, Jeffress, and Tryon (Tryon 1940 as reported by McGill 1965). In these studies a heterogeneous stock of rats was tested on learning tasks in an automatic maze. The observed performance in these learning tasks led the experimenters to classify each rat into one of two groups—'dull' or 'bright'—within which the rats were allowed to mate and produce offspring for the next generation. After accounting for some deleterious effects (e.g., inbreeding and regression), these investigators showed that after successive selection of relatively few generations this behavioral selection led to distinct strains of 'maze-bright' and 'maze-dull' rats. With no change in rearing environments, these experiments provided direct evidence for the role of genetic components in learning.

Galton also initiated large-scale studies based on family resemblance. He studied physical features (e.g., height) in parents and children and reported the law of 'filial regression (reversion) towards mediocrity' better known as 'regression to the population mean.' He also studied intellectual achievement and reported that individuals born to higher social classes attained higher educational and social standing than individuals born to lower classes. These family studies were not controlled experiments, so scientific inferences about the separate impacts of changing genotypes and changing environments remained obscure. Selective breeding experiments were impossible with human populations for ethical reasons. Nevertheless, experimental concepts derived from animal research gradually led to creative use of a novel class of methods based on correlations observed among atypical groups of families—'twin studies,' 'adoption,' 'half-sibs,' and other 'genetically informative' family groupings were discussed in the early twentieth century by many researchers including Thorndike, Siemans, Merriman, and Fisher (see Spector et al. 2000).

Multivariate statistical methods were developed to accommodate unobserved or latent genotypes. Figure 3 is a path diagram from the classical work of geneticist Wright (1920) showing a complex system of influences. Here Wright suggested the environmental factors (E) were mixed with the heredity influences (H) together with 'Chance' factors (and those labeled D), and all were required to account for the 'coat-pattern' of a guinea-pig litter. More complex models of inheritance were defined by the quantitative rules of Mendelian inheritance and the factors of the correlation decomposition of the statistician R. A. Fisher and later by many other including R. B. Cattell, and J. Jinks and D. Fulker (see Falconer and Mackay 1996).

Advanced versions of these genetic methods are now widely used in behavioral research. The search for specific genetic abnormalities linked to behavioral disorders continues at a rapid pace in research from the Human Genome Project (e.g., Strachan and Read 1995). Genotypes are now directly related to behaviors using experimental techniques based on direct manipulation of the genome in model organisms, as well as statistical techniques of linkage and association (Elston 1998). In advanced statistical studies, substantial progress has been made in the use of mathematical and statistical techniques for studies of family resemblance based on latent genotypes and latent phenotypes (McArdle and Goldsmith 1990, Neale and

Cardon 1992). In the next four sections we highlight contemporary issues in these methodologies and present some issues for future research.

2. Contemporary Experimental Methods with Measured Genes

The last decade of the twentieth century has seen a revolution in molecular genetic methodology. Techniques based on recombinant DNA methods permit direct measurement and experimental manipulation of variations in genotype or gene expression. It is useful to divide these techniques on the basis of whether they primarily use information about a gene's position within the genome, a gene's function, a gene's DNA sequence, or a gene's protein product to develop our understanding about the genetic influences on behavioral characteristics.

Studies based on positional information are usually aimed at the long-term goal of positional cloning of the gene or genes that influence some trait (Burmeister 1999, Schork and Chakravarti 1996). In positional cloning, one capitalizes on the fact that Mendel's law of independent assortment does not hold for genes that are physically near each other on the same chromosome, so genes do not assort randomly. Rather they tend to co-segregate within families to a degree that is monotonically and directly related to their physical proximity to each other on a chromosome. A genetic marker is a specific physical location on a strand of DNA. If co-segregation is observed between a genetic marker and a phenotype, it is reasonable to conclude that the marker is physically close or linked to a gene influencing the phenotype. When this co-segregation is examined as an association between markers and phenotypes within families one is generally testing for linkage. When this co-segregation is examined as an association between markers and phenotypes across nominally unrelated individuals one is often testing for linkage disequilibrium (Elston 1998).

There are many types of studies that use positional information to find genes influencing traits in non-human animals (e.g., Lynch and Walsh 1998). Selective breeding studies with measured genes are the conceptually simplest method used to locate genes influencing traits. These studies are based on measurements of the frequencies of alleles at genetic loci in an outbred population. Any loci exhibiting statistically significant changes in allele frequency over generations are either locus influencing the trait or linked to loci influencing the trait.

A more complex but efficient method uses recombinant inbred strains. Once several inbred strains are produced, one animal from each line can be genotyped at a set of marker loci. Because all animals within a strain have the same genotype, it is only necessary to genotype one animal within each strain to know all animals' genotypes. At each locus of interest, one simply tests whether the different lines (i.e., AA vs. BB genotypes) have different phenotypic outcomes. If significant differences are found, this is evidence that the locus under study either causes variation in the phenotype or is linked to a locus that causes variation in the phenotype. In certain situations, the use of inbred strains can be quite efficient in terms of both genotyping and phenotyping costs. More complex methods involve intercrosses between inbred strains, multipoint analyses with simultaneous uses of multiple markers, multivariate analysis of multiple phenotypes, selective sampling, and other enhancements.

One limitation of all approaches based on positional information is that they only work with existing variation within a gene, that is, they do not tell us what genes are fundamentally important to a phenotype when all individuals within a population have the same alleles at that gene. In one kind of experimental approach, molecular geneticists are now able to alter or remove particular genes and observe the behavioral effects. One of the most common types of induced mutation studies is the development of a 'knockout' animal where a specific targeted gene is disrupted or 'knocked-out.' Another type of study is based on the use of random insertion clones—expressed sequences tag (EST) acting as a molecular homing device—to disrupt genes throughout the genome. Another type of direct genetic manipulation is referred to as the generation of a 'transgenic' model where additional genetic material or a new gene is inserted into an animal. A final approach to studying gene function uses microarrays, small silicon chips, which have been set up to simultaneously measure the amount of expression of literally thousands of genes. This approach is permit study the circumstances under which certain genes come in to play and, in future research, may permit analyses of multiple interactions of measured genes with measured environments.

3. Contemporary Statistical Methods Based on Measured Genotypes in Humans

In human studies, one classical way to find genes that influence behavioral traits is to look for cytogenetic anomalies among rare individuals with respect to those traits (see Dixon and Johnson 1980). If a cytogenetic anomaly affects a large segment of the chromosome (i.e., missing, transposed, or present in extra copies), and it can be detected in the individuals, then it is plausible that the genomic region of the anomaly harbors a gene influencing the trait. While this method may provide a good starting point for further study, most behaviors are not marked by discontinuity, are not related to abnormality, and cannot be examined in this simple way.

In the method known as linkage analysis, associations (e.g., correlations) between transmission of a

measured allele and some measured phenotype are studied within families (Elston 1998). The variety of methods for testing for linkage in human studies is now quite large, but the prototypical study is the sib-pair study. Two classes of tests are commonly applied to quantitative genetic information. Neither relies on the assumption of a particular genetic model (Elston 1998) and both are based on allele sharing at marker loci (Lander and Schork 1994). Covariance-based tests are based on standard statistical principles and are commonly used (Allison et al. 2001). Under the null hypothesis of no linkage between a marker and a quantitative trait locus (QTL), there will be no association between the number of alleles a sib-pair shares identical by descent (IBD) and their degree of phenotypic similarity for the traits under study. In contrast, when there is linkage between the marker and the QTL, there will be a positive association between (a) the number of alleles a pair shares IBD and (b) the degree of phenotypic similarity of the pair. Thus, by evaluating whether there is evidence of positive covariance between IBD and phenotypic similarity one can test for linkage. Another class of tests is based on inferring such association by observing a departure of the 'marginal' distribution of IBD from its null expectation in samples of sib-pairs that have been selected on the basis of their intra-pair phenotypic similarity or dissimilarity.

One of the key issues in both human and non-human genetics, particularly linkage analysis with large numbers of candidate genes, is the determination of an appropriate threshold for statistical significance given the enormous amount of multiple testing conducted (Lander and Schork 1994). Although no definitive consensus now exists, probability or p-values greater than about 10^{-4} are rarely taken as compelling evidence for a linkage effect. Reports of p-values smaller than 10^{-4} are common and are often persuasive. However, as in any statistical model, these p-values are calculated under various untested assumptions (e.g., distributional assumptions, accuracy of genotyping, accuracy of phenotyping, independence of observations) and the test procedures may lead to empirical Type I error rates which exceed (or fall below) nominal alpha levels (e.g., Allison et al. 1999). It follows that small p-values alone should not be taken as convincing evidence for a linkage finding in the absence of additional evidence. Researchers looks for careful documentation of the statistical robustness of the finding (e.g., bootstrap and sensitivity analyses), replication in an independent data set, or confirmation through independent methods such as gene knockouts in animal models.

Perhaps the chief limitation of linkage analysis in the context of human studies is low statistical power. For example, to detect a QTL explaining 5 percent of the variance in a phenotype would usually require that one study, or at least screen, thousands of sibling pairs. The reasons for this low power are manifold, but chief among them is the fact that in linkage analysis in humans, one is generally not examining the direct effect (in terms of mean displacement) of different alleles at a QTL, but rather the variance explained by the QTL (i.e., a variance component). In addition, this is only an indirect estimate based on the observation of IBD scores at a marker allele, and these alleles are only partially correlated with IBD scores at the QTL. This being the case, initial analyses focus on a simple test of mean differences in phenotype by genotype at a putative QTL; such a test is referred to as an 'association test' in the human genetics literature. Under most circumstances, association tests will be enormously more powerful than linkage tests when one has the polymorphisms of the actual QTL in hand. The typically greater power of association tests has led some to suggest that linkage analyses be replaced by association tests, but this position has led to controversy (e.g., Risch and Merikangas 1996).

One special concern comes from the need to control for background correlations or other confounding, usually referred to as admixture or population stratification in the context of human genetic studies. These confounds arise from the fact that the effects of alleles at any one locus may be stochastically dependent upon the effects of alleles present at another locus even if these two loci are not physically linked on the same chromosome. To avoid the inappropriate conclusions that could stem from such confounding, Spielman et al. (1993) developed what is termed the transmission disequilibrium test (TDT). In the TDT, one counts the number of times a particular allele is transmitted to an offspring affected with the disease from a parent that is heterozygous at the marker locus. If a particular allele is transmitted from a heterozygous parent (i.e., a parent that has two different alleles at the candidate locus) more than 50 percent of the time (i.e., the chance expectation) this is evidence that, not only is that allele associated with the disease, but that it is also linked to a locus which causes the disease. This is because, under ordinary circumstances, each allele a parent has at a genetic locus has an equal (50 percent) chance of being transmitted to an offspring, so any difference from this expectation is evidence of both linkage and association. Since then, numerous extensions of the TDT approach have been developed for both qualitative and quantitative traits (e.g., Allison 1997, Allison et al. 1999).

4. *Contemporary Statistical Methods Based on Latent Genotypes in Humans*

The vast numbers of human genotypes are very difficult to measure and are necessarily treated as latent attributes. This is especially true when a complex of influences and a complex of outcomes are considered. In behavior genetics and genetic epidemiology

all statistical methods utilize classical biometric genetic theory to define and specify the relations among the underlying components thought to represent a continuously distributed trait (e.g., Falconer and Mackay 1996). Most models start by asserting that individual genes combine to produce the underlying genotype. Furthermore, these genes are assumed to be: (a) numerous or 'polygenic,' (b) additive in effect, (c) segregating independently (i.e., the probability of receiving one gene is independent of the probability of receiving others), (d) not sex-linked (i.e., located on the X or Y chromosomes), and (e) not influencing characteristics used for spousal selection.

These biometric assumptions are important because they lead to statistical expectations about the observed correlations among pairs of relatives. If all resemblance between relatives in a family is due to genetic sources, the correlations between their observed phenotypic scores are expected to be proportional to the correlations between their genotypic backgrounds, and unrelated individuals would be uncorrelated. More specifically, the genotypes of first-degree relatives correlate (on average in a large sample) at 1/2. First-degree relatives include fraternal (also called dizygotic or DZ) twins, ordinary siblings, and parent-child relationships. The genotypes of second-degree relatives (grandparents-grandchildren, aunt-niece, half-siblings) are expected to correlate 1/4, and the third-degree relatives (cousins) at 1/8, etc. Unrelated individuals living together, such as stepsiblings or parents and their adopted children would be expected to have zero genetic correlation. In the special case of identical (monozygotic or MZ) twins, the genotypes are considered identical and correlate 1.0. Of course, the adequacy of inferences from the sampling of people that are related or unrelated is a major concern of behavior genetic researchers.

Nongenetic influences are often considered as well. In some models the similarity in behavior is expected to be correlated with the similarity in relevant environments (e.g., home of rearing or social background). If true, the highest correlations would be found in scores from twins and similar-aged siblings, lower correlations from parents and children, and still lower correlations from related individuals who have not lived together. Thus, under a purely environmental hypothesis, adoptees should be as similar to their adopting parents and siblings as any other children, and related individuals living in different environments would have scores with no correlation. A model that includes both genetic and environmental sources of familial resemblance would produce estimates intermediate between these other two. All such models can include measurement error and the influence of experiences unique to particular individuals, so even MZ twins living together are not expected to have scores that correlate perfectly.

Different family data permit different assessments of the theoretical components. Figure 4(a) illustrates this kind of latent path model for a 'family' design. In this model, additive genetic and shared environmental sources of resemblance are confounded, that is, any interpretation of family resemblance cannot be partitioned into separate additive genetic scores (A) and common family environment scores (C). In the adoption design of Fig. 4(b) these A and C components are (theoretically) separable, and the only source of parent-child resemblance is through cultural transmission, resulting in an estimate of common family effects. By comparing differences in resemblance among family members in the adoptive versus biological family designs, an estimate of the additive genetic (A) contribution to the phenotypic deviation is obtained.

Figures 4(c) and 4(d) show models for resemblance between siblings or fraternal (DZ) twins and identical (MZ) twins. In this kind of model, pairs of DZ twins and pairs of full siblings are assumed to have the same genotypic correlation, even though the siblings can differ widely in chronological age. Unlike the family and adoption models, here we cannot estimate the effects of assortative mating based on the correlation between A and C. In most twin models these correlations are assumed to be zero although fixed estimates from other sources can be inserted. Using these assumptions, the twin design can provide estimates based on the difference between the resemblance between MZ and DZ pairs. The special case of MZ twins raised apart (MZA) is important because, in theory, these data permit a complete separation of these latent components.

Structural equation models developed in several areas of science are used in behavior genetics to organize, estimate, and test aspects of these theories of underlying influences (McArdle and Goldsmith 1990, Neale and Cardon 1992). In some recent models an observed phenotype (P) is often cast as a simple combination of true scores (T) and error of measurements (W). Behavioral genetics models go further and include unobserved variables representing the biometric hypotheses: Additive genetic scores (A), nonadditive genetic scores (D), common family environments (C), other nongenetic (E) components, and possibly other complex sources. The goal of the structural analysis is independently estimate the total impact (contribution to the variance) of the separate components using the available data from relatives. The statistical basis of the models relies on features common to many scientific analyses—maximum likelihood estimation and likelihood ratio comparisons (e.g., chi-square tests). A sequence of nested statistical tests is routinely used to examine the necessity of each latent component and confidence boundaries can be created.

After specific model coefficients are estimated at a specific value, it is possible to standardize the unobserved variables, assume the latent scores (A, C, and E) are uncorrelated (orthogonal), and calculate the

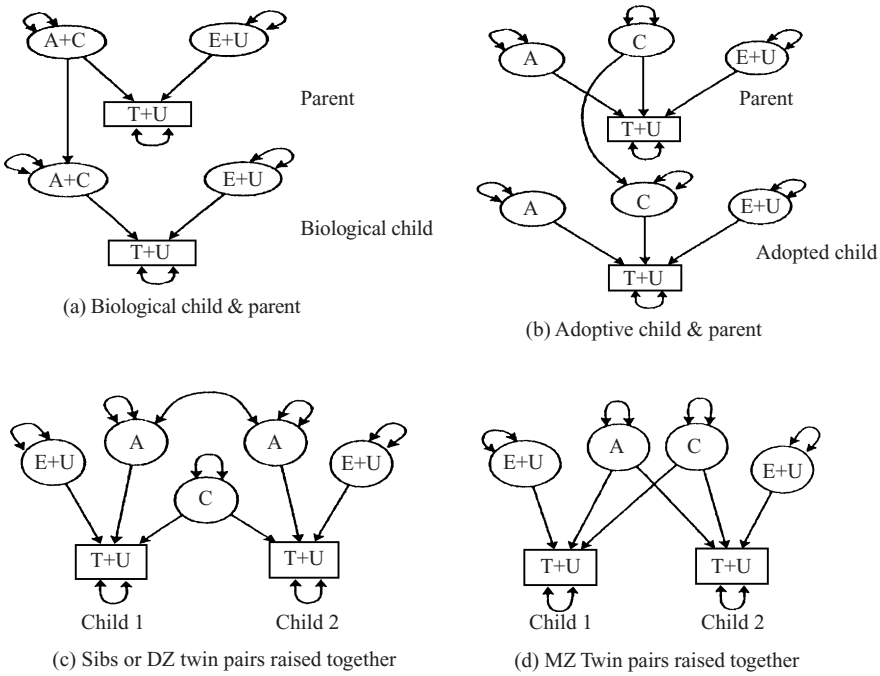

Figure 4
Latent variable path models for the analysis of family resemblance (from McArdle and Prescott 1996). (Note: In these diagrams: (1) Squares are used to represent observed or measured variables. (2) Circles are used to represent unobserved or theoretical variables (3) One-headed arrows are used to represent fixed regression coefficients. (4) Two-headed arrows are used to represent estimated variance or correlation coefficients.)

unobserved genetic variance. In order to compare results for phenotypes with different measurement scales, it has become common to calculate the 'heritability' ratio—the unobserved additive genetic variance divided by the observed phenotypic variance ($h^2 = V_a/V_p$). This kind of calculation may or may not take into account the common family environment, other genetic factors (dominance, epstasis), and is limited by the sample phenotypic variance. In most published research this sample statistic (h^2) is reported and the utility of this routine practice has been seriously questioned (e.g., Gottlieb 1997, Wachs 1999, Hooper in Spector et al. 2000).

There has been a great deal of recent progress towards addressing more complex problems and creating methods that permit the study of more complex patterns of influences. Contemporary behavioral genetic methods do include the possibility of evaluating these more complex patterns of influence, including: parent to child environmental transmission of shared effects, further decomposition of the nonshared environments, the correlation between spouses, and the separation of measurement error via multiple phenotypes. The possibility of a correlation between *A* and *C* variables is an important feature of contemporary models. These correlations can be meaningful but they often cannot be estimated without more extensive data collections and more complex models. In this recent work, parameters for all effects are estimated by assuming the model parameters are invariant over all the multiple groups of Fig. 4. Other important model features include nonadditive genetic effects such as dominance (nonadditive effects of alleles at the same locus) and epistasis (nonadditive effects of genes at different loci) and the general utility of mean intercepts, subject sampling corrections, and multivariate phenotypes (see McArdle and Goldsmith 1990, Neale and Cardon 1992, McArdle and Prescott 1996).

5. *Future Issues in Genetic Methods for Behavioral Studies*

As with any methodology, there are both strengths and weakness of the genetic methods discussed here. Direct gene measurement permits usually strong experimental designs—in gene association analyses, the appropriate dependent variable is phenotype and the appropriate independent variable is genotype. However, in most studies the genotype is a nearby marker and not the gene of interest. This produces several

unresolved questions that generate skepticism about the value of association tests: (a) Can candidate loci and polymorphisms be identified with sufficient confidence to be useful? (b) Can enough separate polymorphic loci be studied to do genome-wide association studies when candidate loci cannot be confidently identified on an *a priori* basis? (c) Can confounding by population admixture/stratification be controlled? (d) Will applied investigators conduct association studies with adequate rigor even when the tools are available to them?

The challenge of creating hypotheses about candidate genes must be acknowledged, but this is more a limitation than a flaw of the method. The availability of sufficient markers when one wants to go beyond prespecified candidate genes, the ability to economically assess a large number, and the ability to determine how large that number needs to be are currently being addressed by many researchers. The occasional lack of rigor in the conduct of association studies should not be trivialized, but is a problem that potentially affects all areas of research.

There is ongoing research in trying to understand genes and their role in producing phenotypic effects by studying the more complete gene sequence and gene-protein structure. This latter field is often referred to as proteomics and is one of the most active fields in molecular biology at the beginning of the twenty-first century. These methods will become increasingly important in the future and, moreover, will be combined with studies involving information about gene position, gene function, and behavioral measurements to provide a much richer picture about the way genes work to produce their effects. In these more detailed analyses of measured genes we may also be able to study about how gene-protein functions are stimulated by specific environments.

In a similar vein, the statistical methods for latent genotypes rest on assumptions that may not be reasonable representations of the underlying mechanisms, and some of these assumptions are not testable. Statistical biases arise in adoption studies when the effects of selective placement lead to a correlation between the genotypes of the biological and adoptive parents. In twin data, positive assortative mating leads to underestimates of additive genetic effects and overestimates of downward and common family effects. In addition, the standard twin model assumes that both kinds of twins experience similar environments to an equal degree, but the similarity of MZ pairs is in part due to sharing a more similar common environment, the additive genetic effects will be raised spuriously.

But other modeling assumptions are more fundamental and problematic. For example, the classical models above routinely assume no interaction of the genetic and non-genetic components, initially because these additional components are hard to estimate without detailed measurements of both the genotypes and the environments. Similarly, the concept of genotypes as immutable characteristics with asymmetric direct effects on phenotypes is often challenged. Researchers often suggest that both genes and the environments play an important role in creating constraints on developmental processes (e.g., Waddington 1962, Gottlieb 1997, Wachs 1999).

More complex genetic methodologies are needed to answer these complex questions about complex human characteristics. These are not questions that can be simply answered in terms such as, 'we have found the gene(s) *Q* for behavior(s) *P*,' or, 'recent twin research shows *P* is largely genetic.' Further answers for the questions raised here will, no doubt, combine the best aspects of both measured and latent genotype approaches. Recent research demonstrates the natural convergence of QTL analysis and latent variable structural models (e.g., Neale in Spector et al. 2000). Newer models are likely to include basic information about the phenotypic outcomes in terms of multivariate measurement models combined with longitudinal dynamic modeling.

A clearer understanding of the methods, models and results from these newer analyses may illuminate the complexity of genetic theory and behavioral theory, as well as the appropriateness of scientific inferences that can be made from such studies. We are given pause when we consider that, even after the entire human genome is accurately mapped, we will still be confronted with an enormously complex set of analytic problems—of linking genotypes to proteins, and proteins to physiology, and physiology to basic levels of behavior, and basic behaviors to more complex ranges, constraints, and the resulting limitations on multiple behaviors. The prior methodological advances have carried us very far, but we now look forward to the future development of novel methods for these monumental tasks.

See also: Asymmetry of Body and Brain: Embryological and Twin Studies; Depression; Ethical Issues in the 'New' Genetics; Genes and Culture, Coevolution of; Genetics in Anthropology; Genetics of Complex Traits Through the Life Cycle; Human Evolutionary Genetics; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Schizophrenia

Bibliography

Allison D B 1997 Transmission disequilibrium tests for quantitative traits. *American Journal of Human Genetics* **60**: 676–90

Allison D B, Heo M, Kaplan N, Martin E R 1999 Development of sibling-based tests of linkage in the presence of association for quantitative traits that do not require parental information. *American Journal of Human Genetics* **64**: 1754–64

Burmeister M 1999 Basic concepts in the study of diseases with complex genetics. *Biological Psychiatry* **45**: 522–32

Dixon L K, Johnson R C 1980 *The Roots of Individuality: A Survey of Human Behavior Genetics*. Brooks/Cole, Monterey, CA
Elston R C 1998 Linkage and association. *Genetic Epidemiology* **15**(6): 565–76
Falconer D S, Mackay T F C 1996 *Introduction to Quantitative Genetics*, 4th edn. Longman, London
Gambaro G, Anglani F, D'Angelo A 2000 Association studies of genetic polymorphisms and complex disease. *Lancet* **355** (9200): 308–11
Gottlieb G 1998 Normally occurring environmental and behavioral influences on gene activity: From central dogma to probabilistic epigenesis. *Psychological Review* **105**: 792–802
Johnson R C, McClearn G E, Yuen S, Nagoshi C T, Ahern F, Cole R R 1985 Galton's data a century later. *American Psychologist* **40**(8): 875–92
Lander E S, Schork N J 1994 Genetic dissection of complex traits. *Science* **265**(5181): 2037–48
Lynch M, Walsh B 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Association, Sunderland, MA
McArdle J J, Goldsmith H H 1990 Some alternative structural equation models for multivariate biometric analyses. *Behavior Genetics* **20**(5): 569–608
McArdle J J, Prescott C A 1996 Contemporary models for the biometric genetic analysis of intellectual abilities. In: Flanagan D P, Genshaft J L, Harrison P L (eds.) *Contemporary Intellectual Assessment: Theories, Tests and Issues*. Guilford Press, New York, pp. 403–36
McGill T E 1965 *Readings in Animal Behavior*. Holt Rinehart & Winston, New York
Neale M C, Cardon L 1992 *Methodology for Genetic Studies of Twins and Families*. Kluwer, Dordrecht, The Netherlands
Risch N, Merikangas K 1996 The future of genetic studies of complex human diseases. *Science* **273**(5281): 1516–17
Schork N, Chakravarti A 1996 A nonmathematical overview of modern gene mapping techniques applied to human diseases. In: Mockrin S C (ed.) *Molecular Genetics and Gene Therapy of Cardiovascular Disease*. Marcel Dekker, New York, pp. 79–109
Spielman R S, McGinnis R E, Ewens W J 1993 Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**(3): 506–16
Spector T D, Sneider H, MacGregor A J (eds.) 2000 *Advances in Twin and Sib-pair Analysis*. Oxford University Press, London
Strachan T, Read A P 1995 *Human Molecular Genetics*. Wiley-Liss, New York
Wachs T D 1999 *Necessary But Not Sufficient: The Respective Roles of Single and Multiple Influences of Individual Development*. American Psychological Association, Washington, DC
Waddington C H 1962 *New Patterns in Genetics and Development*. Columbia University Press, New York
Wright S 1920 The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. *Proceedings of the National Academy of Sciences* **6**: 320–32

J. J. McArdle and D. B. Allison

Genetic Studies of Personality

Genetic studies of personality focus on (a) evaluating the extent to which genetic influences are important for individual differences in personality traits and (b) demonstrating whether or not specific genes are associated with personality traits. The thought that genetic effects may be important for personality is not new. Indeed, many cultures have sayings, for example, in Swedish and English, 'the apple does not fall far from the tree,' which suggest that genetic effects are important. Formal studies of genetic effects on personality started with Sir Francis Galton, who examined familial similarity for a variety of human characteristics. During the twentieth century, multitudes of behavioral genetic studies of personality have quantified the relative importance of genetic effects. Genetic differences among individuals account for on average 40 percent of the variance (individual differences) observed in most personality traits. Many specific genes are likely to be associated with personality, each with small effects.

1. *Behavioral Genetic Studies of Personality*

Twin, family, and adoption studies have been the primary source of information on the relative importance of genetic effects ('heritability') for personality. Behavioral geneticists study the similarity of individuals who share genes and environments to different degrees to quantify heritability. In the classical twin study, identical twins share all of their genes in common whereas fraternal twins share on average only half of their segregating genes in common. Thus, if genetic effects are important for a personality trait, identical twin pairs should be twice as similar as fraternal twin pairs. Similarity between parents and offspring or among siblings may also indicate the influence of genetic effects. The importance of the shared rearing environment can be evaluated by assessing the similarity of adoptees and their adoptive parents. Environmental influences can also be individual specific, or 'non-shared,' i.e., contributing to differences among family members. Behavioral geneticists typically apply structural equation modeling techniques to paper and pencil self-reports of personality when estimating heritability and the importance of shared and non-shared environmental influences (see *Genetic Studies of Behavior: Methodology*; *Genetics of Complex Traits Through the Life Cycle*).

1.1 *Univariate Results*

Most genetic studies of personality are based on very large twin and family samples. Personality has been assessed on over 50,000 pairs of twins in the national twin registries in Australia, Finland, Norway, Sweden, and The Netherlands, on volunteer twin registries in the USA, and on the 30,000 individuals in the Virginia study of twins and their families. Numerous smaller twin, family, and adoption studies have also been performed. Many of the studies (particularly those in

the national twin registries) have assessed personality, using versions of the Eysenck Personality Inventory. However, during the 1980s and 1990s increasing amounts of data have become available concerning other personality inventories, such as the Five-Factor Model of Costa and McCrae (McCrae and Costa 1987) and the Seven-Factor Model of Cloninger (Cloninger 1998).

Remarkably, there is considerable consensus across studies and cultures. Genetic effects account for on average 35–40 percent of the variation in personality traits such as neuroticism and extraversion. (In other words, the heritability of these traits is 0.35–0.40.) Evidence from studies of twins reared apart and adoption studies indicate that shared environmental influences account for another 5–10 percent of the variation in most personality traits, and the remaining variation is individual specific or 'non-shared' environmental variation. In other words, people differ for personality. Most of the individual differences we observe are due to individual specific experiences not shared by other family members and about 40 percent of the differences are due to genetic differences among us. The extent to which family members are similar to one another reflects primarily their genetic similarity, and only in part being raised in the same home.

There is some heterogeneity in genetic effects on personality. Traits related to extraversion, such as impulsivity, openness to experience, and sensation seeking tend to have slightly higher heritability estimates than those related to neuroticism and instability. Furthermore, the genetic influences for the extraversion-related traits may in part be operating in a nonadditive manner. That is, there may be interactions within or between genetic loci, such that the effect of one gene may depend on another gene. The latter effect, known as epistasis (see *Genetics of Complex Traits Through the Life Cycle*), may complicate efforts to identify specific genes that are of importance for personality traits.

Heritability estimates may be on average 40 percent for most personality traits, but there are traits for which genetic effects are of less importance. Examples of these are 'agreeableness,' 'guilt,' feelings of control over 'luck,' and 'cynic hostility.' For these measures, familial similarity is primarily due to environmental influences shared by family members rather than genetic similarity.

1.2 Genetic Influences on Stability and Change

There has been a great deal of controversy as to whether personality is stable throughout the lifespan. Some developers of personality inventories argued that finding significant heritability for personality would support hypotheses of stability in personality. Their argument rested on the misconception that if genetic effects are important they are immutable and do not change. Despite the fact that humans are born with a full complement of genes, not all genes may be operating at all phases in life. Some genes turn on and others turn off at various transitional periods, while others may react dependent on external stressors. On the other hand, some genes may have a long-term effect that is stable across the lifespan (see *Genetics of Complex Traits Through the Life Cycle*). Thus, the issue of genetic contributions to stability for personality is empirically testable.

During infancy, there is less stability in temperament than is typically reported for adulthood. Nevertheless, results from longitudinal twin studies indicate that the moderate stability observed for infant temperament is influenced in part by genetic effects. During childhood, genetic effects are also important contributors to phenotypic (i.e., observable) stability. Longitudinal studies of adult twins confirm findings of substantial phenotypic stability in personality. Furthermore, the genetic effects themselves are very stable, i.e., the same genes appear to be influencing personality throughout adulthood. Nonetheless, genetic and environmental influences are equally important for phenotypic stability. Even though the genetic effects themselves are very stable, they are relatively less important than environmental effects, and thus the net influence on stability is reduced. Environmental effects are the most important influences on change. Thus, genetic influences on personality contribute to stability rather than to change. However not all personality stability can be explained by genetic influences alone.

1.3 Structure of Personality

Developments in theories of personality have been paralleled with developments in statistical analyses of genetically informative data. During the 1980s and 1990s, the three-factor approach to personality was gradually extended to the five-factor theory and the seven-factor theory (McCrae and Costa 1987, Cloninger 1998). The factor analytic approach generally applied at the phenotypic level is also applicable at the level of genetic and environmental influences. The question to be addressed is whether there is a genetic factor structure or an environmental factor structure that is similar to the phenotypic factor structure. Do genetic or environmental effects mediate the associations among various aspects of personality? This line of analysis has only recently been applied and promises to be an important area for future research concerning the structure of personality.

1.4 Genetic Mediation of Associations with Personality

Health psychologists have demonstrated that personality is important for health outcomes (such as

cardiovascular disease, cancer, and osteoarthrosis) and response to treatment such as recovery from surgery. Current research efforts are focusing on the extent to which genetic and environmental effects mediate the associations between personality and health outcomes. By evaluating these associations, much can be learned about the mechanisms by which personality affects health. For example, neuroticism in young adulthood is predictive of joint pain and osteoarthritic symptoms some 25 years later in midlife. Genetic influences in common to neuroticism and joint pain and shared familial environmental influences mediate this association. There are no genetic influences unique to joint pain. Because neuroticism preceded the joint pain, it is likely that genetic effects are important for neuroticism which in turn influences joint pain and osteoarthritic symptoms through response to stress, immune, and inflammatory processes (Turk-Charles et al. 1999).

2. *Molecular Genetic Associations with Personality Traits*

The second major approach to the study of genetic influences on personality is to test whether specific genetic variants, or alleles, are associated with mean level differences in personality traits. Because personality traits are most often continuous measures and heritability analyses suggest that these traits are polygenic, the challenge is to find genes of medium effect size that influence personality levels. These genes are known as Quantitative Trait Loci or QTLs. Most efforts to identify which genes may be associated with personality traits are currently focusing on genes involved in neurotransmission, such as dopamine receptor genes and serotonin receptor and transporter genes. Many of the findings are only preliminary, and often with contrasting results or lack of replication in other studies or populations.

Finding genes that have relatively small effects will present great challenges to behavioral and genetic scientists, for several reasons. First, because many genes are likely to be involved, each with a small effect, very large sample sizes will be necessary to obtain sufficient statistical power. Second, many of the genes may be interacting with other genes, a phenomenon known as epistasis. For example, a particular allele for gene A may only influence personality negatively if a specific variant of gene B is present. Preliminary findings from large population-based studies suggest that some epistasis exists for personality and genes in the dopaminergic system. Third, there may well be substantial interactions between genes and environments that hamper the potential to find genotypic or even environmental associations. The impact of negative life events is greater in individuals with a genetic susceptibility to major depression, suggesting an interaction between genetic and environmental influences. Similarly, the effects of some allele may be enhanced or reduced depending on environmental circumstances. Again, large samples with genetic and detailed environmental characterization will be necessary to delineate the respective roles of genes and environments. In this fashion, scientists will come closer to understanding not only how much, but how genes influence personality.

3. *Future Directions*

Behavioral genetic analyses have provided a firm foundation for further studies on genetic influences on personality. Genetic influences are indeed important, and appear to be important for stability rather than change in personality. Further analyses of the genetic and environmental architecture of personality will expand our understanding of the associations within and among domains, such as the structure of personality and how personality influences health and other behavioral outcomes. Now that the Human Genome Project has provided a genetic map, the challenge to scientists will be to understand the function of the residents at the 30,000 addresses. Many genes, each of small effect, in interaction with each other and in interaction with the environment, are likely to influence variation in normal personality.

See also: Behavioral Genetics: Psychological Perspectives; Cultural Evolution: Overview; Evolution, History of; Gender Differences in Personality and Social Behavior; Genetic Studies of Behavior: Methodology; Genetics of Complex Traits Through the Life Cycle; Personality Development and Temperament; Personality Development in Adulthood; Personality Structure; Temperament: Familial Analysis and Genetic Aspects

Bibliography

Cloninger C R 1998 The genetics and psychobiology of the seven-factor model of personality. In: Silk K R (ed.) *Biology of Personality Disorders*. American Psychiatric Press, Washington, DC

Eaves L, Eysenck H, Martin N 1989 *Genes, Culture and Personality: An Empirical Approach*. Academic Press, London

Loehlin J C 1992 *Genes and Environment in Personality Development*. Sage, Newbury Park, CA

McCrae R R, Costa P T Jr 1987 Validation of the Five-Factor Model of personality across instruments and observers. *Journal of Personality and Social Psychology* **52**: 81–90

Turk-Charles S, Gatz M, Pedersen N L, Dahlberg L 1999 Genetic and behavioral risk factors for self-reported joint pain among a population-based sample of Swedish twins. *Health Psychology* **18**: 1–11

N. L. Pedersen

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Genetic Testing and Counseling: Educational and Professional Aspects

The sequencing of the human genome has enabled scientists to accelerate substantially the rate of discovery of gene mutations associated with disease. As these discoveries continue to stimulate the development of clinical genetic tests for disease susceptibility, there is a growing need to effectively educate and counsel individuals about the implications of this information. Genetic counseling is the professional practice established to provide this service. The unique familial aspects of genetic susceptibility and conditions have led to the development of medical genetics counseling subspecialties in medicine and nursing, as well as specialized master's degree programs in genetic counseling. While primarily a health communication process, genetic counseling also involves the discussion of relevant psychosocial issues. Thus, it may best be described as psychoeducational. Genetic counseling providers communicate with individuals and their relatives to improve their technical understanding of the cause of genetic conditions and genetic risk. In addition, they address feelings and coping strategies related to the condition and provide support in making preventive, reproductive, and other health-related decisions.

1. *Genetic Counseling*

1.1 *The Role of Genetic Counselors*

Members of families with inherited disease often ask not only why the condition occurred, but also why it happened to them. They seek an opportunity to find personal meaning in the senseless nature of it, and search for ways to live with the risk of future disease (Kessler 1979). In this sense, many of the psychological issues associated with inherited diseases are not unique. However, the shared, familial nature of genetic risk raises additional issues that are distinctive. Inherited risk implies that what is a health risk to one individual may also be a risk to their biological relatives. In addition, when a client seeks information about inherited patterns of disease risk, it may have unanticipated consequences. Information previously kept private (such as nonpaternity or adoption) may be revealed, producing enduring effects on family relationships. Yet there also may be potential harms from failing to disclose genetic information. Individuals who remain uninformed of their risk of a preventable disease, such as advanced-stage colorectal cancer, may suffer needless morbidity or mortality.

Because of their training and expertise, genetic counselors are uniquely positioned to assess and explain the genetic nature of a disease and to address its implications. The process aims to be therapeutic in exploring with clients their feelings of guilt, shame, fatalism, or social isolation. Genetic counseling requires health care practitioners with expertise in medical genetics, knowledge of the disease for which the individual is at risk, effective communication skills, and short-term therapeutic counseling skills.

1.2 *Genetic Counseling as a Profession*

Until recently, genetic counseling focused primarily on risks for and consequences of rare conditions. Consequently, there are few genetic counseling providers worldwide. Also, because most of the evidence concerning the inherited causes of many diseases is relatively new, the practice of genetic counseling is only about 50 years old (Resta 1997). The short history of the field is reflected in the limited, but growing empirical evidence concerning the process of genetic counseling and its effectiveness in facilitating informed decision making and psychological adjustment.

The type of training received by genetic counseling providers is diverse, and thus it is likely that practices vary accordingly (Biesecker and Marteau 1999). In the United Kingdom and Australasia, genetic nurses often work alongside medical geneticists to provide genetic counseling services. In the USA, Canada, South Africa and increasingly, in the UK, genetic counselors receive specialized training at the master's level to provide genetic counseling. In the vast majority of European and South American countries, medical geneticists or social workers provide genetic counseling. Genetic counseling services are not universally available, as underdeveloped nations usually have no specialized genetics health care providers.

1.3 *The Principle of Nondirectiveness*

Nondirectiveness is often presented as a guiding ethical principle of genetic counseling due to the importance of personal autonomy in reproductive decision-making. Genetic information even outside the prenatal setting often has implications for future childbearing decisions. The principle of nondirectiveness is intended to suggest that within genetic counseling, patients are not persuaded or coerced into specific reproductive outcomes. However, there exists disagreement in the profession about the meaning of nondirectiveness and its importance in practice (Kessler 1997). An international survey of genetic providers revealed that nondirectiveness is a nearly universal guiding ethical principle in the majority of countries that offer prenatal genetic testing (Wertz et al. 1990). An exception is China, where providers are more inclined to advise their pregnant patients to terminate affected pregnancies in the interest of reducing the incidence of children affected with genetic conditions (Mao and Wertz 1997).

2. Genetic Testing

Genetic testing describes any test for inherited factors that may influence one's health or the health of one's relatives. The method of performing a genetic test ranges from DNA sequencing or direct mutation detection, to chromosomal analysis and protein function testing. In addition, the circumstances in which a genetic test may be done vary substantially. Therefore the genetics education and counseling that surround decisions about testing or implications of test results differ in content and emphasis.

2.1 Diagnostic Testing

Genetic tests have most often been used to diagnose genetic conditions. This has been done in three different ways. First, a genetic test may be used when a child or an adult is suspected to be affected with a certain condition based on the clinical features and pattern of inheritance. Second, newborn screening for a subset of rare largely treatable recessive genetic conditions has been conducted in the USA and most other developed counties for several decades. Third, prenatal genetic testing has been available since the early 1970s, used most often to diagnose Down syndrome and other chromosomal trisomies in the developing fetus.

The diagnostic tests used in a pediatric or adult medical genetics clinic are most often considered a part of good medical management when relatives (typically parents) or their health care providers are seeking to make a diagnosis. Genetics education and counseling typically addresses the genetic and medical implications of the diagnosis, as well as the psychological impact on the family of having to live with an affected family member, and the increased risk to other relatives.

Genetic counseling occurs infrequently in conjunction with newborn screening unless a diagnosis is made in an infant. When a genetic diagnosis is made, the parents usually require extensive education about medical management as well as recurrence risks. Although most of the conditions tested for are treatable (unlike the vast majority of single-gene conditions), they require exhaustive efforts by parents to keep their child on a specially restricted diet to ensure his/her health.

Prenatal genetic testing is most often preceded by genetic counseling due to the highly personal nature of reproductive decision making. In most countries where prenatal testing is available, it is offered as an option to pregnant women who are at increased risk of having an affected child. Most commonly, this concerns those over the age of 35 years. Prenatal genetic counseling addresses the personal values of the client or couple, as they are guided in the exploration of the potential consequences of adverse test results. This includes envisioning both what their lives might be like if they had an affected child, or how they would feel subsequent to terminating a desired pregnancy.

2.2 Predictive Testing

Genetic testing is also used to predict genetic risk. Carrier testing, for example, is used to identify those who carry a gene mutation that does not affect them directly, but may place their children at risk. Relatives of individuals affected with certain genetic conditions and members of certain populations (e.g., Ashkenazi Jews) known to be at increased risk for certain conditions, are frequently offered carrier screening. Most carrier screening programs are accompanied by pretest genetics education, since it is important that those found to be carriers do not misinterpret the test result as indicating they have the disease.

Presymptomatic genetic testing is used to predict certain risk to an individual. Presymptomatic testing for Huntington's disease, a degenerative neurological disorder, is one example. It is now possible for those at increased risk for Huntington's disease to undergo presymptomatic testing prior to the onset of clinical symptoms. Inevitably, those found to carry a gene mutation develop the disease, for which there is no treatment. Due to the dire nature of these circumstances, the counseling that proceeds such testing is often extensive. Before presymptomatic testing was available, the majority of individuals at increased risk of developing Huntington's disease reported that they would choose to undergo testing if it was developed. However, once testing was available, only a small minority of those at increased risk chose to go forward with testing.

The newest type of predictive genetic testing, predisposition or susceptibility testing, results in a risk estimate about one's own future health. The first predisposition tests developed assess inherited susceptibility for cancer. An increased risk for at least two different types of colon cancer and breast and ovarian cancer can now be determined in members of hereditary cancer families. There are also populations of individuals more likely to carry susceptibility mutations due to shared heritage.

Susceptibility tests are being developed for many other common diseases and will be offered to an increasingly large portion of the population of developed nations. As demand for these tests increases, alternatives to in-person genetic counseling for pretest education will need to be developed and evaluated. Computer and Internet-based resources are likely to grow in their importance as sources of genetics information. These technologies allow information to be tailored to the characteristics of the individual user, allowing genetic counselors to focus on counseling and decision making rather than education and information delivery.

Initial studies of cancer susceptibility testing suggest that the majority of people express high levels of interest, but that many members of high-risk families do not utilize testing once it is offered (Croyle and Lerman 1999). The exploration of the psychological and behavioral costs and benefits of this type of testing have only just begun. Although individuals found to be noncarriers of cancer gene mutations report an alleviation of distress after testing, many mutation carriers do not engage in health behaviors that would allow the early detection of disease. These observations suggest that social and behavioral scientists will continue to play a key role in identifying ways to translate the achievements of genetics and biotechnology into improvements in public health.

See also: Bioethics: Examples from the Life Sciences; Cancer Screening; Ethical Issues in the 'New' Genetics; Ethics and Values; Eugenics as an International Movement; Eugenics, History of; Eugenics in Asia; Eugenics in Europe; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Engineering; Genetic Screening for Disease-related Characteristics; Mental Illness, Genetics of; 'New' Eugenics, The: Role in Society; Reproductive Medicine: Ethical Aspects; Risk Screening, Testing, and Diagnosis: Ethical Aspects

Bibliography

Biesecker B B, Marteau T 1999 The future of genetic counselling: An international perspective. *Nature Genetics* **22**: 133–7

Croyle R T, Lerman C 1999 Risk communication in genetic testing for cancer susceptibility. *Journal of the National Cancer Institute* **25**: 59–66

Kessler S 1979 *Genetic Counselling: Psychological Dimensions*. Academic Press, New York

Kessler S 1997 Psychological aspects of genetic counseling. XI. Non-directiveness revisited. *American Journal of Medical Genetics* **72**: 164–71

Mao X, Wertz D C 1997 China's genetic services providers' attitudes towards several ethical issues: A cross-cultural survey. *Clinical Genetics* **52**: 100–9

Resta R 1997 Historical perspective: Sheldon Reed and 50 years of genetic counseling. *Journal Genetic Counseling* **6**: 375–8

Wertz D C, Fletcher J C, Mulvihill J J 1990 Medical geneticists confront ethical dilemmas: Cross cultural comparisons among 18 nations. *American Journal of Human Genetics* **46**: 1200–13

R. T. Croyle

Genetics and Development

This article describes the major historical misconceptions of the role of genetics in the development of the individual, whether human or nonhuman animal, and also presents our current understanding of the necessarily co-operative roles of genes and environments in the developmental process.

1. *Nature vs. Nurture*

Many people believe that behavioral characteristics such as personality or intelligence are determined primarily, if not exclusively, either by heredity or environment. This belief, known as the nature-nurture controversy, has been with us for quite a long period of time and is the cause for much that is sorrowful and tragic in human history (e.g., the notion that some groups are behaviorally inferior or superior to other groups because of hereditary/genetic predispositions). The idea that genes are primarily responsible for some traits and that the environment is primarily responsible for other traits is now known to be scientifically invalid. This has been a hard-won scientific insight that has not yet percolated down to the mass of humanity. It is now known that both genes and environments are involved in *all* traits and that it is not possible to specify their respective weighting or quantitative influence on any trait. These crucial points are documented in the sections that follow.

2. *Behavioral Genetics*

In a remarkable experiment first conducted by E. C. Tolman (1924) and continued by R. C. Tryon (1942), it was found that it was possible to selectively breed laboratory rats for their ability to learn their way through a maze. Tryon selectively inbred the superior learners among each other and the poor learners among each other. After only seven generations of breeding the brightest rats within each of the brightest litters, and the dullest within each of the dullest litters, the performance of the two groups in learning the specific maze was virtually nonoverlapping, leading Tryon to label the superior line 'maze-bright' and the inferior line 'maze-dull.' It was as if Tryon had been able, by selective breeding, to create two lines of animals which differed greatly in their general learning ability. That turned out not to be the case, however. When Searle (1949) later compared the 'bright' and 'dull' lines on a variety of learning and performance tasks, the dull rats were sometimes superior and often equal to the rats from the bright line. Being bright or dull was not a property of the animals, but, rather, was restricted to the particular learning task that served as the basis for selectively breeding the animals. As it turned out, the specifics of the outcome of the breeding program were even further constrained by the manner in which the rats were reared, as will be described in Sect. 3.

Since the time of Tolman's and Tryon's breeding experiments, the field of behavioral genetics has been able to show that it is possible to selectively breed

Table 1
Examples supporting the notion that it is possible to selectively breed animals for differences in any behavior (more–less, better–poorer, higher–lower)

Species	Behavior
Rats	Maze learning
Mice	Speed in traversing runway
Rats	Spontaneous activity (wheel-running)
Rats	'Emotionality' in an open field
Mice	Sound-induced (audiogenic) seizures
Fruit flies	Orientation toward or away from light source (positive and negative phototaxis)
Rats	Saccharin preference
Chickens	Aggressiveness
Fruit flies	Upward or downward orientation with respect to gravity (negative and positive geotaxis)
Fruit flies	Mating speed
Mice	Alcohol preference
Fruit flies	General activity

The literature sources for these experiments can be found in Gottlieb 1992, p. 150.

animals for differences on any trait, solidifying the idea mentioned earlier that genes are involved in *all* traits. As will be documented below, the results of behavioral genetics breeding experiments are also dependent on the rearing environment being kept constant from generation to generation (recall that the rearing environment is involved in all traits, too). As can be seen in Table 1, both favorable and unfavorable behaviors can be made the subject of selective breeding, not only maze learning but spontaneous activity (wheel-running), 'emotionality,' sound-induced seizures, saccharin preferences, aggressiveness, mating speed, and so on.

As mentioned above, since all behavioral activities involve both genetic and environmental contributions, it is essential when doing selective breeding experiments to keep the rearing environment the same from generation to generation, otherwise the selective breeding experiments will be a failure. For example, in an experiment by Hood and Cairns (1989), mice were selectively bred for high and low aggressiveness. The standard rearing procedure was to keep the animals in isolation after they were weaned. After the high and low aggressive lines were established, to show the contribution of the isolation rearing procedure to the results of selective breeding Hood and Cairns reared half the animals from each line in social groups rather than in isolation. As can be seen in the right-hand side of Fig. 1, the high aggressive line did not differ from the low aggressive line when the animals were reared in social groups; the high aggressive line was only highly aggressive when reared in isolation (left-hand side of Fig. 1), the rearing condition under which the original selective breeding took place. These results lead into the topic of Sect. 3.

3. *Reaction Range vs. Norm of Reaction*

Until very recently some biologists and psychologists held the incorrect belief that genes strictly canalize behavioral development, creating a predictable reaction range when rearing environments are changed from the normal or usual (species-typical) to highly

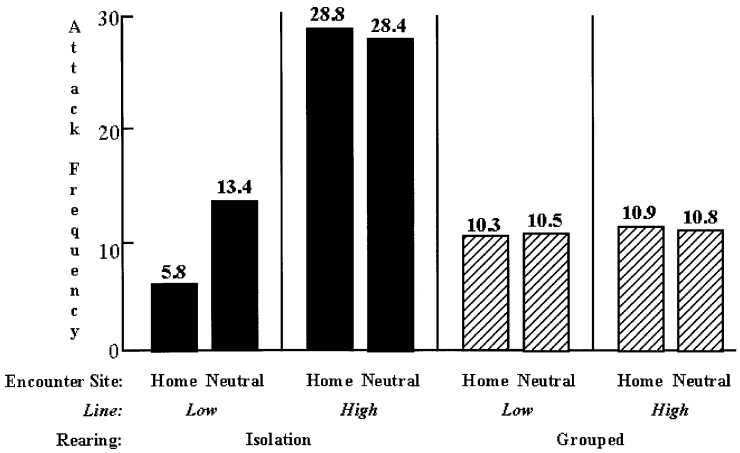

Figure 1
Average attack frequency by male mice in two lines selectively bred for either high or low attack frequency. The original selection experiment involved males reared in isolation prior to testing. In this experiment, half of the males were reared in isolation and the others were group-reared, greatly reducing the attacks by the high line. During the test, the mice met in either the home cage, in which the mouse being tested was an intruder, or in a neutral cage (based on data from Hood and Cairns 1989. Figure modified from Gottlieb et al. 1998, p. 249)

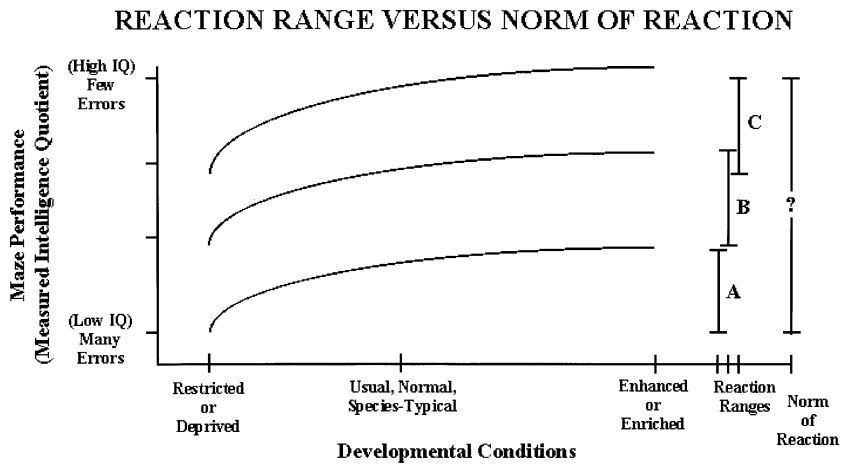

Figure 2
Contrasting predictions of reaction range and norm of reaction for maze performance or IQ of three genotypes or individuals, as a consequence of being reared under three developmental conditions. The reaction-range concept predicts a narrow range of variation for each individual over the three rearing conditions, whereas the norm of reaction predicts an indefinitely broader range of outcomes when the rearing conditions are changed from the original condition (usual, normal, species-typical) (modified from Gottlieb 1992, p. 178)

favorable (enriched) or unfavorable (deprived) conditions.

This point of view was expressed by one authority as follows:

> Canalization is the genetic limitation of phenotypic development to a few possible phenotypes…The genetic restriction of possible phenotypes is a result of coadapted gene complexes (sets of genes that have evolved together) that buffer the developmental pattern against deviation outside of the normal, species range of variation…The limitation of possible phenotypes to a few rather than many is a fact of genotypes' ranges of reaction (Scarr-Salapatek 1976, p. 63).

The belief in the sheerly genetic canalization of behavioral development leads quite logically to the reaction-range idea. Thus, the basic idea of a reaction range is that genes determine (canalize) the upper and lower limits of psychological development so that, given a knowledge of the outcome of development under one rearing condition, one can make linear predictions of the upper and lower limits of outcomes under other rearing conditions (see as an example Fig. 2).

In contrast to the reaction range, the norm of reaction holds that the outcome of development is an emergent (epigenetic) property that is a nonlinear result of co-actions occurring at all levels of the developmental system (including but not restricted to genes and environments, as elaborated in Sect. 4). Thus, it is not possible to predict outcomes from one rearing circumstance to the next. Unfortunately, it is the reaction range, not the reaction norm, that has had, until very recently, wide currency in psychology, as evidenced in many textbooks on developmental psychology. These texts reproduced a version of Fig. 2 depicting only the reaction-range concept for learning ability or mean IQ for three hypothesized genotypes reared in very different environments, and not the norm of reaction (far right).

The difference between the predictions of the reaction-range and reaction-norm concepts is testable. As indicated in Fig. 2, the reaction-range idea predicts that the ranking of the different genotypes on IQ or learning ability will remain the same when they are reared in different environments. In the enriched environment, each genotype will show an improvement but the 'brightest' genotype will still rank above the other genotypes. In the deprived rearing environment each genotype's performance will be poorer but the brightest genotype will still rank above the other genotypes. As shown in the far right of Fig. 2, the norm of reaction makes no prediction on relative ranking when the genotypes are reared in environments (enriched or deprived) that differ from the rearing environment in which the selective breeding took place.

There is only one empirical study in the psychological literature that explicitly addresses the reaction-norm concept, a study by Cooper and Zubek (1958). The results very clearly support the reaction-norm concept. It is interesting to note that the study was carried out with the idea of a reaction range in mind, and it is sometimes erroneously cited as supporting the reaction-range concept. Cooper and Zubek reared maze-bright and maze-dull rats in either an enriched

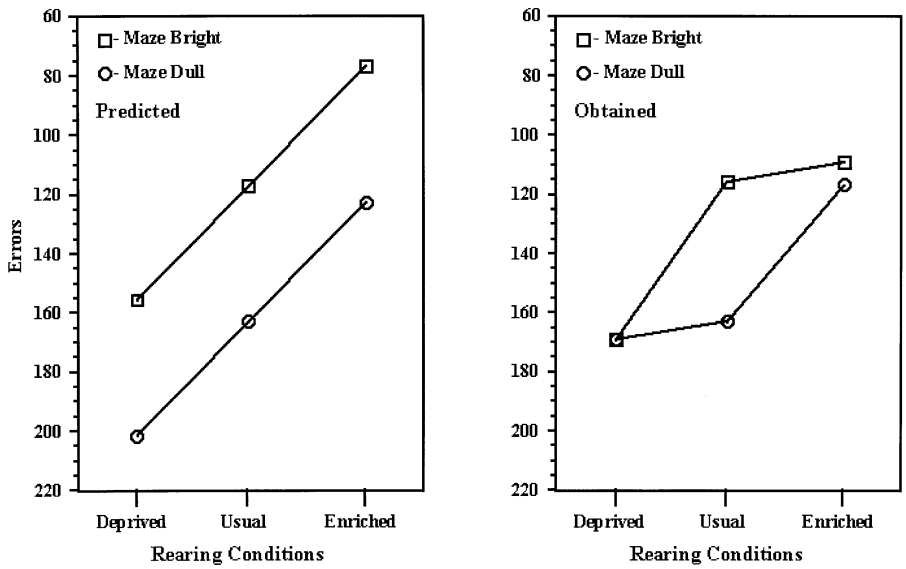

Figure 3
Behavioral reaction range (predicted) and norm of reaction (obtained) for maze-bright and maze-dull rats' performance in a Hebb–Williams maze after rearing in three different environments. Obtained deprived and enriched data points from Cooper and Zubek (1958); obtained usual data points from Hughes and Zubek (1956). Only the obtained usual data points are significantly different from each other (from Gottlieb 1995, p. 137)

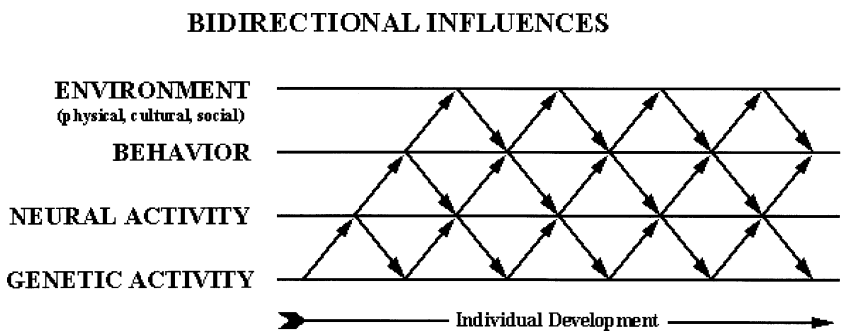

Figure 4
Depiction of the completely bidirectional nature of genetic, neural, behavioral, and environmental influences over the course of individual development (from Gottlieb 1992, p. 186)

or a restricted environment and then tested them in a Hebb–Williams learning maze. Since they had the reaction-range concept in mind in performing the experiment, Cooper and Zubek thought that the learning of both the bright and dull rats would improve relative to each other under the enriched rearing circumstances and would be poorer relative to each other when reared under the restricted (deprived) condition. This prediction is illustrated in the left-hand side of Fig. 3. Instead, as shown on the right-hand side of Fig. 3, they found equality of performance under both of the new rearing conditions. The dull rats made as few errors as the bright rats after enriched rearing and the bright rats made as many errors as the dull rats after restricted rearing.

When the so-called bright and dull rats were tested in the Hebb–Williams maze after being reared in their usual way (neither enriched nor deprived), a significant difference between the lines appeared (middle points, right-hand side of Fig. 3). The reason is that this developmental situation repeats the rearing condition under which the original selective breeding for superior and inferior performance was carried out (Hughes and Zubek 1956). If the reaction-range idea were correct and the genes coded for upper and lower limits of learning ability, when these lines were reared under

Table 2
Normally occurring environmental and behavioral influences and gene activity

Species	Environmental signal or stimulus	Resulting alteration
Nematodes	Absence or presence of food	Diminished or enhanced neuronal *daf-7* gene mRNA expression, inhibiting or provoking larval development
Fruit flies	Transient elevated heat stress during larval development	Presence of proteins produced by heat shock and thermotolerance (enhanced thermal regulation)
Fruit flies	Light-dark cycle	Presence of PER and TIM protein expression and circadian rhythms
Various reptiles	Incubation temperature	Sex determination
Songbirds (canaries, zebra finches)	Conspecific song	Increased forebrain mRNA
Hamsters	Light-dark cycle	Increased pituitary hormone mRNA and reproductive behavior
Mice	Acoustic stimulation	Enhanced c-*fos* expression, neuronal activity, and organization of the auditory system
Mice	Light-dark cycle	c-*fos*-induced mRNA expression in hypothalamus, circadian locomotor activity
Rats	Tactile stimulation	Enhanced c-*fos* expression and increased number of somatosensory (sense of touch) cortical neurons
Rats	Learning task involving vestibular (balance) system	Change in nuclear RNA base ratios in vestibular nerve cells
Rats	Visual stimulation	Increased RNA and protein synthesis in visual cortex
Rats	Environmental complexity	Increased brain RNA diversity
Rats	Prenatal nutrition	Increase in cerebral DNA (increased number of brain cells)
Rats	Infantile handling; separation from mother	Increased hypothalamic mRNAs for corticotropin-releasing hormone throughout life
Cats	Visual stimulation	Increased visual cortex RNA complexity (diversity)
Humans	Academic examinations taken by medical students (psychological stress)	Reduced mRNA activity in interleukin 2 receptor (immune system response)

MRNA = messenger RNA; PER and TIM are proteins arising from activity of per (period) and tim (timeless) genes; activity of c-*fos* genes leads to production of C-FOS proteins. Sources documenting the findings can be found in Gottlieb 1998, p. 796

enriched or restricted conditions, the relative differences between them would be preserved. Instead, the experiment shows that the genes are part of a developmental system. The highly specific consequences of rearing under a certain developmental condition were realized by selective breeding under that condition. As called for by the norm-of-reaction concept, selective breeding under one developmental regimen (environment) does not predict outcomes under different rearing conditions. The results of selection depend on the entire developmental manifold, not merely on the genes that are involved.

4. Understanding Development from Genes to Behavior

To provide some insight into what would today constitute a truly developmental behavioral-genetic analysis, Fig. 4 shows a diagram of the four major levels that are involved in moving from genetic activity to environment and back again. This is a fully bidirectional co-actional system, with traditionally trained developmental psychologists working at the Behavior ↔ Environment level, developmental neuropsychologists and neuroscientists at the Neural

Activity ↔ Behavior ↔ Environment levels, and biologists beginning now to work on the Genetic Activity ↔ Neural Activity levels. Viewed in this way, developmental understanding or explanation is a multilevel affair involving at least culture, society, immediate social and physical environments, anatomy, physiology, hormones, cytoplasm, and genes. Hierarchical, multilevel, or developmental systems' analysis is methodologically not theoretically reductionistic in the sense that psychological understanding or explanation does not come from lower levels (or reside in the lower levels), but that both higher and lower levels of analysis are necessary to explain developmental outcomes. The term co-action is used to emphasize the multilevel systems' nature of developmental analysis.

The notion that genetic activity is necessarily influenced by supragenetic factors (factors above the genetic level) may be unfamiliar, in part because of what has come to be known as the Central Dogma of Molecular Biology. The Central Dogma holds that DNA → RNA → Protein, in which DNA serves as a template for RNA which then serves as a template for the synthesis of a protein. In the Central Dogma there are master genes which contain a blueprint of the organism, so 'information' is said to flow outward from these master genes providing instructions for individual development. The actual story is rather different. Rather than being encapsulated (or 'safely sequestered in the nucleus of the cell,' as some biologists have written), the genes require signals from the internal and external environment to function appropriately. As mentioned above, genes are part of the developmental system of influences; they do not stand outside or apart from the developmental system as the Central Dogma would seem to indicate. Gene expression is influenced by cytoplasmic factors, which provide the immediate environment of the nucleus of the cell. (Structural DNA is found in the nucleus of each cell.) Cytoplasm can be influenced by the external environment of the organism (reviewed by Ho 1984) via the behavior of the organism, which gets the organism into and out of exposure to different environmental influences (physical, social, cultural). Likewise, hormones make their way into the nucleus of the cell and trigger gene expression, and hormones themselves are responsive to the organism's experiences in the external world. In fact, the action of some genes is dependent on sensory stimulation early in development (reviewed by Arenander and Herschman 1993). Finally, different proteins are formed depending on the particular factors influencing gene expression (Davidson 1986). Hence, genes do not stand outside the developmental system acting as independent causes. They are part of the developmental system (for a detailed discussion, see Wahlsten in press). One of the main reasons for the failure of gene therapy as currently conceptualized and practiced (Rosenberg and Schechter 2000) is that the genes do not function as independent causes; they do not stand outside the developmental system.

Since there are in fact no master genes telling other genes what to do, in order for genes to function appropriately they require signals from the internal and external environment. Table 2 presents a sampling of the normally occurring external environmental and behavioral signals that activate genes during the normal course of development in a variety of species including humans. The findings summarized in Table 2 substantiate the main point of this article: genes and environments necessarily cooperate during the course of individual development.

See also: Behavioral Genetics: Psychological Perspectives; Genes and Behavior: Animal Models; Genetic Factors in Cognition/Intelligence; Genetics of Complex Traits Through the Life Cycle; Human Evolutionary Genetics; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation

Bibliography

Arenander A T, Herschman H R 1993 Primary response gene expression in the nervous system. In: Loughlin S F, Fallon J H (eds.) *Neurotrophic Factors*. Academic Press, San Diego, CA

Cooper R M, Zubek J P 1958 Effects of enriched and restricted early environments on the learning ability of bright and dull rats. *Canadian Journal of Psychology* **12**: 159–64

Davidson E H 1986 *Gene Activity in Early Development*. Academic Press, Orlando, FL

Gottlieb G 1992 *Individual Development and Evolution: The Genesis of Novel Behavior*. Oxford University Press, New York

Gottlieb G 1998 Normally occurring environmental and behavioral influences on gene activity: From central dogma to probabilistic epigenesis. *Psychological Review* **105**: 792–802

Gottlieb G, Wahlsten D, Lickliter R 1998 The significance of biology for human development: A developmental psychobiological systems perspective. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1

Ho M-W 1984 Environment and heredity in development and evolution. In: Ho M-W, Saunders P T (eds.) *Beyond Neo-Darwinism: An Introduction to the New Evolutionary Paradigm*. Academic Press, London

Hood K E, Cairns R B 1989 A developmental-genetic analysis of aggressive behavior in mice: IV. Genotype-environment interaction. *Aggressive Behavior* **15**: 361–80

Hughes K R, Zubek J P 1956 Effect of glutamic acid on the learning ability of bright and dull rats. I. Administration during infancy. *Canadian Journal of Psychology* **10**: 132–8

Rosenberg L E, Schechter A N 2000 Gene therapist, heal thyself. *Science* **287**: 1751

Searle L V 1949 The organization of hereditary maze-brightness and maze-dullness. *Genetic Social and General Psychology Monographs* **39**: 279–325

Scarr-Salapatek S 1976 Genetic determinants of infant development: An overstated case. In: Lipsett L (ed.) *Developmental Psychobiology: The Significance of Infancy*. Erlbaum, Hillsdale, NJ

Tolman E C 1924 The inheritance of maze-learning ability in rats. *Journal of Comparative Psychology* **4**: 1–18

Tryon R C 1942 Individual differences. In: Moss F A (ed.) *Comparative Psychology*. Prentice-Hall, New York

Wahlsten D in press Genetics and the development of brain and behavior. In: Valsiner J, Connolly K (eds.) *Handbook of Developmental Psychology*. Sage, London

G. Gottlieb

Genetics and Mate Choice

In a sexually reproducing species, an individual must mate to reproduce. Mates are not chosen at random, but apparently according to sexual preferences. Such preferences may be genetically fixed at birth, or learned—acquired by sexual imprinting, mate-choice copying, etc. To the extent that they are reflected in actual matings, preferences may have immediate consequences for reproductive success and far-reaching effects on genetic evolution.

1. Avoidance of Inbreeding

Consanguinity is an important criterion of mate choice. Throughout the vertebrate, invertebrate, and plant kingdoms, the propensity to inbreed (mate with a blood relative) varies widely (Thornhill 1993). Some species of hermaphroditic plants mate predominantly with themselves—the seed is fertilized by pollen from the same individual. In others, mechanisms such as self-incompatibility have evolved to prevent self-fertilization. Hence, self-fertilization is not the unavoidable consequence of female and male functions occupying the same body. In species where the two sexes are separate, nuclear family incest—matings between father and daughter, mother and son, or full sibs (brother and sister with the same mother and father)—is the most extreme form of inbreeding. For some insects, mother–son and brother–sister matings are routine.

On the other hand, close inbreeding is unusual in birds and mammals. Of special interest are the nonhuman primates, where nuclear family incest, in particular, is rarely observed. Adult males of the chimpanzee, our closest relative, have seldom been seen mating with their mothers. Most, but not all, females refuse the sexual advances of males in their natal group old enough to be their fathers. The incidence of copulations between inferred maternal (possibly full) sibs is low (Pusey 1980). However, matings between paternal half sibs may not be uncommon in the gorilla.

The claim is often made that behavioral mechanisms have evolved in birds and mammals to prevent close inbreeding. However, avoidance of incest is not necessarily based on recognition of family members. Because of a sex difference in dispersal, sons and daughters may not encounter their opposite-sex sibs or parents as adults. In the black-tailed prairie dog, for example, daughters usually remain in their natal area whereas the sons emigrate. Fathers may also leave or be evicted before their daughters mature. Under these circumstances, incest may be prevented even when matings are indiscriminate. The reason for such a dispersal pattern is unknown, but avoidance of inbreeding is a possible explanation. It is perhaps surprising, then, that the incidence of close inbreeding, mostly between half sibs, does not differ from the number expected if matings were occurring at random in the study area (Hoogland 1992).

Nevertheless, kin recognition has been implicated in some field studies and demonstrated in the laboratory. A likely mechanism is 'sexual imprinting,' the phenomenon in which sexual preferences expressed later in life are acquired through exposure at a very young age to other individuals, usually parents or sibs. A prairie dog female, for example, is less likely to come into estrus when her father is still present. Similarly, an older female of the medium ground finch mates with a male who sings a song differing from her father's. Since song is culturally transmitted from father to son in this species, close kin are apparently recognized and avoided (Grant and Grant 1996). Finally, Japanese quail reared with sibs choose the company of first cousins over that of either sibs, third cousins, or unrelated individuals. Bateson (1982) infers from this experiment that quails learn the characteristics of close kin and subsequently develop a preference for mates differing somewhat from familiar individuals.

It is interesting that cousin marriage, which is a comparatively mild form of inbreeding, is highly popular in some human societies. In particular, in south India and in many Muslim populations, the incidence of first-cousin marriages exceeds 30 percent (Bittles 1994). The reasons for this preference are likely different from those operative in quails.

In contrast to cousin marriage which exhibits much regional variation in incidence, nuclear family incest, marital and nonmarital, has been prohibited and usually avoided in almost all human societies past and present that we know about. A well-known counter example comes from Roman Egypt, where one-sixth of all marriages among commoners were between full sibs (Bagnall and Frier 1994). The evidence comprises over 300 census returns preserved on papyri, from which pedigrees have been reconstructed. Although this example is certainly the exception rather than the rule, it cautions against naive explanations for the existence of incest taboos.

Westermarck (1891) proposed that 'there is an innate aversion to sexual intercourse between persons

living closely together from early youth.' In particular, full sibs reared together are expected to avoid mating with one another, but not full sibs reared apart. Avoidance is also predicted when unrelated boys and girls are reared together in the same household. Thus, studies on the Chinese custom of *sim-pua* and on marriage patterns among Israelis reared together as children in kibbutzim seem to support the 'Westermarck hypothesis.' For example, the *sim-pua* is a girl who is adopted, often as an infant, into a family with a young boy, to be his future bride ('little daughter-in-law'). After being brought up as sister and brother, they are forced to marry. Compared to other arranged marriages, the fertility of this type of marriage (minor marriage) is low, and the divorce rate is high (Wolf 1995).

This result is usually regarded as reflecting sexual dissatisfaction and supporting the Westermarck hypothesis. Nevertheless, marriages between foster sibs were in fact consummated. Moreover, in the case of institutionalized sib marriage in Roman Egypt, any aversion that may have existed was overridden. Hence, even if negative sexual imprinting occurs, it is apparently not sufficient to prevent an incestuous mate choice when other cultural pressures intervene.

If behavioral mechanisms such as sexual imprinting and sex-biased dispersal have evolved by natural selection to prevent close inbreeding in birds and mammals, there must be a selective disadvantage to mating with close kin. Sexual imprinting and sex-biased dispersal are assumed to be genetically determined capacities—hence innate. An inbred (born of related parents) offspring is known to have lower fitness than an outbred (born of unrelated parents) one. The decrease in fitness, normalized by the fitness of the outbred offspring, is called the inbreeding depression, and is estimated in practice by comparing survival to some arbitrary age. Inbreeding depression is believed to be caused by homozygosity for (having two copies of) deleterious recessive genes. The more closely inbred an individual is, the more severe the inbreeding depression is expected to be.

Estimates of the inbreeding depression associated with nuclear family incest average 33 percent for 38 species of mammals in captivity (Ralls et al. 1988). In humans, one study with good controls obtained an estimate of 29 percent, although this estimate is inflated by the inclusion of morbidity (Adams and Neel 1967).

The conditions under which avoidance of inbreeding will evolve are difficult to compute, but some theory, mostly heuristic, exists for full-sib matings (Maynard Smith 1980). If all males have the same number of mates, then sib mating is selected against, and avoidance will evolve. This would be true of a strictly monogamous species. However, if the inbreeding depression is less than one-third, and a female by being incestuous can give her brother one extra mating, then it will be in her interests to do so.

Another factor that may offset the natural selection against incestuous matings is 'reproductive compensation'—if offspring lost due to inbreeding depression are replaced. Similarly, if ecological or economic conditions are such that sibs can mate earlier than unrelated individuals, then lifetime reproductive success may be comparable—or even greater—for the incestuous pair. Keller and Arcese (1998) argue that this may be the reason why song sparrows do not purposefully avoid close inbreeding. In fact, the incidence of full-sib and parent–offspring matings was as expected if song sparrows were mating indiscriminately.

In modern human societies, age at marriage is lower when the spouses are cousins—and generally when related—than when they are unrelated. Perhaps because of this, or because of reproductive compensation, cousin marriages enjoy a higher fertility (live births) than nonconsanguineous unions (Bittles 1994). When on-going marriages in Roman Egypt are broken down by age, there is a relatively higher incidence of brother–sister and other close-kin marriages among the younger couples. Therefore, it is possible that sibs were marrying earlier than unrelated couples. Moreover, one of the most fertile couples are sibs, with eight surviving children (Bagnall and Frier 1994). If sib mating permitted an early start in reproduction during much of human evolutionary history, then sib mating rather than its avoidance may have been favored by natural selection.

2. *Preference for a Mate Differing at the Major Histocompatibility Complex*

In (positive) assortative mating, psychologically, behaviorally, or physically similar individuals pair up. Negative assortative, or disassortative, mating occurs as a result of attraction between dissimilar individuals. An example of positive marital assortment is the tendency of deaf persons to marry one another. Another example is the high correlation between the intelligence quotients of spouses. In the latter example, however, it is not known whether someone with an IQ comparable to one's own is actually preferred as a marriage partner. Rather, people who marry one another have often experienced similar environments as children, and since the environment is a major determinant of IQ, spouses may incidentally have similar IQs.

Recently, Wedekind et al. (1995) suggested that humans prefer a mate differing at the major histocompatibility complex (MHC). MHC is an essential part of the vertebrate immune system. It comprises many genes, each of which is highly variable and exists in many alternative forms called alleles. In the human, the genes are called HLA-A, HLA-B, HLA-C, etc. (HLA = human leucocyte antigen.) For illustrative purposes, let us focus on the HLA-A gene and assume

that four alleles are present, which will be called A1, A2, A3, and A4. The pair of alleles in an individual defines the genotype of that individual. The genotypes could be A1A1, A2A2, A3A3, or A4A4, which are called homozygotes, or A1A2, A1A3, A1A4, A2A3, A2A4, or A3A4, which are called heterozygotes.

Mates can share two alleles (as when the genotypes of husband and wife are A1A2 and A1A2), one allele (A1A2 and A2A3, for example), or no alleles (A1A3 and A2A4, say). The hypothesis of disassortative mating asserts that sexual attraction is negatively correlated with the number of alleles shared. Hence, an individual of genotype A1A1, say, given a choice of three different partners whose genotypes are A2A2, A1A2, and A1A1, say, is predicted—all other things being equal—to prefer the first (A1A1 and A2A2 share no alleles) over the second (A1A1 and A1A2 share one allele) over the third (A1A1 and A1A1 share two alleles).

Theoretically speaking, disassortative mating for MHC may make evolutionary sense. Yamazaki et al. (1976) argue as follows. Heterozygotes for MHC genes have a higher fitness than homozygotes, because the presence of two alleles rather than just one would permit an immunological response to a wider range of antigens. Although the genotype that an individual is born with cannot be altered, the genotype of offspring is under some personal control. Namely, by choosing an MHC-dissimilar partner, the chances are improved that offspring will be heterozygotes. A second reason why disassortative mating for MHC might be favored by natural selection is that it leads to avoidance of inbreeding.

A preference for MHC-dissimilar mates is fairly well established in house mice. There is good evidence, at least in mice, that this preference is mediated by body odor, which is influenced by the MHC genes. More precisely, mice imprint on the MHC identities of the individuals by, or with, whom they are raised, who under normal circumstances would be extended family. They subsequently choose a mate differing in MHC from these individuals. Since, close relatives are on average more similar for their MHC than unrelated individuals, the resulting preference would be for MHC-dissimilar mates.

The human evidence is of two kinds. First, in a provocative experiment, male subjects were asked to each wear a T-shirt, and female subjects to rate the odors of these T-shirts for pleasantness (sexiness). The subjects, all of whom were students at a Swiss university, were also typed for their HLA-A, HLA-B, and HLA-DR genes. It turned out that the females preferred the odors of T-shirts worn by MHC-dissimilar males to those worn by MHC-similar males (Wedekind et al. 1995).

Second, a direct test of whether humans mate disassortatively for MHC should be possible by typing married couples. To date, such an analysis has been done on the Hutterites (a North American reproductive isolate of European ancestry), South American Indians from the lower Amazon basin, and the Japanese. Only among the Hutterites has dissortative mating been demonstrated (Ober et al. 1997). In the other two populations, mating is apparently random with regard to MHC.

It is not clear why such contradictory results have emerged. One possibility is that, since the preference for MHC-dissimilar mates is weak, it may sometimes be overwhelmed by other biological or cultural factors. In particular, a preference for cousin marriages will mask any tendency towards disassortative mating. However, there is no evidence of cousin marriages among the couples sampled in the above mentioned studies.

3. *Genetically Fixed Preferences vs. Acquired Preferences*

In all of the above examples, mating preferences are apparently acquired by sexual imprinting. Despite the title of this article, convincing demonstrations of genetically fixed preferences are hard to find. Even in insects, appropriate social behavior is often contingent on learning to distinguish different categories of individuals—nestmates from intruders, for example.

Nevertheless, a possible example of a genetically fixed preference occurs in ladybird beetles. Ladybirds are polymorphic for color and markings, and females show a preference for melanic males. Majerus et al. (1982) subjected this preference to artificial selection by permitting only those females who mated with melanic males to lay eggs. When the experiment was continued for four generations, a substantial increase in the level of preference was observed.

Cultural transmission of preferences is also difficult to prove. Surprisingly, the best evidence comes not from humans, but rather from small tropical fish—guppies, sailfin mollies, and three-spined sticklebacks—where the phenomenon is known as 'mate-choice copying' (Dugatkin 1992). Mate-choice copying is said to occur when a female is more likely to choose a specific male after observing another female doing so. The following three-phase experiment is conducted to demonstrate mate-choice copying.

In phase one, a female (focal female) is given a choice of two males. If the focal female shows a consistent preference for one of these males, then in phase two, a second female (model female) is placed near the nonpreferred male in view of the focal female. Finally, in phase three, the model female is removed, and the focal female is again given a choice of the two males. Mate-choice copying is implied, if the focal female's preference is reversed as a result of seeing the model female and the nonpreferred male together. (It is not clear how long the reversal of preference is maintained.)

The experiment must control for confounding effects. For example, the focal female might be attracted by the nonpreferred male's activity during or after his interaction with the model female, rather than by the presence of the model female. Or the focal female may simply be exhibiting a tendency to join other females, as is apparently the case with fallow deer. White and Galef (1999), who investigated mate-choice copying in Japanese quail, point out other complications in the interpretation of such experiments.

Given this emphasis on socially learned sexual preferences in fish and birds, it is perhaps ironic that most research on humans—by the evolutionary psychologists—has attempted to identify preferences that are relatively impervious to social influence. For example, it is claimed that men are attracted to women with a low waist-to-hip ratio (WHR)—an hourglass figure. This appears to be true in several Westernized cultures. Moreover, since women with lower WHR are apparently more fecund (physiologically able to give birth) (Zaadstra et al. 1993), Singh (1993) argues that it is adaptive for a man to have such a preference, whence it has evolved to become a panhuman—culturally invariant—male characteristic. The logic is fine, but the prediction has yet to be verified. Recently, Yu and Shepard (1998) studied WHR preferences in an isolated Peruvian population unexposed to Western media. Contrary to prediction, an overweight woman with a high WHR was most attractive and desirable as a spouse for the men of this population.

See also: Divorce, Sociology of; Evolutionary Selection, Levels of: Group versus Individual; Family, Anthropology of; Family Theory: Economics of Marriage and Divorce; Incest, Inbreeding, and their Consequences; Incest Prohibition, Origin and Evolution of; Marriage; Natural Selection; Partner Selection across Culture, Psychology of; Personality and Marriage; Primate Behavior: Significance for Understanding Humans

Bibliography

Adams M S, Neel J V 1967 Children of incest. *Pediatrics* **40**: 55–62

Bagnall R S, Frier B W 1994 *Demography of Roman Egypt*. Cambridge University Press, Cambridge, UK

Bateson P 1982 Preferences for cousins in Japanese quail. *Nature* **295**: 236–7

Bittles A H 1994 The role and significance of consanguinity as a demographic variable. *Population Development Review* **20**: 561–84

Dugatkin L A 1992 Sexual selection and imitation: Females copy the mate choice of others. *American Naturalist* **139**: 1384–9

Grant B R, Grant P R 1996 Cultural inheritance of song and its role in the evolution of Darwin's finches. *Evolution* **50**: 2471–87

Hoogland J L 1992 Levels of inbreeding among prairie dogs. *American Naturalist* **139**: 591–602

Keller L F, Arcese P 1998 No evidence for inbreeding avoidance in a natural population of song sparrows (*Melospiza melodia*). *American Naturalist* **152**: 380–92

Majerus M E N, O'Donald P, Weir J 1982 Female mating preference is genetic. *Nature* **300**: 521–3

Maynard Smith J 1980 Models of the evolution of altruism. *Theoretical Population Biology* **18**: 151–9

Ober C, Weitkamp L R, Cox N, Dytch H, Kostyu D, Elias S 1997 HLA and mate choice in humans. *American Journal of Human Genetics* **61**: 497–504

Pusey A E 1980 Inbreeding avoidance in chimpanzees. *Animal Behaviour* **28**: 543–52

Ralls K, Ballou J D, Templeton A 1988 Estimates of lethal equivalents and the cost of inbreeding in mammals. *Conservation Biology* **2**: 185–93

Singh D 1993 Adaptive significance of female physical attractiveness: Role of waist-to-hip ratio. *Journal of Personality and Social Psychology* **65**: 293–307

Thornhill N W (ed.) 1993 *Natural History of Inbreeding and Outbreeding*. University of Chicago Press, Chicago, IL

Wedekind C, Seebeck T, Bettens F, Paepke A J 1995 MHC-dependent mate preferences in humans. *Proceedings of the Royal Society of London Series B* **260**: 245–9

Westermarck E 1891 *History of Human Marriage*. Macmillan, London

White D J, Galef B G 1999 Mate choice copying and conspecific cueing in Japanese quail, *Coturnix coturnix japonica*. *Animal Behaviour* **57**: 465–73

Wolf A P 1995 *Sexual Attraction and Childhood Association*. Stanford University Press, Stanford, CA

Yamazaki K, Boyse E A, Mike V, Thaler H T, Mathieson B J, Abbott J, Boyse J, Zayas Z A, Thomas L 1976 Control of mating preferences in mice by genes in the major histocompatibility complex. *Journal of Experimental Medicine* **144**: 1324–35

Yu D W, Shepard G H 1998 Is beauty in the eye of the beholder? *Nature* **396**: 321–2

Zaadstra B M, Seidell J C, Van Noord P A H, te Velde E R, Habbema J D K, Vrieswijk B, Karbaat J 1993 Fat and female fecundity: Prospective study of effect of body fat distribution on conception rates. *British Medical Journal* **306**: 484–7

K. Aoki

Genetics and the Media

Genetics is in the news as the scientific frontier, the cutting edge of current research. Every new discovery is reported in the media with optimistic expectations as a 'breakthrough'—even when the significance of a particular finding is in question. A survey of the coverage of genetics in the American press from 1970 to 1996 found that reports were overwhelmingly positive as journalists equated research with economic potential and medical progress (Lewenstein et al. 1998). Beyond the news, the gene has also become a prevalent media metaphor appearing in fictional narratives, popular magazines, visual images, and the cultural detritus of sound bites, slogans, and snippets: it's all in the genes. These media attribute extra-

ordinary powers to the gene, conveying a concept of 'genetic essentialism' (Nelkin and Lindee 1995). DNA in media narratives is an essentialist entity—the cause of disease, the definition of personhood, the location of the true self.

DNA is data without dimension, text without context, so that the equation of DNA with personhood requires a profound leap of faith. Yet, the gene as a solid source of identity holds media appeal. So too does the belief that the physical and behavioral traits of individuals, and the social and economic differences among groups are caused by inborn distinctions.

Genetic language has become a code in the media, readily applied and broadly understood. What is genetic is powerful, predictable, and permanent; 'hard-wired' in the human constitution. Underlying the notion of genetic destiny is the old idea of biological determinism: people are the way they are because they are born that way. This article reviews several themes conveyed by media gene talk:

(a) Genetic technologies provide the means to predict health and disease.

(b) Genes determine behavior and personality traits.

(c) Genetic research holds the promises of therapeutic solutions.

(d) But there are also potentials for abuse.

Following a brief description of these media themes, this article offers some reasons for the appeal of genetic stories. It suggests how the media coverage of genetics conforms to popular stereotypes and prevailing ideologies, and reflects the rhetorical skills of scientists who are eager to promote their work in the public arena. The article concludes by suggesting some of the problems with the media message that health, behavior, personhood—indeed, life itself—can be defined in genetic terms.

1. Media Themes

1.1 Predicting Genetic Destiny

Genetics is a science of prediction. Scientists emphasize the predictive possibilities of genetics as they seek to convey the importance of their science as a way to identify predispositions. Using media-appealing terms, geneticists call the gene a 'Delphic Oracle,' a 'time machine,' a 'trip into the future,' a 'medical crystal ball.' James Watson, the first director of the Human Genome Project, has announced in frequent media interviews that 'our fate is in our genes.'

Such metaphors convey several messages: a definition of the gene as the basis of human destiny, an expectation that genetic research will enable the prediction of future disease, and a promise of therapeutic solutions. These messages are disseminated by a receptive press.

The media, for example, cheered the discovery in the mid-1990s of the *BRCA1* and *BRCA2* genes that helped to explain many cases of familial breast cancer particularly among Jewish women of Ashkenazi descent. News headlines announced that scientists had found the gene causing breast cancer, some referring to the 'breast cancer gene.' They suggested that heredity is the major cause of breast cancer and that those found to have the gene were certain to get the disease. In fact, genetic factors contribute to only about 5 percent of all breast cancer cases, and even among those with a *BRCA* gene, modifying factors, including diet and genetic variables, may limit the predictive power of the genetic diagnosis. But widespread publicity, according to former Director of the National Institutes of Health, Bernadine Healy, created an ambience of 'medical bookmaking and fortune telling' that was devastating for women who believe they are at risk (Healy 1997).

Joseph Alper describes the media coverage of genetics as OGOD (One Gene One Disease) reporting (Alper 1996). When journalists report on the discovery of disease genes, they tend to minimize the complex, multicausal, and poorly understood sources of the genetic mutations that lead to disease. They are responding to the public desire for simple causal explanations. But Steven Rose (1995) describes genetic predictions as faulty reductionism based on 'reification, arbitrary agglomeration, improper quantification, misplaced causality, and dichotomous partition between genetic and environmental causes.' Similarly, Philip Kitcher (1997) describes how journalists tend to leap from an association of genes with particular conditions or traits to the attribution of genetic causation. Correlations turn into causes in the press.

Gene discoveries are newsworthy, but qualifications and disconfirmations are ignored. Studying the media reports about the causes of mental illness, Peter Conrad (1999) finds only limited coverage of inconsistent and contradictory information. The media message is rather that genes are destiny and the gene for mental illnesses will eventually be found. Science, however, is provisional and scientists routinely disconfirm their claims. These disconfirmations, like failed experiments, are important to the methods of science, but they are not regarded as news.

1.2 Genes Determine Behavior

Encouraged by the enthusiastic media response to research in medical genetics, behavioral psychologists in the 1990s began to extend the notion of genetic determinism to the inheritance not just of diseases, but of complex behavioral conditions and personality traits.

Such complex attributes as shyness, obesity, directional ability, aggressive personality, caring tendencies, exhibitionism, homosexuality, dyslexia, job success, arson, traditionalism, preferred styles of dressing, tendencies to tease, political leanings, re-

ligiosity, criminality, intelligence social potency, and zest for life have all appeared in media accounts as if they are simple Mendelian disorders, directly inherited like brown hair or blue eyes. Behavioral characteristics, once viewed in psychological or sociological terms, are now framed through genetic lenses (Alper and Beckwith 1993).

Seeking provocative and controversial copy, journalists have focused attention on claims attributing antisocial behavior to biological predisposition (Conrad 1997). There is violence, homelessness, alcoholism because people have genes that predispose them to aggression, depression, or alcoholism. Journalists refer to 'bad seeds,' criminal genes,' and 'alcohol genes.' To a *New York Times* writer, 'evil is embedded in the coils of chromosomes that our parents pass to us at conception' (Franklin 1989).

Journalists are also attracted to controversial theories about genetics and intelligence. In the 1970s, the dispute over Arthur Jensen's claims about the relationship between race and IQ attracted considerable media attention. Two decades later, the similar claims by Richard Herrnstein and Charles Murray in *The Bell Curve* again attracted an attentive press. Media interpretations of these studies, however, were more guarded than most reports on genetics.

Snyderman and Rothman (1988) saw the media as strongly biased against genetic explanations of intelligence, favoring arguments based on nurture or nature. Yet the coverage of the genetics and IQ debates has been extensive and the repeated articles about books like *The Bell Curve* disseminated their ideas very widely and ultimately served to amplify their importance.

In the 1990s, the media disseminated the claims about the genetic basis of behavior that were generated by studies of identical twins reared apart. These were controversial studies that often appeared in the media well before they were published in peer-reviewed journals. Offering a simple explanation of complex behavior and a reinforcement of prevailing stereotypes, this research attributed an extraordinary range of behaviors to 'the genes.' *US News and World Report* (April 13, 1987) published an authoritative-looking table providing precise percentages of how much personality traits were determined by heredity rather than culture: extroversion 61 percent, conformity 60 percent, worry 55 percent, creativity 55 percent, optimism 54 percent, and so on.

In the same issue, *US News* published an article called 'How Genes Shape Personality,' claiming that 'solid evidence demonstrates that our very character is molded by heredity.' It suggested that the future of Baby M, the child in a controversial surrogacy dispute, may not rest on which family got her, but in her genes. In 1992, *Time* was still offering the time-worn explanation of sex differences: 'nature is more important than nurture' and it is just a matter of time until scientists will prove it (January 20, 1992).

Media references to genetic influences draw on research in molecular and behavioral genetics, but they minimize the complexities of genetic and environmental interactions. They offer a distorting lens by ignoring the distance between the molecular level of genetic systems and actual behavior (Wilkins 1993). Beyond the curious notion that every quality has a gene, these media often convey a fatalistic and almost resigned attitude, for genetic predispositions appear to be immutable (Alper and Beckwith 1993).

One quantitative evaluation of magazine articles on genetics from 1919 to 1995 found a decreasing emphasis on genetic determinism over time (Condit et al. 1998). However, the way genetics is reported in the media of the 1990s suggests a retreat from the myth of 'can do' optimism that once led to the notion that every child could be president. The old story about those who make it despite difficult social circumstances is no longer a mystery. Genetic predisposition implies genetic destiny.

1.3 Biological Solutions

Good news, magic bullets, solutions, and cures appeal to public hopes and expectations. Journalists are strikingly optimistic as they cover the promise of 'medical breakthroughs' and 'wonder therapies.' Genetics claimed a journalist is 'the medical story of the century which will dramatically cure cancer, heart disease, aging and much more' (Rosenfeld 1992).

The media have welcomed every new claim about gene therapy with extravagant headlines and promotional hype as reporters convey—often uncritically—a wild array of futuristic scenarios presented by enthusiastic scientists. In the future, said a geneticist to *Discover*, 'present methods of treating depression will seem as crude as former pneumonia treatments seem now' (Wingerson 1982). In the future, said geneticist French Anderson to a *Time* reporter, 'Physicians will simply treat patients by injecting a snippet of DNA and send them home cured' (*Time* January 17, 1994). News headlines have promised: 'Genetic research leaves Doctors hopeful for Cures,' 'New Hope for Victims of Disease.' In a story called 'the Age of Genes,' *US News* reported that 'advances bring closer the day when parents can endow children not only with health but also with genes for height, good balance, or lofty intelligence' (Brownlee and Silberner 1991).

The development of pharmaceutical products and the proliferation of clinical trials on new therapeutic procedures have encouraged genetic optimism. The first official gene therapy experiment in 1990—the injection of cells containing ADA genes in a child with an immune system malfunction—became a major news event, touted as the arrival of 'the long awaited era of genetic therapy' (Culver 1991). The isolation of

the colon cancer gene in 1993 prompted an enthusiastic scientist to tell a *New York Times* reporter of its implications: 'deaths are entirely preventable' (*New York Times* December 3, 1993). The media reports implied that gene therapy would allow doctors to 'do something' about devastating disease. But in fact, diagnostic capabilities have far outreached the possibilities of solutions.

The gap between the promise and the reality of gene therapy has led to increased ambivalence in media reports of genetics. While few questions are raised about the ultimate benefit of research or the credibility of scientific claims, reporters are expressing concerns about the growing availability of genetic information and its potential abuse.

1.4 The Potential for Abuse

Stories extol the benefits of research, but then decry the risk of gathering genetic information. We are told of the dawn of a new genetic era but then cautioned about an impending eugenic nightmare.

Ambivalence is especially apparent in the coverage of genetic manipulation. Jon Turney (1998) notes the continuous use of Frankenstein metaphors in the media. An illustration for a *New York Times* article (September 16, 1990) on gene therapy and the potential of genetic engineering featured a drawing imitative of the famous Edvard Munch painting, 'The Scream.' A figure stands, horrified, mouth ajar, eyes wide open, its hair a mass of coiled DNA. Media reports have often associated experiments in gene therapy with genetic engineering or 'tampering' with genes. 'Lurking behind every genetic dream come true is a possible Brave New World nightmare' says a *Time* reporter. 'To unlock the secrets hidden in the chromosomes is to open up the question of who should play God with man's genes.' An accompanying image portrayed scientists balancing on a tightrope of coiled DNA (Dewitt 1989).

An important feature of these ambivalent media reports on genetics is their concern about the social and economic context in which this field is developing. The ties between the science of genetics and its commercial applications invite cynicism. Reporters call attention to the nonscientific interests—the investments and profits in this intensely competitive field. Science journalists, many of whom were once scientists themselves, have maintained an image of science as a pure and unsullied profession, a neutral source of authority, and an objective judge of truth. They are skeptical of corporate driven science. These days, according to some disillusioned journalists, biology has 'lost its virginity.' Scientists working in genetics are 'greedy entrepreneurs,' 'gene merchants,' or 'molecular millionaires' driven by economic interests that threaten their objectivity and override concerns about abuse (Nelkin 1995). Political cartoons portray geneticists in less than flattering terms as bumbling, naive, and unaware of the social implications of their discoveries. News reports and editorials repeatedly call attention to troubling aspects of the growing links between science and industry, the conflicts of interest that are inevitable when profits and ethics collide (Krimsky 1991). While the media have retained their history of genetic optimism, they are also expressing concern that expanding commercial interests will override important social considerations.

2. The Media Appeal of Genetic Explanations?

The Human Genome Initiative has replaced the space program as a new frontier, the cutting edge of high technology exploration, and, just as in the heyday of NASA, journalists are receptive to the enthusiasm of their scientific sources.

The stories of the 'gene of the week' meet the immediate pragmatic needs of the press for sound bites, newspegs, and events that will attract readers. The notions of genetic determinism offer relatively simple causal explanations of complex behavior—a welcome change from messy and uncertain social and environmental speculations. Genetic explanations conform to prominent cultural biases, intersecting with important American social values. Locating social problems within the individual, they support the ideology of individual responsibility. They are consistent with faith in scientific and technological progress. The status of the gene as a deterministic agent, a blueprint for life, offers certainty, predictability, and control. By promoting genetic explanations of behavior, however, the media are using the imprimatur of science to support a particular world view. They do so with little attention to the substance of science and the limits of theories purporting simple explanation of complex human conditions.

The appeal of genetics to the media largely reflects the promotional efforts of scientists themselves, for journalists rely on scientific sources for information. Scientists seek media coverage as a form of public relations, regarding public visibility as a means to attract funds for their research and support for their priorities (Shinn and Whitley 1985). They have become skilled in rhetorical strategies designed to attract media attention. Using media-ready metaphors, they have described the genes as 'master molecules': we are but 'readouts' of our genes. They describe the body in deterministic terms—as a set of instructions, a blueprint, a map, or a program that is transmitted from one generation to another. They suggest that by deciphering the text, classifying the markers on the map, and reading the instructions, they will unlock the key to human ailments and human nature, revealing the secrets of human life. Geneticist Walter Gilbert, introduces his public lectures on gene sequencing by pulling a compact disk from his pocket and announc-

ing to his audience: 'this is you.' Scientists, he claims, will 'provide ultimate answers to the commandment, know thyself' (Gilbert 1990).

The biotechnology industry has further encouraged media hype about gene therapy. The industry has made a major financial commitment to gene therapy, expecting this will be the basis of future medicine. Over 60 percent of gene therapy studies are directly financed by industry in anticipation of a profitable market in the near future. Corporate advertisements announce 'a great leap in the treatment of disease,' and promise 'a healthy future one gene at a time.' Corporate and university press releases are inclined to turn tentative experimental findings into magic bullets. An advertisement appearing in sports magazines said it all: 'Bad Genetics? Use Opti-genetics—The first genetic optimizer.'

When they report on complex scientific issues, journalists do not create science news; they mainly amplify and disseminate the messages provided by scientists themselves. They are seldom critical, responding with untempered enthusiasm to promotional hype especially when packaged in accessible media-friendly terms. However, the surfacing of hereditarian beliefs, their media appeal must be examined critically, for such beliefs have practical implications for both individual choices and social policies.

3. *The Growing Power of the Gene*

Tracing the history of popular images of genetics, Van Dijck (1998) has documented the changing metaphors that correspond to various stages in the development of this scientific field from the new biology of the 1970s to the rise of the biotechnology business and the Human Genome Project in the 1990s. At each stage, she says, 'performers'—scientists, journalists, and public relations people shape the media discourse (Van Dijck 1998).

Nelkin (1995) found considerable media skepticism in the 1970s when journalists wrote about the 'runaway science of genetic engineering.' But by the 1980s, gene splicing became a 'race toward better human health.' Today, genetics is the 'technological frontier.' Journalists describe geneticists as pioneers, 'unlocking the basic laws of nature,' discovering the 'secrets of life,' solving the problems of devastating disease. Genomic researchers are 'riding the DNA trail.' Geneticists are 'relentless hunters of genes,' involved in a 'race' to find the markers for disease. Media accounts in the 1990s have been reverent—almost religious: a 1994 cover of *Time* depicts a figure on a pedestal, his arms extended in a Christlike pose, his torso inscribed with a double helix. The caption reads: 'Genetics—The Future is Now.' The image, of course, is the Ascension. Though simply a biological entity, the gene in the media has assumed cultural significance extending far beyond its scientific meaning.

The media are mainly promotional, yet they are increasingly conveying mixed messages about the costs and benefits of biotechnology. Journalists seem to welcome the notion of biological determinism—simple, seemingly certain, and easy to convey—yet many writers remind their readers of the history of eugenics and place current research in this threatening historical context. While journalists report with enthusiasm and wonder the promises of gene therapy, they also warn about potential abuses of genetic manipulation.

Why should we be concerned about media images? Common and recurring media images provide insight into beliefs that are deeply embedded in society—beliefs that shape the questions of concern to scientists and the public appropriation of scientific findings. The media portrayal of the powers of the genes can have a powerful influence. For science has long served as a cultural resource appropriated and interpreted to support social beliefs and to legitimate social policies (Nelkin 1999). The language of determinism not only reveals but also creates social attitudes. The sound bites and headlines (gay genes, obesity genes, aggressive genes) attract readers, but they also endorse certain modes of thought and explanations.

Media coverage of science and technology provides a useful window on public attitudes, but also has an important influence on public perceptions. The way people perceive research in human genetics—the way they interpret their costs and benefits—may be influenced less by the details of scientific evidence than by the repeated messages conveyed in the popular press. These media messages help to create the beliefs and assumptions that underlie personal decisions, social policies, and institutional practices.

See also: Science and the Media

Bibliography

Alper J S 1996 Genetic complexity in single gene diseases. *British Medical Journal* **312**: 196–7

Alper J S, Beckwith J 1993 Genetic fatalism and social policy. *Yale Journal of Biology and Medicine* **66**: 511–24

Brownlee S, Silberner J 1991 The age of genes. *US News and World Report* **November 4**: 64

Condit C M, Ofulue N, Sheedy K M 1998 Determinism and mass media portrayals of genetics. *American Journal of Human Genetics* **62**: 979–84

Conrad P 1997 Public eyes and private genes: historical frames, news constructions and social problems. *Social Problems* **44**: 139–54

Conrad P 1999 Media images, genetics and culture. In: Hoffmaster B (ed.) *Bio-Ethics in Context*. Cambridge University Press, Cambridge, UK

Culver K W 1991 Splice of life. *The Sciences* **33**(1): 18–24

Dewitt P E 1989 The perils of treading on heredity. *Time* **March 20**: 70

Franklin D 1989 What a child is given. *New York Times Magazine* **September 3**: 36

Gilbert W 1990 *Current State of the HGI*. Dibner Institute, Cambridge, MA
Healy B 1997 BRCA genes—bookmaking, fortunetelling, and medical care. *New England Journal of Medicine* **336**: 1448–9
Kitcher P 1997 *The Lives to Come*. Simon & Schuster, New York
Krimsky S 1991 *Biotechnics and Society*. Praeger, New York
Lewenstein B U, Allaman T, Parthasarathy S 1998 *Historical Survey of Media Coverage of Biotechnology in the United States*. Presented to the Association for Education in Journalism and Mass Communication, Baltimore, MD
Nelkin D 1995 *Selling Science: How the Press Covers Science and Technology*, rev. edn. Freeman, New York
Nelkin D 1999 Behavioral genetics and dismantling the welfare state. In: Rothstein M (ed.) *Biology and Culture*. Johns Hopkins University Press, Baltimore, MD
Nelkin D, Lindee M S 1995 *The DNA Mystique: The Gene as a Cultural Icon*. Freeman, New York
Rose S 1995 The rise of neurogenetic determinism. *Nature* **373**: 380–2
Rosenfeld A 1992 The medical story of the century. *Longevity* **May**: 42–53
Shinn T, Whitley R (eds.) 1985 *Expository Science*. Reidel, Dordrecht, The Netherlands
Snyderman M, Rothman S 1988 *The IQ Controversy*. Transaction Books, New Brunswick, NJ
Turney J 1998 *Frankenstein's Footsteps: Science, Genetics, and Popular Culture*. Yale University Press, New Haven, CT
Van Dijck J 1998 *Imagenation: Popular Images of Genetics*. MacMillan, London
Wilkins A S 1993 Jurassic Park and the 'Gay Gene': the new genetics seen through the distorting lens of the media. *The FASEB Journal* **7**(13): 1203–4
Wingerson L 1982 Searching for depression genes. *Discover* **February**: 60–4

D. Nelkin

Genetics in Anthropology

Anthropologists of biological orientation began to undertake the study of genetics as soon as it became apparent about 1900 that biological inheritance was indeed particulate, and that the supposed essence of inheritance, genotype, could be inferred from phenotype. Within 'the science of man' theories were soon developed which applied these insights to the solution of problems that had long been central to anthropology, many of which were deeply entangled with political and religious controversies concerning 'race' and evolution. These problems especially included: (a) the extent and significance of human variation, (b) the phylogenetic relationships among humans and other primate species, (c) the migration and adaptation of human populations to different continents and geographical regions since the earliest times, and (d) the relationships between biological variation and other kinds of human variation, especially in language and culture. Despite some episodes of resistance to this new source of information, by the end of the twentieth century it was clear that genetics had become central to the science of anthropology, not only for helping to illuminate these problems, but for discovering new problems and creating new fields of study within the discipline (Crawford 2000).

1. Human Variation

When laboratory techniques for studying Mendelian traits first became available, international anthropology had just emerged from a serious controversy over racial typology during which Franz Boas, recognized by many as the founding father of American anthropology, had repudiated to the satisfaction of most anthropologists the racial theories of Gobineau and his intellectual heirs. Originally trained as a physicist, Boas presented data and arguments between 1910 and 1913 showing that 'race,' 'language,' and 'culture' were three independent phenomena which, he asserted, should all be studied by anthropologists, but using different kinds of theories and techniques. On the heels of this observation, the Hirzfelds, building on the work of Landsteiner and others, showed in 1919 that while the ABO blood system was universal among human populations, the phenotypes and hence the alleles, the alternative forms of a gene, were in different proportions, thereby supporting Boas' contention that the differences among 'racial' populations were not qualitative but quantitative (Weiss 1998). The ability to taste PTC (phenylthiocarbamide), a simple Mendelian trait, also seemed to occur in many different human populations, likewise in different frequencies. But other apparently Mendelian traits examined early in the century—eye color, skin color, and ear lobe shape—turned out to embody genetic relationships with a much more complex genetic architecture (McKusick's current list of over 30,000 Mendelian traits can be found at World Wide Web URL: http://www3.ncbi.nlm.nih.gov/Omim/).

Data on human phenotypes assembled for military purposes in World War I continued to encourage anthropological geneticists in the decades following the war. In addition to complex and composite anthropometric traits such as height and weight, the military was especially interested in those factors that affected the compatibility of blood and blood components used in transfusion. The collection of such data was continued through World War II, when most European and American soldiers and sailors were typed for ABO and Rh factors during initial processing, and this body of data became generally available to scientists after 1945 (Harrison et al. 1964).

For anthropologists, the major theoretical model which emerged to describe human genetic variation in this period was the notion of cline, the gradual geographic change in the frequency of an allele, which also served to undermine further the idea of racial typology. As regularities began to emerge, especially

in the European data, it was found that the alleles of a gene, and also phenotypic traits such as red hair, seemed to vary across the landscape in a regular manner. These were mapped using arrows to show the direction of increasing frequency of one of the component alleles of a gene, or by isographs which identified geographical areas where allele frequencies were within a particular range. Traits mapped in this manner included blood system alleles, phenotypic traits, and also diseases which were suspected of having a simple genetic basis, such as albinism, color blindness, and some types of cancer (Harrison et al. 1964, pp. 234–76).

Some patterns which emerged from the analysis of clines conformed to more general biological principles pertaining to other animals. It was found, for example, that Gloger's, Bergmann's and Allen's Rules for variation within a mammalian species (pertaining to pigmentation, mass, and limb length, respectively) were also valid for the human species, reflecting adaptation to the different latitudes, climates, and environments inhabited by local human populations. Since these traits were so clearly adaptive, it was suspected that other patterned traits, with definite clines, likewise reflected genetic adaptation to local conditions.

2. *Adaptation*

The most forceful new example of human genetic adaptation was presented by Livingstone in 1958 with his study of sickle cell disease, especially as it existed in Africa. In addition to presenting convincing evidence of several types to prove that sickled blood cells were a response to the presence of malaria, Livingstone's work illustrated several other lessons in anthropological genetics. First of all, it was the heterozygote of hemoglobin S (HbS) which had the adaptive advantage, not the HbS homozygote, who often sickened and died. The second lesson was that an apparent 'disease,' sickle cell anemia, was on closer inspection actually a successful adaptive mechanism, a proposition which is still better understood by anthropologists and population geneticists than by their colleagues in clinical medicine (Weiss 1993).

Another highly visible connection between genetic variability, culture, and environment concerns the ability of human adults to digest milk and milk products from dairy animals, an ability that seems to be under genetic control. When mapped geographically, the alleles responsible for adult lactose tolerance in human populations seem to be correlated, in the Old World, with the locations of pastoral peoples. Significantly, populations where lactose tolerance is high include pastoral societies all the way from Scandinavia to southern Africa, producing east–west clines which cut across the north–south clines describing the distributions of all sorts of other phenotypical characteristics, including those related to latitude and climate (Durham 1991).

There are indications that the frequencies of many other diseases, such as cancer and diabetes, correlate with blood types and also with alleles involved in the digestion of food. People of A phenotype, for example (in the ABO system), may carry an increased risk of death from cancers of the esophagus, pancreas, and the stomach, but whether these diseases are in any sense 'caused' by type A blood, as the sickling trait can cause anemia, or whether both disease and blood type are derived from an independently variable genetic or environmental condition, is very problematic to clarify. Northern Scandinavia, for example, has a high frequency of type A individuals and a high incidence of stomach cancer. But their diet is different from other Europeans, and their climatic environment is different as well, so it is difficult to separate the variables which might cause stomach cancer (Molnar 1998, pp. 100–7).

Perhaps the most useful way of looking at the various genetic and environmental factors correlating with disease frequencies is to say that each 'contributes' to the likelihood that an individual will exhibit the disease (Terwilliger and Göring 2000). Many anthropologists argue that from the standpoint of adaptation, no gene of itself is a 'bad' gene. Like HbS, any allele manifested in the phenotype might be beneficial in certain circumstances. Among anthropologists, it is increasingly recognized that the relationship of gene to environment, like the relation of gene to gene, is very complex. Also, the distinction between 'genetic disease' and 'infectious disease' begins to break down when we consider that smallpox, for example, takes a more virulent course in people of blood type A or AB , as opposed to O or B (Stein and Rowe 2000, pp. 427–30).

The study of adaptation to cold climate and high altitude has provided an opportunity to differentiate between the effects of genetic differences among human populations and phenotypical differences engendered by growth and development—more simply, the difference between nature and nurture. Most recently, Beall has shown that Tibetan and Peruvian populations have a different genetic basis for their respective adaptations to thin air (hypoxia), but many other aspects of human adaptation to climatic extremes remain to be examined (Jones et al. 1992, pp. 46–51). These studies require the comparison of highland to lowland Peruvian Indians, and the examination of such characteristics as body type, especially lung capacity, and oxygen saturation rates among migrants and colonists in both locations.

3. *Population Genetics*

With increased funding for science in the United States after World War II, and increased access to field sites,

anthropologists began to investigate isolated, small-scale human populations with renewed vigor, often including a biological anthropologist in the research team. The most visible of these joint enterprises was probably the comprehensive study of the Yanomamö, who live in the upper drainage of the Orinoco River in Venezuela and Brazil. In the 1960s, James Neel of the University of Michigan organized a comprehensive multidisciplinary series of field expeditions which investigated the environment, ecology, biology, culture, and social organization of these then little-known people. After a series of books and articles was published over the next two decades, the Yanomamö became not the least known but one of the best known tribal societies in the world.

For the genetics part of the research, Neel and Ward examined six blood systems relatively well known at the time (MNS, Rh, Kidd, Duffy, Diego, and haptoglobin) among seven villages of Yanomamö, seven villages of nearby Makiritare, and 12 other tribes of Central and South America (Neel and Ward 1970). After comparing the different groups, the authors concluded that human evolution might have been proceeding at a pace '100 times more rapid' than previously thought, if these small-scale societies could be taken as analogues of earlier Pleistocene or Paleolithic human peoples.

In the same period, Friedlaender and Harpending were proceeding with similar investigations in Oceania and Africa, respectively. Working in the Solomon Islands from 1966 to 1973, sometimes alone and sometimes as part of a multidisciplinary team, Friedlaender simultaneously undertook studies of demography, anthropometry, genetics, and social structure, seeking to integrate his results into some general conclusions that cut across these traditional fields of research. Like Neel and Ward, he emphasized the genetics of blood systems, in this case ABO, MNS, Rh, Kell, and Duffy. In these years, blood system alleles were not sequenced directly, but determined from immunological reactions to blood proteins, usually through the intermediary use of rabbit antibodies (Jones et al. 1992, p. 304). From this kind of data, Friedlaender calculated phenotype frequency and then allele frequency by techniques described in his major ethnographic work, *Patterns of Human Variation* (Friedlaender 1975). His most general conclusion, combining results from anthropometry, dermatoglyphics, genetics, and demography, was simply that Bougainville communities tended to exhibit high biological variability within the group, but low variability among the groups, and that some traits, such as skin color, eye color, and hair form, were 'monotonous' on the island, while dermatoglyphic characteristics and some blood system alleles were discontinuous from village to village.

Those expecting that these early studies would provide fundamental insights into the mechanisms of human evolution were soon disappointed. Summarizing a widely shared consensus of opinion in 1974, Harpending concluded (1974, p. 229):

> … studies of the genetic structure of small populations have made particular and incidental contributions to formal genetics, to regional history and prehistory, to epidemiology, and to several other fields to which they are peripheral, but … they have not advanced our understanding of human evolution in a global sense. The sample sizes available have been too small to allow reliable inferences about natural selection; the extensive occurrence of what is presumably local random genetic drift has little or no consequence for evolution over long time periods over large areas; and the presumed selective agents in the various environments of these peoples differ greatly so that few of the generalizations which have been put forward hold for many groups.

4. *Evolution of Primates*

Although anthropologists, even before the publication of Darwin's *Descent of Man* had recognized that humans and apes were closely related, the significance of the differences between them had been both exaggerated and misunderstood. Based on the fossil record, the split of humans from the ape line had been put far back in time, about nine million years, and it was assumed that all the ape species were closely related to each other and quite distinct from humans. Therefore the allegation of Sarich and Wilson, published in 1967 and based primarily on immunological data, that humans and apes had diverged as recently as five million years ago (mya), created a serious confrontation between 'the bone people' (paleontologists) and ' the lab people,' human geneticists such as Sarich and Wilson. The former accused the latter of inaccuracies in their assumptions about the speed of 'evolution,' in this case the evolution of blood albumin, while the latter accused the former of manipulating primate taxa to preserve the supposed profound differences between ape and human. In fact the fossil Ramapithecus, with an antiquity of about 14 mya, soon fell out of the human lineage on osteological grounds, so that even the fossil evidence seemed to point to a more recent divergence between human and apes.

Beginning in the 1970s, Morris Goodman and his associates have become prominent in the reconciliation of fossil and biological evidence which, taken together, provide the best description of the evolution of apes and humans. In 1974 Goodman proposed that humans, chimpanzees, and gorillas should join humans as hominines, a subfamily of the family Hominidae, perhaps leaving the more distantly related gibbons and orangutans as separate families or sub families. Goodman's conclusions, like those of Sarich and Wilson, were based on molecular data, in this case amino acid sequces from blood proteins.

Goodman's current synthesis (1999) includes data from many kinds of direct sequencing of DNA from

humans and other primates, especially the β-globin gene cluster, and from the analysis of morphological features, especially certain diagnostic skeletal features on living and fossil primates. His current picture of relationships among living humans and apes shows a split between the Tribe Hominini and the lesser apes about 18 mya, between orangutans and the Subtribe Hominina about 14 mya, and among gorillas, humans, and chimpanzees within the last seven million years. The timing and significance of the chimpanzee–bonobo separation has not yet been firmly established, since they have only recently been recognized as distinct species, and basic comparative genetic studies of both groups are now underway.

5. *Methods*

Although thousands of laboratory procedures have been developed over the last half-century, especially by medical researchers, to elucidate the structure of the human genome and show how it reproduces itself, certain techniques and certain targets have been more interesting to anthropologists because of their particular research agenda. Ideally, anthropologists interested in adaptation would like to know the locations of all human genes on the chromosomes, the number of alleles for each gene, and the locations of all loci involved in constructing a protein, but in practice researchers have had to settle for much less. Early in the twentieth century, the identity of alleles had to be teased out of complex immunological procedures in the laboratory. Later, proteins could be examined directly, but without much knowledge of the DNA structures, the precise sequence of nucleotides which had produced the protein. But with the development of slow though reliable DNA sequencing procedures in the 1970s, anthropologists became much more selective about which portions of the chromosomes they wanted to study, and for what purposes.

Surprisingly, with the inception of direct sequencing, it was discovered that perhaps 93–97 percent of the mass and length of human chromosomes consisted of apparently inactive regions variously called introns, satellites, tandem repeats, noncoding regions, or even 'junk.' Almost immediately these regions began to attract the attention of anthropologists even though they did not contain genes. It was the presumed inactivity of these regions which was interesting, since accumulated mutations on these regions were apparently selectively neutral, that is neither favored nor disfavored by natural selection. Therefore, it was reasoned, they might in a more accurate, clock-like manner reflect not only the relationships among primate species, but also the migrations and gene flow of ancient human populations. That is, satellite DNA was considered to constitute a 'passive' record of evolutionary events, rather than comprising active agents under selective pressure.

5.1 *Stains*

As soon as chromosomes could be observed directly through microscopes, attempts were made to stain them chemically so that bands would appear which might show differences in structure. Initially, the purpose was to discriminate among the different chromosomes of the human karyotype, or suite of chromosomes (22 autosomes, 1 X and 1 Y), so they could be counted and numbered. During the course of this research, it also became apparent that patterns of stains on a particular chromosome might vary from one individual to another within a population. In addition, when it was noted that certain patterns of staining were characteristic of particular species, yet another method became available for evaluating the phylogenetic relationships among apes and humans. At first the functional significance of the patterns was unknown, but this is changing rapidly as new techniques allow the identification of gene-rich regions identified by stains.

5.2 *Restriction Enzymes*

In 1973 it was discovered that certain enzymes, mostly recovered from bacteria, could sever a strand of DNA selectively at locations defined by a known base pair sequence. For example, the enzyme AluI recognizes the sequence AGCT and cleaves DNA at that location to produce fragments which are on average 300 base pairs in length in humans. The resulting distribution of segments of different length serves to define the structure of a chromosome, even without knowing the entire succession of base pair sequences on each segment. An appropriate enzyme, or several enzymes applied in series, can be selected to produce a small or a large number of segments, long or short, depending on the purposes of the research. By comparing the phenotypes of individuals contributing DNA samples and examining their pedigrees, 'linkages' could then be discovered between particular alleles included in the same segment.

Electrophoresis, which had previously been used to separate blood proteins obtained from serum, can be used to separate DNA segments according to size and net electric charge, with the segments sometimes enhanced or differentiated visually by using a large number of new stains developed in the last two decades of the twentieth century, some of which are fluorescent. In the 1980s, the PCR (polymerase chain reaction) 'revolution' occurred, which allowed researchers to clone or amplify small pieces of DNA indefinitely. PCR procedures make it possible to amplify a tiny bit of DNA clipped from an electrophoresis gel, or even a single sequence from a single cell, to obtain sufficient material for analysis (Hillis et al. 1996). All of these became useful tools for describing genetic variation in a population within a species, or to compare and contrast variations between and among species.

5.3 Chromosome Mapping

A major objective of human genetics research is to locate all genes on their respective chromosomes as well as to identify all the other functional regions of DNA, including some formerly dismissed as 'junk.' One technique for gene mapping involves the use of genetic markers, like the restriction enzyme sites described earlier, or other DNA sequence variants discovered by various means at hundreds of other chromosome locations. Because many of these are relevant to human disease, high levels of funding were made available to the Human Genome Project to identify the location of these variable sequences or 'polymorphisms.' This support has led to the development of very fast computer and molecular technology and machinery, and much of the resultant data appears in easily searched form on the worldwide web.

5.4 Mitochondrial DNA (mtDNA)

Most human genes reside in the nucleus of the cell and occur in two copies, one each inherited from the mother and the father. However, the body or cytoplasm of each cell also contains structures known as mitochondria, which are involved in the production of energy for the cell and contain their own DNA. This DNA is special in many ways. At fertilization, the mitochondria of the ovum are transmitted to the offspring, but the mitochondria of the sperm are not. Thus mtDNA is inherited in a strictly maternal manner, a biological phenomenon which parallels several important sociological phenomena of special interest to anthropologists, such as matrilineal clan membership.

Since mtDNA is haploid rather than diploid in form (a single rather than double strand), it does not recombine during replication, and over time, as sequence variations accumulate through mutation, it produces distinct, diverging, lineages of mtDNA sequences. This variation can be used to reconstruct the mtDNA structure of the common, apical ancestor (the coalescent sequence) from a sample of mtDNAs collected from individuals alive today. Because the mitochondria are so numerous in the cell, mtDNA can be recovered much more easily than DNA from the nucleus.

The mtDNA chromosome takes the form of a relatively small ring of 16,500 base pairs (16.5 kB), which was entirely sequenced by 1981. By 1987, 147 human individuals had contributed additional sequences and the structure of the converged mtDNA lineage was announced, which 'coalesced' on a woman who was promptly dubbed 'mitochondrial Eve,' alleged to have lived in Africa about 200,000 years ago, and whose 'modern human' progeny had emigrated from Africa and repopulated the world, which was at that time allegedly occupied entirely by earlier forms of human, such as Neanderthals and Homo erectus (Cann et al. 1987).

5.5 Y Chromosomes

There is one DNA molecule that is uniquely inherited in the paternal line, a kind of mirror image of mtDNA. When it was discovered that certain regions of the Y chromosome were very stable and apparently did not recombine during meiosis, it was hoped that the Y chromosome would become the analytic reciprocal of mtDNA. That is, Y chromosome lineages should parallel patrilineal social practices in the same way that mtDNA lineages paralleled matrilineal social practices. A dramatic confirmation of the utility of Y chromosome analysis has been provided by Thomas et al. (1998) who showed how Y chromosome lineages confirmed the patrilineal inheritance of the Jewish priesthood among widely dispersed Jewish populations. Seielstad and others have opened another promising line of research by comparing the geographical distributions of mtDNA and Y chromosome lineages in attempts to reconstruct the marriage practices of prehistoric peoples (see *Y-chromosomes and Evolution*; Stoneking 1998).

5.6 Satellites

Although the picture is constantly changing, researchers at this point recognize three general categories of tandemly repeated sequences, which are classified by length. They are satellites, the longest sequences, minisatellites, defined as sequences which are 500–40,000 base pairs in length, usually consisting of five or fewer repeated nucleotides, and microsatellites, which tend to be short and simple in motif, even consisting of repeats of a single base pair, or a triad or tetrad of nucleotides (for example, CACACA or CAGCAG) (Jackson et al. 1996, pp. 171–210).

These short, apparently neutral sequences are useful to anthropologists because their rate of mutational change is much higher than the single nucleotide mutation of DNA, so that minisatellites and microsatellites, like mtDNA, provide a finer control of historical relationships among individuals and the populations they represent. While the mutation rate for single nucleotides is about 10^{-7}, the rate for microsatellites is about 10^{-3} per generation.

6. Grand Synthesis

Before the entry of genetics into anthropology, the recent history of our species, over the last 6,000 years, was determined largely from written accounts or, in

the case of preliterate societies, from oral narratives. To reconstruct events farther back, in the late prehistoric period, archaeologists examined architecture and technology, with architecture becoming less important farther back in time and technology, especially stone tools, becoming more important, especially before the Neolithic period began about 10,000 years ago. In all periods, compared to cultural remains, biological evidence in the form of graves and fossil humans has been rare, and the interpretations have been controversial.

In some ways, the present evolutionary debate among anthropological geneticists can be seen as a continuation of an historic dispute in paleontology between 'lumpers' and 'splitters.' Lumpers have historically emphasized the general evolution of the whole species, while splitters have tended to group human fossils into named regional and temporal categories, which were considered to evolve somewhat separately and independent of other regional populations. Splitters tended to assign new names to new fossil finds, while lumpers were reluctant to inaugurate new taxa, emphasizing instead the ongoing wide diversity within the human species.

The nature of genetic evidence lends itself to the continuation of this kind of debate. To a considerable extent, all modern human populations share similar alleles, and can be distinguished in this respect only by their different frequencies. An array of allele frequencies from a population can in principle be analyzed as a combination of two or more previous arrays, identified by name and allegedly specific to some region. This is a splitter's approach. On the other hand, genetic 'lumpers' can maintain that the human species has always been diverse, and is no more heterogeneous now than it was hundreds of thousands of years ago.

The debate can also be characterized as between cladistic theorists who see local or regional human populations as largely maintaining their social and biological integrity, increasing in size, dividing into daughter groups, and then migrating to replace less successful populations. By contrast, ethnogenetic theorists postulate that gene flow by reciprocal marriages or matings among small adjacent populations might be as significant as long distance migration in changing regional allele frequencies, and that while language and culture might change rapidly within a region, biological characteristics change more slowly.

Luca Cavalli-Sforza and his colleagues have been prominent in this debate, and viewed by many as leaders of the cladistic school of thought. In the final decades of the twentieth century, they have produced several important theoretical books, followed by a huge compendium of maps, tables, and texts showing known global genetic variation (Cavalli-Sforza et al. 1994). In this school of thought, current human genetic variation is explained as the result of historic and prehistoric migrations of named ethnic groups from region to region and continent to continent, especially including the 'African Replacement' by Eve and her progeny.

If these works have not entirely convinced all of their academic readers, they at least have served as a convenient target for 'lumpers' and advocates of ethnogenetic theory. Prominent among these critics has been Milford Wolpoff, who argues, primarily on the basis of osteological evidence, that global human evolution has been mostly *in situ*, since modern populations, in the Far East for example, tend to reflect unique or unusual osteological features exhibited by pre-Eve residents of the region. The other main line of criticism has been from linguists and social anthropologists, who argue that Cavalli-Sforza and his co-workers have got the languages and ethnography wrong, and that their hypothesized migrations and other social and linguistic processes cannot be confirmed independently (MacEachern 2000, Terrell 2000). Such critics argue that human groups, especially the small-scale hunting-collecting societies which existed in Paleolithic times, were more fluid and less bounded than required by the migration hypotheses.

Accumulating genetic evidence seems to support the migration hypotheses. Concerning Eve, other genetic lineages besides mtDNA also seem to coalesce at about 120–200,000 years ago, supporting the Out of Africa hypothesis. However, rates of evolution of sequences on the autosomal chromosomes are difficult to judge, and the particular histories of gene families and other segments of chromosomes, especially their evolutionary odysseys of travel among the chromosomes over the last millions of years, are just beginning to unfold.

The analysis of ancient DNA from human fossils has the potential for resolving many problems of human prehistory and evolution, provided that appropriate and usable DNA can be retrieved from scarce and damaged human bones, of great antiquity. Already the examination of mtDNA sequences from a few Neanderthal specimens has shown deep rather than shallow genealogical relationships to modern human sequences, as if indicating that modern Europeans replaced them, rather than being descended from them. A human of lesser age, the so-called and well-publicized Tyrolean Ice Man, has an estimated antiquity of about 5,000 years, according to carbon-14 dating, and could have been ancestral to modern Central Europeans (Herrmann and Hummel 1994, Kolman and Tuross 2000).

The most formidable but most satisfying task facing anthropological geneticists in the near future is to continue the confirmation and criticism of the fit, or lack thereof, among genetic, linguistic, and archaeological evidence for human migrations. No technique has yet been found for differentiating allele distributions resulting from gene flow from those resulting from a sudden migration after several generations

have passed. Intermarriage and recombination tends to blur the distinction, so that anthropologists for the next several decades will no doubt continue to debate the accumulating evidence for global, continental, and regional migration and gene flow.

7. Problems

The greatest barrier to using variation in DNA sequences—polymorphisms—to trace the antiquity of relationships among human individuals and populations is the unreliability of the molecular clock. While mtDNA sequences, DNA lineages constituting gene families, and specific microsatellite sequences can be arranged into evolutionary trees, no one knows for sure the time between each generation or branch of the tree. It is clear that the frequencies of base pair deletions and other forms of mutation are not the same in different chromosomal regions, and similarly the amount and dynamics of variation seem to vary greatly among regions on the same chromosome, and among the autosomes, the Y chromosome, and mtDNA. There also seem to be 'hot spots' where mutations or recombinations are more frequent, near telomeres and centromeres, for example. An additional concern in calibrating the molecular clock is that the 'proofreading,' repair, and editing functions of genetic reproduction seem to be uneven for different genes and loci (especially, mtDNA seems to have weaker repair systems), so that some sequences are actually much older than they appear, because they have been systematically repaired (Kumar and Hedges 1998). What all of this means is that each gene has its own history, and only by aggregating information from many such histories do some aspects of the picture of human history begin to emerge.

Bibliography

Cann R L, Stoneking M, Wilson A C 1987 Mitochondrial DNA and human evolution. *Nature* **325**: 31–6

Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton NJ

Crawford M 2000 Anthropological genetics in the 21st century: Introduction. *Human Biology* **72**(1): 3

Durham W H 1991 *Coevolution*. Stanford University Press, Stanford, CA

Friedlaender J S 1975 *Patterns of Human Variation: The Demography, Genetics, and Phenetics of Bougainville Islanders*. Harvard University Press, Cambridge, MA

Goodman M 1999 Molecular evolution '99: The genomic record of humankind's evolutionary roots. *American Journal of Human Genetics* **64**: 31–9

Harpending H 1974 Genetic structure of small populations. *Annual Review of Anthropology* **3**: 229–43

Harrison G A, Weiner J S, Tanner J M, Barnicot N A 1964 *Human Biology*. Oxford University Press, Oxford, UK

Herrmann B, Hummel S 1994 *Ancient DNA*. Springer, New York

Hillis D M, Moritz C, Mable B K 1996 *Molecular Systematics*. Sinauer Associates, Sunderland, MA

Jackson M S, Strachan T, Dover G (eds.) 1996 *Human Genome Evolution*. BIOS, Oxford, UK

Jones S, Martin R, Pilbeam D 1992 *The Cambridge Encyclopedia of Human Evolution*. Cambridge University Press, Cambridge, UK

Jorde L B, Bamshad M, Rogers A R 1998 Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *BioEssays* **20**: 126–36

Kolman C J, Tuross N 2000 Ancient DNA analysis of human populations. *American Journal of Physical Anthropology* **111**: 5–11

Kumar S, Hedges S B 1998 A molecular timescale for vertebrate evolution. *Nature* **392**: 917–20

Long J C 1993 Human molecular phylogenetics. *Annual Review of Anthropology* **22**: 251–72

MacEachern S 2000 Genes, tribes, and African history. *Current Anthropology* **41**(3): 357–84

Molnar S 1998 *Human Variation*. Prentice–Hall, Saddle River, NJ

Neel J V, Ward R H 1970 Village and tribal genetic distances among American Indians, and the possible implications for human evolution. *Proceedings of the National Academy of Sciences* **65**(2): 323–30

Sarich V M, Wilson A C 1967 Immunological time scale for Hominid evolution. *Science* **158**: 1200–3

Stein P L, Rowe B M 2000 *Physical Anthropology*. McGraw-Hill, New York

Stoneking M 1998 Women on the move. *Nature Genetics* **20**: 219–20

Terrell J E (ed.) 2000 *Archaeology, Language, and History*. Bergin and Garvey, Westport, CT

Terwilliger J D, Göring H H 2000 Gene mapping in the 20th and 21st centuries: Statistical methods, data analysis, and experimental design. *Human Biology* **72**(1): 63–132

Thomas M G, Skorecki K, Ben-Ami H, Parfitt T, Biadman N, Goldstein D 1998 Origins of Old Testament priests. *Nature* **394**: 138–9

Weiss K M 1993 *Genetic Variation and Human Disease*. Cambridge University Press, Cambridge, UK

Weiss K M 1998 Coming to terms with human variation. *Annual Review of Anthropology* **27**: 273–300

J. H. Moore

Genetics of Complex Traits Through the Life Cycle

The purpose of this article is to present a schema for conceptualizing the sources of individual differences in lifespan developmental processes. The two broad domains of genetic and environmental factors are presented as co-acting in their influence on complex traits, with the relative influence differing from trait to trait. Furthermore, intricate interactions are possible among genes and between genes and environments. The system is dynamic, with the possibility that the relative contributions of genetic and environmental sources can change over time. Newly available

methods assure finer-grained descriptions of the interrelationships of genes and environment in the 'architecture' of complex systems in the near future.

1. Mendelian Genetics

Prior to the research of Mendel, the general conceptualization of the hereditary material was of an amorphous, miscible substance, in which the contributions of the parents were blended and influences from the environment altered the mix. Mendel's ingenious contribution was to postulate the existence of discrete elements (later termed 'genes'). Further postulations, in abridged form, were that an organism possesses many such genes, and each occurs as a pair, with an organism receiving one element from each parent. There are different forms of the genes (later to be named 'alleles'). Just as an individual received one allele for each gene from each parent, it transmits only one of each pair to its offspring. The alleles do not blend and are not modified during the association with their partner alleles or with the alleles of other genes, but are transmitted from the organism in the same functional condition in which they were contributed to the organism. Likewise, environmental factors do not alter the nature of the elements.

A critical distinction in Mendelian genetics is made between the phenotype, the measured attribute, and the relevant genes, or genotype, and the crux of the theory was the set of rules determining the relationships that could obtain between genotype and phenotype.

Mendelian genetics deals with categories. In the ideal case, all individuals are assignable unequivocally to one of two or three categories. A particular reason that the genotype/phenotype distinction is important is that the relationship between the two is not always additive. Knowing the phenotype is not always unambiguously informative about genotypes. For example, in one classical Mendelian situation, those individuals possessing two copies of a particular allele of the gene belong to category A, those possessing two copies of another allele of the same gene belong to category B, and those possessing one copy of each also belong to category B. That is to say, one copy of the allele 'for' category B is as effective as two. This relationship is described as dominance; the allele for which one copy is as effective as two is the dominant allele, and the other is termed the recessive allele. The two genotypes with identical alleles are called homozygotes, and the genotype with one each is called a heterozygote. Other relationships are possible: In some situations, the relationship between allelic 'dose' and phenotype is strictly additive, and intermediate degrees of dominance are found for other phenotypes.

Even when dominance is present, there is, of course, an additive component to the relationship. Thus, the phenotypic variability can be regarded as due to an additive component and to dominance deviations. Another major reason that the distinction between genotype and phenotype is critical is that genes may not be the sole determinants of the phenotypic outcome. The basic principle of multiple influences underlying quantitative genetics is illustrated in Fig. 1. Phenotypic values are shown for a number of individuals of each genotype for a single gene with a partially dominant functional relationship to the phenotype. The within-genotype variability is assumed to be due to environmental circumstances that have uniquely affected each individual. A projection of these phenotypic values onto the right Y-axis reveals a distribution, the variability of which is attributable both to genetic differences (with some of this genetic influence due to the additive effect of allelic 'substitution' (represented by the dashed regression line), and some due to dominance (the deviations of the X's marking the means of each group from the regression line) and also with a portion of the variance due to environmental differences among the individuals. The consequence when more than one gene can affect the phenotype will be discussed below.

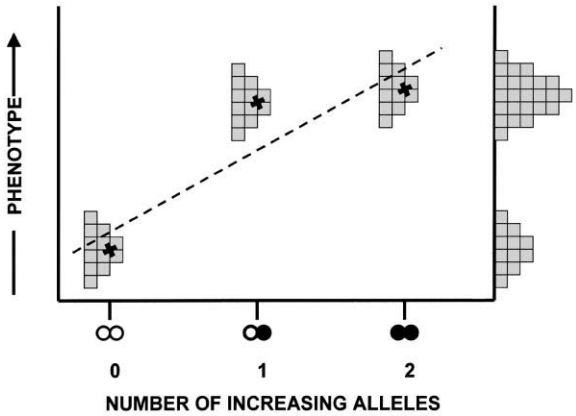

Figure 1
Depiction of a Mendelian partial dominant situation with environmental influences responsible for within-genotype phenotypic variation. Phenotypic values of all genotypes are projected onto a single distribution on the right Y axis

The source of genotypic diversity is, of course, the existence of different alleles. These alternative forms of the gene arise through mutational events, and their ultimate fate in a population depends upon their influence on the reproductive 'fitness' of their possessors, among other matters (see Falconer and Mackay 1996, Hartl and Clark 1997). In any population, the relative frequencies of the alternative alleles (there may be many more than two, but this discussion will be restricted to the simple case of two) will ordinarily be far from equal. Thus, the different genotypes will also not be equally frequent. If p represents the

frequency of the most abundant allele (say A1) and q represents the least frequent one (A2), and if mating is at random, then the relative frequencies of the genotypes *A1A1, A1A2*, and *A2A2* will be p^2, 2pq, and q^2, respectively. Different values of p and q will thus obviously lead to different genotype frequencies, and thus to different means and to different variances.

2. *Polygenic Inheritance*

For many phenotypes, it is not possible to assign individuals to categories; instead, they are arrayed along a continuous distribution. Although for a time there was a conjecture that different genetic systems were engaged in the categorical and continuous, a model (Fisher 1918) that postulated the existence of many loci, each performing according to Mendelian rules, but each having only a small influence on the same phenotype, resolved the issue. The collective effect of these multiple genes (polygenes) yields statistical expectations of resemblance of relatives of different degrees of biological relationship. For example, our understanding of the mechanism of gene transmission informs us that a parent shares half of his or her alleles with a child; grandparents share one quarter of their alleles with their grandchildren; siblings share half of their alleles on average; fraternal twins are just like siblings in this regard, and identical twins share all of their alleles. Comparison of the degree of similarity of phenotypes to the fraction of shared alleles permits the estimation of the proportions of phenotypic variance that can be attributed to genetic sources and to environmental sources. Each of these broad categories can be further divided. The overall genetic component can be described in terms of an additive component and a non-additive one. The non-additive component itself can be subdivided into effects arising from dominance, which can be regarded as the interaction of alleles at a locus, and from epistasis, which is the interaction of genotypes at different loci. This latter effect will be discussed in more detail later.

In the quantitative genetic model, environment is usually defined by exclusion—it is all of the influences that are non-genetic; i.e., not arising in a causal sense from the coded material of the DNA. Thus the scope and range of environment in this conceptualization is enormous, including in its compass factors from cellular biochemistry to peer group dynamics. This environmental contribution to phenotypic variance can also be partitioned in a number of ways, depending on the purpose of the research and on the data available. Long-term environmental effects may be distinguished from short term ones, for example, for developmental studies. One common distinction is that between environmental effects that are shared by family members and non-shared environmental influences that are experienced uniquely by individuals.

A simple, broad-stroke conceptual model represents the phenotype as the additive outcome of genetic and of environmental influences. Algebraically, this can be stated as

$$P = G + E$$

In a system such as this, the variances can be described as

$$V_P = V_G + V_E + 2Cov_{GE}$$

As noted previously, these terms are further divisible. Thus, as noted, V_G includes both an additive component (V_A) and a nonadditive (V_{NA}) component, and V_E is similarly subdivisible. These components yield an estimate of the relative influence of the two general domains. (There are also formulations that include a component of stochastic or of nonlinear dynamic influences [see Molenaar et al. 1993, Finch and Kirkwood 2000].) Measurement error is usually found in the environmental term, and as will be discussed below, there are also interaction terms that can be of great importance. The most commonly encountered index is that of heritability, which is the proportion of total phenotypic variance that is due to genetic differences among individuals. In the narrow sense, heritability is V_A/V_P, and in the broad sense, it is V_G/V_P. These two values are differentially useful for different descriptive and predictive purposes (see, for example, Falconer and Mackay 1996). The same information can be captured by an index of environmentality ($1-h^2$), of course, but this term is infrequently encountered. It is very important to recognize that heritability and environmentality are descriptive statistics describing the state of affairs of a specific population under specific environmental circumstances. Different gene pools, or different environmental forces, might yield quite different estimates. Note also that high heritability does not mean that there is no conceivable environmental influence that can affect the phenotype. It means that in the population that was studied, there was either no sufficiently powerful environmental influence, or that if one existed, it impinged on so few of the population that it had little effect on the overall variability. Thus high heritability for a condition such as Attention Deficit Hyperactivity Disorder should not inhibit the search for environmental interventions. The classical single locus exemplar of this point is that of phenylketonuria, in which the homozygous recessive condition for a particular allele at the phenylalanine hydroxylase locus results in devastating mental retardation in ordinary environmental circumstances. However, the understanding of the metabolic process in which phenylalanine hydroxylase is involved led to the development of a rational intervention—a diet without phenylalanine, which cannot be metabolized normally by the homozygous recessive individuals.

Note that a gene is an unobserved hypothetical element both in a Mendelian system and in a polygenic system. The basic sequence involved in identifying a Mendelian inheritance pattern is to observe a categorical distribution of some phenotype, then to determine if the transmission pattern within families accords to Mendelian rules. In a polygenic, multifactor system, the issue is whether the pattern of resemblance among relatives conforms to expectations of the quantitative model. It is not required to know anything concrete about the gene or genes involved in either of these situations. But in fact, historically, while the range of application of Mendelian rules was being explored and the elaboration of the quantitative genetic model was proceeding, much was also being learned about the physical reality of genes. The key observation was that small bodies in the nucleus of cells behave according to rules similar to those of Mendel's elements. The hypothesis that these small bodies, the chromosomes, were the vehicles of the genes was quickly confirmed, and the field of cytogenetics emerged. Examination of the fine structure of chromosomes revealed distinctive banding patterns that permitted specification of chromosomal location, and subsequent research showed that the Mendelian elements resided in fixed positions on the chromosomes. The position of a gene came to be called its locus, for obvious reasons. The term, locus, was a terminological advance, making it possible to distinguish easily between the concept of gene, which everyone has, and the alleles, which different individuals can have in different combinations.

Throughout the first half of the twentieth century, a major biological mystery concerned the mechanism through which genes exert their influence. In mid-century occurred the revolution of molecular biology that discovered the chemical nature of genes and revealed the processes through which the genetic information in DNA is transcribed into RNA and translated into protein. This work connected the genetics of hypothetical genes to anatomy, physiology, immunology, endocrinology, biochemistry, neurobiology—to all of the disciplines that have explored the biological functioning of the organism. The question of the mechanism of gene action is thus answered in broad principle (although the details of specific cases may be unknown). Genes have no special route of influence to the phenotype; they operate through the complex systems comprising the familiar structures and functions of the body.

Complex systems such as this have some properties of great importance for the developmental sciences. A fundamental feature is the existence of feedback relationships of the sort involved in regulated processes, such as the familiar negative feedback in homeostatic systems. Another related feature is the typical presence of redundant pathways leading to the phenotype. With such properties, deficiencies in a particular element of a pathway might be compensated by up-regulation of other elements in that pathway, or by enhanced activity in the parallel pathway. Kacser and Burns (1979) have emphasized that in biochemical pathways with feedback loops, it becomes impossible to specify the impact of any element without specifying the other elements in the system. In general, the simple notions of linear, unidirectional causation must yield to circular or network causal concepts (Sattler 1986). Specific to interpretation of genetic influences, we are cautioned that the effect of genotypic differences at a particular locus may be greatly influenced by other loci in the system, or by the environment in which the organism exists. The magnitude and complexity of these interactions will be illustrated later.

3. *New Tools*

Recent developments in molecular genetics have provided some remarkable new methods with which to approach the issues identified in this review. Some of the most relevant are the following.

3.1 *Quantitative Trait Loci*

Although the original quantitative genetic model postulated an indefinitely large number of loci with equal small effects, it has long been suspected that within a polygenic set there may exist a distribution of effect sizes. Prior to the genome projects that have provided chromosomal markers throughout the genomes of human beings and several model species, attempts to locate such loci have been fraught with difficulties. With the availability of the new markers, it has become possible to identify segments of the chromosomes in which are located genes that have a detectable influence on the variability of a phenotype. This is obviously a critical step on the way to further characterization of the locus at the molecular level. These individuated polygenes are termed quantitative trait loci (QTLs).

3.2 *Transgenic Procedures*

Among other enticing possibilities, there are now procedures that bring about changes in the DNA molecule constituting a gene so as to prevent its expression (knockout preparations), to alter its regulatory system in order to cause overexpression, or to insert a new gene. These preparations, permitting the examination of the downstream results of altering single loci, provide new avenues of approach to the causal nexus between gene and phenotype.

3.3 *Arrays*

Wedding of the technological advances of microchips with molecular biology has provided chips on which

are arrayed hundreds or thousands of gene segments. When exposed to preparations containing DNA from particular organs or tissues, the chips can identify, through the hybridization of the tissue DNA to the segments on the chips, those genes that are being expressed in the tissue being investigated. Both the bright promise and the methodological issues of this new technology have been well presented by Claverie (1999) and by Lockhart and Barlow (2001).

Although these procedures and techniques are new on the scene, they are already providing powerful complements to the accumulated data on interactions and developmental changes in the architecture of complex phenotypes.

4. *Epistasis*

The phenomenon of interaction among genes has been known since the early days of the last century. An epistatic variance component can be incorporated in the quantitative genetic model, as well, but the detection of interaction effects in polygenic systems is challenging. There are numerous examples of the dependence of single locus effects on a polygenic background, however. A now-classic result can illustrate. Coleman and Hummel (1975) investigated the phenotypic consequences of homozygosity for a particular allele (*db*) at the diabetes locus in inbred mice. Inbred strains approximate the condition of being homozygous at all loci (see McClearn and Hofer 1999); animals within a strain are identical, and different inbred strains differ from each other at some unknown number of loci. It was possible to arrange the *db*/*db* genotype on the background of two different (but related) inbred strains—C57BL/6 and C57BL/Ks. The phenotypic outcome differed dramatically, in respect to body weight, glucose level, pathophysiology (atrophy of the Islet cells in one strain, and hypertrophy in the other), and in effects on lifespan.

The advent of QTL methods has greatly facilitated the study of epistasis. It has become possible to examine the interaction of these as yet unidentified (but approximately located) loci. In mice, to illustrate the point, QTL interactions have been described affecting blood pressure (Rapp et al. 1998), diabetes (Galli et al. 1999), epileptic seizure frequency (Frankel et al. 1995), and lung cancer (Fijneman et al. 1996).

In human research, an epistatic interaction concerning risk for dementia has been reported. The $\varepsilon 4$ allele of the *ApoE* locus has been implicated by many studies as a risk factor for the development of Alzheimer's disease. Overall, an allelic dose-response relationship has been observed, with a significant risk in heterozygous condition and an even greater risk in homozygous condition. Kamboh et al. (1995) describe an interaction of this locus and the *ACT* locus, alleles of which are labeled *A* and *T*. If an individual is homozygous *TT*, the risk associated with two $\varepsilon 4$ alleles is not greater than with one. The $\varepsilon 4$-associated risk is elevated if individuals are also *AA*, and individuals homozygous both for $\varepsilon 4$ and for *A* are especially vulnerable.

These few examples represent the scope for the influence of genetic context on the effect of allelic differences at specific loci (i.e., the effect of allelic substitution, or the allelic dose–response effect). Characterization of a gene as 'the' gene for some phenotype must be understood to involve acknowledging the potential of this contextual influence.

5. *Gene–Environment Interaction*

Under the rubric of gene–environment interaction are those phenomena in which the relation of genotype to phenotype is altered as a function of environment, or, conversely, in which the impact of an environmental agency or condition is dependent upon the genotype of the individual exposed to the environment.

This theme can be illustrated by the results of Fosmire et al. (1993), who, in the context of risk factors in Alzheimer's disease, explored the effects of exposure to high levels of dietary aluminum on brain aluminum levels in five strains of mice. Three strains showed no elevation as a consequence of the exposure, one showed a moderate effect, and one showed a threefold increase.

Logically parallel to evidence such as this on genetic influence on risks of deleterious outcomes engendered by environmental factors is evidence of genetic influence on responsiveness to intended therapeutic interventions. Yaffe et al. (2000), for example, investigated the effects of estrogen used by women 65 years of age and older on risk of cognitive impairment, and found an interaction with *ApoE* genotype. Risk was not affected in women who possessed at least one $\varepsilon 4$ allele, but was reduced in those who were $\varepsilon 4$-negative.

A final example illustrating the potential scope and complexities of gene–environment interactions may be taken from the work of Vieira and colleagues (2000). Seventeen QTLs were identified affecting lifespan in *Drosophila* reared in one of five different environments: three levels of rearing temperature, one involving a brief high temperature exposure, and one involving starvation stress. The results revealed an astounding level of interactions of the QTLs with sex and with environment. Some QTLs were effective only in one or a few environments and not in others; some were effective only in one sex; some were effective in both sexes, but only in one environment; and so on. In a summary statement, the authors reported that *all* of the genetic variance in this system was involved in some type of interaction!

6. *Temporal Considerations*

Dynamic change in the network of genetic and environmental factors over time is a key concept for developmental genetics. A fundamentally important

principle to emerge from investigations of the mechanisms of gene action is that genes may be turned on and off, or may have their level of expression up-regulated or down-regulated. There were temporal aspects to many of the previously cited studies; in a number of researches, the temporal dimension is the major focus. A particularly apposite example is the work of Rogina and Helfand (1995), who measured the expression of a gene in the antennae of *Drosophila melanogaster*. At hatching, there is little or no activity of this gene, but expression rises systematically to a peak in mid-life, with a subsequent gradual decline to near-zero levels at the end of life. When lifespan was shortened under conditions of an elevated rearing temperature or through the introduction of a different gene that shortens lifespan, the expression of the gene followed the same pattern, but in an accelerated manner, scaling to the proportion of the lifespan rather than to chronological time.

The existence of a program that 'knows' what proportion of a lifespan has already passed (and what remains) is an intriguing prospect. But for our present purposes, the important point is that this unambiguous demonstration of lifespan-scaled change in expression of this single gene informs us of the sort of changes that can be taking place in a polygenic system.

An example of dramatic developmental change in a mammalian system is provided by Cheverud et al. (1996) who studied mouse growth. Two sets of QTLs were identified: one affecting early growth but with diminishing influence by 6 weeks of age, and the other coming into play in the interval from 3 to 6 weeks of age.

It remains unclear how many or what proportion of loci can be expected to undergo systematic changes during the lifespan. The new chip technology is poised to generate a flood of data that will illuminate this issue, among many others. As a promissory illustration, Lee and colleagues (1999) have described differences between 5-month and 30-month old mice in respect to gene expression in skeletal muscle. Of 6,347 genes examined, 58 showed a more-than-twofold increase in expression levels and 55 showed a decrease of that magnitude. Although these numbers indicate that only a small percentage of genes are changing expression of this magnitude in this tissue over this time span, the impact of these changes in complex aging systems could be substantial. Information from other tissues is urgently needed, and will undoubtedly be reported soon.

Life stage has also been shown to be critical to the genetics of development in human beings. In a single locus example, Ferrari et al. (1998) showed that the major locus for the vitamin D receptor has an appreciable effect on the bone mass of pre-pubertal girls, but does not have any such influence during post-pubertal adolescence.

Concerning late life, there is a widely shared expectation that with advancing age the cumulative

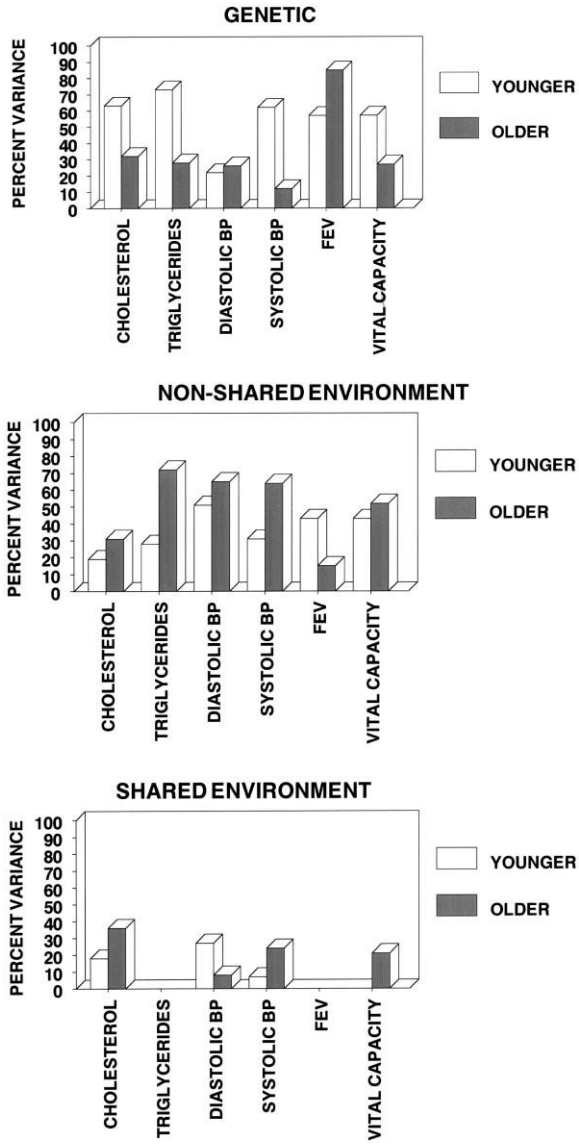

Figure 2
(**a**) Percent of variance of various biomedical phenotypes attributable to genetic sources. Data from the Swedish Adoption/Twin Study of Aging. Estimates are shown for a younger group of twin pairs (aged less than 65 years) and an older group (65 years of age and older). (**b**) Percent of variance of various biomedical phenotypes attributable to non-shared environmental sources. Data from the Swedish Adoption/Twin Study of Aging. (**c**) Percent of variance of various biomedical phenotypes attributable to shared environmental sources. Data from the Swedish Adoption/Twin Study of Aging

effects of uniquely experienced environmental factors will lead to reduced proportional contribution of genetic factors to that variance. Evidence on shifting variance components in aging has been provided by twin studies on the elderly; the classic expectation is fulfilled in the case of some phenotypes, but not all. Fig. 2 gives results from some biomedical variables included in the Swedish Adoption/Twin Study of Aging (SATSA) (Heller et al. 1993, Hong et al. 1994, McClearn et al. 1994). Estimates were derived from structural equation modeling of genetic, shared environmental, and non-shared environmental components of variance.

Participants in the study were Swedish identical and fraternal twins, some of whom were reared together and some of whom were separated in early life and reared separately. Estimates were derived separately for a younger group (less than 65 years of age) and an older group (65 years of age and older). Figure 2a gives the estimates of heritability. The difference between the older and younger cohort clearly conforms to the classical expectation of reduced heritability for measures of cholesterol, triglycerides, systolic blood pressure, and vital capacity. Heritability of diastolic blood pressure is, however, modest and essentially the same in both cohorts, and, with respect to forced expiratory volume (FEV), heritability is very substantial in the younger cohort, and even higher in the older cohort.

Fig. 2b gives the estimates for the influence of non-shared environment for these phenotypes. Most of these phenotypes (FEV being the exception) do display the classically expected result of greater non-shared environmental influence in the older group. Shared environmental influences, shown in Fig. 2c, are generally smaller than the nonshared environmental components. An interesting result is the hint that the effects of early shared rearing environment on cholesterol, systolic blood pressure, and vital capacity may be more manifest in the older cohort than in the younger one.

In addition to the fluctuation of genetic influence over the lifespan, there is growing evidence of shorter-term temporal variability. In rats, for example, the liver mRNA concentration of aldolase B, an enzyme of the glycolytic system, declines during fasting, and increases four- to eightfold upon re-feeding of a carbohydrate-rich diet (Munnich et al. 1985). Investigating the mechanisms involved revealed that regulation of the expression of the relevant gene is accomplished very differently in liver, kidney, and small intestine.

Another example of environmentally mediated alteration of gene expression is provided by the work of McMahon and colleagues (1992), who studied the effects of immobilization stress in rats. Among other results, it was found that a single stress resulted in elevation of levels of tyrosine hydroxylase mRNAs in the adrenal glands.

Thus, the responsiveness of the genetic system t2mo the environmental stimulus can occur quickly. From the point of view of behavioral science, an especially pertinent observation is that the expression of a gene in a particular brain region can be classically conditioned (Smith et al. 1992).

7. *Summary*

The studies reviewed here are but a few, selected, examples from a vast literature pertaining to an enormously broad array of phenotypes observed on species representing the phyletic range from prokaryotes to eukaryotes. Obviously, a brief review such as this can provide only the sketchiest outline. But, hopefully, it provides a hint as to the possibilities of dynamic interaction of genes with genes and of genes with environment and of the changes of this whole complex system over time. It may also serve as a preventive against simplistic interpretations of the old nature vs. nurture variety in issues of lifespan development. The current data are impressive; the anticipated flood of new data from new methodologies will provide us both with sharpened reductionist understanding, and, in the context of the quantitative model, an integrated view appropriate to the complex systems with which the developmental sciences are concerned.

See also: Behavioral Genetics: Psychological Perspectives; Familial Studies: Genetic Inferences; Genes and Culture, Coevolution of; Genes and Languages; Genetic Factors in Cognition/Intelligence; Genetic Studies of Behavior: Methodology; Genetic Studies of Personality; Genetics and Development; Human Evolutionary Genetics; Longevity, Genetics of; Mental Illness, Genetics of; Schizophrenia and Bipolar Disorder: Genetic Aspects; Senescence: Genetic Theories

Bibliography

Cheverud J M, Routman E J, Duarte F A M, van Swinderen B, Cothran K, Perel C 1996 Quantitative trait loci for murine growth. *Genetics* **142**: 1305–19

Claverie J-M 1999 Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics* **8**: 1821–32

Coleman D L, Hummel K P 1975 Influence of genetic background on the expression of mutations at the diabetes locus in the mouse. II. Studies on background modifiers. *Israeli Journal of Medical Science* **11**: 708–13

Falconer D S, Mackay T F C 1996 *Introduction to Quantitative Genetics*, 4th edn. Longman, London

Ferrari S L, Rizzoli R, Slosman D O, Bonjour J P 1998 Do dietary calcium and age explain the controversy surrounding

the relationship between bone mineral density and vitamin D receptor gene polymorphisms? *Journal of Bone and Mineral Research* **13**: 3363–70
Fijneman R J, de Vries S S, Jansen R C, Demant P 1996 Complex interactions of new quantitative trait loci, *Sluc1*, *Sluc2*, *Sluc3*, and *Sluc4*, that influence the susceptibility to lung cancer in the mouse. *Nature Genetics* **14**: 465–7
Finch C E, Kirkwood T B 2000 *Chance, Development, and Aging*. Oxford University Press, New York
Fisher R A 1918 The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**: 399–433
Fosmire G J, Focht S J, McClearn G E 1993 Genetic influences on tissue deposition of aluminum in mice. *Biological Trace Element Research* **37**: 115–21
Frankel W N, Valenzuela A, Lutz C M, Johnson E W, Dietrich W F, Coffin J M 1995 New seizure frequency QTL and the complex genetics of epilepsy in EL mice. *Mammalian Genome* **6**: 830–8
Galli J, Fakhrai-Rad H, Kamel A, Marcus C, Norgren S, Luthman H 1999 Pathophysiological and genetic characterization of the major diabetes locus in GK rats. *Diabetes* **48**: 2463–70
Hartl D L, Clark A G 1997 *Principles of Population Genetics*, 3rd edn. Sinauer, Sunderland, MA
Heller D A, de Faire U, Pedersen N L, Dahlén G, McClearn G E 1993 Reduced importance of genetic influences for serum lipids in the elderly: A study of twins reared apart. *New England Journal of Medicine* **328**: 1150–6
Hong Y, de Faire U, Heller D A, McClearn G E, Pedersen N 1994 Genetic and environmental influences on blood pressure in elderly twins. *Hypertension* **24**: 663–70
Kacser H, Burns J A 1979 Molecular democracy: Who shares the controls? *Biochemical Society Transactions* **7**: 1149–60
Kamboh M I, Sanghera D K, Ferrell R E, DeKosky S T 1995 *APOE*4*-associated Alzheimer's disease risk is modified by α1-antichymotrypsin polymorphism. *Nature Genetics* **10**: 486–8
Lee C-K, Klopp R G, Weindruch R, Prolla T A 1999 Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**: 1390–3
Lockhart D J, Barlow C 2001 Expressing what's on your mind: DNA arrays and the brain. *Nature Reviews Neuroscience* **2**: 63–8
McClearn G E, Hofer S M 1999 Genes as gerontological variables: Uses of genetically heterogeneous stocks. *Neurobiology of Aging* **20**: 147–56
McClearn G E, Svartengren M, Pedersen N L, Heller D A, Plomin R 1994 Genetic and environmental influences on pulmonary function in aging Swedish twins. *Journal of Gerontology: Medical Sciences* **49**: M264–8
McMahon A, Kvetnansky R, Fukuhara K, Weise V K, Kopin I J, Sabban E L 1992 Regulation of tyrosine hydroxylase and dopamine β-hydroxylase mRNA levels in rat adrenals by a single and repeated immobilization stress. *Journal of Neurochemistry* **58**: 2124–30
Molenaar P C M, Boomsma D I, Dolan C V 1993 A third source of developmental diffeces. *Behavior Genetics* **23**: 519–24
Munnich A, Besmond C, Darquy S, Reach G, Vaulont S, Dreyfus J O C, Kahn A 1985 Dietary and hormonal regulation of aldolase B gene expression. *Journal of Clinical Investigation* **74**: 1045–52
Rapp J P, Garrett M R, Deng A Y 1998 Construction of a double congenic strain to prove an epistatic interaction on blood pressure between rat chromosomes 2 and 10. *Journal of Clinical Investigation* **101**: 1591–5
Rogina B, Helfand S L 1995 Regulation of gene expression is linked to life span in adult Drosophila. *Genetics* **14**: 1043–8
Sattler R 1986 *Biophilosophy: Analytic and Holistic Perspectives*. Springer-Verlag, Berlin
Smith M A, Banerjee S, Gold P W, Glowa J 1992 Induction of c-*fos* mRNA in rat brain by conditioned and unconditioned stressors. *Brain Research* **578**: 135–41
Vieira C, Pasyukova E G, Zeng Z-B, Hackett J B, Lyman R F, Mackay T F C 2000 Genotype-environment interaction for quantitative trait loci affecting life span in *Drosophila melanogaster*. *Genetics* **154**: 213–27
Yaffe K, Haan M, Byers A, Tangen C, Kuller L 2000 Estrogen use, APOE, and cognitive decline: evidence of gene-environment interaction. *Neurology* **54**: 1949–54

G. E. McClearn

Genocide: Anthropological Aspects

Genocide is a crime in international law. It was established as such by the United Nations Genocide Convention of December 9, 1948 which committed 'contracting' states to 'prevent and to punish … acts committed with the intent to destroy, in whole or in part, a national, ethnical, racial or religious group.' This legal definition was reached following political compromises between the major powers at the UN, and a more precise, academic definition is adopted here, according to which genocide is 'a series of purposeful actions by a perpetrator(s) to destroy a collectivity through mass or selective murders of group members and suppressing the biological and social reproduction of the collectivity' (Fein 1990, pp. 28–9). After many years of academic avoidance of the topic, there has been a recent surge of interest across disciplines from psychology through political science to anthropology, the best of which work has been interdisciplinary in perspective. Sociological and anthropological investigation of genocide covers four main areas: the definition of genocide and the characterization of particular instances of mass slaughter; the sociology of genocides themselves; the sociology of representations of genocide (from 'survivor testimony' through memorialization and forms of material culture such as architecture to studies of war crimes tribunals): and the sociology of the after effects of genocide. These will be treated in turn.

1. Problems of Definition

The term genocide is a neologism, coined in 1944 by Raphael Lemkin, a legal scholar and Polish Jewish

émigré to the USA. Lemkin took over ancient Greek roots for 'tribe' and 'killing' to create a word that would adequately convey the enormity of Nazi and Axis crimes in Central and Eastern Europe during World War II. This definition inspired a UN General Assembly resolution of 1946 but when, two years later, the same body established a binding Genocide Convention, the range of reference was narrowed in two important ways. First, cases where a culture is destroyed without the extermination of people—now commonly referred to as *ethnocide*—were excluded. Secondly, and much more critically, in order to accommodate concerns of the USSR, which might have faced indictment for mass killing of 'class' enemies (as in the deliberately created famine in the Ukraine in 1932–3, or the then still undiscovered '*gulag*'), the 1948 UN definition excluded the extermination of 'political and other groups' such as social classes.

These questions of legal definition are important because the Convention commits states which subscribe to it 'to prevent and to punish' the crime of genocide by pursuing perpetrators 'whether they are constitutionally responsible rulers, public officials or private individuals.' Moreover, since the Convention explicitly denies that genocide can be a 'political' crime, perpetrators are subject to the laws of extradition.

Perhaps precisely because of its ambitious promise, the Convention has been invoked remarkably little since 1948. Both the desire of member states to preserve inviolate their claims to sovereignty and of others not to have to commit resources to humanitarian interventions have played a role. Leaving aside UN silence during the slaughter of half a million Indonesian communists in 1965 and the death of 1.5 million Cambodians in 1975–9 (since it could be argued that the 1948 Convention excluded such cases) it is a matter of record that US officials at the UN were instructed in 1992 to avoid using the term 'genocide' with respect to Bosnia so as not to trigger obligations under the Convention. In fact, the only major development since 1948 has been the establishment, in 1994, of the *ad hoc* International Criminal Tribunals for former Yugoslavia and Rwanda, though under the peace enforcement provisions of the UN Charter (Chapter VII), not the Genocide Convention, and refused permanent status due to Chinese objections.

While considerable academic energy has been expended in attempting to classify particular mass murders as genocide or otherwise (the destruction of the Albigensian Cathar heretics in France in the thirteenth century, the removal of up to 50 million persons from Africa during the slave trade, the slaughter of the Hereros of South West Africa by German forces in 1904–7, the massacres of the Armenians in 1915, the persecution of political enemies under Stalin in the USSR, the elimination of various indigenous peoples such as the Ache of Paraguay, to name but a few) perhaps the most broadly attended and resonant debates have concerned the uniqueness or otherwise of the Nazi genocide of the Jews. On the one hand, some scholars argue that the Jewish Holocaust is the only case in history of the attempted destruction of an entire people because they were that people. Other genocides were committed as a by-product of the course of war (Hereros) or political struggles (Democratic Kampuchea), or aimed to destroy a culture or way of life by eliminating some of its bearers (Gypsies in World War II), or occurred indirectly as a consequence of actions not *per se* intended to destroy a people (the slave trade). The Nazi persecution of the Jews differed from all these in intent, extent and 'success' (Bauer 1984).

Other scholars, however, resist the implication of the 'exceptionalist' position that the Holocaust occurred in some sense 'outside of normal history' and thus had neither precedent in the past nor lessons for the future. Unsurprisingly, this debate leads on to conflict over the interpretative framework for understanding the causes and the motivations of the Holocaust and its perpetrators. For their part, most anthropologists would favor retaining a comparative, sociological framework and resist the suggestion that any sociological phenomenon can be unique in the strong sense argued by exceptionalist scholars, or require recourse to abstract notions of 'evil' (Conte and Essner 1995, Clendinnnen 1999).

This debate has implications beyond German historiography. An anthropologist working in former Yugoslavia has recently argued that the public understanding of what constitutes genocide, which derives from images of the Nazi Holocaust, prevented observers from recognizing an actually existing genocide as it developed before the world's eyes (Sorabji 1994). Indeed, one of the puzzles of the sociology of genocide is why outside observers tend in general to be so unwilling to face the facts until it is too late, and why even after the event such effort is put into obfuscation and evasion of the brute historical record.

2. *The Sociology of Genocidal Mass Murder*

Though anthropologists aim to study the sources of human social and cultural variation, surprisingly little effort has been contributed by this discipline to the study of genocidal mass murder. This gap in the literature is all the more surprising in that many of the acknowledged genocides of the last century and a half have taken place in parts of the world where anthropologists work. Naturally, part of the explanation lies in the way anthropologists carry out their research. Participant observation in the life of a community is not a methodology that is easily adapted to the study of murderous gangs manning roadblocks or units of

the local military on active service. Beyond methodological considerations, there may also be a sense, shared with those who work for Amnesty International, that to explain is in some sense to forgive, and that the task of observers is merely to record, lest we forget.

Recently, however, anthropological avoidance of this topic—an evasion that was characteristic of most social science, with some noteworthy exceptions like Leo Kuper (1982)—has been coming to an end. Inspired in part by a wave of ethnographic writing about political violence, especially in South Asia (Spencer 1990), a new generation of anthropologists have written on particular genocides (Bosnia, Rwanda, and Cambodia most notably, but also Nazi Germany). In this work, the anthropologists attempt, with more or less success, to bring to bear a sense that violence is rarely simply pragmatic and is often patterned in ways specific to a particular time and place and always interpreted in local, cultural terms.

As in the other social sciences and humanities which deal with cases of mass murder, anthropological analyses aim above all to answer the question, 'how could this happen? Under what conditions did one group of humans attempt to wipe out another?' Responses can be divided along a familiar line into two camps: those who stress cultural factors and those who stress social, political, and economic factors. Taking the former first, for authors in this school the goal is to provide a specifically 'anthropological contribution,' to prove that an ethnographer's interest in 'cultural variation' adds an otherwise missing ingredient to analyses of genocide.

Much of this 'culturalist' work (notably writing on Rwanda and Cambodia) takes a cue, wittingly or not, from an essay by the historian Natalie Zemon Davis (1973) on differential patterns of violence in the religious wars in France in the seventeenth century. Davis showed that the forms of Protestant on Catholic violence were not the same as those perpetrated by Catholics on Protestants. Each religious tradition had its own construction of the nature of demonic power in the other and adherents of the two faiths acted to cleanse the world accordingly. Because of their interest in demonstrating a 'cultural logic,' the anthropological writing on Rwanda and Cambodia does not advance theoretically on Davis's original insight (Taylor 1999, Hinton 1998). Evidence is presented that shows how in some instances cultural patterns seem to have given form to and shaped people's violent behavior, but little or no attempt is made to explain how particular segments of the population or groups of individuals came to participate in the genocide at a particular time and place.

Curiously, this kind of explanation of 'how it happened' has several features in common with the approach adopted by Daniel Goldhagen (1996) in *Hitler's Willing Executioners*, a book which enjoyed runaway popular success across the USA, Europe, and particularly Germany in the mid-1990s (while being almost universally condemned by the historians with period expertise). Goldhagen's answer to the question 'how could it happen' was that it was Germans who killed Jews in World War II (it was not just they, in fact) and they did so because of the activation of a tradition which he dubbed 'eliminationist antisemitism' which had long characterized German culture (while in reality antisemitism had actually been a vote-loser for the Nazis prior to 1933).

For authors working in this tradition, genocide is derived or emerges out of a 'cultural model.' What tends to be missing from such explanations is any attempt to detail political manipulation by those who planned the genocides, to analyze the institutional apparatus which prepared and carried through the genocides, and to establish the sociological and even biographical profile of identified killers. In the case of Rwanda one has to turn away from the anthropologists, to the work of the political scientist Rene Lemarchand to discover, for instance, that the genocide was planned by a core of one extended family and advisors, promulgated through 300–500 middle level, rural cadres, set off by a presidential guard of around 6,000 persons who massacred all opponents of the genocide (Hutu and Tutsi alike) and finally continued by militias who may have made up 1–2 percent of the population, each member of which may have killed 200–300 people (Lemarchand 1997).

Work that searches for the cultural models which underpin and apparently feed directly in to genocide bears a close affinity with the first efforts by anthropologists to deal with issues of genocide during and shortly after World War II. Then adherents of the culture and personality school provided textbook/manual type explanations of the personality of the enemy in an effort to explain the success of totalitarian regimes in Japan and Germany and as a prelude to reconstruction and democratization. At the beginning of the twenty-first century, with decades of detailed analysis of the paradigmatic Nazi case behind us, there really is little excuse for allowing this kind of cultural reductionism to pass for 'anthropological explanation.'

An alternative approach within anthropology is to be found in work on Bosnia. Here the research was carried out by a British-trained *social* rather than US-trained *cultural* anthropologist (Sorabji 1994). This work owes much to the tradition of historical and sociological investigation of the Holocaust represented by such scholars as those around Herbert (2000), as well as the kind of political science of Bosnia pursued by Bougarel (1997). It puts stress less on 'culture' than 'local knowledge,' and tries to provide a fuller sense of 'what happened' before answering the more difficult question, 'why?'. Anthropological expertise is combined with a journalist's sense of political manipulation, a sociological interest in institutions, and a historian's concern with events of the recent and

remembered past. Not only is the Bosnian genocide not represented as being rooted in 'cultural traditions,' but it is shown how this very representation is part of the armory of the perpetrators, used, with great success in this case, to keep outsiders from interfering. As long as US Secretary of State Warren Christopher, or British Prime Minister John Major was advised that 'ethnic cleansing' was not 'genocide' and that deep-rooted cultural patterns on all sides were to blame, they felt justified in evading calls for intervention. Indeed, the very 'disorganized' quality of the genocide (compared with the Nazi model case) came to be taken as evidence of its nonexistence. In reality, however, institutional patterns were at work: Violent criminals were being released from Belgrade jails and sent to join militias. These operated formally independently but under the cover of the official military. Intellectuals were among their first targets, often identified on pre-prepared lists; local schools and other sites of intercommunal, neighborly interaction were chosen as locales for internment and torture; neighbors were deliberately sent to kill neighbors. The 'sites of memory' of Bosnian social life were being systematically and consciously destroyed by the instigators of a genocidal violence that was presented as medieval but was, in reality, wholly modern.

This kind of work points to a need to think more broadly about the psychological and mental underpinnings of the violent behavior of individuals and collectivities and to relate these findings to work elsewhere within anthropology. It is striking that the language of genocide in Bosnia, as in some other cases, was one of sacrifice, an invitation, one might think, to introduce anthropological work on the relation between religious representations of transcendence and political violence (e.g., Bloch 1992).

3. *The Sociology of the Representation of Genocide and Memorialization*

It is possible to argue that much of what has recently passed for analysis of the cultural soil of genocide in anthropology is in reality merely the re-presentation of local, folk explanations of genocidal behavior. This data, which is often part of the discourse around the genocide and ought to be part of the material analyzed and explained by a sophisticated social science, is instead relied upon, more or less uncritically, as if descriptively neutral.

In analyzing perpetrator discourse anthropologists have pointed out how genocide, like political violence more generally, tends to be presented by the perpetrators as a just defense against others who are about to attack, or are already attacking, the perpetrators. But a less sceptical attitude has been adopted, for instance, over accounts which focus on the rural, resentful character of supposed killers (Hinton 1998). Going further into the psychological roots of genocidal violence, the fundamental challenge, as everyone knows who has ever investigated these issues, is that no one ever admits to having personally been a genocidal killer. No researcher has yet overcome this obstacle.

A more fruitful domain has been analysis of survivor testimony and of the reconstructions of generations who have come after (a field which overlaps with broader interests in memory, narrative, and representations of the past). Discussion of the relationship between trauma, personal memory, and reconstruction of the 'self' is still at an early stage in all the disciplines which touch on this area. One feature that is commonly pointed to in victim testimony is a flat, matter-of-fact tone adopted, a technique of distancing and dissociation. In innovative work in transcultural psychiatry, Laurence Kirmayer has developed a notion of *landscapes of memory* to characterize some of the differences in narratives of trauma between child abuse and holocaust victims: 'while holocaust stories involve bearing witness to what is widely … recognized as a human catastrophe, personal stories of abuse are revelatory, shameful and damaging' (Kirmayer 1996, p.188). In this context memorialization may be a key feature of legitimating the experience of suffering. Work is, however, only just beginning on situations where memorialization does not take place (as for instance with most Gypsies in Eastern Europe).

One important new area of investigation is the social and political understandings and implications of particular ways of representing genocides. Early work in material culture on the activity of memory in holocaust memorials has shown how the meanings of memorials are contingent on social and political realities (Young 1993). At a literary level, work by Inga Clendinnen (1999) has opened up one path for a generation who no longer have access to direct testimony from survivors; more challengingly, Peter Novick's (1999) iconoclastic study of the Holocaust in US public life argues that some representations of genocide may act to inoculate a society against its own history and current world duties.

Apart from academic analysis, anthropologists have, at times, become centrally involved in acts of memorialization. In 1989, Polish Prime Minister Mazowiecki convened a commission to reconstruct the Auschwitz–Birkenau museum and monument for a postcommunist world. At the behest of the Polish Ministry of Culture, Jonathon Webber, an anthropologist with field research in Poland, and head of Oxford University's Centre for Hebrew Studies, invited an international body of experts to discuss 'the future of Auschwitz.' Alongside determining major issues of factual history—that 1.6 million had died there, and not four million as previously claimed, that 90 percent of those killed were Jews and not Polish and so on—the commission also addressed such tricky questions as to whether and how one should preserve

the ruins: reconstruction of the gas chambers was out of the question, but what sort of renovation should be permitted?

4. Postgenocide Societies?

Work on the representation of genocide is intimately connected with an interest in the afterlife of such massive social trauma. Many have suggested an association between German and Japanese urban terror in the 1960s with the phenomena of the 'second generation' but the link has never been established systematically. Now, however, a body of clinical work on individual responses to massive trauma is emerging. Imaginative work here often comes at the interface of disciplines, as in Lynne Jones's (2000) ethnographic-psychological examinations of how subjective understandings and political context mediate the impact of political violence on adolescent mental health in Bosnia. Jones, who is based at Cambridge University's Centre for Family Research, has, among other things, investigated children who had watched others being cleansed from their own town. In this case the 'therapeutic remedy' of 'the search for meaning' exposed children to uncomfortable truths about their community and appeared to lead to psychological discomfort, depression, and anxiety. Those children who avoided examining such events had better psychological health, but at the cost of leaving their elders' discourse unchallenged for a second generation.

For many victims of direct attack, the search for meaning, to name the unnameable is unavoidable. Here too, it is to be hoped, ethnographic work will emerge examining the impact of the International Criminal Tribunals. The president of the Yugoslav tribunal has sought to bring as many witnesses as possible to the Hague, so they can see their stories heard and entered in the register. But what effect has this had, and how will the expanding work of the Tribunal change the 'landscapes of memory' in Bosnia? It has been argued that significant use was made in 1990–2 of representations of experiences of World War II in cultivating an environment in which violence came to be seen as a serious option (Hayden 1994). Investigation of how societies deal with these kind of events is thus not merely of academic interest.

At an even broader level of analysis, anthropology would have much to contribute to the study of the experience of societies like Israel—for long unwilling to incorporate the genocide into its collective memory and now, having made it central to the nation, finding it hard to create a unified identity with Arab Jews who did not suffer. In the case of the perpetrator states the issue is almost invariably one of the social consequences of denial. Racist attitudes to Gypsies in Eastern Europe are likely related to denial and ignorance of Gypsy persecution in 1940–5. In this regard the German case is unique—a perpetrator state that has accepted historic responsibility. German foreign policy will be constrained by the Nazi past for the foreseeable future and perhaps as long as there is a German political entity, but it is also clear that for the time being this is a society which cannot shrug off its past. Daniel Goldhagen's work struck a profound chord in Germany perhaps because it offered a new generation a fresh chance to stare at an unadulterated image of the heart of some of their ancestors' 'darkness.' At that very moment there was also anguish over the Berlin Holocaust memorial. One particularly radical proposal was to construct a bus station in central Berlin with timetabled buses taking visitors and pilgrims to and from the death and concentration camps—a reminder of just how close so many of these camps lay to the capital of the Reich and lie to the new capital of a reunified Germany. It was rejected for 'lack of taste' and necessary 'monumentality.' Such phenomena can provide sociologists and anthropologists with rich seams of cultural and social significance.

See also: Anti-Semitism; Collective Memory, Anthropology of; Ethnic Cleansing, History of; Ethnic Conflicts; Ethnocentrism; Ethnonationalism: Cultural Concerns; Genocide: Historical Aspects; Holocaust, The; National Socialism and Fascism; Public Commemoration; Racism, History of; Racism, Sociology of; Social Trauma; Xenophobia

Bibliography

Bauer Y 1984 The place of the Holocaust in contemporary history. In: Frankel J (ed.) *Studies in Contemporary Jewry*. Institute of Contemporary Jewry, Hebrew University of Jerusalem; Indiana University Press, Bloomington, IN, Vol. 1, pp. 201–24

Bloch M 1992 *Prey into Hunter: The Politics of Religious Experience*. Cambridge University Press, Cambridge, UK

Bougarel X 1997 Bosnia and Hercegovina: State and communitarianism. In: Dyker D, Vejvoda I (eds.) *Yugoslavia and After: A Study in Fragmentation, Despair and Rebirth*. Addison-Wesley Longman, New York, pp. 87–112

Clendinnen I 1999 *Reading the Holocaust*. Cambridge University Press, Cambridge, UK

Conte E, Essner C 1995 *La Quete de La Race: Une Anthropologie du Nazisme*. Hachette, Paris

Davis N Z 1973 The rites of violence: Religious riots in sixteenth century France. *Past and Present* **59**: 51–91

Fein H 1990 Genocide: A sociological perspective. *Current Sociology* **38**(1): 1–126

Goldhagen D J 1996 *Hitler's Willing Executioners: Ordinary Germans and the Holocaust*. Knopf, distributed by Random House, New York, London

Hayden R M 1994 Recounting the dead: The rediscovery and redefinition of war-time massacres in late and post-communist Yugoslavia. In: Watson R (ed.) *Memory, History and Opposition under State Socialism*. School of American Research Press, Sante Fe, NM, pp. 167–84

Herbert U (ed.) 2000 *National Socialist Extermination Policies: Contemporary German Perspectives and Controversies*. Berghahn, New York

Hinton A L 1998 A head for an eye: Revenge in the Cambodian genocide. *American Ethnologist* **25**(3): 352–77
Jones L 2000 Adolescent understandings of political violence and their relationship to mental health: A qualitative study from Bosnia Herzegovina. http://www.waraffectedchildren.gc.ca/socscimedpaper-e.asp
Kirmayer L J 1996 Landscapes of memory: Trauma, narrative and dissociation. In: Antze P, Lambek M (eds.) *Tense Past: Cultural Essays in Trauma and Memory*. Routledge, London, pp. 173–98
Kuper L 1982 *Genocide: Its Political Use in the Twentieth Century*. Yale University Press, New Haven, CT
Lemarchand R 1997 The Rwanda genocide. In: Totten S, Parsons W S, Charny I W (eds.) *Century of Genocide: Eyewitness Accounts and Critical Views*. Garland, New York, pp. 408–23
Novick P 1999 *The Holocaust and Collective Memory: The American Experience*. Bloomsbury, London
Sorabji C 1994 A very modern war. In: Watson H, Hinde R (eds.) *War: A Cruel Necessity: The Bases of Institutionalised Violence*. I.B. Tauris, London, pp. 80–95
Spencer J 1990 Collective violence and everyday practice in Sri Lanka. *Modern Asian Studies* **24**(3): 602–23
Taylor C C 1999 *Sacrifice as Terror: The Rwandan Genocide of 1994*. Berg, Oxford, UK
Young J E 1993 *The Texture of Memory: Holocaust Memorials and Meaning*. Yale University Press, New Haven, CT

M. Stewart

Genocide: Historical Aspects

Genocide is any intentional act to physically eliminate a non-combatant group in whole or in part by direct assault, or by inflicting deadly life conditions including measures to prevent biological, economic and cultural reproduction.

This article will outline the development of genocide research as well as the tasks of this young academic discipline (Sect. 1). It will, in Sect. 2, deal with the problem of applying the twentieth-century term genocide to a wide variety of mega-killings which have happened throughout history. The term genocide has generated a tendency not only to rewrite, but also to morally judge anew, historical events to settle recent political accounts. Groups whose ancestors have suffered massacres at a period in which neither they themselves nor their victorious opponents had drawn up or adhered to international legal standards outlawing such atrocities may yet label their past fate as genocide to attract attention to contemporary plights. The entry will try to circumvent this extra-scholarly or ideological use of the term genocide by implicitly or explicitly using a caveat of the sort that an event called genocide 'today' was not seen as such then. It has to be stressed that all figures used in this entry are estimates. There is no such thing as an undisputed figure for even the best researched genocides.

1. *The Term Genocide and the Development of Genocide Research*

Genocide is not only a historical concept but also an international crime defined by the 1948 United Nations *Convention on the Prevention and Punishment of the Crime of Genocide*. In its Article II

> genocide means any of the following acts committed with intent to destroy, in whole or in part, a national, ethnical, racial or religious group, as such: (a) killing members of the group; (b) causing serious bodily or mental harm to members of the group; (c) deliberately inflicting on the group conditions of life calculated to bring about its physical destruction in whole or in part; (d) imposing measures intended to prevent births within the group; (e) forcibly transferring children of the group to another group.

The crimes mentioned in paragraphs (a) to (e) of Article II of the Convention are generalizations of practices employed by Nazi Germany after September 1, 1939 (beginning of World War II) against Slavs and/or Jews, and gypsies. The UN Convention was mostly written by a Polish lawyer and Holocaust refugee, Raphael Lemkin (1900–59), who had first in 1933 proposed a similar measure to the 'League of Nations.' In 1943, he had termed the crime in Polish as *ludobójstwo* (*lud* = people; *zabójstwo* = murder). A year later, he translated it into English as *genocide* (from Greek *genos* = people, and Latin *caedere* = kill, see Lemkin 1944, p. 79).

Since the UN Convention—today signed by some 140 nations—excluded 'politicide'—the murder of any person or people because of their politics or for political purposes—as well as 'ideological genocide'—the murder of any person or people because of their ideology, theory or knowledge—scholars have broadened the concept. Rudy Rummel's term 'democide' may be considered as the most comprehensive one. It is the murder of any person or people including genocide in the international legal sense, politicide as well as ideological mass murder (Rummel 1996, p. 39).

Besides the narrower term genocide and the more inclusive term democide the term 'crimes against humanity' is still in wide use. It was coined on May 24, 1915 by France, Great Britain, Russia, and the USA on behalf of the slaughtered Ottoman Armenians as 'new crimes of Turkey against humanity and civilization.' The term 'crimes against mankind' created, on October 12, 1942, in London (*Declaration of St. James*) to denounce the Nazi Holocaust of European Jewry is also still in use. Less known though quite appropriate is Gracchus Babeuf's French term 'populicide,' a shorthand for his 'système de dépopulation' executed—during the French Revolution—in the Vendée (1793/94) by killing its entire royalist peasantry of some 120,000 (Babeuf 1795).

The discipline of comparative genocide research draws on the entire corpus of historical research of the

last three millennia. Mega-killings have always attracted the attention of laymen and scholars alike. Yet, a first world history approach to genocide was tried by Lemkin who had drafted three volumes on the *History of Genocide* by 1959 (Jacobs 1999). Several publishing houses had turned him down. The manuscripts are only now being prepared for publication by Steven L. Jacobs.

The 1967–1970 massacres of half a million Ibos in Biafra/Nigeria triggered a new interest in the subject. In 1970, Yvan van Garsse published *A Bibliography of Genocide Crimes against Humanity and War Crimes* (Garse 1970). The genocide, in 1971, of some 1.5 million citizens of East-Pakistan (today's Bangladesh) also has drawn a steady albeit not widespread attention to the subject. Both killings had world wide television coverage. Only a year later, in 1972, *The Twentieth Century Book of the Dead* by Gil Elliot became the first major published monograph devoted to comparative genocide research (Eliot 1972) . In 1975, sociology entered the field with Dadrian (1975). Irving Louis Horowitz's *Genocide: State Power and Mass Murder* followed in 1976 (Horowitz 1997). In 1981, Leo Kuper reached a major academic publisher, Yale University Press, with *Genocide: Its Political Use in the Twentieth Century* (Kuper 1982). More systematically and historically minded works followed in 1990 with Fein's (1990) *Genocide: A Sociological Perspective* as well as Chalk's and Jonassohn (1990) *The History and Sociology of Genocide*.

The first—one person—institute devoted to the subject: 'Institute on the Holocaust and Genocide,' was founded in 1979 by Israel W. Charny in Jerusalem. Helen Fein followed in 1982 with the 'Institute for the Study of Genocide,' in New York. Next were Frank Chalk and Kurt Jonassohn with the Canadian 'Institute for Genocide and Human Rights Studies' (Montreal, Concordia University). Australia with the 'Centre for Comparative Genocide Studies' (Macquarie University), and Europe with the 'Raphael-Lemkin-Institut für Xenophobie- und Genozidforschung' (University of Bremen, Germany) followed in 1993. An international 'Association of Genocide Scholars' was established in 1995 with some 150 members by mid-2000. The first encyclopedia of genocide was published in Germany (Heinsohn 1999). A similar endeavor in English followed shortly afterwards (Charny 1999). Both works, with a combined volume of some 1,200 pages, cover most of the contents of this entry.

Historical, sociological, and sociopsychological studies of genocide aim at the assessment of genocidal potential as well as the formulation of means to prevent the crime. Interventions against ongoing genocides are rather dealt with in studies of warfare than in comparative genocide research. The prevention of genocide requires an early warning system. Historically, roughly 25 percent of high risk nations have transformed their perpetrator potential into actual genocide (the rate is higher for democide). Since every genocide needs certain steps of preparation which are difficult to conceal, like, e.g., registering the victims and training the killers, all nations with perpetrator potential (today close to 100) may one day be scanned for these red alert signals indicating the imminence of a genocide.

2. *History of Genocide*

Because of its modern provenance it is difficult to employ the term genocide to the common periodization of history. Homer's *Iliad* (VI: 55–62), e.g., paints, as perfectly legitimate, an intention that today would be considered as genocidal in its full legal sense:

> "My good Menelaos" [king of Sparta], said he [Agamemnon, king of Mycene], "this is no time for giving quarter. Has, then, your house fared so well at the hands of the Trojans? Let us not spare a single one of them—not even the child unborn in its mother's womb; let not a man of them be left alive, but let all in Ilius [Troy] perish, unheeded and forgotten." Thus he did speak and his brother was persuaded by him, for his words were just.

Moreover, a practice of killing considered the expectable as well as acceptable outcome of conflict in an early period of history may continue to this very day. As an example in question one may look at the Yanomami Indians of Southern Venezuela and Northern Brazil. They must—since 1980—be considered a possible victim of genocide by modern settlers and—at the very same time—were successful in nearly wiping out a neighboring tribe (Keegan 1993, p. 97).

2.1 *Tribal Societies*

Genocidal acts in the modern sense were hardly distinguishable from acts of war in tribal societies—covering the longest stretch of human history—because there was no separation of noncombatant civilians and warriors. Warfare was '"ecological" in motivation' and led to the redistribution of hunting and fishing grounds or arable land 'from the weak to the strong' (Keegan 1993, p.101). Means of these wars were 'battles of annihilation' or, if space allowed, 'displacement of the weaker' (Keegan 1993, pp. 29, 387). Genocidal tribal raids continued well into the second half of the twentieth century in territories of the Amazon, New Guinea, the Philippines and elsewhere: 'Typically the goal in such raids is to destroy an entire village or settlement, the common intention being to kill everyone, perhaps excepting a few of the younger women and children who are carried off for incorporation into the home group' (Glick 1994, p. 45).

2.2 Early Antiquity

Early stages of civilization (Early and Middle Bronze Age with priest-kingship feudalism) saw tribal practices of annihilation on a larger scale: 'I slew them without sparing them, they sprawled before my horses, and lay slain in heaps in their blood,' Egypt's Pharaoh Ramses II (conventionally dated 1279 to 1212) boasted after a campaign against the Hittites whose leader, then, begged him for mercy: 'Look, you spent yesterday killing a hundred thousand and today you came back and left no heirs. Be not hard in your dealings victorious king' (Lichtheim 1976, pp. 70 f.). Other powers of the Ancient Near and Far East behaved very much in the same way, as did Bronze Age, i.e., Mycenaean Greece.

2.3 Antiquity/City States

The annihilation practices of time immemorial were cast into a principle by Athenians. When, in the Peloponnesian War (431 to 404 BC), Athens had decided to wipe out Sparta's ally Melos, the admirals in charge tried to formulate an eternal law underlying their action:

> Our knowledge of men leads us to conclude that it is a general and necessary law of nature to rule wherever one can. This is not a law that we made ourselves, nor were we the first to act upon it when it was made. We found it already in existence, and we shall leave it to exist for ever among those who come after us. […] We know that you or anybody else with the same power as ours would be acting in precisely the same way. […] The Melians [in 415 BC] surrendered unconditionally to the Athenians, who put to death all the men of military age whom they took, and sold the women and children as slaves (Thucydides, *The Peloponnesian War*, (1954 trans) V: 85 ff.).

As a first step away from mutually slaughtering their males, the Greeks developed the practice of exchanging prisoners of war.

Within the Roman Empire—a private ownership based society on a much larger scale—the practice of destroying enemies who had caused problems in the past was continued. The elimination of Corinth (metropolis of mainland Greeks) and Carthage (metropolis of Phoenicians) in 146 BC, of Numantia (capital of Spanish Celtiberians) in 134 BC, and the elimination of the Etruscans (92 BC) and the Samnites (82 BC) as major opponents in Italy proper provide typical examples with 50,000 to 150,000 dead in each case. Of course, the opponents acted in the same manner as, e.g., in the slaughter of the inhabitants of the Greek cities Selinus and Himera (Sicily) in 409 BC by the Phoenicians of Carthage. In 88 BC all Romans and Italians, including women and children, living in Greek Asia Minor were put to the sword by orders of Mithradates VI. The Romans of Britannia met a similar fate in 61 BC. Within the Roman Empire a *ius gentium* was developed for strangers who did not hold citizenship with, sometimes, more protection for these minorities than for full citizens.

Parallel to practices of the Greek City States and the Roman Empire, movements against blood sacrifice emerged around 600 BC. Well known are Buddhism in India and the Pythagoreans of Classical Greece. It was, however, monotheistic Judaism whose moral code—albeit slowly—came to influence international ethics most deeply. To overcome expectations of salvation tied to the slaughter rituals of blood sacrifice, Hebrew prophets found a new moral vision: 'I desired mercy, not sacrifice; and the knowledge of God more than burnt offerings' (*Hosea* 6:6). The sanctity of life became the highest principle: 'Thou shalt not kill' (*Exodus* 20:13 and *Deuteronomy* 5:17)] including—in its universal character—the strict outlawing of infanticide.

The Torah—rather a moral compilation than a historical record—emphasizes the sanctity of life as the core of the law and as a value identical with goodness ('the good law'): 'See, I have set before thee this day life and good. […] I call heaven and earth to record this day against you, that I have set before you life and death, blessing and cursing: therefore choose life, that both thou and thy seed may live' (*Deuteronomy* 30:15, 19). From the principle of life's sacredness derived other commandments like 'Thou shalt love thy neighbor as thyself' (*Leviticus* 19:18). Anti-xenophobic teaching followed suit: 'And if a stranger sojourn with thee in your land, you shall not vex him. But the stranger that dwelleth with you shall be unto you as one born among you, and thou shalt love him as thyself' (*Leviticus* 19:33–34). 'Are ye not as children of the Ethiopians unto me, O children of Israel? Saith the Lord. Have not I brought up Israel out of the land of Egypt, and the Philistines from Caphtor, and the Syrians from Kir' (*Amos* 9:7).

Against Judaism, the first recorded order (of Haman) for an ideological, though abortive, genocide (also called religiocide) was issued in the Akhaemenid Empire of fourth century BC: 'And the letters were sent by posts into all the king's provinces, to destroy, to kill, and to cause to perish all Jews, both young and old, little children and women, in one day' (*Esther* 3:8/13). Christians who had adopted the Jewish rejection of blood sacrifice to planetary deities actually suffered the first well recorded religiocides. The edict of Emperor Decius of AD 250 forced them to eat sacrificial meat or suffer death. Four more edicts by Diocletian from AD 303/304 repeated the order upon pain of death. The massacres ended in AD 312 when Constantine the Great converted to the new faith.

2.4 Middle Ages

The Middle Ages saw attempts at a *Pax Romana*, directed by the Papal See. Its concept of a *Pax Dei* (peace of God)—developed since AD 975 and extended over all Christian lands at the Synod of Clermont

(AD 1095)—outlawed war from Wednesday night to Monday morning as well as on Christian holidays. Excommunications and interdicts were the only means of enforcement. It was the beginning of international law based on the Judeo-Christian code of ethics. Yet, the transgressions were severe and, in modern terminology, definitely genocidal.

The crusades—starting in AD 1096—not only brought massacres of thousands of Jews in Germany and tens of thousands of Jews and Muslims in Jerusalem (AD 1099), but also of fellow Christians. Against Pope Innocent III's interdict, Dalmatia's port Zadar, arch rival of Venice, was bathed in the blood of all its inhabitants in November 1202 during the fourth Crusade by French and Italian troops transported on Venetian ships. The same men butchered 2,000 inhabitants of Greek Constantinople on April 13, 1205 before they went on a three-day-rampage of indiscriminate slaughter, not sparing women and children.

Since 1167, the Cathars (from Greek *katharoi* = pure), called Albigensians after the French city of Albi, suffered a series of religiocides by Catholic forces, because they rejected any spilling of blood and did not believe in the resurrection of Christ, and the Holy Communion as the residue of blood sacrifice within Christianity. By 1234 the Cathars were wiped out with the massacre of Béziers as the largest slaughter of this crusade in which all 15,000 inhabitants were killed.

A few years later, in AD 1211, Gengis Khan's Mongols (Morgan 1986)—with no authority to arouse their consciousness by threats of excommunication or invoking the image of all men as protected children of the one God—began their terrifying campaigns from the Japanese to the Baltic Sea. To turn their fertile areas into horse pastures the Mongols have, up to AD 1290, eliminated some 25 million Chinese. Another 5 million perished in Persian, Arab, and Christian territories where the Steppe warriors revived the memory of Attila's fifth century Huns. Behind their lines, the Mongols used to kill everyone perceived as a future threat, i.e., all males not suitable to join their ranks. Defeated armies were allowed a brief lease to life by serving as shields or trail blazers for advancing Mongol troops. Tamerlane, a pious Muslim, supposedly had, in a single day of 1398, some 100,000 Hindus strangled in sight of the walls of Delhi which had rejected surrender. In 1402, he dealt a mortal blow to Arab civilization by killing 90,000 inhabitants of Baghdad.

2.5 Modern Period

A sovereign's right to war (*ius ad bellum*) dominated the period of sixteenth to nineteenth century in which Europe came to dominate the world. What happened within a war was more or less up to the sovereign. Laws of warfare or the definition of crimes of war (*ius in bello*) had to wait for the turn of the nineteenth/twentieth century. In Europe as well as overseas, the results were genocidal. To give only two examples, Tenochtitlan, the Aztec capital in the valley of Mexico with its 100,000 inhabitants, was totally annihilated by Cortes' Spaniards in 1521. All 30,000 inhabitants of German Magdeburg, the richest city of Protestantism, were slaughtered on May 10, 1631 by Catholic troops under General Tilly in the Thirty Years' War.

The European Population Explosion—beginning at the end of the fifteenth century, and still not fully understood—bestowed this small continent with 25 percent of world population by 1900. During the same period, some 60 million young Europeans, with the superior military technology of their ownership-based economies but no attractive positions in an overcrowded continent, conquered their settlement space in the Americas, Australia, Africa, and Northern Asia. In the wake of this 'emigration,' a population of 55 to 85 million native Americans estimated for 1492 (Spanish conquest) was reduced by 85 to 95 percent around 1650. Between 250,000 and 500,000 native Australians estimated for 1788 (British conquest) were reduced to 60,000 by 1921. All means of killing were employed from outright genocide to infections—including intentional ones. Africa, between 1500 and 1870, lost up to 20 million inhabitants into American slavery with 2 million dying on transport. The slave trade within Africa may have cost, for the same period, the lives of 4 million. On the slave routes to—mostly Islamic—Asia another 11 million may have perished in these gigantic democides (Rummel 1996, p. 48).

Though not an invention of modern times, politicide became one of the most heinous activities of this period. It does not necessarily include the killing of the families of the victims. However, in a so-called 'Vespers' in which politicide is turned against a foreign elite, it usually grows into a full blown genocide. It took its name from the Easter massacre in Palermo on March 29, 1282, at the hour of evening prayer (Vespers) when the French in control of the Kingdom of Sicily were killed with their families ('Sicilia Vespers'). It had been preceded, in 1268, by the butchering of some 5000 members of the Hohenstaufen elite of Sicily by the hands of the French conquerers. The best remembered example of politicide was the annihilation, after eight inconclusive civil wars, of the Protestant leadership of France (ca. 100,000 victims), starting on Saint Bartholomew's Day (August 24, 1572) when the Huguenots—easily recognizable by their black clothing—were guests at the royal marriage of Henri of Navarra and Marguerite of Valois. The politicide, in February 1579, against Novgorod, Russia's richest city, by Ivan IV (the Terrible) was peculiar insofar as the Tsar had put together special killing units—the 'Oprichniki'—because the army was not trusted to act ruthlessly enough. In the 'Ulster, Massacre' (October 23, 1641), as the last 'Vespers' of

Table 1
Rough estimates of numbers of victims by genocides, politicides and ideological mass murder of a wide variety of regimes usually lumped together under the term Communism (Rummel 1996).

Soviet Union, 1917–87	40–60,000,000
Red China plus Maoists before 1949	40–60,000,000
Khmer Rouge in Cambodia, 1975–79	2,000,000
North Korea, 1945 onwards	1,700,000
Yugoslavia of Tito, 1941–87	1,100,000
North Vietnam, 1945–87	1,700,000
Poland (deportations of Germans), 1945ff.	1,400,000
Czechoslovakia (deportations of Germans), 1945ff.	240,000
Ethiopia, 1974–91	4,000,000
Albania, 1944–90	100,000
Communist Germany 1945–52	96,000

Western Europe, Irish rebels killed most of the Protestant British settlers and administrators on their island. The Creoles of Haiti delivered their French masters a 'Vespers' in 1791/1805 as did, in 1857/58, albeit with no lasting success, the sepoys to the British masters of India.

2.6 Twentieth Century

The Hague Conventions of 1899 and 1907 with *The Regulations Respecting the Law and Customs of War on Land* prohibited, in Articles 22–8, the murder, ill-treatment, or deportation of enemy civilians and prisoners of war. The use of poison and special ammunitions was banned. However, a nation's harm to its own civilians was not taken into consideration. This had to wait for the 'Istanbul Trials,' installed in January 1919 to inflict punishment on those responsible for the Armenian massacres. On April 10, 1919, Mehmet Kemal Bey, procurator of the province Bogazliyan, became the first person executed for genocide, then called crime against humanity. In the Nuremberg Trials of Nazi-Germany's leadership (1945/46) as well as in the Tokyo Trials of Imperial Japan's leaders (1946–48), 'International Military Tribunals' further developed the legal tools to punish perpetrators of crimes against humanity. Special United Nations courts explicitly devoted to the crime of genocide started, in 1993, to prosecute the perpetrators in Yugoslavia ('ethnic cleansings,' since 1991) and, a year later, in Rwanda (genocide of Tutsis in spring 1994). The statute for a permanent 'International Criminal Court' was passed in Rome on July 17, 1998.

The entire twentieth century is on record for close to 180 million victims of democide according to Rummel's statistics. The extermination of the Ottoman Armenians (1894–96, 1909, 1915/16, 1919, 1923) combined strategic, racial, and ideological intentions (Dadrian 1995). A shrinking empire was to be consolidated by a homogenous pan-Turkic population within an undivided settlement space under Islam as the only religion. That's why not only more than 1 million Armenians suffered genocide but also other Christian minorities like Maronites (1860), Assyro-Chaldaeans (1915), and Greeks (1915–23) with a combined death toll of some 440,000.

Since 1917, single party totalitarian regimes with agendas may have killed up to 110 million people (Rummel 1996, Courtois 1997). After World War II, 'one-party communist states are 4.5 times more likely to have used genocide than are authoritarian states' (Fein 1933b, p. 79)

The figures given in Table 1 are all estimates. Totalitarian regimes are not only known for most of the killings of the twentieth century, but also present the most difficult problems for research because archives were destroyed and many deeds never recorded. Thus, totalitarian regimes have to live with possibly exaggerated figures of their atrocities.

Extreme anti-capitalism slowed economic development though superior prosperity was promised. The shortcomings were frequently blamed on sabotage and/or party comrades still infested with the reactionary spirit of ownership. This view was used to justify massive party purges like the 'Great Terror' in the Soviet Union (1934, 1936–38) with more than 7 million or the 'Cultural Revolution' in China (1964–75) with nearly 8 million dead.

Annihilation through forced labor in camps was initiated by Trotzki on June 26, 1918. By 1922, 23 Soviet camps were at work, forming the nucleus of Stalin's Gulag fully emerging in 1928, and eventually running some 8000 camps. Working people to death became, in the long run, the most devastating killing tool of history. It allowed the elimination of opponents by simultaneously using them for the creation of big projects—like canals, railroads, subways, etc.—to prove by 'miracles' the promised productive superiority of communism that otherwise was wanting.

The traditional siege tool of man-made famines had to wait for the twentieth century to be welded into the most efficient fast-killing techniques ever. To crush national ambitions of the Ukraine, for example, a famine was turned into the mass death of some 7 million in the single winter (4 months) of 1932/33. Ethiopia's Derg managed to kill up to 4 million peasants between 1982 and 1986 by the same means, often intentionally aggravating natural disasters.

The mega-killings by left totalitarian regimes came as a combination of politicides (elimination of like minded rivals for power), genocides (elimination of minority nations, mostly in the Soviet Union and the Peoples Republic of China), and ideological genocides (foremost the elimination of proprietors or property-infested citizens). The latter category probably became the largest victim group in history with some 50

million dead (55 percent in China, 40 percent in the Soviet Union, 5 percent in the other communist regimes).

Of all twentieth-century totalitarian regimes, Nazi Germany was the most genodical. Following September 1, 1939, deformed newborn and mentally-handicapped Germans—as well as some German soldiers severely wounded in the military campaign against Poland—were killed by injections or in gas chambers. Close to 130,000 handicapped eventually ended in this so-called euthanasia operation. Later exterminations struck Gypsies, stigmatized as socially inferior, as well as non-reproductive homosexuals etc. A bigger target, however (outlined in the '*Generalplan Ost*' (Masterplan East)) were 100 million Slavs (Czechs, Poles, White-Russians, Ukrainians, and Russians) living between Germany's Eastern border and the Ural Mountains, a territory marked to become Germany's 'Lebensraum.' Of the Polish elite, 75,000 members, most of them holders of academic degrees, were killed in autumn 1939. The second Eastern campaign, against the Soviet Union (June 22, 1941), was aimed at the immediate killing of 30 million Slavs. Another 30 million were destined to be worked to death to lay the infrastructure for German settlements. The remaining 40 million were to be driven into Siberia with a calculated death toll of another 10 million. With 13 million Slavic civilians and prisoners of war killed by 1945 (Kumanev 1990, Vitvitsky 1990), Nazi Germany had fallen short of its 70 million goal but still had committed the largest twentieth-century genocide explicitly directed against a particular ethnic group. Nazi Germany's Eastern campaigns, thus, were not only massive wars but also militarily enforced genocides (Hillgruber 1993).

In connection with the Eastern wars, 5 to 6 million Jews were killed. Though a 'theory of the Holocaust is lacking' (Herbert 1998, p. 66), Hitler's practice of 'strengthening' Germany by killing its handicapped and of gaining settlement space through genocide of its Slavic inhabitants put him inevitably against the Torah. To him the sanctity of life was a Jewish 'invention,' a weakening ideology that could be cleansed from the German sphere of influence by exterminating its 'carriers.' The Holocaust, in this perspective, was a 'genocide for the purpose of reinstalling the right to genocide' (Heinstohn 2000, p. 424).

Since 1944/45, some 100 acts, fulfilling the definition of the UN Genocide Convention have taken place. In the two decades from 1980 to 2000, some 80 distinct ethnic and religious groups—mostly in Africa, Asia, and Latin America—were victimized with 12 of them going through full blown genocide: Bosnian Muslims, Kurds, Tibetans, Dinka (Sudan), Nuba (Sudan), Shilluk (Sudan), Isaaq (Somalia), Tutsi (Rwanda), Mayans (Guatemala), Albanians (Serbian Kossovo), and Timorese of Indonesia (Harff and Gur 1996, updated by the author).

2.7 The Future

Ecologically and ideologically motivated genocides most probably will have a future in the twenty-first century. Politicides may also occur wherever solidly established democratic structures are missing. Democracies usually do not go to war against each other. The same cannot be said for democide like nuclear and carpet bombing against totalitarian powers with, e.g., a death toll of 1.5 million in Germany (600,000) and Japan (900,000) during World War II.

The prevention of genocide by spreading democracy remains the preeminent goal for the future. Meanwhile, some experience with intervention has been made, in 1999, in Kossovo and Timor. To preempt the cruelties unavoidable in an intervention by force, genocide early-warning stations are now seriously considered. They may be in place within a few years.

See also: Communism; Ethnic Cleansing, History of; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Eugenics, History of; Genocide: Anthropological Aspects; Holocaust, The; National Socialism and Fascism; Religion: Nationalism and Identity; Religion: Peace, War, and Violence; Totalitarianism; Violence, History of; War Crimes Tribunals; Xenophobia

Bibliography

Babeuf F N 1795 *Du système de la dépopulation … avec la guerre de la Vendée*. Franklin, Paris

Chalk F, Jonassohn K 1990 *The History and Sociology of Genocide. Analysis and Case Studies*. Yale University Press, New Haven, CT

Charny I W (ed.) 1999 *Encyclopedia of Genocide*. ABC-CLIO, Santa Barbara, CA

Courtois S (ed.) 1997 *Le Livre Noir du Communisme: Crimes, terreur, répression*. Laffont, Paris

Dadrian V N 1975 A typology of genocide. *International Review of Modern Sociology* **5**: 201–12

Dadrian V N 1995 *The History of the Armenian Genocide: Ethnic Conflict from the Balkans to Anatolia to the Caucasus* (1995), Berghahn. Providence, RI

Elliot G 1972 *Twentieth Century Book of the Dead*. Allan Lane/The Penguin Press, London

Fein H 1993a *Genocide: A Sociological Perspective*. Sage Publications, London

Fein H 1993b Accounting for genocide after 1945: Theories and some findings. *International Journal on Group Rights* **1**: 79–106

Garsse Y v 1970 *A Bibliography of Genocide, Crimes Against Humanity and War Crimes*. Studiecentrum voor Kriminilogie en Gerechtelijke Geneeskunde, Sint Niklaas Waas, Belgium

Glick L B 1994 Religion and genocide. In: Charny I W (ed.) *The Widening Circle of Genocide. Genocide: A Critical Bibliographic Review*. Transaction, New Brunswick, NJ, Vol. 3, pp. 43–74

Harff B, Gurr T 1996 Victims of the state: Genocides, politicides and group repression from 1945 to 1995. In: Jongman A J (ed.) *Contemporary Genocides: Causes, Cases, Consequences*. PIOOM. Leiden, The Netherlands, pp. 33–58

Heinsohn G 1999 *Lexikon der Völkermorde*, 1st edn. 1998, Rowohlt, Taschenbuch Verlag, Reinbek bei Homburg, Germany
Heinsohn G 2000 What makes the Holocaust a uniquely unique genocide. *Journal of Genocide Research* **2**: 411–30
Herbert U (ed.) 1998 *Nationalsozialistische Vernichtungspolitik 1939–1945: Neue Forschungen und Kontroversen*, Fischer, Frankfurt, Germany
Hillgruber A 1993 *Hitlers Strategie: Politik und Kriegfuehrung 1940–1941* [1963]. Bernard & Graefe, Bonn, Germany
Horowitz I L 1997 *Taking Lives: Genocide and State Power*. Transaction, New Brunswick, NJ
Jacobs S L 1999 The Papers of Raphael Lemkin: A First Look. *Journal of Genocide Research* **1**: 105–14
Jonassohn K, Björnson K S 1998 *Genocide and Gross Human Rights Violations in Comparative Perspective*. Transaction, New Brunswick, NJ
Keegan J 1993 *A History of Warfare*. Knopf, New York
Kumanev G A 1990 The German occupation regime in occupied territory in the USSR (1941–1944). In: Berenbaum M (ed.) *A Mosaic of Victims: Non-Jews Persecuted and Murdered by the Nazis*. New York University Press, New York, pp. 128–41
Kuper L 1982 *Genocide: Its Political Use in the Twentieth Century*. Yale University Press, New Haven, CT
Lemkin R 1944 *Axis Rule in Occupied Europe*. Carnegie Endowment for International Peace, Washington, DC
Lichtheim M 1976 *Ancient Egyptian Literature. Volume II: The New Kingdom*. University of California Press, Berkeley, CA
Morgan D 1986 *The Mongols*. Blackwell, Oxford
Rummel R J 1995 *Statistics of Democide: Genocide and Mass Murder Since 1900*. Center for National Security Law, Charlottesville, VA
Rummel R J 1996 *Death by Government 1994*. Transaction, New Brunswick, NJ
Thucydides 1954 *The History of the Peleoponnesian War*, trans. R Warner. Penguin Books, Harmondsworth, UK
Vitvitsky B 1990 Slavs and Jews: Consistent and inconsistent perspectives on the holocaust. In: Berenbaum M (ed.) *A Mosaic of Victims: Non-Jews Persecuted and Murdered by the Nazis*. New York University Press, New York, pp. 101–8

G. Heinsohn

Genotype and Phenotype

The fundamental distinction between genotype and phenotype was first introduced into biology by Wilhelm Johannsen in 1911 to make explicit a distinction that was implicit in Mendel's explanation of heredity, and which marked it off from previous theories of inheritance. Mendel found that when he crossed two pure-breeding varieties of garden pea, the first generation of hybrid offspring were indistinguishable from one of the two parental strains, but if these hybrids were crossed with each other, both of the original types appeared in the offspring of these hybrids. So, for example, a pure bred red-flowered crossed with a pure bred white-flowered form produced all red-flowered plants, but these, when interbred with each other produced both red- and white-flowered plants.

In order to explain this extraordinary result, Mendel postulated the existence of internal, unseen 'factors' contributed by each parent to the offspring, factors that were causal determinants of the outward appearance of the plants, but whose nature was unaffected by their passage from one generation to the next. So, pure bred white-flowered plants had two white 'factors' one contributed by the pollen and one by the egg of their male and female parents, while pure bred red-flowered plants had received two red 'factors' from their two parents. The hybrids of red-flowered and white flowered forms would then contain one red and one white 'factor,' yet these plants were indistinguishable in appearance from a parent that had two red 'factors.' There was then a fundamental distinction between the state of the internal causal 'factors' and the outward state of the plant that carried them. Moreover, to explain the reappearance in the next generation of some white-flowered plants, Mendel had to suppose that the temporary coexistence of red and white 'factors' in the hybrids had no effect on their nature, but that they could be passed on uncontaminated to the offspring of the hybrids. An offspring of the hybrids that happened to contain two white 'factors' was then white like its white grandparent.

This scheme of explanation makes a radical separation between the internal factors that behave according to one set of physical relations in their passage between generations, the laws of *heredity*, and the external appearance of the organism that results from a different set of physiological relations which determine the consequences to the organism of carrying those factors, the laws of *development*. This distinction between the phenomenon of heredity and the phenomenon of development which we usually associate with August Weissman's separation of the germplasm from the somatoplasm had already been made by Mendel twenty years earlier. Earlier theories of heredity had not made such a separation of inner factors from outward manifestation and had been posed as the problem of the inheritance of the organism's characters themselves. It was Mendel's realization that characters are not inherited, they develop. In Johanssen's terminology the description of an organism's internal factors that were passed on to it in the gametes from its parents is a specification of the organism's *genotype*, while the description of the total physical and behavioral state of the organism is a specification of its *phenotype*.

1. Description of Genotypes and Phenotypes

As a result of the discoveries in molecular and cellular biology since Johanssen's time genotypic and phenotypic descriptions have become more concrete, more detailed and greatly amplified in their scope. Originally genotypes had no independent physical descriptors

but were distinguished from each other entirely on the basis of the external phenotypic manifestations and the patterns of similarity in phenotypes of related individuals. 'Genes' were simply abstract placeholders of unknown physical composition, inferred entities that could be characterized only by their phenotypic effects.

The modern identification of DNA as the physical material that behaves in heredity in the manner inferred for Mendel's mysterious 'factors' has allowed a precise physical description of the genotype of an organism. It is the complete spatially-ordered sequence of the four alternative nucleotide bases, adenine, cytosine, guanine and thymine contained in the DNA of the organism's cells. In its most complete form this genotypic description includes not only the DNA in the nucleus of the cell but in certain cellular organelles such as mitochondria or chloroplasts that carry the DNA that is related to their physiological function and that is passed across generational lines. Phenotypic description has also been greatly amplified beyond the external characteristics of individuals. It now includes the gross structure of the various cells of the organism, the physico-chemical description of the various molecules within cells, including their shape and intra-cellular localization, and the network of biochemical reactions in which these molecules participate.

The concepts of phenotype and genotype require a distinction between *types* and *tokens*. Phenotype and genotype are, as the words indicate, types, sets of which individual organisms are members, sets delineated by the complete descriptors of those organisms. Any individual member of these type sets is a token of the type. In addition it is necessary to distinguish the descriptors of organisms from the physical object or behavior that is being described. The concepts phenotype and genotype refer to the descriptors. The physical objects that are being described are the *phenome* and *genome* of an organism, the actual physical material or behavior of some particular individual and its actual DNA molecules. If we take the definitions literally, then no two organisms belong to the same phenotype because there is always some physical difference in morphology, physiology and behavior between any two organisms, even between identical twins, cloned individuals, or cells that result from asexual reproduction. Moreover, except for the results of twinning, cloning or vegetative reproduction, no two individuals belong to the same genotype, as the process of mutation guarantees some variation in their DNA between any two individuals.

The rate of mutation per nucleotide per cell division is of the order of one in a billion, so for mammals, with over a billion nucleotides of DNA making up their genes, even identical twins are likely to have some small difference in total genotype. There is, in addition, an ambiguity in the assignment of an individual to a genotype because mutations occur during the millions of cell divisions that occur in development, so that all cells in the body do not have identical DNA.

In practice phenotypic and genotypic descriptions are not total but partial, referring to some aspect of the genome and phenome that are of particular interest, while ignoring the rest. The partial phenotypic and genotypic descriptions are partial in two senses. First, only some aspect of the morphology or function of the organism is characterized; say the rate of enzymatic splitting of alcohol, a phenotype, and the DNA sequences that code for the enzyme proteins involved in the reaction, a genotype. Second, a decision must be made about the collection of actual partial phenotypes and genotypes that will be regarded as indistinguishable. At the phenotypic level, given the actual variation in enzymatic rates from individual to individual and from time to time, what range of rates will be regarded as belonging to the same phenotypic class? For any phenotypic variable that is continuous, some boundaries of distinguishable phenotypic classes must be laid down in order to make a mapping between phenotype and genotype because genotypes, by their nature, are discrete classes.

The problem of aggregation exists at the genotypic level as well. The relation of DNA differences to differences in the amino acid sequence of a protein is not one-to-one. Because the DNA triplet code is redundant, substitutions of nucleotides in the third position of the triplet often do not change the amino acid that is specified, so for some purposes many different alternative DNA sequences of a gene count as the same genotype. On the other hand, nucleotide substitutions that do not affect the amino acid sequence of a protein may affect the rate of synthesis of that protein and so may, for some questions, be counted as belonging to different genotypes.

The decision about which aspects of total phenotype and genotype are relevant in creating genotypic and phenotypic classes is a deep question of biology that is largely ignored because everything cannot be considered at once. Sometimes biologists act as if the rest of the organism really were constant, sometimes they recognize the variation but claim that it is causally irrelevant to the phenomenon of interest and sometimes they admit its causal relevance but claim that the variation can be treated as background experimental noise that can be averaged out by repeated observations on a large class of individuals. These decisions are usually made either for the sake of observational convenience without a coherent rationale or from a commitment to a reductionist view of the organism that asserts that an understanding of the whole system can be obtained by breaking it down into 'obvious' small components to be studied in isolation.

2. *Relationship Between Genotype and Phenotype*

It is clearly true that organisms do indeed consist of a collection of quasi-independent genotypic and pheno-

typic subsystems, within which there are strong interactions and between which there are very weak interactions. But the boundaries of these genetic and phenotypic subsystems can only be found by actual experiment and differ from function to function. The hand and eye are, as far as is known, developmentally independent systems for which different gene combinations are relevant during embryogenesis, but hand-eye coordination is essential for reaching out and picking up observed objects and so cannot be regarded as phenotypically or genotypically independent in considering this elementary behavior.

The problem for genetics and development is how the genotypic and phenotypic spaces map into each other. It is now certain that Weissman was correct in claiming that the genotype-phenotype mapping is causally asymmetrical. Genotypic differences may cause phenotypic differences during the process of development, but there is no material pathway by which changes in the phenome can result in related changes in the genome. There is no biological inheritance of acquired characters. This causal asymmetry sets the program for developmental, physiological and behavioral genetics which is to discover the mechanisms by which the reading of DNA by the cell machinery is causally efficacious in influencing the phenotypic state of the organism. This is the problem of making the forward mapping of genotype on to phenotype.

There remains, then, the problem of the backward mapping, how the causal knowledge of epigenetic processes can be used to infer the genotype, given the phenotype. Even if a knowledge of the genotype gave perfect information about the phenotype, the inverse would not be true, as discovered by Mendel. The phenomenon of dominance alone makes it impossible to distinguish between heterozygotes and homozygotes from phenotypic information at an arbitrary level. In addition, the same phenotype may appear as a result of changes at different locations in the DNA. There are, for example, more than 200 different nucleotide substitutions known in the genes coding for blood-clotting proteins, any one of which leads to hemophilia. The major difficulty of mapping between genotypes and phenotypes is, however, in the forward direction. The complete DNA sequence of an organism does not contain the information necessary to specify that organism.

3. *Genotype, Phenotype, and Environment*

First, the outcome of development and the ongoing functioning of an individual is the consequence of a unique interaction between the molecules contained in the fertilized egg, including the DNA, and the sequence of environments in which development occurs. Figure 1 shows the result of a cloning experiment in a plant, *Achillea*, grown at three different elevations in the Sierra Nevada. Immature plants were sampled from nature, each plant was cut into three pieces, and one piece was replanted at each of the three elevations. The figure shows the height to which the pieces grew in the different environments, each line connecting the three genotypically identical pieces in the three environments. A line, giving the phenotype of a given genotype as a function of different environments is a *norm of reaction*. The norms of reaction in this case show that there is no genotype with unconditionally better growth and, in fact, there is no way to predict the order of growth of the different genotypes in one environment from their order of growth in another environment. Thus, the phenotype is not only contingent on the environment but in an unpredictable and irregular way.

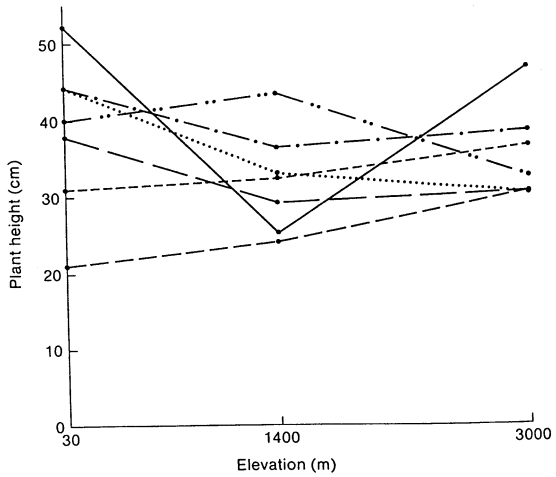

Figure 1
The growth of cloned plants of *Achillea* at three different elevations. The ordinate shows final the plant height and the abscissa the elevation. Each connected line is the pattern of growth for a different cloned genotype. Data of J. D. Clausen, D. Keck and W. W. Hiesey, 1958

Such irregular and particularistic relations between genotype and phenotype in different environments is characteristic of norms of reaction whenever they have been measured for morphological and behavioral characteristics in experimentally accessible organisms. There are, of course, extreme exceptions; genotypes that result in the same phenotype in any environment and environments that are in no way contingent on genotypes for their effect. Eye color variation among humans is entirely a consequence of genotypic differences, uninfluenced by environment, while regional and social class differences in the phonemic structure of speech have no influence of genetic differences. Moreover, major differences between species are a consequence of genetic differences that are manifest in

all environments. There is no environment in which a gorilla could learn to write this article. Within species, however, the vast majority of phenotypic variations in morphology, physiology, and behavior are a consequence of particular contingent relations between genotype and phenotype.

4. Developmental Noise

Second, even a complete specification of the genotype and the environment is insufficient to specify the organism. All cells contain a very small number of copies of biologically active molecules, spatially segregated within the cell, so that the number of various reactions per unit time vary greatly from cell to cell and time to time. Moreover, when cells divide they do not distribute the molecules exactly evenly, except for DNA which has a special machinery for guaranteeing its equal distribution. The consequence of these small molecular numbers is that there is random variation from cell to cell in time to the next cell division, in cell size, in cell shape and in the movements of cells during development. These random events, random at the level of thermal and quantum chemical molecular states, can have considerable effect on the formation and behavior of the whole organism. (Goss and Pecoud 1998)

So-called 'bilateral symmetries' are not, in fact, symmetrical but vary between sides in an irregular way from individual to individual. The fingerprints on the left hand are not identical, although similar to, those on the right. Such *random developmental noise* is a nontrivial source of variation in phenotype. During the development of the central nervous system of a multicellular organism, superimposed on the major regular patterns of neuronal connection are random variations in the growth and connections of specific neuronal pathways, even in organisms like the roundworm *Caenorhabditis* with a very simple central nervous system. It has yet to be determined how much of the behavioral variation among individuals is a consequence of such random variation in the development of the central nervous system, variation that is inborn, but not genetic.

The environmental and chance contingency of the mapping between genotype and phenotype arises from the nature of organisms as physical objects. They are intermediate in size between particles and planets and they are internally functionally heterogeneous. As a result, unlike simple physical systems like the movement of planets around the sun, their functions are not the consequence of a few strong causal forces whose parameters govern the outcome. Organisms are the nexus of a very large number of individually weakly determining forces subject to noise at all levels. The consequence for the understanding of the physical properties of organisms, including their individual and social behavior, is that we cannot expect any simple and universal generalizations like Newton's Laws. That is certainly true for the relation between genotype and phenotype.

See also: Body, Evolution of; Brain, Evolution of; Epigenetic Inheritance; Genetics and Development; Genetics of Complex Traits Through the Life Cycle; Human Evolutionary Genetics; Human Genome Project: Japanese Perspective; Intelligence, Genetics of: Cognitive Abilities

Bibliography

Clausen J D, Keck D, Hiesey W W 1958 *Experimental Studies on the Nature of Species* Vol. 3: *Environmental Responses of Climatic Races of Achillea*. Carnegie Institution of Washington Publication No. 581

Goss P, Pecoud J 1998 Quantitative modeling of stochastic systems in molecular biology using stochastic petri nets. *Proceedings of the National Academy of Science USA* **95**: 6750–5

R. C. Lewontin

Geoarchaeology

Research that uses the techniques, methods, and concepts of the earth sciences to address archaeological problems is called geoarchaeology. In the important early exposition, *Archaeology as Human Ecology: Method and Theory for a Contextual Approach* (1982), Karl Butzer described geoarchaeology as a collection of techniques for contributing substantive data critical to understanding prehistorical and historical human ecosystems (Butzer 1982, p. xi).

There are at least seven major concerns of geoarchaeology that must be integrated into research programs at the design, excavation and analytical stages and include: (a) site survey techniques that use geochemical, electromagnetic and other remote sensing procedures to locate sites or features within a known site; (b) documentation of site-formation processes and the regional context of a site; (c) using a wide array of approaches to differentiate cultural from natural (nonhuman) sediments and features, including disturbance processes by biological, pedological, and geological processes; (d) the development and correlation of intra- and extrasite temporal contexts by relative and/or absolute dating; (e) integration of ancient landforms and biological information for the reconstruction of palaeolandscapes (after Gladfelter 1981, p. 347); (f) reconstructing palaeoenvironments by the identification of pollen, wood charcoal, leaves, and seeds; and (g) the geochemical analysis of artifacts to document interaction between social groups. Exam-

ined here are a few techniques of geoarchaeology that are geared towards: discovering archaeological sites and documenting their internal structure; surveying site formation and disturbance processes; the analysis of soils and sediments; palaeoenvironmental reconstruction and the impact of humans on the landscape; and the physical analysis of archaeological materials.

1. Site Discovery and Definition

Traditionally, locating archaeological sites was accomplished by asking local residents in conjunction with a pedestrian survey; that is, simply walking over the landscape and looking for artifacts and food remains or larger features such as stone house foundations, terraces, or earthworks. In the 1960s with the advent of public archaeology programs—also known as 'contract' archaeology or heritage management studies formulated in response to development projects—archaeological surveys became more systematic, utilizing sophisticated sampling procedures, predictive modeling, and techniques of remote sensing, or archaeological prospecting as it is called in Europe.

1.1 Remote Sensing or Archaeological Prospecting

Archaeological prospecting 'is considered as an exercise in the deliberate planned measurement and evaluation of physical properties which lead to the detection and mapping of archaeological sites' (Scollar et al. 1990, p. 27). Since all surface indications of many archaeological sites have been obliterated by modern urban encroachment or been buried under soils, excavation has been the technique of choice for locating subsurface cultural deposits. Remote sensing techniques are nondestructive and include aerial photography, and resistivity, magnetic, electromagnetic, and thermal prospecting (see Scollar et al. 1990 for a detailed explanation of these techniques).

1.2 Aerial Photography

The oldest form of remote sensing, aerial photography is used in archaeology primarily for discovering field monuments, whether earthen or stone constructions. A huge corpus of aerial photographs exist some of which were taken during various wars for reconnaissance purposes, others for urban planning, to record changing shorelines, and to document vegetation patterns. However, aerial photos made exclusively for archaeological site discovery are the most useful since they are taken at the appropriate scale, angle, and time of day for enhancing surface features. Prehistoric agricultural field systems in areas such as Mexico and the Pacific Islands, consist of low terraces and earthen mounds and embankments which are readily mapped from aerial photographs, saving countless hours of field survey and mapping.

1.3 Other Techniques of Remote Sensing

Once archaeological sites are located, identifying internal structures (intrasite variability) can be accomplished with other techniques of remote sensing including resistivity prospecting, where subsurface anomalies (such as burial pits, compacted ancient floors, or buried house foundations) are mapped by identifying the variation in an electric charge as it passes through the ground (Scollar et al. 1990, pp. 307–421). Magnetic prospecting can locate the shape, size and depth of subsurface anomalies that can then be mapped. Thermal prospecting uses variations in surface temperatures as an indication of subsurface (often cultural) structures.

1.4 Coring Archaeological Sites

Identifying subsurface archaeological site boundaries, internal features, and differential distributions of artifacts and faunal material can be accomplished by coring (Stein 1986), especially in conjunction with the analysis of sediment chemistry. Small cores or bucket augers (about 10 cm in diameter) are used to collect sediment samples for the analysis of cultural content and to differentiate cultural from noncultural layers. Since organic matter, carbonates (e.g., marine shells), and phosphates (deposited in the form of urine, excrement and organic refuse) can increase relative to the intensity of human occupation, these residues are good indicators of past site use (Sjöberg 1976) and their distribution is used to identify subsurface cultural deposits with a minimum of site disturbance.

2. Site Formation Processes

The archaeological record is not static, but changes in response to a number of processes—both cultural and environmental (or noncultural). Once deposited in, or on, the surface of an archaeological site, cultural material (artifacts and food remains) and noncultural matter, such as pollen grains and terrestrial sediments, can shift their position or, in some cases, be removed completely. Since archaeological interpretations gain their strength from identifying and analyzing the vertical and horizontal distributions of material within and between sites, it is vital to understand how the positions of these materials may have changed since they were originally deposited—this is the challenge of site formation research defined as the processes, both natural and cultural, that build, disturb, redeposit, and destroy archaeological sites (Schiffer 1987).

2.1 Nonhuman Disturbance Processes

Since the soils and sediments that make up an archaeological site are part of a dynamic open system, archaeological sites are best viewed in a regional context. It is at this level that soil development is best understood because it is influenced by such large-scale processes as precipitation and temperature and erosional agents—including wind and surface water (e.g., rivers, marine shore waves, and sheet wash). Wood and Johnson (1978) have outlined a number of disturbance processes that affect archaeological deposits. Some of the more important processes include faunalturbation, or the mixing of deposits by animals, especially the burrowing kind (rodents, crabs, insects, earthworms, and birds). Digging burrows not only brings buried material to the surface, but open holes are traps for surface material to accumulate beneath the ground. The mixing of deposits by plants that occurs during root growth and decay is known as floralturbation. Uprooted trees can also bring material to the surface, often from a meter or more below the ground.

Discrete layers within archaeological sites can be homogenized by alternating freezing and thawing episodes which sort sediments by size. This is a problem with sites in the higher latitudes that consist of clay soils that swell and contract in response to precipitation and temperature. Gravity can cause mixing and movement of soil and rock debris downslope, as well as settling or subsidence of relatively large areas. Aeroturbation occurs when wind winnows fine sediments from deposits leaving a 'pavement' of artifacts. This is a significant problem in areas with active sand dune systems. Despite the range of disturbance processes that operate on a regional and site-specific basis, most cultural deposits maintain some order of integrity and, after the site formation processes are understood, valid interpretations are possible.

2.2 Human Disturbance Processes

In addition to natural or environmental site formation processes, people intentionally contribute to the formation of archaeological deposits by bringing to a site raw materials such as fine-grained rock to fashion into tools, finished artifacts of bone, shell, wood and stone, cooking fuel, building materials, and food. Many of these items are discarded on site as waste. Even sand grains adhering to feet, after countless thousands of trips to a site, can add significantly to the cultural matrix. Digging pits for earth ovens and trash disposal, small holes excavated for wall or fence posts, or trenches dug for building foundations can displace subsurface cultural deposits during later periods of occupation. Accurate interpretations of the archaeological record can only be made after understanding both noncultural and cultural site formation processes.

3. Definition of Soils and Sediments

Soils and sediments figure prominently in discussions of site formation processes, so it is essential to know how these terms differ. Soils develop *in situ* at or near the ground surface through the various processes of weathering in conjunction with plant growth and decay, other biological elements, time, climate (temperature and moisture), and local topography. Sediments are particulate matter that has been transported from one location to another by one or more processes such as wind, water, or people.

3.1 Principles of Sedimentation

All particles, including artifacts and food remains, found in archaeological deposits can be viewed as sediments (Stein 1987, p. 339) and, because of this, the four principles of sedimentation (Shackley 1975) apply equally to natural and cultural sediments. First, the source of sediments found at an archaeological site can be unconsolidated particles of bedrock as well as humanly-transported stone tools, shells, bones or building stones, to name but a few examples. Second, the transport history of sediments is usually determined by textural and compositional analysis in that biological agents (people, plants, and animals), wind, and water move sediments according to certain physical constraints. For example, wind normally transports evenly-sorted sand-sized sediments, while water—in the form of rivers, sea waves, and glaciers—can move anything from clay particles suspended in its volume to huge boulders that are pushed under enormous power. Third, the environment of deposition, in this example, is the archaeological site, but offsite examples can include sedimentary basins, aggrading shorelines, sand dunes, and till and eradict boulders deposited from melting glaciers. And last, postdepositional alterations, discussed above, are caused by people, other animals, plants, and geologic erosional processes.

3.2 Sampling Soils and Sediments

Soils and sediment samples are usually taken from a profile in an erosional exposure or from the sidewall of an excavation unit or trench. Beginning from the bottom up, samples are taken from separate stratigraphic layers or in a continuous column. The sample volume varies according to the kinds of analyses undertaken and the size of sediments. For example, small samples of about 20 grams are needed to analyze the texture of clays and silts, while several kilograms are necessary for analyzing cultural content or, if

sediments are gravel, to cobble size. The key is to collect a representative sample of the sediment.

3.3 Analytical Techniques

Among other things, particle size analysis permits the detection of the agent of deposition (e.g., wind, river, sea) and the environment of deposition such as beach, flood plain, and dune (Shackley 1975, p. 87). As mentioned above, well-sorted sand is often wind-deposited and certain agents of deposition have particular sediment signatures. Mineral identification of individual grains can suggest a likely source of the sediments, especially if the regional geology is well known. Grain shape (roundness and flatness) can indicate the transport agent, depositional environment, and history. Grains can be, for example, platy or flat; angular, or granular. Roundness refers to the general grain surface curvature and is more a function of mineral composition (certain minerals are harder or softer than others and respond to friction differently), depositional history, and final depositional environment (Shackley 1975, p. 46). Grain surface textures, examined under high-power magnification, can be: unworn and angular, suggesting a fresh or minimally-transported history; rounded and glossy, the result of transport by running water; or matt-surfaced grains, the result of wind transport. A good review of the analytical methods of sediment analysis can be found in Shackley (1975).

4. Palaeoenvironmental Reconstruction and Human Impact

The relationship of people to the landscape (and here I use 'landscape' to mean the physical and biotic environment) is of worldwide interest, not only to archaeologists, but to biologists and politicians alike, to name just a few. How humans have historically managed their resources—whether successfully or with tragic consequences—is of utmost importance to contemporary society—especially as the world becomes increasingly one 'global village.' Archaeology is better positioned than any other discipline to understand the historical details of human–landscape relationships over tens of thousands of years and across an enormously varied set of environmental conditions that posed opportunities and problems for our ancestors.

To document and understand the variation in human–landscape interactions, the analysis of soils and sediments is especially useful when used in conjunction with the differing distribution of floral remains (see Hastorf and Popper 1988 for a review of analytical methods). The analysis of faunal material is also of importance here and is addressed in *Zooarchaeology*. Whether of carbonized wood fragments, pollen, opal phytoliths (minute silica bodies formed in the cells of living plants that can be diagnostic of certain species), or leaves, plant remains are often indicative of vegetation on or near an archaeological site and, in the case of wind-dispersed pollen, can reveal regional floral patterns. Recovered from dated stratigraphic contexts, vegetation history can be charted over time—before, during, and after human settlement. In essence, archaeological sites hold an historical record of local and regional palaeoenvironmental change, often in response to human modifications to the landscape.

5. Physical Analysis of Archaeological Materials

A growing area of interest in geoarchaeology is the elemental analysis of artifacts such as ceramics and stone tools so links can be drawn from the geological source or, in rare occasions, the site of production, to the place of deposition or archaeological site. The emphasis is on establishing lines of interaction or communication between or within social groups. The distribution of exotic artifacts across space and through time (that is, artifacts recovered from dated contexts or layers) is thought to reflect changing aspects of social organization—a topic of great interest to archaeologists.

5.1 Physical Analysis of Ceramics and Stone Tools

Ceramics are manufactured from clay, water, and nonplastic inclusions (temper). Pots were often decorated with an exterior clay slip, incised or stamped with various intricate patterns. Of interest to interaction or 'sourcing' studies is the elemental composition of the clay and temper. If the latter consists of crushed rock or terrigenous (land-based) sands, then chemical analysis of the temper may isolate its source. Stone tools, manufactured from obsidian, basalt, chert, or other fine-grained rocks, also can be geochemically analyzed to determine the artifact's geological source. Economically-important rock sources were generally well known in prehistory and had a limited distribution, thus facilitating sourcing studies.

There are many useful techniques for determining the elemental composition of ceramics and stone tools. (See Parkes 1986 for a description of various techniques.) These include, for example, x-ray fluorescence (XRF) and lead isotope analysis at the newer and more sophisticated end of the techniques (Weisler 1997), to thin-section petrography which has been used in geologic investigations for over a century.

5.2 X-ray Fluorescence Analysis

Simply stated, XRF analysis consists of bombarding a specimen with primary x-rays. The sample emits fluorescent x-rays that are indicative of the kind and

frequency of its elemental composition (oxides, such as iron or silica, are measured in weight percent, while elements are counted in parts-per-million or ppm). Dedicated computers are used for data analysis, while statistical clustering programs can aid in grouping artifacts with possible sources. Once sources are adequately characterized in terms of their geochemical variability, then unknowns or artifacts can be more confidently assigned to a specific source.

5.3 Thin-section Petrography

Thin-section petrography is used extensively today in ceramic sourcing studies when the mineral identification of temper is desired. Samples measuring up to 25 mm by 40 mm are cut from ceramic sherds and ground thin enough to pass light through them. Placed under magnification, minerals (such as olivine and feldspar) are identified and texture measured. Within a geologically well-known region, temper composition can be assigned to a particular river drainage (in the example of placer sands), a marine beach, or a larger region such as a group of hills or a mountain range.

Geoarchaeology uses the techniques, methods, and concepts of the physical sciences to address archaeological questions. In this short article, I have only touched on a few of the techniques and methods applied to research domains that are of worldwide interest. This subdiscipline of archaeology is continually expanding the range of techniques and their applications to innovative questions. The successful melding of interdisciplinary research typical of geoarchaeological pursuits continues to have a bright future.

See also: Bioarchaeology; Ceramics in Archaeology; Geographic Information Systems; Remote Sensing and Geographic Information Systems Analysis; Zooarchaeology

Bibliography

Butzer K W 1982 *Archaeology as Human Ecology: Method and Theory for a Contextual Approach*. Cambridge University Press, Cambridge, UK

Gladfelter B G 1981 Developments and directions in geoarchaeology. *Advances in Archaeological Method and Theory* **4**: 343–64

Hastorf C A, Popper V S 1988 *Current Paleoethnobotany: Analytical Methods and Cultural Interpretations of Archaeological Plant Remains*. University of Chicago Press, Chicago

Parkes P 1986 *Current Scientific Techniques in Archaeology*. Croom Helm, London

Schiffer M B 1987 *Formation Processes of the Archaeological Record*. University of New Mexico Press, Albuquerque, NM

Scollar I, Tabbagh A, Hesse A, Herzog I 1990 *Archaeological Prospecting and Remote Sensing*. Cambridge University Press, Cambridge, UK

Shackley M L 1975 *Archaeological Sediments: A Survey of Analytical Methods*. Wiley, New York

Sjöberg A 1976 Phosphate analysis of anthropic soils. *Journal of Field Archaeology* **3**: 447–54

Stein J K 1986 Coring archaeological sites. *American Antiquity* **51**: 505–27

Stein J K 1987 Deposits for archaeologists. *Advances in Archaeological Method and Theory* **11**: 337–95

Weisler M I (ed.) 1997 *Prehistoric Long-distance Interaction in Oceania: An Interdisciplinary Approach*. New Zealand Archaeological Association Monograph 21, Auckland, New Zealand

Wood W R, Johnson D L 1978 A survey of disturbance processes in archaeological site formation. *Advances in Archaeological Method and Theory* **1**: 315–81

M. I. Weisler

Geodemographics

1. Introduction

A geodemographic database is a collection of spatially referenced data about individuals or groups, and geodemographics describes a rapidly growing segment of the marketing industry that collects massive amounts of data on consumer characteristics and behavior, constructs statistical models of consumer identity, and maps and analyzes distributions of market-related activity. It is used primarily to identify, or 'target,' potential customers, or 'prospects,' and in business applications such as retail site evaluation, trade area analysis, market share analysis, sales territory management and micromarketing. It can also, however, be used by nonprofit institutions to enhance fundraising, by governments for administrative purposes, such as school districting, and by political campaigns to identify voters susceptible to customized political messages. The sophisticated techniques of geodemographics depend on two simple assumptions: that 'you are what you buy' and 'you are where you live.'

2. The Development of Geodemographics

The collection of customer information and target marketing have been around since the late nineteenth century, and customer segmentation and micromarketing were introduced in the 1950s. It was not until the 1970s, however, that theoretical, technological and institutional innovations made possible the maintenance of electronic databases on consumer behavior, the statistical analysis of consumer identity, and precise geographical location of consumers upon which geodemographics depends. Its invention is attributed to Jonathan Robbin, a sociologist who, in the early 1970s, saw entrepreneurial potential in the

recently introduced Zone Improvement Plan (ZIP) of the US Postal Service. First, he aggregated demographic data for the 240,000 block groups of the Bureau of the Census into the 36,000 postal delivery areas. Then he performed multivariate analyses of this data to produce 40 lifestyle categories describing consumer potential—hence he called the system PRIZM (Potential Rating Index for ZIP Markets). This innovation allowed marketers to target potential customers through heavily discounted bulk direct mail campaigns.

Today, publicly produced census data are still a vital component of geodemographics, and vendors still adopt clever acronyms like ACORN (A Classification of Residential Neighborhoods developed by CACI Marketing Systems, UK) and PSYTE (from CompuSearch, Canada) to describe proprietary segmentation systems. However, the US Bureau of the Census itself provides ZIP level data as a result of lobbying and financial assistance from the geodemographic industry, and its computerized spatial database called TIGER (for Topologically Integrated Geographic Encoding and Referencing) is the basis of the geography of most geodemographic systems. In Canada and the UK, there is also close cooperation between industry and government—in the UK, for example, Pinpoint Ltd. and the General Post Office have digitized every address that receives mail. In these countries, geodemographic databases incorporate diverse sources of data from official records, reported consumer behavior, lifestyle preferences and even psychological profiling ('psychographics'). Elsewhere, the industry has been slower to develop, largely owing to limitations in quantity and quality of public and private data, but it is truly international, with business marketing publications and software solutions available for many individual countries and most regions of the world. A cross-cultural clustering of consumers is already available for Europe.

While a host of specialist corporations and consultants offer individual services, a series of licensing agreements, acquisitions and mergers have resulted in national, and in some cases, transnational 'information conglomerates' (Curry 1993, p. 27) that purvey complex, integrated geodemographic systems. Such systems typically include three principal components: electronic databases composed of public and private, individual and aggregate information; segmentation routines that identify consumer types through cluster analyses of behavioral and psychographic data; and a geographic information system that provides the means for spatial analysis and mapping.

3. *Electronic Consumer Databases*

The geodemographics industry accumulates data from numerous sources to create its Marketing Customer Information Files: purchasing proprietary information on customers from corporations, such as point of purchase surveys, warranty registration questionnaires, customer-preference cards and satisfaction surveys, and from government agencies; purchasing customized data from syndicated national consumer panel surveys, media surveys, opinion polls and credit bureaus; obtaining public transaction records of institutions such as banks, insurance companies and health services, which are kept in compliance with government regulations; and 'mining' public records on births, deaths, marriages, property transactions, crime, and mailing address changes. The scale is enormous, encompassing the vast majority of all of US households, and what were once databases are now 'data warehouses' with more than one trillion entries. Specialist firms and software programs provide means to sift and sort data to eliminate redundancies and match entries, particularly in terms of street addresses, and consumer information is now updated instantly and can be downloaded on line from most of the geodemographic vendors.

Such data are often collected and sold with only the implicit consent of consumers, who are unaware of the further uses to which data gathered in one transactional context might later be put. As a result of informational campaigns and legal actions, consumers are becoming increasingly wary of invasions of privacy and suspicious of 'consumer espionage' (Larson 1992, p. 5), and the industry has adopted voluntary controls. Even individuals who explicitly opt out when presented the option, however, are likely represented in official databases, and their consumer 'profile' can be inferred accordingly. Privacy is a notoriously tricky issue, and surveillance may be a necessary quid pro quo of participation in capitalist society (Robbin, cited in Larson 1992, p. 55). Perhaps public concern should be not so much about specific instruments of surveillance technology nor their applications to particular individuals, but the manner in which social subjects are constituted and represented in the discourses of social and spatial sciences (Foucault 1979). Thus, geodemographics is replete with 'strategic' metaphors of military operations, sexual conquest, and psychological manipulation, based upon reduction of identity and social difference to consumer profiles and residential location through its segmentation schemes and 'market mapping' (Goss 1994).

4. *Consumer Profiling*

Contemporary segmentation schemes go well beyond the abstract demographic variables of the census to offer 'Character: not just characteristics,' and this system, called Focus, for example, offers 'multidimensional portraits of real people,' revealing 'lifestyle and mindset' and promising marketers will 'see how customers view themselves,' and 'meet their customers on a first name basis' (National Demographics and Lifestyles 1993, p. 2). Segmentation schemes draw

upon information on brand preferences, media habits, credit rating, lifestyle and values, to divide neighborhoods in the US, typically into 30–40 categories such as 'Pools and Patios,' 'Shotguns and Pickups,' 'Fur and Station Wagons,' and 'Hard Scrabble' (PRIZM) or 'Cautious Young Couples,' 'Sustaining Ethnic Families,' and 'Young Accumulators' (MicroVision). Marketers 'get acquainted' with these potential customers through visual scenarios representing consumers engaged in a typical activities in front of their home and/or detailed imaginary vignettes describing their typical names and everyday lives—particularly their consumption practices. Although analysts argue their schemes imply no social judgments, the rankings, weightings and premiums charged for names of affluent consumers, and description of segments give them away: to the marketing industry, social identity and value is lifestyle, and 'you are what you buy' (Piirto 1991, p. 23).

5. *Geographic Information Systems*

The 'geo' of geodemographics is essential for several reasons: first, because the map transforms complex data into simple 'actionable' visual images—red typically represents a 'hot' market—ideal for busy executive officers; second, because addresses of households provide a convenient way to match database records; third, because addresses are means by which consumers are reached with marketing messages or products, whether delivered by mail, conventional media or retail outlet; fourth, because a range of analyses and transformations can be applied to spatial data to create 'custom' geographies such as retail trade areas, sales territories and media footprints; and finally because marketers infer characteristics of individual consumers from aggregate data and substitute missing values in their databases based on residential location. This last function is important because in some forms of data (census and vehicle licensing, for example) individual records are not available for confidentiality reasons, and where records are missing, the costs of collecting such data are generally prohibitive.

The importance of the geographical basis of marketing information is attested by the lengths to which the industry goes to infer residential location even for users of the World Wide Web, constructing user profiles from 'search engine queries; online shopping patterns; requests for localized information, such as movies, concerts, lectures and stores; and interested in regional websites such as yourcity.com' (Melville 1999, p. 34). Inference from 'neighboring' records is based on the assumption, according to Jonathan Robbin, that 'humans group themselves into natural areas … They create or choose established neighborhoods that conform to their life-style of the moment' (cited in Burnham 1983, p. 92). Not surprisingly, Claritas' promotional material similarly claims that it is 'based on the fundamental sociological truism that "birds of a feather flock together"' (Claritas/National Planning Data Corporation (NPDC) 1992). The geographical basis of consumer profiling is also a result of the so-called neighborhood effect, such that, for example, 'when one moves into a suburban neighborhood in which everyone has a gas barbecue there is a strong motivation to buy one' (Lea and Heumann 1999, p. 26). There is, of course, an undeniable spatiality to everyday life, but to infer individual characteristics from those of an area is to fall for the ecological fallacy (to assume that individual characteristics can be inferred from group characteristics, which is only true if the population is homogeneous) and the question is how much such an effect is natural or intrinsically motivated, and how much it is the product of the marketing enterprise itself.

6. *Conclusion*

Geodemographics is particularly hype-ridden and the success of its applications limited, for example, to very modest 'lifts' in customer response by 5–10 percent in direct mail campaigns, even if this represents considerable savings in a large-scale operation. It is, nevertheless, the fastest growing segment of the information industry, presumably because of its promise of intimate and actionable knowledge about consumers. Geodemographics first reduces the subjectivity of individuals to objective and quantifiable characteristics, then reconstructs composite 'soap opera' identities (Goss 1995, p. 187), perhaps more consistent with marketers prejudices than social reality—hence one geodemographic system promises that 'as you get to know [our cohorts], you'll discover these are all people you know: your friends, neighbors, relatives, and—most importantly—your customers' (National Demographics and Lifestyles 1993, p. 2). There is a second, more sinister, self-fulfilling promise, however, in that the definition and targeting of consumers based on residential location may effect redlining of consumption, such that marketing campaigns and retail services are provided only to addresses—with known names, inferred names, or merely 'Current Resident'—with the 'right' statistical and behavioral profiles. This is not to argue that there are not cleavages in consumption based on class and residential location, but if, for example, magazine issues are differentiated in terms of advertising content for subscribers in different ZIP codes, as some already are in the US, consumption of these goods and services in turn will reproduce, and perhaps actually produce, further segmentation and segregation of society. Such is the instrumental reason behind the hype of geodemographics.

See also: Economic Geography; Market Areas; Market Research

Bibliography

Burnham D 1983 *The Rise of the Computer State*, 1st edn. Random House, New York

Claritas/NPDC 1992 *Prizm Clusters*. Brochure, New York

Curry D J 1993 *The New Marketing Research Systems: How to Use Strategic Database Information for Better Marketing Decisions*. Wiley, New York

Foucault M 1979 *Discipline and Punish: The Birth of the Prison*. Vintage Books, New York

Goss J D 1994 Marketing the new marketing: the strategic discourse of geodemographic information systems. In: Pickles J (ed.) *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford Press, New York

Goss J D 1995 'We know where you are and we know where you live': the instrumental rationality of geodemographic information systems. *Economic Geography*. **71**: 171–98

Larson E 1992 *The Naked Consumer: How our Private Lives Become Public Commodities*, 1st edn. Holt, New York

Lea T, Heumann D 1999 Key in on variables to improve databases. *Business Geographics*. **7**: 24–6

Melville R 1999 Online marketing goes geographic. *Business Geographics*. **7**: 33–5

National Demographics and Lifestyles 1993 *Character. Not Just Characteristics*. Brochure, Denver, CO

National Demographics and Lifestyles 1993 *NDL FOCUS: Knowing your Customers*. Brochure, Denver, CO

Piirto R 1991 *Beyond Mind Games: The Marketing Power of Psychographics*. American Demographics, Ithaca, NY

J. Goss

Geographic Education

Despite the commonly used name, 'geographic' education is more appropriately thought of as 'geography' education. The distinction is significant. The use of 'geographic' as a modifier might suggest a special type of educational theory or practice. For example, it could imply distinctive and particular pedagogical procedures and therefore unique methods of training for teachers of geography. While the interaction between content and pedagogy in any domain of knowledge requires discipline-specific understandings and skills, there is nothing inherent in the nature of geography that warrants a special type of education, fundamentally different in nature from that practiced in disciplines such as physics, psychology, and history. Geography education is about the teaching and learning of geography (Bednarz et al. in press).

Both geography education and geographic education must be distinguished from the geography of education. As a major sector of the modern economy, the education system employs significant numbers of people and involves substantial capital investments in terms of school buildings, university and college campuses, buses, computers, textbooks, etc. The geography of education is concerned with understanding the spatial organization of the education system. It encompasses, for example, the relationships between patterns of tax revenues and levels of education expenditures, the diffusion of colleges and universities, the optimal design of bus routes to link students with schools, and the equity and efficiency of the provision of education to different socioeconomic and/or racial and ethnic groups. The focus is on understanding the spatial manifestation of the K-22 education system as part of the infrastructure of society.

Although an understanding of the educational infrastructure is not part of geography education as such, the state does have a major impact on the nature of geography education. In the USA, for example, the fundamental belief is that education is best controlled at a local, sub-state level. In Britain, and in France to an even greater degree, education control is more strongly vested in the national government. The ideological and structural differences between decentralized and centralized systems affect the nature of the curriculum, the training of teachers, the selection of educational materials, and the relative emphases on geography education.

1. The Nature and Process of Geography Education

Geography education is concerned with fostering an understanding of the domain of geographic knowledge (its subject matter content, skills, and perspectives). The intent is not necessarily to produce geographers as such but to provide the essential geographic knowledge and skills for an educated person.

1.1 The Nature of Geography

As the science of space and place on the earth's surface, geography seeks to describe and explain the distribution, pattern, and nature of places and regions as seen at scales ranging from the local through to the global. The understanding of places and regions is set in multiple contexts: spatial, environmental, and temporal. The character of places and regions is seen as shaped by the interrelationships between physical phenomena, such as climate, terrain, the hydrologic cycle, biosystems, etc., and human phenomena, such as demographic characteristics, economic functioning, culture, politics, urbanization, etc. The shaping process is understood in a variety of ways, in terms of:

(a) the coalescence of physical and human phenomena to create the particular character of a place (e.g., local topography, microclimates, level of urbanization, ethnic composition, etc.);

(b) the impacts of global environmental and economic changes on local places (e.g., economic restructuring in labor-intensive industries, with the loss of local employment to cheaper 'offshore' production facilities); and

(c) the interdependencies between places on local, regional, and global scales (e.g., chemicals in runoff from farms in upstream areas polluting regional water tables and watersheds).

Therefore, the focus of modern geography is on understanding the evolving character and organization of places and regions on the earth's surface. Geography's approach is distinguished by (a) a sensitivity to context, (b) an appreciation of interconnections between spatial patterns and processes, (c) an emphasis on understanding systemic interrelationships at different spatial scales, and (d) an awareness of the reciprocal relations between physical and human systems.

The skills necessary for geographic understanding are of three types. There are general problem-solving or critical thinking skills which ask and answer geographic questions by means of acquiring, organizing, and analyzing geographic information. There is the use of broadly applicable methods such as spatial analysis (quantitative and qualitative), geographic information systems (GIS), field methods, and cartography. Finally, there is the application of discipline-specific technologies such as satellite image interpretation, computer-based geographic visualization, and global positioning systems (GPS). Perspectives in geography start with the essential spatial and ecological perspectives and have been extended to include, for example, feminist, historical, economic, social and critical theoretic, and humanist perspectives.

1.2 The Process of Geography Education

The process of geography education involves the teaching and learning of geography in formal (e.g., preschool, school, college) and informal (e.g., familial, societal, workplace) contexts. Ideally, the process is lifelong in the context of an individual person—preschool, K-12, and perhaps beyond into higher education (13–16 or even 17–22) and the workplace—and continuous in the context of developments in the discipline of geography. In practice, however, there have been significant historical changes in:

(a) the ages at which geography education has been offered;

(b) its links to other components of the formal education system (e.g., geography as a standalone or as an infused subject);

(c) the role of the state vs. informal organizations in developing materials, training teachers, and setting the educational agenda; and

(d) geography's role in facilitating societal imperatives (e.g., support for exploration and imperialism, mass education, and the development of workforce skills). The field of geography itself has also changed its emphasis on and approaches to geographic education.

2. Approaching the History of Geography Education

Livingstone (1992) characterized the history of geography as a 'contested enterprise.' His strategy for 'telling geography's story' was to focus on the 'intellectual and social context within which geographical knowledge was produced' (Livingstone 1992, p. 23). If we affix 're' to 'produced,' then we can also tell the story of geography education.

While the production and reproduction of knowledge are necessarily linked, they do not move in some elegant *pas de deux*. The field of geography—the social organizations by which domain knowledge is generated, preserved, and transmitted—is complex. As a discipline, geography has an existence as an academic and as a popular subject, the latter exemplified by the activities of the National Geographic Society, which at first were confined to the USA but now have an increasingly international reach. The two parts of the discipline are characterized by approaches, agendas, and audiences that are more often competitive rather than complementary. Moreover, internal, disciplinary views and external, societal views of the purpose and value of geography are different and sometimes at odds with one another. Superimposed on these differences are the valuations placed by professional geographers on the roles for geography education.

3. Geography Education as a Contested Enterprise

The history of geography education is also a contested enterprise. It has been shaped by complex interactions among a variety of factors:

(a) individual and societal valuations of geographic knowledge;

(b) the current state and past history of geography as a discipline (in terms of dominant approaches, concepts, models, and theories);

(c) the state of the formal education system (in terms of societal goals, physical infrastructure, resources, teacher training, textbooks, etc.); and

(d) the state of pedagogical theory and practice in education in general, and geography in particular.

While all of these factors vary across time and space, they coalesce into changing national approaches to geography education.

4. The Rationales for Geography Education

As with many disciplines, the rationale for the importance of geographic knowledge has been intertwined with the definition of an educated person. That definition has been subject to changes in the nature of

geographic understanding, and to the shifting emphases of the larger projects of society.

While it is impossible to unearth the roots of geography education, we can trace the emergence of some of the rationales for expecting people to understand geography. In late-Medieval Europe, for example, one grounding of formal education was in the teaching and learning of the seven sciences of the *artes liberalis*: the *quadrivium* or advanced division (arithmetic, geometry, music, and astronomy) and the *trivium* or lower division (grammar, logic, and rhetoric).

5. Two Geographic Themes

Geography, as in knowledge of the location and character of places and the creation and use of maps, fell partially within the rubrics of geometry and astronomy. Place knowledge was an expected and necessary part of the knowledge base of an educated person and map use was a valuable practical skill in travel, navigation, surveying, business, warfare, and diplomacy. As both foundational knowledge and applied skill, geography therefore had elements of what we would today recognize as a liberal arts subject and as a science discipline, thus accounting for the discipline's varied placement within the college and faculty structures of universities in North America.

The two themes of foundational knowledge and applied skill are interwoven throughout the history of geography education in all countries. On the one hand, geography education is seen as essential to the project of creating an educated person, whether that project is couched in terms of (a) a broad liberal education, (b) citizenship education (in local, patriotic and national, or world contexts as expressed in the slogan 'think globally, act locally'), or (c) geographic literacy (e.g., reading and understanding the contents of a newspaper). Participation in local community affairs, voting, and support for national and increasingly international policies all depend upon a level of geographic literacy. On the other hand, geography education is seen as providing essential skills for operating in society in general (e.g., map-reading skills) and for participating in the workforce in particular (e.g., using geographic information systems). Careers in planning, business, and government agencies all depend upon significant levels of geographic skill.

5.1 Geography as Foundational Knowledge

The foundational knowledge approach to geography education has many variants. Geographic knowledge is seen as a way of counteracting and overcoming the natural and/or socially induced human tendencies towards egocentrism, parochialism, and ethnocentrism. Knowledge of the diversity of patterns of human occupancy of the earth's surface reduces the chances of privileging one's own place in the world and increases the appreciation of cultural diversity and human creativity. Similarly, geographic knowledge underpins an understanding of physical environments, especially important as those environments are under threat from human-induced changes at local and global level. It can reinforce an understanding of the interdependency between human and physical systems. Geographic knowledge is seen as important for the teaching of other subjects, especially history, sociology, anthropology, economics, business, and political science. Without an appreciation of the intersections between space and place and between environment and society, the subject matter of those other disciplines cannot be appropriately grounded and understood. And finally, geographic knowledge need not be seen as exclusively instrumental or utilitarian: '(t)he power and beauty of geography lie in seeing, understanding, and appreciating the web of relationships among people, places, and environments' (Geography Education Standards Project 1994, p. 3).

5.2 Geography as an Applied Skill

The grounding of geography education in skills also has multiple bases. The science, art, and craft of cartography has been fundamental to many social projects, including surveying, documenting land ownership systems, exploration and navigation, colonization and military conquest, regional and local planning, and environmental science. As technologies for data gathering (e.g., satellite remote sensing), ensuring data accuracy (e.g., global positioning systems), data processing (e.g., GIS), and displaying geospatial data (e.g, geo-visualization tools) have become pervasive, so too have been the needs for geographically educated people to produce and interpret an increasing range of graphic displays. Those technologies and their products are central to the daily operations of government (local and national), the military, business and logistics, international politics, and so on. For example, knowledge of the spatial structure of national and world economies, together with the tools of spatial analysis, are seen as essential for individual participation in, and national success in, the business of the increasingly interdependent global economy.

Geography education has, therefore, been based on an intermixing of four principal rationales:

(a) the existential value of understanding one's place in the world;

(b) the ethical imperative of appreciating the fragile human and physical systems upon which we depend;

(c) the intellectual power of understanding the interplay between two pairs of ideas—space and place, and environment and society—in shaping the patterns of the earth's surface; and

(d) the utilitarian motivation of using geography to solve individual and societal problems. These rationales have motivated the process of geography education and have been differentially emphasized during the history of geography education (Douglass 1998).

6. Geography Education: a History

While there had been a strong classical tradition of geography in the eastern Mediterranean as far back as 2000 BC, the efflorescence of geography was contemporaneous with the so-called Age of Exploration and the origins of the Scientific Revolution beginning in fifteenth-century Europe. Built on the maps and texts surviving from the Greek, Egyptian, Roman, and Arab cultures, geography was both a statement about the world and a way of recording the new information about diverse places. Maps were both tools for navigation and means of recording the results of travels. Texts captured descriptions of the location and properties of places encountered.

6.1 Varenius and the Two Geographies

These two roles of maps and texts are a convenient starting point in an account of the history of geography education. In 1649, Bernhardus Varenius (Bernhard Varens), a German doctor living in the Netherlands, published *Descriptio Regni Japoniae et Siam*, followed in 1650 by his *Geographia Generalis*. In the latter work Varenius systematized an understanding of the logic of geography, distinguishing between general and special geography. Although the terms were not original to Varenius, his statement of the distinction was crucial to the development of geography as a discipline in general and to the history of geography education in particular.

General (or universal) geography was explanatory in nature, seeking to understand the patterns and processes of the world via general (or natural) laws. Echoing the links of geography to geometry and astronomy in the *quadrivium*, general geography was linked to the 'mixed' (or applied) mathematics. Its focus on the terrestrial globe was paralleled by astronomy's focus on the celestial globe. General geography sought to understand global patterns of climate, topography, hydrography, etc., by means of forces and mechanisms (the laws of motion, heat, erosion, etc.). It moved from global explanations to accounts of regional variations, emphasized physical over human phenomena, and incorporated both geodesy and cartography.

Special geography was descriptive in nature, seeking to depict the characteristics of particular places. Although Varenius did not offer a detailed description of special geography in *Geographia Generalis*, his outline description matched the contents of his earlier *Descriptio Regni Japoniae et Siam*. While the basis for general geography was demonstration from general laws, special geography derived from experience, systematized in terms of the terrestrial (physical) and human properties of places.

While there is a necessary link between special and general geography, there is also a recognition that general geography is more challenging, more intellectually demanding, and therefore in a sense, more advanced than special geography. General geography seats the discipline squarely within the tradition of natural philosophy and eventually the mathematical and empirical sciences. Varneius' work was influential throughout Europe for over two centuries (used by Newton in England and Kant in Germany, for example) and in the USA (Warntz 1964, 1982).

For geography education, Varenius' distinction between general and special geography establishes the two tracks along which geographic teaching and learning developed. It also helps to account for the trajectory of success, failure, and eventual re-emergence of geography instruction in educational systems. The book itself is important for two additional reasons. First, Varenius was concerned with restoring what he felt was a needed balance between general and special geography, the balance being decidedly in favor of special geography. To Varenius the imbalance meant that students were ignorant of the foundations of their discipline and this lack of attention to general geography placed into question the status of geography as a science. Second, the book was written as a textbook for advanced students, not for scholars. The changing balance between special and general geography, the relative intellectual status of the discipline, and the role of textbooks are themes that recur through the history of geography education.

From the sixteenth to the early nineteenth centuries, geography was part of both the school and college education systems in Europe and North America. In the USA, for example, admission to the colonial colleges was based in part upon knowledge of geography. It was expected that school and college students would know about the world in general and their national worlds in particular, and that they would have some minimal understanding of the maps and globes which were instruments for conveying and using knowledge about the world. Geography instruction incorporated both general and special elements, although there was a preponderance of the latter in schools and the former in college. Geography was a standalone subject in the curriculum of schools and the faculty of colleges and universities. It was supported by a rich infrastructure of textbooks, atlases, wall maps, and globes. Geography instruction led, for example, to new tools. In England in the 1760s, John Spilsbury, a London mapmaker, created 'dissected maps' as a way of teaching geography. In an interesting twist, dissected maps became jigsaw puz-

zles, a major source of entertainment named after the tool used for their construction, and they lost their meaning as a tool of geographic instruction.

During the late-eighteenth and early-nineteenth centuries, geography instruction responded to larger societal projects. In European countries, the twin forces of nationalism and imperialism reshaped territorial relationships throughout the continent and the world. Geography became a way of inculcating an understanding of national identity and international aspirations. Thus there are numerous literary allusions to children being forced to learn maps of the British Empire in which imperial possessions were boldly portrayed in vast expanses of red. Children were subjected to a world being carved into places that were named and owned, demarcated by sharp lines and differentiated by rich physical and human detail. This detail was the cumulative result of several centuries of exploration, conquest, colonization, and settlement. There was also an ideology for structuring this world understanding. Places were ordered as being populated by savages, barbarians, half-civilized, and civilized and enlightened people. Progression along this typology was allegedly a response to the civilizing influences of European society. This ideology was the basis for many eighteenth and nineteenth century textbooks and atlases.

In the USA, two interlocking forces shaped the nature of geography education and led to its gradual demise as a school subject during the first half of the twentieth century. The first force was universal or mass education, a force which applied equally to European societies. The second force was the expansion and integration of the national territory. As the nation grew in size and complexity, there were increasing demands that students should understand the geography of the emerging nation. Universal education brought with it pressures towards access for all students and the fostering of a spirit of national identity. Political independence meant also independence in terms of what was taught and the material on which teaching was based. All of these imperatives lead to an increasing focus on special geography, with the USA being the focal point of the world view and with a sense that understanding one's own country was essential, easy, and accessible to all (Bruckner 1999).

Beginning in 1784, Jedidiah Morse wrote a series of textbooks that shaped American school geography. An 1814 edition was entitled *Geography Made Easy* and it was developed 'for the use of schools and academies in the United States of America.' These books spawned a series of progenitors throughout the nineteenth and twentieth centuries. The texts made all places, but especially the nation and the student's state of residence, seem special and perhaps unique. Children learned by rote, working from the ubiquitous trio of the globe, the textbook, and the wall map of the world. As geography, particularly special geography, literally became easy, it lost respect in the eyes of many in the academy. While the knowledge gained was useful and even essential, it did not challenge children's minds, except in the most trivial sense of factual learning.

From the middle of the nineteenth century to the beginning of the twentieth century, major geographic scholars were centrally involved in the process of geography education at all levels, K-12 included. For example, William Morris Davis in the USA and Halford MacKinder in the UK exemplified a seamless approach to the development of geography as an academic discipline and as a school subject. In both countries, there were organizations, the National Council for Geographic Education and the Geographical Association, whose primary function was the fostering of geography education in society. The National Geographic Society pursued its goal of increasing and diffusing geographic knowledge for mass audiences in as many countries as possible. Geography remained a popular subject, as exemplified by two bestselling books. V. M. Hillyer's (1929) *A Child's Geography of the World* and Hendrik van Loon's (1932) *Van Loon's Geography* remained in print for over 20 years and both were translated into many languages. There was a popular interest in geographic knowledge and therefore modest levels of support for K-12 geography instruction.

7. *Geography Education in the Modern Era*

During the middle to late twentieth century, geography education played different roles in different countries. In the UK, it remained a standalone subject that was part of the curriculum at all school levels and it was available as one of the Ordinary and Advanced level exit exams in high school. In this context, geography was traditionally grouped with the arts rather than the sciences. In both France and Germany, geography also played a comparably steady but modest role in the school curriculum. In the USA, geography languished at all school levels. The integrity of the discipline as a whole was destroyed as a consequence of being split apart as a subject. Human geography was subsumed within the social studies and physical geography within the earth sciences. These curricular emphases were mirrored in the structure of the education training system, with preservice teachers being explicitly trained as specialists in geography in many countries but not in the USA. In all cases, geography was taught as a cross between knowledge of the fundamental properties of places and regions in the world and, to some limited extent, a set of practical skills (e.g., map and atlas use, survey methods, fieldwork, etc.)

There are three principal frameworks underpinning the scope and sequence of geography in the K-12 curriculum. Curricula were place, scale, or concept-

based. The place-based curriculum structure is the oldest structure, dating back to the eighteenth century focus on special geography. Variously known as the 'capes and bays' or the 'principal products of Peru' approach, it employed a simple classificatory framework to capture the characteristics of places around the world. Students learned about the location, boundaries, physical geography, and human geography of places, often in a rote fashion and using simple maps and atlases. In effect, younger children learned less, as in fewer and simpler things, and older children learned more, as in more detail, about the world. Geography education became an A to Z world tour, organized around continents, countries, and regions.

The scale-based structure is also known as the expanding environments or expanding communities approach. It begins in the early grades with a focus on the familiar and on the local world of the child, progressing to increasingly distant places with increasing age of the students. The local-to-global progression is premised on teaching about what is readily accessible to the understanding of children at different ages and levels of cognitive understanding. The third approach is concept-based, developing a systematic understanding of sets of concepts that can be applied to a range of spatial contexts at different scales. Geography is seen as a way of asking and answering questions about what is where in the world, and about why and how it got there.

These three curricular structures fall loosely into a historical sequence, if not progression, and reflect significant changes in the view of the discipline of geography. The essential transition is from seeing geography as a property of the world (the geography of a place) to seeing geography as a perspective on the world, as a way of understanding patterns and processes on the earth's surface. In the place-based approach, there are lists of things to be known about the world; in the concept-based approach, there are alternative ways of knowing about the world. Beginning with the scale-based approach, there is a sense of active involvement with the world, an involvement based on field trips, on maps as tools for recording and investigating places and environments, and students as explorers and creators of their own understanding.

Underpinning these three approaches are, therefore, approaches to child development. Many of the great educational theorists (Pestalozzi, Montessori, Piaget, etc.) treat issues that are central to the child's understanding of the world, although they do not necessarily do so within the formal domain of geography. Thus, for example, Piaget discusses the development of spatial understanding, the sense of nation, the understanding of basic physical phenomena (e.g., clouds, rain, etc.) and he makes use of maps and spatial models in his research designs. Geography education, while not necessarily basing its approaches explicitly on such theorists, is of necessity informed by their ideas.

In the early to middle 1980s, there was a resurgence of interest in many countries in the success—and often failure—of education systems in generating knowledgeable and competent students. The context for this interest was the intersection among several factors, including the effects of globalization on national economies, the desire to remain competitive in a rapidly changing world, and the increasing demands for technical skills in the workforce. As typified in the seminal US report *A Nation At Risk* (National Commission on Excellence in Education 1983), there were significant concerns about the structure of education in terms of teacher training, teacher support, curricula, textbooks, assessment, accountability, etc.

In the case of geography, the focus, especially in the USA, was on unacceptably low levels of geographic literacy among recent high school graduates. International comparisons showed relatively low levels of geographic literacy in a range of countries, but the performance of American students was particularly disappointing. As a consequence of the renewed national interest in education, geography education became the focus of attention from national and private organizations in many countries.

That attention paralleled national concerns about the knowledge and skills of school graduates and it led to a normative focus on what students should know and be able to do. The focus gave rise to national documents specifying the content of the curriculum at various grades and/or ages. Often these documents specified learning outcomes that were expected of students and therefore the documents became the basis for the assessment of student learning. In the case of geography, both the UK and the USA produced national guidelines (or standards).

8. *The Current State of Geography Education*

Geography education remains a contested enterprise. As K-12 educational systems place an increasing emphasis on immediately relevant skills—mathematics, writing, and to some extent science—other subjects are of necessity given less attention in terms of time in the classroom, teacher training, material support, and access to computing resources. Where geography education is infused in the curriculum, this lack of attention is a serious problem and results in geography being squeezed out or at best subsumed within another subject such as history. While there is public support for a level of geographic knowledge, that support is often in terms of place location knowledge, a necessary but insufficient base for geography education.

Geography education is changing in ways that move it away from the negative image associated with 'old-fashioned' rote learning of geographic facts. Several

transitions are underway. The first is a recognition that the process of geography education can be set within a novice–expert framework. This entails understanding the different types of knowledge that underpin increasing expertise: (a) declarative or propositional (e.g., factual and conceptual understanding), (b) procedural (e.g., how to perform some operation), and (c) metacognitive (when, where, and why to link facts, concepts, and operations to answer geographic questions). It involves understanding the organization of the mental models that children develop (e.g., Earth–Sun relations, the water cycle, urban growth), the ways in which different skills are available at different ages (or cognitive developmental levels), and the role of peers and teachers in supporting the student's developing understanding (Downs 1994).

A second transition is a switch from rote learning to active, inquiry-based learning. Geography becomes a way of finding out about the world rather than simply a set of things to be known about the world. Computers and geographic databases allow students to ask and answer geographic questions for themselves. A third transition is from tests to authentic assessments. Assessment is built into the structure of the ongoing learning process and is focused on problems and situations that are likely to be encountered in real-world contexts. Students can demonstrate their developing skills by building portfolios of their work.

9. *The Future of Geography Education*

All of these changes in geography education mirror changes that are occurring in the nature of education in general. The short-term future of geography education lies in the successful linking of the case for geography education with the increasing capacity to do geography. In the same way that geography was fundamental to past societal projects (e.g., exploration and imperialism), it is also fundamental to the current development of an information society. Much of the data that is available can be assigned geospatial coordinates and therefore it can be handled efficiently and effectively via geographic information systems. The larger parent field of geographic information science requires an understanding of maps and graphics, of space and place, of the interactions between environment and society. Those understandings, while not exclusive to geography, are central to the subject and can be taught through geography education. Given the growing importance of information science and technology to all aspects of modern society and economy, there is a potential to reestablish an essential role for geography education in the contest for space—and time—in the educational system (Rediscovering Geography Committee 1997).

See also: Cartographic Visualization; Cartography; Geographic Learning in Children; Geographic Thought, History of; Knowledge, Education, and Skills, Geography of

Bibliography

Bednarz S, Downs R M, Vender J in press Geography education. In: Gaile G L, Wilmott C J (eds.) *Geography in America at the Dawn of the 21st Century*. Oxford University Press, Oxford, UK

Bruckner M 1999 Lessons in geography: Maps, spellers, and other grammars of nationalism in the early Republic. *American Quarterly* **51**: 311–43

Douglass M P 1998 *The History, Psychology, and Pedagogy of Geographic Literacy*. Praeger, Westport, CT

Downs R M 1994 Being and becoming a geographer: An agenda for geography education. *Annals of the Association of American Geographers* **84**: 175–91

Geography Education Standards Project 1994 *Geography for Life: Executive Summary, National Geography Standards*. National Geographic Research & Exploration, Washington, DC

Hillyer V M 1929 *A Child's Geography of the World*. The Century Co., New York

Livingstone D N 1992 *The Geographical Tradition: Episodes in the History of a Contested Tradition*. Blackwell, Cambridge, MA

National Commission on Excellence in Education 1983 *A Nation At Risk: The Imperative for Educational Reform*. US Government Printing Office, Washington, DC

Rediscovering Geography Committee, National Research Council 1997 *Rediscovering Geography: New Relevance for Science and Society*. National Academy Press, Washington, DC

Van Loon H W 1932 *Van Loon's Geography: The Story of the World We Live In*. Simon and Schuster, New York

Warntz W 1964 *Geography Now and Then*. American Geographical Society Research Series No. 25, New York

Warntz W 1982 Geographia generalis and the earliest development of American academic geography. In: Blouet B (ed.) *The Origins of American Geography in the United States*. The Shoe String Press, Hamden, CT, pp. 245–63

R. M. Downs

Geographic Information Systems

A geographic information system (GIS) is best defined as a software system designed to provide services related to geographic data. Such services include storage, analysis, transformation, maintenance, editing, visualization, modeling, and many more. That somewhat dry definition masks a record of rapid growth and widespread applications, beginning some 30 years ago. GIS is now the focus of a substantial software industry and boasts a user community that probably numbers in the hundreds of thousands or

even millions worldwide. While many of those applications lie in areas of social and behavioral science, they also extend across the full range of sciences, industries, and government agencies that deal in one way or another with information about the surface of the earth, the subsurface, and the atmosphere.

1. Overview

GISs are one of several geographic information technologies that have appeared or evolved rapidly in recent years. Geographic information is also handled in image processing systems, in the form of aerial photographs or images from space satellites, and many GISs include image-processing capabilities. The Global Positioning System (Kennedy 1996) and its Russian equivalent use satellites to allow users to measure their positions on the earth's surface, with accuracies that range from millimeters to tens of meters depending on the complexity of the measuring device. Earth-observing satellites (Jensen 1996) capture images of the surface and are now widely used to map, detect change, discover and inventory resources, forecast weather, and measure variables of importance to earth science. The Internet and the World Wide Web (WWW) are widely used to distribute geographic data from data archives and digital libraries, and to provide GIS services to remote users.

Today, the potential user of GIS might take advantage of any of a number of hardware and software configurations. The examples that follow relate to systems available in the year 2000 and are ordered by increasing cost and power.

(a) A simple WWW browser, running on a laptop or office machine, and used to access WWW sites that offer a range of GIS services. Such services include generation of standard map products, and simple analysis and mapping of data.

(b) A desktop system, such as a PC or Macintosh, running a relatively simple GIS such as ArcView (Environmental Systems Research Institute (ESRI), www.esri.com) or MapInfo (MapInfo Corporation, www.mapinf.com). Data would be resident on the system's hard drive, and might have been created locally or obtained from a data supplier. The operating system would likely be a Microsoft or Apple product.

(c) A networked workstation, linked to other workstations and servers in a local area network, perhaps under the Windows NT or Unix operating system. The software would be more complex than in the first option, perhaps ESRI's ArcInfo or Spatial Database Engine.

A social scientist 'doing GIS' might use any one of these configurations, depending on personal skills and needs. The system would provide services that include preparation of simple maps, manipulation of data to overcome issues of map projections, running of statistical tests, and fitting of complex models to data.

Courses in the use of GIS and in methods of spatial analysis and modeling are now offered in almost all universities and colleges, and in many high schools. A wide range of educational materials are available in the form of textbooks (e.g., DeMers 1997; Clarke 1997; Heywood et al. 1998; Burrough and McDonnell 1998), WWW sites (e.g., www.esri.com and www.geoplace.com), and WWW-based courses (e.g., www.unigis.com). GIS software vendors maintain WWW sites with much useful information, and trade magazines (*GeoWorld* and *GeoInfo Systems*) carry information about new products and initiatives.

The history of GIS (Foresman 1998) originated in the 1960s, with the development of the Canada Geographic Information System (CGIS). Analysis of paper maps is notoriously difficult, because of the awkwardness and unreliability of techniques for measuring simple properties such as length or area, and because of the difficulties associated with overlaying different maps of the same area. Computer technology was applied in the late 1960s and early 1970s to the editing of maps by producing agencies such as the US Geological Survey, and to the automation of various functions associated with taking the census. But early efforts were crude and expensive, and GIS became widely available to government agencies and corporations only in the early 1980s. Today, the general public is increasingly aware of GIS through WWW sites such as www.mapquest.com, which generates several million maps a day for users who need driving directions, and through simple mapping packages such as Microsoft's MapPoint 2000.

2. GIS Essentials

General computer packages to support analysis of data began to appear in the 1960s, led by BMD, SAS, and SPSS. The table is a powerful model for organizing data, by arraying cases in the rows and devoting each column to some property of each case, and early packages took advantage of this commonality by organizing housekeeping and analytic functions around it. Thus, means and variances could be computed by column; regression could be readily applied to two columns; and data could be aggregated by computing the mean value of a column for every unique value of some other, categorical column. In essence, these packages exploited an economy of scale in software production: once the housekeeping functions for a simple data model like a table had been built, it became very cost-effective to add new functions based on the same data model.

GIS similarly exploits economies of scale, but in this case by supporting not one but several data models, all of them effective ways of representing the information content of a map or image. The real geographic world is enormously complex, and there are many different ways of capturing its contents in digital form, all of

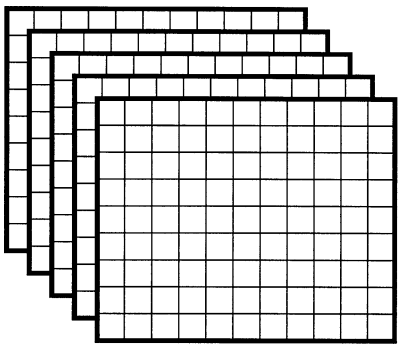

Figure 1
The concept of layers. Geographic reality is represented by a set (five in this case) of co-registered rasters, each representing the spatial pattern of one variable. Example layers might include soil type, vegetation cover type, ownership, mean annual temperature, or elevation

them involving some form of approximation. Remotely sensed images use a raster model, capturing the earth's reflectance in each of an array of square cells. Because of its simplicity, the raster model has been used as the basis for many GISs, each of them representing data as a series of layers of co-registered raster cells (Fig. 1), with each layer capturing some distinct variable, such as land elevation, county name, average income, population, or class of land use.

In the vector model the world is viewed as occupied by collections of points, lines, and areas. Points are located by coordinates, often latitude and longitude or some universal coordinate system such as the Universal Transverse Mercator (UTM). Lines are represented as polylines by connecting an ordered sequence of points with straight lines. Areas are similarly represented as polygons or figures with point vertices connected by straight edges. Each point, line, or area object in the vector model is associated with one or more attributes which capture the object's significant characteristics. A set of objects of the same type is associated with an attribute table by assigning one row of the table to each object, and one column to each characteristic.

Figure 2 shows a simple example of raster and vector data. The same set of census tracts is shown in Fig. 2(a) in the form of a raster, and in Fig. 2(b) in vector form. Most GISs include support for both models, and for transformations between them. They also support various kinds of map projections, allowing the user to switch readily between latitude and longitude, UTM, and other coordinate systems, and to display any part of the earth's surface in the form of a flat paper map.

One of the most compelling arguments for GIS relies on the ability to integrate information around a common geographic location. A project might involve many different kinds of information about different themes or variables for a geographic area. Every point, line, area, or raster cell in a GIS database has an associated location, so it is possible to determine the distance between two locations or to place locations correctly on a map. But more significantly, it is also possible to identify all of the information available about a location, or to link together two data sets through common location. GISs include functions to determine how much overlap exists between two areas, or whether a given point lies inside a given area, and functions to select all of the points lying within a given distance of a line. To return to a point made earlier, the economics of software production provide great advantages to GIS designers once a basic structure has been laid down, because new functions that work on the same structure can be added very easily.

Certain key GIS functions are particularly relevant to social and behavioral science applications.

(a) Geocoding, or the ability to convert street addresses to geographic coordinates, and vice versa.

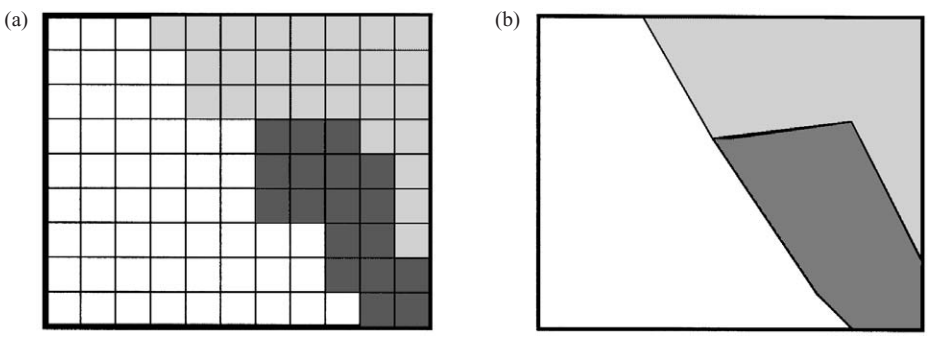

Figure 2
Representation of three adjacent census tracts using raster cells (a) and vector polygons (b). In the raster case, cell size determines the accuracy of boundary positions. In the vector case, each tract is located by specifying the coordinates of its vertices in order

This function might be used to associate coordinates with surveyed households, allowing locations to be mapped, integrated with other data, or subjected to various kinds of analysis.

(b) The point-in-polygon function, or the ability to compare a set of points to a set of polygons, to identify the polygon containing each point, and to combine the characteristics of the containing polygon with those of the point. For example, this function might be used to associate census tract properties with those of the individuals who live there.

(c) The polygon overlay function, or the ability to combine two sets of polygons, creating a new polygon for every unique area of overlap, and associating it with the attributes of the two polygons that overlap to create the area. By overlaying a map of county boundaries on a map of school districts, it would be possible to estimate the number of children living in each school district based on knowledge of the population density of each county (subject to the assumption that population densities are uniform within each county).

(d) The buffer function, which identifies the area lying within a specified distance of a given point, line, or area, and gives it the attributes of this object. By combining this with the geocoding and point-in-polygon functions, it would be possible, for example, to determine the addresses of households within a certain distance of a shopping center.

3. *GIS as Research Infrastructure*

GIS software first became widely affordable in the late 1980s, when personal computers powerful enough to manipulate geographic data became available for a few thousand dollars. Social and behavioral scientists began to invest in its capabilities, particularly in those disciplines that had traditionally adopted a spatial perspective. Archaeologists and anthropologists, for example, began to use GIS software to map and store artifacts at the scale of the dig, to build digital maps of reconstructed landscapes, and to make use of spatial analysis to understand how early societies operated in space (Aldenderfer and Maschner 1996). Criminologists use GIS to map crime, correlate crime rates with social variables, and examine detailed spatio-temporal patterns of criminal behavior. Political scientists use GIS to evaluate districting schemes, and to explore the spatial diffusion of political movements (Eagles 1995). Market researchers and location analysts use GIS to map market potential, model the spatial behavior and choices of consumers, and search for optimal locational strategies (Martin 1996).

Of all the social and behavioral disciplines, geographers have long claimed a special interest in space, and much of the early work on GIS design and application was strongly associated with the discipline. Geography is often the home department of GIS, though many other arrangements exist, especially at institutions lacking a geography department. GIS courses are frequently offered by geography departments, and geography is often the sponsor of campus GIS infrastructure, in the form of laboratories, software licenses, and data repositories.

In recent years efforts have been made to develop wider linkages. In the USA the University Consortium for Geographic Information Science (www.ucgis.org) now links over 60 member institutions in an organization dedicated to advancing GIS and its applications through research and education. In Europe, the Association of Geographic Information Laboratories in Europe (AGILE) has a similar mission, and the European Science Foundation has sponsored a major study of GIS in the social sciences through its GISDATA program. National centers have been funded in several countries to pursue GIS research: in the USA, the National Center for Geographic Information and Analysis (www.ncgia.org) was founded by the National Science Foundation in 1988, while in Canada the GEOIDE network was founded by the three national research councils in 1998 (GEOIDE, *La géomatique pour des interventions et des décisions éclaireés*; see www.geoide.ulaval.ca).

Several journals are devoted to advances in GIS, including the *International Journal of Geographical Information Science*, *Cartography*, *and Geographic Information Systems*; *Geoinformatica*, the *Journal of Geographical Systems*; and *Transactions in GIS*. Descriptions of research using GIS can be found in major journals in many application disciplines.

Comparisons have often been drawn between GIS and statistics, since both can be seen as generic tools and concepts used widely across the sciences (e.g., Goodchild 1987). It is argued that GIS provides the implementation of methods of spatial data analysis in easy-to-use computer tools, just as the statistical packages implement methods of statistical analysis. In this context spatial data analysis can be defined as a set of analytic methods whose results are not invariant under changes in the locations of the objects of analysis—in other words, location matters. This serves to distinguish these techniques from more familiar numerical and statistical analyses, such as the calculation of a mean, or the preparation of a scatterplot, that do not normally include location or respond to it. Spatial analysis predates GIS, and major texts include those by Berry and Marble (1968), Haining (1990), and Bailey and Gatrell (1995). However, the advent of GIS has greatly improved access to these methods and, moreover, it has come at a time when there is increasing interest in visual, intuitive, and exploratory techniques, all characteristics that are readily compatible with GIS. For discussions of the impact of GIS on spatial analysis see, for example, Fotheringham and Rogerson (1994), Goodchild and Longley (1999), and Fischer et al. (1996).

4. Sources of Data

Maps are highly compressed representations of large amounts of information, and extremely tedious to convert to digital form. Two methods are available, with both advantages and disadvantages:

(a) Digitizing, a process in which a human operator traces information from a map, capturing the locations of features and their attributes using a pointing device and a keyboard. The process is very expensive, since it is difficult to capture complex geographic features rapidly and accurately.

(b) Scanning, in which an automated device captures the information from a map. Extensive editing and error correction is necessary, especially if the map is of low quality, such that scanning is typically as expensive as digitizing.

In the social and behavioral sciences, data are often highly specific to applications and must be rendered into digital form using one of these methods. Thus an archaeologist may have to digitize the locations of sites or artifacts from manuscript maps or drawings. In other cases, however, the process of database creation will take advantage of the widespread availability of framework data, or geographic data sets of generic use that already exist in digital form. These include:

(a) Topographic maps, in both raster and vector form (produced in the USA by the US Geological Survey and known as digital raster graphics, or DRGs, and digital line graphs, or DLGs. Various mapping scales are commonly available (in the USA the most detailed scale is, 1:24,000).

(b) Elevation data in raster form (produced in the USA by the US Geological Survey and known as digital elevation models, or DEMs). These data sets vary according to the spacing of the grid: for the entire earth the best spacing available is several kilometers, but 30 m spacing is available for much of the USA.

(c) Street centerline data in vector form. These data sets record the locations, names, address ranges, and other information on streets. The TIGER (Topologically Integrated Geographic Encoding and Referencing) database created and maintained by the US Bureau of the Census covers all of the USA, and includes boundaries of statistical reporting zones. These data sets represent streets by single lines and have positional accuracies of tens of meters.

(d) Imagery from aerial photography. The Digital Orthophoto Quad (DOQ) is typical of these data sets, with a 1 m ground resolution and a 6 m positional accuracy, and available from the US Geological Survey for much of the USA.

(e) Imagery from satellites. These sources include images collected by the Landsat and SPOT satellites, with cell sizes down to 10 m.

(f) Municipal databases created by local government agencies, often in collaboration with utility companies. These are in vector form and often include property boundaries, locations of buildings, and streets represented by double lines. In some countries, notably the UK, they are produced by government for the entire land area. In the USA, coverage is patchy. Positional accuracies are typically in the meter range.

(g) Gazetteers, the digital versions of the indexes found in atlases, linking placenames to geographic coordinates. These indexes make it possible to translate information recorded by placename to map form, for analysis and display.

Many of these data sets are created and maintained by longstanding programs of national mapping agencies. Recently, many countries have attempted to place them within broader policy frameworks that emphasize sharing of data, avoidance of duplication, and the setting of standards. In the USA, the National Spatial Data Infrastructure (National Research Council [NRC] 1993) is a leading example.

These data sets provide the framework on which social and behavioral scientists can often hang their own features and attributes. For example, a criminologist might use a municipal database, adding records of crime locations; a sociologist might use reporting zone boundaries from the TIGER database to create maps and analyze data at the census tract level; and a planner might use imagery to monitor urban growth. By using data that are already in digital form the cost and tedium of database creation are greatly reduced.

Unfortunately, the conditions under which such framework data sets are accessible vary widely. In the USA, policy requires that geographic data produced by the federal government be distributed to anyone at no greater cost than that of dissemination. It is often argued that this policy has stimulated growth in the use of GIS and in the geographic information industry generally. But while US courts have held that geographic data are facts and therefore not subject to copyright, in other countries national governments have adopted cost-recovery policies, protected by copyright. Academic scientists in such countries have found various ways of accessing geographic data at lower cost, through bulk purchase agreements. Lower levels of government in the USA are also not subject to the cost-of-dissemination policy, and have made efforts at cost recovery under trademark protection. For more discussion of these issues, see *Spatial Data Infrastructure*.

While the public sector has traditionally supplied many of the geographic data used by the scientific community, a growing private sector may have substantial impact in the coming years. Very large geographic databases are being created for internal purposes by insurance, retail, banking, and other corporate sectors. Point-of-sale systems now routinely capture data on grocery purchases, and direct-mail and delivery companies make use of detailed geographic data for scheduling and targeting. These activities have raised privacy concerns, and it is clear

that the levels of detail now available to corporations, together with the ability to integrate and merge data based on geographic location, have very significant privacy implications. Much of this activity remains largely outside any regulatory framework.

5. Networks

The Internet and WWW have had major impact on GIS. They have made it much easier to share data and to access data remotely. Data can be published and disseminated easily, as can the more sophisticated results of analysis. Indeed, the entire process of scholarly communication and publication is being radically restructured by networking technologies.

It is useful to distinguish two forms of processing of geographic information. The term 'geographic information system' suggests a combination of hardware and software that may or may not be connected to a larger network. It is probably controlled directly by a single user, and provides capabilities and functions directly related to the user's needs. On the other hand the term 'geographic information services' is used to describe WWW sites that offer specific GIS functions in response to user requests. Instead of sending the data to a GISystem, a GIService retains its data, sending only the results of analysis. The Mapquest site (www.mapquest.com), for example, provides a range of services from the generation of maps at user request to the geocoding of addresses supplied by the user.

GIServices are growing rapidly, and are expected to grow much further in the coming years. In addition to avoiding lengthy transmission of data, they allow the custodian of a sophisticated analysis tool to retain control, and to upgrade the tool, rather than undertake a complex process of distribution to user sites. Pricing based on services is very different, and perhaps simpler, than pricing based on data (Guenther and Mueller 1999). GIServices also assume a level of adherence to standards, and encourage the development of interoperable tools using common formats and protocols (the Open GIS Consortium, www.opengis.org, is a major influence toward greater interoperability in GIS).

Several major initiatives are making geographic data more easily found and accessed via the WWW. In the USA, the Federal Geographic Data Committee sponsors the National Geospatial Data Clearinghouse (www.fgdc.gov), a catalog of sites offering geospatial data, linked together using the Z39.50 protocol. The Alexandria Digital Library (alexandria.ucsb.edu) offers the services of a map library, including search and retrieval of data, over the WWW. Other sites offer more specialized data, for example, a site at the Massachusetts Institute of Technology (ortho.mit.edu) offers DOQs for the Boston area. The term 'geolibrary' has been coined to describe WWW sites that allow search for data based on geographic location, including search for information that is associated with a geographic area, but which is not a map or an image, such as a planning report. The implications of geolibraries are discussed in a recent report of the US National Research Council (NRC 1999).

6. Space and Scientific Explanation

Ultimately, the value of GIS and related technologies in the social and behavioral sciences will be intimately tied to the role of space in scientific explanation and understanding. Before expanding on that point, however, it is useful to distinguish between normative and explanatory applications. Many applications of GIS may be scientific in their objectivity and rigor, but nevertheless aimed at supporting decisions, developing policy, or solving day-to-day problems. Such normative efforts often apply the principles and knowledge of social and behavioral science. Moreover, many do so in a spatial context, by combining principles about behavior and society with the specific conditions existing in a geographic area. Thus, there are abundant normative applications of GIS.

In this section, however, the focus is on those applications that advance scientific knowledge about the world, through observation, experiment, hypothesis-testing, and deduction, because such explanatory efforts lie at the heart of the social and behavioral sciences. To what extent can GIS be used as a tool to advance scientific knowledge?

Ideally, scientific knowledge advances through controlled experiment, by varying the conditions under which society operates and observing the response. But although such experiments are common in areas such as agriculture, and GIS is often used as a tool to manage experimental plots, they are unlikely to be possible in the case of social and behavioral processes operating at geographic scales. Instead, science must advance through inference, either inductively by inferring the nature of a process from its geographic or temporal expression, or deductively by comparing the predictions of theory with reality.

GIS is most strongly associated with cross-sectional approaches to observation and inference. A phenomenon is compared across a geographic area, and its spatial variation is compared with the variation in relevant conditions. Inferences are drawn about process, but since the process is not observed directly it is impossible for inference to be definitive. Alternatively, patterns may be used to provide confirmation of hypothesized processes, though again confirmation is never definite. The tradition of using maps as a basis for inference has a very long history, one of the most compelling historic examples being the map made by Dr. John Snow of a cholera outbreak in the Soho district of London in the 1850s, which provided evidence of the role of drinking water as the disease vector.

It is often argued that longitudinal observation and inference based on change through time is more

definitive than cross-sectional inference. But cross-sectional data are widely available from sources such as the census, and powerful technologies like GIS increasingly make it possible to examine data from both cross-sectional and longitudinal perspectives.

Geographers draw a distinction between the properties of a site, and the site's situation in spatial context. Many social processes respond both to the conditions at a place, and to the conditions in the place's surrounding area. For example, choice of residential location is affected both by the properties of the available sites, and by their surrounding communities and neighborhoods. GIS makes it possible to observe and analyze a site's situation, in the form of maps and images, and also to link one property of a site to other properties through their common geographic location. Further discussions of the spatial dimension of explanation, and of the role of GIS, are available in many sources (e.g., Abler et al. 1992, NRC 1997a).

7. *GIS in the Future*

There is every indication that the use of computers to handle geographic information will continue to increase dramatically in the future, and that the user base will continue to broaden. But trends in the software industry, and in society more generally, suggest that the nature of GIS will continue to evolve.

In 1997 the US National Research Council sponsored a workshop on the Future of Spatial Data and Society (NRC 1997b). It found strong support for continued increases in technical power, in the form of computer speed, network bandwidth, and storage capacity, at constant or diminishing cost. But it also identified a large number of societal trends that are likely to affect GIS, and also fundamental trends in the nature of software.

The report found a discernible trend in recent years towards the solution of problems at the local level. It is reflected, for example, in the popularity of place-based decision making in the US Environmental Protection Agency, a policy that encourages the solution of problems locally, by local stakeholders. It is also reflected in efforts to devolve power to lower levels in the administrative hierarchy.

In the 1960s, the so-called quantitative revolution in geography emphasized the nomothetic search for general laws affecting spatial phenomena, and saw traditional geography's approach as idiographic, with its stress on cataloging the unique properties of places. Today, there is much interest in quantitative techniques that accommodate spatial variation in parameters, and emphasize the role of spatial analysis in exploring local variation. Fotheringham, for example, has identified this as one of the most important themes in contemporary spatial analysis (Fotheringham 1997). GIS provides an ideal platform for such methods, since it makes it easy to analyze spatial data, and to display spatial variation.

At the same time, there is increasing interest in several disciplines in the incorporation of spatial effects directly into theory, on the grounds that a system that is spatially distributed, and forced to communicate over distance, is fundamentally different from one that is infinitely connected. Spatial economics and spatial ecology are now significant areas within the theoretical wings of those disciplines, while geography has a long tradition of explicit theorizing about space. Again, GIS provides an ideal platform for subjecting such spatially explicit theory to empirical test.

Finally, the spatially explicit nature of GIS provides an ideal environment for linking normative and explanatory science. The implications of theories for real geographic areas are readily evaluated, and models that incorporate social and behavioral processes can be combined with geographic boundary conditions to provide readily visualized predictions. Thus, GIS offers an ideal environment for bridging the gap between theory and practice, and a significant clue in the eternal search for relevance.

See also: Computers and Society: Critical Approaches; Spatial Analysis in Geography; Spatial Pattern, Analysis of; Spatial Thinking in the Social Sciences, History of; Technological Innovation

Bibliography

Abler R F, Marcus M G, Olson J M (eds.) 1992 *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography*. Rutgers University Press, New Brunswick, NJ

Aldenderfer M, Maschner H D G 1996 *Anthropology, Space, and Geographic Information Systems*. Oxford University Press, New York

Bailey T C, Gatrell A C 1995 *Interactive Spatial Data Analysis*. Longman Scientific and Technical, Harlow, UK

Berry B J L, Marble D F 1968 *Spatial Analysis: A Reader in Statistical Geography*. Prentice Hall, Englewood Cliffs, NJ

Burrough P A, McDonnell R A 1998 *Principles of Geographical Information Systems*, 2nd edn. Oxford University Press, New York

Clarke K C 1997 *Getting Started with Geographic Information Systems*. Prentice Hall, Upper Saddle River, NJ

DeMers M N 1997 *Fundamentals of Geographic Information Systems*. John Wiley, New York

Eagles M (ed.) 1995 *Spatial and Contextual Models in Political Research*. Taylor and Francis, London

Fischer M, Scholten H J, Unwin D J (eds.) 1996 *Spatial Analytical Perspectives on GIS*. Taylor and Francis, London

Foresman T W (ed.) 1998 *The History of Geographic Information Systems: Perspectives from the Pioneers*. Prentice Hall PTR, Upper Saddle River, NJ

Fotheringham A S 1997 Trends in quantitative methods I: stressing the local. *Progress in Human Geography* **21**: 88–96

Fotheringham A S, Rogerson P A (eds.) 1994 *Spatial Analysis and GIS*. Taylor and Francis, London

Goodchild M F 1987 A spatial analytic perspective on geographic information systems. *International Journal of Geographical Information Systems* **1**: 327–34

Goodchild M F, Longley P A 1999 The future of GIS and spatial analysis. In: Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds.) *Geographical Information Systems: Principles, Techniques, Management, and Applications*, Wiley, New York, pp. 567–80

Guenther O, Mueller R 1999 From GISystems to GIServices: spatial computing on the Internet marketplace. In: Goodchild M F, Egenhofer M J, Fegeas R, Kottman C A (eds.) *Interoperating Geographic Information Systems*. Kluwer Academic, Norwell, MA, pp. 427–42

Haining R P 1990 *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, UK

Heywood I, Cornelius S, Carver S 1998 *An Introduction to Geographical Information Systems*. Addison-Wesley Longman, New York

Jensen J R 1996 *Introductory Digital Image Processing: A Remote Sensing Perspective*, 2nd edn. Prentice Hall, Upper Saddle River, NJ

Kennedy M 1996 *The Global Positioning System and GIS*. Ann Arbor Press, Chelsea, MI

Martin D J 1996 *Geographic Information Systems: Socioeconomic Applications*, 2nd edn. Routledge, London

National Research Council 1993 *Toward a Coordinated Spatial Data Infrastructure for the Nation*. National Academy Press, Washington, DC

National Research Council 1997a *Rediscovering Geography: New Relevance for Science and Society*. National Academy Press, Washington, DC

National Research Council 1997b *The Future of Spatial Data and Society*. National Academy Press, Washington, DC

National Research Council 1999 *Distributed Geolibraries: Spatial Information Resources*. National Academy Press, Washington, DC

M. F. Goodchild

Geographic Information Systems: Critical Approaches

Geographic information systems are computerized information systems, computer hardware and software, used to store, manipulate and map geo-referenced databases of information about the physical and human world. Growing rapidly since their development in the 1970s, GIS are now used in a wide variety of private and public organizations, and increasingly by individuals and grassroots organizations, for analysis, decision-making, and the display of geographical information. In response to this growing influence, critical assessments of GIS have emerged. These have analyzed their positivist bias, their potential to reinforce social inequality, their capacity for surveillance and ethical and legal implications, and have sought to examine these issues and to pursue alternative approaches to contemporary GIS. Critical approaches refer to analyses of GIS that are based in critical theory, i.e., that assess GIS in terms of their relationship to power hierarchies and wealth inequalities in society (Habermas 1984). This article:

traces the origins of such critiques of GIS;

discusses the engagement between critical theorists and GIS researchers that resulted in development of a GIS and society research agenda;

examines critical assessments of the impact of society on GIS, and of GIS on society; and

summarizes attempts to seek alternatives to GIS that may reduce some of the problems identified through these critical approaches.

1. *Early Critiques*

Broad-based claims were made about GIS as the key to a new integrated scientific geography at the end of the first wave of GIS research. This was the point in time when the technology had stabilized around a layer-based GIS architecture, and when the US National Science Foundation had recognized the importance of GIS through establishment of the National Center for Geographic Information and Analysis (cf. Dobson 1983). Written at a time when human geographers were moving away from positivist and critical rationalist approaches to research, such statements catalyzed a series of critical responses (Pickles 1999).

These criticisms focused on both epistemological and practical implications of the spread of GIS. Epistemologically, GIS was argued to be a Trojan horse for the reassertion of broadly positivist approaches within geography, because of its quantitative and empirical nature. Critics argued that geography was too complex and varied a discipline to be analyzed with this approach alone, criticized the claims that positivist science is foolproof, universal and objective, and claimed that the use of GIS would marginalize critical theoretic and hermeneutic approaches to geography, thereby limiting its ability to make sense of the world. They also argued that this kind of scientific approach would not promote human emancipation, and would not leave space for investigation of other possible worlds than those observed and experienced. The benefits for geography of using GIS to enhance its status as a science were seen to be outweighed by the costs associated with a consequent narrowing of geographic inquiry.

It was also argued that embedded within GIS are certain conceptions of space (particularly a geometric and relative space) and certain forms of reasoning (particularly Boolean logic) that are unable adequately to represent non-European conceptions of space and communicative rationality (Habermas 1984, Rundstrom 1995, and Sheppard 1995). Practically, it was argued that increased use of GIS in society would enhance current social and geographical inequalities. It was noted that social actors have differential access to GIS; wealth, gender, race, and geographic location are all strongly correlated with access to the necessary equipment, access to training, and comfort in working

with GIS. Consequently, GIS was seen as facilitating practices of surveillance, social engineering, opinion formation, and warfare by those with access to the technology. These arguments closely parallel more general analyses of the digital divide. These critical attacks provoked equally sharp responses by GIS specialists.

These specialists saw the critiques as simplistic, unduly pessimistic, and even paranoid, and as reflecting a lack of understanding of and experience with GIS and/or a lack of patience or aptitude for the rigors of science. They also resented the implication that GIS specialists are unconcerned with the social issues raised, and/or unaware of the social implications of science. At the same time, they saw such criticisms as a minor irritation that could be ignored given the growing interest in GIS among both students and practitioners. As a consequence, after some initial high profile exchanges, from 1983 to 1993 there was little communication between the two 'cultures of indifference': those critical of and those specializing in GIS (Pickles 1999).

2. *Engagement: GIS and Society*

This intellectual divide was challenged in 1993, when the National Center for Geographic Information and Analysis sponsored a conference organized by both GIS specialists and social theorists who sought a more constructive engagement. This and a subsequent conference in 1995 brought together prominent researchers (Poiker and Sheppard 1995 and Harris and Weiner 1996). Notwithstanding some early tensions, caricatures broke down as participants gained a greater appreciation for the breadth of skills and interests of those coming from the other side. A common desire among those present to learn from one another stimulated development of a GIS and society research agenda composed of the following themes:

the social history of GIS as a technology;

the relevance of GIS for community and grassroots perspectives and lifeworlds;

issues of privacy, access to spatial data, and ethics;

the gendering of GIS;

GIS, environmental justice and political ecology;

GIS and the human dimensions of global change; and

alternative approaches to GIS.

Formal collaboration continued, as collective research catalyzed by these meetings, in jointly organized sessions at GIS and geography conferences, and within the 'Geographies of the Information Society' research initiative workshops between 1997 and 1999 (Sheppard et al. 1999). In addition, critical theorists were asked to participate in key GIS events, and GIS theorists were given voices in critical theory projects (cf. Pickles 1995). By the end of the 1990s, the constructive engagement between GIS specialists and their critics had substituted overlapping cultures of respect for separate cultures of indifference, notwithstanding the continuing reluctance of many critical geographers or GIS specialists to engage with one another.

Tensions between different perspectives remain. For example, the renaming of GIS as Geographic Information Science revived concerns about attempts to reassert an agenda of normal, positivist science with all its trappings (Pickles 1997). Yet the tone of debate had shifted to a search for common ground; much to the relief of graduate students who had been uncomfortable about having to choose between GIS and critical geography (Sheppard et al. 1999). With the shift towards more constructive engagement has come a shift from critique, in which possibilities were suggested and individual cases cited to bolster their plausibility, to analysis of the GIS and society nexus, which attempts to trace the complex ways in which each influences the other in specific contexts. While the GIS and society research agenda consistently has pointed to the dialectical relationship between the two, more research effort has been put into the impact of GIS on society than vice versa.

3. *From Society to GIS*

The starting point of this research has been the proposition that, as for technological change in general, the evolution of GIS has followed only a restricted number of the many paths that could have developed (cf. Arthur 1994). On the one hand, research has investigated the evolution of the software and hardware itself, the rise to dominance, for example, of a layer-based GIS over object-oriented GIS. This research has closely examined processes of technological development at key sites and moments when paths of future development were set, and has speculated on the 'paths not taken'; the alternative designs that never got off the ground or never became popular. The roles of key individuals, technical barriers, disciplinary tradition, and broader social forces in shaping paths of development have been analyzed (Chrisman 1988 and Curry 1998). Critical research has also stressed that the practice of GIS includes not just the technology itself, but the whole set of practices through which it is utilized, adapted, and connected with other related activities on a daily basis in the particular places where it is used. Drawing on insights from actor network theory and science studies, this approach has focused on how users, as well as inventors, go beyond learning-by-doing to shape the technology (Harvey and Chrisman 1998).

4. *From GIS to Society*

Critical approaches to examining the societal implications of GIS have focused on case studies examining

how some of the hypothesized implications are realized in particular contexts. A number of the implications raised in critiques of GIS have been examined:

limits to GIS representations of the world;

limits in access to and in the appropriateness of GIS technologies;

legal and ethical implications of GIS use; and

the applicability of GIS for redressing social and geographical inequalities.

Those examining the limits of GIS representations of the world have taken seriously the criticism that GIS, as we know it, undergoes a process of representational stabilization, foregrounding representations which prioritize a Cartesian view of the world, reify geometric spatial objects, and rely on institutional databases for their information. More generally, such criticisms contrast the instrumental rationality of GIS with the communicative rationality of lifeworlds, and conclude that GIS has difficulty capturing the latter. Rundstrom (1995), for example, argues that American Indian conceptions of space are inconsistent with those used in GIS, concluding that it is inappropriate for representing Indian lifeworlds. Harris et al. (1995) experiment with incorporating into conventional GIS sketch maps made by local peasant farmers in South Africa, in order to determine whether the capabilities of GIS can be extended to incorporate lifeworld-based information. The sketch maps represent African farmers' views of the landscape and their roots in local narratives about land alienation under apartheid. Goss (1995) examines GIS-based representations used in geo-demographic marketing, and how these may shape lifeworlds and the places in which they are pursued.

Those examining limits in access to and in the appropriateness of GIS have often begun with the observation that GIS was developed initially for large organizations, universities, corporations, state agencies, and the military, raising questions about its appropriateness for other (more marginal) social actors. Public participation GIS (PPGIS) research has been examining whether GIS, as we know it, is accessible to grassroots organizations; whether it is appropriate to the goals of such organizations, and thus to their empowerment; and whether it enhances or undermines the participatory decision-making processes characteristic of such organizations (Obermeyer 1998). Barriers to the use of GIS by grassroots organizations have been identified, and strategies to overcome these have been suggested. Studies of the impact of GIS use on the empowerment of, and on democracy within such organizations are beginning. Research into legal issues has focused on undesirable social implications of the spreading use of GIS and geo-referenced databases, for which legal remedies may be desirable. It has also focused on how the laws governing information policy and intellectual property shape the social implications of GIS use. Prominent social implications include:

violations of privacy resulting from the capability to map individual actors and events (such as AIDS cases);

the consequences of inappropriate GIS use, and the question of who bears legal responsibility; and

reduced public access to publicly funded spatial databases, as a consequence of intellectual property laws and the cost recovery strategies of local government agencies (cf. Curry 1998 and Sheppard et al. 1999).

Legal questions frequently revolve around ethical issues, but ethical implications of GIS are far broader than this. Curry (1998) suggests not only that there are significant ethical issues revolving around GIS use, but indeed that a close examination of GIS reveals that conventional models of ethical behavior are themselves problematic. The use of GIS to redress social inequalities has been dominated to date by GIS-based analyses of environmental equity (Nyerges and McMaster 1997). Much of this research employs GIS, as we know it, in fairly conventional ways to gather empirical information about the significance, pervasiveness, and correlation of social and geographical inequities in the potential exposure to toxic hazards. Considerable complexities in the relationship between exposure, social characteristics, and location have been revealed, and measures for determining their significance explored. Research into the historical processes behind such patterns has helped unravel the roles of social actors, and of race and class in creating such patterns (Pulido 2000). Some research is beginning to go beyond this, exploring ways to make information about exposure available to communities, and examining the utility of GIS as a tool for neighborhoods to carry out environmental inventories and develop sustainability indicators.

5. *Prospects*

Critical approaches to GIS are an even newer field of endeavor than GIS themselves, although they draw their inspiration from longer standing traditions of critical social theory and critical geography. The growth of the information society, and the central role of GIS and GIS-related information technologies (such as navigation systems, geo-referenced cellular phones, geographical positioning systems, and on-line real time mapping systems) in shaping the geography of the information society, mean that the questions raised in this research will remain important. Yet, the novelty of this field and its considerable potential for growth make it hard to predict major areas for future research. In addition to further elaboration of current research, however, two topics are likely to become increasingly important. First, the path-dependent nature of how GIS developed, and concerns about its limitations, suggest that an important issue, raised in the early 1990s but as yet little examined, is the potential for using digital technologies to convey

geographic knowledge in ways that are very different from those instantiated in GIS as we know it (Harris and Weiner 1996). Possibilities of alternative GISs have been broached, based perhaps on forms of rationality and information and opinion exchange that have more in common with the communicative logic of the Internet than with the instrumental logic of current GIS (Harris and Weiner 1996). Little research has been done, however, examining how such systems might be developed, and whether their social implications are likely to be more emancipatory than current GIS designs.

Second, we know little about reasons for inequalities in access to and use of GIS as yet. While the great majority of GIS researchers and users are white males, there has been little research into why this is the case as yet, and how it might be mitigated. Research taking a critical perspective on GIS has matured from postulating significant potential problems and limitations, to some selected case studies examining actual interrelationships between GIS and society in individual concrete cases. Such case studies are a vital step forward, and much more of this research is necessary. Yet a fuller understanding of these interrelationships will require comparative case study analysis, to tease out which kinds of contextual conditions affect which kinds of outcomes. Research designs should compare both case studies of similar organizational contexts in different places, *inter alia*, and case studies of different organizational contexts in similar places.

See also: Computers and Society; Geographic Information Systems; Spatial Analysis in Geography; Spatial Pattern, Analysis of; Spatial Statistical Methods; Spatial Thinking in the Social Sciences, History of

Bibliography

Arthur W B 1994 *Increasing Returns and Path Dependence in the Economy*. The University of Michigan Press, Ann Arbor, MI

Chrisman N R 1988 The risks of software innovation: A case study of the Harvard Lab. *The American Cartographer* **15**: 291–300

Curry M R 1998 *Digital Places: Living with Geographic Information Technologies*. Routledge, London

Dobson J E 1983 Automated geography. *The Professional Geographer* **35**: 135–43

Goss J 1995 We know who you are and we know where you live: The instrumental rationality of geodemographic systems. *Economic Geography* **71**: 171–88

Habermas J 1984 *The Theory of Communicative Action, Vol. 1: Reason and the Rationalization of Society*. Beacon Press, Boston, MA

Harris T, Weiner D 1996 GIS and society: The social implications of how people, space and environment are represented in GIS. *Report for the Initiative 19 Specialist Meeting, March 2–5, 1996*, Koinonia Retreat Center, South Haven, MN. Technical Report 96–7, National Center for Geographic Information and Analysis, Santa Barbara, CA

Harris T M, Weiner D, Warner T, Levin R 1995 Pursuing social goals through participatory GIS: Redressing South Africa's historical political ecology. In: Pickles J (ed.) *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford, New York

Harvey F, Chrisman N R 1998 Boundary objects and the social construction of GIS technology. *Environment and Planning A* **30**: 1683–694

Nyerges T, McMaster R (eds.) 1997 GIS and environmental risk assessment. A special section of *Cartography and Geographic Information Systems* **24**(3)

Obermeyer N (ed.) 1998 Public participation GIS. A special issue of *Cartography and Geographic Information Systems* **25**(2)

Pickles J (ed.) 1995 *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford, New York

Pickles J 1997 Tool or science? GIS, technoscience, and the theoretical turn. *Annals of the Association of American Geographers* **87**: 363–72

Pickles J 1999 Arguments, debates and dialogues: The GIS-social theory debate and concerns for alternatives. In: Longley P A, Goodchild M, Maguire D, Rhind D (eds.) *Geographical Information Systems: Principles, Techniques, Applications, and Management*. Wiley, New York

Poiker T, Sheppard E (eds.) 1995 GIS and society. A special issue of *Cartography and Geographic Information Systems* **22**(1)

Pulido L 2000 Rethinking environmental racism: White privilege and urban development in Southern California. *Annals of the Association of American Geographers* **90**: 12–40

Rundstrom R A 1995 GIS, indigenous peoples, and epistemological diversity. *Cartography and Geographic Information Systems* **22**: 45–57

Sheppard E 1995 GIS and society: Towards a research agenda. *Cartography and Geographical Information Systems* **22**: 5–16

Sheppard E, Couclelis H, Graham S, Harrington J R, Onsrud H 1999 Geographies of the information society. *International Journal of Geographic Information Science* **13**: 797–823

E. Sheppard

Geographic Learning in Children

1. Early Childhood Development and Geographic Learning

Most research on early childhood geographic learning has been concerned with the ability of the developing child to understand space, and in particular, to comprehend spatial location. Children begin to be capable of spatial perception—the perception of small-scale spaces they can see at once—shortly after birth. Not long thereafter, they begin to crawl around, and slightly later to toddle through rooms filled with obstacles that impede perception. It is at these slightly later ages that geographic learning really begins, when a child goes from merely perceiving to actively searching the environment. Between 12 and 18 months

of age, children can find objects once seen and now hidden (by a screen, for example) from view. Children between 18 months and two years of age can also deal with 'invisible displacements,' in which a toy, for example, is not just concealed but 'invisibly' transported to another hiding place. Two- and three-year olds can carry out much more complicated searches. However, at these age levels most research has been limited to the scale of an ordinary room, or slightly larger (Matthews 1992).

There is now strong evidence that children acquire basic map-reading skills and the ability to engage in certain aspects of map-reading somewhere between the ages of three and five (see Blades and Spencer 1990, DeLoache 1989, Newcombe and Huttenlocher 1992). Studies have focused on two general levels of spatial ability: (a) micro- or meso-spatial contexts—i.e., table-scale or room-scale representations (e.g., Cohen 1985), and (b) macro- or geographical-scale environments. Micro-spatial cognition involves understanding spatial relationships among objects that can, for the most part, be directly perceived as themselves and at their own scale. For example, micro-spatial cognition is required to put together a jigsaw puzzle. Macro-spatial cognition, on the other hand, involves representations at a perceptual scale of objects and object arrays that occupy much larger spaces in the 'real world,' objects that, together, would not ordinarily be perceived at once on the ground, for example, reading a map to navigate through a city.

The most interesting aspects of spatial development and geographic learning concern the ability of children to understand macro-spatial environments, which are large in extent, and their representation in the form of maps (Blades and Spencer 1990). Almost all research in this area has involved miniature representations of such environments in the form of aerial photos, maps, and modeled landscapes. At issue is how to ascertain very young children's understanding of the connection between these miniature representations and full-scale geographic reality. Children younger than five have limited vocabularies, and efforts to use verbal responses as indices of this understanding have not been very successful.

However, ethnographic studies done in several cultures have begun to document the universality of the use of toys or natural objects to model large-scale environments (Blaut 1999, Stea et al. 1996). It has been noted that some three-and-a-half-year-old children are able to build model landscapes and to solve navigation problems on these landscapes. Such navigation problems involve getting from one place to another on the map in a 'realistic' manner, that is, by a route on the representation that would be reasonable in the 'real world'. However, a series of studies have found that while some spatial skills (micro-spatial puzzle and block-building ability; macro-spatial navigation ability) are present as young as age three-and-a-half, they do not necessarily appear and develop at the same rate. This indicates that while spatial skills are related, they likely reflect different dimensions of cognitive development. Beyond age four-and-a-half, a majority of children studied appear capable of the above tasks, including navigation.

Studies done with children in a number of cultures have concluded that five-year-olds can verbally identify objects in an aerial photograph at scales of from 1:1000 to 1:3000 with minimal errors (although attempts to identify landscape objects at scales larger than 1:10,000 do result in some errors of scale). Further, such children can trace 'maps' from these photos, and can use these maps to solve simple navigation problems.

Slightly older children, in kindergarten through second grade, are capable of drawing quite complex maps. While these maps are plan views, overall, children typically draw the elements of such maps in elevation rather than in plan (e.g. house facades instead of house roofs), leading to the conclusion on the part of some psychologists that these children are unable to 'coordinate perspectives'. However, when seven-year-olds are asked to tell why they have drawn their maps in that way, they show full recognition that theirs is not a literal representation of an 'airplane view' of the world. On the contrary, they say that they are purposely trying to show what's most interesting about each object: a house facade is obviously more interesting than the view of the same house from above. That this is not simply a failure to coordinate perspectives is further indicated by the fact that many maps drawn by adults in the Middle Ages had the same characteristics.

2. *Theoretical Controversies Concerning Spatial Aspects of Geographical Learning*

In determining when children are ready for the introduction of geographic concepts in school, geographic education has been more concerned with the sequence of children's development than with their learning, in a psychological sense. The dominant theoretical paradigm in geographic education is that provided by the noted theorist and researcher in child development, Jean Piaget. In Piaget's view, all cognitive development, including spatial cognition, goes from lower to higher levels of conceptual and symbolic functioning, part of a more general set of competencies acquired in a stagewise, lock-step fashion (see Piaget and Inhelder 1956). Thus, such development is said to progress stepwise through four stages, which he termed sensorimotor, pre-operational, concrete operational, and formal operational, loosely associated with certain age ranges. Children at the sensorimotor and preoperational stages view the environment in relation to themselves and their own actions. Thus, according to this developmental view, children begin to be capable of true geographic learning only upon

reaching the concrete operational stage of development (Liben and Downs 1989). While Piaget stressed that his developmental sequence is to be measured in 'stages, not ages,' some influential educators, more comfortable with ages, have proposed that map concepts cannot be understood by children younger than seven.

An alternative theoretical perspective is termed 'universal mapping' (Stea et al. 1996). Universal mapping, while not directly challenging the general concept of developmental stages, suggests that the kinds of spatial skills that eventually result in the production and use of geographical maps transcend cultural differences and begin to appear very early in the life of the child. Universal mapping states that, throughout the development of the human species, people have had a need to construct mental representations of the environment at a geographical scale. Comprehension of maps requires perspective taking ability, in particular, moving from a three-dimensional real-world perspective to a two-dimensional, micro-level perspective. Thus, the theory proposes that young children have a 'natural' ability to represent the world in symbolic form at a reduced scale, rotated to the overhead perspective. Some proponents of universal mapping theory in its strongest form suggest that children may have an inborn mapping acquisition device analogous to the language acquisition device proposed by the linguistic theorist Noam Chomsky.

There is still some question as to whether and in what way mapping skills and spatial development are related. Piagetian theory has been interpreted by many educators to imply that children are incapable of engaging in the form of geographic learning called mapping at school-entering age. Universal mapping suggests, to the contrary, that some geographic abilities are so fundamental that their mastery by preschool children can form a base for the teaching of other subjects in school.

Deciding between these two theoretical positions has been made more difficult by marked differences in methodologies used by their respective proponents. Studies have differed in dependence on verbal or nonverbal responses by children, interpretations of errors made in verbal responding, degree of generalization from results obtained in meso-scale (room-level) environments to macro-scale (geographical) environments, and widely divergent scales of material used in aerial photos studies, to name but a few. At present, it seems that while cognitive development certainly takes place, and probably in stages, this staged sequence is neither quite as stepwise nor as irreversible as originally proposed. It may be that transition between stages occurs earlier in the development of spatial abilities at the geographic scale than in other cognitive realms, indicating its importance to both human development in general and geographic learning in particular.

3. *Geographic Learning, Geographic Education, and Future Directions*

As children begin to master perspective taking and part–whole relations, they should become more adept at attending to elements that will aid in their learning map information. However, while learning about space at the geographical scale is one important aspect of geographical learning, learning about other characteristics of geographical environment, including its flora, fauna, landscapes, and geology, is also important. Thus, learning to map, while in many ways the keystone of geographic learning, is not all of it. With increased demand in US school systems to develop a systematic program of geography instruction, there is a need for standardized, cost-effective methods for delivering this information. Instruction via electronic media (e.g., videotapes, computer programs) can be both appealing and educational to young viewers. The success of such television shows as *Where in the World is Carmen San Diego*? provides one demonstration that programs can be produced which capture children's attention and at the same time stretch the boundaries of their geographic knowledge.

Geography instruction at the primary school level usually involves, at least in part, the comprehension of pictographic information, in particular, learning the spatial locations and boundaries of continents and geo-political regions such as countries and states. Because of its inherent visual and spatial content, geography lends itself well to instruction via electronic media. In fact, the types of visual and spatial content that are included in televised instruction can determine how readily that information is encoded and later retrieved from memory. For example, the addition of unique visual images can serve as a mnemonic aid for learning geographic shapes. Recent studies have demonstrated that visual icons, such as the Blue Ridge Mountains in the shape of Virginia, provided strong recall cues for state shapes (see Piñon 1999). Similar studies have shown the facilitative effects of icons and imagery in recalling map content (see Webb et al. 1994).

See also: Cartography; Cognitive Maps; Geographic Education; Knowledge (Explicit and Implicit): Philosophical Aspects; Navigation in Spatial Environments, Neural Basis of; Space: Linguistic Expression; Spatial Cognition

Bibliography

Blades M, Spencer C 1990 The development of 3- to 6-year-olds' map using ability: the relative importance of landmarks and map alignment. *Journal of Genetic Psychology* **151**: 181–94

Blaut J M 1999 Maps and spaces. *The Professional Geographer* **51**: 510–15

Cohen R (ed.) 1985 *The Development of Spatial Cognition*. Erlbaum Associates, Hillsdale, NJ

DeLoache J 1989 The development of representation in very young children. *Advanced in Child Development and Behavior* **22**: 1–39
Liben L S, Downs R M 1989 Understanding maps as symbols: the development of maps concepts in children. *Advances in Child Development and Behavior* **22**: 145–201
Matthews M H 1992 *Making Sense of Place*. Barnes & Noble, Lanham, MD
Newcombe N, Huttenlocher J 1992 Children's early ability to solve perspective-taking problems. *Developmental Psychology* **28**: 635–43
Piaget J, Inhelder B 1956 *The Child's Conception of Space* [trans. Langdon J F, Lunzer J L], Routledge, London
Piñon M F 1999 Video imagery and children's recall of televised geographic information: Affecting more than visual content. *Journal of Instructional Psychology* **26**: 24–32
Stea D, Blaut J M, Stephens J 1996 Mapping as a cultural universal. In: Portugali J (ed.) *The Construction of Cognitive Maps*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 345–60
Webb J M, Saltz E D, McCarthy M T, Kealy W A 1994 Conjoint influence of maps and auded prose on children's retrieval of instruction. *Journal of Experimental Education* **62**: 195–208

D. Stea

Geographic Thought, History of

1. General Conditions

1.1 Vernacular, Corporate, and Scientific Geographies

As a scientific discipline, geography has a long history. Its practice goes back to the time of Herodotus (484–425 BC) in the 5th century BC. The term itself was used in *De Mundo* attributed to Aristotle in the 4th century BC. Geography deals with orientation, the location of places, the surface of the Earth, the features which characterize it, and the way people conceive and experience places, the Earth, and the Cosmos. It means that geography existed as a practice of daily life before the development of scientific thought. People had to find places, evaluate the value of soils and the best way to farm them, detect and exploit mineral ores, etc. Many of these vernacular practices and knowledge coexisted simultaneously with the formal development of the discipline.

Since geography gives information on distant places, the ways to reach them, their economy, population, and culture, it is important for rulers. Governments rely either on empirical knowledge accumulated by local persons or on objective evidence gathered by their cartographic, statistical, and ethnographic services for the management of a territory. In the first case, the local agents of the sovereign know the country they are in charge of so well that they can develop policies of their own. Hence, there is a tendency for all governments to develop centralized forms of control, for which they need special forms of geographic knowledge. Modern enterprises also have to develop geographic skills. They have to locate resources, exploit them, and distribute goods. A part of modern geography is linked with their activity.

The history of scientific geographic thought has always been situated in this context, with an eye to the type of practices used by ordinary people as well as to the skills developed by administrative services in gathering and interpreting spatial data. The discipline was institutionalized for several reasons:

(a) the will to provide the state (or other power institutions) with valuable knowledge for domestic control or overseas expansion;

(b) a curiosity for the low culture forms of geographic lore and their preservation; and

(c) the will to provide the citizens of modern nation states with an image of their country and the rest of the world in order to build or strengthen their national identity and explain how the contemporary world is working.

1.2 Geographic Narratives and the History of Geographic Thought

Geographic thought was expressed through standardized forms of narratives. The history of the discipline is partly that of the transformation of geographic discursive fields. What were the publics aimed at, the rhetoric used to attract or convince them, and the type of knowledge judged convenient for them? In Greek and Roman times for instance, according to one of the established geographic genres, geography was often presented as the account of a journey by Icarus.

From the Renaissance onwards, geography developed, as narratives on the newly discovered lands and the strangeness of their inhabitants, universal geographies, i.e., pictures of all the places and countries of the Earth, or quantitative assessments of the population, activities, wealth, and military assets of the provinces of kingdoms in order to facilitate their control. During the nineteenth and twentieth centuries, geographers developed other types of narratives: the description of exotic places for tourists and the analysis of national territories for the civic instruction of young people.

Geography is a complex discipline. Its changes are partly explained by the ways it was institutionalized and partly by the evolution of the discursive fields geographers developed in order to gain a large readership. However, its transformation mainly reflects the dynamics of its scientific purposes. It is convenient to distinguish five different scientific programs that were transformed into traditions later.

2. The Five Approaches in Geography

2.1 Geography as a Science of Terrestrial Locations. Cartography

The first program was linked with the problems of orientation. In order to build a reliable system of place locations, Greek geographers discovered that the best solution was to rely on an abstract grid of reference (Jacob 1991). The idea appeared with the first forms of mapmaking. For the Ionian mapmakers in the late sixth and early fifth centuries BC, all territories had geometric shapes. The idea of spherical coordinates, which appeared in the third century BC, was conducive to the first scientific cartography. At the same time, it provided a coherent grid of reference for locating places. As expressed by Ptolemy (AD 98–168), geography became 'the sublime science which reads in the sky the image of the Earth.'

The Earth is a sphere and terrestrial positions, i.e., longitudes and latitudes, can be measured through the observation of the Sun and stars. Erastothenes (273–192 BC) was the first to develop this line of research. Hipparchus (second century BC) proposed the idea of geographic coordinates. In the second century AD, Ptolemy synthesized the accumulated data in this field. His contribution was preserved and improved by Arabic geographers during the Middle Ages. The rediscovery of his texts in the fifteenth century marked a new start in the history of geographic thought. In the seventeenth century, Bernhardus Varenius (1622–1650) modernized the cosmographic approach through the introduction of the Copernician view of a Sun-centered Universe.

The major problem for this conception of geography came from the impossibility of keeping time, so that astronomical observations were unable to give a measure of longitude—a weakness which subsisted from Antiquity to the discovery of the marine chronometer by John Harrison (1693–1776) in 1735. Until the mid-eighteenth century, one of the main tasks of geographers was to compile data relative to distances in travel accounts. From the mid-eighteenth century, mapmaking became the task of specialized engineers and scientists. They provided geographers with the topographical maps they needed.

Geographers ceased to be responsible for the basic cartography, but the problems of location and mapmaking remained important for them. Their role changed; they learnt in the nineteenth century to use statistical sources to produce thematic maps. During the twentieth century, remote sensing facilitated the gathering of data on land use, vegetation, temperatures, atmosphere, etc. Thanks to computing facilities, the building of geographical information systems has recently become easy; one of the most prized skills of contemporary geographers is to provide administrations or firms with this new cartographic aid to decision-making.

Thanks to cartography, geographers were the first able to describe the Earth as seen from the sky—the Icarus' viewpoint, according to Greek geographers. It was essential for getting an objective description of the face of the Earth and its regional divisions.

2.2 Geography as Description. The Regional Approach

The second agenda developed by geographers was the description of the Earth's surface and its regional differentiation. During Antiquity, it was illustrated by Herodotus, Marinus of Tyre (first or second century AD) and Strabo (64 BC–AD 20). Arab geographers improved much of the knowledge of the Middle East, India, the Far East, and Africa south of the Sahara. With the great discoveries, the program became really worldwide.

The main problem was: What to describe and how to describe it? Thanks to the map and the vertical perspective it offered, geographic description was different from the travelers'; it was not structured along itineraries, but on the apprehension of regional units as seen from above. Such a perspective allowed for the contraction of much information, but it destroyed the link between direct observation and description, hence the dryness of many descriptions until the eighteenth century.

Because there was no standardized vocabulary to speak about the details of topography, rocks, plants, animals, and human activities, geographers were unable to grasp the real world in their narratives. Sometimes, they preferred to provide their readers with information on local mythology (in the Greek and Roman times), the history of Christianity (during the Middle Ages), and the local coats of arms and heraldry (in feudal Europe). In a way, the travel accounts of the Middle Ages and Renaissance are generally more interesting for today's geographers than the geographical publications of the time.

Everything changed in the eighteenth century, with the development of natural history and a new curiosity for human activity. The work of Linné in botany and zoology, a few geologists for rocks, and the first agronomists for rural life, created the technical terms geographers needed in order to write precise descriptions of the world. Thanks to Jean-Jacques Rousseau (1712–1778) and Johann-Heinrich Pestalozzi (1746–1827), geographers began to offer descriptions on two scales: one was based on both fieldwork and direct observation, and the other on the synthetic vision of the map-reader. Carl Ritter (1779–1859), one of the pioneers of modern geography, has been trained in a Pestalozzian school, just as Elisée Reclus (1830–1905). Alexander von Humboldt (1779–1859) discovered the value of first-hand contact with landscapes through his training as a natural scientist and his contacts with

Georg Forster (1754–1794), who was on Cook's second voyage around the world.

The quality of geographic description kept increasing with the development of new scientific fields: pedology, meteorology, and climatology on the natural side of the discipline, and agronomy, demography, and social and cultural anthropology on the human one. Many of these research fields were initially explored by geographers and then turned independent.

Because of the improvement of geographical descriptions, geographers conceived their discipline in a different way: they began to consider that one of its major advantages was to deal with the real world. The idea that geography was a science of concreteness appeared. Geographers did not consider themselves only as scientists; description was an art, so that they had to display a literary bent.

During all of the nineteenth century and a good part of the twentieth, when writing descriptions, geographers felt to provide impartial views of the world since their work was based on objective observations. They considered that the descriptions and modes of regionalization developed by laymen were subjective and as such, not reliable. They preferred to ignore them.

2.3 Geography as Explanation: The Perspective of Natural Philosophies and Natural Sciences

As soon as description progressed in the eighteenth century, a third task appeared for geographers: they had to explain the distributions they observed. A precise description was necessarily conducive to questions about the causes responsible for the patterns read on the maps or the features discovered through fieldwork. Geographers had to give an explanatory dimension to their analyses.

During the first half of the nineteenth century, the main interrogation was about the interaction of natural features, their mutual conditioning, and their influence on the destiny of human groups. In the seventeenth century, Varenius had opened the way through his analysis of position. In order to explain terrestrial distributions, he stressed the role of latitude and of the major topographic accidents. He opposed inland and coastal locations. In the first half of the nineteenth century, this approach was transformed into a coherent paradigm: the analysis of position, well exemplified by Alexander von Humboldt and systematized by Carl Ritter. The major factors in the shaping of the Earth and the explanation of regional distributions were latitude, altitude, continental versus coastal locations, and the patterns of mountain ranges and coastlines. This kind of analysis had to be developed at different scales.

Geographers used these tools to explain the distribution of particular natural features: the altitudinal patterns of vegetation in tropical mountains for Alexander von Humboldt. Ritter wished to explain how the destiny of peoples was shaped by the countries they inhabited. Behind these projects, the influence of early nineteenth century philosophies was obvious. Humboldt stressed the harmony between the different levels of the geographic reality according to the natural philosophies of his time. Ritter followed Herder (1744–1803) when exploring the influence of environment on cultures.

This first type of explanation, linked with the German philosophies of late eighteenth and early nineteenth centuries, was displaced by a natural science perspective when Charles Darwin (1809–1882) stressed the role of the environment in the transformation of life. Geographers of the late nineteenth century, Friedrich Ratzel (1844–1904) in Germany and Paul Vidal de la Blache (1845–1918) in France, built human geography on this evolutionist, either Darwinist or neo-Lamarckian basis. Their ambition was to explain human distributions through the analysis of the mainly 'vertical' (or local) relations people had with their milieu. In fact, they combined the analysis of position inherited from Ritter with an ecological approach. They stressed both the significance of local environments (the ecological approach) and of circulation (the analysis of position). The risk with the ecological approach was to accept a deterministic interpretation of human beings/nature relationships. Circulation offered a way to escape this danger.

Vidal de la Blache used the idea of *genre de vie* as the main concept of the new approach. In order to exploit an environment, people developed sets of techniques and practices that were soon transformed into a way of life because of the force of habits. The main interest of the *genre de vie* notion was to give a social dimension (the force of habits) to human geography and to stress the role of techniques in the mediation between human groups and the environment.

Even if Ratzel and Vidal de la Blache were conscious of the significance of circulation, their human geography tried mainly to explain the influence of physical geography on human distributions and the way people took advantage of natural endowments. The role of social relations and of distance as an obstacle to interaction was practically ignored. Human geography was at its best when it used the *genre de vie* paradigm to explain regional structures through the interaction between physical forces and social groups. Unfortunately, this approach could not work in urbanized and industrialized societies, where *genre de vie* became too complex to be a useful means of investigation.

Because human geography was unable to explain satisfactorily the distribution of social groups and artifacts on the surface of the Earth, some geographers, specially in Germany, preferred to define geography as the science of landscape, which avoided splitting the discipline between natural and social parts.

The analysis of position played a central role in the first efforts to investigate political geography. Alfred T. Mahan (1840–1914) and Halford Mackinder (1861–1947) stressed the opposition between inland locations and coastal ones. This theme became central to geopolitical studies in the first half of the twentieth century.

Thanks to the natural sciences, an explanatory dimension had been introduced in geography. Geographers succeeded in organizing efficiently physical systematic disciplines: geomorphology, climatology, biogeography, pedology, hydrology, oceanography, etc. On the human side of the discipline, the results were less impressive. Geographers capitalized more on the analysis of position than on the ecological paradigm. Even cultural geographers like Carl Sauer (1890–1975) did not try to explore systematically the role of techniques in the distribution of human groups. By the mid-twentieth century, a feeling of dissatisfaction had developed in human geography.

2.4 Geography as Explanation: The Social Science Approach

Many geographers tried to give a new start to their discipline. The most successful was Edward Ullman (1912–1976). His diagnosis was clear. Human geography failed because it was conceived as a natural science. Overemphasis had been given to the study of human beings/milieu relationships. Circulation, the other face of human geography as defined at the end of the nineteenth century, had been neglected. In order to build a modern geography, emphasis had to be given to social interaction (Ullman 1980).

Geographers discovered that economists had developed an interest for spatial economics since the beginning of the nineteenth century. They had explored the influence of distance upon the location of farming, industrial activities, and services. Geographers decided to take advantage of the results of spatial economics to give a more systematic dimension to human geography. It was the basis of the new geography of the late 1950s and 1960s.

In a way, this new geography translated Ritter's position into social terms. Instead of focusing on the opposition between inland and coastal locations, for instance, it focused on the dialectics between center and periphery. The center was defined in terms of social physics as the point or the area that enjoyed the highest accessibility to the totality of a social group. Procedures were invented to measure centrality and to map it. The idea of population or income potential changed the traditional vision. Instead of the center as a point, it presented the center as an area. In the United States, a central core, the manufacturing or industrial belt, appeared as the main feature of the territorial structure.

Geographers relied on spatial economics to develop their new understanding of human geography, but they went far ahead of economists in their analysis of spatial organization. The regional geography of the first half of the twentieth century fell short of a systematic formulation. Thanks to the works of Edward Ullman, Torsten Hägerstrand (born 1916) and a few other geographers, the way space was organized—through specialization, the development of networks, the dynamics of scale and external economies, and the opposition between center and periphery—became clear.

During the 1970s and 1980s, a few geographers, for instance, Paul Claval (born 1932), tried to extend the analysis of spatial interaction to non-economic aspects of social life, either social relations or political structures. Instead of focusing on transport costs, they analyzed communication, the structuring influence of commutation costs, the role of medias, and the significance of symbolism. In this way, they provided a new understanding of the dynamics of globalization, so evident from the 1960s.

The emphasis on social interaction had positive effects on geography in many respects, but it reduced the interest in the ecological approach at a time when ecology came really of age and pollution became a major problem at the local, regional, and global levels.

2.5 Geography and Postmodernity: The Cultural Approach and the Experience of Nature, Place, and Society

From the early 1970s, a strong critical movement developed in geography. Young colleagues struggled against the conception of their discipline as a natural science and the interpretation that was given to its social dimension. The positivist approach considered social relations as natural facts. It ignored their historical dimensions and their creation through human agency. As a result, neo-positivist geographers did not criticize the hierarchical structures and spatial organizations of the groups they observed and explained, even if they were conducive to exploitation and oppression.

From the eighteenth century, a different conception of geography had developed. In his *Géographie politique*, Turgot (1727–1781) wished to promote new geographies much more than to explain the existing one (Turgot 1808–1811). It was one of the main features of the geographic projects of the Enlightenment. The critical tradition was represented during the nineteenth century by anarchists like Elisée Reclus and Peter Kropotkin (1842–1921), who made it illustrious, but the prevailing option for a natural science approach restricted their influence.

At the beginning of the 1970s, David Harvey (born 1935) played a decisive role in the development of a new interest for social justice (Harvey 1973). However,

the break with the neo-positivist positions of the 1950s and 1960s had other correlates. If social realities are the product of human agency, it is urgent to study the ways people experience places and conceive nature and society, the dreams they have about their future, and the significance they give to their lives. Geographers can no more ignore the ways people organize their existence. The rigid division between vernacular or corporate geographies on the one side, and scientific ones on the other, fades out. Scientific geographers have to explore the geographies of lay people in order to understand the landscape they shape and the institutions they create. Ethnogeography becomes a new field of investigation.

There was an old but limited interest in people's conceptions of place thanks to the analysis of choronyms. The development of new attitudes among geographers had not been possible without the influence of Heidegger (1889–1979). Eric Dardel (1902–1968), in France, was well aware of the significance of his philosophy, as Yi-Fu Tuan (born 1930) or Edward Relph (born 1944) a generation later.

When societies are conceived as a product of human activity, i.e., in an historical perspective, it becomes important to understand the practices, attitudes, knowledge, and beliefs shared by social groups, the techniques they master in the fields of material activity and social life, and the significance they give to their lives. The cultural approach, which had been mainly developed by Carl Sauer in the first half of the twentieth century and had experienced a crisis in the 1960s and early 1970s, benefited from a new popularity thanks to Denis Cosgrove (born 1948), James S. Duncan (born 1945), Joël Bonnemaison (1940–1997), and Augustin Berque (born 1942). It was one of the main frontiers in human geography from 1980 onwards. Culture ceased to be considered as a collective heritage. It was analyzed as a continuous process of communication, adaptation, evaluation, and refurbishing.

The criticism of the neo-positivist conceptions of social sciences was by no means restricted to geography. The transformations of the discipline were parallel with new orientations in sociology, anthropology, and history. Geographers discovered the interest of the Clifford Geertz's type of thick description as a tool to explore ethnogeographies and enter the complexities of cultures.

Under the aegis of French deconstructionist philosophers, Jacques Derrida (born 1930) and Michel Foucault (1926–1984) for instance, social knowledge was increasingly analyzed as a tool to legitimize the Establishment and gain support for governmental policies, whatever their content. Geographers began to scrutinize geographical narratives and look for their hidden significance. They got increasingly aware of the role of geography as an instrument to build efficient forms of social control. Thanks to Derek Gregory (born 1951), the study of geographical imaginations and of their role in social life and power situations became popular (Gregory 1994).

The new epistemologies are increasingly critical of the great narratives of the past. Geographers are reluctant to rely on macro-theories or philosophies–historical materialism for instance. They develop a growing interest in the local and privileged micro- or meso-theories. They explore realms that had been previously neglected: gender geography, the geography of young and aged people, the varied forms of segregation, and exclusion or sexual differences that most societies build.

Men and women do not live in a standardized, sterilized, and acclimatized space. They inhabit specific places, develop territorial identities, and use landscape markers as symbols of what unite them and provide them with a sense of shared community. Consequently, geographers are increasingly concerned with the religions, philosophies, or ideologies people create in order to give significance to their lives. They explore the moves the human mind performs in order to discover perspectives of society, nature, and collective destiny, and build values. Transcendence and immanency are wrongly considered as purely philosophical concepts. They have a spatial dimension; geographers have to explore them.

Men and women are part of nature; they live in nature, with nature (Harvey 1996). The ecological perspective has been completely recovered. Geographers are increasingly involved in the study of the energy and material relations which develop between human beings and their environment, and in the analysis of the attitudes towards nature, the significance given to natural landscape, and the management of the planet Earth.

Contemporary geography goes much deeper into the analysis of social and psychological processes than was the case in the 1970s. At the same time, it has become a more critical and modest discipline. For many of the young radicals, scholars have to refrain from building mega-narratives, as in the past. It is better to concentrate on limited problems.

3. *The Coexistence of Perspectives Instead of the Displacement of Paradigms*

Kuhn's interpretation of scientific revolutions was very popular among geographers in the 1970s. They interpreted the new geography of the 1960s as a new paradigm. In fact, the history of geographic thought does not fit well into the Kuhnian model. We have distinguished five traditions, but when a new one appeared, it did not displace the older ones. The interest in location and map making is less central today than in Antiquity or modern times, but it is still fundamental to the discipline. Description as an intellectual venture worth pursuing for its own sake has long ceased to be fashionable, but progress in the

analytical instruments offered by other natural or social sciences allows for new readings of landscapes and maps, and the discovery of new patterns. It makes geographers aware of new problems.

Geography as a natural science and geography as a positivist social sciences are still alive. They provided geographers with many of the tools they use, the analysis of position both relative to physical factors along Ritter's views, or social distributions according to Ullman's conceptions, and the ecological approach to human beings/milieu relationships. Many colleagues try to develop more performing models of economic or social distributions. The emphasis on communication is a new development of an old curiosity.

When building a periodization of the history of geographic thought, the development, fading out, and renewal of geographical traditions constitute only one component of the whole picture.

Greek geography was born out a Kuhnian revolution, the invention of the grid of meridians and parallels, but its successes came partly from its role as a repository of Greek culture. With the development of empires, first in Alexandria and then in Rome, it became a useful tool for the new bureaucracies because of the maps it produced and the descriptions it provided.

The decline of the Classical tradition in the Middle Ages reflected the disappearance of imperial polities in Western Europe and their displacement by feudal structures, which did not involve explicit corporate geographies. Maps were used as symbols of the monarch's power, not as a tool for his or her action. Because of the development of trade from the eleventh century, new cartographic and geographic instruments were needed—hence the invention of Portolano charts. In the Arab world where imperial polities played an important role, the fate of medieval geography was far better.

The new phase of development of geography which started from the Renaissance was linked with the rediscovery of Ptolemy and the cartographic perspective, the voyages of discovery, the rise of the modern state with its proto-bureaucracies, and new forms of corporate economic activities. For two centuries, geography remained based on the two perspectives of cartography and regional description.

The eighteenth century was as a major turning point. The cartographic limitations disappeared thanks to an easier measure of longitudes. Because of the progress of natural sciences and the birth of social sciences, geographic description improved much and ceased to be problematical. This kind of description necessarily led to an interest in explanation. The transition from pre-modern to modern forms of geography was a difficult one and lasted for about a century, since it meant a change in the institutionalization of the discipline: from state cartographic services to universities.

The naturalist orientation that was born at the beginning of the nineteenth century, prevailed until the mid-twentieth. From that time, geography became first a neo-positivist social science, and then a discipline, partly scientific, partly humanistic, or philosophical, based on human experience and the significance of the human presence on Earth. Institutionalization began once again to change. Geography is increasingly dependant on research institutions and private funding.

Geography offers the intellectual tools that are needed to understand the contemporary acceleration of globalization. It explains why technical uniformity is conducive to new forms of regional organization and does not entail the destruction of all spatial differentiation.

What will be the future of the discipline? As in the past, a part of the specialty fields it nurtured will develop as fully-fledged sciences of their own. The technicalities of contemporary physical geographies will deepen the gap between them and human geography. The future of geography as a unitary discipline is perhaps more on the reflective side than on the analytical side of human knowledge.

Bibliography

Berdoulay V 1981 *La Formation de l'Ecole française de géographie*. Bibliothèque nationale, Paris

Capel H 1981 *Filosofia y ciencia en la Geografia contemporanea. Una introducciòn a la Geografia*. Barcanova, Barcelona, Spain

Claval P 1998 *Histoire de la géographie française de 1870 à nos jours*. Nathan, Paris

Dunbar G 1991 *Modern Geography. An Encyclopedic Survey*. Garland, New York

Godlewska A 1999 *Geography Unbound. French Geographic Science from Cassini to Humboldt*. The University of Chicago Press, Chicago

Gould P 1985 *The Geographer at Work*. Routledge & Kegan Paul, London

Gregory D 1978 *Ideology, Science and Human Geography*. Hutchinson, London

Gregory D 1994 *Geographical Imaginations*. Blackwell, Oxford, UK

Hartshorne R 1939 *The Nature of Geography*. The Association of American Geographers, Lancaster, PA

Harvey D 1973 *Social Justice and the City*. Arnold, London

Harvey D 1996 *Justice, Nature and the Geography of Difference*. Blackwell, Oxford, UK

Holt-Jensen A 1988 *Geography. History and Concepts*. Paul Chapman, London

Jacob C 1991 *Géographie et ethnographie en Grèce ancienne*. A. Colin, Paris

James P 1972 *All Possible Worlds. A History of Geographic Ideas*. Bobbs-Merrill, Indianapolis, IN

Livingstone D N 1992 *The Geographical Traditions*. Blackwell, Oxford, UK

Schultz H D 1980 *Die deutschsprachige Geographie von 1800 bis 1970*. Selbstverlag des Geographischen Instituts der Freien Universität Berlin, Berlin

Stoddart D R 1986 *On Geography*. Blackwell, Oxford, UK
Turgot A R 1808–1811 Sur la géographie politique. In: Dupont de Nemoun P S (ed.) *Plans et projets d'ouvrages, Oeuvres completes*, Paris
Ullman E L 1980 *Geography as Spatial Interaction*. University of Washington Press, Seattle, WA

P. Claval

Geography

The practice of geography—literally 'earth writing,' from the Greek *geo* and *graphein*—goes back at least 2000 years, but it has a much shorter history as an academic discipline, with little formal presence in universities before the twentieth century. There have been many attempts to define it: one widely accepted in the mid-twentieth century saw it 'concerned to provide accurate, orderly, and rational description and interpretation of the variable character of the Earth surface' (Hartshorne 1939, p. 21). Its focus remains areal differentiation, but it has moved beyond cataloging and description to *analyzing and providing explanatory accounts for the nature and implications of the material character of the earth's surface and its use by people and societies*. Its key concepts are space, place, and environment, each studied at a variety of scales.

Although most definitions and university departments of geography embrace both physical and human worlds, studies of the two are now substantially separate components of the discipline. This article focuses on human geography, that part normally affiliated to the social and behavioral sciences.

1. Intellectual Origins

As a source of information about the earth, geography has been important to all societies. It was part of the curriculum at Cambridge and Oxford Universities by the late sixteenth century, for example, taught for its practical value with two main themes. The provision of locational information involved cartography, on which a wealth of early material in many cultures has been identified. Alongside was chorology, the description of physical environments and their inhabitants. Geographers portrayed the different parts of the world, meeting the needs of traders, diplomats, and the military and also whetting the curiosity of the 'educated classes.'

Although geographical material was widely disseminated and used, there was no formal discipline with institutional status in universities. This was countered in the nineteenth century by societies founded to promote the dissemination of geographical knowledge (Livingstone 1992): many were closely linked to imperialist projects and actively promoted exploration of extra-European *terrae incognitae*. The first university chairs and departments were established by royal decree in Germany in the 1870s; they were to prepare school teachers, as part of the creation of an educated citizenry, and to sustain colonial and military enterprises. Similar successful campaigns elsewhere included the Netherlands, where the first established chair was in 'colonial geography,' and in England, where the Royal Geographical Society provided financial assistance to establish departments at Oxford and Cambridge, although the first honours degrees in geography were awarded at Liverpool, in 1917. The departments were small but provided the foundation for substantial expansion after World War II, when there was a geography department in nearly every British university and a graduate school in 27 United States' universities.

2. The Creation of a Discipline

The subject matter taught in these new departments focused on the inter-relations among peoples and environments, as conceived by several important German and French geographers (such as von Humboldt, Ritter, Ratzel, and de la Blache: Livingstone 1992). Geographers sought a unique intellectual rationale within the academic division of labor, increasingly so as universities adopted the German tradition of marrying teaching with research. Of the six trends that Freeman (1961) identified up to c.1950 (encyclopedic, educational, colonial, generalizing, political, and specialization), the generalizing trend became increasingly important, with individual specialization as a necessary consequence.

The emerging practice of geography was strongly influenced by its milieux (Livingstone 1992). Debates in the first half of the twentieth century crystallized around the concept of the region. Hartshorne (1939, p. 462) presented geography as:

> … a science that interprets the realities of areal differentiation of the world as they are found, not only in terms of the differences in certain things from place to place, but also in terms of the total combination of phenomena in each place, different from those at every other place.

Regions possess unique characteristics based on their inter-related physical and human environments, and regional geography aimed (Wooldridge 1956, p. 53):

> … to gather up the disparate strands of the systematic studies, the geographical aspects of other disciplines, into a coherent and focused unity, to see nature and nurture, physique and personality as closely related and interdependent elements in specific regions.

This identifies two long-recognized strands in geographical practice: the systematic, focusing on particular subject matter, and the regional, concentrating on specific areas.

There were common elements to geographical practice in those countries where it was established early:

Hartshorne's 1939 essay exemplified the strong German influence on American geography, for example, and much regional geography was based on classic French work. The links between the different language realms weakened after World War II, however (Johnston and Claval 1984): the remainder of this article focuses on the English-speaking world.

2.1 A 'New' Geography

Systematic specialization increasingly dominated the expanding discipline in the mid-twentieth century, with two significant intradisciplinary divides: a split between physical and human geography: and the creation of separate specialisms within each. Regional geography declined in importance as a research field, but remained important in the discipline's teaching. The early systematic strengths were in historical geography, particularly strong in the UK (led by Darby and his major work on the Domesday Book: Darby 1977), and cultural geography in the USA, led by Sauer (1952) and with strong links to anthropology. Economic, political, and social geography were small enterprises until the second half of the twentieth century.

Two strongly linked developments in the 1950s/1960s—termed 'quantitative and theoretical revolutions'—generated an explosion of interest in both physical and human geography. Quantitative methods were adopted to handle spatial data, but the 'revolutionary' component within human geography involved explicit consideration of theory, notably location theories whose key element was a geographical variable—space, usually represented as the cost involved in moving people, goods, and information. Classic works on the location of agricultural land uses industry and retail establishments (notably Christaller's central place theory were appropriated and formed the basis for a 'new geography,' a spatial science whose goal was identifying and accounting for spatial order. This was largely North American in provenance, initially focused on a few departments (Johnston 1997), but the approach soon infiltrated other parts of the discipline, assisted by two seminal British anthologies (Chorley and Haggett 1965, 1967). Further spatial theories were added—on the diffusion of innovations and the widely-applied gravity model for studying flow patterns—and the entire corpus was synthesized in influential text books, notably Haggett (1965) and Abler et al. (1971).

The expansion was accompanied by increased specialization. Economic geography spawned subdivisions of industrial, agricultural, transport, and development geography, for example; urban geography grew a substantial subsector in urban social geography and was paralleled by interest in rural geography; social geography begat population and medical geography; and political geography was both revived and spawned electoral geography. The shared goal was the discovery of spatial order—in both individual behavior patterns and their outcomes in the built environment.

Although diverging systematic interests reduced links between physical and human geographers, they explored common ground in techniques of spatial analysis and approaches to knowledge. The switch from an ideographic focus on unique characteristics of places to a nomothetic search for spatial order largely occurred in a philosophical vacuum, but some geographers promoted the positivist approach to knowledge production (Bunge 1966, Harvey 1969).

These shifts occurred when the discipline was expanding rapidly within universities. Their promoters achieved major intradisciplinary changes by portraying spatial science as meeting contemporary needs among the burgeoning, increasingly quantitative and theoretical, social sciences. They presented the regional approach as outdated: the concepts of areal differentiation and region were retained—though employed in different ways.

2.2 '-Isms,' Schisms, and 'Turns'

Although some promoted this 'new geography' as surpassing its predecessor, there was no complete revolution to a new hegemonic paradigm in the Kuhnian sense. Not all geographers were won over, and some continued to teach and research in their established ways—sustaining the traditions of historical, cultural, and regional geography which were foundations for later responses to the 'shock of the new.' Those responses criticized spatial science on two main grounds. The first attacked location theories for narrow economism, for privileging profit-maximizing (translated geographically as transport-cost-minimizing) as a determinant of decision-making and so denigrating the wide range of cultural and other influences on behavior. Spatial science, it was claimed, removed free will by assuming that humans react to stimuli in predetermined ways while its elevation of space as the predominant influence on locational decisions led to a derogation of place. Some critics explored alternatives to positivism—such as phenomenology, idealism and existentialism—in work categorized as humanistic geography.

The second set of criticisms concerned spatial science's approach to explanation. Location theories, it was argued, could not account for the major contours of the geography of development and underdevelopment at various spatial scales, of inequality and injustice, wealth, power, and discrimination. Explanation required understanding how economic systems, especially capitalism, work and involved more than adding the costs of transport and communications to neoclassical economic models: such explanation could be achieved through Karl Marx's writings. Harvey's (1973) collection of essays initiated

a series incorporating space to Marxian thought (notably Harvey 1982) and stimulated others during a period of social tensions over civil rights and the Vietnam War. Their work became known as radical geography.

Although both critiques attracted adherents, many were as dissatisfied with each as they were with spatial science. As Gregory (1978) demonstrated, whereas some spatial scientists strayed very close to naive economic determinism, much humanistic geography approached another extreme of voluntarism, tearing the individual out of context, while some radical geography offered little more than an alternative economic determinism. Much effort was spent in the late twentieth century integrating aspects of humanistic and radical geography.

3. *The Contemporary Discipline*

Sheppard (1995) divided human geographers into two main groups—spatial scientists and social theorists; strands of contemporary geographical scholarship that operate somewhat independently (Johnston 1997). Some recent textbooks in the 'social theory mold' dismiss spatial science as part of a discredited past (as in Peet 1998), whereas a volume promoting geography within the US scientific community concentrated on it (plus physical geography: NRC 1997). Attempts at comprehensive overviews include anthologies which lack integration (e.g. Gaile and Willmott 1989, 2001) while the editors of a volume on the 'pervasive themes' and 'common elements within a discipline whose practitioners are in danger of forgetting their shared heritage and ideals' concluded that not only had they failed to bridge the human–physical divide but they also feared a similar divide was opening within human geography (Abler et al. 1992).

3.1 *Spatial Science*

Spatial science remains a substantial component of contemporary human geography. It is strongly quantitative, but the formal (geometrical) location theories based on a single causal variable (space) have largely been abandoned: the search for spatial order neither anticipates the discovery of regular structures nor seeks universal laws of spatial behavior. Sayer (1984) drew an important distinction between extensive and intensive research: the former seeks empirical regularities whereas the latter explores the causal chains responsible for particular outcomes. Much spatial science is extensive research, a necessary precursor to many detailed investigations; by not eschewing empirical generalisations, it identifies significant features and trends in the mass of numerical data which characterize modern societies.

In the 1950s–1970s geographers assumed that standard statistical procedures could be applied unproblematically to analyses of point, line, flow, and area patterns. This was challenged by work on spatial autocorrelation, which identified a range of problems and proposed new methods of spatial data analysis (e.g. Haining 1990). Other issues identified included the modifiable areal unit problem. Some geographical analyses study the characteristics of populations aggregated by areas (such as census administrative units), but there is an extremely large number of ways in which such places can be defined, involving both scale (how large are the areas?) and aggregation (how many different ways can areas of the same size be created?) effects. Openshaw (1985) showed that different aggregations can produce divergent statistical results, creating problems in deciding which to employ. There are also related geographical examples of ecological fallacies—assumptions that results for a particular dataset can be generalized to others at different scales and/or aggregations (including individuals). Various procedures for attacking these problems and providing robust solutions are exemplified by essays in Longley and Batty (1996), many significantly assisted by developments in computing power (including applications of artificial intelligence) and stimulating the conception of geocomputation to describe such work in both physical and human geography (Longley et al. 1998).

The most significant technological developments have been in geographical information systems (GIS: Longley et al. 1999), combined hardware and software for the organization, integration, analysis, and display of spatially-referenced data, with the powerful display media underpinning the growth of visualization strategies. These systems have revolutionized spatial analysis and led to the identification of a geographical information science: traditional studies can be undertaken much more readily and quickly; exploratory studies are increasingly feasible, and large-scale modeling strategies integrating datasets collected on different spatial templates made possible.

Much contemporary spatial science, including GIS, is applied in a wide range of public and private sector contexts—as in spatially-targeted niche marketing strategies based on small-area classifications (geodemographics). This has partly been in response to changes in the economic context for universities: the pressure to increase nonstate income has stimulated 'applied research' and has seen the development of such skills as a major selling point in the attraction of students to read for geography degrees (see NRC 1997).

3.2 *Social Theory*

Although spatial science remains a substantial component of the discipline, to many human geographers approaches involving social theory form the disciplinary mainstream (as in Massey et al. 1999). Humanistic and radical geography have metamorphosed

into new forms, although some maintain allegiance to the latter—notably Harvey (1989) and his continued reworking of Marxian approaches for understanding the contemporary world.

Work within this umbrella category focuses on understanding difference. Place now takes precedence over space among the main geographical concepts (on the tension between the two, see Taylor 1999): places provide contexts within which differences are produced and reproduced. But places are not fixed containers; like spaces, they are produced and continually reproduced by human action so that whereas in spatial science space is an exogenous independent variable (though modifiable, as in transport network developments), in social theory it is both independent and dependent variable (as stressed in Martin's 1999 critique of the 'new geographical economics' which relates more to the spatial science of the 1970s than to either strand of contemporary human geography). The interactions between people and places/spaces continually transform both, introducing a geographical component to the structure-agency dialectic for which Giddens (1977) sought a resolution in his structuration theory, very influential in human geography in the 1980s. People make their own histories, but within constraining/enabling spatial contexts—locales, or settings for interaction. (Giddens 1984 was one of the few social theorists who not only used spatial metaphors in his writings but was also influenced by geographical work, notably time-geography—e.g. Hägerstrand 1982—which draws attention to the geographical constraints on how individuals pursue personal and collective projects.)

A very important stimulus to recognition of the role of context/locale in the production and reproduction of difference came from feminist geography. Initially launched to highlight the discrimination against women within a male-dominated discipline and the consequential many silences in what geographers studied, this expanded into a much wider critique of 'single-factor' approaches (such as Marxist's concentration on class and the economic infrastructure) that ignored important other societal cleavages (Women and Geography Study Group 1984, Rose 1993). People viewing society from different positions create separate identities of self and other, which are potential foundations for identity politics. The class struggle is not the only one within capitalist society, and others (based on sexual orientation, race and religion, for example, as well as gender) cannot be folded within it.

Other stimuli came from postmodernism, which also emphasized difference and rejected the search for grand theories. Dear (1988) and introduced it to geographers as a way of addressing the contemporary condition, exemplified in the argued transition from a Fordist regime of accumulation, characterized by mass production of uniform commodities for mass consumption, to post-Fordism, characterized by small niche markets and consumption for status as well as use (sometimes called flexible accumulation). Harvey (1989) countered by arguing that this shift, involving the production of new spaces and places, is but the latest response to a crisis of capitalism.

More important has been the impact of the various strands of poststructuralist thought derived from the humanities and social sciences. Poststructuralists argue that human subjectivity is not given but rather produced in contexts. So much of that production process is spatially and temporally contingent that truth and knowledge are contingent too, varying through time and over space in a range of situated and local knowledges that form the resource foundations for structuration dialectics and identity politics. Representation of these contingencies is crucial to social conduct, and poststructuralists stress the importance of language not only as the communication medium involved in creating human subjectivity but also as a social construction itself, with variations that are fundamental to the creation and understanding of difference (i.e. in the relationship between the signified and the signifier and in its interpersonal transmission—the hermeneutic process). Language is just one means of communicating understandings, however, and as well as interpreting written texts geographers have focused on other images (works of art, films, landscapes) implicated in the creation of situated knowledges—as in Harley's (1989) seminal work on deconstructing maps.

Said's (1978) book on the Western creation of a 'mythical East' was an important stimulus to geographers' studies of the roles of images of places and spaces in the creation and transmission of meanings and identities. They have embraced the postcolonial movement which challenges not only the political and economic components of colonialism and imperialism but also their roles in creating imaginative geographies at variance from those of the peoples who experienced that external power and are creating alternative geographies, knowledges, and identities through their own writings.

Many of these arguments were introduced through cultural geography but not confined to it and have been instrumental in reducing some of geography's fragmentation. Several parts of the discipline have experienced a 'cultural turn': the perceived interdependence of cultural and economic processes, for example, has stimulated broader studies of production, exchange, and consumption (Lee and Wills 1998); critical geopolitics emphasizes the ways images of the world are created, transmitted, and underpin geographies of power and conflict (O'Tuathail 1996); and writing on nonrepresentational theory focuses on the transitory, the 'billions of happy or unhappy encounters' (performances) which characterize the daily interactions among people, places, environments, and 'things' but are never recorded/represented in the traditional source materials for geographical schol-

arship (Thrift 1996). At an extreme of spatial scale heretofore ignored, geographers are also exploring the body as a site of identity-formation, as the most intimate of 'personal spaces' or 'the geography closest in' (Pile 1996).

Research within this area is both catholic and eclectic in its substantive interests. Space and place are being theorized in a wide range of contexts and scales. There are increasing links with other social science and humanities disciplines experiencing a 'spatial turn,' both in their use of spatial metaphors and in their appreciation of the roles of space and place in the structuration of people and societies.

Such work is very far from most spatial science and there is little methodological or philosophical common ground: one area of interaction, however, has emerged from critiques of the representations of the world created within GIS (Pickles 1995) and, more broadly, in cyberspace (Kitchin 1998). Social theory also lacks an 'applied' element as generally appreciated by those promoting geography as a 'relevant' or 'saleable' discipline (hence the absence of much on social theory in NRC 1997). Critical human geography is increasingly used to categorize work committed to emancipation, however, clarifying ways in which space and place are manipulated in the creation and maintenance of power relationships. This has a strong moral/ethical component, concerned not only with how geographical research and teaching are conducted but also with evaluating what is right and wrong, good and bad (Smith 1994, Sack 1997).

One area of traditional geographical interest revived by social theorists is the study of people–environment interactions, on which there has been a continuing stream of studies on various aspects of environmentalism, human impacts on the environment (Turner et al. 1990), environmental policy, and political economy approaches to environmental problems and issues. Poststructuralists argue that nature is a social construction—work on nature–culture interrelationships must take account of the cultural constructions of nature even the binary nature: culture categorization sets up an unreal opposition since each is part of the other. Writing on society–nature interrelationships has thus been substantially broadened, expanding from the technical issues involved in the production of 'environmental problems' and the political ones in their 'solution/resolution' to raise fundamental questions regarding the role of humans in nature. Some argue that studies of the environment can integrate physical and human geography (Simmons 1997) and present a major opportunity for geographers to undertake relevant research.

4. Conclusions

Human geography is a large, expanding, and vibrant discipline, characterized like most others by fragments and tensions: its extent and the nature of the expansion can be appreciated by comparing the four editions of *The Dictionary of Human Geography* (Johnston et al. 2000). It has an established niche without the academic division of labor; its core concerns remain those crystallized over a century ago; and its basic concepts remain place, space, and environment. Within those general characterizations, however, the discipline has changed markedly in recent decades, not only in what it studies but also in how and why. Buttimer (1993) has suggested that throughout their discipline's history geographers have regularly shifted among four foundational metaphors (mosaic, machine, organism, and arena)—without the academic revolutions that are central to Kuhn's (1970) paradigm model of scientific progress. Contemporary spatial science employs the mosaic metaphor with its emphasis on spatial order (without the earlier determinist overtones when the machine metaphor was employed to 'explain' those mosaics), whereas social theory concentrates on organism and, especially, arena—the spatial contexts, or places, which are involved in the production of human-ness while being produced and reproduced by humans.

See also: Cultural Geography; Environment and Development; Geographic Thought, History of; Historical Geography; Place in Geography; Regional Geography; Rural Geography; Space and Social Theory in Geography; Spatial Analysis in Geography; Urban Geography

Bibliography

Abler R, Adams J S, Gould P 1971 *Spatial Organization: the Geographer's View of the World*. Prentice-Hall, Englewood Cliffs, NJ

Abler R F, Marcus M G, Olson J M (eds.) 1992 *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography*. Rutgers University Press, New Brunswick, NJ

Bunge W 1966 *Theoretical Geography*, 2nd edn. Gleerup, Lund

Buttimer A 1993 *Geography and the Human Spirit*. Johns Hopkins University Press, Baltimore, MD

Chorley R J, Haggett P (eds.) 1965 *Frontiers in Geographical Teaching*. Methuen, London

Chorley R J, Haggett P (eds.) 1967 *Models in Geography*. Methuen, London

Darby H C 1977 *Domesday England*. Cambridge University Press, Cambridge, UK

Dear M J 1988 The postmodern challenge: reconstructing human geography. *Transactions, Institute of British Geographers* **13**: 262–74

Freeman T W 1961 *A Hundred Years of Geography*. Duckworth, London

Gaile G L, Willmott C J (eds.) 1989 *Geography in America*. Merrill Publishing, Columbus, OH

Gaile G L, Willmott C J (eds.) 2001 *American Geography at the Millennium*. Oxford University Press, New York

Giddens A 1977 *Studies in Social and Political Theory*. Hutchinson, London

Giddens A 1984 *The Constitution of Society*. Polity Press, Cambridge, UK

Gregory D 1978 *Ideology, Science and Human Geography*. Hutchinson, London
Hägerstrand T 1982 Diorama, path and project. *Tijdschrift voor Economische en Sociale Geografie* **73**: 323–39
Haggett P 1965 *Locational Analysis in Human Geography*. Edward Arnold, London (2nd edn., with Cliff A D, Frey A E 1977)
Haining R Q 1990 *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, UK
Harley J B 1989 Deconstructing the map. *Cartographica* **26**: 1–20
Hartshorne R 1939 *The Nature of Geography: A Critical Survey of Current Thought in the Light of the Past*. Association of American Geographers, Lancaster, PA
Harvey D 1969 *Explanation in Geography*. Edward Arnold, London
Harvey D 1973 *Social Justice and the City*. Edward Arnold, London
Harvey D 1982 *The Limits to Capital*. Blackwell, Oxford, UK
Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK
Harvey D 1996 *Justice, Nature and the Geography of Difference*. Blackwell Publishers, Oxford, UK
Johnston R J 1997 *Geography and Geographers: Anglo-American Human Geography Since 1945*, 5th edn. Arnold, London
Johnston R J, Claval P (eds.) 1984 *Geography Since the Second World War: an International Survey*. Croom Helm, London
Johnston R J, Gregory D, Pratt G, Watts M J (eds.) 2000 *The Dictionary of Human Geography*, 4th edn. Blackwell Publishers, Oxford, UK
Kitchin R 1998 *Cyberspace: The World in the Wires*. Wiley, Chichester, UK
Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago, IL
Lee R, Wills J (eds.) 1998 *Geographies of Economies*. Arnold, London
Livingstone D N 1992 *The Geographical Tradition: Episodes in the History of a Contested Enterprise*. Blackwell Publishers, Oxford, UK
Longley P, Batty M (eds.) 1996 *Spatial Analysis: Modelling in a GIS Environment*. GeoInformation International, Cambridge, UK
Longley P A, Brooks S M, McDonnell R, Macmillan B (eds.) 1998 *Geocomputation: A Primer*. Wiley, Chichester, UK
Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds.) 1999 *Geographical Information Systems: Principles, Techniques, Applications and Management*, 2nd edn. John Wiley, New York
Martin R 1999 The new 'geographical turn' in economics: some critical reflections. *Cambridge Journal of Economics* **23**: 65–91
Massey D, Allen J, Sarre P (eds.) 1999 *Human Geography Today*. Polity Press, Cambridge, UK
NRC 1997 *Rethinking Geography: New Relevance for Science and Society*. National Research Council, Washington, DC
Openshaw S 1983 *The Modifiable Areal Unit Problem*. CATMOG 38, GeoBooks, Norwich, UK
O'Tuathail G 1996 *Critical Geopolitics: The Politics of Writing Global Space*. Routledge, London
Peet R 1998 *Modern Geographic Thought*. Blackwell Publishers, Oxford, UK
Pickles J (ed.) 1995 *Ground Truth: The Social Implications of Geographical Information Systems*. Guilford Press, New York
Pile S 1996 *The Body and the City*. Routledge, London
Rose G 1993 *Feminism and Geography*. Polity Press, Cambridge, UK
Sack R D 1997 *Homo Geographicus: A Framework for Awareness, Action and Moral Concern*. Johns Hopkins University Press, Baltimore, MD
Said E W 1978 *Orientalism*. 1st edn. Pantheon Books, New York
Sauer C O 1952 *Agricultural Origins and Dispersals*. American Geographical Society, New York
Sayer R A 1984 *Method in Social Science: A Realist Approach*. Hutchinson, London (2nd edn. 1992 published by Routledge, London)
Scott A J 1988 *New Industrial Spaces*. Pion, London
Sheppard E S 1995 Dissenting from spatial analysis. *Urban Geography* **16**: 283–303
Simmons I G 1997 *Humanity and Environment: A Cultural Ecology*. Longman, London
Smith D M 1994 *Geography and Social Justice*. Blackwell Publishers, Oxford, UK
Stoddart D R 1986 *On Geography, and its History*. Blackwell, Oxford, UK
Taylor P J 1999 Places, spaces and Macy's: place-space tensions in the political geography of modernities. *Progress in Human Geography* **23**: 7–26
Thrift N 1996 *Spatial Formations*. Sage, London
Turner B L, Clark W C, Kates R W, Richards J F, Matthews J T, Meyer W B (eds.) 1990 *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere Over the Last 300 Years*. Cambridge University Press, Cambridge, UK
Women and Geography Study Group of the IBG 1984 *Geography and Gender: An Introduction to Feminist Geography*. Hutchinson, London
Wooldridge S W 1956 *The Geographer as Scientist*. Thomas Nelson, London

R. J. Johnston

Geopolitics

The idea of geopolitics evokes mixed reaction. It is understood variously as: a set of geographically determined 'laws' governing a state's strategic destinies; the geographical underpinnings of *Realpolitik*; and a scholarly analysis of the geographical factors underlying international relations and guiding political interactions. Widely embraced by policy makers, geopolitics has encountered scholarly criticism for its sometimes deterministic doctrines and lack of moral considerations.

1. *Definitions*

Geopolitics is a product of its times, and its definitions have evolved accordingly. Johan Rudolph Kjellén, who coined the term in 1899, described geopolitics as 'the theory of the state as a geographical organism

or phenomenon in space' (Kjellén 1916). For Karl Haushofer, the father of German Geopolitik, 'Geopolitics is the science of the conditioning of political processes by the earth … of political space organisms and their structure … [and] aims to be equipment for political action' (Haushofer et al. 1928). In 1939, Derwent Whittlesey, the American political geographer, dismissed geopolitics as 'a dogma … the faith that the state is inherently entitled to its "place in the sun".' In contrast, Edmund Walsh espoused an American geopolitics based upon international justice, and 'a combined study of human geography and applied political science … dating back to Aristotle, Montesquieu and Kant' (1944).

Saul Cohen defined geopolitics as a mode of analysis, relating diversity in content and scale of geographical settings to exercise of political power and identifying spatial frameworks through which power flows (Cohen 1973). Recently, Gearoid O'Tuathail argued that 'geopolitics does not have a singular, all-encompassing meaning or identity … it is discourse, a culturally and politically varied way of describing, representing and writing about geography and international politics' (1998).

Because geopolitics straddles two disciplines—geography and politics—its approaches vary according to frameworks of analysis common to each. While most early theories and concepts in geopolitics grew out of geographical thought, historians and political scientists often failed to adapt their later theories to the dynamic, complex nature of geographical settings.

2. *Stage One—The Race for Imperial Hegemony*

Geopolitical thinking can be traced back to Aristotle, Strabo, Bodin, Montesquieu, Kant, and Hegel. Nineteenth century precursors include von Humboldt, Guyot, Buckle, and Ritter. Modern geopolitics has developed through four stages—the race for imperial hegemony, German Geopolitik, the Cold War era, and the post-Cold War years. The theories of its founders, Ratzel, Mackinder, Kjellén, and Mahan, reflected their era of intense nationalism, state expansionism, and overseas empire-building. Their principles and 'laws' reflected national perspectives and experiences, as well as the influence of social Darwinism.

2.1 *Ratzel*

Friedrich Ratzel (1844–1904), German natural scientist and 'father' of political geography, was the first to treat space and location systematically in his comparative studies of states (1897). He provided successor geopoliticians with a 'scientific' basis for state expansionist doctrines that reflected Germany's nineteenth century experiences and ambitions for the future. During the last half of the century, Germany emerged as continental Europe's chief economic and military power. Unified under Bismarck's leadership and victorious in its wars with Austria and France, Germany had enlarged its territory, expanded its heavy industries, and enacted social reform. With the aid of a new powerful naval fleet, Germany posed a serious threat to Britain and France as it acquired an overseas empire in Africa and the Western Pacific, and commercial footholds in East Asia.

Ratzel based his system upon principles of evolution and science (1882/1891). He viewed the state as an organism fixed in the soil, whose spirit derived from mankind's ties to the land. His geographical 'laws' focused on space (*Raum*) and location (*Lage*), the former dependent upon and contributing to the political character of groups living in the space, the latter providing space with its uniqueness. He introduced *Lebensraum* (living space) as the law driving states to expand. Ratzel's 'organic' theories of state growth fitted Germany's view of its future as a youthful, aggressive, capitalist 'Giant State.'

2.2 *Mackinder*

Halford Mackinder (1861–1947), who established geography as a university discipline in Britain, was concerned with safeguarding the British Empire's political, commercial, and industrial primacy, in an era when command of the seas no longer guaranteed world supremacy owing to the wider reach and speed of rail. For Mackinder, geographical realities lay in the advantages of centrality of place and efficient movement of ideas, goods, and people. He theorized that the inner area of Eurasia, characterized by interior or Arctic drainage and impenetrable by sea power, was the 'pivot area' of world politics (1904). He warned that rule of the heart of the world's greatest landmass could lead to world domination. A Eurasian landpower (be it Russia, Germany, or even China, and especially an alliance of the first two) that gained control of the pivot area would outflank the Maritime world. Eleven years later, the English geographer James Fairgrieve (1915), who coined the term 'Heartland' opined that China was in an excellent position to dominate Eurasia.

In *Democratic Ideals and Realities* (1919) Mackinder, now using the term 'Heartland,' and taking into account advances in land transportation, population increases, and industrialization, enlarged his map to include Eastern Europe from the Baltic through the Black Sea as Inner Eurasia's strategic annex. This became the basis for his dictum, 'Who rules East Europe commands the Heartland; who rules the Heartland commands World-Island; who rules World-Island commands the world.' The warning to Western statesmen was clear—the key to world domination lay in the middle tier of German and Slavic states, or Mitteleuropa. Mackinder's world-view guided Lord Curzon's imperial strategies in South Asia and South Russia (Blouet 1987), German Geopolitik between

World War One and Two, and for Western containment strategies of the post-World War Two era. Foreseeing Britain's decline as the world's leading power (Mackinder 1924), he advocated that Western Europe and North America constitute a single community of nations—a forerunner of the North Atlantic community concept.

Whereas Ratzel's theories of the large state were based on concepts of self-sufficiency, closed space, and totalitarian controls, Mackinder was strongly committed to cooperation among states, democratization of the Empire into a Commonwealth of Nations, and preservation of small states. He bridged the academy and politics, serving as a Conservative and Unionist member of Parliament (1910–1922), and as British High Commissioner for South Russia (1919–1920). Initially a Liberal Imperialist and proponent of free trade, he eventually became committed to a preferential tariff system to protect British imperial unity.

During World War Two, Mackinder presented a global strategy based upon two equal power units—Heartland and Midland Ocean (the North Atlantic)—separated from two other regions (Monsoonal Asia and South Atlantic lands) by a mantle of empty deserts and mountains. He felt that, ultimately, Monsoonal Asia (India and China) would emerge to challenge a combined Heartland/Midland Ocean (1943).

2.3 *Kjellén*

Johan Rudolph Kjellén (1864–1922), the political scientist who, in 1899, coined the term 'geopolitics' was influenced by both his Swedish background and Germany's growth into a giant state. He viewed the impending breakdown of the Concert of Europe and the drift towards war and chaos as the death-knell for small states like Sweden (Parker 1998). Adopting Ratzel's organic state concept, he considered Germany's emergence as a great power inevitable and desirable. Sweden's needs would be fulfilled within the framework of a new Mitteleuropean bloc from Scandinavia and the Baltic, through Eastern Europe and the Balkans, dominated by ascendant Germany.

A Conservative member of the Swedish parliament, Kjellén viewed geopolitics as the 'science of the state,' whereby the state's natural environment provided the framework for a power unit's pursuit of 'inexorable laws of progress.' The dynamic organic approach led Kjellén to espouse the doctrine that political processes were spatially determined (1916). Moreover, since giant states in Europe could be created only by war, he viewed geopolitics as primarily a science of war (Gyorgy 1944).

2.4 *Mahan*

The global perspective of Admiral Alfred Thayer Mahan (1849–1914) was also Eurasian-centered (1900). For Mahan, the northern land hemisphere, connected by Panama and Suez, was the key to world power; within that hemisphere, Eurasia was the most important component. Mahan recognized Russia as the dominant Asian landpower whose location made it unassailable. However, he felt that its landlocked position put it at a disadvantage, because, contrary to Mackinder's opinion, sea-movement was superior to land-movement (1890).

Mahan argued that world dominance could be held by an Anglo–American alliance from key bases surrounding Eurasia. The critical zone of conflict lay between the thirtieth and fortieth parallels in Asia where Russian landpower and British seapower met. He viewed the USA as an outpost of European power and civilization, its Pacific shoreland and islands extensions of the Atlantic–European realm. Indeed, he predicted that an alliance of the USA, Britain, Germany, and Japan would one day hold common cause against Russia and China.

3. *Stage Two—German Geopolitik*

German Geopolitik emerged in reaction to Germany's devastating defeat in World War One. Humbled by the Treaty of Versailles, Germany was stripped of its overseas empire and important parts of its national territory, and the social cohesion forged by Bismarck's policies was shattered. The Socialist Weimar Republic was beleaguered by class warfare and attempts to overthrow it by Communists on the Left and racist militant nationalists and aristocratic conservatives on the Right. Unemployment was rampant and inflation raged. Within this setting, Haushofer and his colleagues established the *Zeitschrift für Geopolitik* (1924–1939) and the Institute for Geopolitics at the University of Munich. Undoing Versailles by restoring the lost territories and rebuilding Germany as a world power undergirded the pseudo-scientific 'laws' and principles of Geopolitik that served Nazi Germany.

3.1 *Haushofer*

Karl Haushofer (1869–1946), the general who became a political geographer, was not an original thinker. The Geopolitik of the German geopoliticians whom he led (Otto Maull, Erich Obst, Ewald Banse, Richard Hennig, Colin Ross, Albrecht Haushoffer) was based essentially upon the writings of Kjellén, Ratzel, and MacKinder. Mahan, Fairgrieve and the American geographical determinist Ellen Churchill Semple were also influential. The organismic-Hegelian philosophy of Geopolitik came from Ratzel directly or via Kjellén.

Lebensraum and autarchy became slogans for doctrines whose consequences were conflict and total war. Three geographical settings permeated the literature

of Geopolitik: Ratzel's large states, Mackinder's World-Island, and Pan-Regions. The organic growth of Germany to its west and east was regarded as 'inevitable.' To gain mastery over World-Island, it was necessary to dominate the USSR and destroy British seapower. The geopoliticians posited that German control over Pan-Europe (including Eastern Europe) would force the Soviet Union, regarded as an Asian power, to come to terms.

During most of the 1920s and 1930s, Haushofer espoused continental pan-regionalism based upon complementarity of resources and peoples: Pan-America, Pan-Eurafrica, and Pan-Asia, with the USA, Germany, and Japan as respective cores. His position on the USSR was ambiguous. He proposed variously a German–Russian alliance, a Pan-Russia–South Asia, and a Japan–China–Russia bloc. His call for Germany, the USSR, and Japan to form a Eurasian pan-region that would dominate World-Island, influenced the German–Soviet pact of 1939, made moot by Hitler's subsequent invasion of the Soviet Union. The principles of Geopolitik were designed to fulfill German national and imperial aims, without concern for contradictions.

Haushofer's extraordinary influence derived from his close ties to Rudolph Hess, his aide-de-camp in World War One and subsequent student at the University of Munich. Through Hess, he gained access to Hitler from 1923 to 1938. Doctrines such as *Blut und Boden* (blood and soil), *Rasse und Raum* (race and space), and *Lebensraum* were incorporated into *Mein Kampf* and became ideological tenets of the murderous Nazi regime. He also advised Hitler at Munich 1998. With Hess's flight to England in 1941, the influence of the Geopoliticians upon Hitler ended. Haushofer was imprisoned briefly at Dachau (ironically, he had a Jewish wife), and his son Albrecht, also a geographer with links to aristocrat military circles, was killed by the SS for involvement in the 1944 Generals' Plot to assassinate Hitler. Haushofer and his wife committed suicide in 1946.

4. *Stage Three—The Cold War: State-centered vs. Universalistic Approaches*

Onset of the Cold War re-awakened Western interest in geopolitics by historians, political scientists, and statesmen. Geographers distanced themselves from geopolitics because of the taint of German Geopolitik.

4.1 *State-centered 'Political' Geopolitics*

American Cold Warriors embraced geopolitics as a basis for national policy aimed at confronting the Soviet Union and International Communism. Building on early geographically derived geopolitical theories, and holding static interpretations of global and regional spatial patterns, they introduced into the lexicon of Cold War geopolitics such political–strategic doctrines as containment, domino theory, balancing linkages, and linchpin state.

Germany and Japan emerged as the key military bases for the USA and its allies in the struggle to contain Soviet and Chinese continental power, thrusting the Allies' World War Two enemies into the strategic forefront. Military alliances forged as part of containment produced mixed results. NATO, established in 1949, has proved an enduring strategic instrument. However, SEATO, formed in 1954 to oppose the spread of Communism in Southeast Asia after France's withdrawal from Indochina, collapsed with the US defeat in Vietnam, and was disbanded in 1977.

In the Middle East, the Baghdad Pact (later called the Central Treaty Organization) was even shorter-lived (1955–1959). Its members included Turkey, Iraq, Iran, Pakistan, and Britain. Iraq's subsequent withdrawal to pursue a non-alignment policy left the pact with no Arab representation and therefore limited capacities to counter Soviet influence in Egypt, Syria, and South Yemen.

Western foreign policy was not long confined to containing the Eurasian continental power along Heartland's borders, but adopted a strategy of checking the spread of Communism throughout the Third World. The idealistic vision that had prompted the USA to support freedom and democratization among colonial peoples quickly gave way to expedient realpolitik—propping up right-wing dictatorships in order to stop the threat of Communism.

Another popular geopolitical doctrine, 'domino theory,' was first proposed by William Bullitt (1947), who feared that Soviet Communist power would spread, via China, into Southeast Asia. It was adopted by both the Kennedy and Nixon administrations, rationalizing American intervention in Vietnam to 'save' the rest of Southeast Asia (O'Sullivan 1982).

Domino theory was an important argument for extending Western containment policy well beyond the Southeast Asian and Middle Eastern shatterbelts into the Horn of Africa and Sub-Saharan Africa, Central America and Cuba, South America, and South Asia. These regions became battlefields for both superpowers, each supporting local surrogates militarily, politically, and economically. The goal was to protect or gain sources of raw materials and markets, while denying the enemy overseas military bases.

The imagery of 'dominos' persists. Threat of the spread of Kosovar Albanian irridentism to Macedonia, Bulgaria, and Greece precipitated NATO's air war against Yugoslavia. Ironically, the result has been Serbian ethnic cleansing of the Kosovars, further destabilizing the region.

A third doctrine, 'linkage,' introduced into geopolitics by Henry Kissinger in 1969, was based upon the theory of a network that connected all parts of the world's trouble spots to the Soviet Union, and that US involvement in any single conflict would have an impact upon overall superpower balance (Kissinger 1979). Linkage rationalized continued US fighting in Vietnam long after the war had been lost.

Linkage was applied to detente with the Soviet Union and accommodation with China. To maintain the balance of power, the Nixon administration sought Moscow's agreement on strategic arms limitations and mutual nuclear deterrence, and tried to play China off against the USSR. The logical consequence of this policy was non-intervention in the 1968 Czech uprising, acquiescing to the Brezhnev Doctrine, which held that military force was justified to keep the socialist countries of Eastern and Central Europe within the Soviet camp. The threat of credibility loss also served as a driving force in NATO's war against Yugoslavia.

Zbigniew Brzezinski's geopolitical world-view was based on the struggle between Eurasian landpower and seapower. For him, the key to preventing Soviet world dominance lay in US control of 'linchpin' states (1986). These were defined by their geographical locations which enabled them to exert strategic military and/or economic influence. The designated linchpins were Germany, Poland, Iran or Pakistan-Afghanistan, South Korea, and the Philippines. Their dominance by the USA would effectively contain the Russian 'Imperial' power, protecting Europe and Japan and, in the case of South Korea and the Philippines, preventing encirclement by China.

For Brzezinski the US–Soviet conflict was an endless game, and linchpin control was a necessary part of US geostrategic game plan. In this there is little consideration of the geopolitical complexity of the global system and of the multiplicity of forces beyond superpower reach that had become active agents in the system.

4.2 Universalistic Geographical Geopolitics

When geographers re-engaged in geopolitics in the 1960s and 1970s, they developed theories based upon universalistic/holistic views of the world and the dynamic nature of geographical space. Three approaches predominated: a polycentric, international power system; a unitary, economically based world system; and an environmental and socially ordered geopolitics. These fresh geographical theories which challenged bi-polar Cold War geopolitics had little appeal to the Cold Warriors and failed to make their way into 'political' geopolitics.

The polycentric or multinodal/multilevel power approach rejected the Heartland theory of world domination, as had Halford Mackinder in his later writings. His reinterpretation became a take-off point for those geographers who developed theories of geopolitical polycentrism. In the USA, Saul Cohen proposed a flexible hierarchy of two geostrategic realms, geopolitical regions, shatterbelts, national states, and subnational units within a system that evolved through dynamic equilibrium forces (Cohen 1973). The USA, European Union, and Japan are the cores of the Maritime realm, and Russia and China of the Continental. This developmental theory posited that the structural components of the global system had evolved from a stage of undifferentiation with relatively few parts, to specialization/integration with many parts at different geo-territorial levels (Cohen 1982). Equilibrium is maintained by moving from one stage to another through responses to short-term disturbances.

In England, Chrone presented a geopolitical system of 10 regional groupings that were also hierarchically ordered, and had a historical and cultural basis (Chrone 1969). In Chrone's view, the world power balance was shifting from Europe and the West towards Asia and the Pacific. He predicted that the Pacific Ocean would become the future arena of confrontation for the USSR, the USA, and China.

Peter Taylor, the English geographer, broke away from the 'realistic' school of power-centered geopolitics in applying a world-systems approach based upon global economics (1989). He drew upon the work of Immanuel Wallerstein (1983), who argued that the world economy means a single global society, not competing national economies. Taylor integrated the Wallerstein model with George Modelski's cycles of world power (1978), and presented power and politics within the context of a cyclical world economy in which nation states and localities are fitted.

Taylor, as Wallerstein, viewed global conflict in North/South terms (rich nations vs. poor nations), arguing that capitalist core areas aggrandize themselves at the expense of the periphery. This offered a geopolitical frame for the North–South dichotomy that the 1980 Brandt Report highlighted in its study of Third World issues.

An environmentally and socially oriented geopolitics was promoted by Yves LaCoste in France with the establishment of the *Journal Hérodite*. In moving towards a 'new geopolitique,' he sought to overcome the national chauvinism of the 'old' geopolitics by focusing on the land, not on the state. Hérodite linked geopolitics to ecology and broader environmental issues, as well as such matters as world poverty and resource exhaustion. Much of LaCoste's work was inspired by the French human geographer and political anarchist, Elissé Reclus, who believed it essential to reshape the world's political structure by abolishing states and establishing a cooperative global system. While this French geopolitics did not produce systematic geopolitical theory, it did put the spotlight on applying geopolitics to significant global problems.

5. *Stage Four—The Post-Cold War Era: Competition or Accommodation?*

The end of the Cold War has generated a number of new approaches to geopolitics, in addition to those that continue from the earlier era. For some, it heralds a 'new world order' and the geopolitics of US global hegemony. President George Bush, addressing Congress in 1990 defined the policy behind the war against Iraq as envisaging a new world order led by the USA, 'freer from the threat of terror, stronger in the pursuit of justice, and more secure in the quest for peace … a world in which nations recognize the shared responsibility for freedom and justice' (Bush 1991).

In contrast, Robert Kaplan's geopolitics of anarchy is based on a world divided into the rich North and the poor South. He concludes that the South, especially Africa, is doomed to anarchy and chaos. His map of the future, dubbed the 'last map' is an 'ever mutating representation of chaos' (1994).

Neither of these predictions has come to pass. In most cases, the overthrow of Communist regimes has not led to stable, free-market economies. The restraints upon the unilateral application of US military, economic, and political power are evident from the failures to gain US objectives in Iraq, Somalia, and Haiti. In southeastern Europe, Russia and China have been called upon to help mediate the NATO–Yugoslav conflict.

Chaos theory gives inadequate attention to the regional and global systemic forces that contain turbulence and absorb its positive aspects into the system. Indochina has been stabilized under the leadership of Vietnam. Nigeria has led the community of West African states in attempts to restore order in countries plagued by civil war. The West cannot ignore instability in the South despite regarding it as 'quarter-sphere of strategic marginality.' This is necessitated by humanitarian considerations, as well as fear of the spread of nuclear weaponry and massive out-migrations.

The main thrust of post-Cold War geopolitics continues to follow the two streams of the previous era—the nation-centered/political and the universalistic/geographical. Political geopoliticians advocate projection of Western power into Central and Eastern Europe to weaken Russia's Heartland position at its western edge. They also advance strategies for penetrating the Caucasus and Central Asia, and for playing China off against Russia.

Brzezinki's prescription for maintaining US global hegemony is to achieve primacy in three parts of the 'Eurasian Chessboard': the West—Europe; the South—Middle East and Central Asia; and the East—China and Japan. He advocates pulling Ukraine and the Black Sea into the Western orbit, strong engagement in Central Asia and the Caucasus, and support of Chinese regional dominance in peninsular southeast Asia and Pakistan. China would be limited to regional power status by the globally framed US–Japan strategic alliance. On the assumption that Russia's weakness is a permanent condition, this strategy seeks to prevent Russia from reasserting strategic control over 'near abroad' states, or from joining with China and Iran in a Eurasian anti-US coalition (Brzezinki 1997).

Advancing a geopolitics of 'the West against the Rest,' Samuel Huntington argues that world primacy can be maintained by dividing and playing off the other civilizations (1993). His thesis is that the fundamental sources of conflict will not be ideological, but cultural, and the fault lines between civilizations will be the battle lines. In dividing the world into Western, Confucian, Japanese, Islamic, Hindu, Slavic-Orthodox, Latin American, and possibly African civilizations, he makes little allowance for religious, ethnic, economic, or strategic divisions. He also assumes the permanence of the fault lines, despite the massive changes that are brought about by demography, migrations, and modernization.

Geographical geopolitical theory continues to reflect the universalistic approaches advanced during the Cold War. Building on the works of Taylor and LaCoste, 'critical' geopolitics applies social scientific critical thinking to ask how power works and might be challenged. This school views the following as central to geopolitical studies: analysis of discourse—rhetoric, metaphors, symbolism (O'Tuathail and Agnew 1992); feminist approaches to the subject of national security; and the geographies of social movements, particularly in relation to newly radicalized and participative democracy (Painter 1995).

In contrast, reality-based, polycentric geographical geopolitics emphasizes the continuous proliferation and inter-connectedness of the various parts and levels of the world. The current number of 200 national states could increase to 250–275 within the next quarter of a century. As the pace of devolution quickens, some of these new geo-territorial entities will be highly autonomous 'quasi-states.' Regional geopolitical restructuring is also a continuing process. Just as Southeast Asia has disappeared as a shatterbelt, so may the Middle East. The effort to detach Eastern Europe from the Russian Heartland is underway. However, with Russia's revival, the region could serve as a gateway between Europe and Russia.

In addition, the network of global cities—centers of capital flows and financial services, linked ever more closely by cyberspace, tourism and immigrant communities—will emerge as a major new geopolitical level, promoting policies sometimes contradictory to national interests. International social movements, such as environmentalism, are also influential in shaping national and regional policies, including military ones.

Whatever the course of geopolitical restructuring, we are entering an era of power-sharing among a wide variety of states of different sizes and functions.

Geopolitical theory will continue to be a valuable tool for understanding, predicting, and formulating the structure and direction of the world system.

See also: Globalization: Political Aspects; International Justice: Philosophical Aspects; International Law and Treaties; International Relations, History of; Military and Politics; National Security Studies and War Potential of Nations; War: Causes and Patterns

Bibliography

Blouet B W 1987 *Halford Mackinder*, 1st edn. Texas A & M Press, College Station, TX

Brzezinski Z 1986 *Game Plan*, 1st edn. Atlantic Monthly Press, Boston

Brzezinski Z 1997 *The Grand Chessboard*, 1st edn. Basic Books, New York

Bullitt W 1947 A Report to the American people on China. *Life* **23**: 35–7

Bush G 1991 *The Hard Work of Freedom*, Public Papers of the Presidents of the United States. US Government Printing Office, Washington, DC

Chrone G R 1969 *Background to Political Geography*. Pittman Publishers, London

Cohen S B 1973 *Geography and Politics in a World Divided*, 2nd edn. Oxford University Press, New York

Cohen S B 1982 A new map of global geopolitical equilibrium: A developmental approach. *Political Geography Quarterly* **1**: 223–42

Fairgrieve J 1915 *Geography of World Power*. University of London Press, London

Gyorgy A 1944 *Geopolitics–The New German Science*. University of California Press, Berkeley and Los Angeles, CA

Haushofer K, Obst E, Lautensash H, Maull O 1928 *Foundations of Geopolitik*. Kurt Vowinckel Verlag, Berlin

Huntington S P 1993 The clash of civilizations? *Foreign Affairs* **72**: 22–49

Kaplan R 1994 Coming anarchy. *Atlantic Monthly* **273**: 44–6

Kissinger H 1979 *The White House Years*, 1st edn. Little Brown, Boston

Kjellén R 1916 *The State as a Form of Life*. Hugo Gebers Forlag, Stockholm

Mackinder H 1904 The geographical pivot of history. *Geographical Journal* **23**: 421–37

Mackinder H 1919 *Democratic Ideals and Reality*. Constable, London

Mackinder H 1924 *The Nations of the Modern World*. George Philip, London

Mackinder H 1943 The round world and the winning of the peace. *Foreign Affairs* **21**: 595–605

Mahan A T 1890 *The Influence of Sea Power Upon History*. Sampson Low Marston, London

Mahan A T 1900 *The Problem of Asia and its Effect Upon International Policies*. Little Brown, Boston

Modelski G 1978 The long cycle of global politics and the nation state. *Comparative Studies of Society and History* **20**: 214–35

O'Loughlin J (ed.) 1994 *Dictionary of Geopolitics*. Greenwood Press, Westport, CT

O'Sullivan P 1982 Antidomino. *Political Geography Quarterly* **1**: 57–64

O'Tuathail G, Agnew J 1992 Geopolitics of discourse. *Political Geography Quarterly* **11**: 190–204

O'Tuathail G 1998 Thinking critically about geopolitics. In: O'Tuathail G, Dalby S, Routledge P (eds.) *The Geopolitics Reader*, Routledge, London, pp. 1–12

Painter J 1995 *Politics, Geography and 'Political Geography'*. Arnold, London

Parker G 1998 *Geopolitics–Past, Present and Future*. Pinter, London

Ratzel F 1882 (pt.l)/1892 (pt.2) *Anthropogeographie*. Engelhorn, Stuttgart, Germany

Ratzel F 1897 *Politische Geographie*. Oldenbourg Verlag, Munich, Germany

Taylor P J 1989 *Political Geography*. John Wiley, New York

Wallerstein I 1983 *Historical Capitalism*. Verso, London

Walsh E 1944 Geopolitics and international morals. In: Weigert H, Stefansson V (eds.) *Compass of the World*, George Harrap, London, pp. 12–39

Whittlesey D 1939 *The Earth and the State*. Holt, New York

S. B. Cohen

Geriatric Psychiatry

1. The Field of Geriatric Psychiatry

Geriatric psychiatry deals with disturbances of mental health in older persons. Its practitioners are physicians who have trained first in medicine and general psychiatry, then specialized in the treatment of psychiatric disorders affecting people in later life—usually taken to be from the seventies on. Geriatric psychiatry deals with the following: the diagnosis and accurate assessment of individual patients; the development and use of treatments that are the best available; and the administration of health services for older people, involving family physicians, community health workers, and carers.

The importance of geriatric psychiatry expanded greatly in the latter decades of the twentieth century, mainly because of the unprecedented change in the world's population: many more people are living into old age, and even more are becoming very old. For example, according to predictions published by the United Nations (1998), by 2050, the world's population aged 60 and over will be 74 percent larger than the number of people under 15 years of age. The population aged 80 years and older will rise from 61 million in 1995 to 1,055 million in 2150, *a 17-fold increase*. And it is in less developed countries where the most dramatic increases will occur. The social, economic, and public health consequences for many countries are profound. This is because the very elderly tend to have poor physical health, a greatly increased risk of dementia, and a degree of personal disablement that makes them dependent on others. The prevention and effective treatment of age-related disease has become a compelling need.

The agenda of geriatric psychiatry falls into three categories: how it is practiced; the main disorders treated; and the research questions or gaps in knowl-

edge that are currently the most important. Jeste et al. (1999) have set out the response that is needed to this forthcoming crisis. It is generally acknowledged that, to be effective, the practice of geriatric psychiatry requires close collaboration with other professionals, particularly nurses, social workers, clinical psychologists, family physicians, and specialists. This is because of the nature of the disorders being encountered, the behavior of the elderly person in poor health, and the simultaneous presence of both medical and psychiatric problems that so commonly occurs. Much of the practice of geriatric psychiatry is in the community and in primary care, not in specialist settings or hospitals.

2. *Mental Disorders in Later Life*

2.1 *The Dementias*

The most common group of disorders in geriatric psychiatry is the dementias. In these, changes have taken place in the structure and function of the brain, particularly areas concerned with memory, information processing, and social behavior. 'Dementia' originally meant 'out of one's mind,' from the Latin *de* (out of) and *mens* (the mind). Early in the nineteenth century, Esquirol (1845) gave a succinct definition of dementia as 'a cerebral affection … characterised by a weakening of the sensibility, understanding, and will' (cited by Caine et al. 1995). In describing with such words how the condition can be recognized, Esquirol drew attention not only to the cognitive features of the disorder, with impairment of memory and thinking in day-to-day life, but also to its other manifestations such as apathy, deterioration in social behavior, occasional aggressiveness, delusional ideas, and hallucinations. These show how widespread the changes are in the brain. Following extensive consultations with experts in some 40 countries, the World Health Organization ([WHO] 1992) has described dementia as:

> a syndrome due to disease of the brain, usually of a chronic or progressive nature, in which there is disturbance of multiple higher cortical functions, including memory, thinking, orientation, comprehension, calculation, learning capacity, language and judgement. Consciousness is not clouded. Impairments of cognitive function are commonly accompanied, and occasionally preceded, by deterioration in emotional control, social behaviour, or motivation. This syndrome occurs in Alzheimer's disease, in cerebrovascular disease, and in other conditions primarily or secondarily affecting the brain.

In Alzheimer's disease, an abnormal protein called β-amyloid accumulates in the brain, particularly in the hippocampus. In vascular dementia, areas of the brain become damaged by having their blood supply interrupted, as happens in stroke. Other causes of dementia are Lewy body disease, Pick's disease, Huntington's disease, Creutzfeldt-Jakob disease, and AIDS.

2.1.1 Prevalence. There are published reports on the prevalence of dementia in diverse areas of the world including North America, Europe, Scandinavia, Russia, Japan, China, Singapore, India, Africa, and Australasia. An integrative analysis of 47 surveys across 17 countries has suggested approximate rates under one percent for dementia from any cause in persons aged 60 to 69 years, rising to about 39 percent in persons 90 to 95 years old (Jorm et al. 1987). The prevalence doubles with every five years of age within that range. Throughout the world, many more people than in the past are living to an age when the prevalence of dementia greatly increases, so that the absolute number of persons with dementia is now expanding rapidly. In some countries, the proportion with a dementia is increasing *much faster than the growth of the elderly population*. This is because of the greater increase in the very old, in whom the prevalence of dementia is highest.

2.1.2 Secular change. It is often asked if there is more dementia now than some decades ago. Because there are larger numbers of the elderly now, and more of the very elderly in particular, there has inevitably been a large increase in the number of persons suffering from a dementia. Furthermore, having developed a dementia, an elderly person is now likely to survive longer. This is at least partly attributable to the 'failure of success.' Even with no change in incidence, longer duration of illness leads to an increased prevalence. Studies in Sweden and across Western Europe have found no change in the age-specific incidence of Alzheimer's disease, but it is intuitively plausible that vascular dementia may be decreasing in many countries in parallel with a general decrease in vascular disease.

2.1.3 Age-related cognitive decline. There is no doubt that nearly all people undergo a deterioration in memory and a slowing of mental processes in very late life; and some experience these changes earlier than others for reasons that are not yet well understood. Various terms have emerged to describe such states. The term 'benign senescent forgetfulness' was introduced by Kral to describe one group of such states. Although it was not well defined by him, the term has persisted because it fulfills a need and has no sinister connotations. Other concepts have since

been introduced. A useful review of this important issue has been provided by Ritchie and Touchon (2000).

2.1.4 Risk factors for Alzheimer's disease. As reviewed by Jorm (2000) and Henderson and Jorm (2000), four risk factors can now be regarded as confirmed for Alzheimer's disease: age, a family history of dementia; having the Apolipoprotein E $\varepsilon 4$ genotype; and having Down's syndrome. There are then a number of risk factors where the evidence is at present not fully established: a first-degree relative with Down's syndrome, advanced maternal age, a head injury earlier in life, a previous depressive episode, hypothyroidism, and a long-standing history of physical inactivity. Aluminum has been considered, whether ingested in drinking water or the diet, or absorbed from antiperspirants. From all the evidence available, it remains unproven that aluminum from drinking water or any other source plays a role in causing Alzheimer's disease (Doll 1993).

2.1.5 Possible protective factors. Some factors may confer protection against dementia, or even specifically against Alzheimer's disease. For the latter, four have so far been proposed, each having been identified through epidemiological research: education, anti-inflammatory drugs, estrogen replacement therapy, and statins. Each of these points the way to the possibility of prevention. It is well established in medicine that prevention can precede an understanding of etiology.

2.1.6 Education. A recurrent finding in field surveys is that rates for dementia and cognitive impairment are higher in elderly persons who have had little education. One interpretation is that education may delay the point at which a developing dementia becomes clinically manifest. A recent longitudinal study of American nuns (Snowdon et al. 1996) suggests that intelligence or education, which are inextricably linked, may actually protect against the neuropathological processes in Alzheimer's disease. If such a protective effect does apply, the implications for public health could be appreciable because education has become much more widely available to young persons during the twentieth century.

2.1.7 Geographic and ethnic comparisons. Alzheimer's disease and vascular dementia may occur at different rates in different parts of the world and in different ethnic groups. But it is not easy to be certain what factors are responsible. There may be different patterns of mortality for dementia cases under different living conditions, different genetic vulnerabilities, or different environmental exposures. A finding of potential significance comes from a comparison of prevalence rates for dementia and specifically for Alzheimer's disease in elderly Nigerian Africans and African Americans in Indianapolis by Hendrie et al. (1995). The two groups have much of their heredity in common, but very different lifestyles. The prevalence was only 2.3 percent and 1.4 percent for dementia and Alzheimer's disease respectively in the Nigerians, but 8.2 percent and 6.2 percent in the combined community and nursing home sample of African Americans in Indianapolis. This demonstration of significant differences in rates in two different communities having similar ethnic origins again points to the possibility that some unknown environmental factors promote the neuropathological processes in Alzheimer's disease. The Cree Indians in Canada have also been found to have a lower prevalence of Alzheimer's disease than White Canadians (Hendrie et al. 1993).

In India, Ganguli et al. (1996) and Chandra et al. (1998) undertook a methodologically sophisticated survey of 5,126 individuals aged 55 years and over in Ballabgarh, Northern India. Using the same criteria to define a case, they found a prevalence of only 1.1 percent in persons aged 65 and over for Alzheimer's disease, much lower than in industrialized countries. Neither gender nor literacy were associated with prevalence, but this rose with age for both Alzheimer's disease and all dementias. Their conclusion was that the low prevalence compared to Western estimates could be due to shorter survival of cases, but it is also possible that the incidence is lower, speculatively attributable to differences in the underlying distribution of risk and protective factors.

2.1.8 Anti-inflammatory drugs. An inverse association was noticed between rheumatoid arthritis and Alzheimer's disease, suggesting that persons who have taken aspirin, steroids, or nonsteroidal anti-inflammatory drugs (NSAIDs) over long periods may have a lower risk of dementia, or have slower cognitive decline in late life. The most recent information is that anti-inflammatory drugs probably do prevent or attenuate the symptoms of Alzheimer's disease (Breitner 1996). But it is premature for physicians to prescribe anti-inflammatory drugs for Alzheimer's disease until the size of their effect is established in a randomized controlled trial and the findings are balanced with their risks.

2.1.9 Estrogen replacement therapy. Case-control studies have provided pointers to the possible

protection against Alzheimer's disease afforded to women who take estrogen. But since they also tend to be better educated and to differ in other life-style factors, this finding could be misleading. Again, only randomized controlled trials will establish whether a protective effect exists.

2.1.10 Risk factors for vascular dementia. The assumption is that risk factors are similar to those for stroke. These are age, male gender, a family history of stroke or dementia, hypertension, diabetes mellitus, transient ischemic attacks, elevated cholesterol or lipids, cigarette smoking, heavy alcohol use, oral contraceptives, physical inactivity, obesity, and ethnicity.

2.1.11 Treatment of dementia. The treatment of dementia presently consists of minimizing the deleterious effects of other conditions and of optimizing the person's day-to-day social environment. For Alzheimer's disease, medications have been developed that attempt to correct the deficiency in the neurotransmitter acetylcholine, but none has yet been found to bring about unequivocal and lasting benefit. Intense research is being conducted to find other drugs that enhance cognition. The possibility is emerging that people could be immunized against Alzheimer's disease in earlier adulthood to prevent the deposition of the harmful β-amyloid protein in the brain (St George-Hyslop and Westaway 1999). Meanwhile, remarkable progress has been made by clinical psychologists in developing nonpharmacological methods for treating dementia, including the reduction of severely difficult behavior such as wandering, incontinence, and screaming (Bird in press).

2.2 Delirium

Delirium, a term introduced by Celsus 2,000 years ago, is derived from the Latin *de* (from, away) and *lira* (a track or furrow). It is characterized by incoherence of thought and speech, together with perplexity, in a setting of clouded consciousness that fluctuates in severity. Despite its long-established place in nosology, delirium must number among the most underdiagnosed conditions in medicine. Delirium occurs most commonly at the extremes of life. Its highest incidence is in the elderly, in whom it may indicate the presence of systemic or cerebral disease, or drug toxicity.

2.3 Depression

The elderly might be expected to have higher rates of depressive disorders than younger adults because of their personal and social circumstances. But here a puzzle emerges. Although a high prevalence is usually found in nursing homes, the consistent finding is that the elderly in the community have *on average* less depression than younger adults (Henderson 1994; Henderson et al. 1998). Some think these findings may be attributable to an error in ascertainment. This could be through inappropriately discounting symptoms that may also be due to physical illness or to age itself, such as tiredness, feeling slowed up, or sleep difficulties. Alternatively, a cohort effect has been proposed, in which people who are now elderly may have been protected against depression; or younger adults may have higher rates through being exposed to factors that older adults have been spared.

There is then a further intriguing possibility that is at first counterintuitive: that aging actually protects against depression, and indeed against a range of other mental disorders. It has been suggested that exposure to adverse experience over a lifetime may induce an increased resistance to depression and anxiety through psychological immunization.

2.3.1 Causes of depression in the elderly. When elderly persons have a depressive disorder, one may reasonably ask if it is a recurrence or if it has developed for the first time. Several studies have noted that severe life events precede the onset of depression in at least half of all patients; conversely, depressive disorder may emerge in elderly persons with relatively untrammeled lives. Risk factors for depression in late life are a past history of such a disorder, physical illness or chronic pain, recent adverse experiences, and the absence of a confiding relationship. Contrary to expectation, elderly persons who live alone have not been found to be at risk for depressive disorder. A consistent finding in community surveys is that the complaint of loneliness is associated more with depression than with actual social isolation.

Where depression develops in elderly persons for the first time in their lives, a recent finding is that the brain often shows selective atrophy on neuroimaging.

2.3.2 Treatment. Depression in older persons is highly responsive to psychotherapy and antidepressant medication that is highly effective, nonaddictive, and usually well-tolerated. Often, the main problem lies in having the depression recognized in the first place, instead of its being ascribed by patient, family, and doctor to the inevitable experience of old age.

2.4 Suicide

In most countries, elderly men consistently have higher rates for suicide than other age-groups. Factors

associated with death by suicide in the elderly include depressive disorder, disabling physical illness, chronic pain, and social isolation.

2.5 Anxiety Disorders

Anxiety disorders in late life have been overlooked even more than depression as a significant cause of morbidity. In an integrative review of eight community-based surveys of persons aged 60 years and over, anxiety disorders were found to have a lower prevalence than in younger adults. The explanation is unknown. Co-occurrence of depression with generalized anxiety disorder or phobias was a frequent pattern, the depression usually being inappropriately treated with a benzodiazepine. Anxiety symptoms may also occur in the early stages of dementia and in association with depressive disorders. There is little information on the prevalence of post-traumatic stress disorder in the elderly, risk factors for it or its natural history.

2.6 Schizophrenia

It is recognized that schizophrenia can begin in later life. The term paraphrenia has been used for this condition. A number of distinguishing features may be present, as reviewed by Henderson and Kay (1997): a higher ratio of women, a better occupational history, and a higher frequency of marriage than in younger patients, with the occurrence of sensory loss being more common in those with onset at a more advanced aged. Psychotic symptoms are appreciably more common than formal diagnostic categories. Impaired hearing and vision have both been found to be associated with psychotic symptoms. Social isolation is a further associated factor, but it is unknown if this is a premorbid characteristic, a causal factor, or a consequence. There is consistent evidence that paranoid or delusional symptoms are more frequent in persons who also have cognitive impairment.

3. Services for the Elderly

The extent to which the elderly with mental disorders receive services is itself an issue amenable to systematic examination. The notion of filters proposed by Goldberg and Huxley (1992) is very useful. The first filter determines which persons with mental disorders reach a physician, whether they initiate the consultation themselves or whether it is initiated by their families or the physician. The second filter determines whether the physician recognizes that significant mental disturbance is present. This may happen in only about half the consultations. The third filter determines referral to specialist services. Although family physicians recognize early dementia quite efficiently, they may often need further training in the diagnosis of depressive disorders in elderly persons. This is because symptoms of tiredness, poor appetite, insomnia, and apathy may be mistakenly ascribed to old age. The benefits of correct diagnosis for both the patient and the community are self-evident.

4. Disability

From a public health perspective, information about the elderly and the services they require has to include an assessment of disability. This is because it is impairment of functioning in daily life rather than symptoms or diagnoses that brings about the need for services. Dementia is second only to arthritis as a cause of disability in the elderly. Less is known about the impairment of functioning caused by depressive disorders and depressive symptoms but, since they often occur with physical disease, their interactive effect is considerable. One community-based study conducted by an experienced geriatrician found that neuropsychiatric disorders of cognition, behavior, and mood, together with Parkinson-related disorders of gait, are major predictors of disability in late life.

5. Research in Geriatric Psychiatry

There has been an enormous expansion of research activity in recent decades. Much of it is directed at the following questions:

(a) What determines the course of cognitive decline in old age?

(b) Are there factors in the environment that contribute to dementia?

(c) Can dementia be prevented?

(d) Can drug treatments for dementia be found?

(e) What are the applications of the new neuroimaging techniques to diagnosis and treatment?

(f) What improvements are needed in primary health care for older persons; and how can these be brought about?

(g) What interventions will promote quality of life in the elderly?

The needs of tomorrow's communities, together with the rapid advances taking place in neuroscience, will ensure an exciting future for geriatric psychiatry. Its practitioners will need, as never before, to be highly trained but interpersonally skilled physicians.

Bibliography

Bird M in press A psychosocial approach to challenging behaviour in dementia. In: Mulligan R, Van der Linden M, Juillerat A C (eds.) *Clinical Management of Early Alzheimer's Disease*. Erlbaum, Mahwah, NJ

Breitner J C S 1996 Inflammatory processes and anti-inflammatory drugs in Alzheimer's disease: A current appraisal. *Neurobiology of Aging* **17**: 789–94
Caine E D, Grossman H, Lyness J M 1995 Delirium, dementia, and amnestic and other cognitive disorders and mental disorders due to a general medical condition. In: Kaplan H I, Sadock B J (eds.) *Comprehensive Textbook of Psychiatry VI*. Williams and Wilkins, Baltimore, pp. 705–54
Chandra V, Ganguli M, Pandav R, Johnston J, Belle S, DeKosky S T 1998 Prevalence of Alzheimer's and other dementias in rural India: The Indo-US study. *Neurology* **51**: 1000–8
Doll R 1993 Alzheimer's disease and environmental aluminium. *Age and Ageing* **22**: 138–53
Esquirol J E D 1845 *Mental Maladies: A Treatise on Insanity*, trans. Hunt E K. Lea and Blanchard, Philadelphia
Ganguli M, Chandra V, Gilby J E, Ratcliff G, Sharma S D, Pandav R, Seaberg E C, Belle S 1996 Cognitive test performance in a community-based non-demented elderly sample in rural India: The Indo-US Cross-National Dementia Epidemiology Study. *International Psychogeriatrics* **8**: 507–24
Goldberg D, Huxley P 1992 *Common Mental Disorders*. Tavistock/Routledge, London
Henderson A S 1994 Does ageing protect against depression? *Social Psychiatry and Psychiatric Epidemiology* **29**: 107–9
Henderson A S, Jorm A F 2000 Dementia: Disease definition, natural history and epidemiology. In: Maj M, Sartorius N (eds.) *Evidence and Experience in Psychiatry*. John Wiley, Chichester, UK
Henderson A S, Jorm A F, Korten A E, Jacomb P, Christensen H, Rodgers B 1998 Symptoms of depression and anxiety during adult life: Evidence for a decline in prevalence with age. *Psychological Medicine* **28**: 1321–8
Henderson A S, Kay D W K 1997 The epidemiology of functional psychoses of late onset. *European Archives of Psychiatry and Clinical Neuroscience* **247**: 176–89
Hendrie H C, Hall K S, Pillay N, Rodgers D, Prince C, Norton J, Brittain H, Nath A, Blue A, Kaufert J, Shelton P, Postl B, Osuntokun B 1993 Alzheimer's disease is rare in Cree Indians. *International Psychogeriatrics* **5**: 5
Hendrie H C, O'Osuntokun B O, Hall K S, Ogunniyi A O, Hui S L, Unverzagt F W, Gureje O, Rodenberg C A, Baiyewu O, Musick B S, Adeyinka A, Farlow M R, Oluwole S O, Class C A, Komolafe O, Brashear A, Burdine V 1995 Prevalence of Alzheimer's disease and dementia in two communities: Nigerian Africans and African Americans. *American Journal of Psychiatry* **152**: 1485–92
Jeste D V, Alexopoulos G S, Bartels S J, Cummings J L, Gallo J J, Gottlieb G L, Halpain M C, Palmer B W, Patterson T L, Reynolds C F, Lebowitz B D 1999 Consensus statement on the upcoming crisis in geriatric mental health. *Archives of General Psychiatry* **56**: 848–53
Jorm A F 2000 Risk factors for Alzheimer's disease. In: O'Brien J, Ames D, Burns A (eds.) *Dementia*. Arnold, London, pp. 383–90
Jorm A F, Korten A, Henderson A S 1987 The prevalence of dementia: A quantitative integration of the literature. *Acta Psychiatrica Scandinavica* **76**: 465–79
Ritchie K, Touchon J 2000 Mild cognitive impairment: conceptual basis and current nosological status. *Lancet* **355**: 225–8
Snowdon D A, Kemper S J, Mortimer J A, Greiner L H, Wekstein D R, Markesbery W R 1996 Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. *Journal of the American Medical Association* **275**: 528–32
St George-Hyslop P H, Westaway D A 1999 Antibody clears senile plaques. *Nature* **400**: 116–17
United Nations, Department of Economic and Social Affairs, Population Division 1998 *World Population Projections to 2150*. United Nations, New York
World Health Organization 1992 *The ICD-10 Classification of Mental and Behavioural Disorders. Clinical Descriptions and Diagnostic Guidelines*. World Health Organization, Geneva, Switzerland

S. Henderson

Gerontological Clinical Psychology

Clinical geropsychology is the field that specifically focuses on developing empirically supported skills for the psychological assessment and treatment of older adults. Adults who might merit such attention include those with psychopathology or whose quality of life is negatively influenced by psychosocial factors. Given the increasing population of older adults and the complex life experiences and changing demographics of the population, it is important for mental health professionals to be prepared to assess and treat older clients.

1. Proficiency in Clinical Geropsychology

Until recently, outpatient mental health care services have been underutilized by older adults, despite 1975 legislation mandating specialized services for mentally ill older adults within community mental health centers. The majority of the elderly who receive mental health services are seen as inpatients during hospitalization in a psychiatric hospital or a nursing home. Possible explanations for this phenomenon include the idea that there is stigma attached to receiving mental health services within this cohort, but this concern has not been supported in limited research (Rokke and Scogin 1995). Alternatively, mental health professionals have historically displayed 'professional ageism' dating back to Freud, who was pessimistic about psychological change or the benefits of therapy in later life. Finally, Medicare reimbursement for psychological services in the United States is very limited. The majority of the money spent goes toward psychopharmacology as opposed to psychotherapy, reducing the availability of psychological services.

However, patterns of usage of psychological services by the elderly are changing. Successive cohorts of older persons have higher levels of education and a greater acceptance of psychology. Rokke and Scogin (1995), for example, showed that older adults considered cognitive therapy to be more credible and acceptable than drug therapy for depression (see

Cognitive Therapy). It is likely that psychologists could be much more active in reaching out to older adults to provide services, and that they can expect an increasingly positive welcome.

As more older adults seek or accept psychological services, psychologists must be prepared to meet this need. To address issues of responsibility and competency in providing such care, the American Psychological Association (APA) has approved Clinical Geropsychology as an area of special proficiency. Guidelines addressing the needed competencies for proficiency in clinical geropsychology have been drafted and widely promulgated (APA Interdivisional Task Force 1999). In this document, 13 areas of knowledge and skills that define proficiency knowledge in Clinical Geropsychology are specified, as listed below and then reviewed in more detail:

(a) Research and theory in aging
(b) Cognitive psychology and change
(c) Social/psychological aspects of aging
(d) Biological aspects of aging
(e) Psychopathology
(f) Problems in daily living
(g) Sociocultural and socioeconomic factors
(h) Assessment
(i) Treatment
(j) Prevention and crisis intervention services
(k) Consultation
(l) Interface with other disciplines
(m) Special ethical issues in providing services to the aged.

1.1 Research and Theory in Aging

It is important for psychologists to understand aging in a broad context, as a psychological, biological, and social process. Further, some research methodology issues are especially relevant to aging, in particular understanding the impact of cohort or group differences by age.

1.2 Cognitive Psychology and Change

Changes in cognitive and behavioral function are common with aging; most such changes are relatively minor, such as increased difficulty with short-term memory or recall memory. However, the prevalence rate of more severe cognitive changes, indicating dementia, also increases with age. At age 65 only 1 percent of the population have dementia, compared to 30–35 percent by age 85 and 50 percent by age 90. The concept of dementia has evolved from the nonspecific 'organic brain syndrome' to a more precise picture of a complex disorder encompassing several distinctive disease entities, each with its own specific set of treatment implications (see *Dementia: Psychiatric Aspects*). More detail is provided on cognitive changes with age under proficiencies Sect. 1.5, Psychopathology, Sect. 1.8, Assessment, and Sect. 1.9, Treatment.

1.3 Social/Psychological Aspects of Aging

Aging is a personal process and a social process that occurs in a context of familial, cultural, and political elements. As reviewed above, the demographics of aging reflect a changing social context of aging. This changing context also affects intergenerational relationships, the nature of relationships in later life (e.g., since more adults are living to old age, larger peer networks of older adults exist than in prior generations), and perceptions of aging.

1.4 Biological Aspects of Aging

Important biological aspects of aging include normal biological aging changes, abnormal changes and disease, lifestyle and behavioral factors in health, chronic and terminal illness, and pharmacological issues. Chronic health problems (e.g., arthritis, hypertension, hearing impairment, heart conditions) are often experienced by older adults. Many of these physical conditions are associated with chronic pain, and psychologists can employ strategies to help with pain management. Because of health concerns, many older adults are given medications, often without good explanation of the importance of taking the medication routinely, resulting in erratic usage. Psychologists can work with older adults to encourage consistent use of prescribed medication, to help them understand its importance, and to help identify the health-related belief system of older adults and incorporate those beliefs into the medical regimen. Insomnia, particularly sleep-maintenance insomnia, is an age-related and debilitating condition. Behavioral procedures for treating insomnia in older adults are more effective than relaxation training or imagery training. In addition, they offer the ability to avoid or discontinue sleep-inducing medications that can cause iatrogenic problems (Morin et al. 1999).

1.5 Psychopathology and Aging

The best estimate of the rate of psychopathology in the older population in the United States is approximately 22 percent (Gatz and Smyer 1992), but there are huge differences among different groups of older adults and across different types of psychopathology. Specifying the prevalence and incidence of psychopathology in older adults is complicated by problems with methodology and design. Most studies on older adults do not take racial and or cultural distinctions into account in the manifestation of disease. Finally, studies do not

always agree on diagnostic criteria and terminology, thus yielding imprecise estimates. For all disorders, the accepted criteria for older adults are exactly the same as those for younger adults using the *Diagnostic and Statistical Manual-IV* (*DSM-IV*) diagnostic system. However, interpreting some of the criteria for each disorder may be complicated by other aspects of the aging process. In the following subsections, some of the most common mental disorders will be examined and, when appropriate, issues that make diagnosis more complex in older adults will be addressed.

Older adults have a lower rate of major depressive disorder than middle-aged adults; however, there is a higher prevalence of subclinical symptoms in older adults (Gatz et al. 1996). Data from a number of studies indicate that across the adult life span, the highest depression scores are found among younger adults and persons 75 years and older. As with younger adults, women show higher rates of depression than men, although studies suggest that by age 80, men and women have an equal rate of depressive symptomatology. A subset of depressed older adults that warrant particular concern is white men aged 60–85, particularly those with medical illness or living alone; this group has the overall highest rate of suicide.

Life events, such as deaths of loved ones and change in residence, have shown an inconsistent relationship with depression. Although these negative life events occur more commonly in older adults, some scholars argue that older adults have experienced a number of stressors earlier in life, and/or that older adults are better able to cope and regulate emotion more effectively (e.g., Carstensen and Charles 1998), causing the relationship between negative life events and depression to be less strong among older adults. Physical illness, although a statistically major risk factor for depression, probably is not the causal variable in this relationship. Instead, loss of physical function, such that one cannot take care of basic needs (e.g., dressing, bathing) or must give up valued activities, is the significant vulnerability factor (Zeiss et al. 1996; see also *Depression, Clinical Psychology of*).

Anxiety disorders in older adults occur more often than depression, with estimates of approximately 5.5 percent prevalence (e.g., Beck and Stanley 1997). However, in comparing younger and older adults, the rate of anxiety is lower in older adults than in any other age group. Older women are more likely to be diagnosed with anxiety than older men. Although more prevalent than depression, anxiety disorders have been less studied in older adults. Panic disorders, phobias, generalized anxiety disorder, and post-traumatic stress disorder in older veterans have received the greatest attention.

Dementia is characterized by loss of cognitive function sufficient to impair performance of daily activities. Most dementias are due to irreversible changes in the brain; the most common are Alzheimer's disease and vascular dementias. Reversible causes occur in 10–20 percent of dementia patients, as a response to a medication or physical illness, such as thyroid dysfunction, vitamin B12 deficiency, or a treatable brain lesion. Usually, when reversible causes of dementia are treated, good recovery ensues with return to prior level of cognitive functioning. Since most forms of dementia are progressive, it is important to identify the disease early. However, assessment of dementia in older adults is difficult; in particular, patients differentiating depression and dementing illness can be difficult to differentiate (Kaszniak and Scogin 1995). Alzheimer's disease affects not only the person with the disease, but also the family (Thompson and Gallagher-Thompson 1996). Family members experience a range of emotions that can include guilt, anger, and depression. There are also positive components of providing care for a person with Alzheimer's, and many family members report caregiving as a rewarding experience.

1.6 Problems in Daily Living

Clinical geropsychologists need to be aware of many problems that can interfere with the ability to carry out normal daily activities for older adults. These include deficits in social and daily living skills, decision-making capacity, and elder abuse and neglect. Of these issues, elder abuse is one of the most difficult to handle; law in all states mandates reporting of elder abuse; the primary goal of reporting is to stop the abuse.

1.7 Sociocultural and Socioeconomic Factors

A variety of sociocultural issues are important in understanding the experience of aging, including losses from a variety of sources (deaths, end of employment, limitations in functioning, financial problems, environmental restrictions); cross-cultural and minority aging issues (e.g., interacting with older minorities to convey respect); the role of religion and spirituality; and the roles of intergenerational acquaintances, lifelong friendships, intimate relationships, and other social relations.

1.8 Special Issues in the Assessment of Older Adults

Clinical geropsychologists are asked to evaluate older adults with regard to presence or absence of psychopathology, cognitive capacity, competency to make decisions, substance abuse, etc. As with younger adults, the techniques used in psychological assessments include clinical interviewing, life history record and data review, cognitive and neuropsychological evaluations, behavioral assessments, and situational observations; however, with older adults, psycholo-

gists need to use different norms for tests and to include cognitive testing more frequently. With patients who exhibit disruptive and/or potentially harmful behavior (e.g., wandering, shouting, assaultiveness), behavioral assessment is helpful in determining what type of techniques would be useful for working with the patient (see Lichtenberg 1999 for thorough overview).

1.9 Treatment of Older Adults

Psychotherapy with older adults is as effective as therapy with younger adults. Most research on psychotherapy with older adults has used cognitive-behavioral approaches, and these have been shown to be effective for a wide variety of problems (Scogin and McElreath 1994). There also is a small amount of literature supporting the use of other approaches, such as Interpersonal Therapy and Brief Time-limited Psychodynamic Therapy, especially with depression in older adults (see Zarit and Knight 1996).

There are few major content differences in therapy with older compared to younger adults. Older adults have more health problems resulting in functional impairment, and psychological status is often related to their functional status. In addition, older adults may face obstacles in terms of resources for supporting adequate quality of life, such as limited financial resources or transportation, or the experience of loss of friends or family. The largest content difference is the need to deal more frequently with older adults with therapy issues related to death and dying. Many older adults accept death as part of the life course and engage in active adaptations to death, for example, by making wills, leaving instructions, negotiating interpersonal conflicts, and reintegrating life experiences in a meaningful way, but others have greater emotional difficulty in facing death or dealing with losses of loved ones. Clinical geropsychologists must be sensitive to factors that influence the experience of death and dying, including gender, class, economic variables, and cultural considerations. Recent studies have shown that bereavement practices, once highly structured and culturally determined, are becoming diverse; as a consequence, what practices will be therapeutic is becoming more individualized rather than shared within a common culture (Riley 1992).

With regard to therapy process, chronological age is not a good marker for what adaptation of therapeutic strategies originally developed with younger adults may be needed. Therapy adaptations to enhance utility with older adults might be helpful or even necessary, but assessment of each client determines which adaptations are appropriate (see Zarit and Knight 1996 for a thorough summary of suggestions regarding therapy adaptations). Primary areas of adaptation include changes to compensate for memory and information processing deficits with older adults. As a result of cognitive changes, the pace of therapy may be slower than with younger adults. More repetition of material may be necessary, and processing of new ideas may be slower. Memory aids, such as audiotapes of therapy sessions or written summaries of each session, may be helpful. It may help to present material in multiple ways, both because of potential sensory loss and because repetitions provide multiple routes to memory storage. A key phrase for therapists working with older adults is 'Say it, Show it, Do it': state it clearly, write it down, and help the client use the idea in a specific way.

On the positive side, some changes in therapy process reflect relative strengths of older adults. One such strength can be thought of as wisdom (Baltes and Staudinger 1993), which is an age-related, but not inevitable, consequence of life experience. Older adults have faced many difficult life experiences, and most can abstract helpful information from those experiences and describe personal skills that have helped them handle adversity. Showing respect for the client's accumulated experience can enhance therapy.

1.10 Prevention and Crisis Intervention Services with Older Adults

Outreach, referral, and early intervention are important for making assessment and therapy available to older adults. The organization of many health care systems presents significant obstacles to good care for older adults who may need care provided in their homes, need short, easy drives to reach services, need care for the patient and the caregiving system, etc. In addition, providing health promotion resources to older adults is useful yet often neglected. For example, exercise is effective in improving physical and mental health for older adults, but may not be recommended or supported by health care providers to the extent it is with younger adults.

1.11 Consultation

Since most health care providers are not well trained in providing services to older adults, yet the demand is great and growing, clinical geropsychologists frequently need to provide consultation. Consultation to other professionals is a major component of this work and includes providing specialized information to health care institutions, agencies, and community organizations; providing staff training; and consulting on program development. In addition, consultation to families, lay caregivers, and self-help and support groups is needed.

1.12 Interface with Other Disciplines

Since typical problems older adults face are multimodal, a comprehensive interprofessional approach

to assessment and treatment that encompasses diverse health care providers (e.g., geriatricians, nurses, social workers) as well as psychologists is usually needed. The *biopsychosocial* model combines clinical and scientific approaches to treatment by defining problems in relation to the interactions among biological, social, and psychological systems. This model captures the complexity inherent in addressing the problems of older adults and highlights the need for an interprofessional approach (Zeiss and Steffen 1998).

1.13 Special Ethical Issues in Providing Services to Older Adults

Special ethical issues with older adults include obtaining informed consent for treatment with cognitively impaired elders, balancing patient autonomy and self-determination with apparent needs for care, balancing competing interests between older adults and family members, role conflicts in nursing homes, and confidentiality issues in working with families and teams.

2. Summary and Future Directions

Older adults are a heterogeneous population in terms of values, motives, social and psychological status, and behavior, as well as a quickly growing one. The need for clinical psychologists to be more informed about work with this diverse population of older adults have been summarized in the definition of proficiency in clinical geropsychology. That definition emphasizes broad-based knowledge of aging and adaptations of assessment and treatment for older adults. It also emphasizes the need for psychologists to work as part of interprofessional teams to meet the broad biological, social, and psychological needs of older adults.

However, as helpful as this summary is, much of what we currently know is based on research with the current cohort of older adults, particularly those who are Caucasian Americans. That research does not fully represent the heterogeneity within older adults now and will be less representative of future generations of older adults. In particular, there is considerable speculation that the large number of children born after World War II (now nearing older adult status) may age differently; two competing issues have been raised. First, it is often suggested that this cohort of elders may be more open to psychological interventions, escalating needs for research on assessment and therapy with older adults and for enhanced training of all clinical psychologists regarding work with older adults. On the other hand, coming generations of older adults may be healthier and more functional (Waidmann and Liu 2000), reducing depression and other psychological sequelae of disability, thus reducing the need for psychological interventions. The increasing proportion of minority elders also will create challenges; as current generations of minority Americans age they will carry with them salient features such as a continuing imbalanced sex ratio, segregated distribution, and proportion in poverty. Clearly, researchers and clinicians in clinical geropsychology need to be guided by research to date but open to dramatic shifts in the field.

See also: Geriatric Psychiatry

Bibliography

American Psychological Association Interdivisional Task Force on Qualifications for Practice in Clinical and Applied Geropsychology 1999 Qualifications for practice in clinical geropsychology, Unpublished Working Draft No. 7. Division 12, Section II and Division 20, American Psychological Association, Bethesda, MD

Baltes P B, Staudinger U M 1993 The search for a psychology of wisdom. *Current Directions in Psychological Science* **2**: 75–80

Beck J G, Stanley M A 1997 Anxiety disorders in the elderly: The emerging role of behavior therapy. *Behavior Therapy* **28**: 83–100

Carstensen L L, Charles S T 1998 Emotion in the second half of life. *Current Directions in Psychological Science* **7**: 144–9

Gatz M, Kasl-Godley J E, Karel M J 1996 Aging and mental disorders. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 365–82

Gatz M, Smyer M A 1992 The mental health system and older adults in the 1990s. *American Psychologist* **47**: 741–51

Kaszniak A W, Scogin F R 1995 Assessment of dementia and depression in older adults. *The Clinical Psychologist* **48**: 17–23

Lichtenberg P A (ed.) 1999 *Handbook of Assessment in Clinical Gerontology*. John Wiley and Sons, New York

Morin C M, Mimeault V, Gagne A 1999 Nonpharmacological treatment of late-life insomnia. *Journal of Psychosomatic Research* **46**: 103–16

Riley J W 1992 Death and dying. In: (eds.) *Encyclopedia of Sociology*. Macmillan, New York

Rokke P D, Scogin F 1995 Depression treatment preferences in younger and older adults. *Journal of Clinical Geropsychology* **1**: 243–57

Scogin F, McElreath L 1994 Efficacy of psychosocial treatments for geriatric depression: A quantitative review. *Journal of Consulting and Clinical Psychology* **62**: 69–74

Thompson L, Gallagher-Thompson D 1996 Practical issues related to maintenance of mental health and positive well-being in family caregivers. In: Carstensen L L, Edelstein B A, Dornbrand L (eds.) *The Handbook of Clinical Gerontology*. Sage Publishing, Thousand Oaks, CA, pp. 129–50

Waidmann T A, Liu K 2000 Disability trends among elderly persons and implications for the future. *Journal of Gerontology: Social Sciences* **55B**: S298–307

Zarit S H, Knight B (eds.) 1996 *A Guide to Psychotherapy and Aging: Effective Clinical Interventions in a Life-stage Context*. American Psychological Association Press, Washington, DC

Zeiss A M, Lewinsohn P M, Rhode P, Seeley J R 1996 Relationship of physical disease and functional impairment to depression in older people. *Psychology and Aging* **11**: 572–82

Zeiss A M, Steffen A 1998 Interdisciplinary health care teams in geriatrics: An international model. In: Bellack A S, Hersen M (eds.) *Comprehensive Clinical Psychology*. Pergamon, Oxford, UK, Vol. 7

A. M. Zeiss

Gesture in Linguistics

When we communicate, we not only produce a string of words, but also orchestrate various potentially interpretable body movements such as gesture, gaze shift, holding of a certain posture, interpersonal spacing, self-touch, and bodily symptoms of emotional and physiological states. From the totality of these activities in the communicative situation, this article takes up gesture and examines its relationship to language. In the first section of the article the definition of gesture is discussed and the following sections review its role in conversation, their relation to linguistic semantics and pragmatics, and mental processes underlying speech production.

1. *What is Gesture?*

Though there is controversy as to how to classify nonverbal behaviors (Efron 1941, Ekman and Friesen 1969, McNeill 1992), the central cases of nonverbal behaviors that have been designated by the term 'gesture' can be captured by the following definition. Gesture is nonpractical, nonlinguistic discrete body movement that arises from the intention to express something, and that encodes some aspects of what is intended to be expressed and/or how the accompanying speech should be interpreted.

Gestures are discrete spurts of body movement, which occur in the background of a period of nongesturing. This is similar to verbal utterances occurring in the background of silence. The sporadic nature of gestures contrasts with the inherently continuous nature of other nonverbal semiotic codes such as posture, interpersonal spacing, and gaze.

Gesture is nonpractical in the sense that gestural body movement is not involved in achieving nonexpressive practical goals. For example, the way one holds a cigarette may be used intentionally to express a certain attitude or social affiliation, but it is not considered to be gesture because the body movement serves a practical purpose for smoking.

Gestures arise from the intention to express something. Gesture can be the intended code for transmission of the information. Raising a hand when taking a vote in a meeting is such a case. However, gesture can arise in the situation where gesture seems not to be the code intended for the recipient. For example, it is not uncommon to produce direction-indicating gestures as one gives directions on the telephone. In such cases, the gestural body movement is still caused by the intention to express the instruction for how to get to the destination.

In this respect, gesture contrasts with self-touch, such as rubbing the chin or fiddling with clothing, in which there is no apparent intention to express. Self-touch also appears in certain noncommunicative settings, for example, while solving a crossword puzzle alone. The presence of expressive intention also distinguishes gesture from bodily symptoms for certain emotional and physiological states, such as blushing. However, some of the affect-displays, such as frowning, can be produced under voluntary control. They can be intentionally produced to express certain affect. This type of behavior blurs the boundary between what is a gesture and what is not.

The issue of intentionality has raised both conceptual and methodological concerns in defining gesture. Many authors maintain that intentionality is an essential ingredient of gesture (Ekman and Friesen 1969, Kendon 1981). However, difficult questions arise once one tries to tackle the issue squarely. Does intentionality have binary states (presence vs. absence) or is it a matter of degree? How can one tell what is intended to be expressed? How can one tell whether a given act is intentional, or how can one assess the degree of intentionality? Kendon (1981) suggested that a solution to the problem of identifying intentionality may be to focus on whether communicational partners treat each other's acts as intentional or unintentional. However, this does not solve the problem for gestures that cannot be seen by the communicational partners, as in the case of gesturing during telephone conversation.

Going back to further definitional features of gesture, gesture encodes some aspects of what is intended to be expressed and/or how the accompanying speech should be interpreted. In other words, gestural body movement itself is the carrier of the information to be communicated. In this respect, gesture differs from 'placing,' in which the body moves an object to a meaningful position (Clark in press). For example, putting merchandise on the cashier's counter signals the intention to purchase the merchandise. In placing, the body movement brings about a meaningful spatial configuration, but in gesture, the body movement itself carries the meaning.

2. *In What Way is Gesture Nonlinguistic?*

One of the defining features of gesture is its nonlinguistic nature. Gesture is distinct from sign language, which is language that transmits codes in the visuo-motor modality (see *Sign Language*). It bears different semiotic properties from language (Kendon 1993, McNeill 1992). Gestures vary as to the degree to which they are semiotically different from language.

2.1 Most Language-like Gestures—Emblems

Gestures that are the most language-like are 'emblems' (Efron 1941, Ekman and Friesen 1969) such as the Euro–American OK sign with a ring formed by the thumb and the index finger. Emblems share some properties with words in a language. Emblems have standards of form. Their form-meaning relationship shows a relatively high degree of arbitrariness. Standards of form and form-meaning pairing are shared by a community of users (Calbris 1990, Morris et al. 1979).

Despite the above similarities to words, emblems also exhibit important differences. First of all, emblems are extremely limited in their potential for sequential combination, that is, combinatorial rules that operate over multiple emblems, namely, something equivalent to syntax, are impoverished.

Furthermore, Sparhawk (1978) showed that Persian emblems are not built up on the basis of a unified system of contrasting formational features, namely, something equivalent to phonology. It was found that the palm facing up vs. down is a formational feature that contrasts many Persian emblems. However, for the emblems with a certain class of meaning, namely beckoning emblems (e.g., the 'come here' emblem), this feature is not contrastive, and thus the palm can be facing either up or down. This is because the feature is considered to be irrelevant for the iconic make-up of the emblems.

The conclusion that emblems are not governed by unified 'phonology' has an important consequence for the 'Kinesics' approach to the analysis of semiotic body movements (Birdwhistell 1970), which is inspired by structural linguistics. In this approach, the existence of a system of 'kinemes,' the gestural equivalent of phonemes, is a fundamental assumption. However, even in the most language-like class of gestures, 'kinemes' do not constitute a unified system.

In summary, emblems differ from words in a language in that emblems are not fully unified by formal principles that relate emblems with one another. In other words, formal systems that are equivalent to syntax and phonology organize emblems only to a very limited extent.

2.2 Least Language-like Gestures—Gesticulations

Gestures that are least language-like are called 'gesticulations' (Kendon 1993) or simply 'gestures' (McNeill 1992), which spontaneously accompany speech, and usually share semantic and/or discourse-pragmatic contents with the concurrent speech.

Gesticulation consists of two subtypes: 'beats' and 'representational gestures.' Beats are small rhythmic bidirectional movements, in which hands typically move up and down or flip out and back. They occur when the concurrent speech should be interpreted in a context different from the ordinary flow of discourse (e.g., speech is being repaired, or speech is expressing the beginning of a new discourse unit) (McNeill 1992). Representational gestures express location, direction, action, movement of objects, and shape of object, by means of iconicity (i.e., resemblance between the gestural shape and the referent) and/or spatial indexicality (i.e., the spatial contiguity between gesture and the referent). Representational gestures can be metaphorical in the sense that they can spatialize abstract ideas (Calbris 1990, McNeill 1992).

Gesticulations share fewer semiotic properties with language than emblems. The form of gesticulations is less constrained by conventionalized standards. The iconic representation in a gesticulation is largely idiosyncratic moment-by-moment creation. Thus, gesticulations are flexibly shaped in accordance with the variation in the spatial or actional characteristics of the referent. A sequence of gesticulations can comprise a larger meaning unit only by virtue of iconicity, as a multigesture depiction of a 'picture' of the referent(s).

3. In What Ways Does Gesture Have an Essential Connection to Language?

In the previous section, the difference between gesture and language as semiotic systems was highlighted. Gesture, however, has essential connections to interactional, conceptual, and psychological aspects of language.

3.1 Gesture in Face-to-face Conversation

Gestures contribute to face-to-face conversation in various ways. First, gesture can express some aspects of the content to be conveyed, which may or may not be expressed in the concurrent speech, and gesturally expressed information is taken up by the interlocutor (Kendon 1994). The interlocutor's attention to gesture is also evidenced by the fact that gestures elicit verbal acknowledgement like 'uh huh,' and nonverbal acknowledgement like nodding from the interlocutor (Goodwin and Goodwin 1986, Streeck 1994).

This leads to the second way in which gestures contribute to conversation. Gesture is a part of complex cues as to when nonspeaking interlocutors can jump in and take the speaking turn (Schegloff 1984, Goodwin and Goodwin 1986, Streeck and Hartge 1992). One of the relevant features of gestures is that they are initiated slightly prior to the portion of speech that expresses the related content, and their temporal extent typically overlaps with the co-expressive speech (Schegloff 1984, Morrel-Samuels and Krauss 1992, McNeill 1992). The temporal precedence of gesture projects that the speaker is going to deliver speech further, and this can preempt the interlocutor's attempt to take the turn.

The third way in which gestures contribute to conversation is to mark discourse-pragmatic status of the concurrent speech. Kendon (1995) reports Southern Italian emblems that indicate discourse-pragmatic status of the concurrent speech. More specifically, they mark the topic-comment structure and the illocutionary force (e.g., demanding information, making a plea for understanding) of an utterance.

3.2 Gesture and Deixis

Gesture plays a crucial role in spatial deixis (see *Deixis*), in which the attention of the interlocutor is directed to the direction along which the referent (or the physical clue to the referent) can be found.

Gesture also plays a crucial role in 'abstract deixis,' in which a pointing gesture creates a referent in seemingly empty space in front of the speaker (McNeill et al. 1993). The speaker can point repeatedly to a locus in the empty space, as the same entity is referred to in the concurrent speech. The referent that such gestures spatialize can be something that could exist in space (e.g., a person who is not present), or something inherently nonspatial (e.g., time, plot-line of a story). Thus, the role of gesture in deixis is not only a passive one, in which a gesture is oriented toward a pre-existing target, but also an active and creative one, in which the act of gesturing materializes conceptualization of the speaker as a spatial entity in the extralinguistic context.

3.3 Gesture and Speech Production

Gesture and speech are produced in close temporal and informational coordination. As noted in Sect. 3.1, a gesture typically is initiated prior to the co-expressive portion of speech, and the gesture and the co-expressive speech temporally overlap. In addition, the production of gesture has a systematic relationship to intonational units. A gesture rarely crosses tone unit boundaries, and typically, there is one gesture per tone unit (Kendon 1980, McClave 1991). These findings suggest that there is a mental process that coordinates speech and gesture production (Kendon 1980, Morrel-Samuels and Krauss 1992, McNeill 1992) (see *Speech Production, Psychology of*; *Speech Production, Neural Basis of*).

3.3.1 Gesture and ideophones. The exact meeting point of gesture and speech production processes is a highly debated issue (see McNeill 2000). One phenomenon that sheds some light on this issue is the fact that linguistic entities that have imagistic and emotive meaning are highly associated with iconic representational gesture. One such entity is the class of words called 'ideophones' (also referred to as 'mimetics'). Ideophones are sound-symbolic words, for which native speakers report vivid imagery and emotion as their meaning. They include onomatopoetic words such as animal cries, but their referential domains go far beyond sound emitting entities (e.g., color, smell, wetness, density, emotion, and manner of motion).

Japanese is one of the languages in which ideophones comprise a large word class and they are commonly used. It was found that when Japanese speakers used an ideophone in narrative, it was almost always accompanied by a co-expressive iconic gesture (Kita 1997). This suggests that iconic gestures and ideophones are produced from common underlying mental representation that is imagistic in nature. In other words, iconic gestures are produced from imagistic representation that is evoked in the speaker's mind.

3.3.2 Gesture and psycholinguistics. Gestures that spontaneously accompany speech can be used as a 'window' into speech production processes, through which one can gain insights into the speaker's on-line conceptualization (McNeill 1992). In such a line of inquiry, representational gestures have been in the focus. Since the shape of representational gestures is not completely determined by convention, there are ample degrees of freedom left for idiosyncratic expression of the speaker's thought.

Representational gestures are created on-line by the speaker, and are typically co-expressive with the concurrent speech in the sense that they express the same object, action, event, or spatial relation as the speech, albeit from a different perspective and with a different mode of expression. As discussed above, the form of a representational gesture and its meaning are mediated by iconicity and spatial indexicality. Furthermore, ideophones, whose semantics is inherently imagistic, are almost always produced with a co-expressive iconic gesture. In short, representational gesture is imagistic, idiosyncratic, and non-linguistic; in contrast, speech is mainly propositional, and based on a conventionalized system of signs.

It has been argued that the underlying mental representation for a gesture-speech complex is produced by a single mental process involved in the verbalization of thought (Kendon 1980, McNeill 1992). McNeill stresses the significance of the semiotic duality of a gesture-speech complex. He argues that when the speaker tries to verbalize thought, he or she is engaged in two kinds of thinking: imagistic thinking and linguistic thinking. When the speaker finds a connection between idiosyncratic imagery and socially-constituted linguistic signs, a mental seed from which a gesture-speech complex develops is formed. In this process, imagery is grounded in a socially shared linguistic system, and words are con-

nected to a specific image in the speaker's mind. According to McNeill, the formation of a unified representation that integrates idiosyncratic and social sign systems is a crucial initial step in the verbalization of thought.

It has been suggested that gestures are not only the reflection of the speaker's thought, but they also facilitate verbalization of thought. There is controversy as to the exact nature of the facilitation (see McNeill 2000). Some argue that representational gestures facilitate retrieval of lexical items from the mental lexicon (Morrel-Samuels and Krauss 1992), but others argue that conceptual planning for speaking is facilitated (Alibali et al. 2000).

4. Conclusion

Gesture is a semiotic system distinct from language. At the same time, it is intimately linked to various aspects of linguistic performance. Gesture is an integral part of the organization of face-to-face conversation. Pointing gesture is an essential part of referential acts by means of spatial deixis. The production of intonation and ideophones are tightly coupled with the production of a gesture that is coexpressive. The speaker's imagistic thinking during verbalization of thought is revealed by gestures that spontaneously accompany speech. Because of the interactional, referential, and psychological unity of gesture and language, the understanding of language is not complete unless gesture is taken into account.

See also: Communication and Social Psychology; Communication: Philosophical Aspects; Motor Control; Nonverbal Communication, Social Psychology of; Verbal Communication, Psychology of

Bibliography

Alibali M W, Kita S, Young A 2000 Gesture and the process of speech production: We think therefore we gesture. *Language and Cognitive Processes* **15**: 593–613

Birdwhistell R L 1970 *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania Press, Philadelphia, PA

Calbris G 1990 *The Semiotics of French Gestures*. Indiana University Press, Bloomington, IN

Clark H in press Pointing and placing. In: Kita S (ed.) *Pointing: Where Language, Cognition, and Culture Meet*. L. Erlbaum Associates, Mahwah, NJ

Efron D 1941 *Gesture and Environment*. King's Crown Press, New York

Ekman P, Friesen W V 1969 The repertoire of nonverbal behavioral categories: Origins, usage and coding. *Semiotica* **1**: 49–98

Goodwin M H, Goodwin C 1986 Gesture and coparticipation in the activity of searching for a word. *Semiotica* **62**(1/2): 51–75

Kendon A 1980 Gesticulation and speech: Two aspects of the process of utterance. In: Key M R (ed.) *The Relation between Verbal and Nonverbal Communication*. Mouton, The Hague

Kendon A 1981 Introduction: Current issues in the study of 'nonverbal communication.' In: Kendon A, Sebeek T A, Umiker-Sebeok J (eds.) *Nonverbal Communication, Interaction, and Gesture: Selections from Semiotica*. Mouton, The Hague

Kendon A 1993 Human gesture. In: Gibson K R, Ingold T (eds.) *Tools, Language and Cognition in Human Evolution*. Cambridge University Press, Cambridge, UK

Kendon A 1994 Do gestures communicate? A review. *Research in Language and Social Interaction* **27**: 175–200

Kendon A 1995 Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics* **23**: 247–79

Kita S 1997 Two-dimensional semantic analysis of Japanese mimetics. *Linguistics* **35**: 379–415

McClave E 1991 *Intonation and Gesture*. Unpublished dissertation, Georgetown University

McNeill D 1992 *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago

McNeill D (ed.) 2000 *Language and Gesture*. Cambridge University Press, Cambridge, UK

McNeill D, Cassell J, Levy E T 1993 Abstract deixis. *Semiotica* **95**(1/2): 5–9

Morrel-Samuels P, Krauss R M 1992 Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**: 615–22

Morris D, Collett P, Marsh P, O'Shaughnessy M 1979 *Gestures, Their Origins and Distribution*. Stein and Day, New York

Schegloff E A 1984 On some gestures' relation to talk. In: Atkinson J M, Heritage J (eds.) *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, Cambridge, UK

Sparhawk C M 1978 Contrastive-identificational features of Persian gesture. *Semiotica* **24**: 49–86

Streeck J 1994 Gesture as communication II: The audience as coauthor. *Research on Language and Social Interaction* **27**: 239–67

Streeck J, Hartge U 1992 Previews: Gestures at the transition place. In: Auer P, di Luzio A (eds.) *The Contextualization of Language*. J. Benjamins, Amsterdam

S. Kita

Gifted and Talented Individuals: Developmental and Educational Overview

Some individuals stand out among their peers because of their remarkable abilities; they are called gifted or talented. What do these two terms mean? How many can be labeled gifted or talented? Are some youngsters much more gifted/talented than others? Are they more at risk developmentally than their average ability peers? What are their major educational needs, and how do school systems answer these needs? These are the major themes covered by this brief overview of a field traditionally named Gifted Education.

1. The Nature of Giftedness and Talent

Most published definitions of giftedness and/or talent use these two terms as synonyms (Sternberg and Davidson 1986). Yet, most scholars recognize at least implicitly two distinct manifestations of that phenomenon: (a) a natural, nondeveloped form manifested as precocious early development and presented as a 'potential', and (b) an adult well-developed form characterized by outstanding achievements in a field of human activity. Gagné (1993) proposed a developmental model that uses this distinction to create differentiated definitions for the two key concepts of giftedness and talent.

1.1 Gagné's Differentiated Model of Giftedness and Talent (DMGT)

In the DMGT, *giftedness* designates the possession and use of high natural abilities (called aptitudes) in at least one of four ability domains (see Fig. 1), so that the level of performance places the person among the top 10 percent of same age peers. *Talent* is the demonstration of systematically developed and trained abilities in any field of human activity at a level such that the individual is among the top 10 percent of peers having had equivalent training. The six components of the DMGT bring together in a dynamic way all the recognized determinants of talent emergence in any field of human activity.

Natural abilities, which have a partial genetic origin, can be observed more directly in young children because environmental influences and systematic learning have exerted their moderating influence in a more limited way. However, they still manifest themselves in older children, even in adults, through the *ease and speed* with which individuals acquire new skills. The easier or faster the learning process, the greater the gifts. Many laypersons call these gifts 'natural talent.' Talents progressively emerge from the transformation of high aptitudes into the well-trained and systematically developed skills typical of a particular field of human activity. Thus, natural abilities act as the raw materials or *constituent elements* of talents: one cannot become talented without being gifted. However, the converse is not true: it is possible for outstanding natural abilities not to be translated

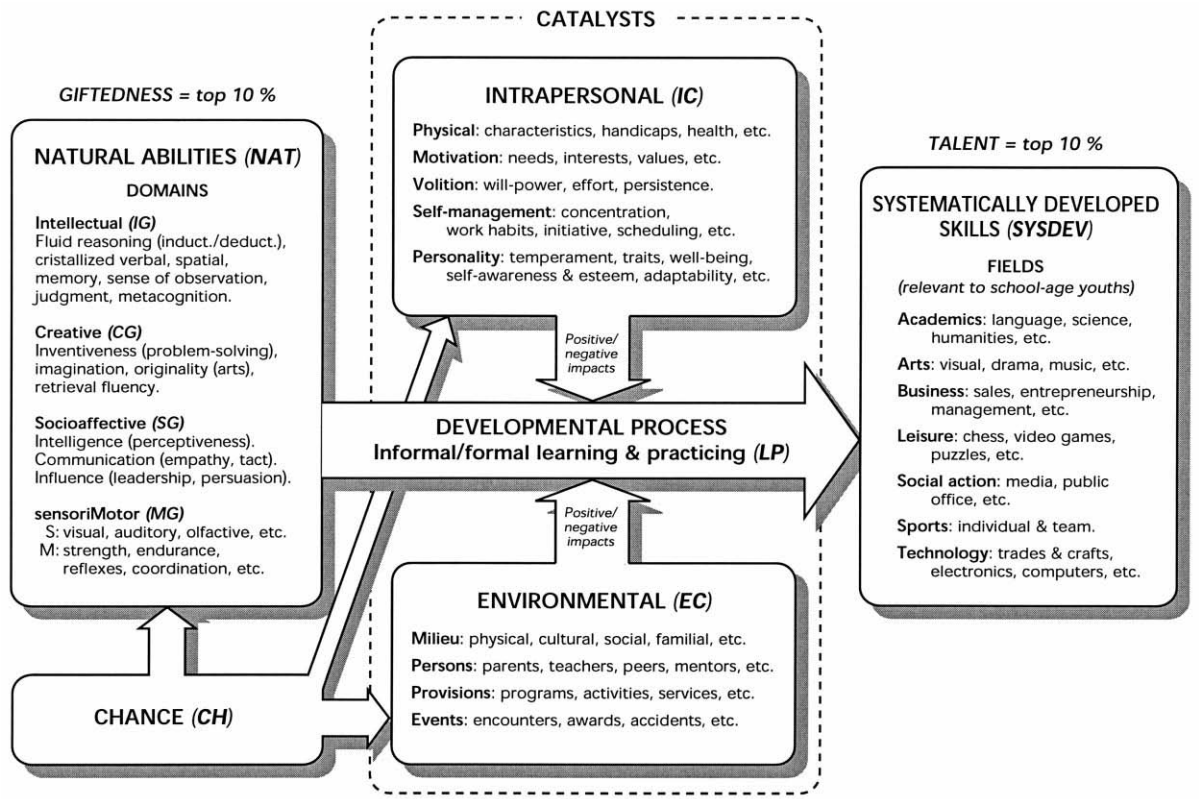

Figure 1
Gagné's differentiated model of giftedness and talent

into talents, for instance when intellectually gifted children achieve poorly in school.

The process of talent development manifests itself when the individual engages in systematic *learning and practicing* (courses, music conservatory, competitive team, etc.); the higher the level of talent sought, the more intensive these activities will be. This process can be facilitated or hindered by two types of *catalysts; intrapersonal* and *environmental*. Among the intrapersonal catalysts, motivation and volition (or persistence, see Corno and Kanfer 1993) play a crucial role in initiating the process of talent development, guiding it, and sustaining it through obstacles, boredom, and occasional failure. Genetic predispositions to behave in certain ways (temperament), as well as acquired personality traits and attitudes, also contribute significantly to the talent development process. The environment manifests its significant impact through four agents, either physical or social. *Chance* appears in the model as a fifth causal factor; it affects the others in different ways (e.g., the randomness of the genetic endowment for abilities or temperament, the chance of having good parents, accidents, etc.). In summary, talent emerges progressively from natural gifts thanks to a complex choreography between numerous causal influences.

1.2 How Many Are There?

Giftedness and talent are *normative* concepts. Such concepts define populations that are different from the norm (e.g., obesity, poverty). Good definitions of normative concepts must specify to what extent the target population differs from the norm and what it means in terms of their prevalence. In the DMGT (Gagné 1998), the threshold for both the giftedness and talent concepts is placed at the 90th percentile. That threshold applies to *each* giftedness domain or talent field; thus, the *total* percentage of gifted and talented persons probably exceeds 30 percent. This generous minimum threshold is counterbalanced by a system of levels of giftedness and talent. Within the top 10 percent of 'mildly' gifted or talented persons, the DMGT recognizes four progressively more selective subgroups, respectively labeled 'moderately' (top 1 percent), 'highly' (top 1:1,000), 'exceptionally' (top 1:10,000), and 'extremely' (top 1:100,000).

1.3 Focus

The field of gifted education traditionally has focused on one form of giftedness (intellectual) and one form of talent (academic). There is also some theoretical interest for creative giftedness, although identification procedures rarely include creativity measures. This article will respect that tradition and redefine the gifted and talented as *intellectually gifted and academically talented* (IGAT) students. Intellectual giftedness usually is measured with individual or group intelligence (IQ) tests, whereas academic talent is assessed through grades or standardized achievement tests.

2. Developmental Overview

Except when noted, statements in this section apply best to mildly and moderately IGAT children, who account for 99 percent of that population.

2.1 Personal and Social Development

Most gifted children manifest some form of precocity (e.g., speaking, reading, counting). The degree of precocity is moderately correlated with the level of giftedness. Within the realm of intrapersonal catalysts, the picture is frequently ambiguous. For instance, gifted students do not appear more intrinsically motivated to learn than their average peers (Shore et al. 1991). But there is a tendency for those who achieve well to manifest slightly more persistence. That slight superiority extends to most measures of personal adaptation, whether global ones, like self-administered personality tests, or more specific indicators, like scales for anxiety, depression, or suicidal ideas. IGAT students tend to manifest fewer behavior problems in school, and the incidence of drop outs and delinquents is lower (Neihart 1999). In the case of self-esteem, study results reveal no clear tendency; differences between studies could be associated with the type of educational environment.

Social adaptation refers to friendships, popularity, leadership, and so forth. Socially, gifted youths fare usually as well, sometimes slightly better, than their average peers. Their achievements will often provoke envy, but, most of the time, it is positive envy (admiration, emulation). Are exceptionally gifted individuals more at risk for social or personal adaptation because of the larger gap with their average peers? The scientific literature shows inconsistent results; some studies indicate an increase in the frequency of maladjusted individuals—who still remain a minority (at worst 20–25 percent)—others do not (Norman et al. 1999). Some of these problems could result from inadequate responses to their educational needs.

With regard to environmental catalysts, statistics show clearly that the parents of IGAT students have on average a higher socioeconomic status. According to Herrnstein's syllogism (Herrnstein and Murray 1994), both the partial genetic inheritance of intelligence and the moderate relationship between IQ and educational attainments could account in large part for that modest over-representation.

In summary, the research base tends to invalidate the 'nerdy' stereotypical image of gifted youths.

Moreover, it proclaims loudly that IGAT students differ much more among themselves than they do when compared, as a group, with average ability peers.

2.2 Educational and Life Achievements

Intellectual giftedness is synonymous with easier and faster school learning, which leads to academic talent for most of them as long as their progress is not impeded by unfavorable intrapersonal or environmental catalysts. The close relationship between IQ scores and academic achievement is probably the best documented causal relationship in education. IGAT students have much better chances of completing high school, and earning a university degree. As the level of talent increases from moderate to exceptional, the DMGT's catalysts and practice component play an increasingly important role, especially will-power (Simonton 1994). Nobel prizes, Olympic medals, or international careers in music require intense and prolonged efforts. But such high achievements are extremely rare. In fact, according to the results of Terman's famous longitudinal study of 1,400 moderately gifted students (Oden 1968), only a minority of IGAT youngsters ever achieve eminence. Other studies with shorter time span confirm that result. This is not surprising since the prevalence of eminent individuals is well below 0.1 percent of the total adult population.

3. Educational Overview

3.1 Educational Needs

Because of their high cognitive abilities, gifted children can easily understand more abstract concepts and more advanced reasoning processes; and they can do that at a much faster pace than average. It is not rare for highly IGAT students to outpace age peers academically by 3 to 5 years. For instance, the US 'talent searches' (Assouline and Lupkowski-Shoplik 1997) identify thousands of mathematically precocious students in grades 7 or 8 by administering them the *Scolastic Aptitude Test*, a test normally used in grade 12 for college admission. Unfortunately, compared with their prevalence in schools, a very small percentage of IGAT students enjoy a school environment that meets their educational needs. The school curriculum targets regular students, and US studies have shown that (a) teachers adjust the normal pace to students close to the 30th percentile, and (b) that the content of school manuals has decreased in density ('dumbing down') since the 1980s (Reis et al. 1993). The short-term consequence is recurring boredom due to lack of challenge. The long-term risks are partial loss of interest for school learning and underdeveloped emotional and strategic skills to overcome obstacles. In summary, IGAT students need a challenging school experience, both in pace and content. Responses recognized as most appropriate differ from the regular curriculum in terms of both content and format.

3.2 A Specific Curriculum: Enrichment

Curricular adaptations specific to IGAT students are grouped under the general concept of *enrichment*, which can take four major forms—the four Ds. The first, enrichment in *Density*, corresponds to a faster learning pace, allowing for more content per unit of time. It corresponds to Renzulli's 'curriculum compacting' (see Reis et al. 1993). Not only does it make the learning process more enjoyable, but it also frees time for other enrichment activities. Enrichment in *Diversity* allows IGAT students to explore subjects outside the regular curriculum of a given grade level (astronomy, genetics, museum visits, history, etc.). Enrichment in *Difficulty* introduces concepts and explanations well beyond the level planned for regular students. For instance, IGAT students will understand very rapidly the extension of the binary system to any form of non-decimal measurement system. Enrichment in *Depth* allows students to pursue a personal interest through a long-term research project. Most large-scale enrichment programs in schools combine the four Ds in varying proportions.

3.3 Appropriate Formats: Grouping and Acceleration

In theory, thanks to curricular individualization, teachers should be able to respond to the particular educational needs of every student in their heterogeneous classroom. In reality, it is rarely done. An important US survey of classroom practices (Archambault et al. 1993) revealed that little differentiation is made, except for occasional special activities when IGAT students complete their schoolwork early. That demonstrated lack of individualization confirms the imperative need for ability grouping, either part-time or full-time. Numerous studies have shown the academic and socioaffective benefits of full-time grouping (Kulik 1992), subject to one essential condition: offering these students a truly enriched curriculum. Full-time grouping is rarely used below high-school level, except in special schools. The most popular grouping format in elementary schools, called 'pull-out classes,' uses resource rooms where students gather, usually for one or two half-day periods a week. A specially trained teacher supervises enrichment activities, combining the four Ds in various ways to taylor activities to the students' interests.

The concept of school acceleration encompasses various administrative procedures that allow students either to begin earlier than normal or to progress faster through the grade levels. It includes, among others,

grade skipping, covering in one year the curriculum of two grade levels, early entrance to school, credit by examination, and radical acceleration (skipping three years or more). Because it is made possible by intensive enrichment in density, it should be considered a true form of enrichment. Acceleration is undoubtedly the most controversial of all format prototypes. The major fear is the personal and social maladjustment of younger students among older peers due to insufficient maturity. Yet, hundreds of evaluative studies have shown little factual support for that fear (Southern and Jones 1991). When school acceleration is carefully planned and implemented, with appropriate follow-up, the large majority of students adapt very well to their new academic and social environment. But, in spite of compelling scientific evidence for its effectiveness, parents and educators alike remain very reluctant to choose it.

4. Conclusion: The Politics of Gifted Education

Worldwide, IGAT students remain the most underserved of all special school populations. Compared with the millions of dollars spent by ministries of education for all forms of disabling conditions, gifted education receives almost no direct subsidies. Advocates for these children face many objections: fear of elitism, fear of increasing the gap between the highest and lowest achievers, the need to prioritize those at risk, feelings that intellectually gifted students will succeed anyway. The accusation of elitism is based on the already mentioned observation that IGAT students come disproportionately from affluent families; but the disproportion in percentages conceals the fact that, in terms of numbers, many more IGAT students belong to families with average or below average socioeconomic status. Moreover, why should fast learners be denied an appropriate learning environment because of their social origin? As stated in the UN Bill of Rights of children, all children have a right to an education that will foster the maximum development of their abilities and personality.

See also: Creativity Training; Education for Gifted and Talented Children; Educational Institutions and Society; Educational Sociology; Giftedness, Psychology of; Higher Education; Intellectual Functioning, Assessment of

Bibliography

Archambault Jr F X, Westberg K L, Brown S W, Hallmark B W, Emmons C L, Zhang W 1993 *Regular Classroom Practices with Gifted Students: Results of a National Survey of Classroom Teachers*. The National Research Center of the Gifted and Talented, Storrs, CT

Assouline S G, Lupkowski-Shoplik A E 1997 Talent searches: A model for the discovery and development of academic talent. In: Colangelo N, Davis G A (eds.) *Handbook of Gifted Education*, 2nd edn. Allyn and Bacon, Boston, pp. 170–9

Corno L, Kanfer R 1993 The role of volition in learning and performance. In: Darling-Hammond L (ed.) *Review of Research in Education*. American Educational Research Association, Washington, DC, pp. 301–41

Gagné F 1993 Constructs and models pertaining to exceptional human abilities. In: Heller K A, Mönks F J, Passow A H (eds.) *International Handbook of Research and Development of Giftedness and Talent*. Pergamon Press, Oxford, UK, pp. 69–87

Gagné F 1998 A proposal for subcategories within the gifted or talented populations. *Gifted Child Quarterly* **42**: 87–95

Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York

Kulik J A 1992 *An Analysis of the Research on Ability Grouping: Historical and Contemporary Perspectives*. The National Research Center of the Gifted and Talented, Storrs, CT

Neihart M 1999 The impact of giftedness on psychological well-being: What does the empirical literature say? *Roeper Review* **22**: 10–17

Norman A, Ramsay D, Martray C, Roberts J 1999 Relationship between levels of giftedness and psychosocial adjustment. *Roeper Review* **22**: 5–9

Oden M H 1968 The fulfilment of promise: 40-year follow-up of the Terman gifted group. *Genetic Psychology Monographs* **77**: 3–93

Reis S M, Westberg K L, Kulilowich J, Caillard F, Hébert T, Plucker J, Purcell J H, Rogers J B, Smist J M 1993 *Why not Let High Ability Students Start School in January? The Curriculum Compacting Study*. The National Research Center of the Gifted and Talented, Storrs, CT

Shore B M, Cornell D G, Robinson A, Ward V S 1991 *Recommended Practices in Gifted Education: A Critical Analysis*. Teachers College Press, New York

Simonton D K 1994 *Greatness: Who Makes History and Why?* Guilford Press, New York

Southern W T, Jones E D (eds.) 1991 *The Academic Acceleration of Gifted Children*. Teachers College Press, New York

Sternberg R J, Davidson J E (eds.) 1986 *Conceptions of Giftedness*. Cambridge University Press, New York.

F. Gagné

Giftedness, Psychology of

1. Historical Perspectives

The history of giftedness is over 3000 years old. Interest in giftedness is evident in several cultures, with examples to be found in ancient Greek (Athenian) and Judeo-Christian traditions as well as in the Chinese Ming Dynasty. In the twentieth century there were fluctuations in public as well as scientific interest in this theme (for greater detail see Tannenbaum 2000).

Although the German personality psychologist William Stern promoted the research and diagnosis (identification) of giftedness in a 1916 publication and the American intelligence researcher Lewis M. Terman started the most comprehensive to date ('Genetic Studies of Giftedness' in California in 1922–1923, for which the most recent follow-up research report was presented in 1995), intensified psychological interest in

giftedness only gathered true momentum in the last 30 years of the twentieth century. This was prompted, above all, by two events and their consequences: the Threat of Sputnik in the USA and the Cultural Revolution on the mainland of China. The influential conflict of the 1960s between egalitarianism (or 'democracy') and excellence was, in the 1970s and 1980s, increasingly overshadowed by the economic demand for the activation of human resources through gifted education. Under these conditions, the number of gifted programs grew globally in the last decades of the twentieth century. Accordingly, research questions at the beginning of the twenty-first century are dominated by applied research rather than the acquisition of knowledge in the sense of basic research. Even so, experimental and field studies in the paradigms of cognitive theory, lifespan, social psychology, learning, and instructional psychology have, in recent years, won growing significance (Sternberg and Davidson 1986, Colangelo and Davis 1997, Heller et al. 1993, 2000).

2. *Definition Problems and Newer Conceptions of Giftedness*

The terms giftedness and talent are used synonymously in everyday language and the scientific literature (Heller et al. 1993, 2000, Colangelo and Davis 1997). Because gifted or talented individuals are a very heterogeneous set of persons, a comprehensive theory is hardly possible. Furthermore, Ziegler and Heller (2000) identify a series of empirical, ontological, and meta-theoretical difficulties with conceptions of giftedness and talent.

> 'For example, several conceptions of giftedness incorporate genetic influences as a significant component of the basic meaning of talent. In fact, there is a large body of research that proves that genetic influences account for some interindividual differences. Wagner (1999) indicates … that these genetic influences have only been proven by untrained persons.' Ericsson drew the plausible conclusion that 'the estimated performance barriers seen as genetically fixed upper boundaries can be overcome through suitable learning processes' (Ziegler and Heller 2000, p. 3).

Eysenck and Barrett (1993) view giftedness as a 'fuzzy concept' that can be defined as synonymous with general intelligence, creativity, and special (e.g., artistic or scientific) ability. Gallagher and Courtwright (1986) point out two distinct meanings based on the psychological context (with a focus on individual differences) and educational practice (with a focus on individual needs and special educational programs). Sternberg and Davidson (1986) analyzed no less than 17 different conceptions, which they divided into explicit and implicit terms. An example of an implicit conception was delivered by Tannenbaum who viewed the interaction of five factors—general ability, special ability, nonintellective (motivational, etc.), environmental, and chance factors—as resulting in gifted/talented performance. Also Renzulli's giftedness model is counted by Sternberg and Davidson among the implicit theories, while Gagné's or the Munich psychometric model as well as Sternberg's cognitively-based triarchic model of intellectual giftedness are examples of explicit (empirical, experimentally confirmable) theories. A common basis shared by these newer conceptions is that giftedness is conceptualized as a differential construct. For greater detail see Ziegler and Heller (2000).

Contrary to the older IQ-based definitions (e.g., Terman defined giftedness by IQ-scores of 135 and higher), models of giftedness developed in the 1980s and 1990s represent multidimensional or typological conceptions. The following is a list of examples:

(a) The *Three Ring Model* by Renzulli (1986). According to this model, giftedness is a result of above average intelligence, creativity, and task commitment. This was expanded by Mönks to include family, school, and peer group, making a dynamic *Six Factor Model*; see Renzulli (1986), Mönks and Mason (2000).

(b) Gardner's *Theory of Multiple Intelligences*. Originally Gardner (1983), Gardner postulated seven distinct human intelligences: linguistic, logical–mathematical, spatial, bodily–kinesthetic, musical, intrapersonal, and interpersonal. Recently he has added 'naturalistic' (the ability to discern patterns in the living world) as an eighth intelligence and two further candidates (8 1/2): 'existential' and 'spiritual' intelligences, which have not yet been confirmed empirically (Ramos-Ford and Gardner 1997). This typological model synthesizes psychometric, cognitive psychological, and neuropsychological approaches to conceptualize *giftedness* (a more accurate label than 'intelligence').

(c) Gagné's *Differentiated Model of Giftedness and Talent*, the newest version of which was presented by the author (2000). Gagné differentiates between three concepts: genetically determined gifts (aptitude domains), talents (fields of exceptional performance), and catalysts (personality vs. environmental supportive factors). The gifts—intellectual, creative, socioaffective, sensorimotor, and other factors—are transformed, with the assistance of catalysts as well as learning and training processes, into domain-specific talents, e.g., in arts, sports, science, and technology.

(d) The *Munich Typological Model of Giftedness* by Heller et al. (see Heller and Feldhusen 1986, Heller 1992/2001). This multidimensional giftedness model consists of seven relatively independent ability factor groups (*predictors*), various performance domains (*criterion* variables), as well as personality (motivational, etc.) and social environmental factors which serve as moderators for the transition of individual potentials into excellent performances in various domains. High ability or giftedness is defined as an individual (intellectual, creative, etc.) potential for

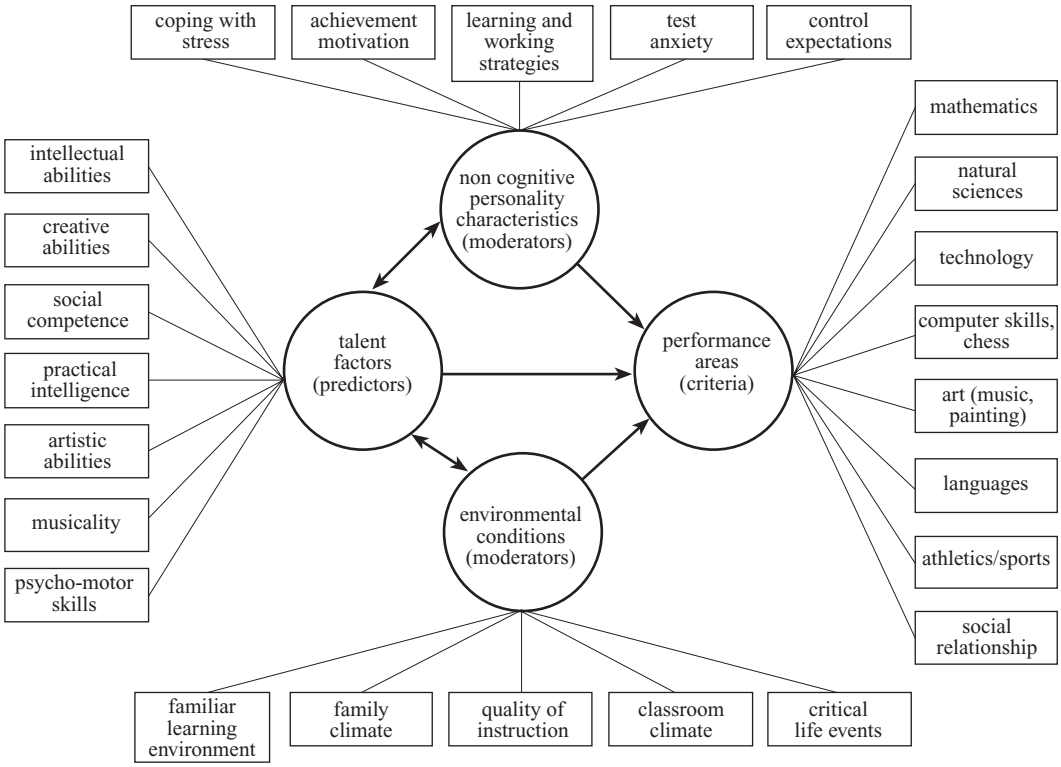

Legend:

talent factors (predictors)
- intelligence (language, mathematical, technical abilities, etc.)
- creativity (linguistic, mathematical, technical, artistic, etc.)
- social competence
- practical intelligence
- artistic abilities
- music ability
- psycho-motor skills

(non-cognitive) personality characteristics (moderators)
- achievement motivation
- hope for success vs. fear of failure
- control expectations
- thirst for knowledge
- ability to deal well with stress (coping with stress)
- self-concept (general, scholastic, of talent, etc.)

environmental conditions (moderators)
- home environmental stimulation ("creative" environment)
- educational style & parental educational level
- demands on scholastic performance made at home
- social reactions to success and failure
- number of siblings and sibling position
- family climate
- quality of instruction at school
- differentiated learning and instruction
- school climate
- critical life events

performance areas (criteria)
- mathematics, computer science, chess, etc.
- natural sciences
- technology, handicraft, etc.
- languages
- music and artistic areas
- social activities, leadership, trade, etc.
- athletics/sports

Figure 1
The Munich typological model of giftedness

outstanding achievements in one—not too narrowly defined—domain (one-sided talent) or two—seldom more—domains (multi-talent). Hence giftedness is conceptualized as a multifactorized ability construct within a network of noncognitive (motivational) and social moderators as well as performance-related variables. For diagnostic purposes (identification), the differentiation between predictor, criterion, and moderator variables (in Gagné's model, so-called 'catalysts') is of particular interest (see Fig. 1). The validity of this model of giftedness has been confirmed in several enlarged studies both in Germany and in international settings (cf. Heller 1992/2001, Perleth and Heller 1994, pp. 77–114). The Münchner Hochbegabungs-Testbatterie (MHBT) [Munich High Abilities Test] by Heller and Perleth (eds.) which measures the various model variables (predictors, moderators) has recently been published.

(e) The *Triarchic Theory of Intellectual Giftedness* developed by Sternberg in 1985 (see Sternberg 1986, 1997) represents a process-oriented conception. In cognitive psychological research paradigms, thinking and problem-solving *processes* stand in the limelight. The triarchic theory consists of three subtheories: a) the *componential-subtheory*, whereby *metacognitive* competencies regarding the planning, monitoring, and control of actions, *performance* components in the sense of encoding and the solution of problems, and *knowledge-acquisition* components are of major significance; b) the *experiential-* or *two-facet-subtheory* with the aspects of *automatization* or *subroutine* vs. insight or creative thinking in the solution of (new kinds of) problems; and c) the *contextual subtheory*, which refers to sociocultural dependence of talent definitions.

Further process-oriented conceptions of giftedness are discussed by Siegler and Kotowsky in their literature survey (1986). Their two levels of giftedness correspond to those delineated by Renzulli through his division into 'schoolhouse giftedness' (intelligence) and 'creative–productive giftedness' (creativity). The diagnostic implications of these have often been confirmed, in that (multifactorial) intelligence tests can predict scholastic success, although they cannot be (exclusively) used to predict academic or career success. On the other hand, creativity tests have proved to be unreliable predictors of scholastic success, although they can often explain significantly larger portions of the variance in career success than intelligence tests. This brings us to our next question regarding the advantages and disadvantages of various research approaches.

3. *Research Paradigms in the Field of Giftedness*

The field of giftedness research differentiates among several types of paradigms, whereby synthetic approaches have recently been accorded a large amount of attention. The psychometric paradigm is based on the trait and factor approach (e.g., Terman, Gagné, Gardner, Heller) already proven in the field of psychology. It focuses on the description of the phenomenology and structure of giftedness and talent (nomothetic function) and is indispensable for identification (diagnosis) of giftedness. Investigations based on the cognitive theory paradigm focus on the explanation of differences in the learning and thinking processes of highly gifted versus less gifted individuals (idiographic function). Such information about the mechanisms of cognitive processes and their individual causes provides not only new theoretical recognition but also useful contributions to the practice of gifted education and counseling. Hence both paradigms are needed, and they should be viewed in terms of how they complement one another. Furthermore, the expert–novice paradigm enriches cognitive theory in terms of the motivational effort variable 'deliberate practice' (cf. Ericsson 1996). While cognitive and metacognitive components are the focal point of cognitive componential approaches, in the expert–novice paradigm these are neglected in favor of the quantity and quality of learning and training processes. This approach has particularly proven its worth in the explanation of expertise in the domains of chess, physics, music, and athletics, but the extent to which this approach can be useful for explaining expertise in other domains must still be empirically confirmed. Even more interesting are recent attempts to combine prospective (psychometric) and retrospective (process-oriented) approaches in the form of the expert–novice paradigm. The intention here is to avoid the weaknesses inherent in the trait-oriented psychometric approach (underestimation of the influences of long-term deliberate practice) in so-called 'testing the limits' conditions (see Howe et al. 1998) and of the incremental (implicit) ability theory (according to Dweck) as well as those of the experience-oriented approach in the expert–novice paradigm (underestimation of innate talents in favor of learning processes); for recent contributions to the genetic influence vs. environmentalism controversy see Howe et al. (1998) including the 30 open peer commentaries. In the genetic and brain research-related approach the focus is on a biological explanation of giftedness and talent (cf. Thompson and Plomin, Eysenck and Barrett 2000).

Although the research paradigms sketched out here conceptualize giftedness more or less as individual resources, the following approaches concentrate on social and learning environment-related conditions. The *interactionist* theories try to balance the position between innate capacities and socialization influences with respect to the development of giftedness and excellence in various performance domains. Here lifespan developmental psychology is interested in the description and explanation of the individual's change of behavior throughout the life course (cf. Mönks and Mason 2000). *Psychosocial* and *cross-cultural*

approaches primarily analyze the external context conditions of the development of giftedness and outstanding achievements (cf. Tannenbaum 1983). From cross-national, cross-cultural, and cross-societal research approaches one can expect not only information about various social-cultural influences on the individual developmental and gifted education processes within specific social settings (-*etic* approach), but also explanations regarding the universality of gifted behavior characteristics (universalism hypothesis in the -*emic* approach). The latter function is, above all, indispensable for examining universality assumptions of many theories—here, in the field of giftedness.

4. *Identifying and Programming (Gifted Education and Counseling)*

Perspectives related to the application of results of research in giftedness include the identification of the gifted and the nurturing and counseling of gifts or talents. Both aim to optimize the development of personality in individual and social developmental conditions. To this end, diagnostic information is necessary, i.e., about characteristics of the individuals and their learning environments. Provided that one can conceptualize the development of giftedness as an interactive product between internal (cognitive and motivational) factors and external socialization conditions, multidimensional tests for the identification of the gifted are necessary. Correspondingly, a comprehensive diagnosis of giftedness cannot be defined by one type (for example within the framework of talent searches for specific promotional courses, scholastic programs, contests, etc.), but rather have to keep individual differences in moderation styles in mind (cf. Fig. 1). It is these moderating variables which frequently offer opportunities to provide effective counseling or intervention, for example underachievers or other at-risk groups. Here, 'at-risk groups' refer to highly gifted and talented girls, gifted handicapped, gifted underachievers, or disadvantaged and culturally different talented, because for these groups a higher risk not to be identified as gifted exists (cf. Yewchuk and Lupart 2000). In order to expand on status diagnostic methods, approaches involving *dynamic assessment* (cf. Kanevsky 2000) are available to assist in the identification of gifted underachievers. Since most norm-oriented ability or intelligence tests, including creativity tests, have proven to suffer from *ceiling effects* in application, they are inadequate for identification, particularly on the basis of IQ-scores. More suitable are multidimensional cognitive abilities tests, such as the Munich High Abilities Test (MHBT) by Heller and Perleth (cf. Heller 1992/2001), developed along the lines of the Munich Giftedness Model (Fig. 1). Here, differential talent predictors with respect to various aspects of intelligence, creativity, musicality, psychomotorics, social and practical intelligence as well as relevant noncognitive (motivational) and social moderators are measured in order to determine outstanding performance (criterion) in various domains.

In order to reduce the bandwidth fidelity dilemma, or so-called alpha and beta errors as defined by Cronbach and Gleser, a *sequential decision strategy* can be used. Initially a *screening* (e.g., with teachers' checklists) comes up with a rough estimation for the reduction of the beta error ('oversights' of gifts and talents), after which in the second and possibly third steps, domain-specific (multidimensional) abilities *tests* etc. are probed in order to reduce the alpha error (in order to avoid false identifications). For more detailed explanations see Heller and Feldhusen (1986), Mönks and Heller (1994), Colangelo and Davis (1997, pt. II) or Heller et al. (1993, 2000, pt. III).

Gifted education strives for both the promotion of effective, creative learning environments versus a fit between stimulating family and scholastic settings and the individual needs for learning, e.g., striving for knowledge, exploratory urges, challenges in the sense of testing one's limits, etc. The Aptitude Treatment Interaction (ATI) Model provides the theoretical basis for these constructs. Quick information processing and accelerated cognitive and metacognitive competencies afford these children and adolescents effective self-regulated, discovery learning without a high risk probability for the acquisition of misconcepts (as among low gifted). In order to avoid understimulation and boredom among highly gifted and talented students, multifaceted scholastic differentiation measures are necessary, whereby two prototypes can be delineated: enrichment and acceleration, which can also be effectively combined. For more information regarding various instructional models and programs of nurturing the gifted and talented see Colangelo and Davis (1997, pts. III and IV), Heller et al. (1993, 2000, pts. IV and VI).

Finally, *gifted counseling* is often needed for enhancing the socialization processes in gifted education, especially with respect of the at-risk groups mentioned above. Opportunities for counseling regarding themes relevant to the highly talented differ with respect to quality and quantity, particularly with respect to gender and age of the clientele, but also regarding family and scholastic socialization as well as regional counseling offerings. Furthermore, relationships to socioeconomic family status have been confirmed. For more detailed information see Colangelo and Davis (1997, pt. V), Heller et al. (1993, 2000, pt. V) or Heller and Feldhusen (1986, Chaps. XI and XII). Counseling the gifted and talented should be an integral part of gifted education.

See also: Creativity and Cognition; Education for Gifted and Talented Children; Genetic Factors in Cognition/Intelligence; Gifted and Talented Individuals: Developmental and Educational Overview; Intelligence, Genetics of: Cognitive Abilities; Intel-

ligence, Genetics of: Heritability and Causation; Intelligence: Historical and Conceptual Perspectives; Intelligence, Prior Knowledge, and Learning

Bibliography

Colangelo N, Davis G A (eds.) 1997 *Handbook of Gifted Education*, 2nd edn. Allyn and Bacon, Boston

Ericsson K A (ed.) 1996 *The Road to Excellence—the Aquisition of Expert Performance in the Arts and Science, Sports and Games*. Erlbaum, Mahwah, NJ

Eysenck H, Barrett P T 1993 Brain research related to giftedness. In: Heller K A, Mönks F J, Passow A H (eds.) 1993 *International Handbook of (Research and Development of) Giftedness and Talent*. Pergamon, Oxford, UK, pp. 115–31

Gagńe F 2000 Understanding the complex choreography of talent development through DMGT-based analysis. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*. Elsevier, Amsterdam, pp. 67–79

Gallagher J J, Courtwright R D 1986 The educational definition of giftedness and its policy implications. In: Sternberg R J, Davidson J E (eds.) 1986 *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK, pp. 93–111

Gardner H 1983 *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books, New York

Heller K A (ed.) 1992 *Hochbegabung im Kindes- und Jugendalter* [*High Ability in Childhood and Adolescence*]. Hogrefe, Göttingen, Germany

Heller K A, Feldhusen J F (eds.) 1986 *Identifying and Nurturing the Gifted. An International Perspective*. Huber, Toronto, Canada

Heller K A, Mönks F J, Passow A H (eds.) 1993 *International Handbook of (Research and Development of) Giftedness and Talent*. Pergamon, Oxford, UK

Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) 2000 *International Handbook of (Research and Development of) Giftedness and Talent*, rev. edn. Elsevier, Amsterdam

Howe M J A, Davidson J W, Sloboda J A 1998 Innate talents: Reality or myth? *Behavioral Brain Sciences* **21**(3): 399–442

Kanevsky L 2000 Dynamic assessment of gifted students. In: Heller K A, Mönks F J, Stenberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*. Elsevier, Amsterdam, pp. 283–95

Mönks F J, Heller K A 1994 Identification and programming of the gifted and talented. In: Husén T, Postlethwaite T N (eds.) *The International Encyclopedia of Education*. Pergamon, Oxford, UK, Vol. 5

Mönks F J, Mason E J 2000 Developmental psychology and giftedness: theories and research. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of (Research and Development of) Giftedness and Talent*, rev. edn. Elsevier, Amsterdam, pp. 141–55

Perleth C H, Heller K A 1994 The Munich longitudinal study of giftedness. In: Subotnik R F, Arnold K D (eds.) 1994 *Beyond Terman: Contemporary Longitudinal Studies of Giftedness and Talent*. Ablex, Norwood, NJ, pp. 77–114

Ramos-Ford V, Gardner H 1997 Giftedness from a multiple intelligences perspective. In: Colangelo N, Davis G A (eds.) *Handbook of Gifted Education*, 2nd edn. Allyn and Bacon, Boston, pp. 54–66

Renzulli J S 1986 The three-ring conception of giftedness: a developmental model for creative productivity. In: Sternberg R J, Davidson J E (eds.) *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK, pp. 53–92

Siegler R S, Kotowsky K 1986 Two levels of giftedness: shall ever the twain meet? In: Sternberg R J, Davidson J E (eds.) *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK, pp. 417–35

Sternberg R J 1986 A triarchic theory of intellectual giftedness. In: Sternberg R J, Davidson J E (eds.) *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK, pp. 223–43

Sternberg R J 1997 In: Colangelo N, Davis G A (eds.) *Handbook of Gifted Education*, 2nd edn. Allyn and Bacon, Boston, pp. 43–53

Sternberg R J, Davidson J E (eds.) 1986 *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK

Subotnik R F, Arnold K D (eds.) 1994 *Beyond Terman: Contemporary Longitudinal Studies of Giftedness and Talent*. Ablex, Norwood, NJ

Tannenbaum A J 1983 *Gifted Children: Psychological and Educational Perspectives*. Macmillan, New York

Tannenbaum A J 2000 A history of giftedness in school and society. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) 2000 *International Handbook of (Research and Development of) Giftedness and Talent*, rev. edn. Elsevier, Amsterdam, pp. 23–53

Thompson L A, Plomin R 2000 Genetic tools for exploring individual differences in intelligence. In: Heller K A, Mönks F H, Sternberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*. Elsevier, Amsterdam, pp. 157–64

Yewchuk C, Lupart J 2000 Inclusive education for gifted students with disabilities. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*. Elsevier, Amsterdam, pp. 659–70

Ziegler A, Heller K A 2000 Conceptions of giftedness from a meta-theoretical perspective. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*, rev. edn. Elsevier, Amsterdam, pp. 3–21

K. A. Heller

Global Environmental Change: Human Dimensions

1. Introduction

Global environmental change is as old as planet Earth. Physical, chemical, and biological processes have been shaping and reshaping the earth's environment since its infancy 4.5 billion years ago. In recent time, however, humankind has been one of the major driving forces of change on the planet, including climate change, stratospheric ozone depletion, deforestation, loss of biodiversity, acidification, pollution, and desertification (see, for example, Turner et al. 1990).

Research on the human dimensions of global environmental change is concerned with the human causes of change, the consequences of such changes

for individuals and societal groups, and the ways in which humans respond to the changes.

The causes include processes like industrialization and urbanization, which themselves are driven by human needs and wants, such as the needs for water, food, and shelter. The impacts include water and food shortages, floods and droughts, health risks, and conflicts about resources. Responses to these changes include adaptation, for example, through planting different crops, and even migration from areas most affected by environmental changes and mitigation, for example, by reducing emissions of gases into the atmosphere.

Individuals and teams carry out research on the human dimensions of global environmental change in many universities and research institutes worldwide.

The range of social science disciplines involved in research related to global environmental change is wide, including political scientists studying the effectiveness of environmental agreements, demographers looking at the links between population growth and environmental change, economists considering the possibilities for delinking economic growth and environmental impacts, anthropologists debating human needs and wants and global environmental change, and scholars looking at the psychological dimensions of global environmental change. In addition, the International Human Dimensions Programme on Global Environmental Change (IHDP) coordinates research at the international level (IGBP/IHDP 1999, IHDP 1999a, 1999b, 1999c).

2. *The Causes of Global Environmental Change*

As pointed out in the introduction, global environmental change has been taking place throughout the earth's history, our focus here is on anthropogenic causes. Changes in the global atmosphere (climate change, stratospheric ozone depletion, and acid rain) are a result of emissions of a number of gases into the atmosphere due to a range of human activities. The increasing greenhouse effect, leading to a warming of the earth's surface and lower atmosphere, is the result of the accumulation of gases including carbon dioxide, methane, nitrous oxide, and chlorofluorocarbons in the atmosphere. The Intergovernmental Panel on Climate Change (IPCC) has documented these changes. Chlorofluorocarbons are also responsible for stratospheric ozone depletion (WMO 1999). Transboundary acidic pollution is attributed to emissions of sulfur dioxide, nitrogen oxide, and other acidifying emissions (e.g., Soroos 1998).

What are the main human activities that lead to these emissions? In the case of climate change the burning of coal, oil, and natural gas, as well as deforestation and various agricultural and industrial practices are altering the composition of the atmosphere and contributing to climate change. It has been estimated that the use of fossil fuels currently accounts for 80–85 percent of the carbon dioxide being added to the atmosphere (UNEP/WMO 1997). Fossil fuels are burned to produce energy used for transportation, manufacturing, heating, cooling, electricity generation, and other applications. Research shows that patterns and environmental impacts of fossil fuel use vary widely by country and national-level consumption varies with technology, population and other factors, such as the level of industrialization. Land use changes, e.g., clearing land for logging, ranching, and agriculture as well as for urban development, also lead to carbon dioxide emissions (e.g., Turner et al. 1990). Over the past several hundred years, deforestation and other land use changes in many countries contributed substantially to atmospheric carbon dioxide increases. Most of the net carbon dioxide emissions from deforestation are currently occurring in tropical regions. The human activities that lead to methane emissions include rice cultivation, cattle and sheep ranching, and creation of landfills. Methane is also emitted during coal mining and oil drilling and by leaky gas pipelines. It is estimated that human activities have increased the concentration of methane in the atmosphere by about 145 percent above what would be present naturally (UNEP/WMO 1997).

Chlorofluorocarbons (CFCs) have been used in aerosol spray cans, refrigeration, air conditioning, and solvents. However, the production of CFCs is being eliminated as a result of international agreements, in particular the Montreal Protocol signed in 1987 and subsequent amendments agreed to in London (1990) and Copenhagen (1992). Soroos (1998) shows the sharp drop in the production and use of CFCs and other ozone-depleting substances in the 1990s and concludes that the level of stratospheric ozone is likely to reach its lowest level around 2000 and then begin a gradual rise back to its natural level.

The accumulation of gases in the atmosphere that can give rise to climatic change, stratospheric ozone depletion, and acidification is increasingly the result of a wide range of human activities like energy production, agriculture, and transportation. These activities in turn result from the satisfaction of human needs and wants. The topic of 'needs and wants' was tackled in detail from a social science perspective by Douglas et al. (1998).

An ongoing debate with regard to the causes of global environmental change is whether population growth in less developed countries is the main cause of environmental change or whether the main factor is the high level of consumption in industrialized countries. As MacKellar et al. (1998) point out, the population-consumption debate has become the focus of the broader debate about equity and fairness, demonstrated, for example, at the UNCED in Rio in 1992. Some case studies have shown that population growth can be correlated with increasing rates of tropical deforestation, while others have shown that

population growth need not lead to increasing deforestation, if alternatives to land-clearing are provided.

Demographic research does point to some important trends that will play a role in determining future global environmental change: there will be further population growth, a further tilt towards less industrialized countries in the world distribution of population, and further population aging (MacKellar et al. 1998). Another important trend is urbanization. At the beginning of the twenty-first century, more people will live and work in the urban centers of the world than in rural areas for the first time in history (Board on Sustainable Development 1999). The 'rush to the cities' of rural populations in developing countries during the second half of the twentieth century was dramatic. The largest cities of the poorer countries are growing much faster than similar cities in the industrialized world.

In addition to the changes occurring in the global atmosphere, global environmental changes are occurring at the earth's surface. In recent years the question of biodiversity loss has been on scientific and political agendas. As defined in the UN Framework Convention on Biological Diversity, biodiversity is the diversity of species, genetic material and ecosystems or, more generally, the variability among living organisms. Human activities contribute to the loss of biodiversity in many ways. Studies of the human causes of biodiversity loss have to deal with issues such as cultural beliefs, religious teachings, social history, and property rights. In fact, of course, it is not only the issue of biodiversity that encompasses these factors. These forces also influence many other issues related to global environmental change. There is a large body of environmental ethics literature in particular that deals with such issues (see, for example, Gardner and Stern 1996).

A related issue for global environmental change is that of land use and land cover change. In addition to playing a role in biodiversity loss, land use and land cover change can affect the global climate system through emissions of greenhouse gases (carbon dioxide, methane, and nitrous oxide) as well as changes in climatic variables, such as the reflectivity of the earth's surface. Over the course of human history, land use change has contributed as much to increases of the atmospheric carbon dioxide concentration as fossil fuel combustion has (Meyer et al. 1998). The human activities that can change land cover include sedentary agriculture, shifting cultivation, navigation, water supply, forest plantations, settlements, transportation corridors, mining, and habitat and scenery preservation (Meyer et al. 1998). The study of Kasperson et al. (1995) identified world regions where misuse of land and water resources seriously threatens human wealth and well being in the near to medium term. There has also been considerable progress in recent years on questions such as the social causes of deforestation in areas such as the Amazon River Basin and Southeast Asia and the role of social, political, and economic institutions in land-use decisions (Entwistle et al. 1998).

Recently, advances have been made in combining social science and remote sensing perspectives, techniques, and data (US National Research Council 1998). Studies have shown how remotely sensed and social data can be combined to understand human–environment linkages, especially with regard to land use and land cover changes on the regional level. The combination of social data and remotely sensed data can lead to improved understanding of the consequences of climate variability, including improving famine early warning systems and dealing with serious human health issues.

3. *The Human Consequences of Global Environmental Change*

Much of the research on the consequences of global environmental changes has focused on the impacts of climatic change, although there have been some studies of the impacts of stratospheric ozone depletion and acid rain.

In recent years assessments of the impacts of climate change have been carried out under the auspices of the UNEP/WMO Intergovernmental Panel on Climate Change, which produced reports in 1990 and 1996 (IPCC 1990, 1996). Given uncertainties in both the expected climatic changes (especially when considering the regional or local scale) in the twenty-first century and the calculations of socioeconomic impacts of such changes, the studies of the impacts of climate change do not produce predictions but generally generate plausible scenarios of future impacts.

The IPCC scientific assessment (UNEP/WMO 1997) estimated that the globally averaged surface temperature will increase by 1–3.5°C by the year 2100, with an associated rise of sea level of 15–95 cm. One of the topics that received wide attention after the IPCC report was the potential impact on human health. The report concluded that the projected increase in the duration and frequency of heat waves would be expected to increase mortality rates as a result of heat stress, especially where air conditioning is not available. McMichael et al. (1996) have discussed the general impacts of climate change on human health.

Many studies have looked at the potential impacts of climate change on agricultural production. The 1996 IPCC assessment concluded that it may be possible for global agricultural production to keep pace with increasing demand over the next 50–100 years if adequate adaptations are made, but that there are likely to be difficulties in some regions. The report concluded that there is likely to be an increased risk of famine, particularly in subtropical and tropical semi-arid and arid locations.

With more than half of the global human population currently living in coastal areas, future sea level rises, alterations in storm patterns, and higher storm surges could have significant effects. About 46 million people are currently at risk by flooding in coastal areas as a result of storm surges. In the absence of measures to adapt, it is estimated that even with current populations a sea level rise of 50 cm would increase the number of people whose land would be at risk from serious flooding or permanent inundation to about 92 million, while a 100 cm rise would increase this number to 118 million.

As the IPCC report also shows, climate change is likely to affect numerous other human activities and resources valued by humans such as river flow and water supply, transportation systems, energy demand and supply, human settlement patterns, and tourism. These impacts are the topics of research. For example, a range of studies has looked at impacts on energy supply and demand from an economic perspective and these studies are reviewed by the IPCC. Multisectoral regional assessments of the consequences of climate change have also been carried out).

The methods for assessing climate impacts have developed considerably since around 1970. As Parry and Carter (1998) point out, the efforts up until the mid-1970s focused mainly on the (one-way) 'impact' of climate on human activity, while more recently there has been more emphasis on the 'interaction' between climate and human activity by assuming that a climatic event is merely one of many processes (both societal and environmental) that can affect humans and their activities. The interaction approach also introduced adaptation into climate impact assessment, such as changes in crops and irrigation at the farm level and policy responses at the subnational, national, and international levels (Parry and Carter 1998).

Until recently, most scientific assessments of global environmental risks devoted little attention to the possibly endangered entities. Now, questions about the 'vulnerability' of social and ecological systems are emerging as a central focus of policy-driven assessments of global environmental risks in a number of arenas.

4. *Human Responses to Global Environmental Change*

There are basically three possible responses to global environmental change. First, humans could decide to 'do nothing,' because they don't believe scientific results, because they think that the changes will be gradual and easily accommodated through technology, or simply because they are convinced that it would cost more to do something than the benefits the environmental changes might bring. In doing nothing, they might find that they are forced to undertake responses, such as migration from flood-prone areas, in the event of global environmental changes (forced adaptation). Second, humans can choose to 'mitigate,' for example, reduce net emissions of greenhouse gases, acidifying substances or ozone-depleting substances into the atmosphere or reduce the rate of deforestation or land-cover change. Third, humans can choose to 'adapt' to global environmental change (anticipatory adaptation).

Humans have shown great capacity to adapt to environmental changes and extreme events in the past and the assessment of the research of past research on adaptation may provide useful insights for the future. However, adaptation in a globalized world with a population of 10 billion people promises to be much more challenging than adaptation in the past.

At the international level, steps to mitigate global environmental change have been taken in recent years. The Convention on Long-Range Transboundary Air Pollution (LRTAP) and its associated protocols have been dealing with emissions reductions of acidifying substances. The Vienna Convention on the Protection of the Stratospheric Ozone Layer and the Montreal Protocol on Substances that Deplete the Ozone Layer, signed in 1987 has led to the phase-out of production of most ozone-depleting substances. At the Rio Conference (UNCED) in 1992 the Framework Convention on Climate Change and the Convention on Biodiversity were signed and have subsequently gone into force with subsequent negotiations on protocols. Clearly, it is not enough to negotiate and sign international agreements to deal with the issues of global environmental change (Underdal 1998, Weiss and Jacobson 1998). Research on policy instruments at the national and international levels of relevance to the implementation of international environmental agreements shows the strengths and limitations of strategies such as regulation, financial incentives and penalties, inducements for technical development, etc. (US National Research Council 1999). Research at the end of the twentieth century and the beginning of the twenty-first century is also showing how social institutions have succeeded and failed in the management of global environmental risks (Ostrom 1990, Young 1999).

Methods for evaluating adaptive responses are discussed by Parry and Carter (1998), who point out that many of the adaptation options are likely to be subject to legislation, influenced by prevailing social norms related to religion or custom, or constrained physically or biologically. This might encourage, restrict, or totally prohibit their implementation. However, it is clear that an assessment of possible constraints must accompany any evaluation of available options. To date, relatively little analysis has been made of the efficacy of different adaptive strategies, partly because of the very wide range of available options.

On balance, however, it seems that much more attention has been paid to the assessment of mitigation

options than to adaptation options. Serious study is still needed of the true costs of adaptation of both plants and people and the differential ability to undertake it.

A controversial topic related to the question of adaptation is the relation between global environmental change and migration. The IPCC report in 1990 suggested that the greatest effect of climate change may be on human migration, as millions of people could be displaced due to shoreline erosion, coastal flooding, and agricultural disruption. Lonergan (1998) attempted to clarify the numerous issues surrounding the linkage between environmental degradation and population displacement. He concluded that generalizations about the relationship between environmental degradation and population movement mask a great deal of the complexity, which characterizes migration decision-making. Environmental degradation and resource depletion may play a contributing role in affecting population movement, often filtered through contexts of poverty and inequity.

5. *Attitudes, Behavior, and Lifestyle*

While there is ample evidence that public concern about ecological problems and technological risks has increased in Western industrialized societies, as well as some other countries, since around 1980, at the same time there has been a continuing deterioration of the environment on a global scale with some dramatic regional consequences (Brand 1997). Data on the development of environmental consciousness have been and are collected in many countries, although they are difficult to compare especially because the results depend on the wording of questions and the cultural contexts within which surveys are made. As Brand (1997) documents, taken together the empirical findings point to a high, and globally growing, concern about the environment. However, a high degree of environmental consciousness does not necessarily mean that there is environmentally sound behavior. Unfortunately there are fewer comparative data on environmental behavior, and different national contexts make it difficult to make intercomparisons. Empirical findings show that the effects of environmental knowledge and consciousness on behavior are insignificant and that other factors have to be taken into account, i.e., psychological, economic, institutional, and contextual factors.

6. *Integrated Assessment*

An important contribution to research on the human dimensions of global environmental change has been provided by 'integrated assessment.' According to Parson (1995):

> Assessment consists of gathering, synthesizing, interpreting, and communicating knowledge from various expert domains and disciplines, to help responsible policy actors think about problems or evaluate possible actions.

An 'integrated' assessment means that the information is assembled from a broader set of domains than would usually be provided by good research from a single discipline. Integrated assessment of global environmental change issues therefore requires knowledge from all relevant disciplines, natural and social sciences, and the assessment is building a bridge between the science and policy realms. Summaries of available results, for instance in the reports of the IPCC, have illustrated the strengths and weaknesses of available approaches. At the end of the twentieth century methodologies have been developed for 'participatory integrated assessment,' in which stakeholders and even lay citizens are involved in the assessment process (Kasemir et al. 1999).

7. *Conclusion*

There is no doubt that global environmental changes have occurred and are occurring. Social sciences are playing an increasingly important role in improving understanding of the human causes and impacts of, and responses to, such changes. In a recent report, the US National Research Council (1999) pointed to the following areas in which major advances in knowledge were achieved through human dimension research:

(a) The nature and role of socioeconomic uncertainties in impact assessments.

(b) The determinants of environmentally significant consumption.

(c) The importance of vulnerability analysis in understanding the consequences of global environmental change.

(d) The role of institutions in determining human–environment interactions.

(e) Estimation of the costs of global environmental change and of policy response options.

The challenges remain large, especially in the area of obtaining, archiving, and using socioeconomic data and in conducting research on a range of unresolved and difficult questions that can only be tackled by international, interdisciplinary approaches.

See also: Air Pollution; Climate Impacts; Consumption, Environmentally Significant; Ecological Economics; Environmental Cognition, Perception, and Attitudes; Environmental Policy: Protection and Regulation; Environmental Security; Environmentalism: Preservation and Conservation; Industrial Ecology; International and Transboundary Accords, Environmental; Land Use and Cover Change; Water Pollution: Marine; Water Pollution: Terrestrial and Groundwater

Bibliography

Board on Sustainable Development 1999 *Our Common Journey: A Transition Toward Sustainability*. National Academy Press, Washington, DC

Brand K W 1997 Environmental consciousness and behaviour: The greening of lifestyles. In: Redclift M, Woodgate G (eds.) *The International Handbook of Environmental Sociology*. Edward Elgar, Cheltenham, UK

Douglas M, Gasper D, Ney S, Thompson M 1998 Human needs and wants. In: Rayner S, Malone E L (eds.) *Human Choice and Climate Change*. Battelle Press, Columbus, OH

Entwistle B, Rindfuss R R, Walsh S J 1998 Land use/land cover and population dynamics, Nang Rong, Thailand. In: Liverman D M, Moran E F, Ruidfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC

Gardner G T, Stern P C 1996 *Environmental Problems and Human Behaviour*. Allyn and Bacon, Boston

IGBP/IHDP 1999 *Land-Use and Land-Cover Change: Implementation Strategy*. IHDP Report No 10. International Human Dimensions Programme on Global Environmental Change, Bonn, Germany

IHDP 1999a *Institutional Dimensions of Global Environmental Change: Science Plan*. IHDP Report No. 9. International Human Dimensions Programme on Global Environmental Change, Bonn, Germany

IHDP 1999b *Global Environmental Change and Human Secutiry: Science Plan*. IHDP Report No. 11. International Human Dimensions Programme on Global Environmental Change, Bonn, Germany

IHDP 1999c *Industrial Transformation: Science Plan*. IHDP Report No. 12. International Human Dimensions Programme on Global Environmental Change, Bonn, Germany

Intergovernmental Panel on Climate Change (IPCC) 1990 *Climate Change: The IPCC Scientific Assessment*. Cambridge University Press, Cambridge, UK

IPCC 1996 *Second Assessment Report of the IPCC*. Cambridge University Press, Cambridge, UK, 3 Vols.

Kasemir B, Schibli D, Stoll S, Jaeger C C 1999 Involving the public in climate and energy decisions. *Environment* **42**(3): 32–42

Kasperson J X, Kasperson R E, Turner II B L (eds.) 1995 *Regions at Risk: International Comparisons of Threatened Environments*. UNU Press, Tokyo

Lonergan S 1998 *The Role of Environmental Degradation in Population Displacement*. Global Environmental Change and Human Security Project Research Report 1. IHDP, Bonn, Germany

MacKellar F L, Lutz W, McMichael A J, Suhrke A 1998 Population and climate change. In: Rayner S, Malone E L (eds.) *Human Choice and Climate Change*. Battelle Press, Columbus, OH

McMichael A J, Haines A, Sloof R, Kovats S 1996 *Climate Change and Human Health*. World Health Organisation, Geneva, Switzerland

Meyer W B, Adger W N, Brown K, Graetz D, Gleick P, Richards J F, Maghalaes A 1998 Land and water use. In: Rayner S, Malone E L (eds.) *Human Choice and Climate Change*. Battelle Press, Columbus, OH

Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, New York

Parry M, Carter T 1998 *Climate Impact and Adaptation Assessment: A Guide to the IPCC Approach*. Earthscan, London, UK

Parson E A 1995 Integrated assessment and environmental policy making: In pursuit of usefulness. *Energy Policy* **23**(4/5): 463–75

Soroos M S 1998 The thin blue line: Preserving the atmosphere as a global commons. *Environment* **40**(2): 6–35

Turner B L, Clark W C, Kates R W, Richards J F, Mathews J T, Meyer W B 1990 *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere over the past 300 years*. Cambridge University Press, Cambridge, UK

Underdal A (ed.) 1998 *The Politics of International Environmental Management*. Kluwer Dordrecht, The Netherlands

UNEP/WMO 1997 *Common Questions About Climate Change*. United Nations Environment Programme, Nairobi, Kenya

US National Research Council 1998 Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC

US National Research Council 1999 *Global Environmental Change: Research Pathways for the Next Decade*. National Academy Press, Washington, DC

Weiss E B, Jacobson H (eds.) 1998 *Engaging Countries: Strengthening Compliance with International Environmental Accords*. MIT Press, Cambridge, MA

WMO 1999 *Scientific Assessment of Ozone Depletion: 1998*. WMO Global Ozone Research and Monitoring Project, Report No. 44. World Meteorological Organization, Geneva, Switzerland

Young O 1999 *The Effectiveness of International Environmental Regimes: Causal Connections and Behavioural Mechanisms*. MIT Press, Cambridge, MA

J. Jäger

Global Governance

Global governance is the combination of intentional and patterned human interactions that regulate action worldwide for the common good. It is a purposeful order that emerges from institutions, processes, norms, formal agreements, and informal mechanisms that provide a field of action for human activities. Governance is based on shared expectations, as well as on intentionally designed institutions and mechanisms (Rosenau and Czempiel 1992).

In the context of national government, the authority of public institutions lies ultimately in the government's monopoly of the legitimate use of force. In the global context, no such authority exists; there is no world governmental body that enjoys a monopoly of the legitimate use of force. Yet commerce among countries and firms expands, information is communicated across national boundaries in an orderly fashion, and expectations about rule implementation are shared through transnational networks. Cooperative action is based on rights and rules that are enforced through a combination of financial and

moral incentives as well as through the use of threat or force. How this happens is the domain of global governance.

Global governance is distinguished from international governance. International refers to relations among states. Global encompasses activity at the international, transnational, and regional levels. For some, the term also includes a recognition of the effects of local activity on the global, such as individual auto emissions of carbon dioxide on the atmosphere, and global effects on the local, such as transnational capital flows on the health of subnational economies. For most, global refers to those activities in the public and private sectors that transcend national boundaries.

1. *Governance in the Context of Rapid Globalization*

Heightened attention to concepts of global governance at the end of the twentieth and beginning of the twenty-first century comes as a result of increasing consciousness of the interconnectedness of human activity on the planet. Basic discoveries about the interdependence of species and habitat for survival, about the effects of human activity on the biosphere, about the volatility associated with rapid capital flows across national borders, and new inventions that permit people to talk to one another around the globe contribute to a sense of the globalization of human society. It is in this context, and in the growing knowledge of the harm that such global processes may bring about, that the idea of global governance takes on greater importance. The problems of shaping societies so that their activities do not produce irreversible harm to the biosphere and to the sustainability of human species prompt increased attention to the possibilities of improved global governance for the public good.

2. *Perspectives on Global Governance*

Ordered and patterned interactions on a global scale are the result of international agreements, of harmonization of laws between countries, of transnational cooperation among functional agencies, of self-enforcing contracts between private firms, and of standard operating procedures encoding daily activities of governmental and nongovernmental organizations. Taken altogether, this array of formal legal arrangements, informal codes of conduct, and habitual practices based on shared expectations about legitimate action, constitute global governance.

So, for example, the United Nations is a formal international arrangement based on legal agreements among states and is considered to be part of global governance. It is one of many formal institutions, however. Others would include the World Trade Organization, the North Atlantic Treaty Organization, the Organization of American States, as well as other regionally based treaty organizations. In addition, transnational cooperative actions among national law enforcement agencies also are part of global governance. Combating illegal criminal activity, for example, is accomplished through cooperation among agencies in apprehending offenders. In civil law, courts set precedents and decide cases based on case law from other national judicial systems. And in commercial law, private mediation is used by business firms from different countries to settle disputes (Slaughter 1997).

Many of these informal institutional arrangements are supported by networks among individuals and agencies. Cross-national professional relationships provide opportunities to learn from the experience of judges, policy officials, or executives from other countries. The annual World Economic Forum at Davos, primarily for business people, is just one rather visible example of informal forums. Through these and other meetings, professionals and executives begin to develop shared goals and values that can contribute to common standards and rules for conducting business, protecting human rights, conserving the environment, or implementing humanitarian missions.

The development of networks has accelerated with the introduction of high-speed electronic communications technology. Even before the advances in this area, however, increased travel and telephone communication contributed to a worldwide human rights movement, a global environmental movement, an international women's movement, the development of transnational corporations, and of international professional societies. Such networks have played a distinctive role in helping to establish common standards in a number of fields. In the case of human rights, for example, the diffusion of norms about the rights of prisoners came about through the use of international media to publicize cases of abuse. Such public 'shaming' of governments is based on a shared norm about the value of individuals and the need to protect their rights. Some of these values have been codified in formal national and international laws, while others are observed because they enhance the public reputations and legitimacy of governments or of businesses.

In sum, the concept of global governance builds on behavioral understandings of law and society. These perspectives suggest that norms of appropriate behavior emerge from daily interactions and the development of community consensus about what is fair and just, and is then codified in law. Law is a dynamic and changing reflection of shared standards. Implementation and compliance are understood as rather more voluntary than enforced, although the threat of sanctions—either criminal or civil—is recognized as an underlying motivating force for compliant action.

The simple example of stop signs illustrates the dual character of rules. Individuals obey stop signs because they may be afraid of getting caught breaking the law; but perhaps more important for some, they obey the stop sign because they wish to avoid being hit by another car. Both the threat of sanctions and an interest in avoiding harm are thought to be at work. In another example—of the private adjudication of transnational business disputes—firms wish to preserve flexibility and avoid undue government interference, which they may view as harmful. So they are willing to comply with the judgement of an agreed-upon private mediator rather than insist on taking the issue to a public institution for settlement.

3. *Elements of Global Governance*

It is possible to identify fundamental elements that constitute the building blocks of contemporary global governance. The most prominent of these are standards, norms, and states.

3.1 *Standards*

One means of coordination among private actors is the setting of standards. For example, the standardization of time through the use of Greenwich Mean Time permits coordination of communications, airline schedules, television broadcasting, and a myriad of daily transactions that we now take for granted. Rules of international navigation and airline procedures ensure the safety of air travel across national borders. A standard railroad gauge eases transborder crossings. Product standardization rationalizes the multinational manufacture of automobiles, televisions, refrigerators, and so forth. We begin to see how useful these standards are when we do not have them. The lack of standardization in electrical plug and socket connections across continents is an example of a lack of common standards.

In general, standards are accepted because they are thought to be voluntary and are arrived at through reason, that is, through the exercise of professional expertise, in which fundamentally equal actors reach decisions through rational deliberations that are open to interested parties. In practice, the process of standard setting by bodies such as ISO 9000 is completed by those with expertise in a field (in accounting or engineering, for example), representing users and pledged to serving the public interest. Efficiency in operations is essential to the acceptance of standards. Furthermore, the legitimacy of standards is based on the desirable consequences of the rule, which allows no exceptions or variations. To the extent that efficiency is a value in itself, and to the extent that professions embody and transmit a set of values, standard setting is not entirely value-neutral. Yet, in comparison with hierarchy, markets, or norm-based communities, standardization is a minimal form of coordination.

3.2 *Norms*

Norms are internalized rules based on commonly held values. In the international security area, non-use of nuclear weapons is cited as a shared norm, as well as a major achievement. The norm is based in part on a shared revulsion at the devastation to human life and society that is inflicted indiscriminately in a nuclear explosion. The success of the human rights movement is based on publicizing human rights abuses perpetrated by national governments. To the extent that 'shaming' governments is a successful strategy it is because a norm is widely shared about the rules for appropriate treatment of individuals by their governments.

Communication of common values through informal and professional networks is critical to the development of shared norms. It is through informal networks, those that include representatives from international institutions, national governments, and nongovernmental voluntary associations, and in some cases from private business, that common understandings and values are established. These commonly held values in turn, contribute to the management of environmental issues, human rights abuses, humanitarian relief, and security relations. Some analysts view these networks as an essential component of global governance, providing a way of governing without government (Reinicke 1998).

3.3 *States*

The organization of nation-states in the twentieth century has included a dependence on and growth of bureaucracy, or hierarchy, as a means of governing. The growth of the modern state has included regulation of markets, of educational institutions, of environmental conditions, of criminal activity—in fact of nearly every aspect of daily life as a means of coordinating action for the common good. While some of this regulation has come through standard setting and norms development, the characteristic mode of hierarchical government is through directives. These directives are enforced through the use and threat of punitive sanctions, including imprisonment and fines.

As noted earlier, the enforcement power of the state is based on its monopoly over the legitimate use of force. This monopoly, in turn, is thought to be based on a notion of a social contract, whereby citizens give up some rights to a sovereign government, in return for protection and through the establishment and preservation of social order. The central problem for

theorists and practitioners of global governance is that no world government exists, yet transnational activity generates needs for coordination as well as problems of collective action. We turn next to a discussion of methods of coordination, problem-solving, and conflict management currently in use—methods that constitute global governance in practice.

4. *Methods of Global Governance*

4.1 *Harmonization of Laws*

The implementation of international agreements is accomplished through the adoption of national laws to bring local and national practices into compliance with international rules. For example, carrying out the provisions of the Montreal Protocol to protect the atmospheric ozone layer is done through legislative action in each country to bring manufacturing practices into conformity with the Protocol's agreed directives to reduce production and release of chlorofluorocarbons (CFCs).

Legislative action is one path to harmonization of laws. Another is through the courts and judicial systems of countries. A body of transnational law is developing informally in which judges in one country are basing their decisions on precedents and case law from other countries (Slaughter 1997). This is taking place in areas of civil and commercial law, where standard rules help to smooth business and legal transactions. In the area of human rights law courts are also citing cases in other national systems, in addition to the United Nation's Universal Declaration of Human Rights, to lend credibility to decisions about the protection of civil and political rights in their own countries.

4.2 *International Regimes*

International regimes are defined as a 'set of implicit or explicit principles, norms, rules, and decision-making procedures around which actors' expectations converge' (Krasner 1983). The effectiveness of a regime rests on the operations of institutions, organizations, governments, and international bodies that share a set of principles, rules, and norms in a particular area of international action. Although regimes include formal treaties and national law, they also rely on informal norms and networks to develop and enforce standard behavior in an area of global policy.

For example, regimes to control the development and spread of nuclear weapons had been relatively successful in the context of the Cold War. They were critical to controlling the US–Soviet arms race, to preventing nuclear war, and to limiting the spread of nuclear weapons technology and weapons. These regimes had their basis in a number of formal treaties and international agreements, such as the Anti-ballistic Missile Treaty of 1972 and the Nuclear Non-proliferation Treaty of 1968, as well as some enforcement powers through the United Nation's International Atomic Energy Agency. For compliance, however, regimes relied on a number of formal and informal surveillance techniques and the trust that emerged from shared norms about the dangers of nuclear weapons, their development, production, and proliferation (de Nevers 1999). A decade after the end of the Cold War, we are beginning to see how much trust the USA and the Soviet Union had in one another, and in the shared norm of the non-use of nuclear weapons, as the arms control regime is being tested by the actions of several parties to these regimes.

4.3 *Global Policy Issue Networks*

Development of global policy and implementation of common laws also depends on informal networks of organizations and individuals. These networks are composed of individuals who share an interest in providing a public good that may not be provided by any single nation-state. Through communication networks, sometimes built around professional associations and scientific bodies, these transnational action networks (Keck and Sikkink 1998) or epistemic communities (Haas et al. 1993) provide information, documentation, and alternative policy solutions to global policy problems.

Some of the most visible networks have formed since the mid-1970s on human rights, environmental protection, and women's issues. In each of these areas, lawyers, scientists, and political and cultural leaders have found counterparts in other countries to form worldwide movements that have publicized the harm to people and the environment brought about by current national policies and practices. In the case of women's issues, the role of nongovernmental organizations at the official United Nations conferences on women and on population has been dramatic. The adoption of women's empowerment as a major path to fertility reduction, rather than an emphasis on birth control, at the population conference in Cairo in 1995 demonstrated the influence and effectiveness of issue networks in shaping global policy.

4.4 *Hybrid Institutions*

As transnational activity has increased, as non-governmental actors have become more prominent, and as national governments have devolved authority to private sector bodies, new institutional arrangements have developed to manage global relations. These arrangements are referred to as hybrid institutions or public–private partnerships because they mix

private sector organizations and a variety of incentive systems with governmental agencies to achieve public goods. They will often combine standard rules, private market arrangements, nongovernmental norm-based organizations, and government agencies—both national and international.

Some of the most inventive have emerged in the area of environmental protection. The certification of sustainably harvested forest products is just one example of a hybrid global governance process. Presented as an alternative to governmental regulation of the timber industry, environmental nongovernmental organizations have worked with major private enterprises in the forest products industry to convince them of the economic benefits of sustainable forest practices. They have demonstrated that consumers will pay premium prices for furniture and other lumber products that are sustainably harvested. They have shown that profits can be achieved over a longer period of time, and that reducing dependence on clear-cutting provides livelihoods for those people, mostly in developing countries, dependent on forests for their survival.

The major method of ensuring compliance with sustainable practices is through certification of wood products. Private firms participate in a council that includes producers, scientists, and other independent interests. This council, in turn, inspects production operations and certifies that manufactured products are derived from sustainably harvested lumber. The manufacturer can use that certification to advertise products and, in most cases, to charge higher prices to consumers. This process takes place without government intervention; it includes market incentives and relies on professional expertise and standardization to enforce rules that will contribute to the protection of natural resources—to the conservation of a public good.

Another example of hybrid institutions can be found in the area of prosecuting transnational criminal activity. Local and national law enforcement agencies, federal bureaus, banking regulation bodies, international police forces, and banking and financial institutions are developing transnational relationships and standard operating procedures to root out and prosecute international criminals. Organized crime has become relatively border-free, especially as liberalization of economic relations has proceeded; yet law enforcement remains within fixed territorial legal frameworks. To deal with this challenge, national law enforcement agencies have developed cooperative relationships with their counterparts in other countries to share information and to apprehend criminals. For example, US and Russian law enforcement agencies—the Federal Bureau of Investigation and the FSB (Federal Security Service, former KGB)—have set up cooperative relations to investigate and help prosecute corruption in banking and industry in their two countries.

To investigate and report money-laundering activity, the US Treasury Department has set up a Financial Crimes Enforcement Network. This network includes US Treasury Department bureaucrats, law enforcement officials, banking regulators, and private banking and financial institutions around the world. Through a series of agreements, government agencies and private sector organizations have established shared authority in identifying suspicious financial transactions. Banks have opened up their accounts to greater scrutiny and, in exchange, government authorities allow banking operations themselves to determine when activity is suspicious, and thus susceptible to further investigation by law enforcement agencies. Standards have been set about what constitutes suspicious behavior, and data about transactions is shared electronically between the private and public sectors (Reinicke 1998, pp. 135–72).

5. *Issues in Global Governance*

While new modes of coordinating transborder activity are emerging that hold the promise of effective global governance, a number of theoretical and normative challenges have been identified. Two are addressed here: the legitimacy of authority and democratic values.

5.1 *Legitimacy and Authority*

Successful governance requires the consent and compliance of those whose activity is being governed, that is, practices and institutions that constrain actions must be seen as legitimate for them to be effective. In contemporary conceptions of governance, this aspect remains problematic. In contrast to theories of governance at local and national levels, the idea of a social contract between citizens and institutions of global governance has not been developed sufficiently to be credible. Such social contracts have depended on theories of participation and representation by citizens in decision-making processes, few of which figure centrally in global governance. Other sources of legitimacy, such as claims of monarchs to 'divine right,' or to hereditary rights, are not recognized in twenty-first century conceptions of global governance. Finally, however, legitimacy claims may also be based on what works well, what is efficient, or what is reasonable; contemporary justifications for global governance mechanisms are often based on efficiency warrants.

Issues of authority and legitimacy become even more vexed with the increased participation of nongovernmental organizations in issue networks and hybrid institutions. While the authority of governments and the legitimacy of private firms are based, in some measure, on implicit contracts with citizens or consumers, respectively, the authority and legitimacy

of nongovernmental organizations is based on their claims of acting in the public good—as they define it. Whether by protecting individuals from the human rights abuses of their own governments, or conserving biodiversity in the face of extractive private industry, or saving women from customary rituals that cause physical harm, nongovernmental actors believe they are upholding standards that are in the long-term interests of humanity. These organizations, however, are not assembled through an election process or any other democratic procedure. They are self-appointed and claim to be acting on the basis of values that serve a global public interest. Without an authoritative institution that provides a process for establishing common values and agreement on what constitutes the public good, issues of authority and legitimacy in global governance will most likely remain unresolved.

5.2 Democratic Values and Global Governance

While some definitions of global governance emphasize the neutral management of human affairs, there is a strong normative component to any discussion of the term, in addition to an underlying assumption about who governs and upon whose standards. Some of the normative implications are acknowledged, but others need to be surfaced and examined before agreement can be reached on a standard definition.

In its current conception, global governance implies democratic governance. Most authors are careful to make the distinction between the term 'world government'—that is, a set of procedures and mechanisms based on a single authority with power to enforce laws, by force, if necessary—and the notion of governance. This conception also recognizes that a greater number and range of organizations, individuals, institutions, and governments participate in decisions that have consequences beyond national borders (Mathews 1997). The informal and pluralistic nature of global governance both recognizes and requires the contribution of many of these nongovernmental actors to policymaking. In its own terms, then, global governance incorporates democratic values including those of equitable participation and the use of representative institutions for resolving conflicts. Yet the reliance on scientific and professional bodies to set standards, rules, and procedures, on bureaucratic agencies to implement policies, and on voluntary organizations to monitor compliance, raises questions about the compatibility of democratic values and the concept of global governance.

See also: Area and International Studies in the United States: Intellectual Trends; Global History: Universal and World; Globality; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; International and Transboundary Accords, Environmental; International Justice: Philosophical Aspects; International Organization

Bibliography

Allott P 1990 *Eunomia: New Order for a New World*. Oxford University Press, Oxford, UK

Brunsson N, Jacobsson B (eds.) 1995 *Standardisering*. Nerenius and Santerus, Stockholm

Commission on Global Governance 1995 *Our Global Neighborhood*. Oxford University Press, Oxford, UK

de Nevers R 1999 *Regimes as Mechanisms for Global Governance*. Rockefeller Brothers Fund, New York

Graham C 1998 *Private Markets for Public Goods*. Brookings Institution Press, Washington, DC

Haas P, Keohane R O, Levy M A (eds.) 1993 *Institutions for the Earth, Sources of Effective International Environmental Protection*. MIT Press, Cambridge, MA

Held D 1995 *Democracy and the Global Order: From the Modern State to Cosmopolitan Governance*. Stanford University Press, Stanford, CA

Kaul I, Grunberg I, Stern M (eds.) 1999 *Global Public Works: International Cooperation in the 21st Century*. Oxford University Press, New York

Keck M, Sikkink K 1998 *Activists Beyond Borders: Advocacy Networks in International Politics*. Cornell University Press, Ithaca, NY

Krasner S 1983 *International Regimes*. Cornell University Press, Ithaca, NY

Mathews J 1997 Power shift. *Foreign Affairs* **76**(1): 50–66

Reinicke W 1998 *Global Public Policy: Governing Without Government?* Brookings Institution Press, Washington, DC

Rosenau J, Czempiel E (eds.) 1992 *Governance Without Government: Order and Change in World Politics*. Cambridge University Press, Cambridge, UK

Slaughter A M 1997 The real new world order. *Foreign Affairs* **76**(5): 183–97

Young O 1989 *International Cooperation: Building Regimes for Natural Resources and the Environment*. Cornell University Press, Ithaca, NY

K. Benedict

Global History: Universal and World

1. Historiographical Traditions

Global histories have been written since Herodotus (495–425 BCE), whose cosmopolitan concerns were commended by Cicero but deplored by Thucydides—who opined that Greeks had no business investigating alien mythologies, religions, customs, and traditions. History writing, Thucydides suggested, should be modeled upon his own study of the Peloponnesian Wars—sharper in focus, shorter in time horizons, based upon verified facts and prescriptive in purpose. Fortunately Herodotus ignored his 'Eurocentric' advice and ranged widely beyond the Hellenic world to include Egypt, India, Babylonia, Arabia, and Persia in his histories in order 'to preserve the memory of the past by placing on record the astonishing achievements of both our own and of the Asiatic peoples.' He used

oral testimony, archeological remains as well as written sources and imposed some chronology and order on streams of events that had occurred on several continents and over long spans of time. His (still popular) book ends in 'European triumphalism.' Its story turns into celebration of the victories of the Greek *polis* (especially Athens) over the Persian Empire, which Herodotus represented as a conflict between occident and orient, despotism and freedom.

Global historians applaud Herodotus for his reflexive interest in barbarian virtues and regret the long hiatus that occurred in the writing of secular world histories until Voltaire and his generation carried the project forward again during those brief decades of Enlightenment, before the outbreak of the French Revolution. They are puzzled, however, to observe that with the exception of a few Stoics (Diodorus, Polybius, and Dionysius) and a handful of geographers and ethnographers (Strabo, Ptolemy, and Pliny), histories of the world did not attract more scholars writing during Graeco-Roman times. Although they lived in empires that maintained contact with an array of cultures (African, Arabian, Persian, Indian, and Chinese) historians of the classical era remained overwhelmingly concerned with Europe. Their histories (e.g., Tacitus) deal mainly with the politics and scandals of Rome, or (e.g., Thucydides) with wars among Greeks. Their equally provincial oriental contemporaries, residing in that celestial empire on the other side of the world, asserted with quintessentially Chinese superiority that they remained unimpressed with anything the Romans had to impart, or to trade for, China's silks, ceramics, medicines, perfumes, and spices.

In Europe, steps in an ecumenical direction occurred when Christian and Jewish historians started their narratives with the Creation and set out to ensure that Greeks and Romans should not be left (in the words of Josephus) 'as the arrogant possessors of antiquity.' That was why Orosius (a student of St. Augustine) proclaimed that 'The Roman Empire arose in the west but was nourished in the east.' Narratives in Christian historiography is dominated by chronicles about the evolution of mankind over stages of history, no longer bounded by the frontiers or cultures of the Greco-Roman world. In 1158 Bishop Otto of Freising recognized that 'all human power and learning had its origins in the East.' Medieval historians included Oriental and African civilizations right down to the Reformation when Catholic and Protestant intellectuals began to look for a usable past to help princes to integrate populations and secure the frontiers required for the formation of nation states. Even then both parties to Europe's conflict over religion and national identities remained aware of places, peoples and cultures not merely outside the borders of their enclosed and ideologically encompassing polities, but beyond Christendom itself, and that is hardly surprising.

Residing on an underdeveloped promontory on the western edge of the great Eurasian landmass, with ice to the north, uncharted oceans and territory to the west, and the hostile power of Islamic communities to the south and east, Europeans had not secured their frontiers, religions, cultures, and technological superiority over other civilizations, long before the eighteenth century. It took centuries to clear the Iberian peninsular of Arab power. Moslem fleets threatened European shipping in the Mediterranean for another two centuries after the surrender of Grenada. Turkish armies crossed into the Balkans in 1350, took Constantinople in 1453 and menaced Vienna as late as 1683. For nearly a thousand years after the fall of Rome, Christendom braced its defences, imported useful knowledge, shaped its collective identity and composed ideologically charged histories in the context of conflicts (and encounters) with the powerful civilization of Islam on all its frontiers. Memoirs, travelogs, diplomatic correspondence and investigations into Arabic medicine and astronomy, together with commercial intelligence on societies under Ottoman dominion, flowed from east to west. Much of this knowledge came through Genoa, Florence and Venice, along with spices, herbs, sugar, botanical drugs, jewels, chinaware, silks, cottons, and metal goods, imported through the Middle East, from India and China. The expense of procuring and transporting luxuries moving slowly, over long distances and subject to extortionate charges for protection along the way, prompted Europeans to seek ways to surmount the barriers to commerce with South Asia and the Far East by their Islamic enemies.

Shortly after the voyages of discovery Europeans broke out of their political, commercial and mental encirclement by Islam. Over the sixteenth century flows of information into Europe widened to include the Americas, and increased exponentially as a result of the establishment of regular and direct trade and contacts by sea with India, China, South East Asia and Africa. Yet nearly two and a half centuries elapsed before Europeans produced a secular and enlightened school of world historians. Its godfather was Voltaire, who composed a history concerned with 'only that which is worth knowing; the spirit, the customs, the practices of the principal nations based on facts which one cannot ignore.' Voltaire and his followers in France, the Netherlands, Scotland, Naples, Germany and other parts of Europe, deliberately constructed histories of their expanding world in ways which they conceived as a departure from the providential narratives of their clerical predecessors and the 'nationalism' of historians writing for princes and their obedient subjects. That aspiration was propitious. As a result of regular contacts and commerce the trickle of information about peoples, geographies, technologies, politics, warfare, customs, manners, and the import of all sorts of commodities and artefacts turned into a river of knowledge about places, populations, societies, and

economies beyond Europe. For radical intellectuals China became an alternative model, favorably contrasted with Europe's political systems, social arrangements, and religious beliefs. European histories became not only wider but deeper as scholarship took over from propaganda, sycophancy, fantasy, religion, and national ideologies.

For something like half a century before the French Revolution and the takeover of India and South East Asia, European historians resituated Europe's and Christendom's place in world history by including more ancient as well as classical Mediterranean civilizations and by widening their gaze to take China, India and Japan as well as the newly discovered Americas into account. Europe's new school of world historians dealt with the natural histories, religions and ethnographies of most parts of an increasingly interrelated universe. Favorable interpretations of oriental polities, societies, cultures, and religions by philosophes and physiocrats did not go unchallenged. Bernier and Leibniz, Montesquieu, Defoe, Winckelman, and Hume remained resolutely unimpressed with the Chinese empire. Nevertheless historians of that era engaged in debates about cultures, geographies, and spans of time that their modern counterparts (confronted with libraries of knowledge and professional interpretations), can only envy.

Alas, the ironic reflexion that their narratives of Enlightenment brought to the task of repositioning Europe in a secular history of the world was overshadowed by a tradition of Western triumphalism which emerged in the aftermath of the Revolutionary and Napoleonic Wars, 1789–1815; when Europeans acquired massive additions to the populations, territories and natural resources of other continents under direct or informal control. During an 'imperial meridian', European naval and military superiority over the states and societies of other continents emerged as irresistible. Furthermore, by 1825, European settlers throughout the Americas and the South Pacific had escaped from all metropolitan and aristocratic restraints upon their economic activities and their relations with the indigenous peoples of colonized worlds. White settlers became free to exploit 'their' heritage under rules of their own making, tempered only by religion. As its technological and economic lead over other continents widened, Europe moved clearly onto a path which led, as time went on, to clear divergences in levels of productivity and living standards between East and West and North and South.

Against this background, the writing of history by Europeans departed from its moment of cosmopolitan enlightenment, moved in a Hegelian direction and returned to the introspection and cultural arrogance associated with imperial Rome. Here and there aspirations to write universal histories survived. Isolated scholars eschewed the condescensions embodied in nineteenth century science, technology, liberalism, and colonialism, and a scatter of historians continued to work outside the encompassing national frameworks, and dominant teleologies associated with the rise, rationality, and hegemony of Europe. Nevertheless, it took the destruction and cultural shock of two world wars in the twentieth century to encourage a return to the rewriting of histories of mankind. That program, conducted by a small group of erudite scholars, began in the climate of pessimism and breakdown in the liberal consensus that followed the Great War, when Spengler, Toynbee, Sorokin, Wells, Mumford, and Dawson wrote universal histories of humanity to help Europeans comprehend the decline and barbarity of the West. In history they attempted to find some substitute for Neitzche's proclaimed demise of Christianity and rationality. Their scholarship is inspiring, but their explicit search for spiritual meanings in the histories of the rather amorphous entities called civilizations did become a compelling framework for scholars and students to work within. The attempt by UNESCO, after the World War II, to fund the writing of a multivolume collaborative history of the world (designed to promote peace across nation states) failed to provide a paradigm for the reconstruction of secular of histories on an ecumenical scale. Long after the war, historical scholarship continued to be located where its founders had situated it from the time of the Renaissance and Reformation onwards—namely, within the confines, chronologies, and parameters established for comprehending the origins and evolution of European (including North American and Australasian) societies and nation states.

2. *Globalization and the Revival of Global History*

Fortunately pockets of global history survived—notably at the universities of Chicago, Northwestern and Leipzig. Complementary branches of scholarship continued to flourish in the forms of: survey courses in Western civilization; area studies, histories of trade and imperialism, investigations into the impact Western technology and culture upon other peoples and above all, as flows of encyclopedias, chronicles and textbooks for schools and popular education concerned with the world as a whole. Thus, 'tribes' of professional historians with interests in fields that transcended nation states and Europe possessed the credentials and remained ready to refashion their intellectual identities, when demands for global history reappeared from the late seventies onwards. That revival could be retraced to canonical publications by M. Hodgson, W. McNeill and L. Stavrianos, but research and teaching has only matured into university programmes for the sustained study of world history over the past quarter of a century of 'globalization.' Although there is nothing particularly new about the current phase of economic integration and inter-

dependence which are the core features covered by that label. Long distance trade goes back millennia to prehistory, and the exchange of commodities across frontiers and overseas had always involved arrangements for counterpart flows of funds, conversions of currencies and the operation of transnational, commercial and financial organizations of a more or less sophisticated kind within and across Europe, Africa and Asia.

Nevertheless real discontinuities had occurred. For example, by the late nineteenth century, supported by steam-powered ships and trains, the volume of commodities and services traded across national borders, continental boundaries and oceans had increased dramatically compared to 1800. Commerce was promoted and sustained by movements of capital, migrations of labour and transfers of technology and information around the world on an unprecedented scale and at ever increasing speeds. Political impediments to international flows of exports, imports, money, credit, capital, labor, technology, and information diminished sharply during the liberal international order that prevailed between 1846 and 1914.

State controls revived again during nearly four decades of neomercantilism that succeeded World War I; when governments everywhere sought to insulate their domestic economies from inflows and outflows of commodities, from the emigration of their nation's capital and skilled labor, and from fluctuations associated with unregulated rates of exchange. In the interests of narrowly defined national competitive advantages and political stability, sovereign states proceeded to de-link economies from participation in an evolving global economy. Between 1914 and 1948 overall levels of world trade fluctuated in volume, but ceased to grow. Warfare and political regulations depressed movements of capital and labor across frontiers. Controls on the convertibility of currencies hampered international exchanges of every kind.

Normalcy returned after World War II when a discernible shift towards liberalism in the political and legal framework for commerce among nations was also promoted by even more radical improvements in new technologies for the transportation of goods and people across space and time; and by vastly more efficient communications systems for the widespread diffusion of commercial intelligence, messages and cultural information between distant locations and separated populations. Revolutions in transport, television and communications widened markets and created opportunities for the organization of transcontinental and multinational business across an expanding range of production, distribution, finance and commerce.

Unsurprisingly, the implications of 'globalization' for hitherto separated communities, bounded spaces, distinctive cultures, partially connected local economies and nominally sovereign states have been analyzed by social scientists. Intensified demands have appeared for historians to supply perspective on trends towards integration—that often appear to economists, political scientists, sociologists, and philosophers (as well as natural scientists) to represent profound discontinuities in the ways that ubiquitous technological, economic, and cultural forces are clearly shaping our interconnected world. This demand is supported by Western governments, uneasily aware that their powers to control societies, nominally under their rule, are transforming their claims to effective sovereignty over the populations, territories, and assets within their borders into something more tenuous than used to be the case during the long era of competing nation states. Their anxieties (eloquently fostered by 'official' intellectuals) have been compounded by recent military, technological, and economic achievements of several Asian societies, especially Japan and, the 'Tigers' but latterly (and more significantly) by the reawakening of China. The rest are perceived to be catching up and modern mercantilists (who continue to conceive of the world in terms of winners and losers or clashes of civilizations), worry about the relative decline of Western hegemony. Politicians and publics turn to historians for explanations of the past and more reassuring predictions for the future.

National frameworks for political action and academic enquiry are now widely recognized as unsatisfactory. That comprehension is patently functional for natural sciences investigating environmental issues, for biological sciences concerned with the health of human populations, and for social sciences analyzing international relations, crime, migration, communications, and whole ranges of economic activity, which can no longer be regulated within the boundaries of nation states. It is also clear that national and local cultures are being subverted by advertising, by fashion, and by the popular arts, especially music. The means and media of modern transportation, travel and communications (cheap and available to the masses), are opening up cosmopolitan discourses (usually in English) that contribute to the reshaping of personal identities around the world and transform frameworks for behavior, especially among younger generations.

3. *Traditional, Postmodern, and Global Histories*

Pressures to make, write and teach histories of the world that might become more contemporary and relevant has encouraged the discipline to become less intellectually, politically, spatially and chronologically confined. They are intensifying, seem irresistible and are being welcomed as a return to the ecumenical programmes of Herodotus, Vico, Voltaire, Leibniz and von Ranke who opined 'there is no history but universal history as it really was.'

Intellectual arguments for resituating national and European histories into frameworks that might better address the problems of the new century have been persuasively rehearsed by proselytisers for global history in recent years—as indeed have their antitheses—recommendations to concentrate attention upon microhistories of difference, diversity and a stance of humility towards the complex histories of 'unknowable others.' That postmodern exhortation is understandably suffused with a concern that the revival of global history will implicitly celebrate 'triumphs of the West' in science, technology, military power, and economic development. Unfortunately that has already occurred in several recent publications by journalists and even by historians whose pretensions to offer overviews of world history seem truncated in time, confined in space and selective in areas of social endeavour emphasized for recognition as universal achievements; and whose understanding of environmental, biological, and other natural constraints on human action look unscientific.

As humanistic scholarship, global history aims to mature into a renaissance and not a return to the arrogance of Rome or to Victorian triumphalism. Modern history has matured into a capaciously catholic subject, willing to take many kinds of hitherto myopic exclusions into its narratives, to reposition knowledge of the past into more relevant spaces, models and chronologies and to plunder the natural and social sciences (as well as archeological remains) for insights into the evolving human condition. In finding voices and viewpoints to construct histories on a global scale, historians (alive to the second cosmopolitan enlightenment that emerged during the closing years of an otherwise dark twentieth century) have not found it difficult to detach their narratives from another politically correct but romanticized appeasement of the underprivileged on the one hand, or from the implicit denigration of non-Western people and cultures, that flows from Eurocentric convergence and modernization theory on the other.

In an ongoing search for strategies to teach and to write compelling histories on a global scale, they will respect the preferences of their colleagues, whose postmodern anxieties about their capacities to comprehend 'others' or to offer plausible conjectures and metanarratives about major trends and significant events, leads them towards dense description, microhistories, and biography. Such flowers should certainly bloom, but the intellectual effort required to landscape the garden also demands attention and imagination from historians.

Global historians confront more challenging intellectual problems in trying to persuade the majority of their academic colleagues (writing bounded histories and bunkered in the security of national archives), that their attempts to write on a global scale could meet the standards of a discipline deeply concerned with detail and documentation that discourages generalization beyond strictly confined geographies and periods. Having rejected Toynbee's quest for spiritual guidance as a tenable mission global historians tend to make a general case behind arresting metaphors extolling the need for histories constructed as: 'maps on the largest scale,' as 'photographs taken through the lenses of wide-angled cameras,' as 'architectural frameworks' and 'as views; contemplated from space.' Many contend that the boundaries and chronologies of parish, let alone national or continental histories, seem equally problematical to define. Nevertheless, coming as most do, as recognized specialists in but one or two fields of history, global historians remain conscious of their credentials and are entirely sensitive to their epistemological vulnerability. It continues to be much easier to conduct research, claim territory and acquire status as national or local historians.

4. *Current Paradigms: Connections and Comparisons*

Nevertheless, historiographical and methodological discussion as well as the volume of publications (off and on line) has increased exponentially. Most of the bibliography is concerned with connections and comparisons across national frontiers and cultural boundaries. Most of it displays a scientifically informed appreciation of evolutionary environmental and biological constraints on all forms of human activity, and represents a welcome reunification of natural and human history. Celebrated and influential books display a talent for the construction of synoptic histories about matters that have been important for humanity as a whole. As more historians risk writing on a grand scale, the field looks set to produce plausible, but competing metanarratives into which the overwhelming flow of parish, regional and national histories could be reconnected.

Meanwhile two distinct approaches have already developed. First (and central) is the model, persuasively exemplified and elaborated in several books by McNeill. He has inspired a programme of historical investigation into connections across continents, countries and geographies as well as through very long spans of time. McNeill's assumption is that such connections, interactions, encounters, contacts with outsiders can be represented as the origins and engine of economic, social, political, military, cultural, religious, technological, and all other conceivable types of change studied by historians; and that connection can be pursued with scholarly detachment and a proper appreciation of the long chronologies required to construct histories that aim to select and analyse *major* forces at work in the evolution of humanity, McNeill's recommendations aim to avoid the condescension of cultures, the restrictions of time and the arrogance of nations built into currently dominant

styles of history. His (and Crosby's) books show that 'interactions' were often malign: the spread of plagues, diseases, and parasites; those waves of destructive invasions by nomads; wars of conquest, plunder, and imperial expansion; forcible conversions to alien systems of religious belief; and the subversion of indigenous traditions, cultures, societies, and communities by strangers, have been part of the universal experience of suffering.

Well-specified types of connection can be investigated over significant periods of contact and transformation. Its taxonomical divisions include: trade, investment, warfare, religion, migrations, transfers of useful knowledge, technological diffusion, botanical exchanges, and the spread of diseases. Proper attention is also accorded to the means of transportation and modes of communication that secured, facilitated and cheapened contacts by sea, land and latterly by air. Once a relevant range of major connections are traced, understood and their significance evaluated global history will broaden our education and provide the spatial and chronological perspectives required for the appreciation of the histories of nations, peoples, localities, and groups embodied within particular histories that the majority of their colleagues study in greater depth and sophistication, but too often in isolation.

Put simply, the second major approach extends geographical catchment areas for comparative histories of topics that could be the subject of illuminating study across national boundaries, continents, and separable cultures. Although the linguistic qualifications for such research are extremely difficult to acquire, and the scholarship needed to make sense of contrasts discovered is formidable, the methodological problems are similar to those encountered for exercises conducted within European, national, or indeed regional frameworks.

Comparative history helps scholars to surmount the complexity and tyranny of local detail and seeks to offer coherent but sensitive answers to a great variety of questions that historians select for investigation. For global historians, the illumination derived from the method tends to be realized when they can concentrate upon well-defined artefacts, institutions, organizations, social practices, attitudes, and beliefs that are found in dispersed cultures, and which have already been studied in scholarly depth, in the context of particular locations and which exhibit comparable, but, more importantly, dissimilar geographical economic, political, social and other features. Then, as Marc Bloch anticipated, the 'comparative method can elicit from the chaotic multiplicity of circumstances those contrasts which were generally effective.'

In order to communicate salient differences global historians will be tempted to aggregate and average contrasts across wider spaces and larger populations (continents, oceans and civilizations) than their colleagues who have the sources to engage in exercises in comparative history within circumscribed boundaries. For purposes of research, designed as building blocks for the production of metanarratives that are truly universal, the comparative method looks set to dominate the field for years to come. It has, for example, already generated a bibliography of global histories concerned with the family, youth, marriage, diet, housing, health, military organization, government, human rights, parliaments, nationalism, religions, fundamentalism, and revolutions, and there are serious debates on two seminal topics: gender and long-run economic growth.

5. *Comparative Histories of Gender and Material Progress*

Gender is perhaps the quintessential 'analytical category' for purposes of universal social history and it interacts with other categories including class, ethnicity, race, nation citizenship, and colonialism for purposes of constructing world histories that 'keep humanity in view.' Yet no surprise will be occasioned by the appearance in recent decades of histories seeking to encompass the world economy—going way back in time and written to expound and to explain the disparate levels of material progress achieved by countries, societies, and communities located on all continents. Most people, in most places for most of history, have been preoccupied with obtaining food, shelter, clothing, and other manufactured artefacts that they required to sustain either a basic, a comfortable, and, only latterly, an agreeable standard of living.

A modern generation of historians is resituating the scientific, technological, and economic achievements of Western societies in global context. They call upon libraries of books and articles dealing with African and Asian agricultures, industries, towns, commercial networks, communications, trade, science, technologies, cultures, business organization, taxation, state systems, Government policies etc., etc., covering the last millennium and written in large part by historians from universities not long emancipated from imperial rule. This enormous, but still far from comprehensive, volume of historical research has, moreover, been communicated to the West by specialists in area studies from North American, European, Australian, and Japanese universities. Not long after World War II and during an era of decolonization, historians were offered the opportunity—provided by the accumulation of a large body of archival knowledge—long available in Europe and North America, but emerging in Asia, the Middle East and Africa and Latin America—to reposition their hitherto disconnected histories of wealth and poverty one against another in order to construct global histories of material progress that might have satisfied the aspirations of Montesquieu, Voltaire,

Smith, and their 'enlightened' followers, and pleased Max Weber.

Recent attempts to survey this monographic literature has been dominated by scholars from Western universities. Syntheses from China, Japan, India, Africa and the Near East would be a salutary correction to what is a plural but nevertheless European perspective. Meanwhile, scholars have carried forward a Weberian program for teaching and writing a truly global history of material progress. They have found that the quantitative evidence for such purposes would certainly not be regarded as sufficient to underpin a serious programme of historical research designed to account for long-term economic success and retardation among the national economies of Western Europe, North America, and Australasia. In the absence of hard data or even of an acceptable range of contemporary observations (which could be included in a narrative designed to track chronologies of divergence in productivities and living standards between Europe and Asia), global histories of material progress continued to be marred—by three powerful Marxist and Weberian assumptions. First, that Europe's technological and economic superiority over Asia (which became first visible in the eighteenth century and then unmistakeable during the nineteenth and twentieth centuries) had evolved gradually, but cumulatively, over several centuries before 1815. Over that time Europe's political, legal, cultural, and institutional heritage for the conduct of economic activity continued to be exceptionally favorable for the promotion, first, of Smithian growth, and, following on from that, for technological progress. Thirdly (and in contrast to Europe), other populations lived and worked in more hostile environments and this big ecological fact restrained and often arrested long-term economic development in Asia and Africa—although our knowledge of pre-colonial African economies remains entirely thin.

Those who follow Weber, Marx, and a tradition of writing that can be tracked back to Montesquieu, insist that Europe's competing economies had been on a potentially more promising trajectory for long-run growth for centuries. They emphasize persistent contrasts between European and Asian institutional arrangements, political systems, legal frameworks, cultural conditioning, and religious beliefs, within which economic activities continued to be embedded for a least four centuries before 1815. Cruder Weberians juxtapose the preconditions for long-run Smithian growth that they found operating efficiently in Europe (or rather in some societies in north western Europe) in sharp contrast to Asia, where political, social, familial, cultural and other negative institutions and beliefs restrained the emergence of competitive markets for commodities, capital, land and labor.

Their critics, who seem more deeply and widely read in the now massive bibliography of economic and social history dealing with South, East and South East Asia, and the Ottoman Empire, have produced samples of data (related largely to population growth, levels of urbanization, balances of commodity trade, yields per hectare and real wages) for selected regions of Asia and the Middle East that at the very least qualify any impression that labor productivity levels and standards of living afforded by European economies to their populations were discernibly (let alone significantly) superior long before the outbreak of the French Revolution. Of course at this level of 'continental' aggregation, the selection of samples of data for comparative history remains problematical because the Asian economy includes much larger shares of the world's cultivable land, resources, income, and population than early modern Europe.

Nevertheless, economic historians are a long way from offering statistically secure comparisons for average levels of productivity and living standards. Meanwhile, the accumulating volume of research now available on Asian agricultures, industries, trades, credit systems, mercantile enterprise, transportation networks, markets for commodities, and factors of production degrades simplistic Weberian perceptions that Europe alone had evolved the political, institutional, legal, cultural, and religious frameworks required for an effective process of Smithian growth long before other continents into dubious propositions. As Marshal Hodgson observed decades ago: 'All attempts to invoke premodern seminal traits in the occident can be shown to fail under close historical analysis.' Braudel, Chaudhuri, Frank, Pomeranz, Goldstone, McNeil, Sugihara, Washbrook and other historians of long-run material change concur. Jones would probably (as he did in *Growth Recurring*, 1988) continue to revise and reformulate some of the positions he adopted in the first edition of the European Miracle in 1981. From his comparisons of levels and types of development achieved by Europe and Asia for the early modern period, Braudel inferred that 'the populated regions of the world faced with demands of numbers seem to us to be quite close to each other.' But there is he continued 'a historiographical inequality between Europe and the rest of the world. Europe invented historians and made good use of them. Her own history is well lit and can be called as evidence or used as claim. The history of non-Europe is still being written. And until the balance of knowledge and interpretation has been restored the historian will be reluctant to cut the Gordian knot of world history.'

That acute observation applies across the entire field because histories of gender and capitalism are but two current areas of controversy that display the unrealized potential of universal history. In truth, it seems difficult to conceive of any grand theme that attracts attention from scholars engaged in writing national and local histories that could not be repositioned within wider geographies and longer temporal periods. For purposes of research, teaching and

debate, comparisons and connections are now the two dominant styles of global history. Pursued with sensitivity they should deepen our understanding of difference and diversity. In time they will meet demands for perspective from those trying to comprehend accelerated trends towards interdependence and integration on a world scale and allow for a less ethnocentric appreciation of the manifold achievements of more peoples, communities and cultures over long spans of human history.

See also: Civilization, Concept and History of; Comparative History; Gender History; Globalization and World Culture; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: East Asia; Historiography and Historical Thought: Indigenous Cultures in the Americas; Historiography and Historical Thought: Jewish Tradition; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Historiography and Historical Thought: South Asia; Historiography and Historical Thought: Southeast Asia; Historiography and Historical Thought: Sub-Saharan Africa; History of Science; History of Technology; History: Overview; Modernization and Modernity in History; Time, Chronology, and Periodization in History

Bibliography

Albrow M 1996 *The Global Age: State and Society Beyond Maturity*. Polity Press, Cambridge

Anderson B S, Zinsser J P 1988 *A History of their Own: Women in Europe from Prehistory to the Present*. Harper & Row, New York

Barnes H E 1962 *A History of Historical Writing*. Dover Publications, New York

Bentley J H 1993 *Old World Encounters*. Oxford University Press, Oxford, UK

Brook T, Blue G (eds.) 1999 *China and Modern Capitalism. Genealogies of Sinological Knowledge*. Cambridge University Press, Cambridge, UK

Chaudhuri K N 1990 *Asia Before Europe, Economy and Civilisation of the Indian Ocean from the Rise of Islam to 1750*. Cambridge University Press, Cambridge, UK

Costello P 1993 *World Historians and their Goals. Twentieth Century Answers to Modernism*. Northern Illinois University Press, De Kalb, IL

Crosby A W 1986 *Ecological Imperialism: The Biographical Expansion of Europe, 900–1900*. Cambridge University Press, Cambridge, UK

Curtin P D 1984 *Cross Cultural Trade in World History*. Cambridge University Press, Cambridge, UK

Diamond J M 1998 *Guns, Germs and Steel: The Fates of Human Societies*. Vintage Random House, London

Frank A G 1998 *ReOrient: Global Economy in the Asian Age*. University of California Press, Berkeley, CA

Goody J 1996 *The East in the West*. Cambridge University Press, Cambridge, UK

Heiser C B Jr. 1990 *Seed to Civilization: The Story of Food*. Cambridge University Press, Cambridge, UK

Hirst P, Thompson P 1996 *Globalization in Question: The International Economy and the Possibilities of Governance*. Routledge, London

Hodgson M G S 1993 *Rethinking World History*. Cambridge University Press, Cambridge, UK

Jones E L 1998 *Growth Recurring: Economic Change in World History*. Oxford University Press, Oxford, UK

Kishlansky M, Geary P, O'Brien P, Wong R B 1995 *Societies and Cultures in World History*. Harper Collins, New York

Landes D S 1998 *The Wealth and Poverty of Nations: Why Some are so Rich and Some so Poor*. Little Brown, London

Lewis B 1995 *The Middle East. 2000 years of History from the Rise of Christianity to the Present Day*. Weidenfeld and Nicholson, London

Livi Bacci M 1992 *A Concise History of World Population*. Blackwell, Oxford, UK

McNeill W 1991 *The Rise of the West: A History of the Human Community*. University of Chicago Press, Chicago, IL

Mazlish B, Buultjens R (eds.) 1993 *Conceptualizing Global History*. Westview Press, Boulder, CO

Mokyr J 1990 *The Lever of Riches: Technological Creativity and Economic Progress*. Oxford University Press, Oxford, UK

O'Brien P K (ed.) 1999 *Philips Atlas of World History*. George Philip Limited, London

Offen K, Roach R R, Rendall J (eds.) 1991 *Writing Women's History: International Perspectives*. Macmillan, London

Pomeranz K 2000 *The Great Divergence: Europe, China and the Making of the Modern World Economy*. Princeton University Press, Princeton, NJ

Sanderson S K 1995 *Social Transformations: A General Theory of Historical Development*. Blackwell, Oxford, UK

Stavrianos L S 1989 *Lifelines from our Past: A New World History*. Pantheon Books, New York

Smil V 1994 *Energy in World History*. Westview Press, Boulder, CO

Weber M 1961 *General Economic History*. Collier, New York

Wong R B 1997 *China Transformed: Historical Change and the Limits of European Experience*. Cornell University Press, Ithaca, NY

Worster D (ed.) 1998 *Perspectives on Modern Environmental History*. Harvard University Press, Cambridge, MA

P. K. O'Brien

Global Population Trends

During their evolution of tens of thousands of years, humans have progressively, although not without setbacks, learned how to better feed themselves, make clothing, provide shelter, cure and prevent diseases, improve their health, and modify their fertility. People have learned how to interact in societies which were at first based on gathering, hunting, and fishing. Many societies then experienced the agricultural revolution combined with the domestication of animals. The evolving increases in labor productivity and scientific advancements generated the industrial revolution and urbanization, the foundations of contemporary soci-

eties. In the process, populations have grown from very small numbers to the six billion global population reached in 1999. Demographically speaking, for many centuries and millennia changes were erratic and slow until several centuries ago mortality started to decline, life expectancy increased, somewhat later fertility commenced its decline, and populations around the globe began to grow at increasing rates. The more pronounced and rapid changes have occurred in the most recent 100–200 years, although the roots of change reach much further into the past.

1. *The World's Population*

In 1695 Gregory King (1973) was among the first scholars who attempted to estimate the size of the world population. His estimates were within a range of 495–700 million—preferring 630 million as the base for his projections—in line with what scholars engaged in similar exercises since then consider as realistic (Table 1). In 1999, the size of the world population was estimated to have reached six billion inhabitants. 'Estimated' because even at the end of the twentieth century uncertainties about the exact size of the world population persist, as there are numerous countries which lack a recent reliable census, and which have inadequate population registration systems or none at all.

Through most of human history the increase in the world population was slow and punctuated by periods of decline. The technologies of food production, as well as the ways and means of health protection, were relatively primitive and humans were thus quite significantly exposed to the vagaries of nature. A notorious case was the epidemic of the plague or Black Death which in the fourteenth century spread from Asia to Europe and Africa and is believed to have decimated Europe's population by one-third.

The growth of the world population gained momentum in the seventeenth to nineteenth centuries. However, the critical acceleration occurred after World War II. During the second half of the twentieth century the global growth rate has been close to 2 percent per year. Improved personal health behavior and public health measures as well as the application of curative and preventive medicine which had been attained and practiced in the industrial countries earlier were rapidly spreading in the developing world (see *Mortality: The Great Historical Decline*). It was during this period that the peak and the turning point in global population growth was reached. Since its climax in the1960s with an annual rate of 2.0 percent, this rate has declined to 1.4 percent in the 1990s. In many parts of the world, in developed and developing countries, fertility has been declining rapidly (see *Fertility Transition: Cultural Explanations* and *Demographic Transition, Second*). It is worth noting that the absolute annual additions continued to increase through the end of the twentieth century. They had also reached their peak and were commencing their decline. Around the year 1900 the annual absolute global population increase was about eight million. In the year 2000 it was around 80 million.

Through the end of the twentieth century the number of years required to add an additional one billion population has been decreasing. The first billion was reached around 1804, the second in 1927 (123 years later), the third in 1960 (33 years), the fourth in 1974 (14 years), the fifth in 1987 (13 years),

Table 1
Estimates of world population, 7000 BC–2000 AD

Date	Population in millions	Average annual increase in % since preceding date	Absolute annual increase in millions since preceding date
BC			
7000–6000	5–10	n.a.	n.a.
AD			
1	200–400	0.0	0.05
1650	470–545	0.0	0.13
1750	629–961	0.4	2.9
1800	813–1,125	0.4	3.5
1850	1,128–1,402	0.5	5.9
1900	1,550–1,762	0.5	7.8
1950	2,521	0.8	17.3
1960	3,022	1.8	50.1
1970	3,696	2.0	67.4
1980	4,440	1.8	74.4
1990	5,266	1.7	82.6
2000	6,055	1.4	78.9

Source: United Nations 1973, United Nations 1999

Table 2
Composition of the world population by major region, in percent, 1750–2000

Region	Year					
	1750	1800	1850	1900	1950	2000
World, in millions	791	969	1,265	1,656	2,521	6,055
World, in percent	100	100	100	100	100	100
Africa	13	11	9	8	9	13
Asia	63	66	64	57	56	61
China	*27*	n.a.	n.a.	*26*	*22*	*21*
India/Pakistan/Bangladesh	*23*	n.a.	n.a.	*18*	*17*	21
Remainder	*13*	n.a.	n.a.	*13*	*16*	*18*
Europe	21	21	22	25	22	12
Latin America and Caribbean	2	2	3	4	7	9
North America	0	1	2	5	7	5
Oceania	0	0	0	0	1	0
More developed countries	25	25	27	33	32	20
Less developed countries	75	75	73	67	68	80

Source: Durand 1974, United Nations 1973, 1999

and the sixth in 1999 (12 years). Indications are that this number of years will start to increase in the twenty-first century (see *Population Forecasts*).

The distribution of the world's population has been changing over time. For several centuries large populations have been living in China and the Indian subcontinent. In the year 2000 each had about 1.3 billion inhabitants constituting 21 percent of the world population (Table 2).

In the course of the nineteenth century it was the population of the industrial countries in Europe and North America which grew more rapidly than in other regions. Its proportion increased from a quarter of the world's population to one-third between the beginning and the end of the nineteenth century. In the twentieth century, population growth in the developed countries slowed down. Subsequently, during the second half of the century, rapidly declining mortality in Third World countries combined with a fertility decline led to rapid population growth so that their proportion of the world's population increased from two-thirds around 1950 to 80 percent around the year 2000.

2. *Mortality*

As best as can be estimated, the average human lifespan in prehistoric and during most of historic times was rather short. The scarce archeological evidence available suggests that most adult deaths occurred between the ages of 20 and 30. In the Iron Age in France average life expectancy was presumably 10–12 years. A life table for ancient Greece, prepared from burial inscriptions, indicates an average life expectancy of 30 years around 400 BC. Life expectancy in Geneva was estimated at about 21 years in the period 1561–1600 and about 26 years in 1601–1700. The latter and other estimates based on fragmentary data in various European countries from the thirteenth to the eighteenth centuries indicate that mortality was very unstable and variable with life expectancy between 20 and 40 years (United Nations 1973). A relatively comprehensive estimate covering Northern and Western Europe, Northern America, and Oceania (The 'West') for the middle of the nineteenth century still resulted in an average life expectancy of 40 years. From thereon health conditions improved and life expectancies grew quite steadily in these countries. Elsewhere in Europe and the former Russian empire mortality started to decline later. In Southern Europe, in Italy, for instance, in 1871–80 life expectancy at birth did not exceed 35 years and a steady fall of mortality started after 1875. Japan was another country where mortality was low quite early. Its crude death rate at the end of the nineteenth century was comparable with many European countries.

Malnutrition, poor hygiene, parasitic infections, and disease were, and in many parts of the world still are, the principal causes for high mortality. It took many centuries of gradual progress to have a visible effect on survival. Changes in diet, personal and public hygiene, reinforced by advances in science and medicine, generated dramatic increases in life expectancy. During the past 100–150 years, life expectancy has essentially doubled from 30–40 years to 70–80 years in those societies where most people at the beginning of the twenty-first century enjoy a decent standard of living.

In the Middle Ages most people lived in primitive living quarters with their domesticated animals and their excrements, with an open fire fueled by wood and manure in the middle of the floor and a hole in the roof to let out smoke and cinders, without any toilets and bathrooms, without soap, windows, furniture, dishes,

Table 3
Mortality measures, world and selected countries, 1950–2000

	Life expectancy at birth in years							Infant deaths in the first year of life per 1,000 live births						
Period	**World**	Sweden	Japan	Hungary	Brazil	India	Uganda	**World**	Sweden	Japan	Hungary	Brazil	India	Uganda
1950–5	**47**	72	64	64	51	39	40	**155**	20	51	71	135	190	160
1960–5	**52**	74	69	69	56	46	44	**117**	15	25	44	109	157	125
1970–5	**58**	75	73	69	60	50	47	**93**	10	12	34	91	132	116
1980–5	**61**	76	77	69	63	55	46	**78**	7	7	20	64	106	118
1990–5	**64**	78	80	69	66	60	37	**62**	6	4	13	47	78	118
2000 estimate	**66**	79	80	72	67	63	42	**55**	5	4	10	40	69	101

Source: United Nations 1999

and cutlery. Individual societies around the world were better equipped in some respects, for instance, the Chinese employed coal as a fuel, and soap was commonly used in the Islamic world. Frequent failures of harvests and major epidemics were commonplace everywhere. It took many centuries for standards of living and hygiene to improve and thus provide favorable conditions for better health. Residential chimneys in Europe became common around 1200 and it was Marco Polo who discovered in 1306 that the Chinese used coal to heat their homes. Soap was manufactured for the first time in London presumably in 1259. Frequent bathing, however, was considered unhealthy and as late as in the sixteenth century Queen Elizabeth I was considered eccentric for her frequent bathing—once a month. It was not until the Industrial Revolution of the eighteenth and nineteenth centuries that the way people lived changed considerably and irrevocably. Significant increases in productivity, including agriculture, provided for the possibility of improved nutrition, shelter, and living conditions in general. In combination with the gradually increasing application of personal hygiene, the installation of indoor plumbing and introduction of sewage systems, the accumulation of scientific and medical knowledge as well as the spread of health services networks, this all led to the gradual improvements in health and the decline of mortality.

The profound long-term changes in patterns of death, disease, and disability have been labeled the *epidemiological transition* (see Omran 1971 and *Mortality, Epidemiological, and Health Transitions*). It has three courses of change: a shift from infectious and parasitic to chronic and degenerative diseases, a shift in the burden of death and disease from the younger to the older population, and a shift from a health picture dominated by mortality to one dominated by morbidity. In general, these directions of change occurred in the developed countries before the middle of the twentieth century. In developing countries, during the latter part of the twentieth century, often the increase in chronic and degenerative diseases coexisted with relatively high levels of infectious diseases (Bobadilla et al. 1993). Most of the morbidity of contemporary advanced societies is preventable because in large part it is due to ills and imperfections of social and individual behavior. It is due to undesirable side effects of industrialization, urbanization, and mass consumption such as atmospheric pollution, environmental degradation, accidents, mental tension and distress, consumption of tobacco, alcohol and drugs, unhealthy diets and obesity (Jozan 1998). Efforts to control and reduce this morbidity are underway, but results are mixed. The complexities are numerous. Consequently, there are large differences regarding success and failure in lowering mortality and morbidity between countries, within countries among regions and social strata.

The progress and problems of mortality and morbidity change in the past 50 years are illustrated by data in Table 3. In most countries around the world major improvements were achieved during the second half of the twentieth century. Nonetheless, there were two types of exceptions. Hungary exemplifies the stagnation in mortality decline of the formerly socialist countries of central and eastern Europe and Uganda represents the countries of sub-Saharan Africa where mortality declines were reversed mainly as a result of the HIV/AIDS epidemic. At the same time, significant improvements in the health conditions of infants have been achieved around the globe, but major differentials persisted through the end of the twentieth century (United Nations 1999).

3. *Fertility*

In the long run, additions of births have to at least match attrition by death if a society is to survive and perpetuate. That is what has been happening throughout humanity's history and before. Fertility could not decline from traditional levels prior to any significant decline of mortality without causing the extinction of human populations.

As long as mortality was high and a large proportion of infants, children, and adolescents died before entering their reproductive ages, women would bear on average 5–8 children. Such high fertility is persisting at present in a number of countries with high mortality, especially in sub-Saharan Africa. There are, however, only few such countries remaining. For the most part, fertility has either already declined to very low levels or countries are on their way between high and low fertility.

The biological maximum a woman can bear during her childbearing period is between 13 and 17 children. As a rule, this does not happen as there are numerous deliberate/conscious as well as unintended/unconscious mechanisms limiting births. Until the nineteenth century in the advanced countries, and until only several decades ago elsewhere, breast-feeding—occasionally combined with taboos in intercourse during lactation—was the decisive factor bringing down the number of children born to between 5 and 8. In premodern times almost exclusive breast-feeding tended to be the rule for infants and for many children at times until they reached their second or even third birthday. For most women this practice prevented a further pregnancy. An extreme example are the San of the Kalahari Desert who were not subject to any taboo and yet they achieved four-year birth intervals through the contraceptive effect of frequent (four times an hour) breast-feeds. The physiological process taking place is such that when the baby suckles on the nipple, nervous impulses pass from the breast to the brain to inhibit the release of hormones responsible for ovulation from the pituitary gland (see Potts and Short 1999 and *Fertility Control: Prevalence and Consequences of Breastfeeding*).

Women breastfeed as it is the most natural and convenient way to feed infants, at the same time the conception preventing effect was apparently known for millennia and in some societies widely realized. Aristotle stated in the *Generation of Animals* that 'while women are suckling children, menstruation does not occur according to nature, nor do they conceive.' In 1792, Mary Wollstonecraft wrote in the *Vindication of the Rights of Women*: 'Nature has so wisely ordered things that did women suckle their children, they would preserve their own health, and there would be such an interval between the birth of each child, that we should seldom see a houseful of babies.' Lactation has other critical salutary effects. Maternal milk is the most beneficial source of nutrition finely adjusted to the needs of the offspring. It is also a living secretion, rich in antibodies that protect against disease. Therefore, particularly in developing countries, breastfed babies tend to be healthier than their bottle-fed peers.

Other critical factors reducing fertility are delayed cohabitation of partners/marriage, the use of contraceptives, and the prevalence of induced abortions. Prior to the latter part of the nineteenth century, these were known, however, they were not widely practiced and therefore they were not decisive in modifying fertility. Western European marriage patterns were the exception. Hajnal (1965) observed that since the eighteenth century fertility had been relatively low and at times declining in countries west of a line between St. Petersburg and Trieste precisely because the age of marriage was high. Later, the considerable fertility decline in western countries in the late nineteenth and early twentieth centuries was apparently brought about by the application of relatively unsafe induced abortions as well as the use of traditional methods of contraception, such as coitus interruptus (Santow 1995). There is, however, scarce statistical hard evidence to support this statement.

During the second half of the twentieth century safe induced abortions became increasingly available for a large proportion of the world population and unsafe abortions were widely practiced in Latin America and Africa (see *Infanticide and Induced Abortion*). A 'contraceptive revolution' was underway throughout the developed countries and increasingly in the developing ones as well. In most countries the use of contraceptives increased, modern contraceptives such as the IUDs and oral contraceptives found extensive application, and female as well as male contraceptive sterilization was widely adopted, although differences between countries were considerable. The prevalence of contraceptive use rose to between 70 and 80 percent of married women in the developed countries, and in countries such as Taiwan and South Korea, the use of contraceptives increased from 20 to 80 percent within a period of about three decades (see Frejka and Ross 2001 and *Family Planning Programs: Development and Outcomes*; *Eyewitness Memory: Psychological Aspects*; *Fertility Transition: Economic Explanations*; *Fertility Transition: China*; *Fertility Transition: East Asia*; *Fertility Transition: Southeast Asia*; *Fertility Transition: South Asia*; *Fertility Transition: Middle East and North Africa*; *Fertility Transition*: *Latin America and the Caribbean*).

Relatively late marriage, the use of contraception, and induced abortion are the *means* of modifying fertility (see *Fertility: Proximate Determinants*). The motivations and decisions to regulate fertility, subconscious or conscious, have numerous economic, social, cultural, and political roots. The following developments are critical. With economic progress children have evolved from being an asset to their parents to become a costly investment. Growing proportions of young women and men remained in school or found employment outside the household and therefore entered marital or consensual unions later than their parents. The family, particularly the younger generation, used to take care of older members of the household; increasingly older people have to take care of themselves possibly assisted by the state. As recently as in the nineteenth century women everywhere were oppressed and disadvantaged, nowh-

Table 4
Fertility measures, world and selected countries, 1950–2000

	Crude birth rate							Total fertility rate						
Period	**World**	Sweden	Japan	Hungary	Brazil	India	Uganda	**World**	Sweden	Japan	Hungary	Brazil	India	Uganda
1950–5	**37.3**	15.5	23.7	21.1	44.0	44.1	51.4	**4.99**	2.21	2.75	2.73	6.15	5.97	6.90
1960–5	**35.2**	14.7	17.2	13.6	42.1	42.0	48.7	**4.95**	2.33	2.01	1.82	6.15	5.81	6.90
1970–5	**30.9**	13.6	19.2	15.7	33.7	38.2	50.3	**4.48**	1.89	2.07	2.09	4.72	5.43	6.90
1980–5	**27.4**	11.4	12.7	12.5	29.6	33.8	49.9	**3.58**	1.64	1.76	1.81	3.63	4.47	7.00
1990–5	**23.9**	13.5	9.7	11.8	22.2	28.7	50.8	**2.93**	2.01	1.49	1.73	2.51	3.56	7.10
2000 estimate	**21.4**	10.0	10.0	9.6	19.8	24.1	50.2	**2.64**	1.57	1.45	1.33	2.21	2.93	6.92

Source: United Nations 1999

Table 5
Population growth and percentage aged 65 and over, world and selected countries, 1950–2000

	Population size, 1950 = 100							Percentage aged 65 and over						
Year	**World**	Sweden	Japan	Hungary	Brazil	India	Uganda	**World**	Sweden	Japan	Hungary	Brazil	India	Uganda
1950	**100**	100	100	100	100	100	100	**5.2**	10.3	4.9	7.3	3.0	3.3	3.0
1960	**120**	107	113	107	135	124	138	**5.3**	12.0	5.7	9.0	3.3	3.4	2.6
1970	**147**	115	125	111	178	155	206	**5.5**	13.7	7.1	11.6	3.7	3.7	2.6
1980	**176**	118	140	115	225	193	276	**5.9**	16.3	9.0	13.4	4.1	4.1	2.5
1990	**209**	122	148	111	274	238	346	**6.2**	17.8	12.0	13.3	4.3	4.3	2.5
2000 estimate	**240**	127	152	107	315	283	457	**6.9**	17.4	17.1	14.7	5.1	5.0	2.1

Source: United Nations 1999

ere did they have the right to vote, for instance. In various parts of the world many women are still battered, sequestered, mutilated, or exploited, but a basic trend supporting their empowerment is in progress. Thus, increases in women's education and employment, their earning potential in the marketplace, and growing equality vis-à-vis laws have enabled women to partake in decisions regarding fertility behavior.

Population policies—deliberate interventions to influence a population development—have frequently been formulated and employed by governments or international institutions (see *Population Policies, Politics of*). With respect to fertility, governments perceiving it to be too low may devise policies which they believe will raise fertility; those regarding high fertility as a burden implement measures to diminish it. Berelson (1977), for instance, demonstrated that there are numerous policies which may conceivably reduce fertility. He clarified how policy makers taking local circumstances into account should and usually do consider whether the policies they wish to employ are *acceptable* to the political elite and to the public, whether they are *feasible* with given resources and administrative infrastructures, and how *effective* they will be.

Particular social, economic, political, cultural, and religious circumstances sometimes in combination with population policies shape people's fertility desires, motivations, and behavior which is expressed in their marriage patterns, breast-feeding practices, contraceptive use, and prevalence of induced abortion and thus result in a certain level and trends of fertility.

The unprecedented changes in fertility which occurred during the second half of the twentieth century are illustrated in Table 4. The total fertility rate, a measure which can be interpreted as the average number of children borne by a woman, on a worldwide scale declined from about five children in the early 1950s to just over two-and-a-half in the year 2000. In the developed countries, represented by Sweden, Japan, and Hungary, fertility declined from an already low level in the early 1950s to significantly below the replacement level of about 2.1 children per woman. The decline was not linear, some fluctuations did occur, however, at the end of the twentieth century not much less than half the world's population lives in countries with below replacement fertility (see Frejka and Ross 2001). In developing countries around 1950, fertility was high, between six and seven children per woman. In a significant number of countries it declined close to the replacement level, such as in Brazil. In

others like in India fertility was considerably reduced. In some countries, however, no discernible change took place. Uganda is typical of sub-Saharan countries with stable high fertility, although there are some where a decline has commenced.

Even though there has been a considerable fertility decline in most developing countries in the past several decades, continued population growth can be expected and will be forthcoming during the first half of the twenty-first century. Past high fertility has produced large cohorts of young people who are presently in the childbearing stage of their life cycle or who will enter their childbearing years. Even if these generations will have low fertility, the total numbers added to the populations of the developing countries are almost sure to be much larger than the number of people dying (see *Population Dynamics: Momentum of Population Growth* and *Population Forecasts*).

4. Concerns

Table 5 illustrates basic demographic developments which are frequently causes for concern among policy makers, scientists, and the general public. The world population has grown more rapidly during the second half of the twentieth century than ever before. Rapid population growth, even though at a moderating pace, will continue for some decades. Extraordinarily fast population growth has occurred in developing countries. Brazil's and India's populations, for instance, have grown threefold, Uganda's 4.5 times. Such growth creates obvious dilemmas and pressures for governments, the private sector, and international organizations striving to raise educational levels, improve health conditions, broaden employment opportunities, and elevate living standards in general. Population growth coupled with increasing consumption in developed and developing countries increases dangers of specific resource depletions, environmental degradation, selective plant and animal extinction, which ultimately could pose problems for and harm the human species. Concern is also voiced that rapid population growth can be a factor exacerbating poverty and political instability.

Hungarian population change exemplifies a shrinking population. There were more than 10 such nations in Europe in the late 1990s and thus concerns for a diminishing labor force, declining international stature, the potential of dying out as a nation, and others have surfaced.

Rapidly declining and low fertility gives rise to changes in the age structure and together with increasing longevity are the root causes of population aging. In the late 1990s in many developed countries 15 percent or more of populations were of age 65 and over. The numbers and proportions of the elderly have been increasing and there is no doubt that they are going to continue to do so. This is one of the major concerns of governments in the developed countries because considerable state expenditures go toward social security benefits of the elderly as well as their health care. As fertility is declining rapidly in many developing countries, population aging is becoming an important social and political issue.

Bibliography

Berelson B 1977 Paths to fertility reduction: The 'policy cube.' *Family Planning Perspectives* **9**(5): 214–19

Bobadilla J L, Frenk J, Lozano R, Frejka T, Stern C 1993 The epidemiologic transition and health priorities. In: Jamison D T, Mosley W H, Measham A R, Bobadilla J L (eds.) *Disease Control Priorities in Developing Countries*. Oxford Medical Publications, Oxford, UK

Chesnais J-C 1992 *The Demographic Transition: Stages, Patterns, and Economic Implications*. Clarendon Press, Oxford, UK

Coale A J 1974 The history of human population. *Scientific American* **231**(3): 40–51

Cohen J E 1995 *How Many People Can the Earth Support?* W W Norton, New York

Durand J D 1974 *Historical Estimates of World Population: An Evaluation*. Population Studies Center, University of Pennsylvania, PA

Frejka T J, Ross A 2001 Paths to subreplacement fertility: The empirical evidence. In: Bulatao RA, Casterline JB (eds.) *The Global Fertility Transition*, Supplement to *Population and Development Review* **27**

Jozan P 1998 Some features of mortality in the Member States of the ECE. In: United Nations Economic Commission for Europe, *Background Papers of the Regional Population Meeting, December 1998* (unpublished)

King G 1973 17th century manuscript book of Gregory King. In: Laslett P (ed.) *The Earliest Classics: Graunt and King*. Gregg, Farnborough, UK

Omran A 1971 The epidemiologic transition: A theory of the epidemiology of population change. *Milbank Memorial Fund Quarterly* **49**: 509–38

Potts M, Short R 1999 *Ever Since Adam and Eve: The Evolution of Human Sexuality*. Cambridge University Press, Cambridge, UK

Santow G 1995 Coitus interruptus and the control of natural fertility. *Population Studies* **49**(1): 19–43

United Nations 1973 *The Determinants and Consequences of Population Trends*. United Nations, New York, Vol. 1

United Nations 1999 *World Population Prospects: The 1998 Revision*. United Nations, New York

Wollstonecraft M 1985 [1792] *Vindication of the Rights of Woman*. Penguin, London

T. Frejka

Global Positioning System

From the beginning of time man has striven to answer three fundamental questions: where am I now, where I am going next, and what is the best route from here to there? The observation of the sun and stars, and then later the use of the compass, the sextant, navigational

maps, and charts of the heavens, have all been essential aids and instruments in this never-ending quest. GPS is the most recent contribution to this search.

1. Origins and Instances

Over the years the need to answer these questions has been the driving force behind the development of many civil and military applications. During World War Two, one of the nations in combat realized that, even if they managed to build warships relatively rapidly, it was much more difficult to find suitable commanders for the ships. Such positions required candidates with a high level of experience, especially for mapping the ship's course. From this dilemma LORAN (Long-range Radio Navigation) was born, a system essentially for use at sea, based on the reception and phase difference of radio waves transmitted from two antennae. These two stations transmitted a continual flow of coordinates to navigators equipped with the correct passive antenna, which they then used to pinpoint their exact location on the earth's surface. This facilitated the commander's decision-making enormously, as he knew the ship's position at any time and, by using additional electronic aids or by simply mapping the points already crossed on a nautical chart, he also knew exactly where the ship was headed. The GPS (Global Positioning System) was the natural evolution of LORAN.

At any given moment and point in the area covered by the system, GPS is capable of providing the three-dimensional coordinates (x, y, z, or latitude, longitude and height). Rapid progress in electronic component design, has miniaturized GPS equipment to the extent that even watches and telephones can contain GPS and can provide the wearer with their exact position, the route they followed, and where they have to go.

On certain commercial flights a screen displays a schematized map of the territory below and, using GPS, a pointer indicates the current position of the airplane, the flight path, and, amongst other information, how much time remains before landing. Passengers draw comfort from knowing where they are and when they will arrive. From the cabin they may be able to use the features mapped on the screen to recognize the physical characteristics of the territory below.

Aside from the passive and temporary use of GPS to determine position, route and destination, the system is also important for activities with long-lasting and consequential effects. The most important example is the use of GPS in cartographic or topographic research, and, by extension, to many forms of inquiry in social science. With GPS, the earth's surface becomes an enormous digitizer, and any point can be identified instantly by its three-dimensional coordinates and archived in a suitable information system, potentially increasing the intensity of sampling events and objects in space and time by several degrees of magnitude. Such measurements affect research and practice in archaeology, economics, geography, planning, and sociology, as well as in the natural sciences. By determining precise coordinates of points on land or on the surface of water, GPS makes it possible to update maps by locating physical and social features from which GPS gathers data instantaneously and/or detecting changes that have occurred in reality but have yet to be recorded (e.g., new construction, physical changes in the environment, transient events, or those in constant transformation such as people, vehicles, coastlines, woods, glaciers and so on).

2. The General Principles of GPS Functioning

GPS receivers use radio waves to determine positional coordinates. The radio waves are emitted by a constellation of orbiting satellites, which function as a reference system for GPS. The constellation consists of 24 satellites, which are located in six orbits, inclined at an angle of approximately 50 degrees, at a height of around 20,000 km. Known as NAVSTAR, the constellation is the property of the US government and is controlled and managed by the Department of Defense. The orbits and the distance between the satellites are organized in such a way that, ideally, every point on the earth's surface can be seen by at least five satellites at the same time, although this is not always this case in practice. The GPS carries out the triangulation of three satellites visible at the same time above the horizon, measuring the distance that separates it from each of these by calculating the time between each satellite (transmitter) to the GPS itself (receiver) using precise time signals emitted by the satellites. The triangulation allows the GPS to determine the latitude and longitude, though a fourth satellite is also necessary to determine a third coordinate giving the height at which the GPS receiver is located. To measure the time required by the signal to travel the distance from each satellite to the receiver, it is essential to use extremely precise equipment to transmit the signal, to calculate the time lapse, and also to pinpoint with extreme accuracy the position of each satellite in space. To this end each satellite is equipped with an atomic clock and its position is constantly monitored and regulated by a series of earth stations.

It is essential to consider the problem of errors incurred by external agents when dealing with instruments as sophisticated as the GPS. To determine the exact time the signal takes to reach the GPS receiver, it is also necessary to understand all the modifications and distortion the signal may undergo while passing through the ionosphere and the atmosphere. The causes of error and/or the malfunctioning of the system are varied, and can range from atmospheric conditions to the position of the satellite constellation in the sky and their visibility from the receiver. The location of the GPS receiver itself may cause problems.

For example the receiver may be unable to detect a satellite if it is located under a canopy of thick vegetation or in a city canyon, where its horizon may be modified by the height of surrounding buildings.

Those managing the system itself may introduce certain deliberate errors. GPS, for example, has been subject to SA (Selective Availability), a randomized error deliberately incurred in the time signal from the satellites to degrade the positional accuracy of the receiver to the order of 10 to 100 meters. Acting on a declaration made in 1996, on May 1, 2000, the President of the USA announced the end to the use of SA in GPS, with the immediate effect of increasing and encouraging the use of GPS for civil and commercial purposes throughout the world. In fact, the demand for GPS precision was such that, as soon as the system was recognized by civilian users as being of major importance, they also wanted instruments of reduced size and low cost to have the same excellent functionality found in the military. While knowing one's position with reasonable approximation (within tens of meters, or at least under 100 m) is satisfactory for most marine or aerial navigation purposes, this level of approximation is not acceptable in other contexts. In an urban situation for example, the user may wish to know what side of the road he is on or which traffic lane to be in, and in the development of technical charts, cadastres, and engineering plans, the precise coordinates of the point on the ground are needed in order to represent it on a topographic map. In the latter case it is necessary to reduce approximation to within a few decimeters or centimeters. In reality, GPS is capable of reaching precision of less than a millimeter, if certain components of the technology are exploited and if the period of observation can be maintained for a sufficient length of time.

The precision of GPS measurements is also influenced by factors such as whether the station is mobile or fixed (at least for the time being) and whether the receiving station functions alone or in connection with other stations. The ability for a station to operate in conjunction with at least one other fixed station provides optimal precision. This technique is called Differential GPS and it is based on the registration of signals transmitted by the satellites and simultaneously picked up by a fixed or base station (Master), of which the precise geographical coordinates are known, and by a mobile station (Rover), located anywhere within a range of around 150 km from the Master station. Post processing the data eliminates the errors in reception by the Rover. The precision of the GPS in the Rover station is able to satisfy all the requirements of civil use from cartographic applications, to support systems for the management and control of transport fleets, and personal mobility.

The late 1990s have seen an increase in the demand for differential correction and it is now easy to find private and public companies that broadcast the data necessary for differential correction in their zone over the radio waves. Certain GPS models are already equipped to receive such radio signals. In the future GPS will continue to improve, in particular after the announcement by the US military controller of the system, with the launching of new satellites, of which 18 are planned in the near future, and with the refinement of certain technical characteristics. The future will also probably see full integration with the Russian-developed GNSS system, which is already integrated with GPS in certain commercial devices.

3. *The Relationship Between GPS and Information Systems*

The most relevant aspect of GPS to the information society in 2001 is the ability to provide the geographical coordinates of any object on the earth's surface and to increase our knowledge of the object's characteristics with ancillary data, collected in real time and directly from the field, at very low cost. Without doubt, the geographical component is an essential part of current databases and is generally deduced from a classification of objects by name (city, road, house number, etc.) or is represented by the cartographic coordinates of the single object gathered from maps (for example the centroid of a manhole cover, telephone or electricity pole, the axis of a water or gas pipe). In each case the geographical component, obtained by traditional means, is approximate, expensive, or unreliable in relation to both time and space. Time, because things can change from one moment to the next (an electricity pole can be moved relatively rapidly in reality but not so rapidly on the charts representing it). Space, because an element located in a certain territory undergoes continual transformation, the shores of a lake or a coastline for example can be modified by environmental factors.

Conferring a geographical address on each item of data allows for the application of the analytical, inquiry, and management functions of a geographic information system (GIS). Real time availability of the geographic address means that the spatial analysis can be carried out instantaneously in the field. Typical spatial analysis functions include: investigating relationships of proximity to, or distance from, the object, superimposition of information layers, the search for the shortest or any route that satisfies specific conditions, identification of spatial environments that group objects with specified characteristics, monitoring transformations of the object, and so on. Typical functions of inquiry analysis are those that allow us to understand the information linked to an object by selecting that object from a spatial database rather than selecting an object from the alphanumeric database and obtaining thematic charts *ad hoc*. Typi-

cal functions of information system management are those that permit us to continually update the constituent elements by refreshing the data and integrating the geographic database whenever necessary.

The continual information flux is one of the major problems in information society and updating coordinates in three-dimensional space is even more challenging. Connecting GPS to other techniques helps solve this problem. For example, we can consider digital cartography, necessary for a decision-making process in a territorial or urban context. This is archived in appropriate databases and territorial information systems use it to undertake analysis and to aid in making the above-mentioned decisions. But what happens when digital cartography no longer reflects reality? A city or region can sometimes change very rapidly and if the digital map is not updated, then we run the risk of the information system elaborating data that is no longer. Invalid information may be merely inconvenient or lead to serious social or legal complications. GPS helps greatly, but it is not only maps that need continual updating but also the geographic address of data that is particularly useful in social science. Among these are census tracts, political districts, administrative units, pollution zones, and epidemiological information, or research concerning traffic and movement. More controversially we may monitor individuals.

4. *Examples of the Use of GPS and its Social Implications*

The development of GPS and its diffusion is a step towards having geographic address (x, y, z coordinates in space, or longitude, latitude and height) and exact time as a component of any type of information. GPS equipment is already widely used by utilities companies for the identification of the components in their networks (poles, pylons, wells, technical components, etc.), while among the earlier applications we should remember the use of GPS for the positioning and identification of submarine oil wells from offshore oil platforms. Additionally anyone, at any moment in time, can visualize the location or destination of any vehicle transmitting its coordinates to be collected in real time by GPS. For example when mounted on a public bus, a GPS can supply immediate information on the vehicle's location, the journey time and anything else that is necessary to satisfy the needs of the driver, passenger, or manager alike. For some time now, insurance companies have been using this system to reduce the incidence of car thefts. Car manufacturers are now equipping cars across the range with navigators: electronic navigation systems containing a digital map of the area in which the car is travelling and on which the positioning and movement of the vehicle itself is traced, thanks to the continual tracking of its location through GPS. Integrated with other systems, GPS permits a vehicle to proceed automatically, without the aid of the driver or pilot. Systems already exist in marine and aeronautical applications that are capable of functioning as autonomous autopilots. Exact details of military applications are not public knowledge, but it is easy to guess that the same basic technology is used with more sophisticated developments. An internal GPS in a cellular phone provides the user with their exact coordinates, and can be used to visualize the location of similarly equipped phones. Such devices are in regular use at large conventions to facilitate (or avoid) meetings. The same device alerts emergency services to incidents on lonely stretches of unfamiliar road or, worn as clothing, GPS may guide the sight-impaired or proof children against becoming lost. All fields that use geographic coordinates may derive benefits from the use of GPS.

There are two principal problems linked to the use of GPS: a reduction in the capability of orientation and the limitation of privacy.

The problem of the reduction in the capacity to orientate oneself is due in part to the use of digital coordinates, rather than analog information such as: go to the square, turn right at the newsstand, go straight on to the white building, etc. With GPS a person can supply a visitor with the coordinates of their position by simply pressing a button; the visitor receives the coordinates and transfers them to his autopilot or navigation system, and the vehicle navigator directs the visitor to the location they require. Nothing will remain in the visitor's memory, unless he has paid attention to the features of his route and not simply paid attention to the screen. It is a question of personal choice whether to learn to navigate in one's environment or to rely wholly on digital assistants.

The privacy problem is a crucial aspect of the use of GPS in society. The power of knowing in a given moment the location of someone equipped with a GPS can be a source of enormous security in certain cases; for example for those suffering from an accident or particular illness who may require treatment as quickly as possible. But knowing someone's movements is potentially a great intrusion on personal privacy. For example, with GPS and GIS we could discover the shop, office or facility someone is currently visiting, or recreate their activities during the past 24 hours or longer. The surveillant society is the flip-side of the information society. Separating valuable monitoring from unwarranted intrusion into personal privacy, and deciding who has the power to distinguish between them, poses a major problem for society.

5. *Summary*

Cartographers, engineers, planners, archaeologists, and geographers have drawn enormous advantages from the active use of GPS and will continue to enjoy

the benefits of a new technology that allows for rapid and much less costly collection and updating of geographic information. The integration of engineering designs realized with CAD, digital mapping, and GPS, introduces the possibility of devices capable of elaborating designs for civil engineering and proceeding through interactive surveys, to the construction of roads, buildings, etc. GPS also integrates well with the analysis of data that has been collected via satellite for the observation of the earth, in that it allows for the exact identification of the coordinates of specific features on the satellite image. Social applications range across all walks of life. Inevitably, as with all technological innovations, its use should be evaluated critically, and the limits of its appropriate use determined.

See also: Applied Geography; Cartography; Geographic Information Systems; Geographic Information Systems: Critical Approaches; Location: Absolute/Relative; Navigation; Remote Sensing and Geographic Information Systems Analysis; Spatial Analysis in Geography; Spatial Data; Time–Space in Geography; Transportation Geography; Transportation Planning; Urban Planning: Methods and Technologies

Bibliography

http://www.trimble.com/gps/
http://giswww.pok.ibm.com/gps/
http://www.navtechgps.com/
http://www.colorado.edu/geography/gcraft/notes/gps/gps-f.html

M. Salvemini

Globality

In the English language the word globality is of relatively recent coinage, although the noun 'globe' has been used rather a lot since the sixteenth century along with the adjective 'global.' The concept of globality has come to be employed widely since the early 1980s (Robertson 1984), along with the concept of globalization. Both globalization and globality became significant terms during the 1980s, in the wake of expanding scholarly use of the adjective global, in primary reference to the dual aspects of the intensifying connectivity (Tomlinson 2000) and of the spreading and deepening consciousness of the world-as-a whole (Robertson 1992, 2001a). This development within intellectual discourse has proceeded in tandem with the ever more conspicuous invocation of the adjectival form in promotional and advertising contexts. Both these trends were largely sparked by the phrase 'global consciousness,' meaning consciousness of the world as a whole, in particular, but not sole, reference to environmental or ecological issues. Note must also be made of McLuhan's extremely influential conception of the global village (McLuhan 1964, p. 93, Gordon 1997). This was centered on the expansion of forms of communication since the invention of movable type and then the printing press in the fifteenth century up to global communication via radio and television, but prior to the World Wide Web, the Internet, and other by now commonplace features of the contemporary period of electronic communication.

Throughout most of human history there have been different civilizational conceptions of the scope, features, and conditions of the world. Scholars have, over thousands of years, speculated and theorized about the place of terrestrial earth in the cosmos. Of particular import in this regard was the period of the European Renaissance, an era of radical restructuring of philosophy, the arts, and the sciences in the fifteenth and sixteenth centuries (Cassirer 1963). A pivotal occurrence of this period was when the Polish scientist-philosopher Copernicus showed in 1512 that the earth was not the center of the universe, but that planets exist in a heliocentric universe. This perspective had truly major historical ramifications, in that the Renaissance as a whole, to put it all too simply, involved within Christendom an increasing valorization of *humanitas*, in contrast to previous worldviews that stressed the preeminence of God. The Renaissance view, well articulated by Cusanus (Cassirer 1963, pp. 7–45), did not, however, elevate humanity to the level of God. For Cassirer, the significance of Cusanus was in the latter's exposition of a move from the pre-Renaissance idea of a metaphysical world of transcendence to the Renaissance view of a scientific world of immanence. This shift represented not a triumph of the latter over the former, but rather a complex form of complementarity (Domandi 1963).

Closely intertwined in the European context with the crystallization of ideas concerning humanity were voyages and travels of 'discovery' to other parts of the world, particularly to Asia and to the Western Hemisphere. The 'discovery' of Africa, from the European standpoint, was not to come until a little later (Wills 2001, pp. 32–44). This expansion of the heliocentric world added a crucial spatial dimension to the more scholastic ruminations of leading figures of the European Renaissance, helped greatly by map-making becoming much more expansive, detailed, and accurate than had been the norm prior to that. It is of great importance to recognize the intimacy of the relationship between the motifs of humanity and globality, which were brought closely together during the Euro-

pean Renaissance and were to come even closer, up to and including, the European Enlightenment of the eighteenth century. This involved, again, an enrichment of the conception of humanity and an even broader and more concrete extension of the ideas concerning global variety. Consciousness of the latter was not often explicitly expressed via the specific concept of globality, this being a latent rather than a manifest theme.

Ever since the Moon landing of 1969, consciousness of what has come to be known as globality has increased. Space travel generally, as well as much of science fiction and the modern cinema, have greatly enhanced our tendency as human beings to gaze upon our global habitat, to become increasingly conscious of and sensitive to our globality. Even though this exact term may not yet have been used much, the ancient Greek notions of *oikoumene* and *okeanos* have remained central to conceptions of what are now considered under the rubric of globality. For example, the recent work of the Swedish anthropologist, Hannerz (1996, pp. 1–13), is centered on the concept of the global *ecumene* (*oikoumene*), a term which he has borrowed from the cultural anthropologist, Alfred Kroeber. Kroeber (1945, p. 9) saw the scope of anthropology to be the global, ecumenical circumstance, suggesting that humanity at large consists of interwoven relationships among a variety of societies and civilizational regions, with particular reference to culture.

In spite of this long etymological history, had it not been for the enormous and still growing degree of attention given to the contested concept of globalization (Robertson 1992, 2001a, 2001b), the notion of globality would not have acquired the contemporary significance that it now has. The discussion of globalization in a considerable number of social sciences and humanities has led to increasing talk about our now living in a global age (e.g., Albrow 1996), even though, as has been stressed, the philosophical, metaphysical, and theological interest in this topic, in a diffuse sense, dates back to ancient speculation about what we would now call the human species and the shape, configuration, and place of the earthly world in the universe.

In fact the phrase, 'the global age,' has become a common, although not unproblematic, characterization of the present condition of humankind. In this regard, there is a controversy in a number of disciplines concerning the relationships between such themes as tradition, modernity, postmodernity, and globality (Robertson 1992, pp. 138–45). Indeed, it is not fully possible to delineate the contours of globality without situating this concept, in contemporary terms, within this array of themes. Likewise, it is not possible to neglect the issue of locality in relation to globality, largely because the two have often, although very simplistically, been viewed as opposites, within the debate about globalization and related issues in recent decades of so-called modernity. In fact the relationship between globality and modernity is particularly controversial, basically because some have claimed that the process of globalization and the consciousness of the condition of globality have been a consequence of modernity (e.g., Giddens 1990), while others (e.g., Robertson 1992, 2001a) have insisted that globalization was a precondition for the rise and spread of different types of modernity in, or to, different parts of the world.

Globalization as an analytic concept first began to be used rather widely in sociology, religious studies, and, and to a lesser extent, anthropology in the early 1980s, although the theme of world, or global, history had been thematized in different civilizational contexts much earlier in the twentieth century than that, falling, however, into temporary decline from the 1920s until the 1980s. The concept of globalization came to be employed in a truncated and increasingly politicized way in some social-scientific, political and journalistic circles, but particularly in the field of business studies, from the late 1980s onward (Robertson and Khondker 1998, Robertson 2001a). This narrowing of the concept of globalization has had the consequence of pushing the theme of globality into the background. Even so, even were this circumscription to be comprehensively embraced, the theme of globality would not thereby be eliminated. This is so in spite of the fact that many academics, not to speak of the 'informed public,' do not appreciate that the concept of globalization was being debated quite extensively before it was more or less independently taken up in business studies, policy-making, and ideological forms. The theme of globality has very little to do in any significant respect with conceptions of globalization that limit it primarily to economic change and politico-economic policy. But the notion of globality is not necessarily marginalized as a consequence. The great attention that has been given to human rights and related concerns since the late 1970s helps to demonstrate this. For the virtually worldwide interest in and controversy about human rights have become closely bound-up with concrete matters of politicoeconomic policies.

When the term globality is explicitly employed, it usually refers to an historical circumstance of human life, and it is the extent to which this circumstance marks a new phase in the history of humankind as a whole that is particularly important, as well as problematic. More specifically, it may be defined as the intensification of more-or-less worldwide connectivity and increasingly reflexive global consciousness (Beck 2000, Robertson 1992). Overall, the condition of globality is accentuated greatly by the compression of the world. To this definitional statement must be added that, in a simply objective sense, this worldly planet was as global millennia ago as it is now. In this regard the world has always existed in-itself, but over many centuries it has become more and more for-itself, this meaning that as human beings we

have become more and more conscious of and have acquired doctrines and views about the future of the world.

Looking at this from the opposite direction, the notion of globaphobia (Burtless et al. 1998) may be invoked. Globaphopia, a term coined originally by a group of economists to characterize resistance to the opening of national economies to free trade and globewide economic competition, may be defined, more generally, as a sentiment involving resistance to globewide, more loosely what are perceived to be foreign or alien, phenomena and trends. Such trends are regarded as threatening the integrity and primordiality of a particular collectivity. Globality is in this sense a form of complexity, a complexity that is seen by some as going beyond the complexity of what has usually been called modernity—or, more recently, postmodernity. For example, in the USA, so-called Christian fundamentalists have spoken vociferously of the evil of (American) secular humanism. In more recent years, however, the enemy has become secular globalism (Bowen 1984). Even though parallel conflicts are to be found in many societies around the world, that between the self-proclaimed antiglobalists and those who subscribe in one way or another to a much more open society within the USA is, given the paramountcy of the USA in the world, of unavoidable—indeed global—significance (Robertson 2001b).

The perceived threat of globalism as a conspiracy against core American values has intermittently been, but particularly in recent years, a conspicuous feature of political discourse in the USA. Many societies have traditions of belief in conspiracy, but the USA has been one of the most prominent in this respect. This has partly been a response to the opposite trend—to be found in many American educational circles—involving the advocacy (often, but by no means only, for economic reasons) of teaching about other cultures and languages, as well as the condition of the world as a whole. This leads to a much more general—indeed, global—issue concerning the ways in which, to greatly varying degrees, people experience existential difficulty in being confronted by alternative ways of life. This is often encapsulated in the idea of alterity, or otherness. Rigid rejection of 'the other' is to be seen most clearly in what has come to be declared by some as fundamentalism (Robertson 1995). Thus, when normatively embraced, globality involves an eager acceptance of the idea of the global-human condition. At the other extreme, fundamentalism entails adamant rejection of the latter, at least insofar as one confines the concept of fundamentalism to specific societies or localities. A complication here, however, is that there is a form of global fundamentalism, one that stresses the priority in everyday life of globality and humanity over such aspects as international relations, features of particular societies or the life-cycles and identities of individuals.

The quotidian problem of alterity arises primarily from processes of relativization. Relativization involves the situational problems arising from the challenges posed to a particular way of life by the problematic realization that there are other, alternative ways of life. In a world marked by its high degree of connectivity, confrontations between different forms of life are ever-increasing. For, as the world becomes more connected and people around the globe become increasingly aware of both their differences and their shared fate, the likelihood of dilemmas and ambivalences concerning 'others' greatly intensifies. Indeed, probably the most general feature of long-term globalization, over many centuries, has been the persistence of relativization processes.

Although the notion of globalism is often used, even by mainstream academic contributors to the discourse of globalization, in its broad, comprehensive—as opposed to its narrow economistic—sense, it is misleading so to do. Specifically, the sentiment, or even ideology, of globaphopia often involves the presentation of 'the enemy' as the ideology of globalism or what is also often called, in a pejorative sense, oneworldism. Thus, even though some social scientists use the idea of globalism with the best of intentions, that is, to indicate their or others' cognitive and moral commitment to a global perspective, they nonetheless unwittingly engage in the rhetoric of resistance to the idea that we now live in a condition of shared globality. So globality is a much more preferable concept in relation to much, but by no means all, of the talk about globalism. The two terms have very different connotations, although they are often used interchangeably, with subsequent confusion. Globality indicates—for the most part—a condition, or circumstance, while globalism is an ambigious term, ranging in its meaning, on the one hand, from the relatively innocuous expression of adopting a so-called global perspective to a pejorative, condemnatory declaration, on the other hand, of the evils of subscription to the view that there is a general, inclusive conception of humanity and that, moreover, we all, every single human being, live in and share a single, global place (Robertson 1992).

When we read or make such statements as 'we now live in a global world,' 'the world is my home,' or 'we must take the feeling of home into exile' (see Iyer 2000), we are drawing attention to global complexity and at the same time welcoming it. This is indeed a personal and seemingly positive kind of globalism, but at the same time, from the analytic standpoint, such attitudes are merely symptomatic of the condition of reflexive globality. The question necessarily arises as to whether these orientations are merely to be found among certain kinds of intellectual or more privileged people, such as those who flit from one habitat, or habitus, to another. Thus it is not unreasonable to ask the question as to whether it is only certain global elites who dwell in and welcome the circumstance of

globality. However, in many ways it is the world's most deprived peoples, not to mention those, such as students, who move back and forth between 'tradition' and 'modernity,' that are most reflexively conscious of the global condition. Moreover, the everyday life of people all over the world, even though they may be relatively, or even absolutely, immobile geographically is becoming increasingly global, not least through representations of 'other cultures' in movies on cable or satellite TV, through popular culture generally, or everyday interaction, including electronic communication. In other word, vicarious, virtual interactional communication is increasingly part-and-parcel of everyday life. Even though celebrities, business elites, and some academics are increasingly conscious of the condition of globality, they do not in fact have a privileged position in this respect; for in one way or another people all over the planet earth are evidently becoming more and more 'global.' While it has to be conceded that there is much resistance all over the world, at the beginning of the twenty-first century, to strangers, it has been argued persuasively that this, often violent, opposition to polyethnicity and polyreligiosity (in the symbolic form of ethnic cleansing), constitutes the dusk of the norm of homogeneity and the contested return of what has been described as the polyethnic norm (McNeill 1986).

Whether 'the average person' is or is not fully conscious of their 'global membership,' many 'average people' nonetheless participate in increasingly reflexive global consciousness. The development of such has been proceeding for a very long time, at least since the crystallization of the so-called world religions in what has been called the Axial Age (Jaspers 1953), for example, through pilgramages, world map projections, voyages of 'discovery,' and so on. In much more recent times such a variety of phenomena as two World Wars, the Cold War, the European and other holocausts, the global AIDS pandemic, awareness of global ecological problems (notably global warming), the World Soccer Cup and other world championships, rapidly burgeoning touristic travel, global festivals, and so on and so forth—the list could be vastly extended—amply illustrate this.

While, in its most simplistic formulation, globality may be interpreted as referring to planetary territoriality and the beings (including nonhuman forms of life) that inhabit this earthly place, it is increasingly clear that globality involves and indicates variety and difference with respect, *inter alia*, to gender and sexuality, 'race' and ethnicity, nationality, and so on. Moreover, each of these, and other, attributes of human beings have become increasingly elective. In our time, to varying degrees, one choses one's body, one's 'roots,' one's gender, one's ethnicity, one's nationality, even one's own form of globality. For some (e.g., Beck 1999) we now live in a global risk society, a world in which we are continously confronted with a multitude of choices, choices that we make without the firm norms and values of tradition and in the absence of secure agencies of authorative trust. In short, our planetary circumstance is 'a runaway world' (Beck 1999, 2001).

When the idea of the global-human condition is invoked, the aim is to underline that we would not now be talking about globality were it not for the connection of this theme with that of humanity. The latter certainly has had a very long, although until fairly recently, rather abstract, history. But since the European Renaissance and particularly since the European Enlightenment, the theme of humanity has become increasingly concretized. The intersection of the concept of humanity with that of globality is to be seen with particular clarity in the French philosophical and sociological tradition from the early nineteenth century through the first quarter of the twentieth century. This was a period during which theological ideas were widely rejected, but also one when a new kind of, or surrogate, religion was deemed to be required. For example, a disciple of Saint-Simon wrote that 'there is in fact only one branch of knowledge, that of humanity, which includes everything and epitomizes everything' (Merle 1987). Saint-Simon's followers elaborated his thinking in a new journal, the *Globe*. August Comte, who had been Saint-Simon's secretary, was to advocate a religion of humanity in his later years. In sum, major French writers of the first half of the nineteenth century maintained not merely that globality and humanity were complicitous, but that a new form of social-science (sociology) inevitably arose from this connection. In the late nineteenth century this relationship was further elaborated by major sociologists such as Emile Durkheim and Marcel Mauss.

Since the 1960s there has been a strong predilection to conceive of 'the local' as the opposite of 'the global'—as reflected in the maxim 'think globally, act locally.' However, recent demonstrations in various countries against global institutions—such as the World Bank, the International Monetary Fund, and the World Trade Organization—have clearly involved a severe modification of this maxim. Protesters against global institutions have come to recognize that global targets require global movements (Robertson 2001a, 2001b). This is probably the most obvious way in which the condition of globality has become inexorably central to the life of many people across the globe. For ostensibly, antiglobal movements have, paradoxically, become part of the process of globalization itself.

See also: Global History (Universal History, World History); World Systems Theory; Globalization and World Culture; Globalization Anthropology of; Globalization, Geographical Aspects of; Globalization, Legal Aspects; Globalization, Political Aspects; Globalization Subsuming Pluralization, Transnational Organizations, Diaspora, Postmo-

dernity; International Organization; International Political Economy; International Trade: Economic Organization; Local–Global Links, Environmental; Capitalism: Global

Bibliography

Albrow M 1996 *The Global Age: State and Society Beyond Modernity*. Stanford University Press, Stanford, CA

Beck U 1999 *World Risk Society*. Polity Press, Cambridge, MA

Beck U 2000 *What is Globalization?* Polity Press, Cambridge, MA

Beck U 2001 Living your own life in a runaway world: Individualisation, globalisation and politics. In: Hutton W, Giddens A (eds.) *On the Edge: Living with Global Capitalism*. Vintage, London, pp. 164–74

Bowen W M 1984 *Globalism: America's Demise*. Huntington House, Shreveport, LA

Burtless G, Lawrence R, Litan R, Shapiro R 1998 *Globaphobia: Confronting Fears About Open Trade*. Brookings Institution Press, Washington, DC

Cassirer E 1963 [1927] *The Individual and the Cosmos in Renaissance Philosophy*. University of Pennsylvania Press, Philadelphia, PA

Domandi M 1963 Translator's introduction. In: Cassirer E (ed.) *The Individual and the Cosmos in Renaissance Philosophy*. University of Pennsylvania Press, Philadelphia, PA, pp. vii–xii

Giddens A 1990 *The Consequences of Modernity*. Polity Press, Cambridge, MA

Gordon W T 1997 *McLuhan for Beginners*. Writers and Readers Publishing, New York

Hannerz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London

Iyer P 2000 *The Global Soul*. Bloomsbury, London

Jaspers K 1953 *The Origin and the Goal of History*. Yale University Press, New Haven, CT

Kroeber A 1945 The ancient Oikoumene as an historical cultural aggregate. *Journal of the Royal Anthropological Institute* **75**: 9–20

McLuhan M 1964 *Understanding Media*. McGraw-Hill, New York

McNeill W H 1986 *Polyethnicity and National Unity in World History*. University of Toronto Press, Toronto, Canada

McNeill W H 1992 *The Global Condition*. Princeton University Press, Princeton, NJ

Merle M 1987 *The Sociology of International Relations*. Berg, New York

Robertson R 1984 Interpreting globality. In: *World Realities and International Studies Today*. Pennsylvania Council on International Education, Glenside, PA, pp. 7–19

Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London

Robertson R 1995 The search for fundamentals in global perspective. In: van Vucht Tijsson L, Berting J, Lechner F (eds.) *The Search for Fundamentals and Modernization: The Process of Modernization and the Quest for Meaning*. Kluwer, Dordrecht, The Netherlands, pp. 239–62

Robertson R 2001a Globalization theory 2000+: Major problematics In: Ritzer G, Smart B (eds.) *Handbook of Social Theory*. Sage, London, pp. 458–71

Robertson R 2001b Anti-global religion? In: Juergensmeyer M (ed.) *Global Religions*. Oxford University Press, Oxford, UK

Robertson R, Khondker H H 1998 Discourse of Globalization: Preliminary Considerations. *International Sociology* **13**(1): 25–40

Tomlinson J 2000 *Globalization and Culture*. University of Chicago Press, Chicago

Willis J E 2001 *1688: A Global History*. Granta Books, London

R. Robertson

Globalization and Health

Recent revolutions in communication have resulted in the development of a global village. Isolated areas of the world have 'joined' the technologically sophisticated world and developed, and sometimes, parochial regions have been increasingly faced with the realities of life in less fortunate parts. Growing awareness of the disparities in lifestyles and health around the world are forcing an increasing recognition that global health policies and practices are a need, not for the future but for the present.

Health care in this increasingly disparate community must shift from the high-technology focus, which has gained predominance in technologically sophisticated societies, to incorporate broader social, cultural, and psychological concerns. For example, it is probably of far greater importance in the fight against disease to provide clean water in most parts of the world than it is to even provide doctors. Concerns with malaria, TB, AIDS, and HIV dominate global health while poverty, discrimination, inadequate education and environmental contamination provide fertile growth for even greater social epidemics. Health-care disparities are extreme with most of the world's health-care money, interest, and attention being focused on those with the least need. Ultimately, however, health care in both the developing and the developed world are interrelated and setting priorities as to how these issues can be addressed with mutual benefit is needed.

A number of themes characterize health in the global environment in the past and future decades. These will be dealt with in turn.

1. *Technological Development*

Health care in the industrialized world has mirrored the dominant ethic of the twentieth century: technology development. The remarkable advances in technology that have resulted in a rapidly changed and changing environment at the close of the millennium have also brought with them increasingly complex challenges.

As technology is developed, so too is the commercial basis for its adaptation and use. Overuse of technology

in health care has been the result, with new advances being applied for the benefit of all rather than only the few for whom the technology is specifically directed.

On the other hand, technology is not readily available in less affluent societies. We need to ask whether it is acceptable to have a two-tiered system of health care with regard to appropriate use of technology for developing and developed countries? How well have we addressed issues of equity regarding access to appropriate medical care? How do financial considerations influence who receives differing types of care? To what extent does remuneration of health care providers and or health care services influence the quality of care provided or the use of available technology? How well are we monitoring the appropriate use of technology and the medical model worldwide? Are the currently available indicators of health sufficient to assess the influence of this level of functioning or are alternative indicators needed to address this issue? How should we evaluate the appropriate use of technology worldwide?

2. *Evidence-based Medicine*

Evidence-based medicine is one response to emerge in reaction to the technological explosion. Demanding that technology be applied when, and only when, it is effective, and ensuring that the strictest scientific rigor be applied to judge such effectiveness is leading caregivers to question the overuse of technology (Cochrane Collaboration 2000, Entin et al 1996).

We have developed a computer-based globe, moving increasingly into cyberspace in addition to its more earthly properties. Health and health care have become enmeshed in this web so that medical information and knowledge are moving more readily into the hands of the layperson than ever before. Medical practitioners are having to share their knowledge with their clients, raising issues such as shared decision making, collaborative care, information exchange, confidentiality, and informed choice by recipients of care. Decision making regarding health care, by both professionals and by families, needs to be evidence based: a fair challenge to achieve. Central to these developments are also legal and ethical issues which surround the management of health care and influence health-care decisions. Health-care services, to cope with these demands, need to move from a physician-focused service to a client-centered approach.

3. *Family-centered Care*

The client too is changing and is being viewed not just as a single entity but as a member of a family unit (whatever that may be), so that care is becoming family-centered care (Chalmers and Levin 2001). The family itself is changing rapidly too, meaning that health care is directed to anything from nuclear, extended, or single-parent families, or to same-gender couples. The traditional role of the doctor–patient relationship is becoming chameleon-like in its complexities.

4. *Psychosocial Health Care*

The need to incorporate holistic components of care into health has never been more strongly felt than now, although this is not a new idea. As long ago as 1948, the World Health Organization's Constitution (World Health Organization 1948) defined health as 'a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.' It is far preferable to care for the client as a whole person rather than as separated fragments. In earlier medical texts, support was often expressed regarding treating the 'patient as a whole person with thoughts, feelings, and fears, and not only as a body progressing through the health care system.' Unfortunately, in technologically sophisticated societies, superspecialization also occurs. Physicians no longer care for all aspects of a client personally: For example, in reproductive health, some specialists are focused on care of the cervix, others on the endometrium, ultrasound diagnosis is delegated to specialized technicians, and even such simple procedures as taking blood are relegated to yet other caregivers. Research has documented that as many as 16 different caregivers in a 6 hour period (Enkin et al. 1996) may look after a woman during a normal birth. A holistic or psychosocial approach to care hardly exists in this structure.

Assessing people's perceptions of their satisfaction with care is one approach to monitoring the integrity of their health care. Feedback regarding educational programs is well accepted in technologically sophisticated societies. No conference ends without a delegate evaluation being completed. Why is this not similarly integrated into health-care services? Is it because the clients' voices are still not acknowledged as an integral part of healing and are weighted as of far less significance than the biological basis of their illness? Becoming partners in care, rather than simply clients, is still the challenge for the future. 'Humanizing' the medical setting has become a clichéd term, yet it is still needed.

5. *Cultural Differences in Understanding of Health and Illness*

Cultures differ widely in their beliefs and understanding of health, illness, and appropriate health care. Many regard the biological level of functioning as of primary significance. In contrast, significant sections of the global village regard this level of functioning as of least importance with spiritual, social, and psycho-

logical concerns playing a primary role in illness etiology, treatment, and prognosis (Chalmers 1990). It is as unacceptable to the latter to neglect consideration of spiritual rites in diagnosis and treatment of illness as it would be to fail to take a woman's blood pressure during pregnancy in technologically sophisticated societies.

While these differing approaches to care have existed peaceably side-by-side for centuries, recent global population movements have encouraged the intermingling of cultures through migration, refugee relief and more temporary forms of translocation, with resulting conflicting health care paradigms. Compounding the problem are language differences combined with status inequalities which may make it extremely difficult for the foreigners' health care needs to be met or, indeed, understood in their adopted country. As multicultural societies continue to develop rapidly in the coming decades, the challenge of integrating their health care needs will become increasingly evident.

Traditional practices of healing should be considered with sensitivity and tolerance, and encouraged unless evaluated as distinctly harmful. Condemning or disallowing such practices because they are different is insufficient for although they may not contribute to biological health, they might be important for psychological health and should be encouraged. They should be evaluated as biologically beneficial, inconsequential, or harmful with only the latter forms of care being discouraged.

6. *Education*

Education to prepare for a multicultural, psychosocial, and evidence-based approach to health is a prerequisite in the twenty-first century. A number of medical institutions have introduced teaching psychosocial issues to medical and allied students. For the most part these have failed. Frequently based in Social or Behavioral Science Departments, they tend to be offered in preclinical years and to focus on theory. They are often perceived by students as programs to be endured, or which are undemanding, and not central to their ability as future caregivers. When offered at a graduate level, as the real challenge of caring for people is encountered in clinical practice, their worth is sometimes reconsidered.

Rather than being offered as distinct, independent course modules, the psychosocial and cultural issues involved in clinical practice should be integrated into mainstream medical teaching programs, e.g., when clinical management of a fetal death is discussed the psychological impact of diagnostic and treatment procedures in addition to longer-term family adjustment issues should be considered. Teaching should be shared by the physician and social scientist concurrently and such joint teaching programs, when tried, have shown much promise of a new and better model to develop for educational success (Chalmers and McIntyre 1993). In addition, providing an integrated model of teaching in the classroom will also establish a model for integrated clinical practice.

It is evident that health science faculties will need to undertake radical curriculum reforms in the twenty-first century to incorporate new, and fundamental, principles of education into their teaching. At the very least, these reform programs need to incorporate evidence-based medicine, family-centered care, multidisciplinary approaches, and psychosocial and multicultural aspects of care. Educational institutes need to be structured in such a way that facilitates the advocacy of both medical and social models of health care. How many medical schools teach social and cultural components of care as an integral component of medicine and how many simply pay lip service to this? Do we even know?

7. *Efficiency of Care*

A significant contributory factor to unease about care are the lengthy delays often encountered when seeking consultations with specialists, diagnostic assessments and their results, and treatment. Once a problem has developed, anxiety levels are high, particularly when fear of severe illness (such as cancer) or potential surgery becomes a possibility. Yet people frequently have to wait weeks or longer for simple diagnostic procedures such as ultrasound and even longer for more sophisticated tests. Further delays before receiving the outcomes of tests compound the problem. It is not surprising that frustration, fear, and helplessness are aroused with concurrent irritation and suspicion of the health-care services.

We frequently use the term 'consumers' to describe health-care clients. If customers in commercial stores were treated with similar delays in service and impersonal attention they would simply shop elsewhere. Within the health care service, with its more threatening and intimate experiences for clients, and its lack of real alternatives from which to choose in many instances, there is reluctance to apply a similar model. This does not mean that all is well in the state of health care, however. That dissatisfaction exists is evident in the increasing reality of litigation.

8. *Conclusion*

This article highlights some of the critical issues surfacing in the health-care services of the developed world from a broader global perspective. These kinds of concerns underscore the importance of those long-held values: concern for the whole person and not only the body, respect for people (and their families) and for their perceptions and needs when undergoing

medical care, a sincerely gentle and caring approach to care with sensitivity rather than routine politeness being the order of the day combined with evidence-based practice. It is apparent that while we have long recognized the importance of providing respectful treatment and care and have gone far in introducing systems which encourage such approaches, we are still grappling with the effective application of this fundamental tenant in many parts of the world, both developed and developing.

See also: Ecology and Health; Health Education and Health Promotion; Health in Developing Countries: Cultural Concerns

Bibliography

Chalmers B 1990 *African Birth: Childbirth in Cultural Transition*. Berev Publications, River Club, South Africa

Chalmers B, Levin A 2001 *Humane Perinatal Care*. Tea Publishers, Tallinn, Estonia

Chalmers B E, McIntyre J A 1993 Integrating psychology and obstetrics for medical students: Shared labour ward teaching. *Medical Teaching* **15**: 35–40

Cochrane Collaboration 2000 The Cochrane Library. Update Software, Oxford, UK

Enkin M, Kierse M, Renfrew M, Neilson J 1996 *A Guide to Effective Care in Pregnancy and Childbirth*. Oxford University Press, Oxford, UK

World Health Organization 1948 *Constitution of the World Health Organization*. World Health Organization, Geneva, Switzerland

B. Chalmers

Globalization and World Culture

1. Globalization and the Construction of World Culture

Globalization involves expanding worldwide flows of material objects and symbols, and the proliferation of organizations and institutions of global reach that structure those flows. World culture refers to the cultural complex of foundational assumptions, forms of knowledge, and prescriptions for action that underlie globalized flows, organizations, and institutions. It encompasses webs of significance that span the globe, conceptions of world society and world order, and models and methods of organizing social life that are assumed to have worldwide significance or applicability.

While many types of global flows have been rising cyclically for centuries, it was only in the second half of the nineteenth century that a transnational cultural complex took a sufficiently organized form to constitute an emerging world culture. The primary locus of this nascent world culture was Europe, particularly the powerful Western European countries whose empires extended to most corners of the globe. Europeans promoted political and philosophical principles, societal and individual goals, modes of organizing, and ways of conceiving and manipulating reality that they deemed universally applicable. Epitomized above all by technical and scientific principles and practical knowledge presumed to be invariant across time and space, emerging world culture also included more historically bound constructs and ideologies, such as nationalism, citizenship, and individualism. This early version of world culture, more properly called 'transeuropean' culture, was carried far and wide by missionaries, traders, military expeditions, colonialists, intellectuals, and travelers.

In this same period, transnational organizations and global structures emerged with increasing frequency, eventually to form a structural backbone or framework for world culture. The vast majority of these transnational organizations were products of international or global 'civil society'—voluntary associations founded and operated by individuals from many countries to pursue specific goals through democratically coordinated action. Typical early examples include the International Charity Association (1855), the International Sugar Union (1864), the Scandinavian Dental Association (1866), the Permanent International Committee of Architects (1867), and the International Meteorological Organization (1873). By the 1890s, such international nongovernmental organizations (INGOs) were appearing at the rate of more than 10 per year, across a wide range of social sectors, drawing participants mainly from Europe and North America but also from Latin American and some Asian countries, particularly India. These bodies defined themselves as global actors and sponsored periodic conferences at which universalistic issues, problems, methods, and solutions were proposed and debated. They came to constitute a formalized global public realm in which world culture was defined, documented, elaborated, and propagated to what the growing number of participants in this public realm were beginning to think of as a single world society.

The calamitous world wars of the twentieth century severely interrupted world-cultural structuration, but after each war the process rebounded quickly. States became increasingly engaged in transnational cooperative relationships through intergovernmental organizations (IGOs, which were rare before 1920). After World War II, the IGO population burgeoned to a total of several hundred organizations, while the number of INGOs soared into the thousands. This expanding complex of global organizations came to center on the United Nations, whose agencies and programs became axes of global governance

regimes in such major institutional areas as education, health, and development.

Until the 1970s, the content of world culture remained primarily Anglo-European, increasingly dominated by the USA. American popular culture and products proliferated throughout the globe—rock music, soap operas, bluejeans, Coca-Cola, fast foods, Hollywood movies. The globalization of tastes in food, dress, and music also promoted a global identity model, that of the freely choosing, pleasure-seeking consumer. At a deeper level, the plethora of global organizations, both nongovernmental and intergovernmental, produced a range of standardized principles, models, and methods for the organization of social life. For example, a basic world model of the modern nation-state crystallized, stipulating that the state was to be responsible not only for internal order and external defense but also for building a modern society and promoting citizen welfare and civic engagement. States were therefore expected to generate national educational systems, health care programs, retirement and unemployment insurance schemes, and cultural promotion programs, among many other duties. In particular sectors, globally legitimated models emerged to guide states in meeting these responsibilities, such as the standardized educational models that took shape under UNESCO's aegis in collaboration with a range of educational INGOs. Despite enormous variation in national circumstances and resources, states have implemented these models in remarkably uniform ways.

By the 1980s, world-cultural structuration had produced largely standardized global models for an enormous range of activities—in science, medicine, health, business, technology, even recreation and leisure (sports, tourism, entertainment). Pushed by the carriers of these models, including INGOs and IGOs but also business consultants, technical advisers, and academic consultants who explain and interpret the models, states, local governments, corporations, social movement organizations, and many other types of social units go to considerable lengths to keep up with the leading edges of world-cultural development. The net result is a tremendous force for the homogenization of the world.

2. *Globalization and Cultural Differentiation*

World culture is not only a homogenizing force; it also engenders and supports diversity and differentiation. Recognition of this feature of world culture was slow in coming; through the 1970s, most analysts interpreted globalization as essentially equivalent to homogenizing Americanization. Five factors are important for understanding world culture's promotion of heterogeneity; ironically, several of these are also important elements of world culture's homogenizing capacity.

(a) *Success of the nation-state political form*. Rapid decolonization after 1945 produced a world organized almost entirely as independent states. Most new states have eagerly joined global governance organizations, debating and helping to shape agreements expressing world-culture principles and prescriptions. Citizens of the new countries became avid joiners of INGOs, expanding the range of their memberships much faster than citizens of older countries. Thus, the voices and views of the Third World are increasingly prominent in world-cultural development, and debates about proper social organization and action are often less consensual than in the past. Cultural controversy has been especially acute in such domains as women's rights, environmentalism, and human rights.

(b) *Cultural relativism and the ideology of cultural authenticity*. Intellectual movements in the social sciences and humanities, coupled with the ideologies of nationalism and national self-determination, have made the principle of the fundamentally equal value of all human cultures a central assumption of world culture. Ethnocentrism has come to be seen as both trap and injustice; tolerance and, indeed, the championing of difference occupy the moral high ground. Of particular moral virtue in contemporary world culture are the poor, the excluded, the oppressed—marginal peoples whose right to their own cultures has been violated by the onslaught of globalization (some see this as a return of 'noble savage' idealization). This universalistic form of particularism impels peoples to emphasize or invent tradition and distinctiveness in counterpoint to universalistic world-cultural principles that are supposed to operate uniformly in all places. In Western societies, such particularism is stimulated by cosmopolitan connoisseurs seeking authentically exotic cuisine or ethnically distinct 'world' music, as the genre is known. At the same time, politically astute groups understand the power of purported cultural authenticity as a fulcrum for prying rewards from global systems, so distinction and difference have become strategic resources for collective mobilization. In consequence, ethnonationalist movements of many sorts (Basques, Biafrans, Bretons, Croats, Eritreans, Quebecois, Scots, Tamils, etc.) and mulculturalist restructuring by states, churches, schools, and other organizations have become the order of the day.

(c) *Regionalism*. Roughly half of the international organizations founded since 1950 have been regional in scope, activating European, Latin American, Asian, francophone, Islamic, Andean, and many other subglobal identities. Like ethnonationalism, regionalism has flourished in the wake of world-cultural intensification. Global structures, ideologies, principles, and models provide an overarching framework of commonality and shared meaning for disparate social units. With the framework well in place, diverse regional organizations and movements have expanded to implement, modify, and argue about the frame-

work's content and implications. We see the same process at work within international organizations and global corporations. For example, Cable News Network International (CNNI) began as a unitary organization bringing a standardized English-language product to televisions around the world, but it eventually established numerous regional centers producing local-language broadcasting adapted to specific target audiences.

(d) *Consumerism as adaptive interpretation*. While global popular culture contains products and symbols that penetrate most local markets, its accompanying ideology of economic freedom and consumer choice encourages varied uses and interpretations of standardized consumption patterns. For instance, consumers in Hong Kong might appreciate a McDonald's restaurant as much for its clean facilities as for its food. They expect fast, not smiling, service, and reserve the right to eat at their leisure. In a city pressed for space, the restaurant can become a gathering spot for youth or a favored locale for birthday parties. By adapting marketing to local attitudes, McDonald's has lost its foreign aura, blending into Hong Kong's diverse, cosmopolitan culinary scene. In Beijing, meanwhile, McDonald's American origins can be attractive to upwardly mobile Chinese as a symbol of participation in global society, but among many Koreans in Seoul those same origins evoke criticism. While not denying the power of multinational corporations to foster acceptance of their products and logos, recent research suggests that they must take into account the distinctive tastes of potentially fickle consumers. Popular culture increasingly creates global tastes, but it also creates critical consumers who keep this culture from becoming entirely uniform.

(e) *Creolization*. In many domains, the homogenizing force of world culture encounters ingrained local tastes and traditions, which may in turn be the residue of earlier such encounters with neighbors or colonizers. Due to creative interaction processes, common elements take different forms in different places. The spread of world culture therefore produces not homogeneity but new cultural melanges, each 'creolized' in its own way. A case in point is West-African popular music. Popular styles that themselves are blends of variegated traditions (Brazilian, blues, British rock) become the ingredients of new and distinctive styles. Adapting indigenous instruments or rhythms, African musicians participate in a global musical movement. Cultural flows come from many directions, intersect in unpredictable ways, and continue on in new forms to yet other places. How they continue, and the publics they eventually reach, depend in part on the global commercial media system's marketing and genre-promotion efforts, but that system is only one factor shaping local aesthetic orientations. Moreover, commercial interests have much to gain from promoting diversity. Thus, as new musical or other fads make their way across the globe, they produce new forms of localized diversity and unforeseen local reactions and interpretations. In the expressive realm, the varied forms of local–global interaction are not inexorably yielding to a standardized, stifling, hegemonic world culture.

3. *Dimensions of World Culture*

Most discussions of world culture focus on its expressive or normative dimensions. Expressive (popular) culture includes media products, consumable goods that assume iconic status as symbols of modernity or avant-gardism, foods and clothing styles originating in particular cultures that become worldwide fads, and so on. Normative culture comprises the values and goals promoted by global corporations and organizations—fetishistic consumerism and capital accumulation, individualism and human rights, democracy and political participation, gender and racial equality. The expressive dimension has received the greatest attention, particularly with respect to the global popularity of American films, television shows, sports figures, and lifestyle elements. Some scholars decry the obliteration of local cultures; others argue in favor of creative interpretation and adaptation, as local cultures integrate new cultural elements while retaining their basic identities (hybridization or creolization). Similarly, analysts decry some aspects of normative cultural globalization and praise others, often sharply disagreeing about the desirability of particular global norms and values and disagreeing as well about the extent to which a global normative consensus is emerging or even possible.

Less attention has been paid to the ontological and cognitive dimensions of world culture. At the ontological level, the reorganization of social life in accordance with dominant world-cultural models engenders increasing individualism, the disenchantment of nature, reliance on rationalized images of society, and an action-oriented model of the individual. Put another way, world-cultural penetration changes prevailing views of 'the nature of things,' promoting the worldview that underlies such major modern institutions as formal education, experimental science, national accounting and statistical systems, and advanced-technology engineering. These and similar institutional complexes bring with them broad bodies of knowledge and distinctive cognitive styles, but the depth of cultural reconstruction they engender remains largely unexamined because few attempts have been made to come to grips with their cultural implications. Science, mathematics, bookkeeping, organizational structure, technology, tools, construction materials, land surveying—all these have achieved the status of almost unassailably universalistic activities, uninteresting because they are mundane, routine, matter-of-fact. Thus, these highly formalized arenas seldom become the focus of ideological debate,

with the notable exceptions of such issues as nuclear power, genetically modified organisms, global warming, and large dams. The most highly institutionalized forms of cultural globalization remain largely invisible.

4. Globalization and Cultural Conflict

World culture engenders conflict in world society through a process that has, as it were, turned the West against itself. By the 1970s, when scores of new states had formed in Africa and Asia, vociferous opposition to continued economic and cultural domination by Western countries (labeled neocolonialism and cultural imperialism, originally formulated in the West) began to penetrate global organizations, especially UN bodies. Various associations of less developed countries called for a restructuring of world society—a New World Economic Order and a New World Information Order that would put restraints on the operations of transnational corporations and shift resources to the poorer countries. Former colonies began assertively invoking world-cultural principles of equality and development as basic human rights. Further tensions arose as the global human rights sector (itself increasingly involving individuals and groups from outside the West) turned its scrutiny on the new states, decrying their selective championing of certain global universals while ignoring or severely violating others, especially those relating to the integrity and political participation of citizens. Thus, both the particularism of world culture (the moral legitimacy attributed to national or ethnic units) and its universalism (the insistence that all units abide by basic principles and values) yield forms of disagreement and conflict that would not emerge in a less globalized world.

As non-Western cultures and regions have become more prominent in the world polity, it makes increasing sense to speak of world cultures (in the plural) rather than a singular world culture. Dominant Western models have penetrated deeply in most places, but they have also evoked resistance as well as efforts to revivify and globalize alternative models. Most notable in this regard is the assertiveness of Islamic cultural carriers, particularly since the 1970s. Many Muslim leaders and organizations promote a societal model that infuses the state with religious precepts and recasts the relationship between state and citizen (codified in 1981 as the Universal Islamic Declaration of Human Rights). African and Asian models of social organization and development have also emerged, and some observers argue that conflict in the twenty-first century will revolve primarily around grand civilizational axes rather than the nation-state clashes that have dominated in recent centuries.

By the same token, more fine-grained analyses of world culture(s) identify multiple models of central world-cultural constructs. For example, derived from the Western tradition are liberal, socialist, corporate, and welfare models of the state; from Asian sources, quasifamilial and state-led development models. Multiple models of the individual, the business enterprise, and the national polity are further examples. World-cultural complexity has increased rapidly in recent decades, perhaps most sharply since the collapse of Communism and the end of the Cold War, as ever more cultural centers generate more alternative cultural constructions. Poorer and more peripheral societies are less able to bring their cultural models to the world-cultural table, but many participants in the global arena from richer societies have become strong advocates of the poor and peripheral, helping to ensure that world culture becomes yet more complex and incoherent but also more significant for national and local structuration and change.

5. Religion and World Culture

Since the early period of European expansion, many of the symbolic flows, organizing institutions, and foundational assumptions that constitute world culture have been religious in origin. Through conquest, evangelization, and migration, Europeans made much of the world 'safe for Christianity,' giving emerging world culture a distinctively Christian cast even though most of the world did not become Christian. In the current phase of cultural globalization, the overall role of religious institutions may have diminished but various religious actors on the global stage vitally affect world culture in new ways, notably by providing alternative worldviews. Religious globalization has been important in constructing world culture, contributing to its heterogeneity, and producing new forms of conflict, as explained below.

Religion played a distinctive role in the construction of world culture, providing motivation for some globalizers and legitimating the actions of others. It also contributed to the foundations of contemporary world culture. The conception of societal progress that prevails almost everywhere stems in part from the Christian legitimation of rational world mastery; the global view of the individual as agent and citizen similarly has roots in Christian views of the person; the global script for organizing society in a rational-legal manner and legitimating authority voluntaristically derives indirectly from Christian precedents. More concretely, missionaries not only spread their faith, they also provided education and health care in far-flung places, diffusing secular commitments that since have become globally entrenched. Their actions prefigured the efforts of later movements in global civil society referred to above. Today distinctly religious views are less obviously influential in world culture, yet they still shape its evolution. According to some, a vigorous Pentecostal movement, successfully expand-

ing to Latin America, Asia, and Africa, carries with it the 'American' gospel of material success and individual well-being, thus supporting global corporate culture. The World Council of Churches, others argue, contributed centrally to global environmental concerns with its initiative on 'Justice, Peace, and the Integrity of Creation,' launched in 1983.

Religion's promotion of world-cultural heterogeneity reflects in part the increasing prominence of Third World voices in global religious organizations. Pope John Paul II has appointed a record number of non-Western bishops and cardinals. In the Anglican Church, non-Westerners significantly aid opposition to gay rights. Non-Western Pentecostals have not only grown dramatically in numbers, they have also devised their own spiritual practices to suit local circumstances. Cho Yong-Gi, pastor of the large Yoido Full Gospel Church in Seoul, initiated a world mission program that has sent hundreds of missionaries and established seminaries in several developed countries, including the USA.

Religion's relation to cultural relativism, another aspect of heterogeneity, varies greatly by tradition and is therefore more complex. In some cases, such as that of Hindu nationalism, religion becomes the primary vehicle for expressions of national distinctiveness. While few Christian churches accept the equal value of all cultures, many have in fact accommodated major cultural differences within loose doctrinal parameters. In Korea and Brazil, for instance, Pentecostal churches incorporate elements of traditional non-Christian spirituality. Islamic fundamentalism presents yet another variant, claiming global legitimacy for its defense of 'authentic' Islam while rejecting the value of other traditions. Finally, religion helps to promote heterogeneity through the creolization process mentioned earlier. Pentecostalism is once again a case in point. An American-inspired movement with a core evangelical message and a secular commitment to material progress has become a mosaic of styles and practices as local congregations have adapted global models to their own needs without direct outside control. As new global movements make inroads in 'national' religious traditions, they produce new juxtapositions that creolize religious cultures. Of course, this process is not unprecedented. Many successful religious traditions have balanced universal ambition with local creativity to produce hybrid religiosity 'on the ground' in counterpoint to unification at the doctrinal level.

Religion also contributes to world-cultural pluralism and conflict. Religious traditions contain diverse views of the good society, leading to varying interpretations of seemingly shared global values. Some view the dominant form of world culture as a coercive, Christian-biased Western force that threatens the integrity of their own traditions. Islamic fundamentalists oppose what they perceive as a godless consumer culture that undermines people's faith and serves the political interests of the West. Pope John Paul II has critized an amoral, neoliberal world order that puts profits ahead of the needs of the poor.

Specific elements of world culture also provoke distinctly religious opposition. In recent decades, for example, the principle that women are citizens—individuals to be granted equal rights and opportunities—has been institutionalized in world culture. Within many traditions, however, equality for women is problematic. The result is severe contestation of women's equality in some places, most notably by the Taliban in Afghanistan. These religiously inspired actors have helped generate a more general backlash against globalization, even while they accept many basic elements of world culture. Religion becomes a vantage point from which to oppose the threat represented by globalization as such. The struggle, as some groups see it, is between globalization and their cultural survival.

In sum, religion is a central aspect of cultural globalization, as traditions spread, transnational networks expand, national cultures become more mixed, and new ways of experiencing the world emerge. In many countries, religion mediates the pluralizing effect of world culture. It plays an important role in the intense contest concerning global values and world order. Yet the construction of world culture has become a mainly secular process; it has no transcendent content in the conventional sense. Cultural heterogeneity and conflict themselves take many forms, only some of them religious. While world religion is intimately connected with globalization and involved in the latter's dynamics, it is by no means a dominant force. Whether it can, or should, take on a greater role in defining the desirable world order is likely to be a central issue in future global cultural contestation.

See also: Capitalism: Global; Globality; Globalization, Anthropology of; Imperialism: Political Aspects; United Nations: Political Aspects; Values, Anthropology of

Bibliography

Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN

Boli J, Thomas G M (eds.) 1999 *Constructing World Culture: International Nongovernmental Organizations Since 1875*. Stanford University Press, Stanford, CA

Brouwer S, Gifford P, Rose S D 1996 *Exporting the American Gospel: Global Christian Fundamentalism*. Routledge, New York

Featherstone M (ed.) 1990 *Global Culture: Nationalism, Globalization and Modernity*. Sage, London and Newbury Park, CA, in association with *Theory, Culture and Society*

Friedman J 1994 *Cultural Identity and Global Process*. Sage, London

Hannerz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London
Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon & Schuster, New York
Jameson F, Miyoshi M (eds.) 1998 *The Cultures of Globalization*. Duke University Press, Durham, NC
King A D 1997 *Culture, Globalization and the World-System: Contemporary Conditions for the Representation of Identity*, Reprint edn. University of Minnesota Press, Minneapolis, MN
Lechner F 1992 Global fundamentalism. In: Swatos W H Jr. (ed.) *A Future for Religion? New Paradigms for Social Analysis*. Sage, Newbury Park, CA, pp. 19–36
McLuhan M 1994 *Understanding Media: The Extensions of Man*, 1st MIT Press edn. MIT Press, Cambridge, MA
Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London
Smith A 1980 *The Geopolitics of Information: How Western Culture Dominates the World*. Oxford University Press, New York
Thomas G M, Meyer J W, Ramirez F O, Boli J 1987 *Institutional Structure: Constituting State, Society, and the Individual*. Sage, Newbury Park, CA
Tomlinson J 1991 *Cultural Imperialism: A Critical Introduction*. Pinter, London
Van Vucht Tijssen L, Berting J, Lechner F J (eds.) 1995 *The Search for Fundamentals: The Process of Modernisation and the Quest for Meaning*. Kluwer Academic Publishers, Dordrecht, The Netherlands, and Boston, MA
Watson J L (ed.) 1997 *Golden Arches East: McDonald's in East Asia*. Stanford University Press, Stanford, CA

J. Boli and F. J. Lechner

Globalization, Anthropology of

1. The Disciplinary Context

Anthropological interest in globalization is difficult to trace precisely but can be recognized as early as Appadurai's work on the global cultural economy (1990, 1996), Hannerz's analysis of cultural complexity and creolization (1987, 1992), and Friedman's work linking global capitalism and cultural processes (1988, 1994), all of which built on earlier anthropological interest in cultural exchange and world systems. From the start, such work has been part of two sets of scholarly exchanges and debates. The first links anthropology with a wider debate about globalization in other social science fields, notably geography, political science, and sociology. The other is internal to anthropology and links it to longstanding concerns with diffusion, comparison, ethnography, and the study of large-scale historical change. A brief look at each of these contexts might help to frame the anthropological perspective on globalization.

As regards the social sciences in general, anthropologists in the late 1980s found themselves, like many other scholars, trying to understand the processes that led to the breakdown of the Soviet world and to guess at the shape of things to come. They soon found themselves having to engage powerful perspectives on the emergent world order, which came out of the Marxist tradition in geography (exemplified by Harvey 1989), by a renewed interest in global political culture, exemplified by the polemical and much discussed work of Samuel Huntington (1996), and some prescient studies of the new forms being taken by global capitalism in the last decades of the twentieth century (Lash and Urry 1987). In addition, anthropologists found themselves in a rich, sometimes competitive dialogue with scholars in literary and cultural studies, notably those influenced by British cultural Marxism, and most powerfully represented by Stuart Hall (1986). In addition, the publication of Benedict Anderson's landmark study of nationalism (1983) provoked a strikingly wide debate about the links between politics, the imagination, and national identification. These stimuli helped to shape anthropological research on globalization, which in many ways is marked by the ongoing effort to link broad structures and processes in the world economy to the subtleties of communication, interpretation, and translation that govern everyday life in all societies.

Within anthropology, the study of globalization built on several well-established currents of interest. In the United States, the study of globalization fit well with Boasian traditions in the study of diffusion, cultural change, and culture contact and with a longstanding interest in urban settings and complex societies. In Europe, anthropologists were slower to take an interest in globalization, but here too longstanding interests in problems of scale, social cohesion, and structural change made globalization a subject of growing interest after 1990. In other parts of the world, such as Latin America, Africa, and India, the anthropological interest in globalization was more closely linked to problems of ethics, development, and inequality.

In spite of these various currents of research and theorization within which the anthropological interest in globalization has developed, from the beginning there were serious doubts about whether globalization by its nature was a topic suitable to the special strengths of anthropology in the study of intimate social relations and of societies governed by non-market social principles. Many of these anxieties have been translated into methodological concerns, reflected in a significant body of methodological work about how anthropology ought to address the emergent world of globalization (e.g., Fox 1991). There is still a considerable body of opinion among professional anthropologists that globalization may well be a mere trend, an artifact of academic fashion, and that in any case, it is not an ideal subject for anthropological research because of its conceptual and social scale. But this rearguard anxiety, not always easy to distinguish

from fear of disciplinary failure, has not prevented a rich set of theoretical debates and empirical studies from emerging out of anthropological traditions, both in the Atlantic world and beyond. These studies, largely a product of anthropological work in the 1990s, are discussed in the following sections.

2. Global Cultural Flows

The image of 'flow' has been a central trope of anthropological work on globalization. In the work of Appadurai (1990, 1996), the idea of 'flow' was used to capture a complex dynamic in which objects, ideas, ideologies, technologies, and images were placed in a single economy of circulation, with an eye to distinguishing different emergent mosaics of cultural form and social design. In Hannerz's use of this concept (1992, 1996), it was fruitfully linked to Kroeber's idea about large-scale civilizational and interactional spheres—ecumenes—as well as to ideas about creolization, cosmopolitanization and the professional networks associated with rapid flows of cultural commodities.

In these early usages of the idea of flow, and in subsequent elaborations of them by others, the idea of flow was sharpened to redress the assumption that cultural flows in the era of globalization were unrestricted in scope and range. A salutary emphasis on boundaries, limitations, selectivities, and exclusions reappeared, thus placing the classic debate between Marxist and cultural approaches once again at the center of attention. This broad interpretive struggle continues, and the choices it poses are reflected in many studies associated with the anthropology of globalization, on such subjects as migration (Basch et al. 1994), nationalism (Chatterjee 1993), colonialism and postcolonialism (Comaroff and Comaroff 2000a), media (Feld 1999), diaspora (van der Veer 1995), and markets (Hart 1999).

The positive contribution of the idea of flows was to open up new questions about the relation between legal and illicit flows of commodities and persons, about the changing nature of borders and boundaries, about the implications of new regimes of financial circulation and cybercommerce, and about the evolution of new forms of diasporic identification and mobile tradition building. The image of flows also opens a potential conversation, yet to be fully exploited, between economists and anthropologists on the question of the changes in the world economy suggested by the idea of globalization, and of the best ways of measuring their scale and significance. In a general way, the engagement with cultural flows, their associated economies and interpretations, has decisively marked a shift, in which anthropologists are no longer confined to the study of simple market societies but are addressing such complex phenomena as cybercommerce, intellectual property, and new forms of genetic engineering and marketing. Likewise, the anthropological focus on cross-border flows, as we will see shortly, has opened up a useful conversation between anthropologists, political theorists and sociologists on new forms of citizenship and sovereignty.

Above all, the focus on flows, regardless of context, commodity, or sphere, has brought anthropology back to one of its earlier insights, namely that most human societies have always been in interactive relations with others, forming spheres of exchange and circulation. Thus the images of social stability, impermeable boundaries, and natural divisions between ethnic groups, polities, and communities, sometimes implicit in anthropological practice, have been put under sharp scrutiny. In turn, the widespread recognition that societies in the era of globalization are inevitably parts of wider circulatory systems and networks, has posed new challenges for how anthropologists will need to theorize classical models of cohesion, consensus, and order. At the end of the twentieth century, there is some uncertainty about how to reconcile the undoubted significance of large scale cultural differences with the reality of the velocity of cross-border flows and the uneven thickness of the membranes between societies, whether organized as nation states or not. It seems probable that the question of how social order and cultural coherence in lived communities are maintained in the face of complex and uneven cross-border flows will be a major theoretical challenge for twenty-first century anthropology.

Many of these concerns have produced a special interest in the anthropology of nationalism and a series of contested views about the salience, form, and future of the nation state as a form of global political organization. The pendulum appears to swing between those (both inside and outside anthropology) who predict the early demise of the nation state, and those who see it as growing stronger than ever. Sometimes these debates are reflections of actual differences between existing states and sometimes they are the result of blurring the lines between nations and nationalism on the one hand, and states and their powers on the other. The emerging consensus seems to be that while nation states are unlikely to disappear in the near future, they are evolving their own powers in a transformed ecology of sovereignty. This ecology is affected by changes in the velocity of global capital flows, in the scale and speed of cross-border shadow economies, and in the geography of dual, mobile, and diasporic political identifications among citizens. These processes are discussed further below (see Sect. 4).

The interest in various forms of global cultural flow, and in the nonisomorphic relationship between them, has provoked a strong debate about whether, and the extent to which, globalization is a new social regime. Many voices within anthropology argue vigorously that globalization is no more than a semiotic con-

spiracy of the media and of global capital and its allies in the world of development. In this view, globalization is only a clever way to repackage colonialism, imperialism, modernization, or various combinations of these ideologies. It is old wine in new bottles. This view runs into the difficulty of having to accommodate too many new animals into an existing theoretical Ark, and suffers from the further disability that it does not provide any criteria for recognizing genuine systemic change. It ends history in the name of theory. The more widespread band of opinion among anthropologists is that something surely has changed. The challenge is, as always, to establish what has changed and how it affects the things that have not changed. Thus, there is a renewed interest in how identities are negotiated across multiple locations, how existing structures like witchcraft (Geschiere 1997) become vehicles and sites for new anxieties and energies, and of how new legal and economic arrangements (those sometimes lumped together under the rubric of neoliberalism) call forth new perceptions of danger, wealth, and risk (Comaroff and Comaroff 2000b).

Finally, ideas and debates surrounding the idea of global cultural flows have cast fresh light on a classic concern of anthropology, namely the process of identification and the production of identities. These processes have to be reexamined in a world characterized by massive information flow, heightened media images of life possibilities for ordinary people, and new fantasies of wealth and mobility. The rapid growth in mass-market media penetration across national boundaries, the associated growth in the battle between global and national media barons, the projected increase in various kinds of bandwidth, and the emergent possibility of new, hybrid instruments for cheap, high-speed communication, all combine to change the grounds on which stable cultural identities are produced and secured. Likewise, these factors make it impossible to presume the image of the local as an unchanging ground against which the tableau of global change is played out. Indeed, many anthropologists now have come to take as great an interest in 'the production of locality' (Appadurai 1996) as in the dynamics of globalization.

With its special interest in language, everyday life, and small-scale social interaction, anthropology is evolving a special purchase on the vast terrain represented by globalization, by paying particular attention to the vagaries of translation and localization and to the special role of language and semiotic mediation in producing new forms of lived experience. This has resulted in a sustained effort to examine the cultural dimension of objects and technologies that have their primary home in the spheres of law, science, and the market. Therefore, anthropologists at the beginning of the twenty-first century study the organ trade, refugee incarceration, NGOs, musical copyright law, and derivatives, to name just a few examples of social and technical forms whose cultural design is sometimes overlooked. In this proliferation of topics (some of them addressed by other authors in the Encyclopedia), there are some major crosscutting themes, three of which are briefly discussed below.

3. *Heterogeneity or Homogeneity?*

Most anthropologists share the widespread presumption that globalization is a new phase in the integration of the world economy, principally powered by new scales and speeds in the circulation of financial capital and the associated explosion in electronic communication technologies. When we recall that the major models of global economic processes have come out of the Marxist tradition, it helps to explain why there is a tendency among some anthropologists to argue that globalization is no more than a new phase in the generally history of capitalist commodification. From this follows the possibility that, from a cultural point of view, the forces of commodification will produce increasing levels of cultural homogenization and social standardization, particularly in the realm of media icons, social styles, and consumption values. The further extrapolation, rarely stated explicitly, is that these processes will in turn erase cultural differences both within and across societies. This is the nightmare of globalization as coca-colonization (Hollywood, McDonald's, and Disney are the favored icons of this line of speculation). Yet the evidence that societies appropriate various forms of commodification in their own terms, that media messages are downloaded idiosyncratically, that weapons of standardization are frequently subverted to produce further brand differentiation, and that the best laid plans of the standardizers are frequently thwarted by local entrepreneurs and customers, is mounting.

A splendid example of this process of differentiation is provided in a collection of anthropological essays on the role and style of McDonald's, as it functions in six distinct East Asian settings (Watson 1997). The ethnographic essays in this collection illustrate that, as a global profit-making organization, McDonald's is certainly eager to create uniformity in its products and in its outlets wherever possible. But the essays also shows that in each of these settings McDonald's encounters a quite different set of needs, anxieties, and possibilities and is drawn into a specific local mosaic of social patterns and cultural orientations, involving age, leisure, work, and freedom. Though McDonald's is a heavily capitalized multinational corporation, with an unerring eye for its local and global profits, its sensitivity to local markets inevitably contributes to the emergence of different desires and practices in its outlets. This does not make it any less a global corporation, but its strategies are regularly bent to fit local wishes and projects. However, these essays also show that this is not a simple matter of Goliath being repackaged to suit David, culturally speaking. In fact,

what McDonald's produces in these settings are new and unintended styles and contexts for consumption, for sociality, and for public interaction, each culturally distinct but not in any obvious or predictable way. Many other studies, especially in the sphere of consumption (involving clothing, music, food, leisure, and housing) also confirm this process.

Heterogenization is not a mechanical product of the sphere of consumption. It is itself produced by local makers of ideas, images, and commodities who give to the culture industries of different societies and nations their own distinctive stamp. Thus, anthropologists have been able to show that the simple distinction that places global producers and uniformity on the one side, and local consumers and difference on the other, is overdrawn. We have a growing series of studies of cultural production worldwide, especially in the areas of music, film, and advertising, which let us look into the sites and institutions through which global commodities are locally interpreted by producers as well as consumers. The study of 'world music' by ethnomusicologists is perhaps the best developed of these subfields. In general, these studies have produced a broad consensus that cultural differentiation tends to outpace homogenization, even in this most interactive of economic epochs. Naturally, this conclusion fits with the methodological and ethical proclivities of anthropology, with its professional investment in cultural difference and its ethical commitment to cultural diversity. Nevertheless, there appears to be substantial empirical basis for this consensus. But anthropological interest has by no means focused solely on the good news of cultural differentiation and has not been uncritically celebratory in its treatment of globalization, as the following section will show.

4. *Sovereignty, Citizenship, and Violence*

Anthropologists have not failed to notice that the world of global cultural flows is hardly free of suffering, injustice, and dislocation. They have certainly been aware that the incidence and brutality of forced population movements have in some respects increased, producing many pockets of human dislocation and frequently of ethnic violence. They have paid attention to the dilemmas of proletarian labor migrants, who struggle with their own translocal loyalties—frequently to religious formations which are themselves increasingly transnational in scope—against the pressures of the varied ideologies of cultural and legal inclusion which characterize their new homes.

Considered transnationally, poorer migrants face bewildering new combinations of democratic ideology and racist reality. In some cases, such as the oil kingdoms of the Middle East, nonimmigrants monopolize most of the benefits of citizenship. In others, the benefits of citizenship are more widely shared, but everyday racism is a constant threat. In some cases, all refugees are drawn into multinational ethnic wars (as in Central Africa in recent years), while in other cases, the revitalization of longstanding ethnonational boundaries flies in the face of intermarriage and ethnic mixing as social realities. In Central Europe after the death of Tito, the new national constitutions frequently encouraged political campaigns that exacerbated genocidal identities parading as 'primordial.' Many nation states now practice various forms of cultural fundamentalism and majoritarianism, as they struggle for new forms of legitimation in a global economy. In the process, they increasingly both produce and demonize 'minorities.' In many of these cases, language, blood, and race are overdetermined icons of peoplehood which force anthropologists to reconcile their own knowledge of the fragility and historicity of these constructions against the popular, lived sense that these realities are timeless, natural, and nonnegotiable.

As anthropologists examine these bloody sites of ethnic violence and violent social purification (Appadurai 1998, Das 1995, Feldman 1991, Malkki 1995), they find themselves having to study a variety of antecedent and emergent social realities. These include new forms of 'flexible' citizenship (Ong 1999); new ideologies of citizenship and law (Comaroff and Comaroff 2000b, Sassen 1996); new battles over indigenes and outsiders (Meyer and Geschiere 1999); and new patterns of state-sponsored violence and legality. In engaging with these topics, anthropologists discover that they need to link longstanding professional interests in conflict, bodily ritual, mythologies of origin, and ethnocultural ideology, with new regimes of law, incarceration, borders, and migration.

In particular, anthropologists are now tackling the ways in which the forms of violence and dislocation produced by globalization may be related to new structures of sovereignty, new forms of nationally sponsored globalization, and new niches for the location of the global economy. They thus find themselves in a productive dialogue with scholars in adjacent fields who are making important contributions to these subjects (Castells 1997, Held and McGrew 2000, Sassen 1998). This engagement promises a more sophisticated answer to the question of how new forms of global linkage are also producing new forms of cultural fundamentalism, a question to which somewhat Manichean answers have been provided in the popular media and its favored public intellectuals in many national contexts.

In the United States, and to some extent in Europe, there have been strong alarms raised by economists and development specialists to the effect that globalization is producing both greater wealth and greater inequality. Yet, in popular opinion, especially among the wealthier countries and classes, there seems to be an emergent sense that globalization is a world of the victors in the Cold War and their global allies who

have cracked the codes of the global market, the information economy, and the new international order. The losers are frequently viewed as cultural holdouts at best, cultural terrorists at worst, people who have not understood the new alphabet of the globalized world. Anthropological work on globalization, in this context, has a challenge and an obligation. The challenge is to take its classical strengths in the study of intimate relations, everyday life, and cultural difference and use them to cast light on circulatory processes that yield local and global results simultaneously. The obligation is to provide a way of thinking about why this era of global integration is yielding at least as much genocidal violence as it is democratic voice. That may not seem like a particularly anthropological obligation, but given that anthropology has long tried to create new ways of showing how broad human processes are found in translation, globalization may need anthropology at least as much as vice versa.

5. Looking Ahead

It is not easy to predict what topics will find special favor among anthropologists interested in globalization in the coming decades. It is likely that some of the broad themes discussed in this essay will continue to elicit empirical research and theoretical debate, but it also seems likely that the institutional and geographical range of the anthropological debate is likely to widen. Research institutes, conferences, workshops, publications, and translations concerned with anthropological contributions to the study of globalization are appearing in Brazil, Turkey, Hong Kong, India, and in many other sites outside of the North Atlantic region. In these many sites, anthropologists are building their own critical resources as they engage with specific kinds of national debate, cultural conflict, and research tradition (Appadurai 1999). Whatever the individual outcomes of this process, two results seem likely. Anthropology is unlikely to remain a bystander in any important national debate about globalization. Moreover, these debates are increasingly unlikely to be determined by scholars, projects, and institutions confined to Europe and the United States. What more appropriate topic could there be for international research than globalization, and what opportunity could be more appropriate for anthropology to contribute to a global scholarly debate about the world we are about to build?

Bibliography

Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Appadurai A 1990 Disjuncture and difference in the global cultural economy. *Public Culture* **2**(2): 1–24

Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN

Appadurai A 1998 Dead certainty: Ethnic violence in the era of globalization. *Public Culture* **10**(2): 224–47

Appadurai A 1999 Grassroots globalization and the research imagination. *Public Culture* **12**(1): 1–19

Basch L, Glick S N, Szanton B C (eds.) 1994 *Nations Unbound: Transnational Projects, Postcolonial Predicaments, and Deterritorialized Nation-states*. Gordon and Breach Scientific Publishers, Langhorne, PA

Castells M 1997 *The Power of Identity*. Blackwell, Oxford, UK, Vol. II

Chatterjee P 1993 *The Nation and its Fragments: Colonial and Postcolonial Histories*. Princeton University Press, Princeton, NJ

Comaroff J, Comaroff J L 2000a Millenial capitalism and the culture of neoliberalism. *Public Culture* **12**(2) Special issue

Comaroff J, Comaroff J L 2000b Millenial capitalism: First thoughts on a second coming. *Public Culture* **12**(2): 291–343

Das V 1995 *Critical Events: An Anthropological Perspective on Contemporary India*. Oxford University Press, Delhi, India

Feld S 1999 A sweet lullaby for world music. *Public Culture* **12**(1): 145–71

Feldman A 1991 *Formations of Violence: The Narrative of Body and Political Terror in Northern Ireland*. University of Chicago Press, Chicago

Fox R G (ed.) 1991 *Recapturing Anthropology: Working in the Present*. School of American Research Press, Santa Fe, NM

Friedman J 1988 Cultural logics of the global system. *Theory, Culture, and Society* **5**(2–3): 477–90

Friedman J 1994 *Cultural Identity and Global Process*. Sage, Thousand Oaks, CA

Geschiere P 1997 *The Modernity of Witchcraft: Politics and the Occult in Postcolonial Africa*. University of Virginia Press, Charlottesville, VA

Hall S 1986 Cultural studies: Two paradigms. In: Collins R (ed.) *Media, Culture and Society: A Critical Reader*. Sage, Beverly Hills, CA

Hannerz U 1987 The world in creolisation. *Africa* **57**: 546–559

Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York

Hannerz U 1996 *Transnational Connections: Culture, People, Place*. Routledge, New York

Hart K 1999 *The Memory Bank: Money in an Unequal World*. Profile Books, London

Harvey D 1989 *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Blackwell, Cambridge, MA

Held D, McGrew A 2000 *The Global Transformations Reader: An Introduction to the Globalization Debate*. Polity Press, Malden, MA

Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon & Schuster, New York

Lash S, Urry J 1987 *The End of Organized Capitalism*. University of Wisconsin Press, Madison, WI

Malkki L H 1995 *Purity and Exile: Violence, Memory and National Cosmology among Hutu Refugees in Tanzania*. University of Chicago Press, Chicago

Meyer B, Geschiere P 1999 *Globalization and Identity: Dialectics of Flow and Closure*. Blackwell Publishers, Malden, MA

Ong A 2000 *Flexible Citizenship: The Cultural Logics of Transnationality*. Duke University Press, Durham, NC

Sassen S 1996 *Losing Control? Sovereignty in an Age of Globalization*. Columbia University Press, New York

Sassen S 1998 *Globalization and its Discontents: Essays on the New Mobility of People and Money*. The New Press, New York
Soares L 1996 *Cultural Pluralism, Identity, and Globalization*. Conjunto Universitario Candido Mendes, Rio de Janeiro, Brazil
van der Veer P 1995 *Nation and Migration: The Politics of Space in the South Asian Diaspora*. University of Pennsylvania Press, Philadelphia, PA
Watson J L 1997 *Golden Arches East: McDonald's in East Asia*. Stanford University Press, Stanford, CA

A. Appadurai

Globalization: Geographical Aspects

1. Introduction

Space, place, distance, proximity, maps, boundaries, territoriality, and spatiality. These are some of the keywords of geography; keywords that distinguish the interest of geographers in the phenomenon of globalization, from that of others in the social sciences. The distinction is not clear cut by any means, not least because 'globalization' so centrally evokes geography: the rise of world-scale processes and phenomena, the intensification of linkage between distant places and cultures, and the associated unmaking and remaking of territorial boundaries and identities. Perhaps where the key difference lies, however, is that the other social sciences are not interested primarily in the intrinsic spatiality of globalization.

In economics and business studies, for example, studies of globalization focus on the implications of the borderless economy (driven by international market integration, global flows of money and information, and transnational production and regulation) on national and subnational economic sovereignty, competitiveness, and growth. The optimists celebrate the offer of new market opportunities, the pessimists warn of the real and discursive dangers of neoliberalism and international dependency, the historically minded reassure us that there is nothing much new (Drache 1999, Hirst and Thompson 1996), while the measured observers of qualitative change reveal the new aspects of combined and uneven development (Dicken 1998, Held et. al. 1999).

In politics and international relations, the debate has centered on the erosion/reformulation of the nation-state as a unit of authority, following the rise of nonstate institutions of regulation and governance, and on the challenge to citizenship and democracy posed by the rise of transnational political organisations (e.g., global NGOs) as well as plural, nonnational sources of political rights and identification (e.g., EU citizenship). Here too there is no consensus. Some herald a new era of reduced nation state influence, eroded national welfare commitment and crisis of parliamentary democracy (Garrett 1998), some assert the continuity of national social models (Hay and Marsh 1999, Weiss 1998), others signal the rise of new forms of unsettled jostling between the nation-state and other organizations organized at local, national, and international scales (Jessop 1994, Keil 1998), while others still find new prospects for cosmopolitan government, democracy, and civic mobilization (Held et al. 1999, Scholte 2000).

In social anthropology and media and cultural studies, the attention has fallen on the implications of global consumerism, media communication, and international mobility and cultural mixture for individual and social identities and lifestyles in different local settings. Early warnings of the erosion of local difference under the weight of the same world products and consumption norms and the same global corporations (i.e., the so-called MacDonaldization of society), have given way to sensitivity to the persistence of local difference, but also the nuances of local change resulting from heightened global exposure and virtual and real linkage with different parts of the world (Robertson 1992, Albrow 1996, Hannerz 1996, Morley and Robins 1995).

Geography, to repeat, is everywhere in these studies of globalization, but the spatiality of social relations, or put differently, how place, space, and spatial difference are being reconstituted is not the prime interest. Geographical theory, in contrast, has been concerned with the spatiality of the contemporary world, and is interested in understanding whether places—cities, regions, and nations—are perforating as geographically contained spaces, how the insertion of places into geographically stretched relations matters, and how new geographical scales of organization and influence associated with globalization are challenging old scales of identification and action.

This geographical imagination, to be clear, is not confined to the discipline of geography. In fact, some of the most influential accounts of the spatiality of globalization have come from social theory outside the discipline. For example, the sociologists Giddens and Held have long argued that a key aspect of globalization is spatial connectivity and reach, such that the conceptualization of place/locale as 'in here' happenings and space as 'out there' happenings is no longer tenable. Held (1995, p. 20), building on Giddens's idea of contemporary time–space distanciation (Giddens 1990), proposes:

> ... globalisation can be taken to denote the stretching and deepening of social relations and institutions across space and time such that, on the one hand, day-to-day activities are increasingly influenced by events happening on the other side of the globe and, on the other hand, the practices and decisions of local groups can have significant global reverberations.

The sociologist Therborn (1998, p. 7) too asserts that globalization should force us to think 'seriously

about space, about the spatiality of the social, about territories and their delimitations.' Two examples of spatiality that he identifies are: first, in a world seen as system, the 'flattening of social processes,' associated with the gradual global spread of dominant social practices (e.g., the triumph of capitalism on a global scale, the hegemony of neoliberalism, the Americanization of culture); and second, echoing Giddens and Held, the 'world as an arena where nationally determined actors meet, interact, and influence each other.'

In contrast, the international studies scholar Scholte (2000) interprets the spatiality of contemporary globalization (compared to preceding periods of intense internationalisation) as 'deterritorialization.' He notes that the rise of a 'social space that transcends territorial geography' is new (composed of such phenomena as transnational production, instantaneous worldwide communication, consumption of the same commodities and images, etc.), such that:

> 'global' relations are social connections in which territorial location, territorial distance and territorial borders do not have a determining influence. In global space 'place' is not territorially fixed, territorial distance is covered in effectively no time, and territorial frontiers present no particular impediment.

He concludes that 'while territoriality may continue to be important, globalization has brought an end to territorialism (that is a condition where social space is reducible to territorial coordinates alone).'

In all three accounts, it is the spatiality of globalization that is stressed, albeit in very different ways, from the emphasis on new territorial scales and connections, to warnings of interterritorial clash or territorial annihilation.

The interpretation shared, and explored further, in this contribution is the one offered by Giddens and Held. It sees the growth of world-level influences (e.g., transnational corporations and banks, common consumption patterns, world ideologies, international authority structures) and the intensification of global spatial connectivity (e.g., flows of people, goods, ideas, and information, facilitated by rapid transport and communications technologies) less of a threat to, or reinforcement of, territoriality, than a loosening and stretching of social geographies of linkage and identification. This is because a key aspect of globalization is:

> A process which embodies a transformation in the spatial organization of social relations and transactions—assessed in terms of their extensity, intensity, velocity, and impact—generating transcontinental or interregional flows and networks of activity, interaction, and the exercise of power (Held et al. 1999, p. 16).

The first part of this article discusses recent claims in geography on the changing relationship between place and space associated with globalization. It argues that the legacy of seeing places in scalar terms, either as geographical sites of proximate links or as territorial units, remains strong within the discipline. The second part—using the example of contemporary urbanity—dissents with the latter interpretation, to advance a very different sense of place. It argues for the need to see places under contemporary globalisation in non-scalar terms, and as perforated entities: as sites immersed in varied geographies of proximate and distant links, within which things no longer cohere as a bounded whole.

2. *Geographies of Scale*

Marston (2000) has reviewed the work of political geographers such as Neil Smith, Eric Swyngedouw, John Agnew, Kevin Cox, Peter Taylor, and Neil Brenner, to conclude that for them spatial scales—from home and locality to city, region, nation and continent—have no pregiven or fixed ontological status, but are produced and reproduced and continually transformed socially by the imperatives of capitalism. Importantly, they associate contemporary globalization with the rise of new territorial scales of organization, and new territorial boundaries, rather than the replacement of old, national scales and boundaries by a unitary global field of meaning and action. They do not see the replacement of an old world 'space of places' with a new 'space of flows' (Castells 1996).

Brenner (1999, p. 435), for example, asserts that 'the post-1970s wave of globalisation has significantly decentred the role of the national scale as a self-enclosed container of socio-economic relations,' and we are witnessing a 're-scaling of territoriality' which includes the increased 'importance of both sub- and supranational forms of territorial organisation.' Similarly, Swyngedouw (1997) mobilizes the term 'globalisation' to indicate that globalisation represents the break down and reconstitution of spatial scales; producing a new 'politics of scale' involving clashes of scale and contested boundaries. Globalization, thus, represents the 'relativisation of scale' (Jessop 1998, p. 90), with governance and political conflicts articulated at urban, regional, national, and supranational level, with no one scale privileged. The result is a re-scaling of territories, such that, for example, national states gear towards enhancing international competitiveness and capturing global investments cities, and cities or regions decouple from national economic circuits, to become locked into global ones (e.g., as investment sites).

The emphasis of these geographers on the produced and contingent nature of territories and scales marks, in Agnew's (1999, p. 504) terms, the 'historicity of spatiality.' He demonstrates, for example, that the present period contains a number of historical models

of power organized spatially: the model of dynastic preservation in a world of distant communities; the model of state–military conflict in a world of rigidly defined territorial units; the model of hierarchical conflict between centers and peripheries in an international economic system; and the model of nodal conflict within the networks of in an integrated world society. The 'geographical embeddedness of power relationships today, therefore, signifies something quite different from what was the case in the past' (p. 512), for example, contemporary globalization elevates the tensions between state territorial relationships (e.g., equating nation with citizenship or property rights) and transterritorial developments (e.g., the rise of globally mobile identities and property rights).

Through these readings, globalization can be seen as the multiplication of geographical scales of social organization, linked to the changing spatial requirements of the latest phase of capitalist development. These are not seen as parallel or mutually exclusive scalar geographies, but intersecting and overlapping ones, with places/territories implicated in the full variety of scalar reorientations. Thus, not an end to geography in a globalizing totality, but a rescaling of boundaries (upwards, outwards, and downwards) and a reconfiguration of territories as they engage in multiscalar processes and politics.

The strength of this interpretation of the spatiality of globalization lies in the scope it leaves for recognizing the persistence of geographical diversity, and the varied ways in which different places, with their unique legacies, are implicated in the balance between territorial and transterritorial influences. Its weakness, however—although Agnew is an exception—is that it fails to address whether the ontological status of place/territoriality itself is significantly altered as a result of the rise of world-scale processes and transnational networks and flows. The language continues to remain that of scalar geographies and territorial units. Reterritorialization follows deterritorialization and spatial scales are relativized under globalization.

Places such as cities and nations continue to exist as territorial units, but now with different external orientations (e.g., as sites in global production networks or places dependent upon international investment or competitiveness) and different scalar involvement (e.g., national welfare policies, continental trade agreements, global environmental regulations, local tax regimes). The boundary between place and space is not seen to be bending or disappearing—a line in shifting sand—but as a remaining line of distinction between 'in-here' relations and institutions, and those 'out-there.' Global connectivity is seen as leading to the nesting of territories into multiple scales, not territorial disorientation away from a geography of scalar nesting.

Cox's (1998) work on localities as particular sorts of spaces (of association and politics) in a multiscalar world provides a good example of this lingering territorial logic. He distinguishes within a locality the spaces of dependence or localized connections that he claims define 'our sense of significance,' from the spaces of engagement within global relationships, which he sees as threatening the place-specific 'essential interests':

> Spaces of dependence are defined by those more-or-less localized relations upon which we depend for the realization of essential interests and for which there is no substitute elsewhere; they define place-specific conditions for our material well being and our sense of significance. These spaces are inserted in broader sets of relationships of a more global character and these constantly threaten to undermine or dissolve them … In so doing they construct a different form of space which I call here a space of engagement: the space in which the politics of securing a space of dependence unfolds (Cox 1998, p. 2, cited in Marston 2000, p. 226).

In this account, localized relations are posed as distinguishable from global relations, and somehow more authentic and politically progressive: the local can be mobilized to resist the global. While Cox is clearly right to note that local societies continue to retain political potential in an age of allegedly globalized power, his dualist distinction between the 'local' and the 'global' as separate scalar fields remains methodologically problematic. To highlight the weakness of this dualist logic, should we, for example, judge the often charged diaspora connections of ethnic minorities or cosmopolitans in a locality to be less significant or progressive than the daily contact among local residents? Surely a key aspect of the transnationalisation of local relations is that we can no longer make an easy distinction between local and global geographies? How localized or global, for example, are the associational politics of worker, immigrant, and NGO groups campaigning for local recognition, but relying on international financial and other support networks?

Or, what makes everyday practices and customs grounded in local encounters and traditions any more 'essential' than nonlocal practices and customs? This desire to see places as territorial units in which local relations are counterposed to global relations leads to a peculiar politics of place in which relations within localities are cast as good and meaningful, and contrasted to bad and totalizing external relations. Its peculiarity follows from the ontological separation of place (read as in here, and intimate) and space (read as out there, and intrusive) as distinctive scalar realms.

Taylor's (1999) recent attempt to conceptualize place and space in nonscalar terms helps us gets around this particular ontological straightjacket, but replaces it with another rather stark binary distinction. Following the pioneering work of the humanist geographer Tuan (1977), Taylor claims that both terms should be used to identify human activity at all geographical scales—local, national, and global—but that they are distinct and always in dialectical tension

with each other. Taylor defines space in modernity as the realm of abstract principles, rules, rationality, science, administration, bureaucracy, and institutions. He judges it to be politically disenabling. In contrast he sees place as the substantive, politically enabling, lived realm involving intimacy, experience, belonging, feelings. The two are said to always exist in tension with each other, for example, the abstract realm of the state, combined with the felt realm of the nation, has yielded the nation-state, in which 'sovereign territory has been merged with sacred homeland to convert a space into the place.' Similarly, Taylor sees the modern home converting the space of the household into a place, and the techno-social space of the one-world (capitalist) system becoming a place through developments such as the rise of a global consciousness (e.g., global ecological fallibility).

Taylor's is an odd usage of the terms space and place, counterposing as it does, the abstract (disenabling) from the lived (enabling). Are not the spaces of science and technology, for example, R&D laboratories or domesticated science, just as lived and 'enabling' as more obviously emotional sites? Isn't the realm of experiences one that blurs the distinction between formal and informal practices? Why, for that matter, choose the labels of space and place to distinguish between the abstract and the lived? What remains helpful in Taylor's schema, however, is his rebuttal of space and place as separate geographical sites or realms, thus allowing the ontological presence of both the proximate and the remote at the same geographical level (e.g., the recognition of systemic rationality or state-sponsored science in the home, or the proximate warmth/odium of nationalist sentiment or universal causes).

The next section pursues this reasoning, but retains the geographical connotations of the terms space and place, rather too quickly abandoned by Taylor. It retains a geographical imaginary because it considers the reconfiguration of the spatiality of social relations a central aspect of contemporary globalization, and importantly, because it sees this reconfiguration shaping the meaning of what goes on in places (localities, cities, regions), how places coalesce or not as entities, how social relations in places are constructed, and how the politics of place matter. In short, it illustrates that place still matters, but no longer as a scalar unit.

3. *'A Global Sense of Place'*

The section heading borrows a deliberately oxymoronic phrase of Massey (1994) to conceptualize localities in nonscalar terms, or, in her words, 'relationally,' as sites of juxtaposed social realities, shaped by the layered historical connections and geographical links that are to be found within and beyond them. Their character as sites of agglomerated activity, thus, has far less to do with the properties of a territorial system (e.g., localized linkage, self-regeneration, local governance, etc.) than with the effects of spatial and temporal exposure and connectivity (e.g., continual and open-ended change, juxtaposition of difference, overlap of networks of different global connections). Two of Massey's now famous examples of places as 'open, porous, hybrid' (Massey 1999, p. 21) are Mexico City, with its varied architectures revealing layers of global Spanish and American-Indian historical connections, and Kilburn High Road in London, with its juxtaposition of English, Irish, and Indian diaspora lifestyles and connections.

It is important to note that since Massey defines the spatial in general 'as the sphere of the juxtaposition, or co-existence, of distinct narratives, as the product of power-filled social relations' (Massey 1999, p. 21), a global sense of place is needed not only for cosmopolitan places, but all places. For, each and every one:

> …may be imagined as particular articulations of these social relations, including local relations 'within' the place and those many connections which stretch beyond it. And all of these embedded in complex, layered, histories. This is place as open, porous, hybrid—this is *place as meeting place*…This is a notion of place where specificity (local uniqueness or sense of place) derives not from some mythical internal roots nor from a history of relative isolation—now to be disrupted by globalisation—but precisely from the absolute particularity of the mixtures of influences found together there (Massey 1999, p. 22, emphasis in original).

Through this lens we might re-imagine Bamako as richly as Bombay, Boston as richly as Bermuda, and mobilize them against persisting views of globalization as place homogenisation. Massey defines places as socially constructed 'envelopes of space-time': relational settings and meeting places, not closed local systems. How does contemporary globalization fit into this picture? Massey's response is:

> …globalisation, imagined through the lens of this conceptualisation of space-time, the globalisation we are facing now, is a thoroughgoing, world-wide, restructuring of those space-times, along particular lines. It is a re-making of those, inherited but always temporary and provisional, spaces, places, cultures which are themselves the hybrid products of previous restructurings.

We can take this to mean that a relational or global sense of place will interpret the spatiality of globalization—however defined—rather differently from a territorial perspective. Let us stay with the example of cities, and focus in turn on the spatiality of two aspects—economic transactions and urban politics—to illustrate this point.

3.1 *Urban Economic Space*

In economic discussions, the language of scales or territorial boundaries, as discussed above, has tended

to interpret the rise of transnational phenomena such as global informatics, international production and finance, and powerful international organizations, in terms of the boundary between a territorial inside and a territorial outside. Consequently, accounts which continue to see an economic place for cities contrast the global as the realm of distanciated links, fast flows, and world-scale processes, with cities as the sites of face-to-face-contact, or centers of agglomeration among global networks (e.g., clusters of HQs, creativity and knowledge, communication hubs). Against the economics of space and distance are posed the economics of agglomeration and clustering, the powers of spatial proximity and interiority. The argument posed is that in the global age, the locally embedded still counts as a source of economic dynamism and competitiveness because spatial proximity encourages trust and reflexivity, reduces transaction and communication costs, facilitates technological and knowledge spill-over, and offers specialized services and institutions to the business and professional communities (Maskell et al. 1998, Krugman 1991, Audretsch 1998, Glaeser 1998).

A sense of place that sees cities as a meeting place for diverse geographies of linkage, in contrast, offers a different reading of the urban economy. The geography of cyberspace is a good example for questioning the distinction between an alleged real world of direct contact and a virtual, telemediated world. As Kitchin (1998, p. 387) argues, 'whilst information on-line might seem geographically dislocated, information is only as useful as the locale within which the body resides', and in any case, 'cyberspace depends on real-world spatial fixity—the points of access, the physicality and materiality of wires.' Consequently, there is a continuity of communication in geographical space and place, such that, as Graham summarizes, 'power to function economically and link socially increasingly relies on constructed, place-based, material spaces intimately woven into complex telematics infrastructures linking them to other places and spaces' (Graham 1998, p. 175). Similarly, in daily contact, there is an unbroken line of communication linking physical and human technologies, and it is of varying geographical reach, mobilized to source information, strike a deal, exchange knowledge, and so on.

The different realms of communication do not offer different social worlds, for example, 'there is not one single, unified cyberspace; rather there are multiple, heterogeneous networks, within which telecommunications and information technologies become closely enrolled with human actors, and with other technologies, into systems of sociotechnical relations across space' (Graham 1998, p. 178). And this is precisely why the speeding up of time within communications networks should not be read as the speeding up of people and their lifestyles (Thrift 1996).

Seen in the context of linked communications architectures, there can be no privileged economic role for spatial proximity. First, co-presence or agglomeration need not imply association and interaction, for firms can mobilize a variety of contact networks to establish economic links with firms, markets, and institutions located elsewhere nationally and internationally. The near immediacy of distanciated communications networks enables economies of association to be established without the need for spatial propinquity. The city is no longer a bounded, but spatially stretched, economic sphere. Second, spatially proximate interactions, where they appear, need not privilege particular types of economic transactions such as those based on trust, reciprocity, and reflexivity through physical interaction.

Intimacy may be achieved through the frequent and regular contacts enabled by the distanciated networks of communication and travel (how else do transnational firms, institutions, and social movements work?), as well as the unbroken interplay between face-to-face and telemediated contact. The city is, therefore, not a place of unique or systemic transactional assets. Third, the agglomeration in some parts of some cities of large volumes of specialised inputs—from knowledgeable and creative people and clusters of inter-related industries, to sites of R&D, corporate power, public bureaucracy, and business services—no longer implies increasing local returns (as theorized by agglomeration economists). These sources of economic dynamism (which may well draw on local resources, for example, schools and college, leisure amenities, transport hubs) are nested into economic networks that stretch beyond the city (from transnational production networks and national or international supply chains, to global market relations).

A global sense of urban economic space recognizes places as clusters of overlapping network sites, without seeing in them only local economies of association feeding a localised economic system. It views with scepticism the territorial metaphors such as 'local fixity vs. global flow,' 'embedding the global,' or 'the local in the global' which have grown in recent years to 'defend' the space of local economy against global 'encroachment.' Instead, it highlights that local economic activity is part of, and inseparable from, proximate and distanciated transactions, and that whatever counts as the local has to be seen as a network site, not a territorial system.

3.2 Urban Political Space

Similarly, a global sense of place implies an interpretation of the implications of globalization on urban politics that is different from one based on a territorial sense of place. The territorial perspective anticipates either the loss of the urban as a political space under the barrage of national and global institutions and

movements, or, as exemplified by the work of Cox above, as a particular zone of engagement. It sees politics as undergoing scalar differentiation, as exemplified by waning interest in universalist causes and nation–state institutions, and the rise of ethnonationalist sentiment at regional level or global political institutions and movements.

In this schema, localities, when viewed positively, are expected to perform a certain kind of place-based politics: tribal, as implied by Cox, through his claims for local authenticity and feelings; formative, through involvement in 'militant particularisms' (Harvey 2000); lived, through conflicts over lifestyle, consumption, and social reproduction issues; experimental, via either the politics of difference forged around the diverse social and cultural voices gathered in the city (Sandercock 1998), or the politics of citizenship based on participation in the city's public spaces and public sphere (Sennett 1977). In short, a new politics of place serving unique purposes in a world of multiscalar politics (note, however, the different position of those political geographers who mobilize the concept of relativization of scales, in order to capture the new combined politics of local, national, and global action).

There is, however, another way of viewing place politics in age of global connectivity. Instead of seeing local political activity as somehow unique, places might be seen as sites that juxtapose or combine the varied politics—local, national, and global—that we find today. What matters is this juxtaposition/combination.

This forces, for one, recognition of actor participation in nonlocal political activity. To return to the urban example, as Sassen (1998) puts it, if globalization involves the urbanization of a world capitalist class (at least in some global cities) as well as people and cultures from different parts of the world, it also generates the possibility of claims in the city which are at once local, national, and transnational in orientation. This might include, for example, business leaders campaigning against the local homeless, national government tax policies, or international trade agreements; ethnic minorities campaigning against local racism, but also through the support of national antiracist groups, and international diaspora organizations; office workers gathered together with other groups outside the municipal offices protesting against local closures, but also linked up with workers and unions elsewhere through Internet discussions of world trade barrier removals. In this perspective the city is no longer a site of place politics, but a place of engagement in plural politics, and through multiple spatialities of involvement.

Why does this shift in emphasis from the politics of place to politics in place matter? Because, it marks the necessity to see political activity in places as plural, open, and contested (Amin et al. 2000). First, the juxtaposition of difference in close spatial proximity generates political challenges associated with this juxtaposition, for example, the contest between the business community and the homeless over the spaces of the city, or the competing demands of different classes, social and ethnic groups over the city's cultural resources. There is no automatic intimacy or consensus here against any perceived global intrusion.

Second, the idea of politics in place marks a significant change in the nature of contemporary politics, from one based on participation in singular, all-encompassing, political institutions and movements, to another based on participation in multiple causes and multiple networks of affiliation, and resulting in people, more than ever before, bearing multiple political identities as well as forging new ones. It recognizes a cosmopolitan politics that features at all spatial scales of organisation and activity (Soja 2000). Third, however, such a politics remains open-ended, and highly dependent upon cultural context and the balance of power, for its outcomes. Thus, no formative place politics of militant particularisms, but a cosmopolitan politics that is xenophobic and reactionary here, progressive and hybrid there, tolerant or particularly militant elsewhere.

4. *Conclusion*

This article has staked out a case for a nonscalar interpretation of the geographies of place and space associated with contemporary globalization. Its principal claim has been that the rise of a world of global flows and processes, and intensified transnational connections, signals the perforation of boundaries and the sharp combination of variety and difference at different geographical sites, rather than a shift in the balance of power between different spatial scales, or the deterritorialization and reterritorialization of activity. Thus, it has stressed the fluidity of cultural/political practices in space and place, against the emphasis elsewhere in geography on new scales of fixity. Of course, in matters of regulation and governance associated with globalization, there is a very real and felt contest of jurisdiction between local, national, and global state and nonstate organizations. It has not been the intention here to deny the state, but simply its legal and institutional inscription on other on-goings within places. Instead, the aim has been to look into how the very terminology of geographical boundaries, caught up in concepts of place and space, proximity and distance, has been challenged by globalization.

Beyond these largely silent differences of opinion within geography, the distinctive contribution of this discipline within the congested study of globalization, has lain in the study of the spatiality—social, economic, cultural, and political—of what is increasingly being seen as a single and interdependent world. The work of geographers has shown that this new world remains as varied, uneven, and contested as ever

before, with little evidence of spatial homogenization or the end of geography.

See also: Borders, Anthropology of; Capitalism: Global; Development and the State; Globalization and World Culture; Globalization, Anthropology of; Globalization: Political Aspects; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Place in Geography; Territoriality: Geographical

Bibliography

Albrow M 1996 *The Global Age: State and Society—Beyond Modernity*. Polity Press, Cambridge, UK

Agnew J 1999 Mapping political power beyond state boundaries: Territory, identity and movement in world politics. *Millenium–Journal of International Statistics* **28**(3): 499–521

Amin A, Massey D, Thrift N 2000 *Cities for the Many not the Few*. Policy Press, Bristol, UK

Audretsch D 1998 Agglomeration and the location of innovative activity. *Oxford Review of Economic Policy* **14**(2): 18–29

Brenner N 1999 Globalisation as reterritorialisation: The rescaling of urban governance in the European Union. *Urban Studies* **36**(3): 431–51

Castells M 1996 *The Rise of the Network Society*. Blackwell, Oxford, UK

Cox K 1998 Spaces of dependence, spaces of engagement and the polities of scale or looking for local politics. *Political Geography* **17**: 1–23

Dicken P 1998 *Global Shift*, 3rd edn. Chapman, London

Drache D 1999 *Globalization: Is there anything to fear?* CSGR Working Paper No 23/99, Centre for the Study of Globalization and Regionalisation. University of Warwick, Coventry, UK

Garrett G 1998 *Partisan Politics in the Global Economy*. Cambridge University Press, Cambridge, UK

Giddens A 1990 *The Consequences of Modernity*. Polity, Cambridge, UK

Glaeser E L 1998 Are cities dying? *Journal of Economic Perspectives* **12**: 139–60

Graham S 1998 The end of geography or the explosion of place? Conceptualizing space, place and information technology. *Progress in Human Geography* **22**(2): 165–85

Hannerz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London

Harvey D 2000 *Spaces of Hope*. Edinburgh University Press, Edinburgh, UK

Hay C, Marsh D 1999 Demystifying globalisation. In: Hay C, Marsh D (eds.) *Demystifying Globalisation*. Macmillan, Basingstoke, UK

Held D 1995 *Democracy and the Global Order: From the Modern State to Cosmopolitan Governance*. Polity, Cambridge, UK

Held D, McGrew A, Goldblatt D, Perraton J 1999 *Global Transformations: Politics, Economics and Culture*. Polity, Cambridge, UK

Hirst P Q, Thompson G 1996 *Globalization in Question*. Polity, Cambridge, UK

Jessop B 1994 Post-Fordism and the state. In: Amin A (ed.) *Post-Fordism: A Reader*. Blackwell, Oxford, UK

Jessop B 1998 The narrative of enterprise and the enterprise of narrative: Place marketing and the entrepreneurial city. In: Hall T, Hubbard P (eds.) *The Entrepreneurial City: Geographics of Politics, Regime, and Representation*. Wiley, Chichester, UK

Keil R 1998 Globalization makes states: Perspectives of local governance in the age of the world city. *Review of International Political Economy* **5**(4): 616–46

Kitchin R M 1998 Towards geographies of cyberspace. *Progress in Human Geography* **22**(3): 385–406

Krugman P 1991 *Geography and Trade*. MIT Press, Cambridge, MA

Maskell P, Eskelinen H, Hannibalsson I, Malmberg A, Vatne E 1998 *Competitiveness, Localised Learning and Regional Development*. Routledge, London

Marston S A 2000 The social construction of scale. *Progress in Human Geography* **24**(2): 219–42

Massey D B 1994 *Space, Place and Gender*. Polity, Cambridge, UK

Massey D B 1999 *Power-geometries and the Politics of Space Time*. Hettner Lecture 1998, Department of Geography, University of Heidelberg, Germany

Morley D, Robins K 1995 *Spaces of Identity: Global Media, Electronic Landscapes, and Cultural Boundaries*. Routledge, London

Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London

Sassen S 1998 Whose city is it? Globalization and the formation of new claims. In: Holston J (ed.) *Cities and Citizenship*. Duke University Press, Durham

Sandercock L 1998 *Towards Cosmopolis: Planning for Multicultural Cities*. Wiley, Chichester, UK

Scholte J A 2000 Global civil society. In: Woods N (eds.) *The Political Economy of Globalization*. Macmillan, London

Sennett R 1977 *Fall of Public Man*. Knopf, New York

Soja E W 2000 *Postmetropolis*. Blackwell, Oxford, UK

Swyngedouw E 1997 Neither global nor local: '*Glocalization*' and the polities of scale. In: Cox K R (ed.) *Spaces of Globalization*. Guilford, New York

Taylor P J 1999 Places, spaces and Macy's: Place-space tensions in the political geographies of modernities. *Progress in Human Geography* **23**(1): 7–26

Therborn G 1998 Globalizations. Swedish Collegium for Advanced Study in the Social Sciences. Uppsala, Sweden

Thrift N 1996 New urban eras and old technological fears: Reconfiguring the goodwill of electronic things. *Urban Studies* **33**: 1463–93

Tuan Y-F 1977 *Space and Place*. Arnold, London

Weiss L 1998 *The Myth of the Powerless State*. Cornell University Press, Ithaca, NY

A. Amin

Globalization: Legal Aspects

1. Introduction

In light of the dramatic intensification of transnational interactions since the 1970s and its impact on the legal field, one might be tempted to conceive of the 'globalization of the legal field' as a radically new phenomenon without any roots in the past. Though

such legal globalization, or, at least, some aspects of it represent a qualitatively new development, it should be borne in mind that the modern world system, within which the globalization of social interactions occurs, has been in place since the sixteenth century and that the roots of the most recent legal, as well as cultural, social, political, and economic transformations are to be located in this historical development. Moreover, the existence of a transnational legal culture antedates the modern world system, as is dramatically illustrated by the reception of Roman law in the twelfth century and onwards.

Even at the end of the nineteenth century, at a time when liberal political theory had imposed the equation among nation, state, and law, the idea of a global or world law continued to flourish as a legal subculture. Though cultivated at the time mainly by legal comparativists who were quite marginal to mainstream legal science, this legal subculture was part and parcel of a widely hegemonic cultural constellation, the idea of the seamless and irreversible uniformization of the conditions of social life throughout the world brought about by progress, capitalism, and scientific and technological development. In contrast with earlier movements for transnationalization that were the products of an intellectual crusade by well-meaning jurists and philosophers, the transnationalization of the legal field at the beginning of the twenty-first century is promoted by multinational corporations, practicing lawyers, state bureaucrats, and international institutions, as well as by popular movements and transnational advocacy NGOs. It is an extremely diverse phenomenon combining uniformity with local differentiation, top-down imposition with bottom-up creation, formal declaration with interstitial emergence. It cannot be accounted for by any monocausal explanation and defies integration in any unidimensional cultural postulate, be it the idea of progress or the critique of progress, capitalist mass culture or anticapitalist popular culture, modern science, or postmodern knowledge. It mirrors the complexity and ambiguity of the much broader seemingly all-encompassing process of transnationalization, of which it is only a very partial manifestation. Seven major types of legal globalization may be distinguished.

2. *Transnationalization of Nation-state Regulation*

Transnationalization of nation-state legal regulation refers to changes in the state law of a given country that have been decisively influenced by formal or informal international pressures by other states, international agencies or other transnational actors. Despite the fact that the impact of the international context on nation-state legal regulation is inherent to the interstate system and can be traced back to the Westphalia Treaty itself (1648), the current process of transnationalization of state legal regulation seems to be a qualitatively new phenomenon for two main reasons. First, it is a very broad and far-reaching phenomenon covering a wide range of state intervention and calling for drastic changes in the pattern of intervention. The core pressure is relatively monolithic as it stems from the 'Washington consensus,' in whose terms the market oriented development model is the only feasible model for a new global regime of accumulation and, accordingly, the structural adjustment it calls for must be carried out worldwide. Other pressures reinforce it, such as the end of the Cold War, the dramatic innovations in communication and information technologies, the new systems of flexible production, the emergence of regional blocs, a newly packaged ideological mix of economic liberalism, and liberal democracy. The scope of these pressures is wider ranging than previous processes of transnationalization, because the current ones take place after decades of active state regulation of the economy in core, peripheral, and semiperipheral countries. The creation of the 'market friendly' model involves massive normative and institutional destruction and construction. The second factor for the novelty of current legal transnationalization is that the asymmetries of transnational power between the core and the periphery of the world system, that is, between the North and the South, are more dramatic today than ever. The sovereignty of the weaker states is now threatened directly not so much by more powerful states, as used to be the case, but rather by international financial agencies and other 'private' transnational actors, such as the multinational corporations.

In the field of the economy the legal transnationalization of state law assumes greatest salience. The policies of 'structural adjustment' cover an enormous range of economic, commercial, and social intervention of the state, provoking turbulence in wide legal fields and institutional settings. Trade liberalization, privatization of industries and services, agricultural liberalization, dismantling of regulatory agencies and licensing mechanisms, deregulation of the labor market and 'flexibilization' of the wage relation, reduction and marketization of social services, less concern with environmental issues, educational reforms geared to job training rather than to citizenship building, family policies that further aggravate the condition of women and children—all these intended or unintended features of the 'Washington consensus' require massive legal changes. Because these changes take place at the end of long periods of state intervention in economic and social life (notwithstanding sizable differences across the world system) the shrinking of the state cannot but be achieved by wide-ranging state intervention. The state must intervene in order not to intervene.

3. *The Law of Regional Integration*

This type of transnationalization of the legal field occurs whenever a group of states combines to create supernational institutions and legal competences which will assume direct regulatory functions that previously did not exist or, if they did, were performed by the individual states as prerogatives of their sovereign powers. Though there have been attempts at regional integration in different parts of the world system, in South and Southeast Asia, in the Middle East and the Magreb, and in the Americas (lastly, the NAFTA), the European Union remains so far the most advanced instance of regional integration and, for that reason alone, deserves to be singled out. Built over four decades by a mixture of diplomacy and interstate democracy, of intergovernment politics and supranational institutionalism, the EU is today an innovative transnational entity both in legal as well as in political terms (see *European Union Law*).

According to the Treaty of Rome, the balance between the market and regulation was to be accomplished by the member states. The harmonization of laws that should have ensued failed, not only because the member states revealed a nationalistic bias, but also because the legislative processes at the European level suffered recurrent bottlenecks. As a consequence, the European Court of Justice, in cooperation with the Commission, had to assume the role of decision maker, a role that has been very active indeed. Performed both with a federalist and a deregulatory bias, it has favored the Union to the detriment of individual members, and the market to the detriment of the state. This development is significant in itself because it reveals a new dimension of the relations between the legal and the political fields. The judicial engineering carried out by the Court from the 1960s onwards decisively affected the European integration process. Through the doctrines of direct effect and supremacy of the European law, the court 'constitutionalized' the community legal structure as a legal order of its own. Much of this occurred without political debate, as a judicial rather than as a political process, which led Weiler (1991) to state the apparent paradox that while European law developed firmly in a continuing process of evolution, the Union went through one political crisis after another.

4. *Global Capital's Own Law:* Lex Mercatoria

Understood as a set of customary principles and rules which are widely and uniformly recognized and applied in international transactions, *lex mercatoria* or law merchant is probably the oldest form of transnationalization of the legal field (see *Lex Mercatoria*). *Lex mercatoria* was a supranational law whose most distinct features were the following: the ease with which it permitted binding contracts; the stress on security of contracts; the speed of adjudication; the variety of mechanisms for establishing, transmitting, and receiving credit; the normative value of customs and usages of the commercial world. Medieval *lex mercatoria* underwent profound changes in the modern period, which indeed caused its seemingly irreversible decline. On the one hand, as trade expanded and more and more diverse trading communities emerged, customs changed and became more diversified themselves. The diversity of customs made *lex mercatoria* less predictable, less transparent, and thus vulnerable to the critique of lacking impartiality. On the other hand, as the modern states gained control over their territories, the existence of a deterritorialized law, which did not derive its normative claims from treaties among sovereign states, was viewed as a threat. As a result, merchant courts were often either forbidden or assimilated into the domestic court system, and transnational business customs were made to comply with law. The expansion of transnational practices and the need to protect them legally led the states to develop a private international law. Since this was a national state law and could therefore conflict with the private international law of other states, efforts were made to harmonize these bodies of laws through the creation of international uniform laws.

The scope of uniform laws has always remained very limited and unable to account for the tremendous growth, in number, complexity, and variety, of transnational contracts and other business transactions, especially after World War II. These relationships, which committed generally larger amounts of money for longer periods than domestic relationships ever did, and engaged partners often separated by great distances, and cultural and linguistic differences, had a greater degree of insecurity and required, therefore, a common normative foundation. Such foundation was sought in a deterritorialized set of normative principles and rules, expressed in such formulas as 'general common principles,' 'principles of equity,' 'principles of good faith and good will,' 'principles of international law,' 'international trade usages,' etc. Because such normative references aimed at circumventing submission to national laws and to the traditional conflicts of law, transnational contracts were deemed to be self-regulatory, subjected only to their own provisions: *contrats sans loi*, *contratti senza legge*, *rechtsordnunglose Verträge*. In fact, a new supranational legal order was emerging. It was the new *lex mercatoria*, which was to expand enormously in the last quarter of the twentieth century, by reason of the sheer intensification of transnational transactions and in the wake of a new worldwide regime of accumulation in search of adequate institutional structures. The new *lex mercatoria* comprises several elements, including general principles of law recognized by commercial nations, rules of international organizations, customs and usages, standard form contracts, and reports of arbitral awards. Though this

is disputed, *lex mercatoria* may be conceived of as comprising also uniform laws and public international law. Concerning the latter, though public international law governs the relations among nation–states, and not among private parties, *lex mercatoria* shares with it the 'general principles of law recognized by civilized nations,' such as *pacta sunt servanda*, *rebus sic stantibus*, and the prohibition of undue enrichment. But the relation between *lex mercatoria* and public international law may also be conflictual whenever the first is used to create forms of immunity which operate both vis-à-vis the national law and the public international law. As regards international uniform laws, once they have been adopted by the nation-states and become part of the national law, they are not, in themselves, part of *lex mercatoria*. Before that process of adoption and to the extent that they represent a crystallization of international trade usages, such laws are *lex mercatoria*.

However informal, the new *lex mercatoria* is neither amorphous nor neutral, the customs and usages being not necessarily universal and much less traditional or immemorial. The new *lex mercatoria*, as an emerging transnational legal field, is constituted by thick cognitive expectations and thin normative allegiances reproduced by the routine repetition of myriads of transnational contractual relations originally framed by transnational corporations and their lawyers, international banks, and international organizations dominated by both. Depending on the power relations between the parties and their stakes in the transaction, the new *lex mercatoria* can operate in either a high rigidity mode (the iron cage mode) or in a high flexibility mode (the rubber cage mode). Even in its rigidity mode, *lex mercatoria* is often ephemeral, an instantaneous customary law, so to speak, as in the case of a new type of contract drafted by a leading transnational corporation and its lawyers, which must be accepted as drafted by the weaker partner, irrespective of the fact that such contract type may not be used again in its entirety. Two issues raised by *lex mercatoria* deserve special attention: *lex mercatoria* and the world system and lex *mercatoria* and legal cultures.

4.1 Lex mercatoria *and the World System*

It has been claimed that transnational contracts are purely contractual in that they contain their own rules of recognition and validation and that, as a result, *lex mercatoria* is nonpolitical, needing no reference to noncontractual elements to sustain itself as a normative order. Grounded on a legal formalistic reasoning, this conception unduly abstracts from the hierarchies and unequal exchanges that characterize the world system. A relative balance of power between partners in the transactions is only rarely obtainable. In most cases, the dominant practices at the basis of this global law are the practices of the dominant actors. The most prominent among these do indeed enjoy state-like political prerogatives, immunity privileges, special access to political resources (tax incentives, special rights on infrastructures, etc.), and they often negotiate directly with the host states the conditions under which the private transactions will be carried out. The political facilitation made possible by the state is a crucial noncontractual dimension of the transnational contracts.

Since stable property rights are exceedingly difficult to establish across national boundaries, the international extension of capital, which is an original feature of the modern world system, has always been an inherently problematic activity. This interstate legal field, created by the core states and imposed on the semiperipheral and peripheral states, has changed substantially in the last 150 years, but has prevailed as a framing, boundary-setting legal and political structure within which *lex mercatoria* has evolved apparently unencumbered by noncontractual constrains. The use of overt or covert interventionary force and economic sanctions by the core countries in the case of expropriations or debt renunciations in the periphery and semiperiphery has been for a long time an integral part of an 'international regime,' closely reflecting the hierarchies in the world system.

4.2 Lex Mercatoria *and Legal Cultures*

The internationalization of capital has always been the motor behind the development of *lex mercatoria*. *Lex mercatoria* is, thus, basically a transnational business law or, more broadly, an economic law. Within the national legal field, economic law traditionally has been one of the laws most permeable to foreign influences and transplants, an area of law in which the ethos of the national legal cultures has always been less determinant. Not by mere coincidence has this area of law been the area of harmonization, uniformization and conventionalization *par excellence*. For this reason, and also because of its deterritorialized character, *lex mercatoria* has often been considered as the expression of a 'global legal culture,' a kind of 'third culture,' independent of the various national legal cultures and hovering above them in its specific domain.

If a new regime of global capital accumulation is indeed emerging, its most visible institutional features, aside from *lex mercatoria* itself, are the rise of transnational corporations and the international organizations that back up their activities, the adaptation of the nation-state regulatory legislation according to the transnational finance capital requirements, the globalization of the market for legal services, and the rise of international commerce arbitration. Particularly visible is the American or American-style law firm. It stands at the core of the

new international market in consultancy and its rise to prominence contributes to the Americanization of international legal services. The type of juridicization it produces is symbiotic with the economic relations it covers and leads toward the infusion of high level competition and market imperatives in the practice of law itself.

5. *Transnational Third Worlds: The Law of People on the Move*

Greatly increased cross-border movements raise myriad sociolegal issues, from international contracts, binational marriages, adoptions of foreign children, legal protection of tourists, and cross-border consumer rights, to civil, political, and social rights of legal and illegal foreign migrant workers, refugees, and asylum seekers. But the international community has paid relatively little attention to the movement of people across national borders, compared with the elaborate regulation of the movements of goods and services. Since traditional international private law only covers a very small number of issues, the international movements of people are in many respects a legal no-man's land. As a result, the legal protection of human beings remains much more territorialized than the legal protection of goods and services. International movements and interactions of people, thus, imply a net loss of legal protection. The challenges posed by this 'black hole' heavily underscore the need to analyze the movements of people across national borders, particularly of those who are less autonomous and run higher risks in moving across borders: migrants and refugees.

5.1 *International Migration*

The slave trade was the first major form of international migration in the modern world system. From the nineteenth century until the 1960s, a period of dramatic emigration from Europe, the major form of international migration emerged based on migrant recruitment through economic inducements. This form of international labor flow was at mid-point between the coerced labor extraction of the slave trade and the self-initiated or spontaneous flows that came to dominate in more recent years.

The state and the national boundaries it controls have, thus, performed crucial roles in the creation of international migration regimes throughout the history of modern capitalism. This control has become more central lately in light of the convergent impact of two recent trends. On the one hand, the trend of core states to favor, if at all, only temporary migration and only in restrictive terms (available only to those whose skills and expertise are in demand), with the consequent tighter control over borders; on the other hand, the increased pressure to emigrate and the potential for large-scale population movements motivated by the struggle for survival resulting from the immiseration of significant regions of the world system and the growing disparities between the North and the South. Whence the potential increase of illegal, clandestine, or undocumented migration and the consequent strengthening of the receiving states as monopoly holders of the legal status of migrants and of migrant life chances and expectations attached to it.

No international consensus has been reached on the possible terms of a compromise between the principles of international human rights applicable to nationals and nonnationals, and to legal and illegal aliens, on the one hand, and the principles of territorial sovereignty, on the other; in other words, no international migration regime has come into existence.

5.2 *Refugees and Displaced People*

As in the case of undocumented migrant workers, refugees concern us insofar as they raise specific new issues in the broad conceptualization of the transnationalization of the legal field outlined here. Such issues have to do, in this case, with two clear-cut distinctions upon which the international refugee regime was built in the postwar period: between voluntary and involuntary international migration, on the one hand, and between economically motivated and politically motivated migration, on the other. These two distinctions have collapsed in recent years, thereby producing new inadequacies between the hegemonic principles of territorial sovereignty and the dynamics of transnational movements.

According to the United Nations Convention of 1951 and 1967, a refugee is any person who, 'owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country; or who, not having a nationality or being outside the country of his former habitual residence as a result of such events, is unable or, owing to such fear, is unwilling to return to it.' This definition excludes most of the world's displaced populations who are mainly victims of massive conflicts, human rights abuses, civil wars, external aggression and occupation, foreign domination, etc. Since they have not been personally persecuted and often have not crossed borders, they are not entitled to the refugee status, even though the political nature of the causes of their flight offers no doubt. They are considered internally 'displaced persons' and at the most may receive some humanitarian assistance. The core countries repeatedly have expressed their opposition to an expansion of the concept of refugee; in recent years, when

confronted with the potential rise of applications for asylum, they have even adopted more restrictive legislation on (or rather, against) refugees and asylum seekers.

The last three decades of the twentieth century have witnessed other 'involuntary' flows of people still more at odds with international and domestic refugee legislation: hundreds of thousands of people fleeing from famine, starvation, and natural disasters (droughts, earthquakes, volcano eruptions). In many cases extreme poverty and environmental crises have acted together with civil war or political repression to provoke massive displacements of people. In these dramatic situations the distinction between economic and political factors have been blurred, if not totally dissolved.

One type of 'involuntary' transnational migration is most likely to grow in the coming years. It will be caused by environmental catastrophes, directly or indirectly caused by anthropogenic climate changes and often coupled with over-population. Refugees will be environmental refugees, seeking ecological asylum. Environmental refugees occupy the outer fringes of a generally precarious system of international protection. Environmental refugees are most likely to collide with the territorial logic underlying social protection as a matter of right.

The high vulnerability of refugees derives from a double adverse condition. On the one side, international human rights applicable to them are highly restricted by a highly restrictive definition of refugees that leaves out most of the current and massive situations of forced migration. On the other side, the few rights they are entitled to are grossly violated with almost total impunity. Refugees are often confined to camps and detention centers where, aside from the lack of freedom of movement and access to the outside world, refugees are subjected to the most degrading human conditions.

6. *Ancient Grievances and New Solidarities: The Law of Indigenous People*

The collective rights of indigenous peoples, is surely an old issue—at least, as old as European expansion from the fifteenth century onwards—which, however, gained a new prominence in the last quarter of the twentieth century. This new prominence is due to a transnational-local linkage which has to do with the most recent step of the secular trend of capitalist world economy to incorporate the whole planet, even its most remote regions, in the process of capital accumulation. As a result, many indigenous peoples saw their traditional lifestyles, customary rights, and peasant economies threatened by new or more powerful development claims by multinational corporations or the national state or, in most cases, the two combined. They represent another and probably more powerful wave of imperial globalization. One of its specific features is the extent to which the national state itself intervenes as a transnational actor. Pressed by the 'Washington consensus' on debt, stabilization, and structural adjustment, the peripheral and semi-peripheral states embark in new 'development projects,' often in joint ventures with transnational corporations and with the support of international financial institutions, be they extractive—oil, minerals, biodiversity—agro-industrial, road construction, or hydroelectric projects. Some of these projects are to be carried out on indigenous lands, further destroying their traditional lifestyles and economies, forcing them into new and massive relocations.

What is truly new in the new prominence of the indigenous question is the emergence of a transformative, transnational coalition consisting of indigenous and nonindigenous NGOs, which have been drawing the attention of world public opinion to the violations of indigenous historical rights and pressing both international organizations and the national states to condemn the violations, stop the destruction, and take active measures to effectively right (or begin to right) such massive historical wrongdoing. The struggles of the indigenous peoples thereby are transnationalized. The legal dimension of this transnational transformative action lies in the struggle for the transnational recognition of indigenous collective rights and local autonomous legal regulation within territorially defined areas.

7. *Cosmopolitanism and Human Rights*

In the post-World War II period a new cosmopolitan, legal culture emerged out of a transnational understanding of human suffering and the transnational constellation of progressive (legal, political, humanitarian) social actions devised to minimize it; it has gradually evolved into an international human rights regime supported by transnational coalitions of local, national and transnational nongovernmental organizations, which have been growing dramatically, in number, range, and effectiveness in recent years.

The international human rights regime is built around three major documents and sets of standards which together comprise 'the global human rights regime,' a system of rules and implementation procedures centered on the United Nations, and whose main organs are the UN Commission on Human Rights and the Human Rights Committee.

This global human rights regime allows for internal differentiation across the interstate system. Four regional subregimes usually are identified and ranked as follows by decreasing order of strength: the European, the Inter-American, the African, and the Asian and Middle-Eastern (lack of) regime. What distinguishes the *European regime*—established by the members of the Council of Europe in 1950—is the very

strong monitoring powers of the European Commission of Human Rights and the binding decision-making powers of the European Court of Human Rights. Without questioning the overall strength of the European human rights regime when analyzed in comparative terms, one of its weaknesses, which is indeed bound to expand in the coming years, refers to the massive violations of human rights that may be in store for noncommunity citizens, as the internal borders are eliminated and the external borders strengthened.

The *Inter-American regime* is substantively narrower than the European one. Similarly to the European regime, its main organs are the Inter-American Commission of Human Rights and the Inter-American Court of Human Rights. In practice, however, the two institutions have been operating very differently from their European counterparts. As to the *African regime*, the African Charter on Human and Peoples' Rights (1981) contains two major innovations: as the title of the Charter itself indicates, to individual human rights it adds collective rights, as well as the right to development; it also introduces the concept of duty of individuals vis-à-vis the family, the community, and the state, but the provisions that establish the African Commission on Human and Peoples' Rights are exceedingly vague, conferring it no enforcement powers of any sort and only very modest investigatory powers, and there is no regional human rights court. Though in *Asia* and *The Middle East* the human rights debate is widening and deepening, there are, among other reasons, important cultural barriers to the establishment of a human rights regime. Human rights are often viewed as an exotic, foreign, Western conception whose worldwide circulation is nothing but a manifestation of the overall cultural imperialism of the West.

International regimes refer to international normative consensuses among nation-states and, as such, they raise two interrelated questions: the extent to which the normative consensus collapses whenever the overriding imperatives of national sovereignty are considered to be better served by the violation of human rights; and the extent to which the inherent statism of implementation and enforcement mechanisms stays in the way of the emergence of new international legal subjects with a more cosmopolitan orientation and a transnational advocacy of greater efficiency. The optimist predictions of a brighter future for human rights, in the wake of democratic transitions and the collapse of the communist bloc, may prove to be displaced and utterly wrong themselves.

8. *The Global Commons:* Jus Humanitatis

Of all forms of transnational legality *jus humanitatis* is the one that pushes the idea of globalization further, for it takes the globe itself as the object of its regulation. *Jus humanitatis* expresses the aspiration to a form of governance of natural or cultural resources which, given their extreme importance for the sustainability and quality of life on earth, must be considered as globally owned and managed in the interest of humankind as a whole, both present and future. In this sense, *jus humanitatis* clashes with two fundamental principles of the dominant paradigm: property, upon which the capitalist world system is based; and sovereignty, upon which the interstate system is based. Little wonder, then, that its application has been so little so far. Nevertheless, its prolegomena are already quite visible in the doctrine of the *common heritage of humankind* that has been adopted by international law. Though the doctrine has its own limitations, it is grounded on principles which, were they to be fully developed, would bring about the bankruptcy of the dominant paradigm, and hence of international law itself as it is conceived today.

The concept of the common heritage of humankind was formulated for the first time in 1967 by Malta's Ambassador to the United Nations, Pardo (1968), in relation to UN negotiations on the international regulation of the oceans and the deep seabed. Pardo's purpose was 'to provide a solid basis for future worldwide cooperation … through the acceptance by the international community of a new principle of international law … that the seabed and ocean floor and their subsoil have a special status as a common heritage of mankind and as such should be reserved exclusively for peaceful purposes and administered by an international authority for the benefit of all peoples.' Since then, the concept of the common heritage of humankind has been applied not only to the ocean floor but also to other 'common areas' such as the moon, outer space, and Antarctica. Five elements usually are associated with the concept of the common heritage of humankind: nonappropriation; management by all peoples; international sharing of the benefits obtained from the exploitation of natural resources; peaceful use, including freedom of scientific research for the benefit of all peoples; conservation for future generations.

The concept of the common heritage of humankind transcends by far the field of traditional international law. International law deals traditionally with international relations among nation-states, which are supposed to be the main beneficiaries of the regulation agreed upon. Such relations are based chiefly on reciprocity, that is, granting advantages to another state or states in return for equivalent advantages for oneself. The concept of the common heritage of humankind is different from traditional international law on two accounts: as far as the common heritage of humankind is concerned, there is no question of reciprocity; and the interests to be safeguarded are the interests of humankind as a whole, rather than the interests of states. The Law of the Sea Convention, signed in Montego Bay on December 10, 1982,

contains in part XI the most developed formulation of the concept of the common heritage of humankind to date.

See also: Global Governance; Globalization and World Culture; Globalization, Anthropology of; Globalization: Geographical Aspects; International Justice: Philosophical Aspects; International Law and Treaties; International Relations, History of; International Relations: Theories; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; Markets and the Law; Science and Technology: Internationalization; World Systems Theory; World Trade Organization

Bibliography

Bilder R 1992 International law in the 'New World Order': Some preliminary reflections. *Journal of Transnational Law and Policy* **1**: 7–21

Brysk A 2000 *From Tribal Village to Global Village. Indian Rights and International Relation in Latin America*. Stanford University Press, Stanford, CA

Cohen R, Deng F 1998 *Masses in Flight. The Global Crisis of International Displacement*. Brookings Institution Press, Washington, DC

Dezalay Y, Garth G B 1996 *Dealing in Virtue. International Commercial Arbitration and the Construction of a Transnational Legal Order*. University of Chicago Press, Chicago

Donnelly J 1992 Human rights in the New World Order. *World Policy Journal* **9**: 249–77

Draetta U, Lake R, Nanda V 1992 *Breach and Adaptation of International Contracts. An Introduction to Lex Mercatoria*. Butterworth, Salem, NH

Harvey C J 2000 Dissident voices: Refugees, human rights and asylum in Europe. *Social & Legal Studies* **9**: 367–96

Hathaway J C (ed.) 1997 *Reconceiving International Refugee Law*. Kluwer, The Hague

Joyner C 1986 Legal implications of the concept of the common heritage of humankind. *International and Comparative Law Quarterly* **35**: 190–9

Keck M, Sikkink K 1998 *Activist Beyond Borders. Advocacy Networks in International Politics*. Cornell University Press, Ithaca, NY

Kiss A 1985 The common heritage of mankind: Utopia or reality? *International Journal* **40**: 423–41

Pardo A 1968 Whose is the bad of the sea? *American Society International Law Proceedings* **62**: 216–29

Picciotto S 1997 Fragmented states and international rules of law. *Social & Legal Studies* **6**(2): 259–79

Rogers R 1992 The politics of migration in the contemporary world. *International Migration* **30**: 33–52

Santos B de Sousa 1997 Toward a multicultural conception of human rights. *Zeitschrift für Rechtssoziologie* **18**: 1–15

Sassen S 1999 *Guest and Aliens*. New Press, New York

Tigar M, Levy M 1977 *Law and the Rise of Capitalism*. Monthly Review, New York

Van Cott D L 2000 A political analysis of legal pluralism in Bolivia and Colombia. *Journal of the Latin American Studies* **32**: 207–34

Weiler J 1991 The transformation of Europe. *Yale Law Journal* **100**: 2403–83

B. de Sousa Santos

Globalization: Political Aspects

Globalization is the process by which the economies of many different nations and regions are being transformed into a single, integrated economy which is worldwide or global in scope. The elimination of national barriers and the expansion of international transactions are being carried to their ultimate conclusion, where the exchange of goods, capital, and services occurs within a global market. This profound economic transformation is having great consequences for national states and their societies. Some nations and groups are winners in globalization and some are losers. The result is that globalization is accompanied by major political actions and by intense intellectual debate.

1. The Term 'Globalization'

The term 'globalization' came into common usage in the early 1990s. The decline of Marxism, the end of the Cold War, and finally the collapse of the Soviet Union in 1991 meant that there was no longer any formidable ideological and political competitor to liberal democracy and the free market. The end of the socialist or Second World meant that the capitalist or First World could expand into the entire world or globe, hence 'globalization.' Also in the 1990s, new information technologies reached the critical mass where they became an entire information economy. The conjunction of computers, telecommunications, the internet, and the worldwide web greatly facilitated and expanded the exchange of goods, capital, and services on a global scale.

Globalization is often described as a process: steadily progressing over time, pervasively spreading over space, and clearly inevitable in its development. In this conception, the economics of globalization are all-important; the politics of globalization are practically nonexistent. But globalization can also be seen as a revolution, one of the most profound revolutions the world has ever known. (It is certainly more of a world revolution than communism ever was.) Revolutions produce reactions, however, and the interactions between revolutionaries and reactionaries produce a particular kind of politics. The first decades of the twenty-first century are likely to be dominated by the politics of globalization, so much so that they could give the name to the era, in the same way that the first

half of the twentieth century was known as the era of the World Wars and the second half was known as the era of the Cold War.

2. Previous Eras of Globalization

The contemporary period is not necessarily the first era of globalization; in some ways it can be seen as the third. There have been at least two previous periods when there was a rapid and massive expansion in economic exchange on a world, even global scale. Consideration of these earlier eras may help in understanding the current one.

The first era of globalization was in the sixteenth century, at the time of the commercial revolution and when European powers opened up the Americas, Africa, and Asia to new or greatly expanded trade. Facilitated by new technologies in shipbuilding and navigation, a vast system of seaborne commerce in goods, capital, and slaves came into being, connecting five continents. This first global economy did not come into being by the spontaneous operations of the free market, however. It had to be promoted and protected by a great political and military power, one that was worldwide in scope. This global hegemon was the Habsburg empire, whose centers were Spain and Austria but which ruled large territories on or near the five continents (Wallerstein 1974, 1980).

The first wave of globalization washed over many old and traditional political authorities. Many feudal and local entities were diminished or disappeared entirely, but others reinvented themselves to become absolutist monarchies and national states. The process was not steady and gradual, but disruptive and violent. It contributed to the Wars of Religion and culminated in the catastrophe of the Thirty Years War (1618–48). At the end of that war and with the Peace of Westphalia, the major surviving powers created an entirely new political order, one composed of sovereign and independent states. Most of these states now developed systematic policies (known as mercantilism) to protect local interests from foreign imports. Thus, the first era of globalization ended in what can be seen as a new era of localization.

The second era of globalization was in the nineteenth century, at the time of the industrial revolution and when European powers and the USA opened-up even larger (now inland as well as coastal) areas of the Americas, Africa, Asia, and Australia to new or greatly expanded trade. Facilitated by the steamship, the railroad, and the telegraph, a vast system of seaborne and overland commerce in goods, capital, and immigrant labor connected six continents. This second global economy also did not come into being by the spontaneous operations of a free market (although the economic theories of the time often assumed that it had). It had to be promoted and protected by a great political and military power that was worldwide in scope. This global hegemon was the British Empire, which in the late nineteenth century ruled one-fourth of the land surface of the globe and which was the supreme naval power in every ocean (Polanyi 1944).

The second wave of globalization also washed over many old and traditional political authorities. Again, many local entities were diminished or disappeared entirely (particularly in Africa, the Middle East, and East Asia), but others reinvented themselves to become strong national states (particularly in Europe but also Japan). The process was not steady and gradual, but disruptive and violent. It contributed to a series of nationalist wars and culminated in the catastrophic World War I and II, which some historians have called a second thirty years war (1914–45).

At the end of that war and with the onset of the Cold War, a bipolar system was created, composed of the USA (liberal–democratic in its politics and free-market in its economics) and the Soviet Union (authoritarian in its politics and statist in its economics). The Americans promoted free markets and open economies in their alliance system (the Free or First World); the Soviets promoted command and closed economies in theirs (the Socialist or Second World). The existence of the Soviet Union and the Second World acted to contain the spread of the free-market policies of the USA, even within the First World itself. Thus, the second era of globalization ended in another era of economies that were closed or protected in important ways.

3. The USA as the Leader of Globalization

We are now in the third era of globalization. With the collapse of the Soviet Union, the USA could pursue its grand project of promoting and protecting free markets throughout the globe, unconstrained and uncontained by any comparable power or 'peer competitor'; it has been the leader and the energizer of the process of globalization. It serves as the new global hegemon, the successor to the Habsburg and the British Empires. Acclaimed as 'the sole superpower,' the 'high-technology economy,' and 'the universal nation,' the USA has vigorously and consistently supported this great transformation by systematically pressing to remove any national barriers to the free movement of capital, goods, and services. Its leadership is facilitated and legitimated by the three great international, now global economic institutions, the International Monetary Fund (IMF), the World Bank, and the World Trade Organization (WTO).

US leadership in globalization is also facilitated and legitimated by the particular American worldview or ideology. Americans have long advocated liberal

democracy and free markets as political and economic ideals. In foreign affairs, this ideology has been known as liberal internationalism, and its most prominent proponents were Presidents Woodrow Wilson and Franklin D. Roosevelt (T. Smith 1994). This familiar ideology has been revised in the era of globalization into a new American ideology which promotes human rights and the open society. This is a version of liberal internationalism, one that is more accurately termed liberal globalism; its most prominent proponents in the 1990s were President Bill Clinton and Secretary of State Madeleine Albright, but it is held by most American political, economic, and intellectual elites. The ideology of human rights and the open society advocates the limitation of state sovereignty and the weakening of the nation state. As such, it poses a fundamental challenge to traditional conceptions of international relations.

This ideology of human rights and the open society—liberal globalism—also provides the justification for a new kind of US military operation. This is humanitarian intervention, as in Somalia, Haiti, and Bosnia; the culmination of this development was the US-led NATO war against Serbia over Kosovo. Intervention in support of human rights is not supported by all parts of the American public; it is primarily advocated by political and intellectual elites, especially those in the media and in academia. Political opposition to humanitarian interventions and more generally to globalization comes from other American groups and interests, especially those who see themselves as losers in the global economy (e.g., labor unions and domestic industries). But the major sources of political opposition to globalization are found outside the USA.

4. *The Resistance to Globalization*

The US-led revolution of globalization has generated resistance and reaction. Immediately after the collapse of Soviet communism, the principal sources of resistance seemed to be major nation states, especially Japan, France, and reunited Germany. However, by the end of the 1990s these nations offered little resistance to globalization; one reason was the stagnation of their economies, compared with that of the USA.

With the paralysis of the major nation states, the principal sources of resistance to globalization have become several of the great religions. Some of these have undergone a revival and transformation into a political religion. Especially strong in their resistance have been revivalist Islam and neo-Confucianism (sometimes referred to as 'Asian values'). Also resistant have been revivalist Hinduism and Eastern Orthodoxy. The great religions see the globalization led by the USA to be closely connected with secularization and therefore a threat to themselves. These four sources of religious opposition correspond to four large countries or areas. Two of these, Russia and the Middle East, have been losers in the new global economy, and two, China and India, have been winners.

Russia and, more generally, most countries with an Eastern Orthodox religious tradition, for a variety of reasons have been unsuccessful in making the transition from a communist regime to a liberal democracy and market economy, which are able to adapt well to an open society and global economy (e.g., Russia, Belarus, Ukraine, Moldova, Romania, Bulgaria, and Serbia). Conversely, most countries with a Roman Catholic or Protestant tradition have been successful in making this transition. This contrast among ex-communist countries, between the more western and the more eastern, between the Roman Catholic or Protestant and the Eastern Orthodox, means that the political and economic developments of the 1990s revived and reinforced an historic divide, one that corresponds to the great schism between Western Christianity and Eastern Christianity and even to the ancient division between the Latin and the Greek halves of the Roman Empire.

With their condition of political and economic weakness, the governments of Russia and the other Orthodox countries cannot mount an effective and sustained opposition to globalization. But among the populations of these countries, there is now substantial resentment and resistance directed at the USA and its global project. The Middle East and, more generally, most countries with an Islamic religious tradition have also been unsuccessful in establishing a viable liberal democracy and market economy, which are able to adapt well to an open society and global economy. For those that experienced the heady years from 1973 to 1985, when OPEC seemed to be a world power, a long decline in oil prices deepened their sense of failure and their alienation from globalization. Here, too, the governments are not willing and able to mount opposition that is effective and sustained. But among their populations there is widespread resentment and resistance, whose most obvious representatives are militant organizations based upon Islamic revivalism.

Conversely, there is growing opposition even from some of the winners in the global economy. These include China and, more generally, most countries with a Confucian tradition. Most Chinese, both in China itself and among the 'overseas Chinese,' see their economic success as resulting from their own culture or 'Asian values,' from engaging in the global economy in their own way. This conception has been reinforced by their interpretation of the Asian economic crisis of the late 1990s. The countries with the most open currency markets (South Korea, Thailand, Indonesia) suffered the greatest disruptions and declines. Those whose currency markets were most regulated (China and, ironically, Taiwan) experienced little disruption and continued their economic growth.

The entry of India into the global economy has been more recent, and its benefits have been unevenly distributed. It has occurred, however, at the same time as the growth of Hindu nationalism, the development of Indian nuclear weapons, and an increased confidence in rejecting certain Western ways. The sources of opposition to globalization led by the USA are thus very disparate. But together these sources of opposition comprise a vast region, really all of Eurasia and more, which stretches from Russia and Eastern Europe, through the Middle East, through South Asia, to China and East Asia. In this vast region are found the four great civilizations that Samuel Huntington has identified as most likely to oppose the West in 'the clash of civilizations' (Huntington 1996). These are, in his terms, 'Slavic–Orthodox,' Islamic, Hindu, and 'Sinic–Confucian' civilizations. In this region, too, are four nuclear powers, Russia, Pakistan, India, and China, each of them seeing themselves as the center of their civilization. Singly or collectively, they may develop the capacity and will to disrupt and contain globalization. If so, it will take American leadership of the highest ability to insure that the third era of globalization does not end with a clash of civilizations which culminates in a third thirty years war.

See also: Cold War, The; Global Governance; Global Population Trends; Globalization and World Culture; Globalization, Anthropology of; Globalization: Legal Aspects; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; International Relations, History of; World Systems Theory

Bibliography

Cox R W (ed.) 1996 *Approaches to World Order*. Cambridge University Press, Cambridge, UK

Eichengreen B 1996 *Globalizing Capital. A History of the International Monetary System*. Princeton University Press, Princeton, NJ

Friedman T 1999 *The Lexus and the Olive Tree*. Farrar, Straus, and Giroux, New York

Gray J 1998 *False Dawn: the Delusions of Global Capitalism*. New Press, New York

Held D, McGrew A, Goldblatt D, Perraton J 1999 *Global Transformations: Politics, Economics, and Culture*. Stanford University Press, Stanford, CA

Hoekman B M, Kostecki M M 1995 *The Political Economy of the World Trading System: From GATT to WTO*. Oxford University Press, Oxford, UK

Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon & Schuster, New York

Kapstein E B 1994 *Governing the Global Economy: International Finance and the State*. Harvard University Press, Cambridge, MA

Luard E 1990 *The Globalization of Politics*. Macmillan, London

Polanyi K 1944 *The Great Transformation*. Beacon Press, Boston

Smith A D 1995 *Nations and Nationalism in a Global Era*. Polity Press, Cambridge, UK

Smith T 1994 *America's Mission: The United States and the Worldwide Struggle for Democracy in the Twentieth Century*. Princeton University Press, Princeton, NJ

Strange S 1996 *The Retreat of the State: the Diffusion of Power in the World Economy*. Cambridge University Press, Cambridge, UK

Wallerstein I 1974 *The Modern World-System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

Wallerstein I 1980 *The Modern World-System II: Mercantilism and the Consolidation of the European World-Economy 1600–1750*. Academic Press, New York

J. Kurth

Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity

Globalization, the historical process by which all the world's people come to live in a single social unit, brings with it religious changes as well. Prime among these is that globalization encourages religious pluralism such that religions identify themselves in relation to one another. Religions also become less rooted in particular places because of diasporas and transnational organization. In postmodern fashion, global religious pluralism creates a situation in which absolute visions end up living side by side.

1. Globalization and Religion

Since the 1980s, the term globalization has taken on a certain currency in social scientific as well as in popular discourses. Its meaning in both these areas varies; but what unites economic, political, cultural, and even theological conceptions of globalization is the sense that, for better or worse, all people around the world live today in a single social unit much more than they ever have before. There is now and has been developing for some time a world society which serves as the encompassing social context of all more local social units. In this regard, most of the debates about globalization, whether scholarly or popular, have centered around the relation and even polarity between the global and the local. Robertson (1992), perhaps the first social scientist to use the term globalization in a technical and consistent sense, talks about the dialogical relation between the universal and the particular. What this means is that the global

context, rather than only leading to a certain homogenization of culture and social structure around the world, also provides new conditions for constructing social and cultural difference. Global society, while single, is inherently pluralistic, not only permitting but even favoring a multitude of local and particular identities, both individual and collective.

The status of religion in debates about globalization has varied, to a significant extent according to academic discipline. Perspectives that see the development largely in economic terms, for instance as the increasing dominance of global capitalist institutions, tend to ignore religion almost entirely. Many of the approaches originating in the fields of international relations, political geography, and communication studies also pay little attention to religion, with the occasional exception of the role of Islam in many Muslim states (see, e.g., Anderson et al. 1995). Overall, it is primarily the works of certain anthropologists and sociologists that accord religion a more visible place. Yet even here, analyses vary a great deal, for the most part depending on how the term is understood. Two overlapping lines of variation appear to be central: the distinction between functional and substantive notions of what counts as religion and whether one sees religion as an aspect of culture or as a differentiated and independent social domain.

A functional conception of religion regards as religion any social process that addresses certain problems, for instance the question of the ultimate meaning of life, or social integration and control. Here religion, or at least the religious, almost automatically becomes a necessary dimension in the formation of world society because problems of meaning or of self-description (integration) have to be addressed in any society, especially one as diverse as the global one. As examples under this heading, some authors have sought to analyze the emergence of a global culture that is the equivalent of a new global 'sacred canopy' (Meyer et al. 1997). One might even include a growing discussion about the development of a global 'civil society' under this heading (Rudolph and Piscatori 1997, Braman and Sreberny-Mohammadi 1996). In addition, a functional approach allows questions of national, regional, ethnic, or civilizational identity in world society to take on religious character (Huntington 1996, Robertson 1992); and thus will tend to see religion largely as an aspect of culture. Substantive ideas of religion, by contrast, insist on the presence of identifiable and institutional sorts of activity that refer to a posited supraempirical, transcendent reality. From this perspective, the relation of various specifically religious groups to globalization becomes the focus of attention, groups ranging from the Roman Catholic Church and Islamic organizations to Western Buddhist groups and Pentecostal movements (Cox 1995, Robertson 1992, Rudolph and Piscatori 1997). Such approaches perforce also see religion as an independent domain of social activity and as more than an expression of something else, whether economics, politics, or culture (see Beyer 1998, Hexham and Poewe 1997).

This variety in ideas about what counts as religion in the global context by itself already points to important aspects of the relation between globalization and religion. The fact that most of the literature which deals explicitly with globalization ignores religion or functionalizes it suggests that substantive or institutional forms are for these observers somewhere between irrelevant and antiglobal. One manifestation of this orientation is the great amount of attention that so-called 'fundamentalisms' receive, and this, ironically, very often from scholars who are oriented quite positively toward institutional or substantive religion. Thus, e.g., the volumes issuing from Appleby and Marty's *Fundamentalism Project* (see Marty and Appleby 1994), although not directly about globalization, nevertheless assume globalizing processes. They judge a wide variety of important religious movements all around the contemporary world as particularistic and usually local reactions against them (see also Juergensmeyer 1993, cf. Beyer 1994). Among some of these authors, however, designating 'fundamentalisms' as, in effect, antiglobal movements parallels a positive attitude to the idea of a single world society, especially as manifest in the search for a specifically global ethos. In these cases, the search is on an explicitly religious basis, albeit not on the basis of a particular religion. What is somewhat ironic is that the resulting global ecumenical ethic (see especially Küng 1991) corresponds reasonably well to the intention of those who functionalize religion at the global level: i.e., it avoids reference to a particular religion while addressing what it assumes to be the urgent and globally relevant religious task of providing ethical ideals and values that people all around the world can accept. In short, this sort of discussion seems to suggest that the distinction between functional and substantive conceptions of religion overlaps in a certain way those between universal/global and particular/local. Since these latter are principal organizing polarities in the entire globalization debate, it is best not to decide for one conception of religion or the other; but rather to consider the potential insights that flow from both.

2. *Theoretical Perspectives*

An excellent example of how a functionalized conception of religion can enter the globalization debate at the theoretical level is to be found in the work of Meyer and his co-workers. In a 1997 article that clearly presents their basic outlook, they carry to the level of global society an analysis that the classical sociology of Max Weber and Emile Durkheim made for modern societies. For these scholars, the world polity, as they call it, is an extension of modern

Western values and their corresponding rationalized structures to the rest of the world. Western religion had a critical role in bringing about this modernization and rationalization; but today the old religious elites have lost their power and have in any case also become believers in salvation through the pursuit of modern global values, especially progress (development) and equality. In their place there is now a new 'religious' elite who are 'the professionals, researchers, scientists, and intellectuals who write secularized and unconditionally universalistic versions of the salvation story, along with the managers, legislators, and policymakers who believe the story fervently and pursue it relentlessly. This belief is worldwide and structures the organization of social life almost everywhere' (Meyer et al. 1997, p. 174). Meyer and his co-workers support the existence of this world polity by showing how the concrete policies of most countries around the world conform to its goals and values. In arguing in this fashion, however, they present a rather alternativeless and homogenizing picture which leaves room neither for the assertion of any meaningful particular or local variation nor, not coincidentally, for any role for institutional religion other than as futile reaction or powerless supporter.

A signficantly more nuanced vision, both as regards the question of homogeneity/heterogeneity and the possible role of institutional religion, is contained in the work of Robertson. This author also emphasizes what he calls the 'globalization of societal modernity' (Robertson 1991), which includes the sort of global religious function that Meyer and co-workers see. In addition, however, Robertson points much more positively to the proliferation of religious movements in the contemporary world, movements ranging from Latin American liberation theology and Küngian 'ecumenism' to Islamic 'fundamentalism,' Japanese Soka Gakkai, and the Unification Church of Sun-Myung Moon. His theoretical model makes room for these more institutional and substantive forms of religion, and not just as local defensive reactions against incorporation into the global system. Instead, Robertson allows that they can also be attempts to define the 'global–human' situation as a whole; and at the same time ways of identifying a particular culture, region, or (in his terms) society in relation to the global whole and other subunits of it. The task of constructing such universalistic religious visions—whether functional or substantive—is always done on the basis of particular traditions, cultures, or movements. In consequence, the forms of institutional religion themselves change in the context of globalization, leading, for instance, to revisions of old traditions, syncretic or hybrid movements, and the rise of entirely new religions. For Robertson, then, the World Parliament of Religions, the Vishwa Hindu Parishad (World Hindu Organization), the Church of Scientology, a peripatetic Pope or Dalai Lama, and Chinese Qi Gong groups are just as much aspects of globality and contributors to the shape of global society as are the aims and pronouncements of social scientists and technocrats. They all reflect and share a common social context; and one is not more modern or global than the other.

3. *Religious Pluralism*

From these theoretical starting points, one of the important questions that emerges concerns religious pluralism. Only with a radically functionalized concept of religion, such as that of the world polity school, can it appear possible for global society to share a single religious vision, a single world view. All other conceptions point to the near impossibility of a hegemonic, global religion. Substantively speaking, contemporary global society shows no sign of moving in this direction. If anything, the opposite is the case: not only is there an increasing number of identifiable religions or religious traditions in the world, the variety of expressions within those religions is also growing. Indeed, some sociologists of religion, notably Luckmann (1967), have long suggested that the typically modern—and therefore now global—form of religious expression is a highly individualized one that carries pluralism to the point where Luckmann can call it 'invisible religion.' Put somewhat differently, the question of religious pluralism is not simply one of how different religions, each one claiming an absolute vision, can live side by side in world society; but also to what extent this society still favors authoritative institutions worthy of the name, religions. A partial answer is possible through a comparison of the idea of religion with those of culture and nation.

The model of global culture argues that, in the contemporary social world, the ideas of culture and nation serve as names for types of legitimate and fundamentally plural collective identities: cultures and nations, like individuals, have the right to equal existence. Within the context of homogenizing forces like global capitalism and mass media, one can assert one's difference and defend that difference in the name of one's culture or nation. Cultures and nations, like individuals, are legitimated agents or actors in global society: it is part of the world polity of which Meyer and co-workers speak, that their right to 'self-determination' should only be limited by the right of other such identities to assert themselves in like fashion (see Lechner 1989). Religion is both a similar and different concept. On the one hand, religion is often a part of cultural and national identities: it can help define them. On the other hand, global culture also allows that religion is something independent of these other two notions. Religions are something to which people assumedly adhere voluntarily as individuals; and this may or may not flow from their cultural or national identities. Religions also enjoy the same 'protection' as cultures and nations; at least such

freedom of religion is an ideal expressed in most of the world's state constitutions. Correspondingly, global cultural ideals regard religions, like cultures and nations, as inherently plural. Two seeming paradoxes emerge from this view. First, religious freedom is grounded in the individual's right to choose, but also in the right of the religious collectivity to exist and determine itself. Religious collectivities can therefore claim authority over their members, but only with the consent of these latter! Second, religions are particular and plural; and yet it is understood that they each offer a universal or cosmic vision, a total way of life. This holist emphasis therefore includes the possibility of exclusivist religions, ones that deny the validity of all others. Thus, in true postmodern fashion, religions provide ways of multiplying ultimate narratives about how the world is and they do this in such a way that each is regarded as of equal value, the choosing of one or the other being deemed a matter of preference, or upbringing, or culture. Ultimately, each is grounded only in the conviction of its adherents, and this in the knowledge that most others in global society have different convictions: even by the most generous estimates, the largest religion, Christianity, can claim far less than half the world's population.

4. *Religion in Postmodernity*

The postmodern character of religion in the context of world society is manifest in other features as well. Embedded in the idea that religion is fundamentally plural are the questions of how religions distinguish themselves from one another and what the difference is between religion and nonreligion. Neither question is susceptible to a simple and clear answer. The idea that there are a plurality of distinct religions is one that dates historically only from the European seventeenth century in the context of post-Reformation religious conflicts and as European expansion brought its people into protracted contact with a much wider variety of other cultures, civilizations, and hence religious traditions. Which traditions emerged from this historical process as 'one of the religions' depended both on what Westerners observed as religions and on what non-Westerners were willing to style according to the implied model of a religion. Different and even diametrically opposite outcomes resulted. Thus, the elite carriers of Hinduism took up the suggestion of Western observers and cooperated in reimagining South Asian religious traditions as somehow a single religion, whereas corresponding Chinese intelligentsia refused, in spite of Western observation, to reconstruct their Confucian thought and ritual traditions as a religion (see Beyer 1998). The contemporary list of (world) religions therefore has an inductive and even arbitrary character that cannot be reduced to abstract characteristics. Like nations and cultures, what comes to be seen as one of the religions depends in large part on how well the supposed members succeed in having others accept it as one.

The boundary between religion and nonreligion exhibits comparable ambiguities, the difficulty of deciding between functional and substantive ideas of religion being one example. At the core of this problem lies the notion that religion refers to an overarching view of the world and thus to a complete way of life; and yet religions are, in principle and in fact, plural. A prime symptom of this ambiguous status of religion in world society is that, over the past century and a half, key representatives of almost all the so-called world religions have claimed that 'ours is not a religion.' Most often, protests of this sort favor a broad functional conception of religion for their tradition, and reject the sort of substantive one that limits religion to a special category of human action, say one dealing only with the interaction with gods or spirits. Similarly, opponents of various new religions have sought to exclude these from the category, in these cases so as to eliminate them from the range of legitimate religious choices. Such conflicts over categorization are, on occasion, ways of circumventing the global cultural value of 'freedom of religion,' justifying in these instances the obligatory adherence of everyone to the tenets and practices of a particular 'religion'; or preventing such adherence to another. Thus, does the category of religion in global culture allow a postmodern juxtaposition of absolute narratives without offering an overarching narrative through which to reconcile them.

5. *Religious Authority, Transnational Organization, and Diaspora*

Accompanying these various ambiguities regarding the category of religion and religions, the carrying out of institutional religion under conditions of globalization brings a number of other challenges and changes, especially with respect to questions of religious authenticity and authority. In any society, there exist a limited number of ways through which conformity to authoritative religious traditions can come about. Thus, e.g., in small scale societies a relative lack of alternatives, and social structures centered on face-to-face interaction could assure reasonable religious conformity. After the advent of writing and large scale political empires, dedicated priesthoods or monastic orders, political state support, along with distinctions between elite and popular religiosity (e.g., religion vs. superstition) could handle the problem according to hierarchical principles. Under conditions of globalization, neither relative social isolation nor hierarchical solutions are possible. Globalized communication media, from telephones and air travel to television and the internet assure that religious social networks need not be limited to local interaction or particular places. In fact, the existence

of these possibilities positively encourages global or translocal religious networks and religious organizations. Moreover, the global cultural valuing of equality for both individuals and collectivities severely militates against the notion that the orthodoxy practiced or advocated by a certain group of religious professionals or adherents in a certain place should serve as the unquestioned standard for all others. Even in religions that have traditional centers of such authority, e.g., Rome for Roman Catholics, al-Azhar for Sunni Muslims, or Salt Lake City for Mormons, these centers have limited resources for enforcing their authority; and those that choose to ignore them can use the same techniques of individual religious freedom, (alternative) tradition, separate organization, and transnational networking to circumvent it. The only real exceptions are situations in which state power helps enforce religious authority; but in those cases, such as Iran or Russia, the state boundaries themselves set effective limits to such compulsion.

A prime illustration of the way that globalization encourages the multiplication of local centers of religious authenticity is the effect on religion of global migration and transnational religious networking and organization. Although the migration of people from one part of the world to others is a constant of world history, what is new under conditions of globalization is both the extent of the phenomenon and the fact that migrants leave one place for another far less definitively. The possibilities for maintaining communicative and even physical ties with the places of origin are far more developed, meaning that migrants are less dependent on their new contexts and can still influence the places that they have left. From a different angle, the difference between migrants and nonmigrants is not as great as it used to be, just as the difference between travel/tourism and migration is not as clear. And the chief source of this new situation is the much greater and more rapid communicative links that now exist throughout the world.

Among the consequences for religion of migration and enhanced transnational or translocal linkages, two are of the greatest potential significance. The first is that, as migrants from a particular religion settle in different parts of the world outside of the traditional 'heartlands,' these diaspora communities will develop religious variations that respond to their new locations, and may furthermore re-export these to their places of origin as authentic and authoritative versions. Thus, e.g., the temples that Hindu migrants typically set up in diaspora communities of Europe or North America often allow the worship of a range of gods and goddesses in the Hindu pantheon, unlike in India where temples usually are dedicated to one deity only. The diasporic 'one-stop supermarket' temple is primarily a response to the fact that Hindus from different parts of India migrate to the same places and that these communities do not have the resources to build temples for each subgroup. In spite of these strategic origins of the variation, however, the model is being transferred back to India, providing another authentic way of 'doing' this religion. Similar analyses could be made for Muslim, Christian, Buddhist, or other migrants as well.

The second major consequence for religion is related but also in a sense opposite. It consists in the possibility of tying together disparate versions of a religion across the globe, expressly in spite of the difference in local variations. The Roman Catholic church is perhaps the prime example of a religious organization that has taken great advantage of this possibility. The links among the world's Islamic populations have also strengthened significantly in this context (see Levitt 1998, Rudolph and Piscatori 1997). Another very illustrative example is global Christian Pentecostalism. Here, the concrete ways of doing Pentecostal Christianity vary enormously when one compares places like Korea, South America, North America, Asia, and Africa. Yet transnational linkages in such forms as common publications, regular world conferences, connections among professionals, and migration encourage a consciousness in most Pentecostals of being part of the same worldwide religious movement, and this in spite of the differences among them (see Van Dijk 1997).

6. *Conclusion*

What the foregoing demonstrates is that globalization creates and enhances conditions in which religion can become more pluralistic and yet at the same time more unified. The local and the global in religion, the universal and the particular, are not opposite poles between which religious forms oscillate. Rather local religion can now increasingly only be done with reference to the global context, especially in the awareness of religious pluralism. And global religious forms expressly include and are the expression of local manifestations. Global society does not present the choice between antiglobal, perhaps fundamentalist resistance and proglobal, liberal accommodation. Instead, it recreates the conditions under which religions form and operate in relation to one another.

See also: Diaspora; Globality; Globalization and World Culture; Globalization, Anthropology of; Multinational Corporations; Pluralism; Pluralism and Toleration; Postmodernism in Geography; Postmodernism in Sociology; Religiosity: Modern

Bibliography

Anderson J, Brook C, Cochrane A (eds.) 1995 *A Global World? Re-ordering Political Space*. The Open University and Oxford University Press, London

Beyer P 1994 *Religion and Globalization*. Sage, London

Beyer P 1998 The modern emergence of religions and a global social system for religion. *International Sociology* **13**: 151–172
Braman S, Sreberny-Mohammadi A (eds.) 1996 *Globalization, Communication and Transnational Civil Society*. Hampton Press, Cresskill, NJ
Cox H G 1995 *Fire From Heaven: The Rise of Pentecostal Spirituality and the Reshaping of Religion in the Twenty-first Century*. Addison-Wesley, Reading, MA
Hexham I, Poewe K 1997 *New Religions as Global Cultures: Making the Human Sacred*. Westview Press, Boulder, CO
Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon & Schuster, New York
Juergensmeyer M 1993 *The New Cold War? Religious Nationalism Confronts the Secular State*. University of California Press, Berkeley, CA
Küng H 1991 *Global Responsibility: In Search of a New World Ethic*. Crossroad, New York
Lechner F 1989 Cultural aspects of the world system. In: Swatos Jr W H (ed.) *Religious Politics in Global and Comparative Perspective*. Greenwood Press, Westport, CT
Levitt P 1998 Local-level global religion: the case of United States–Dominican migration. *Journal of Scientific Religious Studies* **37**: 74–89
Luckmann T 1967 *The Invisible Religion: The Problem of Religion in Modern Society*. Macmillan, New York
Marty M E, Appleby R S (eds.) 1994 *Accounting for Fundamentalisms: The Dynamic Character of Movements*. University of Chicago Press, Chicago
Meyer J W, Boli J, Thomas G M, Ramirez F O 1997 World society and the nation-state. *American Journal of Sociology* **103**(1): 144–81
Robertson R 1991 The globalization paradigm: thinking globally. In: Bromley D (ed.) *Religion and the Social Order. New Developments in Theory and Research*. JAI Press, Greenwich, CT, Vol. 1
Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London
Rudolph S H, Piscatori J (eds.) 1997 *Transnational Religion and Fading States*. Westview Press, Boulder, CO
Van Dijk R A 1997 From camp to encompassment: discourses on transsubjectivity in the ghanian pentecostal diaspora. *Journal of Religion in Africa* **27**: 135–59

P. Beyer

Goddess Worship (Old and New): Cultural Concerns

Goddess worship is integral to many religions, old and new. The ancient Mediterranean world contained abundant Goddess lore; today, Hinduism and Buddhism are richest in female divinities. Academic studies of these traditions trace their texts, histories, rituals, iconographies, theologies, intersections with economics and politics, and—most recently—implications for female status. This article builds upon such scholarship, but it aims to describe the contemporary phenomenon of Goddess worship as a *choice* by women and men who believe that worshipping God in a female form will empower them spiritually, socially, or politically. Who is turning to Goddess worship, and why? What does this say about history, religion, and society? Are Goddesses 'good' for women? Although the material investigated here derives largely from Western contexts, one can detect the same recourse to Goddess figures by people in other cultures who seek transformation through potent symbols and personalities. But the 'new' Goddess worship of today, embracing and refashioning the Goddesses of 'old,' is not everywhere the same. A Wiccan member of The Fellowship of Isis may differ radically from an Indian author involved with the feminist press, Kali for Women. Thus to nuance the picture of this emerging phenomenon, non-Western experiences of contemporary Goddess appropriation—particularly Indian—are brought in as test cases for claims made about Goddess worship here in the West.

1. Seeking New Symbols

In Western countries, Goddess worship is typically viewed as a religious alternative for those disillusioned with the overwhelming maleness of Jewish and Christian religious imagery and social structure. Symbols of God as father, conqueror, judge, and lord are believed to enforce the patriarchal domination of women. Moreover, Western philosophical dualisms—soul/body, sacred/profane, transcendent/immanent, rational/emotional, and changeless/changeable—justify the creation of hierarchy and demonize the lower halves, such that it becomes necessary to subdue sexuality, nature, emotion, and change. The new Goddess traditions also make historical charges—that Jews and Christians have wiped out indigenous Goddess-worshipping societies, Jews during the time of the prophets and Christians in the fourth to fifth centuries, when their emperors closed Goddess temples throughout the Roman empire. Indeed, as many critics claim, Jewish and Christian histories have systematically subordinated Goddesses and women, rather than according them equivalence in the theological and social spheres.

The most cogent justification for the need to reject the Judeo–Christian tradition for that of the Goddess draws upon theories about religious symbols. Following the work of Clifford Geertz, Carol Christ (1987, 1997) and other feminists have postulated that symbols have a direct relationship to sociocultural systems. Since symbols create powerful, pervasive, and long-lasting moods and motivations, shaping our understanding of the world and suggesting patterns of action, patriarchal symbols reinforce patriarchal social structures, which therefore appear uniquely realistic. Changes in the way women are perceived and treated can only come through the creation of new symbols, which empower women and undercut male

dominance. As Birnbaum states, quoting Gloria Anzaldúa, 'nothing happens in the "real" world unless it first happens in the images of our heads' (1993, p. 7).

2. *Varieties of Goddess Votaries*

Western Goddess spirituality is heterogenous and flexible, and sits at the intersection of four large groups (Eller 1993, Puttick 1997). The first, known loosely as spiritual feminism or women's spirituality, started in the early 1970s as a spontaneous grassroots movement. Cynthia Eller describes it as being founded on four premises: religion and ritual *can* be empowering for women; a healthy environment, peace, and community must be sought; Western history needs revising; and gender relations should be explored. Since many spiritual feminists stay within their birth traditions to work for justice and equality, not all women in this category worship Goddesses.

The second umbrella label is (neo-) paganism and Wicca, founded in the 1950s in Britain and based on the recovery of pre-Christian European religions such as Druidry and Odinism. In general, such groups believe in a dynamic balance between male and female; the centrality of fecundity; an immanent deity; care for the earth; the efficacy of ritual and magic; and the leadership of women. Goddesses are central to pagan religiosity, but not necessarily in isolation from male gods.

The concerns of spiritual feminists and Wiccans can be similar, as seen in the writings and practices of feminist witches Zsuzsanna Budapest and Starhawk. However, there are also frictions between the two groups. Since Wicca is not political, does not aim to strip away patriarchal domination, and advocates male/female complementarity, some spiritual feminists view Wiccans as sexist. Wiccans return the compliment; for them, politics and feminism are irrelevant to spiritual growth.

A third set of related groups is referred to collectively as the New Age. Famous for their crystals, chakras, channeling, past-life regressions, astrology, and mind-over-matter affirmations, New Age individuals create a pastiche of religious self-help devices, many of which involve Goddess figures. Spiritual feminists and pagans often critique the New Age for its tendency to whitewash and sweeten Goddesses; for its racism, misogynism, lack of commitment to community struggle, and exploitation of cultures from which it borrows; and for its naive optimism, unhelpful in times of crisis (Christ 1997, Sjöö 1992).

The final group—new since the 1980s—of contemporary Western Goddess worshippers are the eco-feminists, eco-spiritualists, and Gaians (from Gaia, or Earth). Though not all of them gynemorphize the earth as a Goddess (there are Jewish, Christian, and Buddhist eco-feminists as well), most worship the earth as a mother Goddess and decry her desecration by modern industrial patriarchs. For Vandana Shiva, the Indian human rights activist, feminism is a form of ecology, and ecology is a revival of *prakṛti*, the demonized but potent Hindu feminine principle of nature (1988, pp. 38–54).

3. *Who is She?*

Like the variety of the Goddess movement, so also the Goddess herself. Her identity is multiple and fluid, changing depending upon the needs of her worshippers. Some women understand her as the affirmation of the female body, sexuality, and power, all of which have been downplayed in traditional Western religions. Although many feminists caution against too rigid a defense of female sexuality, as if biology were determinative of a women's 'essence' (di Leonardo 1998, pp. 98–108, Gross 1998, pp. 159–70), others argue that creating a space for women to reclaim the positive aspects of their sexuality is so important that the project is worth the risks.

Another method of affirming female strength is through a Jungian perspective, where the Great Goddess is an archetype of the feminine, dark side of life. Following Robert Graves, Erich Neumann, and Joseph Campbell, modern Jungians claim not only to empower women and men through the recognition of the positive aspects of innate femaleness, but also to explain, through the theory of archetypes, why women who have never encountered Goddesses in real life can suddenly start creating and dreaming about them. Not all feminists and Goddess worshippers approve of the use of Jungian archetypes, however. Some protest that in revalorizing the feminine unconscious, the banished 'other,' one is creating a new version of the old male/female dichotomy which subjugates women (Gross 1998, p. 190); others who worship a 'real' Goddess complain that Jungians relegate her to the 'unreal' realm of the mind (Christ 1998, pp. 86–8); and still others, drawing again upon Geertz's work on symbols, prefer to create new ones rather than to adopt those postulated by a patriarchal psychological system.

The Goddess also has material manifestations. As indicated above, some women and men view her as the earth. Since we are all part of nature, we must celebrate, and not attempt to control, its cycles. 'I feel deeply that the flight from finitude and death is at the root of the problems we face… Maybe something went wrong, massively wrong, when Platonism and Christianity became the dominant symbol systems of our culture. Maybe… the Goddess can help us...to love a life that ends in death' (Christ 1987, p. 210). Often linked with this conception of divinity is the assertion that the Goddess is a relational symbol of embodied, immanent love.

Although many Western Goddess votaries might assent to the interpretations given above, the majority

Figure 1
Artemis of Ephesus (reproduced by permission of the Mansell Collection/TimePix)

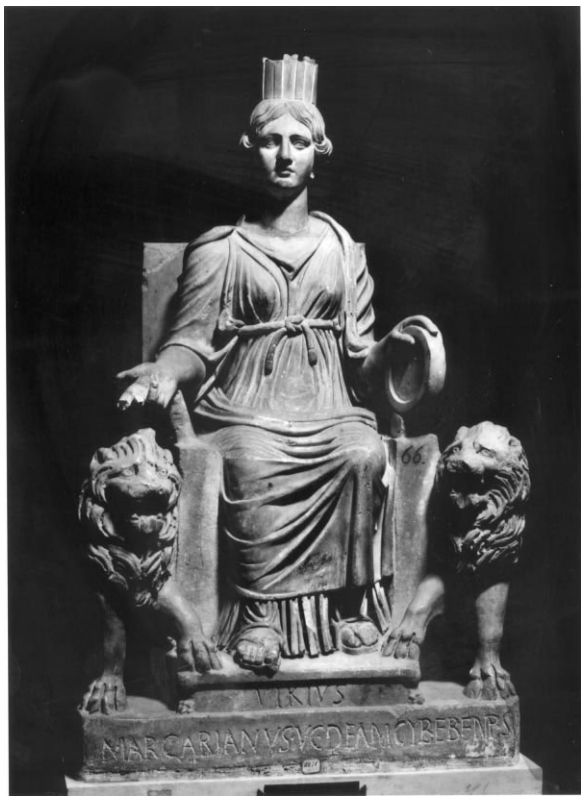

Figure 2
Cybele enthroned with her lions (reproduced by permission of the Mansell Collection/TimePix)

also worship the Goddess as a living presence. Some, choosing to stay within their Jewish and Christian contexts, derive inspiration from the feminine sides of God as revealed in the Bible or from their respective mystical traditions: Isaiah's 'womb language'; the emphasis on God's wisdom; the Kabbalists' Shekhinah, or presence of God; the Virgin Mary; feminine interpretations of the Holy Spirit; and Julian of Norwich's Mother Jesus.

Others, finding the traditions of their birth irredeemably patriarchal, go outside in search of Goddess figures whom they can approach in prayer and ritual. The most frequent destinations for such seekers are Western, often European, traditions, where, even if the Goddesses are no longer the center of living cults, they are believed to be waiting, still puissant, able to heal. Examples include the Greek Demeter and Persephone, Aphrodite, Artemis (see Fig. 1), and Athena; the Egyptian Isis; the Sumerian Inanna; the Canaanite Astarte; and the Anatolian Cybele (see Fig. 2). Sometimes, though more gingerly, for fear of causing offense, Western Goddess worshippers also turn to figures from Native American tribes, such as Old Spider Woman, Corn Woman, and Serpent Woman, as well as to powerful ancestors and spirits from Africa.

Non-Western Goddesses also have devoted followings: the Japanese Amaterasu, the Hindu Durgā and Kālī, the Tibetan Buddhist Goddesses Tārā and Vajrayoginī; and the Chinese Buddhist Kuan-yin. Often these, and the Western deities mentioned above, are lumped together through superficial resemblances, archetypal links, and claims of historical connection, producing what Cynthia Eller calls the 'grab-bag approach to feminist spirituality' (1993, p. 75). Many criticize these dehistoricized approaches to Goddess worship as insulting, romantic, essentialist, and orientalist. The fact that some modern Goddesses have been created afresh, according to elite needs—for example, Asphalta, the parking Goddess (Grey 1989)—does not assuage such fears. Nevertheless, most modern Goddess worshippers are sincere seekers, turning to female embodiments of the divine in a genuine search for new life.

Before concluding this introduction to the meanings of the Western Goddess, it is worthwhile noting that

the reclamations under discussion appear not to be occurring among Muslims. There are three possible explanations: the minority status of Muslims in Western countries, which places them outside the liberal mainstream; the paucity of feminine images of the divine in the Qur'an; and the strict monotheism of Islam, which makes anathema the search for a Goddess figure.

4. *How Old is She?*

One of the most controversial elements of the Goddess spirituality movement is a claim about origins, or the 'Goddess Hypothesis.' Currently championed by authors like Christ, Eisler, Gadon, Gimbutas, Spretnak, and Stone, but relying on the earlier works of Bachofen, Engels, Graves, and Neumann, this hypothesis rests on a belief that 'there were past societies in Europe and the Near East, especially prior to the so-called invasions of the Indo-Europeans circa five thousand years ago, that were Goddess-worshipping, female-centered, in harmony with their environments, and more balanced in male-female relationships' (Conkey and Tringham 1995, p. 206). Drawing upon archaeological excavations of female figurines (fertility 'idols' of 'the Great Mother cult'), grave sites, and cave paintings, as well as anthropological research into modern-day 'primitive' societies who may be considered survivals from earlier eras, such authors see continuities from the time of the Upper Paleolithic period (32,000–10,000 BCE) and the Neolithic Revolution (9,000 BCE), down to the societies of Old Europe (6,500–3,500 BCE). What they have in common is a Great Mother cult and a social organization that is matrifocal (some claim matriarchal), earth-bound, and peace loving. Present-day clues of this earlier time, when the 'earth was believed to be the body of a woman and all creatures were equal' (Birnbaum 1993, p. 3), can be seen in the black Madonnas of Europe and South America, as well as in tribes that revere women for their reproductive powers. All such societies ended abruptly with the arrival of the patriarchal, war-oriented Indo-Europeans in the middle of the fourth millennium BCE; they subjugated human women and tamed, desexualized, or slew their Goddesses.

Although some women and men find such a vision of the past to be liberating for the future, others find much to criticize (Conkey and Tringham 1995, Eller 2000, Motz 1997, Puttick 1997). How do we know what the female figurines represent without an adequate knowledge of the social systems that produced them? Is it possible that they had the same symbolic meaning over nearly 30,000 years? Is it not feminist essentialism to argue that all the figurines connote fertility, as if women have always been defined biologically? Most historical Goddesses, indeed, are not tied to fertility at all. And if it is allowable to read back to Old Europe by citing modern-day 'survivals,' what about contemporary groups in which there is matrilineality but male control, ritualized fertility but no Goddess worship, and Goddess veneration but little respect for women? The brunt of such criticisms is that the past is oppressive, not golden; it ill serves the women's movement to idealize something unproven.

5. *Is She 'Good' for Women?*

Definitions of women and Goddesses that rely too much on biological determinism, Jungian archetypes, and the 'Goddess Hypothesis' are not the only contested issues in the Goddess movement. Another is the relationship between Goddess-worshipping societies and the status of their women. Do Goddesses serve patriarchal interests, or feminist ones? The evidence is mixed; although some Goddesses appear to serve their human devotees by undercutting patriarchal subordination, and although some women find strength in images of the divine feminine, other Goddesses reinforce their societies' domination of women, such that feminists refuse to call upon them.

First, some positive correlations. Many modern Western women do indeed derive inspiration from their worship of Goddess figures: Luisah Teish, who looks to West African ancestor veneration; Charlene Spretnak, Christine Downing, and Carol Christ, who have found spiritual homes with the Goddesses of the Mediterranean; China Galland, whose life was changed by encountering Tārā; and Rita Gross, who testifies about the inner transformation she experiences through the worship of Vajrayoginī. Very often, one woman's metamorphosis leads to another's, as occurs through the women's pilgrimages to Crete led by Carol Christ.

It is not only Western women who are thus empowered. To take a few examples from South Asia, the audacity of the Goddess Rādhā vis-à-vis her lover Kṛṣṇa provides Bengali women with models of assertiveness and resistance (Wulff 1997); forms of Durgā empower Hindu devotees and holy women by their autonomy and puissance (Erndl 1997); women Tantrikas of the Tibetan Buddhist tradition find their spiritual aspirations mirrored in powerful images of the feminine divinity (Shaw 1994); and the activist Mother's Front in Sri Lanka worships Kālī for fortitude in the midst of the civil war (de Alwis 1999). Sometimes the Goddess herself overturns patriarchal values; the South Indian Goddess Kaṟumāriammaṉ allows menstruating women into her temple (Narayanan 1999, p. 69).

But there are also contrary indications. All 10 Goddesses whom David Kinsley (1989) describes in *The Goddesses' Mirror* belong to societies in which men dominate women; the same is true for those

catalogued by Motz (1997). Other scholars find similarly depressing inverse correlations between Goddess worship and freedom for women. Again, from South Asia: William Sax investigates the north Indian devotion to Nandādevī and finds that while the Goddess represents the trials of out-married daughters, her cult enables rather than undercuts their sorrows (1991); China Galland visits Bhaktapur, Nepal, and describes both her delight in the beauty of Durgā worship and her horror at the thriving trade in child prostitution which exists alongside it (1998); and Lisa Hallstrom discovers that although devotees view the modern saint Ānandamayī Mā as an incarnation of the Goddess, she is not considered a 'woman,' and feminist interpretations of her significance are discouraged (1999). The frustration of such a disjunction between expectations and reality is summed up by an eight-year-old girl in a story by Mrinal Pande: 'when you people don't love girls, why do pretend to worship them?' (1990, p. 63).

Many women's activists in countries such as India, therefore, eschew Goddesses as aids in their struggles. For Goddesses can validate a sexist model, as in the case of Sītā, the virtuous wife of Rāma, or represent a romanticization of the past and a glorification of motherhood. Indian feminists, after initially turning to Kālī and Durgā as symbols of empowerment, found them usurped by Hindu nationalists and deployed for divisive ends (Agnes 1995). One example is Durgā Vāhinī, or Durgā's Army, a Hindu nationalist women's group. The Vāhinī has revived female militancy—not in the service of women's issues (caste, divorce, dowry, sati, or marital violence), which are not debated or fought for, but against Muslims (Pathak and Sengupta 1995). The eight-armed Durgā gives women a new identity, but an identity that serves patriarchal values.

6. *Trends, Evaluations*

Goddess worship, wherever it is found, is diverse and controversial. Particularly since 1985 in the West, it is also a place where the popular and the scholarly meet and spar. Sometimes academic authors bemoan the theoretical backwardness that characterizes the public enthusiasm for Goddess spirituality and matriarchal histories, complaining that exhorting women to realize that 'we're all Goddesses within!' vitiates the need for external change (di Leonardo 1998, pp. 104–7). But sometimes the meetings between popular and scholarly are not so antagonistic: Goddess-revering people of all stripes can and do join hands when it comes to issues of social justice and eco-spiritualism.

The revived Goddess spirituality of the West also provides opportunities for encounters of other types. The increasing ethnic mix of Western societies means that 'new' Goddess worshippers come into contact with those for whom Goddesses are a traditional part of their religious backgrounds. Sometimes Western votaries find this an inspiring, deepening encounter, and sometimes they do not. How, for instance, is one to incorporate the traditional practice of goat sacrifice into one's Western worship of Kālī? By the same token, many first- and second-generation immigrants find the appropriation of their Goddesses by mainly white Americans to be offensive. The Goddess movement is obviously a fertile field for intercultural exchange.

How then to evaluate this refashioning of 'old' Goddesses in the service of 'new' ideals and hopes? While from an academic perspective one might deplore the 'Goddess Hypothesis,' biological determinism, the perceived exploitation of 'other people's symbols,' and a naive assumption that all women everywhere benefit from the Goddess, the debate, creativity, and diversity characteristic of 'new' Goddess worship in the West are evidence of laudable cultural vibrancy. Goddesses may not always, or even often, provide alternative models for women in today's global societies. But as various studies have shown, in religious groups where women do have a leadership role, there is typically a feminine aspect to God (Sered 1994). Rita Gross makes the point differently when she asserts that even if Goddesses do not confer status or independence to women, they do contribute to psychological well-being. 'Though the presence of goddesses does not guarantee equity for women, their absence almost certainly guarantees inequity for women' (1998, p. 161). A final example from India: just because the slain Mrs. Gandhi was depicted in the press as a mother Goddess killed by her ungrateful Sikh children, who therefore deserved the violence unleashed upon them, there is no reason that Goddess symbolism, wrested from nationalist agendas and purposely reworked, could not be used for healing and hope. To echo Christ, echoing Geertz: people change as symbols change, and vice versa. That the denizens of the Goddess movement are trying to push this mutual process forward to their own ends is neither surprising nor unprecedented, though where the Goddesses 'take hold' and in what forms are matters for future interest.

See also: Buddhism and Gender; Catholicism and Gender; Feminist Theology

Bibliography

Agnes F 1995 Redefining the agenda of the women's movement within a secular framework. In: Sarkar T, Butalia U (eds.) *Women and Right-Wing Movements: Indian Experiences*. Zed Books, London

Birnbaum L 1993 *Black Madonnas: Feminism, Religion, and Politics in Italy*. Northeastern University Press, Boston

Christ C 1987 *Laughter of Aphrodite: Reflections on a Journey to the Goddess*. Harper & Row, San Francisco
Christ C 1997 *Rebirth of the Goddess: Finding Meaning in Feminist Spirituality*. Routledge, New York
Conkey M, Tringham R 1995 Archaeology and the Goddess: Exploring the contours of feminist archaeology. In: Stanton D, Stewart A (eds.) *Feminisms in the Academy*. University of Michigan Press, Ann Arbor, MI
de Alwis M 1999 Motherhood as a space of protest: Women's political participation in contemporary Sri Lanka. In: Jeffrey P, Basu A (eds.) *Resisting the Sacred and the Secular: Women's Activism and Politicized Religion in South Asia*. Kali for Women, New Delhi, India
di Leonardo M 1998 The Three Bears, The Great Goddess, and the American temperament: Anthropology without anthropologists. In: di Leonardo M (ed.) *Exotics at Home: Anthropologies, Others, American Modernity*. University of Chicago Press, Chicago, pp. 79–144
Eller C 1993 *Living in the Lap of the Goddess: The Feminist Spirituality Movement in America*. Crossroad, New York
Eller C 2000 *The Myth of Matriarchal Prehistory: Why an Invented Past Won't Give Women a Future*. Beacon Press, Boston
Erndl K 1997 The Goddess and women's power: A Hindu case study. In: King K (ed.) *Women and Goddess Traditions: In Antiquity and Today*. Fortress Press, Minneapolis, MN
Galland C 1998 *The Bond Between Women: A Journey to Fierce Compassion*. Riverhead Books, New York
Grey M 1988 *Found Goddesses: Asphalta to Viscera*. New Victoria Publishers, Norwich, VT
Gross R 1998 *Soaring and Settling: Buddhist Perspectives on Contemporary Social and Religious Issues*. Continuum, New York
Hallstrom L 1999 *Mother of Bliss: Ānandamayī Mā (1896–1982)*. Oxford University Press, New York
Kinsley D 1989 *The Goddesses' Mirror: Visions of the Divine Feminine from East and West*. State University of New York Press, Albany, NY
Motz L 1997 *The Faces of the Goddess*. Oxford University Press, New York
Narayanan V 1999 Brimming with Bhakti, embodiments of Shakti: Women of power in the Hindu tradition. In: Sharma A, Young K (eds.) *Feminism and World Religions*. State University of New York Press, Albany, NY
Pande M 1990 Girls. In: Holmström L (ed.) *The Inner Courtyard: Stories by Indian women*. Virago Press, London
Pathak Z, Sengupta S 1995 Resisting women. In: Sarkar T, Butalia U (eds.) *Women and Right-Wing Movements: Indian Experiences*. Zed Books, London
Puttick E 1997 Goddess spirituality: The feminist alternative? In: Puttick E *Women in New Religions: In Search of Community, Sexuality, and Spiritual Power*. St. Martin's Press, New York, pp. 196–231
Sax W 1991 *Mountain Goddess: Gender and Politics in a Himalayan Pilgrimage*. Oxford University Press, New York
Sered S 1994 *Priestess, Mother, Sacred Sister: Religions Dominated by Women*. Oxford University Press, New York
Shaw M 1994 *Passionate Enlightenment: Women in Tantric Buddhism*. Princeton University Press, Princeton, NJ
Shiva V 1988 *Staying Alive: Women, Ecology, and Survival in India*. Kali for Women, New Delhi, India
Sjöö M 1992 *New Age and Armageddon: The Goddess or the Gurus? Towards a Feminist Vision of the Future*. The Women's Press, London
Wulff D 1997 Rādhā's audacity in Kīrtan performances and women's status in Greater Bengal. In: King K (ed.) *Women and Goddess Traditions: In Antiquity and Today*. Fortress Press, Minneapolis, MN

R. F. McDermott

Goffman, Erving (1921–82)

The work of the sociologist Erving Goffman is regarded as a classic by most of his professional colleagues. Randall Collins (1988, p. 41) even considers him 'the greatest sociologist of the latter half of the twentieth century.' Goffman's influence, particularly on many younger researchers, is still powerful at the beginning of the twenty-first century and is by no means restricted to sociology. Goffman's works have been influential in numerous parallel disciplines, above all in social psychology, linguistics, philosophy, literature studies, pedagogics, psychiatry, and media studies. It cannot be denied that Goffman made important empirical, theoretical, and methodological contributions. The following essay will discuss them against the background of a short biographical sketch and relate them to relevant sociological contexts of discussions.

1. *Biographical Sketch*

Erving Goffman was born in the small Canadian town of Manville on July 11, 1922, the son of Russian-Jewish immigrants. He began his academic career at the University of Toronto, where he took his BA in 1945. After completing his sociology studies at the University of Chicago, USA, where he primarily studied under Everett C. Hughes and W. Lloyd Warner, he took up his first post in Scotland, UK. From 1949 to 1951 he worked at the Department of Social Anthropology at Edinburgh University. After that he returned to Chicago, where he gained his doctorate with his thesis *Communication Conduct in an Island Community*, which he had prepared in Scotland. The empirical basis of his dissertation, which was supervised by W. Lloyd Warner, Donald Horton and Anselm L. Strauss, was a field study on the Shetland Islands. Following a spell as research assistant at the Department of Sociology at Chicago University, where his research included social stratification, Goffman was employed as a Visiting Scientist in the Laboratory of Socio-Environmental Studies at the National Institute of Mental Health (NIMH) in Bethesda (Maryland) from mid-1954 to the end of 1957. During this time Goffman carried out a one-year study in a mental hospital. This study on the basis of 'participant observation' formed the essential foundation of his book *Asylums* (1961a). In 1958 Goffman transferred to Berkeley as Assistant Professor of

Sociology and member of the Center for the Integration of Social Science Theory at the University of California. From 1962 he was a Full Professor in Berkeley. In 1968 he accepted the Benjamin Franklin Professorship at the University of Pennsylvania in Philadelphia, where he substantially advanced his research program and wrote his *chef d'oeuvre* '*Frame Analysis*' (1974).

His works, which reached a wide academic and non-academic audience and have been translated into many languages, earned Goffman numerous honors. He received a number of prestigious awards and was Visiting Professor at several universities at home and abroad. In 1981 Goffman became President of the American Sociological Association (ASA). At the peak of his powers after an academic career of almost 30 years, Goffman died on November 20, 1982.

2. *Goffman's Work*

2.1 *Research Program and Intellectual Context*

Goffman's main theme was the forms of immediate social interaction, 'face-to-face' encounters. He saw in them a specific type of social system and a sociological field in its own right. In his Presidential Address delivered at the end of his term as President of the American Sociological Association shortly before his death, he declared: 'My concern over the years has been to promote acceptance of this face-to-face domain as an analytically viable one—a domain which might be titled for want of any happy name, the interaction order—a domain whose preferred method of study is microanalysis' (Goffman 1983, p. 2). This programmatic description of his position does not mean that Goffman regarded the interaction level as completely autonomous or socially fundamental. Rather he stressed the manifold 'relativity' of the autonomy of the interaction order. In many places in his works Goffman pointed out its historical variability as well as links ('a loose coupling') between the micro-level of the interaction and social macro-structures (e.g., social stratification or bureaucratic organization). However, Goffman did not systematically consider the historical and social conditions and dependencies of his subject. Rather he was primarily concerned with observing and discovering seemingly insignificant aspects of daily behavior. He focused on tiny moments of people's behavior, which they themselves find more or less natural and are unaware of: fleeting glances, physical contact and distance, apologies, greetings, compliments, silence, etc.

All Goffman's works, 11 books in total, are consistently oriented towards this empirical reality, but from the start Goffman also adopted means of interpretation from various sociological schools and traditions, including above all the 'classics' William James (see *James, William (1842–1910)*), Emile Durkheim (see *Durkheim, Emile (1858–1917)*) and Georg Simmel (see *Simmel, Georg (1858–1918)*), structural functionalism (Alfred Radcliffe-Brown, Talcott Parsons (see *Mead, George Herbert (1863–1931)*)), representatives of the Chicago School (George H. Mead (see *Mead, George Herbert (1863–1931)*), Charles H. Cooley, Louis Wirth et al.), existentialism (Jean-Paul Sartre), ethology, ethnomethodology (Harold Garfinkel, Harvey Sacks), and sociolinguistics (Dell Hymes). But Goffman never regarded his theoretical references, models and metaphors as ends in themselves, they were rather means to the end of discovering and describing social order.

2.2 *Approaches, Positions, and Results*

In the following attempt to briefly summarize and appraise Goffman's work, a number of key sociological terms either coined or redefined by Goffman will be used as a guideline.

2.2.1 Dramaturgy. Goffman became known for his analysis of social life based on the model of the theater (see 1959), and his model has remained influential. It is too simple to speak of Goffman's work, as for example Jonathan Turner does, as the 'dramaturgical school of interactionism,' but it can be demonstrated that the dramaturgical perspective consistently played an important role for Goffman. It can similarly be proved that Goffman's dramaturgical approach partly anticipated and formed a basis for many of his later constructs. This particularly applies to his works on strategic interaction, ritual, territoriality, identity, and stigma.

Within the context of the theater model Goffman reconstructed above all the strategies of social 'face work': forms of concealment, lying, self-stylizing, mystification, the manipulation of 'facades,' etc. His aim here was not just to show how individuals, selfish subjects in the fight for social recognition, attempt to present themselves in the best light; rather he constantly also saw dramaturgical everyday life as producing and reproducing social order, a restrictedly open and widely endangered process that is to be permanently established.

Goffman's dramaturgical approach and his later reflections on social territories (1971) and frames (1974) also help us understand the spatial organization (social ecology) of modern society. For example, it becomes clear that normality is a fictional impression to be produced by differentiating distinct spheres of perception, audiences, and representations; it is the foundation for the actors' basic social trust. Goffman's differentiation between 'frontstage' and 'backstage' (see 1959) became famous in this context.

2.2.2 Roles and identities. Goffman's early works in particular are to be seen in the context of role theory,

influential at the time, which he criticized and developed further. Goffman treated role analysis primarily as interaction analysis, investigating the functioning and organization of the 'actual' practice of performing a role against the background of its normative frame. Thus Goffman developed concepts of patterns and styles of behavior, including the term 'role distance' (see 1961b), which has meanwhile acquired the status of a basic sociological term and which refers to a way of behavior which comments on the role and primarily serves either the interaction system or the selves relevant in the situation. For example, Goffman described a five-year-old boy riding a roundabout horse who by little irreverences in his behavior demonstrates that his current role does not correspond to his 'true,' 'more adult' self.

In *Stigma* (1963), Goffman differentiated the term 'self' in the context of an identity theory which distinguishes three types of identity in a basic sense that is still valid at the beginning of the twenty-first century: (a) social identity as a person's role set, (b) personal identity as a person's synchronic and diachronic individuality ascribed to him by observers, and (c) ego identity as a person's inner self-reference. Goffman used these concepts of identity together with the instruments of his dramaturgical approach above all in analyzing stigmatized deviations and deviants. He focused on the techniques of information control and emotion management of the stigmatized people as well as on the consequences of the stigmatization for their identities and socialization.

2.2.3 Total institutions. Goffman's book *Asylums* (1961a) represents an exception among his works in that it systematically goes beyond the interaction level. It describes a class of social organizations that has to do with confining, controlling and 'altering' people, such as mental hospitals, prisons, convents, homes, barracks, concentration camps, etc. However, Goffman's (1961a) investigated such 'total institutions', where the most elementary matters of course and norms of everyday interaction are broken, with the aim of exposing the interaction order and the self connected with it. Beyond that, Goffman's analysis demonstrated that total institutions form an ambivalent milieu for socialization. On the one hand they act as machinery for disciplining and normalizing behavior and interaction. On the other hand these institutions damage the moral self, e.g., self-respect and modesty, through various forms of humiliation.

2.2.4 Ritualizations. In his work Goffman repeatedly dealt with the moral sides of the interaction order. For Goffman the self is a moral fact which implies rights and duties when showing respect and which can be observed in ritual practices and codes of everyday interaction. Every person appears as both subject and object of a ritual 'language,' a 'language' of politeness, decency, and tact, but also on occasion of contempt, wounding, humiliation, etc. (see Goffman 1967).

At the end of his career Goffman elaborated his early ritual model to include ethological and frame theoretical considerations, on the basis of which he analyzed gender interactions and the interpretative patterns they are based on (see 1979). Goffman depicted gender as a combination of ritual styles of presentation representing beliefs about gender-specific essential attributes. These beliefs amount to a schematic asymmetry similar to that of the parent-child relationship. In daily interaction life and its media depictions (e.g., in advertising) men appear as superior beings to women by ritualizations of dress, physical contact, smiling, activity, etc.

2.2.5 Strategic interactions. For Goffman, ritualizations represented the most important social condition of profit-oriented strategic action on the level of interaction. Totally ignoring their contents, he analyzed this action within the framework of a model of strategic interaction oriented towards game theory (see 1969). The main point of departure and aim of Goffman's respective considerations were the social situation in which the subjects strive for control of the 'definitions' of the other players or rivals. Following on from his dramaturgical approach, Goffman above all investigated the conditions of competence and limits of 'impression management,' the strategically relevant interpretations of expression (e.g., of the body), the logic of reciprocal strategic observation and assessment, as well as the complexity of meaning and virtuosity of strategic moves. The empirical basis of the investigation consisted in particular in self-descriptions by a wide variety of strategic subjects (spies, confidence tricksters, adulterers, politicians, etc.). Goffman 'translated' their practical knowledge of action (knowledge of cleverness) into the categories of his analytic description, e.g., a typology of strategic moves.

2.2.6 Frames and framing subjects. Frame Analysis (1974) can be regarded as Goffman's main theoretical work, because it integrates, or could integrate, the many various analytic perspectives and terminological distinctions of his earlier works. *Frame Analysis* itself therefore forms something like a frame. It is characterized by two basic orientations. On the one hand, the interest is directed at 'structures' and procedures for establishing meaning which are effective in generating normality in social situations and which are valid, independent of these situations. On this level, which can be compared with the grammar

of a language, Goffman revealed basic types of meaning and complex possibilities of transforming meaning (play, experiment, demonstration, etc.). On the other hand, on the level of the concrete situative events he aimed both at normal uncertainties, dangers, and the pressure to conform, as well as at extreme forms of abnormality, such as the 'negative experiences' of inmates of mental hospitals, whose habitual framings, which have proved their worth in everyday life, fail in the institution. The individual, whom Goffman revealed as the subject of the framing, is characterized by enormous powers of judgment and virtuosity, particularly in functionally complementing rules and also in manipulating them.

2.2.7 Forms of talk. In his last period, Goffman built on earlier studies and particularly concerned himself with conversations. The main result of this work is his book *Forms of Talk* (1981), which Goffman specifically placed in the context of *Frame Analysis*, one chapter of which was already devoted to conversations.

Goffman's late focusing on conversations can be seen in connection with the emergence of conversation analysis as a branch of ethnomethodology, which he approved of, but also criticized (see Bergmann 1991). In contrast to conversation analysis, which tends to conceptualize the process of communication as an autonomously organized system, Goffman stressed the importance of ritual order, which is always culture-specific and which permits selves and consequently psychological states to develop. Goffman basically rejected the assumption by conversation analysts that sections of a conversation can be analyzed without reference to their situational and extra-situational contexts. Moreover, Goffman was mainly interested in a sociological approach to the contextuality of social processes. *Frame Analysis* represents this approach.

2.3 Goffman's Methods

Goffman's work is essentially the result of the author's talents, in particular his personal sensitivity, the subtlety of his powers of observation, and his ability to classify his material in his search for order. He displayed these talents in the course of practical research, which struck most observers as unconventional, even puzzling. Among the most irritating points are Goffman's data, whose heterogeneity can hardly be exceeded. In his works the reader comes across personal everyday experiences, novels, autobiographies, the daily press, essays by film critics, photographs, etc. In many other respects Goffman also deviated from the normal forms of sociological discourse. He abbreviated the traditional terminological vocabulary of his field, avoided lengthy argumentation, frequently changed perspective, focused on seemingly unimportant details, commented in an autoreflective manner on his own work, made ironic remarks and repeatedly told or referred to 'stories.' What is most important is that Goffman did not conform to any methodological canon, rather his methods entirely depended on the subject. This meant first of all that it was necessary to forgo certain traditional research methods. For example, the fact that those patterns of behavior that interested Goffman were generally unconscious ones meant that interviews were largely or totally impossible.

Goffman himself described his procedure as primarily 'naturalistic.' The scope of what he meant by that ranged from 'participant observation' (even in his own everyday life) to work with audio-visual recordings. His aim was always to get as close as possible to 'social nature' and to capture it as accurately as possible. On the other hand Goffman tried to achieve analytic distance and obtain information by various forms of 'alienation.' Alongside systematic work with models such as that of the theater, his most important technique was the strategy of alienation, which Paul Drew and Anthony Wootton (1988, p. 7) called 'the investigation of the normal through the abnormal.' Goffman employed totally different forms of abnormality, such as stigmatizations, espionage, accidents, or celebrities as approaches to the logic of everyday normality.

In order to convert into texts the results that were generated by his methodological dialectic of approaching and distancing Goffman combined two strategies. On the one hand he developed a formal analytical language which contains the basic features of the interaction order and the self connected with it. On the other hand he employed a strategy of 'thick description' (C. Geertz). By using a literary style, a metaphorical language suited to the specific subject, and extensive quotations from his material, cleverly combined with his analytical observations, he succeeded in making his subject appear alive and viewing it in a new light.

3. Goffman's Importance for the Social Sciences Today

It is not easy to judge how relevant Goffman's work is for the social sciences at the beginning of the twenty-first century; its main significance probably lies in his contribution to discovering the interaction order. In this way Goffman opened up a specific research area and at the same time provided central categories and strategies for exploring it.

But quite apart from interaction sociology, Goffman also had a strong influence on a number of other research areas and he remains influential today. This is especially true of many aspects of research into identity and socialization. Of prime importance among them are Goffman's reflections on institutional control and

discipline procedures, and on subjectivity. Goffman is especially appreciated today in research on normality and deviation as well as in gender studies, where some of his works are still standards now. In addition, he has a significant influence not only on the level of the everyday interaction order, but also on the sociology of certain social fields of action (politics, advertising, science, etc.) and in media research, where many of Goffman's analytical instruments have proved to be valid. Furthermore it can be seen that recently Goffman's works have increasingly been read and discussed from the point of view of general theories of civilization and society. For example, parallels have been drawn and connections made with the works of Norbert Elias (see *Elias, Norbert (1897–1990)*), Michel Foucault (see *Foucault, Michel (1926–84)*), and Pierre Bourdieu (see Willems 1997). In general it is a remarkable fact that widely differing and contrasting schools have incorporated parts of Goffman's works. The spectrum of positions Goffman is 'relevant' for, ranges from modern systems theory through critical theory, right down to the rational-choice approach.

Just as the reactions to Goffman's work have been wide-ranging and many-sided, so evaluations of its scope and its identity have differed greatly. While most people see the importance of Goffman's work essentially on the level of empirical micro-observation, others (e.g., Anthony Giddens) believe that in it there is to be found a systematic theory that goes beyond the level of micro-sociology and is highly significant for the development of the social sciences. Similarly, various labels have been attached to Goffman's thinking (symbolic interactionism, phenomenology, role theory, structuralism, ethnomethodology, etc.).

No matter how the situation of his reception is to be judged, it cannot be denied that Goffman left behind a series of studies that are still relevant and influential today and a network of concepts that is both highly complex and highly integrated. Perhaps an equally important part of his legacy is the paradigm he provided by the style and ethos of his research.

Bibliography

Bergmann J R 1991 Goffmans Soziologie des Gesprächs und seine ambivalente Beziehung zur Konversationsanalyse. In: Hettlage R, Lenz K (eds.) *Erving Goffman—ein soziologischer Klassiker der zweiten Generation*. UTB, Bern, Stuttgart, Germany, pp. 301–26

Collins R 1988 Theoretical continuities in Goffman's work. In: Drew P, Wootton A (eds.) *Erving Goffman. Exploring the Interaction Order*. Polity Press, Cambridge, UK, pp. 41–63

Drew P, Wootton A 1988 *Erving Goffman. Exploring the Interaction Order*. Polity Press, Cambridge, UK, pp. 1–13

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday Anchor, New York

Goffman E 1961a *Asylums. Essays on the Social Situation of Mental Patients and Other Inmates*. Doubleday Anchor, New York

Goffman E 1961b *Encounters: Two Studies in the Sociology of Interaction*. Bobbs-Merrill, Indianapolis, IN

Goffman E 1963 *Stigma. Notes on the Management of Spoiled Identity*. Prentice Hall, Englewood Cliffs, NJ

Goffman E 1967 *Interaction Ritual: Essays on the face-to-face Behavior*. Doubleday Anchor, New York

Goffman E 1969 *Strategic Interaction*. University of Pennsylvania Press, Philadelphia, PA

Goffman E 1971 *Relations in Public. Microstudies of the Public Order*. Harper and Row, New York

Goffman E 1974 *Frame Analysis. An Essay on the Organization of Experience*. Harper and Row, New York

Goffman E 1979 *Gender Advertisements*. Macmillan, London

Goffman E 1981 *Forms of Talk*. Basil Blackwell, Oxford, UK

Goffman E 1983 The interaction order. *American Sociological Review* **48**: 1–17

Manning Ph 1992 *Erving Goffman and Modern Sociology*. Stanford University Press, Stanford, CA

Willems H 1997 *Rahmen und Habitus. Zum theoretischen und methodischen Ansatz Erving Goffmans: Vergleiche, Anschlüsse und Anwendungen*. Suhrkamp, Frankfurt a. M., Germany

H. Willems

Goodness of Fit: Overview

Many statistical methods make assumptions about, or have consequences for, particular distributions of certain quantities. Goodness-of-fit methods assess the extent to which such assumptions or consequences are supported by the data. Similarly, in the process of model search or model selection within a particular choice of statistical distribution or model form, goodness-of-fit approaches are used to guide the process of arriving at a sensible parsimonious model that captures the main features of the data.

This article describes the general philosophy of assessing the goodness of fit of models and distributions in the context of statistical analyses. The following sections address, in turn, the distinction between significance tests and goodness-of-fit tests, the role of alternative test criteria, and distributional plots for assessing goodness of fit, the role of frequentist goodness-of-fit criteria in model selection, as well as Bayesian approaches. Because there is an inevitable trade-off between the desire for model parsimony and models that fit well, it seems wise to give up the notion that a model might be 'true,' and instead think about choosing a model because its usefulness is reasonable.

For a detailed consideration of some specific tests and diagnostic measures in the context of assessing the adequacy of distributional assumptions see *Goodness-of-fit Tests and Diagnostics*. For detailed uses of the different approaches to goodness of fit see *Multivariate Analysis: Discrete Variables (Loglinear Models)*; *Linear Hypothesis: Regression (Basics)*; *Linear Hypothesis: Regression (Graphics)*.

1. Goodness of Fit for Categorical Data

Karl Pearson (1900) originally proposed his chi-squared test as a way to test the fit of probability model involving k parameters, by comparing a set of observations $O_1, O_2, \ldots, O_t$, in t non-overlapping classes with their expectations under the model, $E_1, E_2, \ldots, E_t$:

$$X^2 = \sum_{i=1}^{t} \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

In most practical situations, the values of $E_1, E_2, \ldots, E_t$ involve estimates of the k parameters of the model. Then for large samples, the distribution of X^2 is approximated by the χ^2 distribution with $t-k-1$ degrees of freedom when the probability model is in fact correct (see the discussion in *Goodness-of-fit Tests and Diagnostics* and in Bishop et al. 1975, for details on conditions associated with the estimates used, etc.). For categorical data problems such as those arising in the analysis of contingency table data (see *Multivariate Analysis: Discrete Variables (Overview)*), a widely used alternative to X^2 is the log-likelihood ratio test which is also referred to as the 'deviance.' Both of these are members of the power-divergence family of goodness-of-fit statistics proposed by Read and Cressie (1988):

$$K_\lambda(\mathbf{O}, \mathbf{E}) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{t} O_i \left[\left(\frac{O_i}{E_i} \right)^{\lambda} - 1 \right] \tag{2}$$

where λ is a real-valued parameter in the interval $-\infty < \lambda < \infty$. The statistic x^2 corresponds to $\lambda = 1$, and the log-likelihood ratio statistic corresponds to the limit as $\lambda \to 0$. All of the members of this power-divergence family of statistics are asymptotically equivalent. Read and Cressie (1988) describe some rationales for the choice of value of λ.

To test the fit of a specific model, e.g., the model of independence in a two-way contingency table, the chi-squared test uses an omnibus alternative, e.g., dependence. The p-value of the test, is the probability of observing a discrepancy as large or larger than that observed under the hypothesized null distribution, where 'discrepancy' is measured by the particular distance metric $K_\lambda(\mathbf{O}, \mathbf{E})$ of equation (2). But in many circumstances testing the goodness of fit of a model is not viewed as being the same as doing a test of hypothesis or test of significance. Not rejecting the null hypothesis at some arbitrary p-value is often taken as evidence that the model provides an acceptable fit to the data (cf. the discussion in Bishop et al. 1975).

Goodness-of-fit test statistics are, of course, sensitive to the sample size n, and to the real class of alternative models of interest. Suppose the null model is not exactly true (as is nearly always the case). Then, with a large enough value of n, one will almost surely conclude that the null model does not fit the data. For modest values of n the situation is more complex, and will result in small p-values for some alternatives, while for others the model may appear to fit the data reasonably well. And if the data are sparse, i.e., if the sample size, n is small relative to the number of cells t, or the number of parameters being estimated k, then the p-value computed by reference to the corresponding χ^2 distribution will inevitably suggest that the model fits even when it does not (cf., Koehler and Larntz 1980). Thus, the p-value of a chi-square test alone gives limited guidance on whether the hypothesized distribution is close enough to being correct that the conclusions of the analysis can be relied upon. When goodness-of-fit tests are used for model comparison and model search some of these problems can be attenuated, as is described in Sect. 3.

2. Other Goodness-of-fit Approaches Focused on Empirical Distributions

Other classes of goodness-of-fit tests are of special interest when one is focused on the appropriateness of a specific parametric model for some underlying cumulative distribution function, $F(x)$. What is often done in such circumstances is to compute the empirical distribution function $F_n(x)$, whose value at x is the fraction of the sample with values less than or equal to x, i.e., $\{\# x_i \leqslant x\}$, and then compare $F_n(x)$ with $F(x)$. Well known examples of such tests include the Kolmogorov–Smirnov test, the Cramer–von Mises test, the Anderson–Darling, and the Neyman smooth test. These, and related diagnostics, are described in *Goodness-of-fit Tests and Diagnostics*.

One of the most widely used diagnostic methods for checking the fit of a model is the quantile–quantile or $(Q-Q)$ plot, which plots the quantiles of the empirical distribution $F_n(x)$, against the quantiles of the assumed distribution $F(x)$. A straight line through the origin at a 45 degree slope indicates a perfect fit. The advantage of this method is that it can indicate where in the distribution the assumption is weakest. $Q-Q$ and other probability plots are described in Wilk and Gnanidesikan (1968) and are used in statistical practice. The interpretation of deviations from the line in such plots is, however, often subjective.

3. Classical Approaches to Model Comparison and Model Search

3.1 Model Comparison

Model comparison and model search are ubiquitous activities in modern statistical practice. When analysts

consider developing a regression model to describe a set of data, for example, they typically have a large number of possible variables available for inclusion in the model. What is well-known in the regression context is that the use of additional regression variables typically improves the 'predictive fit' of the model to the data (e.g., see Copas 1983) but including as many variables as possible is a poor strategy which does not hold up to scrutiny under cross-validation or when predictions are made for new observations (e.g., see the discussion in Copas, 1983, and Linhart and Zucchini 1986). Thus, at a minimum, statistical analysts engage in a series of model comparisons or tests in order to avoid the inclusion of 'non-significant' variables or terms in a model. This follows the scientific dictum known as 'Occam's razor,' attributed to William of Occam (but see Thorburn 1918), which suggests that the simpler of two models should be the one chosen. For example, *t*-tests and *F*-tests are widely used in model selection procedures for regression and analysis of variance models (see *Linear Hypothesis*). Similarly, there are various chi-square test-based techniques for selecting among the class of loglinear models applicable for contingency table structures, and these typically (although not always) resemble corresponding model selection procedures for analysis of variance and regression models (see the discussions in Bishop et al. 1975, and Fienberg 1980).

3.2 Model Search

Several methods of penalizing large models to avoid overfitting have been proposed, many taking the form of choosing a model to minimize

$$-2\log l + f(p_1, p_2, n) \tag{3}$$

where l is the likelihood ratio, f is the penalty assessed for choosing a model with p_1 parameters instead of a more parsimonious one with p_2 parameters, and n is the sample size. When the models are nested, i.e., one is a special case of the other, the quantity $-2\log l$ is the likelihood ratio test for comparing them. Under the null hypothesis that the simpler model fits, this test statistic follows a χ^2 distribution when the sample size n is sufficiently large.

The choice $f(p_1, p_2, n) = 2(p_1 - p_2)$ corresponds to Akaike's Information Criterion (Akaike 1973) or AIC, while $f(p_1, p_2, n) = (p_1 - p_2)\log n$ corresponds to Schwarz's (1978), also referred to as the Bayesian Information Criterion or BIC. Other classical approaches to this problem include Mallow's (1973) C_p statistic which is similar to AIC, and Allen's (1974) PRESS which comes from a related prediction criterion.

These and other related penalized optimization approaches to model selection have been used in the regression setting involving continuous or categorical variables. Edwards (2000), for example, uses both AIC and BIC and suggests variations on computationally efficient model search that take special advantage of the form of graphical models.

4. Bayesian Approaches to Goodness of Fit

Faced with the question of the adequacy of a particular parametric model, a natural Bayesian impulse is to embed the model in question in a parametric model of higher dimension (see Gelman et al. 1995). Having done so, how should a Bayesian evaluate the special model against the more general alternative?

One answer, espoused especially by George Box (1980) is to use a test of significance. However, the probability of observing an outcome as less or more extreme than that observed is only distantly related to goodness of fit, and violates the likelihood principle (see *Significance, Tests of*; *Likelihood in Statistics*).

A second answer, more in keeping with Bayesian ideas, is to impose a prior distribution on the larger parameter space created by embeddings. Every continuous prior distribution in this larger space has the consequence that the original model would have prior, and hence posterior, probability zero. It is for this reason that L. J. Savage advocated models 'as big as an elephant' (see Lindley 2000). A second type of prior distribution that might be used takes the form of a spike-and-slab, i.e., one that puts positive probability on the original model (the spike) and a continuous prior on the rest of the space (the slab). The posterior odds on the original model can then be expressed as the product of the prior odds and the 'Bayes factor.'

Bayes factors have attracted attention because, in the case of only two alternatives, they give a quantity that depends on only the likelihood function and that indicates the strength of the evidence. Faced with more than two alternatives, and especially with a continuum of alternatives, the Bayes factor critically depends on the prior distribution in the alternative space. When the prior is close to the special model, the alternative is favored; when it puts most of its mass far from the special case, the special case is favored, as is reasonable. Attempts to overcome this dependence using heuristics claimed to be robust to changes in the prior on the alternative space include fractional Bayes factors (O'Hagan 1995) and so-called intrinsic Bayes factors (Berger and Perrichi 1996). Another difficulty with Bayes factors, highlighted by Kadane and Dickey (1980) is that the loss function associated with them is two-valued. Hence if the wrong model is chosen, one is assumed not to care whether the error is quantitatively trivial or not. This is an unreasonable way to think about model choice.

Each of these methods is based on the notion that a better model should fit the data better. If this were true, the very best model for any data set would put a probability of one on the data actually observed, so that whatever happened had to have happened. While such a model is useless for prediction, it does show that

other considerations such as prior beliefs about how the data might have been generated, and utility considerations, must enter into a full explanation of model choice.

Bayesians too have developed approaches that are analogous to the classical approaches discussed above which penalize large models to avoid 'overfitting.' For example, for the problem of variable selection for the normal linear model, George and Foster (2000) provide interpretations of AIC, C_p, BIC, and RIC (for Risk Inflation Criterion) under a specific hierarchical Bayesian formulation. Similarly, Madigan and Raftery (1994) suggest several such strategies using posterior model probabilities, including variations of Occam's razor in order to discard models with low posterior probability. But there is limited agreement on the theoretical basis for such a strategy. Utility considerations should weigh heavily in choosing the form of a penalized search criteria, and there has not been (and it may not be possible for there to be) a convincing argument as to why a Bayesian's utility 'should' be consistent with one of these forms and not another. By contrast, Leamer (1978) presents a more flexible, but *ad hoc* Bayesian view of what he terms 'specification searches.'

Finally, we mention the proposal to use 'Bayesian model averaging' (see Hoeting et al. 1999), which combines posterior information across possible models instead of choosing directly from among them. This approach is closely related to some of the efforts to do Bayesian model selection, but has a further difficulty. The interpretation of parameters in a statistical model depends on other parameters in the model. Thus it remains unclear how to interpret posterior distributions for parameters based on combining results across posteriors for different models.

5. Conclusion

Given the goal of assessing whether a hypothesized statistical model or probabilistic distribution is so far from true as to jeopardize the intended use, it would seem wise to inquire as to the intended use. It would also seem wise not to expect a satisfactory procedure to be independent of the intended use. Second, it seems necessary to address the question of jeopardy, i.e., how much of what kind of departures might make invalid the intended inference. Clearly statisticians and others need to devote more effort to this important topic.

Part of the philosophy behind the objective of model search is to do what our recently deceased colleague Herbert Simon (1959) called 'satisficing,' that is to find a model that is good enough to capture the main features of the data but which we know is possibly imperfect in various dimensions. The late John Tukey often reminded statisticians and others that 'the best is the enemy of the good.' For many data sets and substantive problems, it is often a fruitless task to try to choose the 'best' model—one which demonstrably dominates all others and whose assumptions can be shown to be correct. But there may be many 'good' models that serve the analyst's purposes well without one that dominates. It is a good idea to pick one of them. Thus, it seems wise to give up the notion that a model might be 'true.' If, instead, models are thought of as useful ways of thinking about the world, choosing a model because of its usefulness is reasonable, even if formal ways of explaining that choice are now lacking.

For a detailed consideration of some specific tests and diagnostic measures in the context of assessing the adequacy of distributional assumptions see *Goodness-of-fit Tests and Diagnostics*. Raynor and Best (1989) provide a good technical discussion of specific goodness-of-fit test statistics, and Read and Cressie (1988) give an extended treatment of the power divergence family of statistics and their use for the analysis of categorical data models. Bishop et al. (1975), discuss model search and goodness of fit in the context of categorical data in a form consistent with the philosophical position presented here. Edwards (2000) and Linhart and Zucchini (1986) provide alternative, but useful discussions of related issues for various forms of linear and generalized linear models. Gelman et al. (1995) offer extended advice on Bayesian approaches to model checking.

See also: Goodness-of-fit Tests and Diagnostics; Likelihood in Statistics; Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Discrete Variables (Overview); Significance, Tests of

Bibliography

Akaike H 1973 Information theory and an extension of the maximum likelihood principle. In: Akaike H (ed.) *2nd International Symposium on Information Theory*. Akademia Kiado, Budapest, Hungary, pp. 267–81

Allen D M 1974 The relationship between variable selection and prediction. *Technometrics* **16**: 125–7

Berger O, Pericchi L R 1996 The intrinsic Bayes factor for linear models. In: Bernardo J M, Berger J O, Dawid A P, Smith A F M (eds.) *Bayesian Statistics 5*. Oxford University Press, Clarendon, Oxford, UK, pp. 25–44

Bishop Y M M I, Fienberg S E, Holland P W 1975 *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA

Box G 1980 Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A* **143**: 383–430

Copas J B 1983 Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B* **45**: 311–54

Edwards D 2000 *Introduction to Graphical Modeling*, 2nd edn. Springer-Verlag, New York

Fienberg S E 1980 *The Analysis of Cross-classified Categorical Data*, 2nd edn. MIT Press, Cambridge, MA

George E I, Foster D P 2000 Calibration and empirical Bayes variable selection. *Biometrika* **87**: 731–47

Gelman A, Carlin J B, Stern H S, Rubin D B 1995 *Bayesian Data Analysis*. Chapman & Hall, London
Hoeting J A, Madigan D, Raftery A E, Volinsky C T 1999 Bayesian model averaging: A tutorial (with discussion). *Statistical Science* **14**: 382–417
Kadane Y B, Dickey J O 1980 Bayesian decision theory and the simplification of models. In: Kmenta J, Ramsey J (eds.) *Evaluation of Econometric Models*. Academic Press, New York, pp. 245–68
Koehler K J, Larntz K 1980 An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association* **75**: 336–44
Leamer E E 1978 *Specification Searches: Ad Hoc Inference With Nonexperimental Data*. Wiley, New York
Lindley D V 2000 The philosophy of statistics (with discussion). *The Statistician* **49**: 293–337 (see p. 307)
Linhart H, Zucchini W 1986 *Model Selection*. Wiley, New York
Madigan D, Raftery A E 1994 Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**: 1535–46
Mallows C L 1973 Some comments on C_p. *Technometrics* **15**: 661–76
O'Hagan A 1995 Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society*, *Series B* **57**: 99–138
Pearson K 1900 On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 5th Series* **50**: 157–75
Rayner J C W, Best D J 1989 *Smooth Tests of Goodness of Fit*. Oxford University Press, Clarendon, Oxford, UK
Read T R C, Cressie N A C 1988 *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York
Schwarz G 1978 Estimating the dimension of a model. *Annals of Statistics* **6**: 461–4
Simon H A 1959 Theories of decision-making in economics and behavioral science. *American Economic Review* **49**: 253–73
Thorburn W M 1918 The myth of Occam's razor (in Discussion). *Mind, New Series* **27**: 345–53
Wilk M B, Gnanadesikan R 1968 Probability plotting methods for the analysis of data. *Biometrika* **55**: 1–17

S. E. Fienberg and J. B. Kadane

Goodness-of-fit Tests and Diagnostics

The millennium year 2000 marked the 100th anniversary of the introduction of the first and possibly best known test of modern statistical inference: the chi-squared test of fit due to Pearson (1900). The purpose of a test of fit is to give an objective measure of how well a probability model agrees with observed data. Otherwise, as Pearson (1900, p. 171) said: 'the comparison of observation and theory in general amounts to a remark—based on no quantitative criterion.' Reviews of Pearson's paper may be found in Plackett (1983) and Rayner and Best (1989, Sect. 2.2). Rayner and Best (1989) identify three benefits of having applied a goodness-of-fit test:

- a compact description of the data;
- illumination of the data generation mechanisms; and
- validation, if that is the outcome, of distributional assumptions necessary for the application of powerful parametric procedures.

Pearson's chi-squared statistic compares the *observed* counts from a histogram of the data with the counts *expected* assuming that the hypothesized probability model is correct. In this form, Pearson's test addresses the one sample goodness-of-fit problem (see also *Goodness of Fit: Overview*). The multiple sample goodness-of-fit problem seeks to assess if data from a number of populations are consistent with having come from the same population. If the answer is affirmative, then the one sample question may be appropriate. We shall not address the multiple sample goodness-of-fit problem further here, beyond recommending Scholz and Stephens (1987) and Rayner and Rayner (1988) for references and recent work.

For continuous data and for discrete data for which pooling is used, the grouping of the data used in forming the Pearson chi-squared statistic leads to a loss of information. Another approach that does not group the data is to compare the sample or *empirical distribution function* (EDF) with the distribution function specified by the hypothesized probability distribution. The Russian mathematician Kolmogorov (1933) proposed the first EDF test of fit. EDF tests of fit, especially for continuous distributions, are reviewed in D'Agostino and Stephens (1986). Tests of fit for discrete distributions using the EDF approach and Cramer–Von Mises statistics are considered in Choulakian et al. (1994). See also Spinelli and Stephens (1997).

Another problem with Pearson's chi-squared test is that it is an omnibus test, with some power against general alternative distributions to the hypothesized distribution, but with relatively poor power against important directional alternatives. Alternatives which are often of great practical importance are differences between the data and the mean, variance and higher moments of the specified distribution. A test that does have good power at detecting these moment alternatives is Neyman's (1937) 'smooth' test. It has the useful property that it can readily be partitioned into components that give a powerful and detailed scrutiny of the data. See Rayner and Best (1989, Sect. 4.1) for a review of Neyman (1937).

In a series of papers, Rayner and Best developed the Neyman smooth tests to cover both categorized distributions and distributions involving nuisance parameters. See Rayner and Best (1989) for an overview, and their subsequent related papers. In the main, Rayner and Best use orthonormal functions, as did Neyman (1937), but permit nuisance parameters, following Thomas and Pierce (1979) and Kopecky and

Pierce (1979). The Rayner and Best smooth tests may be used to test for very general distributions, provided orthonormal functions on the specified distribution may be constructed. The resulting test is omnibus, weakly optimal, and the test statistic has a convenient asymptotic chi-squared distribution. Components are readily available, and these are asymptotically independent, each asymptotically having the standard normal distribution. The components are often well known tests of fit in their own right, but more importantly, permit a detailed, convenient and informative scrutiny of the data.

LaRiccia (1991) and Pena (1998) have developed different classes of smooth tests of fit using this approach.

The chi-squared tests, the EDF tests and the smooth tests of fit are the main families of tests of fit. Other tests of fit have been suggested, but we shall not consider these here, except to note that correlation tests are sometimes suggested in conjunction with quantile–quantile or probability plots of the data. The Shapiro–Wilk test of normality, for example, may be considered a correlation test. Again see D'Agostino and Stephens (1986).

1. Chi-squared Tests

For Pearson's χ^2 test, sometimes called the chi-squared test, observations $O_1, O_2, \ldots, O_m$ are assumed to come from m non-overlapping classes that are expected to contain $E_1, E_2, \ldots, E_m$ observations. The test statistic is

$$\chi^2_{\mathrm{P}} = \sum_{i=1}^{m} (O_i - E_i)^2 / E_i.$$

Pearson (1900) initially assumed that for his test of fit, parameters of the probability models were known, and showed that the asymptotic distribution of χ^2_{P} is the χ^2_{m-1} distribution.

There is a considerable literature on how, when there is choice, the cells may best be constructed. The recommendation that for the χ^2 approximation to be adequate, the cell expectation be at least 5, goes back to Sir Ronald Fisher writing in the 1920s. This advice is now seen as extremely conservative. Mann and Wald (1942) suggested that cells be constructed to be equiprobable and that there be a specified number, k say, of such cells. In their recommendation, k increases with the sample of size n, and for even moderate sample sizes this will lead to a relatively large number of cells. It is now known that this can lead to a considerable loss of power, that sometimes the fewer the number of cells, the greater the power. Recent studies suggest that the greatest power can be achieved for heavy tailed distributions with larger k, while lighter-tailed alternatives require smaller k. Roscoe and Byars (1971) recommend that, with equiprobable cells, the average cell frequency should be at least one with a 5 percent significance level, and at least two with a 1 percent significance level. If cells are not approximately equiprobable, these recommended frequencies should be doubled.

Most interesting applications involve estimating parameters. Pearson asserted that estimating parameters makes no difference. If the cell expectations are calculated from parameters estimated efficiently from the grouped data, then Sir Ronald Fisher, again writing in the 1920s, proved that degrees of freedom for the asymptotic distribution of the test statistic, χ^2_{PF} say, needed to be reduced by one for each parameter estimated. The asymptotic distribution is χ^2_{m-q-1} if q parameters are estimated. Chernoff and Lehmann (1954) showed that Pearson's statistic no longer had an asymptotic chi-squared distribution if parameters were estimated from the raw or ungrouped data. There is a 'partial recovery' of the q degrees of freedom 'lost' by estimation. Rao and Robson (1974) gave a test of fit statistic χ^2_{RR} that had an asymptotic χ^2_{m-1} distribution no matter how many parameters were estimated, provided they are estimated efficiently from the ungrouped data. So all the degrees of freedom 'lost' by estimation are recovered by their statistic. Rayner and Best (1989, Sect. 7.4) compared the power performance of the tests based on χ^2_{PF} and χ^2_{RR}, and concluded that neither is uniformly superior: recovery of the degrees of freedom does not necessarily give a more powerful test.

There are many possible generalizations of the Pearson's test. Neyman's version is based on

$$\chi^2_{\mathrm{N}} = \sum_{i=1}^{m} (O_i - E_i)^2 / O_i.$$

One interesting possibility is the Cressie–Read family, based on the test statistic

$$I_\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{m} O_i \left\{ \left(\frac{O_i}{E_i} \right)^\lambda - 1 \right\}.$$

Taking $\lambda = 1$ leads to χ^2_{P}, $\lambda = -1$ leads to a modified likelihood ratio statistic, $\lambda = 0$ is the usual likelihood ratio statistic, $\lambda = 0.5$ is the Freeman–Tukey statistic, and $\lambda = -2$ leads to χ^2_{N}. Other choices can lead to tests with excellent properties; see Read and Cressie (1988). Read and Cressie (1988) also illustrate the application χ^2 to testing fit of loglinear models (see also *Multivariate Analysis: Discrete Variables (Overview)*; *Multivariate Analysis: Discrete Variables (Loglinear Models)*).

2. EDF Tests

If we have n data points, the empirical distribution function (EDF) is

$$F_n(x) = (\text{number of observations} \leqslant x)/n.$$

Tests of fit for a hypothesized distribution function, $F(x)$, can be made by looking at the difference between $F_n(x)$ and $F(x)$ for each data point. Kolmogorov (1933) suggested the statistic

$$KS = \max_x |F_n(x) - F(x)|$$

while a generally more powerful test, is based on the Anderson–Darling (1954) statistic. The formula for this statistic can be simplified for various applications, as we illustrate below.

As, for example, in Stephens (1974), it can be shown that the EDF test statistics are linear combinations of components similar to the smooth test components, with decreasing weights. If the null hypothesis specifies a discrete distribution where the number of classes is fixed and not too large (say five or less) we suggest calculating χ^2 to assess goodness-of-fit. Otherwise, the Anderson–Darling test is a good choice, as it weights the important early moment components heavily and gives a powerful omnibus assessment. Looking at the first few components of Neyman's smooth test in an exploratory data analysis manner complements this omnibus assessment. See the discussion in Sect. 4 below.

2.1 *Example 1: Testing Normality*

Consider the following set of examination marks from 20 students:

53, 15, 70, 73, 79, 48, 91, 20, 24, 91,
87, 15, 3, 78, 78, 62, 16, 15, 20, 32.

Are these data normally distributed? We now apply the Anderson–Darling test to answer this question. Calculate $\bar{x} = 48.5$ and $s = 30.8$ and then find $y_i = (x_i - \bar{x})/s$ for $i = 1, \ldots, 20$, where x_i are the 20 exam marks. Next use a computer routine or tables of the standard normal distribution to find

$$z_i = F(y_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_i} \mathrm{e}^{-\frac{1}{2}x^2} \, \mathrm{d}x$$

and order the z_i as $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$. Then the Anderson–Darling statistic, AD say, is

$$AD = -n - n^{-1} \sum_{i=1}^{n} (2i-1)[\log z_{(i)} + \log(1 - z_{(n+1-i)})],$$

which, for the above data, takes the value 0.931. From tables of critical values such as D'Agostino and Stephens (1986, Table 4.9), or using MINITAB's Normplot routine, we find a p-value of 0.014. This implies the normality hypothesis is doubtful and the marks are not well summarized by $\bar{x} \pm s$. We suggest a better summary below.

2.2 *Example 2: Testing the Poisson Assumption*

Consider the number of people in a shopping mall contributing to a charity collector during five-minute intervals:

Number of contributors	0	1	2	3	4	5	6	7	8
Number of intervals	9	27	26	17	10	8	4	3	1

We have $n = 105$. Let the observed and expected counts assuming a Poisson distribution be o_j and $e_j = np_j$, respectively, for $j = 0, 1, \ldots, 8$. Now let $z_j = \sum_{i=0}^{j} (o_i - e_i)$ and $h_j = \sum_{i=0}^{j} p_i$, in which case

$$AD = n^{-1} \sum_{j=0}^{8} z_j^2 \, \hat{p}_j / [\hat{h}_j (1 - \hat{h}_j)]$$

where the '^' indicates we have estimated the Poisson parameter by the sample mean. We find that $AD = 1.17824$, and, using Spinelli and Stephens (1997 Table 1), the p-value is between 0.025 and 0.050. Thus the distribution is not well fitted by a Poisson model. We discuss a possible alternative distribution in Sect. 4. It seems plausible that contributors come in clusters rather than at random, as a Poisson model would imply.

3. *Neyman Smooth Tests*

Suppose we wish to test for a probability density function $f_X(x; \beta)$, where $\beta = (\beta_1, \ldots, \beta_q)^{\mathrm{T}}$ is a q by 1 vector of parameters (such as μ and σ in the normal case). This null probability density function is nested in the order k probability density function

$$C(\theta, \beta) \exp \left\{ \sum_{i=1}^{k} \theta_i \pi_i (x; \beta) \right\} f_X(x; \beta),$$

where $\theta = (\theta_1, \ldots, \theta_k)$ is a vector of k real parameters, $C(\theta, \beta)$ is a normalizing constant and $\{\pi_i(x; \beta)\}$ is a set of functions orthonormal on $f_X(x; \beta)$, so that for $f_X(x; \beta)$ continuous,

$$\int_{-\infty}^{\infty} \pi_i(x, \beta) \pi_j(x, \beta) f_X(x; \beta) \, \mathrm{d}x = \delta_{ij} = 1$$

for $i = j$ and zero otherwise.

Sometimes $C(\theta, \beta)$ may not exist but the statistic we are about to define will. Using a random sample $X_1, \ldots, X_n$ we test for $f_X(x; \beta)$ parametrically, by testing H_0: $\theta = 0$ versus K: $\theta \pm 0$ with β a vector of nuisance parameters. For small k the alternatives vary 'smoothly' from the null, in the sense of Neyman (1937). As an example of this, in testing for the standard normal distribution, the order one alternative is a normal distribution with a mean shift, the

order two alternative is a normal distribution with both a mean shift and a variance shift, and the third order alternative is more involved, with skewness differences to what might be expected as the distribution normal.

Neyman (1937) developed this approach for the case of no nuisance parameters. In this situation the probability integral transformation converts the problem of testing for a completely specified distribution to testing for the uniform (0, 1) distribution. Thomas and Pierce (1979) and Kopecky and Pierce (1979) considered smooth tests for general distributions involving nuisance parameters. These tests are based on a family of alternatives specified using powers of the cumulative distribution function rather than orthogonal polynomials as Neyman (1937) used, and as a consequence they require tables of coefficients for their implementation. The Rayner and Best formulation permits nuisance parameters and uses orthonormal functions, eliminating the need for tables of coefficients. For a more detailed account of the history of the smooth tests, see Rayner and Best (1989, Sect. 1.2).

When nuisance parameters are present, the smooth tests involve their maximum likelihood estimators. In several important cases, such as the binomial, Poisson and normal, these estimators are the solutions to the equations

$$\hat{V}_r = \sum_{j=1}^{n} \pi_r(X_j, \hat{\beta})/\sqrt{n} = 0, \quad r = 1, \ldots, q.$$

In such cases the smooth test is now based on the statistic

$$\hat{S}_k = \hat{V}^2_{q+1} + \ldots + \hat{V}^2_{q+k}.$$

The $\hat{V}_r$ are components of $\hat{S}_k$, and provide directional tests with high power in one direction and poor power elsewhere, complementing the omnibus test based on $\hat{S}_k$, which has moderate power in the k directions corresponding to the non-zero components. The omnibus test statistic $\hat{S}_k$ has the χ^2_k distribution, central under the null distribution and noncentral under contiguous alternatives. It can be shown that the score test for testing H_0: $\theta_r = 0$ against K_r: $\theta_r \pm 0$ is based on the statistic $\hat{V}^2_r$. Being score tests, these tests are weakly optimal. If there are no nuisance parameters, no estimation is required and the results indicated hold with $q = 0$ and the ^s removed from the test statistics. Boulerice and Ducharme (1995) give an example of what happens when the maximum likelihood estimators are *not* solutions of $\hat{V}_r = 0$ for $r = 1, \ldots, q$.

We should emphasize that different $\{\pi_i(x; \beta)\}$ detect different alternatives using different tests. Which $\{\pi_i(x; \beta)\}$ should be chosen depends on which alternatives one hopes to detect most powerfully. If the distribution were uniform, as in Neyman (1937), we could take the (non-orthogonal) series $\{x^i\}$. The rth component would still compare the data and moments of the hypothesized distribution up to the rth. However, we prefer, like Neyman, to use orthonormal functions because the components are simply defined and are, at least asymptotically, independent. If it were desirable to test for periodic alternatives we would consider using the series $\{\sqrt{2}\sin(i\pi\sqrt{x})\}$. The aim is to find an orthonormal series that represents the alternatives of interest in as few terms as possible. Greater power results from doing this.

Each family of distributions to be detected requires an independent study. Given the choice of orthonormal family, the small sample distributions of the test statistics and their components should be investigated. Usually we use regression techniques to calculate corrections to the asymptotic null critical points; see Rayner and Best (1989) for details.

Again for each family of distributions we may compare, via simulation studies, the powers of the smooth tests we propose and the commonly used competitor tests. We give the general advice that in a testing context the most powerful goodness-of-fit test is likely to be the sum of the squares of the first two, three, or four non-zero components; $\hat{S}_4$ perhaps. We recommend augmenting this by the use of its components $\hat{V}_{q+1}, \ldots, \hat{V}_{q+4}$ in an exploratory data analysis fashion. However, in particular cases there are variations to this advice.

We have so far studied the continuous and discrete uniform, univariate, bivariate and multivariate normal, exponential, geometric, binomial, univariate and bivariate Poisson. Details are in Rayner and Best (1989) and their subsequent relevant papers. In all cases the smooth tests are generally at least as powerful as the competitors studied, and the smooth tests have the added flexibility of meaningful components being available. So if the null hypothesis is rejected, its failure can be attributed to particular θ_i being non-zero, and an alternative model is suggested, namely

$$C(\hat{\theta}, \hat{\beta}) \exp\left\{ \sum_{i=q+1}^{q+k} \hat{\theta}_i \pi_i(x; \hat{\beta}) \right\} f_X(x; \hat{\beta})$$

where the $\hat{\theta}_i = \hat{V}_i/\sqrt{n}$ are unbiased, consistent and efficient estimators of the θ_i. This density estimate is a fortuitous by-product of the testing, and may be used to graph the data or as a basis for further inference, as in Carolan and Rayner (2001).

4. *Diagnostics*

In Sect. 2 we suggested use of the Anderson–Darling statistic as a test of fit. However, this will only tell us whether a given probability model fits the data well or not. It does not suggest or 'diagnose' how the data differs from the probability model. A histogram, density estimate or a Q–Q plot might shed some light on this difference, but interpretation of these plots is subjective. A more objective data scrutiny is given by calculating the components of Neyman's smooth test which are approximately distributed as independent

standard normal random variables. We now illustrate this 'diagnostic' approach in our previous examples and one further example. There is no guarantee the diagnosis is correct (see Rayner et al. 1995), but it will be suggestive.

4.1 Example 1: Testing Normality Revisited

For the exam mark data in Section 2.1, calculate $y_i = (x_i - \bar{x})/s$. The orthonormal polynomials appropriate for testing normality are often called Hermite–Chebyshev polynomials. We have

$$\begin{aligned}
\pi_1(y) &= y \\
\pi_2(y) &= (y^2-1)/\sqrt{2} \\
\pi_3(y) &= (y^3-3y)/\sqrt{6} \\
\pi_4(y) &= (y^4-6y^2+3)/\sqrt{24} \\
\pi_5(y) &= (y^5-10y^3+15y)/\sqrt{120} \\
\pi_6(y) &= (y^6-15y^4+45y^2-15)/\sqrt{720}
\end{aligned}$$

and given these values we can calculate, for the above data, $\hat{V}_1 = \hat{V}_2 = 0$, due to the estimation of μ and σ from $\bar{x}$ and s respectively, and $\hat{V}_3 = -0.0361$, $\hat{V}_4 = -1.4656$, $\hat{V}_5 = -0.0462$, $\hat{V}_6 = 1.8907$. Note that $\hat{V}_3$ and $\hat{V}_4$ are standardized versions of the usual skewness and kurtosis coefficients. As $\hat{V}_3, \ldots, \hat{V}_6$ have asymptotic standard normal distributions, we see that $\hat{V}_6$ is near significance at the 5 percent level and that $\hat{V}_4$ and $\hat{V}_6$ are much larger than the odd-order moment statistics $\hat{V}_3$, $\hat{V}_5$. These observations suggest an alternative probability model that is symmetric and shorter tailed than the normal. Thus a uniform distribution is a possibility so that we could summarize the data as 'evenly' spread over the range (0, 100) rather than as 48.5 ± 30.7. If the odd order $\hat{V}_r$ had dominated the even order values then a skewed alternative model is suggested, while if both odd and even $\hat{V}_r$ are large, $|\hat{V}_r| > 1.65$ say, then a symmetric alternative with longer tails then the normal, such as the Laplace, is indicated.

What should be done if one or more $\hat{V}_r$ are significant but AD is not? One possibility would be to obtain more data and test the significance of those $\hat{V}_r$ previously significant.

4.2 Example 2: Testing the Poisson Assumption Revisited

For the Poisson, the appropriate orthonormal polynomials are known as Poisson–Charlier polynomials. We have

$$\pi_r(y) = \sqrt{(\lambda^r/r!)} \sum_{j=0}^{r} [(-1)^{r-j}]\,({}^rC_j)\,(j!)\,(\lambda^{-j})\,({}^yC_j)$$

where $y = 0, 1, 2, \ldots$, from which the first two orthonormal polynomials are

$$\begin{aligned}
\pi_1(y) &= \sqrt{\lambda}\,(y/\lambda - 1) \\
\pi_2(y) &= \sqrt{\lambda^2/2}\,[y(y-1)/\lambda^2 - 2y/\lambda + 1].
\end{aligned}$$

Due to the estimation of λ, $\hat{V}_1 = 0$. For the data here $\hat{V}_2 = 2.12$, indicating a model which has greater dispersion than the Poisson. Note that $\hat{V}_2$ is a standardized version of Fisher's Poisson Index of Dispersion.

4.3 Example 3: Testing the Ordered Categories Discrete Uniform Distribution

In a market research trial, 40 consumers were asked for their color preference of five orange cordial drinks with orange colors ranging from 'pale' to 'deep.' Arranged in order of increasing orange color, the preference counts were 5, 10, 11, 10 and 4. If there were no color preference, the counts would all be expected to be eight. As there are only five unambiguous classes, a Pearson χ^2 could be used here. We have

$$\chi^2 = (5-8)^2/8 + (10-8)^2/8 + \ldots + (4-8)^2 = 5.25$$

which, on using the asymptotic χ^2_4 distribution, gives a p-value of 0.27. Using the same formula for AD as that used for the Poisson test, we find the Anderson–Darling statistic takes the value 0.89. Consulting tables in Choulakian et al. (1994) we find the p-value is about 0.30. It appears the data are consistent with a discrete uniform distribution. However, we now compare expected and observed moments via components of Neyman's smooth test. Here the orthonormal polynomials are the Chebyshev polynomials and the first two are

$$\begin{aligned}
\pi_1(y) &= \sqrt{[12/(m^2-1)]}\,[y-(m+1)/2] \\
\pi_2(y) &= \sqrt{\{180/[(m^2-1)(m^2-4)]\}}\,\{[y-(m+1)/2]^2 \\
&\quad -(m^2-1)/12\}
\end{aligned}$$

where m is the number of categories and $y = 1, 2, \ldots, m$. The higher order $\pi_r(y)$, if needed, can be found by recurrence relation as in Emerson (1968). As in Stephens (1974), it can be shown that for the discrete uniform model

$$\chi^2 = \sum_{i=1}^{m} V_i^2 \quad \text{and} \quad AD = \sum_{i=1}^{m} [i(i+1)]^{-1} V_i^2$$

where the components are defined by

$$V_r = \sum_{j=1}^{n} \pi_r(X_j)/\sqrt{n}, \quad r = 1, \ldots, m$$

since no nuisance parameters need estimation.

For these data, neither χ^2 nor AD was significant. However $V_1 = \sqrt{0.05} = 0.22$ and $V_2 = \sqrt{5.14} = 2.27$, indicating possible extra dispersion in the data compared with what was expected for the discrete uniform. We find that V_2^2 explains almost all χ^2, so that dispersion differences are all that can be important. The consumers seem to have preferred a mid-range orange-colored drink. Unless it is the first component, AD will not be powerful if only one component is important, as in the present example. However, for many data sets, more than one component will be important, so that AD should usually have reasonable power.

In the above examples, examination of smooth test components gives an informative and thorough scrutiny of the data. Formal tests of significance would require the Bonferoni adjustment to the α level; here we are recommending an informal exploratory data analysis approach.

5. Conclusion

Statistical model assessment is at the heart of good statistical practice, and is the genesis of modern statistics (see *Goodness of Fit: Overview*). In general we suggest using the Anderson–Darling test as an omnibus test, augmented by the use of the components of the smooth tests in an exploratory data analysis fashion. This will assist in diagnosing the type of deviation from the hypothesized distribution. To find out more about the chi-squared and EDF tests see D'Agostino and Stephens (1986). For more about the smooth tests see Rayner and Best (1989).

Bibliography

Anderson T W, Darling D A 1954 A test of goodness of fit. *Annals Mathematical Statistics* **23**: 193–212

Boulerice B, Ducharme G R 1995 A note on smooth tests of goodness of fit for location-scale families. *Biometrika* **82**: 437–8

Carolan A M, Rayner J C W 2001 One sample score tests for the location of modes of nonnormal data. *Journal of Applied Mathematics and Decision Sciences* **5**(1): 1–19

Chernoff H, Lehmann E L 1954 The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals Mathematical Statistics* **25**: 579–86

Choulakian V, Lockhart R A, Stephens M A 1994 Cramer–Von Mises statistics for discrete distributions. *Canadian Journal of Statistics* **22**: 125–37

D'Agostino R B, Stephens M A 1986 *Goodness-of-fit Techniques*. Marcel Dekker, New York

Emerson P L 1968 Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics* **24**: 695–701

Kolmogorov A N 1933 Sulla determinazione empirica di una legge di sitribuzione. *Giornale dell' Instituto Italiano desli Attuari* **4**: 83–91

Kopecky K J, Pierce D A 1979 Efficiency of smooth goodness-of-fit tests. *Journal of the American Statistical Association* **74**: 393–7

LaRiccia V N 1991 Smooth goodness-of-fit tests: a quantile function approach. *Journal of the American Statistical Association* **86**: 427–31

Mann H B, Wald A 1942 On the choice of the number of intervals in the application of the chi-squared test. *Annals Mathematical Statistics* **13**: 306–17

Neyman J 1937 'Smooth' test for goodness of fit. *Skandinavisk Aktuarietidskr* **20**: 150–99

Pearson K 1900 On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophy Magazine*, 5th series **50**: 157–75

Pena E A 1998 Smooth goodness-of-fit tests for the baseline hazard in Cox's proportional hazards model. *Journal of the American Statistical Association* **93**: 673–92

Plackett R L 1983 Pearson Karl and the chi-squared test. *International Statistical Review* **51**: 59–72

Rao K C, Robson D S 1974 A chi-square statistic for goodness-of-fit tests with the exponential family. *Communication in Statistics*. **3**: 1139–53

Rayner J C W, Best D J, 1989 *Smooth Tests of Goodness of Fit*. Oxford University Press, New York

Rayner J C W, Best D J, Mathews K L 1995 Interpreting the skewness coefficient. *Communications in Statistics—Theory and Methods* **24**(3): 593–600

Rayner J C W, Rayner G D 1988 S-sample smooth goodness of fit tests: Rederivation and Monte Carlo assessment. *Biometrical Journal* **40**: 651–63

Read T R C, Cressie N A C 1988 *Statistics for Discrete Multivariate Data*. Springer-Verlag, New York

Roscoe J T, Byars J A 1971 Investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association* **66**: 755–9

Scholz F W, Stephens M A 1987 K-sample Anderson–Darling tests. *Journal of the American Statistical Association*. **82**: 918–24

Spinelli J J, Stephens M A 1997 Cramer–Von Mises tests of fit for the Poisson distribution. *Canadian Journal of Statistics* **25**: 257–68

Stephens M A 1974 Components of goodness-of-fit statistics. *Annals of the Institute of Henri Poincare* **10**: 37–54

Thomas D R, Pierce D A 1979 Neyman's smooth goodness-of-fit test when the hypothesis is composite. *Journal of the American Statistical Association* **74**: 441–5

J. C. W. Rayner and D. J. Best

Government Lawyers

Government lawyers are those officers and agents that provide legal counsel to government policy makers and represent governments in judicial proceedings. The term 'government lawyers' is not common in social science literature, and, generally speaking, studies of government lawyering focus on individual

offices (such as attorneys general, ministers of justice, solicitors general, district attorneys, etc.), or the legal policies of particular governments or administrations (such as the Thatcher Government or the Bush Administration, etc.). The reluctance to treat government attorneys as a group stems from the fact that they perform a wide variety of roles, from routine criminal prosecution to advising presidents or prime ministers on the broad policy concerns embedded in constitutional or statutory law. Moreover, the role of government lawyers, like the role of lawyers and courts generally, varies a great deal from one political system to the next. Nevertheless, commonalities do exist, and scholarly attention to the role of government lawyers has grown as their political importance has expanded alongside the legal-administrative and regulatory state.

1. Size, Structure, and History of Government Legal Offices

The structure, number, and size of government law offices usually reflects the broader constitutional structures and policy making processes of the regime. Thus, federal or confederal systems usually have separate law officers at the national level and at the state, *Länder*, or canton levels. Likewise, separation of power systems tend to have more legal offices than parliamentary systems, as during periods of conflict each branch of government finds it necessary to access independent legal advice and representation. In countries such as the United States, where there is federalism, a separation of powers, and where courts exercise sweeping policy making powers, governmental lawyers are both numerous and important policy makers (Clayton 1995).

There is no accurate count of the total number of lawyers employed by governments around the world, but in the USA alone the Department of Justice employs nearly 11,000 attorneys and has a budget in excess of $21 billion (DOJ website 2000). Thousands more are employed in the general counsels' offices of various executive branch agencies and departments, including the White House Legal Counsel's Office. Congress too employs numerous attorneys, either in the House or Senate Legal Counsel's offices, or as staff counsel to various committees. Even the Supreme Court, since 1972, has employed attorneys in its own in-house legal counsel's office to advise the Court and its members. At the state level, governmental law offices range in size from the Department of Law in California, which employs nearly 1,000 lawyers, to relatively small Attorney General's offices in states such as Montana. In addition, many more attorneys work for cities, counties, municipalities, and other local governmental units (Clayton 1994).

The premier government attorney in the USA is the Attorney General. The office was established under the Judiciary Act of 1789, and originally operated as a quasi-judicial institution. By the time Congress created the Department of Justice in 1870, the office had become thoroughly entrenched as part of the executive branch establishment. The Attorney General is appointed by the President and confirmed by the Senate. The office is a member of the Cabinet and a key advisor to the President on judicial selection and the making of administration policy. The Attorney General's transformation from quasi-judicial to executive, however, left deep tensions in its normative relationship to the President, especially during periods of separation of powers conflict (see Baker 1992). The office's relationship to other government lawyers has also changed over time. Originally, the Attorney General had little authority over the US District Attorneys (prosecutors in each federal judicial district), or over the solicitors that existed in executive branch departments and agencies. The early history of the Attorney General's office is marked by struggles to consolidate control over the federal governments' legal activities. Eventually, a division of labor emerged wherein agency legal counsel have assumed primary responsibility for advising and counseling government agencies and departments, and the Justice Department, under the direction of the Attorney General, has assumed nearly complete responsibility for conducting federal litigation and prosecution (see Horowitz 1977).

The Attorney General is an administrator of a large department. The office is assisted by a Deputy Attorney General, an Associate Attorney General, and a Solicitor General. The latter has evolved into an elite barristers office, responsible for conducting nearly all of the federal government's appellate litigation, enjoying a special relationship with the Supreme Court (Sakokar 1992). The Justice Department is divided into 15 major divisions and offices, including important policing and corrections divisions such as the Federal Bureau of Investigation, the Drug Enforcement Administration, the Marshals Service, the Immigration and Naturalization Service, and the Bureau of Prisons. The most important legal divisions, each headed by an Assistant Attorney General, include: Antitrust, Civil Rights, Land and Natural Resources, Criminal, Civil, and the Tax divisions. Additionally, the Office of Legal Counsel, also headed by an Assistant Attorney General, is responsible for authoring the official opinions of the Attorney General. Typically, these are considered binding within the executive branch but not binding on courts, although they are usually accorded respect by the judiciary.

In the USA, most state attorney generals' offices or law departments mirror the subject-area structure of the Justice Department's legal divisions. Thus, they are likely to have divisions for antitrust, tax, consumer protection, corporate regulation, etc. Several important distinctions exist at the state level in the USA, however. First, in most states the attorney general is responsible for providing legal advice to all branches

of government. Second, responsibility for nonlawyering functions (police, prisons, etc.) and for most criminal investigations and prosecution are not usually under the control of the attorney general, but under the control of other state-level administrators or locally elected prosecutors. Third, most state attorneys general are elected directly and politically independent of the governor and legislature (Clayton 1994).

In most other countries, the lawyering activities that involve broad policy concerns are usually also separated from those related to routine prosecutions, corrections, and policing. In many, an attorney general, or some similar office, is responsible for advising the government on legal matters, and for litigation enforcing public or political rights, while responsibility for prosecution, police, corrections, and the administration of justice is vested in a Minister of Justice or a Home Secretary.

In England, the Attorney General's title dates back to 1472. The modern office typically is held by a member of the House of Commons and an elected member of the majority political party. By convention, the Attorney General is not usually a member of the Cabinet, although most attend Cabinet meetings. The Solicitor General serves as a chief deputy. These two officials are known collectively as 'the Law Officers of the Crown' (Edwards 1964). Unlike its American counterpart, the Attorney General in England is not an administrator or department head. However, the office has supervisory responsibility over the Treasury Solicitor, the Director of Public Prosecutions as head of the Crown Prosecution Service, the Director of the Serious Fraud Office, and the Director of Public Prosecutions in Northern Ireland, and answers for each of these departments in Parliament.

The Law Officers' primary responsibility is to represent the public interest in court and act as the chief legal advisers to the government. On matters of political importance, the Attorney General personally will offer legal advice to the government. Routine legal advice, however, is provided to each department or ministry by the Treasury Solicitor's Department, or by small in-house departmental legal staffs, which are analogous to the general counsel's offices in various departments in the USA. The Treasury Solicitor's Department is the largest in-house legal organization and provides legal advice to administrative officials across the executive, including the departments of Treasury, the Cabinet Office, the Ministry of Defence, and the Departments for Education and Employment, and Culture, Media and Sport. The Department employs several hundred attorneys and it also assists the Attorney General in litigation to protect the public interest.

The Home Secretary is responsible for the police, and for administering immigration, the correctional system, and other law enforcement agencies. The appointment of judges is handled exclusively by the Lord Chancellor's Office, a part of the House of Lords. The chief prosecutor in England is the Director of Public Prosecutions, an officer appointed by the Home Secretary and located in the Home Office, but who is responsible to the Attorney General in the exercise of prosecutorial discretion. The Director heads the Crown Prosecution Service, which is divided administratively into 42 separate areas, each corresponding to the boundaries of local police forces. Each area is headed by a Chief Crown Prosecutor who works with panels of local attorneys and case workers to prosecute crime in the crown courts. Crown Prosecutors report to the Director of Public Prosecutions, but in routine matters they are locally accountable to the public and cooperate closely with the area Chief Constable andother agencies in the criminal justice system. If criminal prosecutions raise serious political concerns, however, the Director of Public Prosecutions or the Attorney General may become involved and assume responsibility for the prosecution. As with much else in English government, prosecutorial independence is protected through custom and convention. Ever since the famed Campbell case, which brought the downfall of the Macdonald government in the 1920s, the constitutional rule is that the Attorney General may seek advice from the Cabinet when making prosecutorial decisions but contacts initiated in the other direction give rise to serious questions of propriety.

In countries influenced by the French and German models, prosecutors are often a part of the judiciary, or under judicial supervision in an essentially inquisitorial system. In such systems, judges and government prosecutors are often located administratively in the Ministry of Justice, but they are buffered from direct political control. In Italy, for instance, public prosecutors are under the constitutional control of the independently elected Supreme Judicial Council, or *Consiglio Superiore delle Magistratura*. Likewise, in Japan, a hybrid system, the Office of the Public Prosecutor is part of the Ministry of Justice (*Homusho*), but the Prosecutor General (*Kenji Socho*) is appointed directly by the Cabinet, and the Minister of Justice (*Homu Daijin*) is prohibited by law from exercising control over specific cases.

Even in systems more clearly based on the Anglo-American model, certain prosecutions, especially those involving government officials, are institutionally protected from political control. In England, the Attorney General's prosecutorial decision-making is protected by custom and convention. In the USA, following the infamous Watergate scandal, Congress established the office of the Independent Counsel. Independent counsels were appointed by a special division of the court and buffered from direct political control of the President or the Attorney General (Harriger 1992). Controversy surrounding the Independent Counsel's investigation of the Whitewater scandal and the eventual impeachment of President Clinton, however, led to the office's disestablishment

in 2000. Nevertheless, Attorney Generals continue to have authority to appoint 'special prosecutors' inside the Department of Justice, who enjoy nearly complete independence from political direction by the elected executive.

2. *The Intersection of Politics and Law in Government Lawyering*

Protecting prosecutorial independence is one part of a larger concern regarding the role of politics in government lawyering. In this respect, two separate problems dominate the scholarly literature in this area. The first has to do with the separation of powers and the litigating role of government lawyers during periods of constitutional conflict. Administrations who use government litigation to challenge established judicial precedents, for example, are often charged by critics with violating the rule of law or other constitutional constraints (Caplan 1988). Moreover, in the USA it is not uncommon for government lawyers representing different branches to disagree in court about the meaning of a statute or a constitutional provision. In *Morrison vs. Olson* (1988), for instance, Justice Department attorneys argued against the constitutionality of the Independent Counsel Act, while counsel for Congress defended the statute.

Conflicts such as these put government attorneys in an awkward ethical position, and raise serious questions about the appropriate role of government lawyers in a system of separated powers. Moreover, in the USA such conflicts have become more common as prolonged periods of divided government have led partisans increasingly to turn to courts and the administrative bureaucracy, rather than the routine legislative process, for policy making. Scholars and practitioners have written a great deal about the principles that should guide government lawyers when two or more branches disagree about the law. However, such principles are usually just as controversial as the contested notions of separation of powers upon which they rest.

While government lawyers in parliamentary systems escape the political entanglements caused by a separation of powers, they cannot avoid the 'politics' of partisan abuse, corruption, and favoritism. There has been a great deal of scholarly debate about which institutional structures best safeguard government lawyers from 'inappropriate' political pressures. Anglo-American systems have experimented with structural reforms, such as the Independent Counsel's Office, as well as informal protections such as conventions that prevent political leaders from contacting government prosecutors in particular cases. All such efforts, however, suffer from the difficulty of distinguishing between proper and improper political influences.

In a well-known case in the USA in 1971, for example, President Nixon ordered the Justice Department to drop a pending antitrust suit against ITT Corporation in exchange for a $400,000 donation to Nixon's re-election campaign. This led critics to argue for the construction of a 'fire-wall' that would prevent White House officials from contacting the Justice Department in particular cases. However, contrast that White House intervention with President Carter's 1978 decision to reverse the Justice Department's position in *Regents of the University of California vs. Bakke* (1978), the first major affirmative action case to reach the Supreme Court. Or consider whether President Clinton should have had no role in deciding whether the Justice Department should file antitrust charges against Microsoft Corporation in 1999. Some cases, whether civil or criminal, present legal issues or involve individuals in a way that the rule of law properly requires direction from democratically accountable officials. Others do not. However, there is no entirely satisfactory way to distinguish *a priori* between proper and improper political influences. Moreover, removing prosecutors from political control only shifts, but does not remove, the influence of partisanship. Critics of the Independent Counsel's office in the USA, for example, long complained that most investigations, such as those surrounding Whitewater or the Iran-Contra scandals, were politically motivated (Eastland 1989).

By contrast, countries relying on the French and German models tend to protect prosecutorial independence by trying to remove it entirely from 'political' influence and lodging the power under the direction of the judiciary. Such systems thus trade off democratic accountability over criminal prosecution for judicial control. However, they too only shift, rather than remove, the sources of political influence and must also confront the difficulty of distinguishing between proper and improper influences. Even judicial officials are accountable to broader regime structures and they too can be subject to corruption and partisan abuse. Determined political leaders have had no more trouble influencing prosecutors administratively lodged in the judiciary as those in the executive. In the well-known Credit Lyonnais case in France, for instance, President François Mitterand was able to use party influence to control decisions in the Ministry of Justice over the prosecution of corruption charges involving government ministers, and the French and Italian Socialist parties, and leading businessmen with alleged Mafia ties. Likewise, in Italy, where the independence of prosecutors is fixed constitutionally, a well-known effort to prosecute political corruption charges against more than 400 political figures, including several ministers and members of Parliament in the mid-1990s foundered when Parliament approved a law fixing a deadline for the disposal of cases. In a system notorious for delay and inefficiency, the law effectively gave immunity to many of those who had been indicted.

The intersection of law and politics in government lawyering, and the administrative structures governing it, will likely continue to dominate scholarly attention in the future. As noted above, existing studies of government lawyering tend to focus on particular offices or administrations. Little work of a comparative nature, either within, or across, different political systems has been undertaken, and this remains perhaps the most fruitful direction of future research.

See also: Counsel, In-house; Governments; Lawyers; Legal Professionalism; Political Lawyering; Professions in Organizations; Public Interest Law

Bibliography

Baker N 1992 *Conflicting Loyalties: Law and Politics in the Attorney General's Office*. University Press of Kansas, Lawrence, KA

Caplan L 1988 *The Tenth Justice: The Solicitor General and the Rule of Law*. Vintage, New York

Clayton C W 1992 *The Politics of Justice: The Attorney General and the Making of Legal Policy*. Sharpe, New York

Clayton C W 1994 Law, politics and the new federalism: State attorneys general as national policymakers. *The Review of Politics* **56**: 525

Clayton C W 1995 *Government Lawyers: The Federal Legal Bureaucracy and Presidential Politics*. University Press of Kansas, Lawrence, KA

Eastland T 1989 *Ethics, Politics and the Independent Counsel: Executive Power, Executive Vice*. National Legal Center, Washington, DC

Edwards J L J 1964 *Law Officers of the Crown*. Sweet & Maxwell, London

Edwards, J L J 1984 *The Attorney General, Politics, and the Public Interest*. Sweet & Maxwell, London

Eisenstein J 1978 *Counsel for the Untied States*. Johns Hopkins University Press, Baltimore

Fionda J 1995 *Public Prosecutors and Discretion: A Comparative Study*. Oxford University Press, Oxford, UK

Harriger N 1992 *Independent Justice: The Federal Special Prosecutor in American Politics*. University Press of Kansas, Lawrence, KA

Horowitz D 1977 *The Jurocracy*. Heath, Lexington, MA

Perrot R 1995 *Institutions Judiciares*. Montchestrien, Paris

Sakokar R M 1992 *The Solicitor General: The Politics of Law*. Temple University, Philadelphia

C. W. Clayton

Government Statistics

Government statistics are vital to advance the economic well-being and quality of life of a nation's citizens. Without such statistics, policy making would be less effective, economic markets would be less informed, and the countless decisions of private businesses and individuals would allocate economic and other resources less efficiently. But more, government statistics are vital to a democracy, as the information they convey provides the means by which citizens can hold their government accountable.

This article provides a general review of government statistics largely in the context of the USA, but it also addresses broader issues relevant to other countries. The focus is on the national government and, therefore, mostly on national statistics, but much is also relevant to governments at subnational levels, which also produce statistics.

The article begins with the purposes of government statistics and the rationale for why governments produce them. Next it takes up the principles by which the statistics should be produced to serve these purposes effectively; this is followed by descriptions of various types of data collection, and of how the responsibilities for collection, analysis, and reporting of statistics are organized among statistical agencies. Effective practices of these agencies are identified. The entry concludes with a discussion of some challenges faced by statistical agencies.

Statistics from government sources are referred to in many entries in this Encyclopedia (see *Censuses: History and Methods*; *Censuses: Comparative International Aspects*; *Economic Panel Data*; *Sample Surveys: Cognitive Aspects of Survey Design*; *Large-scale Social Survey, Analysis of*).

1. What are Government Statistics?

'Government statistics' means factual information that is reported upon for statistical purposes by a unit of government. Some countries also use the term 'official statistics.' In this articles, 'statistics' and 'data' are used interchangeably to mean factual information, and the terms are not restricted to quantitative information. Usually, the unit or agency of government that issues the statistics has collected or compiled the data to produce them. In most cases, the agency is a distinct entity of government entrusted with reporting the statistics. In some cases, the agency may analyze its data to provide more information about the statistics in its reports, especially information that makes the statistics more relevant to public policy concerns. For the agency to be effective, the focus of any analysis must be on ways to improve the relevance and accuracy for use in public policy analysis and decision making, independently of any particular policy.

The *sine qua non* of the definition of government statistics is that they are reported upon only for *statistical purposes*:

> 'Statistical purposes include description, evaluation, analysis, inference, and research. For these purposes, a statistical agency may collect data from individuals, establishments, or

other organizations directly, or it may obtain data from administrative records, but it does not do so for administrative, regulatory, or enforcement purposes. Statistical purposes exclude any interest in an individual person or economic unit. The data are used solely to describe and analyze statistical patterns, trends, and relationships involving groups of persons or other units' (Martin et al. 2000)

2. *Why do Governments Produce National Statistics?*

A government provides statistics to inform its citizens, satisfying an essential requirement of a democracy (see, e.g., United Nations Economic and Social Council 1994). The government, however, is most often the primary user of the statistics it produces. The statistics are used in the formulation of, and decision making on, public policies, as well as in their implementation and evaluation. And the statistics are used for much private-sector and individual decision making. Information from statistics is, for example, vital to inform economic markets.

In most cases, a government collects and reports the statistics itself, either directly or through the services of a private organization. In some cases, the private sector satisfies the demand for national statistics. Examples are the Conference Board's compilation of US leading economic indicators, abortion statistics for many countries compiled by the Alan Guttmacher Institute, and many private market-research surveys and reports of economic forecasting firms. But even when the private sector can do so, there are reasons for a government to collect the statistics. Alternative sources may not exist. Timeliness may be a critical factor. Statistics collected by government agencies are often deemed to be of greater quality. And the government may collect statistics because they contain confidential information.

A more compelling reason may be trust and credibility. Potential subjects of government data collection may be more willing to volunteer their time and effort, and divulge confidential information, for a government survey rather than for a private one. And the public and other important users may find data collected by the government to be more credible.

A fundamental reason, however, for a government to collect data has to do with the very nature of data. If a commodity can be consumed by many and consumption by one does not reduce the consumption of others, then the commodity is a 'public good.' Data are manifestly public goods: use of data by one party in no way diminishes their usefulness to other parties. Individuals benefit from a public good but, unless each individual is charged for the consumption of the good, they are unlikely to pay for them voluntarily. Consequently, the goods may not be produced, unless the government intervenes. In the case of data, the government may intervene directly by collecting the data, or it may intervene indirectly, for example, by defining rights for a license or copyright to provide the means for the private sector to charge for the value added in producing secondary data. For these reasons, a government typically provides primary data and the information infrastructure. The private sector typically provides secondary data and may build on the infrastructure.

3. *Principles of Government Statistics*

The US National Academy of Sciences–National Research Council report (Martin et al. 2000) defines a 'statistical agency' as a unit of government whose principal function is the compilation and analysis of data, and the dissemination of information for statistical purposes. The unit is recognized generally as a distinct entity, which may be located in a cabinet department or independent agency. The mission of a distinct entity or agency of government that issues government statistics may or may not be solely statistical. The USA has a highly decentralized statistical system, with over 70 agencies having direct funding of over $500,000 per year for statistical activities. Only some of these agencies are statistical agencies.

Statistical agencies are established to be a credible source of accurate, timely, and relevant statistics in one or more subject areas. Accuracy requires proper concern for consistency in spatial and temporal dimensions, as well as statistical measures of error. Timeliness requires concern for the frequency of issuing data, as well as disseminating data to users when they need them. Relevance requires concern for anticipating future data needs, as well as meeting current ones. And credibility requires concern for both the reality and appearance of impartiality, as well as the reality and appearance of independence from political control. It is essential to the mission of statistical agencies to work to ensure the goals of accuracy, timeliness, relevance, and credibility of statistical information.

Martin et al. (2000) identify three principles for a statistical agency. First is that the agency is in a position to provide information relevant to issues of public policy. That requires the agency to be knowledgeable about public policy issues and be able to provide objective information that is relevant to the needs of formulating, implementing, and evaluating public policies. Second is that the agency has a relationship of mutual respect and trust with those who use its data and information. To develop that relationship, an agency must strive to maintain credibility for itself and for its data. In particular, the agency must be perceived to be free of political interference and policy advocacy. Other aspects that contribute to credibility are a commitment to quality and professional standards, and openness about all aspects of the data, especially the quality of the data.

The third principle is that the agency have a relationship of mutual respect and trust with respondents who provide data, and with all data subjects whose information it obtains. Statistical agencies rely in large part on information supplied by individuals and organizations outside the federal government, such as state and local governments, businesses, and other organizations. Some of this information is required by law, such as employers' wage reports, and some of it is compiled from records of government programs. But much of it is obtained through the voluntary cooperation of respondents to surveys. Even when the response to a survey is required by law, cooperation by respondents promotes accuracy (Edmonston and Schultze 1995), which, in turn, promotes the credibility of the agency's data. In encouraging cooperation, it is important to statistical agencies that respondents perceive that the information they are being asked for is important; that it is being collected in an impartial, professional manner; and that their personal information will remain confidential.

Confidentiality of individual responses means that the information will be used only for statistical purposes and that, in particular, it will not be used for any administrative, regulatory, or enforcement purpose regarding the individual (see *Confidentiality and Statistical Disclosure Limitation*). Statistical agencies go to great lengths to maintain the confidentiality of individual responses (see Duncan et al. 1993, Mackie and Bradburn 2000). Some, such as the US Bureau of the Census, have legislation to protect confidentiality, and some, such as the US Bureau of Labor Statistics, have a history of successful reliance on executive order, custom, and case law. A Federal Statistical Confidentiality Order (US Office of Management and Budget 1997) directs 12 statistical agencies not to disclose or use data acquired for exclusively statistical purposes in identifiable form for any other purpose unless otherwise compelled by law. In Europe, the 1995 European Community Directive on Data Protection governs the transfer of individually identified or identifiable data.

The respect and trust of data providers are also essential for statistical agencies to achieve credibility for their data. Protecting the confidentiality of individual responses engenders trust, but, as was noted, agencies may be limited in their abilities to do so. In that case, agencies need to provide respondents with realistic promises of confidentiality that the agency can reasonably be expected to honor.

Respect for data providers also means valuing their time and effort to respond—in particular, designing data collection so that it is not unduly burdensome—and valuing their responses so that they are put to important uses. It is important too that respondents perceive that the agency collecting their information is trustworthy and credible; in particular, that the information will not be used for a particular partisan interest. A statistical agency must be perceived, by its data providers as well as its data users, to be working in the national interest.

4. *How are Government Statistics Collected and Reported?*

Statistical agencies collect data through censuses, surveys, and compilations from administrative records. Many countries conduct a census of their population. In the UK and USA, one is conducted every 10 years; in Canada, every five years. Other examples of US censuses are the quinquennial census of agriculture, which covers agricultural production on farms that are defined as places from which $1,000 or more of agricultural products were produced and sold in the census year; the quinquennial economic census, which covers the economic activity of nonfarm establishments; and the quinquennial census of governments, which covers employment and finances of state and local governments.

Examples of surveys in the USA include the Current Population Survey, from which monthly employment and unemployment data are derived; the Consumer Expenditure Survey, from which data on expenditure patterns are used periodically to adjust the weights of goods and services in producing the Consumer Price Index; and the National Health Interview Survey, which provides annual estimates of health conditions and use of healthcare services. Data in government surveys may be collected by field staff of an agency (e.g., the Census Bureau conducts the Survey of Income and Program Participation with its field staff); by state or local governments under a cooperative program (the Current Employment Statistics program of the Bureau of Labor Statistics is a monthly survey, based on payroll records of business establishments, conducted by state employment security agencies); by a private organization under a contract or grant from the agency (the University of Michigan conducts the Health and Retirement Survey under a grant from the National Institute on Aging); or by another federal agency (the Census Bureau conducts the Current Population Survey and the Consumer Expenditure Survey for the Bureau of Labor Statistics and the National Health Interview Survey for the National Center for Health Statistics). Many of these surveys have counterparts in other countries: for example, the Labour Force Survey of the Office of National Statistics in the UK and the Survey on Employment of the National Institute for Statistics and Economic Studies in France.

The units of observation in surveys may be individuals, households, families, schools, establishments, or other economic units. Surveys may be cross-sectional to provide an estimate at a point or interval of time, or they may be longitudinal in design in that they follow the same units of observation over time.

Longitudinal surveys can provide information on changes in a unit over time: for example, transitions of families in an out-of-poverty status from the Survey of Income and Program Participation (see *Longitudinal Data*; *Economic Panel Data*). Some surveys, however, may have a longitudinal component in their design to reduce error or increase efficiency, such as the American Housing Survey.

Statistical agencies also compile information from administrative records. For example, vital statistics are compiled by the National Center for Health Statistics from state agencies, and administrative records are the source of basic statistics for very small firms in the quinquennial economic census, thus reducing or eliminating the need to send these firms census questionnaires.

Statistical agencies may be dependent on the censuses and surveys of other agencies as well as their own data collection in order to produce national indicators. The Bureau of Economic Analysis relies heavily on economic statistics from other sources in compiling the national income and product accounts. And most agencies rely on the Census Bureau to produce the population denominators for reported rates.

Since the US statistical system is so decentralized, important information on a population of interest, such as children or the elderly, is often collected by many agencies and in different ways that may make relating the data difficult, if not impossible. Sometimes, agencies collaborate to l important information gaps. Two examples are key indicators of well-being for older Americans (Federal Interagency Forum on Aging-related Statistics 2000) and for children (Federal Interagency Forum on Child and Family Statistics 2000).

Whatever the source of data, one of the most important activities of a statistical agency is measuring, controlling, and reporting upon error. Statistical agencies have been pre-eminent in the development of survey designs to reduce sampling error. And many require, as a standard for some forms of publication, that statistical errors be reported. The agencies too are concerned about measurement and other forms of nonsampling errors, and employ various means to study them, including reinterview studies, record checks, multiple indicators, and cognitive studies, which focus on how errors arise in the process of answering questions. Three statistical agencies, the Census Bureau, the Bureau of Labor Statistics, and the National Center for Health Statistics, have cognitive laboratories to evaluate questionnaires and conduct other research to better understand and reduce error (see *Sample Surveys: Cognitive Aspects of Survey Design*).

5. *How are Government Statistics Organized?*

Government statistical activities may be centralized into a single agency or not. The organization in the USA, which is described first, is highly decentralized with activities dispersed among many agencies. Some of the trade-offs between centralized and decentralized systems are described at the end of this section.

5.1 *The US Statistical System*

The highly decentralized statistical system in the USA arose historically because, as new subject matter areas emerged as a national priority, new statistical agencies were created to meet the needs for new data. Little attention was paid to the relevant data-collection activities of other agencies. Efforts at collaboration have met with some successes, such as the interagency forums on aging-related statistics, and on child and family statistics, but efforts to assure systemic collaboration have failed. Attempts to achieve collaboration through centralization, or at least a major consolidation, have run into stakeholders with a special interest in retaining strong, relatively independent agencies within a particular department of government (see Norwood 1995).

Responsibility for coordinating the myriad statistical activities of the US government rests with a small unit in the Office of Management and Budget, the Statistical Policy Branch in the Office of Information and Regulatory Affairs. Authorization of that responsibility derives from legislation principally concerned with paperwork reduction. Despite its size, however, the Statistical Policy Branch has brought statistical agencies together to address many matters of common concern.

The 1995 reauthorization of the Paperwork Reduction Act created an Interagency Council on Statistical Policy, formalizing an arrangement whereby statistical agency heads participate with the Statistical Policy Branch to coordinate their activities. Ten agencies are recognized as having statistical activities as their principal mission: the Bureau of Economic Analysis and the Bureau of the Census in the Department of Commerce; the Bureau of Justice Statistics in the Department of Justice; the Bureau of Labor Statistics in the Department of Labor; the Bureau of Transportation Statistics in the Department of Transportation; the Economic Research Service and the National Agricultural Statistics Service in the Department of Agriculture; the Energy Information Administration in the Department of Energy; the National Center for Education Statistics in the Department of Education; and the National Center for Health Statistics in the Department of Health and Human Services. The heads of these principal statistical agencies, together with the head of the Office of Management and Budget Statistical Policy Branch and the heads of statistical units in the Environmental Protection Agency, the National Science Foundation, the Internal Revenue Service, and the Social Security Administration, are the members of the Interagency Council.

Through this council, its Federal Committee on Statistical Methodology, and other interagency committees it convenes, and its encouragement of the Committee on National Statistics of the National Academy of Sciences–National Research Council, the Statistical Policy Branch has addressed many methodological and other issues that cut across the interests of most statistical agencies. These include issues of poverty measurement, and the definitions of race and ethnicity and of metropolitan areas. It is also concerned with international comparability of statistics, and it represents the USA at the United Nations Statistical Commission and other international organizations. Moreover, the Statistical Policy Branch has striven to extend by regulation and law the protection of confidential data in ways that would allow statistical agencies to share needed information under that protection.

The Statistical Policy Branch is the only unit of government charged to be concerned with the government statistical system as a whole. It conducts an annual review of all major statistical programs for budget directors in the Office of Management and Budget, and publishes a special analysis of funds proposed for statistical activities in the President's budget (US Office of Management and Budget 2000).

5.2 Centralized vs. Decentralized Systems

Most large countries of the world have centralized most or all of their statistical activities into a single agency with broad authority for collecting and compiling government statistics. Coordination of the different activities occurs within the agency itself. Examples are Statistics Canada, the statistical agency of Canada, and the Office of National Statistics in the UK.

A central statistical agency is usually seen to be more efficient in resource allocation; in reaping the benefits of scale; in sharing information across different subject matter areas in order to enhance data and to avoid duplicating activities; in coordinating data activities, such as developing a single longitudinal survey of youth that serves the purposes of both health and education policy makers; and in arranging for the effective use of administrative data. Advantages of a decentralized system are that priorities for budget, relevance, and timeliness, and the urgency to meet needs for new data are set within the department with the most concern for the information and with policies that are most affected by it, and that different perspectives and ways of conducting statistical activities can bring strength to the system as a whole.

6. Practices of Effective Statistical Agencies

The operation of an effective statistical agency begins, first, with a clearly defined and well-accepted mission and, second, with a strong standing of independence. With these prerequisites, a number of practices contribute to the agency's effectiveness: continual improvement of and development of more useful data, cooperation with data users, openness about its data, commitment to quality and professional standards, wide dissemination of data, advancing the effective use of statistics, analysis and interpretation, an active research program, a highly professional staff, and coordination and cooperation with other statistical agencies.

6.1 Mission

An agency's mission should include responsibility for assessing and meeting needs for information in its purview, determining the sources of the data and the measurement and collection methods, and ensuring that its data are available to the public. An agency mission statement often goes beyond its charge in legislation. In particular, it should articulate the principles that will govern all of its work. The mission should be well accepted externally by the agency's department and others who oversee the agency, and especially by data users. It should also be well accepted internally, so that all agency employees understand how their work furthers the mission.

6.2 Independence

A strong and widely acknowledged standing of independence is necessary for a statistical agency to have the credibility to carry out its mission effectively. In particular, the agency must be independent of other parts of its department engaged in policy making, enforcement, regulatory, or administrative activities. It must be impartial and avoid even the appearance that its data might be manipulated for political purposes, or that information on an individual might be used for enforcement, regulatory, or administrative purposes.

Forms of independence differ among statistical agencies. Agency heads may be required by law to be professionally qualified in the collection and analysis of the relevant statistics. They may be appointed by the highest levels of government for a specific term not coincident with the administration. And they may have direct access to the head of the cabinet department in which the agency is located.

Other characteristics of independence include authority for the following: to decide the scope, content, frequency, and manner of data collected, compiled, or reported; to select and promote professional staff; to release statistical information without prior clearance and with a clear separation from policy interpretations by others; to represent the agency's statistical program before legislatures and the public; and to establish and follow predetermined schedules for the release of important economic and other indicators.

6.3 Continually Developing More Useful Data

In order to make their data more useful and to meet changing public policy needs, statistical agencies must continually look to develop data systems that can serve broader purposes. Integrating data from surveys conducted by different agencies or integrating data from surveys with data from administrative records, such as tax or social welfare program records, are ways to achieve this goal. When separate data sets are collected and analyzed in such a way that they may be used together, the value of each may be greatly enhanced. Data sharing and statistical uses of administrative records, with appropriate safeguards for confidentiality, also make a statistical agency more efficient as well as more effective.

6.4 Cooperation with Data Users

An effective statistical agency consults with a broad range of users of its data in order to make its products more useful. It seeks advice on data concepts, methods, and products in a variety of formal and informal ways, from data users as well as subject-matter and other experts. It endeavors to meet the needs for access to data while safeguarding the confidentiality of individual responses. And it strives to provide equal access to all potential users.

6.5 Fair Treatment of Data Providers

To maintain a relationship of mutual respect and trust with data subjects and other data providers, an effective statistical agency observes fair information practices, which include: protecting the confidentiality of data identified with an individual, whether the data are collected directly or obtained from administrative record sources; informing respondents of the purposes of data collection and the expected uses of the information; informing those asked to report whether their participation is voluntary or mandatory, and using appropriate informed consent procedures for voluntary data collections; minimizing the contribution of time and effort asked of respondents, which may be achieved in seeking the advice of respondents in the planning and design of data collection; and respecting the willingness of respondents to contribute to society by representing accurately the information they provide, and by making it available for important uses.

6.6 Openness

An effective statistical agency operates in a completely open manner. Openness requires providing full documentation of the data with acknowledgment of any uncertainty, a description of the methods used and assumptions made, and measures or other descriptions of error, for both sampling and nonsampling error. When a mistake occurs, the agency should acknowledge it promptly and publicly. For a correction to an important indicator, the agency should strive to inform as many current and potential users as possible.

6.7 Quality and Professional Standards

An effective statistical agency is committed to the highest standards of quality work and professional practice and ethics, in statistics and other relevant professions. For this reason, a strong and continuous relationship with appropriate professional organizations in statistics and relevant subject-matter areas is important.

6.8 Dissemination of Data

An effective statistical agency seeks the widest possible dissemination of its data through a variety of means, including public-use data files and the agency's website, and has policies to guide what data to archive and how.

6.9 Advancing Effective Use of Statistics

Statistical agencies should also advance the uses of their data through making them easily accessible to users in convenient forms, involving data users in planning data collection and dissemination, and providing full documentation and other forms of technical support to users.

6.10 Analysis and Interpretation

Agencies can make their data more relevant to public policy concerns by analyzing and interpreting the data. Care must be exercised, however, that analyses not are perceived as policy advocacy.

6.11 Research

An active research program is integral to the activities of an effective statistical agency. Research may be conducted on the substantive issues for which the data were compiled, especially concepts and definitions; methods of measurement, collection, analysis, and dissemination; and topics to improve the utility of the agency's data in policy analysis and decision making.

6.12 Professional Staff

A statistical agency must recruit and retain a professional staff of high caliber in statistics and fields relevant to its mission. Opportunities should be provided for staff to maintain and advance their capabilities, such as through appropriate activities of professional organizations.

6.13 Coordination and Cooperation

In the US context, an effective statistical agency continually seeks opportunities to cooperate with other statistical agencies to enhance the value of its own and other agencies' information. Statistical agencies cooperate, not only with each other, but also with state and local statistical agencies in the provision of data for subnational areas. They cooperate, too, with foreign and international statistical agencies to exchange information, on both data and methods, and to develop common classifications and procedures to promote international comparability of information.

Integrating data collected or compiled by different statistical agencies invariably requires much coordination and cooperation among the agencies, each of which have different missions. But the rewards are data more relevant to policy concerns, and a stronger statistical system as a whole. For these reasons, statistical agencies must cooperate as partners with one another, not only on the development of data, but also on the entire panoply of statistical activities, including definitions, concepts, measurement methods, analytical tools, professional standards, dissemination modes, ways to advance the effective use of their data, and means to protect the confidentiality of responses.

7. Challenges Faced by Statistical Agencies

We live in a rapidly changing, increasingly interdependent world with a global economy. Many countries face profound demographic changes. Primarily fueled by technological advances, widespread changes are occurring in society, the environment, and international relationships. Governments struggle to keep up with the changes. They are concerned with the need for new property rights, laws governing new markets, new forms of businesses, and investments in new technology for education. They are concerned too about mitigating human error in transportation, communications, healthcare, and other complex, critical systems; avoiding harmful effects on the environment and sources of food; and guarding against the loss of individual privacy. And governments are also concerned with disparities in economic growth and the distribution of income and wealth, to which disparities in access to technology may contribute, since these disparities could increase tensions among social groups and nations.

An efficient, flexible, and modern information system is essential for a government to address these and other concerns. The greatest challenge facing any statistical agency is for it to keep up with the times. 'Public policy analysis often cannot wait until new surveys are undertaken, but must rely on knowledge drawn from the existing statistical database' (Cook 2000). Those in charge of government statistics must anticipate and prepare for future needs.

Technology, which has induced so many rapid changes, itself can give statistical agencies the means to respond to those changes. The agencies must not only incorporate new technology into collecting, processing, analyzing, and disseminating government statistics, but they must also prepare for the plethora of new data in digital form.

Advances in information technology are also making it more and more difficult to protect confidentiality. If a statistical agency is to respect its data providers by putting their information to use in the best way to benefit society, stronger ways to protect confidentiality must be sought. Statistical agencies should lead in confidentiality research, including research on data encryption and altering or reconstituting data, and in fostering legal and administrative means to protect confidentiality.

8. Conclusion

Statistics are essential for a government to function effectively and to advance businesses and the well-being of its citizens. The nature of data as a public good provides an economic argument for a government to collect statistics, but there are many other reasons for it to do so. One is to assure accurate, timely, and credible information that its citizens may use to hold their government accountable on any issue. A statistical agency must be able to provide information relevant to current and future policy issues, and to establish credibility for itself and its data. Finally, the agency must be respected and trusted by respondents who provide it with data and with all data subjects whose information it obtains. That respect and trust can be fostered by protecting the confidentiality of individually identifiable information and by putting the information it obtains to use in the best ways to benefit society, including research.

More information on the principles and practices of statistical agencies can be found in Martin et al. (2000), and on the organization of US statistical agencies in Norwood (1995). Statistics for over 200 countries and areas of the world are compiled by the United Nations Statistics Division (2000), which provides links for the national statistics of countries online at www.un.org/Depts/unsd/ through a link to 'Global Statistics.' Links to the websites of statistical agencies in the USA are available at www.fedstats.gov.

See also: Databases, Core: Demography and Registers

Bibliography

Cook L 2000 Statement by the first National Statistician on the launch of National Statistics on 7 June, www.statistics.gov.uk/nsbase/about_ns/About.asp

Duncan G T, Jabine T B, De Wolf V A (eds.) 1993 *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academy Press, Washington, DC
Edmonston B, Schultze C (eds.) 1995 *Modernizing the U.S. Census*. National Academy Press, Washington, DC
Federal Interagency Forum on Aging-related Statistics 2000 *Older Americans 2000: Key Indicators of Well-being, 2000*. US Government Printing Office, Washington, DC, http://www.agingstats.gov
Federal Interagency Forum on Child and Family Statistics 2000 *America's Children: Key National Indicators of Well-being, 2000*. US Government Printing Office, Washington, DC
Mackie C, Bradburn N (eds.) 2000 *Improving Access to and Confidentiality of Research Data: Report of a Workshop*. National Academy Press, Washington, DC
Martin M E, Straf M L, Citro C F 2001 *Principles and Practices for a Federal Statistical Agency: Second Edition*. National Academy Press, Washington, DC
Norwood J L 1995 *Organizing to Count: Change in the Federal Statistical System*. The Urban Institute Press, Washington, DC
United Nations Economic and Social Council 1994 Fundamental principles of official statistics of the Statistical Commission of the United Nations. Report of the Statistical Commission on its Special Session, New York, 11–15 April 1994. *Official Records of the Economic and Social Council*, suppl. 9
United Nations Statistics Division 2000 *Statistical Yearbook Forty-third Issue*. United Nations, New York
US Office of Management and Budget 1997 Notice of order providing for the confidentiality of statistical information and extending the coverage of energy statistical programs under the federal statistical confidentiality order. *Federal Register* **62**: 35044–50
US Office of Management and Budget 2000 *Statistical Programs of the United States Government, Fiscal Year 2001*. US Government Printing Office, Washington, DC

M. Straf

Governments

Governments are universal. They exist in every social organization: there is always a body, normally relatively small, which has the task of running that organization; but the most visible governments are those of nations and of other public institutions. With the decolonization process, independent national governments have come to rule practically the whole of the planet and they have more than doubled in number, from well under a hundred in the early 1950s to nearly two hundred at the beginning of the third millennium.

1. *Problems of Definition*

Being at the center of political life, national governments tend to be rather small and, often at least, compact. Yet it is not possible to circumscribe them neatly as their boundaries are not precise. In many cases, indeed most, an official list exists, but this list is typically not exhaustive, as key personalities, who may formally only be 'advisers,' but are in fact highly influential, tend not to be included. There are also often two or more 'levels' of the government: thus in the UK and in countries which have followed closely the British model, for instance, the 'cabinet' constitutes only an inner core, with 'ministers of state' and 'parliamentary secretaries' being below this core. In the USA, the government is typically referred to as the 'administration' or as the 'executive': however, both of these expressions remain imprecise with respect to those whom they cover.

It is therefore more appropriate to look for a definition in terms of the functions which national governments fulfill. In general, these are said to 'run the affairs of the nation,' although they only do this up to a point since they are helped and advised by a multiplicity of individuals and of groups, and, in modern democratic countries in particular, by parties and legislatures, while everywhere the bureaucracy is at their 'disposal.' One can in reality distinguish among three functions of governments. First, they have to elaborate realistic policies which can be implemented: this is the function of conception. Second, they have a function of implementation, as they must find the means of ensuring that the policies which have been elaborated are indeed turned into practice. There can be tensions between these two functions. The skills required to undertake each of them successfully are different: to conceive means to 'imagine'; to implement means to 'manage.' Yet governments have also a third function, that of coordination: they must ensure that the policies which are elaborated and implemented do not pull in different directions but are complementary.

2. *Types of Government, Liberal-Democratic and Authoritarian*

A 'revolution' in the characteristics of governments has occurred since the eighteenth century, when 'modern' government began. This 'revolution' has had two main aspects. First, there has been a massive increase in the complexity of governmental activities, as states have come to be increasingly involved in the economic and social life of countries, especially from the late nineteenth century to the 1980s. Second, many governments have come to be (or have had to be) responsive to the demands of citizens, in particular where the representative system had been introduced: for the first time in human history, these governments had to abide by rules based on liberal democratic principles and enshrined in constitutions. Admittedly, Greek cities, the Roman Republic, and, during the Renaissance, some parts of Italy, had experienced forms of 'limited' government; but only from the end of the seventeenth century did representative govern-

ment take root in a nation, England, which hitherto had been a traditional monarchy, while, almost a century later, American independence marked the emergence of the first liberal Republic.

England and America were indeed the originators of the two main forms of constitutional government which have existed to this day, the cabinet system, which was gradually adopted in almost every Western European country and in the Commonwealth, and the presidential system, which spread to Latin America, though with only limited success. Thus, at the beginning of the third millennium, approximately half the national governments could be described as being broadly responsive to their people. Meanwhile, in the rest of the world, including in many countries of Africa, of Asia and, at least in the past, especially up to the mid-1980s, of Latin America, governments have tended to be authoritarian. The Communist form was particularly harsh, both in what was the Soviet Union and in Eastern Europe up to the late 1980s, as well as in China and a number of other Asian countries: it was based on the dominance of a single party whose leaders formed the real government and controlled the formal executive. Independence from colonial rule which started in the 1950s and gathered momentum in the 1960s, especially in Africa, also led to many dictatorships, sometimes based on single-party arrangements but relying even more on the 'charismatic' figure of a president, often drawn from the military and having come to power through a coup. Admittedly, since the 1980s, a move towards responsive governments occurred in many countries hitherto ruled by dictatorships, in Latin America, in what had been the Communist block, in East and Southeast Asia and in parts of Africa; but authoritarian rule continued to characterize many countries of the contemporary world, sometimes on the basis of old-fashioned monarchical governments but more often on 'modern' types of dictatorships, in many cases of the military type.

3. *The Organization of Governments: Cabinet and Presidential Systems. The Role of Leaders*

Partly as a result of these differences, governments vary widely in organization, duration, composition, and powers. The distinction between the cabinet system and the presidential system remains the main one among liberal democratic governments, although there are variations within each of these two models and some cases are intermediate. The cabinet system is based on the division between a symbolic head of state, often a monarch, and a government, responsible to parliament, headed by a prime minister: the prime minister has to operate in the context of a collegial system in which those members of the executive who are part of the cabinet (normally referred to as 'ministers') are entitled to participate fully and collectively in decision-making: in practice, this is far from being always the case and there are substantial differences from country to country and within each country. Some cabinets are controlled closely by their prime minister, while others are more egalitarian. By and large, the British government and the governments of many Commonwealth countries tend to be tilted toward prime ministerial dominance, while a number of continental European governments, at least those which are based on complex party coalitions, give more elbow room to ministers. Yet, even though cabinet members may be more or less collectively involved, there is a real sense in which the government constitutes a team. A variation on this model is provided by the French system, in which the picture is somewhat complicated by the fact that the president, elected directly by the people, plays a significant part in policy making, at least when the party of the president also dominates parliament: this model has been adopted by some of the post-Communist governments of Eastern Europe as well as in the developing world.

The presidential system, on the other hand, is based on the separation of powers principle, with both president and legislature elected directly by the people. The president, as head of the executive, has thus greater leeway to organize the government in the way he/she wishes: this is particularly true in the USA. In some Latin American countries, on the other hand, the president has to an extent to take into account the desire of party leaders to play a part in the appointment of members of the executive. However, by and large, presidential governments are both constituted more independently from the parties and in particular from the parties in the legislature. They are also more hierarchical in character: the members of the government (typically known as 'secretaries') are responsible to the president for the department of which they are in charge (although in some cases the legislature can also censure them individually). They are not collectively involved in decision-making.

Governments are fashioned by their leaders. The twentieth century was characterized by the emergence of 'charismatic' rulers, including and perhaps above all in Communist countries, despite the fact that the Communist ideology is based on the notion that the economic 'substructure' is held to be paramount. 'Charismatic' or personalized leadership can also be found in liberal regimes, whether of the cabinet or of the presidential type. This has been the case not just of wartime leaders, such as Churchill or de Gaulle, but of rulers who attracted the imagination of whole populations, such as Kennedy.

4. *The Duration of Governments*

The duration of governments varies. The longest are those of the few remaining traditional regimes, in effect absolute monarchies, and those of some of the

more 'modern' dictatorships: Communist governments, their leaders in particular, often remained in power for over a decade; on the other hand, most other leaders are only in office for a few years. The other government members often stay for even shorter periods: the average duration of ministers or secretaries is in the region of three to four years, with a substantial proportion of them being truly transient and remaining in government for a year or less.

5. *Career Patterns*

The composition of governments also varies. Some of these, primarily where there is a cabinet system, include almost exclusively professional politicians, men and women who have been active in a political party, have then been elected to parliament and have joined the ranks of the executive after a few years on the 'backbenches.' Other government members are 'technicians,' more often drawn from the public sector than from the private sector: their career in politics tends to last a few years only and they return to other public or private jobs after their stint in government. Moves of this kind occur less frequently in cabinet systems than in presidential systems, where, as we noted, the head of the state has more freedom to pick and choose whom he or she wants to appoint; they also occur in many authoritarian governments, though, in these, military men are often side by side with civilian 'managers.'

Variations in career patterns are associated with differences in social background. In general, except where social democratic parties are frequently in power, as well as in the past Communist systems, men and women with a working class background are rare: lawyers, teachers, and managers form the bulk of the civilian members of governments. Only in the last decades of the twentieth century and in some countries, mainly Northern European, have women had more than a token share of governmental posts, especially in such key departments as those of foreign, economic, or internal affairs.

6. *The Impact of Governments: An Unresolved Question*

There is debate about the extent to which leaders and governments affect the countries which they rule. As their careers are often short, it is suggested that members of the higher civil service, together with the lobbies which gravitate around them, play a key part in decision making. As a matter of fact, it has not been possible so far to determine precisely the impact of leaders and members of governments even when they stayed in power many years: environmental factors, both internal and external to the country concerned, must be taken into account. Yet some leaders at least, whether 'harsh,' such as Lenin or Stalin, or 'good,' such as Roosevelt or Churchill, had a major influence on the destinies of the countries which they ruled and indeed of the world. What can only be said is that there are massive variations in impact, as there are major variations in the characteristics of governments and in particular in their responsiveness, organization, duration, composition, and complexity.

See also: Authoritarianism; Authority: Delegation; Leadership: Political; Liberalism; Liberalism: Historical Aspects; Parliamentary Government; Parliaments, History of; Power: Political; Presidential Government

Bibliography

Barber J D 1977 *The Presidential Character*, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ

Bass E M 1981 *Stogdill's Handbook on Leadership*. Free Press, New York

Blondel J 1982 *The Organisation of Governments*. Sage, London

Blondel J 1987 *Political Leadership*. Sage, London

Blondel J, Muller-Rommel F (eds.) 1993 *Governing Together*. Macmillan, London, UK

Bunce V 1981 *Do New Leaders Make a Difference*. Princeton University Press, Princeton, NJ

Burns J M 1978 *Leadership*. Harper & Row, New York

Cammack P, Pool D, Tordoff W 1988 *Third World Politics*. Macmillan, London

Castles F G (ed.) 1982 *The Impact of Parties*. Sage, London

Castles F G, Wildenmann R (eds.) 1986 *Visions and Realities of Party Government*. De Gruyter, Berlin

Cronin T E 1975 *The State of the Presidency*. Little, Brown, Boston

Finer S E 1962 *The Man on Horseback*. Pall Mall, London

Heclo H 1977 *A Government of Strangers*. Brookings Institution, Washington, DC

Hook S 1955 *The Hero in History*. Beacon Press, Boston

Kellerman B (ed.) 1984 *Leadership*. Prentice-Hall, Englewood Cliffs, NJ

Neustadt R E 1960 *Presidential Power*. Wiley, New York

Rejai M, Phillips K 1983 *World Revolutionary Leader*. Sage, London

Shugart M S, Carey G M 1992 *Presidents and Assemblies*. Cambridge University Press, New York

Strom K 1990 *Minority Government and Majority Rule*. Cambridge University Press, New York

Weber M 1968 *Economy and Society*. Bedminster Press, New York, 3 Vols

J. Blondel

Grammar: Functional Approaches

The term *functional* or *functionalist* has a meaning in linguistics distinct from the use of the term found in anthropology or philosophy. In linguistics, a func-

tional analysis is an analysis of linguistic structure that puts the primary emphasis on explanation in terms of factors outside the structural form, such as meaning or the function of language in discourse. Functional analysis is in contrast to *formal* analysis, which is an analysis of linguistic structure that puts the primary emphasis of explanation on principles defined purely in terms of structural form. The term *functionalist* has been applied to any school of grammatical analysis which argues for the primacy or centrality of functional explanations over formal explanations.

1. Introduction

Such a broad characterization of functionalism in linguistics, of course, requires further elaboration. We begin by illustrating the commonest types of functional explanations, and then turn to different schools of linguistic theory that call themselves functionalist.

2. Types of Functional Explanations

There are three types of functional explanations that usually are distinguished in the functionalist linguistic literature: processing, cognitive, and discourse or information structure explanations.

2.1 Processing Explanations

Processing explanations make reference to models of the production and comprehension of utterances. A processing explanation argues that some aspect of grammatical structure is due to limitations on the human ability to produce or comprehend the structure.

Processing explanations typically are employed to account for two sorts of linguistic facts. One is the relative frequency of a particular grammatical structure across the world's languages. For example, it has been observed that suffixes are on the whole more frequent than prefixes (or more complex alterations of word forms) in the world's languages. A processing explanation has been proposed for the suffixing preference. It is hypothesized that speakers process word stems before their inflections. Psycholinguistic experiments imply that the beginning of a word is most salient, hence the word stem should ideally occupy the initial position; the ends of words are the next most salient, hence inflections should ideally occupy the final position.

The second sort of phenomenon for which processing explanations are employed is to account for degrees of acceptability of a construction in a single language. For example, in English, subordinate clauses embedded in the middle of a clause are relatively less acceptable than subordinate clauses embedded at the end of a clause: compare *I know the boy that owned the cat that caught your parakeet* to *The parakeet that the cat that the boy I know caught died.* This fact has been explained as a consequence of the difficulty in interpreting the latter sentence's center-embedded clauses. In fact, it is grammatical in English to move a relative clause from the middle of a clause to the end under certain circumstances: *A woman contacted me who I had shared an apartment with during my last year at college*.

Processing explanations have been proposed for constraints on certain types of syntactic structures, constraints on patterns of morphemes within words, and on sound patterns. The latter explanations typically are described in terms of ease of articulation or ease in discrimination (in speech perception). Processing explanations for syntax and morphology, especially in explaining differences in cross-linguistic frequency of word order and morpheme order, are particularly associated with the work of J. A. Hawkins (e.g. Hawkins 1994).

2.2 Cognitive Explanations

Cognitive explanations generally make reference to the conventional meaning of a word, inflection or syntactic construction (e.g. Bybee 1985, Tomasello 1998). A cognitive (semantic) explanation of a grammatical phenomenon essentially hypothesizes that some or all aspects of the analysis of the grammatical phenomenon make reference to conventional semantic properties. For example, some languages have plural inflections for all count nouns, while other languages restrict plural inflections to only nouns referring to animate beings, still others to only nouns referring to humans. This variation is built on a general ranking or hierarchy of nouns according to the semantic class of their referents: human < animate < inanimate. The description of the constraint on the occurrence of the plural inflection in each language, and the cross-linguistic pattern of constraints, requires reference to the semantic classes of entities denoted by the noun, described in terms of animacy.

Cognitive explanations, like processing explanations have been used to describe both grammatical constructions in particular languages and also patterns of cross-linguistic variation. The example of the animacy hierarchy in the preceding paragraph illustrates both sorts of explanations. It hypothesizes that the distribution of the plural inflection in any particular language will make reference to the semantic class of the entity denoted by the noun. It also hypothesizes that the cross-linguistic variation in the distribution of the plural inflection is subject to the hierarchical ranking of animacy categories given in the preceding paragraph. In recent cross-linguistic research, explanations of cross-linguistic patterns have been offered in which the patterns of occurrence of a

construction are mapped onto a single semantic or conceptual space applicable to all languages.

Some cognitively oriented functionalists make a stronger claim about the relationship between grammatical form and semantic function, namely that certain grammatical categories are better described in semantic terms. This hypothesis will be discussed below on the contribution of cognitive linguistics to functionalism.

2.3 Discourse (Information Structure) Explanations

Discourse explanations make reference to the organization and presentation of information in discourse, that is, the structuring of information for the purpose of communication. Such explanations are called discourse explanations by some functionalist, and information structure explanations by other functionalists. For example, the English active and passive voice constructions (*The car hit the tree* vs. *The tree was hit by the car*) both describe the same event and participants, but differ in which participant is encoded as the subject of the construction. It has been argued (based on naturally occurring examples) that the participant that is more topical is encoded as subject. Another example is the choice of construction in response to a question such as *Wasn't that incredible when Mary called the boss a pig?* One can respond with *Yeah, it really shocked me that she called him that*, but not with **Yeah, what really shocked me was that she called him that*. The unacceptable response uses the pseudocleft construction, and the pseudocleft construction is acceptable only if the information 'something really shocked me' is already in the hearer's consciousness. It is not in this case, and hence is unacceptable.

As with processing and cognitive explanations, discourse functional explanations have been offered both to explain restrictions on the acceptability of specific constructions in particular languages, and to explain the frequency of distribution of particular linguistic features across languages. Examples of the former were given in the preceding paragraph. An example of the latter is certain patterns of word order. In the overwhelming number of languages, the subject precedes the object in the most common order of elements in a clause. It has been proposed that the reason for this asymmetry is that the subject tends to be the topic, and the most common discourse pattern is for the topic to precede the 'comment,' or at least other less topical participants.

3. Schools of Functionalist Analysis

One of the most fundamental issues that divides schools of linguistic theory is: what is the relationship between linguistic form and linguistic function? The relationship is not at all straightforward (Croft 1995). Also, depending on the position that a linguist takes on the relationship between form and function, certain additional hypotheses have to be made in order to accommodate linguistic facts. These hypotheses have to do with aspects of language not directly connected to the form-function relationship, such as the organization of grammatical knowledge in a speaker's mind, the nature of semantics, and the role of cross-linguistic evidence. Since these additional hypotheses are closely associated with various schools of functionalist analysis, they will be discussed here.

3.1 Functionalism and Formalism

Functionalist schools generally set themselves up as contrasting with the formalist approach, associated most strongly with N. Chomsky, but also advocated by other linguists who reject some of Chomsky's more specific claims about syntax. A purely formal explanation of a grammatical phenomenon makes reference exclusively to properties of grammatical form or structure. With respect to grammatical structure, a formal explanation of some pattern of sentence structure will make reference only to syntactic categories and structural relationships between the elements of a sentence, and not to the meaning of the words, inflections or constructions of that sentence, their discourse function, or their production and comprehension in language use. With respect to word structure, a formal explanation will make reference only to morphological categories and relations between roots and affixes, and not to aspects of the meaning and use of word forms. With respect to sound structure, a formal explanation will make reference only to the structure of the sound system independent of its articulatory and auditory realization.

This is not to say that formalist analyses have no place for the meaning and use of linguistic constructions. The formalist hypothesis is that there is a significant number of properties of grammatical form which can be explained in a purely formal way. The concept of a formal explanation is closely associated with the theoretical position that the organization of grammatical knowledge in a speaker's mind is divided into components which separate a linguistic form from its conventional meaning and use in discourse. A typical model describes grammatical knowledge as being compartmentalized into phonological, syntactic and semantic components. The phonological component is separated from its phonetic realization, and the syntactic component is separated from its semantic interpretation. Linguistic form is linked up to its phonetic realization or semantic interpretation, eventually: formal models have rules linking syntax to semantics and phonology to phonetics. The formalist hypothesis does however imply that linguistic form is organized in a purely formal way to a high enough

degree that it can be described best as existing in separate components. That is, the formal analysis of grammatical structure is self-contained relative to functional factors.

This is also not to say that formalists reject all explanations that make reference to functional factors. Formalism accepts that some, perhaps many, linguistic phenomena require reference to functional factors. However, formalists argue that there is a large body of linguistic phenomena, those which most linguists would consider to be central to grammatical structure, that can be analyzed without reference to any functional factors. Another way of describing this position is that the phonological and syntactic components represent a substantial core of a speaker's grammatical knowledge of a language.

Functionalist approaches to language challenge or reject the formalist model of the relationship between form and function to varying degrees. However, all functionalists accept that not all aspects of grammatical structure can be accounted for by meaning and use. There will always be some degree of arbitrariness in language structure, where arbitrariness is intended to mean not wholly explainable in terms of language function. The question is, how much arbitrariness is there in language structure, and to what extent does the arbitrary aspect of language structure form a self-contained system? Different functionalist schools give different answers to this question.

3.2 Autonomous Functionalism

The closest approach to formalism which calls itself functionalist can be called *autonomous* functionalism. Autonomous functionalism argues that the number of properties of grammatical form that are explainable solely in grammatical terms are considerably fewer than most formalist theories assume. In particular, the restrictions on the use of particular constructions are sought in functional properties, whether cognitive, processing or discourse-functional. To take a simplified example in the *There*-construction (as in *There is a bird in the bushes*), a noun phrase with an indefinite article (*a bird*) is acceptable but one with a definite article usually is not (compare) **There is the bird in the bushes*). A plausible explanation for this restriction is that the *There*-construction is used to introduce a new referent into the discourse, but definite noun phrases describe referents already known to both speaker and hearer. This would be an autonomous functionalist analysis of the restriction.

However, even in an autonomous functionalist approach, the rules governing the formation of the construction itself, such as the fact that *There* is the first element followed by a form of *be*, are considered to be purely formal. Autonomist functionalists focus their research efforts on accounting for as many of the constraints on the construction as possible in terms of functional properties, while leaving the basic structure of the construction for formalist analysis. The autonomist functionalist research program is found in the work of Kuno (1987), Prince (1988), Ward, and their students and associates. Their explanations generally appeal to properties of information structure rather than semantic properties. Because they accept the basic structural analyses of formalist approaches, autonomous functionalists generally do not reject the view that linguistic knowledge in a speaker's mind is organized into components. Instead, they have proposed that there is another component, the information structure component, which should be included along with the syntactic and semantic components.

In fact, many functional analyses of specific constructions in more radically functionalist approaches to language are actually autonomous functionalist analyses, that is, they account for the restrictions on syntactic constructions in terms of the meaning and use in discourse of the construction. We may call these autonomous functional analyses in whatever sort of functionalist approach they are found in, that is, autonomist functionalist analyses of particular constructions are not incompatible with more radical functionalist approaches.

3.3 Mixed Formalism/Functionalism

Another sort of approach, not clearly identifiable with any particular school of linguistic theory, but sometimes called functional, challenges the hypothesis of a sharp separation between the use of formal or functional properties to describe grammatical constructions. In these theories, the description of both the structure and the constraints on a construction can mix formal and functional properties. Two schools, *Functional Grammar* of S. Dik and co-workers, and *Role and Reference Grammar* of R. Van Valin and co-workers, both use formal and functional properties (typically semantics, but also information structure) to describe the grammatical structure of sentences as well as the constraints on them. In addition, many grammatical descriptions of languages informally describe both syntactic and semantic properties of constructions without implying any sharp division between the two.

3.4 Typological Functionalism

As was noted above, cognitive, processing and discourse explanations have been offered for patterns of cross-linguistic variation as well as in the analysis of the grammars of individual languages. A school of linguistic theory known as typology has arisen which

focuses on cross-linguistic variation and its limits (see, for instance, Greenberg 1990, Comrie 1989, Givón 1979, Hawkins 1988, Croft 1990). Much cross-linguistic variation appears to be built on the base of general functional patterns, as illustrated above for cognitive, processing, and discourse explanations. This observation has led some typologists to argue that all cross-linguistically verified universals of language are susceptible to functional explanation in this way. This approach to analysis is called typological functionalism or the functional-typological approach.

Typological functionalism makes a stronger claim than the schools described in the preceding sections. The autonomous functionalist approach assumes that there still remains a significant amount of syntactic structure that can be described in terms of a self-contained set of rules (which are further assumed to be in part language-universal). The mixed formal-functional theories do not assume that formal principles governing syntactic structure are self-contained, but those theories allow for the possibility that there are formal properties of languages that are universal for all languages. The typological functionalist approach hypothesizes that any universal generalizations about linguistic structure are ultimately explained in functional terms.

The typological functionalist approach does not, in fact cannot, assume that the structure of a particular language's grammar is entirely explainable in functional terms. For instance, one example given above was that the occurrence of a plural inflection on nouns could be accounted for by the hierarchy human < animate < inanimate. The semantic hierarchy is universal, but each language applies the hierarchy to the distribution of the plural inflection in its own way, that is, one cannot predict what semantic class a language will extend (or restrict) its plural inflection to. That is, the exact class of nouns allowing a plural inflection is partly arbitrary, although it is partly explainable in terms of the semantic hierarchy: if the language allows plural inflection for animate nouns, it will allow plural inflection for human nouns. This sort of explanation of a linguistic phenomenon is sometimes described as motivation, in contrast to a set of rules that exhaustively predicts the forms that would be found.

A typological functionalist analysis must allow for some degree of arbitrariness in accounting for cross-linguistic patterns. If there were no arbitrariness, all languages would be alike in their grammatical structure; and they are not. But the typological functionalist approach hypothesizes that the arbitrary aspects of language structure are language-specific, and that any universal generalizations about linguistic structure are ultimately explained in functional terms.

An important element of typological functional explanations is the hypothesis of iconicity or iconic motivation, in which grammatical structure is hypothesized to reflect semantic or conceptual structure to some degree. For example, the closer grammatical relationship between verb and object as opposed to verb and subject in many languages is attributed to the closer semantic relationship between states of affairs and the participants usually encoded as the object. Iconic motivation can be used to account for grammatical structure itself, not just constraints on grammatical structures. In this respect, typological functionalism represents a more radical functionalist approach than autonomous functionalism.

3.5 Cognitive Linguistics and Functionalism

Cognitive linguistics (e.g. Lakoff 1987, Langacker 1987, Talmy 2000) is a broad school of linguistic theory that is usually called functionalist. Cognitive linguistics, as its name indicates, focuses on cognitive (semantic) explanations for grammatical structure. Cognitive linguistics has contributed a number of ideas to other functionalist approaches that are relevant to assessing functional approaches.

Cognitive linguistics has proposed a model of grammatical knowledge, construction grammar, that offers an alternative to the formalist model in which form and function (semantics and information structure) are separated in distinct components. Construction grammar hypothesizes that knowledge of syntax involves the knowledge of individual constructions which combine formal properties (syntactic structures, morphological inflections) and functional ones (semantics and discourse function). For example, the ditransitive construction [X Verb Z Y], as in *Marilyn sent Gina a book*, specifies not only the syntactic structure of the construction but also the semantic relationships among the participants, such as the fact that Z comes to possess Y. Constructions are organized in a network of grammatical knowledge in a speaker's mind. More broadly, grammatical knowledge is organized as a system of signs or symbols, consisting of form (the signifier) and function (the signified). A model organizing grammatical knowledge in terms of constructions/symbols/signs allows for a direct statement of functional properties that are relevant to particular grammatical constructions and other grammatical units.

Cognitive linguistics argues that semantics involves conceptualization or construal of an experience by a speaker for the purposes of linguistic communication. For example, an English count noun can be used in a mass noun grammatical context, as in *There was a huge Buick there; just acres of car* (attested example). The noun *car*, normally denoting an individuatable object, is construed to denote an unbounded quality (size). These and other examples indicate that the English count noun-mass noun distinction has at least some semantic basis. Cognitive linguists also accept that some alternative construals are conventionalized,

such as *chicken* (meat, mass noun) vs. *chicken* (live bird; count noun). Nevertheless, the conventionalized grammatical distinction conforms to the semantics of count noun vs. mass noun: a live bird is an individuatable entity, while meat is a unbounded substance.

Conventionalized construals are of course partly arbitrary (see below); but the semantic regularity of both conventional and novel construals has led cognitive linguists to argue that even grammatical categories such as 'noun' and 'subject' are susceptible to a cognitive semantic explanation. That is, a grammatical category such as 'noun' encodes a construal of an entity in a certain way, which can be discerned by the occurrence of particular word roots directly as nouns, and in the semantic changes that occur when a non-noun root is used as a noun, or can be morphologically derived as a noun.

Finally, cognitive linguistics has developed a model of categorization which has in turn been exploited by functionalist linguists. A classical view of meaning is that the meaning of a word, inflection or grammatical category must be a single definition that will cover all and only the appropriate uses of the word or category. Cognitive linguists have argued that many word and category meanings can be described in terms of polysemy, that is, distinct but semantically related meanings. For example, the meaning of *over* in *The bird flew over the hill* (path of motion) is distinct from its meaning in *The house is over the hill* (location), but they are related systematically. More specifically, cognitive linguists (drawing on research by psychologists) have suggested that many polysemous categories have one meaning that can be considered central or basic, called the category's prototype. Cognitive and other functionally oriented linguists have argued that many if not all grammatical categories are in fact polysemous, and many of these categories have a semantic/discourse-functional prototype which is found to be valid across languages (see *Linguistics: Prototype Theory*).

3.6 More Radically Functionalist Schools

Grammatical structure is commonly assumed to exist in speakers' minds. However, grammatical structure is also directly involved in social interaction in language use, and language use is central to accounting for language acquisition, language variation and language change. In the more dynamic process of language acquisition and language change, functional factors have been argued to play a role. In both language acquisition and language change, it has been argued that competing motivations among functional principles play a major role.

The concept of competing motivations is that (functional) principles may come into conflict such that there is no grammatical system that satisfies all of the functional principles. As a result, change occurs over time in acquisition and in the history of a language, and the languages of the world exhibit structural diversity even though their speakers' linguistic behavior conforms to the same functional principles. (Competing motivation models do not presuppose that the competing principles are functional, and in fact competing motivation models are now used by many formalists; see below).

The most general and commonly offered example of competing motivations is that between economy and iconicity (see *Linguistics: Iconicity*). *Economy* is the principle that a speaker uses the least effort necessary to express himself or herself. For example, English leaves the singular number of nouns unexpressed: *book-ϕ* vs. *book-s*. Economy is considered to be a speaker-oriented functional principle: its influence on language is for the benefit of the speaker. *Iconicity* is, in part, the principle that all of the relevant parts of the meaning conveyed are in fact conveyed by grammatical elements in the utterance (words, inflections, etc.). For example, in some languages, both the singular and the plural of a noun are expressed by overt suffixes. This aspect of iconicity is considered to be hearer-oriented: any aspect of meaning left out by the speaker may not be recoverable by the hearer. Economy and iconicity compete with each other: a linguistic expression that is economical will not be iconic (since it leaves some elements of meaning unexpressed), and an expression that is iconic will not be economical (since certain elements of meaning are not left unexpressed). Hence, across languages, there is diversity in the expression of the category of number, and languages change from one form of expression of number to another over time.

Perhaps the best known competing motivations model in language acquisition is the competition model of E. Bates and B. MacWhinney (1989). In language change, a number of linguists have put forward competing motivations models, in particular J. Haiman (1985).

More recently, some functionalist linguists (P. Hopper, J. Haiman, and J. Bybee) have emphasized the dynamic character of language in ordinary use, and have argued that a speaker's grammatical knowledge should not be considered to be as static and immutable as is usually believed. They argue that a speaker's grammatical knowledge is not a tightly integrated system, but rather a more loosely structured inventory of conventionalized routines that have emerged through language use. The empirical research of the functionalist linguists who advocate this view has focused on the role of frequency of use on the entrenchment of grammatical knowledge, and on inductive models of abstracting grammatical knowledge from exposure to language use, both in acquisition and in adult usage. There have been a number of studies of inflectional paradigms of words that support the hypothesis that frequency of use influences

grammatical representation; however, studies of syntactic construction in this approach are in their infancy.

4. Directions and Prospects for Functional Approaches to Grammar

Taken as a whole, functional approaches to grammar are widely practiced in linguistics. Nevertheless, there are important divisions between different schools of functionalist theory and between functionalists and formalists.

A common misunderstanding (among formalists in particular) is the belief that functionalists do not accept the existence of *arbitrariness* in language, that is, that not all aspects of grammatical knowledge are explainable by functional principles. There is no functionalist that holds that view. However, most functionalist theories have not developed a model of the representation of grammatical form, instead focusing their attention on the analysis of functional domains and their role in accounting for grammatical structure. There are significant exceptions to this observation, notably the mixed formal/functional theories of Dik and Van Valin, the model of lexical and morphological representation of Bybee, and construction grammar in cognitive linguistics. Construction grammar appears to be gaining a wider audience, and some functionalists are beginning to develop more detailed models of grammatical structure that are compatible with functionalist theories.

A significant split in functionalism is between those who are more cognitively oriented, focusing their attention on cognitive explanations, and those who are more discourse oriented. These two functionalist approaches have largely gone their own ways, developing models of language meaning (cognitive linguists) and language in use (the discourse functionalists) However, there is some functionalist work which draws on both cognitive and discourse models (see, for example, Givón (1979 and Chafe (1994). Some functionalists have argued for a rapprochement between the two approaches, since they are compatible: language must be understood from both a psychological and a social-interactional perspective.

Functionalist approaches to language generally have also neglected research in sociolinguistics (see *Sociolinguistics*). Sociolinguistics begins with the hypothesis that language is a dynamic, variable phenomenon, and argues that factors outside linguistic form, namely social parameters, account for variation and change in language. In both of these respects, sociolinguistics shares theoretical assumptions in common with functionalism. The increase of research on social-interactional factors and on relative frequency of use in functionalist research may lead to an integration of social with other functional factors in a social-functional theory of linguistic variation and change.

A rapprochement between functionalism and formalism is less likely, because the two differ in fundamental assumptions about the nature of language and the relation between form and function. There are some respects in which practical research is converging among formalists and functionalists, however. For instance, competing motivations as a mechanism of analysis is now widely used in formalist theories under the name of *Optimality Theory*. The rise of computer-aided corpus research has led to a convergence in interests on naturally occurring language and the role of usage in grammatical structure. Some formal semanticists have made reference to prototypes and polysemy, and some semanticists working in a generative grammar framework (Jackendoff 1983) also give conceptualization a major role in analysis.

See also: Functionalism, History of; Functionalism in Anthropology; Functionalism in Sociology; Generative Grammar; Grammaticalization; Linguistic Typology

Bibliography

Bates E, MacWhinney B 1989 Functionalism and the competition model. In: Bates E, MacWhinney B (eds.) *The Cross-linguistic Study of Sentence Processing*. Cambridge University Press, Cambridge, UK

Birner B J, Ward G 1998 *Information Status and Noncanonical Word Order in English*. Benjamins, Amsterdam

Bybee J L 1985 *Morphology: A Study into the Relation between Meaning and Form*. Benjamins, Amsterdam

Chafe W 1994 *Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press, Chicago

Comrie B 1989 *Language Universals and Linguistic Typology*, 2nd edn. University of Chicago Press, Chicago

Croft W 1990 *Typology and Universals*. Cambridge University Press, Cambridge, UK

Croft W 1995 Autonomy and functionalist linguistics. *Language* **71**: 490–532

Dik S C 1997 *The Theory of Functional Grammar*, 2nd rev. edn. 2 Vols. Mouton de Gruyter, Berlin

Givón T 1979 *On Understanding Grammar*. Academic Press, New York

Greenberg J H 1990 *On language: Selected writings of Joseph H. Greenberg* (ed. Denning K, Kemmer S). Stanford University Press, Stanford, CA

Haiman J 1985 *Natural Syntax*. Cambridge University Press, Cambridge, UK

Hawkins J A (ed.) 1988 *Explaining Language Universals*. Blackwell, Oxford, UK

Hawkins J A 1994 *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge, UK

Jackendoff R 1983 *Semantics and Cognition*. MIT Press, Cambridge, MA

Kuno S 1987 *Functional Syntax: Anaphora, Discourse and Empathy*. University of Chicago Press, Chicago

Lakoff G 1987 *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, Chicago
Langacker R W 1987/1991 *Foundations of Cognitive Grammar*, 2 Vols. Stanford University Press, Stanford, CA
Prince E F 1988 Discourse analysis: A part of the study of linguistic competence. In: Newmeyer F J (ed.) *Linguistics: The Cambridge Survey, vol. II: Linguistic Theory: Extensions and Implications*. Cambridge University Press, Cambridge, UK
Talmy L 2000 *Toward a Cognitive Semantics*, 2 Vols. MIT Press, Cambridge, MA
Tomasello M (ed.) 1998 *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Erlbaum, Mahwah, NJ
Van Valin R D Jr, LaPolla R J 1997 *Syntax: Structure, Meaning and Function*. Cambridge University Press, Cambridge, UK

W. Croft

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Grammatical Agreement

Grammatical agreement is a phenomenon in which word forms co-occurring in a clause are sensitive to each other. Inflected forms often agree in their values of number, gender, or person, as can be seen from the contrasting examples in (1a, b) to (4a, b)

(1) French a. *La*^{fsg} *fille*^{fsg} *est*$^{\text{3sg}}$ *belle*^{fsg}.
'The girl is beautiful'
b. *Les*^{pl} *garcons*^{mpl} *sont*$^{\text{3pl}}$ *beaux*^{mpl}.
'The boys are handsome'

(2) French a. *Je*$^{\text{1sg}}$ *les*$^{\text{3pl}}$ *ai*$^{\text{1sg}}$ *vues*^{fpl}.
'I have seen them'
b. *Tu*$^{\text{2sg}}$ *l'*$^{\text{3sg}}$ *a*$^{\text{2sg}}$ *vu*^{msg}.
'You have seen him'

(3) Finnish a. *Taina löysi minun*$^{\text{1sg}}_{\text{GEN}}$ *kirja-ni*$^{\text{1sgPoss}}$.
'Taina found my book'
b. *Taina löysi heidän*$^{\text{3pl}}_{\text{GEN}}$ *kirja-nsa*$^{\text{3Poss}}$.
'Taina found their book'

(4) German a. *Ein kleines*^{nsg} *Mädchen*^{n} *saß*$^{\text{3sg}}$ *auf der Treppe*, *und es*$^{\text{3nsg}}$ *lachte*.
'A little girl sat on the stairs, and she laughed'
b. *Ein kleiner*^{msg} *Junge*^{msg} *stand*$^{\text{3sg}}$ *daneben*, *und er*$^{\text{3msg}}$ *heulte*.
'A little boy stood beside, and he cried'

In (1), the definite article agrees with the following noun, and both the copula verb and the predicative adjective agree with the whole subject phrase. If any item of (1a) is exchanged with the corresponding item of (1b), the sentence becomes ungrammatical (e.g., **la fille est beaux*). Example (2) shows two agreement relations interwined: the auxiliary agrees with the subject pronoun, while the participle agrees with the object clitic. Sentence (3) is an example of a possessed noun agreeing with the possessor in the genitive. In (4) it is the pronoun in the second clause that agrees with the antecedent noun phrase in the first clause. Notice that *Mädchen* 'girl' is a neuter noun, triggering a neuter pronoun even though it refers to a female being; in contrast, the English pronouns in the translation are selected solely on semantic grounds.

Grammatical agreement can be seen as an interplay of morphological, syntactic and semantic aspects (Anderson 1988, Lapointe 1988, Corbett 1994, 1995, 1998, Pollard and Sag 1994). Morphosyntactically, agreement is a relation between syntactic constituents in virtue of the fact that the word forms they consist of bear similar information, either inherently or by means of morphological affixation (inflection). Semantically, agreement serves to keep record of discourse referents: only constituents that relate to the same referent may overtly agree with each other (Lehmann 1982). The sentences in (1) are about one referent (a single person or a group of persons), while those in (2) are about two referents in a seeing-relation.

Generally, agreement relations are found among the following elements:

(a) NP- (or DP-) internally: determiner and noun, attribute and noun, possessor and possessed noun;

(b) a predicate and its arguments: verb–subject, verb–object, preposition–object, predicative noun/adjective–argument; and

(c) a pronoun or anaphor and its antecedent.

The categories that may appear in these relations are gender (e.g., feminine f, masculine m, animate, or neuter n), number (e.g., dual, plural pl, and singular sg), and person (1st, 2nd, and 3rd, or inclusive 1 + 2).

Nouns and pronouns (or pronominal affixes/clitics) function as the controller (or source) of an agreement relation in virtue of their features. Nouns are classified inherently for gender (or noun class), and sometimes also for number (cf. *scissors*, *trousers*, *brethren*), whereas pronouns are classified inherently for person and number, and sometimes also for gender. In some instances, this inherent classification is semantically based: as a default, female beings are represented by feminine nouns, but male beings by masculine nouns; aggregates are represented by plural nouns; addressees are represented by second person pronouns, but speakers by first person pronouns. In most other cases feature assignment is purely grammatical (consider the distinct genders in French *lune*^{f}, *soleil*^{m} vs. German *Mond*^{m}, *Sonne*^{f}, 'moon,' 'sun') and may even override the default (cf. German *Mädchen*^{n} 'girl') (Corbett 1991).

The grammatical elements that agree with the noun or the pronoun are said to be the agreement targets (or controllees); in (1), the article *la* is the target of *fille*, which is f, and both the copula *est* and the adjective *belle* are targets of the DP *la fille*, which is fsg. The agreement features of controller and target do not have to be identical, but they must be compatible with each other, as in *la fille*^{fsg} and *est*$^{\text{3sg}}$. In order for a DP to be the controller of an agreement relation, the

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

cc cc cc

cc cc

o o

ff

(8)

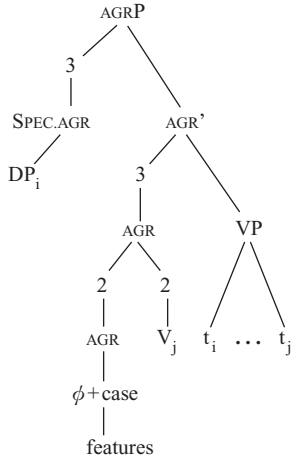

ff

ffi

(9)

D: N: A:

$\lambda N\ \iota z^{fsg}\ \{\text{DISTAL}(z)\ \&\ N(z)\}$ $\lambda x^{fsg}\ \text{WOMEN}(x)$ $\lambda y^{fsg}\ \text{BEAUTIFUL}(y)$

NP: $\lambda x^{fsg \sqcup fsg}\ \{\text{WOMEN}(x)\ \&\ \text{BEAUTIFUL}(x)\}$

DP: $\iota z^{fsg \sqcup fsg \sqcup fsg}\ \{\text{DISTAL}(z)\ \&\ \text{WOMEN}(x)\ \&\ \text{BEAUTIFUL}(x)\}$

(11)

$\lambda P\ \lambda x^{3sg}\ \text{FEAR}(x,P(x))$ $\lambda y^{msg}\ \text{ARRIVED_TOO_LATE}(y)$

$\lambda x^{3sg \sqcup msg}\ \text{FEAR}(x,\text{ARRIVED_TOO_LATE}(x))$

(13)

o

$\lambda y\ \lambda u^{fsg}\ \mathrm{EAT}(u,y)$

$\lambda y^{fsg}\ \lambda u\ \mathrm{EAT}(u,y)$ $\lambda P^{f}\ \lambda x\ \mathrm{WANT}(x,P(x))$

ERG

$\lambda y^{fsg\sqcup f}\ \lambda x\ \mathrm{WANT}(x,\mathrm{EAT}(x,y))$

ERG

ff

ffi

cc

cc

ff

ffi

Grammatical Gender

Gender is a remarkable category, pervasive in some languages, for instance in the Indo-European and Dravidian families, and lacking in others, as in the Uralic family. We start with a brief consideration of terms, then examine how the number of genders in a given language can be established. The number varies considerably: two and three genders are often found, four and five are not unusual, while Fula (a Niger-Kordofanian language) has around 20. Most attention will be given to the problem of how the speaker assigns nouns to genders. Gender systems may have sex as a component, as in languages with masculine and feminine genders; but equally sex may be irrelevant, as in the Algonquian languages where the distinction is between animate and inanimate. Finally we consider prospects for investigating the category further.

1. *Terms*

The defining characteristic of gender is agreement; a language has a gender system only if noun phrases headed by nouns of different types control different agreements. The evidence that nouns have gender in a given language thus lies outside the nouns themselves. There is no substantive difference between 'genders' and 'noun classes'; the different terms are merely the products of different linguistic traditions. For instance, we find systems with three genders, to which nouns are assigned by similar rules, in both Kannada (a Dravidian language) and in Godoberi (a Nakh-Daghestanian language). By tradition the first is said to have three genders, and the second three noun classes. We shall use the term 'gender' in both instances.

2. *The Number of Genders*

The approach to gender sketched in Sect. 1 is based on Zaliznjak's (1964) notion of 'agreement class.' Two nouns are in the same agreement class only if they take the same agreements under all conditions, that is if we control for other grammatical categories such as case and number. Conversely, if two nouns differ in their agreements when factors such as case and number are held constant, then they belong to two different agreement classes and it will normally be the case that they belong to two different genders. While in many languages there is no dispute as to the number of genders, there are other languages where the question is far from straightforward. In the case of Romanian there has been a lengthy debate. The analytical problem of determining the number of genders and the tests for deciding the gender of a given noun depends on separating out the classes into which nouns are divided (the controller genders) from the number of different genders marked on agreement targets (the target genders). Frequently the two match up, but in several languages they do not. A full treatment of the subject is not possible here; for a detailed account and extensive references see Corbett (1991).

3. *Gender Assignment*

We now ask how nouns are distributed over the genders of a given language, in other words, how does a native speaker 'know' what gender a noun belong to? Models of the mechanisms by which nouns are allotted to genders are called 'assignment systems.' Assignment may involve two sorts of information about the noun: its meaning and its form. We shall consider semantic systems first.

3.1 *Strict Semantic Assignment Systems*

This type is found in Dravidian languages like Kannada, spoken in the state of Karnataka, southern India (Sridhar 1990, p. 198). Nouns denoting male humans are masculine, those denoting female humans are feminine. There are also deities, demons, and heavenly bodies in these genders. All remaining nouns, including those denoting infants and animals, are neuter. Thus *appa* 'father,' and *candra* 'moon' are masculine, *amma* 'mother' is feminine, and *na:yi* 'dog' is neuter. Certain other Dravidian languages like Kolami have only two genders: nouns denoting male humans are masculine and all others belong in the nonmasculine gender (Emeneau 1955, p. 73). Elsewhere we also find the reverse. Thus in Diyari, a language of South Australia, there is a gender for nouns with female referents (such as women, girls, doe kangaroos), and the other is for all remaining nouns (Austin 1981). In Alamblak, a Sepik Hill language of Papua New Guinea, the masculine gender includes nouns denoting males and those denoting things like crocodiles, pythons, and arrows, which are long and thin, while the feminine is for nouns denoting females or short, squat items like turtles, frogs, and chairs (Bruce 1984, pp. 96–8, Foley 1986, pp. 80–81). An interesting system with some similar features is found in Taiap (Kulick and Stroud 1992).

3.2 *Predominantly Semantic Assignment Systems*

Many languages have semantic assignment rules, which, however, lack the coverage of those seen in Kannada. An example of what we shall call predominantly semantic assignment is found in Tsakhur, which is a member of the Lezgic group of the Nakh-Daghestanian family. Its four genders are shown in Table 1.

Assignment to genders I and II is straightforward: I is for male humans (also gods, angels, and so on) while II is for female humans (and female mythical beings). Most remaining animates are assigned to gender III,

Table 1
Gender assignment in Tsakhur

Criterion	Gender	Examples	Glosses
male rational	I (masculine)	dak_j	father
		gade	boy
female rational	II (feminine)	jed_j	mother
		sos	bride
animate (also some inanimates)	III (animate)	balkan	horse
		dadal	cock
		(qow)	(nose)
other (including a few animates)	IV (neuter)	t'et'	flower
		niṡe	cheese
		(kabaj)	(butterfly)

though a few belong to gender IV, along with some mythical beings. Inanimates are found in genders III and IV (Melínikov and Kurbanov 1964, Ibragimov 1990, pp. 54–6, and references therein, Kibrik 1999, pp. 48–9).

Languages of this general type are widespread. For some of them, researchers have proposed that abstract semantic criteria partly miss the point, and that if we can gain a better grasp of the world view of the speakers, we can then understand the assignment system. The most discussed case is that of Dyirbal, a language of North Queensland (Dixon 1972), which has four genders, primarily for: male humans and nonhuman animates (I), female humans (II), non-flesh food (III) and the residue (IV). There are numerous apparent exceptions: the moon is in the first, masculine gender and the sun is the second, feminine gender. However, in Dyirbal mythology, as indeed in much of Australia, the moon is the husband of the sun; in Dyirbal the mythological significance of referents is a key factor in assignment. Dixon's data and analysis continue to provoke debate, as shown by the recent paper by Polinsky (1998). There is similar debate as the role of world view in the gender assignment system of Ojibwa (and in other Algonquian language of southern Canada and northern USA).

In all the languages discussed so far, the meaning of the noun determines gender. In the strict assignment systems, the rules are immediately obvious; in the predominantly semantic systems, there may be a minority of exceptions, but these exceptions have been claimed to be largely apparent in some languages, once the cultural setting of the language is taken into account. If we ask which are the semantic criteria on which semantic systems can be based, we see recurring patterns. Quite often we find animate ~ inanimate, human ~ nonhuman, and male ~ female. Sometimes there is a gender for diminutives, as in various Bantu languages. There are also less usual genders, such as that for nonflesh food (Dyirbal) and the gender for insects (found in the Rikvani dialect of the Nakh-Daghestanian language Andi). A criterion which is sufficient to define a gender in one language may be just one factor in the assignment to a gender in another. Thus the Bantu language Chichewa has a gender for diminutives, while in Dizi, diminutives together with nouns denoting females form a gender.

3.3 Formal Assignment Systems

In many languages, assignment by semantic rules would fail to assign many nouns to a gender. While in languages like Kannada the nouns not assigned by the semantic rules (the 'remainder' or 'semantic residue'), all belong to a single gender, in many languages they are spread over more than one gender. In such languages we find additional rules for assigning nouns to genders according to their form. Note the asymmetry: languages may use semantic rules, or semantic *and* formal rules, but not just formal assignment rules. In no language are nouns assigned to genders, as defined earlier, by purely formal rules. An example would be a language in which there were two agreement classes, and the nouns in the first all ended in a consonant cluster, and those in the second did not, and there was no semantic regularity for the distribution of nouns. I claim that this hypothetical type does not exist. Formal assignment rules may appeal to two types of information: phonological and morphological.

3.3.1 Phonological. A clear example of assignment depending on phonological information is provided by Qafar (Afar), an East Cushitic language spoken in north-eastern Ethiopia and in Djibouti (data from Parker and Hayward 1985). Qafar has rather standard semantic assignment rules, namely that for sex-differentiable nouns, those denoting males are masculine and those denoting females are feminine. It is the nouns which fall outside these rules, the residue,

Table 2
Russian nouns belonging to the semantic residue

Masculine	Feminine	Neuter
žurnal 'magazine'	gazeta 'newspaper'	pis′mo 'letter'
dom 'house'	izba 'hut'	zdanie 'building'
čaj 'tea'	voda 'water'	pivo 'beer'
dub 'oak'	sosna 'pine'	derevo 'tree'
ogon' 'fire'	peč′ 'stove'	plamja 'flame'
avtomobil′ 'car'	maš ina 'car'	taksi 'taxi'
stul 'chair'	taburetka 'stool'	kreslo 'armchair'
den′ 'day'	noč′ 'night'	utro 'morning'
čas 'hour'	minuta 'minute'	vremja 'time'
nerv 'nerve'	kost′ 'bone'	serdce 'heart'
nos 'nose'	ščeka 'check'	uxo 'ear'
lokot′ 'elbow'	lodyžka 'ankle'	koleno 'knee'
flag 'flag'	èmblema 'emblem'	znamja 'banner'

Table 3
Inflectional classes in Russian

	I	II	III	IV
NOMINATIVE	žurnal	gazeta	kost′	pis′mo
ACCUSATIVE	žurnal	gazetu	kost′	pis′mo
GENITIVE	žurnala	gazety	kosti	pis′ma
DATIVE	žurnalu	gazete	kosti	pis′mu
INSTRUMENTAL	žurnalom	gazetoj	kost′ju	pis′mom
LOCATIVE	žurnale	gazete	kosti	pis′me
gloss	'magazine'	'newspaper'	'bone'	'letter'

which are of interest. For them the phonological rules apply: nouns whose citation form ends in an accented vowel are feminine (for example, *karmà* 'autumn') while all others are masculine (for example, *gilàl* 'winter,' which does not end in a vowel, and *tàmu* 'taste,' which does end in a vowel, but not an accented one). These rules operate with few exceptions. Moreover, nouns denoting males and females typically fit them too (for example *bàqla* 'husband' and *barrà* 'woman, wife'). It might seem that we could dispense with semantic rules for Qafar, and treat it as being of a quite different type. However, while the phonological rule works for the vast majority of cases, we find the crucial example *abbà* 'father,' which is masculine, even though it ends in an accented vowel, which 'should' be an indicator of feminine gender. Conversely, *gabbixeèra* 'slender-waisted female' is feminine, though the accent is non-final. Qafar is remarkably for the simplicity of its phonological assignment rules, and for the fact that when the semantic and phonological rules are both applicable, they almost always give the same result. However, in cases of conflict, the semantic rules take precedence (as is the normal situation in gender assignment systems).

Other examples can occur in various parts of the world. Even French, often claimed to have no system to its gender assignment, has been demonstrated to have a phonological system, though not as simple as that of Qafar. For example, of 938 nouns ending in $\tilde{\varepsilon}$, 99 percent are masculine (like *le pain* [p$\tilde{\varepsilon}$] 'the bread'). And of 1,453 nouns in /ʒ/, 94.2 percent are masculine (like *le ménage* [mena:ʒ] 'the household' (for details see Tucker et al. 1977).

3.3.2 Morphological. In Russian, which is typical of many other Indo-European languages, nouns denoting male humans are masculine and those denoting female humans are feminine. But unlike the situation in languages like Kannada, it is not the case that the semantic residue is assigned to the neuter gender; in Russian the residue is shared between the three genders, with the neuter gender not even receiving the majority. Searching for additional semantic criteria is not at all promising, as the following data suggests (Table 2).

If we look instead at the morphology of the nouns, then we can make progress. There are four main inflectional classes in Russian, each with several thousands of nouns (for justification of this not uncontroversial view see Corbett 1982, pp. 202–11). There are six cases and two numbers (though no paradigm has 12 distinct forms because of various syncretisms). We give just the singular forms in Table 3.

The native speaker needs to know how a noun inflects, in order to produce grammatical utterances.

On the basis of that information, the assignment rules are straightforward. Nouns in class I are masculine, those in classes II and III are feminine, and those in IV are neuter. At first we might think that the semantic assignment rules are superfluous, since *mal'čik* 'boy' is in class I, while *devuška* 'girl' is in class II, and *mat'* 'mother' is in class III. In other words, many of the sex-differentiable nouns would be assigned to the appropriate gender by the morphological assignment rules. But there are also instances where this is not so, for instance, *djadja* 'uncle,' which denotes a male but is in class II, whose nouns are typically feminine. *Djadja* 'uncle' is masculine. Nouns like this show, once again, that we do not find languages where formal assignment rules are sufficient. Further rules are required in Russian for indeclinable nouns (like *taksi* 'taxi,' which is indeclinable and neuter). Morphological assignment systems are found in various other Indo-European languages. Looking further afield, in Arabic too, gender is assignable in the main according to morphology (Cowell 1964, pp. 372–5).

The different types of assignment criteria may overlap in ways which make it difficult to establish their relative weight in a given language. Thus there may be small clusters of nouns which can be accounted for by semantic criteria (apart from those covered by the main semantic rules) even within systems where formal rules have a major role (as shown for German by Zubin and Köpcke 1986, Köpcke and Zubin (1996). However, whenever gender languages are analyzed in sufficient detail, the gender of the vast majority of nouns turns out to be predictable.

4. Prospects

Studies on grammatical gender continue to appear regularly, including valuable studies of previously unresearched languages. Here we highlight two main areas where we can expect substantial new developments: lexicology and modeling.

4.1 Lexicology

Gender is a key category for lexicology, since it is crucial for understanding the nature of lexical entries. In a gender language the gender of a noun must be available. On the other hand, as we have seen, it does not normally need to be specified, since the information can be derived, by assignment rules, from other information which must be stored. We still do not know the full range of information which may be relevant (for instance, the possible semantic features which may determine gender assignment).

There are further surprises: it is possible too for certain nouns to be of different genders according to the sex of the speaker; this is found in Garifuna, a member of the Arawak family spoken in Belize, Honduras, Nicaragua, and Guatemala (Taylor 1977, p. 60, Munro 1998). These two points raise interesting questions as to the shape of lexical entries, a general question which is of concern to lexicologists and increasingly to psycholinguists (see, for instance, Berkum (1995).

4.2 Modeling

When discussing some types of assignment system, the evidence is so clear-cut that alternative analyses are hard to imagine. However, with morphological assignment systems, particularly when the number of genders and the number of inflectional classes is similar in a given language, there are other possible analyses, in particular some would specify gender in the lexical entry and derive inflectional information from it. Arguments in favour of the approach taken in Sect. 3.3.2 include the following; first, and primarily, it is claimed that in languages like Russian the internal evidence shows that the gender from inflectional class approach is preferable (it proves impossible to predict inflectional class from gender but it is possible to predict gender from inflectional class); and second, there is the typological claim that one or other of the possible types of assignment system (or some combination) will always be in play in a gender language such that gender is predictable for the vast majority of nouns and so if the Russian type system were treated differently this would create an anomaly.

However, there is a further step which we can take, of a rather different nature. Since the Russian type languages are crucial, it worth investing additional time in clarifying them. Hence in Fraser and Corbett (1995) we give an account of the Russian gender and inflectional class relations in the Network Morphology framework. The analysis is implemented in the lexical knowledge representation language DATR (Evans and Gazdar (1996, Gazdar forthcoming), which means that the analyses can be verified by computer. Given minimal lexical entries for nouns our account does indeed predict the right genders for a wide range of nouns. An account like this which is both explicit and testable is not available for the alternative approach.

A more complex type is that in which gender is again assigned according to morphological class, but where the morphological class is itself predictable. One such case is Arapesh; a language of the Torricelli family, spoken on the north coast of Papua New Guinea. Genders and morphological classes form a substantial part of Fortune's (1942) grammar, which led to work based on it by Aronoff (1992, 1994, pp. 89–114). Once again, there are semantic assignment rules based on sex. But then there are morphological assignments rules, which determine 13 genders on the basis of 22 morphological classes (which are in turn predictable largely from phonological information). Again we have demonstrated how this system can be analyzed within the Network Morphology framework and have provided an implementation to demonstrate that our

claims are indeed valid (Fraser and Corbett (1997). There is a complex interaction of assignment to gender and to morphological class in Mayali, a non-Pama-Nyungan language of north Australia, which has four genders and five morphological classes for nouns. For the essential data see Evans (1997) and for details of a formal model of this system see Evans et al. (2001). Finally, there is interesting work on the development of assignment systems over time, including some modelling: see Comrie and Polinsky (1998), Polinsky and Everbroeck (1998), Polinsky and Jackson (1999).

5. *Conclusion*

Gender remains a puzzling category, apparently deeply embedded into the structure of some languages, and missing in other. This is found even in related languages: thus several Indo-Aryan languages have two or three genders, with substantial agreement, while another, Bengali, has no gender even in the personal pronouns. Where we find gender, the assignment rules can be simple, as in Kannada and Qafar, or complex as in French and German. It is an area in which gradual progress is being made and one in which there is still much to learn.

See also: Grammatical Agreement; Grammatical Number; Pronouns

Bibliography

Aronoff Mark 1992 Noun classes in Arapesh. In: Booij G, van Marle J (eds.) *Yearbook of Morphology 1991*. Kluwer, Dordrecht, The Netherlands, pp. 21–32

Aronoff M 1994 *Morphology By Itself: Stems and Inflectional Classes* (Linguistic Inquiry Monograph 22). MIT Press, Cambridge, MA

Austin P 1981 *A Grammar of Diyari, South Australia*. Cambridge University Press, Cambridge, UK

Berkum J J A van 1996 The psycholinguistics of grammatical gender: Studies in language comprehension and production. Ph.D. dissertation. Nijmegen University Press, Nijmegen, The Netherlands

Bruce L 1984 *The Alamblak Language of Papua New Guinea (East Sepik)*. Department of Linguistics, Research School of Pacific Studies, Australian National University, Canberra, ACT

Comrie B, Polinsky M 1998 Gender in a historical perspective: Radial categories meet language change. In: Justus C F, Polomé E C (eds.) *Language Change and Typological Variation: In Honor of Winfred P. Lehmann on the Occasion of His 83rd Birthday: II Grammatical Universals and Typology*. Institute for the Study of Man, Washington, DC, pp. 566–89

Corbett G G 1982 Gender in Russian: An account of gender specification and its relationship to declension. *Russian Linguistics* **6**(2): 197–232

Corbett G G 1991 *Gender*. Cambridge University Press, Cambridge, UK

Cowell M W 1964 *A Reference Grammar of Syrian Arabic (based on the dialect of Damascus) (Arabic series 7)*. Georgetown University Press, Washington, DC

Dixon R M W 1972 *The Dyirbal Language of North Queensland*. Cambridge University Press, Cambridge, UK

Emeneau M B 1955 *Kolami: A Dravidian Language*. University of California Press, Berkeley and Los Angeles

Evans N 1997 Head classes and agreement classes in the Mayali dialect chain. In: Harvey M, Reid N (eds.) *Nominal Classification in Aboriginal Australia*. John Benjamins, Amsterdam, pp. 105–46

Evans N, Brown D, Corbett G G 2001 The semantics of gender in Mayali: Partially parallel systems and formal implementation. *Language* **77**

Evans R, Gazdar G 1996 DATR: A language for lexical knowledge representation. *Computational Linguistics* **22**: 167–216

Foley W A 1986 *The Papuan Languages of New Guinea*. Cambridge University Press, Cambridge, UK

Fortune R F 1942 *Arapesh*. J J Augustin, New York

Fraser N M, Corbett G G 1995 Gender, animacy and declensional class assignment: A unified account for Russian. In: Booij G, van Marle J (eds.) *Yearbook of Morphology 1994*. Kluwer, Dordrecht, The Netherlands, pp. 123–50

Fraser N M, Corbett G G 1997 Defaults in Arapesh. *Lingua* **103**: 25–57

Gazdar G forthcoming Ceteris paribus. In: Kamp J A W, Rohrer C (eds.) *Aspects of Computational Linguistics*. Springer, Berlin

Ibragimov G X (ed.) 1999 *Caxurskij jazyk*. Nauka, Moscow

Kibrik A E (ed.) 1999 *Èlementy grammatiki caxurskogo jazyka v tipologičeskom osveščenii*. Nasledie Press, Moscow

Köpcke K M, Zubin D 1996 Prinzipien für Genuszuweisung im Deutschen. In: Lang E, Zifonoun G (eds.) *Deutsch typologisch: Jahrbuch des Instituts für deutsche Sprache 1995*. de Gruyter, Berlin, pp. 473–91

Kulick D, Stroud C 1992 The structure of the Taiap (Gapun) language. In: Dutton T, Ross M, Darrell T (eds.) *The Language Game: Papers in Memory of Donald C. Laycock*. Department of Linguistics, Research School of Pacific Studies, Australian National University, Canberra, ACT, pp. 203–26

Mel'nikov G P, Kurbanov A I 1964 Logičeskie osnovanija imennoj klassifikacii v caxurskom jazyke. In: Ivanov V V (ed.) *Voprosy struktury jazyka*. Nauka, Moscow, pp. 157–70

Munro P 1998 The Garifuna gender system. In: Hill J M, Mistry P J, Campbell L (eds.) *The Life of Language: Papers in Linguistics in Honor of William Bright*. de Gruyter, Berlin, pp. 443–61

Parker E M, Hayward R J 1985 *An Afar-English-French Dictionary (with Grammatical Notes in English)*. SOAS, University of London, London

Polinsky M 1998 Women are not dangerous things: Grammatical gender and categorization. Paper presented at the Workshop 'Processing of Grammatical Gender,' Max Planck Institute of Cognitive Neuroscience and Centre of Cognitive Sciences, University of Leipzig, Leipzig, 24–25 July 1998

Polinsky M, Everbroeck E van 2000 Noun classes: Language change and learning. Development of gender classifications: Modeling the historical change from Latin to French. Unpublished paper, UCSD

Polinsky M, Jackson D 1999 Noun classes: Language change and learning. In: Fox B A, Jurafsky D, Michaelis L A (eds.) *Cognition and Function in Language*. CSLI, Stanford, pp. 29–50

Sridhar S N 1990 Kannada (Descriptive Grammars Series). Routledge, London

Taylor D 1977 *Languages of the West Indies*. Johns Hopkins University Press, Baltimore, MD
Tucker G R, Lambert W E, Rigault A A 1977 *The French Speaker's Skill with Grammatical Gender: An Example of Rule-governed Behavior*. Mouton, The Hague
Zaliznjak A A 1964 K voprosu o grammatičeskix kategorijax roda i oduševlennosti v sovremennom russkom jazyke. *Voprosy jazykoznanija* **4**: 25–40
Zubin D, Köpcke K M 1986 Gender and folk taxonomy: The indexal relation between grammatical and lexical categorization. In: Craig C (ed.) *Noun Classes and Categorization: Proceedings of a Symposium on Categorization and Noun Classification, Eugene, Oregon, October 1983*. John Benjamins, Amsterdam, pp. 139–80

G. G. Corbett

Grammatical Number

Number is often taken to be a relatively straightforward category. It is often used as the textbook illustration of an inflectional category, and the situation in languages like English is believed by many to be the normal if not the only one. English makes a semantic distinction, and a formal distinction between, for instance, *cat* and *cats*, *child* and *children*. In fact the number category reveals many surprises, and the system found in English is a rather extreme one, when seen within the space of possibilities established from a survey of the languages of the world. We shall consider characteristics of the English system in turn, showing some of the ways in which it is typical or atypical. We shall give greatest prominence to the question of number values.

1. *Nominal and Verbal Number*

In English number is a nominal category, it involves the distinction between instances such as *cat* and *cats*. When it appears on verbs, this is by agreement with a subject noun phrase, as in *the children were playing happily*. The plural verb still indicates the number of children, and not the number of playing events. There are many languages which have verbal number, that is, where the number of the verb indicates the number of events (see Durie 1986 for examples).

2. *Number as an Obligatory Category*

Number is an obligatory category in English. If someone says, *there's a visitor waiting to see you*, when there are several, this is misleading and inappropriate. In many languages, however, the use of the plural is not obligatory, but is used when it is important to mark number. One example is the Cushitic language Bayso:

(1) Bayso (Dick Hayward, personal communications, Corbett and Hayward 1987)
lúban *foofe*
lion.GENERAL watched.1SG
'I watched lion' (it could be one, or more than that)

Another language, from many examples, is the Arawak family language Bare of Venezuela and Brazil (Aikhenvald 1995, pp. 18–19).

3. *The Nominals Involved in the Number System*

In English the majority of nominals can mark number, from the personal pronouns and nouns denoting persons (like *woman*) down to those denoting inanimate objects like *door*, and even to some denoting abstracts like *hour*. Many languages restrict the number opposition to fewer nominals, namely those which are high in animacy (Smith-Stark 1974). Thus in Slave (an Athapaskan language spoken in parts of the Northwest Territories, British Columbia, and Alberta, Canada) number may be marked on the noun phrase provided the noun denotes a human or a dog (Rice 1989, p. 247).

4. *The Semantics of Number*

As long as we stay with innocent examples like English *cat* and *cats*, the semantics of number look clear, since *cats* is appropriate for more than one instance of *cat*. In fact there are many concealed difficulties: a concise overview of work on number within formal semantics can be found in Link (1998), who shows how the study of plurals relates to different movements in linguistic semantics, and provides a substantial bibliography.

If we move from ordinary plurals to consider *I* vs. *we*, it becomes clear that here we have an 'associative' use of the plural; *we* is most often used for 'I and other(s) associated with me.' At the other end of the Animacy Hierarchy, nouns like *coffee* have a plural, provided they are recategorized: if *coffee* denotes a type, or a typical quantity of coffee, then there is an available plural, as in: *she always keeps three different coffees*, and *we've ordered two teas and three coffees*.

5. *Number Values*

A striking way in which English is limited in its number system is in terms of the number values. In English we can distinguish only between singular and plural. Many languages have more distinctions, and it is these on which we shall concentrate.

5.1 *The Dual*

A common system is one in which singular and plural are supplemented by a dual, to refer to two distinct

Table 1
Emphatic pronouns in Sursurunga

	singular	*dual*	'*trial*'	'*quadral*'	*plural*
1 exclusive	iau	giur	gimtul	gimhat	gim
1 inclusive	—	gitar	gittul	githat	git
2	iáu	gaur	gamtul	gamhat	gam
3	-i/on/ái	diar	ditul	dihat	di

Source: Hutchisson 1986 and personal communications
Note: á is used to indicate schwa (ə)

real world entities. Examples are widespread. We will take Central Alaskan Yup'ik, as spoken in the southwestern Alaska (Corbett and Mithun 1996); the number forms are: singular *qayaq* 'kayak,' dual *qayak* 'two kayaks,' and plural *qayat* 'three or more kayaks.' The addition of the dual has an effect on the plural, which is now used for three or more real world entities, and more generally a change in system gives the plural a different meaning

5.2 The Trial

While the dual refers to two distinct real world entities, the trial is for three, and it occurs in systems with the number values: singular, dual, trial, and plural. This system is found in Larike, a Central Moluccan language spoken on the western tip of Ambon Island, Central Maluku, Indonesia. (Central Moluccan forms part of the Central Malayo–Polynesian subgroup of Austronesian.) This example is from Laidig and Laidig (1990, p. 7):

(2) *Kalu iridu-ta-ʔeu au-na-wela*
if 2.TRIAL-NEG-go 1.SG-IRR-go.home
'If you three don't want to go, I'm going home.'

In this language the dual and trial forms can be traced back to the numerals 'two' and 'three,' and the plural comes historically from 'four.' Such developments are fairly common in Austronesian languages. Sometimes the term 'trial' is used according to the form of the inflections (derived historically from the numeral three), even when the forms are used currently for small groups including those greater than three (and so are paucals) and sometimes the term is used according to meaning (for genuine trials). The Larike trial is a genuine trial, used for exactly three (Laidig and Laidig 1990, p. 92).

5.3 The Paucal

Paucals are used to refer to a small number of distinct real world entities, somewhat akin to the English quantifier 'a few' in meaning. There is no specific upper bound on the use of paucals (their lower bound will vary according to the system in which they are found). Let us return to the Cushitic language Bayso. Besides the general number forms (as in (1), where number is not specified), there are also singular, paucal, and plural. Here is the paucal:

(3) Bayso (Dick Hayward, personal communications, Corbett and Hayward 1987)
luban-jaa foofe
lion-PAUCAL watched.1.SG
'I watched a few lions'

The paucal is used in Bayso for reference to a small number of individuals, from two to about six. (Bayso has this system in nouns but not in pronouns.) The Bayso system is rare; normally the paucal is found together with a dual, so that the number values are singular, dual, paucal, and plural. This widespread system is found, for example, in Yimas, a Lower Sepik language spoken in the Sepik Basin of Papua New Guinea (Foley 1991, p. 216).

5.4 The Largest Systems

It is worth asking what is the largest system of number values. There are claims in the literature for languages with a quadral, for four distinct real world entities, thus giving a system of five number values. It is established that five number values can indeed be found, but the evidence specifically for a quadral is less clear. All the claims come from within the Austronesian family and the best documented case is Sursurunga (Hutchisson 1986, and personal communications), spoken in southern New Ireland. The forms labeled quadral are restricted to the personal pronouns, but are found with all of them, the first person (inclusive and exclusive), the second, and third (Table 1).

Here is an example of a quadral form in use (we will retain the traditional label 'quadral,' until we have justified replacing the term for Sursurunga):

(4) Sursurunga (Hutchisson 1986 and personal communications)
gim-hat káwán
1.EXCL-QUADRAL maternal.uncle:nephew/niece
'we four who are in an uncle-nephew/niece relationship'

The 'quadral' has two additional uses. First, in Sursurunga plural pronouns are never used with terms for dyads (kinship pairings like uncle-nephew/niece). Here the 'quadral' is used, for a minimum of four, and

not just for exactly four (Hutchisson 1986, p. 10). Second, in hortatory discourse, the speaker may use the first person inclusive 'quadral' in order to suggest joint action including the speaker, even though more than four persons are involved. These two special uses of the 'quadral' account for the majority of instances of its use. According to its meaning, the term 'quadral' is not appropriate. Rather the forms would be better designated 'paucal.'

Let us now consider the rest of the system of Sursurunga (data and judgments from Don Hutchisson, personal communications). The dual is used quite strictly for two people. It is also used for the singular when the referent is in a taboo relationship to the speaker. This is a special use; the dual's main use is as a regular dual. The trial will be used for three. However, it is also used for small groups, typically around three or four, and for nuclear families of any size. It is therefore not strictly a trial, rather it could be labeled a paucal (an appropriate gloss would be 'a few'). We noted earlier that this is a frequent development for the trial. The 'quadral,' as we have just seen, is used primarily in different ways with larger groups, of four or more (an appropriate gloss would be 'several'). This too would qualify as a paucal; we therefore have two paucals in Sursurunga, a paucal (traditionally 'trial') and a greater paucal (traditionally 'quadral'). The plural, as we would expect, is for numbers of entities larger than what is covered by the greater paucal, however, there is no strict dividing line (certainly not at the number five). Using semantic labels, we might represent the system in Sursurunga as having: singular, dual, paucal, greater paucal, and plural. Sursurunga has a well-documented five-value number category. There are certainly other languages with five number values; we do not have such detailed information as for Sursurunga and there is no certain case of a quadral: it seems that in all cases the highest value other than plural in such systems can be used as a paucal. Besides the split in the paucal, we may also find a split in the plural, with 'greater plurals' of different types, as for instance in Mokilese (Harrison 1976). Greater plurals may imply an excessive number or else all possible instances of the referent. They are as yet poorly researched, except in a few languages.

6. *Conclusion*

Number is indeed a much more complex category than is generally recognized, and the English system, though a common one, is at one extreme of the typological space. We noted that number may be a verbal category as well as a nominal one (but the possibilities are rather more restricted for verbal number). We also saw that number need not be obligatory, that the range of items which distinguish number varies dramatically from language to language, that the semantics of number are affected by the particular item involved, and that the values of the number category can be just two or, in various combinations, up to five. More complex issues, which we have not been able to address here, include various interactions, notably between the forms of marking of number, the particular nominals which distinguish number in a given language, and the systems of number values. For these issues see Corbett (2000).

See also: Grammatical Gender; Number Systems, Evolution of; Pronouns

Bibliography

Aikhenvald A 1995 *Bare* (Languages of the World/Materials 100). Lincom Europa, Munich, Germany

Corbett G G 2000 *Number*. Cambridge University Press, Cambridge, UK

Corbett G G, Hayward R J 1987 Gender and number in Bayso. *Lingua* **73**: 1–28

Corbett G G, Mithun M 1996 Associative forms in a typology of number systems: Evidence from Yup'ik. *Journal of Linguistics* **32**: 1–17

Durie M 1986 The grammaticization of number as a verbal category. In: Vassiliki N, VanClay M, Niepokuj M, Feder D (eds.) *Proc. of the Twelfth Annual Meeting of the Berkeley Linguistics Society: February 15–17, 1986*. University of California, Berkeley, CA, pp. 355–70

Foley W A 1991 *The Yimas Language of New Guinea*. Stanford University Press, Stanford, CA

Harrison S P, Albert S Y 1976 *Mokilese Reference Grammar*. University Press of Hawaii, Honolulu, HI

Hutchisson D 1986 Sursurunga pronouns and the special uses of quadral number. In: Wiesemann U (ed.) *Pronominal Systems (Continuum 5)*. Narr, Tübingen, Germany, pp. 217–55

Laidig W D, Laidig C J 1990 Larike pronouns: Duals and trials in a Central Moluccan language. *Oceanic Linguistics* **29**: 87–109

Link G 1998 Ten years of research on plurals—where do we stand? In: Hamm F, Hinrichs E (eds.) *Plurality and Quantification*. Kluwer, Dordrecht, The Netherlands, pp. 19–54

Rice K 1989 *A Grammer of Slave*. Mouton de Gruyter, Berlin

Smith-Stark T C 1974 The plurality split. In: La Galy M W, Fox R A, Bruck A (eds.) *Papers from the Tenth Regional Meeting, Chicago Linguistic Society, April 19–21, 1974*. CLS, Chicago, pp. 657–71

G. G. Corbett

Grammatical Relations

The linguistic concept 'grammatical relation' with its equivalents 'syntactic relation,' 'syntactic function' belongs to the most important concepts of traditional syntactic theory. By grammatical relations are understood structurally expressed characteristic syntactic relations between sentence constituents (see *Constituent Structure: Syntactic*), in particular between the

predicate and noun phrases (NPs). Thus, in the sentences (1) *Peter sent a message to Mary*, (2) *My friend sent a congratulatory telegram to his boss*, in addition to the predicate *sent* one can identify three NPs linked to the predicate by specific grammatical relations. One of the NPs (*Peter/my friend*) is the subject, the second (*a message/a congratulatory telegram*) is the direct object, while the third (*Mary/his boss*) is the indirect object. (Noun phrases that realize the same grammatical relation occupy the like-named position in the clause. Thus, the NP *Peter* in (1) occupies the subject position.) Each of the NPs in these sentences is structurally distinct from the other two and makes unique reference to a concrete participant in the events denoted by the given sentences. The subject and the direct object of this particular example are expressed positionally—the subject precedes the verb, the direct object follows the verb immediately—while the indirect object contains the preposition *to*.

Grammatical relations fulfill *distinguishing* and *characterizing* functions. First, they permit identification of the various participants in the event (what was sent, to whom and by whom it was sent), secondly they classify the NPs in the given sentence in terms of a restricted set of characteristic properties, allowing the identification of NPs in different sentences. Thus, in (3) *A message was sent to Mary (by Peter)*, (4) *Mary was sent a message (by Peter)*, both semantically related to sentence (1), the subjects will be, respectively, the noun phrases *a message* and *Mary*; in (5) *That he won is fine*, there is a clausal subject *that he won*; while in (6) *It is fine that he won* and (7) *It is dark*, the subject is the pronoun *it* (the so-called dummy subject). The concept 'subject' as a homogeneous syntactic relation unites different kinds of specific constituents. Each grammatical relation has a specific function (see *Grammar: Functional Approaches*) and specific linguistic means for its expression.

1. *What Do Grammatical Relations Mean?*

Despite the widespread use and obvious necessity of the terms subject and object, traditional grammar did not succeed in defining their meaning adequately. Definitions for 'subject' of the type 'what the sentence is about' are at first sight suitable for sentences (1)–(4), but they are dubious for (5), while the subjects in (6) and (7) clearly do not fall under a definition of this type.

In a series of investigations of the last decades (cf., for instance, the pioneering work of Keenan 1976) it has been established that in the most elementary sentences subject NPs correlate statistically with the meanings of various autonomous semantic domains.

First, subjects are linked to the semantic role properties of participants in the given situation. The subject usually encodes the agent, if there is one in the sentence. If there is no agent in the sentence, then the prime candidate for subject position is the patient. In the presence of an agent the patient is the prime candidate for the position of direct object. Sentences with an agent and patient constitute the prototypical transitive clause with a Subject and direct object. Sentences with participants in other roles are assimilated to varying extents to the structure of transitive or intransitive clauses, or use the position oblique object. The participant with the role of recipient/beneficiary usually occupies the position of indirect object.

Second, there are statistical correlations with the discourse (communicative) domain (see *Linguistics: Information Structure*). Thus, subjects often communicate known (given) information, while objects communicate new information; subjects often coincide with the focus of empathy, i.e., with that participant from whose point of view the situation is described, and with the topic (theme), while objects often coincide with the comment (rheme).

Third, the ontological domain, connected with the properties of referents, influences the subject or object position of NPs. Nps denoting animate, and especially human participants tend towards the subject position, inanimates towards the object position (*Grammatical Gender*). Within the set of humans the greatest tendency towards subject position is shown by the speech act participants (concepts from the deictic domain, see *Deixis*).

In addition, concepts from the referential domain, which classify kinds of referents denoted by NPs and facilitate their correct identification by the addressee, also correlate with grammatical relations. Thus, the subject has a presumption of existence, usually has autonomous reference, and is definite and specific, while the direct object is often nonspecific.

However, it is not possible to assign any of the above-mentioned meanings as obligatory for the subject or direct object; in each given sentence these grammatical relations satisfy only some of these meanings.

What, then, is the invariant meaning of grammatical relations, if indeed one exists? There is no generally accepted answer to this question. However, the following can be offered as the most plausible answer. Grammatical relations order NPs in the sentence by rank in accordance with their overall communicative value in a concrete utterance. The ordering takes place by means of a system of rules which takes into account the concrete meanings of NPs according to all the relevant semantic domains, so that the first rank is assigned to the subject, the second to the direct object, then the indirect object and oblique objects, as shown in the so-called 'relational hierarchy':

$$\text{Su} > \text{DO} > \text{IO} > \text{OO}$$

where Su represents 'subject,' DO 'direct object,' IO 'indirect object,' and OO 'oblique object.' The subject

occupies a special position in this hierarchy. The subject is the most essential component of the sentence, and in some languages it is structurally necessary, in contrast to all other grammatical relations. The other grammatical relations are ranked primarily according to their degree of difference from the subject. In this way, grammatical relations have not an absolute, but rather a relative syntactic value in the sentence, and the relation of this value to meanings of various kinds is indirect in nature.

Grammatical relations are a means of cumulative expression of permitted combinations of more elementary meanings, primarily from the role and discourse domains. Given this, then, first, the same grammatical relation in different sentences of the same language can have a different set of meanings; and second, in different languages grammatical relations can be organized differently, since different combinations of these meanings may be grammaticalized (see *Grammaticalization*) into different grammatical relations.

It can easily be seen that the subject and direct object are furthest removed from the semantic role of a NP and that their basic characteristic is their rank position in the hierarchy. Together with the predicate they form the nucleus of the predicate–argument structure. Oblique objects, in contrast to subject and direct object, have a specific semantic role (e.g., place, time, cause, manner), and in a clause there can be more than one oblique object. The indirect object has an intermediate status. On the one hand, it is considerably more loaded, in terms of semantic role and ontological semantics, than are the subject and direct object, and in this it is like oblique objects, but on the other hand it is the easiest to incorporate into the nucleus, compare (1) and (8) *Peter sent Mary a message*. In (8), in contrast to (1), *Mary* occupies the position of direct object.

2. *How are Grammatical Relations Encoded?*

Languages use different means to encode grammatical relations. The best known morphological category specifically designated to encode grammatical relations is case (see *Morphological Case in Linguistics*). Case marking is present in many languages, e.g., in Latin and Greek, in the Slavic and Dravidian languages, and in the languages of Australia. Under case marking the grammatical relation is expressed by the morphological form of the head of the NP (i.e., of the noun, as in Latin and Russian) or by the forms of other structural constituents of the NP (e.g., forms of the article and noun in German). Alongside inflectional morphology many languages make use of grammatical words or particles (usually prepositions or postpositions), which combine with a NP to give analytical constructions that indicate the kind of grammatical relation (e.g., the combination of preposition *to* and NP, such as *to Mary* in (1), encodes the indirect object). The linear order of NPs in the sentence is the positional equivalent of inflectional or analytical marking (see *Word Order*) and is calculated according to the position of the NP relative to the verb, relative to the beginning or end of the sentence, or relative to other NPs. Morphological marking may be indicated not on NPs, but on the verb by means of agreement (see *Grammatical Agreement*): the form of the verb expresses information about the grammatical properties of that NP which stands in a particular grammatical relation to the verb. An example in English is agreement of the verb with a third-person subject in number, combining with the indication of grammatical relations through linear order. Some languages have polypersonal agreement of the verb with two or more NPs. Thus, in Abkhaz the verb has three agreement positions indicating agreement with specific syntactic types of NPs.

3. *The Syntactic Functions of Grammatical Relations*

The question of which grammatical relation a particular NP occupies in the sentence can by no means always be decided on the basis of the coding technique. The same means may be used to encode different grammatical relations and the same grammatical relation may use different coding techniques. Thus, in English an immediately postverbal prepositionless NP is not always a direct object: (9a) *He was sleeping all night*, cf. the equivalent Russian sentence with a time adverbial in the accusative: (9b) *On spal vsju noč′*, where the accusative is the typical case of the direct object. Conversely, in Russian the direct object can stand in the genitive case in the context of negation, e.g., (10a) *On kupil knig-u* (acc.) 'He bought a book' vs. (10b) *On ne kupil knig-i* (gen.) 'He didn't buy a book.' For this reason the question often arises of the criteria by which one can assign different grammatical relations to identically marked constituents or assign the same grammatical relation to differently marked constituents.

The decisive diagnostic role in identifying grammatical relations is played not by coding means, but by syntactic tests. The point is that the coding technique is merely the outer packaging that helps identify grammatical relations in the sentence, whereas the grammatical relations are really components of the syntactic structure of the sentence; their identity or nonidentity is determined in the last analysis on the syntactic level (see *Syntax*). In many languages the system of rules of syntactic organization operates to a considerable extent on the basis of units characterized in terms of grammatical relations, and thus the identity of syntactic behavior of constituents differing in form is an important criterion for uniting them as a single grammatical relation.

Within the clause, voice transformations are oriented to grammatical relations (see *Grammatical Voice*). Thus, the difference between active and passive voices is that different grammatical relations adhere to the same NPs in active and passive versions. Passivization is usually characterized as the promotion of the syntactic status of the direct object to that of subject and/or demotion of the syntactic status of the subject. In English both processes occur in passivization, cf. (1) and (2). The subject NP of the active loses its subject status (it is expressed by the prepositional phrase *by Peter*), while the direct object NP (*a message*) becomes subject (promotional passive). There are also languages in which passivization involves only the demotion of the status of the subject NP (demotional passive). However, in all cases passivization is defined in terms of grammatical relations, irrespective of the coding techniques used in the particular language.

Among nearly universal properties are constraints on reflexivization, operating in cases of coreference between two NPs in the same clause, e.g., (11) *Mary hates herself*, (12) *Mary does not think about herself*. Here, one of the NPs (*Mary*) must be in subject position and is a full identification of the referent (the controller of reflexivization). The other, nonsubject, coreferential NP is replaced by the reflexive pronoun (the target of reflexivization).

Under coordination there are also constraints on grammatical relations. Thus, in English coordination reduction is possible only with coreferential subjects. The second coreferential subject is omitted, e.g., (13) *John washed his hands and—[= John] began to eat*. Another construction formulated in terms of subjects is switch reference, attested in various languages: clauses are expressed differently according to whether their subjects coincide or not (usually by marking the verb of one of the clauses).

As a syntactic test one can use constructions with sentential complements with verbs like 'want,' 'must,' 'begin.' Sentential complements are usually expressed differently depending on the coreference versus noncoreference of the subject of the sentential complement and the subject of the main clause, cf. (14a) *Peter wants to congratulate Mary*, but (14b) *Peter wants Mary to congratulate him*. In (14a) the subject of the sentential complement is coreferential with the subject of the verb *want*, and is therefore not expressed overtly in the subordinate clause. In (14b) the coreferential NP is the direct object, therefore it cannot be omitted, but must be expressed by means of a pronoun.

The formation of relative clauses (see *Relative Clauses*) is constrained, in many languages, by the grammatical relation of the target of relativization. For instance, reduced (participial) relative clauses in English are restricted to subject targets of relatives: (15a) *the student buying a book*, (15b) *the book bought by the student*. In (15a) the target of relativization is in subject position (*The student buys a book*), therefore the relative clause is formed directly from this structure. In (15b) the target of relativization is the direct object (*The student bought the book*), therefore in order to construct the participial version it is necessary first to passivize the relative clause, promoting the status of the target of relativization to subject position (*The book was bought by the student*). In some languages relativization is constrained also by the grammatical relation of the controller of relativization.

Not infrequently constraints on the formation of adverbial clauses (see *Adverbial Clauses*) also rely on grammatical relations. In Russian there is a set of subordinate clauses expressing background activity, the so-called converbal (gerundive) clauses. The subject of the converbal clause must be coreferential with that of the main clause (and is omitted in the converbal clause): (16) *Pokupaja ètu knigu*, *student potratil na poiski massu vremeni* 'Buying this book, the student spent a lot of time in searching.' In (16) the omitted subject of the converbal clause and the subject of the main clause is the NP *student*. Sentence (17) is ungrammatical: (17) **Pokupaja ètu knigu*, *u studenta propal košelek* 'Buying this book, the student's purse got lost,' since *u studenta*, the NP in the main clause coreferential with the subject of the converbal clause, is not the subject (the subject of the main clause is *košelek* 'purse').

To the extent that many syntactic rules are operations in terms of grammatical relations, these rules can be used as tests to determine the syntactic status of a NP in morphologically ambivalent instances.

4. *Initial vs. Derived* (*Primary vs. Secondary*) *Grammatical Relations*

Certain syntactic processes change the grammatical relation of a particular NP or delete it. Such processes change the valency of a sentence (see *Valency and Argument Structure in Syntax*). In such instances it is important to distinguish initial and derived grammatical relations of NPs. Thus, in the passive the initial subject loses this grammatical relation, while the initial direct object becomes the subject of the passive sentence. It follows from this that syntactic processes are often formulated in terms of initial syntactic relations which are not present in the observable, derived syntactic structure. Their existence can nonetheless be reconstructed on the basis of the syntactic process that conditions the observed structure. In a series of instances properties of the initial grammatical relation of a NP are partially retained. For instance, in the examples (18a) *Uncle Tom built this hut for himself* and (18b) *This hut was built by Uncle Tom for himself*, the controller of the reflexive is in both instances the initial subject *Uncle Tom*, even

though in (18b) the derived, observable subject is *this hut*.

5. Nonuniversality of Grammatical Relations

In linguistics the aprioristic assumption of the universality of grammatical relations was long dominant. Many current syntactic theories derive from this assumption, in particular generative grammar (see *Linguistics: Incorporation*) and the theory of Relational Grammar. In the last decades, however, as a result of considerable progress in studying the nature of grammatical relations and thanks to the considerable broadening of empirical linguistic material this assumption is no longer self-apparent.

First of all, the functional and formal properties of grammatical relations characterized above (polyfunctionality of content and cumulative expression) testify to their specific semiotic nature. Why should such a technique of identification and characterization of noun phrases be an obligatory property of natural language? Conversely, it is natural to allow that different languages might grammaticalize various concepts of different semantic domains and package them differently in linguistic form. What we identify as grammatical relations might be filled with different content in different languages and, moreover, one might find not a cumulative, but a separative technique of packaging elementary concepts. Finally, there is no *a priori* basis to assume that the concepts of some or even all the relevant domains should be grammaticalized in each language, given that so far not a single semantic domain has been found that would be grammaticalized in all languages (be it the category of gender, number, tense, aspect, mood etc., see *Grammatical Gender*; *Grammatical Number*; *Tense, Aspect, and Mood, Linguistics of*). Even as apparently as indispensable ontological category as part of speech (see *Word Classes and Parts of Speech*) has recently increasingly lost the assumption of universality.

Second, uncritical attempts to use relational concepts with respect to different languages often encounter serious empirical difficulties. One may start with the observation that the morphosyntactic means of encoding NPs in some languages is not only different from, but sometimes not even comparable with the encoding characteristic of Standard Average European languages. If one uses the widespread notation for the nuclear arguments with the symbols S, A, P (or O), where S is the nuclear argument of an intransitive clause, and A and P the agent-like and patient-like arguments, respectively, of a transitive clause, then in European languages the S/A arguments have identical encoding, grouping them together as the grammatical relation of subject, in opposition to the encoding of the P argument. Such a system is called accusative alignment. Alongside this possibility, many languages found in various continents have the ergative pattern, uniting the S/P arguments in opposition to the A argument. Here the S/P arguments are usually morphologically unmarked, while the A argument is marked (ergative case or its analog). The ergative alignment of the nuclear arguments creates an empirical problem for the establishment of grammatical relations.

In the languages of the world other patterns are also found. One possibility is a pattern that distinguishes agent-like and patient-like arguments of both transitive and intransitive clauses (active alignment). Under active alignment one finds different expression of the single argument of verbs like 'run' and verbs like 'fall.' There is even the tripartite scheme, distinguishing all three arguments, S, A, and P. Finally, there is the neutral pattern, which fails to differentiate the S/A/P arguments by means of coding.

Languages characterized by 'unusual' nonaccusative patterns are quite varied from a syntactic point of view. For instance, languages with ergative alignment may behave as syntactically accusative, i.e., in defiance of the coding technique S and A arguments are united in a single grammatical relation subject. Much more interesting is the situation where ergative alignment is also realized at the syntactic level: S and P arguments are united syntactically in the highest ranking grammatical relation, equivalent to the subject under accusative alignment. The A argument is also a nuclear argument, i.e., it occupies the second position in the Relational Hierarchy. Finally, there are ergative languages in which it is impossible to rank S, A, and P at the syntactic level. In languages of this type it is justifiable to speak of the absence of grammatical relations.

Among languages with active alignment are also found languages behaving accusatively at the syntactic level, as well as languages lacking syntactic relations. So far there is extremely little information available on languages with tripartite or neutral alignment. It is, however, quite clear that they too are quite varied. They may have or not have grammaticalized grammatical relations. In particular, there is evidence for the existence of languages with neutral alignment lacking grammatical relations. But one can assume the existence of a neutral alignment language opposing nuclear and nonnuclear arguments. In such a language the nuclear arguments would be equivalent to subject, and the nonnuclear arguments to oblique objects. If such a language is attested, then the thesis of the uniqueness of the subject will have been overturned.

Thus, grammatical relations both have a narrower sphere of use and are considerably more varied than is usually assumed.

It should be kept in mind that the thesis of the nonuniversality of syntactic relations is not universally held, and is even ignored in many works. However, the assumption of nonuniversality is preferable by far in empirical research of little-studied languages, since it does not impose an aprioristic standard on the description and helps to see the real state of affairs in each language under discussion.

6. Grammatical Relations and the Functional Approach to Language

The concept of grammatical relation has been known in linguistics since the time of the grammarians of antiquity. However, significant progress in understanding the nature of grammatical relations has been achieved only in the last decades thanks to the orientation towards the functional explanation of grammar (see *Grammar: Functional Approaches*), which has permitted establishing the polyfunctional character of the elementary categories which form the clusters known as grammatical relations. The majority of these categories were themselves discovered and investigated in the framework of the functional approach.

In this connection the question of the existence of different patterns of encoding S, A, and P arguments merits special attention. Proceeding from the now widely established assumption of the iconicity of the linguistic sign (see *Linguistics: Iconicity*), replacing the Saussurean assumption of the arbitrariness of the sign, one can allow for the existence of a functional motivation for the various coding patterns. These patterns unite arguments with different elementary roles in different ways. The basic prototype of all these groupings (see *Linguistics: Prototype Theory*) is given by the agent and patient of the transitive construction. From this viewpoint active alignment is a natural metonymic extension of the elementary roles of agent and patient, generalizing them to the hyperroles (macroroles) of actor (the participant in the event which performs, effects, instigates, or controls it) and undergoer (the participant in the event which does not perform, effect, instigate, or control it, but rather is affected by it in some way). Arguments with different elementary roles can be interpreted in terms of these hyperroles.

Accusative alignment is also motivated essentially not by grammatical relations, but by the hyperroles principal and patientive. The principal indicates the main participant, the 'hero' of the situation, who is primarily responsible for the fact that this situation takes place. In the transitive clause the principal is, of course, the agent-like argument, while in the intransitive clause the single nuclear argument has no competitors for the role of principal. The patientive is the most patient-like participant of a multi-participant event. This role unites, in addition to the patient, arguments of other types.

Ergative alignment articulates the S, A, P arguments by means of other hyperroles, absolutive and agentive. The absolutive is the metonymic extension of the patient to intransitive verbs, it is the immediate, nearest, most involved or affected participant in the situation. In the transitive clause the patient is most involved in the situation. In the intransitive clause there is no competition for the role of closest and most involved participant in the situation. The absolutive in the transitive clause is in opposition to the agentive, the most agent-like participant of the multi-participant event.

Tripartite alignment distinguishes agentive and patientive of the transitive clause, distinguishing them from the single nuclear argument of the intransitive clause.

All these means of generalizing the elementary roles are in principle motivated, but a language has the option of grammaticalizing just one of them as its basic pattern. In addition, a language may have different patterns in different contexts, and examples of this kind have been attested.

7. Grammatical Relations and the Cognitive Approach to Language

The 'statistical' correlations between concepts of different semantic domains noted above and which are grammaticalized in grammatical relations and the means of their encoding are not chance, rather they have a cognitive nature, reflecting the human capacity for categorizing reality. Behind the limitless variety of whole grammatical systems lie the same stable cognitive structures of human thought and the pragmatic strategies (see *Pragmatics: Linguistic*) for their verbalization. The reconstruction of these structures and strategies on the basis of data from different languages is the task of cognitive typology (see *Linguistic Typology*).

Bibliography

Aissen J 1999 Markedness and subject choice in optimality theory. *Natural Language and Linguistic Theory* **17**: 673–711

Cole P, Sadock J M (eds.) 1977 Grammatical relations. *Syntax and Semantics*. Academic Press, New York, Vol. 8

Croft W 1991 *Syntactic Categories and Grammatical Relations*. The University of Chicago Press, Chicago

Dik S 1997 *The Theory of Functional Grammar*. Mouton de Gruyter, Berlin

Dixon R M W 1994 *Ergativity*. Cambridge University Press, Cambridge, UK

Dixon R M W, Aikhenvald A Y (eds.) 2000 *Changing Valency: Case Studies in Transitivity*. Cambridge University Press, Cambridge, UK

Dryer M 1986 Primary objects, secondary objects, and antidative. *Language* **62**: 808–45

Dryer M 1997 Are grammatical relations universal? In: Bybee J, Haiman J, Thompson S (eds.) *Studies in Language Function and Language Type*. Benjamins, Amsterdam, pp. 115–43

Givon T (ed.) 1997 *Grammatical Relations: A Functional Perspective*. Benjamins, Amsterdam

Keenan E 1976 Towards a universal definition of "subject". In: Li Ch N (ed.) *Subject and Topic*. Academic, New York, pp. 303–34

Kibrik A E 1997 Beyond subject and object: Toward a comprehensive relational typology. *Linguistic Typology* **1**: 279–346

Li Ch N (ed.) 1976 *Subject and Topic*. Academic, New York
Perlmutter D M (ed.) 1983 *Studies in Relational Grammar*. The University of Chicago Press, Chicago
Plank F (ed.) 1979 *Ergativity: Toward a Theory of Grammatical Relations*. Academic, London
Van Valin R, LaPolla R J 1997 *Syntax: Structure, Meaning, and Function*. Cambridge University Press, Cambridge, UK

A. E. Kibrik

Grammatical Voice

1. Introduction

A typical sentence (clause) of a natural language consists of the predicate and its arguments. Each argument is associated with a certain participant role and is expressed by a certain grammatical function in a clause. Voice is the grammatical category that relates the participant roles associated with arguments on the one hand and the grammatical functions of the noun phrases (NPs) expressing these arguments on the other.

Predicates can be characterized by the number of arguments they take; a one-place predicate is intransitive, a predicate with two arguments is transitive. The ability of a predicate to subcategorize for semantic arguments does not necessarily mean that all these arguments are overtly expressed by NPs; some arguments may be omitted and some arguments may be understood (implicit) but never expressed.

In terms of semantic arguments, the description of grammatical voices relies crucially on the recognition of two types of core arguments of a predicate, agent (logical subject) and theme (logical object). An agent is a participant that has some or all of the following properties: it exists independently of the event named by the predicate, can cause a change of state in another participant, is capable of movement, is volitional and capable of intention (Gruber 1965, Fillmore 1968, Dowty 1991). A theme is a participant that has some or all of the following properties: it may not exist independently of the event, undergoes a change of state in the event, is causally affected by an agent.

Turning now to the grammatical functions subject and object, these functions can be manifested in all or some of the following formal properties: case-marking, linear position in a sentence, agreement, and ability to determine the form and/or linear position of other constituents (for example, the form of reflexives).

Based on how agent and theme are mapped into grammatical functions, linguistic theory recognizes the following main voices: active, passive, middle, and antipassive.

In the active voice, agent is mapped into the syntactic subject of a clause, while theme is mapped into the syntactic object, as in the English example (1).

(1) *Invisible God* [agent, subject]
created a visible world [theme, direct object]

In the passive voice, agent is mapped into an oblique phrase or adjunct, whereas theme may be mapped into the syntactic subject, as in (2).

(2) *The visible world* [theme, subject]
was created by invisible God [agent, adjunct]

In the middle voice, the event denoted by the predicate is construed as developing without external causation; the main participant is both the source and the recipient (beneficiary) of the event. Thus, the main argument of the middle voice is construed as combining features of agent and theme. The contrast between the middle voice and the passive voice can be illustrated by the English (3) and (4); in (3), the ability to wash well is perceived as the property of cotton, and no external agent is permitted:

(3) *Cotton* [theme, subject]
washes well (**by consumers*)
(4) *Cotton* [theme, subject]
is washed (*well*) (*by consumers*)

In the antipassive voice, agent is mapped into subject and theme is mapped into an adjunct. Although English does not have a regular antipassive, the contrast between (5) and (6) is illustrative of the contrast between the active voice (theme = direct object) and antipassive (theme = adjunct):

(5) *The soldiers* [agent, subject]
shot the pirates [theme, object]
(6) *The soldiers* [agent, subject]
shot at the pirates [theme, adjunct]

Alternations in grammatical voice that operate on the same argument structure consistently show information-structural differences. The major distinction in information structure is between topic ('what the utterance is about') and focus (the constituent that, given a well-formed question-answer pair, is the part of the answer that corresponds to the wh-word in the question). Topic-focus differences can be used to motivate the choice of one voice over another. It is erroneous however, to think that each and every language has voice alternations; for a given argument structure, some languages have just one way of mapping arguments to grammatical functions.

2. Active Voice

A verb in the active is often unmarked for any morphological indication of voice; however, this is not a universal characteristic. A number of languages, for instance, many Austronesian languages, mark active voice on the verb with a special morpheme. For

example, in Malagasy, a significant number of verbs form the active by prefixation on the verb stem (7a); for some of these verbs, the passive form is morphologically simplex (7b) (Keenan and Polinsky 1998):

(7a) *sasa-* 'wash' (root)—*manasa* 'wash' (active, present)
(7b) *hita* 'see' (passive)—*mihita* 'see' (active, present)

A common active voice structure is the one where the agent or most agent-like argument of the predicate is subject, while the theme, if there is one, is direct object. Agent appears in the subject case which, depending on a particular language, could be nominative, as in (8a, b), (11), absolutive (9a), ergative (9b), or dative (10). If the verb is transitive and has a theme argument, the case marking of this argument is typically that of direct object (accusative in an accusative language, absolutive in an ergative language); however, theme objects of some verbs can also be marked by a special case, inherent to a given verb, e.g., as in Russian (11).

(8) Latin
(a) *rex* *dēcessit*
king.NOMINATIVE died
(b) *rex* *omnes*
king.NOMINATIVE (all
cōpias *castrīs* *ēducit*
troops).ACCUSATIVE of.camp led
'The king led all the troops out of the camp.'

(9) Chukchii
(a) *ə tlʔa* *wiri-gʔi*
mother.ABSOLUTIVE descend-AORIST:3sg
'The mother came down.'
(b) *ə tlʔa-ta* *mə čə kw-ən*
mother-ERGATIVE shirt-ABSOLUTIVE
tə ni-nin
sew-AORIST:3SG:3SG
'The mother sewed the shirt.'

(10) Hindi
muj^{h}-e *tum-se* *pyaar* *hɛ*
I-DATIVE you-INSTRUMENTAL love AUXILIARY
'I love you.'

(11) Russian
Korol' *upravljaet*
king. NOMINATIVE rules
stran-oj
country-INSTRUMENTAL
'The king rules the country.'

The determination of grammatical functions associated with agent and theme depends on language-internal factors and cannot be done on the basis of isolated examples from a given language.

If the verb in the active voice is intransitive, its single argument may correspond to either topic or focus; focused subjects of intransitives often serve to introduce new information. In a transitive active sentence, the subject/agent argument is more typically associated with topic than the object/theme, as shown by a number of text counts (Mallinson and Blake 1981). However, this information-structural preference can be overridden by a number of contextual factors, and an argument other than agent or even an adjunct can be construed as topic.

3. Passive Voice

Natural languages use passives extensively and show a wide range of variation in the expression of passives. The variation can be captured by the following parameters: (a) difference in overt marking between the active and the passive predicate; (b) expression of the subject of the passive; (c) accessibility of arguments to passivization; (d) expression of the passive agent.

In terms of the encoding of the predicate in the passive construction, passives can be morphological (synthetic) or periphrastic (analytical). In English, the formation of the passive always requires an auxiliary ('be' or 'get,' with slight semantic differences depending on the choice of the auxiliary), thus the passive is periphrastic. In some languages, passives are always formed morphologically, for example, in Swahili, where the passive is marked by the suffix -*w*-(12b):

(12) Swahili
(a) *Juma* *a-li-vunj-a* *dirisha*
Juma he-PAST-break-ASPECT window
'Juma broke the window.'
(b) *dirisha* *li-li-vunj-***w**-*a* (*na Juma*)
windowit-PAST-break-PASSIVE-ASPECT by Juma
'The window was broken (by Juma)'

Although it is common for passives to be morphologically more complex than actives, this is not always the case; in a number of Austronesian languages, morphological passives are simpler than actives, for instance, for subsets of verbs in Toba Batak, Indonesian, or Malagasy; it is also possible for both active and passive to be formed by affixation from a verbal root (stem), as in Malagasy.

Some languages combine morphological and periphrastic passives, depending on the verb categories. For instance, in Latin or Russian, imperfective verbs form morphological passives, and perfectives use periphrastic passives (Latin *capio* 'take' (active), *capitur* 'is taken' (passive imperfect), *captus est* 'was/has been taken' (passive perfect). In addition to passives proper, some languages use the so-called reflexive passive, in which the active form of the verb combines with a reflexive clitic such as Romance *se*, *si*, German *sich*, Slavic *–sja*. Finally, some languages do not use any verb marking to distinguish active and passive predicates; for example, in Acehnese and in Mandarin, the only overt marking of the passive construction comes from the marking of the agent phrase and, optionally, from word order (Perlmutter and Postal 1983).

In terms of argument expression, the mapping of agent into an adjunct is the necessary and sufficient characteristic of passive. It is therefore possible that agent is 'demoted' to an adjunct and the subject position is filled by an 'expletive' ('dummy')—a meaningless element similar to the English 'it' in 'It rains' or 'there' in 'There was a riot.' Passives whose subject is an expletive are known as impersonal passives. Impersonal passives can be formed from intransitive verbs, (13b), or transitive verbs (14b). In the latter case, theme remains in the object position, and the subject of the sentence is expressed by an expletive; in languages where pronominal subjects are typically deleted, the expletive of an impersonal passive is a 3-person null pronoun (*pro*), whose presence is signaled by the agreement features on the verb (14b).

13 German
(a) *Alle* *Kinder* *tanzen*
all children dance
(b) *Es* *wird* (*von* *allen* *Kindern*)
it is by all children
getanzt
danced.PERF.PARTICIPLE
'There is dancing (as performed by all children).'

(14) Lithuanian
(a) *čia* *lankutojai* *klausosi*
here visitors.NOMINATIVE listen
muzikos
music.ACCUSATIVE
'Visitors listen to music here.'
(b) *čia pro* *lankutoju* *klauso-ma-si*
here visitors.GENITIVE listen-PASSIVE-3
muzikos
music.ACCUSATIVE
'There occurs listening to the music by visitors here.'

In a personal passive, the subject position is occupied by a meaningful argument (as in English). In some languages, arguments other than direct object can also become subject of the passive. In some dialects of English, passivization is equally possible for both objects of the verbs 'give' or 'send':

(15a) *The doctor gave the child new medicine*
(15b) *New medicine was given the child* (*by the doctor*)
(15c) *The child was given new medicine* (*by the doctor*)

In Kinyarwanda (Bantu), any prepositionless object can become subject of a passive (Kimenyi 1988, p. 365), thus:

(16a) *umugóre aréerekeesha ímashiîni ábáana*
woman is showing machine children
amashusho
pictures
'The woman is showing pictures to the children with a machine.' (Active)

(16b) *amashusho are´erekeeshwa ímashiîni ábáana*
pictures are being shown machine children
(*n'umugóre*)
by woman
'Pictures are shown to the children with machine (by the woman).' (Theme passive)

(16c) *abáana baréerekeeshwa ímashiîni amashusho*
children are being shown machine pictures
(*n'umugóre*)
by woman
'The children are shown pictures with a machine (by the woman).' (Goal passive)

(16d) *ímashiîni iréerekeeshwa ábáana amashusho*
machine is being shown children pictures
(*n'umugóre*)
by woman
'The machine is being used to show pictures to the children (by the woman).' (Instrument passive)

In languages that allow the passivization of multiple arguments, the accessibility to passivization is determined by the following principle: if a lower (more oblique) object can be passivized, then all objects higher than it should be able to passivize as well. This predicts that there should be no languages where a passive as in (16d) is possible whereas passives as in (16b, c) are ungrammatical.

In terms of the expressions of the agent, no language seems to require that the passive agent be expressed; however, there are languages where an overt by-phrase in the passive is ungrammatical (Arabic, Latvian, Seri).

There are several functional motivations for the use of passives, and although they do not need to co-occur within a single language, these motivations are cross-linguistically consistent. In terms of information structure, passive allows a speaker to indicate that theme, not agent is the main topic of an utterance. Concomitantly, passive may serve to indicate that the passive agent is focus or part of focus. Syntactically, passive is needed if a language has rules that operate only on subjects; in such a case, passive is a 'way-station' for an argument that needs to be made accessible to a given rule. For example, English uses passive in embedded infinitival clauses to put theme in the position of the silent subject argument of the infinitive (PRO), as in (17b), where John is subjected to the interview:

(17a) *We persuaded John*$_i$ [PRO$_i$ *to interview the lawyer*]
(17b) *We persuaded John*$_i$ [PRO$_i$ *to be interviewed by the lawyer*]

Passive may differ from active in pragmatic connotation, for example, politeness or indirectness. Such pragmatic differences are often invoked as the main reason for the prevalence of passives in certain narrative styles and in requests (passive imperatives) in Austronesian languages. A recurrent interpretive

difference between the active and the passive is in that the passive emphasizes the change of state incurred by the referent of the theme. Because of this semantics of change of state and the attendant semantics of the resultant state, passive constructions are more common in the past and perfect; the coalescence of passive and perfect in the history of Indo-European languages is well documented (Benveniste 1960, Kuryłowicz 1964).

Linguistic theories differ in terms of whether the alternation between voices should be accounted for in syntax, in the lexicon or in both modules; for a good overview of grammatical approaches to passive, see Spencer (1991, Chap. 7).

4. *Middle Voice*

The description of the middle voice, based on the Greek grammatical tradition, is that of an event initiated within/by the main participant and having a principle effect on the same participant. For example, in Sanskrit:

(18) Sanskrit
(a) *so* *namati* *daṇḍam*
he.NOMINATIVE bends.ACTIVE stick.ACCUSATIVE
'He bends a stick.'
(b) *daṇḍaḥ* *namate*
stick.NOMINATIVE bend.MIDDLE
'The stick bends.'

A common extension of that meaning is that the volitional and sentient main participant is acting in his own interests. For example in Sanskrit:

(19) Sanskrit
(a) *so* *kataṃ* *karoti*
he.NOMINATIVE mat.ACCUSATIVE makes.ACTIVE
'He makes a mat.'
(b) *so* *kataṃ* *kurute*
he.NOMINATIVE mat.ACCUSATIVE makes.MIDDLE
'He makes a mat for himself.'

The middle voice typically appears in clauses referring to body actions, grooming, speech acts, cognition, emotive events, and meteorological or spontaneous events (Kemmer 1993). Since middles often refer to properties, not events, verbs in the middle voice commonly have a stative meaning, but this is not a necessary characterization of the middle voice. Cross-linguistically, middle voice often coincides or partially overlaps with the passive (hence such terms as medio-passive, e.g., in the history of Indo-European languages), as well as with the reflexive and reciprocal. This conflation can be explained by the commonality of semantics—in the middle, reflexive, and reciprocal, the same participant is perceived as both the cause (source) of an event and as its recipient, thus as agent and theme.

5. *Antipassive Voice*

The antipassive voice is less common than the passive; in those languages where it is found, variation is observed in the following ways: (a) the marking of the antipassive voice on the verb; (b) the encoding of the subject (in contrast with the subject of the active); and (c) the expression of the theme.

The verb marking, as in the other voices, can be analytical or synthetic. Analytical antipassives (formed with an auxiliary verb) are found in some languages of the Caucasus, for instance in Lak (20b); synthetic antipassives appear to be more common, for example, (21b) (in (21a), where the verb is active, it agrees with both ergative subject and absolutive object; in (21b), the verb in the antipassive form agrees only with the absolutive subject):

(20) Lak
(a) *ga-nal* *uʕrčru*
he-ERGATIVE children.ABSOLUTIVE
buvcun-n-i
carry-AORIST-3SG
'He has carried the children.'
(b) *ga* (*uʕrčru*)
he.ABSOLUTIVE children.ABSOLUTIVE
buclej *ur*
carrying is.3SG
'He is carrying children; he is doing the carrying.'

(21) Chukchi
(a) *ətlʔa-ta* *məčəkw-ən*
mother-ERGATIVE shirt-ABSOLUTIVE
təni-nin
sew-AORIST:3SG:3SG
'The mother sewed the shirt.'
(b) *ətlʔa* (*məčəkw-ək*)
mother.ABSOLUTIVE shirt.LOCATIVE
ine-nni-gʔi
ANTIPASSIVE-SEW-AOR:3SG
'The mother sewed (at the shirt).'

As the contrast between (21a) and (21b) shows, in an ergative language, the subject in the active and the subject in the antipassive may appear in different cases; however, this is not a necessary condition on the antipassive, and ergative languages are attested where the subject of the antipassive remains in the ergative case (Australian languages Djaru, Walbiri, Kuniyanti). The theme in the antipassive construction is syntactically optional, thus it is an adjunct. It can be expressed by any of the oblique cases or adpositional phrases available in a given language (for example, in Chukchi, it can appear in the dative, locative, or instrumental case), and it can be freely omitted. However, as (20b) shows, in some languages the case marking of the theme does not have to change; what is crucial though is that the theme be an adjunct syntactically, which may be manifested though 'behavioral' properties of an NP, not case marking. Finally, in

some languages, antipassives require that the theme not be expressed (for example, in Tzotzil), in 'mirror image' to the non-expression of the passive agent in passives of some languages.

The sources of functional motivation of the antipassive are similar to the sources observed for the passive. Syntactically, the antipassive is used when a language has rules that refer to absolutives. For example, in Chukchi only absolutive arguments are accessible to relativization, thus, when an ergative subject has to be relativized, the antipassive is used as a 'way-station' for relativization:

(22) Chukchi
(a) *ə tlʕa-ta* *məčəkw-ən*
mother-ERGATIVE shirt-ABSOLUTIVE
təni-nin
SEW-AORIST:3SG:3SG
'The mother sewed the shirt.'
(b) *[*məčəkw-ən* *təni-lʔən*] *ətlʔa*
shirt-ABSOLUTIVE SEW-PARTICIPLE mother
(c) [*məčəkw-ək* *ine-nni-lʔən*]
shirt-LOCATIVE ANTIPASSIVE-SEW-PARTICIPLE
ətlʔa
mother
'the mother that sewed the shirt'

Interpretive differences between the active and the antipassive construction derive from the presence of an affected participant (theme) in the former and its absence in the latter. A common inference in the antipassive construction is that the theme is not affected or affected just partially. Since the theme participant does not even have to be specified, the event denoted by the antipassive predicate is often construed as generic or habitual, taking scope over a set of possible objects. This in turn explains why the antipassive is more commonly found in the imperfective aspect and in non-past tenses which do not limit an event or limit it indefinitely. In terms of information structure, the antipassive shares with the active that agent is the most likely topic; unlike the active construction, the antipassive makes it more likely that sentential focus takes scope over the predicate and adjunct (if any) or just the adjunct.

Unlike passive, antipassive is less common across languages, and even in those languages that have an antipassive, it is often considered to be a lexical, not a syntactic process, because of its restricted productivity. A syntactic account of the antipassive has been offered by Baker (1988) in terms of incorporation: the object position in the antipassive construction is occupied by an NP which is morphologically realized as a bound affix (including null allomorphs), and this affix incorporated into the verb stem.

6. Phenomena Related to Voice

A number of phenomena overlap with voice or are encoded by some of the same markers as are used for marking voice, in particular the passive. These phenomena include reflexives (predicates whose two main arguments have the same referent, acting upon itself), reciprocals (predicates whose two main arguments refer to participants acting upon each other), inchoatives or anticausatives (intransitive counterparts of 'cause X to P', as in the English *burn*, *melt*) which receive the same marking as regular or reflexive passives (for example, in Romance languages). Inchoatives are sometimes assimilated to middle voice forms. Discussions of impersonal passives sometimes include active impersonal passives to these constructions is problematic.

(23) French
on discipline ces enfants
'One disciplines these children.'

The active-passive alternation is sometimes extended to include the so-called inverse systems attested mainly in Native American languages. In an inverse system, there are two constructions differeing in whether Agent or Theme is higher on the hiearchy Human > Animate > Inanimate (or such similar hierarchy). If Agent outranks Theme on the hierarchy pertinent to a given language, the direct form of the verb is used (24a); if Agent is lower, the inverse form is used (24b).

(24) Navaho
(a) *hastin* *líí′* *yi-z-tal*
man horse 3SG + DIRECT-3SG-kicked
'The man kicked the horse.'
(b) *hastin* *líí′* *yi-z-tal*
man horse 3SG + INVERSE-3SG-kicked
'The horse kicked the man.'

Aside from the presence of an actual alternation, there is no structural similarity between a voice system and an inverse system.

Early descriptions of ergative case-marking systems (see (9) above) assimilated the ergative to the passive. Although historically some ergative systems can be shown to go back to passives (e.g., in Indo-Iranian or Polynesian languages), synchronically the ergative construction is active, not passive.

Finally, most languages have classes of verbs whose morphology is indicative of one voice but whose syntactic behaviour does not support the morphological appearance. Such verbs are typically described as deponent verbs, a name rooted in the Latin grammatical tradition where the term was applied to verbs with passive morphology (vox passiva) which functioned as active verbs (vis activa), for example, *loquor* 'speak', where the inflection *-or* is that of 1 person singular passive.

See also: Morphological Case in Linguistics; Syntax; Syntax–Semantics Interface; Valency and Argument Structure in Syntax

Bibliography

Baker M 1988 *Incorporation*. University of Chicago Press, Chicago

Benveniste E 1960 'Etre' et 'avoir' dans leurs fonctions linguistiques. *Bulletin de la société de linguistique de Paris* **55**: 113–34

Dowty D 1991 Thematic proto-roles and argument selection. *Language* **67**: 547–619

Gruber J S 1965 Studies in lexical relations. Ph.D. dissertation, MIT

Fillmore Ch 1968 The case for case. In: Bach E, Harms R (eds.) *Universals in Linguistic Theory*. Holt, Rinehard and Winston, New York, pp. 1–88

Keenan E L, Polinsky M 1998 Malagasy. In: Spencer A, Zwicky A (eds.) *Handbook of Morphology*. Blackwell, Oxford, UK, pp. 563–623

Kemmer S 1993 *The Middle Voice*. Benjamins, Amsterdam

Kimenyi A 1988 Passives in Kinyarwanda. In: Shibatani M (ed.) *Passive and Voice*. Benjamins, Amsterdam, pp. 355–86

Kuryłowicz J 1964 *The Inflectional Categories of Indo-European*. Carl Winter, Heidelberg, Germany

Mallinson G, Blake B 1981 *Language Typology: Cross-linguistic Studies in Syntax*. North-Holland, Amsterdam

Perlmutter D, Postal P 1983 Toward a universal characterization of passivization. In: Perlmutter D (ed.) *Studies in Relational Grammar I*. University of Chicago Press, Chicago, pp. 3–29

Spencer A 1991 *Morphological Theory*. Blackwell, Oxford, UK

M. Polinsky

Grammaticalization

1. Definition of Grammaticalization

Many definitions of grammaticalization have been suggested according to various theoretical approaches. One of the most diffused and clearest is Kuryłowicz' definition: 'Grammaticalization consists in the increase of the range of a morpheme advancing from a lexical to a grammatical or from a less grammatical to a more grammatical status', Kuryłowicz (1975). Some examples of grammaticalization are: Latin *habere* from a full verb meaning 'to grasp, seize' developed to an auxiliary in periphrastic verbal constructions for the expression of the future (*dicere habet* 'he/she has to say > shall say') or the past (*dictum habet* 'he/she has said'). The prepositional + relative phrase of the Old English *þa hwile þe* 'at the time that' evolved to the adversative and even concessive meaning of the Modern English conjunction *while* (see Traugott and König 1991). The Latin demonstrative *ille* 'that,' with strong deictic function (*ille homo* 'that man,' *homo ille* 'that notorious, famous man') gave rise to the obligatory article of the Romance languages (French *l'homme*, Spanish *el hombre*), where its function is simply to mark the noun phrase. Grammaticalization is thus a process leading from lexemes to grammatical formatives: 'a sign is grammaticalized to the extent that it is devoid of concrete lexical meaning and takes part in obligatory grammatical rules.' (Lehmann 1995, p. 6). Sometimes the terms 'grammaticization' or 'grammatization' have been used instead of grammaticalization without any substantial difference.

Table 1
Six aspects of grammaticalization

	Paradigmatic level	Syntagmatic level
Weight	Integrity	Scope
Cohesion	Paradigmaticity	Bondedness
Variability	Paradigmatic variability	Syntagmatic variability

As is shown by the example of Latin *ille*, grammaticalization may lead to the emergence of new grammatical categories (in the case in point, the article which did not exist in Latin). In all the adduced examples the grammatical output of the process has passed through a syntactic use of the lexeme—though this need not always be the case.

Lehmann (1995, p. 123) has suggested three relevant aspects of grammaticalization, which realize six parameters both at the paradigmatic and syntagmatic level (see Table 1).

While grammaticalization is a process, the six parameters are not processes but properties that linguistic forms possess in increasing or decreasing measure. A linguistic form shows integrity inasmuch as it maintains its phonological and semantic identity. Phonological erosion and semantic bleaching reduce the integrity of a form (Latin *ille* > Italian *il*, Spanish *el*, etc.). The scope is the structural size of a construction: when a full verb is grammaticalized as auxiliary it makes up one phrase with the main verb and thereby reduces its scope to just the main verb (see the example of *habeo cultellum inventum* 'I have found a knife' in Sect. 3.1). By paradigmatic cohesion is meant the formal and semantic integration of a form into a paradigm: the suppletive forms *better*, *best* are less paradigm integrated than *longer*, *longest*. Bondedness refers to the syntagmatic cohesion: the more a form is grammaticalized the more it will be syntagmatically constrained; French *le*, Spanish *el*, etc., are fully grammaticalized noun determiners which cannot appear in isolation. Paradigmatic variability refers, according to Lehmann, to the freedom with which the language user chooses a sign—and, again, variability decreases with increasing grammaticalization, if cases are weakly grammaticalized the case suffixes may vary with relative freedom:

(1) Hungarian
a gyerek az asztal-on *játsz-ik*
The child the table-SUPERESSIVE CASE plays
'the child is playing on the table'

but *asztal-nál* (adessive) and *asztallal* (< *asztal-val*, instrumental–comitative) would also be possible.

Another example of paradigmatic variability is: the use of Latin *ille* was not compulsory in pre- or postnominal position and was semantically governed, but in Romance languages it is in most cases impossible to have a noun without the article. Finally, the syntagmatic variability is the ease with which a form may be shifted in the linear order of a sentence: the more a form has undergone a grammaticalization process the less variable it will be; **I have a knife found* or **I have a found knife* are impossible in Modern English.

It has to be noticed that the processes alluded to under the three labels of weight, cohesion, and variability cut across different levels of language structure; therefore the same items (e.g., *ille*) turn out to be appropriate examples of different parameters (integrity, scope, etc.). Grammaticalization is a pervasive phenomenon whose effects may be seen at all linguistic levels. Consequently, 'There is a growing interest in integrating grammaticalization with theoretical work in descriptive and historical linguistics' (Giacalone Ramat and Hopper 1998, p. 1).

2. *A Historical Sketch of the Research*

The classical examples of *dicere habet* and *l'homme* go back to Antoine Meillet (1916), who probably used first the term grammaticalization—though the concept is implicitly referred to by some previous linguists such as Wilhelm von Humboldt (*Über das Entstehen der grammatischen Formen und Ihren Einfluß auf die Ideenentwicklung*, 1825), William Dwight Whitney (*The Life and Growth of Language*: *An Outline of Linguistic Science*, 1875), Georg von der Gabelentz (*Die Sprachwissenschaft*: *Ihre Aufgaben*, *Methoden und bisherigen Ergebnisse*, 1901) and others (see Lehmann 1995, Chap. 1, Heine et al. 1991, pp. 5–11).

As can be seen from the titles just quoted, grammaticalization has from the very beginning been considered in a large theoretical frame as one of the most characteristic developments languages may undergo. This holds also for the linguists of the next generations. Edward Sapir (1921, p. 102), for one, observed that 'it is possible for a concrete concept, represented by a simple word, to lose its material significance entirely and pass over directly into the relational sphere [i.e., the expression of abstract morphosyntactic relations, PR] without at the same time losing its independence as a word': forms live longer than their own conceptual content, which may be bleached and eventually fade completely. Sapir's examples are the Chinese and Cambodian verbs for 'give' used in a very abstract sense to express the 'indirecte object' relation (see also Sect. 3).

Though not using the term grammaticalization, the French linguist Emile Benveniste (1968) took up again the case of *habere*, and called it a conservative mutation whereby a new (periphrastic) form takes over the function of an older one in the grammatical paradigm (e.g., *habeo dictum* 'I have said' instead of *dixi* 'I said').

3. *The Present State of the Discussion about Grammaticalization*

3.1 *The Continuum of the Grammaticalization Process*

Already Whitney and Benveniste have underlined the fact that linguistic changes do not happen abruptly: 'Present possession often implies past action: *habeo cultellum inventum*, (…) *habeo digitum vulneratum* 'I possess my knife (recovered after loss) [> 'I have found the/my/a knife,' PR], (…) 'I have a wounded finger' [> 'I have wounded my finger,' PR].' Thus, *I have found the knife* became the indicator of 'a peculiar variety of past action contemplated as completed,' (Whitney 1875, p. 91, cf. Heine et al. 1991, p. 7f, Benveniste gives an insightful philological description of how the periphrastic forms with *habere* came first to be used and later spread out.

As a matter of fact, grammaticalization, like all terms ending in *-ization*, alludes to a gradual process, not a state. Grammaticalization is different from grammar. Even what has recently been alluded to as 'synchronic grammaticalization' implies a process whose final stage cannot be simultaneous with the beginning moment. This is not in keeping with the assumption by generative transformational grammar of rigidly separated grammatical categories. Actually, the 'new trend' of grammaticalization studies which has flourished in linguistics since the 1980s is basically oriented toward a functionalist approach, with very few exceptions. Grammatical categories are not seen as waterproof boxes; they permit, on the contrary, the gradual leaking of forms from one category to another.

Being functionally oriented means that attention is paid to the internal mechanisms, which govern linguistic structures, and the external psychological/cognitive causes of grammaticalization. Several linguists have hinted at the fact that grammaticalization very often originates in discourse strategies (i.e., in pragmatics), evolving to syntactic structures and later obligatory morphological markers (up to a possible total erosion of the markers):

Discourse < Syntax < Morphology < Morphophonemics < Zero (see Givón 1979, p. 209, cf. Lehmann 1995, p. 13, Traugott and Heine 1991, p. 3, Bybee et al. 1994, p. 40f).

Habeo cultellum inventum originates at the discourse level: the speaker wants to underline that he does really possess (*habet*) a knife which has previously been found (*inventum*) and uses a construct where every word keeps its own value. This pragmatic, effect-

Table 2
The grammaticalization–degrammaticalization continuum; α represents the evolution towards grammar (grammaticalization) and β the opposite process of degrammaticalization

α.	lexicon⟶	delexicalization⟶	morphologization⟶	morphology
	full words:			full morphemes:
	habet		*di(ce)re (h)a(be)t*	*dir-a*
β.	lexicon⟵	lexicalization⟵	demorphologization⟵	morphology
	full words:			full morphemes:
	ade		*lemon-ade*	*−ade (lemon + ade)*
	perhaps			*per (+ haps)*

loaded expression evolves to a standard analytic construction of the syntax (with inversion and no longer with syntagmatic variability (see Sect. 1 above): *habeo inventum cultellum*) and finally enters the morphological paradigm of the verb *invenire* 'to find' as its periphrastic (agglutinated) past form (note that at this stage of the Romance languages it is no longer possible to have the word order *habeo cultellum inventum* (**j'ai le couteau trouvé*): the auxiliary has agglutinated to the main verb). As an example of morphophonemic evolution one can quote the French future *dirai*, *dir-a*, etc. 'I shall say,' 's/he shall say' from periphrastic future forms with *habeo*: *dicere habeo*, *dicere habet*: the auxiliary has undergone phonetic attrition and has become an inflectional ending.

There is a continuum between the phases of the grammaticalization process: Heine et al. (1991, p. 1) quote the case of Ewe (Kwa branch of the Niger-Congo family):

(2) *me- ná ga kofi*
1Sg-give money Kofi
'I gave Kofi money.'

(3) *me-ple βɔtrú ná kofi*
1Sg-buy door give Kofi
a 'I bought a door and gave it to Kofi.'
b 'I bought a door for Kofi.'

(4) *me-wɔ dɔ́ vévié ná dodókpɔ lá*
1SG-do work hard give exam DEF
'I worked hard for the exam.'

In (2) *ná* has the function of a full verb, meaning 'give'; in (4) *ná* can be interpreted only as a preposition ('for'), whereas in (3) it is ambiguous and allows two different readings. Should one day the Ewe language admit for (3) only the second reading, then we would have a categorial shift (VERB > PREPOSITION), much in the sense of the evolution of Old English *hwil* 'time' (NOUN) to *while* (CONJUNCTION) (see above, Sect. 1). At the end of an uninterrupted process there is really a break and a category shift.

Shifts like VERB > PREPOSITION, NOUN > CONJUNCTION are examples of so-called *transcategorization*. Items belonging to a main (and open) category (VERB, NOUN) pass to a functional (and restricted) category (e.g., ADPOSITION PRE-OR PREPOSITION according to the structure of the language, concerned, CONJUNCTION): these examples clearly hint at a move from a less grammatical to a more grammatical categorial status—it is however clear that the term *transcategorization* (or *recategorization*) is more appropriate than *decategor(ial)ization* used by Hopper (1991, p. 30f) and others.

3.2 The Dynamic of the Grammaticalization Processes

In such a dynamic typology of grammaticalization it is possible to account also for interrupted grammaticalization processes: in Old Italian it was possible to use movement verbs (*andare* 'go' and *venire* 'come') as auxiliaries in far more cases than in Modern Italian (cf. Giacalone Ramat and Hopper 1998, p. 2f, with further references). Further, Hopper has drawn attention to what he calls 'emergent grammar,' i.e., to the incipient stages of grammaticalization at the level of discourse (see above): the negative French adverb *pas* meant originally 'step'; therefore it was used first with movement verbs ('he doesn't go a step'), but later it spread to verbs such as 'eat,' 'see,' etc. and became the general mark of negation, ousting other nouns that denoted a least quantity and were used, like *pas*, to reinforce the negation (e.g., *mie* 'crumb,' *gote* 'drop,' etc.): specialization principle. Thereby the original meaning of *pas* was completely lost (bleaching process).

As noted by Hopper himself, the principles he suggests (specialization, persistence, divergence, etc.) characterize linguistic change in general and are not distinctive for grammaticalization (Hopper 1991, p. 21). This has raised the problem of defining the limits of grammaticalization vs. other dynamic linguistic processes (cf. Giacalone Ramat and Hopper 1998).

3.3 Unidirectionality of Grammaticalization?

Unidirectionality is a much debated question in the literature on grammaticalization. Many linguists maintain that the evolution towards grammar cannot

be reversed (Lehmann 1995, p. 19, Traugott and Heine 1991, p. 4f, Heine et al. 1991, p. 31f). This is true if by this it is meant that once a lexeme has entered the grammar, it will hardly come back and assume again a lexical value. But there are many instances of items or morphemes leaving the realm of grammar and assuming an autonomous value as lexical entries: Hungarian, Spanish, and Italian have a large number of present participles used as nouns, e.g., Hungarian *költö* ('poet' (< *költ* 'to compose'), Spanish and Italian *cantante* 'singer' (< *cantar*(*e*) 'to sing'); nobody can fail to recognize in French *seigneur* (> English *sir*) the comparative of the Latin adjective *senex* 'old' or in English *elder* 'official in the Presbyterian church' a comparative from Old English *ald* 'old.' Both were originally forms belonging to the grammatical paradigm of the adjectives, but later they developed an autonomous lexical meaning. Even syntactic constructs may develop to new lexemes: Dutch *misschien* 'perhaps,' (< (*het*) *mach schien* literally 'it may happen'), Italian *forse* 'perhaps' (< Latin *Fors sit* (*an*) literally 'be the case'), English *perhaps* literally 'per chances (plural).' Grammaticalization and lexicalization are two complementary processes (Moreno Cabrera 1998, Ramat 1992).

Degrammaticalization processes are certainly less frequent than grammaticalization processes, but they exist and have to be explained. The causes for their existence are to be sought in the overall tendency to use labels as economic symbols: e.g., suffix morphemes like *-ade* (from *orangeade*, *lemonade*, or even *gatorade*) can be separately manipulated and become autonomous lexemes: the label *ade* with the general meaning of 'fruit juice' is more economic than the phrase *fruit juice*, though not as transparent (see *bus*, in *schoolbus*, *autobus*, etc., as hyperonym for 'public transportation vehicle'). Economic symbolism and noneconomic iconism are conflicting forces which are at work in every natural human language (cf. Ramat 1992, p. 557, Newmeyer 1994, p. 69). If we had just unidirectional evolution toward grammar we would expect that languages become more and more grammaticalized, which by all evidence is not the case (see Moreno Cabrera 1998, p. 224). On the contrary, we have a double evolution towards grammar as well as towards lexicon (see Kuryłowicz 1975).

Phonogenesis is also a process which shows that loss of grammatical function may give rise to new lexemes (cf. Hopper 1990, 1992). As the German grammatical prefix *ge-*, used to indicate collectives (see *Ge-beine* 'limbs,' plural, *Ge-dränge* 'crowd'), was no longer felt as a prefix, it underwent phonetic erosion to *g-* and gave rise to nouns such as *Glück* 'luck,' *Glied* 'limb,' *Gnade* 'grace' which synchronically are no longer analyzable: phonetic junk, no longer recognizable as a functional morpheme, produced new lexemes. In the same vein of reasoning we have to consider also remnants of disappeared paradigms: this applies to dual forms which survived the loss of the dual declension: cf. Latin *ambo* 'both' (perhaps also *octo* 'eight') with the old dual marker *-o*.

Summing up, we may represent (see Table 2) the grammaticalization and degrammaticalization processes along a continuum which moves into two opposite directions. At both poles of the scale we have either clear lexemes (full words like *habet*, *perhaps*) or morphemes (inflectional *–a* or derivational *–ade*), while it is difficult to state when *habet* was delexicalized and the construct *dicere habet* became the norm to express the future, or *ade* was reanalyzed as a second member of a compound: that is why *di*(*ce*)*re* (*h*)*a*(*be*)*t* and *lemon-ade* appear in the previous scheme midway between (de)lexicalization and (de)morphologization. Transcategorization is a gradual process along a continuum; its final output represents a break in the categorial status of the input (see Sect. 3.1).

Bibliography

Benveniste E 1968 Mutations of linguistic categories. In: Lehmann W P, Malkiel Y (eds.) *Directions for Historical Linguistics: A Symposium*. University of Texas Press, Austin, TX, pp. 83–94

Bybee J, Perkins R, Pagliuca W 1994 *The Evolution of Grammar, Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press, Chicago

Giacalone Ramat A, Hopper P (eds.) 1998 *The Limits of Grammaticalization*. Benjamins, Amsterdam

Givón T 1979 *On Understanding Grammar*. Academic Press, New York

Heine B, Claudi U, Hünnemeier F 1991 *Grammaticalization. A Conceptual Framework*. University of Chicago Press, Chicago

Hopper P 1987 Emergent grammar. *Berkeley Linguistic Society* **13**: 139–57

Hopper P 1991 On some principles of grammaticization. In: Traugott E, Heine B (eds.) *Approaches to Grammaticalization*. Benjamin, Amsterdam Vol. 1, pp. 17–35

Hopper P 1992 Phonogenesis. In: Pagliuca P (ed.) *Perspectives on Grammaticalization*. Benjamins, Amsterdam, pp. 119–35

Kuryłowicz J 1975 The evolution of grammatical categories. In: J K (ed.) *Esquisses Linguistiques* Vol. 2. Fink, Munich, Germany, pp. 38–54

Lehmann Chr 1995 *Thoughts on Grammaticalization. A Programmatic Sketch*. (*Arbeiten des Kölner Universalien-Projekts Vol. 48*). Lincom Europa, Munich, Germany

Meillet A 1948 L'évolution des formes grammaticales. In: A M (ed.) *Linguistique Historique et Linguistique Générale*. Champion, Paris, pp. 130–48

Moreno Cabrera J C 1998 On the relationships between grammaticalization and lexicalization. In: Giacalone Ramat A, Hopper P (eds.) *The Limits of Grammaticalization*. Benjamins, Amsterdam, pp. 211–27

Newmeyer Fr J 1994 Competing motivations and synchronic analysis. *Sprachtypologie und Universalienforschung* **47**: 67–77

Ramat P 1991 Thoughts on degrammaticalization. *Linguistics* **30**: 549–60

Traugott E C, Heine B (eds.) 1991 *Approaches to Grammaticalization*. Benjamins, Amsterdam

Traugott E C, König E 1991 On the expression of spatio-temporal relations. In: Greenberg J, Ferguson Ch, Moravcsik E (eds.) *Universals of Human Language*. Stanford University Press, Stanford, CA, Vol. 3, pp. 369–400

P. Ramat

Gramsci, Antonio (1891–1937)

Antonio Gramsci (born in Ales, Sardinia on January 22, 1891 and died on April 27, 1937) started out as a freelance journalist writing assiduously for *Il Grido del Popolo* and *Avanti!*, newspapers of the Italian Socialist Party in whose ranks he had been active since 1914. In 1919 he founded *L'Ordino Nuovo* which became the daily paper of the Turin Communists, of whom Gramsci was to become undisputed leader. In 1921 he joined the Communist Party of Italy, created from the walkout during the XVIII Socialist Congress in Livorno. The following year he was in Moscow representing the Italian Communists in the Third International. Returning to Italy in 1924 he founded the newspaper *L'Unità*. Two years later he participated in the III Congress of the Italian Communists, held in Lyon, where he opposed Amadeo Bordiga and obtained an overwhelming majority. The same year, despite his parliamentary immunity, he was arrested by the Fascists. Given a summary trial, he was sentenced to 20 years of imprisonment by the special Court for the Defence of the State.

Taken to the prison of Turi in the province of Bari, he remained there until 1933 when he was moved to Formia due to poor health. Obtaining a conditional discharge, in 1935 he was admitted to the Quisisana hospital in Rome, where he died following a stroke. Thanks to his sister-in-law Tatiana, his *Quaderni del carcere* ('Prison Notebooks'), written between 1930 and 1935, were sent underground to Moscow and delivered to Palmiro Togliatti. Published after the collapse of the Fascist dictatorship, they revealed a leading theoretician who, albeit following the path of Marxist–Leninist tradition, had elaborated an original concept of society and history underpinned by the theory of hegemony.

Gramsci's political thinking was deeply influenced by the October Revolution. He immediately saw it as confirming his activist and voluntaristic reading of Marxism, induced by Giovanni Gentile's neo-idealism. For Gramsci, Marxism was something more than a simple, empirical interpretation of history, as Benedetto Croce had defined it; it was a complete philosophy containing, in nuce, a new civilization to be constructed over a long and dramatic process. The industrial proletariat or, more specifically, its conscious vanguard, the Communist party, would be its protagonists. So Gramsci enthusiastically greeted the Bolshevik October as 'the revolution against the Capital.' The revolution against the positivistic and deterministic interpretation of Marxism which the parties of the Second International had adopted while awaiting the inevitable, catastrophic collapse of capitalism. Gramsci also considered that Lenin had been able to grasp the systematic link between Marxism and German classical philosophy, of which Hegelianism had been the speculative summit. Thus Gramisci was prompted to see Lenin not only as the greatest strategist of the international workers' movement but also its greatest theoretician. In his deeds, even before his thoughts, he had expressed a higher degree of consciousness. Therefore, the unity of the theory and practice incarnated by Lenin and by the movement he had created had to be emphasized. Cultural work was thus a priority, to demonstrate that Communism was the positive response to the world crisis of capitalism and the moral dissolution of the bourgeois civilization, precisely because it had, thanks to Marxism, reached a higher degree of theoretical awareness. It was no coincidence that Bolshevism arose from the claim that the workers' consciousness should prevail over their spontaneity and that without a revolutionary theory no really revolutionary action was even conceivable. For Gramsci, the Communist revolution had to be seen as a philosophy which was being put into practice or, more precisely, as an ethical-political will enlightened by a theory whose aim was the reconstruction of the world.

Up to the end of 1923, Gramsci's thinking, expressed in a myriad of short articles on the political situation, was nothing else but an intelligent adaptation of Leninism to the Italian situation. All the typical elements of the Bolshevik model were to be found in it: the supremacy of theory over spontaneity, the concept of the party as a nucleus of activists devoted heart and soul to the revolution and submitted to iron discipline, the intransigent defense of orthodoxy against any form of revisionism, the military vision of the class war, the revolution as a declaration of war on the existing, the regime of transition focused on the dictatorship of the proletariat exercised by its conscious vanguard, the edification of Socialism as a substitution of the anarchy of the market with a single production and distribution plan; one can even find the idea, already expressed by Lenin to justify the red terror, that the Communists' historic mission was to 'purify the social environment with iron and fire,' exterminating the lower and middle bourgeoisie, 'corrupt, dissolute and putrescent humanity.' But, drafting his program-letter of February 1924, Gramsci started to work out the revolutionary theory that, in his *Quaderni del carcere*, was to be called the 'war of position' in opposition to the 'war of movement.'

Gramsci was the first Communist leader to realize that the Bolshevik strategy had few possibilities of success in Western Europe, with its robust civil society, rich in counter-powers and institutions relatively independent of the state. The Third International was

founded on the hypothesis that the world capitalist system had entered a phase of rapid and uncontrollable disintegration and that, therefore, the Communists' task would be that of accelerating this disintegration, mobilizing the proletarian masses against the bourgeois state, and enabling the Bolsheviks to take power. But the Comintern leaders had underestimated the vitality of capitalism that, despite the great postwar crisis, had shown it was strong enough to suffocate any revolutionary attempt before it was born. The 'war of position' was Gramsci's idea for rescuing the Communist movement from the blind alley down which it had gone. He had realized that the dominance of the bourgeoisie was based more on the consensus of the masses than on control of the coercive machinery. The celebrated theory of hegemony, developed in his *Quaderni del carcere*, says that the 'supremacy of a social group is shown in two ways, as dominance and as intellectual and moral leadership.'

Which is like saying that a class is hegemonic only insofar as it is capable of increasing the material and spiritual patrimony of society and obtaining the consensus of the working classes. Therefore there is no hegemony that is not founded on a 'historical bloc,' that is on an organized system of social alliances cemented by a common ideology. The historic bloc is the social basis of hegemony. It is not only a political phenomenon, it is also, and in a certain sense above all, a cultural phenomenon. This led Gramsci to claim that the superiority of a hegemonic class lasts insofar as it is able to keep its system of alliances united and bring in other social groups by satisfying, albeit within precise limits, their material and moral interests. Then, and only then, will it become the 'driving force of a universal expansion, of a development of all the national energies.' Gramsci's approach to the problem of social order was typical of the sociologies of consensus, stamped by a certain Marxism as interesting ideological distortions of reality, and he grafted their fundamental theorems on the trunk of historical materialism. He also took over Vilfredo Pareto and Gaetano Mosca's elitist theories of the late nineteenth and early twentieth centuries. His central premise was that the historical existence of a society is, always and necessarily, concentrated on the minorities–masses dialectic.

In *Quaderni del carcere*, society is conceived along Marxist lines as a hierarchical structure of classes, differing according to their position and function in organizing the production and reproduction of material life. But, for a class to make history, it must distinguish itself and become independent; it has to acquire a critical knowledge of itself. This happens gradually, transforming a class, an immediately corporate reality, into an ethical–political force with a national character that imposes its leadership on society. This is precisely the stage in which ideologies confront each other and enter into battle until one of them, or a combination of them, takes the upper hand and unites not only the economic and political, but also the intellectual and moral, aims of society. Vice versa, a dominant class starts to decline when it gradually loses its ability to interpret the needs of the dominated classes and to mobilize national energies. It is true that a dominant class can keep its power even when it has lost its ability to rule, thanks to its control of the state's coercive machinery; but in the long run it is destined to be replaced by a new ruling class.

This led Gramsci to a conclusion which is, basically, the reformulation of Pareto's famous theory of revolution as a violent and traumatic substitution of an old aristocracy with a new aristocracy. But where Pareto only saw ideologies as rhetorical stratagems through which an elite justified its desire to dominate, Gramsci ascribed an enormously important function to them. Ideologies are not mere derivations, as Pareto called them; they are global visions of reality, essential for constructing a civilization. There is no civilization without an ideology, and without this ideology being absorbed by the lower classes and transformed into popular culture. So it can be said that history is dominated by the clash between ideologies, between hegemonic principles competing for the consensus of the masses. And the theatre of this clash is civil society. Gramsci saw this differently from Marx, that is as the society of industry and general competition, material base of the state and the ideological superstructures. For Gramsci, civil society is the theatre of intellectual and moral relations, where a social group that aspires to dominance constructs its own hegemony by soliciting the consensus of the dominated classes. Civil society, therefore, is a cultural rather than an economic reality: it is where rival hegemonic principles confront each other until one of them eliminates the others and becomes the impersonal engine of historical development.

Now, if what is decisive in history is not economic but spiritual power, then to move the middle classes out of the city of command, you first of all and above all have to earn the active consensus of the lower classes. That is, you have to incite a critical attitude to the existing order and persuade the workers to support the new type of society implicit in Marxist philosophy. The second consequence drawn from Gramsci's theory of hegemony was that the Bolsheviks' Jacobin strategy was not applicable to the type of war that had to be waged in Western societies. It presumed the power vacuum which had effectively occurred in Russia in 1917; in other words, it presumed the disintegration of the ruling class and the general collapse of the political system. Lenin and Trotsky's Blitzkrieg tactics had not, therefore, any possibility of success if there was a sufficiently cohesive historic bloc to repel frontal assault by the armies of the revolutionary movement. It was necessary to change the line of action, to work out a long-term strategy for gaining power.

The Italian Communist leader's answer to this problem is to be found in his *Quaderni del carcere*.

Studying the history of Western civilisation, Gramsci found a perfect model in the victory of Christianity over the Greco-Roman culture. Christianity had been, in its way, a revolutionary movement that had introduced a spiritual scission in ancient society and had triumphed thanks to its intellectual and moral superiority, to the indefatigable teaching of its apologists and to its organization. If it was impossible to conquer the bourgeois citadel manu militarti, it was possible, as the history of the expansion of Christianity indicated, to conquer it by wearing away the ideological defenses of the middle classes. Thus the Communists had to imitate the early Christians by eroding from within, by a methodical, silent and all-embracing conquest of the ideological superstructures and the existing order until the working classes stopped supporting the ruling class. One of the fundamental theorems of Gramsci's sociological theory was that a class which wants to win power has first of all to be hegemonic. That is, it has to be able to exercise a firm and united, intellectual and moral leadership over the working masses and then to dominate them. It has methodically to take possession of the 'fortresses' and 'casemates' of civil society and transform them into the operational bases of its march towards the city of command. Thus, whereas Leninism had theorized the occupation of society through the violent conquest of the state, Gramsci proposes an inverse procedure: the conquest of the state through the cultural occupation of society. To this end, it was all-important to penetrate the mass-media, schools, universities and publishing houses and convert them into agencies for the socialization of Marxism.

In Gramsci's model for the conquest of power, violence was relegated to the background. Nonetheless, the strategic goal remained that of the Bolshevik Revolution: the institutionalization of the mono-party dictatorship as an indispensable condition for totally remodeling the social body. On this point Gramsci remained faithful to Lenin's teaching. The party, and only the party, had to be the leading light of the intellectual and moral reform for creating the new civilization whose institutional foundations had been laid in the Soviet Union. So that Gramsci, after having indicated the Communist party as the only institution which could be compared to the Catholic church, did not hesitate to define Marxism as the 'religion which could kill Christianity' by building the City of Man. And he saw the modern prince (the party of the professional revolutionaries, conscious vanguard of the working classes) as the historical agent. He who, overthrowing the entire intellectual and moral system of liberal civilization with his impassioned teaching, would have 'to take, in people's minds, the place of the divinity and the categorical imperative' (*Quaderni del carcere*).

Gramsci's is a rigorous theorization of the charismatic nature of the revolutionary party, not unlike Gyorgy Lukacs' in *Geschichte und Klassenbewusstein*, where the Communist party is described as the cosmic–historic agent which, in virtue of its superior theoretical knowledge, will have to achieve the 'salvation of humanity.' Gramsci defines the modern prince as the 'focus of the faith' and the 'custodian of the doctrine of Socialism' destined to introduce an epochal caesura in universal history. This prince has to ensure that he becomes the source of all the satisfaction his militants had previously found in a multiplicity of organizations. He must also destroy all these other organizations or incorporate them into a system in which this modern prince will be the 'only regulator.' Hence the distinction, spelt out in *Quaderni del carcere*, between two types of totalitarianism: the regressive kind, with a party opposing the drive of history, and the progressive one, which is precisely that of the modern prince bringing a new and superior culture.

This led Gramsci to affirm peremptorily that 'everything is politics,' and that, consequently, it is not possible to separate politics from philosophy, since the latter is the driving force of history in the making. A history that is always 'between two hegemonic principles, between two religions,' each of which aspires to regulate human life in all its material and spiritual manifestations. The idea of a competitive coexistence between a plurality of social forces and cultural models, so typical of liberal civilization, is not only entirely extraneous to Gramsci's political philosophy but even explicitly rejected. The final aim of his philosophy is the construction of the future city, conceived by Marxism as a 'harmonious society.'

The role of the intellectuals is of fundamental importance within this vision of the Communist revolution. They are the protagonists of history, since history is nothing else but the dramatic process through which a determinate philosophy takes over the world. Gramsci sees the revolutionary party as a 'collective intellectual' and gives the 'organized intellectuals' framed within the rigid discipline of Leninist centralism the function of 'permanent persuaders.' This explains why Gramsci's philosophy, in the decades following the publication of the Khrushchev Report (1956), had such a powerful influence on the Communist intelligentsia. The latter found in the *Quaderni del carcere* what they needed to liberate Marxism from Stalinist dogmatism and to conceive the fight for socialism as a magnificent pedagogical operation to conquer spiritual hegemony by changing how the subordinate classes felt and thought. But after the crisis of Marxism in the 1980s, Gramsci's theory of the peaceful conquest of power rapidly lost its fascination, despite the fact that some basic concepts of this theory—and first of all the concept of hegemony—have entered the lexicon of contemporary sociology.

See also: Bourgeoisie/Middle Classes, History of; Communism, History of; Hegemony: Cultural; Historiography and Historical Thought: Modern History

(Since the Eighteenth Century); History: Forms of Presentation, Discourses, and Functions; Marxism in Contemporary Sociology; Marxism/Leninism; Marxist Social Thought, History of; Power in Society; Power: Political; Russian Revolution, The; Social History; State, History of; Subaltern History; Violence, History of; Working Classes, History of

Bibliography

Anderson P 1977 *The Antinomies of Antonio Gramsci*. London
Buci-Glucksmann C 1975 *Gramsci et l'Etat*. Fayard, Paris
Cammet J M 1967 *Antonio Gramsci and the Origins of Italian Communism*. Stanford University Press, Stanford, CA
Femia J V 1981 *Gramsci's Political Thought*. Oxford University Press, Oxford, UK
Finocchiaro M 1987 *Gramsci and the Methodology of Criticism*. Las Vegas
Finocchiaro M 1999 *Beyond Right and Left*. New Haven
Hobsbawm E J (ed.) 1995 *Gramsci in Europe and in America*. Bari
Jocteau G C 1975 *Leggere Gramsci*. Fettrinelli, Milan
Pellicani L 1981 *Gramsci: An Alternative Communism?* Hoover Institution Press, Stanford
Pellicani L 1991 *Gramsci, Togliatti e il Pci*. Rome
Portelli H 1973 *Gramsci et le Bloc Historique*, 1st edn. Presses universitaires de France, Paris
Tamburrano G 1963 *Antonio Gramsci*. Lacaita, Manduria

L. Pellicani

Graphical Methods: Analysis

1. Introduction

Everyone is accustomed to seeing graphical displays used for presentation purposes: in scientific articles, newspapers, and magazines. Yet graphical displays may be used for exploratory purposes too, for investigating datasets and for uncovering the information hidden in them. Exploratory graphics are different. The aim is not to present information, but to use the graphics to uncover information. Static printed graphics can be useful for exploratory work, but the real gain lies in using interactive graphics on a computer.

In all the following examples it is assumed that exploratory analyses are being carried out, that each graphic is part of a broader analysis, and that they are drawn to aid the analyst, not to present the information to others. This is why there are few scales given. The analyst will know the approximate scaling and can always query the graphic on screen to get exact values, not only along axes, but for any point or object. In general a single presentation graphic is drawn to convey information to a potentially large and varied audience, whereas a large and varied collection of exploratory graphics are drawn to extract information for a single person, the analyst.

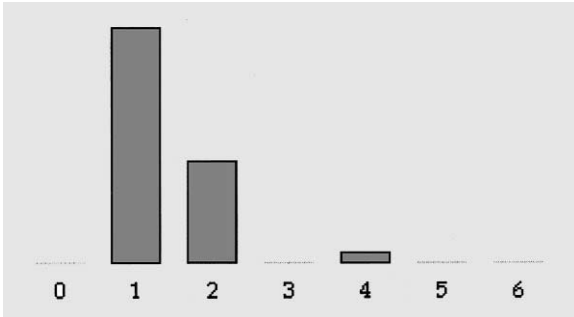

Figure 1
Barchart of gender of car insurance policyholders

Graphics are useful for checking basic assumptions and for getting a feel for data. Figure 1 shows a bar chart of gender from an insurance data set. Unusually there are seven codes for gender! Three made sense (firms are neither male nor female) and the rest were errors.

The histogram for policyholders' birth dates for the same data set was also revealing (Fig. 2). The tall white column to the left represents missing values for the variable (mainly firms again). There is a small isolated column at 1,900 (missings originally coded as '00'?) and there are some very recent dates of birth, surely errors for car-driver insurance data. This is a large dataset (over 50,000 cases) and some errors may be expected, but it is extraordinary how often datasets which have been purportedly quality checked still have errors in them. Incidentally, the dip in the histogram is due to low birth rates during the war.

Assuming major errors have been corrected or removed, graphics are valuable both for checking old ideas and for generating new ones. After the historic 1990 Irish Presidential election it was hypothesized that Mary Robinson had won because she got the women's vote. Data from the last MRBI survey before the election suggests otherwise. Figure 3 shows a bar chart for the preferences for the three candidates linked to a spine plot for gender (in a spine plot all column heights are the same and column widths represent the counts so that highlighted proportions are directly comparable). The proportion of women intending to vote for Mary Robinson was only slightly higher than the proportion of men. Linking to another spine plot shows that urban voters were more important for Mary Robinson's win.

2. Background

Presentation graphics (see *Graphical Methods: Presentation*) have been used for a long time. Exploratory

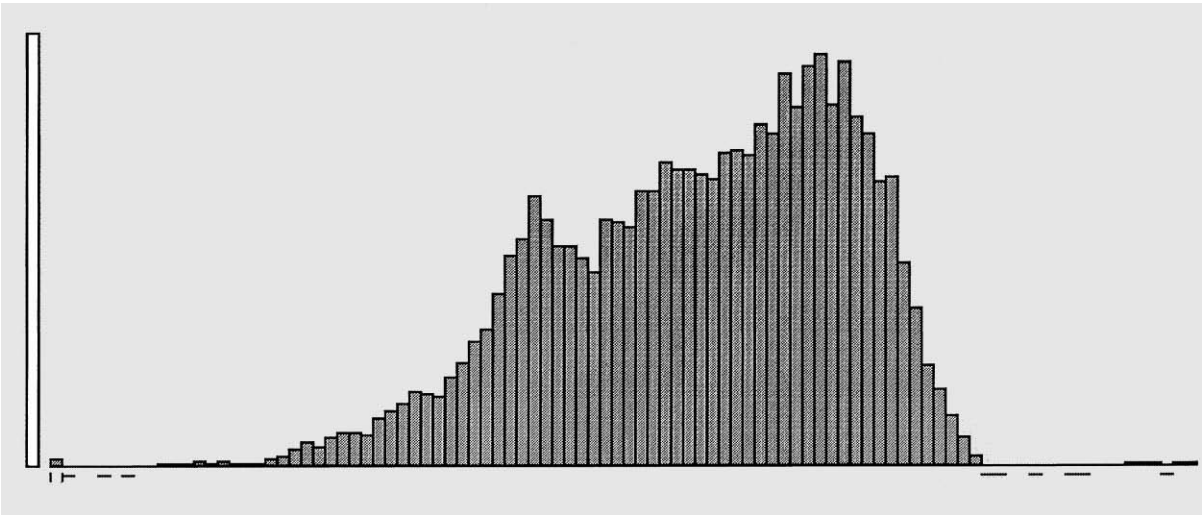

Figure 2
Histogram of years of birth of car insurance policyholders

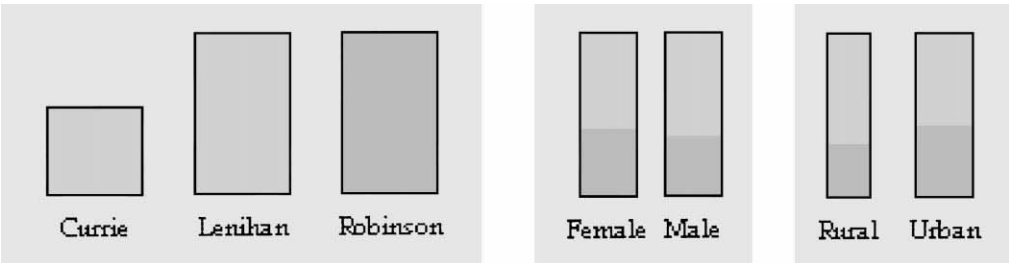

Figure 3
Barchart of candidates' support linked to spineplots for gender and location. (Results from the last MRBI survey before the 1990 Irish Presidential Election. Thanks to Jack Jones and the Market Research Bureau of Ireland for permission to use the data)

graphics are newer, not because people did not appreciate their value, but because without a computer it was not efficient to draw many variations of many different displays. John Tukey led the way with his innovative and provocative book, *Exploratory Data Analysis* (Tukey 1977; see *Exploratory Data Analysis: Univariate Methods*) although, curiously enough, the book was written slightly too early, before personal computers were becoming commonplace. Thus he emphasizes pencil and paper methods like stem and leaf plots for continuous data. These are like histograms laid on their side with each bin having a width of a power of 10. The advantages are the simple construction by hand, and the listing of each value individually. The disadvantages are the dependence on a very limited range of bin sizes, and the unsuitability of the method for anything other than small datasets. Imagine a stem and leaf plot for the car insurance data set of Fig. 2.

More recent computer-oriented references include Cleveland and McGill (1988), Eick and Wills (1995), and Unwin (1999). There has also been much related work outside statistics for which Spence (2001) is a good source. In none of these works is a substantial formalization of exploratory graphics attempted. Considering that presentation graphics have been around much longer and yet Wilkinson's groundbreaking work (Wilkinson 1999) only appeared in 1999, this is not surprising. Furthermore, exploratory analyses are by their nature interactive and dynamic and so the process is difficult to explain on the printed page. Some of the best sources of information and examples

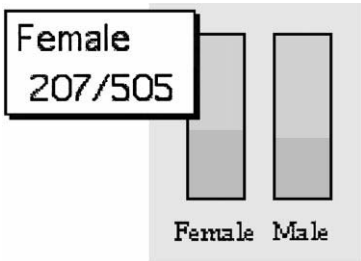

Figure 4
Direct querying of a bar. How many women supported Mary Robinson?

can be found on the web using keywords in this article and/or author names as search parameters.

3. *Interacting with Displays*

3.1 *Interactive Features*

Graphics may mislead. For instance, bar charts will look similar whether they are based on 50 or on 5,000 cases. (The survey in Figs. 3 and 4 was based on 1,000 interviews.) It is important to be able to check any impressions gained. Sometimes this might mean carrying out the corresponding statistical test (e.g., a chi-square test for comparing bar chart proportions), but a first step is to directly query the graphic, as in Fig. 4.

Direct querying is useful in all displays, especially when it is context sensitive. The same action may be used for each part of each plot and yet provide different information according to which part of which plot is queried. Other steps that may be carried out to gain more insight are to rescale the graphic (e.g., change the bin size in a histogram), to reformat the display (e.g., switch the axes of a scatter plot), to reorder the display (e.g., order the categories in a bar chart by count), to zoom in (e.g., concentrate on an outlying cluster of points in a scatter plot), or to select various groups of points and link them to their representation in other displays (see Unwin 1999).

Histogram bin sizes which are small pick out gaps and identify preferred values, while large bin sizes emphasize overall trends. Figure 5 shows two histograms of maths marks for 126 Irish schoolchildren. Notice how one version picks out the typical gap below the pass mark of 40 and the other does not.

Scatter plots can be extremely informative tools for revealing the relationship between two variables, but the scales, the aspect ratio of the axes, the size and color of the points can all affect how they look. For large datasets with extensive overplotting (i.e., many points are drawn at the same location on the plot), some form of density estimation (see *Probability Density Estimation*) is necessary or there is a risk the dataset may be misunderstood. Figure 6 shows a scatter plot of the birth weights of 4,000 babies plotted against the estimated duration of pregnancy. A coarse density estimation has been used whereby black dots represent single cases and the brightest white ones at least 30 cases. It is clear where the majority of the data is concentrated.

Neither drawing sunflowers at each point with the numbers of petals representing the numbers of points nor jittering the data (adding small random components to each point to reduce the overlapping) are serious solutions for large datasets like this (although both suggestions are to be found in the literature). Sounder solutions have been suggested by several authors including Carr et al. (1987) and Hyndman (1996), but should be implemented interactively. Single displays optimized on a particular criterion cannot reveal all the structure in the data. An interesting problem for interactive scatter plots is how to display both highlighted and unhighlighted distributions informatively.

Variations of scatter plots may be used to display subgroup distributions in parallel, e.g., income of those surveyed by job category. The resulting parallel dot plots are a valuable initial exploration tool for smaller datasets, but parallel box plots are more effective for large datasets as in Fig. 7.

Bar charts are good for comparing absolute numbers, but are not good for displaying proportions. Simple pie charts are good for emphasizing shares, though it is difficult to compare the sizes of individual sectors and too often they are drawn with irrelevant three-dimensional features and other distracting distortions. Spine plots are best for comparing proportions based on selections, but can only deal with one

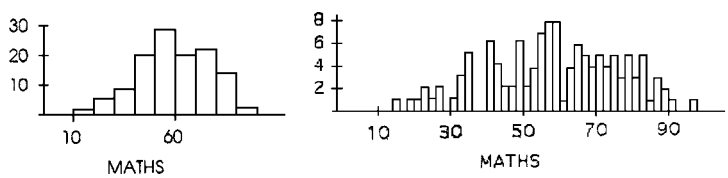

Figure 5
Histograms of maths marks out of 100 for a group of Irish schoolchildren

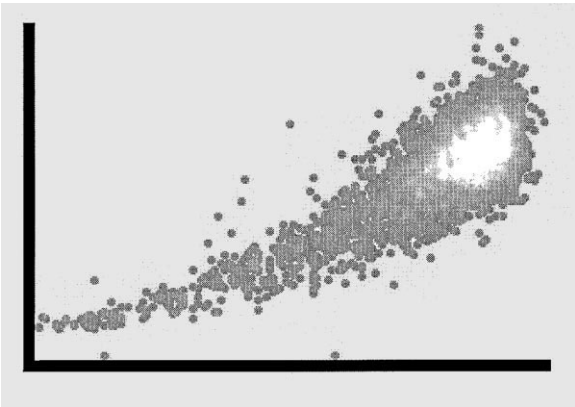

Figure 6
Babies' birthweights plotted against duration of pregnancy

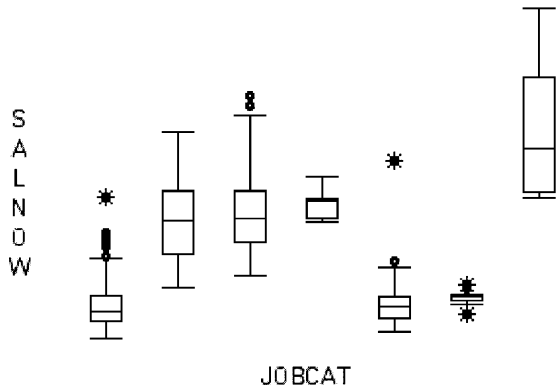

Figure 7
Boxplots of salaries by job category in a bank

selection at a time (see Fig. 3). Switching between bar chart and spine plot representations interactively is one solution.

3.2 Linking

Linking means that points selected in one display are highlighted in all other displays. Selecting a subgroup may identify an outlier within the group in another plot. There may be patterns within subgroups which mask each other when the data for all groups are shown together. Highlighting area-based plots is straightforward; an area corresponding to the size of the selected group is colored. Highlighting point displays is complicated by overplotting, as mentioned above in relation to scatter plots. The best solution for box plots seems to be the idea due to Wills of drawing a box plot for all the data in the background and a box

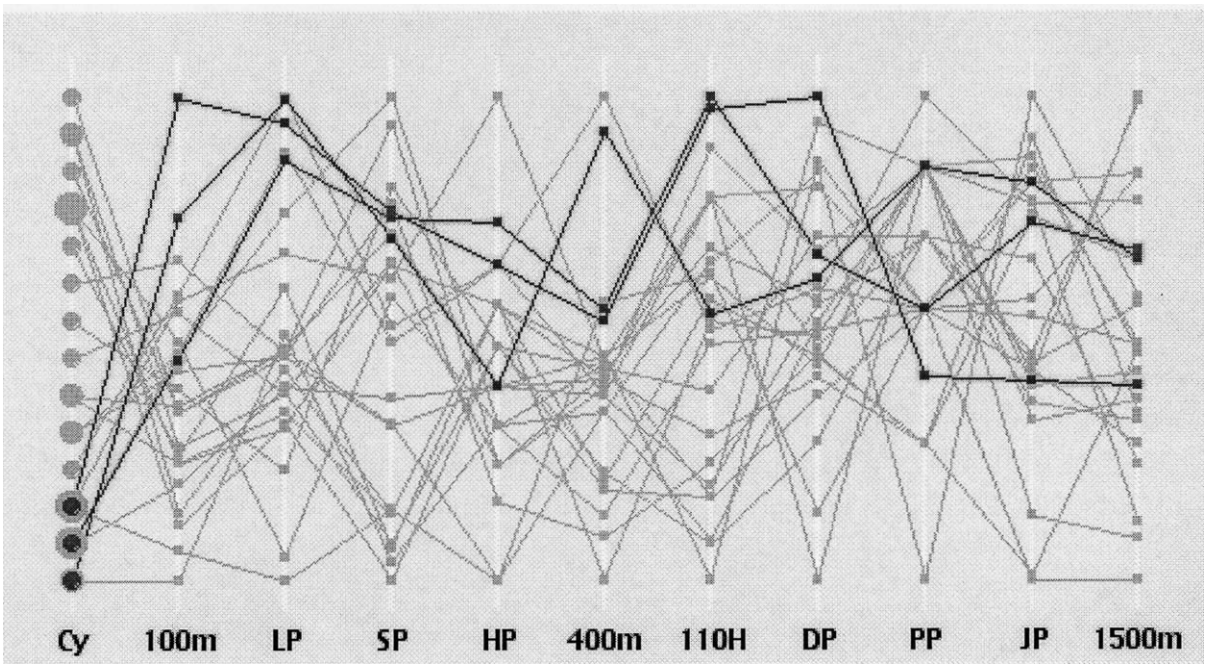

Figure 8
A parallel coordinate plot of the points by event for the mens' decathlon at the Sydney Olympics 2000. The medal winners have been highlighted

plot for the highlighted data on top (Eick and Wills 1995).

Linking is central to successful interactive graphics in the way it provides insight into multidimensional structure. Not just points in graphics may be linked, but cases, scales, orderings, and colors may be linked in any kind of window. In this sense querying is just a form of linking.

3.3 Interfaces

All of these tools are exploratory in nature and only work effectively when they are interactive. If you have to wait for the screen to clear and a graph to redraw, you will have forgotten why you wanted the change in the first place. Interactivity implies an intuitive interface, flexible commands, and a fast response. Command–line interfaces are valuable for the precise control they offer and the ease with which they can be recorded. For exploratory work it is direct manipulation of the objects on screen which is important and it is rarely of interest to know how a conclusion was first discovered, as it is common to find much better ways of showing results after you have first discovered them. All components of a graphic display (bars, axes, points, legend, etc.) may ideally be queried, moved, rescaled, and reformatted. To avoid clutter on screen, it is helpful to use cues to hint at what is possible rather than to fill the screen with visible controls. To protect the user from possible unwanted side effects of their actions, warnings, and undos are beneficial.

3.4 Dynamic Graphics

The term 'dynamic graphics' refers to graphical displays that move such as rotating plots, grand tour displays, or animations controlled by sliders. (A slider is a control which allows the user to change the value of a parameter of a model in real time and see the affected displays update continuously (see Data Desk or Tierney 1990). While the resulting dynamic displays may be informative, it is the interactive control in the hands of the user which is most useful and in this sense they are a specialist subset of interactive exploratory graphics.

4. Multivariate Displays

4.1 Continuous Variables

Modern graphical tools are very good for exploring many variables at once. Scatter plot matrices (a matrix display of scatter plots of all possible pairs of variables) have often been suggested for continuous variables, though they are inefficient in their use of space. For up to about eight variables (the number depending on the screen's resolution and the analyst's eyesight) they can be used for exploratory purposes with linking.

Parallel coordinate displays are primarily for large numbers of continuous variables. Each variable is represented by a vertical axis and individual cases are joined by lines. There is an attractive geometric duality, pointed out by Inselberg (1998), that a point in a scatter plot is represented by a line in a parallel coordinate display and vice versa. Interactive tools are essential for these plots for two reasons: (a) the order of the axes affects what can be seen, so it has to be possible to change the order easily; (b) the large numbers of lines for a dataset of even moderate size makes it essential to be able to temporarily hide points and lines. Categorical variables can also be displayed on their own axes and Wills has suggested using circles whose areas reflect the numbers in a category. Since there is no natural order of categories for most such variables, being able to reorder categories interactively is also valuable. While there is in principle no limit to the number of axes which may be displayed in a scrollable window, it is clear that these displays provide only limited insight to the full multidimensional structure.

4.2 Categorical Variables

Interest in investigating multivariate categorical variables has grown with the rise of data mining. The large datasets which banks, retailers, and other commercial organizations want to study often include many categorical variables. Survey datasets commonly do too, but it is rare to find visualizations of such data. Mosaic plots are one way of uncovering the multivariate structures. Although first discussed in general by Hartigan and Kleiner in 1981, it is only in the late 1990s with the addition of interactive features that they have revealed their true potential (Hofmann 2000). A mosaic plot in its original form is constructed by dividing up the horizontal and vertical axes in turn by each new variable. Figure 9 shows a diagram for the Titanic data set (Dawson 1995). Initially the horizontal axis was divided up according to the numbers in each class, then the individual columns were divided vertically by age, then the resulting cells were each divided horizontally by gender. Finally, the numbers surviving in each group have been highlighted.

This mosaic plot allows the comparison of survival rates by gender within class and age group. The higher survival rates of the female adults compared to the males are obvious. Redrawing with a different order of variables (gender, age, class) permits direct comparisons of survival rates by class within each gender and age group. There are gaps drawn between the cells for two reasons: the display is easier to interpret, and empty cells can be shown (for instance, there were no children in the crew of the Titanic).

Alternative versions of mosaic plots include: (a) the same bin size display, where each cell has the same size;

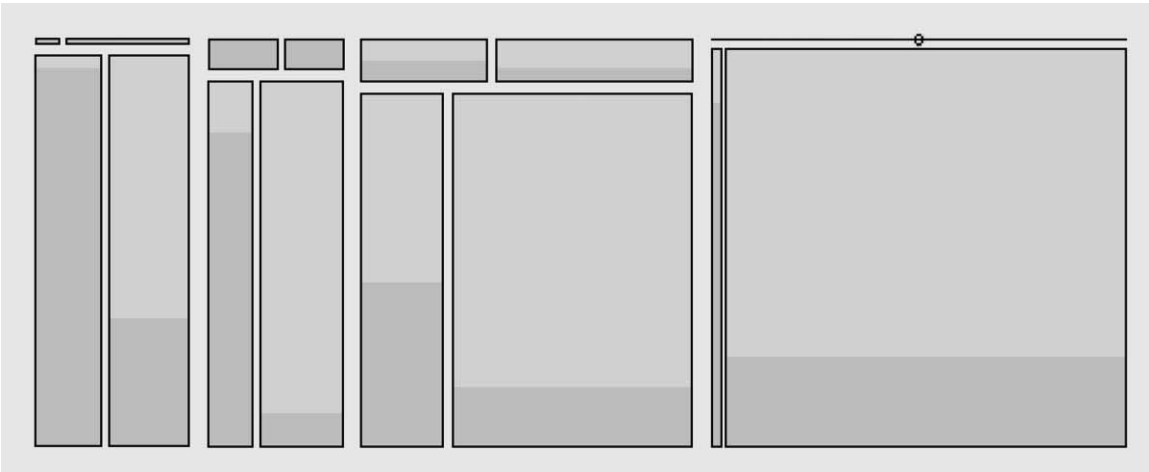

Figure 9
Mosaic plot of the Titanic data in the order class, age, gender with survivors highlighted

this is good for identifying patterns of empty combinations and for comparing highlighted rates for all combinations; (b) fluctuation diagrams, where each cell's size is proportional to the count of the combinations which it represents, but the cell is drawn at a fixed position on the grid. Fluctuation diagrams have the advantage of maintaining a clearer structure, but can be very inefficient in their use of space in comparison to the original format. The alternating of horizontal and vertical axes is an arbitrary restriction and Hofmann has introduced double-decker plots, where only the horizontal axis is divided, as a particularly insightful way of looking at association rules.

One of the most useful approaches for analyzing multivariate data, especially with large datasets, is to study subsets of the data in parallel as in Fig. 7. Trellis displays (Becker et al. 1996) are a more sophisticated version and draw individual displays for each subset in a structured panel display. As in a mosaic plot, the panels are defined by a hierarchical succession of variable splits. Unlike mosaic plots the panels are all of the same size. Each panel may display any one of a number of basic statistical graphics, for example, a box plot or a scatter plot. The art in a successful trellis display lies in the choices of panel plot and of how to split the data set into subgroups. For this reason it would be useful for them to be interactive, but at the time of writing trellis displays are only available in a static form.

4.3 Multivariate Projective Displays

The multivariate displays mentioned thus far show all the raw data. An alternative is to concentrate on projections of the data onto specially chosen lower dimensional spaces. Good projections reveal almost all the information in the data, albeit in the transformed space. Biplots are so called because they show both the projected points and also the axes of the variables (for continuous variables) or the category level points (for categorical variables) in the projection space (Gower and Hand 1996). Figure 10 shows the UCI wine dataset with the second group highlighted. This projection separates the groups fairly well. Long axes show that the corresponding variables are well represented in the projection space. Biplots have many

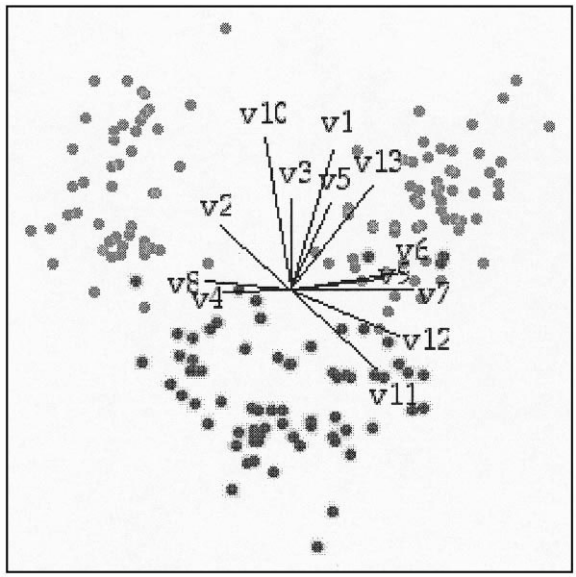

Figure 10
A biplot of the 13 variables and 178 cases of the UCI Wine data set

attractive properties, but remain two-dimensional representations. The larger the number of variables, the less likely it is that a biplot can reveal the major features of a dataset, but the addition of interactive tools enhances the capabilities of biplots substantially (Hofmann 1998).

The first major innovation in exploratory graphics was none of these tools: it was the rotating plot (Fisherkeller et al. 1974 in Cleveland and McGill 1988), in which a succession of two-dimensional images gives the impression of rotation in three dimensions. This is a natural extension of a scatter plot, is excellent for three-dimensional continuous data, and is fun to use, but for higher-dimensional datasets it is not so practical. In that case it can be used in a different way, displaying three-dimensional projections of the data. Many methods of identifying 'interesting' projections have been discussed and the most effective implementation seems to be a directed grand tour. The grand tour was designed to rotate through all possible three-dimensional projections of a dataset of continuous variables and the analyst would decide which of the resulting projections gave him or her insight into the data. Directed grand tours make use of projection pursuit indices to guide the user towards revealing projections (Swayne et al. 1998).

5. *Specialist Displays*

5.1 *Missing Values*

In the histogram in Fig. 2 there was a special column for recording the numbers of cases with no values for that variable. Missing values are a problem for large datasets and any initial overview of the data needs to check for numbers of missings and for any structure among them. Some statistical methods cannot handle cases with missings. A simple approach is the missing value plot as in Fig. 11. There is a bar for each variable. The solid part of the bar on the left represents the cases with values and the blank section on the right the missing values. Bars can be queried and selected to give a rough idea of associations. In the figure it is clear that many cases are missing on a large number of variables.

This kind of display cannot reveal more complex structures and one idea is to use a fluctuation diagram of the corresponding binary variables (missing: yes/no), concentrating on the 10 variables which still have noticeable numbers missing.

5.2 *Weighted Displays*

In datasets arising from surveys or based on geographic areas, it can often be the case that individual data points are of differing importance: the unemployment rate in London should have a greater weight than the unemployment rate in Stratford-upon-Avon. Some authors have suggested using bubbles instead of points in dot plots and scatter plots, where the size of the bubble is proportional to the weighting variable. This can be effective for small numbers of cases, but generally leads to overplotting problems. Area-based displays such as histograms, bar charts, or mosaic plots have no such problems and counts are simply replaced by sums of weights. This can lead to very striking and informative differences in displays. The

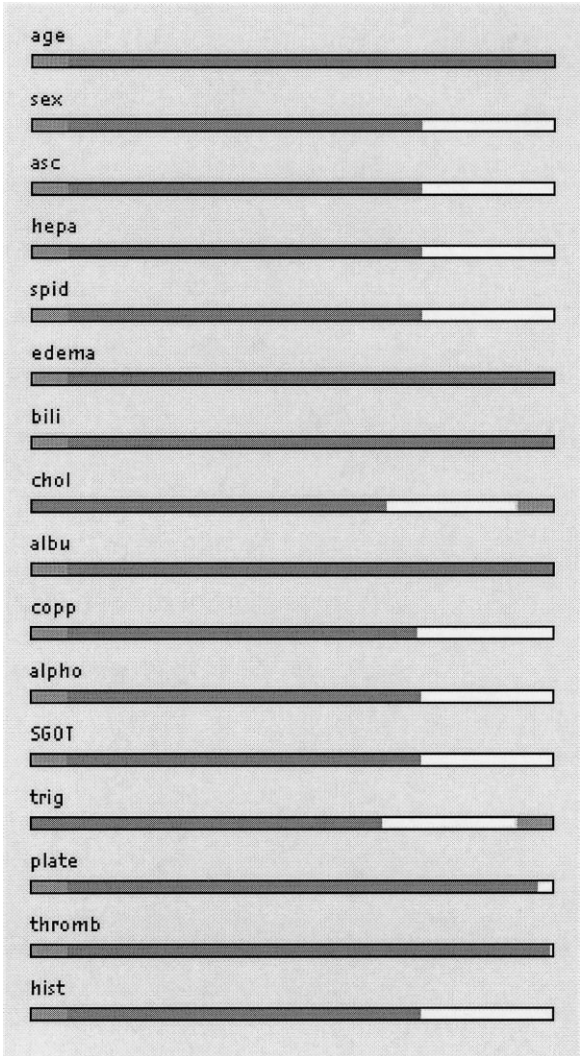

Figure 11
A missing value plot for variables from a patient survival study. A subgroup missing on only two variables is highlighted

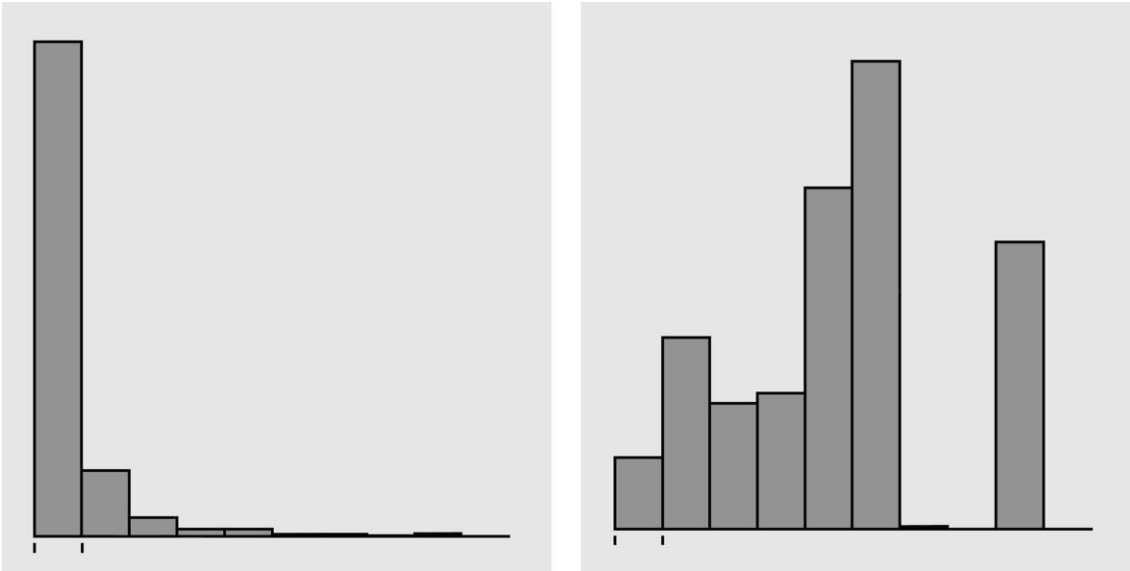

Figure 12
Histograms of black population percentage by US MidWest county, unweighted (left) and weighted by black population (right)

left-hand histogram in Fig. 12 just about shows a single county with the highest percentage of blacks among the over 400 counties in the five US mid-west States. Most counties have low percentages of blacks. The right-hand histogram has the same horizontal scaling, but the height of the bins is now proportional not to the number of counties but to the aggregate black population in the bins. Most blacks live in counties with relatively higher percentages of blacks. It is, of course, necessary that the weighting variable is additive! Population is fine, population density is not (see Unwin and Hofmann 1998).

6. *Statistical Tools and Plots*

Surprisingly, and disappointingly, statistical graphics rarely make use of statistical tools. There are a few exceptions. Some packages offer the option of displaying box plots with confidence intervals around the medians. Some packages offer the chance of plotting error bars round means. SAS have software (*JMP*) where circles are drawn round subgroups as a means to enable graphical detection of whether differences are significant or not. However, the last example illustrates one of the problems. *JMP* offers several different options because there is no agreed optimal way of comparing several means simultaneously. The same problem would arise with bar charts and more seriously with mosaic plots; there are so many possible statistical tests that it is not easy to specify a universally acceptable default. The solution will be to offer the various options interactively.

7. *The Future*

Advances in statistical graphics have not been due to improvements in screen resolution. The best exploratory graphics are simple displays. What makes them powerful is the speed and flexibility with which they can be manipulated and combined. It is the increase in computing power which has been crucial for modern interactive graphics. More powerful computing has also led to the development of computer-intensive analytic procedures and these have sometimes been combined successfully with graphics. Evaluation of projection pursuit indices and rotating plots is one example. Smoothing methods such as Lowess are another.

A major requirement for all statistical methods is that they can be scaled up to deal with huge datasets. Area-based graphical tools such as histograms and bar charts can in principle be used for datasets of any size, but interaction is necessary to catch details and explore links between variables. Much more needs to be done to cope with displays for large numbers

of variables, and interactive capabilities will be essential.

Statistical graphics has grown in fits and starts. There are many attractive ideas, some general principles, but little theory. Real progress will only come with a successful formalization, so that the future can build on the successes of the past. This is something to look forward to in all senses of the phrase.

8. *Software*

The displays in this article have been taken from screenshots. There has been no attempt to pretty them up; they illustrate exploratory analyses in progress. Most are from MANET (Unwin et al. 1996 and www1.math.uni-augsburg.de/Manet/), a couple are from Data Desk (www.datadesk.com). Other software also offers interactive capabilities, but the degree of interactivity varies considerably. For up-to-date information on software, check statistical websites.

Bibliography

Becker R, Cleveland W S, Shyu M-J 1996 The visual design and control of trellis display. *Journal of Computational and Graphical Statistics* **5**: 123–55

Carr D B, Littlefield R J, Nicholson W L, Littlefield J S 1987 Scatterplot matrix techniques for large *N*. *Journal of the American Statistical Association* **82**(398): 424–36

Cleveland W S, McGill M E (eds.) 1988 *Dynamic Graphics for Statistics*. Wadsworth & Brooks/Cole, Pacific Grove, CA

Dawson R J M 1995 The 'unusual episode' data revisited. *Journal of Statistics Education (Online)* **3**(3):

Eick S G, Wills G J 1995 High interaction graphics. *European Journal of Operational Research* 445–59

Gower J C, Hand D J 1996 *Biplots*. Chapman & Hall, New York

Hofmann H 1998 Interactive biplots. In: *NTTS 98* 1. Eurostat, Sorrento, Italy, pp. 127–36

Hofmann H 2000 Exploring categorical data: Interactive mosaic plots. *Metrika* **51**(1): 11–26

Hyndman R J 1996 Computing and graphing highest density regions. *American Statistician* **50**: 120–26

Inselberg A 1998 Visual data mining with parallel coordinates. *Computational Statistics* **13**(1): 47–63

Spence R 2001 *Information Visualization*. Addison-Wesley, New York

Swayne D F, Cook D, Buja A 1998 XGobi: Interactive dynamic data visualisation in the X window system. *Journal of Computational and Graphical Statistics* **7**(1): 113–30

Tierney L 1990 *Lisp-Stat*. Wiley, New York

Tukey J W 1977 *Exploratory Data Analysis*. Addison-Wesley, London

Unwin A R 1999 Requirements for interactive graphics software for exploratory data analysis. *Computational Statistics* **14**: 7–22

Unwin A R, Hawkins G, Hofmann H, Siegl B 1996 Interactive graphics for data sets with missing values—MANET. *Journal of Computational and Graphical Statistics* **5**(2): 113–22

Unwin A R, Hofmann H 1998 New interactive graphics tools for exploratory analysis of spatial data. In: Carver S (ed.) *Innovations in GIS 5*. Taylor & Francis, London, pp. 46–55

Wilkinson L 1999 *The Grammar of Graphics*. Springer, New York

A. Unwin

Graphical Methods: Presentation

Figure 1 shows a graphic of death rates against birth rates per 100,000 population for 27 selected countries in a 1990 UN databank (Wilkinson 1999b). The contours in Fig. 1 show two concentrations of countries. One, to the left, has relatively lower death rates and small-to-moderate birth rates. The second, toward the upper right, has high death rates and extraordinarily high birth rates. The curve in the middle of the contours shows that the overall relation between death and birth rates is curvilinear.

The single plot in Fig. 1 contains three graphs: (a) a point graph (collection of points) whose labels show country names; (b) a set of contours based on a kernel density estimate (Silverman 1986) that represents the concentration of the countries; and (c) a LOESS smoother (Cleveland and Devlin 1988). The graphic also includes three guides for our understanding. The first is a general geometric object called a form that is, in this instance, a line delineating zero population growth. Countries to the left of this line tend to lose population, and countries to the right tend to gain. The other two are guides that demarcate axes for the represented space. Other examples of guides are legends and titles.

Each of the constituent graphical elements in Fig. 1 has a long lineage in the history of statistical charts. Scatterplots of points and contours are several centuries old, and regression smoothers are at least a century old. The use of a rectangular coordinate system to locate graphs within a frame demarcated by axes predates by centuries Decartes' eponymous Cartesian coordinate system. We shall review briefly the history of statistical graphics and charts.

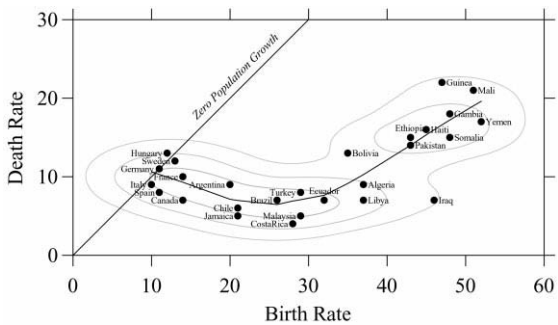

Figure 1
Plot of 1990 death rates against birth rates for selected countries

The lines, colors, typography, proportions, and other aspects of Fig. 1 have been chosen to facilitate communication of the ideas and models underlying the chart. We do not need to resort to platitudes such as, 'A picture is worth a thousand words' to understand that this graphic comprises a relatively substantial analysis of the data. We shall review briefly the topic of graphics principles which can help us devise effective ways to display analyzed data.

The world of charts is enormous. Principles can guide us through that world, but there are many alternatives for representing one or more variables in a graphic. In the final section, we shall present some of these choices and use them to derive additional guidelines for developing effective presentations.

1. History

The contours, or level curves, in Fig. 1 represent a map. This representation is based on a statistical function that maps birth rate and death rate onto a density. Any value on this density can be derived from a corresponding pair of birth rate and death rate values taken from the domain of values within the frame bounded by the two displayed scales. The geometric space in which this mapping is embedded is three-dimensional. The first two dimensions are birth rate and death rate, and the third dimension is density.

The technical mathematical terms in the previous paragraph (contours, level curves, map, function, onto, density, value, domain, frame, bounded, scales, geometric space, and dimensions) are derived from ordinary language, often ancient, describing the physical world. This correspondence between the world of charts and the physical world is not coincidental. Charts are maps of abstract worlds. The word chart and the word cartography have the same root (Latin *charta*, a piece of paper or papyrus). The Greek word for geometry means 'land measurement'. A scale (in Latin) is a ladder. We map in order to organize our world in our mind. However, our world is larger than the domain of geography. As Pinker (1997) has argued, abstract reasoning is built on metaphors for reality. We manipulate abstractions by making them analogous to concrete objects.

Statistical charts are maps, but it does not suffice to simply describe them as abstract geography. The history of statistical charts contains remarkable leaps from a world of continuous rectangular (or spherical) space to other worlds that are categorical (worlds with gaps), multidimensional (worlds beyond three dimensions), and topological (worlds in abstract coordinate systems). Funkhouser (1937), Tilling (1975), Beniger and Robyn (1978), Fienberg (1979), Robinson (1982), Stigler (1983), Tufte (1983, 1990, 1997), Collins (1993), Wainer (1997), and Wainer and Velleman (2001) document these leaps. These references also contain numerous examples of graphics, ancient and modern, that represent significant landmarks in this history. The remainder of this section will highlight a few.

Physical coordinates were used by Egyptian surveyors as early as the third millennium BCE for locating points on land (Beniger and Robyn 1978). By the beginning of the second millennium CE coordinates were used to locate nongeographic entities. Funkhouser (1936) shows an example of a line graphic of planetary movements on a space–time grid produced by an unknown astronomer in the tenth or eleventh century. Not surprisingly, one of the 'planets' is the Sun.

Cartesian coordinates were devised by Descartes for locating points in abstract space, particularly for representing equations underlying his analytic geometry (Funkhouser 1937). However, relatively few seventeenth or eighteenth century embeddings of observed data points in Cartesian coordinates have been found. (Funkhouser 1937, Tilling 1975). Scientists of the Enlightenment sometimes regarded graphics as dilettantish, and if they used them at all, it was for geometrical diagrams or maps of algebraic functions rather than for plots of raw data (Beniger and Robyn 1978). Social scientists of this period focused on collecting tables of state statistics that could be perused in detail without resorting to graphical approximation (Beniger and Robyn 1978, although see Stigler 1983). (Beniger and Robyn (1978) and Wainer and Velleman (2001) discuss notable exceptions to these observations (graphics by Wren, Halley, Huygens, Plot, Lister, Priestley) but conclude that the important varieties of what we now call the statistical chart did not arise until the work of Playfair in the late eighteenth century.

Playfair graphed the data of the state, the root of the word statistics. He invented the pie chart, the circle chart, the categorical bar chart, the area chart, and he adapted other geometric forms in novel ways to graph data over time and space. Wainer and Spence (1997) provide a brief biography of Playfair's fascinating life and the charts he created.

Other widely used graphical forms were developed in the nineteenth and early twentieth centuries. Beniger and Robyn (1978) and Stigler (1983) summarize a number of these. Density bars were used by Guerry in 1833 to display crime by age groupings. Pearson adopted this form for representing a frequency distribution and named it a histogram. Minard developed a number of schematic maps overlaying statistical data on geography. Florence Nightingale introduced polar area charts. The age–sex pyramid (a dual histogram) was introduced by F. A. Walker, Superintendent of the US Census in 1874. The sample cumulative distribution plot was introduced by Galton in 1875.

If charts are maps of abstract worlds, we might conclude that guiding principles for graphics usage could be derived from the psychology of perception and cognition in the physical world. Indeed, recent

psychological research has tended to support the simple statement that the more graphics adhere to the rules of the physical world, the more likely we are to extract and process the information in them, reliably and accurately. Section 3 covers this area.

2. Principles

Over the centuries of their history, various charts have emphasized communication (e.g., Snow's map of cholera deaths in London, reproduced in Tufte 1983), persuasion (Fletcher's map of the distribution of ignorance in nineteenth century England and Wales, reproduced in Wainer 1997), and graphic design (e.g., Playfair's chart of imports and exports of England to and from North America, reproduced in Tufte 1983). These criteria are, of course, not exclusive. Some charts incorporate all three (e.g., Minard's compound chart of Napoleon's troop losses in the Russian campaign, reproduced in Tufte (1983).

Not surprisingly, modern critics of charts have tended to align themselves along these same dimensions. Communicators (e.g., Cleveland, Kosslyn) have conducted experiments to determine rules that improve the transfer of information from chart-maker to chart-viewer. Persuaders (e.g., Holmes 1991) have evaluated charts on the basis of their psychological impact. Designers (e.g., Herdeg 1981) have applied aesthetic criteria. Others (e.g., Tufte 1983) have adopted the criteria of the communicator, and the methods of the designer. None of these approaches can claim exclusive validity. Since scientists are more interested in rules that foster accurate communication of replicable results, however, this section will focus on communication.

Evaluating graphical communication has been the province of psychology, in the subdisciplines of human factors, ergonomics, and applied cognitive science. For almost a century, human factor psychologists have developed methods and criteria for evaluating graphics. Military psychologists, for example, have conducted randomized experiments to test the effectiveness of aircraft cockpit displays.

Psychologists gave much less attention to the perception of statistical graphics until a statistician, William Cleveland, published a series of experiments designed to identity aspects of charts that helped or hindered accurate decoding of quantitative data (Cleveland and McGill 1984). (Beniger and Robyn 1978, cite a number of statisticians' informal studies of the effectiveness of popular charts in the 1920s, but these did not induce many psychologists to study the topic more formally.)

Cleveland used paper-and-pencil tests containing simple graphical elements—points, lines, angles, areas, colors—generated randomly by a computer. His subjects gave magnitude estimates of the values generating the instances of each element. From these estimates, he derived and analyzed error rates. Figure 2 summarizes the main result of these experiments. The top of the figure (BEST) represents elements that resulted in fewer errors and the bottom (WORST) represents those with more.

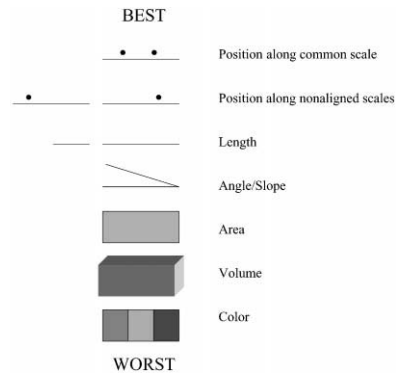

Figure 2
Cleveland (1994) graphical features hierarchy

Cleveland's hierarchy gives us a start in designing effective displays. All other things being equal, we might choose to use a bar chart instead of a pie chart because judgments of position on a common scale and judgments of length are both more accurate than judgments of angle. There are exceptions to these rules, however (see the following discussion of Fig. 3b). A chart is more than the sum of its parts; elements interact in the context of a chart to suppress or enhance errors. Kosslyn (1994) summarizes relevant psychological studies in the extensive notes at the end of his book.

Others have provided numerous general rules for effective graphical communication. The following guidelines are derived from Bertin (1981), Tufte (1983), Cleveland (1994), and Kosslyn (1994).

Avoid clutter. Especially, do not clutter the data region inside the plotting frame demarcated by axes. Tufte constructs a data/ink ratio and urges us to maximize it by removing irrelevant detail. This exhortation has value as long as we do not remove redundant features that can reinforce an accurate perception. Most writers agree that one should avoid jazzy textures (especially stripes), gratuitous colors, excessive tick marks and grid lines, ornate fonts, and unnecessary embellishment. Avoid visual illusions due to pseudo-3D and other gratuitous use of color, angles, area, or volume in a 2D plotting world.

Seek simple geometric forms—straight lines, circles, triangles, squares. When it helps to simplify a relationship, use transformations (logs, square roots) to convert curves to straight lines, for example. Use polar coordinates when variables are cyclical (astronomical perceptual, directional).

Sort and organize. This is especially important for complex graphics. Usually, one should sort the values

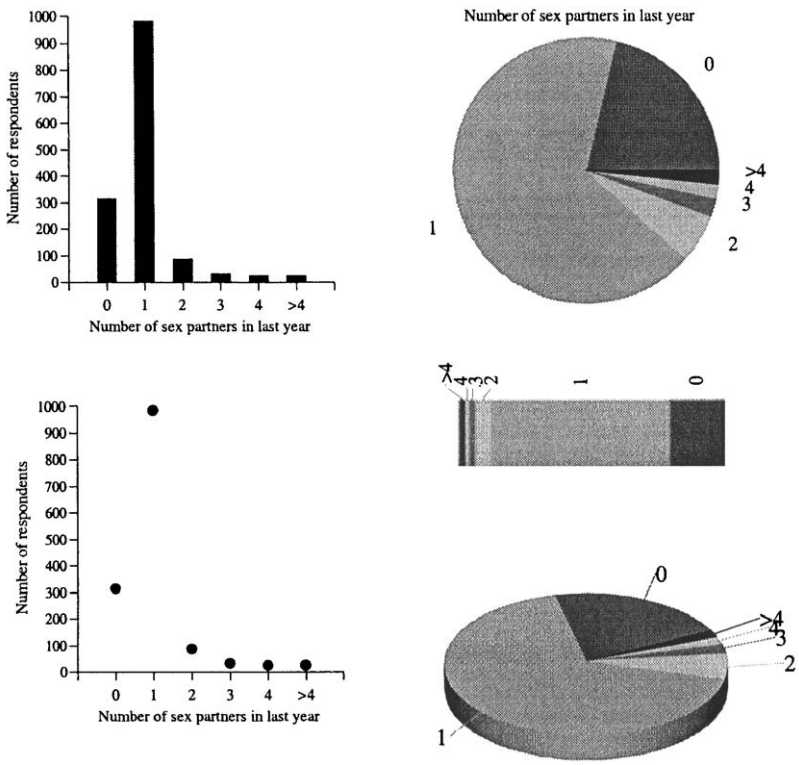

Figure 3
Reported number of sexual partners (General Social Survey, 1993)

of a dependent variable (the range of the graph), but other sorts can help viewers navigate and perceive structure. Bertin (1981) has numerous examples showing the value of simple and multivariate data sorts. Figure 12 shows an example.

Annotate extensively. Make legends comprehensive and informative. Figure captions should describe the source of the data and explain the relation of the data to the graphic. Note exceptions (see Fig. 6).

3. *Choices*

Rules guide, but examples instruct. We need to consider specific choices for specific problems to understand exceptions to rules. The graphic representation problem is often expressed as one of choosing the proper chart type for a given set of data. This is an unhelpful way to express the problem, however. As Wilkinson (1999c) shows, different chart types with their own popular names are often mirror images or simple transformations of each other. Instead, we will examine several basic problems confronting the designer of a scientific graphic: geometry (choosing the type of graphical element for a chart), coordinates (choosing a coordinate system for the chart), and uncertainty (representing random error).

3.1 *Geometry*

The following examples were selected to illustrate principles, not exhaust possibilities. It is organized by the configuration and number of variables, categorical and continuous, that we wish to graph.

3.1.1 One variable. Figure 3 shows five different ways of representing values on one categorical variable. The data are from 1606 respondents to the 1993 General Social Survey (Davis et al. 1993). Respondents were asked, 'How many sex partners have you had in the last 12 months?' Those reporting more than four partners (some reported up to 100) were consolidated into the last category. There were 1,466 responses in the resulting six categories.

The upper left graphic is a bar chart of the frequencies of occurrence of each category. Bars are most suited for displaying either magnitudes referenced against zero (anchored bars, as in this figure) or a range of values on a continuous scale (range bars, as in Fig. 7).

Anchoring bars at zero can sacrifice resolution when no data values exist near zero; this creates substantial white space in the plotting area (Cleveland 1994). In

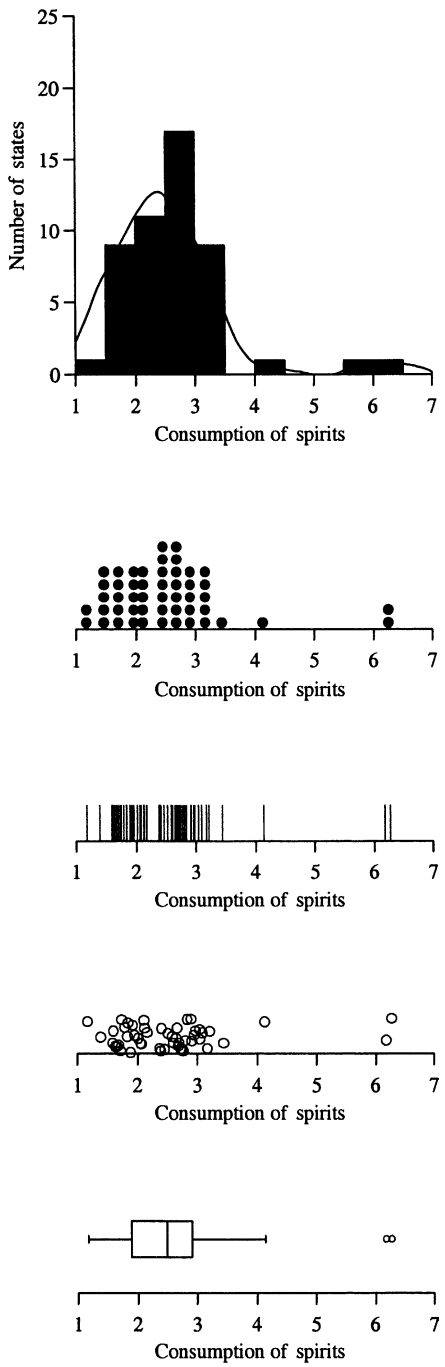

Figure 4
Consumption of spirits in gallons per capita, by US state, 1986

these cases, dots are preferable to bars because we do not need to include zero on the scale. The upper right graphic shows a dot plot of the same variable. In this case, we would prefer the bar version because there are several values near zero and it is easier to anchor the bars visually on the partner categories. The locations of both the dots and the tops of the bars are easily decoded on the vertical scale.

The third graphic in Fig. 3 is a pie chart of the proportions of total category frequencies. The general antipathy statisticians have toward pie charts is misinformed. Pie charts are not always bad, although the issues on when to use them are complex. In general, bar charts are more suited to absolute judgments and pie charts are more suited to judging proportions (Simkin and Hastie 1987). Pie sectors near a fourth or half of a pie in size are judged especially accurately. When there are many sectors, however, pies are not usually appropriate. In this example, the bar chart would be more suitable.

The fourth graphic is a divided bar chart, which is nothing more than a pie chart in rectangular coordinates. Divided bars do worse than pies in studies of graphical perception (e.g., Simkin and Hastie 1987). Cleveland (1994) provides several examples to show why they are generally a bad choice.

The fifth graphic is a 3D pie chart. This is probably the worst of the popular ways to represent proportions. Three-dimensional graphics are not in themselves harmful. What creates the problem with the 3D pie chart is the distortion of angles that the perspective projection produces. In the 3D pie, the single partner category seems to cover almost 75 percent; the actual coverage is 68 percent.

Some users and computer programs apply line or area elements (see Fig. 7) to categorical variables. Popular examples are line graphics of means in analysis of variance, or area graphics of profiles across categories. These elements should generally be avoided with categorical data. Spacings between categories (even ordered categories) are not fixed, so slopes of line segments or profiles between categories are meaningless.

Figure 4 shows five ways of representing one continuous variable. The data are per-capita consumption of spirits for each of the 50 US states (Bureau of the Census 1986). The top graphic is a classic histogram overlaid with a kernel density estimate (Silverman 1986). Histogram bars look like ordinary bars, but they measure areas rather than intervals and they rest on a continuous scale rather than on a set of categories. Histograms are more useful than kernels when we are interested in displaying frequencies within intervals, especially when the intervals themselves are bounded by meaningful, round units. Kernels are more useful than histograms when we are looking for a smooth estimate of a continuous distribution.

The next graphic down is a dot histogram (Chambers et al. 1983, Wilkinson 1999a). This display represents each observation with a dot located at its scale value. If several values coincide on the scale, the

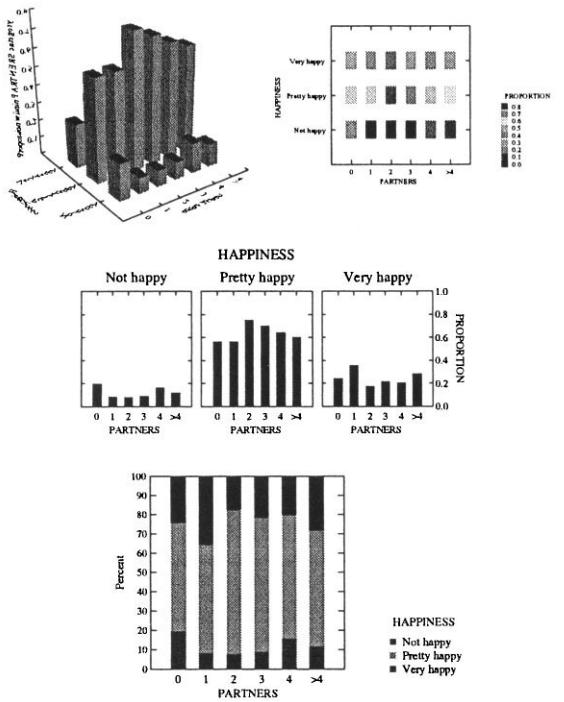

Figure 5
Number of reported sexual partners during previous year by reported level of happiness among 1,462 respondents to 1993 General Social Survey. Third variable is proportion of respondents in each sexual partner category

dots are stacked vertically. This display is most suited for small samples, when every value is to be displayed.

The graphic below that is a stripe plot (Chambers et al. 1983). A vertical stripe occurs at each data value. This display can handle more cases than the dot plot, although it is not suitable for large datasets.

The graphic following is a jittered density (Chambers et al. 1983). Random error is added to each value so that the plotting symbols do not overlap. This display is designed to handle roughly the same conditions as the stripe plot.

The last graphic in Fig. 4 is a box plot, or schematic plot (Tukey 1977). As Tufte (1983) notes, Tukey's plot is based on an earlier display that represents the quartiles of a distribution. In that earlier plot, the whiskers cover the range and the box covers the midrange. Tukey improved on this design by scaling the whiskers to allow for outliers. The display in Fig. 4 shows how important this feature can be. Two states (Nevada and New Hampshire) have unusually large values (see Fig. 6), possibly due to gambling and cheap liquor, respectively. The box plot is a clean representation of the summary values we usually want to see—the median, the twenty-fifth and seventy-fifth

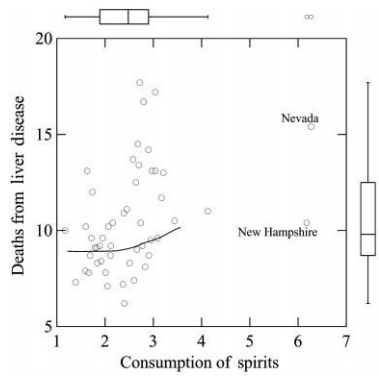

Figure 6
Consumption of spirits in gallons per capita by US state vs. number of deaths from chronic liver disease and cirrhosis per 100,000 by US state, 1986

centiles (hinges), the extreme values (usually bounding the range), and possible outliers (outside values).

3.1.2 Two variables. Figure 5 shows four ways of representing two categorical variables crossed. The data are from the 1993 General Social Survey used in Fig. 3. The additional variable is general happiness, as measured by the response to the question, 'Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?'

The upper left plot shows a 3D bar chart of happiness against number of sexual partners. The vertical scale is the proportion of respondents within each partner category. This graphic has the disadvantage that short bars tend to be hidden by long ones. Generally it should be avoided.

The graphic to the right is a tiled bar chart. This graphic uses color or gray scale to indicate the relative frequency of each category combination. However, as Cleveland's research has shown, the use of color or gray scale to encode continuous values is problematic (Cleveland 1994).

The next graphic below shows a paneled categorical plot. Instead of using an extra dimension in 3D, we use an extra categorical dimension in 2D to panel the plot. This is probably the most effective method for handling the extra dimensionality required for representing the frequencies, plus the two categories. The bottom panel shows a multiple divided bar chart. This layout has the same problems as the single divided bar chart (see Fig. 3). Cleveland (1985) shows an example of this graphic and offers a more effective dot plot alternative.

Figure 6 shows the most popular way of representing two continuous variables crossed with each other—the scatterplot. The data are per capita consumption of spirits for each of the 50 US states, used in

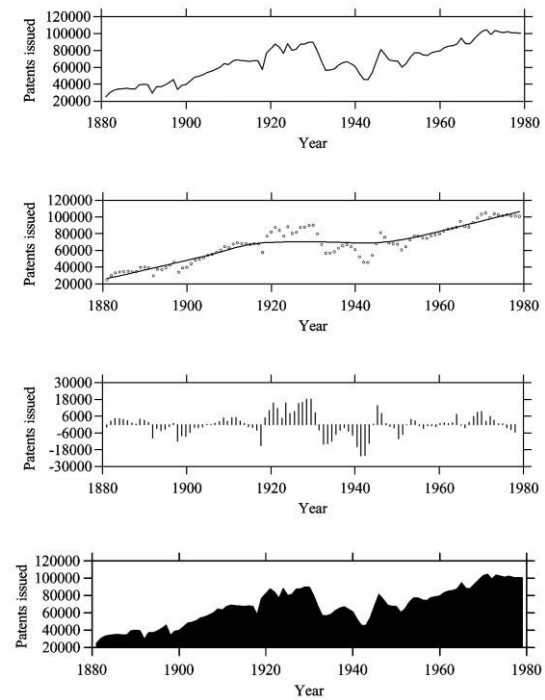

Figure 7
US patents issued from 1880 to 1980

Fig. 4 (Bureau of the Census 1986). The additional variable is number of deaths from chronic liver disease and cirrhosis per 100,000 people, by state. This scatterplot has been enhanced in a number of ways. The bordering box plots help us to assess the marginal distributions of the variables and highlight the Nevada and New Hampshire outliers. The smoother shows the conditional mode of deaths (estimated mode of deaths given consumption). This was computed through kernel regression (Scott 1992). This particular smoother limits itself to areas where there are relatively more data values, so it provides a conservative estimate of trend.

Time series data require special treatment. Figure 7 shows four examples. The data are number of US patents issued in the century from 1880 to 1980 (Wilkinson et al. 1996). The top panel shows a line element representing the series. Lines are most useful when a series is relatively smooth. The next panel shows a point element with a superimposed LOESS smoother. Points are best when we supplement them with a smoother; they tend to obscure time order if used alone, unless the series is especially smooth. The third panel shows spikes. These are vertical lines best used to reveal deviation from a constant level. This panel displays the residuals from the LOESS smooth. The lowest panel shows an area chart for the raw series. This highlights trend but prevents the use of confidence intervals and other enhancements to the plot. Generally, it should be avoided. Cleveland (1994) discusses other graphical representations for time series data.

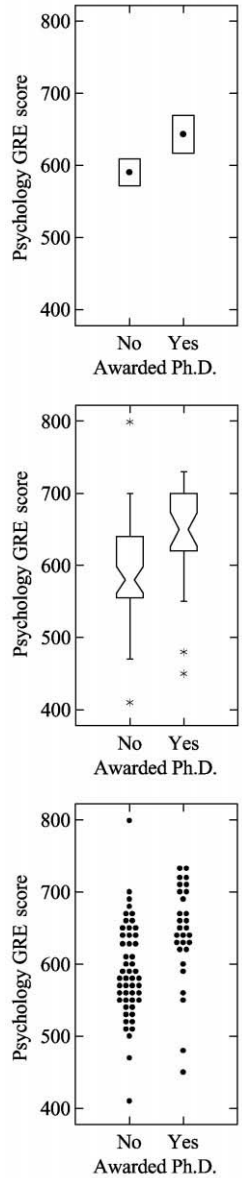

Figure 8
Advanced Psychology Graduate Record Examination (GRE) scores for 80 students taken from 10 years of a graduate psychology program. Students are grouped by whether or not they were awarded a PhD.

Figure 8 shows three ways of representing a categorical variable crossed with a continuous variable.

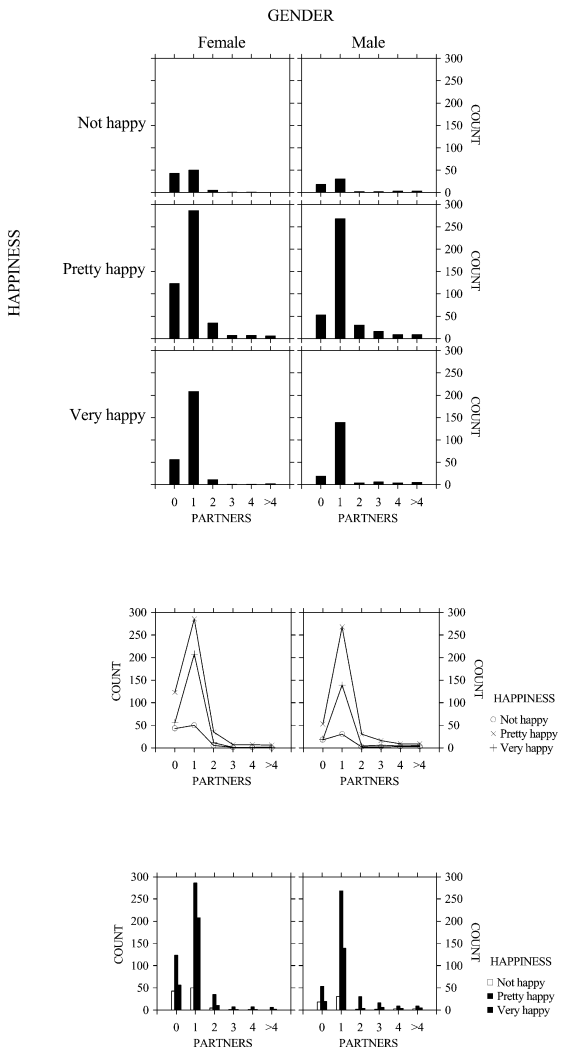

Figure 9
The number of reported sexual partners during last year by reported level of happiness among 1,462 respondents to 1993 General Social Survey. The third variable is the proportion of respondents in each sexual partner category

The data are based on 80 graduate students over a 10-year period in a US psychology department (Wilkinson et al. 1996). Graduate Record Examination Advanced Psychology Test scores are plotted against whether or not the students eventually received their PhD.

The first graphic employs a dot for the mean and a range bar to represent a 95 percent confidence interval on the mean. The bars do not overlap, which suggests that the advanced test can help identify those who achieve a PhD. (Interestingly, neither the Verbal nor the Quantitative test scores predicted graduation in this sample.)

The second graphic is a notched box plot (McGill et al. 1978). This is a variation on the box plot that not only conveys more of the important data landmarks than the classic confidence interval plot but also provides an approximate confidence interval of its own. If the data are independent samples from identically distributed populations that are lumpy in the middle (approximately normal in their midrange), then comparing the notches yields an approximate 95 percent test of the null hypothesis that the true medians in the population are equal. In this example, the notches do not overlap, reinforcing what we concluded from the confidence intervals on the means. Because it relies on the median instead of the mean, the notched box plot procedure is more robust against outliers.

The third graphic is a dot plot (Wilkinson 1999a). Used frequently in the medical literature, this plot shows every data point. In a small sample (e.g., clinical case study), a dot plot can be useful for readers who wish to consider every data value. However, it is of little use in making informal graphical inferences concerning group differences. It is best to think of grouped dot plots as unadorned one-dimensional scatterplots within categories.

3.1.3 Three variables. Figure 9 shows a triple crossing of categorical variables. The data are from the 1993 General Social Survey used in Figs. 3 and 5. The additional variable is gender (observed by the interviewer, not reported by the respondent). We ask whether the relationship between reported happiness and number of sexual partners differs by gender.

The top plot in the figure shows a paneled graphic. The format is similar to a Trellis display (Becker et al. 1996), but the labeling of the paneling variables (happiness and gender) is placed outside the plotting area to improve readability.

The middle plot collapses happiness into a legend in order to reduce the number of panels. This is a popular method for saving space, particularly when representing factorial layouts in ANOVA and other designs. It has several problems. First, the line segments used to highlight trends have slopes that depend on the particular spacing and arrangement of partner categories. Unless we employ an external scaling procedure to determine the spacing of the partners categories, we have scant justification for using these lines. Second, the collapsing introduces a symbol choice problem. It is difficult to find symbols that are easily distinguishable for more than a few categories. The symbols collide, as well, at the upper end of the horizontal scales.

The lower plot introduces an even less appealing alternative. Clustered bar charts are used widely but they have several defects. First, it is difficult to discern

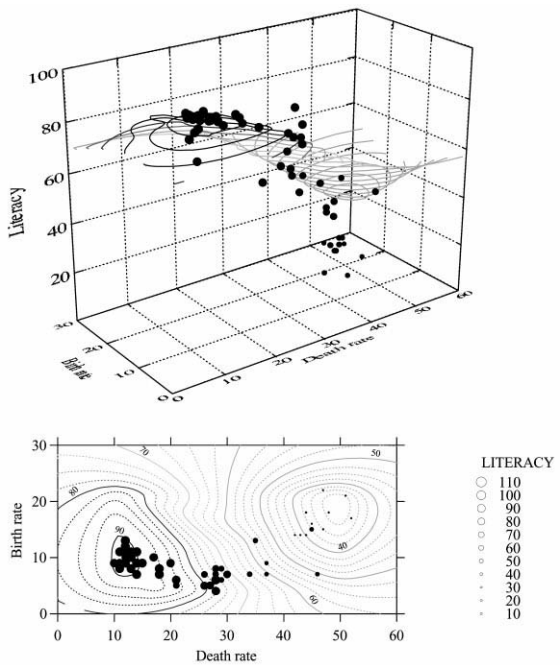

Figure 10
Birth rate and death rate vs. literacy for selected countries

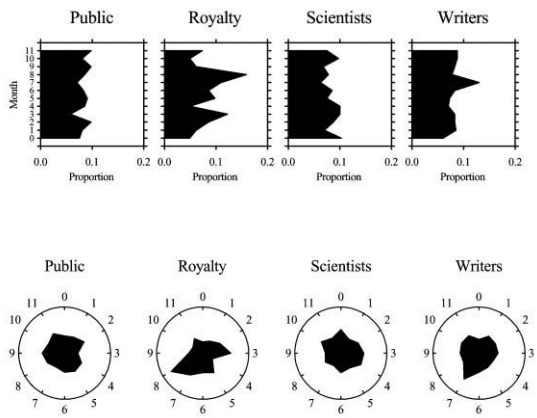

Figure 11
Difference between death month and birthday by occupational category

separate patterns for the categories. One needs to focus on one set of bars to do this, but there is visual interference from the other bars in each cluster. Second, these bars can become quite thin with more than a few categories. Decoding this graphic is problematic.

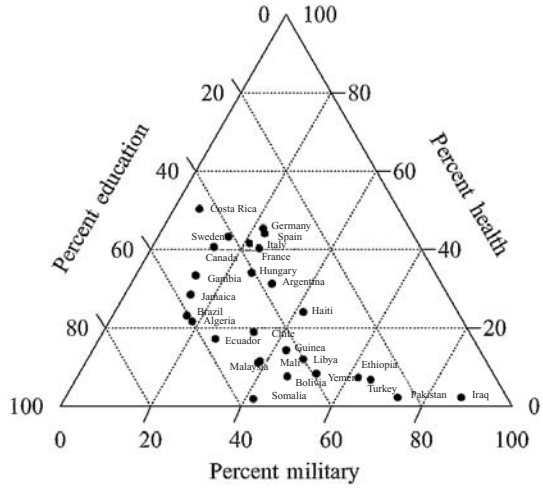

Figure 12
Percentage per-capita spending on education, military, and health

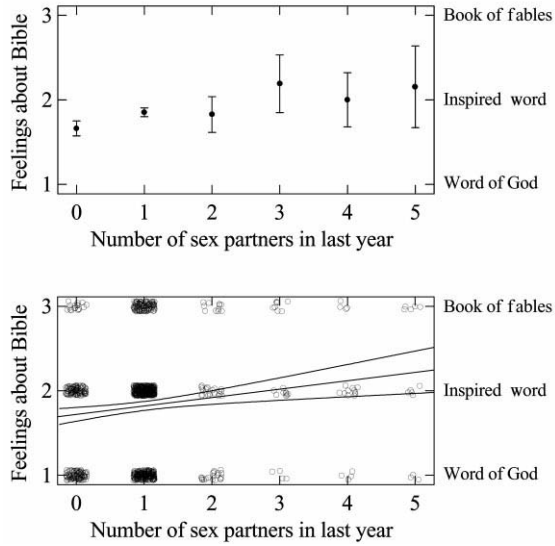

Figure 13
Sex partners vs. feelings about the Bible (GSS survey responses)

Figure 10 shows a triple crossing of continuous variables. The data are from the UN databank used in Fig. 1. The additional variable is literacy (percentage of the population that can read). We ask whether literacy is related to birth and death rates taken together.

The upper plot is a 3D frame that includes a contoured regression surface and a scatterplot. The

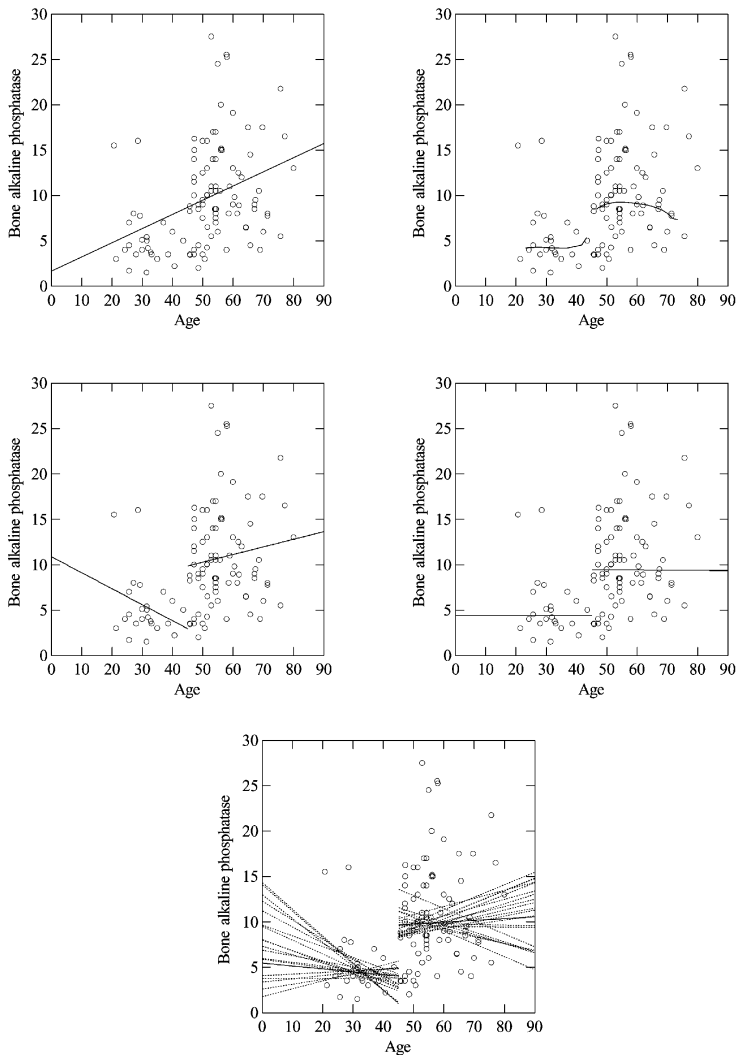

Figure 14
Bootstrapped data from Gonnelli et al. (1996)

contours (as opposed to a wire-frame or rendered surface) help viewers discern literacy levels in different elevations of the plot and keep the surface from hiding the data. Coloring the contours improves the coherence of the surface.

The lower plot is a 2D frame that includes the same contoured surface plus a scatterplot of symbols whose size is proportional to literacy. This plot makes it easier to decode specific *xyz* triplets.

Some critics tend to eschew 3D surface plots, but they have their uses. As Wilkinson (1999c) argues, surfaces elicit a holistic impression of a function. They are less useful for decoding individual values, however. Perhaps the best compromise is to present both displays simultaneously to provide different structural views. Another alternative is to categorize one or more of the variables and display the results in a paneled plot such as those in Fig. 9.

3.2 *Coordinates*

We do not ordinarily think of charts being embedded in different coordinate systems, except for certain scientific graphics such as polar plots. Some chart types, such as radar charts, pie charts, and horizontal bar charts are nothing more than popular charts in unusual coordinate systems. Wilkinson (1999c) provides numerous examples.

Figure 11 shows two different graphics in different coordinate systems, based on data from Andrews and

Herzberg (1985), contributed by C. O'Brien. The data are months since their last birthday during which members of different social groups died. The upper plot shows a paneled horizontal area chart produced by rotating a vertical area chart 90 degrees (exchanging x and y coordinates in this case). This graphic reminds us that horizontal bar charts are a coordinate transformation of a vertical bar chart, not a new chart type.

If it makes sense to consider a death one month before a birthday as being more similar to a death one month after, than to one six months after, then we might want to plot these variables in polar coordinates. The lower plot shows a profile plot of the same results.

Figure 12 contains a triangular coordinate plot. Mosteller and Tukey (1968), Bishop et al. (1975), and Wainer (1997) show examples of this type of plot. The data are from the UN dataset used in Fig. 1. The variables represent expenditures in US dollars per person for health, education, and the military.

The outside tick marks signify how to read the scales. Grid lines collinear with ticks demarcate values on the three scales. Iraq and Pakistan, for example, allocated most to military spending and least to education and health. Costa Rica spent more than 40 percent on education, almost 50 percent on health, and less than 10 percent on the military. While triangular coordinates may be used most frequently in analysis of mixture models in industrial experiments, they can also be useful for social science models involving tradeoffs or compensatory mechanisms.

3.3 Uncertainty

There are two chief methods for representing error in a graphic. The first is sharp: error bounds are represented by clear edges, points, or lines. The second is fuzzy: error is represented by blurring an estimate.

Figure 13 shows two sharp ways of representing error in a graphic. The data are from the 1993 General Social Survey used in Fig. 3. The additional variable is feelings about the Bible, as measured by the response to the question. 'Which of these statements comes closest to describing your feelings about the Bible?' The responses coded are 1 (Word of God), 2 (Inspired word), and 3 (Book of fables). We assume that the dependent variable is continuous (biblical absolutism vs. relativism?) even though the responses are integers.

The upper plot shows error bars representing 95 percent confidence intervals on the means by category. Error bars are also used to represent one standard deviation or one standard error. It is important to make clear in accompanying titles or notes, which type is used. If the spacing of values on the independent variable is meaningful, then we can fit a linear regression model to our data. We shall further assume the number of sex partners to represent values of a continuous variable (promiscuity?). Assuming the assumptions for inference using a t distribution are appropriate, we can fit continuous confidence bounds on the regression line as in the lower panel of the figure. This plot uses jittering to reveal the concentration of data points where they are tied.

Figure 14 shows one way to represent error by fuzziness. The data are from Gonnelli et al. (1996). They represent concentration of bone alkaline phosphatase (BAP) in a sample of women of different ages. The authors fit a linear regression (shown in the upper left panel) to argue that BAP levels increase with age.

A modal regression in the upper right panel indicates that there is a discontinuity in this relationship at age 45 years or so, corresponding most likely to the onset of menopause. Accordingly, we fit separate linear regressions to the two subgroups split at age 45 years (third panel). These regressions appear to be sensitive to outliers, however, so we fit linear models using robust regression with t weighting (fourth panel). This fit indicates that a plausible model for predicting BAP from age involves level differences but no slope differences.

It is not easy to compute confidence intervals on the robust linear fits, so we resort to bootstrapping to provide an estimate of error. The bottom panel shows the result of 20 bootstrap robust fits displayed as faint dashed lines. The non-overlapping envelopes of the fits indicates that our level-change model is reasonable.

4. Conclusion

Figure 15 shows the principal result of a survey by Cleveland (1984) on the use of graphics in scientific articles. The horizontal axis represents the percentage of total page area devoted to graphs in 47 articles sampled from the 1980–81 volumes of selected journals. As Cleveland noted, the substantial differences among disciplines remain when one examines number of graphs per article instead of page area.

In Cleveland's survey, social science journals ranked lowest in the use of graphics. We might speculate that this is a legacy of the aversion toward graphics among

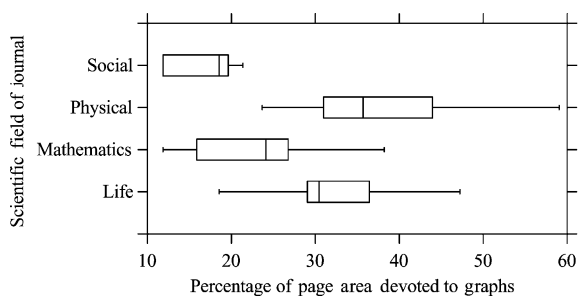

Figure 15
Use of graphics in journal articles (data from Cleveland 1984)

eighteenth century social scientists that Beniger and Robyn (1978) describe. Whatever the reason, there is little justification for this to continue. Software packages such as S-Plus, SAS, SPSS, and SYSTAT (used for the graphics in this article) now offer statistical graphics and publishing standards such as Adobe PDF format now make it possible to display graphics as easily as text.

See also: Bayesian Graphical Models and Networks; Graphical Models: Overview

Bibliography

Andrews D F, Herzberg A M 1985 *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag, New York

Becker R A, Cleveland W S, Shyu M-J 1996 The design and control of Trellis display. *Journal of Computational and Graphical Statistics* **5**: 123–55

Beniger J R, Robyn D L 1978 Quantitative graphics in statistics: A brief history. *The American Statistician* **32**: 1–11

Bertin J 1967 *Semiologie Graphique*. Editions Gauthier-Villars, Paris [English translation Berg W J, Wainer H 1983 *Semiology of Graphics*. University of Wisconsin Press, Madison, WI]

Bertin J 1981 *Graphics and Graphic Information-Processing*. [English translation Berg W J, Scott P. Walter de Gruyter, New York]

Bishop Y M, Fienberg S E, Holland P W 1975 *Categorical Data Analysis*. MIT Press, Cambridge, MA

Bureau of the Census 1986 *State and Metropolitan Area Data Book*. US Government Printing Office, Washington, DC

Chambers J M, Cleveland W S, Kleiner B, Tukey P A 1983 *Graphical Methods for Data Analysis*. Wadsworth, Monterey, CA

Cleveland W S 1984 Graphs in scientific publications. *The American Statistician* **38**: 19–26

Cleveland W S 1993 A model for studying display methods of statistical graphics (with discussion). *Journal of Computational and Graphical Statistics* **2**: 323–43

Cleveland W S 1994 *The Elements of Graphing Data*, rev. edn. Hobart Press, Summit, NJ

Cleveland W S 1995 *Visualizing Data*. Hobart Press, Summit, NJ

Cleveland W S, Devlin S 1988 Locally weighted regression analysis by local fitting. *Journal of the American Statistical Association* **83**: 596–640

Cleveland W S, McGill R 1984 Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* **79**: 531–54

Collins B M 1993 Data visualization—has it all been seen before? In: Earnshaw R A, Watson D (eds.) *Animation and Scientific Visualization: Tools and Applications*. Academic Press, New York, pp. 3–28

Davis J A, Smith T W, Marsden P V 1993 *The General Social Survey*. National Opinion Research Center, Chicago

Fienberg S 1979 Graphical methods in statistics. *The American Statistician* **33**: 165–78

Funkhouser H G 1936 Note on a tenth century graph. *Osiris* **1**: 260–2

Funkhouser H G 1937 Historical development of the graphical representation of statistical data. *Osiris* **3**: 269–404

Gonnelli S, Cepollaro C, Montagnani A, Monaci G, Campagna M S, Franci M B, Gennari C 1996 Bone alkaline phosphatase measured with a new immunoradiometric assay in patients with metabolic bone diseases. *European Journal of Clinical Investigation* **26**: 391–6

Herdeg W 1981 *Graphis Diagrams: The Graphic Visualization of Abstract Data*. Graphis Press, Zürich, Switzerland

Holmes N 1991 *Designer's Guide to Creating Charts and Diagrams*. Watson-Guptill Publications, New York

Kosslyn S M 1994 *Elements of Graph Design*. W. H. Freeman, New York

McGill R, Tukey J W, Larsen W A 1978 Variations of box plots. *The American Statistician* **32**: 12–6

Mosteller F, Tukey J W 1968 Data analysis, including statistics. In: Lindzey G, Aronson E (eds.) *Handbook of Social Psychology*, 2nd edn. Addison-Wesley, Reading, MA, pp. 80–203

Pinker S 1997 *How the Mind Works*. Norton, New York

Robinson A H 1982 *Early Thematic Mapping in the History of Cartography*. University of Chicago Press, Chicago

Scott D W 1992 *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York

Silverman B W 1986 *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London

Simkin D, Hastie R 1987 An information processing analysis of graph perception. *Journal of the American Statistical Association* **82**: 454–65

Stigler S 1983 *The History of Statistics*. Harvard University Press, Cambridge, MA

Tilling L 1975 Early experimental graphics. *The British Journal for the History of Science* **8**: 193–213

Tufte E R 1983 *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT

Tufte E R 1990 *Envisioning Data*. Graphics Press, Cheshire, CT

Tufte E R 1997 *Visual Explanations*. Graphics Press, Cheshire, CT

Tukey J W 1977 *Exploratory Data Analysis*. Addison-Wesley, Reading, MA

Wainer H 1997 *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. Springer-Verlag, New York

Wainer H, Spence I 1997 Who was Playfair? *Chance* **10**: 35–7

Wainer H, Velleman P F 2001 Statistical graphs: Mapping the pathways of science. *The Annual Review of Psychology* **52**

Wilkinson L 1999a Dot plots. *The American Statistician* **53**: 276–81

Wilkinson L 1999b *SYSTAT*. SPSS, Chicago

Wilkinson L 1999c *The Grammar of Graphics*. Springer-Verlag, New York

Wilkinson L, Blank G, Gruber C 1996 *Desktop Data Analysis with SYSTAT*. Prentice-Hall, Upper Saddle River, NJ

L. Wilkinson

Graphical Models: Overview

1. Some General and Historical Remarks

Graphical models aim to describe concisely the possibly complex interrelationships between a set of variables. Moreover, from the description key, proper-

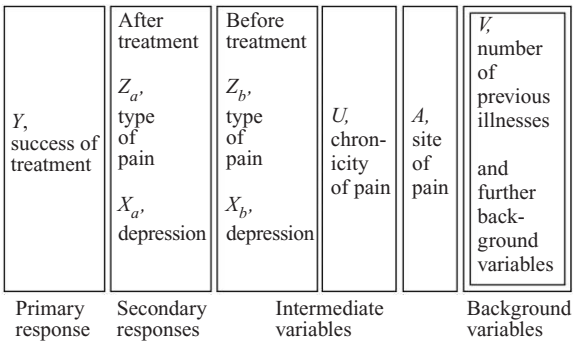

Figure 1
Initial partial ordering of variables in chronic pain study for a multivariate regression chain. Variables in the same box are treated on an equal footing. Variables in any one box are considered conditionally given all variables in boxes listed to their right. Associations among background variables in the double-lined box are not analyzed but taken as given

ties can be read directly. The central idea is that each variable is represented by a node in a graph. Any pair of nodes may be joined by an edge. For most types of graph a missing edge represents some form of independency between the pair of variables. Because the independency may be either marginal or conditional on some or all of the other variables, a variety of types of graph are needed.

A particularly important distinction is between directed and undirected edges. In the former an arrow indicates the direction of dependency from an explanatory variable to a response. If, however, the two variables are to be interpreted on an equal footing then an edge between them is undirected, or cyclic dependencies are permitted. For instance, systolic and diastolic blood pressure would typically be treated on an equal footing because they are two aspects of a single phenomenon, namely the blood pressure wave.

Graphical models started to be developed by Darroch et al. (1980) and Wermuth (1976) as special subclasses of loglinear models for contingency tables and of multivariate Gaussian distributions which are interpretable in terms of conditional independencies and can be represented by undirected graphs. See *Multivariate Analysis*: *Overview*; *Multivariate Analysis*: *Discrete Variables* (*Loglinear Models*).

The first extension was to problems in which the variables can be arranged in sequence so that each variable is regarded as a response to all variables to the right of it in the sequence. This leads to representation by a directed acyclic graph. Then there are joint response models with graphs having blocks of undirected components, the blocks being again connected in a directed and acyclic way. Much recent and ongoing research is for more specialized situations such as models for confounded dependencies, for event history data, for time series data, and for simultaneous dependencies in multivariate Gaussian distributions. The latter correspond, for instance, to independence graphs having directed cyclic components.

As an example of some of the above ideas we report here on a study concerning chronic pain. One main research question is: when investigating the development of this illness, is it necessary to include information on the patient's site of main pain? Figure 1 shows a first partial ordering of a set of variables within boxes. It reflects some of the prior knowledge on the development of chronic pain and its treatment. It also determines important aspects of the types of analysis to be carried out: relations among several background variables are taken jointly as given, univariate conditional distributions are studied when there is a single variable within a box, and joint conditional distributions are studied when there are several variables within a box. We return to a graph resulting from such statistical analyses in Sect. 6

We concentrate in this article on the use of models for the analysis and interpretation of empirical data, and mention some of their other uses in Sect. 7 on suggested further reading. The more recent statistical work quoted here by authors only is referenced in Edwards (2000), Studený and Bouckaert (1998), Richardson (1998), or Wermuth (1998).

2. *Models for Contingency Tables and the Need for Extensions*

For a q-dimensional contingency table, i.e., for counts on q categorical variables, the most complex undirected graph model has no constraints. It is called the saturated model and is represented by a graph in q vertices or nodes, with each node pair having exactly one edge, i.e., being connected by a full line. Each full line represents the conditional association of the connected variable pair given all remaining $q-2$ variables. Removing a line from the graph introduces a particular independence constraint for the corresponding variable pair: the pair are to be conditionally independent given all remaining variables. In a loglinear model formulation this is achieved by setting the two-factor and all higher-order interaction terms involving this variable pair to zero. If all full lines are removed, the graph of q disconnected nodes results. It represents the simplest model in the class, the one of mutual independence of all q variables.

This development built on previous work (quoted in monographs on graphical models) by M. S. Bartlett, M. W. Birch, L. A. Goodman, Y. Y. Bishop, S. E. Fienberg and P. Holland and much earlier work by A. A. Markov, who used the notion of conditional independence around 1900 to formulate simple multivariate models, now called Markov chains. A Markov

chain for q categorical variables can be represented by an undirected graph which is a single path of full lines, i.e., a sequence of nodes $i_1, i_2, \ldots, i_q$ where just each consecutive pair is connected by a full line.

Compared to the larger class of loglinear models the graphical models are attractive for three main reasons: (a) the interpretation of a model can be much simpler if independencies are taken into account compared to only loglinear interactions; (b) parameter estimation can be strengthened by basing it on a sequence of possible small contingency tables which depends on the unique decomposition of a given graph into its prime graphs, i.e., into graphs which in a graphical sense cannot be split any further; and (c) a powerful separation criterion permits one to read all independence statements implied by a model directly off the graph.

This separation criterion for undirected graphs takes any three nonoverlapping subsets a, b, c of nodes in the graph of which the set c may be empty. Then the corresponding variables X_a are conditionally independent of X_b given X_c if every path from a to b has a node in c. This holds provided two conditions are satisfied: all pairwise independence statements specified with the missing edges of the undirected graph hold, and a technical positivity condition is fulfilled. One sufficient condition for checking the latter from an observed contingency table is that there are no zero frequencies.

In graphical models represented by undirected graphs all variables are treated on an equal footing, i.e., response variables and possible explanatory variables are not distinguished. This is appropriate for various different types of data, for instance for symptoms of a given disease, for items which are to measure slightly varying aspects of a particular attitude or of a behavior of persons, or for several aspects which all might contribute to a specific risk factor. However, for much multivariate research some response variables are of primary interest. There is in addition a set of background or context variables, and there are possibly sequences of intermediate variables which play the role of potential response variables to some and of potential explanatory variables to other variables under investigation. Typically, not all variables are only categorical; some are nominal, some are ordinally scaled, and still others are quantitative measurements. Subsets of variables may have to be considered as joint responses instead of univariate responses.

3. *Questions Regarding Applications and Statistical Research*

The original formulation of graphical models has already been extended in various directions to accommodate the need posed by different substantive research questions, to describe properties of different classes of independence models, and to prove corresponding results.

For research in different substantive fields such as genetics, economics, social, and life sciences, some important questions arising in applying graphical Markov models are:

(a) How do graphical Markov models relate to techniques applied traditionally in the given field? What additional possibilities do they offer to gain better insight into any given research question?

(b) How can a given graphical Markov model be best fitted to data? Is software available which permits fast application as well as comparison and integration into existing techniques and strategies of data analysis?

(c) Which of the associated new results in statistical theory are most useful for the substantive research at hand? Are case studies available in which graphical Markov models have been used fruitfully?

(d) Are criteria available to decide among alternative types of model to be considered for a given set of data?

The answers must depend partly on the specific substantive research questions. Also, the kinds of research questions which may be analyzed in the context of graphical Markov models widen as more theoretical results become available, and as new classes of model are developed. But many positive answers have become available, especially since, as described in Sect. 2 for general loglinear models, representations interpretable in terms of conditional independencies often result as a subclass of widely used tools for data analysis.

The following are some of the associated research questions in statistical theory.

(a) Which sets of independence statements can be specified within different classes of independence graphs? In particular, what are the defining independencies, i.e., what are (possibly alternative) sets of independence statements by which a given graph can be characterized? Which independence statements are implied in addition to those in a defining set?

(b) Under what conditions are two different graphs independence-equivalent; that is, when do they characterize the same independence structure, so that they imply an identical list of independence statements?

(c) Under what distributional assumptions do all models in a given class of independence graphs exist? When are they not only independence-equivalent but when are they also distribution-equivalent; that is, when do they define the same joint distribution?

(d) For which graphical Markov models do unique maximum-likelihood estimates exist? With which types of algorithm can maximum-likelihood estimates or parameter estimates in posterior distributions be obtained? When can large estimation problems be decomposed into smaller, simpler ones? When can complicated estimation problems be embedded in larger, simpler ones?

(e) Can the implied independence structure be predicted that results, for instance, after marginalizing over some of the variables and after conditioning on

other variables? When can even the strength and direction of implied conditional associations be derived explicitly?

(f) How can deterministic transformations of variables be utilized systematically in model building? Should definitions of variables, i.e., measurement questions, be integrated into the formulation of a graphical Markov model, or should they be introduced and studied separately?

While the last set of questions involve ongoing research, at least partial answers are available to all the others for most subclasses of graphical Markov models known until now. Here we give only examples of results, and point to the researchers who developed them.

4. Types of Independence Graph and Corresponding Markov Models

As mentioned before, there are a number of different types of graph which differ with respect to their defining sets of independencies. To describe them we denote by Y_V a random vector variable of d_V individual components. A graph for the node set V captures an independence structure for the joint distribution of all d_V components. For nonoverlapping subsets S and C of V, with S having at least two nodes, a graph for the node set S gives an independence structure for the joint conditional distribution of Y_S, given Y_C. When C is empty the graph is for the joint marginal distribution of Y_S. Conditional independence of Y_i and Y_j, given Y_C, is written more compactly as $i \perp\!\!\!\perp j \mid C$.

For each member of the following types of graph, there exists the corresponding independence structure for mixtures of discrete and continuous variables in some exponential family distribution, defined either locally for conditional Gaussian regressions, or globally for a general conditional Gaussian distribution or for one having no higher than two-factor interactions. For simplicity we illustrate here the different types of model only for a mean-centered vector variable Y_V having a joint Gaussian distribution and being partitioned into at most three components $V = (a, b, c)$. The overall covariance matrix as well as its inverse, the overall concentration matrix, are written accordingly, partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \cdot & \Sigma_{bb} & \Sigma_{bc} \\ \cdot & \cdot & \Sigma_{cc} \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} \Sigma^{aa} & \Sigma^{ab} & \Sigma^{ac} \\ \cdot & \Sigma^{bb} & \Sigma^{bc} \\ \cdot & \cdot & \Sigma^{cc} \end{pmatrix} \tag{1}$$

where the dots denote symmetric entries.

The concentration matrix in the marginal distribution of Y_c is denoted by $\Sigma^{cc \cdot ab}$, and the one in the joint marginal distribution of Y_b and Y_c has components $\Sigma^{bb \cdot a}$, $\Sigma^{bc \cdot a}$, $\Sigma^{cc \cdot a}$. In the latter, the notation reminds one that marginalizing is over a. Similarly, the conditional covariance matrix of Y_a is denoted by $\Sigma_{aa \cdot bc}$, and the conditional covariance matrix of Y_a and Y_b given Y_C has components $\Sigma_{aa \cdot c}$, $\Sigma_{ab \cdot c}$, $\Sigma_{bb \cdot c}$. Here the notation reminds one that conditioning is on c. Finally, the matrix of regression coefficients of Y_c when regressing Y_b on Y_c is $\Pi_{b|c}$, while in the regression of Y_a on both Y_b and Y_c the regression coefficient matrix is in partitioned as $(\Pi_{a|b.c}, \Pi_{a|c.b})$. For example, $\Pi_{a|b \cdot c}$ are the coefficients of Y_b. The notation reminds one that regression is also on c.

4.1 Directed Acyclic Graph

A *directed acyclic graph* of Y_V is a graph of arrows in d_V nodes without directed cycles, i.e., starting from any one node it is impossible to return to this node by following any path in the direction of the arrows. The (i, j) arrow is missing in it if

$$i \perp\!\!\!\perp j \mid \text{ parents of } i. \tag{2}$$

Nodes from which an arrow points directly to node i are called the parents of i.

For a Gaussian distribution, the independencies show as zero coefficients in recursive linear equations having independent residuals. More precisely, we have equations $AY_V = \varepsilon$, with residuals ε having zero mean and covariance matrix, cov (ε), being a diagonal matrix T. Further, A is upper triangular with ones along the diagonal, giving $\Sigma^{-1} = A^T T^{-1} A$. Therefore (A, T^{-1}) is a triangular decomposition of the concentration matrix and off-diagonal elements in row i and A are negative values of regression coefficients when regressing Y_i on all its potentially explanatory variables, i.e., variables ordered to have indices larger than i. Thus, the independencies show as zeros in A. If only a directed acyclic graph is given, typically more than one set of recursive equations is compatible with it, i.e., with several different orderings of the variables, an upper triangular matrix can result which reflects the same independence structure as the given graph.

However, often a unique full ordering is provided from substantive knowledge about how the data could have been generated. This was how the geneticist S. Wright introduced and used path diagrams for recursive linear relations. For other than linear relations, a generating process in terms of univariate conditional distributions still determines uniquely the edge matrix in a directed acyclic graph and a corresponding factorization of the joint density. Separation criteria for directed acyclic graphs apply therefore to any joint distribution generated in this way over a directed acyclic graph. A first path criterion was given by J. Pearl, who called it d-separation (for separation in directed graphs). Two other equivalent criteria are due to S. L. Lauritzen and coworkers, and D. R. Cox and N. Wermuth.

4.2 Types of Undirected Graph

A *conditional concentration graph* of Y_S, given Y_C, is an undirected graph of full lines in d_S nodes. The full (i, j) line is defined to be missing in it if

$$i \perp\!\!\!\perp j \mid C, S \quad \text{excluding } i, j. \tag{3}$$

For a Gaussian distribution, the independencies show as zeros in the conditional concentration matrix $\Sigma^{-1}{}_{SS.C}$. Such models have been introduced by A. P. Dempster under the name of covariance selection. The name concentration graph model for Gaussian variables is more common now. It records that for Gaussian distributions the above independence statement is equivalent to the vanishing of a concentration, which is a multiple of the partial correlation of Y_i and Y_j given all remaining variables in S and C.

The name concentration graph model does not imply that for general distributions concentrations vanish, but only that the distribution satisfies a set of independence statements of the above type. The separation criterion in Sect. 2 holds for such general concentration graph models and was given by J. Darroch, S. L. Lauritzen and T. P. Speed.

A *conditional covariance graph* of Y_S, given Y_C, is an undirected graph of dashed lines in d_S nodes. The dashed (i, j) line is missing in it if

$$i \perp\!\!\!\perp j \mid C. \tag{4}$$

For a Gaussian distribution, the independencies show as zeros in the conditional covariance matrix $\Sigma_{SS.C}$. The models are a special case of T. W. Anderson's linear in covariance models. They may not have a unique maximum likelihood estimate.

In general a covariance graph structure can capture all independencies in a joint distribution only if it has, just as a Gaussian distribution, no higher than two-factor interactions. For binary variables this holds, for instance, for the quadratic binary exponential distribution. For mixed variables it can be achieved for the conditional Gaussian distribution with all three and higher-order interactions being zero. Separation criteria were given by G. Kauermann and by Richardson (2001).

4.3 Types of Chain Graph

A *chain graph* of Y_V is a graph in d_V nodes which may be arranged in a sequence of ordered boxes $(1,\ldots, k, k+1,\ldots, K)$ such that there are undirected edges within boxes and arrows between boxes all pointing in one direction, e.g., from k to $1,\ldots, k-1$; the arrows form no directed cycle, i.e., starting from any one box it is impossible to return to this box by following the direction of the arrows.

A *chain graph of block regressions* contains full lines within boxes and full arrows between boxes. The full (i, j) line is missing within box k or the full (i, j) arrow pointing from a node in one of the boxes $k+1,\ldots, K$ to a node in box k is missing if

$$i \perp\!\!\!\perp j \mid \text{ all nodes in boxes } k,\ldots, K \text{ excluding } i, j. \tag{5}$$

For three chain components in a Gaussian distribution, each independence statement corresponds to a zero entry (possibly symmetric) in one of the following concentration matrices: Σ^{aa}, Σ^{ab}, Σ^{ac}, $\Sigma^{bb\cdot a}$, $\Sigma^{bc\cdot a}$, $\Sigma^{cc\cdot ab}$. The name block regression arises in spite of this concentration matrix representation since dividing all elements in row i of, say $(\Sigma^{bb\cdot a}, \Sigma^{bc\cdot a})$, by the element in position (i, i), a transformation into corresponding block-regression coefficients results. Two different equivalent separation criteria were given by M. Frydenberg and by M. Studený.

Models of this type for the special kind of model for mixtures of continuous and discrete variables called CG-regressions are known to have unique maximum likelihood estimates, but it took more than 20 years before an algorithm was made available by D. Edwards and S. L. Lauritzen that converges in reasonable time.

A *chain graph of multivariate regressions* contains dashed lines within boxes and dashed arrows between boxes. The dashed (i, j) line is missing within box k if

$$i \perp\!\!\!\perp j \mid \text{ all nodes in boxes } k+1,\ldots, K \tag{6}$$

the full (i, j) arrow pointing from a node in one of the boxes $k+1,\ldots, K$ to a node in box k is missing if

$$i \perp\!\!\!\perp j \mid \text{ all nodes in boxes } k+1,\ldots, K \text{ excluding } j. \tag{7}$$

For three chain components in a Gaussian distribution each independence statement corresponds to a zero entry (possibly symmetric) in one of the following matrices of covariances or of regression coefficients: $\Sigma_{aa\cdot bc}$, $\Pi_{a|b\cdot c}$, $\Pi_{a|c\cdot b}$, $\Sigma_{bb\cdot c}$, $\Pi_{b|c}$, Σ_{cd}. A separation criterion was given by N. Wermuth and D. R. Cox and by T. S. Richardson.

A *chain graph of concentration regressions* is a mixture of a concentration graph and a multivariate regression graph. It contains full lines within boxes and dashed arrows between boxes. The full (i, j) line is missing within box k if

$$i \perp\!\!\!\perp j \mid \text{ all nodes in boxes, } k,\ldots, K \text{ excluding } i, j \tag{8}$$

the dashed (i, j) arrow pointing to box k is missing if

$$i \perp\!\!\!\perp j \mid \text{ all nodes in boxes } k+1,\ldots, K \text{ excluding } j. \tag{9}$$

For three chain components in a Gaussian distribution, each independence statement corresponds to a zero entry (possibly symmetric) in one of the following matrices of concentrations or of regression coefficients: Σ^{aa}, $\Pi_{a|b\cdot c}$, $\Pi_{a|c\cdot b}$, $\Sigma^{bb\cdot a}$, $\Pi_{b|c}$, $\Sigma^{cc\cdot ab}$. For

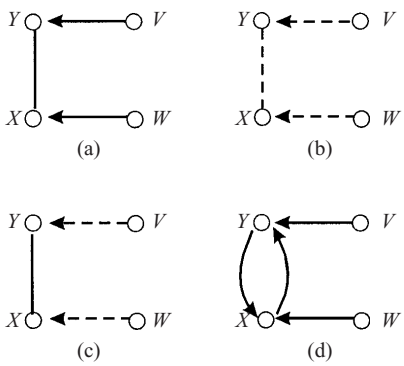

Figure 2
Simple examples of the four types of joint response graphs in Sects. 4.3 and 4.4. In each the joint responses are Y and X, and the explanatory variables are marginally independent: $V \perp\!\!\!\perp W$. The only additional independencies represented by each graph are in the (a) chain graph of block regressions: $Y \perp\!\!\!\perp W \mid (X, V)$ and $X \perp\!\!\!\perp V \mid (Y, W)$, (b) chain graph of multivariate regressions: $Y \perp\!\!\!\perp W \mid V$ and $X \perp\!\!\!\perp V \mid W$, (c) chain graph of concentration-regressions: $Y \perp\!\!\!\perp W \mid V$ and $X \perp\!\!\!\perp V \mid W$; and (d) cyclic graph: $V \perp\!\!\!\perp W \mid (Y, X)$.
Graphs (b) and (c) are examples of two independence equivalent graphs. In fact a distinction in independencies would emerge only with a third response Z in the left-hand box. Then a missing edge between, say X and Z, in (c) would refer to independency conditional on Y, V, W, and in (b) conditional only on V, W

exponential families the concentration graph components refer to canonical parameters and the arrows correspond to (conditional) mean parameters. One intensively studied special case is for categorical data in which joint responses are modeled by loglinear interactions and dependencies via so-called marginal models. A separation criterion was given by S. A. Andersson, D. Madigan and M. Perlman.

Of these three types of chain graph only the second class, the graph of multivariate regressions, contains the general linear models in which every regression is of each component of the response separately on the same set of explanatory variables and the residual covariance matrix contains the remaining parameters. The class also contains seemingly unrelated regressions.

4.4 Cyclic Independence Graphs

Some fully directed graphs may contain directed cycles among a set of different variables and even for the same pair of variables. It has been shown by P. Spirtes and by J. Koster that the d-separation criterion can be applied to these graphs in unchanged form. Graphs of traditional structural equation models may be slightly modified (Koster 1999) to have in our notation dashed lines, arrows, or combinations of both, as well as cyclic dependencies. Then a slightly extended version of the d-separation criterion can be applied to them as well. However, missing edges in these graphs need not point to any independence statement for the variable pair involved, and the meaning of edges in terms of distributional properties may have to be derived from scratch for a given graph.

Figure 2 shows examples of graphs for all four types of joint response models of Sects. 4.3 and 4.4. They have in common two connected joint responses, two marginally independent explanatory variables, and missing edges for two further pairs: (X, V) and (Y, W).

5. Independence Graphs and Generating Processes

As mentioned above, multivariate models which have a univariate recursive generating process can be expressed in a stepwise fashion in terms of univariate conditional distributions. They are sometimes called association models generated over a directed acyclic graph. Because they can be compatible with causal interpretations they are attractive for much substantive research driven by causal hypotheses. See *Causal Inference and Statistical Fallacies*.

Multivariate regression chains are closest to a recursive generating process, since residual association may be regarded as a secondary feature. Also, after replacing each dashed line separately by two arrows starting from a common additional node, called a synthetic common source node, a directed acyclic graph results which implies the correct independence structure for the variables, in the given multivariate regression chain. It also implies covariance graph components which do not have interactions involving more than two variables.

In general, no similar simple generating processes are available for the other two types of chain graph containing chordless undirected cycles. The reasons were that (a) it is not possible to order the nodes in a chordless cycle of a concentration graph to generate it recursively in the given variables and (b) it is also not possible to enlarge the graphs by synthetic common sake nodes to generate undirected chordless cycles. At present the only ways known to generate such cycles are by conditioning on synthetic common responses or by postulating a dynamic process observed cross-sectionally in statistical equilibrium.

Experience suggests that multivariate regression chain structures fit well data from panel studies in several waves. Also, they permit prediction of individual responses without involving response variables considered on an equal footing, but only variables prior to all the joint responses. For many applications in medicine and social science this is an explicit goal of the investigation.

6. A Chain Graph as Partial Summary of Data Analysis

We now return to the variables in Fig. 1. The multivariate regression chain in Fig. 3 is derived by statistical analyses based on data provided by Judith Kappesser of the Department of Psychology at the University of Mainz. The data are for 201 chronic pain patients who have been given a three-week stationary treatment at a chronic pain clinic.

The response of primary interest in Fig. 1 is self-reported success of treatment. It is measured three months after discharge by a score which comprises several aspects of successful treatment. The background or context variables are age, gender, marital status, years of formal schooling, number of previous other illnesses, and duration of pain. There are a number of intermediate variables. Before and directly after treatment, questionnaire scores are available of depression and of self-reported type of pain, ranging from 'no pain' to 'pain as strong as imaginable.' The chronification index is a score incorporating different aspects of time, spreading of pain, use of pain relievers, and the patient's pain treatment history. The main site of pain has two categories: 'back pain' and pain on the 'head, face, or neck.'

The chain graph of Fig. 3 summarizes some important aspects of the results of analysis. It shows, in particular, which of the variables are needed for each response such that adding one more of the potentially explanatory variables does not improve prediction. Site of pain is an important intermediate variable and information on it should therefore be included in studies of chronic pain.

Some of the directions and types of dependency which cannot be read off the graph are as follows. Patients with many years of formal schooling (13 years or more) are more likely to be headache patients, while the others are more likely to be backache patients, while possibly because more of them have jobs involving physical work, and they are more likely to reach higher stages of intensity of pain before treatment. Backache patients reach higher stages of chronification and report higher intensity of pain after treatment. Figure 4 shows that treatment is more successful the lower the intensity of pain, but for scores of intensity of pain of about six or higher this dependence vanishes. One path of development is that patients with shorter periods of formal schooling may get chronic backache, and patients with chronic backache get help too late and respond less well to the type of treatment offered.

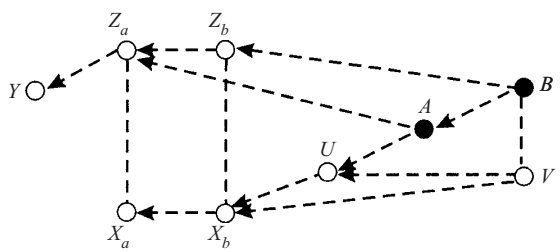

Figure 3
Multivariate regression chain of significant relations, well compatible with the data. Discrete variables are drawn as full circles, continuous ones as open circles

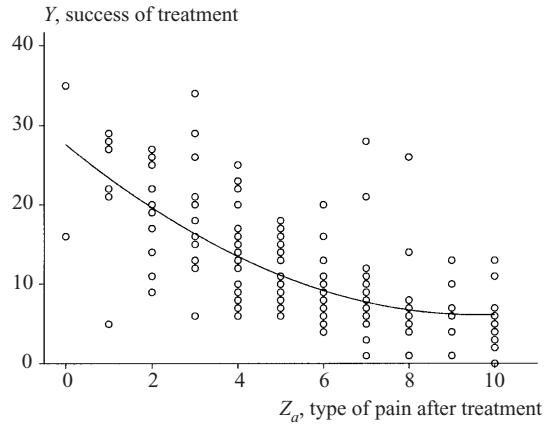

Figure 4
Form of dependence of primary response Y on Z_a

Screening for nonlinear relations and interactive effects did not point to important other relations. The model is said to fit the data well because for each response taken separately no indication was found that adding a further variable would improve prediction. Had there been no nonlinear relations and no categorical variables as responses the overall model fit could have been tested within the context of available software for models of linear structural relations.

7. Suggested Further Reading

We have described some of the statistical aspects of graphical Markov models. Statistical monographs documenting the development are by Wermuth (1978), Whittaker (1990), Edwards (2000), Lauritzen (1996), Cox and Wermuth (1996). For an exposition of important different types of independence structure present in small data sets containing only linear relations, see Cox and Wermuth (1993). Models for single and joint response chain graphs can be viewed as a generalization of Sewall Wright's path analysis different from the generalization to models of linear structural relations (Bollen 1989). The two model classes have some subclasses in common, for instance linear systems represented by directed acyclic graphs and linear multivariate regression chains.

Monographs treating the use of graphical Markov models as a basis for probabilistic calculations in expert systems and for artificial intelligence applications are by Pearl (1988) and by Cowell et al. (1999). For their use in a specifically decision-making context, see Oliver and Smith (1990). For discussions in relation to causal reasoning, see the monographs by Spirtes et al. (1993) and by Pearl (2000).

See also: Bayesian Graphical Models and Networks; Graphical Methods: Presentation

Bibliography

Bollen K A 1989 *Structural Equations with Latent Variables*. Wiley, New York

Cowell R G, Dawid A P, Lauritzen S L, Spiegelhalter D J 1999 *Probabilistic Networks and Expert Systems*. Springer, New York

Cox D R, Wermuth N 1993 Linear dependencies represented by chain graphs (with discussion). *Statistical Science* **8**: 204–18

Cox D R, Wermuth N 1996 *Multivariate Dependencies*. Chapman and Hall, London

Darroch J N, Lauritzen S L, Speed T P 1980 Markov fields and log-linear models for continuous tables. *Annals of Statistics* **8**: 552–39

Edwards D 2000 *Introduction to Graphical Modelling*, 2nd edn. (1st edn. 1995) Springer, New York

Koster J T A 1999 On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scandinavian Journal of Statistics* **26**: 413–31

Lauritzen S L 1996 *Graphical Models*. Oxford University Press, Oxford, UK

Oliver R M, Smith J Q 1990 *Influence Diagrams, Belief Nets and Decision Analysis*. Wiley, Chichester

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA

Pearl J 2000 *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK

Richardson T S 1998 Chain graphs and symmetric associations. In: Jordan M I (ed.) *Learning in Graphical Models*. Kluwer, Dordrecht, The Netherlands, pp. 231–60

Richardson T S 2001 A local Markov property for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*.

Spirtes P, Glymour C, Scheines R 1993 *Causation, Prediction and Search*. Springer, New York

Studený M, Bouckaert R R 1998 On chain graph models for description of conditional independence structures. *Annals of Statistics* **26**: 1434–95

Wermuth N 1976 Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**: 95–108

Wermuth N 1978 *Zusammenhangsanalysen medizinischer Daten*. Springer, Heidelberg, Germany

Wermuth N 1997–99 Graphical Markov models. In: Kotz S, Read C, Banks D (eds.) *Encyclopedia of Statistical Sciences*. Wiley, New York, Vol. 2, pp. 284–300

Whittaker J 1990 *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester, UK

N. Wermuth and D. R. Cox

Green Parties

Green parties have become established political players in most industrialized democracies and are emerging in several industrializing ones. The electoral strides of green parties, especially in Western Europe, have resulted in a situation unimaginable a decade ago: green parties in many national parliaments, green ministers sharing power in several European national governments, greens in the European Parliament forming one of the largest party groups, as well as sustained green representation at subnational levels of government. Yet greens' electoral fortunes are decidedly uneven: electoral success in northern countries contrasts with near invisibility in the South. Moreover, electoral success has brought new strategic dilemmas for the greens as they struggle to maintain their alternative 'green' credentials while joining parliaments and supporting, if not forming, national governments.

1. Emergence

The emergence of green parties needs to be understood in the context of a broader shift from industrial to 'postindustrial' politics. For younger, better-educated citizens in particular, the prosperity and rise in living standards enjoyed in the postwar era formed the backdrop for a change from materialist to 'postmaterialist' values, or from 'old politics' to 'new politics' (see Inglehart 1977, Poguntke 1993). As the basic material needs (food, shelter, etc.) of a far higher share of the population were being satisfied, political attention shifted from materialist concerns to 'quality of life' issues, such as enhanced political participation, gender and racial equality, and, perhaps above all, environmental protection.

Growing environmental concern among the public was heightened by the continued degradation of the environment and increased media attention of environmental issues. Toxic chemical spills in major rivers, air pollution alerts in cities, and oil slicks in pristine coastal waters further increased public anxiety, especially in Europe and the USA. Meanwhile, the accelerated construction of nuclear plants in the wake of the energy crisis, as well as the increased deployment of nuclear weapons, awakened or reinforced fears about the safety of nuclear power, the problems of disposing nuclear waste, and fears of nuclear annihilation.

To many citizens in advanced democracies, none of these concerns was adequately addressed by existing parliamentary structures. A proliferation of extraparliamentary, nonpartisan citizens' movements emerged in the 1960s and 1970s to protest issues of peace, ecology, and women's rights. These new social movements served as the antecedents for a new type of

'green' party, ideologically focused on new politics issues and an alternative paradigm of politics. Vowing to be different from mainstream parties in beliefs and practice, these new parties embraced a 'green' ideology which profoundly challenged conventional politics.

2. Green Party Ideology

The core green values forming this ideology (see *Environmentalism*) encompass a wholesale critique of advanced industrial society and advocate instead a 'sustainable society' (Dobson 1995) based on decentralization and reduced consumption of resources and material goods. For greens, the postindustrial 'good life,' revolving around economic growth and globalization, sophisticated technology, expanding services, and material goods needs to be replaced by a simpler lifestyle offering a cleaner planet, less resource depletion, social equality, a nature-centered spirituality, and local action to solve global problems.

Green parties are characterized not only by their values and beliefs, but by the way they express and pursue these beliefs. Above all, greens emphasize 'grassroots democracy': a decentralized, direct democracy within a radically participatory society. Policy decisions should be made at the lowest possible level and matched with varying local requirements. Diversity is not only tolerated, but celebrated. The organizational structure of most green parties thus tends to be decentralized and nonhierarchical, offering a wide range participatory channels to members, gender equality, and significant autonomy to local branches (Müller-Rommel 1998).

Green party values and practices do not represent a coherent or consistent ideology. Demands for grassroots direct democracy do not always mesh with the need for a sustainable society and the imposition of regulations such a society may require. Structures imposed to deter party hierarchy (e.g., midterm rotation of posts, refusal to appoint party leaders) have proved unworkable. Nor are greens a coherent or monolithic set of thinkers or actors. In particular, several ideological and strategic divisions characterize green parties.

3. Typology of Green Parties

The primary ideological distinction separating green parties is between 'purist' and 'rainbow/leftist' parties (Müller-Rommel 1993). Some refer to this divide as 'green-green' vs. 'red-green' parties. Green-green parties, such as *De Groenen* in The Netherlands, adopt radical environmental policies, reject leftist ideology as too proindustrial, and seek to create a clear distance between their views and views of mainstream parties, left or right. Red-green, or 'rainbow' parties, include the German Greens and the Dutch Green-Left. These parties link ecological issues with leftist concerns of social equality and a critique of capitalism. The predominance of the latter parties has led some writers (Kitschelt 1989) to view green parties merely as an extension of a broader New Left ideology, in which environmental themes are adopted as one theme of many. This view neglects the existence of purist parties or purist elements within green parties. More importantly, it neglects the central role of green ideology and practice which defines green parties and separates them from their leftist counterparts. Green parties are best understood not merely as New Left parties, but as manifestations of an 'ecological cleavage' (Franklin and Rüdig 1995). Despite their differences in emphasis, all green parties have broadly similar core ecological concerns and demands which serve as a pivotal part of their agenda and political identity.

Another broad distinction between green parties is their strategic orientation: the extent to which they embrace parliamentary, mainstream tactics, or favor instead more radical unconventional means to achieve their aims. All green parties pride themselves on alternative strategies, but the intensity of 'radicality' differs across parties. Generally, 'purer' parties eschew traditional forms of political activity such as party alliances, coalitions, or deals with mainstream parties. The small Dutch *De Groenen* party fits this bill. Other parties do not. For example, the French Greens (*Les Verts*) while originally adopting a 'purist' or 'green-green' ideology, were willing to engage in traditional forms of political activity such as participation in presidential elections. The strategic divide applies not only between but also within parties, as illustrated by the tension between German *Realos* (pragmatists) and *Fundis* (fundamentalists). These cleavages are damaging but not debilitating. Despite the obvious difficulties in pulling together different strands and types of green parties, the Green Group in the European Parliament has managed to exercise significant policy influence on issues such as auto emissions and biotechnology controls (see Bomberg 1998).

4. Electoral Fortunes

The varied electoral fortunes of green parties can be explained as a product of the wider political and economic climates, as well as more specific opportunity structures in which these parties operate. Examining the macroeconomic environment in which parties develop is a good starting point for explaining green party success. It is widely recognized that green parties fare better in 'postmaterialist' countries with higher levels of economic development and lower unemployment. Contrast green party emergence and development in the industrialized north with green parties' struggle to emerge and survive elsewhere. But these features are less able to explain differences among developed industrialized states (e.g., why do

green parties fare better in Ireland than in the USA?). More specific opportunity structures better account for such variation.

Electoral laws are often cited (especially by green parties themselves) as a major factor in determining electoral success. Certainly proportional representation (PR) has facilitated the development of new parties in several countries (e.g., Belgium and The Netherlands) while the first past the post electoral system as operates in the USA or British national elections has made it difficult for any small party to succeed unless they are regionally concentrated. But unsuccessful green parties also exist in countries with PR systems (Denmark, Greece) and other factors may be more important in explaining green success. A federal structure offers parties the opportunity first to articulate their demand at subnational level where they can gain experience, publicity, and often funds before launching a nationwide campaign (Müller-Rommel 1998). Green parties have been more successful in federal states (Germany, Belgium Austria, Switzerland, Australia) than in unitary states, and the weak US Greens may be the exception that proves the rule.

To understand change in green party fortunes, other explanations are needed. Evolving party systems help explain the swings in green party support, which often has floundered when major established parties (British Labour Party, US Democrats, Australian Labor Party) or other smaller protest or radical parties (such as the PPR in The Netherlands) either share or poach the ideas of greens and adopt green policies in deed or at least in word. Similarly, green parties face competition from other green actors—environmental interest groups or protest movements—who absorb green demands without themselves seeking government office. To illustrate the point, environmental awareness in the UK and USA is not noticeably lower than in countries with strong green parties, but this support is reflected in swelling membership of environmental interest groups rather than green parties (Rootes 1995).

These various explanations may be supplemented by an examination of political culture (compare France's embrace of radical culture and parties with the UK's tradition of political mediation and compromise) and episodic factors (environmental disasters, rise of new controversies such as genetically modified foods). Finally, a focus on agency and structure is required. Whatever the structural constraints in which they must perform, green parties shape their own fortunes by their (in)ability to develop and articulate a coherent party platform, run a professional electoral campaign, and avoid internal combustion through infighting and personality squabbles.

Taken together these contending explanations provide a plausible if nuanced explanation of green party development and change. Certainly green parties have developed and changed dramatically since their emergence in the 1970s. By the early twenty-first century, green parties were represented in 15 national parliaments (mainly in Western Europe but also in the East), as well as in supranational organizations and hundreds of subnational and local assemblies. Moreover, by the late 1990s greens held ministerial positions in several major European countries. Most successful by far have been the German Greens (*die Grünen*) who first entered federal parliament in 1983, formed a governing national coalition with Social Democrats in 1998 and took on federal ministerial posts, including that of German foreign minister (Joschka Fischer). In late 1999 German Green Michaele Schreyer was the first green to be appointed as a member of the European Commission, the European Union's powerful executive.

5. *Conclusion*

The greens' success has clear policy implications, especially on issues of nuclear power, ecological tax reform, and citizenship rights (see Collier and Golub 1997). But success also has implications for green parties themselves. Greens have always faced a unique 'strategic conundrum' (Bomberg 1998) arising from their unique beliefs and movement roots. Put simply, how can they reconcile their radical alternative politics with participation in mainstream or 'grey' parliamentary and government structures? Throughout the 1990s most green parties shed their radical cloth in an attempt to capture votes, even at the expense of green party unity and purity. Most were rewarded with electoral success well beyond what had been imaginable in the 1980s. The price to pay has been tortured internal debates about strategy, and new questions about green party identity and purpose. Today the key questions facing green parties revolve around not whether to embrace power, but what to do with it. More specifically, green parties face three new challenges in the new millennium: first, how to carve out a policy niche as established parties and governments become wiser to green demands, and as green concerns themselves appear more mainstream. Second, how to take green ideas beyond the confines of rich industrialized states into Eastern Europe and the developing world where green parties remain marginal and environmental problems acute. Third how to ensure that the broader role of green parties—as consciousness raisers, agitators, conscience of parliament and politics—is not sacrificed on the altar of electoral success.

Green parties have come a long way since their emergence and development in the 1970s and 1980s. They have become established players able to shape party competition, government formation, and government policy. But this very 'establishment' carries risks for a party whose core values and identities depend mightily on their ability to challenge the conventional order, to agitate and to annoy. For most green parties,

the greatest fear is not electoral decline so much as the prospect of becoming a party with parliamentary platform, ministerial voice, but nothing new to say.

See also: Environmental Change and State Response; Environmental Planning; Environmental Policy; Environmental Policy: Protection and Regulation; Environmentalism, Politics of; Greening of Technology and Ecotechnology; Participation: Political; Parties: Litigants and Claimants; Party Systems; Political Culture; Political Parties; Political Parties, History of; Third Parties: United States

Bibliography

Bomberg E 1998 *Green Parties and Politics in the European Union*. Routledge, London

Collier U, Golub J 1997 Environmental policy and politics. In: Rhodes M, Heywood P, Wright V (eds.) *Developments in West European Politics*. Macmillan, Basingstoke, UK, pp. 226–41

Dobson A 1995 *Green Political Thought*, 2nd edn. Routledge, London

Franklin M, Rüdig W 1995 On the durability of green politics. Evidence from the 1989 European election study. *Comparative Political Studies* **28**: 409–39

Inglehart R 1977 *The Silent Revolution. Changing Values and Political Styles Among Western Publics*. Princeton University Press, Princeton, NJ

Kitschelt H 1989 *The Logic of Party Formation: Ecological Politics in Belgium and West Germany*. Cornell University Press, Ithaca, NY

Müller-Rommel F 1993 *Grüne Parteien in Westeuropa. Entwicklungsphasen und Erfolgsbedingungen*. Westdeutscher Verlag, Opladen, Germany

Müller-Rommel F 1998 Explaining the success of green parties: A cross-national analysis. *Environmental Politics* **7**: 145–54

Poguntke T 1993 *Alternative Politics. The German Green Party*. Edinburgh University Press, Edinburgh, UK

Rootes C 1995 Britain: Greens in a cold climate. In: Richardson D, Rootes C (eds.) *The Green Challenge. The Development of Green Parties in Europe*. Routledge, London, pp. 66–90

E. Bomberg

Green Revolution

The term 'green revolution' was applied to land reform in Eastern Europe in the 1930s and recently to the attempts to persuade governments and business to adopt policies and practices that are not inimical to the environment. But the term is most commonly used in the social sciences to describe the changes that took place in farming in Asia after the introduction of semi-dwarf wheat and rice varieties in the mid-1960s. If planted on irrigated land with the use of chemical fertilizers and pesticides these crop varieties gave higher yields and greater profits than traditional ones. These innovations led to a substantial increase in the yield of wheat and rice, which was sustained until the 1990s when the rate of increase of yield and output declined; there may be difficulty in keeping cereal output ahead of population growth in the twenty-first century.

1. Early History

The period immediately following World War II in South Asia was characterized by rapid population growth, food shortages, and poverty. A great variety of wheat and rice varieties were in use, each adapted to their local micro-environment; little chemical fertilizer was used and no pesticides, and yields were low. In 1966 a semi-dwarf variety of rice known as IR8, bred at the International Rice Research Institute in the Philippines, was distributed amongst farmers in parts of India, whilst a semi-dwarf wheat originally bred in Mexico was also introduced. Where these High Yielding Varieties (HYVs) were adopted, yields and output increased dramatically. This led W S Gaud to state in a lecture in the USA in 1968: 'this and other developments in the field of agriculture contain the makings of a new revolution. It is not a violent Red Revolution like that of the Soviets ... I call it the Green Revolution'(quoted in Dalrymple 1979). The importance of the Ford and Rockefeller foundations and of the US government in sponsoring the breeding programs in Mexico and the Philippines meant that much of the criticism of the new technology and its alleged consequences was ideological. The term 'revolution' was perhaps unfortunate. The breeding of higher-yielding or disease-resistant varieties had been going on in Europe, North America, and parts of Asia since the rediscovery of Mendel's work at the beginning of the twentieth century, whilst the use of chemical fertilizers and applications of pesticides had been practiced since the 1920s in Western Europe, although adopted at far greater rate after 1945. More relevant were the developments in Japan.

Much Japanese rice had long been irrigated, but the area irrigated was increased in the 1890s when the most successful existing rice varieties were selected and their seed distributed to farmers. Chemical fertilizers were also applied. By the 1930s, rice yields in Japan and its colonies Korea and Formosa were substantially above those elsewhere in Asia. Even in South Asia there had been advances in technology before 1966. Indian grain output increased by 62 percent before 1950 and 1964, and in Pakistan by 59 percent (Farmer 1981). Although rice yields in China in the 1950s were low compared with the rest of East Asia, semi-dwarf rices were bred there independently of the research in the Philippines and distributed to Chinese farmers before IR8 was released.

2. The Nature of High Yielding Varieties (HYVs)

The principal characteristic of the HYVs were their high response to chemical fertilizers. They were also short with a stiff straw, and so, unlike the tall traditional varieties, did not lodge with higher yields. They were insensitive to photoperiodicity and matured in about 110 days rather than 180 days; it was thus possible to grow two or even three crops in a year. The yield potential of these varieties was greater in the temperate regions of Asia and in the dry season in the monsoon region than in the humid tropics, because of the longer hours of sunshine and hence the greater potential photosynthesis available to the plant. The highest yields are achieved in areas of well-managed irrigation and with adequate amounts of chemical fertilizers. At first the new varieties were highly susceptible to disease, and pesticides were necessary to obtain the high yield potential, but later varieties have been bred with a greater resistance to pests. Breeding of new varieties of rice and wheat has continued since the 1960s, increasingly in research stations in the countries where they are to be grown. High yielding varieties of maize have also proved successful and to a lesser extent new varieties of sorghum. Since the 1970s Chinese farmers have used hybrid rice seed. The new varieties introduced since the first HYVs are now often referred to as MVs (modern varieties).

3. The Diffusion of the Modern Varieties

The inputs of the Green Revolution are divisible: seed, fertilizer, and pesticides can be bought in small amounts and so in theory could be obtained by farmers with only very small holdings. In practice however, the first adopters were generally occupiers of larger farms with better credit facilities and a greater awareness of the new technology. By the 1980s however, there was near universal adoption in irrigated areas with no differentiation of inputs by farm size or tenure. The principal incentive to adopt was the greater profit per hectare, in spite of the increased unit costs. At first the use of fertilizers and seeds were subsidized, and the improvement and extension of irrigation was undertaken by government. In the 1980s however, subsidies began to be withdrawn, and the further extension of irrigation had become very expensive. The introduction of modern varieties began at different times in different countries and different regions. Rice was introduced into India, for example, in 1965 but not in South Korea until 1972. The rate of adoption also differed; by 1976 four-fifths of all the rice in China was sown with modern varieties, but only 7 percent in Burma, 11 percent in Thailand, 14 percent in Bangladesh. National differences persisted in the 1990s (see Table 1).

Within countries there have also been major regional differences in the rate of adoption; thus in India in 1990, 95 percent of the rice in Punjab and Tamilnad was sown with modern varieties, but only one-third in Bihar and Kerala (Singh 1997). By the 1990s some of the major international differences in the rate of adoption had been reduced, but a greater proportion of Asia and Latin America's wheat was sown with modern varieties than Africa or the Near East, whilst more of China's was in new varieties than the rest of Asia (see Table 1). But of course there is a difference in the proportion of each continent sown in wheat, rice, maize, and other crops. Little wheat or rice is grown in tropical Africa, so the proportion of these crops in modern varieties is of little significance. This distribution explains why, in the mid-1980s, modern varieties were planted on less than one-third of the total cereal area in the developing countries: only 1 percent in Africa, 22 percent in Latin America and 36 percent in Asia (Tangley 1987).

4. The Consequences of the Green Revolution

In the 1950s and 1960s it was feared that future population growth would outrun cereal output in Asia and increase malnutrition and hunger. This has not occurred. The rate of increase in output of both rice and wheat in Asia exceeded population growth rates between 1966 and 1995. Population increased 76 percent but rice output doubled; and the real price of rice has fallen steadily since the 1950s, halving between 1970 and 1990. Cereal consumption per capita has increased by 25 percent, imports have been greatly reduced, and grain reserves built-up. The increase has

Table 1
The diffusion of modern varieties—percentage of crop area sown in modern varieties

	Rice			Wheat			Maize
	1970	1976–7	1991	1970	1976–7	1990	1990
Sub Saharan Africa	4	3	n.a.	5	23	52	43
West Asian and North Africa	0	4	n.a.	5	17	42	53
Asia (excluding China and West Asia)	12	30	67	42	72	89	45
China	77	80	100	n.a.	25	100	90
Latin America	4	13	58	11	41	82	46

Source: D Byerlee 1996
D G Dalrymple 1979

Table 2
Cereal output in Asia

	1948–52	1961–5	1971–3	1979–81	1987–9	1995–7
			Rice: million tonnes			
Japan	11.9	16.4	15.1	13.3	12.9	12.9
South Korea	2.9	4.8	5.6	6.8	8.0	6.8
China	58.5	86.0	108.1	145.7	175.8	194.3
Indonesia	9.4	12.4	19.0	29.6	41.7	50.5
Philippines	2.8	4.0	5.2	7.9	9.0	11.3
India	33.4	52.7	63.4	74.6	99.7	121.4
Bangladesh	12.4[1]	15.0[1]	16.1[1]	20.1	24.2	26.6
Thailand	6.9	11.3	13.3	17.0	20.8	21.7
			Wheat: million tonnes			
China	18.1	22.2	34.1	59.1	87.4	111.8
India	6.0	11.2	25.0	34.6	48.2	65.9
Pakistan	3.7	4.2	7.0	10.8	13.0	16.8

[1] includes Pakistan
Food and Agricultural Organization *Production Yearbooks*, various years

Table 3
Crop yields in Asia, 1948–1997

	1948–52	1961–65	1971–75	1979–1981	1987–89	1995–97
			Rice: tonnes per hectare			
Japan	4.0	5.0	5.7	5.6	6.1	6.4
South Korea	2.8	4.1	4.7	5.5	6.4	6.5
China	2.1	2.8	3.1	4.2	5.4	6.2
Indonesia	1.6	1.7	2.3	3.3	4.1	2.8[1]
Philippines	1.2	1.3	1.5	2.2	2.7	2.9
India	1.1	1.5	1.7	1.9	2.4	2.7
Bangladesh	n.a	1.7	1.7	1.9	2.4	2.6
Thailand	1.3	1.8	1.9	1.9	2.0	2.2
			Wheat: tonnes per hectare			
China	0.7	0.9	1.4	2.0	3.0	3.7
India	0.7	0.8	1.3	1.5	2.1	2.7
Pakistan	0.9	0.8	1.2	1.6	1.7	2.0

1.44 tonnes 1991–94
Food and Agriculture Organization, *Production Yearbooks*, various years

varied between countries (Table 2), and has been greater in wheat than rice output, but there have been dramatic increases in cereal harvests throughout the region. In the 1950s and 1960s increases in cereal output were largely due to an extension of the cultivated area; since the 1960s, overwhelmingly due to the increases in crop yields following the adoption of the new high-yielding varieties (Table 3).

The shorter growing season for HYVs has also allowed more double cropping in some parts of Asia. Yields have stagnated in the 1990s, however, particularly in Japan and Korea, which have the highest rice yields. The yield potential of the original HYVs has not been increased since the 1960s; the many new varieties bred since then have had greater disease resistance or better quality grain, but not greater yield potential. Additionally, in the high-producing irrigated rice areas, waterlogging, salinization, and disease have led in some places to a decline in yields. The fall in rice prices and the continued increase in import prices has squeezed cereal producers, and many have left farming or turned to more rewarding crops. In the last few years there have been renewed fears for the future of cereal output in Asia.

5. *Criticisms of the Green Revolution*

Initially the Green Revolution was greeted with extravagant praise and high expectations. These were excessive, for it was unlikely that many farmers could reach the yields obtained on experimental farms, or that all farmers could profitably use the new techniques. A barrage of criticism followed to the present, although with changing targets. Controversy remains over the social and economic effects of the Green Revolution, nor is this surprising. There are three technical reasons why interpretations may differ. First,

national and state statistics in India and other parts of Asia are not reliable and so different interpretations of changes are possible, particularly with reference to land tenure, farm size and labor. Second, it would be surprising if there were identical responses to new technology over such a vast area and population with greatly varying physical and socioeconomic environments. Third, research produces different results at different stages of the adoption cycle of agricultural innovations. It is unlikely that the consequences of technological change in the 1960s would be the same in the 1990s.

But there are other criticisms that are a result of the unequal distribution of farm sizes and land ownership. Early critics believed that the adoption of HYVs benefited the larger farmers most, often led small farmers to sell their land, increased landlessness, and forced down real wages. It was argued that while those with larger farms—and in most of Asia more than 20 acres is a large farm and most farmers have less than five acres—were able to rapidly adopt the new technology, and make considerable profits; smaller farmers were unable to do so and often sold their holdings to larger farmers, whilst many tenants were evicted to allow landowners to farm larger units. The number of landless households increased and so forced down real wages. In some areas the larger farmers adopted machinery to replace labor, and real incomes of the landless fell. Consequently income inequality increased.

These criticisms were made most fiercely in the early 1970s. At the time it was pointed out that many of these trends predated the 'green revolution,' or had alternative explanations. Machinery was being adopted before 1965, and rural population growth alone was sufficient to increase landlessness. Since then there have been more sample studies over longer periods (David and Otsuka 1994, Hazel and Ramasamy 1991). They have shown that small farmers have been able to adopt the new techniques, seed, and other inputs; that real wages have not fallen; and that the adoption of machinery has often been the result of farm laborers seeking higher paid jobs elsewhere. Other critics pointed out that the HYVs were only profitable in irrigated areas, and that improved varieties of only wheat and rice had been bred, yet in many parts of India farmers grew other cereal crops so that the benefits of the 'green revolution' occurred only in limited regions. This was true at that time, but subsequently modern varieties were bred that gave higher yields than traditional varieties in rainfed areas and with low amounts of fertilizer. Furthermore, high-yielding varieties of other crops have been bred: high-yielding maizes were bred in Kenya and Rhodesia before 1965, and now occupy half the African maize acreage, whilst in India high-yielding varieties of sorghum and millet have been bred. The early HYV rices had an inferior taste, and although subsequent varieties were of higher quality, richer consumers still prefer the taste of traditional varieties. Other critics were alarmed that the inputs for the new farms were imported, whilst some believed that that monoculture was replacing the diversity of cropping in traditional Asian farming, particularly the replacing of protein rich pulses with cereals.

By the 1990s new fears replaced these traditional concerns. Yields were—and still are—stagnating and even declining as a result of waterlogging and poor drainage in irrigated areas, whilst chemical pollution was becoming a problem. Falling cereal prices and rising costs are now threatening the future of cereal output, whilst Asia's population may increase by 53 percent by 2030. It may be that a new technological advance will be necessary to maintain Asia's freedom from famine that was the result of the 'green revolution' of the 1960s.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Environment and Development; Environmental Planning; Food Security; Greening of Technology and Ecotechnology; Technological Innovation

Bibliography

Burmeister L 1987 The South Korean green revolution: Induced or directed innovation. *Economic Development and Cultural Change* **35**: 767–90

Byerlee D 1996 Modern varieties, productivity and sustainability: Recent experience and emerging challenges. *World Development* **24**: 697–718

David C C, Otsuka K 1993 *Modern Rice Technology and Income Distribution in Asia*. Rienner, Boulder, CO

Dalrymple D G 1979 The adoption of high-yielding grain varieties in developing nations. *Agricultural History* **53**: 704–26

Evenson R E, Herdt R W, Hossain M 1996 *Recent Developments in the Asian Rice Economy: Challenges for Rice Research*. CAB International, Wallingford, UK

Farmer B H 1977 *Green Revolution? Technology and Change in Rice-growing Areas of TamilNad*. Westview, Boulder, CO

Farmer B H 1981 The green revolution in South Asia. *Geography* **66**: 202–7

Griffin K 1974 *The Political Economy of Agricultural Change: An Essay on the Green Revolution*. Macmillan, London

Hazell P B R, Ramasamy C 1991 *The Green Revolution Reconsidered: The Impact of High-yielding Rice Varieties in South India*. Johns Hopkins University Press, London

Karim M B 1986 *The Green Revolution: An International Bibliography*. Greenwood Press, New York

Lipton M, Longhurst R 1989 *New Seeds and Poor People*. Unwin Hyman, London

Pingali P L, Hossain M, Gerpacio R V 1997 *Asian Rice Bowls. The Returning Crisis?* CAB International, Wallingford, UK

Shiva V 1989 *The Violence of the Green Revolution*. Natray Publishers, Dehra Dun, India

Singh J 1997 *Agricultural Development in South Asia. A Comparative Study in the Green Revolution Experiences*. National Book Organisation, New Delhi, India

Tangley L 1987 Beyond the green revolution. *Bioscience* **37**: 176–80

D. B. Grigg

Greening of Technology and Ecotechnology

'Greening of technology' and 'ecotechnology' have similar meanings. Both terms define a perspective regarding the impact of technology as a mediator and/or cause of the environmental problems in modern societies. Both imply a move toward a different kind of technology that is environmentally friendly.

1. *'Green' and 'Ecological'*

'Greening' and the prefix 'eco' refer to technology's association with environmental processes and to its role, and thus humankind's, as a cause of environmental change, from pollution to global climate change. Greening of technology and ecotechnology call for more environmentally benign technology, including systems that operate with a minimum impact on species and habitat. So both terms define a vision for change and refer to a set of criteria to be taken into account when evaluating the qualities of given technologies in terms of their impact on the natural environment, ranging from wastes and emissions to the consumption of non-renewable resources. They call for intentional technological designs to achieve the goals in question. Also in other fields these terms are used to pinpoint intended changes and policy needs, as exemplified by 'greening of industry,' 'ecotourism,' 'greening of consumption,' 'eco-efficiency,' and 'green accounting,' each pointing to specific areas of socioeconomic activity.

Although the term 'green' is used as a synonym for nature and natural in a number of situations, the close link to environmental problems is rather new and follows from the systematic and public recognition of pollution impacts in the 1960s and subsequently. In this sense, the term differs from that used in, for example, the 'green revolution' and its emphasis on increasing crop production from improved varieties and petro-chemical inputs (see *Green Revolution*). Indeed, some of the impacts following from the green revolution and identified by Rachel Carson (1962) in *Silent Spring* are precisely those illustrative of the double-edged character of technology and the impetus for green technology. The origins of the new uses of 'green' and 'eco-' in regard to technology have not been adequately addressed. The linkages between technology and environmental damage, however, were highlighted in the Club of Rome report *Limits to Growth*: 'For the first time, it has become vital to inquire into the cost of unrestricted material growth and to consider alternatives to its continuation' (1972, p. 191) and brought to the international policy discussions in 1972 at the United Nations' first global environment conference in Stockholm.

2. *'Technology'*

Technology is at the same time the engine of economic growth and human well-being and the immediate cause of environmental degradation. Technologies as machine embedded or enacted processes are the basis for transforming materials using energy. Industrial mass production and transport technology extract and concentrate materials, increasing energy demands and amounts of waste, both from the industrial process and from the 'throwaway culture' it supports. This waste includes materials that cannot be reused or recycled but remain concentrated and combined in ways that cause health hazards and reduce the stock of raw materials accessible for future generations. Such resource consumption is accompanied by impacts on the structure and function of the biosphere, be it increasing atmospheric carbon dioxide or the extinction of species.

The definition of 'technology' is important for the scope of its impact on nature. Technology does not only cover machines and other artifacts, but also includes processes and procedures, like the methods used in rural agrarian production and in the structure of settlements (e.g., contemporary continued urbanization leading to mega-city societies). The question extends beyond the domains covered by the notion of 'technology,' however, to the sociology of its use and the organization of economic processes. These systemic uses of technology are registered in such terms as technological system, or socio-technical ensemble (see *History of Technology*). Through these broader uses of the term, technology addresses such human activities as the movement, transformation, and consumption of materials and energy. Many environmental impacts from technology are not only linked to the production of technology *per se* or its basic function, but also to how production and consumption are connected (see *Industrial Metabolism*).

3. *Is 'Green' an Attribute or a Process?*

The labels 'ecotechnology' and 'green technology' are commonly applied to environmentally friendly production technologies and/or products, such as wind turbines and solar cells for renewable energy, natural textiles, and organic farming reducing, for example, the use of inputs from petrochemical industries. This use has been contested on the grounds that, in absolute terms, no technology is truly green and, in some cases, green and technological risks (or perception of risks) are confused. This is the case for genetically modified crops, for example. They are usually labeled as 'non-

green,' in public debate. But from a different perspective, it can be argued that such hybrid crops could produce more food on less land. It is, however, not questioned that specific technologies can be less polluting and greener than others, in relative terms.

What is identified as environmentally friendly, green, or ecological is dependent on how the environmental problems and risks are perceived and prioritized. Every technology has an impact on the environment, and the choice between technologies must be assessed in terms of these impacts and the risks involved when recognizing the limits to existing knowledge. This leads the decision in the direction of a political process that has to deal with uncertainties as well as conflicting interests in the outcome.

Consequently in more sophisticated uses of the notions of 'green' or 'ecological,' they are used to characterize the direction of change based on selected measures referring to environmental performance of technologies. The notions are thereby used to establish distinctions between the outcome of different change processes and their focus.

4. Promotions and Problems

While traditional environmental policies developed and installed in the 1970s focused on the discharge of pollutants from production, the growing awareness for a need to engage industry in a constructive way resulted in new policy strategies in the later part of the 1980s. Here the strategy was to engage companies in environmental protection measures by introducing the concept of cleaner technologies and industrial ecology (see *Industrial Ecology*). The cleaner technology (could also read 'greener') approach value materials and energy reductions based on improved design and production as the highest priority, followed by substitution of nonrenewable materials, recycling, and ending with filtering and waste handling as the absolute lowest priority.

These new policies were supported by the introduction of environmental management as a new subdiscipline and the implementation of new management standards like the international ISO 14000 series and EMAS in the EU together with economic fees on pollution and waste. By focusing their interest on waste and the inefficient use of materials and energy, many companies improved their environmental performance. Again the ambiguous character of technology is illustrated: cleaner is not a simple quality of technology, but a product of a specific direction in industrial transformation.

Following these insights, environmental regulation has also changed from a focus on industrial point sources as the object of regulation (e.g., smokestack emissions) to broader aspects of societal production and consumption, such as the impacts from product use and disposal (e.g., recycling). Product lifecycle and eco-design approaches have begun to be implemented by companies and supported by government environmental policies, especially in the developed world. The lifecycle approach addresses environmental impacts from the perspective of the complete cycle of a product's life: the production of raw materials, the transport of materials and products, the production process, the use of the product, and the final recycling or disposal of the product as waste. Consequently it introduces a rather extended obligation for companies in managing not only their own pollution and resource consumption but in being responsible for the complete chain of supplies and for the minimization of environmental impacts and resource consumption through consumption and disposal. Procedures for eco-design have as their main focus the translation of environmental impacts within the design phase and the reorganization of innovative processes to include both environmental concerns but also user perspectives and new types of expertise (Mitch and Jørgensen 1989).

The problems from which green and ecotechnology have evolved are now also registered in the term 'sustainable development,' signifying the dual needs of providing improved human well being with less environmental impacts (see *Sustainability Transition: Human–Environment Relationship*). This idea was officially recognized in the World Commission on Environment and Development's report *Our Common Future* (The Brundtland Report), which stated that 'When the century began neither human numbers nor technology had the power to radically alter planetary systems. As the century closes, not only do vastly increased human numbers and their activities have that power, but major, unintended changes are occurring in the atmosphere, in soils, in waters, among plants and animals, and in the relationships among all of these' (1987, p. 343).

The introduction of cleaner technologies has in principle shown results in reducing waste and the discharge of pollutants in quite substantial ways. But voluntary actions in industry do not guarantee the introduction of these green technologies (Kemp 1997). And these results do not satisfy those needed for a future whose material and energy demands appear to be enormous. Rather radical reductions in materials and energy consumption and also the reduction of the 'ecological footprint' of individual consumers are required. Phrased as the 'factor 4' and 'factor 10' strategies for reductions in materials and energy use, a new agenda for technological change has been outlined, including also the need for increased global wealth. This agenda will influence product designs, work organization, transport, and culture in radical ways in order to reach the result: a radical reduction in the amount of materials and energy circulated and used per produced unit of good or service (Weizsäcker et al. 1997). Such changes cannot be reached in business and industry alone, but will have to include

behavioral changes of consumers as well (see *Consumption, Environmentally Significant*). This strategy is not seen as different from or in competition with strategies focusing on industrial performance and eco-design, but as supplementary to them (Fisher and Schot 1993).

5. Involved Actors

Such greening and ecological concerns engage a large part of industry. This engagement is partly a response fueled by industrial interests in promoting and marketing their green improvements, and partly a response to pressures from governments and consumers to reduce environmental impacts. During the 1980s a number of new national and international organizations and networks were established in response to these challenges. These included cooperating networks as the World Business Council for Sustainable Development, the European Roundtable for Cleaner Production, the American Coalition for Environmentally Responsible Economics, and many others established on both regional and global levels. Consumers and environmental grassroots movements also took part in the construction of new agendas for green technologies and greener products, some even becoming involved in innovations, for example, Greenpeace's development of a low-energy refrigerator and cooperation with an industrial company about its manufacture, and other organizations' contribution to innovation in the wind turbine. The 'green' impact is also seen in the commercial use of environmental qualities in marketing products, ironically termed 'green speak' and 'green wash' when not based on substantiated environmental performance improvements.

6. Academic and Disciplinary Context

The study of 'greening' and 'ecotechnology' is not the subject of one specialist discipline but involves a large number of academic disciplines, ranging from engineering and design to social perception and innovation management. Specific subfields in the technical sciences advance green or ecological technologies, including energy-efficient cars, urban ecology, eco-textiles, organic farming, and energy efficient household equipment. Special attention has been given to methodologies assessing the impact of a product during its complete lifecycle (in short: LCA for life cycle assessment) and in searching for tools to be used as a basis for environmental prioritization. Design procedures and ways of translating and integrating environmental concerns into design practices have escalated as research interest lately showing both in the design journals and in new design manuals (*Ecological Engineering* is an example). New interdisciplinary research centers have also been established.

Technology studies, drawing on the experiences from the STS (Science Technology Society) field, and the studies of innovation in both management and evolutionary economics also sometimes include studies of the environmental dimensions of technology. The technology studies field, in particular, has developed a special interest in studying green technologies. But in the mainstream academic social science disciplines, as in the mainstream technology design fields, environmental concerns are often not explicitly mentioned or even left aside. Researchers have therefore built their own networks and journals and organize special sessions within the meetings of other formal organizations. One example, however, of a cross- or interdisciplinary organization, is the 'Greening of Industry' Network, established in 1990.

Among the mainstream social sciences, political science and economics have been active in regard to environmental regulation and politics lately. The interest in 'green industrial policies,' 'greening of business activities,' and 'green tax reforms' has been growing. So has the support for 'environmental economics' focusing on how to make external impacts visible in the economic system (i.e., the costs and problems caused by the economic system, but not accounted for in the price and market mechanism). Although 'technology' is often referred to in this context, in most studies both the environmental problems and the technologies stay in the 'black box,' and are not analyzed in detail.

7. Future Directions of Research

The rather dispersed character of research in 'greening of technology' illustrates two fundamental problems: the need for an interdisciplinary approach to tackle the problems, and the need for an integration of environmental concerns into a number of technical and social science disciplines. The need for interdisciplinary approaches is obvious in design and in management, where the traditional focus has been on economic priorities and organization rather than environment matters. These priorities have less to do with the character of the field of research than with the tradition of disciplines. The need for integration among disciplines is obvious, given the range of linked issues, from specific technologies, to business management, to government policy.

The complexity of the problems and interested communities is confounded by the juxtaposition of policymakers' and managers' desire for simple solutions. Such solutions are problematic, of course, especially when they are not measured by assessments of knowledge limitations and uncertainty (see *Precautionary Principle*). More generally, in studies of innovation and technological design there is a need to

develop methodologies to cope with integrating environmental concerns into design and management of processes and to deal with the conditions for use and consumption.

The development of greener technologies is not simply a question of improvements and integrations in the technical sciences and in design practices, though we should not underestimate the importance of these areas. A change in the dominant priorities for technological development is dependent on priorities set in the business sector, and changes therefore must address both management practices and societal governance. Likewise, since business is not unresponsive to public pressures and consumer actions these concerns must be addressed as well. In short, the greening of technology is a multi-sectoral and multidisciplinary subject which is likely to increase in importance as future generations cope with increasing demands to meet human needs, needs that are being redefined to include environmental quality (IHDP 2000).

See also: Agricultural Sciences and Technology; Consumption, Environmentally Significant; Ecotourism; Environmental and Resource Management; Green Revolution; History of Technology; Industrial Ecology; Industrial Metabolism; Land Reclamation; Land Use and Cover Change; Precautionary Principle; Sustainable Development

Bibliography

Carson R 1962 *Silent Spring*. Houghton Mifflin, Boston

Club of Rome 1972 *Limits to Growth*. Universe Books, New York

Fischer K, Schot J 1993 *Environmental Strategies for Industry*. Island Press, Washington, DC

IHDP 2000 *Industrial Transformation: Science Plan*. International Human Dimensions Programme on Global Environmental Change, Report No. 12, Amsterdam

Kemp R 1997 *Environmental Policy and Technical Change*. Edward Elgar, Cheltenham, UK

Mitch W J, Jørgensen S E (eds.) 1989 *Ecological Engineering: An Introduction to Ecotechnology*. Wiley, Chichester, UK

Roome N (ed.) 1998 *Sustainability Strategies for Industry: The Future of Corporate Practice*. Island Press, Washington, DC

Weizsäcker E von, Lovins A, Lovins L H 1997 *Factor 4. Doubling Wealth—Halving Resource Use*. Earthscan, London

World Commission on Environment and Development 1987 *Our Common Future* [The Brundtland Report]. Open University Press, Buckingham, UK

U. Jørgensen

Grounded Theory: Methodology and Theory Construction

Grounded theory is an inductive methodology that provides systematic guidelines for gathering, synthesizing, analyzing, and conceptualizing qualitative data for the purpose of theory construction. In addition, one of its creators, Barney G. Glaser (1978, 1992), has long proclaimed that grounded theory methods may be used with quantitative data. The term, 'grounded theory,' derives from its central premise: theory must be developed from systematic analysis of empirical data. This methodology consists of flexible strategies to guide qualitative data collection, and, particularly, data analysis. The strength of the grounded theory method lies in articulating: (a) logical steps for handling data collection and analysis, (b) a means of correcting errors and omissions and of refining analytic ideas, (c) tools for studying basic social and social psychological processes in natural settings, and (d) strategies for creating middle-range theories.

Prior to publication of Barney G. Glaser and Anselm L. Strauss's *The Discovery of Grounded Theory* (1967), training in qualitative research had largely occurred through an oral tradition of mentoring and immersion in a field setting. Thus, the grounded theory method represented a significant advance because it codified steps in analyzing qualitative data. The distinctive features of the grounded theory method include: (a) simultaneous data collection and analysis, (b) reliance on comparative methods, (c) early development of categories, (d) intermediate analytic writing between coding data and writing the first draft, (e) sampling for developing ideas, (f) delay of the literature review, and (g) a thrust toward developing theory.

1. The Intellectual Heritage of Grounded Theory

Grounded theory derives from the intellectual traditions of each of its founders. Strauss brought Chicago School pragmatism, symbolic interactionism, and field research to grounded theory (see also *Symbolic Interaction: Methodology*; *Field Observational Research in Anthropology and Sociology*). Hence, grounded theorists study processes in natural settings and invoke pragmatic criteria of usefulness to evaluate the completed study. Grounded theory methods themselves echo pragmatist and symbolic interactionist assumptions of social life as emergent and open-ended, and answer Herbert Blumer's (1969) call to study social action in natural settings. Glaser's training in survey research at Columbia University lent grounded theory its systematic approach, positivist proclivities, and procedural language. Glaser codified steps of analysis of qualitative data in an analogous way as quantitative research had been codified. His efforts culminated in an explicit statement of how to handle data analysis (Glaser 1978, Glaser and Strauss 1967). Glaser's approach assumed a knowable world waiting to be discovered, unbiased observers who are uninfluenced by preconceived logico-deductive theories of this world or by prior research about it, and a view of grounded theory

categories as arising from the data. As a result of its divergent origins, grounded theory contains elements of both positivism and constructivism.

Glaser and Strauss developed grounded theory methods at a time when quantification had gained hegemony throughout the social sciences. Theory and research had become separate pursuits. The quest for quantified research findings resulted in the waning of qualitative studies. Except for an occasional classic study, most quantitative methodologists ignored qualitative research and relegated it to disciplinary sidelines, or treated it only as a precursor to rigorous quantitative research. Qualitative research was deemed impressionistic and anecdotal, unfitting the scientific quest for quantified facts. In contrast, Glaser and Strauss (1967) argued that qualitative research could stand as science in its own right, demonstrate rigor, and generate theory.

Glaser and Strauss's arguments found receptive audiences. Their methods appealed to social scientists who wished to conduct qualitative research but lacked tools for doing it. In the late 1960s and throughout the 1970s, grounded theory methodological rationales contributed significantly to re-establishing the legitimacy of qualitative research. This method provided researchers with ready justifications for conducting qualitative studies and strong rationales that their research inquiry was systematic. Not surprisingly, diverse qualitative researchers still claim to use grounded theory to establish their credibility and the legitimacy of their research enterprise. However, their claims of using grounded theory may rest on having conducted some form of qualitative research or on only following the initial steps of the grounded theory method.

2. *Developments and Changes in the Method*

Although *The Discovery of Grounded Theory* called for conducting qualitative research and developing theory, it was less clear on details about using the method. Glaser's *Theoretical Sensitivity* (1978) set forth the most explicit statement of how to conduct grounded theory research, yet its abstract and dense writing made it resonate most with those who have had prior training in grounded theory. Since then, Glaser's position has remained consistent with his early formulations. For Glaser (1992), grounded theory builds on discovered data, avoids preconceiving interpretation of them through extant theories or categories, relies on comparative methods, and aims toward theory development. Thus, the grounded theorist compares data with data, data with concept, concept with concept, and theoretical category with theoretical category. For example, the grounded theorist compares one interview excerpt with another, an interview excerpt with a concept that may explain it, a concept with another concept and so forth.

Strauss with his co-author, Juliet Corbin (1990, 1998), took grounded theory in somewhat different directions than the early formulations. They introduce several new techniques, stress description, and include verification as part of grounded theory. They emphasize adherence to technical procedures which Glaser views as forcing data into preconceived categories rather than letting categories emerge through comparing data with data. Both Strauss and Corbin and Glaser's approaches remain objectivist because they each assume the reality of an external world, discovery of data within that world, a neutral observer of it, conceptual categories as emanating from the data, and representation of data and subjects as nonproblematic.

A constructivist grounded theory (Charmaz 1995, 2000) adopts methods of grounded theory without adhering to earlier objectivist, positivist assumptions (see *Constructivism/Constructionism: Methodology*). It places priority on the studied phenomenon rather than methods of studying it; hence, it adopts grounded theory strategies as useful tools, not as rigid prescriptions. Because these grounded theorists build reflexivity into the research process, they scrutinize their research experience, ways of knowing, and products of knowing (see *Reflexivity: Method and Evidence*). A constructivist grounded theory attends to data collection closely. It does not assume that data simply exist in an external world, awaiting discovery. Nor does it assume that an observer can enter the research scene as a *tabula rasa*. Rather, what observers see and hear depends upon their prior interpretative frames, biographies, and interests as well as the research context, their relationships with research participants, and modes of generating and recording data. The questions that observers ask of the empirical world shape data that they select from it. Similarly, constructivists view their conceptual categories as constructed through their interpretations *of* the data rather than emanating *from* them. Thus, they assume that their resulting theoretical analyses provide interpretive renderings of a reality rather than an objective reporting of the reality.

3. *Data in Grounded Theory Research*

Although grounded theorists have given priority to data analysis over data gathering, the quality of their gathered data affects the final analysis. Full, rich data provide a more thorough view of the studied topic. The kind of data obtained for a study should suit the research topic and research participants. Grounded theory methods have become closely associated with interview studies, however, they are suitable to use with varied forms of data such as ethnographic field notes, written personal accounts, and documents (Clarke 1998).

A hallmark of grounded theory studies is that data collection and analysis proceeds simultaneously. Each

provides a check on the other. Grounded theorists code and categorize data during the initial stages of data collection. This strategy helps them to shape further data collection to gather material that fits the topic and field setting. Hence, they are less likely to pursue their preconceptions about their topic and more likely to follow leads obtained through their early analysis to find new data.

Grounded theory methods are emergent; both the type and nature of data and analysis emerges throughout the research process. Before grounded theory methods were developed, researchers separated data analysis and collection phases of research. Often they discovered later that they had gathered extensive, but thin, data that contained gaps. Grounded theory strategies include returning to the field to gather more complete data that fills gaps and answers questions. Because grounded theory methods are aimed to further the theoretical power of the analysis, these methods tailor data gathering to the researcher's emerging analysis. Thus, adopting these methods makes data collection more efficient and, therefore, streamlines the research process.

4. *Grounded Theory Strategies for Theory Construction*

The purpose of grounded theory strategies is to develop middle-range theories. Each stage of grounded theory analysis moves the work toward theoretical formulations. The techniques involved in three major strategies of the grounded theory method, coding, memo-making, and theoretical sampling, serve to both build theory and to distinguish grounded theory from other kinds of qualitative analysis.

4.1 *Coding Qualitative Data*

Coding is the first phase of analysis. Rather than apply extant concepts to their data, grounded theorists create codes as they study their data and, in abbreviated terms, define what they see in it. Coding helps the researcher begin to conceptualize what basic processes occur in the research setting or situation. While coding, grounded theorists ask the essential question: 'What is happening here?' Objectivist grounded theorists believe that the most significant processes in the setting will be apparent and that research subjects will inform the researcher of them. Constructivist grounded theorists believe that the most significant processes may be liminal and taken for granted by research subjects.

Grounded theory coding includes the following characteristics: (a) a focus on action and process, (b) a practice of line by line initial coding, (c) a simultaneous involvement in coding and further data collection, (d) an emphasis on analytic development rather than description. In initial or open coding the researcher proceeds to code line-by-line to begin breaking bits of data into specific types of action. After assessing which initial codes appear most frequently, grounded theorists then apply these codes to large amounts of data during a second stage of focused coding. Strauss and Corbin (1990, 1998) also recommend 'axial coding' which means coding for the dimensions of a category.

4.2 *Memo-writing*

Memo-writing constitutes the pivotal intermediate stage between coding data and drafting the theoretical analysis. Through memo-writing, grounded theorists fill out their codes and identify gaps in them. They define the code, delineate and analyze its properties, specify conditions under which it exists and changes, demonstrate its relationship to other codes, and weigh its significance for processes discovered in the field. An objectivist grounded theorist proceeds by emphasizing external characteristics of the code and the processes with which it is linked. Constructivist grounded theorists look for their own assumptions and implicit meanings, as well as those of their research subjects. Throughout data collection and analysis, constructivists remain more interpretive and less tied to overt behavior than their objectivist counterparts.

4.3 *Theoretical Sampling*

Theoretical sampling means sampling to develop or to refine emerging theoretical categories, not to describe populations chosen before the research begins. Thus, theoretical sampling occurs after the grounded theorist has defined and analyzed core theoretical categories through focused coding and memo-writing, but needs more data to develop, refine, and check the properties, boundaries, causes, and consequences of these theoretical categories. Thus, theoretical sampling builds precision, density, and complexity into the emerging theoretical statements and keeps them grounded in data.

When engaging in theoretical sampling, the researcher makes brief inquiries into fruitful areas for core categories of the emerging theory. Techniques of data gathering in theoretical sampling vary and are aimed to answer questions about the emerging theoretical categories. Theoretical sampling raises the conceptual level of the analysis and does so through systematic comparative analysis. The scope of the theory increases as the grounded theorist samples and compares different groups for their relevance to the theoretical category.

Gaps and questions may necessitate either returning to the same field setting or research problem or obtaining data in a different setting that possesses the characteristics the grounded theorist needs to test. Conducting comparative work across groups or field

settings allows the grounded theorist to build variation and complexity into the analysis. The emerging theory determines whether seemingly disparate groups make relevant sources to sample for theoretical comparison. For example, a study of establishing trust may sample such groups as mountaineering partners, con-men and women and their marks, and married couples. Theoretical sampling fosters a developing a full range of the properties of the category. Grounded theorists stop theoretical sampling when they have saturated their core categories and find no new data to shed light on them.

5. Current Emphases and Future Directions

The grounded theory method has significantly influenced the development of qualitative research in the social sciences and professions, particularly nursing and education. Strauss and his colleagues trained several generations of graduate students in sociology and nursing, whose students and colleagues have subsequently adopted the grounded theory method. In sociology, this method has, perhaps, had most influence in medical sociology and social studies of science (see Baszanger 1998, Clarke 1998, Star 1989). However, recent debates between Glaser and Strauss and Corbin, as well as larger epistemological debates about scientific inquiry, have intensified interest in examining and using the method. Although questions have been raised concerning objectivist premises within grounded theory, a constructivist revision of the method provides answers to them and points to new directions. In short, the grounded theory method can be expected to gain renewed support from those social scientists and professionals who value conceptual analysis of rigorous empirical qualitative research.

See also: Content Analysis; External Validity; Qualitative Methods, History of

Bibliography

Baszanger I 1998 *Inventing Pain Medicine: From the Laboratory to the Clinic*. Rutgers University Press, New Brunswick, NJ

Blumer H 1969 *Symbolic Interactionism*. Prentice-Hall, Englewood Cliffs, NJ

Charmaz K 1995 Grounded theory. In: Smith J A, Harre R, Van Langenhove L (eds.) *Rethinking Methods in Psychology*. Sage, London

Charmaz K 2000 Grounded theory: constructivist and objectivist methods. In: Denzin N K, Lincoln Y S (eds.) *Handbook of Qualitative Research*, 2nd edn. Sage, Thousand Oaks, CA

Clarke A E 1998 *Disciplining Reproduction: Modernity, American Life Sciences, and the Problems of Sex*. University of California Press, Berkeley, CA

Dey I 1999 *Grounding Grounded Theory*. Academic Press, New York

Glaser B G 1978 *Theoretical Sensitivity*. Sociology Press, Mill Valley, CA

Glaser B G 1992 *Emergence vs. Forcing: Basics of Grounded Theory Analysis*. Sociology Press, Mill Valley, CA

Glaser B G, Strauss A L 1967 *The Discovery of Grounded Theory*. Aldine, Chicago

Star S L 1989 *Regions of the Mind: Brain Research and the Quest for Scientific Certainty*. Stanford University Press, Stanford, CA

Strauss A L 1987 *Qualitative Analysis for Social Scientists*. Cambridge University Press, New York

Strauss A L, Corbin J 1990 *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage, Newbury Park, CA

Strauss A L, Corbin J (eds.) 1997 *Grounded Theory in Practice*. Sage, Thousand Oaks, CA

Strauss A L, Corbin J 1998 *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory Procedures and Techniques*, 2nd edn. Sage, Thousand Oaks, CA

K. Charmaz

Group Decision Making, Cognitive Psychology of

A group is a set of three or more persons who engage in joint activity directed at a common goal. For many groups, the goal is to make a decision. For other groups, the goal is achievement of a political, social, educational, or recreational objective. Regardless of their overall purpose, every group must make decisions so that member's actions are coordinated, and directed toward the common objective. This article presents major developments in the understanding of the cognitive processes involved in group decision making. It begins with a description of the forces leading to the current importance of group decision making, and proceeds to explain how the cognitive processes in decision making groups can be understood with a social information processing model. Next, some of the conceptual and methodological problems surrounding group decision making theory and research are covered. The theme throughout this discussion concerns differences among group members in terms of their information, beliefs, and decision preferences. The significance of two recent advances in addressing differences among group members, the Judge Advisor System and Information Sharing paradigms, is explained.

1. Groups as Information Processors

Group decision making has become increasingly important in the world of work for many reasons. One is that a greater proportion of jobs now involve cognitive work, so just as individual workers shifted from physical to cognitive work, so did groups.

Another force promoting the use of groups is the effort to increase representation and participation in organizational and societal decision making. Decisions that might have once been dictated by an authority are now made by a group representing various levels and segments of the population. Much of the current group research occurs in the context of organizational theory and practice (Sanna and Parks 1997)

The growth in the practice of using groups for decision making has been accompanied by a similar growth in the study of cognition in groups. The benefits of this research extend to all groups because all groups, regardless of their purpose, are dependent on cognitive processes for functioning. Simply put, people must coordinate their individual processes of attention, reasoning, judgment, memory, etc. during interaction in order to achieve their common goal. Group decisions are essential to coordination, and occur at many points during interaction. A group must choose its goal and the nature of the process that will be used to achieve it, select information deemed relevant, and integrate information, beliefs, etc. to determine the group's course of action.

The influence of 'real-world' practice is evident by the efforts at realism in modern group decision making research. Whereas early group research used games, trivial and 'eureka' decision problems, research tasks now capture such essential features of real decisions as the inability to know, with certainty, the quality of the decision at the time it is made. Motivational issues are taken more seriously by, for example, using incentives for making good decisions to reflect better the consequences facing persons in decision making groups. Other advances in group decision making research are the use of ongoing groups, and the study of interactions using technology (email, videoconferencing, etc.).

Part of the current scientific activity pertaining to cognitive processes in groups can be traced to the recent growth of cognitive psychology and the birth of social cognition (Fiske 1992), a subdiscipline of psychology devoted to the study of cognition about social objects (see *Cognitive Psychology: History*). Advances in these areas, combined with the increased prevalence of groups engaged in decision-making tasks, led to the modern view of conceptualizing groups as information processors. As described by Hinsz et al. (1997), key leaders in this development, the information processing approach to groups emphasizes the social processing of information, thereby integrating the group decision making and social cognition research approaches. In the social information processing view, groups exchange ideas, information, resources, strategies, etc. in the service of a processing objective. Research is directed at understanding attention, encoding, processing, storage, retrieval, feedback, and response processes in groups. Although recent group research tends to be guided by an individual-level information processing model, group and individual level processing are not assumed to be identical, and a pervasive research agenda is to understand the qualitative and quantitative differences between individual and group processing.

A primary difference between individual and group decision making arises from the greater complexity of group level processing of information, that is, group decision making requires coordination of individual members' cognitive processes and communication of the outputs of these processes. For example, consider how the memory problem in groups requires development of common strategies for transactive memory (Wegner 1987), a shared representation of the organization of information over members. It is through transactive memory that groups are able to identify and use the unique information that members previously knew or are able to recall from group interactions. Further, groups have more complex processes because they have a number of unique problems from which individual performers are free. At the start, group members face social uncertainty about the other group members. This refers to an inability to predict the behavior of other persons, making it difficult to know the extent to which they are willing or able to work toward the group goal. Group members may have different ideas about what the group goal is or should be. If so, it is necessary to negotiate them before purposeful interaction can occur. At any point in time, groups can have social goals that detract attention, time, and efforts from the task of making a decision. Even if all members are clearly focused on a single objective directly related to their decision problem, they are likely to have differences in information, beliefs, and preferences that need to be integrated in order to reach that objective.

2. *Levels of Behavior in Groups*

Despite the differences between individual and social information processing, it is sometimes possible to apply 'individual' cognitive constructs at the group level. Consider the case of human judgment. Judgment is the location of an object in an abstract representation such as a scale from one to nine, or a set of categories. An example of judgment in group decision making is a committee's recommendation for a salary offer. On occasion, group members' independent judgments may agree closely enough for the purpose at hand. But often, group members have cognitive conflict, that is, differences in beliefs. In some cases, groups may begin with varying judgments but end the task with a consensus group judgment of the amount. Yet, in other circumstances, members will adhere to discrepant judgments, and the judgment for the group must be a compromise such as the central tendency of the members' judgments. Then an accurate picture of the group's judgment needs to include some indicators of the variation in the members' judgments. The

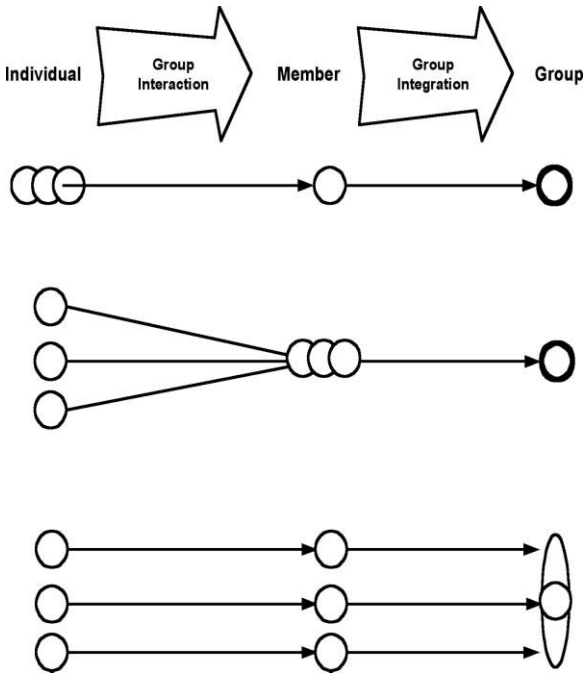

Figure 1
Three types of relationships between individual, member, and group judgments

problem is similar for cognitive tasks with discrete outcomes, such as deciding to uphold or overturn a law. When all are in agreement, there can be a group determination; if not, there can be only majority and minority member opinions. In general, we can describe members' cognitions at the group level if observations of the members of a group are indistinguishable, that is, every group member provides a response identical to that of every other group member. Under such conditions, we can speak about group goals, group confidence, group memory, and so forth. Otherwise, it is necessary to represent the diversity of group members' cognitions, with or without information on the cognitive value for the average or majority.

Figure 1 illustrates the distinction between the application of the cognitive construct of judgment to individuals, group members, and the group. Prior to interaction, there are three independent individual judgments representing individual information processing. These are similar in the first case, but different in the bottom two cases. These individual judgments or the beliefs surrounding them may be discussed during group interaction. Group interaction allows for social influence processes to transform the initial individual judgments into revised member judgments. Thus, the term 'member' is used to refer to individual members' judgments during or following group interaction to distinguish these judgments from the initial individual judgment, whether or not they have changed. In the first case, members learn that they are in agreement on a single judgment that becomes the consensus group judgment. Member judgments have become similar in the second case, but not the third. As in the first case, the integration of member judgments into a single group judgment is straightforward in the second case. But in the third case shown at the bottom of the figure, there is no consensus group judgment, only an average judgment for the group. There may be no alternative but to use this average judgment to represent the group judgment. But an accurate picture of the group's judgment must include the persistent differences among member judgments.

3. *Problems in Conceptualizing Groups*

The study of cognition in decision making groups extends to many aspects of the groups behavior. In practice, groups are comprised of individuals who can have considerable cognitive conflict in their processing of information relevant to their interactions and objectives, that is, the various persons in a group may have different schemas or frames for the problem at hand, attend to different problem features, or know or recall different information (see *Mental Models, Psychology of*). Due to such discrepant views within the group, one of the first problems encountered by group researchers is appropriately representing a group, that is, research needs to begin with a model of the group that describes features of its structure and function that are important to the research question. This is often difficult because the members themselves can differ with respect to their own mental representations of the group. They can lack agreement in their beliefs about who is in the group, what its goals are, and how the group should proceed to achieve them.

There are four ways in which group researchers can deal with the problem of group representation. The first is to use the researcher's model, with the implicit assumption that it represents not only the reality of the group, but also the members' mental models of the group. For example, a group may be characterized as egalitarian without any assessment of members' beliefs about status and authority within the group. Second, the researcher can elicit the perspectives of a representative of the group. Although it is often convenient to collect information from only one group member, that person may not accurately represent the entire group. A third approach to the problem of group representation is to identify and use members' cognitive representations of the group. Typically these are represented mentally by a few discrete structures or categories. A final solution is to impart a model to group members prior to group formation and interaction. This latter option is especially popular in the laboratory study of *ad hoc* groups. With this approach,

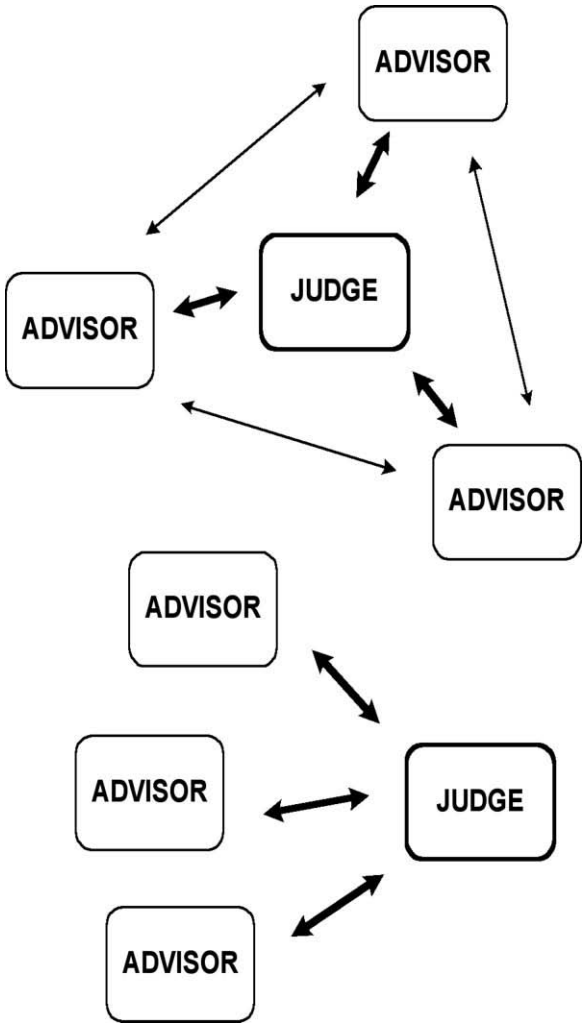

Figure 2
Illustrations of Judge Advisor Systems showing role differentiation in groups

group members may be informed that they were selected for the group because of their knowledge in a given domain, and instructed to discuss a problem until they reach a consensus on its solution. The advantage of this last approach is that data collection begins with a common understanding within the group—and between the group and the researcher—about the group's composition, task, and procedures.

One aspect of groups in natural settings that is not captured by most laboratory research is role differentiation (see *Decision Making (Naturalistic), Psychology of*). What may appear to an outsider or the researcher as five persons sharing information and power to achieve their goal may be better represented as something else. Many groups are structured with a leader who is authorized to act on behalf of the group, or with a single member who emerges to take on the role of leading the group to its goal. In many circumstances, information processing becomes social when an individual seeks advice from others. The traditional group model does not adequately describe either the group with a leader or the individual seeking advice from others. A recent paradigm that accounts for role differentiation reconceptualizes the group as a Judge Advisor System. The Judge Advisor System, illustrated in Fig. 2, is a model of social decision making that was designed to distinguish between the member responsible and accountable for making a decision, known as the Judge, and the other members, called Advisors, who provide input to the Judge (Sniezek and Buckley 1995, Sniezek and van Swol 2001). The top diagram shows a group where each member can interact with any other, but the Judge is the leader who is responsible for the goal. In the bottom diagram, the Judge is also responsible for the goal, but seeks and interacts with a number of independent Advisors to elicit resources needed for goal attainment. Whenever a decision maker interacts with another person in the course of making a decision, the process becomes social, and cannot be portrayed accurately by individual decision making theory and research.

A central cognitive activity according to this view is information search. The Judge must decide to solicit advice, whom to consult, and when to stop. Research on advising in decision making has become an international scientific activity (e.g., Harvey et al. 2000). Although similar to information search from non-social sources, this social information search differs in that the social sources are capable of strategic and reactive behavior in the process.

4. *Cognitive Conflict in Groups*

At this point the reader can reasonably wonder why cognitive tasks should be done by groups and not a single individual. With so many coordination costs groups must be much more inefficient than individuals. Nevertheless, there are many reasons for using groups for cognitive tasks even when the costs of coordinating members' difference are high. Some of these reasons are unrelated to the task itself; for example, democracy requires broad representation of views regardless of costs. In some situations, the task is simply not divisible; multiple persons are required. Yet, the primary motivation for using groups to make decisions is that the problem is important, and there is uncertainty about the quality of the various decision alternatives.

Ironically, the very same cognitive conflict in group inputs that creates challenges often provides the rationale for the use of groups. With a diversity of inputs, there is the potential for the group to follow the

one that best serves members' collective interest. Put simply, there are more ideas and strategies from which to choose. The best input in a group will, by definition, be superior to that of the average member, and superior to those of most members. For the most part, groups performing cognitive tasks have been found to produce outcomes that are better than those of the average member, though not as good as those of the best member. The extent to which the group outcome is better than that of the average individual input increases with the variation in group members' inputs. An important point is that this result has been obtained even when analyses control for the initial bias in group members' inputs. This assures us that the benefit of input variation cannot be attributed to groups simply discounting their poorest member.

In some cases, group outcomes can be even better than those of any group member (Sniezek and Henry 1989). Clearly the potential for such dramatic improvements in results can greatly outweigh the inefficiencies and other costs of group information processing. However, this potential benefit comes with the risk that groups can sometimes perform below the level of the poorest individual member. Therefore, a more realistic picture is that groups have the potential to be superior information processors compared to individuals working independently, but they simultaneously pose the risk of costly, inefficient, and seriously inferior performance.

Heterogeneity of inputs in a group is necessary but not sufficient for process gain. A second requirement is that those inputs be shared with other group members. One stream of research on information sharing in groups, originated by Stasser and co-workers, identifies some barriers to the sharing of information that is unique to individual group members (Stasser and Titus 1985). For starters, there is a sampling advantage for common information that means it will be more likely to be discussed in the group than unique information. Even if both common and unique information is discussed, groups tend to weight common information more than unique information. Unique information that is contributed to discussion is influential only on the judgment of the group member who contributed it. While the reduction of belief heterogeneity is necessary for group consensus, it is not always easy. This is especially true when disagreements go beyond differences in preferences to differences in preference structures. In the latter case, members have cognitive conflict, that is, differences concerning the interpretation or importance of information.

The picture that emerges from research on cognitive processes in groups shows a series of challenges unique to social information processing. Most of these problems arise from the fact that group members are not replications of an individual but varying individuals. We have seen that the persons comprising a group can vary in their beliefs about the nature of the group and its goals as well as their processing objective, focus of attention, encoding format, and memory processes. The essence of the group challenge is to reconcile member differences so that coordinated action is possible.

See also: Group Decision Making, Social Psychology of; Group Processes in Organizations; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Groups, Sociology of; Organizational Decision Making

Bibliography

Fiske A P 1992 The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review* **99**: 689–723

Harvey N, Harries C, Fischer I 2000 Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes* **81**(2): 252–73

Hinsz V B, Tindale R S, Vollrath D A 1997 The emerging conceptualization of groups as information processors. *Psychological Bulletin* **121**(1): 43–64

Sanna L J, Parks D C 1997 Group research trends in social and organizational psychology: Whatever happened to intragroup research? *Psychological Science* **8**(4): 261–67

Sniezek J A, Buckley T 1995 Cueing and cognitive conflict in judge-advisor decision-making. *Organizational Behavior and Human Decision Processes* **62**(2): 159–74

Sniezek J A, Henry R 1989 Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes* **43**(1): 1–28

Sniezek J A, van Swol L 2001 Trust, confidence, and expertise in Judge Advisor Systems. *Organizational Behavior and Human Decision Processes*

Stasser G, Titus W 1985 Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology* **48**: 1467–78

Wegner D 1987 Transactive memory: A contemporary analysis of the group mind. In: Mullen B, Goethals G R (eds.) *Theories of Group Behavior*. Springer-Verlag, New York, pp. 185–208

J. A. Sniezek

Group Decision Making, Social Psychology of

People do many things in groups, of which one of the most significant is making decisions (see Baron et al. 1992, Brown 2000, Hogg and Tindale 2001, Levine and Moreland 1998, Moreland and Levine 1994). Indeed decision making is one of the most important functions of groups (see also *Group Processes, Social Psychology of*). Our lives are largely guided by decisions we make with others (e.g., groups of friends), or are made by groups of others (e.g., parliament), and many of us also spend a significant portion of our

working lives engaged in group decision making (e.g., committee meetings). Group decision making is also influenced in varying ways by the wider intergroup context in which the group exists—that is, by awareness of the existence of other groups (i.e. outgroups) with which one's own group may be in competition (see also *Intergroup Relations, Social Psychology of*). Finally, although group decision making usually involves people getting together to exchange information, it can also involve virtual groups making decisions via computer.

Social psychologists have long been interested in processes involved in group decision making, and whether groups make better or different decisions than do individuals—it is a field with significant applied relevance. There are many different ways to approach the study of group decision making (e.g., social communication approaches, social influence perspectives), there are many different facets of group decision making (e.g., persuasion, negotiation), and there are many different group decision making phenomena (e.g., juries, groupthink). Group decision making almost always involves psychological movement of group members from a relative diversity of positions to a more consensual position or to agreement on a group position. Group decision making reduces opinion differences within groups. Group decision-making can therefore be considered a social influence phenomenon (even a conformity phenomenon) that often also involves discussion, argument, persuasion, and leadership (see also *Group Processes, Social Psychology of*).

1. Social Combination Models

Social combination models relate the distribution of initial opinions in a decision making group to the final group decision (Stasser et al. 1989). These models, unlike social communication models, are not really concerned with interpersonal communication processes that transform individual positions into a group decision, but rather with the decision making rules that determine how group members combine their preferences (initial positions) into a group decision—the focus is on group preferences and their change over time. There is now an array of social combination models (Baron et al. 1992). Some are complex computer-simulation models, whereas others are more immediately related to real groups.

1.1 Social Decision Schemes

Social decision schemes are explicit or implicit decision-making rules that groups can adopt to decide how to combine information into a group decision (Davis 1973). Knowledge of the initial distribution of individual opinions in the group, and what rules the group is operating under allow reliable prediction of the final group decision. These rules include: (a) unanimity—discussion pressurizes deviants to conform; (b) majority wins—discussion confirms the majority position, which is then adopted as the group position; (c) truth wins—discussion reveals the position that is demonstrably correct; (d) two-thirds majority—unless there is a two-thirds majority the group is unable to reach a decision; (e) first-shift—the group adopts a decision consistent with the direction of the first change in opinion by any member of the group. For intellective tasks (where there is a demonstrably correct solution, as in, e.g., a mathematical puzzle) groups tend to adopt the truth wins rule, and for judgmental tasks (with no demonstrably correct solution, e.g., for aesthetic preference) the majority wins rule.

Rules vary in (a) strictness—i.e., the degree of agreement required by the rule (unanimity is very strict and majority less strict), and (b) distribution of power among members—i.e., authoritarian rules concentrate power in one member, whereas egalitarian rules spread power evenly. Generally, the stricter the rule the less the power concentration—unanimity is very strict but low in power concentration, whereas two-thirds majority is less strict but has greater power concentration. The type of rule adopted can, mainly due to its strictness, influence not only the group's decision but also members' preferences, their satisfaction with the group decision, the perception and nature of group discussion, and members' feelings for one another. For example, stricter decision rules can make the process of reaching final agreement exhaustive, slow, and difficult, but can enhance liking among members, and satisfaction with the quality of the decision.

1.2 Social Transition Schemes

Kerr's (Stasser et al. 1989) social transition scheme model focuses on the actual pattern of member positions that a group, operating under a particular decision rule, moves through *en route* to its final decision. To track opinion transitions, member opinions are monitored during the process of discussion, either by periodically asking the discussants, or by getting them to note any and every change in their opinion. These procedures can be intrusive, and therefore one issue concerns the extent to which they affect the natural ongoing process of discussion.

2. Jury Decision Making

Social combination models have often been tested with simulated juries, as the laboratory analogue has similar characteristics to an actual jury (e.g., *ad hoc* collection of people who initially do not know one another). Of course, juries are important decision making groups in their own right (Hastie et al. 1983), not only because they are held up as a symbol of all that is democratic, fair and just in a society, but also

because of the consequences of their decisions for defendants, victims, and the wider community (see also *Justice: Social Psychological Perspectives*). For instance, the 1992 Los Angeles riots were sparked by a 'not guilty' jury verdict, and the murder trial of O. J. Simpson in the mid-1990s became a national obsession in the USA. Juries are groups, and thus prey to all the consequences of group processes involved in group decision making.

Research on juries indicates that the initial majority position usually prevails in determining the final verdict. Close cases, which are associated with an even split of initial juror positions, tend to produce a hung jury. Hung juries, which are considered a relatively favorable outcome from the perspective of the defendant (the case is often dismissed), are less likely as jury size decreases from 12 to six persons, and if the decision rule changes from unanimity to some form of majority. Because jurors are instructed to presume the defendant innocent while reasonable doubt remains, juries have a leniency bias. This bias tends to surface (via acquittal) in close cases where there is an even split of opinions.

Juries are decision making groups that make a decision about a person. Thus, characteristics of the defendant and of the victim can affect the jury. Physically attractive defendants are more likely to be acquitted or to receive a lighter sentence, though biases can be reduced by furnishing sufficient factual evidence, presenting the jury with written rather than *in vivo* testimony, or explicitly directing the jury to consider the evidence alone. In the USA (and probably elsewhere) race can also affect the jury. Blacks are more likely to receive prison sentences, and people who murder a black are less likely than those who murder a white to receive the death penalty. Another issue is the influence of laws and penalties on the jury. Harsh laws with stiff penalties (e.g., the death penalty) tend to discourage juries from convicting. The jury foreman is important in guiding the jury to its verdict, because he/she occupies the role of leader. Research suggests that the foreman is likely to be someone of higher socioeconomic status, who has had previous experience as a juror, or simply initially occupies the seat at the head of the table.

3. *Risky Shift and Group Polarization*

Folk wisdom has it that people in groups make inherently more conservative decisions than they do on their own—individuals are likely to take risks, while group decision making is a tedious averaging process that errs towards caution. This is consistent with traditional social psychological perspectives on conformity and group influence processes—through mutual influence, people's attitudes tend to converge on the group mean (see also *Social Influence, Psychology of*, Turner 1991). However, under some circumstances groups can make more risky decisions than individuals. Specifically, groups that already favor risk to some extent (i.e., the mean of the members' prediscussion opinions favors the risky pole of a decision making dimension) tend, through group discussion, to arrive at a group decision that is even more risky than the mean. This has been called risky shift. Later research showed that if the group initially favors caution then there is a cautious shift, and indeed that group decisions can be more extreme than the mean of the members' prediscussion opinions on a range of decision making dimensions that do not involve risk or caution, provided that the group initially tends towards one direction. Hence, the effect was renamed group polarization. It is defined as a tendency for groups, as a consequence of discussion, to make decisions that are more extreme than the mean of individual members' initial positions, in the direction already favored by that mean.

Since 1960, research has produced many different explanations of group polarization. They can be simplified to three major perspectives: persuasive arguments, social comparison/cultural values, and self-categorization.

(a) *Persuasive arguments theory*. Focuses on the persuasive impact of novel information. A group that leans in one direction will generate, during discussion, more novel arguments favoring that direction than the opposing direction, and, because novel arguments are persuasive, will cause members to become more committed to their original position and thus the group to shift more in that direction.

(b) *Social comparison/cultural values*. Focuses on people's motivation to avoid social censure and to seek social approval. Group discussion reveals the group's views, which are assumed by members to reflect what is socially desirable or culturally valued, and so group members publicly shift in the direction of the group in order to gain approval and avoid disapproval.

(c) *Self-categorization theory* (see also *Social Identity, Psychology of*). Argues that people in decision making groups psychologically identify with the group and use the group's position, which is cognitively represented as a prototype, to define themselves. The prototype is formed not only to capture similarities within the group but also to distance the group's position from outgroup or non-ingroup positions. Thus, extreme groups form polarized prototypes. The process of categorizing oneself as a group member (self-categorization) causes people to conform to this prototype and, thus, produces group polarization.

4. *Groupthink*

Groups sometimes employ defective decision-making procedures that produce poor decisions that can have disastrous consequences. Janis (Janis and Mann 1977) coined the term groupthink to define a style of thinking in cohesive groups in which the desire to reach

unanimous agreement overrides the motivation to adopt proper rational group decision-making procedures. The main cause of groupthink is excessive group cohesiveness, but there are other factors that relate to structural faults in the group (e.g., lack of impartial leadership) and to the immediate decision making context (e.g., high stress). Together, these factors generate a set of symptoms that are associated with defective decision making procedures. For example, there is inadequate and biased discussion and consideration of objectives and alternative solutions, and a failure to seek the advice of experts outside the group. A number of American foreign policy decisions with unfavorable outcomes (e.g., the 1961 Bay of Pigs fiasco, the unsuccessful defense of Pearl Harbor in 1941) have been attributed to groupthink in Presidential decision-making committees. Groupthink is probably a lot more common than one would like to think. It is a serious pitfall of group decision making, and groups need to be particularly vigilant in order to avoid it. A special decision making procedure, called the nominal group technique, is available to help groups avoid groupthink.

5. *Group Remembering and Transactive Memory*

Group decision making involves collective recall of information (see Clark and Stephenson 1995). Group remembering is not simply a retrieval activity in the pursuit of veridical recall, rather it is a constructive process where comparisons among group members shape the resultant group memory. Complex ideas can sometimes be transformed through group discussion into much simpler, and thus distorted, shared knowledge—a process called social representation, that is similar to rumor transmission. Group remembering may not be more accurate than individual remembering, but may be subject to implicit or even intentional bias.

Another feature of group remembering is transactive memory. Groups often need to bring a large amount of information to bear upon a decision, and thus they develop a differentiated memory structure where individuals or subgroups specialize in different memory domains. Where this occurs people not only have to remember their own part of the puzzle, but they also need to know who specializes in what memories—they develop a transactive memory structure. Research shows that group functioning and decision making is compromised where people have not acquired a transactive memory, or where such a system has not developed in the group.

6. *Intergroup Decision Making*

Decision making groups often contain factional rifts. In fact, many decision making groups can be considered interactive decision-making contexts where representatives from different groups are brought together to reach a mutually satisfactory agreement (Brown 2000). In this sense, the group is actually a context for intergroup decision-making aimed at reducing intergroup conflict (see also *Intergroup Relations, Social Psychology of*; *Conflict and Conflict Resolution, Social Psychology of*). Examples of intergroup decision making include negotiations between nations, hostage crises, union/management confrontations, and parliament. The intergroup dimension introduces an additional dynamic to group decision making, because decision makers are charged with the responsibility to argue on behalf of their group, not themselves (Pruitt and Carnevale 1993).

When people are bargaining on behalf of groups to which they belong they tend to bargain more fiercely and uncompromisingly than if they were bargaining for themselves. The effect is enhanced when negotiators are aware that they are being observed by their constituents, either directly or through the media. This bullish strategy of intransigence is less likely to secure a satisfactory compromise than a more interpersonal orientation in which both parties make reciprocal concessions. Direct negotiation between group representatives is therefore likely to reach an impasse where neither group feels it can compromise without losing face.

To break the deadlock a third party can be recruited to mediate between the groups (Pruitt and Carnevale 1993). To be effective, mediators should have power, must be seen by both groups to be impartial, and the groups should already be fairly close in their positions. Biased mediators are ineffective because they are not trusted, and weak mediators are ineffective because there is little pressure for intransigent groups to be reasonable. Although mediators have no power to impose a settlement they can help in several important ways: (a) they can reduce the emotional heat associated with deadlock; (b) they can help reduce misperceptions, encourage understanding, and establish trust; (c) they can propose novel compromises that allow both groups to appear to win—i.e., change a zero-sum conflict (in which one group's gains are the other group's losses—the more one gains the more the other loses) into a nonzero-sum conflict (i.e., both groups can gain); (d) they can help both parties make a graceful retreat, without losing face, from untenable positions; (e) they can inhibit unreasonable claims and behaviors by threatening to publicly expose the group as being unreasonable; and (f) they can reduce intragroup conflict and thus help a group clarify its consensual position.

Where mediation fails, the last resort is arbitration, where a third party is invited to impose a mutually binding settlement. Arbitration can backfire because both groups adopt outrageous final positions in pursuit of a favorable compromise. One way to combat this is through 'final-offer arbitration,' where the third party chooses one of the final offers. This tends to encourage more reasonable final positions.

7. Motivation and Status

Group decision making is influenced by how hard members are prepared to work towards a group decision. Research on performance motivation suggests that people are often less motivated to work hard in groups than they are on their own—a phenomenon called social loafing (see also *Group Productivity, Social Psychology of* and Baron et al. 1992). A final aspect of group decision making is that some people have greater influence over the decision making process because they have higher status or occupy a leadership position. Research on status within groups shows that influence over group decisions and group action stems from possession of specific status characteristics, abilities and skills directly related to the group's function, and of diffuse status characteristics, membership of social categories that have high status in society at large (see Ridgeway 2001). Finally, leadership research suggests that effective leaders are people who can instill and inspire others with their own vision for the group, so called, transformational leaders who can transform a group so that its decisions reflect the leader's preferences (see also *Leadership, Psychology of*).

8. Future Directions

Due to its applied relevance the study of group decision making is always popular. However there are some key directions for future research. One such direction is the study of shared cognition—how groups create shared memories and representations and influence people to think in a collective manner. Another direction is the study of how intergroup relations impact decision making in groups. And finally, the study of computer-mediated decision making by virtual groups is a new area that is already attracting great attention.

See also: Decision Making (Naturalistic), Psychology of; Decision Making, Psychology of; Decision-making Systems: Personal and Collective; Group Decision Making, Cognitive Psychology of; Group Processes in Organizations; Group Processes in the Classroom; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Leadership, Psychology of; Organizational Decision Making

Bibliography

Baron R S, Kerr N L, Miller N 1992 *Group Process, Group Decision, Group Action*. Brooks/Cole, Pacific Grove, CA

Brown R 2000 *Group Processes: Dynamics Within and Between Groups*, 2nd edn. Blackwell, Oxford, UK

Clark N K, Stephenson G M 1995 Social remembering: individual and collaborative memory for social information. *European Review of Social Psychology* **6**: 127–60

Davis J H 1973 Group decision and social interaction: A theory of social decision schemes. *Psychological Review* **80**: 97–125

Hastie R, Penrod S D, Pennington N 1983 *Inside the Jury*. Harvard University Press, Cambridge, MA

Hogg M A, Tindale R S (eds.) 2001 *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford, UK

Janis I L, Mann L 1977 *Decision-Making: A Psychological Analysis of Conflict, Choice, and Commitment*. Free Press, New York

Levine J M, Moreland R M 1998 Small groups. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*. McGraw-Hill, Boston, Vol. 2

Moreland R L, Levine J M 1994 *Understanding Small Groups*. Allyn and Bacon, Boston, MA

Pruitt D G, Carnevale P 1993 *Negotiating and Social Conflict*. Brooks/Cole, Monterey, CA

Ridgeway C L 2001 Social status and group structure. In: Hogg M A, Tindale R S (eds.) *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford, UK pp. 352–75

Stasser G, Kerr N L, Davis J H 1989 Influence processes and consensus models in decision-making groups. In: Paulus P B (ed.) *Psychology of Group Influence*, 2nd edn. L. Erlbaum, Hillsdale, NJ

Turner J C 1991 *Social Influence*. Brooks/Cole, Pacific Grove, CA

M. A. Hogg

Group Processes in Organizations

The success or failure of social and commercial organizations depends not only on specific individuals but also on the quality of the interaction and communication within groups (intragroup relationships) and between groups (intergroup relationships). Groups inside organizations influence individual performance, motivation, and work satisfaction as well as productivity, decision quality, learning, innovation, absenteeism, and fluctuation. Group processes are a central subject of social psychology as well as organizational psychology. During the 100-year history of group research, a substantial amount of literature has been amassed (for current reviews, see West 1996, Hinsz et al. 1997). In the following, we will concentrate on a selection of group research issues that we believe are of particular relevance to the understanding of group processes in organizations.

1. Definition and Structural Dimensions of Groups

Groups are defined as a number of people who are oriented toward the same or similar goals and who interact and communicate in order to achieve these goals. All groups share at least three structural principles: norms, cohesion, and role differentiation.

Norms. Norms can be defined as the shared expectations of the group members about how they should act or behave in certain situations. Norms make the behavior of group members more easily predictable and thus help to clarify social relationships.

Cohesion. When group membership is highly attractive—because goal achievement and need satisfaction are perceived to be achieved better in the group—group members are motivated to sustain the contact with the group. The sum of all the forces that increase group attractiveness is called 'cohesion.'

Role and status. Role differentiation in groups is the social–psychological basis for work differentiation and organization. Role is defined as the expectation of the group members about the appropriate behavior in a certain position. Role-conformant behavior is regularly positively sanctioned, and deviations are sanctioned negatively. Status means that people differ in the hierarchy of prestige, power, influence, and privileges. High-status group members are allowed to deviate from group norms more strongly without being sanctioned than low status members (idiosyncrasy credit).

2. Conformity and Dissent

Conformity and dissent within groups are central aspects of group functioning. Behavior according to the group norm is necessary for relatively conflict-free interaction. Normally, group pressure brings the groups' behavior and attitudes into line with the norms. This is important for group functioning and in order to achieve the relevant goals. However, normative behavior can also be dysfunctional, as has been shown by Asch's studies (1951) in which group members expressed clearly wrong and irrational perceptions just because of the groups' pressure towards conformity.

2.1 Groupthink: The Conformity of Groups

Groupthink describes the phenomenon by which existing information that is relevant to a certain decision is not processed by the group members in a rational and systematic way, but instead information processing is biased in the direction of a preferred decision alternative within the group. This reduces the quality of information processing within the group and increases the risk of faulty decisions. Analyzing historical political disasters following group decisions, such as, for example, Pearl Harbour, the American invasion of Cuba, or the Vietnam War, Janis (1982) has specified the symptoms of *groupthink*. These symptoms can be observed in the economical field too, such as, for example, in project teams, or on boards or committees. The risk of groupthink is amplified when groups are homogeneous (absence of minorities and contradicting views), when the group structure is markedly hierarchical, and when there is a strong commitment to particular decision alternatives before group discussion begins.

The symptoms of groupthink include the following (cf. Janis 1982). (a) *Illusion of invulnerability*; that is, high conformity concerning the views of the group members and concerning their thinking styles makes them optimistic and overconfident. (b) *Willingness to take high risks* resulting in warnings, potential hazards, and other forms of negative feedback being ignored or not taken seriously. This behavioral tendency in groups serves the purpose of avoiding questioning one's own basic beliefs and value systems when facing a challenging decision. (c) *Stereotyping of members with a different opinion* by labeling them as incompetent. (d) *Pressure towards conformity* exerted by the formal leader or the majority of the group in order to stabilize the group's homogeneity. Conformity pressure helps members to neglect counter arguments more easily and supports self-censorship processes which makes group members (falsely) assume that silence means approval.

There is considerable empirical evidence regarding prevention and intervention strategies for counteracting groupthink. Some examples are (for a summary of these interventions, see Frey et al. 1996):

(a) the group leader acts as an impartial coordinator,

(b) systematically strengthening self-confident minority members,

(c) bringing external experts into the group,

(d) implementing artificial conflict in the group decision process (e.g., by devil's advocacy or dialectical inquiry techniques),

(e) holding each group member individually responsible for the decision process and the outcome,

(f) creating different subgroups separately engaging in decision making.

2.2 Minority Influence

The problem of conformity and homogeneity in groups is that it may come at the expense of divergent thinking, which can be helpful when scrutinizing premature conclusions, and innovative thinking. Due to conformity pressure, the default information processing style in groups is consent oriented. Minority influence, however, often generates productive dissent.

According to Moscovici (1976), minority members are able to oppose the pressure towards conformity in groups under certain conditions—and are thereby able to influence the attitudes and behavior of the majority. To achieve this, minority members should present their point of view in a consistent behavioral style, though not rigidly (flexibility of negotiation and argumentation), while signalizing a willingness to compromise.

Why can minorities exert an influence on majority members? According to Moscovici's Conversion Theory (cf. Moscovici 1980), minority members induce an internal cognitive conflict within members of the majority; majority members then ask themselves *What is in fact the right course?* This leads to an intensification of information processing, and is accompanied by divergent and more creative thinking styles, resulting in a less biased information search and greater openness for the minority position. In contrast to this, majorities exert a pressure to conform, producing social conflict between people or subgroups of people within the group rather than a cognitive conflict. Group members ask themselves *Which person is right now*? This leads to a more superficial processing of information, and is accompanied by conventional and conformist styles of thinking.

What is the relevance of the minority research for group processes in organizations? Very often minority groups or minority persons have the feeling that they *cannot* change a big organization. Taking minority research into account, the opposite is in fact true. Minority influence in organizations is a precondition for scrutinizing common patterns of thought and practice, and thus is an important ingredient for organizational learning, change, and innovation. The interesting thing is that a minority can be influential independently of the hierarchical structure.

3. *Bringing Together Individual Contributions*

How can individual contributions be brought together in an optimum way so that production losses (e.g., loss of motivation and coordination loss) can be minimized, and procedural gains or synergies (e.g., motivating each other, learning from each other, and sharing innovations) are maximized?

3.1 *Activating or Impeding Individual Performance*

Whether social inhibition or social facilitation occurs via the mere presence of other persons is a question of the complexity and newness of the task to be performed. If other persons are present, difficult or new tasks are carried out less satisfactorily, since these tasks will lead to uncertainty and fear of negative feedback ('social inhibition'). Simple or well-trained tasks, on the other hand, are usually carried out better in the presence of other people, as the active subject can expect success and a positive judgment from his/her colleagues ('social facilitation').

In the case of social inhibition, no matter whether the task performed is difficult or not, loss of motivation is involved. For example, if the individual contributions are not perceptible in the groups' overall output, or their importance for group productivity is not clearly seen, social loafing occurs. When group members expect other people in their group to be 'free-riders,' they may be inclined to conclude that their own individual contribution might be too high compared with the contributions of other group members. Consequently, individual performance will be decreased since no one wants to be left to carry the can; this is called the 'sucker effect.' Additionally, if group members fear negative comments from others, or if they are not sure about the norms and standards of evaluation as a whole, they will also withhold optimal performance (for fear of a negative evaluation). These motivational impediments on optimum individual performance can be met by transparency of performance feedback, by conveying concrete individual as well as collective performance norms, by fostering intrinsic motivation (e.g., interesting and challenging tasks), by strengthening group cohesion in combination with a high achievement norm, and by pointing out how important each and every individual contribution is for the group's productivity (for an overview see Sheppard 1993).

On the other hand, when group cohesion is high and the group's task is highly important, one can observe that groups will increase individual performance via motivation gain. If the group members' capabilities do not differ very much, and if the success of the group as a whole is very important for all the members, very strong and competent persons often try to compensate for the weaknesses of their colleagues by trying even harder to do their best. The weaker group members then likewise try to increase their output, as they do not want to be the 'losers' bringing the whole team down. These gains in motivation are especially likely if group cohesion is high—that is in case of a very attractive group (cf. Brown 1992), in case of a high mutual sympathy (cf. Latané 1986), and if the group's task is highly important (cf. Williams and Karau 1991).

3.2 *Task Structure and Social Combination of Individual Resources*

According to a very well-known model of group performance in social psychology (i.e., Steiner 1972), the *de facto* group performance is a function of the potential group performance, minus procedural losses, plus procedural gains. In order to find out whether groups are working to their full potential, one needs to determine the potential level of group performance. For this, one needs to know the individual capabilities of the group members and the type of task at hand.

In general, the structure of work tasks may be *additive*, in which case the potential group performance is the sum of all the individual performances, for example in the case of brainstorming. If the task is structured in a *compensatory* way, then the potential group performance is the result of the mean of all individual performances, for example in the case of

guessing tasks or majority decisions). If work tasks are *disjunctive*, the group potential is the result of the best individual performance (for example in the case of problem-solving tasks or decision problems). Thus, determining a group's potential level of performance requires consideration of the group task. In order to further determine whether and to what extent a group is working below (or sometimes even above) its potential, the actual group performance is obtained and compared with the potential level of performance. The differences tell us to what extent group-induced procedural losses and procedural gains have occurred.

3.3 Synergy in Groups

Practitioners very often speak about 'group synergy.' This means that a group is more than the sum of all its parts, or more precisely: that a group generally does better than its aptest single member. However, synergy effects have only rarely been demonstrated in social–psychological group research (for an overview, see Hill 1982, Brown 1993). Stroebe et al. (1992) have argued that the common belief in group synergy may rest on an overestimation of one's own output in group contexts. For example, where brainstorming is concerned, single individuals usually believe they are better and more productive when they work in groups—probably because they feel 'enriched' by the ideas of the others and because many new perspectives emerge. However, Stroebe et al. (1992) have shown that individual group members are inclined to claim other group members' ideas as their own body of thought (and, hence, believe in 'having produced a lot'). Consequently, one feels being 'more productive' as a participant of a brainstorming group compared with working alone. Each group member extrapolates this impression onto the other persons, and in the end believes that the group is much more efficient than the sum of its components (i.e., members) working alone. This overestimation feeds the idea of 'group synergy.'

3.4 Top Performance of Work Teams

There is growing evidence that top performances in groups occur despite dominant forces such as conformity, groupthink, loss of motivation, and coordination loss (Levine and Moreland 1990, West 1996). A high decision quality and top performance of teams is likely under the following conditions.

(a) Homogeneity of values and rules exists, meaning that team members agree on rules for dealing with each other; they are able to communicate honestly and to respect each other, and show loyalty.

(b) A heterogeneity of talents, experiences, education, and backgrounds exists within the team.

(c) The team has a vivid commitment to excellent output and an ethos of achievement. All members take responsibility for team success.

(d) Team members fit together technically and personally.

(e) Team members are able to profit from their individual strengths.

(f) Orientation towards the best practice idea (benchmarking). This technique means comparing oneself with best practice in one's own branch or with the best in the world, but also with the best within the organization.

The team is able to use the tool of team reflection, meaning regular joint reflection on social and task issues. What is good and should be sustained, what is bad and should be improved (West 1996).

4. Group Dynamics

Group interaction may be viewed from three different theoretical perspectives. First, one can look at socio-emotional relations where cohesion, role-forming, open communication, the building of trust, and linear models concerning the development of groups are relevant (forming, storming, norming, performing; Tuckman and Jensen 1977). Second, one can look at the development of goal-directed and output-oriented groups, mainly to be found in work groups in organizations where goal finding, role differentiation, leadership, quality of decisions, coping with problems, and action-oriented models of group dynamics (cf. Pinto and Prescott 1988) are relevant. Third, one can look at the development of interpersonal and task-related structures within and between groups, in organizations where high-quality work, customer orientation, and innovation are sought to be improved. Scientific progress has been made in all these fields (cf. Brodbeck 1996, Hinsz et al. 1997). In the following we will discuss two subjects that are particularly related to the development of goal and task-oriented group dynamics: group learning and approaches to team development in practice.

4.1 Group Learning

Group learning in general has not been explored very well to date (Argote and Epple 1990). However, transactive knowledge systems have been researched quite thoroughly; in other words the way in which a group chooses to encode, store, and retrieve information systematically. Transactive knowledge systems make information accessible for team members, even if this information has been (physically) stored by other members of the group (Wegner 1987). According to Liang et al. (1995), teams that have been trained to do complex tasks together (i.e., who have been trained as a team) do better afterwards than groups where the members have been trained individually to solve the task. The authors hold that this lead in the output is due to the fact that teams who have been trained as a group have established transactive knowledge systems. These structures improve the group's performance, for

example by means of group processes such as feedback about weaknesses, helping each other out, sharing information that others need right now. A model of group learning as a multilevel phenomenon; that is, on the individual level, the group level, and respective interactions between these two levels; has been presented by Brodbeck and Greitemeyer (2000a, 2000b).

4.2 Team Development

Typical techniques of team development usually deal with the analysis and layout of group processes in order to increase work performance, personality development, team climate, and cooperation between departments. Four team development approaches can be distinguished. (a) The interpersonal approach tries to increase competence as well as cooperation by promoting self-reflection within groups. (b) The goal-oriented approach tries to increase motivation, performance, and subjective well-being by optimizing goal finding and goal negotiation, for example by way of participation, mutual feedback, and intelligent reward systems. (c) The role approach tries to reduce misunderstandings and conflicts (due to poor coordination or bad communication): underlying conflicts are to be made visible and a better role understanding achieved using techniques of role building and role reconstruction. (d) The process-oriented approach aims to obtain an increase in performance as well as achievement by restructuring collective work processes (such as problem solving and decision making) by way of well-established methods (e.g., Delphi Technique), to some extent assisted by computer technology (for example, group decision support systems (GDSS)).

Practitioners often combine these approaches. For example, it is common to combine the negotiation of new goals with role negotiation so that a corresponding new understanding of expectations and norms can be established. (For concrete techniques and measures of team development that can be utilized in this process, see West 1997.)

5. Social Identity: In-group–Out-group Processes

Until now we have discussed group processes and their consequences within groups. Of course, in organizations there are also group processes between groups. Here the most important theory which is relevant for these processes is Tajfel's social identity theory (SIT). Interestingly, it is a psychological theory about the social cognition of individuals (not a sociological one that directly addresses entire collectives such as organizations). The predictions made by SIT are among the best proven in social psychology.

Individuals identify with the groups they belong to and derogate others who do not belong to their group (the in-group). Consequently, there is a high degree of identification with the in-group and the respective organization, but at the cost of respect for and cooperation with individuals perceived to belong to other groups (out-groups). Tajfels's theory is very important for understanding the in-/out-group processes underlying conflict and conflict resolution.

Social identity theory emphasizes that all individuals are motivated to shape a positive social identity. The underlying processes are: (a) categorization; that is, a basic cognitive mechanism for structuring the manifold stimuli from the environment in an orderly fashion—in particular, within their social environment, humans categorize on the basis of who belongs to the in-group and who does not (in-group/out-group distinction). (b) Social comparison; that is, a basic sociocognitive mechanism for gaining information about where one stands within the social context—in particular, humans compare their own in-group with other out-groups, preferably in such a way that the dimension on which the comparison is performed produces positive distinctiveness (i.e., a feeling of positive uniqueness). Having the choice of a dimension is important in that when the in-group compares badly (e.g., with a highly performing group), it can be argued that it is more important to have character than to have strong skills and success, thereby protecting the positive distinctiveness of the in-group. Thus, in summary, SIT describes the dynamics between groups on the basis of the individual's motivation to achieve social identity and positive distinctiveness—at the cost of derogating others because they are perceived to belong to an out-group.

Tajfel (1978) shows that even a minimal group context (minimal group paradigm) which may come about through a more or less arbitrary distinction between two groups (e.g., admirers of Vasily Kandinsky vs. admirers of Paul Klee), suffices to produce the in-group favoritism and out-group discrimination processes. People from the out-group are discriminated against more—for example concerning the evaluation of positive attributes, concerning judgments about liking, and concerning the distribution of material rewards (compared to persons from the 'in-group').

These processes are highly relevant for groups within organizations because very often subgroups within organizations think in terms of in-groups and out-groups instead of seeing the overall organization as an in-group. An organization is always in a conflict, on the one hand to increase the in-group feeling because a strong feeling of togetherness is positive, on the other hand the in-group feeling must not be too strong because this could mean shutting down borders so that there are not that many contacts between groups and departments. Therefore the question is what can be done in order to increase both. According to Tajfel's SIT, it is (a) good to have a common goal which can only be achieved by joint action and

cooperation of all groups, (b) important that groups have background information about what the tasks of the other groups are, (c) good to have a rotation of positions between heterogeneous groups, and (d) good to have a lot of informal activities and personal relationships between members of different groups. The important point is that a group must be open and have open borders.

6. Group Processes in Mergers, Acquisitions, and Change Management

When organizations are merged or taken over, or when organizational changes take place, group processes as described in SIT play a central role. In mergers and acquisitions, it is important that the norms and values of the distinct groups can be combined and are not too different from the start. Very often, these groups behave towards each other like an in-group towards an out-group. Very often with mergers or acquisitions, two actual competitors which merge as an in-group or out-group come together and have to continue as one group. Very often, an earlier empty competitor has to be integrated. Many intergroup conflicts can be reduced by taking into account certain psychological findings about the acceptance of change management processes, which are listed below.

(a) There is a cultural fit between the two merging companies and therefore an identification with the new emerging culture is possible (principle of identification and social identity).

(b) The change process is perceived to be definite, inevitable, and irreversible (principle of inevitability).

(c) The vision is clearly perceived and shared, and one understands the reasons why change is necessary (principle of explainability and sense).

(d) Participants have the impression that they can foresee and predict the processes to come (principle of predictability).

(e) There is a consensus about rules and responsibilities (principle of clarity).

(f) Participants feel they are an active part of the decision process (principle of participation).

(g) They feel they are being treated fairly; for example, there is perceived fairness about resources and positions (procedural fairness).

(h) Top managers are perceived as being trustworthy and are positive role models (principle of truthfulness).

Very often, groups and people who are responsible for change processes in organizations, who have to make the relevant decisions and who are responsible for the implementation of those decisions, have a poor understanding of how human beings and groups function in psychological terms. This is one reason why so many mergers, acquisitions, and change processes fail. Very often, the culture and the mentality of groups do not fit with each other; when the above-mentioned aspects are neglected in the change process, it will fail.

It is the inability to judge employees' and groups' anxieties, hopes, and expectations correctly, and to communicate why and how and what changes are necessary (Hogan and Overmyer-Day 1994). Taking the above-mentioned conditions into account will help to make change processes more successful in the future.

7. Future Directions

Group processes in organizations are a typical example of good interdisciplinary cooperation between social psychologists and organizational psychologists. This research shows that nothing is as practical as a good theory. Of course, there are no integrating theories of groups but as we have shown there are different small group theories which can be applied for different heterogeneous phenomena. In the future, it is important to have an integration of these small group theories and in the empirical field it is also necessary to do more longitudinal studies within organizations.

See also: Group Decision Making, Social Psychology of; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Intergroup Relations, Social Psychology of; Problem Solving (Joint), Psychology of; Psychological Climate in the Work Setting; Psychology: Overview; Situated Learning: Out of School and in the Classroom; Social Psychology; Social Psychology, Theories of

Bibliography

Argote L, Epple D 1990 Learning curves in manufacturing. *Science* **247**: 920–3

Asch S E 1951 Effects of group pressure on the modification and distortion of judgements. In: Guetzkow H (ed.) *Groups, Leadership, and Men*. Cemegie, Pittsburgh, PA, pp. 177–90

Brodbeck F C 1996 Criteria for the study of work group functioning. In: West M (ed.) *Handbook of Workgroup Psychology*. Wiley and Sons, Chichester, UK, pp. 285–315

Brodbeck F C, Greitemeyer T 2000a A dynamic model of group performance: Considering the group members' capacity to learn. *Group Processes and Intergroup Relations* **3**: 159–82

Brodbeck F C, Greitemeyer T 2000b Effects of individual versus mixed individual and group experience in rule induction on group member learning and group performance. *Journal of Experimental Social Psychology* **36**: 621–48

Brown R 1993 *Group Processes: Dynamics Within and Between Groups*, 4th edn. Blackwell, Oxford, UK

Frey D, Schulz-Hardt S, Stahlberg D 1996 Information seeking among individuals and groups and possible consequences for decision making in business and politics. In: Witte E, Davis J H (eds.) *Understanding Group Behavior—Small Group Processes and Interpersonal Relations*. Lawrence Erlbaum, Mahwah, NJ, pp. 211–25

Hill G W 1982 Group versus individual performance: Are two heads better than one? *Psychological Bulletin* **91**: 517–39
Hinsz V B, Tindale R S, Vollrath D A 1997 The emerging conceptualization of groups as information processors. *Psychological Bulletin* **121**: 43–64
Hogan E A, Overmyer-Day L 1994 The psychology of mergers and acquisitions. In: Cooper C L, Robertson I T (eds.) *International Review of Industrial and Organizational Psychology*. John Wiley and Sons, Chichester, UK, Vol. 9, pp. 247–81
Janis I L 1982 *Victims of Groupthink, a Psychological Study of Foreign-Policy Decisions and Fiascoes*, 2nd edn. Houghton Mifflin Company, Boston, Atlanta, Dallas
Latané B 1986 Responsibility and effort in organizations. In: Goodmann P S and Associates (eds.) *Designing Effective Work Groups*. Jossey-Bass, San Francisco, pp. 277–304
Levine J M, Moreland R L 1990 Progress in small group research. *Annual Review of Psychology* **41**: 585–634
Liang D W, Moreland R, Argote L 1995 Group versus individual training and group performance: The mediating role of transactive memory. *Personality and Social Psychology Bulletin* **21**: 384–93
Moscovici S 1976 *Social Influence and Social Change*. Academic Press, London
Moscovici S 1980 Towards a theory of conversion behavior. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Bd. 13, S. pp. 208–39
Pinto J K, Prescott J E 1988 Variations in critical success factors over the stages in the project life cycle. *Journal of Management* **14**: 5–18
Sheppard J A 1993 Productivity loss in performance groups: A motivation analysis. *Psychological Bulletin* **113**: 67–81
Steiner I D 1972 *Group Process and Productivity*. Academic Press, New York
Stroebe W, Diehl M, Abakoumkin G 1992 The illusion of group effectivity. *Personality and Social Psychology Bulletin* **18**: 643–50
Tajfel H 1978 *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations*. Academic Press, London
Tuckman B, Jensen M 1977 Stages of small-group development. *Group and Organizational Studies* **2**: 419–27
Wegner D M 1987 Transactive memory: A contemporary analysis of the group mind. In: Mullen B, Goethals G R (eds.) *Theories of Group Behavior*. Springer, New York, pp. 185–208
West M 1994 *Effective Teamwork*. BPS Books, Leicester, UK
West M 1996 *Handbook of Work Group Psychology*. Wiley and Sons, Chichester, UK
West M 1997 *Developing Creativity in Organizations*. BPS Books, Leicester, UK
Williams K, Karau S J 1991 Social loafing and social compensations: The effects of expectations of co-worker performance. *Journal of Personality and Social Psychology* **61**: 570–81

D. Frey and F. Brodbeck

Group Processes in the Classroom

Interest in group learning and group processes has a very long history and remains a key focus of current efforts to improve education. Group processes in the classroom occur when peers engage in a common task and can include a variety of both social and cognitive processes. The focus of this entry is on cooperative or collaborative groups of peers and the various theoretical approaches that attempt to explain group processes in these kinds of group. These approaches propose varying mechanisms for the effectiveness of group learning. Comparisons of empirical findings across studies conducted from different theoretical perspectives are very difficult because of the wide variances in tasks used, duration of studies, roles of teachers and students, outcome measures, and other variables such as group composition. One consequence of the variance in theoretical approaches is the fact that these approaches require different decisions with regard to key variables such as group size or group composition. Group processes can best be understood within particular theoretical approaches. This entry will not address the influences of specific variables but will instead focus on the delineation of various theoretical perspectives and the linkage of research and practice.

1. *Theoretical Approaches to Group Processes*

Different theoretical approaches to peer learning propose different underlying mechanisms or processes that come into play when a group works together (O'Donnell and O'Kelly 1994). Slavin (1996) described two main classes of orientation toward peer learning: a social-behavioral orientation that includes motivational and social cohesion approaches, and a cognitive orientation that includes elaboration and developmental perspectives. There are two distinct developmental theories. Table 1 delineates key differences among these perspectives as they relate to instructional decisions and problems that may arise from such decisions.

1.1 *Social-behavioral Perspectives*

The concept of interdependence among group members is fundamental to the motivational and social cohesion perspectives on cooperative learning, although they differ in how interdependence is created and sustained. When group members are interdependent, the success of one member can occur only if other members are also successful. In competitive situations, one person's success requires that another person fail. In the social motivational approach, interdependence is created by the use of group rewards. In Student Teams Achievement Divisions (Slavin 1995), students work in small groups to master material that the teacher has presented. Each student takes an individual quiz on the material, and points are assigned to the group based on each individual's improvement score. Thus, students are accountable for improving over past performance and learners of varying abilities can contribute equally to the total group score. Group rewards are assigned based on the team's number of points. A key component of this

Table 1
Key considerations in cooperative learning from a variety of theoretical perspectives

	Theoretical perspectives				
	Social-behavioral		Cognitive		
				Developmental	
Considerations	Motivation	Social cohesion	Elaboration	Piagetian	Vygotskian
Goals/incentives	Rewards essential	Rewards nonessential	Rewards nonessential	Rewards nonessential	Rewards nonessential
Group size	Large (4–6)	Large (4–6)	Small (2–4)	Small	Dyads
Group composition	Heterogeneous	Heterogeneous	Heterogeneous	Homogeneous	Heterogeneous
Tasks	Rehearsal	Rehearsal/integrative	Rehearsal/integrative	Exploratory	Skills
Teacher role	Director	Facilitator	Facilitator	Facilitator	Model/guide
Potential problems	Use of reward	Poor social skills	Poor help-giving	Inactive	Poor help-giving
Averting problems	Group size Improvement scores Simpler tasks Social cohesion	Social loafing Cognitive loafing Team building Conflict resolution strategies Discuss group process	Unequal participation Direct instruction in help-giving Model help-giving Scripting interaction	No cognitive conflict Structuring controversy	Providing adequate time/diaglog Direct instruction in help-giving Modeling help-giving

Source: O'Donnell and O'Kelly 1994, p. 327, © Kluwer Academic/Plenum Publishers. Reproduced with permission

approach is the reliance on individual accountability. Evidence of the effectiveness of this approach comes from a large number of studies (Slavin 1996). The kinds of learning task used in this approach, however, tend to focus on basic skills or factual knowledge, and it is unclear whether this approach is effective in promoting higher-order skills.

1.2 Social Cohesion Perspectives

In the social cohesion approach, interdependence derives from a motive of care and concern. Group members help one another because they are concerned about their peers. Teaching social skills that will support this type of interdependence is an important emphasis in social cohesion approaches. The Johnsons's Learning Together technique (Johnson and Johnson 1994) illustrates the social cohesion approach. The empirical evidence supporting this orientation toward group learning as a means of enhancing academic achievement is much weaker than that supporting the motivational perspective. Slavin (1996) notes that cooperative learning methods that emphasize team building and group process, but do not provide group rewards, are not very effective but that when they are combined with the use of rewards or accountability, student achievement from group learning exceeds individualistic learning. Slavin concludes that individual accountability and group rewards are critical to the success of cooperative learning methods.

Not everyone subscribes to the idea that rewards are necessary to create cohesion or interdependence. Cohen (1994) believes that cooperative learning that uses group rewards may be useful for lower-level skills but may not be either necessary or helpful for promoting higher-order skills. Thus, decisions about whether to adopt a motivational or social cohesion approach to peer learning may depend on the desired outcomes from group interaction.

Cohen (1994) emphasizes a different source of social cohesion, suggesting that the interest value of the task in which students engage will create the kind of interdependence necessary for students to succeed. In such situations, difficulties associated with status differences among students are ameliorated. When students have low status in the group, they participate less and in ways that do not promote achievement. Explicit rewards in this context may only exacerbate the problems created by differential status in a group. Cohen and her colleagues have successfully promoted social cohesion in groups by reducing status differences among group members so that they are oriented to accept that all students *can* contribute to task accomplishment. Complex and interesting tasks are necessary for this level of inclusion to occur. Cohen and her colleagues have been very successful in promoting student achievement. Results from Cohen and her colleagues suggest that the nature of the task may have important implications for the kind of accountability that is needed to promote effective group learning.

1.3 Cognitive/Elaborative Perspectives

The third perspective identified by Slavin is a cognitive/elaborative approach and is based on information processing theory. The role of the group interaction from this perspective is to increase active processing by providing opportunities for restructuring and elaborating knowledge. Group learning techniques influenced by this perspective tend to provide some structure to the interaction of students by emphasizing the cognitive processes in which students must engage. Examples of work influenced by this perspective include the use of scripted cooperation by O'Donnell, Dansereau, and colleagues (O'Donnell and Dansereau 1992), reciprocal peer questioning (King 1999), reciprocal teaching (Palinscar and Brown 1984), and the work of Noreen Webb (Webb 1992). The empirical evidence supporting the use of these techniques is strong.

The quality of discourse is critical from this perspective and will have an influence on the outcomes from group interaction. Webb (1992, Webb and Palinscar 1996) has shown that the quality of explanations that are provided in a group are linked to achievement. Giving explanations is more strongly associated with achievement than receiving explanations. King (1999) has also shown that instructional efforts to improve the quality of discourse by teaching students to ask questions that facilitate knowledge construction is very effective.

1.4 Developmental Perspectives

A fourth approach to understanding cooperative groups relies on two different theories of development. From a Piagetian perspective (De Lisi and Golbeck 1999), peers can learn from one another because they can provide opportunities for cognitive conflict and resolution. Groups that experience mutuality of influence and power are most likely to provide a context within which cognitive development can occur because they are more likely to include differences of viewpoint. Children who have not attained the principle of conservation often develop and maintain conservation concepts when they have worked with a child who has already developed this concept (De Lisi and Golbeck 1999). Much of the work on conceptual change in science education has been influenced by this perspective.

From a Vygotskian perspective (Hogan and Tudge 1999), children acquire knowledge and skills that are first modeled in the community and are subsequently internalized by an individual. When cooperative learning is anchored in Vygotskian theory, it is

concerned with the larger community of practice in which individual learning is accomplished. Key processes include the scaffolding of the learner's performance by a more expert individual. The difference between what an individual can accomplish alone and with the assistance of an adult guide or more capable peer is defined as the zone of proximal development.

1.5 Sociocultural and Combined Theoretical Perspectives

A fifth perspective that undergirds such group learning techniques such as Communities of Learners (Brown and Campione 1990) may be identified. It includes the developmental perspective of Vygotsky, with its emphasis on the sociocultural basis of learning, and a social cohesion approach in which such learning is motivated by care and concern. Other forms of cooperative learning borrow from multiple perspectives in making choices about the selection of tasks and the organization of learning (Van Meter and Stevens 2000).

2. Research and Practice

Despite the burgeoning research findings that might assist teachers in making choices about grouping for instructional purposes, teachers do not seem to rely on the empirical research for guidance when making such choices (Antil et al. 1998). However, few studies have examined teachers' beliefs about group processes in the classroom.

2.1 Teachers' Use of Cooperative Learning

The research findings from any of the perspectives previously described seem to have little systematic influence on teachers' choices about groups. Although Slavin (1996) is convinced that empirical evidence supports the use of group rewards, he notes that there is quite a reluctance on the part of classroom teachers to use them.

Antil and his colleagues (1998) conducted a study of elementary school teachers' conceptualization and utilization of cooperative learning. Eighty-five teachers responded to an initial survey, and 93 percent of them reported that they used cooperative learning for both academic and social reasons. In interviews with 21 of these teachers, Antil found that the majority structured tasks to create positive interdependence, and taught students skills for working in groups. Few, however, employed recognized forms of cooperative learning and, in particular, did not link individual accountability to group goals. No data were presented, however, that linked classroom practices in relation to cooperative learning and student achievement. It is clear from this study that teachers do not implement the particular practices of cooperative learning that research has shown to be effective. However, when teachers receive extensive professional development, the effects on student achievement can be powerful (Schultz 2000).

2.2 Usable Models of Group Processes

The availability of multiple perspectives on peer learning and the proliferation of specific techniques may contribute to the disconnection between the available research findings and their use by classroom teachers. The kind of classification system represented in Table 1 is rarely presented to either pre-service or in-service teachers. Usable models of peer learning that can guide practical decisions are needed. Such models need clearly to link expected outcomes to classroom processes and teachers' actions or decisions.

Slavin (1996) proposed a model of effective peer learning with motivation at its core. Motivation is necessary in order for students to engage in the active processing (e.g., explanations) that is linked to achievement. The keys to motivating students are group rewards and individual accountability. Teachers can act to create and sustain children's motivation that will allow them to link will and skill in pursuit of achievement.

Van Meter and Stevens (2000) emphasize the quality of discourse in a group as the key variable in effective group learning. Other instructional choices should be such that the quality of discourse is enhanced. Thus, task instructions and group composition are only interesting insofar as they influence the quality of the discourse among peers.

These two models provide a simple set of goals for teachers using group learning. Coherence among classroom decisions about fundamental aspects of arranging instruction (e.g., group size, group composition, type of task) is more likely to occur if there is a clear goal in place (e.g., motivate students to engage in active learning). Teachers in the Antil et al. (1998) study reported that they did not subscribe to particular cooperative techniques because they wanted to maintain their decision-making autonomy. The two models that are briefly described here with their emphases on motivation and discourse provide a set of superordinate goals that can accommodate teachers' needs for autonomy, while preserving a principled approached to decision making about group processes.

3. Future Directions

Given that teachers seem reluctant to use group rewards (Slavin 1996), it is important to specify the conditions under which alternative methods of creating interdependence among group members can be effective. What are the kinds of tasks for which extrinsic reward are not necessary? Answers to this question lie in the analysis of the kinds of process in which students engage when they are engaged in

interesting and challenging tasks. A clearer understanding of effective processing of complex tasks may make it possible to find alternatives to rewards.

The tension between theories of group processes and classroom practices of group learning presents the richest source of future research questions. Teachers' beliefs about learning lie at the fulcrum of their abilities to implement effective group learning in the classroom. The relationships between teacher beliefs, implementation practices, and student achievement remain untouched as research questions. Research on this set of relationships would provide valuable insights that could inform teacher preparation and in-service professional development, and also improve classroom practice of group learning.

See also: Cooperation and Competition, Psychology of; Cooperative Learning in Schools; Group Processes in Organizations; Group Processes, Social Psychology of; Groups, Sociology of; Hidden Curriculum in the Classroom; Open Classroom: Learning and Teaching

Bibliography

Antil L R, Jenkins J R, Wayne S K, Vadasy P F 1998 Cooperative learning: Prevalence, conceptualizations, and the relationships between research and practice. *American Educational Research Journal* **3**: 419–55

Brown A L, Campione J C 1990 Communities of learning and thinking, or a context by any other name. *Developmental Perspectives on Teaching and Learning Thinking Skills* **21**: 108–26

Cohen E G 1994 Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research* **64**: 1–35

De Lisi R, Golbeck S 1999 Implications of Piagetian theory for peer learning. In: O'Donnell A M, King A (eds.) *Cognitive Perspectives on Peer Learning*. Lawrence Erlbaum, Mahwah, NJ, pp. 3–37

Hogan D, Tudge J H R 1999 Implications for Vygotsky's theory for peer learning. In: O'Donnell A M, King A (eds.) *Cognitive Perspectives on Peer Learning*. Lawrence Erlbaum, Mahwah, NJ, pp. 39–65

Johnson D W, Johnson R 1994 *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning*, 4th edn, Allyn and Bacon, Boston

King A 1999 Discourse patterns for mediating peer learning. In: O'Donnell A M, King A (eds.) *Cognitive Perspectives on Peer Learning*. Lawrence Erlbaum, Mahwah, NJ, pp. 87–115

O'Donnell A M, Dansereau D F 1992 Scripted cooperation in student dyads: A method for analyzing and enhancing academic learning and performance. In: Lazarowitz R, Miller N (eds.) *Interaction in Cooperative Groups: The Theoretical Anatomy of Group Learning*. Cambridge University Press, New York, pp. 120–41

O'Donnell A M, O'Kelly J 1994 Learning from peers: Beyond the rhetoric of positive results. *Educational Psychology Review* **6**: 321–49

Palinscar A S, Brown A 1984 Reciprocal teaching of comprehension monitoring activities. *Cognition and Instruction* **2**: 117–75

Schultz S E 2000 *Are Four Heads Better than One? Effects of Group Work on Students' Achievement in Science*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA

Slavin R E 1995 *Cooperative Learning: Theory, Research, and Practice*, 2nd edn. Allyn and Bacon, Boston

Slavin R E 1996 Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology* **21**: 43–69

Van Meter P, Stevens R J 2000 The role of theory in the study of peer collaboration. *Journal of Experimental Education* **69**: 113–27

Webb N M 1992 Testing a theoretical model of student interaction and learning in small groups. In: Lazarowitz R, Miller N (eds.) *Interaction in Cooperative Groups: The Theoretical Anatomy of Group Learning*. Cambridge University Press, New York, pp. 102–19

Webb N M, Palinscar A S 1996 Group processes in the classroom. In: Berliner D, Calfee R C (eds.) *Handbook of Educational Psychology*. Macmillan, New York

A. M. O'Donnell

Group Processes, Social Psychology of

Group processes encompass such a wide range of phenomena that the social psychological study of group processes has many different focuses. It can focus on individual cognitive and motivational processes that produce group behaviors, interpersonal processes among more than two people, cognitive and social processes that cause people to conceive of themselves as group members and behave accordingly, intergroup relations that affect intragroup processes, the interrelationship of individual, interpersonal, and social processes, the behavior of specific groups or types of group, and so forth (e.g., Baron et al. 1992, Brown 2000, Hogg and Tindale 2001, Levine and Moreland 1998, Moreland and Levine 1994). What unites this diversity is a focus on the group as a social psychological entity.

1. What is a Group?

There is, however, little agreement on a social psychological definition of the group (see Hogg and Abrams 1988). The dominant view (since Floyd Allport in the early 1920s) is that a group is a collection of individuals, and that group processes are actually individual or interpersonal processes among a number of people. For instance, Latané's social impact theory attributes differences between interpersonal and group behavior to, among other things, the effects of increasing numbers of people. An alternative perspective (see Hogg and Abrams 1988) is that groups and group behavior are qualitatively distinct from individuals and interpersonal behavior, and that different or

additional concepts are needed to properly analyze groups—for example, groups produce norms that regulate members' behavior and people reconfigure their self-concept so that it is defined in terms of properties of the group as a collective. This perspective is represented by the early work of Wundt and Durkheim (late nineteenth century), and McDougall and Mead (early twentieth century), the more recent work of Sherif and Tajfel, and presently by social identity theory (see *Social Identity, Psychology of*) and intergroup approaches (see *Intergroup Relations, Social Psychology of*). It is argued that the analysis of groups in terms only of interpersonal processes is reductionist—it does not adequately address questions to do with group processes, but instead deals only with interpersonal processes among more than two people.

The issue of 'what is a group' continues to simmer as an important, and unresolved, metatheoretical debate in social psychology. Many researchers adhere to one view or the other, for example, perspectives that focus on the interdependence of individuals in small face-to-face aggregates vs. perspectives that focus on the collective self-concept that derives from membership of groups of all sizes. Others opt out of the formal controversy—this is particularly true of an increasing number of scholars who research group processes from more applied perspectives (e.g., sport psychology). In recent years, however, there have been promising attempts to integrate interactive small group perspectives, intergroup perspectives, and cognitive self-concept perspectives (see Brown 2000).

2. *Studying Group Processes*

Traditionally the study of group processes has lain at the heart of social psychology. It was in many ways the explosion in the late 1940s and early 1950s of research into what was then called group dynamics that set experimental social psychology off on its current expansion. Even at the beginning of the twenty-first century, early group dynamics researchers like Kurt Lewin and Leon Festinger have significant impact on the discipline. Group processes research thrived until about the mid-1960s when it suffered a fairly rapid decline in popularity (replaced by attribution theories, then social cognition), from which it is showing strong signs of recovery, mainly through cognitive and intergroup perspectives on group processes (Moreland et al. 1994). The apparent demise of traditional group processes research may be due to a variety of factors (Steiner 1974). For instance: (a) it is time consuming and costly to conduct research with interactive groups; (b) data from group research are rarely neat, tidy, and easily explained; (c) the computer revolution shifted social psychological interest from the study of groups to the development of theories of individual human behavior that resembled computer software; and (d) the logical end-point of reductionist perspectives on group processes is to focus exclusively on individual psychology.

3. *Group Formation, Commitment, and Identity*

A central concept in the study of group processes is 'group cohesiveness,' which refers to the overall solidarity of groups, their 'groupness,' and the process that transforms an aggregate of unrelated individuals into a psychological group. Through cooperative interdependence in the pursuit of shared goals, people become attracted to one another, to the group, and to the group's goals. In this way a group forms, and its overall cohesiveness is an arithmetic combination of individual members' attraction to one another and the group. This attraction-to-group—or interpersonal attraction—analysis has not gone unchallenged, and recent perspectives place an emphasis on some form of commitment to group, or self-definition in terms of group membership (Hogg 1992). For instance, social identity theory does not consider attraction-to-group to be the basic process of group formation, but rather one of a range of consequences of a process of identification with the group (i.e., self-categorization in terms of the group's defining features—see *Social Identity, Psychology of*). Moreland and Levine (e.g. 1994) have developed a 'group socialization model' to address what they consider to be a lack of diachronic perspective in many models of group formation and development. Commitment is considered the most basic process. People become committed to a group, and a group to its members, through bilateral cost-benefit analyses of membership in this group relative to other groups. As mutual commitment rises or ebbs, members move in and out of different general roles within the group (e.g., new member, 'oldtimer')—a process characterized by role transitions that are quite marked, and which can involve public rites of passage.

4. *Group Influence*

Groups influence the attitudes and behaviors of their members in many different ways (e.g., Turner 1991—see *Social Influence, Psychology of*; *Attitude Change: Psychological*).

4.1 *Direct Influence and Persuasion*

Direct influence attempts involve individuals, subgroups, or the group as a whole trying to persuade members (e.g., deviants, dissenters) to go along with or come into line with the group. Persuasion attempts may range from relatively polite requests all the way to direct orders (see *Obedience: Social Psychological*

Perspectives) backed up by threats and ultimately force. Persuasion is generally more effective if the source of influence has a degree of legitimate authority and power within the group (see *Leadership, Psychology of*; Chemers 2001). Direct influence generally produces behavioral compliance which is not necessarily matched by underlying cognitive or attitudinal change—people go along with the group for instrumental reasons which may include a desire to be evaluated favorably or to avoid disapproval or punishment.

4.2 Indirect Influence and Group Norms

Groups can also influence members indirectly through the power of social norms that describe and prescribe appropriate behavior for group members in that context (Brown 2000, Turner 1991). In the absence of direct social pressure people conform to group norms. This may happen through 'normative influence,' in which people simply publicly comply in order to obtain social approval or avoid disapproval from others—there is no underlying cognitive change. Another process is 'informational influence' which produces true conformity—underlying cognitive change which may go along with behavioral change (see *Attitudes and Behavior*). Informational influence is effective where people are uncertain about the objective correctness of their beliefs, and in the absence of objective criteria they use group members' behaviors as social criteria (Festinger 1954; see *Youth to Adulthood across Cultures, Transition from*). A third process is 'referent informational influence' (see *Social Identity, Psychology of*) in which people who identify with a group construct a contextually appropriate norm that defines the ingroup in contrast to relevant outgroups, and then enact the norm as an automatic consequence of self-categorization.

Although indirect group influence usually comes from the ingroup, there are at least two ways in which an outgroup can have influence: (a) outgroups influence the nature of the ingroup norm that people conform to, and (b) minority outgroups can adopt strategies to convert people to their viewpoint (e.g., by promulgating a consistent and consensual viewpoint, and by presenting themselves as actually part of the ingroup)—see *Social Movements: Psychological Perspectives*.

5. Group Structure

Whereas norms and social influence processes tend to produce uniformity within groups, other factors introduce unevenness and variability across the group that structures the group into different regions. Aspects of group structure can surface through intragroup patterns of who prefers or likes whom, known as 'sociometric choice' (see Hogg 1992).

5.1 Roles

Groups usually embrace a number of different roles that can be occupied by different people, or by subgroups of people. Roles prescribe behavior, and they emerge to assist the group to function better as an integrated entity. According to Moreland and Levine's (1994) group socialization model, the group socialization process involves people moving in and out of quite distinct general roles within the group as a function of commitment to the group. Roles are rarely equal—some are more desirable than others, and have more power, prestige, and status attached to them. In this way, groups are often internally structured into role hierarchies. High status roles tend to be occupied by people who have specific status characteristics (abilities and skills directly related to the group's function) and/or diffuse status characteristics (they are members of social categories that have high status in society at large)—see Ridgeway 2001.

5.2 Communication Structures

Another feature of group structure is the ease with which members or role occupants can communicate with one another (see *Communication and Social Psychology*; see Baron et al. 1992). Depending on the group's task, different communication structures can emerge—highly centralized structures channel all communication through a single communication hub (often the leader), while less centralized networks allow freer communication among all members. The nature of the communication network affects group functioning, task performance, member satisfaction, and group solidarity. For instance, highly centralized networks can, under certain circumstances, reduce efficiency, cause peripheral members to feel excluded and dissatisfied, and even reduce group cohesion. Communication networks can be traced and described using Bales's 'interaction process analysis' method, which documents who communicates what with whom, and how often.

5.3 Leadership

One of the most important, and certainly the highest status, role in a group is the role of leader (Chemers 2001, see *Leadership, Psychology of*). Research on leadership suggests that effective leadership (often defined in terms of the ability to motivate members to internalize and act on one's own innovative vision for the group) is an interaction between leadership style and the type of leader that is needed in a particular situation. Another factor that may be important in some circumstances is that the leader is perceived to be a good group member who closely fits the defining features of membership. Because leadership is a role, leadership dynamics involve followers—if followers will not be led then the leader cannot lead.

Generally, followers tend to endorse leaders who have good leadership qualities, who have abilities that benefit and promote the group as a whole, and who are good group members or have been good group members in the past. These attributes encourage followers to endow the leader with status and to allow the leader to be innovative.

5.4 Deviance

Whereas leaders are at the pinnacle of the intragroup role hierarchy, deviants are at the bottom (see Hogg and Tindale 2001). Research suggests that groups try very hard to socialize marginal members by making them conform or by finding them useful and valued roles within the group. If these attempts fail, people evaluatively reject deviants, often pathologizing their behavior and attempting to classify them as outgroup members. Research on the 'black sheep effect' suggests that because marginal members pose a threat to the normative integrity of the group and thus threaten members' group membership-based self-concept, marginal ingroup members are actually disliked even more than an identical person who belongs to an outgroup.

6. Group Performance

Since groups often exist to perform specific tasks, an important question to ask is whether groups perform better or worse than individuals (see Baron et al. 1992). This is an extension of an even more fundamental question about how individual task performance is affected by the presence of others. Coactors or a passive audience tend to improve performance of a well-learned or easy task, but impede performance of a poorly learned or difficult task (see *Social Facilitation, Psychology of*). In general, however, groups are interactive entities, and the question becomes one of whether the task is better accomplished by interaction in a group, or simply by individuals working alone. The answer to this question is that it depends on the specific nature of the task—for instance, more cars can be produced by a group effort based on a coordinated division of labour, than by individuals building them from scratch (see *Group Productivity, Social Psychology of*).

6.1 Social Loafing, Free Riding, and Social Dilemmas

Groups often perform less well than would be suggested by the sum of the abilities or performances of individual members (see Moreland and Levine 1994). This can be because of coordination losses, that is, interference between individuals trying to coordinate their actions within the group. Coordination losses may be one explanation of why group brainstorming is often a suboptimal way to generate novel ideas in a group. Groups can also perform less well because of motivation losses, that is, people in groups put in less effort because they are less motivated. This latter 'social loafing' effect seems to be partly due to a reduction in perceived personal identifiability and responsibility in the group (see *Deindividuation, Psychology of*), and can be explained in terms of 'social impact theory.' Social loafing can be reduced by, among other things, task importance, enjoyment, and relevance, and increased group cohesiveness and personal identifiability. Related to loafing is a tendency for individuals in a group confronted by a scarce resource to free ride, that is, maximize their own short-term gains with little regard for the group as a whole and thus their long-term gains (see *Social Dilemmas, Psychology of*). Many of the environmental problems faced by the world today are examples of social dilemmas. Social dilemmas can be (partly) avoided by having a leader manage a resource rather than letting the members compete directly, or by having members identify strongly with the group that accesses the resource.

7. Deindividuation, Collective Protest, and the Crowd

A common assumption about groups is that they cause people to lose their sense of individuality and that this leads to low motivation, lack of responsibility, and antisocial conduct—all captured by the notion of 'deindividuation' (Zimbardo 1970, see *Deindividuation, Psychology of*). Historically, this perspective, which has its critics (see Postmes and Spears 1998), is most evident in research on a particular type of group—the crowd. Starting with LeBon in the late nineteenth century, social psychologists have generally considered crowds to be irrational and antisocial because members of the crowd are deindividuated. A contrasting perspective is one that views crowd behavior as social protest (see *Social Movements: Psychological Perspectives*). Crowds have a purpose that is tied to a specific identity, and so crowd behavior has a logic to it because it is tightly regulated by the context relevant norms of the group. Where a crowd engages in antisocial acts, this is because the event is an expression of protest against a dominant group—the protest takes the form of violation of the dominant group's norms and rules.

8. Group Decision Making

One of the most common tasks that groups perform is decision making (Baron et al. 1992, see *Group Decision Making, Social Psychology of*). Group decision making is potentially superior to individual decision

making because in a group there is a larger pool of expertise, viewpoints, and so forth. Groups may come to decisions slowly, but this may be a good thing, particularly for complex issues, because it indicates proper consideration of all facets of the issue. However, groups are also subject to an array of group processes that may impair decision making, for example, factional conflict, biased leadership, and pressures to be obedient and compliant.

8.1 Groupthink and Group Polarization

One potential problem is the tendency for highly cohesive decision-making groups, that are under stress and do not have impartial leadership, to engage in 'groupthink,' a mode of thinking in which the desire to reach unanimous agreement overrides the motivation to adopt proper, rational group decision-making procedures. Groupthink involves defective decision-making procedures, which generally produce poor decisions that can have disastrous consequences. Many poor-quality crisis decisions made by governments have been attributed to groupthink, for example, the 1961 Bay of Pigs fiasco. Another potential bias in group decision making is 'group polarization' (see Turner 1991). This is a tendency for group discussion to cause groups to come to a decision that is more extreme than the mean of the individual members' prediscussion opinions. Polarization occurs only when the prediscussion mean already tends to favor one pole of the decision-making dimension—the group decision is then more extreme in this direction.

8.2 Social Combination Models

A major focus of research on group decision making concerns how one can predict a group decision from the initial distribution of members' views (Baron et al. 1992, Hogg and Tindale 2001). A number of social combination models have been derived to relate the distribution of initial opinions in a decision-making group to the final group decision. The most enduring of these is Davis's social decision schemes model (Davis 1973, which identifies at least five implicit (or explicit) decision rules that can exist in groups (e.g., unanimity, majority wins, truth wins). Different rules tend to prevail for tasks in which there is a demonstrably correct solution ('intellective tasks') as opposed to tasks which are more a matter of taste and preference ('judgemental tasks'). Rules also differ in terms of how strict they are (i.e., how much intragroup agreement is called for), and how much power concentration they have (i.e., how much power is located in one individual). Unanimity is very strict but low on power concentration, while two-thirds majority is less strict but has greater power concentration. Strictness and power concentration have effects on satisfaction within the group, and on the speed with which a group makes a decision. Social combination models are often tested using a simulated jury paradigm.

8.3 Group Remembering

Group decision making usually involves collaborative recall of information, or group remembering (Baron et al. 1992, Hogg and Tindale 2001). Group remembering is not simply a retrieval activity in the pursuit of veridical recall, rather it is a constructive process where social comparisons among group members shape the resultant group memory. As such, group remembering may be subject to implicit or even intentional bias. Another feature of group remembering is 'transactive memory.' Groups often need to bring a large amount of information to bear upon a decision, and thus they develop a differentiated memory structure where individuals or subgroups specialize in different memory domains. Where this occurs people not only have to remember their own part of the puzzle, but they also need to know who specializes in what memories—they develop a transactive memory structure. Research shows that group functioning and decision making is compromised where people have not acquired a transactive memory, or where such a system has not developed in the group.

8.4 Intergroup Decision Making

Decision-making groups can contain factional rifts, or can be considered to be interactive decision-making contexts in which opposing groups or their representatives come together to reach an agreement. Under these circumstances the group provides a context for intergroup decision making (see *Intergroup Relations, Social Psychology of*). Because individuals are acting as group representatives rather than individuals, face-to-face intergroup bargaining has a relatively low probability of success—intransigence and stalemate is more likely. Third party mediation can help, and as a last resort, so can arbitration (in which a settlement is imposed by a third party).

9. The Intergroup Dimension

The study of group processes has tended to focus on what happens among members of small interactive groups (e.g., Levine and Moreland 1998, Moreland and Levine 1994), and does not fully consider how relations between groups may influence these processes. There are, however, indications that an intergroup perspective may be useful in helping to understand intragroup processes (Brown 2000, Hogg and Abrams 1988, see *Social Identity, Psychology of*; *Intergroup Relations, Social Psychology of*). For inst-

ance, group polarization may be an intergroup phenomenon in which a relatively extreme ingroup differentiates itself from a less extreme outgroup by becoming even more extreme. Group decision making may sometimes actually be intergroup decision making, and the commitment process that underlies group socialization involves an assessment not only of ingroup membership but also outgroup memberships. There is also an array of group behaviors that are more properly termed intergroup behaviors: these may include a more favorable evaluation of the ingroup, its members, and all it stands for, in relation to outgroups (i.e., ethnocentrism); discriminatory behavior against outgroups and in favor of the ingroup; and stereotypic perceptions and assumptions about ingroup and outgroup members (see *Stereotypes*). Also, the extreme forms of collective behavior that occur in crowds are sometimes best explained in intergroup terms—the crowd is a group with a passion that is guided and fuelled by the group's opposition to an outgroup (e.g., football supporter groups). Finally, group influence is not purely one-way, with the group bringing dissenters and deviates into line. It is often an intergroup affair with active minorities within the group trying to sway the majority to its own position.

10. Specific Groups

The study of group processes has tended to have numerically small, face-to-face, demographically homogenous, short-lived, task-oriented groups in mind (Steiner 1974). As such, a great deal of research has used four to six-person laboratory groups of university (often psychology) students. In many cases, experimental studies of group processes have used the dyad as the prototypical group. This is based on an assumption that group processes are merely aggregated interpersonal processes. More naturalistic studies have dealt with a great variety of groups—gangs, housing programs, sports teams, summer camps, scout troops, school classes, military groups, therapy groups, juries, government and organizational decision-making groups, and so forth. Studies in the intergroup tradition have tended to be interested in cognitive processes and so have often, though not exclusively, used rather abstract laboratory experiments where groups are often only nominal and there is no interaction.

11. Future Directions

There has been a marked revival of interest among social psychologists in group processes (Moreland et al. 1994). The key feature of this revival is an integration of intergroup, small group, and social cognitive perspectives in social psychology (see Brown 2000, Hogg and Tindale 2001) that sets a clear course for future developments. Recent developments have, however, tended to focus on generic group processes rather than the phenomenology of specific types of group. Although the study of larger, naturally occurring groups in more naturalistic contexts is time consuming, methodologically problematic, and involves weaker experimental control, future developments in group processes may be assisted by more of this type of research. Indeed there is evidence that although certain group processes may be common to all groups, a full understanding of groups also needs to recognize differences between small interactive groups and large social categories, between groups that vary in the degree to which they are a coherent entity, and between groups as diverse as a family, a sect, a therapy group, an army, a nation, a virtual group, and so forth (Deaux et al. 1995).

See also: Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Processes in Organizations; Group Processes in the Classroom; Group Productivity, Social Psychology of; Stereotypes, Social Psychology of

Bibliography

Baron R S, Kerr N L, Miller N 1992 *Group Process, Group Decision, Group Action*. Brooks/Cole, Pacific Grove, CA

Brown R 2000 *Group Processes: Dynamics Within and Between Groups,* 2nd edn. Blackwell, Oxford

Chemers M M 2001 Leadership effectiveness: An integrative review. In: Hogg M A, Tindale S T (eds.) *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford, UK, pp. 376–99

Davis J H 1973 Group decision and social interaction: A theory of social decision schemes. *Psychological Review* **80**: 97–125

Deaux K, Reid A, Mizrahi K, Ethier K A 1995 Parameters of social identity. *Journal of Personality and Social Psychology* **68**: 280–91

Festinger L 1954 A theory of social comparison processes. *Human Relations* **7**: 117–40

Hogg M A 1992 *The Social Psychology of Group Cohesiveness: From Attraction to Social Identity*. Harvester Wheatsheaf, London

Hogg M A, Abrams D 1988 *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*. Routledge, London

Hogg M A, Tindale R S (eds.) 2001 *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford

Levine J M, Moreland R M 1998 Small groups. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*. McGraw-Hill, New York, Vol. 2, pp. 415–69

Moreland R L, Hogg M A, Hains S C 1994 Back to the future: Social psychological research on groups. *Journal of Experimental Social Psychology* **30**: 527–55

Moreland R L, Levine J M 1994 *Understanding Small Groups*. Allyn and Bacon, Boston

Postmes T, Spears R 1998 Deindividuation and anti-normative behavior: A meta-analysis. *Psychological Bulletin* **123**: 238–59

Ridgeway C L 2001 Social status and group structure. In: Hogg M A, Tindale S T (eds.) *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford, pp. 352–75
Steiner I D 1974 Whatever happened to the group in social psychology? *Journal of Experimental Social Psychology* **10**: 94–108
Turner J C 1991 *Social Influence*. Brooks/Cole, Pacific Grove, CA
Zimbardo P G 1970 The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In: Arnold W J, Levine D (eds.) *Nebraska Symposium on Motivation* 1969. University of Nebraska Press, Lincoln, NE, Vol. 17, pp. 237–307

M. A. Hogg

Group Productivity, Social Psychology of

Group productivity refers to the degree to which working in a group context either enhances or reduces individual motivation and performance. Group productivity is one of social psychology's cornerstones. The earliest documented studies in social psychology examined the effects of groups on individuals, and hundreds of laboratory and field studies have been conducted on this issue.

1. Historical Background

The earliest studies revealed two seemingly opposing consequences of group work. In the 1880s, Max Ringelmann observed that oxen yoked together did not pull plows as fast as would be expected from their individual efforts. In his laboratory, he asked male volunteers to pull on a rope as hard as they could, both alone and as members of groups ranging in size from two to six. He found that as group size increased, group performance was increasingly lower than would be expected from the simple addition of individual performances, suggesting that working in a group could reduce performance (see Kravitz and Martin 1986). In 1898, Norman Triplett noted that competitive cyclists were faster when racing against other cyclists than when racing against the clock. In his laboratory, he demonstrated that children worked harder on a physical task when they were in the presence of other children who were working on the same task compared to when they worked alone, suggesting that working in a group could enhance performance.

Triplett's work helped stimulate a flurry of studies on what came to be known as 'social facilitation,' revolving around the idea that the presence of others could facilitate individual performance. Meanwhile, Ringelmann's findings were essentially ignored, regarded with skepticism, or interpreted solely in terms of poor coordination of group members' activities. It was not until 1974 that Ringelmann's findings were replicated, and not until even later that the motivational component of these findings was examined as an important and reliable phenomenon that came to be known as 'social loafing.'

2. Social Facilitation

Though Triplett's findings were replicated early and often, inconsistencies began to emerge across studies throughout the first half of the twentieth century. Specifically, the presence of others, either as an audience or as coactors (i.e., individuals who work on the same task in each other's presence but who do not combine their outputs into a group product), did not always enhance performance and sometimes reduced performance. Zajonc (1965) devised an elegant solution to this discrepancy. He suggested that the presence of other people increases generalized drive, thereby facilitating dominant responses. On simple or well-learned tasks, dominant responses are likely correct and the presence of others should enhance performance. On complex or unfamiliar tasks, dominant responses are likely to be in error and the presence of others should reduce performance.

Later research has largely supported Zajonc's view (for a review, see Bond and Titus 1983). Indeed, most researchers now define social facilitation as the tendency for the presence of others to facilitate performance on simple or well-learned tasks but to impair performance on complex or unfamiliar tasks. However, there is an active controversy among researchers and theorists as to why and when the presence of others facilitates dominant responses. The mere presence perspective of Zajonc (1965) and others focuses on arousal and assumes that the presence of other people introduces an element of uncertainty to a situation that leads people to become more aroused and alert, regardless of whether or not the other people present have any interest in what the individual is doing. Cottrell's (1972) evaluation apprehension theory states that mere presence is not sufficient to produce arousal, but rather that it is the ability of other people to create apprehension about being evaluated that leads to social facilitation. If others are present but are paying no attention to the individual and have no regard for how well the individual performs, performance should not be affected. Finally, distraction-conflict theory (Baron 1986) focuses on attention rather than arousal as the primary mediator of social facilitation. Research has documented repeatedly that the presence of others can be distracting and can divide an individual's attention between working on the task at hand and monitoring the behavior and reactions of the audience or coactors. Distraction-conflict theory suggests that individuals try to resolve this cognitive and attentional conflict by increasing their effort at the task to overcome the effects of distraction, thereby enhancing performance on simple tasks and reducing it on complex tasks.

Research has found some support for all of these theories. It appears as though mere presence, evaluation apprehension, and distraction are all capable of creating and contributing to social facilitation effects. Arousal levels have been documented to increase in the mere presence of others, and support has been found for social facilitation even when the others present are very unlikely to create any apprehension about evaluation. Arousal does become more pronounced, however, when observers are clearly evaluating the individual. Similarly, studies have shown that social facilitation effects can emerge from situations in which the presence of others creates distraction. Thus, arousal, evaluation apprehension, and distraction can all potentially contribute to social facilitation effects, both alone and in combination with one another (for a detailed discussion of these factors, see Geen 1991).

Research has also documented social facilitation effects across a wide array of settings, tasks, and subject populations. The effect also appears to hold for a wide variety of animals in addition to humans, including armadillos, chickens, monkeys, and rats. In one rather unusual study, the ability of cockroaches to run either a simple runway or a maze was examined either alone, in coacting pairs, or in the presence of other cockroaches who observed from behind a plastic barrier (Zajonc et al. 1969). When in the presence of coactors or observers, cockroaches ran the simple runway more rapidly but took longer to navigate the maze, replicating prior social facilitation research on humans and other animals. Social facilitation has also been widely documented in field settings. As just two examples, social facilitation effects have been documented for gymnastics performance and billiards playing in public settings.

However, although social facilitation effects are well-established, they appear to be somewhat small in magnitude in comparison with other prominent social psychological phenomena and the impact of facilitation on performance sometimes varies across performance measures. Based on their meta-analysis of 241 studies, Bond and Titus (1983) concluded that, on simple tasks, the presence of others increased the rate of productivity but did not improve performance quality. They also concluded that, on complex tasks, the presence of others reduced both the quantity and quality of performance. Finally, they noted that the productivity gains associated with simple tasks were weaker in magnitude than the productivity losses associated with complex tasks performed in the presence of others.

3. *Social Loafing*

Formally, social loafing is a reduction in motivation and effort when individuals work collectively compared with when they work individually or coactively. In order to study the effects of group context on individual motivation while controlling for mere presence effects, social loafing researchers usually compare individual performance on collective tasks to individual performance on coactive tasks, while holding group size constant across these conditions. In addition, in order to separate motivation loss from mere coordination loss (i.e., poor coordination of group member's actions), social loafing researchers often employ pseudogroups in which individuals work alone or produce outputs that can be individually monitored by the researcher, but believe that their outputs are being combined with those of others into a group product.

These methodological features were prominent in a highly influential experiment by Latané et al. (1979). Participants shouted and clapped as loudly as they could both individually and with others. While blindfolded and wearing headphones to mask the noise, participants shouted both in actual groups, and in pseudogroups in which they shouted alone but believed they were shouting with others. Participants still reduced their efforts on the pseudogroup trials, even though they only thought they were working with others. Thus, Latané and colleagues (1979) demonstrated that a substantial proportion of the decreased performance of groups was attributable to reduced individual effort, as distinct from coordination loss. They also coined the term 'social loafing' for the demotivating effects of working collectively.

Since the late 1970s, more than 90 studies on social loafing have been conducted. Both laboratory experiments and field studies have been conducted, using a wide variety of tasks and examining a range of subject populations. The bulk of the available evidence suggests that social loafing is moderate in magnitude and consistently found across studies. A meta-analysis of 78 studies by Karau and Williams (1993) found that the magnitude of social loafing was comparable in size to that of a number of prominent social psychological effects. Social loafing has been found for a wide variety of tasks, including evaluative tasks such as rating the quality of editorials and poems, cognitive tasks such as navigating computer mazes and detecting visual signals, creative tasks such as song-writing and thought-listing, physical tasks such as pulling on a rope and swimming, and work-related tasks such as typing. Furthermore, although national differences are occasionally found in some studies, a number of studies have documented social loafing regardless of participants' gender, nationality, or age. Thus, social loafing is robust and generalizes across tasks, as well as across most subject populations (Karau and Williams 1993).

However, social loafing is not inevitable and a number of factors have been found to moderate the effect. Thus, studies have found that social loafing can be reduced or eliminated by making individual inputs identifiable, enhancing task meaningfulness or personal involvement with the task, providing individual

or group comparison standards, providing incentives for good performance, threatening punishment for poor performance, increasing personal perceptions of efficacy at the task, and increasing the uniqueness of individual contributions (Karau and Williams 1993). Finally, gender and culture both appear to moderate social loafing. In their meta-analysis, Karau and Williams (1993) found that the magnitude of social loafing across studies was lower for women and for people in Eastern cultures such as Japan, Taiwan, and China, than it was for men and for people in Western cultures such as the USA and Canada, although significant effort reductions were still found within each of these groups.

A number of theories have been offered to explain social loafing, of which four perspectives are most prominent in the literature. First, social impact theory (Latané 1981) views people as either sources or targets of social influence and states that the amount of social impact experienced in a situation is a function of the strength (status or power), immediacy (physical or psychological closeness), and number of sources and targets present. On collective tasks, the influence of the outside influence source (the manager or the experimenter) is diffused across all of the group members, making it less potent on collective tasks than on individual tasks.

Second, the arousal reduction viewpoint (Jackson and Williams 1985) builds on social impact theory to suggests that the presence of others actually reduces arousal when those others are co-workers rather than coactors. According to this view, social loafing is the flip side of social facilitation because the presence of other co-workers lead to reduced drive, and therefore to reduced performance on simple tasks but enhanced performance on complex tasks.

Third, the evaluation potential perspective (Harkins 1987) suggests that social loafing arises because individuals can often be evaluated when working coactively, but are harder to evaluate when working collectively. Thus, individuals can 'hide in the crowd' on groups tasks and avoid taking the blame for a poor group performance. Similarly, they may also feel 'lost in the crowd' and unable to take credit for a good group performance.

Fourth, expectancy-value perspectives (Karau and Williams 1993, Kerr 1983, Sheppard 1993) suggest that individuals work less hard on a collective task because they feel their efforts will not be instrumental in obtaining an outcome that they personally value. As delineated in detail in a model by Karau and Williams (1993), collective tasks pose unique barriers to motivation both because they create more obstacles to the perceptions of efficacy and instrumentality, and because they may undermine the value attached to the task. All of these theories appear to have some usefulness, with the expectancy-value approach representing a more comprehensive or integrative approach to understanding individual motivation in groups, and the other viewpoints explaining specific mediators of social loafing quite well, albeit in a more restricted fashion.

4. *The Relationship Between Social Facilitation and Social Loafing*

Both social loafing and social facilitation are concerned with the effects of group contexts on motivation and performance and yet social loafing suggests that this presence decreases performance whereas social facilitation suggests that this presence increases performance for some tasks but decreases performance for other tasks. Indeed, the very first empirical studies of group productivity by Ringelmann and Triplett documented these seemingly contradictory effects of groups.

The key to resolving this discrepancy appears to be in specifying whether the other people present are co-workers or coactors. In social loafing studies, groups are composed of co-workers who pool their efforts together. In social facilitation studies, however, groups are composed of coactors who work in one another's presence but do not pool their efforts. Therefore, while both types of studies explore individual motivation in social contexts, they examine two very distinct aspects of group life. Taken together, they suggest that the impact that other people have on an individual's effort depends on whether or not the individual is actually working with those others on a collective task. Collective tasks have potentially demotivating properties, whereas coactive tasks tend to increase drive, thereby enhancing performance on simple tasks but reducing it on complex tasks. In other words, social loafing research is targeted to collective group work, whereas social facilitation research is targeted to the mere presence of coactors or observers.

5. *Future Directions*

Existing research on social facilitation and social loafing has revealed many important dynamics of group life, highlighted the conditions under which groups are likely to enhance or reduce individual performance, and provided a number of theoretical perspectives for understanding the dynamics of social motivation. Additional research is likely to provide an even richer understanding and greater guidance as to how groups can be designed and managed to bring out the best in individuals.

See also: Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Processes in Organizations; Group Processes in the Classroom; Group Processes, Social Psychology of

Bibliography

Baron R S 1986 Distraction-conflict theory: Progress and problems. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 19, pp. 1–40

Bond C F, Titus L J 1983 Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin* **94**: 265–92

Cottrell N B 1972 Social facilitation. In: McClintock C G (ed.) *Experimental Social Psychology*. Holt, Rinehart, and Winston, New York, pp. 185–236

Geen R G 1991 Social motivation. *Annual Review of Psychology* **42**: 377–99

Harkins S G 1987 Social loafing and social facilitation. *Journal of Experimental Social Psychology* **23**: 1–18

Jackson J M, Williams K D 1985 Social loafing on difficult tasks: Working collectively can improve performance. *Journal of Personality and Social Psychology* **49**: 937–42

Karau S J, Williams K D 1993 Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology* **65**: 681–706

Kerr N L 1983 Motivation losses in small groups: A social dilemma analysis. *Journal of Personality and Social Psychology* **45**: 819–28

Kravitz D A, Martin B 1986 Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology* **50**: 936–41

Latané B 1981 The psychology of social impact. *American Psychologist* **36**: 343–56

Latané B, Williams K, Harkins S 1979 Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology* **37**: 822–32

Sheppard J A 1993 Productivity loss in performance groups: A motivation analysis. *Psychological Bulletin* **113**: 67–81

Triplett N 1898 The dynamogenic factors in pacemaking and competition. *American Journal of Psychology* **9**: 507–33

Zajonc R B 1965 Social facilitation. *Science* **149**: 269–74

Zajonc R B, Heingartner A, Herman E M 1969 Social enhancement and impairment of performance in the cockroach. *Journal of Personality and Social Psychology* **13**: 83–92

K. D. Williams and S. J. Karau

Group Psychotherapy, Clinical Psychology of

Group psychotherapy is the treatment of psychological, behavioral, emotional, and interpersonal problems in a group context. This article traces the development of group procedures in clinical psychology from its roots as a didactic technique to contemporary psychoanalytic, behavioral and cognitive–behavioral, and interpersonal treatment procedures. These approaches stress different therapeutic mechanisms and methods, but nearly all trace the health-promoting effects of groups to a common set of factors that instill a sense of hope, promote social learning, and induce self-disclosure, emotional expression, and self-insight. Research suggests that group psychotherapy, when properly administered, is as effective as individualized treatment techniques.

1. Approaches to Group Psychotherapy

Group psychotherapy emerged as a means of helping people deal with mental and emotional problems in the early years of the twentieth century. Psychotherapy itself was developing during this period, but most pioneering psychotherapists worked with individual patients only. Some innovators, however, applied their basic change principles in group settings, as did some therapists who turned to groups for didactic purposes or because their patient load demanded a massed approach to treatment. These group applications were often confined to a particular psychiatric facility or clinic, but in time they became more dispersed as other mental health professionals heard of and made use of the group methods. These early psychotherapy sessions involved treating individuals 'in groups, with the group itself constituting an important element in the therapeutic process' (Slavson 1950, p. 42).

Group psychotherapy is used currently to treat a variety of problems, including addiction, thought disorders, depression, eating disorders, and personality disorder. Group psychotherapists are usually credentialed mental health professionals and their patients tend to be suffering from diagnosed clinical conditions. The methods used by group psychotherapists are as varied as individual approaches to therapy. Psychoanalysis, psychodrama, systems-theory, object relations, existential, Gestalt, and humanistic are just a few of the general orientations adopted by group psychotherapists, who tend to be more eclectic than nongroup therapists in their approaches to treatment. Most approaches, however, fall into one of three basic categories: psychoanalytic, behavioral and cognitive–behavioral, and interpersonal (Forsyth and Corazzini 2000).

1.1 Psychoanalytic Group Psychotherapy

Psychoanalytic group psychotherapy draws on Freud's (1922) conceptualization of the causes of dysfunction and his recommendations regarding the most effective way to treat those dysfunctions (see *Psychoanalysis in Clinical Psychology*). The analyst, through such traditional analytic methods as free association, interpretation, and the dissection of transference processes, creates a powerful relationship with the client who then gains insight into unresolved conflicts. The psychoanalytic group psychotherapist, however, exploits the group situation to stimulate self-insight, transference, and the cathartic release of unconscious anxieties. In most cases the therapist acts

as the authority in the group and works with each patient in turn. This rotation gives each patient the opportunity to experience the benefits of one-on-one psychotherapy, but also gives them the opportunity to gain from the analysis of co-members' psychological and emotional difficulties. The group setting, as a recapitulation of a family grouping, also provides more opportunities to work through problems that result from early family conflicts. Just as individual therapy usually stimulates parental transference, group psychoanalysis stimulates sibling transference. Indeed, some experts suggest that Freud's Vienna Circle—ostensibly a study group formed by Freud and his students to discuss theories of adjustment—was a form of group psychotherapy (Roth 1993).

Psychoanalytic group psychotherapists diverge on one key aspect of the process. Some stress the importance of the individual in the group. This approach argues that the therapist should stress individual-level processes in treatment, and avoid spending valuable treatment time exploring the dynamics of the therapeutic group. Others, in contrast, integrate the treatment of the individual with the analysis of the group itself. The group-as-a-whole approach, which can be traced to Bion, Rickman, and Foulkes's work at the Northfield Military Hospital in England, is a sterling example of this approach. Although group-as-a-whole therapists embrace psychoanalytic assumptions of unconscious motivations, personality conflicts, and transference, they make use of the tension between the group and the individual to promote growth and development. Bion, for example, stressed the analysis of defense mechanisms used by groups to cope with ego threats, uncertainties, and anxieties. He believed that these strategies are rooted in dependence, for the group members engage in collective projective identification as they struggle to avoid blame for their problems by shifting responsibility from themselves to the therapist. Bion assumed that rapid and lasting adaptive change could be achieved by helping group members understand both individual and collective defensive processes that unfold in their group session, and his work provided the basis for the Tavistock Institute of Human Relations (Harrison 2000).

1.2 Cognitive–Behavioral Therapy Groups

Cognitive and cognitive–behavior psychotherapists, rather than searching for the cause of the problematic behavior in unseen unconscious conflicts or interpersonal transactions, base their interventions on principles derived from learning theories (Dobson et al. 2000) (see *Dialectical Behavior Therapy*; *Cognitive Therapy*). They identify and quantify desirable cognitions and behaviors that will be encouraged and undesirable cognitions and behaviors that will be extinguished, and then modify these behaviors through cognitive restructuring, self-instructional training, and problem-solving instruction. Cognitive–behavioral therapy groups use these principles with two or more individuals (Rose 1993). Cognitive–behavioral group therapy is generally both structured and problem-focused. Pretreatment assessments identify a patient's current level of functioning, and interventions are calibrated to fit the patient's level of functioning and targeted outcomes. The therapeutic interventions usually use a constellation of behavioral procedures, including modeling, rehearsal, and feedback.

1.3 Interpersonal Group Psychotherapy

Many group psychotherapists not only exploit the social processes that occur in groups during treatment, but they also trace psychological disturbances back to social sources—particularly interactions with friends, relatives, and acquaintances. Interpersonal theorists assume that behavior results from individuals failures to take note of, and correct, cognitive and behavioral tendencies that disrupt their relationships with other people (see *Interpersonal Psychotherapy*).

Many group therapists, recognizing the interpersonal bases of psychological problems, use the group setting to help members examine their interpersonal behavior. Yalom's (1995) interpersonal group psychotherapy (also called interactive group psychotherapy), for example, stresses the 'here-and-now' of the group's interaction over the analysis of events and experiences that occur outside of the group (the 'then-and-there'). Members may spend some time, especially early in the group's life span, discussing problems they are facing at home or work, but Yalom feels that much of this discussion will be futile since members' characterize such events inaccurately. Yalom instead recommends that members focus on interactions within the group, with each member eventually responding to one another as they do to others in their lives. The group thus becomes a 'social microcosm,' giving members true insight into the causes of the personal problems.

2. Therapeutic Factors in Groups

Group psychotherapy is based on the Lewin's law of psychological change: 'It is easier to change individuals formed into a group than to change any of them separately' (Lewin 1951, p. 228). But what aspects of the group promote positive change? A common-factors perspective on group therapy suggests that various types of group psychotherapies, despite their many differences, share some common properties, and these commonalities may be responsible for their effectiveness. A psychodynamic therapist uses different techniques than a cognitive-

behavior or interpersonal therapist, but these various formats share mechanisms that promote change and sustain well-being. Yalom (1995) terms these shared qualities curative factors, and his list includes the installation of hope, universality, the imparting of information, altruism, the corrective recapitulation of the primary family group, the development of socializing techniques, imitative behavior, interpersonal learning, group cohesiveness, catharsis, and existential factors.

2.1 Social Motivation in Group Psychotherapy

Groups, by their very nature, satisfy a range of basic human social motives. Although individuals who are suffering from some behavioral or psychological problem may not find 'safety in numbers,' they can find 'reassurance in numbers.' When they assemble in a therapeutic group they quickly realize that the situation they face is shared by others. Yalom (1995) calls this realization universality, for patients realize that they have not been singled out to bear some extraordinary misfortunate; rather, the problem they face is a universal one.

Therapeutic groups also capitalize on social comparison processes to stimulate growth and adjustment. Social comparison occurs when individuals spontaneously compare themselves to other people around them. This comparative process does not occur in individualistic therapies, for patients can only imagine how they are coping relative to others. In groups, in contrast, patients learn, both directly and indirectly, about one another's functioning. These comparative data yield two useful consequences. First, comparison provides patients with a baseline to calibrate their own level of functioning and determine where they stand along the 'coping vs. collapsing' continuum (Yalom 1995). They may learn, through downward social comparison—comparing themselves to those in the group who are performing poorly—that their own situation is not so bleak as they originally assumed. Second, this comparison can increase their sense of optimism about reaching an improved level of functioning. Yalom (1995) terms this process the installation of hope. Whereas individuals facing unpleasant circumstances alone may feel discouraged and pessimistic, when group members recognize that others are improving their motivational levels.

2.2 Interpersonal Learning

Much of the learning that occurs during psychotherapy is social in nature. Under the guidance of the therapist, patients explore and gain insight into sources of anxiety, they identify and practice behaviors that they can use to cope with sources of stress, and they acquire the skills they need to interact with others. The therapist models the behaviors they want clients to adopt, and shape their behaviors through feedback and selective reinforcement. When therapy is delivered in a group context, however, the sources of social learning are multiplied. Group psychotherapy necessarily involves patients interacting with other patients, and so it sets up opportunities for interpersonal learning among patients. Some therapists purposely minimize the impact of the group on the individual patients, but most group therapists exploit these social processes to promote change.

Yalom (1995) argues that much of the curative power of a group springs from such social processes as imparting of information from one person to another (guidance), the development of socializing techniques, imitative behavior, and interpersonal learning. For example, the exchange of information among participants was one of the first purposes of groups, for the earliest group therapies were primarily didactic in nature, with psychotherapists lecturing to large groups of patients about mental and physical health. Interpersonal learning also occurs indirectly, as group members monitor their impact on the other people within their group, and the psychotherapist helps them draw conclusions about their maladaptive interpersonal tendencies. Group members also gain from simple observation of others. By watching others' emotional displays, problem-solving behaviors, coping strategies, and interpersonal maneuverings, patients acquire adaptive social skills.

2.3 Emotional Experiences and Group Development

Groups set the stage for strong emotional experiences for members. As therapeutic groups develop over time, members become closely connected to each other emotionally, developing bonds of friendship that are as strong as any they experience in nontherapeutic relationships. Members also report a strong level of identification with the group, with the result that they are more willing to accept the group's influence.

The range of emotions experienced depends on the group's stage of development. Early in the group interactions tend to be superficial, but in time conflicts tend to surface. In most groups, these conflicts between members and the group therapist give way to the development of cohesion. This progression of emotional experience, although characterized in a variety of ways, generally follows the forming, storming, norming, performing, and adjourning pattern identified by Tuckman (Tuckman and Jensen 1977). During the forming stage members are uncertain how to behave, and do not trust one another enough to disclose personal information or to admit they need help. During the storming stage group members often find themselves in conflict over the purpose of the group, and they may withdraw from the group and the therapist. In groups where the leader takes a nondirective orientation, this conflict stage focuses on the leader's failure to 'take charge' of the interactions.

During the next phase, norming, the group becomes more cohesive, as norms and standards emerge that dictate how each session will unfold. This stage makes the performing stage possible, where group members engage earnestly in the exchange of information that will help them achieve their therapeutic goals. In time, members must deal with issues of termination as the group adjourns. This movement from one stage to another is therapeutic, so much so that a group that does not move through these stages may not benefit its members (Yalom 1995).

2.4 Self Processes in Therapeutic Groups

All group psychotherapies stress the value of groups as vehicles of self-understanding and self-efficacy. Although group members may initially be reluctant to disclose any information about themselves in the relatively public group setting, as they develop stronger bonds of trust their disclosures change from superficial to substantial. Group members also report feeling gains in self-efficacy as they help others in their group with their problems.

Members particularly value the self-insights they achieve in groups. In interpersonal therapy, in particular, patients often act in the therapeutic group as they do in other social contexts, so much so that the behaviors that caused them interpersonal difficulties erode their relationships within the group. When their fellow group members and therapist call their attention to the parallel, patients experience greater insights into their own motives and emotions. In many cases, as Freud's analysis of groups suggests, these tendencies can be traced to early childhood experiences in the family of socialization.

3. The Effectiveness of Group Psychotherapy

When they were initially proposed, skeptics questioned the wisdom of putting people who are suffering from psychological problems together in one group. How, they asked, could troubled individuals be expected to cope in a group when they had failed individually? And how could the therapist guide the therapeutic process in a group, given the subtle nuances that pervade one-to-one therapies? In time, however, the method gained widespread acceptance as therapists added group sessions, either supplementing or completely replacing their individual sessions.

But do members leave the group better adjusted than when they entered? Meltzoff and Kornreich (1970) were guardedly optimistic about the utility of group therapies because 80 percent of the methodologically sound studies reported either major or minor benefits for clients, whereas nearly all of the studies that reported no benefit were methodologically flawed. Bednar and Kaul (1994) were guardedly positive, although they questioned the quality of the outcome research in group psychotherapy. Meta-analytic reviews also suggest that group approaches are as effective as individual methods (e.g., Fuhriman and Burlingame 1994). Fuhriman and Burlingame (1994), for example, reviewed 700 group therapy studies and seven meta-analytic studies before concluding that group methods are effective treatments for a wide variety of psychological problems. Similarly, Faith et al. (1995), in a meta-analytic review of 63 studies of sensitivity training, concluded that these groups generally led to increases in self-actualization and self-esteem and improved interpersonal relations. They noted that these effects increased in larger groups, when the groups met for longer periods of time, and when the measures focused on behavioral outcomes rather than self-reported ones. McRoberts et al. (1998) found that group therapies were more effective with clients (a) who were not diagnosed clinically, (b) suffering from substance abuse problems and/or chemical dependencies, and (c) who attended 10 or fewer sessions. Older studies—those conducted prior to 1980—were more likely to favor group over individual approaches.

This positive conclusion, however, requires some qualification. First, the empirical evidence is not definitive. Whereas a number of reviews are positive, others conclude that group therapy is not as potent as individual therapy. Second, the changes brought about by group experiences may be more perceptual than behavioral. Group members frequently report achieving major insights into their behavior and are positive about the gains they have achieved, but objective indices of outcome are more contradictory. Self-reported insight is also related to actual improvements in adjustment. Third, as Bednar and Kaul (1994) note, a participant may decide to leave the group before they have benefited in any way, and in rare cases patients are significantly harmed by the group experience. Last, far more data of better quality are needed to settle the question of efficacy. As Forsyth and Corazzini (2000, p. 325) conclude, the 'use of varied and undocumented therapeutic methods, with different types of clients, by therapists who differ in skills and experiences, in studies that too frequently lack valid measures and adequate controls, make it difficult to draw firm conclusions.'

Group psychotherapy's empirically supported usefulness is reason enough for its continued growth, but the approach will also prosper as health maintenance organizations require, and patients increasingly seek, group-level treatment methods. Group psychotherapy is an efficient therapy, for a single therapist can treat more than one individual at a time—a characteristic that is considered very valuable in healthcare systems that seek to minimize costs without cutting quality. But group approaches may also become more popular as clients seek out supportive, stimulating environments for achieving personal change. Indeed, the

increasing popularity of self-help groups signals the need for treatments that take advantage of groups for promoting adjustment. Although the use of groups in business and educational settings was once considered innovative, individuals now expect to learn and work in group contexts. These experiences, combined with a growing collectivism in cultures that were once primarily individualistic, should heighten awareness of groups as a means of treatment.

See also: Cognitive Therapy; Group Processes, Social Psychology of; Groups in Special Environments; Interpersonal Psychotherapy; Psychoanalysis in Clinical Psychology; Psychological Treatment, Effectiveness of; Therapist–Patient Relationship

Bibliography

Bednar R L, Kaul T 1994 Experiential group research: Can the canon fire? In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York, pp. 631–63

Dobson K S, Backs-Dermott B J, Dozois D J A 2000 Cognitive and cognitive–behavioral therapies. In: Snyder C R, Ingram R E (eds.) *Handbook of Psychological Change*. Wiley, New York, pp. 409–28

Faith M S, Wong F Y, Carpenter K M 1995 Group sensitivity training: Update, meta-analysis, and recommendations. *Journal of Counseling Psychology* **42**: 390–9

Forsyth D R, Corazzini J G 2000 Groups as change agents. In: Snyder C R, Ingram R E (eds.) *Handbook of Psychological Change*. Wiley, New York, pp. 309–36

Freud S 1922 *Group Psychology and the Analysis of the Ego*. Hogarth, London

Fuhriman A, Burlingame G M 1994 Group psychotherapy: Research and practice. In: Fuhriman A, Burlingame G M (eds.) *Handbook of Group Psychotherapy: An Empirical and Clinical Synthesis*. Wiley, New York, pp. 3–40

Harrison T 2000 *Bion, Richman, Foulkes, and the Northfield Experiments*. Kingsley, London

Lewin K 1951 *Field Theory in Social Science*. Harper, New York

McRoberts C, Burlingame G M, Hoag M J 1998 Comparative efficacy of individual and group psychotherapy: A meta-analytic perspective. *Group Dynamics: Theory, Research, and Practice* **2**: 101–17

Meltzoff J, Kornreich M 1970 *Research in Psychotherapy*. Atherton, New York

Rose S D 1993 Cognitive–behavioral group psychotherapy. In: Kaplan H I, Sadock M J (eds.) *Comprehensive Group Psychotherapy*, 3rd edn. Williams and Wilkins, Baltimore, MD, pp. 205–14

Roth B E 1993 Freud: The group psychologist and group leader. In: Kaplan H I, Sadock M J (eds.) *Comprehensive Group Psychotherapy*, 3rd edn. Williams and Wilkins, Baltimore, MD, pp. 10–21

Slavson S R 1950 Group psychotherapy. *Scientific American* **183**(6): 42–5

Tuckman B W, Jensen M A C 1977 Stages of small group development revisited. *Group and Organizational Studies* **2**: 419–27

Yalom I D 1995 *The Theory and Practice of Group Psychotherapy*, 4th edn. Basic Books, New York

D. R. Forsyth

Groups in Special Environments

1. Introduction

'Special' environments include those that require extraordinary facilities for human survival, as well as those that temporarily pose dangers because of disaster, accident, or violent conflict. Also known as 'extreme and unusual environments' (EUEs), they have always engaged high public interest. Since the 1950s, social scientists have begun the systematic study of how people react in and are affected by experiences in hazardous habitats. These include life-support capsules (under water, in space, near the Poles, on offshore oilrigs), natural and industrial cataclysms, combat zones, refugee and concentration camps.

To date, most of the research emphasis has been on the ways in which people react to their experience (Wilson et al. 1988), on methods for preventing, minimizing, or treating negative outcomes (Holland and Curtis 1998), and on how individuals are selected and trained for some special environments, such as space and submarine missions (e.g., Santy 1994). Methods used in these efforts have included some experiments; but the emphasis has been on simulation and analogue studies, questionnaires, interviews, and the content analysis of archival materials.

Although researchers have focused on the individual, the behavior of human beings in special environments is to a great extent the story of groups. Solitary individuals have found themselves in such situations, sometimes by accident (shipwrecked mariners, sole survivors of other disasters) and sometimes on purpose (aboriginal youths on spirit quests, solo mountaineers, hermits, single-handed sailors, early astronauts, and adventurers of various sorts); but the modal inhabitants of EUEs are not alone.

Many aspects of the environment itself strongly influence group relations and behavior. Some of these are: the dangers involved, life-support equipment, duration of the stay, deviation from accustomed physical parameters (temperature, humidity, air quality and pressure, gravity, noise, crowding, comfort, the availability of food and water, and shelter), remoteness from home, the possibility of communication with one's family and the parent organization, medical care, and rescue if needed. Whether the group is caught up in an extreme environment unwillingly, by accident or superior force, or has volunteered, and its motives for volunteering (excitement, prestige, patriotism, curiosity, money, altruism) are also significant. Last, it is useful to distinguish between organized

groups (teams, crews, units) and random conglomerations of individuals who happen to be thrown together in an unusual environment.

Many issues apply only to a subset of EUE situations; for example, polar, space, undersea, and similar missions require a relatively small group to be chosen from a larger number of applicants. This is obviously less true of the recruitment of combat soldiers, and not at all of groups caught up in accident or disaster. Nevertheless, there are concerns that are common to many kinds of groups, in a range of extreme environments. It is these general aspects of group functioning in EUEs that will be covered in the subsequent examination of the existing literature.

2. *Cohesiveness*

One striking pattern is the deep interpersonal bonding that many groups experience, existing alongside other groups' very high levels of friction and hostility. In extreme situations, self-sacrifice is not unusual, from Captain Lawrence Oates of Scott's South Pole team who, when supplies were running too low to sustain the whole group, left their shelter and went off to certain death in a blizzard to the real-life hero of the World War II ballad, Rodger Young, who 'fought and died for the men he marched among.' Studies of combat soldiers show that their primary motivation for risking their own safety is identification with their leader and unit, and comradeship with their fellows. Soldiers who have not developed such ties—for example, recent replacements—are much more likely to fail the test of battle (Manning 1991). Even in less organized groups, such as civilians in cities undergoing aerial bombardment or a natural disaster, high levels of mutual helpfulness and feelings of solidarity are commonplace.

In capsule EUEs, being confined in a small space with the same people for a long time, under conditions of danger, discomfort, fatigue, lack of privacy, and other stressors, is a potential source of hostility. Admiral Byrd (1938), among others, reflected on the time when a colleague's 'pet ideas become a meaningless drool, and the way he blows out a pressure lamp or drops his boots on the floor or eats his food becomes a rasping annoyance.' Intimate disclosures, encouraged earlier by loneliness, proximity, and boredom, in retrospect seem embarrassing and regrettable. Violations of personal space or possessions seem contemptuous provocations, irritating habits are perceived as deliberate affronts, and minor disagreements can become infuriating. The results range from vehement arguments through ostracism and physical attacks, to murderous assaults (Suedfeld and Steel 2000).

One of the most pervasive findings in isolated groups is that of dislike and disrespect between subgroups, such as polar scientists and support personnel or people from different cultural or national backgrounds (Kanas 1998). This is a growing problem, as many projects involving EUEs are becoming international and/or multicultural.

3. *Workload*

One important aspect of EUE groups is that they are often overworked as well as stressed. In many extreme environments, living and working space is limited, facilities are functional and Spartan, and duties are both important and demanding. Survival as well as mission success may depend on everyone's performance, so that there are superordinate goals—important to all, and unreachable without total effort and cooperation. Social loafing or shirking is, while certainly not unknown, relatively rare and easily detected. In fact, studies have shown that members of submarine, space, and polar crews spend much of their supposed recreational or resting time working, sometimes leading to severe fatigue. In other situations, such as disaster rescue and combat, exhaustion may be the result of extreme physical exertion and danger. Insomnia, which is endemic in many EUEs, exacerbates the problem, as do physical discomforts including noise and duties that interrupt sleep (Stuster 1996).

Even the most dedicated crews may come to resent the demands made on them by the situation or by the organization to which they belong. There have been minirebellions or simple refusals to comply with work schedules by groups in both the American and Soviet space programs, as well as in many other EUE settings (Harrison et al. 1991). Typically, these have been brief and moderate, consisting of a self-declared and temporary 'time out,' although in some more extreme cases in wartime and extended sailing voyages they became full-blown mutinies.

4. *Boredom*

There is a quite opposite situation as well: in many special environments, there may be long stretches of unstructured time when there is little if any important work to be done. This has been true of groups of prisoners of war, ships' crews on multiyear voyages or frozen into the polar ice, personnel in remote stations, civilians in air raid shelters and refugee camps, troops between active duty deployments, resource extraction workers waiting for equipment repairs, resupplies, or further exploration, emergency workers (firefighters, paramedics) between calls, and so on. Spacefarers spending a year or more en route to their destination and then again on the way back will experience this (in contrast to astronauts in short-term, overprogrammed missions such as those of the Space Shuttle) (Stuster 1996).

Under physical and social monotony, free time becomes a time of boredom, lassitude, and apathy.

The desire to explore and manipulate—play with—elements of the environment becomes important. The group may manufacture artificial challenges, sometimes taking unnecessary risks either unthinkingly or deliberately. As individuals lose interest and motivation, they may neglect personal hygiene and withdraw from the group. As a result, and because hostile exchanges at least break the monotony, interpersonal arguments and conflict may intensify.

Alert leaders of organized groups make sure that no one has too much free time: initiating systematic recreational activities, getting people to put in extra work on cleaning and maintenance, and introducing novel work assignments such as cross-training crew members. In Arctic whaling and exploring ships of the eighteenth and nineteenth centuries, sailors and officers put on elaborate theatrical and musical performances, making their own scenery and costumes; competed in athletic events; and published shipboard 'newspapers' filled mostly with stories and poetry, since there was very little real news (Johnson and Suedfeld 1996). In the ghettoes of Nazi-dominated Europe, Jewish authorities arranged elaborate cultural, scientific, religious, and educational programs. Even in lifeboats drifting after shipwreck, officers have organized duty watches, group sing-songs, story-telling, and the like. In all of these environments, such strategies help to maintain morale, ensure that necessary work gets done, and divert group members from their danger, deprivation, and fear.

5. *Group Culture*

Group culture represents an identity, made up of the characteristics of the members and what emerges from their interactions, which identifies the in-group as distinct from all others. This phenomenon is particularly marked in EUE groups: nicknames, jargon, practical jokes, initiation rites, and so on, weld the group together and simultaneously express and improve its morale. So do adornments such as bizarre hair-styles or beards, and special symbols such as badges, patches, pins, special articles of clothing, and other unique identifiers.

This pattern of values, activities, norms, jokes, and general lifestyle characteristics becomes particularly meaningful to isolated groups and frequently incomprehensible to outsiders. Such groups have an awareness of shared hardships and triumphs, and feel removed from the outside. As an example, a group of aquanauts living in an undersea environment came to refer to the rest of humanity as 'Earth people' (Radloff and Helmreich 1968), and of course the designation 'the world' is widely used among polar personnel and foreign-service soldiers when referring to home. The use of in-group jargon is also a way to deliberately stigmatize newcomers and outsiders (Taylor 1987).

Rituals, special foods, and the celebration of major occasions are important aspects of group culture. For centuries, these have been used to strengthen a feeling of solidarity and continuity, and to exercise everyone's ingenuity, have fun, and reduce boredom and they continue to do so in the Space Age.

6. *Leadership*

Leaders of EUE groups have a role that is in many ways more crucial to the group's success than those normally encountered in most modern organizations. They may be cut off from the parent organization, so that they can not receive advice or instructions from their superiors; even when communication is possible, it may be long delayed. In any case, because of the unique circumstances, supervisors or consultants 'back home' may not fully understand all of the relevant factors—or the group may think that they do not, and therefore may be reluctant to follow instructions or advice (Connors et al. 1985). In such cases, the leader on the spot must decide how to handle the uncertainty and conflict.

The leader's formal rank confers the power of reward and punishment; but the impact of these can be attenuated in long-duration remote work, where any material reinforcers may be trivial or far in the future, and in extremely dangerous environments where any 'punishment' may in fact increase the transgressor's safety (e.g., a court-martial that would involve a transfer to the rear echelon). Even worse, the leader may come to be seen as dictatorial, petty, and unfair, with consequences that can range from the unpleasant to the disastrous. The best EUE leaders are those who use formal authority sparingly but are admired, perceived as holding authority legitimately, seen as a role model to be emulated, and believed to be expert in solving the problems that the group faces (Stuster 1996).

Good interpersonal relations are made even more important by the fact that in EUEs the trappings that normally keep leaders remote and aloof—secretaries and aides, imposing offices, separate facilities, and so on—are drastically attenuated (Kanas 1998). Social leveling occurs even in traditionally hierarchical groups, such as military units. The leader is revealed as having human foibles and faults. If the crew dislikes or disrespects the leader, it has many opportunities to express its feelings. Conflict between the leader and a key subordinate or a clique of group members, is common under stress and requires wise strategies balancing tact and forcefulness. Excessive consciousness of rank differences may also lead to disaster for the group: social psychologists have identified a number of airline crashes that could have been avoided if a crew member who noticed that something was wrong had dared to assertively interrupt the captain while the latter was paying attention to something else (Helmreich 1984).

How the leader should function varies with the circumstances. EUE groups have indicated that on matters of a technical nature, like navigation or equipment repair, the leader is expected to make decisions in consultation with the relevant technical experts; on issues of routine policy, like work assignments and recreational schedules, the leader should consult with the entire group in an open way, even if the group's preferences are not followed exactly; and in emergencies, the leader's clear, unhesitating, and unilateral decision is appreciated even by those who otherwise resent direction (Stuster 1996, Suedfeld and Steel 2000).

The difficulties related to EUEs are especially severe for the leader: resolving other people's environmentally caused stresses while one is experiencing the same ones is a true challenge to leadership. One development that arises only too frequently is the leader's becoming a source of stress for the rest of the group. In an EUE, attributes that can be tolerated in a normal environment become unbearable. Some leaders are too unstable, demanding, threatening, perfectionistic, or inconsiderate—traits that may be part of their personality, or that emerge because they too are under stress. These traits lead to decreased leader influence, and to increased resistance on the part of other group members. In long-duration special environments, it is probably not a good idea to rely too much on a crew's ability to tolerate the commander (Kanas 1998).

7. *Sex and Gender Issues*

The presence of growing numbers of women in formerly all-male EUEs has affected morale and performance, and raises some interesting issues. Almost all Western military forces, national Antarctic programs, and space programs have by now included female members and some leaders. These developments have changed group dynamics, in as yet undetermined ways. If a recent all-female Antarctic traverse expedition is typical, a shift toward more emphasis on group cooperation and solidarity may be occurring, perhaps with some decrease in goal orientation, and it certainly seems that women in mostly male groups both provide and are offered more interpersonal support and nurturance than has been usual (Rothblum et al. 1998).

There is a lively controversy among military and polar officials as to whether sexual integration has resulted in additional or reduced sexual tension, and as to just how pair relationships affect the participants and the rest of the group. Some people are less sexually frustrated, but others are not; resentment, jealousy, and envy have surfaced. In at least one Antarctic case, there was a serious drop in station morale when the cook's romantic partner abandoned him and began a relationship with someone else, and meals became unappetizing. One can only speculate what would happen on a long-duration EUE mission (or in an emergency) if, for example, the group leader were involved in this kind of triangle (Suedfeld and Steel 2000).

8. *Selection and Training*

This topic applies to only a subset of special environments, but is important in those. There are two phases of selection: selecting out (eliminating applicants below the threshold) and selecting in (picking the best possible crews). The former is done on an individual basis, using biodata, interviews, and tests; in the second stage, both individual and group selection become relevant. The interaction and performance of potential crewmates are observed in analogue or simulation situations (field and laboratory settings respectively, both incorporating some aspects similar to the eventual EUE). For example, candidates may go through a wilderness survival course or (in the ex-Soviet space program) undertake a long cross-country journey in a small automobile. Many programs choose and train crews as a unit; if for some reason one member must be struck from a mission, the entire crew is replaced (Santy 1994).

9. *Future Directions*

The number of people exposed to EUEs, fortuitously or deliberately, is not likely to diminish in the near future; and as the sheer technical and physical difficulties are solved, more attention will be paid to psychosocial issues. An emphasis on understanding group reactions, not only individual ones, is emerging. As the connections between EUEs and 'general' social science areas become clearer, research will become less atheoretical and applied; EUEs provide unique opportunities for both gathering new knowledge and having an impact on serious real-life problems, which may tempt more social scientists into the field (both literally and metaphorically).

See also: Culture and Emotion; Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Groups, Sociology of; Leadership in Organizations, Psychology of; Military and Disaster Psychiatry; Military Psychology: United States; Violence and Effects on Children

Bibliography

Byrd R E 1938 *Alone*. Putnam's, New York

Connors M M, Harrison A A, Akins F R 1985 *Living Aloft: Human Requirements for Extended Spaceflight*. NASA, Washington, DC

Harrison A A, Clearwater Y A, McKay C P (eds.) 1991 *From Antarctica to Outer Space: Life in Isolation and Confinement*. Springer-Verlag, New York
Helmreich R L 1984 Cockpit management attitudes. *Human Factors* **26**: 583–89
Holland A W, Curtis K 1998 Operational psychology countermeasures during the Lunar–Mars life support test project. *Life Support and Biosphere Science* **5**: 445–52
Johnson P J, Suedfeld P 1996 Coping with stress through creating microcosms of home and family among Arctic whalers and explorers. *History of the Family* **1**: 41–62
Kanas N 1998 Psychosocial issues affecting crews during long-duration international space missions. *Acta Astronautical* **42**: 339–61
Manning F J 1991 Morale, cohesion, and esprit de corps. In: Gal R, Mangelsdorff A D (eds.) *Handbook of Military Psychology*. Wiley, New York
Radloff R W, Helmreich R L 1968 *Groups Under Stress: Psychological Research in SEALAB II*. Appleton-Century-Crofts, New York
Rothblum E D, Weinstock J S, Morris J F (eds.) 1998 *Women in the Antarctic*. Harrington Park Press, Binghamton, NY
Santy P A 1994 *Choosing the Right Stuff: The Psychological Selection of Astronauts and Cosmonauts*. Praeger, London
Stuster J 1996 *Bold Endeavors: Lessons from Space and Polar Exploration*. Naval Institute Press, Annapolis, MD
Suedfeld P, Steel G D 2000 *The Environmental Psychology of Capsule Habitats. Annual Review of Psychology*. Annual Reviews Press, Chippewa Falls, WI
Taylor A J W 1987 *Antarctic Psychology*. DSIR, Wellington, NZ
Wilson J P, Harel Z, Kahana B (eds.) 1988 *Human Adaptation to Extreme Stress: From the Holocaust to Vietnam*. Plenum, New York

P. Suedfeld

Groups, Sociology of

A group is a bounded collection of interacting individuals or a consensual social categorization. It is difficult to distinguish sociological and psychological studies of groups and, in the past, there has been a considerable overlap. Still, one feature may help to make a reasonably accurate distinction. The sociological study of groups always involves attention to collective phenomena in addition to the concept of the group itself. The psychological study of groups sometimes does, but often does not, involve attention to these phenomena. In the work of sociologists, processes of interaction and of individual behavior in social contexts are seen as giving rise to durable or semidurable collective phenomena (such as cohesion; relational structures pertaining to status, communication, friendship, power, control; structural similarity; distribution of information; collective goods provision; social norms; collective representations; styles of conflict resolution; patterns of group decision-making).

These small-scale collective phenomena, in turn, influence the behavior of individuals and their interactions (see Fig. 1). Some sociological work concentrates on relation 1, some on relation 2, and some on 1 and 2, and other combinations. Relation 3b allows explicit feedback mechanisms involving both levels (as in expectation states theory, see below). Psychological studies often focus on the direct influence of interaction (or the presence of others) on attitude change, productivity, satisfaction, self-esteem, depression, etc., occasionally allowing for feedback mechanisms (3a and 3b). The extra arrows in Fig. 1(i) indicate exogeneous factors (such as genetic dispositions, rules, established categories, status differences).

1. *Historical Background of Small Group Research*

The scientifically serious investigation of groups in the sociological tradition started with anthropological studies of tribes, clans, and families in the nineteenth and early twentieth centuries. Sumner, Simmel, and Cooley were early codifiers of that kind of research. The most important turn in this development occurred

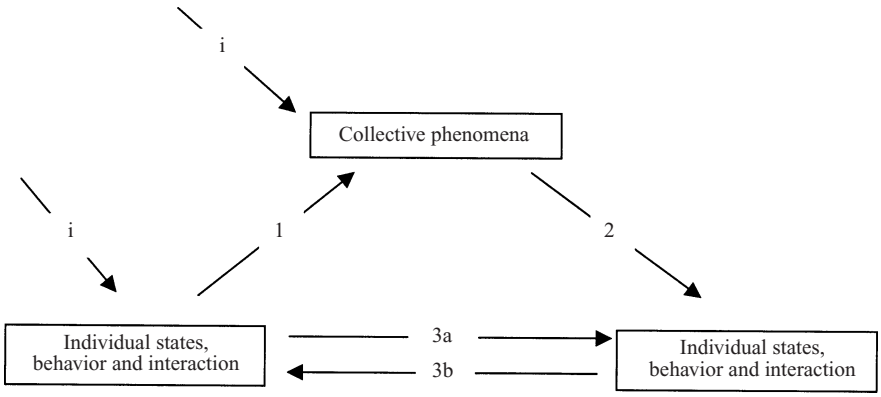

Figure 1
Schematic depiction of the explanation of groups

in the 1930s. Within the span of a few years, four great books appeared which had two very important messages in common: that groups are constituted by a cognitively mediated functional interdependence between individuals, i.e., people who are dependent on each other for achieving a goal interact, and something important happens to their minds and the way their minds operate when they interact. In addition, the actual situation in which their action is placed matters most for behavior (as opposed to traits or other cross-situational constants). In 1934 Mead's (then already posthumous) *Mind, Self and Society* appeared. Two years later, Lewin's *Principles of Topological Psychology* and Sherif's *The Psychology of Social Norms* came out. A year later, Parsons' *The Structure of Social Action* made its appearance. Merton's early work on reference groups was also quite influential.

From this time, group studies continued in two traditions—experimental and field studies—and both traditions bore the mark of this convergent conception of groups. Under the leadership of Lewin and Sherif, 'group dynamics' as the experimental study of groups was then called, achieved great heights. Field studies had Mayo's Hawthorne studies and Whyte's famous *Street Corner Society* as models. Both traditions owe their success to the simultaneous consideration of the three fundamental interdependencies in groups: *functional* (related to goal achievement), *cognitive* (related to common categorizations, beliefs, and knowledge) and *structural* (related to the configurations of social relations) (see Lindenberg 1997 for more detail on this and other aspects of this article).

However, the precise nature of the interdependencies was not yet well investigated and in the late 1950s and early 1960s, specialized groups of studies on each of these interdependencies became dominant, with interesting crossovers between psychology and sociology. Homans (1950, 1958), a sociologist, combined the field and the experimental traditions in the study of functional interdependencies in exchange relations. Inspired by Homans, two psychologists, Thibaut and Kelley (1959), worked out the study of small groups as the study of functional interdependencies. Their book is to this day one of the best on the subject. It influenced Homans' further work on the subject (Homans 1961). Cognitive interdependencies were elaborated by Sherif and turned into a major strand of group research called 'social identity theory' by Tajfel (1982) and Turner (1982), both psychologists but very influential among sociologists and in turn influenced by sociological symbolic interactionism. In a related development, Moscovici, a psychologist, elaborated Durkheim's theory of collective representations (see Flick 1998). Structural interdependencies were more formally elaborated by the psychologist Cartwright and the mathematician Harary and turned into a full-blown network approach by sociologist White and his followers (see White et al. 1976).

2. *Bringing Different Kinds of Interdependencies Back Together Again*

This section deals briefly with the field study approach and in some more detail with experimental studies. No attempt is made to be complete. Rather, the overview is selective. The selection criteria are first, that the studies should be substantively important. Second, they must mainly focus on group processes comprising small groups (i.e. face-to-face interaction with at least one common goal), small networks, and face-to-face encounters. Third, the studies must be seriously concerned with the integration of at least two of the three fundamental interdependencies. Fourth, they must cluster around a program of investigation with some continuity.

With regard to field studies, work inspired by Goffman was chosen. Goffman's studies provided a new integration for field studies in sociology (excluding Mayo-type studies). With regard to experimental (and simulation) studies, expectation states theories, various groupings of social exchange theories, and studies of friendship and cognitive networks were selected. Studies of social dilemmas are covered elsewhere (see *Cooperation: Sociological Aspects*). Helpful discussions of a broader scope can be found in Cook et al. (1995) and Foschi and Lawler (1994). Due to space restrictions, the reader will at times be referred to bibliographies and collections rather than individual works.

2.1 *Fieldwork Studies*

From the beginning of the 1960s, fieldwork studies on face-to-face groups have received their main theoretical guidance from Goffman (1961). He also put short-lived groups, or 'encounters', on the map. There is a relatively small but very productive circle of contributors to this kind of research (a few important examples are Anderson 1978, Cahill and Lofland 1994, Fine 1987, MacLeod 1995). These studies are concerned with conformity, status, group divisions, identities, and their combination in impression management. They are rich in descriptive detail. Aspects of functional and cognitive interdependencies are clearly present in Goffman's work.

2.2 *Experimental Studies*

2.2.1 Expectation states theories. Generic expectation states theory is concerned with the formation and maintenance of power and prestige structures in face-to-face task groups (see literature in Berger and Zelditch 1993). Following the tradition of Bales in the 1950s, power and prestige are conflated and their structure is conceived in terms of behavioral aspects

(opportunities and influence) rather than the distribution of scarce resources. People are higher on the power and prestige structure, the more opportunities to act they receive, the more often they take advantage of these opportunities, and the more often they win disagreements if and when they occur.

The basic mechanism that creates this structure is pairwise interaction (between, say, Ego and Alter) that is governed by performance expectations. It is the (unobservable) structure of these performance expectations which keeps generating the observable power and prestige structure. When Ego has positive performance expectations towards Alter, he or she will give more action opportunities to Alter and Alter will more likely take up each opportunity offered. Alter has influence over Ego. It is assumed that there is an evaluation bias. Ego is more likely to judge Alter's performance to be positive if the performance expectation was positive and vice versa. And if Ego nonetheless voices disagreement, he or she is more likely to eventually yield to Alter if the performance expectation was positive to begin with. Such evaluation biases greatly enhance the stability of the power and prestige structure (provided the task endures).

The theory is meant to apply to task groups. For this reason, functional interdependence is by definition the starting point. Here, a joint task and face-to-face interaction is what makes a collection of individuals into a group. Each group member's contribution influences a group score (mostly outcome interdependence, say, linking bonus to agreement in answers). There are also cognitive aspects. For the theory to apply, group task achievement must be a salient goal in the interaction with other members ('collective orientation'). The central cognitive element in the theory is a performance expectation. It presupposes other cognitive elements, such as goal criteria and causal beliefs about how likely a particular characteristic or behavior is to lead to success or to failure. Cognitive interdependence comes about when the performance expectations (and, prior to that, the goal criteria and causal beliefs) are being shared. Given the evaluation bias, the interdependence means that performance expectations of different group members interlock (see Moore 1985).

Functional and cognitive interdependencies are closely interlinked in this theory. This link is made even stronger through extensions of the theory to include external status characteristics (see Berger and Zelditch 1993) and the generation of status characteristics from nominal distinctions (see Webster and Hysom 1998). Task groups are embedded in a wider culture in which such status and competence beliefs exist (such as 'men are better at mathematics'). People seem to follow the maxim 'when in doubt, assume that a salient characteristic is task-relevant' (the so-called 'burden of proof principle'). With regard to gender, this favors men in general tasks as well as in 'masculine' tasks, and it favors women only in stereotypically 'feminine' tasks. Through the link of performance expectations with worthiness, there is also a link of expectation structures to conceptions of just rewards, which is worked out in the 'status value theory' of distributive justice (see Berger and Zelditch 1993).

Ridgeway has shown that, in turn, such cultural beliefs can be explained by the diffusion of *status-belief-generating mechanisms* in small groups. When interaction partners in task groups are doubly dissimilar in that they differ in valuable resources (say high pay level) and in a salient nominal distinction (such as race or gender), then a process of misattribution will causally link having the valuable resource with the nominal distinction and eventually with assumptions of competence, worthiness, and esteem. The group consensus on this attribution generalizes to the idea that 'most people' believe it to be so (Ridgeway et al. 1998). Group members are likely to transfer such beliefs to other group contexts, thereby spreading and reinforcing them across groups and turning them into cultural status beliefs. Similar mechanisms are now invoked to explain legitimation and delegitimation of power and prestige structures in groups (see Berger et al. 1998).

2.2.2 Theories of exchange in networks. Theories of social exchange by Blau (1964), Homans (1950), and Thibaut and Kelley (1959) have given rise to a number of theories on the influence of social structure on the distribution of exchange ratios or benefits within groups. The results of these theories are also relevant to the explanation of network effects on power advantages on a macro level (see, e.g., Burt 1992). Paradigmatic for these theories is Emerson's (1976) concept of 'network connection,' which points to externalities between exchange relations. Two exchange relations are connected if and only if the exchange in one relation affects the exchange in the other (in terms of frequency and value). Thus, in an exchange network, the specific distribution of ties can have a strong impact on the distribution of benefits, but this influence depends on the nature of the network connections. For example, if there is a rule that one may exchange only with one other partner then the network is negatively connected. Division of labor creates positively connected networks in which exchange with one stimulates exchange with the other(s). Negative connections are easiest to trace both theoretically and empirically, and thus the bulk of research in this area is on negatively connected networks (an important exception is Burke (1997), who has developed an interesting link between network exchange and identities).

One prominent version of such a theory is *exchange network theory* (Cook and Yamagishi 1992). The distribution of benefits (or rates of exchange) within a network is explained on the basis of power ('power/dependence'). In positively connected networks, the power process is determined by brokerage, i.e., by the

ability to connect resources of various actors, and by the value of the focal resource to the actors. In a negatively connected network, power is based on the value and degree of availability of a focal resource from alternative sources. The more value Ego gets from an exchange with Alter compared to the best available exchange alternative, the more dependent Ego is on Alter and the more power Alter has over Ego. With repeated exchange in such negatively connected networks, the exchange ratio is predicted to shift toward the point at which neither of them could improve his or her outcome by exchange with another partner. Power advantages are thus structurally induced, given the 'connectedness' of the network.

A variant of this theory for coercive power has been elaborated by Molm (1997). Interestingly, Molm predicts and finds that power based on the ability to inflict damage on others is not structurally but strategically induced in order to extract rewards from the other. This opens exchange theory to a more explicit consideration of rational strategies, especially with regard to the balance between incentive to coerce and fear of retaliation. Lawler's work is also relevant to this topic (see, e.g., Lawler 1992a).

Elementary theory (ET, sometimes also called 'network exchange theory') is also rooted in the Homans/Emerson tradition and it also explains exchange ratios on the basis of structurally induced power (see Berger and Zelditch 1993, pp. 323). The major difference with exchange network theory is twofold. First, in reference to Weber, ET considers resistance to the exercise of power more explicitly than exchange network theory. Second, ET uses experiments in which the degree of 'exclusion' (i.e., the degree of negative connectiveness) varies. This second possibility gives it more flexibility to approximate various real life phenomena. However, the way resistance and the representation of exchange are worked out imposes strong restrictions, which reduces real life application to some degree (see Van Assen 2001). More recently, interesting links of ET with expectation states theory (Willer et al. 1997) and with game theory (see Willer and Skvoretz 1997) have been developed.

An additional grouping of theories on exchange should be mentioned. It is also related to the Homans/Emerson tradtion but uses different elaborations. There have been long-standing efforts to explain structural network effects on power and exchange rates by means of game theory (especially using the concept of the 'core') by Bonacich and co-workers (see, e.g., Bonacich 1998). See Friedkin (1995) for yet a different model and Fararo and Skvoretz (see Berger and Zelditch 1993, pp. 416) for a comparison of various models. There is also a productive development of models that use exchange in networks for the explanation of power processes in collective decision making, mostly in small groups (see Stokman et al. 2000). Because these models are not tied to a particular experimental paradigm, they are more open to the consideration of other exchange models and strategies (such as Coleman's model; see Coleman 1990) and to an explicit consideration of strategies of persuasion and challenges that accompany exchange strategies.

2.2.3 Friendship and cognitive networks. In recent years, dynamic friendship networks have received an increasing amount of attention. A new line of research has opened up by taking relatively simple behavioral assumptions and tracing their structure-generating consequences through simulation and, at times, also empirical research. What is new is that the interdependence of simultaneous friendship choices is modeled. If choices go on simultaneously in a group of strangers, what structure will emerge from the interdependence of choices under various choice preferences and constraints? How does 'group' size affect the resulting structure(s)? How does heterogeneity of the choosing individuals affect the resulting structure(s)? Preferences have been made more complex by stepwise adding status preferences to similarity preferences, and by explicitly introducing friendship dissolution to friendship formation. As to influence, friends do not just have a preference for similarity, they also become more similar. Paradigmatic for this kind of work is Stokman and Zeggelink (1996) and work presented in Doreian and Stokman (1997). See Van der Bunt (1999) for empirical research. Work linking this kind of dynamic approach to issues of cooperation is exemplified by Flache and Macy (1996) and Flache and Hegselmann (1999).

Networks are also cognitively represented which, in turn, contributes to the generation of joint cognitive and social structures. Research dealing with these aspects of networks link explicitly cognitive and structural interdependencies (see Carley 1986, Freeman 1994, Krackhardt 1987, Pattison 1994). Related research deals with the effects of cognitive interdependencies on learning, search, and general problem-solving strategies in groups and organizations (see Carley 1995, Edmonds and Dautenhahn 1999). Here functional interdependencies in terms of joint production are added explicitly (see also Troyer and Silver 1999). In the future, one can expect innovative impulses from these lines of research for sociological group research in general.

3. *Evaluation*

Sociological research on groups is lively, productive and, in the experimental fields, also cumulative. There is also a slow tendency to open up across the narrow borders of the research paradigms. However, there are a number of important shortcomings (see Lindenberg 1997 for more detail and for suggestions for new directions). First, field and experimental studies generally ignore each other. This is not a

desirable state. Crossovers in the past have been very productive and should be encouraged. Second, the experimental paradigms carry such a strong weight that they hamper new theory formation. For example, in expectation states theories, the functional interdependence is experimentally restricted to outcome interdependence with regard to agreement. There is little or no task interdependence. In the exchange theories, exchanges are often limited to divisions of resource pools. Such exchanges lack many of the properties of the exchange of goods, information, and emotions. On the whole, there is little attention to the effects of joint production on expectations and exchange.

Third, functional interdependencies should have analytical primacy. This primacy derives from the fact that cognitive processes are strongly influenced by goals (see Gollwitzer and Moskowitz 1996) and structural effects only occur in interaction with goals (externalities) as has been seen in exchange network theory. Yet, in most exchange theories this primacy is lacking. Fourth, present research rarely includes 'groupishness' (i.e., categorization and framing effects) in its consideration of cognitive interdependencies (see Lawler 1992b as an exception). For example, in expectation states theory, 'collective orientation' is introduced as a scope condition, not as a variable. Related to this point is, fifth, that the distinction between small groups and small networks is nonexistent or ad hoc. There is little or no theory on the influence of groups on networks or vice versa, or on networks nested in groups. Sixth, the study of intergroup relations for the explanation of what goes on inside the group needs much more attention. A positive example of what is needed is Coleman's classic study of the influence of interscholastic athletic competition on the local status structure in schools (see Coleman 1961).

See also: Collective Behavior, Sociology of; Exchange: Social; Networks: Social; Status and Role, Social Psychology of; Status and Role: Structural Aspects

Bibliography

Anderson E 1978 *A Place on the Corner*. University of Chicago Press, Chicago

Berger J, Ridgeway C L, Fisek M H, Norman R 1998 The legitimation and delegitimation of power and prestige orders. *American Sociological Review* **63**: 379–405

Berger J, Zelditch M (eds.) 1985 *Status, Rewards, and Influence*. Jossey-Bass, San Francisco

Berger J, Zelditch M (eds.) 1993 *Theoretical Research Programs*. Stanford University Press, Stanford, CA

Blau P 1964 *Exchange and Power in Social Life*. Wiley, New York

Bonacich P 1998 A behavioral foundation for a structural theory of power in exchange networks *Social Psychology Quarterly* **61**: 185–98

Burke P J 1997 An identity model for network exchange. *American Sociological Review* **62**: 134–60

Burt R S 1992 *Structural Holes. The Social Structure of Competition*. Harvard University Press, Cambridge, MA

Cahill S E, Lofland H (eds.) 1994 *The Community of Streets*. JAI Press, Greenwich, CT

Carley K 1986 An approach for relating social structure to cognitive structure. *Journal of Mathematical Sociology* **12**: 137–89

Carley K 1995 Computational and mathematical organization theory: Perspectives and directions. *Computational and Mathematical Organization Theory* **1**: 39–56

Coleman J S 1961 *The Adolescent Society*. Free Press, Glencoe, IL

Coleman J S 1990 *Foundation of Social Theory*. Harvard University Press, Cambridge, MA

Cook K S, Fine G A, House J S (eds.) 1995 *Sociological Perspectives on Social Psychology*. Allyn and Bacon, Boston

Cook K S, Yamagishi T 1992 Power in exchange networks: A power-dependence formulation. *Social Networks* **14**: 245–66

Doreian P, Stokman F N 1997 *Evolution of Social Networks*. Gordon and Breach, Amsterdam

Edmonds B, Dautenhahn K (eds.) 1999 Special issue on social intelligence. *Computational and Mathematical Organization Theory* **5**

Emerson R M 1976 Social exchange theory. *Annual Review of Sociology* **2**: 335–62

Fine G A 1987 *With the Boys*. University of Chicago Press, Chicago

Flache A, Hegselmann R 1999 Rationality vs. learning in the evolution of solidarity networks. *Computational and Mathematical Organization Theory* **5**: 97–127

Flache A, Macy M 1997 The weakness of strong ties: Collective action failure in a highly cohesive group. *Journal of Mathematical Sociology* **21**: 3–28

Flick U (ed.) 1998 *The Psychology of the Social*. Cambridge University Press, Cambridge, UK

Foschi M, Lawler E J (eds.) 1994 *Group Processes: Sociological Analyses*. Nelson-Hall, Chicago

Freeman L C 1994 Interpersonal proximity in social and cognitive space. *Social Cognition* **12**(3): 223–47

Friedkin N E 1995 The incidence of exchange networks. *Social Psychology Quarterly* **58**: 213–22

Goffman I 1961 *Encounters*. Bobbs-Merrill, Indianapolis, IN

Gollwitzer P M, Moskowitz G B 1996 Goal effects on action and cognition. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology. Handbook of Basic Principles*. The Guilford Press, London, pp. 361–99

Homans G C 1950 *The Human Group*. Harcourt, Brace & World, New York

Homans G C 1958 Social behavior as exchange. *American Journal of Sociology* **63**: 597–606

Homans G C 1961 *Social Behavior: Its Elementary Forms*. Harcourt, Brace & World, New York

Krackhardt D 1987 Cognitive social structures. *Social Networks* **9**: 109–34

Lawler E J 1992a Power processes in bargaining. *The Sociological Quarterly* **33**: 17–34

Lawler E J 1992b Affective attachments to nested groups. *American Sociological Review* **57**: 327–39

Lindenberg S 1997 Grounding groups in theory: Functional, cognitive, and structural interdependencies. *Advances in Group Processes* **14**: 281–331

MacLeod J 1995 *Ain't No Making It*. Westview, Boulder, CO
Molm L 1997 *Coercive Power in Social Exchange*. Cambridge University Press, Cambridge, UK
Moore J C 1985 Role enactment and self-identity. In: Berger J, Zelditch M (eds.) *Status, Rewards, and Influence*. Jossey-Bass, San Francisco, pp. 262–316
Pattison P 1994 Social cognition in context: Some applications of network analysis. In: Wasserman S, Galaskiewicz J (eds.) *Advances in Social Network Analysis*. Sage, Thousand Oaks, CA, pp. 79–109
Ridgeway C L, Heger Boyle E, Kuipers K J, Robinson D T 1998 Resources and interaction in the development of status beliefs. *American Sociological Review* **63**: 331–50
Stokman F N, Van Assen M, Van Der Knoop J, Van Oosten R C H 2000 Strategic decision making. *Advances in Group Processes* **17**: 131–53
Stokman F N, Zeggelink E P H 1996 'Self-organizing' friendship networks. In: Liebrand W B G, Messick D M (eds.) *Frontiers of Social Dilemma Research*. Springer, New York, pp. 385–418
Tajfel H 1982 Instrumentality, identity and social comparisons. In: Tajfel H (ed.) *Social Identity and Intergroup Relations*. Cambridge University Press, Cambridge, UK, pp. 483–507
Thibaut J, Kelley H H 1959 *The Social Psychology of Groups*. Wiley, New York
Troyer L, Silver S D 1999 Institutional logics and group environments. *Advances in Group Processes* **16**: 219–52
Turner J C 1982 Toward a cognitive redefinition of the social group. In: Tajfel H (ed.) *Social Identity and Intergroup Relations*. Cambridge University Press, Cambridge, UK, pp. 15–40
Van Assen M 2001 Two representations of negotiated exchange: A review and comparison. In: Van Assen M (ed.) *Esssays on Actor Models in Exchange Networks and Social Dilemmas*. Thela Thesis, Amsterdam
Van der Bunt G 1999 *Friends by Choice: An Actor Oriented Statistical Network Model for Friendship Networks Through Time*. Thela Thesis, Amsterdam
Webster M, Hysom S J 1998 Creating status characteristics *American Sociological Review* **63**: 351–78
White H C, Boorman S A, Breiger R L 1976 Social structure from multiple networks: I. Blockmodels of roles and positions. *American Journal of Sociology* **81**: 730–80
Willer D, Lovaglia M J, Markovsky B 1997 Power and influence: A theoretical bridge. *Social Forces* **76**: 571–603
Willer D, Skvoretz J 1997 Games and structures. *Rationality and Society* **9**: 5–35

S. Lindenberg

Growth Curve Analysis

The phrase 'growth curve analysis' denotes the processes of describing, testing hypotheses and making scientific inferences about the growth and change patterns in a wide range of time-related phenomena. The term 'growth curve' was used originally to describe a graphic display of physical stature (e.g., the height or weight) of an individual over consecutive ages. In contrast to some longitudinal data, growth curves have unique features: (a) the same entities are repeatedly observed, (b) the same procedures of measurement and scaling of observations are used, and (c) the timing of the observations is known. Formal models for the analysis of growth curves were developed in many different substantive domains with a common goal—to examine and uncover a fundamental set of regularity conditions or 'basic functions' responsible for the manifest growth and change. Techniques for the analysis of growth curves were initiated in the physical sciences, more fully developed in the biological sciences, and then used in studies of the size and health of plants, animals, and humans. In the behavioral sciences, growth curve analyses have been applied routinely to a wide range of phenomena—from experimental learning curves, to the growth and decline of intellectual abilities and academic achievements, and to changes in other psychological traits over the life span.

Growth curve analyses are among the most widely studied and well-developed mathematical and statistical techniques in all scientific research. These analyses have roots in the seventeenth- and eighteenth-century calculus of Newton and probability of Pascal, but this article is limited to more recent historical developments and other related techniques, such as time-series and dynamical systems analyses are not covered.

1. Growth Curve Data Collections

The first reported growth curve data were collected by the French Count de Montbeillard between 1759 and 1777. These data consist of semi-annual measurements on the growth of the height of his son over nearly 18

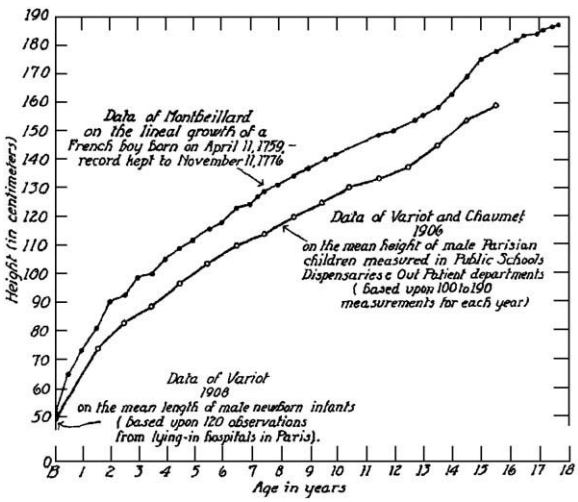

Figure 1
The initial 'growth curves' of human height data from Scammon (1927, p. 334). Note: The vertical (y) axis represents the height in centimeters and the horizontal (x) axis represents the age in years from birth

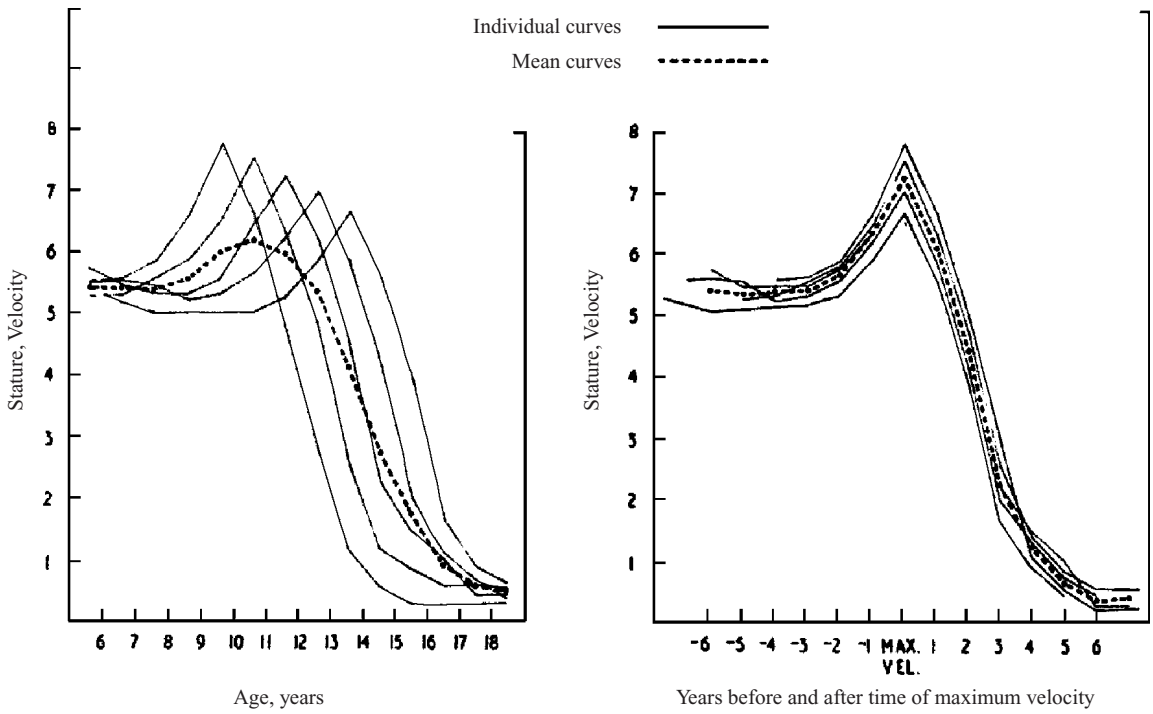

Figure 2
Alternative 'velocity curves' of physical stature from Tanner (1960, p. 22). Note: the left-hand graph shows the change in height as a function of the change in time (y) versus the age in years (x), while the right-hand graph shows the same curves (y) plotted around the time of maximum change in scores (x) for each individual

years. The data were plotted by Scammon (1927) and are presented here in the upper curve of Fig. 1. As Scammon reported

> it will be noted that the curve shows the typical four phases which most modern students have observed in the postnatal growth in stature of man, and which are characteristic of the growth of so many parts of the body (p. 331).

The first analysis of these data was by the naturalist Buffon in 1799 and

> should be given full credit for the discovery of seasonal differences in growth a full hundred years before the modern investigation of this work (p. 334).

The lower growth curve in Fig. 1 is based on group averages obtained by Variot in 1908 and

> it is interesting to note that, while the absolute values of the two series are quite different, the general form of the curve is essentially the same in both instances (p. 335).

These early growth curves were the precursor to an enormous collection of biological data on growth and change. A more recent illustration comes from the important work of Tanner and his colleagues (see Tanner 1960). In the individual plots of height in Fig. 2: (a) the 'velocity' or rate-of-change is plotted at each

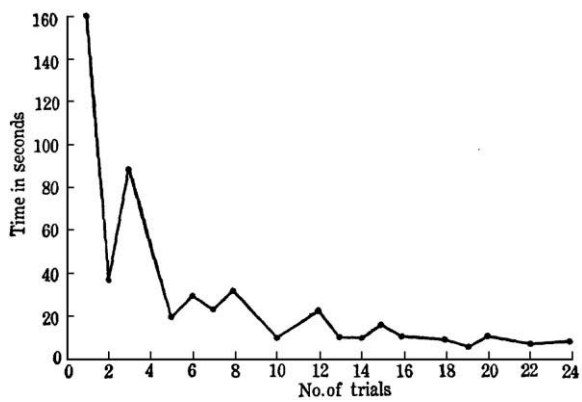

Figure 3
The 'learning curve' of a cat from Thorndike (1911) as reported by Garrett (1951, p. 44). Note: the response time (y) required for a cat to escape from a closed box by pulling on a wire as a function of consecutive number of trials (x)

age and (b) these curves are plotted against their own 'highest peak velocity.' This display demonstrated two interesting features of physical growth: (a) persons

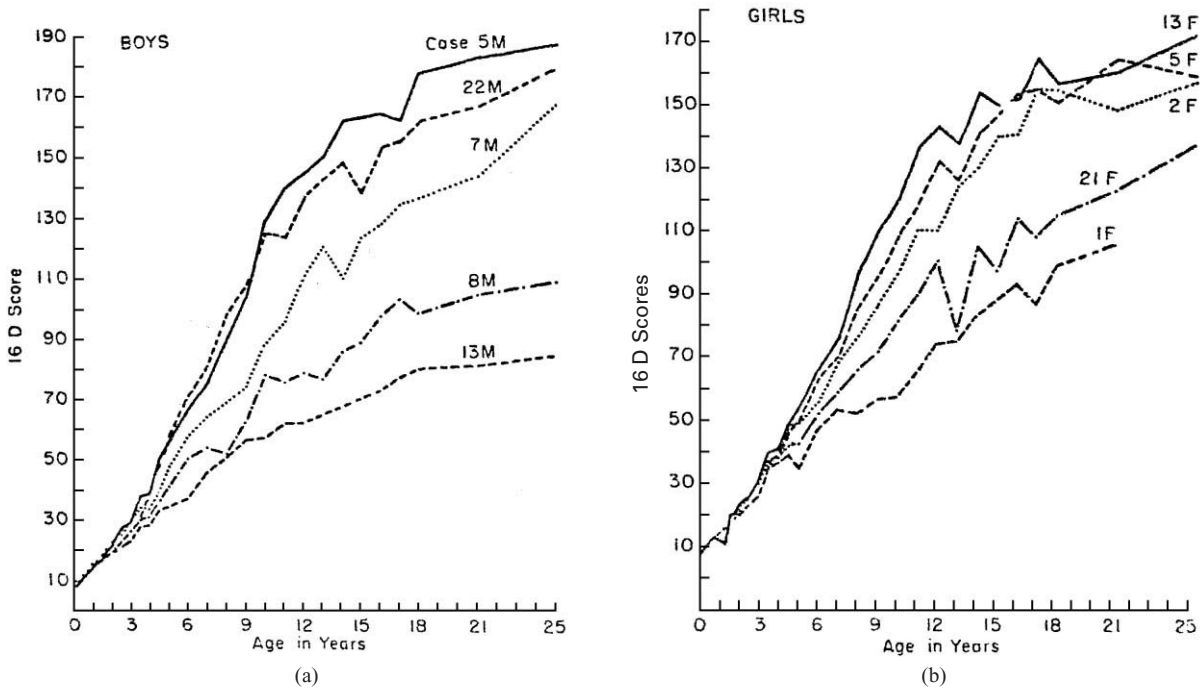

Figure 4
Growth curves of intellectual abilities in selected boys and girls from Bayley (1956, p. 67). Note: Age 16 D scores (y) plotted as a function of age at measurement (x) for (a) five boys and (b) five girls

who start growth at the earliest ages also grow the most but (b) all individuals share a remarkably similar shape in their 'adolescent growth spurt.' Extensions of these basic findings on physical growth systems and the contrast of chronological time vs. biological time, are still used in contemporary research.

Experimental psychologists have routinely collected different kinds of growth curves. Among the first here were the classical 'forgetting' curves collected by Ebbinghaus in the 1880s, whose work introduced the use of quantitative methods in the study of learning and memory and stimulated many experimental data collections. Included here are the 'curves of improvement' in studies on telegraphic language by Bryan and Harter in 1897. These principles were also used in the animal learning curve experiments of Thorndike in 1911, and one example of such data appears in Fig. 3. This plot illustrates an example of 'trial and error' learning: learning was objectively defined by decreasing response time and the lack of smooth function over trials was considered as error. Thorndike used these growth curves to illustrate several classical principles of learning, including the 'law of exercise' and the 'law of effect' (see Garrett 1951).

Differential psychologists have also contributed growth data in many different substantive areas. One good example of this tradition is given in the plots of Fig. 4 from work of Bayley (1956). Individual growth curves of mental abilities from birth to age 25 are plotted for a selected set of boys and girls from the well-known Berkeley Growth Study. Because mental ability was not easily measured in exactly the same way at each age, these individual curves were created by adjusting the means and standard deviations of different mental ability tests (i.e., Stanford-Binet, Terman-Merrill) at different ages into a common metric. As Bayley says

> they are not in 'absolute' units, but they do give a general picture of growth relative to the status of this group at 16 years … These curves, too, are less regular than the height curves, but perhaps no less regular than the weight curves. One gets the impression both of differences in rates of maturing and of differences in inherent capacity (1956, p. 66).

This use of 'linked' measurement scales created a novel set of growth data and permitted the use of growth analyses initially derived in other scientific areas.

2. *Linear Models for Growth Curve Analyses*

The simplest growth models are based on the simplest mathematical functions. The fitting of a straight-line to a set of measurements, a standard procedure in scientific research, was an initial model used by

Wishart (1938) in a classic paper on growth analysis. These linear models were quite practical because the calculations used standard statistical assumptions (e.g., independence and normality of the residuals), standard forms of estimation (e.g., Gauss-Markov least squares, maximum likelihood), and standard statistical tests of goodness-of-fit (F-tests, chi-square tests, and likelihood ratio tests).

Another key idea was also introduced by Wishart (1938) when he showed how power polynomials could be used to fit the nonlinear curvatures apparent in growth data. The individual growth curve (consisting of $t = 1, T$ occasions) is summarized into a small set of linear orthogonal polynomial coefficients based on a power-series of time ($t, t^2, t^3, \ldots t^p$) describing the general nonlinear shape of the growth curve. A model of growth data might require a second-order (quadratic), third-order (cubic), or even higher-order polynomial model fitted to the data, but the basic shape of each individual curve could be captured with a small number of parameters.

The organization of individual differences is an important consideration in any contemporary growth model. In the linear model, each individual is assumed to be following a straight line over time, describable by an intercept term and a slope term. The average of these individual scores can be used to describe the overall growth curve for the group, and the individual differences can be considered in the context of other substantive questions, e.g., do the groups under observation differ in intercepts or slopes? In the more complex polynomial models, Wishart suggested a second stage of analysis where these polynomial coefficients describing each individual curve were considered as the underlying scores that co-vary among different individuals. In a quadratic polynomial, for example, a person's trajectory over time could be summarized using an individual intercept, linear slope, acceleration, and a residual term, and secondary questions could be addressed, e.g., do the groups under observation also differ in acceleration?

This 'curve-fitting' or 'two-stage' approach was relatively easy to implement using a standard statistical basis for hypothesis testing (e.g., Bock 1975). The simplicity of this approach allowed linear growth models to enjoy widespread use in behavioral studies and become a useful starting point for understanding individual differences in change (e.g., Rogosa and Willett 1985). The persistent critiques of this polynomial approach include some key issues:

(a) the individual parameters can be difficult to estimate, especially for higher-order terms and incomplete data,

(b) even high-order polynomials (i.e., $p > 4$) cannot capture 'seasonality' or 'phase-shifts,'

(c) the substantive meaning of these individual parameters is often in doubt, and

(d) the individual level parameters can be transformed (i.e., rotated) to yield alternative meaning.

These standard critiques are appropriate, and they still apply to many other contemporary forms of growth curve analyses.

More complex linear models have been used to represent growth. A linear model applied to different age segments but with connected critical 'knot points' permitted a 'conjoined' or linear or polynomial 'spline' model (e.g., Bryk and Raudenbush 1992). This kind of spline model is not often used, partly because it is atheoretical, and partly because it may requires a relatively large number of parameters to achieve adequate fit. A more parsimonious organization of individual differences was proposed in the form of 'latent curves,' independently defined by both Rao and Tucker in 1958 (cited in Meredith and Tisak 1990). Principal components of the repeated observations were used to reorganize these data into the sum of a small number of unspecified linear functions are fitted to growth data, the shape of the latent curves is determined by the component loadings, and the individual curve parameters are the component scores. The summation of latent curves has roots in classical nineteenth-century work of Fourier as now widely used. In a recent innovation, Meredith and Tisak (1990) showed how these 'Tuckerized curve' models could be represented using confirmatory common factor analysis models, used to represent a wide range of alternative models (including most of the models discussed herein), and fitted using structural equation modeling techniques (also see McArdle 1988, McArdle and Bell 2000, McArdle and Woodcock 1997).

3. *Nonlinear Models for Growth Curve Analysis*

The linear models can be used to describe a variety of nonlinear shapes, but many growth curve models for specific substantive problems have included nonlinear functions of the parameters. For example, Ebbinghaus in 1880 described his forgetting curves using a form of the classic 'exponential growth' model. Here the rate of change is a linear function of the percentage of initial size (e.g., compound interest), and individual trajectories rise or fall in nonlinear fashion towards some asymptotic values. In contrast, the nineteenth-century Velhurst curve of population growth, an S-shaped 'logistic curve,' was used in the early twentieth century by Pearl for many forms of growth. At about the same time Peters advocated an 'ogival' curve of growth in ideational learning, Thurstone found that a hyperbolic curve of learning best to fit the norms of many different tests, and both Ettlinger and Valentine demonstrated direct relationships among these functions (see Seber and Wild 1989, Bock and Thissen 1980).

More complex nonlinear models have been used extensively. In about 1825 Gompertz described the growth curve in terms of two exponential accumulations of different rates towards different asymptotes.

This flexible model was studied in the first half of the twentieth century by Winsor, used by Medwar to study the growth of chicken hearts, and by Deming for human physical growth, and in studies of learning curves (see Browne and du Toit 1991). Another popular nonlinear model, introduced around 1938 by von Bertalanffy, suggested that an individual's change in weight was the direct result of the difference in the forces of anabolism and catabolism, with the exact relationship specified by known empirical properties. In subsequent work, Richards challenged the empirical basis of this model, and formally showed how many prior models can be seen as specific solutions of a 'family' of deterministic differential equations; specific restrictions led to the exponential, logistic, Gompertz, and von Bertalanffy equations. This work was more recently extended by Nelder and Sandland and MacGhilcrest (1979). (See also Seber and Wild 1989, Zeger and Harlow 1987.)

Attempts to fit a single growth model to observations over a wide range of ages with a minimal set of parameters led researchers to combine aspects of other models. Jenss and Bayley in 1937 proposed a simple composite where a linear part was used to fit the early rapid growth phase and this was added to an exponential part to fit negative acceleration of the later slowing-down phase. A more complex composite model, the sum of multiple logistic curves, was suggested by Robertson in 1908 and Burt in 1937 but only made practical and fully developed by Bock and his colleagues (see Bock and Thissen 1980). In each segment the rate of growth (a) increases early, (b) reaches a maximum (peak growth velocity), (c) decreases towards the asymptote, and (d) the final value of one segment is used as the starting value of the next segment. A novel composite model based on multiple logistic functions was developed by Preece and Baines in 1978 and here the model parameters were supposed to be linearly related to meaningful biological parameters. Simple versions of this new family of curves have proven useful in contemporary work (see Hauspie et al. 1991).

The mathematical elegance of these nonlinear formulations often make it seem that the parameters estimated are related directly to the underlying phenomena and there is a tendency to treat these estimates with direct substantive interpretation. Prior research has show this to be a difficult goal and now it is known that all models permit alternative formulations that may fit equally well. Another important property of growth curves comes when the fixed parameters of the nonlinear curves, the 'parameters of the means' also represents the 'means of the parameters' of the individual curves. However, classical research by Merrill in 1931, Estes in 1957, Tucker in 1966, and Keats in 1983 has shown this form of 'dynamic consistency' does not always hold, even for some of the most popular growth models (e.g., a group exponential). Contemporary statistical models are now constructd carefully to clarify the unique substantive interpretation of nonlinear parameters (see Browne and du Toit 1991, Bock and Thissen 1980).

4. *Multivariate Growth Curve Analyses*

The collection of multiple variables at each occasion of measurement leads naturally to questions about relationships among growth processes and multivariate growth models. In biological research, early efforts were directed at characterizing the parallel properties of different growth variables. Models were developed originally to deal with the size of two different organs, and early nineteenth-century work was used by Huxley to form a classical 'allometric' model where two variables have a constant ratio of growth rates throughout the growth period. Many physical processes were subsequently found to grow in parallel, or in an ordered time-sequence (e.g., Tanner 1960). This work led to more sophisticated models based on systems of differential equations for the size of multiple variables. In one comprehensive multivariate model, Turner in 1978 extended the simple growth principles to more variables and permitted an examination of biologically important interactions based on the size and sign of the estimated parameters (see Griffiths and Sandland 1984).

Multivariate research in the behavioral sciences has relied on advanced versions of the linear growth models formalized by Rao, and Pothoff and Roy (see Bock 1975). The structural models developed within this framework have added measurement models and this includes work on multivariate models of latent curves (McArdle 1988). These structural models have proven useful in testing hypotheses about parallel growth processes across time in multiple psychological variables. In recent examples, McArdle and Woodcock (1997) showed how latent growth of a single common factor of intellectual abilities is a poor representation of multiple abilities data and McArdle (2000) showed how a multivariate dynamic sequence could be fittted using standard structural equations. Current research in this area is focused on an improved understanding of the time-sequences and interactions among multiple behavioral and biological variables.

5. *Recent Issues in Growth Curve Analysis*

The prior developments in growth curve analysis represent an important but limited class of longitudinal data analyses (e.g., Nesselroade and Baltes 1979, Bock 1975, Pinherio and Bates 2000). These developments have been distinguished by features of the models themselves: (a) the degree of mathematical specification, (b) the way these statistical models are fitted, and (c) the clarity and substantive meaning of

the results. In future work, most of these problems are likely to remain under active investigation.

Most contemporary growth curve models can be written in a common symbolic form (Seber and Wild 1989). A general model for a change in the scores over time (often using derivatives dy/dt or differences $\Delta y/\Delta t$) can be based on some mathematical functional form ($f\{x\}$) with unobserved scores ($x[t]$) and with unknown parameters ($\beta[t]$) to be estimated. Additional statistical forms not discussed here can be included, including autoregressive residual error structures, Markov chains, and Poisson processes. The growth functions described here are not an exhaustive listing and future extensions to more generalized functions, and to dynamic and chaotic forms are likely.

Recent research in statistical model fitting is important here too. It is now clear that growth curve models of arbitrary complexity can be fitted to the observed trajectory over time (i.e., the integral) and the unknown parameters can be estimated to minimize some statistical function (e.g., weighted least squares, maximum likelihood), and parameter estimates obtained using nonlinear programming. This generic statistical approach avoids some problems in older techniques, such as fitting a model to a log-scale or directly to the velocity or the analysis of difference score data (Fig. 2). These new techniques make it possible to address the critical problems of forecasting future observations, and further research on Bayesian estimation is the focus of current research here (for details, see Seber and Wild 1989).

This generic model fitting approach also permits a wide range of new possibilities for dealing directly with unbalanced, incomplete or missing data. In early work, linear polynomials were used extensively to deal with these kinds of problems. But recent work on linear and nonlinear mixed and multilevel models have surfaced, making it possible to estimate growth curves and test hypotheses by collecting only small segments of data on each individual (Bryk and Raudenbush 1992, Pinherio and Bates 2000, Verbeke and Molenberghs 2000). These statistical models are being used in many longitudinal studies to deal with self-selection and subject attrition, multivariate changes in dynamic patterns of development, and the tradeoffs between statistical power and costs of person and variable sampling. The statistical power questions of the future are not 'how *man*y occasions do we need?' but 'how *few* occasions are adequate?' and 'which persons should we measure at which occasions?' (McArdle and Woodcock 1997, McArdle and Bell 2000).

Some of the most difficult problems for future work on growth curves are related to the elusive meaning of the growth model parameters. Growth curve analyses also rely on basic measurement properties including (a) interval scaling, (b) age-and-time equivalence, and (c) substantively meaningful. These issues are key problems in the behavioral sciences and future research is needed to address these fundamental concerns. Multivariate models may be useful here, especially as these models are further extended to a fully dynamic time-dependent form. In any case, empirical information and experimentation will be needed to judge the utility of any growth curve model.

Given the long history of elegant formulations from mathematics and statistics in this area, it is somewhat humbling to note that many of the most insightful growth curve analyses have come from careful visual inspection of the growth curves (e.g., in Figs. 1–4). The insight gained from visual inspection of a set of growth curves is not in dispute now and in fact obvious visual features should be highlighted and emphasized in future research (e.g., Pinherio and Bates 2000). As in the past, the best future growth curve analyses are likely to be the ones we can see most clearly.

See also: Hierarchical Models: Random and Fixed Effects; Latent Structure and Casual Variables; Learning Curve, The; Longitudinal Data; Statistical Analysis: Multilevel Methods; Time Series: General

Bibliography

Bayley N 1956 Individual patterns of development. *Child Development* **27**(1): 45–74

Bock R D 1975 *Multivariate Statistical Methods in Behavioral Research*. McGraw-Hill, New York

Bock R D, Thissen D 1980 Statistical problems of fitting individual growth curves. In: Johnston F E, Roche A F, Susanne C (eds.) *Human Physical Growth and Maturation: Methodologies and Factors*. Plenum, New York, pp. 265–290

Browne M, Du Toit S H C 1991 Models for learning data. In: Collins L, Horn J L (eds.) *Best Methods for the Analysis of Change*. APA Press, Washington, DC, pp. 47–68

Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA

Fischer G H, Molenaar I W 1995 *Rasch Models: Foundations, Recent Developments, and Applications*. Springer, New York

Garrett H E 1951 *Great Experiments in Psychology*. Appleton-Century-Crofts, New York

Griffiths D, Sandland R 1984 Fitting generalized allometric models to multivariate growth data. *Biometrics* **40**: 139–150

Hauspie R C, Lindgren G W, Tanner J M, Chrzastek-Spruch H 1991 Modeling individual and average growth data from childhood to adulthood. In: Magnusson D, Bergman L R, Rudinger G, Törestad B (eds.) *Problems and Methods in Longitudinal Research: Stability and Change*. Cambridge University Press, Cambridge, GB

McArdle J J 1988 Dynamic but structural equation modeling of repeated measures data. In: Nesselroade J R, Cattell R B (eds.) *The Handbook of Multivariate Experimental Psychology, 2*. Plenum Press, New York, pp. 561–614

McArdle J J 2000 A latent difference score approach to longitudinal dynamic structural analyses. In: Cudeck R, du Toit S, Sorbom D (eds.) *Structural Equation Modeling: Present and Future*. Scientific Software International, Lincolnwood, IL, pp. 342–380

McArdle J J, Bell R Q 2000 Recent trends in modeling longitudinal data by latent growth curve methods. In: Little T D, Schnabel K U, Baumert J (eds.) *Modeling Longitudinal and Multiple-group Data: Practical Issues, Applied Approaches, and Scientific Examples*. Erlbaum, Mahwah, NJ, pp. 69–107
McArdle J J, Woodcock J R 1997 Expanding test-rest designs to include developmental time-lag components. *Psychological Methods* **2**(4): 403–35
Meredith W, Tisak J 1990 Latent curve analysis. *Psychometrika* **55**: 107–22
Nesselroade J R, Baltes P B (eds.) 1979 *Longitudinal Research in the Study of Behavior and Development*. Academic Press, New York
Pinherio J C, Bates D M 2000 *Mixed-effects Models in S and S-PLUS*. Springer, New York
Rogosa D, Willett J B 1985 Understanding correlates of change by modeling individual differences in growth. *Psychometrika* **50**: 203–228
Sandland R L, McGilchrist C A 1979 Stochastic growth curve analysis. *Biometrics* **35**: 255–271
Scammon R E 1927 The first seriatim study of human growth. *American Journal of Physical Anthropology* **10**: 329–36
Seber G A F, Wild C J 1989 *Nonlinear Models*. Wiley, New York
Tanner J M (ed.) 1960 *Human Growth*. Pergamon Press, New York
Verbeke G, Molenberghs G 2000 *Linear Mixed Models for Longitudinal Data*. Springer, New York
Wishart J 1938 Growth rate determinations in nutrition studies with the bacon pig, and their analyses. *Biometrika* **30**: 16–28
Zeger S L, Harlow S D 1987 Mathematical models from laws of growth to tools for biologic analysis: Fifty years of growth. *Growth* **51**: 1–21

J. J. McArdle

Guilt

Guilt has a number of related but distinguishable meanings. The word generally signifies having committed or being responsible for a wrongdoing, sin, or offence. This definition easily translates into the legal use of the word which means having committed a crime or tort as indicated by the offender's plea or the court's decision. A third usage of the word describes the feelings or emotion connected with recognition of having done wrong. Guilt, in other words, has a descriptive sense, a legal sense, and an affective sense.

1. *Descriptive and Legal Meanings of Guilt*

The Oxford English Dictionary traces the word guilt back to Old Teutonic words such as geld, meaning to pay, and Old English words (gylt, gielt, etc.) meaning delinquency, the fact of having committed an offence, being legally liable to a penalty, and a penalty for a crime. Modern definitions, such as are found in the seventh edition of the *Concise Oxford Dictionary*, focus more simply on having 'committed a specific or implied offence.' Thus, in its broadest sense guilt simply signals that someone is responsible for an action or omission that is considered wrong. Implicit to this is the assumption that the guilty person was causally involved, though it has been argued that the person's intentions were not considered important in many times and places (Wierzbicka 1999).

As is evident from the origin and definitions of guilt, the concept is strongly linked to legal, especially criminal frameworks. Within a legal setting, the word guilty is used either by the defendant in pleading to a charge, or by the jury in convicting (a guilty verdict), to indicate responsibility for having committed a crime or tort (see Lely 1902). It often also indicates, as a consequence, liability for some penalty (Walker 1980).

To be guilty of a crime, one needs to commit the *actus reus* (the wrongful act) and to have *mens rea* (a guilty mind). Intention to commit the *actus reus* is the paradigm guilty mind. But over the centuries a variety of defences against a finding of *mens rea* have been refined in the common law: coercion, duress, mistake, self defence, provocation, insanity, necessity, accident and automatism being the main ones. The most sweeping defence against criminal responsibility is infancy, even though the *actus reus* has been committed with intent. Children under eight in most nations cannot be found guilty of a crime and there are generally qualifications on the capacity to hold older children criminally responsible. This is a major distinction between legal guilt and descriptive and affective guilt (which children can experience or at least are learning to experience).

In some contexts, recklessness—a weaker standard than intention but a stronger one than negligence—will be found to justify criminal guilt. In others the criminal law will countenance strict liability—guilt without *mens rea*. Generally this is reserved for contexts in which proving a guilty mind is difficult or when public policy regards the act itself to be sufficiently dangerous. Proving beyond reasonable doubt that the possessor of the illicit item had any criminal intent in relation to it is difficult, so the law resorts to the expedient of the *actus reus* of possession as sufficient for guilt. A second set of circumstances where the criminal law sometimes allows guilt without *mens rea* is where the *actus reus* involves a very great danger to the community and where the criminal actor has a capacity to take extraordinary steps to protect the community from that danger. Manufacturing pharmaceutical products or operating nuclear power plants are examples of where such strict liability crimes are countenanced in many nations.

Increasingly the world's legal systems accept the idea that a collectivity, such as a corporation, can be guilty of a crime. The reading of corporate criminal guilt initially became more expansive in the US; it then spread throughout the common-law world and in the final decades of the twentieth century took hold in the

civil-law world. Corporate criminal guilt can be based on a notion of corporate intention, or it can be determined that the corporation is vicariously responsible for the guilty intention of one of its members. Or corporate guilt can be based on a combination of a harmful corporate act that was enabled by irresponsible corporate policies or procedures (the corporate equivalent of the individual mental element of recklessness) (Bucy 1991, French 1984, May 1987, Fisse and Braithwaite 1993). In some ways the law imposes higher standards on corporations which are able to implement corporate compliance systems than it does on individuals. Arguably this spectrum of expectations is necessary for a 'risk society' (Beck 1992) where large corporations can cause aircraft to crash and nuclear power plants to melt down. At some point during the twentieth century in all industrialized societies, most litigants in the higher courts were corporations, rather than individuals.

Prior to the Middle Ages, when the distinction between tort and crime evolved, guilt was routinely used in civil law. Whether one burnt down the house of another intentionally (arson) or accidentally, one would be guilty of the act. In most legal systems guilt came to be used almost exclusively to refer to criminal guilt. However, when a civil wrong involved some element of fault, it remained appropriate to refer to the defendant as guilty of that civil wrong. While guilt was decreasingly so used when the fault element was only negligence, in US law it remained more common than in other systems to use the term guilt when exemplary damages were set in civil cases. When treble damages are imposed in a tort case, the existence of a punitive as well as a compensatory element in the court order means that it is appropriate to speak of the wrongdoer as guilty of that wrong. In equity and contract, if the conduct of a defendant is 'unconscionable' or shows bad faith, we can refer to the person as being guilty of unconscionable conduct. Finally, in US regulatory law there is the widespread phenomenon, less common in other nations, of civil penalties which can be imposed on evidence of guilt on the balance of probabilities, as opposed to the criminal test of beyond reasonable doubt. Americans speak of such defendants as being guilty of a breach of the law, even though there may not be criminal intent and defendants may not be able to avail themselves of excuses which would be available to them were criminal charges to be laid. Today, in a world where most of the things that can be done for good or ill in the world are done by organizations, most such civil penalty defendants are corporations. (See *Remedies and Damages: Legal Aspects*.)

At an even more macro level than the corporation, international law debates about crimes against humanity, such as genocide, resonate today with concern about 'impunity' as the world moves to establish an International Criminal Court. In this literature collective guilt is often supplanted by the concept of collective shame as something nations might transcend through a process such as the South African Truth and Reconciliation Commission.

At both the macro level of national reconciliation and in the most micro context of juvenile justice and child protection, part of the impact of the growing social movement for restorative justice has been to shift the emphasis in debates around legal institutions from debates about guilt to debates about shame (Braithwaite 1999). We now turn to the affective foundations of this contest.

2. *Feelings of Guilt*

In addition to communicating both general and legal responsibility for wrongdoing, guilt has also come to be discussed as a feeling or emotion. At its core, a guilty feeling simply refers to the knowledge of one's guilt. However, the concept assumes not just awareness of wrongdoing but also discomfort with it. Thus, acceptance of having done wrong seems to be an important precondition of feeling guilty, although it is not when someone accepts or is assigned guilt in a descriptive or legal sense. The feeling of guilt is characterized by negative affect and may include behavioural changes. So, for example, it is sometimes said that someone is acting guilty or even that someone was consumed by guilt. Indeed, much classical literature, such as Shakespeare's *Macbeth* and Dostoevsky's *Crime and Punishment*, has followed the deterioration of characters haunted by past deeds. In such examples of guilt or guiltiness, the emotion is usually endured in secret, as the individual tries to hide his or her transgressions from the view of others. In fact, as will be seen, guilt is described by some theoretical traditions as a distinctively private emotion, especially in contrast to other related emotions such as shame.

Perhaps the most influential, or at least the most widely recognised, use of guilt in the social sciences has occurred within pschoanalytic theory. Freud (1949) argues that a primitive basis for guilt is the fear of losing another's love. In this primitive form feelings of guilt occur when one's transgression is discovered by an authority figure, or when the individual fears its discovery. This, according to psychoanalytic theory, is the only form of guilt young children can feel. However, individuals become able to feel what might be called true guilt with the formation of a super-ego or conscience, which internalizes the values that might have otherwise been enforced by an external authority. In this way individuals come to punish themselves through feeling bad (guilty) for transgressions or, because of the internal nature of the superego, even for offensive thoughts. Within this framework, conscious feelings of guilt, which Freud suggests should be called remorse, occur after committing a transgression. However, Freud also suggests that unconscious feelings of guilt play a much more malevolent role and are

the underlying cause of neuroticism. This is because the superego is characterized as taking on the aggressive and punishing characteristics of the external authority as well as the individual's own feelings of aggression towards the external authority. Ultimately the superego redirects this aggressiveness towards the ego, or conscious self. Ignoring the complexities raised by Freud's analysis of neuroticism and unconscious forms of guilt, what is most significant about this psychoanalytic perspective is that it describes guilt as an internally generated sanction based upon the individual's recognition of wrongdoing. For further detail on Freud's conception of guilt see his Civilisation and its Discontents (1949).

Guilt is also assigned an important role in the comparison of cultures by anthropologists such as Benedict (1946) and Mead (1937). In comparing the way societies maintain social control, a distinction is drawn between those that rely more heavily on internalized feelings of guilt and those that rely to a greater extent on social mechanisms of shame and shaming to maintain conformity. In Benedict's famous comparison, United States society is classified as a guilt culture, while Japanese society is classified as a shame culture. This distinction reflects Freud's conception of guilt as an internally generated sanction which develops as a result of socialization practices. However, it also highlights the alternative emotion of shame, which is described as a reaction to others' criticism and is similar to Freud's conception of 'primitive guilt.' This contrasting of shame and guilt as social versus internal control has been an influential one, with much of the recent research on guilt and shame examining the social vs. non-social dimension (Cheek and Hogan 1983, Gibbons 1990, Wallbott and Scherer 1995, Tangney et al. 1996). Indeed, as will be seen, a growing interest in shame has meant that research on both emotions has been dominated by an interest in distinguishing them.

While the social distinction between shame and guilt has been influential, it has also been questioned by a number of perspectives. Piers and Singer (1953) argue that there is much evidence from everyday life that individuals experience shame in private and without others' intervention. Building on the psychoanalytic framework they argue that shame and guilt are distinguishable because they result from different psychological tensions. Whereas guilt arises from tension between the ego and superego (conscience), shame is the product of tension between the ego and what they call the ego ideal. The ego ideal might be described as the individual's perception of their ideal self, those attributes and characteristics to which he or she aspires. This structure implies a new distinction: shame is about failure, it represents a shortcoming, whereas guilt is about transgression, it represents breaking a rule. Lewis (1971) argues that a second distinction implied by this psychoanalytic structure concerns the object of the two emotions. When feeling guilt the individual's attention is focused upon the act or omission itself, but because shame is about failure to achieve an ideal, it is focused upon a global evaluation of the whole self. Thus, in addition to the emphasis upon transgression evident in the legal, Freudian and early anthropological perspectives, this later work has argued that guilt is felt about specific actions rather than the self.

The distinctions between guilt and shame have become a significant focus of recent empirical research. While empirical studies show that subjects' memories of feeling guilt involve less self-conscious feelings than their memories of shame (Wicker et al. 1983, Tangney et al. 1996), evidence that guilt is less focused upon evaluation of the whole self is equivocal. While at least one study (Niedenthal et al. 1994) has suggested that guilt is more specific in focus, a number of other studies (Tangney et al. 1996, Wicker et al. 1983) have found no significant difference between the emotions on this dimension. Indeed, a recent study (Harris 1999) which clustered subjects' reports of feelings using factor analysis did not distinguish between shame and guilt in either dimension. These results suggest that shame and guilt may often be empirically indistinguishable.

The discussion highlights the importance that has been placed in recent time on distinguishing feelings of guilt and shame. It is equally apparent that the issue has not been resolved and continues to be an important focus of research. This is particularly evident in recent research which explores guilt proneness and shame proneness as personality dispositions that are related to individual coping mechanisms and psychopathologies such as depression (Tangney 1991, Harder 1995, Ferguson et al. 1999). Despite disagreement over the dimensionality of guilt, research has highlighted a number of characteristics of guilt: that it is felt as a result of the acceptance that one has committed a wrongdoing; that it is focused upon one's actions or omissions; and that it often involves the perception of having hurt others in some way (Kugler and Jones 1992).

See also: Confessions: Psychological and Forensic Aspects; Crime: Knowledge about and Prevalence; Criminal Defenses; Criminal Law and Crime Policy; Honor and Shame; Law: Defense of Insanity; Personality and Crime; Responsibility: Philosophical Aspects; Shame and the Social Bond

Bibliography

Beck U 1992 *Risk Society: Towards a New Modernity*. Sage, Newbury Park

Benedict R 1946 *The Chrysanthemum and the Sword; Patterns of Japanese Culture*. Houghton Mifflin, Boston

Braithwaite J 1999 Restorative justice: Assessing optimistic and pessimistic accounts. In: Tonry M (ed.) *Crime and Justice: A*

Review of Research. University of Chicago Press, Chicago, Vol. 25
Brett P 1963 *An Inquiry into Criminal Guilt*. Law Book Company, Sydney
Bucy P H 1991 Corporate ethos: A standard for imposing corporate criminal liability. *Minnesota Law Review* **75**: 1095–184
Cheek J M, Hogan R 1983 Self-concepts, self-presentations, and moral judgements. In: Suls, Greenwald A G (eds.) *Psychological Perspectives on the Self*. Lawrence Erlbaum Associates, Hillsdale New Jersey, Vol. 2
Ferguson T J, Stegge H, Miller E R, Olsen M E 1999 Guilt, shame and symptoms in children. *Developmental Psychology* **35**(2): 347–57
Fisse B, Braithwaite J 1993 *Corporations, Crime and Accountability*. Cambridge University Press, Cambridge
French P A 1984 *Collective and Corporate Responsibility*. Columbia University, New York
Freud S 1949 *Civilisation and its Discontents*. Hogarth Press and Institute of Psycho-Analysis, London
Gibbons F X 1990 The evolution and manifestation of social anxiety. In: Crozier W R (ed.) *Shyness and Embarrassment: Perspectives from Social Psychology*. Cambridge University Press, Cambridge
Harder D W 1995 Shame and guilt assessment, and relationships of shame- and guilt-proneness to psychopathology. In: Tangney J P, Fischer K W (eds.) *Self Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford Press, New York, Chap. 17, pp. 368–92
Harris N 1999 *Shaming and shame: An empirical analysis*. PhD thesis, Australian National University, Canberra
Kugler K, Jones W H 1992 On conceptualizing and assessing guilt. *Journal of Personality and Social Psychology* **62**(2): 18–27
Lely J M 1902 *Wharton's Law-Lexicon* (10th edn.). Stevens, London
Lewis H B 1971 *Shame and Guilt in Neurosis*. International Universities Press, New York
May L 1987 *The Morality of Groups: Collective Responsibility, Group-based Harm and Corporate Rights*. University of Notre Dame Press, Notre Dame, IN
Mead M 1937 *Cooperation and Competition Among Primative Peoples*. McGraw-Hill, New York
Niedenthal P M, Tangney J P, Gavanski I 1994 'If only I weren't' versus 'If only I hadn't': Distinguishing shame and guilt in counterfactual thinking. *Journal of Personality and Social Psychology* **67**(4): 585–95
Piers G, Singer M B 1953 *Shame and Guilt: A Psychoanalytic and a Cultural Study*. Norton, New York
Tangney J P 1991 Moral affect: The good, the bad, and the ugly. *Journal of Personality and Social Psychology* **61**(4): 598–607
Tangney J P, Miller R S, Flicker L, Barlow D H 1996 Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology* **70**(6): 1256–69
Walker D M 1980 *The Oxford Companion to Law*. Clarendon Press, Oxford
Wallbott H G, Scherer K R 1995 Cultural determinants in experiencing shame and guilt. In: Tangney J P, Fischer K W (eds.) *Self Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford, New York, pp. 465–87
Wicker F W, Payne G C, Morgan R D 1983 Participant descriptions of guilt and shame. *Motivation and Emotion* **7**: 25–39
Wierzbicka A 1999 *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge University Press, Cambridge

N. Harris and J. Braithwaite

Gynecological Health: Psychosocial Aspects

Most women do not experience menstrually related changes that interfere with their lives to any appreciable extent. Where such changes are identifiable, they do not have a simple biological basis but are embedded in a complex social and cultural context.

1. *Menstrually Related Distress*

A comprehensive review (Klebanov and Ruble 1994) indicated no relationship between subjective distress and any hormonal measure; there is no consensus on its definition, symptoms, prevalence, effects, or indeed the component of the menstrual cycle (premenstrual, menstrual, or intermenstrual) with which the distress is associated.

Although mood, physical symptoms, and coping with life are claimed to demonstrate menstrually related cyclicity, the empirical evidence is mixed and inconclusive. Decades of research provide evidence for no cyclic variation in any aspect of cognitive functioning, memory, or motor performance (Richardson 1992). Women's reports of symptoms and mood swings are no different from those experienced by men (McFarlane and Williams 1994), and menstrually-related fluctuations in mood are considerably less than those associated with days of the week.

Prevalence estimates of menstrually related distress range from 95 to 5 percent, illustrating a complete lack of definitional consensus. The main standardized self-report survey, the Moos Menstrual Distress Questionnaire, obtains prevalence rates between 55 and 70 percent for physical discomfort, and between 23 and 70 percent for negative affect, with the majority of symptoms extremely mild and causing negligible distress (Logue and Moos 1986).

The American Psychiatric Association included a preliminary definition of premenstrual dysphoric disorder (PDD) in DSM-IV. PDD involves 'symptoms such as markedly depressed mood, marked anxiety, marked affective lability, and decreased interest in activity … during the last week of the luteal phase in most menstrual cycles during the last year. The symptoms begin to remit within a few days of the onset of menses and are always absent in the week following menses.' PDD is diagnosed if the condition is com-

parable in severity to a major depressive episode, involving marked impairment in either social or occupational functioning. PDD must be confirmed with prospective daily ratings over at least two cycles, since retrospective reports are unreliable and there is little consistency across consecutive cycles. Rivera-Tovar and Frank (1990), using a prospective, blind, study over 90 days, found that 5 percent of US students met criteria for PDD. This suggests that some women experience menstrually related distress, but that it is not the norm, and it should be noted that experiencing distress on a cyclic basis does not necessarily mean that it is hormonally mediated.

2. *Postnatal Distress*

It is often assumed that negative psychological outcomes of childbirth result from hormonal changes, but three main categories of childbirth-related psychological distress with different causes and prognoses need to be distinguished.

2.1 *Baby Blues*

Postpartum dysphoria, or 'baby blues,' tearfulness, mood swings, and irritability, occurs in a high proportion of mothers in the few days following childbirth, and normally resolves itself within 10 days. It is not related to any environmental, social, or cultural factor (Kumar 1994), and may be related to changes in estrogen levels, although individuals undergoing surgery experience similar symptoms, suggesting it may be a reaction to physical stress. Prevalence estimates vary from 26 to 85 percent of all mothers. The condition is distressing, especially if women are not expecting it, but resolves quickly and has little long-term impact.

2.2 *Postpartum Psychosis*

Postpartum psychosis, characterized by hallucinations, delusions, and other schizophrenia-like symptoms, occurs in all societies, at a relatively consistent rate of 0.01 to 0.02 percent of all births (Kumar 1994), and can result in suicide or infanticide. Previous psychotic illness is the strongest risk factor for postpartum psychosis; there is no link with life events, stress, or relationship difficulties, supporting the view that it arises, like other psychotic states, in biologically susceptible individuals as a reaction to substantial life change.

2.3 *Postpartum Depression*

Nonpsychotic postpartum depression is mainly influenced by psychosocial factors, and there is no convincing evidence for hormonal causation (Kumar 1994). Postpartum depression is predicted by relationship difficulties, low social support, previous depression, and stress. It is identical in presentation to other major depression, and while DSM-IV allows the specifier 'with postpartum onset' to be applied to depression which occurs within four weeks of the delivery of a child, it has no separate diagnostic category for postpartum depression.

Rates of depression in new mothers, assessed by the Edinburgh Postnatal Depression Scale (Cox et al. 1987), are between 6 and 10 percent. Symptoms may begin at any time up to 12 months after the birth of the baby and may be preceded by normal adaptation. High rates of depression persist as long as the mother has young children at home, and fathers also experience postnatal depression. Looking after a baby is stressful and anxiety provoking, and fatigue and disturbed sleep add to the difficulty. Depression in the postpartum period is best treated as reactive depression, reflecting a need for social and practical support, rather than as unique to childbirth.

3. *Fertility-related Problems*

3.1 *Infertility*

Infertility is generally defined as the failure to achieve a pregnancy after 12 months of unprotected sexual intercourse. Research on failure to conceive focuses on women, even though the cause is equally likely to be a characteristic of the male partner (Rowland 1992). The main causes are biological, although stress and anxiety can exacerbate physical problems.

Infertility and its treatment are distressing, but neither is associated with major psychological disturbance. The long-term outcomes of successful *in vitro* fertilization for children and their families are positive, but most outcomes are less benign. Involvement in *in vitro* fertilization is disruptive, expensive, and stressful. The probability of bearing a living child is around 10 percent per treatment cycle, and the risk of gynecological and obstetric complications, malformations, and multiple births is significantly increased. Repeated cycles of treatment are associated with increasing levels of depression (Rowland 1992).

3.2 *Fertility Control*

Access to adequate contraception and to abortion is important to women's physical and psychological well being. Where abortions are legal, around 2 percent of women aged between 15 and 44 have an abortion in any year. The worldwide annual death rate from illegal abortions has been estimated at 100 000, while the death rate from legal abortions is 0.05 per 10 000, 25 times lower than the rate for childbirth. Provision

of safe abortions and adequate contraception would have a major impact on the well being of women and children in developing countries.

Women may need support in deciding whether to choose abortion, but the majority experience no significant psychological distress. Psychological distress following abortion is considerably lower than that associated with childbirth. Distress is highest during decision making before the abortion, and the most common postabortion reaction is a sense of relief.

Women who cope best with abortion are characterized by optimism, high self-esteem, and a sense of personal control. Social, cultural, and partner support, and confidence in the decision, assist in coping, while women with a history of psychiatric disorder are likely to react negatively (Adler et al. 1992).

3.3 Miscarriage

Miscarriages are relatively common; over 30 percent of 50-year-old women have had at least one miscarriage. Reactions to spontaneous abortion of a wanted pregnancy are generally characterized by short-term grief, anxiety, and depression, normally resolved successfully within a few months (Slade 1994). Reactions to medical abortions follow a similar pattern, with short-term distress which may be severe but which resolves rapidly. Serious negative consequences are rare, and associated with previous psychiatric illness or lack of social support.

4. Menopause

Menopause is a natural physiological process resulting from aging of the ovaries, leading to cessation of ovulation and of estrogen production. Natural menopause occurs over several years. During the perimenopause, estrogen levels decrease and menstruation becomes irregular and ceases. The postmenopause, defined after 12 months without menstruation, is reached at an average age of 50.4 years.

Menopausal symptoms are categorized as: vasomotor (e.g., hot flushes, night sweats), psychosomatic (e.g., headaches, dizziness), and psychological (e.g., tiredness, irritability). The most common symptoms, reported across most cultures, are hot flushes and night sweats, clearly the result of changes in estrogen. However, prevalence of symptoms varies between countries, suggesting that sociocultural factors play a role (Hunt 1994).

Hormone replacement therapy relieves hot flushes and night sweats, but does not influence psychological well being in mid-aged women. Stress, life events, history of depression, and expectations are the strongest predictors of psychological well being during menopause. Women in their middle years often experience major life events (McKinlay et al. 1987). Once life changes are accounted for, there is little evidence to suggest that menopausal women are particularly prone to negative psychological states.

5. Conclusion

Distress associated with gynecological events is better explained by poor social support and by negative expectations than by endocrine function. Future research might usefully examine the experiences of women in a wider range of situations, including women in same-sex relationships, women who choose childlessness, and women with disabilities. An exploration of the experiences of women in different ethnic groups would also provide a broader perspective on psychosocial aspects of gynecological health.

See also: Fertility and Culture: Anthropological Insights; Fertility Control: Overview; Fertility of Single and Cohabiting Women; Fertility: Proximate Determinants; Fertility Theory: Embodied-capital Theory of Life History Evolution; Gender and Health Care; Gender and Physical Health; Gender and Reproductive Health; Gender Role Stress and Health; Reproductive Medicine: Ethical Aspects; Reproductive Rights in Developing Nations; Social Networks and Fertility; Women's Health

Bibliography

Adler N E, David H P, Major B N, Roth S H, Russo N F, Wyatt G E 1992 Psychological factors in abortion: A review. *American Psychologist* **47**: 1194–204

Cox J L, Holden J M, Sagovsky R 1987 Detection of postnatal depression: Development of the 10-item Edinburgh Postnatal Depression Scale. *British Journal of Psychiatry* **150**: 782–6

Hunt K 1994 A 'cure for all ills'? Construction of the menopause and the chequered fortunes of hormone replacement therapy. In: Wilkinson S, Kitzinger C (eds.) *Women and Health: Feminist Perspectives*. Taylor & Francis, London

Klebanov P K, Ruble D N 1994 Toward an understanding of women's experience of menstrual cycle symptoms. In: Adesso V J et al. (eds.) *Psychological Perspectives on Women's Health*. Taylor & Francis, Washington, DC

Kumar R 1994 Postnatal mental illness: A transcultural perspective. *Social Psychiatry and Psychiatric Epidemiology* **29**: 250–64

Logue C M, Moos R H 1986 Perimenstrual symptoms: Prevalence and risk factors. *Psychosomatic Medicine* **48**: 388–409

McFarlane J M, Williams T M 1994 Placing premenstrual syndrome in perspective. *Psychology of Women Quarterly* **18**: 339–73

McKinlay J B, McKinlay S M, Brambilla D 1987 The relative contributions of endocrine changes and social circumstances to depression in mid-aged women. *Journal of Health and Social Behavior* **28**: 345–63
Richardson J T E 1992 *Cognition and the Menstrual Cycle*. Springer, New York
Rivera-Tovar A D, Frank E 1990 Late luteal phase dysphoric disorder in young women. *American Journal of Psychiatry* **147**: 1634–6
Rowland R 1992 *Living Laboratories: Women and Reproductive Technologies*. Pan Macmillan, Sydney, Australia
Slade P 1994 Predicting the psychological impact of miscarriage. *Journal of Reproductive and Infant Psychology* **12**: 5–16

C. Lee

Gypsies and Travelers

The label *Gypsy* refers to diverse groups—each with its own history, culture, and name—who live in scattered communities throughout Europe and North America and in parts of Asia, Central and South America, and Africa. (In Europe, the label *Roma* is commonly used.) The English word 'Gypsy' is derived from 'Egyptian,' based on the false assumption that they had come from Egypt. From linguistic evidence, however, it is apparent that Gypsies originated in India. They began migrating westward in the tenth century, spreading across Europe and absorbing aspects (e.g., religion, language) of the host cultures they lived among. Eventually, they formed the groups found today. The Indic structure of the major dialects of *Romani* spoken by Gypsies has been heavily influenced by European languages, Greek, Armenian, Persian, and possibly Arabic.

Traveler is the term commonly used to refer to indigenous groups who share economic and structural features with Gypsies. They practice (or once practiced) similar peripatetic occupations and occupy a similar minority position vis-à-vis the majority populations of the countries in which they live. In Ireland and Scotland, for example, indigenous groups once known as 'tinkers' and 'tinklers,' but who refer to themselves as Travelers or Traveling People, have occupied a niche very similar to Gypsies for at least 300 years. Similar indigenous Travelers found in other European countries include the *Resande* in Sweden and *Woonwagenbewoners* in The Netherlands. Other indigenous groups are found in South Asia, they include the *Kanjar* in Pakistan and northern India and the *Ghorbat* in Afghanistan. Like Gypsies, members of some Traveler groups have since migrated to other countries. Irish and Scottish Travelers, for example, now also live in the United States, having first emigrated there in the mid-1800s. The designation Traveler is also sometimes used as a broad term to encompass both indigenous Travelers and Gypsy populations.

1. Identity

All Gypsies and Travelers see themselves in opposition to outsiders who they typically categorize with a single label. Many Gypsies, for example, use the term *gadje* (non-Roma in the Romani language); Irish Travelers call members of the majority Irish population *buffers*. Most Gypsies follow, to varying degrees, a system of purity rules which govern women's behavior and how food, eating utensils, clothing, washing containers, and the body are handled. Most believe that the body is ritually unclean from the waist down and, consequently, that anything associated with the lower body is polluting. Individuals and groups who do not follow purity rules are by definition 'polluted' or ritually unclean, which is one factor reinforcing the boundaries between Gypsies and non-Gypsies. Of course, Gypsies and Travelers also distinguish their own group from similar groups, usually on the basis of kin ties (or lack thereof), differences in dialect, and other criteria such as socioeconomic status, occupation, and length of settlement (i.e., before 1945 or during the communist era for East European groups).

Language also acts as a marker of ethnic identity. While Gypsies speak the language of the majority population of the countries in which they live, most also speak a form of Romani. Most indigenous Travelers likewise have specialized argots which they typically use to disguise their communications in the presence of outsiders (e.g., during economic transactions). Other markers of group identity, such as items of material culture, skills, or traditions, are also used to create and reinforce a symbolic separation between themselves and outsiders. Roma in Hungary, for example, assert that it is their 'Gypsy work'—their wit, speech, and ability to hustle—that allows them to survive, rather than the income they make from the wage labor the State forces them to engage in. Irish Travelers pride themselves in their 'cleverness'—their ability to live by their wits—and see this as something which differentiates them from other Irish. It has been suggested that 'disreputable' behavior of Spanish Gitanos—their flaunting of accepted social norms about appropriate public comportment (e.g., loudness, violations of street etiquette)—is a deliberate strategy intended to offend non-Gypsies and keep them at a distance. Other groups use their involvement in horse trading or their distinctive music and celebrations to symbolically draw boundaries between themselves and outsiders.

When possible, Gypsies and Travelers limit their interaction with members of the majority population to economic transactions and brief, necessary encounters with institutional representatives such as legal, social service, or medical personnel. Most academics who have studied these groups have documented their antiestablishment attitudes and efforts to remain outside the direct supervision of the State. Socialization among Gypsies and Travelers instills in

most children a strong sense of ethnic pride, and few express a desire to abandon their identity or group. This pride is reflected in and reinforced by high rates of group endogamy. Nevertheless, attrition in the form of 'passing' into mainstream society as well as intermarriage with non-Gypsies or non-Travelers does occur. Some groups have been subjected to long-term state-sponsored assimilation programs, which have included forced settlement among non-Gypsies, wage labor, and education.

As the quintessential Other, Gypsies and Travelers have historically captured the imagination of artists and writers and figure prominently in the arts and literature of many countries. In Russia, for example, writers such as Pushkin, Tolstoy, and Blok included Roma in their works. Likewise, Travelers ('tinkers') have figured prominently in the work of such literary figures as Irish playwright J. M. Synge.

2. *Economic Organization*

The traditional economy of Gypsies and Travelers was based on petty trade and the provision of craft products, specialized services, and entertainment to customers and clients in the majority populations they lived among. In many countries, they serviced a scattered rural population, conveniently bringing goods and services to their customers' doors. Others provided services which had to be performed on site. Rom wipe-tinners, for example, traveled the United States retinning expensive mixing bowls and massive, relatively immobile counter tops, holding tanks, tubs, troughs, sinks, and separators used in institutional cafeterias, dairies, and commercial bakeries. Blacktopping (or tarmac laying), a common occupation of Gypsies and Travelers in North America, Ireland, and Great Britain today, likewise must be done on site. Sometimes, Gypsies and Travelers provide unique products or services that cannot be acquired from anyone else. The Kanjar of Pakistan, for example, sold handmade terracotta toys, while the Qalandar entertained their clients with trained monkeys and bears. Fortune telling remains a profitable specialization for many American Rom. Because Gypsies and Travelers often accepted payment in kind (e.g., farm produce, fuel), the services they provided were especially attractive to poorer clients.

Occupational flexibility, the ability to do several income-earning activities at the same time, and to change occupations quickly as demand dictates, is a key aspect of the traditional economy. When the Ludar arrived in the United States from Bosnia, for example, they brought trained bears and monkeys, but they also peddled small goods, dealt in horses, told fortunes, and made rustic furniture. Since the 1950s, they have focused on laying blacktop, tarring roofs, and working carnival concessions. Irish Travelers once swept chimneys, made tin containers, bought and sold donkeys and horses, peddled small goods (e.g., brushes, needles, boot polish, religious pictures), told fortunes, and performed agricultural labor. As these rural-based occupations became obsolete, they moved to urban areas and began laying tarmac, selling carpeting from door-to-door, collecting and recycling scrap metal, and selling second-hand furniture, car parts, and other merchandise from the roadside.

Many Gypsy and Traveler populations—or significant segments of them—are still peripatetic, exploiting narrowly-defined markets within the host economy. Families are successful because their overhead is minimal, they are willing to accept a narrow profit margin from multiple sources, and generally lack interest in material accumulation and capital expansion. They also use wage-free household labor. Typically, household income is produced by all family members—men and women, young and old. When larger work units are needed, they are drawn from the kin network or else the lowest socioeconomic stratum of the host society (e.g., the homeless).

While nomadism remains an important feature of many groups—and is the image instantly evoked by the very words 'Gypsy' and 'Traveler'—substantial numbers today are sedentary for all or much of the year. Some Western European groups are also sedentary. Spanish Gitanos, for example, have been sedentary for centuries. Although many Gypsies and Travelers in Ireland and Great Britain remain nomadic, the pressure from government to settle is strong. In Eastern Europe, Gypsies were long ago forced into wage labor—generally low-status, unskilled jobs—and have been sedentary for generations. In North America today, only a few families in each Gypsy and Traveler group are predominantly nomadic. Most own or rent homes, and maintain a mixed pattern of settlement and travel.

3. *Social and Political Organization*

It is difficult to generalize about the social organization of Gypsy and Traveler groups. Many groups reckon descent bilaterally and have no formalized or corporate kin groups. Individuals have equal rights and obligations towards both maternal and paternal kin. At marriage, a couple can choose to affiliate with either spouse's family, and later, may periodically live on their own or affiliate with nonkin. Other groups—such as the Rom and Ludar in the United States—trace descent through the male line, have named patrilineages, and practice preferential patrilocal residence at marriage, at least until the birth of a young couple's first child. Descent in most South Asian groups is also patrilineal.

Residential groupings larger than the household, whether they are travel clusters or stationary enclaves

within cities, are fluid. Kinship is the fundamental basis for association, but group composition and size is continually being renegotiated in response to changing economic opportunities and a host of social and legal concerns. Groups split apart due to personal animosities, the desire for change, the need to scatter in order to find new clients, or as a result of punitive actions (e.g., evictions) by local authorities. Likewise, they coalesce for many reasons. Scattered kin come together to celebrate life crisis events such as marriages and funerals, to protect a valuable economic territory from the encroachment of other groups, to defend a kinsman from insult or attack, and, sometimes, to make evictions by local authorities more difficult.

Political organization is also diverse. The most formalized mechanisms of social control and political leadership are found among some Gypsy groups and South Asian Travelers. American Rom, for example, have a *Rom Baro* (big man) who speaks for his extended family and controls economic activities in the territory they claim. Important disputes are adjudicated by a *kris*,an *ad hoc* Gypsy court, composed of popularly-nominated, male family heads who act as 'judges' and representatives from each kin grouping within the vicinity who act as witnesses and jury. The *kris* seeks consensus; its proceedings are spontaneous and often volatile; and the fines or punishments it metes out are morally, not legally, binding. Other American Gypsy groups such as the Romnichel, however, lack both the *Rom Baro* and the *kris*. Various South Asian groups have loosely structured *panchayat* councils. These are informal meetings convened primarily to settle disputes, with respected elders acting as arbitrators. Both the *kris* and the *panchayat* are mechanisms for airing grievances, alleviating tension, asserting (and sometimes reassessing) group values, arriving at a compromise, and restoring order.

Researchers working with diverse groups of Gypsies and Travelers in Europe, North America, and South Asia have noted the importance members attach to assembly, both at annual events like seasonal fairs and at periodic events like religious pilgrimages, marriages, and funerals. During assemblies, group solidarity is reinforced through informal social interaction (including music, song, and drinking), marriage negotiations, religious ceremonies, and dispute settlements. Such gatherings demonstrate and reinforce a common identity that overrides, at least temporarily, localized kinship and political alliances. English Gypsies appear to recognize the important role fairs play in their collective life. When authorities tried to close Appleby Fair in 1965 and to prevent them from assembling at Epsom in 1967, they united in protest.

Unless prevented from doing so by the state, Gypsies and Travelers seem to respond eagerly to opportunities for assembly and social engagement. Their gatherings also have an emotional intensity—evident in the total absorption of buleria performances, flamenco, and 'deep song' as performed by Spanish Gitanos and in the drinking and aggressive interpersonal interactional style of many groups—that is striking to the researchers who have observed them. It is as if what these Gypsy and Traveler groups lack in formal organization or structure they make up for by periodic assembly and intense emotional engagement.

4. *Harassment and Assimilation*

Gypsies and Travelers have sometimes enjoyed the liberties of other citizens. Under Ottoman rule, for example, Gypsies had the same citizenship rights as anyone else as long as they paid the taxes required of their religious group: Christian Roma paid higher taxes than Muslim Roma. Much more frequently, however, they have been persecuted. In Europe, Gypsies and Travelers have been subject to punitive legislation for many centuries. Individuals as well as entire groups have been denied entry, forcibly settled, had their trades outlawed, forced to adopt wage labor, had their children taken and adopted by non-Gypsy families, been sterilized, imprisoned, enslaved, or deported. In Moldavia and Wallachia—Romania's two historic provinces—Gypsies were sold into slavery as early as the fifteenth century; legal emancipation for Romany slaves or *robi* finally came in 1864. During the Holocaust a half to more than a million Gypsies were exterminated. Deep-seated anti-Gypsy prejudice continues today throughout Eastern Europe and parts of Western Europe, periodically erupting in physical assaults by local residents and the police.

In the second half of the twentieth century, many Western European governments and private agencies became involved in programs ostensibly aimed at providing Gypsies and Travelers with 'choice,' 'options,' or 'self-determination.' Many of these movements came about as modernization in the host society eliminated the need for the rural-based trades and services nomadic groups had provided, forcing them to seek new clients. In Ireland and Great Britain, for example, Travelers and Gypsies converged on cities looking for work, camping along motorways, next to housing estates, and on rapidly diminishing amounts of open land. This led to a public outcry and their definition as a 'social problem.' Public policy has fluctuated over the years. Liberal programs (i.e., providing campsites and mobile classrooms for nomadic families, creating culturally-sensitive curricula) have vied with more straightforward attempts to assimilate Gypsies and Travelers through housing, laws prohibiting camping, and conventional vocational training education. In recent years, however, Gypsies and Travelers have become more politically active. Many organizations have emerged to articulate their concerns and lobby government for the same rights as any other minority. Cooperation across national boundaries with other Gypsy and Traveler groups is also becoming more common.

In Eastern Europe, government policy had focused more on forced assimilation, including aggressively settling Gypsy families that were nomadic, dispersing those who had lived in settled enclaves, enrolling children in school, licensing or banning specific trades, forcing adults to accept wage labor, and officially denying their existence as ethnic minorities. Many Gypsies were able, nevertheless, to preserve their separate identity in spite of official communist policy. The collapse of state socialism in Eastern Europe in 1989 brought new hope to the approximately five million Gypsies in the region who responded by organizing a plethora of cultural groups and political parties to articulate and represent their diverse interests. Some common goals include achieving full recognition and civil rights as distinct ethnic minorities, having schools use Romani as a second language of instruction, and getting state-supported social and economic programs to create opportunities and address poor living conditions. In Macedonia, Gypsies called upon the United Nations to establish a Romany state to be called 'Romanestan.' Most groups efforts, however, have been hampered by internal leadership problems, the multiplicity of competing Gypsy organizations within each country, and, ultimately, their reliance on the State to accord their organizations and demands legitimacy. Anti-Gypsy sentiment in Europe remains strong. Attacks on Gypsies in the 1990s targeted both long-term residents of settled communities in Eastern Europe and recent immigrants fleeing to Western countries from Romania, Bulgaria, Hungary, and the former Yugoslavia.

See also: Counterculture; Diaspora; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Nomads/Nomadism in History; Stigma, Social Psychology of; Subculture, Sociology of

Bibliography

Barany Z 1998 Ethnic mobilization and the state: the Roma in Eastern Europe. *Ethnic and Racial Studies* **21**(2): 308–27

Crowe D 1994 *A History of the Gypsies of Eastern Europe and Russia*. St. Martin's Press, New York

Gmelch S B 1986 Groups that don't want in: Gypsies and other artisan, trader, and entertainer minorities. *Annual Review of Anthropology* **15**: 307–30

Kendrick D 1993 *Gypsies from India to the Mediterranean*. CRDP, Toulouse, France

Liegeois J-P 1994 *Roma, Gypsies, Traveller*, rev. edn. Council for Cultural Cooperation, Strasbourg

Lucassen L, Willems W, Cottaar A 1998 *Gypsies and Other Itinerant Groups: A Socio-Historical Approach*. St Martins Press, New York

Mayall D 1995 *English Gypsies and State Policies*. University of Hertfordshire Press, Hatfield, UK

Muller-Hill B 1988 *Murderous Science: Elimination by Scientific Selection of Jews, Gypsies, and Others, Germany 1933–1945*. Oxford University Press, Oxford, UK

Puxon G 1987 *Roma: Europe's Gypsies*. Minority Rights Report No. 14, Minority Rights Group, London

Rao A 1986 *The Other Nomads: Peripatetic Minorities in Cross-Cultural Perspective*. Bohlau Verlag, Köln, Germany

Salo M 1981 *The American Kalderash: Gypsies in the New World*. Gypsy Lore Society, Hackettstown, NJ

Stewart M 1997 *The Time of the Gypsies*. Westview Press, Boulder, CO

S. B. Gmelch

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Ghetto

It is a paradox that, while the social sciences have made extensive use of the 'ghetto' as a *descriptive term*, they have failed to forge a robust *analytical concept* of the same. In the historiography of the Jewish diaspora in early modern Europe and under Nazism, the sociology of the black American experience in the twentieth-century metropolis, and the anthropology of ethnic outcasts in Africa and East Asia, its three traditional domains of application, the term 'ghetto' variously denotes a bounded urban ward, a web of group-specific institutions, and a cultural and cognitive constellation (values, mindset, or mentality) entailing the sociomoral isolation of a stigmatized category as well as the systematic truncation of the life space and life chances of its members. But none of these strands of research has taken the trouble to specify what makes a ghetto *qua* social form, which of its features are constitutive and which derivative, as they have, at each epoch, taken for granted and adopted the *folk concept* extant in the society under examination—which explains that the notion, appearing self-evident, does not figure in most dictionaries of social science, including previous publications of this encyclopedia.

1. *A Wooly and Shifting Notion*

Thus the semantic range of the 'ghetto' in American society and social science, which has dominated inquiry into the topic both quantitatively and thematically, has successively expanded and contracted in keeping with how political and intellectual elites have viewed the vexed nexus of ethnicity and poverty in the city (Ward 1989). At first, in the latter half of the nineteenth century, the term designated residential concentrations of European Jews in the Atlantic seaports and was clearly distinguished from the 'slum' as an area of housing blight and social pathology. It dilated during the Progressive era to encompass all inner-city districts wherein exotic newcomers gathered, namely lower-class immigrants from the Southeastern regions of Europe and African Americans fleeing the Jim Crow regime of caste subjugation in the US South. Expressing ruling-class worries over whether these groups could or should assimilate into the predominant Anglo-Saxon pattern of the country, the term then referred to the intersection between the ethnic neighborhood and the slum, where segregation combined with physical disrepair and overcrowding to exacerbate urban ills such as criminality, family breakdown, and pauperism, and thwart participation in national life. This conception was given scientific authority by the ecological paradigm of the Chicago school of sociology. In his classic book *The Ghetto*, Louis Wirth (1928, p. 6) assimilates to the Jewish ghetto of medieval Europe the 'Little Sicilies, Little Polands, Chinatowns, and Black Belts in our large cities,' along with the 'vice areas' hosting deviant types such as hobos, bohemians, and prostitutes—all of which are said to be 'natural areas' born of the universal desire of different groups to 'preserve their peculiar cultural forms' and each fulfilling a specialized 'function' in the broader urban organism.

The notion contracted after World War II under the press of the Civil Rights movement to signify mainly the compact and congested enclaves to which African Americans were forcibly relegated as they migrated into the industrial centers of the North. The growth of a 'Black Metropolis in the womb of the white,' wherein Negroes evolved distinct and parallel institutions to compensate for and shield themselves from unflinching exclusion by whites (Drake and Cayton 1945), contrasted sharply with the smooth residential dispersal of European Americans of foreign stock. Writing at the acme of the black uprisings of the 1960s, Kenneth Clark (1965, p. 11) made this relationship of ethnoracial subordination epicentral to his dissection of the *Dark Ghetto* and its woes: 'America has contributed to the concept of the ghetto the restriction of persons to a special area and the limiting of their freedom of choice on the basis of skin color. The dark ghetto's invisible walls have been erected by the white society, by those who have power.' This diagnosis was confirmed by the Kerner Commission (1968, p. 2), a bipartisan taskforce appointed by President Johnson whose official report on the 'civil disorders' that rocked the American metropolis famously warned that, due to white racial intransigence, America was 'moving toward two societies, one black, one white—separate and unequal.' But over the ensuing two decades, the dark ghetto collapsed and devolved into a barren territory of dread and dissolution due to deindustrialization and state policies of welfare reduction and urban retrenchment. And, as racial domination grew more diffuse and diffracted through a class prism, the category was displaced by the duet formed by the geographic euphemism of 'inner city' and the neologism of 'underclass,' defined as the substratum of ghetto residents plagued by antisocial behaviors, acute joblessness, and social isolation (Wilson 1987). By the 1990s, the neutralization of the 'ghetto' in policy-oriented research culminated in the expurgation of all mention of race and power to redefine it as any tract of extreme poverty, irrespective of population and institutional makeup, in effect dissolving the ghetto back into the slum.

The extension of the term to the study of the distinctive sociocultural patterns elaborated by homosexuals in the cities of advanced societies 'in response to both stigma and gay liberation' after the Stonewall riots (Levine 1979, p. 31) and its recent resurgence in Western Europe in heated scientific and political debates over the links between postcolonial

immigration, postindustrial economic restructuring, and urban dualization (Mingione 1996) would seem only to further muddle its meaning. Yet one can extract out of these varied literatures common threads and recurrent properties to *construct a relational concept* of the ghetto as an *instrument of closure and control* that clears up most of the confusion surrounding it and makes it a powerful tool for the social analysis of ethnoracial domination and urban inequality. For this it suffices to return to the historical inception of the word and of the phenomenon it depicted in Renaissance Venice.

2. *A Janus-faced Institution of Ethnic Closure and Control*

Coined by derivation from the Italian *giudecca*, *borghetto* or *gietto* (or from the German *gitter* or the Talmudic Hebrew *get*: the etymology is disputed), the word 'ghetto' initially refers to the forced consignment of Jews to special districts by the city's political and religious authorities. In medieval Europe, Jews were commonly allotted quarters wherein they resided, administered their own affairs, and followed their customs. Such quarters were granted or sold as a privilege to attract them into the towns and principalities for which they fulfilled key roles as money-lenders, tax collectors, and long-distance tradesmen. But, between the thirteenth and the sixteenth century, in the wake of the upheavals caused by the Crusades, favor gradually turned into compulsion (Stow 1992). In 1516, the Senate of Venice ordered all Jews rounded up into the *ghetto nuovo*, an abandoned foundry on an isolated island enclosed by two high walls whose outer windows and doors were sealed while watchmen stood guard on its two bridges and patrolled the adjacent canals by boat. Jews were henceforth allowed out to pursue their occupations by day, but they had to wear a distinctive garb and return inside the gates before sunset on pain of serious punishment. These measures were designed as an alternative to expulsion to enable the city-state to reap the economic benefits brought by the presence of Jews (including rents, special taxes, and forced levies) while protecting their Christian residents from contaminating contact with bodies perceived as unclean and dangerously sensual, carriers of syphillis and vectors of heresy, in addition to bearing the taint of money-making through usury, which the Catholic Church equated with prostitution (Sennett 1994, p. 224).

As this Venetian model spread in cities throughout Europe and around the Mediterranean rim (Johnson 1987, pp. 235–245), territorial fixation and seclusion led, on the one hand, to overcrowding, housing deterioration, and impoverishment as well as excess morbidity and mortality, and, on the other, to institutional flowering and cultural consolidation as urban Jews responded to multiplying civic and occupational restrictions by knitting a dense web of group-specific organizations that served as so many instruments of collective succor and solidarity, from markets and business associations, to charity and mutual aid societies, to places of religious worship and scholarship. The *Judenstadt* of Prague, Europe's largest ghetto in the eighteenth century, even had its own city hall, the *Rathaus*, an emblem of the relative autonomy and communal strength of its residents, and its synagogues were entrusted not only with the spiritual stewardship but also with the administrative and judicial oversight of its population. Social life in the Jewish ghetto was turned inward and verged 'on overorganization' (Wirth 1928, p. 62), so that it reinforced both integration within and isolation from without.

One can detect in this inaugural moment the four constituent elements of the ghetto, viz., *stigma, constraint, spatial confinement, and institutional encasement*. The ghetto is a social-organizational device that employs space to reconcile two antinomic purposes: to maximize the material profits extracted out of a group deemed defiled and defiling and to minimize intimate contact with its members so as to avert the threat of symbolic corrosion and contagion they carry. This same dual rationale of *economic exploitation cum social ostracization* governed the genesis, structure, and functioning of the African-American ghetto in the Fordist metropolis during most of the twentieth century. Blacks were recruited into northern US cities after World War I because their unskilled labor was indispensable to the industries that formed the backbone of the expanding factory economy. But there was no question of them mixing and consorting with whites, who regarded them as inherently vile, congenitally inferior, and shorn of ethnic honor owing to the stain of slavery. As blacks moved in from the South in millions, white hostility increased and patterns of discrimination and segregation that had hitherto been informal and inconsistent hardened in housing, schooling, and public accommodations and were extended to the economy and polity (Spear 1968, Osofsky 1971). African Americans had no choice but to seek refuge inside the bounded perimeter of the Black Belt and to endeavor to develop in it a network of separate institutions to procure the basic needs of the castaway community. Thus arose a parallel city anchored by black churches and newspapers, black block clubs and lodges, black schools and businesses, and black political and civic associations, nested at the core of the white metropolis yet sealed from it by an impassable fence built of custom, legal suasion, economic discrimination (by realtors, banks, and the state), and by violence manifested in the beatings, fire-bombings, and riots that checked those who dared stray across the color line.

This forced institutional parallelism predicated on enveloping and inflexible spatial seclusion—not extreme poverty, housing blight, cultural difference, or mere residential separation—is what has distinguished African Americans from every other group in US history, as noted by leading students of the black urban experience from W.E.B. Du Bois and E. Franklin Frazier to Drake and Cayton to Kenneth Clark and Oliver Cox (Wacquant 1998). It also characterizes the trajectory of the Burakumin in the Japanese city after the close of the Tokugawa era (Hane 1982). As the descendants of the *eta*, the lowest of the four castes forming the estate order of feudal Japan, the Burakumin were untouchables in the eyes of the Buddhist and Shinto religions, and they were legally confined from sundown to sunup in out-of-the-way hamlets (*buraku*), obliged to wear a yellow collar and to walk barefoot, expected to drop on their hands and knees when addressing commoners, and restricted to marrying solely among themselves. Though they were officially emancipated in 1871, as they moved into cities they were funneled against their will into notorious neighborhoods near garbage dumps, crematoria, jails, and slaughterhouses, widely viewed as nests of criminality and immorality. There, they were barred from industrial employment and locked in low-paying and dirty jobs, sent to separate schools, and compelled to remain endogamous by the indelible taint of their blood as traced through 'family registration records' (DeVos and Wagatsuma 1966). In the late 1970s, according to the Burakumin Defense League, they were estimated to number 3 million, trapped in 6,000 ghettos in some thousand cities across the main island.

Spread over three continents and five centuries, the Jewish, African-American, and Burakumin cases demonstrate that the ghetto is not, *pace* Wirth (1928, pp. 284–285), a 'natural area' arising via environmental adaptation governed by a biotic logic 'akin to the competitive cooperation that underlies the plant community.' The error of the early Chicago school here consisted in falsely 'converting history into natural history' and passing ghettoization as 'a manifestation of human nature' virtually coterminous with 'the history of migration' (Wirth 1928, p. 285) when it is a highly peculiar form of urbanization warped by asymmetric relations of power between ethnoracial groupings: a special form of *collective violence concretized in urban space*. That ghettoization is *not* an 'uncontrolled and undesigned' process, as Robert E. Park asserted in his preface to *The Ghetto* (Wirth 1928, p. viii), was especially visible after World War II when the black American ghetto was reconstructed from the top down through state policies of public housing, urban renewal, and suburban economic development intended to bolster the rigid separation of blacks from whites (Hirsch 1983). It is even more glaring in the instance of the 'caste cities' built by colonial powers to inscribe in space the hierarchical ethnic organization of their overseas possessions, such as Rabat under French rule over Morocco and Cape Town after the passage of the Group Areas Acts under the apartheid regime of South Africa (Abu-Lughod 1980, Western 1981).

Recognizing that it is a product and instrument of group power makes it possible to appreciate that, in its full-fledged form, the ghetto is a *Janus-faced institution* as it serves opposite functions for the two collectives that it binds in a relation of asymmetric dependency. For the dominant category, its rationale is to *confine and control*, which translates into what Max Weber calls the 'exclusionary closure' of the dominated category. For the latter, however, it is an *integrative and protective device* insofar as it relieves its members from constant contact with the dominant and fosters consociation and community building within the constricted sphere of intercourse that it creates. Enforced isolation from the outside leads to the intensification of social exchange and cultural sharing inside. Ghettos are the product of a mobile and tensionful dialectic of external hostility and internal affinity that expresses itself as ambivalence at the level of collective consciousness. Thus, although European Jews consistently protested relegation within their outcast districts, they were also deeply attached to them and appreciative of the relative security they afforded and the special forms of collective life they supported: Frankfurt's ghetto in the eighteenth century was 'not just the scene of confinement and persecution but a place where Jews were entirely, supremely, at home' (Gay 1992, p. 67). Similarly, black Americans took pride in having 'erected a community in their own image,' even as they resented the fact that they had done so under duress, as a result of unyielding white exclusion aimed at warding off the specter of 'social equality,' that is, sexual mixing (Drake and Cayton 1945, p. 115).

3. *Disentangling Poverty, Segregation, and Ethnic Clustering*

Articulating the concept of ghetto makes it possible to disentangle the relationship between ghettoization, urban poverty, and segregation, and thence to clarify the structural and functional differences between ghettos and ethnic neighborhoods. It also enables us to spotlight on the role of the ghetto as a symbolic incubator and matrix for the production of a spoiled identity.

1. *Poverty is a frequent but derivative and variable characteristic of ghettos*: the fact that most ghettos have historically been places of endemic and often acute misery owing to the paucity of space, the density of settlement, and the economic exploitation and generalized maltreatment of their residents does not imply that a ghetto is necessarily a place of

destitution, nor that it is uniformly deprived. The *Judengasse* of Frankfurt, instituted in 1490 and abolished in 1811, went through periods of prosperity no less than penury and it contained patches of extraordinary opulence as court Jews helped the city become a vibrant center of trade and finance—part of its glamor to this day comes from it being the ancestral home of the Rothschild dynasty (Wirth 1928, Chap. 4). James Weldon Johnson (1930, p. 4) insisted that the Harlem of the 1930s was 'not a slum or a fringe' but the 'cultural capital' of black America, where 'the Negro's advantages and opportunities are greater than in any other place in the country.' Likewise, Chicago's 'Bronzeville' was far more prosperous at mid-century than southern black communities and harbored the largest and most affluent African-American bourgeoisie of its era (Drake and Cayton 1945). Whether a ghetto is poor or not depends on extraneous factors such as demography, ecology, state policies, and the shape of the surrounding economy.

Conversely, not all dispossessed and dilapidated urban districts are ghettos. Declining white neighborhoods in the deindustrializing cities of the US Midwest and British Midlands, depressed rural towns of the former East Germany and southern Italy, and the disreputable *villas miserias* of the greater Buenos Aires at the close of the twentieth century are territories of working-class demotion and decomposition, not ethnic containers dedicated to maintaining an outcast group in a relationship of seclusive subordination. They are not ghettos other than in a metaphorical sense, no matter how impoverished—if extreme rates of poverty sufficed to make a ghetto, then large chunks of the former Soviet Union and most third-world cities would be gargantuan ghettos. The *favelas* of the Brazilian metropolis often portrayed as segregated dens of dereliction and disorganization turn out to be working-class wards with finely stratified webs of ties to industry and to the wealthy districts for which they supply household service labor. As in the *ranchos* of Venezuela and the *poblaciones* of Chile, the families that dwell in these squatter settlements span the color continuum and have extensive genealogical bonds to higher-income households; they are 'not socially and culturally marginal, but stigmatized and excluded from a closed class system' (Perlman 1976, p. 195; also Quijano 1968). Given that not all ghettos are poor and not all poor areas are ghettos, one cannot collapse and confound the analysis of ghettoization with the study of slums and lower-class districts in the city.

2. Similarly, *all ghettos are segregated but not all segregated areas are ghettos*. The select boroughs of the West of Paris, the exclusive upper-class suburbs of Boston or Berlin, and the 'gated communities' that have mushroomed in global cities such as São Paulo, Toronto, and Miami are monotonous in terms of wealth, income, occupation, and often ethnicity, but they are not for all that ghettos. Segregation in them is entirely voluntary and elective, and for that reason it is neither all-inclusive nor perpetual. Fortified enclaves of luxury package 'security, seclusion, social homogeneity, amenities, and services' to enable bourgeois families to escape what they perceive as 'the chaos, dirt, and danger of the city' (Caldeira 2000, pp. 264–265). These islands of privilege serve to enhance, not curtail, the life chances and protect the lifestyles of their residents, and they radiate a positive aura of distinction, not a sense of infamy and dread.

This suggests that residential segregation is a necessary but not a sufficient condition for ghettoization. For a ghetto to emerge spatial confinement must, first, be imposed and all-encompassing and, second, it must be overlayed with a distinct and duplicative set of institutions enabling the group thus cloistered to reproduce itself within its assigned perimeter. If blacks are the only ethnic group to be 'hypersegregated' in American society (Massey and Denton 1992), it is because they are the only community to have combined involuntary segregation with organizational parallelism entrapping them in a separate and inferior social cosmos of their own, which in turn bolstered their residential isolation. That even involuntary segregation at the bottom of the urban order does not *eo ipso* produce ghettos is demonstrated by the fate of the declining French *banlieues* after the 1980s. Although they have been widely disparaged as 'ghettos' in public discourse and their inhabitants share a vivid feeling of being cast out in a 'penalized space' suffused with boredom, anguish, and despair (Pétonnet 1982), relegation in these depressed concentrations of public housing at the urban periphery is based on class, not ethnicity; as a result they are culturally heterogeneous, typically harboring native French families along with immigrants from two dozen nationalities; and their inhabitants suffer not from institutional duplication but, on the contrary, from the lack of an ingrown organizational structure capable of sustaining them in the absence of gainful employment and adequate public services. Like the British or Dutch inner cities and the immigrant clusters of urban Germany or Italy, the French *banlieues* are, sociologically speaking, *anti-ghettos* (Wacquant to be published).

3. *Ghettos and ethnic neighborhoods have divergent structures and opposite functions*: moving beyond a gradational perspective to scrutinize the peculiar patterning of social relations within the ghetto as well as between it and the surrounding city throws into sharp relief the differences between the ghetto and the ethnic clusters or immigrant neighborhoods such as newcomers to the metropolis have formed in countless countries. The foreign 'colonies' of interwar Chicago that Robert Park, Ernest Burgess, and Louis Wirth—and after them the liberal tradition of assimilationist sociology and historiography—mistook for so many white 'ghettos' were scattered and

mobile constellations born of cultural affinity and occupational concentration. Segregation in them was partial and porous, a product of immigrant solidarity and ethnic attraction instead of being imposed by outgroup hostility. Consequently, residential separation was neither uniformly nor rigidly visited upon these groups: in 1930, when the all-black Bronzeville harbored 92% of the city's Afro-American population, Chicago's Little Ireland was 'an ethnic hodge-podge' of 25 nationalities composed of only one-third Irish persons and containing a paltry 3% of the city's denizens of Irish ancestry (Philpott 1978, pp. 141–145).

What is more, the distinctive institutions of European immigrant enclaves were turned outward: they operated to facilitate adjustment to the novel environment of the US metropolis. They neither replicated the organizations of the country of origin nor perpetuated social isolation and cultural distinctiveness, and so they typically waned within two generations as their users gained access to their American counterparts and climbed up the class order and the corresponding ladder of places (Nelli 1970; a similar process of spatial diffusion via class incorporation of Belgian, Italian, Polish, and Iberian immigrants in the French industrial city is reported by Noiriel 1988), all of which is in sharp contrast with the immutable racial exclusivity and enduring institutional alterity of the Black Belt. This Chicago illustration dramatizes the fact that immigrant neighborhood and ghetto serve diametrically opposed functions: one is a springboard for *assimilation* via cultural learning and social-cum-spatial mobility, and the other a material and symbolic isolation ward geared toward *dissimilation*. The former is best figured by a bridge, and the latter by a wall.

4. *A Tainted Identity Machine*

The ghetto is not only the concrete means and materialization of ethnoracial domination through the spatial segmentation of the city but also a potent *collective identity machine* in its own right, for it helps incrustate and elaborate the very division of which it is the expression in two complementary and mutually reinforcing ways. First, the ghetto sharpens the boundary between the outcast category and the surrounding population by deepening the sociocultural chasm between them: it renders its residents objectively and subjectively more dissimilar from other urban dwellers by submitting them to unique conditionings, so that the patterns of cognition and conduct they fashion have every chance of being perceived by outsiders as singular, exotic, even aberrant (Sennett 1994, p. 244; Wilson 1987, pp. 7–8), which feeds prejudicial beliefs about them. Second, the ghetto is a cultural combustion engine that melts divisions among the confined group and fuels its collective pride even as it entrenches the stigma that hovers over it. Spatial and institutional entrapment deflect class differences and corrode cultural distinctions within the relegated ethnoracial category. Thus Christian ostracism welded Ashkenazic and Sephardic Jews under an overarching Jewish identity such that they evolved a common 'social type' and 'state of mind' across the ghettos of Europe (Wirth 1928, pp. 71–88, 1964). America's dark ghetto similarly accelerated the sociosymbolic amalgamation of mulattos and Negroes into a single 'race' and turned racial consciousness into a mass phenomenon fueling community mobilization against continued caste exclusion (Drake and Cayton 1945, p. 390).

Yet this unified identity cannot but be stamped with ambivalence as it remains tainted by the very fact that ghettoization proclaims what Weber calls the 'negative evaluation of honor' assigned to the group confined. It is therefore wont to foster among its members sentiments of self-doubt and self-hatred, dissimulation of one's origin through 'passing,' the pernicious derogation of one's kind, and even fantasical identification with the dominant (Clark 1965, pp. 63–67). And, because ghettoization is typically bound up tightly with ethnicity, segregation, and poverty, it is difficult to discern empirically which of the properties exhibited by ghetto dwellers are 'ghetto-specific cultural traits' as opposed to properties expressive of class, community, or masculinity (Hannerz 1969, p. 79). Also, cultural forms forged in the ghetto seep across its boundaries and circulate through the surrounding society where they often become outward signs of cultural rebelliousness and social eccentricity—as indicated by the fascination of bourgeois teenagers around the world for black American 'gangster rap.' This makes it difficult to distinguish between cultural forms effectively in currency among ghetto residents and the public imagery of them diffused in the broader society (including via scholarly writings).

It is fruitful to think of ghetto and ethnic cluster as *two ideal typical configurations situated at opposite ends* of a continuum along which different groups can be located or travel over time depending on the intensity with which the forces of stigma, constraint, spatial confinement, and institutional duplication and completeness coalesce with each other and impinge upon them. Ghettoization can then be turned into a *multilevel variable* for comparative analysis and empirical specification. It can become attenuated to the point where, through gradual erosion of its spatial, social, and mental boundaries, the ghetto devolves into an elective ethnic concentration operating as a springboard for structural integration and/or cultural assimilation into the broader social formation. This describes well the trajectory of the Chinatowns of the United States from the early to the late twentieth century (Zhou 1992) and the status of the

Cuban immigrant enclave of Miami that fostered integration through biculturalism after the Mariel exodus of 1980 (Portes and Stepick 1993). It also characterizes the 'Kimchee Towns' in which Koreans have converged in the metropolitan areas of Japan, which sport a blend of features making them a hybrid between ghetto and ethnic cluster (DeVos and Chung 1981): they are places of infamy that first arose through enmity and constraint, but over the years their population has become ethnically mixed and they have enabled Koreans to socialize and intermarry with Japanese neighbors as well as obtain Japanese citizenship through naturalization. This schema also fits the so-called 'gay ghetto,' which is more aptly characterized as a 'quasi-ethnic community' since 'most gay persons can "pass" and need not be confined to interacting with their "own kind"' and none are forced to reside in the areas of visible concentration of gay institutions (Murray 1979, p. 169).

The double-sidedness of the ghetto as a weapon and shield implies that, to the degree that its institutional completeness and autonomy are abridged, its protective role for the subordinate group is diminished and risks being swamped by its exclusionary modality. In cases where its residents cease to be of economic value to the dominant group, ethnoracial encapsulation can escalate to the point where the ghetto serves as an apparatus merely to warehouse the spoiled group or to prepare it for the ultimate form of ostracization, i.e., physical annihilation. The first scenario fits the evolution of the black American 'hyperghetto' in the post-Civil Rights era: having lost its function of reservoir of unskilled labor power, it has become symbolically linked to the hypertrophied carceral system of the United States by a triple relationship of structural homology, functional surrogacy, and cultural fusion (Wacquant 2004). The second scenario is that implemented by Nazi Germany, which revived the *Judenghetto* between 1939 and 1944, first, to impoverish and concentrate Jews with a view toward relocation and later, when mass deportation turned out to be impractical, to funnel them toward extermination camps (Friedman 1980, Browning 1986).

The unchecked intensification of its exclusionary thrust suggests that the ghetto might be most profitably studied not by analogy with urban slums, lower-class neighborhoods, and immigrant enclaves but alongside the reservation, the refugee camp, and the prison, as belonging to a broader class of institutions for the forced confinement of dispossessed and dishonored groups. It is not by happenstance that the Bridewell of London (1555), the *Zuchthaus* of Amsterdam (1654), and the *Hospital général* of Paris (1656), designed to instill the discipline of wage work to able-bodied vagrants, beggars, and criminals via incarceration, were invented around the same time as the Jewish ghetto. And that today's sprawling refugee camps in Rwanda, Sri Lanka, and in the occupied territories of Palestine look ever more like a cross between the ghettos of late medieval Europe and gigantic gulags.

Bibliography

Abu-Lughod J L 1980 *Rabat: Urban Apartheid in Morocco*. Princeton University Press, Princeton, NJ

Browning C R 1986 Nazi ghettoization policy in Poland, 1939–1941. *Central European History* **19**(4): 343–68

Caldeira T 2000 *City of Walls: Crime, Segregation and Citizenship in São Paulo*. University of California Press, Berkeley, CA

Clark K B 1965 *Dark Ghetto: Dilemmas of Social Power*. Harper, New York

DeVos G, Chung D 1981 Community life in a Korean ghetto. In: Lee C, De Vos G (eds.) *Koreans in Japan: Ethnic Conflict and Accomodation*. University of California Press, Berkeley, CA, Chap. 10, pp. 225–51

DeVos G, Wagatsuma H (eds.) 1966 *Japan's Invisible Race: Caste in Culture and Personality*. University of California Press, Berkeley, CA

Drake St C, Cayton H R, [1945] 1993 *Black Metropolis: A Study of Negro Life in a Northern City*. University of Chicago Press, Chicago

Friedman P 1980 The Jewish ghettos of the Nazi era. In: *Roads to Extinction: Essays on the Holocaust*. The Jewish Publication Society of America, New York, Chap. 3, pp. 59–87

Gay R 1992 *The Jews of Germany: A Historical Portrait*. Yale University Press, New Haven, CT

Hane M 1982 *Peasants, Rebels, and Outcastes: The Underside of Modern Japan*. Pantheon, New York

Hannerz U 1969 *Soulside: Inquiries into Ghetto Culture and Community*. Columbia University Press, New York

Hirsch A 1983 *Making the Second Ghetto: Race and Housing in Chicago 1940–1970*. Cambridge University Press, Cambridge, repub. University of Chicago Press

Johnson J W [1930] 1981 *Black Manhattan*. Da Capo, New York

Johnson P 1987 Ghetto. In: *A History of the Jews*. HarperPerennial, New York, pp. 230–310

Kerner Commission [1968] 1989 *The Kerner Report. The 1968 Report of the National Advisory Commission on Civil Disorders*. Pantheon, New York

Levine M P 1979 Gay Ghetto. *Journal of Homosexuality* **4**(4) (Summer). Reprinted in expanded form as '"YMCA": The Social Organization of Gay Male Life,' In: *Gay Macho: The Life and Death of the Homosexual Clone*. New York University Press, New York, pp. 30–54

Massey D, Denton N 1992 *American Apartheid: Segregation and the Making of the Under! class*. Harvard University Press, Cambridge

Mingione E (ed.) 1996 *Urban Poverty and the 'Underclass'*. Basil Blackwell, Oxford

Murray S O 1979 The institutional elaboration of a quasi-ethnic community. *International Review of Modern Sociology* **9** (July): 165–77

Nelli H S 1970 *Italians in Chicago: A Study in Ethnic Mobility*. Oxford University Press, New York

Noiriel G 1988 *Le Creuset français*. Editions du Seuil, Paris

Osofsky G 1971 *Harlem: The Making of a Ghetto—Negro New York, 1890–1930*, 2nd edn. Harper and Row, New York

Perlman J 1976 *The Myth of Marginality: Urban Poverty and Politics in Rio De Janeiro*. University of California Press, Berkeley, CA
Pétonnet C 1982 *Espaces habités. Ethnologie des banlieues*. Galilée, Paris
Philpott T L 1978 *The Slum and the Ghetto: Neighborhood Deterioration and Middle-Class Reform, Chicago 1880–1930*. Oxford University Press, New York
Portes A, Stepick A 1993 *City on the Edge: The Transformation of Miami*. University of California Press, Berkeley
Quijano A 1968 *Notas sobre el concepto da marginalidad social*. Commission for Latin American Report, Santiago (Chile)
Sennett R 1994 Fear of touching. In: *Flesh and Stone: The Body and the City in Western Civilization*. W.W. Norton, New York, Chap. 7, pp. 212–51
Spear A H 1968 *Black Chicago: The Making of a Negro Ghetto, 1890–1920*. University of Chicago Press, Chicago, IL
Stow K R 1992 *Alienated Minority: The Jews of Medieval Europe*. Harvard University Press, Cambridge, MA
Wacquant L 1998 A black city within the white: Revisiting america's dark ghetto. *Black Renaissance* **2**(1): 141–51
Wacquant L 2004 *Deadly Symbiosis: Race and The Rise of Neoliberal Penality*. Polity Press, Cambridge
Wacquant L *Urban Outcasts: Toward a Sociology of Advanced Marginality*. Polity Press, Cambridge (to be published)
Ward D 1989 *Poverty, Ethnicity, and the American City 1840–1925*. Cambridge University Press, Cambridge
Western J 1981 *Outcast Cape Town*. University of Minnesota Press, Minneapolis
Wilson W J 1987 *The Truly Disadvantaged: The Inner City, The Underclass and Public Policy*. University of Chicago Press, Chicago
Wirth L 1928 *The Ghetto*. University of Chicago Press, Chicago
Wirth L [1956] 1964 The ghetto. In: Reiss A J, Jr. (ed.) *On Cities and Social Life*. The University of Chicago Press, Chicago, Chap. 5 pp. 84–98
Zhou M 1992 *Chinatown: The Socioeonomic Potential of an Urban Enclave*. Temple University Press, Philadelphia

Loïc Wacquant

Copyright © 2004 Elsevier Ltd.
All rights reserved.

 ISBN: 0-08-043076-07

H

Habit: History of the Concept

A family of expressions found in many ordinary languages, habit and its linguistic cognates compose a malleable vocabulary that, for over two millennia, has been incorporated into a wide range of intellectual projects in the domains of the social sciences and beyond. Commonly used for the purpose to describing various *recurrent and relatively automatic features* of human conduct, 'the discourse of habit is one of the great, albeit largely unacknowledged, cultural legacies of Western civilization' (J. M. Thomas 1993, p. 8). Until the early twentieth century, the history of this discourse was roughly coextensive with Western intellectual history itself and scarcely more amenable to brief summary, though broad phases may be distinguished.

1. *The Concept of Habit*

In the course of its long history, habit has been a polysemous term. Common to most of its meanings and usages, however, has been the conceptualization of habit as a more-or-less self-actuating disposition or tendency to engage in previously adopted or acquired forms of action. When these forms of action are differentiated in terms of their content, distinctions then arise between moral habits, cognitive habits, emotional habits, and motor habits.

Of greater historical importance, however, has been another basis of differentiation. The forms of actions whose repetition is designated as habit may range from simple and circumscribed to complex and generalized. Among the social thinkers who have invoked the concept, habit has most frequently been applied to recurrent patterns of activity of moderate complexity and generality, i.e., to habits of economic, political, religious, and domestic conduct, habits of obedience to rules and to rulers, habits of sacrifice, disinterestedness, and restraint, and so on. But, at various points in the history of the concept, habit has referred primarily to certain simpler, or at least more elemental, behavioral regularities, such as habits of talking, walking, eating, dressing, routine problem-solving, etc.

At the opposite extreme, habit has sometimes been applied to the highly generalized and durable disposition that suffuses an individual's mode of action and discernment in entire spheres of activity or even throughout an entire lifetime. in which case the concept has been largely synonymous with 'character,' a term encapsulating the whole manner or mold of a human personality. When employed in this extended sense, habit has often been spoken of as *habitus*. That habit has historically taken on this full range of meanings serves as a precaution against projecting on to all instances of the concept connotations only applicable in restricted cases. The contemporary supposition that habit is a fixed, mechanical reaction to discrete stimuli and thus devoid of subjective meaning is, for example, a view that has historically been associated mainly with habit in the sense of simple behavioral regularities, not with more complex habitual forms of action.

At no point has the vocabulary of habit been a free-standing one; to a significant degree, its usage has varied because of the different ways in which habit has been opposed to or aligned with other concepts. At a decisive period in its history, habit was counterposed to instinct, such that all *non*-instinctual behavior was regarded as necessarily habitual. Prior to and subsequent to this period, however, habit was ordinarily set over against purposive, reflective types of action that involve the selection of means and ends according to various normative standards; in this case, habit contrasts with those types of action presupposed by most twentieth-century social-scientific *reflective models* of human conduct. At still other moments, however, habit has been taken not as the opposite of, but as the precondition for reflexive action. The same flexibility is evident with regard to moral conduct, which has sometimes been contrasted with habitual action, other times held to depend upon it. Further, conceptual variability has appeared at the aggregate level, as the similar habits of the several members of a social group have sometime been equated with custom, while other usages reserve custom for practices done in compliance with social sanctions and hence distinct from forms of action that have become somewhat more automatic owing to repetition or other habit-formation processes. (For a fuller treatment of the issues in this section, see Funke 1958, Camic 1986.)

2. *Historical Course*

2.1 *Antiquity to the Early 1800s*

The discourse of habit came naturally to Greek philosophers as they opened pathways for later social

thinkers. Aristotle's *Ethics* (Book 2, Chapter 1) (Aristotle [1973]) laid it down that 'none of the moral virtues arise in us by nature,' that 'we are adapted by nature to receive them and are made perfect by habit'; and the same discourse molded as well to the varied teachings of Plato, Plotinus, Quintilian, Aquinas, Luther, Montaigne, Pascal, and Wolf, among a great number of others (Funke 1958).

In the Age of the Enlightenment, amid determined efforts to launch a science of man, the concept of habit was subject to more systematic formulations which reaffirmed habit's vast empirical role but appraised this role differently. In France, Helvetius celebrated habit as the wellspring of public and private morality and accepted it as a 'principle by which [humans everywhere] are actuated' (Helvetius [1758] 1807, p. 57); in Scotland, Ferguson declared habit the force 'by which the good or bad actions of men remain with them and become part of their characters' (quoting Hill 1996, p. 210); while from Prussia Kant averred that 'all acquired habits are objectionable'—that 'virtue is moral strength in the pursuit of one's duty, which should never be a matter of habit' (Kant [1798], 1978, pp. 32–34). Hume, Berkeley, Reid, Condillac, Rousseau, Hegel, and the German idealists were among the many thinkers who took part in the discussion. (On the thinkers named here and others, see Funke 1958, Camic 1986.)

2.2 A Century of Ferment

From the middle of the nineteenth century until the early decades of the twentieth century, the language of habit was even more ubiquitous in European and American thought. It found expression in the work of a wide range of social thinkers that included James and John Stuart Mill, Spencer, Bagehot, Marshall, Bradley, and Bosanquet in England, Comte and LePlay in France, and Nietzsche, Simmel, Tönnies, and Husserl in Germany. In the United States, the discourse of habit was also drawn upon extensively by, among others, popular reformers, churchmen, and educators, pragmatist philosophers (Peirce, James, and Dewey), and leading thinkers involved in establishing the social sciences as independent academic disciplines (Sumner, Giddens, Veblen, Thomas).

Typical of the plaint ways in which the discourse was used were the views held by Emile Durkheim and Max Weber as they laid the foundations for modern sociology. Throughout his career, Durkheim took the position that 'habits … are the real forces which govern us' (Durkheim [1905–6] 1956, p. 152). For him, as for the American pragmatists, human reflection is a process that generally 'overloads or paralyzes action'; only 'when habit is disrupted,' and the individual or collective 'being is … at a cross-roads situation [facing] a whole range of possible solutions,' does reflection arise in order to restore 'equilibrium' and allow 'habits of all kinds' to reassert themselves (Durkheim [1913–14] 1983, pp. 38, 79–80). Accordingly, Durkheim's plans for the moral regeneration of modern society relied heavily on habit: on developing in children 'the habit of self control and restraint' and 'habits of group life'; on cultivating in occupational specialists 'an absolute regularity in habits' and interactional 'habits [that], as they grow in strength, are transformed into rules of conduct'; and on instilling in society's leaders 'a habitus of moral being' (Durkheim 1902–3 [1961], pp. 149, 249, 1893 [1984], pp. 187, 302, 1904–5 [1977], p. 29).

Driven by a different intellectual agenda, Weber held similar views. Occupied with the various forms of reflective action, he nonetheless affirmed the importance of habit, holding that 'the further we go back in history, the more we find that conduct, and particularly social action, is determined in an ever more comprehensive sphere exclusively by the disposition toward the purely habitual': that 'individuals are still markedly influenced' by this force 'even today,' with 'the great bulk of all everyday action [approaching an] almost automatic reaction to habitual stimuli which guide behavior in a course which has been repeatedly followed' (Weber [1922] 1978, pp. 320, 337, 25 [modified translation]). The vocabulary of habit thus figured significantly into Weber's treatment of economic, political, and religious traditionalism, into his writings on economic, political, legal, and communal action in the modern world, and even into his analysis of the process by which Protestantism promoted the spirit of modern capitalism. Conceptualizing the spirit of capitalism as a 'particular habitus,' Weber credited the early Calvinists with creating, 'out of their religiously conditioned family traditions and from the religiously influenced life-style of the environment, [a distinctive] 'habitus' among individuals which prepared them in specific ways to live up to the specific demands of early modern capitalism' (Weber [1910a] 1968, [1910b] 1978, p. 1124).

Yet, as social thinkers mobilized the language of habit in these ways, contemporary natural scientists increasingly moved to appropriate it to refer to a far more elementary range of activities, both human and subhuman. This, for example, was the practice of Lamarck, and Darwin followed the model, regularly speaking of the flowering habits of plants, the feeding habits of insects, and the flying habits of pigeons (Darwin [1859] 1964, pp. 11, 183, [1872] 1975, pp. 29–31). This was also the practice of many European physiologists and psychologists as they turned attention to the study of the neural foundations of simple animal and human reflexes, widely using the term habit to describe well-established neural pathways (Fearing 1930). This usage gained great currency among late-nineteenth and early-twentieth century American psychologists, convinced with James that even 'the most complex habits [are] nothing but

concatenated discharges in the nerve-centres, due to the presence there of systems of reflex paths' (James [1890] 1950, p. 108). Extending the same view into their own domain, turn-of-the-century American social scientists ascribed vital significance to habits rooted in 'the physiology of the nervous system,' convinced 'that education and social control are largely dependent upon our ability to establish habits in ourselves and others' (quoting Park [1915] 1969, see also Camic 1986).

2.3 The Decline of Habit

The culmination of these developments came with the behaviorist revolution which John Watson launched in American psychology during World War One and which overran the field in the next two decades. Taking the position that *every* 'mode of acting ... not belonging to man's hereditary equipment must be looked upon as a habit,' conceived physiologically in terms of 'muscular and glandular changes which follow upon a given stimulus,' Watson and his followers touted habit-psychology as the way 'to guide society ... towards the control of group and individual behavior' (Watson 1919, pp. 270, 14, 2–3). From this point of view, the science required for the study of the social world was 'not [a social science like] sociology, but psychology,' which derives from 'biology, chemistry, and the other natural sciences' (Allport 1927, pp. 167–168).

The challenge to the legitimacy of their own fledgling academic disciplines now manifest, social scientists swiftly recoiled from the language of habit. Countering the behaviorist effort to eliminate the reflective forms of action that had long stood alongside habitual forms, sociologists of the inter-war period led an aggressive campaign to rescue purposive human conduct, deliberately forsaking their previous use of the concept of habit. This campaign was inaugurated by Thomas (a former proponent of the concept) and Znaniecki, whose work attacked the 'behavioristic school' for its 'indistinct [application] of the term 'habit' to [all] uniformities of behavior' and urged that 'habit' ... be restricted to the biological field, [since] it involves no conscious, purposeful regulation of [conduct], but merely ... is unreflective, [whereas] the uniformity [that constitutes social life] is not a uniformity of organic habits but of consciously followed rules' (W. I. Thomas and Znaniecki [1918] 1958, pp. 1849–1852). Within a generation, this view became dogma, displacing the discourse of habit with the language of 'attitude,' which Thomas and Znaniecki brought to the fore and which subsequently furnished American social scientists with a new vocabulary with which to encompass regular tendencies in human conduct while holding to reflective action models (Fleming 1967, Camic 1986). Parallel movements away from the concept of habit occurred in European thought (as even Durkheim and Weber excised the concept from sociology) and in European and American literature (Fisher 1973, J. M. Thomas 1993). In the aftermath of this turnabout, at least outside the natural sciences, the language of habit became dormant for the half century from the late 1920s to the late 1970s, with very few exceptions. In the heyday of the behaviorist revolution, Dewey resuscitated a broader conception of habit, and Merleau-Ponty in France subsequently did likewise (Kestenbaum 1977, see also Berger and Luckmann 1966); but such efforts did not stem the behaviorist view of habit, nor the resulting widespread reaction against the concept. Only among psychologists working in the behaviorist tradition, and among a few economists focused on consumer behavior, did the language of habit remain a presence (Deaton and Muelbaur 1980, Smith 1986). Otherwise, social scientists essentially abandoned the discourse, conceiving of all human action as a reflective process of fitting means to ends according to various normative standards.

3. Recent Developments

The last quarter of the twentieth century brought little fundamental alteration in the stance that had been taken toward the language of habit during the previous half century. By and large, reflective conceptualizations of human action have continued to dominate the social sciences. This dominance has been less complete, however, as the concept of habit, understood in non-behaviorist terms that simultaneously accommodate reflective action, has periodically been revived: by social psychologists (Ouellette and Wood 1998), by a wider range of economists (Scitovsky 1976, Campbell and Cochrane 1995), and by sociological theorists, some developing earlier pragmatist and phenomenological approaches (Ostrow 1990, Joas [1992] 1996), others aligned also with the traditions of Durkheim and Weber, most notably Bourdieu ([1980] 1990), in whose writings 'habitus' receives extended attention (see also Giddens 1984). Still largely disconnected from one other, these lines of work open promising future possibilities for the eventual transformation of exclusively reflective models of human conduct in the social world.

See also: Habituation; Tradition, Anthropology of; Traditions in Sociology; Traditions: Social; Weber, Max (1864–1920)

Bibliography

Allport G W 1927 The nature of institutions. *Social Forces* **6**: 167–79

Aristotle [1973] McKeon R (ed.) *Introduction to Aristotle*, 2nd edn. University of Chicago Press, Chicago

Berger P L, Luckmann T 1966 *The Social Construction of Reality*. Doubleday, New York
Bourdieu P [1980] 1990 *The Logic of Practice*. Stanford University Press, Stanford, CA
Camic C 1986 The matter of habit. *American Journal of Sociology* **91**: 1039–87
Campbell J Y, Cochrane J H 1995 *By Force of Habit*. National Bureau of Economic Research, Cambridge, MA
Darwin C [1859] 1964 *On the Origin of Species*. Harvard University Press, Cambridge, MA
Darwin C [1872] 1975 *The Expression of Emotions in Man and Animals*. University of Chicago Press, Chicago
Deaton A, Muelbaur J 1980 *Economics and Consumer Behavior*. Cambridge University Press, New York
Durkheim E [1893] 1984 *The Division of Labor in Society*. Free Press, New York
Durkheim E [1902–3] 1961 *Moral Education*. Free Press, New York
Durkheim E [1904–5] 1977 *The Evolution of Educational Thought*. Routledge and Kegan Paul, London
Durkheim E [1905–6] 1956 *Education and Society*. Free Press, Glencoe, IL
Durkheim E [1913–4] 1983 *Pragmatism and Sociology*. Cambridge University Press, Cambridge, UK
Fearing F 1930 *Reflex Action*. Williams and Williams, Baltimore, MD
Fisher P 1973 The failure of habit. In: Engel M (ed.) *Uses of Literature*. Harvard University Press, Cambridge MA, pp. 3–18
Fleming D 1967 Attitude: History of a concept. *Perspectives in American History* **1**: 287–365
Funke G 1958 Gewohnheit. *Archive für Begriffsgeschichte* **3**: 1–606
Giddens A 1984 *The Constitution of Society*. University of California Press, Berkeley, CA
Helvetius C A [1758] 1807 *De l'Esprit*. Jones, London
Hill L 1996 Anticipations of nineteenth and twentieth century social thought in the work of Adam Ferguson. *Archives européennes de sociologie* **37**: 203–28
James W [1890] 1950 *The Principles of Psychology*. Dover, New York
Joas H [1992] 1996 *The Creativity of Action*. University of Chicago Press, Chicago
Kant I [1798] 1978 *Anthropology from a Pragmatic Point of View*. Southern Illinois University Press, Carbondale, IL
Kestenbaum V 1977 *The Phenomenological Sense of John Dewey*. Humanities Press, Atlantic Highlands, NJ
Ostrow J M 1990 *Social Sensitivity*. State University of New York Press, Albany, NY
Ouellette J A, Wood W 1998 Habit and intention in everyday Life. *Psychological Bulletin* **124**: 54–74
Park R E [1915] 1969 Man not born human. In: Park R E, Burgess E W (eds.) *Introduction to the Science of Sociology*. University of Chicago Press, Chicago, pp. 82–4
Scitovsky T 1976 *The Joyless Economy*. Oxford University Press, New York
Smith L D 1986 *Behaviorism and Logical Positivism*. Stanford University Press, Stanford, CA
Thomas J M 1993 Figures of habit in William James. *The New England Quarterly* **66**: 3–26
Thomas W I, Znaniecki F [1918] 1958 *The Polish Peasant in Europe and America*. University of Chicago Press, Chicago
Watson J B 1919 *Psychology from the Standpoint of a Behaviorist*. Lippincott, Philadelphia
Weber M [1910a] 1968 Antikritisches zum 'Geist' des Kapitalismus. In: Winckelmann J (ed.) *Die Protestantische Ethik II*. Siebenstern, Munich, pp. 149–87
Weber M [1910b] 1978 Anticritical last word on 'The Spirit of Capitalism.' *American Journal of Sociology* **83**: 1105–31
Weber M [1922] 1978 *Economy and Society*. University of California Press, Berkeley, CA

C. Camic

Habituation

> If a drop of water falls on the surface of the sea just over the flower-like disc of a sea anemone, the whole animal contracts vigorously. If, then, a second drop falls within a few minutes of the first, there is less contraction, and finally, on the third or fourth drop, the response disappears altogether. Here in this marine polyp with the primitive nerve net is clearly exhibited one of the most pervasive phenomena of the animal kingdom—decrement of response with repeated stimulation. Almost every species studied, from amoeba to man, exhibits some form of response decrement when the stimulus is frequently repeated or constantly applied. The ubiquity of the phenomenon plus its obvious survival value suggests that this kind of plasticity must be one of the most fundamental properties of animal behaviour. (Sharpless and Jasper 1956, p. 655)

With these words, Sharpless and Jasper introduced their now classic study of habituation of the electroencephalogram (EEG) arousal response. Their study had a profound impact on behavioral neuroscience avnd served to kindle the interest of neural scientists in brain mechanisms of habituation. *Habituation* is a decrease in response to repeated stimulation, and *sensitization* is an increase in response as a result of stimulation. Virtually all behavioral responses, as well as autonomic and EEG responses and signs of arousal, habituate rapidly and to a profound degree. The processes responsible for habituation of behavioral responses probably occur largely after the primary sensory pathways of the brain and before the motor neurons that move muscles—somewhere between initial processing of sensory information and the execution of decisions to respond. Sensitization, the increase in response that often occurs when strong stimuli are given, may be similarly localized, although some evidence indicates that sensitization may occur in motor systems and motor neurons.

In early work on brain processes of habituation, it was thought that virtually all measures of neural activity might habituate, ranging from first-order sensory nuclei to the cerebral cortex. An early experiment to report evoked response habituation was that of Hernández-Peon et al. (1957); they recorded responses to click stimulation at several levels of the

auditory system. Trains of clicks were delivered once every two seconds for long periods of time and evoked responses of the cochlear nucleus (the first relay of the auditory system) were reported to habituate.

Careful studies by Worden and associates (see Marsh et al. 1962) demonstrated that click-evoked responses at the cochlear nucleus do not show habituation. Instead, amplitudes of responses at this first relay nucleus in the auditory system are rigidly controlled by the physical properties of the sound stimulus. Because of acoustic factors, the intensity of a sound is often weaker at the floor of a test cage. If an animal gradually became bored and rested his head on the floor, the cochlear nucleus-evoked response to click would decrease because of reduced sound intensity.

In an extensive and careful study, Wickelgren (1968) found that habituation of evoked responses to auditory stimuli was not prominent below the level of the thalamus and cortex in the auditory system. This appears to be a general finding in all sensory systems—the lower ascending relay nuclei do not exhibit habituation. It would be most surprising if there were marked habituation in sensory systems. If our sensory systems habituated to sensory stimuli, we would rapidly cease to have sensation—a most unadaptive state of affairs. However, behavioral responses and phenomena that seem to be related to the probability of behavioral responding, such as EEG arousal and autonomic signs of orienting, show marked habituation. It is very adaptive to cease responding to arousing stimuli that have no particular significance.

Historically, such processes as habituation and sensitization were viewed simply as reflexes and not as phenomena of learning and memory. However, the broad definition of learning as a change in behavior as a result of experience certainly encompasses both phenomena. When this was recognized, analysis of neural mechanisms became 'legitimized' (Sharpless and Jasper 1956, Thompson and Spencer 1966). Habituation and sensitization are now seen as examples of nonassociative learning and memory. Greater progress has been made in analysis of the basic neuronal mechanisms underlying habituation than in more complex forms of behavioral plasticity such as classical and instrumental conditioning. Major reasons for this progress lie in the widespread occurrence of habituation, in its simplicity, and in the reliability with which the behavioral properties of habituation can be specified and determined. Habituation occurs both in spinal reflexes of the vertebrate (Sherrington 1906) and for responses of simple animals, as well as in complex human behavior. Consequently, neural mechanisms of habituation can be studies in simplified systems.

As noted earlier, behavioral habituation is simply response decrement as a result of repeated stimulation. For example, if a moderate electric shock is delivered to the foot of a normal but restrained animal for the first time, the animal will show marked struggling and leg flexion movements. However, if the stimulus is repeated a few times, the responses decrease considerably and may even disappear. Habituation is commonly differentiated from response decrements due to trauma, growth, aging, etc. In the example above, if the shock to the foot is withheld for a period of time after habituation and then given again, the response will reappear with normal strength. This spontaneous recovery is a common control used to show that habituation is a *reversible* response decrement.

Response decrements due to altered sensitivity of receptors are often termed *receptor adaptation*, and response decrements due to decremental effects occurring at the neuromuscular junctions of muscles are often termed *muscular fatigue*. Neither of these processes is considered to be habituation; it is a central rather than a peripheral process. Almost all behavioral responses of the intact organism initially elicited by a given stimulus, ranging in complexity from leg flexion to exploratory and play behavior, show habituation with repeated stimulation.

Response sensitization may be defined simply as an increased response strength to a given stimulus as a result of some other, usually strong, stimulus. The galvanic skin reflex (GSR) is a good case in point. Suppose we were to measure the GSR activity of sweat glands in the palm of the hand in response to a sound of moderate intensity. The response would probably be of relatively low amplitude and somewhat irregular in occurrence. If a strong shock were delivered to the skin, the sound alone now would evoke a much larger GSR response; the response would have been sensitized. The response increase due to sensitization is distinguished from conditioning in that no pairing of shock and sound is necessary to produce the effect. Sensitization itself typically habituates; after a number of shock presentations the response to the sound decreases to the initial weak level. In fact, it seems to decay spontaneously following a sensitizing stimulus.

The study of habituation has a long history; some of its basic properties were described in classic works by Jennings (1906), Humphrey (1933), Prosser and Hunter (1936), and Harris (1943). Thompson and Spencer (1966) surveyed the by then very extensive behavioral literature on habituation and identified some nine basic parametric properties or characteristics exhibited by behavioral habituation.

1. *Properties of Habituation*

(a) Given that a particular stimulus elicits a response, repeated applications of the stimulus result in decreased response (habituation). The decrease is usually a negative exponential function of the number of stimulus presentations.

(b) If the stimulus is withheld, the response tends to recover over time (spontaneous recovery).

(c) If repeated series of habituation training and spontaneous recovery are given, habituation becomes successively more rapid (this might be called potentiation of habituation).

(d) Other things being equal, the more rapid the frequency of stimulation, the more rapid and/or more pronounced is habituation.

(e) The weaker the stimulus, the more rapid and/or more pronounced is habituation. Strong stimuli may yield no significant habituation.

(f) The effects of habituation training may proceed beyond zero or asymptotic response level.

(g) Habituation of response to a given stimulus exhibits stimulus generalization to other stimuli.

(h) Presentation of another (usually strong) sensitizing stimulus results in recovery of the habituated response (dishabituation or sensitization).

(i) Upon repeated application of the sensitizing stimulus, the amount of dishabituation produced habituates (this might be called habituation of sensitization).

It is important to emphasize that these characteristics of habituation are simply the empirically observed properties in a wide variety of behavioral studies on many species and as such do not imply particular theories or inferences. These properties have come to serve as the detailed definition of habituation. Thus, any particular stimulus-response system is said to exhibit habituation if the response decrement to repeated stimulation exhibits these properties.

The most potent variables determining the course of habituation are stimulus frequency and intensity. The frequency range must be behaviorally relevant. Very high frequencies can yield fatigue, potentiation, and other processes. The range of stimulus frequencies over which habituation occurs, i.e., behavioral response decrements exhibiting the properties noted above, is an empirical question dependent upon the organism, the stimulus, and the response system engaged. For the orienting or arousal response, the effective range may be from minutes to hours to days; for the spinal flexion reflex, the range is over seconds.

The fact that rate and degree of habituation are inversely related to stimulus intensity (property (e)) is true if the same intensity is used for test and habituating stimuli (the test stimulus is given sufficiently infrequently that it does not by itself cause habituation). However, if the test and habituating stimuli differ in intensity, more complex effects can occur (Davis and Wagner 1968).

2. *Dishabituation vs. Sensitization*

Thompson and Spencer (1966) employed habituation of the spinal flexion reflex, first used as a model system to study habituation by Prosser and Hunter (1936), to characterize the detailed properties of habituation and analyze possible neuronal mechanisms. They discovered that the process of 'dishabituation' caused by presentation of a strong extra stimuli was not in fact dishabituation. That is, it did not disrupt the process of habituation at all. Instead, it added a superimposed, independent process of increased excitability, namely sensitization. The fact that dishabituation, *per se*, does not appear to exist as such, but rather reflects a separate process of sensitization, appears to hold for all vertebrate preparations studied, ranging from reflexes to the startle response to EEG arousal and the orienting response. One exception to this rule has been reported for a monosynaptic system in the invertebrate *Aplysia* (Marcus et al. 1988). However, the neuronal mechanisms underlying the processes of sensitization/dishabituation appear to differ in vertebrates vs. the *Aplysia* circuit. In the latter, 'dishabituation' appears to be a presynaptic process, occurring at the sensory neuron terminals on the motor neurons (see Sect. 3 and Kandel 1976); whereas, at least in spinal reflexes, sensitization can involve increased excitability in motor neurons (Thompson and Spencer 1966). Actually, recent evidence suggests that changes may occur in motor neurons in the *Aplysia* circuit as well (Murphy and Glanzman 1997).

The fact that dishabituation does not exist as such, but rather reflects a separate independent process of sensitization in vertebrate preparations led Groves and Thompson (1970) to develop the dual-process theory of habituation: the outcome of repeated stimulus presentations is due to the net interaction of the inferred processes of habituation and sensitization. This theory is able to account for an astonishingly wide range of behavioral phenomena (see, e.g., Peeke and Herz 1973).

A somewhat differently construed and widely influential theory of habituation is the stimulus-model comparator theory developed in Russia by Evgeny Sokolov (1963), based largely on data from the human orienting response. In this view, repetition of a given stimulus results in the development of a neuronal 'model' of the stimulus in forebrain structures. If a repetitive stimulus matches the model, then there is little output, only the habituated response, which could be no response, i.e., no EEG arousal. However, if there is a mismatch, then the difference is amplified via the reticular formation and a larger response results.

The dual-process theory is more successful in accounting for basic phenomena of habituation and sensitization in a wide range of species; the model-comparator theory can more easily account for more complex phenomena like the 'missing stimulus' effect (increased response to the absence of one presentation of the habituating stimulus, a phenomenon sometimes seen in EEG arousal but not in reflex systems). But note that a process of synaptic depression could underlie the formation of the neuronal model of the

stimulus in the model-comparator theory; the theories are at different levels and not incompatible. A view somewhat similar to that of Sokolov was developed by Wagner (1979) to account for long-term habituation. Thus, the central model of the habituating stimulus is considered as a relatively long-term memory trace, yielding the basic properties of a learned response. In their classic study of human EEG arousal habituation, Sharpless and Jasper (1956) stressed the distinction between short-term and long-term habituation.

As noted earlier, primary sensory afferent system neurons do not exhibit habituation. This makes perfect sense; in order to 'decide' how much to respond to a stimulus it is necessary to know what the stimulus is. Virtually all behavioral responses elicited by stimuli show habituation to repeated stimulation under at least some conditions. Central brain processes that are behavior-related like cortical EEG arousal show clear habituation, as do brain motor systems. Davis and associates completed a comprehensive analysis of the startle circuit and startle habituation in the rat (Davis 1984). The circuit includes auditory relay nuclei: the cochlear nuclei and the nuclei of the lateral lemiscus; a nucleus in the brainstem reticular system (nucleus reticularis pontis caudalis) and its projections to brainstem and spinal cord premotor and motor nuclei. Evidence suggests that habituation to the acoustic startle stimulus occurred before the motor systems, i.e., at the nucleus reticularis pontis caudalis (Davis et al. 1982). The reticular nucleus is a 'bridge' nucleus between sensory and motor systems.

3. Neural Mechanisms

Unlike associative learning, in habituation the stimulus-response neuronal circuit is hardwired. In systems where some degree of analysis of neuronal mechanisms is possible, the neural substrate of habituation occurs within the stimulus-response circuit itself and is not due to any extrinsic circuit action like external inhibition acting on the habituating circuit.

Greatest progress in analyzing neural mechanisms of habituation and sensitization has come from simplified neuronal systems. Two of these are the gill withdrawal reflex of the mollusk *Aplysia* and vertebrate spinal reflexes. Kandel (1976) and his many associates completed a detailed analysis of the processes of habituation and sensitization in the *Aplysia* monosynaptic circuit. They were able to show that habituation was a presynaptic process involving a decrease in the probability of neurotransmitter release as a result of repeated activation. Similarly, a process of synaptic depression is responsible for habituation of the tail-flip escape response in the crayfish, another very productive invertebrate model system (Krasne 1969). In the *Aplysia* model, synaptic decrement is due to a persisting decrease in calcium ion influx into the terminals. Sensitization, elicited for example by strong shock to the tail of the animal, was due to activation of a polysynaptic circuit that acted presynpatically on the sensory neuron terminals, the same terminals that habituated. The effects on calcium channels are due in turn to complex and not fully understood biochemical cascades within the neurons. Kandel and associates have also described a long-term sensitization effect involving structural changes in the synapses that they have used as a model for long-term memory storage (Hawkins and Kandel 1984).

In the vertebrate spinal cord, the flexion reflex is polysynaptic. Thompson and Spencer (1966) were able to rule out any changes in excitability in the sensory neuron terminals and in the motor neurons themselves during habituation. However, sensitization did involve increases in motor neuron excitability, as noted above. They suggested, but could not prove, that a process of low-frequency depression in the interneurons was responsible for habituation, a process analogous to the presynaptic depression process demonstrated by Kandel and associates in *Aplysia*. Using a monosynaptic pathway in the isolated frog spinal cord, Farel and Thompson (1976) demonstrated a presynaptic process for habituation. The story concerning neuronal mechanisms of habituation is particularly satisfying because the same process of synaptic depression appears to underlie habituation in both invertebrate and vertebrate systems. This simple process of synaptic decrement can account for a wide range of behavioral phenomena of habituation.

The study of habituation continues to be a most active field of research, ranging from responses of nematode *C. elegans* (Wicks and Rankin 1996) to the startle response in mammals (Pilz and Schnitzler 1996) to behaviors in the human fetus and infant (Leader 1995, Kaplan et al. 1990).

See also: Arousal, Neural Basis of; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Eyelid Classical Conditioning; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of

Bibliography

Davis M 1984 The mammalian startle response. In: Eaton R C (ed.) *Neural Mechanisms of Startle Behavior*. Plenum, New York, pp. 287–351

Davis M, Parisi T, Gendelman D S, Tischler M D, Kehne J H 1982 Habituation and sensitization of electrically elicited 'startle' reflexes. *Science* **218**: 688–9

Davis M, Wagner A R 1968 Startle responsiveness after habituation to different intensities of tone. *Psychonomic Science* **12**: 337–8

Farel P B, Thompson R F 1976 Habituation of a monosynaptic response in frog and spinal cord: Evidence for a presynaptic mechanism. *Journal of Neurophysiology* **39**: 661–6

Groves P M, Thompson R F 1970 Habituation: A dual-process theory. *Psychological Review* **77**: 419–50

Harris J D 1943 Habituatory response decrement in the intact organism. *Psychological Bulletin* **40**: 358–422
Hawkins R D, Kandel E R 1984 Is there a cell-biological alphabet for simple forms of learning? *Psychological Review* **91**: 373–91
Hernandez-Peon R, Jouvet M, Scherrer H 1957 Auditory potentials at cochlear nucleus during acoustic habituation. *Acta neurologica latinoamericana* **3**: 144–56
Humphrey G 1933 *The Nature of Learning*. Harcourt, Brace, New York
Kandel E R 1976 *Cellular Basis of Behavior: An Introduction of Behavioral Biology*. W. H. Freeman, San Francisco
Kaplan P S, Werner J S, Rudy J W 1990 Habituation, sensitization and infant visual attention. In: Rovee-Collier C, Lipsitt L P (eds.) *Advances in Infancy Research*. Ablex Publishing, Norwood, NJ, Vol. 6, pp. 61–109
Krasne F B 1969 Excitation and habituation of the crayfish escape reflex: The depolarization response in lateral giant fibers of the isolated abdomen. *Journal of Experimental Biology* **50**: 29–46
Leader L R 1995 The potential value of habituation in the prenate. In: Lecanuet J-P, Fifer W P, Krasnegor N A, Smotherman W P (eds.) *Fetal Development: A Psychobiological Perspective*. Lawrence Erlbaum, Hillsdale, NJ, pp. 383–404
Marcus E A, Nolen T G, Rankin C H, Carew T J 1988 Behavioral dissociation of dishabituation, sensitization, and inhibition of *Aplysia*. *Science* **241**: 210–13
Marsh J T, Worden F G, Hicks L 1962 Some effects of room acoustics on evoked auditory potentials. *Science* **137**: 280–2
Murphy G G, Glanzman D L 1997 Mediation of classical conditioning in *Aplysia Californica* by LTP of sensorimotor synpases. *Science* **278**: 467–71
Peeke H V S, Herz M J (eds.) 1973 *Habituation, Volume I: Behavioral Studies, Volume II: Physiological Studies*. Academic Press, New York
Pilz P D, Schnitzler H-U 1996 Habituation and sensitization of the acoustic startle response in rats: Amplitude, threshold, and latency measures. *Neurobiology of Learning and Memory* **66**: 67–79
Prosser C L, Hunter W S 1936 The extinction of startle responses and spinal reflexes in the white rat. *American Journal of Physiology* **41**: 609–18
Sharpless S, Jasper H H 1956 Habituation of the arousal reaction. *Brain* **79**: 655–80
Sherrington C S 1906 *The Integrative Action of the Nervous System*. Yale University Press, New Haven, CT
Sokolov Y N 1963 *Perception and the Conditioned Reflex*. Trans. Waydenfeld S W, Pergamon Press, Oxford, UK
Thompson R F, Spencer W A 1966 Habituation: A model phenomenon for the study of neuronal substrates of behavior. *Psychological Review* **173**: 16–43
Wagner A R 1979 Habituation and Memory. In: Dickinson A, Boakes R A (eds.) *Mechanisms of Learning and Motivation: A Memorial Volume for Jerry Konorski*. Lawrence Erlbaum, Hillsdale, NJ, pp. 53–82
Wickelgren W O 1968 Effects of acoustic habituation on click-evolved responses in rats. *Journal of Neurophysiology* **31**: 777–84
Wicks S R, Rankin C H 1996 Recovery from habituation in *Caenorhabditis elegans* is dependent on interstimulus interval and not habituation kinetics. *Behavioral Neuroscience* **110**: 840–4

R. Thompson

Halévy, Élie (1870–1937)

Élie Halévy was a French historian and philosopher who made three principal contributions to the social sciences: a study of the English Utilitarians; a historical account of English society in the eighteenth and nineteenth centuries which is noteworthy for the role it attributes to Methodism and Evangelicalism in maintaining political stability; and an analysis of the development of European socialism culminating in what he designated an 'era of tyrannies.'

1. *Biography*

Élie Halévy was born at Étretat to wealthy, artistic parents on September 6, 1870, which marked a turbulent period in French history, with the origins of the Third Republic, France's defeat by Prussia, and the collapse of the Paris Commune. His father, of Jewish descent, was a successful librettist for such composers as Bizet and Offenbach, while his mother was of Swiss Calvinist origins. Both Élie and his brother Daniel were brought up as Protestants under her strong moral influence: like Élie's contemporary Max Weber (see *Weber, Max (1864–1920)*) there was a strong familial influence on the study of religious—and especially Christian Protestant—ideas. Halévy's grandfather had been a disciple of Saint-Simon and succeeded Comte as his secretary for a time. Though the family had important connections in French political life, Élie remained distant from, but also keenly concerned with, political affairs.

While still a student at the *Lycée Condorcet* he planned, with his brother and other associates, the *Révue de Métaphysique et de Morale*, a journal with which he was associated for the whole of his life. At the *École Normale Supérieure* he studied philosophy, completing a thesis on Platonism. Already his political and moral principles had become established as liberal, individualist, rationalist, and secular—principles that were to characterize him for the rest of his career. In 1898 Halévy was invited to lecture at the *École Libre des Sciences Politiques*, which had been founded after France's defeat in 1871 with the goal of diagnozing French political instability, and he remained there all his professional life. The *École Libre* was much admired in Britain by the group of Fabians and their associates who founded the London School of Economics and Political Science, and Halévy came to know them well as he moved in British intellectual circles.

The Dreyfus affair and the defense of the Third Republic were issues which united Halévy with Émile Durkheim (see *Durkheim, Emile (1858–1917)*), though on a philosophical level he rejected Durkheim's positivism and collectivism. On the other hand he attributed his initial interest in socialism, to which he

devoted half of his scholarly life, to Durkheim's lectures on socialism at Bordeaux in 1895. The affinity that Halévy perceived between communist and fascist régimes—since he viewed 1914 as the start of an 'era of tyrannies'—was an unpopular stance in the intellectual climate of the interwar years, but he persisted in his opinion until his death at Sucy-en-Brie on August 21, 1937. The enthusiasm of the Webbs and Shaw for fascist or communist experiments in no way surprised him, and he ended his life firmly committed to the liberal, rational, democratic ideals of his youth.

2. *Intellectual Contributions*

In his thesis, published in 1896 as *La théorie platonicienne des sciences*, Halévy explicated a dialectical approach, which he labeled Platonic, by means of which the contradictions in a subject were stated as systematic dichotomies. By exposing the core assumptions and practical implications behind a doctrine or belief system, which usually involved logical conflicts, his own analysis of it could proceed. This fundamental principle of methodology was deployed in his subsequent studies of Bentham's thought, of Wesleyan doctrine, and finally of socialism. In his study of the English utilitarians he found just such a contradiction between Bentham's political and economic theories and argued that 'the contradiction was continually breaking out in the current formulae of Benthamism' (Halévy 1952, p. 489). In his political theory, Bentham postulated the necessity of a legislator who would reconcile conflicts of interest and thus secure the greatest good of the greatest number: this he called the artificial identity of interests. In his economic theory, however, Bentham assumed with Adam Smith (see *Schumpeter, Joseph A (1883–1950)*) and his successors that there was a spontaneous identity of interests and that individual interests were naturally harmonious. Halévy's book is a reminder that the Bentham who designed the *Panopticon*—'a new application of the principle of the artificial identification of interests, the idea of which he had found in Helvetius' (Halévy 1952, p. 82)—also propagated a belief in *laissez-faire* and the 'hidden hand.'

Soon after the completion of *The Growth of Philosophic Radicalism*, Halévy's interest turned to Methodism in eighteenth- and nineteenth-century England. Sometimes regarded as the complete antithesis of utilitarianism, he regarded Methodism as an ally of utilitarianism in shaping the social morality of modern England. In 1905 he traveled to England to work on material relating to the early history of Methodism and it was at this time that he expressed his intention of writing a history of England in the nineteenth century. The Methodism project had a more immediate outcome, however, in two articles on *La naissance du Méthodisme en Angleterre* in the *Revue de Paris* in 1906. With very few exceptions these articles went unremarked by historians until their translation into English in 1971 (Halévy 1971).

The fundamentals of Halévy's interpretation of the role of Methodism in *La naissance* correspond with his later treatment in *A History of the English People in* 1815 (Halévy 1924)—the first volume of his five-volume *A History of the English People* which was published in French between 1912 and 1926 and in English between 1912 and 1932—but the focus and detail of the earlier work provide a particularly strong statement of the 'Halévy thesis.' He traces the earliest development of manufacturing industry in the first three decades of the eighteenth century and points to the growth of pauperism in the industrial centers. By the late 1730s this was compounded by a commercial crisis of overproduction accompanied by working class agitation, and it was this that stimulated the 'eruption,' in 1739, of Methodism. He argues that working class despair was transformed by Methodist doctrine and discipline into a religious and mystical movement of a conservative nature.

In this process Halévy gave a key role to the *bourgeoisie*, a class which he saw as an essential component of leadership in harnessing the enthusiasm of the proletariat. Had the *bourgeoisie* in 1739 been fired by revolutionary ideals there might well have been the conditions for social revolution, but the English *bourgeoisie* retained a strong element of Puritanism and a revulsion against the excesses of the seventeenth-century Civil War. In Halévy's words, 'There was the raw material for a general revolt. But popular discontent took the form the discontent of the *bourgeoisie* wanted to give it: a religious and conservative form' (Halévy 1971, p. 75).

On this foundation is constructed the broader thesis with which Halévy is more generally known. In *A History of the English People in 1815* he stated the problem as that of explaining 'the extraordinary stability which English Society was destined to enjoy throughout a period of revolutions and crises; what we may truly term the miracle of modern England, anarchist but orderly, practical and businesslike, but religious, and even pietist' (Halévy 1924, p. 339). The explanation he gives is complex and is focused not exclusively on Methodism, nor solely on the working class, for Methodism's impact is traced first on Nonconformity, then on the Evangelical wing of the Church of England, and through them on the ruling class, and finally on secular thought. In contrast to the view of religion as an 'opiate' which Marxian critics of Halévy such as E. P. Thompson have employed, the thesis is better interpreted as a version of *embourgeoisement*. Halévy did not attribute the influence of Methodism to the depths of the working class but saw it as having an impact on upwardly mobile workers. Methodism, originating in the Church of England, occupied an interstitial position between that Church and Nonconformity, offering a means of passage between the two: 'Puritan nonconformity thus tended

to become a transitional creed, a stage in the history of an English family. The unskilled laborer becomes in turn a skilled workman, an artisan, the head of a small business, a businessman possessed of a modest capital, and as he rises out of the barbarism in which the working class was plunged, he becomes a Nonconformist. If he himself rises still higher on the social ladder, or if his children rise after his death, he, or they go over to the Church of England' (Halévy 1924, p. 371). Wesleyan Methodism, he claimed, rendered the transition imperceptible. Paradoxically, as the religious institutions facilitated such upward social mobility they also served to sharpen status distinctions, for while the middle-class Nonconformist might be despised by members of the Church of England, to which his family might one day belong, compensation could be found in contempt shown towards the working class from which the family had come. Hence, Halévy's opinion that 'the élite of the working class, the hard-working and capable bourgeois' had been instilled with values which did not threaten the existing order of society.

Wesleyan Methodism had originated in the Church of England among a group of conservative High Church clergy. In preaching an egalitarian doctrine of salvation for all it had successfully mobilized an important segment of the rising working class. Here is an excellent example of Halévy's utilization of systematic dichotomy at the core of a belief system. He goes on to demonstrate the influence which Methodism had on establishment opinion by proposing that Wesley had left a 'rear-guard' in the Church of England in the form of the Evangelical movement, which sought to realize his original goal of Church regeneration: 'If the Wesleyan sect, with its hierarchic constitution, and frank political conservatism, constituted the High Church of Nonconformity, the new Low Church or evangelical party was a species of Anglican Methodism' (Halévy 1924, p. 379). With powerful representatives at the University of Cambridge, and with wealthy and influential advocates in the 'Clapham sect' (which was comprised principally of Members of Parliament like William Wilberforce), the Evangelicals were in a position to influence not only ecclesiastical reform but also to campaign for a wider reform of morals—including those of the ruling class. In these several ways, argued Halévy, the influence of Wesleyan Methodism had permeated throughout English society and had helped establish the temper of Victorian England. By contrast, no comparable institution or creed had existed in eighteenth-century France.

Halévy's *Histoire du socialisme européen* (1948) and *L'ère des tyrannies* (1938) are the most cited of his works in France. They further extend the central theme of his work on English ideas, where the contrast between reform and revolution provided his comparative focus. Just as the contradiction between natural and artificial interests had been identified in utilitarianism, and that between egalitarian salvation and hierarchical discipline in Wesleyanism, so the contradiction in socialism derived from the conjoint goals of human emancipation and the zeal to organize society. The decisive event in Halévy's view of socialism, which changed from a skeptical liberalism to a critical pessimism, was the First World War. The growth of the power of the state and the spread of bureaucracy were seen not only as threats to parliamentary democracy but also as evidence of the dominance of the organizational over the emancipatory principle in socialism.

Halévy's central thesis is clearly stated: 'Since its beginnings, in the early years of the nineteenth century, socialism has suffered from an internal contradiction. On the one hand, its partisans often present it as the outcome and fulfilment of the Revolution of 1789, a revolution of liberty, a liberation from the last remaining subjection after all the others have been destroyed: the subjection of labor by capital. But, on the other hand, it is also a reaction against individualism and liberalism; it proposes a new compulsory organization in place of the outworn institutions destroyed by the Revolution …' (Halévy 1966, p. 265). It is known that the pessimistic conclusions he drew from this diagnosis surprised many in his intellectual circle at the *Société Française de Philosophie*, and arrangements were being made for a further debate on his propositions when he died in 1937.

3. *Impact and Controversy*

Halévy's work continues to be a focus of research and debate among social scientists. Two biographies have appeared, the first relating him to the French and European contexts of his time and the second emphasizing the major themes of his work: both characterize the 40-year span of his writing with the same term, 'unity' (Bramsen 1978, p. 348, Chase 1980, p. 3). In Chase's work there is a particularly clear exposition of the 'Halévy thesis' which gives effective treatment to aspects of his argument which have sometimes been obscured in subsequent debate, such as the broader significance of Nonconformity and his portrayal of the 'hard-working and capable *bourgeois*' as the 'élite of the working class.' The republication in 1995 of a French edition of his book on English utilitarianism has stimulated critique of a work that had been widely accepted as a definitive analysis of its subject. Thus, Halévy has been taken to task for allegedly interpreting the principle of utility, or seeking of pleasure, as both a positive and a normative statement—as both an 'is' and an 'ought' (Vergara 1998).

Even more extensive has been the attention paid to his claim that the Methodist and Evangelical movements were responsible for England's remarkable stability during a period of political ferment in Europe.

Shortly after the publication of the first volume of *A History of the English People in the Nineteenth Century*, the Hammonds offered their portrayal of Methodism as having a repressive and intolerant stance towards the working class. In 1957, in *History Today*, E. J. Hobsbawm questioned Halévy's claims of the extent to which Methodism penetrated the working class and offered an alternative explanation of England's political stability deriving from the ruling classes's ability to retain control. E. P. Thompson's *The Making of the English Working Class* in 1963, however, attributed even more influence to Methodism in subduing the working class than had Halévy. In a 'Postscript on the Halévy Thesis' in *Victorian Minds*, Himmelfarb (1968) commented on the curious status of the thesis in current historical literature, where it was often assumed that the thesis had been discredited or greatly modified while in reality there had been no substantial exploration of it. After reviewing the main features of the debate she concluded by paraphrasing Halévy: 'Will Halévy defeat Marx as the interpreter of this crucial period in English history?' (Himmelfarb 1968, p. 299).

Himmelfarb's postscript became the preface to two decades of research into 'the Halévy thesis.' Semmel's 1971 translation of Halévy's 1906 articles *The Birth of Methodism in England* (Halevy 1971 was published with an introduction which noted that 'the full dimensions of the thesis, as Halévy stated them, have been imperfectly understood'. This was because the 1906 articles had gone virtually unread. The thesis was further explored in *The Methodist Revolution* (Semmel 1973), which began by noting that in the major discussions of the 'Halévy thesis' the role of Methodist theology had been afforded little significance, in contrast to social scientific discussions of seventeenth-century Calvinism. Semmel devoted several chapters to the elements which contributed to the synthesis of Methodist beliefs, restating Halévy's observation of an underlying contradiction between the revolutionary and counter-revolutionary elements of Wesleyan doctrine.

The thesis continued to be the subject of discussion throughout the 1970s and attracted the attention of sociologists examining religion and social change: in this context the 'Halévy thesis' has been described as 'the political corollary, applied in a more restricted historical and geographical context, of the broader thesis put forward by Weber' (Hill 1973, p.183). As closer attention has been paid to Halévy's work, some of his specific historical observations—and their interpretation by more recent writers—have provided further stimulus for debate. One such area has been the role of foreign missions, and specifically of the Wesleyan Methodist Missionary Society. Did such activity provide a safety valve for social and religious tensions, rechanneling potentially disruptive enthusiasm by exporting it, or was there equal concern for home missions? These questions have been addressed in a revisiting of Halévy's original account (Piggin 1980).

In a wide-ranging review of the debate, Itzkin, like Himmelfarb before her, has noted that many of the contentious issues remain unresolved. While Halévy's antagonists have succeeded in revising some of his claims, the central thesis has remained viable and continues to constitute a 'working historical hypothesis' (Itzkin 1975, p. 56). In the work of revision, attention has focused on the numerical strength of Methodism together with its geographical distribution; on the nature and political stances of the variety of bodies which came within the overall rubric of Methodism; and on the way in which moderate political deviance can make 'an important contribution to public order and political stability' (Gilbert 1979, p. 392).

With the culmination of two decades of lively debate over the 'Halévy thesis,' a collection of most of the important statements was published in Olsen (1990). Extracts from Halévy, from historians of Methodism and Evangelicalism such as Maldwyn Edwards, John Walsh, and Kiernan, and from critics like the Hammonds, Hobsbawm, and E. P. Thompson provide a striking testimony to the continuing vitality of a thesis first advanced in 1906. The parallel with the 'Weber thesis,' first published in 1904–5, is too close to ignore.

Halévy's lectures on the history of socialism, reconstructed from his students' notes and published posthumously as *Histoire du Socialisme Européen* in 1948, remains untranslated into English. *The Era of Tyrannies*, first published posthumously in French in 1938, has, however, been translated and published in English (Halévy 1966). In the latter work the author provides not only a summary account of socialism but also a treatment of the era of tyrannies—using the term in the classical Greek sense of a group of armed men who seize the state—under Stalin, Hitler, and Mussolini. At the end of World War One Halévy had written: '… I recognize that socialism holds the key to the secret of the future. But I cannot decipher this secret nor can I say whether socialism will lead us to a Swiss-type, universalized Republic or to European Caesarism' (Aron 1984, p. 407). In the last years of his life he took an increasingly pessimistic view of Europe's future in the event of another war, which he thought likely. For him, that would represent the final blow to liberal traditions; but for this judgment Aron has criticized him. While cautioning against over-confidence, Aron thought Halévy had 'sinned through over-pessimism' (Aron 1984, p. 422). With the benefit of hindsight Aron's criticism may be justified, but it should be recalled that in the era of the 1930s Halévy was not alone among social scientists who believed that the capitalist élite were becoming 'highly placed bureaucrats.'

See also: Bentham, Jeremy (1748–1832); Durkheim, Emile (1858–1917); Liberalism; Marx, Karl (1818–89);

Socialism; Utilitarianism: Contemporary Applications; Weber, Max (1864–1920)

Bibliography

Aron R 1984 *Élie Halévy*. *Government and Opposition* **19**(4): 407–22
Bramsen M B 1978 *Portrait D'Élie Halévy*. Grüner, Amsterdam
Chase M 1980 *Élie Halévy: An Intellectual Biography*. Columbia University Press, New York
Gilbert A D 1979 Methodism, dissent and political stability in early industrial England. *Journal of Religious History* **10**(4): 381–99
Halévy É 1924 *A History of the English People in 1815* [trans. Watkin E I, Barker D A]. Fisher Unwin, London
Halévy É 1952 *The Growth of Philosophic Radicalism* [trans. Morris M]. Faber and Faber, London
Halévy É 1966 *The Era of Tyrannies* [trans. Webb R K]. New York University Press, New York
Halévy É 1971 *The Birth of Methodism in England* (trans. Semmel B). University of Chicago Press, Chicago
Hill M 1973 *A Sociology of Religion*. Heinemann Educational, London
Himmelfarb G 1968 *Victorian Minds*. Weidenfeld and Nicolson, London
Itzkin E 1975 The Halévy thesis—A working hypothesis? *Church History* **44**(1): 47–56
Olsen G W (ed.) 1990 *Religion and Revolution in Early-industrial England: The Halévy Thesis and its Critics*. University Press of America, Lanham, MD
Piggin S 1980 Halévy revisited: The origins of the Wesleyan Methodist Missionary Society. *Journal of Imperial and Commonwealth History* **9**(1): 17–37
Semmel B 1973 *The Methodist Revolution*. Basic Books, New York
Vergara F 1998 A Critique of Élie Halévy. *Philosophy* **73**(283): 97–111

M. Hill

Hall, Granville Stanley (1844–1924)

The age of the modern research university began with the founding of the Humboldt University in Berlin in 1810. One of the 9,000 Americans who studied at this and other German universities during the nineteenth century was Granville Stanley Hall. He, like they, sought to gain more advanced knowledge and research skills than was available elsewhere. This educational opportunity aided him in becoming the first professor of psychology and pedagogics at the new Johns Hopkins University, and later the first president of research-oriented at Clark University in the USA. During his academic career, Hall launched the American discipline of psychology (Ross 1972, 1979) and the science of education (White 1992), out of the then broader study of philosophy. He achieved these intellectual accomplishments through his long faculty and presidential career, numerous books and more than 400 other publications, creation of research journals, and extensive professional conference presentations. In this way, Hall played the initial pivotal and defining role in founding psychology and education as social and behavioral sciences in American higher education (Mitchell and Haro 1999, p. 50, Lagemann 2000).

1. Hall's Early Life, Career, and Intellectual Pursuits Until 1880

Born in Ashfield, Massachusetts on February 1, 1844 to Granville Bascom Hall and Abigail Beals Hall, Stanley was the oldest of nine siblings. His upbringing on the family farm offered his inquisitive mind opportunities to observe nature intimately, which would so powerfully shape his future intellectual outlook. Eschewing agrarian life, Hall completed college preparatory studies at Williston Seminary in Easthampton, Massachusetts in 1862 and entered Williams College in the following year. There he studied under two famous American educators and philosophers, Mark Hopkins and John Bascom. Their influences shaped his interest in philosophy and theology, respectively. Finishing his bachelor's degree in 1867, Hall next decided to study for the ministry and entered Union Theological Seminary in New York. Yet, two years of study, some pastoral experience, and a frank discussion about his intellectual interests with the imposing American theologian Henry Ward Beecher resulted in Hall's first trip to Germany to study philosophy at the University of Berlin in 1869. In Friedrich Adolf Trendelenburg's seminary, he became an avowed Hegelian and contemplated becoming a philosopher. With the outbreak of the Frank-Prussian war, Hall returned to New York and finished his bachelor's degree in divinity in 1871. Yet, his intellectual and philosophy study there led him to embrace the evolutionary ideas of Charles Darwin and Herbert Spencer (Hall 1923).

Forsaking the ministry then, Hall sought the academic life. He became a professor at Antioch College in Ohio, where he taught English, French, and German from 1872 to 1876. Becoming increasingly disenchanted with Hegel intellectually during this time, Hall openly became an enthusiast for Darwin, Spencer, and Huxley and positivistic science (Hall 1923, p. 199, Ross 1972, pp. 58–9). The turning point in his intellectual life occurred with his reading of Wilhelm Wundt's (1873) first volume of *Physiological Psychology*, after which he decided to specialize in psychology. Leaving Antioch, Hall completed a Ph.D. at Harvard University in 1878 under William James; it was the first doctorate in psychology awarded in the USA. His dissertation explored the physiological

relationships between muscles, nerves, and consciousness from a genetic perspective (Ross 1972, pp. 70–5, Wilson 1914, p. 64). Hall's research showed his fascination with experimental psychology and its links to Darwinian and Spencerian evolutionary thought. To bolster his research in the study of psychology, and because he could not obtain a suitable professorial appointment at a major university, Hall again went to Germany to study under Hermann von Helmholtz and William Wundt at the universities of Berlin and Leipzig, respectively (Goodchild 1996, pp. 72–3).

2. *Hall's Faculty Appointments at Johns Hopkins University from 1882 to 1887*

After returning to the USA in 1880, Hall lectured at Harvard at the invitation of President Charles W. Eliot and then at Hopkins by courtesy of President Daniel Coit Gilman. Later the Hopkins president appointed him a lecturer in 1882 and then Professor of Psychology and Pedagogics in 1884. From these positions, Hall furthered the study of education and psychology through a series of major articles exploring their relationships to human development. First, he surveyed pedagogical study in the USA and then explored the moral and religious training of both children and adolescents in one of his first discussions of child development. In analyzing the needs of education later, the professor stated boldly that 'we must study and follow the child's nature as it actually is' (Hall 1883a, p. 286). Hall further proclaimed his new concept of 'natural' methods of education where education and psychology may both be utilized to understand the relationship between intelligence and human development in his important article, New Departures in Education (Hall 1885a, pp. 147–8). Finally, he and one of his doctoral students outlined the entire field of education in their 1886 book, *Bibliography of Education* (Hall and Mansfield 1886).

Second, in psychology, Hall quickly created one of the first experimental psychological laboratories in America at Hopkins and authored many of its early articles in experimental psychology (Boring 1950, Hawkins 1960, pp. 202–3). One of his important contributions was the publication of 'The new psychology' (1885b) which outlined this developing field. In 1887, he created a new research journal, the *American Journal of Psychology*, to publish psychological research. It was pivotal in establishing this new discipline. Third, his pedagogical and psychological ideas had already coalesced a few years earlier in his first major empirical studies using questionnaires on child development. 'The contents of children's minds' (1883b) and *The Study of Children* (1883c) described 'the real nature' of children, used the reports of actual teachers' observations from the questionnaires, and suggested appropriate pedagogical techniques. They played a major role in launching Hall's national reputation and his central position as the founder of the child study movement in the USA (White 1990). Some of his doctoral students followed him as his career evolved; others, particularly John Dewey and James McKeen Cattell, were not favored and moved onto their own significant accomplishments (Ross 1972, pp. 144–7). Thus Hall's work at Hopkins set the stage for the remainder of his scholarly career and publications as he pursued research in education, psychology, and their mutual implications for understanding human development.

3. *Hall's Presidency of Clark University from 1888 to 1920*

In Worcester, Massachusetts, wealthy industrialist Jonas Clark founded an institution of higher learning with a $1 million gift in 1887. He intended it to have a research mission for older male students and be a 'useful' college for local boys (Koelsch 1987, pp. 9–12). Hall was named its first president in April 1888. He was determined to create a new graduate research university comparable to Hopkins, following the German ideal. The first three years saw considerable fulfillment of his vision. However, conflict with the founder over the college and loss of many faculty members to the new University of Chicago severely constrained the president. The struggling and much smaller institution still embraced the primacy of research as the reform ideal for the mission of all American universities (Goodchild 1996, Goodchild and Miller 1997). Eventually, Hall allowed an undergraduate college to be created when funds from the founder's estate became available. The president remained highly productive in his research and continued to oversee the institution until his retirement in 1920.

4. *Hall's Prolific Scholarly Achievements During his Presidency*

The president's accomplishments in psychology, education, and their implications for a new genetic interpretation of human development established the American prominence of these new scientific fields of study. In psychology, Hall founded the American Psychological Association in 1892, became its first president, and anchored his *American Journal of Psychology* to the organization. It remained the pre-eminent national association for this growing academic discipline. In 1904, Hall's most famous psychological study, *Adolescence: Its Psychology and Its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion and Education*, appeared. This definitive two-volume work on adolescence comprised many of his earlier ideas on the child's next stage of development, established this stage as a scientific

concept, related this developmental period to various human behaviors, and discussed its meaning for other emerging academic fields, especially education. Hall's interest in the sexual development of youth in this work was ground breaking. It led him to invite Sigmund Freud, Carl Jung, and many other leading psychologists to the famous 1909 twentieth Anniversary Conference of the founding of Clark University (Koelsch 1987, pp. 246–7). As if to secure his role in this developing discipline, Hall published the valuable book, *Founders of Modern Psychology*, in 1912. After World War I, the scholarly president turned to his lifelong interest in religion and its psychological meaning. He authored a major reinterpretation of Christianity's founder in his *Jesus the Christ, in the Light of Psychology* (1917) and its application in the advocacy of a new life ethic in his *Morale: The Supreme Standard of Life and Conduct* (1920a) in which Hall espoused a holistic understanding of health and well-being as the rule for human conduct. With retirement from Clark's presidency in 1920, Hall reflected on his accomplishments in his *Recreations of a Psychologist* (1920b), and *Life and Confessions of a Psychologist* (1923). Hall achieved another scholarly triumph by completing his work on the last stage of human development with the publication of *Senescence: The Last Half of Life* (1922). In all this pioneering work, he made some significant missteps, especially claiming child development recapitulates human evolution (White 1990, p. 31) and a too pervasive Darwinian and Spencerian 'evolutionism' (Ross 1972, p. 92). On the other hand, he awarded the first doctorate in psychology to an African-American and encouraged Asian students to study psychology at Clark (Goodchild 1996).

By his death on April 24, 1924, Hall had established psychology as an American discipline and study, almost entirely graduated the next generation of psychologists (Goodchild and Miller 1997), and secured the role of behavioral and genetic psychology as being an integral part. To insure the continuation of his work, Hall's estate established at Clark an endowment for a genetic psychology professorship and for his educational journal that was redirected and renamed *The Journal of Genetic Psychology*.

Similarly, his work in education commanded national attention. Hall believed that education was applied psychology. He followed identical approaches in establishing the importance of education as a field of study, as he had done in psychology. Hall offered a series of university lectures on kindergarten, elementary, secondary, and higher education (Goodchild 1991). He next founded a research journal, *The Pedagogical Seminary: An International Record of Educational Literature, Institutions, and Progress*, in 1891. It provided a new research outlet from studies intended to reform education from American and international perspectives. Further, he extracted his pedagogical ideas from his books on adolescence and released them as *Youth: Its Education, Regimen, and Hygiene* (1906) to serve educators and teachers who were implementing his ideas in the classroom. Five years later, Hall presented his definitive study on education with another extensive two-volume work, *Educational Problems* (1911). He claimed to have fomented an American educational revolution where 'now the school, including its buildings, all its matter and method, revolve about the child' (p. 1: vi). This 'natural' education was to be individualized to the student's developmental stage, sex, and race. Thus pedagogy was to assist in the intellectual development of every person, although higher education was intended for the most able. At Clark, Hall had been instrumental in creating the study of education—particularly in conducting empirical research on children to develop better pedagogy, offering the first courses in higher education, and espousing genetic psychology as the basis for educational practice. His extensive publications in higher education focused on the mission of the new research university and its faculty (Goodchild 1996). Similarly, missteps occurred in education as Hall advocated separate education for women, decried 'identical' coeducation, and espoused elitism in higher education (Ross 1972, pp. 311–12). Nevertheless, the president greatly assisted in the arrival of education as an academic field of study (Cremin 1961).

Hall's most significant contribution may be the creation of developmental psychology and its relationship to education (White 1992). His research on child study, adolescence, and old age enabled him to argue for five stages of human development: (a) childhood; (b) adolescence; (c) middle life, from 25 to 45; (d) senescence, after 40; and (e) senectitude, approximately near the end of life (Hall 1922, p. vii). His psychological and educational ideas and practices were focused on what he believed, would further the human race. His Darwinian and Spencerian concepts of evolution pervaded his understanding of human genetic development. He consistently maintained these ideals, even in the face of demands for coeducation in schools, growing behaviorism in psychology, and expansive progressivism in education. Hall was a man of his times, his scientific worldview centered on the mid- to late nineteenth century ideas and their implications for professional practice.

Finally, Hall's most important achievement when assessing the rise of the social and behavioral sciences was to engender the development of psychology and education from their philosophical progenitor. His patterns consisted of a critical assessment of the condition of the field first, next an assertion of a new position for the field, then a new research journal to publish these findings, and finally definitive works exploring the field of study. Hall thus played an essential role in the establishment of psychology and education as new academic fields of study in the USA. In 1992, the centenary meeting of the American

Psychological Association pointed to his pivotal role in these founding moments (Goodchild 1992, White 1992).

See also: Adolescent Development, Theories of; Aging, Theories of; Education: Phenomena, Concepts, and Theories; Helmholtz, Hermann Ludwig Ferdinand von (1821–94); James, William (1842–1910); Life Course: Sociological Aspects; Lifespan Development, Theory of; Psychology: Historical and Cultural Perspectives; Rousseau, Jean-Jacques (1712–78); Spencer, Herbert (1820–1903); Wisdom, Psychology of; Wundt, Wilhelm Maximilian (1832–1920); Youth Culture, Sociology of

Bibliography

Boring E C 1950 *A History of Experimental Psychology*, 2nd edn. Appleton-Century-Crofts, New York

Cremin L A 1961 *The Transformation of the School: Progressivism in American Education, 1876–1957*. Vintage Books, New York

Goodchild L F 1991 Higher education as a field of study: Its origins, programs, and purposes, 1893–1960. In: Fife J D, Goodchild L F (eds.) *Administration as a Profession. New Directions for Higher Education*, no. 76. Jossey-Bass, San Francisco

Goodchild L F 1992 G. Stanley Hall's Psychological Avocation: Founding the Study of Higher Education. Research paper given at the 100th American Psychological Association annual meeting in Washington, DC

Goodchild L F 1996 G. Stanley Hall and the study of higher education. *The Review of Higher Education* **20**: 69–99

Goodchild L F, Miller M M 1997 The American doctorate and dissertation: Six developmental stages. In: Goodchild L F, Green K E, Katz E L, Kluever R C (eds.) *Rethinking the Dissertation Process: Tackling Personal and Institutional Obstacles. New Directions for Higher Education*, no. 99. Jossey-Bass, San Francisco

Hall G S 1883a Educational needs. *North American Review* **236**: 284–90

Hall G S 1883b The content of children's minds. *Princeton Review* **11**: 249–72

Hall G S 1883c *The Study of Children*. Printed privately

Hall G S 1885a New departures in education. *North American Review* **140**: 144–52

Hall G S 1885b The new psychology. *Andover Review* **3**: 120–36, 239–48

Hall G S, Mansfield J M 1886 *Hints Toward a Select and Descriptive Bibliography of Education*. Heath, Boston

Hall G S 1904 *Adolescence: Its Psychology and Its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion and Education*, 2 Vols. Appleton, New York

Hall G S 1906 *Youth: Its Education, Regimen, and Hygiene*. Appleton, New York

Hall G S 1911 *Educational Problems*, 2 Vols. Appleton, New York

Hall G S 1912 *Founders of Modern Psychology*. Appleton, New York

Hall G S 1917 *Jesus, The Christ, in the Light of Psychology*. Doubleday Page, New York

Hall G D 1920a *Morale: The Supreme Standard of Life and Conduct*. Appleton, New York

Hall G S 1920b *Recreations of a Psychologist*. Appleton, New York

Hall G S 1922 *Senescence: The Last Half of Life*. Appleton, New York

Hall G S 1923 *Life and Confessions of a Psychologist*. Appleton, New York

Hawkins H 1960 *Pioneer: A History of Johns Hopkins University, 1874–1889*. Cornell University Press, Ithaca, NY

Koelsch W A 1987 *Clark University, 1887–1987: A Narrative History*. Clark University Press, Worcester, MA

Lagemann E C 2000 *An Elusive Science: The Troubling History of Educational Research*. University of Chicago Press, Chicago

Mitchell T R, Haro A 1999 Poles apart: Reconciling the dichotomies in educational research. In: Lagemann E C, Shulmlan L S (eds.) *Issues in Educational Research: Problems and Possibilities*. Jossey-Bass, San Francisco

Ross D 1972 *G. Stanley Hall: The Psychologist as Prophet*. University of Chicago Press, Chicago

Ross D 1979 The development of the social sciences. In: Oleson A, Voss J (eds.) *The Organization of Knowledge in Modern America, 1860–1920*. Johns Hopkins University Press, Baltimore, MD

White S H 1990 Child study at Clark University: 1894–1904. *Journal of the History of the Behavioral Sciences* **26**: 31–50

White S H 1992 G. Stanley Hall: From philosophy to developmental psychology. *Developmental Psychology* **28**: 25–34

Wilson L N 1914 *G. Stanley Hall: A Sketch*. Stechert, New York

Wundt W 1873 *Principles of Physiological Psychology*. Engelmann, Leipzig, Germany

L. F. Goodchild

Hallucinations, Neural Basis of

Hallucinations, by definition, are percepts experienced in the absence of sensory input. They can be expressed in any of the sensory modalities and arise from a wide range of pathological conditions. Hallucinations are also a normal component of conscious experience. For instance, they arise routinely as ordinary dreaming during sleep. Sometimes daydreams and relaxed waking consciousness produce imagery so vivid that it is momentarily experienced as real (Foulkes and Fleisher 1975). The main types of hallucinations and their distinctive phenomenological characteristics have been the subject of considerable prior study (Brown 1985, Benson and Gorman 1993). A brief summary of the major types follows:

Visual hallucinations: *Peduncular* hallucinations consist of animals, face, or human figures and are traditionally associated with vascular lesions in the upper brainstem (involving, especially, the cerebral peduncles). They can also occur in cases of encephalitis or during the transition from waking to sleep (Lhermitte 1922, from Brown 1985). *Charles Bonnet syn-*

drome is similar and occurs in the elderly, generally with significant visual impairment. However, human figures are characteristically small and doll-like, and can actually amuse the subject (Benson and Gorman 1993). *Palinopsia* is a pathological persistence of a visual percept originating from external stimulation that arises from lesions of the occipital cortex, seizures, or drug-induced states (Benson and Gorman 1993, Brown 1985). So-called *elementary* hallucinations consist of bands of light or phosphenes that can arise from pathology of the retina or optic nerve, and certain drug-induced states (Brown 1985). *Hypnogagic* hallucinations occur normally during the transition from sleep to wakefulness and are experienced as the continuation of a typical dreamstate. Hallucinations of faces or human figures are reported by approximately 20 percent of patients with schizophrenia. Visual hallucinations also often occur during seizures in patients with epilepsy. These hallucinations can have a geometric design or reflect a recollection of visual images based on past experience. Severe alcohol withdrawal often is accompanied by visual hallucinations of bugs or spiders. Lysergic acid (LSD) and related compounds often produce visual hallucinations consisting of light flashes, colors, geometric figures, and generally are accompanied by distortions of the geometric form of percepts (Crowley 1995). Higher doses produce fantastic forms, distorted facial images, etc. So-called *autoscopic* hallucinations reproduce an actual visual image of the self and are generally associated with epilepsy though these experiences also can occur in patients with schizophrenia or depression (Dening and Berrios 1994).

Auditory hallucinations: Most commonly, auditory hallucinations consist of spoken speech. These hallucinations are reported by 70–80 percent of patients with schizophrenia and less commonly by patients with other psychiatric disorders, such as depression or mania. These patients may also experience hallucinations of non-speech sounds, such as rattling, footsteps, etc. Other common causes of auditory hallucinations of speech include chronic alcoholism and phencyclidine use, or chronic amphetamine abuse. Hallucinations of speech consist of one or more 'voices' which comment on the patient's speech, thought, or behavior, command them to perform certain actions, or converse with each other. A nonpathological variant of this experience is the momentary experience of hearing one's name being spoken in the absence of an actual speaker. Both pathological and nonpathological variants of hallucinated speech are often triggered by indistinct sounds such as background conversation, street sounds, and running water. Often, if visual hallucinations are also experienced by patients with schizophrenia, they are phenomenologically linked with the auditory hallucination, e.g., a face will appear to speak the hallucinated words. Auditory hallucinations of music can occur in psychiatric patients with an affective disorder such as depression (Wengel et al. 1989) as well as in patients with deafness and steroid-induced psychotic states. *Palinacousis* is the auditory analogue of *palinopsia* and consists of the persistence of previously heard sounds, music, or speech.

Tactile or haptic hallucinations: These experiences commonly occur during the *delirium tremens* stage of severe alcohol withdrawal and consist of *formication*, i.e., the sensation of insects crawling on the body. Amphetamine or cocaine-induced psychosis can also lead to formication experiences. *Phantom limb* hallucinations follow arm or leg amputation and consist of the persistent sensation of the lost limb that can include the experience of movement and/or pain. Sensation of altered or duplicated body shapes can

> occur in epilepsy, toxic encephalopathy, or with hallucinogenic drugs. When such hallucinations are accompanied by delusions that are fantastic, they are likely to be related to a schizophrenic process (Benson and Gorman 1993).

Hallucinations of bodily sense and movement have also been reported following direct electrical stimulation of the temporal lobe in humans (Ishibashi et al. 1964).

Olfactory hallucinations: These hallucinations most typically occur during seizure auras, migraine attacks, or as a result of damage to the olfactory bulb at the base of frontal lobe (e.g., by tumors). Olfactory hallucinations generally have a noxious content, consisting of a burning smell or some other foul odor. They are also reported occasionally by patients with schizophrenia and depression.

Mechanisms of hallucinations: Benson and Gorman (1993) outline a simple, but intuitively appealing, approach dividing hallucinations into two categories: the first 'release' hallucinations and the second 'ictal' hallucinations. The former arise when 'normal sensory processing channels are blocked, allowing stored images to be experienced' (pp. 266–267). Typical examples would be musical hallucinations following deafness or the Charles Bonnet syndrome arising from visual impairment in the elderly. These hallucinations tend to be longer in duration, can often be well-formed and quite elaborate, and occur only in the defective receptive field. 'Ictal' hallucinations arise from 'abnormal neuronal discharge' and tend to be briefer, less well-formed, and associated with alterations in consciousness. Benson and Gorman also highlight the importance of alterations in attention/arousal in producing certain types of hallucinatory phenomena (1993, p. 278). Examples include dream-like hallucinations arising from the transition from sleep to arousal, the disturbed consciousness of drug-induced states, sleep deprivation, narcolepsy, or delirium. The postulated mechanism is an alteration in the gating of material into conscious awareness. In addition, limbic disorders and frontal dysfunction were postulated to cause hallucinations. We will discuss below some evidence suggesting a role of the former.

More recent work on the details of hallucinatory phenomena, considered in the remainder of the article, suggests some revision may be necessary in the useful framework introduced by Benson and Gorman.

First, consider again phantom limb phenomena, a classic example of 'release' hallucination. These experiences are almost universal, occurring in approximately 85 percent of amputees (Benson and Gorman 1993) and are experienced as if the cortical representation of the limb is retained somehow as a conscious experience. Phantom limb experiences, especially early on, often are finely articulated with detailed sensations of the digits of the hands or feet. After loss of a hand, the person may still, for instance, re-experience the sensation of a long-worn ring. One puzzling observation is that patients who also lose sensation and use of a limb due to spinal cord injuries do *not* generally have phantom limb hallucinations or, if they do, these are very indistinct and fleeting. Why does loss of limb produce these experiences, while paraplegia does not? In both cases, deafferentation has occurred and 'release' phenomena should, in theory, arise.

Recent work by Spitzer et al. (1995) has provided an updated model suggesting that phantom limb hallucinations arise from a progressive reorganization of the sensory area of the cerebral cortex responsible for limb representation. The critical observation underlying Spitzer's model is that sensory ganglia in the spinal cord are known to fire randomly following denervation. These random impulses are then projected in a sustained fashion to the somatosensory cortex. Random firing from these sensory ganglia is not generated, in contrast, in the case of spinal cord injury. Using computer simulations, Spitzer demonstrated that chronic random inputs cause a deafferentized cortical sensory system to reorganize so that activation patterns becomes self-sustaining. Moreover, a cortical region whose input is limited to noise was shown to assimilate receptive field properties of neighboring cortical areas gradually. Consequently, representations of the original limb representation are gradually usurped. Progressive reorganization of limb cortical maps in this fashion could produce the characteristic subjective 'shrinking' of these hallucinations over time—the phantom limb grows smaller over a period of weeks to months so that, in later stages, it is only experienced as, say, the size of a postage stamp superimposed on the stump.

The limits of the Benson/Gorman approach are also suggested by the hallucinations experienced during amphetamine psychosis. These hallucinations typically consist of spoken speech of whole phrases or sentences or the perception of faces or bodily forms and are often accompanied by paranoid fears. Their phenomenology often resembles that of schizophrenia. Consequently, amphetamine psychosis is studied as an experimental model of this disorder.

These hallucinations do not readily fit a reduced attention/arousal model of hallucinations since amphetamine generally heightens attention and arousal. There are other data to suggest that perhaps amphetamine-induced hallucinations reflect an 'ictal' or 'neuronal activation' model of hallucinations. It is well-known, for instance, that a single dose of amphetamine produces widespread, heightened activation of the cerebral cortex (Porrino et al. 1984). However, amphetamine-induced hallucinations almost invariably do not emerge unless the person is a chronic abuser of the drug (Bell 1973).

Induction of apparent hallucinatory behavior in monkeys also requires repeated exposure to this drug (Castner and Goldman-Rakic 1999). Thus, it seems clear that the hallucinations induced by amphetamine do not arise simply from heightened brain activation—which is readily detected following a single exposure to the drug—but instead reflect complex changes in brain function generated over time by repeated drug exposure. The exact nature of these brain 'sensitization' changes has yet to be determined. However, recent studies suggest that alterations of neural circuitry in brain regions which receive projections from dopamine neurons and ordinarily respond to amphetamine are critical (Pierce and Kalivas 1997). These alterations probably involve a cascade of molecular events that enhance cellular aspects of learning/conditioning mediated by receptors for the excitatory amino acid neurotransmitter, glutamate, as well as elaboration of new neuroanatomic connections (Robinson and Kolb 1999, Wolf 1998). These changes produce widely distributed neurocircuits that consequently become 'primed' to respond to the activating effects of amphetamine.

Reports of direct electrical stimulation in the human brain also suggest that neural activation alone is not sufficient to cause hallucinations. These studies have been conducted in patients with intractable epilepsy who were evaluated for neurosurgical resection of seizure loci. Penfield, a pioneer in this area, reported that complex hallucinations whose content consisted of 'engrams' of previous memories could be elicited in unanesthetized human subjects by electrically stimulating temporal lobe neocortical gray matter (cf. Penfield and Perot 1963). Penfield proposed that such stimulation directly accessed neural networks that actually coded for memories. Later workers found it difficult to obtain similar results, however. Halgren et al. (1978), Gloor et al. (1982), and Horowitz and Adams (1970), for instance, were able to elicit experiential responses but by stimulating deeper limbic structures (i.e., hippocampus and amygdala) rather than neocortical areas.

These experiences tended not to be memories but did include a range of hallucination-like experiences including unformed sensory experiences (lights, streaks, flashes) and complex visual experiences of persons, scenes, or objects or autoscopic experiences of the self. The location of sites that elicited hallucinations is curious since it is generally assumed that only

neocortical structures (not limbic structures) code for actual sensory experiences. This raises the question of how hippocampal and amygdala stimulation, whose known functions pertain to registration of emotional salience, memory storage, and memory reactivation, could produce hallucinations.

One clue was that hallucinations were more likely to emerge when limbic stimulation produced not simply activation, but sustained afterdischarges, i.e., rhythmic activity detected via EEG recordings (Halgren et al. 1978, Horowitz and Adams 1970). After-discharges of this sort are more likely to lead to a temporary functional lesion of neural systems rather than activation of coded information (Gloor et al. 1982). Analogously, rhythmic electrical stimulation of neocortical tissue produces temporary aphasia or speech arrest if directed to language processing areas of the brain (Ojemann and Mateer 1979). Halgren et al. 1978) and Horowitz and Adams 1970) proposed that sustained disruption of limbic structures induced by after-discharges lead to the 'release' of sensory material from neocortical regions into consciousness. Consistent with this view are EEG studies in animals showing that the REM stage of sleep and LSD administration, which each produce vivid hallucinations in humans, cause dramatic *reductions* in hippocampal activation (Halgren et al. 1978). These observations suggest that the actual production of hallucinations by intracerebral stimulation arises from a dynamic interaction between neocortical and limbic systems, rather than simple activation of a group of neurons.

As with other forms of hallucinations, a new picture of auditory hallucinations in patients with schizophrenia is now emerging. Functional neuro-imaging studies indicate that these hallucinations are accompanied by activation from speech perception and auditory neurocircuitry (Silbersweig et al. 1995, Dierks et al. 1999); one of these studies also strongly implicated activation of hippocampal structures as well (Silbersweig et al. 1995). However, it is not clear if this distributed pattern of activation produces auditory hallucinations or arises as a consequence of some other pathophysiological process. For instance, shifts in attention due to the experience of hallucinations (from outward focus to internal imagery focus) may produce regional brain activation effects. If so, brain activation of this type will accompany hallucinations but will not in itself signal a causal mechanism.

Transcranial magnetic stimulation (TMS) is a method whereby an electromagnet placed on a scalp transmits magnetic pulses or waves to a small portion of the brain. Repetitive transcranial magnetic stimulation (rTMS) delivered at a low frequency (once per second) has been shown to reduce the reactivity or excitability of the part of the brain stimulated and other brain regions functionally connected to the region stimulated (Chen et al. 1997, Wassermann et al. 1998). A pilot study using a double-blind, crossover design compared effects of active versus sham low frequency rTMS in schizophrenic patients with auditory hallucinations (Hoffman et al. 1999, 2000). Based on the aforementioned neuro-imaging studies, the left temporoparietal cortex, a brain region critical for speech perception, was stimulated. A total of forty minutes of rTMS was administered over a four-day period. Statistically robust improvements in auditory hallucinations following active rTMS relative to sham rTMS were observed. These rTMS data support the hypothesis that activation of speech processing neurocircuitry contributes to the production of these hallucinations rather than emerging as a 'downstream' consequence of some other pathophysiological process.

These findings still do not address why speech processing areas of the brain, such as temporoparietal cortex, are pathologically activated in patients with schizophrenia. To understand this process better, computer simulations of neural networks have been developed that reproduce some aspects of speech processing of continuous narrative speech (Hoffman and McGlashan 1997). These simulations suggest that the ultimate 'cause' of auditory hallucinations is a loss of neuronal connections within these networks. Reductions in cortico-cortical connectivity are a normal component of 'late' development in human association cortex that continues during human adolescence (Huttenlocher and Dabholkar 1997). There are empirical data suggesting that reductions in synaptic density accompany some forms of learning (Scheich et al. 1991). Along these lines, computer simulations demonstrate that selective pruning of connections within recurrent neural networks produces significant information processing advantages (Hoffman and McGlashan 1997). In this model, pruning connections caused the system to fill in some of the gaps due to ambiguous phonetic input. It is well known that ordinary continuous speech, when produced at normal rates, has significant phonetic ambiguity that we do not 'hear' because our brains are also very skillful in 'filling in the gaps.' When pruning was excessive in our simulated networks, however, the system began to produce outputs in the absence of any phonetic input at all, thereby simulating speech hallucinations. This model therefore suggests that the hallucinations of schizophrenic psychosis arise as a result of an otherwise normal neuro-developmental trend that is excessive or unregulated. An appeal of this model is that it provides an account for the characteristic age of onset of schizophrenia, namely late adolescence and young adulthood.

Overall, the 'release' and 'ictal' models of hallucination described by Benson and Gorman appear to be good starting points. A comprehensive understanding of their nature, however, needs to consider interactions between different brain regions as well as between different 'levels' of neural processes involving molecular and neuroplastic alterations on the one

hand, and large-scale neural networks dynamics on the other. In spite of our current limited understanding of hallucinations, this fascinating area is likely to provide important insights into the pathophysiology of brain disturbances, as well as the normal physiology of perception and consciousness.

See also: Phantom Limbs; Visual Imagery, Neural Basis of

Bibliography

Bell D S 1973 The experimental reproduction of amphetamine psychosis. *Archives of General Psychiatry* **29**: 35–40

Benson D F, Gorman D G 1993 The neural basis of hallucinations. *Current Neurology*, Vol. 13, Mosby-Year Books, pp. 265–87

Brown J W 1985 Hallucinations: Imagery and the microstructure of perception. In: Fredriks J A M (ed.) *Handbook of Clinical Neurology*. Elsevier, New York, Vol. 1, pp. 351–372

Castner S A, Goldman-Rakic P S 1999 Long-lasting pscyhotomimetic consequences of repeated low-dose amphetamine exposure in rhesus monkeys. *Neuropsychopharmacology* **20**: 10–28

Chen R, Classen J, Gerloff C, Celnik P, Wassermann E M, Hallett M 1997 Depression of motor cortex excitability by low-frequency transcranial magnetic stimulation. *Neurology* **48**: 1398–403

Crowley T J 1995 Hallucinogen-related disorders. In: Kaplan H I, Sadock B J (eds.) *Comprehensive Textbook of Psychiatry*. Williams and Wilkins, Baltimore, MD, Vol. 1, pp. 831–8

Dening T R, Berrios G E 1994 Autoscopic phenomena. *British Journal of Psychiatry* **165**: 808–17

Dierks T, Linden D E J, Jandl M, Formisano E, Goebel R, Lanfermann H, Singer W 1999 Activation of Heschl's gyrus during auditory hallucinations. *Neuron* **22**: 615–21

Foulkes D, Fleisher S 1975 Mental activity in relaxed wakefulness. *Journal of Abnormal Psychology* **84**: 66–75

Gloor P, Olivier A, Quesney L, Andermann F, Horowitz S 1982 The role of the limbic system in experiential phenomena of temporal lobe epilepsy. *Annals of Neurology* **12**: 129–44

Halgren E, Walter R D, Cherlow D G, Crandall P H 1978 Mental phenomena evoked by electrical stimulation of the human hippocampal formation and amygdala. *Brain* **101**: 83–117

Hoffman R E, Boutros N N, Berman R M, Roessler E, Krystal J H, Charney D S 1999 Transcranial magnetic stimulation of left temporoparietal cortex in three patients reporting hallucinated 'voices.' *Biological Psychiatry* **46**: 130–2

Hoffman R E, Boutros N N, Hu S, Berman R M, Krystal J H, Charney D S 2000 Transcranial magnetic stimulation of left temporoparietal cortex in schizophrenic patients reporting auditory hallucinations. *Lancet* **355**: 1073–5

Hoffman R E, McGlashan T H 1997 Synaptic elimination, neurodevelopment and the mechanism of hallucinated 'voices' in schizophrenia. *American Journal of Psychiatry* **154**: 1683–9

Horowitz M, Adams J 1970 Hallucination on brain stimulation: Evidence for revision of the Penfield Hypothesis. In: Kemp W (ed.) *Origin and Mechanism of Hallucinations*. Plenum Press, New York

Huttenlocher P R, Dabholkar A S 1997 Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology* **387**: 167–78

Ishibashi T, Hori H, Endo K, Sato T 1964 Hallucinations produced by electrical stimulation of the temporal lobes in schizophrenic patients. *Tohoku Journal of Experimental Medicine* **82**: 124–39

Lhermitte J 1922 Syndrome de la calotte du pédoncule les troubles pysychosensoriels dans les lesions du mesencephale. *Revue Neurologique* **38**: 1359–65 November 9

Ojemann G, Mateer C. 1979 Human language cortex: Localization of memory, syntax, and sequential motor-phoneme identification systems. *Science* **205**: 1401–3

Penfield W P, Perot P 1963 The brain's record of auditory and visual experience. A final summary and discussion. *Brain* **86**: 595–696

Pierce R C, Kalivas P W. 1997 A circuitry model of the expression of behavioral sensitization to amphetamine-like psychostimulants. *Brain Res. Brain Res. Rev.* **25**: 192–216

Porrino L J, Lucignani G, Dow-Edwards D, Sokoloff L 1984 Correlation of dose-dependent effects of acute amphetamine administration on behavior and local cerebral metabolism in rats. *Brain Research* **307**: 311–20

Robinson T E, Kolb B 1999 Alterations in the morphology of dendrites and dendritic spines in the nucleus accumbens and prefrontal cortex following repeated treatment with amphetamine or cocaine. *European Journal of Neuroscience* **11**: 1598–604

Scheich H, Walhausser-Franke E, Braun K 1991 Does synaptic selection explain auditory imprinting? In: Squire L R, Weinberger N M, Lynch G, McGaugh J L (eds.) *Memory: Organization and Locus of Change*. Oxford University Press, London, pp. 114–59

Silbersweig D A, Stern E, Frith C, Cahill C, Holmes A, Grootoonk S, Seaward J, McKenna P, Chua S E, Schnorr L, Jones T, Frackowiak R S J 1995 A functional neuroanatomy of hallucinations in schizophrenia. *Nature* **378**: 176–9

Spitzer M, Bohler P, Weisbrod M, Kischka U 1995 A neural network model of phantom limbs. *Biological Cybernetics* **72**: 197–206

Wassermann E M, Wedegaertner F R, Ziemann U I, George M S, Chen R 1998 Crossed reduction of human motor cortex excitability by 1-Hz transcranial magnetic stimulation. *Neuroscience Letters* **250**: 141–4

Wengel S P, Burke W J, Holemon D 1989 The sounds of silence? *Journal of the American Geriatrics Society* **37**: 163–6

Wolf M E 1998 The role of excitatory amino acids in behavioral sensitization to psychomotor stimulants. *Progress in Neurobiology* **54**: 679–70

R. E. Hoffman

Harlow, Harry Frederick (1905–81)

Open any Introductory Psychology textbook published since 1960, start thumbing through the pages, and most likely you will come across a photograph of a rhesus monkey infant clinging to a cloth-covered dummy adjacent to another dummy, this one with a wire surface and a protruding milk-filled bottle. Along with the photo will be a description of Harry Harlow's classic 'surrogate mother' experiments, in which he

dramatically demonstrated that tactile contact was more important than feeding in establishing a monkey infant's bond to its surrogate. Harlow's conclusions from these studies, that 'there is more to mother love than mere milk' (Harlow 1958), had important and lasting consequences. They effectively demolished the prevailing theoretical consensus that an infant's primary social relationship is formed as a direct consequence of the feeding process, be it through oral gratification, as advocated by orthodox psychoanalysts, or by its reinforcement history, the contemporary behaviorist view. More importantly, Harlow's findings also provided compelling empirical support for specific biological foundations underlying Bowlby's attachment theory, which went on to become the predominant view of social and emotional development over the latter third of the twentieth century. But Harlow's contributions to, and influences on, both psychology and a remarkably diverse range of other disciplines went far beyond his legacy as the father of the surrogate mother. His numerous ground breaking discoveries over a lifetime of research with monkeys have not only stood the test of time and scientific debate, but indeed provided the very basis for many new areas of scientific inquiry that still remain active in the year 2001.

Harry F. Harlow was born Harry F. Israel on October 31, 1905 in Fairfield, Iowa, where he spent his childhood years. He left this small Midwestern town to go to college, initially matriculating at Reed College for a year (as an English major) before transferring to Stanford University, where he changed his major and then earned both his BA (1927) and Ph.D. (1930) in psychology. His graduate advisors at Stanford, concerned that his family name might restrict his academic employment opportunities during a time of economic depression and latent anti-Semitism, suggested a name change. Harlow obliged, and later in life he often quipped that they had given him not only his degree but his very name as well. In the fall of 1930 he was hired as an Assistant Professor in the Department of Psychology at the University of Wisconsin in Madison, where he remained for the rest of his professional career. Following his retirement in 1974 Harlow moved to Tucson, Arizona, where he served as a Emeritus Professor of Psychology at the University of Arizona until his death on December 3, 1981.

Although Harlow's graduate training in comparative psychology had been largely limited to research with rodents, he began working with non-human primates shortly after his arrival at Wisconsin, carrying out his initial studies with a variety of monkeys and apes then residing in Madison's public zoo. He quickly discovered that the standard types of conditioning and memory tests then being used in studies with rats or pigeons were simply not adequate to tap the considerable cognitive capabilities of his non-human primate subjects. He and his students therefore began creating new tasks and test batteries that were substantially more challenging. Shortly after moving into his own laboratory building on the University campus, Harlow designed and built the first prototype of a apparatus that would profoundly alter the course of research in primate cognition and that continues to be used in primate laboratories around the world in the year 2001. This testing device was named the Wisconsin General Test Apparatus (WGTA); it enabled an experimenter to present a large number of discrete learning and memory tasks in a rapid and highly standardized fashion to subject after subject (Harlow and Bromer 1938). At the same time, he began directing most of his efforts away from broad cross-species comparisons to more focused studies using a single species—the rhesus monkey (*Macaca mulatta*), an unusually hardy and adaptive species that thrived in captivity.

Armed with a laboratory specifically designed to house and test primates, in which individual monkeys could be maintained in good physical health for year after year, and a testing apparatus that enabled the same monkey to be tested repeatedly under standardized conditions, Harlow began changing the field of animal learning. In the 1940s he and his students created a battery of learning and memory tasks that provided a standardized 'intelligence test' for his monkeys. He then proceeded to study cortical localization of learning capabilities by lesioning different brain areas and recording subsequent differential patterns of deficits across the various components of the test battery. This research established his credentials as a physiological psychologist and it opened up areas of inquiry and an experimental methodology that continue to be pursued in neuroscience research.

In 1949 Harlow made a major conceptual and methodological breakthrough with his discovery of 'learning sets' in rhesus monkeys. He found that monkeys presented with a long series of six-trial two-choice discrimination problems soon learned to achieve near-perfect performance on the second and all subsequent trials of each problem. This demonstrated unequivocally that the monkeys had acquired a general strategy for problem solution, i.e., an abstract concept (in his words, 'learning to learn'), rather than a product of simple associative learning (Harlow 1949). This finding rendered untenable the neo-Decartes characterization of animals as 'reflex machines,' and it ultimately has led to increasingly more sophisticated studies of primate cognitive capabilities that foster direct comparisons with those of young human subjects. Largely because of this discovery, Harlow was elected to the National Academy of Sciences in 1951.

Harlow became interested in the ontogeny of rhesus monkey cognitive capabilities in the early 1950s when he developed a program of research designed to precisely establish when specific cognitive capabilities first emerged during development and when they subsequently achieved adult levels. He demonstrated

that infant monkeys could be both classically and operantly conditioned and performed associative learning tests at adult levels as early in life as such tests could be administered. In contrast, monkeys did not achieve adult levels of performance on more abstract problems such as learning set and oddity until their juvenile and adolescent years, respectively (Harlow 1959). In the process of carrying out these studies he also observed that young monkeys would solve problems and otherwise engage their physical environment in the absence of appetite rewards. This led him to conclude that in many circumstances curiosity was a more powerful motivator than a reduction of any primary drive (Harlow 1953). These observations provided the conceptual foundation for the 'surrogate mother' studies cited above.

In the 1960s Harlow shifted the major focus of his research from primate cognition to the development of social behavior, both normal and abnormal. He developed the concept of 'affectional systems,' in which he characterized emerging relationships formed by monkey infants with their mothers, peers, and others in the social group as differing in qualitative fashion and serving complementary roles in the socialization process (Harlow and Harlow 1965). He and his students also studied the consequences of blocking the formation of different affectional systems via various degrees and duration of social deprivation initiated at birth (Harlow and Harlow 1962, 1969). These studies established convincingly the importance of early social experiences for the development of species-normative adult social activities, including reproduction and maternal behavior, and they provided the conceptual and methodological basis for current studies of early experience effects on brain as well as behavior. In recognition for this research Harlow was awarded the National Medal of Science, America's highest official scientific award, in 1967.

Harlow continued his investigations of abnormal behavior during his final years at Wisconsin, generating two more areas of research in the process. First, he and his students demonstrated that the devastating behavioral consequences of total social isolation for the first 6 months or year of life, previously thought to be permanent, could be largely reversed by exposing the monkeys upon emergence from isolation to conspecific 'therapists' deliberately chosen to not only be socially normal but chronologically younger than the isolates themselves (Suomi and Harlow 1972, Novak and Harlow 1975). These findings directly refuted a strict 'critical period' interpretation of social deprivation effects, providing compelling evidence of developmental plasticity and once again opening up a new area of study that remains at the forefront of developmental neuroscience research today. At the same time, Harlow set out to create a rhesus monkey model of human affective disorders, relying on earlier observations by Rene Spitz and Bowlby that certain forms of social deprivation were associated with infant and childhood depression. He and his students used a variety of social separation procedures to precipitate analogous behavioral states in young monkeys and then employed both social and pharmacological interventions to reverse to behavioral disorders (Harlow and Suomi 1974, Suomi and Harlow 1977). In recognition of these efforts, Harlow received the Kittay Scientific Award in 1975, the most prestigious award in the field of psychiatry at that time.

Harlow's influence on the field of psychology extended far beyond his empirical contributions, which resulted in over 323 original scientific publications over his 40-year career (a complete bibliography of Harlow's works may be found in Suomi and LeRoy 1982). He served as Editor for the *Journal of Comparative and Physiological Psychology* for over a quarter century. He also took an active role in many scientific societies, including the National Academy of Sciences, the Society for Experimental Psychology, and the American Psychological Association, for which he was elected President in 1958. Harlow served as the Chief Psychologist for the US Army during the Korean War, and he subsequently was appointed to numerous US government planning and review committees. He played a major role in the establishment of the US Regional Primate Research Center System in the early 1960s and was instrumental in establishing the University of Wisconsin as the site for one of the 7 original Centers, which he directed from 1964 from 1971.

Despite all of his research activities and administrative responsibilities, Harlow maintained an active teaching role throughout his career. Not only did he train numerous graduate students and postdoctoral fellows (a total of 35 students completed their Ph.D. dissertations under his tutelage), but he was also legendary as an undergraduate instructor, especially for Introductory Psychology, which he taught nearly every year during his tenure at Wisconsin. In addition, Harlow worked hand-in-hand with veterinary and technical personnel to establish numerous basic advances in primate husbandry, including the development of the first self-sufficient primate colony, the first large-scale primate neonatal nursery, and various quarantine and screening procedures that are still used.

Harlow was a colorful character who enjoyed dealing with controversy. He delighted in generating empirical findings that directly challenged existing theoretical dogma. In addition to his considerable skills as a writer and editor, he was a gifted public speaker, who for years was one of the most sought-after scientific lecturers in the country. As a result, he was able to present his findings not only to psychologists around the world but also to representatives of numerous other scientific disciplines, as well as to lay audiences. Indeed, 20 years after his death Harlow's scientific contributions were still cited not only by

comparative, physiological, developmental, cognitive, and clinical psychologists but also by others in fields ranging from anthropology to zoology. Clearly, he was much more than the 'father of the cloth mother.'

See also: Animal Cognition; Learning to Learn; Psychologism, History of

Bibliography

Harlow H F 1949 The formation of learning sets. *Psychological Review* **56**: 51–65

Harlow H F 1953 Mice, monkeys, men, and motives. *Psychological Review* **60**: 23–32

Harlow H F 1958 The nature of love. *American Psychologist* **13**: 673–85

Harlow H F 1959 The development of learning in the rhesus monkey. *American Scientist* **47**: 459–79

Harlow H F, Bromer J 1938 A test-apparatus for monkeys. *Psychological Record* **2**: 434–6

Harlow H F, Harlow M K 1962 Social deprivation in monkeys. *Scientific American* **207**: 136–46

Harlow H F, Harlow M K 1965 The affectional systems. In: Schrier A J, Harlow H F, Stollnitz F (eds.) *Behavior of Nonhuman Primates*. Academic Press, New York, Vol. II pp. 287–334

Harlow H F, Harlow M K 1969 Effects of various mother-infant relationships on rhesus monkeys behaviors. In: Foss B M (eds.) *Determinants of Infant Behaviour*. Methuen, London, Vol. IV pp. 15–36

Harlow H F, Suomi S J 1974 Induced depression in monkeys. *Behavioral Biology* **12**: 273–96

Novak M A, Harlow H F 1975 Social recovery of monkeys isolated for the first year of life: 1. Rehabilitation and therapy. *Developmental Psychology* **11**: 453–65

Suomi S J, Harlow H F 1972 Social rehabilitation of isolate-reared monkeys. *Developmental Psychology* **6**: 487–96

Suomi S J, Harlow H F 1977 Production and alleviation of depressive behaviors in monkeys. In: Maser J, Seligman M E P (eds.) *Psychopathology: Experimental Models*. W H Freeman, San Francisco, CA pp. 131–73

Suomi S J, LeRoy H A 1982 In memoriam: Harry F. Harlow (1905–1981). *American Journal of Primatology* **2**: 319–42

S. J. Suomi

Hart, Herbert Lionel Adolphus (1907–92)

Herbert Lionel Adolphus Hart was born on July 18, 1907 in the spa town of Harrogate where his family, which was Jewish, ran a furrier's business. Hart's was the first generation of the family to receive a university education: he studied at New College, Oxford, where he took an outstanding First Class degree in Greats (ancient history and philosophy) in 1929. He then qualified as a barrister, and enjoyed a successful practice at the Chancery Bar throughout the 1930s. In 1940, he became a civil servant working in military intelligence; in 1941 he married Jenifer Fischer Williams, a civil servant in the Home Office, who was later herself to become an academic.

During the war, Hart's keen interest in philosophy was stimulated by regular conversations with Stuart Hampshire and Gilbert Ryle, who were working in a related Department, and who introduced him to the new movements in linguistic philosophy. In 1945, after some hesitation, but encouraged by his close friend Isaiah Berlin, Hart accepted a fellowship in philosophy at New College. Despite his 16 years outside the academy, he established himself quickly as a leading member of the influential group of analytic philosophers, including Ryle, J. L. Austin, G. A. Paul, and Friedrich Waissman, working in Oxford at the time. Such was his reputation that, notwithstanding having published only one article and a book review, he was in 1952 elected Professor of Jurisprudence at University College, Oxford. During his tenure of the Chair, he revived what had become an intellectually sterile subject, bringing to jurisprudence the rigor of a philosophical method of which it had been deprived since the days of Bentham and John Austin, and introducing the fresh insights of Wittgenstein's and, primarily, J. L. Austin's philosophy of language.

Outside the UK, his reputation was especially strong in the US, where he spent several periods, notably at Harvard in 1956–7, and in Israel, which he first visited in 1964. He was elected President of the Aristotelian Society in 1959 and Fellow of the British Academy in 1962. In 1968 he resigned from the Chair, taking up a Senior Research Fellowship at University College, Oxford, in order to devote himself to the editing and analysis of the works of Bentham. From 1967–73 he served as a member of the Monopolies Commission; he also chaired a committee that made important recommendations for the reform of Oxford's governance in the light of student unrest in the late 1960s. From 1973–8 he was Principal of Brasenose College; following his retirement he returned to University College as an honorary fellow. He died in Oxford on December 19, 1992.

1. Hart's Contribution to Legal, Moral, and Political Philosophy

Hart's intellectual contribution lies in two distinct but related fields. First, he revived the tradition of analytical jurisprudence, exploring both the concept of law in general and a number of specific concepts—causation, rights, obligation, responsibility, rules—which structure legal reasoning and doctrines. In this work, Hart emphasized J. L. Austin's assertion that one may use a 'sharpened awareness of words to sharpen our perception of phenomena.' He insisted,

however, that the task was not one of simple definition: rather, the aim was to build a theory or elucidation of the relevant concepts by examining their operation within the practical contexts in which they are used, and paying heed to the 'open texture' of language, with its core meaning surrounded by an unsettled penumbra. Perhaps the most outstanding example of these methodological commitments lies in Hart's and Honoré's monumental *Causation in the Law*, first published in 1959, which sets its conceptual analysis of causation within the context of hundreds of legal cases, and which draws criteria for the identification of causes out of the common sense understandings which underlie everyday usage in the relevant contexts.

Hart's second contribution was in the fields of normative legal, moral, and political philosophy. Building out from his analytic conceptions, he made highly influential interventions in normative and policy debates about matters such as capital punishment, abortion and the regulation of homosexuality. His treatment of these issues was located within a tradition stretching back to Hobbes, Hume, Locke, and, most obviously, J. S. Mill, to which Hart contributed, among other things, an elaboration of principles limiting the proper scope of state action, a novel account of the justification of punishment, and an analysis of the relationship between utility and rights.

To understand why Hart's work in analytical jurisprudence had such an extraordinary impact, one has to understand something about the development of jurisprudence in the century preceding his election to the Oxford Chair. From the late eighteenth to the mid-nineteenth century, the utilitarian thinkers Jeremy Bentham and his disciple John Austin had developed an influential theory of law known as 'legal positivism.' Rejecting the idea that law derives its authority from God, or from some metaphysical conception of nature or reason—so-called 'natural law'—Bentham and Austin argued that law is essentially man-made: it is a command, backed by sanctions, issued by a political superior or sovereign, to whom the populace is in a habit of obedience. This positivist theory had, however, been little developed since Austin's death. The one exception was the brilliant Austrian jurist Hans Kelsen, whose theory of law as a system of norms was, however, mainly influential on the continent of Europe. In Britain, the slow development of legal education in universities contributed to the impoverishment of the jurisprudential tradition, which in the first half of the twentieth century concerned itself primarily with the decontextualized analysis of legal terms such as ownership, possession, or the corporation. This analysis was not enriched by any systematic exchange with philosophy, nor was there any attempt to link it to any broader idea of the nature of law, or to consider how technical legal concepts assisted law to serve its various social functions. In the USA, by contrast, jurisprudence had moved in a sociological direction which arguably emphasized the empirical at the expense of a sustained examination of the conceptual dimensions necessary to an analysis of law's distinctive normativity. Prescriptive questions about what purposes law ought to pursue were left to the attention of moral and political philosophy—the latter itself a field that was relatively stagnant at this time.

Hart's first single-authored book, *The Concept of Law* (1961), which stands as his most important work, claimed to provide a general, descriptive theory of law which was at once a contribution to 'analytical jurisprudence' and to 'descriptive sociology.' In other words, he sought to elucidate a concept of law that would be relevant to an understanding of all forms of law, wherever or whenever they arose. In pursuing this project, Hart returned to the insights of Austin and Bentham, but—in a crucial innovation—combined their methods with those of the new linguistic philosophy represented by J. L. Austin and Ludwig Wittgenstein.

The nub of Hart's theory is the startlingly simple idea that law is a system of social rules structurally similar to the rules of games such as chess or cricket. The rules are of different kinds, with complementary functions. Some—'primary rules' such as criminal laws—directly govern behavior; others—'secondary rules'—provide for the identification, interpretation, and alteration of the former. Distinctive among these 'secondary' rules is the 'rule of recognition.' This rule provides criteria by which officials identify valid rules belonging to the legal system. It, therefore, also serves to identify the content of one particular legal system. Yet it is not, Hart argued, itself a valid rule of the system. Rather, it exists in the constellation of official attitudes and practices. The existence of a legal system is hence, fundamentally, a complex matter of social fact.

Like his nineteenth-century counterparts, Hart therefore insisted that law is a social, human invention: though legal rules generate genuine obligations, they are not moral rules. Their authority derives not from their content but from their source, which lies in a distinctively institutionalized system of social recognition. For example, the rule that we should drive on the left is authoritative not because there is any intrinsic value to driving on the left. Rather, it is because the rule can be identified in accordance with an agreed set of criteria for recognition, such as parliamentary enactment or judicial precedent. Precisely the same is true, moreover, of legal rules that overlap with moral standards: the legal prohibition on murder is not the same as, and derives its validity in a different way from, the moral injunction against killing. Hart also teased out, however, what he saw as important grains of truth in the natural law tradition. While he insisted that the connection between law and morality was a contingent one, he also pointed out that a rule-based legal order brings with it a distinctive conception of formal justice. He also argued that,

given certain features of human being—our physical vulnerability and limited altruism, and the limited resources of the natural world, for example—legal systems, given their purpose of facilitating human survival, would have to include a core set of rules governing—justly or unjustly—matters such as physical integrity and property. This he termed, perhaps misleadingly, 'the minimum content of natural law.'

Hart's account of how legal rules are recognized as valid served, on the other hand, to distinguish law from a mere system of force, or 'orders backed by threats.' For—crucially—according to Hart, legal rules have not only an external but also an internal aspect: we know that a rule is in existence not only because it is regularly observed, but also because those subject to it use it as a reason or standard for behavior, criticizing themselves or others for breaches of the rules. Here the linguistic philosopher's focus on the understandings embedded in ordinary usage becomes important. For example, Hart explores the distinction between habitual behavior (going to a café on Sunday morning) and rule-governed behavior (going to Church on Sunday morning); between being obliged to do something (handing over money because someone threatens to kill you if you refuse) and having an obligation to do it (paying your taxes). Of crucial importance to his status as a positivist is his insistence that the decision to adopt an 'internal attitude' to legal rules may be based on a variety of reasons, ranging from moral commitment to mere pragmatism: indeed, for a legal order to exist, it is theoretically only necessary for the officials of the system to adopt an internal attitude to its rules. In his elucidation of the internal attitude, Hart claimed to have detached the concept of law from that of sanction: the normativity of law derives from the internal aspect of rules and not from the threat of punishment for their breach.

By moving from the early positivist notion of law as a sovereign command to the notion of law as a system of rules, Hart produced a theory that resonated with the impersonal idea of authority embedded in modern democracies, providing a remarkably widely applicable rationalization of the nature of law in a secular and pluralistic world. He also offered an account of legal validity that purported to explain the (limited) sense in which citizens have an obligation to obey the law. Unlike the 'command' theory, it accounts not only for laws which impose duties but also the wide variety of laws which confer powers upon or create normative facilities for citizens: the law of property, contract, marriage, or succession, for example.

Hart also succeeded in reviving the idea that the project of jurisprudence should be not only to provide a clear, descriptive concept of law; but also—though as a separate project—to elaborate a framework for systematic criticism of law and for law reform. Hart's work became increasingly concerned with these prescriptive questions: from the late 1950s on he published highly influential papers dealing with issues such as the punishment of those who had committed what would now be regarded as serious crimes in Nazi Germany; capital punishment, the decriminalization of homosexuality, and the legalization of abortion. In these works, and in his widely publicized debate with the conservative judge Patrick Devlin, he regenerated a liberalism that finds its roots in J. S. Mills's *On Liberty* (1859). In *Law, Liberty and Morality*, Hart (1963) argued, as against Devlin, that democratic states are not entitled to enforce moral standards for their own sake. With the exception of special cases where paternalistic legislation can be justified, or where grave public offence is at issue, the state should respect individual freedom, intervening only to punish the commission of tangible harms. This books stands, over 30 years after its publication, as the resounding late twentieth-century statement of the most influential genre of liberal social policy. The philosophical ideas on which it rests are articulated closely with those expressed in Hart's conception of rights as founded in respect for individual human liberty and choice ('Are there any natural rights?'), and in his argument that consequentialist aims of punishment should be tempered by considerations of individual fairness in the distribution and quantum of punishment (Hart 1968).

2. *Hart's Importance for the Social Sciences*

Since the mid-1960s, Hart's work has been subject to intense scrutiny and criticism. The most significant criticism has come from two different directions. First, from the direction of 'natural law' theory, it has been argued that, ultimately, the reason why those subject to legal rules adopt an 'internal attitude' to them has to do with moral factors, and that consequently Hart's restricted sense of 'legal obligation,' and with it his distinction between law and morality, collapse. Second, from the direction of the social sciences, it has been argued that although the techniques of analytic philosophy may establish that law is simply one form of social rule, the further question of just what is distinctive about legal as opposed to other social rules can only be understood in terms of historical and social facts in which Hart shows surprisingly little interest. Notwithstanding these criticisms, there is strong reason to think that Hart's work has real significance for the social sciences.

First, in reviving the tradition of descriptive legal theory—the attempt to provide a systematic elucidation of law—Hart illuminated an institution whose distinctive importance in the organization of modern societies has become a fundamental preoccupation of the social sciences. In spite of his universalistic philosophical ambitions, and of the undoubted tension in his work between his 'analytic' and his 'descriptively sociological' methods, his theory can be argued to provide a persuasive account of the distinctive nature of legal authority—law's normativity—while main-

taining that law's existence is, ultimately, a matter of social fact and hence susceptible of empirical inquiry.

Second, therefore, in elucidating the nature of law as a social institution, Hart arguably provides precisely the sort of legal theory which social scientists need in incorporating law into broader social analyses. In his insistence on the structural continuity between legal and other social rules yet on law's distinctive mode of institutionalization, Hart provides a vision that is at once analytically well delineated yet relatively modest in terms of its substantive assumptions. Hence, the feature of his theory which might at first sight seem least sympathetic from a social science point of view—its pretension to provide a universal theory of law—may in fact be argued to be one of its strengths. For it provides a clear model for the purposes of testing hypotheses about, for example, the relationship between legal and other forms of social authority. This idea of his concept of law as, at root, a social science model is backed up by his own insistence, in terms reminiscent of Weber (though, in a significant ambiguity, also of Aristotle), that he is delineating a 'central case' with reference to which other, 'penumbral' cases can be identified and analyzed rather than 'banished to another discipline.'

Third, in his insistence on the internal aspect of rules, and hence on the idea that the social phenomenon of law can only be understood by reference to the attitudes and understandings of the agents who are its subjects, Hart's thinking meshed with a cluster of theoretical developments in disciplines such as sociology and anthropology. In this respect, his theory is a vast step forward from the cruder empiricism of the nineteenth-century positivists. Although it is not something with which Hart himself was centrally concerned, this feature of his method undoubtedly opens up the possibility of an analysis of the symbolic aspects of legal power and authority such as would be central to any adequate social science account of legal practices. In his emphasis on the importance of linguistic usage, moreover, his ideas relate, albeit obliquely, to the so-called 'linguistic turn' in much social theory of the late twentieth century.

Finally, in his revival of the utilitarian tradition of systematic critique of law and argument for its reform, Hart provided an example of the way in which conceptual analysis can be marshaled in aid of social policy. Clear thinking and reason—Hart's deepest intellectual commitments—are used, in an apparently modest and sometimes painstaking way, to clear the confusion in social and political debate, opening up in often unexpected ways paths forward in the development of policy and the construction of social consensus. Though Hart's own work is surprisingly little concerned with empirical—social, psychological, economic—evidence, his arguments often serve to clarify the relevance of just this sort of social science data. Though the fulcrum of his work lay firmly at the analytical–philosophic rather than the descriptive–sociological end of the spectrum, he can, therefore, justly claim a place among those whose thinking has made a significant contribution to the social and behavioral sciences.

Bibliography

Bayles M D 1992 *Hart's Legal Philosophy*. Kluwer, Dordrecht, The Netherlands

Gavison R (ed.) 1987 *Issues in Contemporary Legal Philosophy: The Influence of H. L. A. Hart*. Clarendon Press, Oxford, UK

Hacker P M S, Raz J (eds.) 1997 *Law, Morality and Society: Essays in Honour of H. L. A. Hart*. Clarendon Press, Oxford, UK

Hart H L A 1955 Are there any natural rights? *The Philosophical Review* **64**: 175–91

Hart H L A 1961/1994 *The Concept of Law*, 2nd edn. Clarendon Press, Oxford, UK

Hart H L A 1963 *Law, Liberty and Morality*. Stanford University Press, Stanford, CA

Hart H L A 1965 *The Morality of the Criminal Law*. Oxford University Press, Oxford, UK

Hart H L A 1968 *Punishment and Responsibility*. Clarendon Press, Oxford, UK

Hart H L A 1982 *Essays on Bentham*. Clarendon Press, Oxford, UK

Hart H L A 1983 *Essays in Jurisprudence and Philosophy*. Clarendon Press, Oxford, UK

Hart H L A, Honoré A M 1959/1985 *Causation in the Law*, 2nd edn. Clarendon Press, Oxford, UK

MacCormick N 1981 *H. L. A. Hart*. Edward Arnold, London

Martin M 1987 *The Legal Philosophy of H. L. A. Hart*. Temple University Press, Philadelphia, PA

Mill J S 1859/1974 *On Liberty*. Penguin, Harmondsworth, UK

N. Lacey

Hate Crimes

1. What is a Hate Crime?

There is no universally accepted definition of what precisely constitutes a hate crime, nor agreement as to the precise conceptual boundaries that delimit hate crimes from mundane crimes. The sole factor stressed uniformly is the motivation of the perpetrator. In hate crimes the principal motivation is one of negative affect (animus) toward the victim because of his or her skin color, ethnicity, religion, sex, national origin, or sexual orientation, for example. It is this animus that drives the criminal behavior rather than more common motives of pecuniary gain, retribution, jealousy, or status seeking. Thus, any crime may become a hate crime if the perpetrator's motivation is animus toward the victim based on some socially constructed category.

Some care must be taken to differentiate two seemingly related circumstances: (a) situations where the fundamental cause of the behavior is hatred; and (b) situations where hatred is coincident, with the primary motivation being something other than hatred. Presumably, it is only a true hate crime if the former condition exists. In other words, having extreme negative affect founded on skin color, ethnicity, religion, sex, national origin, or sexual orientation is both a necessary and sufficient condition for a hate crime.

In terms of behavior, hate crimes range from spray-paint vandalism and verbal harassment to physical assault and murder. The perpetrators may be single individuals, organized or unorganized groups, or entire nation-states. Similarly, the victims may be solitary individuals, small groups, or large aggregates of human population. In fact, a hate-crime universe can be conceptualized in terms of a three dimensional space: along one dimension is the degree of aggregation or organization of the perpetrator (ranging from single persons to nation-states), along another dimension is the degree of aggregation of victims (from single persons to nation-states), and the third dimension is the degree of severity (extending from vandalism to lethal sanctioning). Within this heuristic model, behavior can be located as diverse as a sole actor desecrating a synagogue with swastikas, to nation-states with genocidal policies aimed at despised populations. Yet underlying each is the assumption that the primary motive is negative affect based on a socially constructed status such as race or religion.

While the term hate crime, and the public notoriety it has received, has been particularly discernible during the 1980s and 1990s, the phenomenon is far from recent and has long been a frequent factor in human societies across the globe. This century alone has witnessed the genocide of Christian Armenians at the hands of Muslim Turks (1915–18), the Nazi attempt to exterminate European Jewry and the Roma, the mass murder of Vietnamese and Cham Muslims by the Khmer Rouge in Cambodia in the 1970s, the 1994 Hutus inspired genocide of Tutsis in Rwanda, and the Serbian 'ethnic cleansing' of Muslims from Bosnia. These are but a few examples of mass-hate crimes involving large numbers of perpetuators and with their victims experiencing the most severe levels of sanctioning. Often, however, public attention has been drawn to smaller-scale events involving fewer victims: for example, in 1991 skinheads in Hoyerswerda, Germany attacked several Asians and fire-bombed a hostel used by asylum-seekers (similar attacks were made on immigrants and refugees in many, if not most, Western European countries during the 1980s and 1990s)—the burning of some African-American churches in the American South during the early 1990s—or the 1998 lynching of James Byrd Jr, a 49-year old African-American in Texas by three white supremacists.

In sum, hate crimes can range from killing of genocidal proportions perpetrated by a nation state to verbal harassment by a single actor. They are frequent in human societies globally.

2. *Why Hate Crime?*

Most theories of hate crime involve, implicitly or explicitly, the notion of 'threat': that hate crime perpetrators perceive members of the target group as representing danger or peril to their life chances, lifestyles, social status, or belief systems. This threat does not have to be 'real' in any demonstratively physical sense—it only needs to be thought to be real to motivate some persons to commit crimes of hate.

Macropolitico-economic theories emphasize the importance of large-scale changes in the economic, political, or demographic arenas in generating apparent threats to the status quo. Typically these theories postulate that social change brings uncertainty, disruption, and competition between dominant and emergent groups for economic standing or political power. That contested terrain breeds fear and motivates some to strike out against members of the perceived threatening group. The effect of intergroup competition (threat) seems to be especially pronounced when the target group is culturally dissimilar from the dominant one. This is a point emphasized by cultural explanations of hate crime. Cultural theories stress threats to traditional or dominant belief-systems. Differences between the dominant culture and the target group subculture are seen as incompatible, if not totally hostile—the culture and belief systems of the victim group may be viewed as contaminating the dominant culture and in that sense, representing a threat to the cultural status quo. It is quite likely, for example, that the hate crime directed at guest workers, asylum seekers, and refugees in many counties was rooted in both the fear that new immigrants would take jobs away or lower wages of the native labor force and in the fear that a 'foreign' culture would fundamentally challenge traditional values and beliefs.

3. *Hate Crime and Hate Speech Laws*

Hate crime laws often take the form of sentence enhancements to penalties for crimes motivated by animus. For illustration, in the United Kingdom the 1998 Crime and Disorder Act provides that the maximum sentence for common assault increase from 6 months to two years if the assault was racially motivated. In the USA, the vast majority of states have enacted legislation providing for sentence enhancements for some hate crimes (but not others). At the federal level, the Hate Crimes Sentencing Enhancement Act of 1994 increased sentencing by about one-third when a crime is proven to have been hate motivated and occurred on federal property.

In many countries, hate speech, as well as hate crime, is criminalized. In the UK, speech or writings that may provoke racial animosity are prohibited by the Public Order Act. Similarly, Section 319 of the Criminal Code of Canada prohibits deliberately inciting or promoting 'hatred against any identifiable group.' There are comparable provisions in the criminal codes of Austria, Poland, Sweden, France, and Ireland. The Indian penal code not only criminalizes racial and religious-hate speech but caste-hate speech as well. In Germany it is illegal to incite violence or racial hatred against groups or minorities, and it is a crime to deny the Holocaust. Likewise, 1995 laws in both Switzerland and Belgium prohibit denying or defending genocide. There are no comparable laws against hate speech in the USA, however.

4. *Trends in Hate Crimes*

There is considerable rhetoric by advocacy groups that many countries are undergoing an unprecedented wave of hate crime, yet there is little undisputed empirical evidence supporting that claim. All authorities acknowledge that existing data on hate crime are fraught with errors, so any apparent rise in frequency of hate crimes may be due to: (a) a more sensitized public willing to report incidents of hate that had gone previously unreported; (b) the media and advocacy groups more frequently publicizing hate crimes; (c) an actual increase in the incidence of hate crimes; or (d) some combination of these. Where 'official' statistics are offered, it is suspected that they too may be seriously compromised because of under reporting of hate incidents by victims, the reliance upon local authorities to recognize and define an incident as a hate crime, and changes and ambiguities in the definition of what constitutes a hate crime. Given the shortcomings of official hate crime data and the unknown reliability and validity of anecdotal data from advocacy groups, it would seem impossible to state definitively whether hate crimes are actually increasing in frequency, decreasing, or remaining the same.

Given that caveat, what meager empirical data exists suggests that in the USA the frequency of hate crime incidents increased from the early 1990s until 1996, and has been declining since. In most of Western Europe, the trend appears that hate crimes may have become somewhat less frequent toward the end of the 1990s than they were earlier in the decade, with the possible exception of the UK.

5. *Victims of Hate Crimes*

While time trend data on hate crimes are especially problematic, a cross-sectional inspection of official reports of hate crime can illuminate who were the victims. In many European countries during the time 1970–2000, reported hate crime has targeted immigrants, refugees, asylum-seekers, and guest workers. Additionally, there have been many attacks on Jewish cemeteries and synagogues. In the USA, the situation is somewhat different.

Based on hate crimes reported in 1998 by 10,730 local law-enforcement agencies covering a population of 216,235 million, the US federal government cites 7,755 incidents of hate crime involving 9,722 victims. Of these incidents, 56 percent involved race hatred, and of those over two-thirds were anti-black. While race-hatred incidents were the most common, incidents of religious hatred were frequent, and 78 percent of these incidents involved anti-Jewish activity. Reports of sexual orientation bias were also common, with 67 percent of these incidents targeted at male homosexuals. Of all the hate crimes reported in 1998, 62 percent (4,832 incidents) were incidents that were anti-black, anti-Jewish, or anti-gay male. Of those offenses where the suspected offender's race was known, whites committed 76 percent of the hate crime offenses.

6. *Consequences of Hate Crimes*

Being a victim of virtually any crime extracts a definite psychological toll, but there is some evidence that the emotional mark of hate crimes may be more severe and longer lasting than the impact of normal (nonhate motivated) crime. Furthermore, some argue that when one member of the targeted group is victimized by hate crime, then all members of the group are adversely affected—that is, the negative effect of hate crime permeates the victim community more thoroughly than other forms of victimization. Because of the very nature of hate-motivated crime, a poignant and unmistakable message is delivered to both the victim and his or her community.

See also: Disgust, Psychology of; Genocide: Anthropological Aspects; Genocide: Historical Aspects; Heterosexism and Homophobia; Mass Killings and Genocide, Psychology of; Personality and Crime; Racism, History of; Racism, Sociology of; Religion: Morality and Social Control; Religion: Peace, War, and Violence; Religious Fundamentalism: Cultural Concerns; Social Trauma; Violence and Media; Violence: Public; Xenophobia

Bibliography

Berk R A 1994 What is a hate crime? In: Hamm M S (ed.) *Hate Crime: International Perspectives on Causes and Control.* Anderson, Cincinnati, OH, pp. v–ix

Grimshaw A D 1999 Genocide and democide. In: Kurtz L, Turpin J (eds.) *Encyclopedia of Violence, Peace, and Conflict*. Academic Press, San Diego, Vol. 2, pp. 53–74
Hamm M S 1994 Conceptualizing hate crime in a global context. In: Hamm M S (ed.) *Hate Crime: International Perspectives on Causes and Control*. Anderson, Cincinnati, OH, pp. 173–94
Jones T D 1998 *Human Rights: Group Defamation, Freedom of Expression and the Law of Nations*. Martinus Nijhoff, The Netherlands
Myra H I (ed.) 2000 *Ethnic Violence*. Greenhaven, San Diego
Jacobs J B, Potter K 1998 *Hate Crimes: Criminal Law and Identity Politics*. Oxford University Press, New York
Kelly R J, Maghan J 1998 Introduction. In: J Maghan R J, Maghan J (eds.) *Hate Crime: The Global Politics of Polarization*. Southern Illinois University Press, Carbondale, IL, pp. 1–21
Lawrence F M 1999 *Punishing Hate: Bias Crimes under American Law*. Harvard University Press, Cambridge, MA
Witte R 1995 Racist violence in Western Europe. *New Community* **21**: 489–500

E. M. Beck

Hayek, Friedrich A von (1899–1992)

Friedrich von Hayek is regarded generally as the most influential intellectual figure in the twentieth-century revival of classical liberalism (Kukathas 1990) and as the most articulate critic of socialist doctrines (Caldwell 1997). He also deserves recognition as one of the preeminent social philosophers of that century. In an age of increasing specialization within the social sciences and the humanities, he developed an approach to social analysis and social reform that integrates insights from economics, law, politics, philosophy, psychology, and evolutionary theory into an impressively coherent outlook at the foundations and the evolution of social order. Hayek's scholarly writings span more than six decades and include numerous books and articles, republished in a 19-volume edition of his collected works (Hayek 1988).

1. *Biography*

Hayek was born in Vienna on 8 May, 1899. He earned doctorates in law (1921) and in political economy (1923) at the University of Vienna, where he became *Privatdozent* for Political Economy in 1929. Interrupted by a year of study in New York (1923–4) he worked from 1921–6 as a legal consultant in a government office in Vienna, directed by his mentor Ludwig von Mises, in whose famous *Privatseminar* he participated and with whom he founded, in 1927, the Austrian Institute for Business Cycle Research, of which he was director until 1931. A series of invited lectures that Hayek gave at the London School of Economics (LSE) in 1931 led to his appointment to the Tooke Chair of Economic Science and Statistics at the LSE, a position Hayek held from 1932 to 1950, the year he was appointed Professor of Social and Moral Sciences at the University of Chicago's Committee on Social Thought. After retiring from Chicago in 1962 he taught as Professor of Economic Policy at the University of Freiburg, Germany, until 1967. An honorary professorship at the University of Salzburg from 1969 to 1974 brought Hayek back to his native Austria. In 1977 he returned as professor emeritus to Freiburg where he died on 23 March, 1992. In 1974 Hayek was awarded the Nobel Prize in Economics.

2. *Hayek's Social Philosophy*

About his personal development Hayek (1967, p. 91) has said that he was led 'from technical economics into all kinds of questions usually regarded as philosophical.' It was his early work on monetary theory, capital theory, and business cycle theory that earned him his appointment at the LSE, and it was about these themes that he engaged in controversies with John Maynard Keynes, Piero Sraffa and Frank H. Knight in the early 1930s. An important turning point in his scholarly interests is marked by his work, in the mid-1930s, on an English edition of selected contributions to the 'socialist calculation debate,' including Mises' early (1920) article on the subject (Hayek 1935). As Hayek has noted in retrospect, in preparing this edition he began to develop his growing interest in questions of the nature and unavoidable limitations of our knowledge, and in the issue of what these limitations imply for our theoretical efforts at understanding the social world as well as for our practical efforts at organizing our social life. The concern with these questions, often referred to as the 'knowledge problem,' was to become the core theme of Hayek's life work.

In Hayek's own assessment, the critical step in his move towards social philosophy was his 1937 article on 'Economics and Knowledge' (1948, Chap. 2), an article in which Hayek addressed what he describes as the 'central question of all social sciences,' the question of how the fragments of imperfect knowledge that exist dispersed in individual minds are coordinated in the social process to allow for productive cooperation among multitudes of persons. With his 1945 article on 'The Use of Knowledge in Society' (1948, Chap. 4), his most often cited publication, Hayek restated and extended his earlier argument, stressing that the 'problem of what is the best way of utilizing knowledge initially dispersed among all the people is at least one of the main problems of economic policy—or of designing an efficient economic system' (1948, p. 78). In these two articles, which he considered his most

original contributions to economics, Hayek developed a fundamental critique of two bodies of thought prevalent at the time, of formal equilibrium economics on the one hand and of socialist concepts of central planning on the other, censuring both of them for failing to account for the theoretical and practical difficulties that are posed by the imperfect, fragmentary, and dispersed nature of human knowledge.

In Hayek's view, the models of mathematical equilibrium economics provide only a pseudo-answer to the question that they are meant to illuminate, namely, how markets work. As he put it, with its 'assumption of a perfect market where every event becomes known instantaneously to every member' (1948, p. 45), formal equilibrium analysis begs the very question that an empirically contentful theory of market processes would have to answer, namely how 'a solution is produced by the interaction of people each of whom possesses only partial knowledge' (1948, p. 91). Ironically, the failure committed in concepts of central planning is of a very similar kind. By assuming perfect knowledge on part of the planners, advocates of central planning provide only a pseudo-answer to the practical problem of how effective coordination of economic activities can be achieved in the absence of markets. Assuming that all relevant knowledge is available to, and can be administered by, a planning agency, raises the question of how such an agency is supposed to come to terms with the fact that much of the relevant knowledge 'never exists in concentrated or integrated form, but solely as the dispersed bits of incomplete… knowledge which all the separate individuals possess' (1948, p. 77), as knowledge that can be activated only by these very individuals.

What is needed, Hayek argued, is a theory that accounts for 'the unavoidable imperfections of man's knowledge and the consequent need for a process by which knowledge is constantly communicated and acquired' (1948, p. 91). He called on his colleagues in the economics profession to turn their attention from the formal statements of equilibrium analysis to empirically contentful 'propositions about what happens in the real world' (1948, p. 46), propositions about how knowledge is actually acquired and communicated. While his call did not find much response in the profession, in his own work Hayek very much pursued the research agenda that he suggested.

Hayek has especially addressed the issue of how we acquire knowledge in his *The Sensory Order* (1952a), a most unlikely book for an economist and one of his least known, but nevertheless one that is of greatest systematic significance for his entire work. In this book (based on an early manuscript from Hayek's student days), subtitled 'An Inquiry into the Foundations of Theoretical Psychology,' Hayek develops a theory of the human mind, the essence of which is in amazing accord with modern cognitive science. Seeking to explain how the human mind establishes its connections 'between the input of (external and internal) stimuli and the output of action' (1988, p. 288), Hayek interprets the working of the mind as an adaptive process of constant classification and reclassification on many levels (sensory perceptions, emotions, concepts), as a process in which conjectural internal models of the outside world are adaptively formed and guide actions.

The issue of how knowledge is communicated and advanced in society is the theme of Hayek's theory of the market as a communication system (1976, Chap. 10) and of competition as a discovery procedure (1978, Chap. 12), a theory that forms the centerpiece of his work. The prices formed in the competitive market process, Hayek argues, signal to all market participants the relevant general information about changes in relative scarcities that they need to know in order to adjust their behavior adequately, utilizing their unique local knowledge of the particular circumstances of time and place, knowledge that could never be communicated to, and utilized by, a central planning agency. Beyond serving in such fashion as a device for utilizing dispersed knowledge, the market serves also as an arena for the continuous competitive exploration of new and potentially better solutions to economic problems, thus inducing the discovery and creation of new knowledge.

The relevance of Hayek's work, not only for economics but for the social and behavioral sciences more generally, is most clearly visible in his theories of spontaneous social order and of cultural evolution that he has developed as generalizations and extensions of his arguments on market processes and competition. The thrust of these theories is that human societies are highly complex phenomena, and that the processes of social coordination and societal evolution are inappropriately simplified and misconceived if they are approached in the spirit of the mechanistic models of physics (1952b). He argues that acknowledging the complexity of social phenomena means to acknowledge that all we can reasonably hope to achieve in the social realm, as in all areas of complex phenomena, are explanations of the principle and pattern predictions, by contrast to the explanation and prediction of specific events (1967, Chap. 2).

Hayek places his own social theoretical outlook in the tradition of the Scottish moral philosophers (David Hume, Adam Smith, Adam Ferguson and others), whom he credits with the discovery of 'the twin concepts of the formation of spontaneous orders and of selective evolution' (1988, p. 146). Central to this paradigmatic tradition is a methodological individualism and subjectivism, that is, on the one hand, the notion that social aggregate phenomena are, ultimately, to be explained in terms of the actions and interactions of individual human beings and, on the other hand, the insight that 'the analysis of what people do can start only from what is known to them' (1948, p. 44), that is, from their subjective view of the world. Equally central to it is the notion that the

productive coordination and division of labor that we observe in society is to a large measure not due to deliberate planning and conscious design, but to forces of spontaneous mutual adjustment and evolution, as are the rules and institutions on which societal cooperation is based. From these premises important conclusions are drawn with regard to the limits of our ability to improve our social condition by deliberate construction.

According to Hayek, there are essentially two basic ways in which the actions of individuals can be socially coordinated, namely either through spontaneous mutual adjustment within a framework of general rules of conduct, or through the commands of a central coordinator (1973, Chap. 2). As he argues, the respective merits and limits of the two kinds of order—spontaneous order and organization—have to do with the nature of their principal means of coordination, namely general rules versus commands. Command-based organization is an effective method of social coordination where specific, limited purposes are to be achieved. Because of its reliance on central decisions it is, however, severely limited in its capacity to utilize the knowledge dispersed among its individual participants. The advantage of rule-based spontaneous social orders is that, by allowing individuals to pursue their own ends within the general confines defined by the rules, they can utilize and activate to a much higher degree localized, dispersed knowledge. The price that has to be paid for their greater capacity for utilizing knowledge is that the particular outcomes that they generate must inevitably remain largely unpredictable. In Hayek's view, considering the respective merits and limits of the two kinds of order should lead us to conclude that, while organization can be a useful method for solving a multitude of specific, limited tasks, it cannot possibly be appropriate for the extended order of an economy or a society at large. There is ample room for organizations within the spontaneous order of society, but efforts to turn society itself into an organization can only end in disaster.

An important element in Hayek's theory of spontaneous social order is the insight that, while the particular outcomes of a rule-based spontaneous order remain necessarily largely unpredictable, the general patterns of its outcomes are determined systematically by the nature of its rules. There exists, Hayek emphasizes, a systematic interplay between the 'order of rules' and the 'order of actions' (1967, Chap. 4), that is, between the nature of the 'rules of the game' on which the interactions of the participating individuals are based, and the nature of the action-patterns that emerge from the individuals' behavioral choices within the rules, just as the rules of a sport allow us to predict—if not particular outcomes—the general patterns that we observe on the playing field. In fact, Hayek argues, our knowledge of the systematic relation between the order of rules and the order of actions is the principal instrument that we can employ to improve our condition, by choosing rules of the game that can be predicted to generate more desirable overall patterns of outcomes.

Hayek's emphasis on the basic facts of 'the dispersion and imperfection of all knowledge' leads him to be highly critical of a mind set that he labels 'rational constructivism' or 'constructivist rationalism' (1978, Chap. 1). Under this rubric he subsumes approaches to issues of social policy and social reform whose recommendations are based on a 'pretence of knowledge' (1978, Chap. 2), that is, on unjustified presumptions concerning the extent of our knowledge and our intellectual powers. The primary targets of Hayek's critique are approaches that, like socialist concepts of central planning, amount to the 'claim that man can achieve a desirable order of society by concretely arranging all parts in full knowledge of all the relevant facts' (1967, p. 88). Against such claims Hayek points out that, because it is simply impossible for us to take 'conscious account of all the particular facts which enter into the order of society' (1973, p. 13), we need to rely on general rules 'as the means to create order in social affairs' (1948, p. 19), and that our efforts at building a better society must largely rely on the indirect method of improving the rules and institutions under which we live, instead of on attempts to directly produce desired outcomes through specific interventions.

To be sure, in our efforts to improve the framework of rules and institutions under which we live, we need, Hayek stresses, to be no less careful to avoid the pitfalls of a pretense of knowledge. Since the desirability of the outcome patterns that particular rules or institutions tend to generate can only be judged appropriately in terms of their working properties over extended periods of time, it would be presumptuous to claim that we can easily design, at the drawing board, rules and institutions that serve us better than those that have evolved over generations. Just as, in the ordinary market, the discovery process of competition helps us to find out how our 'economic' problems can be solved better than before, competition in the realm of rules and institutions can serve the function of allowing us to find out what kinds of rule arrangements are better suited to allow for mutually beneficial cooperation among multitudes of persons. A process of competitive selection can here, just as in ordinary markets, help gradually to evolve solutions that are much more effective than what any rational design could have ever invented at the outset. In fact, so Hayek argues, the institutions on which our civilization rests, including the rules of the market, must be viewed as the outcome of a process of cultural evolution that extends over thousands of generations and that reflects the experience of many more trials and errors than the most wise and intelligent rule designer could ever consider. Our efforts at institutional reform and rule design should, therefore, always

be informed by due respect for the potential wisdom implicit in inherited institutions, and we should expect further improvement 'more from slow experimental piecemeal evolution than from redesign of the whole' (1964, p. 8).

The theory of cultural evolution is the theme to which Hayek devoted much of his later work (1979, Epilogue, 1988), and it is this part of his theory that has become a primary focus of discussions on his ideas. A central issue in these discussions is how the notion of competition among institutions—a notion that is at the heart of Hayek's ideas on cultural evolution—is to be specified, and under what conditions such institutional competition can be expected to generate 'good' rules, that is, rules with working properties that are desirable to those who are affected by them.

3. *Impact and Significance of Hayek's Work*

With his arrival at the LSE in the early 1930s Hayek became almost instantaneously a well-recognized figure in the economics profession, engaged in prominent controversies on economic theory with such opponents as John Maynard Keynes, and on the issue of rational economic calculation under socialism with Oskar Lange and others. As his involvement in the latter debate induced the shift in his interest towards general issues of social and political philosophy, he increasingly removed himself from the standard professional discourse. He turned his attention to a fundamental critique of the philosophical foundations and the political implications of socialist doctrines, reflected most prominently in the publication of his most famous book *The Road to Serfdom* in 1944, a book that quickly made him a widely known public figure, in particular in the USA, where a condensed version appeared in the *Reader's Digest* in 1945. While the book earned him public fame, it was met with considerable hostility among his colleagues in the economics profession, many of whom were, at the time, much more inclined to interventionist views than to Hayek's warnings against socialist concepts of social and economic planning. With the rise of Keynesianism in the 1950s and 1960s, Hayek's work fell almost into oblivion, in academia as well as in public debate, and when in 1960 his major restatement of the principles of classical liberalism, *The Constitution of Liberty*, appeared, it was widely regarded, even by sympathetic observers, as an outdated treatise, out of step with modern economic and political thought. It came as a surprise to many, including Hayek himself, when in 1974 he was awarded the Nobel Prize in Economics. Even if the fact that he had to share the prize with Gunnar Myrdal, an economist of a quite different orientation, reflected significant ambivalence on part of the Nobel committee, the award was an early indicator of a reawakening interest in classical liberal thought in general, and Hayek's work in particular, an interest that grew significantly in the later half of the 1970s, as a response to an increasing disillusionment with Keynesian economic policies and to accumulating evidence of the failure of 'real socialism.' In fact, Hayek became very much the symbol of what has been described as the 'modern rebirth of classical liberalism.' His work regained increasing attention in the academic community, his ideas influenced politicians like Margaret Thatcher and Ronald Reagan, and his publications found growing interest among dissidents and reformers in the disintegrating communist world. Hayek lived long enough to witness the final collapse of the communist regimes in Europe as a confirmation of his lifelong critique of socialist constructivism.

In modern theoretical economics Hayekian ideas play an important role in a number of approaches that seek to advance an alternative theoretical outlook to the orthodox equilibrium paradigm. His ideas on the nature of human knowledge and learning are finding increasing attention among scholars who seek a corrective for deficiencies of the traditional rational choice model. Recent developments in evolutionary economics, in the new institutional economics, and in the economics of complex adaptive systems, either explicitly draw on, or give support to core ideas of his work, to his theory of the market as a spontaneous order, his outlook on competition as discovery process, and his evolutionary approach to institutions.

The reception of Hayek's work in social sciences other than economics is still in its infancy, yet his approach to social theory may well come to play a significant role as a paradigmatic contribution towards a theoretically integrated social science. It was one of Hayek's major concerns that the disciplinary fragmentation of modern social science left what should be its central issue, namely 'the problem of an appropriate social order,' without adequate attention. As he put it, even though this issue 'is studied today from the different angles of economics, jurisprudence, political science, sociology, and ethics, the problem is one which can be approached successfully only as a whole' (1973, p. 4). More than any of his contemporaries in the social sciences Hayek has made an effort in his own work to address the fundamental issue of an 'appropriate social order' from a general social theoretical perspective that cuts across traditional disciplinary boundaries. With his ideas on the two kinds of social order, on the function of rules and institutions, and on cultural evolution he has made important steps towards a theoretically integrated social science.

See also: Economics, History of; Knowledge, Sociology of; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Socialism: Historical Aspects

Bibliography

Caldwell B 1997 Hayek and socialism. *Journal of Economic Literature* **35**: 1856–90
Hayek F A (ed.) 1935 *Collectivist Economic Planning*. Routledge, London
Hayek F A 1944 *The Road to Serfdom*. Routledge, London
Hayek F A 1948 *Individualism and Economic Order*. University of Chicago Press, Chicago
Hayek F A 1952a *The Sensory Order*. University of Chicago Press, Chicago
Hayek F A 1952b *The Counter Revolution of Science*. Free Press, Glencoe, IL
Hayek F A 1960 *The Constitution of Liberty*. University of Chicago Press, Chicago
Hayek F A 1964 Kinds of order in society. *New Individualist Review* **3**: 3–13
Hayek F A 1967 *Studies in Philosophy, Politics and Economics*. University of Chicago Press, Chicago
Hayek F A 1973, 1976, 1979 *Law, Legislation and Liberty*, 3 vols. University of Chicago Press, Chicago
Hayek F A 1978 *New Studies in Philosophy, Politics, Economics and the History of Ideas*. University of Chicago Press, Chicago
Hayek F A 1988 *The Collected Works of F. A. Hayek*. University of Chicago Press, Chicago
Kukathas C 1990 *Hayek and Modern Liberalism*. Clarendon Press, Oxford, UK
Mises L von 1920 Die Wirtschaftsrechnung im sozialistischen Gemeinwesen. *Archiv fuer Sozialwissenschaft* **47**: 86–121
Vanberg V J 1994 *Rules and Choice in Economics*. Routledge, London

V. Vanberg

Healing

In the context in which healing is studied here, it applies to any human or socially organized intervention having as its purpose the improvement of a physical or mental condition. In sociological analysis healing is judged by its intention, not by its effect. However, the differential success rates of opponent forms of healing are the basis of conflict between their practitioners and of the legitimization of the organizations within which they are practiced. The evidence is that while religious organizations are in other spheres of human life losing ground to secular agencies, there has been a recent increase of the resort to forms of healing that are alternative to secular medicine.

1. Types of Religious Healing

The widely used concept of shaman is a precise one, having been crystallized in a classic work by Mircea Eliade in the 1960s. A shaman is a person who enters a state of abnormal consciousness, being induced either by rhythmic music of by psychedelic drugs in the form of snuff or mushrooms. As a functionary in the business of healing, the shaman may be distinguished from looser and more popular concepts such as witch doctor and medicine man that suggest skills in the use of locally available medications without the element of trance or ecstasy. Again, the role of the shaman in healing is not that of magician: whereas the magician uses ritual or other mysterious means to control the course of natural events, the shaman is merely a channel or borrower of the powers of the universe through whom some benign outcome is effected. The word shaman is taken from a tribe in north Siberia but shamanism within this narrow conception may be found also in, for example, the upper Amazon, the Orinoco basin, Australia, and in south-east Asia.

Mental healing, which is based on the diagnosis of illness as a condition of the mind, may be traced within modern Christian practice to the theology of Emanuel Swedenborg (1688–1772), a Swedish scientist and initiator of the religious sect which bears his name. He perceived healing as a harmony of physical and spiritual realities, achievable by the adjustment of the body to the mind, the mind to the soul, and the soul to the angelic order. In mid-nineteenth century America the Swedenborgian idea of mental adjustment was current in the fashionable mesmerism of Charles Poyen, in the teachings of Phineas Quinby, and subsequently in the writings of his most famous patient, Mary Baker Eddy, the founder of Christian Science. In her seminal *Science and Health with Key to the Scriptures* (1875) she turned to Matthew 9, 2–8 to demonstrate the central tenet that sickness and pain were delusions curable by the correction of mental error.

The notion of faith healing implies a degree of prior belief, even confident expectation, either in the seeker or in the healer. Healing which is reckoned to depend on the commitment of the patient lends itself to explanations in terms of the will power of the individual and other psychological factors. But the term faith healing is sometimes applied also to practices in which the healer claims authority over evil spirits or practices the laying on of hands in confidence of a miraculous outcome.

Divine healing, by contrast, is believed to be an intervention that does not rely on the faith of those involved. These conceptions of healing are prevalent in various of the Christian sects though not exclusive to them. Most notable among these are the 'Foursquare' pentecostal churches of north America, one of the four squares being 'Christ the Healer.'

In recent years there has been a burgeoning interest in healing in the mainstream churches in which the term healing is used in a more general way to include acts of reconciliation of individuals or groups, either as a result of conflicts between themselves or in changing circumstances such as tragedy and bereavement. The ministry of healing which is particularly associated with the Church of England does not connote the sense of the miraculous which distinguish

divine healing and faith healing. It recognizes the skills of the minister who will accordingly remain conscious during the consultation.

The conciliation of mind and spirit that takes a variety of forms and is pursued by a still greater variety of methods is the essence of spiritual healing. It is not to be confused with spiritualist healing in which the practitioner is said to be in contact with spirits in another domain, such as that of the dead.

2. *The Social Context of Faith and Healing*

The pentecostal churches have flourished lately among the dispossessed of central and south America (Martin 1990, Chesnut 1997). The routine Marxist explanation is that the hope of a life hereafter enables the faithful to endure the discomforts of the present world. In Brazil, however, Chesnut found relatively less conviction in remote consolation than hope of immediate relief from the illnesses that attend poverty. The gravitation to the Pentecostal churches may to that extent be compared with the pilgrimages to Lourdes, to the sacred waters of the Ganges, and to the healing shrines and centers that abound in India and in the Catholic world.

What is more distinctive of recent years is the interest in ritual healing among the American middle class. Maguire (1988) recognizes in 'the metaphysical movement' the confluence of a number of nominally Christian traditions from New Thought, mesmerism, the mind-cure groups of the 1880s, Spiritualism and Christian Science, and an experimental interest in currents of spirituality that are or were native American, Latin American, Caribbean, and Polynesian. The turning to metaphysical healing in the suburbs has an intellectual basis involving an alternative pathology, a redefinition of illness and disharmony, and an open-minded disposition to eastern thought emanating from student subcultural movements in the late 1960s.

The host of life and death concerns which are addressed by hands-on therapy rituals in the religious life of middle America provide the agenda too for the ministry of healing in the Church of England. This neither trades on the expectation of the spectacular nor publishes its results against those of the medical services. The ministry operates, for example, in the rehabilitation of counseling and psychotherapy within the sacred domain, in the hospice movement pioneered in Britain by Dr Cicely Saunders, and in the provision of healing services within the regular ministrations of the mainstream churches.

Church-type healing is distinguishable from that of the sects in a number of ways: (a) the sectarian cult of the personality tends to elude church-type practitioners; (b) the expectation is not of the miraculous, nor are successes celebrated; (c) ministry to the dying will have the purpose of enabling them to die well, not to recover to full health; (d) needs and benefits are largely reserved within the private domain; (e) the ministry is integrated with the Church's sacraments; (f) practitioners tend not to claim divine gifts so much as seek to enhance their effectiveness through conventional training of knowledge and skills; (g) the Church's ministry of healing is reckoned to complement conventional medicine and not to supplant it. While the miraculous healing tradition is rooted in the Pauline formulation of spiritual gifts—notably in Corinthians 1, 12—and denotes the healer as a direct agent of divine power, the Churches' ministry of healing looks to the Gospel for a motivation but to contemporary science for a method.

This movement by the mainstream churches may be seen as an attempt to evolve a new role and purpose in a secularizing world. In Mexican folk religion, too, *curanderismo* has adapted to address current issues and aspirations such as love, marriage, and happiness.

Some religious communities in which the practice of healing is central may stand or fall by their results. Others, however, lay hands on the sick because they are directed to do so (Mark 16, 18) and not with the primary purpose of a validating outcome. In an age of accountability, of the perception of an audience as a market, and of the importance of reputation and the fear of litigation, there has been a systematic movement to professionalize the alternative practice of healing. Practitioners have formed themselves into associations, established courses of study and qualifications, written ethical codes, and introduced a corporate discipline. These developments have been much more evident within complementary medicine outside the churches than within them.

3. *The Charisma of the Healer*

There is a tendency for the personalities of those involved in healing in sects and new religious movements to be rather higher in public profile than those who practice therapy in the suburbs and within the ministrations of the churches. Their names and faces are better known and are used in drawing crowds to religious meetings. The consequences of divine workings through them are more spectacular. Their disposition to secular medicine is more rejectionist and there is a triumphalism about their healing stories. When Melvin Banks testifies to the miracles God has worked through him, he shows the picture of a woman holding aloft her walking stick in defiance of the doctors who said she would always need it. Similarly, Her Holiness Mataji Nirmala Devi Srivastava, the 'Divine Mother' and teacher of Sahaja Yoga passes on to her disciples healing skills, including the use of vibration therapy in the cure of cancer. In the rhetoric, if not always in the understanding of the faithful, it is made explicit that the power is divine and the healer a mere agent. The claim to be a channel of divine power

is of course a greater enhancement of charisma than the claim to possess skills. The shaman possesses knowledge about healing which is not available to the community and which signifies an exclusive relationship with the spiritual domain. However, no man or woman would describe himself of herself as a shaman: the sense is that the power is invested in the shaman and an ethic of humility restrains self-promotion (Doore 1988). Of course, humility can itself be an element of charisma. Many *curanderos*, the folk-healers of Mexico, believe their healing powers to be a gift from God, invested in them and giving them certain responsibilities to alleviate misery. They claim no credit for themselves and, far from rejoicing in their powers, they may even regret the burden of their responsibilities: albeit they are full-time and charge fees.

But charisma invites suspicion from the secular world and from mainstream churches. It is in view of the capacity to exploit the charisma in the role of healer, both Church and the medical establishment distance themselves from alternative healing practices of new religious movements.

4. *Conflicts of Faith and Secular Practice*

It is reported (2 Chronicles 16, 12) that King Asa 'in the thirty and ninth year of his reign was diseased in his feet, until his disease was exceeding great: yet sought he not unto the Lord but the physicians: ... and in the one and fortieth year of his reign Asa slept with his fathers.' The Chronicler implies an error of judgement on Asa's part with a fatal consequence. Asa lacked faith. In modern times the resort to secular medicine frequently is perceived to compromise religious conviction.

The *curanderos* are condemned with suspicion by the fundamentalist churches which attribute to the devil the mysterious source of their power. For its part the medical establishment has invoked both research and the law to counter the rising tide of religious healing. The British Medical Association (1986) issued a report warning against the use of healing in new religious movements to confirm allegiance; it contended that they had caused harm to recruits and their families. The report was not well received, either by politicians or by the BMA's own members, but it survives as a significant expression at the high point of a conflict between the medical profession and alternative practitioners.

In the tradition of King Asa, Christian Scientists and the Peculiar People in England and others have at various times appeared in courts of law for refusing medical attention to the children. One of the more publicized conflicts between secular medicine and religious principle relates to the belief of Jehovah's Witnesses that blood transfusions are unscriptural and dangerous; in the United States, Canada, and Australia medical decisions to transfuse children without parental consent—for example, by making children wards of court—have been more frequent than in the United Kingdom.

The rejection of secular medicine is, for some groups, axiomatic in the rejection of the world and its wisdom. For others, medicine does not need to be rejected with the world: the boundaries that exclude electricity and the motor car from Amish territory do not resist conventional medicine.

The conflict of religious and secular may be overcome by a strategy of incorporation: the religious organization may reckon to control not only its own ministrations but also those which are conducted outside it. So Christians in the church tradition may pray for those in hospital 'that by thy blessing upon them and upon those who minister to them they may be restored ...' while those in the Pentecostal church may ask, 'We pray thee, use that surgeon Lord to thy great glory.' And again, 'We prayed about the medicine and it began to work.'

5. *Conclusion*

The process by which religious organizations lose control of aspects of human life is familiar to students of religion as secularization. In recent years conventional medicine has suffered a similar attrition and the revival of religious healing has traded upon it.

There have emerged in the western world two movements that function as alternatives to conventional medicine and compare themselves with it. The religious tradition typified by but not exclusive to the practice of the Pentecostal churches has compelling parallels with shamanism from which it does not directly draw. The second tradition is that of complementary or alternative medicine which has burgeoned since the 1960s. The alternative is distinguished by the claim of its therapists to understand and control the process of cure while the religious involves the engagement of rituals in the expectation of a miraculous effect. There are several features common to both movements, including a suspicion if not rejection of conventional medicine and the role of physical contact between the healer and the healed, be it in the form of massage or manipulation or of the ritual laying on of hands. However, there are significant differences between the two movements. Complementary medicine has highly organized professional structures and the authority to practice comes from formal study and qualification. Clients of alternative medicine seek practitioners because of their methods, not for their charisma or personal reputation.

Such a typification highlights the peculiar case of the Church's Ministry of Healing which conforms much more to the secular alternative model than to the religious: its practitioners cultivate skills that have been developed in secular practice: it draws on the

Gospel for the spirit of its ministry but not for the letter of its technique.

See also: Religion and Health

Bibliography

British Medical Association 1986 *Alternative Therapy*. BMA, London
Chesnut R A 1997 *Born Again in Brazil*. Rutgers University Press, New Brunswick, NJ
Doore G (ed.) 1988 *Shaman's Path: Healing, Personal Growth and Empowerment*. Shambhala, Boston
Eliade M 1964 *Shamanism: Archaic Techniques of Ecstasy* (trans. Trask W R, Bollingen Series no. 76). Bollingen Foundation, New York
Maddocks M 1981 *The Christian Healing Ministry*. SPCK, London
Maguire M B 1988 *Ritual Healing in Suburban America*. Rutgers University Press, New Brunswick, NJ
Martin D 1990 *Tongues of Fire*. Blackwell, Oxford, UK
Pullar P 1988 *Spiritual and Lay Healing*. Penguin, London
Trotter R T II, Chavira J A 1997 *Curanderismo: Mexican American Folk Healing*, 2nd edn. University of Georgia Press, Athens, GA

R. Homan

Health and Illness: Mental Representations in Different Cultures

Mental representations of health and illness are the cognitive structures that guide people's response to symptoms and illness as well as to various kinds of health threats. These mental representations have variously been described as 'cognitive schemata,' 'mental prototypes,' 'structured vocabularies,' and 'social representations.' However one describes them, they provide the foundation for responding to states of health and illness. These mental representations also vary from culture to culture and provide a basis for the wide variety of ways in which people in different cultures respond to health and illness related events. This article explores the nature of these mental representations, the ways in which they vary from culture to culture, and some of the implications of this variation.

1. Understanding Mental Representations of Health and Illness

Although the biological nature of disease remains constant from one culture to another, the evidence is very clear that how diseases, and health and illness in general, are understood varies dramatically. This variation occurs from culture to culture as well as between individuals and subgroups within the same culture. As such, illness, which refers to the experiential component of disease and ill health, is best understood as a social and cultural construction. Whereas disease refers to objective pathology, illness, as experience, is constructed on the basis of shared understandings in addition to individual experience (Angel and Thoits 1987, Landrine and Klonoff 1992).

This construction of health and illness has been described in various ways. At one level it can be described in terms of structured vocabularies used by people to define and express feelings of distress. The basic idea is that culture and experience provide people with categories for defining experience as well as a language for expressing feelings of distress. For example, the common concept of influenza provides many people with a category for understanding the experience of fever, chills, headache, nasal congestion, sore throat, and general achiness, whereas the concept of *susto pasado* is a Mexican-American category used for understanding the experience of stomach trouble, diarrhea, lack of appetite, lethargy, irritability, and loss of weight after experiencing a strong fright.

A useful approach to understanding the shared nature of health and illness concepts is via the theory of social representations. The key idea in this approach is that in the process of attempting to understand their bodily experience, people create socially shared representations of health and illness that then become part of the heritage of the society as a whole or of a subgroup within the society. These representations influence the way that individuals think about health and illness but are a property of the larger social collective. This approach emphasizes the essential social nature of concepts of health and illness and the way in which they are shared between individuals, a key prerequisite for communicating private bodily experience to others. A good example of this approach is Herzlich's (1973) examination of illness beliefs in France.

A contrasting but complementary perspective is the information processing approach that examines the cognitions which individuals have about health and illness. This approach has its roots in cognitive psychology and, in particular, work on cognitive schemata and prototypes. Research using this approach has shown that people have implicit theories of disease that exert a strong influence on how they cope with health problems as well as how they deal with treatment. For example, Leventhal and Diefenbach (1991) have shown that the mental representations held by hypertension and cancer patients strongly influence their behavioral response to treatment. Research using this approach has also pointed out that such mental representations operate at different levels (sensory, affective, and conceptual) and have a number

of components. These component include the identity of the condition (the label it is given as well as the symptoms involved), its cause, time line (expected length and course), consequences, cure, and perceptions of the type of person who gets the condition. Taken together these components make up the prototype that a person has of a particular disease. For example, a common prototype of 'heart disease' among British young people includes symptoms of lack of stamina, breathlessness, and chest pains as well as the beliefs that it is caused by lifestyle and high blood pressure, is serious but can be cured by surgery, and mainly affects older people (Lalljee et al. 1993). These prototypes, in turn, affect the recognition of disease conditions (see *Symptom Awareness and Interpretation*) as well as the person's illness behavior (see *Illness Behavior and Care Seeking*). Research from the information processing perspective has also explored the cognitive organization of disease concepts, noting that these concepts are organized cognitively in characteristic ways that have implications for how people interpret illness threats and how they respond to people with particular diseases (Bishop 1991).

2. *Cultural Differences in Mental Representations*

One of the more striking features of health and illness concepts is the way in which they vary from culture to culture. This variation is obvious even to the casual observer, as concepts common in one culture or society may not even exist in another. Although there are any number of ways in which health and illness concepts may vary, three major ways stand out. First, cultures show substantial variation in conceptions of the basic nature of health and illness. For example, in Western societies placing a strong emphasis on the scientific understanding of health and illness, these concepts are understood in natural terms with disease considered as pathology deriving from natural causes, such as germs, injury, or genetic aberrations. By contrast, in other societies, as well as among some ethnic-cultural minorities within Western societies, illness is understood more in interpersonal as well as supernatural terms. For example, illness may be seen as resulting from violations of standards of interpersonal behavior or other social norms, particular emotions (such as jealousy or envy), transgressions of moral and/or religious taboos, or such quasinatural agents as hot–cold foods, drinks, weather, 'balance,' or particular characteristics of one's blood (e.g., thin, weak, or bad) (Landrine and Klonoff 1992). These differences in the general nature of health and illness are then reflected in the beliefs people hold about specific conditions.

Second, cultures often differ in their disease vocabularies. In some cases the differences may result from using different terminology for what appears to be the same condition. For example, Kleinman (1980) notes that symptoms that are often diagnosed as depression in the West are interpreted as neurasthenia in China. In other cases, however, a condition defined in one culture may have no referent in another. For example, conditions such as *susto*, *empacho*, *caida de mollero*, and *mal de ojo*, experienced by some Mexican-Americans, have no direct referent among conditions known to Anglo–Americans or to medical professionals.

Third, the underlying structure of illness representations often differs between cultures. Work using techniques such as multidimensional scaling and factor analysis has shown that the mental representations people have of different diseases show characteristic organization in memory and that this organization differs between different cultures. For example, research with Anglo-Americans has found that disease representations appear to be organized along with dimensions of contagiousness and severity, whereas those of Mexican-Americans were organized according to concepts of 'hot' vs. 'cold' and whether a disease was associated with children or old people (D'andrade et al. 1972). As one might expect, the organization of disease concepts is closely related to the general models of health and illness that people are exposed to. Thus among Anglo-Americans whose folk model of health derives primarily from the biomedical model, contagiousness, and severity, central concepts in the biomedical model, have been found to be prime organizing dimensions. However, among Chinese in Singapore where Chinese traditional medicine exists side-by-side with Western scientific medicine, disease concepts are organized according to dimensions deriving from both Chinese and Western medicine (Bishop 1998).

3. *Implications*

The mental representations that people have of health and illness, and cultural differences in these representations, have a number of implications for health and illness related behavior. First, the disease concepts that people have can be expected to have a strong influence on their interpretation of symptoms and bodily experience in general. In brief, having specific disease concepts make its possible, and perhaps likely, that a person will interpret bodily experience in line with that concept. Thus, having available the concept of *susto* makes it possible for Mexican-Americans to experience this condition, an experience that is not available to others not possessing this concept. Along these same lines, the relative currency of the concept of neurasthenia in China and of the concept of depression in North America increases the likelihood that the similar symptoms associated with these two conditions

will be interpreted, and perhaps experienced, differently in these two cultures.

Second, and following from this, health and illness representations have implications for what people do when they experience distress. For one thing, these representations provide a definition for when bodily distress represents illness and requires assistance from a medical professional or other healer. For example, research with heart attack patients points out that one reason for delay in seeking treatment is that the person misinterpreted the symptoms and failed to recognize the seriousness of the situation. In addition, disease representations serve to direct the person's help-seeking behavior. For example, research in Singapore has noted that degree of adherence to beliefs associated with Chinese traditional medicine is a key determinant of whether the person seeks help from a practitioner of Western or Chinese traditional medicine (Bishop 1998).

Third, once help is sought, disease representations play an important role in determining whether the person stays in treatment and follows the treatment advice. A good example of this is research with hypertension patients showing that the cognitive 'model' they had of hypertension and its relationship to specific symptoms was a major determinant of whether they stayed in treatment (Leventhal and Diefenbach 1991).

Finally, illness representations play a key role in how people respond to potential health threats. Specifically, the beliefs that people have about particular diseases, how they are contracted, and their consequences have a strong influence on how they respond to people having the condition as well as what precautions, if any, they take to prevent the disease. For instance, beliefs among some Africans that anyone who has ever had sex with a person who has died of AIDS will automatically die as well appear to lead to refusal on the part of those holding such beliefs to seek medical advice, be tested for HIV, or change their sexual behavior so as to reduce their risk of infection (Ankrah 1991).

4. *Remaining Questions*

In many respects work to date has only scratched the surface of this fascinating topic. As noted, progress has been made in the theoretical understanding of mental representations of health and illness but much more can be done to flesh out this understanding and, in particular, explicitly account for cultural differences. Also much remains to be done in further exploring the content of health and illness concepts in different cultures and relating these to other aspects of the cultures in question as well as to dimensions, such as individualism-collectivism and power distance, used to describe cultures.

See also: African Studies: Health; Cultural Psychology; Culture as a Determinant of Mental Health; Illness Behavior and Care Seeking; Latin American Studies: Health; Psychosomatic Medicine; South Asian Studies: Health; Southeast Asian Studies: Health; Western European Studies: Health

Bibliography

Angel R, Thoits P 1987 The impact of culture on the cognitive structure of illness. *Culture Medicine and Psychiatry* **11**: 465–94

Ankrah E M 1991 AIDS and the social side of health. *Social Science Medicine* **32**: 967–80

Bishop G D 1991 Understanding the understanding of illness: Lay disease representations. In: Skelton J A, Croyle R T (eds.) *Mental Representation in Health and Illness*. Springer-Verlag, New York

Bishop G D 1998 East meets West: Illness cognition and behaviour in Singapore. *Applied Psychology* **47**: 519–34

D'andrade R G, Quinn N R, Nerlove S B, Romney A K 1972 Categories of disease in American–English and Mexican–American Spanish. In: Shepard R N, Romney A K, Nerlove S B (eds.) *Multidimensional Scaling*. Seminar Press, New York

Herzlich C 1973 *Health and Illness: A Social Psychological Analysis*. Academic Press, London

Kleinman A 1980 *Patients and Healers in the Context of Culture*. University of California Press, Berkeley, CA

Lalljee M, Lamb R, Carnibella G 1993 Lay prototypes of illness: Their content and use. *Psychology and Health* **8**: 33–49

Landrine H, Klonoff E A 1992 Culture and health-related schemas: A review and proposal for interdisciplinary integration. *Health Psychology* **11**: 267–76

Leventhal H, Diefenbach M 1991 The active side of illness cognition. In: Skelton J A, Croyle R T (eds.) *Mental Representation in Health and Illness*. Springer-Verlag, New York

G. D. Bishop

Health and Mortality, Body Stature as a Measure of

In recent years a rapidly expanding research on the determinants and consequences of human stature has become central to the study of health and mortality processes. Human stature—as measured by height, weight, body mass (which combines height and weight), and a host of other anthropometric measures—has been shown to reflect health status and mortality risk across the life course. Stature is an indicator of health *in utero*, in infancy, and in childhood, and is an important predictor of chronic disease and mortality among adult populations. Moreover, anthropometric measures can be used to com-

pare health and nutritional status across populations, social classes, and historical time periods.

1. Stature as an Indicator of Nutrition, Health, and Mortality

1.1 The Importance of Height

Although auxology (the study of human growth) began as early as the seventeeth century, it was not until the nineteenth century that scholars began realizing that environment can systematically affect human growth. For instance, Villerme argued, based on a study of French soldiers, that poverty was much more important than climate in influencing growth. A few decades later, Charles Roberts estimated fitness for factory employment using frequency distributions of key anthropometric measures such as weight-for-height and chest circumference. In the early 1890s, Franz Boas conducted a national growth study, which was used to develop national standards in the USA. In later years, Boas was a pioneer in the use of statistical methods to analyze growth and furthered the work of linking environment and heredity to growth (Steckel 1995).

In the twentieth century, it became clear that a child's growth rate is likely the best single indicator of his or her health and nutritional status. Furthermore, mean values of childhood heights and weights provide powerful summary statistics of the public health of the population, and childhood anthropometric measures can be used to make meaningful comparisons of health and nutritional status across populations (Eveleth and Tanner 1990).

Adult height can also be used as a measure of public health, since adult height is a proxy for the childhood nutritional status of a population. However, adult height reflects a complex interaction of genetics, environment and the timing of nutritional shortfalls. If nutritional deprivation occurs during the 'sensitive periods' of development, differences in adult height may be quite striking. The sensitivity of growth to nutritional deprivation depends upon the age at which it occurs. Catching up is possible, but prolonged and severe deprivation results in stunting or a reduction in adult size.

Height is a significant predictor of older age mortality. Overall, mortality risk declines as height increases. One of the most comprehensive modern studies was completed by Waaler (1984) on the entire Norwegian population. He found that men measuring 165 cm in height face a mortality risk that is 71 percent greater than those measuring 182.5 cm. Obviously height is also strongly influenced by genetic factors, but genetics cannot explain the significant differentials in height that exist today according to socioeconomic status, nor the narrowing of differentials in industrialized nations (Floud et al. 1990).

1.2 Weight and BMI

In both children and adults, weight is an important anthropometrical measure. Whereas height is typically used as a proxy for past nutritional development, weight is seen as a measure of current health and nutritional status. In the developing world, low weight remains a significant indicator of nutrition and health. In the developed nations, high weights are the principal concern. Obesity rates, especially in the USA, have risen over recent decades and obesity is frequently found to be a risk factor for serious chronic illnesses, including cardiovascular disease, a leading cause of adult mortality in developed nations. Obesity has also shown to be correlated with race and social class.

Obviously weight can be more effectively exploited in epidemiological analysis if it is standardized for height. A common standardization is to use body mass index (BMI), which is defined as height (in meters) divided by the square of weight (in kilograms). The mortality risk is relatively constant in the range of 21–28. However, as BMI falls below 21, adult mortality rises sharply. Mortality also rises for BMI higher than 28. BMI is associated with higher mortality from all causes, though the relative importance of BMI may be confounded by a history of smoking and by previous diseases, and the impact of BMI declines with age. Some studies suggest that much of the increased mortality risk from low BMIs in the USA can be attributed to smoking. Interestingly, excess mortality among African-American women in the USA is not attributed to BMI, even though they have high rates of obesity.

2. Determinants of Stature

Over the last century, scholars have made considerable strides in characterizing and understanding human growth and how various genetic, nutritional, and environmental factors interact with the growth process. Clearly the most important time is infancy, in which the growth rate is greater than at any point thereafter. Not coincidentally, infancy is also the time when humans face the greatest risk of mortality. After infancy, the growth rate declines irregularly until adolescence, during which it rises, at its peak, to about one-half the rate during infancy. The most important study on human growth is by Eveleth and Tanner (1990), and the following sections draw heavily from this work.

2.1 Genetics

Within a community facing similar environmental influences, the distribution of heights is dominated by genetic influences. Systematic studies of heights within families confirm this everyday observation. Studies of twins and other sibling relationships provide the most

convincing evidence. Wilson (1976) reports that monozygotic (MZ) twins, who share the same genes and, in most cases, a common family environment, have, on average, only 1.1 cm difference in height at age four, while dizygotic (DZ) or fraternal twins have a 3.2 cm difference at the same age. Furthermore, differences in height in MZ twins decline as the twins get older, but the difference increases for DZ pairs.

Numerous studies of twins and other siblings provide strong evidence that the closer the genetic relationship, the closer the growth patterns. MZ twins are more closely intercorrelated than DZ twins, but siblings in general are much more similar to each other than they are to nonrelated individuals. Furthermore, the heights of parents and children are highly correlated, though no evidence exists that one parent is more important than the other.

2.2 Nutrition

Even though genetic influences are an important determinant of *individual* variation in height, differences in average height across *populations* are almost entirely the product of environmental influences, not genetics. Although many environmental factors affect body stature, chief among them is nutrition. Of particular importance is nutrition in early life. Numerous scientific studies have shown that severely malnourished children have a greatly increased mortality risk relative to normal children (Behrman and Deolalikar 1988).

Infancy is the period of life in which both the human growth rate and the risk of mortality is at its highest. It is not surprising, therefore, that it is also the time during which nutritional intake is the most critical. Indeed, the concern for nutrition may begin *in utero*, and several studies have shown that supplementing the diet of pregnant women in developing nations significantly increases birth weights—the most important predictor of infant mortality. Even in developing nations, infant mortality is higher than any other period of life. By the age of two, much of the deficit characteristic of adults has already accumulated. Nutrition during the adolescent growth spurt may be important, though the evidence is mixed. In any case, the impact on morbidity and mortality of adolescent malnutrition is nowhere near as dramatic as it is in early childhood.

Many sub-populations in developing nations have mean birth weights that are similar to those in developed nations. In these developing nations, however, weight gain begins to diminish at about six months, the time when mothers' lactating ability declines and children are often weaned. Although food supplements can help, the prevention and treatment of infectious diseases is also critical in this period in developing nations.

2.3 Infectious Disease

Infectious disease interacts closely with nutrition in the growth process. Even in nutrition-rich environments, the undeveloped immune systems of children make them more susceptible to disease. When diet is poor, most of the body's resources are devoted to basic maintenance; malnourished children, therefore, are even more susceptible to disease. Furthermore, infection reduces the body's ability to absorb nutrients, which can make children even more malnourished. Infectious disease in childhood, therefore, can have long-term consequences on growth. Some controversy does exist on this topic, however. An alternative 'small but healthy' hypothesis has been promoted by a few researchers, though it has not received widespread support.

2.4 Socioeconomic Correlates

Economic growth (which results in better nutrition, public sanitation, improved hygiene, improved medical treatment, etc.) has been the driving force behind the secular increases in height around the globe. Variation in national income can also explain much of the variation in mean heights across nations. Steckel (1995) calculates correlation coefficients between mean height and the log of per capita income that are between 0.8 and 0.9. National income remains a powerful correlate of height, even after controlling for other demographic characteristics of the nations such as urbanization and age. And in nearly all nations, both developed and developing, children with moderate to high socioeconomic status are larger than their counterparts with low socioeconomic status, where socioeconomic status is determined by education, occupation, and income.

Of course the effect of socioeconomic status is interwoven with nutritional intake and infectious disease environments. Consequently, socioeconomic differentials are much more pronounced in developing nations. Socioeconomic differentials in height are most pronounced in childhood, but exist for adults as well. Several studies show that the social class effects are greater for boys than for girls.

3. Stature and the Study of Health and Mortality

3.1 Historical Populations

In the last decades of the twentieth century scholars began using body stature, and height in particular, to determine the living standards and health status of populations for several historical time periods and in various locations around the globe. Health and mortality details, particularly at the individual level, are generally unavailable in historical data, but height data are available (typically from military service records) in several American and European popu-

lations since the seventeenth century. Height, therefore, is the most important indicator of health status we have when investigating the historical record.

These studies have confirmed traditional notions of public health, such as the abysmal living conditions in eighteenth- and nineteenth-century urban areas. There have also been some unexpected findings. For example, the USA, in contrast to European nations, achieved near-modern average height as early as the mid-eighteenth century, though heights declined somewhat over the nineteenth century (from 173 cm in 1800 to 171 in 1900 for adult men), mostly during the middle part of the century. Mortality evidence from genealogies and from plantation records indicates that life expectancy deteriorated while heights declined during the ante-bellum period. The cause of the nineteenth-century decline in height is still a puzzle. Hypotheses include increased urbanization, nutritional decline, an increase in infectious diseases created by trade, migration, and public schooling, and the disruption of the US Civil War.

Before the twentieth century, European men were smaller than Americans, probably due to far poorer diets, but there was significant diversity in height across European nations. In the early nineteenth century, residents of Ireland and Scotland were the tallest (168 cm) followed by Norway, Sweden, England, France and Austria-Hungary (163 cm). In the twentieth century, geographic distributions of height shifted. Based on data collated by Eveleth and Tanner (1990), the Dutch and Norwegians were now the tallest, followed by Americans, British, French and Austrians. Socioeconomic differentials were pronounced in Europe. Floud et al. (1990) report, for instance, that in nineteenth-century England, 14-year-old Sandhurst boys (mostly the children of aristocrats, high-ranking officers, and professionals) exceeded the stature of boys from the Marine Society (mostly from the slums of London) by 10 to 15 cm.

3.2 The Modern World

The period from the late-nineteenth to the mid-twentieth century was characterized by dramatic increases in stature and equally dramatic reductions in mortality. Life expectancy was about 40 years in 1860 but had risen to 68 years by 1950 (see *Mortality: The Great Historical Decline*). There is no single cause for this dramatic increase, but several contributing factors have been identified. Widespread public health programs were implemented, such as vaccinations, purifications of water supplies, sewage disposal, and milk pasteurization. Medical advances, such as an understanding of the germ theory of disease and the development of antibiotics, as well as increasing access to medical care, also contributed. Increases in income also facilitated better diets and housing.

Stature remains an important variable in modern studies of health and mortality, though its use varies by the economic development of the population under study. In developing nations, stunting and wasting are still critical indicators of nutritional status and the overall health of the population. In developed nations, most of the focus is on increased mortality of the obese, as indicated by high levels of BMI. Significant differentials in all health indicators, including anthropometric measures, still exist across cultures and social classes.

Central to the improvements in the twentieth century was the fight against infectious diseases. In a wide-ranging study of numerous populations from 1900 to the early 1960s, Preston (1976) attributed 25 percent of the mortality decline to influenza, pneumonia, and bronchitis; 15 percent to 'other infectious and parasitic diseases'; 10 percent to respiratory tuberculosis; and 10 percent to diarrheal diseases (the remaining portion of the decline is most likely attributable to cardiovascular disease). As noted earlier, the impact of infectious disease, particularly among children, is impacted significantly by nutrition.

Many factors, such as the development of antibiotics, contributed in the twentieth century to the decline in infectious disease prevalence and in the associated reduction in mortality. During the period between 1950 and 1970, life expectancy in developed nations increased only slightly (see *Life Expectancy and Adult Mortality in Industrialized Countries*), which many attributed to the successful fight against infectious diseases. Some argued that because so much early mortality was due to infectious disease, modern populations were facing an era of increased life but worsening health. People with frailer constitutions were living to older ages and, it was thought, would face higher rates of chronic illness than previous cohorts of elderly (Verbrugge 1984). However, most recent evidence suggests that age-specific morbidity and disability rates have been declining in the elderly population. Also surprising to most demographers has been the sharp increase in life expectancy after the period of stagnation between 1950 and 1970. The sources behind this increase remain largely a puzzle, and little consensus exists on future projections of life expectancy.

Towards the end of the twentieth century, birth weight, as well as other infant anthropometric measures, was linked not only to infant and child mortality, but also to disease and mortality in later life. The work of Barker (1994) suggests that development *in utero* may be critical to long-term health. This and other research suggests that anthropometric measures will remain an important component in an understanding of the complex and changing relationship between nutrition, health and mortality.

See also: Genetics and Development; Longevity, Genetics of; Mortality, Biodemography of; Mortality Differentials: Selection and Causation; Mortality, Epidemiological, and Health Transitions

Bibliography

Barker D J P 1994 *Mothers, Babies and Disease in Later Life*. BMT Publishers, London

Behrman J R, Deolalikar A B 1988 Health and nutrition. In: Chenery H, Srinivasan T N (eds.) *Handbook of Developmental Economics*. North Holland, Amsterdam

Eveleth P B, Tanner J M 1990 *Worldwide Variation in Human Growth*, 2nd edn. Cambridge University Press, Cambridge, UK

Floud R, Wachter K, Gregory A 1990 *Height, Health and History: Nutritional Status in the United Kingdom*, 1750–1890. Cambridge University Press, Cambridge, UK

Fogel R W 1986 Nutrition and the decline in mortality since 1700: Some preliminary findings. In: Engerman S L, Gallman R E (eds.) *Long-term Factors in American Economic Growth*. University of Chicago Press, Chicago

Preston S H 1976 *Mortality Patterns in National Populations*. Academic Press, New York

Rogers R G, Hummer R A, Nam C B 2000 *Living and Dying in the USA: Behavior, Health and Social Differentials of Adult Mortality*. Academic Press, San Diego, CA

Steckel R H 1995 Stature and the standard of living. *Journal of Economic Literature* **33**: 1903–40

Verbrugge L M 1984 Longer life but worsening health?: Trends in health and mortality of middle-aged and older persons. *Milbank Quarterly* **62**: 475–519

Wilson R S 1976 Concordance in human growth for monozygotic and dizygotic twins. *Annals of Human Biology* **3**: 1–10

Waaler H T 1984 Height, weight and mortality: The Norwegian experience. *Acta Medica Scandinavica* Suppl. **679**: 1–51

S. E. Wilson

Health: Anthropological Aspects

Health and health practices are part of the inmost complexities of social existence, permeating the domains of politics, economics, and religion and almost always connected with dimensions that go beyond the body, such as interpersonal, family, and community relationships. Health is at once an end and a means. Explanations of the causes and patterns of health and disease often convey value judgments, senses of right and wrong, and of accountability and blame, as well as revealing what is morally at stake in definitions of health and its failures. Medical anthropological studies have built upon historical and cultural analyses over a long period of time that indicate an array of difference with respect to the metaphors and meanings that signify health. These studies have emphasized cultural fabrics that give coherence and depth to these meanings such that cosmology and ethical traditions come to define the body in states of sickness and well-being.

Anthropological studies of health focus on local contexts where health and illness are recognized and responded to. Such studies also trace the effects of global flows of commodities, information, finance, images, people, and pathogens on such worlds. Anthropologists have been as interested in the social roots and consequences of health (and illness) as in their cultural representations. But what most particularly characterizes the anthropological perspective is the use of ethnography to understand health, illness, and health care. Anthropological perspectives on health bring together individual and collective realities in the way they are organized, narrated, contested, and in every sense lived as social trajectories (Kleinman 1995).

What anthropologists have learned through research in different societies is a story of diversity in experiences of health and health care giving, as well as of the ways social relationships and emotional processes interfuse in the mediation of health-relevant experiences. The meanings of health range from holistic and organic interpretations to hyperindividualistic ones. While biomedicine has treated health as a separate domain, popular meanings and interpretations of healthy states are often inseparable from other moral, political, and economic domains. Anthropologists have been particularly concerned with the transformations in health as those transformations express differences in power, social position, and social inequality, particularly as experienced by marginal groups and individuals. But in recent years, equal interest has been devoted to studies of health professionals, bureaucrats, and scientists who contribute to the public definition of health and health problems. Culture affects health as much through the culture of biomedicine and bureaucratic institutions as through the popular culture; indeed in the age of the Internet the two are closely related.

Applied medical anthropology addresses issues such as the ones raised and addresses them at the policy level. In heterogeneous societies such as North America and Western Europe, cultural differences continue to matter in the health field in large part because of differences in health conceptions which affect the panoply of health-relevant practices, from popular assessments of risks and responses to them to the uses and misuses of health care resources. Those differences contribute to different health outcomes for people, but they also contribute to different health systems, which are themselves undergoing transformation owing to both local and global economic and social changes.

1. Images of Health and Diverse Experiences of Health

The contemporary world contains impressive variations of images of health, modes of health maintenance, efforts at prevention, and illness experiences. Much of this variation is subsumed under what demographers and epidemiologists have termed the

'health transition.' This phrase refers to an increase in life expectancy, decline in mortality, and overall decrease in infectious diseases in the twentieth century by means of interventions such as vaccinations and dietary improvements. Health status measures have been introduced and are continually revised to develop policy, planning, and evaluation; these measures shape ideas of health and health improvement according to specific indices that promote interventions and sociobehavioral change. The health transition has also been accompanied by major changes in patterns of morbidity, increases in chronic degenerative diseases, old and new epidemics and behavioral and mental pathologies and is thus more complex than its original conceptualization (Chen et al. 1994). The very economic, technical, and industrial development that has benefited so many coexists with ongoing massive societal disruptions. Developing countries, for example, are experiencing the social and health costs of rapid urbanization and economic restructuring, costs that are confirmed in statistical increases of preventable disease and illness, and borne in terms of health problems such as infant mortality, life-threatening dehydration, urban violence, and hunger (Scheper-Hughes 1992).

Such acts as war, political persecution, and homicide contribute to a deliberate production of suffering and death. The era of infectious diseases is by no means past as the spread of HIV/AIDS, multidrug resistant tuberculosis, and other infectious diseases clarify. With the collapse of authoritarian rule came the collapse of public health measures and treatment regimes, demographic changes, travel, and other human actions that intensify the flows and resistance of pathogens. It is through such mechanisms that Paul Farmer has shown the ways social forces alter disease distribution and contribute to the persistence of new microbial conditions; indeed, the way 'social forces and processes come to be embodied as biological events' (1999, p. 5). The appearance of these disorders, closely correlated with poverty and social inequality, challenges the notion of a single biological etiology for disease and illustrates that these infectious diseases are mediated by multiple interacting causes that require anthropological inquiry and policy intervention.

The burden of chronic diseases—heart disease, cancer, stroke, diabetes, asthma—continue to be distributed unevenly across minority groups. Yet statistics also disclose impressive variations. In the United States, for example, there is the so-called Latino Epidemiological Paradox: Latinos, who generally experience a low socioeconomic status, nonetheless, have health outcomes that are either equal or better than non-Hispanic Whites. African Americans, by comparison, experience significantly higher rates of infant mortality, death, diabetes, HIV/AIDS, heart attacks, strokes, and homicides. Differences in levels of institutionalized racism, dietary patterns, and the influence of the Spanish language as a protective cultural barrier have been pointed to as possible explanations for this paradox. Notably, health indices of Latino immigrants worsen from generation to generation in the United States, owing perhaps to a loss of balanced diet or tobacco or substance abuse as well as limited accessibility of health care services (Hernandez and Charney 1998).

The latest cluster of disorders to receive attention in the health transition is the rising rates of substance abuse, violence, psychological trauma, depression, suicide, and related mental health problems (Desjarlais et al. 1995). This group of health problems has increased to epidemic proportions both in developed and developing societies. Anthropological studies indicate that these problems often cluster together in settings of deep poverty, and societal breakdown. They are compounded by the fact that services tend to be limited, and health care financing and resources are inadequate. Many psychiatric hospitals in China, Russia, and Latin America continue to rely on a model of care involving prolonged stays in locked wards with antipsychotic medication and physical treatment (Phillips 1998).

Anthropological interventions have gone beyond explication of the 'soft' variables underpinning mental health problems, and focus on the ways social, economic, and political contexts shape the formation, distribution, and manifestation of disorders. Thus in the anthropological perspective, health problems and social problems are inseparable; health policies and social policies are viewed as necessary complements in defining criteria and techniques of successful rehabilitation.

2. *Local Variations in Illness and Health*

In response to emerging sets of health issues, anthropologists have identified the need for cross-cultural studies of health, healing, and patterns of morbidity in local contexts. One thread linking these studies is the importance of understanding patterns of everyday life experience, local knowledge, and social networks that influence personal agency and access to health care, as well as modes of transmission of diseases. From this perspective, health is seen as not simply originating in individuals themselves or deriving strictly from social forces, but as being engendered in complex interactions between individuals and their social contexts. Such lived responses may occur in the contexts of medical pluralism where several therapeutic systems coexist in one cultural setting, or in circumstances in which symptoms are subject to several distinct classifications, diagnoses, and hence, therapeutic courses of action or, once again, where different moral and religious values shape health promoting and disease preventing activities related to diet and lifestyle practices and support (Leslie and Young 1992). By stressing contingencies and detailing local histories of

particular places and peoples undergoing many kinds of social change, these ethnographies highlight the way particular worlds define health and conceive of health equity, while also illuminating the ways in which risks to disease and illness inter-relate.

It is not surprising that ethnographers have focused on illness whose event structure is relatively defined. Emphasized is the experience of illness as an interpersonal phenomenon as it is narrated through stories of illness. Ethnographies disclose how suffering is transformed through language and cultural meanings into distinctive illness narratives whose shared metaphors, voices, and genres represent distinctive experiences and political processes. This transformation, in turn, shapes the lived experience of illness through embodiment of cultural meanings and social relations into body-self processes. Ethnographers have focused on different mediating processes—ritual, devotional practices, music, poetics—that structure a sociosomatic continuum between moral and emotional processes. This approach relates to the concept of 'local biologies' by which is meant the different means of relating social worlds and biological processes such that those processes become particularized. Examples include Margaret Lock's demonstration of fundamental differences in the experiences of menopausal symptoms and the aging process among women in Japan and North America (1993); Lawrence Cohen's contrast of distinctive interpersonal experiences of dementia in India and the West (1998); and Rayna Rapp's description of how experiences of class in the United States and technologies of the female reproductive life cycle come to alter experiences of pregnancy, infertility, abortion, and delivery (1999).

Ethnographers have also analyzed the trajectory and outcome of illness experiences on the social course of illness. Thus, Benedicte Ingstad and Susan Reynolds Whyte (1995) have shown how disability is as much a product of social definitions and restrictions and opportunities of access as it is a physical or functional limitation or disease, thus turning disability into an experience of potential wholeness and health. Other anthropologists have investigated societal responses to highly stigmatized disorders that respond to feelings of threat with practices aimed at marginalizing sufferers. Still others have studied experiences of chronic conditions like chronic fatigue syndrome, chronic pain syndrome, and fibromyalgia—disorders that are in a limbo of partial or minimal authorization where patients feel discredited or discreditable because of the lack of social legitimacy. Research here shows how patients often resist biomedical responses that are perceived as undermining and delegitimating.

More recently, this area of research has expanded well beyond biomedically defined conditions to include a variety of forms of violence—structural, political, domestic—that represent significant threats to health. Here researchers of violence such as Veena Das (1995) have shown how different societal idioms of those institutions authorized to deal with social problems—legal, welfare, religious, and of course medical—construct distinctive objects of programs and policies that not only differ greatly but that coerce different courses of health and disease and outcomes of these conditions.

Perhaps the largest field of anthropological inquiry has examined the varieties of community healing systems that parallel state institutions. Studies of herbalists, shamans, and other folk practitioners (and practices) illustrate how experience is manipulated, how meaning is negotiated, and how illness comes to serve some societal ends such as social control, cultural remembering, and political resistance. The point being that the collective significance of illnesses goes beyond individual and even familial concerns and plays a role in how societies remember, how social reality is legitimated, and how social worlds change.

The anthropology of health and medicine largely began as an applied field. There continues to be important segments of the field that seek application in public health and biomedical settings and in the policy arena. Such applied work classically demonstrated that health beliefs could either facilitate or become a substantial barrier to implementing policies and progress (Paul 1955). More recently, anthropologists have engaged the nongovernmental, administrative, and political dimensions of international health agencies and the policy development process, indicating how community organization can positively inform health and social programs in arenas such as HIV/AIDS prevention, cancer or infectious disease control, and family planning (Hahn 1999). Hence, anthropology has moved from a stance of critique of application to a stance of generating novel applications combining anthropological knowledge and public health technologies to enhance health.

3. *Culture of Biomedicine*

Distinctions between health and disease or illness are patterned according to norms. Norms of health and disease often emerge through a combination of statistics patterns, public consensus around desirable health outcomes, and social definitions and responses to deviance (Mechanic 1997). The extent to which biomedical norms have achieved a position of preeminence in health care systems and definitions of disease is undeniable. Biomedicine's success in eradicating certain diseases and prolonging life shows that norms, however value-laden and socially constructed they may be, can indeed be health-enhancing (DelVecchio Good 1995). Yet anthropological studies have questioned biomedicine's reductionism, primary value of efficiency, and authority in mediating interpretations of health, illness, and suffering (Good 1994, Kleinman 1995); they point to the fact that physiological measures alone fail to capture the

subjective dimensions of health. The growth of biomedicine, its characteristic quality of redefining social problems as health problems needing biomedical intervention, and its global cultural reach have also been analyzed. That biomedicine has been closely allied with the growth of industrial capitalism has led many ethnographers to examine biomedicine as a conduit of social regulations and discipline influencing all sectors of life, from definitions of disability to norms of occupational health, to definitions of mental illness and fitness (Turner 1995). This process has led to strong critiques regarding the necessity and indeed at times harmfulness of biomedical interventions, particularly with respect to the unpredictable side effects of technical procedures and drugs used to diagnose and treat patients. Concerns for iatrogenesis are reflected in current US debates regarding the consumption of psychiatric drugs by preschool children for example.

Attention to the international political economic context in which biomedicine operates is central to understanding how ideas of health modify, as well as to understanding the modes through which they are integrated as a set of values and practices in local contexts. Cross-cultural studies reveal the extent to which this integration with local experience is contested or resisted by highlighting: (a) the many discrepancies between biomedical knowledge and subjective accounts of illness and health, particularly in non-Western countries; (b) the moral and religious tensions underlying choices of medical systems; and (c) the degree to which distinctions between health and disease are rooted (or uprooted) by global ideologies and beliefs.

4. *Biotechnology and Health*

In response to the large influence that public health has achieved in our times due to the advances of biomedicine and advances in technology, the power of state action in the public domain, and market forces, health has increasingly become a specialized domain. The history of public health has been marked by major advances in population-level health in the nineteenth and early twentieth centuries through sanitary and workplace improvements, housing, and nutrition. Public health has shifted priority from social medicine to individually targeted health promotion. The launching and completion of the largest biology project ever conceived—the mapping and sequencing of the human genome—has already accelerated this process of individualizing health vis-à-vis the genetic profiling of diseases and increased prognostic information. For example, silicon chips embedded with hybrid bits of DNA that can test for various genetic conditions in individuals are currently in development. Such diagnostic tools are refocusing medical practice away from intervention to risk factor analyses and disease prevention.

The diseases individuals might be predisposed to, how they might react to particular medicines, and what health-promoting behaviors they might take to prevent future illness are all encompassed in this revolution (Rabinow 1996, Eisenburg 1999). Genomic information and the multitude of genomic sequences represent novel targets for therapeutic intervention. In short, the 'life sciences industry'—knowledge of the genetic code and how to manipulate it and therapeutic advances—is transforming public health and biological knowledge. The prospect of widespread genetic testing has prompted concerns about how such information is being used to discriminate against those labeled as having genetic risks in terms of restricting or denying insurance coverage, as is already the case with women who are BRCA1 or BRCA2 positive.

These developments also highlight the depth of demand for 'health.' They attest to the power of biomedicine as a rationality in everyday life (in the organization of health care and delivery, legal procedures, policy analysis) that comes up against other forms of knowledge such as religion, and that coexists with older power structures related to family, gender, ethnicity, and class. Health as commodity, to be bought and sold in the marketplace, has led to new freedoms to pursue health but also raise dilemmas and create new risks: who is caretaker of genomic information? Who will have access to genetic therapies? These questions highlight some of the dangers regarding the social uses of genetic knowledge in the genomic and postgenomic era; they make it difficult to say what counts as success in this drive toward health and profits, a drive which has both sacred and secular sources and consequences.

The tradeoff between benefits to individuals and risks to populations of any given new technology must also be carefully assessed. Xenotransplantation, which promises to end the so-called shortage in donor organs by engineering transplantable animal organs, offers one prominent example. While such breakthrough technologies may save the life of a few, they pose dangers to populations in terms of the resistant viral strains and antibodies that might develop as a result of such transplant procedures. Another example of the way scientific advancement produces new social risks is found in the case of the traffic in human organs. Such traffic has been afforded by the availability of cyclosporine, among other things. This availability, combined with demand for human organs, has turned vulnerable populations in Brazil, South Africa, and India, into potential organ suppliers. Inconsistencies across the globe regarding the rights of donors as well as the lack of follow-up care for many of them have led to prolonged suffering or premature deaths, or both. Processes related to the globalization of the life sciences industry demand ethnographic attention to technical advances in terms of their development and diffusion; patterns of accessibility and differences in improved health outcomes between the poor and the

nonpoor; and the new forms of health and vulnerability these advances engender.

Images of health, illness, and healing have become key media, a 'fragile salvation,' through which societies and groups struggle with central values and beliefs, and negotiate social and political progress. The invention and implementation of techniques of health improvement have taken on many forms, including epidemic control, disease eradication, and combinations of environmental and individual behavioral change and medical advance. Yet definitions of health depend on normative and oftentimes conflictual values. Health as an individual goal coexists with new threats to health worldwide. The spread of advertising for cigarettes or the dumping of expired pharmaceuticals in low-income countries are all part of the globalization of disease. Under such conditions, what was once health now becomes disease. Health maintenance and promotion, once so influential a concept, can be undermined owing to new political and economic realities.

Contexts of health are themselves undergoing changes and are influenced by transformations in economy and polity, in science and technology, and in culture. These contexts shape the conditions of health and are in turn affected by it. These processes demonstrate that health as an experience is irreducible to a set of norms of physiological and mental activity, or strictly to a set of cultural differences. Only through concrete understandings of particular worlds of suffering and the way they are shaped by local histories and political economies can we possibly come to terms with the complex human dimensions that protect or undermine health. Health, seen in this way, is a construction as well as a contested way of being in the world.

See also: Community Health; Cultural Variations in Interpersonal Relationships; Culture and Emotion; Development: Sustainable Agriculture; Ecology and Health; Emotions and Health; Environment and Anthropology; Food in Anthropology; Globalization and Health; Healing; Health Care Systems: Comparative; Health Psychology; Medical Geography; Medical Sociology; Postmodernism: Philosophical Aspects

Bibliography

Chen L, Kleinman A, Ware N 1994 *Health and Social Change in International Perspective*. Harvard University Press, Cambridge, MA

Cohen L 1998 *No Aging in India: Alzheimer's, the Bad Family, and Other Modern Things*. University of California Press, Berkeley, CA

Das V 1995 *Critical Events: An Anthropological Perspective on Contemporary India*. Oxford University Press, Delhi, India and New York

DelVecchio Good M-J 1995 *American Medicine: The Quest for Competence*. University of California Press, Berkeley, CA

Desjarlais R, Eisenberg L, Good B, Kleinman A 1995 *World Mental Health: Problems, and Priorities in Low-Income Countries*. Oxford University Press, New York

Eisenburg L 1999 *The Implications of Genetics for Health Professional Education*. Josiah Macy, Jr Foundation, New York

Farmer P 1999 *Infections and Inequalities: The Modern Plagues*. University of California Press, Berkeley, CA

Good B 1994 *Medicine, Rationality, and Experience: An Anthropological Perspective*. Cambridge University Press, Cambridge, UK and New York

Hahn R 1999 *Anthropology in Public Health: Building Differences in Culture and Society*. Oxford University Press, New York

Hernandez D, Charney E 1998 *From Generation to Generation: The Health and Well-Being of Children in Immigrant Families*. National Academy Press, Washington, DC

Ingstad B, Reynolds Whyte S 1995 *Disability and Culture*. University of California Press, Berkeley, CA

Kleinman A 1995 *Writing at the Margin: Discourse Between Anthropology and Medicine*. University of California Press, Berkeley, CA

Leslie C, Young A 1992 *Paths to Asian Medical Knowledge*. University of California Press, Berkeley, CA

Lock M 1993 *Encounters with Aging: Mythologies of Menopause in Japan and North America*. University of California Press, Berkeley, CA

Mechanic D 1997 The social context of health and disease and choices among health interventions. In: Brandt A, Rozin P (eds.) *Morality and Health*. Routledge, New York

Paul B 1955 *Health, Culture, and Community: Case Studies of Public Reactions to Health Programs*. Russell Sage Foundation, New York

Phillips M R 1998 The transformation of China's mental health services. *China Journal* **39**: 1–36

Rabinow P 1996 *Essays on the Anthropology of Reason*. Princeton University Press, Princeton, NJ

Rapp R 1999 *Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America*. Routledge, New York

Scheper-Hughes N 1992 *Death Without Weeping: The Violence of Everyday Life in Brazil*. University of California Press, Berkeley, CA

Turner B 1995 *Medical Power and Social Knowledge*, 2nd edn. Sage, London

A. Kleinman and A. Petryna

Health Behavior: Psychosocial Theories

The term 'health behavior' is used very broadly in this article to refer to any behavior that influences, or is believed to influence, physical health outcomes, either by increasing or decreasing their risk or severity. A number of psychosocial theories (or models—the two terms are used interchangeably in this article; see also *Health Behaviors*) has been developed to predict, explain, and change such health behaviors. These theories can be divided into two main groups which

are commonly referred to as social cognition models and stage models, respectively. These are discussed in turn.

1. Social Cognition Models

The term 'social cognition models' is used to refer to a group of similar theories, each of which specifies a small number of cognitive and affective factors ('beliefs and attitudes') as the proximal determinants of behavior. These models do not deny that behavior is influenced by many other factors (e.g., social structural, cultural, and personality factors), but they assume that the effects of such distal factors are largely or completely mediated by the proximal factors specified by the model. Unlike the distal factors, the proximal factors are assumed to be amenable to change, for example by provision of relevant information. Thus social cognition models can be used as the basis for health behavior interventions.

The five models that have been used most widely by health behavior researchers in recent years are shown in Fig. 1 in the form of causal diagrams. The diagrams depict the core constructs in each model; numerous variations appear in the literature.

1.1 The Health Belief Model

The health belief model (HBM; Becker 1974) was developed in the 1950s by a group of social psychologists working in the field of public health who were seeking to explain why some people do not use health services such as immunization and screening. The model is still in common use. There are four core constructs: the first two refer to a particular disease whereas the second two refer to a possible course of action that may reduce the risk or severity of that disease. Perceived susceptibility (or perceived vulnerability) is the individual's perceived risk of contracting the disease if he or she were to continue with the current course of action. Perceived severity refers to the seriousness of the disease and its consequences as perceived by the individual. Perceived benefits refer to the perceived advantages of the alternative course of action including the extent to which it reduces the risk of the disease or the severity of its consequences. Perceived barriers (or perceived costs) refers to the perceived disadvantages of adopting the recommended action as well as perceived obstacles that may prevent or hinder its successful performance. These factors are commonly assumed to combine additively to influence the likelihood of performing the behavior. Thus, high susceptibility, high severity, high benefits and low barriers are assumed to lead to a high probability of adopting the recommended action. Another factor that is frequently mentioned in connection with the HBM is cues to action (events that trigger behavior), but little empirical work has been conducted on this construct.

There have been two meta-analyses (quantitative reviews) of research using the HBM. Janz and Becker (1984) calculated significance ratios showing how often each HBM construct was statistically significant in the predicted direction across 46 studies. These ratios were 81 percent for susceptibility, 65 percent for severity, 78 percent for benefits, and 89 percent for barriers; the pattern of findings was similar when only the prospective studies were examined. Thus, barriers is the most consistent predictor of behavior and severity is the least consistent. Harrison et al. (1992) used extremely strict inclusion criteria; they included only 16 of the 234 studies they originally identified. Across these 16 studies, the mean correlations between HBM components and behavior were 0.15, 0.08, 0.13, and -0.21 for susceptibility, severity, benefits, and barriers, respectively. While statistically significant, these correlations are small in substantive terms. Harrison et al. found that benefits and barriers had significantly larger effect sizes in prospective compared with retrospective studies, whereas severity had a significantly larger effect size in retrospective studies.

1.2 Protection Motivation Theory

Protection motivation theory (PMT; Rogers 1983) was originally developed to explain how people respond to fear-arousing health threat communications or 'fear appeals.' It can be regarded as an adaptation of the HBM. Protection motivation refers to the motivation to protect oneself against a health threat; it is usually defined operationally as the intention to adopt the recommended action. Of the determinants of intention specified by the model, the four that have received the most empirical attention are vulnerability and severity (equivalent to perceived susceptibility and severity in the HBM), response efficacy (the belief that the recommended action is effective in reducing the threat), and perceived self-efficacy (the belief that one can successfully perform the recommended action; Bandura 1997). Thus, a person will be more motivated to protect himself or herself (i.e., have a stronger intention to adopt the recommended action) to the extent that he or she believes that the threat is likely if the current course of action is continued, that the consequences will be serious if the threat occurs, that the recommended action is effective in reducing the likelihood or the severity of the threat, and that he or she is able to carry out the recommended action.

In many studies using this model (e.g., Wurtele and Maddux 1987), specific PMT variables are experimentally manipulated in a factorial design and their effects on intention (and sometimes behavior) are measured. In fact, PMT is unique among social cognition models with respect to the relatively large

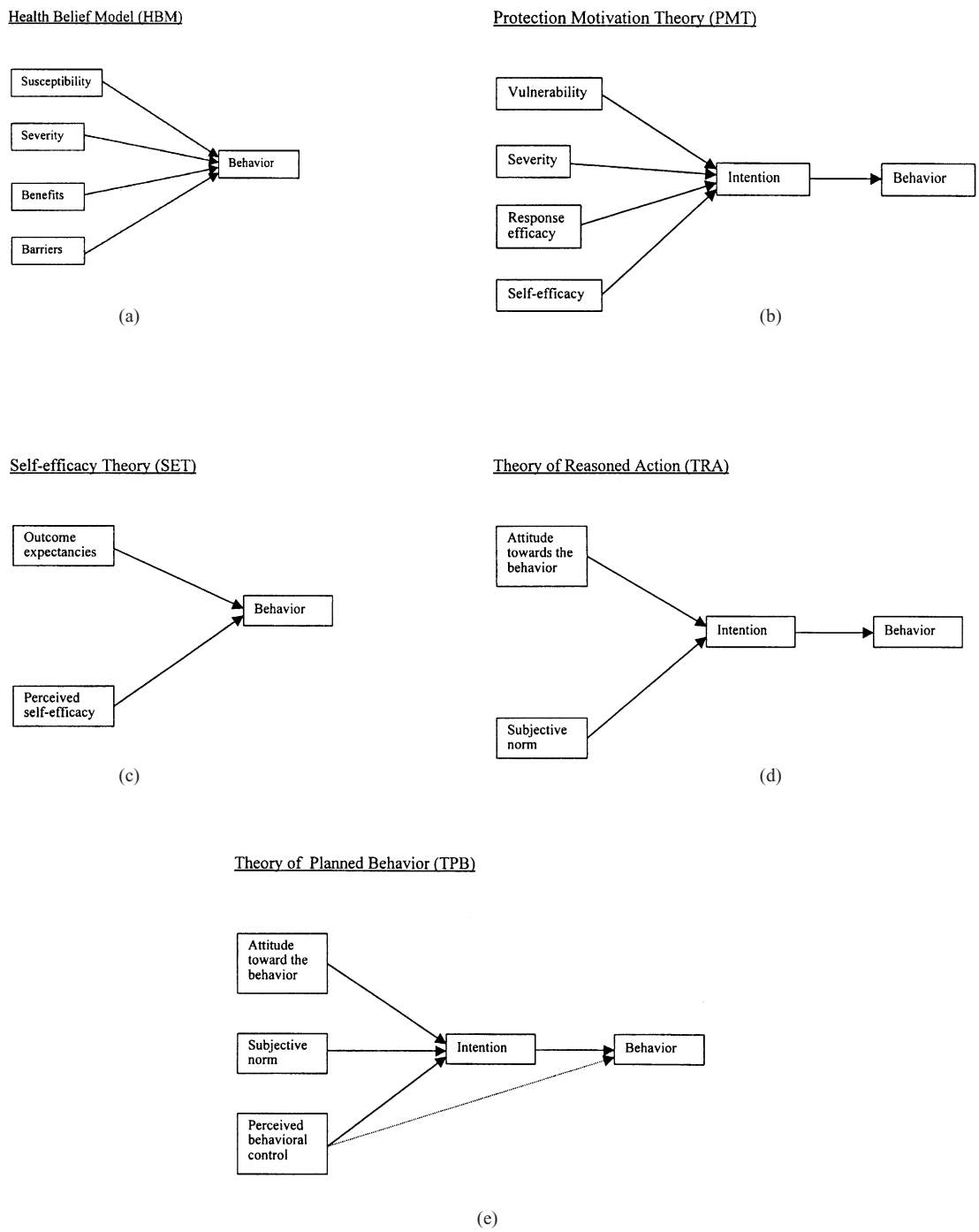

Figure 1
Causal diagrams for five social cognition models. In some cases these are simplified versions of the original models: (a) HBM; (b) PMT; (c) SET; (d) TRA; (e) TPB

number of experimental tests that have been conducted.

To date, two meta-analyses of PMT studies have been conducted (Floyd et al. 2000, Milne et al. 2000). The analyses used different study inclusion criteria and different effect size measures. Floyd et al. analyzed 65 studies with about 30,000 research participants whereas Milne and colleagues included 27 studies with about 8,000 participants. There were only 12 studies in common. Both analyses found support for each of the main PMT variables as predictors of intentions and/or behavior. Self-efficacy had the strongest, most consistent, and most robust effect.

1.3 Self-efficacy Theory

Self-efficacy theory (SET) is a subset of Bandura's (1986) social cognitive theory. According to this approach, the two key determinants of behavior are perceived self-efficacy and outcome expectancies. The latter construct refers to the perceived positive and negative consequences of performing the behavior. See Schwarzer and Fuchs (1996) for a version of this model that incorporates risk perceptions and behavioral intention, as well as components of the action phase of behavior change. No meta-analysis of SET has been published, though there is substantial evidence for the predictive validity of self-efficacy (Schwarzer and Fuchs 1996).

1.4 The Theory of Reasoned Action and the Theory of Planned Behavior

The theory of reasoned action (TRA; Ajzen and Fishbein 1980) developed out of social–psychological research on attitudes and the attitude–behavior relationship. The model assumes that most behaviors of social relevance (including health behaviors) are under volitional control, and that a person's intention to perform a behavior is both the immediate determinant and the single best predictor of that behavior. Intention in turn is held to be a function of two basic determinants: attitude towards the behavior (the person's overall evaluation of performing the behavior) and subjective norm (the perceived expectations of important others with regard to the individual performing the behavior in question). Generally speaking, people will have strong intentions to perform a given action if they evaluate it positively and if they believe that important others think they should perform it. The relative importance of the two factors may vary across behaviors and populations.

Although not shown in Fig. 1, the TRA also specifies the determinants of attitude and subjective norm. Attitude is held to reflect the person's salient behavioral beliefs concerning the possible personal consequences of the action. For example, a person who believes that performing a given behavior will lead to mostly positive personal consequences will hold a favorable attitude towards the behavior. Specifically, attitude is held to be a function of the sum of the person's salient behavioral beliefs concerning the outcome of the action each weighted by their evaluation of that outcome. An indirect, belief-based, measure of attitude can be created by multiplying each behavioral belief by its corresponding outcome evaluation and then summing over outcomes. In a similar way, subjective norm is a function of the person's beliefs that specific individuals or groups think he or she should, or should not, perform the behavior. A person who believes that most significant referents think he or she should perform the behavior will perceive social pressure to do so. Specifically, subjective norm is held to be a function of the person's salient normative beliefs with respect to each referent, each weighted by their motivation to comply with that referent. An indirect measure of subjective norm can be created by multiplying each normative belief by its corresponding motivation to comply and then summing over referents.

Many behaviors cannot simply be performed at will; they require skills, opportunities, resources, or cooperation for their successful execution. The theory of planned behavior (TPB; Ajzen 1991) was an attempt to extend the TRA to include behaviors that are not entirely under volitional control, for example giving up smoking or using a condom. To accommodate such behaviors, Ajzen added a variable called perceived behavioral control to the TRA. This refers to the perceived ease or difficulty of performing the behavior, and is assumed to reflect past experience as well as anticipated obstacles. According to Ajzen, perceived behavioral control is a function of control beliefs in just the same way as subjective norm is a function of normative beliefs. It is assumed to have a direct influence on intention. For desirable behaviors, greater perceived behavioral control should lead to stronger intentions. Perceived behavioral control may also have a direct predictive effect on behavior, through two different mechanisms. First, holding intention constant, an individual with higher perceived behavioral control is likely to try harder and to persevere for longer than an individual who has lower perceived control. Second, people may have accurate perceptions of the amount of actual control they have over the behavior.

A number of meta-analyses of the TRA/TPB have been conducted. The findings show that when intention is predicted from attitude and subjective norm, or from attitude, subjective norm and perceived behavioral control, between 40 and 50 percent of the variance is explained, on average. When behavior is predicted from intention alone or from intention and perceived behavioral control, between 19 and 38 percent of the variance is explained (Sutton 1998).

1.5 Comparison of Social Cognition Models

The five social cognition models outlined above show a number of important similarities and differences. Some constructs are common to more than one model. For example, perceived susceptibility or perceived vulnerability occurs in both the HBM and PMT. Other constructs appear to be very similar, for example, perceived behavioral control and self-efficacy. Resolution of current controversies concerning the extent of overlap between such constructs requires the development of clear definitions, so that similar constructs can be distinguished on conceptual grounds, and more frequent tests of discriminant validity to investigate whether sets of apparently similar measures are tapping the same or different constructs.

All the models assume that individuals are future oriented and that they weigh up the costs and benefits of possible future courses of action. They all incorporate to a greater or lesser extent the expectancy–value principle which derives from the classical ex pected utility model, a normative model of decision making. In its weak form, this principle states that expectancies (subjective probabilities) and values are important determinants of behavior. In its strong form, the principle states that expectancies and values are combined multiplicatively, or at least that people behave as if they combine their cognitions in this way. Both the TRA and the TPB employ the strong form of the expectancy–value principle. The most systematic application of the principle is found in the subjective expected utility model (SEUM; Edwards 1954) which is based directly on expected utility theory. Versions of the SEUM have been applied to a number of health behaviors (e.g., Sutton et al. 1987). Researchers who use models such as the TPB and the SEUM need to be aware of the problems that arise from the use of multiplicative composites (Evans 1991).

This raises the thorny issue of rationality. Social cognition models are often criticized for offering an unrealistically rational account of how people form intentions and make decisions. However, the models do not imply that individuals always deliberate carefully and always make optimal decisions. People may not be aware of all the options available to them and of all the consequences that may follow from their actions. They may hold incorrect beliefs about the outcomes. They may make rapid decisions based on a few salient considerations. Having made a considered decision (e.g., to go jogging every Sunday morning), they do not necessarily have to weigh up the pros and cons again unless circumstances change; they may simply retrieve their previously formed intention from long-term memory and act on it. Thus, social cognition models imply a more limited rationality than is sometimes suggested by their critics.

Social cognition models are also sometimes criticized for being static. This criticism is unfounded: social cognition models summarize dynamic causal processes. In the TRA, for example, changes in behavioral beliefs and/or outcome evaluations are assumed to produce changes in attitude which in turn lead to changes in intention which ultimately produce changes in behavior. Models of health behavior should specify the time lags involved in these causal processes but most of them do not. However, it is often assumed implicitly that effects on intention are almost instantaneous whereas effects on behavior may be delayed.

Social cognition models differ in the degree to which they specify the content of the cognitions they identify. With the TRA, for example, once the behavior of interest has been defined, it is possible to generate questionnaire items for intention and for the direct measures of attitude and subjective norm. However, in order to generate items for behavioral beliefs, outcome evaluations, normative beliefs and motivations to comply, it is recommended that researchers gather information on salient beliefs from members of the target population. These constructs remain 'content-free' until such information is obtained.

Compared with the other models, research on the TRA and the TPB shows a relatively high degree of standardization of measures based on published recommendations (Ajzen and Fishbein 1980). In addition, these authors emphasize the principle of correspondence (or compatibility) which, put simply, states that for maximum prediction the measures of all the constructs in the model should use similar wording. This principle is not widely applied in research using the other social cognition models.

The models also differ with regard to their scope of application. Key constructs in the HBM and PMT include perceived susceptibility and perceived severity with respect to a given health threat. Although these components can be extended to nonhealth-related events, for example the risk of financial loss, the scope of both models is necessarily limited by the nature of these two constructs. By contrast, the other three models are general theories that can be applied to any domain or behavior, for example voting, career, and purchasing decisions. Stated differently, SET, and the TRA and the TPB, regard health behaviors as having the same proximal determinants as other kinds of behavior. Thus, they offer the potential benefit of parsimony.

2. Stage Models

Stage models use similar concepts to social cognition models but organize them in a different way. They are fundamentally different in structure from social cognition models (Weinstein et al. 1998). According to this approach, behavior change involves movement through a sequence of discrete, qualitatively distinct, stages. Different factors are assumed to be important at different stages. Hence, people in different stages are

Table 1
Stage definitions (from DiClemente et al. 1991)

Precontemplation
Currently smoking and 'not seriously considering quitting within the next 6 months'
Contemplation
Currently smoking and 'seriously considering quitting within the next 6 months'; 'were not considering quitting within the next 30 days, had not made a quit attempt of 24 h in the past year, or both'
Preparation
Currently smoking, 'were seriously considering quitting in the next 6 months and were planning to quit within the next 30 days,' and 'had made a 24 h quit attempt in the past year'
Action
Currently not smoking; quit in last 6 months
Maintenance
Currently not smoking; quit > 6 months ago

assumed to require different interventions to encourage or help them to move to the next stage in the sequence.

Current models of health behavior that incorporate stage assumptions include: the transtheoretical model (TTM; Prochaska et al. 1992) and variants of it (e.g., De Vries and Mudde 1998), the precaution adoption process model (Weinstein and Sandman 1992), the health action process approach (Schwarzer and Fuchs 1996), and the health behavior goal model (Gebhardt 1997). This section focuses on the TTM because it is the dominant model in the field.

The TTM developed from studies of the processes of change in psychotherapy and smoking cessation. Smoking still accounts for the majority of applications of the model but it has been applied to a wide range of other health behaviors, including condom use, exercise, sunscreen use, and healthy eating (Prochaska and Velicer 1997). Although it is often referred to simply as the stages of change model, the TTM includes 15 different theoretical constructs drawn from different theories of behavior change. These include the stages of change (which provide the basic organizing principle), the 10 processes of change, the perceived pros and cons of changing, and self-efficacy and temptation. The TTM was an attempt to integrate these different constructs in a single comprehensive framework—hence the name transtheoretical.

The version of the TTM used most widely in recent years specifies five stages: precontemplation, contemplation, preparation, action, and maintenance (DiClemente et al. 1991). Table 1 gives the operational definitions of the stages as applied to smoking cessation. Using a 'staging algorithm,' participants are classified into stages on the basis of their responses to a small number of questionnaire items. The first three categories contain the current smokers, the remaining two the ex-smokers; people who have never smoked are not represented in this scheme. Precontemplation, contemplation, and preparation are defined in terms of current intentions and past behavior (whether or not the person has made a 24 h quit attempt in the past year), whereas action and maintenance are defined purely in terms of behavior; ex-smokers' intentions are not taken into account.

Prochaska et al. (1992) represented the stages of change as a spiral. People start at the bottom in precontemplation. They then move through the stages in order (contemplation, preparation, action, maintenance) but may relapse to an earlier stage. They may cycle and recycle through the stages several times before achieving successful long-term behavior change.

Of the other constructs in the TTM, the 10 processes of change refer to the things that people think and do to help them move through the stages, the pros and cons are the perceived advantages and disadvantages of changing, self-efficacy is borrowed from Bandura's social cognitive theory, and temptation is a related concept that refers to the degree to which a person expects to feel tempted to lapse in different situations.

The TTM is popular among clinicians and practitioners and there is a large body of evidence which is interpreted by its proponents as supporting the model. However, a closer examination of this literature reveals a number of serious problems with the TTM itself and with much of the research based on it (Sutton 2000). These problems include: lack of standardization of measures, particularly of the central construct of stages of change; logical flaws in current staging algorithms; inadequate specification of the causal relationships among the different constructs; misinterpretation of cross-sectional data on stages of change; and confusion concerning the nature of stage models and how they should be tested.

It is unfortunate that the model that has dominated the field to date is surrounded by a 'thicket of problems,' to use Bandura's (1997) phrase. The notion that behavior change involves movement through a sequence of qualitative stages is an important idea that

deserves further consideration. Health behavior researchers should turn their attention to the other stage models that have been proposed.

3. Conclusion

One focus of current interest is the 'intention–behavior gap.' Researchers in the health behavior field are using Gollwitzer's (1993) concept of implementation intentions (the intention to do X in situation Y) and related ideas to explain why some people who have strong intentions to achieve a goal succeed whereas others fail, and to try to close this gap using simple interventions (e.g., Orbell et al. 1997). Related to this is the increasing interest in models of self-regulation (Abraham and Johnston 1998). Another trend, which has been apparent in social psychology in recent years and is now being imported into research on health behaviors, is to regard attitudes as being activated automatically (e.g., Fazio 1990). Although this seems to pose a challenge to social cognition models, the two approaches are not necessarily incompatible, especially if the idea is extended to automatic activation of previously formed intentions.

There is a plethora of models of health behavior and new models are continually being developed. Although this can be interpreted as a sign of a field rich in conceptual and theoretical development, it makes it increasingly difficult to accumulate research findings into a coherent body of knowledge. The field would benefit from clearer definition of concepts, greater standardization of measures, more tests of convergent and discriminant validity, greater concentration on a small number of models, and more empirical comparisons of models. The vast majority of studies use nonexperimental between-subjects designs. Obtaining repeated measures on the same individuals would allow models to be tested at the individual level (Hedeker et al. 1996). Experimental manipulation of constructs, as has been done in many studies of PMT, would provide much stronger evidence of causal effects.

See also: Control Beliefs: Health Perspectives; Health Behaviors; Health Behaviors, Assessment of; Health Risk Appraisal and Optimistic Bias; Health: Self-regulation; Illness Behavior and Care Seeking; Patient Adherence to Health Care Regimens; Self-efficacy and Health; Vulnerability and Perceived Susceptibility, Psychology of

Bibliography

Abraham C, Johnston M (eds.) 1998 Special issue on self-regulation and health. *Psychology and Health* **13**: 565–766

Ajzen I 1991 The theory of planned behavior. *Organizational Behavior and Human Decision Processes* **50**: 179–211

Ajzen I, Fishbein M 1980 *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall, Englewood Cliffs, NJ

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ

Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman, New York

Becker M H (ed.) 1974 *The Health Belief Model and Personal Health Behavior*. Slack, Thorofare, NJ

De Vries H, Mudde A N 1998 Predicting stage transitions for smoking cessation applying the attitude–social influence–-efficacy model. *Psychology and Health* **13**: 369–85

DiClemente C C, Prochaska J O, Fairhurst S K, Velicer W F, Velasquez M M, Rossi J S 1991 The process of smoking cessation: an analysis of precontemplation, contemplation, and preparation stages of change. *Journal of Consulting and Clinical Psychology* **59**: 295–304

Edwards W 1954 The theory of decision making. *Psychological Bulletin* **54**: 380–417

Evans M G 1991 The problem of analyzing multiplicative composites. *American Psychologist* **46**: 6–15

Fazio R H 1990 Multiple processes by which attitudes guide behavior: the MODE model as an integrative framework. *Advances in Experimental Social Psychology* **23**: 75–109

Floyd D L, Prentice-Dunn S, Rogers R W 2000 A meta-analysis of research on protection motivation theory. *Journal of Applied Social Psychology* **30**: 407–29

Gebhardt W A 1997 *Health Behaviour Goal Model: Towards a Theoretical Framework for Health Behavior Change*. University of Leiden, Leiden, The Netherlands

Gollwitzer P M 1993 Goal achievement: the role of intentions. *European Review of Social Psychology* **4**: 141–85

Harrison J A, Mullen P D, Green L W 1992 A meta-analysis of studies of the Health Belief Model with adults. *Health Education Research* **7**: 107–16

Hedeker D, Flay B R, Petraitis J 1996 Estimating individual influences of behavioral intentions: an application of random-effects modeling to the theory of reasoned action. *Journal of Consulting and Clinical Psychology* **64**: 109–20

Janz N K, Becker M H 1984 The health belief model: a decade later. *Health Education Quarterly* **11**: 1–47

Milne S, Sheeran P, Orbell S 2000 Prediction and intervention in health-related behavior: a meta-analytic review of protection motivation theory. *Journal of Applied Social Psychology* **30** 106–43

Orbell S, Hodgkins S, Sheeran P 1997 Implementation intentions and the theory of planned behavior. *Personality and Social Psychology Bulletin* **23**: 945–54

Prochaska J O, DiClemente C C, Norcross J C 1992 In search of how people change: applications to addictive behaviors. *American Psychologist* **47**: 1102–14

Prochaska J O, Velicer W F 1997 The transtheoretical model of health behavior change. *American Journal of Health Promotion* **12**: 38–48

Rogers R W 1983 Cognitive and physiological processes in fear appeals and attitude change: a revised theory of protection motivation. In: Cacioppo J T, Petty R E, Shapiro D (eds.) *Social Psychophysiology: A Sourcebook*. Guilford Press, New York, pp. 153–76

Schwarzer R, Fuchs R 1996 Self-efficacy and health behaviours. In: Conner M, Norman P (eds.) *Predicting Health Behaviour: Research and Practice with Social Cognition Models*. Open University Press, Buckingham, UK, pp. 163–96

Sutton S 1998 Predicting and explaining intentions and behavior: how well are we doing? *Journal of Applied Social Psychology* **28**: 1317–38

Sutton S, Marsh A, Matheson J 1987 Explaining smokers' decisions to stop: test of an expectancy-value approach. *Social Behaviour* **2**: 35–49

Sutton S A 2000 A critical review of the transtheoretical model applied to smoking cessation. In: Norman P, Abraham C, Conner M (eds.) *Understanding and Changing Health Behaviour: From Health Beliefs to Self-Regulation*. Harwood Academic Publishers, Amsterdam, pp. 207–25

Weinstein N D, Rothman A J, Sutton S R 1998 Stage theories of health behavior: conceptual and methodological issues. *Health Psychology* **17**: 290–99

Weinstein N D, Sandman P M 1992 A model of the precaution adoption process: evidence from home radon testing. *Health Psychology* **11**: 170–80

Wurtele S K, Maddux J E 1987 Relative contributions of protection motivation theory components in predicting exercise intentions and behavior. *Health Psychology* **6**: 453–66

S. Sutton

Health Behaviors

1. Introduction

Interest in behaviors that have important impacts on our health and well being is based upon two assumptions; (a) that a significant proportion of the mortality from the leading causes of death is caused by the behavior of individuals, and (b) that such behavior is modifiable (Conner and Norman 1996). Behavior is held to exert its influence on health in three basic ways: by producing direct biological changes, by conveying health risks or protecting against them, or by leading to the early detection or treatment of disease (Baum and Posluszny 1999).

This chapter begins by offering a definition of health behavior and the different types of health behavior. The prevalence of key health behaviors and their relationship to morbidity and mortality is then examined. The chapter then considers who performs these different behaviors, why they might do so, and how health enhancing behaviors might be encouraged and health impairing behaviors discouraged.

2. The Definition of Health Behaviors

Health behaviors have been defined in various ways. For example, Conner and Norman (1996) define them as any activity undertaken for the purpose of preventing or detecting disease or for improving health and well being. Gochman (1997) in the *Handbook of Health Behavior Research* defines them as 'behavior patterns, actions and habits that relate to health maintenance, to health restoration and to health improvement' (Vol. 1, p. 3). Behaviors within this definition include medical service usage (e.g., physician visits, vaccination, screening), compliance with medical regimens (e.g., dietary, diabetic, antihypertensive regimens), and self-directed health behaviors (e.g., diet, exercise, smoking, alcohol consumption). All have received considerable attention from social and behavioral researchers and we now have a good understanding of the factors influencing how and why individuals engage in such behaviors.

In describing health behaviors it is common to distinguish health enhancing from health impairing behaviors. Health impairing behaviors have harmful effects on health or otherwise predispose individuals to disease. Such behaviors include smoking, excessive alcohol consumption, and high dietary fat consumption. In contrast, engagement in health enhancing behaviors convey health benefits or otherwise protect individuals from disease. Such behaviors include exercise, fruit and vegetable consumption, and condom use in response to the threat of sexually transmitted diseases.

3. Key Health Behaviors

Numerous studies have examined the relationship between health behaviors and health outcomes (e.g., Blaxter 1990) and have demonstrated their role in both morbidity and mortality. One of the first such studies identified seven features of lifestyle which were associated with lower morbidity and higher subsequent long-term survival: not smoking, moderate alcohol intake, sleeping 7–8 h per night, exercising regularly, maintaining a desirable body weight, avoiding snacks, and eating breakfast regularly (Belloc and Breslow 1972). Health behaviors also impact upon individuals' quality of life, by delaying the onset of chronic disease and extending active lifespan. Smoking, alcohol consumption, diet, gaps in primary care services and low screening uptake are all significant determinants of poor health, and changing such behaviors should lead to improved health. For example, in the USA, Healthy People 2000 (US Department of Health and Human Services [USDHHS] 1990) lists increased physical activity, changes in nutrition and reductions in tobacco, alcohol and drug use as important for health promotion and disease prevention.

3.1 Smoking

Smoking is the health behavior most closely linked with long-term negative health outcomes (see *Smoking and Health*). Morbidity and mortality from coronary heart disease (CHD) are increased among smokers (Doll et al. 1994). Moreover, there is a strong positive relationship between the number of cigarettes smoked per day and the incidence of CHD (Friedman et al. 1979). Smoking has also been linked to a number of cancers, including cancer of the lung, throat, stomach and bowel as well as a number of more immediate

negative health effects such as reduced lung capacity and bronchitis (Royal College of Physicians 1983). Despite the array of negative health outcomes, smokers often report positive mood effects from smoking and the use of smoking as a strategy for coping with stress.

The number of people smoking in the USA and UK has shown a steady decline over the past twenty years. Data from the General Household Survey showed that 28 percent of people over the age of 16 smoke in the UK. Smoking is more common among men and among unskilled manual workers (General Household Survey 1994). A similar pattern is evident in the USA and other First World countries, with smoking more common among less educated, lower income and minority groups (Rigotti 1989). Those who quit smoking reduce the risk to their health, particularly if they quit before 35 years of age (Doll et al. 1994) (see *Smoking Prevention and Smoking Cessation*).

3.2 Diet

The impact of diet upon morbidity and mortality are well established (USDHHS 1988). In the Third World, the problems related to diet and health are ones of under-nutrition; in the First World, the problems are predominantly linked to overconsumption of food. In Western industrialized countries excessive fat consumption and insufficient fiber, fruit and vegetable consumption are related to health problems. In addition, excess consumption of calories combined with insufficient exercise has made obesity a major health problem. Diet has been implicated in cardiovascular diseases (CVDs), strokes and high blood pressure, cancer, diabetes, obesity, osteoporosis, and dental disease.

It is generally agreed that elevated blood cholesterol level is a major risk factor for the development of CVD (Consensus Development Conference on Lowering Blood Cholesterol to Prevent Heart Disease 1985). Nutbeam and Catford (1990) estimate that 26 percent of men and 25 percent of women in the UK have cholesterol levels greater than 6.5 mmol l^{-1} (a level considered to be excessive). While in the USA, it is estimated that 50 percent of the adult population is at risk of CHD by virtue of elevated blood cholesterol levels (Sampos et al. 1989). The reduction of blood cholesterol via dietary change is now widely accepted as an important way of tackling CHD. Dietary recommendations include reducing fat in the diet and increasing soluble fiber intake. However, their impact upon cholesterol levels may be limited.

3.3 Exercise

The potential health benefits of engaging in regular exercise include reduced cardiovascular morbidity and mortality, lowered blood pressure, and the increased metabolism of carbohydrates and fats, as well as a range of psychological benefits such as improved self-esteem, positive mood states, reduced life stress and anxiety (see *Physical Activity and Health*). Nevertheless, a significant proportion of the population lead a sedentary lifestyle. The General Household Survey (1989) indicated that only one in three men and one in five women in the UK participate in any sport or recreational physical activity. Moreover, the Allied Dunbar Fitness Survey (1992) of 6,000 English adults reported that one in six adults had done no exercise (i.e., for 20 minutes or more at a moderate or vigorous level) in the previous four weeks. Participation in regular exercise is strongly related to a number of sociodemographic variables. In particular, young people and males are more likely to engage in regular exercise. For example, the 1988 Welsh Heart Health Survey (HPAW 1990) reported that among 18–34 year olds, 61 percent of men engaged in moderately vigorous exercise at least two times a week compared with only 35 percent of women. For 35–64 year olds, the percentages drops to 37 percent for men and 17 percent for women. Overall, across First World countries the typical exerciser is likely to be young, well educated, affluent and male.

3.4 Health Screening

Individuals may seek to protect their health by participating in various screening programs which attempt to detect disease at an early, or asymptomatic, stage. In the UK, screening programs have been set up for various diseases, including anemia, diabetes, bronchitis, cervical cancer, and breast cancer. Considering breast cancer, it has been estimated that breast screening programs which include mammograms can reduce breast cancer mortality by up to 40 percent among women aged 50 and over (Strax 1984). However, participation rates in breast screening programs show great variability across different countries, ranging from 25 percent to 89 percent (Vernon et al. 1990). Participation tends to be negatively related to age, and positively related to education level and socioeconomic status.

3.5 Sexual Behaviors

Sexual behaviors are considered health behaviors because of their impact upon the spread of sexually transmitted diseases (STDs) such as gonorrhea and syphilis (see *Sexually Transmitted Diseases: Psychosocial Aspects*). More recently, the role of sexual behaviors in the spread of human immunodeficiency virus (HIV) has been a focus of attention. While early health education campaigns emphasized the need to reduce the number of sexual partners or avoid par-

ticular sexual practices (e.g., anal sex, penetrative sex), more recently the focus has been upon the use of condoms during penetrative sex to reduce the risk of HIV transmission. Condom use is particularly recommended for those with multiple partners or those who do not know their partners' sexual history. For these reasons, much of the health advice concerning condom use has been focused on young people (see *Sexual Risk Behaviors*).

There seems to be considerable variation in the use of condoms in response to the threat of HIV/AIDS. For example, studies in the UK and USA report rates of between 24 and 58 percent (Fife-Schaw and Breakwell 1992, Gerrard et al. 1996). The General Household Survey (1993) in the UK reported changes by age group in the use of condoms for the period 1983 to 1991. Among 16–24 year olds, condom use increased from around 6 percent to around 12 percent during this period, while among 40–49 year olds it dropped from around 18 percent to around 12 percent over the same period.

3.6 Alcohol Use

Moderate alcohol consumption has been linked to positive health outcomes. However, high alcohol consumption has been linked to a range of negative health outcomes including high blood pressure, heart disease and cirrhosis of the liver. High levels of alcohol consumption have also been associated with accidents, injuries, suicides, crime, domestic violence, rape, murder and unsafe sex (British Medical Journal 1982). While many of the adverse effects of high alcohol consumption are due to continued heavy drinking (e.g., cirrhosis of the liver, heart disease), others are more specifically related to excessive alcohol consumption in a single drinking session (e.g., accidents, violence).

The General Household Survey (1992) reported that the average weekly consumption of alcohol in the UK was 15.9 units (1 unit = 1 glass of wine, 1 measure of spirits, or 0.5 pints of beer) for men and 5.4 for women. In addition, 27 percent of men and 11 percent of women were drinking more than the recommended weekly sensible limits (21 units for men, 14 units for women). Heavy drinking is also more likely among younger age groups. In a survey of 12,000 Welsh adults, Moore et al. (1994) reported that 31.1 percent of drinkers aged 18–24 engaged in binge drinking (i.e., drinking half the recommended weekly consumption of alcohol in a single session) at least once a week.

4. Relationship of Health Behaviors to Sociodemographic Factors

A clearer understanding of why individuals perform health behaviors might assist in the development of interventions to help individuals gain health benefits. A variety of factors have been found to account for individual differences in the performance of health behaviors. Demographic variables show reliable associations with the performance of health behaviors. For example, there is a curvilinear relationship between many health behaviors and age, with high incidences of many health risking behaviors such as smoking in young adults and much lower incidences in children and older adults (Blaxter 1990). Such behaviors also vary by gender, with females being generally less likely to smoke, consume large amounts of alcohol, engage in regular exercise, but more likely to monitor their diet, take vitamins and engage in dental care (Waldron 1988). Differences by socioeconomic status and ethnic group are also apparent for behaviors such as diet, exercise, alcohol consumption and smoking (e.g., Blaxter 1990) (see *Socioeconomic Status and Health*).

Generally speaking, younger, wealthier, better educated individuals, under low levels of stress, with high levels of social support are more likely to practice health protective behaviors. Higher levels of stress and/or fewer resources are associated with health risking behaviors such as smoking and alcohol abuse (Adler and Matthews 1994). Social factors seem to be important in instilling health behaviors in childhood. Parent, sibling and peer influences are important, for example in the initiation of smoking. Cultural values also have a major impact, for instance in determining the number of women exercising in a particular culture. For example, Steptoe and Wardle (1992) report that between 34 and 95 percent of women in their European student sample had exercised in the past 14 days.

5. Understanding the Distribution/Prevalence of Health Behaviors

Perceived symptoms control health behaviors when, for example, a smoker regulates his/her smoking on the basis of sensations in the throat. Access to medical care has been found to influence the use of such health services (e.g., Black Report 1988). Personality factors have also been associated with health behaviors (Adler and Matthews 1994). Cognitive factors also determine whether or not an individual practices health behaviors and may explain how other factors influence behavior. Knowledge about behavior–health links is an important factor in an informed choice concerning health behaviors. Various cognitive variables have been studied including perceptions of health risk, efficacy of behaviors in influencing this risk, social pressures to perform the behavior, and control over performance of the behavior. The relative importance of various cognitive factors in determining who performs various health behaviors constitutes the basis of different models. Such models have been labeled social cognition models (SCMs) because of their focus

on cognitive variables as the primary determinant of individual social behaviors. These SCMs provide a basis for understanding the determinants of behavior and behavior change (see *Health Behavior: Psychosocial Theories*). Each of these models emphasize the rationality of human behavior and assume that behavior is based upon elaborate, but subjective, cost–benefit analysis of the likely outcomes of differing courses of action. It is assumed that individuals generally aim to maximize benefits and minimize costs in selecting a behavior.

5.1 Health Belief Model

The Health Belief Model (HBM) attempts to conceptualize two types of health beliefs that make a behavior in response to illness more or less attractive (Sheeran and Abraham 1996): perceptions of the threat of illness and evaluation of the effectiveness of behaviors to counteract this threat. Threat perceptions depend upon the perceived susceptibility to the illness and the perceived severity of the consequences of the illness. Together these variables determine the likelihood of the individual following a health-related action, although their effect is modified by demographic variables, social pressure and personality. The particular action undertaken is determined by the evaluation of the possible alternatives. This behavioral evaluation depends upon beliefs concerning the benefits or efficacy of the health behavior and the perceived costs or barriers to performing the behavior. Hence, individuals are likely to follow a particular health behavior if they believe themselves to be susceptible to a particular condition or illness which they consider to be serious, and believe the benefits of the behavior undertaken to counteract the condition or illness outweigh the costs. It is assumed that this whole process is set in motion by cues to action. Cues to action include a diverse range of triggers to the individual taking action and are commonly divided into factors that are internal (e.g., physical symptoms) or external (e.g., mass media campaigns, advice from others) to the individual. Other influences upon the performance of health behaviors, such as demographic factors or psychological characteristics (e.g. personality, peer pressure, perceived control over behavior) are assumed to exert their effect via changes in the components of the HBM.

5.2 Theory of Planned Behavior

The Theory of Planned Behavior (TPB) was developed by social psychologists and has been widely employed as a tool to aid our understanding of a variety of behaviors including health behaviors (Ajzen 1991, Godin and Kok 1996). The TPB details how the influences upon an individual determine that individual's decision to follow a particular behavior. Within the TPB, the determinants of behavior are intentions to engage in that behavior and perceived behavioral control (PBC) over that behavior. Intentions represent a person's motivation. The construct is conceptualized as an individual's conscious plan or decision to exert effort in order to engage in a particular behavior. Perceived behavioral control is a person's expectancy that performance of the behavior is within his/her control. Intentions are determined by three variables. The first is attitudes, which are an individual's overall evaluation of the behavior. The second is subjective norms, which consist of a person's beliefs about whether significant others think he/she should engage in the behavior. The third measures the extent to which the individual perceives that the behavior is under their personal control and is labeled PBC.

The attitude, subjective norm and PBC components are determined by underlying beliefs. Attitude is a function of a person's salient behavioral beliefs; which represent perceived likely consequences of the behavior (e.g., taking exercise will reduce my risk of heart disease). Subjective norm is a function of normative beliefs, which represent perceptions of specific salient others' preferences about whether one should or should not engage in a behavior (e.g., my family think I should take exercise). PBC is based on beliefs concerning access to the necessary resources and opportunities to perform the behavior successfully (e.g., I have easy access to a place where I can exercise).

So, according to the TPB, individuals are likely to engage in a health behavior if they believe that the behavior will lead to particular outcomes which they value, if they believe that people whose views they value think they should carry out the behavior, and if they feel that they have the necessary resources and opportunities to perform the behavior.

5.3 Perceived Self-efficacy

Self-efficacy is one of the most powerful predictors of health behavior (Bandura 1997, Schwarzer and Fuchs 1996). It has its origins in Bandura's social cognitive theory, which states that behavior is a function of both incentives (i.e., reinforcements) and expectancies. Three kinds of expectancies can be identified, these being: situation–outcome expectancies, which refer to beliefs about how events are connected; outcome expectancies, which refer to beliefs about the consequences of performing a behavior; and self-efficacy expectancies, which refer to beliefs about one's ability to perform the behavior. Thus, in order to perform a health behavior, individuals must value their health (i.e., have an incentive), believe that their current lifestyle poses a threat to their health (i.e., situation–outcome expectancy), believe that adopting the new behavior will reduce the threat to their health (i.e., outcome expectancy), and believe that they are capable

of performing the behavior (i.e., self-efficacy expectancy). While all these beliefs are seen to be important in the initiation and maintenance of health behavior, perceived self-efficacy is seen to be the most important. Individuals with a strong sense of self-efficacy are believed to develop stronger intentions to act, to expend more effort to achieve their goals and to persist longer in the face of barriers and impediments. Perceived self-efficacy is therefore believed to play a crucial role in the determination of health behavior (see *Health: Self-regulation*). According to Bandura (1997), self-efficacy can be conceptualized and measured in terms of three parameters; magnitude, strength and generality, with strength being the key parameter. This is reflected in its measurement (e.g., 'I am confident that I can refrain from smoking, even if someone offers me a cigarette').

5.4 Stage Models of Health Behavior

A number of researchers have suggested that there may be qualitatively different stages in the initiation and maintenance of health behavior, and that to obtain a full understanding of the determinants of health behavior it is necessary to conduct a detailed analysis of the nature of these stages. One of the first stage models was put forward by Prochaska and DiClemente in their transtheoretical model of change (TTM). Their model has been widely applied to analyze the process of change in alcoholism treatment, smoking cessation, and psychotherapy. In its most widely applied form, Prochaska et al. (1992) identify five stages of change: precontemplation, contemplation, preparation, action, and maintenance. Individuals are seen to progress through each stage to achieve successful maintenance of a new behavior. Taking the example of smoking cessation, it is argued that in the precontemplation stage the smoker is unaware that his/her behavior constitutes a problem and has no intention to quit. In the contemplation stage, the smoker starts to think about changing his/her behavior, but is not committed to try to quit. In the preparation stage, the smoker has an intention to quit and starts to make plans about how to quit. The action stage is characterized by active attempts to quit, and after six months of successful abstinence the individual moves into the maintenance stage characterized by attempts to prevent relapse and to consolidate the newly acquired nonsmoking status.

6. Changing Health Behaviors

The above models detail the key cognitive determinants of health behavior. To the extent that these models outline the key social cognitive determinants of health behavior, interventions which target these variables should lead to associated changes in behavior. For example, the HBM would suggest that encouraging health behaviors is best achieved by increasing individuals' perceived susceptibility to negative health outcomes and making individuals aware of the severity of such outcomes. Such approaches have been commonly employed in health promotion messages. In addition, the HBM might suggest the need to focus on the benefits of health behaviors and the fact that barriers to action are easily overcome. However, there has been little systematic evaluation of their effectiveness, perhaps due to the common sense appeal of these approaches.

Relatively few studies have used the TPB as a framework for developing interventions, despite the fact that the theory would suggest a number of interventions focusing on different components of the model. Brubaker and Fowler (1990) did examine the effect of persuasive messages tackling behavioral beliefs upon men's intentions to perform testicular self-examination in response to the threat of cancer. A persuasive message was found to increase intentions to perform testicular self-examination compared to a no message control. The TPB would also suggest the need to tackle normative beliefs and control beliefs in any attempt to change behavior. A number of studies have attempted to use persuasive messages aimed at tackling normative pressures. For example, in relation to preventing adolescent smoking, prevention programs commonly attempt to tackle the perceived pressure from teenage peers who smoke. Tackling control beliefs has been seen to bear many similarities to changing perceptions of self-efficacy (see below).

Another interesting approach has focused directly on the immediate determinant of behavior in the TPB: intentions. Where individuals do have an intention to engage in a health behavior (goal intentions), but are having trouble implementing their intention, forming a specific plan about where and when to act has been found to help (Gollwitzer 1993). For example, Orbell et al. (1997) gave out a questionnaire about breast self-examination in response to the threat of cancer. Half the women were asked to indicate when and where in the next month they intended to perform breast self-examination (a specific plan or implementation intention). A one month follow-up found that 64 percent of these women had performed breast self-examination that month compared with only 16 percent of women who hadn't made an implementation intention, despite having similar goal intentions.

Several studies have focused on enhancing feelings of self-efficacy as a means for encouraging health behavior change. As Bandura (1997) outlines, there are four main sources of self-efficacy, each of which could be addressed in interventions. First, individuals can develop feelings of self-efficacy from personal mastery experience (e.g., it may be possible to split a behavior into various subgoals, such that the easiest subgoals are achieved before more difficult tasks are attempted). Second, individuals may develop feelings

of self-efficacy through observing other people succeed on a task (i.e., vicarious experience). Third, it is possible to use standard persuasive techniques to try to instill feelings of self-efficacy. Finally, one's physiological state may be used as a source of information, such that high levels of arousal or anxiety may indicate to the individual that he or she is not capable of performing a given action (e.g., relaxation techniques may be employed to help maintain feelings of self-efficacy).

Stage models have also been used as the basis for designing interventions. The main distinction from the approaches already presented is that this approach suggests that interventions need to be matched to the needs of each individual. So, for example, an individual at the pre-contemplation stage in relation to exercise needs to be made aware of the health problems associated with a lack of exercise. A different individual in the contemplation stage needs information about the pros of changing their exercise behavior and the cons of not changing. While an individual in the preparation or action stage needs helps with acting on their plans to engage in exercise. Whilst these ideas appear entirely plausible, to date relatively few studies have managed to demonstrate that such stage-matched interventions are more effective than presenting the same intervention to all individuals.

7. Conclusions

Health behaviors have important consequences for both the quality and length of life by influencing disease outcomes. Nevertheless, there is still considerable variation in those individuals who perform these behaviors. Social cognition models provide one approach to understanding the variation in who performs health behaviors. These models are also useful because they suggest ways in order to change health behaviors in order to improve health. Perhaps the greatest challenge for social and behavioral research on health behaviors is the demonstration that such theory-based interventions can produce effective and long-lasting behavior change that will lead to real health benefits for all individuals.

See also: Health Education and Health Promotion; Health Risk Appraisal and Optimistic Bias

Bibliography

Adler N, Matthews K 1994 Health psychology: why do some people get sick and some stay well? *Annual Review of Psychology* **45**: 229–59

Ajzen I 1991 The theory of planned behavior. *Organizational Behavior and Human Decision Processes* **50**: 179–211

Allied Dunbar Fitness Survey 1992 Health Education Authority, London

Bandura A 1997 *Self-efficacy: The Exercise of Control*. W. H. Freeman, New York

Baum A, Posluszny D M 1999 Health psychology: mapping biobehavioral contributions to health and illness. *Annual Review of Psychology* **50**: 137–63

Belloc N B, Breslow L 1972 Relationship of physical health status and health practices. *Preventive Medicine* **9**: 409–421

Black Report 1988 *Inequalities in Health: The Black Report and the Health Divide*. DHSS, London

Blaxter M 1990 *Health and Lifestyles*. Routledge, London

British Medical Journal 1982 *Alcohol Problems*. BMJ Publications, London

Brubaker R G, Fowler C 1990 Encouraging college males to perform testicular self-examination: evaluation of a persuasive message based on the revised theory of action. *Journal of Applied Social Psychology* **20**(17): 1411–22

Conner M, Norman P 1996 (eds.) *Predicting Health Behaviour*. Open University Press, Buckingham, UK

Consensus Development Conference on Lowering Blood Cholesterol to Prevent Heart Disease 1985 Lowering blood cholesterol to prevent heart disease. *J. of the Australian Medical Association* **253**: 2080–6

Doll R, Peto R, Wheatley K, Gray R, Sutherland I 1994 Mortality in relation to smoking: 40 years' observations on male British doctors. *British Medical Journal* **309**: 901–11

Fife-Schaw C R, Breakwell G M 1992 Estimating sexual behaviour parameters in the light of AIDS: a review of recent UK studies of young people. *AIDS Care* **4**: 187–201

Friedman G D, Dales L G, Ury H K 1979 Mortality in middle-aged smokers and non-smokers. *New England Journal of Medicine* **300**: 213–7

General Household Survey 1989 OPCS, London

General Household Survey 1992 OPCS, London

General Household Survey 1993 OPCS, London

General Household Survey 1994 OPCS, London

Gerrard M, Gibbons F X, Bushman B J 1996 Relation between perceived vulnerability to HIV and precautionary sexual behavior. *Psychological Bulletin* **119**: 390–409

Gochman D S (ed.) 1997 *Handbook of Health Behavior Research*. Plenum, New York, Vols. 1–4

Godin G, Kok G 1996 The theory of planned behavior: a review of its applications to health-related behaviors. *American Journal of Health Promotion* **11**: 87–98

Gollwitzer P M 1993 Goal achievement: the role of intentions. In: Stroebe W, Hewstone M (eds.) *European Review of Social Psychology*. Wiley, Chichester, UK, Vol. 4, pp. 141–85

Health Promotion Authority for Wales 1990 *Health in Wales*. Health Promotion, Cardiff, UK

Moore L, Smith C, Catford J 1994 Binge drinking: prevalence, patterns and policy. *Health Education Research* **9**: 497–505

Nutbeam D, Catford J 1990 Modifiable risks for cardiovascular disease among general practitioners in Wales. *Public Health* **104**: 353–61

Orbell S, Hodgkins S, Sheeran P 1997 Implementation intentions and the theory of planned behavior. *Personality and Social Psychology Bulletin* **23**: 945–54

Prochaska J O, DiClemente C C, Norcross J C 1992 In search of how people change: applications to addictive behaviors. *American Psychologist* **47**: 1102–14

Rigotti N A 1989 Cigarette smoking and body weight. *New England Journal of Medicine* **320**: 931–3

Royal College of Physicians 1983 *Health or Smoking: Follow-Up Report*. Pitman, London

Sampos C, Fulwood R, Haines C et al. 1989 The prevalence of high blood cholesterol levels among adults in the United

States. *Journal of the American Medical Association* **262**: 45–52
Schwarzer R, Fuchs R 1996 Self-efficacy and health behaviors. In: Conner M, Norman P (eds.) *Predicting Health Behaviour*. Open University Press, Buckingham, UK, pp. 163–196
Sheeran P, Abraham S C S 1996 The health belief model. In: Conner M, Norman P (eds.) *Predicting Health Behaviour*. Open University Press, Buckingham, UK, pp. 23–61
Steptoe A, Wardle J 1992 Cognitive predictors of health behaviour in contrasting regions of Europe. *British Journal of Clinical Psychology* **31**: 485–502
Strax P 1984 Mass screening for control of breast cancer. *Cancer* **53**: 665–70
US Department of Health and Human Services 1988 *The Surgeon General's Report on Nutrition and Health*. US Department of Health and Human Services, Public Health Service, Washington, DC
US Department of Health and Human Services 1990 *Health People 2000: National Health Promotion and Disease Prevention Objectives*. US Department of Health and Human Services, Public Health Service, Washington, DC
Vernon S W, Laville E A, Jackson G L 1990 Participation in breast screening programs: a review. *Social Science and Medicine* **30**: 1107–18
Waldron I 1988 Why do women live longer than men? *Journal of Human Stress* **2**: 2–13

M. Conner

Health Behaviors, Assessment of

1. Introduction: Health Behaviors as a Multidimensional Construct

In the field of health psychology, health behaviors are most often defined as behaviors that people engage in to maintain or improve their current health. These definitions focus on subjective health beliefs since they include any behaviors performed by a person in order to protect, promote, or maintain his or her health, whether or not such behaviors are objectively effective towards that end. In contrast, most empirical studies in public health and epidemiological research focus on behaviors which are defined by experts as having a substantial relationship to health. Therefore, the term health behavior has two different connotations. First, a subjective one which focuses on which behaviors lay people believe may affect their health, and second, an objective one which focuses on what experts evaluate and define as health-affecting behaviors. However, lay and expert views on which behaviors are related to health may not be consistent. In addition, several studies have found little consistency in people's health habits. Therefore, a person who exercises will not necessarily go for a regular check-up at the dentist. One reason people's current health habits are not more interdependent is the fact that health behaviors differ on a number of dimensions and may be influenced by different factors. The modest link between health behaviors and the high variability in their diffusion raises the question of whether it makes sense to talk about health behavior as a unitary construct.

Furthermore, health behaviors could be examined as antecedents or causes of health status. Such an approach is taken by social epidemiologists. In contrast, health educators are interested in health behaviors as objective dimensions for systematic interventions directed at producing behavioral changes. Finally, as subjects of scientific study, health behaviors are analyzed as consequences or outcomes, as well as correlates of personal factors. For an accurate and reliable measurement of health behaviors it is essential to distinguish between their different dimensions and to define clearly the topic to be investigated.

2. Measurement of Health Behaviors

There are various methods to assess health behaviors. On rare occasions, physiological methods can be used. This is the most accurate way to measure alcohol consumption (via blood or urine sampling), drug consumption (via immunoassay, hair or sweat bioassay procedures), habitual dietary intakes (via biochemical markers), or physical activity (via doubly labeled water). However, such bioassay methods are only required when a high level of accuracy about recent health behavior is needed (e.g., for workplace drug testing) or used in addition to self-reported data to provide a means of confirming or falsifying self-reported information (e.g., about recent drug use). However, in some circumstances it may be appropriate only to make respondents believe that there is a way of finding out about their behaviors via physiological measures to reduce misreporting.

Other direct methods are behavioral observation techniques, which are used to assess physical activity among children or a driver's choice of speeds, to name two examples.

Another alternative for observing 'true' health behaviors is unobtrusive, seminaturalistic measures, such as video-based simulation which involves participants viewing a video film of various situations relating to the task at hand and measuring responses to that task (e.g., a video-simulation measure of drivers' speed choice). This method could provide ecologically valid measures which simultaneously satisfy the demands of internal and external validity. Another possibility for seminaturalistically measuring health behaviors has been proposed by Rothman et al. (1993), who offered participants the opportunity to order a free sample of suntan lotion and to indicate their preferred protection factor rather than asking for behavioral intention. However, these techniques can only be approximations of 'reality' because they are often confounded with other variables. In the study of

Rothman et al. (1993), one could argue that the participants ordered the suntan lotion because it was for free. In addition, behavioral observation techniques are often impractical for broader samples as well as for complex and habitual health behaviors. Therefore, self-reports are a crucial—and often the only—source of information about health behaviors. There are different kinds of self-reports, which can be distinguished roughly into qualitative and quantitative methods.

3. *Qualitative Methods*

Unstructured or semistructured interviews are useful qualitative techniques for researchers interested in understanding individuals' cognitive and conceptual models of health behavior and the frames of reference that individuals use to organize health behaviors. Therefore, qualitative methods are mainly concerned with exploration and analysis of health behavior because they allow the interviewee room to address the issues which he or she feels relevant to the topics being raised by the investigator. For instance, detailed interviews and ethnography could give insight into subjective and experiential dimensions of those behaviors related to substance misuse. One major disadvantage of qualitative methods is the fact that generality is, by definition, not quantifiable. Furthermore, since anonymity is not given, self-reports elicited by interviews may be affected by social desirability biases which lead to over-reporting of socially desirable behavior as well as under-reporting of socially undesirable behavior. The results from the National Household Survey of Drug Abuse (NHSDA) showed that the proportion of the respondents reporting marijuana use was higher when questions were self-administered than when they were administered by an interviewer.

Alternatively, one could introduce a 'reality check' on the respondent's account by using the *method of focus groups* (Morgan 1997). A focus group is a collective, discussion-based interview orchestrated by a moderator who uses the small group setting to achieve a group dynamic to construct a supportive social context that allows respondents to reveal thoughts and behaviors they would normally be reluctant to disclose. Hence, this method could be useful for generating information about behaviors that are rarely talked about (e.g., sexual behaviors and drug use).

4. *Quantitative Methods*

Stone and Shiffman (1994) have labeled strategies for collecting self-reports of respondents' momentary or current state as *Ecological Momentary Assessment* (EMA). EMA studies usually consist of repeated assessment of participants' momentary state as they go about the tasks of daily life in their natural environment. There are several ways to approach this sampling task. *Interval-contingent* assessments require assessment at regular intervals. One example is the method of *interactive voice response* (IVR), where alcoholics are asked to call in on a regular basis to report their drinking to interviewers. Another way is asking subjects to record every episode of smoking, eating, or some other behavior of interest. This *event-contingent* approach may not lead to a representative sample of the participant's general state and requires a clear definition of the triggering event. In contrast, *signal-contingent* sampling supplies participants with an external signal cue, which is usually timed to be emitted at random, to prompt subjects to complete a written assessment or an electronic diary. As signal devices, beepers, electronic watches, and palmtop computers have been used. Investigations of temporal and contextual patterns of smoking have shown that global self-reports are largely inaccurate in comparison with a computerized EMA. Hence, EMA is a method which precisely assesses recent health behaviors. The major advantage of this method is that it minimizes reliance on recall from memory by relying on respondents' reports of their experience at the moment of the inquiry.

Diary log techniques are data collection strategies which gather information in relation to the passage of time. Therefore, the distinctive feature of this method is that it yields information which is temporally ordered. It tells the sequence of events and the profile of actions across time. Diary techniques can be particularly useful when data from the same person are required over a considerable period of time and/or very frequently, such as for assessing smoking behavior, alcohol consumption, or dietary habits to provide a general estimate of an individual's consumption. For example, alcohol consumption diaries often include questions about the frequency of consumption, the type of drink taken, and the quantity consumed on each occasion. In comparison to questionnaires, the diary log format minimizes recall biases associated with retrospective reporting, but daily reporting may be more reactive. In addition, diaries could be valuable for getting access to so-called 'intimate' information (e.g., sexual behavior).

Timeline Followback Method Reports (TLFB) have been developed by Sobell and Sobell (1995). They provide a detailed insight into health behaviors such as smoking, drug abuse, or drinking over a designated time period. Participants are required to provide retrospective estimates of their daily behavior by using a calendar over a certain time period, which can range up to 12 months prior to the interview. With this method, the pattern, variability, and level of drinking or smoking can be profiled, which is especially useful when relatively precise estimates are needed or when

researchers wish to evaluate specific changes in health behaviors before, during, and after interventions.

A questionnaire, which assesses past behavior frequency, is the most commonly used method to measure health behaviors. There are numerous questionnaires which ask for the average or typical quantity and frequency of alcohol use (see Sobell and Sobell 1995), dietary habits, or physical activity (see Pereira et al. 1997). However, the information provided by quantity and frequency measures (QF estimates) is limited by the fact that respondents must abstract an estimate from highly variable experiences. QF estimates often reflect less drinking and tend to misclassify drinkers compared with daily diary or timeline reports. They also provide lower absolute food intake estimates than a longer, interviewer-administered diet history.

5. *Distortions in Self-reports of Health Behaviors*

There are a series of methodological problems shared by all surveys relying on self-reports which could seriously decrease the internal and external validity of a survey, such as the ambiguity surrounding health behavior terminology. The term 'sexual partner,' for instance, may be interpreted in many different ways by the respondents, ranging from describing someone with whom one has experienced orgasm to someone with whom one has shared a bed while naked. In addition, the target population should be carefully selected because health behaviors have different meanings for adolescents and adults.

Furthermore, short-term fluctuations, for example in substance use, produced by environmental (e.g., social settings) and psychological (e.g., mood or stress) variables, may affect the psychometric properties of use measures. Hence, seasonal effects and short-term fluctuations may lead to superficial behavioral changes which could be misinterpreted by researchers as genuine changes.

Survey questions must be fair and unbiased to obtain valid answers. Asking leading questions, which suggest an expected answer, or loaded questions, which have emotionally laden connotations, can quickly invalidate a survey. In addition, by their very nature, questions about past behaviors assume accurate memory for events as well as willingness to report these to a researcher. However, respondents may not recall the actual events, employing instead various cognitive heuristics (rules of thumb) to estimate frequencies. This could result in certain biases. People use different strategies to answer frequency questions over different time spans. Episodic enumeration (recalling and counting individual incidents) is more likely to be used when shorter time spans are involved in frequency reports, whereas rate-based estimation (projecting the typical rate over the length of the recall period) is more likely to be used when longer time spans are involved. Reported behavioral frequencies for a year are generally less than 12 times the equivalent frequencies for a month. This phenomenon has been coined as the *time span effect*. Use of different time spans over or within studies may lead to inconsistent or even misleading results.

Beside memory-based information, respondents may use context-based information to make frequency judgments. One context-based heuristic involves the use of information provided by response alternatives. For example, respondents may use the range of the response alternatives provided to them as frame of reference in estimating their own behavioral frequency.

Accurate and reliable measurements of health behaviors, especially drug usage and sexual activity patterns, have proven difficult to obtain because of social desirability influences. People's tendency to under-report smoking and to underestimate their alcohol consumption are two well-established facts. In addition, studies that focus on behavioral frequencies consistently yielded *illusory superiority*: people report lower frequencies of unhealthy behaviors and higher frequencies of healthy behaviors for themselves than for an average peer.

6. *Conclusion*

If high accuracy is needed, health behavior should be assessed by physiological measures or (if these are not available) by behavioral observation, EMA, or diary techniques. However, these assessment strategies are resource consuming and often not practicable for researchers in the field studying large samples or complex, habitual health behaviors. Although self-reports may have a decreased validity and reliability in comparison with these methods, due to memory or motivational biases, they can nevertheless provide very useful information for scientific studies of the consequences of health behaviors. To ascertain the accuracy of self-reports, researchers should consider different possible connotations of questions in different samples, and construct questions carefully in order to reduce autobiographical memory biases and social desirability influences.

See also: Adolescent Health and Health Behaviors; Health Behavior: Psychosocial Theories; Health Behaviors; Health Risk Appraisal and Optimistic Bias; Illness Behavior and Care Seeking

Bibliography

Morgan D L 1997 *Focus Groups as Qualitative Research*, 2nd edn. Sage, Thousand Oaks, CA

Pereira M A, FitzerGerald S J, Gregg E W, Joswiak M L, Ryan W J, Suminsk R R, Utter A C, Zmuda J M 1997 A collection of physical activity questionnaires for health-related research.

Medicine and Science in Sports and Exercise **29**(6 Suppl.): S1–205
Rothman A J, Salovey P, Antone C, Keough K, Drake M C 1993 The influence of message framing on intentions to perform health behaviors. *Journal of Experimental and Social Psychology* **29**: 408–33
Sobell L C, Sobell M B 1995 Alcohol consumption measures. In: Allen J P, Columbus M (eds.) *Assessing Alcohol Problems: A Guide for Clinicians and Researchers*. NIAAA Treatment Handbook Series 4. National Institutes of Health, Bethesda, MD, pp. 55–73
Stone A A, Shiffman S 1994 Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine* **16**: 199–202

B. Renner

Health Care Delivery Services

Delivery systems for health care services are nested within health care systems, which are in turn nested within health systems. In fact, the topic of this article is best defined with reference to those more macrolevel systems, which are themselves addressed by other articles in the Health section. *Health care delivery services* consist of that segment of any society's *health system* which is specifically focused on the delivery of health care services. A *health system* encompasses a complex set of structural relationships among populations and institutions that have an impact on health, broadly defined. Further, as described by Ruggie (see *Health Care Systems: Comparative*), each country's *health care system* is defined as those specific macrostructures focused on financing, delivering, and evaluating health care, all of which are contained within a singular context of culture, history, politics, and economics.

This article will focus on defining and elaborating the dimensions of *health care delivery services* in a manner that is generally consistent with the literature on health services research, and comparable to discussions of 'health care delivery systems' in the United States and most developed nations. Commonly accepted definitions of health care delivery system emphasize the formal arrangements constructed to deliver health care services to specific populations, identified either by market, geographic location, type of service, or subpopulation characteristics (such as age or gender). Thus, this is a level of structure that is slightly more narrow than Ruggie's treatment of 'health care systems,' and represents a specific substructure within Frenk and Murray's definition of a 'health system.' Further, there are multiple types of health care delivery systems within any health system, depending on the type of health care service being delivered, whether preventative, diagnostic, or therapeutic, and the type of population or subpopulation being served.

This article will explore, in order: (a) the principle actors or components within health care delivery systems; (b) the variety of system types often present in highly developed health systems; (c) major characteristics of contemporary health care delivery systems; and (d) strategies for assessing health care delivery system performance. A number of recent trends in the health sector and among health care organizations will be examined, particularly since they highlight the difficulties in assessing delivery system performance in the twenty-first century, using standard measures of performance, most of which were developed during simpler times.

Given the brevity of this article, bibliographic emphasis will focus largely (although not exclusively) on US health care systems. Although the concepts to be explored in this article are applicable across most developed countries, the enormity of changes unfolding in many countries would require several chapters of dedicated comparative analysis. For example, economic and political turmoil in Eastern Europe is currently wreaking havoc with fragile public health and health care delivery systems. A prime example is Russia, where economic resources and basic medical supplies are scarce, government promises to cover basic health care go unfulfilled, and physicians must barter and beg to acquire needed medical equipment and drugs (Shishkin 1999). An 'underground' medical system has developed providing care for those few who can pay privately. Massive change also characterizes Latin American health care systems. In Chile free market reforms adopted during the Pinochet regime privatized much of what had been an extensive and effective government supported health care system. What remains now is a nonsystem with considerable problems of access (Reichard 1996). For more extensive comparative discussions of contemporary changes in health care systems, the reader is encouraged to consult recent work by Field (1989), Light (1999), or Day et al. (1998).

1. *Components of Health Care Delivery Systems*

The primary actors involved in delivering health care services are the members of *the health care workforce* (physicians and non-MD professionals, technicians, paraprofessionals; see *Health Professionals, Allied*), various *health care organizations* where care is either delivered or coordinated, and the agencies or firms which *finance* the delivery of services. Throughout the last half of the twentieth century, the medical and health division of labor has expanded dramatically, an expansion that has been fueled by the development of sophisticated medical technologies and specializations. Similarly, a variety of health care organizations deliver or coordinate care services along every step of the care continuum, and the number and type of

provider organizations mushroomed during the 1980s and 1990s. Diversification of health care organizational forms has both produced new organizational types (ambulatory surgery centers, women's health centers) and led to the adoption of new product lines by traditional organizational forms. For example, hospitals have diversified by adding a component of geriatric care, or vertically integrated into a comprehensive health care system, complete with ambulatory clinics and long-term care units (Fennell and Alexander 1993).

The *agencies and firms which finance health care service delivery* are also part of this system. In the US, a mix of actors, from both public and private sectors, play this role. Insurance carriers and the federal government have historically been important financiers of health care; more recently private sector corporations have been exerting considerable control over service delivery as major purchasers of health care for their employees. Toward the end of the 1990s the US health system came to embrace the concept of 'managed care,' an arrangement in which the financial partner (the health plan) becomes directly linked to both the providers of care and the client, through risk contracts. These arrangements have obfuscated the roles of the providers of care and the third-party payers of care, health insurance carriers or plans (Scott et al. 2000). Managed care arrangements contractually link providers to health plans, ensuring that an enrolled group of patients will seek care only from contracted (and somewhat controlled) providers. The providers are often paid a preset per capita rate for treating enrolled patients, and agree to 'gate-keep' or control the patient's use of services through utilization management strategies (see *Managed Care*). Large employers seek contracts with health plans that can promise low premiums and a variety of covered services; health plans compete for large employee groups to enroll as they are the most preferred, low-cost, healthy patient panels. Variations on this general strategy have resulted in a range of complex managed care organizational structures, in which the health plan is mediator between patients and providers, and between providers and payment. Managed care organizations combine provider and plan functions into one structure (Flood et al. 1998).

Although health care workers, organizations, and financing agencies constitute the major actors of health care delivery systems, there are a number of external factors which exert very important influences on these actors. Some influences are so important that it is sometimes difficult to separate the system actors from the external influences. Medical technology, for example, is both part of the environment of health care delivery and part of the health care organizations delivering the care. Continued expansion of medical diagnostic and treatment equipment, medications, biological agents, and devices has driven medical costs upward. The new technologies have created new types of technical and patient care jobs and professions within the health delivery system, which in turn lead to more demand for those technologies (Osterweis 1996). Both medical professionals and patients still expect new technologies to be available, despite their certain impact on costs, and their sometimes uncertain effect on outcomes. The availability of new technologies is used as a primary marketing device by health care organizations to attract patients, professionals, and healthplan contracts.

Several other environmental forces influence the major delivery system actors, including changes in the age structure of the population, shifts from acute to chronic illnesses as the leading causes of morbidity and mortality, the explosion of information and telecommunications technologies, and the role of the state as both purchaser and regulator of health services delivery. Each of these factors has shaped delivery system structures and performances, and interact with system structures and processes so as to change expected outcomes and performances (Fennell and Flood 1998). Although a full discussion of their impact is beyond the scope of this article, the reader is encouraged to consult recent work by Scott and coauthors for a comprehensive case study of change in one local health care system. Using an institutional theory perspective, these authors examine both the material resources (technologies, population characteristics, economic shifts, market characteristics) and the institutional structures and expectations surrounding health care (regulatory changes, norms about care delivery, cultural beliefs) in the San Francisco Bay Area from 1945 through 1995. They conclude that material resources and institutional environments have interacted to change the local health system over time, with different factors more or less dominant during the historical eras of professional dominance, federal involvement in health systems, and the current era of market-oriented delivery. Scott et al. (2000) remind us that it is difficult to draw boundaries around systems in the process of change, as one set of changes begets other changes, which then churn into new combinations of forces, counterforces, and evolving structures (see also Light's discussion of countervailing powers, 1993). Further, change at the local level is influenced by changes in national policy as well as changes in social and cultural norms.

2. *Varieties of Health Care Delivery Systems*

Within any given health system there can be a variety of delivery systems, identified by the type of health service being delivered. For example, care systems can be defined by where in the continuum of care the service is located, such as primary care, secondary, tertiary, rehabilitative, and/or custodial care. Two frequently made dichotomous distinctions, based on the general type of care, are acute vs. chronic

care and mental vs. physical health systems. Additionally, care systems have been described on the basis of *where* that care is provided, and by which actors. Categorization by location used to be a fairly reliable and straightforward method. For example, complex surgical care was once always provided in the hospital. Ambulatory care meant that care happened in clinics and doctors' offices. Long-term care always occurred in nursing homes. Hospice care at the end of life was most often provided in freestanding hospice centers or by a home care team working with the family (Watt 1996).

More recently, however, those clear boundaries of care location have faded out of focus, and the sectors of care often blend into each other. What used to be considered inpatient care is now routinely delivered on an outpatient basis (many surgical procedures, for instance, and their adjuvant treatments), and significant rehabilitative care following an acute hospital episode is delegated to either the skilled nursing facility, the rehabilitation facility, or home with nursing visits. Early innovations in hospice care as an alternative setting for end of life care, which moved the dying patient away from the technically-oriented hospital setting, first developed in the United Kingdom in the late 1960s, and then spread to the US in the 1970s (Boling and Lynn 1998, Keay and Schonwetter 1998). More recently in the US, hospice services are increasingly provided by special units embedded within hospitals or nursing homes. Efforts to evaluate the quality of hospice care within such a broad range of care settings are, of course, made more complex given the inevitable clash between the goals of acute care settings and end-of-life care (Mor 1987).

Similarly, the traditional definitions of acute, subacute, or postacute care services can no longer be made on the basis of care setting. Hospitals are increasingly providing more long-term care related services, and nursing homes have been providing care for more acute and subacute patients. It has even become more difficult to clearly differentiate the type of care services provided in particular care settings on even the most basic of distinctions: medical vs. mental health care. The prevalence of depression, for example, among older patients, whether located in primary care settings, inpatient settings, or long-term care settings has come under closer scrutiny. Some estimates suggest that up to 65 percent of elderly patients in nursing homes suffer from depression. Whether those institutionalized elderly actually receive appropriate mental health diagnoses or mental health care, however, is not at all certain.

Diversified health systems market their superiority over other health care providers, of course, on the basis of the 'seamless delivery system,' where all needed services are available in one location or from one large, multiunit provider (Shortell et al. 1996). Theoretically, the diversified system should be able to ensure better coordination of care and a much smoother transferal of information from one division to another. But whether or not this type of delivery system structure actually results in net gains for care quality or continuity has yet to be demonstrated in any conclusive fashion. In fact, whether diversified delivery systems actually result in shared or compatible patient record systems, thus, assuring coordination of information about the patient within the system, still needs to be examined. Recent work on the difficulties of guaranteeing security and privacy of electronic patient records revealed that the systems presumed to coordinate electronic information across health care organizations are much further behind in development than expected. When health care organizations merge or are acquired by corporate systems, the details of developing shared information systems through which patient care can be coordinated across care settings are only attended to as an afterthought or far later in the merger process (National Research Council 1997).

3. *Characteristics of Contemporary Health Care Delivery Systems*

Traditional studies of the structure and process of health care organizations have provided a model for the study of health care delivery systems. The basic *structural characteristics* of size, ownership, profit orientation, range of service provision, and primary governance structure provide a template for plotting variation in delivery systems. Similarly, *process characteristics* describe systems in terms of what is done to and for the patient over the course of treatment. Key variables of process include system coordination mechanisms, information flows, decision-making processes, quality control, and communication flows between practitioners and patients. In the past two decades, health care delivery systems have been characterized by three major trends, the first of which was mentioned earlier: (a) an increase in the diversification of structural types and products; (b) change in traditional ownership and management configurations; and (c) the development of new interorganizational patterns of care delivery and complex multitiered governance structures.

Changes in ownership and management patterns of delivery systems and their component organizations over the past two decades have drastically changed how we think about the delivery of health care services. There have been basically four major facets of ownership/management reconfiguration, affecting different health care organizations and systems in various ways: (a) an increased concentration of the acute care delivery system, through hospital closures and mergers; (b) a rise in the prevalence of corporate restructuring across both acute and long-term care sectors, and the spread of managed care philosophy and organizational structures; (c) a blurring of the once basic distinction between profit and not-for-

profit ownership (see *Health Care Organizations: For-profit and Nonprofit*); and (d) a drastic and pervasive change in the management 'climate' of most health care organizations, from an emphasis on nonprofit service and community responsibility to an emphasis on competitive business strategies. These trends have been discussed at length elsewhere (see Bibliography).

The development of new interorganizational patterns of care delivery and more complicated governance structures is best exemplified by the growth in multihospital systems, loosely structured provider networks or coalitions, and formally structured ownership arrangements linking hospitals, physician groups, health plans, and other care settings (best exemplified by the proliferation of managed care organizations, as discussed above). Both horizontally and vertically integrated forms have made substantial inroads in most health care markets, and they vary by size, composition, and level of integration. The governance structures that have developed along with these complicated multiunit structures are generally described as business-oriented, no matter what their profit status. Governing boards within integrated delivery systems often take a two-level form: the separate units (such as a hospital) may retain an internal board, with at least some members drawn from the community. But these local boards tend to focus only on community fund raising and local public relations. Significant strategic planning and investment decisions are usually under the control of the 'corporate' board of the system.

4. *Assessing Health Care Delivery System Performance*

Two criteria for evaluating the performance of health care delivery systems and the organizations within them have dominated the health services literature: the quality of care received by patients (including their satisfaction with services) and the cost efficiency with which health care services are delivered. Quality of care can be evaluated by examining actual care outcomes (changes in health or functioning) or by benchmarking, that is, comparing actual performance or structures to state-of-the-art standards. Evaluations of care delivery focused on processes or structures assume, of course, that an excellent structure in place enhances the likelihood of high-quality processes being performed and, consequently, the likelihood of better outcomes occurring. These assumptions therefore rest on correctly identifying state-of-the-art standards or validating that their use does in fact lead to improved outcomes.

Fueled by physician profiles and public report cards, the 1990s have brought about an increased sophistication in measuring outcomes. Most health services researchers are also careful to control for the myriad of patient-level differences that can affect appropriate accountability for care delivered within a single organization within the delivery system, such as adjusting for patient health factors while assessing outcomes following surgical care. Very sophisticated measures attempt to include posthospital outcomes to assess hospital care. Nonetheless, most of these outcome measures have been designed to focus on health statuses typical of young and middle-aged patients, and rarely do they focus on the most commonly seen age group of health service users—the elderly. Outcome measures are also most likely to be built around acute medical problems, and rarely are they modeled on the health problems of the elderly or of the chronically ill, who typically have several concurrent, complex, interrelated chronic diseases with the potential for acute episodes. Nor are they able to measure accurately the outcomes for people with interrelated mental and physical health problems, people who are facing end-of-life scenarios, or people for whom the goals of care delivery need to emphasize pain relief, palliation, or functional maintenance (Lamers 1990).

4.1 *Problems with Outcome Measure Validity*

Indeed, the convergence of several recent trends in health care delivery structures and environments (discussed briefly above) are of particular concern when assessments of service delivery system performance are considered: the graying population of most developed countries, shifts in the prevalence of acute and chronic diseases, the proliferation of new interorganizational patterns of care delivery, and the spread of managed care. A variety of methodological and design issues need to be addressed in order to properly model and evaluate health care delivery system performance under this confluence of conditions. As an example, consider the situation of terminally ill patients: what outcomes should be used to evaluate hospices, which are designed to provide psychological, social, and spiritual care for dying persons and their families? Evaluations in the past have typically focused on the costs of care and the range of services provided (Mor 1987). Quality of hospice care is usually defined in terms of its effect on the quality of life for the patient receiving hospice services and is often 'customized' to reflect differences in each state's goals for the hospice program. But what truly constitutes a 'positive outcome' for dying patients? Whose care needs to be assessed—should we also consider the care of the dying patient's family? How should 'quality of life' be modeled? What is the set of 'costs' that needs to be included in evaluating the cost-effectiveness of hospices themselves or hospices compared to other sites where people die?

In addition to these issues regarding the proper 'outcomes' to measure, terminally ill patients receive their care from a wide range of providers and in a variety of settings. Despite the typical complexity of

care for any given patient, evaluations of the performance of health care delivery systems almost always have focused on a particular portion of that care attributable to a single provider or organization. Common examples include the Resident Assessment Instrument (RAI) for nursing homes (Morris et al. 1990), accreditation standards for hospitals, and NCQA standards for managed care organizations (Zinn and Mor 1998). How well, in fact, have these measures identified the portion of care that can be attributed to a single provider, when the true process of health care delivery involves multiple actors?

What is needed most is a reliable and valid measure of continuity of care, which should take into account the extent to which care is coordinated across care settings, care providers, and transitions between providers and settings. Even outcome measures based upon an illness episode (rather than a specific and standard unit of time, or service consumption event) do not necessarily capture care outside of a particular setting, such as when a hospitalization is the unit of analysis for an episode.

4.2 Problems in Defining the Unit/Level of Analysis for System Performance

Although there are obviously many possibilities for defining levels of systems of care delivery, it is conventional to identify at least three: (a) a single health care organization; (b) a larger, socially defined unit that contains the health care organization, such as a community or a system of health care organizations; and (c) the subunits contained within the organization, such as individual departments or practitioners. Researchers examining organizational issues typically inspect only one facet of care delivery at a time, or one level of analysis. All complex systems, however, tend to be nested units, that is, systems within systems within systems. While most analyses of delivery system performance attempt to bypass these issues by focusing on one or more of the standard three levels, they can instead fail to identify fully the processes and outcomes they are intended to study. The critical point is not that all nested levels must always be studied, but that one should be as clear as possible in identifying the level of analysis selected and about ensuring that the choice of level is aligned with the characteristics of the type of care delivered, the subpopulation of patients, and the goals of the delivery system under examination.

It is also important to recognize that system performance at any given level may not be analyzable as a simple aggregation of performance at lower levels. This is one of the principal features of any system: its performance is determined as much (if not more) by the arrangement of its parts, the types of coordination and integration achieved across units, as by the performance of the individual components. Specifying the appropriate levels within the model, however, is not a simple matter, since multiple contexts can influence patient outcomes simultaneously and those effects may be difficult to distinguish due to nesting or overlapping. For example, a physician's clinical practice behavior can be nested within his/her clinic's setting, which itself may be nested within a particular diversified health care system, and a particular region or market with characteristic practice patterns.

5. Conclusion

Continuity of care, or clinical integration, has been recommended by Shortell and other researchers as the model on which managed care should be based, so that providers are explicitly recognized as able to bridge care settings along with the patient, so that information is coordinated and shared explicitly across care setting transitions, and so that the care manager may change with these transitions, even if the same actors are involved. Managed care at its best would embrace such a model, and by creating effective mechanisms such as case managers, real integration and coordination would result, presumably leading to better delivery outcomes. Similarly, the multi-institutional, multiprovider, diversified health care delivery system should, on the face of it, make the sharing of information a matter of routine, through uniform patient records and centralized record systems across care settings and providers, thus improving care continuity and a variety of outcomes.

This is the basis of the 'seamless delivery system.' But whether or not such changes in care structures actually result in net gains for care quality or continuity has itself yet to be demonstrated in any conclusive fashion. Understanding the complexity of service delivery systems, and the other systems within which they are embedded, will not simplify this task. But it will lead to better measurement and modeling, and thus, perhaps, to a more realistic assessment of what is needed to achieve 'seamless' system functioning.

See also: Caregiver Burden; Caregiving in Old Age; Gender and Health Care; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care Organizations: For-profit and Nonprofit; Health Care, Rationing of; Health Care Systems: Comparative; Health Care Technology; Illness Behavior and Care Seeking; Managed Care; Minority Access to Health Services: United States

Bibliography

Boling A, Lynn J 1998 Hospice: Current practice, future possibilities. *Hospice Journal* **13**(1–2): 29–32

Day P, Klein R, Miller F 1998 *Hurdles and Levers: A Comparative US–UK Study of Guidelines*. Nuffield Trust, London
Fennell M, Alexander J 1993 Perspectives on organizational change in the US medical care sector. *Annual Review of Sociology* **19**: 89–112
Fennell M L, Flood A B 1998 Key challenges in studying organizational issues in the delivery of health care to older Americans. *Health Services Research* **33**(2): 424–33
Field M G (ed.) 1989 *Success and Crisis in National Health Systems: A Comparative Approach*. Routledge, New York
Flood A B, Fremont A M, Jin K, Bott D M, Ding J, Parker R C Jr 1998 How do HMO's achieve savings? The effectiveness of one organization's strategies. *Health Services Research* **33**(1): 79–99
Friedson E 1994 *Professionalism Reborn: Theory, Prophecy and Policy*. University of Chicago Press, Chicago
Gold M R (ed.) 1998 *Contemporary Managed Care: Readings in Structure, Operations, and Public Policy*. Health Administration Press, Chicago
Gray B H 1999 The changing face of health care. In: Clotfleter C F, Ehrlich T (eds.) *Philanthropy and the Nonprofit Sector in a Changing America*. Indiana University Press, Bloomington, IN
Keay T J, Schonwetter R S 1998 Hospice care in the nursing home. *American Family Physician* **57**(3): 491–4
Lamers W M Jr 1990 Hospice: Enhancing the quality of life. *Oncology-Huntington* **4**(5): 121–6
Light D 1993 Countervailing Power: The changing character of the medical professions in the United States. In: Hafferty F W, McKinlay J B (eds.) *The Changing Medical Profession: An International Perspective*. Oxford University Press, Oxford
Light D 1999 *Future Challenges for the NHS: An International Perspective on the 50th Anniversary*. Nuffield Trust, London
Mor V 1987 *Hospice Care Systems: Structure, Process, Costs, and Outcomes*. Springer, New York
Morris J N, Hawes C, Fries B E, Phillips C D, Mor V, Katz S, Murphy K, Drugovich M L, Friedlob A S 1990 Designing the National Resident Assessment Instrument for nursing homes. *Gerontologist* **30**: 293–307
National Research Council: Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure 1997. *For the Record: Protecting Electronic Health Information*. National Academy Press, Washington, DC
Osterweis M, McLaughlin C J, Manasse H R, Hopper C L (eds.) 1996 *The U.S. Health Workforce: Power, Politics, and Policy*. Association of Academic Health Centers, Washington, DC
Reichard S 1996 Ideology drives health care reforms in Chile. *Journal of Public Health Policy* **17**(1): 80–98
Shishkin S 1999 Problems of transition from tax-based system of health care finance to mandatory health insurance model in Russia. *Croatian Medical Journal* **40**(2): 195–201
Scott W R, Ruef M, Mendel P J, Caronna C A (eds.) 2000 *Institutional Change and Health care Organizations: From Professional Dominance to Managed Care*. University of Chicago Press, Chicago
Shortell S M, Gillies R R, Anderson D A, Erickson K M, Mitchell J B 1996 *Remaking Health Care in America: Building Organized Delivery Systems*. Jossey-Bass, San Francisco
Watt K 1996 Hospice and the elderly: a changing perspective. *American Journal of Hospital Palliative Care* **13**(6): 47–8
Wennberg J (ed.) 1998 *The Dartmouth Atlas of Health Care in the United States*. American Hospital Association, Chicago
Wholey D R, Burns L R 2000 Tides of change: the evolution of managed care in the United States. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ
Zinn J S, Mor V 1998 Organizational structure and the delivery of primary care to older Americans. *Health Services Research* **33**(2): 354–80

M. L. Fennell

Health Care: Legal Issues

After years of relative stability, the relationship between law and health care in the USA is changing rapidly. These changes are both a response to, and a reflection of the transition to, a market-based health care system. How the courts and legislatures respond to these changes will, in part, determine the success of managed care's cost containment initiatives. This article broadly examines the relationship between law and health care, with a particular focus on how the courts have responded to challenges to managed care's cost containment innovations.

1. *Law and Health Care: Framing the Issues*

It should not be surprising that two of the USA's most important institutions, the legal and health care systems, are intertwined and intersect at various points. Since the early 1800s, medical jurisprudence, where physicians are called as expert witnesses and where the courts act to regulate the quality of medical care, has been an integral part of the growth of the health care enterprise. As Mohr noted, '… medical jurisprudence … figured prominently in early American medical history …; neither subject can be fully understood in its modern context without taking the history of legal medicine into account' (Mohr 1993, p. 251). As this quote indicates, the involvement of the courts in health care delivery is not new. Yet the relationship between law and health care has never been more visible or manifest than it is today.

Despite the inevitable ebbs and flows of the relationship between law and medicine, it is arguable that the relationship was essentially stable from the mid-nineteenth century until the mid-1960s (Stryker 1932). As recently as the late 1960s and early 1970s, health care attorneys and health care executives had a relatively easy set of legal rules to live with. Both sides could predict the nature and scope of the litigation since the essential rules establishing the standard of care and the types of litigation initiated varied little during this period. A typical court case involved one patient suing one physician, guided by liability rules that reflected judicial deference to physicians in setting the standard of care. That relative stability was largely shattered by the advent of managed care. At both the state and Federal levels, all three branches of the legal

system, the legislatures, regulatory agencies, and courts, have persistently been involved in setting health care policy and monitoring health care delivery since the mid-1960s. The framing of the legal issues in the managed care era thus reflects the dramatic changes engulfing the health care system since the mid-1980s and the resulting policy conflicts that have emerged.

1.1 The Litigation Context

For many years, courts have actively monitored quality of care through determining the standard of care for medical liability (Jacobson 1989, Furrow 1997) and interpreting contractual definitions of medical necessity (Eddy 1996, Hall and Anderson 1992). Courts have also been called upon to interpret the vast array of state and federal legislation regulating health care delivery.

In responding to litigation challenging managed care cost containment practices, judges face a threshold question of what role courts should play in monitoring the managed care environment relative to legislatures. Courts face the tension of applying common law liability principles to new situations (perhaps by depicting public policy considerations) or deferring to elected representatives to set public policy. In the new health care environment, courts will be asked to distinguish the economic aspects of managed care that order relationships between health plans, physicians, and patients, from the incentives that contribute to below-standard care. The former are more traditionally legislative prerogatives while the latter are traditionally within the judicial purview. For those cases resolved in the courts, judges will need to decide whether they should be resolved by contract or tort (civil wrongs such as negligence) law. The difference is that recovery under contract is more difficult and damage awards are lower. What complicates matters is that in managed care, the insurance (i.e., contract) and medical care (i.e., potential negligence) are combined in one entity.

1.2 Liability

Common law liability doctrine has traditionally emerged incrementally to adapt to changing circumstances. Thus, it will take some time before the actual contours of liability can be assessed. Nevertheless, certain themes have emerged in the litigation so far.

Despite variation in the state cases decided to date, the direction appears to be that courts are applying traditional liability doctrine to the new organizational forms, including health maintenance organizations (HMOs), independent practice associations (IPAs), and preferred provider organizations (PPOs) (Furrow 1997). For staff or group model HMOs, the law had already developed holding them liable for torts (civil wrongs) committed by their employees. The primary issue is whether courts will apply indirect liability doctrine (i.e., vicarious liability and agency principles) to the other MCO organizational forms for malpractice committed by an independent physician.

In determining vicarious liability, the courts look to the MCO's control over the physician or how the plan markets its physicians. The greater the indications of control and the more the plan markets the quality of its physician panel, the greater the likelihood that the MCO will be held vicariously liable. In the leading case of *Boyd v. Albert Einstein Medical Center*, 547 A.2d 1229 (Pa. Super. 1988), for example, the court held that an IPA could be found liable for an individual physician's medical malpractice because the patient reasonably believed that the IPA controlled the physician. The court rejected the plan's argument that it was not liable because the physician was an independent contractor, holding that the physician could still be the plan's agent with respect to the patient. A key element is whether the MCO has sufficient control, such as through utilization management arrangements, to override a physician's clinical decision.

1.3 Challenges to Cost Containment Practices

The changes in the health care environment have inevitably resulted in a new set of issues to be litigated, particularly the conflict between population-based cost containment and access for individual subscribers (sometimes to experimental procedures), the multiplicity of actors in a given case, and the evolving nature of the organizational structures. Just as the courts had to adapt medical liability principles to the emergence of hospitals as central to health care delivery in the 1960s, the same process is now underway with regard to the emerging organizational forms comprising managed care. There is inevitably a learning curve as courts become educated about the underlying changes and begin to adopt principles derived for a different model to the new arrangements (Jacobson and Pomfret 1999). For example, instead of the simple model of litigation involving one physician and one patient, managed care litigation may include the patient, the physician, the health plan, the utilization review firm, and the administrator of an ERISA-covered plan. Sorting out which party is legally responsible for delayed or denied care and the acceptable level of care resulting from cost containment strategies remains a work in progress.

Nonetheless, the cases decided so far suggest that courts are not systematically impeding the implementation of cost containment initiatives (Jacobson 1999). While it may be that many of the most challenging cost containment innovations have yet to be litigated, courts seem willing to uphold cost containment programs, except, perhaps, if such programs are seen as

limiting access to life-sustaining treatment. For example, courts have generally rejected challenges to financial incentives (though some courts have allowed such challenges to proceed to trial), utilization management programs, and many contractual challenges to benefit denials.

Few courts seem willing to usurp legislative choices in formulating health policy or to obstruct the market in organizing and delivering health care services, with several courts specifically deflecting allegations that managed care incentives violate public policy to the legislatures. One reason for this is that courts have interpreted the Employee Retirement Income Security Act (ERISA) narrowly to block state court challenges to cost containment programs that deny benefit coverage. Another reason why cost containment initiatives have not been overturned is that courts are looking to contract law to determine the extent of the parties' obligations and responsibilities, especially in physicians' litigation against MCOs. Though the trend toward contract is neither unlimited nor uniform, it represents a potentially profound shift from the early 1970s, where courts were perceived as protectors of individuals, to courts as deferring to market-based changes.

1.4 Physician Autonomy

The initial litigation trends suggest that courts will not subordinate cost containment programs to physician autonomy. Courts are ruling that people who are dissatisfied with managed care, including physicians, should take their case to the legislatures. This represents a dramatic departure from the era when courts deferred to physician dominance.

It would be understandable if physicians expected that courts would protect the physician–patient relationship from interference by managed care plans. After all, the essence of the health care system lies with the physician–patient relationship, and courts have protected physician dominance in the past (Kapp 1985, Hall 1988). But courts have not done so in the managed care era. To the extent that physicians have been able to secure some gains through the courts, it has largely been through fair process protections and retaining the often-maligned legal standard of care.

The question is why courts are not deferring to physicians; what has changed over the past decade or so to cause the courts' shift from protecting physician dominance to reinforcing managed care's dominance? At this point, the answer is somewhat speculative because of the limited number of cases. One potential answer is that courts respond to changes in the underlying environment. As the nature of health care delivery has changed, it stands to reason that courts would begin to look at and incorporate those changes. Slowly but surely, courts are internalizing cost containment goals in developing legal doctrine, rather than ensuring the primacy of the physician–patient relationship. Since the dominant force in health care is institutional, it is not surprising that the legal dominance physicians maintained is now waning.

2. Future Directions and Challenges

During the next few years, the legal system faces four key challenges regarding health care: (a) how to allocate liability given the multiplicity of decision makers involved in managed care; (b) how to balance between the health care needs of individual patients and the cost containment goals of the health care system; (c) how to deal with the consequences of ERISA pre-emption, which now prevents state courts from resolving challenges to how cost containment goals are implemented; and (d) how can conflicting social policy goals (i.e., the cost/quality/access tradeoffs) be reconciled.

2.1 Allocation of Liability and ERISA Pre-emption

The allocation of liability and ERISA pre-emption issues are both about imposing accountability on a system that has managed to shift legal accountability to physicians while maintaining the fiction of non-control. Either voluntarily or through legislation, MCOs should be held accountable for their actions.

2.2 Balancing

There is no easy solution to balancing the individual patient's access to health care with conserving the plan's assets for the entire patient population. One approach might be to articulate a set of fiduciary obligations that a plan must adhere to in balancing the two. The burden would be on the plan to justify denial of care to the individual patient by a preponderance of the scientific evidence and to demonstrate that the individual's medical care is not being subordinated to the patient population (Jacobson and Cahill 2000).

2.3 Social Policy Conflicts

The plethora of legislation has placed health care executives in somewhat of a bind. Some regulations, such as fraud and abuse, are specifically designed to reduce health care costs. Others, such as EMTALA, are designed to improve access regardless of cost. Taken alone, each individual law is defensible. Taken together, however, the social policy goals of cost containment, improving access to health care, and improving quality of care conflict across statutes and

regulations. Complying with all of the laws may be impossible, and may reflect incoherent social policy.

3. Implications for Other Countries

While other countries may not, for domestic social policy reasons, want to follow the managed care model, it seems likely that most countries will face the need to impose cost constraints on their national health care systems. These choices are already evident in the British and Canadian systems. What this brief survey of US health law trends suggests is that reconciling cost containment with various institutional objectives and interests will take careful planning to ensure that the changes are implemented fairly and equitably.

See also: Doctor–Patient Interaction: Psychosocial Aspects; Health Care Markets: Theory and Practice; Health Care Organizations: For-profit and Nonprofit; Health Care, Rationing of; Liability: Legal; Litigation; Litigation: Effects on Medical Research; Psychotherapy: Legal Issues (United States)

Bibliography

Eddy D M 1996 Benefit language: Criteria that will improve quality while reducing costs. *Journal of the American Medical Association* **275**: 650–57

Furrow B R 1997 Managed care organizations and patient injury: Rethinking liability. *Georgia Law Review* **31**: 419–509

Hall M A 1988 Institutional control of physician behavior: Legal barriers to health care cost containment. *University of Pennsylvania Law Review* **137**: 431–536

Hall M A, Anderson G F 1992 Health insurers' assessment of medical necessity. *University of Pennsylvania Law Review* **140**: 1637–712

Jacobson P D 1989 Medical malpractice and the tort system. *Journal of the American Medical Association* **262**: 2230–7

Jacobson P D 1999 Legal challenges to managed care cost containment programs: An initial assessment. *Health Affairs* **18**(4): 69–85

Jacobson P D, Cahill M T 2000 Redefining fiduciary responsibilities in the managed care context. *American Journal of Law and Medicine* **26**(2 & 3): 155–74

Jacobson P D, Pomfret S D 1999 Establishing new legal doctrine: A model of judicial response to industrial change. *University of Michigan Journal of Law Reform*

Kapp M B 1985 Medicine and law: A symbiotic relationship? *American Journal of Medicine* **78**: 903–7

Mohr J C 1993 *Doctors and the Law: Medical Jurisprudence in Nineteenth-century America*. Oxford University Press, New York

Rosenbaum S, Frankford D M, Moore B, Borzi P 1999 Who should determine when health care is medically necessary? *New England Journal of Medicine* **340**: 229–332

Stryker L P 1932 *Courts and Doctors*. Macmillan, New York

P. D. Jacobson

Health Care Markets: Theory and Practice

This article describes the origins, development, and proliferation of health care markets. It then reviews their performance, summarizes recent theory and research on their limitations, and concludes by suggesting new ways to conceptualize and research economic behavior in health care.

1. Origins

'Health care markets' is a relatively new term, framed by economical, structural, and social forces. It hardly existed 25 years ago but rose to prominence in response to the years of recession and high inflation ('stagflation') in the 1970s. The costs of social and welfare programs, such as health care, kept rising as revenues stagnated, prompting a 'crisis in the welfare state' (Esping-Andersen 1994). In addition, there were increasing complaints about government-run programs and about hospital-based specialists running up large bills without regard to the needs of payers or the population (Bengoa and Hunter 1991). In response, health economists recommended that payers (governments, employers, and insurers) apply the well-established principles and techniques of price competition theory to health care in order to constrain costs, increase efficiency, and make providers more responsive to the priorities of payers. This constituted a historical shift from professional dominance to buyer dominance among the countervailing powers of providers, purchasers, patients, and the state, over the control of power, costs, information, and the distribution of risks (Light 1995a).

1.1 Problems of Market Failure

Applying price competition theory to health care was unprecedented, for medicine had long been regarded by leading economists as lacking the basic qualities for markets (Arrow 1963, Fox 1979). The idea of health care as a commodity to be traded in a market was anathema to them. Clinical medicine is rife with uncertainties and too often contingent on diagnosis and response. Thus 'products' often cannot be defined, and property rights may be unclear. Physicians vary substantially in how they treat the same problem. Good market information is poor and highly asymmetrical. The absence of these requirements for beneficial competition can mean that competition is harmful to patients and society.

1.2 Managed Competition as a Possible Solution

These serious misfits between the attributes of clinical medicine and the requirements for competition led

Enthoven 1988) to develop the theory of managed competition. To overcome these forms of market failure, Enthoven emphasized competition between comprehensive health care plans and stipulated several requirements for a fair and beneficial market, such as universal health insurance or access, coverage of a common set of benefits, risk-adjusted contracts and good comparative information on quality, efficiency, and costs. He also held that health care markets had to be actively managed, because providers had proved so clever in manipulating or circumventing rules. This he called 'managed care' rather than 'regulated competition.'

Managed competition won critical acclaim and spawned international efforts to proselytize it as the 'only practical solution' to soaring costs, though the model had serious flaws. It assumed that some people (managers) can be trusted and others (physicians) cannot; it did not solve the problems of uncertainty or contingency but rather tucked them inside the health plans; and it led to oligopolies, which minimize price competition (Light 1995b). New businesses and a new industry developed to create health care markets and varieties of managed competition by monitoring services and bills, changing clinical behavior, and organizing providers into 'preferred provider organizations' and 'network HMOs.'

1.3 Proliferation and Variations

A signal event occurred when Prime Minister Margaret Thatcher decided in 1989 to transform the world's largest health care organization, the National Health Service, from an administered service to a market model of competitive contracts between purchasers and providers. Other governments, notably New Zealand and the Netherlands, followed suit. By the mid-1990s, however, most had pulled back from implementing markets in health care, as they came to realize that price competition could cause service dislocations, cost shifting, increased inequality, reduced choice, reduced care, and increased costs over government-run programs (Light 1997, *Social Science & Medicine*, special issue 2001). Some countries such as the United States did little to prevent adverse selection, service dilution, and cost shifting. In the meantime, important European variations on American models developed such as 'planned markets' and 'quasi-markets' (LeGrand and Bartlett 1993, Saltman and von Otter 1992). Behind them was the realization that markets themselves are never 'free' but rather structured and regulated to reflect a society's values and judgments about the limits of competition.

2. Performance Assessment

It has proven difficult to assess the performance of health care markets and different competitive schemes because changes are introduced wholesale with no controls, or markets are introduced simultaneously with changes in coverage. American employers, for example, have simultaneously shifted costs to employees, increased forms of risk-rating and exclusions in policies, used policy churning and within-group underwriting, reduced the amount of service or reimbursement in areas like mental health, and limited access to specialty services and then credited the savings to 'competition.' Most stakeholders do not seem interested in well-functioning markets like the one developed by Enthoven. They want to save or make money, as the case may be. Reflecting a policy bandwagon more driven by conviction than scientific evaluation, the most systematic external analysis of research on US health care markets found that among hundreds of research articles, only a small percent systematically compared HMO performance to some alternate delivery model (Robinson and Steiner 1998). Evaluation of the market reforms in New Zealand also found the measures seriously flawed, efficiency no greater, health gain no greater, and fewer services leading to greater unmet need (Ashton 1999).

Operationally, the best tracking system in the United States reported that 'few private employers offer employees a choice of plans, give employees financial incentives to choose economical plans, or provide information so that employees can evaluate the quality of the care they receive—three hallmarks of managed competition' (Long and Marquis 1998, p. 1). The well-designed studies showed that managed care organizations reduced the high utilization rates and prices found in the United States, but there were no controls for better health status of enrollees which could explain lower utilization. Reinhardt (1996) summarized the American scene as employers hiring corporate health plans as 'bounty hunters' to reduce costs—no questions asked. Buyers have paid little attention to effects on quality. Zelman and Berenson (1998), two of the architects of managed care markets, concluded that the corporations took early large profits by selecting healthier and wealthier markets and patients, while compromising on quality. Now that most of the excess has been squeezed out of prices and utilization, it is unclear how corporate health plans will reward investors. Stung by widespread criticism and political backlash, competing plans have eased up on managing utilization to such a degree that they resemble the open choice of fee-for-service that originally prompted the development of health care markets and managed care organizations (Freudenheim 2000).

In other countries, Hsiao (1994) evaluated the impact of introducing health care markets in several Latin American and Asian systems and reported higher costs, reduced access, increased inequality, and profiteering. Evans' (1997, p. 427) international assessment concluded that '…greater reliance on the market is associated with inferior system perform-

ance—inequality, inefficiency, high cost, and public dissatisfaction…' Stocker et al. (1999) reported that market-based reforms in Latin America are driven largely by American health insurance or other corporations who have taken the early profits out of American markets and now seek high profits in new countries by buying government support. Meantime, backlash to health care markets in the UK and New Zealand led to ousting the ruling party that installed health care markets, and substantially dismantling the health care markets in order to foster cross-sector partnerships and develop community-based programs that integrate public health with clinical services. Competitive markets are regarded as inherently anathema to community and public health.

A central challenge posed by health care markets is that they are costly to start up and operate. Good market information, contracting, monitoring, and managing together cost about 10 percent of revenues, so that comparable savings on clinical services must be achieved just to stay even. Another problem, related to what Hsiao found in several other systems, is that proper markets are often not set up with the necessary information and safeguards against inefficient but profitable strategies.

2.1 Contributions

In response to criticisms, market-oriented research has made a number of contributions. Health care markets evolved earliest and most extensively in the United States, as employers self-insured and began to ask new questions such as, 'Which services are most cost-effective? Which providers have the lowest costs and prices?' To answer such questions, new statistical techniques, methods, and theory were developed to measure effectiveness and efficiency, identify low-cost providers, estimate risk adjustments, and compensate for imperfect information (Folland et al. 1997). Accountability, performance, consumer satisfaction, and quality have become much more visible and have led to serious efforts to measure them. Health care market thinking has spawned a new and wide range of strategies for delivering more cost-effective care, even if those strategies do not themselves entail much market-like behavior. The major reports from WHO-Europe and the London School of Economics, for example, describe new developments in priority setting, contracting, capital formation, performance-related pay, decentralized management structures, and consumer involvement that would not have occurred without the radical challenge posed by market-like thinking (Saltman et al. 1998, Mossialos and LeGrand 1999).

2.2 Limitations of Health Care Markets

Restructuring clinical services into markets of corporations bent on growth and profit is likely to drive up costs in the long run, abetted by the important natural advantages of sellers (providers). They diagnose, prescribe, and treat. They control the clinical record that determines charges. They also have the patient's loyalty, if not dependency. A second structural problem in many health care markets is that each insurer or plan contracts with most of the providers in a given market in order to offer subscribers the widest choice possible. But that means no real competition is possible, because each provider in effect 'competes' against himself. This practice can lead to the third limitation, that there are so many plans and policies that no one can identify or evaluate real differences. In a sample of 2,277 patients, Grembowski et al. (2000) found there were 189 different provider organizations and 755 different policies. Behind so many policies and plans are high costs for designing, marketing, and managing them, and such numbers far exceed what one finds in real competitive markets.

At a micro-, empirical level, Rice (1998) has summarized studies documenting several practical problems. First, the preferences of patients and employers are not fixed but are affected by prior experience, current interactions, and advertising. This poses 'a very big problem' for researcher and competitors as well as for economic theory, as do problems posed by uncertainty and contingency. Second, consumers' choices do not necessarily reveal their preferences and may be poor judges of them. Third, Rice observes, consumers have poor market information, and fourth, when they are given it, studies find they sometimes ignore the most relevant data. Finally, a majority of users are also unable to understand basic labels, instructions, and policies (Parker et al. 1999). What, then, actually takes places in health care markets, especially when 'consumers' do not consume or buy, even though standard texts characterize them that way (Follard et al. 1997, p. 21)?

3. Casting a Wider Research Net

Studies reviewed by Rice indicate there are variables not now being captured in economic models and research. These have been the focus of the burgeoning field of economic sociology, which attends to the influence of affiliations, networks, institutions, and culture, as well as to the role of power, hierarchy, and the past in shaping 'path-dependent' behavior (Smelser and Swedberg 1994). In a classic article, Granovetter (1985, p. 481) showed how economic behavior and institutions 'are so constrained by ongoing social relations that to construe them as independent is a grievous misunderstanding.' What passes for 'rational decision-making' differs substantially from one nation to another (Dobbin 1994) and even from one corporation to another.

Economic behavior begins with cognitive and cultural categories that get combined into scripts which actors routinely use in deciding what they value, buy, or sell (DiMaggio 1997). These scripts allow actors whose capacity for rationality is bounded by their ability to filter, sort, and apply what they know through 'selective attention' (Lindenberg 1998). The sociology of microeconomics, or actual market behavior closely observed, looks more like a negotiated construction of a reality that the parties can live with, even if front-stage behavior conforms to the scripts and symbols of market competition (Hughes et al. 1997). Institutionalized routines, once established, define paths which others follow, making future actions 'path dependent.' Most industries (and certainly health care) are 'oligopolies guarding carefully differentiated niches through strategies that aim to preserve rents by avoiding ruinous competition' (DiMaggio 1998, p. 700).

Key dimensions for analyzing economic interactions in health care are: (a) the relative frequency of transactions; (b) their relative complexity; (c) their relative specificity; (d) their relative size and risk; (e) their relative quality of information; and (f) the degree of social relationship over time. Each of these continua must be assessed in terms of each major party affected. Sociological dimensions of the relationships between parties in an economic transaction include: (g) the degree of autonomy vs. mutual obligations between the; (h) the degree of tightness or looseness in the network or organization; (i) the size of the organizations or networks; and (j) the structure of power relationships. This approach is likely to deepen our understanding of health care markets in the future.

At a second level, the concept of 'health care markets' and its recent rise to dominance in certain countries, but not others, needs explaining. In a historical study of the 'conceptions of control' that prevail in a given industry from one period to another, Fligstein (1990) has described how a crisis in an industry prompts its leaders to construct or select a model of their business and its realities that will provide them with a coherent account of what went wrong and how to succeed in the future. The theories of price competition and managed competition in the late 1970s and 1980s seemed to have been embraced by payers for these reasons. Fligstein notes that the prevailing conception is interpreted as suits individual leaders, an observation that conforms to the selective and partial ways in which corporations and governments implement theories for how to have cost-effective models of health care markets. As a result, health care markets has become an international business and policy movement, spurred by international consulting firms, even as some countries have spurned them. The World Bank, the International Monetary Fund, and other major lenders began to require, or strongly urge, scores of developing countries to restructure their services into health care markets as a condition for receiving financial help with their economies (World Bank 1993, Stocker et al. 1999).

Selective and partial implementation probably explains why managed care and managed competition have not resulted in greater efficiency or better services in some countries and why they have held cost increases at bay in the United States for only a few years. As the crisis of sharply increasing costs during the 1980s resumed in 2000, employers were concluding that health care markets as a 'conception of control' could not hold down costs and were losing faith in it (Freudenheim 2000). Like leaders in some other countries, they may start to search for a new conception of control, one that will explain why the health care markets of the previous 20 years were only partially successful and how to manage costs more effectively.

See also: Health Care Delivery Services; Health Care: Legal Issues; Health Care Organizations: For-profit and Nonprofit; Health Care, Rationing of; Health Care Systems: Comparative; Health Care Technology; Health Insurance: Economic and Risk Aspects; Managed Care; Minority Access to Health Services: United States; National Health Care and Insurance Systems; Pecuniary Issues in Medical Services: Ethical Aspects; Psychiatric Care, Economics of; Services Marketing; Socioeconomic Status and Health Care

Bibliography

Arrow K 1963 Uncertainty and the welfare economics of medical care. *American Economic Review* **53**: 941–73

Ashton 1999 The Health reforms: To market and back? In: Boston J, Dalziel P, St. John S (eds.) *Redesigning the Welfare State in New Zealand: Problems, Policies and Prospects*. Oxford University Press, Auckland, pp. 134–53

Bengoa R, Hunter D J (eds.) 1991 *New Directions in Managing Health Care*. World Health Organization Regional Office for Europe, Copenhagen

DiMaggio P 1997 Culture and cognition. *Annual Review of Sociology* **23**: 263–87

DiMaggio P 1998 The new institutionalism: Avenues of collaboration. *Journal of Institutional and Theoretical Economics* **154**: 696–705

Dobbin F 1994 *Forging Industrial Policy*. Cambridge University Press, New York

Enthoven A 1988 *Theory and Practice of Managed Competition in Health Care Finance*. North-Holland, Amsterdam

Esping-Andersen G 1994 *After the Golden Age: The Future of the Welfare State in the New Global Order*. United Nations Research Institute for Social Development, Geneva

Evans R G 1997 Going for the gold. *Journal of Health Politics, Policy and Law* **22**: 427–65

Fligstein N 1990 *The Transformation of Corporate Control*. Harvard University Press, Cambridge, MA

Folland S, Goodman A C, Stano M 1997 *The Economics of Health and Health Care*. Prentice-Hall, Upper Saddle River, NJ

Fox D M 1979 *Economists and Health Care: From Reform to Relativism*. Prodist, New York
Freudenheim M 2000 H.M.O. costs spur employers to shift plans. *New York Times* 6 September: A1
Granovetter M 1985 Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* **91**: 481–510
Grembowski D E, Diehr P, Novak L C et al 2000 Measuring the 'managedness' and covered benefits of health plans. *Health Services Research* **35**: 707–34
Hsiao W C 1994 'Marketization'—the illusory magic pill. *Health Economics* **3**: 351–57
Hughes D, Griffiths L, McHale J V 1997 Do quasi-markets evolve? Institutional analysis and the NHS. *Cambridge Journal of Economics* **21**: 259–76
LeGrand J, Barrett W (eds.) 1993 *Quasi-Markets and Social Policy*. Macmillan, London
Light D W 1995a Countervailing powers: A framework for professions in transition. In: Johnson T, Larkin G, Saks M (eds.) *Health Professions and the State in Europe*. Routledge, London pp. 25–41
Light D W 1995b *Homo Economicus*: Escaping the traps of managed competition. *European Journal of Public Health* **5**: 145–54
Light D W 1997 From managed competition to managed cooperation: Theory and lessons from the British experience. *Milbank Quarterly* **75**: 297–341
Lindenberg S 1998 The cognitive turn in institutional analysis: Beyond NIE and NIS? *Journal Institutional and Theoretical Economics* **154**: 716–27
Long S H Marquis M S 1998 How widespread is managed competition? *Health System Change Data Bulletin*. Center for Studying Health System Change, Washington, DC
Mossialos E, LeGrand J (eds.) 1999 *Health Care and Cost Containment in the European Union*. Ashgate, Aldershot, UK
Parker R M, Williams M V, Weiss B D 1999 Health literacy: Report of the Council on Scientific Affairs. *Journal of the American Medical Association* **281**: 552–7
Reinhardt U E 1996 A social contract for 21st century health care: three-tier health care with bounty hunting. *Health Economics* **5**(6): 479–99
Rice T 1998 *The Economics of Health Reconsidered*. Health Administration Press, Chicago
Robinson R, Steiner A 1998 *Managed Health Care*. Open University Press, Buckingham, UK
Saltman R E, Figueras J, Sakellarides C (eds.) 1998 *Critical Challenges for Health Care Reform in Europe*. Open University Press, Buckingham, UK
Saltman R E, vonOtter C 1992 *Planned Markets and Public Competition: Strategic Reform in Northern European Health Systems*. Open University Press, Buckingham, UK
Smelser N J, Swedberg R (eds.) 1994 *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ
Social Science and Medicine 2001 **52**(8) (special issue)
Stocker K, Waitzkin H, Iriat C 1999 The exportation of managed care to Latin America. *New England Journal of Medicine* **340**: 1131–6
World Bank 1993 *World Development Report 1993: Investing in Health*. Oxford University Press, Washington, DC
Zelman W A, Berenson R A 1998 *The Managed Care Blues & How to Cure Them*. Georgetown University Press, Washington, DC

D. W. Light

Health Care Organizations: For-profit and Nonprofit

This article must begin with an explanation of the meaning and boundaries of its subject. First, its focus is on organizations, not on the motivation of individuals. Second, it pertains to two ownership categories that by convention, partly rooted in the US experience, stand in distinction to public or governmental ownership. That is, this article pertains to nongovernmental organizations, which may be established as for-profits or as nonprofits.

Third, it pertains to a legal characteristic of organizations that is defined under the law of the jurisdiction in which they exist. In the USA, health care organizations are typically incorporated under state laws. Variations among these state laws are not important for our purposes, but the fact that fundamental aspects of the ownership and control of health care organizations (as well as the roles they play in the health care system) are defined by the laws and traditions that are specific to particular countries is. A recent effort to measure and compare the size of the nonprofit sectors in different countries had to define 'nonprofits' in terms of certain organizational characteristics (not profit-distributing, self-governing, voluntary) rather than legal form, because nations do not make the same legal distinctions between private and governmental or between for-profit and nonprofit (Salamon et al. 1999). To cite an example from health care, ownership by identifiable individuals is incompatible with the nonprofit form under US law, while Japan, where most hospitals are owned by physicians, is often said to prohibit for-profit ownership of hospitals because of a prohibition of ownership by corporations that pay dividends (Rodwin and Okamoto 2000).

Fourth, this article pertains primarily to types of health care organizations in which a mix of for-profit and nonprofit ownership is found. Thus, attention will not be directed to organizations that manufacture pharmaceuticals or hospital supplies (which are typically all for-profit corporations), to physicians' practices (which are seldom incorporated as nonprofit organizations), or to charitable fundraising organizations (such as the American Cancer Society or the Royal Commonwealth Society for the Blind), which are all nonprofit. Our focus is on the types of organizations among which both for-profits and nonprofits can be found, as is the case with hospitals, nursing homes, health maintenance organizations, and home care agencies in the USA.

Most literature on for-profit and nonprofit health care comes from the USA, which has an unusually pluralistic health care system on both the provider and purchaser sides (Starr 1982). Although the health care systems of other countries are much more typified by governmental or quasigovernmental ownership, the private ownership of important components, such as

hospitals, is not peculiar to the USA. In some countries there are doctor-owned clinics, small hospitals, or nursing facilities that are outgrowths or extensions of private physicians' offices. Some countries that have universal health insurance systems nonetheless have private hospitals. These countries may either prohibit for-profit ownership, as does Canada, or they may exclude private hospitals (including for-profits) from participation in the national health insurance system, while permitting for-profit hospitals to serve people who have the desire and wealth to opt out of that system. Several investor-owned hospital companies from the USA also own hospitals in other countries, serving such niches. And at least one German hospital company owns some US hospitals. In recent years, American companies have begun to export versions of managed care to other countries, notably in Latin America (Stocker et al. 1999).

In the main, however, in Western countries other than the USA for-profit organizations that provide health services are extensions of doctors' practices or constitute a sector that serves a small market outside of the governmental health system (and is often not included in statistical information about that system). This, plus the nation-specific nature of the laws that define the legal form of health care organizations, probably explains the paucity of research comparing for-profit/nonprofit organizations in countries other than the USA or making cross-national comparisons of ownership forms. Policy debates about privatization have stimulated research on public versus private provision of services in a number of countries, but these are outside the scope of this article.

1. *The Nonprofit/For-profit Distinction*

In this article, the terms 'for-profit' and 'nonprofit' (or its less economical synonym, 'not-for-profit') are used as they are commonly defined under state statutes and federal tax law in the USA. The term 'nonprofit' invites confusion, since nonprofit organizations are not forbidden from generating surpluses of revenues over expenses. Indeed, such surpluses, which are called profits in the for-profit world, are essential to economic health and survival of nonprofit organizations. It is important, therefore, to be clear about the shared characteristics of for-profit and nonprofit organizations and the ways that they differ.

Table 1 shows that there are important similarities and differences between for-profit and nonprofit health care organizations in the USA (Gray 1992). Two key families of legally defined characteristics are particularly important in distinguishing these organizations. The first pertains to ownership, governance, and purpose. For-profit organizations have owners (proprietors, partners, or the shareholders of corporations), who share in the profits and increases (or decreases) in the value of the enterprise and who, in the corporate form, elect the board of directors. Those who operate the for-profit organization are thus either owners or are managers who are directly accountable to owners. Thus, those who control the organization have a direct stake in the financial performance of the organization.

By contrast, nonprofit organizations do not have owners (except insofar as they are subsidiaries of other nonprofit organizations) and are legally prohibited

Table 1
Basic legal and economic differences between for-profit and nonprofit hospitals

For-profit	Nonprofit
Corporations owned by investors	Corporations without owners or owned by 'members'
Purpose: Management has legal obligation to promote wealth of shareholders within the boundaries of law; does so by providing services	*Purpose*: Has legal obligation to fulfill a stated mission (provide services, education, research, etc.); must maintain economic viability to do so
Can distribute some proportion of profits (net revenues less expenses) to owners as dividends; owners also share in increases in the value of the enterprise	Cannot distribute surplus (net revenues less expenses) to those who control the organization; there are no owners entitled to proceeds if organization is sold
Management ultimately accountable to owners (stockholders) through boards of directors elected by stockholders	Management accountable to voluntary, often self-perpetuating boards of trustees
Sources of capital include:	Sources of capital include:
Equity capital from investors	Charitable contributions
Debt	Debt (generally tax-exempt)
Retained earnings (plus depreciation and deferred taxes)	Retained earnings (plus depreciation)
	Government grants
Revenues derived from sale of services	Revenues derived from sale of services and from charitable contributions
Pay property, sales, and income taxes	Generally exempt from most taxes

from distributing surplus revenues to those who control the organization—a characteristic that Hansmann (1980) called the nondistribution constraint. Nonprofit organizations are governed by boards of trustees whose members are typically chosen by the existing board and who are responsible for assuring that the organization operates soundly and adheres to the mission (charitable, educational, religious, etc.) for which it was chartered. Importantly, the boards to which management is accountable have no direct stake in the profitability of the institution.

The second family of differences pertains to the tax treatment of for-profit and nonprofit organizations (Simon 1987). Nonprofit organizations that meet qualifying standards have a variety of tax-related benefits under federal, state, and local laws. The tax treatment of nonprofits varies according to the kind of organization it is (for example, there are differences in how churches, grantmaking foundations, charitable organizations, and professional associations are treated) and according to governmental jurisdiction (criteria for exemption from local property taxes may differ from criteria for exemption from federal corporate income tax). However, most nonprofit health care organizations are exempt as charities from federal corporate income tax under Section 501(c)(3) of the Internal Revenue Service Code. Charity in this case is defined in terms of benefiting the community at large, not in terms of providing free care (Fox and Schaffer 1991). (Some health maintenance organizations and insurers are exempt under Section 501(c) (4) as social welfare organizations and receive less-favorable tax treatment than do the charitable organizations.) Other tax benefits also apply to organizations that are exempt in the eyes of the Internal Revenue Service. Charitable nonprofits are generally able to receive contributions that are tax deductible to the donor, to borrow at favorable rates under bond issues whose interest is not taxable to the lender, and to be exempt from local and state property taxes and certain excise taxes.

However, in health care there are important financial similarities between for-profits and nonprofits. First, the revenues of both for-profit and nonprofit health care organizations come overwhelmingly from the sale of services. For most nonprofit health care organizations, charitable contributions represent no more than one percent of revenues. To survive, these nonprofit organizations must compete successfully for business in a very crowded and heavily regulated marketplace. Though not businesses, they must operate in a very businesslike fashion.

Second, although nonprofit hospitals and HMOs have benefited in earlier periods from governmental grant and lending programs, today's nonprofits and for-profits depend upon similar sources of capital—primarily retained earnings and debt (Institute of Medicine 1986). Thus, just as with for-profit organizations, the primary capital sources on which nonprofit health care organizations depend are a function of the organization's economic performance. The fact that the typical nonprofit health care organization is tax exempt as a charitable organization under the Internal Revenue Service code suggests that they are subject to complex and somewhat conflicting social expectations.

Notwithstanding the importance of their common reliance on retained earnings and debt, for-profit and nonprofit organizations each have access to unique sources of capital. Unlike nonprofits, for-profits have the paid-in equity capital of individual and organizational investors or stockholders. Access to equity capital provides a substantial advantage during periods of growth, particularly for publicly traded firms whose stock price in relationship to earnings may be very high. The unique form of capital to which nonprofits have access is charitable contributions, as well as debt financing from instruments whose interest payments are not taxable for the lender.

These differences in sources of capital are enormously important regarding the expectations placed on these organizations—with stockholders seeking value, economic returns, and capital growth, and contributors seeking charitable behavior. The two families of differences have driven theoretical and empirical work in the social sciences (Powell 1987, Powell and Clemens 1998) and stimulated an array of public policy debates.

2. *The Role and Performance of Nonprofit Organizations*

The question of why some segments of our capitalist economy are predominantly for-profit while others are predominantly nonprofit has intrigued social theorists and economists, who generally start with the premise that the incentives internal to the for-profit structure create a drive toward efficiency. So, what is the role of nonprofits? Several answers have been offered. Weisbrod (1990) suggests that nonprofits occupy a niche created by, on the one hand, the failure of for-profit organizations to provide services that benefit the community at large, not just the purchasers, and, on the other hand, the failure of governmental programs to meet the demand for other than average services. Hansmann (1980) posits that the nonprofit form provides a comparatively efficient alternative in situations of contract failure in which purchasers are not well situated to monitor the performance of sellers of services. The nondistribution constraint makes nonprofits less likely than for-profits to exploit the purchaser's vulnerability. For services to be financed by donations, the nonprofit form gives potential donors some confidence that their funds will be used for the promised purpose (e.g., for charitable services). For services that are difficult for purchasers to monitor (and medical care is an example), the nonprofit form may provide some assurance that vulnerabilities of patients and payers will not be exploited.

Such theories suggest three directions in assessing the performance of nonprofit organizations—their costliness (in the absence of accountability to owners, are they less efficient?), their charitability (very important for tax-exempt organizations in a health care system with more than 40 million uninsured patients), and their effectiveness or trustworthiness as providers of services/care. The empirical literature that has attempted to assess these topics in different types of health care organization is large, growing, and impossible to summarize adequately in a brief article. Citations to most studies can be found, however, in the Institute of Medicine (1986) report on for-profit and a 1999 literature review that is available on-line (Gray et al. 1999) and that summarizes some 75 studies.

First, in health care, for-profit organizations are not necessarily less costly to purchasers than are nonprofits. Most studies of hospitals have found that costs to purchasers are higher in for-profits (Gray et al. 1999). This appears to be a response to 'reimbursement maximization' strategies based on exploiting the weaknesses of third-party payment systems, not to some inherent inefficiency disadvantage of for-profits vis-à-vis nonprofits. For example, several studies have shown that expenses incurred in the production of hospital services are similar in for-profit and nonprofit hospitals, but that markups of price over expenses are higher in for-profits.

Second, taken as a group, nonprofits are more likely to engage in charitable activities and public goods than are for-profits (Gray 1992, Claxton et al. 1997, Gray et al. 1999). For example, nonprofits care for more uninsured patients (Frank et al. 1990), provide more uncompensated care (Lewin et al. 1988), provide more services focused on vulnerable populations (LeBlanc and Hurley 1995, Shortell et al. 1987), and engage in education and research (Institute of Medicine 1986), and they are less likely to discourage admission of uninsured patients (Schlesinger et al. 1987) and less likely to avoid location in areas with large numbers of the uninsured (Norton and Staiger 1994). However, the charitable performance of nonprofits is quite variable, with charitable activities and public goods heavily concentrated in urban teaching hospitals. The performance of nonprofit hospitals as charitable organizations has come under criticism in recent decades, and in the 1990s eight US states established requirements for nonprofit hospitals to report on their community benefit activities. There is evidence that the charitable performance of nonprofits is enhanced by articulated community expectations and active public oversight (Schlesinger 1998).

Third, although the Institute of Medicine (1986) study found no clear pattern of ownership-related quality differences among hospitals, several more recent studies have found a quality advantage for nonprofits regarding hospital mortality (Hartz 1989, Kuhn et al. 1994, Sloan et al. 2000) and post-hospital recovery (Sloan et al. 2000). Quality advantages have been reported among nonprofit nursing homes (Hawes and Phillips 1986) and HMOs (Himmelstein et al 1999), two fields characterized by serious contract-failure problems. In nursing homes, many residents are unable to look after their own best interests and have no one else to do so and so must rely on their notoriously difficult-to-monitor relationship with the facility to protect their interests. In HMOs, the complexity of the incentive relationship between health plans and physicians may make it difficult for patients to be certain that contracted-for services are actually being provided.

At a more general level, the research findings to date suggest that for-profits are indeed more responsive to economic incentives than are nonprofits, which can be a problem in a field with significant information asymmetries between sellers and purchasers of service (Arrow 1963). In view of the similarities of for-profits and nonprofits in revenue capital and revenue sources and the regulatory regimes under which they must operate, the overall pattern of performance differences that are often small but generally in the predicted direction is notable. It is clear that the comparative performance of nonprofits and for-profits is not a simple function of ownership form, but of an interaction of ownership form with market forces, payment systems, and regulatory environments.

3. *Ownership Trends in Health Care*

The concentration of different forms of ownership varies in the USA across fields (nursing homes are predominantly for-profit; hospitals are mostly nonprofit) and over time (Marmor et al. 1987, Gray 1999). Among acute care community hospitals, surveys conducted early in the twentieth century reported that about half of all hospitals were for-profit (Stevens 1989, Starr 1982). The for-profit share of hospitals declined through the century, falling to 15 percent in 1965, with nonprofits comprising 60 percent and the remainder governmental—a level that has remained roughly the same, with a slight increase in for-profit and a slight decrease in governmental. However, the character of the for-profits has changed since the mid-1960s, with ownership by large, powerful, publicly traded investor-owned firms replacing ownership by local entrepreneurs (often doctors) as the predominant for-profit form, and with the average for-profit facility being larger as measured by the number of beds.

There are, however, a number of signs that the US health care system is moving toward an increased for-profit character. Even among hospitals, where the nonprofit share has been stable at 60 percent for 30 years, a shift toward for-profit has occurred beneath the surface. First, if beds rather than hospitals is used as the measure, the for-profit share of acute-care hospitals doubled between the mid-1960s and the mid-1990s (from 6 percent to 12 percent, largely at the

Table 2
Health care organizations by form of ownership, USA 1985 and 1997

Provider type	1985		1997		% of change >85–>97
Community hospitals[a,b]					
For-profit	805	14%	797	16%	−1
Nonprofit	3,349	58%	3,000	59%	−10
Government	1,578	28%	1,260	25%	−20
Private psychiatric hospitals[a,c]					
For-profit	149	64%	233	77%	56
Nonprofit	84	36%	69	23%	−18
Private rehabilitation hospitals[a,c]					
For-profit	11	18%	91	58%	727
Nonprofit	51	82%	65	42%	27
Home health agencies[d]					
For-profit	2,055	36%	6,290	67%	206
Nonprofit	3,653	64%	3,123	33%	−15
Nursing homes[e,f]					
For-profit	14,300	75%	11,250	65%	−21
Nonprofit	3,800	20%	4,792	28%	26
Government	1,000	5%	1,134	7%	13
Dialysis centers[g]					
For-profit	616	42%	2,322	68%	277
Nonprofit	847	58%	1,101	32%	30
HMOs[h,i]					
For-profit	137	35%	481	74%	251
Nonprofit	256	65%	167	26%	−35

a American Hospital Association (1986) American Hospital Association Hospital Statistics, Chicago. b American Hospital Association (1999) American Hospital Association Hospital Statistics, Chicago. c American Hospital Association (1998) American Hospital Association Annual Survey of Hospitals, Chicago. d Health Care Financing Administration (1998) 1998 Profile of Medicare Chart Book, Rockville. e National Center for Health Statistics (1995) NCHS National Nursing Home Survey, Rockville. f National Center for Health Statistics (1997) NCHS National Nursing Home Survey, Rockville. g HCFA Office of Standards Planning and Quality (1999) National Listing of Medicare Providers Furnishing Kidney Dialysis Transplant Services, Baltimore. h InterStudy Publications (1985) InterStudy Guide to National Firms, Minneapolis. i InterStudy Publications (1999) InterStudy Competitive Edge 8.1, Minneapolis

expense of public hospitals). Second, even with hospitals as the unit of analysis, if one includes nonprofit and public hospitals that are managed by for-profit firms, as well as non-acute care hospitals (e.g., psychiatric and rehabilitation facilities), the for-profit sector among hospitals grew to about 25 percent of the total by the late 1990s.

Great increases in for-profit ownership have occurred elsewhere in the US health care system since the 1980s, as is shown in Table 2. Health maintenance organizations (HMOs), which were almost exclusively nonprofit in 1980, are now predominantly for-profit (Gray 1999). Strong trends in the for-profit direction also occurred among home health agencies and dialysis treatment centers. Within the trend toward for-profit ownership has been a second trend—ownership not by individual proprietors but by investor-owned corporations that own multiple units (Institute of Medicine 1986, Corrigan et al. 1997). Similar trends toward multi-institutional arrangements have also occurred among nonprofits. This, along with increased entrepreneurial activity by nonprofits (Sloan 1998) has led to concerns among many observers that the character of many nonprofit health care organizations has changed in ways that mirror the for-profits.

Comparing the cases of community hospitals and HMOs, some of the factors that are involved in the persistence and change in ownership patterns among health care organizations can be seen (Gray 1999). On the side of persistence, many nonprofit hospitals, with their multiple important community roles (care giver, employer) and their governing boards often made up of prominent citizens, have received both charitable and governmental support in bad economic times. The difficulty of closing a nonprofit hospital has been frequently noted in health policy circles as hospital occupancy has declined since the 1980s. Nonprofits have also engaged in a variety of strategies to compete and grow—entering into multihospital systems, merging, joint-venturing, and establishing for-profit subsidiaries. Use of these strategies has carried a price, however. The increased commercialism of the nonprofits has attracted criticism and fanned doubts about their deservedness of tax exemption.

The most important factor affecting the ownership composition of the health care field is public policy. It

could hardly be otherwise since the basic corporate structures of for-profit and nonprofit organizations are themselves products of public policy. But the role that public policy plays in influencing the composition and performance of health care organizations has become increasingly apparent since the 1950s. The creation and growth of the investor-owned hospital industry was a direct result of capital reimbursement policies built into the Medicare legislation in 1965 (Gray 1992). The interest costs incurred in borrowing money to purchase hospitals was a reimbursable expense under Medicare, and gave purchasers annual payments for the noncash expense called depreciation. Since Medicare payments to hospitals were based on reimbursing them for costs incurred, the legislation also provided for giving for-profit hospitals generous 'return-on-equity' payments so that investors would receive a return on their investments. All of these provisions stimulated the growth of the investor-owned companies into the early 1980s, when, in separate actions, Congress reduced return-on-equity payments and changed the depreciation rules so that Medicare would pay for the depreciation of any particular asset only once. After these changes were made, the leading investor-owned companies all stopped growing at a time in which it was being predicted that they were going to take over the system.

Similarly, the growth of the nonprofit HMO industry in the 1970s and the for-profit transformation of the field in the 1980s were direct results of public policy. First, provisions of the HMO Act of 1973 made capital available to nonprofit HMOs. But the field was transformed in the for-profit direction in the 1980s as a result of the Reagan administration's decision to phase out the HMO Act's support of nonprofit HMOs and to undertake a program to stimulate investment in for-profit HMO companies, which had not previously existed; a surge in demand for managed care from the corporate sector which was experiencing large increases in health care costs; and inattentive state regulators who allowed many nonprofit HMOs to be acquired for a pittance by their leadership and converted into for-profit organizations (Gray 1999).

Permissive state regulation also created the environment for a surge in the acquisition of nonprofit hospitals by large investor-owned firms (Columbia/HCA, Tenet) in the early and mid-1990s (Gray 1997). This, along with conversions involving HMOs and state Blue Cross insurance companies, led more than 25 states to pass legislation in the late 1990s to regulate nonprofit conversions. Some 120 new foundations were created in the 1980s and 1990s from the proceeds of the sale of nonprofit health care organizations (Isaacs et al. 1997). Studies of the consequences of hospital conversions have generally shown them to have occurred among financially troubled institutions and that few negative consequences could be documented, either because the institutions had previously engaged in only low levels of charitable activity or because strong contractual protections had been built into the sales agreement (Mark 1999, Young and Desai 1999). Interestingly, there have also been examples of for-profit hospitals that reverted to the control of nonprofit organizations, though the trend remains in the direction of greater for profits.

4. *Conclusion*

The American health care experience has shown that for-profits and nonprofits can coexist in the same field. It has also shown that the comparative performance of the two forms of organizations are influenced by contextual factors such as the incentive structures built into the systems by which they are paid, regulatory regimes, professional values, and public expectations. Although for-profits generally appear to be more responsive to economic incentives than are nonprofits, nonprofits cannot be indifferent to economic realities. It is increasingly apparent that policymakers need to pay attention to the characteristics of the organizations that provide health services as they design programs, payment systems, and regulatory structures to achieve desired health policy goals.

The US experience also shows that the for-profit/nonprofit composition of fields can change, sometimes quite rapidly, as can the ownership form of particular organizations. In a policy-sensitive field such as health care, the future of for-profit and nonprofit health care organizations is likely to be affected quite substantially by public policies. These policies may be focused specifically on some aspect of the ownership issue, but the most important ones may have an ownership-neutral motivation, such as the desire to control costs in large governmental programs. For segments of the health care industry in which government is the dominant payer, it may be difficult for investor-owned companies to generate revenues in the long run that will be satisfactory to investors. High profitability is likely to lead to cuts in the programs that fund that profitability. For-profits also face suspicion that has been driven by highly publicized fraud-related scandals involving some leading health care companies, plus widespread distrust of the predominantly for-profit HMO industry.

For their part, nonprofits face a twin challenge. In a highly competitive health care system dominated by large purchasers and without direct governmental subsidies, they must perform well enough economically to be able to continue paying their bills and to maintain their access to capital, which depends upon lenders' confidence in their ability to repay their debts. As they do this, they must also contend with a growing skepticism, fed in part by their own economically driven policies and practices, regarding whether they engage in sufficient charitable activity to deserve the tax-related benefits they receive.

With health care accounting for almost 15 percent of the US economy, investor money is likely to be available to fund new ideas for making money from the provision of health services, and there is interest among American companies in making investments in other countries as well. The clear advantages that the for-profit form might offer are the ability to raise capital, a willingness to take risks and innovate, and, perhaps, a reservoir of experience that might be transferable into new contexts. But will countries that have had little experience with this organizational form in their health care system find such factors outweigh possible costs in terms of a changed ethos of the health care system, major shifts in power and control, and potential serious problems of exploitation of the informational asymmetries that typify health care?

See also: Alternative and Complementary Healing Practices; Gender and Health Care; Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care, Rationing of; Health Care Technology; National Health Care and Insurance Systems; Psychiatric Care, Economics of

Bibliography

Arrow K 1963 Uncertainty and the welfare economics of health care. *American Economic Review* **53**: 941–73

Claxton G, Feder J, Shactman D, Altman S 1997 Public policy issues in nonprofit conversions: An overview. *Health Affairs* **16**: 9–28

Corrigan J M, Eden J S, Gold M R, Pickreign J D 1997 Trends toward a national health care marketplace. *Inquiry* **34**: 11–28

Fox D M, Schaffer D C 1991 Tax administration as health policy: Hospitals, the IRS, and the courts. *Journal of Health Politics, Policy, and Law* **16**: 251–79

Frank R G, Salkever D S, Mullan F 1990 Hospital ownership and the care of uninsured and Medicaid patients: Findings from the National Hospital Discharge Survey 1979–1984. *Health Policy* **14**: 1–11

Gray B H 1992 *The Profit Motive and Patient Care: The Changing Accountability of Doctors and Hospitals*. Harvard University Press, Cambridge, MA

Gray B H 1997 Conversions of HMOs and hospitals: What's at stake? *Health Affairs* **16**: 29–43

Gray B H 1999 The changing face of health care. In: Clotfelter C T, Ehrlich T (eds.) *Philanthropy and the Nonprofit Sector in a Changing America*. Indiana University Press, Bloomington, IN

Gray B H, Carrino G, Collins S, Gusmano M 1999 *The Empirical Literature Comparing For-Profit and Nonprofit Hospitals, Managed Care Organizations, and Nursing Homes: Updating the Institute of Medicine Study*. The New York Academy of Medicine, New York

Hansmann H 1980 The role of nonprofit enterprise. *The Yale Law Journal* **89**: 835–901

Hartz A J 1989 Hospital characteristics and mortality rates. *The New England Journal of Medicine* **321**(25): 1720–5

Hawes C, Phillips C D 1986 The changing structure of the nursing home industry and the impact of ownership on quality, cost, and access. In: Institute of Medicine (ed.) *For-Profit Enterprise in Health Care*. National Academy Press, Washington, DC

Himmelstein D U, Woolhandler S, Hellander I, Wolfe S M 1999 Quality of care in investor-owned vs not-for-profit HMOs. *Journal of the American Medical Association* **282**: 159–63

Institute of Medicine 1986 *For-Profit Enterprise in Health Care*. National Academy Press, Washington, DC

Isaacs S L, Beatrice D F, Carr W 1997 Health care conversion foundations: A status report. *Health Affairs* **16**: 228–36

Kuhn E M, Hartz A J, Krakauer H, Bailey R C, Rimm A A 1994 The relationship of hospital ownership and teaching status to 30- and 180-day adjusted mortality rates. *Medical Care* **32**(11): 1098–108

LeBlanc A J, Hurley R E 1995 Adoption of HIV-related services among urban US hospitals: 1988 and 1991. *Medical Care* **33**(9): 881–91

Lewin L S, Eckels T J, Miller L B 1988 Special report. Setting the record straight: The provision of uncompensated care by not-for-profit hospitals. *The New England Journal of Medicine* **318**(18): 1212–16

Mark T L 1999 Analysis of the rationale for, and consequences of, nonprofit and for-profit ownership conversions. *Health Services Research* **34**(1): 83–101

Marmor T R, Schlesinger M, Smithey R W 1987 Nonprofit organizations and health care. In: Powell W W (ed.) *The Nonprofit Sector: A Research Handbook*. Yale University Press, New Haven, CT

Norton E C, Staiger D O 1994 How hospital ownership affects access to care for the uninsured. *RAND Journal of Economics* **25**: 171–85

Powell W W (ed.) 1987 *The Nonprofit Sector: A Research Handbook*. Yale University Press, New Haven, CT

Powell W W, Clemens E S (eds.) 1998 *Private Action and the Public Good*. Yale University Press, New Haven, CT

Rodwin M A, Okamoto A 2000 Physicians' conflicts of interest in Japan and the United States: Lessons for the United States. *Journal of Health Politics, Policy, and Law* **25**: 343–75

Salamon L M, Anheier H K, List R, Toepler S, Sokolowski S W 1999 *Global Civil Society: Dimensions of the Nonprofit Sector*. Johns Hopkins Center for Civil Society Studies, Baltimore, MD

Schlesinger M 1998 Mismeasuring the consequences of ownership: External influences and the comparative performance of public, for-profit, and private nonprofit organizations. In: Powell W W, Clemens E S (eds.) *Private Action and the Public Good*. Yale University Press, New Haven, CT

Schlesinger M, Bentkover J, Blumenthal D, Musacchio R, Willer J 1987 Privatization of health care and physician's perceptions of access to hospital services. *The Milbank Quarterly* **65**(1): 25–58

Shortell S M, Morrison E M, Hughes S L, Friedman B S, Vitek J L 1987 Diversification of health care services: The effects of ownership, environment, and strategy. *Advances in Health Economics and Health Services Research* **7**: 3–40, 111–14

Simon J G 1987 The tax treatment of nonprofit organizations: A review of federal and state policies. In: Powell W W (ed.) *The Nonprofit Sector: A Research Handbook*. Yale University Press, New Haven, CT

Sloan F A 1998 Commercialism in nonprofit hospitals. In: Weisbrod B A (ed.) *To Profit or Not to Profit*. Cambridge University Press, Cambridge, UK

Sloan F A, Picone G A, Taylor D H, Chou S Y 2000 Hospital ownership and cost and quality of care: Is there a dimes' worth of difference? *Journal of Health Economics* forthcoming
Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York
Stevens R 1989 *In Sickness and in Wealth: American Hospitals in the Twentieth Century*. Basic Books, New York
Stocker K, Waitzkin H, Iriart C 1999 The exportation of managed care to Latin America. *The New England Journal of Medicine* **340**(14): 1131–6
Weisbrod B A 1990 *The Nonprofit Economy*. Harvard University Press, Cambridge, MA
Young G J, Desai K R 1999 Nonprofit hospital conversions and community benefits: New evidence from three states. *Health Affairs* **18**(5): 146–55

B. H. Gray

Health Care, Rationing of

The term 'rationing' is used with different meanings in health care discussions (Mechanic 1979). Most generally, it refers to processes of allocation when demand for care exceeds the available supply or the willingness of payers to increase it. In its more narrow application, it refers to how a scarce medical resource, such as hearts, kidneys, or livers, is allocated among competing recipients, only some of whom can receive it. Medical care has always been rationed in both senses, but the concept itself is associated with much affect and denials that rationing occurs.

To the extent that medical care is treated substantially as a commodity, as in the USA, then the market distributes services. Individuals with more comprehensive insurance or the ability to pay out-of-pocket command a greater quantity of and more sophisticated services. But even in market systems, some life and death services in scarce supply transcend the market and are governed by different rationing schemes. In publicly funded medical systems, where most of the population has health care entitlement, whether through national health insurance or a national health service, the availability of services depends on decisions about what services will be financed and their comprehensiveness and intensity (Klein et al. 1996). Some such systems, for example the British National Health Service, allow competing private systems of care. In contrast, Canada prohibits private provision of services covered by its health insurance program.

Rationing processes and decisions are implemented at different levels of government, by sick funds, large insurance plans, provider organizations such as hospitals, clinics, and health maintenance organizations, and at the point of service delivery between clinician and patient. Disagreements are common on where in the hierarchy of health care organization varying types of rationing decisions should be made (macro, meso, or micro) and on appropriate processes for making such decisions (Mechanic 1997). At each level, rationing may be more or less explicit.

1. *Explicit vs. Implicit Rationing*

Explicit rationing refers to the detailed specification of how resources will be allocated by a rationing authority. Explicit rationing includes specifying eligibility rules, the specific services to be covered, the amount of coverage for recipients, intervals for receiving services, copayment requirements, the proportion of the budget to be allocated to varying services, the geographic allocation of services, the specific form and location of many services, and appliances and drugs that may or may not be allowed.

Implicit rationing, in contrast, involves setting budgetary limits but allowing the processes of allocation to be worked out at lower organizational levels (Mechanic 1979). Implicit rationing includes prospective reimbursements or capitation payments. In each case the authority sets the financial limits but then allows the budgeted or capitated entity to establish the processes for decision-making. Total responsibility is sometimes delegated to the lower unit but more typically there are some explicit guidelines and constraints over discretion. All medical systems combine explicit and implicit processes but the balance may differ considerably from one nation to another and within varying national contexts.

Klein et al. (1996) describe a range of strategies for implicitly limiting resources allocated to individuals. These include denial as beneficiaries or for the services claimed; selection of beneficiaries or services on the basis of need or likely success, or demographic characteristics such as age; deflection of care to other sectors or providers; deterrence through bureaucratic barriers such as waiting lists and other delays; diluting services by reducing intensity or restricting specialty care; and premature closure or early discharge.

2. *Concerns About Rationing*

Despite general acceptance of the inevitability of rationing, major issues of controversy include who should determine rationing strategies, the openness of decision processes, and whether decisions should be made by managers or clinicians (New 1997). Common criticisms are that the implicit rationing processes used are not open, often favor the privileged, and result in unequal care for persons of lesser status, education, and sophistication, and with less skill in managing bureaucratic systems (Churchill 1987). Critics also argue that implicit rationing gives doctors free reign to act on their prejudices. The lack of openness makes it difficult to assess the fairness of decision-making or to

allow aggrieved individuals to successfully appeal exclusions or dilution of care. Implicit rationing is also seen as undermining democratic processes through which communities can bring their preferences to bear in the allocation of important public resources.

A significant difficulty with explicit rationing, in contrast, is that it mobilizes interest groups who seek to affect decisions about coverage. Bringing medical decisions into the public arena energizes a political process and instability in the medical care system (Mechanic 1995). While a particular treatment may be of little value in general, there may be instances, difficult to specify beforehand, where it is appropriate. Explicit rationing can not, in most instances, accommodate the complex contingencies associated with the use of a particular treatment for an individual patient.

Questions about transparency apply at all levels of decision-making. The criteria used when a national government divides a health budget among regions can have significant consequences. Funds can be allocated politically, on the basis of historical funding, and on the basis of algorithms that take account of population characteristics and dispersion, and varying concepts of health status and need. The availability of facilities and specialists differs widely among population areas and accounts for large variations in expenditures among varying population groups. Nations can seek to bring greater equity among regions, types of services, and even individuals by how they decide to distribute resources.

Within organizations, decisions are made to allocate funding among primary and specialty services, among specialty units, and between medical and related health services. These decisions can have large consequences for the likelihood that a person with a particular need can receive responsive care. At the clinical level, decisions about waiting time, access to specialized services, availability of specialized treatments, or home or hospital treatment can profoundly affect patterns of care. Debate persists as to the extent such decisions should be governed by public discussion and decision-making or left to the discretion of clinicians based on their own professional judgment and experience (Symposium The Law and Policy of Health Care Rationing 1992).

Concerns about fairness and facilitating public discussion of priorities would support making practices explicit but there are significant competing concerns. Despite increasing interest in knowledge-based medical practice, much remains uncertain. There is no consensus, in most areas, on standards that apply to a diverse patient population with varying needs and inclinations and unique medical histories. Medical knowledge and practices are rapidly changing and it is uncertain whether explicit standards, once established, can be modified to keep pace with new knowledge. Moreover, medical care is an iterative process with assessment and disease management developing over time. As a complex process of discovery and negotiation, clinical practice is not easily subjected to definitive technical rules.

The notion of explicit decision-making at the level of direct patient care implies that individuals with comparable conditions have similar needs and preferences. But treatment must be flexible and take account of varying social contexts, cultural inclinations, and personal preferences. Treatments of comparable medical effectiveness may have vastly different implications for work and family life, or for comfort under varying circumstances. To treat medical care simply as a technical process, devoid of meanings, preferences, and differential consequences, contributes to dehumanizing medicine, and is unacceptable to many professionals and patients.

3. *Rationing of Specific Resources*

Rationing becomes especially salient when problems result, not simply because of the unwillingness to provide more resources or allocate them fairly over some population (which may result in increased illness and death in a statistical sense), but when there is an inability to provide a life-saving resource to all identified individuals who seek it. Such situations involving the allocation of hearts, kidneys, livers, and lungs have been called 'tragic choices' because there are no fair or generally accepted rules for making the decisions as to who receives the resource and who dies (Calabresi and Bobbitt 1978). A variety of criteria have been applied from one occasion to another, such as allocation by age, by community contribution, by the largest opportunity to benefit, by family circumstances, by ability to cooperate in the treatment process, and many more, but all raise serious ethical conflicts (Elster 1992). Age is commonly suggested as a suitable criterion (Callahan 1990) and has been used in making tragic choices (Aaron and Schwartz 1984) but has been strongly contested. The most satisfactory criteria are medical ones based on likely technical success. But even these criteria do not adequately discriminate among potential beneficiaries, and social judgments and biases inevitably intrude (Halper 1989). For example, judgments about who can best cooperate in treatment typically give preference to patients with more education, higher intelligence, attractive personality characteristics, and strong social attachments.

Several analysts have concluded that because the equity issues cannot be resolved intellectually, the fairest mode of allocation is a lottery where persons with comparable medical qualifications have equal chances to be selected. There is little professional or public support for such selections because most people believe that consensual judgments can be made that would improve on any random process. Many rationally constructed strategies of allocation fail also

because of public perceptions. The public places greater value on rescuing identified lives than on saving statistical lives (Hadorn 1991).

4. Specific Rationing Policies/Practices

As evidence of inequalities in care increases, motivation to distribute limited health resources more equitably has led several countries to address rationing and seek explicit allocation principles and rules. The Oregon rationing program under Medicaid (the USA's federal–state partnership to provide health care to poor and disabled subgroups) was followed with great interest in Europe and represents perhaps the most extreme example of explicit rationing (Leichter 1999). This effort sought to increase enrollees under its Medicaid program by selecting to rank treatments consistent with judgments of efficacy and other community values. Through an elaborate process of evaluating condition–treatment pairs, a list was developed that prioritized these pairs. The basic idea was to establish a funding line, given available funds, that excluded from coverage the lowest ranking condition–treatment pairs. The formal evaluation and ranking process and the use of community meetings to involve consumers and professionals in defining the values for performing the rankings excited the imagination of many outside observers. The process was seen as a rational and fair way to make difficult public decisions about what would and would not be paid for. The Oregon program, however, was not quite what it appeared to be. It was a very public and explicit format for designing the benefit structure. The Oregon Medicaid program is predominantly a managed care program in which health maintenance organizations (HMOs) are capitated to provide all funded services. Once capitated, HMOs are free to provide whatever services they deem necessary and need not conform to the list. Thus, the list was more an instrument for bargaining over rates than an explicit set of rules for rationing health care. Early implementation was aided by increased state funding and a favorable economy, which allowed the establishment of acceptable capitation rates. The political sustainability of such an approach under more difficult times is yet to be tested.

The British National Health Service (NHS) exemplifies implicit rationing. There are no explicit guaranteed benefits in the NHS but the system seeks to provide all necessary care within the resources available. Care is rationed by the availability of professionals and treatment facilities, by professional decisions about intensity of care and referrals, and by waiting lists for various types of services (Klein et al. 1996, Aaron and Schwartz 1984). The extraordinary professional discretion within the NHS has led to efforts for greater equity and accountability in resource use. Under recent reforms, health authorities were given purchasing responsibilities and could limit the treatments covered. The political difficulty of doing so explicitly in any but marginal aspects of medical care became evident. One exception was the effort to ration Viagra explicitly, a popular but expensive drug for treatment of impotence.

5. Ethical and Political Issues with Rationing Policies/Practices

A major ethical issue concerns the appropriateness of having physicians, acting as agents of their patients, also responsible for allocation decisions. Traditionally, doctors have been seen as sole agents of their patients, trying to get them what they need. Some believe it is unethical to dilute such agency with allocation responsibilities at the bedside. Others believe that physicians must have a dual responsibility for both their patients and the 'commons.' With the development of HMOs and managed care strategies in the USA, financial incentives are increasingly used to encourage doctors to allocate resources efficiently. There are worries that such incentives create ethical conflicts between the doctor's financial interests and patient care. Such conflicts are also common in unmanaged care when physicians have financial interest in diagnostic and treatment facilities to which they make referrals or in other practices for which they receive financial incentives. Studies have consistently found that such arrangements increase the use of ancillary services (Rodwin 1993).

There is widespread agreement that it is appropriate and necessary for some governing authority to establish financial constraints on health care expenditures. All national systems have such constraints in place for at least public expenditures. Few believe that it is any longer possible, even in the richest of nations, to provide open-ended financing, allowing physicians to provide all they deem useful. A more-contested issue is the extent to which such an authority, once having established such limits, should specify how the budget should be divided among different types of facilities, modes of care, professional groups, types of patient needs, and treatments and procedures, and by what criteria. The extent of authority given to the nation, as compared with regions, health districts, states, or even smaller population entities, elicits tensions in many contexts. Issues also arise as to whether organizational rationing decisions are best made by management, speciality services, or professionals exercising their best individual judgment.

A possible model that blends explicit and implicit rationing is one emphasizing clinical advocacy within a structure of procedural justice. Such a model would explicitly constrain some clinician and patient decisions through clear rules that are reasonably justified, while also allowing doctors and patients to challenge decisions and argue for revision through dispute

resolution mechanisms. Such compromise models offer opportunity for achieving more consensus.

Public views of rationing processes are important because any rationing process that lacks political legitimacy becomes unstable. One of the attractions of the Oregon program was its openness and the participation of large numbers of people. In contrast, in many areas of the USA, the implementation of utilization review (a form of explicit rationing) by managed care organizations overruling physician and patient definition of need has resulted in a significant political backlash. Managed care organizations are finding that transferring risk to professional groups (implicit rationing) kindles less friction than imposing explicit decisions on professionals and patients.

It will take decades for rationing issues to be sorted out. Health care systems and facilities are experimenting with a range of strategies that may improve decisions and help achieve legitimacy for rationing. These include community representation and involvement in decision-making, surveys to elicit community preferences, focus on evidence-based services, use of treatment guidelines and pathways, involvement of patients in treatment decisions, and implementation of disease management approaches. Rationing and the ability to sustain whatever system is preferred depends on achieving a sufficient level of value consensus so that the difficult choices are widely perceived as fair and responsive to individual need.

Bibliography

Aaron H J, Schwartz W B 1984 *The Painful Prescription*. The Brookings Institution, Washington, DC

Calabresi G, Bobbitt P 1978 *Tragic Choices*. Norton, New York

Callahan D 1990 *What Kind of Life*. Simon and Schuster, New York

Churchill L R 1987 *Rationing Health Care in America: Perceptions and Principles of Justice*. University of Notre Dame Press, Notre Dame, IN

Elster J 1992 *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. Cambridge University Press, Cambridge, UK

Hadorn D C 1991 Setting health care priorities in Oregon. *JAMA—Journal of the American Medical Association* **265**: 2218–25

Halper T 1989 *The Misfortunes of Others: End-stage Renal Disease in the United Kingdom*. Cambridge University Press, Cambridge, UK

Klein R, Day P, Redmayne S 1996 *Managing Scarcity: Priority Setting and Rationing in the National Health Service*. Open University Press, Buckingham, UK

Leichter H M 1999 Oregon's bold experiment: whatever happened to rationing? *Journal of Health Politics Policy and Law* **24**: 147–60

Mechanic D 1979 *Future Issues in Health Care: Social Policy and the Rationing of Medical Services*. Free Press, New York

Mechanic D 1995 Dilemmas in rationing health care services: the case for implicit rationing. *British Medical Journal* **310**: 1655–9

Mechanic D 1997 Muddling through elegantly: finding the proper balance in rationing. *Health Affairs* **16**: 83–92

New B (ed.) 1997 *Rationing: Talk and Action in Health Care*. BMJ Publishing Group, London

Rodwin M 1993 *Medicine, Money, and Morals*. Oxford University Press, New York

Symposium, The Law and Policy of Health Care Rationing 1992 *University of Pennsylvania Law Review* **140**: 1505–2095

D. Mechanic

Health Care Systems: Comparative

1. Introduction

Each country's health care system is contained within a singular configuration of culture, history, politics, and economic capacity, and from these derive a specific mode of financing, delivering, and even evaluating health care. Yet, despite the uniqueness of each health care system, all are experiencing similar exogenous pressures, albeit to different degrees, arising from unprecedented developments in science and technology, demographic distributions, and patterns of disease, to name a few. And, despite their differences, many countries are responding to these pressures in remarkably similar ways. As Brown (1998, p. 36) puts it, 'Everywhere one hears the weary refrain: "We can no longer afford the [health care] system we have."'

To some scholars these similarities in pressures and responses suggest an impending convergence in health care systems, defined as 'a certain macro process in which a narrowing of system options takes place' (Mechanic and Rochefort 1996, p. 242). Mechanic and Rochefort emphasize, however, that convergence does not imply a blurring of social, cultural, and ideological distinctions among nations. Indeed, each nation's response necessarily builds on 'preexisting social institutions and professional organizations' (1996, p. 243). A number of scholars (cited throughout this article) have identified a specific kind of convergence in health care: many nations are turning to some type of market reform in order to control expenditures. This trend entails an unusual paradox: many of the ideas on market reform emanate from the United States, and yet the United States remains, not just an 'exception' among advanced industrial societies in its commitment to a welfare state in general and to health care in particular (Ruggie 1996), it is also a pariah from the perspective of universal coverage. At present, over 44 million Americans—approximately 18 percent of the population—are without health insurance, and millions more are underinsured. Adding to the paradox of the convergence, it is not altogether clear that the new methods offer a long-term solution to the problem of rising health care costs (Light 1997). For instance, an increase in administrative expenses is frequently as-

sociated with the need to assess competitive bids and manage the varieties of insurers and providers in a competitive market. In addition, concern about the consequences of market-based reforms on equity and access may already be diffusing, if not halting, recent innovations.

These inconsistencies in processes and outcomes suggest that the march toward marketization and the convergence it implies may be more apparent than real (OECD 1994, Gray 1998). Jacobs (1998, p. 5), among others, observes that while marketization connotes a direction of change and implies a choice of policy instruments, empirically it can become a 'highly eclectic venture.' His study shows that reforms—which introduced provider competition, greater cost-consciousness, and more patient choice—in the United Kingdom, Sweden, and the Netherlands actually resulted in divergence. Contrasting goals among political actors embedded in different institutional systems led to outcomes that were more significant for their dissimilarities than their resemblances.

Focusing on data from the 29 OECD countries and experiences from selected cases, this article explores the convergence/divergence debate by examining some of the main changes occurring in health care systems at present and some of the key issues driving change. It argues that convergence may be occurring at the level of interest among governments in experimenting with market-based forms of organizing health care, but whether this development constitutes what Max Weber would call a 'master trend' (a dominant movement in social change), sufficient to coalesce other distinctions, is far from clear. For an equally important concurrence is the resort on the part of governments to various types of regulation to mitigate the impact of market mechanisms. Hence, what we are witnessing is an oxymoron of modern society first identified by Karl Polanyi decades ago: planned markets. In general, this thesis also captures the experience of developing countries beyond the OECD, many of which are trying to introduce market reforms, and not only in health care. However, since planned markets require a relatively strong state, developing countries may be hindered in their capacity to harness viable market-based innovations.

2. *Financing and Spending: An Overview*

Health care systems are most commonly categorized according to three main methods of financing: general taxation, compulsory contributions to social insurance, and voluntary private insurance or direct payment (see Table 1). Health care systems funded through taxation tend to achieve the highest degree of universal coverage and, in principle, equity in access and provision. Within each category, however, significant differences occur in the delivery of health care services and these may affect patterns of coverage.

For example, Britain, Canada, Denmark, and Sweden all raise funds for health care through general taxation. But Britain considers its system to be a national health service, a concept more commonly associated with socialist countries. The other countries have national health insurance systems. The main difference is that under a national health service, ownership and/or control of the factors of production is public. Ownership of the factors of production in national insurance systems varies, but in general most hospitals are in the public domain whereas physicians consider themselves to be independently employed. Another difference pertains to the degree of centralization. In Canada, the national insurance system is funded by a combination of federal and provincial taxes. Also, the Canadian system is administered by the provinces, the Danish and Swedish systems by counties. While adminstration of the British national health service is decentralized, overall priorities emanate from the central government insofar as local governments have no independent source of revenue.

National or social health insurance systems are much more frequent than national health service systems. In most national health insurance systems, unlike in Canada and Scandinavia, financing is based on employer and employee contributions to an insurance fund (Belgium, France, Germany, the Netherlands, Norway); these contributions are compulsory for all employers and many workers. The funds may be administered privately (this is most likely if the funds are closely tied to employer or employee organizations) or publicly. Governments sometimes contribute to the funds as well and invariably set up some contributory mechanism to provide for special groups (such as low income or unemployed individuals). Social insurance systems generally achieve their universality by being tied to social security systems. Ownership of the factors of production varies considerably, but in general there is more private control than in systems funded through taxation.

Voluntary private health insurance is purchased by either employers or individuals or, in some cases, by a combination of the two. While this system predominates in the United States, it occurs in many other countries as a supplement to the national system. Depending on the country, private insurance may cover either private services only or some portion of public services as well.

Most countries house some mix of these models. For instance, in New Zealand hospital and specialist services are funded through general taxation, whereas financial support for ambulatory services and pharmaceuticals derives from an earmarked social security fund. In the United States, the Veterans Administration resembles a national health service and is funded through general taxation; Medicare is a national insurance system for those aged 65 and over as well as for the disabled, and is based on compulsory employer and employee contributions; and Medicaid is an

Table 1
Financing methods

General taxation	Compulsory social insurance	Voluntary private insurance
Britain (national)	Belgium	USA
	France	Elsewhere to supplement national system
Canada (federal and provincial)		
Denmark (primarily county)	Germany	
New Zealand (hospital and specialist, national)	Netherlands	
	New Zealand (ambulatory and pharmaceutical)	
Sweden (county)	Norway	
USA (VA—federal; Medicaid—state and federal)	USA (Medicare)	

insurance program for the poor funded through general taxation.

Some recent efforts to control rising costs have tinkered at the margins of basic financial arrangements, but none has significantly changed a country's distinctive modality. Many cost-control efforts have sought to increase the contribution of private payments to overall expenditures. Even though all advanced industrial societies except the United States have 'national' systems, there is a considerable range in the public/private mix (Ruggie 1996). The term privatization suggests a direction of change away from the state, but actual cases of privatization demonstrate that the process is complex and does not necessarily shed government responsibilities.

Most countries have encouraged an increase in the role of private insurance by, for instance, granting a tax deduction to employers or individuals (Australia, Britain) or reducing the amount of public provision of certain services, leaving the private sector to fill the gap (Canada). Most countries have also increased charges to patients each time they use certain services (including hospital care in Sweden and general practitioners in New Zealand), but these additions tend to be nominal and aimed at enhancing cost consciousness and deterring frivolous or unnecessary use. Efforts to shift financial responsibility invariably introduce not only consumerism and commercialism but also stratification and selectivity. For instance, private insurance companies may discriminate by health status and social characteristics (such as age and class), charging certain individuals more while offering a greater range of services to those who can pay higher premiums. As a result, new government regulations have been deployed to correct these inequities.

For their part, private insurance companies are also engaging in cost-shifting by raising the price of premiums, deductibles, copayments, and so on. Similar increases in recipient contributions to social insurance funds have occurred as well. Some social insurance systems have attempted to increase cost consciousness by offering consumers greater choice among plans (Germany, the Netherlands, and Sweden). While this fundamental market principle—choice—has increased competition within health care systems, its impact on costs is dubious because, unlike the predictions of market models of rationality, patients are not inclined to choose less care.

It is difficult to assess the effect of these privatization measures on the overall burden of health care expenditures, because they rarely appear alone. Also, while privatization by definition connotes some form of marketization, what exactly occurs is too variable and vague to be analytically useful. To confuse the issue even more, marketization can occur without privatization (as in Britain). However, just as privatization may in the end not entail a decline in the role of government, so too marketization is sometimes coupled with greater government regulations. Far from being contradictory, this combination may well provide the key to achieving the new triad of goals in health care: efficiency, effectiveness, and equity.

3. *Factors of Production: State vs. Market*

There is general agreement among scholars in the field of health policy that market mechanisms alone are not appropriate in matters of health care. Not only are consumers driven by need and not want, but when in need consumers do not exercise rational choice and prefer to spend to their limits (Evans 1999). However, few consumers are free actors in health care: governments and other third-party insurers buffer individual decisions and indeed shape them. What exists in all health care systems, then, are degrees of marketization and regulation, with pendulum-like swings between them reflecting the inevitable difficulty of staying too long on one side. That we are currently witnessing a (re)turn to market-based mechanisms indicates certain shortcomings in government controls, to be sure. But we ought not expect the market to replace the state, simply to stimulate correctives. Markets in health care, and especially those introduced by governments,

Table 2
Physician reimbursement

Fee-for-service	Salary	Capitation
Belgium	Britain (hospital specialists)	Britain (GPs)
Canada	Norway (some municipalities and counties)	Denmark ($\frac{1}{2}$ GP income)
Denmark (specialists and $\frac{1}{2}$ GP income)	USA (some HMOs)	Netherlands (GPs and for lower-income patients)
France (ambulatory)	Elsewhere in public hospitals and health centers	
Germany		
Netherlands (for higher income patients)		
New Zealand (ambulatory)		
Norway (ambulatory)		
Sweden		
USA (some Medicaid and Medicare, and much private insurance)		

are bound to be 'quasi' at best and reflective of the dominant social values and norms implicit in their construction.

Market forces have been introduced in various areas of the delivery and provision of health care. However, either the introduction of market reforms has been conducted within the framework of government regulations, especially as these pertain to issues of equity and access, or government regulations have been reintroduced when the negative consequences of market mechanisms on equity and access have exceeded acceptable norms.

3.1 Physicans

Of the three main methods of reimbursing physicians—fee-for-service, salary, and capitation—fee-for-service is the most open-ended and has been the most difficult to control. Governments (as payers) have attempted either to limit overall increases in fee schedules (as in Canada) or to specify prospective rates for procedures (as in the American Medicare Fee Schedule) (Table 2). In countries with social insurance systems, efforts to exercise these controls are frequently encumbered by corporatist negotiations involving representatives of employers, workers, and physicians. Private insurance companies have found it very difficult to impose reimbursement controls or develop other disincentives to increased utilization of physician services under fee-for-service plans, perhaps because the main competitive edge of private insurance is the greater choice it offers consumers who are willing to pay for higher-priced physician services. Because fee-for-service reimbursement controls encourage physicians to increase the number of patient visits in order to maintain their level of income, governments in Canada and Germany have begun to monitor the volume of physician services.

Physicians tend to be salaried when they work in public hospitals, health centers, or certain types of health maintenance organization (HMO). While salaries are inherently easier to control, they may elicit reduced service provision, especially when specialists also have private fee-for-service practices outside the health care facility. Capitation payments (a specified annual or monthly amount for each patient) commonly accrue to primary care physicians or general practitioners. Although there are outer limits on the number of patients on a physician's list, in order to control costs there must be strong disincentives to refer patients to specialist physicians or services.

Greater control over physician reimbursement has been facilitated by the introduction of managed care. Managed care is also responsible for ushering in a shift away from fee-for-service reimbursement toward either modified fees (with prospective fee schedules of some sort) or increased capitation if not salaried reimbursement within a physician's portfolio. Managed care plans rely on primary care physicians to serve as gatekeepers to the system of care, referring patients only when they deem it necessary. Various incentives exist to reward physicians for their efforts, ranging from bonuses in American HMOs to managerial independence among fundholding GPs in Britain. Fundholding GPs are allotted a lump sum payment for their practices and allowed to make all expenditure decisions themselves. Developments such as GP fundholding also represent an attempt to increase provider cost consciousness by making providers purchasers of health care with significant budgetary discretion.

Managed care has also introduced increased forms of competition among providers. Some managed care organizations negotiate contracts with selected providers for their services, encouraging providers to devise attractive offers that promote the goal of 'value

for money.' Some managed care organizations simply enlist providers who are willing to offer their services for a discount; physicians are motivated to participate in these plans because of the promise of increased volume. In many countries competition is occurring among providers within the public sector, but, while reimbursements levels are affected, the practice is oriented more toward enhancing cost consciousness than achieving any significant reduction in overall expenditure on physician services.

Physicians and patients in the United States have undergone a dramatic and relatively abrupt transition to managed care, but at the same time, this revolution has been stymied by a shortage of primary care physicians. Unlike other countries the government in the United States exercises only weak control over physician supply and distribution, based entirely on monetary incentives. It is expected that the still growing market for managed care will continue to influence physicians to choose primary care as a field of specialization. Meantime, however, experiments are underway (for example, in New York City) in the use of highly trained advanced nurse practitioners as primary care providers. These nurses perform all of the functions of primary care physicians with regard to prescription, referral, and hospital admitting authority, and they are paid at the same rate. They and their employers insist that they are not threatening the livelihood of physicians, simply filling a gap. But their appearance is part of a wider movement to demote the level of skill in certain jobs. In all countries, most types of health care worker are undergoing a deskilling as hospitals cut middle-range positions while hiring lower-skilled replacements, some of whom work only part-time.

Other controls on physicians aim to standardize what in some cases has been wildly divergent practice patterns (Wennberg 1999). Clinical or practice guidelines, also known as medical audits, contain specifications for how physicians should go about treating certain ailments. In most countries these guidelines are developed by the medical profession and monitoring is based on peer reviews, but in other countries (such as the United States) government officials or other interested parties have a hand in the process as well. As protocols for treatment options, clinical guidelines have a dual purpose, for they are also central to efforts to improve quality assurance. Accordingly, in some places they have been extended to include primary care. Finally, in a number of countries, physicians' expenditures on pharmaceuticals are coming under greater control. Britain and Germany are holding physicians directly responsible for spending over a specified limit.

These and related developments have led to the suggestion that the dominance and authority of the medical profession is declining (Mechanic and Rochefort 1996, Raffel 1997). Certainly, managed care has removed much clinical decision making from the sole domain of the physician and muddied it with administrative rules and budgetary constraints. Further proof derives from the British experience. Having once enjoyed considerable clinical autonomy (especially compared to the United States), physicians in Britain are complaining about the introduction of a planned market with its greater managerial control over medical decision making (Harrison 1995). Alternatively, managed care has enhanced the role of primary care physicians in relation to the secondary sector. Even in countries that already had well-developed primary care services, the incorporation of managed care has strengthened the vertical integration of sectors and placed primary care physicians at the intersection.

There can be little doubt that the heyday of physician dominance over health care is past. But we must look beyond the medical bureaucracy to understand the forces behind this development. For example, better patient knowledge, triggered not only by successful health promotion programs but also by the media and increased access to the Internet, has greatly altered patient–physician relations. Medical authority is also being tested by negative reactions against increasingly more invasive technologies and procedures as patients are turning away from overly medicalized treatments and institutional settings. While we can note change, we ought not jump to conclusions about the endpoint. There is no reason why demedicalization cannot exist alongside a continuation of medical authority. Furthermore, demedicalization has generated greater accountability criteria on the practice of medicine. Accountability requirements have perhaps gone furthest in the United States, where some state and local governments are publishing physican (and hospital) 'report cards,' ostensibly to provide consumers with the information they need to make their health care choices. The status of physicians is undoubtedly coming under greater scrutiny, due as much to marketization and its competitively-based principles as to broader social changes in the culture of health care.

3.2 Hospitals

In most countries hospitals consume the greatest portion of health care budgets. By now, various forms of prospective payment systems to hospitals exist in most countries. Prospective payment systems frequently include additional incentives to contain costs, such as allowing hospitals that underspend to keep the difference and average out their costs. Some countries, such as Germany, have prospective *per diem* rates. Unless they are combined with disincentives to extensive lengths of stay, *per diem* rates do little to hold down hospital expenditures. Many more countries have already adopted or are shifting to prospective per case rates based on the American Diagnostic-Related Group (DRG) system. In this system each of several

hundred categories of diseases and disorders is assigned a numerical value based on service utilization that is then translated into a reimbursement level. Most hospitals responded to the incentives within the DRG system to contain costs by reducing lengths of stay (for example, in Sweden the decline was from an average of 21.3 days in 1985 to 7.8 days in 1995, OECD 1998). The number of hospital beds per population also decreased as a result. However, both of these indicators of reduced hospitalization have occurred in all OECD countries except Korea and Mexico, whether or not countries adopted the DRG system or similar cost containment methods, indicating that a more generalized decline may be occurring in the role of the hospital.

Another prospective payment system, used for years in Britain and Canada, and now being widely adopted elsewhere (France, the Netherlands, Sweden, Finland) is global budgeting on an annual basis. Spending decisions are usually decentralized to the local and even hospital level. In some countries there are strict controls on public sector capital investment in hospitals, sometimes with separate budgeting mechanisms.

A novel form of independent decision making by hospitals has occurred in Britain where hospitals are now self-governing trusts. The move was prompted by the expectation that hospital managers would become more cost conscious if they were responsible for all decisions—purchasing, servicing, delivering, administering, and even paying specialist physicians. Hospitals advertise their services and compete for contracts with, among others, fundholding GPs and local authorities. Similar developments have occurred in Italy and Denmark.

The introduction of cost control measures in hospitals has raised concern about declines in intensity of service and concomitant quality of care. Because of the great diversity among countries in the size of the hospital sector, goals regarding efficiency differ. Countries that are attempting to reduce total hospital output are faced with the challenge of closing hospitals or at least beds that are not being used efficiently (US, Britain). These countries nevertheless join others in striving to encourage hospitals to produce more so as to achieve certain goals, such as reducing waiting times for nonemergency treatments. The ability to regulate hospital costs goes beyond administration, however, for what happens in hospitals is highly dependent on technological advances. Yet, everywhere technology is undergoing contradictory developments.

3.3 *Technology: From High to Low*

While cause–effect relationships are always difficult to impute, it is safe to say that hospital cost containment has been made possible because of new technologies that have both reduced the risks of certain surgeries and enabled an increase in outpatient surgeries. Paradoxically, rather than expanding megacenters of health care, as in the past, technological innovations may be contributing to the opposite effect. For instance, with the ability to perform more office-based procedures, physicians may be spending less time in traditional hospital settings.

At the same time, reactions against certain technological developments have engendered a number of nonmarket forms of cost containment and a search for less invasive and more personal forms of care, all of which signal new conceptions of what constitutes quality in health care. For example, there has been an exponential increase in the home health industry, which in many countries is contained within the private sector.

In some countries there is also renewed interest in supplementing both primary and secondary care with community care facilities. These provide a range of caregiving services either at home or in clinics, or day centers not only for those recovering from hospital treatment but also for the chronically ill, the mentally ill, and the elderly. Palliative care facilities and hospices are also proliferating to meet the demand for more personal and less technologized care for the dying.

Community care goes hand in hand with a renewed emphasis on preventive medicine, which was first heralded as a cost containment mechanism by the early health maintenance organizations. Insofar as community care, preventive medicine, and related alternative forms of health and healing entail a greater role for patients, both in self care as well as in determining the course of treatment, they also represent a demedicalization of sorts. Nevertheless, as with other evolving relations in health care, we may be witnessing less a replacement of allopathic medicine than a complement to standard modalities.

4. *Managing Markets in Health Care*

An American economist, Alain Enthoven, first introduced the idea of 'managed competition' in health care in 1978. For the American context emphasis was put on the term 'managed,' which carried a dual meaning. First, Enthoven was as much an advocate for the development of managed care as for the competition that he felt should occur among managed care organizations based on issues of quality and cost. As his formulations have been imported into other countries, this emphasis has shifted somewhat, but not entirely. Countries with strong primary health care sectors—whether these have been based on general practitioners or community care clinics, or both—have always had the foundations for managed care. For these countries competition among primary, secondary, and tertiary

care providers has been more important for enhancing cost consciousness than reducing health care expenditures.

Enthoven also suggested that competition itself should be managed, that is to say regulated, by public agencies who would act as sponsors for consumers. Each sponsor was to be: 'an active, intelligent collective purchasing agent on the demand side… that creates, develops, administers and enforces the rules of competition in a never-ending effort to root out market failures and to perfect the market' (1994, p. 1416). Countries with systems of financing based on social insurance funds already had the foundations of such sponsorship. The incorporation of managed competition in these countries (Germany, the Netherlands, New Zealand) has primarily enhanced patient choice of fund, although some selective contracting by the funds has also occurred. The consequent increased competition among insurers has raised some predictable problems. In Germany the federal government has introduced risk adjustment schemes to compensate funds that have been negatively affected by the infusion of new patients. Some observers suggest that the equalization and standardization that is resulting may lead Germany toward a single national sickness fund—an odd consequence of managed competition (Freeman 1998). In other words, as new methods of competition are introduced into health care systems, their potentially unsettling consequences are eventually, if not immediately, controlled by new regulatory mechanisms. Even in the United States, proposals for a national Patients' Bill of Rights promise to reestablish fundamental criteria of justice over the exigencies of competition. Everywhere the result has been to tame market externalities without completely debilitating market mechanisms.

5. *Policies, Politics, and People*

Given the pressure to expand medical services and the reality of other demands on strained government budgets, countries have been struggling with the dilemma of trying to maintain full and equitable access to high quality health care while at the same time controlling cost increases if not actually cutting expenditure. During the 1980s and 1990s, most countries experimented with some form of market mechanism in order to balance the equation. Interestingly, the effort to contain costs has occurred regardless of level of expenditure. Among those OECD countries actively engaged in health care reform, expenditures on health as a percentage of GDP range from 6.4 in Denmark and 6.9 in Britain to 10.4 in Germany and 14.2 in the United States (OECD 1998). The rate of growth in expenditures has slowed in a number of countries. Many scholars attribute this consequence to the diffusion of a single idea—that the market is best suited to achieve efficiency. Because marketization has occurred regardless of the degree of state vs. market penetration across diverse national settings, some scholars have suggested that a convergence in health care systems is occurring. To some extent countries have accepted new policy instruments cautiously and rationally, learning from each others' successes and mistakes. But to a larger extent as countries have borrowed ideas and instruments, and adapted these to their own situations, the inevitable tensions between market and state, state and society, professions and politics, values and rationalities have given full play to the rich diversities among health care systems. The current swing toward market solutions to economic problems in health care is limited in its penetration of health care systems, for few countries tolerate for long the negative consequences of market forces on equality and access. Indeed, 'planned markets appear as an attempt to find more effective administrative instruments, rather than as a conscious or intended surrender of political control over the main objectives of health policy to market forces' (Saltman and Otter 1995, p. 10).

In the end we must ask, however, whether we are witnessing actual reform of health care systems or something less tangible. If markets are constantly subject to government control, perhaps the underlying goal of reform is cost consciousness and the efficiency it purportedly entails. But beyond the buzzwords, a more profound eventuality is in the offing. If actors are expected to become more aware of the economic consequences of their decisions to utilize health care services, they must also become more attuned to their willingness to forego medical options, for themselves and for others. That health care must be rationed is probably the most difficult realization ecountered by governments and citizens. It rests on the premise that health care is a scarce or at least finite resource, a premise that is hotly debated among medical professionals and ethicists. But as the discussion above has indicated, implicit forms of rationing abound—inability to pay; primary care practitioner 'gatekeeping'—limited number of professionals; limited number of technologies; queues for treatment–to name but a few. It is unclear how fully these practices are understood to be forms of rationing. Few governments have confronted the issue head on and developed explicit protocols.

The state of Oregon in the United States has one of the only plans specifically denying coverage for certain treatments. But the plan applies only to a population unable to purchase private insurance—Medicaid (poor) recipients. No matter what type of rationing exists, someone has to make the decision about denying access. How a country goes about this task reveals as much about prevailing values and norms as do decisions about financing. The Oregon plan, which on the surface appears to be a Leviathan in a land of individual freedom, was actually the product of considerable grass roots participation. In Britain,

government control over the macrolevel of expenditure and priority setting is coupled with confidential case-by-case rationing decisions made by physicians, a process that relies on an unusual level of trust. The persistence of these cultural features of health care systems explains the incomplete successes of market reforms.

Ultimately, however, the story of market-based reform, and the answer to the questions raised about convergence or divergence, rest on an understanding of the role of political institutions in health care systems. For reform to be effective, governments must have the political will to control the demands and behaviors of all relevant actors. But at the same time, for health care to be adequately and justly distributed, governments must marshall citizen involvement in decision making processes in order to construct and legitimate collective goals in health care. This is a universal point, applicable to newly industrializing countries or countries with struggling economies no less than to those discussed in this article.

See also: Health Care, Rationing of; Health Policy; National Health Care and Insurance Systems; Public Psychiatry and Public Mental Health Systems

Bibliography

Brown L D 1998 Exceptionalism as the rule? US health policy innovation and cross-national learning. *Journal of Health Politics, Policy and Law* **23**(1): 35–51

Enthoven A 1994 On the ideal market structure for third-party purchasing of health care. *Social Science and Medicine* **39**(10): 1413–24

Evans R 1999 Health reform: What 'business' is it of business? In: Drache D, Sullivan T (eds.) *Market Limits in Health Reform: Public Success, Private Failure*. Routledge, London

Freeman R 1998 The German model: the State and the market in health care reform. In: Ranade W (ed.) *Markets and Health Care: A Comparative Analysis*. Longman, London

Gray G 1998 Access to medical care under strain: New pressures in Canada and Australia. *Journal of Health Politics, Policy and Law* **23**(6): 905–47

Harrison S 1995 Clinical autonomy and planned markets: The British case. In: Saltman R B, Otter C von (eds.) *Implementing Planned Markets in Health Care: Balancing Social and Economic Responsibility*. Open University Press, Buckingham, UK

Jacobs A 1998 Seeing difference: market health reform in Europe. *Journal of Health Politics, Policy and Law* **23**(1): 1–33

Light D W 1997 Lessons for the United States: Britain's Experience with Managed Competion. In: Wilkerson J D, Devers K J, Given R S (eds.) *Competitive Managed Care: The Emerging Health Care System*. Jossey-Bass, San Francisco, CA

Mechanic D, Rochefort D A 1996 Comparative Medical Systems. *Annual Review of Sociology* **22**: 239–70

OECD 1994 *The Reform of Health Care Systems: A Review of Seventeen OECD Countries*. Organization for Economic Cooperation and Development, Paris

OECD 1998 *OECD in Figures*. Organization for Economic Cooperation and Development, Paris

Raffel M W 1997 Dominant issues: Convergence, decentralization, competition, health services. In: Raffel M W (ed.) *Health Care and Reform in Industrialized Countries*. Pennsylvania State University Press, PA, pp. 291–301

Ruggie M 1996 *Realignments in the Welfare State: Health Policy in the United States, Britain, and Canada*. Columbia University Press, New York

Saltman R B, Otter C von 1995 Introduction. In: Saltman R B, Otter C von (eds.) *Implementing Planned Markets in Health Care: Balancing Social and Economic Responsibility*. Open University Press, Buckingham, UK

Wennberg J 1999 Understanding geographic variations in health care delivery. *New England Journal of Medicine* **340**(1): 52–3

M. Ruggie

Health Care Technology

1. Definition

According to the US Office of Technology Assessment, healthcare technology consists of 'the drugs, devices, and medical and surgical procedures used in medical care, and the organizational and supportive systems within which such care is provided,' (US Congress 1982) including diagnostic, preventive, therapeutic, rehabilitative, organizational, information, educational, and supportive technologies.

Social scientists routinely distinguish 'advanced' medical technologies from the more mundane healthcare technologies. The distinction between advanced and normal technology often implies that technology has become increasingly central in modern medicine. But in a broader historical perspective, technology has always been central to healthcare, indeed social life. Features of life once popularly regarded as advanced technology, such as patient records or needles, have now been absorbed into routine.

2. The Importance of Healthcare Technology for Social Scientists

The broad definition of healthcare technology encompasses any tool from a mundane shake-down thermometer to the latest version of the computerized patient record, presenting the healthcare field as layered with technologies. Social scientists have been mainly interested in the upper layers of the medical geology: they have written extensively about advanced

life-supporting technologies, organ transplantation innovations, reproductive advances, information technologies, and genetic breakthroughs that pose profound ethical and sociological questions about equity. The introduction of new biomedical technologies not only maps medical advances but also reflects changes in cultural understandings of health, illness, life, death, the body, and even the family (e.g., when reproductive technologies offer an alternative to sexual procreation), religion, politics, crime, nature, race, and gender.

A good example of how medical technologies influence cultural notions is the institution of brain death criteria by the Harvard Ad Hoc Committee on Death in 1968. In an archival study of the committee's correspondence, Mita Giacomini confirms the general suspicion that a significant factor tailoring the redefinition of death were the interests of the transplant community (Giacomini 1997). In 1967, after the first heart transplant in South Africa, the general public had become sensitized to the potential conflict of interest of transplant surgeons to declare patients dead before their heart had stopped beating. Up until that time, physicians and lay people had been unable to come up with criteria for death beyond the century old criterium of the heartbeat cessation. The tacit consensus seemed to be to view death as a bedside physician prerogative. The transplantation community needed more flexibility to receive timely access to comatose bodies and was actively involved in the creation of the new criteria. For example, clinical criteria were adapted to facilitate organ procurement. Even the term 'brain dead' rather than 'irreversible coma' fit with the transplant community's needs because only 'dead' people could serve as cadaver organ donors. In addition to the interests of the transplant community, the redefinition of death was influenced by the emergence of the electroencephalograph and the application of resuscitative and life support technologies.

While most sociological writings have had analytical purposes, some scholars have entered the field of health care technologies with interventionist intentions. The purpose of such socially engaged studies by feminists and other patient advocates is to influence health care policy. In their pioneering study of dying in hospitals, Barney Glaser and Anselm Strauss discussed how technological advances at the end of life created complicated lingering dying trajectories, posing special problems of information management for nurses and physicians (Glaser and Strauss 1965). Glaser and Strauss painted a poignant picture of dying patients and their relatives engaged in a painful game of extracting information from healthcare providers reluctant to reveal the terminal condition. As part of a general death awareness social movement which swept the United States in the 1960s, this study contributed to a revision of the medical school curriculum to include doctor–patient communication courses. Similar studies discussed how the rapid introduction of pharmaceutical, surgical, and technological innovations during the post-Second World War decades featured in the socialization process of individual physicians.

3. *Different Conceptualizations of Healthcare Technology*

Social scientists have viewed healthcare technology in at least three different ways, each of which comes with a number of assumptions about the role of technology, human actors, and the social world. At stake in the different conceptualizations is the autonomy of the technology and its power to shape social relationships. These different approaches greatly impact the critical potential of the analysis.

3.1 *Social Science with Technology Perspective*

In the first conceptualization, technology is not the central focus of the analysis but more a catalyzing variable in a network of social relationships. Healthcare technology remains unproblematized, assumed in the same way that most sociologists take the physical layout of hospitals for granted. What matters is how humans (as professionals, relatives, patients) cope with the new possibilities generated by an influx of instruments in technology-intensive medical settings. The primary goal of social scientists working in this approach is to analyze well-established topics of interests to observers of health and illness, that is, professionalization, social equality, social control, or medicalization, often relying upon the canon of a broad social constructivist perspective.

A good example of this approach is Renée Anspach's study of decision-making about life and death in a neonatal intensive care unit (Anspach 1993). She situates her ethnography in a technology-intensive environment and states that most of the dilemmas about keeping premature and sick infants alive result from the availability of the technology and expectations of staff and relatives. Yet, she does not name or describe what technology she is talking about, discuss its origins, or elaborate on the staff's work to keep the technology functioning and interpretive. Instead, Anspach writes about the neonatal intensive care unit as a complex social organization where different groups of people rely on diverse sources of information and weigh varying priorities to reach a decision about the continuation of aggressive treatment. Physicians, for example, rely more on technology-derived data than nurses who depend on closer, interactive contact with infants, leading to conflicts in life-and-death decisions. The healthcare technology confirms the power differences between an emerging and established profession.

Even scholars who make technology more central often rely on this socially determined conceptualization of medical technology. Susan Bell's interactive diffusion model of the artificial hormone DES (diethylstilbestrol) applies a social constructivist social network approach to healthcare technology (Bell 1989). Bell shows how the artifical hormone originated from within scientific and health communities with distinctive ways of defining problems and offering solutions. A broader political-economic context (US research funding post-Second World War) structured the relationships of these communities and reflected the way the technology diffused among physicians and their women patients. In this conceptualization, the focus of the social scientist's analysis remains with the social networks advancing the new technology. Bell explores how social communities gain or lose power because of their support for DES. Yet, how the meaning of the hormone itself changes as a result of those interactions remains unexplored.

The main assumption underlying this approach to healthcare technology is that the tools and instruments of medicine are shaped by social forces. Technology is a given (Anspach 1993), or a contingently diffused product following a trajectory set by social circumstances (Bell 1989). Dominant social groups invest values, power, and preferences in the healthcare technology, and it is these norms and paradigms that dominate the analysis. This approach offers interesting insights into the *social* aspects of healthcare technology but often glosses over the *technological* characteristics of medical technology. In other words, what makes health-care technology such a unique case, is often left unexamined.

3.2 Humanistic Social Science Argues Against Technology

The second conceptualization of healthcare technology originates from a humanist concern about the dangers of technology and is primarily intended as a critique of the healthcare field. In contrast with the first approach, the spotlight of the analysis is firmly centered on healthcare technology. Healthcare technology is singled out, its effects are scrutinized, compared to a technology-free past, and usually found wanting. Healthcare technology symbolizes here the power and arrogance of the medical professions, the hegemony of the medico-industrial complex, and the excesses of Western modernity. Technology becomes a modernist instrument by which some in society enforce power over others.

Authors working in this tradition emphasize how the presence of medical technology turns caregivers into technicians and leads to depersonalized and dehumanized care. Technology results in distancing and alienation in patient–care-provider interactions, a loss of dignity and integrity, and a colonization of the body. This kind of analysis receives its persuasive power from a historical or cross-cultural comparison. The technology-saturated present is compared with idealized tropes of the 'natural' (the natural body, birth, or death) of past epochs and of others, unpolluted by modern healthcare technology.

This antitechnology analysis has been fueled by the Luddite legacy and the threat of a nuclear apocalypse, and is popular with both conservative scholars lamenting the moral, spiritual, and existential crisis in the secular West (Aries 1977) and Marxist scholars who analyze healthcare technology as an integral part of the capitalist profit structure (Waitzkin 1979).These politically diverse analyses of healthcare technology have in common that they believe that technology is a powerful force shaping social relationships. This belief has been referred to as *technological determinism*. Technological determinism comes in different strengths (Winner 1980): strong technological determinists believe that technology develops as the sole result of an internal dynamic and molds society to fit its patterns. In his book *Medical Nemesis*, for example, libertarian social critic Ivan Illich views the exportation of Western medical technologies as part of an international conspiracy to spread illness across the world and establish Western medical hegemony. According to Illich, international collaboration to eradicate virus-carrying mosquitoes resulted in short-term pyrrhic victories, in the long-term technological inference led to more virulent strains of tropical diseases and an aggravation of malnutrition (Illich 1976, Chap. 7). In a weaker version, medical technology is viewed as a political phenomenon in itself. Technology as politics occurs when a particular technological device or system becomes a way of settling an issue in a particular community or when technologies require particular kinds of political relationships. For example, physician-generated demand for expensive diagnostic technologies such as MRI scanners led to the consolidation and integration of large medical centers at the expense of many local community hospitals that did not have the patient load to justify the large capital investment. In both versions, the focus of the analysis is on what has been lost after the introduction of medical technologies. The purpose of these critiques is to warn of over-technologization and to recapture a mystified technology impoverished past.

This conceptualization of healthcare technology often works from an analysis of extremes and is therefore limited in the analytical insights it brings in the everyday workings of healthcare technology. Yet, because of its strong rhetorical packaging, this approach to technology has often lead to more social mobilization and actual social change than the other two conceptualizations. One can detect technological determinist tenets in the hospice movement, rallying for palliative care for terminally ill people and in the 'natural' birth movement, pointing out the physical

dangers of an overmedicalized birth. In both social movements, antitechnology sentiments united activists and led to concrete social change.

3.3 Technological Practice

A third conceptualization of healthcare technology finds a middle ground between the view that an interplay of social structure and culture shapes the use of technology and the belief that technology determines society. Drawing from the interdisciplinary field of science and technology studies, researchers have turned to what specifically is gained and lost in medical encounters when new instruments and drugs are introduced. This approach led to a number of detailed studies of the practice of healthcare technology in which the technology takes center stage. Healthcare technology is here not an *a priori* established entity but both the technological elements and the social characteristics emerge during the design process and while the technology is used. Social and technological characteristics are continuously reconfigured in interaction, to the point that it is difficult to tell where technology ends and the social begins.

As a prime example of this approach, sociologist Marc Berg addressed whether medicine has become more rational and scientific since the Second World War because of its reliance on formal tools such as standardized protocols, expert systems, clinical decision analysis, and automated decision systems (Berg 1997). Berg neither supported the Luddite camp which argues that such technologies primarily lead to deskilling nor bolstered the protechnology argument that the formal tools necessarily reduce the staff's workload, but instead showed how technologies have shifted medicine's understanding of rationality. Rationality was not a pregiven quality but achieved in tandem with the production and implementation of formal tools. Berg analyzed how the designers of the automated clinical decision analysis systems and protocols went beyond solving technical conundrums. The designers of formal tools studied the work activities and places in which their tools needed to function and made assumptions about users, environment, input, and expected output. For example, the designers of an automated cardiac decision-making system aimed for a computer that calculated a diagnosis and suggested treatment options after a healthcare provider entered diagnostic data about chest pain. The designers prescribed medical practice with guidelines (the algorithms of the diagnostic system), prerequisites (which data is mandatory or optional), and a reshuffling of existing professional relationships to make their devices work (not only physicians but nurses could enter data). In the end, the automated decision-making system provided a diagnosis and the tool worked, but not in the way the designers intended it because the tool's requirements became just one element in the management of patients' medical conditions. For example, nurses quickly learned to enter the 'right' data based on the availability of beds in the intensive care unit. Medical practice and tools emerged simultaneously to constitute new forms of knowledges, new challenges, and problems to keep both the tools and the users on track, and ultimately new opportunities for diagnosing and curing.

The most radical conceptual innovation of the science and technology scholars was to extend agency from humans to things. According to the French sociologist Bruno Latour, this ontological extension of human agency occurs through the *delegation* of agency from humans to nonhumans (Latour 1987). Machines and instruments, such as thermostats and tape recorders, perform as agents because they substitute for the actions of people. As Stefan Hirschauer has shown, during surgery, pumps and monitors become a patient's 'externalized organs,' taking over vital functions (Hirschauer 1991). Because a malfunctioning ventilator might mean the end of life, there is no easy distinction between the human and the technical realm. Technology acts similarly to people. This does not mean that humans and nonhumans are equal. Andrew Pickering, for example, draws the line at intentionality; intentions cannot be attributed to nonhumans (Pickering 1995). Some critics have charged that analyzing technology as actors is disempowering, because ultimately humans need to be held accountable. But other researchers have linked the premise of technological practice to a more critical and interventionist agenda (Timmermans 1999).

The key characteristic of this approach is that changes in the social and the technical realm are investigated symmetrically as they enfold over time. Instead of assuming the predominance of social or technological forces in advance, the social scientist studies how technological innovations advance new medical practices and understandings about health and illness.

4. A Selection of Research Topics in Healthcare Technology

4.1 Organ Transplantation Technology

Like ethicists and physicians, social scientists have been particularly puzzled by the Janus face of biomedical technology: the instruments of medicine have the potential to postpone mortality and enhance the quality of life, yet, also possess the possibility of destroying the world as we know it. Nowhere is the Janus face of biotechnology better researched than in the area of the biomedical organ transplant technologies. The transplant of kidneys, hearts, lungs, heart and lungs, liver, and pancreas benefitted from the development of immunosuppressive drugs that counteracted the body's tendency to attack the foreign

organ and, in the case of US kidney transplants, from Medicare funding in 1972. Since the technology moved down a continuum from experimental to therapeutic, its utilization has prolonged and saved lives that would otherwise almost definitely be lost, but it also spawned many unintended, complex ethical, financial, legal, and social problems.

Renée Fox and Judith Swazey point out in a classic study that more than just organs are exchanged during organ transplantation (Fox and Swazey 1978). The technological procedure poses important questions about the corporeality of personal identity, the relationship of a body to its parts, and the boundaries that define social ties. Organ procurers present the donation of cadaver organs or, in the case of kidneys, live organs as the ultimate 'gift of life,' invoking Judeo-Christian values of altruism, generosity, humanitarianism, and freedom of choice. For the recipient, the donated organ might not only be under continuous threat of immunological rejection but might also carry a heavy symbolic load. Because a donated organ can not be reciprocated with another gift, the organ exchange locks two people and their relatives in an unequal relationship, creating what Fox and Swazey term 'the tyranny of the gift' (Fox and Swazey 1978, pp. 20–32). Relatives of donors might insist on developing kin relationships with recipients because a piece of their loved one 'lives on' in another person. For recipients, the reconstruction of the body requires an active reshaping of the self (Sharp 1995).

These social complications of organ donation are exacerbated by the 'scarcity of organs.' In the US, organ donation seems to have leveled off at 5,000 while the demand for organs has topped more than 33,000 and keeps rising (Sharp 1995). With a much higher demand than supply, who is more likely to receive organs? According to the American Medical Association, medical need should prevail. But does medical need mean that the person who is the sickest should get priority or the person who has the best chance of long-term survival? According to social scientists, the chance to receive an organ is positively correlated with socioeconomic status and patient's race (Kutner 1987). African Americans are less likely to be donors, organ procurers, or recipients of organs. In addition, people with history of mental, emotional, or family problems are less likely to receive organs, raising ethical questions about disability when patients with Down's Syndrome were initially refused a position on the organ donation waiting list. Economists have tried to move away from the gift giving ideology towards a market ideology that considers organs as any other commodity to be bought and sold. Several proposals circulating in medical circles propose to pay a financial benefit to the estate of the donor to motivate donation. In other countries and some US states, the governments have tried to enhance organ donation with a system of presumed consent: everyone is presumed to be an organ donor unless documented otherwise.

Anthropologist Margaret Lock's study of organ transplants and brain-death in Japan and the US (Lock 1989) offers an important counterargument against the idea of the 'technological imperative.' The technological imperative implies an inherent force in healthcare technology; once a medical technique is available there is virtually no objection possible for not using it. The concept, introduced by economist Victor Fuchs, has caught on in other studies to explain the quest for more and better technologies and the quick adoption of questionable technologies (Rothman 1997). Lock explained that any technological imperative is not inherent to technology but to the broader culture. Although Japanese physicians and scientists possess all the technical know-how and skills to implement organ transplants, Japanese people have until recently chosen not to take the cadaver organ transplant route but instead focus on the development of artificial organs. The reason for Japan's reluctance to jump the organ transplant bandwagon is that brain death is not accepted as a sufficient criterion of death for religious reasons and because Japanese families nurse their terminally ill and would be reluctant to accept that a breathing body with a heartbeat is 'dead.' In addition, Japanese have a different concept of the self and feel reluctant to just emulate every technological innovation from the West.

4.2 Reproductive Technologies

Another area flush with research attention is the field of reproductive technologies. Different infertility treatments, biotechnologies, and fetal surgeries have opened the categories of parent, child, fetus, human, family, and even population to some renegotiation. Key questions include control over the human body, the manipulation of eggs, sperm, and embryos, and the threat these technologies pose for the identity and legal status of the 'products' created. Reproductive technologies have been largely analyzed from feminist inspired perspectives but the emphasis of the analysis has shifted over time. Initially, feminist scholars looked at reproductive technology as a tool to liberate women because it could sever the link between sexuality and reproduction. A technological fix would erase some of the differences between men and women and create equal sexuality (Firestone 1970).

While reproductive possibilities seem to be more about fulfilling than rejecting the traditional female role, some feminist scholars gaze at new technologies as the next installment of patriarchal exploitation of women's bodies and reproductive processes. Reproductive technologies are viewed as attacking women, a modern incarnation of eugenics. The technologies become 'technical delegates,' tools used to compensate for the perceived deficiencies of its users. Women's bodies are opened up, manipulated, and controlled to be ultimately bypassed. The consequence is an alien-

ation, commodification, and depersonalization of women's bodies (Corea et al. 1985).

More recently, feminists have looked at the same reproductive technologies from a postmodern perspective (Casper and Clarke 1998), giving more agency to women themselves. The postmodern strategies concentrate upon the design and transformation of reproductive bodes. Poststructuralist feminists have rejected the notion that the 'real female body' is passively 'acted upon,' instead preferring to view it as being both inscribed and constituted through reproductive discursive practices and processes. Reproductive technologies generate different kinds of subjectivity. For example, Charis Cussins analyzed how undergoing infertility treatments figures in women's self making and at the same time transforming their body conceptions (Cussins 1998). She located the fertility treatment in women's biography. Before entering the infertility clinic, many women have already a history of trying not to get pregnant using anticonceptiva, followed by a period of increased awareness of what could have gone wrong while trying to conceive. Cussins showed that the objectification, naturalization, and bureaucratization of the women's bodies during pelvic exams, ultrasound, explanatory surgery, and embryology laboratory is associated with particular forms of agency. If the activities in the infertility clinic lead to pregnancy, the patient exercises agency in her active participation of the different forms of objectification. If the treatment cycle has been unsuccessful, women reported their lack of agency or insufficient attention to the specificity of the patient. The consequence of the process in which women render themselves compatible with instruments, drugs, and material surroundings is what Cussins called an 'ontological choreography' with the potential of transforming the women's long-term self-perceptions as potential mothers.

Most feminist writing, both past and recent work, is ultimately aimed at creating awareness of the reproductive technologies cost and benefits for women, a women centered reappropriation of the technologies, or to offer alternatives for social change (see the book *The New Our Bodies*, *Ourselves* (Boston Women's Health Book Collective 1992).

4.3 Technology and Healthcare Costs

With the explosion of healthcare costs worldwide, social scientists have also investigated the financial role of medical technologies in the healthcare system. Although there seems to be a general consensus that technology has contributed to rising healthcare raised costs, exact figures of how much cost has increased or mechanisms of cost raising are not well documented. Researchers at the US Office of Technology Assessment, for example, estimated that new technologies accounts for anywhere between one-third to three-quarters of the cost of hospital care—a range of several billions of dollars (Banta et al. 1981). Others have estimated that healthcare technology takes up 10 to 40 percent of the increase in healthcare expenditures over time, again a very wide range (Neumann and Weinstein 1991).

Problems with calculating the cost of technology include the broad definition of healthcare technology and the scope of health expenditures. Should medical technology be limited to 'big ticket' items or include the entire technical geology? When calculating health expenditures, some researchers take only the capital cost, while other scholars combine the capital cost with more expensive operating costs and increased use of auxiliary services. Other problems include measuring cost saving and balancing health outcomes of life-saving technologies against long-term institutional expenditures. In some studies, all increases not explained by other factors are attributed to technology, while others focus on the cost of treating certain illnesses or the cost of particular technologies (Neumann and Weinstein 1991).

Relying on total healthcare costs, Frans Rutten and Gouke Bonsel note that labor intensive, low-technological interventions such as mental healthcare are two to four times as expensive as the cost of high cost technology, such as kidney dialysis and transplantation, megavolt therapy, CT tomography, cardiac surgery, *in vitro* fertilization, and picture archiving. They also found that even when one isolates technology, it is often the specialized and highly salaried personnel that take up the major part of the total cost. They conclude that 'high cost technology is expensive per item, but that its contribution to total healthcare expenditure is limited' (Rutten and Bonsel 1992). Economist Victor Fuchs calculated that technology contributed 0.6 percentage points of the 8 percent annual rate of increase in health expenditures from 1937 to 1967 (Fuchs 1972).

Pulitzer-prize winner sociologist Paul Starr has also offered a caveat to the reasoning that technology is to blame for financial healthcare woes. Big-ticket technologies such as new imaging machines, organ transplantation, intensive care units, and renal dialysis have undoubtedly brought higher costs 'But, only about a third of the higher levels of spending (in the United States), compared with Canada, reflects the expense of hospital care—and of that, only a portion is due to greater use of technology' (Starr 1994, p. 22). Starr instead suggested that we adopt a systemic view to explain the rising healthcare costs in the US. He located expanding costs in the historical development of a number of incentives for private decision makers to expand and intensify medical services.

The factors singled out to explain the importance of technology in the healthcare field relate to technological characteristics and the wider healthcare field. In the innovation literature, healthcare technologies are considered 'dynamic,' that is, they develop as they are diffused. This means that technological

innovations are regularly introduced before information about effectiveness and cost is established. Social observers also point to the billion-dollar medical industry, the public's seemingly unsatiable appetite for more technology, and varying incentive structures. Countries with stronger centralized government regulation have been able to curtail costs via planning and technology allocation. Countries following more of a market model create a situation where hospitals invest in technologies to remain competitive. The consequence of these incentives is, paradoxically, an underuse of services.

> Consider the case of early detection of breast cancer through the use of mammography [in the US]. With fully utilized mammography machines, a screening mammography examination should cost no more than $55 … But because machines are typically used far beneath capacity, prices run double that amount. With prices so high, many women cannot afford a mammogram … In other words, *because we have too many mammography machines, we have too little breast cancer screening* (Starr 1994, p. 25, emphasis in original)

5. Future Trends

The electronic patient record, telemedicine, artificial organs, and genetics will continue to raise the specter of healthcare technologies in the coming decades. The challenge for social scientists remains to keep up with the multitude of micro and macro changes healthcare technology poses for societies without falling into simplistic forms of technological determinism or social essentialism.

See also: Biomedical Sciences and Technology: History and Sociology; Health Care Delivery Services; Health Care: Legal Issues; History of Technology; Innovation and Technological Change, Economics of; Reproductive Medicine: Ethical Aspects; Technological Determinism; Technological Innovation; Technology and Social Control; Technology, Anthropology of; Technology, Social Construction of

Bibliography

Anspach R 1993 *Deciding Who Lives: Fateful Choices in the Intensive Care Nursery*. University of California Press, Berkeley, CA

Aries P 1977 *The Hour of Our Death*. Allen Lane, London

Banta H D, Behney C J 1981 *Toward Rational Technology in Medicine: Considerations for Health Policy*. Springer, New York

Bell S E 1989 Technology in medicine: Development, diffusion, and health policy. In: Freeman H E, Levin S (eds.) *Handbook of Medical Sociology*, 4th edn. Prentice Hall, New York, pp. 185–204

Berg M 1997 *Rationalizing Medical Work: Decision Support Techniques and Medical Practices*. MIT Press, Cambridge, MA

Casper M J, Clarke A E 1998 Making the pap smear into the 'right tool' for the job: Cervical cancer screening in the USA, circa 1940–95. *Social Studies of Science* **28**(2): 255–90

Boston Women's Health Book Collective 1992 *The New Our Bodies, Ourselves*. Simon and Schuster, New York

Corea G, Duelli R, Klein R 1985 *Man-made Women: How New Reproductive Technologies Affect Women*. Hutchinson, London

Cussins C 1998 Ontological choreography: Agency for women in an infertility clinic. In: Berg M, Mol A (eds.) *Differences in Medicine: Unraveling Practices, Techniques and Bodies*. Duke University Press, Durham, NC, pp. 166–201

Firestone, S 1970 *The Dialectic of Sex: The Case for Feminist Revolution*. William Morrow, New York

Fox R C, Swazey J P 1978 *The Courage to Fail: A Social View of Organ Transplants and Dialysis*. University of Chicago Press, Chicago

Fuchs V R (ed.) 1972 *Essays in the Economics of Health and Medical Care*. Columbia University Press, New York

Giacomini M 1997 A change of heart and a change of mind? Technology and the redefinition of death in 1968. *Social Science and Medicine* **44**(10): 1465–82

Glaser B G, Strauss A L 1965 *Awareness of Dying*. Aldine, Chicago

Hirschauer S 1991 The manufacture of bodies in surgery. *Social Studies of Science* **21**: 279–319

Illich I 1976 *Medical Nemesis*. Random House, London

Kutner N G 1987 Issues in the application of high cost medical technology: The case of organ transplantation. *Journal of Health and Social Behavior* **28**(March): 23–36

Latour B 1987 *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, Cambridge, MA

Lock M 1989 Reaching consensus about death: Heart transplants and cultural identity in Japan. *Society-Society* **13**(1): 15–26

Neumann P J, Weinstein M C 1991 The diffusion of new technologies: Costs and benefits to health care. In: Gelijns A C, Halm E A (eds.) *The Changing Economics of Medical Technology*. National Academy Press, Washington, DC

Pickering A 1995 *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press, Chicago

Rothman D J 1997 *Beginnings Count: The Technological Imperative in American Health Care*. Oxford University Press, Oxford, UK

Rutten F F H, Bonsel G J 1992 High cost technology in health care: A benefit or a burden? *Social Science and Medicine* **35**(4): 567–77

Sharp L A 1995 Organ transplantation as a transformative experience: Anthropological insights into the restructuring of the self. *Medical Anthropology Quarterly* **9**(3): 357–89

Starr P 1994 *The Logic of Health Care Reform: Why and How the President's Plan Will Work*. Whittle, New York

Timmermans S 1999 *Sudden Death and the Myth of CPR*. Temple University Press, Philadelphia

US Congress, Office of Technology Assessment 1982 *Strategies for Medical Technology Assessment*. US Government Printing Office, Washington, DC

Waitzkin H 1979 A Marxian interpretation of the growth and development of coronary care technology. *American Journal of Public Health* **69**(12): 1260–8

Winner L 1980 Do artifacts have politics? *Daedalus* **109**: 121–6

S. Timmermans

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Health Economics

An individual writing an entry summarizing the field of health economics in 1960 would have had difficulty finding much to say. By contrast, an individual writing at the turn of the twenty-first century faces an embarrassment of riches: the field has exploded since the 1960s. There are now several textbooks, as well as a two-volume *Handbook of Health Economics* (Culyer and Newhouse 2000) and two international journals: the *Journal of Health Economics* and *Health Economics*. This article will thus necessarily be selective.

At least four reasons can be identified for the growth in health economics. The first is the enormous expansion of the health-care sector in almost every developed country during the second half of the twentieth century. Between 1960 and 1997 real health-care spending per person in the G-7 countries grew at annual rates between 3.7 percent (the United Kingdom) and 7.4 percent (Japan) (Anderson and Poullier 1999 give figures in nominal US dollars, and I have deflated by the US Gross Domestic Product, GDP, deflator). As a result of this growth, the G-7 countries in 1997 spent in real terms between 4 and 14 times more on health care per person than they spent in 1960. Because their economies did not grow nearly as rapidly, all the countries saw a rise in the share of GDP going to health care. By 1997 that share ranged from 6.7 percent in the UK to 13.6 percent in the USA.

Second, developments in economic theory, especially the growth of information economics, have aided the development of health economics. Third, although this is not really taken up below, the development of computational power has greatly facilitated empirical work by permitting analysis of large administrative databases. Finally, in several countries, especially the USA, the field has been generously supported when compared with most other subfields of economics, although not when compared with most natural or biomedical sciences.

1. *Uncertainty, Medical Insurance, and Medical Care Services*

The seminal article in health economics is generally considered to be by Arrow (1963). In explaining the institutions surrounding the delivery of medical care, Arrow emphasized two related features: the pervasiveness of uncertainty in medical care, and the asymmetric information between the physician and the patient. Uncertainty gave rise to a demand for insurance, but because insurance markets were seriously incomplete, nonmarket institutions evolved to compensate. In particular, medical ethics dictated that physicians were not to act in a narrow, profit-maximizing fashion, and the patient trusted that the physician was acting in the patient's interests. Much of what was to follow in health economics over the next four decades flowed out of these notions.

The subsequent development of information economics in the 1970s stimulated developments in health economics; indeed, medical care was an example used in several of the early seminal articles in information economics (for example, Rothschild and Stiglitz 1976). These articles showed that if buyers in competitive insurance markets knew more about their risk than sellers, a market equilibrium would not exist and hence certain markets would not exist. One necessary condition for lack of equilibrium in the Rothschild–Stiglitz model was that the proportion of bad risks be relatively small, which is descriptively accurate about health insurance. Typically the highest five percent of spenders account for 50 percent or more of resource use (Berk and Monheit 1992). Relaxing the assumption of symmetric information between buyer and seller thus gave a theoretical rationale for Arrow's observation that insurance markets were often incomplete. It also gave a policy rationale for the public sector interventions in health insurance markets—or the direct provision of services by the public sector—that had by then occurred in every developed country.

Zeckhauser (1970) pointed out that health insurance—and risk-spreading arrangements more generally—inevitably created a second-best situation: although insurance offered protection against risk, it did so at the expense of increasing moral hazard. The traditional meaning of moral hazard in the insurance literature was failure of insured individuals to take preventive measures. In the case of medical care, that insured individuals might forgo prevention seemed of less concern because their health, including pain and suffering, was not insured. But medical care insurance typically reimbursed a certain proportion of spending rather than reimburse a lump sum contingent upon a certain event's occurring. Hence, moral hazard came to mean the additional consumption stimulated by the subsidy implied by insurance. Although the additional consumption was often construed as creating a deadweight loss (Pauly 1968), this is not necessarily the case. One could want resources transferred to certain states of the world precisely so that one could consume medical care that would otherwise be unaffordable in those states (DeMeza 1983). As a result, the conventional test of a welfare loss from health insurance is whether a consumer, to insure a given state of the world, would purchase an actuarily fair insurance policy with an appropriate lump-sum reimbursement (Goddeeris 1984). Any willingness to pay on the part of others (i.e., an externality) should also be added.

There is, however, a literature in health economics that rejects the conventional test of a welfare loss from insurance for various reasons. Two prominent reasons given in this literature are that the observed demand curve does not reflect the preferences of an informed consumer (about which more below) or that income should not affect the delivery of medical services

(Williams and Cookson 2000). The applicability of the conventional test thus remains a controversial issue.

The understanding of how the health insurance contract itself affects health care and even health has progressed over several years. Following Zeckhauser, many early US studies focused attention on the extent of the moral hazard introduced by health insurance, or more concretely the demand response of the consumer to more or less generous insurance coverage. Indeed, the RAND Health Insurance Experiment, one of the largest social science projects to date, was in effect a controlled trial that was in part aimed at quantifying the consumer's demand response to varying levels of coverage (Newhouse and the Insurance Experiment Group 1993). (The results of this experiment are covered below.)

In addition to specifying the price to the consumer or the degree of subsidy, the health insurance contract also specifies the terms at which the insurer would reimburse the supplier. (Terms must also be specified when the government is a direct provider.) In many cases the supplier agrees to accept the insurer's terms as payment in full and does not bill the consumer an additional amount; indeed, in some instances, such as in Canada, billing additional amounts to the consumer is illegal.

Numerous studies have showed that the amount paid to the supplier, especially the physician, affects the quantity of services, holding constant the price paid by the consumer or patient (for example, Rice 1983). These studies have given rise to a lengthy debate among health economists about whether supplier-induced demand rendered standard demand theory, which postulated an exogenous set of consumer preferences, inapplicable to medical care.

The dominant empirical evidence purporting to show that supplier-induced demand violated standard theory was that the volume of health-care services often rose when physician fees were reduced. This was sometimes interpreted as physicians' delivering more services in order to maintain their income, and was referred to as the target-income hypothesis. But it subsequently was shown that this behavior could be accommodated by a standard neoclassical model, provided the physician's utility function was assumed to include total income, work effort or leisure, and the degree of inducement (the latter was assumed to be negatively valued) (McGuire and Pauly 1991). Under these conditions, if the cut in fees implied a large enough cut in income, the physician would increase inducement and services delivered would rise, but if the reduction in fees had only a modest effect on income, services would fall. (In the former case an income effect would dominate a substitution effect, and conversely in the latter case.) This modification, although potentially leaving the construct of a demand curve intact from the point of view of positive economics, left the applicability of conventional welfare economics on shaky ground.

The role of the physician as an agent for the patient gave rise to a literature that emphasized altering financial incentives facing physicians to improve the trade-off between risk spreading and moral hazard. Specifically, it was proposed to put the physician (or other supplier such as a hospital) at some risk for the quantity of services used by a group of patients (Ellis and McGuire 1986). Fee-for-service reimbursement of the physician, with fees set in an administrative transaction between the insurer and the physician, was thought to contain economic rents in many cases, with a resulting incentive for the physician to overprescribe services and procedures. For if fees were set at marginal cost there would be no incentive to overprescribe, and no concern about supplier-induced demand (Pauly 1980). Thus, the use of two-part pricing, wherein the physician (or other supplier) received a lump sum for each enrollee—a capitation—as well as a diminished fee, could in principle reduce the moral hazard from rents in fees without increasing the financial risk borne by the patient. (But some cost sharing on the demand side would still presumably be useful to address the patient's incentive to seek care in the first place.) Of course, if the consumer were well informed and could monitor the services delivered, the basis of payment to the physician, capitation, or fee-for-service, would not matter, but this was a strong assumption. Indeed, what the consumer often wanted from the physician was diagnostic information. The principle of a two-part payment could be extended to other kinds of services, for example paying a lump sum for each hospital admission, with perhaps additional payments for very expensive cases, rather than a varying amount for each admission, depending on the exact services delivered. This latter example shows that the fee portion of the payment could in principle be paid in nonlinear fashion.

The importance of pricing in affecting supplier behavior has blossomed in the USA with the development of managed care (Glied 2000). Health-maintenance organizations, which provide all necessary medical services in return for a monthly premium, had existed since the 1930s. Until the 1980s, however, they had a small market share (under 10 percent). But as medical capabilities and the resulting costs of health care rose, the scale of moral hazard and potential welfare loss of the then dominant insurance form rose concomitantly. Probably as a result, starting in the mid-1980s managed care grew to dominate the US market, in part because it was seen as a set of techniques to reduce the moral hazard without increasing the financial risk on the consumer.

Managed care consists of two broad techniques. The first is financial incentives to physicians, usually groups of physicians, to reduce services. A common form of such incentives is capitation, but there are also various kinds of risk-sharing arrangements with payment according to whether utilization is over or under a target. The second technique is command-and-

control-type regulation, such as utilization review or prior authorization of hospitalization, in order to reduce the delivery of low-valued services. Baumgardner (1991) showed how, as the technical capabilities of medicine increased, command-and-control techniques could become more valuable in addressing moral hazard.

But capitation had a downside as well. Because in practice the same capitation amount was paid for patients whose expected costs varied, there was an incentive to select good risks and avoid bad risks. This issue has been the subject of much analysis (Van de Ven and Ellis 2000). Some of the analysis stressed situations in which regulation prohibited charging different risks different amounts even when information was symmetric. Other analysis stressed asymmetric information and the attraction of sicker individuals to more generous benefits even if insurance plans made no active effort to discriminate among risks. The latter point of view gave rise to a literature on risk adjustment, meaning that an external party (government, an employer, a coalition of employers) would alter premiums facing consumers choosing among plans in accordance with the risk mix of their enrollees (Newhouse 1996, Van de Ven and Ellis 2000). After risk adjustment, plans that had a disproportionate number of bad risks would not be more expensive to the consumer than other plans.

The Rothschild–Stiglitz model of selection in insurance markets, which assumes asymmetric information, has been influential, but makes two predictions that do not correspond with experience. First, it predicts no pooling equilibrium (meaning heterogeneous risks purchase the same insurance plan). Second, it predicts that high-risk individuals can obtain the insurance they want, whereas low-risk individuals cannot. Real-world insurance markets obviously exhibit some pooling; furthermore, it is typically high-risk rather than low-risk individuals who have difficulty obtaining as much insurance as they want in competitive insurance markets. Newhouse (1996) showed that if it is costly for insurers to discriminate among varying risks or persons with varying expected costs, there will be some pooling of risks (there will be more pooling with higher costs of discriminating). Moreover, under these conditions it is high-risk rather than low-risk individuals who will not be able to buy all the insurance they wish. Although these predictions appear to accord better with reality, it is not clear that the magnitude of costs to insurers to discriminate among different risks is large enough to carry the burden assigned to them. Theoretical understanding of insurance markets remains imperfect.

If insurers (or health plans) compete for individual members, as is observed in the USA and in certain other countries (e.g., The Netherlands, Israel), competition will take place on risk type if premiums do not approximate expected cost at the individual level. In other words, if paid a uniform premium, it is in a plan's financial interest to enroll relatively healthy persons and not to enroll the sick. A similar phenomenon can occur among physicians to the degree that they are reimbursed a fixed capitated amount per patient and are at risk for spending above that amount (e.g., primary care groups in the UK National Health Service). Selection based on expected costliness of patients implies that premium spirals may occur, in which certain plans are left with poorer and poorer risks, and ultimately may go out of existence (Cutler and Reber 1998). It also implies that prices quoted by health plans reflect the average enrollee risk in addition to any product differences that would affect cost for a given risk. In turn this means that prices facing an individual enrollee do not generally reflect the marginal cost of providing services to that enrollee, and that the allocation of persons to plans differs from that of a perfect market.

Remedying this situation by charging premiums geared to individual risk means an individual cannot avoid the financial risk of a deterioration in health if premiums are adjusted periodically, say annually, to reflect risk in the next period. For example, persons newly diagnosed with diabetes mellitus would face an increased premium and would thus bear the financial risk of contracting diabetes. In principle, the risk would be avoided if insurance policies paid a lump sum to an individual diagnosed with a disease at the time of diagnosis (Cochrane 1995). Such a feature is not observed, however, probably for the reason that health insurance tends not to be written as a fixed indemnity policy in the first place, namely that the large number of states of the world (i.e., the severity and combinations of various diseases) makes it impractical to enumerate the states. Furthermore, such insurance would fail to protect against the risk that a better but more expensive treatment will become available in the future, a common occurrence in medicine.

A monopoly public insurance program or the direct public provision of services effectively insures against the possibility of becoming a bad risk and also avoids the selection problem inherent in competitive arrangements, but potentially gives up whatever benefits in more favorable prices or quality may be available from competition. Moreover, it cannot readily cater to diverse preferences or willingness to pay for alternative treatments of varying expense. Also, even a monopoly insurance plan may leave competition among providers with which it contracts (e.g., physicians) and thereby introduce selection issues.

One middle ground between a monopoly insurer and competitive individual insurance markets is competition to insure groups of individuals who have formed for purposes other than obtaining health insurance (Diamond 1992). Provision of insurance through the place of employment, as is common in several countries, is one method of achieving this. Employer-provided insurance, however, breaks down

for small firms or the self-employed, as well as for individuals not attached to the labor market.

Another middle ground is so-called carve-out arrangements, whereby an employer or a public health insurance plan asks for competitive bids from suppliers of particular services such as mental-health services. A contract for the given service is then awarded to a single supplier, and the contract can be periodically rebid. For the service that is carved out, there is a single supplier to an entire group, but for services that are not carved out there can be competitive suppliers within the group. Carve-outs eliminate selection for the service that is carved out, but at the price of potential coordination problems among various suppliers. For example, it may be unclear whether a visit by a depressed patient who is complaining of fatigue should be charged to the supplier of carved-out mental-health services or the health plan that is responsible for all other health services.

Currently there is much interest in the consequences of different methods of reimbursing suppliers, especially physicians. The interest stems in part from the wide variety of methods used around the world, including fee-for-service, capitation (a fixed payment per period of time for 'necessary' services), and salary. Similarly, hospitals may be paid a fee for each detailed service, an all-inclusive fee per day, or an all-inclusive fee per admission. Several countries have changed their method of reimbursement (e.g., greater use of capitation in the USA). The consequences of different methods of paying providers will almost certainly be one of the main topics of research over the next several years.

2. *Health*

In 1972 Grossman developed a model that construed the demand for medical care services as derived from the demand for health, which he treated partly as an investment good and partly as a consumption good; for an exposition of the model and subsequent developments, see Grossman (2000). As an investment good, health care was assumed to increase the stock of health and thereby the flow of healthy days and productivity; as a consumption good, health care was assumed to improve well-being. Health, of course, was not traded in the market, but factors such as age and schooling altered the optimal level of health by altering its shadow price. Grossman's model, by emphasizing health rather than medical care as what was desired, was consistent with a stream of empirical work that sought to measure clinical outcomes of health care and did not simply take spending on medical services as equivalent to output. Starting from a different position, Sen (1992) has also emphasized the importance of health as a good in its own right.

The importance of health has given rise to two types of literatures in health economics, one on cost-effective interventions to change health and the other on the determinants of health. One strand of literature evaluates various clinical interventions according to conventional economic criteria. Valuing health outcomes in dollar terms, however, is controversial; as a result, alternative methods for measuring benefits that are independent of willingness to pay have evolved. Two prominent methods are Quality Adjusted Life Years (QALYs) and Disability Adjusted Life Years (DALYs). These methods attempt to attach utility weights to various states of the world in order to compare them. In effect, they are measures of life expectancy adjusted for quality of life. For explications of methods and issues refer to Garber (2000).

It was natural for economists to study how additional medical services affected health outcomes. The RAND Health Insurance Experiment showed that free medical care induced participants to consume some 30–40 percent more services relative to a plan with a large deductible, but the additional services did little or nothing to affect the average person's health (Newhouse and the Insurance Experiment Group 1993). Such a finding was consistent with the prediction of a standard economic model, namely that ever-greater subsidies would lead to the consumption of ever-lower valued services. That additional services could have a small marginal effect on health, of course, is consistent with their having a large average effect; reducing medical services to zero could have catastrophic consequences.

Nonetheless, closer analysis of the experimental data suggested that the allocation of services by person did not fit the standard prediction. Fully insured consumers were more likely than those with cost sharing to receive services of substantial clinical value, which should have left them on average in better health, but they also were more likely to receive services of negative value (e.g., antibiotics for viral conditions). On average, the additional positive and negative services consumed by those with free care appeared roughly to offset, although for the poor who were chronically ill the additional services when care was free appeared to have a net benefit. In this group, of course, the prior odds that a given service would be beneficial rather than harmful was considerably greater.

The apparent misallocation of services to persons is consistent with one of the best established facts in health economics: there is enormous variation across areas—both small and large—in the delivery of various medical services, even conditioning on diagnosis (Phelps 2000). For example, rates of specific procedures per Medicare beneficiary in the USA generally varied by more than a factor of two from high to low across 13 large areas in 1981 (Chassin et al. 1986). And areas that had high rates for one particular procedure did not necessarily have high rates for another. Moreover, the percentage of 'appropriate' services in the various areas was roughly constant, rather than

declining in the areas with higher rates, and there were seemingly high rates of zero or negatively valued services in all areas, one-sixth to one-third for several procedures (Chassin et al. 1987). In other words, allocation of services to persons did not appear to be following the principle of maximizing health benefits in the population. Further studies have revealed high rates of zero or negatively valued services in other health systems as well (Brook 1993), which implies that the efficiency of health-care delivery worldwide may be well within the frontier.

In addition to the effect of medical services on health, economists have studied other determinants of health, in particular the consequences of various personal behaviors. The use of tobacco has probably been the most studied (Chaloupka and Warner 2000). Abuse of alcohol has also been studied intensively (Cook and Moore 2000). For both tobacco and alcohol, economists have shown that higher taxes reduce use substantially. In the USA, however, cigarette taxes are already at a level that roughly offsets the average and probably the marginal externality of smoking. By contrast, alcohol taxes are well below the average and probably the marginal externality created by excessive drinking (Manning et al. 1991).

Recently there have been a number of studies of the potential effect of income inequality on health outcomes, although most of these have been carried out by noneconomists. Many studies purport to show that increased inequality, controlling for the level of income, damages health, but there are issues as to which way causality runs and whether any inferences can be drawn from the aggregate data that many studies have employed. A useful review and critique is found in Smith (1999).

3. *Policy Considerations and the Future*

Health economists in several countries have been active in policy debates over financing medical services and measures to promote health. In some European countries, especially the UK, an important thrust has been to break apart the government role as both buyer and seller of services in order to promote competition on price among sellers. Similarly in the USA the spread of managed competition has been an effort to promote price competition among sellers of medical services (Enthoven 1988). The essence of managed competition is that a consumer should bear much or all of the incremental cost of choosing one health plan rather than another, which in turn motivates the health plan to seek economical suppliers of care.

The competition needs to be managed, however, so that it does not degenerate into a competition for good risks. As mentioned above, the 'manager' could be an employer, a coalition of employers, or government; in fact, all three are observed. Current techniques for management or risk adjustment, however, are imperfect, explaining something less than half of the systematic or predictable variance across individuals in annual spending (Newhouse 1996, Van de Ven and Ellis 2000).

Although institutional arrangements in all countries attempt to control moral hazard, the efforts are imperfect for several reasons. First, medical knowledge is continually changing, so that what was thought to be the treatment of choice today may be discredited tomorrow. And sometimes the opposite happens as well. Because the spread of knowledge is imperfect, physicians may unwittingly be delivering a service that is known to be no longer effective or which is dominated by a newer service. Second, any efforts the physician makes to elicit the patient's preferences among alternative treatments may be imperfect. In extreme cases they may be futile, as in the case of a demented patient. As a result, the physician may give the patient a service a more informed patient would not have wanted or vice versa. Third, there is often lack of consensus about how health-care services should be rationed and in particular what role, if any, income should play (Garber 2000, Williams and Cookson 2000). Indeed, the advent of managed care in the USA and its effort at controlling moral hazard has led to a demand for 'patient protection.' In particular, under managed care arrangements patients have no ready way at the point of service to express the intensity of their preference or demand for a service. As a result, it is certainly possible that rationing would occur among highly valued services. Fourth, the degree to which others are willing to pay for an individual's treatment is variable across societies and across individuals. Improving methods for rationing services is likely to remain on many policy agendas for a long time.

One of the features of the late twentieth and early twenty-first century has been the aging of most developed countries. Several factors are contributing to this trend including increased life expectancy among those over 65, the post-World War II baby boom, and the ensuing baby bust. The result will be an increased demand for medical care (those over 65 typically spend from three to five times as much on medical care as those under 65) and an increased demand for personal care services as individuals become unable to live independently without assistance. Much of this demand will be financed from public sources, but the decline in fertility in developed countries means there will be more retired elderly per worker in most countries, and hence a greater tax burden for workers.

Historically, as noted above, real per capita medical care costs have gone up four percent or five percent per year in many countries over several decades. It is now generally accepted that this increase is mainly due to technological change or the increased capabilities of medicine (Newhouse 1992, Fuchs 1996). If this rate of increase in cost continues, medical care, much of

which is publicly financed, will consume an ever greater share of GDP. There is, of course, nothing inherently wrong with this, provided the benefits justify the costs, although the presence of insurance raises a question about whether, at the margin, the benefits do justify the costs. To the degree that the growth in costs is attributable to new capabilities, the issue is willingness to pay for those capabilities, both for oneself and others. Given the similar rates of increase in medical costs across countries with dissimilar institutions, there is some presumption that much of the increase is desired.

Continued biomedical research, especially the revolution in molecular biology and the mapping of the human genome, promise further technological change in medicine. Unfortunately, the implications of the change for cost are highly uncertain. Consequently, the future of health care and what of the increased menu of valuable services various societies will wish to purchase remains highly uncertain.

See also: Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care Organizations: For-profit and Nonprofit; Health Insurance: Economic and Risk Aspects; Health Policy; Health Promotion in Schools; Health Surveys; Medical Geography; Medical Sociology; Occupational Health and Safety, Regulation of; Welfare State; Well-being and Burnout in the Workplace, Psychology of

Bibliography

Anderson G F, Poullier J-P 1999 Health spending, access, and outcomes: Trends in industrialized countries. *Health Affairs* **18**: 178–92

Arrow K J 1963 Uncertainty and the welfare economics of medical care. *American Economic Review* **53**: 941–73

Baumgardner J 1991 The interaction between forms of insurance contract and types of technical change in medical care. *RAND Journal of Economics* **22**: 36–53

Berk M L, Monheit A C 1992 The concentration of health expenditures: An update. *Health Affairs* **11**: 145–9

Brook R H 1993 Maintaining hospital quality: The need for international co-operation. *Journal of the American Medical Association* **270**: 985–7

Chaloupka F, Warner K E 2000 The economics of smoking. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 1539–627

Chassin M R, Brook R H, Park R E, Keesey J, Fink A, Kosecoff J, Kahn K, Merrick N, Solomon D H 1986 Variations in the use of medical and surgical services by the medicare population. *New England Journal of Medicine* **314**: 285–90

Chassin M R, Kosecoff J, Winslow C M, Kahn K L, Merrick N J, Keesey J, Fink A, Solomon D H, Brook R H 1987 Does inappropriate use explain geographic variation in the use of health care services? A study of three procedures. *Journal of the American Medical Association* **258**: 2533–7

Cochrane J H 1995 Time-consistent health insurance. *Journal of Political Economy* : 445–73

Cook P, Moore M 2000 Alcohol. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 1629–73

Culyer J, Newhouse J P (eds.) 2000 *Handbook of Health Economics*. North-Holland, Amsterdam

Cutler D M, Reber S J 1998 Paying for health insurance: The tradeoff between competition and adverse selection. *Quarterly Journal of Economics* **113**: 433–66

DeMeza D 1983 Health insurance and the demand for medical care. *Journal of Health Economics* **2**: 47–54

Diamond P 1992 Organizing the health insurance market. *Econometrica* **60**: 1233–54

Ellis R P, McGuire T G 1986 Provider behavior under prospective reimbursement. *Journal of Health Economics* **5**: 129–51

Enthoven A C 1988 *Theory and Practice of Managed Care Competition in Health Care Finance*. North-Holland, Amsterdam

Fuchs V R 1996 Economics, values, and health care reform. *American Economic Review* **86**: 1–24

Garber A 2000 Recent developments in CBA/CEA. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 181–221

Glied S A 2000 Managed care. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 707–53

Goddeeris J H 1984 Medical insurance, technological change, and welfare. *Economic Inquiry* **22**: 56–7

Grossman M 2000 The human capital model. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 347–408

Manning W G, Keeler E B, Newhouse J P, Wasserman J, Sloss E 1991 *The Cost of Poor Health Habits*. Harvard University Press, Cambridge, MA

McGuire T G, Pauly M V 1991 Physician response to fee changes with multiple payers. *Journal of Health Economics* **10**: 385–410

Newhouse J P 1992 Medical care costs: How much welfare loss? *Journal of Economic Perspectives* **6**: 3–21

Newhouse J P 1996 Reimbursing health plans and health providers: efficiency in production versus selection. *Journal of Economic Literature* **34**: 1236–63

Newhouse J P, Insurance Experiment Group 1993 *Free for All? Lessons from the RAND Health Insurance Experiment*. Harvard University Press, Cambridge, MA

Pauly M V 1968 The economics of moral hazard: Comment. *American Economic Review* **58**: 531–7

Pauly M V 1980 *Doctors and Their Workshops*. University of Chicago Press, Chicago

Phelps C E 2000 Information diffusion and best practice adoption. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 223–64

Rice T 1983 The impact of changing Medicare reimbursement rates on physician-induced demand. *Medical Care* **21**: 803–15

Rothschild M, Stiglitz J 1976 Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* **90**: 629–50

Sen A K 1992 *Inequality Re-examined*. Harvard University Press, Cambridge, MA

Smith J P 1999 Healthy bodies and thick wallets: The dual relation between health and economic status. *Journal of Economic Perspectives* **13**: 145–66

Van de Ven W P M M, Ellis R P 2000 Risk adjustment in competitive health plan markets. In: Culyer J, Newhouse J P

(eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 755–845
Williams A, Cookson R 2000 Equity in health. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 1863–1910
Zeckhauser R J 1970 Medical insurance: A Case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* **2**: 10–26

J. P. Newhouse

Health Education and Health Promotion

Health education is any combination of learning experiences designed to facilitate voluntary actions conducive to health (Green and Kreuter 1999). Health promotion is the combination of educational and environmental supports for actions and conditions of living conducive to health (Green and Kreuter 1999), thereby including health education. The two definitions represent a historical development in approach from more individual to more ecological where the role of the environment has acquired increased relevance in understanding and changing conditions for health. For this reason, we will use the term health promotion in the remainder of this article. Health promotion can be characterized by four other main developments: the need for planning, the importance of evaluation, the use of social and behavioral science theories, and the systematic application of evidence and theories in the development of health promotion interventions. Finally, we will describe recent developments in information technology (IT) and their effect on health promotion.

1. *Health, Environment, and Behavior*

In the ecological approach to health promotion, health is viewed as a function of individuals and the environments in which individuals are embedded, including family, social networks, organizations, and community and public policies. The relation between social economic status and health is a clear example of environmental influences on health. At the same time it also constitutes a major challenge. The central concern of health promotion is health behavior. However, health behavior refers not only to the individual's behavior but also to the behavior or actions of groups and organizations. Stress at work may be related to individual coping behavior, but also to managers' decision-making behavior (organization). Richard et al. (1996) describe the various environmental levels as embedded systems. They indicate that individuals exist within groups, which are in turn embedded within organizations and higher order systems. The individual is influenced by, and can influence directly or through groups and organizations, the higher order systems. The picture that emerges is a complex web of causation as well as a rich context for interventions. In the stress example, the individual as well as the manager will both be targets for health promotion interventions. Moreover, at the society level, the intervention may target politicians' decision-making related to a healthier organization of labor. We see managers and politicians as agents in the environment who serve as targets for health promotion interventions (Bartholomew et al. 2000).

2. *Health Promotion Planning*

Health promotion is a planned activity. The most widely used health promotion planning framework is Green and Kreuter's (1999) PRECEDE/PROCEED model (see Fig. 1).

The model begins on the right with the assessment of quality of life and health problems, changes in which should be the proposed outcomes of a health promotion intervention. It then guides the planner to assess the behavioral and environmental causes. In the behavioral assessment typically one asks what the individuals at risk are doing that increases their risk of death from the health problem. In the environmental assessment we ask what factors in the environment are related to the health problem directly or to its behavioral causes. In the previous subsection, it was explained how the environmental causes can be viewed as behaviors or actions at various environmental levels: groups, organizations, communities, and society.

In the next phase of PRECEDE/PROCEED, the determinants of the behavioral and environmental factors are assessed. Green and Kreuter (1999) describe determinants affecting behavior as predisposing, reinforcing, and enabling. Predisposing factors relate to the motivation of an individual or a group to act. Notice that the target may also be an agent in one of the environmental levels, such as a manager or a politician. These factors mostly fall into the psychological domain and include the cognitive and affective dimensions of knowing, feeling, believing, valuing. Reinforcing factors are those consequences of action that determine whether the actor receives positive (or negative) feedback and is supported socially after it occurs. Reinforcing factors thus include social support, peer influences, and advice and feedback by healthcare providers. Notice that the perception of social support would be a predisposing factor. Reinforcing factors also include physical consequences of the behavior such as well-being or pain. Enabling factors, often conditions of the environment, facilitate the performance of an action by individuals, groups, or organizations. Factors included are the availability, accessibility, and affordability of healthcare and community resources. Enabling factors also include new skills that a person, organization, or community needs to learn to carry out a behavioral or environmental

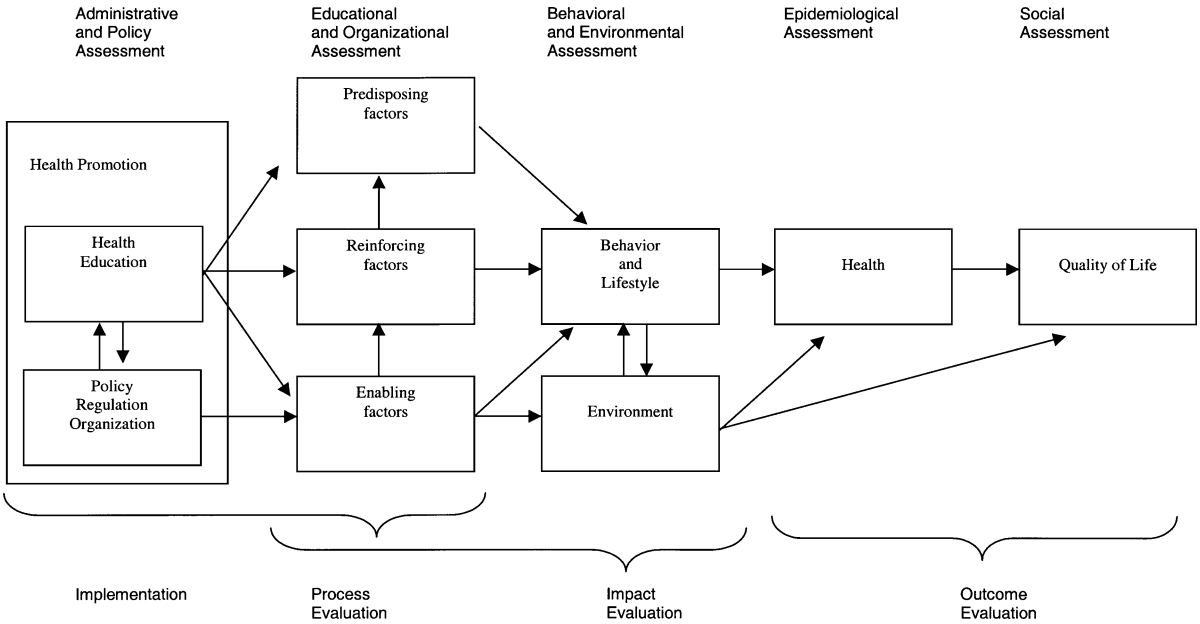

Figure 1
PRECEDE/PROCEED model (Green and Kreuter 1999)

change. Notice that self-efficacy expectations are a predisposing factor, the actual skills an enabling factor.

PRECEDE/PROCEED suggests that program planners start with the assessment of quality of life and health problems. In practice, that is a rare situation. Often health educators start somewhere within the model, depending on their task description or the specific moment in the ongoing health promotion planning process. For example, a health educator may be hired to develop a smoking prevention program for schools. We then enter the model at the behavior box. However, we first want to go back, to the right, and (re-)assess the quality of life and health problems as well as the behavioral and environmental conditions before we go on to the assessment of determinants of the behavior and environmental conditions and the development of the program.

The next phases of PRECEDE/PROCEED involve the administrative and policy assessment. The planner will now develop the plans into health education and other health promotion interventions, such as policy, regulation, or organization. The assessments include identifying barriers to overcome in implementing the program and policies that that can be used to support the program. We then proceed to program implementation. Implementation is an essential element in the planning process. Without implementation, the intervention will not have any impact on determinants, behaviors, or health. The final phases of PRECEDE/PROCEED describe the evaluation levels: process, impact, and outcome. Thinking about implementation and evaluation has to start as early as possible, not after the intervention development. In the Intervention Mapping protocol, anticipating implementation and evaluation are steps in the development process.

3. *Program Evaluation*

Even the best planned health promotion intervention may not have much effect. It is important to always evaluate the effect of an intervention, its impact and process. When the program has failed, we need to know why it failed. Rossi et al. (1999) suggest four levels of evaluation:

(a) *Program conceptualization and design* refers to the planning process and the use of theory and evidence in the development of the program.

(b) *Program monitoring* provides feedback on the actual implementation of the program and its reception by the target population.

(c) *Impact evaluation* is defined by Rossi et al. as the extent to which a program causes change in the desired direction. They use the word impact to make clear that the desired effect of a program may be defined at different levels, and often not at the health and quality of life level. It is useful to agree on the operationalizations of desired effects before the actual implementation of the program.

(d) *Program efficiency* evaluates the program in terms of costs and effects, such as cost-benefits and cost-effectiveness analyses.

Rossi et al. are very outspoken on when to evaluate and when not to evaluate: 'Clearly, it would be a waste of time, effort and resources to estimate the impact of a program that lacks measurable goals or that has not been properly implemented.'

Effect evaluation relies heavily on quasi-experimental designs (Rossi et al. 1999). The basic principle in effect evaluation is the random assignment of participants to the program group and a control group. Only then can we give a straightforward answer to the question if an intervention was successful. In health education practice the random assignment of individuals is often impossible. For instance, students from secondary schools cannot be assigned randomly to a school program or a control program, because most school programs are school-wide or at least class-wide. In that case, we ask various schools if they are willing to participate in the program, and we then assign schools randomly to the program condition or the control condition. A variety of quasi-experimental designs allow us to compare two or more groups that are as similar as possible. We have to expect that the groups are not completely equivalent on a number of relevant characteristics, and we cannot even assume that we know all the relevant characteristics. Statistically, we can control for most of these differences and levels, but only when we measure them before the program starts (see Bartholomew et al. 2000).

4. *Social and Behavioral Science Theories*

A health promotion program is most likely to benefit participants and the community when it is guided by social and behavioral science theories of health behavior and health behavior change (Glanz et al. 1997). Theory-driven health promotion programs require an understanding of the components of the theory as well as the operational or practical forms of these theories. Finding and applying relevant theories is a professional skill that health educators have to master (Bartholomew et al. 2000).

Notice that we assume that all theories are potentially applicable to all levels and also to adoption and implementation. The Theory of Planned Behavior, for example, is often applied to individual health behavior (Godin and Kok 1996) but also to politicians' behavior (Flynn et al. 1998) and to implementers' behavior (Paulussen et al. 1995).

The behavioral science theories that try to explain behavior and behavior change have two types of roots: health and health promotion in particular or behavior and behavior change in general (the last mainly being social psychological theories). Health and health promotion oriented theories are often related to perceptions of health risks, for example, the Health Belief Model (Sheeran and Abraham 1996, Strecher and Rosenstock 1997) or Protection Motivation Theory (Boer and Seydel 1996). Some other theories were developed in a health setting, but have evolved into a general theory such as the Transtheoretical Model of Stages of Change (Prochaska et al. 1997) or Relapse Prevention Theory (Marlatt and Gordon 1985). Most general social psychological models were developed for a broad range of behaviors, but are easily applicable to health behavior and change. Some examples are Learning Theories (Westen 1996), Information Processing Theory (Hamilton and Ghatala 1994), Theory of Planned Behavior (Connor and Sparks 1996, Montano et al. 1997), Social Cognitive Theory (Bandura 1997), Goal Setting (Strecher et al. 1995), Attribution Theory (Weiner 1986), Self-regulatory theories (Clark and Zimmerman 1990), Social Networks and Social Support theories (Heaney and Israel 1997), Persuasive-communication Model (McGuire 1985), and Diffusion of Innovations Theory (Rogers 1995).

At the organizational level, health educators draw from various Organizational Change theories (Goodman et al. 1997). Typically we use constructs from these theories for considering policy development within organizations and for promoting program adoption and implementation.

At the community and society level, theoretical input can be found in the classic models of community organization and community development and the current perspectives being used for health promotion (Minkler 1997). Issues of power, participation, and goals, for instance, have received much discussion among health educators in recent years. In selecting the constructs to use for community change, planners must clarify their assumptions and values about the nature of the change process and select and implement strategies congruent with these. Finally, there are theories of policy making that have been used primarily at the national, state, and governmental level, such as Policy Window Theory (Kingdon 1995).

Theories are very important tools for professionals in health education and promotion. On the one hand, theories have become available to health promotion practice through textbooks such as Glanz et al. (1997) or Connor and Norman (1996). On the other, the application of theory has long been a challenge, for researchers as well as practitioners. Students of health promotion learn of theories and learn how to apply theories to well-selected practical problems. However, in real life the order is reversed: the problem is given and the practitioner has to find theories that may be helpful for better understanding or changing behavior (Kok et al. 1996). Recently, a protocol was published that describes a process for developing theory-based and evidence-based health education programs: Intervention Mapping (Bartholomew et al. 2000).

5. *Intervention Mapping*

Intervention Mapping distinguishes five steps: define proximal program objectives, select theoretical methods and practical strategies, design the program,

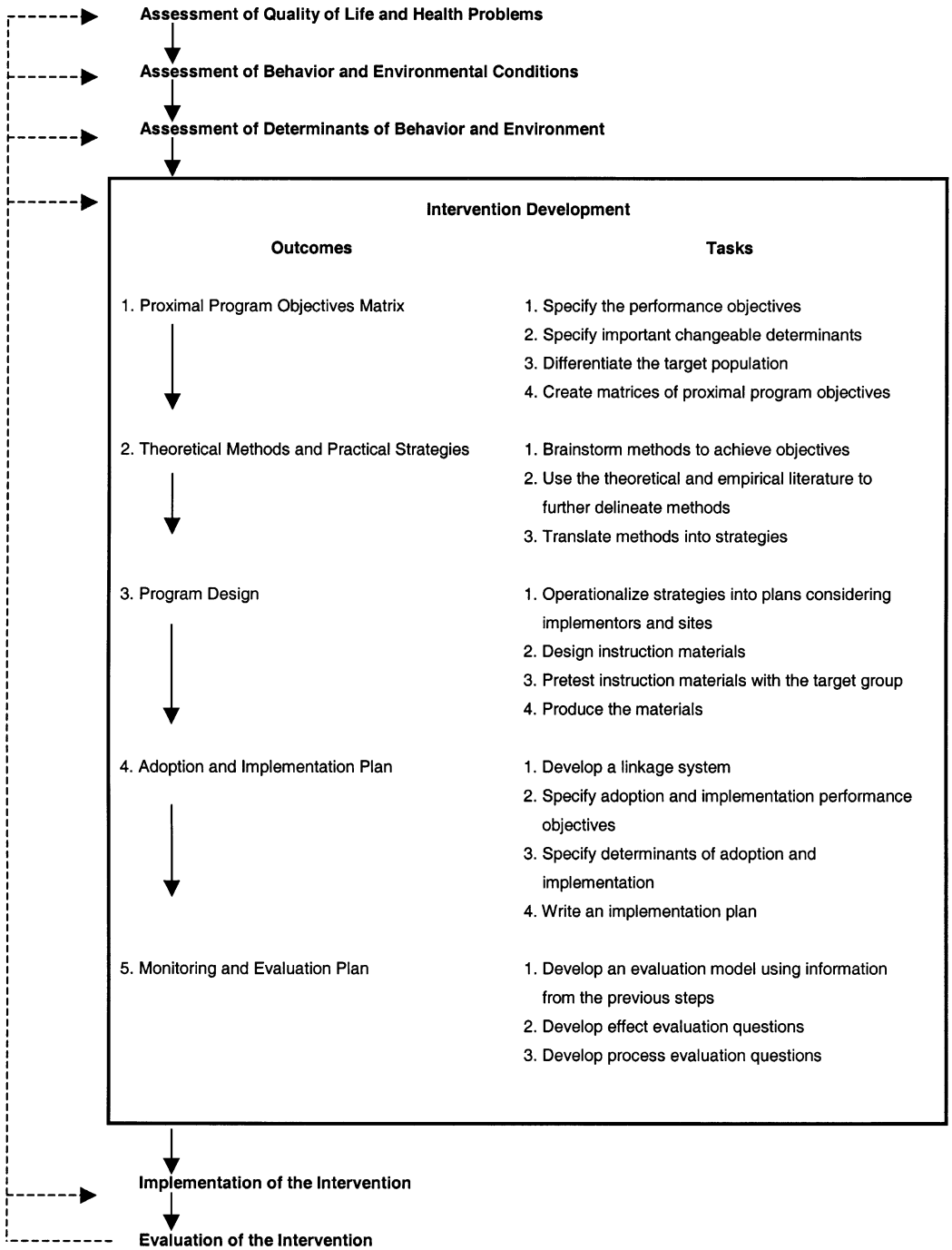

Figure 2
Intervention Mapping (Bartholomew et al. 2000)

anticipate adoption and implementation, and anticipate process and effect evaluation (see Fig. 2). Bartholomew et al. (2000) see planning as an iterative process: two steps forward and one step back. Bartholomew et al. (2000) describe three core processes for Intervention Mapping, tools for the professional

health promoter: searching the literature for empirical findings, accessing and using theory, and collecting and using new data. From the literature search, a provisional list of answers is developed, which is often not adequate in finding solutions for the problem. The planner must go further to search for theory, using three approaches: issue, concept, and general theories approach. The issue approach searches the literature for theoretical perspectives on the issue. The concept approach begins with the concepts in the provisional answers, linking these concepts to theoretical constructs and theories that may be useful. The general theory approach considers general theories that may be applicable. Finally, it is important to identify gaps in the information obtained and collect new data to fill these gaps.

The first step in Intervention Mapping is the development of proximal program objectives, a crossing of performance objectives, determinants, and target groups. For instance, one proximal program objective for an HIV-prevention program in schools would be: 'adolescents (target population) express their confidence (determinant) in successfully negotiating with the partner about condom use (performance objective).' Performance objectives are the specific behaviors that we want the target group (or the environmental agents) to 'do' as a result of the program. For example, in the case of HIV prevention: buy condoms, have them with you, negotiate with the partner, use them correctly, and keep using them (Bartholomew et al. 2000). Determinants of behavior can be personal or external. Personal determinants are, for instance, outcome expectations, social influences, and self-efficacy expectations. External determinants are, for instance, social norms and support, and barriers. Target groups can be subgroups of the total group, for instance, men/women, people in different stages of change. Proximal program objectives are numerous and should be ordered by determinant. So we end this step with a series of lists, for instance, all proximal program objectives which have to do with skills training.

Intervention Mapping Step 2 is the selection of theoretical methods and practical strategies. The theoretical method is the technique derived from theory and research to realize the proximal program objective; the strategy is the practical application of that method. For instance, the method for self-efficacy improvement could be modeling and the strategy could be peer modeling by video. An important task in this step is to identify the conditions or parameters that limit the effectiveness of theoretical models, such as identification, observability, and reinforcement as necessary conditions for the effectiveness of learning by modeling.

Intervention Mapping Step 3 is the actual designing of the program, organizing the strategies into a deliverable program, and producing and pretesting the materials. In the example of HIV-prevention in schools, the program comprised five lessons, an interactive video, a brochure for students, and a workbook for teachers (Bartholomew et al. 2000).

A solid diffusion process is vital to ensure program success. So, in Intervention Mapping Step 4, a plan is developed for systematic implementation of the program. The first thing to do, actually at the start of intervention development, is the development of a linkage system, linking program developers with program users. Then, an intervention is developed to promote adoption and implementation of the program by the intended program users. It may be clear that the anticipation of implementation is a relevant process from the very beginning of planning, not only at the end.

Finally, Intervention Mapping Step 5 focuses on process and effect evaluation. Again, this process is relevant from the start, not only at the end. For instance, 'adolescents express their confidence in successfully negotiating with the partner about condom use' is an objective, but is also a measure of that objective, that can be asked in pre- and postinterviews with experimental and control group subjects.

6. *Recent Developments in IT*

Mass media health promotion is able to reach large numbers of people, but a major weakness of mass media programs is their limited effect in terms of impact and outcome. However, more effective individualized health promotion is costly and time-consuming. Computer technology may help break several important barriers in health promotion related to individualization and program delivery. During the 1990s, health promotion researchers have been testing simple computer expert systems that enable the tailoring of communications to match certain characteristics that have been measured in individuals in the target group (Velicer et al. 1993, Brug et al. 1998, Dijkstra and De Vries 1999).

Velicer et al. define an expert system as a collection of facts and rules about something and a way of making inferences from the facts and the rules. The most common types of expert systems in health education are computer programs that are tailored to specific characteristics of the receiver. Thus, the expert system comprises one or more databases of messages based on theoretical constructs that vary as they apply to different characteristics of individuals, and on algorithms for matching the messages to individuals. The message channel could in general be anything that facilitates delivery of the message to the target: a report, a letter, computer-assisted instruction or any other channel. In the work of Velicer et al. (1993), messages were based on the Transtheoretical Model of Stages of Change and included processes of change tailored to the stage of the individual in regard to quitting smoking. Feedback included current status

and stage of change, current use of change processes, suggested strategies, and high-risk situations. Feedback was compared against a normative database as well as against participants' progress. All examples of computer-tailored health promotion in the literature are based on a similar configuration comprising (Brug et al. 1998):

(a) A theoretical framework and specification of relevant hypothesized determinants of the health behavior.

(b) Use of the determinant model to create a data collection tool and a series of messages that are tailored to these determinants.

(c) Several databases including, at least, a determinants file and a delivery message file.

(d) A set of decision rules and a tailoring program that executes these decisions.

(e) The actual communication messages.

(f) Delivery vehicles, such as a printed letter.

It is to be expected that computer tailoring will become more popular in the future and will also be applied using the Internet.

7. Conclusion

Health education and health promotion are planned activities. PRECEDE/PROCEED provides a very useful framework for the systematic planning of health promotion and Intervention Mapping provides a very useful protocol for evidence-based intervention development. Both models acknowledge the role of the individual as well as the relevant ecological levels in the environment.

See also: Health Behavior: Psychosocial Theories; Health Behaviors; Health Interventions: Community-based; Health Promotion in Schools

Bibliography

Bandura A 1997 *Self-Efficacy: The Exercise of Control*. Freeman, New York

Bartholomew L K, Parcel G, Kok G, Gottlieb N 2000 *Intervention Mapping: Designing Theory and Evidence-based Health Promotion Programs*. Mayfield, Mountain View, CA

Boer H, Seydel E R 1996 Protection motivation theory. In: Conner M, Norman P (eds.) *Predicting Health Behaviour: Research and Practice with Social Cognitive Models*. Open University Press, Buckingham, UK

Brug J, Glanz K, Van Assema P, Kok G, Van Breukelen G P 1998 The impact of computer-tailored feedback and iterative feedback on fat, fruit, and vegetable intake. *Health Education and Behavior* **25**: 517–31

Clark N M, Zimmerman B J 1990 A social cognitive view of self-regulated learning about health. *Health Education Research* **3**: 371–9

Connor M, Norman P (eds.) 1996 *Predicting Health Behaviour: Research and Practice with Social Cognition Models*. Open University Press, Buckingham, UK

Connor M, Sparks P 1996 The theory of planned behavior and health behaviors. In: Connor M, Norman P (eds.) *Predicting Health Behaviour: Research and Practice with Social Cognitive Models*. Open University Press, Buckingham, UK, pp. 121–62

Dijkstra A, De Vries H 1999 The development of computer-generated tailored interventions. *Patient Education and Counseling* **36**: 193–203

Flynn B S, Goldstein A O, Solomon L J, Bauman K E, Gottlieb N H, Cohen J E, Munger M C, Dana G S 1998 Predictors of state legislators' intentions to vote for cigarette tax increases. *Prevention Medicine* **2**: 157–65

Glanz K, Lewis F M, Rimer B K 1997 *Health Behaviour and Health Education: Theory Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA

Godin G, Kok G 1996 The theory of planned behavior: A review of its applications to health related problems. *American Journal of Health Promotion* **11**: 87–98

Goodman R M, Steckler A, Kegler M C 1997 Mobilizing organizations for health enhancement: Theories of organizational change. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA, pp. 287–312

Green L W, Kreuter M W 1999 *Health Promotion Planning: An Educational and Ecological Approach*, 3rd edn. Mayfield, Mountain View, CA

Hamilton R, Ghatala E 1994 *Learning and Instruction*. McGraw Hill, New York

Heaney C A, Israel A 1997 Social networks and social support. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA, pp. 179–205

Kingdon J 1995 *Agendas, Alternatives and Public Policies*, 2nd edn. Harper-Collins, New York

Kok G, Schaalma H, De Vries H, Parcel G, Paulussen T H 1996 Social psychology and health education. In: Stroebe W, Hewstone M (eds.) *European Review of Social Psychology*. Wiley, Chichester, UK, Vol. 7, pp. 241–82

Marlatt G A, Gordon J R 1985 *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviours*. Guilford Press, New York

McGuire W J 1985 Attitudes and attitude change. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*, Vol. 2. Random House, New York, pp. 233–346

Minkler M (ed.) 1997 *Community Organizing and Community Building for Health*. Rutgers University Press, New Brunswick, NJ, pp. 244–58

Montano D E, Kasprzyk D, Taplin S H 1997 The theory of reasoned action and the theory of planned behavior. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA, pp. 85–112

Paulussen T, Kok G, Schaalma H P, Parcel G S 1995 Diffusion of AIDS curricula among Dutch secondary school teachers. *Health Education Quarterly* **22**: 227–43

Prochaska J O, Redding C A, Evers K E 1997 The transtheoretical model and stages of change. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA, pp. 60–84

Richard L, Potvin L, Kishchuk N, Prlic H, Green L W 1996 Assessment of the integration of the ecological approach in health promotion programs. *American Journal of Health Promotion* **10**: 270–81
Rogers E M 1995 *Diffusion of Innovations*, 4th edn. Free Press, New York
Rossi P H, Freeman H E, Lipsey M W 1999 *Evaluation: A Systematic Approach*, 6th edn. Sage, Newbury Park, CA
Sheeran P, Abraham C 1996 The health belief model. In: Connor M, Norman P (eds.) *Predicting Health Behaviour: Research and Practice with Social Cognition Models*. Open University Press, Buckingham, UK, pp. 23–61
Strecher V J, Rosenstock I M 1997 The health belief model. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research, and Practice*. Jossey Bass, San Francisco, CA, pp. 41–59
Strecher V J, Seijts G H, Kok G, Latham G P, Glasgow R, DeVellis B, Meertens, R M, Bulger D W 1995 Goal setting as a strategy for health behavior change. *Health Education Quarterly* **22**: 190–200
Velicer W F, Prochaska J O, Bellis J B, DiClemente C C, Rossi J S, Fava J L, Steiger J H 1993 An expert system intervention for smoking cessation. *Addictive Behaviors* **18**: 269–90
Weiner B 1986 *An Attributional Theory of Motivation and Emotion*. Springer, New York
Westen D 1996 *Psychology: Mind, Brain, and Culture*. Wiley, New York

G. Kok

Health, Global Budgets for

Introducing global budgets for health, i.e., setting prospective limits to expenditure on health care, is by definition the most effective means of controlling spending. If expenditure levels are fixed for the coming year or years, and if the limit is made to stick, then tautologically the rise in spending on health will be contained within the planned total. Not surprisingly, therefore, an increasing number of governments adopted this strategy in the 1980s, as a reaction to the largely unplanned and unanticipated rise in health care spending throughout the 1960s and much of the 1970s: what came to be known as the health care cost explosion. This article examines, first, the feasibility, form and effectiveness of this strategy in different health care systems; second, the conceptual and practical issues involved in fixing the global budget and then allocating it to populations or health care sectors.

1. *Strategies and Mechanisms*

In 1960, total health expenditure (privately as well as publicly financed) in the world's 18 richest countries—those belonging to the Organisation for Economic Co-operation and Development—accounted for 4.2 percent of their combined gross domestic product (GDP). Ten years later, the figure was 5.8 percent, rising to 7.2 percent in 1980 (OECD 1987). Both the rate of increase, and the proportion of GDP devoted to health, differed among countries. Between 1960 and 1980, spending in the United States increased from 5.3 to 9.5 percent of GDP, rising to 10.7 percent by the mid-1980s, whereas the figures for the UK were 3.9 and 5.6 percent respectively, leveling out thereafter. From this flowed a widespread perception that spending on health was out of control, imposing damaging fiscal burdens on the economy and a consequent interest in mechanisms for halting, or at least slowing down, the seemingly inexorable rise.

The problem of ever-escalating costs could not be solved by invoking the usual market mechanisms and restraining demand by raising prices. Not only is health care financed predominantly through insurance schemes of various types; i.e., third parties rather than the consumers pay most of the costs. Equally, there is only limited scope for using co-payments or user charges to limit demand: the long-term effects on population health of deterring use may outweigh any short-term fiscal gains. Controlling the supply of health care thus seemed a more feasible and effective option, all the more so because of evidence that much of the demand for health care was generated by the system itself; i.e., was provider-driven. While it is consumers who make the initial contact with the health care system, it is the professionals who decide how they should be treated. And it is these decisions—about the number of tests to be carried out, the drugs to be prescribed and so on—which drive spending. Enter global budgeting.

Here the UK's National Health Service (NHS) was the paradigmatic case. Outside the communist bloc countries, the UK was the only health care system funded overwhelmingly by central government out of general taxation. In this it differed both from the tax-funded Scandinavian model, where both control and fiscal responsibility was divided between central and local government, and the social insurance based West European model, where responsibility was divided between central government and an often large number of insurance funds. In the UK, therefore, the Treasury controlled the fiscal tap, for both current and capital spending, to the health care system. Almost from the launch of the NHS in 1948, the system operated what in effect was a global budget; i.e., the Treasury set the limits of total spending for the next fiscal year. In turn, the UK's record suggested two conclusions. First, it appeared to demonstrate that global budgeting could restrain the rate of increase in health care expenditure: the UK continued, and continues, to be a restrained spender relative to other countries with a comparable national income. Second, it seemed to show that there was a direct relationship

between fiscal centralization and the successful operation of a global budget.

Both conclusions were borne out, if to a differing extent, as an ever higher proportion of countries adopted variations on the global budgeting theme in the 1980s and early 1990s as part of their strategies for cost containment. The rate of increase in health care expenditure, as a proportion of national incomes, did indeed slow down (Abel-Smith 1992, OECD 1992, 1994), and much of this could be attributed to the introduction of global budgeting. In all this, the USA was the most conspicuous exception; it did not introduce global budgeting and its spending on health care was almost 14 percent of GDP by the end of the 1990s (twice the UK figure). However, success did not necessarily correlate with fiscal centralization. Sweden not only halted, but reversed, the upward trend, even though its model of health care divides responsibility between central and local government: no doubt because the former carries by far the heavier fiscal clout. And the Netherlands, perhaps the most pluralistic variant on the European social insurance model, introduced global budgeting covering both the public and the private insurers (Maarse 1996) with the same set of fee schedules applying to both sectors.

This suggests that there are various mechanisms for achieving global budgeting—i.e., control of the supply side of health care—although the UK's single-payer model offers the most direct leverage. Without the simplicity of the UK model, these mechanisms vary with the nature of the reimbursement system used in different sectors of health care services and the political determination of governments to make their policies stick. In the case of the hospital sector, it proved relatively easy to impose global budgets (Glaser 1993). In the case of service payments for general practitioners or office-based specialists, controlling spending turned out to be less easy. Freezing the fee schedule tends to multiply activity as physicians safeguard their incomes by doing more. Here the strategy adopted was to impose a limit on the total amount to be paid out in fees, with retrospective downward adjustments of the payments to individual physicians if the total exceeds the target figure: Germany providing one such example, leaving responsibility for administering the pool to the medical profession (Schwartz and Busse 1996). And while the USA is a conspicuous failure in macro-terms, it has generated innovative models for microglobal budgeting; i.e., various forms of capitation payments to providers.

Systems also vary in the extent to which global budgets are underpinned by, or complementary to, other supply side measures. Some systems, such as that in the UK, control the production and distribution of the medical labor force. Most systems seek to control the number and distribution of hospitals and to exercise birth control over the diffusion of expensive new technologies. In the case of health care systems based on the social insurance model, governments can further buttress policy by regulating—in some instances freezing—the level of contributions charged by insurance funds, so controlling the flow of money into the system.

Global budgets have their pathologies as well as their advantages. By removing the incentive to maximize income by maximizing activity (possibly a desirable policy goal, since it also lessens the risk of overtreatment), global budgets for hospitals may also undermine the incentives to maximize efficiency (so leading to the undesirable policy outcome of queues for treatment). Hence Britain's brief experiment with a mimic market in the 1990s. This separated the purchaser and provider roles. Previously, health authorities had been responsible for managing hospitals and other services directly for the population within their geographical boundaries. Under the new system, they became responsible only for buying the necessary care from semi-autonomous agencies. Both before and after the changes, however, health authorities worked within budgets fixed by central government. The aim (never fully achieved) was to introduce an element of competition between providers—and thus provide incentives to greater efficiency—within the framework of a global budget for the NHS as a whole.

Some other issues also require noting. Global budgets for item-of-service payments raise awkward questions about the distribution of income between physicians: between those who maximize their activity and those within a fixed total who do not. Above all, global budgets tend to be associated with rationing: if the supply is limited, waiting lists may be the result. And while the phenomenon of rationing is far from exclusive to health care systems with global budgeting, it does raise questions about the adequacy of the budget level set.

2. *Setting the Budget*

The adoption of global budgets represents the pragmatic reaction of governments searching for policy instruments to restrain the growth of collective spending on health. A global budget for the health care system of a country is thus the byproduct of a government's decisions about priorities between expenditure on health and other programs competing for funds, between maintaining public services and cutting taxation. Its level reflects as much calculations of political feasibility—assumptions about how much spending can be restrained without provoking a political backlash from professional providers and the public—as calculations about what the appropriate level of spending on health should be. As such, it is an entirely atheoretical exercise. For no formula for determining a global budget, independent of the actual level of spending in different countries, has yet been devised: there are no agreed benchmarks against which the adequacy or otherwise of a given level of spending

can be assessed. Nor is it likely that such benchmarks can be devised, given the heterogeneity of the services produced under the 'health' label, the scope for substitution between them, and the productive factors involved and variations in productivity.

Global budgets are therefore best seen as history modified by politics, and often rationalized by appeals to demography. The inherited level of spending on health inevitably provides the baseline for setting global budgets. This is then modified in the light of factors either extrinsic to the health care system (usually the prospective capacity of the economy to produce extra wealth), or intrinsic to it (usually demography and the cost of technological innovation). The first strategy puts the emphasis on what a country can afford. The second puts the emphasis on what a country needs. Neither is as objective as it appears (Smee 1996). 'Affordability' is an elastic concept. And so is need. Even the most commonly used indicator of 'need,' the demographic structure of any given population, is problematic. For sure, people need more health care as they age. But projecting the cost of demographic change into the future assumes that patterns of both morbidity and service use will remain the same as in the past: a highly questionable assumption. And many of the same problems arise when it comes to apportioning global budgets, whether geographically or by health care sectors.

3. *Dividing up the Budget*

The issue of distributing global budgets geographically arises only in the case of health care systems—on the UK and Scandinavian models—where public agencies are responsible for delivering health care to a defined population. In these cases, the equity principle demands that resources should be distributed according to need. In the past, need was often equated with use: if use was high so, it was assumed, was need. But given increasing evidence that demand (or use) was strongly influenced by supply, emphasis has switched to devising other indicators. The demographic composition of the population provides, as noted above, the core of the various formulas that have been devised to allocate resources. Standardized Mortality Ratios are another. So is morbidity data, where available. Debate continues as to whether social deprivation is in itself a need indicator, over and above its contribution to the mortality and morbidity figures. In all this, need is, of course, a relative concept. That is, the various distributive formulas that have been developed purport to measure the needs of population A as against those of population B, not to measure the absolute level of provision that either requires.

In the case of the West European social insurance model of health care provision, the question of geographical distribution clearly does not arise. Responsibility for funding health care is divided between funds whose members—as often defined by occupation as by place of residence—do not necessarily live within the same geographical areas. However, some of the same distributional issues do arise. One frequent element in social insurance systems is a central pool of money which is then distributed to individual funds according to the risks that they carry. Designed to weaken the incentives of funds to discriminate against potential members who may—because they are older or sicker—require more expensive health care, risk compensation in turn means devising a formula for assessing relative need. This raises the same conceptual issues as devising a formula for distributing global budgets geographically and leads to many of the same solutions, notably heavy weighting for demographic factors.

Global budgets are, in most systems, also broken down by sectors for accounting and control purposes. Indeed, a national global budget may be the product of a step-by-step process of imposing expenditure limits on specific sectors; primary care, acute hospital care and so on. However, in this case there have not been—and probably cannot be, given the absence of a conceptual basis for such an exercise—any attempts to devise a formula for allocating funds between sectors. The sectoral distribution of global budgets, even more than the determination of their total, remains the product of history modified by incremental policy decisions.

See also: Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care, Rationing of; Health Care Systems: Comparative; Health Economics; Health Policy; National Health Care and Insurance Systems; Pecuniary Issues in Medical Services: Ethical Aspects; Public Health

Bibliography

Abel-Smith B 1992 Cost containment and new priorities in the European Community. *The Milbank Quarterly* **70**(3): 393–416

Glaser W A 1993 How expenditure caps and expenditure targets really work. *The Milbank Quarterly* **71**(1): 97–128

Maarse H A M 1996 Fixed budgets in the inpatient sector: The case of the Netherlands. In: Schwartz F W, Glennerster H, Saltman R B (eds.) *Fixing Health Budgets: Experience from Europe and North America*. Wiley, Chichester, UK

OECD 1987 *Financing and Delivering Health Care: A Comparative Analysis of OECD Countries*. OECD, Paris

OECD 1992 *The Reform of Health Care: A Comparative Analysis of Seven OECD Countries*. OECD, Paris

OECD 1994 *The Reform of Health Care Systems: A Review of Seventeen OECD Countries*. OECD, Paris

Schwartz F W, Busse R 1996 Fixed budgets in the ambulatory care sector: The German experience. In: Schwartz F W, Glennerster H, Saltman R B (eds.) *Fixing Health Budgets:*

Experience from Europe and North America. Wiley, Chichester, UK
Schwartz F W, Glennerster H, Saltman R B (eds.) 1996 *Fixing Health Budgets: Experience from Europe and North America*. Wiley, Chichester, UK
Smee C H 1996 Setting regional allocations and national budgets in the UK. In: Schwartz F W, Glennerster H, Saltman R B (eds.) *Fixing Health Budgets: Experience from Europe and North America*. Wiley, Chichester, UK

R. Klein

Health in Developing Countries: Cultural Concerns

Human health and wellbeing worldwide are profoundly affected by complex interactions between economic, social, and political forces as well as by behavioral patterns and the application of technology (Wilkinson 1996). When considering health in countries around the world, use of the terms 'developed' and 'developing' countries implies a linearity of progress towards a condition that has been achieved by some countries but not others. It is clear that such a simplistic approach equating development with economic status in terms of average per capita gross national product (GNP) or the level of industrialization is both inaccurate and unhelpful. So-called developed countries may lack what many would consider very important aspects of development; for example, a sensitivity to the impact of consumption on ecology, or concern for the adverse effects of opulent and aggressive lifestyles on billions of people in other countries, or more obviously in relation to health, the achievement of universal access to health care. Similarly, some so-called developing countries are more developed than may be apparent from their economic status; for example in Kerala, India, the degree of social solidarity and health status achieved exceeds that expected from economic status alone—revealing the impact of human capital resources such as female literacy and empowerment.

A more sophisticated model of development (such as that used in the United Nations Human Development Report 1991) emphasizing achievements in health, education, female literacy, gender equality, and political liberty, more accurately reflects the real human development required. However, it should be acknowledged that even such development indices are incomplete, as they do not measure some important aspects of human flourishing. When comparing development and health in countries around the world, it is also necessary to appreciate that consumption patterns and the ways of life in industrialized countries cannot become the norm for the whole world. For example, while the world's population has grown fivefold in the past 150 years, human consumption has increased 30-fold. The USA, with 5 percent of the world's population, accounts for 24 percent of annual global energy consumption compared to China with 20 percent of the world's population using 10 percent. Within health care, even a nation that spends $4,000 per person per year on health care cannot afford to provide access for all citizens to all that modern medicine can offer (Iglehart 1999). Modern consumption patterns and health care practices are thus neither universalizable nor sustainable. Such considerations reveal the inadequacy of the terms 'developed' and 'developing' countries.

1. A Polarized World

While the twentieth century is considered by many to be the most successful in world history, with spectacular achievements in science, technology, and health care, the above considerations make it necessary to focus on the down side of progress—as reflected in widening disparities in economic and health status, the extent of expenditure on weapons of mass destruction, and the social construction of human rights abuses. Improvements in human health globally in the twenty-first century and achieving real development of all people throughout the world will be dependent on acknowledging such adverse features of modern life, and on developing new paradigms of thinking that will promote widespread human flourishing.

1.1 Economic Disparities

At the beginning of the twentieth century the wealthiest 20 percent of the world's population were nine times richer than the poorest 20 percent. This ratio has grown progressively—to 30 times by 1960, 60 times by 1990, and to over 70 times by 1997. The extent of absolute poverty has also increased. Today almost half the world's population lives on less than US$2 per person per day and more than a billion people live on less than US$1 per day. Tens of millions of people, including many children, die every year of starvation—in a world with enough food to feed all.

1.2 Disparities in Health and Disease

At the end of the twentieth century patterns of diseases and of longevity also differ markedly across the world. Of the approximately 52 million people who die each year, 17.5 million die of infectious and parasitic diseases (16 million of these—many in their youth—in

Table 1
Causes of death in 1997 by main causes and by level of development (The World Health Report 1997)

	Developed world (12 million deaths)	Developing world (45 million deaths)	World
Infectious and parasitic diseases	1 percent	43 percent	33 percent
Perinatal and maternal causes	1 percent	10 percent	8 percent
Cancers	21 percent	9 percent	12 percent
Diseases of the circulatory system	46 percent	24 percent	29 percent
Diseases of the respiratory system	8 percent	5 percent	6 percent
Other and unknown causes	23 percent	9 percent	12 percent

the developing world), 10 million die of diseases of the circulatory system (4.5 million of these in the developing world), and 6 million die of malignant diseases (3.5 million of these in the developing world) (see Table 1).

Among the poorest quintile of people in the world, 55 percent die of communicable diseases, as compared with 5 percent of the richest. The World Health Organization (WHO) estimated that in 1998, 11 million children and adults of working age died of six infectious diseases that could have been prevented at the cost of US$20 per life saved. Poor countries bear over 80 percent of the global burden of disease in disability adjusted life years (DALYs). This burden is likely to increase as the epidemiological transition progresses, with added disability and suffering from noncommunicable causes of disability such as vascular disease, malignant neoplasms (especially of the lung associated with smoking), neuropsychiatric disease, accidents, and trauma.

Life expectancy at birth ranges from well over 70 years in highly industrialized countries to below 50 years in many poor countries. Wide disparities in life expectancy are also observed between rich and poor in rich countries. In sub-Saharan Africa gains in longevity achieved during the first half of the twentieth century are rapidly being reversed by the HIV/AIDS pandemic.

1.3 Mental Health

Improvements in physical health over the past 50 years have not been matched by improvements in mental health—even in developed countries. Mental illness accounts for 8 percent of the global burden of disease (measured in DALYs), and a further 34 percent of the global burden of disease is accounted for by behavior related problems such as violence, diarrhoeal diseases, AIDS and other sexually transmitted diseases, and motor vehicle accidents and other nonintentional injuries (Desjarlais et al. 1995). The inadequacy of the knowledge base about mental disease (compiled almost exclusively in North America and Europe) to explain psychiatric diseases in other parts of the world poses formidable challenges for progress.

1.4 Militarization and the Social Construction of Human Rights Abuses

Industrialized countries spend on average 5.3 percent of GNP on the military but only 0.3 percent on economic aid to developing countries (Sivard 1996). The arms trade and its link to economic aid has contributed significantly to the escalating number of wars, conflict, and widespread torture since the Second World War which have rendered over 25 million people to refugee status and displacement—with profound disruptive effects on life and health. Global military expenditure in 1990 amounted to US$1 trillion (Sivard 1996). The annual budget of the WHO is equivalent to 2.5 hours of global military expenditure. While money is always available to make war, it is less readily available to promote peace. Military expenditure has decreased in recent years, but released resources have not been channeled towards effective human development processes.

2. Relationship Between Economics and Health

It is now well established that there is a definite relationship between wealth/poverty and health/disease, although this relationship is complex and not linear. For example, one of the wealthiest nations in the world (the USA) has worse health statistics (infant mortality and longevity) than some other highly industrialized countries, and a particularly poor state

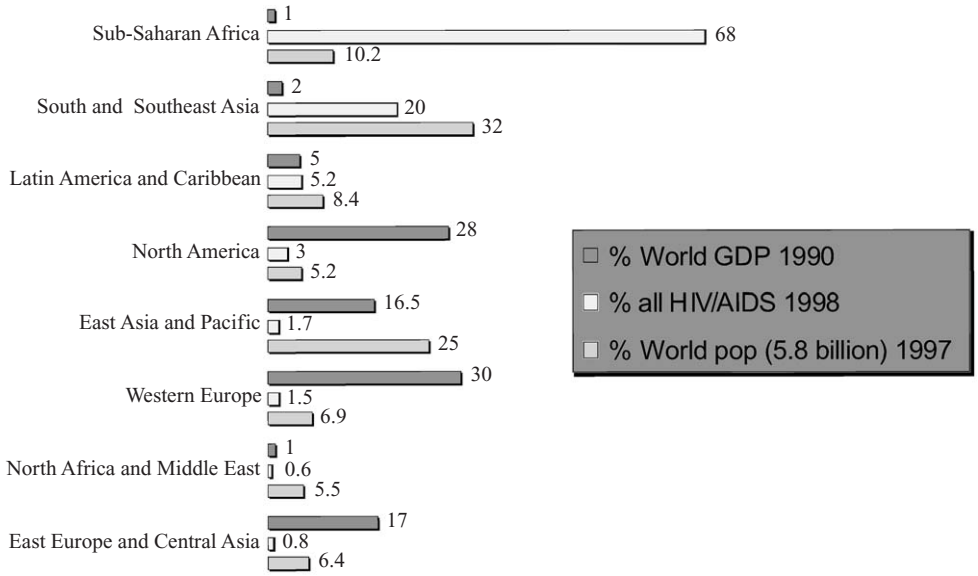

Figure 1
Distribution of economic resources (top bar), the burden of HIV/AIDS (middle bar) and world population (lower bar) in the world

Table 2
Health and other statistics in the 40 poorest countries in the world (WHO 1995)

Annual per capita GNP (US$)[a]	168	282	404	570
Life expectancy at birth (years)[a]	47	51	56	60
Infant mortality rate per 1,000 live births[a]	126	112	90	70
Under 5 year mortality rate	158	169	132	109
Population access to safe water[a]	37%	53%	53%	63%
Population access to sanitation[a]	29%	26%	39%	48%
Population per physician (1970)	33,925	39,381	23,009	15,956
Adult literacy rate (1992)	42%	44%	46%	63%
Average annual population growth[a]	2.7%	2.8%	2.6%	2.3%

a 1991 data.

(Kerala) in India has achieved lower infant mortality rates and greater longevity than many wealthier nations (Wilkinson 1996). Both absolute and relative wealth affect health. Among the developed countries it is not the richest societies that have the best health, but rather those with the smallest income differentials between rich and poor (Wilkinson 1996). Despite the nonlinearity of the relationship between wealth and health, its existence underscores the need to see health and disease as intimately linked to social and economic conditions (Fig. 1).

Poverty (defined as lack of economic resources, lack of education, lack of access to basic life resources such as food water and sanitation, and lack of control over the reproductive process) directly accounts for almost one-third of the global burden of disease. Poverty leads to poor health, which in turn aggravates poverty. Of the 40 poorest countries in the world, 11 have annual per capita GNPs ranging from US$80 (Mozambique) to US$210 (Sierra Leone), 10 from US$220 (Bangladesh) to US$340 (Nigeria), 9 from US$370 (China) to US$450 (Nicaragua), and 10 from US$500 (Sri Lanka) to US$600 (Zimbabwe) (WHO 1995). The relationship between income and some health indicators for these groups are shown in Table 2. Despite the difficulty in making accurate intercountry comparisons, it is clear that improvements in economic status are associated with better health statistics, and that

economic growth is necessary to improve the living conditions conducive to good health, independent of benefits that may arise directly from health care.

3. Health Care Expenditure

Since the 1960s, major advances in medicine and technology have been associated with escalating expenditure on their application in health care—most of this in highly industrialized countries. Annual per capita expenditure on health care ranges from $4,000 in the US to less than $5 in the poorest countries in Africa. Half the world's population lives in countries that cannot afford annual per capita health expenditures of more than $5–10, and many people do not have access to even basic drugs. The WHO has estimated that in 1990 the annual cost of providing a basic package of public health and essential clinical services in a low-income country was at least $12. In most such countries health care expenditure is typically less than $6. Even though some parasitic diseases (for example, onchocerciasis, schistosomasis, and lymphatic filariasis) could be controlled by mass treatment campaigns using inexpensive drugs, the infrastructure required to provide such coverage is inadequate. Effective treatment of diseases such as tuberculosis and leprosy, and HIV/AIDS requires more complex infrastructures than can be afforded on current health care budgets in poor countries.

In the 1990s, 89 percent of annual world expenditure on health care was spent on 16 percent of the world's population who bear 7 percent of the global burden of disease (in DALYs) and less than 10 percent of medical research expenditure was on health problems of the developing world. These are examples of global injustice that we should surely find intolerable if we have a real commitment to universal human rights and human dignity. Vaccine development programs and drug donation programs for poor countries, admirably promoted and supported by pharmaceutical companies, are necessary but insufficient responses to such injustices. Modest increments in economic status combined with a provision of a basic health care package have the potential to improve health in poor countries significantly. However, it should be noted that health is not merely a resource issue even in rich countries—as exemplified by comparative health statistics around the world (Wilkinson 1996).

4. The Global Political Economy and Health

The world at the end of the twentieth century is thus characterized by widening disparities in economic and health status (within and between countries), and by suffering, conflict, and alienation associated with pervasive social forces (Wilkinson 1996, Sivard 1996, Benatar 1998). Erosion of the economy of many poor countries, under the impact of globalization and the 'debt trade,' has obstructed the introduction of effective forms of modern medicine into many poor countries and thus prevented achievement of widespread access to even basic health care for billions of people (Falk 1999). Third world debt requires special mention. The ways in which such debt have been created, the fact that third world debt is a small component of total world debt and can never be repaid, and that in sub-Saharan Africa four times as much is spent on debt repayment each year than on health and education combined, makes insistence on repayment unconscionable (Pettifor 1996).

The emergence of new diseases such as AIDS, that afflict predominantly those marginalized by poverty (80 percent of HIV-positive persons live in the poorest countries in the world), has been attributed to the social and environmental niches created by the nature of the global political economy and its ideology (Lee and Zwi 1996). Failure to appreciate such associations will result not only in failure to control such diseases as HIV/AIDS (as with tuberculosis in the past, Benatar 1995), but more importantly, will probably favor the emergence of new infectious diseases in the future. The changing spatial dimensions of human relations resulting from urbanization, migration, travel, ethnic conflict, and displacement of people and refugees, facilitates the spread of such diseases from which no one should feel immune.

5. Access to Health Care, and the Nature of Health Care Systems

Today many countries consider access to at least a minimum package of health care as a basic human right that nation-states should be committed to providing for their citizens. Some form of socialized and equitable health care is provided in all Western European nations and in Canada.

Medical care, delivered as a marketable commodity—albeit with considerable state assistance for the poor and the aged—is the example that has been set by one of the wealthiest nations in the world. However, this results in many citizens having inadequate access to health care. In considering how to overcome this, while recognizing that private/public alliances cannot be avoided and indeed may be beneficial, Buchanan has provided cogent arguments for the nine elements he considers essential for a just health care system. He argues further that achieving a just health care system as a cooperative between the public and private sectors requires an organization, such as the state, to act as a coordinator and guarantor of equitable access to care and fair distribution of costs. Such a balance between private and public medicine with appropriate coordination could over-

come the deficiencies of either system when one dominates (Buchanan 1995; see also Benatar 1996).

The effects of intensive privatization, aided and abetted by the structural adjustment programs promoted by the International Monetary Fund (IMF) and the World Bank, have been much more marked in poor countries, with severe adverse effects on health. Acknowledging the need for, and the right to, universal access to a basic health care package and achieving this goal pose challenges for the future. The WHO's renewal strategy for health-for-all places emphasis on equity, solidarity, and appropriate technical, political, and economic strategies that could promote health and sustainable health care systems as central to development. Here too an appropriate balance needs to be struck between public and privately funded medicine. How this balance can be struck and how health can be coordinated within an effective strategic alliance between many parties remain hotly debated subjects.

6. *Traditional Healers*

Despite the pervasiveness of Western influences, traditional healers have survived and play an increasing role in some countries. For example, while there are about 25,000 Western trained medical practitioners in South Africa, and Western medicine has a powerful foothold in this middle-income country, over 200,000 traditional healers carry out a thriving trade. Increasing contact and collaboration between Western medicine and traditional healers is being encouraged and the potential impact of their contributions to health, in particular through development of a primary health care approach, in many countries in Africa in the future should not be underestimated (Van Rensburg and Benatar 2000).

7. *Prospects for the Future*

As we stand at the threshold of a new era in which exponential increments in genetic information will be applied to medical practice, we need to contemplate whether such new power will be used for the benefit of all humankind or whether it will be reserved for the privileged. Acquiring a deeper understanding of the forces promoting health and diseases in developing countries needs to be coupled to a commitment to address these if health and human flourishing are to be achieved on a wider scale (Benatar 1999).

See also: African Studies: Health; Economics: Overview; Health Economics; Health Policy; Infant and Child Mortality in the Less Developed World; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Third World

Bibliography

Benatar S R 1995 Prospects for global health: Lessons from tuberculosis. *Thorax* **50**: 487–9

Benatar S R 1996 Editorial on Buchanan (1996). *British Medical Journal* **313**: 1567–8

Benatar S R 1998 Global disparities in health and human rights. *American Journal of Public Health* **88**: 295–300

Benatar S R 1999 A perspective from Africa on human rights and genetic engineering. In: Burley J (ed.) *The Genetic Revolution and Human Rights*. Oxford Amnesty Lectures. Oxford University Press Oxford, UK, pp. 159–89

Buchanan A 1995 Privatization and just health care. *Bioethics* **9**: 220–39

Desjarlais R, Eisenberg L, Good B, Kleinman A 1995 *World Mental Health: Problems and Priorities in Low-income Countries*. Oxford University Press, New York

Falk R 1999 *Predatory Globalization: A Critique*. Polity Press, Cambridge, UK

Iglehart J 1999 The American health care system: expenditures. *New England Journal of Medicine* **340**: 70–6

Lee K, Zwi A B 1996 A global political economy approach to AIDS. *New Political Economy* **1**: 355–73

Pettifor A 1996 *Debt, the Most Potent Form of Slavery*. Debt Crisis Network, Christian Aid, London

Sivard R L 1996 *World Military and Social Expenditures*, 16th edn. WMSE, Leesburg, VA

Van Rensburg H C J, Benatar S R 2001 History of medicine in Africa. In: Last J, Lock S (eds.) *The Oxford Illustrated Companion to Medicine*. Oxford University Press, Oxford

Wilkinson R G 1996 *Unhealthy Societies: The Afflictions of Inequality*. Routledge, London

World Health Organization 1995 Poverty and health in developing countries. No. 16 in the WHO series *Macroeconomics, Health and Development*. World Health Organization, Geneva

S. R. Benatar

Health Insurance: Economic and Risk Aspects

Over the course of a year, individuals are at risk for a variety of unfortunate and unpredictable events, e.g., property damage due to fire or hurricane, property loss due to burglary, or physical and mental pain and suffering due to adverse health events. In most cases, affected individuals can purchase goods or services to replace the loss or to alleviate the adverse effects of an untoward event. In some cases, it is possible to reduce the probability that an adverse event will occur. In many instances, through the purchase of insurance, it is also possible to reduce financial loss associated with an adverse event. In general, people dislike risk and are willing to trade a small amount of money, in the form of an insurance premium, for protection against a potentially large loss in income. (Depending on what kind of health insurance system, particularly whether it is a private or public system, the payment can be

either a private insurance premium paid out of pocket or through a reduction in wages or a public premium in the form of taxes. For simplicity, we will refer to the payment as a premium. Similarly, the bearer of risk under insurance can be a private company or a public governmental body. We will refer to the bearer of risk as the insurer.)

This entry outlines the basic economic models of health insurance, emphasizing the relationship between risk and insurance demand. It also discusses the distortions in the market for health care services created by health insurance.

1. Risk and the Demand For Insurance

The relationship between risk and insurance demand is derived from the relationship between utility and wealth (for additional discussion see Arrow 1963). Utility is a function of wealth, with more wealth leading to higher levels of utility or satisfaction. Although utility rises with wealth, it is generally assumed to do so at a decreasing rate. This property is called diminishing marginal utility. Each incremental increase in wealth provides a smaller incremental increase in utility. It therefore follows that the gain in utility associated with any incremental gain in wealth is less than the loss in utility associated with an equivalent loss of wealth.

The property of diminishing marginal utility implies individuals are *risk averse*. That is, individuals would prefer to have any level of wealth with certainty than a gamble providing the same level of wealth on average. For example, a risk averse individual would prefer $100 dollars with certainty as opposed to a gamble with a 50 percent chance of winning $200 and a 50 percent chance of winning $0. Risk aversion is an inherent property of a concave utility function. Because risk averse individuals prefer certainty, the premium they are willing to pay for an insurance policy that removes risk exceeds the actuarially fair premium (AFP) of the insurance, which is the amount that the insurer would have to pay out, on average, for this policy. The gap between the premium individuals are willing to pay and the AFP is termed the *risk premium* (Phelps 1997).

The risk premium allows insurers to cover expenses above the medical payout, e.g., claims processing. Several factors determine the size of the risk premium an individual will be willing to pay. First, risk aversion is likely to vary with wealth. For very different reasons, both individuals with very high and those with very low levels of wealth are less willing to pay a risk premium (Feldstein 1999). The marginal disutility of an incremental decrease in wealth falls at high levels of wealth, reducing any utility gain to avoiding risk. In contrast, the marginal disutility becomes very large at very low levels of wealth, making the opportunity cost of purchasing insurance too high.

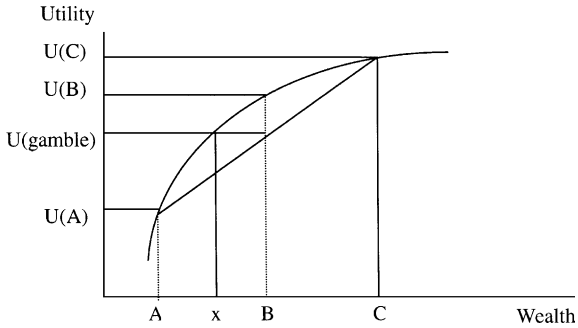

Figure 1
Expected utility and risk

Second, the probability of a loss will influence the size of the risk premium. As the probability of the loss approaches 1, the willingness to pay for insurance rises, but more slowly than the increase in the AFP for that individual. Thus the risk premium falls. In the extreme case, when the probability of a loss equals 1, the risk premium goes to zero. In this case, there is no risk and individuals would not be willing to pay any risk premium. Similarly as the probability of a loss goes to zero, both the willingness to pay and AFP fall, but the willingness to pay falls faster and eventually the risk premium equals zero. Individuals would not pay for insurance if the probability of a loss equaled zero.

Third, the magnitude of the loss affects the risk premium individuals are willing to pay. A greater loss represents an increase in the variance of income. An individual will be willing to pay a higher risk premium for a higher cost illness.

The utility curve depicted in Fig. 1 can be used to analyze the demand for insurance described above. Throughout the analysis we assume individuals know the probability they will suffer a loss (or benefit from a gain). Because greater wealth leads to higher utility, the utility at point C, U(C), is greater than the utility at points A or B. Diminishing marginal utility implies that if point B is equidistant from points A and C, the utility gained from moving from B to C is less than the utility lost from moving from B to A.

Consider an individual at wealth B evaluating a gamble with outcomes A and C, each of which has a 50 percent probability of occurring (e.g., tossing a fair coin). If point B is equidistant from points A and C, the expected wealth resulting from the gamble, i.e., where *on average* the individual could expect to be if tossing the coin, is the initial level of wealth, B. The expected utility of such a gamble, EU (gamble), is the probability weighted sum of the two outcomes A and C and, because each outcome has a 50 percent chance of occurring, can be determined by finding the midpoint on the chord connecting A and C. Because the utility curve is concave, even though the initial level of

wealth is the same as the expected wealth of the gamble, the utility of B, U(B), will exceed the expected utility of the gamble, EU (gamble). This utility gain of avoiding risk is key to the demand for insurance.

Models of health insurance fit exactly into this model. For simplicity, assume a world in which there is only one type of adverse health event and spending in the unhealthy state is unaffected by the presence of insurance. Individuals are assumed to start with wealth level C and remain there if healthy during the year. If they suffer an illness shock, they will then spend C − A on health care services. Assuming the probability of a loss is 50 percent, the expected loss is C − B. Expected wealth, under uncertainty, is B. Without insurance the individual would have expected utility of EU (gamble).

Now imagine individuals could purchase an insurance contract that would pay the costs of medical care in the event of an illness. With a 50 percent chance of each individual incurring the loss, the expected cost to the insurer of each enrollee is C − B. The AFP therefore is C − B. If individuals pay the AFP, their wealth level is B, regardless of whether the illness occurs. In other words, the risk of financial loss has been eliminated. The individual has turned a potentially large loss in income into a smaller known loss of income (in the form of the AFP). In this situation, the utility would be U(B) because the income level B is achieved with certainty. As is clear from the figure, the individual prefers to purchase insurance relative to self-insuring. The amount of the utility gain is equal to the vertical distance between U(B) and EU (gamble). The greater the concavity of the utility curve, the greater the risk aversion and the greater the utility gain from insurance.

The utility curve depicted in Fig. 1 also can be used to assess the amount that an individual would be willing to pay for insurance. Utility of a wealth level of x, with certainty, is equivalent to the expected utility of remaining uninsured (facing the risk of having a loss that leaves them at wealth level A). Thus an individual would be indifferent between being uninsured or paying a premium of C − x for insurance. This premium is their maximum willingness to pay for insurance (WTP), with C − B the AFP and B − x the risk premium.

2. *Moral Hazard*

Up until now, the magnitude of the loss associated with an illness was not affected by the presence of insurance coverage for the desired medical care services. Substantial theoretical and empirical work suggests that this is not the case (Pauly 1968, Manning et al. 1987, Newhouse 1992). Specifically, because health insurance lowers the price of medical care services faced by the consumer at the time of purchase, individuals will often consume more care than they would if they paid the full price. The providers of this care *are* paid the full price, however, receiving out of pocket payments from the patient and reimbursement from the insurer.

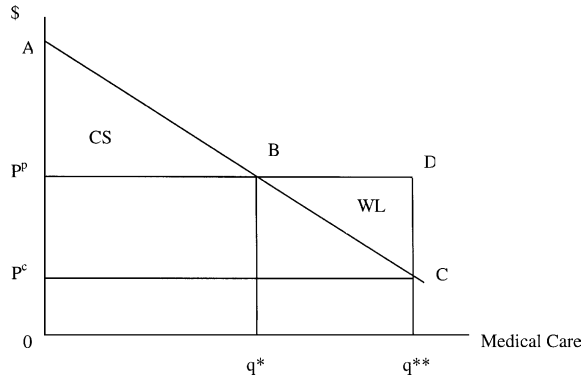

Figure 2
Welfare loss and moral hazard

The consequences of this behavior, known as moral hazard, depend in part on the difference between the cost of the care and the value individuals place on that care. Because individuals do not demand this same level of care without insurance, standard economic models assume that the value they place on the care is less than the full cost of that care. If this is true, insurance causes a loss in consumer welfare because of the insurance-induced increase in consumption, whose costs are passed back to the consumer in the form of higher insurance premiums. The welfare loss is thus incurred by the enrollee at the time the insurance premium is paid, not when care is received. This welfare loss accrues even if, as is the case in most complex insurance markets, the individual is not purchasing health insurance directly and may in fact not be aware of the welfare loss. In aggregate, adjusting for transfers between individuals, the amount paid through premiums exceeds the value of care financed by the insurance.

This analysis is depicted graphically in Fig. 2. The demand curve tells us how much medical care will be demanded for any given price. For example, assume the price of care, as paid to the provider of care, is denoted P^p. In this case, uninsured individuals would purchase q* units of care. If insurance lowered the individual's out of pocket price to P^c, insured individuals would then consume q** units of care.

The demand curve is taken as the measure of value because it also represents the maximum amount an individual would be willing to pay for any given quantity of care. For example the maximum amount an individual would pay for q* is P^p. The overall value placed on all care up to any quantity is the area under the demand curve up to that point. For an uninsured individual, the value of care consumed would be the area under the demand curve up to q* (ABq*0). For

an insured individual, it would be represented by ACq**0.

The total cost of care, paid by the individual and insurer, is the price paid to the provider times the quantity of services provided. In the case of an uninsured individual, it would be P^p times q* (P^pBq*0). Notice that the value of care to the consumer exceeds the costs of the care by the portion of the area under the demand curve not included in the box (P^pAB). This net value to the consumer is termed consumer surplus (CS).

If the individual purchases insurance and then increases consumption, the total cost of care increases to P^pDq**0. Notice the incremental cost of the care purchased, BDq**q*, exceeds the incremental value to the consumer, BCq**q*. The amount by which the incremental cost exceeds the incremental value represents the welfare loss (WL) of insurance and is represented by the triangle BDC.

At the time that care is consumed, the individual only pays a price of P^c, so the out of pocket cost is P^cCq**0. The actuarial cost of the care to the insurer, however, is P^pDCP^c. It is this cost, not the out of pocket cost, that determines the AFP. As noted above, the welfare loss is thus incurred at the time of insurance purchase. If the demand curve were inelastic, that is, if the level of consumption were not very responsive to changes in price, the welfare loss would be relatively small. Consumers, if aware of the welfare loss, would prefer to consume less in exchange for an appropriately lower premium, particularly in cases in which the welfare loss is large (elastic demand).

There are three important assumptions that would affect the size of the welfare loss. First, the analysis assumes that individuals other than the patient receiving the care do not receive positive utility from that consumption. If this positive externality exists, the welfare loss is reduced. Second, the analysis assumes that the price paid to providers is equal to marginal cost (as would be the case if it were set in a competitive market). If the price were set in an imperfectly competitive market, so that it exceeded marginal cost, the welfare loss from a societal prospective would be diminished. Third, the preceding analysis assumes that income effects associated with paying the premium and the transfer of income from healthy to sick individuals have only small effects on the consumption of medical care. Nyman and Maude-Griffin (2001) suggests that if these effects are large, the welfare loss is considerably reduced.

Many different approaches might be taken to minimize the welfare loss associated with insurance. For example, increasing consumer cost-sharing for medical care services would reduce utilization but has the disadvantage of increasing the size of the financial risk faced by the consumer. Increasing provider cost-sharing through a variety of financial incentive systems would reduce utilization, but has the potential disadvantage of too little care being prescribed. Private regulatory mechanisms, such as utilization review, can be implemented to constrain both where and what kinds of covered treatments an individual can receive. Managed care plans in the United States have been using these and other approaches to restrain use. In some cases, plans select providers whose preferred practice style leads to lower levels of utilization, particularly of higher cost alternatives. Some plans also educate both enrollees and providers to keep utilization below that which would otherwise occur. The extent to which various strategies are used will depend on the extent to which plans are in competition with other plans and the preferences of purchasers.

Nationalized systems can use similar strategies, limiting the resources available to provide care or providing financial and nonfinancial incentives to both providers and consumers to keep utilization at lower levels. Nationalized systems may also use their bargaining power to obtain favorable prices for medical care services.

3. *Insurable Hazards*

Given the analysis above, the theory of insurance would make several predictions regarding for which types of risks individuals would likely seek insurance coverage. First, they would seek insurance for services to treat events for which the probability of occurring is neither zero nor one. In other words, individuals would not want to buy insurance for protection against events for which they are not at risk, either because they know they will seek these services or because they know that they are at minimal risk of needing those services. Theory would predict that events with intermediate probabilities, that is, services with high-risk premiums, would generate the most demand. Thus the model would predict that most individuals would not purchase insurance to cover preventive services. The use of preventive services is not stochastic; that is, individuals know the probability of using these services in advance. Second, services that are costly would also have high risk premiums and would be attractive to insure. Third, individuals would not seek coverage for services for which demand is very elastic because such insurance would generate substantial welfare loss.

4. *Adverse Selection*

The preceding discussion was based on the utility curve for a single individual. To move the analysis to the market level, one must consider the behavior of many individuals, each with their own utility curve, initial wealth, and risk. Some individuals will be relatively low risk with low AFPs and others will be higher risk with high AFPs. Most models assume that individuals have more information about the risk that

they face than do the insurers. If an insurer were to base the premium it charged to a defined group of individuals on the estimated average AFP of all these individuals plus some mark-up, there would be some whose WTP was below the premium and others whose WTP was above the premium. If insurance purchase were voluntary, the individuals whose WTP was below the premium would opt not to purchase coverage, leaving a smaller pool of higher risk buyers. The average AFP for those remaining in the pool would rise. Of course as the AFP rose, the insurer would need to raise the premium in the next cycle. Some of the individuals who had remained in the first cycle would find the new premium above their WTP and would also opt out of buying coverage. This nonrandom selection out of the pool would result in an increase in the average AFP for the decreasing number of ever-higher risk individuals still interested in insurance.

Several papers outline scenarios in which insurance markets may collapse due to adverse selection (Akerlof 1970, Rothschild and Stiglitz 1976, Cutler and Reber 1998). In their seminal work, Rothschild and Stiglitz (1976) demonstrate that the set of policies that will be offered in the market is constrained by the ability of insurers to enter the market with new policies. This situation arises because insurers cannot perfectly ascertain the characteristics of the individuals they enroll. Certain policies cannot be offered in equilibrium because new insurers could enter the market and attract the lowest cost enrollees, leaving incumbent insurers with the higher cost enrollees, whose premiums do not cover their expected losses. In this environment, if an equilibrium does exist, it will be a separating equilibrium characterized by high risk individuals purchasing their most preferred actuarially fair policy and low risk individuals purchasing an insurance policy inferior to their most preferred actuarially fair policy, which will not be offered. Equilibrium is more likely to exist when there is a relatively large number of high-risk individuals and when the variation in risks is great.

Subsequent work indicates that the constraint on the set of policies offered depends on the equilibrium concept employed. Specifically, alternative equilibrium concepts can give rise to pooling equilibria (Wilson 1977) or even equilibria in which insurers may use profits on one product to subsidize losses on another (Spence 1978). All of the work in this area emphasizes the point that free entry into a market with asymmetric information limits the set of policies available because of adverse selection. Feldman et al. (1998) demonstrate that compulsory partial coverage may improve consumer welfare relative to a competitive equilibrium or compulsory comprehensive coverage.

Nationalized systems of insurance can reduce or eliminate adverse selection, depending on the degree to which coverage is compulsory and competition occurs within the system. Typically this would benefit those with the highest expected expenditure, at the expense of those with less expected expenditures. The benefits of reducing adverse selection must be weighed against reductions in the ability of individuals to select their most preferred benefit design.

5. *Empirical Evidence*

5.1 *Magnitude of Moral Hazard*

The magnitude of moral hazard associated with the purchase of health insurance depends directly on the elasticity of the demand for medical care, i.e., how responsive consumers are to changes in the price. Manning et al. (1987) report estimates from a randomized trial indicating that the overall elasticity of medical care is -0.17, that is, a 0.17 percent increase in utilization occurs for every 1 percent decrease in price. Demand for hospital services was estimated to be the least elastic (-0.14) and demand for well visits was estimated to be the most elastic (-0.43). These estimates are broadly consistent with other estimates derived from non-randomized settings (Feldstein 1999).

Estimated elasticities for other services tend to be somewhat higher than those for hospital and physician services. For example, elasticity of demand for pharmaceuticals has been estimated at about -0.40 (Smith and Garner 1974). Demand elasticity for nursing home care has been estimated at -0.76 (Lamberton et al. 1986).

5.2 *The Demand for Insurance*

The theory outlined above suggests that insurance demand is likely to be greatest for services with a high expected expenditure. In general, the evidence is consistent with the theory: services with a large estimated loss were more likely to be covered with insurance (Feldstein 1999). In addition, we would expect insurance demand to be greatest for services with highly inelastic demand. Again, evidence is consistent with theory. That is, the less price responsive the demand for the services, and therefore the smaller the moral hazard and associated welfare loss, the higher the level of insurance coverage (Phelps 1997).

Taken together, these factors explain much of the variation in service coverage witnessed among working adults in the U.S. For example, 76 percent of all employees participate in the medical care benefit offered by their employer, whereas only 59 percent participate in the dental care benefit, and 26 percent in the vision care benefit (USDOL, BLS, p. 6). Of those participating in the medical care benefit, 100 percent have coverage for hospital room and board, inpatient surgery, and inpatient physician visits; 85 percent have

coverage for home health care, 60–66 percent for well-baby care, physical exams, and hospice care, 52 percent for immunization and inoculation, and only 35 percent for hearing care (same source, p. 48).

Although measuring risk aversion is difficult, Barsky et al. (1997) used survey responses from the Health and Retirement Study regarding a respondent's willingness to take a new job with uncertain earnings to classify individuals based on their degree of risk aversion. They then related this measure of risk aversion to a variety of behaviors, including the purchase of health insurance. Stratifying by employment status (self-employed, employed, and not employed), they find that within each strata, the more risk averse respondents are more likely to purchase insurance.

5.3 Adverse Selection

Considerable evidence suggests that when given a choice individuals make health plan decisions based on their anticipated medical expenditures, thus generating nonrandom and, from the insurer's perspective, adverse selection. Much of the literature examines enrollment patterns when individuals face multiple insurance options that vary in terms of generosity. Even though a variety of data sources and methods have been used to investigate the self-selection into a range of plan options, the evidence suggests that adverse selection is quantitatively large (Cutler and Zeckhauser 2000).

6. Conclusions

Health insurance provides value to individuals by reducing the risk of a large financial loss associated with an adverse health event. This allows individuals to obtain medical care that they otherwise might not be able to afford. The payment from the insurer serves as a transfer of income to individuals who become ill.

However, if left unchecked, health insurance markets generate two deleterious effects. First, by lowering the patient's cost of consuming medical care, insurance leads to an increase in utilization (moral hazard). This increased consumption results in welfare loss. Second, insurance markets have a tendency to be unstable because of adverse selection. In the extreme, non-random selection could cause insurance markets to unravel, leaving some or all individuals uncovered and distorting the set of plans offered in the market.

Most health insurance systems, whether private or public, incorporate a variety of strategies to limit the welfare loss from the overconsumption of medical care and the effects of adverse selection. Inherently these strategies involve restrictions on individual behaviors. The challenge in most systems is to mitigate the unwanted consequences of insurance as much as possible while still accomplishing the fundamental goal of reducing risk.

Bibliography

Akerlof G A 1970 The market for 'lemons.' *Quarterly Journal of Economics* **84**: 488–500

Arrow K 1963 Uncertainty and the welfare economics of medical care. *American Economic Review* **53**: 941–73

Barsky R B, Juster F T, Kimball M S, Shapiro M D 1997 Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *The Quarterly Journal of Economics* (May): 537–79

Cutler D M, Reber S J 1998 Paying for health insurance: The trade-off between competition and adverse selection. *The Quarterly Journal of Economics* **113**(2): 433–66

Cutler D M, Zeckhauser R J 2000 The anatomy of health insurance. In: Culyer A J, Newhouse J P (eds.) *Handbook of Health Economics*. Elsevier, Amsterdam, Vol. 1A, Chap.11, pp. 564–643

Deaton A, Muelbauer J 1987 *Economics and Consumer Behavior*. Cambridge University Press, Cambridge, UK

Feldman R, Escribano C, Pellise L 1998 The role of government in health insurance markets with adverse selection. *Health Economics* **7**: 659–70

Feldstein P J 1999 *Health Care Economics*, 5th edn. In: Williams S J (ed.). Delmar publishers.

Lamberton C E, Ellingson W D, Spear K R 1986 Factors determining the demand for nursing home services. *Quarterly Review of Economics and Business* **26**: 74–90

Manning W G, Newhouse J P, Duan N, Keeler E B, Leibowitz A, Marquis M S 1987 Health insurance and the demand for medical care: Evidence from a randomized experiment. *The American Economic Review* **77**(3): 251–77

Miyazaki H 1977 The rat race and internal labor markets. *Bell Journal of Economics* **8**(3): 394–418

Newhouse J P 1992 Medical care costs: How much welfare loss? *Journal of Economic Perspectives* **6**(3): 3–21

Nyman J A, Maude-Griffin R 2000 The welfare economics of moral hazard. Presented at the *11th Annual Economics Conference*, Washington, DC

Nyman J A, Maude-Griffin R 2001 The welfare economics of moral hazard. *International Journal of Health Care Finance and Economics*. **1**: 23–42

Pauly M V 1968 The economics of moral hazard. *American Economic Review* **58**(3): 531–7

Phelps C E 1997 *Health Economics*, 2nd edn. Addison-Wesley, Reading, MA

Rothschild M, Stiglitz J 1976 Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 629–49

Smith M C, Garner D D 1974 Effects of a medicaid program on prescription drug availability and acquisition. *Medical Care* **12**(7): 571–81

Spence A M 1978 Product differentiation and performance in insurance markets. *Journal of Public Economics* **10**(2): 427–47

US Department of Labor, Bureau of Labor Statistics 1997 *Employee Benefits in Medium and Large Private Establishments, 1993, 1994*

Wilson C 1977 A model of insurance markets with incomplete information. *Journal of Economic Theory*. **16**: 167–207

C. G. McLaughlin and M. E. Chernew

Health Interventions: Community-based

This article considers theories and methods underlying successful community-based health interventions, emphasizing four exemplars. Three exemplars relied heavily on modern means of mass communication and social marketing, relatively low-cost approaches with the potential to reach and change lifestyle behaviors of entire populations, in contrast to traditional individual or group counseling. One exemplar, studying alcohol-involved trauma, used community organizing for adherence to existing laws, rather than public education for behavior change. The article also describes barriers to change and lessons learned for future extensions of such endeavors.

By definition a community-based program is organized locally, achieving its goals through the community's institutions and communication networks. A community here is defined as a residential area with legally defined geographic boundaries, where a governmental system regulates certain local aspects of schools, businesses, transportation, policing, and recreational activities. A community is ordinarily last in a nation's regulatory chain, where education must ultimately occur, although for rural areas the *county* becomes the governing agent for education.

The relevance and intellectual context of community-based health education are addressed through these questions. (a) What is the rationale for community-wide interventions rather than in more limited locales, such as clinics, hospitals, work sites, or schools? (b) What theories, methods, and resources underlie success? (c) How does a community become committed to a novel health education campaign? (d) How can barriers to success be overcome in such interventions? (e) What roles do community activism, laws, and regulations play? (f) What is the future of community-based health education?

1. *Rationale*

Health education in schools, work sites, and clinics have a long history of individual success, but synergistic interactive effects can occur when they are imbedded in a total community campaign that adds inherently cost-effective mass media and environmental change (Schooler et al. 1997). Lifestyle changes, so powerfully influenced by custom and by the media in developed countries, cannot be countered effectively through simple means. Community-wide approaches fit the public health model because the traditional medical model can neither prevent most chronic disease nor reach the entire population in need. The community provides influence through locally produced electronic and print mass media and through work sites, schools, medical and hospital settings, and social gatherings of many sorts. It offers opportunities for health-promoting regulations, such as providing opportunities for physical activity for all, school fitness classes, healthful school lunches, alcohol sales limits, and preventing tobacco product marketing to children (see *Health Promotion in Schools*).

Studies at the Stanford Center for Research in Disease Prevention (SCRDP) have shown that only multiple influences produce meaningful changes in the dietary, exercise, or tobacco-use behavior of adults, adolescents, and children. For adolescents, always resistant to health behavior change, the following influenced success: parent and teacher roles; school environments; amount and quality of school-based education on tobacco, fitness, and nutrition; peer influence; and amount and quality of community-wide health education. (Also, certain personal characteristics were influential, for example, self-efficacy toward one's behavior-change abilities; Fortmann et al. 1995.)

2. *Theoretical Formulations*

Community-based health education carried out by SCRDP and by some analogous projects have been guided by three major theoretical and management models (Flora et al. 1989).

(a) The community organization framework provides methods of identifying the health problem (and resources needed or available), mobilizing the community's opinion leaders and organizations, gaining populace support, forming coalitions, launching and maintaining education programs, achieving regulatory changes, and empowering communities to reach and maintain their goals.

(b) Social marketing provides methods of designing and sequencing messages for population subgroups based on their health needs, cultural attributes, social networks, media habits, and readiness for change.

(c) The mass communication–behavior change model describes theories underlying educational content, such as social cognitive theory (Bandura 1986) and methods of reaching and influencing the total population in multiple ways (see *Self-efficacy and Health*).

3. *Intervention Principles, Methods, and Resources Needed for Success*

3.1 *Acceptance*

Both relevant organizations and political and opinion leaders must be recruited into a coalition that sanctions any novel health education project. This

coalition must identify the health problem (needs assessment), provide a resource inventory, obtain populace support, and plan the intervention.

3.2 Comprehensive

Interventions must go beyond attempting to change knowledge, the usual goal of an educational system, by providing training in behavior-change skills. They must also go well beyond the individual, enlisting multiple community organizations in campaigns for change and seeking changes in the social environment and in regulations that promote access to the facilities and resources needed for healthful practices. Multiple communication channels (i.e., both mediated and face-to-face) are needed to reach different audiences, recognizing their preferences, stage of knowledge, and readiness for change.

A comprehensive intervention should involve schools, work sites, churches, and facilities for sport, recreation, and health. These organizations and others can serve as education conduits, with the community's electronic and print mass media organizations assisting in message design and delivery. The Internet provides a new channel, whose community education role is as yet poorly defined. Interactive computer learning, now becoming common in classrooms, can be designed for large groups—an emerging variant of mass media.

3.3 Variety

Comprehensiveness requires variety. For example, in tobacco control components of Stanford's CVD campaigns, local medical clinics, dental offices, pharmacies, and libraries distributed a low-cost skills-training Quit Kit; a local smoking cessation class, was shown on television; many newspaper articles and columns appeared; a local business supported costs of a smoking cessation contest; and all newspapers and electronic media 'cross-advertised' activities designed for mass audiences.

3.4 Integration

Success requires an appropriate mix and sequence of programs delivered through varied channels. This integration, with goals set in advance and goal changes based on early results, is analogous to a commercial marketing campaign, hence the term 'social marketing.'

3.5 Message Design and Audience Segmentation

Social marketing requires message tailoring to fit each subgroup's needs and preferences, taking into consideration cultural differences, learning styles, and preferred learning sites. A message sequence should increase awareness, then, increase knowledge, and last, increase motivation and provide training in the skills needed for adoption and maintenance of a new behavior (Bandura 1986). Electronic media can carry out the first two parts of this sequence and stimulate use of the more information-dense print media of newspapers and booklets, which are inherently more effective in skills training than are electronic media (Flora et al. 1997).

Messages must be clear, focused, and salient. Salience requires broad reaching media, arousing interest and awareness—topics must break through passive indifference engendered by the information overload of many societies and become 'on the public agenda.' Given the large advertising budgets of today's mass media, health agencies' messages must be of sufficient production quality to compete for the public's attention.

3.6 The Message Must be Effective

Behavioral research confirms the rather self-evident rule that 'learning by doing' is more effective than 'learning from observing' (modeling a behavior), and both are more effective than an 'information-only' approach that changes knowledge alone. These principles are contained in Bandura's social cognitive theory, which posits that guided practice in a new behavior can lead to increased self-efficacy and to greater behavior change (Bandura 1986). Thus, 'knowledge-only' campaigns have been found less effective than those that apply Bandura's recommendations (see *Self-efficacy and Health*).

3.7 Intervention Dose

The dose needed depends on many factors: lesser amounts are needed in smaller communities, at earlier stages in a country's adoption of a 'health innovation,' and when the advocated behavior change is reasonably simple (such as mammography and hypertension screening and immunization campaigns). Clearly, more complex changes are needed in individuals and in society's norms to alter eating or exercise patterns or to control tobacco use. Complexity in respect to nutrition arises from many sources, including long-standing cultural beliefs and practices; entrenched methods in agriculture, food production, and retailing; advertising of 'unhealthful' foods; and the advent of widespread fast-food chains that are dominated by commercial interests unresponsive to local demands and needs.

Few projects have measured intervention dose. One excellent method records the total number and duration of messages distributed over a defined time period, albeit with a defect due to lack of message

quality measures (Farquhar et al. 1990). However, even this rough method, were it widely adopted, would be of great value to public health practitioners.

3.8 Role of Activism and Advocacy

Successful community-based health education requires effective leaders, community activists with the courage and charisma to advocate health innovations. Waves of advocacy filtering down from international, national, state, or provincial sources can wash up on the shore of an inert community to provide a local activist leader with the popular support to 'fight city hall' or other entrenched bureaucrats who defend the status quo. Tobacco control in Australia and the US provide examples. National and state advocacy groups with access to mass media created a strong mass movement for change, allowing advocates to enlist popular support for local tobacco control measures. In both California and Australia's State of Victoria this popular support led to statewide increases in tobacco taxes, with some retained for education against tobacco, a measure that had been resisted by state legislators who had long been influenced by tobacco lobbyists—an example of community activism filtering up the chain to a higher political level (Catalonia Declaration 1996).

3.9 Role of Policy, Laws, and Regulations (PLR)

Local actions can affect alcohol and tobacco sales and create environments that improve nutrition and enhance physical activity. However, national, state, or provincial actions can magnify local PLR and education efforts on topics such as tobacco taxation, automobile seat-belt laws, food and drug safety, school nutrition, school physical activity policies, and (in the United States) laws on firearms. As described in Section 4.8, widespread popular attitude changes in numerous communities can also affect the political process at the state or federal level.

4. History of Comprehensive Community Health Education

The history since the 1970s is restricted largely to 13 formal research projects designed to affect CVD risk factors. They involved entire populations of at least one education community, compared with at least one control community. Three of the four exemplars to follow are drawn from these 13 projects. Dissemination into practical applications of community organizing and mass communication technologies, derived in part from these research projects, occurred throughout many parts of the world.

4.1 The First Two Exemplars: The First Decade

The first and second exemplars, Stanford Three-community Study (TCS), in three small agricultural marketing towns in California (total population 45,000), and the North Karelia Study (NKS), in two adjoining predominately rural Finnish counties (North Karelia population about 180,000), each began in 1972. TCS, the first Stanford project, was carried out from 1972–75 in both English and Spanish, comparing effects of mass media alone in one community and mass media plus 10-session risk reduction classes for high-risk adults in a second, with a third as a control (Schooler et al. 1997). Groups exposed to varied education amounts showed a dose-response change in smoking, blood pressure, and blood cholesterol, with a proportionately larger effect in the Spanish-speaking residents than in the Anglo majority. (This minority population outcome required a compensating increase in intervention resources.) A composite CVD risk reduction of about 23 percent and 30 percent occurred in the mass media—only and mass media—plus classroom conditions, respectively. Thus, a relatively modest amount of mass media (about 30 television and radio 'spots,' weekly newspaper columns on heart health, and four separate mass mailings of booklets) was sufficient to change the population's body weight, cholesterol and blood pressure levels, and smoking prevalence. The education followed Bandura's principles but adapted to the mass media (as was also the case with exemplars two and three).

NKS, evaluated two, matched, rural Finnish counties that contained many villages with farming and lumbering as the main occupations. North Karelia (about 180,000 population) received an education campaign that began in1972, continuing to the present. After 5 years, CVD risk factor changes comparable to the TCS occurred, and significant net reductions in CVD events occurred (Schooler et al. 1997). This study was marked by extensive community organizing, resulting in strong partnerships with residents and their organizations. The NKS influence on its country's policies was unparalleled among the CVD projects, providing its most important lesson—that a well executed project led by respected scientists can move an entire country. As examples, Finland's food and agricultural industries made large changes: Fertilizers were supplemented with selenium (a substance low in Finland's soil that is needed for health), milk pricing was changed (based on protein instead of fat content), programs were created to replace dairy farms with berry farms, a new canola industry replaced jobs lost in dairying, and increased production of low-fat foods occurred (Catalonia Declaration 1996).

In 1972 population-wide nutrition change and smoking cessation interventions were an innovation internationally, which may partly explain the success of these two pioneering programs (TCS and NK).

4.2 Early International Diffusion, 1977–1983

Studies patterned after the TCS done in Italy, Australia, Switzerland, and South Africa all had significant risk factor changes, adding evidence for the effectiveness of the TCS model (Farquhar et al. 1991).

4.3 The Second and Third Decades: Projects Begun in the 1980s and 1990s

The third exemplar is the Stanford Five-city Project (FCP) (USA). Its intervention phase from 1980 to 1986 extended TCS methods to larger populations (total population of about 360,000) with multifactor CVD prevention directed at two northern California cities. There were three control cities (Farquhar et al. 1990). It differed from TCS in greater use of community organizing and in greater collaboration with the communities' health, media, and education organizations in planning and implementing programs. It was similar in generous use of mass media, both print and electronic (63 percent television and 36 percent print). It was unique in measuring total dose of education (about 5 hours/year and about 100 episodes of exposure/year to all forms of media and classroom education).

FCP's initial year of television messages (defined as 'high reach/low involvement') stimulated the public's use of print media, which supplied more effective skills training than television (Flora et al. 1997). Results were comparable with the TCS (about a 15 percent fall in composite risk of CVD), with a major impact on blood pressure (Farquhar et al. 1990). Health bureaucracies, usually timid, should gain courage from the FCP's 'David and Goliath' demonstration that only 3 hours/year of high-quality television health education can counteract the public's exposure to about 100 hours/year of television advertising devoted to unhealthy nutrition. Although all preceding CVD studies showed effects in small towns and/or rural districts, FCP showed benefit in cities (with populations as high as 100,000). Also, these effects occurred despite the advent (at least in the state of California) in the 1980s of dual working families and increasing public use of fast food, factors that made achieving nutrition-behavior change more difficult.

FCP's modest resources and greater effects, as compared with some comparable projects, support the benefits of the skills training aspects of Bandura's social cognitive theory. Perhaps the most practical lesson to policymakers is that adult residents saved 30 times more money from their decreased cigarette purchases ($120/adult/year) than the cost of the campaign ($4/adult/year)—savings retained by the community. Lastly, the communities adopted FCP's technologies, later applying them to seat-belt promotion and violence and adolescent pregnancy prevention.

Other successful CVD projects, both large and small, occurred in these decades in the US, Sweden, Denmark, Canada, the Czech Republic, and China. Also, World Health Organization-sponsored projects began in about 23 other countries (Schooler et al. 1997). In all instances, they borrowed heavily from the experiences of the three exemplars and in many instances received training from either the Stanford or the North Karelia groups.

4.4 Community Projects in Other Health Topics

Interventions on alcohol, mammography, tobacco control, motor vehicle injuries, and HIV/AIDS are prominent examples of other community-based projects, but with fewer well-controlled studies than in CVD. One of these, Preventing Alcohol Trauma (PAT), is the fourth exemplar. This 5-year study of three US communities showed a 10 percent decrease in alcohol-related traffic injuries and a 50 percent decrease in adolescent alcohol use (Holder et al. 1997). Coalition building and organizational behavior change among the police, alcohol sales outlets, and alcohol servers were the main interventions. The public responded to a fear, instilled through publicity, of the penalties of greater enforcement of existing laws on underage alcohol purchases or drinking and driving. Therefore, in contrast to the three CVD exemplars, PAT showed that major public education is not needed for large public behavior changes. PAT estimated a cost saving of $2.88 for every dollar invested, a number close to that found in many worksite health promotion studies. PAT found their communities had the required infrastructure for the campaigns, requiring only training provided by one indigenous community coordinator, a part-time clerk, an imaginative plan, and the will to proceed.

TCS, the first, and PAT, the fourth exemplar, are opposites: TCS had a maximum of mass-media education and a minimum of organizing, whereas PAT had the opposite. Thus, either model works, but to change complex behaviors, one needs sophisticated behavior-change methods, such as those adapted from Bandura (see *Self-efficacy and Health*). If fear of arrest for breaking existing laws on alcohol use suffices to change personal drinking-and-driving habits, then the education process is simpler.

5. Lessons Lead to a 'Master Plan'

A five-step approach emerges from these studies.

(a) Define the problem. Does the community need a health education campaign? A decision can be made from national or local survey data.

(b) Organize the community, creating a campaign that includes education, continued organizing, training of community organizations, empowerment of the public and the community's organizations, and future

maintenance of programs developed during an initial phase of about 2 years.

(c) Implement an intervention, delivering 3–5 hours of education exposure/year for 2 years, using initially about 50 percent television to arouse interest, followed by about 25 percent from more specific print media (emphasizing newspapers, if available) and 25 percent from community events and programs (such as health fairs, contests, and classes).

(d) Institutionalize programs. The community becomes a demonstration project, with its 'empowered' organizations functioning as health promotion resource centers for a wider region.

(e) Use the new community resources and its residents' potential power to advocate for local, regional, and national governmental regulations and laws that will increase local intervention effects and extend them beyond the community.

Any individual or organization that wishes to engage in community-based health education should gain courage from the words of Margaret Mead, as quoted in the closing passage of the Catalonia Declaration (1996, p. 75), 'Never doubt the capacity of a few dedicated individuals to change the world, in fact, it is the only way it ever has.'

6. Barriers and Challenges

A healthy city has been defined by the absence of crime, crowding, and poverty and by the presence of educated residents and enlightened (and trained) organizations. Together these lead to a community empowered to solve its social problems (i.e., to increase its social capital) (Travers 1997).

Considering barriers, MacIntyre (2000) found Glasgow's environmental factors to be major barriers to healthful exercise behavior. Certain macroeconomic factors inherent in globalization, such as capital flight and increased wealth and income gaps, have been described as barriers to planned change (Bezruchka 2000, Cahill 1983). All such barriers threaten community stability, inhibiting greatly the success of health promotion attempts. However, wise compensatory resource allocation can overcome many barriers, as was shown in the Hispanic minority of the TCS. Therefore, the challenge for the future is: Responsibility for success in community-based health interventions lies with the interventionist, not with the community's residents!

7. Conclusions

Three decades of the 'total community' health education approach strongly support the feasibility, at relatively low cost, of achieving transfer of public education technologies to a community's infrastructure (public health, media, schools, etc.), resulting in significant changes in health habits of populations. Although most studies derive from small communities, recent successes in Tianjin (China), a city of 400,000 (two exemplars examined populations of > 100,000), suggest that the model also works in large populations (Schooler et al. 1997). Organizing and educating communities requires advocacy, activism, coalition building, and leadership; success is enhanced by regulatory change. Theory matters: when the population gains self-efficacy through education, the result—community efficacy—enhances capacity to change institutional policy and practice, thus maintaining community change.

Science cannot serve society if its evidence for educational benefit is ignored. This is not a new concept. As written over 2,000 years ago, 'If a virtuous and learned scholar aims to influence the people as a whole, one must first educate the people' (Legge 1983).

See also: Community Health; Drugs: Illicit Use and Prevention; Health Behaviors; Health Education and Health Promotion; Health Psychology; Mental Health: Community Interventions; Public Health; Public Health as a Social Science; Smoking Prevention and Smoking Cessation; Socioeconomic Status and Health Care; Substance Abuse in Adolescents, Prevention of; Sun Exposure and Skin Cancer Prevention

Bibliography

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ

Bezruchka S 2000 Is globalization dangerous to our health? *Western Journal of Medicine* **172**: 332–4

Cahill J 1983 Structural characteristics of the macroeconomy and mental health: Implications for primary prevention research. *American Journal of Community Psychology* **11**: 553–71

The Catalonia Declaration: Investing in Heart Health 1996 Declaration of the Advisory Board, Second International Heart Health Conference 1995. Department of Health and Social Security of the Autonomous Government of Catalonia, Barcelona, Spain

Legge J 1983 *The Chinese Classics: Confucian Analects, The Great Learning, The Doctrine of the Mean*. Southern Materials Center, Taipei, Republic of China

Farquhar J W, Fortmann S P, Flora J A, Maccoby N 1991 Methods of communication to influence behaviour. In: Holland W W, Detels R, Knox G (eds.) *Oxford Textbook of Public Health*, 2nd edn. Oxford University Press, Oxford, UK, Vol. 2

Farquhar J W, Fortmann S P, Flora J A, Taylor C B, Haskell W L, Williams P T, Maccoby N, Wood P D 1990 The Stanford Five-city Project: Effects of community-wide education on cardiovascular disease risk factors. *Journal of the American Medical Association* **264**: 359–65

Flora J A, Maccoby N, Farquhar J W 1989 Communication campaigns to prevent cardiovascular disease: The Stanford studies. In: Rice R, Atkin C (eds.) *Public Communication Campaigns*. Sage, Newbury Park, CA, pp. 233–52

Flora J A, Saphir M N, Schooler C, Rimal R N 1997 Toward a framework for intervention channels: Reach, involvement and impact. *Annals of Epidemiology* **S7**: S104–S112

Fortmann S P, Flora J A, Winkleby M A, Schooler C, Taylor C B, Farquhar J W 1995 Community intervention trials: Reflections on the Stanford Five-city Project. *American Journal of Epidemiology* **142**(6): 576–86

Holder H D, Salz R F, Grube J W, Treno A J, Reynolds R I, Voas R B, Gruenewald P J 1997 Summing up: Lessons from a comprehensive community prevention trial. *Addiction* **92**(S2): S293–S301

MacIntyre S 2000 The social patterning of exercise behaviours: The role of personal and local resources. *British Journal of Sports Medicine* **34**: 6

Schooler C, Farquhar J W, Fortmann S P, Flora J A 1997 Synthesis of findings and issues from community prevention trials. *Annals of Epidemiology* **S7**: S54–S68

Travers K D 1997 Reducing inequities through participatory research and community empowerment. *Health Education and Behavior* **24**: 344–56

J. W. Farquhar

Health Outcomes, Assessment of

Health is probably the most valued human asset. In fact, studies on the preference for different states of being show that virtually everyone rates health as most important. Despite the perceived importance of health, health status has remained difficult to define. There are two common themes in definitions of health. First, premature death is undesirable, so one aspect of health is the avoidance of mortality. The health status of nations is often evaluated in terms of overall mortality rates or infant mortality rates. Second, quality of life is also very important. Disease and disability are of concern because they affect either life expectancy or life quality. For example, cancer and heart disease are the two major causes of premature death in the United States. In addition, disease or disability can make life less desirable. A person with heart disease may face restrictions in daily living activities and may be unable to work or participate in social activities. Even relatively minor diseases and disabilities affect quality of life. A cold, for example, may interfere with the ability to concentrate, work, or attend school. The cold, however, lasts only a short time. A chronic disease, such as arthritis, may affect the quality of your life for a long time (Brown et al. 2000).

Within the last few years, medical scientists have come to realize the importance of measuring patient-reported health. Many major diseases, including arthritis, heart disease, and diabetes, or digestive problems are evaluated in terms of the degree to which they affect life quality and life expectancy (Koloski et al. 2000). One can also evaluate treatments for these conditions by the amount of improvement they produce in everyday life. The Food and Drug Administration now considers patient-reported outcomes in their evaluations of new products and nearly all major clinical trials in medicine use quality of life assessment measures. Several approaches to health outcome measurement will be reviewed. Chapter 130 by O'Boyle covers quality of life measures in more detail.

1. Historical Perspective

For many years, researchers and clinicians have struggled to reach consensus on the measurement of health and health outcomes. Different definitions emphasize related, but divergent, concepts. Some terms used to describe health are positive concepts, such as wellness or normality. Other definitions emphasize negative concepts, such as disability and illness. Much of the debate has centered on whether health is a continuum. Are disability and illness distinct from health, or are they opposite ends of the same continuum? (Patrick and Erickson 1993).

Because the concept of health lacked conceptual clarity, the World Health Organization (WHO) proposed a comprehensive definition of health in their charter document. Health was defined as, '… a state of complete physical, mental, and social well-being and not merely the absence of infirmity' (World Health Organization 1948). Since the introduction of this definition there has been some convergence of thought. It is now widely recognized that health has multiple dimensions. Further, it is now accepted that measures of mortality alone cannot summarize the health status of populations (Field and Gold 1998). However, there is still considerable debate about what constitutes health and health status.

Some of the debate about the definition of health originates from sociologists. From the sociological perspective, illness represents deviation from society's standards for physical and mental well-being. Often, this deviation is recognized when individuals cannot perform activities usual for their social roles. In addition to observable functioning, illness may cause reports of symptoms of pain and these deviate from society's norms on wellness. Parsons defined illness as '… a state of disturbance in normal functioning of the total human individual including both the state of the organism as a biological system, and/or his personal and social adjustments' (Parsons 1951, p. 431). For Parsons, health represented the capacity to perform valued tasks.

The traditional biomedical model is predicated on finding a specific biologic problem and repairing it. In contrast, measures of health outcome consider benefit from the patient's perspective. Successful treatments are those that improve the quality of life or extend life. Most surgical interventions are justified on the basis of saving lives. However, the great majority of surgical procedures have no effect on life expectancy. For

example, four of the most common indications for surgery in the United States are transurethral resection prostatectomy (TURP) for benign hyperplasia of the prostate (BPH), extractions of cataracts, joint replacement for severe osteoarthritis of the knee or hip, and hysterectomy for diseases of the uterus. Although, it is often argued that these procedures prevent deaths, reviews of the evidence suggest that these procedures are rarely performed to extend life expectancy. In each case, however, the surgery may lead to improvements in functioning or the reporting of symptoms. For example, TURP may reduce the symptom of urinary frequency in men. Cataract extraction may help older patients function better with reading, night-driving, and other activities of daily living. Joint arthroplasty may increase mobility for older patients. For most of these cases, traditional physiological measures are limited as outcome assessment. Patients seek these procedures because they want relief of symptoms. Outcomes assessment must consider whether the symptoms are, indeed, relieved. Ultimately, measures of health outcome can be used to evaluate the benefits of treatment from the patient's perspective.

2. *Assessment of Health Outcomes*

Methods of assessment of health-related quality of life represent at least two different conceptual traditions. One grows out of the tradition of health status measurement. Several efforts to develop measures of health status were launched in the late 1960s and early 1970s. All the projects were guided by the World Health Organization's (WHO) definition of health status. The projects resulted in a variety of assessment tools, including the Sickness Impact Profile (Bergner et al. 1981) the Quality of Well-Being Scale (Kaplan et al. 1996) the SF-36 (Ware and Gandek 1998) and the Nottingham Health Profile (Lowe et al. 1990). Many of the measures examined the effect of disease or disability on the performance of social roles, the ability to interact in the community, and physical functioning. Some of the systems have separate components for the measurement of physical, social, and mental health. The measures also differ in the extent to which they consider subjective aspects of life quality (Brown et al. 2000).

Perhaps the most important distinction between methods to assess quality of life is the contrast between psychometric and decision theory approaches. The psychometric approach attempts to provide separate measures for the many different dimensions of quality of life. Perhaps the best known example of the psychometric tradition is the Sickness Impact Profile (SIP). The SIP is a 136-item measure that yields 12 different scores displayed in a format similar to an MMPI profile (Bergner et al. 1981).

The decision theory approach attempts to weight the different dimensions of health in order to provide a single expression of health status. Supporters of this approach argue that psychometric methods fail to consider that different health problems are not of equal concern. A runny nose is not the same as severe chest pain. In an experimental trial using the psychometric approach, one will often find that some aspects of quality of life improve while others get worse. For example, a medication might reduce high blood pressure but also produce headaches and impotence. Many argue that the quality of life notion is the subjective evaluation of observable or objective health states. The decision theory approach attempts to provide an overall measure of quality of life that integrates subjective function states, preferences for these states, morbidity, and mortality.

3. *Common Methods for the Measurement of Quality of Life*

A variety of methods have been proposed to measure quality of life, but we cannot review and critique them all here. Instead, we present some of the most widely used psychometric and decision theory based methods. Readers interested in more detailed reviews should consult Walker and Rosser (1993) and McDowell & Newell (1996). Quality of life measures are also covered in Chapter 130 (by O' Boyle).

3.1 *Psychometric Methods*

(a) SF-36. Perhaps the most commonly used outcome measure in the world today is the Medical Outcome Study Short Form-36 (SF-36). The SF-36 grew out of work by the RAND Corporation and the Medical Outcomes Study (MOS) (Ware & Gandek 1998). Originally, it was based on the measurement strategy from the RAND Health Insurance Study. The MOS attempted to develop a very short, 20-item instrument known as the Short Form-20 or SF-20. However, the SF-20 did not have appropriate reliability for some dimensions. The SF-36 includes eight health concepts: physical functioning, role-physical, bodily pain, general health perceptions, vitality, social functioning, role-emotional, and mental health (Kosinski et al. 1999). The SF-36 can either be administered by a trained interviewer or self-administered. It has many advantages. For example, it is brief, and there is substantial evidence for its reliability and validity. The SF-36 can be machine scored and has been evaluated in large population studies. The reliability and validity of the SF-36 are well documented (Keller et al. 1999, Scott-Lennox et al. 1999, Stewart and Ware 1992, Ware and Gandek 1998).

Despite its many advantages, the SF-36 also presents some disadvantages. For example, it does not have age-specific questions and one cannot clearly

determine whether it is equally appropriate at each level of the age continuum. The items for older retired individuals are the same as those for children (Stewart and Ware 1992). Nevertheless, the SF-36 has become the most commonly used behavioral measure in contemporary medicine.

(b) Nottingham Health Profile. The Nottingham Health Profile (NHP), is another profile approach that has been used widely in Europe. One of the important features of the NHP is that the items were originally generated on the basis of extensive discussions with patients. The NHP has two parts. The first includes 38 items divided into six categories: sleep, physical mobility, energy, pain, emotional reactions, and social isolation. Items within each of these sections are rated in terms of relative importance. Items are rescaled in order to allow them to vary between 0 and 100 within each section. The second part of the NHP includes seven statements related to the areas of life most affected by health: employment, household activities, social life, home life, sex life, hobbies and interests, and holidays. The respondent indicates whether or not a health condition has affected his or her life in these areas. Used in a substantial number of studies, the NHP has considerable evidence for its reliability and validity.

An important strength of the NHP is that it is based on consumer definitions of health derived from individuals in the community. The language in the NHP is simple and the scale requires only a low level of reading ability. Psychometric properties of the NHP have been evaluated in a substantial number of studies. However, the NHP, like most profile measures, does not provide relative-importance weightings across dimensions. As a result, it is difficult to compare the dimensions directly with one another (Lowe et al. 1990).

3.2 Decision Theory Approaches

Within the last few years, interest has grown in using quality of life data to help evaluate the cost/utility or cost-effectiveness of healthcare programs. Cost studies have gained in popularity because healthcare costs have grown so rapidly in recent years. Not all healthcare interventions equally return benefit for the expended dollar. Objective cost studies might guide policymakers toward an optimal and equitable distribution of scarce resources. Cost-effectiveness analysis typically quantifies the benefits of a healthcare intervention in terms of years of life, or quality-adjusted life years (QALYs). Cost/utility is a special case of cost-effectiveness that weights observable health states by preferences or utility judgments of quality. In cost/utility analysis, the benefits of medical care, behavioral interventions, or preventive programs are expressed in terms of years of life adjusted for reduced quality of life or quality-adjusted life years (QALYs).

If a man dies of heart disease at the age of 50 and we expected him to live to age 75, we might conclude that the disease precipitated 25 lost life-years. If 100 men died at age 50 (and also had a life expectancy of 75 years), we might conclude that 2,500 (100 men $\times$ 25 years) life-years had been lost. Yet death is not the only relevant outcome of heart disease. Many adults suffer myocardial infarctions that leave them somewhat disabled for a long time. Though still alive, they suffer diminished quality of life. Quality-adjusted life-years take into consideration such consequences. For example, a disease that reduces quality of life by one-half will take away 0.5 QALY over the course of each year. If the disease affects two people, it will take away one year (2×0.5) over each year. A medical treatment that improves quality of life by 0.2 for each of five individuals will result in the equivalent of 1 QALY if the benefit persists for one year. This system has the advantage of considering both benefits and side effects of programs in terms of the common QALY units.

The need to integrate mortality and quality of life information is clearly apparent in studies of heart disease. Consider hypertension. People with high blood pressure may live shorter lives if untreated, longer if treated. Thus, one benefit of treatment is to add years to life. However, for most patients, high blood pressure does not produce symptoms for many years. Conversely, the treatment for high blood pressure may cause negative side effects. If one evaluates a treatment only in terms of changes in life expectancy, the benefits of the program will be overestimated because one has not taken side effects into consideration. On the other hand, considering only current quality of life will underestimate the treatment benefits, because information on mortality (death) is excluded. In fact, considering only current function might make the treatment look harmful because the side effects of the treatment might be worse than the symptoms of hypertension. A comprehensive measurement system takes into consideration side effects and benefits and provides an overall estimate of the benefit of treatment (Russell 1986).

Of the several different approaches for obtaining quality-adjusted life years, most are similar. The three most commonly used methods are the EQ-5D, the Health Utilities Index (HUI) and the Quality of Well-being Scale (QWB).

(a) EQ-5D. The approach most commonly used in the European community in the late twentieth century was the EQ-5D. This method has been developed by a collaborative group from Western Europe, led by Paul Kind and associates, known as the EuroQol group

(Kind 1997). The intention of this effort was to develop a generic currency for health that could be used commonly across Europe. The concept of a common EuroQol was stimulated by the desire for a common European currency—the Euro dollar. The original version of the EuroQol had 14 health states in six different domains. In addition, respondents placed their health on a continuum ranging from death (0.0) to perfect health (1.0). The method was validated in postal surveys in England, Sweden, and the Netherlands. More recent versions of the EuroQol, known as the EQ-5D, are now in use in a substantial number of clinical and population studies (Gudex et al. 1996, Hurst et al. 1997). Although the EQ-5D is easy to use and comprehensive, there have been some problems with ceiling effects. Substantial numbers of people obtain the highest possible score.

(b) Health Utilities Index. Another approach has been developed in Canada by Torrance et al. (1996). This method, known as the Health Utilities Index (HUI), is derived from micro-economic theory. There have been several versions of the measure, typically identified by 'Mark.' The HUI Mark I was developed for studies in the neonatal intensive care unit. The measure had 960 unique health states. In 1992, the HUI Mark II was developed and included 24,000 unique health states. The HUI Mark III, released in 1995, had 972,000 health states. Eight components of the HUI Mark III include vision (six levels), hearing (six levels), speech (five levels), ambulation (six levels), dexterity (six levels), emotion (five levels), cognition (six levels), and pain (five levels). Multiplying the number of levels across the eight dimensions gives the 972,000 states.

Using multiattribute utility scaling methods, judges evaluate levels of wellness associated with each level of each domain. A multiattribute model is used to map preference for the 972,000 possible states on to the 0.0 to 1.0 continuum. The HUI has been used in many population and clinical studies. For overall health status, men obtain higher scores early in the life cycle. However, after about the age of 45, women obtain higher scores and this difference grows systematically through the remainder of the life span (Kaplan and Erickson 2000).

(c) Quality of Well-being Scale (QWB). A third method, known as the Quality of Well-being Scale integrates several components into a single score. First, patients are classified according to objective levels of functioning. These levels are represented by the scales of mobility, physical activity, and social activity. Once observable behavioral levels of functioning have been classified, each individual is placed on the 0 to 1.0 scale of wellness, which describes where a person lies on the continuum between optimum function and death.

To accomplish this, the observable health states are weighted by quality ratings for the desirability of these conditions. Human value studies have been conducted to place the observable states onto a preference continuum, with an anchor of 0 for death and 1.0 for completely well. Studies have shown that the weights are highly stable over a one-year period and that they are consistent across diverse groups of raters. Finally, one must consider the duration of stay in various health states. Having a cough or a headache for one day is not the same as having the problem for one year. A health measure must take these durations into consideration. Using this information, one can describe health-related quality of life in terms similar to years of life. For example, one year in a state assigned the weight of 0.5 is equivalent to 0.5 of a quality-adjusted life-year.

3.3 Cost-effectiveness Evaluations

Utility-based measures of quality of life are often used to evaluate the cost-effectiveness of health-care programs (Kaplan et al. 1997). For example, one study using the QWB showed a new medication for patients with arthritis produced an average of 0.023 QALY per year, whereas a new medication for AIDS produced nearly 0.46 of these units per year. However, the benefit of the arthritis medication may last as long as 20 years, ultimately producing 0.023×20 years $= 0.46$ year. The AIDS treatment produced a benefit for only one year, so its total effect was 0.46×1 year $= 0.46$ year. In other words, the general system allows the full potential benefits of these two very different treatments to be compared (Kaplan et al. 1995).

QALYs are generic measures of life expectancy with adjustments for quality of life (Gold 1996). QALYs consider both benefits and side-effects of programs in terms of the common outcome units. Although QALYs are typically assessed for patients, they can also be measured for others, including care givers who are placed at risk because they experience stressful life events. The Institute of Medicine recommended that population health metrics be used to evaluate public programs and to assist the decision making process (Field and Gold 1998).

In addition to health benefits, programs also have costs. Resources are limited and good policy requires the allocation of resources to maximize life expectancy and health-related quality of life. Thus, in addition to measuring health outcomes, costs must also be considered. Methodologies for estimating costs have now become standardized (Gold 1996). From an administrative perspective, cost estimates include all costs of treatment and costs associated with caring for any side

effects of treatment. From a social perspective, costs are broader and may include the costs of family members staying off work to provide care. Comparing programs for a given population with a given medical condition, cost-effectiveness is measured as the change in costs of care for the program compared with the existing therapy or program, relative to the change in health measured in a standardized unit such as the QALY. The difference in costs over the difference in effectiveness is the *incremental cost-effectiveness* and is usually expressed as the cost/QALY. Since the objective of all programs is to produce QALYs, the cost/QALY ratio can be used to show the relative efficiency of different programs (Gold 1996).

Some traditional interventions, such as bypass surgery for one-vessel coronary heart disease, may cost as much as $700,000 to produce a QALY. Screening programs, such as mammography, may also require many resources to produce a QALY. On the other hand, public health programs such as pneumonia vaccines for the elderly or laws requiring children to be in infant seats and adults to be in seat belts in motor-cars may produce a QALY at a very low cost. It might be argued that programs to the left of the pay line should be funded, but those with cost/QALY ratios to the right of the line should be examined more carefully.

4. *Disease-specific Measures*

This chapter has focused on generic measures of health outcome. These measures can usually be applied to any age, gender, or ethnic group. In most cases age and gender-related norms are available. However, there are a significantly larger number of situation-specific measures. Some of these measures are appropriate for a particular age, gender or ethnic group. For example, several measures are designed specifically to assess health outcomes in children. Most often, the measures are disease-specific, meaning that they are designed to assess outcomes of a particular health problem. Validated measures of health outcome are available for illnesses such as arthritis, diabetes, heart disease, kidney failure, and virtually every other major health condition. Two examples are the Arthritis Impact Measurement Scale (Meenan 1982) and the UCSD Shortness of Breath Questionnaire used in studies of patients with emphysema (Eakin et al. 1998).

The Arthritis Impact Measurement Scales (AIMS) is a health index designed at the Multipurpose Arthritis Center at Boston University. It is intended to measure physical health and social well-being for patients with rheumatoid arthritis (Meenan 1982). The resultant scale includes 67 items, with questions about functioning, health perceptions, morbidity, and demographics. The AIMS contains scales for mobility, physical activity, social activity, activities of daily living, depression and anxiety and arthritis-related symptoms. In effect, it is an adaptation of an early version of the QWB (Eakin et al. 1998), with a series of items designed to tap more specifically the effect of arthritis upon functioning and the quality of life. Factor analysis of the AIMS has produced three subscales: physical function, psychological function, and pain. Most current applications of the AIMS use composite scores for these three areas.

The University of California, San Diego, Shortness of Breath Questionnaire (SOBQ) includes 25 items that evaluate self-reported shortness of breath during the performance of various activities of daily living. Evaluations of the measure show it highly correlated with other quality of life measures such as the Quality of Well-Being Scale and the Center for Epidemiologic Studies Depression Scale. The measure has high internal consistency (alpha = 0.96) and is significantly correlated with performance measures such as the distance that can be walked in six minutes (Eakin et al. 1998).

5. *Summary*

The assessment of patient-reported experience has become a standard component of health outcome evaluation. There are several important distinctions between commonly-used approaches. Generic measures are used to evaluate the health outcome for any illness or disease of state. Generic measures can typically be classified as derived from psychometric or decision theory. Psychometric approaches include the Sickness Impact Profile, the SF-36, and the Nottingham Health Profile. Decision theory approaches are used to estimate outcomes in terms of quality-adjusted life years (QALYs). Methods used for this purpose include the EQ-5D, the HUI, and the QWB. Disease-specific measures are available for a wide variety of health conditions. Although disease-specific measures may be more sensitive for outcomes of a particular condition, they cannot be used for cross-illness comparisons or for cost-effectiveness analysis. The assessment of health outcome is a rapidly-developing field and we anticipate major new developments over the next decade.

See also: Chronic Illness: Quality of Life; Quality of Life: Assessment in Health Settings; Well-being (Subjective), Psychology of

Bibliography

Bergner M, Bobbitt R A, Carter W B, Gilson B S 1981 The sickness impact profile: Development and final revision of a health status measure. *Medical Care* **19**(8): 787–805

Brown M, Gordon W A, Haddad L 2000 Models for predicting subjective quality of life in individuals with traumatic brain injury. *Brain Injury* **14**(1): 5–19

Eakin E G, Resnikoff P M, Prewitt L M, Ries A L, Kaplan R M 1998 Validation of a new dyspnea measure: The UCSD Shortness of Breath Questionnaire. University of California, San Diego. *Chest* **113**(3): 619–24

Feeny D, Furlong W, Mulhern R K, Barr R D, Hudson M 1999 A framework for assessing health-related quality of life among children with cancer. *International Journal of Cancer Supplement* **12**(5): 2–9

Field M J, Gold M R 1998 *Summarizing Population Health*. Institute of Medicine. National Academy Press, Washington, DC

Gold M R 1996 *Cost-effectiveness in Health and Medicine*. Oxford University Press, New York

Gudex C, Dolan P, Kind P, Williams A 1996 Health state valuations from the general public using the visual analogue scale. *Quality of Life Research* **5**(6): 521–31

Hurst N P, Kind P, Ruta D, Hunter M, Stubbings A 1997 Measuring health-related quality of life in rheumatoid arthritis: Validity, responsiveness and reliability of EuroQol EQ-5D. *British Journal of Rheumatology* **36**(5): 551–9

Kaplan R M, Anderson J P, Patterson T L, McCutchan J A, Weinrich J D, Heaton R K, Atkinson J H, Thal L, Chandler J, Grant I 1995 Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection. HNRC Group HIV Neurobehavioral Research Center. *Psychosomatic Medicine* **57**(2): 138–47

Kaplan R M, Erickson P 2000 Gender differences in quality-adjusted survival using a Health-Utilities Index. *American Journal of Preventive Medicine* **18**(1): 77–82

Kaplan R M, Sieber W J, Ganiats T G 1997 The Quality of Well-being Scale: Comparison of the interviewer-administered version with a self-administered questionnaire. *Psychology & Health* **12**(6)

Keller S D, Ware J E Jr, Hatoum H T, Kong S X 1999 The SF-36 Arthritis-Specific Health Index ASHI: II Tests of validity in four clinical trials. *Medical Care* **5**(Suppl MS): 51–60

Kind P 1997 The performance characteristics of EQ-5D, a measure of health related quality of life for use in technology assessment [abstract]. *Annual Meeting of International Society of Technology Assessment in Health Care* **13**(5): 81

Koloski N A, Talley N J, Boyce P M 2000 The impact of functional gastrointestinal disorders on quality of life. *American Journal of Gastroenterology* **95**(1): 67–71

Kosinski M, Keller S D, Ware J E Jr, Hatoum H T, Kong S X 1999 The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: Relative validity of scales in relation to clinical measures of arthritis severity. *Medical Care* **37**(5 Suppl MS): 23–39

Lowe D, O'Grady J G, McEwen J, Williams R 1990 Quality of life following liver transplantation: A preliminary report. *Journal of the Royal College of Physicians of London* **24**(1): 43–6

McDowell I, Newell C 1996 *Measuring Health: A Guide to Rating Scales and Questionnaires*, 2nd edn. Oxford University Press, New York

Meenan R F 1982 The AIMS approach to health status measurement: Conceptual background and measurement properties. *Journal of Rheumatology* **9**(5): 785–8

Parsons T 1951 *The Social System*. Free Press, Glencoe, IL

Patrick D L, Erickson P 1993 *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*. Oxford University Press, Boston, MA

Russell L B 1986 *Is Prevention Better Than Cure?* Brookings Institution, Washington, DC

Scott-Lennox J A, Wu A W, Boyer J G, Ware J E Jr 1999 Reliability and validity of French, German, Italian, Dutch, and UK English translations of the Medical Outcomes Study HIV Health Survey. *Medical Care* **37**(9): 908–25

Stewart A L, Ware J E 1992 *Measuring Functioning and Well-being: The Medical Outcomes Study Approach*. Duke University Press, Durham

Torrance G W, Feeny D H, Furlong W J, Barr R D, Zhang Y, Wang Q 1996 Multiattribute utility function for a comprehensive health status classification system. Health utilities index mark 2. *Medical Care* **34**(7): 702–22

Walker S R, Rosser R 1993 *Quality of Life Assessment: Key Issues in the 1990s*. Kluwer Academic Publishers, Dordrecht

Ware J E Jr, Gandek B 1998 Overview of the SF-36 Health Survey and the International Quality of Life Assessment IQOLA Project. *Journal of Clinical Epidemiology* **51**(11): 903–12

R. M. Kaplan

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Health Policy

In Western industrialized nations, the term health policy has come to mean government policies that subsidize and regulate the *economic* relations between patients, insurers, and health care providers—doctors, hospitals, pharmacies, pharmaceutical companies. Within this general framework—an economic model of medicine—the role of government in health care varies considerably from country to country and has been the subject of considerable and continuing political contestation. In order to understand these differences, one must return to the political circumstances that led to the enactment—or rejection—of specific health programs.

1. *Types of Health Policy*

The role of government in health care may be classified according to methods of financing health care provision. Governments may subsidize medical treatment by providing for tax exemptions for medical expenditures. They may provide government subsidies to private insurers. They may provide health insurance coverage by creating government insurance programs. Or they may provide health services through government-employed doctors and government-owned hospitals. Because each type of financing has implications for the role of government in regulating the medical market, government financial provisions in the health area can be ranked from less to more

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

invasive: tax expenditures are considered the least interventionist form of health policy; government health services are considered the most interventionist.

1.1 Government Subsidies

Tax deductions for medical expenditures were generally introduced after the Second World War, and were often the result of political conflicts preventing a more ambitious role for government in health, such as the failure of the Truman national health insurance proposal of 1948 in the United States. A policy of tax expenditure impacts the medical market by reducing the costs of health care to the consumer. This financial incentive allows governments to promote health policy goals, for example, by deciding which types of medical treatment will be deductible, or which types of insurance benefits qualify as tax exempt. Tax deductibility has been an important element in the spread of health insurance coverage as part of collective bargaining agreements since the 1940s, although the current trend is towards reduction rather than expansion of insurance benefits provided by employers.

A second form of government financing for health care is the payment of *subsidies* to insurance carriers. This policy dates from the mid-nineteenth century and was a reaction to the growth of voluntary mutual aid societies (also known as friendly societies or sickness funds) that provided their members with death and sickness benefits (cash payments to cover the costs of burial and income loss during illness, respectively). Typically, governments provided the mutual societies with a fixed subsidy per member. In return, the societies were required to follow specific financial practices—for example to build up larger financial reserves—and to provide medical benefits in addition to cash payments. In addition, political authorities used financial resources as a means of political control, monitoring the activities of the societies, and sometimes intervening in the selection of the members of the societies' governing boards, as was the case in nineteenth century France, Italy, and Belgium.

1.2 National Health Insurance

Under national health insurance programs, governments provide *compulsory* health insurance to all or parts of the population. Historically, such programs were first introduced to provide insurance to industrial workers at the end of the nineteenth century, with Bismarck's social insurance laws of 1883 as the first example (see *Social Security*; *Welfare State*, *History of Social Insurance*: *Legal Aspects*). Over the course of the twentieth century, compulsory health insurance programs have been expanded to cover ever greater portions of national populations—hence the term 'national.' National health insurance programs are in place in most of Europe, Latin America, Canada, Australia, and Japan. When the entire population is covered as a right of citizenship, these programs are said to be universal. Nevertheless, while the term 'national health insurance' connotes the compulsory provision of health insurance to broad sectors of the population, there is no generally accepted definition for the portion of the population to be covered or for the exact benefits to be provided. These are matters of political negotiation and political choice.

National health insurance programs provide governments with greater incentives and instruments for regulating the medical market than programs of government subsidies. Government decisions determine who is to be publicly insured—often defined by income or age limits, or by specifying occupational groups, such as farmers, salaried employees, civil servants, or the unemployed, that are to be compulsorily insured. These decisions, in turn, set limits to the market for private insurance. National health insurance programs may be directly administered by government agencies, or by independent insurance carriers, often the voluntary mutual aid societies or nonprofit group plans that existed prior to the government program. In either case, however, governments regulate benefits, conditions of membership, portability of benefits, and contribution rates. Rather than paying a premium for their insurance coverage, citizens are said to pay a contribution, one which is often collected in the form of a payroll tax, as part of the nation's social insurance or social security system. Employers are generally required to pay for half or more of the insurance contribution for their employees. Consequently, many public insurance programs provide for some representation of these contributors on the boards of the insurance carriers, a practice that has been termed 'self-administration'.

The regulation engendered by national health insurance is not limited to citizens and insurers. Political decisions—often the result of negotiations between representatives of providers and insurance authorities—determine the specific treatments and medications that will be covered by a particular national health insurance program, as well as the price that may be charged. Governments may simply set maximum fees to be charged to national health insurance patients, or they may negotiate fees with representatives of medical or hospital associations. These lists of negotiated fees are known as fee schedules. Hospitals or doctors may be required to accept the national health insurance fee as payment in full, and thus prohibited from adding surcharges (balance-billing or extra payments).

Because government price controls may be circumvented by increasing the volume of services, newer methods of cost containment have been developed. Case-based payment or diagnostic-related groups remunerate hospitals a fixed sum per patient with a particular diagnosis rather than reimbursing the full

costs of treatment for a specific individual. In this way, providers are given an incentive to reduce the volume (and hence the costs) of treatment. The global budget bases hospital payments on a larger entity. Either a department or even the entire hospital will be given a yearly (or quarterly or semi-annual) budget, within which the hospital managers may allocate funds at their discretion. In the area of pharmaceuticals, common cost-containment measures include the development of positive (allowed) and negative (prohibited) lists of prescription drugs, as well as efforts to negotiate directly with manufacturers to obtain discounts on pharmaceutical and other medical products. In order to reduce the volume of prescriptions, some national health insurance authorities have introduced economic monitoring of physicians (reviews whereby physician behavior is compared to statistical averages) and prescription budgets for physicians. Further, many national health insurance programs rely on cost-sharing on the part of patients to reduce the demand for health services.

1.3 National Health Services

In a national health service, governments provide health services directly. Citizens are guaranteed a range of health care benefits, which are generally provided at no or at low cost at the time of treatment. Medical benefits are typically financed by tax revenues (although there may be a health insurance component as well) and provided under government auspices. Just what this 'public' provision is to entail, however, is highly controversial. In the British National Health Service, founded in 1948, and in Sweden, a country that can be considered to have a *de facto* national health service, the overwhelming majority of hospitals are government owned, and hospital doctors are paid a full-time salary. In Southern Europe (Italy, Portugal, and Spain), the more recently created national health services have incorporated privately-owned hospitals, such as those owned by churches and social security institutions, into the public service on a contractual basis. Within national health service hospitals, patients may or may not be allowed to pay additional monies for treatment in a 'private' bed or private room. Similarly, hospital doctors (especially chiefs of staff) may or may not be allowed to treat 'private' patients for additional fees.

The outpatient or ambulatory sector has traditionally been a battleground for the demarcation of the public–private boundary. Local health clinics or hospital outpatient departments compete with private office practice by providing health treatments at virtually no charge through doctors that are paid on a salaried basis. Even in private offices, doctors in national health services are often under contract to the government, and paid according to fee schedules or on a capitation basis (a set fee per patient per year). Efforts to pay private office practitioners a yearly salary for their national health service work have generally failed in market economies. In state socialist countries, on the other hand (China, Cuba, the former German Democratic Republic, and the former Soviet Union, for example), health care was long provided through government hospitals and outpatient clinics (polyclinics or Ambulatorien) that employed doctors as salaried civil servants.

In comparison to systems of government subsidies and national health insurance—both of which increase the demand for health services—national health services provide governments with relatively effective tools of cost-containment and health planning. Particularly in a system such as the British, where health expenditures are financed by the national government and the national health budget constitutes a hard constraint on health spending, less regulation of providers has been necessary to hold expenditures in check. Moreover, centralized public authority for investment decisions has facilitated national planning for medical technology and the regionalization of the hospital system. These policies limit and guide the distribution of expensive technologies, such as nuclear-magnetic-resonance-imaging devices, and create a division of labor among hospitals, with more highly-specialized personnel and equipment concentrated at regional centers (see *Health Care Technology*). As expenditures are curtailed, however, waiting lists and complaints about the quality of services may arise.

1.4 Health Policy Issues

From this description, it is clear that the relationship between state and market has been a central issue of health policy. But, while this division dovetails neatly with political ideologies—and hence with political discourse—it is too simplistic for describing the complex mixture of private and public provision and financing of health care in place today. Even in relatively 'market-oriented' systems, government financing and regulation of health care is extensive, including funds for hospital construction and research, as well as a myriad of regulations governing health insurance, hospitals, and medical equipment. Furthermore, health care decision-makers—whether public or private—are turning increasingly to quite similar policy instruments. Private insurance companies, hospital corporations, and planners interested in regulating the private market promote managed care and managed competition that involves the same budgetary controls, fee schedules, and salaried employment that have been developed to contain costs under national health insurance and in national health services. On the other hand, government health systems now strive to use market incentives and increasing

competition to improve the quality, responsiveness, and efficiency of public systems in programs of public competition and internal markets. Consequently, terms like 'socialized medicine' and 'medical rationing' are better considered as political slogans than as accurate descriptions of health policy. The public component of medical care is very large in all countries—and has been so historically—and efforts to influence the distribution of health care services are prevalent throughout both the public and the private sectors. Instead, the politically more neutral monopsony (or existence of a single buyer) may better capture differences in health systems. To the extent that purchasing power becomes concentrated, the buyers of health care can exert more leverage over the providers. This tendency is at the root of conflicts between providers, such as doctors, and the organized buyers of medical services—private insurance companies, voluntary mutual aid societies, hospitals, and governments.

In sum, health policy entails government's pursuit of three goals through both public and private means: the provision of citizens with access to health care; the assurance of the quality of health care; and the control of the costs of health care. Because it is virtually impossible to meet these three goals simultaneously, health policy will continue to be contested. (See also: *Health Care Systems*: *Comparative*; *Health Care Delivery Services*; *Health Care Markets*: *Theory and Practice*; *Health Economics*; *Managed Care*; *National Health Care and Insurance Systems*; *Health Care*, *Rationing of*.)

2. *Determinants of Health Policy*

2.1 *Public Health Provision*

Conflicts over the role of the state in medicine go back as far as the Reformation. At that time, responsibility for health care was divided among an assortment of authorities with overlapping and contradictory jurisdictions. Medical services were delivered by a diversity of practitioners, with widely varying degrees of training, and through a range of institutions, including the Catholic church. In their quest for external and internal sovereignty (see also *Absolutism*, *History of* and *Policy History*: *Origins*), monarchs such as Henry VIII of England or Gustav Wasa of Sweden appropriated hospices along with other Church lands, and began to offer health services under government auspices. Particularly in countries with an absolutist tradition, the rise of the nation state (see *Nations and Nation-states in History*; *State Formation*; *State and Society*; *State*, *Sociology of the*; *State*, *History of*) was associated with an expansion of military medicine, and the establishment of public health officers as well as government bureaucracies for tracking contagious diseases. In Russia, the state established public health clinics called *Zemstvos*; in many other countries, public health centers were blocked by medical opposition. By the end of the nineteenth century, governments throughout the West had acknowledged some role for the state in the provision of health services, but its scope varied widely.

2.2 *The Medical Profession*

Political authorities also interceded in the relationships between churches, universities, and medical corporations by giving government approval to specific institutions of medical training and licensing, and, in some cases, by outlawing alternative practitioners. The Royal College of Physicians of London, for example, was granted a charter in 1518 to license physicians. This charter did not interfere, however, with the rights of Cambridge and Oxford universities or ecclesiastical authorities. As in many European countries, three strata of practitioners co-existed: the academically-trained physicians; the more practically-trained surgeons—who had originally been associated with barbers, and whose development was linked to military medicine, like the *Feldshers*, *Wundärzte*, and *Fältskäraren* on the Continent; and the apothecaries—who, expanding their trade to prescribing medicines, had split off from the grocer's guild in 1617. The medical profession was not unified until the Medical Act of 1858, which provided for a General Council of Medical Education and Registration, composed of representatives of the medical corporations and the universities, that was to supervise medical education and register holders of medical licenses. All previous licensing bodies could continue to grant degrees, but there were no restrictions on who could practice medicine other than penalties for improper use of a medical title.

The process by which specialized routes of training and barriers to entry were established has been called professionalization (see *Medical Sociology*; *Professions*, *Sociology of*; *Professionalization/Professions in History*). In some countries, such as Spain and Sweden, professional unification came earlier than in Britain, and professional monopoly was 'tighter,' as unlicensed practice was actually outlawed. In many cases, liberal ideology—with its stress on free markets, individual freedom, and a minimal state—was a barrier to professionalization. In the United States and Switzerland, liberal and democratic ideals precluded medical licensing, limits on medical school admissions, and even prohibitions on medical quackery until well into the twentieth century. In Prussia, the liberal Berlin Medical Society opposed the introduction of mandatory licensing in 1869, which led to a system of state certification for doctors with no penalties for unlicensed practice when Germany was

unified. In France, the revolution (which firmly placed hospitals in the hands of the state) led to new laws that annulled the *ancien régime* requirement of a university degree for the practice of medicine, outlawed medical associations and abolished the university medical faculties, as well as a host of special rights and privileges granted to various subgroups of the profession. Only at the end of the nineteenth century was medical monopoly firmly re-established in France and placed under the jurisdiction of the state universities, without a role for professional bodies. Perhaps for this reason, professional organization in France has been weak, although medical monopoly is strong (see also *Western European Studies*: *Health*).

2.3 Medical Associations

One can distinguish between several different types of medical organization. Medical corporations, like the British Royal Colleges or the German *Bundesärztekammer*, are professional bodies with legal jurisdiction over medical licensing and professional ethics. Medical societies are associations formed for the purposes of disseminating medical knowledge, and have often founded medical journals. Despite their efforts to maintain some distance from economic and political issues, they have also promoted the interests of their members. Medical associations and medical unions are most directly involved with economic negotiations and lobby efforts, having been founded explicitly as interest associations. In addition, some nations require doctors participating in national health insurance schemes to belong to associations that represent doctors in bargaining with health authorities, such as the *Kassenärztliche Bundesvereinigung* in Germany. Separate organizations often exist for hospital doctors and for various medical specialties. Medical licensing is now only rarely left to medical corporations, but specialty licensing is often left to specialty societies, which may explain their high membership rates and influence.

Professionalization and the development of medical organizations both spurred and was spurred by the expansion of the market for medical care. By defining medicine as a distinct body of knowledge—in order to gain public legitimacy for professional privilege, on the view of some interpretations—medical professions promoted the demand for their services. As governments acknowledged these claims to expertise, they moved to provide medical services to the public, turning increasingly to programs of medical insurance to supplement direct public provision. To be sure, these policies of state medical care were not motivated purely by health policy concerns—the fear of working-class unrest and the need for healthy soldiers were just as important. Indeed, the voluntary mutual societies themselves were not formed solely to provide sickness and death benefits, but instead served as a way to form associations at a time when many working-class organizations were prohibited by various anticombination acts. In any case, the provision of medical benefits through state-subsidized mutual associations and compulsory health insurance plans further stimulated the formation of medical associations, as doctors sought to prevent the development of disadvantageous working conditions.

2.4 Professional Autonomy

The increasing prevalence of collective payment for medical services stimulated normative and theoretical discussions about the relationship between public service and professional freedom (see for example the debate between Harold Laski and T. H. Marshall described in Marshall 1964 [1939]). Freidson's seminal work defines the key feature of professional freedom as 'technical autonomy.' Occupations that have been able to convince the public that their practitioners are privy to indispensable knowledge gain a privileged position, defined by the exclusive competence to judge professional work. Whereas Freidson specifically states that economic autonomy ('control over the socioeconomic terms of work,' 1970, p. 25) has nothing to do with professional freedom, this has been a central issue in health policy. Medical associations across Western Europe and North America fought for economic autonomy, arguing that economic freedom such as the right to set fees and receive private patients was a precondition of therapeutic freedom (the physician's ability to choose freely among medical therapies) and of the right of patients to a 'free choice of doctor.' As private medical practice has become subject to burdensome government regulation and corporate management controls, however, some argue that physicians are better off in national health systems, where more generalized budgetary controls obviate the need for such detailed micromanagement, preserving a greater degree of clinical autonomy (the ability of physicians to make individual medical decisions).

2.5 The Politics of Health

National health policies are thus the result of centuries of political conflict over the proper role of government in health. Both governmental decisions and those made in private negotiations have influenced the complex set of institutions and regulatory framework that governs contemporary health care financing and provision. In order to make sense of these patterns, scholars have turned to some of the same sociological and political theories that have been applied to the politics of the welfare state (see *Social Welfare Policy*:

Comparisons; *Welfare State*, *History of*). The development of different national health policies has been understood in terms of differences in: leading political ideologies and political cultures; social structure and class relations; state traditions and state structures; left and right political representation; national political institutions. A particular focus of the study of health politics has been the issue of professional dominance, that is, whether the medical profession as the sole experts qualified to deliver medical services can parlay their professional expertise into sufficient political power to dominate health policy decisions.

2.6 Future Challenges

Because of its historical development, health policy has been focused very much on the role of nation-states in facilitating and regulating medical services that are provided by national medical professions. This traditional definition now faces two central challenges: the internationalization of the economy, and the revolution in molecular biology. Internationalization poses a challenge to national institutions and programs of health care. Not only the free movement of persons and services across borders, but the tremendous international differences in access to medical treatment, highlighted by crises such as the Aids pandemic, cannot be easily addressed through national health policies. Nor is the Western-industrial approach, which is focused on individual medical services, necessarily appropriate for the problems of the developing world.

Even more profound, perhaps, are the implications of the cracking of the genetic code. Discussions about whether the goal of health policy should be to provide access to medical services or whether investments in public health—hygiene, vaccines, and nutrition—might not have more impact on the health states of populations have been perennial, to be sure. Nevertheless, the discovery of the genetic basis for disease poses a new threat both to the position of doctors as the holders of medical expertise and the role of curative medicine. Furthermore, the potential for genetic engineering, and even the development of more effective pharmaceuticals through recombinant DNA methods, entails tremendous ethical and cost problems.

See also: Health Care Delivery Services; Health Care: Legal Issues; Health Care Organizations: For-profit and Nonprofit; Health Economics; Health in Developing Countries: Cultural Concerns; Health Insurance: Economic and Risk Aspects; Health Outcomes, Assessment of; Medical Sociology; National Health Care and Insurance Systems; Professions, Sociology of; Welfare State

Bibliography

Abelsmith B 1985 Who is the odd man out? The experience of Western Europe in containing the costs of health care. *Milbank Memorial Quarterly/Health and Society* **63**: 1–17

Anderson O W 1972 *Health Care: Can There be Equity? The United States, Sweden and England*. John Wiley, New York

Björkman J W, Altenstetter C 1998 *Health Policy*, The International Library of Comparative Public Policy 7. Edward Elgar, Cheltenham, UK

Freidson E 1970 *Profession of Medicine: A Study of the Sociology of Applied Knowledge*. Dodd, Mead & Co., New York

Glaser W A 1991 *Health Insurance in Practice: International Variations in Financing, Benefits and Problems*. Jossey-Bass, San Francisco

Hollingsworth J R, Hage J, Hanneman R A 1990 *State Intervention in Medical Care: Consequences for Britain, France, Sweden, and the United States, 1890–1970*. Cornell University Press, Ithaca, NY

Klein R 1989 *The Politics of the National Health Service*, 2nd edn. Longman, London

McLachlan G, Maynard A (eds.) 1982 *The Public/Private Mix for Health: The Relevance and Effects of Change*. The Nuffield Provincial Hospitals Trust, London

Marmor T R 1994 *Understanding Health Care Reform*. Yale University Press, New Haven, CT

Marshall T H 1964 [1939] The recent history of professionalism in relation to social structure and social policy. In: Marshall T H (ed.) *Class, Citizenship and Social Development: Essays*. Doubleday, Garden City, NY

Maxwell R J 1981 *Health and Wealth: An International Study of Health Care Spending*. Lexington Books, Lexington, MA

Organisation for Economic Co-operation and Development (OECD) 1994 *The Reform of Health Care Systems: A Review of Seventeen OECD Countries*. OECD, Paris

Powell F D, Wessen A F 1999 *Health Care Systems in Transition: An International Perspective*. Sage, Thousand Oaks, CA

Raffel M W (ed.) 1997 *Health Care and Reform in Industrialized Countries*. Pennsylvania State University Press, University Park, PA

Rimlinger G V 1971 *Welfare Policy and Industrialization in Europe, America and Russia*. John Wiley, New York

Rodwin V 1984 *The Health Planning Predicament: France, Québec, England and the United States*. University of California Press, Berkeley, CA

Schieber G J, Poullier J-P 1991 International health spending: Issues and trends. *Health Affairs* **10**(1): 106–16

Schneider M, Biene-Dietrich P, Gabanyi M, Huber M, Köse A, Scholtes L, Sommer J H 1993 *Gesundheitssysteme im internationalen Vergleich*, Ausgabe 1992. Beratungsgesellschaft für angewandte Systemforschung (BASYS), Augsburg

Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York

Stone D A 1980 *The Limits of Professional Power*. University of Chicago Press, Chicago

White J 1995 *Competing Solutions: American Health Care Proposals and International Experience*. The Brookings Institution Press, Washington, DC

Wilkerson J D, Devers K J, Given R S 1997 *Competitive Managed Care: The Emerging Health Care System*. Jossey-Bass, San Francisco

E. M. Immergut

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Health Professionals, Allied

Health care services experienced remarkable growth in the latter half of the twentieth century in industrialized countries. Advances in medical science and technology, the expansion of health insurance, the aging of the population, increased public expectations and disposable income, and expanding social definitions of the boundaries of medical care all contributed to increased utilization of health services and changing patterns of care. These changes fostered the development of an expanded division of medical labor in which allied health professionals emerged to complement and compete with physicians in the provision of a broader array of health services. Globalization of health care has fostered renewed interest in strengthening the roles of allied health providers in developing countries and in the post-Soviet states in order to increase access to primary medical care and improve population health.

1. *The Changing Context of Health Care*

Health and medical services have undergone dramatic change since the mid-twentieth century. Medical care has become more specialized while also expanding in scope. At the beginning of the twentieth century, physicians accounted for about one in three healthcare workers; by the 1980s the comparable ratio was about 1 in 16 (Mechanic 1983). Health care occupations are highly sex-segregated with most nonphysician occupations dominated by women. Indeed, in many parts of the world physicians are also predominantly women reflecting societal sex role links between women and caring/health functions.

1.1 *Changing Disease Patterns*

Advances in medical science including antibiotics and vaccines as well as improved standards of living contributed to shifting the burden of illness from acute, infectious illnesses to chronic conditions. While advances in modern medicine have produced important treatments for chronic illnesses, most have resulted in 'halfway' technologies that extend life but do not cure the underlying disease (Thomas 1977). Examples include long-term renal dialysis for the treatment of kidney failure and, more recently, antiretroviral drugs that extend the lives of people with acquired immunodeficiency syndrome (AIDS). Chronic conditions are accompanied by a myriad of functional limitations and social support needs that fall outside the conventional definition of medical practice. Nurses, social workers, occupational therapists, physical therapists, and other nonphysician health professionals have demonstrated interest and expertise in helping the chronically ill and frail elderly manage the sequella of chronic conditions and medical treatments, thus creating new roles for themselves in the medical division of labor.

1.2 *Medicalization of Problems of Everyday Living*

The successes of modern medicine in treating life-threatening illnesses have resulted over time in the redefinition of a range of health-related issues and problems of everyday life as medical concerns. The boundaries of medicine have expanded to encompass health promotion, diet and exercise, child rearing, school performance, occupational health, and a wide array of personal and social problems. Medicalization of problems of everyday living lend legitimacy to social interventions (such as behavior control of disruptive school children), allow those seeking help to adopt a socially sanctioned sick role (Parsons 1951), and provide access to health insurance funds for financing treatment. Physicians are not well suited by interests or training to providing quasimedical social services, thus opening up new roles for allied health professions including social workers, psychologists, and chiropractors, as well as a myriad of new occupations such as personal exercise trainers and weight loss consultants.

Indeed, physicians have continued to pursue more specialized medical training and practice as help-seeking patterns in the general population have redefined the boundaries of medical care to encompass a much broader array of health concerns. Thus, in addition to the expanded opportunities afforded the allied health professions, alternative and complementary health practitioners such as naturopaths, acupuncturists, and herbalists have grown in number and availability. In 1997, American consumers paid an estimated $27 billion for unconventional therapies, which is roughly comparable to out-of-pocket expenditures for physician services (Eisenberg et al. 1998). Increasingly, conventional physician-dominated settings, including prestigious teaching hospitals, are establishing alternative and complementary health clinics employing nonphysician providers.

1.3 *Organizational Changes Conducive to Expansion of Nonphysician Professions*

Professions routinely aspire for full control over well-documented jurisdictions in which workplace authority is secured formally through legislation or courts of law and validated by public opinion (Abbott 1988). Physicians have successfully defended their dominance over the medical division of labor through legislative

and legal actions, controlling referral patterns, and by limiting direct access to health services payments (Sage and Aiken 1997). Yet, nonphysician health professions are constantly seeking to define themselves, grow, survive, and contest with medicine within an organizational environment that is continually subject to change due to exogenous social and technological trends.

The organizational context of medical practice changed substantially in the latter third of the twentieth century. The corporatization and consolidation of health services resulted in loss of autonomy for physicians (Starr 1982), but new opportunities and a more favorable environment for the allied health professions. Managerial priorities on cost containment gave allied health professions their first opportunity to compete with physicians on the basis of price and client satisfaction, and they have been successful on both.

1.4 Rapid Growth of Nonphysician Clinicians

The changing nature of health care and its organization has resulted in the rapid growth of allied health professions even in a context of a rapidly increasing physician supply. An analysis of growth of nonphysician clinicians in the USA revealed that the aggregate number of nonphysician clinicians in 10 disciplines (nurse practitioners, physician assistants, nurse-midwives, chiropractors, acupuncturists, naturopaths, optometrists, podiatrists, nurse anesthetists, and clinical nurse specialists) doubled between 1992 and 1997, and was expected to increase by two-thirds more by 2005 (Cooper et al. 1998). These 10 occupations were selected for study because they most overlap the scope of medical and surgical services provided by physicians. The growth of allied health professions is not unique to the USA, and evidence of expansion of nonphysician clinicians can be found in many industrialized countries and also in the developing world as health care becomes an increasingly global enterprise.

1.5 Globalization of Health Services

Global trade and business interests have increasingly impacted upon health services and health care personnel through cross-national commercial ventures, trade agreements affecting the movement of workers across national boundaries, and international aid activities. The North American Free Trade Agreement (NAFTA), linking Mexico, the USA, and Canada, and the European Union (EU) created the impetus to begin reconciling the training and practice of health professions across countries (Freeman et al. 1995). Health services is a major vehicle for transforming economic markets in the now independent states of the former Soviet Union, and many Western concepts of organization and financing of health services, and education of medical personnel are being introduced (Mossialos and McKee 1998/99). The growing threat of new emerging infectious diseases in parts of the developing world has attracted international attention and focus on strengthening public health and primary care in poor countries (Institute for the Future 2000). Since the health workforce is the infrastructure upon which health services initiatives are built, global initiatives inevitably impact on the education and composition of the medical workforce.

2. Nursing

Nursing is the largest of the health professions. Nursing historically has been closely tied to the profession of medicine, and the boundaries between the two have blurred over time in interesting ways. Nurses now participate in activities that in the past were the sole domain of physicians, such as physical examinations, diagnosis, treating illness, and prescribing drugs. The obverse is also true. Physicians have claimed jurisdiction over some types of services provided routinely in the past by nurses, as in the case of attending at normal births (Fagin 1990), and provision of anesthesia for surgery which only since the 1980s has become a physician specialty (Bankert 1989).

Expanded roles for nurses is a global phenomenon. The USA has led in developing expanded roles for nurses in care exploiting high technology (Fairman and Lynaugh 1998) and in independent practice roles in ambulatory care with the development of advanced practice nurses (Mezey and McGivern 1999). Other countries are now following similar patterns. British nurses have been granted wider authority by the National Health Service to prescribe drugs (Norman 2000). Canada is training nurse practitioners based on the US model of advanced practice nursing. Role expansion for nurses is occurring in a context of sufficient or excess supply of physicians, suggesting that nurses are not substituting for physicians as much as they are responding to a broader definition of health and medical care and market forces that provide new opportunities for role expansion for nonphysicians.

2.1 Historical Perspectives

Modern nursing dates back to mid-nineteenth century England when hospitals were disorganized, unsanitary, staffed by unskilled and undisciplined workers, and were generally held in low esteem. Only the destitute sought help in hospitals, except in times of war. Florence Nightingale, a wealthy and educated English woman, accepted the challenge of bringing order to military hospitals during the Crimean War and in doing so began a social experiment that would change the nature of hospitals around the world. Nightingale identified trained nurses as the vehicle for transforming hospitals but she recognized that nurses capable of skilled work and responsible ward man-

agement would have to be recruited from the better-educated classes. To legitimize the work of nursing to more highly educated women, she drew on the well-respected traditions of religious education and military-style authority. Young nurses motivated by a sense of dedication and service to mankind worked long hours for few economic rewards under strict discipline with total obedience expected (Abel-Smith 1960).

Nightingale combined her vision of order and management led by trained nurses, with the opportunity created by the Crimean War and her social position and statistical prowess to forge a new future for hospitals. Florence Nightingale, while well known as the founder of modern nursing, was also among the first to use proportional statistics (Cohen 1984). By tracking deaths by cause in military hospitals in Crimea she demonstrated that more soldiers died of hospital environmental causes than of injuries received in battle. Her volume, *Notes on Hospitals*, has been reissued in several versions as an example of one of the first medical outcomes studies (Rosenberg 1989).

2.2 Vestiges of History in Contemporary Nursing Roles

Nursing's history leaves a legacy that gives contemporary professional nurses special legitimacy in the medical division of labor, and a high level of public trust. However, that same legacy creates substantial barriers to the continued evolution of the profession of nursing in contemporary health care. Nursing developed as an ancillary occupation supportive to physicians. Nurses gained much of their professional status and public legitimacy from their close association with physicians. The subordinate status of nursing relative to medicine in the medical labor hierarchy, however, has hindered their quest for greater autonomy (Friedson 1970). The dominance of medicine over nursing is often cited as the archetype of settlement of competing occupational jurisdictional claims by subordination (Abbott 1988). Medical dominance over nurses is linked to the organizational frame of the modern hospital where the majority of nurses practice and where physicians exercise considerable professional authority. Indeed, nursing's quest for professional autonomy was not realized to any great extent until they began to practice in larger numbers and in a broader array of roles outside the hospital.

The association of early nurse training with religious orders has reinforced a continuing expectation of service and charity. These attributes are no longer valued as highly in Western culture, and no longer serve as sufficient motivating forces to recruit and retain professional nurses in today's world. Despite a shift from caring and nurturing functions to more technological responsibilities akin to those associated with male roles in other occupations, nursing has retained a feminine image and is thus subject to sex-role stereotyping and sex discrimination in professional relationships and labor practices. The large-scale entry of women into the labor force and into higher education has facilitated the advancement of nursing by increasing the absolute number of nurses, raising overall educational levels, and broadening scope of practice. However, women now have many more career options than in years past and nursing is attracting a smaller share of new career aspirants as suggested by a steady rise in the average age of practicing nurses (Buerhaus et al. 2000). Prospects for attracting enough qualified nurses to meet projected future demands for health care are uncertain, which poses a threat to the professional advancement of nursing. In the past when the supply of nurses has been insufficient to meet market demand, nursing has been vulnerable to the substitution of less-educated personnel and pressures to lower educational standards.

2.3 Nursing Education

The education of professional nurses has undergone a dramatic transformation since the 1970s that is still underway world-wide (Lynaugh and Brush 1996). Nursing education developed in hospital apprentice-like programs with a large service component. Indeed, until the Great Depression forced many graduate nurses into hospital employment, hospital nursing care was supplied largely by students in training. Leaders in American nursing advocated early in the twentieth century the standardization of nursing education within institutions of higher education. The Flexner Report of 1910 had brought about similar reform in medical education in the USA. The analogue of the Flexner Report for nursing, known as the Goldmark Report, was published in 1923 and recommended the baccalaureate degree as the minimum entry-level educational requirement for professional nursing. Subsequent reports and commissions supported the recommendation culminating in the American Nurses Association in the late 1970s setting 1985 as the target year for requiring the baccalaureate degree. While progress has been made toward the goal of baccalaureate education for nurses, the new century dawned without this goal having been achieved in the USA, although other countries including Australia, New Zealand, Iceland, Chile, and United Kingdom have implemented an academic degree requirement for new entrants into nursing.

Freidson (1970) described nursing as an incompletely closed profession because of its inability to set a minimum educational requirement. Three generic types of education programs prepare professional nurses: three-year hospital diploma programs, two-year associate degree programs, and four-year college/university programs. Neither employers nor legal practice acts distinguish between professional nurses by type of education. The proportion of US nurses

trained in hospital diploma programs declined from over 80 percent in 1960 to less than 8 percent in 1996. Community college programs instead of baccalaureate education replaced the diploma programs and now train over 60 percent of US nurses (Aiken and Gwyther 1995). By 1996, only about 40 percent of US nurses held a baccalaureate or higher degree, and 10 percent held master's or doctoral degrees (Moses 1997).

3. *Nurse Practitioners and Physician Assistants*

Nurse practitioners, also known as advanced practice nurses, and physician assistants are two relatively new additions to the medical division of labor, both originating in the USA in the 1960s. Nurse practitioners represent more of an evolution in the profession of nursing than a distinct new occupation while physician assistants constitute a new allied health profession.

Nurse practitioners (NPs) are registered nurses with advanced clinical training (Mezey and McGivern 1999). In the USA a master's degree is required for certification in advanced practice nursing. Nurse practitioners provide generalist care including routine physical assessment, treatment of uncomplicated episodic illness, continuing care for persons with stable chronic conditions, and acute care in hospital settings. Nurse midwives, a type of advanced practice nurse, provide routine obstetrical care including normal deliveries as well as routine health care for women. Nurse practitioners have successfully defined a legal scope of autonomous practice although attempts by physicians to exert control over their practice and remuneration is a continuing source of conflict (Sage and Aiken 1997).

Physician assistant (PA) education is generally 24 months in length. Programs exist in four different academic pathways: certificate programs, associate degree programs, bachelor's degree programs, and master's degree programs although the PA component is the same regardless of degree level (Hooker and Cawley 1997). A major difference between physician assistants and nurse practitioners is that physician assistants have accepted a legally dependent role working under the direction of a physician while nurse practitioners have fought for and obtained a legal scope of practice that does not require physician supervision (Sage and Aiken 1997).

3.1 *Historical Context*

While nursing's history and evolution is strongly tied to the development of the modern hospital, nurses have a long and distinguished history in public health and home care, settings in which they exercised considerable autonomy and professional leadership. Lillian Wald invented the term 'public health nurse' in 1893 to describe a role for nurses that was a combination of health promotion, illness care, and social reform (Buhler-Wilkerson 1993). Public health nursing and public health in general underwent substantial erosion in the USA with the rise of health insurance after World War II. However, the foundation for the emergence of nurse practitioners was the legacy of public health nursing.

The evolution of expanded roles for nurses and the development of physician assistants is often linked to a perceived shortage of generalist physicians in the USA in the 1960s and the market demand for physician substitutes. In the case of nurse practitioners, however, these factors created a favorable environment for expanding the scope of nursing practice that had long been advocated by nursing's leaders and that was exemplified by the early public health nurses. Thus, nurse practitioners, while often considered new allied health professionals, are really not new at all but nurses with legally sanctioned expanded scope of practice including some domains formerly controlled solely by physicians, such as drug prescription authority (Lynaugh and Brush 1996, Mezey and McGivern 1999). While educational and legal requirements for advanced nursing practice vary across countries, there is a global trend to replicate the successful US model (Fagin 1990, *British Medical Journal* 2000).

The origins of the new occupation of physician assistants in the USA stem from the return to the civilian workforce of military medical corps personnel in the 1960s at a time of a perceived domestic physician shortage. Most initial recruits to physician assistant programs were experienced medical corps personnel and the initial training programs were short-term on-the-job training. While the profession was predominantly male in its inception, it is now almost equally male and female, and the most recent graduating classes are over 60 percent female (Hooker and Cawley 1997). The early programs were located at medical schools but did not have academic standing or lead to degrees. Over time, degree-granting programs have been incorporated into schools of allied health.

Nurse practitioner education was rapidly incorporated into the large existing university-based infrastructure of schools of nursing. Enrollments in nurse practitioner programs grew steadily and by 1996 over 70,000 nurses in the USA had formal preparation as nurse practitioners, primarily at the master's level (Moses 1997). There was no comparable existing educational infrastructure for physician assistant education, and thus their numbers have grown more slowly than those of nurse practitioners, reaching 29,000 in 1997 (Cooper et al. 1998).

3.2 *Practice Issues*

A large research literature has developed on the safety and effectiveness of care provided by nurse practitioners and physician assistants. The Office of Tech-

nology Assessment of the US Congress conducted a synthesis of research published through 1985 and concluded that nurse practitioners and physician assistants working within their areas of competence provided comparable care to that provided by physicians, and in some aspects of care, such as patient satisfaction, nurse practitioners and physician assistants had better outcomes than physicians (Office of Technology Assessment [OTA] 1986). A special issue of the *British Medical Journal* (2000) reports similar findings from four randomized clinical trials evaluating care provided by nurse practitioners in the UK.

Competing jurisdictional claims between professions are unlikely to be solved by scientific evidence. Indeed, no matter how many studies have been undertaken including randomized controlled trials (Mundinger et al. 2000), many physicians have not accepted nurse practitioners (Kassrer 1994, Sox 2000). The opposition of physicians notwithstanding, the market demand for nurse practitioners has grown steadily in the USA, a trend that appears to be spreading to other countries (Norman 2000).

4. Other Allied Health Professionals

Over 200 different occupations and professions comprise the allied and auxiliary health care workforce. In the USA, this broad category accounts for 60 percent of the nation's 10.5 million health care workers and includes personnel who provide therapeutic, diagnostic, informational, and environmental services across the full array of health care delivery settings (Ruzek et al. 1999). The non-nursing allied health occupations are diverse in roles, functions, and education (Hamburg 1975). Many of these occupations are products of the twentieth century, having evolved in response to demands created by new technologies and niche roles created by changing patterns of medical care. Some of the alternative and complementary medicine occupations such as acupuncturists have roots in cultures that predate the development of modern Western medicine but are nevertheless considered newcomers within the Western division of medical labor (Eisenberg et al. 1998).

Education, work setting, and license/regulatory requirements vary vastly across the allied health occupations. Some fields require six months to one year training and are institutional on-the-job training programs, while others are university based. Physical therapy (PT) and occupational therapy (OT) are illustrative of the allied health professions that emerged primarily in the twentieth century.

4.1 Roles and Work Settings

Physical therapists plan and administer treatment to relieve pain, improve functional mobility, maintain cardiopulmonary functioning, and limit the disability of people suffering from physically disabling injury or disease. Approximately one-third are employed in hospitals; other major employers include rehabilitation facilities, home health agencies, nursing homes, school systems, and clinics. About 20 percent of physical therapists are in private practice.

Occupational therapists promote learning of skills necessary to perform daily tasks, diminish or correct pathology, and promote and maintain health. Occupational therapy focuses on both mental and physical disabilities. About 60 percent of occupational therapists work with people with physical disabilities. Some specialize in particular age groups such as the elderly or children. Occupational therapists working with the mentally ill or disabled in institutional and community-based settings typically provide activities that help mentally ill and disabled people learn to cope with daily stresses and manage their work and leisure time more efficiently (Institute of Medicine 1989). Most occupational therapists are employed by healthcare organizations and private practice is not common.

4.2 Evolution of Occupational and Physical Therapy

Occupational therapy traces its origins back 200 years to a French physician's observation that mental patients given menial tasks to perform improved in functional capacity faster than those who were idle. In the USA at the end of the eighteenth century, physician Benjamin Rush introduced work as a treatment for mentally ill patients at Philadelphia's Pennsylvania Hospital, and in 1906 the first training course for occupational therapists was established in Boston (Institute of Medicine 1989). World War I created a demand for rehabilitation of large numbers of injured soldiers, which served as a major impetus for expanding the scope of practice of occupational therapy to physically as well as mentally disabled persons, and for establishing the new occupation of physical therapy.

4.3 Education and Characteristics

Not surprisingly, given the growth impetus for physical and occupational therapists associated with the two World Wars, the early recruits to both occupations were predominately women. Over time both have remained predominately female occupations, although 25 percent of physical therapists are men compared with 5 percent of occupational therapists and 6 percent of nurses. The first training program in physical therapy in the USA began in 1917 as a short course for women aides. The success of these aides in the aftermath of World War I resulted in federal funding to increase the supply of therapists during World War II through accelerated programs to train college graduates in related fields such as physical education for roles in therapy. University-based degree-granting programs in physical therapy were

established concurrently. Physical therapy has thus been a university-based educational program almost from its inception.

Since 1979, the American Physical Therapy Association has advocated graduate level preparation for the profession. As of 2002, the Association will no longer accredit baccalaureate professional programs and is moving to master's and doctoral level professional education. While practitioners with a baccalaureate degree may still practice, the purpose of the change in accreditation is to shift the educational patterns for future cohorts. Physical therapy does not support career mobility from the physical therapy aide category to the professional category and has established non-overlapping educational requirements for the two groups thus avoiding the problem nursing has faced. Physical therapy has also developed viable options for private practice, which allows the occupation to continue to evolve independently, to some extent, from institutional managerial control of salaried practice. Physical therapy is considered to be more visible and autonomous than occupational therapy for these reasons (Institute of Medicine 1989).

Occupational therapy in many ways more closely resembles nursing in its composition and education than physical therapy, perhaps because its roots are in the nineteenth century with nursing and its knowledge and intervention domains are less discrete than physical therapy. The first formal 'professional' standards for education in occupational therapy were established in 1923 as 12-month training programs for high school graduates, in contrast to physical therapy, which moved directly to university-based education. There are presently two educational levels for practice: technical training at the two-year associate degree level and professional education at the baccalaureate degree for registered occupational therapists.

Occupational therapy has had difficulty attracting students reportedly because of its focus on mental health services, which are highly stigmatized, and poorer remuneration opportunities than alternative fields such as physical therapy. In 1995, there were 54,000 registered occupational therapists compared to 102,000 physical therapists in the USA (Bureau of Labor Statistics 1996).

5. *Global Perspectives*

Developing countries, which contain 84 percent of the world's population, account for only 11 percent of the world's health professionals (Schieber and Maeda 1999). Physicians, the most expensive category of health personnel, are often in relatively generous supply compared to nurses and allied health personnel in countries with modest annual health expenditures. An imbalance in the ratio of physicians to allied health workers has a number of consequences including excess expenditures on secondary and tertiary care leaving inadequate resources for primary care and public health as well as geographic imbalances in health care because physicians tend to locate in urban areas. The range in the ratio of nurses to doctors is enormous across countries: from one doctor to three or four nurses in northern European countries, to one doctor to 1.5 nurses in south-eastern Europe, and from one nurse to two doctors in Argentina, to one nurse to five doctors in Pakistan (Bankowski and Fülöp 1987).

Allied health workers provide much of the medical care in poor countries but are in short supply and often poorly trained. The potential of nurses and nurse midwives to improve access to and quality of health care in developing countries is well recognized. However, many countries have insufficient resources to create employment opportunities for qualified nurses (Bankowski and Fülöp 1986). Thus, nurses trained in developing countries migrate in substantial numbers to more affluent countries. Additionally allied health workers with some medical training such as feldshers in Russia are also in short supply because of the practice of admitting experienced health care workers to physician training programs (Ryan and Thomas 1996).

Developing countries tend to emulate the medical care systems in industrialized countries by building hospitals, purchasing technology, and training large numbers of physicians. Ultimately the prospects for improving health in the developing world are in investments in frontline allied health workers, primary care, and public health and safety.

Thus, nursing and other allied health occupations have good prospects for future growth in developing countries as well as in the sophisticated and highly technological health systems of industrialized countries.

See also: Alternative and Complementary Healing Practices; Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care Organizations: For-profit and Nonprofit; Health Care, Rationing of; Health Insurance: Economic and Risk Aspects; Managed Care; Medical Profession, The; National Health Care and Insurance Systems; Nursing Homes

Bibliography

Abbott A 1988 *The System of Professions*. University of Chicago, Chicago

Abel-Smith B 1960 *A History of the Nursing Profession*. Heineman, London

Aiken L H, Gwyther M E 1995 Medicare funding of nurse education: the case for policy change. *Journal of the American Medical Association* **273**(19): 1528–32

Bankert M 1989 *Watchful Care: A History of America's Nurse Anesthetists*. Continuum, New York

Bankowski Z, Fülöp T 1987 *Health Manpower Out of Balance*. Council for International Organizations of Medical Sciences, Geneva, Switzerland

British Medical Journal 2000 Doctors and nurses. Special issue (7241: 15 April)
Buerhaus P I, Staiger D O, Auerbach D I 2000 Implications of an aging registered nurse workforce. *Journal of the American Medical Association* **283**(22): 2948–54
Buhler-Wilkerson K 1993 Bringing care to the people: Lillian Wald's legacy to public health nursing. *American Journal of Public Health* **83**: 1778–86
Bureau of Labor Statistics, US Department of Labor 1996 *Occupational Outlook Handbook*. US Government Printing Office, Washington, DC
Cohen I B 1984 Florence Nightingale. *Scientific American* **250**: 128–32
Cooper R A, Laud P, Dietrich C L 1998 Current and projected workforce of nonphysician clinicians. *Journal of the American Medical Association* **280**(9): 788–94
Eisenberg D M, Davis R B, Ettner S L, Appel S, Wilkey S, Van Rompay M, Kessler R C 1998 Trends in alternative medicine use in the United States, 1990–1997: results of a follow-up national survey. *Journal of the American Medical Association* **280**(18): 1569–75
Fagin C M 1990 *Nursing Leadership: Global Strategies*. National League for Nursing, New York
Fairman J, Lynaugh J 1998 *Critical Care Nursing: A History*. University of Pennsylvania Press, Philadelphia, PA
Freeman P, Gómez-Dantés O, Frenk J 1995 *Health Systems in an Era of Globalization*. National Academy of Sciences, Washington, DC
Freidson E 1970 *Professional Dominance: The Social Structure of Medical Care*. Aldine, Chicago
Hamburg J (ed.) 1974 *Review of Allied Health Education*. University of Kentucky Press, Lexington, KY
Hooker R S, Cawley J F 1997 *Physician Assistants in American Medicine*. Churchill Livingstone, New York
Institute for the Future 2000 *Health and Health Care 2010: The Forecast, the Challenge*. Jossey-Bass, San Francisco
Institute of Medicine 1989 *Allied Health Services: Avoiding Crises*. National Academy Press, Washington, DC
Lynaugh J E, Brush B L 1996 *American Nursing: From Hospitals to Health Systems*. Blackwell, Cambridge, UK
Kassirer J P 1994 What role for nurse practitioners in primary care? *New England Journal of Medicine* **330**: 204–5
Mechanic D 1983 *Handbook of Health, Health Care, and the Health Professions*. Free Press, New York
Mezey M D, McGivern D O (eds.) 1999 *Nurses, Nurse Practitioners: Evolution to Advanced Practice*. Springer, New York
Moses E B 1997 *The Registered Nurse Population: Findings from the National Sample Survey of Registered Nurses*. Division of Nursing, US Bureau of Health Professions, Rockville, MD
Mossialos E, McKee M (eds.) 1998/99 Health care reform, public health, pharmaceutical policy, focus on Russia and Central Asia. *Eurohealth* **4**(6): Special Issue, Winter
Mundinger M O, Kane R L, Lenz E R, Totten A M, Tsai W, Cleary P D, Friedewald W T, Siu A L, Shelanski M L 2000 Primary care outcomes in patients treated by nurse practitioners or physicians: A randomized controlled trial. *Journal of the American Medical Association* **283**: 59–68
Norman S 2000 Making sense of higher level practice. *Nursing Standard* **14**(35): 47–56
Office of Technology Assessment (OTA) 1986 *Nurse Practitioners, Physician Assistants, and Certified Nurse Midwives: A Policy Analysis*. Health Technology Case Study No. 37. US Government Printing Office, Washington, DC
Parsons T 1951 *The Social System*. Free Press, Glencoe, IL
Rosenberg C E (ed.) 1989 *Notes on Hospitals: Florence Nightingale on Hospital Reform*. Garland, New York
Ruzek J Y, Bloor L E, Anderson J L, Ngo M 1999 *The Hidden Health Care Workforce: Recognizing, Understanding and Improving the Allied and Auxiliary Workforce*. University of California, San Francisco Center for the Health Professions, San Francisco, CA
Ryan T M, Thomas R 1996 Trends in the supply of medical personnel in the Russian Federation. *Journal of the American Medical Association* **276**(4): 339–42
Sage W M, Aiken L H 1997 Regulating interdisciplinary practice. In: Jost T (ed.) *Relation of the Health Professions*. Health Administration Press, Ann Arbor, MI, pp. 71–101
Schieber G, Maeda A 1999 Health care financing and delivery in developing countries. *Health Affairs* **18**(3): 193–205
Sox H C 2000 Independent primary care practice by nurse practitioners. *Journal of the American Medical Association* **283**: 106–8
Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York
Thomas L 1977 On the science and technology of medicine. *Daedalus* **106**: 35–46

L. H. Aiken

Health Promotion in Schools

Health promotion in schools is a topic of both practical and empirical interest. Schools which provide comprehensive school health promotion programs are more effective in encouraging children to adopt health-enhancing behaviors and in reducing health-compromising behaviors than schools that provide health education alone. Research into the effectiveness of comprehensive school health programs provides evidence of the positive influence that such programs can exert on student health behaviors. However, conditions that enable effective school health promotion programs to be initiated and maintained depend on several school organizational and program implementation factors.

1. *Changing Concepts of Health Promotion*

The meaning of health and health promotion have been transformed over the 1990s. Influenced by the World Health Organization (WHO), health is seen as more than avoiding illness. Similarly, health promotion involves more than helping people to choose healthy lifestyles.

Health promotion in the USA, Canada, and Western Europe has been influenced by WHO concepts contained in the Ottawa Charta for Health Promotion (1986). Health is viewed as a resource of daily living, a positive concept emphasizing social and personal

resources, as well as physical capacities. Responsibility for health rests not only with the health sector, but includes building healthy public policy, creating supportive environments, strengthening community action, developing personal skills, and reorienting health services. Accordingly, health promotion professionals are expected to work for health through advocacy, enabling individuals, and mediating among diverse sectors. Developments since the Ottawa Charta suggest increasing awareness of the role values play in health-related activity as well as increasing emphasis on social justice, reducing inequalities, and enhancing social cohesion. The Jakarta Declaration on Health Promotion in the 21st Century (1997) finally outlined priorities for promoting social responsibility for health, increasing investments for health development, increasing community capacity, empowering the individual, and securing an infrastructure for health promotion.

Currently, the ever-widening gap between rich and poor is an issue of great concern, and compelling evidence confirms that exclusion, both societal and material, is highly linked to health problems. The potential effects of inequalities on the health of populations have first been stressed by The United Kingdom Charta for Health Promotion, followed by reports from Canada and the US which reported profound variation in health and well-being between poor and nonpoor children. Health inequalities exist in level of mortality and morbidity, accidents and injuries, risk behavior, mental health, school achievement, and family violence. These differences have traditionally been interpreted in terms of risk behaviors resulting from lack of education, and material deprivation. Yet, careful analyses indicate that most of the variance in health outcomes among differing social classes is not accounted for by lifestyle factors. Recent analyses challenge the 'risk group' metaphor, showing that American and European children and adolescents in high-risk settings (low income families, high unemployment of the caregivers, poor neighborhoods) were in poorer health than those living in wealthier settings.

2. *Health-promoting Programs*

Numerous health-promoting programs have been published; the focus diverges with some publications emphasizing public health concerns and other reflecting an education perspective. In addition, the programs vary with respect to age range, comprehensiveness, theoretical background, and the importance of evaluation. While some studies comprised primarily of survey data, others examine the effect of intervention programs. Several papers incorporate a substantial body of epidemiological and demographic data as basis for their conclusions. Most discuss policy options and recommendations. Thus, substantial variation exists with respect to such factors as level of detail, use of research data, intended audience, and political viewpoint.

Rather than giving an overview of the current state of research in school-based health promotion (see David and Williams 1987), in the following, some examples of comprehensive programs will be given, prominent trends will be presented, and, finally, frequent domains in health promotion programs will be evaluated.

To date, not many prevention and intervention programs have incorporated the important developments in health promotion values and principles. The Dartmore Health Promotion Study (McIntyre et al. 1996) is one of the rare programs that coordinated school health instruction, health services, and a healthful environment to enhance the program's effect on the health of the heart and mental health of children in grades four, five, and six. Several studies suggest that children are not engaging in physical activity of sufficient length and intensity to improve cardiovascular fitness; during adolescence, the amount of physical activity further decreases. This has resulted in the development of community-wide Youth Exercise Promotion in order to prevent heart diseases and to influence blood pressure in pupils with risk for hypertension.

To give an example, the CATCH (Child and Adolescent Trial for Cardiovascular Health) is the most comprehensive multilevel program which has been extensively evaluated (Perry et al. 1997). The interventions targeted changes at the school level (food services offerings and minutes of class time devoted to physical activity) and at the student level (psychosocial factors, dietary behaviors, physical activity levels, smoking intentions, physiological indicators). The CATCH interventions consisted of: (a) Eat Smart, a school food service program to provide children with meals reduced in fat and sodium; (b) Catch PE, to increase the amount of time students spent in physical activity at school and for other times of the day; (c) Hearty Heart and Friends, 15 classroom lessons to teach dietary and physical activity information through a series of stories; and (d) Family Fun Nights, during which pupils and family members participated in 90 minutes of games, fitness-related activities, and healthy snacks.

Generally, health promotion programs have been confined to highly selective programs designed to identify specific health risks, such as tobacco smoking, alcohol consumption or drug use, and to train health-promoting behaviors, for example with respect to prevention of HIV. Most of these programs are designed for late childhood and early adolescence and pursue multiple strategies for preventing health damage, including more traditional strategies (for example, teaching drug information and/or general communication and problem-solving skills) and approaches based on the social influence model (which emphasizes

strategies for identifying and resisting social pressure to use drugs or alcohol). Time reserved for cognitive instruction was usually low, with most curricula ranging between 5 and 15 lessons. Although the overall cognitive progress in most programs was impressive, this was not sufficient to produce corresponding reduction in use of alcohol, drug or nicotine. Mostly, the effects were stronger and longer lasting for girls than for boys. Frequently, developmentally relevant variables are incorporated, for example increasing peer pressure or a risky lifestyle during adolescence. Resistance-Skill Training(s), however, was not always found to be effective strategies for preventing the onset of alcohol, drug or tobacco use; sometimes counterproductive effects were found.

Health promotion in middle schools is a neglected area, although the middle school years are a time of increasing risk for negative health related outcomes. Health promotion in the middle school mostly focuses on two approaches, a social decision approach (e.g., teaching strategies and skills that help children to make favorable decisions when confronted with health damaging situations and offers) and social skills training (in which children are prompted, coached, and guided to practice using the skills). These approaches are based on the observation that children are unlikely to retain and use information conveyed solely through formal lessons; they need opportunities for reinforcement and practice. Further, not many health promotion or health education programs are developed for elementary school children. The Claxton Elementary School Health Program (Landis and Janes 1995) is one of the rare examples. In this project, school administrators, parents, and local physicians worked together to discover what types of healthy and unhealthy activities children and parents are engaged in, and identified which health problems are of greatest concern to parents and teachers.

Professionals in school health education have frequently emphasized parent and family involvement in school health education. Unfortunately, preparing prospective educators for such involvement often has been overlooked. Teacher training for involving families in school health education were designed to provide staff members with the background necessary to incorporate family involvement concepts and activities into their classes. This may include, for example, (a) newsletters to provide information about health and the school health program to family members, (b) family calendars about quick activities—less than five minutes—that families can do daily, and (c) information about children's books, television programs and movies that address health topics.

Recently, the interaction with families of high-risk children became a major focus, particularly in settings outside school. The Triple P Program, for example, aims at helping parents of kindergarten children with sleeping and eating disorders to cope with these challenges by developing positive parenting attitudes and behaviors. Although it targets categorical symptoms rather than their common antecedents, it is highly valuable for preventing more severe future health problems in offspring.

To summarize, health promotion and education programs focus on late childhood and adolescent subjects and target problem behavior (which is highly inter-related such as drug and alcohol use) and unhealthy life styles (with respect to nutrition and physical activity). Stress prevention programs or programs teaching functional coping skills are less frequently applied. Most programs are centered in and around schools, but preventive approaches sometimes reach out to families with preschool children at risk

3. *Evaluation of School-based Prevention and Intervention Programs*

Quantitative research summaries are needed to deal with questions such as the relative efficacy of prevention and intervention programs with different aims and program orientations. In assessing program outcomes, five criteria should be met: (a) the use of appropriate comparison groups; (b) the use of appropriate pretests, post-tests, and follow up assessments; (c) examining research attrition; (d) the validity of dependent variable assessments; and (e) the use of statistical procedures appropriate for program evaluation research. In the following, quantitative reviews of prevention and intervention programs targeting adolescent drug use, smoking, and alcohol consumption were presented. They represent a new approach in that program classification schemes were used to categorize outcome, which were then employed in a quantitative meta-analysis of study effect size, based on the above criteria.

3.1 *Meta-analyses of Smoking Prevention Programs*

A substantial number of research report evaluating programs designed to prevent adolescent smoking have been published in the 1980s and 1990s. A very important set of questions deals with: (a) how successful these programs have been in general; (b) whether programs of different orientation show different levels of success; and (c) what variables, in addition to program orientation, may account for different levels of success. The meta-analyses conducted by Bruvold (1993) was based on 94 separate intervention studies published between 1970 and 1990. These studies evaluated school-based programs designed to deter adolescents from smoking tobacco products and were screened for methodological rigor. Each study reviewed was independently evaluated on the five methodological criteria detailed above. Forty-eight studies with adequate and defensible method-

ology were included in the meta-analyses. Program classification revealed four different types of orientation of the prevention program: (a) rationally-oriented information approach; (b) developmentally-oriented affective education; (c) social norms-oriented alternatives; and (d) social reinforcement-oriented social pressures approach. The rational model supports traditional classroom interventions where information about the short- and long-term health effects of tobacco is didactically presented. Reinforcement, social norms, and developmental interventions have two features in common: all are relatively new, and all were developed because of dissatisfaction with the uncertain effectiveness of the traditional didactic approach. Weighted average effect sizes were computed for all 48 better methodology studies for knowledge, attitude, and behavior, respectively. Positive results with similar high effect sizes for knowledge were found in programs with developmental, social norms, and social orientation. Results for attitude were generally not as positive. Programs classified primarily as social reinforcement-oriented seem to have had the most impact on attitudes, followed by those classified primarily as developmental. This suggests that newer orientations were more successful in modifying attitudes than more traditional rational approaches. Behavioral effect sizes were found to be largest for interventions with a social reinforcement orientation, moderate for interventions with either a developmental or a social norm orientation, and small for interventions with the traditional rational orientation. Altogether, the results of this meta-analyses suggest that rational orientation had very little if any impact on behavior and attitudes, and that the developmental and social reinforcement orientation revealed similarly positive effects on attitudes and behavior. Not much difference, depending on program orientation, emerged with respect to the cognitive outcome.

3.2 Comparison of Tobacco and Alcohol Intervention Programs

Frequently, school-based interventions were designed to deter children and adolescents from starting to consume alcohol and tobacco, and to convince those already consuming these substances to stop or reduce consumption. Bruvold and Rundall (1988) conducted a meta-analysis on those programs that combined tobacco and alcohol prevention. Again, they classified the programs according to their theoretical orientation into rational, social reinforcement, developmental, and social norms. In addition, the major methodological requisites for research evaluating school-based tobacco and alcohol interventions detailed above were applied to the studies incorporated in the meta-analysis. Altogether, 19 studies were found which had a control or comparison group, reported attrition rate, employed pretest, post-test and follow-up measures, and used adequate statistical analyses. A very consistent result was that the intervention program, based upon the rational model produced fewer positive behavior and attitude effects but more positive knowledge effects than those in which the rational model was *not* dominant. Knowledge effects were consistently positive over all studies reviewed, but attitude effects were mixed, some positive and some negative. For smoking behavior, the effect size was satisfactory for long- and short-term measures, but behavioral results for alcohol were meager for long- and short-term measures. Thus, while the alcohol interventions had an effect in the desired direction on alcohol consumption, the average impact was not as large as that for smoking.

3.3 Effectiveness of Drug Abuse Resistance Education

School-based drug prevention programs have been an integral part of the US and European antidrug campaign during the 1980s and 1990s. Although programs have proliferated, Project DARE (Drug Abuse Resistance Education) is the most widely implemented school–based, drug use prevention program in the United States, and has been adopted by approximately 50 percent of school districts nationwide. The drug prevention curriculum offers 17 lessons, usually once a week, and is applied in elementary schools, as well as in junior and senior high schools. Lessons focus on teaching pupils the skills needed to recognize and resist social pressures to use drugs. In addition, lessons focus on providing information about drugs, teaching decision-making skills, building self-esteem, and choosing healthy alternatives to drug use. Didactics include lectures, group discussion, question and answer sessions, audiovisual material, workbook exercises, and role playing. Ennett et al. (1994) used meta-analysis to review methodological rigorous evaluations of short-term effects. The DARE effect size for drug use behavior was very low across the studies. For all outcomes considered, the DARE effect size means were substantially smaller than those of programs emphasizing social and general competencies and using interactive techniques. Considering the enormous amount of public funding for DARE, the short-term effectiveness of DARE for reducing or preventing drug use behavior can not be considered as satisfactory.

4. Implications for School Health Programs and Personnel

The School Health Education Evaluation indicated that for school health education to promote health for children, schools should devote at least 50 hours of classroom time each year to health instruction. How-

ever, in many industrialized countries, particularly in southern and eastern Europe, these standards are not met (Takashini and Hamburg 1997). Limited resources, including curricular time, make it unlikely that health promotion in schools will realistically increase in strength. Thus, it is important to develop more efficient ways to promote children's health, for example by adding complementary interventions to school health education from other sources.

Meta-analyses of more recent school-based studies show that it is necessary to carefully evaluate the programs. Overall, improvements of methodology and conceptualization of programs is noticeable, but still a need exists to upgrade substandard procedures regarding experimental attrition and establishing the validity of dependent variable measures. Further, there is a need for developing program intervention models that clearly and fully follow tenets of the guiding theory. In addition, prevention and intervention programs which have been designed and evaluated so far are largely confined to certain age groups (frequently to adolescents) and limited to the treatment of problem behaviors (such as drug, alcohol, or tobacco use). The health needs of younger ages, for example preschool children or children from grades one to five, are largely neglected.

Based on the recent developments of health promotion values and principles detailed in the Jakarta Declaration, health promotion in schools should acknowledge societal factors such as inequalities and poverty as important determinants of health in children and adolescents. Health promotion in schools therefore has to incorporate the community level of life quality in designing programs. School health personnel need to deal with issues that influence children and youth's health, but is beyond the traditional focus of health education and behavior change. They need to assess the quality of life as it is perceived by children and adolescents, and identify the specific needs of individual children. Then, community-based action can be developed to address these issues.

See also: Adolescent Health and Health Behaviors; Alcohol Use Among Young People; Childhood Health; Health Education and Health Promotion; Health Interventions: Community-based; Mental Health: Community Interventions; Physical Activity and Health Promotion; Public Health; Smoking Prevention and Smoking Cessation

Bibliography

Bruvold W H 1993 A meta-analysis of adolescent smoking prevention programs. *American Journal of Public Health*. **83**: 872–80

Bruvold W H, Rundall T G 1988 A meta-analysis and theoretical review of school-based tobacco and alcohol intervention programs. *Psychology and Health*. **2**: 53–78

David K, Williams T (eds.) 1987 *Health Education in Schools*. Harper and Row, London

Ennett S T, Tobler N S, Ringwalt C L, Flewelling R L 1994 How effective is drug abuse resistance education? A meta-analysis of project DARE outcome evaluations. *American Journal of Public Health*. **84**: 1394–401

Jakarta Declaration on Health Promotion into the 21st Century 1997 World Health Organization, Geneva

Landis S E, Janes C L 1995 The Claxton elementary school health program: Merging perceptions and behaviors to identify problems. *Journal of School Health*. **65**: 250–4

McIntyre L, Belzer E G, Manchester L, Blanchard W, Officer S, Simpson A C 1996 The Darmouth Health Promotion Study: A failed quest for synergy in school health promotion. *Journal of School Health*. **66**: 132–7

Ottawa Charta on Health promotion 1986 World Health Organization, Geneva

Perry C L, Sellers D E, Johnson C, Pedersen S, Bachman K J, Parcel G S, Stone E J, Luepker R V, Wu M, Nader P R, Cook K 1997 The child and adolescent trial for cardiovascular health (CATCH): Intervention, implementation, and feasibility for elementary schools in the United States. *Health Education & Behavior*. **24**: 716–35

Takanishi R, Hamburg D A (eds.) 1997 *Preparing Adolescents for the Twenty-First Century: Challenges Facing Europe and the United States*. Cambridge University Press, New York, pp. 82–107

I. Seiffge-Krenke

Health Psychology

Health psychology can be defined as the application of psychological knowledge and methods to the study, prevention, and management of physical diseases and disorders. Evident in this definition is the fact that health psychology is both a science and an applied discipline intended to promote human welfare. However, unlike many psychological disciplines, such as clinical or counseling psychology, health psychology addresses issues of physical rather than mental health. Emotional adaptation and mental health are considered in health psychology only to the extent that they influence—or are influenced by—physical health and illness (Schwarzer and Gutiérrez 2000).

The formal emergence of health psychology in the 1970s was influenced by a variety of factors. By that time, for industrialized nations infectious illnesses such as smallpox and influenza were no longer the leading cause of death. Through advances in basic biomedical science and its application in public health, vaccines, antibiotics, and acute clinical care, infectious illnesses had been replaced by several chronic diseases as the leading sources of morbidity and mortality. Cancer, coronary heart disease, cerebrovascular disease, and diabetes differ from infectious diseases in three fundamental ways that assured the emergence of health psychology. First, these illnesses are influenced by behavior. Daily habits or 'lifestyle' factors are

major contributors to their development and course. Second, unlike the acute crisis of infectious illness, chronic illnesses have far-reaching impacts on emotional, social, and vocational adaptation over long periods of time. Finally, behavior change is a key component of care for such conditions. Hence, behavior plays a prominent role in the development, impact, and management of the current major health threats in industrialized nations, making psychological research and practice essential in comprehensive medical research and care. This is the case for our most common chronic diseases, such as cancer and heart disease, and for more recently emerging diseases such as HIV/AIDS.

1. *History, Foundations, and Related Disciplines*

The emergence of health psychology was made possible by a variety of developments at the interface of behavioral and biomedical science. The work of Walter Cannon, Hans Selye, and other physiologists working in the 1930s, 1940s, and 1950s described physiological changes induced by environmental stressors and negative emotions (e.g., the 'fight or flight' response). This identified a pathway through which psychological factors could impact physical health. Shortly thereafter, epidemiologists determined that several behaviors—notably smoking, excessive intake of fat and calories, and low levels of physical activity—were risk factors for the subsequent development of cancer and cardiovascular disease. Elsewhere, psychologists were beginning to explore the application of learning principles to the treatment of physical conditions, as in the seminal work by Neal Miller on the operant conditioning of physiological responses and the work of William Fordyce on the use of behavioral contingencies in the management of pain and disability. At the same time, clinical psychologists—led primarily in the United States by Joseph Matarazzo—were beginning to establish research and clinical roles within academic medicine through appointments to medical school faculties and the establishment of related academic medical departments and services.

In response to a variety of government and organizational initiatives, health psychology became a recognized subdiscipline in the late 1970s. For example, it was established as a formal division within the American Psychological Association in 1978, and the first health psychology handbook was published in 1979 (Stone et al. 1979). At the same time, an interdisciplinary field—behavioral medicine—was established, sharing the focus on behavior in the development, prevention, and management of medical conditions (Schwartz and Weiss 1978). Although most health psychologists view themselves as engaging in behavioral medicine, that field includes concepts, methods, and professionals from medicine, public health, and other disciplines. Hence, health psychology is a component of this broader field. Because health psychology addresses the impact of behavior on health beyond its role in established disease and disability, its emergence represented a departure from older fields such as medical and rehabilitation psychology. The former refers traditional clinical psychological services (e.g., assessment, consultation, and psychotherapy) within medical settings, whereas the latter refers to similar services and research in the context of physical disabilities (e.g., spinal cord or brain injuries).

The biopsychosocial model is a conceptual cornerstone in health psychology and behavioral medicine. In the traditional biomedical model which underlies much of modern medical research and practice, disease is seen as stemming from disordered biological systems. Aberrations in biochemistry or physiology cause disease, alter its course, and are intervention targets. Psychological or social factors are seen as largely irrelevant. In a landmark paper, Engel (1977) articulated the biopsychosocial model as an alternative. Here biological, psychological, and social processes are seen as inter-related influences on and effects of health and illness. The onset, course, and treatment of illness is best understood as involving each of these interacting, reciprocally determined levels of analysis.

Current research in health psychology consists of three general topics. The first—health behavior and prevention—focuses on the determinants of behaviors that influence disease (e.g., diet, exercise, smoking), and the effectiveness of interventions to reduce these risks. The second—stress and illness—focuses on direct connections between psychological processes and the development of illness. Rather than the intervening effects of habits and lifestyle factors, this research examines the impact of environmental stress and related personality and social processes on psychobiological mechanisms and the pathophysiology of disease. The third research area—the psychosocial impact and management of illness—focuses on the emotional and social consequences of illness and psychological interventions intended to improve adjustment and supplement traditional medical care.

2. *Health Behavior and Prevention*

Decades of epidemiological research have demonstrated the impact of daily habits on health (see *Health Behaviors*). A variety of conceptual models have been developed to describe the determinants of these behaviors, and thereby provide targets for interventions intended to reduce health risks (Armitage and Conner 1999). For example, the health belief model suggests that several cognitive factors influence health behav-

ior: the value or importance an individual places on health; beliefs about vulnerability conferred by a specific behavior; beliefs about the severity of the related illness; the individual's beliefs about the effectiveness of a behavior change in reducing risk; and a view of the relative costs and benefits of that change. From this perspective, interventions designed to increase the perceived importance of good health, strengthen the belief that a specific behavior impacts health, make clear the severity of the disease or condition to be prevented, emphasize the risk reducing impact of behavior change, and make salient the relative benefits of this change in promoting more healthy lifestyles. Based on this and closely related models of health behavior, interventions ranging from individual counseling to national advertising campaigns have been developed.

Other models of health behavior describe the adoption of a healthier lifestyle as a cyclical process, in which people consider and initiate behavior change, often revert to less healthy habits, and perhaps later renew their efforts (Weinstein et al. 1998). Different interventions may be optimal at different stages in this process. Informational strategies derived from the health beliefs model may be best for individuals who are in the earliest stages, either before or as they begin to consider a change in lifestyle (see *Health Behavior: Psychosocial Theories*). In contrast, active planning about the nature of 'high risk' situations where a return to old habits is likely and rehearsal of strategies for avoiding or managing such relapses may be useful after they have initiated changes. A specific health behavior change can also be addressed to many different intervention levels. For example, to increase regular physical activity, individual level counseling can be supplemented by small group or organizational level interventions such as couple, family, or workplace programs. Further, public policy level interventions can be effective, such as taxes to increase the cost of cigarettes or the construction of public recreational facilities.

For many health behavior changes, the biological level of analysis is a critical consideration. Nowhere is this more evident than in the case of smoking cessation. It has become quite clear that nicotine is an addictive component of tobacco, and that habitual smoking can be seen, in part, as the self-regulation of this addictive agent. Hence, nicotine replacement therapy (e.g., nicotine gum or trans-dermal patches) is a valuable addition to psychological or behavioral approaches.

The biopsychosocial approach to health behavior change is also evident in our current understanding of the nature and treatment of obesity. Body weights 25 percent or more above normal are associated with increasing risk of diabetes, high blood pressure, coronary heart disease, and stroke. Further, in industrialized countries, the prevalence of obesity is increasing. Vulnerability to obesity is influenced by genetic factors. There are several likely physiological 'sites' of these genetic vulnerabilities, including resting metabolic rate, the degree of weight gain in response to overfeeding, and the location of fat deposition in response to excess energy intake. The energy balance model identifies two additional influences on obesity—the amount of energy (i.e., calories) consumed and the amount expended. Simply, individuals gain weight when the amount of energy consumed exceeds energy expended. Hence, caloric restriction (i.e., dieting) and increased energy expenditure through physical activity are critical determinants of obesity and essential components of treatment. However, larger sociocultural and economic factors also affect obesity. In industrialized nations, the availability of high-fat, energy-rich foods has increased dramatically in recent decades, and levels of leisure time and work-related physical activity have decreased. Especially for the genetically vulnerable, this sociocultural and economic context for the individual's energy balance can be a 'toxic environment' promoting obesity (Hill and Peters 1998). Hence, our understanding of health behavior influences on obesity is facilitated by the comprehensive perspective of the biopsychosocial model.

Smoking and obesity are very difficult to treat in adulthood, even with multicomponent, multilevel interventions. Research and practice on health behavior and prevention has begun to address the prevention of unhealthy behavior. That is, rather than focus on improved methods of smoking cessation and weight loss in adults, more attention has been paid recently to methods of preventing smoking and obesity in childhood and adolescence, and on the modification of these and other risk factors in youth. This trend has necessitated a developmental view of health habits and behavioral risk factors, as well as age-appropriate interventions at the individual, family, school, and community levels. Even though the modification of health behaviors in adulthood remains a priority in health psychology, early prevention is likely to become more widespread in the future.

The emergence of health psychology and its focus on health behavior change preceded by only a few years the identification of Acquired Immune Deficiency Syndrome (AIDS) and the HIV infectious agent. Quickly, epidemiological and clinical research identified intravenous drug use and unprotected sexual activity as primary risk factors. As a result, health psychologists have been closely involved in efforts to prevent HIV infection (see *HIV Risk Interventions*). As in other health behavior change efforts, multiple levels of intervention have been employed. Individual level programs include educational interventions based on the health behavior model and related approaches such as enhancing the individual's skill in negotiating condom use during intercourse. Community level approaches have focused on changing norms regarding sexual behavior and advocacy of 'safer sex' by

visible members of high-risk communities. Institutional approaches have involved increased access to clean needles for drug users.

3. Stress and Disease

A second major focus of research and practice in health psychology involves the impact of stress, negative emotions, and other psychosocial factors on the development of physical illnesses (see *Stress and Coping Theories*). In contrast to the effects of health behavior described above, this area addresses more direct connections between mind and body. In the general view guiding this research area, the stresses and strains of everyday life exact a toll on the individual's physiology. Accumulated over many days, months, or even years, such physiological changes can initiate, hasten, or exacerbate the development of illness. Further, some characteristics of the individual's personality or social environment can make them more or less susceptible to these psychobiological influences on disease.

A growing array of environmental and personality characteristics have been identified in epidemiological research as risk factors for physical illness and premature death. For example, individuals who are socially isolated or report low levels of available social support are at increased risk of cardiovascular disease, cancer, and premature death (House et al. 1988). High levels of job stress also place people at greater risk of illness (Schnall et al. 1994), as does the chronic stress of caring for seriously ill family members (Schulz and Beach 1999). Individuals who are prone to anger and hostility are at increased risk of developing cardiovascular disease and dying prematurely, as are people who report symptoms of anxiety, depression,pessimism, and hopelessness (Smith and Gallo 2001).

People who are socially isolated, experiencing high levels of job stress, or prone to negative emotions may be more likely to smoke, overeat, consume too much alcohol, or avoid regular exercise. These health behaviors could account for some of the effect of the social environment and personality on health. However, the association of these psychosocial risk factors with subsequent morbidity and mortality remains significant even when the effects of health behavior are controlled (Adler and Matthews 1994). The prevailing view in this research area is that the physiological effects of stressful environments and negative emotions are the link between psychosocial risk factors and subsequent disease. These mechanisms are best understood in two general pathways—the effects of stress on the cardiovascular system (Rozanski et al. 1999) and its effects on the immune system (Kiecolt-Glaser and Glaser 1995).

Environmental threats and demands evoke transient increases in heart rate, blood pressure, and concentrations of various hormones (e.g., epinephrine, norepinephrine, cortisol, etc.). In human and animal research, over time these stress-induced physiological changes appear to promote more enduring levels of high blood pressure, and initiate and hasten the development of atherosclerosis in the coronary and carotid arteries. Atherosclerosis in these sites increases the risk of coronary heart disease and stroke, respectively. Environmental stressors and the brief physiologic changes they evoke can also precipitate acute manifestations of cardiovascular disease (e.g., temporary reductions in oxygen supply to the heart muscle) among individuals with pre-existing atherosclerosis (Rozanski et al. 1999). Further, psychosocial risk factors identified in epidemiological studies as risk factors for cardiovascular disease tend to be related to this psychophysiological mechanism. For example, social support generally reduces the magnitude of these stress-induced physiological reactivities (Uchino et al. 1996), and personality characteristics associated with increased risk of cardiovascular disease such as hostility are associated with more pronounced cardiovascular and neuroendocrine reactivity (Smith and Gallo 2001). Additional evidence of this mechanism comes from nonhuman, primate models of environmental stress, individual differences in social behavior, and the development of atherosclerosis. For example, subjected to the chronic stress of being housed with new cage-mates every few weeks, socially dominant but not subordinate male monkeys develop more severe atherosclerosis than when they are left in stable, less stressful housing conditions. Presumably, the recurring threats to status among dominant animals and the effort required to reassert status evoke physiological reactivity. Further, this effect can be prevented when the input of the sympathetic nervous system to the heart is blocked pharmacologically (Manuck et al. 1995). That is, reduction of the usual cardiovascular 'fight or flight' response though pharmacologic blockade prevents the otherwise unhealthy effect of psychosocial factors on atherosclerosis. Such findings provide the outlines of an answer to age-old questions about the influence of the mind on the body, and suggest that stress reduction could be useful in the treatment of cardiovascular disease. Interventions designed to reduce psychological stress have been found to produce beneficial effects on the health of patients with coronary heart disease (Rozanski et al. 1999).

A variety of illnesses from the common cold to some forms of cancer are influenced by the immune system, as is the rate at which wounds heal. The immune system is a complex set of structures and processes, the overall function of which is to identify and destroy foreign bodies (e.g., viruses) and aberrant cells (e.g., malignancies), and to facilitate repair of damaged tissues. Reductions in immune system functioning render the individual more susceptible to conditions it otherwise controls. A growing body of evidence suggests that environmental stress and factors that

influence the individual's susceptibility to stress impact immune functioning (Rabin 1999). Factors as diverse as the stress of final examinations to marital discord have been found to reduce immune functioning. This interdisciplinary field, psychoneuroimmunology (PNI), is the second area in which the study of mind–body connections is quite advanced. Nearly all of the psychosocial variables identified in epidemiological research as risk factors for disease have been found to be related to immune functioning (Kiecolt-Glaser and Glaser 1995).

Health behaviors might be responsible for some effects of stress on immune functioning. Smoking, inactivity, poor diet, and even sleep deprivation can produce deficits in immune functioning, and are therefore plausible links between stress and immune functioning. However, as in the case of research on cardiovascular disease, statistical control of health behaviors does not eliminate the effect of life stress and chronic negative emotions on immune functioning. As described above, environmental stress and negative emotions are associated with heightened circulating levels of several neuroendocrine factors such as cortisol and catecholamines. These neuroendocrine responses can disrupt immune functioning. It is not clear that the type and degree of immune disruption produced by stress is sufficient to account for the effects of such psychosocial risk factors on the development and course of even minor illnesses, let alone conditions as serious as cancer. Further, although some studies have demonstrated that stress-reducing interventions can have positive effects on immune functioning even among high-risk groups such as individuals with HIV, the evidence of substantial health benefits of interventions based on the PNI literature is preliminary. Nonetheless, the sophistication and clinical implications of this research area are growing rapidly.

4. *Psychosocial Aspects of Medical Illness and Care*

The third major topic in health psychology involves the psychosocial impact of acute and chronic medical illness, and the usefulness of psychological interventions as adjuncts to standard medical care. Some research is intended to identify the type and extent of psychosocial impacts of medical illness such as pain, disability, emotional distress, and disruption of personal relationships. Other work seeks to identify modifiable psychosocial influences on these outcomes (e.g., social support, coping styles, etc.), as well as psychosocial influences on the disease process itself (e.g., depression). Such studies can guide the development of psychosocial interventions. Although the patient is typically the focus of such interventions, some are directed at family members, caregivers, or even members of the healthcare team.

An early example of this topic was Irving Janus's psychological study of surgical patients. Based on patient interviews, Janus described a curvilinear association between the patient's anxiety before an operation and subsequent physical recovery. Those with moderate levels of presurgical anxiety experienced the least pain and the most rapid recovery compared to patients with very low or high levels. Janus argued that moderate anxiety facilitated the 'work of worry,' helping patients to develop realistic expectations and prepare to cope with later discomfort. In contrast, patients with high levels of preoperative anxiety were too distressed to engage in adaptive preparations, and patients with very low levels preoperative distress were not sufficiently motivated to prepare. Decades of subsequent research have indicated that presurgical anxiety is related to postoperative recovery in a simple linear manner (i.e., greater preoperative distress is associated with slower and more complicated recovery), providing the impetus for now routine psychological preparation of patients undergoing many potentially stressful medical procedures (Contrada et al. 1994). The provision of social support and information, hypnosis and relaxation therapies, and rehearsal of distraction and other coping strategies have been found to be effective in reducing emotional and physical discomfort, reducing medical complications, and accelerating wound healing and recovery after invasive diagnostic procedures, chemotherapy, childbirth, and even open heart surgery (Kiecolt-Glaser et al. 1998). In the case of young children, interventions intended to reduce parents' distress can facilitate the child's recovery, presumably by enhancing the parent's ability to comfort the child. These interventions are cost effective; interventions as brief as 30 minutes can reduce the length of hospitalizations by a day or more (Devine 1992). Research in this area suggests it is important to match interventions and the patient's coping style. Additional information is useful for patients who prefer involvement in their care, whereas training in distraction is best for patients who manage stress with avoidance.

Chronic illnesses such as coronary heart disease, cancer, and diabetes create many burdens for patients. These conditions pose the threat of further limitations and early death, often are associated with discomfort, disrupt social and vocational functioning, and involve unpleasant medical interventions and demands for patient behavior change. These threats to quality of life are in and of themselves important targets for health psychology interventions. However, emotional distress and poor adherence to prescribed medical regimens can contribute directly to poor medical outcomes. Interventions to facilitate coping with chronic disease, managing unpleasant symptoms, and making necessary changes in health behavior can improve functioning among these individuals and in some cases improve prognoses (Compas et al. 1998, Dusseldorp et al. 1999). For some diseases (e.g.,

essential hypertension, diabetes), behavioral interventions for changes in diet, increased physical activity levels, weight loss, and adherence to medical regimens are particularly effective, as these patient behaviors are key components of medical management. The patient's success in implementing and maintaining these behavior changes is a central determinant of health (see *Chronic Illness, Psychosocial Coping with*).

In other conditions (e.g., arthritis, headache, chronic back pain), psychological interventions are useful additions to standard care. Stress management, training in pain coping techniques, reinforcement for increasing levels of physical activity, and interventions designed to improve adaptive communication in couples and families have been found effective for reducing pain and reliance on pain medication, and in increasing functional activity levels (Compas et al. 1998). HIV/AIDS poses nearly all of these challenges in that patients are faced with a life-threatening and sometimes disabling disease, painful and distressing symptoms, and the necessity of adherence to complex medical regimens. Behavioral interventions to reduce stress, control unpleasant symptoms and treatment side effects, and improve adherence are valuable additions to HIV/AIDS care (Sikkema and Kelly 1996).

5. *Emerging Challenges and Opportunities in Health Psychology*

The rapid growth and evolution of health psychology seems likely to continue. Developments in basic and applied research in each area of the field have included many contributions to our understanding of health and illness, and to efforts to prevent and manage illness. Three factors are likely to shape health psychology in years to come. First, rising costs of medical care will continue to encourage the development of effective methods to prevent illness, maximize the benefits of care and limit its costs, and reduce the economic impact of illness. Although psychological interventions face growing competition from traditional medical services in a tightening climate in healthcare financing, effective interventions of any type will be well received.

Second, all areas of medical science and practice, including health psychology and behavioral medicine, must increase attention to the nature and management of health problems across diverse segments of the population. The growing concern with women's health and the health of ethnic minorities are obvious examples of this issue. Behavioral influences on health, the impact of illness on psychosocial functioning, and optimal intervention strategies in prevention and management are likely to vary across these groups. More recently, it has become clear that socioeconomic status is also a critical consideration. Within industrialized nations, socioeconomic status is related to health and life expectancy. Health behavior, environmental stress, and access to medical care are important factors in this association (Adler et al. 1993). Hence, research and applied agendas in health psychology are likely to shape and be shaped by the emerging emphasis on the socioeconomic gradient in health. Of course, behavioral issues are no less important in the health of poorer nations, and an emerging 'world health' perspective will identify new roles for health psychology.

Finally, advances in basic and applied medical science will continue to shape health psychology. For example, developments in medical imaging have made it possible to study the impact of stress and other psychological factors on disease in more definitive ways. The revolution in medical genetics is making it possible to study gene behavior and gene environment interactive effects on disease more precisely, and may create the opportunity to focus behavioral risk reduction on those at greatest risk. The development of the artificial kidney and organ transplantation procedures created the need to understand the psychosocial aspects of new medical technologies. Similarly, advances in genetic testing to assess risk of illnesses such as cancer have created a new area of psychosocial research. The last three decades of theory, research, and practice in health psychology have produced remarkable growth and innovation. Although the future is difficult to predict, it appears to hold even more progress and impact for this recent variety of social and behavioral science.

See also: Adolescent Health and Health Behaviors; Behavioral Medicine; Biopsychology and Health; Health and Illness: Mental Representations in Different Cultures; Health Behavior: Psychosocial Theories; Health Behaviors; Health Behaviors, Assessment of; Illness Behavior and Care Seeking; Pain, Health Psychology of; Public Health

Bibliography

Adler N, Boyce W T, Chesney M A, Folkman S, Syme S L 1993 Socioeconomic inequalities in health. *Journal of the American Medical Association* **269**: 3140–5

Adler N, Matthews K 1994 Health psychology: Why do some people get sick and some stay well? *Annual Review of Psychology* **45**: 229–59

Armitage C J, Conner M 2000 Social cognition models and health behaviour: A structured review. *Psychology and Health* **15**: 173–89

Compas B E, Haaga D A, Keefe F J, Leitenberg H, Williams D A 1998 Sampling of empirically supported psychological treatments from health psychology: Smoking, chronic pain, cancer, and bulimia nervosa. *Journal of Consulting and Clinical Psychology* **66**: 89–112

Contrada R, Leventhal E, Anderson J 1994 Psychological preparation for surgery: Marshaling individual and social

resources to optimize self-regulation. In: Maes S, Leventhal H, Johnston M (eds.) *International Review of Health Psychology*. Wiley, New York/Chichester, UK, Vol. 3, pp. 219–66
Devine E 1992 Effects of psychoeducational care for adult surgical patients: A meta-analysis of 191 studies. *Patient Education and Counseling* **19**: 129–42
Dusseldorp E, van Elderen T, Maes S, Meulman J, Kraaij V 1999 A meta-analysis of psychoeducational programs for coronary heart disease patients. *Health Psychology* **18**: 506–19
Engel G L 1977 The need for a new medical model: Challenge for biomedicine. *Science* **196**: 129–36
Hill J O, Peters J C 1998 Environmental contributions to the obesity epidemic. *Science* **280**: 1371–4
House J S, Landis K R, Umberson D 1988 Social relationships and health. *Science* **241**: 540–5
Kiecolt-Glaser J K, Glaser R 1995 Psychoneuroimmunology and health consequences: Data and shared mechanisms. *Psychosomatic Medicine* **57**: 269–74
Kiecolt-Glaser J K, Page G G, Marucha P T, MacCallum R C, Glaser R 1998 Psychological influences on surgical recovery: Perspectives from psychoneuroimmunology. *American Psychologist* **53**: 1209–18
Manuck S B, Marsland A L, Kaplan J R, Williams J 1995 The pathogenicity of behavior and its neuroendocrine mediation: An example from coronary artery disease. *Psychosomatic Medicine* **57**: 275–83
Rabin B S 1999 *Stress, Immune Function, and Health: The Connection*. Wiley-Liss & Sons, New York
Rozanski A, Blumenthal J A, Kaplan J R 1999 Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. *Circulation* **99**: 2192–217
Schnall P L, Landsbergis P A, Baker D 1994 Job strain and cardiovascular disease. *Annual Review of Public Health* **15**: 381–411
Schulz R, Beach S R 1999 Caregiving as a risk factor for mortality: The caregiver health effects study. *Journal of the American Medical Association* **282**: 2215–19
Schwartz G, Weiss S 1978 Yale Conference on Behavioral Medicine: A proposed definition and statement of goals. *Journal of Behavioral Medicine* **1**: 3–11
Schwarzer R, Gutiérrez B 2000 Health psychology. In: Pawlik K, Rosenzweig M R (eds.) *The International Handbook of Psychology*. Sage, London, pp. 452–65
Sikkema K J, Kelly J A 1996 Behavioral medicine interventions can improve the quality of life and health of persons with HIV disease. *Annals of Behavioral Medicine* **18**: 40–8
Smith T W, Gallo L C 2001 Personality traits as risk factors for physical illness. In: Baum A, Revenson T, Singer J (eds.) *Handbook of Health Psychology*. Erlbaum, Mahwah, NJ, pp. 139–72
Stone G, Cohen F, Adler N 1979 *Health Psychology*. Jossey-Bass, San Francisco
Uchino B N, Cacioppo J T, Kiecolt-Glaser J K 1996 The relationship between social support and physiological processes: A review with emphasis on underlying mechanisms and implications for health. *Psychological Bulletin* **119**: 488–531
Weinstein N D, Rothman A J, Sutton S R 1998 Stage theories of health behavior: Conceptual and methodological issues. *Health Psychology* **17**: 290–9

T. W. Smith

Health Research, Qualitative

Qualitative health research is social research that uses qualitative methods to study problems in health, illness, and medicine. These methods include participant observation, in-depth interviewing, document analysis, and focus groups. The data collected are qualitative in nature including fieldnotes, open-ended interviews, narratives, text, organizational documents, and news and media excerpts. The data are usually in text form, but also may be visual or pictorial, and are analyzed by a variety of strategies (see Miles and Huberman 1994, Strauss 1987).

Qualitative health research is distinguished from quantitative and clinical research. Quantitative health research typically utilizes some kind of numerical data (e.g., from censuses, surveys, or experiments) and relies on statistical techniques for analysis. Clinical research is usually based on patient populations and is conducted as an adjunct to the delivery of medical care, focusing on clinical (e.g., treatment) problems. Some clinical research may use qualitative data, but typically qualitative health research does not focus on clinical or medically defined problems.

1. *Background and Context*

The roots of modern qualitative research can be found in early twentieth century social science (Vidich and Lyman 1994). While anthropologists had used ethnographic field work to study nonliterate and tribal cultures, sociologists at the University of Chicago in the 1920s pioneered using participant observation studies for examining aspects of urban and industrial societies (Bulmer 1984). Very little of this early fieldwork focused on health or medicine, although the classic study of *Middletown* did devote a chapter to it (Lynd and Lynd [1929] 1956).

During the 1950s and 1960s a new generation of University of Chicago trained sociologists including Everett Hughes, Anselm Strauss, Howard S. Becker, Eliot Freidson, Julius Roth, and Fred Davis, among others, laid the groundwork for the modern development of qualitative health research (Charmaz and Olesen 1997). This research was characterized by ethnographic approaches and the use of qualitative data, especially from interviews and observation, focusing on the subjective and interactional aspects of social life.

Early studies focused on medical education, organization of health care, medical institutions (especially hospitals), doctor–patient interaction, and death and dying. Several of these studies, especially by Becker et al. (1961) and Glaser and Strauss (1967) contributed significantly to the development of more rigorous and systematic data gathering and analytic methods.

Others like Goffman's (1961) *Asylums* became classics in the field.

In the latter years of the twentieth century qualitative health research prospered and continued to evolve. While social scientific research on health and medicine was dominated by quantitative studies, qualitative studies first found an intellectual niche and later flourished.

While sociologists pioneered qualitative health research, other disciplines adapted it to their own research needs. Anthropologists, who had long used qualitative methods, began to focus more research on modern societies and conducted studies on ethnicity, illness, and health care in community settings (Good and Good 2000). Qualitative studies permeated nursing research, in part because of the development of a nursing Ph.D. program at the University of California, San Francisco by Anselm Strauss and his colleagues. Feminist researchers have used qualitative methods for research on gender and health, reproductive issues, care-giving and care-receiving (Olesen 1994). Beginning in the late 1960s, British academics adopted qualitative methods to conduct research in health and medicine. In ensuing years it was adopted by researchers in France, Australia, Germany and other countries (Anderson and Bury 1988).

Qualitative methods are now well established in social scientific health research. Any number of useful primers for qualitative research now exist (e.g., Lofland and Lofland 1995), as well as compendiums for more advanced researchers (e.g., Denzin and Lincoln 1994, Hammersley and Atkinson 1995). Several journals feature qualitative studies, especially *Sociology of Health and Illness*, *Medical Anthropology Quarterly*, *Health*, and *Qualitative Health Research*. Dozens of health-related books based on qualitative data from researchers in several fields appear each year.

2. *Types of Qualitative Methods*

2.1 *Participant Observation*

Participant observation is the classic form of qualitative research. In this mode, the researchers collect observational data by observation in a situation or organization, after gaining access to their research sites. The balance between participation and observation varies, depending on the researcher and the site. The goal of the research is to understand the situation from the perspective of the participants and gaining this type of social scientific empathy can be a lengthy process. Participant-observers take fieldnotes based on their observations, which become the core data and the basis of their analysis and report. Observation is often combined with informal and formal in-depth interviewing to broaden data sources.

Participant observation is very labor- and time-intensive, but allows a close-up view of social life. In health research it has been used with great benefit for the study of medical education (Becker et al. 1961), the delivery of health care (Zussman 1992), uses of medical technology (Timmermans 1999), professional–patient interaction (Waitzkin 1991), patient and self-help groups, alternative health care, and interaction among professionals (Bosk 1979).

2.2 *In-depth Interviewing*

In-depth interviewing is a central method of qualitative data collection, characterized by the researcher asking respondents a series of open-ended questions or raising topics for discussion. This is contrasted to survey interviewing, where researchers use mostly fixed-response questions (e.g., yes or no, multiple choice, Likert scale questions). Many researchers see in-depth interviewing as a 'guided conversation,' lasting perhaps from 15 minutes to many hours. While some use a fixed list of questions, others prefer to have an interview guide of topics, and then allow respondents to present their view or 'tell their story.' In either case, the goal here is to obtain narrative or textual data in the respondent's own words. Some studies are longitudinal and conduct several interviews with the same subjects of a period of time. Researchers often use tape recorders or similar devices and then transcribe the tapes into written textual data.

Researchers need to be careful to construct their interview guides so as not to unnecessarily bias the data. Questions should be asked in as neutral a form as possible and interviews should take place in a situation that is comfortable for the researcher and the respondent. One of the benefits of this kind of interviewing is that it allows the researcher to be sensitive to respondents and adjust the interview to new paths in the research inquiry.

The number of respondents in such studies varies greatly. Some studies use just a few selected respondents and interview them in great detail and depth. At the other end of the spectrum, some studies use a randomly selected sample of hundreds of respondents, each of whom is interviewed (in this case, usually by members of a research team). Most frequently, in-depth interview samples range from 20 to 100 respondents, depending on research question and the researchers' resources.

In-depth interviews have been used to study experience of illness and disability; patient and/or family view of treatment or care giving; health care providers' views of their work; the development and dilemmas in health careers; biographical perspectives on health and illness; and medical issues like adherence, seeking care, use of services, and quality of life (e.g., Conrad 1985, Kleinman 1988, Schneider and Conrad 1983).

2.3 *Focus Groups*

Since the early 1980s, focus groups have come into fashion as a method for collecting qualitative data,

adapted from marketing research. Focus groups are a collection of people who meet together with the investigator to discuss some topic or issue. Depending on the research topic, the members of the focus group may be already acquainted or unknown to each other, may or may not have direct experience with the topic at hand, and the groups may be of different sizes and composition. Typically, focus groups are small (five to 10 members) and meet from a single to a few occasions with the researcher. The researcher proposes the question or topic and then lets the group discuss it; depending on varying strategies, the leader may guide groups' discussion toward the topic of interest or let the group take its own course. The conversation is usually tape-recorded and then transcribed.

One advantage of focus groups is that they are a very economical method of collecting data. Several well-designed focus groups meeting for a couple of hours each can produce interactive data representing dozens of people. Compared to interviews, focus groups allow for considerable interaction among participants and thus resemble a multifaceted conversation. However, focus groups are artificial occasions, created by the researcher, and thus may not very well reflect everyday life, tend to remain on a superficial level, and are constrained by the group experience.

Focus groups are useful for discovering what a range of views are on different issues, what people's reactions are to certain topics or stimuli (e.g., type of drug or treatment); how people discuss health issues collectively; and to get a sense of the experience related to an illness, care-giving situation, or some other health-related matter (e.g., Kerr et al. 1998).

2.4 Documents

Some qualitative health data comes from sources other than observation, interviews, and focus groups. Most of these data come in the form of some type of document. This can range from governmental or 'official' documents, to organizational memos, to patient records, to news media stories, to personal letters and diaries. They can include pictorial or visual documents but are usually in some form of text. Such documents are ubiquitous to our modern world and are common in health settings and related to health issues.

Document analysis may be used as an adjunct to other methods such as participant-observation or may be the central data of a research inquiry. In some studies the data are collected very systematically, with a rigorous sampling strategy and a tight coding scheme; in other studies, often with limited and scarce data, the data collection is more of a convenience sample meant to illustrate some empirical or conceptual point. In most cases, for textual materials, one of several forms of qualitative content analysis is used to analyze the data. Often the documents can be analyzed with the same strategies used for other qualitative data.

One advantage of documents is that they already exist and are available if one can get access to them. However, documents were created for other purposes than research and thus may not contain all the information an investigator requires or may represent more the view of the producer than reflect the reality of the situation.

Document analysis has been used to examine the emergence of medical conceptions in the professional literature, the changes in official statements about an illness or treatment, the news media's coverage or depiction of a health problem, an organization's statements about operation or purpose, or the content and presentation in patient records.

3. Qualitative Analysis

Qualitative analysis tends to be inductive rather than deductive. Instead of working from specific hypotheses and with predetermined codes, the researcher codes the data while or after it has been collected. Coding schemes can often be very complicated but are key to the data analysis (see Lofland and Lofland 1995, Strauss 1987).

Once the data are coded, researchers use qualitative analytic strategies such as 'grounded theory' (Glaser and Strauss 1967) to develop a substantive and a conceptual analysis. Grounded theory is a strategy that uses elaborate coding and continuous comparison to develop the analysis, allowing the categories to 'emerge' from close reading and analysis. An important aspect of qualitative analysis is the development of concepts, abstractions from the data that can be used to generalize the findings to other situations or cases. Examples are Glaser and Strauss's (1964) concept of 'awareness contexts' which they developed to understand the different situations and stances around information about dying, and Conrad's (1985) 'self-regulation' as an alternative to the commonly held notion of 'noncompliance.' Qualitative studies rarely claim to develop entire theories but at their best create insightful and useful conceptual understanding of aspects of the empirical world.

The best qualitative health studies enlighten the substantive issue at hand by the application or development of a conceptual analysis. If qualitative studies are to be case studies, the researchers must ask, what is it a case of? This will necessitate a conceptual or theoretical abstraction from the particulars of the case.

4. Contributions of Qualitative Health Research

Qualitative health studies have made important contributions to our knowledge. By insisting that researchers stay close to the data in their analysis and by

emphasizing first-hand data collection, qualitative studies have enlightened areas around health and health care. Qualitative studies are well suited to examining the social context of health care, organizational culture (Anspach 1993), the impacts of technology on patients (Timmermans 1999, Baszanger 1998), bioethics in medical care (Bosk 2000), the experience of illness (Conrad 1987, Anderson and Bury 1988), the functioning of medical organizations (Zussman 1992), and the process of medical education (see Hafferty 2000). Many important studies on patient experience and professional–patient interaction are qualitative (see Conrad 1987, Waitzkin 2000).

For many health issues, qualitative studies provide a unique 'insider's approach' to studying a problem. In their review of ethnographic studies in medical sociology, Charmaz and Olesen (1997) conclude that qualitative studies have contributed in six ways: (a) maintaining a critical focus on the institution of medicine; (b) bringing subjective experience into medical sociology; (c) providing empirical corrections to public, professional, and sociological assumptions about medical care and health and illness; (d) challenging accepted ideas in subfields of medical sociology; (e) deepening other social scientific analyses; and (f) extending the range of medical sociology.

Qualitative research usually stands on its own but can also be used in conjunction with quantitative research; typically, either the quantitative or qualitative research predominates. For example, quantitative research can be used to set the context of a qualitative study or qualitative data can be used as cases to illustrate quantitative findings. True integration of qualitative and quantitative data is achievable although difficult (e.g., Broadhead and Heckathorn 1994).

Qualitative health research is most often of the ethnographic or analytically descriptive form described here. But other varieties have developed, including conversational analysis (e.g., Frankel 1984), narrative research (Kleinman 1988) and biographical reflection (Zola 1982). While these are not discussed separately here, they have also made important contributions to health research.

5. *Critiques of Qualitative Health Research*

There are several critiques of general qualitative research that also apply to qualitative health research. Since most such studies are of specific sites or nonrandom samples, the question is often raised about how generalizable findings are from the case to the larger population. Generalizablity in this sense of sampling is a concept derived from quantitative studies that is often misapplied to qualitative studies. Most qualitative studies attempt to examine social processes that can be revealed by a close investigation and analysis, and it is the processes, not the site or sample, that may be generalizable. Another critique, which frequently is a self-critique, is that qualitative researchers rarely tend to study 'up.' That is, investigators tend to study on their own social level or below, but rarely study the 'seats of power.' In health, that means we have many studies of patients, doctors, and health service organizations, but few studies of health insurance organizations, the medical industries, or the high level policy makers. This is in part a problem of access, and, it can be said, true of quantitative social scientific studies as well.

In the late 1980s, a number of scholars advanced the so-called postmodern critique of ethnography, calling into question the authority of the research and researcher and creating what some have called 'the crisis of representation' of qualitative research (see Denzin and Lincoln 1994). These critics argued that ethnographies are stories (even 'fictions') and narratives rather than authoritative representations of reality. The postmodernist emphasis on the validity of multiple perspectives rather than privileging the authors' authority and singular truth raises significant questions about data collection, presentation, and validity of research. Some critics even suggest social scientists should abandon the pretense that ethnography adequately represents the 'real world.' Postmodernism's disenchantment with 'realism' is not evident in most qualitative health research. Most of the research data are still realist tales (Van Maanen 1988), albeit more reflexive and circumscript than the qualitative research of earlier decades. Qualitative research remains accepted as a social scientific method of investigating social reality.

Having survived challenges from quantitatively oriented disciplines and criticisms from postmodernists, social scientific qualitative health research is by now well-established. New data sources like the Internet, an increasing number of computer-based analytic packages, and adoption of the method by more academic disciplines should allow for continued expansion and development of qualitative health research in the twenty-first century.

See also: Archival Methods; Biographical Methodology: Psychological Perspectives; Case Study: Methods and Analysis; Clinical Assessment: Interview Methods; Clinical Treatment Outcome Research: Control and Comparison Groups; Ethical Dilemmas: Research and Treatment Priorities; Experimenter and Subject Artifacts: Methodology; Field Observational Research in Anthropology and Sociology; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Laboratory Experiment: Methodology; Narratives and Accounts, in the Social and Behavioral Sciences; Objectivity of Research: Ethical Aspects; Person-centered Research; Qualitative Methods, History of; Research Subjects, Informed and Implied Consent of; Single-subject Designs: Methodology; Social Survey, History of

Bibliography

Anderson R, Bury M (eds.) 1988 *Living with Chronic Illness: The Experience of Patients and Their Families*. Unwin Hyman, London

Anspach R 1993 *Deciding Who Lives*. University of California Press, Berkeley, CA

Baszanger I 1998 *Inventing Pain Medicine: From the Laboratory to the Clinic*. Rutgers University Press, New Brunswick, NJ

Becker H S, Geer B, Hughes E C, Strauss A 1961 *Boys in White*. University of Chicago Press, Chicago

Bosk C L 1979 *Forgive and Remember: Managing Medical Error*. University of Chicago Press, Chicago

Bosk C L 2000 The sociological imagination and bioethics. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ, pp. 398–410

Broadhead R S, Heckathorn D D 1994 AIDS prevention outreach among injection drug users: Agency problems and new approaches. *Social Problems* **41**: 473–95

Bulmer M 1984 *The Chicago School of Sociology*. University of Chicago Press, Chicago

Charmaz K, Olesen V 1997 Ethnographic research in medical sociology: Its foci and distinctive contributions. *Sociological Methods and Research* **25**: 452–94

Conrad P 1985 The meaning of medications: Another look at compliance. *Social Science and Medicine* **20**: 29–37

Conrad P 1987 The experience of illness: Recent and new directions. In: Roth J, Conrad P (eds.) *Research in the Sociology of Health Care*. JAI Press, Greenwich, CT, Vol. 6, pp. 1–32

Denzin N K, Lincoln Y S (eds.) 1994 *Handbook of Qualitative Research*. Sage Publications, Thousand Oaks, CA

Frankel R M 1984 Physicians and patients in social interaction: Medical encounters as a discursive process. *Discourse Processes* **7**: 103–5

Glaser B G, Strauss A L 1964 Awareness contexts and social interaction. *American Sociological Review* **29**: 269–79

Glaser B G, Strauss A L 1967 *The Discovery of Grounded Theory*. Aldine, Chicago

Goffman E 1961 *Asylums*. Doubleday Anchor, New York

Good M D, Good B 2000 'Parallel sisters': Medical anthropology and medical sociology. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ, pp. 377–88

Hafferty F 2000 Reconfiguring the sociology of medical education: Emerging topics and pressing issues. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ, pp. 238–57

Hammersley M, Atkinson P 1995 *Ethnography: Principles and Practice*, 2nd edn. Routledge, London

Kerr A, Cunningham-Burley S, Amos A 1998 Drawing the line: An analysis of lay people's discussions of the new genetics. *Public Understanding of Science* **7**: 113–33

Kleinman A 1988 *Illness Narratives*. Basic Books, New York

Lofland L, Lofland J 1995 *Analyzing Social Settings*. Wadsworth, Belmont, CA

Lynd R S, Lynd H M 1956 [1929] *Middletown*. Harcourt Brace, New York

Miles M B, Huberman A M 1994 *Qualitative Data Analysis: A Sourcebook of New Methods*, 2nd edn. Sage, Thousand Oaks, CA

Olesen V 1994 Problematic bodies: Past, present and future. *Symbolic Interaction* **17**: 231–7

Schneider J W, Conrad P 1983 *Having Epilipsy: The Experience and Control of Illness*. Temple University Press, Philadelphia, PA

Strauss A L 1987 *Qualitative Analysis*. Jossey-Bass, San Francisco

Timmermans S 1999 *Sudden Death and the Myth of CPR*. Temple University Press, Philadelphia

Van Maanan J 1988 *Tales from the Field*. University of Chicago Press, Chicago

Vidich A J, Lyman S F 1994 Qualitative methods: Their history in sociology and anthropology. In: Denzin N K, Lincoln Y S (eds.) *Handbook of Qualitative Sociology and Research*. Sage, Thousand Oaks, CA, pp. 23–59

Waitzkin H 1991 *The Politics of Medical Encounters*. Yale University Press, New Haven

Waitzkin H 2000 Changing patient–physician relationships in the changing health-policy environment. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ, pp. 271–83

Zola I K 1982 *Missing Pieces*. Temple University Press, Philadelphia, PA

Zussman R 1992 *Intensive Care*. University of Chicago Press, Chicago

P. Conrad

Health Risk Appraisal and Optimistic Bias

1. Importance of Health Risk Perceptions

Perceptions of health risks influence the actions people take to prevent illness and injury, the decisions they make to seek medical care, their choices among treatment options, and their adherence to treatment requirements. Public controversies over the risks that may be created by government and corporate actions are also familiar. During these controversies, scientists and members of the public frequently disagree, demonstrating that laypeople do not passively accept the risk judgments of experts.

These various reasons suggest that it is important to understand how risk perceptions are formed, how accurate they are, and how they influence decisions and behavior. Furthermore, as research continues to identify new risks and 'risk factors' (attributes of individuals or environments that alter a person's likelihood of harm), the need grows to explain this information clearly so that individuals can make better decisions about their health. Such education about risk is the goal of the specialty called 'risk communication.'

2. Defining 'Risk' and 'Risk Perceptions'

Before pursuing this topic it should be mentioned that the term 'risk' is used in several different ways. Sometimes 'risk' is used synonymously with the word 'hazard', denoting a specific undesirable outcome. In this sense, cancer, the common cold, and automobile accidents are all 'risks.' 'Risk' is also used as a synonym

for 'probability', referring to the probability (or likelihood) of the undesirable outcome. Thus, one can speak of the risk that a person will develop cancer. 'Risk' is frequently used in still another way, as a synonym for 'danger' or 'threat' (for example, 'How risky is smoking?'). The term 'risk perceptions', then, is a general term that refers to the various beliefs that people hold about health and safety hazards.

3. How do People Think about Risk?

3.1 Dimensions of Hazards

Hazards have many dimensions. According to classical theories of decision-making, the key attributes of a hazard are the probability that it will occur and the severity of the undesirable consequences should it occur.

Pioneering research by Lichtenstein et al. (1978) demonstrated that laypeople are also influenced by hazard attributes that are not so obviously important. When scientists are asked about the riskiness of different illnesses and technologies, their judgments strongly reflect the number of people who die annually from each one, thus reflecting both seriousness (deaths) and likelihood (number of events in a particular population). Lay judgments are only moderately related to the number of fatalities. In addition, laypeople respond to whether the risk is familiar or unfamiliar, manufactured or natural, voluntary or involuntary, and other factors. Summarizing the many attributes that influence risk perceptions, researchers have suggested that lay judgments of risk reflect two main dimensions, the extent to which the hazard is perceived to be known or unknown and the extent to which it is dreaded or not (Slovic 1987).

Leventhal and colleagues suggest that each person copes with illness threats in accord with their own 'common sense' model of that illness (see *Illness Behavior and Care Seeking*). These models involve five attributes: the identity of the risk, its timecourse (e.g., chronic vs. acute), its consequences, its causes, and its potential cure or control (Leventhal et al. 1992).

The preceding perspectives emphasize various hazard attributes. Another approach emphasizes the interconnections among the hazard properties. Beliefs about the cause of an illness, its prevalence, routes of exposure, and personal risk are all important, but it is also important to know how they are related to one another. Misunderstandings about the causes and prevalence of hazard, for example, may result in misperceptions about the likelihood of victimization. These different aspects of knowledge and belief are viewed as a 'mental model' that describes both the separate pieces and the connections between these pieces (Bostrom et al. 1992).

3.2 Hazard Probabilities

Much attention has been given to how people think about hazard probabilities. (Apparently researchers think that hazard severity is a simpler issue, for this topic has received little attention.) Although proportions (e.g., '0.25'), percentages (e.g., '25 percent'), and odds (e.g., '1 in 4') are the units used by scientists to measure probability, there is abundant evidence that laypeople have great difficulty understanding and using such information. For instance, people sometimes give a higher probability to the likelihood of dying of breast cancer than to the likelihood of developing breast cancer. Thus, the ability to recite the correct percentage, rate, or odds should not be used as the criterion to decide whether people truly understand the probability. Except for percentages given in weather forecasts, the public is rarely exposed to probability statistics, so it is not surprising that many people find numerical information about hazard probabilities unhelpful.

Some evidence indicates that people find verbal designations of probability (e.g., 'small chance of happening') to be easier to use and to better reflect their own views than numerical risk measures (Diefenbach and Weinstein 1993), but people vary greatly in the numbers they associate with particular verbal labels, making the conversion of numbers to words highly subjective. Expressions of probability in terms of odds (e.g., 1 in 50 chance) are especially disliked by laypeople.

Because it seems impossible to navigate an uncertain world without forming estimates of the likelihood that particular events will occur, we tend to assume that people must form and encode judgments of hazard probabilities. In fact, it is possible that people rarely do encode such information. Instead they may generate risk judgments from past experiences and from their own implicit theories as needed. It is also possible that people store in memory only very crude probability information (e.g., no chance, small chance, large chance, certain) and never use the finer distinctions that concern scientists and policy makers.

Because social comparison is a central feature of human existence, in everyday life people may find it more natural to make decisions on the basis of risk comparisons than on beliefs about absolute probability (the likelihood that an event will occur). Thus, a woman's interest in preventing heart disease may depend less on her beliefs about her absolute likelihood of a heart attack (e.g., 40 percent chance) than on her belief that her risk is greater than the risk of her friends. She may also compare one risk (e.g., heart disease) with another (e.g., cancer).

4. Goals of Risk Perceptions

Obviously, accurate risk perceptions can help people make appropriate decisions about actions to avoid

illness or injury. However, accuracy is only one of the goals guiding people as they formulate their risk judgments (Kunda 1990). Risk perception is not an unbiased appraisal of information, but an attempt to seek the most comforting view of one's personal vulnerability that fits within the bounds of the evidence. Beliefs about our vulnerability to harm can lead to positive or negative emotions (pleasure about successfully avoiding harm; anxiety about potential illness or disability). Perceptions of vulnerability can also affect our feelings of self-esteem (see *Vulnerability and Perceived Susceptibility, Psychology of*). Since we want to believe that we are intelligent, competent, and foresighted, to admit that our actions place us at greater risk than our peers would be a threat to our self-esteem. Thus, minimizing negative affect and maximizing self-esteem are also goals of risk perceptions, and these goals may be at variance with the goal to be accurate.

5. *Optimistic Bias*

This bias in risk perceptions is revealed especially clearly when people are asked to compare their risks with the risk of the rest of the population. Whatever they may believe about the risk of others, they tend to believe that their own risk is less (Weinstein 1987). For example, when people are asked how their risk of a particular illness compares with the risk of others of the same age and sex, many claim to have below average risk and few admit to having above average risk. Across a variety of hazards, it is typical to find 40–70 percent of a group asserting below average risk; another 30–50 percent saying that their risk is average; and less than 10 percent acknowledging that their risk is above average. Obviously, this distribution cannot be correct. The 'average' person has, by definition, an 'average' risk. So when the people who claim below average risk greatly outnumber those who say their risk is higher than average, implying that the average person is below average in risk, something must be wrong with their risk judgments.

Optimistic bias is not just denial, a blanket defense against anxiety. If it were, we would expect to find the greatest optimism among hazards, such as cancer, that generate the most fear. The amount of optimistic bias (that is, the extent to which people claiming below average risk outnumber those who say their risk is above average) does vary from hazard to hazard (Weinstein 1987), but it is not related to the severity of the hazard. Instead it is greatest for events that are believed to be preventable by individual action and events with which people have little personal experience. Bias is especially great if people think that vulnerability appears early, leading them to conclude that if they have not experienced the problem so far, they are free from risk.

The belief that we are less at risk than others is apparently reassuring, since people resist attempts to correct their misperceptions. Most interventions designed to eliminate bias in comparative risk judgments—for example, having people think of the ways in which the negative event could happen or think of the precautions that they have not taken—have proved unsuccessful (Weinstein and Klein 1995).

It may appear obvious that underestimating risk will expose people to unnecessary harm. However, a different view of this issue, known by the label 'positive illusions,' suggests that optimistic biases are really adaptive (Taylor and Brown, 1988). It points out that people who are overly optimistic about their ability to fight off the AIDS virus, for example, may be encouraged to maintain a healthy lifestyle and to bear with burdensome treatments. Thus, even if the perception of resistance is false, it may support other actions that are themselves beneficial. Although definitive evidence is lacking, it seems likely that underestimations of personal vulnerability decrease interest in prevention, especially when people see themselves as members of a group that does not need to be concerned about the risk, but optimism about surviving a life-threatening illness can help patients avoid depression and hopelessness (Wood et al. 1985).

6. *Effects of Risk Perceptions on Health-related Behaviors*

Perceived probability and/or perceived severity occupy an important role in nearly all theories that are used to explain the health-related behavior of individuals (see *Health Behavior: Psychosocial Theories*). The literature on health behavior generally supports the idea that increases in perceived risk increase precautionary behavior (Janz and Becker 1984), but there are also numerous instances in which risk perceptions and behavior turn out to be unrelated.

This should not be surprising. There are many factors other than risk perceptions that influence behavior, and people who think that they have a high risk may nevertheless doubt their ability to carry out precautions or the efficacy of the precautions. In addition, most health behavior studies have used correlational research designs, which are problematic for examining risk perception–risk behavior linkages. Finally, as mentioned earlier, experimental interventions may fail to alter peoples' perceptions of their invulnerability.

The link between perceptions and action may depend on the type of action. Precautions that prevent a health problem from occurring (such as vaccinations) should reduce both perceived risk and worry about that risk. In contrast, being tested for the presence of an existing health problem (as in cancer screening) offers the immediate potential of extremely distressing information, especially if the problem is

difficult to treat, so high perceived risk may be associated with high fear. The links between risk perceptions and different types of action still need to be elucidated.

See also: Attitudes and Behavior; Depression, Hopelessness, Optimism, and Health; HIV Risk Interventions; Personality and Risk Taking; Sexual Risk Behaviors

Bibliography

Bostrom A, Fischhoff B, Morgan M G 1992 Characterizing mental models of hazardous processes: A methodology and an application to radon. *Journal of Social Issues* **48**: 85–100

Diefenbach M A, Weinstein N D, Oreilly J 1993 Scales for assessing perceptions of health-hazard susceptibility. *Health Education Research* **8**: 181–92

Janz N K, Becker M H 1984 The health belief model: A decade later. *Health Education Quarterly* **11**: 1–47

Kunda Z 1990 The case for motivated reasoning. *Psychological Bulletin* **108**: 480–98

Leventhal H, Diefenbach M A, Leventhal E A 1992 Illness cognition: Using common sense to understand treatment adherence and affect cognition interactions. *Cognitive Therapy and Research* **16**: 143–63

Lichtenstein S, Slovic P, Fischhoff B, Layman M, Combs B 1978 Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory* **4**: 551–78

Slovic P 1987 Perception of risk. *Science* **236**: 280–5

Taylor S E, Brown J D 1988 Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin* **103**: 193–210

Weinstein N D 1987 Unrealistic optimism about susceptibility to health problems: Conclusions from a community-wide sample. *Journal of Behavioral Medicine* **10**: 481–500

Weinstein N D, Klein W M 1995 Resistance of personal risk perceptions to debiasing interventions. *Health Psychology* **14**: 132–40

Wood J V, Taylor S E, Lichtman R R 1985 Social comparison in adjustment to breast cancer. *Journal of Personality and Social Psychology* **49**: 1169–83

N. D. Weinstein

Health: Self-regulation

Many health conditions are caused by risk behaviors such as problem drinking, substance use, smoking, reckless driving, overeating, or unprotected sexual intercourse. Fortunately, human beings have, in principle, control over their conduct. Health-compromising behaviors can be eliminated by self-regulatory efforts, and health-enhancing behaviors such as physical exercise, weight control, preventive nutrition, dental hygiene, condom use, or accident-preventive actions can be adopted instead. Health self-regulation refers to the motivational, volitional, and actional process of removing such health-compromising behaviors in favor of adopting and maintaining health-enhancing behaviors (see *Health Behaviors*). Health self-regulation encompasses a broad range of social cognitions and behaviors that are subject to developmental changes. This article, however, is limited to models that address the current debate within the discipline of health psychology, excluding the larger fields of social and developmental psychology (Baltes 1997, Baumeister and Heatherton 1996, Brandtstädter and Greve 1994, Carver and Scheier 1998).

1. Health Motivation Factors

Before changing their habits, people need to become motivated. This is seen as a process towards an explicit intention (e.g., 'I intend to quit smoking this weekend'). Three variables are considered to play a major role in this process: (a) risk perception, (b) outcome expectancies, and (c) perceived self-efficacy (Bandura 1997, Carver and Scheier 1998, Conner and Norman 1996).

Perceiving a health threat seems to be the most obvious prerequisite for the motivation to remove a risk behavior. If one is not aware at all of the risky nature of one's actions, motivation cannot emerge. Usually, people are aware of some level of risk although the accuracy of their perception may be biased. When it comes to a comparison with similar others, one's view of the risk is somewhat distorted ('Compared to others of my age and sex my risk of getting lung cancer is low/medium/high'). This has become known as the 'optimistic bias' (see *Health Risk Appraisal and Optimistic Bias*). Nevertheless, persons also acknowledge some degree of risk when confronted with objective data. There is a realistic component that keeps the positive illusions in leash. For example, smokers not only know that smoking can cause adverse health in others, they also perceive that they are more at risk themselves for lung cancer and other diseases than nonsmokers.

Risk perception has two aspects: perceived severity of a health condition and personal vulnerability toward it (see *Vulnerability and Perceived Susceptibility, Psychology of*). The first refers to the amount of harm that could occur, and the second pertains to the subjective probability that one could fall victim to that condition. Thus, it has been recommended that people should be informed about the existence of a health risk, and moreover, that they should imagine themselves as possible victims if they do not take the necessary precautions.

Fear arousal is an emotion that can accompany risk perception. Scaring people into health behaviors, however, has not been shown to be effective. In general, initial risk perception seems to be advan-

tageous to put people on track for developing a motivation to change, but later on other variables are more influential in the self-regulation process, such as outcome expectancies, self-efficacy, and behavioral intentions.

People not only need to be aware of the existence of a health threat, they also need to know how to regulate their behavior by understanding the contingencies between their actions and subsequent outcomes. These *outcome expectancies* are influential beliefs in the motivation to change (Bandura 1997). A smoker may find more good reasons to quit than good reasons to continue smoking ('If I quit smoking then I will save money'). This imbalance does not lead directly to action but it can help to come up with the intention to quit. The pros and cons represent positive and negative outcome expectancies that are typical in rational decision making. However, such contingencies between actions and outcomes need not be explicitly worded and evaluated, they can also be rather diffuse mental representations, loaded with emotions. Outcome expectancies can also be understood as methods, or means–ends relationships, indicating that people know proper strategies to produce desired effects.

Moreover, the efficacy of such a method has to be distinguished from the belief in one's personal efficacy to apply the method. *Perceived self-efficacy* (Bandura 1997) portrays individuals' beliefs in their capabilities to exercise control over challenging demands and over their own functioning (see *Self-efficacy and Health*). Self-regulatory strength (Baumeister and Heatherton 1996) is a similar notion of one's capacity or stable resource to cope with demands. These beliefs are critical in approaching novel or difficult situations or in adopting a strenuous self-regimen. People make an internal causal attribution when imagining their behavior (e.g., 'I am certain that I can quit smoking even if my friend continues to smoke'). Such optimistic self-beliefs influence the goals people set for themselves, what courses of action they choose to pursue, how much effort they invest in given endeavors, and how long they persevere in the face of barriers and setbacks. Some people harbor self-doubts and cannot motivate themselves. They see little point in even setting a goal if they believe they do not have what it takes to succeed. Thus, the intention to change a habit that affects health is to some degree dependent on a firm belief in one's capability to exercise control over that habit (see *Control Beliefs: Health Perspectives*).

2. Self-regulatory Efforts

Changing one's health behavior is considered to be a difficult self-regulation process. After people have become engaged in a goal they need to prepare action and, later, maintain the changes in the face of obstacles and failures. Thus, goal setting and goal pursuit can be understood as two distinct processes, the latter of which requires a great deal of self-regulatory effort.

The process of self-regulation can be subdivided into sequences such as planning, initiation, maintenance, relapse management, and disengagement although these are not perfectly distinct categories. Plans specify the when, where, and how of a desired action, and carry the structure of 'When situation S arises, I will perform response R' (Gollwitzer 1999). People take *initiative* when the critical situation arises and give it a try. This requires firm self-beliefs to be capable to perform the action. People who do not hold such beliefs see little point in even trying.

The adoption and *maintenance* of the health behavior is not achieved through an act of will but involves the development of self-regulatory skills and strategies. This embraces various means to influence one's own motivation and behaviors such as the setting of attainable, proximal subgoals, creating incentives, drawing from an array of coping options, and mobilizing social support. Action control processes include focusing one's attention on the task, while avoiding attention to distractors, resisting temptations, and managing unpleasant emotions. Perceived self-efficacy is required to overcome barriers and to stimulate self-motivation repeatedly.

Competent *relapse management* is needed to recover from setbacks. Some people rapidly abandon their newly adopted behavior when they fail to get quick results. When entering high-risk situations (e.g., a bar where others smoke), they cannot resist due to a lack of self-efficacy. The competence to recover is different from the competence to initiate an action (Marlatt et al. 1995). Restoration, harm reduction, and renewal of motivation are serviceable strategies within the context of health self-regulation.

Disengagement from the goal can be evidence for lack of persistence and, thus, indication of self-regulatory failure (Carver and Scheier 1998). But in the case of repeated failures, disengagement or scaling back the goal might become an option which can be adaptive, depending on the circumstances. For example, if the goal was set too high or if the situation has changed and become more difficult than before, it is often not worthwhile to continue the struggle. In the case of health-compromising behaviors, giving up is not a tenable option. Better self-regulatory skills have to be developed and unique approaches to the problem need to be taken. The experience of failure can be a useful learning experience to build up more competence, on the condition that the individual makes a beneficial causal attribution of the episode, and practices constructive self-talk to renew the motivation.

3. Models of Health Behavior Change

In health psychology (see *Health Psychology*), models are built to describe, explain, and predict how people become motivated to change their risk behaviors, and how they become encouraged to adopt and main-

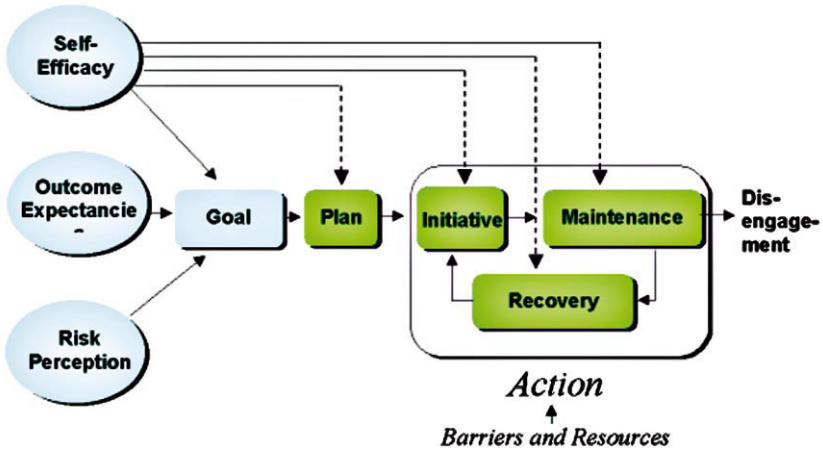

Figure 1
A social-cognitive process model of health self-regulation (Schwarzer 1992, 1999)

tain health actions (see *Health Behavior: Psychosocial Theories*).

3.1 The Transtheoretical Model of Behavior Change

Stage theorists have made an attempt to consider process characteristics by including the postintentional phase. The *Transtheoretical Model of Behavior Change* (DiClemente and Prochaska 1982), for example, has become the most popular stage model. Its main feature is the implication that different types of cognitions may be important at different stages of the health behavior change process. It includes five discrete stages of health behavior change that are defined in terms of one's past behavior and future plans (precontemplation, contemplation, preparation, action, maintenance). For example, at the precontemplation stage, a drinker does not intend to stop consuming alcohol in the future. At the contemplation stage, a drinker thinks about quitting sometime within the next six months, but does not make any specific plans for behavior change. At the preparation stage, the drinker resolves to quit within the next six months. The action stage includes individuals who have taken successful action for any period of time. If this abstinence has lasted for more than six months, the person is categorized as being in the maintenance stage.

The model, which also includes self-efficacy and other relevant features, has been applied successfully to a broad range of health behaviors, with particular success for smoking cessation. However, it has been argued that the notion of stages within this theory might be flawed or circular, in that the stages are not genuinely qualitative, but are rather arbitrary distinctions within a continuous process. Instead, these 'stages' might be better understood as 'process heuristics' to underscore the nature of the entire model. That is, the model can serve as a useful heuristic that describes a health behavior change process which has not been the major focus of health behavior theories so far. In redirecting the attention to a self-regulatory process, the transtheoretical model has served an important purpose for applied settings. The postintentional, preactional phase ('preparation') may be the most challenging stage for researchers and professionals because it is exactly in this phase where an intention is or is not translated into an action—depending on the circumstances. Planning and initiative, but also volatility, hesitation, and procrastination characterize this phase. The identification of stages has been shown to bear important implications for treatment. Interventions have been tailored to communicate stage-matched information to smokers who were in different motivational stages (Dijkstra et al. 1998).

3.2 The Health Action Process Approach

The *Health Action Process Approach* (HAPA model; Schwarzer 1992, 1999) suggests a distinction between preintentional motivation processes and postintentional volition processes. Within these two phases, different patterns of social-cognitive predictors may emerge (see Fig. 1).

In the initial *motivation phase*, a person develops an intention to act. In this phase, *risk perception* is merely seen as a distal antecedent within the motivation phase. It may include not only the perceived severity of possible health threats but also one's personal vulnerability to fall prey to them. Risk perception in itself is insufficient to enable a person to form an intention. Rather, it sets the stage for further elaboration of

thoughts about consequences and competencies. *Outcome expectancies* operate in concert with *perceived self-efficacy*, both of which contribute substantially to the development of an intention, whereas this pattern may change later in the postintentional stage (Schwarzer and Renner 2000).

After goal setting, people enter the postintentional stage or *self-regulation phase* in which they pursue their goal by planning the details, trying to act, investing effort and persistence, possibly fail, and finally recover or disengage. The guiding principle here is that each substage advances from the previous one and is, in addition, facilitated by perceived self-efficacy. Thus, at each point there are two predictors, namely the successful completion of the previous substage in conjunction with an optimistic sense of control over the next one. Risk perception and outcome expectancies are no longer of much influence when goals have been set and when it comes to predicting actual behaviors. Certainly, there may be various influential factors in the self-regulation phase, but according to the present knowledge, stage progression and self-beliefs appear as the most parsimonious and powerful set of predictors (Bandura 1997, Schwarzer and Renner 2000).

This health behavior change model is regarded as a heuristic to better understand the complex mechanisms that operate when people become motivated to change, when they adopt and maintain a habit, and when they attempt to resist temptations and recover from setbacks. It applies to all health-compromising and health-enhancing behaviors and could be extended to behavior change in general (see Conner and Norman 1996 and Weinstein 1993 for other models).

4. Future Directions

Health self-regulation encompasses a broad range of cognitions and behaviors. Research is needed to identify the key variables and the essential mechanisms that operate in this process. Certainly, different perspectives lead to an emphasis on different variables and mechanisms. A social psychological perspective would underscore social influence and the social context of self-regulation. A life-span developmental perspective would accentuate personal history, life events, as well as gains and losses. The present health psychology view pertains to rather short time periods that are typically involved in coping with addictive behaviors and in thorny cessation attempts. In the last two decades, social-cognitive prediction models had been preferred to explain the mechanisms of health self-regulation. These approaches are being replaced now by stage models that imply a cycling and recycling process of individuals across two or more stages. This advance in the field has led to some difficulties and controversies. One refers to the demand for parsimony because variables become easily inflated. For example, the Transtheoretical model includes five stages and 10 processes. According to the present view, it is suggested to reduce the motivation stage to only three predictors and the self-regulation stage to only two components: stage progression and self-beliefs appear as the most parsimonious set of components. Further elaborations, however, are required and could benefit from other fields, in particular from relapse prevention theory (Marlatt et al. 1995) and self-regulation theories (Baltes 1997, Baumeister and Heatherton 1996, Brandtstädter and Greve 1994, Carver and Scheier 1998, Karoly 1993).

See also: Adolescent Health and Health Behaviors; Control Beliefs: Health Perspectives; Health Behavior: Psychosocial Theories; Health Behaviors; Health Behaviors, Assessment of; Self-efficacy; Self-efficacy and Health; Self-regulation in Adulthood; Smoking Prevention and Smoking Cessation

Bibliography

Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of development theory. *American Psychologist* **52**: 366–80

Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman, New York

Baumeister R F, Heatherton T F 1996 Self-regulation failure: An overview. *Psychology Inquiry* **7**(1): 1–15

Brandtstädter J, Greve W 1994 The aging self: Stabilizing and protective processes. *Developmental Review* **14**: 52–80

Carver C S, Scheier M F 1998 *On the Self-regulation of Behavior*. Cambridge University Press, New York

Conner M, Norman P (eds.) 1996 *Predicting Health Behavior: Research and Practice with Social Cognition Models*. Open University Press, Buckingham, UK

DiClemente C C, Prochaska J O 1982 Self-change and therapy change of smoking behavior: A comparison of processes of change in cessation and maintenance. *Addictive Behaviors* **7**: 133–42

Dijkstra A, De Vries H, Roijackers J, van Breukelen G 1998 Tailored interventions to communicate stage-matched information to smokers in different motivational stages. *Journal of Consulting and Clinical Psychology* **66**: 549–57

Gollwitzer P M 1999 Implementation intentions: Strong effects of simple plans. *American Psychologist* **54**: 493–503

Karoly P 1993 Mechanisms of self-regulation: A system view. *Annual Review of Psychology* **44**: 23–52

Marlatt G A, Baer J S, Quigley L A 1995 Self-efficacy and addictive behavior. In: Bandura A (ed.) *Self-efficacy in Changing Societies*. Cambridge University Press, New York, pp. 289–315

Schwarzer R 1992 Self-efficacy in the adoption and maintenance of health behaviors: Theoretical approaches and a new model. In: Schwarzer R (ed.) *Self-efficacy: Thought Control of Action*. Hemisphere, Washington, DC, pp. 217–42

Schwarzer R 1999 Self-regulatory processes in the adoption and maintenance of health behaviors. The role of optimism, goals, and threats. *Journal of Health Psychology* **4**: 115–27

Schwarzer R, Renner B 2000 Social-cognitive predictors of health behavior: Action self-efficacy and coping self-efficacy. *Health Psychology* **19**(5): 487–95
Weinstein N D 1993 Testing four competing theories of health-protective behavior. *Health Psychology* **12**: 324–33

R. Schwarzer

Health Surveys

Information collected by health surveys is used by public health officials to plan, conduct, and evaluate programs; by elected officials to inform policies and legislation; and by researchers to understand the health status of the population, the determinants of health, and the health care system. While such surveys are sponsored and carried out by many private entities, core information is generally obtained through relatively large, multipurpose, government-sponsored data collection systems; these will be the focus here, although most of the issues discussed pertain to all health surveys.

The World Health Organization definition of health as a state of physical, mental, and social well-being and not merely the absence of disease or infirmity provides the starting point for health surveys. The multidimensionality of this definition requires different survey methodologies. Since health is both a biologic and a social construct, health surveys must reflect this duality. Major topics to be covered include aspects of survey design, conceptualization and measurement of health status, determinants of health, summary measures of health, health care utilization and health behaviors. The effect of recent technological advances on data collection methods are discussed as are the growing opportunities for data linkage and the future challenges of providing access to data while protecting confidentiality and privacy.

In the late 1990s, 9.2 percent of the United States population reported they were in fair or poor health; 14 percent of children under 18 years of age had no health insurance coverage; 79 percent of children 19–35 months old were up to date on immunizations; 25 percent of adults were smoking cigarettes; and 24 percent were hypertensive (National Center for Health Statistics, NCHS 1999). Such information is used by public health officials to plan, conduct and evaluate programs; by elected officials to inform policies and legislation; and by researchers to understand the health status of the population, the determinants of health and the health care system. Information on the health of a population can be obtained through a variety of data collection approaches; health surveys represent one mechanism.

While such surveys are sponsored and carried out by many private entities, core information is generally obtained through relatively large, multipurpose, government-sponsored data collection systems; these will be the focus here, although most of the issues discussed pertain to all health surveys. These surveys are used to profile the health of the population, to track demographic shifts and trends, and to monitor change in health and health care. In addition, the data are also used to investigate epidemiologic relationships between risk factors and health outcomes and to provide information for prevention, evaluation, research, planning, and policy. Many nations include the conduct of general purpose health surveys in their statistical systems. Table 1 provides a list of countries and the title and year of their most recent general health survey (CDC, NCHS 1999). It is tempting to directly compare the results of these surveys, but cross-national cultural differences and variation in the nature of health care systems as well as in survey designs must be taken into account when making international comparisons.

1. *Definition and Operationalization of Health*

The World Health Organization definition of health as a state of physical, mental, and social well-being and not merely the absence of disease or infirmity provides the starting point for health surveys (WHO 1948). However, because this definition is broad, complex, and multidimensional, it is difficult for any single survey to collect all the needed information. Depending on the specific objectives of a particular survey, there will be variation in the universes from which the samples are drawn and in the health and health-related information obtained. Since health is both a biologic and a social construct, health surveys must reflect this duality. In fact, these inter-relationships between the biological and social aspects of health are often best approached through health surveys, and advances in technology have made it increasingly possible to incorporate direct measures of biologic aspects of health into more traditional surveys.

2. *General Design Issues*

Health surveys can either obtain information directly from individuals (the subject or an informant) or can rely on administrative records. A wide range of sampling methodologies, sampling frames, and data collection strategies are used depending on survey objectives and the characteristics of the area in which the survey is done. Countries that have universal health care with centralized administrative systems can utilize designs that would be inappropriate in countries with less well-developed or decentralized health care systems. The simplest sampling scheme is a simple random sample of a population obtained from a complete and up-to-date list such as from a census of

Table 1
National population-based health surveys

Country	Title and year
Australia	Australia Health Survey, 1983 National Health Survey, 1995
Austria	Special Health Survey, 1999
Canada	Canada Health Survey, 1978–79 National Population Health Survey, 1996–97
China (People's Republic of)	Health Services Household Survey of China, 1998
Czech Republic	Health Interview Survey, 1996
Denmark	The Danish Health and Morbidity Survey, 1994
Estonia	National Health Interview Survey, 1996
Finland	The Finnish National Health Survey, 1995 Mini-Finland Health Survey, 1977–81
France	France Health Survey, 1991–92 Health and Social Protection Survey, 1988
Germany	National Health Examination and Interview Survey, October 1997–March 1999
Hungary	Health Interview Survey, 1984
Iceland	Multipurpose Survey (surveys on health promotion), 1996
Ireland	The National Health and Lifestyle Survey, 1998
Italy	Survey of Health Conditions of the People and the Use of Health Services, 1994
Japan	Comprehensive Survey of Living Conditions of People on Health and Welfare, Triennial since 1986
Korea (Republic of)	National Health Interview Survey, 1983
Latvia	Living Conditions Survey, 1994
Netherlands	National Health Interview Survey, Continuous since 1981
New Zealand	New Zealand Health Survey, 1996–7
Norway	Health Interview Survey, 1998
Poland	Family Health Survey, 1987 Health Interview Survey, 1996
Portugal	National Health Survey, 1998–9
Singapore	National Health Survey, 1998
Spain	National Health Survey, 1993
Sweden	Survey of Living Conditions, Continuous since 1975
Switzerland	Swiss Health Survey (SOMIPOPS), 1981–2 Swiss Health Survey, 1997
Ukraine	National Health Survey, Annual since 1986
United Kingdom	Health Survey for England, Continuous since 1991 Scottish Health Survey, Triennial since 1995
United States	National Health Interview Survey (NHIS), Continuous since 1957 National Health and Nutrition Examination Survey (NHANES), Continuous since 1999

Source: National Center for Health Statistics, International Health Data Reference 1999.

the population or a centralized health care system. In the case of surveys of persons, if such lists do not exist, area sampling can be used with clustering to reduce costs. In order to be able to report information on subpopulations of interest as defined by geography and/or demographic, socioeconomic or health characteristics, designs often include oversampling of these groups. It is often necessary to incorporate screening into designs to achieve this objective. Information can be obtained either by mail, by phone, by in-person interview, or by some combination of modes depending on the ability of the researchers to contact the population using these means. The latter two modes can also incorporate computer-assisted technologies, and the use of the Internet for the administration of health surveys is being investigated.

In general, obtaining information directly from the survey subject provides more reliable and valid health data. However, this is not universally true and depends on the nature of the information sought and the characteristics of the subject (Moore 1988). For example, while accurate information on risk behaviors among adolescents can probably only be obtained from the adolescent in a private environment, adolescents are not good reporters of health care utilization or household income. Even more problematic is the situation where a health condition does not permit a subject to respond for him- or herself. Since eliminating the subject from the survey would seriously bias the results, a proxy respondent is often used.

While necessary, the use of proxy respondents does introduce a source of error into health surveys, particularly longitudinal surveys. Obtaining information on the relationship of the proxy to the subject as well as the conditions leading to the use of a proxy can reduce this error. In other cases, subject reporting is fraught with a great deal of measurement error. For example, information on the characteristics of an insurance plan is usually better obtained from the insurer rather than the person holding the insurance. Finally, reliable assessment of specific types of morbidity, risk factors, and mortality requires direct measurement through physical examination or access to administrative records from health care providers.

Many health surveys limit their population of interest. For example, it is common for health surveys to only include the noninstitutionalized population. The dramatic differences in the living conditions of the institutionalized and noninstitutionalized populations make it difficult to design survey methodologies that would apply in all situations. The military population is also often excluded from general health surveys. The scope of the universe needs to be clearly defined, especially if populations that differ in their health status are omitted—such as persons residing in nursing homes.

When obtaining information directly from the population is inefficient or would result in either poor data quality or high costs or both, an alternative is to sample from existing records that were created for other purposes. Information on health care utilization for specific conditions is more easily obtained from hospital or provider records than from a population sample given the relative rarity of the phenomena of interest. The vital statistics system can also be considered a survey of administrative/legal records for the purpose of describing the health of the population. Administrative systems, in addition to providing basic health information, are often used as sampling frames for population based surveys. They are very attractive for this purpose as they eliminate the need to identify members of the population who have the characteristics of interest.

3. *Content*

Information collected on health surveys can be divided into three general types: (a) health status; (b) determinants or correlates of health including health behaviors; and (c) health care utilization. In addition, information is also collected on demographic and socioeconomic factors that affect health. For illustrative purposes, examples will be provided from two major health surveys conducted in the United States by the National Center for Health Statistics, CDC: The National Health Interview Survey (NHIS) (http://www.cdc.gov/nchs/nhis), and the National Health and Nutrition Examination Survey (NHANES) (http://www.cdc.gov/nhanes). Many statistical agencies work closely with counterparts in other nations in the design of their health surveys, thus, many observations regarding the NHIS and NHANES apply to similar surveys in other countries. Attempts to standardize various parts of health surveys have increased in recent years.

The NHIS is a principal source of information of the health of the civilian noninstitutionalized US population. Conducted annually since 1957, the NHIS currently collects information through computer assisted in-person interviews from approximately 40,000 households covering 100,000 people. Conducted about once a decade, NHANES is based on detailed physical examinations in addition to in-depth interviews. A nationally representative probability sample of the US civilian, noninstitutionalized population is interviewed at home and then invited to visit a Mobile Examination Center (MEC) for a standardized physical examination of about four hours, including a physician's exam, physical measurements, blood and urine tests, and nutritional assessments.

4. *Overall Health Status*

As noted above, health status is a multidimensional concept, requiring multiple indicators and multiple methodologies for adequate measurement. Several

different indicators of health status are usually included in health surveys, including single summarizing measures; questions relating to disease incidence and prevalence; and questions relating to functioning (physical, cognitive, emotional, and social). Health status measures also vary by whether they are based on objective information obtained from standardized examinations or medical records or from information obtained from the individual or a proxy. To be most effective, individual health surveys should capture a variety of aspects of health status so as to provide a more comprehensive and complete assessment of health status than would be possible from any single strategy. Measures of health status should also be constructed to be useful in epidemiologic analyses of risk factors as well as for monitoring trends.

The subject's self-assessment of his/her health status as excellent, very good, good, fair, or poor is a popular summary indicator of health status. This measure has been shown to be highly correlated with other measures of health status and is predictive of mortality and admission to long-term care facilities (Idler and Angel 1990). However, this measure can be problematic when used to monitor change over time. The means by which individuals evaluate various aspects of health have been shown to be affected by contextual parameters. Implicit in self-perceived health status is the individual's evaluation of his/her health status against some unstated standard. Societal norms act to define the standard, but the norms and the standards change in response to a variety of conditions. The reporting of general limitation of activity has also been shown to be affected by changes in such things as the criteria used by government agencies in determining eligibility for disability benefits (Wilson and Druny 1984). Thus, observed changes in the indicator may not reflect changes in underlying health status.

5. *Medical Conditions*

A variety of approaches can be used to measure disease incidence and prevalence including: diagnosis by a health care provider; symptomatology (if an appropriate symptom battery exits); or medication use. Additional information of interest related to conditions includes date of onset, the use of health care services, and the impact of the condition on the individual's ability to function. Severity of condition is often inferred from the type of medications or health care services used or the nature of the impact of the condition. In examination surveys, diagnostic testing adds to the information available on the prevalence of conditions. Most important, direct physical exams allow for the measurement of previously undiagnosed conditions.

A major issue in the collection of condition data is the number of conditions to be considered. Condition checklists can be used to obtain information on a great number of conditions, but in most cases, only a minimal amount of information on any given condition is collected. In addition, long checklists are difficult to administer and result in reporting errors. However, the use of comprehensive lists does allow for the ascertainment of co-morbidities. An alternative is to obtain more detailed information on a select group of conditions. Criteria such as the prevalence of the condition or the potential impact for the individual and the health care system are used to select the target condition. In the NHIS, for example, information is obtained for adults on heart disease, diabetes, cancer, pulmonary diseases, depression, and arthritis.

Some surveys, such as the NHIS (until 1997 when a major questionnaire redesign was fielded), limit their focus to conditions that result in the receipt of medical care or in activity restriction, thus placing more emphasis on the impact of the condition than on the fact that a diagnosis had been made or symptoms reported. This provides a somewhat more concrete referent for the questions being asked which should result in improved data validity and reliability. However, limiting the definition of health problems to those that result in behavior such as seeking care or restricting activity is more appropriate when ill health is primarily the result of acute conditions where onset is clear and there is a closer association between the condition and the behavior. As a result of the 'epidemiologic transition,' acute conditions have given way to chronic conditions as the major sources of morbidity. Chronic conditions are more difficult to diagnose, and longer periods are spent in the disease state both prior to and after diagnosis. As a result, the collection of information on symptoms takes on greater importance when investigating chronic conditions that may or may not have been diagnosed. Diagnostic examination or testing is necessary for some important conditions (e.g., hypertension, cancer) which may be asymptomatic in their early or less severe stages.

Requiring that health care use or activity limitation be present in identifying conditions also makes it difficult to investigate social correlates of health status since the definition incorporates a social dimension. Access to sick leave or affordable medical care may be limited by social or economic factors. Persons who do not demonstrate the required behaviors because of these social factors will not be considered as having a health condition.

Examination surveys like the NHANES provide a complementary mechanism for obtaining information on conditions. Self-reported information, similar to that collected on the NHIS on self-perceived health status, disease symptomatology, physician diagnosis, and level of functioning, is obtained from in-home or MEC interviews. In addition, the results of the physical exam, and of the blood and other tests, can be used to objectively measure health status. For example, information on diagnosed diabetes is

obtained from the interview, and information on undiagnosed diabetes is obtained from blood tests. Information on functional limitations related to diabetes (e.g., vision and mobility problems) coupled with similar information available from the other components of the NHANES allows for a detailed characterization of an individual's health status. The information is also used for the investigation of risk factors for disease and for monitoring of functional outcomes.

Information on conditions can also be obtained from surveys of health care providers (hospitals, private physicians, hospital outpatient departments, emergency departments, ambulatory surgery centers, and long-term care). Sampling frames of providers must first be developed. Random samples of discharges, stays, or visits are then selected from the records of a random sample of providers. Information on conditions is routinely found on the medical records in addition to other descriptive information concerning the patient and the encounter. Such survey designs are generally used to obtain information on health care utilization and the health care system, but the information available can be used to study conditions that are associated with medical care.

Health care or provider surveys offer an independent assessment of disease prevalence by approaching the problem from the provider side rather than the patient side. Health care surveys are particularly important since they obtain information on rare conditions that would not be picked up in population based surveys. They also provide larger sample sizes than would be obtained in surveys such as the NHIS and the NHANES. While information is collected on all conditions for which health care is obtained, the nature of the information available is limited to that which would routinely be obtained from records. In addition, the event (e.g., the discharge), rather than the person, is the unit of analysis. For example, while it is possible to estimate the number of hospital stays associated with hip fracture, it is not possible to estimate the number of persons with hip fracture even if one could assume that all hip fractures were admitted to the hospital. An enhancement to health care surveys would be to follow patients after the health care encounter.

Cause of death statistics obtained from a vital statistics system also provide essential information on medical conditions. While not error free, a great deal of effort is put forth to standardize the collection and coding of cause of death (http://www.cdc.gov/nchs/icd9).

6. *Functioning*

An important component of health status concerns functioning—physical, cognitive, emotional, and social. While conditions relate more to the biologic aspects of health, functioning moves more into the social realm. Information on functional status can be ascertained through objective tests but is most often obtained through direct reporting by the individual or a proxy. Information on functional status encompasses a wide range of activities from the very basic such as raising an arm to the more complex such as working and going to school. Obtaining information on functioning is complicated by the fact that performance is affected not only by physiologic abilities but by the environment as well. A person's ability to walk may be impaired by a condition that leads to weakened muscles, but with the use of braces or a wheel chair along with appropriate accommodations in the physical environment (e.g., ramps), that individual is mobile and able to carry out appropriate social roles. For some purposes, information is needed on functional ability without the use of assistance of any type. In other cases, functioning ability as measured by usual performance is important. In any case, questions must be specific concerning how performance should be measured. A large number of question batteries have been developed to measure function (McDowell and Newell 1996). Some use only a few questions, focus on the more complex social roles, and do not differentiate between different aspects of functioning (physical, cognitive, emotional, or behavioral). Others are quite lengthy and obtain detailed information on a wide range of activities both with and without assistance.

In the NHIS, information is obtained on limitation of activity (i.e., the subject's ability to perform age-/sex-appropriate roles such as working, going to school, keeping house), activities of daily living (e.g., bathing, dressing, toileting, eating, and transferring), and physical functioning (e.g., walking, bending, and standing). Information is also obtained on the conditions causing the limitation. A subset of the NHIS functioning questions is provided in Table 2.

As for self-perceived health status, measures of functioning are affected by contextual parameters as well as unstated standards against which respondents compare their own functioning. The less explicit the referent standard, the more likely that the response will be affected by external factors. For example, no explicit standard is provided when asking subjects if they are limited in the kind or amount of work they can do. Responses to this question will be affected by the kind of work that the individual thinks he or she should be doing as defined by their own expectations and by the requirements set by government agencies such as disability programs as well as by more objective states of health. This is less of a problem if one is interested in monitoring the impact of health on work. However, it is much more of a problem if the aim is to understand and monitor a more objective measure of health.

Many health surveys are limited to the noninstitutionalized population and therefore do not include

Table 2
Examples of questions on functioning from the National Health Interview Survey

Family members under five years of age	Are/Is {fill in name of child} limited in the kind or amount of play activities he/she/they can do because of a physical, mental, or emotional problem?
	Is {subject name} able to take part at all in the usual kinds of play activities done by most children {subject name}'s age?
	Do any of the children under 18 in this family, {fill in names of children under age of 18} receive Special Education or Early Intervention Services?
	Has a doctor or health professional ever told you that {subject name} had Mental Retardation or any other developmental delay?
	Has a doctor or health professional ever told you that {subject name} had Attention Deficit Hyperactivity Disorder (ADHD), Attention Deficit Disorder (ADD), Mental Retardation, or any other developmental delay.
Family member between the ages of 4 and 17	Do any of the children under 18 in this family {fill in names of children under age of 18} receive Special Education or Early Intervention Services?
Family member older than 17 years of age	Do you now have any health problem that requires you to use special equipment, such as a cane, a wheelchair, a special bed, or a special telephone?
	By yourself, and without any special equipment, how difficult is if for you to walk a quarter of a mile—about 3 city blocks, walk up 10 steps without resting, stand or be on your feet for about 2 hours, sit for about 2 hours, stoop, bed, kneel, or reach up over your head? (each asked separately)
	By yourself, and without using any special equipment, how difficult is it for you to use your fingers to grasp or handle small objects, lift or carry something as heavy as 10 pounds such as a full bag of groceries, push or pull large objects like a living room chair? (each asked separately)
	By yourself, and without using any special equipment, how difficult is it for you to go out to things like shopping, movies, or sporting events, participate in social activities as visiting friends, attending clubs and meetings, going to parties, do things to relax at home or for leisure (reading, watching TV, sewing, listening to music …)? (each asked separately)
	Because of a physical, mental, or emotional problem, {do/does} {you/anyone in the family} need the help of other persons with personal care needs, such as eating, bathing, dressing, or getting around inside of the home?

Table 2
(Continued)

	{Do/Does} {you/subject name} need the help of other persons with bathing or showering, dressing, eating, getting in or out of bed or chairs, using the toilet, including getting to the toilet, and getting around inside of the home? (each asked separately)
	Because of a physical, mental, or emotional problem, {do/does} {you/anyone in the family} need the help of other person in handling routine needs, such as everyday household chores, doing necessary business, shopping, or getting around for other purposes?
	Does a physical, mental, or emotional problem now keep {you/any of these family members} from working at a job or business?
	{Are/Is} {you/anyone in the family} limited in any way because of difficulty remembering or because {you/they} experience periods of confusion?
	What condition or health problem causes you to have difficulty with activities/these activities?
	How long {have/has} {you/subject name} had this condition?

Source: Centers for Disease Prevention, National Center for Health Statistics. 2000 NHIS Family Section, Sample Adult, and Sample Child Questionaires.

those segments of the population in poorest health or with the highest level of functional limitations, such as those in long-term care settings. Special studies of these populations can be carried out and there should be some attempt to combine results from the non-institutionalized and long-term care populations (http://www.cdc.gov/nchs/about/major/nnhsd/nnhsd.htm, http://www.cds.duke.edu/text/nltcs.html). Changes in the delivery of long-term care with the expansion of transitional establishments have made this an even more important issue for survey designers.

7. *Determinants of Health*

Health surveys generally include measures of risk factors, health behaviors, and nonhealth determinants or correlates of health such as socioeconomic status. The range of measures that can be included is wide and varies by survey. Basic demographic and socioeconomic variables are usually included (i.e., age, race/ethnicity, education, income, urbanicity, region). Tobacco use, alcohol use, diet, and physical exercise are common health behaviors. In order to identify causal relationships between determinants and outcomes, longitudinal designs are needed. The value of health surveys would be increased by the inclusion of a broader array of potential determinants of health (biologic, psychological, and social).

8. *Summary Measures of Health*

There has been considerable interest in the development of summary measures of health in order to more easily characterize and monitor the health status of a population and for use in cost-benefit and cost-effectiveness analyses. There is also increasing interest in developing basic health status indicators comparable to the economic indicators currently in use. These indicators would be used to monitor the effects of public policy and to identify areas where interventions are needed. An overall health measure is very appealing in this context. Many health surveys include the information necessary to calculate the summary measures that have been developed. One such approach is to use a measure of morbidity or health status to transform life expectancy into healthy life expectancy. Examples include Disability Free Life Years, Healthy Life Expectancy, and Years of Healthy Life. Constructs such as these combine measures of the different aspects of health into a single number using a conceptual model that views health as a continuum. The health states along the continuum are assigned numbers that represent the values that either society as a whole or individuals place on that health status. Various methods have been used to determine

the values of the different health states. These measures are used to modify duration of life (Erickson et al. 1989).

9. Health Care Utilization

The use of the health care system is a major dimension of health. Information on utilization and related characteristics such as health insurance are collected from person-based surveys and through surveys of administrative records. In person-based surveys, respondents are asked to recall the number of contacts they have had with various health care providers and the nature of the provider and the contact. Information on the receipt of particular services, such as immunizations or mammography, is also obtained. This is often difficult for respondents. Research has shown that reporting of doctor visits drops off considerably when the recall period extends beyond two weeks (Edwards et al. 1996). To address this recall problem, some surveys such as the Medical Expenditure Panel Survey (MEPS) ask respondents to keep detailed diaries of their medical care contacts, obtain the name of the provider, and then contact the providers (with the subject's permission) to verify the self-reported information (Monheit et al. 1999). Information on utilization can also be obtained from surveys of providers, usually providing more accurate information on the nature of the contact, including the reason for the contact and the services provided. Surveys of providers can also yield valuable information on the structure of the health care system itself and how the characteristics of the system affect access to care and the cost of care.

10. Longitudinal Designs

Most health surveys use a cross-sectional design which provides data for the calculation of population averages and is most appropriate for monitoring changes in health over time. A cross-sectional design is not appropriate for investigating causal patterns or for documenting transition probabilities from one health state to another where change at the individual level needs to be measured. A solution would be longitudinal studies which are expensive and more complicated to field. One alternative is to transform cross-sectional studies into longitudinal studies by linking information from mortality records to the survey data. The United States has developed a National Death Index which allows investigators to track the vital status of their study cohorts by matching identifiers to a complete listing of death certificates. Using a probabilistic matching algorithm, information is provided on the fact, date, and cause of death ((http://www.cdc.gov/nchs/r&d/ndi). Cross-sectional surveys also can form the baseline cohorts for future active longitudinal follow-ups (US DHHS 1987). These longitudinal activities present extensive opportunities for understanding the relationship between risk factors and disease outcome as well as the natural history of disease. Generation and analysis of such longitudinal data is greatly facilitated in countries (e.g., Sweden) where various governmental information systems and surveys can be linked.

11. Technological Advances

Advances in technology have affected the conduct of all surveys. Computer-assisted telephone and personal interviewing have allowed for increasing levels of complexity of survey administration. Errors associated with keying and coding have been replaced by errors associated with programming. The use of these technologies has resulted in the need for longer lead times prior to fielding a study but has reduced the time from the end of data collection to the availability of results.

The increase in electronic as opposed to paper files has greatly expanded the opportunities for linking data from various sources. Ecological or contextual information about the geographic area in which survey respondents live can now be linked to the survey data. This has the advantage of allowing researchers to address different levels of analysis. Survey data can also be augmented with administrative data from mortality files and from health care providers. In addition to reducing survey costs, this often leads to data of improved quality.

12. Methodological Research and Analytic Techniques

The quality of health survey data has been greatly increased by methodological research on survey techniques (Bradburn and Sudman 1979, Fowler 1995, Lessler and Kalsbeek 1992). A major advancement has been the development and application of cognitive methods for questionnaire construction (Sirken et al. 1999). Drawing from the field of cognitive psychology, cognitive research laboratories have been established and can investigate how respondents perceive and answer questions. Another fruitful area of research involves studying the behavior of both interviewers and respondents within the social context of the interview situation (Cannel et al. 1977). A more structured approach to the errors found in health surveys will enhance data quality. The results of this research will lead to improved design that minimizes error but will also lead to ways to adjust for known errors in the analysis and interpretation of results (Jabine 1987).

Advances have also been made in the development of appropriate analytic tools for use with health survey data. These include methods to account for such issues as survey nonresponse and complex survey design. As

it is possible to collect data of increasing complexity, it will be necessary to develop the analytic approaches appropriate for the collection methods used (Korn and Graubard 1999, Biemer et al. 1991, Groves and Couper 1998, Rubin 1987).

13. Privacy and Confidentiality

For ethical reasons and in order to obtain high response rates and valid information, most health surveys closely guard the information provided by respondents. In some cases, the requirement to protect confidentiality is legislatively mandated. This is particularly important for health surveys, given the personal nature of the information collected. Confidentiality can also be protected by not releasing data that could identify a respondent. Most survey sponsors subject files that are to be released for public use to rigorous disclosure review. The risk for inadvertent disclosure has increased in recent years. It is no longer necessary to have access to large mainframe computers to utilize survey data, much of which are provided on the Internet or on CD-ROMs. In addition, databases not related to the survey data, but which can be used to identify individuals in the survey data files, are more available and more easily accessible.

It is the potential linking of survey data to these external databases that increases the risk of disclosure, and the risk increases with the amount of information that is available. The ability to link external data to survey responses presents an analytic breakthrough as the utility of survey data can be greatly increased through the appropriate linkages, however, this ability also greatly increases the risk of disclosure especially when the linkages are done in an uncontrolled and inappropriate way. The increased sensitivity to issues of privacy in many countries, especially as related to health care, is affecting how confidential data are being protected. There has been a decrease in the amount of data that can be released as public use files. Other mechanisms are being developed so that access to data can be maximized while protecting confidentiality. The use of special use agreements, licensing and research data centers are examples of these approaches (Journal of Official Statistics 1998).

See also: Health Research, Qualitative; Public Health; Public Health as a Social Science; Quality of Life: Assessment in Health Settings; Statistical Systems: Health

Bibliography

Biemer P P, Groves R M, Hyberg L, Mathiowetz N, Sudman S (eds.) 1991 *Measurement Errors in Surveys*. Wiley, New York

Bradburn N M, Sudman S 1979 *Improving Interview Method and Questionnaire Design*. Jossey-Bass, San Francisco

Cannel C F, Monquis K H, Launent A 1977 A summary of studies of interviewing methodology. *Vital and Health Statistics* (Series 1, No. 69)

Edwards W S, Winn D M, Collins J G 1996 Evaluation of 2-week doctor visit reporting in the National Health Interview Survey. National Center for Health Statistics. *Vital and Health Statistics* (Series 2, No 122)

Erickson P, Kendall E A, Anderson J P, Kaplan R M 1989 Using composite health status measures to assess the nation's health. *Medical Care* **27**(3): 566–77

Fowler F J 1995 *Improving Survey Questions*. Sage, Thousand Oaks, CA

Groves R M, Couper M P 1998 *Nonresponse in Household Interview Surveys*. Wiley, New York

Idler E L, Angel R J 1990 Self-rated health and mortality in the NHANES I. Epidemiologic follow-up study. *American Journal of Public Health* **80**(4): 446–52

Jabine T B 1987 Reporting chronic conditions in the National Health Interview Survey: A review of tendencies from evaluation studies and methodological tests. National Center for Health Statistics. *Vital and Health Statistics* (Series 2, #105)

Journal of Official Statistics 1998 Disclosure limitation methods for protecting the confidentiality of statistical data. *Journal of Official Statistics* **14**: 337–573 (Special issue)

Korn E L, Graubard B I 1999 *Analysis of Health Surveys*. Wiley, New York

Lessler J T, Kalsbeek W T 1992 *Nonsampling Error in Surveys*. Wiley, New York

McDowell I, Newell C 1996 *Measuring Health: A Guide to Rating Scales and Questionnaires*. Oxford University Press, Oxford, UK

Monheit A C, Wilson R, Arnett R H (eds.) 1999 *Informing American Health Care Policy: The Dynamics of Medical Expenditure and Insurance Surveys, 1977–1996*. Jossey-Bass, San Francisco

Moore J 1988 Self/proxy response status and survey response quality: A review of the literature. *Journal of Official Statistics* **4**(2): 155–72

National Center for Health Statistics 1999 *Health United States*. National Center for Health Statistics, Hyattsville, MD

National Center for Health Statistics 2000 *International Health Data Reference Guide 1999*. National Center for Health Statistics, Hyattsville, MD

Rubin D B 1987 *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York

Sirken M G, Herrmann D J, Schecter S, Tourangeau R (eds.) 1999 *Cognition and Survey Research*. Wiley, New York

Stewart A L, Ware J E (eds.) 1992 *Measuring Functioning and Well-being: The Medical Outcomes Study Approach*. Duke University Press, Durham, NC

United States of America, Department of Health and Human Services 1987 Publication No. (PHS) 92-1303, Series 1, No. 27. Plan and Operation of the NHANES I Epidemiologic Followup Study

Wilson R W, Druny F 1984 Interpreting trends in illness and disability: Health statistics and health status. *Annual Review and Public Health* **5**: 83–106

World Health Organization 1948 *World Health Organization Constitution in Basic Documents*. World Health Organization, Geneva

J. H. Madans

Published by Elsevier Ltd.

Hebb, Donald Olding (1904–85)

Donald Hebb was responsible in no small part for the profound change in psychological outlook that emerged in the 1950s. His influence on the field of cognitive neuroscience is still apparent.

1. Education and Early Life

Hebb was born, the eldest of four siblings, on July 22, 1904 in Chester, Nova Scotia. Both his parents were physicians. He was educated at home until the age of eight by his mother, using Montessori methods. He then attended the local school, which he hated.

Hebb's ambition as a young man was to become a novelist. He graduated from Dalhousie University (Halifax, Nova Scotia) in 1925 with a BA in English and philosophy. While awaiting inspiration for the 'Great Canadian Novel' he returned to his village school as a teacher. In search of experience he spent the next summer as an itinerant farm worker in western Canada and tried, unsuccessfully, to get a job as a deckhand in Vancouver.

In 1927 Hebb was back East, teaching at a Montreal high-school. Becoming interested in Sigmund Freud's ideas, he applied to McGill University to study psychology part time. The next year he was appointed principal of an elementary school and enthusiastically set about instituting and testing new teaching methods. In 1930 he was bedridden for a year with a tubercular hip, during which time he conceived a theory of how spinal reflexes might be learned *in utero*. This formed the basis of an MA thesis, now lost. Hebb (1980) later described it as nonsense.

In 1932 Hebb began doctoral research, still part-time, at McGill. The experiments, which involved conditioning dogs, were supervised by Leonid Andreyev, a former student of Pavlov, but Hebb soon became disillusioned with the technique and when his first wife was killed in a motor accident he abandoned the research.

Hebb enjoyed teaching and administration and was tempted to make a career of being a headmaster. He finally chose to continue his studies, however, and in 1934 he was accepted as a student at the University of Chicago by Karl Lashley, whose book, published a few years earlier (Lashley 1929), had established him as the pre-eminent physiological psychologist of the period.

Hebb spent a year in Chicago where, in addition to his close association with Lashley, he attended lectures by Wolfgang Köhler, Paul Weiss, C. J. Herrick, and the statistician L. L. Thurstone. Then Lashley accepted the offer of a professorship at Harvard and Hebb accompanied him. He was granted a Harvard PhD in 1936 for research on the effects of early experience on vision in the rat.

2. Montreal Neurological Institute

After a post-doctoral year of research with Lashley, Hebb was awarded a two year Fellowship to study mental changes resulting from brain operations at the newly established Montreal Neurological Institute (MNI). The operations were mostly frontal and temporal lobectomies performed by Wilder Penfield to combat epileptic seizures. Hebb made what is probably the first observation of a visual deficit after human right-temporal lobe removal. To his surprise he found no measurable effect of frontal lobectomy on standard measures of intelligence (Hebb and Penfield 1940). Attempts to understand this result lay at the foundation of much of Hebb's later work.

3. Primate Research with Lashley

After his MNI Fellowship, Hebb obtained a post as Lecturer at Queen's University (Kingston, Ontario). There he devised a variable rat-maze so that he could continue his brain lesion research using rats. In 1942 Lashley offered him a five year Fellowship to study effects of brain lesions in chimpanzees at the Yerkes Laboratories of Primate Biology, Florida. Partly because of his heavy teaching load at Queens, he accepted it.

The project was to study the effects of brain lesions on intelligence and emotion in chimpanzees. Lashley was to work on intelligence tests, Hebb on tests of emotion, but no lesions were made before 1947 when Hebb's Fellowship ended. Hebb did, however, assemble a good stock of chimpanzee stories that contributed to the great popularity of his subsequent lectures, and in the intellectually stimulating surroundings he wrote his most influential work, *The Organization of Behavior*, which was published in 1949.

4. McGill Once More

After World War II, the McGill psychology department had dwindled to one part-time child psychologist, so in 1947 Hebb was one of four psychologists who were recruited to rebuild the department. The next year he was appointed chairman.

Having made a highly successful foray into the world of science, Hebb returned to teaching and administration with enthusiastic vigor. He taught the introductory course to overflow classes for the next 25 years, and wrote three editions of a highly respected textbook. He also conducted a seminar for all incoming graduate students, which gave him the opportunity to coach them in the presentation of succinct and closely timed talks, as well as to instill a thorough understanding of a range of theoretical issues. To test some predictions of his theory, he supervised research on early learning in rats and dogs, and effects of sensory deprivation on human thought processes.

In later years Hebb received many honors; he was President of both the Canadian (1952) and the American (1960) Psychological Associations. He was elected a Fellow of the Royal Society of London (1966), a Foreign Associate of the National Academy of Sciences (USA), and an Honorary Member of the American Academy of Arts and Sciences. In 1970 he was appointed Chancellor of McGill University. In 1974 Hebb retired to his family home in Chester Basin and became active in the psychology department at Dalhousie University, where he was appointed Honorary Professor. He published his last monograph, *Essay on Mind* in 1980 and died in Halifax on the 20th of August, 1985 of heart failure during hip-replacement surgery.

5. *Hebb's Contributions to Neuroscience*

During his years as a graduate student Hebb learned the hard way what was wrong with contemporary psychology. At McGill, as is evident from the topic of his Masters thesis, he flirted with the extreme empiricism then rampant in North America. Influenced by Boris Babkin and Leonid Andreyev, students of Pavlov, he passed through a Pavlovian phase. When he moved to Chicago he entered the Gestalt enclave of Wolfgang Köhler and Karl Lashley; and at Harvard he encountered the historian, Boring, who taught him what psychology had been like before the Behaviorists. Sooner or later, for reasons that are cogently presented in the first few chapters of *The Organization of Behavior* (Hebb 1949), he turned against all the established schools of psychology.

Because of his research on cortical lesions and his neural theories, Hebb is often labeled as a physiological psychologist, but his interests were really much broader. *The Organization of Behavior* was an important catalyst, aiding the revival of cognitive psychology, and although he was a zealous proponent of the idea that the brain is the organ of mind, Hebb's interest in neurophysiology took second place to his interest in thinking and other aspects of behavior. In his hands neurologizing was a tool to restore respectability to the study of mental functioning in psychology.

As a psychologist, Hebb was frustrated by the behaviorists' refusal to acknowledge such important processes as attention, expectancy, and thought, lest they should be accused of vitalism. Hence his eagerness to show that those mental processes could be performed by an entirely material brain. He based his neural model as far as possible on established physiological principles, but neuroscience was on the brink of a revolution and many of those principles were soon to be revised.

For example, Eccles, whose defense of the electrical theory of synaptic transmission Hebb had embraced, recanted and demonstrated excitatory and inhibitory chemical synapses. Furthermore, the behavioral importance of subcortical structures such as the reticular arousal system, the amygdala and the hippocampus became clearer, and improved staining techniques revealed a much more complex visual system than was known to Hebb. The fact that these revelations had almost no effect on the impact of Hebb's book shows to what extent the psychological implications of his neurologizing outweighed the underlying physiology. By proposing a material basis for concepts, Hebb demolished barriers to the study of cognitive processes that the behaviorists had erected a generation earlier. In his Presidential address to the Experimental Psychology Division of the American Psychological Association in 1954, Hebb introduced the reticular activating system into his model to extend his explanation of emotion, but apart from that he did little to revise the theory in the light of new neuroscientific data.

His most famous contribution to neuroscience was not based on neurophysiological research, it was inferred from behavior. Most physiologists thought that learning synapses were strengthened by use, but Hebb along with other psychologists realized that association involved at least two activities. Thus he postulated that learning changes at synapses require simultaneous firing of both presynaptic and postsynaptic neurons. Synapses with properties similar to those postulated were subsequently identified experimentally and are now known as Hebb synapses.

Hebb held strong views about the 'nature–nurture' question. After his early gyrations he adopted the unassailable position that both inheritance and experience are totally necessary for intelligent behavior, using the analogy of area as the product of length and breadth. Of course, one may still be interested in whether an area is square, long and narrow, or short and fat. In spite of Hebb's bipartisanship, his concept of the development of vision was long in the 'nurture' dimension. Hebb's visual cortex bore a strong resemblance to Locke's *tabula rasa*.

Thus, Hebb assumed that neurons are randomly connected in the newborn visual cortex. With experience, synapses are strengthened between neurons activated simultaneously by sensory input, producing groups of connected neurons (*cell assemblies*) that then represent the input. Hebb further postulated that a complex figure elicits a series of eye movements, thereby activating a succession of associated cell assemblies that Hebb called a *phase sequence*.

Anatomical diagrams of the cortex suggested to Hebb that cell assemblies consist of networks of closed loops. This led him to speculate that once a cell assembly is activated impulses continue to circulate around the loops for some time (reverberation), constituting a short-term memory. In this way Hebb explained the persistence of an idea or mental image in the absence of the corresponding object. Today, short-term memory may be explained more plausibly by

persistent chemical traces, but Hebb never abandoned his reverberation theory, which still dominates the thinking of many neural model builders and neurophysiologists who study the visual system.

According to Hebb, the impulses reverberating around the cortical loops of an activated cell assembly produce synaptic changes that bind the neurons together more permanently. This explains long-term memory. Ideas arouse other ideas by association, thus cell assemblies must acquire associations with each other whenever they are active at about the same time. Hebb never satisfactorily explained how, after linking in this way, they avoid merging to become bigger cell assemblies.

The capacity to explain intention, attention and other mental processes was an important goal of Hebb's theorizing but one that, in *The Organization of Behavior*, he was content to achieve in principle rather than by rigorous neural modeling. Short-term expectancy and attention were explained in terms of the phase sequence getting ahead of the sensory input. Thus a previously experienced sequence of stimuli sets up a central process that facilitates cell assemblies in advance of the sensory input they represent. Hebb postulated higher order central processes that facilitate more extended sequences. In much of his writing Hebb treats the cell assembly as equivalent to an idea; its neural basis is tailored to fit behavioral observation.

6. *Summary*

The Organization of Behavior is now a classic. Half a century after it was written it is still frequently cited. Many psychologists regard it as the herald of an era of cognitive research that blossomed in the 1950s. Hebb's ideas about the representation of concepts and their associations with each other are the common currency of almost all present-day psychologists, even those who prefer to ignore the brain, but devotees of neural networks and neurophysiologists studying sensory systems are those who most often refer explicitly to his theory.

After Hebb left Lashley's 'think tank' in Florida and took up his duties as chairman of the McGill psychology department, most of his energies turned to building and running a strong department there, including teaching. Although Hebb took every opportunity to promote neurologizing in psychology, he showed little further inclination to keep pace with the rapid advances in neuroscience or to make any major changes to his neural theory.

Apart from his own characteristically stylish and witty account (Hebb 1980a), there have been several more recent commentaries that complement this entry for those who wish to learn more about Hebb's life and work. These include nine essays in his honor, written by former students and edited by Jusczyk and Klein (1980), and a delightful chapter by Stephen Glickman (1996) which accurately portrays the man and his achievements. Fentress (1999) and Klein (1999) edited a special issue of the *Canadian Journal of Experimental Psychology* that contains a further collection of papers on Hebb and on research inspired by his ideas. Orbach, a student of both Hebb and Lashley, published a book (1998) in which the contributions of his two mentors are discussed and compared.

See also: Behaviorism; Brain Stimulation Reward; Cognitive Neuropsychology, Methodology of; Cognitive Psychology: History; Cognitive Psychology: Overview; Learning and Memory, Neural Basis of; Lesion and Behavior Approaches in Neuroscience; Linear Algebra for Neural Networks; Long-term Depression (Cerebellum); Long-term Depression (Hippocampus); Long-term Potentiation and Depression (Cortex); Long-term Potentiation (Hippocampus); Memory: Synaptic Mechanisms; Neural Development: Mechanisms and Models; Neural Networks, Statistical Physics of; Neuropsychological Functioning, Assessment of; Synaptic Efficacy, Regulation of

Bibliography

Fentress J C 1999 The organization of behavior revisited. *Canadian Journal of Experimental Psychology* **53**: 8–20

Glickman S E 1996 Donald Olding Hebb: Returning the nervous system to psychology. In: Kimble G A, Boneau C, Wertheimer M (eds.) *Portraits of Pioneers in Psychology*. Erlbaum Press, Mahwah, NJ, Vol. 2, pp. 227–44

Hebb D O 1949 *The Organization of Behavior: a Neuropsychological Theory*. Wiley, New York, pp. 1–335

Hebb D O 1955 Drives and the C.N.S. (conceptual nervous system). *Psychology Review* **62**: 243–54

Hebb D O 1980 *Essay on Mind*. Erlbaum Press, Hillsdale, NJ

Hebb D O 1980a D. O. Hebb. In: Lindzey G (ed.) *A History of Psychology in Autobiography*. Freeman, San Francisco, CA, Vol. 7, 273–303

Hebb D O, Penfield W 1940 Human behavior after extensive bilateral removal from the frontal lobes. *Archives of Neurology and Psychiatry* **44**: 421–38

Jusczyk P W, Klein R M (eds.) 1980 *The Nature of Thought. Essays in Honor of D. O. Hebb*. Erlbaum Press, Hillsdale, NJ

Klein R M 1999 The Hebb legacy. *Canadian Journal of Experimental Psychology* **53**: 1–7

Lashley K S 1929 *Brain Mechanisms and Intelligence*. University of Chicago Press, Chicago, pp. 1–186

Orbach J 1998 *The Neuropsychological Theories of Lashley and Hebb*. University Press of America, Lanham, MD

P. M. Milner

Hegel and the Social Sciences

G. W. F. Hegel (1770–1831), born in Stuttgart, achieved his full fame in Berlin between 1818 and 1831 as a professor of philosophy. Much of his fame rests

on the posthumously edited and published lectures on the philosophies of art, religion, and history, along with his epochal book of 1807, the *Phenomenology of Spirit*. However, despite Hegel's immense influence on modern thought, he has remained one of the most controversial and difficult to interpret philosophers of the modern period. It is nonetheless widely agreed that Hegelian-inspired theories of society rest on certain fundamental theses: first, they give fundamental priority to the 'totality' of social life over more individualist accounts; second, such theories claim that we can only understand the kinds of demands social life makes in terms of its historical embeddedness; and, finally, such theories claim that there can be no value-free approach to history and society. There is, however, not nearly as much consensus as to what each of these general claims meant for Hegel or about what follows from them. In some interpretations, Hegel has been taken as an ominous figure, a kind of dark warning for modern academics to stay away from metaphysics; in other interpretations, Hegel is taken as one of the original theorists of European modernity and as an ongoing source of insight and inspiration.

Although part of the difficulty of interpreting Hegel has to do with his notoriously obscure style and his sometimes ferociously idiosyncratic jargon, much of the difficulty has to do with how one is to take his statements about the role of mentality (*Geist*, perhaps better rendered in English somewhat artificially as 'mindedness') in nature, society, and history (and often one's interpretation of that depends how one understands Hegel's ultimate stance toward religion). This is also tied into Hegel's general holism, the thesis that the politics, art, religion, and economy of a society are so intimately related to each other that not just any concatenation of such practices and institutions is compatible with any others, and that the point of such practices and institutions is to provide the vehicles through which a way of life can collectively reflect on and become self-consciously aware of what ultimately matters to it. In Hegel's own lifetime and afterwards, people took his holism more or less in one of two ways: either as a grandly metaphysical statement about an absolute reality (called *Geist*, mindedness) that assumes different forms in history and which progressively reveals itself to itself; or as a more normative account of agency that does not invoke that kind of metaphysical entity acting behind the scenes to explain things.

To understand what is at stake in that debate, one must at least have some idea of the rather heated atmosphere present in German philosophy at the time in which Hegel came to prominence.

By the end of the eighteenth century, Kant's critical philosophy had changed the very terms of debate in Germany, and while certainly not everybody became a Kantian, almost all aspiring philosophers in Germany felt that they had to come to terms with Kantian thought. Kant had, according to his own description of things, essentially demolished the longstanding feud between empiricists—those who believed that science was only a sophisticated and methodical purification of our more ordinary responses to the givens of experience—and rationalists—those who thought that Galilean and Newtonian physics showed that the mind was capable of mathematically articulating the structure of the universe independent of experience. Kant argued that there were no givens of experience at all, that our conscious experience of the world always involved an active and spontaneous 'taking-up' of that experience (a rationalist theme), although our activities of 'taking-up' experience depended for that experience for its empirical content (an empiricist theme). He also argued that in the moral realm, we are completely self-legislating; there are no givens within human nature or in the cosmos at large that can determine what may count as a moral principle for us. The reality of this radical human spontaneity and freedom was secured, so Kant had argued, by distinguishing between the way in which we necessarily had to describe and order our experience of the world and the way that the world existed in itself independent of any description we might give it. We could never know, in Kant's famous phrase, the way things were in themselves, only the way in which they necessarily appear to us. Thus although we had to appear to ourselves as standing under deterministic natural causal laws, we also had to postulate for practical reasons the necessity of our freedom as we were 'in ourselves,' even though we could never theoretically demonstrate the existence or lack of existence of such radical freedom.

For Hegel's generation, who had come to believe, however indistinctly, that they were leading unprecedented lives and for whom the French Revolution was the defining event, Kant's philosophy of freedom and spontaneity had an electrifying effect. Kant's immediate successors were also convinced that the outstanding problems with his own solution called for an effort to 'complete' the project he had begun. This took the forms of attempting to establish Kantianism on better foundations, or of focusing on the way in which Kant had (or should have) rejected all forms of 'givenness' in epistemology and morality, or of showing how Kant's own dualisms were to be reconciled by showing them to be only mutually conflicting 'appearances' of some deeper underlying 'absolute' reality that itself was only to be apprehended through some non-discursive act of intellectual intuition. The latter turn was famously carried out by Hegel's old friend, F. W. J. Schelling, who crafted an entire theory of nature and mind as modes of appearance of the 'absolute,' or as the way the absolute gradually realizes itself in history. (In part, much of the debate about Hegel's philosophy has since centered on just how 'Schellingian' or 'non-Schellingian' his thought really was.)

In much of his writing prior to the 1807 *Phenomenology of Spirit*, Hegel had worried about what shape a 'modern' way of life would take and in particular what role religion, especially a reinterpreted form of Christianity, might play in such a modern way of life. Throughout his development, however, he held fast to the post-Kantian notion that the modern, post-revolutionary world was something qualitatively new and demanded a 'new sensibility' appropriate to itself. To that end, Hegel's 1807 *Phenomenology* was intended to demonstrate what would be required of 'we moderns' if we to conceive of our lives without any secure, 'given' foundations. In his *Phenomenology*, Hegel sarcastically characterized Schelling's notion of the Absolute as 'the night in which all cows are black,' signaling his rejection of the idea of an absolute reality that is beyond all conventional descriptions and differentiations and which can only be grasped in some non-discursive mode. So Hegel argued, rejecting all 'givens' and affirming Kantian notions of spontaneity implied there could be no inherently privileged viewpoint without at the same time also implying that we were stuck with any kind of pure relativism about viewpoints.

In an early chapter in the *Phenomenology* (drawing on some themes taken from Fichte and Schelling), Hegel constructed an imaginary encounter between two agents, each thinking of himself as seeking only to satisfy the desires given to him by organic life. To the extent that each agent legitimates his own judgments about what is valid for him only in light of locating himself within the totality of 'life' and its demands, he demands of the other that the other recognize his point of view as authoritative, that he may do what he pleases to satisfy his own desires, even to use the other to get what he wants. In making that demand, though, each finds that it is not 'life' itself that is setting the terms of judgment but his own self-conceived project for his life, of his bestowing on *his* desires a prior claim to satisfaction over the desires of others. Each thus demands 'recognition' from the other of the authoritativeness of his own point of view—not merely that he has pressing desires but that he determines which desires and whose desires have a legitimacy to satisfaction. Those reciprocal and incompatible demands necessitate a struggle for such recognition, and to the extent that one of the parties decides that his own self-conception is more important than life itself and is willing to risk his life to secure the authoritativeness of his self-conception, it becomes a struggle to the death.

If in such a struggle for recognition, one of the combatants out of fear for his life submits to the other's demand for recognition, he then enters into the relationship of servitude to that other, who now becomes the 'master.' The vassal then comes to judge the validity of his own desires, tastes, and plans in terms of how they fit into the master's point of view. Or, as Hegel put it, the vassal's own *self-relation* is now mediated through an 'other.' Through the discipline of his work for the master, the 'vassal' comes to distinguish more clearly between his own subjective point of view and a more impersonal, normative point of view represented by the master. The master's point of view represents the 'totality' in terms of which the vassal must orient himself. However, as only a particular point of view whose sole claim to authority has to do with its ability to force itself on the vassal, it cannot ultimately sustain any normative allegiance to itself. As the vassal reflects on his status, he comes to understand the sheer contingency of the master's hold on him, and the master comes to understand that the recognition he requires from the vassal is only compelled, not the free recognition he requires.

Self-consciousness in Hegel's scheme thus only comes about through struggles over mutual recognition. To be self-conscious in Hegelian terms is thus not to have some inner sensation of oneself or merely to be consciously alive; it is rather to locate oneself and others (and similarly be located by others) in a complex 'social space' of norms, commitments, entitlements, and the like, and for both parties to act in light of that mutual self-locating and being located. This conception of self-consciousness as emerging only in mutual recognition and of its being the result of a struggle turned out to be one of Hegel's most enduring and widely interpreted legacies. Although others had often made claims about the inherently social nature of man, Hegel was the first to make this claim about the very structure of reasoning and agency itself. Each agent needs the other not to help him satisfy some given desire but to *be* an agent at all. Agency is thus itself a reciprocally conferred social *normative status*, not a metaphysical *fact* about people: to be an agent is to be the kind of organic creature that binds himself to and is held by others to epistemic, practical, even religious and aesthetic *responsibilities*. We become agents, almost paradoxically, by treating each other as agents.

The rest of the *Phenomenology* presented a historical account of how various modes of sociality—of complex reciprocal holding of self and others to basic authoritative norms—have failed to be able to sustain themselves and became thereby unlivable, compelling their participants to come to recognize that they had over time failed to sustain their own status as agents and (without intending it) had come to require of themselves a different form of collective life. In such transitions, what had seemed like an 'absolute' point of view had turned out to be incoherently unlivable or to be only the point of view of some particular group, and its capacity to sustain allegiance to itself unraveled. In those terms, Hegel reinterpreted not only the great transition from the ancient to the modern world but also certain particular historical transitions such as that from ancient Greece to Rome or from the aristocratic society of the *ancien régime* to the moralistic society of postrevolutionary Europe. Each such transition came to be understood as part of the

development of reason itself, of what counted as a basic norm and therefore what counted as agency itself. This was in Hegel's account not the development of a latent capacity somehow already present within us but which was only gradually actualizing itself nor was it the result of any kind of natural law compelling such historical movement, as it were, from the outside; it had to do with the inherent difficulties involved in binding each other to certain types of norms and in the way in which we came to be required to hold ourselves to different norms—that is, in *being* certain types of agents—and in the specific ways being those types of agents had failed and quite unintentionally resulted in different collective ways of being agents.

In his own lifetime and until today, people have tended to read Hegel's philosophy of history in the *Phenomenology* as a kind of story about how some appropriately grand absolute reality (identified with mindedness itself, *Geist*) gradually develops itself and actualizes its own potentiality by using various world-historical individuals and nations as a way to follow its own destined path of maturation. However, Hegel himself argued that the sense of destiny in history had nothing to do with laws of history or any such grand metaphysical scheme but with the ways in which human agents can understand the kind of agents they concretely *are* only in light of the kinds of options that their history has opened up for them. The realization that that there is no grander show behind the curtain of appearance pulling the strings for us, either in the form of historical laws (metaphysical or material) or in the form of gradual actualizations of longstanding latent potentialities, is itself distinctly modern and, as an *unsurpassable* form of self-understanding, as marking perhaps *the* decisive event in human time, is itself *absolute*. It also represents, so Hegel thought, the human community coming to a more adequate collective self-consciousness about the kinds of norms to which it had committed itself and to which it had to keep faith; and, quite famously, Hegel held that art, religion, and philosophy are to be understood as historically embedded social practices whose underlying point has to do with the manner in which they embody a way of life's concrete reflection on who they have come to be and on the matters that have come to be authoritatively valuable for them.

The different ways of taking Hegel's thought also apply to the different interpretations given of his mature work in social and political philosophy, the 1820 *Elements of the Philosophy of Right*. Although the book was unfairly criticized both during and after Hegel's lifetime as a mere 'apology' for Prussian absolutism, it was actually an attempt on Hegel's part to articulate the rational form of the kind of reformed, modern European state that modern reformers like Baron von Stein, and later, Prince von Hardenberg, had tried to establish in Prussia. The core idea of the book is that what counts as 'right' in general is what is necessary for the realization of freedom.

Hegel thought of freedom, though, not as the exercise of some special form of causality (as Kant had done) but as a kind of 'self-relation,' as having one's actions come from oneself (always though in terms of such self-relations being mediated by relation to like-minded others); and, quite radically, Hegel held that for 'we moderns,' freedom involves being held to norms that are rationally redeemable. The opposite of such freedom is to be pushed around by factors or considerations not really our own but stemming from something else that we cannot rationally redeem (for example, brute desires or mere social conventions). In Hegel's almost paradoxical way of putting it, one can be an agent only by being recognized as an agent; the conditions of agency are the social conditions under which such agents can reciprocally hold themselves and each other to being certain kinds of agents.

For agents to be free, they must be able to practically reason about their activities, and that requires that they have some conception of some 'good' that they are seeking to actualize. Hegel wished to argue that all genuine practical reasoning has as its major premise some statement about what is ultimately good and best, and when an individual acts rationally, he acts on the basis of some (again, perhaps sometimes implicit) deliberation about what is necessary for him to achieve what is good and best for himself.

For 'we moderns,' that good must be that of 'freedom,' but stated so abstractly, it seems to offer little guidance for deliberation on what would be required to achieve such 'freedom.' To that end, Hegel argued, in the modern world, the realization of freedom must be articulated into three more determinate spheres, which he characterized as 'abstract right,' 'morality,' and 'ethical life' (*Sittlichkeit*). Each of them embodies a way in which very specific institutions and practices underwrite and sustain the ways in which our freedom is actualized in that each of them provides individuals with more concrete, specific first premises about 'the good' (freedom) on the basis of which they may then rationally deliberate what they are required to do.

If we are to have any concrete first principles for moral reasoning, therefore, we must grasp them not as specifications of some 'master rule' but as elements of a social *practice*, ways in which we prereflectively learn to orient and move ourselves around in the social world. Hegel called this sphere, 'ethical life,' *Sittlichkeit*. 'Moral' individuals exercising their 'abstract rights' to life, liberty, and property require a location in these kinds of social practices, which embody within themselves determinate conceptions of what is 'ultimately best,' namely, as the way in which individuals exercise their rights, manage their moral obligations, and come to be 'at home' in the social world by virtue of acquiring a kind of 'ethical virtuosity' in being brought up and socialized in these practices. *Sittlichkeit* thus provides us with determinate principles and a kind of practice-oriented ethical 'know-how,'

and such institutions and practices form what Hegel called 'concrete universals.'

There are three such institutionalizations of *Sittlichkeit* in the modern world, each serving to give individuals a concrete specification of these final goods upon which they can then rationally deliberate. These are the modern family, civil society, and the constitutional state. Together they form a social 'whole' in terms of which individuals orient themselves and which reconciles them to modern social life, gives them good grounds for believing that modern life really is, although imperfect and finite, nonetheless for the best.

Since the modern state appeals to neither God nor natural law for its legitimacy, it must appeal to some sense of what a 'people' collectively establish as rational. This drives political philosophy into a philosophy of history, which, so Hegel argued, demonstrates that from the vague intimations in the ancient eastern states that 'one' (the emperor, the pharaoh) is free, and from the Greco-Roman political conceptions that 'some' (aristocratic males) are free, history has culminated in the modern world's recognition of the principle that 'all' are free. This, Hegel argued, is the *meaning* of world history, and Europe has been its penultimate staging ground. European culture, with its inherent sense of 'negativity,' self-doubt, and skeptical questioning, has by virtue of the kind of culture it is, propelled itself progressively to abandon those criteria, whose lack of rationality made itself manifest until in the modern period it has come to recognize, at least implicitly, that freedom has always been its collective goal, and that such freedom can only be vouched for in an institutional setup much like the one outlined in the *Philosophy of Right*. Freedom itself is thus reinterpreted as a kind of social norm, a historical *achievement*, not the better realization of a potency that, for example, the Greeks also possessed but failed to actualize. (In a completely empty sense of 'potentiality,' such capacities were, of course, always empirically present exactly in the sense that, for example, empirical science or abstract art has always been a human possibility. The kind of 'potency' that is being realized is more like that involved in a narrative than it is like an orthodox Aristotelian realization of a potentiality.)

After his death, in the rather tense atmosphere in Berlin in the 1830s and 1840s, there was a lot of pressure to interpret Hegel in either conservative or progressive terms. The 'July Monarchy' in France and the subsequent revolutionary upheavals in other parts of Europe (along with social unrest throughout Germany) made the ruling authorities anxious while raising the hopes of the reformers. There was a line of thought, especially popular among liberals, to the effect that a kind of historical law was at work that mandated countries to pass through a revolutionary upheaval, establishment of a republic, years of war, and then roughly fifty years later to settle into a constitutional monarchy (first in England, later in France); on that view, England represented the endpoint of modernity, France was almost there, and progress seemed to call for the establishment of constitutionalism in Germany by revolution if necessary; many of Hegel's own students enthusiastically embraced this view of the 'law' of history (despite Hegel's own rejection of such a view). His school thus split famously into 'right' and 'left' Hegelians. The 'right' Hegelians made Hegel into more of a Schellingian and had him espousing a view of history that basically endorsed the (non-constitutional) Prussian monarchy as metaphysically destined, as the only form of social organization compatible with the necessary structure of the cosmos. The 'left' Hegelians, led first by Hegel's friend and follower, Eduard Gans, tended to stress the elements of work and recognition as accomplishing the necessary historical tasks. In the hands of later 'left' Hegelians, however, such as Ludwig Feuerbach, the question began to be raised as to whether it was necessary to have any kind of 'absolute' at all. If history was a matter of labor and recognition, and people's labors were always socially and therefore materially mediated, then why could Hegelianism not be turned into something more like sociology, or into some form of empirical social theory?

The young Karl Marx, coming to maturity in this left-Hegelian atmosphere, transformed Hegel into a materialist thinker. (He was helped in this endeavor by the deservedly forgotten H. M. Chalybäus's characterization of Hegel's thought as following the strict scheme of thesis/antithesis/synthesis, a bowdlerization that has been applied ever since.) Eschewing all talk of the 'absolute' and of 'mindedness,' Marx read into Hegel a kind of evolutionary doctrine of the gradual realization of human capacities that were themselves historically mediated and were themselves the achievements of labor and struggle. The material conditions of life—both its organic demands for sustenance and its social demands for the organization of labor and distribution of wealth—set the terms in which people interact with each other. Hegel's doctrine of the 'struggle for recognition' thus became shorn almost completely of its epistemological bearings and became a doctrine of materialist social reality and history. In the writings of the mature Marx, this was complemented by a theory of how human consciousness is itself socially mediated and is there fore dependent on the ways in which that society is structured, particularly in terms of the relevant forces of production in society and the patterns of ownersihip of the means of production. Famously, Marx also transformed Hegel's philosophy of history into a set of dialectical laws to the effect that revolutions come about when the forces of production and the relations of production come into irrevocable conflict with each other. Hegel had argued that history was not best understood as a matter of divine providence, natural law, or even as the consequence of the decisions of

great men, but rather as a kind of broad scale drama of people collectively trying to come to terms with the people they had come to be in terms of the norms to which they had collectively bound themselves; Marx transformed that into a broad scale story about the ways in which human societies are organized around key material matters such as the production of the necessities for organic and social life, and he held that these follow a law-like pattern in a manner analogous to natural events.

Like Hegel, though, Marx thought that one had to understand social and political phenomena as a whole, even if he ascribed more or less independent importance to the facts about ownership of the means of production; and like Hegel, he put great emphasis on how a given social formation understands itself—what it basically values and counts as authoritative for itself, even though there too he ascribed a more or less independent value to the economic organization of society in determining how such communities arrive at this kind of collective self-consciousness. Marx's own ideas about the historical destiny of the proletariat, of course, went far beyond anything to be found in Hegel's thought. (Later 'Western' Marxists such as Georg Lukacs tried to synthesize Marx's and Hegel's alternative understandings of historical development.)

Simultaneous with Marx's materialistic development of Hegel's thought, there were ongoing 'liberal' developments of Hegel's views. Lorenz von Stein (whose impact on Marx is difficult to prove but who developed certain key 'Marxist' ideas before Marx had developed them himself) developed a kind of liberal post-Hegelian theory of society and history in the 1840s and 1850s through a series of historical writings on the social and political developments in France from 1789 to 1848. Stein argued that 'society' (*Gesellschaft*) had come to assume a prominent role in modernity such that one could best understand the way in which a people comes to reflect on itself not in terms of a 'people' or a 'nation' but in terms of the way in which the 'social' has in modernity separated itself from the 'familial,' the 'political,' and even from 'religion' to make itself central to modern life. The basic norm of freedom had historically generated a sense of individualism that, together with the emerging industrial revolution, made 'social' matters (and, more specifically, economic matters) absolutely central to modern forms of self-understanding. In particular, notions of 'possession' and 'property' assumed all the more importance in the self-understanding of such modern individuals; in 1842 Stein argued that a new social class, the 'proletariat,' had thus emerged within modern 'society,' and the tensions emerging between propertied owners and unpropertied proletarians was to be the central social issue of modern times; since the propertied classes and families would quite naturally seek to take over state power for themselves, it was crucial to construct Hegel's notion of the political state around an idea of a socially oriented state that would secure the fair distribution of the basic goods of society for all its members. (Like Hegel, Stein also thought constitutional monarchy was consistent with such a state.) Retaining the Hegelian emphasis on the 'totality' and the ways in which collectivities understand themselves, Stein nonetheless transformed Hegelian idealism into a kind of quasi-empirical sociology and sociological theory of history; the dynamics of social change and contemporary social structures were to be understood historically in terms of the different modes of self-understanding at play in them; the crucial role of the economy in modern self-understandings was not a timeless natural law (as with Marx) but the result of the kind of self-understandings that modern individualism and industrial society had given rise to. (The next step, although Stein himself did not take it, was to transform Hegel's notion of social practices of family, society, and state into a theory of social *roles*. Hegel's own conception of agency as itself a reciprocally conferred normative status opened the way to understand social action in terms of the more theatrical notion of role-playing, since for Hegel there is no self 'behind' or 'beyond' the deeds that express the norms to which we mutually bind ourselves.) Hegel's own more philosophical worries about how, if all knowledge is socially embedded, the thinker could be in a position to know that or to truthfully assert it were also absent in Stein's transformation of Hegelianism into sociology.

In sum: for Hegel, mindedness, *Geist*, formed a sphere distinct from nature without itself being any kind of non-natural thing; indeed, *Geist* was, for Hegel, no *thing* at all but instead a socially mediated normative status. Hegel's view, moreover, toward the incipient social sciences of his own time was, on the whole, positive; he seemed to think, for example, that Adam Smith and the other Scots had made economics into a science. There is no reason for a Hegelian to deny that there are many interesting causal facts or correlations about societies that can be discovered and established only by empirical investigation. (Thus, the post-Hegelian distinction between the 'understanding' appropriate to the human sciences vs. the 'explanation' appropriate to the natural sciences would be viewed by Hegelians as yet another inadequately dualistic approach.) However, Hegel's key idealist notion—that we best understand societies by understanding how they collectively interpret themselves and the kinds of authoritative values to which they bind themselves and with how those acts of mutual binding must be understood historically—remains forever at odds with all doctrines that posit that some set of basic, natural facts fully determine how we are to understand the course of human events (such as the overall material geography of a place, as the vastly influential historian, Ferdinand Braudel, has argued). Such natural facts, while clearly forming the context in which such collective attempts at self-interpretation occur, cannot fully determine those interpretations. To understand

those interpretations, we must understand why that way of life came to regard those specific norms and not others as binding for them, and to understand that, we must understand how such commitments emerged as necessary in light of the way in which past attempts at collective self-definition had proved themselves to be unlivable. Underlying that is the 'Idea,' as Hegel put it, of freedom, of humanity realizing its destiny in a way of life whose members could be at home in the social world they create, although of course never under the conditions chosen by themselves. For Hegel, all social science only made sense as a smaller part of a larger story about modernity's coming to terms with the radical sociality of our own mindedness and with our very modern awareness of modernity's decisive rupture in human time.

See also: Cognitive Theory: ACT; Emotion in Cognition; Evolution of Cognition: An Adaptationist Perspective; Functionalism, History of; Hegel, Georg Wilhelm Friedrich (1770–1831); Human Cognition, Evolution of; Marr, David (1945–80); Phenomenology in Human Science; Phenomenology: Philosophical Aspects; Situated Cognition: Origins

Bibliography

Adorno T W, Horkheimer M 1991 *Dialectic of Enlightenment* [trans. Cumming J]. Continuum, New York

Hegel G W F 1956 *Philosophy of History* [trans. Sibree J]. Dover Publications, New York

Hegel G W F 1971 *Werke in zwanzig Bänden*. [eds. Moldenhauer E, Michel K M]. Suhrkamp Verlag, Frankfurt, Germany

Hegel G W F 1991 *Elements of the Philosophy of Right* [trans. Nisbet H B; ed. Wood A W]. Cambridge University Press, Cambridge, UK

Honneth A 1995 *The Struggle for Recognition: The Moral Grammar of Social Conflicts* [trans. Anderson J]. Polity Press, Cambridge, MA

Lukacs G 1971 *History and Class Consciousness: Studies in Marxist Dialectics* [trans. Livingstone R]. MIT Press, Cambridge, MA

Marcuse H 1960 *Reason and Revolution: Hegel and the Rise of Social Theory*, 2nd edn. Beacon Press, Boston

Marx K 1959–1974 *Werke*. Dietz, Berlin

Pinkard T P 1994 *Hegel's Phenomenology: The Sociality of Reason*. Cambridge University Press, Cambridge, UK

Pinkard T P 2000 *Hegel: A Biography*. Cambridge University Press, Cambridge, UK

Pippin R B 1997 *Idealism as Modernism: Hegelian Variations*. Cambridge University Press, Cambridge, UK

Stein L von 1921 *Geschichte der Sozialen Bewegung in Frankreich von 1789 bis auf unsere Tage*. Drei Masken Verlag, Munich, Germany

Stein L von 1971 *Proletariat und Gesellschaft* [ed. Hahn M]. Wilhelm Fink, Munich, Germany

T. Pinkard

Hegel, Georg Wilhelm Friedrich (1770–1831)

1. Biographical Data

Hegel, Georg Wilhelm Friedrich the philosopher was born on August 27, 1770, in Stuttgart (Germany), and died on November 14, 1831, in Berlin. In 1788, Hegel, son of a medium-grade civil servant, entered the Protestant Seminary in Tübingen, where he won the friendship of his roommates Hoelderlin and Schelling. He held a ducal scholarship for studying Theology and Philosophy at the University of Tuebingen. After graduating in 1793, he earned a living as a private tutor in patrician families, first in Bern, Switzerland (1793–1796), and then in Frankfurt (1797–1800), where he regularly met Hoelderlin. At the beginning of 1801, Hegel joined Schelling in Jena and qualified for lecturing with his habilitation thesis *On the Orbits of the Planets*. While at Jena (1801–1806), Hegel published his *The Difference Between Fichte's and Schelling's System of Philosophy* (1801). In the years 1802 and 1803, his essays *The Relationship of Skepticism to Philosophy, Faith and Knowledge*, and *On the Scientific Ways of Dealing with Natural Law* came out in the *Critical Journal of Philosophy*, which he copublished with Schelling. The major work of this period was, however, the *Phenomenology of Spirit* (1807), which was intended to be the first part of his system of philosophical science.

After the Battle of Jena, Hegel, having no income, took over as editor of a newspaper in Bamberg in 1807 and one year later he became principal of a secondary school in Nuremberg. There, Hegel married Marie von Tucher with whom he had two sons. During the Nuremberg period, Hegel published his *Science of Logic* in three volumes (1812, 1813, and 1816), the first of which, the *Doctrine of Being*, reappeared in heavily revised edition in 1832. In 1816, Hegel was finally granted the privilege of earning a secured income as a full professor in Heidelberg. In 1818, owing to the personal intervention of Minister Altenstein, he became Fichte's successor in Berlin.

From 1816 on, Hegel published—apart from several minor writings—the *Encyclopedia of Philosophical Science* (1817 in Heidelberg, and in two greatly enlarged editions 1827 and 1830) as well as his *Elements on the Philosophy of Right* (1821). His lectures on Esthetics, the History of Philosophy, and the Philosophy of History and Religion were not only attended by students but also by members of the Berlin society hungry for education. Until the mid-1940s, Philosophy in Germany could not be exercised without taking a position on Hegel's philosophy. Moreover, his philosophy influenced the formation of those special sciences, which Dilthey later called '*Geisteswissenschaften*.' Soon after Hegel's unexpected death, his school split into different wings as a consequence of

philosophical disputes on religion. In 1837, David Friedrich Strauss classified this splitting into the left (e.g., Karl Marx), the right (e.g., Friedrich Wilhelm Hinrichs), and the central (e.g., Karl Rosenkranz).

2. *Aspects of his Philosophy*

2.1 *Interests of the Young Hegel*

The young Hegel studied the ancient authors, the literature of the Enlightenment and above all Kant's critical philosophy (see *Kant, Immanuel (1724–1804)*), Rousseau's works on cultural criticism and theory of politics (see *Rousseau, Jean-Jacques (1712–78)*) as well as the dispute on Spinoza, which had been initiated by F. H. Jacobi. Furthermore, he enthusiastically paid close attention to the course of the French Revolution. At that time, he mainly worked on three questions. Firstly, how must a national religious education be conceived for it to strengthen the motivation for ethical and political action? This question had strongly suggested itself after Rousseau had failed to come to a convincing result concerning the question: which type of religion can be legitimated from the political point of view of a rational community? In his works on religion and ethics, Kant had then shown that a conflict is avoidable between religious and ethical duties on the one hand and religious and political duties on the other. Therefore, what mattered was to translate the idea of such a union of all those duties into an educational program.

Hegel, Hoelderlin and Schelling devoted themselves to finding a corresponding way of life, which they called 'invisible church.' Related to these interests is the second, historical question: what explains the fact that Christian faith developed into a 'positive,' authoritarian religion, which became a tool for despotism to misuse? It was the strategic behavior of the scientific community that gave rise to the third question. In the mid-1990s, the theological Orthodoxy of Tübingen began to misuse Kantian philosophy in an attempt to justify anti-liberal church dogmas in a new manner. This led to the question: how can Kant's critical philosophy be protected argumentatively from such a dogmatic misuse? From his penetrating critique, first of Kant's doctrine of postulates, emerged a system of philosophy in the end whose position may be called 'speculative idealism.' But it did not replace his original interests. Indeed, this system was to provide for their answers.

2.2 *Phenomenology of Spirit*

The *Phenomenology of Spirit* (PS) exhibits the 'way of despair' which 'natural consciousness' has to go in order to justify its knowledge under the conditions set forth by the conscious contrast constitutive of itself: its own subjectivity—made into a first principle—in opposition to the objectivity not reducible to it. On the way, it becomes clear that the knowledge intended by various shapes of consciousness does not satisfy its own, immanent truth criteria. For example, it is regarded as impossible to guarantee the intended immediate knowledge of perceptual data by means of indexical references alone ('sense-certainty'). This self-examination is the first step of the procedure. In a second step, '*we*,' who contemplate consciousness philosophically, have to identify this negative result—taken positively—with a new shape of consciousness. The self-examination is not complete until the contrast of subject and object has been overcome and the contents of the consciousness fulfill its truth criteria. Hegel calls this ultimate shape of consciousness 'absolute knowing,' which is at the same time an undeveloped knowing of the Absolute.

Apart from the much discussed chapter 'Lordship and Bondage'—which exhibits how two people self-consciously struggling for recognition end up developing the specific shape of intersubjectivity of lordship and bondage—the chapters on practical reason, spirit, and religion of the PS should be of some interest to the social scientist. The chapter on *Reason* exhibits that any effort of self-actualization—i.e., gaining pleasure, making inner laws universally valid, and spreading virtue—must necessarily fail. Even a self-conscious practical reason cannot reach its goal of self-actualization if it merely advances the concealed Good of the world instead of rearranging the world in the first place.

The chapter on *Spirit* contains the doctrine of the basic convictions of a society concerning the purpose of society or its central elements. The internal problems of occidental societies—exhibited in their historical order—transformed them into one another, so that—in them—the Spirit reached new basic convictions about itself. The relevant problems, however, are not those which an objective observer of a later stage in history can discover, but those which the members of the societies in question could experience themselves.

In the chapter on *Religion*, Hegel carries out a similar analysis for the religions in history. As Max Weber after him, Hegel believed that the religions above all—*qua* their internally developed convictions—promoted a process of rationalization in world history (see *Weber, Max (1864–1920)*), a process which Hegel comprehended as an advance in the consciousness of freedom. For him, Christianity was the religion, which first gave effect to the principle of subjectivity and thereby pushed through the consciousness that all men as men are free. At the end of the PS it is supposedly possible to expose a concept of cognition based on the historical experiences which natural consciousness has made in the course of

history of occidental societies and religions. This concept, which Hegel called 'absolute knowledge,' has the process of these experiences as its contents and the type of rules concerning the transformation of the shapes of consciousness as its form. It is the purpose of this knowledge to promote humanity, which, in turn, Hegel determines to be the process 'to press onward to agreement with others' at the end of the preface to the PS. The knowledge arrived at in the PS, however, so far only concerns the historical genesis of the present as an epoch which has to bring out humanity. In order to satisfy this demand, knowledge of what emphatically *is* has to be acquired, too. (Preface to the Philosophy of Right.) Since the PS has shown that certain fundamental concepts are the central constituents of each shape of consciousness, Hegel conceives of a doctrine of such 'determinations of thinking' in his *Science of Logic*: It has to examine whether these determinations are true for themselves.

2.3 Science of Logic

To logically unfold the absolute knowledge developed in the PS is what Hegel declared to be 'proper metaphysics.' It is its task to settle an adequate concept of philosophical recognition and thereby of itself ('absolute Idea'). To fulfill this task it primarily criticizes the aims at knowledge of pre-Kantian metaphysics and Kant's transcendental philosophy. In the *Science of Logic* (SL), which is this metaphysics, the traditional concepts of Metaphysics and Logic incapable of defining a concept of the Absolute are improved in an iterated process: abstract understanding begins with determining the traditional concepts in a way that makes them as distinct as possible from other concepts and hence suitable to define the Absolute. In a second step, dialectical reasoning demonstrates that no concept, however well determined by understanding, can be distinguished even from its opposite, and demands to skeptically suspend any judgment in the face of this negative result. But then, speculative reason identifies this negative result with a new term of the history of philosophy, which the understanding tries to determine again (see Sect. 2.4). Since the results of the first two steps are included in the new concept, the concepts of the SL become richer in content in the process. Once dialectical reason cannot find an opposite concept anymore, the process reaches its end. Hegel calls the ultimate concept 'absolute idea,' which is at the same time the concept of the philosophical recognition carried out in the SL (Fulda 1991).

In the final step, this concept is arrived at by overcoming the dualism of our theoretical epistemic attitudes ('Idea of Truth') and our practical ones ('Idea of the Good'). In these attitudes of finite recognition—paradigmatically conceived of by Kant as theoretical and practical reason—we regard the world as being a well-structured touchstone of our knowledge, but also as insignificant and worthless in the face of the Good which lies only in the will of the perceiver. The concept of the absolute idea is intended to deliver us from this conflict between our references to the world. But in order not to give rise to a new conflict between the finite knowledge (prescientific and of the special sciences) and philosophical recognition, Hegel required translation rules mediating between both types of attitudes (see Sect. 2.4) and obliged philosophical recognition of the Real to demonstrate the correspondence of its results to 'empirical appearances' (see Sect. 2.4). According to Hegel, a philosophy of the Real also depends on integrating new expressions from other sciences to denote the developed concepts. To that extent recognizing the Real is underdetermined by the SL, and a philosophy of the Real as well as Logic itself depends on the status quo of human knowledge.

2.4 Philosophy of the Real

The *Encyclopedia of the Philosophical Science in Outline* was meant to instruct the students who attended Hegel's lectures, and therefore only contains the rudiments and basic concepts of all philosophical sciences. The logic outlined above makes up the first part of the *Encyclopedia*, the second and third parts being the philosophy of Nature and Spirit. It is primarily through these two parts that we have access to Hegel's philosophy of the Real.

2.4.1 Philosophy of Nature. In his philosophy of Nature Hegel tries to answer the question of how we can overcome the duality of our theoretical and practical interests concerning Nature and thereby satisfy our needs without turning Nature—and thereby aspects of ourselves—into the Other per se, which would merely become controlled by technology. This philosophy looks for a third way between the modern objectification and quantification of phenomena in the natural sciences on the one hand and a romantic subjectification of Nature on the other. In doing so it does not, however, regard itself as competing with the natural sciences, but rather tries—on the basis of the results of scientific research—to conceive of Nature as an idea in which we can finally recognize ourselves.

2.4.2 Philosophy of Spirit. In so far as Hegel gains his concept of Spirit from his doctrine of living natural beings, which contrasts the sphere of Spirit from that of Nature, he tries to overcome Descartes' dualism between *res cogitans* and *res extensa* in his whole

Philosophy of Spirit, without, however, reducing the phenomena of one sphere to those of the other.

In particular, Hegel teaches in his *Doctrine of the Subjective Spirit* how the subjective spirit of individuals rises step-by-step from its embodiment (Anthropology) and its object-boundedness (Phenomenology) towards higher cognitive and voluntative competencies which culminate in the knowledge of and will to its own freedom (Psychology). In the first part, the *Anthropology*, the phenomena to be comprehended are e.g., dreaming, habits, aging, gestures, and mental disorders, while the *Phenomenology* treats the different levels of consciousness of objects. Sensory consciousness, perceptual consciousness, and comprehending consciousness are all subsumed under the heading 'consciousness,' since these forms of consciousness do not imply any knowledge the subject has of itself. Under the heading 'self-consciousness' are reviewed those forms into which some knowledge of the subject of consciousness of itself necessarily enters (desire, social approval). Finally, 'Reason' is the consciousness of objects, in which it is comprehended that knowledge of general objects is only possible when the subject understands that the functions of his subjectivity constitute these objects in the first place. In conclusion, the higher cognitive performances (intuition, recollection, imagination, memory, and thinking) and voluntative ones (practical emotions, instincts, arbitrary acts, and striving for happiness) are discussed in *Psychology*. Within the 'free spirit,' the highest form of the subjective spirit, the attitude is reached and comprehended in which an individual knows of itself that it is only free and can only realize its freedom when it has overcome the dualism of interests which are merely of a theoretical or practical nature.

The *Doctrine of the Objective Spirit*, which Hegel has also worked out as *Elements of the Philosophy of Right* (PR), comprehends the objective institutions in themselves, which men who know of their freedom give themselves—right in a narrower sense (property, contract, punishment of felonies), moral norms of acting individuals who strive towards the individual or collective Good ('*Moralität*'), and the forms of social life in family, civil society, state, communities of states, and world history ('*Sittlichkeit*'). In order to accomplish this project, Hegel did not try to set up as large a set of consistent principles as possible, but instead conceived of concepts adequate to the subjects in question. The adequacy of the concepts is to be achieved by penetrating the fundamental conceptual material of effective norms and existing convictions of common sense and thus correcting the concepts.

Hegel's analysis of civil society has been especially influential (see *Marx, Karl (1818–89)*). According to this analysis, modern society tends to accumulate great wealth in the hands of a few, alienate workers in the working processes from their products and from themselves, press large parts of the population below the subsistence level, undergo times of over- and underproduction, form global markets, and colonize other countries (Elements of Philosophy of Right (PR)). But at the same time, civil society implements institutions to administer justice, provide for the common welfare, and unite people in trade associations and municipalities. These institutions of society and others of the state make comprehensible how civil societies can develop and stabilize themselves in spite of their inherent tendency to crises. In his doctrine of state, Hegel tried to comprehend institutions which represent the general interests of a rational community and enforce them in the face of the special needs of estates and individuals of the civil society. So the state becomes the authority which alone can guarantee the 'concrete freedom' of the individual of self-determination.

According to Hegel, world history, the doctrine of which makes up the final part of the PR, is constituted by the state. The principle organizing world history is the 'world-spirit,' the process of the developing consciousness that all men are free. In this history, several states have made—to Hegel's mind—decisive contributions to this progress at certain times. In these Periods, Hegel thought, they have the right—and indeed, the power—to politically and organizationally establish and expand a new principle. France was an example of Hegel's time: a country which had carried out a progress in the consciousness of freedom in its Revolution and had expanded this new order through the Napoleonic Wars in great parts of Europe. Since the doctrine of the objective Spirit did not reach the highest form conceivable of freedom, Hegel's philosophy of spirit does not end with it, but rather with the doctrine of the absolute spirit.

Hegel's *Doctrine of the Absolute Spirit* is a doctrine of the three leading forms of culture in which men have at all times tried to understand what is significantly true and who they are within the whole of actuality that depends on such a 'Truth.' According to Hegel, these forms are art, religion, and philosophy. However, they have not always had the same importance since the time they came into being. Each form rather has its particular period in which it becomes the model form to exhibit the Truth. Hegel believed that for the ancient Greeks, art was the medium through which they best understood themselves and were able to express what they cherished most. For the Christians, religion was the best form for men to represent their relation to the Truth and the Truth itself. Both forms, however, have the drawback that their means of explication are at most somewhat discursive—or, indeed, not at all. It is the purpose of philosophy as a science to conceive the content of Christianity's true representations in an argumentative discourse and thereby to gain insight into the deepest reasons for which men come to agreement with each other and afford a stable basis for their life with each other. Therefore, Sect. 2.4 closes with a concept of phil-

osophy, which is also to justify in retrospect the recognition carried out in the system of the philosophical sciences.

3. *Some Lines of Reception in the Social Sciences*

Karl Marx is the most prominent recipient of Hegel's philosophy in the domain of the social sciences; therefore, see *Marx, Karl (1818–89)*.

3.1 *Lorenz von Stein (1815–1890)*

Lorenz von Stein was one of the most politically influential Hegelians among the theoreticians of society in the nineteenth century. With his diverse writings on modern society, he influenced central economic terms of Marx' theory (e.g., surplus value, and accumulation) on the one hand, and considerably influenced Bismarck's social policy on the other. In the sciences, however, he was almost forgotten when Positivism became dominant. It was not until the 1920s that he was studied by theorists of public and constitutional law (Rudolf Smend and Carl Schmitt), a discussion that was continued in the German debates on the principle of the welfare state after the second World War (Stein 1972). Stein took up the question posed in Hegel's doctrine of '*Sittlichkeit*': how must a state be structured for the principles of the French Revolution to be realized considering the conditions of the industrializing society? He regarded the political constellation of his times as an unavoidable antagonism between the civil society based on the satisfaction of individual needs and the state aiming at general freedom. To his mind, this antagonism leads the propertied to try to occupy all state functions through unequal elections, in order to prevent the socioeconomical rise of the nonpropertied. Therefore, Stein thinks it to be the duty of a legitimate state and his institutions (monarch, officials, army) to implement social reforms resulting in a rise of the nonpropertied to the economical level necessary for them to participate in the process of forming a public will. In addition to this, the state should counteract the inequalities of the civil society by introducing universal and equal suffrage and guaranteeing basic rights. To this extent, Stein's doctrine of the state—like Hegel's—aims at more than just the guarantees of the liberal constitutional state, because it does not confine itself to the task of protecting the freedom of the individual from the state, but instead goes further to protecting the political freedom of the individual against the inequalities of the civil society through the state.

3.2 *St. Louis Hegelians*

This American philosophical school was founded by William Torrey Harris (1835–1909) and Hans Conrad Brokmeyer (1828–1906). After the Civil War, this school was institutionalized by the foundation of the *St. Louis Philosophical Society* (1866) and the *Journal of Speculative Philosophy* (1867), the first philosophical journal in the USA. In contrast to the Transcendentalists at Yale and many of the German intellectuals, questions concerning religion did not play an important role for the St. Louis Hegelians. Instead, questions concerning social justice in a modern society were central for them. In addition to this, another one of Hegel's basic interests became important again, on which, however, Hegel had only remarked in passing in his systematic works: how must a pedagogy be in order to make the individuals growing up able to competently participate in forming modern society? Besides being influenced by Hegel, Denton Snider (1841–1925) and Thomas Davidson (1840–1900), two St. Louis Hegelians, were also influenced by the pedagogy of the German Friedrich Froebel (1782–1852) when they realized several educational projects: the *Communal University* in Chicago, the *Chicago Kindergarten College*, the *Goethe School* in Milwaukee, and the *Breadwinner's College* in New York City, which was devoted to the education of the working class. In 1889, Harris became the United States Commissioner of Education. Brokmeyer played a key role in the Constitutional Convention of Missouri in 1875, which guaranteed the right to education for all aged six to 20. Besides these important roles of the St. Louis Hegelians in the formation of the modern educational system of the USA, the *Journal of Speculative Philosophy* offered a platform for the publications of important American philosophers such as John Dewey, William James, Charles S. Peirce, and Josiah Royce. Beyond offering this platform, they successfully animated the study of Hegel's philosophy (Flower and Murphy 1977).

3.3 *Wilhelm Wundt (1832–1920)*

Wundt is commonly regarded as the founding father of Psychology, which established itself as a science around the beginning of the twentieth century. In 1879, Wundt founded the first psychological laboratory of the world in Leipzig, Germany, where he mainly studied sensations and feelings by employing experimental methods. This foundation also became influential outside Germany due to the many visits of foreigners—especially American students and scientists (see *Hall, Granville Stanley (1844–1924)*). Although Wundt often mocks at the methodology of Hegel's philosophy, he explicitly praises Hegel's achievement of 'granting the right to exist to all 'things in themselves,' whether they are to count as metaphysical borderline concepts or as practical postulates' (Wundt 1911, p. 736). Without this basis, it would have been impossible for Wundt to pursue Psychology as an empirical special science, because the soul would have been withdrawn from the applic-

ability of methods borrowed from the natural sciences, as long as it were conceived of as an independent substance removable from the body (Descartes) or were postulated as such on practical grounds (Kant). Hegel, on the other hand, 'has regarded the individual soul as the immediate relationship of the experiences of the soul, as a piece of that infinite reality of the world spirit, which still draws only on its very own actual reality to obtain all that it means' (Wundt 1911, p. 736). So Wundt arrives at the conclusion that Hegel 'had also declared the liberty of Psychology when it proclaimed for all that all mental coming into being—and, therefore, all happening in the soul as well—is actuality, is immediate experienced reality, and that essence and appearance of the Spirit are one and the same, and only mean something different in so far as we comprehend the essence as the relationship of the appearances which is correctly recognized' (Wundt 1911, p. 737; own translations) (see for more *Wundt, Wilhelm Maximilian (1832–1920)*).

3.4 George Herbert Mead (1863–1931)

The posthumously published book *Philosophy of the Present* (1959) primarily promoted an interpretation of Mead's thinking that considers the last phase of it as on the way towards a speculative metaphysics of Hegel's kind. This is surely exaggerated. It is nevertheless possible to prove that Mead named Hegel as an inspirer of his own thoughts throughout his working period. Hegel's philosophy especially stimulated the scientific program of the young Mead. His early outline of a program of philosophical disciplines puts forward two main tasks of Psychology, one of which was inspired by Hegel's PS: Psychology is to give 'an analysis of the objective world in terms of the consciousness of the objective individual' (Mead 1900, p. 12). In pursuing this task, Hegel—according to Mead's view—e.g., proved that the organization of different characters, which constitute 'social objects,' comes from 'our mutual relations' (Mead 1903, p. 106). In contrast to the explaining and describing Psychology, Logic is a normative science (Mead 1903, p. 107), into which 'the general theory of the intelligent act as a whole would fall,' and which is as such of Hegel's kind (Mead 1900, p. 2). However, apart from this praise, Mead also continued to criticize Hegel's doctrines throughout his life: e.g., he falsely subsumed Hegel under the philosophy of romanticism, which, to his mind, only discussed the rationalization of bygone attitudes—quite contrary to the orientation towards the future of the American society and philosophy. Nevertheless, Mead praises Hegel's philosophy of history as '[t]he most grandiose attempt to combine the three great motives of speculation (that of Greek contemplation, that of the church's plan of salvation, and that of experimental science)' (Mead 1964, p. 505) (see *Mead, George Herbert (1863–1931)*).

3.5 Juergen Habermas (1929–)

In his essay *Can Complex Societies Form a Rational Identity?* (Habermas 1976), which Habermas chose as the basis of his speech on the occasion of accepting the Hegel Award of the city of Stuttgart, Germany in 1974, he explicitly classifies his proposal as standing in the tradition of Hegel's and Marx' thinking (p. 118). Although his interpretation of Hegel's answer to the question posed in the title of his essay is affirmative in many points, Habermas mainly criticizes—firstly—that for Hegel, 'only the idea of freedom' is to be conceived as a constituent of modern consciousness. In addition to this, 'the unrestricted objectifying thinking and a radical orientation towards the future' are also constitutive for it (p. 105). Therefore, Hegel did not succeed in abolishing the threefold division between the modern subjectivity and external nature, society, and inner nature (p. 102) in *all* respects. Habermas objects, second, that modern society did not gain its rational identity in the constitutional state; and he objects, third, that philosophy is unable to exhibit this identity as rational for all individuals—as religion has done in the past (p. 106 ff.). It is Habermas' opinion instead that modern societies could gain their rational identity in 'the consciousness of the general and equal opportunities of participation in leaning processes of forming values and norms' (p. 119) by realizing communicative structures 'by themselves,' in whose 'autonomy' alone a new collective identity can crystallize 'today' (p. 120; own translations). Critical remembrance of tradition, science, philosophy, and art propose rival identity projections (p. 121). However, Habermas seems to have partially taken back his criticism of Hegel with his concept of 'constitutional patriotism' (Habermas 1990). In addition to this, his interpretation of the rational function of Right to deliver moral discourse (Habermas 1996) can be understood as a continuation of Hegel's thoughts on the relationship of '*Moralität*' and '*Sittlichkeit*.'

See also: Civil Society, Concept and History of; Civil Society/Public Sphere, History of the Concept; Dialectical Thinking, Psychology of; Hall, Granville Stanley (1844–1924); Kant, Immanuel (1724–1804); Marx, Karl (1818–89); Mead, George Herbert (1863–1931); Phenomenology: Philosophical Aspects; Rousseau, Jean-Jacques (1712–78); Weber, Max (1864–1920); Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Flower E, Murphy M G 1977 The absolute immigrates to America: The St. Louis Hegelians. In: *A History of Philosophy in America*. Putnam, New York, Vol. 2

Fulda H F 1991 Philosophisches Denken in einer spekulativen Metaphysik. In: Pätzold D (ed.) *Hegels Transformationen der Metaphysik*. Dinter, Cologne, Germany
Habermas J 1976 Können komplexe Gesellschaften eine vernünftige Identität ausbilden? In: *Zur Rekonstruktion des Historischen Materialismus*. Suhrkamp, Frankfurt a.M, Germany
Habermas J 1990 *Die nachholende Revolution*. Suhrkamp, Frankfurt a.M, Germany
Habermas J 1996 *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Polity Press, Cambridge, UK
Hegel 1968 ff Gesammelte Werke. Hrsg. In: *Verbindung mit der Deutschen Forschungsgemeinschaft von der Nordrhein-Westfälischen Akademie der Wissenschaften*. Felix Meiner, Hamburg, Germany
Hegel 1969a *Science of Logic*. Translated by Johnston W H (int. al.). Allen and Unwin, London
Hegel 1969b ff *Werke*. In: Moldenhauer E, Michel K M (eds.) *20 Bände und ein Registerband*. Suhrkamp, Frankfurt a.M, Germany
Hegel 1979 *Phenomenology of Spirit*. Translated by Miller A V and introduced by Findlay J N. Clarendon Press, Oxford, UK
Hegel 1983 ff *Vorlesungen: Ausgewählte Nachschriften und Manuskripte*. Felix Meiner, Hamburg, Germany
Hegel 1988 *Aesthetics. Lectures on Fine Art*. Translated by Knox T M. Clarendon Press, Oxford, UK
Hegel 1990 *Encyclopedia of the Philosophical Science in Outline*. Behler E (ed.). Continuum, New York
Hegel 1991 *Elements on the Philosophy of Right*. In: Wood A W (ed.) Cambridge University Press, Cambridge, UK
Hegel 1995a *Lectures on the History of Philosophy*. Translated by Haldane E S. University of Nebraska Press, Lincoln, NE
Hegel 1995b ff *Lectures on the Philosophy of Religion*. Hodgson P C (ed.). University of California Press, Berkeley, CA
Mead G H 1900 Suggestions toward a theory of the philosophical disciplines. *Philosophical Review* **9**: 1–17
Mead G H 1903 The definition of the psychical. *Decennial Publications of the University of Chicago*. First Series, Chicago, Vol. III, 77–112
Mead G H 1959 In: Murphy A E (ed.) *The Philosophy of the Present*. Open Court Publishers, La Salle
Mead G H 1964 *Philosophy of the Act*. Morris C W (ed.). The University of Chicago Press, Chicago
Stein V L 1972 *Gesellschaft-Staat-Recht*. Forsthoff E (ed.). Propyläen Verlag, Frankfurt a.M, Germany
Wundt W 1911 *Grundzüge der physiologischen Psychologie*. Leipzig [*Principles of Physiological Psychology*. Titchener E B (ed.). Sonnenschein and Macmillan, London 1904]

M. Wildenauer

Hegemony: Anthropological Aspects

1. Introduction

By the time anthropologists discovered Antonio Gramsci's idea of hegemony in the late nineteen eighties, it had been reworked by others so that it barely resembled Gramsci's. Although anthropologists cite Gramsci's *Prison Notebooks* (1971) when they refer to hegemony, most are inspired by revisionist interpretations of hegemony.

2. Hegemony

Hegemony is the central idea among many by which Antonio Gramsci established a humanistic, neo-Marxist approach to revolutionary change. Instead of subordinating the superstructure of ideas to the force of the economic base, Gramsci empowered the influence of ideas over Marxism's economic determinism. At the core of Gramsci's methodology is the dialectical relationship between hegemony (consent–support) and domination (coercion–force). His primary theoretical concern was how to resolve this contradiction, for in political practice one is always tempered by the other. He tried to reach resolution in part by redefining hegemony.

Lenin defined hegemony as domination. Gramsci redefined it as an *intellectual and moral leadership* directed by contradictory political and cultural agents and organizations he called *organic* and *traditional* intellectuals. Organic intellectuals represented the interests of working classes. One of their tasks was to make the proletariat aware of capitalist domination and exploitation. Traditional intellectuals represented the interests of the bourgeoisie. One of their tasks was to quell counter-hegemonic resistance to capitalism.

More than other Marxist thinkers, Gramsci recognized the significance of culture in revolutionary practice. Gramsci defined culture as 'the exercise of thought, the acquisition of general ideas *the habit of connecting cause and effect* … enlivened by organization' (Kurtz 1996a, 1996b). Culture was central to Gramsci's political strategy. He argued that culture was a political product of intellectuals, for another of their tasks was to develop the people's culture through education and ideological indoctrination. Traditional intellectuals worked to convince people of the virtues of capitalism and obtain their consent to capitalist rule. Organic intellectuals worked to help people develop an alternative to a capitalist culture. Among other strategies, Gramsci suggested how organic intellectuals might use popular cultural forms and signs to marshal working class revolutionary energies, depose capitalism, and construct a unitary proletarian culture.

3. Revisionists

Of the many works that explore Gramsci's thinking (Kurtz 1996a), anthropologists rely largely on those of Williams (1977) and Laclau and Mouffe (1985).

Williams is concerned with the relationship between culture and hegemony. He argues that hegemony goes beyond culture, is synonymous with culture, serves as a form of 'cultural domination,' and, as a structure of cultural forms, provides a repository for traditional values. But, by equating hegemony with tradition, Williams divests it of Gramsci's political specificity.

Laclau and Mouffe are concerned with the heterogenous social organization of contemporary capitalist nations. They equate hegemony with ideology grounded in discourses by which oppressed categories in capitalist societies—women, gays, ethnic categories, and the like—can be made aware of their oppression. For Laclau and Mouffe hegemony (read ideology) provides the 'organic cement' of political formations that will transform western capitalism into a 'radical socialist democracy.'

These authors appropriated Gramsci's idea of hegemony and made themselves its authority, yet little that they say corresponds to Gramsci's idea of hegemony. They also ignore the role of intellectuals in political practice. Still, the beauty and bane of their interpretations are the extent to which hegemony becomes amenable to diverse interpretations by anthropologists, but sometimes at the cost of the political intent with which Gramsci imbued it.

4. *Anthropology and Hegemony*

Anthropologists were attracted to the idea of hegemony for several reasons. It provided a neo-Marxist approach that complemented the idealist orientation among anthropologists who emphasize the power of ideas over material forces; it provided a new way to explore political relationships; it enabled reconsideration of old ideas, such as culture. Above all, it was thought provoking. Consequently there is little consistency regarding how anthropologists use it.

Most often anthropologists use hegemony as a trendy buzz-word that dilutes the concept's power. The idea of hegemony becomes more complicated when anthropologists use it seriously to analyze political relations. Only rarely do anthropologists use Gramsci's definition of hegemony to explore relations of domination and subordination. Instead, most anthropologists cite Gramsci, if only in passing, as the source of the idea of hegemony. Then they relegate his idea of hegemony to a secondary status and make it a point of departure to use or develop other interpretations of hegemony (Kaplan and Kelly 1994, Kurtz 1996a).

As a tool for analysis, anthropologists most often equate hegemony with a political force in society which usually emanates from those in power. Some anthropologists use revisionist ideas of hegemony to focus on other of Gramsci's ideas, such as *common sense* (peoples' awareness of their condition) (*American Ethnologist* 1993). However they approach hegemony, anthropologists infer that a hegemony of the proletariat to counter domination is rare (*American Ethnologist* 1993, Kurtz 1996b).

When anthropologists explore the relationship between hegemony and culture, it is Williams' ideas of hegemony and culture, not Gramsci's, that informs these explorations. Others use Laclau and Mouffe's idea of hegemony as ideology in conjunction with Williams' ideas of hegemony as culture and tradition to rework Gramsci's ideas. In these instances anthropologists become the source and authority of new ideas of hegemony. Almost no anthropologist places either hegemony and domination or traditional and organic intellectuals in the dialectical context that Gramsci thought to be important in hegemonic encounters.

This diversity of application of hegemony by anthropologists can be explained by their frequent argument that Gramsci did not define hegemony clearly. It is true that Gramsci's *Prison Notebooks* are fragmentary and open to interpretation. But anthropologists also do not read the *Notebooks*' informative footnotes, Gramsci's preprison writings, or the authoritative exegetes of his work. As a result they privilege others' interpretations of hegemony. Regardless how anthropologists define and locate hegemony theoretically, it is most applicable to two related concerns: the consequences of state and colonial policies and political relations of domination and subordination.

4.1. *State–Colonial Policies*

Only a few anthropologists apply Gramsci's formulation of hegemony to explore the impact of colonial policies. Carstens (1991) used Gramsci's idea of hegemony to account for how unequal relations between dialectically opposed intellectuals, British missionaries and Okanagan Indian chiefs, changed Okanagan culture and relegated the Indians to an inferior economic status in contemporary Canada. Kurtz (1996b) extrapolated from Gramsci's idea of hegemony to develop the idea of *hegemonic culturation* to explain why political leaders strive to instill an ideology of work in their political communities and why resistance to these efforts is rare.

More commonly, analyses rely on Williams' idea of hegemony as cultural domination. Fox (1989) used Williams' idea of hegemony as culture to explore how cultural meanings arose in colonial India and compelled colonizer and colonized to conform to certain behaviors. Contrary to Williams, hegemonic agents are central to Fox's idea of hegemony. He demonstrates how intellectuals continually experiment with existing cultural meanings which both enable and constrain resistance to colonial rule. But hegemony is not a source of revolutionary change for Fox. Instead,

as Williams' cultural domination, hegemony continues to subordinate behaviors to the meanings it encodes.

In analyzing cultural confrontations in colonial South Africa, the Comaroffs (1991) emphasize Williams' ideas of hegemony as tradition, and Laclau and Mouffe's idea that discourse has the power to transform indigenous culture. For the Comaroffs, hegemony represents a tradition which, through discourses, acquires an ideology that helped to reorganize indigenous South African societies in accord with colonial rule. Ideology, not hegemony, provided the force of change. As tradition, hegemony had power only to shape actions, perceptions, and changes to the extent that it provided the source of new ideological challenges to existing traditions. As a process, hegemony (read tradition) is reproduced by its interplay with ideology into a functionalist 'hegemonic order' that provides a repository from which cultural forms may be created through the discursive power of ideology.

4.2 Domination and Subordination

Anthropologists who analyze the impact of state and colonial policies on anthropological subjects do not ignore the relationship between resistance and domination. Domination and subordination are implicated in almost all application of hegemony. But anthropologists who emphasize these relations explore them largely in post colonial and authoritarian societies, such as Sri Lanka and Bolivia, through ideas of hegemony that are largely of their own making. They agree only that hegemony is always a process.

Among these explorations hegemony is perceived as a *synthesizing principle* for welding class interests into a coherent world view (McDonogh 1991). In another context hegemony is a *social process* embedded in discourses aimed at instilling an ideology of nationalism (Brow 1988). The persistence of patron–client relations is maintained by hegemony identified as a *force* that maintains a social structure through the cultural shaping of experience (*American Ethnologist* 1993). Hegemony, conceived as a *historical moment*, explains how a ruling class shapes perceptions of ideas, such as female beauty, to legitimate their rule over the working class (*American Ethnologist* 1993). Some anthropologists load the idea of hegemony with an array of theoretical considerations. In one instance, hegemony functions as the source of *education* for Gramsci's idea of 'common sense' (Woost 1993). In another it represents an ongoing *formative process* that may promote resistance to domination (*American Ethnologist* 1993).

Most anthropologists who consider the role of agents in these relations emphasize the role of elites, and argue that proletarian agents are not very effective in developing counter-hegemonies. Even those few anthropologists who emphasize the role of working class intellectuals in confrontations with elites conclude that proletarian intellectuals have little impact on changing the sources of rule and domination. As Laclau and Mouffe suggest, this is because the proletariat is too alienated and submerged under a dominant culture and rule. A logical consequence of this tendency is to privilege hegemony as tradition and render it apolitical.

Although anthropologists use hegemony as understood by others than Gramsci, and often in more convoluted fashion, they tend to agree on what could only be disheartening to Gramsci: the hegemonic impact of colonial and some contemporary governments has sufficiently restructured the culture and society of subordinate peoples that counter-hegemonies to domination are difficult to attain. Hegemony, however conceived, seems to accommodate best the interests of the elites.

5. The Future of Hegemony in Anthropology

The concern of anthropologists with the idea of hegemony peaked in the early 1990s. Now hegemony is firmly ensconced in anthropological discourse and theoretical orientations concerned with culture, structuralism, a poststructuralism *cum* postmodernism, even a revised historical materialism that objectifies ideas as material forces. Regardless how anthropologists conceive hegemony, it holds a privileged place among a cadre of non-materialist, neo-Marxist anthropologists and provides them a humanistic methodology to explore novel cultural and political relations. The collapse of the Western European and, more recently, Soviet empires and the energy of the expanding world capitalist order have produced global disjunctions in social, cultural, economic and political relations. To explain these epiphenomena, anthropologists are likely to continue to develop creatively the explanatory power of the idea of hegemony.

See also: Civil Society/Public Sphere, History of the Concept; Class Consciousness; Gramsci, Antonio (1891–1937); Hegemony: Cultural; Marxism in Contemporary Sociology; Marxist Social Thought, History of

Bibliography

American Ethnologist 1993 20 (1)

Brow J 1988 In pursuit of hegemony: representations of authority and justice in a Sri Lankan Village. *American Ethnologist* **15**: 311–27

Carstens P 1991 *The Queen's People*. University of Toronto Press, Toronto

Comaroff J, Comaroff J L 1991 *Of Revelation and Revolution*. University of Chicago Press, Chicago

Fox R G 1989 *Gandhian Utopia*. Beacon Press, Boston

Gramsci A 1971 *Selections from the Prison Notebooks*. Lawrence and Wishart, London
Kaplan M, Kelly J D 1994 Rethinking resistance: Dialogics of 'disaffection' in colonial Fiji. *American Ethnologist* **21**: 123–51
Kurtz D 1996a Hegemony and anthropology: Gramsci, exegeses, reinterpretations. *Critique of Anthropology* **16**: 103–35
Kurtz D 1996b Hegemonic culturation and work in state formations. In: Claessen H J M, Oosten H G (eds.) *Ideology and the Formation of Early States*. Brill, Leiden, The Netherlands
Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy*. Verso, London
McDonogh G W 1991 Culture and categorization in a turn-of-the-century Barcelona elite. *Cultural Anthropology* **6**: 323–45
Williams R 1977 *Marxism and Literature*. Oxford University Press, Oxford
Woost M D 1993 Nationalizing the local past in Sri Lanka: Histories of nation and development in a Sinhalese village. *American Ethnologist* **20**: 502–21

D. V. Kurtz

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Hegemony: Cultural

The concept of 'hegemony' has provided one of the most influential theoretical frameworks in contemporary cultural and media studies, and Marxist intellectual work in particular. While imprisoned by Mussolini's Fascist regime in the 1920s and 1930s, the Italian Marxist Antonio Gramsci developed the concept of hegemony to address the question of why the working class failed to make a communist revolution in the advanced capitalist countries despite Marx's prophecy. By 'hegemony', Gramsci (1971) intended to conceptualize the pivotal role of cultural persuasion in acquiring, exercising, maintaining, or challenging power without resorting to physical coercion. Stated briefly, the concept of hegemony refers to a historical process in which a dominant group exercises 'moral and intellectual leadership' throughout society by winning the voluntary 'consent' of popular masses.

1. *The Concept of Hegemony*

The concept of hegemony can be more effectively understood when it is compared to Gramsci's notion of 'domination,' as each term represents a different kind of power. Whereas domination rests on a dominant group's coercion of the less powerful, hegemony relies on a dominant group's ability to procure subordinate groups' consent to the existing social order. Domination belongs to the state, which Gramsci called 'political society,' and is equipped with instruments of coercion such as the police and armed forces. In contrast, hegemony takes place in the realm of private institutions—including those outside of the spheres of economic production and the state, such as families, churches, trade unions, and the media—which Gramsci called 'civil society.' Gramsci attributed the political stability and obedience of the working class in Western European countries to the hegemony built into the vigorous civil societies of those countries. Consent and coercion always coexist, but typically, dominant groups in contemporary capitalist societies rule through hegemony unless severe social and political crises require directly coercive interventions.

1.1 *The Concepts of Consensus and Historical Bloc*

The essence of the concept of hegemony lies in the idea of subordinates' consent. According to Gramsci, the consent of subordinates is constructed when a dominant group's particular world view is accepted by the subordinate groups as their world view, and thus the dominant group's interests appear to be the interests of the society at large. As a result, the dominant group is seen as entitled to its ethical and intellectual leadership and arises as a hegemonic group in a particular historical period. Moral and political authority legitimated by popular consent characterizes hegemony as power from below, distinguished from power wielded by coercion or by manipulation from the top down.

However, the paradox of subordinates' consent to their subordination in the long run raises suspicions about the nature of popular consent (Lears 1985). Indeed, some critics of the concept of hegemony have questioned how subordinates can give consent to what will adversely affect their interests unless they are self-deceived in false consciousness. The very idea of 'false consciousness' threatens to undermine the concept of hegemony in at least two ways. First, if one accepts the idea of false consciousness as the basis of popular consent, subordinates' consent is the outcome of sheer manipulation, and Gramsci might be accused of elitism for his implicit assumption that the popular masses are dupes. Alternatively, if one rejects the idea of false consciousness by arguing that people know the purpose of their actions and consistently base their actions on rationally determined self-interest, then the consensus may represent genuine or true consensus. Then, the characteristic of hegemony as dominance over subordinates would be invalidated, and their actual grievance or resistance may slip from view.

Both Jackson Lears (1985) and Stuart Hall (1996a) responded to this dilemma by rejecting the binary opposition of false and true consciousness in understanding the nature of consent. They instead embraced Gramsci's idea of 'contradictory consciousness' grounded in 'common sense.' According to Gramsci, common sense as everybody's spontaneously arrived

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

ff

ff ff

ff

ff

ff

ff

ff

ff

ff

ff

ffi

ff

ff

ff

ffi

ff

ff

ff

identities (Martín-Barbero 1993) (see *Soap Opera/Telenovela*).

Gramsci's concept of hegemony has been continually reinterpreted and applied to different areas of cultural and media studies, depending on the immediate social and cultural questions of theorists in specific social and historical contexts. However, all research relying on the concept of hegemony shares a common pursuit in criticizing culture as a source of control, and, simultaneously, in envisioning it as a source of resistance.

Bibliography

Condit C M 1994 Hegemony in a mass-mediated society. *Critical Studies in Mass Communication* **11**(3): 205–30

García Canclini N 1995 *Hybrid Cultures: Strategies for Entering and Leaving Modernity*. University of Minnesota Press, Minneapolis, MN

Gaventa J 1980 *Power and Powerless: Quiescence and Rebellion in an Appalachian Valley*. University of Illinois Press, Urbana, IL

Gitlin T 1980 *The Whole World Is Watching*. University of California Press, Berkeley, CA

Gramsci A 1971 *Selections from the Prison Notebooks*. Lawrence & Wishart, London

Hall S 1980 Encoding/decoding. In: Hall S, Hobson D, Lowe A, Willis P (eds.) *Culture, Media, Language*. Hutchinson, London pp. 128–38

Hall S 1996a The problem of ideology: Marxism without guarantees. In: Morley D, Chen K (eds.) *Stuart Hall*. Routledge, London pp. 25–46

Hall S 1996b Gramsci's relevance for the study of race and ethnicity. In: Morley D, Chen K (eds.) *Stuart Hall*. Routledge, London pp. 411–40

Hall S, Critcher C, Jefferson T, Clark J, Roberts B 1978 *Policing the Crisis*. Macmillan, London

Hall S, Jefferson T (eds.) 1976 *Resistance Through Rituals*. Hutchinson, London

Hallin D C 1989 *The Uncensored War*. University of California Press, Berkeley, CA

Hallin D C 1994 *We Keep America on Top of the World*. Routledge, London

Hebdige D 1979 *Subculture: The Meaning of Style*. Methuen, London

Herman E S, Chomsky N 1988 *Manufacturing Consent*. Pantheon, New York

Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy*. Verso, London

Lears T J J 1985 The concept of cultural hegemony: Problems and possibilities. *American Historical Review* **90**: 567–93

Liebes T 1997 *Reporting the Arab–Israeli Conflict*. Routledge, London

Lukes S 1974 *Power: A Radical View*. Macmillan, London

Martín-Barbero J 1993 *Communication, Culture, and Hegemony: From the Media to Mediations*. Sage, London

Morley D 1980 *The 'Nationwide' Audience*. British Film Institute, London

Rose T 1994 *Black Noise: Rap Music and Black Culture in Contemporary America*. University Press of New England, Hanover, NH

Williams R 1977 *Marxism and Literature*. Oxford University Press, Oxford

Willis P E 1977 *Learning to Labor*. Columbia University Press, New York

S. Kim

Heider, Fritz (1896–1988)

Fritz Heider was born in Vienna on February 19, 1896, spent his childhood in Graz, and received a classical education at the official state Gymnasium. Unable to serve in the army during World War I because of the loss of vision in one eye due to an injury during his childhood, he entered the University of Graz (where ironically at one point he studied depth perception). He did not have a professional goal in mind, and in the European tradition attended a wide range of lectures open to all University students. His interests eventually centered on the study of philosophy and psychology (in opposition to the pursuit of architecture desired by his more practical father).

As a student, Heider met Alexius Meinong, an influential figure in European philosophy, who suggested the topic of Heider's doctoral thesis with the obscure question: 'Why do we say we see the house, rather than saying that we see the sun, for the reflected rays of sunlight are the stimuli that actually strike the eye?' In discussing this problem, Heider distinguished between 'thing,' or the objective structure in the environment, and 'medium,' or that in the environment providing information about the object. This theme reappears in much of his work on person perception. He believed in a 'real' world, with the task of the observer being to come to know this reality. In Graz, Heider was also influenced by the psychologist Vittorio Benussi, one of the first people to publish experiments in the field of gestalt perception.

After finishing his dissertation, written during 1919, Heider worked for a year as an applied psychologist. But he soon became restless and moved to Berlin. At the suggestion of Bernussi, he made contact with the psychologists at the University of Berlin. An air of excitement permeated the Berlin Department of Psychology. The lectures in perception, especially those of Max Wertheimer, were very popular among students and intellectuals. There was a feeling that Gestalt principles would have an important influence on the development of psychology. Heider attended the lectures of Wertheimer, as well as those of Wolfgang Köhler. He found many Gestalt concepts, such as laws governing 'unit formation,' useful in his later work in person perception and interpersonal relationships. He also formed a lasting social and intellectual

friendship with Kurt Lewin, at that time a younger member of the Berlin faculty.

After Berlin came the *Wanderjahre*, a period when Heider spent a good deal of time meandering and reading philosophy, especially Spinoza and Nietzsche, as well as psychology and literature. Then, in 1927, he became an assistant to Wilhelm Stern at the University of Hamburg. In 1930, supported by a recommendation from Stern, Heider took leave from Hamburg to accept what he thought of as a one-year position with the Gestalt psychologist Kurt Koffka, who was on a research fellowship at Smith College in Massachusetts, USA. Heider's principle work was at the Clark School for the Deaf, where Grace Moore was also a member of Koffka's research group. They were married shortly thereafter and Heider remained in the USA.

Heider settled at Smith College as an Assistant Professor until 1947, where he combined research at the school for the deaf, teaching, and raising three sons. In 1947, he accepted an offer from the University of Kansas as Professor of Psychology. Roger Barker, the new Chairperson of the Psychology Department, had brought with him a group of people like himself who had been closely associated with Kurt Lewin. Then, with the aid of Beatrice Wright, Heider completed his long-awaited book, *The Psychology of Interpersonal Relations* (1958). A number of awards followed, including the Distinguished Scientific Contribution Award from the American Psychological Association. Fritz Heider died in Lawrence, Kansas, on January 2, 1988 at the age of 92. He was a gentle man with a sly twinkle in his eye, shy and introspective, with a love of art and literature, and a wide-ranging intellectual curiosity.

1. *Scientific Contributions*

Unlike other influential psychologists, Heider published few articles, did little experimental work, had few students, and was removed from the mainstream of American psychology. He apparently sought neither power nor fame. Nonetheless, his book *The Psychology of Interpersonal Relations* (1958), published when he was 62 and nearing retirement, had enormous impact. In this work, Heider applied the laws of the perception of physical objects and events developed by Gestalt psychologists to the perception of persons and their actions. Synthesizing object and person perception is one of Heider's important scientific accomplishments.

1.1 *Balance Theory*

Two central and associated thematic statements were most fully articulated in *The Psychology of Interpersonal Relations* (1958). One of these relates to the concept of balance. Balance theory as formulated by Heider should be considered within the context of Gestalt views regarding the perception of objects. Gestaltists contended that there is a tendency toward simplicity in the perceptual system such that simple structures are preferred. For example, faces are perceived as symmetrical although this is not exactly the case, and even when a circular form is not quite closed, it is nonetheless perceived as a circle. Furthermore, parts of a system become organized and the groupings form a particular structure in an uncontrollable and inevitable manner. For example, two similar objects in close proximity to one another are joined to form a unit, as in the following example: ** ##. In the perceptual field, organisms impose particular arrangements.

Heider reasoned that object configurations, including social objects, could be classified into those that are in a balanced or in an imbalanced state. Balanced states are preferred to imbalanced states because they form 'good' figures or simple structures. For example, in a two-person system, if a likes b and b likes a, then the system is in balance. But if a likes b while b does not like a, then there is disequilibrium—imbalance exists and there is a force to change in the direction of balance. This can be brought about by changing attitudes toward the other, misperceiving the sentiments of the other, changing the attitudes of the other, and the like.

Configurations with three entities, such as those involving two persons and an object, are more complex. If, for example, a likes b as well as object x (e.g., the elected president), then if b does not like x, the situation is again imbalanced and a force will be created to bring the system into a state of equilibrium by, for example, a change in one's attitude toward b or toward x. In Heider's characteristic manner of integrating everyday observations, literary examples, and scientific analysis, he noted that many literary tragedies make use of balance principles (e.g., Romeo loves Juliet, Juliet loves Romeo, Juliet loves and obeys her parents, but her parents do not accept Romeo). An example of 'latent' imbalance, which contributed to the dramatic tension in Ibsen's Wild Duck, occurs when a scrupulous husband loves his wife who, unbeknownst to him, has engaged in a dishonest act for his benefit. Here we can see the relations between object perception, unit formation, the preference for particular arrangements, and social behavior.

Balance principles have been widely applied in the study of interpersonal relations, attitude change, social cognition, and other areas of research. They also provided a new approach in motivation inasmuch as motivation to change was derived from the configuration of cognitive elements, rather than from viscerogenic need states, which were involved in most studies of motivation. In accord with Gestalt principles, the whole pattern of the structure determines its motivational effects.

1.2 Causal Attributions

Causal attributions, or beliefs regarding the causes of events, were the second major focus in *The Psychology of Interpersonal Relations*. Attribution theory was even more impactful than Heider's balance ideas, and became the dominant theme in social psychology for nearly fifteen years, between 1970–1985.

Heider postulated that people are motivated to understand and to master their environment—understanding is adaptive and instrumental to future behavior, and also there is a basic curiosity and desire 'to know.' Understanding and mastery require knowledge of the causes of events.

Pursuing the distinction in his thesis between thing versus medium, Heider stated that we are often only in contact with immediate facts or raw date (the 'medium') but we search for the underlying core processes or dispositional properties to explain these facts (the causal 'things'). These perceived causes are generally enduring aspects of the world. Covariation information regarding the presence and absence of the inferred cause and the effect are important sources of causal knowledge. But when explaining the actions of persons, Heider reasoned that we often ascribe their behavior to some stable dispositional quality or trait, rather than to situational factors. The underestimation of the situation as a perceived cause of the behavior of others, and over-attribution to the person, subsequently was labeled 'the fundamental attribution error.'

Although the topics of balance and attribution have been treated separately in the psychological literature, Heider saw them as intertwined. He reasoned that balance principles guide causal attributions. Thus, we infer that a good act was done by a good person, and ascribe success more than failure to the self (the 'hedonic bias'), for in both of these examples the inferences result in balanced states or simple structures.

1.3 Scientific Method

The sources of Heider's ideas were largely observational and literary: he often relied upon 'thought experiments' in which interpersonal situations were imagined and altered in a systematic manner (e.g., assume Romeo does not like Juliet, or that Juliet does not care about her parents). The psychological consequences of these new cognitive configurations were then inferred. This method may seem atavistic given the movement of psychology toward an experimental science, but it proved fruitful for Heider.

Heider also made use of a second methodology: Examining common speech in order to capture *a priori* logical relations. Concepts such as belonging, benefit, harm, and sentiment were subject to his logical analysis. This was particularly the case regarding perceptions concerning the causes of success and failure, where he articulated the relations between can and try as causal determinants of achievement outcomes. Because of his mental experiments (rather than manipulation of variables and responses by other), and the use of everyday language in his theoretical analysis, Heider is often labeled a 'naïve' psychologist. This characterization is not entirely correct inasmuch as his theoretical analyses embrace abstract principles that transcend any one concrete situation, providing laws that a layperson could not articulate.

2. Contemporary Relevance

As already indicated, attribution theory dominated social psychology in the 1970s. However, its influence did not end there. It remains a presence in social psychology and has made deep inroads as well in the fields of clinical psychology, motivation and emotion, cognitive psychology, educational psychology, and personality. Furthermore, it is almost impossible to discuss interpersonal relationships without making some reference to the concepts of balance or attribution. When assessing the contemporary impact of Heider, it must be recognized that this influence is now carried through more contemporary figures in social psychology who interpreted and elaborated Heider's ideas, while adding their own insights. Among the figures one generation removed from Heider are Edward Jones, Harold Kelley, and Bernard Weiner, three among the authors of the book: *Attribution: Perceiving the Causes of Behavior* (1972). More recently, other researchers two generations removed from Heider have, in turn, replaced these figures as leaders in their field.

What follows is a sampling of some of the attribution-related research areas outside of social psychology that can be traced to the foundation laid by Heider. Rather than assess the 'truth' of the various hypotheses that have been suggested, these are presented to show the presence of research activity stimulated by Heider's ideas.

2.1 Clinical Psychology

There is substantial research examining whether causal ascriptions are precursors to depression. It has been contended that when the causes of aversive outcomes are perceived as unchanging, unchangeable, and/or implicate the self (i.e., there is learned helplessness), then depression may follow. Furthermore, perceptions of the causes of dysfunctional symptoms by others may contribute to the course of the illness. Critical and hostile emotional expressions, spawned by causal beliefs that the other is responsible for the illness, affect the recidivism rates of schizophrenics as well as depressives.

2.2 Aggression and Prejudice

A number of different forms of aggression have been subject to attributional analyses. Some types of child abuse, wife battering, and marital hostility have been traced to the aggressor ascribing intentional misbehavior to the other (e.g., 'My child refuses to obey me;' 'She purposively made me a bad meal'). In general, aggressive individuals tend to over-attribute intentionally to others when experiencing a negative outcome. Prejudice also is a product of (or may give rise to) different attributions for in-group and out-group members. For example, aggression by an out-group member may by perceived as controllable by others and intentional, thus producing further out-group derogation.

2.3 Coping

Following a negative life stressor, such as rape, unwanted pregnancy, or illness, there often is a search for causality ('Why me?'). Some individuals ascribe the aversive event or outcome to their character ('I am too impulsive'), whereas others place blame on a particular behavior ('I was at the wrong place at the wrong time'). Those causal beliefs elicit different coping mechanisms, with the more adaptive behavioral blame resulting in active coping, whereas the less adaptive avoidance coping is in part evoked by attributions to properties of the self that are unchanging.

2.4 Motivation and Emotion

A large literature documents that attributions for achievement success and failure have important consequences on subsequent striving. Guided by the can/try distinction articulated by Heider, it has been found that attributions for failure to low aptitude are particularly debilitating, in contrast to ascriptions to insufficient effort, which promote action. In a similar manner, the belief that ability is changeable is more adaptive than the belief that one's ability level is fixed. In addition, reactions of others are guided by beliefs about the causes of failure. Non-attainment of a goal because of lack of aptitude elicits sympathy and help from others, whereas failure due to low effort gives rise to anger and neglect.

A great deal of research under the rubric of appraisal approaches to emotion has contended that how one feels is determined by what one thinks. As intimated above, beliefs about the causes of events comprise an important aspect of this literature. For example, guilt is often preceded by the belief that one 'could and should have done otherwise' whereas shame often follows from the causal inference that the causes of personal failure mark one as different from others and cannot be changed. Anger, sympathy, regret, and gratitude also are typically preceded by specific causal beliefs.

2.5 Cognitive Psychology and Psycholinguistics

Many researchers are examining the antecedents of causal beliefs, ranging from individual differences in cognitive preferences to the role of covariation information and the probabilistic co-occurrence of antecedents and consequences. Not all research attempting to understand the contents and process of causal inferences can be traced back to Heider. Nonetheless, some of this research is directly related to his thinking. Linguists have pointed out that verb forms also implicate causality. For example, the phrase 'I admire him' results in causality residing in the admired person, whereas 'I hit him' places causality in the person committing the act. Some people interested in gender issues have pointed out that journalists may influence lay causal beliefs when they state: 'The flirtatious woman was raped by him' as opposed to 'He raped the flirtatious woman.' In sum, it is evident from this very select review that, like his ideas about perceived interpersonal relations represented in balance theory, the understanding of perceptions of causality, represented in attribution theory, had a broad influence in psychology. Fritz Heider, an armchair psychologist with knowledge of philosophy and literature, and armed with theoretical acuity, has left a flourishing psychological legacy.

See also: Attributional Processes: Psychological; Authoritarian Personality: History of the Concept; Causation (Theories and Models): Conceptions in the Social Sciences; Emotion: History of the Concept; Khler, Wolfgang (1887–1967); Lewin, Kurt (1890–1947); Psychology: Historical and Cultural Perspectives; Social Psychology; Social Psychology, Theories of

Bibliography

Heider F 1927 Ding und Medium. *Symposium, Philosophische Zeitschrift fur Forschung und Ausprache* **1**: 109–57

Heider F 1944 Social perception and phenomenal causality. *Psychological Review* **51**: 358–74

Heider F 1946 Attitudes and cognitive organization. *Journal of Psychology* **21**: 107–12

Heider F 1958 *The Psychology of Interpersonal Relations*. Wiley, New York

Heider F 1976 A conversation with Fritz Heider. In: Harvey J H, Ickes W J, Kidd R F (eds.) *New Directions in Attribution Research*. Erlbaum Hillsdale, NJ, Vol. 1, pp. 3–21

Heider F 1983 *The Life of a Psychologist: An Autobiography*. University of Kansas Press, Lawrence, KS

Heider F 1988 *The Notebooks of Fritz Heider*. M. Benesh-Weiner (ed.) Psychologie Verlags Union, Munchen-Weinheim, Germany, Vols. 1–6
Jones E E, Kanouse D E, Kelley H H, Nisbett R E, Valins S, Weiner B 1972 *Attribution: Perceiving the Causes of Behavior*. General Learning Press, Morristown, NJ
Weiner B 1995 *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. Guilford, New York

B. Weiner

Helmholtz, Hermann Ludwig Ferdinand von (1821–94)

1. *Life*

Hermann Ludwig Ferdinand von Helmholtz was born in Potsdam, Germany on 31 August, 1821. His father taught humanistic subjects at the local gymnasium, and encouraged Hermann's keen interests in music and the arts generally, in enjoying nature, and in philosophy. At the gymnasium Helmholtz distinguished himself in every subject of study. Although his great intellectual love was physics, his father persuaded him that there was little call for professional physicists and that, moreover, he could not afford to send Hermann to university. Helmholtz thus enrolled (in 1838) at the Königliches medicinisch-chirurgisches Friedrich-Wilhelms-Institut in Berlin, the Prussian military medical school. The school provided full support to its students in exchange for eight years of military service.

After graduating in 1842 and serving a year as an intern in Berlin's Charité hospital, Helmholtz served nearly five years (1843–8) in Potsdam as an army surgeon. The position afforded him much free time, and he spent it publishing papers in physiology and conducting research in physics. In 1848–9, he served as Johannes Müller's assistant and as an anatomy instructor in Berlin, before being released early from his remaining military obligations in order to begin an academic career as Extraordinary Professor of Physiology at the University of Königsberg.

While at Königsberg, Helmholtz invented the ophthalmoscope (1850). It soon revolutionized ophthalmology by allowing doctors to examine the living retina, and thus diagnose the eye far more thoroughly and help treat it for disease. It also led to Helmholtz's becoming a full professor, boosting his reputation among the Prussian educational authorities, and giving him a European-wide reputation. In 1855, he transferred to the University of Bonn, where he spent the next three years as Professor of Anatomy and Physiology. In 1858, he accepted Baden's offer to become Professor of Physiology at the University of Heidelberg. A year after arriving there, his wife, Olga, died, leaving him a widower with two young children to raise. He soon married Anna von Mohl, with whom he had three children and who helped introduce him to the world of German high society. The university built him a new institute of physiology; he served as the university's prorector in 1862–3, and, more generally, became one of Heidelberg's stellar figures during its golden era.

With the establishment of the German Reich in 1871, the Prussian authorities were determined to reacquire Helmholtz as part of their effort to turn the University of Berlin into the centerpiece of German academic life. Helmholtz felt he could not refuse the call to Berlin, where he now became Professor of Physics and Director of the Institute of Physics. Prussia built him an entirely new physics building, the largest such in Germany; his institute became a mecca for German and foreign students alike. In going to Berlin, Helmholtz also left physiology and devoted himself almost exclusively to physics. In Berlin, Helmholtz became an even more prominent figure in German academic and intellectual life. Since the early 1850s, he had regularly been called upon to give popular lectures on science, which he continued to do throughout Germany and in Britain. His home became one of Berlin's leading *salons*. He regularly served the Prussian and Reich governments as a representative at national and international scientific meetings. In 1883, the German emperor ennobled him for his services. Four years later, he was appointed the founding President of the Physikalisch-Technische Reichsanstalt (PTR), a nonacademic government institute devoted to research in physics and technology, including the setting of physical standards and testing. He spent the last seven years of his life directing the PTR and pursuing theoretical physics. He died in Berlin on 8 September, 1894.

2. *Scientific Contributions Pertinent to the Social Sciences*

Helmholtz's contributions to science concentrated above all on physics and physiology. He produced over 150 scientific papers, which were reproduced in his three-volume *Wissenschaftliche Abhandlungen* (1882–95); a three-part *Handbuch der physiologischen Optik* (1856–67) that analyzed and synthesized the results of physiological optics; a pathbreaking musicological treatise, *Die Lehre von den Tonempfindungen, als physiologische Grundlage der Theorie der Musik* (1863); and a two-volume collection of some twenty-five popular and philosophical essays (*Vorträge und Reden*, 5th edn., 1903) on various aspects of science. The hallmarks of Helmholtz's efforts in science included a gift for generalizing and synthesizing results;

using the methods or results of one discipline to advance another; showing the importance of instrumentation and measurement for advancing science; turning philosophical problems into scientifically addressable questions; analyzing the epistemological foundations of science; relating science and art to one another; and popularizing science. Although Helmholtz was not a social scientist, several highly selective aspects of his work in physics and physiology merit attention for their pertinence to the social sciences, above all psychology.

2.1 The Conservation of Energy and Socioeconomic Thought

In the mid-1840s, Helmholtz and several other scientists sought to establish an 'organic physics,' i.e., to explain life processes exclusively on the basis of the laws of physics and chemistry, and thereby rid physiology of the vague notion of 'life force.' Helmholtz centered his early scientific efforts on the problem of heat and its mechanical equivalent. In 1847, he published his epoch-making conception of the principle of conservation of force. Although he was only one of several scientists in the 1840s who conceived of the idea of force conservation in physical or physiological systems, his presentation offered by far the most systematic, mathematical, and general account, thereby helping to unite various disparate branches of physics. The work was part of his larger program of providing a mechanical interpretation of nature, organic as well as inorganic. Nonetheless, Helmholtz never maintained that human beings, whom he believed to possess free will, could be understood solely in material and mechanical terms.

During the 1850s and 1860s several physicists transformed Helmholtz's principle of force conservation into the first law of thermodynamics. This new energy-based physics and the associated concept of 'work' generally inspired a number of economic and social thinkers in the second half of the century. Thanks in good measure to Helmholtz's work, physical scientists and social thinkers alike now unequivocally recognized the impossibility of perpetual motion. Social and economic thinkers began to see both the individual human body and socioeconomic systems in general in terms of individual or societal energy production, consumption, transformation, and output. Helmholtz helped give economists a greater sense of the importance of the conservation and general circulation of energy. His work in thermodynamics also helped deepen understanding of the human body as a mechanical system, a human motor. The new thermodynamic viewpoint helped give new meaning to the terms 'work,' 'labor,' and 'value' (see Rabinbach 1992, Mirowski 1989).

2.2 Nerve–Impulse Measurements

Helmholtz's work on the propagation velocity of the nerve impulse spoke directly to the interests of the emerging set of experimental psychologists. In 1849–50, Helmholtz conducted a series of experiments on the propagation velocity in frogs, demonstrating that the impulse was propagated at a finite, and measurable, velocity. He and others saw that reflex nervous action could also be measured in humans, and that there was a small, noninstantaneous time interval between human thought and bodily response.

As a result, the mind–body problem took on a new dimension, and a vigorous program of stimulus–response studies in physiological psychology began. By implication, Helmholtz's work provided a measure of a subject's reaction time to sensory stimulation, e.g., to visual or auditory cues, to voluntary movements, and to the association of ideas. Wilhelm Wundt and others built on Helmholtz's work by constructing methods for measuring the time intervals in thought processes. Helmholtz's work, along with Gustav Theodor Fechner's on psychophysical experiments, and Wundt's mark an important milestone in the emergence of experimental psychology as a field. It meant that at least some psychological phenomena could be treated quantitatively and become the subject of laboratory research.

2.3 Physiological Acoustics and Optics

Beyond neurophysiology, Helmholtz's importance to the emergence of experimental psychology lay generally in his work in physiological acoustics and optics, especially in his two great books, the *Handbuch* and the *Tonempfindungen*. These two books became the standard reference works for psychologists and others. Both had a similar structure: they first gave analyses of the physical aspects of their respective subjects, physiological optics and acoustics; then physiological analyses of the respective sensations of light and sound; and finally arguments for the importance of the psychological factor as mentally integrating sensations into experienced perceptions or sounds. In particular, Helmholtz believed that unconscious inferences transformed the physical and physiological sensations received by the bodily organs and nervous systems into perceptions. From birth onwards, the individual learned through empirical experience that certain types of behavior produced certain types of responses, which, in the course of time, the mind transformed into unconscious inferences.

Helmholtz's work in physiological acoustics and optics began with his theory (borrowed from Müller) of specific fiber energies. Müller had maintained that the stimulation occurring in each type of sensory organ is responsible for a specific sensory experience; each type of sensory nervous system is qualitatively

different from the others. Helmholtz extended this theory for vision by claiming, as had Thomas Young before him, that there were three sets of optical nerve fibers, each of which provided its own specific nerve energy that corresponded to the primary colors. Similarly, he extended Müller's theory for hearing by maintaining there were several thousand specific auditory energies. He announced his so-called resonance theory of hearing in the *Tonempfindungen*, showing that the ear is a set of resonators that analyzes homogeneous sounds into their components. Although he thought that sensory processes of hearing and seeing were in good part mechanical ones, he did not think that sensations were copies, let alone mirror images, of the external world; rather, they were 'signs' of it. Although Helmholtz's theory of specific energies was modified or rejected by some, it came to constitute part of the foundations of experimental psychology.

2.4 *The Nativist–Empiricist Controversy*

Helmholtz was not only one of the two principal antagonists in the nativist–empiricist controversy over the proper foundations of vision research. He also coined the labels of these two fundamental positions and did much to shape the controversy between himself and Ewald Hering, and their respective followers. Although their differences were many, the two sides mainly disagreed on whether the eye possessed and required a mind to see. The controversy began in the 1860s and lasted until the 1920s, though in several ways it has continued down to the present.

In the *Handbuch*, Helmholtz labeled Hering and his followers 'nativists,' since they claimed that spatial perception was innate to newborns. By contrast, he called himself and his followers 'empiricists,' since they claimed that spatial perception was an acquired phenomenon, one that each individual gradually learns through daily experience and which requires the mediation of unconscious inferences. The two sides also disagreed about the receptor mechanisms affecting color vision, and about explanations for contrast phenomena, adaptation, optical illusions, eye movements, the constancy of color, retinal correspondence, and so on.

The controversy helped shape the intellectual substance and disciplinary formation of physiology and psychology. Experimentally minded psychologists like Carl Stumpf, Georg Elias Müller, and Franz Brentano opposed the empiricist view insofar as it employed unconscious inferences and so harkened back to psychology's roots in philosophy. On the other hand, nativism's stress on the biological foundation and explanation of perception threatened to replace psychology with physiology. Yet nativism also emphasized the role of phenomenological experience, and that appealed greatly to a number of experimental psychologists, including Stumpf, Hermann Ebbinghaus, Oswald Külpe, Erich Jaensch, and David Katz.

3. *Impact on Psychology and Beyond*

Unlike Fechner and Wundt, Helmholtz had no desire to create a new, institutionally independent discipline of psychology. His contributions and overall importance to the discipline came largely (and indirectly) from his work in physics and physiology, and from his general prestige. His contributions to the 'New Psychology'—a movement that arose in Germany between 1860 and 1910 and that was associated with names like Wundt, Ebbinghaus, Külpe, and Stumpf—consisted largely in the observational and experimental study of, and speculation about, sensations and perceptions. The New Psychology in good measure replaced the older philosophical psychology, and in so doing adopted, adapted, or extended many of its methods and results from sensory physiology and psychophysics, not least those introduced by Helmholtz.

Helmholtz's work deeply influenced and must be seen against the emergence and course of the New Psychology. During the 1860s and 1870s, a number of psychologists aligned themselves with (sensory) physiology, and it was during these decades that Helmholtz had his greatest influence on the field. For example, Wundt, who was Helmholtz's assistant at Heidelberg, shared Helmholtz's empiricist theory of perception, and Helmholtz later drew on Wundt's work on vision for his *Handbuch* and supported him professionally. As advocates of psychology as an experimental discipline, they were allies. Yet there were also differences between the two: Wundt later rejected Helmholtz's mechanistic approach to sensation, the two men had a priority dispute concerning the theory of unconscious inferences, and Helmholtz criticized several of Wundt's views. The young Ebbinghaus was similarly influenced and supported by Helmholtz. Helmholtz's work on the measurement of the velocity of nerve impulse served as a starting point for Ebbinghaus's famous work on measuring memory. Helmholtz was one of Ebbinghaus's examiners for the right to teach at Berlin, where Ebbinghaus lectured in psychology (1880–94) and where Helmholtz supported him in other ways.

In 1890, when Ebbinghaus and Arthur König founded the *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, they included Helmholtz on their editorial board. The new journal played a leading role and constituted an important milestone in physiological psychology. Yet the discipline of psychology was itself changing: as it gradually separated from physiology, a number of its practitioners gradually abandoned its elementalist, sensory-based approach in favor of the phenomenalistic approach of the

Gestalt psychologists (especially after 1920), who rejected Helmholtz's mechanistic approach to psychology and, like the nativists, his theory of perception. In short, from the 1890s psychology moved away from Helmholtz's views and gained a new sense of intellectual and institutional autonomy.

Helmholtz also had a strong impact on other young psychologists, two of whom warrant special mention. As a medical student and young doctor, Sigmund Freud was especially taken with the program of the 'organic physicists.' (He had once hoped to study with Helmholtz, though nothing came of this.) Freud sought to link neuropsychological to physiological phenomena, and his early psychological language is the language of physiology (e.g., that of 'energy'). He called Helmholtz 'one of my idols.'

Likewise the young William James had intended to study with Helmholtz but never did. He called Helmholtz 'immortal' and said that he was 'perhaps the first scientific genius now [in 1868] above ground.' Helmholtz's work and views on physiological psychology constituted an important source for James: as a young scientist, he assiduously studied Helmholtz's *Handbuch*, the *Tonempfindungen*, and the popular and philosophical essays. Helmholtz was one of the most-frequently-cited authorities in James's *Principles of Psychology*. Yet although James often cited him positively, especially as a source of empirical evidence, he rejected Helmholtz's notion of unconscious inference and his sensation-based theory of perception, and he found Helmholtz's mechanistic and mathematical predilections in science alien to his own. James wrote of the *Handbuch*: 'Can I find fault with a book which, on the whole, I imagine to be one of the four or five greatest monuments of human genius in the scientific line? If truth impels I must fain try, and take the risks. It seems to me that Helmholtz's genius moves most securely when it keeps close to particular facts.'

Beyond psychology, Helmholtz also affected several of the future leaders of the next generation of social science in ways that are perhaps impossible to state. The young Franz Boas had originally intended to become a physicist and to study with Helmholtz, but he soon turned away from physics and towards geography before settling on anthropology. (Helmholtz did, however, serve as one of Boas's examiners when he [successfully] sought to obtain the right to teach at the University of Berlin.) The young Wilhelm Dilthey was a frequenter visitor to the Helmholtz home in Berlin. Although Dilthey's holistic and dynamic theory of consciousness stood in sharp contrast to Helmholtz's sensationalist approach, he studied Helmholtz's writings carefully; considered Helmholtz's popular scientific essays as the best such set that had ever appeared in German; engaged him in friendly discussions of their philosophical differences; and graced his *Einleitung in die Geisteswissenschaften* (1883) with an opening quote from Helmholtz.

By the 1920s Helmholtz's substantive achievements and outlook had lost their attraction for some psychologists. Nonetheless, for many in Germany and abroad, Helmholtz still stood, as he did for Max Weber in 'Science as a Vocation' (1919), as a model scientific researcher. And as late as 1950, Edwin G. Boring, in the second edition of his monumental *A History of Experimental Psychology*, ranked Helmholtz, along with Darwin, James, and Freud, as one of the four most important figures in the history of psychology.

See also: Empiricism, History of; Experimentation in Psychology, History of; Fechnerian Psychophysics; Neuropsychological Functioning, Assessment of; Psychology: Historical and Cultural Perspectives; Psychology: Overview; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Boring E 1950 *A History of Experimental Psychology*, 2nd edn. Appleton Century Crofts, New York

Cahan D 1994 Helmholtz and the civilizing power of science. In: Cahan D (ed.) *Hermann von Helmholtz and the Foundations of Nineteenth-Century Science*. University of California Press, Berkeley, CA, pp. 559–601

Cahan D (ed.) 1994 *Hermann von Helmholtz and the Foundations of Nineteenth-Century Science*. University of California Press, Berkeley, CA

Cahan D 1995 Helmholtz als Kulturträger. *PTB-Mitteilungen* **105**(4): 249–54

Helmholtz H von 1856–67 *Handbuch der physiologischen Optik*. Voss, Leipzig, Germany

Helmholtz H von 1863 *Die Lehre von den Tonempfindungen, als physiologische Grundlagen für die Theorie der Musik*. Friedrich Vieweg und Sohn, Braunschweig, Germany

Helmholtz H von 1882–95 *Wissenschaftliche Abhandlungen*, 3 Vols. Barth, Leipzig, Germany

Helmholtz H von 1903 *Vorträge und Reden*, 5th edn., 2 Vols. Friedrich Vieweg und Sohn, Braunschweig, Germany

Helmholtz H von 1995 *Science and Culture: Popular and Philosophical Essays* (Cahan D ed.). University of Chicago Press, Chicago

Koenigsberger L 1902–3 *Hermann von Helmholtz*, 3 Vols. Friedrich Vieweg und Sohn, Braunschweig, Germany

Mirowski P 1989 *More Heat Than Light. Economics as Social Physics: Physics as Nature as Economics*. Cambridge University Press, Cambridge, UK

Rabinbach A 1992 *The Human Motor: Energy, Fatigue, and the Origins of Modernity*. University of California Press, Berkeley, CA

Turner R S 1994 *In the Eye's Mind: Vision and the Helmholtz-Hering Controversy*. Princeton University Press, Princeton, NJ

Weber M 1919 Science as a vocation. In: Gerth H H, Mills C W (eds.) *From Max Weber: Essays in Sociology*. Oxford University Press, New York pp. 129–56

D. Cahan

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Hempel, Carl Gustav (1905–97)

Carl Gustav Hempel was a leading figure of the logico-empiricist movement that dominated much of American philosophy in the second third of the twentieth century, and has been a principal source of present day analytical philosophy. In the spirit of logical empiricism Hempel aimed at a scientific philosophy with the principal objective to arrive at a purely logical methodology of the empirical sciences, including the social sciences. His most important contribution in this respect is what can be regarded the first theory of scientific explanation. This theory is referred to as 'the Hempel–Oppenheim model,' the 'covering-law model' or the 'standard model of scientific explanation.' In his later years Hempel gave up the idea of a purely logical (i.e., syntactical and/or semantically) 'rational reconstruction' or 'explication' of basic methodological concepts like explanation, confirmation, law of nature, probability, type, etc. Both the internal failure of this program and the influence of Thomas S. Kuhn caused Hempel to acknowledge for a satisfying reconstruction of methodological concept also 'pragmatic' aspects of science in the sense of social and historical circumstances of its development.

1. *Biography*

Hempel was born on 8th January, 1905, in Oranienburg (near Berlin, Germany). He studied mathematics, physics, and philosophy at the universities of Göttingen, Heidelberg, and Berlin from 1923 to 1929. After graduation ('*Staatsexamen*') that qualified him to teach in secondary schools ('*Gymnasien*') Hempel went to Vienna to study with Moritz Schlick and, particularly, Rudolf Carnap. Both of them, together with the mathematician Hans Hahn and the economist Otto Neurath and others, had founded in 1923 the 'Vienna Circle' ('*Wiener Kreis*'), that was to become the nucleus of logical empiricism. After his return to Berlin in 1930 Hempel entered teacher's training and teaching and worked at the same time with Hans Reichenbach, head of the 'Berlin Group' of logical empiricism, on a philosophical dissertation on the concept of probability. After graduation in 1934 Hempel preferred to leave Nazi Germany and went to Brussels to assist Paul Oppenheim, a private scholar, in carrying out philosophical research that related principally to concept formation in empirical science. In 1937 to 1938 Hempel worked for a year as research associate with Carnap in Chicago, and again went back to Brussels until he found a post of instructor at City College, New York in 1939. The academic year 1940 to 1941 saw him at Queens College, New York, from where he went as associate professor to Yale in 1948. In 1955 Hempel became Stuart Professor of Philosophy at Princeton. After retirement at Princeton (1973) he worked as university professor at Pittsburgh University from 1977 until 1985. He died at Princeton on 9 November, 1997.

2. *Major Contributions*

Hempel has given important contributions: (a) to the rational reconstruction of methodological concepts, particularly the concept of scientific explanation; (b) to the 'universality thesis,' i.e., the thesis that the social sciences do not, and should not, use methods that are essentially different from the methods of empirical science; and (c) to concept formation and taxonomy in empirical science.

2.1 *Scientific Explanation*

Hempel reconstructs scientific explanations of concrete events in nature, history, and society as arguments, i.e. as logical deductions containing as premises ('*explanans*') at least one (general) scientific law L_1, *as* well as initial and boundary conditions C_r, and as conclusion ('*explanandum*') the statement E describing the event in question. We thus obtain the following model of 'deductive-nomological (*D-N*) explanation'

$L_1, L_2, \ldots, L_n$	The *explanans*
$C_1, C_2, \ldots, C_r$	[entails logically]
E	the explanandum

In case the *explanandum* contains essentially a statistical law (e.g., for all x: the probability of G, if F is given, equals r and as boundary condition that F is given for some b) the explanatory argument does not deliver a deduction of the *explanandum* statement E but rather gives inductive support for E, i.e., the *explanans* shows that E was to be expected with a certain degree r of probability. In simplified form (one statistical law, one boundary condition) we thus obtain the following model of 'inductive-statistical (*I-S*) explanation':

$p(G, F) = r$	The *explanans*
	[lends inductive support of
$F(b)$	probability r to]
$G(b)$	the *explanandum*

In the case of *I-S* explanation, one must further assume that the *explanans* is maximally specified, i.e., the *explanans* must contain all available information that is possibly relevant for the explanation of the *explanandum* E. Or, in other words, the *explanandum* must contain all pertinent statistical laws and those special facts that can be connected with the E-event by statistical laws. *I-S* explanations thus relate to the body of scientific knowledge at a given time, i.e., they

become 'epistemically relative.' The postulate of maximum specificity avoids the 'explanatory ambiguity,' i.e., the possibility to give acceptable *I-S* explanations based on true premises for mutually exclusive *explananda*.

In case one deduces as *explanandum* a general or a statistical law of lesser generality than the one(s) in the *explanandum* one obtains the reduction of a law or theory in the first case (e.g., explaining the law of free fall by means of the law of gravitation), and a deductive-statistical or *D-S* explanation of a statistical law by a more comprehensive statistical law in the second.

Originally Hempel claimed the structural identity of explanation and prediction. Later he had to admit that, although every adequate explanation is (under pragmatically changed circumstances) also a prediction, not every adequate prediction yields also an adequate explanation, for symptoms are often adequate for a prediction, but not for the corresponding explanation. So is, for example, a sudden fall of the barometer (together with suitable laws) sufficient for predicting a thunderstorm, but it does not explain it, since the fall of the barometer is not the cause of the thunderstorm, but rather a symptom of its arrival.

2.2 Unity of Science

Hempel claims the methodological unity of empirical science. With respect to scientific explanation this amounts to a universality thesis of scientific explanation: there is no essential difference between explanations in the natural sciences and those in psychology, the social sciences, or history. Explanations in these fields aim to show that the event in question was to be expected on the basis of antecedent and boundary conditions and general laws. In history those general laws are usually taken from psychology, sociology, economics, the natural sciences, etc. Hempel leaves the question open whether genuine historical laws exist. Normally explanations in history and sociology fail to include an explicit statement of the laws they presuppose. This failure is due to the fact that those laws are part of folk psychology and seem to be tacitly taken for granted. Apart from that it is difficult to formulate in a sufficiently exact way the underlying general assumptions about e.g., the outbreak and course of revolutions. Furthermore, most of the regularities in the social sciences are statistical. Therefore one can only expect explanation sketches in these fields.

The method of empathic understanding that is often claimed to distinguish the social from the natural sciences is neither sufficient, nor necessary for explanations in these fields. It is not sufficient, because the individual act of understanding is only a heuristic device in order to suggest general psychological hypotheses that might subsequently serve as explanatory principles in the case under consideration. It is not necessary, because a historian might well be able to explain the deeds of a paranoiac historic personality by reference to the principles of abnormal psychology without being able to arrive at an empathic understanding of such a personality.

2.3 Concept Formation in Empirical Science

Hempel's work on concept formation in empirical science originates in his early cooperation with Paul Oppenheim and consists basically in the application of principles of logical empiricism to the formation of scientific concepts. Hempel distinguishes: (a) classificatory; (b) comparative; and (c) quantitative concepts. These distinctions normally characterize at the same time the historical development of a science from its initial descriptive 'natural history' stages to its mature theoretical stages. Classification divides a given set or class of objects (e.g., human beings) into subclasses (e.g., beings that suffer from certain mental disorders). Each of these subclasses is defined by means of a certain concept (e.g., schizophrenia) that represents the complex of characteristics essential for the membership in that subclass. Hempel speaks also of 'classificatory types.' The elements of these subclasses are those individuals to whom the respective concept applies (e.g., schizophrenics). The characteristics that form the defining concept are ascertainable fairly directly by observation. Classificatory concepts as well as scientific concepts in general have to fulfil two requirements. They have: (a) to be objective; and (b) to have systematic import. Objectivity is attained first by using operational criteria and (often only partial) operational definitions in a large sense that includes observation as operational and second by eliminating criteria with valuation overtones. Scientific concepts have systematic import if they lend themselves to the formulation of general laws or theoretical principles. Whereas classificatory concepts are a yes-or-no affair, i.e., a certain individual belongs to some subclass or it does not, comparative concepts admit of more or less with respect to a certain trait (e.g., x is warmer than y with respect to cold–hot). They lead to (quasi)linear orderings as soon as one includes the relation of coincidence (e.g., x is equally warm as y). Hempel calls the extremes of such orderings 'ordering types.' The next step would consist in giving these linear orderings a metric and thus arriving at quantitative concepts (e.g., temperature).

For Hempel the 'ideal types' that have become popular in the social sciences since Max Weber are neither classificatory nor ordering types. Ideal types are interpretative or explanatory schemas that contain a set of empirical generalizations that establish subjectively meaningful connections between different aspects of some kind of phenomenon, e.g., purely

rational economic behavior, a capitalist society, a religious sect. Accordingly ideal types must rather be reconstructed not as concepts but as theoretical systems that are intended to provide explanations and therefore must contain testable hypotheses. Despite their lack of clarity and precision there thus are also with respect to ideal types no essential methodological differences between psychology or the social sciences and the corresponding methods in the natural sciences.

3. *Influence, Impact, and Current Significance*

Hempel's ideas constitute the mature form of logico-empiricist thinking. His work on scientific explanation 'stands out as the benchmark to which all subsequent studies of this topic must be referred' (Salmon 1998, p. vii). His reconstruction of concept formation in empirical science has become a sort of standard for objective procedures of concept formation. It has greatly influenced subsequent editions of the *Diagnostic and Statistical Manual* (*DSM*) of the American Psychiatric Association (Houts 2001). His later work reflects the openness of logical empiricism to new developments in the philosophy of science (particularly T. S. Kuhn's work) that put a strong emphasis on the historical and sociological ('pragmatic') factors of the development of science (Wolters 2000).

See also: Empiricism, History of; Explanation: Conceptions in the Social Sciences; Ideal Type: Conceptions in the Social Sciences; Idealization, Abstraction, and Ideal Types; Logical Positivism and Logical Empiricism; Mathematical Models in Philosophy of Science

Bibliography

Hempel C G, Oppenheim P 1936 *Der Typusbegriff im Lichte der Neuen Logik. Wissenschaftstheoretische Untersuchungen zur Konstitutionsforschung und Psychologie*. A. W. Sijthoff, Leiden, The Netherlands

Hempel C G 1952 Fundamentals of concept formation in empirical science. *International Encyclopedia of Unified Science*. University of Chicago Press, Chicago, Vol. II, no. 7

Hempel C G 1965 *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, New York

Hempel C G 1967 *Philosophy of Natural Science*. Prentice-Hall, Englewood Cliffs, NJ

Hempel C G 1979 Scientific rationality: Analytic vs. pragmatic perspectives. In: Geraets T F (ed.) *Rationality To-Day—La Rationalité Aujourd'hui*. University of Ottawa Press, Ottawa, ON, pp. 46–66

Hempel C G 2000 *Selected Philosophical Essays*. Jeffrey R (ed.). Cambridge University Press, Cambridge, UK

Houts A C 2001 The diagnostic and statistical manual's new white coat and circularity of plausible dysfunctions: Response to Wakefield. *Behaviour Research and Therapy* **38**: 315–45

Salmon W C 1989 *Four Decades of Scientific Explanation*. University of Minnesota Press, Minneapolis, MN

Salmon W C 1998 *Causality and Explanation*. Oxford University Press, Oxford, UK

Wolters G 2000 Die pragmatische Vollendung des logischen Empirismus—In memoriam Carl Gustav Hempel (1905–1997). *Zeitschrift für allgemeine Wissenschaftstheorie* **31**: 205–42

G. Wolters

Henry, Louis (1911–91)

Henry was born on February 19, 1911 and was 35 years old when he joined the newly created French Institut National d'Études Démographiques (INED) in 1946; he retired in 1973 and died on December 30, 1991. Most of his work was published under the auspices of INED, either in the journal *Population*, or in the Institute's monograph series. He was one of several retired artillery officers hired by INED because of their rigorous mathematical education at the national *Ecole Polytechnique*; others were Jean Bourgeois-Pichat and Sully Lederman, who would also leave their mark in the field, although they had no formal training in the social or historical sciences. Henry soon joined forces with Paul Vincent in the study of fertility, and this became his area of specialization. He made an early contribution to its measurement by inventing a new index, the parity progression ratio (Henry 1953). This is defined as the proportion of women with a particular number of children, or parity, who have gone on to have another child, and is computed on the basis of the number of children ever born recorded at a particular age.

Henry intended to go beyond the mere measurement of phenomena, and undertook to decompose reproductive processes into their component factors, what later became known as the proximate determinants of fertility. He decided to study the 'physiological' level of fertility, that is, the fertility that would prevail among married women or couples in the absence of deliberate efforts to limit their number of births; this he called 'natural' fertility. In a contemporary population, however, contraception and abortion that are subject to human decisions are widespread, and it becomes difficult to investigate physiological factors. In his quest to find reliable data on populations where family limitation was not practiced, Henry studied records from the European past, at a time when the recording of vital events was complete and accurate. He first used a series of genealogies on the ruling classes of Geneva, going all the way to the sixteenth century (Henry 1956); later he analyzed the parish records for a village of Normandy, Crulai, that had been extensively compiled by a local historian, Etienne Gautier (Gautier and Henry 1958). The records for Geneva demonstrated the sharp contrast between the

earliest marriage cohorts, where marital fertility by age seemed to be independent from age at marriage, and the later cohorts where the behavior of couples was clearly modified by the duration of their marriage and the number of children they had attained. It could be shown conclusively that the bourgeoisie of Geneva had adopted family limitation as early as the second half of the seventeenth century. In contrast until the end of the eighteenth century, Crulai, the first parish analyzed by the new method of family reconstitution that Henry had invented, showed the features of a regime of natural fertility, where parity does not affect a woman's subsequent behavior.

In a manual written with Michel Fleury (Fleury and Henry 1956), he laid out—to the minutest details—the rules for the processing of parish data. These rules have been followed faithfully by demographic historians, not only in France but also in other European countries such as England, where they guided the work of the Cambridge Group for the History of Population and Social Structure, and in Canada. Henry would publish another textbook on historical demography (Henry 1967). Moreover, applying the vast experience accumulated in his early parish studies, he initiated a study of rural France, using a representative sample of French parishes during the eighteenth century. He analyzed this material by aggregative methods, and for a smaller sample of parishes used nominal information to reconstitute families. The project provided the first national series of estimates of fertility, mortality, and nuptiality for a Western European country in the eighteenth century. This body of work makes Louis Henry the uncontested father of historical demography as it is practiced today.

Mining the past for data on populations that did not consciously limit their family size was initially only a tool towards measuring the physiological components of human reproduction. Henry identified the factors to be considered as fecondability, the probability that a conception would result in a live birth; the delay in the resumption of ovulation after a delivery or nonsusceptible period, and total or partial sterility. He provided estimates of the value of these parameters on the basis of his own historical studies of population, but also of various accurate measurements made on other selected populations. His work was picked up by English and American demographers and biometricians working on fertility models, and among them by Mendel Sheps and Jane Menken who edited a collection of translations from Henry's most important articles (Henry 1972.) The seminal importance of Henry's work was acknowledged by the Population Association of America when it made him the first recipient of the Irene B. Taeuber Award for Excellence in Demographic Research in 1977. Just before his death in 1991, the International Union for the Scientific Study of Population had named him its first Laureate.

The concept of natural fertility is often considered as one of Henry's most significant contributions to population studies. He did not invent the term, but he defined it rigorously and provided means to identify natural fertility regimes. It refers specifically to the behavior of married couples, and is therefore not affected by variations in the proportions married and the age at marriage. Natural fertility is compatible with behavior that may affect the levels of natural fertility (frequency of intercourse, length of breast feeding) or even with attempts to space birth that are not depending on parity. Although the notion remains controversial, it has been used widely in the study of the fertility transition, both in the West and in today's developing countries.

See also: Demography, History of; Demography: Twentieth-century History; Fertility: Proximate Determinants; Historical Demography

Bibliography

Fleury M, Henry L 1956 *Des registres paroissiaux à l'histoire de la population. Manuel de dépouillement et d'exploitation de l'état civil ancien*. INED, Paris

Gautier E, Henry L 1958 *La population de Crulai, paroisse normande. Etude historique*. Presses universitaires de France, Paris

Henry L 1953 *Fécondité des mariages. Nouvelle méthode de mesure*. Presses universitaires de France, Paris

Henry L 1956 *Anciennes familles genevoises. Etude démographique, XVIe–XXe siècle*. Presses universitaires de France, Paris

Henry L 1967 *Manuel de démographie historique*. Droz, Geneva, Switzerland

Henry L 1972 *On the Measurement of Human Fertility. Selected Writings of Louis Henry*. Elsevier, New York

E. van de Walle

Hermeneutics, History of

The term 'hermeneutics' refers to the science, art, or technique of interpretation, paradigmatically of written texts but by extension of human actions and other social phenomena. In this extended sense, it has come to denote a variety of approaches in the social and behavioral sciences. The word refers to the Greek messenger-god Hermes, and reflection on the problems of interpretation and criteria for the truth, validity, or adequacy of interpretations, goes back to ancient Greek thought in the European tradition and a similar historical distance in the other world civilizations. We can consider this as the prehistory of hermeneutics, revived in the philological criticism of classical texts in the Renaissance, the interpretation

of Roman Law, the interpretation of the Bible in (especially Protestant) Christianity and the philosophical analysis of texts.

The systematization of hermeneutics occurred largely in German-speaking Europe in the course of the nineteenth century, although two earlier thinkers deserve particular mention. Johann Martin Chladenius (1710–59) published in 1742 a general account of interpretation which brought together traditional Aristotelian rhetoric and more recent Enlightenment themes, developing in particular the notion of point-of-view or perspective (*Sehe-Punkt*), crucial to later philosophy of history and social theory (Mueller-Vollmer 1985). Second, Giambattista Vico (1688–1744) formulated the basic principle that our knowledge of what we ourselves have made (individually or collectively) is different from what we have not made. The world of human society and culture is in some sense 'our' product.

This distinction, crucial to later hermeneutics and 'hermeneutic' social science, between what we know from the inside and what we know because we have learned about it is taken up in a rather more speculative form in Hegel's differentiation between reason (*Vernunft*) and the understanding (*Verstand*), although the maker here is the world-spirit coming to recognize its own productions (including, ultimately, the world itself) and its learning processes as rational, in contrast to the essentially contingent states and relations found in nature and described by the mathematical and natural sciences. Something more like Vico's idea returns again in the second half of the nineteenth century, when Hegel's concept of objective mind is extracted from its surrounding developmental 'grand narrative' and treated more as a descriptive category.

It was Friedrich Schleiermacher (1768–1834), following Friedrich Schlegel, who consolidated hermeneutics in a systematic form, establishing the term understanding (*Verstehen*), which had been used by Chladenius and which has survived as standard usage in English-language social science discourse, and making it central to interpretation—understood as a more systematic activity. Interpretation here involves both the linguistic understanding of meaning and a psychological understanding of the author's intention. Schleiermacher introduced the distinction between technical and more developed or speculative forms of interpretation and formulated the much-discussed principle that it should be possible to understand an author better than he understood himself. Schleiermacher's contribution was made the central and culminating point of the account of 'the rise of hermeneutics' given by Wilhelm Dilthey (1833–1911), with whom hermeneutics becomes central to the self-definition of what he called the human sciences or *Geisteswissenschaften*. In these sciences, as Dilthey put it, the mental activity of humans and of some other animals, and its products, can be understood.

Dilthey and his contemporary, the philosopher of history J. G. Droysen (1808–84), developed what we would now call a research program for history and the other human sciences based on the distinctiveness of human psychic expressions and the understanding of those expressions. In a move that was to become a definitional feature of later interpretive social science, Dilthey, like Schleiermacher, emphasized the continuity between everyday understanding and more formal processes of interpretation. His distinction between the natural and human sciences was developed in large part in opposition to Comtean positivism, which had become influential, even in the German-speaking countries, by the middle of the nineteenth century. In a parallel, but more methodological formulation, two other neo-Kantian thinkers, Wilhelm Windelband and Heinrich Rickert, argued that the study of culture is essentially concerned with individual processes and relating them to shared human values, whereas the natural sciences are concerned with general laws about objects which are essentially remote from questions of value. We are interested, for example, in the French Revolution, not just as a member of a class of revolutions exhibiting certain common features (this would be, for Rickert, a natural-scientific mode of approaching it), but as a unique event embodying, and perhaps also violating, certain crucial human values.

This opposition between positivism and methodological dualism, and more particularly between causal explanation, analyzed in terms of universal regularities, and 'understanding' comes to structure the emergent human or social sciences, as the term 'culture' increasingly gives way to 'society' or 'sociation' (*Vergesellschaftung*). There is a fairly strong line of influence from Rickert to Max Weber, both directly and also through his friend Georg Simmel, who discusses our knowledge of the social world in terms that foreshadow social constructionist theory (see *Constructivism/Constructionism: Methodology*).

Max Weber's personal development from economic and legal historian to sociologist continues the encounter between hermeneutics and the social sciences which had been a dominant feature of the late nineteenth century, and he also develops a middle position on the issue of methodological dualism, as he had earlier on the related opposition in economic theory (the *Methodenstreit*) between the pursuit of a systematic laws and a more interdisciplinary and historical approach. For Weber, as he put it in an early essay, 'the course of human action and human expressions of every sort are open to an interpretation in terms of meaning which in the case of other objects would have an analogy only at the level of metaphysics' (Weber 1975, pp. 217–8). He therefore later defines sociology, in the first sentence of his major work 'Economy and Society,' as a science which aims at an interpretative understanding of action in order thereby to understand its course and its effects.

Whether by this Weber means that explanatory understanding is itself a form of causal explanation, or merely complementary to it, the crucial point for him is that explanations of social phenomena must be both 'causally adequate' and 'meaningfully adequate.'

For Weber, then, our access to knowledge of the social world is importantly different from our knowledge of nature. However, it is not, in his view, any less objective. He heroically attempts to hold together Rickert's principle that our perspectives on cultural phenomena, and our knowledge of them, are shaped by values (based for Weber on ultimately ungroundable existential choices), with the idea that the social sciences can attain a bedrock of solid and 'value-free' knowledge which would have to be accepted, as he sometimes curiously puts it, 'even by a Chinese.' Weber insists that it is the intentions and purposes (subjective meanings) of human actors which define their actions and which therefore have to be understood by the historian or sociologist, but he moves rapidly to the construction of a system of ideal types of action-orientation less limited than that found in economic theory but still substantially dependent on it.

Weber's synthesis was pulled apart from both sides in the decades following his early death in 1920. At one pole, there was now a more stridently naturalistic and indeed reductionist variant of positivism: the logical empiricism of the Vienna Circle, in whose 'unified science' the statements of all sciences should be ultimately reducible to material-object language or to statements in physics; *Verstehen* was of no more importance, in Otto Neurath's vigorous formulation, than a good cup of coffee which sustains the social scientist.

From the other direction, Alfred Schutz initiated the tradition of social phenomenology, and what is often loosely called 'hermeneutics' in the context of the social sciences, with a book published in Vienna in 1932 with the title *The Meaningful Constitution of the Social World*. Schutz felt that the problem with Weber's ideal types was not that they were insufficiently scientific, but precisely the opposite: Weber was too quick to impose them on the phenomena he described, paying insufficient attention to their grounding in acts of typification performed by ordinary members of society. For Schutz, the social scientist is merely constructing second-order typifications based on those already carried out in the lifeworld. This theme was taken up by the phenomenologist Aron Gurwitsch in the USA, where Schutz had also settled, and by Peter Berger and Thomas Luckmann (1967), who finally put the term 'social construction' on the map, offering in the guise of a sociology of knowledge a paradigm more directly adapted to use in social research. By the time Schutz's first book was republished, in Germany in 1970 and in the USA in 1967, the way had been prepared by his own later work and by Berger and Luckmann (1967).

The middle decades of the twentieth century replay many of the debates which had dominated the second half of the nineteenth century. Droysen's complaint, in a letter of 1852, about the rise of what he called 'crass positivism,' and his crusade against it in a course which he taught from 1857 onwards, is echoed by antinaturalist social science just over a hundred years later. Social phenomenology also of course had affinities with the well-established minority North American tradition of symbolic interactionism, which also experienced a certain resurgence in the 1960s with, for example, the republication of the work of G. H. Mead and that of Herbert Blumer, and also the publication of various studies by Erving Goffman. But where the interactionists tended not to spend time on formal critiques of empiricism, Harold Garfinkel's ethnomethodology (Garfinkel 1967) was more aggressive. A former student of the structural functionalist Talcott Parsons, Garfinkel was led by his study of the deliberations of a trial jury to emphasize, like Schutz, the importance of practical reasoning in everyday situations. He showed that our drive to interpret leads us to impute a meaning even to random processes in the social world, as in one of his later experiments in which an interviewer playing the role of a counsellor replied to the clients' questions by a random 'yes' or 'no,' leading them into more and more contorted interpretative efforts. The production of meaning is at the same time the production of social order—Parsons' major concern. Unlike his former teacher, however, Garfinkel insisted that social actors are not simply bearers of their social roles ('cultural dopes'), but active subjects obliged to practice social analysis in order to function in everyday society.

Social phenomenology found a somewhat unexpected ally in a convergent move within analytic philosophy. Beginning on the margins of the Vienna Circle, Ludwig Wittgenstein came to abandon the simple conception of a picturing relation between propositions and the world, and was drawn into a more sensitive and holistic analysis of the practicalities of 'language-games' based on implicit rules and embedded in what he enigmatically called 'forms of life.' An important book by the Wittgensteinian philosopher Peter Winch (1958) drew the consequences for social theory, using Max Weber, as Schutz had done, as one of the foils for his argument. For Winch, knowing a society meant learning the way it is conceptualized by its members. He thus revived the central principle of nineteenth century German historicism, according to which every age must be understood in its own terms. Winch directly identified himself with the German idealist tradition by insisting that social relations are 'like' logical relations between propositions (1958, p. 126) as well as, more concretely, with an ethnographic field-work approach.

Hermeneutic theory itself also took a new turn with the 'philosophical hermeneutics' of Hans-Georg Gadamer (1900–), whose *Truth and Method*, published in

1960 and translated into English in 1975, insists, in opposition to historicist hermeneutics, on the practical dimension of interpretation, conceived in Heidegger's sense of an 'encounter' between the 'horizon' of the interpreter and that of the text itself. Gadamer's philosophical hermeneutics is thus conceived in opposition to the methodological emphasis of traditional hermeneutic theories and their concern with the accuracy of interpretation. Gadamer's aim is to describe the underlying process, an existential encounter between two perspectives or horizons of expectation, which makes interpretation possible in the first place. Understanding is not just a matter of immersing oneself imaginatively in the world of the historical actor or text, but a more reflective and practical process which operates with an awareness of the temporal and conceptual distance between text and interpreter and of the ways in which the text has been and continues to be reinterpreted and to exercise an influence over us. This effective history (*Wirkungsgeschichte*), which traditional historicist hermeneutics tends to see as an obstacle, is for Gadamer an essential element that links us to the text. Our pre-judgments or prejudices are what make understanding possible.

Although Gadamer has often stressed that his philosophical hermeneutics, with its origin in Heidegger's hermeneutic ontology, is distinct from hermeneutics as a technique of interpretation, his approach clearly poses a challenge to more traditional conceptions of hermeneutics. These differences are brought out in particular in Gadamer's exchanges in the 1960s with Emilio Betti. The alternative conception of the human sciences or *Geisteswissenschaften* put forward in Gadamer's work also made it central to Jürgen Habermas' reformulation of the logic of the social Sciences (Habermas 1983. Habermas welcomed Gadamer's critique of hermeneutic objectivism, which he saw as the equivalent of positivism in the philosophy of the natural sciences, and also his stress on the totalizing character of understanding. For Habermas, however, Gadamer's stress on the fundamental nature of language, expressed in his claim that 'Being that can be understood is language,' amounted to a form of linguistic idealism. Together with Gadamer's stress on the importance of tradition and his rehabilitation of the category of prejudice, this suggested an ultimately conservative approach which was unable to deal with the systematic distortion of communicative processes by relations of power and domination. Habermas and Gadamer debated these issues in the late 1960s and early 1970s (Bleicher 1980); more recent theorists have tended to stress the compatibility of hermeneutics and critical theory (see *Critical Theory: Frankfurt School*) in a conception of critical hermeneutics (Thompson 1981, Outhwaite 1987). More recently, Gadamer also engaged briefly with the French deconstructionist philosopher Jacques Derrida, whose conception of interpretation is more skeptical.

Habermas' version of 'critical theory' can be seen, along with critical realism and Anthony Giddens' structuration theory, as one of three particularly influential attempts in the final third of the twentieth century to reconcile, as Max Weber had done at the beginning of the century, the rival claims of explanation and understanding in the social sciences. Habermas, along with his close collaborator Karl-Otto Apel, argued for a complementarity between an empirical-analytic approach oriented to the explanation, prediction, and control of objectified processes and a hermeneutic approach concerned with the extension of understanding, in an emancipatory model of critical social science, instantiated by psychoanalysis and the Marxist critique of ideology, which aims at the removal of causal blocks on understanding. Much of this remains in his more recent theories of reconstructive science and communicative action.

Critical theory tends towards a dualism of natural and social science, but in an increasingly muted form. The assumption that opposition to positivism also entailed dualism or antinaturalism was however also put in question in the late 1960s and early 1970s by the realist metatheory of science developed by Rom Harré and Roy Bhaskar (1975). Both Harré and Bhaskar, like Habermas, were substantially motivated by the desire to undermine positivistic theories and approaches in the social sciences. Harré and Secord (1972) developed a philosophy for social psychology based on the work of the later Wittgenstein and the analytic philosophy of language practiced at Oxford by J.L. Austin. Ordinary language, they argued, is better suited to the description of the mental processes of social actors than apparently more scientific artificial terminology, and they drew attention to models of research practice of this kind in the work of Goffman, Garfinkel, and others.

Harré and Bhaskar were in any case interested in giving a more adequate account of science as a whole, in a world composed of relatively enduring structures and mechanisms. Some of these could be isolated in scientific experimentation, given the contingent existence of *Homo sapiens* and *Homo scientificus*. An important aspect of the realist programme developed by Harré, Bhaskar, and others was a conception of explanation as involving not an essentially semantic reduction of causal statements to general laws but a reference to the causal powers of entities, structures, and mechanisms. Causal tendencies might or might not be outweighed by countervailing tendencies, and two causal tendencies may neutralize one another, as do the centrifugal force of the earth's rotation and its gravitational attraction, with the convenient consequence that human beings and other animals are safely anchored to the earth's surface.

This and other features of realism meant that the whole issue of naturalism could be rethought. Human beings could be seen as having causal powers and liabilities, just like other entities; it no longer mattered

so much that their relations rarely sustained any universal generalizations of an interesting kind, but only sets of tendencies regular enough to be worth exploring. The fact that many of the entities accorded causal force in social scientific explanations were necessarily unobservable was not, as it was for empiricism, a problem of principle. And the understanding of meaning could, as Bhaskar put it, be seen as in some ways equivalent to measurement in the natural sciences. Finally, it seemed natural to include among the causes of human action the agents' reasons for acting—reasons which must be understood as far as possible.

The realist critique of traditional epistemology found an echo in social theory, notably in the work of Anthony Giddens, who had become similarly impatient with the residues of positivist social science as well as the more radical contentions of social constructionism. Giddens' conception of the 'duality of structure' was designed to replace the traditional dichotomies between theories of social structure and social change, the micro–macro divide, and between interpretive and more structural approaches. Approaches like these which aim to mediate between pure hermeneutics and more naturalistic conceptions of social science coexist, at the dawn of the twenty-first century, with more explicitly hermeneutic or phenomenological conceptions (see e.g. Soeffner 1989, Müller-Doohm and Jung 1993). Hermeneutics in a broader sense continues to exist as a major research tradition in the humanities, as well as a minority one in the social sciences (Shapiro and Sica 1984). More importantly perhaps, social scientists that would not sign up to an explicitly hermeneutic program have at least accepted the importance of hermeneutic issues.

See also: Ethnomethodology: General; Hermeneutics, Including Critical Theory; Interactionism: Symbolic; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Schtz, Alfred (1899–1959); Semiotics; Weber, Max (1864–1920); Wittgenstein, Ludwig (1889–1951)

Bibliography

Apel K-O, Bubner R, Gadamer H-G, Giegel H J, Habermas J 1971 *Hermeneutik und Ideologiekritik*. Suhrkamp, Frankfurt, Germany

Berger P L, Luckmann T 1967 *The Social Construction of Reality*. Allen Lane, London

Bhasker R 1975 *A Realist Theory of Science*. Leeds Books, Leeds, UK

Bleicher J 1980 *Contemporary Hermeneutics*. Routledge & Kegan Paul, London

Chladenius J M 1742 *Einleitung zur richtigen Auzlegung vernübftiger Reden und Schifted. Mit einer Einleitung von Lutz Geldsetzer*. F Lanckisch, Leipzig, Germany [1969 Stern-Berlag Janssen, Düsseldorf, Germany]

Dallmayr F R, McCarthy T A 1977 *Understanding and Social Inquiry*. University of Notre Dame Press, Notre Dame, IN

Dilthey W 1964 *Die Geistige Welt*. Tenbner, Stuttgart, Germany

Droysen J G 1974 *Historik: Vorlesungen uber Enzyklopadie und Methodologie dei Geschichte*. Oldenbourg, Munich, Germany

Gadamer H-G 1960 *Wahrheit und Methode*. Mohr, Tübingen, Germany [1975 *Truth and Method*. Sheed and Ward, New York]

Garfinkel H 1967 *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ

Giddens A 1976 *New Rules of Sociological Method*. Hutchinson, London

Giddens A 1984 *The Constitution of Society*. Polity, Cambridge, UK

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, Gaiden City, NY

Habermas J 1978 *Knowledge and Human Interests*. Heinemann, London

Habermas J 1987 *The Theory of Communicative Action*. Polity, Cambridge, UK, Vol. 1 (1984), Vol. 2: (1987)

Habermas J 1988 *On the Logic of the Social Sciences*. Polity, Cambridge, UK

Harre R, Secord P 1972 *The Explanation of Social Behaviour*. Blackwell, Oxford, UK

Muller-Doohm S, Jung T (eds.) 1993 *Wirklichkeit im Deutungsprozeß*. Suhrkamp, Frankfurt, Germany

Mueller-Vollmer K 1986 *The Hermeneutics Reader*. Blackwell, Oxford, UK

Outhwaite W 1987 *New Philosophies of Social Science*. Macmillan, London

Schutz A 1932 *Der sinnhafte Aufbau der sozialen Welt*. Springer, Vienna [1972 *The Phenomenology of the Social World*. Heinemann, London]

Shapiro G, Sica A 1984 *Hermeneutics: Questions and Prospects*. University of Massachusetts Press, Amherst, MA

Soeffner H-G 1989 *Auslegung des Alltags–Der Alltag der Auslegung*. Suhrkamp, Frankfurt, Germany

Thompson J B 1981 *Critical Hermeneutics*. Cambridge University Press, Cambridge, UK

Weber M 1975 *Roscher and Knies*. Free Press, New York

Winch P 1958 *The Idea of a Social Science and its Relation to Philosophy*. Routledge, London

W. Outhwaite

Hermeneutics, including Critical Theory

1. Johann Conrad Dannhauer (1603–66)

The term 'hermeneutics' is generally defined as 'the science or art of interpretation.' Although it denotes a practice which stretches back to the Greeks, modern hermeneutics emerges in the seventeenth century with the publication in 1654 of J. C. Dannhauer's *Hermeneutica sacra sive methodus exponendarum sacrum litterarum*. The significance and novelty of Dannhauer's approach lies in the fact that interpretation is for him no longer treated simply as a tool for biblical exegetes or textual philologists, but as a universal phenomenon. Dannhauer was led to this conclusion by virtue of the fact that in the seventeenth century there was an enormous proliferation of texts in all disciplines due to

major innovations in the world of printing. Consequently, critical attention was now directed toward the various worldviews which were circulating in textual form, and away from those systems of belief which had hitherto escaped the scrutiny of a less than literate public. Knowledge was being construed as an activity which sought to penetrate behind appearances neither to ultimate reality, nor to a stable and immutable order of things, but as a practical mode of engagement with the texts and writings of those who had made popular the view that there was such an order. Hence, texts which had hitherto assumed divine status—such as the Bible—were now in open competition with multifarious other books and treatises. The veracity of any text depended, from a hermeneutic perspective, upon how well both it and its author withstood interpretative scrutiny.

Such a procedure did not seek to engender confusion or to render truth and meaning inaccessible. For Dannhauer, on the contrary, hermeneutics amounted to a 'universal science' which enables us to appraise the truthfulness of any discipline by clearing away prejudice and falsehood. As such, getting at meaning or truth requires an approach which is never content simply to take things at face value. The master of suspicion will see meaning as something to be disclosed through a recovery of an author's intention, and not as something which can be grasped intuitively.

Dannhauer's universal science of hermeneutic inquiry was a rudimentary attempt to come to terms with the uncertainties regarding the status of truth, knowledge, and meaning occasioned by the dawn of the Renaissance. By giving interpretation a central epistemic role, he anticipated what was to become a major philosophical school of thought.

2. *Romantic Hermeneutics: Schleiermacher and Dilthey*

Although Dannhauer's significance ought not to be underestimated in the history of hermeneutic thought, much of what comes under the headings of 'hermeneutics' and 'critical theory' today has its direct antecedents in the eighteenth and nineteenth centuries. It is to scholars of the Romantic period, such as Friedrich Schleiermacher and Wilhelm Dilthey, that one must look in order to gain an understanding of the many contemporary movements which are classified as being largely hermeneutic in origin and ethos.

2.1 *Friedrich Schleiermacher*

Schleiermacher and Dilthey are usually credited with having made hermeneutic inquiry a central *philosophical* preoccupation. Schleiermacher (1768–1834) sought to advance Dannhauer's attempt at a universal hermeneutics by trying to identify what was common to the technical operations and apparatuses of both the philologist and classical exegete. Consequently, he disregarded subject matter—whether classical document or scripture—in an effort to pinpoint the fundamental and universal techniques or methodology of the interpretative process.

Schleiermacher concluded that any act of interpretation involved both a 'grammatical' and a 'technical' form: while grammatical interpretation tries to understand a document by studying the broader linguistic milieu from which it derives, technical (stemming from 'techne'—meaning artistic creativity or making) interpretation, on the other hand, looks to the author and the author's personality and unique genius as a means of understanding the text in question. While both types of interpretation are significant for Schleiermacher, he nevertheless insists that they cannot be carried out concomitantly. For to interpret grammatically is to focus attention on language *stricto sensu*, and not on the individual speaker as such; conversely, technical interpretation demands that the reader look to the peculiarities of a particular author, to the author's style and use of expression. In the latter case, the general language used by the author is overlooked in favor of the *way* in which the author uses it.

For Schleiermacher, the ultimate task of the hermeneutic project is served best by this technical approach to questions of meaning and understanding. Truly to understand a text, the readers must strive to reappropriate its subjective dimension. They must endeavor, that is, to extrapolate from the document the individual author's 'psychological' intentions and motives. In so doing, the readers seek to come to terms with the text's central purpose and aims. Technical interpretation allows us to tap into the soul of the work in question.

As stated above, Schleiermacher considered his way of dealing with the complexity of texts as more than simply a means of alleviating the difficulties confronted by biblical exegetes and classical philologists. He followed Dannhauer's lead in trying to identify a hermeneutic methodology (*Kunstlehre*), which would govern not only specialized interpretative procedures, but all forms of understanding. In so doing, Schleiermacher insisted that we must always begin by presupposing misunderstanding, that misunderstanding is not something which may bedevil our reading at some future point, but that which bedevils us at every stage of the interpretative process. Hence, in the words of Jean Grondin, Schleiermacher 'universalizes misunderstanding as the situation and occasion of interpretation' (Grondin 1994, p. 70). The prospect of their not having fully understood a work, in other words, is what persuades the readers to read and reread the subject matter, to assess and reassess their findings. Furthermore, to presuppose misunderstanding sharpens the interpreters' acuity, in that such an approach

ensures that they will not take anything for granted; that the readers will always admit to the possibility of being disproved.

This is why technical interpretation is, for Schleiermacher, to be preferred over and above grammatical interpretation. For if misunderstanding is to be surmounted, there ought to be as complete a reconstruction of the authorial intention as it is possible to undertake. It is only by trying to stand in the shoes of the author that we can best understand the meaning and raison d'être of the text. Indeed, he goes further by suggesting that we ought not only understand as well as the author, but that we must, if we are really to reconstruct a foreign viewpoint, cultivate an even more profound understanding of the work than that possessed by its author. Such is the ultimate objective of all hermeneutic inquiry for Schleiermacher.

To understand a text more thoroughly than its author seems a strange ideal for someone who emphasizes so emphatically the fundamental role of misunderstanding. It should be read, however, as simply that—an *ideal*. Due to the fact that one can never assume that all misunderstanding has been eradicated, it is the duty of the reader to purvey ever more valid and fruitful interpretations of a work. Such an ideal provokes the reader to overturn previously held interpretations in favor of more plausible alternatives. The desire to know the author and the author's world better than the author is what prevents, in Schleiermacher's opinion, hermeneutic complacency. It is what drives and inspires the inquisitive reader who cannot settle for the commonly held interpretation.

Such attempts to penetrate behind the mind of an author through a technical interpretation of that author's work caused many to charge Schleiermacher with psychologism. While such a criticism is valid to the extent that Schleiermacher's hermeneutics did evince a tendency to reconstruct a psychological profile of the author, or to highlight the thought processes which led certain authors to write as they did, it could be argued alternatively that such a procedure was a nascent attempt to crystallize what was to become the bedrock of all subsequent hermeneutics—the recreation of an imaginary dialog with those of our past in an effort to make sense of *who* we are today. Given that Schleiermacher did not have at his disposal the insights of Freud and other subsequent theorists of the 'self,' his tentative steps toward a universal hermeneutic methodology founded on a dialogical conception of identity should be seen as a brave attempt to give the role of interpretation a philosophical credibility which it hitherto lacked.

2.2 Wilhelm Dilthey

While Schleiermacher's contribution to hermeneutic inquiry was not insubstantial, Wilhelm Dilthey (1833–1911) was a thinker whose work, and its subsequent appropriations and applications on the part of leading philosophers, has had a most profound impact on the formation of twentieth-century hermeneutics. For it was Dilthey who argued that if hermeneutics is to be genuinely universal, it must have as its object all of reality. The text of reality is history itself, as that which interconnects the multifarious and disparate strands of human endeavor. History, on this reading, is mankind's document—its grand narrative. As all texts have a place in this document, it behooves the inquirer to try to understand how history coheres before trying to understand how individual texts cohere. In problematizing history thus, Dilthey dramatically rendered more complex the job of the interpreter.

Dilthey's methodological approach was determined in large measure by the advent of positivism—a school of thought which insisted on the primacy of the natural sciences (*Naturwissenschaften*) as a model for understanding the world and our relation to it. For it was his conviction that if the human sciences (*Geisteswissenschaften*) were to fend off the challenge coming from the natural sciences, they would have to adopt analogous methodological procedures as their natural counterparts. Only in so doing could one come up with an adequate defense for the efficacy of historical knowledge as encapsulated in the various human sciences. Dilthey is, therefore, faced with the problem of how to rescue the laws of history from the laws of science, or of how to distinguish scientific *explanation* from historical *understanding* (*Verstehen*).

In his quest to pinpoint what it is that marks understanding off from explanation, Dilthey argues that a human being is not known solely as a physical object amongst others; it cannot, that is, be reduced to the level of scientifically observable phenomena. In contradistinction to scientific data, the human being has a mind which exceeds the purview of the natural scientist or the positivist, a mind which attests to its being. Consequently, a person may be understood by other people who are free to interpret and decode that person's intentions and expressions, desires and volitions. It is the mental faculty of human beings, in other words, which permits them to be *understood* and not merely *explained*. Thus, for Dilthey it is *psychology*, as the science of the human as historical and sociological agent, which ought to act as the bedrock of the human sciences. In this, he remains faithful to Schleiermacher's privileging of interpretative psychology as the most efficacious means of understanding the intentions of an other.

Unlike our knowledge of the external world, which requires the mediation of the senses, knowledge of the inner world of psychological life is free from temporal vicissitudes. As such, it forms the basis of a universally fixed mental structure which may be accessed immediately. But the quandary which Dilthey faced in advancing such a theory was quite considerable. For

example, in order to legitimate his claims he had to put forward incontrovertible proof that there were in fact complex mental structures which were indeed self-evident, and which could be accessed in a neutral way by any thinking subject. Given that there has never been agreement among the practitioners of psychology as to the precise nature of such structures, let alone proof of their existence, it is hardly plausible to assert that they are universally self-evident.

In an effort to counter such dilemmas, Dilthey elects to prioritize hermeneutics as that which has the best chance of permitting access to the inner life of the mind, by way of external media such as signs and texts. For it is through such documents that the mental life of one different to the self is made manifest for analysis by the human scientist. The inner life of an other, in other words, is objectified in and through the traces which he or she leaves behind. Consequently, the interpreter has the capacity to reconstruct the life of the other by recovering what the other has deposited in textual form. The text, therefore, is that which bears witness, in external form, to the inner mental life of one who was formerly alien to the self. It is that by way of which the interpreter, or the human scientist, can arrive at a credible understanding of a life. The ultimate aim of the hermeneutic enterprise is to reproduce the life that one is endeavoring to understand. If this could be achieved, the objective validity of the human sciences would no longer be in any doubt. It would also serve to avert the pending crisis, which historicism and relativism threatened to engender.

3. *Phenomenological Hermeneutics: Heidegger and Gadamer*

3.1 *Martin Heidegger*

Dilthey's legacy was far-reaching and extensive. In the domain of philosophy, it was Martin Heidegger (1889–1976), a German student of the eminent phenomenologist, Edmund Husserl, who would do most to build upon and advance Dilthey's hermeneutic insights. While Dilthey emphasized the role of hermeneutics in the human sciences, Heidegger was to go one step further by inquiring into the ontological conditions of such sciences. In other words, Heidegger considered the human sciences to be a branch of epistemology (Greek: *episteme*) or the theory of knowledge. As such, they ought not to be the focus of genuinely fundamental philosophical analysis. Rather, the task of the hermeneutic phenomenologist was to inquire into the actual *being* of such sciences, their ontological (Greek: *on*) status. Indeed, this move towards *being*, and away from *knowledge*, was Heidegger's most significant contribution, not only to phenomenology and hermeneutics, but to the history of twentieth-century philosophy.

Central to Heidegger's thought was the belief that the dualism between subject (the self) and object (the world beyond the self—the 'external' world) was misguided and mistaken. Although such dualisms are a regular feature in the history of philosophy, they are nowhere more apparent than in the thought of the seventeenth-century French philosopher, René Descartes. Descartes's contention that the thinking self (*Cogito*) could doubt away its environment in an effort to establish its own certainty, was primarily responsible for concretizing the subject–object divide. Even Heidegger's philosophical master, Husserl, structured his thought around this dualism; his emphasis on the role of intentional consciousness in the formation of the self's constitution was seen by Heidegger to be yet another unfortunate consequence of the Cartesian legacy, even if Husserl's original aim was to surmount the subject–object divide by leading us back to an originary experience of the world. If Husserl emphasizes consciousness as that upon which our primordial experience of the world is founded, Heidegger, in an effort to reconcile both subject and object, interprets this originary experience as a *being-in-the-world*. In so doing, Heidegger replaces the question common to Descartes, Kant, Dilthey, and Husserl—'What does it mean to *know*?,' with the question, 'What does it mean to *be*?' For Heidegger, there is no way the subject can, or should, try to extricate itself from the world in which it finds itself—from its natural environment. For it is through our basic interaction with the world that we come to a genuine and fundamental understanding of the 'who' of being.

The question of being, the question which first stimulated the early Greek metaphysicians, is thus what Heidegger intends to retrieve. He does so by underscoring our *being-there* (*Dasein*), our temporal situatedness *in* the world, as distinct from our consciousness *of* the world. Due to the fact that a human being is the only being capable of asking the question: 'What does it mean to be?,' any examination of being must therefore proceed by way of an inquiry into the human self as a being-there (*Dasein*), as a being who is concretely submerged in the practical affairs of the everyday world. This helps to explain the title of Heidegger's *magnum opus* of 1927, *Being and Time* (1962).

One such way Heidegger approaches the question of the being of *Dasein*, is through an analysis of the category of 'understanding.' For Heidegger, 'understanding' does not denote conscious awareness or assimilation, but rather those 'moods' (*Stimmungen*) which characterize our basic lived experience. These moods take the form of concern, anguish, fear, guilt, excitement, etc., and are determined by Heidegger as ontological acts of *pre*understanding (*Vor-Verständnis*), or the means by which we come to grips with the

world prior to the process of making the world an object for thought. In other words, Heidegger's phenomenological analysis strives to bring the moods of *Dasein*'s preunderstanding to the level of a reflective self-awareness. Like his nineteenth-century predecessor, Soren Kierkegaard, Heidegger turns Descartes's dictum—'I think, therefore I am' (*Cogito ergo sum*), on its head. For both Kierkegarad and Heidegger, we exist before we are objectively aware that we exist; our existence is preunderstanding, in the sense that we prereflectively interpret the world as a project of possibilities for our existence, before coming reflectively to understand it as such.

Heidegger's philosophical hermeneutics, unlike the methodological hermeneutics of Dilthey, will not, therefore, reflect upon the human sciences, but will attempt an examination and explication of the very foundations of those sciences. Moreover, for Dilthey, understanding has always to do with the understanding of another psyche or mind. For Heidegger, however, the problem of other minds is replaced with *Dasein*'s preunderstanding of the world into which it is 'thrown.' Understanding the world, thus, and not an other mind, is the fundamental goal of Heideggerian hermeneutics. To understand is not to understand facts, but to understand what possibilities are open to us in the situation in which we find ourselves. As such, human existence constitutes what Heidegger famously terms 'a hermeneutic circle,' in that it implicitly interprets itself in terms of its everyday moods and projects before it raises this interpretation to the level of philosophical speculation.

For Heidegger, it is by way of language that we interpret a thing *as* something which is intimately related to our project of being-in-the-world. Language should not be thought of in this context as mere assertion, but as interpretative of hermeneutic discourse. For on Heidegger's reading, the world never shows up for *Dasein* in some raw manifestation, but is always disclosed in and through the projects which *Dasein* sets itself. For example, Heideggerian discourse does not simply assert that an object possesses a series of objective characteristics such as weight, color, and width. More essentially, it interprets the object *as* something which is useful and meaningful for its existence. Hence, the object is described, interpreted, or defined, in its relation to *my* specific projects. A rock is thus seen *as* a weapon or a sculpting stone, depending on the use *I* make of it. If objective assertion treats things as objects present-at-hand (*vorhanden*), hermeneutic discourse recognizes them as instruments ready-to-hand (*zuhanden*)—or as things which aid us in our attempts to fulfill our everyday projects. Hermeneutic discourse, that is, reveals things in terms of their possible serviceability for me, and it does so by releasing them from the abstraction of a timeless present into the temporal horizons of my concrete historical concerns—retrieving the meaningfulness of these entities from my past experience or projecting such meaningfulness into my future possible experiences.

Heidegger's great contribution to the history of hermeneutics was his insistence that the world discloses itself through interpretative discourse. He failed, however, to explain what repercussions and ramifications this might have for the human sciences and for culture as a whole. It would be left to his student, Hans-Georg Gadamer, to fulfill this task.

3.2 Hans-Georg Gadamer

In Gadamer (1900–), we see a movement away from Heidegger's preoccupation with the question of being, back toward more epistemological concerns. His *magnum opus*, *Truth and Method* (1975), is dedicated to asking how the Heideggerian question of truth, as developed through a hermeneutic framework, can accommodate Dilthey's methodological hermeneutics of the human sciences. According to Gadamer, it was not essential for the human sciences to develop an exact method if they were to compete with the natural sciences. For while the natural sciences proceeded by way of logical induction, the human sciences had no such starting point or *method*. The human scientist, in other words, used insight and tact as tools rather than the laws of induction. Gadamer places the blame for the decline of the human sciences at the feet of those who, like Kant and his successors, considered the natural sciences as being eminently more objective than their subjective and aesthetic humanistic counterparts.

In his attempt to resuscitate the human sciences, Gadamer relies heavily on Heidegger's contention that we come to understand ourselves through our projects and possibilities. Preunderstanding, thus, is the condition of self-awareness. As such, and this is what it means to be caught in the hermeneutic circle, we come to an understanding of ourselves and our world by virtue of the prejudices of our specific time and place; that is, because *Dasein* is always already temporally situated, its preunderstanding will be conditioned and prejudiced by the world into which a person has been born. For instance, the reason why I interpret a rock *as* a weapon is because those who have preceded me in my tradition interpreted it *as such*. Interpretation, we will recall, is never neutral.

The question which Gadamer prioritizes, however, concerns the veracity of such interpretation; in an effort to avoid a crude form of relativism, Gadamer asks what we must do in order to avoid being duped by fallacious preunderstandings. As a devout hermeneut, he does not, of course, subscribe to the belief that misunderstanding can be fully eradicated. Interpretation, by its very nature, involves misreading and misconstrual. What we need to guard against, however, is the general propensity to take our prejudices at face value. While we cannot escape the prejudices of

our preunderstandings, we can nevertheless subject those prejudices and biases to critical scrutiny. Gadamer is aware that the interpreters have no vantage point outside their time and place to evaluate neutrally the efficacy of their preconceptions and prejudices. They rely, therefore, on what he terms 'temporal distance'; with the insight of hindsight, we can discern what interpretations have positively withstood the test of time. While he is cognizant of the fact that temporal distance is by no means fail-safe or fully reliable, Gadamer nevertheless acknowledges that it is perhaps the best available strategy for critically discerning between rival interpretations. The task for the individual is to try to develop a dialectical accommodation between the self that the individual has become by being on the receiving end of a tradition, and the indivdual's capacity to assume a critical disposition in relation to that tradition. Becoming aware that one is historically conditioned, and that one's beliefs derive not from a divine source, but from one's predecessors, is the first step towards such an accommodation.

It is through the process of responding to the claims which my tradition makes upon me that I come to understand the self that I am. Self-understanding, for Gadamer, involves dialoging with all those whose voices speak through me. I come to an understanding of *who* I am by reading and responding to those into whose heritage and legacy I have been situated. Hence, I am always already engaged in conversation with others who have gone before. The self is, in other words, dialogical through and through.

The problem with method, according to Gadamer, is that it never takes sufficient account of dialog, of the ongoing dialog within and beyond tradition. The form such dialog takes is that of a dialectic between question and answer. This ensures that no position can ever assume that it is beyond critical appraisal. The answer in such a dialectic will always beg yet another question. By standing back from one's tradition, not outside it, one can ask the relevant questions of it from one's particular vantage point, thus paving the way for our successors to ask vital questions of our own position. If there were no such dialog, tradition would simply freeze over. It is precisely by virtue of the fact that we endeavor to understand ourselves through a critical dialog of this type, that we can talk of historical progress.

Temporal distanciation is, therefore, that which allows me to observe a world which, although related in many respects to my own, is in many other senses different from the one I inhabit. The dialogue which pertains between the two worlds—my own and that of my forebear—is realized in what Gadamer calls a 'fusion of horizons.' This suggests that no tradition, horizon, or world is totally closed in upon itself, that it is indeed porous and open-ended. It demonstrates also that tradition cannot innoculate itself against the tensions which emerge in the course of a critical dialog with its successor traditions. Such is the way we are 'affected' by history (*Wirkungsgeschichtlichkeit*).

4. *Critical Theory and Critical Hermeneutics: Habermas and Ricoeur*

4.1 *Jürgen Habermas and Critical Theory*

One way of appraising the Heidegger–Gadamer tradition of hermeneutic thought is to observe the way in which it interacts with what is known as the 'critical theory' school, the main protagonists of which are Georg Lukacs, Theodor Adorno, Walter Benjamin, Herbert Marcuse, and Jürgen Habermas. Primarily, critical theory appropriates the dialectical methods of Hegel and Marx, so as to develop a distinctive brand of socially committed analysis. The ultimate meaning of our world, they contend, cannot be divorced from its historical and political context. Benjamin and Habermas each developed the dialectical method of humanist Marxism in its dual capacity as critique of ideology and a project of emancipation. As such, they owed their greatest debt to Marx's original recognition that it was only by combining our practical and reflective activities that we might effectively surmount prevailing forces of domination and control.

It is within the context of Habermas's critique of ideology that the main differences between hermeneutics and critical theory emerge most clearly. For Habermas (1929–), what blocks the way to genuine social relations are the dominant illusions that inhibit fair and equal communication between social beings. The way to achieve such a goal is to keep in focus the utopia of, what he calls, an 'ideal speech situation.' This amounts to a democratic system of communication in which all social agents would have the chance to speak and make themselves heard in an undistorted fashion. It would be made possible when all participants in discourse possess an equal opportunity to initiate communication and become transparent by their words and actions. Hitherto, communication has been distorted by unequal distribution of dialog opportunities which sustain privileged positions.

In articulating the demands of the ideal speech situation, Habermas utilizes the psychoanalytic technique of identifying the forces which systematically distort communication. This model of *Sprachanalyse* ('the analysis of speech') amounts to a 'metahermeneutic' which endeavors to disclose the subject's quest for self-understanding through a process of desymbolization and resymbolization. Such a technique goes beyond, in Habermas's view, both the romantic and phenomenological deployment of hermeneutics which are founded upon a spontaneous recovery of intersubjective agreement. Where Gadamer sought to surmount misunderstanding by way of a question and answer retrieval of some original dialogical understanding, Habermas opts for the more arduous route

through a critique of ideology leading to the possibility of genuine language. The notion of unrestrained communication remains for Habermas, therefore, a utopian ideal.

Romantic and phenomenological hermeneutics are, according to Habermas, hindered by their refusal to think through sufficiently the sociopolitical dimension of understanding. He believes that Gadamer's ontological and poetic approach to the question of language is naive, to the extent that it is predicated upon the erroneous assumption that a consensus of dialogical communication could be achieved by recovering our traditional 'belongingness' to being. For Habermas, the construction of genuine consensus cannot dispense with a project of historical action. Hence, where the romantic hermeneutic spoke of a retrieval of lost meaning, Habermas speaks of an anticipation of possible meaning. In contrast to Gadamer's view that misunderstanding presupposes a prior or preunderstanding, Habermas insists that where there is misunderstanding there may exist a posterior understanding if and only if we first undertake a critique of those ideological distortions which reside at the origin of misunderstanding. What is given is always already deformed communication. Habermas, that is, wishes to show how the ideal of communicative competence is not an ontological precondition of understanding but a historical possibility yet to be achieved. Only in this manner can we reconcile the philosophers who interpret the world and the social agents who change it.

4.2 Paul Ricoeur and Critical Hermeneutics

If Habermas criticizes romantic and phenomenological hermeneutics for being too focused on the past, and not sufficiently attentive to its ideological deformations, French philosopher Paul Ricoeur (1913–) tries to find a mid-passage between the metahermeneutics advanced by Habermas, and the hermeneutics of tradition evoked by Gadamer. In so doing, he develops a 'critical hermeneutics' which dissents from the romantic idea that we understand the past by reproducing in the present some original production of meaning, as if the temporal distanciation of meaning could be in some way erased. A critical hermeneutics of tradition insists on the necessity to discriminate between true and false interpretations of history. This, of course, raises the crucial question of legitimation. To resolve this problem, Habermas appealed to an ahistorical ideal of undistorted communication. For Ricoeur, however, such a criterion of legitimacy may be deferred to an indefinite future without any grounds in history. He recommends, accordingly, that we interpret tradition's claim to truth in a nonabsolutist sense of a *presumption* of truth. This suggests that we accept the truth claims of tradition until such a time as a better argument prevails. The presumption of truth refers to our basic attitude of credit or trust in the propositions of meaning legacied by the past—a primary response which precedes the critical moment of distanciation and reminds us that we are not the originators of truth but that we always already belong to a context of what he terms, 'presumed truth.' Ricoeur believes that such a model bridges the divide between the finitude of hermeneutic imagination (Heidegger and Gadamer) and the validity of the ideal undistorted communication (Habermas).

Ricoeur inserts Habermas's critique of ideology into the heart of the hermeneutics of tradition. This is born of his conviction that such a project demands a fundamental respect for tradition if it is to safeguard itself against arbitrary or ahistorical voluntarism—that is, a future project completely divorced from the historical heritage of the past. Where Gadamer invokes the notion of a common understanding which precedes us, the critical consciousness of Habermas reinterprets it in terms of the ideal of unrestricted and unconstrained communication. This antithesis is, however, broken down if one espouses a critical hermeneutics that realizes that critical theory cannot take up a view from nowhere, but must stick to its professed principles by presupposing some form of historical memory. For critical theory, a historical memory would not be that of romanticism (as it was for Gadamer) but rather of the Enlightenment, understood as a project of emancipation.

Tradition, Ricoeur concludes, must be understood in the dynamic perspective of our being-affected-by-the-past, which in turn is related to our historical horizon of expectancy. In this larger dialectic between tradition and expectation, we rediscover suppressed potentialites of past meaning which give substance to the ideal of undistorted communication. Indeed, it is only in terms of such an interplay between memory and anticipation that the ideal image of a reconciled humanity can be invested with an effective history. A critical hermeneutics has thus the ethical duty of ensuring that the expectancies for the future are brought nearer to the present by being sensitive to the concrete actions which need to be undertaken so as to bring to a realization what is 'desirable and reasonable.' Second, it must continually strive to liberate the still untapped potentialities of inherited meaning.

In this way, Ricoeur's particular brand of hermeneutics provides an essential link between romantic hermeneutics and critical theory. His incisive analysis of myth throughout his multifarious writings shows precisely how such a critical hermeneutics can be deployed in the service of concrete sociopolitical objectives. The modern project of unmasking myth frequently takes its cue from the investigative methods developed by Marx, Nietzsche, and Freud—the 'three masters of suspicion' as Ricoeur likes to call them. Once a hermeneutics of suspicion has unmasked the alienating role of myth as an agency of ideological conformism, there remains, for Ricoeur, the task of a

positive interpretation. Hermeneutics, that is, has the double task of *suspecting* and *listening*. Having demythologized the ideologies of false consciousness it labors to disclose the utopian symbols of liberating consciousness. It thereby seeks to rescue mythic symbols from reactionary domination, and to show that once the mystifying function has been dispelled we may recover genuinely utopian anticipations of 'possible worlds' of liberty and justice. A positive hermeneutics of this sort offers an opportunity to rescue myths from the ideological abuses of doctrinal prejudice, nationalism, class oppression, or totalitarian conformism; and it does so in the name of a universal project of freedom—a project from which no creed, nation, class, or individual is excluded. The utopian content of myth differs from the ideological in that it is inclusive. It opens up human consciousness to a common goal of liberation instead of closing it off in inherited securities. That is, having eliminated the ideological abuse of myth as a false explanation of how things are, we are now free to appreciate the properly symbolic role of myth as an exploration of how things might be.

See also: Critical Theory: Contemporary; Critical Theory: Frankfurt School; Hermeneutics, History of; Husserl, Edmund (1859–1938); Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects; Verstehen und Erklren, Philosophy of

Bibliography

Bruns G 1992 *Hermeneutics Ancient and Modern*. Yale University Press, New Haven, CT

Dannhauer J C 1654 *Hermeneutica sacra sive methodus exponendarum sacrarum literarum*. Augsburg, Germany

Dilthey W 1976 *Selected Writings*. Cambridge University Press, Cambridge, UK

Gadamer H-G 1975 *Truth and Method*. Seabury Press, New York

Grondin J 1994 *Introduction to Philosophical Hermeneutics* [trans. Weinsheimer J]. Yale University Press, New Haven, CT

Habermas J 1971 *Knowledge and Human Interests*. Beacon Press, Boston

Habermas J 1974 *Theory and Practice*. Heinemann, London

Habermas J 1979 *Communication and the Evolution of Society*. Beacon Press, Boston

Habermas J 1984 *The Theory of Communicative Action*. Beacon Press, Boston, Vol. 1

Habermas J 1987 *The Theory of Communicative Action*. Beacon Press, Boston, Vol. 2

Heidegger M 1962 *Being and Time*. Blackwell, Oxford, UK

Ricoeur P 1974 *The Conflict of Interpretations*. Northwestern University Press, Evanston, IL

Ricoeur P 1981 *Hermeneutics and the Human Sciences*. Cambridge University Press, New York

Ricoeur P 1984–7 *Time and Narrative*. University of Chicago Press, Chicago

Ricoeur P 1986 *Lectures on Ideology and Utopia*. Columbia University Press, New York

Ricoeur P 1992 *Oneself as Another*. University of Chicago Press, Chicago

M. Dooley and R. Kearney

Heterosexism and Homophobia

Homophobia is a colloquial expression that refers to negative, fearful, or hateful attitudes and behavior toward gays and lesbians. Although homophobia is a common term used to indicate antilesbian and antigay prejudice, many social scientists object to the concept. Herek (1995), for instance, objects that prejudice against homosexuality is not a phobia in the clinical sense of the word; it is not based necessarily on fear, nor is it inevitably irrational or dysfunctional for the individuals who manifest it. He defines heterosexism as 'an ideological system that denies, denigrates, and stigmatizes any nonheterosexual form of behavior, identity, relationship or community.' Adam (1998), Plummer (1992), and Stein and Plummer (1994) find it useful to distinguish analytically and conceptually between heterosexism, which is defined as societal prejudice located in the structures, institutions, and material conditions of society, and homophobia, which is a manifestation of the individual-level attitudes and behavior reflective of heterosexism. Herek (1995) prefers the term 'psychological heterosexism' to denote individual-level antigay and antilesbian prejudice. Heterosexism and homophobia are mutually constitutive in that individuals feel that their antigay attitudes and behaviors are legitimate to the extent that these prejudices are entrenched in social institutions. Individuals are more likely to act out their prejudices when they consider them to be culturally legitimate and shared by others (Berrill 1992). Expressions of antigay prejudice reinforce ideologies and practices of intolerance and hostility and reinscribe homosexuality as something that should be despised and feared and should remain hidden.

1. Attitudes and Expressions of Homophobia

One common stereotype is that homosexuality is related to gender-role nonconformity. It is assumed frequently that lesbians manifest characteristics defined as masculine and gay men exhibit behavior associated with feminine characteristics. For many people, this presumed gender nonconformity is reason enough to be suspicious and intolerant of homosexuals. Gay men are also often assumed to be hypersexual, animalistic, promiscuous, overly visible, and conspiratorial. Lesbians are assumed to be manhaters, aggressive, socially nonconforming, and sex-

ually cold (Herek 1995). Until 1973 the American Psychiatric Association considered homosexuality to be a form of mental illness. Many people continue to hold this belief and also believe homosexuality to be immoral and antisocial. Expressions of homophobia include verbal assaults and derisive joking, negative stereotypical media representations, discrimination in employment, education, housing, medical research, legal defense, and custody rights, physical attack, and tacit approval of all forms of discrimination, including gay bashing. Various studies, including surveys of gays and lesbians based on probability samples, indicate that nearly half of all respondents have experienced some form of discrimination in employment or housing based on their sexual orientation. A meta-analysis of 24 questionnaire studies showed that a median of 44 percent of the respondents have been threatened with violence and 80 percent have been verbally harassed (Berrill 1992). All of the respondents attribute these discriminatory actions to the perpetrators' perceptions that the target was homosexual. In 1987 the US Department of Justice reported that homosexuals are the most frequent victims of hate violence.

1.1 Containing Sexuality

Several theorists have observed that homophobia is an expression of a culture that is confused and inconsistent about sexual behavior. Attitudes and actions reflect both obsession with and repression of sexual desire. One basis of tacit agreement is that sexuality is something that must be contained. Homosexuality, along with adultery, prostitution and pederasty, mark the boundaries by existing outside culturally approved forms of sexuality—what poet Rich (1980) refers to as compulsory heterosexuality. Social historians (e.g., Boswell 1980, D'Emilio and Freedman 1988, Katz 1976) note that the emergence of a sexual binary in the nineteenth century posits heterosexuality as natural and socially desirable by contrasting it with homosexuality, which is characterized as unnatural and immoral. This binary represents the general tendency in modern western thought to conflate morality and nature. According to this cultural equation, sexual desire is a manifestation of a natural drive to reproduce. The moral, disciplined individual contains this desire within the bounds of heterosexual monogamy. These attitudes are reflected in clinical, social and religious definitions of the homosexual as cursed, perverted, sexually obsessed, and evil and, until recently, mentally and physically sick.

2. Institutional Heterosexism

Pervasive negative stereotypes regarding lesbians and gay men suggest that these attitudes and beliefs are rooted in cultural ideologies and social practices. All of these attitudes and expressions reflect the prevailing belief that sexual desire for the opposite sex is expected and normal; sexual desire for someone of the same sex is abnormal and problematic. The lingering belief that homosexuality is dysfunctional and that it constitutes a state of anguish for the individual can be traced to the medicalization of homosexuality and its treatment as an illness. Many medical practitioners and clinicians continue to offer treatments intended to 'cure' lesbians and gays of their 'affliction.' These practices reflect the cultural belief that individuals will be unhappy and maladjusted if they do not experience sexual desire for persons of the opposite sex. Medical assessments that treat homosexuality as a form of disease reinscribe the notion that sexuality is primarily physiological and that heterosexuality is the healthy manifestation of human physiology.

Most contemporary religions also treat homosexuality as problematic. In some cases a distinction is made between homosexual feelings and homosexual acts (e.g., Catholic Catechism 1989). The experience of homosexual inclinations may be considered an affliction—a 'cross to bear'—but to act on one's homosexual feelings is considered a sin. Many religious leaders advocate celibacy for those 'afflicted' with homosexual tendencies. Some religious organizations sponsor intensive therapy programs intended to help the individual who is 'suffering' from homosexuality to recover. Similar to medical institutions, religious institutions take heterosexuality for granted as the normal and desired form of sexual organization and treat homosexuality as deviant, undesirable, dysfunctional, and, additionally, sinful.

Legal practices also reflect a cultural ideology that supports heterosexuality as normal and denigrates homosexuality as abnormal. The most prevalent indication of compulsory heterosexism is the widespread lack of legal protection from antigay/lesbian discrimination. Lesbians and gays who experience discrimination in employment, housing, education and immigration frequently find it difficult to obtain justice. In the US it is still legal to ban lesbians and gays from various public institutions including service in the military and employment in various educational domains. Another indication of cultural heterosexism is public resistance to attempts by lesbians and gay men to reduce discrimination and to seek justice. In the 1990s US citizens in several states voted against ballot initiatives that would ensure equal legal protection for lesbians and gays. The rhetoric employed to defeat these initiatives was that lesbians and gays were seeking 'special rights and privileges.' The logic underlying this argument was that the US constitution already guarantees every individual equal protection under the law. While this is the case in theory, advocates of the equality initiatives point out that in practice this protection is often not extended to lesbians and gays. In fact, many social officials,

including employers, housing agencies and health care providers actively refuse to grant lesbians and gays equal rights and openly cite deviant sexual orientation as the basis for this discrimination. US courts have routinely upheld these decisions

Not only does the law fail to uphold the rights of lesbians and gays, but it has been enforced differentially as a means of punishing lesbians and gays. One example of differential enforcement is the enactment of sodomy laws. A majority of states still have laws that proscribe oral or anal sex. In most cases, these laws refer only to the behavior and not to the gender of the participants. However, the only recorded arrests in the past several decades based on these archaic laws have been gays and lesbians. When challenged, several states (Kentucky, Louisiana, Michigan, and Texas) have ruled sodomy laws as unconstitutional. However, in 1986, in the case of Bowers v. Hardwick, the US Supreme Court upheld a Georgia sodomy law in a narrow vote of 5 to 4. The majority opinion, written by Chief Justice Warren Burger expresses the general legal failure to separate the expression of private individual sexual behaviors from perceived cultural expectations. In his brief he writes: 'To uphold that homosexual sodomy is somehow protected as a fundamental right would be to cast aside millennia of moral teaching' (quoted in Singer and Deschamps1994).

Court decisions such as Bowers vs. Hardwick galvanized many US citizens who were concerned with the implications of court-mandated moral behavior and invasion of privacy. As a consequence, in the 1990s lesbians and gays have fought actively against discrimination. Several gains have been made, but these gains have been matched in intensity by antigay and antilesbian legislation. The intent of initiatives such as Colorado's Amendment 2 (enacted in 1992) is to legally prohibit lesbians and gays from seeking legal protection for discrimination based on their status as homosexuals. This is the first proactive legislation in the history of US law that actively denies the pursuit of civil rights based on a single specific category—homosexuality. In an address to Congress regarding legislation intended to protect persons from discriminatory violence (hate crimes), Senator Dannemeyer makes clear this categorical prejudice. He notes: '(I)t is a Federal responsibility to ensure the equal protection of all citizens regardless of their race, religion, or ethnic origin. It is not a Federal obligation to protect citizens in their sexual orientation' (US Congress 1988, p. 12).

2.1 Permissible Prejudice

The extent to which homophobia is an institution is indicated in the decisions of law enforcement officials and judges who consider homosexuality sufficient grounds for justifying acts of prejudice, discrimination, and violence. This tacit approval constitutes a form of permissible prejudice. Persons presume their prejudices to be legitimate or permissible when these prejudices are condoned by persons in positions of authority. Not only do the words of Justice Burger and Senator Dannemeyer reveal the extent to which heterosexuality is considered normal and legitimate, they suggest that homosexuality is a form of deviance that is wrong and shouldn't be tolerated. These institutional practices buttress individual feelings of fear and hate and send the message that discrimination and violence will be tolerated. In contrast, legal and cultural practices in many countries and some states in the US reflect the intent to extend equal rights to all citizens. Gay and lesbian marriages are legally recognized in Denmark and the Netherlands, for instance, and of the 16 countries that comprise the North Atlantic Treaty Organization, only the UK and the US ban lesbians and gays from service in the military.

2.2 Internalized Homophobia

In a climate of permissible prejudice, many lesbians and gays live in a continual state of fear. The entrenchment of heterosexist expectations and the corresponding active prejudice against homosexuality also induces feelings of 'internalized homophobia.' Internalized homophobia is a form of self-disgust and hatred experienced by many lesbians and gays. The phrase 'internalized homophobia' suggests that individuals who are trying to come to terms with their homosexual desires absorb or take in attitudes of cultural prejudice and use these attitudes as a basis of self-judgement. If the prevalent cultural story-line is one of intolerance and debasement, then individual lesbians and gays are likely to have to deal with this in formulating their own sense of themselves. In a culture that presumes heterosexuality, gays and lesbians must decide continually whether or not to voice information indicative of a deviant sexuality. 'Coming out' is an ongoing process. This process involves wrestling with and reconciling family, peer and religious expectations with personal feelings and experiences. Many lesbians and gay men fear negative consequences for revealing their sexuality. The majority experience rejection by their families, friends and religious and working communities when they reveal their homosexuality. Fear of rejection and a lack of positive role models can lead to confusion and self-hatred. Information compiled by the US Department of Health and Human Services in 1989 indicates that lesbian and gay youth are two to three times more likely to attempt suicide than their heterosexual peers. Gay and lesbian youth account for 30 percent of all completed youth suicides. In contrast, religious and community organizations that endeavor to understand and support gay and lesbian youth have been successful in reconciling family conflict and in integrating these youth into their schools and communities.

3. Explanations for Homophobia/Heterosexism

Adam (1998) delineates three theoretical approaches to understanding homophobia. Each approach is distinguished by a conceptual term denoting a specific theoretical orientation to anti-homosexual prejudice: Homophobia, Heteronormativity, and Heterosexism. Adam links these concepts to specific disciplines: Psychology, Cultural Studies (literary), and Sociology, respectively. The distinguishing aspect of each approach is the analytical unit: Personality, Discourse, and Social Structure. Explanations can be grouped accordingly. Within the field of psychology, Herek (1995) reviews the 'functional' explanations ('people hold and express particular attitudes because they derive psychological benefit from doing so'). Herek proposes three principle functions underlying attitudes about homosexuality: experiential, social identity, and ego-defense. A person's attitude may be a consequence of personal experiences (positive or negative); a prejudiced attitude may be intended as an expression of allegiance to a particular social group; or it may be a defensive response to personal psychological conflicts (such as gender expectations). He emphasizes that it is important to recognize the nexus between these functional attitudes and cultural heterosexism. The source and expression of these attitudes directly reflects a person's perceptions of cultural expectations.

The development of literary cultural studies in the 1990s has generated several interrelated theories which come together under the term, 'queer theory.' Queer theorists (e.g., Butler 1999, Sedgewick 1990, Seidman 1997) emphasize discursive practices that promote certain lines of thought and action and discourage or punish others. Heteronormativity is one of the most prevalent discourses in contemporary western culture. These theories focus on the manifestations of heterosexuality in art, media, education, legal discourse, religion, and so forth. According to these theories, the dominant cultural discourse is saturated with images and ideals that promote heterosexuality and condemn homosexuality (or anything that does not fit the expected monogamous, heterosexual, family idea). Queer theorists have particularly focused on the experience of the 'closet'—physical and discursive practices intended to keep homosexuality hidden and contained in separate spheres.

Adam subsumes the psychological and cultural studies approaches within the sociological. This approach emphasizes social structure, particularly the ways in which psychological attitudes and the values revealed in cultural discourses are manifest and maintained through institutional structures and practices. Change, he suggests, will be a result not only of education against prejudiced attitudes and transgressive acts intended as critiques of taken-for-granted cultural practices, but the extent to which this education and alternative expression permeates and unsettles existing institutional arrangements (such as the family, law, and religion). Seidman et al. (1999) suggest that recent gay and lesbian social movements have not only served to challenge some of these existing arrangements, but that they may have the paradoxical effect of 'decentering' lesbian and gay identities. According to their logic, the conditions of the closet may compel individuals to make their homosexuality a primary focus and thus, to solidify it as a primary identity. Paradoxically, normalization of homosexuality may enable individuals to de-emphasize this aspect of their behavior.

3.1 Gender

Understanding homophobia is complicated by the strong link between gender and sexuality in this culture. Gender is the basic criterion by which we select sexual partners. The violation of expected gender behavior elicits taunting, prejudice, and physical and social abuse. Frequent targets of this prejudice include the mannish lesbian and the effeminate gay male. Both evoke stereotypes that represent violations of gender expectations and hint at sexual transgressions. Transgendered persons are frequent targets of violence (see *Transsexuality, Transvestism, and Transgender*). It is difficult to know if attackers are provoked by what they assume to be a sexual aberration or a gender violation. Given the tight connection between gender and sexuality, it is likely that any deviation is perceived as a threat to heterosexuality. In the 1990s Lamba Legal Defense documented several cases in the US of murdered transgendered persons in which the murderer was released by the courts. The rationale offered in each case was that to arouse sexual desire in another under 'false gender pretense' was to invite violent 'self-defense' when the pretense was revealed (Singer and Deschamps 1994). This rationale can be interpreted as a manifestation of homophobia; individuals who do not conform to the expected norms of gender and sexuality do not deserve social justice.

4. Social Movements

The social protests and political activism associated with lesbian and gay history can be interpreted as a response to widespread, organized manifestations of homophobia. Several observers have suggested that the struggle for gay and lesbian rights was the civil rights issue of the 1990s. This history represents a fight for rights that are constitutionally guaranteed to all Americans but frequently withheld from homosexuals. Many local and national lesbian and gay movements have been organized as a direct response to political initiatives designed to legalize discrimination against homosexuals. The Defense of Marriage Act introduced into US federal legislation in 1996 is one such case. Marriage between same-sex couples poses no legal threat or loss of benefits to heterosexual couples.

It is difficult to interpret the urgency with which states have pursued legislation to restrict same-sex marriages as anything other than a display of mass homophobia. In the early 1990s antigay/lesbian legislation appeared on the ballots of 41 states and the District of Columbia. This sort of legislation is indicative of the extent to which homophobia is culturally legitimated as an institutionalized form of prejudice. Critics who suggest that antidiscrimination laws are 'special rights' fail to comprehend the extensive reach of homophobic actions into the everyday lives of lesbians, gays, bisexuals, and transgendered persons.

See also: Feminist Movements; Feminist Theory: Radical Lesbian; Foucault, Michel (1926–84); Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender, Feminism, and Sexuality in Archaeological Studies; Homosexuality and Psychiatry; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Libraries: Digital, Electronic, and Hybrid; Masculinities and Femininities; Queer Theory; Rationality and Feminist Thought; Regulation: Sexual Behavior; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Preference: Genetic Aspects; Sexuality and Gender; Sexuality and Geography; Social Movements and Gender; Transsexuality, Transvestism, and Transgender

Bibliography

Adam B 1998 Theorizing homophobia. *Sexualities* **1**: 387–404

Berrill K T 1992 Anti-gay violence and victimization in the United States: An overview. In: Herek G, Berrill K (eds.) *Hate Crimes: Confronting Violence against Lesbian and Gay Men*. Sage, Newbury Park, CA pp. 19–45

Boswell J 1980 *Christianity, Social Tolerance, and Homosexuality*. University of Chicago Press, Chicago

Butler J 1999 *Gender Trouble*. Routledge, New York

Catechism of the Catholic Church 1989 *Article Six: Offenses Against Chastity*. Paulist Press Mahwah, NJ, pp. 566–7

D'Emilio J, Freedman E 1988 *Intimate Matters: A History of Sexuality in the United States*. Harper and Row, New York

Herek G 1995 Psychological heterosexism in the United States. In: D'Augelli A, Patterson C (eds.) *Lesbian, Gay and Bisexual Identities over the Lifespan*. Oxford University Press, Oxford, UK, pp. 321–45

Katz J 1976 *Gay American History*. Crowell, New York

National Institute of Justice, US Department of Justice 1987 *The Response of the Criminal Justice System to Bias Crime: An Exploratory Review*

Plummer K 1992 Speaking its name: Inventing a lesbian and gay studies. In: Plummer K (ed.) *Modern Homosexualities*. Routledge, London, pp. 3–28

Rich A 1980 Compulsory heterosexuality and lesbian existence. *Signs: Journal of Women in Culture and Society* **5**: 631–60

Sedgewick E K 1990 *Epistemology of the Closet*. University of California Press, Berkeley, CA

Seidman S 1997 *Difference Troubles: Queering Social Theory and Sexual Politics*. Cambridge University Press, Cambridge, UK

Seidman S, Meeks C, Traschen F 1999 Beyond the closet? The changing social meaning of homosexuality in the United States. *Sexualities* **2**: 9–34

Stein A, Plummer K 1994 I can't even think straight: Queer theory and the missing sexual revolution in sociology. *Sociological Theory* **12**: 178–87

Singer B L, Deschamps D 1994 *Gay and Lesbian Stats*. New Press, New York

US Department of Health and Human Services 1989 *Report of the Secretary's Task Force on Youth Suicide*

US Congress 1988 Hate Crimes Statistics Act. *Report from the Committee on the Judiciary, 100th Congress, 2nd Session*. Government Printing Office, Washington, DC

J. O'Brien

Heuristics for Decision and Choice

Heuristics are approximate strategies or 'rules of thumb' for decision making and problem solving that do not guarantee a correct solution but that typically yield a reasonable solution or bring one closer to hand. As such, they stand in contrast to algorithms that produce a correct solution, given complete and correct inputs, if one exists. More specifically, heuristics are usually thought of as shortcuts that allow decisions or solutions to be reached more rapidly, even in conditions of incomplete or uncertain information—often because they do not process all the available information. Heuristics are most commonly studied in psychology within the domains of judgment and decision making, social cognition, and problem solving, and in computer-based applications in artificial intelligence (AI) and operations research (OR). This entry focuses on heuristics proposed as models of how people make decisions and choices.

1. Historical Overview

The term 'heuristic' (from a Greek root, 'to discover') was used for most of the twentieth century to refer to useful, even indispensable, strategies for finding solutions to problems that are difficult to approach by other means. Gestalt psychologists spoke of heuristic reasoning methods such as 'looking around' and 'inspecting the problem' to guide the search for useful information in the environment, while mathematicians employed heuristics including 'examining special cases', 'exploiting related problems', 'decomposing and recombining', and 'working backwards' (Groner 1983). These rather vague strategies were made more precise in computer-based models of human problem

solving and reasoning largely based on the means-ends analysis heuristic, which sought some way to reduce the distance between the current partial-solution state and the goal state (e.g., in Newell and Simon's early General Problem Solver system; see *Artificial Intelligence: Search*). Such general purpose or 'weak' methods proved insufficient to tackle many problems, so research in AI in the 1970s turned to collecting domain-specific rules of thumb from specialists in a particular field and incorporating these into expert systems. Around the same time, mathematicians working in OR faced new results from computational complexity theory indicating that efficient algorithmic solutions to many classes of challenging combinatorial problems (such as the Traveling Salesman Problem) might not be found; as a consequence, they too turned to the search for problem-specific heuristics, though through invention rather than behavioral observation (Müller-Merbach 1981).

After 1970, though, heuristics gained a different connotation in psychology: fallible cognitive shortcuts that people often use in situations where logic or probability theory should be applied instead. The 'heuristics-and-biases' research program launched by Tversky and Kahneman (1974, Kahneman et al. 1982) emphasized how the use of heuristics can lead to systematic errors and lapses of reasoning (see *Decision Biases, Cognitive Psychology of*) indicating human irrationality. The heuristics studied (see Sect. 2) were often vaguely defined and broadly applicable to judgments made under uncertainty in any domain, akin to the weak methods explored earlier in AI. This negative view of heuristics and of the people who use them as 'cognitive misers' employing little information or cognition to reach biased conclusions has spread to many other social sciences, including economics (Rabin 1998) and law (Hanson and Kysar 1999). More recently, a new appreciation is emerging within psychology that heuristics may be the only available approach to decision making in the many problems where optimal logical solutions are computationally intractable or do not exist (as OR researchers realized), and that domain-specific decision heuristics may be more powerful than domain-general logical approaches in other problems (as AI found). This has led to the study of precisely-specified heuristics matched to particular decision tasks (see Sect. 3), and the ways that learning and evolution can achieve this match in human behavior (e.g., Payne et al. 1993, Gigerenzer et al. 1999). The existence of such evolved adaptive heuristics has already been widely accepted for other animals in research on rules of thumb in behavioral ecology (Gigerenzer et al. 1999).

2. *Heuristics for Probability Judgment*

Since the rise of the heuristics-and-biases research program in the 1970s, the heuristics most widely studied within the social sciences are those that people use to make judgments or estimates of probabilities or other quantities in situations of uncertainty where they do not have complete, exact information (Tversky and Kahneman 1974). The representativeness heuristic is a means to assess the probability that an object A belongs to a particular category B (e.g., that a person described as meek is a pilot) or that an event A is generated by a particular process B (e.g., that the sequence HHTTTT was generated by throwing a fair coin). This heuristic produces probability judgments according to the extent that A is representative of, or similar to, B (e.g., meekness is not representative of pilots, so a meek person is judged as having a low probability of being a pilot). The availability heuristic can be used to produce assessments of class frequency or event probability based on how easily instances of the class or event can be mentally retrieved (e.g., plane crashes may seem like a frequent cause of death because it is easy to recall examples) or constructed (via the simulation heuristic). The anchoring-and-adjustment heuristic produces estimates of quantities by starting with a particular value (the anchor) and moving away from it (e.g., people asked to quickly estimate either $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ or $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$ give a higher value in the former case, which could come from estimation based on multiplying the first few values together, creating a higher or lower anchor, and then adjusting upwards, which yields a higher overall estimate of the first product).

Most researchers agree that such heuristics are widely used because they usually lead to good choices without much time or mental effort. Despite this, most of the large body of evidence amassed that is consistent with the use of these heuristics comes from showing where they break down and lead to errors (e.g., availability judgments can be manipulated by vivid examples like plane crashes, and anchoring can lead to different choices solely due to the order of information presentation—see *Decision Biases, Cognitive Psychology of*), leading many to conclude that people are poor decision makers after all. These heuristics have also been criticized as being vague redescriptions of observed choice behavior that are therefore difficult to falsify; however, more specific versions (e.g., testing whether availability works in terms of ease of recall or number of items recalled) are also being explored.

3. *Heuristics for Choice between Alternatives*

Specifying decision tasks in more detail—for instance, choosing between two (or more) alternatives on the basis of cues or pieces of information known about each—allows still more specific heuristics to be investigated. The typical normative approach to such choice decisions involves weighting and combining (typically adding) all of the available cues, with the weights determined by a process such as multiple regression. Heuristics can shortcut this process in two main ways.

First, rather than weighting all cues differentially, positive and negative cues for each alternative can merely be tallied (Dawes 1979). Second, not all cues need be used—instead, a stopping rule can indicate when enough cues have been considered to allow a choice (see below). Despite these simplifications, such heuristics often perform very well when compared to expert human decision makers (Dawes 1979), traditional benchmarks (Payne et al. 1993), or objective measures of correct answers (Gigerenzer et al. 1999).

There are several forms of stopping rules that can be used to limit the information considered by choice heuristics. The recognition heuristic (Gigerenzer et al. 1999) only ever considers the single cue of recognition: when one alternative is recognized and the other is not, this heuristic chooses the former (and otherwise guesses). One-reason decision mechanisms (Gigerenzer and Goldstein 1996) compare both alternatives on a single cue-dimension (or reason) at a time and stop at the first dimension that distinguishes between the alternatives (i.e., where their cue values differ, allowing a decision to be made). Accumulator or horse-race models (Busemeyer and Rapoport 1988) tally the cues in favor of each alternative until a specified difference in positive evidence is reached (e.g., until one alternative is three reasons ahead of the other). Elimination heuristics (Tversky 1972, Gigerenzer et al. 1999) begin with a set of alternatives and successively remove subsets that do not match a specified cue value until a single choice remains (e.g., to choose a restaurant, eliminate all those in town over some price threshold, then eliminate all those beyond some distance away, and so on until one restaurant is left). In every case (aside from the recognition heuristic), a search rule must also be specified to determine the order in which cue dimensions are considered; this order can (among other possibilities) be random, or based on recently used cues, or on cue validity or correlation with the choice criterion (which, when combined with a one-reason stopping rule, yields a lexicographic strategy such as the Take The Best heuristic—Gigerenzer and Goldstein 1996).

4. *Heuristics for Sequential Search*

When choice alternatives are encountered sequentially over time rather than simultaneously (e.g., in comparison shopping by driving from one store to the next), heuristics can be used to decide when to stop searching and select an available alternative. The traditional normative approach is to search until one finds an alternative below a precalculated reservation price balancing expected benefit of further search against its cost; this requires full knowledge of the search costs and distribution of available alternatives. Heuristics that use a simplified version of the reservation price calculation (e.g., replacing an integral with a weighted sum) come very close to normative performance (e.g., at selecting good prices, Moon and Martin 1990). Other heuristics require less knowledge, such as 'Keep searching until the total search cost exceeds 7.5 percent of the best price found' (Moon and Martin 1990). Herbert Simon's bounded rationality principle of satisficing (see *Bounded Rationality*) suggests setting an aspiration level equal to an alternative that is good enough (rather than optimal) and searching until that aspiration level is met; exactly how the aspiration level can be set varies with the search setting (e.g., whether it is a one-sided search like shopping, or a two-sided mutual search like finding a mate, Gigerenzer et al. 1999). Finally, another type of search heuristic that people use stops search after a particular pattern of alternatives is encountered, rather than after some threshold is exceeded (despite the fact that pattern should not matter from a normative perspective). For instance, the 'one-bounce' and 'two-bounce' rules say to keep searching for a low price until prices go up for the last or two last alternatives, respectively (Moon and Martin 1990).

5. *Further Directions*

Decision heuristics have been studied in different research traditions, in particular one that has focused on when and where verbally described heuristics can break down and yield deviations from classical norms of rationality (Kahneman et al. 1982), and another that has investigated how specific computationally modeled heuristics can exploit structured information to yield fast and accurate decisions (Gigerenzer et al. 1999). Further research should explore not only those aspects of decision environments that lead to different levels of performance, but also what factors lead people to use heuristics at all (e.g., time pressure, Payne et al. 1993), to choose between different possible heuristics (e.g., the influence of domain-specific emotions or social norms), and to learn or develop new heuristics (e.g., through social imitation).

See also: Bounded and Costly Rationality; Decision Biases, Cognitive Psychology of; Decision Making (Naturalistic), Psychology of; Decision Making: Nonrational Theories; Decision Research: Behavioral; Decision Support Systems; Decision Theory: Classical; Heuristics in Social Cognition; Informal Reasoning, Psychology of; Problem Solving (Everyday), Psychology of; Sequential Decision Making

Bibliography

Busemeyer J R, Rapoport A 1988 Psychological models of deferred decision making. *Journal of Mathematical Psychology* **32**: 91–134

Dawes R M 1979 The robust beauty of improper linear models in decision making. *American Psychologist* **34**: 571–82

Gigerenzer G, Goldstein D G 1996 Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* **103**: 650–69
Gigerenzer G, Todd P M, ABC Research Group 1999 *Simple Heuristics That Make Us Smart*. Oxford University Press, New York
Groner R (ed.) 1983 *Methods of Heuristics*. Erlbaum, Hillsdale, NJ
Hanson J D, Kysar D A 1999 Taking behavioralism seriously: The problem of market manipulation. *New York University Law Review* **74**: 630–749
Kahneman D, Slovic P, Tversky A 1982 *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York
Müller-Merbach H 1981 Heuristics and their design: A survey. *European Journal of Operational Research* **8**: 1–23
Moon P, Martin A 1990 Better heuristics for economic search. Experimental and simulation evidence. *Journal of Behavioral Decision Making* **3**: 175–93
Payne J W, Bettman J R, Johnson E J 1993 *The Adaptive Decision Maker*. Cambridge University Press, New York
Rabin M 1998 Psychology and economics. *Journal of Economic Literature* **36**: 11–46
Tversky A 1972 Elimination by aspects: A theory of choice. *Psychological Review* **79**: 281–99
Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* **185**: 1124–31

P. M. Todd

Heuristics in Social Cognition

'Heuristics' are strategies of simplifying judgments that allow individuals to make decisions under suboptimal circumstances. The discovery of heuristics has had a profound impact on social psychology (see *Social Psychology*), especially in the field of 'social cognition,' which studies the attempts by individuals to make sense of others (see *Mental Representation of Persons, Psychology of*). This article describes the general idea of heuristic information processing, looks at three specific heuristics, and examines the importance of heuristics for social psychology.

1. *Judgments under Suboptimal Conditions*

Whereas traditional theories of human decision-making (see *Decision Making, Psychology of*) have focused on the normative aspects of valid judgments while neglecting the context in which they occur, the 'heuristics' approach has directed its attention to the psychological processes that enable individuals to make judgments and decisions under situational, motivational, and cognitive conditions that are less than optimal. Because decisions typically are made in a state of uncertainty, it has proven fruitful to identify the strategies that individuals actually use to arrive at solutions that do not represent the best possible outcome but simply meet specific criteria. The study of these simplifying rules of thumb, called 'judgmental heuristics,' has been spearheaded chiefly by Tversky and Kahneman (1974). Because the concept of heuristics is associated closely with these authors, the application of this term is often limited to the specific mechanisms they have identified. However, the underlying logic of heuristic processing also applies to other strategies that serve the same purpose, that of simplifying human judgments and making them feasible under suboptimal conditions.

Because generally it is not possible to identify the use of a heuristic procedure from the outcome of a judgment, researchers have had to create specific conditions under which heuristics lead to errors. As a result, heuristics have often been associated with the supposed irrationality of human reasoning (e.g., Nisbett and Ross 1980). However, just as visual illusions do not testify to the deficiency of human perception, the errors produced by judgmental heuristics are not proof of the inadequacy of human judgment in general; instead, they point to mechanisms that allow individuals to make acceptable decisions under natural constraints.

The discussion that follows will introduce the three most prominent heuristics.

2. *The Availability Heuristic*

In assessing the frequency or probability of an event (or the co-occurrence of several events), individuals often employ a strategy that is based on the ease with which bits of information can be retrieved or generated from memory. An employer wishing to gauge the rate of unemployment in their community may go to the trouble of obtaining the relevant information from official sources. But if they are not motivated or able to do that, they can try to think of unemployed friends or acquaintances. The more easily they are able to do so, the higher will be their estimate of the rate of unemployment. Tversky and Kahneman (1973) called this judgment strategy the 'availability heuristic.'

Additional judgmental phenomena in the social domain are connected to the availability heuristic. One example is a risk whose assessment depends on the frequency with which a type of event occurs. Thus, riding a motor bicycle is risky to the extent that accidents occur frequently. However, the actual frequency of an event and the ease with which it comes to mind may be dissociated if certain events are more likely to be reported in the media. For example, Lichtenstein et al. (1978) found that causes of death frequently reported in the press were greatly overestimated in terms of their frequency. While heart disease causes 85 percent more deaths than accidents,

only 20 percent of those surveyed thought that heart disease was the greater risk. The conspicuousness of events also influenced their availability. For example, in the study by Lichtenstein et al. (1978), the overestimated causes of death were especially dramatic and sensational (murder, flood, automobile accident), while the rather inconspicuous causes of death (heart disease, cancer, diabetes) were underestimated. Events that are known from personal experience are also more readily available (and therefore judged more likely to occur) than events that are only known through third parties (Greening et al. 1996).

It should be noted that 'availability' has two psychological components that usually are confounded: the content that comes to mind and the ease (or effort) experienced while retrieving the information from memory (Schwarz et al. 1991). Several studies from the domain of health psychology demonstrate the direct influence of experienced ease of retrieval on the assessment of risk. For example, Rothman and Schwarz (1998) found higher ratings of risk for contacting a heart disease when participants were asked to list three (easy to generate) instead of eight (difficult to generate) factors that increased their own risk.

Attitudinal judgments (see *Attitude Formation: Function and Structure*) about everyday topics are also made on the basis of the availability heuristic under certain conditions. Wänke et al. (1996) reported more positive attitudes towards public transportation on the part of subjects if they had been asked previously to generate three instead of seven arguments in favor of public transportation. These experimental examples were constructed such that the retrieval of more information was associated with less experienced ease of recall. The results suggest that under suboptimal conditions, judgment is based on the experience of information retrieval and not on the content of what is retrieved.

Another example refers to the ease with which we are able to imagine a different course of events. Let us look at the following example of counterfactual thinking: 'If I had gotten up five minutes earlier this morning I would not have missed the train, I would not have been late for the exam and would have been able to read the one additional problem I needed to pass the exam.' What is crucial is that the ease with which an event can be undone mentally influences affective reactions to this event. While the person in this example would presumably feel a great deal of anger over being five minutes late, a person who overslept an hour instead of five minutes would be less angry. For a detailed account of recent findings on contrafactual thinking see Roese and Olson (1995).

Apart from these examples of judgments and affective reactions on the basis of the perceived ease of cognitive operations, the 'availability principle' in its general form, that is, the finding that increased accessibility of contents (see *Priming, Cognitive Psychology of*) and cognitive structures influences judgments, has stimulated a host of research in social psychology. This pertains, for example, to work on the categorization of persons, causal attribution (see *Attributional Processes: Psychological*), the constancy of opinions after they have been discredited, or the testing of hypothesis.

3. *The Representativeness Heuristic*

The representativeness heuristic refers to people's tendency to simplify categorical judgments by relying solely or excessively on similarity. For example, a person who wants to determine the profession of another person may use the target person's similarity to the typical member of this profession. Consequently, a student might decide that a fellow student at the next table in a university cafeteria is a business administration major if they display the characteristics of a typical MBA student, such as reading the business section of the paper and talking about the stock exchange. Although this strategy has a logical basis, it tends to neglect other types of relevant information, such as base rates. For example, this university may have many more students of law than of business administration. Information about such a frequency distribution should affect the judgment in the same way as information about the target person. However, research by Tversky and Kahneman (1982) has demonstrated that such base rates largely are neglected even if the individuating information is not very diagnostic.

This neglect of base rates is strikingly similar to findings from attribution research. A number of studies discovered that when people identify the causes of an observed behavior, they attribute it to characteristics of the person and neglect situational influences. This response tendency implies that judges underestimate 'consensus information' (Kelley 1967) that reflects the power of situations. 'Consensus information' describes how other people behave under the same circumstances and serves as a basic determinant of causal attributions (see *Attributional Processes: Psychological*). Specifically, if everybody shows the same behavior in a given context there is little reason to attribute an action to the unique characteristics of the actor. The tendency to neglect such base-rate information in social judgments and to give it less weight than it should have under normative considerations can be understood as a manifestation of the representativeness heuristic because it consists of drawing inferences on the basis of similarity ('a person who behaves aggressively is aggressive') at the expense of base-rate information (see *Person Perception, Accuracy of*).

The relationship between this attributional bias and heuristic processing was demonstrated subsequently by varying the conduciveness of the judgmental

situation (for a review, see Gilbert and Malone 1995). Researchers demonstrated that the neglect of situational influences depended on judges' cognitive resources, such that this bias was more likely to occur when people were distracted.

The use of the representativeness heuristic may also result in violations of logical principles. In one study by Tversky and Kahneman (1983), judges had to assess the probability of the joint occurrence of two characteristics of a target (a liberated woman). One feature was highly representative ('She is a feminist') while the other was not ('She is a bank teller'). In line with the representativeness heuristic, judges considered it less likely that the target person was a bank teller than that she was a feminist bank teller, which is, of course, logically impossible.

Other effects of the representativeness heuristic concern the misperception of chance and the neglect of sample size (Tversky and Kahneman 1974).

4. *Anchoring and Adjustment*

'Anchoring and adjustment' describes the phenomenon that judgments are assimilated toward a value that was initially considered. For example, a person who has to estimate the proportion of African nations in the UN may arrive at a higher percentage if they have been exposed previously to a high rather than a low standard of comparison. While such an influence would be hardly surprising if this 'anchor' is offered as a piece of information that is relevant to the judgment in question, Tversky and Kahneman (1974) were able to show that the resulting assimilation effect occurred even if relevance was ruled out by presenting the anchor as the outcome of a probabilistic process.

The psychological process that was suggested by Tversky and Kahneman (1974) for this phenomenon is an insufficient adjustment of the final judgment. Alternatively, Strack and Mussweiler (1997) have proposed a mechanism that is related to the availability heuristic, the so-called selective accessibility model. Here the anchoring effect involves two stages. In the first stage, judges engage in biased hypothesis testing when they consider the anchor as a possible value of the target. In this process, semantic information consistent with the anchor is activated. As a consequence of this selective activation, consistent information will be more accessible in a second stage when information for the final judgment will be retrieved. In this perspective, the anchoring effect is not the result of a numeric influence or a mere effect of insufficient adjustment but caused by a mechanism of semantic priming that has been demonstrated in many studies in social cognition (e.g., Higgins et al. 1977).

These supporting findings are not only important because they reflect the dynamics of many negotiation situations in which an initial offer greatly affects the final outcome but also because they are helpful in understanding the consequences of merely considering a possibility before forming a judgment. For example, it has been demonstrated that explaining a hypothetical event leads people to expect its actual occurrence (Ross et al. 1977). Similar outcomes were obtained for imagining a particular event (Carroll 1978). More dramatically, it was shown that merely thinking about an attitude object may lead a person to adopt more extreme attitudinal positions if he or she holds a schema about the object (e.g., Tesser et al. 1995). Perhaps the most relevant social domain is that of social comparisons (see *Social Comparison, Psychology of*). Here, work on the anchoring heuristic suggests a mechanism that explains why and under what conditions comparing ourselves to another person will make us appear similar to that person (Mussweiler and Strack 2000). The anchoring effect has also been invoked to explain the so-called hindsight bias (Fischhoff 1982).

5. *Other Heuristics in Social Psychology*

In the previous sections three basic judgmental strategies were introduced that were identified by Kahneman and Tversky as devices employed to form sufficiently accurate judgments under adverse conditions and with little cognitive effort, although under certain circumstances these strategies can lead to systematic distortions. The idea of heuristic information processing has been expanded subsequently to other heuristic cues, with lasting influence on theory formation in broad areas of social psychology.

A prominent example is the use of feelings as information. Specifically, it has been argued that under suboptimal circumstances, judges resort to their subjective experiences to generate a judgment on a different dimension. This has been demonstrated for both affective (e.g., Schwarz and Clore 1983) (see *Emotions, Psychological Structure of*) and nonaffective feelings, such as familiarity and mental effort (e.g., Strack and Neumann 2000). As discussed before, it has been shown that the availability heuristic operates through the use of the mental effort experienced by an individual engaged in a specific mental operation (Schwarz et al. 1991).

The idea of heuristics has become effective in a second line of research in social psychology. The study of persuasive communication (see *Communication and Social Psychology*), in particular, has profited from the distinction between heuristic and systematic processing (Chaiken 1987). In this domain, the reliance on peripheral cues (such as the expertise of the communicator or the length of the message) instead of on central features (i.e., the strength of the arguments) was shown to occur under suboptimal conditions. Similarly, it has been shown that the use of stereotypes

(see *Stereotypes, Social Psychology of*) obeyed the same regularities, such that people were more likely to use them when they were unwilling or unable to process individuating information (Bodenhausen and Lichtenstein 1987).

A final application and extension of the heuristics idea can be found in the domain of social influence (see *Social Influence, Psychology of*). As Cialdini (1993) has argued, people's proclivity to simplify their judgments can be exploited by agents of social influence. For example, an advertising agency may deliberately attempt to increase the ease with which certain information (presumably about their own product) may be brought to mind or try to associate positive feelings with the product. Cialdini has identified and described many more such strategies, which he called the most effective 'weapons of influence.'

6. Conclusion and Future Directions

Over the years, the concept of heuristics has generated many insights into the cognitive dynamics of social behavior. At the same time, however, the idea of heuristic processing has gone far beyond the three mechanisms described by Kahneman and Tversky and it has been claimed that the use of heuristics may possess biologically adaptive value (Gigerenzer and Todd 1999). In a more general perspective, a universal psychological principle has emerged that sheds light on the flexibility of human information processing under divergent epistemic goals. However, the work of finding the laws of this flexibility and integrating it into psychological theorizing still needs to be done.

See also: Decision Making, Psychology of; Heuristics for Decision and Choice

Bibliography

Bodenhausen G V, Lichtenstein M 1987 Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology* **52**: 871–80

Carroll J S 1978 The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of Experimental Social Psychology* **14**: 88–96

Chaiken S 1987 The heuristic model of persuasion. In: Zanna M P, Olson J M, Herman C P (eds.) *Social Influence: The Ontario Symposium*. Erlbaum, Hillsdale, NJ, Vol. 5, pp. 3–39

Cialdini R B 1993 *Influence. Science and Practice*, 3rd edn. Harper Collins, New York

Fischhoff B 1982 For those condemned to study the past: Heuristics and biases in hindsight. In: Kahneman D, Slovic P, Tversky A (eds.) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK, pp. 335–51

Gigerenzer G, Todd P M 1999 *Simple Heuristics that Make Us Smart*. Oxford University Press, New York

Gilbert D T, Malone P S 1995 The correspondence bias. *Psychological Bulletin* **117**: 21–38

Greening L, Dollinger S J, Pitz G 1996 Adolescents' perceived risk and personal experience with natural disasters: An evaluation of cognitive heuristics. *Acta Psychologica* **91**: 27–38

Higgins E T, Rholes W S, Jones C R 1977 Category accessibility and impression formation. *Journal of Experimental Social Psychology* **13**: 141–54

Kelley H H 1967 Attribution theory in social psychology. In: Levine D (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, Vol. 15, pp. 192–240

Lichtenstein S, Slovic P, Fischhoff B, Layman M, Combs B 1978 Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory* **4**: 551–78

Mussweiler T, Strack F 2000 Consequences of social comparison: Selective accessibility, assimilation, and contrast. In: Suls J M, Wheeler L (eds.) *Handbook of Social Comparison: Theory and Research*. Plenum, New York, pp. 253–70

Nisbett R E, Ross L 1980 *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall, Englewood Cliffs, NJ

Roese N J, Olson J M 1995 *What Might Have Been: The Social Psychology of Counterfactual Thinking*. Erlbaum, Mahwah, NJ

Ross L, Lepper M, Strack F, Steinmetz J L 1977 Social explanation and social expectation: The effects of real and hypothetical explanations upon subjective likelihood. *Journal of Personality and Social Psychology* **35**: 817–29

Rothman A J, Schwarz N 1998 Constructing perceptions of vulnerability: Personal relevance and the use of experiential information in health judgments. *Personality and Social Psychology Bulletin* **24**: 1053–64

Schwarz N, Bless H, Strack F, Klumpp G, Rittenauer-Schatka H, Simons A 1991 Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology* **61**: 195–202

Schwarz N, Clore G L 1983 Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology* **45**: 513–23

Strack F, Mussweiler T 1997 Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology* **73**: 437–46

Strack F, Neumann R 2000 Furrowing the brow may undermine perceived fame: The role of facial feedback in judgments of celebrity. *Personality and Social Psychology Bulletin* **26**: 762–8

Tesser A, Martin L L, Mendolia M 1995 The impact of thought on attitude extremity and attitude-behavior consistency. In: Petty R, Krosnick J A et al. (eds.) *Attitude Strength: Antecedents and Consequences*. Erlbaum, Mahwah, NJ, pp. 72–92

Tversky A, Kahneman D 1973 Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* **54**: 207–32

Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* **185**: 1124–31

Tversky A, Kahneman D 1982 Judgments of and by representativeness. In: Kahneman D, Slovic P, Tversky A (eds.) *Judgments under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York, pp. 84–98

Tversky A, Kahneman D 1983 Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* **90**: 293–315

Wänke M, Bless H, Biller B 1996 Subjective experience versus content of information in the construction of attitude judgments. *Personality and Social Psychology Bulletin* **22**: 1105–13

F. Strack

Hidden Curriculum in the Classroom

> Long after I've forgotten to use the power rule to derive an equation, or that the animals are classified by kingdom, phylum, class, order, family genus and species ... I will remember the other lessons I've learned, the lessons that taught me how to live (Jenny Lobasz, senior student at Dreyfoos School of the Arts).

The hidden curriculum generally refers to student learning that takes place in school as a result of actions by school personnel, parents, students, and other community agents. What is learned in the classroom is largely a function of the prevailing school conditions and classroom dynamics as well as such demographic factors as the social class, race, ethnicity, language, gender, and condition of disability of all school participants. Compared with the formal school curriculum, the hidden curriculum is not written, does not have explicit objectives, and varies considerably from one school setting to another. The hidden curriculum often referred to as the 'covert' or the 'informal' curriculum, is a most powerful influence in the classroom; it is estimated that up to 90 percent of what is learned is due to this influence (Massialas 1996).

1. *The School and Classroom Dynamics*

In many cases the formal curriculum is antithetical to the hidden curriculum. For example, the formal curriculum of the school explicitly promotes such values as academic learning, fairness, and good citizenship; the latter expressed as democratic participation in decisions affecting the individual. The hidden curriculum, on the other hand, teaches the students how to manipulate the system, i.e., how to cheat without being caught, how to 'apple polish the teacher,' and how to comply with the rules without questioning the legitimacy of the authorities for making them (Jackson 1983). The formal curriculum preaches democracy but the hidden curriculum imposes autocracy. The formal curriculum emphasizes scholarship and an equitable grading system but the hidden curriculum teaches students that scholastic attainment expressed through grades is a political process of give and take and of bargaining with the teacher. In this environment it is clear that if students do not master the lessons and the techniques of the hidden curriculum they are destined to fail.

Jackson (1983), being one in the first groups of researchers who studied the hidden curriculum in depth concluded that 'Learning how to make it in school involves, in part, learning how to falsify our behavior' (p. 50). He found out that teacher–student relations in the classroom are hierarchical—the teacher being at the top of the hierarchy and the students at the bottom. He observed that classroom activity in elementary schools usually centers on seatwork, group discussion, teacher question and answer sessions, as well as teacher demonstrations. Students may learn about the three Rs but they also learn other things, things that are not spelled out in formal instruction. For example, they learn to work in crowds and they learn to accept praise or reproof. They also learn how power is distributed in the system, that is, how certain students gain special privileges. Students also learn how to be patient and how to conform to institutional requirements and to tolerate minor frustrations.

Many studies confirm the claim that schools indirectly teach students to be obedient, passive, and compliant. In this they are supported by other social institutions that directly relate to the life of the students, e.g., the family. One author recognizes that, on the one hand, modern society seeks to develop self-reliant and critical thinking individuals. On the other, it promotes, through the schools and other socialization agencies, obedience. 'Little notice is taken of the legions of overly obedient children in the schools; yet, for every overly disobedient child, there are probably twenty who are obeying too much. There is little motivation to encourage the wheels to develop as noisy, creative, independent thinkers who may become bold enough to disagree' (McCarthy 1998, pp. 56–7).

A review of research on the influence of the hidden curriculum in schools in the Arab world revealed that these schools, much like their Western counterparts, impart passivity and obedience among the students (Massialas and Jarrar 1991). For example, in a study conducted in Morocco it was found that the 'primary school imparts the values and behaviors of the society, especially the home—absolute dependence on patriarchal commands, suppression of creative thought, strict compliance with rules, both written and unwritten, acceptance of things as they are' (Massialas and Jarrar 1991, p. 137). Similar values to those transmitted in the Moroccan schools were reported in studies in other Arab schools, e.g., in Saudi Arabia, in Lebanon, in Qatar. On the basis of these studies it is apparent that the Arab classroom replicates to the extent possible the patriarchal family structure. The teacher is invariably the dominant figure, the patriarch who conveys awe, respect, and obedience.

2. *Social Class and the Hidden Curriculum*

Student background factors such membership in a social class invariably affect how students are treated in school and their relations with the teachers and other school personnel. A study conducted in New Jersey, USA, found that social class clearly influenced

the values learned in school through the hidden curriculum (Anyon 1983). The study identified four types of schools—working class, middle class, affluent professional, and executive elite. Working class schools emphasized mechanistic, step-by-step instruction conducive to memory type of work. Class work was quite structured. There was very little in trying to provide an explanation of the world. The middle class schools emphasized the 'right answer'; instruction was largely teacher-directed; the majority of the lessons were drawn from the textbook. In the professional-affluent schools the emphasis was on fostering creativity and independent thought in the classroom. Students were constantly asked through projects to apply concepts in explaining their world. The executive elite school emphasized critical thinking, conceptual development, and decision making. Instruction stressed analysis of social issues and problem solving. The study concluded with the observation that the working-class schools prepared students to assume positions as laborers, whereas the professional and executive schools prepared students to be intellectuals and professionals. Thus the school, in a major way, reproduced the larger system of unequal social relations. One of the few advantages that students in the working-class schools could have was learning the 'methods of resistance.' 'These methods are highly similar to the "slowdown," subtle sabotage, and other modes of indirect resistance carried out by adult workers in the shop …' (Anyon 1983, p. 163).

Several other research studies on the influence of social class in learning confirm Anyon's conclusions above. Although some studies recognize that there are variations in community and school expectations of what students should learn, there is evidence that working-class communities placed their priority 'on acquiring credentials and following rules, whereas the higher-SES communities … placed emphasis on students' gaining particular kinds of knowledge and skills' (Knapp and Woolverton 1995, p. 553). The research clearly indicates that in the working-class school students are taught through the hidden curriculum 'that little is expected of them … they are not highly valued; they cannot be trusted to guide their own learning or conduct themselves responsibly in the class, and they must accept low-level positions in a rigidly controlled hierarchy' (Knapp and Woolverton 1995, p. 553). The higher-status school, on the other hand, encourages students to develop a greater sense of self-worth, higher expectations for school success, and a wider range of possibilities for fitting into hierarchies of control at different levels' (Knapp and Woolverton 1995, p. 563).

3. *Cultural Minorities and the Hidden Curriculum*

Students who are members of ethnic, racial, or linguistic minorities are usually adversely affected by the hidden curriculum of the schools. The grading system, the standardized tests, the reward and punishment measures, the textbooks, the school rules, all influenced by the hidden curriculum, normally have a negative effect on minorities. Available research clearly indicates that these and other educational instruments discriminate against minorities. For example, about 45 percent of Hispanics in the United States are expected to drop out from high school (Sleeter E et al. 1998). 'In-school factors such as poor academic performance and disliking school account for 56 percent of dropouts among Black, male students' (Massialas 1996, p. 69). School suspensions and expulsions, strong instruments of the hidden curriculum, also impact adversely racial, ethnic, and linguistic minorities. There is usually a disproportionate number of students from minority groups being punished through these measures. Language minority students are also adversely affected by the hidden curriculum in the classroom. For instance, one study on the topic indicated that in a classroom which had a mixture of mainstream and limited English proficient students, 'the teachers interacted less frequently with the ESL students than their native-English counterparts … the interactions with ESL students tended to be managerial in nature rather than instructional' (McKeon 1994, pp. 25–6). Also reported in the literature is the negative result that ensues from the disability of teachers to understand the flow of discourse between them and students who have different patterns of linguistic interactions. '… when the patterns of language use or the patterns of social interactions that children bring to the school are ignored or denigrated by their teachers, the ensuing relationship may threaten the students' academic development' (Pease-Alvarez and Vasquez 1994, p. 93).

4. *Gender and the Hidden Curriculum*

The hidden curriculum has a very strong influence in school learning when mediated through gender. Gender-appropriate roles are conveyed through the formal curriculum as well as the hidden. Upon reviewing the works on the subject, Banks (1995, p. 622) stated, 'Research indicates that sex-role attitudes and gender associations develop early, and that teaching materials, the mass media, and society at large often reinforce sex-role stereotyping ….'

5. *Condition of Disability and the Hidden Curriculum*

Students with learning disabilities, e.g., students with visual or hearing impairment, emotional disturbance are also affected by the hidden curriculum. Whether placed in classes with mainstream students (inclusion classes) or in special education classes there is always a built-in bias in their relations with teachers, other

students and in the school-related expectations which influence in a significant way what is to be learned, how, and for what purposes. There is an over-representation of cultural minorities in the special education programs that clearly point to the influence of the hidden curriculum in student selection and placement.

6. Conclusion and Outlook

Recently there have been many attempts to attend to the negative influences on learning of the hidden curriculum. Programs for teachers seek to make them aware of these often-negative influences and to take appropriate measures to alleviate them from the school settings. Instructional approaches such as 'cooperative learning,' 'peer tutoring,' and portfolio assessment are all attempts to combat the negative effects of the hidden curriculum on student learning.

See also: Education and Income Distribution; Educational Institutions and Society; Educational Systems: North America; Gender and School Learning: Mathematics and Science; Group Processes in the Classroom; Minorities; Obedience: Social Psychological Perspectives; School as a Social System; Self-concepts: Educational Aspects

Bibliography

Anyon J 1983 Social class and the hidden curriculum of work. In: Giroux H, Purpel D (eds.) *The Hidden Curriculum and Moral Education*. McCutchan, Berkeley, CA, pp. 146–67

Banks J A 1995 Multicultural education: Its effects on students' racial and gender role attitudes. In: Banks J A, Banks C A Mc (eds.) *Handbook of Research on Multicultural Education*. Macmillan, New York, pp. 617–27

Jackson P 1983 The daily grind … In: Giroux H, Purpel D (eds.) *The Hidden Curriculum and Moral Education*. McCutchan, Berkeley, CA, pp. 28–60

Knapp M S, Woolverton S 1995 Social class and schooling. In: Banks J A, Banks C A Mc (eds.) *Handbook of Research on Multicultural Education*. Macmillan, New York, pp. 548–69

Lebasz J 1999 Learning the lessons of living. Sun-Sentinel **August 31**; 1E, 7E

Massialas B G 1996 The hidden curriculum and social studies. In: Massialas B G, Allen R F (eds.) *Critical Issues in Teaching Social Studies*, K-12. Wadsworth, Belmont, CA, pp. 61–79

Massialas B G, Jarrar S A 1991 *Arab Education in Transition*. Garland, New York

McCarthy S J 1998 Why Johnny can't disobey. In: Shapiro H S, Harden S B (eds.) *The Institution of Education*. McCutchan, Berkeley, CA, pp. 28–60

McKeon D 1994 Language, culture, and schooling. In: Genesee F (ed.) *Educating Second Language Children*. Cambridge University Press, Cambridge, UK, pp. 15–32

Pease-Alvarez C, Vasquez O 1994 Language socialization in ethnic minority communities. In: Genesee F (ed.) *Educating Second Language Children*. Cambridge University Press, Cambridge, UK, pp. 82–102

B. G. Massialas

Hierarchical Models: Random and Fixed Effects

1. The Basic Statistical Model and its Probability Interpretation

The model that has dominated statistical practice recently supposes that data x have, for each value ψ of a parameter, a known probability distribution dependent on the unknown ψ. The distribution may be described by the probability density

$$p(x|\psi) \tag{1}$$

where p is a generic symbol for a density of the quantity, here x, before the bar; the symbol after the bar describing the conditions applying to the distribution, here the supposed value ψ of the parameter. Statistical inference then consists in making some statement about the parameter on the basis of the data; such as a point estimate for ψ, or a test that ψ assumes specified values. An important, special case has x a vector $(x_1, x_2,\ldots, x_n)$, of n components, and ψ chosen such that, given ψ, the components are independent and identically distributed (i.i.d), when Eqn. (1) assumes the special form

$$\prod_{i=1}^{i=n} p(x_i|\psi) \tag{2}$$

The Bayesian approach would include, in addition to the density (1), a probability distribution for ψ expressed through a density $p(\psi)$. Statistical inference then consists of calculating $p(\psi|x)$ by Bayes's theorem, inverting the roles of x and ψ in (1).

A distinction is often made between the nature of the probabilities involved in Eqn.(1), and more particularly Eqn. (2), and those in $p(\psi)$. In view of the repetitions under similar circumstances of $x_1, x_2,\ldots, x_n$, their probability is associated with frequency and the x_i thought of as successive realizations of a single random variable. Alternatively, $(x_1, x_2,\ldots, x_n)$ is viewed as a random sample from a population. In contrast, ψ is a fixed, but unknown, quantity and, when introduced, $p(\psi)$ expresses beliefs about the values it might assume, without reference to frequency or populations. The disparity between random, for x, and fixed, for ψ, has played an

important role in the development of hierarchical models, leading to different inferences depending on whether the effect is supposed fixed or random. In the Bayesian approach there is no distinction and all probabilities express beliefs; the special case Eqn. (2) reflecting the belief in exchangeability among the observations x_i.

2. The Hierarchical Model

To illustrate the need for a hierarchical model, consider data consisting of repeated observations divided into groups. Let x_{ij} denote the jth observation in group i ($i = 1, 2,\ldots, m$; $j = 1, 2,\ldots, n$), with n observations in each of m groups, and suppose that within each group the values are conditionally i.i.d with distributions that may vary from group to group. A popular, special case has

$$x_{ij} \sim N(\theta_i, \sigma^2) \tag{3}$$

with the x_{ij} being independent given $\psi = (\theta_1, \theta_2,\ldots, \theta_n, \sigma^2)$, so that they are normally distributed with common variance σ^2 but with means that depend on the group. This model is the basis for one-way analysis of variance where inference may well concentrate on whether the means are truly different and, if they are, of what these differences consist. As it stands, this model is just a special case of Eqn. (1) with x the set of x_{ij} and ψ as above; but now contrast two cases that differ in respect of the nature of the groups. In the first, the study is one investigating the potency of a drug and the groups correspond to dilution levels. In the second, the milk yields of cows are measured and the groups correspond to different sires. In both cases, it may be reasonable to think of the duplicate measurements within a group as i.i.d, forming a random sample from a possible population, in agreement with Eqn. (3). But there are differences between the nature of the groups; for in the dilution experiment, the levels are fixed, whereas, with the sires, those selected may be thought of as a sample from a population of sires. In the language of the last paragraph, the θ_i are fixed in the dilution experiment, but random with the cows. In the latter, it might therefore be reasonable to suppose that the means are capable of a frequency description and a probability distribution. A possible, simple case is to suppose that, in addition to Eqn. (3),

$$\theta_i \sim N(\mu, \tau^2) \tag{4}$$

with the θ_i being independent given μ and τ^2. The frequency interpretation is inappropriate in the dilution experiment, where the levels have been organized, for example at levels 1, $\frac{1}{2}$, $\frac{1}{4}$ etc. It is also possible that inferences in the two experiments could be directed to different ends. Thus the variability among sires expressed through τ^2 is meaningful, but not in the dilution experiment. There are many other possibilities besides Eqn. (4) for the second-stage distribution. For example, one group may be a control and the remainder correspond to new treatments. In this case, the distribution for the control mean may differ from that for the others. Outside the Bayesian view, only the latter may have a frequency interpretation. It is this flexibility that helps make hierarchical modeling so attractive.

The probability judgements in Eqns. (3) and (4) together form a hierarchical model with two stages to the hierarchy, the first expressing the dependence of the data on the group means in Eqn. (3), the second describing the variability of those means in Eqn. (4). The hierarchical model has been thought by some to be appropriate in studies of milk yield, but not in dilution studies, because the means are random in the former but fixed in the latter. The general form of a two-stage, hierarchical model has data x dependent on parameters (θ_1, ϕ_1) expressed through

$$p(x|\theta_1, \phi_1) \tag{5}$$

at the first stage, with θ_1 having a distribution

$$p(\theta_1|\theta_2, \phi_2) \tag{6}$$

at the second stage. The parameters of the second stage, θ_2, ϕ_2 are often called hyperparameters. Notice that in this formulation some of the parameters at the first stage are fixed, ϕ_1, whilst others, θ_1, are random and have their probability structure expressed in the second stage in terms of the fixed hyperparameters. In the normal example, Eqns. (3) and (4), θ_1 here corresponds to the group means there, and ϕ_1 to the common variance. The hyperparameters are $\theta_2 = \mu$, $\phi_2 = \tau^2$, the values σ^2 and τ^2 describing the within- and between-groups variances. There may not be any fixed parameters at the first stage, as when the within-variances are supposed different, at σ_i^2, and given a distribution at the second stage, as with the means. It is possible to extend the number of stages in a hierarchical model, for example by adding a third stage to Eqns. (5) and (6) by supposing θ_2 depends on (θ_3, ϕ_3) expressed through

$$p(\theta_2|\theta_3, \phi_3) \tag{7}$$

Such an extension would be appropriate for a two-way analysis of variance; the first stage describing the variation within cells of the two-way layout, the second concerning the row and column means (the main

effects), the third involving the interactions between the main effects. Any Bayesian model is, in a sense, hierarchical, the two stages being $p(x|\psi)$ and $p(\psi)$ respectively, but the terminology is usually confined to the models just described. A further extension is to include regression effects. For example, Eqn. (5) may be extended to

$$p(x|z, \theta_1, \phi_1)$$

where z is an observed regressor variable. Similar regression effects may be introduced at other stages in the hierarchy.

Returning to the two-stage model in Eqns. (5) and (6), notice that x and θ_1 are random, whereas ϕ_1, θ_2 and ϕ_2 are fixed, and the stages describe the joint distribution of the random quantities, given the fixed ones, θ_2 and ϕ_2 having no effect on x except through θ_1. This lack of effect can be expressed in the form of a directed graph, or influence diagram, which clarifies the hierarchical form. The two-stage model is represented in the figure, the lack of links from x to θ_2 and ϕ_2 expressing the feature. The extension to more stages is easily apparent.

As with the basic model in Eqn. (1), the Bayesian approach would add a probability distribution $p(\phi_1, \theta_2, \phi_2)$ for the fixed parameters and hyperparameters. A joint distribution for all the quantities, random and fixed, is not thereby specified unless the additional assumption is made that, given (θ_1, ϕ_1), x is independent of (θ_2, ϕ_2). This assumption is usually made, often tacitly. The independence assumption is also reflected in the directed graph by the lack of links already mentioned. There may be additional links unless the quantities $(\theta_1, \theta_2, \phi_2)$ are independent. Since the probability structure is complete, it follows that every quantity has its probability structure defined, under any conditions expressed in terms of the quantities of the problem, which probability can be found following the usual rules of the probability calculus. In particular, the posterior distribution of any of the parameters or hyperparameters, given the data, may be determined.

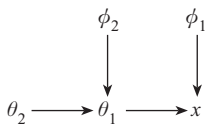

3. *Inference in Hierarchical Models*

We now return to the simple case where the observations are divided into m groups, each of size n, and possess the normal, independent distributions described by Eqn. (3). Suppose that the common variance within groups, σ^2, is known. The extension to unknown variance presents complications that do not affect the general thrust of what follows. If the means $\{\theta_i\}$ are fixed, then (3) completes the specification of the model and inference about them is ordinarily performed either by least-squares, or by maximum likelihood, the two methods here being equivalent, with the result that θ_i is estimated by the group mean

$$x_{i\cdot} = \sum_{j=1}^{n} x_{ij}/n \tag{8}$$

with variance σ^2/n. Suppose, in contrast, the means are random and this randomness is expressed by the second-stage assumption (4), where again τ^2, the between-group variance, is supposed known. Applying either of the inferential methods here gives another estimate for θ_i, namely

$$\left(\frac{nx_{i\cdot}}{\sigma^2}+\frac{x_{\cdot\cdot}}{\tau^2}\right)\Big/\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right) \tag{9}$$

with variance $(n/\sigma^2+1/\tau^2)^{-1}$, in the form of a weighted average of the group mean $x_{i\cdot}$, which was the estimate in the fixed case, and the overall mean $x_{\cdot\cdot} = \sum x_{i\cdot}/m$, the weights being inversely proportional to the variances, σ^2/n and τ^2 respectively. It is often said that the effect of randomness in the group means is to shrink the fixed estimates towards the overall mean. An important feature of the shrinkage is that the apparently best (worst) groups with highest (lowest) values of $x_{i\cdot}$ will have their estimates decreased (increased) and, as a result, the group differences will be diminished. Notice that if τ^2 is large, the shrunk means approach the usual ones. The shrinkage feature is important in comparisons of the θ_i, for example in the construction of league tables (Goldstein and Spiegelhalter 1996), or in census returns for different communities (Ericksen et al. 1989).

In the Bayesian paradigm all the unknown quantities will have a probability distribution, in particular $(\theta_1, \theta_2,\ldots, \theta_n, \sigma^2)$ in the normal, one-way case in Eqn. (3), irrespective of whether the means are fixed or random. One possibility is for the θ_i to be provided with the second-stage distribution (4) and finally (μ, σ^2, τ^2) given a distribution. If, as above, the variances are known, only the distribution of μ needs additionally to be specified. If it too is supposed normal and the variance taken to be large in comparison with σ^2 and τ^2, Eqn. (9) is the mean of θ_i, given the data (and σ^2, τ^2). Thus the Bayesian approach favors the shrunk estimate whenever the second-stage distribution is thought to be of the form in Eqn. (4), which might be appropriate for sires but not with the dilution experiment. For a Bayesian, there is no real distinction between the fixed and random effects; what differences

there are depends on the specific form of the second-stage distribution, and perhaps on the quantities of interest. For example, interest may centre on τ^2, the variance between means, rather than on the individual values θ_i. The situation is in marked contrast to other views where, as we have seen, the estimates of the means depend on whether or not the means are fixed or random; also the sample variance-components in the analysis of variance have different interpretations, in the fixed case being a measure of the differences between the specific means, in the random case referring to the population variance τ^2.

Returning to the general, statistical model in Eqn. (1) with data x and parameters ψ, it is often hard, especially when the data are numerous, to specify the required density $p(x|\psi)$. In general, whenever there are a number of quantities $u_1, u_2, \ldots, u_s$, it is difficult to describe a joint density $p(u_1, u_2, \ldots, u_s)$; an easier course often being to think about the quantities in some order and structure a series of conditional distributions $p(u_1)$, $p(u_2|u_1)$, $p(u_3|u_1, u_2)$ etc. This is essentially a hierarchical model except that, as explained earlier, it is usual to have the distribution of one quantity dependent only on the immediately previous one and not on all the previous ones; $p(u_3|u_2)$ without reference to u_1. In general, the order in which the quantities are introduced may be arbitrary, but in the hierarchical case this does not arise since the order is naturally dictated by the situation. For these reasons, hierarchical models have proved very popular in the description of complicated cases, enabling them to be broken down into stages, each of which is simple, and yet where they combine together to form a whole that reasonably describes complicated reality.

4. Types of Hierarchical Models

There are few hierarchical models in which it is possible to obtain explicit, analytical results for the inference desired. An exception is the normal, linear model of which Eqns. (3) and (4) provide a special case. Generally, at the first stage

$$x \sim N(A_1\theta_1, C_1) \tag{10}$$

where x $(n \times 1)$ and θ_1 $(s \times 1)$ are vectors, and A_1 $(n \times s)$ and C_1 $(n \times n)$ are matrices. A_1 is known, θ_1 is unknown. Usually C_1 is known, the simplest case, or known up to an unknown multiplier, though it is possible to describe the distribution of its elements in a second stage. Equation (10) expresses the usual linear dependence of x_1 on θ_1 and is basic to least squares, where θ_1 is estimated by $(A_1^{\mathrm{T}}C_1^{-1}A_1)^{-1}A_1^{\mathrm{T}}C_1^{-1}x$. The second stage imposes a similar structure on θ_1,

$$\theta_1 \sim N(A_2\theta_2, C_2)$$

with obvious extensions to further stages as required. Explicit results are provided by Lindley and Smith (1972) using Bayesian ideas, though many of the results can be interpreted sensibly without invoking a distribution on the final hyperparameters, as was seen in the special case.

Similar ideas had previously been used, starting in the 1960s, by Henderson in genetics and specifically in animal breeding; see Henderson (1984). These were outside the usual statistical paradigms and remained unappreciated in broader contexts. An admirable statistical survey of Henderson's and related work is provided by Robinson (1991).

Another field in which explicit results have been obtained, though usually of an asymptotic form, is that of Empirical Bayes. This is not always considered a hierarchical model but does fit into the general framework above. We illustrate with a simple case again involving several groups and data $\{x_{ij}\}$ but this time with independent, Poisson distributions of means μ_i, the same for all observations in group i. This describes the first stage of the hierarchy. At the second stage, the μ_i are supposed i.i.d with some unknown density $p(\mu)$. The model might be suitable when the groups refer to runs in an industrial process. A distinction from the model above is that hyperparameters are avoided and a nonparametric approach adopted at the second stage; no restrictions, other than of a mild nature like continuity, being placed on $p(\mu)$. The observed distribution of the group means $x_{i.}$ provides information about $p(\mu)$. Despite the nomenclature, the analysis is ordinarily not Bayesian. A good account is provided by Maritz and Lwin (1989). The term has also been used in connection with the equation-based approach mentioned later in Sect. 5.

Hierarchical models have also been successfully applied to stochastic processes. These may be spatial but here an exposition in the more usual temporal terms will be used. It is often reasonable to suppose that there is an underlying development process taking place over time, but that one's observation of it is disturbed by error, these two features providing the two stages of the hierarchy. A simple example is to suppose an observation x_t at time t depends on the state θ_t of the process then, with a superimposed error ε_t,

$$x_t = \theta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, v_t), \tag{11}$$

whereas the process is linear, autoregressive with

$$\theta_t = a_{t-1}\theta_{t-1} + \eta_t, \quad \eta_t \sim N(0, w_t), \tag{12}$$

$t = 1, 2, \ldots$. Equation (11) is termed the observational equation, (12) the system equation; the two sets of error terms being supposed individually independent and independent of each other. The approach is

usually Bayesian with $\{v_t\}$, $\{w_t\}$ and $\{a_{t-1}\}$ supposed known, at least in the simpler cases, and an initial distribution supposed for θ_0. If this last is normal, it is straightforward to update by Bayes to obtain successively the distributions of θ_t given $x^{(t)} = (x_1, x_2, \ldots, x_t)$, the observations up to time t. Often the primary purpose of the study is to forecast the future state of the process, θ_{t+1}, given the current data $x^{(t)}$,

$$p(\theta_{t+1}|x^{(t)}) = \int p(\theta_{t+1}|\theta_t)\, p(\theta_t|x^{(t)})\, \mathrm{d}\theta_t$$

involving the system equation and the posterior distribution of θ_t. A full account is provided by West and Harrison (1989).

As previously mentioned, hierarchical models are especially useful in describing complicated, but realistic, situations, where the individual stages in the hierarchy are relatively easy to assess. Nevertheless, apart from the cases outlined here, analytical results are difficult to obtain because the integrals that arise in the probability calculations are not expressible in standard forms, so that resort has to be made to numerical methods. In one sense this is no hardship, since the data are already numeric and in any application one only requires answers for that specific data set. On the other hand, the integrals are typically in high dimensions where numerical results are hard to come by. An important, recent advance has been to use Markov-chain, Monte Carlo (MCMC) methods which themselves depend, like the hierarchical models, on conditional distributions. There has therefore been a recent explosion of papers using MCMC to handle these models. A summary is given by Brooks (1998).

5. *Equation-based Approaches*

Hierarchical models are often described in an alternative terminology as multilevel, our stages being replaced by levels. An example is provided by students, at the first level, being in classes at the second, which form schools at the third. It is also common to alter the emphasis and not talk in terms of probability, as here, but in terms of equations which incorporate a probabilistic element, so that Eqn. (3) would be written

$$x_{ij} = \theta_i + \varepsilon_{ij} \tag{13}$$

with ε_{ij} being an error term, of zero mean. Complete adherence to Eqn. (3) would in addition require the error terms to be independent, normal and of constant variance. It is usual to employ matrix notation, as in Eqn. (10), and a popular model is

$$y = X\beta + Zu + \varepsilon \tag{14}$$

Here y, replacing x, is the observed data, dependent on observed covariates X and Z; whilst β is a set of unknown, fixed parameters, u a set of random quantities and ε the error term. Thus ε is at the first level and u at the second level. A further complication, that X is only observed with error, is sometimes included, so that

$$X = x + \eta$$

in terms of the unknown, true value x and error η at another level. This generalizes the problem of regression with both quantities subject to error. We shall refer to this whole approach as equation based, rather than probability oriented, though it is possible to pass from one to the other, as between Eqns. (3) and (13). These ideas have been used in longitudinal studies where several observations are made on an individual over time, forming one level of variation, another level being that between individuals.

The distinction between the approaches is not confined to the model description but includes the analysis, where manipulations of equations replace the probability calculus. Thus Eqns. (14) leads to estimates $\hat{\beta}$, $\hat{u}$ β, u given by

$$X^{\mathrm{T}}X\hat{\beta} + X^{\mathrm{T}}Z\hat{u} = X^{\mathrm{T}}y$$

and

$$Z^{\mathrm{T}}X\hat{\beta} + (Z^{\mathrm{T}}Z + I)\hat{u} = Z^{\mathrm{T}}y$$

under simple assumptions about u and ε. Some of the methods that are used do not employ the full probability structure but only refer to the first and second moments, a popular variant of least squares being quasi-likelihood. Other methods are based more closely on likelihood and variants thereof and often embrace ideas associated with the general, linear model. Lee and Nelder (1996) provide a good account with a lively discussion in which R. L. Smith suggests disagreement between their results and those produced by the fully Bayesian approach. This is not surprising since, as has been shown in Sect. 3, Bayesian procedures employ more shrinkage than some other methods.

An issue of the *Journal of the Royal Statistical Society A* (**162**, 1999, 269–448) is devoted to articles on 'Applications of random effects/multilevel models to categorical data in social science and medicine' and provides many examples of the ideas here briefly described. The book by Goldstein (1995) is relevant; he and his colleagues have provided a computer package (1998) to handle many models. Similar ideas have been put forward by Bryk and Raudenbush (1992). Best et al. (1996) present a Bayesian approach,

whilst Spiegelhalter et al. (1998) supply some computer facilities.

Bibliography

Best N G, Spiegelhalter D J, Thomas A, Brayne C E G 1996 Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society* A **159**: 323–42

Brooks S P 1998 Markov chain Monte Carlo method and its application. *The Statistician* **47**: 69–100

Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park, CA

Ericksen E P, Kadane J B, Tukey J W 1989 Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association* **84**: 927–44

Goldstein H 1995 *Multilevel Statistical Models*. Arnold, London

Goldstein H, Rasbash J, Plewis I, Draper D, Browne W, Yang M, Woodhouse G, Healy M 1998 *A User's Guide to MLwiN*. Institute of Education, London

Goldstein H, Spiegelhalter D J 1996 League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society* A **159**: 385–443

Henderson C R 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph, Ontario, Canada

Lee Y, Nelder J A 1996 Hierarchical generalized linear models. *Journal of the Royal Statistical Society* B **58**: 619–78

Lindley D V, Smith A F M 1972 Bayes estimates for the linear model. *Journal of the Royal Statistical Society* B **34**: 1–41

Maritz J S, Lwin T 1989 *Empirical Bayes Methods*. Chapman and Hall, London

Robinson G K 1991 That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**: 15–51

Spiegelhalter D J, Thomas A, Best N G 1998 *WinBUGS User Manual, Version 1.1.1*. Medical Research Council Biostatistics Unit, Cambridge, UK

West M, Harrison J 1989 *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York

D. V. Lindley

Hierarchies and Markets

Whether the first market entailed the exchange of nuts for berries on the edge of the forest or the first hierarchy involved leadership for the hunt of a wild boar, it is clear that markets and hierarchies go back to antiquity. It is also clear that organizational innovations in both markets and hierarchies have contributed significantly to economic progress.

As between the two, markets have the better reputation among economists. That goes back to Adam Smith (1922) and continues today. Thus, whereas Smith celebrated autonomy, specialization, and the marvels of the market, he had grave concerns over the separation of ownership and control in large joint stock companies and the excesses of managerial discretion that accrued thereto. Indeed, but for grudging acceptance of economies of scale due to technology, hierarchy has been under a cloud since. Only recently has a symmetrical approach to organizations of all kinds—markets, hierarchies, bureaus—begun to emerge.

1. Foundations

The neoclassical theory of the firm that had taken shape by the 1930s described the firm in technological terms—as a production function—to which a profit maximization purpose was ascribed. The adverse consequences of monopoly, monopsony, and monopolistic competition were all readily displayed with this abstract apparatus, but issues of organization—hierarchy—were scanted by this black box construction.

That, however, began to change with the appearance of radical new ideas about organization by economists (Coase 1937, Commons 1932), lawyers (Llewellyn 1931), and organization theorists (Barnard 1938). Coase advanced the proposition that firms were not pre-existing technological entities but that firms and markets were alternative modes of organization. Whether the firm should make-or-buy was, thus, not technologically determined (although the condition of technological separability between stages was relevant) but needed to be derived. Coase further averred that the choice between market and hierarchy was decided principally by transaction cost differences. The canonical problem, then and since, for the theory of markets and hierarchies, was that of vertical integration.

In principle, this choice was easy: holding technology constant, the firm should produce to its own needs when the transaction costs in the firm were less than in the market, and it should buy in the market when this difference was reversed. Awaiting, however, a predictive theory of when transaction costs went one way rather than another, that remained a tautological construction. Thus, although after-the-fact choices could always be rationalized in transaction cost terms, that was undisciplined and such ad hoc explanations were understandably met with grave skepticism. The twin ideas of treating firms and markets as alternatives and of appealing to transaction costs to explain the allocation of activity between them nevertheless turned out to be very fruitful.

Coase went on, moreover, to pose two deep puzzles. If the market is a marvel, why do we see so much activity organized in firms? And if the firm enjoys management advantages over markets, why is not everything organized in one large firm? Any would-be

theory of firms and markets needed to come to terms with both of these issues.

Commons' approach to economic organization was different than but also complementary to that of Coase. Commons was intensely aware of the tensions that beset organization and averred that 'The ultimate unit of activity…must contain in itself the three principles of conflict, mutuality, and order. This unit is a transaction' (1932). Not only does the triple to which Commons referred—conflict, mutuality, order—prefigure the idea of governance, in that governance is a means by which to achieve order, thereby to mitigate conflict and realize mutual gains, but viewing the transaction as the unit of analysis turns out to be very instructive.

Llewellyn took exception with the legalistic way in which both lawyers and economists viewed contract. Rather than interpret contract in the prevailing legal rules tradition, Llewellyn advanced the more elastic concept of contract as framework—according to which a contract between two parties 'almost never accurately indicates real working relations, but… affords a rough indication around which such relations vary, an occasional guide in cases of doubt, and a norm of ultimate appeal when the relations cease in fact to work' (1931). Llewellyn's ideas prefigure the concepts of incomplete contracting (as against the complete contracting tradition in both law and economics) and of private ordering (as against court ordering).

Barnard brought his experience as a businessman and his instincts as a social scientist to the study of complex organization. Interestingly, both the economist Hayek and the organization theorist Barnard were in agreement that adaptation was the central problem of economic organization. But whereas the adaptations to which Hayek (1945) referred were autonomous adaptations in the market, the adaptations to which Barnard referred entailed 'the kind of cooperation among men that is conscious, deliberate, purposeful' (1938). These two work out of different mechanisms. Autonomous adaptations of a spontaneous kind are made in response to changes in relative prices signaled by the market; co-operative adaptations of an intentional kind are implemented through administrative processes within firms. Adaptations of both kinds being crucial to a high performance economy, there is a need to treat hierarchies and markets in a combined way.

This collection of deep insights—from economics, law, and organization theory—provided an interdisciplinary foundation for a science of organization. Aside, however, from work at Carnegie (by Simon (1947, 1957) and by Cyert and March (1963)), much of which was inspired by Barnard, such a program was slow to take shape. Instead, many of the prescient insights from the 1930s remained fallow for the better part of the next 35 years. The pressing but unmet needs were (a) to combine and (b) to operationalize these profound insights. Law, economics, and organization theory needed to inform one another, thereafter to be joined in a predictive theory of economic organization.

2. *Operationalization*

Work of a combined kind got under way in the 1970s (Williamson 1971, 1975, 1979; Alchian and Demsetz 1972; Jensen and Meckling 1976; Klein et al. 1978) and has grown exponentially since. Six key moves in the progressive operationalization of the markets and hierarchies research program are sketched here.

2.1 *Human Actors*

According to Simon (1985), 'Nothing is more fundamental in setting our research agenda and informing our research methods than our view of the nature of the human beings whose behavior we are studying.' Both the cognitive and self-interestedness features of human actors are crucial to the economics of organization. Bounded rationality, according to which behavior is intendedly rational but only limitedly so, gradually has supplanted hyper-rationality as the basic cognitive condition. Opportunism, whereby economic actors are sometimes given to strategic behavior, has come into common use to describe self-interestedness. Bounds on rationality notwithstanding, economic actors are also presumed to possess foresight, which is crucial to the Commons triple. Taken together, these three attributes lead to the following: (a) all complex contracts are unavoidably incomplete (bounded rationality), (b) contract as mere promise unsupported by credible commitments poses hazards (opportunism), and (c) parties who are alert to prospective hazards will work out the ramifications and factor these into the organizational design (foresight).

2.2 *Transactions*

If the transaction is the basic unit of analysis, then the key dimensions with respect to which transactions differ need to be named and the ramifications worked out. Among the main attributes of transactions are: (a) the frequency with which transactions recur, (b) the uncertainty (disturbances to which transactions are subject), and (c) the condition of asset specificity. This last attribute has reference to the degree to which the assets that are used in support of a transaction can be redeployed to alternative uses and users without loss of productive value. Asset specificity takes several forms, of which physical asset specificity (specialized dies), human asset specificity (firm-specific learning), site specificity (locational), dedicated assets, brand name capital, and temporal specificity are leading types. A condition of bilateral dependency develops

when significant investments in transaction-specific assets are made. Intertemporal governance complications arise as a consequence.

2.3 Organization

Alternative modes of organization are described not in marginal but in discrete structural terms, which is to say that each mode of organization—market, hybrid, firm, bureau—is described as a syndrome of internally consistent attributes. Key dimensions for describing alternative modes of governance include (a) incentive intensity, (b) administrative controls, and (c) the contract law regime (Williamson 1991). (The market mode—with strong incentives, limited controls, and legalistic dispute settlement—favors autonomous adaptation; the firm mode—with weak incentives, extensive controls, and internal dispute resolution—favors co-operative adaptation.) The upshot is that alternative modes have distinctive strengths and weaknesses. They differ from one another in kind rather than in degree.

Whereas standard economic theory describes the firm as 'a point or at any rate a black box,' the economics of organization maintains that 'firms are palpably not points. They have internal structure. This structure arises for a reason' (Arrow 1999). The markets and hierarchies research paradigm describes the firm as a complex hierarchy in which both governance and specialization purposes are served. Cognitive and behavioral variance among the human actors in the population presents opportunities for specialization; and transactions for which co-operative adaptations are needed (namely, those for which assets are highly specific and need to be adapted to successive disturbances, given that contracts are incomplete) are ones for which added governance supports are needed. The concept of firm-as-governance structure—as a means by which to induce order, relieve conflict, and realize mutual gains for transactions that are in need of co-operative adaptations—thus comes to the fore.

2.4 Refutable Implications

Most of the refutable implications pertaining to governance are derived from the discriminating alignment hypothesis, according to which transactions, which differ in their attributes, are aligned with governance structures, which differ in their cost and competence, so as to effect a (mainly) transaction cost economizing result. Vertical integration in intermediate product markets is the canonical case, in relation to which a large number of variations on a theme have been worked out—for labor markets, capital markets, vertical market restrictions (such as franchising), regulation (and deregulation), economic reform (especially privatization), and public bureaus. More generally, any problem that arises as or can be posed as a contracting problem can be examined to advantage in transaction cost economizing terms.

A second class of predictions arise in conjunction with specialization. Upon making provision for variance among human actors (in cognitive and self-interestedness respects), many opportunities to realize economies by matching the attributes of human agents to the needs of a task present themselves. This second class of predictions has been less thoroughly worked out but constitutes a promising research opportunity.

2.5 Empirical

Some theories of economic organization make little effort to advance refutable implications. Among those that do, few are empirically tested. Simon (1991) evidently believes that transaction cost economics is remiss in empirical respects: awaiting empirical testing, the new institutional economics and related approaches are acts of faith, or perhaps of piety.

Coase had registered similar concerns about the dearth of empirical work on contract and organization 20 years earlier, but that was before the operationalization of transaction cost economics had begun and predicted alignments were advanced. Empirical applications of transaction cost economics got under way in the 1980s and have grown exponentially since. It could have been otherwise, but the theory and evidence display remarkable congruity (Masten 1995). As Joskow (1991) puts it, 'this empirical work is in much better shape than much of the empirical work in industrial organization generally.'

2.6 Full Formalization

The 1930s was the decade when the informal ideas for a new science of organization took shape, the 1970s was the decade when these ideas were worked up in a preformal way, the 1980s was the decade of semiformal analysis, and the 1990s witnessed a burgeoning of work of a fully formal kind. The paper by Grossman and Hart (1986) on *The Costs and Benefits of Ownership* was the first fully formal model of an incomplete contracting kind. Hart's book on *Firms, Contracts, and Financial Structure* (1995) elaborates upon and develops these ideas. Additional work of this kind has been published since and more is in progress. As matters stand presently, however, there is a huge gap between the semiformal and fully-formal stages.

A leading reason for the break is that the fully formal analysis of incomplete contracts relies entirely on property rights assignments to assess comparative efficiency. Not only are the discrete structural differences in organization among alternative modes of governance denied, but the proposition that adaptation is the central problem of organization (of which

two types are distinguished) is ignored. These differences are consequential, whence Kreps (1999) concludes that 'If Markets and Hierarchies has been translated into game theory using notions of information economics, it is a very poor translation ... In particular, mathematical-based theory still lacks the language needed to capture the essential ideas of bounded rationality, which are central to ... transaction costs and contractual form.' Kreps (1999), nevertheless, remains sanguine that formal efforts will succeed eventually, possibly with the benefit of new mathematical apparatus.

3. Conclusions

Whereas once it was customary to talk about markets or hierarchies, where the former denoted decentralization and the latter central planning, that has changed. Now it is more customary to talk about markets and hierarchies. That entails much more than a replacement of the conjunction 'or' by 'and.' It represents a veritable sea change in the way economic organization is studied. The combined study of law, economics, and organization that is herein described (a) involves comparative institutional analysis in which economizing is the main case and the action resides in the attributes of human actors, transactions, and governance structures, (b) generates numerous refutable implications, (c) invites empirical analysis, (d) has many public policy ramifications, and (e) invites progressive formalization. Many economists are now persuaded that the study of economic organization needs to be studied in a combined way in which law, economics, and organization are joined. Legal scholars have incorporated many of these ideas into contract law, corporate governance, antitrust enforcement, and regulation. And organization theorists have come to appreciate the pervasive reach of economizing reasoning.

Real headway notwithstanding, new problems and challenges appear to be unending. Productive tension between transaction cost economics and each of the parts on which it stands—law, economics, and organization—can be projected safely into the future.

See also: Administration in Organizations; Llewellyn, Karl Nickerson (1893–1962); Organization: Overview; Organizational Control; Organizations: Authority and Power; Organizations, Sociology of; Transaction Costs and Property Rights

Bibliography

Alchian A, Demsetz H 1972 Production, information costs, and economic organization. *American Economic Review* **62** (December): 777–95

Arrow K 1999 Forward. In: Carroll G, Teece D (eds.) *Firms, Markets and Hierarchies*. Oxford University Press, New York, pp. vii–viii

Barnard C 1938 (1982) *The Functions of the Executive*, 15th edn. Harvard University Press, Cambridge, MA

Coase R 1937 The nature of the firm. *Economica* **4**: 386–405. (Reprinted. In: Williamson O E, Winter S (eds.) 1991 *The Nature of the Firm: Origins, Evolution, Development*. Oxford University Press, New York, pp. 18–33)

Commons J R 1932 The problem of correlating law, economics, and ethics. *Wisconsin Law Review* **8**: 3–26

Cyert R M, March J G 1963 *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ

Grossman S J, Hart O D 1986 The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* **94** (August): 691–719

Hart O 1995 *Firms, Contracts, and Financial Structure*. Oxford University Press, New York

Hayek F 1945 The use of knowledge in society. *American Economic Review* **35** (September): 519–30

Jensen M, Meckling W 1976 Theory of the firm: Managerial behavior, agency costs, and capital structure. *Journal of Financial Economics* **3** (October): 305–60

Joskow P 1991 The role of transaction cost economics in antitrust and public utility regulatory policies. *Journal of Law, Economics, and Organization* **7** (Special Issue): 53–83

Klein B, Crawford R A, Alchian A A 1978 Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics* **21** (October): 297–326

Kreps D M 1999 Markets and hierarchies and (mathematical) economic theory. In: Carroll G, Teece D (eds.) *Firms, Markets, and Hierarchies*. Oxford University Press, New York, pp. 121–55

Llewellyn K N 1931 What price contract? An essay in perspective. *Yale Law Journal* **40**: 704–51

Masten S 1995 Introduction to Vol. II. In: Williamson O, Masten S (eds.) *Transaction Cost Economics*. Edward Elgar, Brookfield, VT

Simon H 1947 *Administrative Behavior*. Macmillan, New York

Simon H 1957 *Models of Man*. Wiley, New York

Simon H 1985 Human nature in politics: The dialogue of psychology with political science. *American Political Science Review* **79**: 293–304

Simon H A 1991 Organizations and markets. *Journal of Economic Perspectives* **5** (Spring): 25–44

Smith A 1776 [1922] *The Wealth of Nations*. Dent, London

Williamson O E 1971 The vertical integration of production: Market failure considerations. *American Economic Review* **61** (May): 112–23

Williamson O E 1975 *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York

Williamson O E 1979 Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics* **22** (October): 233–61

Williamson O E 1991 Comparative economic organization: The analysis of discrete structural alternatives. *Administrative Science Quarterly* **36** (June): 269–96

O. E. Williamson

High Performance Computing

Many scientific and engineering applications require a higher computing power than standard computers can provide. So in the last three decades of the twentieth century specialized computer architectures for high

performance computing (HPC) have been built. They are mainly optimized to execute floating-point operations at very high speed. The fastest HPC systems are also called supercomputers (the term supercomputing is sometimes used as a synonym for HPC). High performance is obtained by vector computing, parallel computing, use of special fast technologies or by a combination of these techniques. This article reviews the history of HPC systems, explains the most important high-performance computer architectures, and presents some typical HPC applications.

1. History of HPC

Figure 1 demonstrates that the performance of HPC systems based on the LINPACK benchmark (Dongarra 1990) was raised from about 0.1 GFlop s^{-1} (billions of floating-point operations per second) in the 1970s to more than 1000 GFlop s^{-1} in the last decade of the twentieth century; so every decade the performance was increased by two orders of magnitude.

In the second half of the 1970s the introduction of *vector computers* marked the beginning of the HPC era. They offered a performance advantage of an order of magnitude compared to standard computer systems of that time, reducing the cycle time by introducing *pipelining* in the processor architecture and supporting high-level operations on complete vectors (e.g., component-wise addition or multiplication) by hardware. Typically, several functional units (FU), each dedicated to a particular arithmetic operation, are available in a vector processor (see Fig. 2(a)) and two or even more result vectors can be computed simultaneously. To achieve a high performance, the pipelines of all FUs must be filled and emptied again at very high speed. Thus, input/output data and intermediate results are read from and written to fast vector registers each holding a fixed number (mostly 64 to 256) of vector elements. Separate load/store units perform an independent and overlapped data transfer from/to memory.

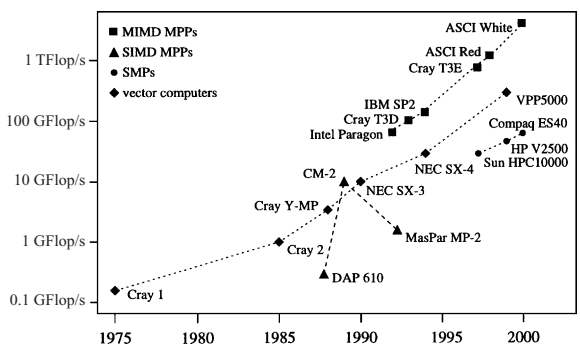

Figure 1
Performance of various high-performance computers from 1975 to 2000 for the solution of linear equation systems, measured in billions or trillions of floating-point operations per second (GFlop s^{-1} or TFlop s^{-1}, respectively)

The first vector computer, Cray 1, was introduced in 1976. It is composed of six FUs and one load/store unit and is realized in fast emitter coupled logic (ECL) technology operating at 80 MHz. In later vector processor architectures, such as Cray 2 (1985), Cray Y-MP (1988), NEC SX-3 (1990), and NEC SX-4 (1994), performance was further increased by using up to 16 FUs and up to eight load/store units. Also the clock cycle time was decreased (e.g., by using liquid-cooled processors). During the 1980s, the number of vector processors installed in vector computers grew from one to eight and later to 16. Although they could communicate via a shared memory, it was difficult to use several processors on a single job, so mainly the throughput was increased by running several jobs in parallel.

As an alternative to vector computers, in the 1970s massively parallel processing (MPP) systems appeared as a new class of high-performance computers. Here a large number (typically ranging from 256 to 65,536) of processors, interconnected by a static network, operate in parallel to compute one task. The memory is distributed among the processors. In the first MPP systems the SIMD (single instruction multiple data) operation principle was used. The instructions are broadcast from a control unit (CU) to very simple processors (called processing elements or PEs) where they are executed simultaneously in lock-step mode on different local data items (see Fig. 2(b)). PEs are based on simple arithmetic units operating on words of limited length (often only of 1 bit). Operations on integers or floating-point numbers are performed by software routines that take several clock cycles depending on the length of the operands. Thus the power of such machines does not result from the PE itself but from the enormous number of installed PEs. So SIMD-type MPPs perform excellently on applications based on large matrices of short data elements, such as signal or image processing.

One of the first commercially available SIMD-type MPP systems was the distributed array processor (DAP) with up to 4096 PEs arranged in a 64×64 grid. The most famous machine of this class is the Connection Machine CM-1 (1985) and CM-2 (1987) with up to 64k simple one-bit processors connected by a hypercube network (Hillis 1989). For certain applications the CM-2 offered a higher performance than the CRAY Y-MP (see Fig. 1). MasPar MP-1 (with 4-bit PEs, 1990) and MP-2 (with 8-bit PEs, 1992) represent the last important machines of this architecture class, which disappeared from the market in the first half of the 1990s.

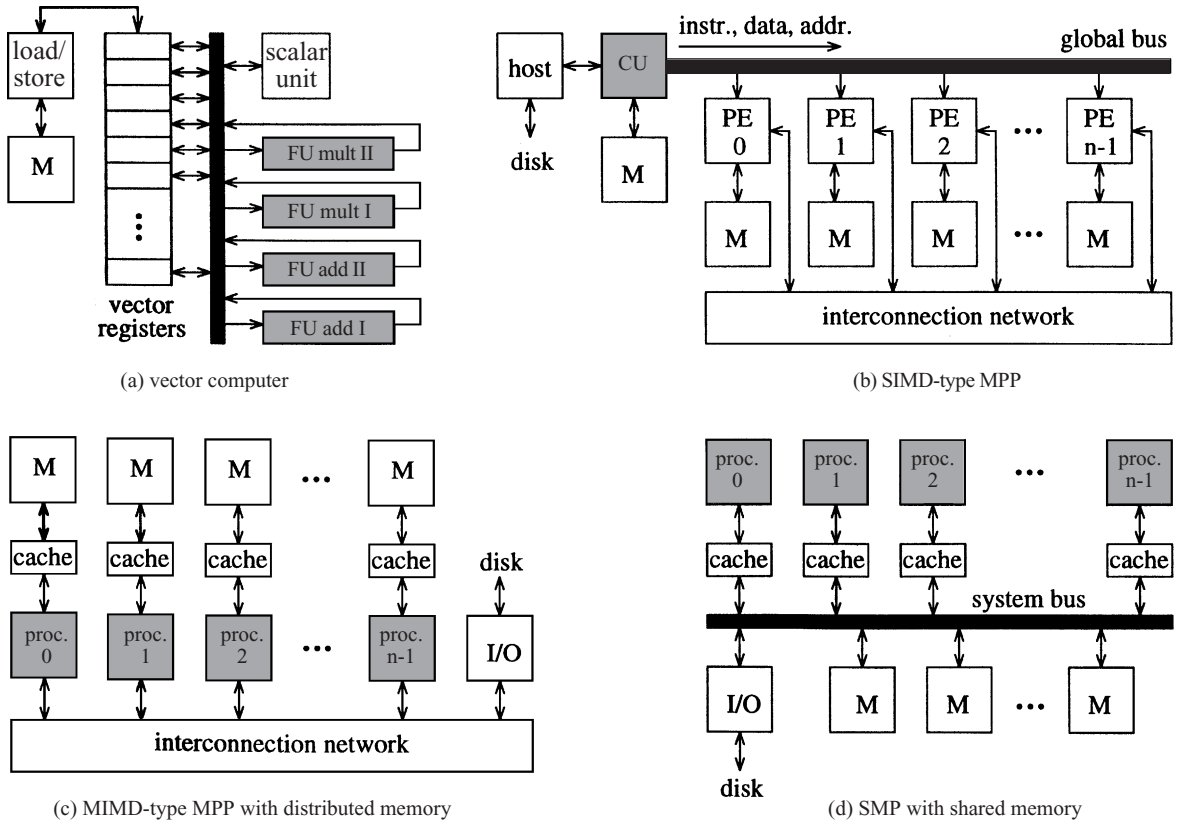

Figure 2
Architectures of high-performance computers

At the same time a new kind of HPC systems was proposed: MPPs based on advanced microprocessors operating according to the MIMD (multiple instruction multiple data) principle. In contrast to SIMD architectures the processors execute different tasks asynchronously. A few hundred or thousand cheap off-the-shelf processors designed for high clock-frequencies are interconnected by a network of high bandwidth. However, the communication is much slower than in SIMD-type MPP systems because there is a high overhead for initializing each communication and for synchronizing the processors.

In the first MIMD-type MPP systems, such as Intel's iPSC series with up to 128 processors or the nCube/2 system (1990) with up to 8,096 processors, the d-dimensional hypercube was the preferred topology. Here each processor can be reached in at most d steps; also two- and higher-dimensional grids can be embedded easily. Later systems were provided with additional communication controllers in each node and used new routing techniques so that the communication time became less dependent on the distance between nodes. Topology was no longer important (if contention is ignored) and cheaper topologies such as a grid or a torus were favored. The Intel Paragon (1992) consists of up to 4,000 Intel i860 processors connected by a two-dimensional (2D) grid. The Cray T3D (1993) and T3E (1997) systems are based on a three-dimensional (3D) torus with up to 2,048 Alpha processors. In the second half of the 1990s, sometimes several workstations or PCs connected by an Ethernet or fiber Channel network were used instead of many integrated processors. Conceptually, this network of workstations (NOW) is not different from a distributed-memory MPP system, but the communication latency is higher.

The advantages of MIMD-type MPP systems are a good price/performance ratio and a good scalability of the architecture. On the one hand cheap and small systems with a few processors and a medium level of performance can be built easily; on the other hand the high performance of expensive vector computers can be reached or even exceeded at lower cost by using a sufficiently large number of processors. Due to the

large communication overhead, MIMD-type MPP systems are well suited for applications with a coarse-grain parallelism.

Since about 1995 several companies started to sell symmetric multiprocessor (SMP) models of their workstation families. In the first systems a rather small number (typically 4 to 16) of microprocessors access a shared memory built from several memory modules by using a central system bus (see Fig. 2(d)). Here each variable can be read and updated by each processor, which is supplied with a local cache to reduce the memory access time. The main problem in SMPs is to ensure that each variable remains consistent in all caches, especially when it is updated by one processor. This *cache coherency problem* is solved via a special protocol. Instead of a single bus often a set of several buses or even a crossbar network are used to reduce the risk of bus contention and to increase the collective bandwidth between memory and processors.

At the end of the twentieth century SMPs with a larger number of processors and a hierarchical interconnection structure became available. Here clusters of tightly-coupled processors are interconnected by a less costly and also slower network. Typical examples are Sun's HPC series (1997) and Hewlett-Packard'sV-class series (1998). SMPs offer a high performance at low price. However, the scalability is limited: only a medium number of processors can be installed. So the performance levels of MPPs and vector computers cannot be reached (compare Fig. 1).

2. Programming HPC systems

On vector computers any existing sequential code can be used although it should be optimized manually to achieve high performance. For MPPs, SMPs and NOWs, however, the program must be parallelized based on either the data-parallel programming model or the message-passing model.

Data-parallel programs express all operations on vectors and matrices from a global point of view. The compiler automatically maps the data structures onto the parallel architecture. No manual synchronization or communication between the processors must be programmed, so the design of such data-parallel programs is rather simple. Examples of data-parallel languages are the several dialects of C offered by the manufactures of SIMD systems and High Performance Fortran (HPF).

In parallel programs based on the message-passing model the user must distribute tasks and data structures among the available processors. The user must also program each synchronization and communication explicitly by sending messages between the parallel tasks running on different processors. Each task can be implemented in a standard programming language such as C, C++ or Java. For the exchange of messages between tasks several standardized libraries such as PVM (parallel virtual machine) or MPI (message passing interface) are available.

3. HPC Applications

In the beginning of the HPC era most applications involved the solution of large systems of linear or (partial) differential equations. Later other computation-intensive applications have emerged, such as the simulation of chemical and nuclear physical models, image processing, virtual reality, and optimization. Also artificial intelligence applications can be accelerated by parallelism—for instance, in 1997, IBM's MIMD-type MPP system Deep Blue consisting of 32 processors and 512 chess co-processors defeated world chess champion Kasparov. Deep Blue used a parallel alpha–beta tree search algorithm, tuned with expert knowledge, and could generate and evaluate 200 million chess moves per second (Hamilton and Garber 1997).

The simulation of artificial neural networks (ANN) represents another typical HPC application. Many ANN models (such as multi-layer perceptrons or radial-basis function networks) have been proposed that can be applied to many classification, approximation and optimization tasks without a formal treatment of the underlying problem. The training of ANNs by examples requires a high computational power, and many researchers have successfully simulated such ANN training on various parallel computer architectures (for an overview, see Šerbedžija (1996)). ANNs involve an inherent massive parallelism that can be exploited for a parallel implementation on different levels: either the *neurons* (basic ANN operating elements), the *synapses* (weighted connections between neurons), the *layers* (in case of a multi-layer network), or the *patterns* of the training set can be mapped onto different processors. However, especially at the first two levels, the communication/computation ratio is rather high because only simple operations are performed in each synapse or neuron between two communication steps. So, in particular, SIMD-type MPPs are well suited for a parallel implementation on all levels because of their good communication capabilities. High performance on MIMD-type MPPs, SMPs and NOWs can only be achieved by a pattern-parallel implementation; however, this technique can be applied to just a few neural learning algorithms. So for ANN simulation the paradoxical situation exists that older (and no longer available) SIMD-type systems are more suitable than modern HPC computer architectures.

The importance of high-performance computers is also steadily increasing in computational neuroscience. Complex biology-oriented neural models can be simulated on HPC systems in acceptable periods of time. Examples are: highly detailed single-cell models (such as a cerebellar Purkinje cell with 4550 compart-

ments (De Schutter and Bower 1994), medium-sized networks of neurons modeled at a less detailed level (such as hippocampal circuits with up to 128 pyramidal cells described by 70 compartments (Traub et al. 1996) or very large networks of neurons modeled at a higher abstraction level (such as the cat's primary visual system with 16384 integrate-and-fire neurons (Niebur and Brettle 1994). In all cases discrete-time approximations to the solutions of a large set of nonlinear differential equations must be found.

Furthermore, HPC systems can accelerate several computation-intensive tasks in psychology such as simulating complex models of cognitive processes, simulating realistic models of visual perception, or analyzing electromagnetic activities in the brain. Especially the evaluation of psychological models and the estimation of model parameters yielding a best fit to data requires high computing power (for recent such applications, see Edwards (1997)).

4. *Perspective*

In the last three decades of the twentieth century, high performance computing was marked by a strong dynamic with a continuous appearing and disappearing of manufacturers, systems, and architectures. This rapid change will probably continue in the near future. Due to technological progress faster microprocessors will become available each year and cheap HPC systems with a few processors will offer a performance that could formerly be achieved only by expensive vector computers or MPP systems. Thus, high-performance computers will no longer be specialized architectures but be the high end of a continuum of fast desktop systems and servers. Scalable architectures can easily be adapted to the performance requirements of most computation-intensive HPC applications. However, for algorithms with high interprocessor communication demands (such as the simulation of ANN), it is not obvious if future HPC systems will be able to provide a sufficiently low communication latency to allow efficient parallel implementation.

A few special applications (such as weather prediction and computational fluid dynamics) require a still higher computing power than near-term HPC systems can provide. In 1996, the US Department of Energy's Accelerated Strategic Computer Initiative (ASCI) launched a program for a three-fold increase in computing performance every 18 months over a 10-year period using commercial hardware only. The first systems, ASCI Red (1997) and ASCI White (2000), consist of more than 7000 microprocessors and exceed $1\,\mathrm{TFlop\,s^{-1}}$ (see Fig. 1).

The scope of HPC applications will be further extended in the future. New applications will emerge, not only in science and engineering, but also in commerce (in areas such as information retrieval, decision support, financial analysis, or data mining). *Data-intensive* algorithms are used here; they must operate on very large databases containing massive amounts of information (e.g., about customers or stock prices) that have been collected over many years. In particular, the rapid future increase in the amount of data stored worldwide will make HPC indispensable.

See also: Artificial and Natural Computation; Information Processing Architectures: Fundamental Issues; Information Technology; Science and Technology, Social Study of: Computers and Information Technology

Bibliography

De Schutter E, Bower J M 1994 An active membrane model of the cerebellar Purkinje cell. *Journal of Neurophysiology* **17**: 375–419

Dongarra J J 1990 The LINPACK benchmark: an explanation. In: Van der Steen A J (ed.) *Evaluating Supercomputers. Strategies for Exploiting, Evaluating and Benchmarking Computers with Advanced Architectures*. Chapman and Hall, London, pp. 1–21

Edwards L (ed.) 1997 Special issue on high performance computer applications in the behavioral sciences. *Behaviour Research Methods, Instruments, and Computers* **29**: 11–124

Fox G C, Williams R D, Messina P C 1994 *Parallel Computing Works*. Morgan Kaufmann, Los Altos, CA

Hamilton S, Garber L 1997 Deep Blue's hardware–software synergy. *Computer* **30**: 29–35

Hillis W D 1989 *The Connection Machine*. MIT Press, Cambridge, MA

Leighton F T 1992 *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes*. Morgan Kaufmann, San Mateo, CA

Niebur E, Brettle D 1994 Efficient simulation of biological neural networks on massively parallel supercomputers with hypercube architecture. *Advances in Neural Information Processing Systems* **6**: 904–10

Šerbedžija N B 1996 Simulating artificial neural networks on parallel architectures. *Computer* **29**: 56–63

Traub R D, Whittington M A, Stanford I M, Jefferys J G R 1996 A mechanism for generation of long-range synchronous fast oscillations in the cortex. *Nature* **383**: 621–4

A. Strey

High School Dropout

Dropping out of high school, that is, leaving school without graduating, is a central educational issue in almost all industrialized countries, although large differences between educational systems make cross-national or cross-cultural comparisons problematic. Nevertheless, there is broad international agreement that dropping out increases the risk of subsequent criminal behavior and drug abuse, lower occupational

and economic prospects, lower lifetime earnings, and becoming a member of the underclass. Furthermore, dropping out increases the probability of lower levels of mental and physical health. Studies investigating the consequences of high school dropout at a macro-level have shown that higher dropout rates result in lower tax revenues and increased expenditure for governmental assistance programs. Against the backdrop of all these negative consequences, sociological, educational, and psychological researchers have investigated the reasons for dropping out. The reasons identified can be classified as school-related, family-based, or personal. Researchers agree that the risk of dropping out begins to develop early in a student's life and that intervention programs to prevent dropout should therefore be initiated at an early age.

1. *Problems in Calculating the Dropout Rate*

A major challenge facing any study of high school dropout is how to define and identify the dropout population. There is some agreement, among American researchers at least, that students who were enrolled in grades 10, 11, or 12, but left school without graduating are to be defined as dropouts. However, this definition is often not viable in other countries. In international comparison, school systems not only differ with respect to the organization of primary and secondary schools, forms of ability grouping, and curricula, they also differ in terms of the age at which students are allowed to leave school. While students in the UK and most states of the USA are legally obliged to attend school until the age of 16, for example, students in other countries (e.g., Germany) are not allowed to leave school until they are 18. As a consequence of the different lengths of compulsory schooling, the probability that a student will drop out between the ages of 16 and 18 varies from country to country. Furthermore, there is some evidence that high schools tend to overestimate the percentage of dropout (e.g., Rumberger et al. 1990), and it has been argued that 'dropout data are probably the least reliable information available today regarding the reality of schools' (Frymier 1996, p. 4).

Beyond these problems of definition, it is characteristic of almost all industrialized countries that school participation rates increased continuously during the twentieth century. In the USA, for example, the proportion of 14- to 17-year-olds enrolled in high school increased from 11 percent in 1900 to 94 percent in 1978. However, dropout rates have been relatively stable since the beginning of the 1970s (between 5 and 10 percent in industrialized countries).

2. *Factors Influencing High School Dropout*

Three major directions of research into the antecedents of high school dropout can be distinguished, although there is substantial theoretical overlap between the three approaches. While the first two types of research concentrate on family background and school variables from the sociological perspective, the third—more psychological—approach focuses on individual characteristics in terms of achievement and motivation.

2.1 *Family Background and High School Dropout*

Where family background is concerned, the determinants found to be of importance for academic achievement have also been considered with respect to dropout (e.g., Rumberger 1983, Rumberger and Thomas 2000). Parents' level of education represents a potential predictor of dropout, that is, better educated parents spend more time with their children, thus supporting their accomplishments and increasing the likelihood that they will remain in school. Furthermore, these parents influence the academic performance of their children by imparting the values, aspirations, and motivation needed to succeed and remain in school. With regard to parental income, children from poor families probably feel more pressure to contribute to the family income, and thus leave school to seek work at an earlier stage. Family structures are also important, in that children from broken homes where one or both parents are absent are less likely to receive the support and encouragement needed to stay in school. The absence of parents, however, is often combined with financial problems that might explain the dropout. In a more detailed analysis of family influences, Rumberger et al. (1990) investigated the interaction styles between parents and children and the consequences for dropout. Dropouts were more likely to live in households characterized by a permissive or an authoritarian parenting style, their parents were more likely to use extrinsic punishments as a reaction to poor grades, and reacted with more negative emotions to both good and bad school grades. Furthermore, dropout students reported that their parents were less involved in their education and gave them less help with their homework.

In the USA, in particular, relations between ethnic group membership and high school dropout have been investigated, and higher percentages of dropout have been reported for Hispanic and African-American students. In a study investigating primary reasons for dropout, Rumberger (1983) found out that African-American females most often cited marriage and pregnancy as reasons for leaving school. However, if SES and individual achievement are held constant, the relationship between ethnic background and dropout diminishes or reverses, thus suggesting that the association is mainly mediated by family background and individual levels of achievement (e.g., McNeal 1995).

2.2 School Characteristics and High School Dropout

School characteristics have been investigated in research on school effectiveness. Authors agree that not only academic achievement, but dropping out of school is a reliable indicator of school effectiveness. Rumberger and Thomas (2000) identify school resources, structural characteristics of schools, and school processes as factors, which can influence high school dropout. With regard to the financial resources of schools, research shows that the student-teacher ratio is positively related to high school dropout. Furthermore, there is some evidence that a higher percentage of teachers with advanced degrees increase the likelihood that students will remain in high school.

Structural characteristics such as school location (urban, suburban, rural), size, and type (state vs. private) also seem to influence dropout, even when important individual and family characteristics are controlled. Research by Pittman and Haughtwout (1987) revealed a positive correlation between high school size and dropout rate. Mediation analyses demonstrated that this effect was mediated primarily by the social climate (sense of cohesion, level of student participation in school activities, interaction with faculty, the magnitude of certain problems at the school), and indicated that smaller schools had a better social climate and consequently lower dropout rates. Private schools usually have lower dropout rates than state schools (e.g., Rumberger and Thomas 2000), and students attending urban schools are more likely to drop out than their counterparts in suburban schools (e.g., Frymier 1996).

Research on school processes has identified a number of school-related variables that influence dropout rates. A higher level of cooperation between teachers and students seems to reduce dropout rates and, with regard to the academic climate, Bryk and Thum (1989) found that dropout rates were lower in schools in which students took more advanced courses and reported doing more homework. Bryk and Thum also reported that high schools with serious staffing problems had higher dropout rates.

Rumberger and Thomas (2000) suggested that because schools are anxious to show that their students' test scores are improving, they actively reduce the membership of low-achieving students through a process of discharge. This process affects not only low-achieving students, but also hard-to-educate students. Since both variables are confounded with SES and ethnicity, these school processes usually lead to higher dropout rates for students with low SES, and for Hispanic and African-American students.

2.3 Individual Determinants of High School Dropout

Although many background and school variables are important antecedents of dropout, a more psychological view of the problem suggests that the concrete decision of whether or not to carry on or leave school is guided by individual characteristics such as attitudes and motivational states. Schools, parents, and teachers provide an academic environment, and students' perceptions of this environment determine their achievement motivation, their behavioral intentions, and their academic behavior and choices. In one of the few psychological studies on reasons for dropout, Vallerand et al. (1997) proposed a model assuming self-determined academic motivation to be an important factor influencing behavioral intentions and dropout behavior. Their study suggests that students who perceive their parents, teachers, and school administrations as supportive develop a higher level of perceived competence and academic autonomy, and that this leads to higher intrinsic academic motivation and a lower likelihood of leaving school without graduating.

Expectancy-value models (e.g., Wigfield and Eccles 2000) suggest that academic choices are based mainly on expectancies of success in the domain of choice, and on how students value the domain. Both expectancies and values are based on family background, school and class variables, and previous individual achievement-related experiences. Low expectancies and values with respect to the future school career increase the probability of dropping out.

3. Prevention of High School Dropout

Low levels of academic achievement (e.g., Caliste 1984) and high absenteeism (e.g., Bryk and Thum 1989) are the strongest behavioral predictors of high school dropout. Several authors (e.g., Simner and Barnes 1991) have pointed out that the earlier a student with low achievement levels and high absenteeism can be identified, the more likely is it that dropout prevention will be successful. An effective 12-week prevention program proposed by Caliste (1984) first identifies at-risk students (grade-point averages below 1.0 on a 4.0 scale). These students receive extra tuition from their teachers and participate in group counseling sessions. Topics discussed during the weekly counseling sessions include motivation, relevance of school, academic problems, career goals, and study habits. In the tutoring sessions, students are primarily assisted with current coursework and given some remedial instruction.

Participation in out-of-school activities, particularly sports clubs, can also prevent dropout. McNeal (1995) argues that higher social integration in both school and in non-academic clubs influences whether or not students remain in school. His analyses show that for low-achieving students (one standard deviation below the population mean) not involved in sports clubs, the probability of dropout is $p = 0.25$, while for members of sports clubs the probability is only $p = 0.17$.

Finally, authors agree that children from disadvantaged family backgrounds need particular social and financial support. Females, especially African-Americans, often leave school without graduating because of marriage and pregnancy. Day-care programs are viable measures to prevent drop out here.

See also: Adolescent Behavior: Demographic; Adolescent Development, Theories of; Adolescent Vulnerability and Psychological Interventions; Career Development, Psychology of; Childhood and Adolescence: Developmental Assets; Families as Educational Settings; Family Health; Gangs, Sociology of; Internet: Psychological Perspectives; Media Effects; Media Effects on Children; Motivation, Learning, and Instruction; Popular Culture; School Achievement: Cognitive and Motivational Determinants; School as a Social System; School Learning for Transfer; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values; Schooling: Impact on Cognitive and Motivational Development; Urban Poverty in Neighborhoods; Violence and Media; Youth Culture, Sociology of; Youth Gangs

Bibliography

Bryk A S, Thum Y M 1989 The effects of high school organization on dropping out: An exploratory investigation. *American Educational Research Journal* **26**: 353–83

Caliste E R 1984 The effect of a twelve week dropout intervention program. *Adolescence* **19**: 649–57

Frymier J 1996 Determining dropout rates in large city school districts: Problems and accomplishments. *The High School Journal* **80**(1): 1–10

McNeal R B Jr 1995 Extracurricular activities and high school dropouts. *Sociology of Education* **68**: 62–81

Pittman B P, Haughtwout P 1987 Influence of high school size on dropout rate. *Educational Evaluation and Policy Analysis* **9**: 337–43

Rumberger R W 1983 Dropping out of high school: The influence of race, sex, and family background. *American Educational Research Journal* **20**: 199–220

Rumberger R W, Ghatak R, Poulos G, Ritter P L, Dornbusch S M 1990 Family influences on dropout behavior in one California high school. *Sociology of Education* **63**: 283–99

Rumberger R W, Thomas S L 2000 The distribution of dropout and turnover rates among urban and suburban high schools. *Sociology of Education* **73**: 39–67

Simner M L, Barnes M J 1991 Relationships between first-grade marks and the high school dropout problem. *The Journal of School Psychology* **29**: 331–5

Vallerand R J, Fortier M S, Guay F 1997 Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology* **72**: 1161–76

Wigfield A, Eccles J S 2000 Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology* **25**: 68–81

O. Köller

Higher Education

In most countries, teachers of the most prestigious institutions of higher education are assigned both teaching and research tasks, whereby 'academic freedom' is held in high esteem. A trend towards systematization notwithstanding, curricular approaches have remained varied across countries, fields and institutions, regarding the academic or professional emphasis, their adaptive or critical role to the world of work, the degree of specialization and the role higher education plays for personality development beyond the cognitive domain. Recent reforms react to trends of internationalization and globalization, increase the role of practical experiences, lifelong education, distance education and use of new technologies.

Universities are often called surprisingly persistent institutions, though their organization is often weak and anarchic. The continental European tradition of a weak university leadership amid a major role of government and academic freedom of scholars gave way in the 1980s to reduced governmental bureaucratic control and increased formal autonomy of the central management of higher education institutions amid growing pressures for efficiency, accountability, and social relevance.

1. Definition

'Higher education' became a popular term in the second half of the twentieth century depicting the intellectually most demanding stage of pre-career education. Terms such as 'post-secondary,' 'tertiary,' or 'third-level' education underscore stages of learning: upon completion of primary and secondary education, i.e., after 10–14 years of schooling, varying according to national systems, students might enroll in the third stage of education, as a rule prior to embarking regular employment. The term 'higher,' however, suggests a specific quality, expecting students to learn questioning prevailing rules and tools and to understand theories, methods, and substance of 'academic' knowledge. Higher education also often serves a twofold function: teaching and 'research,' i.e. the creation and preservation of systematic knowledge.

2. History

Institutions and programs of advanced learning existed in various countries more than 1,000 years ago. However, the core elements of today's higher education, i.e., teaching and learning of 'analytic,' 'rational,' 'systematic,' 'critical,' 'skeptical,' and 'innovative' thinking, emerged from the European universities of the Middle Ages.

Around 1800, a new historical stage can be noted when substantial changes of the European universities occurred. The British model of the Oxford and Cambridge approach regained strength which puts emphasis on the trained mind and the personality development through close communication between the teachers and the students. The French university, reformed under the Napoleonic regime, combined strong state coordination of organization and curricula leading to state-coordinated degrees with a decentralized organization within the university, thus reinforcing a divide between the faculties and notably intellectual and cultural approach of the facultés des lettres and sciences on the one hand and on the other the professional emphasis of other faculties inside the universities and écoles outside the universities. The German model, mainly the 'unity of teaching and research' approach of the Humboldtian university, was based on the idea that academic freedom and institutional autonomy guaranteed by the state help to secure the university's most appropriate service to society, and also underscored the freedom of learning, i.e., the student's ability and responsibility of independent learning in communication with the professor. The German model became the internationally most influential in the nineteenth century.

The models spread through adaptations and misunderstandings which turned out to be creative in some instances. Notably, the US approach of being 'as British as possible' as far as undergraduate education is concerned and 'as German as possible' by combining teaching and research, and in establishing graduate schools as a completely new element, tends to be viewed as the most successful one in the twentieth century (Ben-David 1977).

The development of higher education after World War II is often regarded as the third stage of the history of modern higher education. Higher education was no longer expected to serve a small number of academics and socially exclusive professions, but rather became a 'mass' or an almost 'universal' phenomenon (see Trow 1974). Whereas less than 5 percent of the respective age group enrolled around 1950 in most industrial countries, the respective proportion was even more than 50 percent in some countries at the end of the twentieth century. The number of doctoral-granting institutions in the world had risen to about 10,000 and of other institutions to even a higher number, and the total number of students was more than 80 million in 1995.

3. *Quantitative and Structural Developments*

3.1 *Access and Admission*

According to OECD statistics (OECD 2000), about 80 percent of the corresponding age group on average of the OECD member states successfully completed upper secondary education in the late 1990s. Upper secondary education varied between countries structurally, i.e., 'horizontal' (graduates of all institutions and programs are eligible for entry to higher education) or 'vertical' (specific types and programs preparing for higher education), as well as according to the role vocational training plays and the way it is organized.

Modes of admission to higher education can be classified according to three types (see Teichler 1988).

(a) If secondary education is organized horizontally, the large numbers of those successfully completing secondary education are eligible to apply for admission to higher education, whereby entry to higher education is the major point of selection within the education system. The entry decisions rest with the higher education institutions (in the USA) or their departments (e.g., in Japan), and might be based on school grades, combinations of school subjects, admission tests, interviews, etc.

(b) If secondary education is vertically organized, those successfully completing the academic track (being awarded the Abitur, Matura, Baccalaureat, etc.) tend to be entitled to enroll at institutions of higher education (e.g. Germany and The Netherlands). As a rule, entry is a nondramatic point of allocation, and selective admission applies only for few fields and institutions.

(c) Finally, some countries provide for both selective secondary education and selective admission at entry to higher education (e.g. the UK).

At all modes, most of those applying for higher education will eventually enroll. The admission system in industrialized countries is not primarily a gatekeeper of the overall quantity of enrolment, but rather allocates primarily with respect to type, institutional prestige or field of study.

3.2 *Enrolment, Drop-out, and Graduation*

The 'youth' in most countries acquire entry qualification for higher education at the age of 17 to 19 years. The average age of new entrant students varies in industrial countries between 18 and 23 years.

The OECD reports an average 'net entry rate' to higher education in terms of programs leading to a (at least bachelor) degree of 40 percent in industrialized countries for 1998. In addition, almost 20 percent begin short tertiary-education studies. Altogether, women have enrolled as frequently as men, on average, in industrialized societies since about 1990. UNESCO calculated a gross enrolment rate, by comparing enrolment figures of young people within five years of the normal school-leaving age: 63 enrolled percent in 'more developed regions,' 38 percent enrolled in countries in transition and 7 percent enrolled in 'less developed regions' in the mid-1990s.

About two-thirds of those enrolling in degree programs in industrialized countries are eventually

awarded a degree. The estimated dropout rates are 10 percent in Japan (lowest) and about 65 percent in Italy (highest).

In 1998, 22 percent of the respective age group on average across industrial countries eventually was awarded a degree (in comparison, 1 percent on average were awarded a doctoral degree). In addition, 11 percent completed other tertiary education programs.

3.3 Diversity of Levels, Programs, and Institutions

Universities, i.e., multidisciplinary institutions in charge of both teaching and research, entitled to award advanced academic degrees (notably the doctorate) and, where applicable, entitled to award subsequent degrees qualifying for senior academic positions (the 'doctor scientiae' or the 'Habilitation'), are considered the key institutions of higher education. Except for France, where Grandes Écoles lead to the highest professional careers, entry to university is viewed as most prestigious. Most other institutions, i.e., mono-disciplinary institutions of technology, theology, etc., initially viewed as less prestigious, often gained the same status over the years.

Since the 1960s, two developments reinforced a diversification of higher education. Pressure towards increasing systematic knowledge in the middle-level occupations led to an 'up-grading' of the institutions serving the training of these occupations. Concurrently, the university sector came under pressure to diversify because the growing student body became increasingly heterogeneous in terms of motivations, competencies and career prospects and because resources were not made available for an expansion of research-oriented higher education in tune with increasing students' enrolment (Teichler 1988, Meek et al. 1996).

In various countries, an additional sector of higher education was established, called for example 'polytechnics' in the UK (which eventually became universities in 1992), 'instituts universitaires de technologie' in France and 'Fachhochschulen' in Germany which initially were expected to provide applied educational programs, but not to undertake research. This sector was named 'short cycle,' 'non-university' or 'alternative' without any of these terms being generally appreciated.

Countries varied substantially, however, according to the extent of which differentiation of higher education was realized through:

(a) types of higher education institutions;

(b) curricular approaches of programs (e.g. academic versus professional–vocational);

(c) levels of programs and degrees; (e.g. the OECD's 'non-university certification, 'short first university degree,' 'long first university degree,' 'second university degree,' and doctoral or similar degrees);

(d) length of programs (in terms of years of study or credits–points, etc.); and

(e) varied reputation and prestige among formally equal institutions and programs.

Higher education systems, for example, were characterized as 'unitary,' 'binary,' 'dual' or 'multi-type,' 'stage' or 'level,' 'comprehensive,' etc.

4. Knowledge, Teaching, and Learning

4.1 Learning and Pursuit of Knowledge

Students are expected to learn questioning the available wisdom and the established ways for problem solving and to be prepared for indeterminate work and other life tasks. A close link between teaching and research should ensure that teaching is at the 'cutting edge' of research development. In some countries, curricula are very open and teaching is conceived more or less as a by-product of research.

4.2 Trend Towards Systematization of Curricula

With the expansion of higher education, however, a trend could be observed towards systematization of content, structures, and processes. Study programs were divided into sequences and stages. Curricular regulations grew, and measures of informing and guiding students were extended. In many countries, teachers were expected to be trained in techniques of teaching, and regular evaluation of teaching practices often was introduced. The trend towards systematization affected initial course programs most strongly. But the US concept of graduate school for those preparing for a doctoral degree was also taken up or modified in most other countries in the latter half of the twentieth century (Clark 1995).

Yet major curricular approaches remained varied across and within countries, notably according to seven dimensions. First, curriculae may be directed more strongly towards preparation for research and the creation of knowledge or towards the reproduction of knowledge available. Second, curriculae may be geared closely to occupational preparation or not related to job roles at all. Third, objectives of higher education curricula differ according to the degree of specialization. Fourth, fields of study may be designed according to disciplines or take on a multidisciplinary character. In some countries (fifth), a common core of knowledge might be offered for all students. Sixth, views differ as regards the responsibility of higher education for the personality development of students beyond the cognitive domain. Finally, curricula vary according to the extent they aim to contribute to cultural enrichment of society.

The variety of curricular approaches had their pendant in the roles of the teachers and the students. Students might be viewed as young learners for which the university serves 'in loco parentis,' as customers or

as independent and self-responsible young members of the university community. Teachers might view themselves as scholars in the pursuit of knowledge, or as educators and counselors. Available research suggests that students have more similar views across countries on 'good' teaching and learning, and teachers on their role and function (Altbach 1997) than one would expect, taking into account the diversity of national systems of higher education.

4.3 *Expectation of Relevance and Utility Amid Uncertainty*

In the 1980s and 1990s, the conditions both for teaching and learning, as well as for research in higher education, changed substantially. First, pressure grew on teaching, learning and research to be more utilitarian than in the past. Public and private funds were moved increasingly towards research and establishments outside higher education, thus reducing the share of research expenditures spent in higher education in many countries to less than 20 percent. Thus, modes of research spread aiming to cut across disciplines, in spite of the persistence of distinct cultures of the disciplines (Becher 1989), and became more context aware and problem oriented (Gibbons et al. 1995).

Efforts to respond to the societal needs were often met by reinforcing a divide between basic and applied research or by a divide between academic and professional or vocational study programs, but these distinctions became blurred over time. While utilitarian pressures turn out to be imaginative for basic research in some cases, research aiming not to be applied often lay the foundation for practical use. Study programs not geared to certain professions are enriched by a confrontation with practical problem solving, both in classes and internships, while study aiming to be professionally useful is forced to become broader, owing to the uncertainties of future labor market conditions and work assignments—'key qualifications' is a widely used term underpinning this trend.

4.4 *Internationalization and Globalization*

Higher education could be viewed as very international in terms of cosmopolitan values, pride based on international reputation and global search for knowledge. However, the regulatory and funding context tends to be national, and curricula vary strikingly between countries. Kerr (1990) described a conflict between 'the internationalization of learning and the nationalization of the purposes of higher education.' International agencies, e.g. UNESCO, OECD or the World Bank often advocate concepts of convergence, but seem to be counterbalanced by a persistence of varied traditions and political options.

5. *Steering, Governance, and Organization*

5.1 *The Deliberately Weak and 'Fuzzy' Organization*

Universities are often called surprisingly persistent organizations and have survived for several centuries, though they seem to be weak or anarchic institutions in most countries. The links between the higher education institutions and the government or the society can be described as a peculiar mix of partial control and vague pressures, on the one hand, and safeguards of institutional 'autonomy' on the other hand. In the internal organization, accommodation prevailed in most countries between administrative supervision, a strong position of the professorate as a whole and the 'academic freedom' of the individual scholar.

The weakness was widely viewed as supportive for an ethos within the institution which allowed scholars to focus on the pursuit of knowledge without major concern about the social and organizational environment. The Humboldtian ideal was widely shared that a university protected from pressures of immediate utility was most likely to serve generation of knowledge and innovation eventually valuable for society.

5.2 *Traditional Modes of Governance and Administration*

Analyses of higher education governance often intend to describe both the relationships between higher education and society and the modes of governance within the institutions, thus underscoring that the internal mechanisms of power, cooperation, and communication are linked to the ways how higher education is steered from outside. Notably, US scholars have described traditional higher education systems according to four models:

(a) the collegial model—emphasizing nonhierarchical cooperative decision making, and a significant degree of self-determination by academic staff;

(b) the bureaucratic model—emphasizing legal-rational authority and formal hierarchies;

(c) the professional model—emphasizing the authority of experts and the importance of horizontally differentiated units linked to loose confederations;

(d) the political model—conceptualizing governance in terms of political conflict among interest groups with competing views and values' (Harman 1992, p. 1282).

Often, three actual modes of academic governance and administration are pointed out (see Ben-David 1977, Clark 1983). Both in the French and in the German traditions, governments, though largely respecting or protecting academic freedom, pursued certain visible ways of bureaucratic control of higher education institutions. The university leadership was

kept deliberately weak whereby the faculties in France and the chairs in Germany were the stronghold of academic freedom and institutional autonomy. In the British tradition, funds were allocated to the higher education institutions with little strings, the strongest power of the institution rested in academic collegiality, and the university leadership, though stronger than in continental Europe, exerted moderate control. The US model was characterized by a weak position of visible macrosocial forces, whereby the market seemed to play a stronger role than government and the trustees of the individual institutions of higher education, and by a strong power of the president and the deans along limited decision-making power of the academics.

Clark (1983) defined the patterns of control and coordination according to a triangle of market, state authority and academic oligarchy. He allocated US higher education as relatively close to the market, Germany to state authority, and Italy to academic oligarchy.

5.3 In Search of New Modes of Governance and Administration

Since the 1970, the climate around and within institutions of higher education has changed substantially. Obviously, institutions of higher education have lost much of their special character. They are expected to be more similar to other institutions, and the functions they perform are less exclusive. Higher education experienced:

(a) a loss of trust in the academic profession and in academic freedom. Counterbalances by other forces, scrutiny of job performance and of academic quality as well as direct incentives for social relevance and utility are viewed now as appropriate;

(b) a loss of confidence in governments' potentials of steering higher education successfully through detailed regulations and control of administrative processes; and

(c) a growing demand for efficiency in terms of serving extended duties of teaching and research with stagnating resources and for visible effectiveness.

Four major changes of governance and administration of higher education occurred. First, in the wake of the student protests in the late 1960s, various continental countries introduced models of participation whereby students, junior academic staff, and administrative staff got a substantial say, as was the case previously in various Latin American countries. This reform approach, however, lost momentum rapidly because the process of decision making seemed to become controversial and time consuming and because the actors benefiting often were disappointed as regards the limited real influence the formal decision-making process actually allotted to them.

Second, various governments in Europe moved from a 'control' approach, and in the UK from an 'institutional autonomy' approach, towards a 'supervision' approach (Neave and van Vught 1991). Mechanisms employed were, among others, political target setting, steering through indicator-based funding, contracts between the state and higher education institutions, and the establishment of evaluation systems allowing to identify the 'output' of higher education (Westerheijden et al. 1994).

Third, managerial actors (university leadership, deans) were strengthened and managerial mechanisms were established or extended. The internal processes of decision making and administration were expected to get close to industrial entrepreneurship, while respecting the specifics of a professional organization serving unpredictable innovation of knowledge.

Fourth, the setting of key actors increasingly grew complex. In addition to the state, the market and the professorate (Clark 1983), the management of the higher education institutions, and other internal actors (students, junior academic staff, and administrative staff) came into play as well as other 'stakeholders' (i.e., all other forces who could have a legitimate interest to be involved in higher education as a 'visible hand').

See also: Education and Income Distribution; Education and Learning: Lifespan Perspectives; Education (Higher) and Gender; Educational Institutions and Society; Educational Institutions, History of; Human Capital: Educational Aspects; Policy Knowledge: Universities; Universities and Science and Technology: Europe; Universities and Science and Technology: United States; Universities, in the History of the Social Sciences

Bibliography

Altbach P G (ed.) 1997 *The International Academic Profession*. Carnegie Foundation for the Advancement of Teaching, Princeton, NJ

Becher T 1989 *Academic Tribes and Territories*. SRHE/Open University Press, Buckingham, UK

Ben-David J 1977 *Centers of Learning: Britain, France, Germany, United States*. McGraw-Hill, New York

Clark B R 1983 *The Higher Education System: Academic Organization in Cross-National Perspective*. University of California, Berkeley, CA

Clark B R 1995 *Places of Enquiry: Research and Advanced Education in Modern Universities*. University of California Press, Berkeley, CA

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M 1995 *The New Production of Knowledge*. Sage Publications, London

Harman G 1992 Governance, administration and finance: Introduction. In: Clark B R, Neave G R (eds.) *The Encyclopedia of Higher Education*. Pergamon Press, Oxford, UK, pp. 1279–93

Kerr C 1990 The internationalization of learning and the nationalization of the purposes of higher education: Two 'laws of motion' in conflict? *European Journal of Education* **25**: 5–22

Meek V L, Goedegebure L, Kivinen O, Rinne R (eds.) 1996 *The Mockers and Mocked: Comparative Perspectives on Differ-*

entiation, Convergence and Diversity in Higher Education. IAU Press/Pergamon, Oxford, UK
Neave G, van Vught F (eds.) 1991 *Prometheus Bound: The Changing Relationships Between Government and Higher Education in Europe*. Pergamon Press, Oxford, UK
OECD 2000 *Education at a Glance*. OECD, Paris
Teichler U 1988 *Changing Patterns of the Higher Education System*. J. Kingsley Publishers, London
Trow M 1974 Problems in the transition from elite to mass higher education. In: OECD (ed.) *Policies for Higher Education*. OECD, Paris, pp. 51–101
Westerheijden D F, Brennan J, Maassen P A M (eds.) 1994 *Changing Contexts of Quality Assessment*. Uitgeverij Lemma, Utrecht, The Netherlands

U. Teichler

Hindu Law

Hindu law claims one of the longest continuous histories of any legal system in the world. Indeed, for about 2,500 years it was based on the same primary sources. Yet, the attitude toward these sources has varied considerably over the centuries, with the result that Hindu law in 2,000 CE, is very different from what it may have been in 500 BCE.

1. *The Smṛtis*

Until the 1950s the main sources of Hindu law were contained in texts written in India's classical language, Sanskrit (Derrett 1973a, 1973b). The earliest of these texts, the *dharmasūtras* (mostly in prose, and probably earlier) and *dharmaśāstras* (exclusively in verse, and probably more recent) may be dated to approximately between 500 BCE and 500 CE. The Indian concept of *dharma*, for which there is no Western equivalent, is much broader than the Western concept of 'law.' A Hindu's *dharma* encompasses each and every rule people are supposed to live by and, even though the legal aspects of *dharma* are more and more treated in isolation in the *dharmaśāstras*, 'legal' prescriptions in these texts remain inseparable from other obligations, especially those which we would label 'religious.' A case in point is adoption, which is not valid unless accompanied by a ritual act, the *dattahoma*.

To understand the Hindu view of *dharma*, and of law in particular, it is essential to keep in mind that the *dharmasūtras* and *dharmaśātras* are part of the revealed Veda. They make up the *smṛti* 'remembrance,' the authority of which is only slightly lower than that of the older Vedic texts, the *śruti* 'revelation.' They too, were considered to have been transmitted to mankind by *rishis* ('seers'). Like the Vedas themselves, the texts on *dharma* are perfect, eternal, and unchangeable; they should not be interfered with by human beings.

The most important *smṛti* text is the *Manusmṛti* or *Mānavadharmaśāstra* (ca. 200 BCE–200 CE), attributed to Manu, the progenitor of all Hindus (Bühler 1886). Chapters 8 and 9 of the *Manusmṛti* are entirely devoted to legal matters. Together with Chapter 7 they constitute the *rājadharma*; the king is indeed the supreme justice in his kingdom. For the first time, an effort is made here to classify the causes of legal disputes, under 18 titles: failure to pay debts, deposit and pledge, sale without ownership, partnerships, resumption of gifts, failure to pay wages, breach of contract, rescission of sale or purchase, disputes between the owner and keeper of cattle, boundary disputes, physical assault, verbal assault, theft, violent crimes, sex crimes, rules for husband and wife, partition and inheritance, and gambling and betting (8.4–7). The subdivision into 18 'heads of litigation' is maintained in all later *dharmaśāstras*, with the difference that their eighteenth chapter is often called *prakīrṇaka* 'miscellaneous,' to incorporate aspects of substantive law that were not dealt with by Manu. Only one *dharmaśāstra* is entirely devoted to law, the *Nāradasmṛti* (Lariviere 1989). The *Manusmṛti* deals with procedure under the first head of litigation; Nārada devotes three introductory chapters to it. Most notable, in the chapters on evidence, is the highly detailed treatment of witnesses and ordeals.

In view of the later development of Hindu law, reference must be made here to one characteristic trait of *smṛtis* that has provoked much discussion in the scholarly literature. There are, in these texts, obvious contradictions, not only between different texts, but also within single texts such as Manu's translation of Lingat's 1967 work (see Lingat 1967). In one verse (9.104) Manu prescribes that, after the death of both parents, their sons divide the paternal estate equally. Without any transition, the following verse (9.105) says that the eldest brother takes the entire paternal property, while the others shall live under his authority as they did under their father's. After a few verses in praise of the eldest son, Manu continues: 'The eldest son takes an additional share of five percent, as well as the most desirable part of the estate; the next son gets half of that, the youngest one quarter' (9.112). Or, Manu acknowledges and describes in detail the duties of partners in a levirate marriage (9.60), and elsewhere rejects levirate marriages as *paśu-dharma* 'the *dharma* of cattle' (9.66). Scholars have attempted to expurgate the preserved Manu text so as to avoid the contradictions. In reality, Manu's three ways of inheriting paternal property, and his opposite views on levirate, may result from the fact that they were all *dharma* in different places, among different groups of people, etc. They all knew exactly which rule applied in their community, but all these rules were *dharma*, and deserved a place in authoritative texts on the subject (Rocher 1993).

2. The Commentaries

From about the seventh century CE onward, and well into the eighteenth century, the *smṛtis* become the object of voluminous commentaries. Both the commentaries on individual texts and the digests which treat of particular aspects of *dharma*, drew into their discussions the entire body of ancient *smṛti* texts. Among the several, often book-length, sections of the digests most often one or two are devoted to law: some deal with procedure (*vyavahāra*) and substantive law (*vivāda*) separately; others combine the two in a single treatise. There are even digests on specific aspects of substantive law, mainly on inheritance and on adoption.

The commentaries and digests follow the basic trends of the revealed, eternally valid *smṛti* texts. Even though the digests discuss the legal aspects of *dharma* in separate volumes, 'law' remains closely linked to other aspects of *dharma*: an adoption is still considered invalid if it is not accompanied by the *dattahoma*.

On the other hand, drawing on the entire body of *smṛti* literature as equally authoritative, the commentators and digest writers are able to fill in gaps in one *smṛti* by drawing on a different *smṛti*. With two exceptions, the *smṛti* texts tell us little about what happens to the inheritance of a father when he does not have any descendant beyond a great-grandson. The commentaries and digests appeal to all *smṛti* texts to establish complete, if rarely uniform, lists of heirs, from the son and other relatives, down to the king. The commentaries and digests do provide coherent, albeit different, systems of Hindu law.

Nowhere is that tendency expressed more clearly than in the contradictions referred to earlier. For the commentators, all *smṛti* texts are equally authoritative; there cannot be any contradictions, and it is their task to show this. They either present the rules as applicable under different circumstances, or they interpret one text literally and do not hesitate to force a meaning on the others, so that all texts are in perfect harmony. In the scholarly literature the often important and far-reaching difference between the interpretations put on the *smṛti* texts by the commentators are most often interpreted as attempts on the part of the latter to adapt the former to local circumstances and customs. Or, rather than being concerned with local customs, the commentators apply the same principles as their interpretations of *smṛti* texts in any other field of *dharma*: subtle grammatical explanations borrowed from Pāṇini's *Aṣṭādhyāyī* and other grammatical treatises, intricate rules of textual exegesis laid down in Jaimini's *Mīmāṃsāsūtras*, principles of logic contained in Gautama's *Nyāyasūtras*, etc. Each individual commentator brings his vast amount of traditional, panditic learning to bear on the *smṛti* texts, to mold them into one coherent system, to interpret them more correctly than his predecessors, with whom he does not hesitate to argue in the strongest possible terms. Differently from the *smṛti* texts, which were revealed once for all through the intermediary of ancient *rishis*, the commentaries and digests were produced by human beings with whom it was possible to disagree as well as to agree.

3. Anglo-Hindu Law

In search of ways to govern Indians according to their own laws rather than by British Law, in 1772, under the impetus of the Governor of Bengal, Warren Hastings, the East India Company adopted the principle that 'in all suits regarding inheritance, marriage, caste and other religious usages or institutions, the laws of the Koran with respect to the Mohamedans and those of the Shaster with respect to the Gentoos shall invariably be adhered to.' The British administrators were familiar with Muslim law, but the law of 'the Shastar,' i.e., the *dharmaśāstras*, was totally unknown to them; in fact, at that time not a single Englishman had access to the language of the *dharmaśāstras*, Sanskrit. What followed, throughout the British period, is a series of remarkable attempts, on the one hand, to maintain the integrity of shastric law and, on the other hand, to interfere with it and abolish it whenever it came into conflict with basic principles of British justice.

In a first attempt to gain access to the sources of Hindu Law (1772), pandits were attached to the Anglo-Indian courts of law. The Court Pandit received the legal question via a translation from English or Bengali into Persian, and hence into Sanskrit. He consulted his law books, and made a decision, which was again translated, through Persian, into English. Besides being cumbersome and time-consuming, this procedure soon created uneasiness in British legal circles in India: distrust in the pandits and their decisions became a common topic in contemporary reports and correspondence. Even though the office of Court Pandits was not abolished until 1864, meanwhile some Englishmen decided that for them to get access to the sources of Hindu law, they needed to study Sanskrit. One of these early Orientalists was Sir William Jones, a judge in the Calcutta Supreme Court, who produced an English translation of the *Manusmṛti* (Jones 1794).

Even prior to Jones's translation, Hastings called to Calcutta a group of 11 pundits, to provide him with *the* code of Hindu law. While the pundits' composition, *Vivādārṇavasetu*, 'the bridge across the ocean of disputes,' was still going on, their work was translated into Persian, and from there into English by one of Hastings's protégés, Nathaniel Brassey Halhed. Halhed's book became known as *A Code of Gentoo Laws, or, Ordinations of the Pundits* (1776).

However, neither Halhed's *Code* nor Jones's *Institutes* proved to be helpful as law books in the Anglo-Indian courts of law. The former suffered much in the two successive translations (and is, in fact, very

different from the Sanskrit original); the latter displayed all the characteristics of an ancient *smṛti* text mentioned earlier: no answers to important legal questions, conflicting answers to the same questions, and so forth.

At that point the British turned away from the *smṛtis*, toward the commentaries which at least provided coherent legal systems and gave clear answers to questions raised in Court. On the other hand, each commentator's coherent legal system was, in varying degrees, different than that of any other commentator. Henry Thomas Colebrooke was first to introduce the concept of Schools of Hindu Law (Rocher 1972), a concept that was to dominate the history of Hindu law until the middle of the twentieth century. According to Colebrooke, there was one digest that was applicable in Bengal, and another commentary that was followed in the rest of India.

In 1810 he published a volume entitled *Two Treatises on the Hindu Law of Inheritance*, in which he translated *Jīmūtavāhana's* digest *Dāyabhāga* (early twelfth century) as the text for Bengal, and the chapter on inheritance from the *Mitākṣarā*, a contemporareous commentary on the *Yājñavalkyasmṛti*, as the text for India outside Bengal. In more recent years, as more commentaries and digests were translated, the 'Mitakshara School of Hindu law' was further subdivided into four subschools: the Benares, Mithila, Maharashtra or Bombay, and Dravida or Madras schools of Hindu law, each school and subschool administering Hindu law on the basis of its own Sanskrit text (or texts). The result was a pyramidal system in which the decision of any lower court could be appealed to the regional High Court, whose decisions were binding on the lower courts within its territory. The Judicial Committee of the Privy Council in London, the final court of appeal, made binding but different decisions depending on the region from where the appeal originated.

The basic difference between the two major Schools of Hindu law is that, in the *Mitākṣarā* school, any male individual born into a family, by his very birth acquires an unspecified right of ownership in joint family property. According to the *Dāyabhāga*, on the contrary, no one becomes the owner of any part of the family property before the time at which the right of the previous owner lapses, either by death or for any other reason that prevents him from exercising his obligations as the manager of the family property. This distinction inevitably had a profound impact on the lesser (*Mitākṣarā*) or greater (*Dāyabhāga*) power of the head of the family to manage the family property at his pleasure.

Notwithstanding these efforts on the part of the British to administer Hindu law to Hindus, and even though they continued to pay homage to the Sanskrit *dharmaśātras*, at the same time they steered Hindu law in directions that were alien to the ancient *rishis* and the commentators.

First, based on the principle that 'we are here to administer law, not religion,' the Anglo-Indian courts drew the line between what they considered legal actions and the ritual activities that were an integral part of them in the Sanskrit texts: in Anglo-Hindu law, adoption became valid even without the performance of the *dattahoma*.

Second, since the commentaries and digests in each region became the principal sources of law, the Courts no longer considered it necessary to look for the correct meaning of the *smṛti* texts on which the commentaries were based; they relied invariably and solely on the commentators' interpretations, thereby reversing the original scale of authority between the human commentaries and the texts transmitted in the revealed *smṛtis*.

Third, even when Indians joined the Courts, the number of judges and advocates who knew Sanskrit remained small. All arguments and judgements were based on English translations. If one of the parties demonstrated that a translation was in error, the judges rejected that argument, and referred instead to precedents contained in the massive Indian Law Reports. From being based on texts, Hindu law now became a case law, and the law reports were interspersed with Latin phrases such as *stare decisis* and *communis error facit ius*.

Finally, there were traits of Hindu law that ran counter to basic British conceptions of law and justice. In many of these cases the judges overruled Hindu law on the basis of the principle of 'justice, equity, and good conscience.' When the sisters of a murderer claimed the inheritance of his victim, which in Hindu law their brother was entitled to because he was physically and mentally capable of performing the prescribed rituals (*śrāddha*) for the deceased, the case went through several appeals with opposite results, until the Privy Council rejected outright the murderer's right to inherit.

Conversely, Hindu law was expanded by the Courts' recognition of wills, which were unknown in the traditional system. In other cases, the British passed a number of Acts which overruled and abolished some existing practices of Hindu law. As the titles of the earliest Acts (*The Caste Disabilities Removal Act, 1850; The Hindu Widow's Remarriage Act, 1856*; etc.) indicate, they aimed at improving the living conditions of outcasts (a Hindu who renounced his religion or was excluded from his caste lost any right to property and inheritance) and women (when her husband died, his widow, however young she might be as a result of the common practice of child marriage, was not allowed ever to remarry).

4. *Hindu Law in Independent India*

In 1947 Hindu law as established by the British passed unchanged into the courts of Independent India.

Implementing article 44 of the Constitution: 'The State shall endeavour to secure for the citizens a uniform Civil Code throughout the territory of India,' was delayed. Even the more modest attempt at providing all Hindus in India with a uniform 'Hindu Code,' an idea that preceded Independence, failed. However, in 1955 the Indian Parliament enacted *The Hindu Marriage Act*, which applied not only to Hindus, but also to Buddhists, Jainas, and Sikhs; in short to any Indian citizen 'who is not a Muslim, Christian, Parsi or Jew by religion' (art. 2 [1]). In 1956 *The Hindu Marriage Act* was followed by *The Hindu Succession Act*, *The Hindu Minority and Guardianship Act*, and *The Hindu Adoptions and Maintenance Act*. All four acts state, in almost identical terms, that:

> Save as otherwise expressly provided in this Act,—
> (a) any text, rule or interpretation of Hindu law or any custom or usage as part of that law in force immediately before the commencement of this Act shall cease to have effect with regard to any matter for which provision is made in this Act;
> (b) any other law in force immediately before the commencement of the Act shall cease to have effect in so far as it is inconsistent with any of the provisions contained in this Act.

The impact of these modern Acts, overriding Hindu law as it had been practiced for a century and a half, is still a matter of dispute. The point is that, officially, they made large sections of traditional Hindu law obsolete. Any Hindu (in the broad sense) can legally marry any other Hindu, monogamous marriage becomes the rule of law, and so do annulment of marriage, divorce, and remarriage for women; succession to inheritance is made uniform, male and female heirs are treated equally; and women, who were accorded a limited estate in the past, now become full owners of their property and can dispose of it as they please.

There remains, however, an undetermined residue of Hindu law for which the Acts do not make any provision; in areas such as joint family law the Courts continue to administer Hindu law as it was prior to 1955 (Derrett 1963).

See also: India, Religions of; Law: Imposition, Reception, and Colonial; Postcolonial Law

Bibliography

Bühler G 1886 The laws of Manu. In: *Sacred Books of the East*, 25. Clarendon Press, Oxford, UK

Derrett J D M 1963 *Introduction to Modern Hindu Law*. Oxford University Press, London

Derrett J D M 1973a Dharmaśāstra and juridical literature. In: Gonda J (ed.) *A History of Indian Literature*. Harrassowitz, Wiesbaden, Germany

Derrett J D M 1973b History of Indian Law (*Dharmaśāstra*). In: Spuler B, et al. (eds.) *Handbuch der Orientalistik*, II.3.1. Brill, Leiden, The Netherlands

Halhed N B 1776 *A Code of Gentoo Laws, or, Ordinations of the Pundits*. East India Company, Calcutta, India

Jolly I 1885 *Outlines of an History of the Hindu Law of Partition, Inheritance and Adoption, as Contained in the Original Sanskrit Treatise*. Thacker Spink, Calcutta, India

Jolly J 1896 *Recht und Sitte* (*einschliesslich der einheimischen Litteratur*). Trübner, Strassburg, Germany

Jones W 1794 *Institutes of Hindu Law; or, the Ordinances of Menu*. Government Press, Calcutta, India

Kane P V 1930–77 *History of Dharmaśāstra* (*Ancient and Mediaeval Religious and Civil Law*) 5 vols in 7 parts. Bhandarkar Oriental Research Institute, Poona, India

Lariviere R W 1989 The *Nāradasmṛti*. *University of Pennsylvania Studies on South Asia* **4–5**, University of Pennsylvania, PA

Lingat R 1967 *Les Sources du Droit dans le Système Traditionnel de l'Inde*. Mouton, Leiden, The Netherlands [1973 *The Classical Law of India*. trans. Manu, University of California Press, Berkeley, CA]

Rocher L 1972 Schools of Hindu law. In: Ensink J, P Gaeffke (eds.) *India Maior Congratulatory Volume Presented to J. Gonda*, pp. 167–76. Brill, Leiden, The Netherlands

Rocher L 1993 Law books in an oral culture: The Indian *Dharmaśāstras*. *Proceedings of the American Philosophical Society* **137**: 254–67

L. Rocher

Hintze, Otto (1861–1940)

The Berlin historian Otto Hintze continues to fascinate those engaged in international historical research. In Germany, there was a particularly noticeable increase of interest in Hintze in the 1960s, in conjunction with an enduring resurgence of social historical schools. The rare combination of detailed investigation through decades of sources on the one hand and the elaborated approach of history's systematic contiguous disciplines on the other brought to light results which are still of great interest in modern research.

This is especially true for the comparative approach to political structural history, particularly that of the European states from the height of the Middle Ages to the nineteenth century. From an early stage, Hintze was at pains to incorporate the categories of political science and economics, constitutional law, and sociology into historiographical analysis, thus broadening the scope of the discipline. Despite modernization theory and more recent concepts of the comparative analysis of nation building, Hintze's advocacy of the comparative approach to constitutional and governmental history and the methodological initiatives taken in his works are still of great significance. Indeed, Hintze's approach remains indispensable to the investigation of the processes of state making.

Otto Hintze was born in the small Pomeranian town of Pyritz on August 27, 1861, as the son of a middle-

ranking civil servant. After studying for two years at the University of Greifswald, Hintze transferred to the Friedrich-Wilhelms-Universität of Berlin in 1880–1. There he studied, for example, under Johann Gustav Droysen, who was still in his final years of life. Hintze completed his doctoral studies in the spring of 1884 with a dissertation on medieval history (*König Wilhelm der Holländer. Ein Beitrag zur Geschichte des deutschen Interregnums*), supervised by the historian Julius von Weizsäcker. Gustav Schmoller, the economic and political scientist who was soon to have a considerable impact on Hintze, was not yet named as one of his academic teachers. Hintze followed his first rather narrow course of study in history with a second in legal and constitutional history (sciences of state). According to the records, this was a formative stage for Hintze, during which Schmoller introduced him to archival research on the Prussian administrative files. Even this, his very first research on unpublished source materials, was carried out in connection with early comparative projects. At the same time, it was Hintze's first practical preparation for his work on the *Acta Borussica*, Gustav Schmoller's editing and research project on Prussian history in the eighteenth century in particular.

Hintze was engaged in this exhaustive project, which was based at the Prussian Academy of Sciences, in various functions over a period of decades. He was initially contracted in 1888 as a research fellow. In this capacity, he acquired an unparalleled command of the published and unpublished source material on the Brandenburg-Prussian state, its regions and provinces, classes and estates, before expanding the scope of his research to include the sixteenth and nineteenth centuries. He completed his postdoctoral thesis, a monograph on the history of the Prussian state economy (*The Prussian silk industry in the 18th century*), in 1895. In 1899 he was appointed professor at the University of Berlin, where—turning down several offers of chairs at other universities—he advanced to the status of full professor from 1900 to 1902–7, and continued to teach until prematurely given emeritus status in 1920. Besides his broad research, which had long since extended beyond the boundaries of Prussia, he owed this distinguished academic career to Gustav Schmoller's excellent contacts in political and administrative circles which is of vital importance in the view of the various animosities among contemporary historians.

The initial focus of Hintze's extensive writings was on constitutional and economic issues of Prussian history, culminating in 1901 in a lengthy monograph on the organization of the Prussian state administrative authorities around 1740. In this volume Hintze analyzed the making of the Prussian state, its development from a territorial state to a 'grand state.' He then reluctantly accepted an official commission to compile a comprehensive survey of the history of Brandenburg-Prussia. Several editions of this work were published in 1915–16, leading to heated (interdenominational) controversies during World War I.

In the course of archival research spanning two decades, Hintze acquired an excellent grasp of the source material, establishing an empirical basis that enabled him, even in further-reaching investigations, to take concrete historical life into consideration, particularly that of more recent centuries. For Hintze, the exploration of Prussian history was never a limited end in itself, but a paradigm for the historical process *per se*. This does not mean to say that he took this historical example as a yardstick by which the development of other states was to be measured, however. By his own account, it was at a early stage of his career—during his studies of legal and constitutional history under Schmoller—that Hintze had identified a much broader area as his future field of research: general constitutional and administrative history, from the comparative perspective. Furthermore, in choosing his Prussian topics of inquiry, he had always considered their usefulness for the 'entirety of my academic plans and studies' (1892).

Long before the World War I, which marked a turning point in Hintze's life, he became engaged in other, broader fields of research. In the course of his dispute with the protagonist of the older German historical school of economics, Wilhelm Roscher, Hintze had presented his conception of a comparative approach to the analysis of world history as early as 1897. In contrast to Roscher, whom he accused of neglecting the fundamental social aspects of state structures, Hintze devised a procedure that integrated both social conditions and the effects of foreign policy and power politics into the analysis of the formation of political structures. Hintze's main concern was to identify the recurring patterns in the development of states by means of a typological and comparative procedure. 'In the social region, there is a tendency towards regulated organic development: the making of states follows this developmental trend to a certain extent, simultaneously regulating it, but is by no means dependent on it alone.' Hintze maintained that the implications of 'power interests' were not to be underestimated, that the establishment of a state was 'at least as dependent on its external position of power as on the preconditions of the particular social structure for domestic governmental activity,' and that the 'external conditions cannot fail to influence social development.' It was this 'political side' of the formation of political structures which Hintze felt was neglected in any typological theory of states which, following the Aristotelian example, merely tried to identify analogous sequences of state systems. Hintze considered the process of state making to be largely dependent on the 'interrelations of the states, on the forces prevailing amongst them, on the rise and fall of the neighbouring states,' i.e., on what Leopold von Ranke had called the 'great world relationships.' (see

Ranke, Leopold von (1795–1886)) Hintze considered power as a given category, which—like the phenomenon of state rivalries—he did not attempt to deduce in any further detail. For him, the central aim was to define and register 'types conditioned by world history': phenomena observed in groups of states that had maintained close political relations. In this sense, he explained afterwards, he understood state making to be a 'general, regulative principle,' thus implying that he did not intend strict rules to be postulated. It seems that in his early writings, Hintze still assumed a cautious primacy of foreign policy factors in the process of state making, but that—remarkable in the history of science—he turned away from this explanatory model over the course of his further research and analysis.

In a whole series of individual studies, whose scope was gradually to broaden, reaching the dimensions of world history in about 1930, Hintze applied and tested his typologically-based comparative approach. He developed what could be termed a spatial-typological procedure, which makes (European) historical development appear to be that of a bonded unit of regions. Above all, Hintze was keenly interested in the differentiation of a 'core Europe' (with a close genetic and spatial relationship to the Carolingian Empire) and a ring of 'peripheral' countries. He tested this spatial typological and structural historical differentiation at various junctures in his comparative studies of constitutional history—in his works on the history of the European estates, for example. It also becomes apparent in the comparison of continental western and central European absolutism on the one hand and the insular parliamentarianism of England on the other, a differentiation which has admittedly been called into question in recent Anglo-Saxon research. The 'peripheral' states of England, Sweden, Poland, and Hungary were of great interest to Hintze on account of their numerous structural parallels, especially where the 'parliamentarian' tradition was concerned.

Hintze set himself the aim of writing a comprehensive 'general comparative constitutional and administrative history of the modern states, namely the Latin and Germanic peoples,' but the manuscript was no longer published from around 1930 onwards. Central parts of this work were lost after Hintze's death; the fragments that survived are currently being edited by an international team. New sources reveal that Hintze remained keenly interested in the 'individuality' of states and political processes. Of course, the historical legacy of Hintze's work should not be allowed to conceal his innovative potential.

Otto Hintze sought to establish contact with the systematic contiguous disciplines at an early stage of his career, and not merely in reaction to the traumatic upheaval of political cultures subsequent to the 1914/1918 war. On the contrary, he had already begun to lose interest in the Prussian paradigm in the years preceding the war. His openness to the more modern approaches taken by social history, psychology, and sociology became evident as early as 1900, in the mediating statement he made during the so-called Lamprecht dispute about the relationship of 'the individualistic and the collectivist conception of history' (1897). Hintze pointed to the significant contribution made to historiography by Comte and Spencer (see *Comte, Auguste (1798–1857)*; *Spencer, Herbert (1820–1903)*), both of whom he referred to in his lectures. For Hintze, individuality and collectivity constituted the two poles of historical insight, and he maintained that the methods of social psychology afforded great opportunities to go beyond Ranke. Admittedly, Hintze's growing interest in the concepts of sociology characterizes him as an outsider among the historians of his time. Hintze's great 1920s discussions on the theory and philosophy of history, not only with Weber (see *Weber, Max (1864–1920)*), but also with Oppenheimer, Sombart, and Troeltsch, as well as with the jurists Kelsen and Smend, merely marked the culmination of a continuous development in his work. This is demonstrated, for example, by Hintze's discussion of Sombart's *Modern Capitalism*, in which the Berlin historian also referred to Marx. The extent to which Hintze started from the assumption of a primary, autonomous concept of society is a matter of debate. While he regarded the state as the dominant factor, his later studies and the surviving fragments of manuscripts written prior to 1933 reveal marked changes and a broadening of approach—he now tended towards a concept of society which was not derived from the state alone. For Hintze, social and socioeconomic factors began to gain in causal-analytic significance; the same applies to the concept of class, though Hintze never approached Marxist conceptions with anything but a distant interest. It is thus also a matter of contention whether Hintze's typological procedure, along with the principle of 'clear abstraction' (*anschauliche* Abstraction), *ipso facto* neglected the element of dynamics. Hintze's procedure was related (or rather contiguous) to Weber's conception of the ideal type (*Idealtypus*), though always referring to the central question of the individualities of historical developments. It should be noted that Georg Jellinek had made pertinent suggestions for a typological methodology as early as 1900 in his *Allgemeine Staatslehre* (*General Theory of State*), and that Hintze was able to develop these ideas in his search for phenomena of 'relative comparability' (Kocka). In any case, Hintze saw his work as a necessary supplement to both the legacy of Ranke and modern sociology.

Hintze's reception of Weber's writings also predates 1914. Both the strength of his systematic interests and the positioning of methodological boundaries are evident in Hintze's second unpublished magnum opus, which can only be reconstructed to some extent from lecture notes. After 1920, Hintze had intended to

publish this work as a 'general theory of state and society on a historical basis,' a continuation of and supplement to his general constitutional history. He had intended to paint a 'typical picture of the modern state in terms of its shared properties and developmental trends, its various historical stages and individual forms.' In this 'policy' of Hintze, which he evolved in his middle period, the primacy of the state and of the law was (still) very apparent. 'The state must be considered from the legal, the philosophical, and the historical perspective'—in this order. Social factors were suited and subordinated to this conception. In contrast to his classical studies of comparative constitutional history, in which industry and workers (or the labor movement) played at most a marginal role, Hintze integrated the social developments of the nineteenth century into his 'policy' by means of critical discussion of Marx and Lassalle. In so doing, Hintze was able to identify developments in the middle classes and nonmanual occupations that were to be examined in greater detail by later social historians.

All this makes Hintze still of interest to present-day historians. Of course, not all of his works are relevant today, as a brief glance at his political writings will show. However, his openness to the contributions of the social and political sciences—highly unusual among his contemporaries in the field—and the scope of his view, which finally extended beyond eurocentrism, mean that he remains an important, innovative figure in modern political structural history. Admittedly, the reconstruction of his *oeuvre* is essential to apply his approach in practical research also for further developing his theories, yet this is forever hindered by the fact that central works were never published and can only be substituted and inferred. After 1933 Otto Hintze, who had married the Jewish historian Hedwig Guggenhämer, became a *persona non grata* in Germany. He ordered that his personal papers should be destroyed after his death, which for the most part they were. Hintze died in his Berlin apartment on April 25, 1940, long since isolated from academic life.

See also: Comparative History; Constitutionalism; Constitutionalism, Comparative; Historiography and Historical Thought: Christian Tradition; Middle Ages, The; Nation: Sociological Aspects; Nations and Nation-states in History; Policy History: Origins; Policy History: Regimes; Policy History: State and Economy; Political History (History of Politics); State and Society; State Formation; State, History of; State, Sociology of the

Bibliography

Büsch O, Erbe M (eds.) 1983 *Otto Hintze und die moderne Geschichtsschaft* [*Otto Hintze and Modern Historiography*]. Colloquium Verlag, Berlin

Hintze O 1908 *Historische und politische Aufsätze* [*Historical and Political Essays*]. Verlag Deutsche Bücherei, Berlin, 4 Vols.

Hintze O 1962–1967 In: Oestreich G (eds.) *Gesammelte Abhandlungen* [selected works], 2nd and 3rd edns. Vandenhoeck & Ruprecht, Göttingen, 3 Vols.

Hintze O 1998 Di Costanzo G, Erbe M, Neugebauer W (eds.) *Allgemeine Verfassungs- und Verwaltungsgeschichte der neueren Staaten. Fragmente* [*General Constitutional and Administrative History of the Newer States. Fragments*]. Palmar, Naples, Italy, Vol. 1

Kocka J 1972 Otto Hintze. In: Wehler H-U (ed.) *Deutsche Historiker*. Vandenhoeck & Ruprecht, Göttingen, Vol. 3, pp. 41–64

Kocka J 1981 Otto Hintze, Max Weber und das Problem der Bürokratie [Otto Hintze, Max Weber and the problem of bureaucracy]. *Historische Zeitschrift* **233**: 65–105

Neugebauer W 1993 Otto Hintze und seine Konzeption der 'Allgemeinen Verfassungsgeschichte der Neueren Staaten' [Otto Hintze and his conception of the 'General consitutional history of the newer states']. *Zeitschrift für historische Forschung* **20**: 65–96

Neugebauer W 1998 Die wissenschaftlichen Anfänge Otto Hintzes [The academic origins of Otto Hintze]. *Zeitschrift der Savigny-Stiftung für Rechtsgeschichte*, *Germanistische Abteilung* **115**: 540–51

Neugebauer W 1999 Zum schwierigen Verhältnis von Geschichts, Staats- und Wirtschaftswissenschaften [On the difficult relationship between history, political science and economics]. In: Kocka J (ed.) *Die Königlich Preußische Akademie der Wissenschaften im Kaiserreich*. Akademie Verlag, Berlin, pp. 235–75

Neugebauer W 2000 Zur Quellenlage der Hintze-Forschung [On source material in research on Hintze]. *Jahrbuch für die Geschichte Mittel- und Ostdeutschlands* **45**: 323–38

Oestreich B 1989 Otto Hintze. In: Erbe M (ed.) *Berlinische Lebensbilder. Geisteswissenschaftler* [*Berlin Lives. Arts Scholars*]. Colloquim Verlag, Berlin, pp. 287–309

Smith L S 1967 Otto Hintze's comparative constitutional history of the West. Doctoral dissertation, Saint Louis, MI

W. Neugebauer

Hippocampus and Related Structures

The hippocampus is part of the cortex of the forebrain in mammals and is closely linked to a restricted number of related cortical areas which are here collectively referred to as the parahippocampal region. This collection of cortical regions, which in humans is located in the most medial portion of the temporal lobe, is involved critically in learning and memory processes, in particular in long-term explicit memory. The hippocampus also has strong functional links with subcortical structures believed to play a role in the selection of appropriate behavior to favor either survival of the individual or of the species. In this

article, the wiring of the hippocampus and the parahippocampal region underlying these two proposed functions will be described, and experimental data that support the proposed functions will be summarized.

1. What is the Hippocampus and What are Related Structures?

The hippocampus is a macroscopically defined cortical structure which is present in the brains of all mammals. *Grosso modo*, its appearance looks alike in all species in that it is an enrolled continuation of the most medial rim of the cortical surface of the hemisphere. The cortices directly adjacent to the hippocampus, which are referred to collectively as the para (next-to) hippocampal region, are considered as the related structures. The particular arrangement of these different cortical structures can be appreciated easily in sections taken perpendicular to the long axis of the hippocampus (see Fig. 1).

The hippocampus and the parahippocampal region of different species are positioned differently in the brain, such that in species with a clearly developed temporal lobe, the hippocampus is positioned more ventrally and anteriorly (humans and monkeys) compared to the situation in, for example, the rat, where the hippocampus looks more like a c-shaped structure positioned in the caudal third of the hemisphere, such that its dorsal one-third has a marked spatial relation with the corpus callosum. However, these differences in position do not alter the major characteristics and the topological relations between the hippocampus and the parahippocampal structures (see Sects 2 and 3). Nor does it influence the fact that the most anterior/ventral portion of the hippocampus has a close spatial relationship with the amygdala (see *Amygdala* (*Amygdaloid Complex*)).

2. General Features of the Hippocampus/ Parahippocampal Region and Species Differences

The hippocampus consists of three major subdivisions, which can be recognized easily in all species (see Fig. 1). The first field, the dentate gyrus, is a c-shaped, three-layered cortical structure which is characterized by a densely packed cell layer, mainly consisting of granular cells, i.e., cells with a round soma, and an apical dendritic tuft and no, or only sparse, basal dendrites. Enclosed in this c-shaped cortical blade is a polymorphic cell group generally referred to as the hilus of the dentate gyrus. The second subfield is the Ammon's Horn or Cornu Ammonis, generally abbreviated as CA-field. The CA-field is generally subdivided into CA1, CA2, and CA3. The third subdivision is the subiculum. Note that, according to some authors, a prosubiculum can be differentiated from the subiculum proper. Also, readers may find texts in which reference is made to the subicular complex, consisting of (pro)subiculum, presubiculum, and parasubiculum. Based on connectional arguments, it is preferred to consider the presubiculum and parasubiculum as being functionally different from the subiculum (cf. Amaral and Witter 1995; and for a very detailed comparison between species the reader is referred to Stephan 1975).

The parahippocampal region, in contrast to the hippocampus, comprises cortical regions with more than three layers—generally, five or six layers are distinguished. In all species, the parahippocampal region comprises five different regions, the pre- and parasubiculum, the perirhinal and entorhinal cortices, and a fifth area which in primates is commonly referred to as the parahippocampal cortex. In nonprimates, the most likely homolog for the latter area is the so-called postrhinal cortex (cf. Burwell et al. 1995, Witter et al. 1989). The respective names of some of the areas that make up the parahippocampal region refer to the rhinal sulcus, which can be distinguished in all mammals. However, its particular features differ considerably between species. Whereas in rats it is a distinguishable fissure that runs along the rostrocaudal extent of the hemisphere, in humans it is only present in about 50 percent of the cases, as a shallow and short indentation in the most medio-anterior temporal lobe. It is of interest that the spatial relationships among the different fields of the parahippocampal region appear similar in, for example, mouse, rat, monkey, and human (see Fig. 2).

A remark is appropriate concerning the question of whether nonmammalian species do have a hippocampus and related structures. In reptiles, cortical areas have been described which resemble, both in terms of celltype as well as connectivity, the mammalian dentate gyrus and Ammon's Horn. Part of the olfactory-related cortex of reptiles may be considered homologous to the mammalian parahippocampal region (see Ten Donkelaar 1997 for further reading). Interestingly, there is some evidence that lesions of these structures in reptiles interfere with learning. With respect to birds, there is convincing evidence that a particular region in the forebrain, homologues to the reptilian cortex, is functionally important for spatial memory. This region is generally referred to as the hippocampus, but structurally it looks quite different from the mammalian hippocampus. Connection-wise, though, it is quite similar (see Clayton and Krebs 1995 for further details).

3. Wiring of the Hippocampus and Parahippocampal Region

Most detailed information about the connectivity of the system comes from anatomical tracing studies in

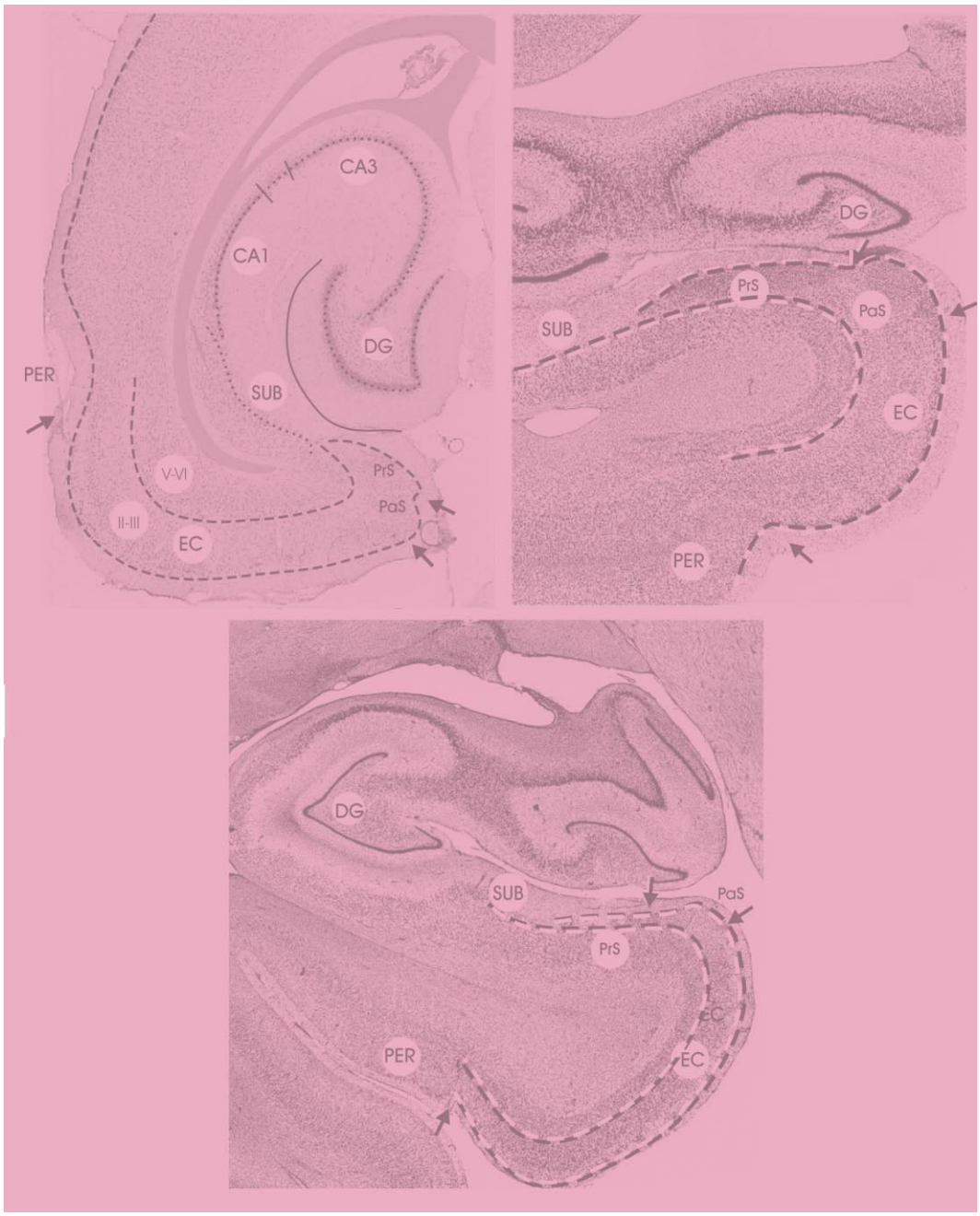

Figure 1
Three sections of the hippocampus and related parahippocampal structures cut perpendicular to the antero-posterior axis of the forebrain of the rat (upper left), the monkey (upper right), and the human (lower panel; modified after Scharfman et al. 2000, p. 11, reproduced by permission of the New York Academy of Sciences)

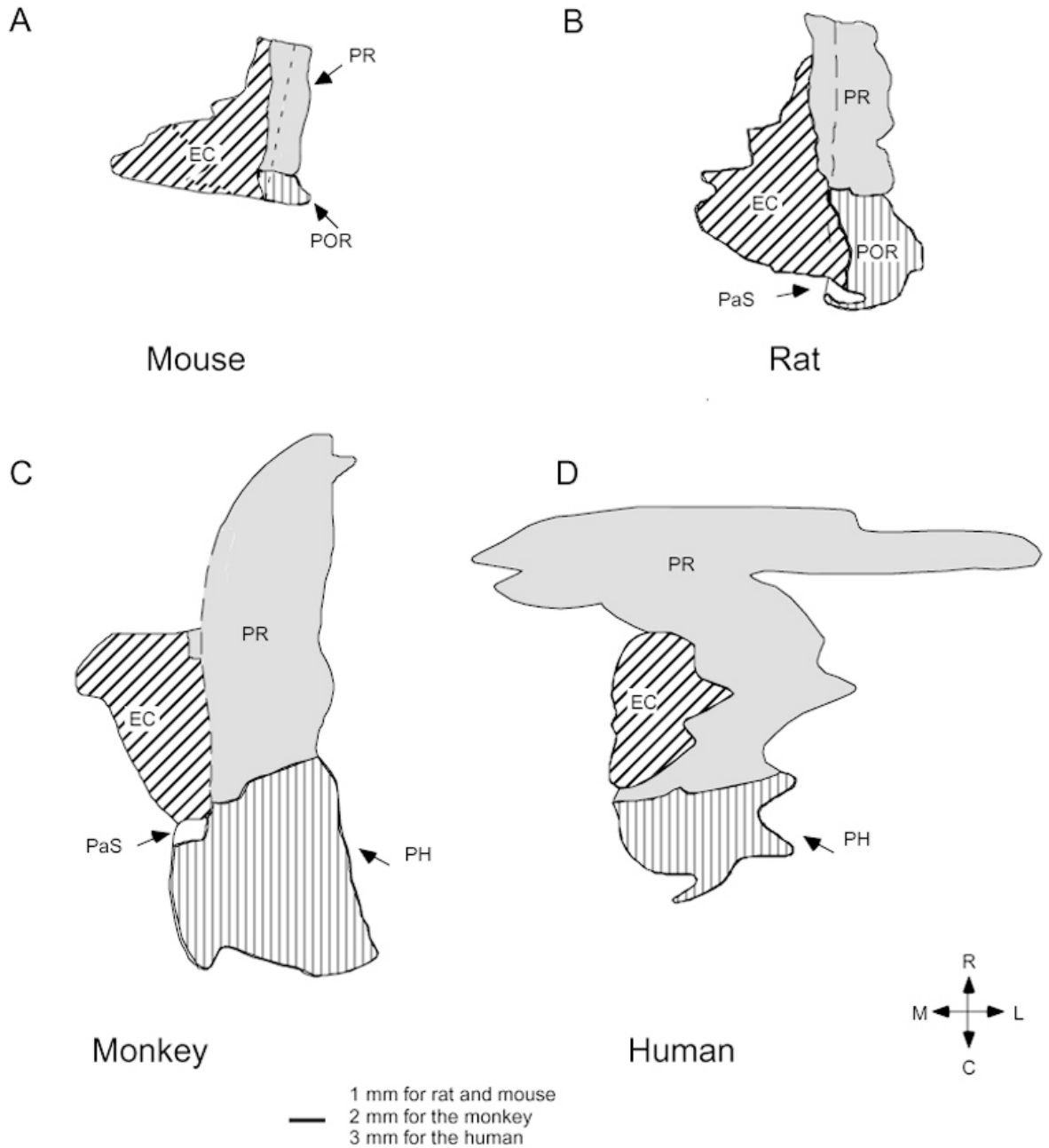

Figure 2
Two-dimensional representations of the parahippocampal region in the mouse, rat, monkey, and human, illustrating that in all four species this region comprises the same subdivisions but that their spatial arrangements and individual relative sizes differ between species. Note that the pre- and parasubiculum are not represented in these figures (from Burwell 2000, p. 29, reproduced by permission of the New York Academy of Sciences)

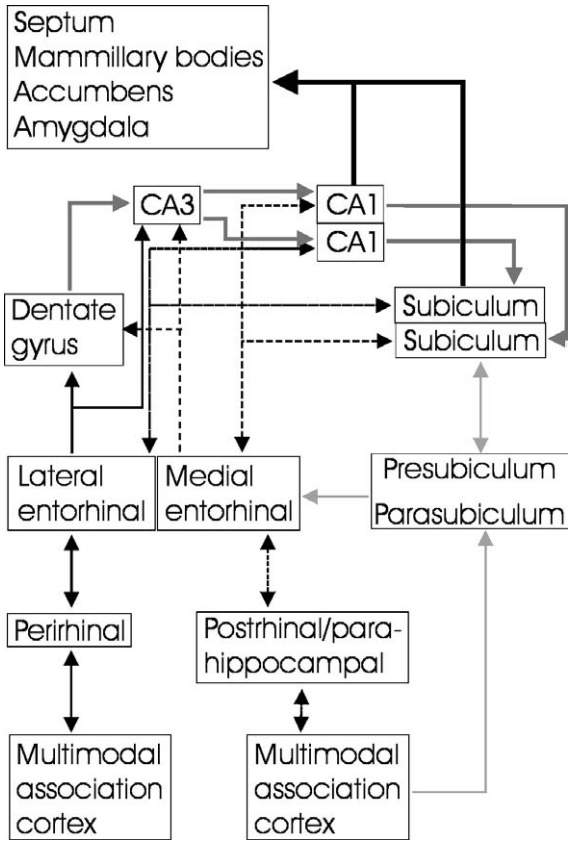

Figure 3
Summary diagram of connectivity of the hippocampus illustrating the two major communication pathways: 1, the cortico–hippocampal circuitry through the hippocampal region; 2, the fornix

the rat, although for some pathways, relevant detailed information is available for the monkey, and for the cat as well. The overall picture which emerges from all these studies is that this connectivity is rather conservative, such that a general non-species specific description may suffice. For further details, readers are referred to Witter et al. (1989, 2000a, 2000b), Amaral and Witter (1995), Burwell (2000), and Suzuki and Amaral (1994a, 1994b). The hippocampus has two major pathways through which it is connected to the rest of the brain. The first pathway makes use of the so-called cortico–parahippocampal–hippocampal circuit, and the second is formed by the fornix (see Fig. 3).

3.1 Cortico–Parahippocampal–Hippocampal Circuit

A striking feature of the cortico–parahippocampal–hippocampal circuit is that the hippocampus does not have direct connections with the majority of the functionally different cortical regions. In contrast, this communication is mediated by way of the parahippocampal region. Of this region, the perirhinal and postrhinal/parahippocampal cortices maintain the highest number of cortico–cortical connections, including connections within all major sensory realms, such as visual, auditory, somatosensory, and gustatory ones. Also, massive connections exist with the other two major higher-order association domains, the parietal and prefrontal cortices, but connections with the motor cortex are less prevalent. The perirhinal and postrhinal/parahippocampal cortices, in turn, provide a major input to the superficial layers II and III of the entorhinal cortex, which are further reached by inputs from olfactory structures, the amygdala, and from the remaining two components of the parahippocampal region, i.e., the pre- and parasubiculum. Neurons in entorhinal layers II and III, which are the recipients of this wide variety of inputs, give rise to the major cortical input to the hippocampus, the so-called perforant pathway. This name is derived from the traditional descriptions by Ramon y Cajal (1911), who noted that fibers from the entorhinal cortex perforate the underlying white matter and the cortical lamina of the subiculum to gain access to the dentate gyrus molecular layer. In order to understand the potential functioning of the system, it is critical to point out that the perforant pathway projection harbors two different systems. Layer II cells distribute their axons to the dentate gyrus and CA3, whereas layer III cells send their axons to CA1 and the subiculum. In addition to this intriguing difference in target specificity, the terminal distributions of these two components of the perforant pathway are entirely different. Layer II cells give rise to a widespread projection to all dentate granular cells. In contrast, axons of layer III cells target only restricted groups of the available neurons in CA1 and the subiculum (see Sect. 5.1 for functional relevance; see also Fig. 3). This particular feature is of interest, since field CA1 and the subiculum constitute the major output structures of the hippocampus. Hippocampal output from these two hippocampal fields is distributed by way of projections back to the entorhinal cortex, from where ongoing output projections originate to a host of cortical and subcortical structures, including projections to perirhinal and parahippocampal/postrhinal cortices, prefrontal areas, amygdala, ventral striatum, and midline thalamus.

A further refinement has to be added to the overall organization of the connections between the parahippocampal region and the hippocampus. It has been established that the input–output system consists of at least two parallel pathways. One of the pathways carries inputs from the perirhinal cortex into and away from the hippocampus, whereas the second mediates transfer of information from and to the parahippocampal/postrhinal cortex. In view of the anatomical

organization outlined above, it has thus been suggested that in the hippocampus these two pathways converge at the level of the dentate gyrus and CA3, whereas they are kept more or less separate at the level of CA1 and the subiculum (cf. Witter et al. 2000b).

3.2 The Fornix

The fornix is a major fiber bundle which connects the hippocampus to the hypothalamus, in particular the mammillary bodies. On its way, the fornix also issues fibers to the septal complex, the ventral striatum, and the amygdala. As indicated above, the fornix is not a pure output pathway, since projections from the septal complex (see Sect. 5.2) to the hippocampus and in part to the entorhinal cortex travel by way of the fornix. In addition, the commissural connections between the left and right hippocampi also partially travel by way of the fornix. The hippocampal connection to the mammillary bodies is part of the traditionally described limbic or Papez circuit, which includes the mammillo-thalamic tract connecting the mammillary bodies with the anterior complex of the thalamus, which in turn project to large portions of the limbic cortex, including anterior- and posterior-cingulate cortex, and pre- and parasubiculum (see also *Cingulate Cortex* and *Limbic System*). All these structures, in turn, provide input to the hippocampus, either directly, or indirectly by way of the entorhinal cortex.

4. The Hippocampus is not a Homogeneous Structure

When one looks at the overall anatomical organization of the parahippocampal–hippocampal system, it appears that the hippocampus is rather homogeneous along its long axis. However, differences in numbers of neurons between the dorsal/posterior and ventral/anterior hippocampus have been reported, indicative for differences in the overall neuronal network. Moreover, in the rat and in the monkey, it has been established that the dorsal/posterior portion of the hippocampus receives its major cortical input from the more sensory-related or exteroceptive lateral and caudal portions of the entorhinal cortex, whereas the ventral/anterior hippocampus receives inputs from more medial parts of the entorhinal cortex, which most likely convey information that reflects the interoceptive status of the individual. In addition, direct reciprocal connections with the amygdala are limited to the ventral/anterior and intermediate levels of the hippocampus and are absent for the dorsal/posterior portion of the hippocampus (cf. Witter 1986). Behavioral data in support of this notion are the findings that in rats the dorsal hippocampus is critical for the initial learning and long-term memory of spatial information, whereas the ventral hippocampus is not essential (cf. Moser and Moser 1998). Recently, brain imaging studies have revealed that in humans functional differences are also present between anterior and posterior halves of the hippocampus; however, the precise nature of these differences is still an issue for debate (Lepage et al. 1998, Schacter and Wagner 1999).

5. Functions of the Hippocampus and Parahippocampal Region

5.1 Functions of the Cortico–Hippocampal System

Ample clinical and experimental evidence supports the view that the hippocampus and parahippocampal structures are essential for memory processes, although the unique contributions of each of the different subfields are as yet incompletely understood (Scoville and Milner 1957, Squire 1992, Eichenbaum et al. 1994). The results of the behavioral studies, in particular, have provided insights into how each of the components of the hippocampal system contributes to memory processes. Instead of contributing in a unitary way to memory, it appears evident that individual structures make selective contributions. Such studies indicated that the hippocampal formation itself does not contribute critically to object recognition memory. Instead, it is the laterally positioned component of the parahippocampal region, the perirhinal cortex, that appears to be involved critically in this task (Murray et al. 2000). In contrast, the more posterior portions of the parahippocampal region, the parahippocampal cortex in primates or the postrhinal cortex in nonprimates, may be involved critically in spatial or topographical memory (Bohbot et al. 2000, Vann et al. 2000, cf. Witter et al. 2000b). Concerning the role of the entorhinal cortex, the results of lesion studies appear inconclusive; still, it is most likely that the entorhinal cortex operates cooperatively with the hippocampal formation, such that both structures process information in different, yet complementary ways. With respect to the hippocampal formation, it has been proposed that it plays a critical role in mediating memory for relevant relationships among conjunctions or associations unique to a particular episode (Eichenbaum et al. 1994).

5.2 Functions of Hippocampus–Fornix Connectivity

As indicated above, the fornix provides the hippocampus with another pathway, beside the cortical route, to communicate with a host of subcortical brain structures, including the mammillary bodies. Behavioral data indicate a clear functional dichotomy between the cortical versus the fornix route. However,

the role of the fornix pathway in learning and memory is as yet poorly understood. Lesions of the main target of the fornix, i.e., the mammillary complex in animals, do not result in marked memory deficits, although in Korsakoff's patients, which do show a marked anterograde as well as retrograde memory deficit, the mammillary bodies show a marked decrease in volume (cf. Aggleton and Brown 1999). In addition, there is convincing human evidence from patients with thalamic infarctions in favor of a role in memory processes of the connection from the mammillary complex to the anterior complex of the thalamus (van der Werf et al. 2000).

The fornix also mediates connectivity from the hippocampal–parahippocampal structures to the ventral striatum and the amygdala. In the ventral striatum, interactions may take place between hippocampal, amygdaloid, and prefrontal outputs, leading to the selection of appropriate behavioral strategies, for example, in learning paradigms (consult Groenewegen et al. 1999 for further reading). The functional relevance of the reciprocal connections between the amygdala and hippocampus, and for that matter also those between the amygdala and the parahippocampal region, is still poorly understood (Pitkanen et al. 2000, Stefanacci et al. 1996). It is assumed that amygdala inputs may provide the hippocampal system with a marker for relevance to a particular memory episode.

The connections to and from the septal complex of both the hippocampus and parahippocampal structures use the fornix as well as a number of other white matter pathways. Although much about these interactions has yet to be resolved, it is clear that the normal functioning of the hippocampal–parahippocampal–cortical interactions depends strongly on a functional septal complex. It is in particular the fact that the septal complex supplies the hippocampus and parahippocampal region with most of its cholinergic input, as well as the role of the septal complex in the synchronization of electrical activity which relates to its functional relevance (see Sect. 5.3).

5.3 Oscillatory Activity

The hippocampus shares with the entorhinal cortex the presence of strong rhythmical activity, in particular theta rhythm, which is coexpressed with trains of gamma waves during exploring behavior in the awake animal as well as during a particular sleep phase called Rapid Eye Movement (REM) sleep. Both structures also share the occurrence of so-called sharp waves, a high-frequency discharge which occurs predominantly during the alternating sleep phase, the so-called slow-wave sleep phase. Although the electrophysiological mechanisms underlying the occurrence of such oscillatory phenomena have been clarified to some point, the functional relevance, for example, in relation to learning, is still debated. There appears to be consensus though that the normal functioning of the cortico-hippocampal learning system depends strongly on the normally occurring alternations between the theta and the sharp wave state (cf. Buzsaki 1996).

5.4 Clinical Relevance

The hippocampal–parahippocampal structures play an established role in learning and memory processes, which is supported by reports concerning pathological changes seen in diseases that are associated with learning and memory disorders such as Alzheimer's disease. However, other neurological/psychiatric disorders, such as epilepsy, schizophrenia, Huntington's, Parkinson's and Pick's disease are associated with pathological alterations in these regions as well. The most striking observation, however, is the specificity of the pathology in each of these diseases. In all except epilepsy and schizophrenia, the first pathological changes are seen in the anterolateral part of the entorhinal cortex, also called the transentorhinal cortex. However, in all these diseases, the layer showing the most dramatic alterations seems to be disease specific (Braak et al. 2000). In the case of epilepsy, the most striking initial pathology is observed in layer III of the medial entorhinal cortex (Schwarcz et al. 2000), whereas in the case of schizophrenia, the pathology appears to start in layer II throughout the entorhinal cortex (Arnold 2000). In all these disorders, the hippocampus becomes involved as well, leading to an overall dramatic reduction in neuronal numbers and/or connections.

See also: Alzheimer's Disease, Neural Basis of; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Long-term Depression (Hippocampus); Long-term Potentiation (Hippocampus)

Bibliography

Aggleton J P, Brown M W 1999 Episodic memory, amnesia and the hippocampal–anterior thalamic axis. *Behavavioral and Brain Sciences* **1999**: 425–89

Amaral D G, Witter M P 1995 Hippocampal formation. In: Paxinos G (ed.) *The Rat Nervous System*. Academic Press, San Diego, CA

Arnold S 2000 Cellular and molecular neuropathology of the parahippocampal region in schizophrenia. *Annals of the New York Academy of Science* **911**: 275–92

Bohbot V D, Allen J B J, Nadel L 2000 Memory deficits characterized by patterns of lesions to the hippocampus and parahippocampal cortex. *Annals of the New York Academy of Science* **911**: 355–68

Braak H, Del Tredici K, Bohl J, Bratzke H, Braak E 2000 Pathological changes in the parahippocampal region in select non-Alzheimer dementias. *Annals of the New York Academy of Science* **911**: 221–39

Burwell R D 2000 The parahippocampal region: corticocortical connectivity. *Annals of the New York Academy of Science* **911**: 25–42

Burwell R D, Witter M P, Amaral D G 1995 Perirhinal and postrhinal cortices of the rat: a review of the neuroanatomical literature and comparison with findings from the monkey brain. *Hippocampus* **5**: 390–408
Buzsaki G 1996 The hippocampal–neocortical dialogue. *Cerebral Cortex* **6**: 81–92
Clayton N S, Krebs J R 1995 Food-storing memory and the hippocampus. *Current Opinion in Neurobiology* **5**: 149–54
Eichenbaum H, Otto T, Cohen N J 1994 Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences* **17**: 449–72
Groenewegen H J, Mulder A B, Beijer A V J, Wright C I, Lopes da Silva F H, Pennartz C M A 1999 Hippocampal and amygdaloid interactions in the nucleus accumbens. *Psychobiology* **27**: 149–64
Lepage M, Habib R, Tulving E 1998 Hippocampal PET activations and memory encoding and retrieval: The Hiper model. *Hippocampus* **8**: 313–22
Moser M B, Moser E I 1998 Functional differentiation in the hippocampus. *Hippocampus* **8**: 608–19
Murray E A, Bussey T J, Hampton R R, Saksida L M 2000 The parahippocampal region and object recognition. *Annals of the New York Academy of Science* **911**: 166–74
Pitkanen A, Pikkarainen M, Nurminen N, Ylinen A 2000 Reciprocal connections between the amygdala and the hippocampal formation, perirhinal cortex and postrhinal cortex in the rat. *Annals of the New York Academy of Science* **911**: 369–91
Ramón y Cajal S 1911 *Histologie du Système Nerveux de l'Homme et des Vertebrés*. Maloine, Paris
Schacter D L, Wagner A D 1999 Medial temporal lobe activations in fMRI and PET studies of episodic encoding and retrieval. *Hippocampus* **9**: 7–24
Scharfman H E, Witter M P, Schwarcz R 2000 Preface. *Ann. New York Acad. Sci* **911**: ix–xiii
Schwarcz R, Eid T, Du F 2000 Neurons in layer III of the entorhinal cortex. A role in epileptogenesis and epilepsy? *Annals of the New York Academy of Science* **911**: 328–42
Scoville W B, Milner B 1957 Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry* **20**: 11–21
Squire L R 1992 Memory and hippocampus: A synthesis of findings with rats, monkeys and humans. *Psychology Review* **99**: 195–231
Stefanacci L, Suzuki W A, Amaral D G 1996 Organization of connections between the amygdaloid complex and the perirhinal and parahippocampal cortices in macaque monkeys. *Journal of Comparative Neurology* **375**: 552–82
Stephan H 1975 Allocortex. In: Borgman W (ed.) *Handbuch der Mikoskopischen Anatomie des Menschen* IV-9. Springer-Verlag, Berlin, pp. 1–998
Suzuki W A, Amaral D G 1994a The perirhinal and parahippocampal cortices of the Macaque monkey: Cortical afferents. *Journal of Comparative Neurology* **350**: 497–533
Suzuki W A, Amaral D G 1994b Topographic organisation of the reciprocal connections between the monkey entorhinal and the perirhinal and parahippocampal cortices. *Journal of Neuroscience* **14**: 1856–77
Ten Donkelaar H J 1997 Reptiles. In: Nieuwenhuys R, Ten Donkelaar H J, Nicholson C. (eds.) *Central Nervous System of Vertebrates*. Springer-Verlag, Berlin, Vol 2, pp. 1315–524
van der Werf Y D, Witter M P, Uylings H B M, Jolles J 2000 Neuropsychology of infarctions in the thalamus: A meta-analysis. *Neuropsychologia* **38**: 613–27
Vann S D, Brown M W, Erichsen J T, Aggleton J P 2000 Fos imaging reveals differential patterns of hippocampal and parahippocampal subfield activation in rats in response to different spatial memory tests. *Journal of Neuroscience* **20**: 2711–18
Witter M P 1986 A survey of the anatomy of the hippocampal formation, with emphasis on the septotemporal organization of its intrinsic and extrinsic connections. In: Schwarcz R, Ben-Ari Y (eds.) *Advances in Experimental Medicine and Biology, Vol. 203: Excitatory Amino Acids and Epilepsy*. Plenum Press, New York, pp. 67–82
Witter M P, Groenewegen H J, Lopes da Silva F H, Lohman A H M 1989 Functional organisation of the extrinsic and intrinsic circuitry of the parahippocampal region. *Progress in Neurobiology* **33**: 161–254
Witter M P, Naber P A, van Haeften T, Machielsen W C M, Rombouts S A R B, Barkhof F, Scheltens Ph, Lopes da Silva F H 2000a Cortico-hippocampal communication by way of parallel parahippocampal–subicular pathways. *Hippocampus* **10**: 398–410
Witter M P, Wouterlood F G, Naber P A, van Haeften T 2000b Anatomical organization of the parahippocampal–hippocampal network. *Annals of the New York Academy of Science* **911**: 1–24

M. P. Witter

Historical Archaeology

Historical archaeology is the concurrent study of material culture and documentary evidence. This article outlines the scope of the practice of historical archaeology, some of its principal features as a discipline, and some of the major research questions for which historical archaeologists seek answers.

In a strict sense, historical archaeology should begin with the discovery of writing, for it is from this time onwards that societies are able to express themselves both in the ways in which they fashion the world around them—their artifacts—and also through recorded self-expression: the written record. But in practice, these early literate periods have been left to classical archaeology, and historical archaeologists have defined their domain as the last 500 years. This half millennium saw the creation of extensive 'world systems' and, in particular, the colonization of the Americas, Africa, and parts of Asia by Europeans. This focus has given historical archaeology its particular character, and some have gone so far as to describe it as the archaeology of European colonial expansion. But more recently—as we shall see—these conventions have been challenged, opening up the possibility of a dynamic and expanding area of enquiry.

The practice of historical archaeology is rooted in the study of the European settlement of North America and, in particular, of the early seventeenth century colonial settlement in New England and the Chesapeake. Here, a rich heritage of material culture

has long been closely associated with a sense of national identity, best represented by the archaeology of Colonial Williamsburg, the eighteenth century capital of Virginia which has been preserved and reconstructed as an interpretation of daily life in America's early years (Colonial Williamsburg 2000). This focus has resulted in a rich narrative of early settlement, best represented by Ivor Noel Hume's excavation and interpretation of Martin's Hundred (Noel Hume 1983). Here, we see the carefully excavated traces of daily domestic life matched with an imperfect documentary record to reveal a far fuller narrative history than words alone could tell. Martin's Hundred is matched by numerous other studies that provide an ever-emerging picture of life in the early American colonies.

More recently, this emphasis on the settler and colonist has been balanced by the emergence of an archaeology of slavery. Archaeologists of slavery—whether working on former plantations in the south or on slave living sites in northern towns—have made the point that archaeology's claim to be the history of ordinary people must remain hollow if all dimensions of colonial life are not fully explored (Singleton 1995). In this area, historical archaeology has made notable gains in revealing domestic life on plantations and some of the aspects of identity, including the persistence of elements of African material culture in house design and personal artifacts, distinctive culinary practices, and patterns in composite material culture assemblages.

Historical archaeology in North America now extends well beyond this 'classic' focus and there is a variety of specialized subfields. These include the archaeology of western frontier settlements, battlefield archaeology, the archaeology of industrial sites, and urban archaeology. The practice of historical archaeology has also been strongly influenced by the professional field of cultural resource management, (CRM) as large State and federal projects have required numerous surveys and excavation projects ahead of development and urban renewal projects. CRM work has ranged from small studies of individual archaeological features to comprehensive surveys and extensive excavations in the face of freeway construction. Such work tends to be characterized by a heavy emphasis on fieldwork and recording, and less attention given to analysis, interpretation and historical synthesis, leading to a division between field-oriented archaeologists working in cultural resource management and archaeologists working in museums, universities, and other public institutions.

Despite a fondness for the self-image of a rugged fieldworker, uncluttered by intellectual pretensions, historical archaeology has reflected the major theoretical trends of the social sciences over the past half century (Hall 2000, Orser and Fagan 1995). Early work—such as Noel Hume's study of Martin's Hundred—was avowedly narrative. In this approach (which has continued through to the present) the results of archaeological excavations serve to amplify and illustrate an incomplete documentary history, contributing to telling a story that is made richer by the tangible traces of the past. This empirical emphasis was challenged from several directions. Firstly, the 'New Archaeology' of the 1960s and 1970s sought law-like generalizations of human behavior through the evidence of material culture—an approach represented in Stanley South's search for standard patterns across archaeological sites (South 1977).

This approach was challenged by structuralism, best represented by James Deetz' widely read *In Small Things Forgotten* (1996). Here, material culture in its full breadth—from pipe-stems to houses—was organized as evidence of the human mind at work in creating systems of binary oppositions, together constituting 'world views.' Structuralism reasserted the principles of narrative history in seeking to show how European settlers 'became American' when they adopted an embedded mental template that determined how their world was arranged—the essence of cultural identity. But, in emphasizing the commonality of world views, structuralism lost a clear sense of cause and effect, and the processes of history unfolding. This prompted new interpretative approaches that turned to historical materialism for theoretical models. Well represented by Mark Leone's work in Maryland's Annapolis (1988), this approach has seen in the archaeological record the evidence for the emergence of merchant capitalism and social and economic control through the manipulation of material culture to create a false consciousness that disguises the true location of power. For example, Leone showed how garden design served to naturalize the authority of the new class of merchant capitalists in Annapolis, and how such naturalization extended to other categories of material culture, including musical instruments and ceramics.

More recently, the overwhelming association between historical archaeology and North America has been challenged by strong developments in other parts of the world. Each regional focus has tended to develop its own identity while remaining linked to the core concepts of this discipline. This is opening up the possibilities of a richer body of theory and of new areas of comparative study. One such area of development has been in Australia, where the archaeology of colonial settlement has gained increasing attention (Connah 1988). Here, studies have ranged from the historical archaeology of colonial buildings and early settler footholds on the continent to women's prisons in Tasmania. This has been matched by the development of historical archaeology in South Africa, where Dutch settlement in the mid-seventeenth century was followed by a century of British colonial settlement and a diverse material record (Hall 1993). Newer work in areas such as the Caribbean and in South America is adding a comparative richness to the field.

Despite these developments, though, historical archaeology still remains overwhelmingly the archaeology of European colonial expansion. Despite notable exceptions (for example, Johnson 1996), the archaeology of Europe itself in recent centuries still remains a separate field—'post-Medieval archaeology.' The archaeology of colonial contact—and the impact of European colonial expansion on the other side of colonial frontiers—remains largely unexplored. Claims for other forms of historical archaeology, for example the use of oral traditions as verbal sources (Andah 1990), remain largely unacknowledged. And the historical archaeologies of other large-scale, literate systems such as those of China or of Islamic expansion and settlement remain outside the scope of the field.

For some, historical archaeology is no more than a transition zone between preliterate periods, in which material culture is the only source of evidence, and documented history. But for others, historical archaeology has a distinctive character that makes it a discipline in its own right. There is some justification in this second point of view. Increasingly, historical archaeology has a distinctive set of methodologies and a suite of particular research questions of its own. Historical archaeology has its own professional journals, and its practitioners attend their own conferences. If this were the world of natural organisms, we could say that speciation is well advanced.

Historical archaeology shares with the rest of archaeology the basic methodologies of excavation. But although historical archaeologists often work with the precision of the prehistorian, teasing evidence from the slightest detail, they also often work to a larger scale. This is because the archaeological sites are large, with complex histories of deposition and disturbance which make the mechanical excavator a more appropriate research tool than the trowel. In addition, a large proportion of historical archaeological evidence is not beneath the ground. Standing buildings, field systems, roads and town layouts can reveal as much about the past as a conventional excavation. This brings historical archaeology close to fields such as ecology, cartography, architectural studies, and historical geography in its methodologies. Indeed, while retaining its ties with the time-honored skills of excavation, historical archaeology is quintessentially interdisciplinary in its search for, and analysis of, evidence.

See also: Art History; Cartography; Cultural Resource Management (CRM): Conservation of Cultural Heritage; Environmentalism: Preservation and Conservation; Historical Geography; Writing, Evolution of

Bibliography

Andah B W 1990 The oral versus the written word in the cognitive revolution: Language, culture and literacy. *West African Journal of Archaeology* **20**: 18–45

Colonial Williamsburg Foundation 2000 Williamsburg restored. http://www.history.org. 10 January 2000

Connah G 1988 *'Of the Hut I Builded': The Archaeology of Australia's History*. Cambridge University Press, London

Deetz J 1996 *In Small Things Forgotten. An Archaeology of Early American Life*, (expanded and revised). Anchor Books, New York

Hall M 1993 The archaeology of colonial settlement in southern Africa. *Annual Review of Anthropology* **22**: 177–200

Hall M 2000 *An Archaeology of the Modern World*. Routledge, London

Johnson M 1996 *An Archaeology of Capitalism*. Blackwell, Oxford, UK

Leone M 1988 The Georgian order as the order of merchant capitalism in Annapolis, Maryland. In: Leone M, Potter P (eds.) *The Recovery of Meaning. Historical Archaeology in the Eastern United States*. Smithsonian Institution, Washington DC, 235–61

Noel Hume I 1982 *Martin's Hundred: The Discovery of a Lost Colonial Virginia Settlement*. Knopf, New York

Orser C E, Fagan B M 1995 *Historical Archaeology*. HarperCollins, New York

Singleton T A 1995 The archaeology of slavery in North America. *Annual Review of Anthropology* **24**: 119–140

South S A 1977 *Method and Theory in Historical Archaeology*. Academic Press, New York

M. Hall

Historical Change and Human Development

Human development refers to patterns of change and stability that characterize a person from birth to death. Many theories maintain that development is highly plastic, which means that it is not predetermined in its processes or outcomes. In fact, human development is highly responsive to social and historical change. This article examines how behavioral scientists interrelate history and biography: how the particularities of historical circumstance lead to unique patterns of development.

Historical data are also used to test general theories of the family, life course, and life span. This approach does not deny that historical forces operate in unique ways to shape development, but it does consider the generalizability of theories across historical settings. Accordingly, this article then examines how historical data are used to test general theories of behavior.

1. The Historical Time Frames of Human Development

Human development reflects mechanisms that operate in three historical time frames, each of which is studied from different disciplinary standpoints. First, human behavior reflects evolutionary mechanisms that operate across historical epochs covering many

millions of years. These mechanisms are studied by biologists, anthropologists, and evolutionary psychologists who share a set of Darwinian assumptions. Second, individual development reflects the unique influences of historical periods that are comparatively limited in their duration. For example, social historians and students of psychohistory examine how the organization and ideology of Medieval society led to specific biographical patterns. Finally, life-span psychologists and life course sociologists examine connections between human lives and the 'historical present,' which encompasses the historically specific circumstances of a person's life (e.g., the Great Depression or the 1970s).

1.1 Historical Epochs: Phylogeny, Ethology, and Evolution

One of the most influential (albeit erroneous) contributions to the study of human development has been the principle of recapitulation, which emerged in the late eighteenth and early nineteenth centuries. Although many nineteenth century biologists subscribed to recapitulation, its greatest proponent was Ernst Haeckel, whose biogenetic law maintained that ontogeny (the life history of the person) recapitulates phylogeny (the evolutionary history of the lineage leading to the person). That is, the morphological development of each person quickly repeats—indeed, is caused by—the adult forms of the evolutionary history that preceded it. For example, Haeckel argued that because 'the human egg is a simple cell, we may at once infer … a unicellular ancestor of the human race resembling an Amoeba.' Similarly a later embryonic form in humans points to a 'primitive worm-like ancestral form which is related to the sea squirts or Ascidians …' (quoted in Gottlieb 1992, pp. 29–30). Because this recapitulation takes place in the embryonic and juvenile stages of development, the further evolution of a species depends on terminal addition, the appending of new characteristics to 'the end of an unaltered ancestral ontogeny'(Gould 1977, p. 74).

The biogenetic law fails principally because there is no known mechanism that links the evolution of a lineage with the development of an individual. And yet this law is important for two reasons. First, even though it was commonly rejected in biological circles in the early twentieth century, various forms of psychological recapitulation continued to influence many developmentalists, including Baldwin, Freud, Hall, Piaget, and Preyer (Gould 1977). For example, the biogenetic law (as it addressed both morphological and behavioral development among its proponents) and Larmarkian biology influenced Freud in pervasive ways. Thus, Freud held that the psychoanalytic stages of development (e.g., the oral and anal stages) were recapitulations of the phylogenetic sexual stages of our remote ancestors (Sulloway 1979).

This influence also is illustrated clearly in *Totem and Taboo* (Freud [1913] 1962) and such later works as *Moses and Monotheism* ([1939] 1967). In these works, Freud elaborates his theory of psychosexual recapitulation—this time focused on the Oedipus complex—involving an early trauma, a defensive reaction, a latency period, and then the outbreak of a neurosis. This theory, he believed, characterized the history of the human species, the history of the Judeo-Christian tradition, and the life history of the individual.

In the history of the human species, following the Ice Age a horde of brothers killed and consumed their father because of their sexual desire for his wives. Later, the brothers atoned for this act by deifying the slain father, now represented by a totem, and by observing prohibitions against parricide and incest. This sequence of acts was probably repeated over the millennia and was traumatic; consequently, by way of Lamarckian inheritance (which posits that acquired characteristics can be transmitted to successive generations), the feelings that surrounded these events were passed to later generations and now reside deep in the memory of all people. That is, according to Freud, the Oedipus complex is an inherited psychological structure rooted in phylogeny.

The life history of the individual repeats this sequence. Childhood sexual traumas (either real or conjured from phylogenetic fantasies that often involve the Oedipus complex) are repressed and forgotten during a period of latency. Additional traumas or disappointments in adulthood can then lead to high levels of distress and often to a neurosis. Thus, mental disturbances in adulthood need not trace back to a trauma that is consciously experienced by people during their own childhood; all humans are born with inherited traumas that reside deep in the unconscious and that can compromise pathways of adult development. Unfortunately, Freud's theoretical edifice is flawed because of its strong and erroneous biological assumptions. Referring to Lamarkian biology, Freud himself remarked in *Moses*, 'If things are different, then we are unable to advance one step further on our way, either in psychoanalysis or in mass psychology.' (Freud [1939] 1967 p. 128, Sulloway 1979, see also *Freud, Sigmund (1856–1939)*).

Second, although the biogenetic law was a flawed attempt to link the evolutionary time frame with individual development, the link is still being pursued at the beginning of the twenty-first century, albeit in very different ways, including, for example, through ethology and life history theory. Ethology is the study of animal behavior, its causes, development, functions, and, importantly, its evolution. Developmentalists have drawn on ethological research to provide principles whose applicability to humans can then be assessed (Hinde 1982). This approach is vividly illustrated by Bowlby's attachment theory, which relies heavily on ethological research (e.g., animal im-

printing). In *Attachment and Loss* (1969–80), Bowlby explains that the attachment behavior of young children, which includes variations in proximity seeking to the primary caregiver, is an instinctual behavior observed in many birds and mammals because it facilitates survival (but see Belsky 1999). According to Bowlby, the attachment relationship in humans between child and primary caregiver leads to an 'internal working model' of felt security, which in turn has pervasive implications for one's later relationships and psychological development (see *Attachment Theory: Psychological*).

Life-history theory refers to evolutionary explanations for the specific timing of biological events during the life course. Life-history theory assumes that natural selection has influenced when and how energy is allocated to growth, maintenance of the self, and reproduction so as to maximize reproductive output. Hence, life-history theory attempts to explain, for example the age of maturation, fertility rates and the spacing of children, menopause, senescence, and the length of the life span. For example, 'disposable soma theory' provides an evolutionary explanation for why people senesce (i.e., experience a decrease in bodily functions with increasing age). According to this theory, death from unavoidable causes (e.g., accidents, predation) in the course of evolution was inevitable and thus energy was diverted from attempts to live for exceedingly long periods to reproductive efforts earlier in the life span. Indeed, empirical evidence supports the proposition that the rate of senescence is directly related to the probability of death due to difficult-to-avoid causes. One of the major challenges to life-history theory, however, is its application beyond nonhuman species and primitive peoples. Many of its expectations refer to environments that were typical of the evolutionary past and so require further elaboration or restriction when applied to modern societies, especially societies that have experienced the demographic transition (Hill and Kaplan 1999; see also *Education and Learning: Lifespan Perspectives*; *Psychological Development: Ethological and Evolutionary Approaches*).

1.2 Historical Periods: Social History and Psychohistory

Compared to evolutionary epochs, historical periods are more limited in both time and place. Historical periods typically refer to intervals in history that represent relatively cohesive and distinct patterns of material living conditions, ideologies, norms, social organizations, and institutions.

Social history is often concerned with how these factors jointly shape biography, especially factors such as childhood socialization, school, work, and family. For example, many historians have studied how changing social circumstances have altered the meaning of 'childhood,' 'adolescence,' and indeed the functions of the family before, during, and after the Early Modern Period, which is often defined in terms of the Industrial Revolution. Ariès's path-setting *Centuries of Childhood* (1962) provided the impetus for many contemporary inquiries. According to Ariès, in medieval society (and much later among the lower classes), childhood and adolescence did not exist; rather, 'as soon as the child could live without the constant solicitude of his mother, his nanny or his cradle-rocker, he belonged to adult society' (p. 128).

Beginning in the sixteenth century, parents began to educate their children before they could enter the adult world. As educational systems developed, youth were taught in age-segregated settings according to curricula that were increasingly less concerned with vocational training than the medieval guilds had been. With this prolongation of education and segregation from the adult world, childhood and adolescence emerged as distinct age-graded identities. While *Centuries of Childhood* has been criticized in matters large and small, its basic analytic orientation—one that emphasizes the interconnectedness of work, school, family, and the phases of the life course—remains a point of departure for historians interested in human development (e.g., Kett 1977).

Psychohistory emphasizes the psychological development of the individual. Its primary focus is on the complex relationships between historically unique processes and the psychological functioning of individuals or, in some cases, mass psychology. Numerous biographies that utilize psychological data or theories have followed Freud's early attempts. Most draw on some form of psychoanalytic theory, although there are notable exceptions (e.g., B. F. Skinner's application of behavioral theory to his autobiography).

Erikson's *Young Man Luther* ([1958] 1993) is often acknowledged as one of the great contributions to psychohistory. He focuses on the adolescent identity crisis because, according to his general theory of the life span, the resolution of this crisis ideally leads to the rejuvenation of the person and, potentially, to the rejuvenation of society. Erikson claimed that Luther's crisis consisted of an intense Oedipus complex that prompted Luther to question the moral authority invested in father figures, be they his own father or the Pope. Both his father and the Catholic Church attempted to control Luther with a 'negative external conscience' since both had contributed significantly to his superego, which emphasized his inherently flawed nature: In the case of his father, Luther did not find support in his choice of occupations, and, in the case of the Church, he felt sinful and unworthy of his priesthood. According to Erikson, Luther's uniquely strong sense of conscientiousness led to an affirmation of his inner voice as a valid instrument of both self-determination and religious belief, thereby justifying salvation based on faith rather than on obedience to an ecclesiastical (paternal) authority. 'In laying the

foundation for a religiosity for the adult man,' Erikson argued, Luther 'displaced the attributes of his own hard-won adulthood; his renaissance of faith portrays a vigorous recovery of his own ego-initiative.' (Erikson [1958] 1993 p. 206).

Erikson's biographical approach is unusual for its reciprocal view of history and human development: historical forces clearly shaped Luther and yet, in resolving his identity crisis, Luther in turn shaped history. There are, however, several basic difficulties common to analyses such as Erikson's, including, for example, a lack of pertinent biographical information about the subject's childhood, difficulty establishing causal links between early and later events in a person's life, and identifying why one person—and not another among many people in similar circumstances—becomes a historically great figure (Runyan 1982; see also *Erikson, Erik Homburger (1902–94)*).

1.3 The 'Historical Present': The Life Span and Life Course

Social historians and psychohistorians emphasize common elements of context that people share in a historical period. Yet within historical periods, unique trends and events occur that differentiate age groups and even individuals within age groups. Life-span psychology and life course sociology represent two approaches to the historical present and development.

Life-span developmental psychology refers to a series of interrelated concepts and propositions that emphasize the complexity of development in historically changing contexts. According to this perspective, age-graded development is markedly influenced by existing sociocultural conditions and by social change. Indeed, development can be viewed as the interactions among age-graded, historical, and nonnormative (or idiosyncratic) influences (see Baltes et al. 1998). The life-span paradigm has been especially influential in charting connections between age trends in psychological phenomenon and historical time.

For example, Nesselroade and Baltes (1974) examined the relative contributions of age and history to the development of adolescent personality and ability between 1970 and 1972 in the USA. They reasoned that because youth are highly influenced by social circumstance and because personality reflects situational factors, trends in variables such as adolescent ego strength, independence, and ability may be explained by (a) unique historical experiences shared by young people born in the same year (a birth cohort effect) and (b) historical experiences that exert a similar influence across many age groups (a period effect). To examine these possibilities, they studied a sample of adolescents of diverse ages (ranging from 13 to 18) and, within age groups, adolescents from different birth cohorts (e.g., 15 year olds born in 1955, 1956, and 1957).

Empirical analyses reveal significant declines in self-rated ego strength and achievement and a significant increase in independence between 1970 and 1972, irrespective of age, which suggests that the development of adolescent personality and ability are less dependent on age than on historical setting. According to Nesselroade and Baltes, the observed effects may reflect several historical features that characterized the early 1970s and affected many adolescents (i.e., a period effect), including the tendency of youth to value moral, social, and political questioning (especially of the Vietnam war) rather than to value cognitive achievement, and the widespread decline in confidence in public and educational leadership. This research and related studies suggest that psychological development does not proceed as 'an orderly, sequential unfolding of universal, stage-like behavior patterns' (Nesselroade and Baltes 1974, p. 57), but rather is shaped in significant ways by historically based experiences (see also Schaie 1996, for notable cohort and period effects on adult cognitive development).

The life course refers to the sequence of age-graded roles that typically characterize development. These roles (e.g., student, worker, mother) bring with them expectations, opportunities, and limitations, all of which shape the contours of people's lives. This sequence of roles, their timing, and content are sensitive to immediate historical circumstances. Thus, the study of the life course typically begins with social changes and traces the implications of these changes to the developing person (see *Life Course in History*).

Drawing on data from a wide variety of sources, for example, Modell (1989) shows that the 'normative' sequence of roles often associated with the transition to adulthood—the completion of school, entry into work, marriage, and becoming a parent—became less common through the twentieth century. Modell juxtaposed his quantitative findings with the qualitative study of US cultural practices (dating, for example) to suggest that as youth were freed from the constraints of family and as they acquired financial resources of their own by wage labor, they increasingly became 'shapers of youth,' making choices about when and how they would enter adulthood. The aggregate of these choices creates a 'demonstration effect' for subsequent cohorts, until the next demonstration effect comes along to replace it.

The study of the life course owes much to Elder's watershed *Children of the Great Depression* ([1974] 1999), which integrated a focus on historical and social change with the study of the family and the social and psychological development of the person. A key analytic tool in Elder's research is the birth cohort (i.e., the year a person is born), which situates the person's development in the context of historical events. For example, Elder compared the lives of older children (ages 11–13 in the depths of the Depression) and younger children (ages 3–5 in the depths of the Depression) during and after this event. He further

distinguished between families who did and did not experience serious financial burdens because of economic downturn. Although the Depression led to similar changes in the deprived families of both cohorts—involving the division of labor, altered family roles, and social strains—its developmental effects varied by the age of the children. For example, children in the younger cohort from deprived families were more vulnerable to family strains and indeed were less likely to be hopeful, self-directed, or confident about their future than were their counterparts in families not adversely affected by the Depression. In contrast, the older boys from deprived families were likely to take active steps to help their families and themselves; these experiences in turn enhanced their social independence, which had lasting effects on their lives (see Elder et al. 1993).

Elder's research established a general analytic model that interrelates economic downturn and individual development. The basic model posits that economic stressors lead to marital tensions, which in turn have negative implications for parenting and children's well-being. Although the model was developed in the context of the Great Depression, it has proven useful in understanding other economic downturns (e.g., the Farm Crisis of the 1980s). Elder's research also established a series of analytic principles for the study of change and human development across historical circumstances. For example, the life-stage principle asserts that the influence of a historical event depends on the age at which the person experiences it. This principle is illustrated by the differences observed among the older and younger children of the Great Depression, but it has also been used, for example, to study the effects of military service, German reunification, and China's Cultural Revolution on people's lives.

2. *Historical Data and Theories of Human Development*

While the theories and empirical research discussed in the previous section emphasize how particular historical forces lead to specific patterns of development, some scientists use historical data to test general ideas about human behavior. This approach has yielded valuable insights, for example, about the factors that affect physical height (Fogel et al. 1983), human creativity (Simonton 1999), and the timing of puberty (Mitterauer 1993).

One of the landmark contributions in the use of historical data to test general psychological principles is Sulloway's *Born to Rebel* (1996). Sulloway developed an elaborate theory that interrelates birth order, family dynamics, and aspects of personality. For example, his theory predicts that later-borns are more likely to subscribe to radical ideas than first-borns. He tested this and other expectations by drawing on a comprehensive database of 6,566 participants in famous scientific, religious, and political movements in history. A noteworthy feature of this research is the use of experts to rate the participants in these events in terms of their personality and social attitudes.

Drawing on these unique data, Sulloway found support for his theory. For example, the Copernican revolution and Darwin's theory of evolution by natural selection are both rated as radical scientific theories and in fact the leaders of both were likely to be later-borns, including Copernicus and Darwin themselves. (Alfred Russell Wallace, sometimes considered the co-originator of evolution by natural selection, was also a later-born.) In contrast, scientific breakthroughs of a technical nature are likely to be achieved by first-borns as found, for example, in the cases of Watson and Crick, the cofounders of the structure of DNA (as well as in the case of their chief rival, Linus Pauling, also a first-born). Sulloway's research is exemplary for its attempt to document person-by-situation interactions involving the historical context, such as birth order and the social implications of a scientific revolution.

3. *Integrating History and Development*

The preceding discussion suggests two directives for future research. First, accounts of human development will ideally incorporate explanatory factors from all three historical time frames. The case of attachment illustrates the potential of this approach. In terms of epochs, ethological research has contributed to an understanding of the instinctual basis of attachment behaviors. In terms of historical periods, the parent–child relationship is heavily conditioned by social institutions such as education, work, and family. And, in the historical present, parental loss or changes in the parent–child relationship are often precipitated by immediate circumstances of historical import. Indeed, attachment theory was inspired in part by the countless children left without parents after World War II.

Similarly, Sulloway's work on sibling strategies and family dynamics unites evolutionary theories about sibling competition with a consideration of historical practices (such as primogeniture), and finally with statistical patterns associated with individual development.

Second, interdisciplinary collaborations will benefit from the integration of diverse methods of inquiry. Many developmentalists rely on statistical analyses of data that were recently collected; these accounts typically emphasize how contextual and personal factors predict behavior. In contrast, many historians rely on the qualitative study of archival materials; these accounts typically emphasize the complex combinations of structural contingencies and cultural

meanings that shape biography. If developmentalists often ignore the interactionist, cultural stance of historians, historians are often skeptical as to whether their interests can be measured and subjected to statistical analyses. Yet a growing number of studies illustrate the value of integrating diverse methods, including, for example, Modell's (1989) study of the transition to adulthood and Elder's study of the Great Depression (Elder et al. 1993; for another example see Hareven 1982).

In short, as research on attachment and family dynamics illustrates, all three historical time frames come together to shape one's life, and thus render the study of human development an interdisciplinary task that encompasses diverse historical influences. And as research on the transition to adulthood and the Great Depression illustrates, the quantitative and qualitative methods of the behavioral sciences can be used in complementary ways to uncover these many historical influences as they shape development and the biography.

See also: Human Development, Bioecological Theory of; Life Course in History; Life Course: Sociological Aspects; Lifespan Development, Theory of

Bibliography

Ariès P 1962 *Centuries of Childhood: A Social History of Family Life*. Vintage Books, New York

Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Damon W (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. 1, pp. 1029–43

Belsky J 1999 Modern evolutionary theory and patterns of attachment. In: Cassidy J, Shaver P R (eds.) *Handbook of Attachment*. Guilford, New York pp. 141–61

Bowlby J 1969–80 *Attachment and Loss*, 3 vols. Basic Books, New York

Elder G H Jr [1974] 1999 *Children of the Great Depression: Social Change in Life Experience*. University of Chicago Press, Chicago

Elder G H Jr, Modell J, Parke R (eds.) 1993 *Children in Time and Place: Developmental and Historical Insights*. Cambridge University Press, New York

Erikson E [1958] 1993 *Young Man Luther: A Study in Psychoanalysis and History*. W W Norton, New York

Fogel R W, Engerman S L, Floud R, Friedman G, Margo R A, Sokoloff K, Stekel R H, Trussel T J, Villaflor G, Wachter K W 1983 Secular changes in American and British stature and nutrition. *Journal of Interdisciplinary History* **14**: 445–81

Freud S [1913] 1962 *Totem and Taboo*. Norton, New York

Freud S [1939] 1967 *Moses and Monotheism*. Vintage, New York

Gottlieb G 1992 *Individual Development and Evolution: The Genesis of Novel Behavior*. Oxford University Press, New York

Gould S J 1977 *Ontogeny and Phylogeny*. Belkaapp Press of Harvard University Press, Cambridge, MA

Hareven T 1982 *Family Time and Industrial Time: The Relationship Between the Family and Work in a New England Industrial Town*. Cambridge University Press, Cambridge, UK

Hill K, Kaplan H 1999 Life history traits in humans: Theory and empirical studies. *Annual Review of Anthropology* **28**: 397–430

Hinde R A 1982 *Ethology: Its Nature and Relations with Other Sciences*. Oxford University Press, New York

Kett J F 1977 *Rites of Passage: Adolescence in American, 1790 to the Present*. Basic Books, New York

Mitterauer M 1993 *A History of Youth*. Blackwell, Oxford, UK

Modell J 1989 *Into One's Own: From Youth to Adulthood in the United States, 1920–1975*. University of California Press, Berkeley, CA

Nesselroade J R, Baltes P B 1974 Adolescent personality development and historical change: 1970–1972. *Monographs of the Society for Research in Child Development* **154** (vol. 39) no. 1

Runyan W M 1982 *Life Histories and Psychobiography: Explorations in Theory and Method*. Oxford University Press, New York

Schaie K W 1996 *Intellectual Development in Adulthood*. Cambridge University Press, New York

Simonton D K 1999 *Origins of Genius: Darwinian Perspectives on Creativity*. Oxford University Press, New York

Sulloway F J 1979 *Freud, Biologist of the Mind: Beyond the Psychoanalytic Legend*. Basic Books, New York

Sulloway F J 1996 *Born to Rebel: Birth Order, Family Dynamics, and Creative Lives*. Pantheon Books, New York

M. J. Shanahan

Historical Demography

Some of the central questions historical demographers are concerned with may be traced back to the writings of Thomas Robert Malthus. His influential thesis was that human beings are able to reproduce themselves 'in a geometrical ratio' while the capacity of the earth to produce subsistence increases 'only in an arithmetical ratio.' In the long run, therefore, population increase of the past must have been kept controlled either by the 'positive check' of higher mortality or by the 'preventive check' of abstinence or late marriage. However, it was not until the post-World War II period that historical demography was born as a separate discipline in historical science: in the 1950s concepts and techniques of modern demography were applied by French scholars to historical material. Its progress and achievements since then have changed our perceptions on the Malthusian and other questions.

1. A Social Science History

Historical demography is a social science history. It is distinguished from other branches of history by its explicit use of concepts and tools derived from a

discipline in social sciences, i.e., demography in this case. Its goal is to establish historical facts by applying such conceptual frameworks to historical sources, and its methodological novelty lies in quantification. The research results of historical demography as such, have often questioned the conventional pictures of past populations, thereby throwing doubts about assumptions many theories rest upon. The position of historical demography in social and behavioral sciences thus resembles that of New Economic History, but it should be realized that there are a few notable differences in emphasis between these two social science history fields. While the application of theoretically constructed econometric models is given prime importance in New Economic History, emphasis in historical demography has always been on measurement. Attempts to formulate models to explain historical facts have never been put aside, but more effort has been directed to the development of finer measures and the identification of patterns, temporal and cross-cultural, by using those quantitative measures.

2. *Sources, Methods, and Impacts*

The innovation in methodology often goes with the discovery of new data sources. In historical demography too, its birth was closely tied with the French scholars' application of family reconstitution method, a form of nominal linkage, to registers of baptism, burial, and marriage kept in Christian parish churches. Parish register demography, as one may call this type of research, is an attempt to reconstruct a demographic profile of the past based on a set of family histories reconstituted according to the rules set out by the demographer Louis Henry who turned to historical material to identify what he called 'natural fertility,' that is, fertility before the deliberate effort to reduce family size came to be practised.

Information that historical demography started supplying since then is a kind that they had always thought would only be available from modern vital records or sample surveys. Thanks to laborious works by historical demographers, demographers are now able to describe past populations in much the same vocabulary as for the present. However, there are a few drawbacks. In this kind of research, analysis ought to be restricted to persons whose burial dates are known, or who are known to have survived until their reproductive period was over. Since more people tended to leave the community with advancing age, it is very difficult to compute death probabilities across all the age groups, although a large enough sample, such as the one constructed by E. A. Wrigley and his associates at the Cambridge Group for the History of Population and Social Structure, does enable researchers to estimate age-specific patterns of mortality even from family reconstitution (Wrigley et al. 1997).

Furthermore, demographic research based on historical micro data has enabled the researchers to work with 'cohort' measures. Most demographic concepts, such as life table and marital fertility, are designed to portray the life events of a group of people who were born or married in the same period (termed 'cohort'), but usually it is not possible for present-day demographers to single-out people born in a particular period and to trace them to the moment they die. Instead of using cohort measures, they have to rely on 'period' rates to account for demographic events in a specified period, in which the behavior of one cohort is bound to be mixed with that of other cohorts. In this sense, historical micro data have given demographers an opportunity to work with methodologically ideal populations.

For historians, on the other hand, historical demography has offered them a different set of merits. First, population becomes tractable as 'population change' and can now be decomposed into demographic components such as births, deaths, and migrations. This in turn enables them to relate those demographic phenomena to changes in economic circumstances. For example, one of the results of the earliest family reconstitution studies in France showed us how changes in grain prices were correlated with those in births and deaths. Similarly, one may examine interrelationships between time series of demographic rates and real wages in detail.

Another impact historical demography has had is being felt in social history. Since parish register demography reconstructs families and their life events that occurred in familial contexts, it has offered social historians invariable insights into the demographic domain of family life. Marriage, for example, is no longer a subject tied only with legal and ecclesiastical concerns. Who married whom, at what age, with what age gap between spouses—all these questions can now be answered. In addition, information about prenuptial pregnancy, gender differentials in infant and child mortality, and other related topics, can shed light on the intimacy of family life in the past.

Also related, but somewhat different approaches, have been adopted by another group of social historians. Statistical studies of the family and household, launched by Peter Laslett at the Cambridge Group in the 1960s, have been conducted in close association with research in parish register demography (Laslett 1983, Laslett and Wall 1972). The research along this line strongly suggests that types of family system, whether or not a new household was formed on marriage, whether the presence of live-in servants in the household was frequent or not, the mean age at marriage, and the frequency of celibacy, were all closely interrelated.

These links with social history have led to the burgeoning of family and household studies in history. Topics studied thus far include servanthood, gender inequality within the household, care for the elderly,

widows and orphans, and the succession of headship and property. There are also issues that have enabled historians to get further into the inner domain of the individual. Abortion, infanticide, child neglect, and other means of family limitation embedded in traditional reproductive culture, are such issues.

However, source material being used in historical demography and its related research fields is not confined to the parish registers. In fact, there is a serious drawback in the use of parish registers as a source material. Since only people who registered baptism, marriage, and burial in a single community were recorded, it is impossible to reconstruct families who eventually left the community. Also impossible is to know how many people existed in that community at a particular point in time. All this implies that aggregative studies are extremely difficult to conduct, implying that the size of a settlement population, its birth and death rates, and in- and out-migration ratios can hardly be computed. Also difficult to conduct, for the reasons already mentioned, are mortality studies.

It is not surprising, therefore, that in many west European countries other micro data are being used. Especially useful are listings of households and population, most of which were one-off 'censuses' taken for various purposes. In other cultural regions, there are places where census-type records were regularly taken. In traditional Japan, for example, registers of village population were compiled by officials every year while in some parts of imperial China similar records were taken triennially. Such repetitive census-type data are useful not only for household analysis, but also for demographic studies since probabilities of entry or disappearance can be calculated at a regular interval, which, as monographs such as Smith (1977) and Lee and Campbell (1997) illustrate, allow researchers to draw demographic interpretations. One of the few shortcomings of such population registers is that it is very difficult to estimate the infant mortality rate since those born after the previous compilation, but died before the next, were not recorded at all.

3. *Two Dimensions: Micro and Macro*

Two broad areas of issues have been occupying a central place in the historiography of historical demography. One is the Malthusian models of positive and preventive checks, and the other the thesis of demographic transition. The former concerns pre-modern settings. The latter is about the transition processes from the pre-modern to the modern age, as Demeny (1972) put it: 'In traditional societies, fertility and mortality are high. In modern societies fertility and mortality are low. In between there is the demographic transition. In both fields of research, however, it is useful to distinguish micro-behavioral from macro-aggregative dimensions.

3.1 *Micro Demography*

As noted in Sect. 2, parish register demography was born as a fertility-oriented research field. Malthus's notion of positive check posits that overpopulation tended to be checked by a rise in mortality, which postulates that reproductive behavior of couples was governed by the invariable 'passion between the sexes.' The classic theory of demographic transition assumes that pre-transition populations were under 'strong pressure on their members to reproduce,' and hence that marital fertility was uncontrolled and its level close to the physiological maximum. Both Louis Henry, who coined the term 'natural fertility,' and Ansley Coale, who developed a technique to measure a deviation from the natural, uncontrolled fertility pattern in the course of his own project on the European fertility decline, assumed that if the couple has no target size of family and is not concerned with the existing number of children (termed 'parity'), marital fertility is not controlled and hence its level must be invariably high. In other words, the decision made by a couple to end childbearing (termed 'stopping') is assumed to mark the onset of fertility limitation. Much effort, therefore, has been made to measure fertility levels and to identify patterns by examining age, parity, birth interval, and other factors associated with individual couples' reproductive behavior (see Coale and Watkins 1986).

However, according to a large body of empirical studies so far conducted the original assumptions about reproductive behavior prior to the demographic transition are no longer tenable (see Cleland and Wilson 1987, Friedlander et al. 1999). First, pre-transition of marital fertility was far more variable than previously imagined. Few populations exhibited levels comparable to that exemplified by the Hutterites, a small group of anabaptist settlers in north America, whose total marital fertility rate (TMFR, defined as the number of children a married woman would have within her reproductive period) stood as high as 11. In most cases, in fact, it was kept at moderate levels with their TMFR ranging from 6 to 9.

Second, it is confirmed that stopping behavior was never common among pre-modern populations. It should be noted, however, that some elite subpopulations, such as Catholic aristocrats in France and Italy and bourgeois families in Calvinist Geneva, exhibited a distinct stopping behavior, so that they can be regarded as forerunners of modern family limitation. Third, moreover, the couple's behavior to 'space' births at particular intervals, as contrasted with stopping, seems to have been more important in accounting for the moderate levels of marital fertility in traditional times. Unlike stopping, birth spacing could have been achieved either through deliberate means such as contraception, or through unconscious actions such as breastfeeding, and many research

results suggest that the role played by differences in feeding practices of the infant was unmistakable. Fourth, it is now widely recognized that in many instances marital fertility was rising, rather than falling, in the period immediately prior to the onset of modern fertility decline. This was the case in the UK and other European populations as well as in twentieth-century developing countries (Knodel 1988, Wrigley 1998). One of the factors suggested as having affected this rising tendency is a reduction in the level of breastfeeding, which was often associated with early phases of urbanization. Another is a decline in the incidence of stillbirths, which has recently been suggested for eighteenth-century England. The latter is a promising area of micro-demographic study, because even from family reconstitution, it is possible to investigate into the interplay between fertility and mortality measures at very early phases of life.

The conclusion that pre-transition fertility was not uncontrolled, be it consciously or unconsciously, demolishes the conventional divide between the modern and the pre-modern regime. Furthermore, findings from micro-demographic data analysis have made it difficult to see the demographic transition in a dichotomous framework. To what extent the transition was a behavioral revolution, still remains debatable.

3.2 Macro Demography

The concept 'reproduction' can also be applied to a population as a whole. It means the way in which the population replaces itself through natural processes of birth and death. The overall level of such replacement may be indicated by the total number of children that an average woman will have during her childbearing period. This is called the total fertility rate (TFR). Since it does not take mortality into account, the population's natural process should be measured in terms of both fertility and mortality such as life expectancy at birth (e_0). Thus one may describe historical populations by referring to the two, TFR and e_0. According to Livi-Bacci (1997), for example, Paleolithic populations and the Russians in 1879 happened to roughly exhibit the same level of TFR, a little lower than 6.5 children per woman. But mortality was different. The life expectancy is estimated to have been at below 20 for the Paleolithic, whereas that for Russia was a little below 30. As a result, the difference in the rate of natural increase was considerable. It was only marginally above zero in the Paleolithic case whereas, more than 10,000 years later, the Russian population was growing at somewhat over 1 percent per annum.

This, however, should not be taken to imply that the driving force in population history was always mortality decline, with fertility remaining high throughout the long, long period before the modern age. Indeed, over time both fertility and mortality may have exhibited substantial volatility. Neolithic settlers are supposed to have been characterized by fertility and mortality higher than Paleolithic hunters and gatherers, reflecting a combination of higher population density and lower mobility in the farming population. A much recent case in point is England from 1550 to 1800. Wrigley and Schofield (1981) have documented that from the late sixteenth century, when England's rate of natural increase stood at below 1 percent per annum, the country's TFR and e_0 simultaneously declined towards the late seventeenth century. From that period to the second half of the eighteenth century the two measures came back to a position of higher fertility and higher life expectancy again. This time, however, the speed of rise in fertility exceeded that in life expectancy, so that the country's population grew faster than 1 percent.

Since information from family reconstitution studies does not add up to macro, aggregate measures, complex and sophisticated estimation techniques, such as back projection, are needed to reconstruct the history of a past population at large. Judging from a few well-grounded back projection works, such as Wrigley and Schofield (1981) for England from the sixteenth to the nineteenth century and Weir (1994) and Bonneuil (1997) for eighteenth- and nineteenth-century France, first, there is now agreement that the role of nuptiality should be stressed in accounting for how overall fertility changed in pre-transition western Europe. The age at marriage was instrumental in adjusting the actual level of fertility in relation to changing economic circumstances. This what the classic Malthusian model of the preventive check predicted, thus suggesting that such a homeostatic mechanism was at work in the long- to medium-run time frame of the European past.

Second, this does not imply that mortality crises were not important. Short-term fluctuations in mortality were frequent in various regions in Europe, most of which were subsistence crises occasioned by crop failure. The frequency of famines, therefore, could account for the long-run level of death rates, but it should also be noted that changing levels of 'exposure' to viruses and other micro-organs were equally—sometimes more—responsible in determining that level of mortality in the past. The supposed link between mortality and the general standard of material life was thus less direct than often imagined (see papers in Bengtsson and Saito 2000).

It is true that during the period of fertility decline, neither nuptiality, nor mortality, could offset the declining trend of marital fertility and, hence, the corresponding rise in population. This, however, should not be taken to imply that economic growth, which accompanied this demographic transition in most cases, changed the whole relationship between the demographic variables. Neither industrialization, nor urbanization, was the crucial determinant of the

onset of fertility decline. Even the modern decline of mortality was not in close association with an increase in the general standard of living.

Finally, mention must be made to the role of migration. The basic demographic equation tells us that a population will change according to the addition of births and immigrants, and the subtraction of deaths and emigrants. In other words, without information about in- and out-migration it is impossible to account for any macro change. Migration, therefore, is as important a topic as nuptiality, fertility, and mortality in historical demography.

Unfortunately, however, as historical sources rarely recorded actual population movements, not much has been done on this important topic. The historiography reveals that the volumes of in- and out-migration, distance, pull and push mechanisms are issues relatively better covered thus far. Much less known is the demographic impact of such migration flows. Were death probabilities in areas migrants headed for, different from those in their native communities? Was their marriage and reproductive behavior altered by their geographical mobility? Or was even their parents' demographic decision-making affected by the prospect that their children would eventually leave for a place of, presumably, better opportunities? It is only in cases where no effect of these was substantial, that we do not need to know anything other than the rate of net migration (immigrants minus emigrants), which tended to be rather small. More specifically, any interpretation of results from family reconstitution studies rests upon the assumption that 'leavers' would behave just as 'stayers' did. All the demographic measures derived from this method only relate to the population of stayers, i.e., those born locally who continued to live in the their native community. The mean age at marriage, for example, could have differed between these two populations. Although it turns out that the difference was minimal in the case of English family reconstitution, the possible effects of migration on marriage as well as the other demographic variables can only be determined empirically (Wrigley et al. 1997, Chap. 5).

One of the few findings we have in this regard is about rural-to-urban migration. In a traditional setting the urban population was distinguished by a greater exposure to diseases and, hence, higher mortality than in the countryside. Obviously, even in such a mortality-dominant urban demography, the marriage market may also have been affected by the size of in-migration as well as that of labor demand. Moreover, the rural sector should not be considered demographically homogenous. There are European case studies suggesting that 'proto-industrial' villages exhibited an earlier age at marriage and higher fertility rates than farming villages. To what extent this was universal is debatable. What is certain, however, is that the destination and volume of each migration stream between the rural and urban sectors could affect the overall rate of population change in a pre-industrial setting (de Vries 1984).

The picture becomes less clear for the period of modern fertility transition, which in many cases coincided with extremely large volumes of migration streams, both long- and short-distance movements. It is probably safe to suppose that rural-to-urban migration in this period also meant deterioration in the disease environment, while trans-oceanic migration to land-abundant settlements led to lower age at marriage and higher fertility. What is less clear is its influence on the population sending those migrants in the course of change. In the literature of demographic transition, migration, especially townward migration, tends to be regarded as having facilitated the decline of fertility. On the other hand, it is often noted that rural populations did not readily respond to changing economic and social circumstances by lowering their fertility. One suggested reason why it was still beneficial for peasant families to keep their fertility relatively high is the expanding prospects for the younger generation. When it was felt that they could emigrate and form their own households elsewhere, the older generation's immediate need for the limitation of fertility may have been reduced. In nineteenth-century Scotland, for example, people left when circumstances allowed, and the same set of factors accounting for the emigration forced stayers to postpone marriage. Once married, however, they were less pressurized to limit the family size (see Andersdon and Morse 1993, Friedlander et al. 1999).

4. The Demographic Landscape in Cross-Cultural Perspective

At a more generalized level, historical demographers have developed a notion of homeostatic demographic regime. Many societies in the past had an ability to maintain equilibrium between their resources and the demographic characteristics. It is the ability to restore the original state of equilibrium, even if an external shock forces the system away from that position. When the ways in which this homeostasis is achieved are similar for a group of populations, it is probably appropriate to call that particular mode of interrelationships between the demographic variables a 'demographic regime.' Given empirical evidence so far accumulated, therefore, we may see the traditional west European populations as characterized by a particular mode of demographic regime. In this regime, both fertility and mortality levels were relatively low, and the key to its adjusting mechanism was nuptiality acting to ensure the system to restore equilibrium.

Turning to the nuptiality patterns, John Hajnal in his influential articles has shown that in traditional Europe, west of the line from St Petersburg to Trieste,

the mean age at marriage was relatively high and celibacy not negligible, and that all this was closely tied with the nuclear family system and its associated principle of neo-localism (Hajnal 1965, 1983). In contrast, non-European countries exhibited comparatively low ages at marriage and very low proportions never married, suggesting that different demographic regimes were operating in other parts of the world.

Take China and Japan. According to in-depth regional studies in traditional periods, the level of life expectancy at birth (e_0) in northern China was 32, while it stood at 39 in central Japan (see Lee and Wang 1999 for China, Bideau et al. 1997, Chap. 8 and, Liu et al. 2001, Chap. 7 for Japan). Compared with the western European average, China's mortality was on the high side and Japan's was on the low side. On the other hand, TMFR in both countries fell between 6 and 6.5, unmistakably on the low side of the European standard.

The Chinese and Japanese results raise some interesting questions. First, it is clear that the much lower level of marital fertility did not reflect stopping behavior, but spacing, thus confirming much of what the work on Europe has suggested. What is not certain at this stage is whether the observed long birth interval was a mere reflection of low fecundity, or a product of deliberate actions. While the factors accounting for the former vary from breastfeeding to malnutrition, the latter raises further questions: to what extent abortion, infanticide, and child neglect were practised, which of the three was most responsible, and whether it was motivated by despair or guided by the notion of 'planning,' however, rudimentary it would have been. Second, the nuptiality pattern was undoubtedly different from that of western Europe. There was no neo-local principle working. Nor was the age at marriage functioning so as to adjust overall fertility to economic conditions in Asia. Rather, nuptiality appears to have been more closely related to mortality. Indeed, the female mean age at marriage was unmistakably higher in Japan than in China, reflecting the aforementioned difference in e_0, although both countries seem to have shared the characteristics of universal marriage.

Third, however, if attention is turned to the status of females and family systems, we realize that there were some fundamental differences between China and Japan. For example, while no particular anomalies are observed from Japanese sex-specific life tables, the Chinese studies point out that there were persistent gender inequalities in mortality: the probability of girls dying between ages 1 and 5 was about 20 percent higher than that of boys, and such gender imbalance continued up to age 45. Clearly the status of girls and women was very different in traditional Chinese society. More specifically, it is the family system that differed. Predominant in Japan was the stem family system, under which the family household was vertically structured, and customs such as adoption, marriage, and remarriage were all for the family line to perpetuate. In China, on the other hand, what was operating was the rules of joint family. Under that system the household was laterally structured and the composition tended to be perennially complex.

To conclude, historical demography as it now stands, is both interdisciplinary and comparative and has raised new questions of its own, as it has evolved over time as well as cross-culturally.

See also: Demography: Twentieth-century History; Families and Households, Behavioral Demography of; Family and Kinship, History of; Family Systems in Europe; History and the Social Sciences; Life Course in History; Quantification in History; Social History

Bibliography

Andersdon M, Morse D J 1993 High fertility, high emigration, low nuptiality: Adjustment processes in Scotland's demographic experience, 1861–1914. Parts I and II. *Population Studies* **47**: 5–25, 319–43

Bengtsson T, Saito O (eds.) 2000 *Population and Economy: From Hunger to Modern Economic Growth*. Oxford University Press, Oxford, UK

Bideau A, Desjardins B, Perez-Brignoli H (eds.) 1997 *Infant and Child Mortality in the Past*. Oxford University Press, Oxford, UK

Bonneuil N 1997 *Transformation of the French Demographic Landscape 1806–1906*. Oxford University Press, Oxford, UK

Cleland J, Wilson C 1987 Demand theories of the fertility transition: An iconoclastic view. *Population Studies* **41**: 5–30

Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ

Demeny P 1972 Early fertility decline in Austria-Hungary: A lesson in demographic transition. In: Glass D V, Revelle R (eds.) *Population and Social Change*. Arnold, London

de Vries J 1984 *European Urbanization 1500–1800*. Methuen, London

Friedlander D, Okun B S, Segal S 1999 The demographic transition then and now: Processes, perspectives, and analyses. *Journal of Family History* **24**: 493–533

Hajnal J 1965 European marriage patterns in perspective. In: Glass D V, Eversley D E C (eds.) *Population in History*. Arnold, London

Hajnal J 1983 Two kinds of pre-industrial household formation system. In: Wall R, Robin J, Laslett P (eds.) *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK

Knodel J E 1988 *Demographic Behavior in the Past: A Study of Fourteen German Village Populations in the Eighteenth and Nineteenth Centuries*. Cambridge University Press, Cambridge, UK

Laslett P 1983 *The World We Have Lost—Further Explored*. Methuen, London

Laslett P, Wall R (eds.) 1972 *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK

Lee J Z, Campbell C D 1997 *Fate and Fortune in Rural China: Social Organization and Population Behavior in Lioning 1774–1873*. Cambridge University Press, Cambridge, UK

Lee J Z, Wang Feng 1999 *One Quarter of Humanity: Malthusian Mythology and Chinese Realities*. Harvard University Press, Cambridge, MA

Liu Ts'ui-jung, Lee J, Reher D, Saito O, Wang Feng (eds.) 2001 *Asian Population History*. Oxford University Press, Oxford, UK
Livi-Bacci M 1997 *A Concise History of World Population*, 2nd edn. Blackwell, Cambridge, UK
Smith T C 1977 *Nakahara: Family Farming and Population in a Japanese Village, 1717–1830*. Stanford University Press, Stanford, CA
Weir D R 1994 New estimates of nuptiality and marital fertility in France, 1740–1911. *Population Studies* **48**: 307–31
Wrigley E A 1998 Explaining the rise in marital fertility in England in the 'long' eighteenth century. *Economic History Review* 2nd ser. **51**: 435–64
Wrigley E A, Schofield R S 1981 *The Population History of England, 1541–1871: A Reconstruction*. Arnold, London
Wrigley E A, Davies R S, Oeppen J E, Schofield R S 1997 *English Population History from Family Reconstitution 1580–1837*. Cambridge University Press, Cambridge, UK

O. Saito

Historical Explanation, Theories of: Philosophical Aspects

There are three fairly distinct kinds of theory about the nature of historical explanations. First, there are reflections upon accounts people have given of basic processes of historical change which determine the general direction of history. These accounts are what J.-F. Lyotard calls 'the grand narratives' or 'meta-narratives' of history. In Western culture, these have generally been theories of historical progress and theories of cyclical change. Such theories have been proposed by theologians, political theorists, and philosophers, but seldom by professional historians, who are understandably wary of generalizations of such immense scope. In recent years both the truth and the point of such explanatory theories have been queried, and as a result such theories of world history are now seldom produced. Some historians still offer quite general explanatory interpretations of national and international history, but these too are often held to be partial, in both senses. For convenience historical explanations of this first kind will be referred to as 'general interpretations of historical change.' The closest thing to a theory about such explanations is the discussion by J.-F. Lyotard, so his writing will be used as a platform for some remarks about them.

Second, there are four carefully articulated theories of how historians should explain events. These are written by philosophers rather than historians. Some think historical explanations should be scientific, employing general laws. Some declare that history is not a scientific but a hermeneutic discipline looking for the meaning of events rather than their causes. A few favor comprehensive, many-layered structural history, noting how basic geographic and social structures influence historical change. Finally, there are those who, in a postmodern vein, see history as no more than a form of literature, with historical events acquiring meaning from their place in a discourse. The philosophers who promote these theories of historical explanation would say they are describing historical practice, but in fact they are much less interested in analyzing historical practice than in promoting a form of explanation which they think very important in understanding history. For that reason these will be called 'prescriptive theories of historical explanation.'

The third kind of theory of historical explanation is a theory which claims to be descriptive and analytic, focusing on historical practice to discover the variety of historical explanations historians actually produce, and their criteria of acceptability. The task of this theory is to provide clear analysis of typical historical explanations, pointing out the logical standards which it seems they aim to satisfy.

The article will briefly describe each of these theories of historical explanation, and make one or two comments upon them.

1. General Interpretations of Historical Change

In the culture of Western Europe, there have been several grand theories of historical change. Christianity, more particularly the Bible, brought with it a theory of progress in history, of God establishing his kingdom on earth, first through his prophets, then through the Israelites, then through Jesus, and finally through the Church, until Christ returns in triumph. St. Augustine presented an immensely influential version of this story in *The City of God*. With the progress in science achieved by the application of reason in the seventeenth and eighteenth centuries, a secular theory of historical progress was created by Kant, Condorcet, and others, who saw reason and the methods of science driving human progress towards ever-increasing, this-worldly happiness. Hegel and Marx described the process of human progress as dialectical, contradictions being overcome by new stages of development, either in the advancement of reason, freedom, and self-consciousness in the case of Hegel, or in autonomy and prosperity, in the case of Marx. The optimism implicit in all these theories of progress was largely shattered by the experiences of the twentieth century, of hideous world wars, of vast economic depression, of rampant infectious diseases, of tyranny, greed, and madness defeating enlightenment in so many countries. As Lyotard puts it: Auschwitz invalidates the Hegelian assumption that the real is rational; Budapest 1956 invalidates the Marxist claim that what is communist is proletarian (Lyotard 1987, p. 162).

A few writers have argued that progress has indeed continued throughout the twentieth century. With the defeat of fascism and the disintegration of Soviet

communism, F. Fukuyama sees the emergence of liberal-democratic capitalism as the final stage of historical development, 'The End of History.' He says that it best satisfies the two driving forces in history, the desire for material prosperity and the desire for personal recognition. Tyrannies had to go, as tyrants showed no respect for the common people. In democracies everyone has a say, or at least a vote. It seems obvious, however, that if the vote is not enough to ensure the recognition and respect people desire, then history has not ended after all.

Lyotard has three main objections to the grand theories of historical progress mentioned above. The first has already been stated: all are contradicted by the horrors of the twentieth century. Second, Lyotard objects that these theories have constrained historical writing in an unacceptable way. Under the influence of such theories, historians have written particular histories as instances of the theory of progress they most favored. For example, they depicted political revolutions as steps towards political freedom, and scientific revolutions as evidence of the growth of reason. In this way Lyotard finds metanarratives oppressive, and prefers that people should write their narrative stories as they like (Lyotard 1984).

Finally, Lyotard says that 'all great historical upheavals' such as the French Revolution 'are formless and figureless in historical human nature' (Lyotard 1987, p. 174). The enthusiasm such an upheaval generates is sublime, he says, and incapable of true representation, although we like to suppose that all who participated were motivated by a desire for greater freedom. All that historians can truly describe, according to Lyotard, are individual human actions and responses, 'little narratives' of particular events.

A few comments on Lyotard's ideas seem called for. First, he is right to deny inevitable progress toward a more rational, free, self-conscious, prosperous, and autonomous society. Theories of progress all underestimate the desire and ability of powerful people to exploit others for their own gain. More realistic, it seems, are some cyclical theories of historical change. Aristotle and Polybius produced such theories which were revived during the Renaissance in the writing of Louis Le Roy and Jean Bodin in France, and of Machiavelli and later Vico in Italy. In Aristotle they found a doctrine of the natural succession of political structures from monarchy to aristocracy to democracy and back to monarchy again, each stage established to resolve injustices which had arisen in the previous one. From Polybius they came to see cycles in the growth, flourishing, and decay of states, some attributing decay to states overextending themselves, others seeing it as a result of moral failure among its leaders. However, as Bodin said, the causes of a state's collapse can be varied, including military conquest by its enemies (1965, p. 59). There is no general theory of historical change which is true without exception. Cyclical theories were popularized in the twentieth century by Manuel Spengler's (1926–8) *The Decline of the West* and Toynbee's (1951–61) *A Study of History*.

Second, while metahistories may have been used to legitimate certain accounts of particular historical events, to the exclusion of others, some generalizations about the motivation for political revolutions and other events seem in order. The leaders of the French Revolution and the pamphlet writers who supported them did legitimate their attack on the monarchy by an appeal to reason, liberty, and human rights. Of course there were other, more self-interested motives at work as well, but that is not to deny the legitimation given to the revolutionary movement by these ideas. Rather than dismiss generalizations about the motives for revolutions and other large-scale events as all mistaken, historians should take care to qualify them so that they are fairly accurate. Generalizations are useful ways of summing up individual stories and describing social movements.

Finally, one of the chief objections to metahistorical explanations of historical change is something that Lyotard does not mention. They frequently assume that a whole series of events were all justified by the same ideals, when in fact this is false. For example, many political events and revolutions have been justified in the name of liberty and human rights, but the ideas of liberty and human rights were not always the same in every case. The rights and liberties sought in the English Revolution, the French Revolution, and the Russian Revolution, for example, differed significantly. Therefore, it is misleading to suggest that they were all legitimated by the same ideals.

This typically historicist objection to metahistory is important, but not entirely decisive. While different particular rights and liberties were sought at different times and places, all can still be called rights and liberties. Indeed, all drew upon traditions of thought about human reason and dignity found in classical and biblical writing, which were applied to the particular situations. It is wrong to suggest that these events all had precisely the same legitimation, but it is also wrong to deny that the values they appealed to were quite unrelated.

2. *Prescriptive Theories of Historical Explanation*

In the last section we considered some philosophical responses to theories of historical change, that is to theories which identified certain more or less constant sources of historical developments. These theories provided substantive explanations of history. We turn now to theories of historical explanation of a quite different kind. These are theories about the logical structure of explanations of historical events. They are not theories about what produces historical change, but theories about how historians should explain what has happened in the past. These theories are prescriptive, describing how their authors think historical

events should be explained. There are in fact many forms of explanation commonly used by historians, but these theories focus upon just those which their authors judge, for different reasons, to be the most important.

2.1 Theories of Scientific Explanation

In the nineteenth century both John Stuart Mill and the historian Henry Thomas Buckle argued that historians should try to explain historical events as scientists do, by discovering the laws of society which governed their occurrence. With these laws the occurrence of historical events could be deduced, given certain antecedent conditions, just as the occurrence of natural events can be deduced from the laws of nature.

When K. R. Popper came to propose a theory of historical explanation, he quoted Mill's theory with approval, but thought it could be made more precise. 'I suggest,' he wrote, 'that to give a causal explanation of a certain *specific event* means deducing a statement describing this event from two kinds of premises: from some *universal laws*, and from some singular or specific statements which we may call the *specific initial conditions*' (Popper 1961, p. 122). Popper went on to point out that the laws used by historians are familiar, everyday ones of no particular interest. He opposed those historicists who believe there are general laws of historical development; and he opposed those who claim that historical events are unique, and so have nothing general about them whatever. Those who present general laws of history, such as the Marxist theory that all history is of class struggle, are really offering interpretations, Popper said, describing those changes in history which interest them.

The philosopher who most clearly and forcefully presented a deductive model of historical explanation was C. G. Hempel, in his article 'The function of general laws in history.' Drawing upon the tradition of Mill, Popper, and others, he produced the following analysis: The explanation of an event of some specific kind E

> consists of
> (1) a set of statements asserting the occurrence of certain events $C_1 \ldots C_n$ at certain times and places,
> (2) a set of universal hypotheses, such that
> (a) the statements of both groups are reasonably well confirmed by empirical evidence,
> (b) from the two groups of statements the sentence asserting the occurrence of event E can be logically deduced.
> (Hempel 1965, p. 232)

The letters C and E are intended to suggest cause and effect. The aim of historical explanations, he wrote, is that of 'showing that the event in question was not "a matter of chance," but was to be expected in view of certain antecedent or simultaneous conditions' (Hempel 1965, p. 235). He admitted that the relevant laws are often difficult to formulate precisely, and that often the consequence of the initial conditions occurs only with a degree of probability (Hempel 1965, p. 237). Consequently he called the explanations offered in history incomplete 'explanation sketches' (Hempel 1965, p. 238).

Hempel's formulation of the deductive theory of explanation was subject to much discussion and criticism. The main objections were as follows:

(a) The model does not state conditions *sufficient* to produce a good explanation,

(i) because sometimes events which enable one to predict an outcome are not the events which caused it. For instance, from a sudden fall in a barometer one can predict a storm, but that event does not explain the occurrence of the storm, because the fall in the barometer was not necessary for the storm's occurrence.

(ii) Furthermore, for an explanation to be satisfactory, it must describe the kinds of cause which are of interest to the reader. If a man with a gastric ulcer is suffering from stomach pain after eating parsnips, a doctor would attribute the pain to his gastric ulcer, whereas his wife would attribute it to his eating parsnips. For the doctor, eating parsnips is normal, but having an ulcer is not, whereas the wife is familiar with the man's ulcer and picks out his eating parsnips as the event which made the difference. This objection assumes that the function of causal explanations is to explain a contrast between a given event and another expected, normal event, rather than to show why an event was predictable.

(b) The model does not state conditions *necessary* for a good explanation,

(i) because often all that is required to explain an event is to show that the event was more probable than a contrasting state, which might not be very probable at all. Consider the question: 'Why was the monarchy replaced by a republic in France in 1792?' This question could be asking: why was there a change in constitution and not just a change of government, why was the republic set up in 1792 and not earlier, and why was the establishment of a republic in France more likely than any other kind of constitution, such as a limited monarchy of the sort which resulted from the English Revolution?

(ii) Furthermore, many explanations in history are not intended to show that an event was very probable, or that it was more probable than a contrasting state. For instance, historians often want to know the reasons why people behaved as they did, reasons which may or may not have made their actions very probable.

There is some merit in Hempel's theory that historical explanations are meant to show events to have been very probable. Admittedly, some explanations have different aims. And certainly causal explanations must refer to events which were necessary for the occurrence of the event being explained. Never-

theless, some causal explanations are not contrastive, but are meant to explain what brought about a certain event by pointing out events and circumstances which made its occurrence very probable. However, this can very seldom be done by citing a single law relating cause and effect. Normally historical events are the product of several tendencies at work in a situation, and are explained as an outcome of those. Historians try to make the outcome appear as highly probable as possible, but often that is not very probable, just more probable than anything else in the circumstances.

2.2 Theories of Hermeneutic Explanation: Explaining the Meaning of Actions

While G. Vico and J. G. von Herder had written of the importance of understanding human cultures as complex, almost organic units, the theory of hermeneutic understanding, in opposition to scientific understanding, was first fully applied to history by W. Dilthey. The theory had its origin in the lectures on hermeneutics by F. Schleiermacher, whose biography Dilthey published early in his career (in 1871). Following Schleiermacher, Dilthey said that the way to understand human actions is by using empathy to imagine the experiences and thoughts which moved people to act as they did. We interpret what people do and say, he said, by drawing upon our own experiences of life to imagine what was going on in their minds to make them produce the expression which is being interpreted. After he had developed this psychological theory of understanding, Dilthey became interested in the meaning human actions have in furthering the aims of the social or cultural structures to which they contribute. He wrote:

> To start with, the parts of a life have a meaning according to their relation to that life, its values and purposes, and according to the place they occupy in it. Furthermore, historical events become significant through being links in a system of interactions in which they co-operate with other parts to bring about values and purposes in the whole.
> (Dilthey 1961, p. 148)

The structures he mentions as being 'the simplest and most homogeneous' are 'education, economic life, law, politics, religions, social life, art, philosophy, science' (Dilthey 1961, p. 146). These different structures often express similar values and beliefs, and so are interconnected. These common ideas can create a common tendency in a particular culture for an age. People are constantly improving the structure and effectiveness of these organizations, and their actions can be interpreted as having such historical significance.

The main objection to Dilthey's theory of historical understanding was clearly stated, and addressed, by Gadamer in *Truth and Method* (1986). Dilthey had said that we understand other people's actions on the basis of our general knowledge of life in the world. Gadamer pointed out that often the historian's world is remote from the world of the people being studied. Dilthey, he said, 'took no account whatsoever of the historical nature of experience' (Gadamer 1986, p. 212). Heidegger explained that understanding things is not something we do after we have apprehended them, but rather it is what we do *in* apprehending them, in reading a text or in recognizing a machine. And our interpretation of things is a function of our own historical nature (Gadamer 1986, pp. 230–2): 'that we study history only insofar as we are ourselves "historical" means that the historicalness of human There-being [*Dasein*] in its expectancy and its forgetting is the condition of our being able to represent the past' (Gadamer 1986, p. 232). Our understanding of what things mean is first of all provided by the traditions in which we have been raised (Gadamer 1986, p. 250).

Once the historicity of historical understanding is fully understood, the question of the objectivity of history becomes quite urgent. Is there any sense in which one's understanding of the past could be called 'objective'? Gadamer has a couple of things to say about this. First, he insists upon coherence as a condition of correct understanding: 'The movement of understanding is constantly from the whole to the part and back to the whole. Our task is to extend in concentric circles the unity of the understood meaning. The harmony of all the details with the whole is the criterion of correct understanding' (Gadamer 1986, p. 259). He remarks that when it is difficult to find a coherent interpretation, then we suspect that we are dealing with a world whose patterns of meaning we do not really understand. This leads to his second important point. Gadamer employs the concept of a 'horizon,' which he describes as 'the range of vision that includes everything that can be seen from a particular vantage point' (Gadamer 1986, p. 269). Historians occupy a particular vantage point in the history of the tradition which formed them. To *understand* a past event, they must do so from the point of view of those involved. They can find that historical horizon because they have learned of it from within their own tradition. Then when historians come to *report* their understanding of the past, they do so in terms of their present worldview. Gadamer explains: 'In the process of understanding there takes place a real fusing of horizons, which means that as the historical horizon is projected, it is simultaneously removed' (Gadamer 1986, p. 273). It is but a short step from Gadamer's account of historical understanding to postmodern theories, which describe history as a cultural construct, usually in narrative form, having no particular relation to the real events of the past whatever.

At the same time as German philosophers were reacting to attempts to make historical explanations scientific by developing theories of meaning, British

philosophers were pointing out that historical explanations are typically accounts of the reasons for which people acted as they did. R. G. Collingwood wrote as follows:

> For science, the event is discovered by perceiving it, and the further search for its cause is conducted by assigning it to its class and determining the relation between that class and others. For history, the object to be discovered is not the mere event, but the thought expressed in it. To discover that thought is already to understand it. After the historian has ascertained the facts, there is no further process of inquiring into their causes. When he knows what happened, he already knows why it happened. (Collingwood 1961, p. 214)

Collingwood said that knowledge of past minds is arrived at by drawing inferences from observable facts, more particularly by forming and testing hypotheses about what people had in mind when acting as they did.

Collingwood's attack on scientific models of explanation and espousal of rational explanations was given fresh expression by W. H. Dray in his first book *Laws and Explanation in History* (1957). He criticized the 'covering-law models' of Popper and Hempel, and developed an account of rational explanation which was less idealistic and more logical than Collingwood's. He said that in deciding reasons for actions, historians relied on general knowledge of the form: 'When in a situation of type $C_1 \ldots C_n$ the thing to do is x' (Dray 1957, p. 132). Such statements, he said, were 'principles of action' which are universally known and respected. They explain actions by showing that the action was the rational thing to have done, in the circumstances.

Today we would give more respect to Collingwood's account of explanation than to this early statement of Dray's, because people's reasons for acting are not always universally appropriate. Indeed, even from the agent's point of view, they are sometimes not very rational at all. All that matters when explaining the reasons for an action is that they were the reasons for which the agent acted.

Hempel responded to Dray by arguing that reasons only explain actions if we know that people holding such reasons were likely to act on them (Hempel 1963). According to Dray, historical explanations should show that actions were reasonable responses to a situation. Hempel said they must always show that the action was to be expected in the circumstances. I think reasons can cause actions by inclining people to perform them, but they do not always make the actions appear either very rational or predictable.

2.3 *Theories of Structural Explanation in History*

There are two distinct kinds of structural explanation in history, both closely associated with the work of the French Annales school in the twentieth century.

2.3.1 Braudel's structural theory. The first kind of structural explanation is most famously exemplified by F. Braudel's large history, *The Mediterranean and the Mediterranean World in the Age of Philip II* (1972b). In it, Braudel explains the geographic and social structures in which observable actions occurred, and how those structures limited or promoted certain possibilities. He wrote:

> when I think of the individual, I am always inclined to see him imprisoned within a destiny in which he himself has little hand, fixed in a landscape in which the infinite perspectives of the long term stretch into the distance both behind him and before. In historical analysis as I see it, rightly or wrongly, the long run always wins in the end … it indubitably limits both the freedom of the individual and even the role of chance. (Braudel 1972b, p. 1244)

The constraints which Braudel had in mind were largely geographical and technological, but he included frameworks of belief as well (see Braudel 1972a, p. 18).

Braudel showed how geography and technology limited agricultural production and trade, and so the growth of wealth in the sixteenth century. The size of populations, and their vulnerability to famine and disease, also played a part. In addition to these slowly changing structures of geography and technology, Braudel also described cycles of economic and social change which occurred more frequently. These constituted processes of change within economic, social, and institutional structures which historians could model and describe in a theoretical way. He provided a careful analysis of the cycles of mercantile capitalism between the fourteenth and eighteenth centuries, which formed the background for the history of Philip II in the sixteenth century. As Braudel showed, these structures limited the possibilities of human action, as well as making the effects of certain events very probable. However, Braudel gave no systematic account of the relation between these two kinds of structures and the actions of individuals, which constituted the third level of historical description he identified.

2.3.2 Structures of meaning. The second kind of structural explanation favored by historians of the Annales school, such as Lucien Febvre, Marc Bloch, and more recently by Emmanuel Le Roy Ladurie and many others, is one which explains the meaning of an utterance, a gesture or a symbol by referring to a structure of meanings, a language, of which it is a part. Dilthey had anticipated this theory of meaning in some of his later writings, although he is best known for relating meaning to intention. This French structuralist theory of explanation emerged from work by Saussure in linguistics and Lévi-Strauss in anthropology, which explained the mean-

ing of one part of a language by referring to other parts to which it was systematically related. Historians found such patterns of meaning in social events. As Jonathan Culler has put it:

> The cultural meaning of any particular act or object is determined by a whole system of constitutive rules [such as rules of a marriage ceremony or of football]—rules which do not so much regulate behaviour as create the possibility of particular forms of behaviour … make it possible to marry, to score a goal, to write a poem, to be impolite. It is in this sense that a culture is composed of a set of symbolic systems. (Culler 1973, p. 22)

People are seldom aware of the rules governing the symbols they use, any more than they are aware of the rules of language, but the rules can be discovered by historians and anthropologists studying their behavior. The rules can be implicit or explicit. Historians prefer to study the explicit rules governing behavior, usually found in the discourse by which it is described, because then the meaning given to the behavior is truly historical, and not something merely inferred by the historian.

2.4 Theories of Narrative Explanation

Once historical explanations were thought to display the meaning of historical events, rather than just describe their causes, the way was open to view historical narratives as forms of explanation. This is precisely what Hayden White did in his seminal book *Metahistory* (1973). In the theoretical introduction to that book White describes 'explanation by emplotment,' 'explanation by formal argument,' and 'explanation by ideological implication.' A historical event which is merely listed in a chronicle, he said, has no particular meaning or significance. But once it is located in a story, it is given both meaning and significance, by being related to what preceded and followed it. Following Northrop Frye, White identified four forms of emplotment: Romance, Tragedy, Comedy, and Satire, each with its characteristic structure. Thus, for example, 'Romance is fundamentally a drama of self-identification symbolised by the hero's transcendence of the world of experience, his victory over it, and his final liberation from it.' (White 1973, p. 8). As well as telling stories of different kinds, he said, historians also describe the relations between events in the past in several different ways. Following Stephen C. Pepper, he identified four kinds of 'argument' about the nature of historical events. The Formist depicts them as unique; the Organicist sees particular events as parts of a whole; the Mechanist presents events as instances of general laws at work in history; and the Contextualist describes events as products of tendencies at work in their environment. Finally, narratives can represent certain ideological views of historical change, and give events meaning by locating them within one of those. Following Karl Mannheim, White picks out four such ideologies: the Conservatives,' who see history as the evolution of existing structures; the Liberals,' who think that reason and goodwill will eventually see that certain ideals are attained; the Radicals,' who think utopian conditions should be achieved by revolutionary means; and finally the Anarchists,' who believe that an ideal state will be achieved by destroying existing structures and leaving people free to act on their own humanity. White found 'affinities' between these three groups of four. For example he said Romantic plots often employed a Formist presentation of events, and embodied an Anarchist ideology.

White correctly points out that explanations of these kinds can be viewed as interpretations. Interpretations are one of several more or less equally well justified accounts of a subject which give meaning and significance to its parts. White did not consider whether one interpretation could be better than another, but F. R. Ankersmit did in his book *Narrative Logic* (1983, Chap. 8), in which he declared that narrative interpretations should aim for maximum scope, not only in the number of events they describe but in the number they illuminate. This seems to be on the right track.

3. Descriptive Theories of Historical Explanation

Hayden White denied that his analysis of historical explanation was exhaustive, but it was far less prescriptive than previous theories, indicating a large range of ways in which historians could account for events in the past. An avowedly descriptive, analytic approach to historical explanations is that provided by McCullagh in *The Truth of History* (1998). McCullagh analyzes forms of explanation common in historical writing to discover both their structure and their conditions of acceptability. He contrasts explanations, which are answers to fairly specific questions guided by the preconceptions implicit in the question, and interpretations, which look for patterns among events with no particular preconception of what that pattern will be. The adequacy of explanations, on this account, is judged by how adequately they answer the questions they are addressing. The questions provide both the logical and conceptual parameters of the required explanation. The questions, he says, can ask *what* happened during a certain event, *how* a certain change took place, *who* a certain person was, *what* a text or symbol meant, *what* the structure of a society or institution was, as well as *why* a certain event occurred.

An interesting outcome of this approach is his analyses of causal explanations in history. McCullagh finds that causes are events which trigger causal powers, or tendencies, and he depicts causal explanations as setting out the active tendencies which, taken together, made probable the event being explained. A complete causal explanation, McCullagh says, will mention all the causes which directly

increased the probability of the outcome, and these could include both reasons for an action and circumstances which made such an action probable. It does not include all events necessary for the outcome, however, as these go back to the beginning of time.

At present there is some uncertainty as to whether the concept of a correct explanation is intelligible, and whether explanations always provide interpretations of the past. The answer depends largely upon our concepts of explanation and interpretation, and there is no doubt that these need further refining.

See also: Explanation, Cognitive Psychology of; Explanation: Conceptions in the Social Sciences; Functional Explanation: Philosophical Aspects; History: Forms of Presentation, Discourses, and Functions; History: Overview; History: Theories and Methods; Narratives and Accounts, in the Social and Behavioral Sciences; Rational Choice Explanation: Philosophical Aspects

Bibliography

Ankersmit F R 1983 *Narrative Logic. A Semantic Analysis of the Historian's Language*. Martinus Nijhoff, The Hague, The Netherlands

Braudel F 1966 *La Méditerranée et le Monde Méditerranéen à l'Epoque de Philippe II*, 2nd edn. Librairie Armand Colin, Paris

Braudel F 1972a History and the social sciences. Annales: economies, sociétés, civilisations. In: Burke P (ed.) *Economy and Society in Early Modern Europe. Essays from Annales*. Routledge and Kegan Paul, London

Braudel F 1972b *The Mediterranean and the Mediterranean World in the Age of Philip II*. Collins, London

Collingwood R G 1961 *The Idea of History*. Oxford University Press, Oxford, UK

Culler J 1973 The linguistic basis of structuralism. In: Robey D (ed.) *Structuralism: An Introduction*. Clarendon Press, Oxford, UK

Dilthey W 1961 *Meaning in History. W. Dilthey's Thoughts on History and Society* [ed. Rickman H P]. George Allen and Unwin, London

Dosse F 1987 *L'Histoire en miettes: Des 'Annales' à la 'nourelle histoire.'* Éditions La Découverte, Paris

Dosse F 1994 *New History in France. The Triumph of the Annales* [trans. Conroy P V, Jr]. University of Illinois Press, Urbana, IL

Dray W 1957 *Laws and Explanation in History*. Oxford University Press, Oxford, UK

Dray W 1974 The historical explanation of actions reconsidered. In: Gardiner P (ed.) *The Philosophy of History*. Oxford University Press, Oxford, UK

Gadamer H-G 1960 *Wahrheit und Methode*. J C B Mohr, Tübingen, Germany

Gadamer H-G 1986 *Truth and Method*. Crossroad, New York

Hempel C G 1963 Reasons and covering laws in historical explanation. In: Hook S (ed.) *The Philosophy of History: A Symposium*. New York University Press, New York, pp. 143–63

Hempel C G 1965 The function of general laws in history. In: Hempel C G (ed.) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press, New York

Lyotard J-F 1979 *La Condition postmoderne: rapport sur le savoir*. Editions de Minuit, Paris

Lyotard J-F 1984 *The Postmodern Condition: A Report on Knowledge* [trans. Bennington G, Massumi B]. Manchester University Press, Manchester, UK

Lyotard J-F 1987 The sign of history. In: Attridge D, Bennington G, Young R (eds.) *Post-structuralism and the Question of History*. Cambridge University Press, Cambridge, UK

Manuel F E 1965 *Shapes of Philosophical History*. Stanford University Press, Stanford, CA

McCullagh C B 1998 *The Truth of History*. Routledge, London

Owensby J 1994 *Dilthey and the Narrative of History*. Cornell University Press, Ithaca, NY

Popper K R 1961 *The Poverty of Historicism*. Routledge and Kegan Paul, London

Spengler O 1926–8 *The Decline of the West* [trans. Atkinson C F]. Allen and Unwin, London, 2 Vols.

Toynbee A J 1951–61 *A Study of History*. Oxford University Press, London, 10 Vols.

White H 1973 *Metahistory: The Historical Imagination in Nineteenth-century Europe*. The Johns Hopkins University Press, Baltimore, MD

White H 1978 Interpretation in history. In: White H (ed.) *Tropics of Discourse. Essays in Cultural Criticism*. The Johns Hopkins University Press, Baltimore, MD

C. B. McCullagh

Historical Geography

Partly because of its richly diverse and spatially variable traditions, historical geography is not amenable to ready definition. While it is concerned with the study and reconstruction of past geographies and understanding processes of ecological and landscape change, its remit also includes the political, cultural, and economic interactions between past and present. Thus historical geography examines how our interpretations and representations of the past are constructed in contemporary circumstances and shaped by current needs. There are, however, significant differences between societies in the traditions of historical geography and the ways in which it is currently construed, researched, and taught. This article deals less with the history of historical geography than with certain of the prevailing trends in its current status as an academic project. A discussion of the nature of contemporary historical geography is followed by an assessment of three themes, which permit an elaboration of this general context. These are: interconnected historical geographies; differential historical geographies; and the past in the present.

1. The Nature of Historical Geography

Two overlapping perspectives long dominated traditional approaches to historical geography. First was an intensely academic concern, epitomized, for ex-

ample, through the work of H. C. Darby in Britain and Ralph Hall Brown in the USA, with the scholarly reconstruction of past geographies for their own sakes. The second, encapsulated, for instance, in the work of Andrew H. Clark in North America, emphasized vertical themes, the essentially spatial changes and continuities that occur through time and help explain the present. More recently, however, historical geography has been revitalized by the wider concern throughout the social sciences and humanities with the relativity of knowledge. Thus the fundamental assumption informing much current historical geography is that representations of the past are constructed in the present to serve the demands of contemporary societies. The content and methodology of this work, particularly in the anglophone world, also reflects the theoretical shifts in many of the social sciences away from the 'objectivist fallacy' of positivist models of research and knowledge production (Wishart 1997). Conceptualizing the past as a resource for the present also inevitably invokes notions of contestation as various groups compete for 'ownership' of the past. Consequently, much current work in historical geography acknowledges the situated nature of research and the political difficulties of reconciling contested interpretations of the past in multicultural societies.

While this work builds on historical geography's long-standing empirical focus on topics including industrialization, rural settlement and land ownership, agriculture and agrarian change, migration and population patterns, urbanization, and transport networks, it is also moving in new directions. Many contemporary historical geographers, informed by perspectives such as feminism, poststructuralism, antiracism, and postcolonialism, share concerns about questions of power and meaning with other researchers more readily located within the traditional systematics of cultural, economic, political, and social geography. The key themes evident in much recent historical geography include: a more overtly theorized exploration of the material, political, and symbolic dimensions of travel and exploration; imperialism and colonialism; nationhood and state formation; and ideas of nature and environmental change (Graham and Nash 1999, Nash 1999).

The exploration of the philosophy, content, and methodology of historical geography has itself a long tradition (Butlin 1993). This introspection has often been anxious, not least when faced with the positivistic revolution in geography in the late 1960s, a paradigm shift to which historical geography seemed ill equipped. The delineation of the boundaries between historical geography and geographical history has also occasioned much debate, often driven by a perceived need to stake out precise disciplinary territories. For example, Guelke (1997) and Donkin (1997) have expressed their fears at the deconstruction and potential disappearance of historical geography. To them it lacks an academic core, defined particularly through the ways in which people relate to the earth. Nevertheless, much current historical geography is concerned less with the establishment of precise boundaries within geography, or between geography and history, than in engaging in the arguably more valuable task of creating a body of work that is both geographical and historical. In addition, concerns over the erosion of historical geography by other systematics within the subject, or its dilution through the influences of other disciplines, ignore the interdisciplinary and hybrid nature of both past and present historical geography. Where traditionally it drew on archaeology, economic and social history, ecology, and social theory, it now interacts with feminism, environmental history, postcolonialism, and cultural history.

Thus in his capacity as editor of the *Journal of Historical Geography*, Michael Heffernan (1997, p. 2) argues for an historical geography that is 'a haven of intelligent eclecticism.' He offers 'an alternative and more optimistic prospectus for the future' to those 'who worry that eclecticism breeds conceptual chaos and threatens the intellectual integrity of historical geography as an academic project.' Heffernan sees it as a hybrid discipline open to the widespread contemporary questioning of 'the conventional intellectual categories through which the modern world has been interpreted and conceptualized.'

While this eclecticism is undoubtedly reflected in a more cultural historical geography, that content should be no less concerned with the material shaping of places, spaces, landscapes, and lives, even if the need to ground historical geography in material contexts of power, politics, and economics is sometimes neglected. The past is a political resource and its control is fundamental to issues of hegemony and empowerment in societies. If historical geography is about the interplay between the past and the present, between symbolic and material worlds, it must also be open to the hard-edged realities that, for many people, 'history hurts.'

2. *Interconnected Historical Geographies*

Interconnectedness possesses at least three separate dimensions. First, the history of historical geography itself reflects complex webs of intellectual interconnections, which have produced different outcomes in different places. It is also true, however, that linguistic and intellectual parochialism often limits the potential of these interconnections. Second, historical geography possesses a lengthy tradition of locating regional studies within broader processes operating at macrospatial scales, and of paying attention to both the specificities of the local and their interaction with wider economic, cultural, and political domains and institutional structures. Finally, there are interconnections between places—for example, the modern networks of communication, transport, or trade that

connected Europe and its imperial sphere of influence—but also in the kinds of geographical discourses and practices that have structured the ways in which the relationships between places or regions are understood.

2.1 Interconnections within Historical Geography

Ogborn (1999, p. 99) argues that a key dimension to interconnectedness is that of showing how and why issues, ideas, and intellectual interactions 'cross connect and reshape a kaleidoscopic interdisciplinary terrain.' Thus cultural and material geographies combine, for example, to produce environmental histories that are also histories of power, colonialism, and imperialism (Wynn 1997). That historical geography is seeking to understand the social or cultural construction of spaces, knowledges, and powers also impacts on the discipline itself, which has been, and remains, a spatially variable process. The contents and methodologies of historical geographies vary both through time and across space, reflecting the current —if transient—trends defining the broader intellectual climates within which they are practiced. Consequently, the subject matter of contemporary historical geography in the anglophone world is dominated by the cultural turn in human geography and by the poststructuralist emphasis on situated knowledges. The early modern and modern periods are preferred to the medieval and the contingent to theories of a past shaped by cyclical processes of historical transformation of varying temporality. Marxian theories of world systems and modes of production that emerged from long-term cycles of history and their periodic crises (e.g., Wallerstein 1974–89) are deeply unfashionable in a discipline that has often embraced Foucault's dictum that nothing is fundamental with uncritical enthusiasm.

The temporal and spatial variability in historical geography is nowhere better illustrated than in the contrasting perspectives on landscape, historical geography's oldest and most enduring concept. For example, the evolution of British historical geography has, in many ways, been dominated by efforts to provide alternative directions to the source-driven methodology used by H. C. Darby in his systematic reconstruction of past landscapes. Although open to vertical themes and the cycles of history behind geography, Darby's methodology invoked an empirical dependence on the inevitably élitist context of cartography and documents, and consequently excluded dimensions of power and culture in the societies and landscapes reclaimed. In some contrast, the French geographer, Paul Vidal de la Blache (1926), emphasized the significance of ordinary people and their environment: to him, the region was not simply a convenient framework, but rather a social reality (Claval 1984). Vidal's belief that landscape reflected a harmony between human life and the milieu in which it was lived was also fundamental to the work of the Berkeley geographer, Carl Sauer. For him, the richest source of evidence lay 'in the material content of the cultural landscape,' the 'task of human geography [being] nothing less than comparative study of areally localized cultures' (Conzen 1993, p. 28).

Vidalian ideas were also crucial to the *géohistoire* of the French *Annales* school, and its concerted attempt to map and explain the complex reality of human life by reference to local and regional studies. They also underpinned the geographical philosophy of Ireland's most influential geographer, E. Estyn Evans, partly explaining why Irish historical geography evolved along a very different trajectory to that in Britain. This was one in which the meanings of historical place and landscape were seen as fundamental to a people's genetic and cultural identity, albeit conceptualized in overtly more anthropogeographical terms than through the traits of polyvocality and hybridity that attend identity studies today. As in Sauer's methodology, the critical evidence was that of the visible and material landscape. Similar ideas also influenced the study of landscape in North American, Scandinavian, and Australasian historical geographies, and clearly retain a contemporary resonance and vibrance, given the hybrid and multi stranded processes that link the discursive and material worlds, which constitute the raw resources of historical geography.

2.2 Local Particularities and General Processes

Historical geography has long been concerned with the processes of historical specificity and historical transformation across space and through time. The intellectual attraction of the 'grand' structuralist models lies in their identification of general power-related processes, which interact with the specificities of local place to produce an almost infinite chiaroscuro of individual outcomes and experiences. Historical geography becomes the product of this complex hybridity and polyvocality of places and peoples, while the contingent is reconciled—as societies seem to require—with temporal structures that create explanations of change, meanwhile observing a cultural and political need for continuity and social memory. The outcome is spatial and temporal diversity, which historical geography has long sought to reconcile with generality of process through a methodology of regionally based analysis.

This is perhaps most coherently encapsulated in a French *géohistoire* seeking to explain a country defined by diversity in which unity—as elsewhere in Europe—was achieved only by force, civil war, and the creation of ethnic nationalisms. Vidal de la Blache argued that it was this diversity—rooted in the physical en-

vironment—which paradoxically provides France with its identity. The country's unity evolves from a beneficent force of commonality, derived from the ways in which social life combines with—and transcends—its divisions through time. This relationship was encapsulated in the idea of the *pays* (literally an area with its own identity derived, not only from divisions of physical geography, but also from ethnic and linguistic divisions imposed on a region by its history) as the geographical mediation of synthesis and continuity, the product of human interaction with the environment over many centuries.

In his monumental (if unfinished) investigation of the identity of France, Fernand Braudel (1988–90) takes up the same theme. He sees a patchwork France, a jigsaw of regions and *pays* in which the vital thing for every community is to remain other from the next tiny *patrie*. But Braudel, recognizing a similar fragmentation elsewhere, was also critical of Vidal's exceptionalism: 'Yes, France is certainly diverse…it is a diversity which breaks up, divides, and sets one region against another.' But it is not an 'unparalleled' diversity: 'Germany, Italy, Britain, Spain, Poland can all lay claim to diversity' (Braudel 1988–90, p. 669). So too can smaller countries. Pursuing the same symbiosis of physical environment and human society, Evans (1981) believed that even fractious Ulster could seek unity in a diverse past to which all its present inhabitants might subscribe.

Although Braudel and Evans saw the natural world as providing the backdrop to this human diversity, both also recognized the decisive contribution of the economy and social intercourse. Moreover, diversity itself was more apparent because it often reflected what Braudel referred to as *l'histoire événementielle*, short-term episodic history or the history of events. Behind all that, however, lay *l'histoire structurelle*, the long-term changes of *la longue durée*, 'the slow speed of the secular trend, moving scarcely at all' (Braudel 1988–90, p. 678). The real essence of this European conceptualization of diversity lies in the ways in which these different histories, trajectories of time, and social change and geographical particularities have fused in varying ways to produce distinctive places and peoples.

The same theme of region and regionalism—albeit differently expressed and generally lacking the Marxian connotations of *géohistoire*—is also a fundamental feature of North American historical geography. Clark's immensely influential emphasis on historical change was expressed through a methodological approach that was 'explicitly historical, attentive to the economic basis of regions, devoted to field and archive, well integrated with professional history, analytically adventuresome, and book-oriented' (Conzen 1993, p. 61). The outcome, readily apparent in much historical geography in both the USA and Canada, was a heavily empirical emphasis on regional spatial patterns, which often conspicuously failed to ask fundamental questions regarding broader social, political, and economic relationships (Wynn 1993).

2.3 Interconnections Between Places

This final dimension to interconnectedness refers both to the physical linkages between places, but also to the ways in which the relationships between places and regions are understood in geographical discourses and practices. It is concerned with the ways in which technological developments and their exploitation allowed particular countries to achieve at least a transient hegemony in the world system. Hugill (1993) argues that the postmedieval nations that developed and marketed new technologies best were the nations that rose to world power, revolutions in transport and communication technologies actually constituting the moments of transformation from one world economy to another. Globalization is not a recent process, the creation of European imperial empires effectively leading to the integration of large areas of the world into the emerging national space economies of Europe. These global networks and flows of information, capital, people, culture, plants, animals, and objects that linked the world in the past continue to rebound on its present, not least through the persistent patterns of disparity that mock the concept of global free trade.

If imperialism and colonialism were hierarchical, they also represented reciprocal networks of influence and power. European nationalism and the emerging nation-state system were framed within these imperial rivalries and later exported to the former empires. Thus the resistances to imperialism made anti-imperial citizens as well as imperial ones (Ogborn 1997). Moreover, a consideration of the connections between places also involves stressing the ways in which the historical geographies of colonial countries themselves were shaped by colonialism. The commonplace interpretation of the historical geography of colonialism as the imposition of the dominant logic of the colonial political and cultural economy unto non-Western time and territory, requires destabilization (Yeoh 1996). The difficulty in so doing, however, lies in the lack of knowledge about colonized groups, given the asymmetries of the historical record.

One final dimension to the ways in which the interconnections between places help shape geography concerns the linkages between geographical education and the making of citizens. Imperialism and colonialism also entailed political, intellectual and cultural, as well as material, domination. Historical geographers have extended the contention that imperial ideologies remain at the heart of modern Western culture to explore the linkages between the cultures of geography and imperialism. Thus 'knowledges of geography are themselves part of often complex geographies of knowledge' (Ogborn 1997, p. 416). Hence it can be

argued, for example, that both popular and formal geographical education in Britain and notions of Britishness were constructed in direct opposition to imaginative geographies of other places and peoples (Graham and Nash 1999).

3. *Differential Historical Geographies*

Inevitably, the previous discussion has already touched upon the general argument that despite the interconnected nature of world systems, the processes of social change through time, with which historical geographers are concerned, are still experienced differently by peoples in different social groups and places and articulated in different ways. Moreover, these dissimilar knowledges are articulated at a variety of scales ranging from the local to the global. The first dimension to this issue to be considered here concerns problems of social inclusion and exclusion. Second, however, it is also the case that knowledges themselves are rarely consistent and contemporary historical geography is itself internally differentiated or even contested between different ideological and methodological imperatives.

3.1 *Social Inclusion and Exclusion*

Narratives of the relationships between places are characteristically defined through processes that privilege particular accounts and peoples while marginalizing or excluding others. Contemporary historical geography recognizes that earlier generations did not sufficiently understand the privileging of Europe in their accounts and is aware of the need to avoid the naturalization of Western models of modernity and development. Again, Conzen (1993) identifies the 'startling silences,' including slavery, the African American experience, and the results of war, in American historical geography. The combined importance given to gender and racism in contemporary social and cultural theory helps explain the polyvocality and hybridity of contemporary historical geography. Just as racism ensured that colonized indigenous peoples were excluded from the narratives of empire, all human geography has traditionally privileged masculine discourses and conceptualized modernity in masculine, middle-class, urban, and Eurocentric terms (Nash 1999). Women—especially rural women—were equated with the authentic and premodern, while, in contrast, the modern was associated with a male-directed logic of rationalization, objectification, and developmental progress. Thus traditionally women have been largely invisible and misrepresented in historical geography, while men have 'monopolized the transmission of history' (Lowenthal 1998, p. 48).

However, it is important not to equate marginalization and exclusion with gender and race alone. A contemporary historical geography has to be open to the nuances of multiple resistances, and to the enduring importance of class, ethnicity, and nationalism as criteria of identity. The implication of the past in the construction of identity narratives is but one example of the way in which historical geography has become a more pluralist and disputed enterprise since the beginning of the 1970s. There are many different modernities and many diverse dimensions to othering and otherness.

3.2 *Historical Geography as a Contested Enterprise*

As noted in the discussion regarding the interconnectedness of historical geographies, the result of differentiation may be knowledges that are internally inconsistent and contested between opposed ideological and methodological imperatives. Heffernan's 'intelligent eclecticism' does not preclude disputation, inevitably so because of the privileging of particular perspectives within the geographical discipline itself. If a methodologically diverse historical geography traditionally seemed eclectic, its focus on areal differentiation, modernist notions of development and progress along continuities from past to present, and its reliance on empirical, material evidence drawn from the landscape, did link it to many of the principal strands that defined the emergence of geography as a discipline.

In the 1960s, however, the positivist revolution seemed to have little to offer historical geography. Largely lacking the means to meaningfully quantify data and indulge in model building, its practitioners found little resonance in arguments that geography was a formal deductive science charged to describe and understand the real world. The important exception to this generalization was the reversal of historical geographers' traditional aversion to the urban realm, although this was expressed—as the prevailing *zeitgeist* demanded—through a commitment to the morphological and functional relationships of the historic city and to its modeling. Questions of power and social relationships remained peripheral as they had in the regional studies of the rural world.

Given historical geography's vulnerability to this paradigm shift, Harris's (1978) argument that an understanding of the synthesis of regions, places, and landscapes might come from the development of an historical mind that could encompass experience, memory, imagination, and reflection was a welcome harbinger of the revitalization of historical geography by the cultural turn in geography in the late 1980s and 1990s. It is still important, however, to be aware of the dangers of cultural exceptionalism in what remains a polyvocal enterprise. While the cultural theme was dominant in British historical geography and, if to a lesser extent, throughout the anglophone world for much of the 1990s, its content and methodologies do not define historical geography as a global enterprise,

or even describe the research of many historical geographers within the English-speaking world. The privileging of the cultural perspective at the expense of alternative methodologies and languages disguises the survival of enduring issues in historical geography—frontiers and borderlands, migration, settlement morphology, settlement—the city in particular—as an indicator of cyclical social change and social relations, placenames, and the making of the physical environment.

Again, despite its openness to previously disenfranchised peoples and their narratives, the cultural turn in historical geography is often neglectful of power and economy as they control material lives. To some extent, the denigration of the material world is a result of the extension of the source materials used in researching contemporary imaginative historical geographies to encompass a diverse range of written and artistic representations and texts—diaries, biographies, novels, poetry, and paintings. This otherwise welcome development has, however, led to a denigration of traditional sources of data such as the material physical and human landscape. Ironically, the contemporary emphasis on nuanced qualitative readings of texts and representations is paralleled by the previously unimaginable power of Geographical Information Systems and Remote Sensing methodologies to handle and manipulate huge repositories of statistical historical data.

Historical geography now possesses the technological means of far more effectively addressing its traditional responsibilities of reconstructing past environments, spatial analysis, and regional differentiation. Often, however, these tasks have largely been abandoned to a technologically proficient historical archaeology, which combines historical archival research with excavation and geophysical evidence of past societies in both terrestrial and maritime landscapes. More important, historical geography now also possesses the capacity, but has largely shirked the challenge, of producing and managing integrated masses of geographical information, which would substantially enhance understanding of social change through time but also elucidate the regional and local specificities that underpin the constructs of identity through which peoples project and comprehend their relationships with place. Intelligent eclecticism can encompass the deconstruction of the empirical and the abandonment of the objectivist fallacy, but must remain open to using far more elaborate means of processing of historical data, not least to avoid these being subverted by an apparently enhanced capacity for objectivism.

4. *The Past in the Present*

Finally, historical geography is also increasingly informed by the ways in which the past is remembered and represented in both formal or official senses and within popular forms, and the implications, which these have for the present. It is in this respect that the past can be most clearly visualized as a domain of contested meanings, perhaps the most significant way in which contemporary historical geography differs from the archival, 'objective' reclamation of the past that characterized its previous existences. Interpretations of the past are seen as being context bound and power laden, an issue briefly explored here through a consideration of the functions of heritage (Graham et al. 2000).

4.1 *Heritage and its Functions*

The stories and places of contemporary heritage, and the ways in which they are told and described, often represent the most overt examples of the contemporary politics of the past in the present. Heritage is that part of the past selected in the present for contemporary purposes, be they economic, cultural, political, or social. If heritage knowledge is situated in particular social and intellectual circumstances, it is time-specific and thus its meaning(s) can be altered as texts are re-read in changing times, circumstances, and constructs of place and scale. Consequently, such knowledges are inevitably fields of contestation of meaning and function. As such, the past can be visualized as a resource but simultaneously, several times so. Clearly, it is an economic resource, one exploited everywhere as a primary component of strategies to promote tourism, economic development, and rural and urban regeneration. Heritage—variously defined—is the most important single resource for international tourism although the economic commodification of the past can so trivialize it that arguably the result may be the destruction of the heritage resource which is its *raison d'être*.

But heritage is also a knowledge, a cultural product, and a political resource and thus possesses crucial sociopolitical functions. In particular, the past validates the present by conveying an idea of timeless values and unbroken lineages and through restoring lost or subverted values (Lowenthal 1998). Hence there are archetypal national landscapes, which draw heavily on geographical imagery, memory, and myth. Continuously being transformed, these encapsulate distinct home places of 'imagined communities,' comprising people who are bound by cultural—and more explicitly—political networks, all set within a territorial framework that is defined through whichever traditions are currently acceptable, as much as by its geographical boundary. Such national traditions are narratives that are invented and imposed on space, their legitimacy couched in terms of their relationship to particular representations of the past. So important are these that peoples cut off from their past through migration or even by its destruction in war, often recreate it, or even 'recreate' what could—or should have been there—but never actually was. Inevitably,

therefore, the past as rendered through heritage also promotes the burdens of history, the atrocities, errors, and crimes of the past which are called upon to legitimate the atrocities of the present. Consequently heritage is always a duality—a resource of economic and cultural capital. This is less a dialectic than a continuous tension, these broad domains generally being in conflict with each other.

5. Conclusion

A hybrid and interdisciplinary historical geography has thus widened its interests alongside history itself. As historians increasingly explore popular forms of history today as well as cultural sources for the past, historical geographers have also extended what counts as source material in researching contemporary imaginative historical geographies. These include objects, practices, and discourses of popular culture as well as the past itself. If this makes historical geography more cultural, it should not make it any less concerned with the material shaping of places, spaces, landscapes, and lives, or with issues of empowerment and economy. These current intersections reflect the ways in which historical and cultural geography have overlapped in the past—in the study of landscape for example—just as they do in the present, whatever the labels academics chose to define their research interests. The opposition between an empirically grounded historical geography and a purely qualitative cultural geography is clearly reductive and redundant. Similarly, the contemporary emphasis on the ways in which the past features in the present does not make the work of historical geography any less concerned with happenings of history. Rather, historical geography is about the interplay between the past and the present, between symbolic and material worlds.

See also: Cartography, History of; Geographic Thought, History of; Geography; Historical Change and Human Development; Historical Demography; Historical Sociology; Historiography and Historical Thought: Current Trends; History and the Social Sciences; History: Forms of Presentation, Discourses, and Functions; Postmodernism in Geography; Public Sphere: Eighteenth-century History; Public Sphere: Nineteenth- and Twentieth-century History

Bibliography

Butlin R A 1993 *Historical Geography: Through the Gates of Space and Time*. Edward Arnold, London

Braudel F 1988–90 *The Identity of France* (trans. Reynolds S), 2 Vols. Fontana Press, London

Claval P 1984 The historical dimension of French geography. *Journal of Historical Geography* **10**: 229–45

Conzen M P 1993 The historical impulse in geographical writing about the United States, 1850–1990. In: Conzen M P, Rumney T A, Wynn G (eds.) *A Scholar's Guide to Geographical Writing on the American and Canadian Past*, University of Chicago Press, Chicago, pp. 3–90

Donkin R A 1997 A 'servant of two masters'? *Journal of Historical Geography* **23**: 247–66

Evans E E 1981 *The Personality of Ireland: Habitat, Heritage and History*, 2nd edn. Blackstaff, Belfast, UK

Graham B, Nash C (eds.) 1999 *Modern Historical Geographies*. Prentice Hall, Harlow, UK

Graham B, Ashworth G J, Tunbridge J E 2000 *A Geography of Heritage: Power, Culture and Economy*. Arnold, London and Oxford University Press, New York

Guelke L 1997 The relations between geography and history reconsidered. *History and Theory* **36**: 216–34

Harris C 1978 The historical mind and the practice of geography. In: Ley D, Samuels M S (eds.) *Humanistic Geography: Prospects and Problems*, Croom Helm, London and Maaroufa Press, Chicago, pp. 123–37

Heffernan M 1997 The future of historical geography. *Journal of Historical Geography* **23**: 1–2

Hugill P J 1993 *World Trade Since 1431*: Geography, Technology, and Capitalism. John Hopkins University Press, Baltimore and London

Lowenthal D 1998 *The Heritage Crusade and the Spoils of History*. Cambridge University Press, Cambridge, UK and New York

Nash C 1999 Historical geographies of modernity. In: Graham B, Nash C (eds.) *Modern Historical Geographies*, Prentice Hall, Harlow, UK, pp. 13–40

Ogborn M 1997 (Clock)work in historical geography: autumn 1995 to winter 1996. *Progress in Human Geography* **21**: 414–23

Ogborn M 1999 The relations between geography and history: work in historical geography in 1997. *Progress in Human Geography* **23**: 97–108

Vidal de la Blache P 1926 *Principles of Human Geography* [English translation by Bingham M T]. Constable, London and H. Holt and Company, New York

Wallerstein I 1974–89 *The Modern World-System*, 3 Vols. Academic Press, New York

Wishart D 1997 The selectivity of historical representation. *Journal of Historical Geography* **23**: 111–18

Wynn G 1993 The writing of Canadian historical geography. In: Conzen M P, Rumney T A, Wynn G (eds.) *A Scholar's Guide to Geographical Writing on the American and Canadian Past*, University of Chicago Press, Chicago, University of Chicago Research Paper no. 235, pp. 91–124

Wynn G 1997 Remapping Tutira: contours in the environmental history of New Zealand. *Journal of Historical Geography* **23**: 418–46

Yeoh B S A 1996 *Contesting Space: Power Relations and the Urban Built Environment in Colonial Singapore*. Oxford University Press, Kuala Lumpur and New York

B. Graham

Historical Linguistics: Numerical Methods

1. Introduction

The main goal for numerical methods in historical linguistics has always been to answer the question of how closely two languages are related. Applied pairwise to languages, this can be used to reconstruct a tree

for a whole language family (lexicostatistics). More ambitious methods have additionally assigned dates to the splits within those trees (glottochronology). Recently the questions and methods have been extended to involve distant (hence controversial) relationships between languages, so the basic question becomes not just how closely two languages are related, but whether they are related at all. All methods in this area, no matter how sophisticated, rest on the simplest of observations: since languages change over time, two originally related languages gradually become less similar over time. This enables the linguist to infer something about diachrony from synchronic information only, particularly attractive in situations where little is known about the history. The methodological difficulties come in quantifying rates of change, measuring difference, and accounting for the myriad of factors that can interfere with so-called normal change processes (e.g., mutual influence between languages). The attractions of numerical methods in historical linguistics are the same as those attributed to quantitative methods in almost any discipline in the humanities or social sciences: speed, objectivity, replicability, and ability to handle large volumes of data efficiently.

2. *Early Methods*

The earliest precursor to a numerical method in historical linguistics is a statement made by Robert Gordon Latham (1850, p. 565, see also Hewes in Hymes 1960, p. 338): 'the average rate at which languages change is capable of being approximated' and 'the maximum difference, at a given period, between two or more languages is also capable of being approximated.' He never pursued this, and the next milestone in the development of numerical methods in historical linguistics was a statement by Sapir (1916, p. 76): 'The greater the degree of linguistic differentiation within a stock the greater is the period of time that must be assumed for the development of such differentiation.' Sapir's influence on later linguists was great, so even though he himself never followed this statement up by taking the basic steps of quantifying linguistic differentiation or rate of change, this can be seen as the genesis of the most widely known numerical method in historical linguistics, that of Swadesh (1950, and see below).

The first explicit numerical treatment of genetic proximity of languages was Kroeber and Chrétien (1937), following a method created by two anthropologists, Jan Czekanowski (1928) and Stanislaw Klimek (1935), for use in physical anthropology and ethnography. For each pair of languages, on a list of N features, calculate a (number of features shared by both languages), b (number of features in the first language but not the second), c (number of features in the second language but not the first), and d (number of features in neither language). This produces a 2×2 contingency table, a well-studied mathematical entity enabling various forms of analysis. In particular, the quantity Nr^2, where $r = (ad-bc)/\sqrt{(a+b)(c+d)(a+c)(b+d)}$, is χ^2 (one degree of freedom), enabling one to test the statistical significance of the relationship between the two languages. The values of r for all pairs of languages under consideration can become the input to various methods (e.g., hierarchical cluster analysis) for family tree reconstruction. The method is simple but not unproblematic (e.g., construction of the feature-list, difficulties with non-independence of features, interpretation of correlations especially if negative). (For fuller treatment and commentary, see Embleton 1986.)

The second numerical method in historical linguistics is Ross's root-retention model (1950). For N features and m languages, construct an $N \times m$ table in which a cell has a cross if that language exhibits that feature. The question of relationship of languages can then be recast as whether or not the crosses are distributed randomly in the table. For each pair of languages, calculate n_1 (number of features found in the first language is represented by the number of crosses in the first column), n_2 (number of features found in the second language is represented by the number of crosses in the second column), and r (number of features shared by both languages is represented by the number of rows with a cross in both columns). Then the probability for that pair of languages of obtaining the shared number of crosses or more is

$$\sum_{R=r}^{N} = \frac{\binom{n_1}{R}\binom{N-n_1}{n_2-R}}{\binom{N}{n_2}}$$

for which a normal approximation exists. These probabilities can be tested for significance, and could become the input for various family tree reconstruction methods. Interestingly, Ross's method can be shown to be mathematically equivalent under certain conditions to Kroeber and Chrétien's, and thus subject to the same problems. (For fuller treatment and commentary on this and the method in general, see Embleton 1986.) Although certain elements of these models recur today, both have largely been abandoned, although the problems encountered and lessons learned remain relevant.

Historically, the next method to evolve was Swadesh's, but given its centrality to the field and ongoing relevance, I will deal with that in a separate section. A final numerical method to consider is that of counterindications originated by Gleason (1959), later rediscovered by others. All topologically possible family trees for the languages in question are constructed, each is evaluated for how well it fits against a fixed list of features (every time the tree does not fit the distribution of that feature, a counterindication is

recorded), and the tree with the best score (even if not perfect) is chosen. This method is difficult to implement for two reasons. Even with modern high-speed computing and sophisticated algorithms, the number of topologically possible trees for even a moderately-sized language family quickly becomes unworkably large. The method is also highly vulnerable to the effect of borrowing across languages. A variant of Gleason's method, relying on hierarchical clustering and attempting to estimate branch lengths (i.e., used glottochronologically) is shown in Embleton (1986), along with fuller discussion of the method itself. Taylor et al. (2000), using recent advances and sophisticated techniques from computer science and biological cladistics, are developing a character-based (rather than distance-based) method for family-tree reconstruction with some similarities to Gleason's method. Despite difficulties with dealing with borrowing in realistic amounts, such methods are promising partly because they closely parallel the cognitive processes of skilled practitioners of traditional comparative-historical reconstruction techniques, while being more objective and simultaneously handling vast amounts of data.

3. *Swadesh*

By far the best known numerical method in historical linguistics is that introduced by Swadesh (e.g., 1950), actualizing Sapir's statement by regarding lexical replacement as analogous to radioactive decay. As with carbon-14 dating, dates could then be provided for branching points in the tree (glottochronology). Take a list of N meanings of supposed basic core vocabulary, found universally in all cultures, expressing basic noncultural concepts, expected to be more resistant to change and borrowing than other vocabulary items. Such a list, now known as a Swadesh list, contains such items as lower numerals, simple kinship terms, body parts, topographical terms, widespread flora and fauna, pronouns (personal, demonstrative, interrogative), verbs denoting basic actions, and naturally occuring noncultural phenomena. Assuming a constant rate of replacement, R, for words representing meanings on this list, one can calculate the time elapsed (time depth) since any pair of languages split using the formula

$$\frac{-1}{2R} = \ln\left(\frac{c}{N}\right)$$

where c is the number of cognates on the Swadesh-list between the two languages compared. By calculating time depths for all pairs of languages under consideration, we can construct a family tree with dates.

Swadesh's method was greeted with widespread enthusiasm and widely applied in the late 1950s, 1960s, and early 1970s, almost with a sense of euphoria about finally having an exact tool for obtaining insights into hitherto unknowable linguistic prehistory. The debate over Swadesh's method quickly became emotional, but regardless of the researcher's subjectivity, a common core of problems can be identified, sometimes with potential solutions.

Difficulties with particular meanings on the Swadesh-list.

(a) Some meanings cannot be translated into the target language. This is fairly rare, and best treated as missing data, reducing N accordingly.

(b) Sometimes there are multiple possible translations in the target language for a meaning on the list. Different researchers might make different but equally valid choices from the range of possibilities ('multiple synonymy'). Solutions include always counting only the most frequent possible translation, choosing at random, always choosing to maximize the cognate count, always choosing to minimize the cognate count, or various fractional scoring systems.

Difficulties with cognacy judgments (remember one is not necessarily dealing with languages where the history, and hence cognacy are well known).

(a) Due to accumulated sound change, cognates may not be recognizable, causing us to overestimate time depth.

(b) Chance similarities may be wrongly interpreted as cognates, causing us to underestimate time depth. In practice, one might hope that this and the preceding cancel out.

(c) Sometimes only one part of a word is cognate to a word in another language ('partial cognation'); how should one score this? Solutions include always counting this as positive, always counting this as negative, or various fractional scoring systems.

Problems with the universality of R, which was shown to vary over time, between languages, and according to meaning. R had been estimated from languages with written historical records as about 14 percent per millennium for the standard 100-word Swadesh-list and about 19 percent per millennium for the standard 200-word Swadesh-list. Not only might the very fact that these languages are written bias the rate (one might expect less replacement in a language with a lengthy written history), but 11 of the 13 pairs used to establish the rate were Indo-European (representing only three of its branches).

Problems with the family tree (Stammbaum) model, generally centering around the splits not being clean (thus giving any precise date is artificial and meaningless). Languages are often in close physical proximity even after they split, mutually influencing one another; tendencies in the two languages predating the split may come to fruition only afterwards. Both these interfere with the model, but are more properly problems with the family tree model, and Swadesh's

adoption of it, rather than problems with Swadesh's method itself.

Borrowings between languages (related or unrelated) can alter replacement rates quite substantially. For some languages, even the Swadesh-list contains a high amount of borrowed vocabulary.

For various reasons (see Embleton 1986 and elsewhere), the use of Swadesh-style lexicostatistics and glottochronology waned by the mid-1970s, although applications lingered on in several language areas, and the methods continued to be refined by applied mathematicians.

4. *Sankoff*

Most refinements introduced by applied mathematicians involved adding further parameters to the basic Swadesh model. For example, parameters were added to account for drift and for 'chance recurrent cognation' (i.e., when a word was replaced twice, such that the original word was restored), replacement rates were allowed to vary from meaning to meaning, a very crude allowance was made for borrowing, and there was talk of allowing the rates to vary from language to language and from time to time. (Further details, along with some models, can be found in Embleton 1986.) Some of these are useful refinements (e.g., allowing replacement and borrowing rates to vary over time, between languages, and between meanings), but these are very much a mathematician's way of solving the problem, producing quite elegant solutions. But in the real world in which the linguist works, all those parameters have to be estimated—and this is surprisingly difficult even for well studied languages for which we actually have both long and extensive written records. It is also the case that an extraordinary amount of data is required to estimate so many parameters, and this is most often simply impossible. So these elegant theoretical solutions have very little practical import.

However, the single most important development was the incorporation of a geographical dimension into the family tree, thus allowing for borrowing to be largely dealt with. The basic insight here is due to Sankoff (1972), with correction and further refinement by Embleton (see 1986 for the most comprehensive treatment). The resulting set of differential equations required to reconstruct the tree and estimate the branch lengths is complicated, reflecting the increased power of the method (equations are given in Embleton 1986 and elsewhere). The parameters required all have some reasonable chance of being estimated, even for lesser studied families; these are a similarity measure for each language pair, replacement rates for each language, borrowing rates between each language pair (not necessarily symmetric), and knowledge of which languages are in contact with one another.

5. *Starostin*

Largely independently of the work described above, research on numerical methods in historical linguistics was proceeding in the Soviet Union (later Russia, still later 'in diaspora'). Unlike the impetus for Swadesh's work, which was largely to aid in providing histories and trees for relatively little studied language families (e.g., Amerindian), much of this work gained its impetus from questions of more distant linguistic relationship. This had its effect on the development of the methods, because it was crucial that the methods be able to produce useful results even at much greater time depths, where Swadesh's method typically has difficulty distinguishing the signal (cognacy, true genetic relationship) from the noise (chance resemblance, contact relationship).

Historically the first advance here was due to Dolgopolsky (1986), who capitalized on the fact that different meanings change at different rates. He compiled a list of the 15 most stable meanings, based on examination of 140 languages. One can then calculate the probability of fortuitous coincidence of the words representing those 15 most stable meanings for any language pair; if the probability is significantly less than that expected by mere chance, this can be taken as evidence of relationship. This method is most striking in its extreme simplicity, but can also be related to some more complex and ongoing work, largely carried out in eastern Europe, founded on the observation that several characteristics of a word (e.g., length, number of meanings, age) are intercorrelated with its frequency. Words of more recent origin tend to be less frequent, longer, and have fewer meanings, whereas older words tend to be more frequent, shorter, and have more meanings. The age-frequency relationship in particular is being extensively researched, as there may be a small hard core of ancient short vocabulary which may reveal distant relationships, and may even be able to be harnessed into a lexicostatistical or even glottochronological method.

Another advance has been made by Sergei Starostin, with work first published in the 1980s and continuing to the present, known as root glottochronology or etymostatistics (see Starostin 1999). Compared to standard Swadesh-style glottochronology, the main innovation is the observation that while glottochronology regards any change (however small) in a word's meaning as a loss of that word from that meaning, root glottochronology does not. Thus while standard glottochronology sees English *dog* and German *Hund* in the meaning 'dog' as noncognate, root glottochronology notices the existence of the English word *hound*; this more conservative approach to the maintenance of a language's root stock enables one to work better at greater time depths. Starostin's method also gives more accurate results for known cases where the time depths are more shallow (standard glottochronology tends to give dates which are too recent).

Root glottochronology takes a text, containing a given set of words in the parent language, and then measures the number of these words still extant (in any derivative, with any semantic shift) in the daughter language. It proceeds only with detailed comparative-historical and etymological analysis (eliminating look-alikes and loans), and also can be repeated for more than one text (hence a different set of words), giving greater precision and potentially some estimate of error (see Starostin 1999 for formulas and more detailed explanation).

6. Ringe

Various linguists have used probabilistic methods to assess similarity between vocabularies of two different languages, usually focusing on phonological similarities (an early example is Ross 1950). The basic question is whether the observed degree of similarity between the two vocabularies is greater than what could be expected based on chance. Although the problem is conceptually simple, it quickly becomes mathematically intractable even between just two languages, as one has to wrestle with different phoneme distributions, different phoneme frequencies, how to measure phonetic distance between different phone(me)s, different phonotactic structures, and lengths of roots, quite apart from the questions (familiar from glottochronology) of how much semantic latitude to allow between the words compared. The most recent and sophisticated treatment of this problem is by Ringe (e.g., 1998 and references therein), who attempts to account for many of the above issues, albeit with simplifying assumptions. This research remains active, yet without consensus as to its utility, partly because of the degree of the simplifying assumptions, but also partly because it has mostly been applied in the realm of long-distance relationships, where the emotions over the results can sometimes cloud the objectivity with which the method is viewed.

7. Future Directions

Research in numerical methods in historical linguistics is gradually reentering the mainstream after several decades on the sidelines. The most active areas of current research are improvements to the basic Swadesh model, exploration of the age–frequency relationship, root glottochronology, computer-implemented tree reconstruction using methods borrowed from biological cladistics, and Ringe's probabilistic methods. The degree of mathematical, statistical, or computational expertise required is extremely high, which is likely to prevent the majority of historical linguists from ever mastering the methods, and consequently render them unable to do more than use results produced by others or participate in research teams where others control the methods. There seems to be a growing realization, however, that numerical methods, while not providing any magic solutions to any long-standing problems, can make a useful contribution as one of the tools that we use to explore our linguistic history and prehistory.

See also: Archaeology and the History of Languages; Historical Linguistics: Overview; Lexicology and Lexicography

Bibliography

Czekanowski J 1928 Na Marginesie Recenzji P. K. Moszyńskiego o Książce: Wstęp do Historji Slowian. *Lud*, Series II, Vol. 7. Reprint Lwow, 1928

Dolgopolsky A 1986 A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia. In: Shevoroshkin V, Markey T (eds.) *Typology, Relationship and Time: A Collection of Papers on Language Change and Relationship*. Karoma, Ann Arbor, MI, pp. 28–50

Embleton S M 1986 *Statistics in Historical Linguistics*. (*Quantitative Linguistics*, *Vol. 30*). Brockmeyer, Bochum, Germany

Gleason H A Jr 1959 Counting and calculating for historical reconstruction. *Anthropological Linguistics* **1**: 22–32

Hymes D H 1960 Lexicostatistics so far. *Current Anthropology* **1**: 3–44; 338–45

Klimek S 1935 The structure of California Indian culture, culture element distributions: 1. *University of California Publications in American archaeology and Ethnology* **37**: 1–70

Kroeber A L, Chrétien C D 1937 Quantitative classification of Indo-European languages. *Language* **13**: 83–103

Latham R G 1850 *The Natural History of the Varieties of Man*. Van Voorst, London

Ringe D R Jr 1998 Probabilistic evidence for Indo-Uralic. In: Salmons J C, Joseph B D (eds.) *Nostratic: Sifting the Evidence*. Benjamins, Amsterdam, pp. 153–97

Ross A S C 1950 Philological probability problems. *Journal of the Royal Statistical Society* B **12**: 19–59

Sankoff D 1972 Reconstructing the history and geography of an evolutionary tree. *American Mathematical Monthly* **79**: 596–603

Sapir E 1916 *Time Perspective in Aboriginal American Culture* (*Memoir 90,Anthropological Series*, *Vol. 13*). Government Printing Bureau, Ottawa, Canada

Starostin S 1999 Comparative-historical linguistics and lexicostatistics. In: Renfrew C, McMahon A, Trask L (eds.) *Time Depth in Historical Linguistics*. McDonald Institute for Archaeological Research, Cambridge, UK, Vol. 2, pp. 377–413

Swadesh M 1950 Salish internal relationships. *International Journal of American Linguistics* **16**: 157–67

Taylor A, Warnow T, Ringe D 2000 Character-based reconstruction of a linguistic cladogram. In: Smith J C, Bentley D (eds.) *Historical Linguistics 1995. Selected Papers from the 12th International Conference on Historical Linguistics, Manchester, August 1995, Vol. 1, General Issues and Non-Germanic Languages*. Benjamins, Amsterdam, pp. 393–408

S. Embleton

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Historical Linguistics: Overview

Historical linguistics is the study of language through time, and involves the consideration of changes from some earlier stage towards the present; the construction of earlier, unattested states; and the determination and depiction of relatedness among languages and language groups. Although the main focus of twentieth-century linguistics was synchronic, there is currently increasing interest in historical linguistics, both from linguistic theory and cognate historical disciplines, notably archaelogy and genetics.

1. *The Scope of Historical Linguistics*

Historical linguistics involves the study of languages as they change through time. It includes tracing developments from an earlier stage of a language to a later one: for instance, English 'silly' meant 'blessed' at the time of Chaucer; Latin sentences often had the word order Subject Object Verb, as in *Helena Marcum amat* 'Helena loves Marcus,' whereas its daughter language French typically has Subject Verb Object, as in *Hélène aime Marc*; and many Old English noun and verb endings with final nasal consonants, like *-an*, *-um*, *-on*, dropped gradually about 1,000 years ago. Historical linguists are interested in when, how, and ultimately why changes of this sort happened.

However, this diachronic perspective also operates in the opposite direction, encouraging historical linguists to look back from present-day or recorded languages to earlier, unattested stages in an attempt to reconstruct unrecorded features and characteristics. This may involve filling in gaps in the historical record, for instance in hypothesizing the form a word might have taken in Old English or in Latin, although that word does not actually appear in any surviving manuscript. At greater time depths, features of unattested, hypothetical ancestral languages, like Proto-Indo-European, the probable source of most languages of modern Europe and northern India, may also be reconstructed. Work of this kind also explicitly or implicitly involves consideration of language classification: reconstruction necessarily includes the hypothesis and testing of linguistic family groups, which are taken to be demonstrably descended from a single common source. At present, it is uncontroversial to say that English, French, and Latin are all related, as members of the Indo-European family; but it would be more controversial to contend, as some historical linguists do, that these languages are in turn related to Semitic languages like Arabic, Akkadian, and Hebrew; or that Chinese is related to Chechen and other Caucasian languages.

In this article, the development of historical linguistics will first be considered briefly, and its place within modern linguistics assessed. Then the two subdisciplines of language change, and linguistic classification and reconstruction, will be examined in turn, with the emphasis on outlining currently lively and productive areas of research. In the final section, some trends common to change and reconstruction will be identified, with their implications for future work in historical linguistics.

2. *Historical Linguistics Within Linguistics*

In the heyday of traditional work in the late nineteenth century, historical-comparative linguistics was linguistic theory. Its focus and methodology were in line with the strongly historicizing tendency of the day, and with the development of parallel sciences such as geology and evolutionary biology, so that although the foundations for language comparison and study had been laid considerably earlier, the discipline advanced rapidly and attracted considerable scholarly attention in this period (see Davies 1998). The cornerstones of the discipline were set with the classification and reconstruction of Indo-European in particular, and the Comparative Method on which such reconstructions are based (see Fox 1995) is still the foremost tool of the reconstructionist's trade, although many of the specific results obtained have been modified, as knowledge of language change has deepened and developed, and more data have become available. The Comparative Method requires sets of cognate forms in descendant languages: these will be similar in form and meaning, and will show regular, recurrent correspondences of particular sounds across the group of languages compared. We can then hypothesize that these have been derived historically from a single most plausible common source, by the operation of generally regular sound changes. Thus, given English *father*, Irish *athair*, Latin *pater* and Greek *pater*, alongside a large number of similar comparands which equally lack an initial consonant in Irish, and have English /f/ corresponding to Latin /p/, it seems reasonable to hypothesise Proto-Indo-European **pətér*. It is also possible to set out the contexts in which particular sound changes have operated in the daughter languages, and therefore to begin to identify the factors in the surrounding sounds which may have led to these developments (see *Linguistics*: *Comparative Method*).

Although many of these methodological and theoretical assumptions still hold good in historical linguistics today, the twentieth century has seen a radical shift in perspective in linguistics as a whole. The development of structuralism led to a move away from the study of individual languages and families, and towards identifying language universals, those aspects of human language which recur regardless of genetic linguistic affiliation. Consequently, the historical linguistic concentration on variation as a key to linguistic

relatedness gave way to attempts to identify theoretical limits on variation. Finally, whereas most earlier work had concentrated on vocabulary and phonology, the 'Chomskyan revolution' of the 1960s centered attention on syntax.

However, these rapid and far-reaching changes have not led to the demise of historical linguistics; although it is true that historical work became less of a core part of linguistics for some years in the mid-twentieth century, it is now again one of the liveliest areas of the discipline. This is for three main reasons. First, advances in linguistic theory, while typically based on synchronic study and data, often allow the sources of change to be identified and explained in a way which was not previously possible: much nineteenth-century work, although allowing individual elements to be plausibly reconstructed, was piecemeal and did not trace related changes to a common cause. Explanation comes center stage in modern linguistic work. Second historical data provide an additional means of testing and refining these linguistic theories, which may benefit from a demonstration that they can analyze and incorporate language change. Finally, related disciplines like genetics and archaeology may not always find developments in linguistic theory *per se* particularly useful or enlightening; on the other hand, interactions of data and hypotheses from these areas with those of historical linguistics are promising. All these areas will be considered, with respect to both change and reconstruction, in the following sections.

3. *Areas of Debate in Language Change*

3.1 *Phonology*

Much early work on sound change was based on the assumption of relative ease of articulation: that is, it is presumed to be more straightforward for a speaker to utter two similar adjacent sounds, or indeed to make them identical, than to produce a sequence of sounds which require major vocal tract movements between the two. This explanation for the extremely common sound change of assimilation, as seen in Latin *somnum* → Italian *sonno*, was put forward in the last century. However, advances in phonetic technology has allowed historical linguists to pinpoint more accurately the types of sound sequences which are likely to cause such problems, and the advent of experimental phonology has allowed these hypotheses to be tested, not only from the point of view of the speaker's production problems, but also taking the hearer's perception difficulties into account (see *Phonetics: Articulatory*). One of the pioneers in this area, John Ohala (1993), has suggested that certain sound changes like dissimilation, which increases the difference between close or adjacent sounds in the stream of speech, may be provoked by the hearer's expectations. Thus, a hearer receiving the Proto-Quechumaran word */t'ant'a/* 'bread' might erroneously have assumed that the speaker has produced two ejective [*t*']s close together by assimilating a second, nonejective stop to the first truly ejective one. Undoing this perceived but not real mistake by the speaker produces the modern Quechua form /t'anta/. Advances in phonological theory may also be relevant here, for instance in helping us understand what it means for sounds a and b to be closer in their constituent features than sounds b and c; or in showing that the apparently disparate environments triggering a particular sound change in fact form a unified class in terms of some deep principle of phonology. In this respect it is encouraging that, while certain phonological theories current in the 1980s, such as Government Phonology and some forms of Declarative Phonology, explicitly ruled historical processes out of their remit, Optimality Theory, a currently popular model which has been developing since the early 1990s, already includes a considerable body of work addressing problems of both change and variation.

3.2 *Morphology and Syntax*

Morphology has often been left out of historical linguistics as a separate domain of enquiry, being seen as simply an interface: phonological changes, like the loss of final unstressed syllables, may have morphological consequences if case or tense markers are lost, and this in turn may have a knock-on effect on the syntax, changing or fixing word order. However, morphology has more recently been the focus of attempts within Natural Morphology (Wurzel 1989) to identify those features which are favored universally, for instance in terms of ease of production, perception, or acquisition, and the factors underlying that naturalness, as well as assessing why in many cases these universally natural morphological features are subordinated to language-specific requirements which may reduce naturalness. Morphological change is also one component of grammaticalization, the extremely common process whereby independent words become grammatical markers, as in German *Drittel* 'third' and *Viertel* 'quarter,' formed from the numerals *drei* 'three' and *vier* 'four' plus a reduced form of *Teil* 'part'; or the famous French negative *ne … pas*, where *pas* is derived from Latin *passum* 'a step,' and became an integral part of the negative construction in earlier French, perhaps via gradual bleaching of reinforcing constructions meaning 'I won't go a step.' Grammaticalization has become a major focus of research in linguistic theory (see Hopper and Traugott 1993), introducing many students and researchers to issues in linguistic change (see *Grammaticalization*).

In syntax, historical work has become enormously more sophisticated over the last 30 years, reflecting the

newly ascendant status of syntax in synchronic work and in linguistic theory. Work at first may have been hampered by lack of knowledge of the range of syntactic changes, and uncertainty over quite fundamental questions, such as whether syntactic change can be seen as essentially regular in the same way as many sound changes. These issues are now being addressed by major cross-linguistic surveys (see Harris and Campbell 1995). On a more theoretical level, Lightfoot (1991, 1999) uses historical data and current models of syntactic theory in a mutually informing way, attempting to show that children learn syntax in the way they do because of particular factors in their innate, universal grammar, but that certain superficial changes in syntactic structures may also cause them to learn differently, setting up new structures which will form the basis for subsequent development in their language.

3.3 Semantics

Developments in linguistic theory may also increase understanding of semantic change, which has long been seen as a domain where description is the highest achievable goal. Earlier approaches have attempted to classify changes, but seeing developments as, for instance, extensions, or restrictions of meaning does not necessarily explain why these occur, and linguists have tended to fall back on the old adage that 'every word has its own history.' Developments in pragmatics and in cognitive linguistics may improve prospects here. For instance, Traugott (1982) suggests that many words develop from an external to an internal, subjective sense: thus, English *feel* begins as descriptive of some characteristic external to the speaker, in *It felt soft*, but comes to signal internal states, in *I feel miserable*. Sweetser (1990, p. 2) suggests that linguists must crucially adopt 'a cognitively based theory which takes not the objective "real world," but human perception and understanding of the world to be the basis for the structure of human language': this approach allows recurring semantic changes, such as those involving mind-as-body metaphors of the *grasp*, *cast light on*, and *see* type, to be unified.

3.4 Variation and Change

One of the most productive recent developments in historical linguistics has been the recognition of a connection between variation and change, and the development of sociohistorical research (see *Sociolinguistics*). First, this substantially increases the database of changes available for study, since changes in progress can now be considered alongside the more traditional category of change from earlier, written data to the present day. However, for spoken, colloquial data to be accepted, it is necessary to delimit and define what constitutes a change in progress, requiring new quantitative techniques for identifying and measuring ongoing change. Innovations, which may provide the seeds of future developments, must now be distinguished from changes, and innovators from early adopters, the speakers who begin to use a new variant and spread it to their own social group (see Milroy 1992, Labov 1994).

This sociohistorical work has also directed attention away in part from internal, systemic factors and towards external factors in the search for explanations for change. Internal factors may belong to the language system itself, or to some other level of the grammar, or involve the innate language faculty which many linguists see as unique to humans; such factors have typically been favored by theoretical linguists in their attempts to explain change. However, sociolinguistic work places great emphasis on the social characteristics of speaker and interlocutor, including age, sex, and social class, and on characteristics of the interaction itself, such as level of formality and topic of exchange. This in turn encourages focus, not only on interaction between speakers, but between varieties and languages. The consequences of language contact have therefore become vitally important in recent historical linguistics, with increasing interest in contact phenomena like borrowing and convergence; the development of pidgins and creoles; and sociolinguistically determined language breakdown and loss in language death (Bickerton 1981, Thomason and Kaufman 1988, Dorian 1989). These are no longer being seen as peripheral and uncommon, but as a central part of language change and germane to a greater understanding of language itself (see *Language Contact*; *Areal linguistics*; *Pidgin and Creole Languages*; *Language Obsolescence*).

4. Areas of Debate in Reconstruction and Classification

4.1 Typology and Reconstruction

In the closely related subfields of classification and reconstruction, matters have advanced considerably since the various nineteenth-century attempts to write a fable in Proto-Indo-European. There is still much interest in what is reconstructed: the reliability and realism of reconstructed systems remain open questions, as does the issue of whether syntax can be reconstructed using the same methods as phonology, or whether early, unattested areas such as word order are beyond recovery (Lass 1997). Innovations in phonological reconstruction continue to occur: for instance, the 'new look' of Proto-Indo-European proposes a set of ejectives alongside the more traditional voiceless and voiced stops, on the basis that the resulting system would be more common cross-

linguistically, even though the alleged ejectives have disappeared, largely without trace, in the daughter languages. This opens a wider debate over the roles of typology and uniformitarianism in reconstruction (see *Linguistic Typology*): that is, to what extent should reconstructions be driven by generalizations that hold for modern-day languages? Is it more important to reconstruct a proto-form or proto-system which accords with the comparative evidence from its daughters, but which might appear skewed when compared with other attested systems, or to produce a balanced, typologically normal system although considerable changes may be required to derive the daughter forms?

4.2 Classification and Methodology

However, classification has recently threatened to overtake reconstruction as the main focus of interest in this area, and here the questions are primarily methodological. The Comparative Method (see *Linguistics: Comparative Method*), with its concentration on regular, repeated correspondences and its incorporation, at least for many practitioners, of reconstruction as a test of the family groups proposed, has been challenged by Greenberg's much simpler method of multilateral comparison, based on straightforward surface inspection of word-lists from many languages. Family groups are then proposed on the basis of etymologies drawn from these lists: forms within an etymology are similar to an extent in form and meaning, but the similarities need not be recurrent, and reconstruction is not necessarily undertaken. These families tend to be considerably larger (and hence potentially older) than those conventionally reconstructed using the Comparative Method. For instance, Greenberg (1987) uses multilateral comparison to propose only three language families for the Americas, including Amerind, a phylum or superfamily containing more than 1,000 languages. Even more controversial connections are drawn by other practitioners of multilateral comparison, for instance between the Sino-Tibetan and Caucasian languages; and there have even been attempts to produce universal etymologies, including forms from all or most language superfamilies, which might suggest forms of Proto-World, the alleged first language of mankind (see Ruhlen 1994).

4.3 Connections with Archaeology and Genetics

Multilateral comparison remains extremely controversial. Its apparent ease of use may be deceptive, not compensating for problems of reliability of data, and potential laxness of the criteria governing what counts as a phonetic or semantic match. Many historical linguists are reluctant to accept the method or its results, and therefore doubt whether language groupings above the traditional family level, and beyond a time threshold of perhaps 5,000–8,000 years, can be conclusively established.

Long-range comparativists are, consequently, increasingly turning to data from outside linguistics itself in an attempt to validate their work, and exploration of the potential for historical linguistics and related disciplines like archaeology and genetics to produce mutually informing and enforcing data is becoming central. For instance, Greenberg et al. (1986) cite genetic and archaeological data which seem to favour a three-way migration of peoples into the Americas, mirroring Greenberg's tripartite classification of New World linguistic groups. Cavalli-Sforza et al. (1994), in work which caught the popular scientific imagination, analyzed genetic data from 142 indigenous populations worldwide, and argued that the resulting population tree closely matched a linguistic family tree including controversial linguistic phyla.

However, these correlations are themselves problematic. The lack of a one-to-one correspondence between language and archaeological culture means interpretation is all important, and increases the need for data from the fields under comparison to stand independently and to be subject to rigorous validation within their own areas before cross-disciplinary conclusions are drawn. In genetics, in particular, advances in the data-types which can be considered may produce a range of possible population histories rather than a single definitive answer: thus, Cavalli-Sforza et al.'s (1994) picture of the Americas, based on classical analysis of proteins, does not agree with the genetic groupings derived from more recent analysis of mitochondrial DNA and Y chromosomes; these will therefore produce population trees which need not match a particular linguistic tree.

Connections of historical linguistics with genetics and archaeology are also relevant beyond the domain of long-range comparison and influences move in both directions. For instance, Dixon (1997) applies the biological model of punctuated equilibrium to language, arguing that Comparative Method and the associated family tree model may be relevant only to periods of concentrated change and expansion in the histories of language families. At intervening periods of relative equilibrium, language contact and gradual diffusion of features would determine the course of language change. Geneticists and archaeologists are also increasingly seeking a historical linguistic methodology which is truly independent and repeatable, and which might transcend arguments among its practitioners, for comparison with their own data: the development of quantitative techniques of language comparison therefore remains a pressing goal within historical linguistics (see *Historical Linguistics: Numerical Methods*).

4.4 The Evolution of Language

If our attempts to reconstruct might be blocked beyond a certain time-depth on methodological grounds, some progress might be made by starting from the other end: that is, by considering the origin of language in our species. This is a topic which logically belongs with historical linguistics, at least from a nonspecialist point-of-view, but which has traditionally been excluded from the field, being famously banned by the Société de Linguistique de Paris in the 1860s. However, some rehabilitation has taken place during the 1990s, with increasing interchanges of ideas between historical linguists, computational linguists, primatologists, neurobiologists, and psychologists; Hurford, Studdert-Kennedy and Knight (1998), for instance, includes contributions on the possible emergence of phonology and syntax, with some use of computer simulations to show how learning of these new structures might have taken place, as well as work invoking more social and anthropological considerations to explain the rise of human language. Again, this is by definition interdisciplinary research, which raises questions fundamental to historical linguistics (such as how the, probably simple, initial constructions of early language developed into the modern range of variants), and to linguistics in general (such as the limits which might be set on what aspects of language are innately specified and which might be learned, this in turn depending in part on whether the underlying systems responsible for language evolved by natural selection or are in some way maladaptive).

5. Common Trends in Reconstruction and Change

In the sections above, change and reconstruction have been presented separately for convenience. However, these areas are not truly independent, and some common themes can be discerned which may be important for the future of historical linguistics in general.

First, it must be emphasized that change and reconstruction influence one another. For instance, if an early text or unknown language is deciphered, the database usable for drawing conclusions about changes is increased, but the same data may also confirm or refute the results of work on reconstruction.

Second, both subdisciplines need to become reconciled to the development and use of quantitative, computational approaches, and of tools. Work of this kind is probably more advanced in change, where techniques derived from sociolinguistics are already well established in some domains. The need is more pressing, and the issues involved more controversial, in classification.

Finally, it is encouraging to note that historical linguistics is becoming a major focus of research within linguistics. This means that historical linguists in turn must be willing to work with synchronically-based theories, and to consider what cognate disciplines like genetics and archaeology might require, as well as what they can provide.

See also: Archaeology and the History of Languages; Genes and Languages; Historical Linguistics: Numerical Methods; Languages: Genetic Classification; Linguistics: Internal Reconstruction

Bibliography

Bickerton D 1981 *Roots of Language*. Karoma, Ann Arbor, MI

Campbell L 1998 *Historical Linguistics: An Introduction*. Edinburgh University Press, Edinburgh, UK

Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

Davies A M 1998 Nineteenth century linguistics. In: Lepschy G (ed.) *History of Linguistics*. Longman, London, Vol. 4

Dixon R M W 1997 *The Rise and Fall of Languages*. Cambridge University Press, Cambridge, UK

Dorian N (ed.) 1989 *Investigating Obsolescence: Studies in Language Contraction and Death*. Cambridge University Press, Cambridge, UK

Fox A 1995 *Linguistic Reconstruction: An Introduction to Theory and Method*. Oxford University Press, Oxford, UK

Greenberg J 1987 *Language in the Americas*. Stanford University Press, Stanford, CA

Greenberg J, Turner C G, Zegura S L 1986 The settlement of the Americas: A comparison of the linguistic, dental and genetic evidence. *Current Anthropology* **27**: 477–97

Harris A C, Campbell L 1995 *Historical Syntax in Cross-Linguistic Perspective*. Cambridge University Press, Cambridge, UK

Hopper P J, Traugott E C 1993 *Grammaticalization*. Cambridge University Press, Cambridge, UK

Hurford J R, Studdert-Kennedy M, Knight C (eds.) 1998 *Approaches to the Evolution of Language*. Cambridge University Press, Cambridge, UK

Labov W 1994 *Principles of Linguistic Change. I: Internal Factors*. Blackwell, Oxford, UK

Lass R 1997 *Historical Linguistics and Language Change*. Cambridge University Press, Cambridge, UK

Lightfoot D 1991 *How to Set Parameters: Arguments from Language Change*. MIT Press, Cambridge, MA

Lightfoot D 1999 *The Development of Language: Acquisition, Change and Evolution*. Blackwell, Oxford, UK

McMahon A M S 1994 *Understanding Language Change*. Cambridge University Press, Cambridge, UK

Milroy J 1992 *Linguistic Variation and Change*. Blackwell, Oxford, UK

Ohala J 1993 The phonetics of sound change. In: Jones C (ed.) *Historical Linguistics: Problems and Perspectives*. Longman, London, pp. 237–78

Ruhlen M 1994 *On the Origin of Languages: Studies in Linguistic Taxonomy*. Stanford University Press, Stanford, CA

Sweetser E 1990 *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press, Cambridge, UK

Thomason S G, Kaufman T 1988 *Language Contact, Creolization and Genetic Linguistics*. University of California Press, Berkeley, CA

Trask R L 1996 *Historical Linguistics*. Arnold, London
Traugott E C 1982 From prepositional to textual and expressive meanings: Some semantic-pragmatic aspects of grammaticalization. In: Lehmann W P, Malkiel Y (eds.) *Perspectives on Historical Linguistics*. Benjamins, Amsterdam, pp. 245–71
Wurzel W U 1989 *Inflectional Morphology and Naturalness*. Kluwer, Dordrecht, The Netherlands

A. McMahon

Historical Sociology

In 1844, Auguste Comte's *Lecture on the Positivist Outlook* (*Discours sur l'esprit positif*) proposed the name 'sociology' for the general science of humanity. In Comte's vital vision, the hierarchy of sciences proceeded from mathematics at the base through astronomy, physics, chemistry, and biology to sociology at the summit. The maturing of basic sciences, he declared, now made it possible to construct the capstone science, sociology. In that defining moment for the sociological discipline, Comte's conception of sociology included history, in fact consisted largely of analyzing the development of humanity through historical stages. From that point onward, however, professional history and professional sociology moved in very different directions.

As the historical discipline organized later in the nineteenth century, professional historians came to specialize in particular times and places, for example, modern Latin America or ancient China. This specialization built up clusters of expertise in the sources, languages, and institutions peculiar to different time–place settings, which in turn promoted collective research programs, debate among experts, education of newcomers, and exclusion of nonprofessionals from the historical guild. It also made syntheses of all human history in a Comtean style less attractive to ordinary historians.

Meanwhile, sociologists divided up their own specialties mostly by structures and processes—families, religions, industrialization, crime, and so on—and drew their evidence chiefly from observation of the countries they currently lived in. Sociologists' historical essays divided among broad evolutionary and stage schemes, scattered examples illustrating general arguments, and sketchy prologues to studies of contemporary social phenomena. From these contrasting bases of intellectual organization sprang a number of mutual misunderstandings between historians and sociologists, including the sociologists' common presumption that sociology is an explanatory, generalizing science and history a descriptive, particularizing science fated to provide raw material for sociologists' generalizations. Historians returned the compliment by complaining that the present studied by sociologists was itself a narrow historical moment with no claim to universal significance, and that to pluck examples from history without thorough knowledge of the relevant sources, languages, and institutions courted intellectual disaster.

Despite increasing separation between the disciplines, to be sure, occasional thinkers since Comte have ventured into grand syntheses that do not obviously belong to a single discipline. The ideas and accomplishments of Karl Marx and Max Weber transcend any single discipline. Grand syntheses addressing major historical questions have continued to appear in the twentieth century, with inquiries by Barrington Moore, Jr., Norbert Elias, Immanuel Wallerstein, and William H. McNeill overflowing disciplinary containers. Such works belong to the common heritage of history and the social sciences rather than to the narrower field of historical sociology.

The idea of a distinct enterprise called historical sociology only formed during the twentieth century, as sociologists themselves began to distinguish sharply between evidence coming from direct observation of the present and from indirect observation of the past. Historical sociologists sought to connect the two through serious investigation of other settings than the ones they lived in. During the twentieth century, historical sociology has differentiated into four rather different pursuits. We might call them *social criticism*, *pattern identification*, *scope extension*, and *process analysis*. All are legitimate, effective forms of sociological analysis. They differ substantially, however, in aims, procedures, and results.

Historical social criticism reconstructs the past on the way to informing human choices in the present and future. It rests on the plausible assumptions that history contains a record of human successes and failures in dealing with problems persisting into the present; that if long-term trends exist, their causes are likely to continue; and that recent history limits what will happen next. Thus sociologists have created competing accounts of capitalism's development as liberation, as oppression, or as something in between. Similarly, sociologists of the family review its history as a means of criticizing current trends and charting alternative futures.

Although the description and criticism of contemporary social conditions preoccupied many nineteenth century sociologists, a large share of nineteenth century sociology consisted of historical social criticism—attempts to make better sense of current dilemmas and general directions of human development by situating the present in long, large social processes. Landmarks of nineteenth century social thought such as Karl Marx's *Capital*, Max Weber's *Economy and Society*, or Ferdinand Tönnies' *Gemeinschaft und Gesellschaft* embedded critiques of contemporary social life in vast historical syntheses. The tradition continues into our own time, as sociologists in the

varied styles of Göran Therborn, Theda Skocpol, Jürgen Habermas, and Orlando Patterson contribute to current social criticism by reworking history. At its best, historical social criticism helps people act wisely and effectively in the present by displaying possible futures and paths to them.

Historical social criticism overlaps with historical pattern identification, which it helped bring into being as a sociological specialty. Pattern identification searches for recurrent structures and sequences across time and space. Before World War II pattern identification often involved natural histories for large social aggregates, for example, Pitirim Sorokin's claim that each civilization passes through Ideational, Idealistic, and Sensate phases. (Sorokin's analyses edged over into social criticism as he argued that western civilization had reached the last, corrupt stage of sensate culture, and that only spiritual renewal could save the West.) Pattern identification also took the form of natural histories for wars, revolutions, social movements, and other smaller-scale phenomena.

After World War II, ideas of development and modernization sometimes followed the natural history model, but pattern-identifying historical sociologists generally moved toward the study of processes within countries or regions such as secularization, industrialization, revolution, democratization, and population growth. Pattern identifiers usually seek to discover necessary and sufficient conditions for large, recurrent social transformations and/or to establish general sequences within those transformations. Much pattern identification consists of lining up comparable cases to determine whether they conform to the proposed pattern. Thus students of state formation as different in approach as Reinhard Bendix, Michael Mann, and Bernard Silberman collect historical evidence on central bureaucracies in multiple countries to see whether state bureaucratization regularly occurs in the course of economic expansion. At its best, historical pattern identification facilitates recognition of recurrent social processes with sufficient accuracy to permit intervention or anticipatory action.

Scope extension applies techniques, models, or generalizations that sociologists have developed in studies of contemporary social life to historical situations. Demographers, for example, have often verified or refined models of fertility change developed initially to explain differences or short-range shifts in contemporary populations by examining longer runs of historical experience. Similarly, network analysts have sometimes extended ideas and techniques from the study of connections within currently observable populations to the reconstruction of changing connections within historical populations. Scope extension to historical cases can remain quite ahistorical or even antihistorical if the analyst disregards the time–space contexts of the social processes under investigation. But it also offers grounds for extensive collaboration between historians and sociologists, as well has for adaptation of sociological ideas by historians. Historians have, for example, conducted valuable investigations of social mobility in past times, examining evidence and time spans rarely available to their sociological cousins.

Within disciplinary history, in fact, the special fields of social history, family history, demographic history, and economic history have grown up as interfaces with the social sciences, including sociology (Zunz 1985). In these fields, some work actually consists of sociological scope extension or its equivalent in relation to economics and political science. More takes up questions of general concern to historians by means of models, theories, or techniques grounded in social science, for example, using collective biography to trace changes in composition of the English gentry (Stone and Fawtier Stone 1984).

Scope extension has a mirror-image version. It consists of using historical evidence to challenge widespread suppositions of contemporary sociological analysis. The historical work of Viviana Zelizer offers an excellent case in point. Although her later writings about the social meanings of money balance between scope extension and process analysis, her treatments of life insurance and valuation of children in the United States during the nineteenth and twentieth centuries demonstrate that economic transactions commonly portrayed as operating independently of culture—or even to the detriment of culture—actually rest on extensive, changing webs of shared understanding. It took extensive cultural work, for example, to make the insuring of human lives acceptable to Americans. At its best, scope extension accomplishes two distinct objectives: it corrects historical misconceptions as it broadens the domain of social scientific knowledge.

Process analysis examines how social interactions impinge on each other in space and time. Instead of considering space and time as additional variables, it presumes that space–time connections define social processes, and that social processes operate differently as a function of their placement in space and time. Sociologists studying urbanization, for example, have long sought to identify recurrent sequences and geographic arrangements of urban growth (pattern identification) and to determine whether explanations of urban growth that work reasonably well in contemporary capitalist countries also apply in noncapitalist history (scope extension). But they have also tried to determine how, why, and under what conditions large cities host highly differentiated occupations (process analysis). Although theorists have sometimes drawn their ideas on the subject from biological ecology or from the comparison of contemporary populations, this sort of process analysis is intrinsically historical. The analyst has no choice but to locate the processes under examination in time and space, then to ask how time and space affect them. Process analysts typically deny the premise of pattern

identifiers: that large structures and processes repeat themselves in essentially the same form across history. Process analysts search instead for explanations of variation.

A bold example comes from Janet Abu-Lughod's study (1989) of the thirteenth century world system. Reacting to Immanuel Wallerstein's claim that under European capitalist auspices the first economic world system formed during the fifteenth century and has prevailed ever since, Abu-Lughod examined connections among such far-flung cities as Baghdad, Calicut, Canton, and Malacca three centuries earlier. She offered evidence of an active trading system, connected by both land and sea, of which Europe was then a peripheral appendage. Although Abu-Lughod's analysis centers on the same sorts of mechanisms—generation of interaction by trade in high value-for-bulk commodities—as Wallerstein's, it resets the questions. Abu-Lughod forces us to ask how and why a vast Asian-centered world system declined and a European-dominated system arose in its place. At its best, process analysis identifies causal mechanisms of broad scope as well as conditions that affect activation, interaction, and outcomes of those mechanisms (Hedström and Swedberg 1998).

Social criticism, pattern identification, scope extension, and process analysis take somewhat different approaches to historical evidence. Social criticism relies almost entirely on selective use of existing historical accounts, often privileging earlier social criticism and social thought. Thus Jürgen Habermas' *Theory of Communicative Action* reinterprets modern western history (especially that of Germany) via critiques of Karl Marx, Max Weber, Emile Durkheim, Talcott Parsons, and other thinkers, arriving at proposals for a critical theory to serve human liberation. In a more obviously historical vein, Manuel Castells' *The Information Age* reviews changing information systems and identities in the world at large during the later twentieth century in order to identify choices and challenges facing today's humanity. Castells' sources include statistical compilations and his own observations, but beyond general theory and the immediate past they center on works by historians and historical sociologists.

By and large, pattern identifiers draw their evidence from historians' syntheses, extracting comparable elements from the detail. In effect, they apply a uniform questionnaire to distinct historical episodes. Robert Wuthnow's *Communities of Discourse* (1989) asks how ideological innovations take root by comparing the Protestant Reformation, the European Enlightment, and the formation of Marxist socialism. His implicit questionnaire's broadest headings concern environmental conditions (e.g., pace of economic growth), institutional contexts (e.g., extent of division within ruling elites), and action sequences (e.g., communication of new ideologies to broad audiences). Wuthnow draws evidence under these headings from wide reading of relevant histories, refashioning the evidence to establish uniformities, for example, the regular emergence not of a unitary belief but of competing ideologies, from which a political selection process singles out one or a few for predominance. Such implicit questionnaires discipline inquiry, but they also run two serious risks: of treating as equivalent events, institutions, and social relations that actually differ fundamentally, and of ignoring the causal connections of events, institutions, and social relations within their actual historical settings.

Among historians, authors of textbooks and of grand narratives often proceed in a manner superficially resembling that of sociology's pattern identifiers; they incorporate other historians' detailed studies into large syntheses. Historians' syntheses, however, typically hang together not through the detection of similar elements in distinct episodes but through the establishment of connections *among* episodes. Historians' connections arise in different ways: from evolution of political or economic systems, from the impact of leaders, from transformation of the dominant culture, or perhaps from critical events such as creation of the first Chinese empires or the Russian revolutions of 1917. But on the whole historical synthesizers, unlike their sociological cousins, insist on connections in space and time.

Within historical sociology, extensions of scope often employ procedures and sources resembling those of historical researchers. Historians make a rough distinction between *primary* and *secondary* sources. If the historical processes under study directly generated the sources, they qualify as primary; administrative correspondence, triumphal arches, artisans' tools, and contemporary portraits can all serve as primary sources. Secondary sources include later efforts to reconstruct what happened, especially efforts by other historians. The distinction must remain blurred, since chronicles, autobiographies, ambassadors' dispatches to their home countries, newspaper accounts, and similar sources occupy a middle ground.

Sociological scope extenders characteristically become familiar with the character of primary sources in their chosen historical eras even when their own analyses rely chiefly on secondary work by specialist historians. Thus historical demographers have frequently compiled and analyzed registers of births, deaths, and marriages for themselves in search of clues concerning determinants of changes in fertility, mortality, or nuptiality (Willigan and Lynch 1982). As historical demography illustrates, sociologists working with a recognized body of historical sources are more likely than their historical brethren to center their analyses on models, methods, or arguments drawn directly from contemporary studies of ostensibly similar phenomena.

Not all sociological scope extenders, however, apply formal models to the past. In common with their cousins in anthropology, for example, sociologists

with historical orientations have often engaged in retrospective ethnography, taking the questions and procedures of local community studies back in time. In a pioneering work of historical sociology, *English Villagers of the Thirteenth Century* (1941), George Caspar Homans reconstructed the organization of different sorts of villages much as contemporary observers of communities were reconstructing the organization of villages in the Americas, Europe, and elsewhere. Homans substituted historical records for the interviews and first hand observation of community ethnographers. Don Kalb (1997) has adopted retrospective ethnography to follow transformations of class, family structure, and politics in industrial North Brabant during the nineteenth and twentieth centuries. Both Homans and Kalb made simultaneous contributions to sociologists' and historians' treatments of their subject matters.

Of our four varieties of historical sociology, process analyses engage sociologists most directly in thinking about history as history, that is, as the embedding of human action in time and space. Here the distinction between primary and secondary sources begins to dissolve, and the sources rarely come ready made. Reconstructing processes by which science changes, for example, engages historical sociologists of science in looking simultaneously at organizational settings, individual biographies, interpersonal networks, contested bodies of thought, and connections among all of them. No simple exportation of contemporary models into the past will accomplish that sort of synthesis.

Process analysts concentrate on connections among events, which means that, whether explicitly or implicitly, they typically construct matrices having multiple dimensions: time, place, type of event, form of connection, and so on. Wherever they find their historical evidence, then, process analysts snip, splice, paste, and rearrange it extensively. John Markoff's *Abolition of Feudalism*, a case in point, recasts local statements of grievances from the France of 1789 into regional and class distributions of demands, then integrates a vast catalog of peasant actions into an analysis of popular mobilization as a dynamic process during the early Revolution. Similarly, Roberto Franzosi pursues a search for causes of industrial conflict by threading his way among aggregate Italian strike data between 1950 and 1980, increasingly refined econometric models, and close examination of shifts in institutional and political settings. In both cases, the analysts ended up showing that when and where political processes occurred significantly affected *how* they occurred, and what results they produced.

As sociology and history now operate, relatively little flow occurs between sociologists' historical social criticism and adjacent fields of historical endeavor. The same is generally true of historical pattern identification. Neither endeavor attracts much attention from historians. Nor do sociologists respond readily to historians' efforts along similar lines. More ideas flow in both directions across the boundaries of scope extension, but the chief effect of collaboration between sociologists and historians at that interface has been to create a series of specialized subdisciplines: historical demography, demographic history, and their equivalents. As a result, sociologists and historians at large remain woefully ignorant of each other's sources, methods, models, ideas, and discoveries.

In the zone of process analysis, however, we find intense interaction between historians and sociologists. We also find sharp disagreement. Questions of epistemology, ontology, and method align practitioners with competing answers to such questions as 'What is an event?,' 'Can we detect causes in history?,' 'Do all social processes result from individual choices?,' and 'Do organizations really exist?' Debates over such issues roil current discussions of state formation, population change, revolution, capitalist development, and other large processes (see, e.g., Abbott 1994, 1998, Bates et al. 1998, Hamilton 1996, McDonald 1996, Steinmetz 1999). That is as it should be; where relations among historical researchers run harmoniously, we should fear that they have worked out a comfortable but ultimately fruitless division of labor. We can take some satisfaction from noticing that the deep epistemological, ontological, and methodological divisions of our time do not pit historians against sociologists, but cut across the once-forbidding frontier between history and sociology.

See also: Anthropology and History; Comparative History; Comparative Method, in Evolutionary Studies; Comte, Auguste (1798–1857); Historical Archaeology; Historical Demography; Historicism; History and the Social Sciences; Modernization and Modernity in History; Social Evolution: Overview; Spencer, Herbert (1820–1903); Theory: Sociological; Weber, Max (1864–1920)

Bibliography

Abbott A 1994 History and sociology: the lost synthesis. In: Monkkonen E (ed.) *Engaging the Past. The Uses of History Across the Social Sciences*. Duke University Press, Durham, NC

Abbott A 1998 The causal devolution. *Sociological Methods & Research*. **27**: 148–81

Abu-Lughod JL 1989 *Before European Hegemony: The World System*. Oxford University Press, New York

Bates R H, Greif A, Levi M, Rosenthal J-L, Weingast B 1998 *Analytical Narratives*. Princeton University Press, Princeton, NJ

Bendix R 1964 *Nation-Building and Citizenship: Studies of our Changing Social Order*. Wiley, New York

Castells MI 1996–98 *The Information Age: Economy, Society, and Culture*. Blackwell, Malden, MA, 3 Vols.

Comte A 1963 *Discours sur l'esprit positif*. Union Générale d'Editions, Paris [First published in 1844]

Elias N 1978, 1982 *The Civilizing Process*. Pantheon, New York, 2 Vols. [German edition first published in 1939]
Franzosi R 1995 *The Puzzle of Strikes: Class and State Strategies in Postwar Italy*. Cambridge University Press, Cambridge, UK
Habermas J 1984, 1987 *The Theory of Communicative Action*. Beacon, Boston, MA, 2 Vols. [German original first published in 1981]
Hamilton R F 1996 *The Social Misconstruction of Reality*. Yale University Press, New Haven, CT
Hedström Pr, Swedberg R (eds.) 1998 *Social Mechanisms. An Analytical Approach to Social Theory*. Cambridge University Press, Cambridge, UK
Homans G C 1941 *English Villagers of the Thirteenth Century*. Harvard University Press, Cambridge, MA
Kalb D 1997 *Expanding Class. Power and Everyday Politics in Industrial Communities, The Netherlands, 1850–1950*. Duke University Press, Durham, NC
Mann M 1986, 1993 *The Sources of Social Power I. A History of Power from the Beginning to A.D. 1760; II. The Rise of Classes and Nation-States, 1760–1914*. Cambridge University Press, Cambridge, UK
Markoff J 1996 *The Abolition of Feudalism: Peasants, Lords, and Legislators in the French Revolution*. Pennsylvania State University Press, University Park, PA
Marx K 1970–72 *Capital: A Critique of Political Economy*. Lawrence & Wishart, London, 3 Vols. [German version first published in 1867–94]
McDonald T J (ed.) 1996 *The Historic Turn in the Human Sciences*. University of Michigan Press, Ann Arbor, MI
McNeill W H 1982 *The Pursuit of Power: Technology, Armed Force and Society Since A.D. 1000*. University of Chicago Press, Chicago
Moore B Jr 1966 *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon Press, Boston
Moore B Jr 1978 *Injustice. The Social Bases of Obedience and Revolt*. M.E. Sharpe, White Plains, NY
Patterson O 1991 *Freedom in the Making of Western Culture*. Basic Books, New York
Silberman B S 1993 *Cages of Reason: The Rise of the Rational State in France, Japan, the United States, and Great Britain*. University of Chicago Press, Chicago
Skocpol T 1979 *States and Social Revolutions. A Comparative Analysis of France, Russia, and China*. Cambridge University Press, Cambridge, UK
Skocpol T 1992 *Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States*. Harvard University Press, Cambridge, MA
Skocpol T (ed.) 1998 *Democracy, Revolution, and History*. Cornell University Press, Ithaca, NY
Sorokin P A 1962 *Social and Cultural Dynamics*. Bedminster Press, New York, 4 Vols. [First published 1937–41]
Steinmetz G (ed.) 1999 *State/Culture. State-Formation after the Cultural Turn*. Cornell University Press, Ithaca, NY
Stone L, Fawtier Stone J C 1984 *An Open Elite? England 1540–1880*. Clarendon Press, Oxford, UK
Therborn G 1995 *European Modernity and Beyond: The Trajectory of European Societies, 1945–2000*. Sage, London
Tönnies F 1957 *Community and Society*. Michigan State University Press, East Lansing, MI [German version first published in 1887]
Wallerstein I M 1974–89 *The Modern World System*. Academic Press, New York, 3 Vols.
Weber M 1968. In: Roth G, Wittich C (eds.) *Economy and Society: an Outline of Interpretive Sociology*. Bedminster Press, New York, 3 Vols. [German version first published in 1922]
Willigan J D, Lynch K A 1982 *Sources and Methods of Historical Demography*. Academic Press, New York
Wuthnow R 1989 *Communities of Discourse: Ideology and Social Structure in the Reformation, the Enlightenment, and European Socialism*. Harvard University Press, Cambridge, MA
Zelizer V A 1979 *Morals and Markets: The Development of Life Insurance in the United States*. Columbia University Press, New York
Zelizer V A 1985 *Pricing the Priceless Child: The Changing Social Value of Children*. Basic Books, New York
Zelizer V A 1994 *The Social Meaning of Money*. Basic Books, New York
Zunz O (ed.) 1985 *Reliving the Past. The Worlds of Social History*. University of North Carolina Press, Chapel Hill, NC

C. Tilly

Historicism

1. Origin and Meaning of the Term

The term 'historicism' comes from the German Historismus, and first appeared in German as far back as the closing years of the eighteenth century. It was only imported into other European languages—French, Italian, English, Spanish, and so on—much later, from the 1930s. In German the term is to be found (albeit only occasionally) in writers such as Novalis and Friedrich Schlegel, and in the nineteenth century, in spite of having an uncertain definition, it came to be used quite widely, assuming a wide variety of meanings, most of them negative. On the whole it was used to refer to the conception of history and society worked out by the German historical school—in its writings on law and (especially) on the economy. Thus for example Carl Menger entitled one of his works 'Die Irrtümer des Historismus,' to criticize the reduction of economic theory to economic history, or at least the confusion of the two at the hands of the historical school of economics (see *Political Economy, History of*). A few decades later, Edmund Husserl, in his essay on 'Die Philosophie als strenge Wissenschaft' (1911), argued that naturalism and historicism were two parallel forms of thought which reduced philosophy to a Weltanschauung. Even when the word was used more neutrally—as occurred in the early years of the twentieth century, in writers like Ernst Troeltsch or Friedrich Meinecke or, in Italy, Benedetto Croce—its connotation was by no means univocal. It is no accident, therefore, that it was used mainly to refer to movements and theoretical approaches belonging to the preceding centuries.

This use of the word primarily to refer to the past explains why it was employed to refer to different things: (a) the conception of history present in the German historical school; (b) the positions of (often widely differing) writers such as Wilhelm Dilthey, Georg Simmel, Ernst Troeltsch, Friedrich Meinecke, and (to some extent) Max Weber; (c) the conception of history found in Hegel or Marx (this is the use Popper makes of the term); and (d) (a usage confined to Italy) a speculative tradition going back to Vico and Hegel, and leading up to Benedetto Croce. Up until the years between the wars, there were no writers who described themselves as 'historicist,' and up until that time, the word had rather negative connotations. The first attempt to provide a thoroughgoing overhaul and redefinition of historicism as a historical movement was Friedrich Meinecke's *Entstehung des Historismus* (1936, trans. *Historicism*, London 1972), a work which, however, produced more criticism than consensus (see *Historiography and Historical Thought: Modern History* (*Since the Eighteenth Century*)).

2. Historicism and the Historical School

There is no doubt, however, that historicism is not just a historiographical tendency—even though it has certainly had widespread influence on historians—for it is primarily a general conception of reality, or at least of Man and the human world, a conception whose roots are to be found in late eighteenth-century German culture. (The inclusion of Vico as a historicist is a retrospective operation, and one that is highly controversial.) Historicism in fact emerged in close connection with that reaction against French culture that started to make its voice heard clearly with Johann Gottfried Herder. In *Auch eine Philosophie der Geschichte zur Bildung der Menschheit* (1774), Herder argued against views of history which saw it as the progress of humanity out of a savage state towards civilization, but espoused a developmental perspective which maintained there was an analogy between the life of humankind and the life of an individual. According to Herder the various epochs that came after one another—seen as paralleling the gradual passage from childhood to youth and then virile manhood—correspond to the succession of peoples. Each historical epoch was said to bear the imprint of the national character of a particular people, which was destined, in the same way as humanity as a whole, to run through a life cycle. Herder linked up this organicist perspective with the idea that history was the working out of the design of Providence. Thus each people was seen as carrying out an indispensable task in the history of humanity, imparting in the process the mark of its peculiar nature—its 'spirit'. The idea of 'a people's spirit' (Volkgeist) therefore came to mean the essence of a particular people, the fundamental traits which underlay and united all the forms that people took in the course of its development.

It was precisely this idea which was taken up by the historical school. Whereas Hegelian philosophy of history developed Herder's providential perspective, and tried to show the essential rationality of history, demonstrating that underneath the apparently accidental succession of events was the development of spirit, Friedrich Carl von Savigny interpreted law as the individual creation of a particular people—thus rejecting the natural law tradition. In a similar way, Wilhelm Roscher and other exponents of the historical school among economists saw economic life as the expression of a particular people, and thus a pattern which could not be understood independently from other aspects of that people. Most influentially of all, Leopold von Ranke gave a theoretical framework to romantic historicism in a series of works—from *Geschichten der romanischen und germanischen Völker von 1494 bis 1535* (1824) to his unfinished *Weltgeschichte* (1881–5). Rather than seeing each people as having an essence, Ranke saw each epoch as possessing its own individual character. He believed each age was characterized by dominant 'ideal' tendencies, and that the people who lived in a particular age were shaped by these tendencies. He denied that it was possible to interpret the succession of epochs as having any teleological development, and denied that earlier ages could be seen as laying the way for later ones, or later ones the necessary outcomes of earlier epochs. Each epoch has its own value, and is, as Ranke put it, in an immediate relationship with God. The divine presence in history did not imply, therefore (as Hegel believed) that there was a 'progressive' evolution from one age to another.

A common thread in Romantic historicism was the belief that all knowledge about Man, society and the state must be historically based. Historicism thus put forward an alternative approach to that which had become dominant in the second half of the eighteenth century—exemplified in writers such as Montesquieu or Adam Smith—which took modern natural science as the pattern to follow. Its epistemological model not only implied that historical knowledge was different from knowledge of natural phenomena, but also insisted on the need to bring all disciplines studying Man back to a historical basis. This meant that knowledge of this kind was inevitably 'partial', for it took objects—economic life, say, or law, or the organization of the state—which could only be abstractly isolated from the historical process. Only historical knowledge could lead to understanding of the historical process, it was claimed, or understanding of any epoch in its totality. And only historical knowledge could reveal the deep underlying meaning of events and of the relationship between them. The historicism of the historical school thus remained faithful to the organicist framework it inherited from

Herder. Indeed, it ensured that a shadow of divinity fell upon history—since it was only God who could give a sense to humanity's journey.

3. *Modern Historicism*

In the last two decades of the nineteenth century a decisive turn took place in historicism. There was an encounter with the neo-Kantian movement and it was felt that what were called the 'sciences of the spirit' or 'sciences of culture' needed to be given an epistemological foundation. This turn started with Dilthey's *Einleitung in die Geisteswissenschaften* (1883). Dilthey positioned himself in the tradition of the historical school, but he abandoned its organicist assumptions. Still more importantly, he rejected any attempt to use history as the basis for a metaphysical perspective which saw history as containing a deus ex machina—whether this consisted of a transcendent will (as in Christian ideas) or an immanent one (as in Hegel). Dilthey tried to apply the kind of critique that Kant had made of natural science to the sciences of the spirit. He therefore thought it was necessary to supplement the critique of pure reason with a 'critique of historical reason'—a critique which would 'provide foundations' for the disciplines concerned with Man, society and the state. However, this enlargement of the critical enterprise demanded a shift in approach, according to Dilthey, to permit a non-metaphysical nalysis of the historical world.

A kind of historicism thus began to emerge which was different from, and in many ways opposed to, that of the Romantic period. This new historicism was active in two fields—in epistemology and methodology, and in analysis of the structure of the historical world. In both fields work was produced which was often very diverse. Dilthey was more tied than were other new historicist writers to the historical school, and he saw knowledge of history as part of the *Geisteswissenschaften* (seen as including psychology and the various social sciences as well as history). He thus tried to identify what it was about knowledge of this kind that made it different from knowledge in the natural sciences. He saw the natural and human sciences as differing both in their object and in their method. Whereas natural sciences took as their object phenomena which belonged to a world outside human beings, and were rooted in external experience, in the sciences of the spirit the knowing subject and the object belonged to the same world; knowledge thus came from the lived experience knowing subjects had of their own inner states. Dilthey took up Droysen's idea that historical knowledge was 'understanding' (Verstehen) rather than explanation (the kind of knowledge which could be achieved in the natural sciences). Other writers, however, such as Wilhelm Windelband and his pupil Heinrich Rickert, who were more directly influenced by the neo-Kantian movement, took up Ranke's idea that history is a search for individual identities (the identity of an age) and transposed it into the theory that history was a kind of knowledge oriented towards the individual—as against natural science, which is the search for general laws. Still others, such as Simmel, focused on the social sciences and especially sociology, and posed a problem which was analogous to Kant's problem regarding how nature was possible—asking how society was possible.

These various positions are related to different approaches to the structure of the world of history. Dilthey developed the critique of 'historical reason' into a historical critique of reason, and thus came to a theory of historicity which later influenced Martin Heidegger's *Sein und Zeit*. The historical world—that is to say, systems of culture, systems of social organization, historical epochs—was seen by Dilthey as the outcome of a process whereby life was objectified (became 'objective spirit'). Yet life as an attribute specific to human beings was itself a process which evolved in time. Man is thus a historical being, and his life is historically conditioned by the tendencies of a particular epoch. This led Dilthey to deny that there could be any absolute values—a position which was quite different from that of Windelband and Rickert, who took the neo-Kantian movement in the direction of a theory of values. Dilthey's denial of the possibility of absolute values was echoed on the other hand by Simmel, who produced a psychological interpretation of the Kantian *a priori* and asserted that knowledge was just as relative as other forms of life. After 1910 Simmel came to espouse an immanent-type philosophy of life, founded on the dialectic between life, 'more-life' and 'more-than-life.'

In the period between the two world wars, German historicism again changed track and returned to the metaphysical perspectives of Romantic historicism. The work of Oswald Spengler, published just after the First World War, stressed the radically relativistic implications of historicism, and portrayed history as the terrain in which heterogeneous cultures played themselves out—each one carrying its own set of values, which were peculiar only to the people living within the particular culture in question. Spengler was influenced by Goethe and Nietzsche: from Goethe he adopted the idea of the opposition between the 'world of nature' and 'the world of history,' and also the analogy between cultures and organisms, and he drew on Nietzsche to make a distinction between culture and civilization (Zivilisation), with civilization being seen as the inevitable decline of any culture (see *Civilization, Concept and History of*). Troeltsch and Meinecke opposed the relativism of Spengler and tried to re-assert the absolute nature of values from within the historical process itself. Troeltsch advocated a 'material' philosophy of history that would create a 'cultural synthesis' between the various national components of European civilization. Meinecke went

back to Ranke's work and investigated the relationship between the structure of the historical process and God. For both writers, the fact that values were situated in history did not imply that they could not be transcendent, and thus escape becoming outmoded by the course of historical development or becoming.

4. Beyond Historicism

In the first two decades of the twentieth century Max Weber took another path, going 'beyond historicism' as Wolfgang Mommsen put it (see *Weber, Max (1864–1920)*). Weber broke decisively with Romantic historicism, took up Menger's critique of the historical school in economics and rejected the organicist assumptions of so much historicism. Weber did not draw on Dilthey—who was still linked to the historical school—but rather on Rickert's distinction between historical knowledge and natural science. For Weber, however, historical knowledge consisted of a set of disciplines that needed not only the explanation of historical phenomena but also what he called nomological knowledge. The focus on individual phenomena therefore in no way meant that general concepts and laws could not be used: these are, indeed, essential if events and the relations between events are to be explained. These general concepts and laws differed from those of natural science in that they took the form of ideal-types. However it was they that made up the basis of 'pure' economics and sociology—these being seen as generalizing disciplines which tended to distinguish them from historical research proper. Weber's position was similar to, but more rigorous than, Simmel's. Both wished to restore the relationship with the social sciences which the emphasis on the individual nature of historical phenomena (the inheritance of Romantic historicism) seemed to imply was impossible.

Weber also broke with Rickert's theory of values on another terrain—that of the interpretation of the status of values. Rickert had sought to ground historical knowledge in universal values, which were transcendent and independent of the course of history. For Weber, on the other hand, values were the criteria for choosing from the infinite range of empirical data—and thus a criterion whereby a historical object was constituted. Far from being absolute, therefore, values possessed a validity that coincided with the possibility of their finding concrete form in human action. Weber also denied that values were organized in systems, as Rickert believed, for values are multiple, as are the spheres of values. Between the various value spheres, and between individual values within each sphere, there was 'collision' of values. This is true even within the sphere of ethics, which is dominated by the split between ethics of intention or conviction (Gesinnungsethik) and ethics of responsibility (Verantwortungsethik). Given that there is a plurality of conflicting values, Weber argued that people have to choose; and choosing certain values inevitably means rejecting others. For Weber, history was thus the stage on which people positioned themselves vis-à-vis values, adopting certain values as criteria in their own action, thus putting them into effect.

5. Historicism outside German culture

Like that of Troeltsch and Meinecke, Weber's thought was cast in the framework of modern historicism. This was not the case for Croce, who was in intellectual dialogue with Hegel's philosophy of history—although he also put forward a 'reform' of the Hegelian dialectic. Croce denied the existence of a nature existing outside spirit—so even ideas are included in spirit. So reality as a whole was identified with spirit—and this, in turn, was identified with history. In epistemological terms, knowledge was seen as historical knowledge. The sciences (whether abstract sciences such as mathematics or political economy, or empirical sciences, such as the natural sciences) were denied any cognitive validity: Croce argued they simply had an 'economic' function. Croce termed his thought 'absolute historicism'—that is to say, it asserted the thoroughly historical character of reality and knowledge as a whole. Croce's thought was absolute in another, perhaps deeper, sense. For Croce saw history as the 'unfolding' of a first principle, spirit, which becomes real by moving from one to another of its eternal forms—whether art, logic, economic activity or ethical action. He saw individuals as—at best—simple instruments of these processes. History for Croce was about spirit, not about individuals, even though it did deal with individual's 'works'.

Croce worked out the principles of his 'absolute historicism' in the years following the late 1920s—so at a time when modern German historicism had virtually come to an end. The distance between the two types of historicism became clear in the debate which followed publication of Meinecke's *Entstehung des Historismus*. This debate made it clear that Croce and the German historicists thought of historicism in different terms, and also had differing interpretations of the history of historicism. Croce saw his historicism as opposing not only the 'abstract rationalism' of Enlightenment culture, but also the irrationalism which he saw as being present in the historical school, and also in Meinecke. Meinecke saw historicism as originating in the passage from the natural law tradition to Romantic thought, and thought of its greatest triumphs as being the works of Goethe and Ranke. Croce, in contrast, was critical of Ranke's approach to history, which he saw as being unproblematic, and he imagined a direct line of inheritance running from Hegel to his 'absolute historicism' (although the genealogy was complicated by Vico—the national progenitor).

In this same period, historicism (especially German historicism) started to spread to the Latin and the English-speaking countries. This produced some significant accounts and re-statements, e.g., at the hands of José Ortega y Gasset in Spain, or Raymond Aron in France. Croce's historicism (although not fully distinguished from Dilthey's) found an echo in Britain, in Robin G. Collingwood's work. However, by this time, historicism was almost at an end. In 1944 Karl Popper (see *Popper, Karl Raimund* (1902–94)) published a book with the significant title, *The Poverty of Historicism*, which took up the criticisms which neo-positivists (particularly Otto Neurath) had made of the traditional distinction between historical knowledge and natural science. After the war historicism was subjected to head-on attack from various sides (e.g., in the critiques of Leo Strauss and Karl Löwith). In the second half of the twentieth century it has been rare for writers to admit to describe their philosophical positions as 'historicist', even though some have taken up themes and forms of analysis which are similar.

6. *From Historicism to 'Historical' Social Science*

At the same time as its influence in philosophy was declining, historicism became important again for historians, especially in Germany. Central was the reference to Max Weber and his insistence on the need to link historical research and the social sciences, and on the need for history to employ concepts and theories developed by the social sciences. Instead of making the traditional distinction (and contraposition) between historical disciplines and the natural sciences, or that between historical understanding and 'systematic' knowledge of society, Weber favored a strong link between the latter two kinds of knowledge; this approach was taken up by historians such as Hans-Ulrich Wehler, Jürgen Kocka, Wolfgang J. Mommsen and others. In a 1969 essay Wehler proposed a definition of history as 'historical social science' and argued that history ought to be seen as one of the social sciences, yet at the same time have its specificity recognized. He thus emphasized historiography's affinity to economics, sociology or political science, in so far as it necessarily drew on theories, and took most of them from other social sciences.

One of the consequences of this current of thought was that it opened the way for a return to Marxism: not Marxism as a general theory of history, but as a set of theories which could be tested in the interpretation of specific historical processes. In other words, Marxist concepts were used in a 'Weberian' fashion, as ideal types. Theories thus had a purely functional value and could not guarantee the validity of research. This brought out the problem (stressed especially by Kocka) of how to check the findings which theoretically guided research led to, and the related problem of how the validity of theories could be assessed. Kocka's solution was that it is not theories which legitimize findings, but rather findings which demonstrate the heuristic efficaciousness of theories.

These works left aside many of historicism's classic problems, and rejected many of its methodological stances, such as its opposition between explanation and understanding. Their aim was to show that historical writing had its scientificity, albeit scientificity of a special, 'sui generis' kind. They had nothing in common with the revival of Dilthey in hermeneutics championed by Hans-Georg Gadamer. And they were firmly opposed to the notion of history as 'narrative,' proposed by Hayden White and many other American scholars (in work that, not surprisingly, also often embraced a hermeneutic perspective) (see *History: Theories and Methods*; *Bourgeoisie/Middle Classes, History of*). Nonetheless, this did not stop writers like Jörn Rüsen recognizing the narrative character of historical writing, and seeing this as the feature which marks history out from other social sciences. But Rüsen does not deny either the need for theories nor the scientific character of historical research: on the contrary, he aims to show that the narrative structure of historiography is perfectly compatible with the use of general theories, and does not absolve historians from the duty to empirically check the results of their research.

Historicism has thus not so much been extended as 'superseded.' For not only does 'historical social science' reject Ranke's ideal that the historian should be purely 'objective'. It goes so far as to claim that history can and should be critical (or 'emancipating' in Habermas' terms): that is to say, critical vis-à-vis traditional history-writing and its methodological principles, including the separation between history of ideas and social history, but also critical towards traditional historiographical interpretations of Germany's past. This has led to a significant shift in attitudes towards the Enlightenment (see *Enlightenment*). As we have seen, historicism was born partly out of polemic with Enlightenment culture, which it considered anti-historical. In contrast, Horst Walter Blanke has recently tried to link 'historical social science' up to Enlightenment historiography and the Enlightenment conception of history. This marks a sharp break with the traditional conceptions and categories of historicism—especially (but not exclusively) those of Romantic historicism. Thus, historicism did not only decline as a general philosophical approach, but also as a paradigm of historical research and writing.

See also: Hegel, Georg Wilhelm Friedrich (1770–1831); Historiography and Historical Thought: Current Trends; History and the Social Sciences; History: Overview; Simmel, Georg (1858–1918); Weberian Social Thought, History Of

Bibliography

Antoni C 1959 *From History to Sociology. The Transition in German Historical Thought*. Wayne State University Press, Detroit, MI

Aron R 1938 *Essai sur la théorie de l'histoire dans l'Allemagne contemporaine*. J Vrin, Paris (2nd edn., 1950 entitled *La philosophie critique de l'histoire. Essai sur une théorie allemande de l'histoire*)

Blanke H W 1991 *Historiographiegeschichte als Historik*. Fromman-Holzboog, Stuttgart-Bad Cannstaff, Germany

Engel-Janosi F 1944 *The Growth of German Historicism*. Johns Hopkins University Press, Baltimore

Hughes H S 1958 *Consciousness and Society. The Reorientation of European Social Thought, 1890–1930*, 1st edn. Knopf, New York

Iggers G G 1968 *The German Conception of Historical Thought from Herder to the Present. The National Tradition of History*. Wesleyan University Press, Middletown, CT (orig. edn. *Deutsche Geschichtswissenschaft*, DTV, Munich, Germany)

Iggers G G 1984 *New Directions in European Historiography*, rev. edn. Wesleyan University Press, Middletown, CT

Kon I S 1964 *Die Geschichtsphilosophie des 20. Jahrhunderts—Kritischer Abriss*. Akademie Verlag, Berlin

Lee D E, Beck R N 1953–4 The meaning of historicism. *American Historical Review* **59**: 568–77

Rand C G 1964 Two meanings of historicism in the writings of Dilthey, Troeltsch and Meinecke. *Journal of the History of Ideas* **25**: 503–18

Rossi P 1956 *Lo storicismo tedesco contemporaneo*. Einaudi, Turin, Italy (new edn., 1994 *Edizioni di Comunita*, Milan)

Rossi P 1960 *Storia e storicismo nella filosofia contemporanea*. Lerici, Milan (new edn., 1991, Il Saggiatore, Milan)

Rossi P 1989 'Historismus' und 'storicismo': zwei Denktraditionen. In: Esch A, Petersen J (eds.) *Geschichte und Geschichtswissenschaft in der Kultur Italiens und Deutschland*. Max Niemeyer, Tübingen, Germany

Rothaker E 1960 Das Wort Historismus. *Zeitschrift für deutsche Wortforschung* **16**: 3–6

Rüsen J 1993 *Konfigurationen des Historismus*. Suhrkamp, Frankfurt, Germany

Tessitore F 1991 *Introduzione allo storicismo*. Laterza, Rome and Bari, Italy

Wittkau A 1992 *Historismus*. Vandenhoeck & Ruprecht, Göttingen, Germany

P. Rossi

Historiography and Historical Thought: Christian Tradition

Ever since the beginnings of orderly reflections on the topic of history in Christianity and the first literary attempts to write history (by the author of *The Gospel According to St. Luke*, c. 81–96 AD, cf. Luke 1, 1–4 and Eusebius of Caesarea, c. 264–340 AD, *Ecclesiastical History*), four basic problems have been characteristic of all such attempts. Their uninterrupted significance from the ancient world up to the present allows development of a history of problem areas of Christian historical thought. All Christian historical thought and also all Christian historiography can be characterized by basic decisions concerning the following four problem areas: (a) the classification as a specific inside or outside perspective on Christianity (Sect. 1); (b) the respective understanding of the position of one's interpretation on a range between historical and theological standards (Sect. 2); (c) specific mechanisms of selection (Sect. 3), and (d) a small number of models by which the past is reconstructed see (Sect. 4).

1. The Problem of Inside Perspective and Outside Perspective

A Christian understanding of history describes Christianity from an inside perspective; but in many cases such descriptions can, to some extent be regarded as outside perspectives. This is because, together with an outside perspective on Christianity, e.g., by non-Christians, there are also outside perspectives of Christian minorities on Christian majorities, and vice versa. Such perspectives have a special forming influence on the understanding of history: The *Ecclesiastical History* by Theodoret of Kyrrhos (c. 393–466 AD), one of the most definitive works of Christian historians of the second generation after Eusebius, regards historical developments as formed by the debates between orthodox church ('orthodoxy') and apostate heresy. *The Ecclesiastical History* by his contemporary Philostorgius (c. 368–425 AD) has the same perspective, however with the opposite judgment: the theological tradition of the Alexandrine presbyter Arius (c. 260–336 AD), which has influenced theological thought mainly in Great Britain up to the present, is regarded as true orthodoxy by Philostorgius but as apostate heresy by Theodoret. Besides these, there have always existed more or less critical descriptions from an outside perspective by non-Christians (e.g., by the Roman historian Ammianus Marcellinus, c. 330–395 AD, in the heyday of Christian historiography in late antiquity) in which Christian historians were tackled implicitly or explicitly.

It was only in modern Europe that Christian historians have made determined attempts towards a reduction of inside-perspective elements in Christian historiography. This is, of course, mainly with respect to the narrow horizon of thought formed according to the self-awareness of the three Christian denominations that arose in Central Europe after the religious wars of the sixteenth and seventeenth centuries (the Religious Peace of Augsburg 1555 and the Peace of Westphalia 1648). At the beginning of the modern era, reflecting denominational antagonism, the Lutheran, Reformed, and Roman Catholic interpretations of

history entered into competition. The protagonist of the overcoming of this denominational approach to the history of Christianity was Gottfried Arnold (1666–1707; Professor of History in Giessen 1697–1698). He tried to write a '*unpartheiische*,' which is a non-denominational church history. In the twentieth century, following the institutionalization of the ecumenical movement, which has tried to overcome the denominational splittings of the sixteenth century, the call for a general 'ecumenical church history' (Ernst Benz) arose. In a widespread late twentieth-century attempt to conceptualize the alternatives, 'church history' as opposed to 'history of Christianity' has been used. 'Church history' describes a perspective orientated towards the Christian denominations, and 'history of Christianity' describes a trans-denominational perspective.

2. *Christian Interpretation of History—Historical and Theological Standards*

A Christian interpretation of history experiences a tension between historiography (or history) on the one hand and Christian theology on the other. This tension has existed from the very beginning in so far as theology (which in this context means the theory of the Christian faith) naturally favors certain theoretical options when reconstructing a history, for example, its assumption of a god acting in history, which other interpretations of history may hold but are by no means bound to hold. Such specific theoretical options always have methodical consequences; this is what the religious philosopher Ernst Troeltsch (1865–1923) showed in his clear analysis and what has been called 'historical or dogmatic method' (Troeltsch 1922, pp. 729–753). But one must not overestimate the 'theological' elements of Christian interpretations of history before the modern era: from the fourth century they had as a matter of course been orientated towards the literal and methodical standards of the interpretation of history of their respective times. This theoretical background was broken only very partially by specific religious or theological concepts. The first author of an ecclesiastical history, Eusebius of Caesarea, already followed the forms and standards of the historiography of his time in his historical writing, as can be seen in the prologue (Eusebius, historia ecclesiastica I 1,3f). Consequently, his account leaves out central questions of a theological interpretation of history. Because pagan ancient historiography very rarely explicitly treated the difficult relation between divine acting in history through fate (*θεοτειχη/theoteiche* or *fatum*) and human freedom, Eusebius, too, never picked out the exact relation between man's freedom and God's acting in history as a central theme.

The fundamental problem of the relation between an interpretation of history according to modern standards on the one hand and specific theological premises on the other hand intensified in modern Europe. In the course of the modern secularization of historical thought, under the premise of *etsi Deus non daretur* (God is not necessary), new determinants of history (such as fate, chance, reasons of state, and ethics) were established in place of the Christian approach to a history regarded as being under the law of God. In consequence, the problem of how—under these conditions—a theological view of history was possible, increased. Most Christian interpretations of history after the Enlightenment have, however, tried to take up the problem, often through a religious or theological interpretation of the philosophical framework of the historiographic theory. The Berlin theologian Friedrich Schleiermacher (1768–1834) for example made the following point in his notes of the years 1796–1799: '*Die Historie ist immer religiös und die Religion muß ihrer Natur nach historisch sein*' ('history is always religious and religion must be historical by nature,' Schleiermacher 1984, p. 25). At the same time the possibility of a Christian interpretation of the world, which was lost during the period of Enlightenment, was meant to be restored by means of such a religious charging up of history.

3. *The Mechanisms of Selection of Christian Interpretation of History*

All interpretation of history and, furthermore, all historical writing, is characterized by acts of selection from a huge amount of material. Within the framework of a Christian interpretation of history, first the basic choice of subject must be made: to date, it has been heavily disputed whether a Christian interpretation of history merely ought to deal with certain fields of general history such as Christianity, the history of certain denominations or ecclesiastical groups, or a particular history formed in a particular way by a transcendental power ('Heilsgeschichte', 'salvation history'), or whether a Christian interpretation of history ought rather to come up with a distinctive Christian perspective on history as a whole called 'world history' since the end of the eighteenth century. The controversies over whether the chosen ideological perspective necessarily leads to a concentration on an ecclesiastical history, or on a history of Christianity, or alternatively on the wider framework of a universal history, has always been an issue in Christian historical writing.

The first Christian author of a historical monograph, Eusebius, chose to concentrate on an ecclesiastical history without any detailed consideration of the political history of his time. A universal history, now lost, was presented by the Greek presbyter Philippus of Side less than a hundred years later, before AD 439. The Christian interpretation of history

that has had the strongest influence up to the modern era, and which was fundamental in the middle ages, was written by the North-African theologian Augustine (AD 354–430). It takes the perspective of a universal history: after the city of Rome had been captured by the Goths in AD 410, both pagan and Christian interpretations of history, which until then had their focus in the Empire, broke down because of the obvious collapse of the Roman Empire. As a consequence, the bishop was bound to develop his own distinctive concept of a universal history which would be able to convey a new sense of history. Augustine presented it with his work *De Civitate Dei*, '*The City of God*' (AD 413–426; for details see Sect. 4).

While the Middle Ages mainly expressed concepts of universal history based on Augustine's, e.g., by Otto of Freising (c. 1112–1158), the traditional Christian universal—historical interpretation broke down during the Enlightenment. Theological categories, such as 'God's revelation' or 'God's acting', which had been systematic premises of the Christian interpretation of a universal history, and had consequently formed a unity of transcendental and earthly reality, became problematic. After the attempts of the nineteenth century to recreate such a concept of unity on a new basis, and thus to once more present a Christian interpretation of a universal history (in very different ways by Schleiermacher, Hegel, and Ranke), the twentieth century progressively withdrew from such theological interpretations of history. Christian interpretation and presentation of history in the twentieth century were more or less confined to the specific context of Christianity and church, even if the theoretical basis was sometimes broader. That was the systematic reason for the modern criticism of concepts which attempted to create a close correlation between the history of human development and God's acts of salvation, by way of 'salvation history', the theological interpretation of certain incidents.

Along with the fundamental question of whether a Christian interpretation of history must be understood universally, or be restricted to a history of Christian institutions, there are other mechanisms of selection taken from general historiography which are of importance to a Christian interpretation of history. Historians of the ancient world and the middle ages used, as important mechanisms of selection, the questions of the meaning of an incident, of its benefit for future generations, or of its documentation in the respective privileged sources, and so forth. The bias thus created is obvious. However, in general, the critical sense with which historians before the modern era treated their own modes of selection is underestimated. Such reflections were usually made in the prologue of historiographical works.

The vehement calling into question of traditional mechanisms of selection in the general history of the twentieth century (see e.g., the criticism of any historical interest with a biased concept of histories of ideas or politics, the criticism of the concentration on Europe, and of the neglect of feminist research) reached the historiography of Christian denominations with a certain delay; its consequences on a general Christian historical awareness only gradually become apparent.

4. *The Basic Models of a Christian Interpretation of History*

Christian conceptions of history up to the modern era usually have reconstructed it on the basis of a small number of comparatively simple models. Eusebius describes the history of the church according to a model of a process of constant 'advances,' something which is in strict contrast to the historiography of his times with its stress on cyclic processes and the decadence of traditional Roman values. The history of his own religion teleologically leads to the Christian empire of Constantine. Augustine breaks with Eusebius' model of a progress with his concept of two cities (*civitates*) wandering through the ages. He differentiates between the *civitas Dei*, the 'citizenry of God' as an earthly community of the heavenly citizens, and the *civitas diaboli*, 'the citizenry of the devil.' Of course, the Roman state and the community of those who do not live under God and towards him are not identical in the sense of the state as such being the *civitas diaboli*. But Augustine takes up the traditions of radical Christian criticism of the state, and with his anthropologically based historiographic model gives a clear analysis of the social deficiencies of his time. The European Reformation breaks with this conception. It mainly follows the models of decadence developed by the medieval church reformers, which can also be found with humanistic authors: the promising beginnings of the first Christians had decayed in a middle period which are only now brought to light again. This conception was monumental, and had a strong anti-Catholic impetus in the so-called 'Magdeburger Zenturien'—a team effort by protestant theologians (Basel 1559–1574) led by the strict Lutheran Matthias Flacius Illyricus (1520–1575), which has been known under that name since the eighteenth century, and which has remained a fragment. Along with a comparatively simple modeling, the work is characterized by the use of an enormous number of sources. Cardinal Caesar Baronius (1538–1607) reacted to this challenge with his *Annales Ecclesiastici*. But this catholic reply, in terms of method only, mirrored the protestant work, and it tried to answer the reproach of having broken with the truth with a proof of uninterrupted continuity from the age of the early Christian church.

The model of decadence gained special influence (e.g., in Goethe's work); Arnold applied as a basis—despite his programmatic orientation towards an

impartial standpoint. Arnold explained the history of Christianity according to a model of progressive decadence, beginning after the pure earliest church. Despite his persistent interest in counterevidence to official protestant ecclesiastical historiography, he held to his conception of decadence, and to a structural principle of the denominational history of Christianity. Such simple basic models for the reconstruction of the history of Christianity were by no means, or first, only questioned by the European Enlightenment. Augustine's criticism of Eusebius' conception of progress in *De Civitate Dei* is an example, because it was motivated by an empirical diagnosis of society and by critical work on biblical texts. Humanist historians of the Reformation must also be mentioned, particularly Philipp Melanchthor (1497–1560) and Johannes Sleidanus (1505–1556). In his short world history, and in a work on contemporary history entitled *Denkwürdigkeiten*, *Betreffend Staat und Religion Unter Kaiser Karl V*, strict models of decadence are missing. Sleidanus always endeavored to describe religious changes with regard to imperial power and state power; he can be regarded as one of the progenitors of a strong orientation towards political history.

Along with the orientation towards simple conceptions of progress or decadence, the history of Christian historiography, and the Christian awareness of history, has called into question over-simplified models up to the present. In Ranke's formula, for example, each epoch is immediate to God; divisions into periods are good indicators for the respective underlying models. In this respect, denominational Christian interpretation of history tends to differ from the standards of general history; protestant conceptions for example speak of a 'period of Reformation' as a threshold between the middle ages and the modern era.

The four basic problems outlined above have characterized Christian concepts of history from its beginnings, and they show that the crucial changes in Christian interpretation of history, caused by the increasing influence of the methodical standards of modern historiography and finally by the victory of historicism, must not be overlooked in theology, either. The victory of historicism has had a significant consequence on academic theology as a whole, because the historization of all theological disciplines was demanded (among others by Troeltsch), and was partially carried out. It is equally clear and almost undisputed that the general 'crisis of historicism' also had an impact on theological form of Christian interpretation of history, something which had already been diagnosed by contemporary church historians like Karl Heussi (1877–1961). Mainly in continental Protestant theology, this general crisis was accompanied by a theological criticism of general historicism. This development is now called an 'anti-historical revolution' (Graf 1988, pp. 377–405), a term which points to its consequences: the removal of church history as a leading subject of academic theology.

In the twentieth century there was an interesting revitalization of apocalyptic thought on a large scale, the relation of which to the 'crisis of historicism' is not immediately clear. This form of Christian theology, which goes back to a movement in Hellenistic Judaism, and which had already left deep marks on certain biblical texts, is characterized by a distinctive awareness of history. The present time is interpreted as the 'last age,' and forms part of a certain fixed scheme of the course of history, whilst there is no historical interest in the identification of certain 'signs of the times.' In the ancient world, the formation of a Christian awareness of history was accompanied by a drastic loss of meaning of apocalyptic elements for Christianity. The eruptive revival of apocalyptic trends in the middle ages, the early modern era, and the present has in most cases been accompanied by a protest against Christian interpretations of history orientated towards the standards of the respective time.

5. *Current Problems and Questions of Christian Interpretation of History*

The systematization of current schools of Christian interpretation of history meets with the difficulty of the actual presence of pluralism (Nowak 1997, pp. 3–12), perceived a 'new confusion.' Such concepts are present in the institutional fields of denominational church history as well as independent of these. The Christian inside perspective, considered to embody prejudice in some accounts, presents a challenge to other approaches to the problems of Christian historicism; they have not, however, discarded it. This shows the limited significance of the presently popular alternatives of a 'church history' and a 'history of Christianity.' In both cases, attempts to take an outside perspective are, more or less admittedly, wrapped in an inside perspective.

Within the horizon of the standards of history at the beginning of the third millennium on the one hand, and of Christian theology on the other, almost all schools have tried to achieve careful de-theologization. This, however, is rarely picked out as a central theme. It is only very occasionally that a decidedly theological contribution to the general discussion on history is demanded (Herms 1997, pp. 305–30); a 'resort to pragmatism' can be noted (Nowak 1997, p. 12). Even if some historians still demand a concept of universal history in Christian terms as regards the selection of subject matter, most publications actually present contributions to a discipline of general history. That such presentations are an most cases characterized by the leading assumptions of universal history (and theology) is obvious to any sensitive examination; but

it is only rarely admitted or made explicit. The presently vigorous call for an approach to church history, or history of Christianity, as historical–cultural studies also implies certain universal–historical structures in the shape of an '*histoire totale*'.

The critical calling into question of historiographical mechanisms of selection and of modeling, which mainly arose in the twentieth century, has increasingly been accepted by the institutionalized Christian interpretation of history. But it is very rare that it is understood as one of the consequences of fundamental decisions made in the course of the European Reformation movement which have created a high level of ideology-critical potential.

These four basic problems of Christian interpretations of history cannot be solved by means of isolated special theological hermeneutics. They call for an original contribution to the general discussion of history and of the interpretation of history. This is especially true in view of the latest developments in 'writing history in the postmodern era': can the concept of a literal sense as the regulatory concept of the process of interpretation really be given up in the course of the deconstruction of its own narrative? does not a Christian interpretation of history which does away with the 'grand narratives', and the 'master processes' dissolve itself? is the notion of 'God's acting' dispensable in a Christian interpretation of history? Here, a new and interesting challenge is put to the long-neglected field of a theory of Christian interpretation of history. It is possible that a theoretically renewed, clear-cut, Christian interpretation of history helps to preserve in a positive sense two functions of the Christian interpretation of history under the conditions of the postmodern era: the function of forming society, and the function of critically analyzing society. But with all this, the traditional objections to a simplistically designed history of salvation remain important.

See also: Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: Modern History (Since the Eighteenth Century); History: Theories and Methods

Bibliography

Graf F W 1988 Die "antihistorische Revolution" in der protestantischen Theologie der zwanziger Jahre. In: Rohls J, Wenz G (eds.) *Vernunft des Glaubens. Festschrift W. Pannenberg*. Vandenhoeck & Ruprecht, Göttingen, Germany, pp. 377–405

Herms E 1997 Theologische Geschichtsschreibung. *Kirchliche Zeitgeschichte* **10**: 305–30

Markschies C 1995 *Arbeitsbuch Kirchengeschichte*. JCB Mohr, Tübingen

Markschies C 1998 Kirchengeschichte. In: Goertz H-J (ed.) *Geschichte. Ein Grundkurs*. Rowohlt, Reinbek bei Hamburg, Germany, pp. 408–22

Meinhold P 1967 *Geschichte der kirchlichen Historiographie*. Karl Alber, Freiburg, Germany

Nowak K 1997 Wie theologisch ist die Kirchengeschichte? *Theologische Literaturzeitung* **122**: 3–12

Schleiermacher F 1984 *Schriften aus der Berliner Zeit*. Meckenstock G (ed). W. de Gruyter, Berlin

Troeltsch E 1922 Ueber historische und dogmatische Methode in der Theologie. In: *Zur religiösen Lage, Religionsphilosophie und Ethik*. JCB Mohr, Tübingen, pp. 729–53

C. Markschies

Historiography and Historical Thought: Classical Period (especially Greece and Rome)

Historiography, the rational reconstruction of the past in the medium of a prose that claims to literary quality, was developed in Greece during the fifth century BC. Its emergence reflects the peculiar intellectual and political developments of this age. The high cultures of Egypt and Mesopotamia had known lists of rulers, their foundations, festivals, and wars that represented the official version according to which the monarch inevitably would crush enemies. There was no need and no chance to develop a conception of the past in which the course of events would be ascribed to the contending efforts of multiple actors, the result of which was unpredictable and thus needed to be reconstructed and narrated in each case. The historical books of the Old Testament, in a certain sense, took account of the contingency of human affairs but at the same time attributed the outcome of events to God's intervention.

1. The Origins of Greek Historiography

After the development of literacy during the eighth century BC, Greek commemoration of a distant past had materialized in the epic tradition (Homer, Hesiod). Its combination of 'historical' and 'mythical' elements became subject to criticism in an age when Ionian sixth century philosophy (as represented by the Milesians Thales, and Anaximander) had attempted at a 'scientific' explanation of the physical word and of human society. About 500 BC Hecataeus of Miletus tried to reconstruct the genealogical facts out of the mythological traditions. He wrote a pioneering work of geography which reflected the expansion of the geographical and historical horizon of the Greeks as a result of their colonization (since c. 750 BC). This

included their expansion on the shores of the Mediterranean and Black Seas, their voyages of discoveries, and the information that the Eastern Greeks (as subjects of the vast Persian Empire) had acquired about parts of Asia. Historiography, in a narrower sense, developed out of the experience of the Persian Wars. However, the realization that the great victories of the Greeks in 490 and 480–479 BC had been achieved, so to speak, against all probability apparently developed only with a certain distance to the events themselves.

Herodotus (c. 485–after 425 BC; a widely-traveled native of Halicarnassus in Asia Minor) opened his account of the Persian Wars with the statement that he wanted to present the results of his research (*historie*) in order to commemorate the deeds of both Greek and Barbarians and to establish the cause of their conflict. *Historie* is a general term for a rational inquiry. It was only later (especially with Aristotle's *Poetics*) that it became the label for the literary genre which Herodotus had created. Herodotus pointed out that his study was the work of an individual and that he felt obliged to an impartial approach. Greek historiography was developed by free intellectuals and did not start as officially committed accounts. Apparently, especially exiles like Herodotus—and later *inter alia* Thucydides, Xenophon or Polybius—were able to take on the position of an (at least, subjective) neutral observer. Herodotus did not only provide the first work of an *histoire événementielle* based on stories which might have been generated two generations before his own time; he also continued the tradition of geographical and ethnographical accounts by incorporating digressions on the various peoples (from the Egyptians to the Scythians) under Persian control. Here he relied on local traditions and autopsy and sometimes reported stories, the authenticity of which he himself doubted.

Herodotus' younger contemporary Thucydides of Athens (c. 460—c. 400 BC) created a quite different model of historiography. He deliberately distanced himself from Herodotus (though without naming him) since he thought Herodotus to have shown too much confidence in his sources. Thucydides' statements on his methodological principles are (with the notable exception of Polybius) almost without parallel in ancient historiography. He chose as his subject the Peloponnesian War between Athens and Sparta and their respective allies (431–404 BC; Thucydides' account, however, is incomplete and breaks off at 411 BC). According to his own words he had started to take notices already with the outbreak of the war since he had anticipated the unprecedented dimensions of this war which affected the whole Greek world. Later, he used his time in exile (the consequence of his defeat as an Athenian general in 424) to collect all available facts by interrogating protagonists and eyewitnesses of both sides. Being confident that he was able to reconstruct the objective course of events he did not name his sources or quote varying, or even conflicting, versions by different witnesses but digested all information into an authoritative account of 'how it had really happened' (an aside of his which was later to be taken up by Ranke). He wanted to differentiate the long-term causes, which led to a climate of mutual distrust between the two powers, from the immediate occasions, which made them take military action. War and civil war were seen by Thucydides as revealing the *condition humaine*, love of power, ambition, greed, and fear as constituents of people's invariable nature; thus the thoroughgoing analysis of human action should provide insights of everlasting value.

Already Herodotus had incorporated some speeches into his narrative in order to explain the motives behind the decisions of protagonists but Thucydides went much further by systematically using this stylistic device based on the techniques of rhetoric that were taught by contemporary Sophists. As Thucydides explained, his methodical principle was 'to make the speakers say, what, in my opinion, was called for by each situation while keeping as closely as possible to the general sense of the words that were actually used.' Those speeches served to present opposing arguments (in pairs of speeches ascribed to antagonists) and to discuss the political, moral, and juridical implications of certain decisions or to express the self-image of a society (especially in the 'Funeral Speech' attributed to Pericles). The unity of style, the high level of abstraction, the implicit cross-references, and the shortness of the speeches point out that they were primarily a medium to carry the historian's interpretation. Since Thucydides, the inclusion of more or less fictitious speeches was a standard feature of ancient historiography.Thucydides made two exceptions to his principle of writing a strictly contemporary history. He gave an account of the development of early Greek history in order to prove that due to the level of its material culture, a war of such dimensions like the Peloponnesian War would have been impossible. That was a sort of conjectural history based on theories of societal development and on the interpretation of cultural survivals. He also gave a selective description of the 50-year period between the end of the Persian Wars and the outbreak of the Peloponnesian War in which he stressed the rise of Athenian power and the distrust it fostered with the Spartans. This account also served to bridge the gap between Herodotus' narrative and his own.

Thucydides' model of historiography differed from the Herodotean one by eliminating the features of cultural history and ethnographical discourse. But both corresponded with respect to concentrating on large-scale wars, which involved great parts of the Greek world, and with regard to treating matters of constitutional history and internal politics of the states concerned only insofar as they were relevant for the course of events. Their different attitudes to the period

of time that could be reconstructed by drawing on oral traditions notwithstanding, they both primarily relied on oral sources. As a rule, down to the Christian historians of Late Antiquity ancient historiographers did not consult documentary evidence systematically like inscriptions or state papers in archives and did not engage seriously in 'antiquarian' research on constitutional, legal, and religious institutions. That is why they could not think of replacing the narratives of former historians by a new account based on other better sources. Thus Thucydides was a continuator of Herodotus, Xenophon, and other fourth century historians who started their Greek histories at the point where Thucydides' narrative had broken off. Historians who wanted to treat former times again, cannibalized (often in combination with harsh criticism) the accounts of their predecessors.

2. *The Variety of Greek Historiographical Production*

Thucydides had shaped the trend-setting model of military and political history that was followed by a great number of Greek writers including the historians of Sicily. However, his successors, embellished the rhetorical elements or amplified scenes with emotive appeals to readers. Throughout antiquity, Herodotus enjoyed a dubious reputation as being at the same time the father of history but also a narrator of fabulous stories (in Cicero's famous phrase). The broader Herodotean approach, however, was not totally outdone; it was taken up again by the historians of Alexander the Great and survived in later works with strong geographical and ethnographical interests like that of Posidonius (c. 135–c. 50 BC) and in the specialized geographical works from Eratosthenes (third century BC) to Strabo (Augustan age).

Besides the mainstream of military and political history, a number of specialized historical genres developed since the late fifth and during the fourth centuries. These included: universal history (Ephoros of Cyme, fourth century BC, whose—later lost—work was used by Diodorus Siculus, first century BC); foundation histories, local chronicles, and constitutional histories (e.g., on Athens); cultural histories (Dicaearchus, pupil of Aristotle, late fourth century BC); biographies (late fifth century BC sketches later to be used by Plutarch, c. 45–c. 120 AD; and Xenophon's life of the Spartan king Agesilaus II, fourth century BC).

During the Hellenistic Age, an enormous number of historical works of various kinds have been produced by nearly a thousand historians known to us by name. This tremendous increase of historical production reflected the political reorganization of the Mediterranean world by the establishment of the vast Hellenistic empires (and later the rise of Rome to a world power). It was fostered by the patronage of Hellenistic monarchs and the establishment of libraries and scholarly institutions, especially in Alexandria and Pergamum. Almost all of this extremely rich production has, however, got lost and is only known from quotations and references in later works. For example, the various reports on Alexander's campaigns by court historiographers, officials, and generals only emerge with the compilation of Diodorus; comprehensive accounts of Alexander's history based on the works of his companions date from the second and third centuries AD (Arrian and Curtius Rufus).

Polybius (c. 200–c. 120 BC) is the only Hellenistic historian of whose historical work a substantial amount has survived. Polybius, a military of the Achaean Confederacy (on the Peloponnese) was one of 1,000 leading Achaeans who were deported to Italy after 167 BC. He came into close contact with prominent figures of the Roman nobility, especially Scipio Aemilianus (the destroyer of Carthage). The subject of his historical work was Rome's rise to world dominion in the time from the Punic Wars to the destruction of the Macedonian state (167 BC). Polybius calls it a 'pragmatic history'; it focuses on political and military actions and is especially based on eyewitness reports and first-hand knowledge of the theaters of war. His intentions and explicit methodological reflections (combined with polemic against several former historians) imply a conscious return to the Thucydidean model of historiography. But Polybius understood Rome's military success as having depended on its uniquely stable, and at the same time flexible, political system. In order to explain this point to his Greek readers he incorporated a description of the Roman constitution and military system and thus gave a systematical treatment of institutions otherwise not to be found in ancient historiography. (His analysis of Rome's 'mixed constitution' should become one of the most important subjects in the reception of ancient political theory since the Renaissance.)

3. *Roman Republican Historiography*

Roman history was first made the subject of historiographical works by Greek authors like the Sicilian Timaeus of Tauromenium (c. 350–c. 260 BC) and other Greek writers who were especially interested in the stories about the foundation of Rome. Latin literature only developed since the mid-third century BC as a product of the reception of Greek culture. Certain records of the past, however, had been preserved since early times: public documents like laws, decrees of the senate and international treatises, and family records of the aristocracy. And there were the *annales* kept by the pontiffs in which the names of consuls (the eponymous magistrates) were listed, triumphs and religious prodigies were noted, and some

miscellaneous information was collected. About 200 BC, the senator Fabius Pictor provided the first history of Rome written by a Roman yet in Greek. Drawing on the models of Hellenistic historiography Fabius Pictor covered the time from the founding of the city until 200 BC. He probably gave elaborate accounts of the early, and of the contemporary periods and only a sketchy one of the time in between, a proportion that should become typical of many later writers. He was followed by other historiographers writing in Greek until the mid-second century. The older Cato (234–149 BC) was presumably the first to write a historical work in Latin. His *Origines* applied the interest in foundation history not only to the earliest history of Rome but also to the emergence of the cities in Italy and paid particular attention to the unification of Italy. His treatment of political history focused on the justification of his own policy and thus provided the model for the *ex parte* tendency of Roman contemporary historiography. Roman historiography was ethnocentric and it was the product of the leisure-time activity of senators. Their intentions were to demonstrate Rome's cultural equality with the Greeks, to justify her domination over the Mediterranean world, to praise the wisdom and success of the ruling aristocracy, and to take sides with respect to contemporary conflicts. From the late second century onwards, a number of Latin historiographical works were composed which followed in their structure the pontifical *annales* (which were published in a full version in 120 BC) and gave a year-by-year account from the origins of the city. They apparently embellished their descriptions of the early republic and especially the 'Struggle of the Orders' between patricians and plebeians by projecting back features of the internal conflicts that had become almost endemic since the tribunate of Tiberius Gracchus (133 BC). Some historians seem to have taken into account certain results of serious antiquarian research, others had no scruples and they invented 'documents.' (It was the Greek author Dionysius of Halicarnussus who, living in Augustan Rome, showed the greatest interest in evidence of antiquarian nature.) All these Latin 'annalistic' works (as well as the earlier historiographical attempts previously mentioned) are only indirectly known. They were used by Livy (c. 59 BC–AD 17) who was the first Roman historian without a record of political and military experience of his own. His history covered the time from the foundation of the city to 9 BC. Out of his 142 books, only 35, which reach down to 167 BC (with *lacunae*) have survived. Livy often struggled with the contradictory accounts of his predecessors and the quality of his own descriptions obviously depends on the quality of the sources available to him. He did not invent his own versions of the course of events but his work shows the problematic nature of a history totally made on the basis of other, often overtly partisan, historical works that were mostly accepted at face value.

The Thucydidean model of contemporary history was imitated by Sallust (86–35 BC) who after the failure of his political career wrote monographs on the conspiracy of Catiline (63 BC) and on the Roman Wars against the Numidian king Jugurtha (111–106 BC). He interpreted the political crisis of the Late Republic as the consequence of a moral decline following from the abandonment of traditional values. (No other contemporary work on the Late Republic has survived; a comprehensive account of this period is only available in works by later authors such as Appian, second century, and Cassius Dio, third century, who both wrote in Greek.)

4. *Roman Imperial Historiography*

Writing contemporary history might become a dangerous job under the Emperors as the fate of the historian Cremutius Cordus demonstrates. Under the reign of Tiberius (in AD 25) he was indicted for *lese majesty* because of the republican sympathies he had shown in his work on the Civil War period (he committed suicide; his work was burnt). The short sketch by his contemporary Velleius Paterculus, a former military, covered Roman history from the very beginnings to his own days in only two books and culminated in the panegyric of the emperor Tiberius.

The historiography of Tacitus (AD c. 55–c. 120) reflects a new climate of intellectual freedom during the reigns of Nerva and Trajan in comparison to the experience with the tyrannical régime of Domitian (under whom Tacitus had started his public career). His (only partly preserved) works, the *Histories* and the *Annals*, cover the time from AD 69–96 and AD 14–68, respectively. Tacitus kept the annalistic form and he paid special attention to the proceedings of the senate. But the form was often at odds with the contents of the works since the policies of the rulers could not easily be fitted to a chronological scheme that had been adequate for a senatorial régime with yearly changing consuls at the head. Tacitus showed certain nostalgia for republican liberty without indulging into illusions about a return to a republican constitution. Though claiming to write *sine ira et studio* his personal judgments on rulers (especially the gloomy picture of Tiberius) should serve to uphold the mirror of an honest ruler and his comments on senatorial proceedings reflected his standards for dignified behavior under a monarchy. Tacitus' work with its stress on the psychology of emperors and senators marks the final point of Latin senatorial historiography. (However, later there was an annalistic history of Rome from the beginnings to AD 229 written in Greek by the senator Cassius Dio.)

Apart from short summaries of Roman history, during the second century serious historiography gave way to imperial biographies. Suetonius (AD c. 70–c. 140, secretary for the library and subsequently for the imperial correspondence under Hadrian) gave the life

of 12 rulers from Caesar to Domitian. They were arranged in a more topical than strictly chronological manner and drew on archival materials as well as traditions stemming from court gossip. Modeled on Suetonius but of a highly fictitious character (with numerous references to imaginary sources) was the *Historia Augusta*, a series of biographies of emperors for the time from Hadrian (AD 117–138) to the accession of Diocletian (AD 285) which was probably produced in the late fourth century by one or more anonymous writers who took on the names of bogus authors.

5. *The Impact of Christianity*

The religious tensions of Late Antiquity due to the rise of Christianity did only partly affect pagan historiography which continued to concentrate on short summaries of Roman history. The general restraint from religious polemic holds also true for the work of Ammianus Marcellinus (c. 330–395), a Greek from Syrian Antioch and former military. His history, written in Latin, is otherwise a great exception since he embarked on a continuation of Tacitus that treated the time from Nerva's accession to the throne (96) to the emperor Valens' disastrous battle of Adrianople against the Goths in 378 (only the exhaustive parts on the years from 353 onwards have survived).

Christian authors did not enter into a competition with pagan ones on the traditional fields of political and military historiography but they developed models of historical writing that should become the most important innovations until Renaissance times. Since the third century they compiled chronicles of universal history which synchronized the data of biblical, Greek, and Roman history. The chronicle—in Greek—by Eusebius (c. 260–339, bishop of Caesarea, adviser, and later biographer of Constantine the Great), its Latin translation and continuation (until 378) by Hieronymus and several successor works provided the model for a view of world history since the creation. Eusebius, again, created with his *Church History* a new historiographical genre not only with concern to the subject, the history of the Christian 'nation,' but in formal respects as well by his dispensing with speeches, his giving bibliographical surveys, and his verbatim quoting from letters, official documents, and other sources. The awareness that their theological approach might make their accounts challengeable led Eusebius and later *inter alia* Socrates and Sozomenus (both early fifth century) to prove thus the accuracy of their factual reports.

6. *Concluding Remarks*

Given the complexity and variety of ancient historiography it is almost impossible to summarize its impact on the later Western tradition, especially since the Renaissance, in a few sentences. The most important legacy of Greek historians was the development of a critical attitude to the past, the invention of methods that made it possible to discern between facts and fiction. The shortcomings of their attempts in view of modern eyes should not obscure this unique contribution to the development of the Western mind. The adoption of historiography by the Romans provided the model for constructing the past in the shape of national history. The Christian writers invented ecclesiastical historiography as a subject of its own and a model for universal chronicles and histories starting from the creation. Since the Renaissance, Thucydides and Tacitus were taken as the masters of political history, Polybius as the model military historian. With the growing interest in Oriental cultures Herodotus gained a new reputation as a cultural historian. Ancient historiography, which had covered republican city-states as well as world monarchies, was as a whole considered as teaching political and moral lessons of everlasting value (and thus was also of great importance for political theory). The tradition that historiography and antiquarian studies were two quite different subjects survived until the late eighteenth century. There were a vast number of commentaries on ancient texts, collections of inscriptions and coins, learned dissertations on the constitutional, legal, military, and religious institutions of the ancient world but nobody thought of replacing the ancient historiographical works by writing anew narrative histories of antiquity. Gibbon was one of the first to do this and to integrate the results of antiquarian research but he still avoided the period covered comprehensively by ancient historiography and thus started his own narrative only with the second century AD. The doubts on the reliability of ancient historiography, especially of the Roman annalistic tradition, that had been developed in learned dissertations since the seventeenth century led only in the early nineteenth century, starting with Niebuhr, to compose a modern historiography of antiquity based on source criticism and systematic researches on the monumental remains of the ancient world.

See also: Antiquity, History of; Aristotle (384–322 BC); Christianity Origins: Primitive and 'Western' History; Classical Archaeology; Historiography and Historical Thought: Christian Tradition; History: Forms of Presentation, Discourses, and Functions; History: Overview; History: Theories and Methods

Bibliography

Dorey T A (ed.) 1966 *Latin Historians*. Routlegde and Kegan Paul, London

Fornara C W 1983 *The Nature of History in Greece and Rome*. University of California Press, Berkeley, CA

Hornblower S (ed.) 1994 *Greek Historiography*. Clarendon Press, Oxford, UK

Meier C 1975 Geschichte (II. Antike). *Geschichtliche Grundbegriffe* **2**: 595–610

Meister K 1990 *Die griechische Geschichtsschreibung*. Kohlhammer, Stuttgart, Germany
Mellor R 1988 Roman historiography and biography. In: Grant M, Kitzinger R (eds.) *Civilization of the Ancient Mediterranean*. Charles Scribner's, New York, Vol. III, pp. 1541–62
Momigliano A D 1955–1992 *Contributi alla storia degli studi classici e del mondo antico* (9 Vols.). Edizioni di storia e letteratura, Rome
Momigliano A D 1978 Greek historiography. *History & Theory* **27**: 1–28
Momigliano A D 1990 *The Classical Foundations of Modern Historiography*. University of California Press, Berkeley, CA
Strasburger H 1977 Umblick im Trümmerfeld der griechischen Geschichtsschreibung. In: *Historiographia Antiqua. Commentationes Lovanienses in honorem W. Peremans septuagenarii editae*. Leuven University Press, Leuven, Belgium, pp. 3–52
Usher S 1988 Greek historiography and biography In: Grant M, Kitzinger R (eds.) *Civilization of the Ancient Mediterranean*. Charles Scribner's, New York, Vol. III, pp. 1525–40
Winkelmann F 1991 Historiographie. *Reallexikon für Antike und Christentum* **15**: 724–66
Woodman A J 1988 *Rhetoric in Classical Historiography*. Croom Helm, London

W. Nippel

Historiography and Historical Thought: Current Trends

1. Introduction: Dominance of Social Science Models of History around 1970

The years 1970–2000 have seen important reorientations in historical thought and historical writing. These changes are closely related to fundamental transformations in the conditions of modern life.

Briefly comparing the intellectual and historiographical situation in the year 1970 with that of 2000, we can distinguish certain dominant trends then and now. In neither case, however, is it possible to reduce historical thought to simple common denominators. There are no paradigms in the Kuhnian sense that determine historiographical practice but a great plurality of outlooks that are perhaps even more numerous now than they were then. Nevertheless, in 1970 most historians were still convinced that history was a science, albeit a social science, engaged in recreating a real past as it had actually occurred. Today this confidence in the objectivity of science, including historical inquiry, has come under severe attack, as have the notions still dominant in the early 1970s of a coherent process of history. To be certain, faith in objective truth, in historical inquiry, and confidence in history as the ongoing development of civilization had been undermined as early as the second half of the nineteenth century, e.g., by Jacob Burckhardt and Friedrich Nietzsche, and even in the United States, which remained a citadel of belief in science and progress, by leading historians such as Carl Becker and Charles Beard in the 1930s.

Historical science, as it was generally understood in the Western world around 1970, viewed history as a social science. It replaced an earlier orientation sometimes labeled 'historism' (which had been dominant in much of the nineteenth and well into the twentieth century), that had argued that history as a *Geistes-* or *Kulturwissenschaft* required a logic of inquiry different from that of the hard sciences, namely one which takes into account the elements of meaning and intention in human life which cannot be reduced to abstract generalizations. These conceptions of historical study that viewed history as a special sort of *Geisteswissenschaft* were increasingly replaced in the twentieth century by conceptions of history as a 'social science.' The new social science orientations went in several directions discussed below, but they shared certain assumptions. They criticized the older historiography for narrating a history of events, primarily political, focusing on leading personalities in positions of power, and called for a history that proceeded analytically, seeking to examine the social context in which events occurred. While the older historiography relied heavily on philology as a tool for the examination and understanding of texts to fathom human intentions, the new history turned to the neighboring social sciences, sociology, economics, and psychology, and in France also geography, which analyzed social structures and processes and stressed the interdisciplinary nature of historical studies. They strove for a history that was more comprehensive than that of the conventional narrative political historiography. They also wanted to include broader segments of the population into the historical accounts.

But it was only after 1945 that forms of social science history emerged as a dominant model for historical studies. Again it is difficult to reduce social science-oriented history to a common denominator; it ranged from the studies of the material and biological bases of historical life in the French *Annales* in the 1960s, to the new historical demography in France and Great Britain, to quantitative urban history and electoral behavior in the United States, to the formulation of quantifiable models of economic cycles in the United States, Great Britain, and France, to a 'historical social science' in West Germany strongly influenced by Max Weber. Nevertheless, despite their differences, they all agreed that history had more kinship to the social sciences than to the humanities. This assumed, as the earlier social science-oriented history had also done, that there was a real historical past that constituted the object of historical studies. It sought to make historical method conform to the methodologies of other social sciences. Hence, it generally preferred different kinds of sources than did the older historiography; for example, hard economic

and demographic data. This made Geoffrey Barraclough comment in a survey written in the 1970s for UNESCO on trends in recent historical studies that:

> the search for quantity is beyond all doubt the most powerful of the new trends in history, the factor above all others which distinguishes historical attitudes in the 1970s from historical attitudes in the 1930s (Barraclough 1978, p. 39).

These historians tended to sacrifice individual human beings to structures and processes thus largely ignoring the individuals who compose history.

A key to almost all forms of social science history, with the possible exception of many of the historians in the *Annales* tradition, was the notion of history as an ongoing directional process. This was true of the various advocates of history as an empirical, analytical social science as it was for the different shades of Marxists and for the German school of 'Historical Social Science.' A broadly shared concept was that of modernization. It went back to nineteenth-century positivism and assumed that history, or at least the history of the modern Western world, was marked by an ongoing process towards greater scientific enlightenment followed by technological progress and by the concomitant development of a 'modern' society. While most forms of social science history tended to downgrade political history, the German school of 'historical social science' (Wehler, Kocka) sought to locate politics in the context of social structures and processes to understand how it was possible for National Socialism to gain power in Germany.

2. *Challenges to Social Science Approaches to History: Postmodernism*

This widely held confidence in social science and modernization was radically challenged in the 1970s and 1980s, most vocally by postmodernist philosophers and literary critics, but more quietly also by historians, particularly cultural historians. While it is difficult to reduce these critics to a common denominator, they agree in repudiating two assumptions that they consider to be central to the modern historical outlook. The first concerns the belief in the coherence and continuity of world history, the invention of a 'master' or 'metanarrative' (Lyotard) which sees history as an onward development to the modern Western world with its promise of enlightenment, emancipation, and well-being for the world as a whole. Practitioners of the established historiography, critics charged, had overlooked the close relationship between Western progress and imperialism abroad and the consolidation of power relations based on race, gender, and class at home taken for granted in their view of history. But history, they argued, had no coherence, thus eliminating the possibility of a grand narrative and opening the way to innumerable separate histories.

Closely related to this critical view was the rejection of the possibility of objective historical inquiry into the past. According to Peter Novick the idea and ideal of objectivity lay at the very center of the historical profession as it was constituted with its 'commitment to the reality of the past, and to truth as correspondence to that reality' and consequently 'the separation of history and fiction' (1988, pp. 3–4). This commitment, Hayden White noted, rests on the 'illusion that there is a past out there that is directly reflected in the texts' (1987, p. 209). Postmodernists according to Keith Jenkins seek to free modern historical consciousness from the notion that behind appearances there lies a solid, foundational, real world. 'In fact history now appears to be just one more foundationless, positioned expression in a world of foundationless, positioned expressions' (Jenkins 1997, p. 6).

For the postmodernists history consists of texts that are linguistic constructs with no direct relation to reality. It is language that creates reality. Jacques Derrida went further in denying that there is anything beyond the text (*il n'y a pas de hors-texte*) (see Megill 1985, p. 266). Moreover, for Derrida the text has no stable signification, but is open to an infinity of significations because there is no Archimedean point from which a clear meaning can be assigned. The text, moreover, according to Foucault is seen not only independently of its relation to the external world, but also independently of its author. Derrida criticized what he considered to be the 'logocentrism' of the classical Western philosophic tradition since Socrates that rested on the error that reality could be reduced to rational concepts, ignoring that these concepts reflected systems of domination. The idea that reality was not given but was constructed through language was further developed in feminist theory. Gender, for Joan Scott, is not found as such in nature but is socially constructed through language, texts, and cultural symbols that reflect or challenge power relationships. Thus the deconstruction of language is a necessary tool for revealing and dismantling these unequal power relationships.

All this leads to epistemological relativism. Perhaps the most influential proponent of such relativism as it relates to historiography has been White, who argued that historical thought is 'the captive of the linguistic mode in which it seeks to grasp the outline of objects inhabiting its field of perception' (1973, p. xi). Consequently it is not the purportedly objective investigations of the historian into a real subject matter which lead to knowledge about history but rather the knowledge at which the historian arrives is conditioned by the linguistic mode in which she/he operates. Professional historiography for White generates no more objective knowledge of the past than does speculative philosophy of history or the historical novel. Any attempt to go beyond the facts to a coherent narrative is a 'poetic act' so that the 'best

grounds for choosing one perspective on history rather than another are aesthetic or moral rather than epistemological' (1973, p. xii). 'Viewed simply as verbal artifacts,' he notes, 'histories and novels are indistinguishable from one another' (1978, p. 122). Similarly, 'the opposition between myth and history … is as problematical as it is untenable' (1978, p. 83).

This negation of a distinction between history and fiction is at the core of the French literary critic Roland Barthes's essay—*Le Discours de l'histoire* (*Discourse of History*). For Ankersmit, while 'science was the alpha and omega of the modernists' (1995, p. 166), for postmodernism 'the metaphorical dimension of historiography is more powerful than the literal or factual dimensions … criticizing metaphors is indeed an activity which is just as pointless as it is tasteless. Only metaphors "refute" metaphors' (p. 180). Hence research is irrelevant because 'the "real past" does not enter into historiography except rhetorically' (Jenkins 1995, p. 18).

3. *The Turn to Cultural History and the Return to Narrative*

Few practicing historians would agree with Barthes, Derrida, White, Ankersmit, or Jenkins that there are no criteria of objectivity. They would continue to hold that historical works in contrast to historical novels must proceed from research that requires standards of objectivity shared by the scholarly community. Postmodernism had reflected a general mood that was shared by many historians. Part of this mood was a general skepticism regarding the nature and quality of modern civilization and its science; the recognition that the historical science of the West, whether in its historicist or its social science form, had legitimized relationships of power in which a middle and upper-class elite had been privileged at the expense of the broad masses of the population, including women, and of the nonWestern world. Thus, by the 1970s the old scientific ethos of the social science historians was shaken. In 1979 in a now famous article, *The Revival of Narrative*, Lawrence Stone noted that in the 1970s a change had occurred in the way history was viewed and written. The belief, central to social science history, that 'a coherent scientific explanation of change in the past' (1979, p. 19) was possible, was widely rejected. In its place there emerged an interest in the most varied aspects of human existence, accompanied by the conviction that 'the culture of the group, and even the will of the individual, are at least as important causal agents of change as the impersonal forces of material output and demographic growth' (p. 9). Yet, although Stone emphatically rejected the illusion of 'coherent scientific explanation,' he nowhere suggested that historical narrative should surrender its claim to rational inquiry and realistic reconstruction.

4. *The Impact of Cultural Marxism on the Writing of History*

The new directions which historical studies took must be understood within the context of the political climate of the 1960s and after, such as the emergence of a women's movement, the rise of ethnic consciousness, the struggle for racial equality, anti-colonialism, and environmentalism which all pointed at the inadequacies of the social and economic order of an industrialized, modernizing Western world and the social science which served it. The transformations within Marxist theory and historiography provide important indicators for the change of attitude among historians. Marxism itself was in a deep crisis as indicated in the Soviet power block by the suppression of the Prague Spring and by the rejection of Marxist orthodoxy in the West.

But the significance of a non-dogmatic Marxism on historical thought in Western Europe as well as in Japan, India, and Latin America should not be underestimated. Marxism–Leninism was of little significance there as compared to Western forms of Marxism informed by Antonio Gramsci, Georg Lukács, and the Critical Theory of the Frankfurt School. Western Marxism preserved the idea of class conflict, the role of domination as a major force in social relations, but gave up the idea of a necessary development to a classless society.

The reorientation of Marxism began in the 1920s and involved an examination of the traditional economic interpretation of history. Gramsci's writings in Mussolini's prisons in the 1920s and 30s, which became known in the West only in the 1960s, explained the defeat of the communist workers movement in Italy in terms which involved factors insufficiently developed in traditional Marxist philosophies of history. For him the strength of capitalism rested ultimately not exclusively on its economic strength but on its cultural hegemony over the working classes. Lukács already in 1923 had seen in the section on 'The Fetishism of Commodities' in *Capital* the core of Karl Marx' critique of economistic conceptions of capitalism. The empiricism that dominated scientific and social science thought in his opinion failed to understand the broader context of society and culture. A similar note was pursued in the 1930s and 1940s by the 'critical theorists' of the Frankfurt School (Horkheimer, Adorno, Marcuse) who, proceeding from a critique of capitalism that they argued had turned culture into an industry in the service of economic ends, strove for a science which went beyond quantifiable data to the qualitative aspects of society and culture. They nevertheless gave up Lukács' historical optimism.

Perhaps no book had as great an impact on the reorientation of social history, at least in the English-speaking world but also beyond, as E. P. Thompson's *The Making of the English Working Class* (1963).

Thompson maintained the concept of class and the notion that 'the class experience is largely determined by the productive relations into which men are born—or enter involuntarily' (p. 9) But he stressed that class must be seen not as 'structure,' or even 'category' but as something which in fact happens in human relationships (p. 9). The English working class was thus not primarily the product of economic forces, as it had been in the *Communist Manifesto* or in structuralist, functionalist sociology, but an active agent, coming from a specific cultural tradition, which made itself as much as it was made. Hence, his reliance on qualitative elements that make up a culture such as literature, art, folklore, and symbolism. Moreover, Thompson emphatically rejected the orthodox Marxist notion that sees the past as a step to the future.

Yet Thompson's historiographical approach soon came under attack from cultural Marxists who argued that he had not broken sufficiently with classical Marxist notions of class. Even if he saw class in terms of culture, they charged, this culture still focused on an industrial working class in which those not directly linked to the industrial work process played little of a role. The *History Workshop: A Journal of Socialist Historians*, founded in 1976 in Great Britain—which in 1982 changed its subtitle to *A Journal of Socialist and Feminist Historians* and in 1995 dropped the subtitle altogether—tried to overcome this by expanding the concept of work in a changing society and including women in the family and the workplace.

From the perspectives of much post colonialist history of the 1980s and 1990s, such as put forward by Latin American *Dependencia* Theory or Indian Subaltern Studies, Marxism operated with a concept of historical development, akin to that of capitalist writing, for which modernization was understood as Westernization. Edward Said in *Orientalism* (1978) argued that Western scholarship had invented an Arab culture that rested not on historical knowledge but on preconceptions, pitting an enlightened West against a backward Muslim East which needed to be Westernized. Postcolonialist writings integrating feminist, psychoanalytical, and Marxist theories argued that categories of class, ethnicity, and gender must be seen in their interaction. They held that a Marxist conception of class conflict misses the analysis of power to which Foucault had pointed as an infinitely complex network of micro powers that permeate every aspect of social life and culture.

5. *From Macro- to Microhistory*

While few social historians could be identified as postmodernists or Marxists, certain central ideas of postmodernism and of cultural Marxism impacted a great deal of writing in social history which increasingly moved in the direction of cultural history. The idea of the unity or direction of history gave way to the concentration on small units, on the local and the marginal.

While this same orientation destroyed the credibility of orthodox Marxism, the Marxist commitment to the oppressed, even if oppression was now understood differently, occupied a great deal of the new culturally oriented social history whose subjects were most often 'the lives and feelings and behaviour of the poor and obscure rather than the great and powerful' (Stone 1979, p. 19). Anthropology and semiotics replaced sociology and economics as the key auxiliary sciences of social history. But anthropology underwent a sea change. While classical anthropology, e.g., that of Bronislaw Malinowski in the 1920s and 1930s, had tended to a static analysis of 'primitive' and 'timeless' societies, anthropologists had now become increasingly historical in outlook, recognizing that all, even the most 'primitive' cultures, evolve and that modern societies as well are proper subjects of anthropological study.

At the same time, historians increasingly turned to an interpretative, cultural anthropology. As the cultural anthropologist, Clifford Geertz, phrased it:

> Believing with Max Weber that man is an animal suspended in webs of significance he himself has spun, I take culture to be those webs, and the analysis of it to be therefore not an experimental science in search of laws but an interpretive one in search of meaning (Geertz 1973, p. 5).

In Germany, this anthropological turn in the 1980s took the form of *Alltagsgeschichte*, which sought to reconstruct the experiences and emotions of common people (Hans Medick, Alf Lüdtke, Lutz Niethammer) in earlier preindustrial as well as in modern industrial times. In Italy, historians (Carlo Ginzburg, Eduardo Grendi, Carlo Poni, Giovanni Levi) around the journal *Quaderni Storici*, began in a series of books and articles to give a firm theoretical foundation to a microhistory (*microstoria*) which reduces historical research to identifiable individuals and from these creates a history from which the net of relationships of these individuals would emerge. In place of the thousands of quantifiable data on which the older social history constructed a history of mentalities, the microhistorians wanted to recapture a fuller picture of individual experiences using transcripts and inquisitorial records in which the voices of the subjects were recorded. Outside of Italy, Emmanuel Le Roy Ladurie in *Montaillou* (1975) used such records to recapture the lives and emotions of individual villagers in a fourteenth-century village while Natalie Davis in *The Return of Martin Guerre* (1983), using judicial records, focused on gender in sixteenth-century rural France.

Nevertheless, the new orientations toward microhistory did not necessarily mean a break with the social sciences. Foucault had rejected the idea that theories can be verified because in his view the standards of verification came from a social scientific discipline which represented institutions of power;

Geertz had argued that any attempt to approach a historical subject matter with theoretically oriented questions distorts this subject and has to be replaced by 'thick description' which immerses itself untheoretically into an alien world. This, in effect, invalidated traditional forms of historical inquiry. But neither Ginzburg nor Levi were willing to go so far. For them, *microstoria* does not reject the empirical social sciences *in toto*, but stresses the methodological need for testing their constructs against existing reality on a small scale. Levi's concern in *L'eridità immateriale: Carriere di un esorsista nel Piemonte del Seicento* (1983) (*Inheriting Power: The Story of an Exorcist*) is social scientific insofar as he wishes to test established hypotheses on a local level. His central concern is patterns of power relationships in the village. In studying land purchases, Levi questions the extent to which impersonal market forces and the development of the modern state machinery determine these power relationships and argues that the decisive element in the understanding of the peasant world is 'the preservation or transmission of intangible or symbolic goods: power and prestige' (back cover of English paperback edition) Here Levi was much closer to Pierre Bourdieu's conception of 'symbolic capital' than to Geertz' 'thick description.'

The gap between social and cultural history began to close markedly in the 1980s and 1990s. On the one hand, there were important works in microhistory emphasizing anthropological aspects of culture (Medick, David Sabean). Despite the reservations these historians expressed in regard to the systematic social sciences, they relied heavily in their work on economic and demographic data and made use of computer analyses. On the other hand, historians coming from a social science orientation recognized the significant role of the cultural dimension in society. The work that the German school of 'Historical Social Science' had undertaken in the social analysis of politics now increasingly gave consideration to the interaction of society and culture. An example of this is the major comparative project on the nineteenth-century European bourgeoisie (*Bürgertum*), headed by Jürgen Kocka, which defined the latter as much in terms of culture and education as in terms of economic status. The journal *Geschichte und Gesellschaft* became an important forum for the presentation and discussion of these tendencies.

The economic analysis of class also began to play a decreasing role in the study of the revolutionary movements since the eighteenth century. By the 1970s not only the economic interpretations of the French Revolution by Marxists such as those of Georges Lefebvre and Albert Soboul, but also of their non-Marxist critics such as Alfred Cobban and George Taylor, were replaced by new viewpoints (Régine Robin, François Furet, Lynn Hunt) which placed greater stress on culture and language. As Lynn Hunt explained in the introduction to her *Politics, Culture, and Class in the French Revolution* (1984), this work conceived in 1976 began as 'a social history of Revolutionary politics' but 'increasingly turned into a cultural analysis in which the political structures … became but one part of the story' (p. xi).

Linguistic analysis plays an important role in William Sewell's analysis of the 1848 Revolution in France, in Gareth Stedman-Jones' study of British Chartism, and Thomas Childress' examination of electoral propaganda in the Weimar Republic. None of them, however, would share Foucault's radical position that 'reality does not exist, that only language exists' (quoted in Berman 1988, p. 183). Rather, they would agree with Carroll Smith-Rosenberg that 'while linguistic differences structure society, social differences structure language' (Spiegel 1990, p. 84).

6. *The Transformation of Political History*

With the move to the social sciences in the 1950s and 1960s and to the cultural sciences since the 1970s, the significance of political history has been downgraded. Various forms of social science history, as early as the *Annales* before 1945, reacted against the focus on political history in conventional historiography. Moreover, the new cultural history that sought to move away from conceptions of centralized power also consciously neglected the state. Yet this led to a certain blindness regarding the realities of the present world. The rapid changes in international relations since the collapse of communism in 1989–1991 have made the significance of politics for society and culture apparent. Recent historical theory with its emphasis on the pluralism of historical narratives has assigned the process of modernization and globalization to the sphere of macronarratives that it rejects. But one cannot escape the reality of processes of modernization, however complex and varied they may be, as significant aspects of modern historical development.

The history of politics, of international relations, of peace and war, and the study of political regimes in the twentieth century continued to occupy an important place in the historiography of the last third of the twentieth century, notwithstanding the new emphases on cultural history. And political history continued to work closely with documentary sources.

Nevertheless, as seen above in recent histories of the French Revolution, greater emphasis was given to cultural factors and to language. Something similar has happened in the studies of the Holocaust. Earlier studies had seen the destruction of the Jews as a vast and complex administrative process, as Raoul Hilberg had described it in *The Destruction of European Jews* (1961), carried out from their desks by bureaucrats like Adolf Eichmann acting on orders from above. The long debate between the 'Intentionalists' and 'Functionalists' centered on the question whether the genocide had its roots in the conscious decisions of the

Nazi leadership or developed without a clear plan in the course of the war. Both approaches failed to deal concretely with the Holocaust as it involved perpetrators and victims on the local level. More recent studies such as Christopher Browning's *Ordinary Men: Police Reserve Battalion 101 and the Final Solution in Poland* (1993) have sought to study the killing through the concrete participation of individual perpetrators.

7. Conclusion

The concerted attack on ideas of objectivity and historical method on the part of postmodernist critics has by no means resulted in the abandonment of serious historical study based on research into the sources. Although the borderline between scholarship and literature has become more fluid, as for example in Simon Schama's *Dead Certainties: Unwarranted Speculations* (1991) which points at the tremendous problems in telling a truthful history, few historians, whether in the areas of political, social, or cultural history, would abandon the commitment to honestly reconstructing the past. At the same time, the confidence of nineteenth-century professional historians in the possibility of telling definitely *wie es eigentlich gewesen* has yielded to the recognition that the same sources and the same set of events lend themselves to a variety of interpretations, but that these interpretations do not necessarily disprove one another but cast different perspectives on a subject. The years 1970–2000 have seen the collapse of illusions dear to modern historical thought until then: the linear progression of Western history and the accumulation of historical knowledge. The scope of historical study has been expanded immensely. The concentration on centers of political power and macroeconomic and macrosocial processes has been supplemented by a new interest in many aspects of life and culture which previously had not entered the imagination of historians, or at least of professional historians, and which require special research strategies. These two scales, the macro and the micro, by no means exclude each other, and in fact, as we have indicated, political and social historians have become increasingly aware of the role of culture, and cultural historians have become aware of the social and political context in which collective human existence takes place.

See also: Gender History; Historicism; Subaltern History

Bibliography

Ankersmit F 1995 *History and Tropology. The Rise and Fall of Metaphor*. University of California Press, Berkeley, CA

Appleby J, Hunt L, Jacob M 1994 *Telling the Truth about History*. Norton, New York

Barraclough G 1978 *Trends in History*. Holmes & Meier, New York

Berman A 1988 *From the New Criticism to Deconstruction. The Reception of Structuralism and Post-Structuralism*. University of Illinois Press, Urbana, IL

Geertz C 1973 *The Interpretations of Cultures*. Basic Books, New York

Green A, Troup K 1999 *The Houses of History. A Critical Reader in Twentieth-century History and Theory*. New York University Press, New York

Hunt L 1984 *Politics, Culture and Class in the French Revolution*. University of California Press, Berkeley, CA

Hunt L (ed.) 1989 *New Cultural History*. University of California Press, Berkeley, CA

Iggers G G 1997 *Historiography in the Twentieth Century. From Scientific Objectivity to the Postmodern Challenge*. Wesleyan University Press, Hanover, NH

Jenkins K 1995 *On 'What is History?'* Routledge, London

Jenkins K (ed.) 1997 *The Postmodern History Reader*. Routledge, London

Lyotard J F 1979 *La Condition postmoderne: rapport sur le savoir*. Editions de Minuit, Paris. (English edn. 1984 *The Postmodern Condition: A Report on Knowledge*. University of Minnesota Press, Minneapolis, MN)

Megill A 1985 *Prophets of Extremity. Nietzsche, Heidegger, Foucault, Derrida*. University of California Press, Berkeley, CA

Novick P 1988 *That Noble Dream. The 'Objectivity Question' and the American Historical Profession*. Harvard University Press, Cambridge, MA

Scott J 1988 *Gender and the Politics of History*. Columbia University Press, New York

Spiegel G 1990 History, historicism, and the social logic of the text in the Middle Ages. *Speculum* **65**: 59–86

Stone L 1979 The revival of narrative: Reflections on a new old history. *Past and Present* **85**: 3–24

Thompson E P 1963 *The Making of the English Working Class*. Vintage Books, New York

White H 1973 *Metahistory. The Historical Imagination in Nineteenth-century Europe*. Johns Hopkins University Press, Baltimore, MD

White H 1978 *Tropics of Discourse. Essays in Cultural Criticism*. Johns Hopkins University Press, Baltimore, MD

White H 1987 *The Content of the Form, Narrative Discourse and Historical Representation*. Johns Hopkins University Press, Baltimore, MD

Windschuttle K 1997 *The Killing of History: How a Disicipline is Being Murdered by Literary Critics and Social Theorists*. Free Press, New York

Wolfe P 1997 History and imperialism: A century of theory from Marx to postcolonialism. *American Historical Review* **102**: 388–420

G. G. Iggers

Historiography and Historical Thought: East Asia

1. Features of East Asian Historiography

The art of recalling the past has developed very differently in Europe and East Asia. Whereas, since the time of Herodotus and Thucydides, Western

historiography has been centered on histories written by individuals for individuals, since ancient times East Asian historiography has developed a historical culture pivoting on public historiography. The distinctive characteristic of East Asian historiography could be described as publicly sponsored history, a tradition that continues today as a *basso continuo*.

'East Asia' as a historic–cultural concept, comprises present-day China, Japan, Korea, and Vietnam, where history developed around China, which overwhelmed the peripheral countries with its physical size, and political and cultural ascendancy. This historic East Asia has been termed 'the Chinese world order' or a system of international relations with China at the center. The system reached Korea and Vietnam in the early Han period (202 BC–9 AD), and Japan in the first century AD.

This Sino-centric system also formed the common cultural heritage of East Asia: the classical Chinese language, and Chinese writing system; the Confucian social and familial order; Buddhism through Chinese translations; the legal and administrative systems, and the art of historiography. The peripheral states of the East Asian world might also be called Sino-graphic cultures, in the sense that they used classical Chinese as a common language, which is, in a sense, comparable with the use of Latin in mediaeval and early modern Europe.

2. *Role and Purpose of Historiography*

For the most part, the task of historiographical compilation in East Asia was traditionally a state-run project. The 'official history' produced by that compilation, along with materials collected for the purpose, constituted the core of East Asian historiography. It could be said that historiography was the primary cultural undertaking in East Asia. This is in contrast to the cultures of Europe, India, and Islam, where the concentration of cultural power has not been fixed in history.

For 2,000 years, Chinese historiography centered on the 'official history' compiled by each successive dynasty as a state enterprise. Later generations positioned the *Shiji* (*The Records of the Grand Historian*) by Si-ma Qian (145 – ca. 86 BC) as the first official history and since then, 24 official histories have been compiled. A characteristic of these official histories is that they have an encyclopedic tinge; that is, the body of the work originated by Si-ma Qian brought an entire culture, its politics, economics, society, culture, technology, etc., into one unified structure. History was written as a means of comprehensively describing such a world system.

Historiography in East Asia is perhaps equivalent to such 'primary cultural undertakings' as Biblical commentary and the *Corpus Iuris Iustanianus* in the West, the *Laws of Manu* in India, and the *Koran* in the Muslim world.

The purpose of writing history in East Asia was based on the Chinese philosophical premise that historical facts were the only certain and immutable reality. Chinese metaphysics was not premised on a revealed religion based on the existence of a unique, almighty God, like Judaism, Christianity, and Islam. Rather, it took the world as an ever-changing phenomenon, as presented in the *Yijing* (*The Book of Changes*). Therefore, it sought immutable reality in history, because human beings could not alter that which had already happened. This belief took history as its axis in Chinese civilization.

This culture of history later spread throughout East Asia in tandem with Confucianism, creating a common historical culture throughout East Asia. The proclamation of this philosophy may be found in Confucius' statement that, 'All the empty words I want to write down are neither so clear nor so startling as seeing their meaning in action' (*Shihji*).

The Chinese had developed their own way of creating an immutable past. In China and Korea, it was standard practice that, once the state's historical compilation bureau had completed compiling the official history of the previous dynasty, the bureau destroyed all the sources it had collected. This was to prevent the revision or rewriting of the official history, for once it was published by the government, the history itself took on the character of a sacred text. The most certain way to endow the official history with the imprimatur of authority was to destroy the sources on which it had been based. During the Korean Chosŏn dynasty (1392–1910), for example, the source materials were burned after use. When that was done, the account embodied in the official histories became the facts of history.

In this fashion, East Asian cultures preserved the ideal that history was the sole immutable basis for human judgment. The biographies that comprise over half the material in the official histories maintained this tradition of objective narration in their own way. In the biographies, as in other sections of the official history, they first set forth what they believed to be 'fact,' and following that, the historians added their own evaluation. This vast corpus of biographies forces us to consider why the historians believed the biography to be a necessary part of a history. It is because in a culture that lacks a unitary supreme being, the records of the lives of eminent individuals are the only true sacred texts.

3. *Historical Objectivity*

The tradition was maintained of seeking objectivity in historical narration in the lives of the historians who wrote it even more than the narration itself. We can

find an episode that presents the historian as a lofty figure in the life of a fifteenth-century Korean historian.

It is recorded that in 1431, as the compilation of the *T'aejong sillok* (*the Veritable Records of King T'aejong*) was nearing completion, his successor King Sejong (1418–1450) asked the compilers to show him their work in advance. 'In the previous dynasty, every monarch personally reviewed the veritable records of his predecessor; but King T'aejong did not review the Veritable Records of King T'aejo.' Sejong's senior ministers replied that, 'If Your Majesty were to review [the work in progress], later monarchs would surely revise [the historians' work]. Then, [future] historians would suspect that the monarch might look at the draft, and they would inevitably fail to record the facts completely. Then, how would we transmit [facts] faithfully for the future?'

In traditional East Asia, the role of the past was to serve as a normative history. This forms an interesting contrast to Western historical practices, where it evolved as a cognitive discipline. That is, in the West, the historian found his *raison d'être* in rewriting the past. The discipline of history developed as a competition among the interpretations and approaches of different historians.

'History' in modern East Asia is generally parsed with the two-character compound Chinese word *lishi* (Japanese, *rekishi*; Korean, *yoksa*), but until the middle of the nineteenth century it was expressed in the single-character term *shi* (Japanese, *shi*; Korean, *sa*). This word *shi* itself originally denoted the historian himself, and only later, by analogy, did it come to refer to the product of the historian's craft, the written work of history itself. This is particularly important when we note that in the languages of Europe, it was precisely the reverse: The word 'historian' was formed to refer to the maker of the written 'history.'

It is a notable characteristic of historical consciousness in East Asia that it tends to focus far more on the attitude of the historian, than on actual works of history themselves.

4. Development of East Asian Historiography

4.1 China

In China, Sima Qian substantially established the tradition of Chinese historiography by inventing his own format for compilation. He succeeded his father as the grand historian of the Han emperor, and with his access to court archives, he completed the *Shiji* (*The Record of the Grand Historian*) in 130 volumes. It covers a history of three thousands years from the (legendary) Huang-di dynasty to his own period, and covers China and Central and South Asia. The format consists of five sections: 'Benji,' or imperial annals (12 volumes); 'Biao,' a chronology of the successive dynasties (10 volumes); 'Shu,' treatises on economy, technology, astronomy, etc., (eight volumes); 'Shi-jia,' annals of feudal states and nobles (30 volumes); and 'Liezhuan,' biographies of famous persons and foreign peoples (70 volumes). Sima's format is encyclopedic in nature, systematizing the world in a single, unified structure; his historiographical format dominated East Asian historiography until the mid-nineteenth century.

Ban Gu (32–92 BC) followed Sima Qian's work, but limited himself to the history of a single dynasty, which he completed as the *Hanshu* (*History of the [former] Han dynasty*). His practice, that each Chinese dynasty compiled the history of its predecessor, became standard for all of the following official histories. It was in the Tang dynasty (618–906) that historical culture truly flourished, as history became an independent subject in the Chinese classification of learning. Historians were closely linked with governmental bureaucracy and their team method of compilation started in this period and continued until the twentieth century.

A new format for compiling historiography was proposed by Sima Guang (1019–1086) in his 294-volume *Zizhi tongjian* (*The Comprehensive Mirror for Aid in Government*) in 1084. It was an extensive historiography that employed the method of chronological order to help readers find certain historical events through dynasties. It covers sixteen dynasties of 1,362 years and includes information from dynastic histories, chronicles, biographies, novels, etc. It is necessary to refer here also to the Hanxueh (the Han School), the development of which established historical study based on textual criticism of historical documents in the Qing dynasty. Qian Daxin (1728–1804) was one of the great figures among them, whose *Nianershi kaoyi* (*Verification of the text of the 22 histories*) was an examination of characters and language, dates and places, genealogies and institutions in the text of the 22 official histories.

4.2 Japan

The idea of official history spread from China to the other East Asian countries that compiled their own official histories, modeled on the Chinese format and written in classical Chinese, not in their native languages.

The first Japanese state historiography was the *Nihon shoki* (*The Chronicle of Japan*, 720) compiled by Prince Toneri, and modeled after the Chinese format. It covers the period from the Age of Gods to the end of the reign of Empress Jitô (686–697), arranged in chronological order, in annalistic form. Toneri used history as a forum to establish Japanese national identity and consciousness. This is in marked contrast to the *Kojiki* (*Record of Ancient Matters*, 712) by Ôno Yasumaro, which was also a historiography based on a transcription of oral tradition.

This tradition continued until the late nineteenth century. The *Honchô tsugan* (*Comprehensive Mirror of Japan*, 1644–1670), compiled by Hayashi Razan and his son Hayashi Gahô, was also in the same line of official history sponsored by the Tokugawa shogunate, and the *Dai Nihonshi* (*the History of Great Japan*, compiled 1657–1906) was compiled under the support of the Mito clan.

Apart from the historiographical tradition, it is important to add such auxiliary sciences as bibliography, classics, textual criticism, and philology, which had developed in Japan to the same levels as in the West by the nineteenth century. These originated, in Japanese tradition, from the *Chronicle of Japan* (720), of handing down variant texts or conflicting readings and traditions intact to posterity. In this important respect, China differed markedly from Japan. In China, history was transmitted as an authentic, authoritative text, which admitted no variant. This difference is richly illustrated by the work of Hanawa Hokiichi (1746–1821). His *Gunsho ruijû* (*Classified collection of Japanese classics and documents*), begun in 1779 and still in progress, first appeared in 1819, with 1,270 titles in 530 volumes; another 2,103 titles in 1,150 volumes appeared in 1822. The objective was not to establish the authoritative text and obliterate the others, but to make available all the variant texts, which later generations might judge. This collection compares favorably with such European endeavours as the *Monumenta Germaniae Historica*, the *Rerum Italicarum Scriptores*, etc.

The Meiji government, established in 1868, which was responsible for the restoration of the emperor as the focus of sovereignty, and led the subsequent modernization of Japan, also continued the tradition of 'official histories,' founding the Office for the Compilation of National History. The aim here was to create an official history of Japan, but for various reasons the objective was amended to that of compiling historical documents, and the original idea of compiling a national history was abandoned. The work of the Office continues today in its successor, the Historiographical Institute, now part of Tokyo University. Moreover, virtually every prefecture, district, and locality in Japan, even down to the level of individual villages, continues the tradition of official history at the local level today, compiling local historical sources, and publishing local histories. Similarly, many individual government bureaus, as well as corporations, universities, and other corporate entities, regularly commission their own histories, extending the model of 'official history' to the private realm.

4.3 Korea

It was around the fourth century that Korea founded its historiographical tradition, though the epoch-making historiography did not appear until the twelfth century. Kim Pu-sik (1075–1151) compiled the *Samguk sagi* (*History of the Three Kingdoms*) in 1145, the oldest extant history of Korea, covering from antiquity to the decline of unified Silla in 935. It follows the format of Sima Quan's official history as is shown by the name *sagi* (the Korean pronunciation of *Shiji*), and he intended to provide moral lessons for posterity. Iryŏn (1206–1289), a high-rank Korean Buddhist monk, compiled the *Samguk yusa* (*Memorabilia of the three Kingdoms*) in 1281, which is the only extant history of Korea before the Chosŏn dynasty (1392–1910), together with the *Samguk sagi*. This work is unique in the sense that he wrote it privately, including myths, tales, and social issues, and even refers to the sources he used.

So Kŏ-jŏng (1420–1488), compiled the *Tongguk t'onggam* (*Comprehensive Mirror of Korea*) as a Korean official history in 1485, modeled on Sima Guang's *Zizhi tongjian*. It covers the period from ancient times to the end of the Koryŏ dynasty in 1392. This work was compiled under the influence of neo-Confucianism, and reflects the moralistic purpose of praising good and condemning evil. An Chŏng-bok (1712-1791) likewise published his *Tongsa kangmok* (*Outline of Korean History*), 20 volumes, in 1778, a chronological treatment of Korean history from Tan'gun, the legendary founder of Korea to the decline of Koryŏ in 1392.

4.4 Vietnam

In Vietnam, just as in Japan and Korea, some official histories were published including *Dai Viet Su Ky Toan Thu* (*Complete Book of the Records of Vietnam*) in 1479, which is based on the style of the *Zizhi tongjian* of Sima Guang.

5. Historiology in East Asia

The study of historical theory in East Asia began with Liu Zhiji's *Shitong* (*Comprehensive Historiography*), written in China in the eighth century (708). It was the first thorough treatise in Chinese—or any other language, for that matter—on historical criticism, which also constituted a history of Chinese historians and a theory of historiography, and deals with such questions as selection and criticism of historical materials, historical objectivity, judgment in history, historical causality, etc. The work was a constant reference for Chinese historians until the nineteenth century, and gave rise to the meta-historical tradition in Chinese historiography.

Liu's theory of history was further developed by Zheng Qiao (1104–?1162) in his *Tongzhi* (*Comprehensive Monographs*, 200 volumes). In the introduction to the *Tongzhi*, he proposed his core idea of 'synthesis' and 'inter-relatedness and causality,' through which he argued the importance of the continuity of his-

torical development. Based on this theory he completed the voluminous *Tongzhi*, including the 'Twenty Summaries,' his noteworthy encyclopedic chapters.

Zhang Xuecheng (1738–1801) wrote his *Wenshi tongyi* (*General Survey of Literature and History*), a work on the theory and history of historiography (only published in 1920, long after his death). The first chapter starts with 'The Six Classics are all history,' in which he expressed his philosophy of history, that the very idea of the Sacred (*tao*) can only be known through men's actions (*shi*), which are recorded in the *Six Classics*.

It is thought provoking to realize how few works of historical criticism were written outside China, in comparison with the voluminous amounts of historical works. It signifies that traditional East Asian historiography was normative in nature, in the sense that it required little discussion of the nature of history. It is also necessary to refer to the fact that there were relatively few works of metaphysical interpretation of the course of history—save perhaps Wang Fuzhi's (1619–1692) discussion of historical logic and historical tendency in work published only after his death. East Asians valued detailed precision in history, rather than the philosophical interpretations that were developed in Europe. Some exceptions, informed by Buddhism, were found in medieval Japan. Jien (1155–1225), a prominent Japanese Buddhist cleric, wrote his self-deprecatingly titled *Gukanshô* (*Notes of the Views of a Fool*, 1220), in which he proposed his philosophy of history with the term *dôri*, an idea of historical inevitability, through which he explained historical processes in terms of the state of development and proposed his own periodization. Similarly, a century later Kitabatake Chikafusa proposed in his *Jinnô Shotoki* (*Chronicle of the True Descent of the Divine Sovereigns*, 1339) that Japan is a land founded by the gods and maintained by the blessings of an unbroken lineage of divinely descended emperors.

6. *The Rise of Modern Historiography in East Asia*

Western concepts of 'civilization and enlightenment' were introduced systematically to East Asia beginning in the mid-nineteenth century had a transformative effect on notions of history and historiography as well. The movement toward 'modern' and 'scientific' history in East Asia began with a variety of intellectual experiments among Japanese scholars in the 1870s and 1880s. As Korean and Chinese intellectuals ventured to Japan toward the end of the nineteenth century, 'modern' Western notions and methods of historical inquiry and explanation gradually filtered across the entire East Asian region.

It is worth special mention that Western cultures and technologies including historiography were introduced through translations of European and American works into Japanese. Japanese intellectuals created their own neologisms to translate Western terms by combining two or three Chinese characters together. Those newly invented terms in Chinese characters were later exported to China and Korea, and in consequence, East Asian scholars now share most academic terms in the liberal arts and sciences.

The very fact that such translations were feasible, and successfully received, demonstrates the stage that Japanese historiography had achieved. By introducing Western historiography through translation, rather than keeping it as the preserve of a few language specialists, Western history and historiography were transformed into common intellectual property for all East Asian historians, and, in particular, made accessible to those who worked on their national histories.

As far as historiography is concerned, there already existed a traditional Japanese historiography, and the auxiliary sciences made possible the introduction and importation of Western historiography. Native pioneers in the introduction of Western historiography, historians like Shigeno Yasutsugu (1827–1910) and Kume Kunitake (1839–1931), were initially trained in *Kôshôgaku* tradition (Chinese, *kaozhengxue*: textual criticism, which originated from the so-called 'Han School' in the Qing period), and were thus well prepared to accept the empirical methods of Leopord von Ranke. They were particularly receptive to his notions of a rigorously critical approach to sources and the avoidance of value judgments, when his student Ludwig Riess, who was invited to Tokyo Imperial University as a professor of history in 1887, introduced Ranke's ideas to Japan.

The acquisition and acceptance of European historical theory was started by Japanese translations of leading contemporary European works on historical method and theory, such as those by J. G. Droysen, E. Bernheim and C. V. Langlois and C. Seignobos. Under the strong influence of Bernheim, Tsuboi Kumezô published his *Shigaku kenkyûhô* (*Methods of Historical Research*, 1903) in Japan, and Liang Qichao (1873–1929) published the *Zhongguo lishi yanjiufa* (*Methods of Chinese Historical Research*, 1922) in China.

'To study history' in traditional East Asia had meant 'to become conversant with the official Chinese histories,' and the introduction of Western ideas of historiography opened new horizons for students of history, who began to inquire, 'what are the historical facts,' in the sense of going beyond the historical narrative to examine the underlying facts themselves. It was an indication of the opening of modern historical scholarship in East Asia, in which the introduction of Western learning played an important role.

A typical example of this consciousness is found in Shigeno Yasutsugu, nicknamed 'Doctor Obliteration.'

Applying the European historical method to Japanese history, he re-examined the *Taiheiki* (*Chronicle of the Great Peace*, 1370–1371), and proved that the fourteenth-century military commander Kojima Takanori never existed.

In China, Gu Jiegang (1893–1980) was one of the most eminent figures, whose *Gushibian* (*Critiques of Ancient History*, seven volumes, 1926–1941) contributed greatly to the modernization of Chinese historiography by taking a finely honed critical scalpel to the world of ancient Chinese legend, which had long been accepted as fact. His school was named 'ancient doubters.' His method was rooted in the Qing dynasty (1644–1911) school of ancient doubters; however, it was at the same time backed with the new European historical method.

Yi Pyŏng-do (1896–1989) was the first Korean historian to have received training in the modern historical method. He stressed critical and objective methods of treating historical sources in his *Han'guk-sa taegwan* (*Grand Survey of Korean History*, 1963).

With the introduction of Western culture to Japan, history became re-institutionalized from the mid-nineteenth century onwards, from the governmental section to the newly established Imperial University. However, it should not be forgotten that the introduction of Western historiography reduced East Asian historiography from its former position as a state enterprise to merely a single academic field within the university. Individual historians started to have their own 'specialty' in order to survive in this new 'university' setting, and in order to distinguish themselves from eminent amateur historians, their researches were published as monographs in the newly established historical journals.

On the one hand, in the West, the 'professionalization' of historical research resulting in the establishment of history as an independent university academic field was a great achievement for this institutionally neglected discipline. In East Asia, on the other hand, the introduction of modern Western historical research heralded the end of the prestigious traditional East Asian historiography, which had aimed at a comprehensive description of the entire world. Since then, the subsequent evolution of East Asian historiography has developed largely along the lines of 'modern' historiography, in the same basic methodological and theoretical terms as the rest of the 'advanced' world.

7. *Development of Modern Historiography*

Since modern historiography was introduced to East Asia, its development varied according to the situation in each country.

In China, along with Gu Jiegang's ancient doubters, newly discovered archaeological results urged a new perspective of ancient history on Chinese historians. In particular, the discovery of oracle-bone inscriptions laid the foundations for the new Chinese ancient history led by such historians as Luo Zhenyu (1866–1940) and Wang Guowei (1877–1927). These new works were later synthesized with Marxist materialism by Guo Moruo (1892–1978). His *Zongguo gudai shehui yanjiu* (*Study of Ancient Chinese Society*, 1930) is the first Marxist work on ancient Chinese society. Thereafter, especially with the establishment of the People's Republic of China in 1949, Marxism became dominant among Chinese historians. Traditional historiographical achievements were mostly replaced by reverse interpretations, a trend which was highlighted in the period of the Cultural Revolution (1966–1976). After the 1990s, the worldwide information revolution and the introduction of the latest historical methods of social history are changing Chinese historiography, opening it to the rest of the world.

In Japan, soon after the introduction of modern European historiography, the emperor-centered state promulgated a new political orthodoxy, and began to restrict intellectual freedom, especially in the study of ancient Japanese history. Since the 1930s, social and economic history began to apply the Marxist framework to Japanese history, mostly to the study of Japanese capitalism. After World War II, Marxism emerged as the dominant historiographical school, and it was only in the 1970s that Japanese historians began to accept the idea of social history as their latest historical method from France, Germany, and England. Since the late 1980s, many Marxist historians quietly shifted from socioeconomic history to social history.

8. *Metamorphoses of Public History*

Currently, in East Asia, the aura of official history still floats on the surface of peoples' collective historiographical consciousness. The past is still seen as the trustworthiest 'mirror on mankind.' What is being sought from official historiography is not 'a' history, but 'the' history, the reliability of which, people undoubtedly believe, is guaranteed by public editorship.

In China, 24 official dynastic histories have been published and the tradition of public history continues to the present day. The *Qingshigao* (*Draft of the History of the Qing Dynasty*) was completed in 1927, and was finally published in Beijing in 1977. This applies to local history too; the publication of prefectural histories sponsored by local governments began in the Ming dynasty (1368–1644), and continues even in contemporary China.

In Japan, this idea is seen in school history textbooks and local histories. The Japanese Ministry of Edu-

cation vets all school textbooks, the idea of which seems a reincarnation of official history as far as history textbooks are concerned. Japanese schools teach students 'what' rather than 'why.' They cram students' heads with a fixed course of history – Japanese history, world history, whatever – without much attempt to have students recognize that there are differing viewpoints. Forty-two Japanese prefectures out of 47 have compiled their own prefectural histories – some of which have been compiled two or three times over the last 120 years – and there is hardly a city, town, or village anywhere in Japan that has not done likewise. These local histories are premised on the traditional historiographical notion that there is a historical truth that is embodied in authoritative narratives.

Korea also continued to chronicle an official history and this tradition is alive in present-day South Korea. The 25-volume *History of the Republic of Korea*, whose publication was completed in 1979 by the National History Compilation Committee (established in 1948), is a state-produced history of the homeland.

Since the late nineteenth century, the practice of history has undergone professionalization and specialization on a worldwide scale, and East Asia too has been drawn into this whirlpool. Since then the style of historiographical practice in East Asia has, in most respects, been the same as that in Western countries. The work of the historian, as an individual, is to write many research papers and books. However, many historians, as far as historians working on their national history are concerned, are also at the same time involved in publicly sponsored historiography; countless university-based Japanese historians, for example, serve on the editorial boards of local and prefectural history projects. We can thus observe the continuing 'coexistence' rather than 'conflict' of the two differing traditions of historical study.

To put it another way, East Asian historians simultaneously inhabit two worlds of historiography, one a world of Rankean (or traditional East Asian) factual objectivity and another world of post-Rankean interpretations 'varied and opposing voices.'

See also: Area and International Studies: Development in Southeast Asia; China: Sociocultural Aspects; East Asia, Religions of; East Asian Studies: Politics; East Asian Studies: Society; Japan: Sociocultural Aspects; Korea: Sociocultural Aspects; South Asia: Sociocultural Aspects

Bibliography

Beasley W G, Pulleyblank E G (eds.) 1961 *Historians of China and Japan*. Oxford University Press, London

de Bary W T, Wing-tsit Chan, Watson B (eds.) 1960 *Sources of Chinese Tradition*. Columbia University Press, New York

Han Yu-shan 1955 *Elements of Chinese Historiography*. W H Hewley, Hollywood, California

Ku Chieh-kang (Gu Jiegang) 1931 *The Autobiography of a Chinese Historian*. Trans. Hummel A W, E J Brill ltd., Leyden

Maruyama M 1974 *Studies in the Intellectural History in Tokugawa Japan*. Tokyo University Press, Tokyo

Nivison D S 1966 *The Life and Thought of Chang Hsueh-ch'eng*. Stanford University Press, Stanford, CA

Schneider L 1971 *A Ku Chieh-kang and China's New History*. University of California Press, Berkeley, CA

Watson B 1958 *Ssu-ma Ch'ien: Grand Historian of China*. Columbia University Press, New York

M. Sato

Historiography and Historical Thought: Indigenous Cultures in the Americas

1. Terminology

History is information handed down from generation to generation. Since information can be inherited in the form of natural objects as well as by culture itself, a basic distinction needs to be drawn between natural history and the history of civilization. The way people develop consciousness of their history is by taking information from their natural environment, their cultural artifacts, and their memory, which is then collectively conceptualized by means of interpreting, reconsidering, and agreeing on its meaning. Historical concepts resulting from this always reflect the people's experience of their natural environment as well as their specific cultural achievements. In exceptional cases a historical concept can be limited to the individual. However, historical research is more concerned with collective concepts, since it is characteristic of human societies that mainly those dominant concepts are handed down. Thus each society cultivates a particular historical concept, which again needs to be handed down in order to become and remain effective. All cultures are alike in inheriting their historical concepts in the form of oral (stories and myths) and figural traditions (monuments). In addition to this, civilizations often use writing (books and inscriptions) as the preferred means for handing down historical concepts to future generations.

2. America

The only people of pre-colonial America to develop a fully functioning writing system were the Lowland Maya. They used the written records to maintain knowledge of a wide thematic range and thus stored and conserved an enormous amount of cultural

information from which they formed their own distinctive historical concept. Rather incomplete writing systems (picture-writing) were characteristic of their northern and western neighbors—that is in historical order the Epi-Olmecs, the Zapotecs, the Mixtecs, and the Aztecs. Taken together, these cultures form the core of the cultural area Mesoamerica, characterized by similar, and in many respects uniform, conceptions of history, and therefore can be treated to all purposes of this article as a unit.

On their first contacts with indigenous peoples of the Americas, the European conquerors came across various forms of picture writing that was used to note down historical events. This even occurred in some areas of North America, predominantly in the eastern woodlands and on the plains and prairies west of the Mississippi. South America was in this respect quite the reverse, even though pictorial representations in painting (Moche, Nasca), weaving (Paracas, Nasca), and rock engraving (Nasca) were highly elaborated.

The description in Sect. 3 is mainly restricted to Mesoamerica, which is distinguished from all other cultural areas by its emergent writing systems.

3. *Components of Mesoamerican History*

The Mesoamerican *Weltbild* shows three basic components, inasmuch as the world is conceived of as consisting in discrete and ordered entities, often arranged in cycles. Connections and explanations of meaning beyond the historical facts make use of analogy, i.e., interpreting something that is identical or similar in form to be also identical in substance or related in a causal way. Another method is the reification of originally mental constructs and their material (picture-written) form of expression. This is particularly significant if information is handed down over a long time and cannot be clarified independently from written or pictorial sources. These and other methods of interpretation were operating in historical traditions, which is why they contain to a great extent patterns, topics, and stereotypes extraneous to the historical facts.

Mesoamerican cultures have also developed methods of dating. A cyclical and public calendar based on the solar year of 365 days, in addition to a ritual 260-day almanac for private and personal use that is likewise cyclical and parallels the solar calendar, are the foremost methods of dating historical events. The Lowland Maya included in their historical writing the calculation of temporary celestial phenomena like the lunar, Venus and Mars cycles, and observed the positions of planets in constellations of fixed stars; comparable to the Zodiacs of the Mediterranean cultures. They even managed the complex arithmetic of interrelations between some of these celestial cycles. Yet they accomplished only the arithmetic calculation of it, but never presented a geometrical model of the world's system corresponding to the Ptolemaic and Copernican models.

4. *Historical Topics*

By means of chronology Mesoamerican people created histories (theogonies and myths) of anthropomorphic and theriomorphic deities. This is conveyed in detail particularly for the Aztecs through the *Legend of the Suns* (Lehmann 1938) and the famous calendar stone. Mesoamerican people believed in the existence of four or five successive world ages, each with duration of a few thousand years. Each age is quite distinct with different creatures inhabiting a world, which is then destroyed in cataclysm by water, fire, and air (wind)—the reader may note the analogy to the ancient Greek theory of the elements. A new creation marks the beginning of each world age. Not even the contemporary world age of the Aztecs around 1500 was considered to be of unlimited duration. Within the chronological framework of this age man appears relatively late, and more detailed chronological units (years, year bundles of 52 years) apply to his history. Yet Mesoamerican people connect themselves occasionally, and in a speculative way with the world of their deities anteceding the human epoch. This holds especially true for the ruling dynasties of the Maya who very often consider themselves to be descendants of the gods of preceding world ages (Palenque). Such a Mayan world age lasts exactly 1,872,000 days, which is about 5000 years. The duration of the world ages in Central Mexico is not exactly known, since their chronology is technically less developed. The interrelation of the people with their pantheon becomes, furthermore, apparent in the tutelary deities of each tribal group. These are commonly organized in city-states during Postclassic times: Quetzalcoatl for the Toltecs (capital: Tollan) around AD 900–1100 and for the Chololtecs (capital: Cholula) around AD 1000–1500, Tohil for the Quiché (capital: Utatlán) around AD 1000–1500, and Huitzilopochtli for the Aztecs (capital: Mexico) around AD 1350–1500. In this manner origin-related, trans-historical interrelations take form in a similar way as they are known for the civilizations in the Old World (Sumer, Ancient Egypt, Babylon, Jewish culture of the Old Testament).

Tribes with their leaders, the ruling dynasties arising from them during the process of getting sedentary and their tutelary deities are the main actors of Mesoamerican history. Rulers are described with their most important *rites de passage* (birth, inauguration, marriage, procreation of offspring, death), martial triumphs, and—to a lesser extent—civilizing deeds (introduction of cultivated plants or technical skills). There exists the idea among Mesoamerican tribes that they originated from a distinct place, which is the tree

of Apoala for the Mixtecs, the Seven Caves or the island of Aztlan in the center of a lake for the Central Mexicans and the Totonacs; and this origin is placed in an apparently rather historical not very remote time; according to the Aztec main tradition this happened in the year AD 1064. In some areas, the version of an origin in an even more historical setting, namely the city of Tollan can be encountered (Cakchiquel). This Tollan was the capital of the famous Toltecs, who were actually a historical people, admired for their skills and accomplishments. Migration myths, apart from being descriptions of historical events, which are to a certain extent factual—as far as can be concluded from language history and geography—seem to have functioned as a means of political legitimization. The successfully migrating and thus surviving tribal group legitimizes its right to rule over the traversed territories and the cities within, or at least to exert indirect political control in case the areas in question are too distant for factual dominance. After settling down, topics such as marriage alliances, conspiracies, insidious martial acts, and heroic self-sacrifice appear in the form of historical episodes. It is in these later times, when major wars between the different Mesoamerican polities become a subject of history. In Central Mexico it is learned that the Mexicans throw off the yoke of Tepanec rule by Azcapotzalco and gain completely independent status, prestige and property of lands. In the Southern Maya Lowlands the city-states of Calakmul and Tikal are at war with each other for hegemony of the region; and the Guatemalan Highlands are the scene of frequent warfare between the Mam, Quiché, and Cakchiquel. Since the range of historical accounts rarely transgresses the own polity and is immediate neighbors and never exceeds the limits of the Mesoamerican core area, 'world history' from the precolonial Indian perspective is very limited. Nevertheless this compares, albeit on a more narrow basis, with antique and medieval concepts of the world in Near Eastern and European cultures. No statements revealing knowledge or speculation about countries and peoples beyond this horizon can be found, even though archaeological investigations indicate external contacts in the form of trade relations with peoples far outside Mesoamerica, namely with the North American Southwest, with Panama, and with Ecuador. This seemingly odd and willful limitation of the historical horizon is characteristic for many early states and for most chiefdoms worldwide.

It is men who predominate the historical records as well as the actual political scene, and these are generally members of the nobility as all Mesoamerican societies were ranked. The historical records give special emphasis to particularly successful individuals: such as to ruler '8 Deer Tiger Claw' among the Mixtecs or to the rulers 'Pakal' of Palenque and 'Kakupakal' of Chichén Itzá among the Lowland Maya. In addition to politically active members of the ruling nobility, historical records of the Maya and the Aztec convey a few names of artists (painters, sculptors, poets, and performing artists), thus demonstrating to some extent interest in cultural and art history.

Taking into account that political history is predominantly the subject of men, the reader's attention must be drawn to the occasional significance of women among the Lowland Maya, especially in the city-states of Palenque and Yaxchilán around AD 600/700. These women presumably never equaled men in rank, but gained prominence in their patrilineal societies only in order to substitute for a male heir and to uphold dynastic continuity. It remains unclear whether this role involved real implementation of power or whether it was confined to representational duties, although—in keeping with ethnological theory of gender studies—a rather representative and symbolic function is suggested. Apart from this, noble women appear as objects in marriage alliances that serve to enlarge the ruling territory and become instruments of political provocation (loot, rape, and sacrificial death).

5. *Indian Reflections on History*

It is impossible to give a definite description of historical thinking among the Mesoamerican Indians, since no explicit statements related to this can be found in the historical sources. It is, however, manifest that man is controlled and directed by deities to whom, he for his part, is compelled to pay tribute in the form of prayers, offerings, and cult (there are 18 major monthly feasts per year) in order to maintain and ensure the world's existence. This is most evident in the Aztec's dogmatic obligation of invigorating the sun on its course with sacrificed human hearts. This compulsion to human sacrifice led to those sacrificial excesses that were later witnessed by the Spanish conquerors. Prior to the conquest, several Indian philosophers had already criticized and perceived the practice of human sacrifice as incorrect by expressing their desire for the idealized Toltec times (around AD 1000) to return, when people sacrificed flowers and animals instead of human beings.

The general Mesoamerican outlook on life is pessimistic and fatalistic, which manifests itself especially in the divinatory rituals of the Maya, Zapotec, and Aztecs where negative prophecies are outnumbered by the indifferent or positive ones. The collective fatalistic behavior of the colonial Yucatec Maya and—two centuries later—of the Itzá concerning their own political fate might serve as an example of this common way of thinking. However, it needs to be taken into account that the Yucatec case could well be the result of traumatic experiences during the Spanish conquest; since only colonial sources give information on this topic, it is difficult to decide, whether fatalism is a traditional or an adopted behavior. Human life is

considered to be hard and fateful (like in Old Jewish faith), controlled by deities, and improvement can only be gained by religious manipulation. Consolatory is hence the Aztec concept of the otherworld, where sweet paradise awaits those who endure dangerous labor and hardship on earth, favoring warriors and childbearing women. Yet we lack a theologically differentiated comprehension of the Aztec idea of an afterlife, since this is not further described in the historical accounts. The Lowland Maya left more written and pictorial documents regarding this topic, although these have not been systematically analyzed thus far.

Rarely, reflections on antecedent cultures can be found among the descriptive historical traditions of the Central Mexican Aztecs. The major city Teotihuacan, which had been abandoned prior to Aztec times, is taken by myth as the place of origin of the gods. Earlier cultures are remembered by conserving their artifacts and imitating their products. The idealization of such a past culture—the Toltecs—has been mentioned earlier.

6. *Traditions and Sources*

The numerous monuments from lasting material (stone), although not yet completely understood, are an important source for historical thinking. They date back as far as 1000 BC: Beginning with sculptures portraying Olmec rulers and slabs illustrating Zapotec conquests (so called *danzantes*). The first real inscriptions date from the first two centuries of the Christian era, called the Epi-Olmec period. Especially noteworthy is the stela from La Mojarra, although its long inscription presumably representing the Proto-Zoquean language of the Isthmus region is not fully deciphered yet.

Vase paintings of the Maya with their pictorial and partially narrative depictions and hieroglyphic explanations are one of the chief sources for reconstructing the Mesoamerican classic from AD 300–800 AD. These vessels show painted scenes from royal palaces, royal visits, payment of tribute, or the hunting of deer and birds. All illustrations do probably depict scenes from the otherworld. Since daily life can be assumed to have served as the *calcque* for the concept of this otherworld, it is possible to reverse this relation in deducing every-day life of the Maya from the vase paintings. Vase paintings have been similarly exploited for reconstructing daily life of the South-American culture of Moche, which was contemporaneous but unrelated to the Maya.

Further sources of historical records are manuscripts with historical contents, which—due to their fragility—generally do not date back too far. About a dozen precolonial manuscripts and perhaps a hundred copies and extracts compiled during early colonial times have been preserved. Most are of Central Mexican (*Codex Huitzilopochtli*) and Mixtec (*Codex Nuttall*) origin.

Christian missionaries introduced the writing with the Latin alphabet in the early colonial times (AD 1520–1650). Written documents of this type are especially numerous in Central Mexico. Basically European in form, and to some extent even in content, the object of these written sources is almost completely related to the colonial situation (lawsuits dealing with inheritances, properties, and privileges). The writings obviously connect pre-Spanish traditions of picture writing with oral traditions and form a new integrated genre. In the Guatemalan highlands (*Popol Vuh*, *Annals of the Cakchiquels*), as well as with the Yucatec Maya (Books of *Chilam Balam*) these sources are not only less numerous but show less European influence. From other Mesoamerican regions only a few historical documents have been preserved, written in the Spanish language by Spaniards who drew on remembered information and interviews. Thus Gonzalo Fernández de Oviedo y Valdés reported on the Nicarao who lived at the southern periphery of Mesoamerica; the Franciscan missionary Juan de Torquemada conveyed comparable traditions of the Pipil; and the accounts of the Mixtecs by Francisco de Burgoa were comprehensive but a little concise.

A major European element in sources, which have been written in the indigenous languages but with Latin letters, is the synthesis regarding world history and natural history. This synthesis can be noted, for instance, in a general chronology drawn from the uncoordinated lists recording royal succession in individual Indian city-states as presented by the *Annals of Quauhtitlan* and the works of Domingo Chimalpahin. As suggested by its subdivision into 12 books, the topics, and the illustrations, the *Historia General* of the Franciscan friar Bernardino de Sahagún obviously resembles medieval European natural history and its antique origins (Pliny the Elder).

7. *Problems of Research*

Without explanations from the creators themselves, it remains difficult to understand form and symbolism of the great range of historical objects and sources, even though researchers have presented seemingly comprehensive and highly sophisticated interpretations (in particular Eduard Seler around 1924–5). Hieroglyphic captions of Mayan vase paintings, which have been subject to serious research since the 1980s, are—although seemingly well understood—overly standardized (e.g., 'this is the beaker for cacao belonging to X') and, therefore, still meager in cultural content and fuzzy in context. Most inscriptions that designate singular objects like 'earring' or '[sack full of] beans' are short and do neither contribute to a differentiated understanding of the depicted scene, since they only name the object, nor do they help much in a broader

comprehension of local history, since they only mention the owner or receiver of said object.

Although theoretically achievable, the decipherment of pictorial and hieroglyphic writing systems (Zapotecs, Epi-Olmecs, Maya, Mixtecs, Aztecs, Tetzcocans) has not been completed yet and not all documents have been thoroughly studied. The research strategy regarding these documents assumes a uniform paradigm underlying all different regional forms of pictorial and syllabic writing (Maya, Mixtec, Aztec, Tetzcocan, etc.). Nevertheless, painstaking and detailed analysis of each document, as well as methodological reasoning in decipherment and interpretation are required. Investigation should strategically begin with thematical complexes that are especially revealing because their cultural contexts are well understood, such as burials at major archaeological sites containing inscribed goods.

Texts from early colonial times (until about 1650) that have been written down in the indigenous languages by means of the Latin alphabet are easier to approach, although they are still not very well understood either. Most of these documents originated with the Aztecs, some of them with the Yucatec and Highland Maya, while there are only very few from other language groups and regions. Whereas the more technical problems of translation can be met by increasing language competence, the main problem of ethnohistory (the technical term for the discipline of history that analyses sources in the writing-systems and languages of the people studied) is to grasp concepts and metaphors, etc. from the *Weltbild* of the Indians. These are then interpreted in terms of European historical thinking, and their authenticity elucidated with regard to events that have actually come to pass.

Devoting more time to research of chronological problems is most likely rewarding, since Mesoamerican chronology is not only of high accuracy but also omnipresent in historical records. For the time of the Maya Classic from AD 300–900 there is not a single year without several exactly datable historical accounts. Not close by region but by time, Central Mexican historiography from approximately AD 600 until the Spanish conquest is structured with about the same amount of dates. However, due to the calendrical system, Central Mexican dates are chronologically ambiguous, since they recur every 52 years. The interrelations of dates in the different calendar systems (of Mixtec and Aztec and between Aztec and Lowland Maya), and the comparison of dates with chronological reconstructions in archaeology, and the absolute dating by way of historical astronomy, making use of permanent celestial phenomena, some day will render it possible to compile a wide-ranging historical chronological chart for the Mesoamerican core area.

Because of Mesoamerica's high civilizing and political complexity as well as its dynamics, the analysis of overarching political systems and their changes is a very important field of research. A great amount of effort has already gone into acquiring knowledge on the Post-classic times in Central Mexico, Oaxaca, and Northern-Yucatan; the Lowland Maya Classic is currently subject to politico-historical analysis (Nikolai Grube and Simon Martin). No historical events can be reconstructed for the Formative Period, in which the later Mesoamerican states emerged, since it lacks sizable written records. Yet the abundant archaeological findings of this period, including monumental sculpture and buildings, serve for assessing general theories on the emergence of the state. Such theories have made use of the respective Indian cultures, mostly Aztec an Inca, since Adolph Bandelier (1880) classified the Aztecs within the scope of an evolutionary scheme, followed over 100 years later by Stefan Breuer's (1990) more sophisticated analysis. American cultures are particularly relevant for this kind of research because it is assumed that they developed independently of Old World civilizations.

8. *Survey of Other American Regions*

With its politically tight organization and its technical achievements (metallurgy, road system) the South American cultural area of the Central Andes, which was last dominated by the Incas, can be perceived as superior to Mesoamerica. The study of historical concepts is nevertheless here rendered more difficult due to a lack of pre-colonial written records. All records of Indian history originated in early colonial times, are based on oral traditions and the system of knotted cords (*quipu*) only, and have been written mostly by Spaniards (Juan de Betanzos, 1551) or mestizos (Garcilaso de la Vega). They thus show the categories and forms of the European conception of history. This holds true also for writings that are sometimes considered as 'ancient Indian' such as the preciously illustrated *Nueva Corónica y Buen Gobierno* of *Guaman Poma de Ayala*, which is formally European in its illustrations and text, even though with mainly Indian contents. The anonymous historical drama *Ollantay* is according to its form and subject rather similar to a baroque theater play from Europe than being an expression for Indian thinking, albeit written in a stylistically pure form of *Quechua*, the language of the Inca empire, which makes it appear very Indian. The general assertion alone that something like an 'Andean Historical Concept' exists is not helpful. If more substantial information on the Indian historical concept of this cultural area is to be revealed, progress can be achieved only by concentrating on less spectacular administrative sources (*visitas*) and by diagnostic separation of European and ancient Indian components in all sources. It should also content itself with subjects of manageable size in respect to thematical scope and regional breadth, before proceeding to generalizations.

In conception and structure similar to the Mesoamerican historical tradition are the eighteenth-and nineteenth-century historical traditions of the Lenape (Algonkin language family) and the Indians of the Great Plains in North America. The Lenape (*Walam Olum*, the authenticity is controversial) and other related groups, e.g., Creek, cultivated an oral and possibly picture-written tradition of the tribe's migrations led by certain chiefs; while the Plains Indians depicted individual events from the group's daily life following the year's seasons (*Winter Count*).

See also: Central America: Sociocultural Aspects

Bibliography

Adorno R 1989 Die visuelle Vermittlung zwischen Oralität und Literalität in einer Anden-Chronik. In: Birgit Scharlau (ed.) *Bild-Wort-Schrift*. Gunter Narr Verlag, Tübingen, Germany, pp. 1–17

Barrera V, Rendon A, Rendón S 1948 *El libro de los libros de Chilam Balam*. Fondo de Cultura Económica, México [abridged in later editions]

Breuer S 1990 *Der archaische Staat. Zur Soziologie charismatischer Herrschaft*. D. Reimer Verlag, Berlin

Brinton D G (ed.) 1885 *The Annals of the Cakchiquels*. (Brinton's Library of Aboriginal American Literature, 6). Philadelphia 1885 [reprinted: AMS Press, New York]

Crónica M 2000 *Quellenwerke zur alten Geschichte Amerikas, 14*. Gebrüder. Mann Verlag, Berlin

Eschmann A 1976 *Das religiöse Geschichtsbild der Azteken*. (Indiana, Beiheft 4). Gebrüder Mann Verlag, Berlin

Ethnohistory 1972–5 In: *Handbook of Middle American Indians*. University of Texas Press, Austin, TX, Vols. 12–15

Guaman P de A F 1936 [1615] *Nueva Corónica y Buen Gobiemo*. (*Codex péruvien illustré*). (*Travaux et mémoirs de l'Institut d'Ethnologie, 23*). [reprinted: 1968] Institut d'Ethnologie, Paris

Lehmann W (ed.) 1938 *Die Geschichte der Königreiche von Colhuacan und Mexico*. (*Quellenwerke zur alten Geschichte Amerikas*, 1). W. Kohlhammer, Stuttgart [rev. edn. by Gerdt Kutscher 1974]

Marcus J 1992 *Mesoamerican Writing Systems. Propaganda, Myth, and History in Four Ancient Civilizations*. Princeton University Press, Princeton, NJ

Middendorf E W (ed.) 1890 *Ollanta. Ein Drama der Keshuasprache*. Leipzig, Germany

Prem H J 1989 *Geschichte Alt-Amerikas*. (*Oldenbourg Grundriss der Geschichte, 23*). R. Oldenbourg Verlag, Munich, Germany [English edition in preparation]

Riese B 1997 Zeitstrukturen in Mesoamerika. In: Müller K E, Rüsen J (eds.) *Historische Sinnbildung*. Rowohlt, Reinbeck bei Hamburg, pp. 240–50

Scharlau B, Münzel M *Qellqay. Mündliche Kultur und Schrifttradition bei Indianern Lateinamerikas*. Campus, Frankfurt am Main, Germany

Schultze J L (ed.) 1944 *Popol Vuh. (Quellenwerke zur alten Geschichte Amerikas, 2*). W. Kohlhammer Verlag, Stuttgart & Berlin, Germany [reprinted: 1975]

Seler E 1924–5 *Gesammelte Abhandlungen zur amerikanischen Altertums- und Sprachenkunde*. 5 Bände. Berlin [reprinted with additional Index-volume: Akademische Druck- und Verlagsanstalt, Graz 1960–7]

Taylor G 1999 *Ritos y tradiciones de Huarochiri*. (*Travaux de l'Institut Français d'Etudes Andines, 116*). Paris

Voegelin C (ed.) 1954 *Walam Olum or Red Score*. Indianapolis

B. Riese

Historiography and Historical Thought: Islamic Tradition

'Islamic historiography' is a useful but somewhat misleading term commonly applied to the historical writing produced in the predominantly Muslim societies of Asia and Africa between the rise of Islam (post-622 CE) and the mid-nineteenth century. In contrast, history written after that point may have a distinctive regional or cultural orientation, but increasingly it falls within the post-Enlightenment historiographic discourse established and dominated by Western Europe. The term is useful because it allows us to discern the degree to which the enormous historical literature produced within Muslim-dominated societies and polities embodies an unbroken (albeit complex and constantly changing) tradition of debate over the meaning and interpretation of the past. It is misleading insofar as it beguiles us into thinking that this literature was produced entirely by Muslims, or that it is exclusively grounded in Islamic doctrine and focused on religious issues. In fact non-Muslims produced important works (admittedly a small proportion of the total) within the 'Islamic' historiographic tradition. More importantly, much historical writing by Muslims focuses on decidedly nonsacred topics and presents these topics in a manner that owes little to Islamic religious concepts and concerns.

1. History as a Form of Knowledge and a Profession

Quite apart from the ambiguity of the term, Islamic historiography represents an intriguing paradox. Historical works were produced in great profusion from at least the mid-eighth century on, and it is evident from the number and distribution of manuscripts that these works were widely read. Technical terms for historical writing emerged very early and remained in constant use thereafter: *ta'rikh* (literally 'dating,' hence 'chronicle'), *akhbar* ('recitals of events'), and less commonly, *sirah* ('biography'). '*Ta'rikh*,' the most common term, referred not only to chronologically organized works but to historical writing of every kind, including alphabetically arranged biographical dictionaries. However, even though historical writing permeated medieval Islamic culture and was universally seen as a distinct form of knowledge and literary practice, it never achieved the status of a 'science' (Arabic, '*ilm*')—i.e., it never became a formal

subject of study and training in the scholarly curriculum. Within the normative classification of knowledge constructed by medieval Islamic culture, history remained always on the margins of scholarship and literature. For religious specialists it was ancillary to the core disciplines of legal studies (*fiqh*), Prophetic tradition (*hadith*), and Qur'anic exegesis (*tafsir*). Secular writers might well regard history as *magister vitae*, a trove of practical wisdom, but only rarely did they regard it as the foundation for a systematic investigation of politics and society. Even in the humble roles of handmaid and schoolmaster, however, history still generated many works of great literary and methodological sophistication. The political (and occasionally social and economic) analyses of the best Muslim historians yield nothing to their counterparts in the ancient, medieval, and Renaissance West. All this suggests that history played a much more fundamental role in defining Islamic self-understanding than that culture's spokesmen knew how to articulate. History showed Muslims how God was working his purposes out in the deeds and sufferings of the community of believers; in particular, it measured the fidelity (or lack thereof) of the believers to the revelation brought by Muhammad. History challenged Muslims to measure their progress on the path to salvation or perdition as the inevitable Day of Judgment approached.

Historians were in effect self-taught amateurs, bureaucrats, or religious scholars who composed *ta'rikh* in their spare time. They learned how to do it not through formal study with established masters, but simply by reading the works of other historians and adapting the forms and approaches of these writings to their own purposes. Some histories were written at the behest of rulers anxious that their deeds be properly memorialized for future generations, but even these do not represent a true professionalization of history. We do find official court historiographers under the Ottomans between the sixteenth and mid-nineteenth centuries, but for the most part the job was handed over on an *ad hoc* basis to any member of the ruler's entourage who was thought to possess sufficient political acumen and literary skill.

Two outstanding scholars did attempt to make history into a generally recognized formal science, but in spite of their great prestige and influence, neither achieved his goal. The Qur'anic exegete, jurist, and historian Abu Ja'far al-Tabari (838–923) tried to demonstrate in his *Chronicle of Prophets and Kings* (*Ta'rikh al-rusul wa'l-muluk*) that historical reports could be deployed according to the rigorous criteria applied to *hadith* (the dicta ascribed to the Prophet). However, very few followed him down that path, even though his reconstruction both of pre-Islamic history and of Islam's first two centuries provided an authoritative framework for almost all later chroniclers. Far from adopting al-Tabari's strict method, most of his successors simply condensed and paraphrased his work. Five centuries later, the North African Ibn Khaldun (1332–1406)—a gifted legal scholar and logician as well as a historian—argued in his famous *Muqaddimah* that history was the science of society and as such ought to be treated as a branch of philosophy. Again, his argument did not suffice to move historiography up to that exalted status, in spite of the substantial circulation and influence that his work enjoyed.

2. *The Producers of Historical Knowledge*

As we have seen in Sect. 1, history was written (and read) by highly literate men with some time on their hands. Given the social and educational structures of medieval Muslim societies, that means it was written chiefly by chancery secretaries and religious scholars (Arabic, '*ulama*'). It is tempting to describe these two groups in terms of well-defined ideal types and thus to posit opposing historical perspectives of palace and mosque. In reality these perspectives overlapped in complex and constantly changing ways.

Throughout Islamic times the '*ulama*' were given state appointments as judges or mosque officials, though it is true that many of the most prestigious scholars prized their independence and refused to accept such positions. But after the mid-eleventh century, even the independent-minded found it difficult to refuse professorships in the burgeoning religious colleges (*madrasahs*) that were endowed and to some degree controlled by the rulers and their senior military commanders. However scrupulous they might be, scholars could hardly avoid some entanglement with the princely courts. Quite apart from questions of material interest, most also had a profound fear of social disorder—as a popular dictum had it, one night of anarchy was worse than sixty years of tyranny. Thus they seldom challenged or even questioned the underlying political order of their societies, however severely they might criticize particular rulers.

On the other hand, their relations with the princely courts were always ambivalent. They knew that rulers came and went like the clouds. They might collaborate with the regime in power but rarely identified themselves closely with it. Their job was to embody Islamic beliefs and values in their own lives and to propagate these within society at large. Religious scholars came from a wide range of social backgrounds and interacted constantly with the urban population. They indeed derived much of their prestige (the only solid basis of their social influence) from their accessibility to all.

As historians, then, the religious scholars were the spokesmen for the community and its concerns. As such, they stood not so much against the princely courts as apart from them. They naturally placed the rulers and their doings at center stage in their historical writing, since the rulers were the heads of the Muslim

community within the territories they ruled—responsible not only for defense, public order, and justice, but also for ensuring that Islamic religious practices and moral norms were maintained. Without a Muslim ruler, however incompetent or corrupt, an authentic Islamic way of life was impossible. However, they did not regard rulers as the sole guardians of these matters. On the contrary, scholar–historians asserted the indispensable role of their own class by recording the lives and words of eminent men of religion in countless biographical dictionaries.

In contrast to the religious scholars, chancery secretaries were men of the regime, who shared and indeed articulated its perspectives on rulership and society. The middle and upper ranks especially tended to live in a world apart from the people they governed. Their lives and fortunes were held at the whim of an autocrat, whose purposes and moods were of far greater importance to them than the life of the larger society outside the palace walls. In spite of their narrow focus, the secretaries often took a surprisingly critical view of the regimes they served. They typically judged them in accordance with the political traditions of pre-Islamic Iran, though these were often blended with specifically Islamic values. Iranian political tradition emphasized the crucial obligation of the ruler to uphold justice within his realms. In practice that meant respect for established social hierarchies, equitable taxation, enforcing right religion, and mastery of the army. The chancery secretaries might well be sincere Muslims, but for them the ruler and his senior officials were the sole actors who really counted. Hence their historical writing was strongly focused on the court and on the struggle for power. The best historians provide searching analyses of this struggle, to be sure, but with a few notable exceptions they portray it as the affair of a few score individuals, not of broad social groups and forces.

3. *The Problem of Origins, 622–825 CE*

Islamic historiography was born in obscurity. The earliest historical texts that we now have only acquired their present form in the late eighth century. For earlier stages of development we are reduced to inference or, more accurately, speculation. We can at least be sure that Islamic historical writing emerged within Arabophone Muslim circles, for the most primitive elements in this literature are couched in archaic Arabic and reflect almost exclusively the internal concerns of the Muslim community. Islamic historiography emerged out of the triumph and trauma of the seventh century: the sudden rise of a new religion among the peoples of Arabia, the vast conquests which they made under the banner of that religion, the profound changes which religion and empire wreaked on their traditional culture and socio-political organization, and the bitter civil strife which rent a unified community of believers into hostile sects and factions. To make sense of these tremendous events, early Muslims had to devise some way of recalling, ordering, and interpreting them.

By the early eighth century they had come to believe that the Qur'an held the key to the underlying meaning of these events. The Qur'an (the collection of the revelations received by Muhammad between his call to prophethood ca. 610 and his death in 632) is far from a historical record of the earliest Muslim community. It alludes to a few crucial events which formed that community but does not narrate them. Rather, it places Islam within the long sequence of prophetic messages through which God has repeatedly called his creatures back to obedience. In the Qur'an, Muslims could see how a prophet had been sent to every people, bearing the promise of salvation for those who believed and the threat of damnation for those who refused. Most peoples rejected their prophets and suffered destruction for their obduracy. In contrast, a very few (most notably the Jews and Christians) did accept their prophets and prospered so long as they adhered to the messages these prophets had brought them. Over time, however, even these believing peoples tended to fall away, and God called them to repentance by periodic chastisement and humiliation. Through this repeated pattern the Qur'an laid down a paradigm which could explain the course of historical events. This paradigm contained three elements: the covenant offered by God to his people, the ever-present threat that this covenant might be betrayed by its recipients, and the desperate hope for redemption even after betrayal. In the circumstances of the seventh and eighth centuries, the Qur'anic paradigm of Covenant–Betrayal–Redemption was very compelling indeed.

Paradigms do not create literary genres. To recount the things that had befallen them, the early Muslims developed short prose narratives (Arabic, *akhbar*, singular, *khabar*) focused on discrete events. The earliest historical narratives about the seventh-century conquests and the internal struggles within the community of believers sprang up in a spontaneous, uncontrolled manner. They represented the contradictory memories and claims of innumerable tribal spokesmen and sectarian advocates, not a master narrative propounded by official spokesmen of the caliphal court. They were originally (and perhaps remained for some decades) oral accounts, and changed constantly as they were told and retold. Moreover, highly partisan narratives inevitably provoked counter-narratives. With a very few exceptions, it is no longer possible to recover the primitive form of any of these early narratives. That was no doubt the case even by the early 700s.

The earliest methodical effort to identify and collect the most reliable (i.e., religiously and politically acceptable) of these floating, infinitely plastic narratives took place around 700 under the aegis of the Umayyad caliphs. But the Umayyads could impose no

monopoly on the process of collection and redaction; indeed their version of Islam's first century was nearly obliterated when they were overthrown by the Abbasids in 749–50 and subjected to a systematic *damnatio memoriae*. Between 750 and 830 a group of scholars with informal ties to the Abbasid court did generate a stable corpus of historical narratives. This corpus represents what we now regard as the 'mainstream' version of early Islamic history. But competing versions produced by rivals and opponents of the Abbasids survive, albeit in fragmentary or masked form. However obscure the realities of early Islamic history may be, the historical debate of the late eighth and early ninth centuries is very well known.

The narratives collected, redacted, or (sometimes) created in the early Abbasid period were usually gathered in 'monographs' devoted to a single event (typically a battle) or category of information (e.g., lists of provincial governors). However, this period also witnessed the first major syntheses. Ibn Ishaq (d. 767) produced his famous biography of the Prophet Muhammad, the oldest complete work of Islamic historiography to reach us, and the following decades saw a host of works on the early conquests, the first caliphs, etc. Structurally, these syntheses are compilations of discrete narratives (*akhbar*), not continuous integrated narratives. On the other hand, the editors shaped their compilations in accordance with a conscious (though usually unstated) design and agenda. Mid-range syntheses of this kind dominated Islamic historiography down to the late ninth century, and they never ceased to be produced altogether.

The early Abbasid period also witnessed a plethora of efforts to develop a reliable absolute chronology, though the first essays in this direction seem to have been made under the later Umayyads. Chronographers naturally used the *hijri* calendar, traditionally instituted in 642. But of course events that had occurred before 642 had to be backdated, while the precise dates for a host of later events could only be guessed at. Nevertheless, Muslim historians quickly came to feel that precise dating was an essential element of their craft; the effort to pin things down and put them in the correct chronological order became a permanent characteristic of Islamic historical practice. By the 850s at latest, we begin to see chronicles of the kind long familiar in Byzantine historiography—viz., a year-by-year calendar of significant events (usually recounted in very summary fashion), beginning at a fixed point and coming down to the present. In the Byzantine tradition, that fixed starting point was typically the Creation, while early Muslim historians favored the Hijra (622 CE). The annalistic structure of the chronicle was a rigid but useful way of ordering events, and by the eleventh century it had become by far the most widely used format in Arabic-language historiography.

The early ninth century witnessed the appearance of one last genre of synthetic historiography, the biographical dictionary. The first major work of this class was the *Great Book of Generations* (*Kitab al-tabaqat al-kabir*) of Ibn Sa'd (d. 845). His goal was to compile concise biographies of all those persons who had transmitted Islamic religious knowledge from generation to generation, beginning with the Prophet himself. These biographies seldom provided a detailed account of the events of a person's life, let alone any effort to probe his personality. Rather, they aimed at evaluating his religious standing and his reliability as a scholar. Ibn Sa'd had clearly addressed a deeply felt need, for succeeding centuries (including the twentieth) saw a profusion of such works. In addition to countless dictionaries dedicated exclusively to religious scholars, there were dictionaries dealing with poets, men of letters, mystics, physicians, bureaucrats, and ultimately all those (including numerous women) who had earned a place within the cultural memory of Islam.

By the last quarter of the ninth century Muslim historians produced a series of grand syntheses, which claimed as their subject matter all of Islamic history, or even the history of the world since the Creation. Among many remarkable works, pride of place undoubtedly belongs to al-Tabari's massive *Chronicle of Prophets and Kings*, a work still in progress at his death in 923 (see Sect. 1). Apart from its methodological rigor, al-Tabari's compilation is distinguished by its vast scale, its meticulous attention to biblical and Iranian (though not Roman) history, and its inclusion of conflicting perspectives on the critical events of Islam's first 250 years. Al-Tabari was a somewhat controversial figure in his own time, but his authority among later generations was immense. Most of his successors were content simply to abridge and condense his account of early Islamic history, but from him they did learn to place Islam within God's plan of creation and salvation by linking it to the religions and empires which had preceded it.

4. *The Mature Tradition, 950–1800*

By the time al-Tabari died, the Covenant–Betrayal–Redemption paradigm that had shaped historical discourse throughout the eighth and ninth centuries seemed only remotely connected to the new crises confronting the Islamic world. Historians now had to grapple with the collapse of the unitary caliphate and the rise of a host of warlord-dominated regimes throughout the Islamic lands—in short, the political disintegration of the community of believers. Moreover, the religious conflicts which had arisen in Islam's first two centuries could no longer be healed; indeed, by the mid-tenth century what had been fluid if bitterly contested allegiances were transmuted into rigidly defined sectarian divisions. Political and religious fragmentation were unwelcome facts, but facts they incontrovertibly were, and they deeply shaped the historical writing of the following centuries.

First of all, the new environment provoked a change in subject matter. In striking contrast to al-Tabari and his predecessors, historians began to focus on recent and contemporary events rather than the community's early history. Likewise, they increasingly framed their work in local and regional rather than universal terms; their subject was more often the history of Isfahan or Damascus than of the Islamic world as a whole. This emphasis on recent and local events encouraged historians to adopt dynastic history as a framework for their narratives, for dynasties (and the warlords who created them) were the most visible political reality in their world. The turn to dynastic history was linked to new interpretive concerns. For historians observing the rise and fall of the warlord states, the central problem was not a ruler's role in the grand drama of salvation, but rather the reasons for the success or failure of his regime.

Within this new historiographic culture, universal history—i.e., history which claimed to include the whole Islamic world since the time of the Prophet, or even the Creation—continued to be written. But even the finest work of this genre, the *Complete History* (*al-Kamil fi al-ta'rikh*) of Ibn al-Athir (1160–1234), is strongly inflected by the distinctive concerns of its era. Within its strict annalistic framework is imbedded a quite visible sequence of dynastic histories. It covers all the lands of Islam with remarkable balance and detail, but it does so by examining each region *seriatim* as it surveys the events of each year. Finally, its primary focus is the mundane political process and the pragmatic lessons that can be inferred from that process.

Two other universal histories of great originality are equally products of their time and place. The *Comprehensive Chronicles* (*Jami' al-tawarikh*) composed under the supervision of the vizier Rashid al-Din (d. 1310) is a work which symbolizes the aspirations to universal rule of the Mongol rulers for whom it was written, for it deals with the history of every people about whom Rashid al-Din could obtain any information. Even the usually despised and neglected Franks are included within his vast tableau. It is an irreplaceable encyclopedia of the historical knowledge available to Iranian scholars in the early fourteenth century. Finally, the *Book of Examples* (*Kitab al-'Ibar*) of Ibn Khaldun (1332–1406) is distinguished not only by its famous introductory volume (*al-Muqaddimah*) but by its analytic survey of peoples and cultures and its author's grasp of the socioeconomic structures which shaped events.

Apart from these shifts in subject matter and emphasis, the mid-tenth century witnessed the bifurcation of historical writing into two linguistic branches. Arabic (the traditional language of Islamic scholarship) was now joined by Persian, the language spoken at the emerging princely courts of Iran and Muslim Central Asia. Arabic continued to be used throughout the Iranian cultural zone (which came to include Central Asia, Anatolia, and much of South Asia) for the religious sciences, but Persian rapidly supplanted it in other arenas: secular and religious poetry, government administration, didactic prose, and history. The use of Persian was more than a linguistic phenomenon, for Persian historiography quickly developed distinctive characteristics of its own—e.g., a preference for lengthy continuous narratives, an explicit appeal to the political precepts and institutions of ancient Iran, a certain vagueness or carelessness about chronology. By the eleventh century Arabic and Persian historiography clearly embodied two distinct traditions, each developing in its own way. This picture was further complicated by the emergence of Turkish as a vehicle of literary expression in the late fifteenth century. From that point down to the mid-nineteenth century Turkish-language works were composed throughout the Middle East—no surprise, since the political and military elites from the Nile to northwest India were overwhelmingly of Turkish origin during that period. However, the main centers of Turkish literature were located in predominantly Turkish-speaking areas—i.e., Uzbek-dominated Central Asia and the core Ottoman lands in Anatolia and the eastern Balkans. In spite of some specific characteristics of its own and many works of outstanding quality, Turkish historiography tended to model itself upon the Persian tradition, and in some regions the two languages coexisted on almost an equal basis as vehicles of literary expression.

5. *The Transition from Islamic to European Models, 1800–2000*

By 1800 the Muslim peoples of Asia and North Africa faced severe military and economic challenges from Europe. However, most Muslim polities (with the important exceptions of the Mogul Empire in India and the sultanates of Java) still retained their political independence, and the deep-rooted societal and cultural structures of the Islamic lands seemed intact. By 1900, however, almost every Muslim state had been swallowed up or made a client regime, and Muslim societies were enmeshed in massive, uncontrollable transformation on every level.

Change on such a scale inevitably compelled the creation of a radically new historiography, though the process was slow and complex. Works of great merit continued to be produced within the existing traditions throughout the first half of the nineteenth century. Indeed, excellent historical scholarship of a decidedly traditional cast was being written well into the twentieth century. Nineteenth-century historians were fully cognizant of the new world they inhabited, but for many decades traditional historiographic forms and approaches seemed adequate to the task. By the 1870s that was no longer the case. First, existing models of historical writing simply did not possess the analytic

tools that were needed to address the profound cultural and ideological problems created by the European challenge. For this purpose a new historiography was essential. No less important, by this time many Muslim intellectuals had studied in Europe or with European teachers, and they felt the powerful allure of Enlightenment and post-Enlightenment thought. To them it was self-evident that European military and economic power was grounded in the power of European ideas. Just as Muslim armies needed to adopt European weapons and tactics, so Muslim intellectuals had to appropriate European concepts and methods of inquiry.

European ideas and literary models were easier to perceive than to internalize, and the later nineteenth century witnessed a difficult struggle by Muslim intellectuals and scholars to develop new ways of writing history. Most late nineteenth-century historiography is thus a hybrid product. The point is illustrated by two major works, both marked by broad scope, critical use of documents, and awareness of European scholarship. Ahmet Cevdet Pasa's history of the Ottoman Empire (12 Vols., 1855–84) adheres to the traditional chronicle form, while 'Ali Mubarak's historical geography of Egypt (20 Vols., 1886–88) is modeled on a classic work by the fifteenth-century Cairene scholar al-Maqrizi (d. 1442). A model for historiography in a modern (i.e., European) style was provided by the *History of Islamic Civilization* (*Ta'rikh al-tamaddun al-islami*, 5 Vols., 1902–6) of Jurji Zaydan, a Syrian Christian novelist and man of letters. For both political and cultural reasons, however, Zaydan had no immediate imitators, and only after World War I did Middle Eastern historiography begin to internalize contemporary European modes of research and presentation. This shift was first visible in works dealing with modern history, but by the end of the 1930s the same trend had taken hold for studies of early and medieval Islamic times. The 'Europeanization' of historiography affected not only research methods and literary structure, but also the kinds of questions that Muslim scholars asked about their past. Even where (for a wide range of scholarly, political, and religious reasons) they rejected the specific analyses and interpretations of Middle Eastern and Islamic history asserted by European historians, they could not ignore the problems to which these scholars had called attention.

This transformation coincided with the professionalization of history after World War I—a sharp departure from traditional practice in the Islamic world. This process began with the establishment of European-style universities in Cairo and Istanbul just before the war, along with the creation of officially sponsored societies for historical research. Simultaneously, history was enlisted in the struggles for national independence and cultural autonomy of the interwar and postwar periods. Finally, by World War I a few Muslim scholars were taking advanced degrees from the leading universities in Europe, and their numbers increased greatly in the following decades. The prestige and authority that these foreign-trained scholars enjoyed when they returned home in effect imposed a new model of historical thought and writing on their students.

Cross-cultural influence in historiography has not flowed in a single direction, however. Even in the 1920s and 1930s, a few Muslim scholars published their work in Western languages, and since the 1950s this trend has grown enormously. A growing number of scholars from Islamic Asia and Africa have not only taken their advanced degrees in Europe and North America, but have also been appointed to many of the most prestigious teaching posts there. This situation has begun to create a real dialogue among a plurality of theoretical and cultural perspectives. The history of the Islamic world is increasingly in the hands of its own scholars—the natural state of things, perhaps, but one which was hardly the case in the nineteenth and twentieth centuries.

See also: Islam: Asia; Islam: Middle East; Islam: Sub-Saharan Africa; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Religion; Near Middle East/North African Studies: Society and History

Bibliography

Crabbs J A 1984 *The Writing of History in Nineteenth-century Egypt: A Study in National Transformation*. Wayne State University Press, Detroit, MI

Delanoue G 1982 *Moralistes et politiques musulmans dans l'Egypte du XIXe siècle (1798–1882)*. Institut Français d'Archeologie Orientale, Cairo, Egypt

Donner F M 1998 *Narratives of Islamic Origins: The Beginnings of Islamic Historical Writing*. Darwin Press, Princeton, NJ

Ende W 1977 *Arabische Nation und islamische Geschichte: die Umayyaden im Urteil arabischer Autoren des 20. Jahrhundert*. Orient-Institut der Deutschen Morgenländischen Gesellschaft, Beirut, Lebanon

Encyclopaedia of Islam, New Edition 1960 – in progress Entries: *ta'rikh* and individual historians. E J Brill, Leiden, Netherlands

Fleischer C 1986 *Bureaucrat and Intellectual in the Ottoman Empire: A Biographical Study of the Historian Mustafa Ali, 1541–1600*. Princeton University Press, Princeton, NJ

Hourani A H 1962 *Arabic Thought in the Liberal Age, 1798–1939*. Oxford University Press, London

Hoyland R G 1997 *Seeing Islam as Others Saw It: A Survey and Evaluation of Christian, Jewish and Zoroastrian Writings on Early Islam*. Darwin Press, Princeton, NJ

Humphreys R S 1991 *Islamic History: A Framework For Inquiry*, revised edn. Princeton University Press, Princeton, NJ

Khalidi T 1994 *Arabic Historical Thought in the Classical Period*. Cambridge University Press, Cambridge, UK

Lewis B, Holt P M 1962 *Historians of the Middle East*. Oxford University Press, London

Meisami J S 1999 *Persian Historiography to the End of the Twelfth Century*. Edinburgh University Press, Edinburgh, UK

Mustafa, Shakir 1983–93 *Al-Ta'rikh al-'arabi wa'l-mu'arrikhun*, 3rd edn., 4 Vols. [Arabic Historiography and Historians] Dar al-'Ilm li'l-Malayin, Beirut, Lebanon
Noth A 1994 *The Early Arabic Historical Tradition: a Source-Critical Study*, 2nd edn. Darwin Press, Princeton, NJ
Radtke B 1992 *Weltgeschichte und Weltbeschreibung im mittelalterlichen Islam*. [World History and the Description of the World in Medieval Islam] Orient-Institut der Deutschen Morgenländischen Gesellschaft, Beirut, Lebanon
Rosenthal F 1968 *A History of Muslim Historiography*, 2nd edn. E J Brill, Leiden, Netherlands
Wellhausen J 1899 *Prolegomena zur ältesten Geschichte des Islams*. [Prolegomena to the History of Early Islam] In: Wellhausen J, *Skizzen und Vorarbeiten* Georg Reimer, Berlin Vol. 6

R. S. Humphreys

Historiography and Historical Thought: Jewish Tradition

1. Biblical Traditions of History Writing

Together with the historical traditions of ancient Greece the Hebrew Bible contains the oldest corpus of historical writings. While not all of the various books which entered the Jewish biblical canon reveal a strong sense of history, quite a few—such as Samuel, Kings, and Chronicles—are primarily historical accounts. In contrast to the Greek historians, the authors and editors of the Bible were interested not in recording history but in giving a theological sense to history. The numerous biblical passages which command the Jews to remember (*zakhor*) do not imply a special interest in history itself but express a definition of the Jews as a historical people held together first and foremost by its will to observe its historically revealed laws and to remember its common history. God is described in terms of His historical deeds. He is the 'God of the fathers' and 'Your God who brought you out of the Land of Egypt.'

With the canonization of the Hebrew Bible in the second century CE (Christian Era) Jewish historical writing undertakes a clear shift. The first century CE produces the last major Jewish historical writings for about 15 centuries. In a historical work, of which only two out of five parts survived, Philo of Alexandria describes the persecution of Egyptian Jews by the Roman governor Flaccus and the delegation of Alexandrian Jews (including himself) to the emperor Caligula. The most important historian in this period is Flavius Josephus, whose account of the Jewish War against Rome in 66–73 CE was written a few years later in Roman exile and—together with his *Antiquities*—stands in the tradition of Roman rather than Jewish historiography.

2. Historical Thought in Rabbinic Literature and Medieval Sources

There is hardly a more telling symptom for the decline of Jewish historical writing than the fact that Josephus' histories fell into oblivion among the Jews during the Talmudic period. The same holds true for many historical accounts which did not enter the Jewish Biblical canon, such as the Books of the Maccabees, which describe the rise of the Hasmoneans against Hellenism. Preserved by non-Jews, they were rediscovered by Jews only in the sixteenth century. During the Middle Ages the history of the Second Temple period—beyond the Biblical accounts themselves—was known among Jews in rather distorted versions of the original accounts, such as the *Book of Antiochus* and the tenth-century *Book of Yossipon* (believed to be Josephus' original account), both produced in southern Italy.

The Talmud, which was to become the single most important source for Rabbinic Judaism over the centuries, is unhistorical in its nature. Although this vast corpus contains quite a few passages of historical value, its editors reveal no special interest in the chronology of events. Thus, rabbis of different centuries are made to engage in conversations, and biblical events are often presented out of their chronological order.

Medieval Jewish historical thought is framed by the memory of the biblical past on the one hand and by the longing for a messianic future on the other hand. The time span between a golden past and an ideal future is usually interpreted with clear reference to the biblical course of events. Biblical place names like Ashkenaz and Sefarad designate European countries, Christianity is Edom, and Islam is Yishmael. Vicious rulers are called Haman, like the archetypical biblical villain from the Book of Esther, and saviors are termed Mordechai after the hero of the same story. Messianic expectations led to concrete calculations about the End of the Days, and contemporary tragedies were often seen as the birthpangs of the Messianic Age.

Altogether, traditional Jewish historical thought as expressed in Rabbinic literature and medieval sources reveals at least a partly cyclical understanding of time. Every year when celebrating Passover observant Jews remember the exodus from Egypt, and during three annual days of fasting they mourn the destruction of Jerusalem. Each rescue from persecution is regarded a new Purim (in commemoration of the Book of Esther). Every Sabbath morning a weekly portion of the Pentateuch is read in the synagogue, thus reviving sequentially each year the historical accounts of the Five Books of Moses.

The most prominent genre of historical writing during the Middle Ages was the chronologies, which often tried to establish a reliable 'chain of tradition' (*shalshelet hakabbala*) from Moses through the leading Rabbis of the Mishna and Talmud and up to the

authors' own time and place. They constituted a refutation of the anti-Talmudic challenge made by the Karaite sect and at the same time underlined the importance of the local Jewish community in which they were produced. Examples of this genre are the *Letter of R. Sherira b. Hanina* from the academy at Pumbedita in Babylonia, who in the late tenth century responded to scholars from Kairouan, and the twelfth-century *Sefer Hakabbala* (Book of Tradition) of the Spanish scholar Abraham ibn Daud. A stronger emphasis on contemporary events can be seen in the detailed accounts of the Crusade massacres concerning the cruelties which befell the Jewish communities of the Rhineland in the eleventh and twelfth centuries.

Among the most popular sources reflecting medieval and early modern Jewish historical thought in Central Europe were Memor Books, which contain collections of prayers together with a necrology of distinguished persons, and recorded local martyrs and persecutions. The oldest known Memor Book is from Mainz and dates back to 1296. A few hundred Memor Books are extant, mainly from seventeenth-century rural communities in southern Germany.

3. *Sixteenth-century Jewish Historiography as a Response to Catastrophe*

The sixteenth century saw a sudden upsurge of Jewish historical writing with the publication of 10 important historical works, more than all of those known to us from the Middle Ages. Quantitatively as well as qualitatively the development was significant, with a new interest in history itself, as Yosef Hayim Yerushalmi demonstrated in his seminal study *Zakhor* (1989). Writers like Yosef ha-Kohen, Abraham Zacuto, and Samuel Usque wrote in the wake of the expulsion of Jews from Spain and Portugal (1492 and 1497), which affected not only their own lives but also completed the wave of expulsions of Jews from Western Europe. The new search for explanation is apparent in their writings, perhaps most pointedly in Solomon ibn Verga's *Shevet Yehuda* (The Staff of Judah, 1554), but also in the historical oeuvre of the only Ashkenazic author of this group, David Gans from Prague. His *Tsemah David* (Offspring of David, 1592) contains the most precise account of Jewish and world history undertaken by a Jewish author until that time.

The Italian Jew, Azaria de' Rossi, who completed his *Me'or Enayim* (A Light to the Eyes, 1573–5) at about the same time, has the merit of recovering ancient sources that had been lost to Jewish tradition because they did not enter the biblical canon. Among them are works by Philo of Alexandria and the apocryphal *Letter of Aristeas*. For the first time the historical interest of Jewish writers did not concentrate on the biblical past or speculations about the messianic future but dealt with the concrete history of the Jews during recent centuries and in their own time. These authors also showed a genuine interest in the histories of other nations. The impact of their work on later Jewish historiography was, however, minimal. It was only as a result of the encounter with academic scholarship in the early nineteenth century that modern Jewish history writing developed.

4. *The Beginnings of Modern Jewish Scholarship and Critical Historiography in the Nineteenth Century*

The beginnings of modern Jewish scholarship can be dated to the small circle of Jewish university students from Berlin who founded the *Verein für Cultur und Wissenschaft der Juden* in 1819. Leopold Zunz, the leading spirit of this group and the dominant figure in the field which would become known as *Wissenschaft des Judentums*, published in 1818 a systematic agenda for modern Jewish research under the title '*Etwas ueber die rabbinische Literatur*' (On Rabbinic Literature). The study of history, which had been marginal in the works of eighteenth-century Jewish enlighteners (*maskilim*), now became a central pillar of Jewish scholarly enterprise. From the beginning, a certain apologetic character was inherent in many attempts at reclaiming the Jewish past. As Immanuel Wolf, a founding member of the *Verein* expressed it: 'Scientific knowledge of Judaism must decide on the merits and demerits of the Jews, their fitness or unfitness to be given the same status and respect as other citizens.'

In a more direct way, the scholarly studies of Leopold Zunz served as instruments to achieve emancipation for German Jews. Like Wolf, Zunz believed that 'the equality of the Jews in customs and life will follow from the equality of *Wissenschaft des Judentums*'. In his *Gottesdienstliche Vorträge der Juden, historisch entwickelt* (1832), he documented the long tradition of Jewish sermonic literature, thus rebutting the Prussian government's dismissal of German synagogue sermons as innovations in Jewish liturgy. Five years later, in his study on Jewish surnames, *Die Namen der Juden*, Zunz proved the ancient origin of many 'non-Jewish' names among Jews and thus reverted to a decree which forbade the use of Christian first names among Jews.

It would, however, not do justice to the enormous enterprise of nineteenth-century *Wissenschaft des Judentums* to view it predominantly under the aspect of apologetics, as did later critics, especially Zionists, most prominently Gershom Scholem. Another important motif for the pioneers of Jewish historiography was to retrieve the lost treasures of the past for a rapidly assimilating Jewish community as a means of strengthening their particular identity as modern Jews. As Yosef Hayim Yerushalmi writes, 'History becomes what it had never been before—the faith of fallen Jews' (1989, p. 86). Finally, the purely scholarly

interest in the Jewish past at a time in which history writing and reading became one of the most popular intellectual endeavors should not be disregarded.

There was no bridge between those first scholarly accounts of Jewish history and earlier Jewish writing. It was through the non-Jewish world that historiography entered Jewish writing. Consequently, it was not a Jew, but a Protestant French author in Dutch exile, Jacques Basnage, who wrote the first multivolume postbiblical Jewish history. He characterized his *Histoire du peuple Juif depuis Jésus Christ jusqu'à présent*…(1706–11, rev. edn., 1716–21) in his subtitle as a continuation of Josephus, and there was indeed some truth to this statement. To be sure, his account would in no way measure up to modern scholarly standards, and was still full of missionary intentions, but it was nevertheless the first serious attempt to write a comprehensive postbiblical Jewish history.

5. *Political Agendas and the Writing of Jewish History*

The first Jewish writer to embark on such an endeavor was Isaak Markus Jost, a member of the *Verein für Cultur und Wissenschaft der Juden* and one of the pioneers of *Wissenschaft des Judentums*. The modern secular education he received as a student at Berlin and Goettingen is visible in his attempt to present a sober and objective view of Jewish history in his nine-volume *Geschichte der Israeliten seit der Zeit der Maccabäer bis auf unsere Tage* (1820–28). Jost's work is clearly centered on the German-Jewish experience and draws a picture of Judaism as a religion which has lost all its original national characteristics. Depicting modern Jews as a purely religious community was part of the larger German-Jewish struggle for Emancipation. The title of his later work, the three-volume *Geschichte des Judentums und seiner Sekten* reveals the urge to present Judaism as a religion analogous to Christianity.

When Heinrich Graetz wrote his 11-volume *Geschichte der Juden von den ältesten Zeiten bis auf die Gegenwart*—the most popular such history ever written—one generation later (1853–76), he deviated from the early works of *Wissenschaft des Judentum* representatives in his stronger emphasis on the existence of a Jewish people and by a more passionate presentation. Although he was later claimed as a proto-Zionist writer, it is important to note that his definition of the Jewish nation is that of a community held together by an essentially religious idea. Like Jost, for whom modern Jewish history begins with the enlightened Prussian monarch Frederick the Great, Graetz located the beginnings of Jewish modernity in Germany, with the appearance of the philosopher Moses Mendelssohn and the Berlin Jewish Enlightenment.

It was not until the late nineteenth century that a clear break from the earlier Germanocentric view of Jewish history and from the strong emphasis upon a 'history of suffering and scholarship' (Graetz) appeared in the work of the Russian-Jewish historian Simon Dubnow. His 10-volume *World History of the Jewish People* was written in Russian but published first in German translation (1925–29). Motivated partly by his own political agenda—he was the founder of the autonomist Jewish movement which represented Jewish diaspora nationalism in eastern Europe—his version of Jewish history centered on institutions of Jewish life in the diaspora, most notably the *kehilla*, the semi-autonomous Jewish community.

Dubnow was also among the founders of the YIVO Institute for Jewish Research—an institution established in Berlin and Vilnius in 1925 to systematically research the Jewish past and present in Eastern Europe and other Ashkenazic communities. In contrast to the traditional German dominance of nineteenth-century *Wissenschaft des Judentums* and to Hebrew-language Zionist scholarship, the YIVO deliberately presented its research in Yiddish, the language spoken by the Jewish masses of Eastern Europe.

From the end of the nineteenth century, Jewish historiography was increasingly influenced by new subdisciplines of *Wissenschaft des Judentums*, such as ethnography, sociology, and demography. The *Gesellschaft fuer Juedische Volkskunde*, a society of Jewish folkloristics founded in 1898 and led by the Hamburg rabbi Max Grunwald, played a pioneering role in the scholarly research of Jewish folk traditions around the globe. Its journal, the *Mitteilungen fuer Juedische Volkskunde*, published the most important findings in the field for almost three deacades (1898–1929). In Russia, an expedition to the traditional communities of the Ukraine led by the playwright and folklorist An-Ski from 1912 to 1915 was the culmination of a long search for the remnants of rural Jewish life.

Jewish demographers created their own institutional framework when they established an Office for Statistics among the Jews in Berlin, which from 1904 published its own journal and was closely related to the burgeoning interest in Jewish sociology best expressed in the pioneering works of Arthur Ruppin. Most of these endeavors were clearly related to the 'Jewish Renaissance' propounded by the emerging Zionist movement. Zionist scholars reacted against what they alleged to be the orientation of nineteenth-century Jewish scholars exclusively to the Jewish past and broadened their interest to include contemporary issues within their research.

After the establishment of the Hebrew University in 1925, scholars with a Zionist outlook formed a vaguely connected group of Jewish historians, sometimes referred to as the 'Jerusalem School,' emphasizing the centrality of Palestine in the course of Jewish history. Their most outspoken representative was the later

Israeli minister of education Benzion Dinur (Dünaburg), who in 1935 helped to establish the most important Hebrew-language historical journal, *Zion* (an earlier series of the journal had been aborted), together with the first professor of Jewish history at the Hebrew University in Jerusalem, German-born Yitshak (Fritz) Baer.

From the very early works of Zionist historiography, usually written by Zionist activists who were not professional historians (such as Nahum Sokolov and Adolf Böhm), a teleological view of history leading to the return to the Jewish homeland is clearly visible. At the same time, the Jewish past in the Land of Israel is placed in the center of interest at the expense of diaspora history. In contrast to most historians, Benzion Dinur thus commences the period of exile not with the destruction of the Second Temple in the first century CE but much later, with the Muslim conquest of Jerusalem in the seventh century CE. In a similar vein, he concludes the period of diaspora domination as early as 1700, when a group of pious Jews from Eastern Europe made their way to Palestine. In his *History of Zionism, 1600–1918* (1919) Nahum Sokolov anachronistically dates the Zionist movement back to the seventeenth century.

While Dubnow's conception of Jewish history represented a corporate view of Jewish diaspora life and the Jerusalem School emphasized the centrality of Palestine, the last historian who attempted a single-handed multi-volume universal Jewish history, Salo Wittmayer Baron, stood for a more positive approach to individuals' diaspora experience. His 16-volume *A Social and Religious History of the Jews* (1952–76) reaches only the year 1650. His more affirmative view of the diaspora was certainly influenced by his own experience of spending most of his life in the United States, where he was appointed in 1930 as the first professor of Jewish history at a Western university, Columbia University in New York.

Jewish historiography has thus to a large extent been determined by the political context, the geographical situation, and the ideological and religious outlook of the respective historians. Since all these factors differed considerably in the Jewish world, the outcome was a highly diverse array of Jewish historical writing – even though, at least up until the second half of the twentieth century, most of those who undertook to write Jewish history were themselves Jews. The destruction of European Jewry, the establishment of the State of Israel, and the flourishing of Jewish studies in the United States led to an entirely new situation for Jewish historiography after World War II.

6. *Twentieth-century Trends*

With the exception of Baron no single historian embarked on a monumental Jewish history project in the tradition of Jost, Graetz, and Dubnow in the course of the twentieth century. There were a few attempts to publish Jewish world histories as collective enterprises, as early as the ambitious plan developed by Ismar Elbogen of the Berlin *Hochschule für die Wissenschaft des Judentums* in the late 1920s; these, however, fell victim to the economic and political crisis of the early 1930s. More successful, but for financial reasons also never completed, was a multi-volume project initiated by scholars of the Hebrew University.

After World War II Jewish history became more and more a field for specialists who would concentrate on limited areas and periods. This tendency toward specialization was also the result of an increasing professionalization and acceptance of the subject as a distinct field to be taught at university level, with new centers emerging in Israel and the United States. Until the 1960s, the teaching of Jewish history in the US had been restricted to a handful of universities and a few institutions with a strong emphasis on Jewish education, such as rabbinical seminaries and Jewish teachers' colleges or the Jewish-sponsored Brandeis University. With the establishment of new Israeli universities the singular role of the Hebrew University in Israel diminished, while a true explosion in the field of Jewish studies enabled a similar diversification at American universities beginning in the 1970s. European institutes of higher learning followed to a more modest degree in the 1980s and 1990s. By 2000, Jewish history had become a highly specialized and diversified field represented at almost every major institution of higher learning in the United States and many European countries.

This development led to a more complex picture of Jewish historiography. The personalities and political programs of individual historians and the formations of more or less clearly defined schools no longer dominate the professional scene. Zionist historians who had fought an ideological battle over Jewish history, mostly against their non-Zionist Jewish colleagues, up to the 1930s were essentially left without opponents as a result of the Holocaust and the destruction of alternative Jewish options, such as diaspora nationalism and Bundist socialism. In the State of Israel, apologetical historiography (now mainly directed against Arab and British versions of the history of Palestine) did not entirely disappear but was more and more balanced by the emergence of a younger generation of historians, by whom the existence of the Jewish state was already taken for granted. This resulted in a gradual break with long-accepted notions rather than an intellectual revolution as more and more historians and sociologists of a new postwar generation attacked the older schools of Zionist historiography in the 1980s and 1990s.

While mainly questioning issues of Israeli history or prehistory, such as the role of the Zionist leadership in Palestine with respect to attempts at the rescue of European Jewry in World War II, and the accepted notions surrounding the first Arab–Israeli War in 1948–49, the views of the so-called 'New Historians'

also challenge more general Zionist notions of Jewish history. In a highly politicized context the two sides of the 'Israeli Historians' Debate' stand not only for different views on Jewish or Israeli history but mark also different paths for Israel's political future. In contrast to Israeli historiography, which has been dominated by a classical left-liberal Zionist intellectual establishment, the 'new' historians and sociologists often represent a post-Zionist stance and as such constitute a major challenge to the master narrative of Jewish and Zionist history.

In the last decade of the twentieth century, Jewish historiography moved from its early Germanocentric and later Eurocentric emphases to a stronger integration of widely neglected communities. In particular, the histories of Jews in the Muslim world have slowly been integrated into the general picture. This is partly the result of the emergence, mainly in Israel, of historians of oriental background, but also of the establishment of a few academic positions concentrating on Sephardic Jewry at American universities. It was in the United States, where the history of Jewish women emerged as a subfield of gender studies, with significant publications in the 1990s about Jewish women in Imperial Germany and in the United States, and a historical encyclopedia of Jewish women. New trends among general historians, including postmodern ideologies, have also left their imprint on Jewish historiography.

Europe, too, witnessed a revival of Jewish historical scholarship in the 1990s. The larger Jewish communities of Western Europe, France and Britain, emerged as new centers of Jewish scholarship, although most of the historical studies there deal with their respective national Jewish histories. Central and East European institutions relating to the Jewish past were destroyed by the Nazi onslaught, together with those who participated in them. Ironically, the Nazis entered into much more systematic Jewish historical research than non-Jewish scholars had done in the Imperial and Weimar periods; they established research institutes for Jewish history in Munich, Frankfurt, and Berlin. Studies undertaken there, however, were preoccupied not with Jewish history *per se*, but with the so-called 'Jewish Question' and sought to justify Nazi discrimination and persecution against the Jews from a pseudo-scholarly perspective.

It is interesting to note, though, that some of the research begun there—especially on the Court Jews and on Sephardic Jews in Hamburg—was published with only slight alterations in the 1950s. More modern Jewish historical research in Germany developed as a result of the establishment of the first Jewish Studies institutes and of the emergence, starting in the 1960s, of a new interest in German–Jewish history among modern German historians. While Jewish history is mainly written by Jewish historians and for a Jewish public both in prewar Europe and in today's Israel and America, the postwar European situation is different. With only small Jewish communities, research is mainly undertaken by non-Jewish scholars, and mainly read by non-Jews.

7. *Major Issues in Modern Jewish Historiography*

Perhaps even more than scholars of other particular histories Jewish historians have been divided by a set of fundamental questions. First of all, there is the question of whether Jewish history exists at all or whether the history of Jews merely constitutes a chapter within the histories of the nations among whom they lived. If Jewish history exists in a broader sense, the question arises whether it is the history of a people or a purely religious history—or whether at some point the national history turned into a religious history. Finally, an additional problem of Jewish history is its theological interpretation by some modern Jewish and non-Jewish thinkers, which makes Jewish history appear as though it escapes the regular laws of history. This last argument is strengthened by the conspicuous role of antisemitism throughout the centuries and the claim of the uniqueness of the Holocaust which brought about the near destruction of European Jewry.

The doubt whether Jewish history exists *per se* seems rather superfluous in an age of expanding programs in Jewish history and within the framework of an article discussing the rich tradition of Jewish history writing. Nevertheless, one can still find this opinion among some modern historians. It is based partly on the Christian theological notion that active Jewish history came to an end with the early stages of Christianity, and partly on the secular claim that Jews lack a clearly defined political history based on a common territory. In the nineteenth century such a belief was inherent in the theological characterization of ancient Judaism as 'late Judaism,' while on the other side of the spectrum of historical outlook Marxist historians denied a separate course of Jewish history on the basis that Jews not only lacked a territory of their own but also a division into classes of society in the modern period. Nowadays, the questioning of the existence of a Jewish history throughout the ages is less explicit. More indirectly, however, such a view can be detected in the almost exclusive interest of many historians in the histories of the Jews in their respective societies. Thus, German Jewry or French Jewry are often seen as isolated entities, as part of German or French history, but without any connection to the developments of other Jewish communities. On the other hand, there has been a contrary tendency among some Zionist historians to neglect the respective contexts of Jewish histories and view them only through a national Jewish perspective.

If we accept the notion that Jewish history exists as such and is stretched over numerous centuries and

over five different continents, the question remains whether it is primarily a national or a religious history. The pioneers of Jewish historiography in nineteenth-century Germany often preferred to depict Jewish history in terms of the development of a religious idea (the purest form of monotheism) and its bearers. This concept was in line with the demands of Emancipation that Jews integrate into European societies as citizens who only differed in their religious creed. Thus, the notion of a Jewish people was characterized as an ancient idea long left behind. With diaspora nationalists like Simon Dubnow and the emerging Zionist school the national element was restored to Jewish historiography. Both schools emphasized that Jews never ceased to constitute a nation even if they lacked their own territory. Dubnow went so far to stress that a nation which had overcome the need for its own territory constituted the highest and most sophisticated form of nationalism. Zionists, on the other hand, emphasized the unbroken longing for return to the Land of Israel and the use of the Hebrew language over the centuries.

A final peculiarity of Jewish history is the theological weight it enjoyed especially in a Christian context. In medieval and early modern Europe Jews often constituted the only non-Christian minority. The traditional position Jews enjoyed in Christian Europe was a mixture of protection and discrimination. Jews served as the classical 'other,' being protected because they bore testimony to Jesus Christ while at the same time discriminated against as Christ-killers and as unbelievers. On the basis of this peculiar role, many Christian authors of postbiblical Jewish history regarded this particular history as outside the course of 'regular' history. They claimed accordingly that it should be judged by theological rather than historical criteria. On the Jewish side, the most prominent proponent of a Jewish '*Sonderweg*' theory, albeit with a different background, was the philosopher Franz Rosenzweig, a student of Friedrich Meinecke, who stressed in his main work *The Star of Redemption* (1921) that general rules of history do not apply to Jewish history. More recently, the discussion of the uniqueness of the Holocaust has brought the question of a peculiar course of Jewish history to a different level. In this view, it is not Jewish history that is incomparable to other histories, but rather the persecution of the Jews which culminated in the attempt of their total annihilation.

In his programmatic and provocative essay, 'Ghetto and emancipation' (1928), the historian Salo Baron fiercely objected to what he called a 'lachrymose version of Jewish history.' At the beginning of the twenty-first century, and half a century after the greatest tragedy of the Jewish people, numerous voices once again express their reservations against a one-sided picture of Jewish history reduced to a history of persecution. While they stress the importance of research on antisemitism and the Holocaust, they also warn not to read Jewish history backwards starting from its most tragic chapter.

See also: Historiography and Historical Thought: Christian Tradition; Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: Islamic Tradition; Jewish Diaspora; Jewish Law; Judaism; Judaism and Gender; Religion, History of

Bibliography

Funkenstein A 1993 *Perceptions of Jewish History*. Berkeley University Press of California, Berkeley, CA

Liberles R 1995 *Salo Wittmayer Baron: Architect of Jewish History*. New York University Press, New York

Meyer M A 1974 *Ideas of Jewish History*. Behrman House, New York

Myers D 1995 *Reinventing the Jewish Past: European Jewish Intellectuals and the Zionist Return to History*. Oxford University Press, New York

Myers D N, Ruderman D B (eds.) 1998 *The Jewish Past Revisited, Reflections on Modern Jewish Historians*. Yale University Press, New Haven, CT

Schorsch I 1994 *From Text to Context: The Turn to History in Modern Judaism*. University Press of New England, Hanover, NH

Yerushalmi Y H 1989 *Zakhor: Jewish History and Jewish Memory*, 2nd edn. Schocken, New York

M. Brenner

Historiography and Historical Thought: Modern History (Since the Eighteenth Century)

1. Reorientation of Historical Thought in the Eighteenth Century: The Emergence of a Secular Outlook

There is good reason for beginning this article on Western historical thought and historical writing in the modern age with the eighteenth century. The Western tradition is much older, going back at least to the Hellenic Age. There is much continuity between the older traditions, particularly those of the humanistic historiography of the Renaissance and historical writing in the eighteenth century. Yet there are also profound changes in the conceptions of history and of historical knowledge and in the institutional context in which history is written that mark a sharp break with previous periods.

Henri Berr and Lucien Febvre in the article 'History' in the first edition of the *Encyclopedia of the Social*

Sciences in 1932 identified two characteristics of the modern conception of history as basic to the modern historical outlook and practice: the commitment to establishing the 'truth' and to placing historical accounts within 'universal history.' A third is the transformation of historical studies into a professional discipline. The older ancient and humanistic historiography, they argued, saw history as a form of rhetoric in the service of politics and morality and as an art form. 'History as it is conceived today,' they noted, 'may blossom into art, may be crowned with philosophy; but it is primarily and necessarily the solid establishment of facts' (Berr, p. 358), however difficult this may be.

This formulation presupposes a sharp distinction between history as literature and as a scientific discipline, even if its approach to science takes into account the elements of contingency and particularity that distinguish history from the generalizing natural sciences; but it assumes that historical accounts based on a solid factual basis narrate a coherent development that reflects historical reality.

There is a good deal to be said for this description of the transformation which historical studies underwent in the eighteenth and particularly in the nineteenth century; but especially from the perspective of the late 1990s it seems one-sided and incomplete.

On two levels, historical thought and historiographical practice in the eighteenth century moved toward greater historical realism. On the one hand, an increasingly secular outlook insisted on the exclusion of the supernatural and held that historical explanation must not violate a common understanding of natural reality. This secular approach was relatively new in the eighteenth century. On the other hand, it coexisted often with a residual religiosity, as it still did even in the nineteenth century with a conception of a hidden divine plan, as in Leopold von Ranke's trust in the 'hand of God' (1973, p. 138), or in a less hidden form in G. W. F. Hegel's dialectic. Thus Enlightenment historians such as J. C. Gatterer and A. L. Schloezer did not challenge the biblical chronology or the biblical account, although for the post biblical period, or for the non-Hebraic ancient peoples, they insisted on an approach in accord with natural reality.

Apart from the great works of synthesis in the eighteenth century, David Hume, Montesquieu, Voltaire, Edward Gibbon, G. T. Raynal, and Adam Ferguson, there exists a tradition of erudition that introduced standards of source criticism, which went back to the humanist scholarship of the Renaissance and Reformation and to the erudition of seventeenth-century historical scholarship. The belief that great philosophic works of synthesis are not enough and that scholarly history cannot be written until the authenticity and veracity of the sources on which it rests have been established expressed itself in the great methodical collection of sources begun in the seventeenth century with the Jesuit *Acta Sanctorum*, published after 1643, to provide a firm basis for ecclesiastical history, and with Ludovico Muratori's efforts in the *Scriptores rerum italicarum* (1723–51) to lay the foundation for an Italian national history. The German scholars at the universities, first J. L. Mosheim and J. S. Semler at Halle and then J. A. Ernesti and J. O. Michaelis at Göttingen, established a tradition of historical exegesis that laid the foundations for historical hermeneutics and the new discipline of historical philology which emerged in the latter part of the eighteenth century and became the key elements of the critical method of B. G. Niebuhr and Ranke in the nineteenth century.

A key concept of eighteenth-century thought was the linear character of history. History was no longer understood as a cumulation of events but as a continuous story possessing inner coherence. Reinhart Koselleck has suggested that between 1750 and 1850 the concept 'history' (*Geschichte*) replaced that of 'histories' (*Geschichten*). Each of the great historical works of the eighteenth century followed a continuous, coherent story whether in Gibbon's *Decline and Fall of the Roman Empire*, Hume's *History of England*, or Voltaire's *Essai sur les moeurs*.

Yet the conception of world history that arose with the Enlightenment constituted not only an extension of the historical outlook but at the same time also a narrowing. And this conception dominated almost all of Western historical thought and historical writing until well into the second half of the twentieth century. The history of the world was identified with the history of civilization and the latter rather narrowly with that of the West. Two variations emerged in the conception of what constituted world history, one identified with Condorcet, centering around an idea of progress on a world scale, the other identified with Herder, which stressed the diversity of peoples and cultures. Yet on closer examination, these two conceptions shared many characteristics, specifically in their stress on development.

The Enlightenment has too many aspects to permit one to reduce it to a simple formula. Two ideas were central to it, but neither was free of contradictions. The first assumed the essential equality of all human beings in all cultures; the second maintained that only human beings who had attained a high degree of culture and education deserved to be free and equal citizens. The latter view distinguished between societies that were in the vanguard of civilization, understood as Western civilization, and less-developed peoples. If G. W. Leibniz and Voltaire had still regarded China as a civilization equal and possibly superior to that of the West, Christoph Meiners considered the yellow peoples as inferior to Europeans, and the Blacks as barely different from animals. The notion emerged of a distinctive Western civilization with roots in classical Greece. J. J. Winckelmann's rediscovery of Greek art in 1764 contributed to an idealization of Greek standards of beauty and personal

culture which were adopted as models for a modern culture distinguishing the West from other cultures and giving it its superior standing. Meiners' scientific racism represented an extreme position, but the idea of the unique position of the West and the relative inferiority of all non-Western cultures and peoples became a central assumption in Western historical consciousness.

2. Nineteenth-century Conceptions of History and Historical Science: Between Romantic Historism and Scientistic Positivism

This outlook was dominant well into the twentieth century. It reflected the changes that had taken place in the world in political and military strength, economic power, and scientific and technological development. The two divergent historical outlooks in the first half of the nineteenth century, one represented by Hegel's dialectical philosophy of history, the other by Ranke's historism, shared this outlook despite their philosophic differences. Hegel's perception of the world as an ongoing dialectical process by which, through progressive contradictions, the absolute mind, reason, fulfills itself and takes on concrete forms in the institutional life of the modern European world, and specifically in the Prussian political and social order after the French Revolution and Napoleon, reflects its origins in specifically German political and philosophical traditions. But it has parallels in less involuted intellectual currents of thought outside of Germany, from Turgot, Condorcet, and Thomas Paine in the late eighteenth century to Henri de Saint Simon, Auguste Comte, J. S. Mill, Thomas H. Buckle, Herbert Spencer, and Benedetto Croce in the nineteenth century. For all of them world history is the history of the European world, a world which for some may include the United States but for all excludes the Slavic peoples. For Hegel only the West has a history. China, India, and Persia may have been the earliest civilizations but they have remained stagnant.

According to Hegel, Black Africans are incapable of cultural creativity. Karl Marx with his conception of static Asiatic modes of production took over much of Hegel's conception of the inferiority of non-Western societies and he and Friedrich Engels used this to justify Western imperialism as the only way to bring the non-Western world into the realm of modern civilization.

Theoretically Ranke's conception of the historical world represented the antithesis of Hegel's. Ranke rejected the idea of progress as well as the schematization of historical development as it occurred in Hegel's philosophy of history. History for him deals with the particular and the contingent and avoids generalizations and determinism. In fact, however, Ranke saw history as a linear process at the culmination of which stood the post-Napoleonic Western and Central European world. Ranke proclaimed that 'every age is equally immediate to God' (1973, p. 52), but at the same time denied that China or India had a true history. Except for his early work on the Serbs, his writings deal with the countries he defined early as the Latin and Germanic. His incomplete *World History* is a history of the West.

Whatever the philosophical and political differences, nineteenth-century thinkers as well as the educated public generally turned to history as the key to understanding things human that replaced traditional philosophy. The charges, often repeated, that the eighteenth century, particularly the Enlightenment, ignored history, generalized, and moralized, and did not recognize individuality and uniqueness of historical contexts is disproved when we look at the great historical works of the eighteenth century which we have already mentioned. Despite the impact of romanticism in the first part of the nineteenth century, with its stress on imagination and poetry, a cult of science dominated historical study and writing, which was not necessarily in conflict with a romantic outlook.

This search for a scientific approach to history took two different directions: one that has frequently been called positivism, the other historism, neither of which can easily be reduced to a common denominator.

The first of these orientations took the natural sciences as its model. Building on Turgot and Condorcet, Auguste Comte sought to formulate laws of development by which societies moved from simpler to more complex forms of organization. The driving force was the transformation of knowledge from theological and then metaphysical stages to positive science that would provide the means for the scientific organization of society. But positivism remained largely grand theory with little empirical application to historical inquiry. More significant was Karl Marx's approach to history. Marx saw the driving force of history in the lawful development of economic forces, which constituted the basis of historical transformation, and the revolt of this basis against the social, political, and intellectual superstructure. In the course of this conflict the contradictions within the social order, which involved inequality, would be overcome in a classless society. Marx thus intended to replace Hegel's idealist dialectic with a materialistic one.

If positivism both in its Comtean and in its Marxist form in its search for a scientific history envisaged a utopia at the end of history, the main currents of historical scholarship sought to avoid a speculative approach to history but nevertheless did not free themselves from metaphysical and even theological presuppositions which they refused to acknowledge. Ranke may perhaps serve as the most influential representative of this orientation. He too was committed to a strictly scientific approach to history. For him, however, this approach must recognize that general laws have no application to history that deals

with the particular and the contingent. For him the task of the historian was to recapture the past as 'it actually was' (*wie es eigentlich gewesen*), free of any bias, requiring the application of philological methods to the critical analysis of the sources as F. A. Wolf and Niebuhr had applied them. But for him history was more than the reconstruction of the past. It involved interpretation and synthesis, as it did for J. G. Droysen. Ranke was convinced that the particular event contained the general, and that the general became apparent to the researcher who immersed himself in the sources, because—and here we return from the realm of strict science to that of metaphysics—the historian can observe the general tendencies that manifest themselves in history, tendencies which do not represent God's direct intervention in the world, but nevertheless reflected divine purpose.

3. *Professionalization of Historical Studies*

The stress on the scientific character of historical studies led to the formation of a professional discipline of history. Professionalization was not restricted to history but was part of a process in which scholarly studies in all fields, in the liberal arts as well as the natural sciences, were institutionalized at centers of higher learning, primarily the universities and research institutes. The German model was imitated elsewhere, relatively early in France with the introduction of the German seminar method at the Ecole Pratique des Hautes Etudes in Paris in 1868 and in the United States with the introduction of a research oriented graduate program at The Johns Hopkins University in Baltimore in the 1870s leading to a Ph.D. With some delay the German model was adopted in Japan, Great Britain, the Netherlands, Italy, Spain, Russia, Scandinavia, Greece, Romania, Latin America, and elsewhere, each time, it must be stressed, with major modifications that reflected different scholarly traditions. Thus, the sharp distinction between the cultural and the natural sciences which dominated German thought played less of a role elsewhere.

A basic tenet of professional historians was the total objectivity or what Ranke called the 'impartiality' of the historian. In fact, however, the new discipline was highly politicized everywhere. This intensified interest in archival research went hand in hand with the wave of nationalism that gripped all Western countries and also modernizing non-Western countries such as Japan. The broad cosmopolitan view of many historians in the eighteenth century gave way to a new focus on national history. Ranke still wrote the histories of a variety of European states, Serbia, England, France, Prussia, and also the Reformation in Germany, and the Papacy. Later historians, who took him as a model for their own scholarship, in most cases concentrated on their own nation.

However, the development of modern historical scholarship was not a German monopoly. The opening of archives and the collection of documents, primarily those of the Middle Ages, became a prerequisite for scholarly national history. In Paris the *Ecole des Chartes* was founded in 1821 to train scholars in the critical use of documents. Muratori's collection of Italian documents in the eighteenth century has already been mentioned; this was followed by the *Monumenta Germaniae historica* launched in Germany in 1819 and by similar undertakings in France, Great Britain, Spain, Poland, Bohemia, Greece, and elsewhere.

The Rankean belief that immersion into the sources guarantees an impartial view of the interrelation among historical events turned out to be an illusion, something which was openly recognized by the scholars of the Prussian school such as Heinrich von Sybel, J. G. Droysen, and Heinrich von Treitschke who frankly saw historical scholarship as a tool for the legitimation and propagation of the German national cause under Hohenzollern rule. Moreover, the process by which scholars were recruited guaranteed a high degree of homogeneity in social composition and political outlook.

Professionalization was to guarantee objectivity, but also to provide a scientific or scholarly discourse distinct from that of literature. Ranke had stressed the scientific character of history, but at the same time he also recognized that history combines science with art. The great prestige which history enjoyed everywhere in the nineteenth century—a number of the leading French statesmen from François Guizot to Alexis de Tocqueville, Alphonse de Lamartine, Adolphe Thiers, and Jean Jaurès were important historians—rested on the literary quality of their histories. History could not escape being literature, and imaginative literature at that. In fact there existed a close parallel between the great histories written by professional historians and the great realistic novels of the nineteenth century—Stendhal, Honoré de Balzac, Charles Dickens, Gustave Flaubert, Gottfried Keller, Gustav Freytag, Ivan S. Turgenev, Fyodor Dostoyevski, and even Thomas Mann's *Buddenbrooks* (1902). They all constructed a plot that followed a temporal line and involved individual subjects as agents. And the great historians of the century, Jules Michelet, Denis Fustel de Coulanges, Thomas B. Macaulay, and Theodor Mommsen, who in 1902 received the Nobel Prize for literature, were read as great literature, as was Ranke himself. Michelet is a striking example of the crossing of borders between critical scholarship and imaginative literature, the exhaustive mining of the archives for the purpose of creating a national memory and identity.

But the role professionalization played in nineteenth-century historical writing should not be overestimated as it often has been. A great deal of history did not follow the dominant paradigm of narrative

national history such as three of the most important and lasting works of the mid-century, Fustel de Coulanges's *La Cité antique* (The Ancient City), de Tocqueville's *L'Ancien régime et la Révolution Française* (The Old Regime and the French Revolution), and Jacob Burckhardt's *Die Kultur der Renaissance in Italien* (The Civilization of the Renaissance in Italy). Moreover, many leading writers of history such as Thomas Carlyle, Macaulay, John Green, George Bancroft, Lamartine, Louis Blanc, and, of course Tocqueville, were not professional historians.

The great literary histories of the eighteenth century, for example, those of Hume, Gibbon, and Voltaire, were intended to entertain, educate, and edify a broad educated public. Professional historians wrote often for specialists. But in fact the division between literary and professional history and their audiences was not as deep as it may appear. Certainly Hume, Gibbon, and also William Robertson were serious scholars and the great professional historians of the nineteenth century, such as Ranke, Fustel de Coulanges, Theodor Mommsen, and Lord Acton were read avidly by a broad public for whom history had become the key science replacing philosophy as a guide to understanding the human world.

4. Twentieth-century Reorientations of Historical Thought and Scholarship: The Turn to the Social Sciences

Outside the discipline proper, critical methods were applied to many areas: the history of religion (F. D. Strauss, Ernest Renan), the history of law, the history of literature and art, and other fields. While the historians proper neglected the study of society and economics, an important group of economic historians, the so-called 'Historical School of National Economy' in Germany (Bruno Hildebrand, Karl Knies, Wilhelm Roscher, and Gustav Schmoller) pursued a middle course between an empirical approach, seeking generalizations of economic behavior and development, reminiscent of positivism, and an intense regard for the particular historical setting of culture and society in which economic behavior occurred, which reflected historicist principles.

The late nineteenth and the turn to the twentieth century witnessed profound reexaminations of the historical outlook that had been dominant until then. On the one hand, there was a continued, even a heightened faith in science, technology, and the progress of Western civilization which largely survived the trauma of World War I. On the other hand, there was an increasing sense of discomfort with the quality of modern civilization. Darwinism, or rather Social Darwinism, played a fundamental role in both of these outlooks. The world was seen as a struggle for the survival of the fittest that could forecast the progressive victory of the white races over the rest of the world, which for Oswald Spengler was a desperate struggle for survival. At the turn to the twentieth century, some Europeans and Americans foresaw progress in terms of growing tolerance and democracy, while others preached racism and ultra nationalism.

As far as historical thought was concerned, both Burckhardt and Friedrich Nietzsche voiced their discomfort with a modern industrial society with its concomitant social and cultural aspects, a mass civilization moving to commercialism and democracy, and extended this dismay to a critique of the modern faith in science. Nietzsche, Wilhelm Dilthey, Henri Bergson, and Georges Sorel in differing ways called for a restoration of the vital against the mechanical aspects of life and culture. Established patterns of historical scholarship, as practiced by the historical profession, were called into question simultaneously at the end of the nineteenth century from Russia and Poland to North America. Generally, however, the critics of the historical profession did not criticize the latter because it was too scientific but rather because it was insufficiently so and had not taken into consideration the tremendous changes in modern society and civilization in an industrial and democratic age.

The most spectacular confrontation with the established profession occurred in Germany with the so-called *Methodenstreit*, occasioned by the publication of the first volume of Karl Lamprecht's *Deutsche Geschichte* in 1891. Lamprecht criticized his colleagues for their narrow focus on the state and their neglect of broader cultural and social aspects of history. He also accused them of working with an outdated concept of science and historical science that restricted itself to events and personalities rather than taking into consideration the structural context of historical change. The German professoriate, reflecting the peculiar semi autocratic character of the Prussian German state, vociferously rejected Lamprecht's initiative and saw in it a threat to the established social and political order. Lamprecht's attempt to formulate laws of historical development was rejected internationally but his call for a broadening of historical perspective to encompass society, economics, and culture was generally welcomed outside of Germany. Historians in Russia (Paul Vinogradoff, Vasiliev Kliuchevskii, Paul Miljukov), Romania (Nikolae Iorga), Poland (Jan Bujak, Franticzek Rutkowski), Norway (Halvdan Koht), Belgium (Henri Pirenne), as well as the circle around the journal *Revue de synthèse historique* founded by the philosopher Henri Berr in 1900 as a forum for an international discussion, and the American 'New Historians' (F. J. Turner, J. H. Robinson, Charles Beard, Carl Becker, H. E. Barnes) called for an expansion of the scope of history and a closer cooperation with the empirical social sciences.

Two variations of a social science approach to history need to be mentioned. The first is Max Weber's historical sociology that moves between the historistic

emphasis on the recognition of the uniqueness of cultures and a historical social science working with clearly defined concepts and a rigorous logic of inquiry. Weber's conception of social causation and historical development was in some ways indebted to Marx, but unlike Marx emphasized the interrelation between culture, including religion, and economics, and placed much greater weight on contingency. Weber recognized that societies constituted meaningful relationships that required an interpretive sociology but avoided the sharp distinction that idealistic philosophers such as Wilhelm Windelband, Heinrich Rickert, and Wilhelm Dilthey made between the generalizing natural sciences and the cultural sciences (*Geisteswissenschaften*) that favored empathetic understanding (*Verstehen*).

A second approach to a history taking into account the social sciences was offered by the so-called '*Annales* school.' It is difficult to reduce the *Annales* to a common denominator. The journal by this name, launched in France in 1929 by Lucien Febvre and Marc Bloch with the support of Henri Pirenne, continued initiatives begun early in the century in Henri Berr's already mentioned *Revue de synthèse historique*. A similar project was initiated in Poland in the journal *Roczniki Dziejow Społecznych i Gospodarczych*, the editors of which, Rutkowski and Bujak, were in close contact with the *Annales*. In the place of the concentration of the established historiography on events and great personalities, the *Annales* explored impersonal forces and structures.

The 'human geography' of Vidal de la Blache played an important role in the early works of Lucien Febvre and Marc Bloch. Human history had to be understood in the concrete context of space and time. But unlike the German Friedrich Ratzel, who saw soil and race (*Volk*) as determining factors in history, Paul Vidal de la Blache, as did Febvre, Bloch, and Fernand Braudel later, recognized the role which human agency and culture played in the formation of the geographic setting. The historians of the *Annales* attempted to pursue a 'total history,' generally limited to a restricted region, in which all aspects, geographic, economic, political, religious, and demographic, were to be included. Quite consciously the dividing line between the established disciplines were overcome and in the *Maison des Sciences de l'Homme*, opened in Paris in 1968, economists, sociologists, anthropologists, linguists, psychoanalysts, and literary critics were housed with history as the integrating science.

In other ways too, the *Annales* represented a radical break with older historical conceptions. The nation no longer represented the basic unit of history; it was replaced by the region or by a broad comparative approach that extended beyond Europe (Braudel, Pierre Vilar). The concept of a linear time, so basic to all forms of nineteenth century historical thought, and with it the idea of progress, were abandoned and a historical epoch viewed in its own terms, not as part of a historical process. Burckhardt had already done something similar in 1860 in *Die Kultur der Renaissance in Italien*, and had worked with a very broad concept of culture, as had also Jan Huizinga in *Herfsttij der Middeleuwen* (1919), but unlike the *Annales* historians they focused heavily on elites. In contrast the *Annales* dealt with the population at large, but ran the danger that at times their history became impersonal and void of concrete human beings.

In retrospect, the racist historiography of German National Socialism of the period 1933–1945, although it built on widespread Social Darwinist ideas and ultra nationalism, remained an interlude, which even in Germany had limited influence. The Marxism-Leninism of the Soviet Union and its dependent states in its dogmatic forms, in many ways a perversion of Marxism, similarly represents an interlude. As early as 1956, historians in Poland and Hungary, and until the freeze after the Soviet occupation in Czechoslovakia, and even in the Soviet Union (Aaron Gurevich), participated in the broader historical discussions of the international community.

5. *New Orientations after 1945: The Turn to a Cultural Dimension*

A subsequent article will deal with trends in historical thought and study in the second half of the twentieth century. The victory of the Allies in World War II and the dominance of the United States in the postwar world revived older beliefs in progress and science. The 1950s and 1960s witnessed a turn to the quantitative social sciences in historical studies.

The years since the late 1960s have seen a reorientation in historical outlook in the Western world generally, including to a lesser extent also Eastern Europe and Russia. The conflicts of the 1960s, the racial confrontation in the United States, the Vietnam War, the student revolts throughout Europe and North America, and the suppression of the Prague spring all contributed to a reorientation as did the fundamental technological changes which resulted in the transformation of an industrial society concentrated on production into an economy based on information technology, in which consumption and services play a key role and the old socioeconomic classes gave way to new ones. The confidence in science and economic growth was severely tested as a new generation became increasingly conscious of ecological concerns and called for the emancipation of subordinated segments of the population such as women and ethnic minorities at home and the peoples of the former colonies abroad. The very quality of the modern Western world, including its science, came under attack. All of this had consequences for historical studies after the 1960s.

See also: Historicism; Historiography and Historical Thought: Christian Tradition; History and the Social Sciences; History: Forms of Presentation, Discourses, and Functions; History, Professionalization of; History: Theories and Methods; Modernity; Modernization and Modernity in History

Bibliography

Bann S 1984 *The Clothing of Clio: A Study of the Representation of History in Nineteenth-century Britain and France*. Cambridge University Press, New York

Barnes H 1962 *A History of Historical Writing*. Dover, New York

Barraclough G 1978 *Trends in History*. Holmes & Meier, New York

Bentley M 1999 *Modern Historiography*. Routledge, London

Berr H, Febvre, L 1932 History In: *Encyclopedia of the Social Sciences*, Macmillan, New York, Vol. 7, pp. 357–68

Blanke H W 1991 *Historiographiegeschichte als Historik*. Frommann-Holzbook, Stuttgart-Bad Canstatt, Germany

Breisach E 1983 *Historiography. Ancient, Medieval, Modern*. University of Chicago Press, Chicago

Butterfield E 1955 *Man on His Past. The Study of the History of Historical Scholarship*. Cambridge University Press, Cambridge, UK

Cassirer E *Philosophie der Aufklärung* (*The Philosophy of the Enlightenment*). Mohr, Tübingen, Germany

Chickering R 1993 *Karl Lamprecht: A German Academic Life* (*1856–1915*). Humanities Press, Atlantic Highlands, NJ

Fueter E 1911 *Geschichte der neueren Historiographie* (*History of Modern Historiography*). Oldenbourg, Munich, Germany

Gazi E 2000 '*Scientific*' *History. The Greek Case in Comparative Perspective* (*1850–1920*). Peter Lang, New York

Gooch G 1913 *History and Historians in the Nineteenth Century*. Macmillan, London

Iggers G 1968 *The German Conception of History. The National Tradition of Historical Thought from Herder to the Present*. Wesleyan University Press, Middletown, CO

Kelley D 1998 *Faces of History. From Herodotus to Herder*. Yale University Press, New Haven, CI

Keylor W 1975 *Academy and Community. The Foundation of the French Historical Profession*. Harvard University Press, Cambridge, MA

Koselleck R 1975 Geschichte, Historie. In: Brunner O, Conze W, Koselleck R (eds.) *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland Vol. 2*. Klett, Stuttgart, Germany, pp. 593–717

Novick P 1988 *That Noble Dream. The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK

Ranke L von 1973 The theory and practice of history. In: Iggers G, von Moltke K (eds.) Bobbs Merill, Indianapolis, IN

Reill P H 1975 *The German Enlightenment and the Rise of Historicism*. University of California Press, Berkeley, CA

Ringer F 1969 *The Decline of the German Mandarins. The German Academic Community 1890–1933*. Harvard University Press, Cambridge, MA

Thompson J W 1942 *A History of Historical Writing*, 2 vols. Macmillan, New York

White H 1973 *Metahistory. The Historical Imagination in Nineteenth-century Europe*. The Johns Hopkins University Press, Baltimore, MD

Woolf D R (ed.) 1998 *A Global Encyclopedia of Historical Writing*. 2 vols. Garland Publishing, New York

G. G. Iggers

Historiography and Historical Thought: South Asia

This article addresses two interlinked subject matters, the history of South Asia to which scholars from all over the world have contributed, and historiographical developments in the region of South Asia itself. It briefly covers the historical tradition in pre-modern times and primarily focuses on developments since World War II.

1. Historiography of Ancient India

1.1 Historical Traditions in Ancient India

Compared to classical Europe, ancient India was rather deficient in historical traditions. The Hindu mythological texts called the *Puranas* (circa first to fourteenth century) contained geneological lists of Indian dynasties. These lists, by way of narrating the achievements of the kings, also recorded rudimentary information about wars, conquests, and the rise and fall of empires. A very small number of works chronicle the history of individual kings and dynasties. Of these, the most famous is Kalhana's *Rajatarangini* (Stein 1900), a history of the Kashmir princes. Incidentally, the author recognized one historical principle, namely, that the historian must try to be unbiased in his/her judgment, 'like a judge.'

The Buddhists and Jains developed a tradition of sacred history, especially hagiographies centered on the life-stories of their founders. Some of the stories concerning the Buddha, like the account of his death (*Mahaparinivvana-sutta*, Bongard-Levin 1986) bear the stamp of authenticity.

1.2 European Scholarship on Ancient India

In the nineteenth century, European scholars like Pargiter tried to reconstruct the dynastic history of ancient India using the Puranic lists, inscriptions, coins, archaeological remains, and foreign accounts of India like those of Alexander's historians and the Chinese Buddhist pilgrims. European Orientalists also tried to reconstruct the religious and social history of India, using religious texts and the *smritis*, normative

writings laying down the rules by which the Brahminical Hindu was expected to live in matters both ritualistic and secular (like inheritance and punishment for crimes). They also undertook the massive task of publishing critical editions of religious texts which laid the foundation for serious work on religious, cultural, and social history. A large number of German, French, and East European scholars have contributed to this enterprise.

1.3 Historiography of Ancient India: The Indian Contribution

A large body of Indian scholars started research on ancient Indian history following the lead of the European scholars. Their work was heavily focused on archaeological evidence, especially the study of inscriptions. The archaeological work, by Europeans and Indians, helped move back the known beginning of civilization in South Asia to about 3000 BC through the discovery of the Indus valley civilization. The work of the Indian Indologists also reflected the new-found nationalistic pride in the glories of ancient India. Some of it focused on traditions of democratic/oligarchic politics and developed statecraft to challenge the notion that these were European imports into India. However, the central concern of the earlier work was with the meticulous reconstruction of dynastic history, especially the correct chronology. Some scholars like R. K. Mukherji and R. C. Majumdar did research on the history of Indian cultural influence in south-east Asia and invented the now discarded notion of Greater India.

1.4 'Orientalism'*: A Controversy*

Edward Said's *Orientalism* (1995), which projects the view that western scholarship on Asia projected the notion of an inferior 'other' of Europe and thus reflected the new relationships of power generated by imperialism, has had a seminal influence on the historical work on South Asia (since 1980) even though Said's primary point of reference was the Middle East. The controversy generated by this work has drawn attention to the many strands of Orientalist studies in the West and to the fact that French and German studies of South Asia do not project any unquestioned superiority. For ancient India, the controversy has led to more research on western historiography than to any reorientation in exploration of substantive themes.

1.5 Indology after World War II

The post-World War II historiography of ancient India has moved away from these earlier preoccupations and, under the influence of Marxism as well as anthropology, concerned itself with discovering structures like feudalism and patterns of large-scale social change over time. The writings of D. D. Kasambi (1965), R. S. Sharma (1990), Romila Thapar et al. (1990) are prime examples of this initiative. The multivolume *History and Culture of the Indian People* edited by R. C. Majumdar (1951–1969) reflects a more traditional approach, while Louis Dumont's *Homo Hierarchicus* (1972) applies Levi-Strauss' structural analysis to the study of the caste system.

2. Historiography of the Indo-Islamic Period

The thirteenth century saw Turkish and Afghan dynasties and tribes, Muslims by faith, firmly established as rulers in many parts of India. The sultanate based in Delhi, where there was a succession of dynasties, became the seat of a pan-Indian empire with fluctuating boundaries. These dynasties were succeeded by the Mughals or Timurids in the sixteenth century and their effective rule lasted until the late eighteenth century. This long period, marked by the encounter between the older Indian civilization and the religion and culture of Central Asia's Islamic peoples, has been the subject matter of a rich historical literature.

2.1 The Medieval Persian Chronicles

India's central Asian invaders brought with them a rich tradition of historiography. This literature, written mostly in Persian, is known as *tarikhs*, a term, translated as annals. The *tarikhs* are narratives of dynastic successions, wars and conquests. Some of the writers had leanings towards Islamic fundamentalism and hence were inclined to present the wars of the Muslim invaders and Hindu dynasts as *jihad* or 'holy' wars. These accounts were later used by some British writers to project the Indo-Islamic era as an age of Muslim tyranny. This perception informed the ideology of India's proto-nationalism, which saw the struggle of the Rajput clans and other Hindu chiefs to retain their territories as the Hindu struggle for independence. However, chronicles provide ample evidence of a pragmatic bent in the policies of the Muslim kings, anxious not to alienate their Hindu subjects. The historians at the court of the third Mughal emperor Akbar, project his policy of peace with all, which was not to the liking of the more orthodox elements at his court. One component of the historical literature of the period is the memoirs written by the emperors, Babar and Jahangir and by Babar's daughter, Gulbadan. These sensitively written accounts offer a portrait of the age and the family life of the Mughals.

2.2 Regional Histories

The Indo-Islamic era also produced a rich and varied body of historical writings written in the regional languages. Best known among these are the Marathi *bakhars*, modeled on the Persian *tarikhs*, the chronicles of the Ahom kings of Assam, known as *buranjis*, and the *Madla Panji* retained in Orissan temples which recorded the annals of the local dynasties. Besides these, the Vaishnavas in Bengal produced a body of hagiographical literature celebrating the lives of their founder, Chaitanya and other pioneers of the move ment. All these products of the period have been later used to reconstruct the history of the period.

2.3 Colonial and Post-colonial Historiography of the Indo-Islamic Era

The earliest 'modern' historical writings on the Indo-Islamic era go back to the nineteenth century. Works like Orme's *History of Indostan* are amateur efforts focused on the political history of the period. The famous collection of selected source material, translations from the Persian chronicles, Eliot and Dawson's *History of India as told by her own Historians* (1981), contain introductions which are essentially historical statements. One of the compiler's stated objects was to prove how very inferior 'Muslim rule' was compared to its British counterpart. In general histories of India, written by administrator–scholars like Vincent Smith (1981) the same concern to prove the superiority of British rule is less obviously present. It only appears as a view of Indian history which projects a long-term proneness towards political unification that finally succeeds only under the British. Such works contain an implicit evaluation of moral and material welfare of the people in pre-British times. The twentieth-century economic historian W. H. Moreland (1920) undertook a comparison between India at the death of emperor Akbar with conditions in the nineteenth century. Serious academic research on the Indo-Islamic era, based on the study of the chronicles as well as documentary evidence like the *akhbars* or newsletters of the Mughal period, belong to the twentieth century. It produced magnificent works of scholarship like Sir Jadunath Sarkar's multi-volume histories of Aurangzeb and the later Mughals focused on political narratives. There were other studies dealing with the religious policies of the Mughals and art and architecture of the period. The burgeoning nationalist consciousness sought to emphasize the syncretist traditions of the Indo-Islamic era, implying that the conflict between communities was a product of British rule.

Since World War II, a major development in this field of study resulted from the work of a group of scholars based at Aligarh. Irfan Habib's *Agrarian System of Mughal India* (2000), drawing upon a wide range of source material, was hailed as a path-breaking contribution to the study of agrarian history. The research of the Aligarh 'school' uses the methods of modern historiography and has drawn upon a vast quantity of new material. While some of the scholars subscribe to Marxist theories of history, the impact of Marxism is not very obvious in their work.

In recent years some Western historians have projected the view that the Mughal empire, like all its predecessors in India, was a 'segmentary state' under a thin veneer of centralization. This has led to a controversy with some of the Aligarh historians.

3. Historiography of the Colonial Era

3.1 Colonial Historiography under Colonial Rule

British historians writing in the colonial era produced a large body of works which recorded the history of the conquest, the wars with indigenous powers, and histories of the local powers like the Marathas, Sikhs, and Mysore with whom they had to contend for power. These histories are in the purely narrative mode, supplemented by ethnographic accounts of various elements in Indian society. They are informed by a simple assumption regarding the legitimacy and naturalness of the British conquest. One of the pioneers in the field, James Mill (Mill and Wilson 1997), writing in the early decades of the nineteenth century, first projected a clear notion of cultural and racial superiority, significantly absent from earlier British writings on India.

British writings in the latter half of the nineteenth and early decades of the twentieth century often assumed a defensive posture in response to the growing nationalist criticism of their rule, especially the charge that the colonial nexus had induced a drainage of resources from India, thus steadily impoverishing the country. British historians challenged this view and instead projected a picture of economic transformation and development as a result of colonial rule. This controversy has not yet died out. The British historiography of colonial rule before World War II almost entirely focused on the activities of the government. A number of studies dealt with individual viceroys. Some twentieth-century works briefly refer to the burgeoning nationalist movement (one historian described it under the heading 'Political Sentiment') but with one exception they dismissed it as a phenomenon not to be taken seriously until the very eve of the transfer of power. However, the British colonial historiography increasingly adopted the methods of modern research with its emphasis on 'objective' analysis of records. But since the records in question were for the most part products of the colonial government, the resulting research inevitably reflected the built-in bias of the documents, a bias which the

historian generally shared. Indian historians, excepting those writing on economic history, rarely ventured into comments on the history of British rule. Exceptions to this pattern were the publicists writing on the rising of 1857 (one of whom described it as India's First War of Independence) and authors who considered British rule to be beneficial for India. The strong ideological and political undertone manifest in the historiography of the colonial era has assumed more subtle forms in the post-colonial period.

3.2 Historiography of Modern India after Decolonisation

Since the transfer of power in 1947, attention has shifted from the working of the British administration to the structure and dynamics of indigenous politics, though a number of monographs on the political history of the Raj, as well as its administrative and constitutional initiatives, were written during this period. Indian historians now came to be deeply concerned with the history of the Indian nationalist movement and its social–cultural origins. A group of historians approached the subject from the overt perspective of nationalism; others were influenced by Marxist class analysis and saw Indian nationalism as an expression of colonial bourgeois aspirations. A group of British historians centered at Cambridge explored the close relationship between British administrative and colonial initiatives on the one hand, and the structure and dynamics of indigenous politics on the other. The competition between those who were admitted to a share in power by the colonial ruler and those who were not was seen to be the central feature of indigenous politics. Nationalism itself was perceived to be a strategy reflecting the aspirations of the 'have-nots.' This view was questioned by Indian historians who saw the thesis as a continuation of the old colonialist perception dismissing nationalism as a conspiracy of the self-seeking few. However, the thesis was directly, or indirectly, questioned outside India as well. A group of historians under the leadership of Professor Low of the Australian National University recognized the reality of Indian nationalism and proceeded to unravel its local and class roots. American scholars brought to the subject the insights of social sciences, especially sociology and anthropology. There were considerable developments in economic history, using economic theory as well as statistical methods. The new techniques were used to tackle the old questions regarding the negative or positive impact of colonialism. The debate, as noted, remains inconclusive. The inadequacy of statistical evidence is one reason for this state. It should be noted here that Indian historical research has now become an international enterprise to which the Japanese, Russian, and continental scholars have made significant contributions.

One major new development in the subject is centered on the series of publications entitled *Subaltern Studies* (Guha et al. 1988). Borrowing the concept of the Subaltern from the Italian Marxist, Gramsci, the contributors to this series have focused on the history of the underprivileged elements in South Asian society and used the post-modernist techniques of discourse analysis to unravel the autonomous consciousness of the dispossessed and the record of their resistance to exploitation and domination by the elite. This movement, too, has become an international enterprise with historians and social scientists from all parts of the anglophone world contributing to the *Subaltern Studies*.

Pakistani historians and their pre-independence predecessors who focused on the development of Muslim consciousness and aspirations in the subcontinent have been deeply influenced by the theory of two nations, the ideological basis of the demand for Pakistan. Their perception is in sharp contrast to the nationalist perception of co-existence and mutual accommodation between the major ethnic cultures of the region. The history of Sri Lanka, despite its many links with the civilization and society of the mainland, has been unjustifiably peripheral to the concerns of the historians of India. But there was a strong tradition of historiography in the region drawing, *inter alia*, on the Buddhist religious texts for the ancient period. Some of the products of this historiography are influenced by Marxist theories of history.

One marked feature of developments in the historiography of modern India is the diversification of its subject matter. History of art and its links with nationalist sentiment, women's studies, and ecological history are some of the areas explored in recent years. This has helped bring the subject within the sphere of contemporary intellectual interest and brings to an end its intellectual isolation.

See also: Colonization and Colonialism, History of; Historical Sociology; Historiography and Historical Thought: Current Trends; History and the Social Sciences; History: Overview; Imperialism, History of; Orientalism; South Asian Studies: Culture; South Asian Studies: Economics; South Asian Studies: Politics; Subaltern History; Time, Chronology, and Periodization in History

Bibliography

Bongard-Levin G M (ed.) 1986 *Tripitaka, Sutraϕitaka, Mahaparinivvanasutra*. International Institute of Buddhist Studies, Tokyo

Dumont L 1972 Homo Hierarchius: *The Caste System and its Implication*. Paladin, London

Elist H M, Dowson J 1981 In: Nizami K A (ed.) *The History of India as Told by its own Historians: The Muhammadan Period*. Delhi, India

Frykenberg R E *India to 1858*; Raychaudhuri T. *India, 1858 to the 1930s*; Robin J. Moore, *India in the 1940s Winks, R W.*, 1999 *The Oxford History of the British Empire, Historiography*, Oxford University Press, Oxford Vol. 5
Gallagher J, Seal A, Johnson G 1973 *Locality, Province and Nation Essays on Indian Politics 1870 to 1940*. Cambridge University Press, Cambridge, UK
Guha R, Chakravarti S, Gayatri 1988 *Selected Subaltern Studies*. Oxford University Press, New York
Habib I 2000 *Agrarian System of Mughal India 1556–1707*, 2nd edn. Oxford University Press, New Delhi, India
Kasambi D D 1965 *The Culture and Civilisation of Ancient India in Historical Outline*. Routledge, London
Majumdar R C 1951–1969 *History and Culture of the Indian People,* 10 Vol. Bharatiya Vidya Bhavan, Bombay, India
Mill J, Wilson J H 1997 *The History of British India*. Routledge, London
Moreland W H 1920 *India at the Death of Akbar*. Macmillan, London
Philips C H (ed.) 1961 *Historians of India, Pakistan and Ceylon*. Oxford University Press, London
Said E W 1995 *Orientalism*. Penguin, Harmondsworth, UK
Sarkar S 1985 *Modern India: 1885–1947*, 2nd ed. Macmillan, New Delhi; 1997 *Writing Social History,* Oxford University Press, Delhi
Smith V 1981 In: Spear T G (ed.) *The Oxford History of India*. Oxford University Press, Delhi, India
Stein M A (trans.) 1900 *Kalhana's Rajatarangini: A Chronicle of the Kings of Kasmir*. Westminster, UK
Sharma R S 1965 *Indian Feudalism, 300–1200*. Calculta University Press, Calculta
Thapar R 1990 *A History of India*. Penguin, Harmondsworth, UK

T. Raychaudhuri

Historiography and Historical Thought: Southeast Asia

Embracing the southeastern peninsula of Eurasia and the Indonesian and Philippine archipelagoes, Southeast Asia has recorded its history in a dozen written languages and countless oral traditions. Indic models of writing were influential in the millennium before 1200 CE, and Chinese and Islamic ones in the subsequent six centuries, while European enlightenment interpretations of history became increasingly dominant in the nineteenth and twentieth centuries. Diversity has always been a hallmark of this region. The Vietnamese emulated the Chinese literary tradition with their dynastic chronicles; cyclical Indic ideas had a longer career in the Theravada Buddhist societies of the mainland, while linear Muslim and Christian ideas became influential in the islands. Nevertheless, synchronicities are apparent in the developments within seemingly autonomous literary and historiographic traditions.

Oral poetry and traditions were important for most peoples, but particularly for those in the highlands and eastern islands (including the Philippines) which did not absorb Indic or Chinese literary forms. As late as the 1960s, at least one people in Roti (near Timor) retained oral historical narratives which accurately recorded the names of chiefs for three or four centuries. Most of the written chronicles which took shape in the sixteenth and seventeenth centuries incorporated older traditions with the repetitive features of oral narrative.

1. Chronicle Traditions

The dynastic chronicles of Vietnam were expressed in Chinese and followed the model of didactic Chinese chronicles heavily informed by the moral perspective of Tang and Sung Confucianism. Some such work survives from the fourteenth century, but the most important is the strictly Confucian chronicle composed by Ngo Si Lien in 1479. This in turn incorporates the *Dai Viet su ky* of Le Van Huu (1272), which included a historical justification for the independent imperial status of Dai Viet—then occupying little more than the Red River delta.

The remaining written traditions grew out of Indic models in Sanskrit and Pali, or in the Malay case from Islamic writing in Arabic and Persian. Sacred histories, in Mon (the then language of what is today southern Burma), Burmese, and Khmer, as well as in Pali, traced the establishment of Buddhist law and practice from the time of the Buddha to the founding of local temples in Southeast Asia. Among the earliest examples are the northern Thai *Mulasasana* (around 1420) and the Pali *Jinakalamalipakaranam* (1516), both written by monks in Chiang Mai.

Royal and dynastic chronicles are almost as old. The *Pararaton*, a unique chronicle in old Javanese of the Javanese kingdoms of Singosari and Majapahit down to 1481, was written after the fall of Majapahit to Muslim forces around 1527. The earliest extant *yazawin* (royal genealogy or chronicle) in Burmese was written by the monk Shin Thiawuntha in 1520. Dynastic chronicles (*phongsawadan*) were being compiled in Thai from at least the sixteenth century, although the Luang Prasoet text conventionally dated 1680 is the oldest of the extant versions which have come down to posterity. Written by an astrologer at the cosmopolitan court of King Narai (1657–1688), this chronicle marks a trend to more secular dynastic history well aware of Siam's position among its neighbors as well as European, Indian, and Chinese visitors.

The advent of Islam created more discontinuities in the south than Theravada Buddhism did in the north. As Southeast Asian Islam shifted from an Arabic to an indigenous idiom in the fifteenth and sixteenth centuries, the Malay language written in Arabic script

became its favored vehicle. Only in Java and South Sulawesi, and of course in non-Muslim Bali, did the older literary idiom and alphabet continue alongside the Arabic/Malay one. Probably the oldest extant Malay chronicle is that of the Kings of Pasai, in north Sumatra, the first substantial Islamic state in Southeast Asia, from the 1290s. This text, *Hikayat Raja-raja Pasai*, is known in only one copy, made in 1814 from an older version held in Demak, the fifteenth-century capital of Muslim Java. Since it ends with the story of the fourteenth-century conquest of Pasai by Hindu-Buddhist Majapahit (Java), and the bringing of many Pasai Muslims as captives to Java, its kernel may have originated with this group of Malay-speaking Muslim exiles. It represents a collection of semi-mythical stories about the Pasai kings, much more reflective of local oral traditions than of mainstream Islamic historiography.

Several Malay texts celebrating other royal lineages, first written down in the sixteenth or the early seventeenth century, take this story-telling mode to a high art form extremely illuminating for the historian, though usually devoid of chronology. The *Hikayat Banjar*, for example, tells the story of the establishment of the leading Islamic state in southern Borneo through a conservative ethic extolling the virtues of a Javanese cultural model. Much the best-known text of this genre is the 'Genealogy of (Melaka) Kings' (*Sulalat'us-Salatin*), a version of which Thomas Raffles printed and popularized in 1821 under the name 'The Malay Annals' (*Sejarah Melayu*), in a conscious attempt to provide a national history for Malays. It weaves through a host of lively anecdotes its political message that the great metropolis of Melaka fell to the Portuguese in 1511 because the fundamental contract between ruler and ruled was violated by the last Melaka ruler. Originating in some form a generation after this Malay disaster, the text appears to have been essentially created in the early seventeenth century by the Johor writer Tun Seri Lanang, though subsequently much modified. Not much later in time, but a great deal more Islamic in spirit, was the encyclopedic 'Garden of Kings' (*Bustan as-Salatin*), written by the Gujarat (India)-born theologian Nuruddin Ar-Raniri around 1640. Written in Malay at the behest of the Sultan of Aceh, in northern Sumatra, this includes a substantial chapter on the history of Aceh, and another on Melaka, and represents the most skilful use in Malay of Islamic traditions of historiography.

Java's conversion to Islam was a turbulent and drawn-out process centered in the fifteenth and sixteenth centuries. The syncretic Islamo-Javanese courts which emerged from it produced a chronicle tradition in Javanese verse known as *babad*, part mythical, part prophetic and instructional, and part genealogy. The central narratives designed to show the legitimacy of kings descending through the complex upheavals of the wars of Islamization are collectively known as the *Babad Tanah Jawi*. All modern historians, beginning with Raffles, have drawn on this set of traditions, although the task of editing a canonical text has so far defeated modern scholarship. The most chronologically accurate of the early Javanese histories, the Kartasura verse chronicle *Babad Ing Sangkala*, was written in 1738 but thought to incorporate a particularly accurate text originally compiled before 1670.

A particularly interesting school of historical writing developed in South Sulawesi (Celebes) in the Makasar and Bugis languages. The written syllabary was only introduced (probably from southern Sumatra) in the fifteenth century, and appears to have been especially valued for the recording of aristocratic genealogies. Marriage alliances among the numerous Bugis and Makasar states required a careful chronicle of previous marriages and begettings going back to the supernatural origins of the leading families. From around 1600, genealogies were developed into state chronicles at Makasar, then a major maritime power much influenced by Islamic and European literary models. From the 1620s, this chronicle tradition could draw on court diaries, which recorded royal births, deaths, and marriages, as well as notable foreign visitors, wars, and innovations. Keenly aware of chronology, the state histories adopt a critical tone towards their sources, a distinctly linear notion of the progress of the kingdom through various innovations, and a desire to learn from failures as well as successes. The Goa (Makasar) chronicle began with an explanation of why it was important that future generations remember the past, 'to avoid the twin dangers, either of our thinking that we were all great kings, or of others alleging that we were worthless.'

2. *An Autonomous Modernity?*

While the seventeenth century witnessed remarkable new departures in historical writing, there was a further development in the late eighteenth century of more abundant historical writing in a self-conscious, secular, and historically careful pattern in several Southeast Asian traditions. While the better survival of manuscripts written after the wars of the eighteenth century may be one explanation, there also appears to have been a creative sense of crisis among Tai and Vietnamese writers in particular which gave rise to reflection on the past. An early sign of this trend was U Kala's *Great Chronicle of Kings* (*Mahazawingyi*), the first comprehensive and chronologically reliable Burmese chronicle, written in the 1720s. In the late eighteenth century this was incorporated into the first history which might be called national, the 'New Chronicle of Myanma' (*Myanma Yazawinthit*), also the first to use its sources in a critical sense.

Among Vietnamese, the new trend was embodied by the scholar-official Le Quy Don (around 1726–1784), who wrote a Le Dynasty history in biographical form,

often also referred to as a national history—*The Complete History of Dai Viet* (*Dai Viet Thong Su*). His encyclopedic *Miscellaneous Records of Pacification in the Border Area* (*Phu Bien Tap Luc*) is a careful accumulation of primary sources relating to technological, economic, cultural, and political developments in the southern Vietnamese kingdom (known to foreigners as Cochinchina) which was a rival of his own. The Tai kingdoms endured a mauling from Burma in the late eighteenth century, and virtually all responded in the period 1790 to 1830 by carefully reconstructing their histories with a view to avoiding such disasters in the future. In Java for a similar reason there was a spate of chronicle-writing after a long period of warfare ended with the peace of Gianti in 1755.

3. *European Influences*

These elements of modernity all preceded any direct European influence on the writing of history. The nineteenth century, however, gave rise to numerous writers who were in contact with European scholar-officials with their own agendas for gathering histories of Southeast Asian societies. Among the earliest of these were writers in Malay who cooperated in the quest for manuscripts of English and Dutch officials. Munshi Abdullah bin Abdul-Kadir (1797–1854) was a Malay writer for Stamford Raffles, and used his understanding of British mores to launch a critique of the Malay society of the sultanates. In his vivid autobiographical works, *Hikayat Abdullah* (1849) and *Hikayat Pelayaran Abdullah* (Chronicle of Abdullah's Voyages), he introduced such concepts as progress, race, freedom and the rule of law into Malay historical writing. The Riau court writer Raja Ali Haji (around 1809–1870) was in regular contact with the Dutch Malay scholars of his day and learned something of their perspective on history. The spirit of his two outstanding histories, *Genealogy of Malay and Bugis Princes* (*Silsilah Melayu dan Bugis dan sekalian raja-raja-nya*, 1865) and *The Precious Gift* (*Tuhfat al-Nafis*, around 1866) is nevertheless heavily didactic in the best Islamic tradition.

European writer-officials brought to the study of Southeast Asia an approach often extremely skeptical of chronicles, and more concerned with the written records of European travelers and the epigraphic and archaeological evidence from the early states. Beginning with Stamford Raffles' *History of Java* (1817) there was a fascination with the ruins of vanished kingdoms, and a tendency to find their successors in a state of decline which could only be reversed by European rule. The Edinburgh-trained doctor John Crawfurd (1783–1868) deserves particular attention as an historian and encyclopedist of Southeast Asia as a whole. His *History of the Indian Archipelago* (3 Vols., 1820), followed by the *Descriptive Dictionary of the Indian Islands and Adjacent Countries* (1856) covered almost the whole region, based on first-hand knowledge of Java and Singapore as an administrator, and the official missions he led to Burma, Siam, and Vietnam in the 1820s, each of which yielded published reports which began the rendering into English of the chronicle traditions of these countries.

The Philippines was the area best endowed with early European writings because of the prominence in the colonization of the islands of Spanish religious orders, each of which produced chronicles. Among the best of these, containing much material about the pre-Hispanic Philippines, were those of the Franciscan Marcelo de Ribadneyra (1601), the Jesuit Pedro Quirino (1604), and the Dominican Diego Aduarte (1640). An outstanding later example of the genre was the 14-volume chronicle of Spanish and missionary activity in the islands by the Recollect Juan de la Concepcion (1788–1792). As nineteenth-century Spanish accounts became ever more negative about the achievements of Filipinos, the first generation of nationalists turned for inspiration to the oldest of the Spanish secular writers, Antonio de Morga, whose *Sucesos de las Islas Filipinas* (1609) received its first modern edition in English in 1868. The intellectual pioneer of Philippine nationalism, Jose Rizal, published in Paris a new Spanish edition of the work in 1890, with the professed intention 'to awaken your [Filipino] consciousness of our past, already effaced from your memory, and to rectify what has been falsified and slandered.'

As Southeast Asia became partitioned among the colonial powers in the second half of the nineteenth century, the more talented colonial officials pioneered modern-style 'national' histories in English, French, and Dutch. These usually combined some of the chronicle record of ancient kingdoms, the interventions of foreigners, and the rise of colonial power within the borders later marked out by the colonial powers. Sir Arthur Phayre, an army officer who spent the years 1835 to 1867 in key posts in Burma, was the first of these with his *History of Burma* (1883), followed in 1926 by a book of the same name by G. E. Harvey. Adhémard Leclère wrote a *Histoire du Cambodge* in 1914, some time before the more demanding equivalent for Vietnam, *Histoire Moderne du Pays d'Annam* (*1592–1820*) (1920), by Charles Maybon. A history of Siam was first published in French on the basis of missionary accounts by Turpin in 1771, but the first of the modern European histories was that of the British Consul in Chiang Mai, W. A. R. Wood (1926), relying much on the chronicles.

Sir Richard Winstedt consciously shaped the Malay-language school syllabus with his revealingly-named *Tawarikh Melayu* ('Malay,' not 'Malayan' History, 1918), in collaboration with Daing Abdul Hamid. Winstedt's English-language *Malaya and its History* began its many editions in only 1948, long after the more blatantly imperial *British Malaya* of Sir

Frank Swettenham (1906). The best of the early Dutch writing on the archipelago was devoted to particular kingdoms and culture areas, particularly Java, though the encyclopedic *Oud en Nieuw Oost-Indien* (8 Vols., 1724–1726) of the protestant pastor Francois Valentijn ranged widely. Only with the need for text-books in Netherlands-Indian schools was there a coherent set of 'national' textbooks written, of which the most influential work was the Dutch-triumphalist *Geschiedenis van Nederlandsch Indie* (1930) of F. W. Stapel.

4. Nationalist History

All of Southeast Asia except Siam/Thailand was under Western colonial control from 1900 to 1941, and was ruled by the Japanese for nearly four turbulent years (1942–1945). The sudden Japanese surrender in August 1945 inaugurated a period of revolutionary struggles in the successor states, with relative peace imposed on Indonesia only around 1966, Vietnam after 1975, and Cambodia in the 1990s. The first half of the twentieth century was a period of extraordinary debate and rethinking of the meaning of history, as a Western-educated generation increasingly detached from cultural connections to the past developed new ways of using it for their modernizing purposes. First Social Darwinism and then Marxism were seized upon as means to explain what had to be done to save the nation from colonial humiliation. By the 1950s, a nationalist, anticolonial historiography was dominant everywhere except Malaya/Malaysia and Thailand, turning militant opponents of the colonial order into the most exemplary heroes. Marxist categories were selectively adopted, often to devalue the precolonial past as generically 'feudal.' Debates remained intense between communist, religious, and ethno-nationalist readings, and often between local and national perspectives.

In Indonesia, nationalist writers suddenly occupied center stage when the Japanese occupation removed all Dutch textbooks. Muhammad Yamin (1903–1962), a Sumatran convert to a Java-centric view of nationalist history, became the most influential writer and patron of a nationalist history which glorified Sriwijaya and Majapahit as precursors of modern Indonesia. His work on the fourteenth-century Javanese kingdom of Majapahit, *Gajah Mada* (1945) and *Tatanegara Majapahit* (3 Vols., 1962), has long-term significance.

In Thailand a similarly influential figure through the most nationalist phase was Wichit Wathakan (1898–1962), a commoner who became an immensely popular writer of historical novels and plays, as well as a 12-volume world history (1931). After the anti-monarchy revolution of 1932, he replaced Prince Damrong Rajanubhab (1862–1943) as the government's arbiter of culture and history. Where Damrong had chronicled a Thai identity developed from the monarchy's gradual adoption of liberal modern values, Wichit glorified the purity and heroism of the Thai race.

In Vietnam the debate between traditionalists, nationalists, and Marxists centered around the evaluation of the Nguyen dynasty which came to power with some French assistance in 1802, and of the Tay Son rebellion which preceded it in the late eighteenth century. Tran Trong Kim (1883–1945), author of the first modern Vietnamese history (1928, constantly reprinted until the 1960s) and briefly Prime Minister under the Japanese, broke with the dynastic chroniclers in accepting the Tay Son leader Quang Trung as a national hero. Nevertheless he accepted the first Nguyen emperor Gia Long, who defeated Quang Trung's forces, as a dynastic founder and unifier. Dao Duy Anh's *Short History of Vietnamese Culture* (1938) was the first and perhaps best of the Marxist-influenced studies which saw Vietnamese culture as constantly evolving through social forces. For later Marxists, the Tay Son were heroic peasant rebels, and the last dynasty a disaster which failed to adapt to historical conditions.

5. Historiography of Southeast Asia as a Whole

These nationalist preoccupations allowed little incentive for indigenous historians to enquire into larger Southeast Asian historical patterns except through the Marxist mold. One of the few exceptions is Nguyen Van Huyen (1908–1975), later Education Minister in the Hanoi government. As a student and Vietnamese teacher in Paris he discovered Dutch scholarship on Indonesia, and used it to good effect in his study of common patterns of housebuilding in the region—*Introduction à l'Etude de l'Habitation sur Pilotis dans l'Asie du Sud-est* (1934).

For the most part it was European scholars interested in the precolonial history of the region who found a broader regional perspective useful. German-speaking prehistorians like Heger and Heine-Geldern were the first to use the term in describing the presence of bronze-age Dongson drums throughout the region, or similarities in temple architecture. French scholarship of the 1920s and 1930s began to uncover the relations between the ancient kingdoms of Angkor, Champa, Pagan, and Java. George Coedès (1886–1969) first brought this material together in a path-breaking book he entitled *Histoire Ancienne des États Hindouisés d'Extrême-Orient* (Hanoi, 1944), though better known in its English translation (1968). Coedès broke with the 'Greater India' school which regarded the Southeast Asian kingdoms as Indian colonies, but continued to see Indian agency through visiting traders and scholars. Later historians such as Paul Mus, J. C. van Leur, Bernard Groslier, Ian Mabbett, Oliver

Wolters, and Claude Jacques, and still more the nationalists, were inclined to see Southeast Asians borrowing selectively and creatively what they found useful in Indian civilization, and even sharing in the creation of that civilization.

Histories of the whole region began to appear in English in the 1950s, the most influential being that of D. G. E. Hall (1955), who came to the first British chair of Southeast Asian History, that of the school of African and Oriental Studies (SOAS) in 1949, after an early career (1921–1934) at the University of Rangoon. The other colonial seedbed of historians of the region was the University of Malaya, opened in Singapore in 1949 and eventually spawning a range of institutions in independent Malaysia and Singapore. Such regional historians as Northcote Parkinson, C. D. Cowan, John Bastin, Wang Gungwu, Alistair Lamb, David Bassett, Jan Pluvier, Anthony Reid, Leonard Andaya, and Barbara Andaya learned their craft in these institutions.

The US borrowed from the British tradition initially, with first D. G. E. Hall and then Oliver Wolters pioneering the teaching of the subject at Cornell University. The first generation of American-trained historians, generally studying the dynamic of diverse nationalisms, were less inclined to perceive the region as a whole. The most important US history of the region to date, *In Search of Southeast Asia* (Sleinberg 1971), required the collaboration of six country specialists and is basically divided by country. In contrast, the *Cambridge History of Southeast Asia*, (Tarling 1992) took the heroic line that each chapter should be chronologically and thematically demarcated, but cover the whole region. This ambitious experiment was dominated by historians based in Australia, New Zealand, and Southeast Asia, reflecting a shift which had occurred in the previous two decades.

The approach of Fernand Braudel and the Annales School was brought to Southeast Asia in the 1980s in relation to the Early Modern, a period which had until then been largely abandoned to the 'European expansion' school. Denys Lombard (1936–1998), a colleague of Braudel and founder of the interdisciplinary journal *Archipel*, developed in *Le Carrefour Javanais: Essai d'Histoire Globale* (3 Vols., 1990) the kind of 'total history' pioneered by Braudel. Anthony Reid (b. 1939) found both the broad interdisciplinarity and the ocean-centered focus of Braudel useful in giving coherence to his *Southeast Asia in the Age of Commerce, c.1450–1680* (2 Vols., 1988–1993).

In global terms, the practice of history in Southeast Asia since 1950 has been relatively weak in economic and institutional history, and has produced few outstanding biographies or intellectual histories in the strict sense. Its strengths have been its interdisciplinarity and its deep concern to explicate cultural diversity and autonomy. The two Southeast Asianists who have most influenced historians were trained in other disciplines. The cultural anthropologist Clifford Geertz (b. 1926) defined his task as 'to uncover the conceptual structures that inform our subjects' acts,' and likened it to 'trying to read (in the sense of 'construct a reading of') a manuscript—foreign, faded, full of ellipses, incoherencies, suspicious emendations, and tendentious commentaries.' Though history is never absent from his numerous studies of Java and Bali, it is most salient in *Negara: The Theatre State in Nineteenth Century Bali* (1980), which influentially argued that elaborate royal rituals were the defining acts of Bali states, not simply their decoration or arcane cosmology.

Anglo-Irish Benedict Anderson (b. 1936) made his career at Cornell University, where his brand of culturally-informed political history had more influence on historians than on his own discipline of political science. His pathbreaking *Imagined Communities: Reflections on the Origins and Spread of Nationalism* (1983) was preceded by a series of more specific studies on Javanese and Indonesian political culture, and the influence of the Japanese occupation on them. In the 1980s, deeply alienated from Suharto's Indonesia, he produced revisionist essays which placed the histories of Thailand and the Philippines in a broader political economy perspective. Among the able historians trained at Cornell, Michigan and elsewhere were a number of Southeast Asians who had an impact in their own countries as well as internationally. Reynaldo Ileto (b. 1946) pioneered a new kind of history from below by beautifully demonstrating how popular Christianity provided the idiom of resistance in his *Pasyon and Revolution: Popular Movements in the Philippines* (1979). Nidhi Aeusrivongse wrote almost entirely in Thai after returning to teach in Chiang Mai, but his revisionism has shifted Thai historiography away from the agency of kings and Westerners to the social basis of cultural change.

See also: Area and International Studies: Development in Southeast Asia; Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: East Asia; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Islam: Asia; South Asian Studies: Culture; South Asian Studies: Politics; Southeast Asia, Archaeology of; Southeast Asia: Sociocultural Aspects

Bibliography

Coedès G 1968 *The Indianized States of Southeast Asia*. University of Hawaii Press, Honolulu, HI

Cowan C D, Wolters O W (eds.) 1976 *Southeast Asian History and Historiography: Essays Presented to D. G. E. Hall*. Cornell University Press, Ithaca, NY

Hall D G E 1955 *A History of South-East Asia*. Macmillan, London

Hall D G E (ed.) 1961 *Historians of South East Asia*. Oxford University Press, London
Reid A, Marr D (eds.) 1979 *Perceptions of the Past in Southeast Asia*. Asian Studies Association of Australia/Heinemann, Singapore
Steinberg D J (ed.) 1971 *In Search of Southeast Asia: A Modern History*. Praeger, New York
Tarling N (ed.) 1992 *The Cambridge History of Southeast Asia*, 2 vols. Cambridge University Press, Cambridge, UK
Woolf D R (ed.) 1998 *A Global Encyclopedia of Historical Writing*. Garland, New York/London, 2 Vols.
Wyatt D K, Woodside A (eds.) 1982 *Moral Order and the Question of Change: Essays on Southeast Asian Thought*. Yale University Southeast Asia Studies, CT

A. Reid

Historiography and Historical Thought: Sub-Saharan Africa

The discovery of Africa as an object of scholarly research is part of a wider expansion of academic pursuit. It coincided with the emergence of women's studies. The focus of investigation shifted to the underdogs, including slaves, colonized nations, and other silent majorities of our day as well as the Middle Ages, in Africa, and elsewhere (Wachtel 1971, Clancy 1979, Ginzburg 1980, Feierman 1993).

This article concentrates on sub-Saharan Africa. It covers recent origins of written history, the 'discovery' and valuation of its oral traditions, the interaction of anthropology and history, the Anglo-American debate, and the role of the diaspora.

1. *Origins of History*

The first Sahel states emerged on the northern fringe of the savanna: Ghana (eighth to the eleventh century), Mali (thirteenth to fourteenth century), and Songhay (twelfth to sixteenth century).

Arab authors were the most knowledgeable on Africa's past. Those who made the largest contributions to reconstructing the history of Africa, and western and central Africa in particular, include al-Bakri (1029–1094), al-Idrisi (1154), al'Umari (1301–1349), and Hassan Ibn Mohammed al-Wuzza'n, known as Leo the African (1494–1552). Ibn Khaldun (1332–1406) is unquestionably the 'Father of History,' even though Ibn Battuta (1304–1369) remains the best thanks to his personal accounts and the detail in his descriptions (Levtzion and Hopkins 1981).

Written European sources followed, beginning in the fourteenth century. Their authors were Portuguese navigators. These documents preceded the substantial addition made by codification or oral records in Arabic by Mahmoûd Kâti and Es-Sa'di, both Timbuktu scholars.

In the course of the ninteenth and twentieth centuries, accounts written by Africans in indigenous languages have abounded. A large share of these histories was in Kiswahili, Hausa, Fulfulde, Kanembu, Diula, and Malagasy.

The first professional historians, caught in the Hegelian trap, discounted the history of Africa (Neale 1986, Jewsiewicki and Mudimbe 1993, p. 1). However, the continent and its 'people with no history' were deemed worthy of a present. Ethnography was called upon (Jewsiewicki and Mudimbe 1993, p. 8) to justify the backwardness of the region. Underdeveloped Africa was expected to make do with a 'prehistory' (Moniot 1991, p. 197).

The International African Institute was established in London in 1928. The School of Oriental and African Studies began in 1948 to teach a new subject called history of Africa (Vansina 1992, p. 77). An overdue offspring of academic research was born. Curiously, African universities in French colonies, the local intelligentsia, and their backers in France did not see any need for the discipline (Vansina 1992, Moniot 1995, p. 647).

2. *History and Anthropology*

Almost eight years after publication of *African Historiographies* (Jewsiewicki and Newbury 1986), Jewsiewicki and Mudimbe published *History Making in Africa* (1993). The majority of the authors who contributed to the former work were historians. Anthropologists dominated the latter. This shift in the weight accorded to the two disciplines is a sign of the times. When focused on Africa, anthropology and history do not fit neatly into the university compartmentalization scheme. Rebellious grassroots reality has gradually overcome the rigidity that had governed the allocation of professorships.

As time passed most of the convenient cubbyholes have been banned to the realm of memory. It was once the fashion to contrast societies with no history, acephalous, precapitalist societies, with societies blessed with a state and a history culminating in capitalism. The first group was viewed as inherently handicapped by communalism, polytheism, oralism, and tribalism. The second group of nations was endowed with the virtues of individualism, monotheism, literacy, and polity, to name just a few. Criticism of the concepts of tribe and ethnicity took two different approaches, a French school and an Anglo-Saxon school (Coquerel 1983, Young 1986, Chrétien and Prunier 1989, Barber and Moraes Farias 1990, Coplan 1991, 1993, van Jaarsveld 1992, Harries 1993, Jenkins 1997.

The adoption of precolonial Africa as a concept for a catchall analysis of the continent marked a milestone. This approach created the illusion of a different, idealized world. Postcolonial Africa was viewed as the child of a past to which the umbilical cord had been severed. This is the country cherished by the nationalists during their struggle for independence and the image invoked by politicians. Cheikh Anta Diop was fond of climbing the continent's family tree back to its glorious Egyptian ancestors, while Joseph Ki-Zerbo (1972) and J. F. Ade Ajayi convey us to a not so distant past, labeled the precolonial era (Mudimbe and Jewsiewicki 1993, pp. 1–2). Historians responsible for this breakdown into periods have changed their minds and now support the anthropologists' view of the existence of a continuum between the colonial societies and their worthy successor, sovereign Africa (Bayart 1989, Piault 1987).

The focus of historians on the past and the concentration of ethnologists on the present explain the farsightedness of the former and the shortsightedness of the latter. The outlook of these two isolated disciplines changes radically as soon as historians discover humanity as the center of their field of investigation and anthropologists take account of time, the framework for human behavior, and activity.

3. *The Silent Make Themselves Heard*

Africa is known for its chanted, sung, or simply recounted oral history. The *griots*, professional chroniclers of West Africa, the *biiru*, specialists in Rwandan royal ritual, ordinary storytellers in the cities and the country have provided a wealth of material to researchers accustomed to studying written material.

In 1962, Jan Vansina published his authoritative work, *De la tradition orale*. His name is invariably associated with the acceptance of oral accounts as a source of African history and with the elaboration of a rigorous scientific method. It took historians, or at least part of their fraternity, a long time to learn to appreciate the 'auriture' of oral genres (Coplan 1991, 1993, p. 8).

Zealous pursuit of sources engenders such excesses as equating the death of an old timer with the gutting of a library. This attitude refers us to a kind of precolonial library, to paraphrase Mudimbe. It is also called the memory of the griots and 'oral archives' (Bazin 1979, p. 450). Writing history Western style entails the fault of reducing documents recorded by interviewers to simple sources. This is denounced by Jewsiewicki (1986, p. 16) and Schoenbrun (1993, p. 55).

An increasing hue and cry (Farias 1992) has been raised in opposition to this tendency. The critics are anxious to upgrade the role of reputedly passive traditionists (Ba 1980). While David W. Cohen (1989, p. 12) insists on the contrast between material provided by the Africanist community and that produced by Africans who collect data on a daily basis, Jan Vansina was an early advocate of another reading of these documents, as revealed in a text of Jewsiewicki (1979, p. 81).

As soon as we escape the pull of the two fetishes constituted by the roving reporter and the Western-style historiographer, we are attracted by other types of history, and vice versa. We can then pay heed to nonformalized oral texts, unofficial samples taken from lower rungs of the social ladder. Their particular form and their non-verbal character require another method of analysis, yielding a more fertile approach to Africanist research (Olivier de Sardan 1976, Diawara 1990, Barber 1991). The field is further enriched by the involvement of anthropologists who tap the upstream end of oral production and characterize its interface with society (*surface sociale*). Their tools bridge a major gap neglected by historical studies which concentrate entirely on the oral text and lose sight of the life, environment, and experiences of participants in society (Moniot 1986, p. 58).

The history of the continent has become a professional endeavor. It involves 'interrogating and revealing the historian's operation' (Moniot 1995, p. 647). From now on, the Africanists can forget the 'automatic legitimation of being specialists in a small exotic field' (Cooper 1995, p. 236).

4. *The Anglo-American Debate*

Vansina (1992, p. 91) takes advantage of the lessons of historiography consisting in recognizing the past in the present and vice versa, and aims his harshest criticism at the postmodern drift of David W. Cohen. The latter (1994, p. XV) stresses the fact that historical knowledge is not the monopoly of poetry or formalized narratives but that it is constantly being proclaimed and cited. Henige (1995, p. 314) vehemently attacks what he calls Cohen's 'predilection […] for subordinating the substance of sources to their interpretation'.

5. *The Diaspora Includes Africa; in Fact, it Is Africa*

What position do African specialists take in this controversy, which almost boils down to an 'American-American' debate? Jewsiewicki (1986, p. 9) makes a statement that has lost none of its pertinence for historians and other Africanists. 'The Africanist historian and the African historian share neither the same responsibilities nor the same existential constraints' (cf. Vansina 1994, p. 221). This is certainly true, except that another schism has split the African scholars' camp, divided between a diaspora and a home com-

munity, and occasionally bridged by a coming and going between the two shores. Africans abroad still focus more on debates that are often kindled overseas, in America or Europe.

Nevertheless, practically everything written is in the language of academics, which is not the idiom of indigenous populations. A UNESCO policy to remedy this situation, adopted in the 1970s, remains a pious vow even though issues of *l'Histoire Générale de l'Afrique* have appeared in Kiswahili and Hausa. Anyway, who is able to read them? Where can you get a copy and what would it cost? Does the body of literature constitute history for the average person? The question posed by the *African Historiographies* subtitle, 'What history for which Africa?' is as relevant as ever. The role of Africans in this international forum appears to be rather marginal. The empire of Western historiography is bigger than ever (Vansina 1994, p. 220). The hopes raised by studies in the English-speaking world, which benefits from the North American alternative to the intellectual heirs of the British Empire and from a more substantial organization (Jewsiewicki 1986, p. 15), continue to fade. In the meantime, the New World option has become an attractive outlet for the brain drain from tropical dictatorships. Zaire has forfeited its chances to develop a school of thought liberated from the domination of Brussels. The ex-colonies reproduce the models elaborated in the former centers of colonial power (Chrétien 1997) despite the fact that English-speaking and French-speaking African authors have begun to read one another's works (Coquery-Vidrovitch 1997). Their marginality becomes even more pronounced as the tendency to consume (È Nzièm 1986, p. 25) concepts developed outside Africa grows. It makes little difference whether they are the products of non-Africans or of Africans of the diaspora, which never stops expanding (Jewsiewicki 1979, p. 81).

Intellectual dependence resembling the economic impasse of the continent is worsening. While additions to the circle of recognized Africanists are becoming rare, a heavy storm of productivity is brewing in Africa, as far as reading about and interpreting the past are concerned. Student theses, cassettes recorded by individuals in different villages, radio stations, and other relevant institutions are mushrooming. The harvest has never been so bountiful (Cohen 1994, p. XV, Diawara 1997).

This analysis is not intended to deny the essential, unanticipated impact of African studies on social sciences in general. The 'dissolution of world history' by the history of Africa is a reality (Feierman 1993). In evaluating African studies, one is inclined to dismiss the crucial role played by the diaspora, which actively propagates its share of the African experience (Macamo 1999). Expatriates are deprived of their birthright by attributing any success they achieve to their severance of ties to their dismal native lands in the tropics. Mudimbe and Appiah (1993, p. 113), referring to the impact of African studies on American disciplines, cite the necessity 'to locate ourselves as much in the current history of those disciplines as in the current state of African Studies […].'

In fact, a wider influence of African studies on intellectual life is detected—in Western Europe and elsewhere as well as in America (Feierman 1993, p. 200, note 13). The impact of Africanist historiography on that of France, the former métropole, is undeniable (Coquery-Vidrovitch and Jewsiewicki 1986, pp. 143–44).

A glimpse at what is on the horizon suggests the tide is turning, the groundswell stemming from the unprecedented creativity of local bards, both griots and other producers of historical material. This resurgence is amplified by a foreign legion, which gives a boost to those who have remained on their native soil. Hope is raised by the fact that disciplines applied to Africa are inevitably linked to academic trends in Europe and the US. No longer cramped by a narrow view of a particular field of study, Africanists have demonstrated the benefits of multidisciplinary collaboration. In addition, Africanists have been instrumental in fostering a dialog and creating an interface between traditional academic approaches in Anglo-Saxon and French-speaking countries.

See also: African Studies: History; Anthropology, History of; Historiography and Historical Thought: Christian Tradition; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Society and History; Oral History; Urban Ethnography

Bibliography

Ba A H 1980 La tradition vivante. In: *Histoire Générale de l'Afrique*. UNESCO, Vol. 1, pp. 191–230, Paris

Barber K 1991 *I Could Speak Until Tomorrow. Oriki Women and the Past in a Yoruba Town*. Edinburgh University Press, Edinburgh, UK

Barber K, Farias M, Fernando de P (eds.) 1990 *Self-Assertion and Brokerage. Early cultural Nationalism in West Africa*. Birmingham University African Studies, Series 2, Birmingham

Bayart J-F 1989 *L'État en Afrique. La Politique du Ventre*. Fayard, Paris

Bazin J 1979 La production d'un récit historique. *Cahiers d'Études Africaines* 73–76. **19**(1–4): 435–83

Chrétien J-P 1997 Une crise de l'histoire de l'Afrique en langue française. *Politique Africaine* **68**: 141–49

Chrétien J-P, Prunier G (eds.) 1989 *Les Ethnies ont une Histoire*. Karthala-ACCT, Paris

Cissé Y T, Kamissoko W 1988 *La Grande Geste du Mali des Origines à la Fondation de l'Empire*. Karthala-Arsan, Paris

Cissé Y T, Kamissoko W 1991 *Soundjata, la gloire du Mali*. Arsan Karthala, Paris

Clancy M T 1979 *From Memory to Written Record. England* 1066–1307. Arnold, London

Cohen D W 1989 The undefining of oral tradition. *Ethnohistory* **36**: 9–17

Cohen D W 1994 *The Combing of History*. University of Chicago Press, Chicago
Cooper F 1995 Work, class and empire: An African historian's retrospective on E. P. Thompson. In: *Social History*, pp. 235–41
Coplan D B 1991 Fictions that save: Migrants' performance and Basotho national culture. *Cultural Anthropology* **6**: 164–91
Coplan D B 1993 History in Eaten Whole: Consuming Trapes in Sesotho Auriture. In: Mudimbe V Y, Jewsiewicki B (eds.) *History and Theory*, Beiheft 32, Wesleyan University, Middletown, CT, pp. 80–104
Coquerel P 1988 *Le verouillage ethnique en Afrique du Sud*. UNESCO-OUA
Coquerel P 1989 Les Zulu dans l'Afrique du Sud contemporaine. In: Chrétien J-P, Prunier G (eds.) *Les Ethnies ont une Histoire*. Karthala-ACCT, Paris, pp. 417–25
Coquery-Vidrovitch C 1997 Réflexion comparée sur l'historiographie africaniste de langue française et anglaise. *Politique Africaine* **66**: 91–100
Coquery-Vidrovitch C, Jewsiewicki B 1986 Africanist historiography in France and Belgium. Traditions and trends. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 139–50
Cuoq J 1975 *Recueil des sources arabes concernant l'Afrique occidentale du VIIe au XVIe siècle* (*Bilâd al-Sûdan*). Editions du CNRS, Paris
Diawara M 1989 Women, servitude and history: The oral historical traditions of women of servile condition in the kingdom of Jaara (Mali) from the fifteenth to the mid-nineteenth century. In: Barber K, Farias P F de Moraes (eds.) *Discourse and its Disguises. The Interpretation of African Oral Texts*, Centre of West African Studies, Birmingham, pp. 109–37
Diawara M 1990 *La Graine de la Parole*. Franz Steiner Verlag, Stuttgart, Germany
Diawara M 1994 Der Blick vom anderen Ufer. Oder: Die Entdeckung des Weissen. In: Schlumberger J A, Segl P (eds.) *Europa—aber was ist es? Aspekte seiner Identität in interdisziplinärer Sicht*. Böhlan, Cologne, Germany, pp. 255–83
Diawara M 1997 Über die Bedeutung der Arbeit in der Welt der Mande. In: Spittler G, Beck K (eds.) *Arbeit in Afrika*, LIT, Munster, pp. 49–68
Diop C A 1960 *L'Afrique Noire Précoloniale*. Présence Africaine, Paris
Diop C A 1967 *Antériorité des Civilisations Nègres: Mythe ou Vérité Historique?* Présence Africaine, Paris
Duby G, Lardreau G 1980 *Dialogues*. Flammarion, Paris
È Nzièm N 1986 African historians and Africanist historians. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 20–7
Farias P F M 1992 The oral traditionist as critic and intellectual producer: An example from contemporary Mali. In: Falola T (ed.) *African Historiography—Essays in Honour of Jacob Ade Ajayi*. Longman, London, pp. 20–44
Feiermann S 1993 African histories and the dissolution of world history. In: Bates R H, M V Y, O'Barr J (eds.) *Africa and the Disciplines*. The University of Chicago Press, Chicago, pp. 167–212
Ginzburg C 1980 *Cheese and the Worms. The Cosmos of a Sixteenth-century Miller*. Routledge & Kegan Paul, London
Harries P 1993 Imagery, symbolism and tradition in a South African Bantustan: Mangusuthu Buthelezi, Inkatha, and Zulu history. In: Mudimbe V Y, Jewsiewicki B (eds.) *History Making in Africa, History and Theory*. Beiheft 32, Wesleyan University, Middletown, CT pp. 105–25
Henige D 1995 Homphaloskepsis and infantilizing of history. *Journal of African History* **36**: 311–18
Jaaresveld F A 1992 Recent Afrikaner historiography. *Itinerario* **16**(1): 93–106
Jenkins R 1997 *Rethinking Ethnicity: Arguments and Explorations*. Sage, London
Jewsiewicki B 1979 L'histoire en Afrique ou le commerce des idées usagées. *Revue Canadiénne des Études africaines*. **13**(1–2): 69–87
Jewsiewicki B 1986 Introduction. One historiography or several. A requiem for Africanism. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 9–17
Jewsiewicki B 1991 Le primitivisme, le postcolonialisme, les antiquités 'nègres' et la question nationale. *Cahiers d'Études Africaines* 121–122, **31**(1–2): 191–213
Jewsiewicki B, Newbury D (eds.) 1986 *African Historiographies. What History for Which Africa?* Sage, London
Ki-Zerbo J 1972 *Histoire de l'Afrique Noire: d'Hier à demain*. Hatier, Paris
Levtzion N 1973 *Ancient Ghana and Mali*. Methuen, London
Levtzion N, Hopkins J F P 1981 *Corpus of Early Arabic Sources for West African History*. Cambridge University Press, Cambridge, UK
Macamo E S 1999 *Was ist Afrika? Zur Geschichte und Kultursoziologie eines modernen Konstrukts*. Duncker & Humblot, Berlin
McIntosh R J, McIntosh S K 1979 Terra-cotta statuettes from Mali. *African Art* **12**(2): pp. 52–3, 91
Miller J 1980a Listening for the African past. In: Miller J (ed.) *The African Past Speaks*. Dawson-Archon, Folkstone, UK, pp. 1–59
Miller J (ed.) 1980b *The African Past Speaks*. Dawson-Archon, Folkstone, UK
Moniot H 1970 Les sources de l'histoire africaine. In: Deschamps H (ed.) *Histoire Générale de l'Afrique*. T I, PUF, Paris, pp. 123–47
Moniot H 1986 Profile of a historiography. Oral tradition and historical research in Africa. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 50–8
Moniot H 1991 L'histoire africaine, un observatoire de la periodisation *Périodes: la construction du temps historique. Actes du Ve colloque d'histoire du temps présent*. Éditions de l'École des Hautes Études en Sciences Sociales et Histoire au Présent, Paris
Moniot H 1995 L'histoire à l'épreuve de l'Afrique. *Cahiers d'Études Africaines* 138–139, **35**(2–3): 647–56
Mudimbe V Y, Appiah K A 1993 The impact of African studies on philosophy. In: Bates R H, Mudimbe V Y, O'Barr J (eds.) *Africa and the Disciplines*. The University of Chicago Press, Chicago, pp. 113–35
Mudimbe V Y, Jewsiewicki B 1993 Africans' memories and contemporary history of Africa. In: Mudimbe V Y, Jewsiewicki B (eds.) *History Making in Africa, History and Theory*. Beiheft 32, Wesleyan University, Middletown, CT, pp. 1–11
Neale C 1986 The idea of progress in the revision of African History, 1960–1970. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 112–22
Olivier de Sardan J-P 1976 *Quand nos Pères Etaient Captifs*, Paris

Piault M (ed.) 1987 *La Colonisation, Rupture ou Parenthèse*, Paris
Shoenbrun D L 1993 A past whose time has come: Historical context and history in Eastern Africa's Great Lakes. In: Mudimbe V Y, Jewsiewicki B (eds.) *History Making in Africa, History and Theory*. Beiheft 32, Wesleyan University, Middletown, CT pp. 32–56
Stock B 1983 *The Implication of Literacy*. Princeton University Press, Princeton, NJ
Vansina J 1962 *De la Tradition Orale*. Musée Royal de l'Afrique Centrale, Annales, Sciences Humaines, No. 36, Tervuren
Vansina J 1985 *Oral Tradition as History*. University of Wisconsin Press, Madison, WI
Vansina J 1992 Some perceptions on the writing of African History: 1948–1992. *Itinerario* **16**(1): 77–91
Vansina J 1994 *Living with Africa*. University of Wisconsin Press, Madison, WI
Wachtel N 1971 *La Vision des Vaincus*. Gallimard, Paris
Young C M 1986 Nationalism, ethnicity and class in Africa: A retrospective. *Cahiers d'Études Africaines* **103**: 421–95

M. Diawara

History and Literature

Literature and history have many things in common, and one of them is the aesthetic form of narration. The many contemporary experiments not withstanding, modern historiography and the modern novel still share the common feature of narration. In narratology, some theoreticians have differentiated between the narrative modes in the two fields; others have shown how they conflate and that they can hardly be distinguished. This article wants to show what narration in literature and history has in common and in what respect the two modes differ from each other.

For Aristotle (see Chapter 9 of his *Poetics*) the matter of the division of labor between historiographer and poet was easily decided: the historiographer relates 'what has happened' and the poet 'what might happen'. In contrast to Plato's devastating criticism of literature, Aristotle considered poetry to be of high value. In his eyes, literature dealt with problems of a more general and thus more significant nature, while the historian had to report about specific, and thus, less important matters.

Until approximately 1750, historiography was considered a branch of literature. Since the eighteenth century (when Aristotle's distinctions were rediscovered), the differences between the two disciplines have been stressed, and it was only then that the term 'literature' became synonymous with 'poetry' and 'fiction.' Gottsched and Lessing still shared Aristotle's prejudice against history. For them the field of history was merely a quarry from which they mined some of the raw material they used to build their preconceived literary plots. Of real importance to them was a moralistic story, and examples were taken from the repertoires in the areas of history, legend, and mythology. However, the late eighteenth century is an important turning point. As Koselleck has demonstrated, it is the time of the development of a genuinely new philosophical awareness about history (Koselleck 1976). The concept of a continuously progressing history as such replaces both the religious concept of a Christian teleology as well as the idea of history as a chaotic conglomerate of countless individual stories of an exemplary nature. Through this philosophical development, literature is seen as part of history, and thus it loses its Aristotelian superiority.

Today the relationship between fiction and historiography poses a number of hermeneutic questions, and there is no modern Aristotle who might offer a simple and acceptable solution fitting for our time. Narration can be viewed as one of the primary tools of knowledge, and it determines the structure and aesthetic form in the presentation of a real or a fictional event. Narration discloses the dimension of historical reality as well as the sphere of poetic fiction. The logic of narration provides access to the logic of historiography and fiction; it makes the structure of both historical consciousness and fantasy tangible. In every narrative, real or fictional events are supported by a structure through which happenings are accentuated, selected, or eliminated according to the law of relevance. Neither in historiographical nor in fictional narration are we dealing with a reproduction or duplication of events, but with a certain organization of actual or imaginary occurrences or experiences.

1. *Narration in Historiography*

Perhaps because it is so difficult to liberate it from its established forms, narration is constantly under attack. In the discussion of modern historiography from 1970–2000, it looked as if 'narration' would be replaced by 'description' and 'explanation.' 'Historiography,' Szondi argued, must 'divest itself increasingly of its narrative character'; narrative historiography must 'turn into description if a new concept of history is to do justice to our knowledge of history as an anonymous process, as a sequence of events and changing systems' (Szondi 1973). The arguments of the antinarrativists can be summarized in the following way: the narrative recording of events in historicism was based on the belief that history was made by individuals. However, this categorization did not adequately take into account the development of social processes. Processes and structures can only be described, not narrated. Thus the old '*histoire événementielle*' has to be replaced by the new '*histoire structurale*.' 'The nouvelle histoire' of French historians in particular, with their descriptions of structures

underlying long periods of time, and their quantifying methods, forced the old paradigm of writing history by narrating events into a defensive position (Furet 1982, LeGoff and Nora 1974).

However, according to some historiographers, history as a narration of events and history as a description of structures are not mutually exclusive. These historiographers concede that history as a narration of events is no longer the center of the historiographer's work and that it no longer provides the obvious support of historical writing. Nevertheless, they also concede that narration is allowed to stand as a possibility alongside the writing of history as a structural analysis of processes (Borst 1976). Other scholars will have nothing to do with a strict separation of the two methods and point out that narration and description are intertwined. The argument is that the fact that modern history is a series of processes, it can only be understood through the alternate explanation of events through structures and vice versa (Koselleck 1976 and Kocka 1984). Although the structures become more accessible through description and events through narration, nevertheless in practice it would be impossible to maintain a demarcation between the two methods. The complementary characters of description and narration are increasingly emphasized (Stone 1979).

The defenders of the descriptive method like to confront the narrators among the historiographers with their lack of theory. They assert that the historiographer must not only narrate but also explain, and that this is only possible by means of a theoretical concept of the realities, as it has long been practiced by social scientists. The proponents of the narrative method in writing history defend themselves vehemently against the charge of being anti-theoretical. From their initially defensive position, they have turned to attack. Danto has asserted that narration is vital for description and explanation; that the form of narration is at the same time a form of explanation. The narrative form always surpasses the given, and categorically uses all-encompassing concepts. Narrative structures influence our historical thinking in the same way theories do in the case of the sciences (Danto 1965). Ricoeur expresses the same idea in his phrase 'to tell a story means to explain it' (Ricoeur 1983). Rüsen agrees that narration is part of the logical structure of historiography and that consequently 'description and explanation have a function subordinate to the narrative scheme' (Baumgartner and Rüsen 1982).

2. *Narration in Literature*

In connection with Gottfried Benn's poetics, Grimm described a type of novel from which the narrator had disappeared. It was called 'the novel of the phenotype' or 'the inward directed novel,' labels indicating a tendency toward lack of action and lack of characters (Grimm 1963). The world and its events are banned here. The reduction of the 'I,' the isolated subject, is significant for this novel: the 'I' speaks only of itself. Consequently, one reads in Rilke's *Aufzeichnungen des Malte Laurids Brigge* from 1910: 'It must have been before my time that stories were narrated, really narrated. I have never heard a person narrate a story.' Benjamin's discussion of the decline of storytelling in his essay *Der Erzähler* about the Russian writer Nikolai Lesskow echoes Rilke's statement (Benjamin 1972).

The aesthetic designation of the phenotype novel—exemplified in the novels of Virginia Woolf, Benn, and Rilke—is also appropriate for the writing of some contemporary authors; the authors of the French '*nouveau roman*' (like Alain Robbe-Grillet) as well as postmodern American writers (like William Gass) and contemporary German authors (like Peter Handke) come to mind. In the novel of the avant-garde (Joyce, Dos Passos, Gide, Döblin, Musil, Broch), narration was either complemented or largely replaced by montage techniques and cognitive means, by reflection or essayistic writing. Here it was not so much a matter of narration of events, action, and mimesis as one of the presentation of ideas and the provision of knowledge (Hinck 1995).

However, the best-known postwar novelists (like Garcia-Marquez, Vargas Llosa, Philip Roth, Bellow, or Grass) uphold traditional established narrative models. One can write purely narrative novels. However, nonnarrative epic fiction is nonexistent. Like historiography, the novel obviously cannot dispense entirely with narration, and thus it makes sense, now as ever, to adhere to the science of narrative technique (Lützeler 1986). In the field of narrative technique, literary scholars are far ahead of the narratologists in the camp of the historiographers. Historians could learn from reader theories as developed by such literary theorists as Jonathan Culler ('the competent reader'), Umberto Eco ('the model reader'), Stanley Fish ('the informed reader'), Wolfgang Iser ('the implied reader'), Michel Riffaterre ('the superreader'), and Gerald Prince ('the narratee') (Culler 1975, Eco 1979, Fish 1970, Iser 1974, Riffaterre 1971, and Prince 1982).

3. *The Conflation of Historiography and Fiction*

In *Metahistory*, Hayden White draws interesting parallels between historiography and fiction. He strives for a poetics of historiography and attempts to describe the forms of historical presentation in order to compare them with literary forms (White 1973). White starts with the premise that no sequence of events—by itself—is a story; it is the task of the

historiographer to transfer the events, the facts, to the narrative framework of a story. The connections between happenings do not occur spontaneously; they are the result of the historiographer's reflection. In each historiographical text, White distinguishes between the 'what' (the narrated events) and the 'how' (the type of narration, the 'emplotment'). He was not the only scholar to see that historiographical and fictional narration are related, but he was the first to make a systematic attempt to study the structure of historical consciousness—to show the fictional elements and their formal manifestations in historiography.

The situation in nineteenth century historiography and philosophy of history serves as an example. White could have given his book the title *The Rhetoric of Historiography* as a parallel to Wayne C. Booth's 1961 study *The Rhetoric of Fiction*. In *Metahistory* he concludes that the writing of history has little to do with such qualities as true or false, correct or incorrect, and that ultimately historical writing cannot be distinguished from fiction. The historian's style, White argues, depends on the combination of the following elements: first, there are three different strategies; he calls them 'explanation by formal argument,' 'explanation by ideological implication,' and 'explanation by emplotment.' Specific modes of articulation correspond to each strategy. In this context we are only interested in the strategy of 'emplotment' (narration) with its four corresponding 'poetic' modes of 'romance,' 'comedy,' 'tragedy,' and 'satire.'

White's references to certain structural similarities between the unfolding of tragedies or comedies and that of historical representations are interesting, but these similarities should by no means be construed as an actual fictionalization of history. Demandt has shown how thought-structures were mirrored in the poetic metaphors used by historiographers, and how, in turn, the ideologies of the historiographers influenced their use of rhetorical figures (Demandt 1978). However, the use of metaphors does not signify a fictionalization of history. There exist basic narrative concepts that are equally valid for historical and for fictional narration. Without action, images, and metaphors, narration *per se* is unimaginable.

Returning to White's example of the different depictions and interpretations of the French Revolution, it is evident that this is not a case of fictionalization because the historical process is not depicted as an 'as if,' but as something that actually did occur. No one will deny that the depiction of history is a creative process. However, its qualitative difference from fictional narration remains, despite the use of literary interpretation patterns and flowery metaphors. While fiction has as its basis the so-called 'as-if' structure, and thus proves to be pretense, illusion, and unreality, historiography always relates to actual events. Hamburger clarified the difference through the example of a historical report about Napoleon and a novel about the French emperor. The historiographer cannot see Napoleon 'in the subjectivity, the ego-originality of his inner thoughts, of his very "existence."' If he does so, however, we 'find ourselves in a novel about Napoleon, a piece of fiction' (Hamburger 1968).

The narrative techniques with which novelists depict the psyches of their characters have been analyzed by Dorrit Cohn. She differentiates between psycho-narration, quoted monologue, and narrated monologue (Cohn 1978). These are narrative techniques reserved for fictional narration, which cannot occur in historical writing. Just as the use of a literary (tragic or comic, for example) interpretive category cannot turn a historical narrative into a fictional one, a historical novel does not become a piece of historiography simply by including historical reality in the fictional field. Käte Hamburger validly defines the line between fiction and reality, 'where there is no crossing over from one category to the other,' when she discusses the different perceptions of mimesis in novel and historiography. According to her, there is a definite relationship between the narrating process and the narrated material in historiography, while characters created by novelists are simply narrated persons. In fiction, there is no relationship, only functional context, between the narrating process and the narrated material. In other words, the fictionalization of the persons described consists in the fact that the characters are not presented as objects but as subjects, as 'ego-origines' (Hamburger 1968).

This difference between historiography and fiction also becomes evident when we look at the results of more recent studies of the theory of functionality. Assmann has shown that scholars in the fields of analytic language philosophy and of pragmatic communication theory insist on this distinction (Assmann 1980). In the analytic philosophy of language the point is made that fiction has a very peculiar status as far as its logic of statements is concerned: the authors of fiction are not obliged to come up with any proof; their judgments are quasi, pseudo, or as-if statements. In communication theory, fictionality is studied with respect to its reception effects: it involves the elimination of the normal connection that exists between the participants of an act of communication. Dealing with fiction leads to a temporary liberation from thinking in terms of fixed institutions; it means the absence of social sanctions and a lack of verification.

4. *The Concepts of Truth*

Historia vero testis temporum, *lux veritas*, *vita memoriae*, *magistra vitae*, *nuntia vetustatis*: Cicero's classic definition of the function of historical writing is still valid today—with a few deviations and reservations. The modern formulations are somewhat less pre-

tentious, but given such terms as 'demand for truth,' 'imparting of meaning,' 'provision of identity,' and 'memory,' the function of historiography is not defined so differently from Cicero's wording. Only the postulate '*magistra vitae*' is viewed more skeptically. Authors of novels are equally concerned with the pursuit of truth, the imparting of meaning, the provision of identity, and memory. Friedrich Schlegel wrote in his introduction to the *Wissenschaft der europäischen Literatur*: 'As history aims at recognition and truth it approaches science, but as it is also representation (…) it is related to art.' Here the double-faced aspect of historiography has been precisely formulated.

No historiographer would be so arrogant as to assume that his or her report could encompass the whole truth of a historical event. Historiography can claim with justification that its statements are 'truthful' insofar as they have been gained 'scientifically,' that is in so far as they can be proven. And they can be proven because they can be proven wrong: the truth capacity of the narrative statements of historical writing is based on the fact that they can be falsified. In this respect, one can imagine no greater difference between historical writing and fiction.

Since the statements of a novel are fictional to begin with, they cannot be falsified. While writers of fiction may embellish an event from the past with fantasy, historiographers must check the sources, which dictate what they cannot say. The truth of a poetic statement is not achieved by fiction imitating the principles of historiography. The propensity for truth and truthfulness itself are presented differently in the two narrative modes. The intent to provide the truth does exist in both literature and historical writing, but it is aimed in different directions. Historiographers attempt to achieve the most exact comprehension and representation of actual historical events; novelists have no such ambition. On the contrary, their strength lies not in the realm of facts but in that of possibilities.

The aesthetics of the modern as well as of the postmodern novel are focused on the opening up of possibilities. Modernist literature aims to distance itself from the mimetic-realistic principle that dominated the novel of the second half of the nineteenth century. Musil, for example, in *Der Mann ohne Eigenschaften*, plays with several types of utopias and arrives at the category of 'possibility,' in opposition to confidence in the power of history. The same is true in the postmodern novels of Thomas Pynchon.

5. *The Concept of Totality*

In examining the truth postulate in literature and historiography, one must consider the premises on which historical and novelistic narration are based. In traditional narration the novelist, like the historiographer, is guided by ideas of totality as we know them, on the one hand from the aesthetics of Hegel to Lukács, and on the other hand from Hegel's and Marx's substantialist philosophy of history. In both cases, the attempt is made to see events in such a way that they make sense within the context of a totality. Here the historical and artistic totalities are structurally identical. Danto has criticized as outdated the substantialist philosophy of history, with its concept of a teleological, continuous development. He states that this philosophy acts as if it knew of a 'divine plan' of history, as if one could determine past and present from the perspective of the future. Danto wants to replace the speculative and dogmatic substantialist philosophy of history with an analytical (descriptive and coordinating) theory of history. In this theory, the vantage point would not be from a postulated future goal back to the present and past, but rather from the past to the present, with no certainties about the future (Danto 1965). The ancient trust, reaffirmed by Nietzsche, in the continuity of history is questioned, as are the Christian and Marxist tenets of the eschatological or utopian direction of history.

Such beliefs once formed the foundation for the concept of continuity in history. Many contemporary historians shun these labels and are consequently faced with two choices: giving up the concept of continuity altogether, or adopting it from the tradition of historiography and perhaps developing it in their own fashion. Historians at present are more aware than ever of the heuristic character of the concept of continuity; they see it merely as a 'regulative idea' in historical narration (Rüsen 1982). Today the focus is more on discontinuity. In this respect the impact of Foucault as an outright anti-Hegelian thinker (especially with his 1969 study *L'archéology du savoir*) is not to be underrated.

The dismissal of the concepts of continuity and totality has also had an impact on the writing of the novel. The idea of the 'self-contained' work of art that provides totality by way of typical characters is dead. Its place has been taken by the 'open' novel with its open-endedness, as first envisioned by the early German Romantics. In his *Theorie des Romans* (1916), Georg Lukács declared that the modern novelist could outdo the historiographer if, after the loss of totality of meaning and being, he or she would counter the partial insights into the course of history with concern for a new, homogeneous cosmology. In contrast, however, the modern novel plays an important role as an instrument of cognition because such concepts of systems are now distrusted. Not a new cosmology of being, but a view of the disparity of what is to be, of the diversity of possibilities, is what this medium can provide.

In twentieth-century philosophy, a similar discussion to the one between Plato and Aristotle about the truth-value of fiction took place. Adorno was convinced that, due to its autonomy, art 'as a creative, nonalienated work' was the last vestige of freedom.

According to him, art and literature alone can unveil what is hidden in reality and can destroy ideology and lies (Adorno 1970). What Adorno criticized as a web of ideological deceptions is close to what Roland Barthes calls the 'mythicizing' of everyday life. Like Adorno, Barthes affirms that poetic language maintains direct contact with reality and, unlike historiography, escapes mystification (Barthes 1957). In contrast to Adorno and Barthes, Eagleton and Jameson argue that the writing of literature is just another form of ideological production (Eagleton 1976). Jameson maintains the following position: 'Ideology is not something which informs or invests symbolic production; rather the aesthetic act is itself ideological, and the production of aesthetic or narrative form is to be seen as an ideological act in its own right' (Jameson 1981).

All four authors make extremely one-sided statements. The common ground of historiography and literature is narration. And when narration, through the eyes of Jameson, is seen as an ideological act per se, we cannot expect much truth in either historiography or fiction. But if, on the other hand, literature is able to destroy ideology and deception—to cite Adorno—this virtue should also be attributed to historiography. Narration can lead in both directions: it can help to approach truth, or it can produce ideological veils. Neither literature nor historiography is immune to the virus of error or the distortion of truth, but both are capable of making reality be seen more clearly.

6. *Meaning, Memory, and Identity*

Among the functions of historiography and the novel are, along with the search for truth, the imparting of meaning and the provision of identity. The imparting of meaning (*Sinnbildung*) is far from the creation of meaning (*Sinnstiftung*) (Rüsen 1982). Surely there were, and still are, historiographers and novelists who have declared themselves ideological creators of meaning, but neither modern novelists nor critical historiographers are interested in such ambitious goals. In historical narration, the imparting of meaning is achieved through memory alone. And although novelistic narration cannot be reduced to memory, it plays an important role here as well. This is especially evident when historians fail to bring to mind certain aspects of historical experience through memory; then it is often the novelist who must assume this task. The 1950s in West Germany provide a clear example of this phenomenon. With regard to the National Socialist past, repression rather than grief was prevalent. It was a novel that broke this taboo: *Die Blechtrommel*, by Günter Grass (1959). This literary breakthrough was fallowed by the works of the historiographers. George Steiner listed a number of examples of this function of the 'writer as remembrancer' (Steiner 1973). It is particularly in cases of the conscious falsification of history by historians that the writer steps in.

Writers and historiographers can help define the identity of a generation, a nation, or a culture. Since the Enlightenment and secularization, historiography and fiction have been burdened increasingly with the function of providing an identity. By dealing with memories, the experiences of the present attain a coherent perspective that provides the reader with an orientation for the future. Repeatedly correlating experiences of the present with those of the past can bring about the dissolution of stagnant patterns of identity, with a critical view of ideological prejudices (Lützeler 1998). The factual orientation of the historiographer as well as the novelist's interest in what is possible are part of the provision of identity. By means of memory, a potential of historical and literary meaning is activated that, along with the demonstration of actual limits and the allusion to yet untested possibilities, makes the provision of identity possible. Historical and literary consciousness is a complex construct in which memory is the prerequisite for interpretation and expectation. To structure an identity means, first, to identify with stories. During an identity crisis, society may show greater interest in historical remembering (as was the case in Europe in the nineteenth century), or demand a fictional provision of identity (as is presently the case in Latin America). The weaker the link with the historical past and the more suspect it appears, the stronger will be the desire for fictional patterns of identity.

See also: History and Memory; History, Theories and Methods; Anthropological Universals

Bibliography

Adorno T W 1970 *Ästhetische Theorie*. Suhrkamp, Frankfurt, Germany

Assmann A 1980 *Die Legitimität der Fiktion*. Fink, Munich, Germany

Barthes R 1957 *Mythologies*. Editions du Seuil, Paris

Baumgartner H M, Rüsen J 1982 Erträge der Diskussion. In: Lämmert E (ed.) *Erzählforschung. Ein Symposium*. Metzler, Stuttgart, Germany

Benjamin W 1972 *Gesammelte Schriften 2*. Suhrkamp, Frankfurt, Germany

Borst A 1976 Das historische Ereignis. In: Baumgartner H M, Rüsen J (eds.) *Seminar: Geschichte und Theorie: Umrisse Historik*. Suhrkamp, Frankfurt, Germany

Cohn D 1978 *Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction*. Princeton University Press, Princeton, NJ

Culler J 1975 *Structuralist Poetics. Structuralism, Linguistics and the Study of Literature*. Routledge and Kegan Paul, London

Danto A C 1965 *Analytical Philosophy of History*. Cambridge University Press, Cambridge, UK

Demandt A 1978 *Metaphern für Geschichte*. Beck, Munich, Germany

Eagleton T 1976 *Criticism and Ideology. A Study in Marxist Literary Theory*. Humanities Press, Atlantic Highlands, NJ
Eco U 1979 *The Role of the Reader*. Indiana University Press, Bloomington, IN
Fish S 1970 Literature in the reader: Affective stylistics. *New Literary History* **2**: 123–62
Furet F 1982 *L'atelier de l'histoire*. Flammarion, Paris
Grimm R 1963 *Strukturen*; *Essays zur Deutschen Literatur*. Sachse und Pohl, Göttingen, Germany
Hamburger K 1968 *Die Logik der Dichtung*. E Klett, Stuttgart, Germany
Hinck W 1995 *Geschichtsdichtung*. Vandenhoeck & Ruprecht, Göttingen, Germany
Iser W 1974 *The Implied Reader*; *Patterns of Communication in Prose Fiction from Bunyan to Beckett*. Johns Hopkins University Press, Baltimore, MD
Jameson F 1981 *The Political Unconscious: Narrative as a Socially Symbolic Act*. Cornell University Press, Ithaca, NY
Kocka J 1984 Zurück zur Erzählung? Plädoyer für historische Argumentation. *Geschichte und Gesellschaft* **10**: 395–408
Koselleck R 1976 Wozu noch Historie? In: Baumgartner H M, Rüsen J (eds.) *Seminar: Geschichte und Theorie*. Suhrkamp, Frankfurt, Germany
LeGoff J, Nora P 1974 *Faire de l'histoire*. Gallimard, Paris
Lützeler P M 1986 *Zeitgeschichte in Geschichten der Zeit*. Bouvier, Bonn, Germany
Lützeler P M 1998 Nomadentum und Arbeitslosigkeit. Identität in der Postmoderne. *Merkur* **52**: 908–18
Prince G 1982 *Narratology: the Form and Functioning of Narrative*. Mouton, Berlin, Germany
Ricoeur P 1983 *Temps et récit*. Seuil, Paris
Riffaterre M 1971 *Essais de stylistique structurale*. Spanish, Seix Barral, Barcelona
Rüsen J 1982 Die vier Typen des historischen Erzählens. In: Koselleck R (ed.) *Formen der Geschichtsschreibung*. Deutscher Taschenbuch Verlag, Munich, Germany
Steiner G 1973 The writer as remembrancer: A note on poetics 9. *Yearbook of Comparative and General Literature* **22**: 51–7
Stone L 1979 The revival of narrative reflections on a new old history. *Past and Present* **85**: 3–24
Szondi P 1973 Für eine nicht mehr narrative Historie. In: Koselleck R, Stempel W D (eds.) *Geschichte—Ereignis und Erzählung*. Fink, Munich, Germany
White H 1973 *Metahistory: The Historical Imagination in Nineteenth-century Europe*. The Johns Hopkin University Press, Baltimore, MD

P. M. Lützeler

History and Memory

1. The Alliance Between History and Memory

1.1 Remembered History

Before history is registered, chronicled, and interpreted by professional historians, it is witnessed and remembered by those that are contemporary to it. Such a memory of experienced history is necessarily partial, biased, subjective, and, therefore, also highly variegated. While the present of experience is witnessed as a rather formless, dissonant, and even chaotic mass of perceptions, actions, and events, it assumes a much more unified shape and meaning when it is transformed into what is called the 'past.' To some extent, the multifarious, disrupted, and inchoate experiences of the former present are preserved in personal memories, if they manage to resist the tendency to adapt to the retrospective interpretation and normalization of the past. For many historians, there is very little overlapping between the present as experienced and witnessed and the past as reconstructed by historians (Koselleck 1997). While individual experience is particular, partial, and extremely limited in perspective, the reconstructed past provides a coherent and inclusive view of what was never in this way to be witnessed. In contrast to the past, the present does not produce a common reality that can be experienced by the witnesses in a similar way.

The plurality of lived experience and the unified reconstruction of history, however, are not fixed in a radical polarity. These poles are mediated in many ways, bringing the realms of present experience and retrospective history closer together. The personal experience of history is not only radically subjective, it is also shaped by the generalizing and collectivizing factors of social norms and cultural knowledge. Individual behavior and memory is not wholly erratic because it is stamped by generational patterns of experience, which means that it is already to some extent embedded in and framed by collective and cultural constructions of history.

It is true that personal memories of history differ as one individual life differs from another life. This, however, does not exclude certain similarities arising from biographical patterns. These life patterns with their respective perspectives, challenges, and responsibilities have their impact on the constitution of memories. The second world war, for instance, is remembered differently by the different nations as it is obviously remembered differently by the victims, the perpetrators, and the bystanders. But even among the Germans, it was remembered differently by different generations, by men and women, by those who remained in place and those who were forced to migrate from east to west. It is the task of the oral historian to acknowledge these individual and collectively typical memories as historical sources and to investigate history from below, collecting the various points of view of those who did not plan and control history but who were affected and afflicted by it.

There are other reasons why the singularity and interiority of memories should not be overestimated. Memories are generated in the individual very much like the linguistic capacity and in accord with it. They are elaborated and stabilized to the extent that they are transformed into narratives and exchanged in social

interaction. The memories of an individual are always supported and warranted by the memories of other individuals. Subjective memories are embedded in familial interaction and social communication. As individuals share the memories of their parents and grandparents, the range of remembered history is extended over three generations. Thus, an embodied and participatory historical memory of approximately a 100 years, is built up by oral transmission.

The three-generational oral memory is transient and perishes continually with every new cycle of oral memory that is built up. This continuous fading out of remembered history is a normal, natural, and necessary process of social forgetting which wipes the slate clean for following generations and their experiences. An exception to this rule occurs when the remembered history has an heroic or traumatic character. In cases of historical triumph or trauma, arts of memory are devized to stabilize the otherwise transient memory and anchor it in the consciousness of following generations (Giesen 2000). The survivors of the holocaust are bearers of such a memory which by a general cultural consensus is to be made permanent. The transformation of personal memories into cultural memory is achieved by media of memory such as writings, films, museums, monuments, and video testimonies.

In societies without writing, historical consciousness is divided into a recent and a distant past. The recent past is the three-generational oral memory which is transformed continually over time while the distant mythical past is a heroic memory which is preserved by epic poetry and other ritual and ceremonial arts of memory. This cultural heritage is re-enacted periodically and celebrated in public festivals by the community and carefully preserved and transmitted by bards and other memory specialists.

1.2 Commemorated History

Experiential subjective memory is embedded in collective memories of a widening scope: that of social interaction and that of cultural identity. The boundary lines between individual, collective, and cultural memory are not water-tight, but, as Ricoeur has put it: '*mémoire privée et publique se constituent simultanément selon le schema d'une constitution mutuelle et croisée*' (Ricoeur 1998, p. 32). While private memory is informal and erratic, public memory shows a much higher degree of organization. It is mediated, ritualized, and based on social institutions. Other than remembering, the term 'commemorating' refers to an intentional, formalized, and collective action. Public memory is built on a collective will and communal actions. In the transition from private to public and cultural memory, the media play an important role. It is particularly writing that has revolutionized cultural memory. Following the anthropologist André Leroi-Gourhan, Jacques Le Goff has described the history of cultural memory in five stages as the history of changing media from orality and writing to ever more mechanical technologies of storing information (Le Goff 1986). Not only did it vastly extend the range of remembered history, it also liberated the poets from their traditional office of 'remembrancers,' setting them free for other, more individual tasks. There are two kinds of media with which the historical memory of a group is stabilized and transformed into an extended and continuous past: material media and performative media. Among the material media are books, photographs, films, monuments, relics, and other material carriers of signs that retain traces of the past and can serve as documentary sources in the reconstruction of history. The performative media are symbolic forms of action, such as rites, festivals, and ceremonies linked to symbolic dates. Traditions and celebrations periodically renew and revitalize the memory of the past through acts of repetition and commemoration. While the material media store information for later use, they tend to generate specialized knowledge and to externalize the memory of a group. It is, therefore, the function of institutions of education and the performative media to check this tendency and to restore the mnemonic quality of historical knowledge, to bring it back to popular consciousness, and to reconnect it with the realm of lived experience. Usually, material as well as performative media are involved in acts of commemoration. Halfway in between material and performative media are the modern mass media such as newspapers, cinema, radio, and television, not to forget the large historical expositions that are becoming more and more important in shaping an image of the past and providing occasions that trigger collective memory through public debates.

In addition to their personal, local, and familial recollections, individuals also hold a share in that 'corporate memory' (Lewis 1975) which is constituted by public symbols and rites of commemoration. 'Commemorative ceremonies are distinguishable from all other rituals by the fact that they explicitly refer to prototypical persons and events, whether these are understood to have a historical or a mythological existence' (Connerton 1989, p. 61). The creation of a corporate memory was the aim of ancient societies and was revived by the modern nation–state and other politicized groups. A corporate memory is extremely limited in scope. Unlike books and libraries, it cannot contain many and diverse historical events, nor does it do justice to finer distinctions. It selects only very few items and reshapes these historical events for its purposes. The long chronicle of history has to be condensed into a few poignant moments that acquire a high symbolic value for the identity of a nation and state. Examples for history telescoped in national memory are the 4 July in the US or the 14 July in France. Such regularly recurring dates provide occa-

sions for speeches and celebrations which re-invigorate both the corporate memory and the collective identity. In Israel, the historical event of the fall of Massada became a national cult in the new state. For many centuries, Massada had been forgotten in the Jewish tradition and was restored by historical and archaeological scholarship only in the 1960s. For an historical event to become an item of public memory, the event has to be taken out of its context and to be transformed into a symbol. Among the first to emphasize this was the sociologist Maurice Halbwachs who wrote in his book on the social frames of remembering: 'In order for a person or historical event to enter social memory, it has to be transformed into an instruction, a concept, a symbol. It is invested with meaning, it becomes an element in the system of ideas of a society' (Halbwachs 1985a, pp. 389–390). In the case of Massada, the political message of the historical event was coined in a poem by a Palestinian Jew who in the 1920s had written the line: 'Massada shall not fall again.' The memory of Massada was restored from Hellenic sources that had not been part of the heritage of remembered Jewish tradition (Vidal-Naquet 1981). This memory was 40 years later corroborated by archaeological findings which in themselves, however, were rather scarce. Before these long forgotten external sources could become a cornerstone in the construction of a national myth, they had to be reanimated and kindled by the spark of historical imagination. The story of Massada became a central point of reference in Israeli corporate memory, because in the new state, a new future required a different past (Lewis 1975).

In antiquity, the commemoration of historical events took the form of religious feasts which were celebrated annually like the feast of Passover in which the Jews commemorated their liberation and exodus from Egypt. Such events were integrated into the religious calendar with the effect of turning history into myth. The English still celebrate annually the memory of the attempted catholic assault on their parliament of 1605, with the folkloristic rites of the Guy Fawkes-day starting as early as in 1606. Another way to commemorate historical events is to celebrate them according to the 'round' decimal dates of chronology. In contrast to annual feasts, these cyclical patterns seldom recur in a life-time and, therefore, lack the repetitive structure which is necessary for a continuous memory. This form of long term commemoration already goes back to antiquity. There is a stela of Ramses II commemorating the establishment of the cult of Seth in Avaris after 400 years. As a practice of communal commemoration, however, the decimal pattern was introduced only in the early modern era. One of the earliest examples is the commemoration of Luther's reformation in 1615, 100 years after Luther's actions of protest at the Augsburg Reichstag. This mode of recalling past events via anniversaries in the decimal system has become the dominant mode of remembering history in the nineteenth and twentieth century. Examples are the 500th anniversary of the Turkish conquest of Constantinople which was celebrated in Turkey in 1953 or the 500th anniversary of Columbus' discovery of the new world celebrated in 1992. Such events do not recur within the lifetime of individuals. Therefore, it is rather a mode of spotlighting historical dates than a form of creating a continuous, communal memory.

1.3 Myth and History

Myth has been established by theoreticians of the nineteenth and twentieth centuries as the inveterate antagonist of history. Myth and history were considered as two incompatible modes of presenting the past, the one in a cyclical, the other in a linear time frame. In this polemical constellation, the term myth was also associated with fabrication and lie, while truth and objectivity were reserved for history as presented by critical historians. Beyond these well-known connotations of the cyclical and the untrustworthy, the term myth has in the meantime also acquired the connotation of 'meaningful history.' 'Myths create meaning through which the past is linked to the present in such a way that it can orient the future' (Münkler 1994, p. 21). This description is not far from Assmann's definition of myth as 'imagined and remembered history' (Assmann 1992, see also Kirsch 1998). Myths are conceived by him as a complex of icons, symbols, and narratives about the past that define the cultural identity of a community. Such a communally imagined and transmitted past provides for the group the basic values that it needs for an explanation of the present and an orientation for the future. It is exactly this symbolic use of the past for the construction of an identity that was suspended in the process of professional historical scholarship as developed in academic institutions in the nineteenth century. In this specific sense, the term 'myth' is used as an antonym to (modern) histor(iograph)y; while the former refers to the presence of the past, the latter insists on the pastness of the past, enforcing 'a liberation of the present from the past' (Arnaldi 1974). In the era of historiographic scholarship, however, remembered history is not the opposite of factual history, on which it has to rely as its basis and source of correction. If myth is defined as remembered history, it is severed from other connotations such as 'lie,' 'imposture,' or 'sacred history.' It refers to 'a foundational narrative with the aim to illuminate the present by an account of origins. It is through remembering that history turns into myth. This does not mean that it becomes unreal. On the contrary, it is by turning history into myth that history becomes reality and assumes the quality of a normative and formative power' (Assmann 1992, p. 52).

Myth in this new sense of remembered history stresses mnemonic qualities such as symbolic power, narrative form, iconic impact, and a highly focussed and sensual form of relevance. Invested with a memorable shape, symbolic meaning, and pragmatic relevance, myth provides qualities that professional history had explicitly discarded. With a developed sense of the shortcomings of both myth and history, they are no longer confronted as antagonistic and mutually exclusive but rather seen as complementary modes of referring to the past. The mnemonic and symbolic use of the past is no longer perceived as an abuse but as an important social function which cannot be provided by scholarly historiography.

2. *Memory and Historiography*

2.1 *The Identity of Memory and History*

The relation between historiography and memory has itself a history which has evolved over various stages. Before the foundations of critical historical scholarship were laid in the eighteenth and nineteenth centuries, history, memory, and myth were not yet clearly distinguished. On the contrary, it was considered the central function of the writing of history to preserve the memory of a dynasty, the church, or a state in order to legitimize such institutions and to ensure their continuity by providing for them an honorable past. In this first stage, historiography was the foundation of the community and the 'handmaid of authority' (Plumb 1969). The *res gestae* of heroes and kings were sung by bards and written down by chroniclers in order to rescue them from oblivion and to establish an honorable memory for the noble dead.

Although a consciousness of a specific function of historiography was established in the writings of Herodotus, he emphasized in the proemium to his work the traditional link between history and memory. While giving up a strictly ethnocentric stance, history, for him, is still 'an art of memory' where 'memory' is closely related to 'fame.' His project claims to ensure 'that the actions of people shall not fade with time' and that 'great and admirable monuments produced by both Greeks and barbarians shall not go unrenowned' (Herodotus 1992, pp. 3–4, see also Hutton 1993). His historiographical project is aimed at fixing the events of the past in the memory of those to come. For Herodotus, the function of historiography was already twofold: it was divided into a memorial function and a cognitive function ('to set forth the reasons why they, i.e., the Greeks and the barbarians, waged war on each other'). The memorial function of historiography was never given up throughout Greek and Roman antiquity. Cicero still defined historiography as a weapon against oblivion, a formula that became a topos and was elaborated in the Renaissance. Official historiography of the Tudor dynasty in sixteenth century England still counters oblivion which is personified as 'the ancard enemie,' 'the suckyng serpent,' 'the deadly darte,' and 'the defacer' (Campbell 1964).

The intimate link between memory and history shaped the historiographic genres in a specific way. It narrowed the criteria for the selection of persons and events to be memorized; only those of highest rank were singled out for a continuation in memory and only those feats and achievements were selected that contributed to the honor and fame of those that were remembered. In addition to this, only such events were selected that supported the opinions and interests of the ruling class.

The non distinction between memory and history is not only a feature of early stages of civilization but also a characteristic of authoritarian and totalitarian rule. But, whereas in the early stages neither media nor institutions had been established to back up an authority-independent account of the past, in later stages censorship had the function to destroy the rivaling media and carriers of counter histories threatening the security of the uniform view and authoritarian voice of history.

2.2 *The Polarization of Memory and History*

2.2.1 The discrediting of memory by historians. The second stage is marked by a separation between memory and histor(iograph)y. This distinction was established when professional historiography arose as an independent and specialized institution. It was not before the nineteenth century that historiography turned into an autonomous discursive discipline which set up its own standards of truth and authority, developing specific rules of verification and intersubjective argument. The development of modern historiography as a cognitive enterprise has its roots in classical antiquity and was revived in early modernity. Histor(iograph)y started from memory and it reached its new status in emancipation from memory. Before the rise of critical historiography, history had been adapted to the demands of the present; it served specific functions for the state or community such as justifying the present, controlling the future and legitimizing authority. A critique of memory, therefore, was also a critique of power and authority. Critical historians insist on their distance to the institutions of corporate memory. Lukian called them 'apolis,' thereby indicating that they are not committed to the norms and ideals of a particular community. In spite of their ideal of objectivity, of course, even professional historians are not completely free from loyalties and prejudices. But the professional historian differs from others in that 'he is aware of this fact, and instead of indulging his prejudices seeks to identify and correct them' (Lewis 1975, p. 52). The criticism of historical memory not

only aims at filling its gaps and correcting its errors but also to expose the intentional fallacies and forgeries of memory, the constructions that secure the establishment and maintenance of power. Critical historiography sharpened its tools in the period of Renaissance humanism in the struggles for power and truth, but it developed its ideal in disinterested objectivity only in the nineteenth century. It was in this period that the links between history and identity were severed and historical knowledge turned into a rapidly growing mass of abstract, disconnected, and disembodied information.

2.2.2 The rehabilitation of memory by philosophers and historians. It was Nietzsche who saw a connection between the massive production of abstract historical knowledge and a growing apathy of cultural memory, which led him to express an urging demand for new and invigorated identities (Nietzsche 1988). According to Nietzsche, the historical disciplines had become a dangerous burden for the wellbeing of society. Scholarship produced a pathological form of memory without the possibility of forgetting, an ever growing archive of irrelevant data. Every group and collective unit, however, was in need of a clear cut and limited knowledge of the past, i.e., of a usable past. Nietzsche diagnosed for his age a severe crisis and pathology of cultural memory which had lost its capacity of limitation. For him, the important function of memory was not the storing of knowledge but the limitation of knowledge, the creation of profile and relevance in an indifferent mass of information. Historical knowledge was no longer a tool for the creation of a meaningful life but an oppressive burden that crushed identity and extinguished the sparks of creative energy.

Nietzsche's polarization of history and memory became a topos among historians of the twentieth century. The specificity of histor(iograph)y was defined in contrast to the demands and mechanisms of memory. When Maurice Halbwachs defined the mechanisms of social memory in the 1920s, he introduced history as a negative foil for the process. Halbwachs investigated the function of shared memories for the cohesion of groups. He spoke of a collective memory and showed that the group was sustained and nurtured by common memories as these memories were in turn sustained by the group. This means that the stability of memories is linked to the stability of groups; when its members are disseminated, the memories are also dissolved. Erratic and dissonant memories had no place in Halbwachs's functionalist and constructivist theory of memory. Halbwachs compared historical discourse to an ocean which swallowed up all the partial narratives and memories (Halbwachs 1985b, pp. 72–3). For him, histor(iograph)y is the universal memory of humanity, while collective memories are embodied by specific groups and therefore always partial and biased. In order to integrate these memories into the total view of history, they have to be severed from their carriers and their social milieus. With this transformation of memory into the abstract scheme of history goes the evaporation of live experience and meaning.

Sixty years after Halbwachs, the French historian Nora has revived the concept of collective memory and dispelled lingering critical doubts about its theoretical status. He has extended Halbwachs' notion of an oral memory shared by a social group to a cultural memory of symbols, values, rites, and local traditions that provide the cohesive cement of a society. These shared elements Nora called 'lieux de mémoire,' reinterpreting a term from the tradition of Roman mnemotechnics, the 'imagines' and 'loci' of the '*ars memorativa*.' Like Halbwachs, Nora stresses the nexus between memory and group identity, and like Halbwachs, he distinguished the symbols of collective memory from the scholarly discourse of historiography. For Nora as for Nietzsche, embodied collective memory is threatened to disappear under the burden of academic historiography. Nora defines memory and history as polar opposites. For him, the decisive distinction is that between a lived past which he calls memory and an abstract past which he calls history. Memory links the past to a group, it creates an emotional and symbolically charged bond between the past and a collective identity in the present, while history severs this bond; it is universal in character, therefore it belongs to everybody and nobody (Nora 1984/1992, pp. 23–43).

Whereas the pioneers of critical historiography discredited memory as a rival of history, theorists such as Nietzsche, Halbwachs, and Nora rehabilitated memory over against the ideal of objective and abstract histor(iograph)y. They rediscovered the social functions of memory as an important and indispensable mode of assessing the past. In spite of their evaluations of memory and history, however, they share with the representatives of critical historiography the attitude of polarizing memory and history. In their different writings, this polarization takes the following form:

Memory
–is an embodied form of memory
–stresses differences and exists in the plural
–is linked to a carrier which may be an individual, a group or an institution
–bridges the past, present and future
–is highly selective, includes forgetting
–creates a profile of values supporting an identity and providing orientation for action.

History
–is a disembodied form of memory
–provides a universal frame and exists in the singular
–is disengaged from specific carriers

–separates the past from present and future
–develops an even and impartial attention
–is in search for truth, therefore values are suspended.

2.3 Mnemohistory—New Affinities Between Memory and History

The third stage in the relationship between memory and historiography is marked by a blurring of the distinction between history and memory. This, however, does not imply a simple return to the first stage. Despite postmodern qualms and revaluations, the self-imposed disciplinary regulations that were installed with the modern institutionalisation of critical historiography as a scholarly and intersubjective project may well be considered an irreversible evolutionary achievement. Historiography, writes Le Goff, 'must aim at objectivity and must be built on a belief in historical truth' (Le Goff 1986, foreword). If the third stage is not to be described in terms of a regression, it is perhaps more adequately described as a response to certain tendencies towards an extreme historical positivism inherent in the second stage. The professional historians as exclusive guardians of the past are challenged by those who invoke memory as alternative modes claiming an access to the past. The term 'history' is a collective singular; to this concept corresponded the idea of history as a universal memory of mankind. To say that history is universal memory may be another way of saying that it is indifferent to memory. The price that had to be paid for this generalized memory was a disregard of living and preserved memories.

The object of modern historiography was a past that was neatly separated from the present and the future. The historian Koselleck (1997) has coined the terms 'space of experience' (*Erfahrungsraum*) and 'horizon of expectation' (*Erwartungshorizont*) and stressed the point that in modernity, both dimensions have become radically separated. Along the same lines, history was defined by the philosopher Blumenberg as 'the separation of the space of experience from the horizon of expectation,' i.e., as a fundamental rupture between past and future (Blumenberg 1986, p. 61). From this point of view, memory was the other of history because collective memory tends to bridge past and future, creating a usable past for a changing present.

After the long period of polarizing history and memory, new affinities were discovered between both modes of rendering the past. They were due to a number of reasons: the discontents with positivist historiography and transcendental historiosophy, a reorientation in methodology, and the establishing of new themes for historical investigation. To begin with the discontents. In the 1980s and 1990s the universalist frame of academic historiography was severely shaken when repressive and homogenizing ideologies such as communism and colonialism lost their power. Since then, many divers memories of political groups have asserted their right. Historians were baffled by the enormous impact of these living memories that they had considered to be a negligible entity and which had reappeared with great aplomb at historical turning points. At such moments of sociopolitical reversal, social institutions as well as the cultural frame of perception and values changed. In the course of such changes, new memories emerged and old memories were seen in a different light. It is at such historical moments that the conflict between history and memory rises to consciousness and becomes a political issue. With the reconstruction of the political framework of society, there is a chance to narrate the past in a different way. Such a moment of retrospection is often a moment of revelation; then it suddenly becomes obvious that what had been presented and taken as objective history had only been a particular and biased form of memory. The experience of a fundamental change of values exposes the contingence of earlier accounts of the past. In such situations memory becomes self reflexive, developing a sense of its own constructedness.

Another way of approximating history and memory was to be observed in changes in the methodology of the discipline. The journal *History and Memory* which was founded in 1989 signals a shift in historical interest. Two years before that date, an important debate had taken place between the nestor of German *Zeitgeschichte*, Martin Broszat, and the Jewish Holocaust survivor and Holocaust scholar Saul Friedländer. The controversy which they discussed in their exchange of letters revolved around the key concepts of memory and history (Broszat and Friedländer 1988). Broszat argued for a form of presenting history that moves from the past into the present, thus adopting the perspective of the contemporary observers who experience it as an open process. Friedländer, on the other hand, argued for a presentation of history starting from its manifest results and moving backwards to the beginnings, thus adopting the perspective of the memory of the victims. For Dan Diner, Friedländer's co-editor in *History and Memory*, historiography is no longer an objective universalist discourse but one that is ethically committed to the participatory perspective of the victims. Thus, the two terms converge, history becoming a form of memory and vice versa. In his latest book on the persecution of the Jews in Nazi Germany, Friedländer develops a form of historiography that is not confined to the description of structures and processes but also seeks to give a voice to individual victims and agents thus translating the anonymous statistics of the holocaust back into the memories of those who participated in and suffered from history (Friedländer 1998). In books such as this one a new methodology of history is being developed that seeks

to enlarge the scope of historiography by integrating personal memories into historiography. Instead of being unified by the authorial voice of the interpreting historian, historiography thus becomes multivocal, mixing history with memories. In blending what is known with what is remembered, historical discourse is opened up to include contingencies ad dissonances, forgoing coherence and closure in favor of marking what otherwise remains unrepresented and unrepresentable.

A third possibility of crossing memory and history lies in the discovery of new themes for historiography. Since the 1980s, historians have become more and more interested in modes of remembering as a form of social practice. What was known of the past in the past, which images of the past were constructed, which acts of commemoration were devised? In asking such questions, memory became itself the object of historiography. This form of historiography was favored by medievalists, who had already been used to investigating social forms of religious and secular memorializiation (Oexle 1995). Such interests brought about a form of historical investigation which can be termed mnemohistory. Mnemohistory is what could also be called 'the reception history of history.' 'It concentrates exclusively on those aspects of significance and relevance which are the product of memory' (Assmann 1997, p. 9). To speak in this context of memory rather than of tradition is to acknowledge the constructive as well as the distorting power of memory. It takes into account the ambivalence of the past both as a conscious choice and as an unconscious burden, tracking the voluntary and involuntary paths of memory. 'The task of historical positivism consists in separating the historical from the mythical elements in memory and distinguishing the elements which retain the past from those which shape the present. In contrast, the task of mnemohistory consists in analyzing the mythical elements in tradition and discovering their hidden agenda' (Assmann 1997, p. 10).

An early and impressive project in mnemohistory was organized by the French historian Nora who published seven volumes on what he called the French 'lieux de mémoire' (Nora 1984/1992). Nora designed this historical enterprise which he carried out as team work with eminent historians in reaction to the publications of the Annales school of French historiography. In the scholarly works of this school, which had concentrated on structures, figures and processes in a longue durée perspective, history had more and more lost its meaning and memorability. The palpable dimension of history was restored by Nora who created an inventory of mnemo-historical items which included historical sites such as Vezelay, rites and songs of folklore, national symbols, commemoration dates and festivals, and more than 300 types of cheese. They are the 'lieux de mémoire' that, according to Nora, have constituted the French cultural memory and identity since 1789.

2.4 Memory as an Empirical Basis for Historiography

It is not easy to predict the next phase in the relationship between memory and history. It is possible that after their convergence, new efforts will be directed at their separation. There is also a new interest in memory from the point of view of a new kind of empiricism that is probing the criteria for historical truth. While one variation of mnemohistory tried to integrate personal memories into history and another sidelined factual history in order to focus on a remembered, reconstructed, or even repressed and forgotten past, there are some signs now of a new anthropological empiricism which turns away from memory as a social and cultural construct and concentrates on the neurological operations of remembering and forgetting. In an article with the title *The Veil of Memory*, the historian Johannes Fried noted 'a frightening backwardness among historians as far as questions of memory are concerned' (Fried 1997, p. 8). After the growing fascination and even obsession of historians with questions of memory since the 1980s, this statement can only make sense when it is applied to questions of memory from the point of view of the human brain and the neurosciences. He is interested not in cultural memory but 'the factuality of memory, its conditions and its effects on the historian's sources' because 'the neural connections or the mental processes determine what is preserved and remembered, and thus all knowledge and all history. The same applies to forgetting. The veil of memory, the river of oblivion is constantly in flow between the historian and its object' (Fried 1997, pp. 14, 16–17). This knowledge of the operations of memory also includes the mechanisms of distortion as explored in cultural studies and the neurosciences in the last ten years (Schacter 1999, Schudson 1992). For Fried, however, the crucial question for the historian to ask is no longer: how are texts connected? but: how does memory work? At this point, we return to the lived experience of individuals witnessing history as the starting point for historiography which was treated in the first paragraph of this article. Fried, however, no longer discusses memory in an existential or in a social but in a biological and anthropological perspective. He argues that everything that the historian deals with is primarily mediated by the human memory. Therefore, the historian must have recourse to anthropological empiricism in order to 'see through all this memory-activity' and probe the truthfulness of his one and only reliable source which is, after all, human experience.

See also: Collective Beliefs: Sociological Explanation; Collective Memory, Anthropology of; Collective Memory, Psychology of; Commemorative Objects; Historiography and Historical Thought: Current Trends; History and Literature; History: Forms of

Presentation, Discourses, and Functions; History: Overview; Identity: Social; Media and History: Cultural Concerns; Memory: Collaborative; Narrative, Sociology of; Oral History; Ritual; Ritual and Symbolism, Archaeology of

Bibliography

Arnaldi G 1974 La storiografia come mezzo di liberazione dal passato. In: Cavazza F L, Graubard S R (eds.) *Il Caso Italiano*. Garzanti, Milan, pp. 553–62

Assmann J 1992 *Das kulturelle Gedächtnis. Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. Beck, Munich, Germany

Assmann J 1997 *Moses the Egyptian. The Memory of Egypt in Western Monotheism*. Harvard University Press, Cambridge, MA

Blumenberg H 1986 *Weltzeit und Lebenszeit*. Suhrkamp, Frankfurt, Germany

Broszat M, Friedländer S 1988 Diskussion um die Historisierunges Nationalsozialismus. Ein Briefwechsel. [*On the historicization of national socialism—an exchange of letters*.] *Vierteljahrshefte für Zeitgeschichte* **36**: 339–72

Campbell L B 1964 *Shakespeare's Histories—Mirrors of Elizabethan Policy*. Methuen, London

Connerton P 1989 *How Societies Remember*. Cambridge University Press, Cambridge, UK

Fried J 1997 *The Veil of Memory. Anthropological Problems When Considering the Past*. The German Historical Institute London, 1997 Annual Lecture, London

Friedländer S 1998 *Nazi Germany and the Jews* (Das Dritte Reich und die Juden, Vol. I: Die Jahre der Verfolgung 1933–9). Beck, Munich, Germany

Giesen B 2000 *Triumph and Trauma*. University of California Press, Berkeley/Los Angeles, CA

Halbwachs M 1985a *Das Gedächtnis und seine sozialen Bedingungen*. Suhrkamp, Frankfurt, Germany (French original: *Les Cadres Sociaux de la Mémoire*. Paris 1925)

Halbwachs M 1985b *Das kollektive Gedächtnis*. Fischer, Frankfurt, Germany

Herodotus 1992 *The Histories* (Blanco W (ed.)). Norton, New York/London

Hutton P H 1993 *History as an Art of Memory*. University Press of New England, Hanover, UT

Kirsch J H 1998 Mythos, Geschichte und Geschichtswissenschaft. In: Tepe P et al. (eds.) *Mythologica 6. Düsseldorfer Jahrbuch für interdisziplinäre Mythosforschung*. Verlag Die Blaue Eule, Essen, Germany, pp. 108–27

Koselleck R 1997 Vom Sinn und Unsinn der Geschichte. In: Klaus E M, Rüsen J (eds.) *Historische Sinnbildung. Problemstellungen, Zeitkonzepte, Wahrnehmungshorizonte, Darstellungsstrategien*. Rowohlt, Reinbek, Germany, pp. 79–97

Le Goff J 1986 *Histoire et Mémoire*. Gallimard, Paris

Lewis B 1975 *History—Remembered, Recovered, Invented*. Princeton University Press, Princeton, NJ

Münkler H 1994 Politische Mythen und nationale Identität. In: Frindte W, Pätzolt H (eds.) *Mythen der Deutschen. Deutsche Befindlichkeiten zwischen Geschichten und Geschichte*. Leske & Budrich, Opladen, Germany

Nora P 1984/1992 Entre mémoire et histoire. In: Agulhon M, Boutry P (eds.) *Les Lieux de Mémoire*. Gallimard, Paris, Vol. 1

Nietzsche F 1988 Unzeitgemässe Betrachtungen. Zweites Stück: Vom Nutzen und Nachtheil der Historie für das Leben. In: Colli G, Montinari M (eds.) *Friedrich Nietzsche: Sämtliche Werke. Kritische Studienausgabe in 15 Einzelbänden*. Deutscher Taschenbuch Verlag, Munich, Germany, Vol. 1, pp. 243–334

Oexle O G (ed.) 1995 *Memoria als Kultur*. Vandenhoeck & Ruprecht, Göttingen, Germany

Plumb J A 1969 *The Death of the Past*. Macmillan, London

Ricoeur P 1998 Passé, mémoire et oubli. In: Verlhac M (ed.) *Histoire et mémoire*. Centre Régional de Documentation Pédagogique de l'Académie de Grenoble, Grenoble, France

Schacter D 1999 The seven sins of memory. Insights from psychology and cognitive neuroscience. *American Psychologist* **54**(3): 182–203

Schudson M 1992 *Watergate in American Memory. How We Remember, Forget and Reconstruct the Past*. Basic Books, New York

Vidal-Naquet P 1981 *Les Juifs, la mémoire et le présent*. Maspero, Paris

A. Assmann

History and the Social Sciences

History and the social sciences have developed and maintained complex relationships made up of commonalities and differences as well as divergences and convergences. In principle, both disciplinary areas are concerned with the scientific study of the human lifeworld and the whole range of human behavior, individual and collective, in time and space as well as in the past and the present. Both disciplinary areas originated with the rise of the modern world as a common enterprise to develop systematic, secular, and empirically validated knowledge about the historical-social reality. In this fundamental sense, history can be and should be seen as part of the social sciences or vice versa the social sciences as part of the historical sciences—both constituted by substantially interrelated modes of sociohistorical research (Gulbenkian Commission 1996, Hall 1999).

Despite their fundamental commonality, however, history and the social sciences have followed quite different and diverging epistemological and methodological tracks during their academic institutionalization in Western Europe and North America in the nineteenth and early twentieth centuries. On the one hand, the social sciences (economics, sociology, and political science as well as anthropology) came to be oriented towards the successful epistemological model of the natural sciences and defined themselves primarily as nomothetic and generalizing disciplines. On the other, history predominantly influenced by the humanities and critical against speculative social laws developed traits of an idiographic and particularistic discipline. Against these traditional disciplinary definitions, a converging interdisciplinary cooperation

between history and the social sciences has developed only in the second half of the twentieth century. The social-scientific orientation of history was programmatically formulated in approaches such as the 'new social history,' 'la nouvelle histoire,' 'Social Science History,' or 'Historische Sozialwissenschaft,' developing in leading journals such as *Annales Economies, Sociétés, Civilisations* (1946–), *Past and Present* (1952–), *Journal of Social History* (1967–), *Journal of Interdisciplinary History* (1970–), *Social Science History* (1976–), and *Geschichte und Gesellschaft* (1975–). In a parallel movement, the historical orientation of the social sciences crystallized in new areas like historical sociology, historical economics, historical political science, and historical anthropology. Sometimes even the epistemological identity of history and the social sciences has been stated in a fundamental sense.

More recently, however, this *rapprochement* between history and the social sciences was halted and partly reversed. Within the study and presentation of history a 'revival of the narrative,' a 'new cultural history,' and concepts of the 'historical cultural sciences' have emerged, emphasizing what distinguishes them from the social sciences proper. But it should be stressed that, in a parallel movement, the social sciences have been similarly challenged from within by alternative cultural, interpretative-hermeneutic, and deconstructive-postmodernist epistemologies. Often the so-called 'cultural turn' in history and the social sciences is seen as a replay of the traditional juxtaposition of idiographic and nomothetic positions and conservatively used to reconfirm the traditionally separate institutionalization of both disciplinary areas. But properly understood, this cultural turn can and should be rather viewed as a challenge to the still prevailing principled opposition of generalizing or particularizing disciplinary epistemologies and methodologies. The future task, then, would be to open the historical and social sciences towards a common and multivocal, intra- and interdisciplinary discourse of sociohistorical inquiry.

1. *Disciplinary Divergence*

In its originally institutionalized—but nowadays increasingly challenged—meaning, history is primarily concerned with the study, description, interpretation, and presentation of historical facts or relics from the past, focusing on: idiographic methods, the analysis of historical contexts and temporal change, and the conceptualization of historical phenomena as particular individualities, their inductive interpretation, and narrative presentation (Burke 1992, Zunz 1985). In the opposite—but also increasingly questioned—direction, the social sciences such as sociology, economics, political science, and anthropology have been traditionally understood as concerned with the study, generation, and explanation of social facts in the present emphasizing nomothetic methods, systematic analysis, and theoretical explanation by reference to general laws or social regularities (Monkkonen 1994, Skocpol 1984).

This problematically polarizing division between history and the social sciences developed with the institutionalization of the university systems during the nineteenth century under the impact of the natural sciences. Science, since then, came to be equated primarily with the natural sciences and particularly with its conspicuous inorganic and organic branches of physics and biology. Taking the natural sciences as epistemological models, the social sciences emerged as an attempt to discover the general laws and evolutionary mechanisms of modern society (Lepenies 1988, Gulbenkian Commission 1996). Thus, political or classical economy as developed by Adam Smith or David Ricardo tried to find the general laws of modern capitalist market society or in a critical direction as formulated by Karl Marx to uncover the anatomy of emerging modern capitalism. In the same steps, the early founding fathers of sociology such as Claude-Henri Saint-Simon, Auguste Comte, or Herbert Spencer intended to find the structural and evolutionary laws of modern society. At the same time, with the advent of modern mass politics in the American and French revolutions, the emerging political science as in Alexis de Tocqueville or John Stuart Mill attempted to formulate the general features and bases of modern democratic polities. Finally, with the growing awareness of non-Western societies under the impact of European colonialism and imperialism, anthropology tried to analyze the archaic or 'primitive' societies as structural counterparts and evolutionary pre-stages of modern society.

By contrast, history in view of the contingent flow of historical events and human actions as well as the role of cultural and spiritual forces came to be skeptical of the search for quasi-natural laws of the historical-social world. Instead, history followed the humanities, art, and literature as culturally oriented disciplines that seemed to be closer to the spiritual, creative, and poetic aspects of historical development. Particularly under the impact of German 'historicism,' the epistemological and methodological principles of historical knowledge developed into a radical opposition to the positivist search for historical-social facts. The principle of Leopold von Ranke to narrate history 'wie sie eigentlich gewesen ist' formulated the conviction that the historical method has to grasp the historical reality in its manifold particularities. In epistemological terms as developed by Johann Gustav Droysen and Wilhelm Dilthey, history cannot be subsumed under deterministic-mechanistic laws, but has to be understood in its concrete, experience-near, and life-full quality. On this basis, not only were the historical subdisciplines such as political, legal, cultural, social, and economic history principally viewed as hermen-

eutic-interpretative disciplines, but also economics, sociology, and political science as promoted among others by Gustav Schmoller, Werner Sombart, or Max Weber were primarily conceptualized as cultural sciences rather than as social sciences in order to grasp the individual and collective meanings structuring socioeconomic and political actions and processes. These epistemological and methodological differences and divergences were characteristic for the relationships of history and the social sciences not only in Germany, but in all Western societies. However, through the particular influence of German 'historicism' refuting on ontological grounds the applicability of explanatory social-scientific methods to historical events and cultural individualities, these divergences often developed into a principled opposition (Iggers 1993, Lepenies 1988).

The interdisciplinary bridging of this epistemological, analytical, and methodological opposition only became possible with the increasingly historical orientation of the social sciences as well as the growing social-scientific and sociological orientation of history (Abrams 1982, Lipset and Hofstadter 1970, Skocpol 1984). In the first half of the twentieth century, this converging movement was seldom although of crucial significance in preparing the ground for the later developing convergences between the historical and social sciences. Of lasting influence was, first, Emile Durkheim and the continuation of his approach by Marcel Mauss and Henri Berr, attempting to combine the sociology of social change with anthropological comparisons and historical developments (Raphael 1994, Stoianovich 1976). Of fundamental importance was also, second, the German historical-comparative sociology, particularly in the hands of Max Weber, Otto Hintze, Werner Sombart, and in their continuity also of Norbert Elias, who tried to explain the specificities of the Western civilization, its unique economic, political, and cultural order by historical and universal comparisons (Lepenies 1988). Third, strongly influenced by European sociological traditions, modernization theory developed in the USA as a universal social-scientific approach closely connected with historical-comparative modernization research on the historical prerequisites and developments of modern societies on a world scale.

2. *Interdisciplinary Convergence*

Influenced by these precursors, the historical and social sciences combined programmatically, first in the Annales school (Raphael 1994, Stoianovich 1976) and then later in the American Social Science History approach (Tilly 1981), some proponents of English Marxist historiography (Kaye 1984) and in their emulation in the West German 'Historische Sozialwissenschaft' (Kocka 1977, Wehler 1980). This interdisciplinary convergence led to different forms of 'social-scientific history' or 'historical social science'—despite many national particularities. They materialized in the 1960s and 1970s in a parallel movement in the Western societies (Iggers 1993). In critical opposition to the received tradition of narrative and exclusively interpretive history, the major common methodological aims of social-scientific history or historical social science can be summarized in the following principles. First, social-scientific historical approaches should be oriented to a structural and processual perspective on the various (economic, sociopolitical, and cultural) dimensions of macrohistorical change. Second, they should shift historical analysis from a 'top-down' elite perspective to a 'bottom-up' popular or people's history. Third, social-scientific historical approaches should give up the traditional emphasis on political and intellectual history and turn towards economic and social history. Fourth, they should make systematic use of the available social-scientific, theoretical, comparative, and quantitative methods in history and favor systematic explanation instead of hermeneutic interpretation. Finally, this movement to social-scientific history or historical social science was often motivated by a critique of traditional elitist-conservative historiography and a commitment to emancipatory, democratic, or enlightened values.

The common methodological aims of social-scientific history or historical social science, at the same time, were colored and refracted by the respective national intellectual and cultural environments. In France, the 'nouvelle histoire' promoted within the long established interdisciplinary tradition of the Annales school (Raphael 1984, Stoianovich 1976) was characterized by a rather balanced multilevel (*longue durée, conjonctures, évènements*) and multidimensional (economy, social structure, and civilizational mentalities) mode of historical analysis. In England, a strong neo-Marxist influence shaped the rise of economic and social history with a particular emphasis on social inequality and class history (Kaye 1984). In the United States the interdisciplinary social science history was influenced by the development of historical-comparative multidimensional modernization research and the special attraction of econometric, statistical, and social research methods for the quantification of historical processes (Tilly 1981). In Germany, Marxian and Weberian traditions of scholarship combined with elements of Western modernization theories and led to 'Historische Sozialwissenschaft' and 'Gesellschaftsgeschichte,' in which the analysis of social inequality and political power with a particular emphasis on the German 'special path' were treated (Kocka 1977, Wehler 1980). At the same time, these national variations reflect not only differing intellectual and theoretical traditions, but also a different emphasis on the use of qualitative and quantitative methods in historical-social inquiry. This is also indicated by the paradigmatic oscillations between the concepts of

(a) historical social science, (b) structural history, and (c) social-scientific history.

'Historical social science' ('Historische Sozialwissenschaft') serves as a loosely defined summary category. Still, the historical-comparative analysis of large-scale and multidimensional processes of social change from traditional life-worlds to modern society has been at the center. In particular, the concepts of historical-comparative modernization research, historical sociology, societal history, and synthetic structural history, with different theoretical and methodological orientations, belong to this category. In the French Annales school in a mixture of Durkheimian and Marxian traditions, the comparative analysis of feudalism by Marc Bloch and the social history of capitalism by Fernand Braudel (to be followed by Immanuel Wallerstein's history of the modern world system) are important examples. Among English Marxist historians, the debate on the transition from feudalism to capitalism in a line from Maurice Dobb to Perry Anderson and the social history of modern classes, developed in different directions by Eric Hobsbawm and Edward Thompson, carried the field. In the United States, modernization research guided by a mixture of Max Weber's comparative sociology and Talcott Parsons' structural functionalist theory of social change was of major importance, resulting in path-breaking studies like Neil Smelser's analysis of social change in the English industrial revolution; Shmuel Eisenstadt's civilizational comparison of empires; Barrington Moore's historical explanation of the development of twentieth-century dictatorship and democracy; Tilly's edited volume on the formation of national states in Western Europe; and Reinhard Bendix's study on corresponding historical changes in the legitimacy of political domination (see as an overview Skocpol 1984). In Germany, the modernization paradigm in a mixture of Weber and Marx directed Ralf Dahrendorf's sociological analysis of Germany's troublesome process of democratization or Hans-Ulrich Wehler's (1987–95) and Jürgen Kocka's (1993) comparative social-historical investigations of the German 'special path' and the role of the social classes within it.

'Structural history,' again, has had different meanings with different outlooks. It materialized in innovative areas of a variety of 'new histories' along the lines between the main subsystems of modern society and the related social-scientific disciplines. In the 'New Economic History' of the 1960s, the application of economic concepts and theories has given rise to the history of economic growth, technology, capitalist markets, and institutions; here, David Landes' (1969) *Prometheus* or Sidney Pollard's (1981) *Peaceful Conquest* are examples of a qualitative type, whereas Gianni Tonioli's (1991) comparative history of European industrialization—debating Alexander Gerschenkron's theory of economic backwardness—represents a quantitative type of economic history. In the 'New Social History' in the 1960s and 1970s, sociological concepts were used in the historical analysis of demographic development, social structure, and social mobility; social groups, occupational structures, and social classes; social institutions, associations, and interest groups; as well as social movements, social protest, and political mobilization (Kocka 1977). This was combined with the application of political-scientific approaches to the history of political modernization, state formation, nation building, the development of democratic regimes and party systems, as well as political power and contestation (Tilly 1984). Further, social-scientific approaches influenced by social anthropology have also been applied to cultural phenomena and developments such as family and socialization, religion and secularization, as well as value systems and mental structures (see different essays in Rürup 1977).

The concept of 'social-scientific history' has been used primarily for the application of rigorous quantitative research methods in structural or social history (Landes and Tilly 1970, Schröder 1994). Not always, but often it has been equated with historical social research or the systematic quantitative operationalization and measurement of historical-social variables. In addition, this often combined with a deductive application of theoretical models to historical data and their explanation through statistical-mathematical and increasingly computer-assisted methods. Also this type of social-scientific historical research followed, either in close cooperation or in separate organization, the main areas of the 'new histories.' Particularly, the 'New Economic History' has been suited to cliometric, statistical, and econometric research methods as for instance in Paul Bairoch's and Maurice Lévy-Leboyer's (1981) analysis of economic growth disparities since the industrial revolution. But also in the new social history, social-demographic family, mobility, and migration studies such as Richard Wall, Jean Robin, and Peter Laslett (1983) as well as time-series of social protest such as Charles Tilly, L. Tilly, and R. Tilly (1975) are prominent examples. In the new political history, the historical change of voting behavior, in particular, became an important field of historical social research (e.g., Best and Thome 1991). But also in the cultural history of mentalities as carried through particularly in the Annales tradition, the application of quantitative methods were fruitfully developed (e.g., Shapiro and Markoff 1998).

3. *Interdisciplinary Challenges*

Successful as these different new historical subdisciplines in their qualitative and quantitative orientations have been, the guiding vision of social-scientific approaches to history has been the combination of the various structural histories in a unifying societal

history, 'histoire totale' or 'Gesellschaftsgeschichte' (Iggers 1993, Wehler 1980). This vision follows the epistomological ideal of merging the historical sciences and the social sciences in a common paradigm—as essentially expressed in the paradigmatic concepts of social science history, historical sociology, or 'historische Sozialwissenschaft' (Abrams 1982, Smith 1991, Wehler 1980). With this unifying vision, however, the disciplinary traditions and methodological differences between historicized conceptions of the social sciences and social-scientific approaches to history become easily blurred. On the one hand, historical approaches in the social sciences basically proceed within a generalizing frame of reference (Monkkonen 1994). For instance, historical economics follows the basic idea of uncovering time/space bound economic laws or regularities (Kindleberger 1990, Monkkonen 1994, Wallerstein 2000). Also historical sociology is primarily oriented to the general features of time/space bound social processes, patterns, and figurations (as exemplary studies see Eisenstadt 1996, Mann 1986/1993, Smelser 1992). In parallel, historical approaches in the political sciences concentrate on general features of political development (Monkkonen 1994). Also, historicized versions of anthropology are interested in the general evolutionary patterns of human behavior (Monkkonen 1994). On the other hand, social-scientific approaches to history as multidimensional structural or combined societal history usually remain in the confines of a narrative frame of reference—even when applying rigorously social-scientific theoretical concepts and research methods. Thus, in spite of all interdisciplinary convergence, there have been limits to a full merger between historically oriented social sciences and social science oriented history.

A second *problematique*, expressed in the oscillating meanings of historical social science, structural history, and social-scientific history, concerns the relationships between theory and method. Structural approaches to history often use social-scientific concepts in order to systematize, interpret, and explain long-term historical processes in one or another societal dimension. Social-scientific approaches to history are often oriented to theoretical models in order to explain deductively quantified historical-developmental data. Approaches in historical social science and related synthetic orientations to history often use multidimensional theories of social change in order to explain causal connections between the various societal dimensions. Many varieties of historical social science, structural history, and social-scientific history use theories and concepts, deductive and inductive modes of explanation or interpretation, as well as qualitative, quantitative, and comparative methods of analysis. In an effort to systematize these methodological strategies in historical-social inquiry, Theda Skocpol proposed to distinguish model-theoretical, causal-analytical, and interpretive modes of explanation (Skocpol 1984, pp. 356–91). Charles Tilly, concentrating on the types of comparison and the number of cases and variables involved, differentiated between individualizing, generalizing, encompassing, and variation-finding forms of comparison (Tilly 1984, 84ff.). Charles Ragin (2000), attempting to bridge qualitative and quantitative modes of analysis, distinguished between case-analytical and variation-oriented methodological strategies. Fruitful as these methodological distinctions are, particularly the methodological relationships between explanatory and interpretive research strategies, the theoretical sources and analytical consequences of these methodological distinctions for sociohistorical inquiry still remain unexplored (Hall 1999, p. 173).

A third increasingly salient but divisive *problematique* concerns the role of culture in the differing modes of historical social science, structural history, and social-scientific history. Not only in practice, but also often programmatically, social-scientific approaches to sociohistorical inquiry tend to favor the analysis of the economic, social, and political dimensions and to limit the analysis of culture to value systems or cultural institutions. The contents and forms of culture in their constitutive or causal role in history are often neglected and the corresponding hermeneutic or deconstructive methods to interpret their meanings, symbols, values, and identities are often seen as secondary. Against these prevailing tendencies of social-scientific approaches to culture, there emerged an increasingly influential challenge by the 'new cultural history' or cultural-scientific versions of history (Hunt 1989, Dirks et al. 1994, Mergel and Welskopp 1996). This 'cultural turn' consists of different countercurrents against the social-scientific vision of a systematic, generalizing, and explanatory structural and societal history. First, the revival of the narrative insists on the interpretive, literary, and narrative tasks of history (Stone 1979). Second, the development of micro-history or 'Alltagsgeschichte' emphasizes the complexity of life-worlds as such or as microfoundations of historical macrostructures (Luedtke 1989). Third, the 'linguistic turn' presupposes the all-pervasiveness of language in the social world, thus stressing the analytical task of deconstructing discourses and decoding linguistic-symbolic structures (Hunt 1989). Fourth, the 'cultural turn' is consciously oriented to those social groups such as women, ethnic minorities, or the underclass, seen as neglected by the predominating social-scientific approaches to history (e.g., *History Workshop Journal* 1976– or *Radical History* 1975–). Finally, some authors are influenced by a postmodernist perspective emphasizing the fragmentation and decentering against the structuring and ordering of the social world (Iggers 1993). The result is not only the generation of a whole new set of historical subdisciplines such as women's and gender history, ethnic and minority history, regional and local history as well

as the history of identities and memory, but also an influential shift from social-scientific history to cultural-scientific history (see however Bonnell and Hunt 1999).

In a sense, the 'cultural turn' signifies a revitalization of historicism and cultural-scientific approaches to history—re-emphasizing qualitative, interpretative, ethnographical, and hermeneutic methods against systematic-explanatory orientations and, with it, reversing the causal or structural determination of history from 'society' to 'culture.' At the same time, this theoretical and analytical movement reflects a parallel shift in the social sciences from social-structural to institutional and cultural paradigms (Skinner 1985, Matthes 1992, McDonald 1996). For instance, in economics and economic sociology the new institutionalism has gained wide recognition, but also cultural neo-Weberian approaches reappear (North 1981, Monkkonen 1994, Swedberg 1998). In various domains of sociology as well, the 'new institutionalism' and various forms of a 'new cultural sociology' become influential (Monkkonen 1994, Crane 1994). Further, in political science the traditional institutional approaches become challenged by cultural paradigms (Monkkonen 1994). All this is influenced by a crucial shift in anthropology from social and structural anthropology to various forms of interpretative, ethnographic, historical, and cultural anthropology (Monkkonen 1994). As a result, social-scientific history is not merely confronted with systematic, generalizing, and explanatory forms of the social sciences, but also with individualizing, interpretative, and reflexive varieties of the cultural sciences. In other words, new opportunities of interdisciplinary cooperation emerge.

4. *Interdisciplinary Perspectives*

The interdisciplinary cooperation and convergence of history and the social sciences have opened up novel and dynamical fields of sociohistorical research. At the same time, the *idées directrices* of that historical-social-scientific convergence have often privileged quasi-natural-scientific epistemologies and methodologies. As a consequence, generalizing modes of sociohistorical inquiry stand in tension with particularizing modes; structural and macro-analytical theories used in historical-social analysis conflict with action and micro-analytical theories; causal explanation is often limited to the application of theoretical models and causal-analytic methodologies in opposition to interpretation and hermeneutic strategies; and, accordingly, there are tensions between qualitative and quantitative methods, including generalizing and particularizing methods of comparison. These tensions are revealed, on the one hand, within the social-scientific convergence of history and the social sciences in the paradigmatic oscillations between the guiding concepts of historical social science, structural history, and social-scientific history. These tensions develop, on the other hand, within the wider historical sciences, in the often polarized opposition to renewed historicist and culturalist, narrative and hermeneutic or deconstructivist modes of historical-social inquiry. At the same time, these theoretical and methodological oppositions and polarizations are not solely characteristic for the historical sciences. Parallel disciplinary developments can also be observed in the social sciences and the widespread and often sharp oppositions to cultural-scientific approaches in the various social-scientific disciplines.

These epistemological, theoretical, and methodological oppositions and conflicts have often led into a stalemate in both disciplinary areas of the historical and social sciences between renewed forms of narrative-interpretative and anthropological-cultural history in contradistinction to the nomothetic social sciences. However, these oppositions in the historical as well as the social sciences can and should be used to re-energize a wider intra- and interdisciplinary cooperation between the historical, social, and cultural sciences. These renewed cooperative efforts should concentrate on the crucial epistemological, theoretical, methodological, and analytical dimensions in sociohistorical inquiry within and across the established academic disciplines: on the epistemological level, combining social-scientific and cultural-scientific as well as generalizing and particularizing modes of knowledge; on the theoretical level confronting systematically social-structural and cultural approaches to historical-social reality; on the analytical level, mediating between the social-structural, institutional and cultural dimensions as well as between the global, national, regional and local macro-/microdimensions in sociohistorical inquiry; and on the methodological level, clarifying the relationships between qualitative and qualitative as well as historical, social-scientific, and cultural-scientific comparative research strategies (Bourdieu 1995, Haupt and Kocka 1996, Spohn 1998). All these renewed epistemological, theoretical, methodological, and analytical efforts of an inter- and intradisciplinary cooperation should not be done in segmented isolation, but in a multidisciplinary and multivoiced context of practiced sociohistorical inquiry. Such an opening of the historical and social sciences, however, is not only a matter of an intellectual intra- and interdisciplinary cooperation, but also demands an institutional cross-cutting between and across the segmented academic disciplines of the historical and social sciences (Gulbenkian Commission 1996).

See also: Anthropology and History; Cultural History; Demography, History of; Econometrics, History of; Economic History; Economics, History of; Family and Kinship, History of; Gender and Feminist Studies in History; Historical Sociology; History: Forms of Presentation, Discourses, and Functions; History:

Theories and Methods; Marxist Social Thought, History of; Social History; Weberian Social Thought, History Of

Bibliography

Abrams P 1982 *Historical Sociology*. Open Books, Shepton Mallet, UK

Bairoch P, Lévy-Leboyer M (eds.) 1981 *Disparities in Economic Development Since the Industrial Revolution*. St. Martin's Press, New York

Best H, Thome H (eds.) 1991 *Neue Methoden der Analyse historischer Daten*. Scripta Mercaturae Verlag, St. Katharinen

Bonnell V E, Hunt L (eds.) 1999 *Beyond the Cultural Turn: New Directions in the Study of Society and Culture*. University of California Press, Berkeley, CA

Bourdieu P (ed) 1995 Histoire sociale des sciences sociales. *Actes de la Recherche en sciences sociales* **106/107**

Burke P 1992 *History and Social Theory*. Polity Press, Cambridge, UK

Crane D (ed.) 1994 *The Sociology of Culture. Emerging Theoretical Perspectives*. Blackwell, Oxford, UK

Dirks N B, Eley G, Ortner S B (eds.) 1994 *Culture/Power/History: A Reader in Contemporary Social Theory*. Princeton University Press, Princeton, NJ

Eisenstadt S N 1996 *Japanese Civilization. A Comparative View*. The University of Chicago Press, Chicago

Gulbenkian Commission on the Restructuring of the Social Sciences 1996 *Open the Social Sciences: Report of the University Press*. Stanford University Press, Stanford, CA

Hall J R 1999 *Cultures of Inquiry. From Epistemology to Discourse in Sociohistorical Research*. Cambridge University Press, Cambridge, UK

Haupt H-G, Kocka J (eds.) 1996 *Geschichte und Vergleich*. Campus Frankfurt

Hunt L (ed) 1989 *The New Cultural History: Essays*. University of California Press, Berkeley, CA

Iggers G G 1993 *Geschichtswissenschaft im 20. Jahrhundert. Ein kritischer Überblick im internationalen Zusammenhang*. Vandenhoek und Ruprecht, Göttingen

Kaye H J 1984 *The British Marxist Historians An Introductory Analysis*. Cambridge University Press, Cambridge, UK

Kindleberger C P 1990 *Historical Economics. Art of Science?* University of California Press, Berkeley, CA

Kocka J 1977 *Sozialgeschichte; Begriff, Entwicklung, Probleme*. Vandenhoek und Ruprecht, Göttingen

Kocka J (eds.) 1993 *Bürgertum und bürgerliche Gesellschaft im 19. Jahrhundert*. Beck, München

Landes D S 1969 *The Unbound Prometheus: Techological Change and industrial Development in western Europe from 1750 to the Present*. Cambridge University Press, Cambridge, UK

Landes D S, Tilly C 1971 *History as Social Science*. Prentice-Hall, Englewood Cliffs, NJ

Lipset S M, Hofstadter R (eds.) 1970 *Sociology and History: Methods*. Basic Books, New York

Lepenies W 1988 *Between Literature and Science: The Rise of Sociology*. Cambridge University Press, Cambridge, UK

Luedtke A (ed.) 1989 *Alltagsgeschichte: Zur Rekonstruktion Historischer Erfahrungen und Lebernweisen*. Campus, Frankfurt

Mann M 1986/1993 *The Sources of Social Power*. Cambridge University Press, Cambridge, UK, 2 Vols

Matthes J (ed.) 1992 *Zwischen den Kulturen? Die Sozialwissenschaften vor dem Problem des Kulturvergleichs*. Vandenhoek und Ruprecht, Göttingen

McDonald T J (ed.) 1996 *The Historic Turn in the Human Sciences*. Michigan University Press, Ann Arbor, MI

Mergel T, Welskopp T (eds.) 1997 *Geschichte zwischen Kultur und Gesellschaft. Beiträge zur Theoriedebatte*. Beck, München

Monkkonen E H (ed.) 1994 *Engaging the Past. The Uses of History. Across the Social Sciences*. Duke University Press, Durham, NC

North D C 1981 *Structure and Change in Economic History*. Norton, New York

Pollard S 1981 *Peaceful Conquest: the Industrialization of Europe 1760–1970*. Oxford University Press, Oxford, UK

Ragin C C 2000 *Fuzzy-set Social Science*. University of Chicago Press, Chicago

Raphael L 1994 *Die Erben von Bloch und Febvre*. Klett und Cotta, Stuttgart

Rürup R (ed.) 1977 *Historische Sozialwissenschaft: Beitr. zur Einf in d. Forschungsprais*. Vandenhoek and Ruprecht, Göttingen

Schröder C 1994 *Historische Sozialforschung*. Minerva, St. Katharinen

Shapiro G, Markoff J 1998 *Revolutionary Demands: A Content Analysis of the Cahiers De Doléances of 1789*. Stanford University Press, Stanford, CA

Skinner Q (ed.) 1985 *The Return of Grand Theory in the Human Sciences*. Canto, Cambridge, UK

Skocpol T (ed.) 1984 *Vision and Method in Historical Sociology*. Cambridge University Press, Cambridge, UK

Smelser N J 1991 *Social Paralysis and Social Change: British Working-Class Education in the Nineteenth Century*. University of California Press, Berkeley, CA

Smith D 1991 *The Rise of Historical Sociology*. Temple University Press, Philadelphia, PA

Spohn W (ed.) 1998 Kulturanalyse und vergleichende Forschung *Comparativ. Leizpiger Beiträge zur Universalgeschichte und vergleichenden Gesellschaftsforschung* **8**(1)

Stoianovich T 1976 *French Historical Method: The Annales Paradigm*. Cornell University Press, Ithaca, NY

Stone L 1979 The revival of the narrative: Reflections on a new old history. *Past and Present* **85**: 3–24

Swedberg R 1998 *Max Weber and the Idea of Economic Sociology*. Princeton University Press, Princeton, NJ

Tilly C 1981 *As Sociology Meets History*. Academic Press, New York

Tilly C 1984 *Big Structures, Large Processes, Huge Comparisons*, Russel Sage Foundation, New York

Tilly C, Tilly L, Tilly R 1975 *The Rebellious Century 1830–1930*. Harvard University Press, Cambridge, MA

Toniolo G Sylla R (ed.) 1991 *Patterns of European Industrialization in the 19th Century*. Routledge, London

Wall R, Robin J, Laslett P (eds)1983 *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK

Wallerstein I M 2000 *The Essential Wallerstein*. New Press, New York

Wehler H-U 1980 *Historische Sozialwissenschaft und Geschichtsschreibung*. Vandenhoek and Ruprecht, Göttingen

Wehler H-U 1987–1995 *Gesellschaftsgeschichte Deutschlands*. Beck, München, 3 Vols.

Zunz O 1985 (ed.) *Reliving the Past: the World of Social History*. University of North Carolina Press, Chapel Hill, NC

W. Spohn

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

History: Forms of Presentation, Discourses, and Functions

History as an institutionalized, academic discipline is a continuation of extradisciplinary forms of dealing with the past and, therefore, has never enjoyed a monopoly on its representation. Narrative representations of origins—be it in the form of myths, legends, religions, or ideologies—are found in most cultures. These different forms of representing the past offer collectives an orientation in time: they give a meaning to the present by locating the present on a time axis between a future and a past.

Next to the variety of disciplinary forms of historical representation (Stern 1970, Kelley 1991, Lorenz 1997), these extradisciplinary representations of history still exist, often competing with others and disciplinary history. The problem of how to demarcate the extradisciplinary from the disciplinary ways of dealing with the past, has therefore haunted history as a discipline from its very beginning. Traditionally, the discipline's identity has been based on its truthclaim, its critical methods, and its archival foundation (see *History: Theories and Methods*). This article will only deal with the disciplinary representations of the past in the West. It will only take extradisciplinary representations into account in order to cast light on the origins and the official borderlines of the discipline (Fain 1970, McNeill 1986).

1. Historical Representation and its Function as a Historical Problem

Although the concept of representation has a clear-cut meaning in philosophy—that of an image, in which the structure of the represented phenomenon is maintained (Mormann 1997)—in history the term is usually used as an equivalent of concepts like image or construction. In titles like 'The representation of Hitler in the US press, 1933–1945' or 'The representation of the Black Woman in the literature of the Antebellum South,' historians emphasize the constructed and interpretative character of historical reconstructions: history cannot be directly presented, but is *re*presented by the historian in language. This emphasis has roots in an awareness that historical representations are not mirror images of history itself. It is usually accompanied by an awareness of the fundamental plurality of historical reconstructions. At the same time it is sometimes accompanied by an awareness of the practical functions (or the politics) of representation.

For the consciousness of this 'representational' character of history, two developments were crucial. The first development was the perception of history as an object of knowledge, independent of its individual representations. According to Koselleck this development originated in Europe, in the mid-eighteenth century, when history as a process—*historia*—was first distinguished from the traditional *historiae*, i.e., the stories told about the past. History could only be conceived of as an independent object of knowledge when it was no longer regarded as a binding tradition or as a practical craft belonging to literature or rhetoric, i.e., as the art of telling (his)stories in order to persuade a public, as had been the case in Europe from classical antiquity until mid-eighteenth century. History's functions only developed into a problem after the traditional idea, that past experiences were relevant for present purposes and histories could teach you by example what to do and what not to do (*historia magistra vitae*), began to loose its plausibility.

The second development, that promoted the awareness of the representational character of history, was the 'discovery' of its perspectival character during the Enlightenment. Although the possibility of multiple stories about the same events was known already —and a central concern of rhetorics—Enlightenment historians were the first to recognize explicitly that the historical representations necessarily vary with the perspectives of the historians and with the development of time. The Enlightenment also stimulated the tendency to identify cognitive, scholarly activity with empirical disciplines, which was a consequence of the competition of the *Geisteswissenschaften* with the modern natural sciences (Scholtz 1991, Momigliano 1990). Koselleck dates the fundamental change in the relationship between Europeans and their sense of the past halfway through the eighteenth century; according to Pocock these changes had already started in the seventeenth century (Koselleck 1979, Pocock 1961). The traditional idea of the continuity between the past, the present, and the future was gradually replaced by the experience of permanent change, so typical of modernity (Scholtz 1991, Hardtwig 1990). The past thus gradually turned into a 'foreign country' (Plumb 1973, Lowenthal 1985, Rüsen 1983–9).

Early nineteenth century, this experience of difference between the past and the present would lead to the association of history with the notions of nonrecurrence and nongeneralizability, later expressed in the concepts of uniqueness and individuality (see *Historicism and History: Theories and Methods*). History thus came to be conceived of not just as change through time—which could be cyclical and thus repeatable—but as *change as a process of irreversible and nonrepeatable development*.

With the emergence of this specific new sense of the past, the stage was set for the development of history and the social sciences into separate disciplines. The object of history was increasingly identified with nonrecurring phenomena in the human past, while the social sciences were identified with recurrent phenomena in the human present. Because recurrence is a precondition for generalization and thus for (structural and processual) theories, history and theory gradually turned into opposites, excluding one another. Thus the idea was born that the forms of

representation in history differed fundamentally from those in the social sciences. While historical representation was usually characterized as time-oriented and narrative, social scientific representational forms were characterized as analytical and theoretical (see *History and the Social Sciences*, Meran 1985, Burke 1993, Lorenz 1997). The modern question of narrative in history was thus born in the eighteenth century as a concomitant of the rise of modernity.

2. The Classical Roots of the Modern Debate

2.1 The Aristotelian Heritage

Both the Enlightenment conception of history and its critique in Herder and in later, Romantic thinkers—the two conceptions of history that still define the terms of the modern debate—must be viewed in the context of Aristotelian philosophy, because the Aristotelian definitions and distinctions of disciplines were still very influential in the eighteenth century (Aristotle 1963). In Aristotle's (ca. 384–322 BC) philosophy, there is a basic distinction between empirical knowledge of singular facts and universal knowledge of general facts. Concrete, empirical knowledge just formulates the contingent and particular characteristics of its object, while theoretical knowledge formulates its necessary and universal characteristics. These latter characteristics are viewed as recurrent and unchanging and were therefore limited to the domain of nature. Theoretical knowledge was therefore limited to the domain of nature too. Theoretical knowledge, according to Aristotle, does not only inform us *what* reality looks like, but also *why* this is the case: theoretical knowledge is always explanatory knowledge, because it informs us about the causes of what happens. Therefore, explanations in science are causal explanations and are ideally based on theoretical knowledge. This idea would dominate later philosophy of science in the form of positivism from Auguste Comte to Carl Hempel (see *History: Theories and Methods*; Meran 1985).

Within Aristotelian philosophy the domain of human activity or culture (praxis) held a fundamental different status from nature, because contingency, particularity, and change characterize the first domain. This domain, therefore, was not open to theoretical knowledge and to scientific explanation in principle ('*de singularium non est scientiae*,' as Duns Scotus wrote in the thirteenth century). History, according to the Aristotelian system, belonging to the domain of culture, could never be regarded as a science and was located in the domain of practical knowledge, more akin to poetry (including literature).

Although both poets and historians dealt with the same phenomena, i.e. human actions and events (albeit respectively in verse and in prose), history only dealt with what was real and poetry with what was imagined. The historian, according to Aristotle, 'describes what happened' while the poet 'describes what can happen' (Aristotle 1963). This distinction was later codified in the distinction between the *res factae* [Latin: *res* = affairs; *factae* = done] or 'things done,' which was the domain of history, and *res fictae* [Latin: *fictae* = imagined], which only concerned poets.

The duty of the historian to stick to the facts and to tell the truth, the whole truth and nothing but the truth, derived mainly from the model of Thucydides. It received its canonical formulations in Cicero (ca. 106–43 BC) and Lucian (ca. 125–200 AD), who also linked truth in history directly to impartiality by avoiding moral blame and envy (*sine ira et studio*). The German historian Leopold von Ranke (1795–1886) would only repeat these classical ideals in the nineteenth century, with great force and effect (Iggers 1997).

According to Aristotle, the poet enjoyed a 'constructive liberty' not shared by the historian. While the unit of the historian is defined by what had happened to specific subjects at a specific time at a specific place, the unit of the poet was not defined in such a way. In poetry actions and events are put together by a unit of meaning, that is by constructing a whole with a beginning, middle, and an end, held together by a plot. The parts of the poetic unit all fulfill a function in—i.e. derive their meaning from—this whole. Because of its constructive quality, Aristotle held that poetry was more philosophical than history. This characterization was also based on the idea that poetry—for instance, tragedy—was concerned with general problems—types—and not only with particulars, like history.

The Aristotelian definition of history would define the parameters of the debates on historical representation up to the present. The opposition of nature and culture and the identification of nature with necessity and repetition, and culture with contingency and change, has ever since defined the methodological battlefields in philosophy of history. This opposition linked nature to (universal) theory while excluding (particular) history from theory *a priori*. As influential has been the opposition of poetry and history and the identification of history with the domain of the factual in opposition to the domain of fictionality (see *History: Theories and Methods*). Last but not least the Aristotelian opposition of historical units and poetical units has put the relationship between chronological and narrative ordering of events on the agenda of philosophy of history, i.e. the relationship between chronology and narrative. This opposition raised the question what narrative explanations are and how they relate to causal explanations.

2.2 The Rhetorical Tradition

Another fundamental root of the modern debate on historical representation can be traced back to classical rhetoric (Koselleck et al. 1982). Accounts of the past

were dealt with by classical authors like Aristotle, Cicero, Lucianus and Quintilianus (ca. 35–96 AD) primarily in the context of rhetoric, especially forensic rhetoric, i.e., rhetoric used to persuade a judicial court. So, the problem how to compose and write an account of the past was primarily dealt with as a problem in a communicative context and not as a cognitive problem, as is the case since the nineteenth century. Therefore the question how to produce a convincing history loomed larger than how to write a truthful history—which in turn explains why rhetoric later acquired a doubtful reputation until it was officially expelled from the scholarly domain by Descartes only to return late twentieth century in the train of postempiricism and post-positivism (see *History: Theories and Methods* and McCloskey and Megill 1987). The origin in forensic rhetoric also explains why the domain of history was identified with the domain of human action.

Now to be convincing, according to rhetoric, truth may help, because Plato argued that true arguments are the most convincing. Quintilianus, however, later pointed out that truth as such was neither a necessary nor a sufficient condition for being convincing. Being convincing is directly related to being perceived as probable rather than as being factually true, because some truths are hard to believe. In order to be convincing, according to rhetoric, past action primarily must be explained in (motivational) terms understandable and plausible for the audience addressed. Therefore, the reconstruction of past actions was basically conceived of in the rhetorical tradition as an act of translation of factual information by a speaker, who has a specific aim and who addresses a specific audience.

The rhetorical emphasis on probability instead of on truth was a consequence of its focus on communication and the awareness that factual truth is not a communicative category. It did not imply any plea for blurring the distinction between facts and fiction in history, nor the idea that truth in history fulfilled no practical function. On the contrary, the truth of the past was sought after, because only true accounts of past actions were useful as *exempla* for later generations (Momigliano 1990, Hay 1977). History was thus usually regarded as inherently pragmatic for some 20 centuries until the presuppositions of this pragmatic conception lost their plausibility in the second half of the eighteenth century.

3. The Modern Debate on Historical Representation

3.1 The Problem of Narrative

Basically, the theories of historical representation developed since the eighteenth century can be classified on basis of their relationship with the Aristotelian and the rhetorical characterizations of history. There is a genus of theories, that can be seen as a continuation of Aristotelian theory, and there is a genus of theories, that opposes Aristotelian theory. Both genus contain several species.

The first genus consists of theories that have transformed the Aristotelian analysis of poetical or literary narrative into theories of historical narrative. According to these theories, historical narratives are structured by plots (P. Ricoeur, P. Veyne, H. White, D. Carr), colligatory concepts (W. Walsh), configurational ideas (J.-G. Droysen, L. Mink), or metaphors (F. R. Ankersmit), This view implies that historical narratives are not structured by a chronological order; instead they are structured by linguistic instruments, which synthesize chronologically disparate elements into a synchronic unity (as, e.g., is done by concepts like 'The Enlightenment,' 'feudalism,' or 'imperialism'). Narratives, thus, are fundamentally different from chronologically ordered historical representations, such as chronicles and annals. The fundamental difference between literary and historical narrative, according to most of these narrative theories, is the difference between fiction and fact: while the construction of literary narratives is only restricted by what is plausible, historical narratives are restricted by what is factually true. Therefore, history could and did claim a scholarly status in being empirical or a *Wissenschaft*.

This position was first developed by the Historical School from Niebuhr and Ranke onwards and is philosophically legitimized by historicism that posits a fundamental connection between the narrative form of representation and the notions of time and of individuality. The basic idea is that narrative is the linguistic medium, which shows how a phenomenon (a person, nation, idea, etc.) develops its defining characteristics—its identity or individuality—in time, just as a biography shows how an individual has developed its identity in time (Lübbe 1977, Veyne 1978, Ankersmit 1994). Although most narratives appear to follow the natural flow of time, it is, in fact, the other way around: the central idea of a narrative (e.g., the Enlightenment, the Age of Revolutions, etc.) determines which temporal units and orders (periodization, etc.) are the relevant ones for the narrative in case. Narrative time is thus always constructed time. Droysen formulated this insight in 1857 (Droysen 1977).

Within this first genus of narrative theories, there are several species which show important differences. The first species, exemplified by Louis Mink, Herman Lübbe, and Frank Ankersmit, has developed the argument that explanation by narrative is *sui generis* and is logically independent of causal explanation. All advocates of this species hold that the narrative mode of representation is explanatory in itself.

The second species, exemplified by Arthur Danto and Jörn Rüsen, holds that explanation by narrative,

although different from lawlike explanation, is based on causal explanation nevertheless. The narrative mode of representation is analyzed here as a set of causal chains. A third species of narrative theories, exemplified by Paul Ricoeur, David Carr, and Paul Veyne, has developed the argument that historical explanation has a narrative form and a plot structure. Ricoeur and Carr, moreover, argue that human experience itself is narratively structured and that there is an *a priori* fit between human experience and narrative explanation. Both individuals and communities can be said to have the same narrative form of self-constitution, because their identities can only be represented in the form of narratives.

Whatever their differences, all of these three species of narrative theories regard narratives primarily as explanatory schemes. All three try to answer the Aristotelian question how narratives explain, if not straightforward causally.

A fourth species of narrative theories is not primarily focused on issues of explanatory logic, but on the rhetorical problem why narratives are regarded as plausible and convincing. Ann Rigney and Hayden White in his *Metahistory* exemplify this type of narrative theory. First, White argues that all historical narratives fulfill an ideological function, aimed at specific audiences. Second, he argues that historical narratives contain one of four literary plot structures: comedy, tragedy, satire, or romance. Third, he argues that all history can be emplotted freely in these four manners. Narratives are accepted by an audience as probable and as plausible, according to White, when the narrative plot structure and its ideological implication are recognized and accepted as such.

The acceptability of historical narratives, so White, is thus relative to audiences and relative to ideological aims and not only dependent on the truth of the factual statements in the narrative. In this respect, White's theory has a rhetorical ring, although his fixation on plot structures is at the same time Aristotelian. However, on the issue of narrative truth, White deviates fundamentally from both Aristotelian and rhetorical presuppositions. Contrary to Aristotle, he blurs the borderline between fact and fiction in history by stating that historical narratives are 'fictions of factual representation.' Consequently, according to White, there are no epistemological grounds for preferring one historical representation over another, only political and aesthetic ones. So, along with rhetoric, he holds that plausibility bears no necessary relation to truth. However, contrary to both Aristotelian and rhetorical theory, White severs the crucial link between research and composition of historical narrative: the composition of narratives is not treated as following from—and restricted by—research, but as a fully independent process (Lorenz 1998).

The second genus of theories of historical representation somehow opposes Aristotelian theory. Historically this genus derives from eighteenth century Enlightenment conceptions of history. Contrary to Aristotle, Enlightenment historians like Voltaire, Condorcet and Chladenius held that history could be both theoretical—in the sense of containing knowledge of a general character—and scientific—in the sense of being empirical. 'Theoretical' or 'hypothetical' history was thus claimed as a legitimate 'scientific' project, in which 'theory' and empirical knowledge could be integrated. Theoretical knowledge was no longer exclusively identified with the domain of nature and no longer exclusively interpreted as knowledge of what is necessary and universal. All knowledge of a general character was regarded as theoretical, including empirical generalizations and theories of the middle range. In the twentieth century this type of position has been defended by the proponents of the conception of history as a social science, who conceived of history as a mix of particular and general characteristics. As the general characteristics could be theorized, they emphasize the importance of theories and models in historical reconstructions (Wehler 1973, Kocka 1977, Tilly 1981, Burke 1993, Iggers 1997). This type of representation, being explicit about the concepts and arguments used, is often called analysis instead of narrative, emphasizing its affinity to the social sciences. This happened especially when 'traditional' history and its function were seriously questioned between 1960 and 1980.

4. *The Nineteenth-century Origins of the Modern Debate on the Functions of History*

From classical antiquity until the eighteenth century the pragmatic presupposition that past experience was meaningful for present purposes was not questioned seriously, nor that the (empirical) truth coincides with the (ethical) good. From the end of the eighteenth century onwards, however, historians and philosophers began to doubt the salutary nature of history seriously and therefore the function of history: when history is no longer experienced as a meaningful and salutary process, what good is the study of history then? Nietzsche's all too nasty question in *The Use and Disadvantage of History for Life* (1874), whether the study of history had any use at all, was the logical endpoint of this vanishing faith in history. European historians basically countered these doubts by two strategies of demarcation. First, by attempts to enhance the scholarly status and 'objective' credentials of history by increasingly expulsing the rhetorical and artistic dimensions from its disciplinary domain. An increasing tendency to conceive of history as a *Wissenschaft*, opposed to *Kunst* or art, by identifying the discipline with its empirical research methods (*Forschung*) at the cost of the literary and artistic methods of its presentation (*Darstellung*) was the result (see Hardtwig 1990, Scholtz 1991 and *History: Theories and Methods*).

Second, these doubts were countered by demarcating disciplinary history sharp from any substantial philosophy of history, especially from its Enlightenment and Hegelian variants (including Marxism), and from any practical function. In the famous words of Ranke historians were not to pass any judgement on the meaning of the past—as the Enlightenment had done—but only 'to show what actually had happened.' By this demarcation strategy the reflection on the practical functions of history was excluded from the disciplinary discourse and was expelled to philosophy of history.

This epistemological demarcation had become possible while history was professionalizing and attaching itself to the national states both institutionally and financially. Because both the state elite's and the majority of the professional historians presupposed that education in (national) history was essential for 'nation building,' the practical functions of history stopped being a problem. The practical function of history was now conceived of as indirect—through individual and collective identity formation alias *Bildung*—rather than as direct through *exempla*. This view was also held by proponents of a 'contemplative' function of history, like Ranke.

The ideal of objectivity was silently identified in historical discourse by promoting the practical cause of the nation state. This identification was not seen as a problem, because in the dominant conceptions of history the historical process itself was identified with the genesis of nations and states. It would take two world wars before this identification turned into a serious problem due to an increasing lack of plausibility. Nietzsche's nasty question, until then comfortably put aside by many professional historians as 'philosophical,' had to be faced at last: what good is the study of history anyway?

Nietzsche's analysis and classification of the functions of history in *The Uses and Abuses of History* although framed as a critique of nineteenth century discipline of history in Germany, has proved to be very fruitful up to the present day (Rüsen 1983–89; Kocka 1989, Hardtwig 1990, Tollebeek 1990). Basically, Nietzsche distinguishes three ways to deal with the past and three types of history that correspond with three different attitudes and basic needs of human beings. Therefore, history has three basic functions, which cannot be reduced to each other and which serve as mutual correctives.

The first way to deal with the past is pragmatic, as an active *acting* being. Actors may address the past in search of examples, which, according to Nietzsche, deserve imitation. The past addressed in this way is treated as 'a collection of monuments.' That's why Nietzsche calls this pragmatic type of history *monumentalistic history*.

The second way to deal with the past is full of piety, as a *contemplative* being, when one is looking for one's roots in the past in search of one's identity. The past addressed in this way is treated as 'a collection of antiquities' or as 'a museum'; therefore, this 'pious' type of history, i.e., history just for the sake of the past, is labeled *antiquarian history* by Nietzsche. Because man is not only directed towards the past, but also towards the future, it is impossible always to treat the past as a 'museum.' In order to create something new it is necessary to break with the past to some extent and to replace it. Therefore, it is also necessary to address the past in a critical way, as something to be criticized as fit for replacement. This type of handling the past, which Nietzsche compares to 'cutting one's roots loose with a knife,' is called *critical history*.

All three types of history have their own type of dangers and deficiencies, which can only be remedied by history of the other types. Monumental history, for instance, tends to equalize the past and the present and to be very selective, obliterating differences and huge parts of the past, unfit for the monumental function. This tendency can only be counteracted by the antiquarian impulse, which in turn suffers from 'historical myopia' and from a 'blind drive to collect.' This tendency, in turn, can only be counteracted by the critical impulse, which in turn must be held in check by the other two in order to prevent critical history from degenerating into a war against one's past.

Historiography can be seen as the combined product of these three impulses, in varying mixes. Disciplinary history since the nineteenth century, however, has usually—although not exclusively—embraced the antiquarian self-image and legitimization, because the recognition of both the monumentalistic and the critical approaches to the past would raise the problem of its practical functions and its objectivity. In accordance with this contemplative self-image of most historians, the antiquarian presentation of identity—i.e., the narrative reconstruction of the genesis of individuality in time—is usually presented as the discipline's only practical function. As in Nietzsche's time, the monumental and the critical functions of history are usually not discussed within the disciplinary boundaries (Angehrn 1985). At times, however, as in the 1970s, the situation has been different, because of an influential number of 'critical' historians in the profession (Kocka 1989).

Symptomatically for this state of affairs is that recently the monumental function of history has been delegated to a separate domain, i.e., the domain of collective memory and of *lieux de mémoire* (see *History and Memory*). The same holds for the critical function of history, which is central to projects like gender history, postcolonial studies, and Foucault's critical history (which replaces the Enlightenment self-image of the West by an image of a disciplined 'carceral society'). These forms of critical history were often not regarded as belonging to the domain of history proper or are at best tolerated in its margins (Foucault 1980, Scott 1988). Mainly in one historiographical domain—that of the Holocaust—reflection on the functions of

history has included both its critical and monumental functions, because the very possibility of the antiquarian approach is fundamentally questioned here. Historical representation as such has turned into a problem, because it is doubted whether the normal categories and procedures of historical representation in order to 'objectify' the past can function in this 'unspeakable' domain. Some hold that psychoanalytic categories like trauma, repression, transfer, splitting off, and working through are more apt to characterize the relationship between the historical representations and represented historical reality. Others have suggested that the 'unspeakable' and 'overwhelming' dimensions of the Holocaust can be better characterized through the notion of the sublime, suggesting a deep analogy between the confrontation with the Holocaust and the confrontation with an earthquake or an avalanche (see *Holocaust, The*, Friedländer 1992, LaCapra 1994).

Mainly in this historiographical domain, the discussion about the functions of history is directly connected to the discussion on representation and its limits. Whether this more inclusive type of debate will alter the dominant discourse on history remains as yet to be seen.

See also: Historicism; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); History and Memory; History and the Social Sciences; History: Overview; History: Theories and Methods; Holocaust, The; Linguistic Turn and Discourse Analysis in History; Syntax–Semantics Interface

Bibliography

Angehrn E 1985 *Geschichte und Identität*. de Gruyter, Berlin

Ankersmit F R 1994 *History and Tropology. The Rise and Fall of Metaphor*. University of California Press, Berkeley, CA

Aristotle 1963 *The Philosophy of Aristotle*. Bambrough R (ed). New American Library, New York

Burke P 1993 *History and Social Theory*. Cornell University Press, Ithaca, NY

Carr D 1986 *Time, Narrative and History*. Indiana University Press, Bloomington, IN

Danto A C 1965 *Analytical Philosophy of History*. Cambridge University Press, Cambridge, UK

Droysen J-G (1977) *Historik*. Fromman–Holzboog, Stuttgart

Fain H 1970 *Between Philosophy and History: the Resurrection of Speculative Philosophy of History Within the Analytic Tradition*. Princeton, NJ

Fay B, Pomper Ph, Vann R (eds.) 1998 *History and Theory. Contemporary Readings*. Cambridge

Foucault M 1980 *Power/Knowledge*. Harvester, Brighton

Friedländer S (ed.) 1992 *Probing the Limits of Representation. Nazism and the 'Final Solution'*. Harvard University Press, Cambridge, MA

Gardiner P (ed.) 1959 *Theories of History; Readings from Classical and Contemporary Sources*. New York

Gossman L 1990 *Between History and Literature*. Harvard University Press Cambridge, MA

Hardtwig W 1990 *Geschichtskultur und Wissenschaft*. DTV, Munich, Germany

Hay D 1977 *Annalists and Historians. Western Historiography from the 8th to the 18th Century*. Methuen, London

Iggers G 1997 *Historiography in the Twentieth Century. From Scientific Objectivity to the Postmodern Challenge*. Wesleyan University Press, London

Kelley D (ed.) 1991 *Versions of History from Antiquity to the Enlightenment*. Yale University Press, New Haven, CT

Kocka J 1977 *Sozialgeschichte. Begriff—Entwicklung—Probleme*. Vandenhoek und Ruprecht, Göttingen, Germany

Kocka J 1989 *Geschichte und Aufklärung*. Vandenhoek und Ruprecht, Göttingen, Germany

Koselleck R 1979 *Vergangene Zukunft. Zur Semantik geschichtlicher Zeiten*. Frankfurt, Suhrkamp, Germany

Koselleck R, Lutz H, Rüsen J (eds.) 1982 *Formen der Geschichtsschreibung*. Deutscher Taschenbuch, Munich, Germany

LaCapra D 1994 *Representing the Holocaust. History, Theory, Trauma*. Cornell University Press, Ithaca, NY

Lorenz C 1997 *Konstruktion der Vergangenheit. Eine Einführung in die Geschichtstheorie*. Cologne

Lorenz C 1998 Can histories be true? Narrativism, positivism and the "metaphorical turn" *History and Theory* **37**: 309–29

Lowenthal D 1985 *The Past is a Foreign Country*. Cambridge University Press, Cambridge, UK

Lübbe H 1977 *Geschichtsbegriff und Geschichtsinteresse. Analytik und Pragmatik der Historie*. Basel

McClosky D, Nelson J S, Megill A (eds.) 1987 *The Rhetoric of the Human Sciences: Language and Arguments in Scholarship and Public Affairs*. University of Wisconsin, Madison, WI

McNeill W H 1986 *Mythistory and Other Essays*. University of Chicago Press, Chicago

Megill A (ed.) 1994 *Rethinking Objectivity*. Duke University Press, Durham, NC

Meran J 1985 *Theorien in der Geschichtswissenschaft*. Vandenhoek und Ruprecht, Göttingen, Germany

Mink L O 1987 *Historical Understanding*. Cornell University Press, Ithaca, NY

Momigliano A 1990 *The Classical Foundations of Modern Historiography*. University of California Press, Berkeley, CA

Mormann T 1997 Ist der Begriff der Repräsentation obsolet? [Is the concept of repression obsolete?]. *Zeitschrift für Philosophische Forschung* **51**: 349–66

Nietzsche F 1981–1874 Vom Nutzen und Nachteil der Historie für das Leben. In: Nietzsche F (ed.) *Unzeitgemäße Betrachtungen*. Insel, Frankfurt, Germany, pp. 95–185

Novick P 1988 *That Noble Dream. The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK

Plumb J 1973 *The Death of the Past*. Penguin, Harmondsworth, UK

Pocock J 1961 The origins of the study of the past: a comparative approach. *Comparative Studies in Society and History* **4**: 209–46

Ricoeur P 1983–85 *Temps et récit*. Seuil, Paris

Rigney A 1990 *The Rhetoric of Historical Representation: Three Narrative Histories of the French Revolution*. Cambridge University Press, Cambridge, UK

Rüsen J 1983–89 *Historische Vernunft. Grundzüge einer Historik*. Vandenhoek und Ruprecht, Göttingen, Germany
Scholtz G 1991 *Zwischen Wissenschaftsanspruch und Orientierungsbedürfnis. Zu Grundlage und Wandel der Geisteswissenschaften*. Suhrkamp, Frankfurt, Germany
Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York
Stern F (ed.) 1970 *Varieties of History. From Voltaire to the Present*. Macmillan, London
Tilly Ch 1981 *As Sociology Meets History*. Academic Press, New York
Tollebeek J 1990 *De toga van Fruin. Denken over geschiedenis in Nederland 1860–1960*. Weseldbibliotheck, Amsterdam
Veyne P 1978 *Comment on écrit l'histoire?* Seuil, Paris
Walsh W H 1967 *Philosophy of History: an Introduction*. Harper and Row, New York
Wehler H-U 1973 *Geschichte als Historische Sozialwissenschaft*. Suhrkamp, Frankfurt, Germany
White H 1973 *Metahistory. The Historical Imagination in Nineteenth Century Europe*. Johns Hopkins University, Baltimore

C. Lorenz

History of Science

The history of science studies the emergence and development of systematic knowledge. Just what kind of knowledge that is—whether it is defined by subject matter, reliability of results, contradistinction to technical know-how, cumulative progress, disciplinary consensus, and/or method of inquiry, demonstration, or explanation—is a point of persistent debate and reflection within the history of science and its cognate fields in philosophy, sociology, and anthropology. Linguistic and historiographic traditions diverge sharply on this point. Whereas Francophone and Anglophone scholars have since the mid-nineteenth century largely reserved the honorific 'science' for the natural sciences and mathematics (influenced by Auguste Comte's hierarchy of sciences in his *Cours de philosophie positive* 1830–42, in which only some had reached the stage of 'positive' knowledge), the German *Wissenschaft* still refers to the range of fields covered by the Latin *scientia*, from physics to philology. Moreover, recent explorations of experiment and other practices have blurred the venerable boundary between *ars* and *scientia*, technique and systematic knowledge (whether understood narrowly as 'science' or broadly as *Wissenschaft*); close investigation of the sciences even in the restrictive English and French usage has furthermore revealed considerable variety among methods and explanations in, for example, biology and physics. Finally, studies of pre-seventeenth-century and non-Western intellectual traditions have made the creation of categories like 'science,' as well as the classifications of knowledge that order such categories, themselves objects (rather than presuppositions) of historical research. Hence this article will not attempt to delineate sharply the boundaries of 'science,' but rather briefly survey the (a) emergence and status of the history of science as a recognized field of inquiry; (b) principal divisions of the field and disciplinary affinities; (c) periodizations and narratives that have shaped those divisions; and (d) major historiographical developments since about 1970 (for an earlier survey see Kuhn [1968] 1972).

1. Emergence and Divisions of the Discipline

The history of science is an ancient pursuit, but a relatively young discipline. From Aristotle through the early nineteenth century, practitioners of one or another branch of knowledge have variously used the history of their field to argue for its dignity and importance, introduce it to beginners, situate it within a broader cultural milieu, summarize the literature to date, position themselves in relationship to that literature, praise and blame predecessors, give evidence of progress, extrapolate a program for future research, and draw lessons concerning the nature of knowledge and the conditions for its flourishing. In works like Joseph Priestley's *The History and Present State of Electricity* (1767) or Georges Cuvier's *Rapport historique sur les progrès des sciences naturelles depuis 1789* (1808), the history was inseparable from the science. Particularly in fields dominated by empirical research, the judicious sifting and ordering of past results was a precondition for coherence and a means for achieving consensus on what was reliably known and where the major challenges for future research lay. These functions are preserved today in the scientific review article and the posing of key 'problems' (e.g., the celebrated Hilbert problems), and practicing scientists occasionally still appeal to the past history of their field for present guidance, especially in times of crisis (Graham et al. 1983). By the mid-nineteenth century, however, histories of science had become distinct from scientific publications, although they were still written primarily by scientists, including prominent figures such as William Whewell, Marcellin Berthelot, Ernst Mach, and Pierre Duhem. Their histories often criticized the current state of science by establishing the genealogy of a controversial hypothesis (e.g., atomism), analyzing the origins of a suspect concept (e.g., absolute space) for hidden flaws, or pleading the superiority of one approach to science over another (e.g., Kantian ideas over Comtean facts). By 1900, histories of science had become a genre distinct from science, but they were still motivated by, and deeply engaged with, contemporary scientific developments (Laudan 1993).

History of science coalesced only gradually as a recognizable discipline in the twentieth century, with its own distinctive program of training, institutions (journals, professional societies, university positions), and scholarly standards. Spurred by the organiza-

tional efforts of such scholars as Paul Tannery in France and Karl Sudhoff in Germany, there was a spurt of professionalizing activity around 1900, when the first international congress on the subject was held in Paris. The German Gesellschaft für Geschichte der Medizin und der Naturwissenschaften was established in 1901; the first chairs in the history of science also appeared at European universities around this time (Kragh 1987). The Belgian historian George Sarton, who emigrated to the USA after World War I, brought with him the journal *Isis* (est. 1912), which became the official organ of the American History of Science Society after its establishment in 1924.

Although Sarton's Comtean view of the history of science as a saga of intellectual and moral progress that should evaluate the past in light of present scientific understanding (Sarton [1927] 1948) has left little trace in current historiography, his tireless attempts to institutionalize and organize the discipline in the form of university departments, specialized journals, and research tools like the *Isis Current* (earlier *Critical*) *Bibliography* (History of Science Society 1967) and *Dictionary of Scientific Biography* (Gillispie 1981) had a lasting impact on the field (Thackray and Merton 1972). During the 1950s graduate programs dedicated to the history of science (sometimes in conjunction with the philosophy of science) multiplied; the 1992 *Guide to the History of Science* lists 188 programs in 25 countries, led by the USA with 72 institutions granting doctorates in the subject.

By no means all history of science is done under the auspices of university departments or published in journals expressly devoted to the subject. The disciplinary affinities of the history of science are many and mutable, with distinctive national differences. Until the 1970s the primary connection, both in Europe and North America, was probably with the sciences themselves. This connection was rooted deep in the origins of the field and sustained by the continuing importance of some form of the history of science in the intellectual and social initiation of young scientists, as well as by the recruitment of historians of science from the ranks of those with scientific training. In the UK and North America, history largely replaced philosophy during the 1980s and 1990s as the discipline to which historians of science felt most akin, although interactions with sociology and, more recently, anthropology have also exerted a powerful influence. In France, Germany, and Italy, philosophy has remained a significant framework for the history of science, in diverse forms ranging from neo-Kantianism to logical empiricism. In addition to its disciplinary polyvalence with respect to choice of problems and methods, the history of science is closely linked by its subject matter to the histories of medicine and technology, albeit to differing degrees depending on the historical period in question. The phrase 'science and technology studies' bears witness to these criss-crossing ties to other disciplines, serving as an abbreviation for the conglomerate 'history, philosophy, sociology, and anthropology of science, medicine, and technology'—which, however cumbersome, accurately reflects the ecumenical perspective of many historians of science.

2. *Narratives and Periodization*

Because of its links to philosophy, sociology, and anthropology, as well as to the natural sciences and mathematics, the history of science has been highly theorized for centuries. Theories about what scientific knowledge is, how it develops over time, where and when it expands and contracts, and in which ways it interacts with extrascientific contexts have profoundly shaped the way in which the history of science has been periodized and narrated. The more recent developments (since about 1970) in the historiography of science are incomprehensible without some background concerning the dominant theoretical schemes that had molded the field previously.

All history is inevitably written to some extent from the standpoint of the present, but present science—its orthodoxies and heresies, its controversies and critics—has exerted a mighty influence on the history of science since the Enlightenment. The great bulk of the history of science has been written explicitly or implicitly to illuminate, justify, or criticize the current state of science. In the Enlightenment histories, e.g., Jean-Étienne Montucla's *Histoire des mathématiques* (1758) or Jean-Baptiste Delambre's *Rapport historique sur les progrès des sciences mathématiques depuis 1789* (1810), the aim was to demonstrate that at least some of the sciences had shown steady and irreversible progress, especially since the Renaissance, and that the history of science was therefore a more edifying genre than the chronicles of human folly recounted by political and military history. The notion of progress was originally applied to art (by Giorgio Vasari in 1550) and to technology (by Francis Bacon in 1620), not to science. Once this interpretation of scientific change took hold in the eighteenth century, however, science became the premier example of progressive advance and scientific progress; a criterion by which to judge the comparative worth of civilizations. Whatever may have been the cultural achievements of the Latin Middle Ages or the Chinese empire, the stagnation of the sciences in these milieux, so ran the argument of the Enlightenment historians, revealed fundamental intellectual and social limitations—limitations that in the nineteenth and twentieth centuries came to be described as characteristic of 'traditional' or 'premodern' societies. Hence narratives of scientific progress eventually became part and parcel of larger narratives about modernization and European superiority.

Although the assumptions underlying the Enlightenment vision of scientific progress are seldom invoked

explicitly by contemporary historians of science, the reigning periodization of the field and the definition of specialties as central or peripheral still bears the traces of the traditional/modern division. Despite their glaring differences in many substantive aspects, the study of science in antiquity, the Arabic and Latin Middle Ages, and in non-Western civilizations such as China and India has often been amalgamated into a shared problematic: to what extent did the knowledge of these highly disparate cultures anticipate/approximate the science of post-Renaissance Europe? And to the extent that it did not, what were the intellectual and social obstacles that blocked the way (e.g., Huff 1993)? Hence the pioneering works in these areas (e.g., Duhem 1908, Maier 1949–58, Clagett 1959, Needham et al. 1954) foregrounded possible forgotten contributions to, and missing preconditions for, the Scientific Revolution.

Knowledge traditions that did not lead to modern science, understood to commence in early modern (a term as saturated with hindsight as 'the Middle Ages') Europe, were generally relegated to the periphery of the discipline, whatever their intrinsic interest and however excellent the scholarship devoted to them. Conversely, the Scientific Revolution has remained oddly central to the historiography of science: oddly, because recent scholarship on the sixteenth and seventeenth centuries has emphasized the heterogeneity and regional variations in the developments in anatomy, astronomy, natural history, mechanics, chemistry, mathematics, and physics once summarily lumped together as 'the Scientific Revolution' (Lindberg and Westman 1990, Porter and Teich 1992, Shapin 1996, Osler 2000), and because the steep increase of studies of nineteenth- and twentieth-century science since about 1980 has highlighted the novelty and significance of post-1800 developments, as the steady increase in the number of entries under these rubrics in the *Isis Current Bibliography* bears witness.

Yet the Scientific Revolution retains much of the original status first granted to it by the Enlightenment historians as the mold in which modernity was cast, one of the few points upon which both the apostles of progress and the critics of modernity agree. This periodization has left traces in the division of subspecialties within the history of science, which tend to be chronological or geographic ('ancient,' 'Southeast Asian') for knowledge traditions prior to, or uninfluenced by, the Scientific Revolution, but disciplinary ('physical sciences,' 'biomedical sciences') for those following that original event—even for disciplines (e.g., astronomy) with histories that long antedate the sixteenth century and with geographical distribution well beyond Europe.

The progress of the sciences (especially mathematics and the exact sciences) set forth by the Enlightenment historians was cumulative but not revolutionary. In their accounts, knowledge steadily expanded around a stable core of permanent truths, which may have taken millennia to discover, but which were thereafter firmly established beyond challenge. The language of 'revolutions' declined among early nineteenth-century scientists, perhaps in order to distance the permanent improvements of natural knowledge from the brusque reversals of politics; instead, the language of 'law' (e.g., Comte's 'Law of Three Stages') was applied to the historical development of science (Cohen 1985). In the middle decades of the nineteenth century, however, this view of durable, inexorable scientific progress gave way to one of the rapid succession of scientific theories, even in areas (e.g., optics) long believed to rest upon unshakable foundations. Scientific progress had become vertiginous and open-ended; today's truth might be discredited as tomorrow's error. Under these conditions, it became increasingly difficult for historians of science to portray the steady growth of knowledge as culminating in present triumphs, since the endpoint was constantly changing. Instead, the history of science turned critical and conceptual: in the hands of Mach, Ostwald, and other scientist-historians, history of science revealed the past roots of present incoherences and taught the mutability of scientific theories. Within the revised interpretation of scientific progress, the purpose of the history of science was to use past science to change, not confirm, present science. Despite the dramatic shift in the narrative of scientific progress and in the goals of the history of science, its presentist orientation remained unchanged.

Almost all the strands of early twentieth-century historiography of science can be seen as attempts to come to grips with the peculiar character of scientific knowledge, which seemed to be simultaneously mutable and cumulative, driven by an internal dynamic and yet sensitive to ambient social conditions. These attempts often tugged in opposite directions. For example, the logical positivist program that stamped the history and philosophy of science in the middle decades of the twentieth century, especially in the modified form imported from the Vienna Circle to the UK and the USA, fixed upon the empirical findings of science as its enduring core in the Heraclitean flux of hypotheses and theories. These 'protocol statements' (Carnap 1928) were likened to simple observations, bringing science into the neighborhood of common sense. In contrast, the French historian and philosopher Gaston Bachelard, drawing upon the insights of psychoanalysis, insisted upon the necessity of overcoming everyday experience and intuitions in the formulation of successful scientific concepts (Bachelard 1938).

A many-cornered debate was ignited by the attempts of Marxist historians to relate the rise of modern science to specific socioeconomic conditions (Hessen 1931, Bernal 1939), out of which emerged the opposition of 'internal' to 'external' approaches to the history of science (Kuhn [1968] 1972, Shapin 1992). External approaches found a non-Marxist model in

Robert K. Merton's influential study of how science came to be valued in seventeenth-century England (Merton 1938, Shapin 1988). Informed by German and French neo-Kantian traditions in the history of philosophy, Alexandre Koyré's studies of the development of science in the broader intellectual context of early modern Europe stamped the work of historians of science, especially in France and the USA, trained in the 1950s and 1960s (Koyré 1939–40, 1961, Jorland 1981).

Common to all of these otherwise highly diverse perspectives on the history of science was an engagement with contemporary science and its lineage. Whether positivist or psychoanalytic, externalist or internalist, the history of science took the form of a genealogy leading to ideas and discoveries canonized by contemporary science. However much they may have disagreed on historical methods and forms of explanation, the combatants agreed on what and who in the history of science was worthy of explanation, namely that which had turned out to be pregnant with the future. In these histories a potent telos was at work, making present science appear to be not only the actual but also the inevitable outcome of past developments.

3. *Developments Since 1970*

Thomas S. Kuhn's *The Structure of Scientific Revolutions* (Kuhn [1962] 1970) broke decisively with this historiography, and stimulated new trends (not all to Kuhn's liking) in the history, sociology, and philosophy of science that have transformed the understanding of science and its growth in these fields since 1970. Kuhn argued for a plurality of scientific revolutions, in addition to *the* Scientific Revolution, which alternate with periods of 'normal science' framed by a 'paradigm' that models what can count as an acceptable problem, method, and solution in a given science in a specified period. It was less, however, Kuhn's 'structures' or patterns of long-term scientific development that exercised a deep and enduring influence on the history of science (although this feature of the book dominated its reception in other disciplines, particularly the social sciences) than his frontal attack on progressive narratives, his sketchy but evocative notion of 'paradigm' that embraced practices as well as theories, and the thoroughgoing historicism of many of his examples. All of these lines of thought were further cultivated in the history, philosophy, and sociology of science (the three disciplines were drawn into intensive interaction in part through their differentiated reception of Kuhn's book) during the subsequent three decades, often with consequences that led in directions diverging markedly from those plotted in *The Structure of Scientific Revolutions*.

Kuhn's rejection of progressive narratives in the history of science derived from his view that successive paradigms established and later overthrown by scientific revolutions were, first, internally coherent, and second, in a strong sense incommensurable with one another. This implied that science could not be cumulative in any progressive sense, one paradigm extending and building upon its predecessors. Kuhn's claim that conflicts between rival paradigms were not necessarily resolved by paradigm-independent criteria (e.g., predictive accuracy of explanatory unity) unsettled the inevitability of scientific development still further. Sociologists of science Barry Barnes and David Bloor took these claims as the basis of an approach called 'constructivism' (Golinski 1998), in which the development of scientific knowledge was to be explained without appeal to its truth or falsehood, on the grounds that the truth of a proposition is neither a necessary nor sufficient condition for its acceptance (Bloor [1976] 1991).

For historians of science, this program proved stimulating in at least three respects. First, in conjunction with Kuhn's focus on periods of conflict between rival paradigms, it prompted detailed studies of how scientific controversies were in fact resolved. Second, it enlarged the definition of what could count as 'science' for the purposes of historical investigation by ruling out hindsight judgments about what was and was not legitimate by present standards (to the benefit of specialties such as the history of premodern science and of the human sciences). Third, it encouraged a broader exploration of the grounds for scientific conviction by enforcing a 'symmetric' treatment of reasons for embracing false as well as true beliefs. This deliberate refusal to adopt the perspective of present science in the investigation of past knowledge went under the name of 'anti-Whiggery' among historians of science (in allusion to Herbert Butterfield's critique of 'Whig' political history). Among sociologists of science, the symmetry principle led to an increasingly distanced approach even to present science, to the point of adopting the ethnographer's estranged perspective (Latour and Woolgar [1979] 1986). The overall impact of the constructivists in both sociology and history of science was to highlight the contingency of the development of scientific knowledge, in stark contrast to earlier narratives of inexorable progress (Pestre 1995, Rheinberger 1997).

Kuhnian paradigms drew upon the philosophical reflections of Ludwig Wittgenstein (Wittgenstein 1953) and Michael Polanyi (Polanyi 1958) to stress the importance of knowledge that could not be reduced to rules or algorithms or even to words in a normal scientific tradition. In the history, sociology, and philosophy of science a rich literature on 'scientific practices' has taken the inadequacy of the formal aspects of theory to explain the distinctive aspects of scientific knowing as its departure point. Although scientific practices may embrace everything from a

knack for integrating equations to the deft adjustment of finicky instruments, studies of experiment have taken on an exemplary character in this branch of the field. Interactions among historians (e.g., Shapin and Schaffer 1985, Galison 1987, Gooding et al. 1993), philosophers (e.g., Cartwright 1983, Hacking 1983), and sociologists (e.g., Collins 1985) created a discussion that blended epistemological, historical, and social analyses.

Experiment and scientific practices more generally have emerged as autonomous forms of knowledge, no longer subsidiary to and in lockstep with theory, that depend crucially on local factors (including skill, material culture, values) that travel only with difficulty. This emphasis upon the locality of scientific knowledge has spawned a new set of problems: How is local science made global? In the earlier historiography, science had been assumed to be universal because nature was so. For historians of scientific practice, in contrast, further explanation is required to show the ways in which localized achievements can be made at least partially communicable among practitioners divided by geography, history, or culture (Schaffer 1992, Galison 1999). Sociologists of science have attacked this problem from the standpoint of 'actor-networks' (Latour 1987).

Both the rejection of presentist perspectives and the close study of scientific practices have pushed historians of science toward a more thoroughly historicist approach. Whereas their predecessors practiced anachronism openly and with conviction, on the view that past science must be understood by the lights of present science, most historians of science now writing insist upon some attempt to reconstruct the conceptual and cultural categories of historical actors. This historicism has expressed itself in two important movements in recent historiography of science: the aim to understand 'science in context,' and attempts to write histories of entities previously assumed to be transhistorical.

'Science in context' (which lent its name to a journal) thickens the connections of science with its historical place and time, as well as blurring the boundaries between factors previously regarded as either internal or external to a discipline. Indeed, one of the consequences of the studies of science in context has been to make the emergence, persistence, and occasional disappearance of such boundaries, topics of historical investigation in their own right. One of the most important of the new strands of contextual history of science, because of its potentially wide application, has been the investigation of how ideas and practices of gender have shaped the modern sciences (e.g., Haraway 1989, Jordanova 1989, Schiebinger 1993).

Although some critics of this embedded view have feared that context impugns scientific autonomy and therefore validity, many recent studies have depicted the social, technological, and economic milieu of science as resources for scientific creativity, not merely as ideological contaminants (e.g., Wise 1989–90). Histories of supposedly transhistorical entities such as experience (Dear 1995), truth (Shapin 1994), and objectivity (Daston and Galison 1992, Porter 1995) make use of the tools of science in context to anchor these Kantian abstractions in specific historical circumstances; they are also at least indirectly indebted to the remarkable studies of Georges Canguilhem (Canguilhem [1943] 1966) and, especially, Michel Foucault (Foucault 1969) that challenged the universality and permanence of fundamental modern categories like normalcy and sexuality.

The most recent developments in the history of science might be briefly described as materialized and embodied. The material culture of science, ranging from the architecture of the spaces in which science is done (Rheinberger et al. 1996, Galison and Thompson 1999) to the machines that do the science, have shifted from the periphery to the center of historical attention. Hands-on experience with objects is no longer understood as a means to a greater end, but as a force that shapes the culture and penetrates into the metaphysics of science (Galison 1997). Embodiment has entered the history of science through several doors: the study of the interaction of experiment and craft traditions (Sibum 1995), the investigation of the scientific persona in historical context (Lawrence and Shapin 1998, Shortland and Yeo 1996), and the making visible of the role of women in science (Rossiter 1982, Rossiter 1995). All of these explorations of embodiment challenge the Cartesian image of the disembodied knower, as well as the hagiography of great individuals; they aim at the reconstruction of collective identities of scientists, as well as their consequences for who was allowed to do what kind of science.

Given the profusion of topics and approaches in the history of science, it is understandable that one historian might wonder 'whether the history of science is a coherent discipline or just a collection of scholars aggregated by the accidents of history and the accretion of a common historiography' (Rosenberg 1988, p. 570). This is a doubt that in large part revolves around the difficulty of defining the *sine qua non* of scientific knowledge, if it has one. Yet in this still somewhat undisciplined discipline, the reward of permeable boundaries has been a rare dialogue, ongoing and sometimes heated, among the natural sciences, social sciences, and humanities.

See also: Cognitive Psychology: History; Comte, Auguste (1798–1857); Cultural Studies of Science; Enlightenment; History of Science: Constructivist Perspectives; History of Technology; Knowledge, Sociology of; Kuhn, Thomas S (1922–96); Medicine, History of; Norms in Science; Objectivity, History of; Paradigms in the Social Sciences; Physical Sciences: History and Sociology; Progress: History of the Concept; Science and Religion; Science and Tech-

nology, Anthropology of; Science, Sociology of; Scientific Disciplines, History of; Scientific Revolution: History and Sociology; Situated Knowledge: Feminist and Science and Technology Studies Perspectives; Sociology, History of; Statistics, History of; Technology, Social Construction of; Truth and Credibility: Science and the Social Study of Science; Wittgenstein, Ludwig (1889–1951)

Bibliography

Bachelard G 1938 *La Formation de l'Esprit Scientifique*. Vrin, Paris

Bernal J D 1939 *The Social Function of Science*. Macmillan, New York

Bloor D [1976] 1991 *Knowledge and Social Imagery*. University of Chicago Press, Chicago

Canguilhem G [1943] 1966 *Le Normal et le Pathologique*. Presses Universitaires de France, Paris

Carnap R 1928 *Die logische Aufbau der Welt*. Weltkreis-Verlag, Berlin

Cartwright N 1983 *How the Laws of Physics Lie*. Clarendon Press, Oxford, UK

Clagett M 1959 *The Science of Mechanics in the Middle Ages*. University of Wisconsin Press, Madison, WI

Cohen I B 1985 *Revolutions in Science*. Harvard University Press, Cambridge, MA

Collins H 1985 *Changing Order: Replication and Induction in Scientific Practice*. Sage, Beverly Hills, CA

Daston L, Galison P 1992 The image of objectivity. *Representations* **40**: 81–128

Dear P 1995 *Discipline and Experience: The Mathematical Way in the Scientific Revolution*. University of Chicago Press, Chicago

Duhem P 1908 *ΣΩΖΕΙΝ ΤΑ ΦΑΙΝΟΜΕΝΑ: Essai sur la Notion de Théorie Physique de Platon à Galilée*. A. Hermann et Fils, Paris

Foucault M 1969 *L'Archéologie du Savoir*. Éditions Gallimard, Paris

Galison P 1987 *How Experiments End*. University of Chicago Press, Chicago

Galison P 1997 *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, Chicago

Galison P 1999 Trading zone: Co-ordinating action and belief. In: Biagioli M (ed.) *The Science Studies Reader*. Routledge, New York, pp. 137–60

Galison P, Thompson E 1999 *The Architecture of Science*. MIT Press, Cambridge, MA

Gillispie C C (ed.) 1981 *Dictionary of Scientific Biography*. Charles Scribners' Sons, New York

Golinski J 1998 *Making Natural Knowledge: Constructivism and the History of Science*. Cambridge University Press, Cambridge, UK

Gooding D, Pinch T, Schaffer S (eds.) 1993 *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge University Press, Cambridge, UK

Graham L, Lepenies W, Weingart P (eds.) 1983 *Functions and Uses of Disciplinary History*. D. Reidel, Dordrecht, The Netherlands

Hacking I 1983 *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, Cambridge, UK

Haraway D J 1989 *Primate Visions: Gender, Race, and Nature in the World of Modern Science*. Routledge, New York

Hessen B 1931 The social and economic roots of Newton's 'Principia.' In: Bukharin N I (ed.) *Science at the Crossroads*. Kniga, London, pp. 146–212

History of Science Society 1967 *Isis Current Bibliography* (originally *Isis Critical Bibliography*). University of Chicago Press, Chicago

Huff T E 1993 *The Rise Of Early Modern Science: Islam, China, and the West*. Cambridge University Press, Cambridge, UK

Jordanova L 1989 *Sexual Visions: Images of Gender in Science and Medicine between the Eighteenth and Twentieth Centuries*. University of Wisconsin Press, Madison, WI

Jorland G 1981 *La Science dans la Philosophie: Les Recherches Épistémologiques d'Alexandre Koyré*. Éditions Gallimard, Paris

Koyré A 1939–40 *Études Galiléennes*. Hermann, Paris

Koyré A 1961 *La Révolution Astronomique: Copernic–Kepler–Borelli*. Hermann, Paris

Kragh H 1987 *An Introduction to the Historiography of Science*. Cambridge University Press, Cambridge, UK

Kuhn T S [1962] 1970 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Kuhn T S [1968] 1972 History of science. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan and Free Press, London and New York, Vol. 13, pp. 74–83

Latour B 1987 *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, Cambridge, MA

Latour B, Woolgar S [1979] 1986 *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, NJ

Laudan R 1993 Histories of the sciences and their uses: A review to 1913. *History of Science* **31**: 1–34

Lawrence C, Shapin S (eds.) 1998 *Science Incarnate: Historical Embodiments of Natural Knowledge*. University of Chicago Press, Chicago

Lindberg D, Westman R (eds.) 1990 *Reappraisals of the Scientific Revolution*. Cambridge University Press, Cambridge, UK

Maier A 1949–58 *Studien zur Naturphilosophie der Sptscholastik*. Edizioni di Storia e Letteratura, Rome

Merton R K 1938 Science, technology and society in seventeenth-century England. *Osiris* **4**: 360–632

Needham J et al. 1954 *Science and Civilization in China*. Cambridge University Press, Cambridge, UK

Osler M J (ed.) 2000 *Rethinking the Scientific Revolution*. Cambridge University Press, Cambridge, UK

Pestre D 1995 Pour une histoire sociale et culturelle des sciences: nouvelles définitions, nouveaux objets, nouvelles pratiques. *Annales: Histoire, Sciences Sociales* **50**: 487–522

Polanyi M 1958 *Personal Knowledge: Towards a Post-critical Philosophy*. University of Chicago Press, Chicago

Porter T M 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ

Porter R, Teich M (eds.) 1992 *The Scientific Revolution in National Context*. Cambridge University Press, Cambridge, UK

Rheinberger H J 1997 *Towards A History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford, CA

Rheinberger H J, Hagner M, Wahrig-Schmidt B (eds.) 1997 *Räume des Wissens: Repräsentation, Codierung, Spur*. Akademie-Verlag, Berlin

Rosenberg C 1988 Woods or trees? Ideas and actors in the history of science. *Isis* **79**: 565–70
Rossiter M W 1982 *Women Scientists in America: Struggles and Strategies to 1940*. Johns Hopkins University Press, Baltimore
Rossiter M W 1995 *Women Scientists in America: Before Affirmative Action, 1940–1972*. Johns Hopkins University Press, Baltimore
Sarton G [1927] 1948 *Introduction to the History of Science*. Williams & Wilkins, Baltimore, MD
Schaffer S 1992 Late Victorian metrology and its instrumentation: A manufactory of ohms. In: Bud R, Cozzens S (eds.) *Invisible Connections: Instruments, Institutions, and Science*. SPIE Optical Engineering Press, Bellingham, WA, pp. 23–56
Schiebinger L 1993 *Nature's Body: Gender in the Making of Modern Science*. Beacon Press, Boston
Shapin S 1988 Understanding the Merton thesis. *Isis* **79**: 594–605
Shapin S 1992 Discipline and bounding: The history and sociology of science as seen through the externalism–internalism debate. *History of Science* **30**: 333–69
Shapin S 1994 *A Social History of Truth: Civility and Science in Seventeenth-century England*. University of Chicago Press, Chicago
Shapin S 1996 *The Scientific Revolution*. University of Chicago Press, Chicago
Shapin S, Schaffer S 1985 *Leviathan and the Air-pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, Princeton, NJ
Shortland M, Yeo R (eds.) 1996 *Telling Lives in Science: Essays on Scientific Biography*. Cambridge University Press, Cambridge, UK
Sibum H O 1995 Reworking the mechanical value of heat: Instruments of precision and gestures of accuracy in early Victorian England. *Studies in the History and Philosophy of Science* **26**: 73–106
Thackray A, Merton R K 1972 On discipline building: The paradoxes of George Sarton. *Isis* **63**: 473–95
Wise M N 1989–90 Work and waste: political economy and natural philosophy in nineteenth century Britain. *History of Science* **27**: 263–301, 391–449; **28**: 221–61
Wittgenstein L 1953 *Philosophical Investigations* [trans. Anscombe G E M]. Blackwell, Oxford, UK

L. Daston

History of Science: Constructivist Perspectives

The last quarter of the twentieth century saw the emergence of an important and challenging new perspective on the sciences and their history. Drawing largely on developments in the social studies of science, but also on new trends in labor and social history, historians began to look at science as local practice, rather than as an abstract, disembodied, and universal set of ideas. Rather than seeing science as somehow outside culture, historians started looking at science as part of culture, locally constituted in particular material, social, and even political practices. Constructivist historians are interested in the detailed processes of scientific knowledge production, dissemination, and reception. They differ from more traditional social historians of science in their insistence that the content as well as the context of science is a proper focus of cultural historical analysis. Crucially, constructivist historians see science as being a species of skilled practice and therefore inquire into the cultural origins of those skills, the means through which they are produced, validated and disseminated, and the ways in which particular forms of inquiry—and particular kinds of knowledge—have historically been characterized as scientific.

1. Local Cultures

In their exemplary study of the early modern culture of experiment, Shapin and Schaffer lay out some of the central concerns of the constructivist history of science (Shapin and Schaffer 1985). Focusing on the dispute between Robert Boyle and Thomas Hobbes concerning the validity of Boyle's famous air-pump experiments, they use Hobbes's critique of Boyle's experiments to lay bare the interests underlying them and the strategies that Boyle and his allies adopted to convince others of the experiments' reliability. They demonstrate how even such an apparently universal cornerstone of the scientific method as experiment has its origins in particular local and contingent circumstances. Making science local and examining the interests that underlie the articulation and resolution of scientific debates have been preoccupations of constructivist history. Other pioneering studies include accounts of the ways that disillusionment with rational philosophy and mechanistic notions of causality in postwar Weimar Germany informed interpretations of quantum theory (Forman 1971), and of the role social interests played in debates concerning the anatomical structure of the brain in early nineteenth-century Edinburgh (Shapin 1979). Such studies emphasize the way in which such interests are constitutive rather than contingent features of even the most apparently abstract and technical forms of knowledge.

One important outcome of constructivist history is the recognition that even the issue of what kind of activity might count as science is often the product of locally articulated debates and controversies. Not only did Boyle and Hobbes disagree as to what might reliably be inferred from the outcome of air-pump experiments, but they also disagreed as to whether or not doing such experiments should be counted as a reliable way of generating scientific knowledge at all. Similarly, the scientific and the medical establishment who opposed the introduction of continental ideas concerning evolution and transmutation into early

Victorian England did so both because they did not recognize the practices or the practitioners involved as being properly scientific and because they disagreed with the theories themselves (Desmond 1989). As the labels 'science' and 'scientific' gained in cultural value, so too did different groups of knowledge producers disagree more vociferously as to what kinds of people and what kinds of activity should be accorded those labels (Gieryn 1999). Constructivist historians have identified the business of policing the constantly changing boundaries of science as a crucial part of science's history. Through institutions, regimes of training, control of resources, and access to networks of patronage, natural philosophers and scientists have struggled to circumscribe and discipline their own and others' knowledge producing activities.

This localist historical perspective leads to recognition of the extent to which the production and validation of scientific knowledge are contingent upon more or less informal networks of trust. In order for particular groups or individuals to be considered reliable spokespersons for nature, they have to be trustworthy. They have to be the kinds of people who can be relied upon to tell the truth. This opens up the possibility of producing a cultural history of truth and objectivity. The marks of being trustworthy vary according to time and place. In early modern England for example, trustworthiness was considered the hallmark of the gentleman (Shapin 1994). As a result, many of the informal (and more formal) rules of early modern gentlemanly behavior became enshrined as codes of conduct for the nascent scientific community as well. Some historians have argued that the increased scientific reliance on precision measurement and self-registering instrumentation since the beginning of the nineteenth century may be regarded as efforts to circumvent the need for trusting individuals with knowledge production (Schaffer 1988). Others have pinpointed the increasing quantification of science as a means of overcoming the problem of trust in the production and particularly in the transfer of knowledge (Porter 1995).

2. *The Material Culture of Experiment*

The constructivist history of science pays attention not only to experiment as a form of inquiry, but also to the process of experiment. Experimentation is taken to be a matter of human agents actively engaged in constructing the natural world around them (Gooding et al. 1989). From this perspective, the material culture of experiment becomes a crucial focus of attention. Historians have worked to recover the kinds of material resources required for successful experimentation. They have traced the emergence of the laboratory as a specialized space within which experiments take place. They have emphasized the ways in which successful manipulation of nature requires the experimenter to manipulate material and social resources successfully as well (Latour 1988). Michael Faraday's discovery of electromagnetic induction, for example, involved as much his marshalling of the Royal Institution's considerable resources as it was the outcome of his artful manipulations in the laboratory (Morus 1998). Constructivist historians have identified as well the important ways in which individual scientific reputations hinge upon recognition of the property relationship between experiment and experimenter. Priority disputes in science are frequent and virulent precisely because of the ways in which experiments embody the skills and reputations of their makers (Iliffe 1992).

Seeing experiment as a skilled activity, constructivist historians have focused much attention on understanding the historical origins of the kinds of skills required in order to carry out experiments. They have investigated how particular skills have been transferred into the laboratory from other settings, such as workshops or factories. Historical experiments have been replicated in an effort to better understand the tacit, unarticulated skills required for their successful performance. Such studies have shown how iconic experiments such as James Prescott Joule's famous paddle-wheel experiments establishing the mechanical equivalent of heat during the 1840s, were contingent upon the technical know-how that Joule and his workers derived from their experiences in the family brewery (Sibum 1995). Increasingly also, historians have looked at the politics of skill in the laboratory and elsewhere, investigating the division and distribution of hierarchies of skilled activity within science at different times and places. Much attention has been devoted to the historical transmission of material resources and skilled personnel that accompanies the standardization of experimental practices. Work in this vein has shown how experiments move from being instances of highly skilled, difficult to replicate, and delicate procedures performed in highly circumscribed settings to being robust, standardized, and readily reproducible embodiments of group knowledge (Galison 1987).

Apparently self-evident features of experimentation, such as precision measurement, are similarly shown by constructivist historians to have a history. Far from being a routine and unproblematic aspect of scientific activity, measurement can be taken to be an expression of particular cultural values (Wise 1995). The rise of metrology, for example, was an expression of science's increasing embeddedness in and contribution to a rapidly bureaucratizing and industrializing Western culture. Measurement could be regarded not only as a means of disciplining nature, rendering it more amenable to standardized routines of inquiry, but also as a means of disciplining the scientific workforce. Nineteenth-century scientific managers such as George Biddell Airy, the English Astronomer Royal, regarded their laboratories and observatories as

industrial workplaces and used regimes of routinized measurement and calculation to impose an industrial discipline on their workforces (Smith 1991). In late nineteenth century laboratories at Cambridge or Glasgow, natural philosophers such as James Clerk Maxwell or Lord Kelvin saw the imposition of regimes of precision measurement in teaching and research as an expression of their affinity with industrializing society (Schaffer 1992). Increasingly, constructivist historians have investigated how adherence to cultural values like precision came to be taken as a hallmark of what it meant to be properly scientific. More recently, some historians have returned to matters of theory, using the analytical tools developed to understand experiment to look at theory as a form of cultural practice as well. From this perspective, processes of theory construction and dissemination are seen in terms of the pedagogical training and practical skills required to participate in highly specialized theoretical cultures (Galison and Warwick 1998).

3. *Science as Performance*

One notable outcome of constructivist historical work has been to make the identity of the scientific practitioner itself problematic. Being a scientist is not a self-evident activity: the word itself is of comparatively recent, early nineteenth century origin. Scientific practitioners in the past occupied a number of different cultural niches. Crucially, constructivist historians have shown how natural philosophers at different times and places have been engaged actively in carving out social spaces for themselves and their work. Galileo's position at the Medici court in Renaissance Florence, for example, was not ready made. Galileo had to work hard to promote the idea of a philosopher or mathematician courtier that he could then fulfill (Biagioli 1993). He had to fashion himself into the kind of person who could easily inhabit that cultural niche and then to orient his work so that it would fit into the milieu of Italian Renaissance court life. The same could be said of Humphry Davy and his efforts to forge a place for the fashionable natural philosopher in Georgian London (Golinski 1992). Constructivist approaches share with the 'new history' an emphasis on the important role of individual agents' artful engagement in constructing both themselves and their worlds.

Recent constructivist work has paid particular attention to the body of the scientific practitioner (Lawrence and Shapin 1998). This is an outgrowth of the constructivist insistence on viewing scientific practices and the knowledge they produce as being in crucial ways embodied. Scientific practitioners at different times and places have needed to be particular kinds of human beings with specific bodily attributes. The ways in which those attributes were held to be best produced—whether by birth, self-discipline, or rigorous training—have been a focus of increasing historical attention. Historians have examined, for example, the emergence of particular regimes of mental training in Victorian Cambridge and the ways in which candidates for mathematical honors were encouraged to train their bodies into being fit receptacles for their minds (Warwick, in Lawrence and Shapin 1998). The bodily interaction of practitioners with their experimental apparatus and the spatial organization of their laboratories have come to be recognized as playing an important role in the production of knowledge. The organization of such spaces and the kinds of material interactions between practitioners, and between them and their instruments, that have been taken to produce reliable knowledge are now seen as contingent and open to historical investigation.

Not only the identity and activities of the practitioner producing knowledge, but also the role of the audience for such productions are now recognized as constitutive. Audiences for scientific knowledge do not simply act as passive receivers of settled and unproblematic matters of fact. They have an important and active role to play in the articulation and resolution of controversies. During the early modern period and well into the nineteenth century, experiments were often performed before an audience. This was not simply a matter of demonstration. Members of the audience—prestigious and often named individually—played an important role in validating the knowledge produced. Their reputations could be used to underwrite the matters of fact they had witnessed. Audiences have participated in and responded to the processes of knowledge production through a variety of media, from direct witnessing to the virtual witnessing of reading a scientific paper at first hand, to the reception of second or third-hand reports in museums, newspapers, reviews, or popular handbooks. In the hands of states, public instrumental uses of science and technology have become devices for legitimating political power (Ezrahi 1990). Public scientific performances are an important part therefore of the dynamic process of knowledge-production and validation.

4. *Science at Work*

Constructivist historical analysis looks as well at the many ways in which nature has been used historically as a model for social organization. The constitution of the natural economy has frequently been used as a powerful argument in favor of particular ways of ordering society. Natural theology in the late eighteenth and early nineteenth centuries, for example, was widely used as an argument in favor of existing social hierarchies. Evolutionary theories were similarly cited to support new, meritocratic visions of society. The proper social economy was often held to mirror the natural order. Constructivist historians have shown

how models of nature and society in the past were frequently mutually sustaining. They were produced together. Discussions in mid-nineteenth century thermodynamics surrounding issues of work and waste in natural economies were coextensive with arguments concerning the proper organization of work and the minimization of waste in political economy (Smith and Wise 1989). Proponents of new forms of industrial organization, for example, looked to nature for their models. At the same time, scientists looked to factory organization as a model for the best means of ordering the search for truth.

In turn, different models for the organization of scientific work often expressed different understandings not only of how nature was ordered but also of the proper politics of knowledge production. Nineteenth-century scientific gentlemen, for example, regarded scientific work as a highly disciplined, hierarchical, and increasingly specialized activity. In principle, it was open to all. In practice, however, it could only be conducted properly either by, or under the supervision of, properly trained and vocationally committed 'gentlemen of science' (Rudwick 1985). Other social groups had different notions of the politics of knowledge. The gentlemen's vision excluded women and artisans, for example, from anything other than a supportive role in making knowledge. Working-class men, on the other hand, could construct their own very different accounts of how knowledge making ought to be organized (Secord 1994).

Through constructivist historical analysis, many of the traditional analytical categories employed by historians and sociologists of science have become recognized as being themselves in need of careful historical contextualization. Pronouncements by historical actors concerning matters such as the scientific method, experimentation, objectivity, or even more novel analytical categories such as skill, are not to be taken at face value. Such pronouncements usually carry considerable weight. The issue of what kind of activity might properly be considered as carrying out an experiment, for example, is not straightforward. At different times, experimentation has been regarded as both a public and a private process. It has been characterized as both a fundamentally communal and an essentially individual activity. What it means to be a skilled worker, again, has changed a great deal since the early modern period. Skill, like method, or even science itself, is a historical construction. By looking at science as a pre-eminently cultural activity, constructivist histories of science have made it possible to raise a whole host of new questions concerning the ways in which human beings, in the past, have tried to make sense of the world around them. Rather than seeing science as a static, monolithic edifice, constructivist history recognizes it as a constantly changing process of production, reproduction, and consumption, its boundaries continually shifting and its meaning always open to renegotiation.

See also: Constructivism/Constructionism: Methodology; Constructivism in Cognitive Psychology; History and the Social Sciences; History of Science; History: Overview; Scientific Knowledge, Sociology of; Universities, in the History of the Social Sciences

Bibliography

Biagioli M 1993 *Galileo, Courtier: The Practice of Science in the Culture of Absolutism*. University of Chicago Press, Chicago

Desmond A 1989 *The Politics of Evolution: Morphology, Medicine, and Reform in Radical London*. University of Chicago Press, Chicago

Ezrahi Y 1990 *The Descent of Icarus: Science and the Transformation of Contemporary Democracy*. Harvard University Press, Cambridge, MA

Forman P 1971 Weimar culture, causality, and quantum theory, 1918–1927: Adaptation by German physicists and mathematicians to a hostile intellectual environment. *Historical Studies in the Physical Sciences* **3**: 1–115

Galison P 1987 *How Experiments End*. University of Chicago Press, Chicago

Galison P, Warwick A (eds.) 1998 *Cultures of Theory. Special Issue: Studies in History and Philosophy of Modern Physics*. Pergamon, Oxford, UK

Gieryn T 1999 *Cultural Boundaries of Science: Credibility on the Line*. University of Chicago Press, Chicago

Golinski J 1992 *Science as Public Culture: Chemistry and Enlightenment Britain, 1760–1820*. Cambridge University Press, Cambridge, UK

Gooding D, Pinch T, Schaffer S (eds.) 1989 *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge University Press, Cambridge, UK

Iliffe R 1992 In the warehouse: Privacy, property and priority in the early Royal Society. *History of Science* **30**: 30–68

Latour B 1988 *The Pasteurization of France*. Harvard University Press, Cambridge, MA

Lawrence C, Shapin S (eds.) 1998 *Science Incarnate: Historical Embodiments of Natural Knowledge*. University of Chicago Press, Chicago

Morus I R 1998 *Frankenstein's Children: Electricity, Exhibition and Experiment in Early Nineteenth-century London*. Princeton University Press, Princeton, NJ

Porter T 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ

Rudwick M 1985 *The Great Devonian Controversy: The Shaping of Scientific Knowledge among Gentlemanly Specialists*. University of Chicago Press, Chicago

Schaffer S 1988 Astronomers mark time: Discipline and the personal equation. *Science in Context* **2**: 115–45

Schaffer S 1992 Late Victorian metrology and its instrumentation: A manufactory of ohms. In: Bud R, Cozzens S (eds.) *Invisible Connections: Instruments, Institutions and Science*. SPIE Optical Engineering Press, Bellingham, WA

Secord A 1994 Science in the pub: Artisan botanists in early nineteenth century Lancashire. *History of Science* **32**: 269–315

Shapin S 1979 The politics of observation: Cerebral anatomy and social interests in the Edinburgh phrenology disputes. In: Wallis R (ed.) *On the Margins of Science: The Social Construction of Rejected Knowledge*. University of Keele Press, Keele, UK

Shapin S 1994 *A Social History of Truth: Civility and Science in Seventeenth-century England*. University of Chicago Press, Chicago
Shapin S, Schaffer S 1985 *Leviathan and the Air-pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, Princeton, NJ
Sibum H O 1995 Reworking the mechanical value of heat: Instruments of precision and gestures of accuracy in early Victorian England. *Studies in History and Philosophy of Science* **26**: 73–106
Smith C, Wise N 1989 *Energy & Empire: A Biographical Study of Lord Kelvin*. Cambridge University Press, Cambridge, UK
Smith R A 1991 National observatory transformed: Greenwich in the nineteenth century. *Journal of the History of Astronomy* **22**: 5–20
Wise N (ed.) 1995 *The Values of Precision*. Princeton University Press, Princeton, NJ

I. R. Morus

History of Technology

Historians interested in technology have written articles and books about the subject for at least since 1900. Only since World War II, however, has the subject taken on some of the attributes of a discipline through the establishment of graduate programs and a scholarly journal of international reputation. History of technology includes the history of invention, the history of engineering, and social history of technology. Scholars with varied backgrounds, some formally trained in the history of technology, contribute to the field.

1. *Defining the History of Technology*

Because technology defies simple definition, its history as a field of study is not easily defined. Editors of the five-volume *History of Technology* (1954–1958, Oxford University Press, London) broadly define 'technology' as 'how things are commonly done or made … [and] what things are done and made.' Finding this definition too diffuse, the editors of the two-volume *Technology in Western Civilization* (1967, Oxford University Press, New York,) reject, as well, the oft-encountered definitions of technology as applied science or as a rational attempt to control nature. They associate technology with human work, 'with man's attempts to satisfy his wants by human action on physical objects.'

Technology and its history are as difficult to define as history of politics or history of economics which experienced scholars refuse to define except by countless examples.

It is also hard to characterize a historian of technology. Unlike general historians, who usually have graduate training in history departments, historians interested in technology enter upon the study of its history from a variety of backgrounds, besides graduate study in the history of technology. Persons trained, for example, in engineering, history of science, economic history, general history, and sociology, do outstanding works in the field. Professional writers also contribute works addressed to a general public including essays that appear in widely circulated journals such as the American *Heritage of Invention and Technology*.

2. *Handbooks and Histories of Invention and Engineering*

During the nineteenth and early twentieth centuries, scholars, especially German, nurtured the history of technology by publishing information-rich encyclopedias and handbooks. Franz Marie Feldhaus's *Die Technik der Vorzeit, der Geschichtlichen Zeit und der Naturvölker* (1914, Englman, Leipzig) and Ludwig Darmstaedter's *Handbuch zur Geschichte der Naturwissenschaften und der Technik* (1908 Springer, Berlin) continue to be useful sources for the history of invention and discovery.

In the early years, biographies of engineers and histories of engineering appeared alongside histories of invention. Samuel Smiles, a railroad manager and medical doctor, in his influential *Lives of the Engineers* (1861–2, Murray, London,) cast problem-solving, project-managing engineers of the British Industrial Revolution in a heroic mold. They surmounted technical, political, economic, and organizational hurdles to transform undeveloped land into industrial regions by means of canals, roads, and improved harbors. In the twentieth century engineers have played a more enigmatic role than did Smiles's heroes. Karl-Heinz Ludwig in *Technik und Ingenieure in Dritten Reich* (1979, Princeton University Press, Kronigstein), Kendall Bailes in *Technology and Society under Lenin and Stalin* (1978, Harvard University Press, Princeton, NJ), and Loren Graham in *The Ghost of the Executed Engineer: Technology and the Fall of the Soviet Union* (1993, Harvard University Press, Cambridge, MA) follow the tragic path of engineers under totalitarian regimes.

3. *Early Monographs*

Lewis Mumford, a public intellectual and professional writer, remains a seminal contributor to the field. Painting with broad-brush strokes on a canvas of western history, his approach is imaginatively interdisciplinary as he draws upon secondary works in western history, sociology, and anthropology. Prolific, he appealed to an interested, educated public, and

simultaneously influenced the scholarly world with his insightful concepts. *Technics and Civilization* (1934, Harcourt Brace, New York), his most enduring history of technology, surveys the development of technology from pre-historic to modern times. He acknowledges the deterministic role technology often plays in history, but he also stresses that values shape technology. His later works, including *The Myth of the Machine: The Pentagon of Power* (1970, Harcourt Brace Jovanovich, New York, NY), lament the rise of megamachines, or large bureaucratic, military, technological systems that deny individual's freedom and sweep them towards catastrophic wars.

Swiss architect and historian Siegfried Giedion's *Mechanization Takes Command* (1948, Oxford University Press, New York) ranges from medieval to modern times. A highly original study, it seeks to explain the mechanization of dwellings, food, furniture, and other aspects of everyday life. Anticipating the attitude of post-World War II historians of technology, he laments in particular the way in which mechanization has displaced organic processes and substituted abstract reason for feeling.

Louis C. Hunter's *Steamboats on the Western Rivers, an Economic and Technological History* (1949, Harvard University Press, Cambridge, MA), and Lynn White Jr.'s *Medieval Technology and Social Change* (1962, Clarendon Press, New York) also anticipate the contextual and social history approaches of a later generation of historians of technology. Hunter shows how the natural environment of the Mississippi River shaped the design of nineteenth century steamboats. On the other hand, White believed technology brought manoralism and feudalism (see *Technological Determinism*).

4. General Histories of Technology

The publication of the information-rich *A History of Technology* (1954–8, Oxford University Press, London) edited by Charles Singer, E. J. Holmyard, A. R. Hall, and Trevor Williams gave the field substantial scholarly credentials. Museum curators, engineers, and natural scientists figure prominently among the authors of the more than 100 chapters that tell of the development of technology from its beginnings to 1900. The authors do not integrate political, social, and economic history into their deterministic account of a nature-mastering, society-shaping, western technology that, they believe, had become science-based by 1900.

Published about a decade later, *Technology in Western Civilization* (1967, Oxford University Press, New York), another collective work, differs notably in approach from *A History of Technology*. Trained in general history before turning to history of technology, editors Melvin Kranzberg and Carroll Pursell, Jr., in contrast to the editors of *A History of Technology*, urged their 34 authors to integrate history of technology with social, economic, and political history. The majority are historians of science, historians of technology, and economic historians, and they tend to take a contextual approach and to focus upon the British and American industrial revolutions. By stressing the interplay of technological and social factors, Kranzberg and Pursell move away from technological determinism encountered in *A History of Technology*.

Kranzberg and Pursell anticipate the recent trend among historians of technology to emphasize the interaction of technology and the natural environment by including three chapters on land use and resources. In a concluding chapter about the challenge of technology for the future, they also take a more critical view of the interaction of technology and culture than did the editors of *A History of Technology*.

Other broad ranging histories of technology have emerged from several national sources. Under the general editorship of Maurice Daumas, *Histoire Générale des Technique* appeared between 1962 and 1979 (Paris). *Propyläen Technikgeschichte* (Berlin 1990–1992), a German collective survey under the general editorship of Professor Wolfgang König, also moves from ancient times into the twentieth century. A. A. Sworykin, N. I. Osmova, W. I. Tschernyschew, and J. W. Schuchardin published a Soviet version of the general history of technology translated into German as *Geschichte der Technik* (Leipzig 1964).

Recognized as a major scholarly achievement, Joseph Needham's multivolume *Science and Civilization in China* (Cambridge, UK, 1954) integrates from a Marxist perspective the history of medicine, technology, and science in a non-Western civilization. Among national histories of technology is *Geschiedenis van de Techniek in Nederland: de Wording van een moderne Samenleving, 1800–1890* (Zutphen 1992–1993) under the general editorship of Harry W. Lintsen.

5. Society for the History of Technology

The year of 1958 brought a turning point in the development of the history of technology with the establishment of the Society for the History of Technology (SHOT) and its journal *Technology and Culture* (*T&C*). Earlier, in 1920, the British Newcomen Society for the Study of the History of Engineering and Technology began publishing *The Transactions* which attracted, in particular, engineers interested in the history of their profession. The influential French journal, *Les Annales*, occasionally carried articles about the history of technology, and in 1935 published a special issue on the subject with a wide-ranging article on the history of water mills by its distinguished editor Marc Bloch.

Originally edited by Kranzberg and subsequently by Robert Post and John Staudenmaier, *Technology and Culture* attracts essays and readers from history of science, technology, and economics. The annual bibliography of the history of technology prepared initially by Eugene Ferguson proved especially valuable for researchers in the field. Within a few years, SHOT and *Technology and Culture* became the primary organizational framework for historians of technology in the USA and Western Europe. With several thousand individual members, SHOT includes non-US members on its governing committees and regularly has annual meetings in Western Europe. The International Committee for the History of Technology (ICOHTEC) provides through its colloquia and publications, a forum not only for US and Western European historians of technology, but also those from Eastern Europe, Russia, and elsewhere.

6. Graduate Programs in the History of Technology

Not only did SHOT and *Technology and Culture* provide an institutional framework for the history of technology, but the establishment of several graduate programs did as well. In 1954 the Eleutherian Mills-Hagley Foundation of Wilmington, Delaware with its archives, library, and museum, concentrating on history of technology and business organized, in conjunction with the University of Delaware, a graduate program for historians of technology. In 1973 the University of Pennsylvania initiated a doctoral program for the history of technology within its History and Sociology of Science Department. Other US universities followed suit, including the Massachusetts Institute of Technology. In Europe, Svante Lindqvist founded a history of technology and science program at the Royal Institute of Technology in Stockholm which has become a model for other institutions of higher learning in Western Europe. The research centers at Munich's Deutsches Museum and London's Imperial College also cultivate graduate training in the field.

7. Emerging Themes

7.1 Technology–Science Relationship

Since its founding, the pages of *Technology and Culture* express thematic concerns of scholars publishing and teaching in the field. In the 1960s, they often explored the relationship between technology and science. Motivation for this interest came in part from their wish to establish an identity for their subject other than as the history of applied science, a linear model often adopted by historians of science and science policy makers. Historians of technology point out how often scientific theory rationalizes earlier technological practice.

Historian of technology Edwin Layton argued cogently in the 1970s that engineers develop their own applicable theoretical structures from practical experience and experimentation. Recently Walter Vincenti, eminent aeronautical engineer and historian of technology, has expounded this position persuasively in *What Engineers Know and How They Know It* (1990, Johns Hopkins University Press, Baltimore, MD).

Besides writing biographies of Thomas Edison and other independent inventors who experimented beyond scientific frontiers, historians of technology and science publish works about industrial research. John Beer provided a model for such studies with his essay on [German] 'Coal Tar Dye Manufacture and the Origins of the Modern Industrial Research Laboratory,' *ISIS* **49**, (1958). Histories of the General Electric, Bell, and other major industrial laboratories followed. In these, fundamental research, in contrast with pure research, brings improvements in existing technology such as the three-element vacuum tube and incandescent lamp filaments.

7.2 Invention, Development, and Innovation

Technological change and creativity, a perennially intriguing human activity, became another early thematic concentration of historians of technology, especially those interested in economic and business history who see innovation as a source of economic development. After World War II, *The Sources of Invention* (Jeukes et al., 1969, W.W. Norton, London), which includes more than 50 case histories, stimulated a shared discourse about invention, as did historian Abbott Payson Usher's *A History of Mechanical Invention* (Cambridge, MA, 1954) and sociologist S. Colum Gilfillan's *Supplement to the Sociology of Invention* (San Francisco, 1971). Sociologist of science Robert K. Merton's seminal study, *Science, Technology & Society in Seventeenth Century England* (New York, 1970), broadens the discourse by demonstrating that attitudes and values stimulate technological innovation which in turn stimulates the search for underlying scientific knowledge.

The shared discourse usually defined technological creativity as consisting of three phases: invention, development, and innovation. Professional inventors and researchers in industrial laboratories were seen as presiding over the process of invention from a concept to a physical model. Development, a collective venture, moved the simple model through testing stages until a full-scale version could be tested. Innovation brought the device, process, or system into use. *Technology and Culture* authors usually situated the

three-phase process in an economic, political, and social context. The ongoing publication of the extensive Thomas Edison papers, along with Paul Israel's *Edison, a Life* (2000, New York) based on these, provides historians with an unprecedented rich source of information on a master of invention, development and innovation.

7.3 Industrial Revolutions, Systems, and Management

Industrialization also stimulated research and teaching among historians of technology. The rise of the USA to a position of preeminence among industrial nations raised questions about the nature of the industrialization process, especially as embodied in putative industrial revolutions in the UK and the USA.

Influential scholars exploring the process tend to see it as systematic. In *Das Kapital* (1867), Karl Marx portrays a factory as a system of textile machines driven by a single steam engine. Widening his perspective, he insightfully understands that textile factories interact with transportation, energy, and other systems to produce an industrial revolution, one scholars now label the 'British Industrial Revolution.'

David Landes an American economic historian, in *The Unbound Prometheus* (1972, Cambridge University Press, Cambridge, UK), writes of an interrelated succession of technological changes culminating after 1750 in the British Industrial Revolution and its spread throughout Western Europe. He shows how the coming of the steam engine, for example, depended upon developments in coal mining, iron manufacture, and railroad transportation. Ken Alder discusses the systematic interaction of military culture, industrialization, and politics in *Engineering the Revolution: Arms and the Enlightenment in France, 1763–1815* (1997, Princeton University Press, Princeton, NJ). *Histoire des Techniques* (Paris, 1978), edited and mostly written by Bertrand Gille, presents a history of successive technological systems and industrial revolutions from the earliest times into the twentieth century.

Historians customarily deal with the industrialization of the USA in the early nineteenth century as the spreading of the British Industrial Revolution by technology transfer and adaptation. David Hounshell in *From the American System to Mass Production, 1800–1932* (1984, Johns Hopkins University Press, Baltimore) challenges this interpretation by stressing the originality of the American system of production that developed in the nineteenth century and reached its apogee in the Ford system of mass production. Historian of technology Merritt Roe Smith in *Harpers Ferry Armory and the New Technology* (1977, Ithaca, NJ) sees the early nineteenth century American system as both a technical and managerial enterprise, and places it within a shaping community, or social, context.

Only a relatively few historians explore a second Industrial Revolution in the USA and Germany after 1870 with its technical core of electrical and internal combustion technology.

Hugh G. J. Aitken surveys the spread of radio in *The Continuous Wave: Technology and American Radio, 1900–1932* (1985, Princeton University Press, Princeton, NJ) and Susan Douglas amplifies the story in Susan Douglas, *Inventing American Broadcasting, 1899–1922* (1987, Johns Hopkins University Press, Baltimore, MD). Within the frame of a Second Industrial Revolution, Leslie Hannah gives the history of electrical supply in the UK in *Electricity before Nationalisation* (1973, Johns Hopkins University Press, Baltimore, MD). *Networks of Power: Electrification in Western Society, 1880–1930* (1983, Johns Hopkins University Press, Baltimore, MD) by Thomas P. Hughes gives a history of electric power systems since 1880 in major urban centers in the USA, the UK and Germany. He defines systems as embodying in a seamless web a multitude of technical and organizational components and follows their development through stages characterized respectively and successively by the problem solving of inventive, managerial, and financial/political system builders. Hughes's methodology has spread among historians and sociologists who characterize it as a pragmatic, systems approach.

US historian of management Alfred D. Chandler in his seminal work *The Visible Hand: The Managerial Revolution in American Business* (1977, Cambridge, MA) stimulated a number of historians of technology interested in a Second Industrial Revolution to consider his proposition that innovative management since the late nineteenth century has done more to increase productivity than technical advances. Jurgen Kocka describes in a broad context large-scale, enterprise management at the Siemens company during German industrialization in *Unternehmensverwaltung und Angestelltenschaft am Beispiel Siemens, 1847–1914* (1969, Stuttgart), James R. Beniger in *The Control Revolution: Technological and Economic Origins of the Information Society* (1986, Harvard University Press, Cambridge, MA) moves from a second to an information revolution.

David Noble in *America by Design: Science, Technology, and the Rise of Corporate Capitalism* (1977, Alfred A. Knopf, New York) takes a strikingly different approach from Chandler on the rise of modern management. Reasoning from a Marxist perspective, Noble contends that US corporate culture represented by engineers and industrial scientists co-opted invention, research, and technical education in the interest of capitalism. Philip Scranton, a historian of technology and labor, offers in *Endless Novelty: Specialty Production and American Industrialization, 1865–1925* (1997, Princeton University Press, Prince-

ton, NJ) a correction to both Chandler and Noble by pointing out that small-scale, proprietary management and batch production continued to flourish in the era usually associated with large-scale mass production.

Military technology has been a constituent factor of modern history, especially during and since the Second Industrial Revolution. William H. McNeill in *The Pursuit of Power: Technology, Armed Force, and Society Since A.D. 1000* (1982, University of Chicago Press, Chicago) integrates military technology into general history. Technological developments during World War II and after, are described in Edward W. Constant's *The Origins of the Turboiet Revolution* (1980, Johns Hopkins University Press, Baltimore, MD). Richard G. Hewlett and Oscar E. Anderson, Jr. provide a detailed account of wartime atomic energy in *A History of the United States Atomic Energy Commission: The New World, 1939/1946* (1962, Pennsylvania State University Press, University Park, PA) and Richard Rhodes dramatizes the story effectively in *The Making of the Atom Bomb* (1986, Simon and Schuster, New York). Walter A. McDougall *The Heavens and the Earth* (1985, Basic Books, New York) places space technology in a political context.

Donald MacKenzie in *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance* (1991, MIT Press, Cambridge, MA) and Gabrielle Hecht in *The Radiance of France: Nuclear Power and National Identity after World War II* (1998, MIT Press, Cambridge, MA) place political decision-making in the shaping context of evolving nuclear technology. A number of historians are exploring the recent history of the computer and biotechnology, but the definitive history has not yet appeared. Paul Edwards *The Closed World: Computers and the Politics of Discourse in Cold War America* (1996, MIT Press, Cambridge, MA) goes beyond narrative history to suggest that information systems are increasingly the shaping context within which military and political decisions are made.

7.4 Environment: Natural and Built

During the 1990s, the history of technology and environmental history intersected in the works of a number of historians. Earlier Leo Marx in *The Machine in the Garden: Technology and the Pastoral Ideal in America* (1964, Oxford University Press, New York) discussed the ideas of early American visionaries, such as Ralph Waldo Emerson who envisioned the harmonization of nature and technology in a pastoral landscape. In *Nature's Metropolis* (1991, W.W. Norton, New York). William Cronon describes the way in which Chicago's transportation and manufacturing technologies interacted seamlessly with the exploitation of the hinterland's natural resources to produce enormous quantities of marketable goods and nurture the enormous growth of Chicago.

Richard White, a social and environmental historian, further explores the seamless web theme in an imaginatively titled book *The Organic Machine* (1995, Hill and Wang, New York). It tells of the transformation of the Columbia River Basin in the states of Oregon and Washington into an artifact blending the natural and human built. White believes that the history of technology and the history of the environment can no more be told separately than the story of a marriage.

7.5 Gender, Race, and Technology

In the 1990s historians also began to explore assiduously the mutual shaping of gender and technology. 'Gender Analysis and the History of Technology,' a bibliographic essay in a special issue of *Technology and Culture* **38** (1997) documents the extensive literature emerging in the field. Ruth Schwartz Cowan in *More Work for Mother: the Ironies of Household Technology from the Open Hearth to the Microwave* (1983, Basic Books, New York) demonstrates that the household is a context within which factory-produced goods and consumption interact. Other historians further develop the consumption junction theme by showing that consumers, male and female, are not simply passive users of technology, but shape artifacts through their consumption preferences.

Judith McGaw opens another area of gender studies in *Most Wonderful Machine: Mechanization and Social Change in Berkshire Paper Making 1801–1885* (1987, Princeton University Press, Princeton, NJ) in which she demonstrates that women have gendered niches in the workplace which often perpetuate inequalities. Other historians argue similarly that some technological developments are shaped by ideas about racial differences and conversely that technology often defines racial roles, especially in colonial and developing areas. Michael Adas in *Machines as the Measure of Man: Science, Technology, and Ideologies of Western Dominance* (1989, Cornell University Press, Ithaca) demonstrates that nineteenth century colonial ideology associated technological progress, or the lack of it, with racial characteristics.

8. Methodology: Social Construction and Actor Networks

Until the 1980s histories of technology were generally written as narratives with little formal analysis. Then, European sociologists and anthropologists began shaping the field through the introduction of social construction and actor network methodologies. (See *Technology, Social Construction of*; *Actor Network Theory*.) French sociologist Michel Callon in several

influential articles including 'The State and Technical Innovation: A Case Study of the Electrical Vehicle in France,' *Research Policy*, **9** (1980) presents an actor-network model. It draws upon, but differs from, the Hughes systems approach previously described.

Callon uses an abstract concept 'actor' to signify—without distinguishing between—animate and inanimate components in a network. An actor-network creates and sustains a complex technological artifact such as an electric automobile. French professor of sociology Bruno Latour in *Aramis, ou l'Amour des Techniques* (1993, Paris) utilizes actor-network theory to present the history of a failed effort to introduce a radically new subway system in Paris. Deploying relativist sociology, Latour lets the actors, animate and inanimate, speak for themselves and offer conflicting, but equally valid, accounts of the evolution of a fluid project characterized by conflicting visions and policies.

Together with the systems and actor-network, the social construction approach of Trevor Pinch and Wiebe Bijker became a widely adopted methodology among historians and sociologists of technology in Europe and the USA. Influenced by the sociology of scientific knowledge cultivated at the University of Edinburgh, Pinch and Bijker offer a seminal example of social construction in an essay entitled 'The Social Construction of Facts and Artifacts,' *Social Studies of Science*, **14** (1984). They demonstrate that social environments shape the technical characteristics of an artifact, in this case the late nineteenth century bicycle. Social groups of varying class and different gender pursuing their interests and values define and determine the solution of bicycle-development problems thus embodying meaning in bicycle design.

Essays by Pinch and Bijker, Hughes, Callon, John Law ('Technology and Heterogeneous Engineering: the Case of Portuguese Expansion'), Donald MacKenzie ('Missile Accuracy: A Case Study in the Social Processes of Technological Change'), and by others are collected in *The Social Construction of Technological Systems; New Directions in the Sociology and History of Technology* (1989, MIT Press, Cambridge, MA). From this volume a language of analysis has spread, a language which includes 'seamless web,' 'systems,' 'actors,' 'network,' 'closure,' 'stabilization,' 'conservative and radical innovation,' 'technological momentum,' 'system builder,' 'heterogeneous engineer,' 'reverse salients,' 'interpretive flexibility,' 'translation,' and 'black box.' *The Social Shaping of Technology* (1998, Milton Keynes), edited by Donald MacKenzie and Judy Wajcman, also explores methodological and social construction themes. These concepts and the new approaches to the history of technology did much to discredit technological determinism so widely accepted among the laity, so persuasively argued by Karl Marx, and so eloquently enunciated by White in *Medieval Technology and Social Change*.

See also: Actor Network Theory; Business History; Economic History; History of Science; Industrialization, Typologies and History of; Science, Technology, and the Military; Scientific Revolution: History and Sociology; Technological Determinism; Technological Innovation; Technology and Organization; Technology and Social Control; Technology, Anthropology of; Technology, Social Construction of

Bibliography

Buchanan R A 1991 Theory and narrative in the history technology. *Technology and Culture* **32**: 365–76

Daniels G H 1970 The big questions in the history of American technology. *Technology and Culture* **11**: 1–21

Daumas M 1969 L'histoire des techniques: son objet, ses limites, ses méthodes. *Documents pour l'histoire des techniques* **7**: 5–31

Ferguson E S 1974 Toward a discipline of the history of technology. *Technology and Culture* **15**: 13–30

Hindle B 1966 The exhilaration of early American technology. In: Hindle B (ed.) *Early American Technology*. University of North Carolina Press, Chapel Hill, NC, pp. 3–28

Hounshell D A 1981 On the discipline of the history of American technology. *The Journal of American History* **67**: 854–65

Hounshell D 1989 Rethinking the history of 'American technology'. In: Cutcliffe S H, Post R C (eds.) *In Context: History and the History of Technology: Essays in Honor of Melvin Kranzberg*. Lehigh University Press, Bethlehem, PA, pp. 217–20

Hounshell D 1995 Hughesian history of technology and Chandlerian business history: parallels, departures, and critics. *History and Technology* **12**: 205–24

Pursell C 1978 History of technology. In: Durbin P (ed.) *A Guide to the Culture of Science, Technology, and Medicine*. Free Press, New York

Rürup R 1972 Die Geschichtswissenschaft und die moderne Technik. Bermerkungen zur Entwicklung und Problematik der technikgeschichtlichen Forschung. In: Kurze D (ed.) *Theorie und Praxis der Geschichtswissenschaft*. Berlin, pp. 49–85

Science Museum 1992 *Guide to History of Technology in Europe*. Lists history of technology programs, journals, and research scholars. London

Staudenmaler J M 1985 *Technology's Storyteller: Reweaving the Human Fabric*. The MIT Press, Cambridge, MA

Staudenmaier J 1990 Recent trends in the history of technology. *The American Historical Review* **95**: 715–25

T. P. Hughes

History: Overview

1. Anthropological Universals

Human beings must interpret their world, themselves, and their coevals in order to live. Human activities, sufferings, and desires are formed by definitions of meaning and of purpose, which incorporate such

interpretations. An important area for such interpretation is the experience of time, and here memory plays a central role. The interpretations of the present that disclose it to us as a world, and allow us to act in this world, always refer to the past; the past is differentiated from the present and is thus related to it in that the future can be shaped as a perspective on action.

On this general anthropological level, history is the totality of the cultural practices through which human beings interpret their past in order to understand the present and anticipate the future. In other words: history is the generation of sense out of the experience of time.

But not every interpretation of time is historical. What determines the historical quality of an interpretation is the fact that time confronts human beings in two ways: as *real change within the world*, in which contingent events occur and are experienced, and as an *interior, temporal extension of consciousness* (and of the unconscious). Here, intentions and objectives, guided by the elemental mental forces of fear, anxiety, desire, hope, and longing, play a central role. We interpret the 'objective' time of real change (the present) through the recollection of another time (the past); thus the 'subjective' time into which we project a future is bound to our approach to past experiences. This is human culture's specifically historical procedure for experience and interpretation. What defines the historical interpretation of time is its implicitly meaningful and thus orientative conception of the course of time. The latter comprises the past, the present, and the future, and at the same time articulates experiences and intentions; it is thus empirical and normative at the same time. The philosophy of history explicates this conception of the course of time as a particular mental structure. Even without philosophical reflection or justification, such a conception of the course of time organizes historical thought, and shapes its content. History interprets the contingent events of the past and defines their significance for the present. It organizes historical narrative as a particular kind of explanation (Danto 1965). It is this 'explanation through narration' that defines the logic of historical thought in its academic form as well.

2. *Conceptions of Time*

The different conceptions of the course of time important for historical interpretation employ various temporal categories in various constellations (e.g., cyclical time, linear time, the concept of the *kairos*), each encouraging particular expectations: decline, progress, permanence, recurrence, revolution, destruction, and so on. They relate real temporal processes to the active intentions and self-conceptions of the subjects that make use of them. In other words, history mediates the subjectivity of acting and suffering human beings with the objectivity of the temporal process of their world. In early societies, this was accomplished by an appeal to divine forces, which impel the transformation of the world and the human spirit at the same time; in other societies these divine forces are replaced by humankind as a spiritual and moral being, its ability to create symbolic worlds or to plan the conquest of the natural and social world through knowledge. Finally, nature, too, can be invoked as a sort of metaphysical instance for the continuity between the two temporal dimensions of human historicity. Nature can thus be understood as that which compels the human race, by means of social antagonisms, to create culture (as in Kant, for example), or as struggle of the races in Social Darwinist terms. Psychoanalysis, on the other hand, posits an interior, psychological nature, in order to explain the historical movement of the human world by means of the dynamics of the unconscious.

The idea of a 'history' comprising all temporal change within the human world is of modern origin (Koselleck 1985). However, it has a premodern equivalent in older cosmogonies, theogonies, and anthropogonies, as well as in the concept of redemptive history.

3. *History as Event and Narration*

In every culture and epoch, 'history' has a double meaning. On the one hand, it designates temporally organized events in the past. At the same time, and inseparably from the first meaning, 'history' also means the account of these events that articulates their significance for our understanding of ourselves and of our world. The first meaning has to do with temporal succession and the second with its narrative representation.

Not everything that has happened can be called history, but rather only that which is somehow meaningful in the present; for example, the foundation of a city or a kingdom is significant to its inhabitants, a person's birthday is significant to him or her, and so on. Usually, this event is described as a sequence of occurrences. These may often be the actions of human beings, heroes, or divine beings, as in myth or epic; however, they may also involve more complicated phenomena, such as the development of a manufacturing process, the evolution in the meaning of a concept, the epochal transformation of a mentality, an 'axial time,' and so on.

In its second meaning, history is the representation of the past in the culture of the present. Here, 'history' is the product of historical consciousness: through memory, the past has become present and meaningful in narrative form. In this way, the past 'in itself,' that is, everything that has happened, and persists only in the form of trace, relic, memorial, tradition, ethic, and so

on, becomes history 'for us.' The basic mental procedures involved can be organized somewhat artificially into four:

(a) the *perception* of 'another' time as different: the fascination of the archaic, the obsolete, the mysterious trace, the insistent memorial, and so on;

(b) the *interpretation* of this time as temporal movement in the human world, according to some comprehensible aspect (e.g., as evidence of the permanence of certain values, as examples of a general rule, as progress, etc.);

(c) the *orientation* of human practice through interpretation—both 'outwardly,' as a perspective on action (e.g., as the increase of political legitimacy through political participation, as the restitution of the world before its destruction, as the institution of 'true' conditions against the decline of morals) and 'inwardly,' as identity conceptions (e.g., 'we are the children of the sun,' or 'we as a nation stand for the universality and fulfillment of human rights,' or 'we belong to the communion of saints,' or 'we represent true spirituality as against others' materialism');

(d) and finally, the *motivation* for action that an orientation provides (e.g., a willingness to sacrifice, even die, for the sake of historical ideas of national greatness, the missionary spirit, etc.).

4. Memory, History, and Metanarratives

The relationship between memory and historical consciousness continues to be debated. Memory in the everyday sense is directed toward the experiences of individuals in their own lives, whereas historical consciousness primarily thematizes a past that exceeds an individual life span. Both, however—the individual memory on which individuality and social affiliation are built as well as the expansion of 'history' beyond the limits of a single lifespan back into the past—are as inseparable as two sides of a coin: human beings tend to transform their own identity into temporally more inclusive structures, such as the communion of saints, the nation, the people, the culture, and so on. The basis of individual subjects' self-esteem and of their practical temporal orientation tends to exceed their lifespan or to outlive them. It therefore usually extends far back into the past.

'History' generally performs its orientative function through a 'master narrative' or 'metanarrative.' This describes the conception and evolution of the addressee's world in order to legitimate its normative framework. At the same time it provides an experiential inventory, which is available whenever problems of orientation require consensus. It legitimates not only systems of politics, economics, and the environment, but also the collective and individual subjectivity and identity of the people. Metanarratives represent social affiliation and differentiation internally and externally.

Metanarratives generally organize social inclusion and exclusion *ethnocentrically*, assigning one group a positive normative weight, and other different groups a negative weight. One's own life world appears as culture or civilization, other worlds are looked upon as savagery and barbarism. This difference can be articulated in different ways: as an opposition, as deviation from a norm, as atavism on a progressive scale, etc. Just as history presents past experiences as a 'mirror' in which the present perceives itself, the image of others in the same historical mirror has a contrastive function: it defines a self-image. Under the special conditions of an orientation crisis, the 'other' can assume admirable features, as in the image of the 'noble savage.'

5. Historical Culture

The interpretive work of historical consciousness and its product called 'history' is concretely manifested in a society's historical culture. Historical culture is multidimensional, like every other culture. It has religious, moral, pedagogical, political, and rhetorical expressions; its cognitive substance is always the knowledge of '*wie es eigentlich gewesen*,' how it really was. We can distinguish three basic dimensions of historical cultures as an ideal type, each different in its logic and thus accountable to different criteria of meaning.

(a) The *political* dimension is concerned with the legitimation of a certain political order, primarily relations of power. Historical consciousness inscribes these, so to speak, into the identity concepts of political subjects, into the very construction and conception of the I and the We, by means of metanarratives that answer the question of identity. There is no political order that does not require historical legitimation. The classical example, applicable to every culture and every epoch, is genealogy. Even the pure rule of law that appeals only to the applicability of formal decision procedures must be historically based if these procedural rules are to be plausible to the participants. Charismatic leadership also cannot do without historical elements. Generally, the vehicle of political charisma will refer to the spiritual or natural forces that, within his culture, guarantee the world's temporal coherence.

(b) The *esthetic* dimension is concerned with the psychological effectiveness of historical interpretations, or that part of its content that affects the human senses. A strong historical orientation must always engage the senses. Masquerades, dances, and music can all have historical content. Many older metanarratives are composed in poetic form and are celebrated ritually. A formal defect can destroy the effect of such a presentation, and even endanger the world's continued coherence. Historical knowledge must employ literary models to become discursive. In

many cultures, historical narrative occupies a secure place in the literary canon as a genre of its own. In modern societies, memorials, museums, and exhibitions are among the familiar repertoire of historical representation. In older kinds of social systems, objects such as relics, tombs, temples, and churches obligate the present to the legacy of the past, indeed make the present, in its relationship to the future, responsible for the vitality of historical memory.

(c) The *cognitive* dimension is concerned with the knowledge of past events significant for the present and its future. Without the element of knowledge, the recollection of the past cannot effectively be introduced into discourses concerned with the interpretation of current temporal experience. Mythical metanarratives, too, have a cognitive status, although science would eventually deprive them of it. They lose their orientative power when confronted with a science of the past that possesses a more elaborated relation to experience. Metanarratives then become prosaic, as they already did in antiquity, with Herodotus and others. In modern societies, the cognitive dimension of history has become an academic discipline.

6. *Typologies of the Historical*

Synchronic and diachronic typologies of the historical can be generated out of the particulars of historical thought, and inquire into its carriers, functions, class and gender specificity, and other factors within the social context's framework of determinations (Rüsen 1996). They become historically specific, however, only when they differentiate the specific interpretative work of historical consciousness and of its product—'history' as the cultural orientation system of temporal consciousness—according to the different possibilities of historical meaning construction. Such differentiation from a metacultural perspective has been little attempted. Previous efforts have usually focused on functional and formal aspects, and have thematized less the total domain of human temporal interpretation than simply the domain of historical writing, and if at all they only inspected the Western tradition. Nonetheless, such typologies can also be used as analytical tools for research in the wide areas of historical culture and for intercultural comparisons.

Nietzsche (1985), for example, differentiated the monumental, the antiquarian, and the critical uses of history in terms of their advantages and disadvantages for life:

(a) the *monumental* approach inspires action by recalling the creative deeds and great moments by which the world was actively transformed, and emphasizes the greatness of man as the agent of historical change;

(b) the *antiquarian* approach serves the preservation of tradition and the affirmation of existing conditions by its reverence for and transfiguration of their origins;

(c) the *critical* approach, finally, defeats the power of memory, breaks the control of tradition, and liberates life for new orientations.

Hayden White (1973) presented a very influential model for the analysis of historical writing, and applied it to the nineteenth century's most significant historians and philosophers of history. It consists of a complex system of emplotment, formal argument, ideological implication, and tropes of historical discourse.

(a) The narrative forms that connect a sequence of historical events can follow the patterns of romance, tragedy, comedy, or satire.

(b) Formal argument no longer concerns so much the narrative combination of events into a story or narrative, but rather the explicit explanations within these narratives. They can be formalist, emphasizing the uniqueness of an event; they can be organicist, integrating isolated phenomena into larger totalities for explanatory purposes; they can argue mechanistically, in terms of causal regularities; and they can operate contextualistically, in terms of functional relationships between events.

(c) Ideological implications relate to the standpoint within the present from which a history develops its perspective on time. Such a perspective (White is most concerned with the nineteenth century) can be anarchistic, conservative, radical, or liberal.

(d) What is decisive for the extensive and foundational meaning of a narrative is the verbal factor that brings the different elements of historical narrative into a coherent whole. White identifies and explains this factor in terms of rhetorical figures or tropes that underpin the four possibilities for historical meaning: the latter is generated out of metaphorical, metonymic, synecdochic, or ironic uses of language.

White developed this typology for the analysis of historiographical texts; nevertheless, it can be applied to other manifestations of history, such as historical museums (Bann 1984).

Rüsen (1993) also developed a universally applicable typology of historical sense generation according to its most basic principles: its particular referents, its conception of temporal movement, its communicative form, and its identity concept. All these factors are communicated through and constituted by the principle of historical sense.

(a) The *traditional type* recalls the origins grounding present relations. Traditional history represents the course of events in terms of the persistence of these origins and their normative power, and represents identity by affirming prior interpretive models of human subjectivity (e.g., social roles). Past and present fuse into the permanence of currently effective social relations that are conveyed by time yet are above transience. Traditional narratives eternalize time as meaning. This type predominates in mythical thought.

(b) The *exemplary type* recalls past facts that concretize the rules of current conditions. Exemplary

history represents a course of events in terms of the transtemporal applicability of these rules, and constructs identity by generalizing various temporal experience into rules for action (e.g., as rule-based competence). This narrative type is represented in Europe by the classical motto *historia vitae magistra*, and in China by the metaphor of the mirror. This type predominates in the historical thought of great civilizations up to the modern period.

(c) The *critical type* recalls past circumstances in order to interrogate present conditions. Critical history represents a course of events as the transformation of prior and, in the orientative framework of life praxis, effective interpretive models (i.e., as rupture or as alternation). It constructs identity by negating pregiven identity-forming interpretive models; in other words as the power to say 'no.'

(d) The *genetic type* recalls qualitative changes in the past by which other, strange conditions become present familiar ones. The genetic construction of meaning presents a course of events as evolution, in which conditions change in order to become dynamically permanent. It constructs identity as the synthesis of permanence and change; in other words as a dynamic process. This type predominates in modern societies.

7. Media

The diverse manifestation of history, in its dual significance as fact and as interpretation, depends crucially on the medium through which temporal experience and interpretation are communicated. Language is always predominant, although there are other creative areas of historical perception and representation, such as the visual and musical arts, architecture, and dance. The modes by which language orders human communication are decisive for the expression of historical consciousness.

(a) History was originally communicated orally, as it is even now—at least in the private sphere. In preliterate cultures, history becomes oral history or oral tradition (Vansina 1985). Such a history is usually politically conservative: the traditional world order becomes obligatory; new experiences have to be integrated into it so as to confirm the collective's basic norms and identity. Such representations are sensually immediate and concrete; they are communicated face to face and correspondingly have immediate effects. Cognitively, they contain the cumulative experience of many generations; experientially, they reach into a mythical time that genealogically connects the present to the all-determinative origin of the world and include the more recent generations' memory which contain real events. Its criteria of significance are of a mythical nature; in other words, topical and in the narrow sense 'historical' events are not the vehicles of orientative normativity. Its validity claims are raised and made plausible through ritual and repetition. In modern historical research, oral history is the method for gaining historical knowledge from the recollections of acting and witnessing people. It provides an important source for research into the mechanisms of biographical memory as an element of historical culture in the intersection of the public sphere and individual biography. The memories of witnesses to historically significant events, communicated orally, acquire an important role in public historical culture—especially the testimony of survivors of great catastrophes such as the Holocaust (Langer 1991).

(b) The medium of *writing* has dominated for millennia and continues to dominate hegemonic historical culture (in which, of course, oral elements continue to be important and effective, even constitutive—at least in education). Written history frees participants from the immediacy of the communicative situation and creates a distance between history as content and history as a form of communication (Ong 1982). This distance increases considerably the experiential horizon of historical consciousness, and facilitates new methods for accumulating and managing experience. Literacy liberates memory, fixes facts, creates new communicative forms, disconnects the cosmos of temporal meaning from immediate practical relationships, and allows it to become a phenomenon *sui generis*. Along with new opportunities for objectivation in the approach to historical experience, it also discloses new opportunities for subjectivation through historical interpretation: the narrator can become an author, and readers can obtain expanded scope for critical readings. Validity claims either are raised through canonization, or become an opportunity for discursive justification. In either case, interpretation becomes a discrete practice of historical meaning construction, with its corresponding specializations and the attendant problems of mediation and popularization. The original relationship between poetry and truth dissolves; mythology is now subject to a fundamental critique. Explicit conceptualization becomes an essential cognitive element in historical interpretation, so that history can, at least in the long term, be an object of study and of theory.

Entirely new media have transformed the historical mode of thought. We cannot yet discern distinct lines of development and stable structures, but we can certainly indicate new developments from which we might expect transformations of the most basic kind.

In the historical culture of the public sphere, collective memory is being overpowered by the torrent of historical images. The forms of consciousness created by literacy—and above all the distancing effect of rationality—may quickly decrease in significance, and especially in political efficacy. The grammar of history is becoming an imagology of presentations in which every era is contemporaneous, and the fundamental idea of a single linear movement of time is

disappearing. The constitutive difference of temporality can be suspended into a universal contemporaneity that can no longer be narratively put in order. Whether there can then be a specific 'historical order' within the orientative temporal continuity between past, present, and future has at least become a problem. The very term *post-histoire*, and the related discussion of a mode of life without genuinely historical interpretation (Niethammer 1992), suggest that these questions are now open. At the same time, there has been an immense increase in access to data. New storage media allow new modes of historical experience, and radically call into question previous criteria of significance. At the same time, new media of communication such as the Internet permit no single, politically sanctioned decision. The abundance of possibilities and the diversity of new voices require new strategies, new forms, and new contents of historically grounded participation and exclusion. In each case, fixed ideas of the permanence and substance of individual and collective identity are being outstripped by the diversity of global communication, in favor of more dynamic and more open differentiations. This process then provokes reactions, often expressed through these new media, that stubbornly insist on ethnocentric distinctions.

8. *Developments*

The transformation in the media of historical thought sketched above suggests a transcultural, macrohistorical transformation, that will ultimately leave no culture unaffected. This provides us with a support for a diachronic classification for historiographical variety. Such a classification could be organized in terms of epochal distinctions, yet still have a typological quality. For example, White's or Rüsen's typologies could both be used to characterize historical processes. White's sequence of tropes can be understood as cyclical, and Rüsen's distinction between traditional, exemplary, and genetic historical thought can describe the great transcultural epochs where the critical type of historical thought mediates the transition between the other types.

One of the most important distinctions concerns the relationship between myth and 'history,' where the latter is understood as mundane occurrence. Myth relates to a divine prehistory as the source of meaning (Eliade 1991). In contrast, 'historical thought' in the stricter sense discerns the meaning of history in the immanent dimension of a chain of events (*res gestae*). The Judeo-Christian sacred history mediates between these two modes of thought, and allows events in the world to appear as the manifestation of God's power and God's will. In other cultures, the divine and the mundane are mediated differently: mundane but somehow significant events have to be integrated into a comprehensive cosmic order. History becomes meaningful through nature, which is conceived as divinely ordered. With increasing complexity in human praxis, more and more mundane incidents are integrated into this order until it is finally 'historicized' and 'denaturalized.' Yet this process is often opposed by attempts to renaturalize it as biological racism.

Comprehensive large-scale developments of historical thought have often been conceptualized as leading to the current status of historical studies. That is often criticized as too one-sided and the diversity of new media and the different dimensions of the historical culture are emphasized instead; nevertheless plausible accounts of a developmental process can be articulated (in the form of typological hypotheses). These accounts can be used to discern and analyze transformations of historical thought in the diversity of different cultures.

Max Weber's conception of a universal process of rationalization and disenchantment seems as relevant as ever (Weber 1988). The classical philosophy of history of Voltaire, Kant, and Hegel had presented these criteria in the form of the actual course of events. To the extent that historical studies acquired increased competence in investigating and representing real events in the past, the question of meaning was investigated within the framework of a logic of historical thought (Dilthey, Max Weber, Rickert). This question is now perceived primarily as the concern of the central verbal procedure of historical thought, as historical narrative (Danto 1965). Issues of scholarly standards and truth claims (Appleby et al. 1994; Evans 1999) can be obscured by only focusing on the rhetorical and poetic strategies of historical writing (Rigney 1990).

Despite this formalization of historical understanding, the structures of real historical processes remain important points of reference. Here, the global and often conflictual transformations brought about by modern industrial societies are in the foreground. Their developmental dynamics can be understood only by a concept of time in which the horizon of experience and expectation structurally diverge, and can only be reconnected by a fundamental temporalization of the human world (Koselleck 1985).

Narratives of progressive rationalization and disenchantment must, of course, be complemented by an account of a critical and transcendent re-enchantment resistant to rationality, such as the endlessly renewed rationality of sense (value rationality); it refuses to dissolve in methodical rationality of the control of experience. Historical studies, the development of which in Europe from the eighteenth century has been the most effective factor in the modernization of historical culture, must employ sense criteria that methodical rationality alone cannot generate.

The rationalization of modern historical culture is constituted by a universalization of the norms and values that make the past significant for the present.

The ethnocentric logic of identity formation always has a universalizing tendency, insofar as it describes its own world as both unique and normatively obligatory. 'Humanity' is a common frame of reference for historical identity. In the focus of historical experience, humanity is extended to the limits of the known world. This creates a fundamental, structural conflict of cultures in historical thought, which has repeatedly led to violence, to the legitimation of domination, and even to genocide.

This leads us to a fundamental problem in contemporary historical thought. It is shaped by the rationalism that culminates in historical scholarship and related human sciences (Bentley 1997). This rationalization of history as scholarship corresponds to the modern rationalization of all knowledge, without which the organization of modern societies would be inconceivable. Is historical thought inseparable from the value system of Western culture in its current form along with this methodological rationality? If so, how can it conceive of history in such a way as to account for the diversity in different cultures' articulations of 'history' as the orientative framework of human praxis? Does this diversity, once it has been researched and analyzed by scholarship, not always appear in the light of a particular interpretation—namely the interpretation to which methodological reason essentially belongs?

Current tendencies in historical thought (postmodernism, postcoloniality, the 'cultural turn') are determined to overcome modernity in this sense. They want to create a pluralism of 'histories' that can break the spell of uniformizing rationality and generalizing normativity. They thereby destroy the orientative power of modern historical thought with its notion of a unified direction of development, or 'progress.' At the same time they problematize the claims of the scholarship that describes the laws of such development along with the desire for mastering them, e.g., in Marxism.

In the age of globalization, in spite of the critique of the claim for hegemony inherent in Western historical thought, we cannot do without either methodological rationality or the universal regulative of communication. We need this regulative to understand the present by the interpretation of the past to define inclusion and exclusion, and to act in anticipation of the future. Yet, if the ethnocentric logic of such a universalization is to be broken or overcome, the truly universal regulative of historical thought and of the historical culture can only be a mutual and thereby universal recognition of difference and particularity. This idea is and will remain counterfactual, even utopian; nevertheless, it could disclose new dimensions of historical experience and lead to new and politically significant orientations.

Certainly the terrors and catastrophes of the twentieth century, such as the Holocaust, provide a new challenge to historical thought. Only with difficulty can they be integrated into older conceptions of historical meaning. They require new strategies of interpretation and representation. The question of human 'nature' must be radically reconsidered, and senselessness itself should be taken as a constituent of a meaningful conception of history (Friedländer 1998).

Bibliography

Appleby J, Hunt L, Jacob M 1994 *Telling the Truth about History*. Norton, New York

Assmann J 1992 *Das kulturelle Gedächtnis: Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. Beck, Munich

Bann 1984 *The Clothing of Clio*. Cambridge University Press, Cambridge, UK

Bentley M (ed.) 1997 *Companion to Historiography*. Routledge, London

Danto A C 1965 *Analytical Philosophy of History*. Cambridge University Press, Cambridge, UK

Eliade M 1991 *Myth of the Eternal Return, or Cosmos and History*. Princeton University Press, Princeton, NJ

Evans R J 1999 *In Defense of History*. Norton, New York

Friedländer S 1998 Writing the history of the Shoa: some major dilemmas. In: Blanke H-W, Jaeger F, Sandkühler T (eds.) *Dimensionen der Historik: Geschichtstheorie, Wissenschaftsgeschichte und Geschichtskultur heute*. Böhlau, Cologne, Germany, pp. 407–14

Hutton P H 1993 *History as an Art of Memory*. University Press of New England, Hanover, NH

Koselleck R 1985 *Future's Past: On the Semantics of Historical Time*. MIT Press, Cambridge, MA

Langer L L 1991 *Holocaust Testimonies. The Ruins of Memory*. Yale University Press, New Haven, CT

Lowenthal D 1985 *The Past is a Foreign Country*. Cambridge University Press, Cambridge, UK

Megill A 1998 History, memory, identity. *History of the Human Sciences* **11**(3): 37–62

Müller K E, Rüsen J (eds.) 1997 *Historische Sinnbildung—Problemstellungen, Zeitkonzepte, Wahrnehmungshorizonte, Darstellungsstrategien*. Rowohlt, Reinbek bei Hamburg, Germany

Niethammer L 1992 *Posthistoire: Has History Come to an End?* Verso, London

Nietzsche F 1985 *The Use and Abuse of History*. [Trans. Collins A]. MacMillan, New York

Nora P 1989 Between memory and history: les lieux de mémoire. *Representations*. **26**: 7–25

Ong W J 1982 *Orality and Literacy*. Methuen, London

Ricoeur P 1984–87 *Time and Narrative*, 3 vols. University of Chicago Press, Chicago

Ricoeur P 1998 *Das Rätsel der Vergangenheit. Erinnern—Vergessen—Verzeihen*. Essener Kulturwissenschaftliche Vorträge, Vol. 2. Wallstein, Göttingen, Germany

Rigney A 1990 *The Rhetoric of Historical Representation: Three Narrative Histories of the French Revolution*. Cambridge University Press, New York

Rüsen J 1993 *Studies in Metahistory*. Human Sciences Research Council, Pretoria, South Africa

Rüsen J 1996 Some theoretical approaches to intercultural comparative historiography. *History and Theory*. **35**: 5–22

Vansina J 1985 *Oral Tradition as History*. University of Wisconsin Press, Madison, WI

Weber M 1988 Science as a variation. In: Lassman P, Velody I, Martins H (eds.) '*Max Weber's Science as a Vocation*'. Unwin Hyman, London
White H 1973 *Metahistory. The Historical Imagination in Nineteenth Century Europe*. Johns Hopkins University Press, Baltimore

J. Rüsen

History, Professionalization of

It is evident that the term professional in this connection has two different meanings, which are related to each other but are not equivalent. The first (a) is professional in the sense of being able to live by the trade (in the meaning of occupation, such as the French word 'profession'). The second (b) is professional as a qualified person who is recognized by society (not least by his or her fellows in academia) as competent to practice the trade with a closure directed against nonprofessionals, i.e., amateurs and those writing on history without being 'real' researchers.

1. *The Historian's Profession as Bread-winning*

In the first sense of professionalism (see *Professionalization/Professions in History*), there have been professional historians for a long time but there is no distinct collective process of professionalization. The writing of history is age-old. Some famous history writers were active in the European antiquity, and especially Herodotus and Thucydides of Greece have won lasting fame. These and other early European historians were not writing history for their living. In China from at least the second century BC, and later in Korea and Japan, historians were engaged to describe the history of their country under a dynasty. This usage involved a real employment in China and it even entailed a personal responsibility of the historian for what he wrote. Sima Qian (born 145 or 135 BC), known as the Grand Historian, was employed to make justice to the Han dynasty but he also maintained a theoretical discourse on history (Wang 1995). He and other Chinese dynastic historians were obviously fully occupied and cannot even normally have worked alone but must have had a small office with clerks at their disposal. The Chinese professionalism among its historians had no known counterparts. In Korea the first well-known historians (eighth century AD) were politicians who wrote on their own behalf. In India there were no real historians until the thirteenth century because of the dominant cyclical time perception in the Hindu culture. The late Roman and medieval European historiographers had little profit from their writing and could not live from it. 'A book belonged to its author for as long as it stayed in his hands. After that it was free as air' (Smalley 1974). Some late medieval chroniclers in Europe however managed to live by their work as historians. Such chroniclers were employed by princes or cities in order to relate the important events that took place within the principality or city.

In Europe, the chroniclers employed by princes and cities (those active in monasteries were not employed for the purpose and formally not living by their historical writings) were followed later by official historiographers hired by kings and states for the 'proper' recording of the wars where their masters were engaged. Titles like Historiographer of the Realm (in variations all over Europe) make clear the official mission of the historian. The French Historiographe du Roi, was a title used from 1554 to 1824, and 104 persons bore the title. It was not so much a function as a title with a reward, and the historiographers were not highly regarded at court (Boer 1985). For most such official historiographers, the revenue from the title was only a side income and not enough to live on. Quite another type of professional historians in this sense arose with the great wars in Europe in the seventeenth century: those who lived from their reports from the war scenes. They established a form of market situation for themselves, and especially those behind the *Theatrum Europaeum* were successful in their branch, a sort of intermediary between history and journalism.

Already from the late eighteenth century another type of professional historians, in sense (a) of the word, appeared in Europe: the literary men writing history and more or less living by their literary accomplishments. Men like Voltaire, Gibbon, Robertson, and, to an extent, Schlözer and Herder made themselves known as historians (though primarily as learned and literary men) and their works were sold in editions of considerable numbers. Their ambition was to a large extent to give to the reader an impression of historical personalities and their struggle with moral problems, a sort of applied political science or applied moral philosophy. These historians did not strive for new factual knowledge at first hand but rather a better insight into the moral problems of politics pursued by states and kings. Such literary professionals who worked with historical subjects were more successful than the contemporary professors of the universities of Europe, who also earned their living from history and who were mostly not well known.

Still in the (a) sense of the word professional the authors of the 'broad' literature on history which grew up from the early nineteenth century should be mentioned. Many narrators of history wanted to give to their countrymen a better understanding of the past of their nation. Publishing narratives on national history for the broad layers of the population sometimes also gave a living to the authors already in the first half of the nineteenth century. It is easy to see the continuation of such popular narrative into the twentieth century. The number of persons who have had

the possibility to earn their living by telling history in different media and not only in printed books has increased enormously. The demand for history as entertainment, originating in the eighteenth century among the upper classes, has grown and spread in society. In the twentieth century, publishers have been interested in different types of books, such as specialized historical works, textbooks for schools and universities, historical journals, historical fiction and pure historical entertainment. Adaptations of history in various forms for radio and television have also gradually increased their share in the total turnout, and art film has been used for telling stories with a historical setting. The comprehensive number of persons who make their living from these forms of writing and processing history has never been calculated.

In the twentieth century, though starting earlier, there has also been a great growth of the number of professional historians in sense (a) of the word, who have been employed in teaching. Here must be included all teachers specializing in history, both in secondary schools and at the university level. Historians working in archives and museums also belong to this category. With the great expansion of secondary and tertiary education, the sector has been steadily growing in the twentieth century, especially after World War II. For France in 1910, Boer has estimated the number of historians in these categories to 1045 (Boer 1985). On the eve of the 1950s the number of professional historians in France has been approximated as less than 3,000 but then this number does not include archivists and historians employed in museums. In 1967, the number of the same limited category of professional historians had grown to close to 8,000 (Charle 1995). German and Australian calculations go in the same direction. (Raphael 2000, Macintyre 2000).

2. *Establishing Professionalism Among Historians as a System of Norms*

The second sense in which the term professional is used, 'professional as a qualified person who is recognized by society (not least by his/her fellows in academia) as competent to practice the trade,' has a more complicated historical reference. Historians who were recognized as competent historians by their contemporaries have existed for long. But in regard to the recognition by fellow historians, especially those with university posts, this is something much newer. If recognition is taken to mean an inclusion into a network, group or a formal association, only the history of the period from 1750 provided any means for the realization of such professionalism. If, further, a sort of common program of the network or group is required, the field is narrowed even more.

From the middle of the eighteenth century we can discern the beginnings of a formation of schools among academics in general and also among historians in Europe. It is difficult to say if anything comparable has been at hand in other parts of the world earlier. This was the period when German universities were flourishing and universities in other parts of Europe and its colonies were not. Because of the low level of state power in Germany not only free authors could express their views but also civil servants, such as professors. This made German universities, especially those in Halle and Göttingen, serve as a nursery for new notions in the humanities. Linguistics and history prospered. In the nineteenth century, 130 Protestant historians (plus an unknown number of Catholics) became full professors of history at German universities and for most of them this chair meant a culmination of their academic career. For the first time an occupational group of historians developed. Their competence for the professorship was mostly carefully examined and a discourse on history emerged (Blanke 1991). Mostly, however, these were individual accomplishments and the level of coordination between the learned men was low. No common rules for historical works were acknowledged, though an effort to write a manual of historiography was made by G.A. Will. Many other historians, e.g., Johann Martin Chladenius and Johann Christoph Gatterer, lectured on rules and wrote on the theory of history (Reill 1975). The process of standardization of methodological rules was, however, regarded with skepticism by some of the most important historians, such as August Ludwig Schlözer, who talked about imprisonment in scholarly rules, as Blanke has noted (Blanke 1991). In the early years of the nineteenth century at least four introductions to historical studies were published. As to rules for the critical use of sources and general methodological emphasis the introductory book by Friedrich Rühs from 1811 is noteworthy.

There were hardly any 'schools' of historians before well into the nineteenth century. Academic teaching of history and research in history were well established in many countries in the eighteenth century, though without a canon of methodology or a professed normative system, which applied to history as a discipline. History was rather regarded as a part of a wider discipline of morality, the French *sciences morales et politiques*. Past events could be used to show the consequences of different types of conduct and new enlightenment notions made it possible to analyze historical processes without referring to God or providence (Heilbron 1990). For centuries history had been used for this purpose, but the morality concerned changed its content and its stability in the eighteenth century. The new directions were not diffused to students in standard formulations or transformed into methods and manuals. The cohesion among enlightenment historians was low and this continued to be the case up to around 1830.

The change brought about primarily by Leopold von Ranke was to create the conditions for a network

of historians under the guidance of the master. He had meetings with his disciples, later known as his seminar, where problems of history were discussed. Ranke himself was a spokesman for three main principles of historical interpretation: the state as a central entity, the state system as the polity of international relations, and the necessity for the historian to evaluate each single entity for its role in history and not from its relation to contemporary politics. The first of these principles was a common ground for many thinkers of the period, but combined with the other two it formed a specific analytic system. The third principle has been partly traced back to forerunners in the eighteenth century (Meinecke 1936, Blanke 1991, Blanke and Rüsen 1984) and it created the basis of the thought system, which became known as *Historismus*.

In the German academic community of historians the combination of *Historismus* and teaching through seminars became an instrument for the professionalization of historians. A common evaluative system was created in the latter half of the nineteenth century. A methodological part was often combined with individualization of persons and collective entities of the past, especially among the historians of the Middle Ages who were concerned with the editing of the *Monumenta Germaniae historica*. This work elaborated further methodological principles for editing which had been developed in France by Jean Mabillon and others from the seventeenth century and onwards, and which gave rise to the *École des Chartes* at Paris. Methodology became important and the principles for the evaluation of sources became more and more systematized. A methodological direction of historians within the *Historismus* has therefore been discerned, with a philosophical direction as its counterpart. The former was characterized by its devotion to empirical studies with a sincere application of methodological principles in order to obtain objective knowledge, while the philosophical direction generally rejected the possibility of reaching any form of certainty and proclaimed relativity of knowledge. In fact it will seem that the differences within *Historismus* were great. On the one hand were those who saw historical relativism as fundamental and regarded history primarily as a means of orientation in the present. On the other hand were those who wanted to overcome as much as possible the constraints of the sources for understanding the past, though still with a considerable relativism in regard to values (Oexle 1996, Scholtz 1997).

In several respects a new turn of historical professionalism grew out of the empirical *Historismus*. It was important that new methodological manuals were published and widely accepted in the European community of historians. J. G. Droysen had published his *Grundriss der Historik* (1868), but its emphasis was rather on the principles of *Historismus* than on the strictly methodological side. Ernst Bernheim's *Lehrbuch der Historischen Methode* (1889) and Charles-Victor Langlois' and Charles Seignobos' *Introduction des Études Historiques* (1898) were in this regard epoch-making. These books were in the first hand methodological (from the 1903 edition Bernheim added a section on philosophy of history), and they received a very wide circulation. They taught how to find secure knowledge in history. The authors sought somewhat different philosophical bases for their standpoints but the authors of both books turned their back to the relativism of knowledge. Their differences mattered less than their methodological aims, which found appeal among historians.

The first decades of the twentieth century witnessed a change in the general approach to historical professionalism. The emphasis had been on the theory of the state or the theory of state systems or the theory of the (individual) nation. Now the methodological element became central. Without the proper knowledge of how to evaluate sources a historian was not accepted as a professional. Several factors helped to make methodology a central element in the professionalism of historians. One was the rise of national historical associations. This was a process that took place all over Europe (and also beyond Europe in some countries) in the latter half of the nineteenth century. Philosophical and political issues divided all such associations and, thus, only methodology was really a common basis. This was even more the case with the international organization of historians, later known as the *Comité International des Sciences Historiques*, which can be traced back to 1898. In the early programs of its international congresses from 1900 and 1903 methodology played an important role and the object of the discussion was rather which were the conclusions for the philosophical foundations that must be drawn from the established methods than the other way round. After the congress of 1908, the French historian Simiand complained in a review article that methods had not been sufficiently emphasized in the congress contributions (Erdmann 1987).

In the last decades of the nineteenth century and the early part of the twentieth century the historical associations of the European countries also in most cases started journals (Stieg 1986). The influential German Historische Zeitschrift was, however, founded in 1859 as a private initiative by Heinrich von Sybel. The earliest still existing seems to be the Danish from 1838. During the twentieth century the creation of journals of history has become usual in all parts of the world, and some of these are much more than national journals but quite general (such as the *American Historical Review*, founded in 1895) or thematic for studies of the history a major part of the world (such as the *Journal of Asian Studies*, founded in 1957). Other journals and reviews with a thematic or rather broad methodological focus, such as the *Revue de Synthèse Historique* (1900–31); the *Annales d'Histoire Économique et Sociale* (founded in 1929, from 1946 called *Annales*: *Économies*, *Sociétés*, *Civilisa-*

tions); the *Past and Present* (founded in 1952); the *Journal of Contemporary History* (founded in 1966); the *Journal of Modern History* (founded in 1929), etc., were mostly later creations. In most of the journals, articles on themes of history were combined with critical reviews of books. An important element in such reviews was to examine the methods used and the shortcomings of authors in respect to methodological rules. Thus, in the form of critical reviews of the works on history, the methodological norms were spread. These norms were also commonly accepted. In the case that a historian wanted to contradict criticism he or she had to show that the principles had been observed rather than that these principles were not valid. This is something that characterizes methodology but is not valid for the kind of substantive norms that formed the major part of *Historismus*.

For the formation of the profession of historians, the debates on the application of historical methods were very important. A series of such debates took place in different countries in the first decades of the twentieth century. There was a big debate in Germany, called the *Methodenstreit*, on the correct use of methods in regard to any scientific activity. Generalization and interpretation were advanced as alternatives. Karl Lamprecht, with his concept of social and cultural history, was rather lonely in Germany and is often regarded as defeated. In fact, part of the issue was the theories of *Historismus* and the alternative professed by Lamprecht. Another part dealt with methods and their application in the strict sense, where individualization (W. Windelband, H. Rickert, M. Weber, F. Meinecke, and others) was opposed to views based on structures and collectivities (W. Sombart, H. Pirenne, F. Simiand and others) (Iggers 1984). The struggle was reflected in other countries such as Finland. In other countries, e.g., Sweden, Norway, and China, there were other straightforward struggles about the use of methodological principles and the consequences for national history. The historians who advanced new and radical consequences of methodologies, which were already well known, had a program. They wanted to establish firm knowledge in history, a knowledge that was contrasted to guesses and conjectures, not to historical relativism. In the USA, the focus of the debate was the possibility of establishing objectivity (Novick 1988).

Historical professionalism also increased through the ongoing specialization of all historical fields. In the beginning of the twentieth century, it was no longer the normal thing for a historian to take as the theme of an investigation the history of a country during a certain period. This had been quite usual in the middle of the nineteenth century, but in the first decades of the twentieth century this had changed. History had begun its separation into specific subfields, and economic, social and cultural history were regarded as major such parts. Specialized research institutes came into existence and specialized journals were published: the *Economic History Review* from 1927, the *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte* from 1903, the *Journal of Contemporary History* from 1966, and the *Vierteljahrshefte für Zeitgeschichte* from 1953. Similar journals were published in many countries with specific reference to the speciality in the national context. Specialization went further, chronologically and thematically. Investigations became very specialized in regard to field and object of study, and period of investigation. There was a dramatic change in the themes regarded fit for doctoral dissertations from the middle of the nineteenth century to 1920. Dissertations then were also outcomes of extensive archival research where the author could and should show his or her (some women were achieving the doctoral degree at the time) acquaintance with sources and with the appropriate methods. Academic professionalism aiming at firm knowledge had found its way of expressing itself in such specialized studies.

At the same time as the second phase of historical professionalism in the sense of recognition among equals was triumphing in the interwar years, a distrust in the validity of its central notion, the idea of secure knowledge, was winning ground among historians. The philosophical direction of the *Historismus* had long since brought about doubts as to the possibility of reaching to security in historical research. Its proponents had allied to the old tradition of hermeneutics, grown out from Bible studies, for acquiring a system of understanding. It contained two main requirements: pre-knowledge as the conceptual apparatus and direction of interest; and empathy as the means of acquiring an understanding of human thoughts and actions. Hermeneutics was regarded as a means to acquire more or less probable statements on the past and the most general formulations included an ambition to maximize the probability through research. Collingwood (1946) worked in this direction, and after World War II, Hans-Georg Gadamer brought out his influential interpretation of hermenutics, *Truth and Method* (1960), which was later overshadowed in some circles by Paul Ricoeur's *Time and Narrative* (1983–85).

Common to these works was that many historians accepted them as important foundations for the work of the historian. This meant that a new professional identity arose among some historians, conflicting with the earlier established group identity based on firm knowledge. As long as a basic norm in the hermeneutical camp was to increase probability of historical statements through empirical research, the conflict could be kept within close bounds. After the 1980s, this was no longer possible. Various theoretical directions have taken the relativistic claim further. The debate in the 1980s and 1990s has gone out from Hayden White and his rhetorical analysis of historiography (White 1973, 1978). The analyses of his standpoint have been numerous, above all in the USA, and have led to a variety of relativistic positions as regards the theory of

history based on the assumption that history is basically a rhetorical and aesthetic discipline. At the same time, in Europe, postmodernism has influenced other philosophers of history. A far-going relativism in different forms thus has more or less superseded the earlier basis of professional identity among historians.

Social theories in the broad sense of theories of the social sciences have since long occasionally been used by historians. More regularly has this been the case from the 1960s and onwards. In economic history it was already earlier the rule to rely on and use theories of economics, but theories of management and organizations and sociological theories have been used to a large extent by historians in historical studies only in the 1970s and 1980s. The professionalism of economic historians has come to rely on the ability to apply such theoretical instruments as well as specific economic (statistical) methodology. The analysis of social structures and social dynamics has also developed through the use of social theory. This approach had one of its origins in the Annales circle, then most often directed towards demography and geography and their social consequences. More explicitly, social science theory in general, including sociological and organizational theory, was proposed as the proper instrument for historical research by the so-called Bielefeld school, consisting of a group of historians who had some relation to the university of Bielefeld in the 1970s.

What they proposed was in line with what had been done by several American historians earlier, such as Merle Curti and Stephan Thernstrom, but their efforts had not been coordinated into a program. The Bielefeld historians advanced the concept of a 'historical social science' (Kocka 1977, Wehler 1980). This type of history found much appeal in the 1970s and 1980s. Although parts of this enthusiasm for social sciences were criticized later, it is rather the theories used than the close link to social science (including social anthropology, which found growing appeal among historians) that have met opposition. The use of social science theories has thus been one source of professional identity among historians in the recent decades.

The conflict between ideals inspired by traditional approaches, relativism and social science has not been resolved. In fact historians have no longer (if they ever had) one source of professional identity. This is not only a consequence of the fights between different convictions within the academic community. All through the twentieth century historical professionalism has also been challenged from political quarters. Autocratic regimes all over the world have tried to make historians into docile instruments for the propaganda of the ruling ideology. Many have suffered for their professional ideals, others have accepted the norms of the state instead of the norms of the profession.

By the end of the twentieth century, when historians tried to base their professionalism on interpretative theories, methodological rules and social science theories, all without general acceptance, the history of historiography seems to have met a new interest. Different standpoints to historical knowledge may unite in the interest in investigating the history of historiography, and it is equally important for all different standpoints. The role of this disciplinary self-understanding for a renewal of professionalization should not be overrated, however. The history of the discipline cannot meet the requirement of providing the norms through which a closure can be upheld against nonprofessionals and, as regards norms there is no unanimity. In many countries it seems that the professional closure has become softened in recent decades and no sharp line is often drawn between historians and persons of literature. If history is at a point of crisis as a discipline, as is frequently said, it might be a consequence of two facts. No single normative theory is prevailing and the profession seems to rely on the quasi-theoretical field of its own history to reproduce a common professional identity.

See also: Historicism; History: Forms of Presentation, Discourses, and Functions; History: Overview; History: Theories and Methods; Professionalization/Professions in History; Professions, Sociology of

Bibliography

Bernheim E 1889 *Lehrbuch der Historischen Methode*. Dunker and Humblot, Munich, Germany

Blanke H W 1991 *Historiographiegeschichte als Historik*. Fromann-Holzboog, Stuttgart-Bad Cannstadt, Germany

Blanke H W, Rüsen J (eds.) 1984 *Von der Aufklärung zum Historismus. Zum Strukturwandel Historischen Denkens*. f. Schöningh, Paderborn, Germany

Boer P den 1985 *Geschiedenis als Beroep: De Professionalisering van de Geschiedbeoefening in Frankrijk (1818–1914)* [*History as a Profession. The Study of History in France*]. SUN, Nijmegen, The Netherlands

Charle C 1995 Être historien en France: une nouvelle profession? In: Bédarida F (ed.) *L'Histoire et le Métier d'Historien en France 1945–1995*. La Maison des Sciences de l'Homme, Paris, pp. 21–44

Collingwood R G 1946 *The Idea of History*. Clarendon Press, Oxford, UK

Droysen J G 1868 *Grundriss der Historik*. Leipzig, Germany

Erdmann K D 1987 *Die Ökumene der Historiker. Geschichte der Internationalen Historikerkongresse und des Comité International des Sciences Historiques*. Vandenhoeck and Ruprecht, Göttingen, Germany

Gadamer H-G 1960 *Wahrheit und Methode: Grundzüge einer Philosophischen Hermeneutik* [*Truth and Method*]. Mohr, Tübingen, Germany

Heilbron J 1990 *Ontstaan van de Sociologie* [*The Rise of Social Theory*]. Prometheus, Amsterdam

Iggers G 1984 The 'Methodenstreit' in international perspective. The reorientation of historical studies at the turn from the nineteenth to the twentieth century. *Storia della Storiografia* [*History of Historiography*], **6**: 21–32

Kocka J 1977 *Sozialgeschichte: Begriff, Entwicklung, Probleme*. Vandenhoeck und Ruprecht, Göttingen, Germany
Langlois Ch-V, Seignobos Ch 1898 *Introduction aux Études Historiques*. Hachette, Paris
Macintyre S 2000 History in a new country. Australians debate their past. In: Torstendahl R (ed.) *An Assessment of Twentieth Century Historiography*. Royal Academy of Letters, History and Antiquities, Stockholm, Sweden, pp. 71–86
Meinecke F 1936 *Die Entstehung des Historismus*. R. Oldenburg, Munich, Germany
Novick P 1988 *That Noble Dream: The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK
Oexle O G 1996 *Geschichtswissenschaft im Zeichen des Historismus*. Vandenhoeck und Ruprecht, Göttingen, Germany
Raphael L 2000 Flexisble response? Strategies of academic historians towards larger markets for national historiographies an growing scientific standards. In: Torstendahl R (ed.) *An Assessment of Twentieth Century Historiography*. Royal Academy of Letters, History and Antiquities, Stockholm, Sweden, pp. 127–47
Reill P H 1975 *The German Enlightenment and the Rise of Historicism*. University of California Press, Berkeley, LA
Ricoeur P 1983–85 *Temps et Récit* [*Time and Narrative*]. Le Seuil, Paris, Vols. 1–3
Scholtz G (ed.) 1997 *Historismus am Ende des* 20. *Jahrhunderts. Eine Internationale Diskussion*. Akademie, Berlin
Smalley B 1974 *Historians in the Middle Ages*. Thames and Hudson, London
Stieg M F 1986 *The Origin and Development of Scholarly Historical Periodicals*. University of Alabama Press, Tuscaloosa, AL
Wang Q E 1995 Time perception in Ancient Chinese historiography. *Storia della Storiografia* **28**: 69–86
Wehler H-U 1980 *Historische Sozialwissenschaft und Geschichtsschreibung: Studien zu Aufgaben und Traditionen deutscher Geschichtswissenschaft*. Vandenhoeck and Ruprecht, Göttingen, Germany
White H 1973 *Metahistory. The Historical Imagination in Nineteenth-century Europe*. Johns Hopkins University Press, Baltimore, MD
White H 1978 *Tropics of Discourse. Essays in Cultural Criticism*. Johns Hopkins University Press, Baltimore, MD

R. Torstendahl

History: Theories and Methods

Since history turned into an academic discipline in the nineteenth century and professional historians claimed a privileged epistemological status for their approach vis-à-vis non-professional approaches to history, debates about the role of methods and theories in history have been characterized by a certain conjuncture. In periods of relative uncertainty and crisis about the status of history as a distinctive, academic discipline—such as the periods between 1880 and 1920 and 1965 to 1985 in the West—a substantial body of historians participated in these debates. In those periods, *Methodenstreit* is more or less intense and visible, while in periods of relative certainty and academic recognition, this type of debate usually recedes into the background. So, seen from a historical point of view, reflection on the role of methods and theories in history is in a sense a symptom of crisis; at least outside the small group of historians and philosophers who permanently reflect on history in their capacity as philosophers of history. This direct connection between a 'reflective mood' among professional historians and their sense of professional uncertainty makes sense because, as will be elucidated below, method and/or theory have usually been presented as the discipline's cognitive foundations, especially its claim to produce 'objective' knowledge of the past. Therefore, doubts and debates about method and theory in history are always doubts and debates about the discipline's cognitive status, its academic credentials, and its claim to objectivity.

Seen from the point of view of the sociology of science, debates about the role of methods and theories in history have usually been directly linked to battles waged by the proponents of different conceptions of history for academic legitimacy and supremacy, that is, for academic recognition and reputation. In periods of serious competition for 'intellectual capital' (Bourdieu) between several competing groups, operating within the same 'academic field,' debates about methods and theories in history have been intense. This not only holds true for history, but also for other social sciences such as sociology, economics, psychology, etc., and, therefore, are a characteristic of pluralistic or poly-paradigmatic disciplines.

Characteristic for the periods of increased interest in theory and method is also an increased interest in the history of the discipline in case from both an intellectual and from an institutional angle. This interest in disciplinary history is usually motivated by a desire to locate the origins of the preferred conception of historical theory and method in the discipline and to generate academic credentials by locating prestigious founding fathers or precursors of the preferred conception in the discipline's pedigree. Therefore, in history, an increased interest in the history of history writing alias historiography can usually be observed alongside an increased interest in theory and method: theoretical and methodical reflections usually are accompanied by reflections on the origins, nature, and variety of historiography.

Seen from a philosophical point of view, debates about the role of theories and methods in history often have a blurred character due to the vague philosophical status of the central concepts of theory and method. Therefore, some elementary conceptual clarifications are called for in order to answer the question why theory and method in history have turned into a heavily contested area at times since history turned into an academic discipline.

1. Theory and Method in History: Some Preliminary Distinctions

In discussing the role of theory and method in history, it is necessary to introduce some basic distinctions. Due to the double meaning of history as *res gestae* and as *historia res gestae* (see *History: Overview*), theory in history can refer both to history as a process and to the knowledge of that process. Therefore, the first basic distinction to make is the distinction between:

(a) theories, which deal with characteristics of history as a process (such as, for instance, Marx's theory of history as class struggle); and

(b) theories, which deal with the characteristics of knowledge of history (such as, for instance, that scientific knowledge of history has the form of law-like knowledge or the form of concrete particulars).

Theories of type (a) can be called substantial or material theories of history, while theories of type (b) can be called cognitive theories of history, i.e., theories of historical knowledge.

Cognitive theories of history can further be subdivided into:

(a) epistemological theories and

(b) methodological theories, while substantial theories of history can also be subdivided in finer subsets (see *History: Forms of Presentation, Discourses, and Functions*).

Substantial and cognitive theories usually have been interrelated, because presuppositions about what can be known about the past are linked to presuppositions of how the past can be known. Substantial theories, for instance, which hold that the historical process is characterized by underlying mechanisms and regularities, analogous to theories of classical mechanics, have been interrelated with cognitive theories that posit that real knowledge of history implies knowledge of general laws (see *Positivism, History of*). And the substantial theory that the historical process is characterized by a fundamental chaos or contingency and not by hidden regularities alias a 'motor,' has been interrelated with cognitive theories which posit that real knowledge of history is knowledge of particular facts instead of general laws (see *Historicism*).

Substantial theories, in this way, not only have consequences for methodological positions, but also for epistemological positions. Substantial theories, which posit some kind of affinity or even identity of the knowing subject and the object of history in contrast with nature (as was the case with Hegelian idealism and its offshoots in historicism, such as Droysen and Dilthey's versions of historicism), have usually been interrelated with methodological theories. These posit a distinct kind of method for the human sciences—such as understanding, or *Verstehen*, or interpreting—in contrast with the methods of the natural sciences (usually defined as some form of causal or law-like explanation). And these distinct methodological theories have, in their turn, usually been interrelated with distinct epistemological theories, linking methods of understanding to hermeneutic or phenomenological theories of knowledge and methods of explanation to empirical theories of knowledge. Before taking a better look at this terrain, however, it is first necessary to clarify the central concepts of method and theory as well as their interrelationships in history. When analyzed more closely, both method and theory in history appear to have a broad variety of meanings.

2. Historical Method(s): Concept and History

Analyzing the meaning of 'method' in history, it is useful to start with a clarification of the term 'method' in other disciplines first (Meran 1985). In the mathematical sciences it is common to refer to axiomatic-deductive, to infinitesimal, and to inductive methods, meaning a set of formal, technical procedures in order to solve problems within a specified range. Outside the mathematical sciences, the meaning of method is usually far less defined. In philosophy, for instance, there is a long tradition referring to methods of all sorts, ranging from Descartes' discourse on method to Kant's transcendental method, from Marx's dialectical method to Husserl's phenomenological method, and to De Saussure's structuralist method, to name just a handful. These 'methods' in philosophy, however, do not constitute defined formal rules for specified ranges of problem solving, but constitute complex sets of ideas, concepts, and rules of thumb on how to analyze specified domains, such as:

(a) scientific knowledge;
(b) the human consciousness;
(c) the history of mankind;
(d) the meaning of concepts;
(e) the meaning of language in use; or
(f) language as a system of signs.

'Method' in philosophy, thus, is definitely not identical with a set of formal techniques, but a loosely circumscribed research program. This diffuse and nonformal meaning of method is also characteristic of the human sciences.

In history, the term historical method was first introduced in a systematic way in the sixteenth century by Jean Bodin in his treatise of source criticism, *Methodus ad facilem historiarium cognitionem* (1566). Characteristically, Bodin's treatise intended to establish the ways by which reliable knowledge of the past could be established by checking sources against one another and by so assessing the reliability of the information conveyed by them, relating them to the interests involved. Historical method was thus at the same time intended to defeat the skeptical doubts with regard to the possibility of reliable knowledge of the past. Skepticism (alias historical Phyrrhonism) had been fed by a large quantity of scriptural forgeries parading as original historical documents, primarily produced by competing churches and nobles to back

up their claims to rights and property. Therefore, the traditional claims to impartiality, truth, and trustworthiness of historians, dating back to classical Antiquity, were no longer believed, and were now replaced by a claim to impersonal and intersubjective method.

This direct relationship between history's claim to truth and method would characterize all later treatises on historical method, although what was meant by method did change over time. In the eighteenth century, historians of the Enlightenment, such as Schlözer, included not only the source critical techniques as the historical method (*Forschung*), but also the way the results of historical research were synthesized into a textual whole (*Darstellung*) and presented to an audience. This broader conception of method had its roots in the classical rhetorical tradition and would only recently be rediscovered in historical theory, that is from the 1970s onwards. The German historian and historical theorist Johann-Gustav Droysen (Droysen 1977) has become one of the pivotal historians behind this 'rediscovery' of the textual and rhetorical dimensions of history, because his treatment of historical method in the earlier and the later versions of his *Historik* embodies the transition from the broader to the narrower conception of historical method (Meier and Rüsen 1988).

The larger part of the nineteenth and twentieth centuries would be dominated by the research-oriented conception of historical method of the so-called Historical School in Germany, led by historians as Leopold Ranke and Berthold Niebuhr. Their conception of history, long been regarded as the beginning of modern, 'scientific' history, harked back to the 'narrow' conception of historical method, limiting the methodical character of history to source criticism. The historical composition and presentation—the *Darstellung*—was no longer regarded as methodical and was now seen as the literary, artistic, or aesthetic component of history. Therefore, the question of whether history belongs to the arts or to the sciences could and did arise from the nineteenth century onwards, a question that became part of the continuing *Methodenstreit* (Stern 1970, Fay et al. 1998).

In the nineteenth century, during the process of rapidly increasing disciplinary differentiation and academic institutionalization, it became customary in the human sciences to identify disciplines with one specific method, which was presented as the guarantee of its 'truth' and 'objectivity.' In this process, sociology was identified with the sociological method and history with the historical method. Canonical formulations of this conception of historical method can be found in E. Bernheim's *Lehrbuch der historischen Methode* (1889) or in Ch. Langlois and Ch. Seignobos' *Introduction aux études historiques* (1897).

Historical method in the nineteenth century usually referred to the techniques on how to conduct historical research of written materials, the so-called sources. These techniques were traditionally subdivided in three parts, corresponding to three subsequent steps in research:

(a) heuristics, that is the techniques how to locate the relevant sources;

(b) source criticism, the techniques by which the temporal and spatial origins of the sources are established as well as their authenticity; and

(c) interpretation, by which the 'cleansed' information derived from the sources is put together by means of interpretation in order to infer 'what really happened.'

Usually a distinction was made between primary and secondary sources. Primary sources consist of the written material produced more or less contemporaneously with the object of research, ideally recorded by eyewitnesses (model: the official records of parliamentary debates). The time span between primary sources and what they represent is thus as small as possible, because distance in time is regarded as a major source of unreliability of the information conveyed. Secondary sources consist of records written a considerable time after the fact and not by eyewitnesses (model: 'hearsay'). As the information in secondary sources is based on other sources, it is, at best, 'second hand'; therefore, it is regarded as less reliable than primary source information. According to the dominant, Western conception of history since the nineteenth century, serious academic history can only be based on primary sources, because those are the only reliable source of information on the past. Therefore, the archives (see *Archives and Historical Databases*) were usually considered the workshop of the professional historian. According to the same nineteenth century Western conception, non-scriptural or 'low' cultures were primitive by definition and simply had no history comparable with 'high cultures.' In this way the scriptural, source based conception of historical method defined, and limited, the object of academic history to the history of 'high cultures,' especially the history of the West. It would last until the second half of the twentieth century before these Western biases in the very conception of historical method would be seriously criticized, especially by postcolonial theory (see *Subaltern History*).

Another traditional distinction was that between internal and external source criticism. Internal source criticism is actually textual criticism, i.e., the hermeneutic procedures by which a text is reconstructed (if it is passed down in an incomplete form or in several variants) as a whole and the textual procedures by which authorship and textual authenticity are checked and its meaning established. External source criticism consists of those procedures by which a source is checked by information derived from other sources, written and material. The so-called auxiliary sciences, such as paleography, chronology, sillography, toponomy, numismatics, and diplomactics, usually play an important role in this phase when historians are

dealing with older periods. And when modern historians are dealing with nonscriptural, oral cultures or with cultures that left few or no scriptural traces, archeology and anthropology often function as auxiliary science. Reference to multiple methods within one discipline became the rule only in the course of the twentieth century, thereby reopening the problem of truth and objectivity and paving the way for a new wave of skepticism. For if the method can no longer be regarded as the procedural guarantee for the truth and objectivity of the discipline, what could, so an influential argument ran (Bernstein 1983, Megill 1994)? Recognition of plural methods within one discipline, therefore, has often been regarded as a threat of the disciplines academic credentials.

From the late 1960s onwards, the process of methodological differentiation within all of the human sciences took an unprecedented pace, leading to a temporary fragmentation of most disciplinary fields and a questioning of most traditional disciplinary boundaries and self definitions. Significantly, the new, 'poly-paradigmatic' self-image was at the same time regarded by many as a crisis of their discipline. This also held for history, in 1960s and 1970s regarded by many as one of the social sciences (Novick 1988, Appleby et al. 1994, Iggers 1997 see *History and the Social Sciences*).

Historians identified the quantitative method (see *Quantification in History*), the biographical method (see *Biography: Historical*), the comparative method (see *Comparative History*), the ideal typical method, understanding as a method (*Verstehen*), the causal method (*Erklärung*), discourse analysis as a method (see *Linguistic Turn and Discourse Analysis in History*), the micro-historical method and a psychoanalytical method in history (see *Psychohistory*). Other historians claimed that the concept of method had become severely overstretched, holding that most advertised methods in fact were theories about historical knowledge, which could only be checked by the historical method of source criticism. Therefore, a closer look at the relationship between historical method and theories is necessary.

3. Theories of Historical Knowledge

3.1 Empiricism and Anti-empiricism

The connections between historical method and historical theory can be located both on the level of epistemology and on the level of methodology. As epistemology is the theory of knowledge, historical epistemology deals with the theory of historical knowledge, i.e., the conditions of its possibility, range, sources, and foundations. Two epistemological positions have been defended in relation to historical knowledge from the nineteenth century onwards, empiricism and anti-empiricist idealism.

The problem of whether and how it is possible to acquire knowledge of other minds, i.e., the mental states of others, such as thoughts and intentions, separated from the knowing subject in both time and space—the typical problem of the historian—has been a fundamental one since mid-nineteenth century. According to empiricism, mental states could only be known as far as they correlated with empirically observable epiphenomena. The empirical dispositions or expressions of other minds are seen as the only access to other minds (Hempel 1966).

The proponents of phenomenological and hermeneutic views (hermeneutics is the theory of interpretation, originally of text interpretation), to the contrary, have always affirmed this possibility of a non-empirical access to mental phenomena. Phenomenology claimed introspective methods as the access to the mental universe; hermeneutics, far more influential in theory of history than phenomenology, claimed that 'reading' other minds could be conceived of as analogous to reading texts.

The fundamental cognitive activity in all hermeneutic theories is the interpretation of meaningful symbols and symbolically structured phenomena (like actions and cultural objects), whereas all empiricist theories regard the observation of objects as the fundamental cognitive activity. Just as it is possible to know the ideas expressed in a text by reading and interpreting, it is possible to know the ideas expressed in actions and in all human made objects by inferring the unobservable ideas from their observable expressions through re-enactment and interpretation by the historian (Droysen 1977, Dilthey 1981, Collingwood 1993). This possibility was guaranteed, according to the hermeneutic views, by the basic unity of humanity, implying the basic unity of subject and object of knowledge in the human sciences. Therefore, each alter was an alter ego, accessible through a process of intuitive identification and transfer in thought or in idea (therefore: idealism). So in historical epistemology, just as in the other human sciences, the basic divide between 1850 and 1970 has been the one between empiricist and anti-empiricist, idealist positions. It would last until the 1980s, with the rise of postmodernism and its offshoots in the discipline of history, before the unitary presuppositions of hermeneutic theories would be seriously criticized an questioned. Western (white) male was no longer accepted by all as the model of all humankind. This has led to a renewed emphasis on the basic diversity of knowing subjects and their epistemological and methodological relevance for disciplines like history (Scott 1991) (see *Gender History*; *Subaltern History* and *Linguistic Turn and Discourse Analysis in History*).

3.2 Unity vs. Duality of Method

Between 1850 and 1970, this basic divide in the domain of epistemology roughly corresponded with the basic

divide in the domain of historical methodology, i.e., the divide between the proponents of the theory of the unity of scientific method and the proponents of the theory of a fundamental methodological dualism of the human and of the natural sciences.

Methodology as such aims at the rational reconstruction of the normative rules, which are constitutive for a specific discipline and its development. Rational reconstruction, in contrast with historical reconstruction, presupposes that disciplinary practices and their development can be reconstructed in terms of logical and empirical arguments; this presupposition is, of course, an idealization of the rationality of real disciplinary practice, because apart from rationality, power and privilege also play their role in reality. Methodology abstracts from those non-rational factors in disciplinary life and aims to reconstruct the cognitive criteria and ideals, by which the practitioners of a discipline judge one another's products and evaluate their relative quality. Methodology, therefore, presupposes the existence of a basic cognitive consensus and rationality among the practitioners of a discipline. This presupposition, of course, has been criticized both for the natural as for the human sciences (Goodman 1978, Rorty 1986, Novick 1988).

Next to the basic ideal of truth, other cognitive ideals of a discipline are reconstructed by its methodology, such as the explanatory ideals of the discipline and its criteria of progress. At this level the basic methodological divide between the proponents of the unity of scientific method and the proponents of the dualism of scientific methods must be located, because the divide revolves around the question whether there is one explanatory ideal for all disciplines or not. This issue has dominated the methodological debates for more than a century since 1850 and is known under the labels of the explanation vs. understanding controversy (Gardiner 1959, Hausmann 1991). According to the unity of science view, also often labeled as positivism, there was just one model or logic of explanation and development for all empirical disciplines. Classical mechanics in the nineteenth century was usually considered the paradigm for all other disciplines, including human sciences like history. The only difference between the various disciplines was their phase of development. Underdeveloped disciplines (like history) showed little similarity to Newtonian mechanics while developed ones did. Immature disciplines only produced 'explanation sketches' (Hempel 1966), which had to be developed into real explanations. The only difference between disciplines was thus a matter of scientific maturity and not a matter of explanatory logic.

Real scientific explanation boiled down to subsuming an event to be explained under a general law. Therefore, the event to be explained can be logically deduced from the general law in combination with the given circumstances (the so-called initial conditions) and is at the same time, given the law and the initial conditions, predictable. According to this view on the logic of scientific explanation, that is also known under the headings of the subsumtion model, the deductive nomological model, or the covering law model of explanation, explanation and prediction in science are basically the same; explanation is really prediction in the past alias retrodiction. Scientific explanation, in this view, is showing why what happened, was—given the known laws and circumstances—expected to happen because it was bound to happen: explanation boiled down to show what happened was logically necessary. Given the fact that the general laws were usually (though not exclusively) interpreted as causal laws, formulating universal and invariant connections between causes and effects, scientific explanation was usually seen as causal explanation and explanatory necessity also as causal necessity: scientific explanation specifies the causes of what happened by formulating its causal determinations.

Now although this unitary theory of scientific explanation was amended and modified in time, in order to account for statistical and probabilistic forms of explanation in the natural sciences, its status was based on the assumption that 'positivist' theory at least provided an adequate reconstruction of natural science and its development from the seventeenth century onwards. This assumption has been increasingly undermined by postempiricist and postpositivist philosophies of science roughly since the 1960s and 1970s. Both the empiricist underpinnings of the covering law model and its logical claims were undermined by historical examples, arguing that the supposedly paradigmatic cases in the history of science did not fit the basic logical models of philosophy of science. Therefore, this 'turn' in the philosophy of science is often called the historical turn, because history was mobilized against explanatory logic (Lakatos and Musgrave 1970). However, before examining the implications of postempiricism and postpositivism in the philosophy of science for the opposition between classical unitary and dualistic methodological theories, an overview of the dualistic positions and arguments is called for.

3.3 The Autonomy of History

The philosophical defense of the autonomy of historical explanation vis-à-vis the explanatory models of natural science usually has started off with a critique of the unitary claim. From Droysen and Dilthey, through Windelband, Rickert, Weber, Collingwood, and Gadamer, to White and Ricoeur, the argument has been developed that historical explanation is not only empirically different, but logically different from the covering law model. This claim of autonomy has

traditionally been defended in three variants: an ontological version, an epistemological version, and a methodological version.

The ontological argument for methodological dualism was first systematically developed by Droysen and later by Collingwood and his school. It usually harks back to Vico and Hegel, because it is based on an idea of a subject-object identity in history: in history, humankind confronts themselves, and history, therefore, is the autobiography of humanity. History is fundamentally the process in which humankind expresses and develops themselves in the form of diverse cultures; and because humankind has made history, they are capable of knowing history in a direct way, according to this argument, just as the writer of an autobiography knows the history in case in a direct way. Historians try to reconstruct cultures and the particular ideas expressed in its cultural practices (such as religions and rituals) and its material artifacts (such as its temples and sacred objects). When historians offer explanations for the past, they construct intelligible and plausible connections between the culture and the ideas expressed in it on basis of their research of the sources in case. When they succeed in this task, they offer an understanding of the phenomena in case (*Verstehen*). Understanding history is conceived of as similar to understanding a text; the major difference is that the texts of history have more or less disappeared and therefore must be reconstructed first based on their traces by research. This type of intelligibility is possible, because in history, human beings confront themselves and therefore intuitive understanding is possible in contrast to the situation in the sciences of nature, where the knowing subject confronts an external object.

The intelligibility sought for by historians—historical understanding—is thus centered on connecting ideas to their expressions in particular spatio-temporal contexts, and proceeds by going back and forth between parts and whole in the so-called hermeneutic circle or spiral. Historical understanding is therefore logically different from scientific explanation: the type of connections, which historians present, is not causally determined and cannot be deduced from general laws. Historical connections are not necessary but contingent connections, which always could have been different from their actual form.

Neither are they based on empirical observation, but on plausible inferences, based on critical research of the sources and on analogical reasoning. Historical understanding at best produces plausible and probable knowledge and never certain or founded knowledge, as scientific explanation, based on sensory experience and logic presumably does. And being dependent on intellectual, interpretative activities by the historian, which can be trained but not formalized, historical understanding, thus, remains 'subjective,' i.e., embedded in the culture, time, and place of the historian. Based on this argument, Gadamer in his *Truth and Method* (1972) would radically historicize the idea of historical understanding.

The second type of argument in defense of methodological dualism is not ontological, but epistemological. It was first systematically developed by Wilhelm Dilthey and bears a resemblance to Droysen's argument, because it also harks back to Hegel's idea of a subject-object identity in history with its implied 'history as the autobiography of mankind model.' However, Dilthey does not ground historical method directly in the subject-object structure of the historical process itself, but in the 'inner experience' ('*innere Erfahrung*') of historical facts as 'mental facts' (*geistige Fakten*). The 'mental facts' of history and of other *Geisteswissenschaften* consist of feelings, thoughts, and strivings of human beings living in the past, which are only accessible through the mind and the inner experience of the historian. Like Droysen, Dilthey posits that the historian makes the mental facts of the past intelligible as expressions of ideas, by putting them in a particular spatio-temporal context alias in a *Lebenszusammenhang*. And, like Droysen, he contrasts historical understanding with scientific explanation by rejecting causal, law like reasoning for history. But unlike Droysen, Dilthey emphasizes the contrast of the 'inner experience' of the historian with the 'outer experience' of the scientist as the foundation of methodological dualism.

The third type of argument in defense of methodological dualism is of a methodological character. It is not based on an ontological claim that history deals with an object fundamentally different from nature, and neither is it based on an epistemological claim that the historical facts are different from natural facts, being accessible to an 'inner' way of knowing. According to the methodological argument, as developed by Windelband, Rickert, and Weber, two fundamental ways of knowing or methodological approaches can be distinguished, which are connected by two fundamentally different aims or knowledge interests. The first way to approach an object is a generalizing method; in this case, one abstracts from the particularity of the objects of knowledge and tries to detect and explain its general characteristics. According to Windelband, this generalizing method was specific for the nomothetic sciences (Greek: *nomos* = law, *tithemi* = to posit), such as the natural sciences.

The second fundamental approach to an object is the particularizing method; in this case one abstracts from the general characteristics of an object of knowledge and tries to describe its specific characteristics. According to Windelband this particularizing method was specific for the ideographic sciences (Greek: *ideos* = particular, *graphein* = to describe), such as history. However, both the generalizing and the particularizing methods are independent of the nature of their objects of knowledge; nature can be studied in an ideographic way (for instance, in geology) and history can be studied in a nomothetic way

(for instance, in developmental psychology). It all depends on one's methodological interest in case.

Rickert and Weber would develop Windelband's basic ideas by introducing the dichotomy between sciences of nature (*Naturwissenschaften*) and sciences of culture (*Kulturwissenschaften*). Rickert introduced the idea that knowledge interests behind disciplines were guided by values (*Wertbezogenheit*); Weber pointed out that values usually change through time. According to him, the change of the guiding values behind the sciences of culture could explain why those disciplines did not accumulate knowledge in the systematic form like the natural sciences, and why they seemed to enjoy an 'eternal youth' (Weber 1973). At best, the sciences of culture could accumulate knowledge at the intermediate level between individualizing descriptions and universal explanations in the form of ideal types, according to Weber. He suggested comparison as the method at this intermediate level. Lübbe, Veyne, and Mink would develop the Weberian methodological position later, identifying history with the particularizing methodological interest (Lübbe 1977, Veyne 1978, Mink 1987).

All in all, different conceptions of historical method are connected to different theories of historical knowledge. Since the characteristics of historical knowledge are usually defined in contrast with scientific knowledge as the major 'contrast class,' theories of natural science are also presupposed in theories of historical knowledge. Fundamental changes in theories of natural science, therefore, affect theories of historical knowledge by detour. The rise of postempiricism, postpositivism, and of the historical turn in the philosophy of natural science, thus have undermined most traditional, classical defenses of methodological dualism in an indirect way by undermining the classical unitary model of science. A few remarks must suffice.

Postempiricism has effectively undermined the presupposition that scientific knowledge can ever be grounded on certain foundations in sensory experience. The opposition between grounded and certain knowledge at the one side—traditionally attributed to the natural sciences—and inferential and probable knowledge at the other—traditionally attributed to the human sciences—has thus vanished. This opposition has been replaced by the insight that all sensory experience is guided and framed by concepts and theories. All empirical knowledge, therefore, is based on interpretative activities of the knowing subject because it is 'theory-laden.' Therefore, all knowledge is fallible and open to later reinterpretations. Postempiricism, thus, has dissolved the presupposed absolute contrast between a 'subjective' kind of knowledge, based on interpretation—characteristically for the *Geisteswissenschaften*, the *Kulturwissenschaften*, or the historical sciences—and an 'objective' kind of knowledge, based on empirical observation and experiment, not in need of interpretation. Postempiricism has thus relativized both the distinction between theoretical knowledge and empirical knowledge, at the one side, and the distinction between interpretative knowledge and empirical knowledge, at the other. Thus, it has brought conceptualization and interpretation to the center of the stage in all disciplines. In doing so, postempiricism has undermined traditional underpinnings of the idea of fundamentally distinct scientific and humanistic methods and the idea of a fundamental contrast between the 'objectivity' of science and the 'subjectivity' of the humanities. Both scientific domains aim at empirical argument and inter-subjectivity (Megill 1994, Oexle 1998).

Postpositivism, based on postempiricism, in turn, has effectively undermined the presupposition that all scientific explanations can be reduced to one basic model, i.e., the subsumtion model of explanation. By discrediting this presupposition for natural science, (Salmon 1990), postpositivism *a fortiori* has discredited the subsumtion model for the humanities. Moreover, the fact that the explanation of mental states by nonmental states has not succeeded up to the present has also discredited unitary views. (Bohman 1991).

In the theory of history, postempiricism and postpositivism have led to a renewed reflection on the nature of historical knowledge in postmodernism and in a renewed interest in approaches inspired by literary theory (see *History and Literature*). Now the classical science model was discredited, the roots of history in rhetoric and literature were rediscovered, leading to a renewed interest in the linguistic forms in which history is represented, i.e., in theories of historical representation. The postmodern vanguard, oft identified with the linguistic turn (see *Linguistic Turn and Discourse Analysis in History*), pushed both theory and method in history into the background, locating historical representation beyond epistemology and methodology, on the terrain of aesthetics, rhetoric, and politics (White 1973, Ankersmit and Kellner 1995, Fay et al. 1998).

This move, in fact, harks back to the premethodical and predisciplinary roots of history, cutting history loose from its a claim to constitute a truth-seeking discipline and to produce 'objective' knowledge (Rüsen 1983–89, Lorenz 1997). The history of theory and methods in history, thus, has come full circle at the turn of the twentieth century. Nevertheless, the pendulum is already swinging back, leading to broader approaches of history, which keep up the cognitive claims of the discipline based on both research (*Forschung*) and composition (*Darstellung*), while integrating the literary and the practical dimensions of history (see *History: Forms of Presentation, Discourses, and Functions*).

Bibliography

Ankersmit F R, Kellner H (eds.) 1995 *A New Philosophy of History*. London

Appleby J, Hunt L, Jacob M 1994 *Telling the Truth about History*. New York
Bernheim E 1889 *Lehrbuch der historischen Methode*. Leipzig, Germany
Bernstein R 1983 *Beyond Objectivism and Relativism: Science, Hermeneutics and Practice*. Oxford
Bodin J 1595 *Methodus ad facilem historiarium cognitionem*. Geneva, Switerland
Bohman J 1991 *New Philosophy of Social Sciences: Problems of Indeterminacy*. Cambridge, UK
Collingwood R G 1993 *The Idea of History*. Oxford, UK
Dilthey W 1981 *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften*. Frankfurt, Germany
Droysen J-G 1977 *Historik*. Stuttgart, Germany
Fay B, Pomper P, Vann R (eds.) 1998 *History and Theory. Contemporary Readings*. Cambridge, UK
Gadamer H G 1972 *Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik*. Tübingen, Germany
Gardiner P (ed.) 1959 *Theories of History: Readings from Classical and Contemporary Sources*. New York
Ghandi L 1998 *Postcolonial Theory: A Critical Introduction*. New York
Goodman N 1978 *Ways of Worldmaking*. Indianapolis, IN
Hausmann T 1991 *Erklären und Verstehen. Zur Theorie und Pragmatik der Geschichtswissenschaft*. Frankfurt, Germany
Hempel C 1966 Explanation in science and history. In: Dray W (ed.) *Philosophical Analysis and History*. New York, pp. 95–127
Iggers G 1997 *Historiography in the Twentieth Century. From Scientific Objectivity to the Postmodern Challenge*. Hanover, USA
Koselleck R 1989 *Vergangene Zukunft. Zur Semantik geschichtlicher Zeiten*. Frankfurt, Germany
Kuhn T 1962 *The Structure of Scientific Revolutions*. Chicago
Lakatos I, Musgrave A (eds.) 1970 *Criticism and the Growth of Knowledge*. Cambridge, UK
Lorenz C 1997 *Konstruktion der Vergangenheit. Eine Einführung in die Geschichtstheorie*. Cologne, Germany
Lübbe H 1977 *Geschichtsbegriff und Geschichtsinteresse. Analytik und Pragmatik der Historie*. Basel, Switzerland
Megill A (ed.) 1994 *Rethinking Objectivity*. Durham, USA
Meier C, Rüsen J (eds.) 1988 *Historische Methode*. Munich, Germany
Meran J 1985 *Theorien in der Geschichtswissenschaft*. Göttingen, Germany
Mink L 1987 *Historical Understanding*. Ithaca, NY
Nagl-Docekal H 1982 *Die Objektivität der Geschichtswissenschaft*. Vienna
Novick P 1988 *'That Noble Dream.' The 'Objectivity question' and the American Historical profession*. Cambridge, UK
Oexle O-G (ed.) 1998 *Naturwissenschaft, Geisteswissenschaft, Kulturwissenschaft; Einheit—Gegensatz—Komplementarität?*. Göttingen, Germany
Rickert H 1986 *Kulturwissenschaft und Naturwissenschaft*. Stuttgart, Germany
Ricoeur P 1983–5 *Temps et récit* (3 Vols). Paris
Rorty R 1986 *Philosophy and the Mirror of Nature*. Oxford, UK
Rüsen J 1983–9 *Historische Vernunft. Grundzüge einer Historik* (3 vols.). Göttingen, Germany
Salmon W 1990 *Four Decades of Scientific Explanation*. Minneapolis, MN
Scott J 1991 Women's history. In: Burke P (ed.) *New Perspectives in Historical Writing*. Cambridge, UK, pp. 42–66
Seignobos C 1897 *Introduction aux études historiques*. Paris
Stern F (ed.) 1970 *Varieties of History. From Voltaire to the Present*. London
Veyne P 1978 *Comment on écrit l'histoire?* Paris
Weber M 1973 *Gesammelte Aufsätze zur Wissenschaftslehre*. Tübingen, Germany
White H 1973 *Metahistory. The Historical Imagination in Nineteenth Century Europe*. Baltimore, MD
Windelband W 1915 Geschichte und Naturwissenshaft. In: *Präludien. Aufsätze und Reden zur Einleitung in die Philosophie und ihrer Geschichte*, Vol. 2, Tübingen, Germany, pp. 136–61

C. Lorenz

HIV and Fertility

It is widely recognized that HIV had a catastrophic impact on mortality from the early 1990s in much of sub-Saharan Africa, and significant impacts in a few high prevalence countries of South East Asia. The interactions between HIV and fertility are not so well known, but they deserve the special attention of epidemiologists, demographers, and behavioral scientists since they influence our understanding of the epidemic trends and the sexual behavior which underlies this. Limited evidence of the impact of HIV on fertility is available from developed nations with low-level epidemics concentrated in high risk groups such as intravenous drug users, recipients of blood transfusions, and men who have sex with men, but this review concentrates on the situation in African countries with generalized, heterosexually transmitted epidemics. The fertility effects in these populations are much more clear cut, both because the high numbers involved make it relatively easy to collect representative data, and because historically low rates of contraceptive use make it feasible to distinguish between involuntary, biological effects and behavioral responses to the epidemic.

1. *The Nature of HIV and Fertility Interactions*

Early speculation on the likely effects of the HIV epidemic on fertility tended to concentrate on change in intentional behavior, and predicted resulting fertility increases. Some authors believed that increases in child mortality might fuel increases in childbearing activity, as parents would want to have many children to ensure that at least a few survived (Caldwell et al. 1992). Others feared that in their quest for safe, uninfected partners, men would seek partnerships with progressively younger girls, thereby lowering the starting age for childbearing. Later writing (e.g., Gregson 1994) has given much more weight to biomedical factors and to unintended consequences of changing behavior.

1.1 *Biomedical Factors*

HIV lowers fertility directly in several ways, not all of them yet fully understood. Most importantly, it causes

spontaneous abortions and stillbirths (Brocklehurst and French 1998), and in the later stages of the disease women may suffer from amenorrhea (Widy-Wirsky et al. 1988). In men, it may reduce spermatogenesis (sperm production) (Krieger 1991), and lower coital frequencies have been reported by couples in which the male partner has begun to suffer from opportunistic infections.

Fertility impairments are also caused by other sexually transmitted diseases (STDs), such as syphilis, chlamydia and gonorrhea. HIV is much more prevalent among those who have other STDs, partly because of common behavioral risk factors, partly because HIV-positive individuals are more vulnerable to these infections, but most importantly because the presence of ulcerative STD infections increases the likelihood of HIV transmission (Fleming and Wasserheit 1999). These direct biological factors affect only a minority of couples in the population, namely those in which one or the other partner is already infected with the HIV virus. Therefore they may not have a large impact on fertility at the population level, but they are important determinants of fertility differentials between HIV-positive and HIV-negative women.

1.2 Biosocial Responses to the Epidemic

A range of behavioral responses to the epidemic may alter fertility in the whole population, even though they affect HIV-positive women more than the HIV-negative. Studies in Zimbabwe have indicated higher rates of widowhood, higher divorce rates as a result of knowledge or suspicion of HIV infection in either partner, and lower remarriage rates (Gregson et al. 1997). Increased condom use was reported in Uganda, especially in casual relationships, and there was a decrease in extramarital sexual activity (Kilian et al. 1999) and a postponement of entry into sexual unions by adolescents. These factors would tend to decrease nonmarital fertility. Other behavioral changes in response to the epidemic may tend to cause increases in fertility—for example, in Zimbabwe, some women have chosen to curtail breastfeeding, since it is becoming more widely known that the virus may be transmitted to infants through breastmilk (Gregson et al. 1997). Traditional periods of postpartum abstinence may be shortened to curb the tendency for men to indulge in extramarital affairs—Cleland et al. (1998) showed that in Benin this was more likely to happen when the period of abstinence was lengthy.

1.3 Behavioral Change Consciously Intended to Affect Fertility

There is little evidence that insurance or replacement effects motivate fertility increases: on the contrary, opinion surveys in Tanzania, Rwanda, and Zimbabwe show that the majority believe HIV-infected women should refrain from childbearing in order to limit vertical transmission of the virus and to avoid leaving behind large numbers of orphaned children (Setel 1995, Keogh et al. 1994, Gregson et al. 1997). Followup studies of women in Zaire who were aware of their HIV status have indicated more contraceptive use by the HIV-positive (Ryder et al. 1991), and higher rates of induced abortion have been reported in some developed countries (Thackway et al. 1997).

1.4 Changes in Population Structure

Changes in the birth rate can come about not only because of changes in age-specific fertility rates but also because of structural changes in the population. Gregson et al. (1999) showed that HIV-related mortality will tend to make the population structure younger, and with proportionately more women in the peak childbearing ages of 20–29 this would tend to push up the crude birth rate. Similarly, over time, the excess mortality suffered by HIV-positive women with primary or secondary sterility due to coinfection with other STDs would tend to raise the proportion of fecund women in the population and thereby raise the overall birth rate, even if the fertility rates of HIV-positive and HIV-negative women did not change. Fecundity could also rise if interventions to treat classical STDs were successful.

2. The Scale of the Effect of HIV on Fertility

Empirical evidence for fertility differentials between HIV-positive and HIV-negative women has accumulated from various longitudinal studies. It is summarized below. It is rather more difficult to find unambiguous direct evidence for population level effects, though approximate effects can be estimated by assuming that in the absence of HIV infection all women would experience the fertility patterns currently experienced by the HIV-negative. However, this does not allow for the fact that HIV-positive women may be selected for high levels of other STDs and recent involvement in sexual activity.

2.1 Differences in Fertility Between HIV-positive and HIV-negative Women

Figure 1 is based on data from the longest running observational study of HIV in a developing country setting, in the Masaka district of Southern Uganda (Carpenter et al. 1997). It shows the typical relationship between fertility in HIV-infected and uninfected women. At ages under 20, HIV-infected women have higher fertility than the uninfected, but at age 20 and

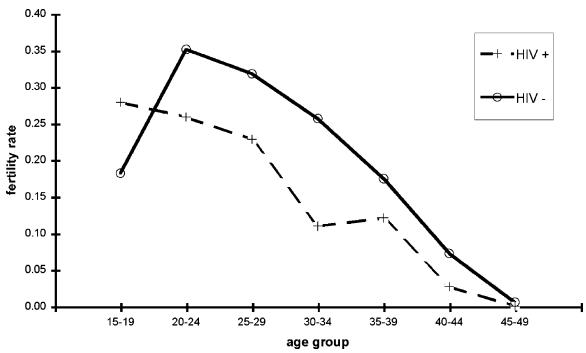

Figure 1
Age specific fertility by HIV status, Masaka, Uganda, 1989–1996

over HIV-positive women have significantly lower fertility than the HIV-negative. This is because in the youngest age group those women who have borne children are selected for an early start to sexual activity—virgins are not at risk of childbearing nor are they exposed to the risk of sexually transmitted HIV infection. At later ages, when almost all women are—or have been—sexually active, these selection effects are not so important, and the fertility-reducing effects of HIV, described above, clearly predominate.

2.2 Population Level Effects

Table 1 (adapted from Zaba and Gregson 1998) indicates the proportionate reduction in fertility attributable to HIV in a number of severely affected populations in Eastern Africa. HIV prevalence in these populations ranges from 8 percent to 30 percent, and fertility in the population as a whole is 4 percent to 12 percent lower than the fertility of HIV-negative women. This represents an average reduction of around 0.4 percent in total fertility for every percentage point increase in HIV prevalence. These observations are all based on studies in populations with low acceptance of family planning—the populations in Table 1 all have contraceptive prevalence rates of less than 15 percent.

There is some indication (Zaba et al. 2000) that fertility differentials between HIV-positive and HIV-negative women are not as pronounced in African countries with higher rates of contraceptive use, such as Kenya and Zimbabwe, where over 30 percent of women use modern contraception. However, in these populations fertility had already begun to fall before the HIV epidemic took hold, and it would appear highly likely that the impetus towards lower fertility in both HIV-positive and HIV-negative women would be reinforced by the behavioral changes described above. It would be difficult to ascertain whether fertility would have fallen more slowly in the absence of the epidemic. The 'attributable change' statistic is an indicator of the overall effect of differential fertility. It does not take account of the common secular trends for infected and uninfected women.

3. Consequences of HIV and Fertility Interactions

As a result of the overall fertility reduction in populations severely affected by the HIV epidemic, population growth rates will be even lower than would be expected on the basis of the dramatic increases in mortality. A fall of 10 percent in total fertility would cause a reduction of the order of half a percentage point in the growth rate, making it more likely that some of the worst affected countries could experience periods of negative growth. Since falling fertility causes population aging, this will counter the tendency of HIV-related mortality to make the age structure of these populations even younger.

4. Measurement of HIV Prevalence in Sentinel Surveillance Populations

One of the more problematic effects of the relationship between HIV and fertility is on the measurement of HIV prevalence. The most widely used source of prevalence data—indeed the source on which all national-level estimates of HIV prevalence are based (US Bureau of the Census 2000)—is sentinel surveillance in antenatal clinics. Since HIV positive women have lower fertility, they will be underrepre-

Table 1
Population level effects of fertility reduction associated with HIV

	Mwanza, Tanzania 1992–1994	Kisesa, Tanzania 1994–1997	Masaka, Uganda 1989–1996	Rakai, Uganda 1994–96	Ft Portal, Uganda l994–1995	Lusaka, Zambia 1995–1996	Mposhi, Zambia 1995–1996
HIV prevalence (%)	15.6	8.4	11.8	21.2	26.0	29.9	17.4
Total fertility							
HIV-women	4.29	6.41	6.80	4.75	7.95	5.69	7.40
all women	4.09	5.89	6.54	4.34	6.97	5.09	6.86
% attributable change	−4.7	−8.1	−3.8	−8.7	−12.3	−10.6	−7.3

sented in antenatal clinics, so that the resulting estimates of HIV prevalence underestimate the true extent of infection among women of childbearing age. Where community-based data are available it has been noted that HIV prevalence among adult males is generally lower than that in the female population (UNAIDS 2000), but conventionally the estimates based on surveillance of pregnant women continue to be used to represent prevalence for adults of both sexes. To obtain reliable estimates of HIV prevalence levels in women of childbearing age, one needs to adjust prevalence data collected in antenatal clinics (Zaba et al. 2000), particularly if one seeks to derive estimates of HIV incidence from trends in prevalence in young women. Prevalence estimates for young women may also be affected by changes in the age of sexual debut.

4.1 Trends in Orphanhood and Child Mortality

Reduced fertility in HIV-positive women implies lower effects of HIV on child mortality than were originally anticipated, and fewer orphans will be left behind at the death of an HIV-positive mother than for one who is HIV-negative. At the individual level, many of the fertility reducing factors (such as HIV-associated illnesses or the likelihood of divorce and widowhood) strengthen with duration of infection. As a result, the downward pressure on fertility, and the associated secondary effects on orphanhood and child mortality, will change with the duration of the epidemic, as the balance between women who have been recently infected and those who have been infected for longer periods shifts in favour of the latter. Gregson et al. (1999) have modeled the effect that such trends will have on the number and age distribution of orphans and child deaths.

4.2 Family Planning in the Era of AIDS

Family planning programmes in sub-Saharan Africa have traditionally promoted hormonal contraception, namely oral contraceptive pills, injectibles, and implants. Barrier methods are regarded with suspicion, as they are considered to be less reliable than hormonal methods and because of their association with commercial and illicit sex. However, effective and systematic use of condoms for disease prevention would also make them highly effective contraceptives. Conversely, an acceptance of condoms for family planning purposes in stable marital unions would destigmatize them and make them more acceptable in other circumstances also. The AIDS pandemic will require a major shift in the contraceptive mix offered by family planning programs and a new willingness to discuss STD prevention in the context of family planning services.

See also: AIDS, Geography of; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants; Infant and Child Mortality in the Less Developed World; Reproductive Rights in Developing Nations

Bibliography

Brocklehurst P, French R 1998 The association between maternal HIV infection and perinatal outcome: A systematic review of the literature and meta-analysis. *British Journal of Obstetrics and Gynaecology* **105**: 836–48

Carpenter L, Nakiyingi J, Ruberantwari A, Malamba S, Kamali A, Whitworth J 1997 Estimates of the impact of HIV infection on fertility in a rural Ugandan population cohort. *Health Transition Review* **7**(suppl.): 113–26

Caldwell J C, Orubuloye I O, Caldwell P 1992 Fertility decline in Africa: A new type of transition? *Population and Development Review* **18**: 211–42

Cleland J, Ali M, Capo-Chichi V 1998 Post-partum sexual abstinence in West Africa: Implications for AIDS-control and family planning programmes. *AIDS* **13**: 125–31

Fleming D T, Wasserheit J N 1999 From epidemiological synergy to public health policy and practice: The contribution of other sexually transmitted diseases to sexual transmission of HIV infection. *Sexually Transmitted Infections* **75**: 3–17

Gregson S 1994 Will HIV become a major determinant of fertility in sub-Saharan Africa? *Journal of Development Studies* **30**: 650–79

Gregson S, Zhuwau T, Anderson R M, Chandiwana S K 1997 HIV-1 and fertility change in rural Zimbabwe. *Health Transition Review* **7**(suppl.): 89–112

Gregson S, Zaba B, Garnett G 1999 Low fertility in women with HIV and the impact of the epidemic on orphanhood and early childhood mortality in sub-Saharan Africa. *AIDS* **13** (suppl. A): S249–58

Keogh P, Allen S, Almedal C, Temahagili B 1994 The social impact of HIV infection on women in Kigail, Rwanda: A prospective study. *Social Science and Medicine* **38**: 1047–53

Kilian A, Gregson S, Ndyanabangi B, Walusaga K, Kipp W, Salmuller G, Garnett G, Asiimwe-Okiror G, Kabagambe G, Weiss P, von Sonnenburg F 1999 Reductions in risk behavior provide the most consistent explanation of declining HIV-1 prevalence in Uganda. *AIDS* **13**: 391–98

Krieger J N 1991 Fertility parameters in men infected with HIV. *Journal of Infectious Diseases* **164**: 464–69

Ryder R W et al. 1991 Fertility rates in 238 HIV-1 seropositive women in Zaire followed for 3 years post partum. *AIDS* **5**: 1521–7

Setel P 1995 The effects of HIV and AIDS on fertility in east and central Africa. *Health Transition Review* **5**(suppl.): 179–90

Thackway S V, Furner V, Mijch A, Cooper D A, Holland D, Martinez P, Shaw D, van Beek I, Wright E, Clezy K, Kaldor J M 1997 Fertility and reproductive choice in women with HIV-1 infection. *AIDS* **11**: 663–8

United States Bureau of the Census 2000 *HIV/AIDS Surveillance Data Base* (see also the companion web site, http://www.census.gov:80/ipc/www/hivaidsn.html). US Bureau of the Census, Washington, DC

UNAIDS 2000 *The Status and Trends of HIV/AIDS in the World.* Technical report of the Monitoring the AIDS Pandemic Symposium. Durban, 13th International AIDS Conference, July 9–14, 2000

Widy-Wirski R et al. 1988 Evaluation of the WHO clinical case definition for AIDS in Uganda. *Journal of the American Medical Association* **260**: 3286–9
Zaba B, Gregson S 1998 Measuring the impact of HIV on fertility in Africa. *AIDS* **12**: S41–50
Zaba B, Carpenter L, Boerma T, Gregson S, Nakiyingi J, Urassa M 2000 Adjusting ante-natal clinic data for improved estimates of HIV prevalence among women in sub-Saharan Africa. *AIDS* **14**: 2741–50

B. Zaba and S. Gregson

HIV Risk Interventions

Human Immunodeficiency Virus (HIV) infection, the cause of Acquired Immune Deficiency Syndrome (AIDS) has become a significant threat to global public health faster than any previous epidemic (Mann and Tarantola 1996). The genetic nature of HIV evades the development of a preventive vaccine and a cure for HIV infection remains a distant hope. HIV is transmitted through direct contact with HIV infected blood, semen, and vaginal secretions. Although HIV is transmitted during birth from mother-to-infant and through contaminated blood products the majority of AIDS cases in the world have resulted from HIV transmission between adults engaged in high-risk practices. Behavioral interventions therefore remain the most realistic means for curtailing the spread of HIV infection. Effective HIV risk reduction interventions target two principle behaviors: (a) sharing HIV contaminated drug injection equipment and (b) decreasing exposure to HIV infected semen, vaginal secretions, and sexually derived blood. Interventions to change injection equipment sharing and high-risk sexual practices can, therefore, dramatically effect the spread of HIV. In this article, factors associated with HIV transmission risks and interventions directed at reducing risks associated with injection drug use and sexual relations are examined.

1. Factors Associated with HIV Transmission Risks

Effective HIV risk reduction interventions are tailored to the immunological, epidemiological, cultural, and social correlates of HIV transmission risk behaviors. HIV transmission risk is affected by both biological and sociocultural co-factors.

1.1 Biological Co-factors

A variety of genetic subtypes of HIV exist, with each having its own unique features relevant to transmission susceptibility. Genetic subtypes of HIV also differ in their geographic distribution. Therefore, genetic subtypes of HIV that are most efficiently transmitted during vaginal intercourse are more prevalent in some regions of the world than others and partially account for the regional differences in epidemic patterning (Mann and Tarantola 1996). Sexually transmitted infections, particularly syphilis, chancroid disease, and gonorrhea, cause ulcers and degradation of mucous membranes increasing both HIV transmission and susceptibility. Another biological factor associated with HIV transmission risk is the infectious state of infected persons. Individuals with lower blood concentrations of HIV, or viral load, may be at lower transmission risk than persons with greater concentrations. Therefore, the availability and use of HIV suppressing medications may influence the transmission of HIV. These biological influences on HIV transmission risks will likely interact with specific prevention strategies targeted by given risk reduction interventions.

1.2 Sociocultural Co-factors

Culture and social structures create a context for HIV transmission risk. Gender biases, power in sexual relationships, social norms supportive of risk and protective behaviors, poverty, drug addiction, and homophobia are known factors that influence HIV subepidemics (Kalichman 1998). HIV transmission occurs in personal relationships. Therefore, the social connections between persons, or social networks, play a critical role in both risk producing and risk reducing behaviors (Friedman et al. 1999). Social environments necessitate tailoring behavioral interventions to subpopulation characteristics. Also, behavioral interventions must adapt over time to the ever-changing social context of HIV-AIDS.

2. Interventions for Injection Drug Use Risks

HIV and other diseases are spread through sharing contaminated needles and syringes for injecting drugs. Tainted injection equipment, therefore, not the drugs themselves transmit HIV. The use of sterile needles and syringes assures prevention against HIV and other diseases. Syringe access and exchange programs have demonstrated substantial reductions in new HIV infections among drug injectors (NIH Panel 1997). Another effective means of preventing HIV transmission via injection drug use is availability of drug treatment on demand. Obviously, when one stops injecting drugs the risks for injection-related HIV also ceases. Yet despite the known public health value of providing access to syringes and to drug treatment, political forces and limited economic resources have

prohibited policies for wide-scale implementation of these intervention strategies (National Academy of Sciences 1995). Cleaning injection equipment is considered a reasonable alternative when sterile syringes are unavailable. Behavioral interventions have shown success in training injection drug users in proper needle and syringe hygiene skills.

HIV risk reduction for injection drug users, therefore, occurs along a spectrum of health objectives. Drug treatment is the ideal intervention because it addresses the risks posed by drug addiction as well as HIV infection. Persons who do not receive drug treatment, however, can reduce the harm of drugs by using sterile injection equipment. When sterile equipment is unavailable, persons are advised to practice needle and syringe cleaning. The philosophy of accepting incremental steps to reduce risks when risks cannot be eliminated is called harm reduction and has been highly effective in framing the expected results of HIV risk reduction strategies.

3. *Interventions for Sexual Relationship Risks*

The vast majority of AIDS cases in the world have resulted from exposure to HIV in sexual relationships. Interventions designed to reduce sexual risk practices that permit HIV transmission—vaginal, anal, and oral intercourse without condoms or other barriers—have occurred at three primary levels: individual risk reduction counseling, small group workshops, and community mobilization campaigns. In each case, effective interventions that reduce high-risk behavior almost universally include three primary elements: educational information about HIV and how it is transmitted and prevented; motivation to induce individuals to make behavioral changes; and interactive exercises to improve risk reduction behavioral skills, such as sexual assertiveness skills and condom use skills (Fisher and Fisher 1992).

3.1 *Components of Sexual Risk Reduction Interventions*

Skills enhancement interventions consist of education, sensitization, and instruction in specific behavioral skills to reduce risk for HIV infection. Educational exercises and activities designed to help individuals take steps toward risk reduction are tailored to reflect the context, values, customs, and language of targeted populations. Therefore, similar intervention content is used for men and women of various backgrounds, but the contextual issues and examples vary accordingly.

3.1.1 HIV risk information and motivation enhancement. HIV prevention interventions provide basic information related to HIV transmission, illness, and prevention. Risk education can include facts about viral transmission, local prevalence of HIV and AIDS, clarification of misconceptions, dispelling AIDS myths, and description of HIV antibody testing. Discussions often couch information in personal life situations, relationships, risk behaviors, and increased motivations to reduce risk. Education also encourages individuals to assess their own behavioral risk histories and initiate discussions of individual life-situations that have placed them at risk. Theoretically, increased understanding of risk and personalization of vulnerability to contracting HIV are believed to increase one's readiness to change and enact behaviors to reduce risks.

3.1.2 Behavioral self-management skills. Behavioral self-management skills related to HIV risk reduction are acts that individuals can perform to reduce their risks for HIV infection with minimal involvement of their sexual partners. Individuals are instructed in techniques for examining their risk and cues that precede sexual risky situations. For example, persons are instructed in ways to identify environmental and cognitive-affective cues that serve as 'triggers' for high-risk situations, including mood states, substance use, settings, and sexual partner characteristics. Strategies for avoiding or managing high-risk triggers and situations are discussed with individuals generating their own risk avoidance behaviors. Coping strategies for high-risk situations may include performing safer sex acts, redirecting sexual activities toward safer sex alternatives, carrying condoms, and avoiding sex after substance use. Identification of barriers to risk reduction, such as substance use and access to condoms are placed in the context of problem solving.

A key to behavioral self-management skills is condom use during sexual intercourse. Interventions educate individuals about the efficacy and proper use of condoms, but the skills training component goes beyond education by including safer sex exercises. Instruction, modeling, practice, and guided feedback desensitizes individuals to condoms, capitalizes on changing negative attitudes toward condoms. Increased experience with condoms has been shown to decrease rates of condom failure, a potential benefit of condom skills training. It should be noted, however, that condom use is a behavioral skill that includes interaction with one's partner, and in the case of women in heterosexual relationships as well as men who are the receptive partner during anal intercourse, use of male condoms can require techniques for putting condoms on partners as well as discussing and persuading partners to use condoms.

3.1.3 Sexual communication skills. Sexual negotiation, sexual assertiveness, refusal of high-risk sex,

and nonverbal communication are the most common types of communication skills included in HIV risk reduction interventions. Increasing skills for resisting partner coercion to engage in sexual intercourse without condoms, and increasing comfort discussing safer sex with partners in advance of sexual activity occur through instruction, modeling, and practice in communicating sexual decisions and discussing individual limits on behavior. Role playing situations allows for experience in what it may be like to talk with partners about risk and risk reduction. Modeling effective responses, practice, and guided feedback again form the core for the skills training model. Individuals learn techniques to communicate feelings and sexual behavior limits through practice sessions conducted to increase proficiency in communication skills and comfort in discussing sexual alternatives with partners.

4. *Levels of Sexual Risk Reduction Interventions*

HIV risk reduction intervention elements that include risk education, risk sensitization, and behavioral and communication skills building are found in all effective HIV risk reduction interventions delivered to individuals, groups, and communities. The most common HIV prevention activities occur in face-to-face counseling sessions, usually in conjunction with HIV antibody testing. Risk reduction counseling can be effective when it includes information, motivation, and skills building components, even when delivered in as short a time as 20 minutes (Kamb et al. 1998). However, there is now compelling evidence that HIV testing delivered with information-based counseling that does not include behavioral skills building components is ineffective at reducing HIV risks (Weinhardt et al. 1999).

Workshops delivered to small groups of individuals, typically 6 to 12 persons, that include information, motivation, and skills building activities have been shown effective in rigorously controlled studies with men who have sex with men, heterosexual men, women, and adolescents (Kalichman 1998). Finally, HIV prevention efforts delivered at the community level have also been shown effective when they include information, motivation, and behavioral skills. For example, peers and natural opinion leaders trained in skills for disseminating information and motivational messages can successfully change patterns of risk behaviors for entire populations (Kelly et al. 1997). Thus, sexual risks for HIV infection can be effectively reduced when they provide information, risk sensitization, and behavioral skills building opportunities in highly interactive formats.

See also: AIDS (Acquired Immune-deficiency Syndrome); AIDS, Geography of; Drug Use and Abuse: Cultural Concerns; Drug Use and Abuse: Psychosocial Aspects; Health in Developing Countries: Cultural Concerns; Mortality and the HIV/AIDS Epidemic; Public Health; Sexually Transmitted Diseases: Psychosocial Aspects

Bibliography

Fisher J D, Fisher W A 1992 Changing AIDS-risk behavior. *Psychological Bulletin* **111**: 455–74

Friedman S R, Curtis R, Neaigus A, Jose B, Des Jarlais D 1999 *Social Networks, Drug Injector's Lives, and HIV/AIDS*. Kluwer Academic/Plenum, New York

Kalichman S C 1998 *Preventing AIDS: A Sourcebook for Behavioral Interventions*. Erlbaum, Hillsdale, NJ

Kamb M, Fishbein M, Douglas J, Rhodes F, Rogers J, Bolan G, Zenilman J 1998 Efficacy of risk-reduction counseling to prevent Human Immunodeficiency Virus and sexually transmitted diseases. *Journal of the American Medical Association* **280**: 1161–7

Kelly J A, Murphy D, Sikkema K, McAuliffe T, Roffman R, Solomon L, Winett R, Kalichman S C 1997 Outcomes of a randomized controlled community-level HIV prevention intervention: Effects on behavior among at-risk gay men in small US cities. *Lancet* **350**: 1500–5

Mann J, Tarantola D J 1996 *AIDS in the World II*. Oxford University Press, New York

National Academy of Sciences 1995 *Preventing HIV Transmission: The Role of Sterile Needles and Bleach*. National Academy Press, Washington, DC

NIH Panel 1997 *National Institutes of Health Consensus Development Statement on Interventions to Prevent HIV Risk Behaviors*. NIH Office of Medical Applications Research, Bethesda, MD

Weinhardt L, Carey M P, Johnson B, Bickman N 1999 Effects of HIV counseling and testing on sexual risk behavior: A meta-analytic review of the published research, 1985–97. *American Journal of Public Health* **89**: 1397–1405

S. C. Kalichman

Hobbes, Thomas (1588–1679)

Thomas Hobbes was born on Good Friday, April 5, 1588, in Malmesbury in Wiltshire, UK. It was not until some months later that the Spanish Armada appeared in the Channel, but the threat that it would do so was widely known and there was widespread apprehension. Hobbes was fond of referring to this conjunction of his birth with the Armada:

> For Fame had rumour'd, that a Fleet at Sea,
> Wou'd cause our Nations Catastrophe;
> And hereupon it was my Mother Dear
> Did bring forth Twins at once, both Me, and Fear.
> (*The Life of Mr. Thomas Hobbes of Malmesbury, written by himself.*)

This statement of Hobbes' is often referred to because of the role that fear played in his account of human nature and the genesis of political order.

Hobbes was born of a poor family, but obscure birth was no bar to the career of a clever young man at that time. His education made him a fluent linguist and translator, and though he wrote a considerable amount in Latin, which was still the common language of the learned, he became the first philosopher to write his most significant works in English. After study at Magdalen Hall in Oxford, he was recommended by the Principal of his hall to take the role of tutor and secretary in the household of William Lord Cavendish, who ten years later became, Earl of Devonshire. Hobbes spent almost all of the rest of his life in the employment of the Earls of Devonshire or their close relatives the Earls of Newcastle.

As a member of such households, Hobbes several times took sons of his employers on tours of Europe, which gave him a probably unparalleled chance to meet intellectuals: he met Galileo, whose work greatly influenced him as a response to the widely accepted scepticism of the time, Gassendi, Mersenne, and later Descartes, amongst others. These acquaintanceships continued through correspondence and when he fled to France in 1640 in reasonable fear that his political writings might result in threats to his life or liberty. In England, too, he met important people and intellectual leaders; Aubrey records that at one time he acted as secretary to Francis Bacon and also helped to translate some of Bacon's essays into Latin, with Bacon commenting that he liked Hobbes best of all those who had taken dictation from him because Hobbes understood what it was that he was writing.

Modern science was in a formative stage at that time, with no generally accepted scientific method. As an intellectually lively man, Hobbes in middle life became caught up in the arguments about this set of problems about the investigative methods appropriate to science, and Aubrey records him as saying that it was first reading Euclid, whom he took to produce definitive proofs, which made him alive to the problem and possibility of establishing a scientific method. He was also quite explicitly influenced by Galileo's *Dialogue Concerning the Two Chief Systems of the World*, published in 1632. Part of his concern with method came out in his dispute with Boyle (an account of which is given in Shapin and Schaffer 1985) about the significance of experiment, which Hobbes downplayed. Hobbes distinguished quite sharply between knowledge and opinion, knowledge requiring the sort of certainty he found in geometry. He held to the view that people were not capable of genuine knowledge of the external world or its characteristics, but that we can know, with the certainty required for knowledge, the contents of our minds. Genuine knowledge will therefore be of ideas and the relations between them. Observation can give us opinion, but cannot provide genuine knowledge. Nevertheless, Aubrey says, 'He was wont to say that he had rather have the advice, or take Physique from an experienced old Woman, that had been at many sick people's Bed-sides, then from the learnedst but inexperienced Physitian,' so he did not ignore the importance of experience. His interest in optics and ballistics, which developed his interest in mathematics, arose from the Newcastle family's interest in the production of an effective telescope for military purposes, and he generally took science to be important because of its practical consequences.

This practical approach probably explains at least part of the dispute about the historians' Hobbes and the philosophers' Hobbes. During the 1960s and 1970s, Quentin Skinner produced a set of articles (1964, 1965, 1966, 1972) about the interpretation of Hobbes's writings on politics, the writings for which Hobbes is still remembered. Skinner's thesis was that more detailed consideration needed to be given to Hobbes's relationship to other writers of his time and the responses that he and they showed to the political events they lived through. Philosophers, it was implied, had taken Hobbes ahistorically and hence misinterpreted him. The most detailed reply (Warender 1979) does not deny that Hobbes was concerned to contribute to practical political debate of his time, but does deny that the differences between philosophers in their interpretations of Hobbes simply reflect a failure to put Hobbes in his historical context: they are, rather, differences about the account to be given of the philosophical backing for Hobbes's position. Skinner is clearly right in his understanding of why Hobbes took up the issues he took up, why he thought that certain problems needed to be dealt with, what allusions he might be making, and what he was trying to contribute to the current political debate, etc., that require the historical context. However, it is also true that Hobbes was not simply contributing to the practical debate but thought that he had a special contribution to make because he was producing a science of politics, the science involving timeless truths. He thought that his views were just as relevant to French disturbances of the time as to those in the UK. There is philosophers' Hobbes to study, and omission of the philosophical aspect of his work leads to misinterpretation.

The events at issue in the dispute he saw himself as entering, though Hobbes did not see himself as only, or always as primarily, concerned with politics, were political events. Hobbes was born in the reign of Elizabeth I and died on December 3, 1679, in the reign of Charles II, during which time he had been a tutor in mathematics to the Prince of Wales (accepting the position in 1646), had been awarded a pension by Charles II, was probably invited to stand as a candidate for the Short Parliament, and had lived through the English civil wars of the 1640s and the Commonwealth. He had had his books burnt in the University of Oxford and had been forbidden to publish his ideas in England, had (with some reason),

felt threatened with accusations of heresy, and had seen much turmoil in Europe. The issues that concerned him in politics were basically issues of stability. He was concerned with what was required for a state to remain stable and peaceful so that people could live their lives with the security they needed in order to be able to pursue their projects.

Hobbes's initial intellectual interests were in such fields as optics, ballistics, and mathematics, especially geometry. He is remembered by many for his attempts to square the circle, and he took part in vituperative disputes with the Savilian Professors of Mathematics at Oxford and especially with Wallis. His enduring contribution to our intellectual heritage, though, is in the fields of morality and politics, which he tried to establish in a scientific manner. His main publications in these fields were *The Elements of Law; De Cive* (originally in Latin, but there soon appeared an English translation about which there is some disagreement concerning the extent to which Hobbes approved of it); *Leviathan* (originally in English but translated with changes into Latin so that it could be read elsewhere); and *De Homine* (which still lacks a complete English translation, though there is a partial translation in Gert 1972). The translations show Hobbes's keenness to make his ideas available as a contribution to the current political debates in a number of countries, and this is also reflected in the changes in his styles of presenting his ideas (see, for example, Johnston 1986). His *Behemoth*, an account of the causes of the English civil wars of the 1640s, also reflects his ideas on politics. He discusses methods of proof and the nature of science and true knowledge frequently, the most extended discussion being in *De Corpore*, an English translation of which appears in volume 1 of the Molesworth edition of *Hobbes: English Works*.

Hobbes's best-known work is *Leviathan*, the last of his English writings directly on the science of politics. It appeared in 1651, at the time of Hobbes's return to England. There are two different stories about how it fits into the historical context and what it was intended to achieve. The Royalist Earl of Clarendon (in *A Survey of Mr Hobbes his Leviathan*) reports Hobbes as saying

> … that he knew when I read his Book I would not like it, and thereupon mentioned some of his Conclusions; upon which I asked him, why he would publish such doctrine: to which, after a discourse between jest and earnest upon the Subject, he said *The truth is, I have a mind to go home*.

Many Royalists did take the book that way, as a defence of Cromwell and his Commonwealth. Aubrey, on the other hand, says

> He wrote and published the *Leviathan* far from the intention either of disadvantage to his Majestie, or to flatter Oliver (who was not made Protector till three or four yeares after) on purpose to facilitate his returne; for there is scarce a page in it that he does not upbraid him

and goes on to suggest that the purpose was to provide justification for those who had fought on the Royalist side until all was lost and now faced the prospect of losing all their property if they did not recognise Cromwell's authority. The most that can be done to sort out this disagreement is to see what the arguments of *Leviathan* will support.

Hobbes's moral and political theory is presented as an account of how and why people emerge from their natural condition and live together in civil society. This involves giving an account of human nature, of the natural condition of mankind, and of the laws of nature that make the transition possible. There is still argument about just what Hobbes's position is on each of these matters. It is clear that Hobbes takes the natural condition of mankind (he rarely uses the expression 'state of nature') to be a state of insecurity and constant fear. Though not always a condition of actual fighting, it is a condition in which people will take an aggressive stance towards each other and in which nobody can be trusted and fighting is always likely to break out. The reasons for this are to do with human nature, and the solution is the making of a contract to establish a sovereign.

Hobbes has more to say about the actual origins of states in *A Dialogue Between a Philosopher and a Student of the Common Laws of England*, but what he is intending to do with his story of the contract is to set up a model for the establishment of a state. This, on his account, is what is involved in explaining something or giving its causes. Establishing the cause of something, on Hobbes's account, is not the searching after particular matters of fact in the situation that we might go into in investigating a traffic accident. He gives an example to explain his method in chapter 1 of *De Corpore*. A figure placed before us might look like a circle, but mere observation cannot assure us that it is one. Knowledge of its cause assures us of that.

> For let it be known that the figure was made by the circumduction of a body whereof one end remained unmoved, and we may reason thus; a body carried about, retaining always the same length, applies itself first to one *radius*, then to another, to a third, a fourth, and successively to all; and, therefore, the same length, from the same point, toucheth the circumference in every part thereof, which is as much as to say, as all the *radii* are equal. We know, therefore, that from such generation proceeds a figure, from whose one middle point all the extreme points are reached unto by equal *radii*. And in like manner, by knowing first what figure is set before us, we may come by ratiocination to some generation of the same, *though not perhaps that by which it was made, yet that by which it might have been made;* [italics added] for he that knows that a circle has the property above declared, will easily know whether a body carried about, as is said, will generate a circle or no. (Hobbes 1839, Vol. 1, p. 6)

The last sentence here makes clear that Hobbes does not seek causes as we usually do at the beginning of the twenty-first century, but seeks a model that will explain

what it is to be a circle, or a civil society. He is not committed to the idea that states actually do begin from contract, and he certainly recognizes that many begin with, or are expanded by, conquest. A civil society, then, will be something that has the features of an institution established by contract. It will involve obligations on people, and those obligations will depend on reciprocity. One of the distinguishing features of Hobbes's account is that the reciprocity would not primarily be between sovereign and citizen, but between citizens (as it is in such present-day theories as Rawls 1972); the sovereign, importantly, is outside the contract. Were he not outside it, then somebody else would have to hold the authority to determine whether the sovereign was behaving properly, and that person would then have the ultimate authority and be the actual sovereign. This is the absolutism of the Hobbesian sovereign to which so many have objected. Hobbes, both in not depending on an actual contract and in making the mutual obligations between citizens rather than between citizen and sovereign, was moving beyond the other contract theorists of his time.

Why would the natural condition, in which people have no common sovereign, be the unpleasant state that Hobbes describes? He gives several significantly different examples in discussing the natural condition, American Indians and the relationships between sovereign states amongst them, which can be confusing. The core of the notion is that people are in their natural condition if they have no common sovereign, so that somebody might be in his natural condition with respect to one person but not with another. It is not that the whole world must be in either the natural condition or civil society or that an individual person must be either in the natural condition or not. Within a family we might not be in our natural condition with respect to each other, but each of us might be in the natural condition with respect to people outside our own family.

Why should people who lack a common sovereign be unpleasant to each other? One thesis is that it is because they are self-interested and will therefore naturally clash in the absence of a coercive power to prevent them. This is the interpretation that has led to games theorists taking up Hobbes (e.g., Gauthier 1969, Hampton 1986, Kavka 1986) and to his being taken as providing a 'realist' model for international relations (e.g., Bull 1980) with reasons of state and pursuit of national self-interest at the center of concern, that is, of international politics as power politics. This is, perhaps, a plausible picture of international politics. Along with this idea goes the interpretation of the laws of nature as policies to be followed out of self-interest, fitting with Hobbes's description of them as rules of prudence.

Bernard Gert (1972) argues strongly against the claim that Hobbes is a psychological egoist who believes that people always seek their own narrow interest. Certainly it seems clear that Hobbes did not believe that people could do no other than try to preserve themselves when their lives were threatened: not only did he comment on those who give their lives too cheaply in duels because of a confused notion of honor, but he admired some of those who gave their lives in defence of the Royalist cause in the civil wars and made laudatory remarks about one of them in dedicating *Leviathan* to his brother. This position, then, requires a different account of the laws of nature.

Hobbes was clear that the laws of nature are effective only in conditions of reciprocity and preferably conditions in which there is a sovereign to guarantee reciprocity. This at least looks as though it supports the idea that the only reason for obeying them is self-interest and particularly fear of the sovereign, but drawing that conclusion means ignoring the fact that Hobbes treats prior performance by one party to a contract as making an obligation effective on the other party just as much as does having a sovereign who will force one to comply. The laws of nature, he says, obligate *in foro interno*, in conscience, in the natural condition, that is, they bind in the intention, but they bind *in foro externo*, in the act, only in civil society. Virtues are more complex than they are sometimes treated as being: Hobbes's point is that benevolence, for example, is a good thing because of the role it actually plays in civil society. Handing a knife to somebody who wants to kill me is not benevolent; it is simply stupid. Handing over money to a grocer who refuses to provide me with the goods I want is not required by justice. The virtues actually require a setting of relatively peaceful and secure life if they are to be possible to exercise, and Hobbes's stronger thesis that they can be exercised only by people who are subject to a common sovereign obviously bodes ill for a morality of international relations and explains why Hobbes has been taken up by some proponents of world government. (On this interpretation of Hobbes, see, for example, Bull 1980 and Ewin 1991.) On this account, the story Hobbes tells of people agreeing on principles of peace is a model of the sort he describes in the account of the circle, and the model is for virtues. Certainly, for all his objections to Aristotle, his objections to what Aristotle said about virtues is that he identified them as means, and not in terms of their being, as Hobbes put it, 'not properly Lawes, but qualities that dispose men to peace, and to obedience' (Chap. 26 of *Leviathan*). That is, he suggests agreement with Aristotle's general position about morality, but disagrees with the principles by which virtues are to be identified. And this fits, not with games theory, but with the accounts of ethics that now seem to be emerging from evolutionary theory about the characteristics that make our social species capable of surviving (see, e.g., Alexander 1987).

If people have these virtues as part of human nature, even if they do not have them perfectly, why would the

natural condition of mankind be as horrible as Hobbes suggests? Because, in the absence of a sovereign, the virtues cannot be operable; the laws of nature will not bind *in foro externo*, that is, to any overt action. We are all equal in the natural condition not only in that we are all vulnerable to sudden death, but in that there is no particular reason why any of us should subject his or her reason to that of another. If we differ and neither of us can persuade the other, then there is no reason why either should submit. And in this world we do differ: we see the world in different ways from different backgrounds, are struck by different features of the situation, and we calculate differently. We have differences about the nature of justice, and we differ about who did what to whom. When we differ about something like justice, there is no reason why one should submit to the view of another and do what he simply regards as unjust. It is differences of judgment, rather than clashing interests, that the sovereign must deal with, providing an artificial reason to replace the right reason that fails to do the job. Those who insist on right reason in a dispute, Hobbes says, are simply insisting on being judges in their own case, for each man takes for right reason what is his own.

His first argument for the sovereign in *Leviathan* appears early in Chap. 5 in the discussion of disagreement and the need to establish an arbitrator or judge to determine disputes (though the arbitrator or judge might be a meeting of the populace voting on the matter or any number of other things, and is best thought of as a procedure rather than as a person), an arbitrator whose judgment we agree to accept before we know what it is, thus establishing rights. Without those rights the virtues will not be operable, and without the virtues, at least to a degree, the rights could not be established. The recognition of authority, on Hobbes's account, is something without which humanity could not survive. Our moral views have the status of private views in which we can clash, but people have a capacity for politics that makes it possible for them to establish a public morality (or law) of rights that determines the sphere within which each person's private judgment can operate.

This shows some of Hobbes's relationship to modern liberalism. And because people do have failings, and because we need security even though we know that there are dangerous people whom we could not recognize at first sight, the sovereign must have coercive powers including the powers to tax people to cover the costs of his activities; there will be clashes of interest as well as clashes of judgment. The sovereign must be absolute in that his function is to resolve disputes, and he cannot do that if his authority depends on his agreeing with both parties. But that, Hobbes suggests, is a logical truth about any system of government and is compatible with our choosing government even by a meeting of all the citizens; it does not require subjection to the arbitrary will of one person, but does require subjection to the decisions of the community however they are determined. This brings out some of Hobbes's relationship to modern democratic theory. The sovereign makes it possible to resolve disputes peacefully and be sure that the resolution will be enforced, so he makes it possible for us to act as one; he determines for us the will of the community, and thus makes us one person. This Hobbesian idea of sovereignty as ultimate and absolute remains influential.

See also: Altruism and Self-interest; Democratic Theory; Liberalism: Historical Aspects; Sovereignty: Political

Bibliography

Alexander R 1987 *The Biology of Moral Systems*. Aldine de Gruyter, Hawthorne, NY
Aubrey J 1898 *Brief Lives*. Clarendon Press, Oxford, UK
Boonin-Vail D 1994 *Thomas Hobbes and the Science of Virtue*. Cambridge University Press, Cambridge, UK
Bull H 1980 *The Anarchical Society*. Macmillan, London
Ewin R 1991 *Virtues and Rights*. Westview, Boulder, CO
Gauthier D 1969 *The Logic of Leviathan*. Clarendon Press, Oxford, UK
Gert B (ed. and intro.) 1972 *Man and Citizen*. Harvester Press, Brighton, UK
Hampton J 1986 *Hobbes and the Social Contract Tradition*. Cambridge University Press, Cambridge, UK
Hobbes T 1839 *English Works* (ed. Molesworth W). John Bohn, London
Hobbes T 1961 *Opera Latina*. Scientia, Aalen, Germany
Hobbes T 1968 *Leviathan* (ed. Macpherson C B). Penguin Books, Harmondsworth, UK
Hobbes T 1969 *The Elements of Law* (ed. Tonnies F.). Cass, London
Hobbes T 1998 *On the Citizen* (eds. Tuck R., Silverthorne M.). Cambridge University Press, Cambridge, UK
Johnston D 1986 *The Rhetoric of Leviathan*. Princeton University Press, Princeton, NJ
Kavka G 1986 *Hobbesian Moral and Political Theory*. Princeton University Press, Princeton, NJ
Rawls J 1972 *A Theory of Justice*. Clarendon Press, Oxford, UK
Rogow A 1986 *Thomas Hobbes: Radical in the Service of Reaction*. W. W. Norton, New York
Shapin S, Schaffer S 1985 *Leviathan and the Air Pump*. Princeton University Press, Princeton, NJ
Skinner Q 1964 Hobbes's. *Leviathan*. *The Historical Journal* **7**: 321–3
Skinner Q 1965 History and ideology in the English Revolution. *The Historical Journal* **8**: 151–78
Skinner Q 1966 The ideological context of Hobbes's political thought. *The Historical Journal* **9**: 286–317
Skinner Q 1972 Conquest and consent. In: Aylmer G E (ed.) *The Interregnum*, Macmillan, London
Tuck R 1989 *Hobbes*. Oxford University Press, Oxford, UK
Warrender 1979 Political theory and historiography. *The Historical Journal*, pp. 931–40

R. Ewin

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Holocaust, The

'The Holocaust' is the term most widely used in reference to the systematic extermination of the Jews of Europe by the Nazis during World War II. Usually, the term also alludes to the historical background of the extermination and to related contemporary circumstances. The first section of this entry will deal with some aspects of the historical background, the second will briefly present the immediate context of the 'Final Solution to the Jewish Question,' while the third and central part will address the extermination *per se* and the reactions to it. This chronological presentation does not imply an unavoidable passage from one phase to the next. However, while the link between the evolution of German history during the first two decades of the twentieth century and the onset of the Nazi regime was in no way evident, the causal relation between the anti-Jewish policies of the early war period and the extermination was much closer.

1. *The Background*

The historical origins of the Holocaust remain an object of ongoing debate. Whereas some scholars consider the extermination of the Jews as the consequence of a specific evolution of German history, at least from the end of the nineteenth century onwards, others interpret the Holocaust as an outgrowth of modernity, the implementation of a pitiless instrumental rationality at the service of a totalitarian racial utopia. Both approaches are not incompatible.

The rise of racial anti-Semitism, the spreading of social Darwinism, and the publication of the first theories of racial anthropology and eugenics turned the late nineteenth century into a breeding ground of ideological trends that were to become central tenets of the Nazi worldview. And, although political anti-Semitism declined in the Wilhelminian Reich at the beginning of the 1900s, an increasingly vocal anti-Jewish hostility spread throughout wide sectors of German society. In this evolution, the outbreak of World War I became a major turning point.

After a brief period of national unity, the unforeseen prolongation of the conflict and the growing political polarization which it induced triggered a radicalization of anti-Jewish attitudes, mainly in the nationalist-conservative camp. The defeat of Germany, the revolutionary upheavals that followed, and the birth of a republic hated and rejected by an important sector of the population, as well as the vast majority of the traditional elites, fanned the already virulent anti-Semitism even further. The Jews were accused of bearing the responsibility for the national humiliation (the 'stab in the back' myth), of controlling the new republic, of instigating bolshevism and, above all, of conspiring to dominate the world, as described in the most widely read anti-Semitic forgery of the century, *The Protocols of the Elders of Zion* (1919).

It was in this postwar atmosphere that Adolf Hitler began his political career by joining the tiny German Workers' Party (DAP) in Munich, in late 1919. In 1920, the DAP became the National Socialist German Workers' Party (NSDAP) with anti-Semitism figuring prominently in its twenty-five point program. Notwithstanding the failed putsch of 1923 and Hitler's brief incarceration, the resumption of his political agitation in 1925 along with systematic organization, unrelenting propaganda, and, more than anything else, the impact of the Great Depression, propelled the party into prominence on the national scene.

Anti-Semitism lay at the center of Hitler's worldview from the outset of his political career and remained so after his accession to power on January 30, 1933. Reduced to its core, this anti-Semitism viewed historical evolution as determined by a merciless racial struggle dominated by the confrontation between Aryan and Jew. The outcome of this struggle was to decide the fate of mankind and particularly of the Aryan race. Hitler's faith had an apocalyptic dimension, presented in terms of 'scientific' theories. But how far this worldview concretely led to the systematic planning of the anti-Semitic policies of the regime, or, obversely, to what extent these measures were the result of uncoordinated initiatives stemming from a growing radicalization of the party and its agencies, remains an open question. One could argue that the ideological guidelines indicated a general framework for action, but that the 'twisted road' followed by the implementation of the anti-Jewish policies was dictated by changing circumstances. In any case, during the 1930s, no plan for a physical extermination of the Jews existed; the ever-more extreme anti-Jewish policies of the prewar years had two successive and identifiable aims: the exclusion of the Jews from the national racial community (the *Volksgemeinschaft*) and, thereafter, their expulsion from the Reich.

From the outset of the new regime the Jews were dismissed from the civil service, from a growing number of professions, and from all major fields of German cultural life. Moreover, all that belonged to or had been influenced by the 'Jewish spirit' had to be cleansed from the German cultural sphere: Freud's or Einstein's theories, Heine's poetry, Mendelssohn's music, books, works of art, memorial plaques, films, libretti—anything. A stream of anti-Jewish incitement was daily spewed out by all the propaganda channels of party and state, under the guidance of Propaganda Minister Joseph Goebbels and chief party ideologist, Alfred Rosenberg; it found its most rabid expression in *Der Stürmer*, the 'specialized' anti-Jewish organ of the Gauleiter of Franconia, Julius Streicher. The entire education system soon absorbed the anti-Semitic

message, as did the Hitler Youth (HJ) and the League of German Girls (BdM). The indoctrinated youngsters of the 1930s were to become the soldiers of the 1940s.

In September 1935, the Nuremberg Laws added strict biological separation to social and cultural segregation by forbidding marriage and sexual relations between Jews and Germans (the Law for the Protection of German Blood and Honor). The Reich Citizenship Law reduced the Jews to second-class citizens. Finally, the first 'Supplementary Decree' to the Reich Citizenship Law defined who was a Jew or, alternatively, a 'mixed breed' of the first or second degree (November 1935). Any transgression of the 'Blood Law' was considered as 'race defilement' (*Rassenschande*), and the offenders (usually the Jewish male offenders) were sent to jail or a concentration camp.

In 1936, the *Reichsführer SS*, Heinrich Himmler, was put in command of all German police forces. This reorganization of the control and terror apparatus led to the appointment of SS General Reinhard Heydrich, previously chief of the security service of the SS (*Sicherheitsdienst* or SD) to the command of both the SD and the newly established Security Police (*Sicherheitspolizei*), which included the Gestapo and the Criminal Police. After the outbreak of the war, these agencies were further centralized within the Reich Security Main Office (*Reichssicherheitshauptamt* or RSHA). While the planning of direct anti-Jewish terror was mainly Himmler's and Heydrich's preserve, the plundering of Jewish assets and enterprises was orchestrated by Hermann Goering's Four Year Plan Administration, and, after the dismissal of the conservative minister of the economy, Hjalmar Schacht (who had managed to protect some measure of Jewish economic life), by his Nazi successor, Walther Funk.

After the annexation of Austria, in March 1938, the drive to compel the Jews to leave the 'Greater German Reich' ('Old Reich' and Austria) became overwhelming. During that year, 'compulsory Aryanization' and, ultimately, the prohibition of any Jewish economic activity destroyed the material basis of Jewish existence in Germany. Notwithstanding the widespread reticence of most countries to take in Jewish refugees (easily identifiable since, at the demand of the Swiss government, their passports were marked with a 'J'), forced emigration and even expulsion over the borders became standard Nazi practice. Finally, on November 9 and 10, following the killing of a German diplomat by a Jewish youth in Paris, a nationwide eruption of state-organized violence, the so-called 'Night of the Broken Glass,' destroyed the remnants of Jewish life in Germany. The Jews of the Reich now desperately attempted to flee. The panic soon extended to the Jews of the 'Protectorate of Bohemia and Moravia,' after the German occupation of Prague in March 1939. Yet, it was the German attack on Poland, on September 1, 1939 and the outbreak of World War II that entirely changed the fate of European Jewry.

2. *The Immediate Context of the 'Final Solution' (1939–41)*

During the first two years of the war, the goal of Nazi anti-Jewish policies still appeared uncertain. Jewish emigration from the Reich and occupied Europe remained a possibility, although wartime conditions soon reduced it to a trickle. The goal frequently alluded to was a territorial concentration of Europe's Jews, first in the easternmost part of Poland, in the Lublin area of the newly established 'General Government' (the name used for the nonannexed area of Poland which was put under the authority of the Nazi 'Governor General,' Hans Frank), then, after the victory over France, in the French island colony of Madagascar, and, finally, after the attack on the USSR, in Arctic Russia or Siberia. None of these plans ever materialized, except for some short-lived attempts to deport Jews from Vienna and the Protectorate to the Lublin area, in the fall of 1939 and in early 1941.

A few months after the occupation of Poland, the Nazis started to deport hundreds of thousands of Jews and Poles from territories annexed to the Reich, such as the Warthegau (a part of western Poland), into the General Government; they drove the Jews into ghettos in larger towns and even dumped Jews from the Reich into Vichy France, in the fall of 1940. The deportations within ex-Poland may have been linked in part to the regime's policy of ingathering ethnic Germans from the Baltic countries, the Russian occupied territory of Poland, and the Balkans in the 'Great Germanic Reich.' Thus, the deportations of Jews and of Poles from the Warthegau were meant to make space for the incoming *Volksdeutsche*. The policy was pursued under the orders of Himmler, as newly appointed 'Reich Commissar for the Strengthening of Germandom' (RKF). It was meant to be the first stage of what became known as 'General Plan East' (*General Plan Ost*), aimed at establishing German domination in the East for centuries to come. The enslavement of entire nations and the death of tens of millions of local inhabitants was taken into account. This policy was probably not directly linked to the genesis of the 'Final Solution' since it cannot explain the deportations from the west (to be described later on).

Simultaneously, from October 1939, an entirely separate but not unrelated operation had started in the Reich: the systematic mass murder of the mentally ill, the so-called euthanasia or 'Aktion T-4': personnel were trained and killing methods were developed that were to be used for the mass extermination of Jews after the scaling down of T-4, in the summer of 1941, to a less visible 'euthanasia' operation (Aktion 14f13).

In Western Europe, the much smaller Jewish populations were not secluded (except for the establishment of a semi-ghetto in Amsterdam) or as brutally treated as the Jews of Poland. But their exclusion from public service and the liberal professions, growing curtailment of their economic acti-

vity, the registration of persons and property, limitations on travel, and police surveillance, turned them into pariahs. Their persecutors were not only the Germans, but also satellite regimes such as Vichy France.

After the German occupation of Yugoslavia and Greece in April 1941, Serbia remained under direct military control and Croatia became a satellite; in both countries, harsh anti-Jewish measures were introduced. Slovakia's anti-Jewish policy followed strict German orders. Romania implemented fierce anti-Semitic measures of its own, and only Hungary and Bulgaria maintained semi-independent policies in this domain.

In order to facilitate their control over the Jews—beyond territorial concentration—as well as to degrade them, the Nazis compelled all those over the age of six to display a 'Jewish star' on their clothing, first in Poland, then, from September 1941 on, in the Reich and the Protectorate, and, in the spring of 1942, in all countries of occupied Western Europe. But the principal instrument of Nazi domination was the 'Jewish Councils.' These councils, composed of representatives of the Jewish communities, received German orders, conveyed them to the Jewish populations, and were responsible, upon pain of death, for their implementation. Moreover, in Eastern Europe the Jewish Councils became agencies of a Jewish self-administration of sorts, struggling against impossible odds; justly or unjustly, they often became targets of sarcasm, resentment, and even hatred in their own communities. Some historians have adopted a highly critical stance toward the councils, considering them as instruments in the hands of the Nazis for the destruction of their own people. Others have questioned this overall characterization as the differences in the behavior of the councils were considerable. All in all, the councils' members were ordinary people confronted with entirely unprecedented circumstances: their own reasoning was incompatible with the murderous strategy of the Nazis.

3. *The Extermination*

Systematic mass extermination of Jewish populations began with the German attack on the USSR, on June 22, 1941. Whereas in most newly occupied Soviet territory mainly Jewish men were executed until August 1941, in the Baltic countries and in Transnistria (where the Romanian army was the principal ally of Germany), entire Jewish communities were butchered from the outset. The killing of Jews in party or state positions had been ordered by Heydrich on the eve of the campaign; the perpetration of pogroms by local inhabitants was also explicitly fostered by the Germans. Whether the killing that engulfed the entire Jewish population from August 1941 onwards was independent of any instrumental goals, or whether, as with the starving to death of two million Russian POWs before the spring of 1942 (around 3.3 million died before the end of the war), it derived from difficulties experienced by the Wehrmacht in securing sufficient food supplies for the army is not clear. In any case, the mass murder of Jews on Soviet territory (around 500,000 victims by the end of 1941) was not perpetrated solely by the 'intervention units' of the Security Police and the SD (*Einsatzgruppen*), by local auxiliaries, and by the Romanian army in the south, but also by tens of Reserve Police Battalions and regular Wehrmacht units. In other words, not only the 'troops of the ideological war' (the SS) but ordinary men, possibly ten of thousands of them, were involved, often willingly so, in the mass murder of the Jews.

Few historians would now argue that the decision to exterminate all the Jews of Europe predated the attack on the USSR. Some contend that this decision was taken by Hitler soon after the German onslaught, during the period of 'euphoria,' due to the apparent success of the invasion, and that it found its expression in Goering's letter to Heydrich of July 31, 1941, ordering him 'to take all necessary measures to prepare the Final Solution of the Jewish Question in Europe.' Others argue that such an order was issued in the early fall, when doubts may have risen in Hitler's mind about the prospects of a quick victory in the East. Others still link the order to the USA's entry into the war, in other words to the transformation of the European war into the 'world war,' which, as Hitler had 'prophesied' in his Reichstag speech of January 30, 1939, would see 'the extermination of the Jewish race in Europe.' This last interpretation seems bolstered by documentary evidence. On December 18, according to Himmler's appointments diary, the SS chief met with Hitler and wrote down: 'Jewish question/eradicate as partisans.' The reference to 'partisans' is clearly a way of describing an enemy lurking behind the lines, an internal foe, who, just as in 1917/18, would stab Germany in the back if given the chance. Finally, some historians doubt the existence of any clear order given at the highest level and consider total extermination as resulting from the cumulative effects of local initiatives.

By October 1941, the killings had extended to Jews from the Greater Reich and the Protectorate who were deported mainly to Lodz, Riga, Kovno, and Minsk. In Riga and Kovno, thousands were shot on arrival. The others were exterminated later with the inhabitants of their ghettos. They had also spread to Serbia, where the Wehrmacht executed some 6,000 Jews in reprisal for partisan attacks. During the same month, Jews were forbidden to leave the continent, 'in view of the coming final solution of the Jewish Question' and the construction of the Belzec extermination camp (possibly that of Sobibor too) began in the General Government.

In early December, the first extermination site became operational at Chelmno, near Lodz, principally for the murder of Jews from the Warthegau. On January 20, 1942, in a meeting in Berlin (*am Grossen Wannsee*), Heydrich informed an array of state secretaries, senior SS officers and other selected officials of the general extermination program that, according to his computations, was to apply to 11 million European Jews. From then on total extermination became a problem of logistics and of some measure of haggling between the Germans and various governments or local authorities about handing over their Jews. In almost all cases, foreign Jews were arrested and delivered first, local Jews followed (as will be indicated further on, in some rare cases Jews were not handed over at all).

The arrests of Jews, mainly in Western Europe, and the logistics of the deportations were coordinated between the Higher SS and Police Leaders (HSSPF), Himmler's delegates to each occupied country or major area, local authorities and police forces, and the Jewish Affairs section at the RSHA (IVB4) under Adolf Eichmann. Furthermore, the deportations demanded joint planning with the Reich Railways Authority (*Reichsbahn*) and its territorial subdivisions (*Ostbahn*, etc.). Once the Jews reached the camps, they passed under the command of the SS Economic and Administrative Main Office (*SS Wirtschafts- und Verwaltungshauptamt*) under SS General Oswald Pohl (except for the *Aktion Reinhard* camps). In the camps, the Jews were destined for extermination; however, mainly in Auschwitz, selections separated the healthiest and the strongest from the mass of the inmates, submitting them to slave labor before sending them to their death.

The first stage of the extermination chiefly included the Jews of the Warthegau and of the General Government who, through 1942, were killed in the *Aktion Reinhard* camps (Belzec, Sobibor, Treblinka, and Majdanek), as well as in Chelmno. The deportation of the Jews of Warsaw to Treblinka began in July 1942. Simultaneously, deportations from Western Europe (France, Belgium, and Holland) started. By then, the principal Auschwitz extermination camp (Auschwitz II or Birkenau) had become fully operational. The killing method shifted increasingly from the use of carbon monoxide gas to that of Zyklon B, a powerful pesticide utilized in the fumigation of army barracks and ships' holds. Before they were killed, the victims' valuables and clothes were taken away and their hair was shorn; after the gasings, gold teeth, and artificial limbs were removed from the corpses. The German economy recycled the 'booty': gold was smelted and delivered to the *Reichsbank*; hair was used for mattresses, slippers, etc.; clothes were distributed by the *Winterhilfe* (the German winter-relief organization). Homes, furniture, and other belongings left behind by the deportees, were distributed to deserving *Volksgenossen* or acquired, usually at dirt-cheap prices. This allocation of spoils spread to occupied countries in the West. Particularly valuable objects, mainly art, looted all over Europe by the *Einsatzstab Reichsleiter Rosenberg* or ERR, were offered to Nazi leaders (Hitler and Goering headed the list), to German museums, and various other German institutions.

In 1943, the last batch of German Jews was deported to the extermination camps; the deportations from Western Europe continued, and, simultaneously, the Jews of the Balkans and, by the end of the year, of Italy (following the German occupation of the greater part of the peninsula) were included in the killing process. From March 1944 onward, after the occupation of Hungary by German troops, this country's Jews became the last large national group of victims (over 400,000) to be sent to Auschwitz–Birkenau. On November 2, 1944, Himmler ordered the cessation of the mass exterminations; his order was frequently disregarded, even after the liberation of a growing number of camps by the Allies. 'Death marches,' organized by camp commanders or SS personnel, went on throughout the winter of 1944–5 and practically until the German surrender in May 1945. By the time the war in Europe drew to a close, approximately 5.5 million Jews had been exterminated.

The exterminations on Soviet territory had become widely known by the fall of 1941. Information about the mass killing of European Jewry beyond the Soviet border spread throughout the continent, and in Allied and neutral countries from the second half of 1942 onward, although the technical details of the 'Final Solution' may not have been widely known until 1944. The overall reaction was one of indifference and inaction, sometimes of outright collaboration. While Vichy authorities and police collaborated with the Germans in the arrests and deportations, Switzerland hermetically closed its borders to the Jews who attempted to flee. In general terms, the Christian churches intervened feebly, mainly on behalf of converted Jews. Thousands of Jewish children were hidden in Christian institutions; in most cases, their lot was conversion. The silence of Pope Pius XII, notwithstanding some help from the Vatican in hiding Jews in the fall of 1943, has, to this day, remained a matter of heated debate.

In some cases, the reactions were different. Spain and Portugal allowed the transit of Jews through their territory. Sweden attempted to take in the Jews of Norway when the deportations started and it took in almost all the Jews of Denmark (most of whom were transported to safety by the Danes themselves). Protests against the anti-Jewish measures were staged in Holland, in early 1941; the Italians tried to protect the Jews in their own country as well as those under their control. Bulgaria refused to deliver its own Jewish citizens, although it did deliver the foreign Jews. Finland rejected all German demands regarding its small Jewish population.

'The abandonment of the Jews' by the Western Allies is largely undisputed nowadays. Rejected or fatally delayed rescue plans involved transfers of ransom money to the Germans or their allies, or the bombing, in 1944, of the Auschwitz gas chambers and of the railway line connecting the camp with Hungary. Such refusal to help, whether known or merely sensed, could but contribute to the feeling of total isolation of the Jews of Europe in the face of the German extermination process.

The concrete possibilities of Jewish resistance were almost nil. Moreover, by the time most European resistance movements became active, the bulk of European Jewry had been exterminated. Hence Jewish armed revolt in the ghettos (the best-known instance being the Warsaw ghetto revolt of April 1943) or even in the camps, although hopeless, carried the utmost symbolic significance. As for the immense majority of European Jews, they had practically no options even when impending murder was obvious: flight and hiding were limited options; despair and numbness, overwhelming.

See also: Anti-Semitism; Collective Memory, Psychology of; Contemporary History; Dictatorship in History; Ethnic Cleansing, History of; Genocide: Historical Aspects; Historiography and Historical Thought: Jewish Tradition; Judaism and Gender; National Socialism and Fascism; Racism, History of; Second World War, The; Violence, History of; War Crimes Tribunals

Bibliography

Aly G 1995 *Endlösung*. S. Fisher, Frankfurt am Main, Germany

Bankier D 1992 *The Germans and the Final Solution*. Oxford University Press, Oxford, UK

Bauer Y 1994 *Jews for Sale?* Yale University Press, New Haven, CT

Browning C R 1992 *Ordinary Men*, 1st edn. HarperCollins, New York

Burrin P 1994 *Hitler and the Jews*. Edward Arnold, New York

Friedländer S 1997 *Nazi Germany and the Jews*, 1st edn. HarperCollins, New York

Gerlach C 1997 *Kalkulierte Morde*. Hamburg, Germany

Gutman Y 1982 *The Jews of Warsaw 1939–1943*. Indiana University Press, Indiana, IN

Herbert U (ed.) 1998 *Nationalsozialistische Vernichtungspolitik*. Frankfurt Mainam, Germany

Heer H, Naumann K (eds.) 1995 *Vernichtungskrieg*. Hamburg, Germany

Hilberg R 1993 *Die Vernichtung der Europäischen Juden*, 3 Vols. Fischer, Frankfurt, Germany

Longerich P 1998 *Politik der Vernichtung*. Piper, Munich, Germany

Marrus M R, Paxton R O 1981 *Vichy France and the Jews*. Basic Books, New York

Mommsen H 1991 The Realization of the unthinkable. In: Mommsen (ed.) *From Weimar to Auschwitz*. Princeton University Press, Princeton, NJ

Trunk I 1972 *Judenrat*. Macmillan, New York

Wyman D S 1984 *The Abandonment of the Jews*, 1st edn. Pantheon Books, New York

S. Friedländer

Home in Anthropology

1. Definition and Meaning of 'Home'

The term 'home,' being of Anglo-Saxon and Germanic origins, often carries rational and irrational connotations at the same time: house, landscape, habitat vs. place of belonging, object of nostalgia, place of origin. The German equivalents, *Heim* and *Heimat*, originally meant 'house' and 'owned land,' but later on denoted ideal and material spaces offering homeliness or German *Gemütlichkeit*. Not coincidentally, the term *Heimat* was re-introduced into the German written language through the reception of British gothic ideas and romanticism. Moreover, *Heimat* was adapted by many other languages as a lean word signifying a mixture of connotations all tending towards an emotionalized interpretation of the place a person feels adapted to: 'my home is my castle.' While loaded with emotional charge, the term 'home' usually also carries allusions to very profane and unromantic connotations, such as in the euphemistic names given to institutions catering for the poor, old, and disabled, or in the meanings expressed by the untranslatable German term *unheimlich* (the 'uncanny' of Freudian philosophy).

Anthropologically relevant references to 'home' initially appeared during the nineteenth century in Anglo-Saxon and German romantic and gothic styles of writing in which it denoted *Ur-Heimaten* or certain groups of humankind, such as the 'Aryan' origins in Asia or people's origins in other 'cradles' (Kuper 1988). Later on the term 'home' acquired a more specific political meaning as in the case of the 'homelands' assigned by the South African apartheid regime to ethnicized groups of 'black' citizens, which had the purpose of containing them in reservations or similar constructions of 'home rule.' More recently, the term has gained some significance again in anthropological studies of migrant, diaspora, and fugitive populations (Malkki 1995, Vertovec and Cohen 1999).

The related discussions move between two poles, one pole being represented by the 'invention of

tradition' approach (Hobsbawm and Ranger 1982), and the other pole tending towards sociobiological and ethological claims about the existence of a human instinct which orientates us towards distinct 'environs' or natural habitats (Wilson 1975). From an anthropological perspective, the often rather unreflected and untheorized use of the term 'home' in scientific texts offers a good opportunity for checking the process of translation between European and non-European cultures. For the discussion of universalistic assumptions about a human drive for being 'at home' involves very heterogeneous forms of belonging, alienation, or indifference towards a 'home.' Often the 'drive' to be at home seems quite easily contented with illusions of a spiritual home; yet in some cases historical processes lead to violent conflicts about presumed 'home countries' between certain groups.

Mary Douglas (1966) introduced anthropologists to a world of symbolic attributions, in which the body can signify everything, and everything can mean the body. The same applies to homes, when looked at as a rational construction of territorial belonging on the one hand, and as an image of embeddedness in a particular social and geographical landscape, or nostalgia about such a situation, on the other.

2. *Ways of Being 'At Home'*

Obviously, the portable shelter constructed by a group of Kung hunters and gatherers is not the same as the proverbial British or German 'castle.' Anthropologists observed, however, that so-called bush people in Africa, traditional Australian aborigines, as well as groups inhabiting the rainforests of South Asia and South America, have developed elaborate systems of navigation in their regions, based for example on landmarks or tactile qualities of the ground (Widlok 1997). In the nineteenth century, Australian aborigines imagined their ancestors to wander around and with their actions invent the current topographical shape of their habitat. In so doing they related to the 'ground' in a way not at all alien to European conceptions of identity defined by geographical, social, genealogical, and individual physical characteristics of a given group in a given area (Spencer and Gillen 1938). Attracted by some of the benefits of provincial town life or driven away by colonial powers from their places, their twentieth-century descendants may insist on returning to their promised lands, if their means of living allow such a step (Berndt 1977). Welfare programs can thus paradoxically lead to a partial return to traditional ways of living and especially of nurturing their ancestors' grounds. However, often we are no longer able to discern whether indigenous nativists adhere to old projections of a social and spiritual life onto certain 'hunting grounds' and places of survival through gathering activities, or whether this just follows Western concepts of belonging, grounded and capitalizing on geographical features of a given area.

There is varied ethnographic and mythological evidence of responses to loss of old nurturing grounds, including the wish to re-enchant the relationship between humans, heaven, and nature, although so far there is not enough comparative research available on the symbolism of land and house. Yet there is some evidence that human groups that succeed in storing greater amounts of food develop more closely-knit local mythologies bound up with the topography of landmarks, human dwellings, places for growing and harvesting food, as well as holy places. Horticulturists and cattle-breeding populations living in segmentary societies or low-grade forms of centralization tend to organize the household, the surrounding area, and far-away places that are of interest to the group by socially constructing ancestors and territorial spirits or divinities (Lienhardt 1961, Lévi-Strauss 1955). The 'juridic domain' (Fortes 1969) concerning the simple facts of life—dwelling together, dividing up food, working a territory, keeping a herd, and maintaining property rights—is thus organized by an explicit ideology of belonging, descent, holy grounds, and the profane (Kramer and Kramer-Marx 1993). A supernatural interpretation of shrines and natural features, even of artificially organized biotopes, can be the territorial expression of a temporal cult of vegetation which unites social groups in a harmonious view of the ego, the household, the family, and other related persons with its ground, including all the plants and animals living or to be raised therein (Turner 1969).

A 'system of total exchange' (Mauss 1960), of generalized reciprocity, seems to pervade and bind human groups and their natural surroundings together. Paradoxically, this can privilege those groups which have reasons to claim to have been the first humans in a particular region and can give them a superior status, heavily defended in conflict, or lost and exchanged with the status of a pariah. The conflict between Hutu and Tutsi in Central Africa has had such a component (Maquet 1961). In a time of international arms traffic, and armed interventions into local cultures, tensions of this kind can lead to severe cases of mass violence.

There is a definite difference between such closely knit groups or networks of groups and the nation states of world society, insofar as the 'total social fact' (Mauss 1960) of traditional societies is organized into a tight overlapping of the domains of landscape, social structure, cosmology, etc. (Fortes 1969). The father or uncle can be the earthmaster, celebrating a holy shrine, and the next moment he can be the leader of a group of young warriors. This does not mean, though, that the social reality of these groups is more homogeneous than ours. Whilst subscribing, like their fellows, to a certain degree to the governing harmonious ideology of space, time, and home, some outsiders, skeptics,

and dissidents always disagree and hold rational or disenchanted views of their group's attempts to maintain a homeland (Radin 1957).

By exchanging 'stigma' for 'charisma,' such outsiders can strongly affect or change their society. Thus, the classical 'big men' of New Guinea who constructed the 'pig festival' through the vegetational and reproductive cycles draw on a whole system of credits and cliental relations countering the heroic culture of the 'great men' who search for the spiritual gratification of the warrior (Godelier 1982). Early imperialist states seem to have been based on a similar mixture of exploiting nature, fellow, and other groups, on the one hand, and heroic fantasies about extending a holy landscape to other regions, on the other (Wolf 1999).

The study of nomadism might be particularly instructive with respect to human ideas about finding or having a home (Mauss 1960). For the layperson, nomads just seem to wander around without clearly relating to certain features of the landscapes. The experience of wandering and fugitive groups in Europe, such as the Sinti, Roma, and many others, may support this conception of a 'wandering' part of humanity that does not feel bound to certain cycles of exchange. Wanderers are always suspect to sedentary populations, and this resentment can remain through long periods of sedentarization, if it is supported by political powers. One of the starting points of German anthropology in the nineteenth century was a series of heated discussions about whether or not Jews felt any obligation of fairness and civil behavior towards non-Hebrew populations (Gilman 1990).

In the anthropological data, nomadic constructions of a unified horizon of landscapes abound, as does information on nomads treating their tents, carriages, or cars as 'mobile homes' (Rao 1987). Motifs in Oriental carpets, manufactured by nomads, often appear to carry the signum of descent and territorial rights as well as some kind of report of the group's history and historical claims for grounds and herds (Schuster and Carpenter 1996).

Not only nomadic populations but many other groups convey an ideal of homogeneity, tradition, and *mise en scène* of an environment of 'being at home' through their material culture, such as dress and costumes. This continues right up to European medieval, Ottoman, and today's international regulations of clothing and corporeal signs. Also the often painfully achieved imprints on the skins of members of more marginal populations seem to have the same rational and spiritual aim of demonstrating a form of belonging together (Clastres 1973). It is the invention of an 'environment,' in the sense that this term is used in modern art. Human environmental creativity ascribes to a certain space the property of being a home. These autopoietic processes of representing and creating an order of time, space, and body are perhaps most obvious in rural societies with their regional and class-bound costumes, their festivals, shrines, and clear-cut regional polities of power.

But even in the case of feudal societies, which often enough did not allow their underclass members to leave their 'homes'—even in this continuum of slavery on the one hand and enthusiasm for one's own 'ground' on the other—we can find distinct groups who, through historical experience or the subtleties of property rights, seem to have acquired a very distanciated attitude towards their houses, fields, and cattle (Lopreato 1967). In times of famine or collective violence, these groups were probably among the first to leave their grounds and wander to the cities or the promised lands, over the mountains and across seas, as soon as the opportunity was offered by the modernization of the world economy into the global capitalistic project.

3. *Modern and Postmodern Homes*

In cities and other modernized landscapes, anthropologists have observed distinct techniques deployed in designing the 'home' of migrated groups or former peasant inhabitants of the area (Hannerz 1980). Conscious city design, the building of ghettos for reasons of controlling the 'bio-power,' and the process of identity formation in subgroups of a given population all go hand in hand here. Homeland associations as well as informal networks amongst industrial workers and employees ascribe to their private and sometimes also their professional life characteristics of tribal life. The elite of industrialized areas tend to form closely-knit networks, sometimes secret, sometimes formally recognizable, which transform their actors into something reminiscent of the earthmasters and segmentary war chiefs as encountered in traditional societies.

In the third millennium, two-thirds of the world population are expected to live such a metropolitan way of life, and from what we know about the older metropolitan areas in the world, we can deduce that the postmodern metropolis will be partly made by neighborhoods defining and defending themselves like traditional 'villages in the city,' like those in Naples, Cairo, Rio de Janeiro, or New York. These 'lieux' and 'milieux' of trust and similarity will be defined by criteria of class and taste as well as ethnicity and religious confession. Homelessness will abound as well as those of the better off desperate to (re)create a home as an anchor to a very mobile style of life (Dehavenon 1999). The inhabitants of big cities try to overcome the anomic atmosphere of 'non-lieux' by staging and topographically creating city life through festivals and customs, charity and sport events, as well as a home life through decoration, invitations of friends, meals, and so on (Augé 1992, Cieraad 1999).

In addition to the withdrawals and the small refuges of identity, however, an extension of the word 'home' has become important. With the advent of the nation states, millions and even billions of humans can unite in very large conglomerations of landscapes, and by means of national symbolism they try to find not only a common political but also symbolic denominator. An organicist idea of the state—'a sound family makes a sound state' (Norman 1991)—may help to transfer feelings which were originally connected with house, family, regional polity, and local religion to the higher spheres of a fatherland united in the sign of a mother language, a home country (Gellner 1998).

This is where ethnicizing or nationalizing politics have often overlapped with anthropological science, a science often intended historically as a kind of school for learning how to invent nations. Not only under Hitler and Stalin did the 'nations of anthropology' (Borneman 1995) stand in close conjunction with their host nations, so did American cultural anthropology and other well-respected disciplines.

Today's 'United Nations' comprise groups that are divided by very different models of inventing traditions as well as interior constructions. Some of them are founded on the basis of the one and only territory alone, others claim that a common language, cultural heritage, race, or historical traumatization provide the fundament of their existence. However, the whole system of nations nowadays demands that no single person, house, or territory can be without a host nation. This is supported by older ethnographic constructions, which always attributed a group of people to a certain territory. Today, such origins of ethnic topographies frequently turn into a topography of terror, while at the same time international movements of refugees and migration make such attributions obsolete. Another form of homebuilding, however, will apparently continue to exist even in the most technologically advanced systems on Earth. Against all the prognostics of the social sciences, the 'home' seems to survive even its most important carrier, the family. Statistical evidence indicates that single households abound in all metropolitan areas of the world, yet this 'yuppie' culture maintains that people ought to belong to some house, some household (often lovingly decorated and furnished), a certain area, a group of friends and colleagues. Big festivals like Christmas—which is, indeed, the only feast of a truly international character celebrated today, not only in the West but in most parts of the world including China and Africa—provide this 'home' with an anchor in time and space. Groups of families and friends reunite during these days to have traditional forms of communality, exchange, and mythopoesis about a bearded old man leaving his occult home in the last area on the world that still proves unreachable for globalization, the polar circle, carrying with him gifts and goods that will help to make people feel more at home (Miller 1993).

4. Final Thoughts

In the era of 'moving targets' (Marcus 1999), anthropological fieldwork has partly lost its central referent: the small group inhabiting a restricted area since immemorial times—the field. However, it seems that humans all over the world are struggling to recreate homes under the circumstances created by today's form of globalization. The anthropological 'field' therefore re-emerges over and over again, sometimes 'under fire' (Nordstrom 1995) and sometimes in a peaceful dialogue with the anthropologist, that 'professional stranger' who leaves his home to study others' homes.

See also: Collective Beliefs: Sociological Explanation; Collective Memory, Anthropology of; Community, Social Contexts of; Community/Society: History of the Concept; Community Studies: Anthropological; Everyday Life, Anthropology of; Family, Anthropology of; Family as Institution; Folklore; Household in Anthropology; Household Production; Kinship in Anthropology; Nationalism, Sociology of; Place in Geography; Tradition, Anthropology of

Bibliography

Augé M 1992 *Non-Lieux. Introduction à une anthropologie de la surmodernité*. Éditions du Seuil, Paris

Berndt R 1977 *Aborigines and Change: Australia in the 70s*. Australian Institute of Aboriginal Studies, Canberra, ACT

Borneman J 1995 American anthropology as foreign policy. *American Anthropology* **97**: 663–72

Cieraad I (ed.) 1999 *At Home: An Anthropology of Domestic Space*. Syracuse University Press, New York

Clastres P 1973 De la torture dans les sociétés primitives. *L'Homme* **13**: 114–20

Dehavenon A L 1999 *There's No Place Like Home. Anthropological Perspectives on Housing and Homelessness in the United States*. Bergin and Garvey, Westport, CT

Douglas M 1966 *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. Routledge and Kegan Paul, London

Fortes M 1969 *Kinship and the Social Order*. University of Chicago Press, Chicago

Gellner E 1998 *Nationalism*. New York University Press, New York

Gilman S 1990 *Jewish Self-Hatred: Anti-Semitism and the Hidden Language of Jews*. Johns Hopkins University Press, Baltimore, MD

Godelier M 1982 *La Production des Grands Hommes*. Fayard, Paris

Hannerz U 1980 *Exploring the City*. Columbia University Press, New York

Hobsbawm E, Ranger T (eds.) 1992 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Kramer F W, Kramer-Marx G 1993 *Zeitmarken. Die Feste von Dimodonko*. Trickster, Munich, Germany

Kuper A 1988 *The Invention of Primitive Society: Transformations of an Illusion*. Routledge, London

Lévi-Strauss C 1955 *Tristes Tropiques*. Plon, Paris

Lienhardt G 1961 *Divinity and Experience. The Religion of the Dinka*. Oxford University Press, Oxford, UK

Lopreato J 1967 *Peasants No More: Social Class and Social Change in an Underdeveloped Society*. Chandler, Scranton, PA
Maquet J 1961 *The Premise of Inequality in Ruanda*. Routledge, London
Marcus G (ed.) 1983 *Elites. Ethnographic Issues*. University of New Mexico Press, Albuquerque, NM
Marcus G 1999 *Ethnography Through Thick and Thin*. Princeton University Press, Princeton, NJ
Malkki L 1995 *Purity and Exile. Violence, Memory and National Cosmology among Hutu Refugees in Tanzania*. University of Chicago Press, Chicago
Mauss M 1960 Essai sur le Don. In: Mauss M (ed.) *Sociologie et Anthropologie*. Presses Universitaires de France, Paris
Miller D (ed.) 1993 *Unwrapping Christmas*. Clarendon Press, Oxford, UK
Nordstrom C (ed.) 1995 *Fieldwork Under Fire. Comparative Studies of Violence and Survival*. University of California Press, Berkeley, CA
Norman K 1991 A sound family makes a sound state. Ideology and upbringing in a German village. *Stockholm Studies in Social Anthropology, 24*. Almqvist and Wiksell International, Stockholm
Radin P 1957 *Primitive Man as Philosopher*. Dover, New York
Rao A 1987 *The Other Nomads: Peripatetic Minorities in Cross-Cultural Perspective*. Böhlau, Vienna
Schuster C, Carpenter E 1996 *Patterns that Connect. Social Symbolism in Ancient and Tribal Art*. Abrams, New York
Spencer B, Gillen F J 1938 *The Native Tribes of Central Australia*. Routledge, London
Turner V 1969 *The Ritual Process. Structure and Antistructure*. Routledge and Kegan Paul, London
Vertovec S, Cohen R (eds.) 1999 *Migration, Diaspora and Transnationalism*. Elgar, Cheltenham, UK
Widlok T 1997 Orientation in the wild: The shared cognition of Hai//om bush people. *Journal of the Royal Anthropological Institute* **3**: 317–32
Wilson E 1975 *Sociobiology*, abr. edn. Belknap, Cambridge, MA
Wolf E 1999 *Envisioning Power. Ideologies of Dominance and Crisis*. University of California Press, Berkeley, CA

T. Hauschild

Homelessness: United States

Homelessness has now been a fixture on the political scene in the United States and other countries for two decades. Yet the scope of the problem and the appropriate remedies for it still prompt fierce debates. This article examines the difficulties of defining homelessness and establishing useful estimates of population size and dynamics. It then provides basic descriptive information about homelessness in the United States and looks at its patterns and antecedents. The article ends by describing the growth of homeless-specific service programs and considering the probable future directions for homeless research and policy. Information about homeless populations and services for the United States is based on data collected in October and November of 1996 by the National Survey of Homeless Assistance Providers and Clients (NSHAPC). NSHAPC provides the most comprehensive overview available for the country as a whole (Burt et al. 1999).

1. *Defining Homelessness*

Homelessness did not exist as a public policy issue from the end of the Great Depression until the early 1980s. Certainly there were people without homes, or with very inadequate or dangerous homes. In some poor areas of many cities known as 'skid rows,' shelters, or 'missions' run by religious groups offered a bed and sometimes a meal in exchange for the chance to save a soul. These skid row areas also contained hotels where one could rent a small room, with perhaps a hotplate and a small sink, for a day, a week, or a month depending on how much money one had. In addition, rooming or boarding houses offered inexpensive accommodations.

Before 1980, the US Census classified people as 'homeless' if they lived by themselves and not with family members. Thus, all of the single people living alone in skid row areas (and elsewhere) might have been called homeless—a definition quite different from the one in use today, when very large numbers of the general population live alone by choice. Because of definitional problems and the absence of a concept close to what today would be called homelessness, there are no good estimates of the size of the literally homeless population for the 1960s and 1970s. The best guess is that in the late 1970s formal shelters (i.e., missions) for the homeless might have been able to accommodate about 50,000 people at any one time in the entire country, and that the number of literally homeless people probably was no more than 100,000 and could have been considerably less.

Since 1987, official US policy defines people as literally homeless if they sleep in a shelter for homeless people or in places not meant for human habitation such as vehicles, abandoned building, outdoor locations, or transportation facilities or stations. However, the people included in any actual study of the homeless, and therefore the descriptions of homeless populations and estimates of their size, will vary greatly depending on study methods (Taeuber 1991). If studies go only to shelters they will miss all homeless people who do not stay in them. Even studies that include searches of nonshelter locations have a very hard time finding most unsheltered homeless people. The best methods developed to date use services such as soup kitchens and outreach programs to find the nonsheltered homeless, and succeed better than street searches. In some European countries with a right to housing, homelessness is sometimes defined quite differently, as people who come to a housing authority needing housing (which they usually receive). Ob-

viously both definitions of homelessness and methods used to find homeless people will affect estimates of population size, while public policies will affect the types of remedy available and the probable length of homeless careers.

2. Homeless Numbers and Rates

The best point prevalence estimate of the size of the homeless population in the US before the early 1980s was 100,000 or fewer. By 1984 the estimate was up to 250,000 to 300,000, and by 1987 to 500,000 to 600,000. Estimates for 1996 of homeless people *using homeless assistance programs* (*not all* homeless people) range from 444,000 for October to 842,000 for February (Burt et al. 2001). All of these rates include any children accompanying a homeless adult, but not runaway and homeless youth.

The best way to compare estimates over time and place is to convert them to rates, which for homeless estimates is usually a rate per 10,000 population. Homelessness rates based on the 1987 estimate of 500,000 to 600,000 homeless people were 19.4/10,000 to 24.3/10,000. Projecting the February 1996 estimate of homeless service users to the whole homeless population assuming that service users during this winter season were 80 percent of the population produces an estimated rate of 38.0/10,000. This significant increase in the estimated rate of homelessness at any given time owes a good deal to the more than 200 percent expansion of homeless shelter capacity between 1987 and 1996 (see below), all of whose residents are classified as homeless.

Annual prevalence rates are a far better vehicle that point prevalence rates for understanding the proportion of a population affected by homelessness over time, because they account for the very large number of homeless spells lasting less than one year. Estimates for New York and Philadelphia indicate that 1 percent (100/10,000) of the entire population of those cities used a homeless shelter during a year's period (Culhane and Hornburg 1997, pp. 101–34). The same level of homelessness was estimated for the entire US for 1996 by Burt et al. (2001). When considered as a proportion of the *poor* population, the rate is close to 10 percent experiencing homelessness in a year's time.

3. Major Characteristics of Homeless People in the United States

3.1 Demographic Characteristics

In 1996, information from NSHAPC indicated that about one-third of homeless services users were members of homeless families; 23 percent were the children in these families, and 11 percent were their parents. These families comprised 15 percent of all homeless *households*; 86 percent of which were headed by women. Among family households, 38 percent were white non-Hispanic, 43 percent were black non-Hispanic, 15 percent were Hispanic, and the remainder were from other races/ethnicities. Forty-one percent had never married, and only 23 percent were currently married. Educational attainment was very low, with 53 percent having less than a high school education. These families had a mean monthly income of $475, or less than half of the US federal poverty level for a family of three (Burt et al. 1999).

Nonfamily homeless households were 61 percent men by themselves, 15 percent women by themselves, and 10 percent men or women with at least one other adult (Burt et al. 2001). Among nonfamily homeless service users, 41 percent were white non-Hispanic, 40 percent were black non-Hispanic, 10 percent were Hispanic, and 9 percent were from other races/ethnicities. Half of these homeless adults had never married, and only 7 percent were currently married. Equal proportions (36–37 percent) had completed high school and had less than a high school education. Mean monthly income was $348 or 51 percent of the federal poverty level for a single person (Burt et al. 1999).

3.2 Other Characteristics

In 1996, less than 10 percent of homeless adults worked in a regular job while homeless. However, 44 percent did some paid work for a month. Homeless adults were almost disconnected from the financial support of family and friends, with only 21 percent receiving money from these sources in an average month. Homeless families were somewhat better connected to public benefits, with 52 percent receiving the primary public welfare benefit of Aid to Families with Dependent Children and associated food and healthcare benefits. About 1 in 10 homeless adults received Supplemental Security Income, a program for people with serious long-term disabilities preventing them from working. Many studies find high levels of alcohol, drug, and mental health problems among homeless populations; the NSHAPC included. During the year before the study, 46, 38, and 45 percent of homeless adults reported alcohol, drug, and mental health problems, respectively. Only one in four did not report any of these problems. Homeless women were more likely to report mental health problems without either alcohol or drug problems, while homeless men were most likely to report having problems only with alcohol and not with drugs or mental health issues.

3.3 Patterns of Homelessness

Peoples' experiences of homelessness vary considerably. Many researchers have differentiated three patterns of homelessness—first-time/crisis/transitional, episodic, and chronic—although the exact classification criteria vary across studies. The first group includes individuals or families who are homeless only once or twice and usually for a relatively short period of time (e.g., less than 3, 6, or 12 months). Often they have been precariously housed and find themselves homeless after an unexpected crisis such as unemployment, divorce/separation, or eviction. They usually secure some type of housing relatively quickly and often do not become homeless again. Episodically homeless people are defined by their pattern of cycling in and out of homelessness over significant periods of time. Many spend time in various institutions such as inpatient treatment centers, detoxification programs, and jail. Chronically homeless people have relied on shelters or lived on the streets for many months or years, and usually have multiple barriers to securing stable employment and housing. They may have only a few distinct spells of homelessness, but each one lasts for a very long time.

Point-in-time studies, which overestimate the proportion of homeless people with long spells, usually find that about one-quarter of homeless adults are in the first group, and that half or more have extensive histories of either episodic or continuous homelessness. Studies that are able to observe flows into and out of the homeless condition over time give a different picture, showing only 10 to 20 percent with long-term homelessness and up to half with quite short-term or crisis patterns (Culhane and Hornburg 1997, pp. 101–34).

4. Antecedents of or Factors Associated with Homelessness

Both societal and personal factors affect homelessness. The societal factors predict *when* and *where* homeless populations will increase or decrease; the personal factors help to predict *who* will be vulnerable to these societal pressures. The two sides of the equation determining the ability to afford housing are household income and housing costs. In the US, housing costs for the lowest-cost units has increased significantly in the past 2 decades. Also, homelessness most recently became a public issue in the early 1980s, during a time of record-high unemployment (the recession of 1981–1982) that reduced household income among both low-wage workers and workers in manufacturing, many of whom lost their jobs. The record prosperity of the US during most of the 1990s has been accompanied by the development of a two-tiered labor market that has relegated many people without a college or technical education to low-paying service jobs. At the same time, wealthier households have pushed housing costs to record-high levels. The differential between earning power and housing costs sets the societal stage for homelessness. A public safety net of housing and other programs could assure that households do not founder under such circumstances. Some European countries have such a safety net, but the US does not. Some level of homelessness is the result.

That being said, certain people are more likely than others to become homeless. Poverty is the *sine qua non* of homelessness. The explanatory value of demographic characteristics such as race/ethnicity diminish or disappear in many analyses that simultaneously include income variables plus characteristics such as physical health, childhood experiences, and variables relating to alcohol and drug use and mental health problems. Regression results from an increasing number of studies on both families and singles confirm that sexual or physical abuse by a household member, foster care, or other out-of-home placement in childhood differentiate both homeless families and homeless singles from their housed counterparts. Physical and mental health conditions in adulthood, and early and current alcohol or drug use are also usually found to increase the probability that a person will experience homelessness when other things are held constant.

5. Growth in Services and Evidence of Effectiveness

5.1 Shelter/Housing Capacity

NSHAPC offers the first opportunity in many years to update knowledge about homeless assistance programs from the national perspective. Estimates based on NSHAPC data indicate that the nation's shelter/housing capacity within the homeless assistance system grew by 220 percent between 1988 and 1996, from the 275,000 beds estimated by the US Department of Housing and Urban Development (HUD) for 1988 to more than 600,000 beds in 1996. This growth came on top of the 275 percent increase in shelter capacity between 1984 and 1988 from 100,000 to 275,000 beds. One of the most striking things about the growth of shelter/housing capacity is the extent to which the system has gone beyond emergency shelter to include large numbers of transitional housing units, and permanent housing units for disabled people who were once homeless. In fact, relatively little growth occurred in emergency shelter programs, or beds specifically, between 1988 and 1996. HUD estimated 5,400 shelter programs in 1988, with a capacity of 275,000 beds. Virtually all of these programs and their beds supplied emergency accommodation, either through emergency shelters or through vouchering arrangements with hotels and motels. In 1996, NSHAPC data indicate that the number of beds available through more than

8,800 emergency shelter plus voucher distribution programs had grown to only 333,500, or a mere 21 percent over the level in 1988.

In contrast, transitional and permanent housing units exploded, going from basically none in 1988 to about 275,000 in 6,300 programs by 1996. Thus, this entirely new capacity to help people with housing needs beyond emergency shelter was essentially equal, in 1996, to the entire homeless assistance system in 1988. This growth coincided with the initial availability and eventual expansion of federal funding for new types of shelter/housing stimulated by homeless legislation.

5.2 Soup Kitchen/Meal Capacity

No national data exist to describe soup kitchen and other prepared meal distribution programs for 1987 that are equivalent to HUD's 1988 survey of shelter capacity. However, a 1987 Urban Institute study (Burt and Cohen 1989) did provide a roughly equivalent look at soup kitchen and mobile food program capacity in cities over 100,000, which can be contrasted to NSHAPC's estimates for central cities in 1996. Big-city prepared meal capacity increased almost four-fold between 1987 and 1996, from 97,000 up to 382,100 meals a day. Nationally, NSHAPC estimates are that soup kitchens and mobile food programs expected to serve almost 570,000 meals a day in February 1996, of which one-third were served in suburban/urban fringe areas or rural areas outside of central cities.

Unlike most shelter/housing programs, users of which are presumed to be literally homeless, soup kitchens and mobile food programs commonly serve many nonhomeless people (although some may have had earlier episodes of homelessness). In addition, people may eat more than one meal a day at these programs. For both these reasons, the estimate of meal service levels in 1996 cannot be interpreted as a level of homelessness. However, it certainly can be interpreted as a level of need. Also, soup kitchen/prepared meal programs have not had even one-tenth of the federal financial incentive to increase their service level that has been available to open or expand shelter programs over the past decade. This makes the service growth in the emergency food arena between 1987 and 1996 even more remarkable than the growth of the shelter/housing segment of homeless assistance. It attests to the responsiveness of many communities' emergency capacity, as well as to the ongoing problem of hunger even in a decade when most Americans enjoyed remarkable prosperity.

5.3 Evidence of Effectiveness

Effectiveness should only be assessed against expectations that are realistic to the programs and services examined. If the goal is to supply shelter and food, then shelters and soup kitchens are effective in meeting that goal. If the goal is to end, or prevent, homelessness, then the growth in shelter capacity in the US has probably had the opposite effect. This perverse result arises from the definition of homelessness, which includes residence in a homeless shelter. It is unclear whether the underlying level of need—the pool of people desperate enough to seek shelter from the homeless service system—grew in the 1990s. But the expanded shelter capacity gave such people a place to go that did not exist in the 1980s, and they have used it.

In 1999, a federally sponsored symposium featured papers summarizing ten years of research on the effectiveness of services for homeless people (Fosburg and Dennis 1999). Most of the findings revealed that intermediate program goals were being achieved. For instance, outreach programs were able to engage homeless people with serious disabilities, and eventually connect them to services. Case management programs were also able to connect homeless people to services. Psychiatric services were able to help mentally ill homeless people with their mental health problems. However, on average, none of these programs were able to end homelessness for their participants unless they supplied housing (Shinn and Baumohl 1999). If they supplied housing, especially with supportive services as required, programs could help even the most disabled homeless people leave the streets and maintain themselves in stable residential situations.

6. Future Directions for Policy and Research

In the US the argument has come full circle after almost 2 decades. The notion that homelessness was a problem of 'housing, housing, and housing' was made forcefully in the early 1980s, and has now proved to be quite accurate. Service programs have helped many homeless people in many ways. But to help people leave homelessness for good, or to divert people from entering into homelessness, the resources required are housing resources. These include help with landlords, back rent, evictions, and move-in expenses as diversion outlays, and actual housing as a means to end homelessness. A big question exists as to the value of supportive services to accompany the housing, once the housing itself is supplied. Research on this issue is summarized by Shinn and Baumohl (1999). Some research, mostly on homeless families, suggests that housing is the key and that services add little or nothing. Other research, more focused on chronically homeless people with significant alcohol, drug, and/or mental health problems, indicates an important role for supportive or supervisory services to accompany housing.

Future research could profitably take two directions. The first direction would focus on what supports

or services made available *just before* a homeless episode or at the moment of its precipitating event could avert the transition to literal homelessness by stabilizing the situation or simply redefining the issue. For instance, some community support services for persons with chronic mental illness offer 'respite' facilities where a person can stay for several days during a crisis in their illness. Respite stays avert complete dissolution of the person's life in the community, including loss of housing. It would be important to learn how these types of interventions can work best, for whom, and when. The second direction for research would focus on what services *in addition to housing* would make the most difference for people who are already homeless. Some communities in the US have already shifted resources from emergency shelter to affordable and supported housing, actually reducing the need for emergency shelter beds. More attention needs to be focused on developing a range of such approaches, because it should be clear that continuing to increase the capacity of the homeless service network will only serve to bring more and more people into the category of homeless.

See also: Drug Addiction; Drug Addiction: Sociological Aspects; Drug Use and Abuse: Psychosocial Aspects; Housing Economics; Mental Health Programs: Children and Adolescents; Mentally Ill: Public Attitudes; Poverty, Culture of; Social Stratification; Social Support and Health; Social Support, Psychology of; Social Welfare Policy: Comparisons; Street Children: Cultural Concerns; Street Children: Psychological Perspectives; Unemployment: Structural; Welfare; Welfare Programs, Economics of; Welfare State

Bibliography

Burt M R, Aron L Y, Douglas T, Valente J, Lee E, Iwen B 1999 *Homelessness—Programs and the People They Serve: Summary and Technical Reports*. Interagency Council on the Homeless/Department of Housing and Urban Development, Washington, DC

Burt M R, Aron L Y, Lee E 2001 *Helping America's Homeless: Emergency Shelter or Affordable Housing?* Urban Institute Press, Washington, DC

Burt M R, Cohen B E 1989 *America's Homeless: Numbers, Characteristics, and Programs that Serve Them*. Urban Institute Press, Washington, DC

Culhane D P, Hornburg S P (eds.) 1997 *Understanding Homelessness: New Policy and Research Perspectives*. Fannie Mae Foundation, Washington, DC

Fosburg L B, Dennis D L (eds.) 1999 *Practical Lessons: The 1998 National Symposium on Homeless Research*. US Department of Housing and Urban Development and US Department of Health and Human Services, Washington, DC

Shinn M, Baumohl J 1999 Rethinking the prevention of homelessness. In: Fosburg L B, Dennis D L (eds.) *Practical Lessons: The 1998 National Symposium on Homeless Research*. US Department of Housing and Urban Development and US Department of Health and Human Services, Washington, DC

Taeuber C M (ed.) 1991 *Enumerating Homeless Persons: Methods and Data Needs*. US Department of Commerce, Bureau of the Census, Washington, DC

M. R. Burt

Homosexuality and Psychiatry

Psychiatry has served both to reflect and to help define Western societal attitudes toward homosexuality for much of the past two centuries. Perspectives on homosexuality—concerning its origins, its social and personal meanings, its diagnostic implications, and its clinical relevance—persistently represented one of the most contentious issues within American psychiatry during the latter part of the twentieth century. With a primary emphasis on the American experience, this article presents current definitions related to homosexuality and sexual orientation and examines the history of psychiatric theories and findings about homosexuality. The relationship between homosexuality and psychiatry in other countries and cultures has often paralleled what took place in the US but also reflects the diverse attitudes and policies regarding both sexuality and mental illness throughout the world. An example of such diversity is the recent legislation of gay and lesbian relationships by some European countries in contrast to the continued opposition to such acceptance in the United States.

1. Definitions

Increased understanding of sexuality in general has led to greater specificity in the use of terms related to sex, gender, sexuality, and sexual orientation. Terms such as gender identity and gender or sex role tended to be confused with sexual orientation in early writing about homosexuality; however, it is now recognized that these are related but discrete characteristics that may interact with but are not synonymous with homosexuality.

Sexual orientation refers to a person's erotic response tendency, homosexual, bisexual, or heterosexual, toward persons of the same or other sex as reflected in such indicators as the proportion of dreams and fantasies directed to one or the other sex, the sex of one's sexual partners, and the extent of physiological response to erotic stimuli associated with one or both sexes. Sexual orientation consists of three components, desire, behavior, and identity, which may or may not be congruent in an individual. The Kinsey scale—a 7-point continuous scale, with 0

representing exclusive heterosexuality, 6 representing exclusive homosexuality, and 3 representing equal amounts of both, or bisexuality—is the most widely used scale for describing sexual orientation. The terms homosexuality, heterosexuality, and bisexuality have not been applied consistently in theoretical, research, or popular discourse; they have been used to refer to a wide range of constructs, including categories of sexual desire, gender role attributes, forms of sexual behavior, personal and social identities, types of personalities and persons, degrees of normality and abnormality, and the presence or absence of mental illness.

The presence of erotic desire for someone of the same or other biological sex does not imply the concurrent existence of any other characteristics in that individual. Thus, for example, a woman may engage in homosexual behavior, demonstrate typical feminine gender role characteristics, be married to a man, and experience a heterosexual identity; or a man may have homosexual desire, have sex only with women, and show gender role nonconformity.

The terms gay, gay man, lesbian, and bisexual are used to refer to men or women who have developed a sexual orientation identity that is homosexual or bisexual. While gay sometimes is used as an inclusive term to refer to both men and women, reference to two distinct groups, gay men and lesbians, is more appropriate because of the significant differences between these groups in their development and experiences.

Homosexual, or lesbian and gay, and bisexual identities are the subjective experiences of being homosexual or bisexual—being significantly or exclusively attracted to someone of the same sex—with the identities being both self- and other-ascribed. Models of lesbian, gay, and bisexual identity development usually portray a series of linear progressive stages involving tasks such as coming out (the process of recognizing one's homosexual or bisexual attraction and acknowledging it to oneself and to others), involvement in lesbian, gay, and bisexual communities, establishing same-sex relationships, and integration of one's sexual identity into other aspects of the self. Not all persons who experience homoerotic desire or participate in homosexual behavior also develop a lesbian, gay, or bisexual identity.

In contrast to sexual orientation, gender identity generally is understood to refer to a persistent sense of oneself as being male, female, or ambivalent and is contrasted with sex, which refers to the biological attributes of being male or female. While gender identity describes an inner experience of being male or female, gender role and social-sex role are defined as the degree to which an individual's outer behavior or appearance can be described as masculine, feminine, or androgynous. The large majority of persons, regardless of their sexual orientation, have a gender identity and gender role consistent with their biological sex, though sometimes gay and lesbian persons may demonstrate a greater range of gender role attributes.

2. *Historical Conceptualizations of Homosexuality within Psychiatry*

Prior to the time homosexuality was first studied as a psychiatric condition in the second half of the nineteenth century, the meaning of homosexual behavior was very different than it is today in Western culture. Homosexual behaviors take place in all societies and homosexual acts have been both ritualized and prohibited in various cultures. However, the contemporary idea of identifying a type of person based on regular participation in homosexual behavior and on development of an identity associated with this behavior is generally believed to be a product of the medicalization of homosexuality during the nineteenth century.

2.1 *Medicalization of Homosexuality*

During the second half of the nineteenth century a number of individuals in Europe began to study homosexuality scientifically. The term homosexual was first coined by the Hungarian physician Kertbeny in 1869. The conceptual basis for homosexuality arose from the earlier ideological construction of two genders, male and female, with dichotomous characteristics, and the attempt to find explanations for homosexuality derivative from these male and female categories. Thus, the homosexual was conceived of as a person of one biological sex with the sexual longing of the other sex, as portrayed, for example, in the German jurist and homosexual rights advocate Karl Heinrich Ulrich's description of the male homosexual as a female soul in a male body.

Subsequently, the German psychiatrists Carl Westphal and Richard von Krafft-Ebing adopted many of Ulrich's ideas, but they considered the traits he described to be unnatural. The German physician and sexual emancipator Magnus Hirschfeldalso believed that there was a genetic basis for homosexuality. While he understood homosexuality to be a malformation in development and ascribed to a number of different biological etiologies, he also asserted that it was natural. Similarly, the English sex reformer Havelock Ellis based his efforts to alter attitudes about homosexuality on the belief that homosexuality was an inborn trait. Specific biological theories about the origins of homosexuality began to be examined during this period, including a variety of constitutional factors, such as different anatomical structures in the brains and in nerves in the rectums of homosexual men, and endocrinological determinants, including atrophied and abnormal testicles in homosexual men.

These early biological theories about the origin of homosexuality, with their emphasis on heredity, endocrinological influences, and anatomical or structural differences between homosexual and heterosexual persons, presage many current efforts to find a biological basis for homosexuality. Their almost exclusive emphasis on the study of homosexuality in men also parallels later twentieth-century biological research, which largely has ignored homosexuality in women. Further, the biological arguments helped to create the notion of a homosexual person by emphasizing the existence of fundamental constitutional differences between those who are attracted to persons of the same sex and those who are attracted to persons of the other sex.

2.2 *Psychoanalytic Approaches*

During the latter part of the nineteenth and the first half of the twentieth century, psychoanalysis, with its emphasis on intrapsychic and interpersonal development, represented an alternative approach to the biological sciences for studying and understanding homosexuality. Freud believed that heterosexuality was the natural outcome of normal development and that homosexuality represented an arrest in development from an instinctual bisexuality to a mature heterosexuality that arises as the result of a variety of factors. However, he did not view homosexuality as a sign of degeneracy and even asserted that homosexuality could occur in persons who had no other signs of deviation and no impairment in their functioning. Later psychoanalytic theorists, including Sandor Rado, who refuted Freud's model of innate bisexuality and developed his own theory of homosexuality as a type of phobia, emphasized far more heavily the beliefs that homosexuality is the result of problems in psychosexual development and that its occurrence is a sign of severe psychopathology.

The outcomes of the growth of interest in studying homosexuality within the biological and psychoanalytic fields during this period were similar: to establish a focus of interest about homosexuality outside of the realms of morality and religion within science and medicine; to reify a category of person, the homosexual, associated with homosexual behavior; and to confirm the abnormality and the pathology inherent in homosexuality.

2.3 *Normalization of Homosexuality*

The publication of the two volumes on male and female sexuality by Kinsey et al. (1948, 1953) marked the beginning of the move away from the view of homosexuality as pathology and toward an understanding of homosexuality as a normal variant of sexual desire and behavior. Kinsey's studies demonstrated the widespread presence of homosexual feelings and behavior among samples of several thousand American men and women thereby suggesting that homosexuality was not an isolated and aberrant phenomenon. Kinsey's conceptualization of sexual orientation as a continuum also introduced the idea that homosexuality and heterosexuality were not dichotomous categories, but rather occurred in varying degrees in individuals.

Ford and Beach (1951) demonstrated that homosexuality was common across cultures and occurred in almost all nonhuman primate species. The psychologist Hooker's (1957) publication of the results of her ground-breaking study comparing the projective test results from 30 homosexual men and 30 heterosexual men showed that experienced judges could not distinguish between the two groups based on their test results, a finding which established for the first time that homosexuality in men was not associated with mental illness. The effect of this significant finding was to require that future views of homosexuality as an illness could only be based on an *a priori* assumption that homosexuality itself was pathological. By the 1970s, influential psychiatrists like Judd Marmor, a psychoanalyst and president of the APA, began to publish works that moved the psychiatric profession further away from the perspective of homosexuality as a type of mental illness and toward a consideration of the perspectives offered in the newer research findings.

During this period of transition and in spite of the availability of contradictory findings from empirical studies, some psychologists, particularly behaviorists who attempted to extinguish homosexual expression through aversive techniques, and psychiatrists, primarily psychoanalysts such as Irving Bieber and Charles Socarides, persisted in publishing works that presented homosexuality as the outcome of developmental disturbance and as a sign of severe psychopathology. Bieber's work (Bieber et al. 1962), which has been criticized extensively for its methodology, was particularly influential in its portrayal of a pathogenic family type, consisting of a detached and rejecting father and a close-binding and domineering mother, that led to homosexuality in adult men.

3. *Classification of Homosexuality by the American Psychiatric Association*

Homosexuality was classified as a type of mental disorder in the first *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-I*) (1952). It was designated, along with the other conditions labeled as sexual deviations, as a type of sociopathic personality disturbance, defined by the existence of certain behaviors and not by the presence of distress or dysfunction. This classification of homosexuality as a

form of mental illness was not controversial at the time *DSM-I* was published, reflecting as it did the negative societal attitudes about homosexuality and the prevailing medical views of homosexuality. In *DSM-II* (1968), homosexuality continued to be listed as a type of sexual deviation.

Beginning in 1968, gay activists began to confront the APA about its classification of homosexuality as a type of mental disorder. Following a series of dramatic encounters between gay rights advocates and psychiatrists at several annual meetings of the APA, the organization began officially to re-evaluate its position. Following further political protest and significant scientific study of the issue, a proposal from the Nomenclature Committee of the APA to eliminate homosexuality from the *DSM* was approved by the APA's Council on Research and Development, Reference Committee, and Assembly of District Branches before being accepted by the APA's Board of Trustees in December 1973. As a result, for the first time in modern history homosexuality *per se*, was not classified as a type of illness. This action had enormous social significance in shifting the predominant and official view of homosexuality from one of sickness to one of acceptable behavior. The decision to declassify homosexuality was accompanied by passage of a position statement by the APA that supported the protection of the civil rights of homosexual persons:

> Whereas homosexuality in and of itself implies no impairment in judgment, stability, reliability, or vocational capabilities, therefore, be it resolved that the American Psychiatric Association deplores all public and private discrimination against homosexuals in such areas as employment, housing, public accommodations, and licensing, and declares that no burden of proof of such judgment, capacity, or reliability shall be placed on homosexuals greater than that imposed on any other persons. Further, the APA supports and urges the enactment of civil rights legislation at local, state, and federal levels that would insure homosexual citizens the same protections now guaranteed to others. Further, the APA supports and urges the repeal of all legislation making criminal offenses of sexual acts performed by consenting adults in private.

The rationale for removing homosexuality from the *DSM* was based, in part, on the clarification of the definition of a mental disorder as meeting the criteria of either experience of subjective distress or of generalized impairment in social effectiveness or functioning as a result of the condition. Neither of these criteria applied to homosexual persons who were satisfied with their sexual orientation and who did not demonstrate impaired functioning. In what appeared to be a compromise between opposing views, a new diagnostic category, 'sexual orientation disturbance,' was introduced to apply to those homosexual persons who 'are either disturbed by, in conflict with, or wish to change their sexual orientation.' When the *DSM-III* (1980) was published, sexual orientation disturbance was replaced with a new diagnostic category, 'ego-dystonic homosexuality,' a type of general psychosexual disorder that labeled homosexuality as a mental disorder only when it was accompanied by persistent distress. Finally, after further review of the appropriateness and utility of this diagnosis, all reference to homosexuality was removed from *DSM-III-R* (1986), representing the final acknowledgment by the American psychiatric profession that homosexuality is not a type of mental illness. This change in the official view of homosexuality signified a dramatic conceptual shift in the cultural meaning and significance of homosexual behavior within American society.

4. *Effects of Declassification*

The decision to remove homosexuality from the *DSM* led to widespread effects within psychiatry. At the organizational level, groups of lesbian, gay, and bisexual psychiatrists formed and many issues relevant to the mental health of lesbian, gay, and bisexual persons could now be considered. At a clinical level, while controversy has persisted among psychiatrists about how to conceptualize homosexuality, new understandings and techniques for evaluation and treatment based on gay and lesbian affirmative approaches were developed.

These changes were accompanied and shaped by larger contextual shifts that occurred as a result of removing homosexuality from the realm of pathology. Thus, for example, once homosexuality was no longer viewed as a disease, it became a topic that could more legitimately be studied within a variety of fields outside of psychiatry and medicine with their emphasis on pathology, such as the law, the social sciences, and the humanities. The expansion of gay and lesbian identities, communities, and culture in the years since the APA's action produced a plethora of new issues that interacted with and influenced mental health, such as coming out, relationship concerns, childbearing and raising, family interactions, and spirituality. Finally, new health concerns, particularly AIDS, substance abuse, and reproductive issues, have created a whole set of additional mental health issues relevant to lesbian, gay, and bisexual persons.

Following the APA's action to depathologize homosexuality, the formerly secretive 'GayPA,' consisting of an informal group of closeted homosexual psychiatrists, began to evolve into a more open and extensive social and political network of gay and lesbian psychiatrists. At the 1978 annual meeting, a formal structure was adopted for the Gay, Lesbian, and Bisexual Caucus of the American Psychiatric Association, which was subsequently re-named the Association of Gay and Lesbian Psychiatrists (AGLP). Consisting of several hundred members, the AGLP is an affiliated group of the APA and is one of

the largest organizations of gay and lesbian physicians in the world. Many of its members served as founding members of other gay and lesbian medical groups, and it has continued to provide support to members and to raise political and clinical issues relevant to gay, lesbian, and bisexual mental health within the psychiatric profession. In addition, a Committee on Gay, Lesbian, and Bisexual Issues was established within the structure of the APA to study and make recommendations regarding mental health concerns. As a designated minority and underrepresented group within the APA, members of the Caucus of Lesbian, Gay, and Bisexual Psychiatrists elect representatives to the Assembly of District Branches, and openly lesbian and gay psychiatrists have been appointed and elected to many leadership positions within the APA.

The substantive issues discussed within the APA relevant to gay, lesbian, and bisexual concerns have involved a variety of topics, such as parenting and custody, education, reparative therapy, and AIDS. This activism and openness could not have occurred without the radical paradigm shift that took place when homosexuality was no longer considered a disease. The reconstruction of homosexuality as a normal variation of sexual desire, behavior, and identity allowed its introduction as a legitimate topic for discussion within the field of psychiatry.

Within the 1980s and 1990s psychiatrists in academic settings have been acknowledged and promoted for their work on gay and lesbian topics. Openly gay and lesbian researchers and clinicians have helped to establish a whole new field of gay and lesbian mental health. Publication of a section on homosexuality in the *Review of Psychiatry* (Oldham et al. 1993), a comprehensive *Textbook of Homosexuality and Mental Health* (Cabaj and Stein 1996), a curriculum for learning about gay, lesbian, and bisexual issues in psychiatry training programs (Stein 1994), a *Journal of Gay and Lesbian Psychotherapy*, and numerous scholarly articles and volumes on affirmative approaches to working with gay, lesbian, and bisexual patients has provided a breadth of valuable resources for helping gay men and lesbians within mental health settings. As a result of this material, many mental health facilities and practitioners are increasingly sensitive to gay and lesbian needs and informed about gay and lesbian concerns.

Problems have remained even after the revolution in perceptions of homosexuality precipitated by the diagnostic change made by the APA in 1973. Some psychiatrists and other mental health professionals continue to view homosexuality and gay, lesbian, and bisexual persons in negative ways and to perpetrate harmful practices, such as involuntary hospitalization of gay, lesbian, and bisexual youth because of their sexual orientation and presentation of unsubstantiated claims for 'curing' homosexuality. Groups within the APA concerned with gay, lesbian, and bisexual mental health concerns have worked to enact new policies that will eradicate the vestiges of heterosexism and antisexuality. In December 1998 they were successful in having the APA approve a Position Statement on Psychiatric Treatment and Sexual Orientation which states that:

> … the American Psychiatric Association opposes any psychiatric treatment, such as 'reparative' or 'conversion' therapy which is based on the assumption that homosexuality *per se* is a mental disorder or based on the *a priori* assumption that the patient should change his/her homosexual orientation.

5. *Future of Homosexuality and Psychiatry*

While the discourse within psychiatry regarding homosexuality has been expressed primarily in the form of scientific arguments about developmental and therapeutic conceptualizations related to sexual orientation, these arguments reflect the larger social debate within the US about attitudes and policy toward homosexuality. The psychiatric profession specified a clear governing principle—that homosexuality is not a type of mental illness—in its 1973 decision to depathologize homosexuality, but struggled for the next 25 years to articulate a consistent strategy for dealing with the implications of this decision. Perhaps official psychiatric policy was ahead of the larger society in its unequivocal rejection of a view of homosexuality as a type of disease. Like American society in general, however, psychiatry continued to maneuver to accommodate two totally conflicting viewpoints about sexuality, morality, and social values that were condensed in attitudes toward homosexuality. It remains to be seen if the definitive opposition to reparative therapy approaches to homosexuality expressed in the 1998 position statement represents final closure within psychiatry on this enduring debate.

The impact of diagnostic and treatment changes in relation to homosexuality within psychiatry during the past three decades has extended far beyond the mental health field. First, in 1973, by removing homosexuality itself from the medical spheres of etiology and pathology, the psychiatric profession yielded dominion over sexual orientation to those who dealt with normal variations of behavior, such as psychology, sociology, and other biological and behavioral sciences. The freeing of homosexuality from its century-long ownership by psychiatry can be seen as a major impetus to the expansion and deepening of understanding about the topic within many fields and to the burgeoning growth and expression of gay and lesbian identities and communities. Once the gate to the prison of pathology had been unlocked, those who had previously been incarcerated because of their sexual identities lost no time in claiming their newly found freedom and, at the beginning of the twenty-first century, are demanding assertively equal access to fundamental social institutions and rights such as

marriage, parenting, and legal protection based on an assumption of their normality.

See also: Ethics and Psychiatry; Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Heterosexism and Homophobia; Mental Health and Normality; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexual Risk Behaviors

Bibliography

American Psychiatric Association 1952 *Diagnostic and Statistical Manual: Mental Disorders*. American Psychiatric Association, Washington, DC

American Psychiatric Association 1968 *Diagnostic and Statistical Manual: Mental Disorders*, 2nd edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1980 *Diagnostic and Statistical Manual: Mental Disorders*, 3rd edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1986 *Diagnostic and Statistical Manual: Mental Disorders*, rev. 3rd edn. American Psychiatric Association, Washington, DC

Bayer R 1987 *Homosexuality and American Psychiatry: The Politic of Diagnosis/Ronald Bayer; with a New Afterword on AIDs and Homosexuality*. Princeton University Press, Princeton, NJ

Bell A P, Weinberg M S 1978 *Homosexualities: A Study of Diversity among Men and Women*. Simon & Schuster, New York

Bell A P, Weinberg M S, Hammersmith F K 1981 *Sexual Preference: Its Development in Men and Women*. Indiana University Press, Bloomington, IN

Bieber I, Dain H J, Dince P R, Drellich M G, Grand H G, Grundlach R H, Kremer M W, Rifkin A H, Wilbur C B, Bieber T B 1962 *Homosexuality: A Psychoanalytic Study*. Basic Books, New York

Cabaj R P, Stein T S (eds.) 1996 *Textbook of Homosexuality and Mental Health*. American Psychiatric Press, Washington, DC

D'Augelli A R, Patterson C J (eds.) 1995 *Lesbian, Gay, and Bisexual Identities Over the Lifespan: Psychological Perspectives*. Oxford University Press, New York

Ford C S, Beach F A 1951 *Patterns of Sexual Behavior*. Greenwood Press, Westport, CT

Hooker E A 1957 The adjustment of the male overt homosexual. *Journal of Projective Techniques* **21**: 17–31

Kinsey A C, Pomeroy W B, Martin C E 1948 *Sexual Behavior in the Human Male*. Saunders, Philadelphia, PA

Kinsey A C, Pomeroy W B, Martin C E 1953 *Sexual Behavior in the Human Female*. Saunders, Philadelphia, PA

Lewes K 1988 *The Psychoanalytic Theory of Male Homosexuality*. Simon and Schuster, New York

Magee M, Miller D C 1997 *Lesbian Lives: Psychoanalytic Narratives Old and New*. Analytic Press, Hillsdale, NJ

Oldham J M, Riba M B, Tasman A 1993 *Review of Psychiatry*, Vol. 12. American Psychiatric Press, Washington, DC

Ryan C, Futterman D 1998 *Lesbian and Gay Youth: Care and Counseling*. Columbia University Press, New York

Stein T S 1994 A curriculum for learning in psychiatric residencies about homosexuality, gay men, and lesbians. *Academic Psychiatry* **18**: 2–59

Weinberg M S, Williams C C J, Pryor D W 1994 *Dual Attraction: Understanding Bisexuality*. Oxford University Press, New York

T. S. Stein

Honor and Shame

Anthropologists conducting research into Mediterranean societies first developed the notions of honor and shame to describe the particular complex of values they found there. Honor comprised both an individual's sense of self-worth and this person's reputation in the surrounding community. Shame, by contrast, arose from the failure to act according to social values and it entailed public disgrace. This article traces the development of these concepts within anthropology and considers their current applicability.

1. Honor in Anthropology

By the middle of the twentieth century the word 'honor' was well on its way to obsolesence in ordinary English language usage. The practice of dueling to redress injuries to one's honor had ended and consequently honor had ceased to be a topic of public concern. The term was mainly frozen into stock expressions such as 'word of honor' or fixed in the aspic of formalized conventions such as the Queen's Honours List, or the school honor roll. Honor was still a familiar word, yet people were not absolutely certain what it meant; it was rich in connotations, but poor in denotation.

The Oxford-trained anthropologist Julian Pitt-Rivers, scion of a venerable English family whose forebears may well have participated in the culture of honor dueling in the last century, capitalized on the evocativeness of 'honor.' He reintroduced the word in a wide-ranging essay that drew principally on data from early modern European history. In this historical period honor drifted away from its medieval sense of a monarch's gift of fiefs in return for (military) services. In the seventeenth century it came to mean a general social reputation and then in the early nineteenth century it took on a more interiorized sense of integrity or moral dignity. If publicly insulted—called a liar or cuckolded, for example—a man was within his rights to demand the satisfaction of a duel to restore his honor.

Pitt-Rivers used a polysemous amalgamation of the historical senses of honor to interrogate his own contemporary ethnographic data from a Spanish village. His essay appeared alongside studies of other Mediterranean societies in a book entitled *Honour and Shame: The Values of Mediterranean Society* (Peri-

stiany 1965) that launched honor and shame in anthropology. The premise of the volume was that '[M]editerranean peoples … are constantly called upon to use the concepts of honor and shame in order to assess their own conduct and that of their fellows' (Peristiany 1965, p. 10).

For the most part honor and shame were conceived to be opposites and furthermore to orient different gender-linked ideals. Men were most likely to claim honor aggressively through acts of violence or cunning that earned them a fearsome reputation. This reputation was maintained by an ongoing performance of swagger and gruffness that staked out an inviolable space around a man's property, livestock, and women. As Peristiany noted (1965, p. 9), if provoked, 'a Greek Cypriot, a Bedouin and a Berber may answer "I also have a moustache".'

For their part, women protected this honor by comporting themselves with the modesty born of a sense of shame. Among the Sarakatsani shepherds of northern Greece described by Campbell (1964, p. 287), women do not run for fear that they might fall over backwards and expose their underwear, thereby shaming themselves irretrievably. Ideally a young woman should stay at home working on her dowry, well away from the public gaze so that her name will not be on people's lips.

2. *Comparative Considerations*

The Mediterranean focus of the *Honour and Shame* volume spurred further ethnographic research within this 'culture area' and also attempts to theorize why such a similar set of values had been arrived at in this region. Virtually all anthropologists assumed that honor and shame were relatively consistent in form throughout the Mediterranean until Michael Herzfeld (1980) pointed out, on the basis of research in two very different Greek communities, that honor amounted to 'socially appropriate behavior.' In a quiet, law-abiding village, macho swagger and aggressivity do not form the basis of honor.

The appreciation of the relativity of honor forced specialists into a theoretical stock taking (Gilmore 1987). Two definable tendencies emerged: (a) honor and shame vary in form in different communities, and indeed may be found in completely non-Mediterranean societies such as Japan where there is a great emphasis on shame; (b) although not exclusive to the Mediterranean, Mediterranean societies do, nonetheless, seem to converge on some strikingly similar cultural expressions of honor and shame. The gestural sign of the billy goat's horns made with extended index and baby finger, for example, signifies that a man has been cuckolded in a number of Mediterranean societies. The northern European resort to the image of the cuckoo bird to represent marital betrayal indicates the imagery of a different cultural area. In Mediterranean societies goats furnish the relevant vocabulary. A man who has been betrayed by his wife wears horns in Greek (*keratás*) and in Italian (*cornuto*) while in Spain he is just a big goat (*cabrón*). Unlike rams, billy goats allow their females to be mounted by other males without protest.

3. *Performance: Class and Gender*

In early modern Europe honor was mainly an affair of the aristocracy; slights from social inferiors could be ignored without consequences. The majority of early Mediterraneanists worked in small rural communities where everyone was putatively a social equal. All men began with an equal portion of honor that they could increase or lose. A single system of honor and shame thus permeated the whole community but in larger, stratified towns or cities there could be several parallel systems of honor by class. Honor itself was not economically determined, but the circles within which it was contested, and the modes of contestation, were.

There can be few societies lacking altogether in notions of precedence and social failure. The particular expressive performance of prestige in Mediterranean societies was surely what attracted the attention of anthropologists. These performances might involve elaborate verbal displays such as the spontaneous exchange of insulting rhyming couplets, or the skilled telling of tales of sexual prowess or daring animal theft. The anthropologists recording these stories usually came from the middle classes of northern European or North American societies. Their fascination with difference precluded the recognition of their own participation in systems of honor (Pina-Cabral 1989, p. 402). If transgressed a middle class person might conceivably report the incident to the police and perhaps bring charges. If defamed in public, this person might well register an accusation of libel and fight the case through the legal system rather than by direct, violent retribution. This configuration is no less a system of honor and shame, but it is not *the* system of honor and shame that Mediterraneanists had in mind.

Mediterranean honor assumed its particular form as a mode of direct action in the absence of moderating state institutions. The place to find similar configurations of honor and shame in many Western societies is often among the working classes or in criminal subcultures. American inner-city gangs have developed an honor culture that is perhaps even more 'Mediterranean' than that of Mediterranean societies as can be seen in the following statement by a Los Angeles Chicano gang member:

> Honor is important to me and every homey [gang member] because if you ain't got honor, you got nothing, man. I mean,

> no amount of money can buy you honor, so you do everything you can to protect it. I mean, without honor, nobody will respect you, not your friends or your family (Jankowski 1991, p. 143).

That middle class anthropologists were apparently oblivious to these alternative systems of honor in their own societies exemplifies the earlier contention that parallel, discrete codes for honorable comportment may operate in stratified societies.

Feminist studies of honor and shame have rejected the idea that honor is exclusive to men, and contended that women are not to be construed primarily as the passive defenders of male honor. Although their silence has been taken as a sign of submissiveness it might actually be a form of subversive irony through which women mock the supposed gender hierarchy. Furthermore, women's performances of suffering (their own pain, or on behalf of family members) can be understood as claims to respect in the community of fellow women sufferers. Dubisch (1995) has observed that Greek women engage in competitive suffering as a means of gaining prestige in the community. A gender studies model, then, where performances and social constructions of masculinity and femininity are treated independently, is increasingly replacing the honor and shame model which linked them together in a version of normal patriarchy.

4. *The Migration of Honor*

Just as the initial optimism in the revelatory power of honor and shame began to wear out in anthropology, scholars in other fields discovered the formulation and adopted it with excitement. Where only a few books about honor and dueling in European history had been written before 1970, by the 1990s many titles had been published analyzing these themes in several European countries. In his *Southern Honor* (1982), Wyatt-Brown acknowledged the influence of Pitt-Rivers and speculated that the similarities between the American South and the Mediterranean arose because 'primitive conditions of life often produce elemental social values sometimes roughly parallel from one society to another.'

Like *mana* (to which Pitt-Rivers compared it), taboo, or shaman, honor belongs to middle-range anthropological theory—terms discovered ethnographically in one region and then applied to guide investigation in other societies. The Azande distinction between witchcraft (innate) and sorcery (learned), for example, has been used in the identification of witchcraft in numerous other societies. Honor and shame appear to have been discovered in Mediterranean anthropology and to have traveled out to other societies and other disciplines. Yet if we look more closely honor and shame had already traveled from early modern European history into Mediterranean anthropology via Pitt-Rivers' essay. That these concepts now seem more attractive to Europeanist historians than they do to Mediterraneanist anthropologists may well be because they actually arise directly from these historical societies. Honor and shame were indigenous concepts in early modern European society.

The future of honor and shame in Mediterranean studies is murky precisely because they do not emerge from, or perfectly correspond to, local vocabularies. It is striking how few Mediterranean societies actually have words for 'honor' in the sense advanced by Pitt-Rivers and how infrequently equivalent terms are employed in daily life. In Greece the words *timí* 'honor' and *philótimo* 'sense of honor' are rarely heard while in Italy the operative concept is neither shame nor honor, even though the latter would seem to be readily accessible in the Italian word *onore*. Instead, exemplary moral fiber is displayed by following 'the code of silence' (*omertá*) and not disclosing details of criminal or other in-group activities to inquisitive authorities. In his early monograph describing an Andalusian village, Pitt-Rivers (1954) focused attention on 'shame' (*vergüenza*) and made no mention of 'honor.'

The conception of honor brought into anthropology by Pitt-Rivers was always slightly vague and polysemous and this was its theoretical attraction. Honor could involve criminal aggression or Christian virtue, its forms extended into gender relations and equally into the politics and economics of patronage and mafia protection. Anthropologists now find that the vagueness of honor and shame and their mismatch with comparable local concepts blunt their usefulness for ethnographic research and for comparison. Perhaps the most important legacy of the honor and shame formulation may prove to be in the area of the anthropology of self and emotion.

See also: Culture and Emotion; Emotions, Psychological Structure of; Emotions, Sociology of; Feud and Internal War: Legal Aspects; Guilt; Mafia; Shame and the Social Bond; Tribe; Violence, History of; Violence in Anthropology

Bibliography

Campbell J K 1964 *Honour, Family and Patronage: A Study of Institutions and Moral Values in a Greek Mountain Community*. Clarendon Press, Oxford, UK

Dubisch J 1995 *In a Different Place: Pilgrimage, Gender and Politics at a Greek Island Shrine*. Princeton University Press, Princeton, NJ

Gilmore D D (ed.) 1987 *Honor and Shame and the Unity of the Mediterranean*. American Anthropological Association, Special Publication 22, Washington, DC

Herzfeld M 1980 Honor and shame: problems in the comparative analysis of moral systems. *Man* **15**: 339–51

Jankowski M S 1991 *Islands in the Street: Gangs and American Urban Society*. University of California Press, Berkeley, CA
Peristiany J G (ed.) 1965 *Honour and Shame: The Values of Mediterranean Society*. Weidenfeld and Nicolson, London
Pina-Cabral J 1989 The Mediterranean as a category of regional comparison: a critical view. *Current Anthropology* **30**: 399–406
Pitt-Rivers J A 1954 *The People of the Sierra*. Weidenfeld and Nicolson, London
Pitt-Rivers J A 1965 Honour and social status. In: Peristiany J G (ed.) *Honour and Shame: The Values of Mediterranean Society*. Weidenfeld and Nicolson, London
Wyatt-Brown B 1982 *Southern Honor: Ethics and Behavior in the Old South*. Oxford University Press, New York

C. Stewart

Hope and Hopelessness

According to mythology, only hope remained in Pandora's jar after envy, spite, and revenge had been unleashed. But, how are we to interpret this ending? Was hope to be the antidote for these evils, or a counterproductive force? Although Judeo-Christian writers espoused the virtues of hope, Plato, Sophocles, Nietzche, and others have criticized it for prolonging suffering. Even more problematic, however, are the prevalent historical views of hope and hopelessness as being vague and useless concepts. Contrary to these traditional perspectives, however, hope and hopelessness have emerged in recent decades as having clearly articulated definitions, reliable and valid instruments for their measurement, and relationships with crucial life outcomes.

1. Definitions and Measures of Hope

In the 1950s and 1960s, psychologists (e.g., Stotland 1969) and psychiatrists (e.g., Menninger 1959) described hope in terms of positive expectations for goal attainment. In the 1970s and 1980s, researchers used experiential, spiritual, cognitive, and interpersonal frameworks for their positive expectancy hope models (Farran et al. 1995).

Three approaches to hope have received the most attention. First, Stotland (1969) hypothesized that hope reflected a greater than zero expectation of attaining important goals, and that it helps in understanding how antecedent events lead to subsequent outcomes. To measure hope, people are queried about their goal attainment expectations, or observed by raters. Although Stotland allowed for the possibility of trait propensities, he focused upon hope in particular situations.

Second, Averill et al. (1990) used a semistructured questionnaire approach, and found that people view hope as being appropriate when it is: (a) a short-term response to specific environmental conditions, and (b) tapping important, societally endorsed goals with 50 percent attainment probabilities.

Third, Snyder (2000) and colleagues, in what is called 'hope theory,' have posited that goal-directed expectancies involve two separable cognitive components—agency, or the person's perceived capability to initiate and sustain movement toward a goal, and pathways, or the perceived capability to produce effective routes to goals. Using this model, trait and state measures of hope in children and adults have been developed and validated (e.g., the 12-item dispositional Hope Scale for adults, Snyder et al. 1991).

2. Similarities and Differences of Hope Relative to Other Constructs

Because of its increasing scholarly attention since the early 1990s, hope theory will be used in order to explore the similarities and differences of hope to other theories. Viewing hope as agentic and pathways, goal-directed thinking provides an overlap with the Scheier and Carver (1985) generalized positive outcome expectancies model of optimism (indeed, this latter model is similar to the definitions of hope in the 1950s through 1980s); however, the pathways component of hope theory adds an explicit recognition of mental 'routing' capabilities in goal-directed thought.

In contrast to the Seligman (1991) model, in which learned optimism is viewed as an attributional process for distancing oneself from undesired outcomes, hope theory emphasizes a person's self-referential thoughts about reaching desired goals.

What about hope and self-esteem? It is posited (and supported in research) that perceptions of successful goal-directed activities (i.e., hope) drive a person's self-esteem. Thus, hope is a causal determinant of self-esteem.

Hope also resembles the self-efficacy model, with the pathways component being similar to what Bandura (1997) has called outcome expectancies, and agency being similar to his efficacy expectancies. A fundamental difference in the two models, however, is that the self-efficacy thoughts are given the pivotal driving role in Bandura's theory, whereas the agency and pathways goal-directed thoughts of hope theory are given equal weight as they iterate throughout the thinking process toward the attainment of goals.

3. Definitions and Measures of Hopelessness

Beck et al. (1974) defined hopelessness as a lack of enthusiasm, a motivational tendency to give up, and a dearth of positive expectancies. Using this definition, they developed the Beck Hopelessness Scale. Patterned theoretically after the Beck Hopelessness Scale, Kazdin et al. (1983) developed and validated the Hopelessness Scale for Children (ages seven and older).

Reasoning that elderly persons often perceive pessimism and futility about their futures, Fry (1984) conceptualized this perspective as hopelessness; moreover, he developed the Geriatric Hopelessness Scale to measure diminished abilities, lost interpersonal worth, and difficulty in recovering spiritual faith.

4. Hope and Hopelessness as Related to Other Variables

Dispositionally high-hope, as compared to low-hope, persons appraise their goals more in terms of successes than failures, they have higher perceived probabilities of attaining goals, and are more likely to experience positive emotions. Also, the positive appraisals of high-hope relative to low-hope people become especially advantageous when coping with stressors. Likewise, higher hope correlates positively with subsequent successful academic outcomes throughout the range of grades (elementary school through college), and better athletic performances (Snyder 2000).

Turning to the antithesis of hope, elevated hopelessness has correlated significantly with higher depression, as well as to more suicidal intentions and suicide completions in adults (Beck et al. 1990); moreover, greater hopelessness in children has related to more depression and suicidal ideation (Asarnow and Guthrie 1989).

5. Developmental Processes

5.1 Building Hope in Children

Newborns supplant raw sensations with perceptual representations. This learning of 'What's out there' leads to the selection of important objects (e.g., mother's face), thereby reflecting the emergence of desired goals. Infants also learn to recognize co-occurrence in events, and babies can anticipate events and act with intention (Schulman 1991). Also, babies learn 'what goes with what' survival skills by attending to those persons and things that fulfill their needs. Additionally, by age three, children can articulate their expectations about desired outcomes.

From ages three to six, hopeful thinking is enhanced as the child's vocabulary grows from approximately 50 to 10,000 words. By providing the means for describing goals and communicating with others, words and language help to frame hopeful thinking.

Through their reading or physical exploration, 7 to 12-year-old children expand their knowledge for use in goal pursuits; moreover, they increase their memory storehouses. These children also realize that their hopeful thoughts must be negotiated in the contexts of other persons (Flapan 1968).

Spanning the years 13 to 18, adolescents test their hopeful thoughts by gaining more understanding and experiences with interpersonal relationships.

5.2 Undermining Hope in Children

The events that diminish hope typically are more damaging when encountered earlier in the child's life, or when they continue for longer periods of time over childhood. Neglectful caregiving interferes with hopeful thinking because children in such circumstances do not have anyone educating them about goal-directed thinking. Likewise, abuse is particularly damaging because children fear the very caregivers who are supposed to instruct them in hopeful thinking. Also, parental loss by divorce or death can diminish hopeful thinking because children come to doubt the stability of relationships, as well as the fact that one can meet goals in the contexts of such relationships. Lastly, caregiving in which children's efforts are not supported lessens hope (Dweck 1999).

6. Increasing Hope

6.1 Children

In regard to goals, babies are aided by caregivers in identifying, naming, and using specific words for desired objects. The caregiver's interest and conversations about the child's goals are productive, as are reinforcements of goal-making.

To increase agentic thinking, children profit by being reminded that they are responsible for making things happen. Praise when children show determination to reach their goals solidifies their subsequent agentic thoughts. Similarly, helping children to practice positive self-referential talk (e.g., 'I can do this') enhances agency (Kanfer et al. 1975).

Pathways thinking can be promoted by showing babies how causality works (e.g., using drums). Songs, stories, and rhymes illustrate how words are sequenced, and how one thing leads to another. Teaching mental scripts about what children are to do in given situations is helpful, as are instructions about how to break routes into smaller, more easily attained steps.

For all of the aforementioned hope lessons, it should be emphasized that close attachments with caregivers are essential for fostering children's hopes.

6.2 Adults

It is important in producing hope that adults set goals based on their desires rather than what other people want for them. Additional helpful strategies include producing an array of goals, setting goals that stretch one's talents, and clearly articulating and setting concrete goal markers so as to facilitate the noticing of progress (Nezu and D'Zurilla 1981).

To enhance agentic thinking, particularly during difficult times, positive self-talk is useful. By learning to anticipate blockages, and to think of these impediments as challenges, people can enhance their determination (especially for when they encounter prob-

lems). People also gain mental lifts by remembering previously successful experiences. Furthermore, switching from truly blocked goals to more reachable ones increases agency.

To enhance one's pathways thinking, breaking a long-range goal into smaller subgoals is helpful. Other productive strategies include developing mental maps to goals, planning alternative routes, and rehearsing mentally. Friendships also provide powerful guidance for pathways-related thinking.

7. Future Directions

Research is needed into parental and school environments that are maximally conducive to hope, as well as the environments that must be avoided so as to lessen hopelessness. Also, additional hope and hopelessness research is warranted in regard to health, psychotherapy, interpersonal relationships, worker satisfaction and performance, stress and coping, and aging.

See also: Childhood Depression; Depression; Depression, Clinical Psychology of; Depression, Hopelessness, Optimism, and Health; Health Risk Appraisal and Optimistic Bias; Learned Helplessness

Bibliography

Asarnow J R, Guthrie D 1989 Suicidal behavior, depression, and hopelessness in child psychiatric inpatients: A replication and extension. *Journal of Clinical Child Psychology* **18**: 129–36

Averill J R, Catlin G, Chon K K 1990 *Rules of Hope*. Springer-Verlag, New York

Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman, New York

Beck A T, Brown G, Berchick R J, Stewart B L, Steer R A 1990 Relationship between hopelessness and ultimate suicide: A replication with psychiatric outpatients. *American Journal of Psychiatry* **147**: 190–95

Beck A T, Weissman A, Lester D, Trexler L 1974 The measurement of pessimism: The hopelessness scale. *Journal of Consulting and Clinical Psychology* **42**: 861–5

Dweck C S 1999 *Self-theories: Their Role in Motivation, Personality, and Development*. Psychology Press, Philadelphia

Farran C J, Herth A K, Popovich J M 1995 *Hope and Hopelessness: Critical Clinical Constructs*. Sage, Thousand Oaks, CA

Flapan D 1968 *Children's Understanding of Social Interaction*. Teachers College Press, New York

Fry P S 1984 Development of a geriatric scale of hopelessness: Implications for counseling and intervention with the depressed elderly. *Journal of Counseling Psychology* **31**: 322–31

Kanfer F H, Karoly P, Newman A 1975 Reduction of children's fear of the dark by competence-related and situational threat-related verbal cues. *Journal of Consulting and Clinical Psychology* **43**: 251–8

Kazdin A E, French N H, Unis A S, Esveldt-Dawson K, Sherick R B 1983 Hopelessness, depression, and suicidal intent among psychiatrically disturbed inpatient children. *Journal of Consulting and Clinical Psychology* **51**: 504–10

Menninger K 1959 The academic lecture on hope. *The American Journal of Psychiatry* **109**: 481–91

Nezu A, D'Zurilla T J 1981 Effects of problem definition and formulation on decision making in the social problem-solving process. *Behavior Therapy* **12**: 100–6

Scheier M F, Carver C S 1985 Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology* **4**: 219–47

Schulman M 1991 *The Passionate Mind*. Free Press, New York

Seligman M E P 1991 *Learned Optimism*. Knopf, New York

Snyder C R (ed.) 2000 *The Handbook of Hope: Theory, Measures & Applications*. Academic Press, San Diego, CA

Snyder C R, Harris C, Anderson J R, Holleran S A, Irving L M, Sigmon S T, Yoshinobu L, Gibb J, Langelle C, Harney P 1991 The will and the ways: Development and validation of an individual-differences measure of hope. *Journal of Personality and Social Psychology* **60**: 570–85

Stotland E 1969 *The Psychology of Hope*, 1st edn. Jossey-Bass, San Francisco

C. R. Snyder

Horticultural Societies

A horticultural society is one in which people's livelihoods depend on the cultivation of small plots, commonly called gardens. These frequently support a wide range of crops and meet staple subsistence needs. The proportion of produce sold on the market varies, to finance the purchase of commodities manufactured elsewhere (e.g., clothing, processed food, hardware, etc.). The scale of society and its organization may vary widely too, from small tribal groups to large peasant populations.

1. The Term 'Horticulture'

Horticulture is distinguished from agriculture by scale of production, specialization, and commercialization. Agriculture features in industrialized societies supplying food to large urban populations. It is, in some regards, the reverse of horticulture. It involves the farming of large fields, usually under a monocrop (all wheat, potatoes, beans, etc.), whereas the gardens of horticulturists usually feature intercropping, the mixing up of several different crops, as opposed to their segregation in separate rows or plots, for which there are sound ecological reasons (e.g., pest and disease control, maximizing nutrient and water availability, increasing photosynthetic efficiency, exploiting microenvironmental variations to best effect, and so on). Agriculture features highly mechanized production techniques, whereas horticultural societies rely more on human labor and animal drought power. Farmers in agricultural systems meet few of their own subsistence requirements, specializing in the large-scale production of a few foodstuffs for sale on the market. They engage in agribusiness, intensive production making wide use of inputs to boost yields, frequently

obtained from off-farm and industrially manufactured (inorganic fertilizers, biocides, etc.). The term horticulture has a quite different meaning in urban-industrial societies, which can confuse the unwary. It refers to the intensive production, often under artificial glasshouse conditions, of high-value crops, such as early vegetables, salads, and flowers by commercial growers.

Other food production regimes from which horticultural societies are distinguished include gatherer-hunters and pastoralists, which support lower population densities (Forde 1934), industrialized urban ones having higher densities. These categories were bequeathed to us by Victorian scholars who thought of them comprising an evolutionary hierarchy, putting horticultural societies somewhere in the middle of the range between the gatherer-hunter and industrial poles. This scheme continues to inform thinking to the current day (as evidenced by entries in this encyclopedia), and underpins notions of development conceived as facilitating the passage of communities to the urban-industrial pole. Anthropologists continue to use it, although unhappy with the continued evolutionary implications, arguing that we should not judge urban-industrial life, notably liberal market democracy, as the superior socioeconomic order to which all human beings should aspire. It is an approximate classification scheme. At some ill-defined point, populations are thought to manage their natural resources sufficiently closely to move beyond gatherer-hunting to farming, this is accompanied by a more sedentary lifestyle. They qualify as horticultural societies when they emphasize crop cultivation over animal herding, although mixed arable and livestock farming is common. The label horticultural society covers a wide range of different subsistence regimes and sociocultural orders. The term is more commonly used in North America than Europe, where instead of lumping these societies together, it is more usual to distinguish the different agronomic regimes that comprise horticultural orders.

2. *Agronomic Regimes*

We find varying degrees of sedentariness among people who depend largely on cultivation to supply them with their food. The variation informs a preliminary distinction between different horticultural strategies. There is also some correlation with increasing population densities.

Where we find the least permanent settlement and physically most direct technologies, we have cultivation systems that fallow the land not by rotating crops but by rotating land use. We call this type of horticulture *shifting cultivation*, because after people have grown crops, and fertility declines, and yields start to fall (these cultivators not using any fertilizer to supplement the soil's nutrients), they abandon the plot to the forest and move elsewhere to clear a new garden on fresh land. Secondary regrowth, meanwhile, recolonizes the abandoned area, which is left fallow for tens of years to allow nature to restore its fertility; following which it may be cleared and planted again in the next generation. Shifting cultivation is labor intensive, featuring hand tools, such as axes, machetes, hoes, and pointed digging sticks. We find a sexual division of labor, orchestrated into a joint effort. Men usually do the heavy axe-related work. They fell the trees and clear the forest. They also enclose the area to be gardened with rough fencing if this is necessary to keep out animals. The subsequent burning off of the dry and dead vegetation cleared from the plot is often done by both men and women working together. This burning is an important step, reducing standing vegetation to ash, which is readily mixed with the soil, liberating locked up nutrients in a form more available to planted crops. The release of nutrients (such as nitrogen, potash, phosphorus, and so on) through burning has led to the coining of the term 'swidden' to describe this type of cultivation. (Swidden is an Old English word meaning to burn.) Following clearing and burning off, women usually take over and are responsible for breaking up the soil, planting crops, weeding, and finally, harvesting the plot. People living around the world in Asia, Africa, the Pacific, and Latin America grow a large variety of crops using this cultivation technique: for example, grains such as corn, millet, and sorghum; roots such as manioc, yams, and sweet potato; and fruits such as bananas, pawpaws, and melons.

Many Amazonian Indian populations practice shifting cultivation, horticultural crops supplying the bulk of their food regardless of the time spent gathering and hunting in the surrounding forest (Watters 1971). Among the Yanomamo, plantains and bananas are the staple crop. They select well-drained sites, relatively clear of thorny underbrush. They do not attempt to burn large tree trunks, left strewing plots in a jumble, serving as boundaries between patches, the crops planted there belong to those who cultivate them. Newly established gardens yield crops in cycles, the rapidly growing ones such as maize coming first, followed by others including manioc, taro, sweet potato, and mapuey. The Yanomamo also grow chewing tobacco, cotton for hammock yarn, drugs for shamanistic rituals, and many fruits including papaya, avocados, gourds, and hot peppers. A garden produces constantly as it matures, but a regular supply of food demands careful management. Gardens have direction, moving little by little as people gradually abandon older areas as thorny underbrush encroaches. They constantly clear adjacent land and transplant secondary plantain suckers as they appear. Sometimes villages are obliged to establish new gardens elsewhere. This is usually a political decision, communities moving when threatened by raids from enemy villages. It requires a large effort to

establish a new garden some way off, to transport needed cuttings and seeds. New gardens also produce sporadically at first, and people may face hunger, whereas established 'rolling gardens' offer predictable subsistence.

At some ill-defined point, agronomists distinguish *bush fallow* systems. These are broadly similar to shifting systems with regard to technology and cropping strategies, but with a reduced fallow interval and longer periods under cultivation. Instead of one or two years under cultivation followed by two or more decades under fallow forest, these horticulturists may farm land for 5–10 years with a fallow interval of similar duration under bush or savanna vegetation. Stock may graze on the fallow, cattle, goats, sheep, and so on. Settlement is increasingly stationary, and land holdings are usually more clearly defined with agreed field boundaries. These fallow systems sometimes evolve from shifting regimes with increasing population densities. They may demand management of soil fertility through the application of manure or fertilizer, otherwise yields will decline over time. Degradation of the natural resource base may occur with increasing population pressure and demand to grow cash crops.

Again, at some ill-defined point, we distinguish *ley* systems. These are ones in which farmers systematically use the land to pasture livestock during the non-arable phase. At one end are unregulated leys where people agist their animals on unmanipulated natural regrowth of grasses and herbs. They have communal grazing rights, and practice little pasture management. At the other end are regulated leys where farmers manage the grazing vegetation, frequently putting down grasses and legumes as fodder. They maintain individual pastures which they may fence off. They may graze meadows directly or use them to produce hay or silage. At this point we usually think they have evolved into agricultural regimes.

Bush fallows and unregulated leys are common in interlacustrine East Africa (Allan 1967). Lugbaraland, for example, comprises fields and homesteads scattered across open upland country. Homesteads comprise compounds, clusters of family huts and granaries, under which are ghost shrines and magical plants. Fields surround them. Each homestead cultivates three types of fields. Men do the heavy work of opening the land, women weed and harvest. Both sow, depending on the crop, men handling more cash crops. They intercrop in all. Postharvest care and use of crops is the responsibility of women. The 'riverine fields', irrigated by channels from streams, are highly fertile and need only short fallows, they support sweet potato, maize, sugar cane, and bananas. The 'home fields', fertilized with ashes and manure, support nutrient-demanding crops such as beer-brewing white sorghums. The 'outside fields' support staple eleusine, together with sorghum, simsim, and pigeon pea, cultivated for two or three years, after which they are left under bush fallow to restore fertility. People allow unregulated grazing of cattle and goats on fallow. They recognize different fallows: from recent and regaining fertility; a few years old, the presence of certain plants showing restored fertility; cleared land left for a few months before hoeing; through to uncultivated land. Each woman in a homestead should have areas in all fields, of equal fertility. If not, quarrelling is likely. The fields of a homestead are likely to be scattered, due to differences in soil and water. This sets a limit on the maximum size of farming groups. A household short of certain farmland may move and take up residence with kinsmen elsewhere, usually a wife's or mother's brother.

It is generally agreed that with the invention of the plough we approach the realm of *agriculture*. The plough allows people to harness either the strength and energy of draught animals or engine-powered machines to assist in their preparation of land for the planting of crops. The result is that one person, together with their team of oxen or whatever, can accomplish considerably more work alone, releasing others to follow different occupations. We have the emergence of increasingly specialized labor, a move from so-called organic to mechanical solidarity. We find the practice of crop rotation and occasional fallowing to maintain soil fertility with the plough and associated agricultural machinery, together with the increasing use of amendments such as manure, fertilizer, biocides, etc.. In regions that suffer water deficits, people may have irrigation technologies. In some places they may have plantations of perennial commercial tree crops. The upshot is permanent farming, with clearly demarcated fields. While it is inappropriate to go into the growth of the first civilizations among the early cultivators of the fertile crescent in the Middle East, of Mesoamerica and of the Far East, we should note some of the consequences of these developments, founded on a settled agricultural way of life (Sauer 1969). They allowed for greater population densities, which, combined with the relative permanence of settlement that results with agriculture, led to the emergence of wider-scale political groups than those based on the family or tribe. These developments opened the way over millennia to the evolution of industrial-urban societies, sometimes with associated peasantries.

Before turning from the agronomic aspects of horticultural societies to consider the social orders associated with them, it needs to be stressed that while the foregoing regimes have been presented with an evolutionary dimension, it would be wrong to consider them as discrete steps leading from shifting cultivation to mechanized agriculture. This typology is no more than a low-level generalization by which we can usefully classify cultures. When we look closely we find that many societies do not fit comfortably into any one

category, but span two or more, as the previous East Uganda example illustrates. In the Baliem valley of New Guinea too, where sweet potato is the staple and people herd pigs, we find three sorts of cultivations. The largest are the permanent plots on the valley floor, some of which the Dani recultivate immediately upon harvest and others they leave fallow for months, even years. These gardens feature elaborate ditch systems which drain away/store water, delineate and protect cultivations from pigs, and serve as mulching basins supplying organic rich mud for slopping over beds to ensure continued fertility. Other gardens men clear in a classic shifting mode on the mountain slopes, burning off plots and planting for a year or two. The third and smallest are kitchen gardens maintained adjacent to houses, supporting a wide range of valuable crops such as tobacco, sugar cane, and bananas, continually fertilized with household waste. Households commonly have several different widely separated fields at different stages.

3. *Sociocultural Orders*

We now turn briefly to look at the gross nature of the social orders found associated with horticulturists. It is difficult to distinguish between horticultural regimes according to political economy (how people distribute resources, etc.). They occur in one of three broad categories: egalitarian with balanced reciprocity, chiefdoms with redistribution systems, or peasantries in nations with markets. There is some correlation between permanence of settlement, cultivation regime, and polity, with shifting cultivators frequently living in egalitarian orders and ley farmers in chiefdoms.

The *egalitarian societies* featuring balanced reciprocity are characterized by groupings commonly called tribes. These are relatively large groups of people (that is, in comparison to hunter-gatherer bands), who occupy demarcated territories, to which resources they lay specific claim and rights. Characteristically these societies recognize all individuals as equals and place no persons above others, with authority and power to rule over them. They are what we call stateless or acephalous, and have evolved various strategies to maintain social order without formal government, judiciary, and so on; employing, for instance, descent and age in Africa, and ceremonial exchange in Melanesia. The productive group is not the tribe, but usually the family, with some help from extra kin in occasional large cooperative ventures. Under this household mode of production farmers consume most of their produce, although they distribute some to others. The formal distribution of food and goods is likely to be significant, taking place predominantly in reciprocal transactions. Relatives and friends who give things to one another keep a careful account. They give, not with the expectation of receiving an immediate return gift (or payment), but with the expectation of an equal return sometime later. There is often a competitive element in these transactions, those doing best winning higher status. They sometimes constitute grand institutions and are attended with some ceremony. Exchange is a facet of social and political relations and may exist independently of material wants and needs. This takes us away from subsistence and horticulture towards sociopolitical issues.

In Melanesia, ceremonial exchanges of balanced reciprocity are central to the maintenance of social order. New Guinea highlanders live in an egalitarian society. Each family supplies its own food and provides its own shelter. It occupies a compound often comprising several related families and friends, with men's and women's houses, and pig sty. Sometimes friends and neighbors make up a cooperative work party to garden or build a house. But there are no leaders. The Dani combine for conviviality, each person doing the same thing, the output comprising the sum total of their efforts (it is more fun mudding gardens together, for example, even though this is a one-man job). Feuding is common, clans fighting with bows and arrows and spears. Social order is promoted through exchange, primarily of pigs, which are only killed on ceremonial occasions. The giving and receiving of pigs constantly refurbishes the social network, people are bound to one another through their lifetimes by the exchange of hundreds, even thousands, of pigs. Clan affiliation determines access to land, all have equal rights. These are rights of usufruct, not land ownership. Men do the heavy, spasmodic garden and construction work, and are more involved in public activities such as politics, warfare, and rituals. Women undertake routine cultivation work and domestic responsibilities, herding pigs, and tending family needs. There is little call for distribution of goods with this minimal sexual division of labor. There is some trade with people elsewhere for a few imported goods.

The *chiefdoms* with redistribution systems are hierarchically organized societies, where members are ranked one above the other. In the top rank are the rulers, commonly called chiefs. These persons have authority vested in them, often sanctioned by some supernatural connection, and govern those living in their chiefdom, kingdom, tribe, or whatever. Everyday rule is usually delegated to lower rank persons, who are locally responsible for day-to-day government, administration of justice, and so on. These chiefdoms are often feudal in nature, with these local leaders owing allegiance to the chief and supplying a military force if required by him. Societies organized in this way exploit raw materials on a larger scale. They do this, for example, by territorial conquests (as epitomized by the Zulu under their king Shaka) and through extensive trading networks (for example, the Aztecs controlled mining activities as far away as northern Mexico). These societies can also organize

larger productive groups through the authority of their rulers, who by so doing not only assist people but also provide some sort of rationale for the existence of their power (for example, the farming terraces, irrigation works, storehouses for produce, and so on, constructed by the Inca). Redistributive systems predominate in these societies, where those of lower rank give food and goods to those of higher rank in the form of tribute. The chief will rearrange everything he receives and then redistribute it to his subjects (usually via local-level officers in his government). He may add to it any surplus goods which his probably superior productive position supplies (i.e., from his large, communally-worked gardens).

The chiefdoms of South Africa are good examples of this political order. The horticulture practiced is broadly similar to East Africa, the main crops being maize and millet, together with ground nuts, gourds, sugar cane, and pumpkins. Farming is seasonal, dependent on the summer rains. One source of rulers' political authority is the belief that they can control the advent of the rains. People may have a precarious existence, food supplies fluctuating annually between plenty and scarcity bordering on famine. The Swazi, for example, talk of lean summer months as the time 'to swallow the pickings of the teeth.' Horticulture is not a prestigious occupation and is left primarily to women, whose main tool is the iron hoe. Except, that is, for ploughing, which involves highly valued oxen handled by men, and the cultivation of cash crops, notably tobacco and cotton. Households cultivate large fields using ox-drawn ploughs. The power rulers wield also relates to their right to allocate land to their subjects, who depend on it for their livelihood. The Swazi say that their political overlords 'serve' them the land, securing their rights by allegiance and usage. Women are allocated garden areas by their husbands on their marriage. Homestead plot size depends primarily on supply of resident labor. People use grazing lands communally. They value pastoralism highly, having the so-called African 'cattle complex': these animals have great symbolic value, being exchanged to ratify marriage, sacrificed to propitiate ancestors, and paid in tribute to rulers. They call the king the 'Bull' of the nation, and the royal herd far exceeds any other, supplying animals redistributed to subjects at national events and sacrifices. In addition to sex and age, pedigree also structures the division of labor. Men and women cooperate in agriculture, man's work is heavier and more spasmodic, woman's routine and more continuous. Aristocrats do not engage in manual labor to the same extent as their subjects, they are responsible for successfully organizing their efforts, supervising work parties and arranging for specialists to doctor land and seed. The rulers have several 'gardens of kingship' in different localities, of which local headmen arrange communal cultivation. Homesteads rely on kinsmen and neighbors to assist at bottlenecks, customarily organized into two competitive work parties they spur each other on in songs, rhythmically coordinating their work through music. The size of communal work parties reflects status: they enhance prestige, being rewarded with beer and food at the host's home.

The *peasantries* associated with markets are also incorporated into hierarchically organized societies. They are usually encapsulated within nation states where they occupy marginal and low positions, relatively powerless and commonly exploited. These are horticultural communities in peripheral, under-privileged situations, frequently taken advantage of, even abused by the dominant society. They usually feature extensive poverty. The larger political formations which rule significant aspects of their lives may supply them with certain services. Peasants depend in some regards on economic and political institutions beyond their local control. They may maintain informal reciprocal exchange networks within their communities, but outside they are obliged to participate in impersonal commodity markets. While they supply their own subsistence needs (similar to tribespeople, from whom many of the Third World's peasants have descended in recent times), they are obliged also to cultivate a surplus to sell on outside markets to earn needed cash. In addition to using this money to purchase goods manufactured elsewhere, they need some cash to pay the taxes levied on them by the colonial/national governments to which they are subject. They may also need to purchase outside inputs, such as inorganic fertilizers and biocides, to maintain their yields, especially if their farming system is under increasing demographic pressure, perhaps resulting in increasing environmental degradation. This may contribute to growing poverty. What peasants receive financially in return for their surplus horticultural produce is on the whole unfair, the élite higher up the social hierarchy, controlling access to imperfect markets and the arrangements for the distribution of produce, make a disproportionate profit to maintain their privileged position and lifestyle.

The Mayan Indians of Mexico have occupied a peasant status increasingly since the Spanish conquest (Wilken 1987). They sell crops surplus to their subsistence needs, largely maize and beans, and trade salt from the Ixtapa works at fiesta day markets. In turn they purchase goods such as coffee, sugar, fish, liquor, and bread from markets. They have been involved in marketing for centuries, historical evidence showing that they traded with Aztecs in prehispanic times. Maize is their staple, believed to have an 'inner soul' in the ear, together with beans, squashes, chillies, chayotes, and fruits such as peaches. They keep chickens and pigs, sheep for wool woven into *chamarras*, blankets and shawls, and mules for carrying heavy loads. The Mayans cultivate three types of land. Small, individually owned garden plots located near highland homes, their ancestral patrimony. Larger *ejido* plots

acquired during the 1940s agrarian reform program, allocated by local committees to households. And plots in the lowlands rented from wealthy *Ladino* landowners by patrilineal groups, which they divide equally among themselves. They use centuries' old swidden techniques, working with axes, machetes, bill hooks, and hoes. After cutting and burning vegetation, they plant using pointed digging sticks. In old fields they cut and rake maize stubble into piles for burning, and then hoe sheep manure into the turned topsoil. They prepare fields for planting at the onset of the rainy season. Ritual attends planting, a critical task in which a shaman offers candles, incense, liquor, and prayers to the Earth Owner, who controls wind and clouds. People cultivate fields for varying periods of time, depending on elevation and soil quality, before allowing brush to grow again. They store threshed maize in granaries or indoor grain bins prior to sale.

Again this sociocultural typology, like the previous agronomic one, is only gross, the elements constituting approximate cultural sets which have proved useful in organizing our ideas over many years. It continues to structure our thoughts in a preliminary fashion. There is some broad correlation between peoples' horticultural arrangements, that is how they produce food and other items necessary for life, and how they are socially organized, both to effect production and to achieve distribution of its fruits. But this is not to imply, as some writers argue, that mode of production determines the nature of social arrangements.

4. *Future Directions*

Several questions remain to be answered about horticultural societies. Many aspects are still poorly understood, and insight sometimes blinkered by orthodoxies. The Baliem Valley, for example, supports one of the highest population densities in New Guinea ($\sim$160 persons per km^2), with no evidence that it is pushing against the environmental limits. How can the region's intensive horticulture system support such a high population density without any outside fertilizer amendments? The accepted shifting cultivation model which pertains to this part of the world predicts evidence of environmental strain, of the kind commonly associated with overpopulation. Yet the sophisticated horticultural regime does not seem to exhaust the soil (Sillitoe 1996).

A major contributor to our understanding of horticultural societies has been 'farming systems research' conducted under the aegis of international development agencies (Shaner et al. 1982). This involves multidisciplinary terms, using survey techniques largely, to detail the ecological, technical, and socioeconomic aspects of horticultural regimes and identify constraints on production amenable to scientific research. It was initially associated with 'modernization,' research intended to advance the development of horticultural societies into agricultural ones, to enmesh them in commercial farming and specialized production for the market place. It was thought that this would alleviate poverty, leading to increased production of a food surplus necessary to support increasingly urban populations specializing in industrial, administrative, and service sectors. To this end, agricultural research institutes worldwide embarked on programs to breed new 'high-yielding varieties' of crops and develop associated technologies to increase production. The scientifically resoundingly successful 'green revolution' diminished the threat of famine. Nonetheless, it increasingly became evident that while the 'revolution' had benefited resource-rich and wealthy landowners, it had largely passed poor farmers by, increasing wealth differentials. It also worryingly increased dependency on outside manufactured inputs (inorganic fertilizers, fossil-fueled machinery, chemical pesticides, and herbicides) and gave rise to concern about long-term sustainability with evidence of increasing environmental pollution, susceptibility to mutated pathogens, insecure supply of needed inputs, and so on. The move towards cash cropping also put farmers at the mercy of market fluctuations, sometimes causing their families considerable hardship, further increasing resistance to modernization and adoption of commercial production. We become aware that 'modernization' not only scarcely helped poor households farming in marginal and fragile, resource-poor environments, but may have made some wealthier ones vulnerable too.

The realization that not only were some small-scale subsistence-focused horticultural societies proving very resistant to change as envisioned by modernization, but that other communities were increasingly exposed to serious perturbations beyond their control in attempting to accommodate its interventions, prompted a rethink of development priorities in the 1980s. It was increasingly acknowledged that poor people need to be involved in the planning and implementation of any development interventions intended to assist them, with the consequences of what international bankers euphemistically called economic restructuring taking effect (a dramatic instance of outside market forces impacting negatively on farmers' fragile commercial enterprises, as their nations' burgeoning debts were rescheduled). The 'participation' approach to development came of age. This promises to advance understanding and the interests of horticultural societies in the future, particularly with the concomitant emergence of 'indigenous knowledge' research, which represents the coming together of applied farming systems research and academic disciplines such as anthropology, geography, and ecology which have long esteemed understanding of horticultural societies (Warren et al. 1995). The assumption is that any development assistance needs to be sympathetic to local sociocultural and environmental contexts and to accommodate

peoples' aspirations. Understanding these implies their cooperation. It also furthers meaningful appreciation of the complex relations that inform natural resources management, as understood by local farmers, those who contend daily with them—the collapse of many interventions has been attributed to a failure to comprehend the complexity of their livelihood systems. These changes in research and development agendas bode well for furthering understanding in the future, and promoting the interests of horticultural societies *along their own lines*.

See also: Agricultural Change Theory

Bibliography

Allan W 1967 *The African Husbandman*. Oliver and Boyd, London

Altieri M A 1995 *Agroecology: The Science of Sustainable Agriculture,* 2nd edn. Westview Press, Boulder, CO

Bayliss-Smith T P 1982 *The Ecology of Agricultural Systems*. Cambridge University Press, Cambridge, UK

Carlstein T 1982 *Time Resources, Society and Ecology*. Allen & Unwin, London

Forde C D 1934 *Habitat, Economy and Society: A Geographical Introduction to Ethnology*. Methuen, London

Gliessman S R (ed.) 1990 *Agroecology: Researching the Ecological Basis for Sustainable Agriculture*. Springer Verlag, New York

Grigg D B 1974 *The Agricultural Systems of the World*. Cambridge University Press, London

Norman M J T 1979 *Annual Cropping Systems in the Tropics*. University Presses of Florida, Gainesville, FL

Ruthenberg H 1976 *Farming Systems in the Tropics*, 2nd edn. Clarendon Press, Oxford, UK

Sauer C O 1969 *Agricultural Origins and Dispersals*, 2nd edn. MIT Press, Cambridge, MA

Shaner W W, Phillipp P F, Schmehl W R 1982 *Farming Systems Research and Development*. Westview Press, Boulder, CO

Sillitoe P 1996 *A Place Against Time: Land and Environment in the Papua New Guinea Highlands*. Harwood Academic, Amsterdam

Warren D M, Slikkerveer L J, Brokensha D 1995 *The Cultural Dimensions of Development: Indigenous Knowledge Systems*. Intermediate Technology Publications, London

Watters R F 1971 *Shifting Cultivation in Latin America*. Food and Agriculture Organization, Rome

Wilken G C 1987 *Good Farmers: Traditional Agricultural Resource Management in Mexico and Central America*. University of California Press, Berkeley, CA

P. Sillitoe

Hospitals and Health: Psychosocial Aspects

During the second half of the twentieth century, the dominant mode of health care delivery in the United States has evolved from individual, patient–physician relationships to care provided by organizations (Starr 1982). At the same time, social scientists have moved from examining hospitals as closed systems, which operate independently of their environment, to seeing them as open systems, both affected by and responding to external forces (Scott 1998). This perspective assumes that structural forms and organizational behavior in the healthcare sector are influenced by factors such as regulation, change in professional power, competition among providers, and mutualistic behavior among different types of providers. The context in which hospitals operate has been described as having two essential components: the technical environment and the institutional environment. The technical aspects of an organization's task environment involve the production of products and services and the market forces that reward the efficient and effective production of those products and services. Institutional environments encompass the rules and requirements to which organizations must conform if they are to receive the social and political support and legitimization needed to survive.

Organizations in the health care sector have traditionally been subject to both strong institutional and strong technical pressures. These forces have operated in separate but parallel fashion to shape the practices and structures of hospitals (Alexander and Scott 1984). For example, hospitals traditionally maintained very loose control over physician staff as a reflection of strong professional requirements for autonomy. At the same time, they were accountable for the 'technical' quality of care through legal and quality assurance controls. Recently, market forces have played a more visible role in shaping national policy debate concerning healthcare access, quality, and costs. In view of this, some have argued that the influence of traditional institutional structures and practices, such as regulation, professional norms, and local control, have weakened and are now subordinated to the technical demands of the market. However, there is sustained evidence that historical institutional forces have not been eliminated, but instead are interacting with the new 'market' ethos to constrain and shape the structures and practices of hospitals, including those originating in the corporate sector. These interactions have given rise to a unique set of changes and practices that differ from both those common to the corporate sector and those common to hospitals in more institutionally dominant eras.

1. Institutional Environment of Hospitals

Institutional environments feature elaborate rules and requirements to which individual organizations must conform if they are to receive support and legitimacy. Attention is directed away from control and coordination of technical processes and towards conformity to externally defined requirements or regulations. The more institutionalized the environment of an

organization, the more its structures and procedures are governed by external controls. This control requires specific arrangements in the organization that are independent of contributions to technical performance (Scott 1995). For much of the twentieth century and up until today, hospitals and other healthcare organizations have been subject to strong institutional pressures, including a broad array of governmental regulations and requirements. Further, numerous professional specifications have governed what types of personnel might be hired, how tasks are distributed among them, and what procedures must be followed in performing these tasks.

The healthcare field in the US has long been characterized by institutional pressures that are both strong and heterogeneous (Stevens 1989). All branches of the federal government make decisions centrally that affect the entire healthcare sector. Local and state governments also exercise authority to regulate health care. This fragmented structure of authority makes it possible for various actors to pass laws to meet particular needs, and regulations are therefore uncoordinated, at best, and inconsistent or conflicting, at worst. Further, a long tradition of volunteerism in American hospitals has complemented the inculcation of professional norms into the institutional structure of these organizations. The notions of local control, community accountability, and philanthropic support of hospitals have served to keep decision-making from becoming overly centralized. At the same time, this has discouraged the formation of a cohesive, integrated healthcare system.

2. *Technical Environment of Hospitals*

Notwithstanding the institutional demands on hospitals, the technical requirements of modern medicine are considerable and often require tight internal controls and careful coordination if performance is to be effective. Hospitals have also increasingly been pressured to improve efficiency in the use of resources. Retrospective cost reimbursement was singled out as the primary explanation for both rising costs and inappropriate utilization. Critics also identified professional specialization and hospital-based care as major contributors to rising costs. Growing specialization and fragmentation in delivery rendered services increasingly impersonal, while the predominance of inpatient services threatened accessibility and provision of comprehensive, family-oriented care.

Historically, political attention had focused on expanding access to health care, as exemplified by the passage of legislation in the 1960s extending health insurance coverage to many of the elderly (Medicare) and the poor (Medicaid). The 1980s, however, marked the beginning of a new era that shifted the focus to controlling costs (Scott et al. 2000). Several years of double-digit increases in prices threatened the economic interests of actors who paid for health care; in response they initiated policies and practices to control costs. In 1984, the federal government's Medicare program became the first major payer to attempt to control hospital costs by paying a fixed fee for each service, rather than the 'usual and reasonable' fees that hospitals routinely charged. This key change in payment policy had significant financial and market effects for hospitals. The resulting reductions in payments to hospitals and other organizational providers cut their budgets and contributed to cash flow problems. Hospitals and other providers tried to compensate by increasing their patient and service volume. In turn, this created more competition for patients.

A series of remedies have been proposed to deal with the cost escalation crisis. The solutions imposed have often clashed with the historical institutional functions of hospitals, namely, providing care to the community, running on a voluntary basis and engaging in philanthropic enterprise. Direct regulatory approaches to containing costs in the 1970s have shifted to greater emphasis on legal, budgetary, and market responses to addressing these issues. Taken as a whole, these remedies fall under the general rubric of a competitive approach to controlling costs, even though all involved substantial government oversight and occasionally intervention. Such approaches are consistent with a general societal shift toward privatization and market solutions to larger social problems. While recently court rulings have relaxed some regulations stifling collaboration among doctors and organizational providers, new regulations have reduced price-fixing and other forms of anticompetitive behavior. Together these rulings have provided both inducements and constraints on organizational behavior, such as severe limitations on hospital acquisitions and mergers.

Budgetary remedies have been applied at both the federal and state levels. At the federal level, the government has budgeted a fixed reimbursement for each inpatient admission based on diagnosis. States have also enacted rate-setting programs that prospectively set reimbursement levels for hospitals and other providers. These budgetary interventions have forced healthcare providers to attend to efficiency and have raised the level of competition among these providers. Finally, market remedies have included the introduction of so-called 'managed care' strategies, designed to reduce utilization and thereby costs (see *Managed Care*). Examples of these strategies have included competitive bidding and selective contracting in government programs and the use of preferred providers and risk-based financial incentives by private insurance firms. Both state and federal governments have begun to publish information on the performance of healthcare providers, aimed at redressing the traditional asymmetry of information between providers and consumers of health care.

As expected, these interventions have lowered healthcare costs and decreased utilization (Robinson 1996). But there have also been other, unanticipated effects. At the field level, the 1990s have witnessed an increasing number of entries and exits of new and existing organizational forms into the healthcare system. Although most hospitals and other organizational providers that have closed have tended to be primarily inner-city and rural institutions and have done so for primarily technical/rational reasons, many others have remained operating despite financial losses and technical inefficiencies. The survival of so many marginal institutions in the face of increasing pressures from a market-based system clearly reflects the importance of local/community support for these institutions, and the salience of legitimacy in organizational survival (Rohrer 1989). This illustrates well the complex combination of technical and institutional forces that shape organizational life in the healthcare sector.

3. *Hospital Responses to Hybrid Environments*

A fundamental question becomes, how do organizations operating within such hybrid environments accommodate different environmental demands that are often diametrically opposed to each other? The convergence of traditional institutional and new technical forces impinging upon hospitals has created new organizational, strategic, and survival challenges for healthcare providers. Neo-institutional theory suggests that organizations functioning in sectors that are highly developed both institutionally and technically will develop more complex and elaborate administrative systems and will experience higher levels of internal conflict (Scott 1995). Historically, we know this to be true for many organizations in the healthcare sector. Hospitals, for example, have typically operated within a dual hierarchy of control, represented on the one hand by hospital administration and on the other by hospital medical staff and physicians. These dual lines of control operate in parallel with each other but rarely have intersected in such a way as to disrupt the relative spheres of influence over issues related to control, resource allocation and professional prerogatives.

Conflict has been expressed in hospitals largely through increasing parity between managerial and professional logic and ideologies. The emphasis of managerial ideology on rational systems and cost-effective care has increasingly conflicted with the professional orientation toward individual doctor/patient relations and the micro delivery of specific medical services (Goodrick et al. 1997). Conflicting claims to the right way to deliver health care, the presence of alternative rationales for organizing work (professional vs. managerial), and the inherent conflicts between technical and institutional pressures in the healthcare sector have eroded the legitimacy of hospitals. This has opened the way for a number of new organizational forms, practices, and strategies, particularly those that have symbolically represented the new ethos of privatization and corporate rationality. For example, health organizations are increasingly lobbying government and engaging in marketing campaigns in order to influence legislation and public opinion. Rather than being passive recipients of institutional changes in the broader environment, they seek actively to shape societal beliefs toward health care and the way health care is organized.

4. *Organizational Heterogeneity*

Health care delivery has been dominated traditionally by hierarchical organizations such as hospitals and healthcare systems. However, as competitive pressures and the demands of managing care under limited resources have increased over the past decade, a wide variety of new organizational forms have emerged (see *Health Care Delivery Services*). These range from managed care organizations, physician–hospital arrangements, and physician practice management companies, to diagnosis-specific service contracts with specialized providers, such as home health care and rehabilitation firms.

As traditional institutional norms and practices have eroded and have been replaced by a stronger corporate rationality, we have observed greater heterogeneity among organizations in the healthcare sector. This new ethos suggests there is no one best way to organize in order to achieve technical efficiency and positive performance in the marketplace. More variation in organizational forms is exhibited as hospitals become less concerned with mapping the characteristics of the institutional environment onto their own structures, and more concerned with achieving technical efficiencies and buffering their technical cores from environmental disturbances.

A clear response to the new market forces in health care has been the adoption of structures and practices from the corporate sector (Burns 1990). These structures include multidivisional organizational designs, parent holding-company arrangements, contract management of facilities, and vertical integration across a number of different service lines. While these practices clearly have gained popularity during the 1990s, there is limited evidence that they actually improve the quality or lower the cost of care. Neo-institutional theorists have argued that whereas early adopters of these corporate structures may have done so for technical/rational reasons, later adopters may have viewed them as the prevailing institutional practices under the new market-based health care environment.

5. Hospital Governance

Historically, most nonprofit hospitals were overseen by autonomous governing boards composed largely of representatives of the communities the hospitals served. This type of governance arrangement reflected a longstanding institutional belief that the control of nonprofit hospitals should be vested in local communities. However, during the 1980s and 1990s, two developments have altered this traditional governance model of nonprofit hospitals, both of which are part of a larger trend toward the so-called 'corporatization' of health care. One development is the consolidation of over 40 percent of US hospitals into complex healthcare systems. In such systems, the locus of control for member hospitals is shifted to varying degrees outside local communities to system-level governing boards and corporate offices (Toomey and Toomey 1993). Another development is corporate restructuring of hospitals where hospital assets and decision-making authority are transferred to multiple corporate entities that may include a parent/holding company and various subsidiaries. Such restructuring often entails the creation of multiple governing bodies that can attenuate the decision-making authority of the provider's community-based board.

Under a traditional model of governance, nonprofit hospitals resemble consumer cooperatives in that community representatives are in a position to determine hospital policies. As such, a traditional form of governance is arguably a constraint on hospital price inflation, since community representatives have an incentive to resist price increases that they ultimately will incur in the form of higher insurance premiums. Also, form of governance may impact the accessibility of hospital services since a community-based board may be inclined to invest resources to expand the accessibility of hospital services to its own community (e.g., through the provision of uncompensated care). Policymakers have expressed concern that system membership and corporate restructuring are diminishing the role of communities in the control of nonprofit hospitals, and government agencies that have oversight responsibilities for nonprofit hospitals are currently examining alternatives (see *Health Care Organizations: For-profit and Nonprofit*).

6. Physician–Organization Structures

As managed care and competitive pressures have compelled hospitals to consolidate and provide services along a broad continuum of care, the need to integrate formerly loosely coupled elements of these organizations has increased. Perhaps the most problematic aspect of such integration is the alignment of incentives and strategic goals between delivery organizations and physicians. Hospitals seek to engage in strategic relations with physicians in order to: (a) protect against the risks of agency from physicians who have traditionally borne little risk for the care of patients, (b) reduce transaction costs by developing economies of scale and integrating physicians into the structure of the delivery organization, (c) achieve market power both up and down stream over suppliers and customers respectively, and (d) develop capabilities that differentiate systems from competitors in the areas of innovation, technology, and quality. While these market-driven reasons are clearly critical to hospitals and healthcare systems, the alignment of economic and strategic interests between organizations and physicians is difficult given the differences in cultural context, socialization, orientation to patient care, and economic incentives.

As managed care organizations and health care regulators have recognized the need to align the economic incentives of providers with the goals of cost containment and high quality, they have changed methods of provider compensation from simple discounting to risk sharing (e.g., capitation, *per diem*, or per case reimbursement for hospitals). Providers under these systems assume actuarial and utilization risk and reward. In principle, risk sharing promotes appropriate behavior, since high utilization and costs will result in deficits for either the physician group or its physicians. In practice, the physician group's form of compensation, incentives, and governance moderate these effects. For example, the incentives inherent in risk sharing can be lost if individual physician compensation is not strongly linked to holding costs below the group's fixed budget. Alternatively, if physicians feel they have no say in the formulation of incentives in compensation systems, they may be less willing to abide by them. These examples suggest that the incentives created by risk sharing are effective only under certain conditions. Several authors have recognized that the effects of such payment mechanisms may partially depend on organizational factors such as type of governance, physician leadership, information dissemination, sanctions, and organizational culture (Kralewski et al. 2000, Alexander et al. 2000).

Physicians and organizational healthcare providers recognize the need to work together to acquire managed care contracts and to manage risk. To facilitate such contracting, avoid institutionalized constraints inherent in hospital-medical staff relations, and develop the administrative infrastructure necessary for managing service delivery, many healthcare systems have established market-based organizational arrangements that link the delivery organization with one or more physician groups. Market-based arrangements are different from internal hospital structures in that they tend to be formally structured, contractual in character, corporate in organization, and include physicians outside the formal membership of the medical staff.

Given fundamental differences between methods of organizing by physicians and delivery organizations, however, it is highly questionable whether structural solutions to achieving greater integration between physicians and delivery organizations can be completely effective. There is only limited evidence to suggest that structural solutions work and considerably more evidence that self-selection by physicians into closely aligned structural arrangements may account for more of the explained variance in alignment than structure itself. Hospitals' attempts to integrate and align with physicians have not always rationalized the delivery of health care, but instead in many cases have made it more complex. Physicians and their interests vary widely, and hospitals have discovered that no one structural vehicle best accommodates the interests of all physician subgroups. Consequently, it is not uncommon for multiple market-based arrangements to exist within a single organization. This makes the coordination of care and alignment of incentives at least as difficult, if not more so, as it has been under the traditional medical staff model.

7. *Organization of Work in Hospitals*

Largely as a result of increasing competitive pressures and external emphases on cost containment, the organization of work within hospitals has shifted. Two trends are particularly in evidence: (a) the use of teams to perform managerial and clinical work, as opposed to performing such work individually, and (b) the use of contracts to obtain labor externally rather than to hire employees.

Because of demands by managed care firms and integrated healthcare systems to provide comprehensive care to patients and/or enrollees in health plans, there is growing recognition that patients need comprehensive care in order to address their multiple problems effectively. Given the broad spectrum of knowledge necessary for comprehensive care, there is growing belief that teams of practitioners with complementary roles and skills are best suited to providing high quality, cost-effective patient care (Fried and Rundall 1994). Teamwork in the clinical round is also perceived as an antidote to fragmentation in services, a traditional problem in healthcare delivery. For example, repeated patient assessments by various healthcare providers add unnecessary costs to patient care, when one combined assessment would suffice. Advocates of team-oriented care cite studies that consistently show that collaboration among physicians, nurses, and other healthcare personnel enhances the quality of care and patient satisfaction (Nikolaus et al. 1999, Le et al. 1998).

Team-based care stands in marked contrast to the traditional hierarchical work relations in most hospitals, which accord the physician a role of absolute autonomy and control over patient-care processes. The growing emphasis on prevention, wellness, and cost-effectiveness in care delivery make this hierarchical structure outmoded. Flatter structures and more collegial relations among providers appear to be the developing norm. For example, many nurses advocate a collaborative relationship with physicians where nurses have primary responsibility for health maintenance, prevention, and diagnosis and treatment of routine cases, while physicians retain responsibility for the diagnosis and treatment of complex cases and the prescription of medications.

Contracting for labor services as an alternative to employment stems from the need for more flexibility in an increasingly competitive environment. Hospitals need to be able to make rapid changes in their workforce in response to demands from a volatile market. Rather than staffing to accommodate periods of peak demand with attendant down time for slack periods, hospitals have increasingly adopted a contingent approach to staffing, responding to the particular needs of the organization at a given time (Katzman 1999, Rizzuto et al. 1999). Contracting typically occurs in housekeeping, business, and clinical services. Proponents argue that contracting frees top managers to concentrate on core services, brings specialized expertise to the organization, improves flexibility through the ability to terminate contracts, reduces supply costs through superior purchasing power, and promotes the recruitment of specialists who may be in short supply. Critics argue, however, that the contract-management firm and the contracting organization may fail to clarify performance goals or that contracted staff may have difficulty working with other departments and full-time personnel. Perhaps most importantly, some believe that contracting for a wide variety of services without a clear strategy risks damaging organizational coherence and control (D'Aunno et al. 1996).

8. *Quality of Care in Hospitals*

Another illustration of the organizational response to technical and institutional pressures lies in the area of healthcare quality. On the one hand, decline in public support and trust for hospitals has reduced their impetus to assure quality and discipline members of their organizations for lapses in quality. Responsibility for quality has begun to shift to the payers of health care and their agents (e.g., managed care organizations), as well as to the courts which adjudicate an increasing number of malpractice suits against hospitals and physicians. These trends reflect a growing tendency to think of health care as a commodity rather than an entrusted service provided by professionals and professional organizations. On the other hand, there is greater emphasis placed on quality

improvement systems within organizations as a counterbalance to external controls over quality and as a response to market pressures to increase the efficiency and effectiveness with which care is delivered. Again, this duality of external control vs. internal control is a juxtaposition based in the competing demands placed on hospitals by institutional and technical forces.

Hospitals, in particular, are being challenged by increasingly competitive markets and regulatory changes to produce high quality care without raising costs. Given the multiple pressures and problems associated with healthcare quality, hospitals face the additional challenge of developing policies, procedures, and incentives to improve quality within a domain that has traditionally been the province of highly autonomous professionals. Quality improvement (QI) practices have emerged in recent years partially as an alternative to quality assurance and as a reaction to the shortcomings of outcome-based control. With mounting pressure to maintain or improve quality in the face of reduced resources, hospitals have begun to look to process improvements as a mechanism to provide high quality care while containing healthcare costs. While the adoption of QI appears to be an internally generated response to changing market conditions, for the most part, accrediting bodies such as the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) have also encouraged hospitals to adopt QI.

Given the traditional emphasis on provider–patient relations, individualized approaches to improving quality have received the vast majority of attention to date. Although the knowledge and practices of individual physicians are important for high quality care, increasingly clinicians practice within groups or systems of care. As such, the quality of care delivered depends critically on the ability of these groups or systems to prevent errors, coordinate care among settings and practitioners, and ensure that relevant, accurate information is available when needed.

One of the most promising recent developments is the effort to institutionalize quality improvement in hospitals as a coordinated, organization-wide approach to care strategies (Berwick et al. 1990, Gaucher and Coffey 1993, Carman et al. 1996). These practices, collectively, are designed to systematically examine and improve processes of care and care support in organizational settings. Typically, these quality improvement efforts are strongly rooted in evidence-based procedures and rely extensively on data collected from hospitals about care processes and patient outcomes.

9. *Hospital Behavior in Other Nations*

The focus of this article has been to associate change in hospital structure and behavior to change in the broader market and institutional contexts in which these organizations operate. The United States has been prominently featured in this discussion because of the somewhat unusual combination of strong technical and institutional pressures to which hospitals must respond. Other countries, however, may diverge markedly in the degree to which either institutional or market forces play a role in shaping hospital behavior. For example, in countries where healthcare delivery is highly centralized and funded exclusively by the government, we would expect to find much more structural homogeneity among hospitals, in marked contrast to the increasing diversity and differentiation among US hospitals. In other nations, where privatization and multiple sources of healthcare funding predominate, hospitals are more likely to resemble in their structure and behavior other commercial vendors of services to the market.

Recent trends in OECD countries would seem to indicate that they are facing technical pressures for increased efficiency similar to those in the United States (Normand 1993). Significant reductions both in the number of hospital beds per 1,000 and in average length of hospital stay have resulted, particularly in Scandinavian countries. In the countries of the former Soviet Union, on the other hand, average length of stay has not dropped, despite mounting financial strain due to economic instability. One potential explanation for this is lingering institutional norms regarding appropriate clinical practice (Hensher et al. 1999). This may illustrate how specific institutional beliefs about health and health care in a given nation may shape hospital behavior in ways that differ substantively from the ways in which institutional forces shape hospital behavior in the US. For example, in Taiwan, the deeply seated belief that family is the primary care giver is reflected in the hospital practices of incorporating family members in patient education, extended visiting hours, and even boarding with the patient during the hospital stay. Cross-national differences in hospital structure and behavior are thus likely to be a reflection of both the structure and content of the broader environment from which hospitals obtain patients, critical resources, and legitimacy.

See also: Doctor–Patient Interaction: Psychosocial Aspects; Gender and Health Care; Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care, Rationing of; Health Care Systems: Comparative; Health Care Technology; Health Professionals, Allied; Medical Profession, The

Bibliography

Alexander J A, Scott W R 1984 The impact of regulation on the administrative structure of hospitals. *Hospital and Health Service Administration* **29**: 71–85

Alexander J A, Vaughn T, Burns L 2000 Organizational, strategic and marketing determinants of operating policies of physician organization arrangements. *Health Services Management Research* **13**: 1–15
Berwick D M, Godfrey A B, Reossner J 1990 *Curing Health Care: New Strategies for Quality Improvement*. Jossey-Bass, San Francisco, CA
Burns L R 1990 The transformation of the American hospital: From community institution toward business enterprise. *Comparative Social Research* **12**: 77–112
Carman, J M, Shortell S M, Foster R W, Hughes E F, Boerstler H, O'Brien J L, O'Conner E J 1996 Keys for successful implementation of total quality management in hospitals. *Health Care Management Review* **21**(1): 48–60
D'Aunno T, Alexander J A, Laughlin C 1996 Business as usual? Changes in health care's workforce and organization of work. *Hospital and Health Service Administration* **41**: 3–18
Fennell M, Alexander J 1993 Perspectives on organizational change in the US medical care sector. *Annual Review of Sociology* **19**: 89–112
Fried B, Rundall T 1994 Managing groups and teams. In: Shortell S M, Kaluzny A D (eds.) *Health Care Management: Organization Design and Behavior*, 3rd edn. Delmar Publishers, Albany, NY
Gamm L D 1992 Health care markets as interorganizational fields: A conceptual perspective. *Health Services Management Research* **5**(1): 44–53
Gaucher E J, Coffey R J 1993 *Total Quality in Healthcare: From Theory to Practice*. Jossey-Bass, San Francisco, CA
Goodrick E, Meindl J R, Flood A B 1997 Business as usual: The adoption of managerial ideology by U.S. hospitals. *Research in the Sociology of Health Care* **14**: 27–50
Hensher M, Edwards N, Stokes R 1999 International trends in the provision and utilization of hospital care. *British Medical Journal* **319**: 845–8
Katzman C N 1999 Outsourcing keeps growing. Contract management survey finds times are good for firms serving healthcare industry. *Modern Healthcare* **29**(35): 42, 44, 46–50
Kralewski J E, Rich E C, Feldman R, Dowd B E, Bernhardt T, Johnson C, Gold W 2000 The effects of medical group practice and physician payment methods on costs of care. *Health Services Research* **35**(3): 591–613
Le C T, Winter T D, Boyd K J, Ackerson L, Hurley L B 1998 Experience with a managed care approach to HIV infection: Effectiveness of an interdisciplinary team. *American Journal of Managed Care* **4**(5): 647–57
Nikolaus T, Specht-Leible N, Bach M, Oster P, Schlierf G 1999 A randomized trial of comprehensive geriatric assessment and home intervention in the care of hospitalized patients. *Age and Ageing* **28**(6): 543–50
Normand C 1993 Comparisons among national health care systems in the European marketplace. *Hospital Formulary* **28**(Supp.1): 6–11
Rizzuto G T, Yallapragada R R, Roe C W, Authement J P 1999 Contract labor use in healthcare. *Hospital Topics* **77**(4): 15–19
Robinson J C 1996 Decline in hospital utilization and cost inflation under managed care in California. *Journal of the American Medical Association* **276**(13): 1060–4
Rohrer J E 1989 Closing rural hospitals: Reducing 'institutional bias' or denial of access? *Journal of Public Health Policy* **10**(3): 353–8
Scott W R 1995 *Institutions and Organizations*. Sage, Thousand Oaks, CA
Scott W R 1998 *Organizations: Rational, Natural and Open Systems*, 4th edn. Prentice-Hall, Upper Saddle River, NJ
Scott W R, Ruef M, Mendel P J, Caronna C A (eds.) 2000 *Institutional Change and Healthcare Organizations: From Professional Dominance to Managed Care*. University of Chicago Press, Chicago
Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York
Stevens R 1989 *In Sickness and in Wealth: American Hospitals in the Twentieth Century*. Basic Books, New York
Toomey R K, Toomey R E 1993 The role of governing boards in multihospital systems. *Health Care Management Review* **18**(1): 21–30

J. A. Alexander and M. E. Comfort

Hotelling, Harold (1895–1973)

Harold Hotelling (1895–1973) was one of a handful of statistical researchers who in the first half of the twentieth century developed the solid intellectual foundations of the then newly developing discipline of statistics. In addition, he was also an important contributor to several areas of mathematical economics. His academic career includes: professorships at Stanford University, Columbia University, and the University of North Carolina at Chapel Hill, and his professional leadership involved major statistical and economic societies. He is perhaps best known for his statistical work in multivariate analysis, introducing the concepts of principal components, canonical correlations and Hotelling's-T^2. In mathematical economics, he contributed to natural resource economics, spatial economic competition, and introduced the welfare equilibrium proposition.

1. Early Life

Born on September 29, 1895 in Fulda, Minnesota, Harold Hotelling, whose parents were Clair Hotelling and Lucy Rawson, was the oldest of six children. Having moved with his family to Seattle at age nine, Hotelling studied journalism at the University of Washington, taking time from his studies to work on a weekly paper. In 1919, he received his BA in Journalism, graduating Phi Beta Kappa. Encouraged, in part, by the mathematician and well-known historian, Eric Temple Bell, he went on to graduate study in mathematics at the University of Washington, receiving an MA in 1921. It was during this time that he married Floy Tracy in 1920. He received his Ph.D. in Mathematics in 1924 from Princeton University, doing dissertation research on dynamic topology with Oswald Veblen, although his original intent was to study mathematical economics. His interest in economics took root while growing up and in his belief in

'the importance … of a study of economics [in] that [it] might possibly impress some people to the point of making some changes in the institutions with which they had become familiar [and] which I [Hotelling] was already prepared to condemn vigorously.' Alternatively, he 'couldn't escape the lure of mathematics' and notes that he 'hoped and [it] appeared that the mathematics would be extremely prolific of new economic theorems and ideas if followed out' (Hotelling 1963).

2. *Beginning at Stanford University*

Invited to join the Food Research Institute at Stanford University, Hotelling was a research associate there from 1924 to 1927, following which he was an associate professor in Stanford's Department of Mathematics (1927–31). An early indication of the breadth of his intellectual interests is illustrated by the fact that in 1927 he taught courses on mathematical statistics, differential geometry, and analysis situs (a forerunner of topology), and in that year published papers on the application of the latter to statistics, and on differential equations subject to error. While working in the Food Research Institute, he was introduced to the work of R. A. Fisher (five years his elder), and through their mutual correspondence they became friends. In fact, Hotelling wrote the book review in the *Journal of the American Statistical Association* (Hotelling 1927) of the first edition of Fisher's *Statistical Methods for Research Workers* (and also did subsequent reviews there of the Second through Seventh Edns.). After the planned visit of Fisher to Stanford in the summer of 1928 was abandoned, Hotelling visited Fisher at the Rothamsted Experimental Station during the second half of 1929. Their connections continued and at the outbreak of World War II, Hotelling, along with three others, wrote to Fisher volunteering to care for Fishers' six daughters in the United States during the war. However, a last minute cancellation kept the Fisher children in Britain (Box 1978, p. 377).

While at Stanford, Hotelling wrote a number of path-breaking papers in statistics and in mathematical economics. The most notable among these statistical results was his generalization of Student's t-ratio to handle multivariate response data (Hotelling 1931). This generalization remains known as Hotelling T^2. His paper with Holbrook Working (Working and Hotelling 1929) on modeling and linear prediction first demonstrated that the standard error of the estimate of a linear regression depended on the independent variable's value, thereby giving rise to the well-known Working–Hotelling curved simultaneous confidence regions for a linear regression. His research contributions to mathematical economics were equally notable. His 1929 paper provided a game-theoretic solution given by two differential equations to a problem of price and location competition between two entities in a spatial setting. His 1931 paper on the economics of exhaustible resources is recognized as the seminal paper in the field of natural resource economics. In that paper, Hotelling demonstrates that in an equilibrium setting, prices of exhaustible natural resources will tend to rise over time at a percentage rate equaling the interest rate.

3. *Columbia University and the Statistical Research Group*

Hotelling moved to Columbia University in 1931 as a Professor of Economics, having been recruited to further the development of mathematical economics, and also to initiate mathematical statistics. Throughout his career at Columbia and later at the University of North Carolina at Chapel Hill, Hotelling was superlative at both attracting outstanding researchers and mentoring young researchers destined for stardom. The Nobel-prize winning economist, Milton Friedman, was among the earliest mentored by Hotelling, spending 1933–4 at Columbia learning mathematical economics from him. Later mentoring Kenneth Arrow, another Nobel-prize winning economist, Hotelling was also instrumental in attracting Abraham Wald, W. Allen Wallis, and Jacob Wolfowitz to the statistics facility at Columbia. As a teacher, Hotelling was perceived to be organized and comprehensive, interweaving theory with application. Importantly, an appreciation of research was imparted to his students, and he was able at an interpersonal level to instill in his students their own confidence to be successful at research (Madow 1960).

Hotelling was one of three original voting fellows of the Institute of Mathematical Statistics on September 12, 1935, and was President of that Institute in 1941. He began his association with the Cowles Commission, participating in its first Summer conference in 1935, and served as President of the Econometric Society during 1936–7. In 1939–40, he was a Visiting Lecturer at the Indian Statistical Institute and was President of the Indian Statistical Congress (1940).

Along with Wallis and Wolfowitz, Hotelling was a charter member of the Statistical Research Group (SRG), at Columbia, for which Wallis was appointed Director of Research. The SRG, founded on July 1, 1942, and in existence until September 30, 1945, was an extraordinary research group of statisticians whose goals were to provide research support and analyses for the armed forces in furthering the quality and accuracy of their war efforts. The origins of sequential analysis stem from Wald's work for the SRG, and among Hotelling's contributions were techniques for control charts for multivariate data (Kruskal 1980). (A detailed history of the SRG is provided by Wallis 1980).

Early in Hotelling's career at Columbia, his wife, Floy, died in 1932, leaving behind two small children.

In 1934, Hotelling married Susanna Edmundson and together they had six children, five sons and a daughter (who died in her infancy).

Hotelling, during his prewar years at Columbia, continued to produce ground-breaking research in statistics and economics. His 1933 and 1936 papers introduced and lay the groundwork for principal component analysis as a multivariate statistical technique for simplifying the structure of large numbers of correlated measures (Hotelling 1933, 1936a). In 1936, he extended notions of multiple correlation to canonical correlations in order to measure and study the relationships between two dependent sets of correlated responses (Hotelling 1936b). In the field of economics, his 1935 paper provided a consistent theory uniting demand and utility by means of a constrained optimization problem and his presidential address to the Econometric Society in 1938 (Hotelling 1938) introduced a welfare equilibrium proposition which leads in certain societal settings to the argument that goods be priced at their marginal costs. In 1940, his innovative and insightful paper on the teaching of statistics in universities was published (Hotelling 1940). In this prescient paper, Hotelling observes the inevitability of the increased demand for statisticians, and argues that the teaching of statistics in universities would be best done solely by statisticians in separate departments of statistics (a quite novel idea in 1940!). With regard to this article, Jerzy Neyman (1960) later wrote 'in my opinion the establishment of the excellent departments of statistics in such universities as Chicago, Columbia, Harvard, North Carolina, and Stanford is due to a considerable extent to the publication of Hotelling's article.'

4. *Career at University of North Carolina*

After the war in 1946, Hotelling was recruited by Frank Graham, President of the University of North Carolina at Chapel Hill to start the Department of Mathematical Statistics there upon his arrival. In addition to being appointed professor of mathematical statistics and professor of economics, he was named Associate Director of the Institute of Statistics (directed by Gertrude Cox), which was constituted with Chapel Hill's department and the Department of Experimental Statistics at North Carolina State, as well as related groups in psychometrics, social sciences and biostatistics. It is notable that when Hotelling was planning to go to Chapel Hill, he proposed that Wald join him there. In order to entice Wald to remain at Columbia, a Department of Mathematical Statistics was promptly proposed and officially approved in December 1946 (Hoxie 1955). As at Columbia, Hotelling attracted outstanding statisticians to the Chapel Hill faculty including R. C. Bose, Wassily Hoeffding, P. L. Hsu (who had been at Columbia with Hotelling), William Madow, Herbert Robbins, S. N. Roy, and Walter Smith. Hotelling relinquished the chairmanship to George Nicholson in 1952, and devoted the remainder of his career there to his research and his students.

His research at North Carolina involved, on the whole, elaborations and in-depth further study of some of his earlier research. Noteworthy among these is his work in 1947 and 1951 on hypothesis tests related to multivariate analysis of variance, which lead to the testing criterion referred to as the Lawley–Hotelling trace. He formally retired in 1966, but continued to be active in the department. He was elected to the National Academy of Sciences in 1970, and received a number of other academic honors during this period. Subsequent to a serious stroke in mid-1972, Harold Hotelling died in Chapel Hill, North Carolina on December 26, 1973.

5. *Hotelling's Contribution to Multivariate Statistical Analysis*

One of the origins of statistical hypothesis testing is a test that the mean μ of a normal distribution has a specified value μ_0. The resultant Student's t-test, obtained in 1908, remains as one of the most used statistical procedures. The t-statistic is

$$t = \sqrt{N}(\bar{x}-\mu_0)/s$$

where $\bar{x}$ and s, respectively, represent the mean and standard deviation of a sample of size N. Hotelling recognized that many experiments will have correlated measures $x_1, \ldots, x_p$, which arise from a multivariate normal distribution, and he provided the Hotelling's T^2 test to handle the means of multiple measurements (Hotelling 1931). For a sample of size N let $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_p)$ denote the sample mean vector and $S = (s_{ij})$ represent the sample covariance matrix.

The T^2-statistic is

$$T^2 = N(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)\,S^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)'$$

where $\boldsymbol{\mu}_0 = (\mu_{10}, \ldots, \mu_{p0})$ is the specified mean vector in the hypothesis. When the hypothesis is true, $(N-p)T^2/p(N-1)$ has an F-distribution with p and $N-p$ degrees of freedom.

Analogous to Student's t-test, the T^2-statistic can also be used to test for the equality of the mean vectors of two normal distributions that have the same unknown covariance matrix but unequal means, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. Now the statistic becomes

$$T^2 = \frac{N_1 N_2}{N_1+N_2}(\bar{\boldsymbol{x}}-\bar{\boldsymbol{y}})\,S^{-1}(\bar{\boldsymbol{x}}-\bar{\boldsymbol{y}})'$$

where $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ are mean vectors from samples of size N_1 and N_2, respectively, and S represents the pooled sample covariance matrix. Although there are many parameters in this model, under the null hypothesis the

distribution of $(N_1+N_2-p-1)T^2/p$ has an F-distribution with p and N_1+N_2-p-1 d.f.

The T^2-statistic additionally arises as a linear discriminant, as noted by Fisher, and is the sample analog of the Mahalanobis squared distance.

Quite remarkably, Hotelling noted that for the two-sample problem, the T^2 statistic is invariant under the transformation $\bar{\boldsymbol{x}} \to \bar{\boldsymbol{x}}A$, $\bar{\boldsymbol{y}} \to \bar{\boldsymbol{y}}A$, where A is any nonsingular matrix. This anticipated a general theory of statistical invariance developed much later. In particular, the T^2-test has a number of optimal properties among invariant tests: (a) the T^2-statistic is the only invariant function of the sufficient statistics $\bar{\boldsymbol{x}}$ and S; (b) as a test statistic it is the uniformly most powerful test among all invariant tests; and (c) the noncentral distribution depends on the scalar parameter $(\boldsymbol{\mu}-\boldsymbol{\nu})\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\nu})'$, which is the only invariant of the parameters.

Large data sets either in exploratory data analysis, data mining, or model building are accompanied by a large number of variables. The reduction of numbers of variables is often a first step in trying to uncover patterns in the data. Because multivariate measures $x_1, \ldots, x_p$ are related, there is the need to disentangle the correlational structure. To accomplish this, Hotelling (1933, 1936a) introduced the p principal components $z_1, \ldots, z_p$, which are linear combinations of the underlying measures, namely $z_k = \Sigma_{i=1}^{p} a_{ki} x_i$, $k = 1, \ldots, p$, where the coefficients are normalized by $\Sigma_{i=1}^{p} a_{ki}^2 = 1$. The first set of coefficients $a_{11}, \ldots, a_{1p}$ is chosen so as to maximize the variance of z_1. The second set of coefficients $a_{21}, \ldots, a_{2p}$ is chosen to maximize the variance of z_2 subject to the constraint that z_1 and z_2 are uncorrelated. At each step the variance is maximized, but with the additional constraints that each z_k be uncorrelated with $z_1, \ldots, z_{k-1}$. Their ordered variances $\mathrm{Var}(z_1) \geqslant \cdots \geqslant \mathrm{Var}(z_p)$ turn out to be the characteristic roots of the population covariance matrix.

In effect, principal components provide transformed coordinates that are uncorrelated, thereby disentangling the measures. Because the principal components are ordered according to their variances, this permits the researcher to focus on a subset of principal components that contributes substantially to the total variance. In multiple regression, the predictor variables often are closely related, and the problem of multicollinearity results. Principal components is one important procedure that introduces orthogonality and reduces dimensionality, thus removing multicollinearity.

To build a hierarchy of correlational measures one can start with the ordinary Pearson product moment correlation between x and y. With multiple measures $x_1, \ldots, x_p$ and y one generates the multiple correlation of y and the linear combination $y^* \equiv \Sigma a_i x_i$, which is the maximum correlation between y and y^*. Suppose now there are multiple measures $x_1, \ldots, x_p$ and $y_1, \ldots, y_q$. To analyze the relation between these two sets of variates, Hotelling (1936b) defined canonical correlations between u_k and v_k, which are the correlations of linear combinations $u_k = \Sigma a_{kj} x_j$ and $v_k = \Sigma b_{kj} y_j$ where $k = 1, 2, \ldots, K \equiv \min(p, q)$. The coefficients are normalized so that the variances of u_k and v_k are 1, and are chosen so as to maximize the correlations, and such that $u_1, \ldots, u_k$ are mutually uncorrelated and $v_1, \ldots, v_k$ are mutually uncorrelated $k = 1, \ldots, K$. The new variables u_k and v_k, $k = 1, \ldots, k$ are called the canonical variables; furthermore, it turns out that u_i and $v_j (i \neq j)$ are uncorrelated. The discovery of the corresponding matrix factorization is quite remarkable. A special case of this factorization leads to the singular value decomposition, which is a very important representation in numerical linear algebra.

As a result of this decomposition, one can transform the original variables to new coordinates $(u_1, v_1), \ldots, (u_K, v_K)$, that exhibit a simple structure: (a) the pairs are uncorrelated; (b) the pairs are ordered decreasingly in accordance to their correlations (which are non-negative); that is, (u_1, v_1) are maximally correlated; and (c) the variances of the u_k's and v_k's are normalized to be one. In a 1935 paper, Hotelling had earlier focused on the first pair (u_1, v_1), and he called u_1 the most predictable criterion and v_1 the best predictor.

Principal components and canonical correlations can be viewed in terms of roots of a determinantal equation and estimated from their sample analogues. It was this context that opened an entire area of statistical research on the distribution of the roots of random determinantal equations. In 1939, a number of papers appeared simultaneously in the statistical literature in which the central distribution of the roots were obtained. From then until the present, there is a continuing flow of theory concerning the noncentral distributions of roots of sample determinantal equations.

Hotelling's training as a mathematician provided him with a warehouse of tools that he applied to statistical problems. Among his first papers was a presentation in 1926 to the American Mathematical Society on the use of topological principles in the derivation of distributions. Many of his later results were obtained with geometrical arguments. He was able to exploit asymptotic theory to obtain normal approximations and limiting results. Hotelling, in 1930, was the first to attempt to state precise conditions for the consistency and asymptotic normality of the maximum likelihood estimator. With Lester Frankel in 1938, he obtained expansions for the distribution of Student's t-statistic, thereby leading to a useful normal approximation.

With his interest in relations between variables, Hotelling recognized that the required normality assumptions may be unrealistic. This led him to study the nonparametric version of the product moment correlation. Together with Margaret Pabst, Hotelling

studied the Spearman rank correlation coefficient and they showed that for large samples it approximately follows a normal distribution. In 1953, Hotelling continued his analysis of correlations and obtained expansions of Fisher's z-transformation of Pearson's product moment correlation. This work remains among the definitive results in this area.

Hotelling wrote on matrix calculations and developed methods that still stand. His work on weighing designs led to a number of combinatorial generalizations. Although Hotelling had a mathematical focus for his research, he was not averse to working on applied problems. The range of his applied papers include results concerning the physical state of protoplasm, duration of pregnancy, causes of birth rate fluctuations, anthropomorphic measures of Southwest Indians, and effects of pregnancy on dental caries.

Harold Hotelling's originality and imagination is seen in both the genius of his research that produced innovative methods that stimulated whole areas of modern statistics and in his prescient leadership in the discipline of statistics. Further details concerning Hotelling, including his complete bibliography, can be found in Olkin *et al.* (1960), in Pfouts (1960), and in three articles appearing as 'Three papers in honor of Harold Hotelling at 65' in *The American Statistician* (**14**), published in 1960 (pp. 15–25).

See also: Fisher, Ronald A (1890–1962); Multivariate Analysis: Overview; Statistics, History of

Bibliography

Box J F 1978 *R. A. Fisher: The Life of a Scientist*. Wiley, New York

Hotelling H 1927 Review of 'Statistical methods for research workers.'*Journal of the American Statistical Association* **22**: 411–12

Hotelling H 1931 The generalization of Student's ratio. *Annals of Mathematical Statistics* **2**: 360–78

Hotelling H 1933 Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**: 417–41 and 498–520

Hotelling H 1936a Simplified calculation of principal components. *Psychometrika* **1**: 27–35

Hotelling H 1936b Relations between two sets of variates. *Biometrika* **27**: 321–77

Hotelling H 1938 The general welfare in relation to problems of taxation and of railway and utility rates. (Presidential address to the Econometric Society). *Econometrica* **6**: 242–69

Hotelling H 1940 The teaching of statistics. *Annals of Mathematical Statistics* **11**: 457–70

Hotelling H 1963 'Remarks by Harold Hotelling at a Seminar in which Honorary Degree Recipients were asked to give brief intellectual biographies.' As transcribed from a tape recording — not corrected or edited. University of Rochester. Private Communication

Hoxie R G 1955 The Department of Mathematical Statistics. In: *A History of the Faculty of Political Science, Columbia University*. Columbia University Press, New York, pp. 250–5

Kruskal W H 1980 Comment: First interaction with Harold Hotelling; testing the Norden bombsight. *Journal of the American Statistical Association* **75**: 331–33

Madow W G 1960 Harold Hotelling as a teacher. *American Statistician* **14**: 15–17

Neyman J 1960 Harold Hotelling: A leader in mathematical statistics. In: Olkin I, Ghurye S G, Hoeffding W, Madow W G, Mann H B (eds.) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Stanford, CA, pp. 6–10

Olkin I, Ghurye S G, Hoeffding W, Madow W G, Mann H B 1960 *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Stanford, CA

Pfouts R W 1960 *Essays in Economics and Econometrics: A Volume in Honor of Harold Hotelling*. University of North Carolina Press, Chapel Hill, NC

Wallis W A 1980 The Statistical Research Group, 1942–45. *Journal of the American Statistical Association* **75**: 320–30

Working H, Hotelling H 1929 Applications of the theory of error to the interpretation of trends. *Journal of the American Statistical Association* **24**(March supp.): 73–85

I. Olkin and A. R. Sampson

Household Archaeology

1. Introduction: The Emergence of Household Archaeology

Household archaeology emerged in the late 1970s and early 1980s as a late development of social archaeology on both sides of the Atlantic. The aim of household archaeology is to create a context in which a humanized reconstruction of the past may be nurtured, through the study of intrasettlement relations. Such an investigation at the microscale involves the intensive recording of archaeological observations and detailed excavation of an area broad enough to include architectural and other features of the built environment. The criticism of such an effort is that it produces an isolated picture of one part of a social group—perhaps only one household—that does not provide a statistically valid extrapolation to the rest of a settlement, or from one settlement to the regional scale.

Household archaeology developed both as an extension of and a reaction, by both processual ('new') and Marxist-oriented archaeologists to the social evolutionary features of social archaeology. Social archaeology provided data and inferences for explanations of change and variation of past cultural systems, focusing on intersettlement relations, especially in terms of the interaction of political leaders on a

regional basis within the parameters of ecological features such as distance and resource distribution. Marxist-oriented studies added the dimensions of ideology and power relations in production and distribution.

A criticism of the macroscale models of social change has been that, in identifying the longer-term evolutionary trajectory of, for example, resource control, hierarchical social relations, and population growth, and in establishing the general social form of a group, the context of shorter term changes and variability is given a lower priority in explanation. In this way a dehumanization of the past has been created in order to 'write history (or prehistory) large.'

Household archaeology developed in the 1980s and 1990s as the result of a growing interest in intra-settlement relations, but it is a sub-discipline that is neither unified, organized, nor recognized within archaeology. The aims and agendas of the archaeologists who carry out studies that are termed 'household archaeology' are as variable as are the theoretical strands of archaeology itself.

2. *Household Archaeology: Archaeology at a Small Scale*

When it was initially developed in the late 1970s the study of households aimed at filling in the most detailed level of settlement pattern and activity analysis, which could then be extrapolated to make more general statements about demographic trends, specialized production, class structures and so on (Flannery and Winter 1976). Later, in the important volumes edited by Wilk and Rathje and Wilk and Ashmore, the importance of households as 'fundamental elements of human society' was emphasized (Ashmore and Wilk 1988, p. 1, Wilk and Rathje 1982). At this time it was suggested that, since households are the level at which social groups articulate directly with economic and ecological processes, their study would offer a chance for archaeologists to examine social adaptation (in a functionalist sense) with direct reference to the empirical details of the archaeological record. It was assumed this would be done within the methodological framework of scientific logical positivism. In other words, household analysis would allow archaeologists to 'bridge the existing 'mid-level theory gap' in archaeology' (Wilk and Rathje 1982, p. 617). Thus household archaeology developed alongside the empirical theory-building in archaeology, known also as middle range research, that lies at the heart of the history of processual archaeology at the end of the twentieth century.

It is clear that household archaeology in the 1980s implied examining and analyzing social change at a microscale as a complement to analyses of social and economic change at a broader scale of regions and continents. Household archaeology served to provide a richness of detail, but was not an end in itself. The ultimate target was to understand social evolutionary change.

3. *Sources of Inspiration for Household Archaeology*

The explicit inspiration for household archaeology comprised the studies of households and multiscalar histories written by historians (including historians of the so-called Annales school) (Braudel 1972), and by anthropologists, especially those interested in historical demography and historical anthropology (Arnould and Netting 1982, Laslett and Wall 1972, Netting et al. 1984), (Hammel and Laslett 1974, p. 76). It is the latter who coined the term 'co-resident domestic group,' as a way to bring together the enormous diversity of the minimal social group—family, household, housefuland to use a term that was not subject to 'cultural bias' and would have value for 'objective scientific analysis.' The "'co-resident domestic group'' consists of those who share the same physical space for the purpose of eating, sleeping and taking rest and leisure, growing up, childrearing and procreating' (Hammel and Laslett 1974, p. 76).

Archaeologists envied the historians and social anthropologists their knowledge about the kinship relations and the demographic composition of the members of a domestic group that allowed them to explore the 'co-resident domestic group' in terms of variability in their size and kinship composition, and the cooperative activities of their members. Traditional scientific archaeological methodology puts limitations on the investigation of the social composition of or relations within the domestic group. In view of the dangers inherent in using architectural features or any other archaeological data to identify a 'household' on a prehistoric site, Wilk and Rathje pointed out that it was more productive for archaeologists to investigate questions such as what co-residents of a domestic group (households) do, rather than what their social form is (who lives there and how they are related) (Wilk and Rathje 1982). Thus, for example, it was argued that to understand transformations in the social relations of production in Neolithic Southeast Europe, it was not essential to be able to identify 'households' in the historical and anthropological sense of the term (Tringham 1990, p. 603). It was sufficient to be able to investigate changes in co-residence and the cooperative activities of kin-based co-resident domestic groups.

It is not surprising that the impetus for household archaeology came from archaeologists who had focused on the excavation and interpretation of well-preserved architectural remains. Architecture—more broadly, the built environment—provides the essential

context of co-resident domestic groups. Archaeologists working in prehistoric Mesoamerica led the field (Ashmore and Wilk 1988, Wilk and Rathje 1982). Simultaneous and parallel developments were occurring in the Southwest US with archaeologists working with spectacular Pueblo architecture, such as Grasshopper Pueblo (Ciolek-Torrello 1986), and in Europe and the Near East with archaeologists working in Neolithic architectural contexts (Bogucki 1988, Tringham 1990, Tringham 1991). The Chacmool conference on Household and Community provided a welcome synthesis of the diverse routes to household archaeology in place by 1989 (MacEachern et al. 1989).

4. *Modeling Household Activities in Archaeology*

In modeling the economic and social role of co-resident domestic groups, participants of 'household archaeology' have followed social anthropologists in their models of the domestic mode of production (Chayanov 1966 (reprinted 1986), Netting et al. 1984, Sahlins 1972). These models characterized self-sustaining agriculturists and horticulturists who live in permanent villages and carry out relatively small-scale production. Such studies have emphasized the importance of the function of households as the organizing units of cooperative labor, production, consumption and sharing of resources, the transmission of property and rights from one generation to another, and the creation and maintenance of ties and alliances with other units through marriage and other forms of exchange. These same essential features of social reproduction became the objects of investigation of household archaeology (Ashmore and Wilk 1988, Wilk and Rathje 1982).

The study of 'what households do' has been enriched by the Marxist focus on ideology, power and social inequality. In the domestic mode of production the pattern of domination and inequality between and within different households of a settlement and between different settlements, although temporary and flexible, is thought to comprise a crucial variable in stimulating and enabling larger-scale transformations such as the intensification of production, the increase of the labor supply, and the increase in cultural complexity (Meillassoux 1978).

Inequalities certainly exist, along with innumerable variations, between households and are the result of differences in the cycles of demographic and economic evolution through which households pass—cycles in which they expand, accumulating wealth, power, property, and members, and then contract. Thus, at any one time the households of a settlement will demonstrate differences in membership, the composition of the labor force, and the structure of dominance within the co-resident group and within the settlement as a whole. They will demonstrate differences in relations with groups outside the village, and, most importantly, differences in access to the products and processes of production. This concept of small-scale social inequality is the basis of a number of archaeological models that use the emerging importance of households as the unit of social reproduction to explain larger social and economic transformations (Rathje and McGuire 1982, Tringham 1990).

These differential patterns of social and economic relations reflecting social inequalities, however, have such a low archaeological visibility that these societies have traditionally been regarded as 'egalitarian.' The difficulty in recognizing inequality in 'egalitarian' societies is due to the fact that it is ephemeral and not cumulative. An additional significant factor, however, is that inequality in such societies is not based on such criteria as the ownership of land and other material property, which are easily definable or easily recognizable in the archaeological data. Control of labor —not land, capital or property ownership—is the crucial variable on which inequality is based in precapitalist self-sustaining societies. Thus, those who control the circulation of labor and its products, control the social reproduction of society. At the scale of the domestic unit or household, this may be an individual within the household, or it may be 'senior' households who control the labor of 'junior' households. It has been argued that inequality at this scale creates tensions in numerous spheres and is a crucial variable in the process of socioeconomic transformation that can be observed in the archaeological record (Tringham 1990, p. 608).

5. *Archaeological Evidence of Household Inequality*

The primary requirement for the examination of the social context of production and economic life at an archaeological site (in other words, of household archaeology) is that a number of domestic architectural units be exposed by excavation. This will enable a comparative view of co-resident domestic groups across the settlement at any one time. This information is enhanced by the provision of such a picture through time. Only when a large number of domestic structures is exposed is it possible to tell whether, for example, all the structures in a given excavated occupation horizon represent residences or parts of residences; whether the differentiated units of co-residence can also be differentiated as units of cooperative production; whether all the domestic structures were occupied at the same time; whether they all became part of the archaeological record at the same time or whether they were destroyed or abandoned at different times, and, finally whether they appear to be in a parallel stage of their use-lives at the

time of destruction or abandonment or whether some seem to have been more recently established than others. The study of archaeological architecture in terms of the use-life of the buildings has greatly enhanced these efforts for the purposes of household archaeology (Tringham 1994). In this way the context and nature of the construction, maintenance, occupation, modification, and abandonment/destruction, and replacement of a residence all become objects of knowledge. Traditionally, they would have comprised elements that were subsumed under the formal features of the building. The investigation of a building through different stages of its use-life, brings to light many important aspects that would otherwise have been ignored; for example, modifications such as the addition or removal of rooms during the cycle of growth and decline of a domestic unit; efforts to prolong the use-life of a building by repairs and renewals; symbolic emphasis placed on particular use-life stages such as re-plastering, the destruction of old houses (for example by fire), or founding of new houses (Tringham 1994).

Inequalities within and between domestic units are manifested archaeologically by the spatial differentiation of the production process (that is, in the production, consumption, and distribution of food and nonfood resources). In each case the pattern of economic activity (and inequality) should have a distinctive appearance, demonstrating differential access to the materials, means, process, and products of production. The expectation is that different households will show differential access not only to the products of exchange with other groups but also to the raw materials, techniques, and equipment for making use of resources, such as food items, minerals, and ceramics. Such a differentiation of economic activity is frequently, but not necessarily, accompanied by specific styles of material elaboration, such as ceramic decoration.

6. *Trends in Household Archaeology*

For many years household archaeology remained marginal to the mainstream of processual (new) archaeology. In his review of Ashmore and Wilk's edited volume on Household Archaeology in Mesoamerica, David Freidel noted that for the interest in and impact of 'household archaeology' to be sustained, it would need to be incorporated into a more ambitious theoretical program in archaeology (Freidel 1989). Its incorporation into Marxist models of power and social inequality certainly helped in this respect. The theoretical basis that has contributed most to sustaining interest in household archaeology, however, has been that of feminist anthropology and theories of social practice (Bourdieu 1977, Moore 1988). A common theme of these diverse post-structural theories has been their development beyond neo-Marxian models of power and labor control to models of how ideologies of gender, class, labor, production, identity, etc., are constructed, maintained and transformed. They focus on domestic arenas and the contexts of daily practice as being the most significant places where these negotiations between social agents take place. During the later 1990s these models have also come to characterize the many faces of post-processual archaeology, one of which is an archaeology of households that has transformed into an archaeology of gendered places (Tringham 1994).

In the archaeology of gendered places, the archaeological record is not only analyzed and observed at a microscale, but also interpreted at a microscale. It challenges the search for the universal category of 'kin-based co-residential domestic group' and embraces the richness of the diversity of the social context of domestic action that has been forefronted in social anthropology for many years (Moore 1988). It lifts the household domain out of its previous position of 'assumed general knowledge' and as marginal to the 'great events' of history. It provides a richly-based, multiscalar theoretical framework to consider the role of the labor of players in prehistory: women, men and children (Tringham 1991).

The theoretical basis of feminist anthropology and social practice theory obviously creates some serious epistemological challenges to the scientific logical positivist methodology that characterizes mainstream processual (new) archaeology. It is as difficult to identify 'gender' in archaeological architecture as it is to find 'households' and 'families.' Responses to this challenge have been articulated in many publications (Hodder 1999, Wylie 1992). They are based in a celebration of the ambiguity of archaeological data, a critical awareness of the scientific process of inference, and a richly researched utilization of sources for interpretation—ethnography, history, art and literature. Above all in the post-processual epistemological framework, archaeology is identified as an interpretive discipline rather than a hypothetico-deductive search for true (pre)history. Thus in order to think about the prehistoric construction and practice of gender ideologies and dominance relations within or between households, a researcher may interpret the archaeological data and their meaning in many ways. It is not required to identify gender or 'domestic unit' in the archaeological record.

The database for an archaeology of gendered places remains especially but not exclusively in the record of the built environment. Along with the nature of household archaeology itself, however, the excavation and interpretation of archaeological architecture has changed. According to post-processual concepts, the built environment can be thought of as both a passive container that has a use-life and reflects social action and as an 'active' arena or medium for and symbolic expression of social actions. Thus social actions, practices, negotiations, and dominance ideologies of

whoever passes through a building are constrained and channeled by the materiality of that building. (Tringham 1994). In practice this has led to the necessity of retrieving a highly detailed microstratigraphic record of the traces of human activity, in which the sequence of every depositional event is recorded and interpreted. It is time-consuming, expensive, but immensely rewarding (Hodder 1999, Tringham 1994).

Along similar lines, household archaeology in the late 1990s has embraced the ambiguities recognized by practice theory anthropologists about kinship categories, including 'family' and 'household.' In their view, it is the meaning and cultural value ascribed to a relationship that are significant rather than some set of biological ties (Gillespie 2000). The enactment of kinship often involves the negotiation of identity and the identification of relationships through particular places on the landscape—buildings and sets of material culture (Joyce 2000).

At the beginning of the twenty-first century, 'what households do' is still a focus of household archaeology, but the 'doing' is far more complex than physical action. Moreover, a 'co-resident domestic group' is no longer a faceless economic category, but comprises the complex life-histories of many people as well as buildings and places. Finally, household analysis is no longer a step along the ultimate goal of understanding social evolutionary change. It plays an essential role in the creation of a multiscalar history that considers such questions as how—in practice—was a sense of place created; how was social memory about this group maintained; what were the myths about garbage; what would the walls of a house say if only they could speak (Tringham 2000).

See also: Ancestors, Anthropology of; Anthropology, History of; Ethnoarchaeology; Family and Kinship, History of; Gender, Feminism, and Sexuality in Archaeological Studies; Historical Demography; Horticultural Societies; Kinship in Anthropology; Marxist Archaeology

Bibliography

Arnould E J, Netting R M 1982 Households: Changing form and function. *Current Anthropology* **23**: 571–6

Ashmore W, Wilk R R 1988 House and household in the Mesoamerican past: An introduction. In: Wilk R R, Ashmore W (eds.) *Household and Community in the Mesoamerican Past*. University of New Mexico Press, Albuquerque, NM, pp. 1–28

Bogucki P 1988 *Forest Farmers and Stockherders: Early Agriculture and its Consequences in North-Central Europe. New Studies in Archaeology*. Cambridge University Press, Cambridge, UK

Bourdieu P 1977 *Outline of a Theory of Practice*. [trans. Nice R]. Cambridge University Press, Cambridge, UK

Braudel F 1972 *The Mediterranean and the Mediterranean World in the Age of Philip II*. [trans. Reynolds S]. Harper and Row, New York

Chayanov A V 1966 (reprinted 1986) *The Theory of Peasant Economy*. University of Wisconsin Press, IL

Ciolek-Torrello R 1986 Room function and household at Grasshopper Pueblo. In: Benson C, Upham S (eds.) *Mogollon Variability*. 107–19. New Mexico State University, The University Museum. Occasional Papers 15, Las Cruces, NM

Flannery K V 1976 Analyzing household activities. In: Flannery K V (eds.) *The Early Mesoamerican Village*. Academic Press, New York

Freidel D 1989 Household and community in the Mesoamerican past. *Science* 19 May: 863–4

Gillespie, S 2000 Beyond kinship: an introduction. In: Gillespie S D, Joyce R A (eds.) *Beyond Kinship: Social and Material Reproduction in House Societies*. University of Pennsylvania Press Philadelphia, PA, pp. 1–21

Hammel E A, Laslett P 1974 Comparing household structure over time and between cultures. *Comparative Studies in Society and History* **16**: 73–109

Hodder I 1999 *The Archaeological Process: An Introduction*. Blackwell, Oxford, UK

Joyce R 2000 Heirlooms and houses: Materiality and social memory. In: Gillespie S, Joyce R (eds.) *Beyond Kinship: Social and Material Reproduction in House Societies*. University of Pennsylvania Press, Philadelphia, PA, pp. 189–212

Laslett P, Wall R (eds.) 1972 *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK

MacEachern S, Archer D J W, Garvin R D (eds.) 1989 *Households and Communities*. The 21st Annual Conference: Chacmool Calgary, AB

Meillassoux C 1978 'The economy' in agricultural self-sustaining societies: A preliminary analysis. In: Seddon D (ed.). [trans. Lackner H]. *The Relations of Production: Marxist Approaches in Economic Anthropology*. Frank Cass, London, pp. 127–57

Moore H L 1988 *Feminism and Anthropology*. University of Minnesota Press, Minneapolis, MN

Netting R McC, Wilk R R, Arnould E J (eds.) 1984 *Households: Comparative and Historical Studies of the Domestic Group*. University of California Press, Berkeley, CA

Rathje W L, McGuire R H 1982 Rich Men … Poor Men. *American Behavioral Scientist* **25**(6): 705–16

Sahlins M 1972 *Stone Age Economics*. Aldine Chicago

Tringham R 1990 Conclusion: Selevac in the wider context of European prehistory. In: Tringham R, Krstic D (eds.) *Selevac: a Neolithic Village in Yugoslavia, Monumenta Archaeologica No. 15*. UCLA Institute of Archaeology Press, Los Angeles, pp. 567–616

Tringham R 1991 Households with faces: The challenge of gender in prehistoric architectural remains. In: Gero J M, Conkey M W (eds.) *Engendering Archaeology: Women and Prehistory*. Basil Blackwell, Oxford, UK, pp. 93–131

Tringham R 1994 Engendered places in prehistory. *Gender, Place, and Culture* **1**: 169–203

Tringham R 2000 The continuous house: A view from the deep past. In: Gillespie S, Joyce R (eds.) *Beyond Kinship: Social and Material Reproduction in House Societies*. University of Pennsylvania Press, Philadelphia, PA

Wilk R R, Rathje W L 1982 Household archaeology. *American Behavioral Scientist* **25**(6): 617–40

Wylie A 1992 The interplay of evidential constraints and political interests: recent archaeological research on gender. *American Antiquity* **57**: 15–35

R. Tringham

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Household in Anthropology

Anthropologists use the term household loosely to refer to a residential group whose members cooperate in some activities of production, consumption, or child rearing. They have distinguished the household from the family by defining the former as a residential unit and the latter as a kinship unit. Close scrutiny of this distinction reveals key assumptions that have shaped the way anthropologists and other social scientists have conceptualized and studied not only households, but *families*, *kinship*, and *domestic groups*.

1. Problems in Defining the Household

The grounds for analytically separating families and households lie in the broader claim that kinship and locality are two distinct principles of organization (Keesing 1958). This claim, in turn, derives from the discipline's focus in the nineteenth and early twentieth centuries on the principles organizing social groups and relations in nonstate societies. Kinship, which was assumed to be universally defined on the basis of genealogical linkages rooted in biological reproduction, was considered the most important among these principles. As a key mode of recruiting members to (and limiting membership in) politically important groups such lineages and clans, kinship (genealogical propinquity) was distinguished from *locality* (spatial propinquity), which defined residential units such as hamlets and villages. As anthropologists expanded their range of ethnographic research to include peasant communities, they drew on the distinction between kinship and locality to differentiate family from household. The family as a corporate kin group was distinguished from the household as a collection of kin and sometimes nonkin who share a common residence. The corporate character of the peasant family was viewed as rooted in its jural rights to property, usually land, which its members held in common. Family members supposedly retained these rights regardless of whether they were members of the same household. Hence, family and household, it was argued, are both 'logically distinct' and 'empirically different' (Bender 1967, p. 493).

Actual usage of the term household in the anthropological literature, however, calls into question both the logical distinction and the empirical difference between families and households. For one, residential units composed of individuals who are not genealogically linked, such as orphanages, college dormitories, men's houses (in societies where gender segregation among adults is practiced), convents, and army barracks, usually are not considered households. At the same time, however, families whose members move between residential units that are located on different continents are referred to as 'transnational households' (Basch et al. 1994). In other words, anthropologists attach the term household only to groups of people sharing both a common living space and activities that are identified as *domestic*. *Domestic activities* include those activities related to food production and consumption or to reproduction, including childbearing and child rearing. As these are the very same activities that are considered the functional basis of the 'family,' in practice, households are identified using both 'kinship' and 'residence' criteria. Thus, the claim for the logical distinction and empirical difference between family and household is not affirmed by scholarly practice.

Because 'domestic' activities are sometimes carried out by groups of people who do not reside together, some anthropologists (Bender 1967) have suggested that it would be better to employ alternate terms in our ethnographic descriptions. Among the alternatives that have been suggested are 'co-residential group,' 'budget unit,' and 'domestic group.' Less important than the terms employed, however, is the need for ethnographers to describe precisely the composition, organization, activities, and functions of these units. In addition, the cultural meanings people attach to these units need to be explicated. In the absence of this knowledge, anthropologist and others are likely to project their own cultural meanings and expectations on them.

The difficulties of identifying households are illustrated by innumerable cases (see Segalen 1984) which demonstrate that neither the boundaries nor the functions of household are easily discerned. Questions have been raised as to how to treat residential groups that move through a seasonal cycle of dispersal and concentration, how to handle the movement of people between dwelling units, whether to define as a single household the huts or houses that share a common yard or those that share a common stove, and whether to include servants, apprentices, boarders, lodgers, and nonkin in the household. More recently, studies of transnational migration (Rouse 1991, Basch et al. 1994) have raised the question as to whether residential units located in different nations and even on different continents constitute a single household. Studies of gay and lesbian kinship (Weston 1991), on the other hand, have challenged the assumption that heterosexual marriage and parenthood are the universal, natural core of families and households.

2. Problems in the Classification of Households Types

Much of the literature on households has been concerned with explaining variations among households that are distributed over time and space. In describing this variation, most scholars have focused on the character of the genealogical links between members of the household as its defining feature. Terms such as 'stem family household,' 'nuclear family

household,' and 'patrilateral extended household' are based on the configuration of genealogically-defined kin types included in the household—in other words, its *household composition*. In addition, the most genealogically 'close' members of the household are assumed to form its structural core. For example, if a household contains two adult brothers, their wives and children, it is labeled a fraternal-joint household. The assumption here is that the relationship between the brothers constitutes the structural core that shapes the relations of all the members of the household. Accordingly, studies of household change over time track changes in the genealogical composition of the household.

The classification of households on the basis of their genealogical composition rests on the shaky assumption that people who are similarly linked genealogically will have the same or greatly similar social relations. In other words, it is assumed that the labeling of a household as a 'stem-family household' captures its organizational structure and function. But a stem-family household found at one time and place does not necessarily share the same organizational dynamics as a stem-family household at another time and place. A household, or any social unit, is much more than the sum of its genealogical links. It is the negotiated outcome of the culturally meaningful actions and relations of its members. The actions and relations of household members with identical genealogically defined links can differ radically. Even households with the same genealogical composition in the same community may have significantly different structures of authority and divisions of labor (Yanagisako 1984).

3. *What Explains Variation among Households?*

A key question that has fueled comparative studies of households by anthropologists, historians, demographers, and other social scientists is what explains variations among them. Two related questions have been asked of variations observed in household composition. First, why do a significant proportion of the households in a society differ from what people report to be the ideal type of household? Second, irrespective of these ideals, what explains the distribution of household types within and between societies? Three types of variables have been offered to answer both these questions: demographic variables, economic variables, and social stratification variables.

3.1 *Demographic Variables*

Demographic variables affect the size and composition of households, thus acting as constraints on the attainment of the culturally ideal household type. Age at marriage, life expectancy, and fertility shape the composition of households at any given time in a population. The economy of the household is also shaped by demographic processes. Among peasants for whom the household is the unit of production, changes in the demographic structure of the household as it moves through the developmental cycle entail changes in the ratio of consumers to workers (Chayanov 1966).

Fortes's (1958) model of the developmental cycle of domestic groups draws on these demographic processes to explain the varieties of domestic groups found in a society. Fortes hypothesized that the variety of household types found in a community or society at any point in time are merely different phases in the developmental cycle of a 'single general form.' As all households pass through different stages of the developmental cycle at different times, a census taken at a particular point in time will catch only a few in the ideal, complete phase. Hence, an ideal household type such as the patrilineal stem family household of Japan (which includes a married couple, their adult son and his wife and children) is better regarded as 'a transitory state in the development of the household, rather than as a form' (Hammel 1975, p. 142).

Fortes's model has been useful in helping us understand the impact of events such as birth, marriage, and death on the composition of households. It cannot account for all the observed variation in households, however, because other cultural and social forces (see below) are also involved. Fortes himself later (1978, p. 18) conceded that a model of a uniform developmental cycle could not explain why the actual histories of different families entail different developmental sequences. The model may also blind us to historical shifts wherein more recently formed households embark on a different trajectory than older ones. In addition, while demographic factors undoubtedly shape households, it is a mistake to treat them as exogenous, biological constraints on the composition and economy of households. To do so confuses social replacement with biological reproduction and overlooks the strategies people actively employ to control household size and composition, including their 'kinship' practices of marriage, fosterage, adoption, timing and spacing of births, and timing of family division. The collection of longitudinal data on households, whether by utilizing historical sources or through retrospective oral histories, is necessary for understanding both continuity and change. Cohort analysis is an effective tool for helping ascertain similarities and differences in the household histories of members of successive cohorts, defined on the basis of year of birth, marriage, or other historically significant events experienced by the members of a community.

3.2 *Economic Variables*

Economic variables are also adduced to explain cross-societal variation in household composition and organization. In some cases a mix of 'economic' and

'ecological' factors, including labor needs of production, defensive needs, care of children, taxation, and conscription practices of the state, are identified as the functional basis for the prevalence of certain household types. In others, the labor requirements of a particular productive activity are singled out as the primary determinant of household type. In their cross-societal analysis of 60 randomly selected societies, Pasternak et al. (1976) conclude that labor requirements are the most powerful determinant of the formation of extended family households. They hypothesize that in societies where work outside the home makes it difficult for a mother to tend to children and other domestic tasks or when the outside activities of the father make it difficult for him to perform his subsistence work, extended family households will prevail. Although they are able to demonstrate a correlation between activity requirements and presence of extended family households, they lack diachronic data to support the hypothesized causal relation between their independent variable (activity requirements) and their dependent variable (presence of extended family households). Consequently, they must resort to arguing that it is more plausible that activity requirements cause extended family households rather than vice versa. However, it is even more plausible that household composition and productive activities do not lie in a simple, unidirectional causal relation but rather shape each other.

Attempts like that of Pasternak et al. to identify the determinants of household composition and structure through statistical analysis of cross-societal comparisons have declined significantly over the past 30 years as the inherent problems and limitations of such studies have been increasingly recognized. Among these problems is that of discerning whether the units that have been selected for comparison are appropriate. As all societies contain different frequencies of a range of household types, it is misleading to classify entire societies into those with 'extended family households' vs. those with 'independent family households.' A better understanding of the relation between household composition and 'economic' factors may be gained by investigating the ways in which diverse types of households in a society articulate with one another, rather than by reducing the entire society to a single household type.

Labor requirements are often used to explain the frequency with which a particular household type appears in a society or section of society. Netting (1974) cites several ecological studies as demonstrating that household composition among horticulturalists and agriculturalists varies with the type and amount of labor required for effective crop production. His own study attributes the differences between the large, extended families of shifting cultivators and the nuclear family households of intensive agriculturalists in the same Nigerian plateau environment to different labor requirements (Netting 1965).

A shaky assumption underlies both cross-societal and intrasocietal studies that seek the determinants of household variation in productive labor requirements. In all of them, it is assumed that once having identified the subsistence activity or primary productive activity of the household, researchers can ascertain the labor requirements of that activity and proceed to show how the household efficiently meets them. In doing so, however, they overlook the fact that it is not simply technological requirements that necessitate a household type, but the entire manner in which production is socially organized. In other words, a productive activity can be accomplished by a variety of household types depending on how that production is socially organized, including the division of labor by sex and age, the use of hired labor, cooperation between households in productive tasks, and the exchange of labor between households. In assuming that the household is a self-contained labor unit, researchers fail to sufficiently investigate the existence of other forms of labor organization. But the social units engaged in production may not be the same at different phases of the productive cycle. For example, different social units may work together in the planting, weeding, and harvesting of crops. To attribute household composition to labor requirements begs the question, because it assumes the prior existence of an entire social organization of production. In order to overcome this narrow focus on technical requirements of production, the interplay of a wider range of cultural, political, and economic forces shaping productive relations within and between households needs to be considered.

3.3 Social Stratification

Differential access to resources, in particular differences in ownership and control of the means of production, also shape household variation in both peasant communities and among urban, industrial populations. A past tendency to neglect inequality in studies of households and other domestic groups can be attributed to the assumption that the communities studied by ethnographers were basically homogeneous. When anthropologists have studied peasant communities, for example, the implicit contrast with the landowners—who were generally not studied—led them to overlook significant differences among the peasants. Consequently, there has been a tendency to dismiss stratification as a significant factor in the analysis of household variation among peasants. Yet peasant communities are nowhere homogeneous, but are always internally differentiated. Differences in wealth among peasants are also overlooked because they are often obscured by idioms of kinship that bind the landed to the landless and the land-rich to the land-poor. Significant wealth differences may exist, however, between the smaller co-residential units of a single kin-based compound or between brothers who head

the main and branch households of a family. In other words, ideas and ideologies of family and kinship may obscure stratification among kin.

4. Changes in Households over Time

Anthropological studies of household and family change over time have focused generally on the effects of industrialization, urbanization, and the increasing integration of communities into capitalist market relations. Many of these studies challenge or argue for the refinement of the prevailing sociological theory that family, household, and kinship decline in importance as a society 'modernizes' and becomes increasingly urbanized and industrialized. They demonstrate that in some cases urbanization and industrialization have strengthened the extended family rather than weakened it and that new functions are assumed by extended families and extended family households. For example, urban migration may contribute to the maintenance of extended family households. While these studies provide an important correction to the theory of modernization and kinship decline, they are marred by the inconsistent use of ambiguous terms, including 'extended family household.' The argument over whether extended families have declined with urban-industrialization is in part the result of differences in the way in which researchers define and conceptualize household type. Quite often it is assumed that people who live in extended family households pursue common strategies and cooperative actions under the direction of a central authority—the head of the household. This not only gives undue weight to centralized authority as the defining feature of the extended family, but it also overlooks the heterogeneity in the goals and strategies of the members of a household.

5. Recent Critical Approaches to Household

The adaptationist, utilitarian approach to households that prevailed in much of the anthropological literature up until the 1980s has been challenged by new theoretical approaches that have had an important impact on the discipline in general. Among these perhaps the most significant has been feminist theories of gender.

In the 1970s, one of the first questions feminist anthropologists raised in rethinking the anthropological analysis of women's status and roles was whether ethnographers had overlooked women's goals, strategies, and forms of action, especially in the 'domestic sphere.' M. Wolf's (1972) study of Taiwanese joint family households, for example, revealed that family and household division was brought about as much by women's attempts to advance their own goals and those of their 'uterine family' as it was by conflicts among brothers. Collier (1974) argued that this process occurs commonly in societies where men gain power by holding together large bodies of co-resident kin, but where women (especially younger women) gain power by breaking up such units. By attending to the informal power women often wield in societies where men are granted formal authority, feminist anthropologists alerted us to the tensions and conflicts and, therefore, the complex processes shaping households and other domestic groups. Likewise, investigations of women's relations with kin and others outside their household demonstrated that it is not just men who link mother–child units to larger institutional structures. The involvement of women in informal exchange networks between families and households had been overlooked by androcentric studies that had equated male relations with the 'public sphere.' Once studies showed that the female and male members of households—as well as the young and the old, those born in the household and those who had married into it—often do not share the same goals and strategies, it was no longer possible to retain the naive concept of a unified 'household strategy.' Finally, by calling into question the presumed universality and immutability of gender identities and the gender division of labor, feminist anthropologists denaturalized gender systems and called for the study of its historical and cultural production. They revealed the gender assumptions embedded in concepts of the 'public' and 'private' spheres (Reiter 1975), in definitions of the 'domestic sphere' and 'domestic groups,' including family and household (Yanagisako 1979) and in theories of kinship (Yanagisako and Collier 1987).

6. Rethinking Households

Functionalist analyses of 'household adaptation' reified the household, treating it as a social actor endowed with human consciousness and motives. Households were described as having goals, fashioning strategies, and adapting to the changing pressures of a modernizing society. Such language demonstrates the ease with which an analytic construct can be made into a social actor. The imputation of consciousness and will to households is rooted in the assumption that those who live together share a collective motive and will—in other words, a unified household strategy. It obscures the complex social processes through which people negotiate the relations that produce households, despite different and sometimes conflicting goals, strategies, and ideas about what it means to 'live together.'

Neither the composition nor the dynamic structure of households can be understood without a thorough understanding of the cultural meanings and sentiments that shape the social actions and relations of the people who live in them. This does not mean that households cannot be compared across societies, but rather that the comparative study of households must include an explication of the culturally specific and

heterogeneous meanings and motives shaping peoples' residential preferences and relations. Such an expanded strategy of research and analysis will enable anthropologists to move beyond crude comparisons of distributions of genealogically defined household types within and across societies to understanding households as the negotiated outcomes of culturally meaningful action.

See also: Collective Beliefs: Sociological Explanation; Community, Social Contexts of; Community/Society: History of the Concept; Community Studies: Anthropological; Everyday Life, Anthropology of; Family and Kinship, History of; Family, Anthropology of; Family as Institution; Home in Anthropology; Household Production; Kinship in Anthropology; Marriage; Matrifocality; Place in Geography; Residential Segregation: Geographic Aspects; Social Stratification

Bibliography

Basch L, Schiller N G, Blanc C S 1994 *Nations Unbound: Transnational Projects, Postcolonial Predicaments, and Deterritorialized Nation-States*. Gordon and Breach, Langhorne, PA

Bender D R 1967 A refinement of the concept of household: Families, co-residence, and domestic functions. *American Anthropologist* **69**: 493–504

Chayanov A V 1966 *The Theory of Peasant Economy*. R D Irwin, Homewood, IL.

Collier J F 1974 Women in politics. In: Rosaldo M Z, Lamphere L (eds.) *Women, Culture, and Society*. Stanford University Press, Stanford, CA

Fortes M 1958 Introduction. In: Goody J R (ed.) *The Developmental Cycle of Domestic Groups*. Cambridge University, Cambridge, UK

Fortes M 1978 An anthropologist's apprenticeship. *Annual Review of Anthropology* **7**: 1–30

Hammel E A 1975 Reflections on the zadruga. *Ethnology Slavery* **7**: 141–51

Keesing F M 1958 *Cultural Anthropology: the Science of Custom*. Rinehart, New York

Netting R M 1965 Household organization and intensive agriculture: The Kofyar case. *Africa* **35**: 422–9

Netting R M 1974 Agrarian ecology. *Annual Review of Anthropology* **3**: 21–56

Pasternak B, Ember C R, Ember M 1976 On the conditions favoring extended family households. *Journal of Anthropological Research* **32**: 109–23

Reiter R R 1975 Men and women in the south of France: Public and private domains. In: Reiter R R (ed.) *Towards an Anthropology of Women*. Monthly Review Press, New York

Rouse R 1991 Mexican migration and the social space of postmodernity. *Diaspora* **1**: 8–23

Segalen M 1984 Nuclear is not independent: Organization of the household in the Pays Bigouden Sud in the nineteenth and twentieth centuries. In: Netting R McC, Wilk R R, Arnold E J (eds.) *Households: Comparative and Historical Studies of the Domestic Group*. University of California Press, Berkeley and Los Angeles

Weston K 1991 *Families We Choose: Lesbians, Gays, Kinship*. Columbia University Press, New York

Wolf M 1972 *Women and the Family in Rural Taiwan*. Stanford University Press, Stanford, CA

Yanagisako S J 1979 Family and household: The analysis of domestic groups. *Annual Review of Anthropology* **8**: 161–205

Yanagisako S J 1984 Explicating residence: A cultural analysis of changing households among Japanese-Americans. In: Netting R McC, Wilk R R, Arnold E J (eds.) *Households: Comparative and Historical Studies of the Domestic Group*. University of California Press, Berkeley and Los Angeles

Yanagisako S J, Collier J F 1987 Toward a unified analysis of gender and kinship. In: Collier J F, Yanagisako S J (eds.) *Gender and Kinship: Essays Toward a Unified Analysis*. Stanford University Press, Stanford, CA

S. J. Yanagisako

Household Production

Household production is the production of goods and services by the members of a household for their own consumption, using their own capital and their own unpaid labor. Goods and services produced by households for their own use include accommodation, meals, clean clothes, and childcare.

The process of household production involves the transformation of purchased intermediate commodities (e.g., supermarket groceries and power-utility electricity) into *final consumption commodities* (meals and clean clothes). Households use their own capital (kitchen equipment, tables and chairs, kitchen and dining room space) and their own labor (hours spent in shopping, cooking, laundry, and ironing). The total economic value added by households in household production has been aptly named Gross Household Product (GHP) (Eisner 1989, Ironmonger 1996).

1. The Use of Labor and Capital in Household Production

In the language of economics, labor and capital are the two factors of production. However, labor, the time and effort provided by household members, is really the use of human capital; the second factor of production, capital, is the use of physical or tangible nonhuman capital (the services provided by land, dwellings, vehicles, and equipment). Economics conceptualises the production process as the use of capital (human and other) together with energy, to transform raw materials and 'unfinished' commodities (intermediate inputs) into finished goods and services ready for use by people (final consumption).

As countries industrialize, a large part of household production of food, clothing, furniture, and housing is transferred to business organizations and then purchased by households. Nevertheless, even in a world

apparently dominated by the market, a large amount of household production is necessary. Households need to add further value to put purchased commodities in the possession of ultimate consumers in the right place and at the right time.

2. Household Production in Contrast to Market and Subsistence Production

2.1 Market Production

While the market economy produces goods and services not produced by households (television sets and battleships), in many cases the market and the household are in direct competition, (restaurant meals vs. home prepared meals, and taxis vs. self-driven owned cars). Restaurants and taxis usually require immediate payment, if not by cash, then by credit card. In households, users do not pay each other by cash or credit card for meals served or transport provided.

Four separate modes of production are possible depending whether the household or the market provides the capital or the labor. Household production occurs when the household provides both its own capital and its own labor. However, the market can provide either one or both of these two factors of production. For example, a housekeeper could be employed to prepare meals, clean the house, and do the laundry. This paid housekeeper's time would be the use of 'market labor' in conjunction with 'household capital'—the kitchen, dwelling, and household equipment. Again, a household could rent a dwelling and a vehicle (market capital) but do its own cleaning, cooking, laundry, and driving (household labor).

In many countries, the use of paid domestic servants has almost disappeared; servants have been replaced by the household's own labor combined with more and better household capital and equipment. Many recent developments, such as supermarkets and automatic banking machines, involve a mode of production with households providing the required unpaid labor in lieu of the paid labor of shop assistants or bank tellers.

2.2 Subsistence Production

In developing countries many millions of households use their own capital and unpaid labor by way of fishing, collecting wood and water, growing vegetables and other food, building shelter, and making clothing. Thus, subsistence production fits the definition of household production when the goods produced are used within the household that produced them.

However, the wood, water, food, fiber, and clothing outputs from subsistence production could be sold on the market to other households or to business enterprises. When such commodities are in fact sold, the value added in their production conceptually becomes part of market production.

3. Household Production and the Household Economy

The household economy describes the collective economic activities of households. Often the household economy is called the household sector as distinct from the business, government, and foreign sectors. However, the household sector is large enough to deserve the term household economy. The rest of the economy can then be called the market economy. Thus, the transactions between the household and the market are perhaps more akin to international trade between two economies than transactions between different industrial sectors of a single economy. The two major types of inter-economy trade are the sale of labor time by the household and the sale of household goods and services by the market.

3.1 Households as Producers

With few exceptions, economic textbooks focus on households as consumers and fail to discuss households as producers using their own labor and capital.

One of the earliest writers about the process of household production was Charlotte Perkins Gilman (*Women and Economics*, 1898) who questioned the traditional gender division of labor. She proposed moving even more household production to the market. This would yield gains from greater specialization and economies of scale and enable women to choose their work on the basis of inclination and talent.

Margaret Reid (*Economics of Household Production*, 1934) played a significant role in the development of household economics as a discipline, particularly for curricula in some North American universities. She held that, although the household is our most important economic institution, the interest of economists was concentrated on 'that part of the economic system which is organized on a price basis.' Reid is regarded as the first writer to specify the often-used third person criterion to distinguish between productive and nonproductive (consumption) activities. Her test was: 'If an activity is of such character that it might be delegated to a paid worker, then that activity shall be deemed productive' (Reid 1934). Thus, preparing a meal is productive work but eating it is nonproductive consumption or leisure. Of course eating and other consumption activities are productive in the sense that they are necessary for the continuation and re-invigoration of life. Another problem with this criterion is that student educational activities are

classed as unproductive since you would not learn to play the piano by paying someone to take the lessons for you.

A further problem is that production almost always requires a contribution from capital as well as from labor. The criterion needs to be extended to include services that might be obtained from rented capital as well as from a paid worker. Thus, accommodation services provided by a household's own dwelling are production; so are the transport services provided by a household's own vehicles and the entertainment services provided by a household's own sound and vision equipment. This extends the range of household production considered by Reid in 1934 and mostly since then by other writers. The expanded market alternative criterion would be 'An activity shall be deemed productive if it is of such a character that it might be obtained by hiring a worker or by renting capital equipment from the market.'

Gilman and Reid had little impact on mainstream economics. Economic theory continued to portray households as places only of consumption and leisure, with production of goods and services occurring only in business or public enterprises.

3.2 *The New Household Economics*

In the mid-1960s a major theoretical development took place, known as the 'new household economics' (Becker 1981, Ironmonger 1972, Lancaster 1971). In this theory the household is regarded as a productive sector with household activities modeled as a series of industries.

In this new approach, households produce commodities that are designed to satisfy separate wants such as thirst, hunger, warmth, and shelter. The characteristics, or want-satisfying qualities, of the commodities used and produced can be regarded as defining the production and consumption technology of households. With changes in incomes and prices, households still alter expenditures as in the earlier theory. However, in the new theory, households adjust their behaviour as they discover new commodities and their usefulness in household production processes.

The activities approach derived from the theory of the new household economics readily combines with the earlier input-output approach of Leontief (1941) to establish a series of household input–output tables as the framework for modeling household production.

4. *Measuring the Household Economy*

The measurement of the household economy emerged as a focal point for many researchers once the household was recognized as a major centre of production, not just consumption.

4.1 *Importance of Measuring the Household Economy*

In the 1970s a number of writers drew attention to the macroeconomic magnitude of the household economy. Boulding (1972) conceptualised households not only as the main driving force for the market economy, with household purchases covering about 60 percent of GNP, but also as the most important agent in the grants economy. This is the economy of one-way transfers, grants, or gifts, given mainly within households from those earning money incomes to other members, children, spouses, and dependants not earning a money income. Dependants receive clean, serviced accommodation, cooked meals, and clean clothes; the transfers are of transformed commodities purchased from the market which embody value added by household capital and labor.

In a perceptive book, *The Household Economy*, the American writer Scott Burns (1977) observed that 'the hours of work done outside the money economy rival those done inside and will soon surpass them.' Burns saw the household as the strongest and most important economic institution—healthy, stable, and growing. Alvin Toffler (1980) took a similar optimistic view of the household economy in his classic *The Third Wave* where he describes the many ways in which tasks are being driven back from the market to the household.

The pressure to transfer paid labor costs from the market to the unpaid labor costs of the household leads to the development of self-service petrol stations, automatic bank tellers, and Internet shopping. It also leads governments to support unpaid household-based care of sick, disabled, and elderly people instead of professional care in hospitals and nursing homes. Taxation of paid labor helps drive these technology and policy changes.

4.2 *Deficiencies in the National Statistics of Work and Production*

During the latter half of the twentieth century an almost unrecognized statistical revolution took place. At its heart was the System of National Accounts (SNA) that provides the framework for estimating GDP. Governments now spend millions of dollars collecting and publishing regular quarterly statistics of GDP and of the numbers employed in this production. These data permit comparisons of economic performance through time and between countries.

Statistics of GDP and employment are not only the common discourse of economists but have been elevated universally as major tools of economic and social policy including those operating at the international level through the International Labour Office, the International Monetary Fund, and the World Bank.

In the 1980s a number of writers exposed the deficiencies in statistics of work and production. Among these are Luisella Goldschmidt-Clermont (1982) and Ann Chadeau (1985). Perhaps the most strongly advanced reason for measuring household production is to make women's work visible as exemplified by Marilyn Waring's books *Counting for Nothing* (1988) and *Three Masquerades* (1996). Throughout the world, women provide most of the labor needed in household and subsistence production. All of this work is unpaid. On the other hand, in spite of increasing participation of women in paid jobs, most paid work is still done by men.

With few exceptions, the national statistics of work and production continue to ignore the unpaid labor and economic output contributed by women (and men) through household production. At least two-thirds of the work and economic production of women, half of the world's adult population, is excluded. Discussion of public issues such as gender equality, labor market policies, wages, and income policies, to name a few, is statistically misinformed. The point is made frequently that the increase in women's participation in paid work leads to overstatement of the increase in measured economic activity, because the reduction in household production is not counted (Nordhaus and Tobin 1973, Weinrobe 1974).

The SNA definitions used to measure production cover market transactions only. They exclude household production and make no allowance for the destruction of natural resources.

These omissions have been much criticized by the women's and environmental movements. Consequently, the UN Statistical Commission, in the 1993 revision of the SNA, has recommended that national statistical offices prepare accounts for economic activities outside the presently defined production boundary. 'Satellite' accounts separate from, but consistent with, the main SNA accounts of the market economy can be prepared to show the use of natural resources or the extent of economic production by households.

Demands for the full recognition of women's economic production culminated in the Platform for Action adopted in September 1995 at the Fourth World Conference on Women in Beijing. It enjoined 'national, regional and international statistical services and relevant governmental and United Nations agencies' to develop a more comprehensive knowledge of all forms of work and employment by:

> Improving data collection on the unremunerated work which is already included in the United Nations System of National Accounts' and developing methods for 'assessing the values, in quantitative terms, of unremunerated work that is outside national accounts, such as caring for dependants and preparing food, for possible reflection in satellite or other official accounts (Platform For Action, Fourth World Conference on Women, Beijing, 1995).

As statistics currently used to measure work and valuable production are incomplete and misleading, new measures are required to produce more accurate statistical pictures.

4.3 Methods of Estimating the Household Economy

Some of the earliest estimates of the value of household services were those made for the United States (by Mitchell for 1919 and Kuznets for 1929) and for Sweden (by Lindahl, Dahlgren, and Korb for 1929 and earlier years); they simply multiplied the total number of households in rural and urban areas by the corresponding annual cost of hiring a domestic servant (Hawrylyshyn 1976).

The statistical offices of Denmark, Norway, and Sweden included household production in their national income estimates prior to World War II. Norway's estimates, started in 1912, were quite elaborate, valuing the unpaid work of married women on average wage income for unmarried women and widows in various age groups. They were discontinued in 1950 when the United Nations recommendations were adopted (Aslaksen and Koren 1996).

The Scandinavian view that the real economic production of goods and services included all goods and services produced by households, whether for their own use or for others, was unable to prevail in the debates about the initial standards for national income and product accounting. The market-based approach to GDP estimates continues. However, the way is clear through satellite national household accounts for separate estimates to be made of the value of household production, Gross Household Product (GHP).

The initial information base for such estimates is provided by surveys of the use of time on various household activities (Vanek 1974). International comparability of statistics of the uses of time was given a great boost by the cross-national time budget study in 12 countries conducted in the 1960s under the sponsorship of UNESCO and the International Social Science Council (Szalai 1972). Since then national statistical offices in many OECD countries have followed the Szalai methodology collecting diary-based national surveys of time use from representative samples of households.

The results show that the present statistics of employment greatly understate the volume of economic work needed. Surveys across 12 OECD countries covering various years from 1985 to 1992 show that the mean time spent in paid work was 24 h per adult per week. The mean average unpaid work in the household economy was 26 h per adult per week (Goldschmidt-Clermont and Pagnossin-Aligisakis 1995, Ironmonger 1995).

The representations of work provided by official statistics of employment are thus at variance with those from official statistics of work from time use surveys. In the main, official statisticians still report an

Table 1
Household productive activities, Australia, 1975–76 $million

INPUT	Intermediate inputs		Factors of production		Total
	Materials	Energy	Labor	Capital	
OUTPUT					
Meal preparation	6219	144	7922	1140	15425
Cleaning and laundry	271	32	8140	813	9256
Childcare	1208	187	4717	588	6700
Shopping	745	224	4969	453	6390
Repairs and maintenance	483	7	1825	283	2599
Other domestic work	362	7	1743	172	2283
Education and community	392	8	2088	122	2609
Gardening	46	–	1496	182	1725
TOTAL	9726	609	32900	3753	46986

Source: Ironmonger D S (1989).

outmoded view that unpaid domestic work is somehow not 'work'—only those working in the monetized market sector are counted as being 'economically active.'

Three methods of imputing a value to the time used in household production have been used, all taking a wage per hour from the market economy. The first is the 'opportunity cost' wage that a person could be paid for doing an extra hour of work in a market job rather than an hour of unpaid household work. The cost of an hour of household work is the forgone opportunity to earn in the market. This method is usually rejected since it yields different values depending on who performed the task. The second method uses the wages of specialist paid workers who come to the household, e.g., a cleaner, a cook, a nanny, a gardener, to value the same tasks performed by household members. This 'specialist replacement cost' method can be criticized because these workers work more efficiently than a usual household member can, taking less time to perform the same task. Finally, a 'generalist replacement cost' method of valuation uses the wage rate for a generalist worker or housekeeper. This is regarded as more appropriate since the working conditions and range of activities are similar to those of household members.

Time-use surveys provide only one source of information about household production since they omit the contribution from nonhuman capital (the land, dwellings, and equipment owned by households). A more complete national accounting approach to measuring and modeling the household economy is required to cover all factors of household production, all intermediate inputs, and all the principal outputs.

5. *National Household Accounts and Modeling the Household Economy*

5.1 *Extended National Accounts*

A number of scholars have suggested that national accounts needed extending to cover missing nonmarket household production. Essentially, they viewed household production as a missing part of the estimates, not a separate economy. Robert Eisner and co-workers at Northwestern University conducted the most extensive work on these lines in the 1980s. Their research culminated in the book *The Total Incomes System of Accounts* (1989). Eisner's estimates for the United States make clear the greatly expanded role, in total economic product, of households. United States GHP in 1981 was put at $1709 billion, 37.5 percent of the extended GNP of $4560 billion.

5.2 *Household Input–Output Tables and Satellite Household Accounts*

Household production is best considered as the production of goods and services by a separate economy, complementary to and competitive with the market economy. Every country can be analyzed as two linked economies. The household sells labor to the market; it uses the money income to purchase intermediate commodities from the market which it transforms into items of final consumption through the use of its own unpaid labor and its own capital goods.

The household economy can be studied as a set of six industries providing accommodation, meals, clean clothes, transport, recreation, and care in competition with parallel market industries. Activities such as shopping and cleaning are simply ancillary activities to the principal final outputs of the household economy.

The input–output approach to measuring and modeling the industries of the market economy associated with the work of Leontief (1941) has been applied to measuring and modeling the industries of the household economy. This was started in Australia for Australian households (Ironmonger 1989). Input–output tables showing the uses of labor, capital, energy, and materials in each of the main and ancillary industries of the household economy have now been

prepared for Australia, Canada, Finland, Norway, and the United States.

How the household economy uses time, materials, energy, and capital can be seen from the Table 1 adapted from the first published household input–output table. This is for the Australian household economy. The household input–output tables are exactly the satellite accounts that have been recommended in the 1993 revision of the SNA and in the resolution of the Platform for Action of the 1995 Beijing Women's Conference. The world statistical agencies need to establish standardized methods of estimating GHP.

6. *Household Production and a World of Binary Economies*

The major scientific achievement of this field has been the measurement of the magnitude of household production through surveys of the uses of time. Household production is now recognized as an alternative economy to the market; in many countries the household economy absorbs more labor and at least one-third the physical capital used in the market economy.

In future, national statistical organizations will produce regular estimates of GHP. Data on outputs of household production—accommodation, meals, clean clothes, and the care of children and adults—will complement data on inputs of unpaid labor and the use of household capital.

Proper recognition of the household economy will have arrived when national household accounts are published each quarter alongside national accounts for the market economy. These data will enable greater scientific research on the organization of household production, the interactions with the market economy, the role of households in building human capital, on the effects of household technology, alternative social and economic policies on gender divisions of labor, and on family welfare.

See also: Boserup, Ester (1910–99); Childbearing, Externalities of; Economic Development and Women; Families and Households, Formal Demography of; Family and Consumer Sciences: United States; Family and Gender; Family and Kinship, History of; Family as Institution; Family Planning Programs: Feminist Perspectives; Family Theory: Feminist–Economist Critique; Feminist Economics; Gender and Feminist Studies in Economics; Gender and Technology; Gender, Economics of; Gender History; Household Archaeology; Household in Anthropology; Land Rights and Gender; Motherhood: Economic Aspects; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Rural Industrialization in Developing Nations and Gender; Time-use and Gender

Bibliography

Aslaksen I, Koren C 1996 Unpaid household work and the distribution of extended income: the Norwegian experience. *Feminist Economics* **2**(3): 65–80

Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Boulding K 1972 The household as Achilles' Heel. *Journal of Consumer Affairs* **6**(2): 110–19

Chadeau A 1985 Measuring household activities: some international comparisons. *Review of Income and Wealth* **3**: 237–53

Eisner R 1989 *The Total Incomes System of Accounts*. University of Chicago Press, Chicago

Goldschmidt-Clermont L 1982 *Unpaid Work in the Household: A Review of Economic Evaluation Methods*. International Labour Office, Geneva, Switzerland

Goldschmidt-Clermont L, Pagnossin-Aligisakis E 1995 Measures of unrecorded economic activities in fourteen countries. Human Development Report Office Occasional Papers 20, United Nations Development Programme, New York

Hawrylyshyn O 1976 The value of household services: a survey of empirical estimates. *Review of Income and Wealth* **22**: 101–31

Ironmonger D S 1972 *New Commodities and Consumer Behaviour*. Cambridge University Press, Cambridge, UK

Ironmonger D S 1989 Research on the household economy. In: Ironmonger D (ed.) *Households Work*. Allen & Unwin, Sydney, Australia

Ironmonger D S 1995 Modeling the household economy. In: Dutta M (ed.) *Economics, Econometrics and The LINK: Essays in Honor of Lawrence R Klein*. Elsevier, Amsterdam

Ironmonger D S 1996 Counting outputs, capital inputs and caring labor: estimating gross household product. *Feminist Economics* **2**(3): 37–64

Lancaster K 1971 *Consumer Demand: A New Approach*. Columbia University Press, New York

Leontief W W 1941 *The Structure of the American Economy, 1919–1939*. Oxford University Press, Oxford, UK

Nordhaus W D, Tobin J 1973 Is growth obsolete? In: Moss M (ed.) *The Measurement of Economic and Social Performance*. National Bureau of Economic Research, New York

Reid M G 1934 *Economics of Household Productions*. Wiley, New York

Szalai A 1972 *The Uses of Time: Daily Activities and Suburban Populations in Twelve Countries*. Mouton, The Hague, The Netherlands

Vanek J 1974 Time spent in housework. *Scientific American* **231**: 116–20

Weinrobe M 1974 Household production and national production: an improvement of the record. *Review of Income and Wealth* **20**: 89–102

D. Ironmonger

Housing Economics

Housing has a set of characteristics which together make it different from all other commodities.

(a) *Necessity*: Housing satisfies a basic human need, shelter.

(b) *Importance*: Most households spend more on housing than on any other class of commodity.

(c) *Spatial fixity*: Most housing units are rooted to a particular location. Living in a particular housing unit, a household therefore consumes not only the characteristics of the housing unit *per se*, but also the characteristics of its *location*—its neighbors and neighborhood, its accessibility to other locations, its taxes and public services, and its locational amenities and disamenities.

(d) *Durability*: Housing is the most durable of major commodities. If built soundly and maintained well, a housing unit will last virtually indefinitely.

(e) *Heterogeneity and complexity*: To describe a housing unit reasonably accurately requires enumerating very many characteristics (complexity). Also, housing units exhibit considerable heterogeneity in characteristics, and households considerable heterogeneity in tastes.

(f) *Housing as an asset*: Because housing is durable and the expenditure share on housing high, the housing stock constitutes a substantial proportion of a nation's wealth and of most households' portfolios. The demand for housing therefore combines consumption and asset motives.

(g) *Housing and the macroeconomy*: Since housing is durable, modest fluctuations in housing demand translate into large fluctuations in housing investment. Due to the importance and volatility of housing investment and the magnitude of the wealth effects generated by movements in housing values, the performance of the housing market has a major impact on the macroeconomy.

The interactions of these characteristics cause the economics of housing to be significantly different from the economics of other commodities. Section 2 will examine the supply of housing and Sect. 3 the demand. Section 4 will discuss how supply and demand interact to determine equilibrium in the housing market. Section 5 will look at the housing market from the perspective of real estate economics. Section 6 will consider the role of government with respect to housing.

1. Housing Supply

Before discussing how housing is supplied, a quantifiable definition of 'housing' is needed. The appropriate definition depends on context. In some contexts, highly aggregated definitions that ignore housing heterogeneity—the number of units or total floor area—are most useful. In other contexts, each housing unit is usefully viewed as a bundle of many characteristics. In yet other contexts, intermediate levels of aggregation are most helpful. The definition employed strongly influences the way the market is conceptualized and housing supply and demand described.

Most housing economics is conducted at a high level of aggregation. A nation's monetary authority is interested in the value of the nation's housing stock and the level of housing investment, not in where housing units are located within cities or how many units have two bedrooms. Similarly, the government's housing ministry is principally concerned with housing affordability, crowding, sanitary facilities, and structural soundness, and not in how many units have fireplaces.

At the highest level of aggregation, a housing unit is described simply by the flow of housing services it provides, which can be measured by its rent. Holding its characteristics fixed, an increase in the unit's rent corresponds to a rise in the price of housing services. A housing supply curve may then be defined relating the units of housing services supplied to the price of housing services. Because housing is durable, the elasticity of the housing services supply curve depends on the time horizon. Within, say, a month, the housing supply curve is almost vertical; little can be done beyond maintenance to increase the quantity of housing services supplied. As the time horizon increases, the housing services supply curve becomes increasingly elastic; within six months, a rise in the price of housing services will induce some rehabilitation; within a year, some new housing will be constructed; etc.

In many situations, the analyst is concerned with the interaction between particular housing submarkets. For example, the government might be interested in the effects of a subsidy on owner-occupied housing. To analyze this policy, the housing market would be divided into owner-occupied and rental submarkets. The two submarkets would interact on both the supply (the subsidy would encourage the conversion of housing from rental to owner occupancy) and demand (the subsidy-induced fall in the consumer price of owner-occupied housing services would cause a decrease in rental housing demand) sides. Other divisions of the housing market into submarkets are appropriate in other contexts. If the government is concerned with the living standards of the poor, the housing market is usefully divided into quality submarkets; if it is concerned with downtown revitalization, a division into downtown and suburban submarkets is suitable; etc.

The discussion thus far has treated the supply for housing services—a flow. Lurking in the background is the supply of housing stock. The two are related through the user cost of capital, which is the ratio of the annualized cost of employing a unit of capital to its supply price, and equals the interest rate plus the rate of depreciation plus the maintenance cost rate plus the effective tax rate on capital less the rate of capital gain.

In the short run, rent adjusts to clear the market for housing services. This translates, via the user cost of capital, into the equilibrium asset price or market value of housing. If this exceeds construction cost, new

construction is triggered, which increases the supply of housing stock and of housing services, causing rent to fall. This process continues until a long-run equilibrium is attained where the market-clearing rent equals construction cost times the user cost of capital.

The above conceptualization of the housing market is especially well-suited for aggregate analysis. The remainder of the section reviews several alternative conceptualizations, each of which has its particular domain of application.

The focus of the Mills–Muth model (see Mills (1972) and Muth (1961, 1969), and Brueckner (1986) for an integrative treatment) is on urban spatial structure—how housing density (the ratio of floor to land area), housing rent, and land rent vary with location. Ignore the spatial fixity and durability of structures, so that static analysis can be applied. A landowner adds capital to his land until marginal cost equals marginal revenue. Under competition land rent equals the corresponding producer's surplus. This analysis relates land rent and housing density to housing rent at a particular location. Now turn to the market. Housing rents adjust so that in equilibrium at each location the quantity of housing supplied equals the quantity demanded, with more attractive and accessible locations commanding higher rents. Then from the relationship between housing rent and housing density, the equilibrium urban spatial structure is obtained.

The Mills–Muth model ignores the durability of structures. A related class of models, reviewed in Brueckner (2000), examines the spatial structure of a growing city, accounting for durability. A developer who owns vacant land then faces the decision of *when* to build as well as at what density. The relationship between predevelopment land value and postdevelopment housing value and density can be straightforwardly derived. Normally vacant land is developed from the center outwards. Subsequently, there are successive waves of redevelopment. Each wave entails construction at higher density with the time between waves depending on the city's growth rate. This process is evident in some cities but in others has been curtailed by redevelopment restrictions.

The Mills–Muth model and dynamic durable housing models treat location, housing density, and unit size. Housing units also differ in quality. If two housing units of the same size at the same location command different rents, the higher-rent unit is presumably of higher quality. Since the poor typically live in low-quality housing, quality differentiation is particularly important to consider when evaluating policies aimed at improving low-income housing. Models of housing quality differentiation examine the filtering process (Grigsby (1963)). An urban location goes through a life cycle. Housing is of high quality when constructed but depreciates with age, and as it does so is occupied by successively poorer households. Eventually, rehabilitation/renovation or redevelopment of the site at higher density becomes profitable, and the cycle repeats. Since the typical neighborhood was initially developed in a short time span, the filtering process gives rise to neighborhood cycles too.

More recent filtering models (Sweeney (1974), Arnott et al. (1983), and Rothenberg et al. (1991)) view the housing market as a hierarchy or continuum of quality-differentiated submarkets. At any instant, the housing market is described by the housing stock at the various qualities. Rent as a function of quality adjusts to clear the market. Construction occurs at those qualities where housing value equals or exceeds construction costs. Maintenance is carried to the point where the last dollar spent increases housing value by one dollar, and a housing unit's price is the present discounted value of rent less maintenance expenditures. These models can be enriched to treat land, as well as housing differentiation by unit size, location, and density (Anas and Arnott (1991).

Filtering models provide a strong conceptual basis for policy analysis and are intuitively appealing, accommodating all the standard housing conversion processes—construction, maintenance, rehabilitation, upward and downward conversion (combining and subdividing housing units), abandonment, and demolition and reconstruction. Despite their popularity among academic housing economists, filtering models have not yet been widely adopted by the housing policy community.

Working with detailed descriptions of housing units is normally distracting, but in measurement is essential. The minister of housing may be uninterested in whether the poor have fireplaces (as long as they receive heat), but will be interested in whether the average quality of low-income housing has increased. This requires obtaining a quality index, an aggregate of many housing characteristics—including fireplaces. Investors, housing policy analysts, and macroeconomists are all concerned with the time trend of housing prices, which requires developing price indices. And for property tax purposes, housing values need to be estimated accurately, which necessitates detailed description. Hedonic analysis utilizes such detailed description. The market values (or rents) of housing units are regressed against a long list of their characteristics (including public services, neighborhood and neighbor characteristics, accessibility measures, and environmental variables). The regression coefficients are interpreted as implicit market prices for characteristics (Rosen (1974)). Hedonic analysis has proved very useful not only in housing valuation and housing price and quality index construction, but also in inferring the market's valuation of various neighborhood characteristics, public services, air quality, etc. But the implicit conceptualization it employs—separate markets for each of hundreds of housing characteristics—is unwieldy for most housing policy analysis.

Industry studies focus on market organization—industry concentration and the form of strategic interaction between firms. The conventional wisdom

that the housing market is competitive or monopolistically competitive, while inadequately researched, appears sound.

2. Housing Demand

The previous section indicated that how housing supply is described depends on how housing is conceptualized. The same is true of housing demand. Most housing demand studies employ the housing services conceptualization. Accordingly, there is a vast empirical literature estimating the income and price elasticities of demand for housing services. Excellent reviews are provided in Goodman (1996) and Olsen (1986). Muth, whose classic *Cities and Housing* (1969) provides the foundation of modern housing analysis, has argued that since the housing expenditure share has shown remarkable stability over time and location—about 25 percent of after-tax income—the long-run price and income elasticities of housing must be close to unity. Most empirical studies, however, estimate significantly lower price and income elasticities. How can the results of these empirical studies be reconciled with Muth's observation?

The literature has concentrated on four classes of explanations.

(a) With high moving costs (including psychic costs), a household does not instantaneously change its housing when its income or the price of housing changes; rather, housing decisions are long-term. This explains why experiments investigating how households alter their housing consumption under temporary rent subsidies have found a generally price-inelastic response. Relatedly, the demand for housing should depend more on permanent than on transitory income.

(b) That many moves are triggered by demographic change (e.g., marriage, births, retirement, and divorce) makes it more difficult to identify the economic determinants of housing choice.

(c) Tenure choice with borrowing constraints (e.g., Jones (1993), Ortalo-Magné and Rady (1999)). The typical household starts its lifecycle by renting, often under-consuming housing to save up for a down-payment. Having saved up enough, it moves into a starter home, aiming to trade up when it can afford to. The timing of these moves depends on the realization of the household's uncertain income prospects and demographic changes and on the evolution of housing prices, interest rates, and down-payment requirements.

(d) The asset or portfolio demand for housing. When deciding how expensive a home to buy, a household considers how much housing it would like to include in its wealth portfolio. A household's demand for owner-occupied housing achieves a balance between its consumption and portfolio demands. Households adjust the discrepancy between these demands through tenure choice, customization and furnishing (Henderson and Ioannides (1983)). A household whose asset demand exceeds its consumption demand may purchase a large home but spend little on decoration, or may purchase additional housing units and rent them out. A household whose asset demand falls short of its consumption demand may choose either to own a cheaper house and decorate it nicely or to rent.

Estimating price and income elasticities accounting for all these complications, as well as the tax treatment of housing, is indeed a daunting task. Despite these difficulties, considerable progress has been made in the empirical analysis of housing demand, reflecting advances in data collection, computing power, applied econometrics, and housing economic theory, and the use of controlled experimentation. Regrettably, however, little work has been done on estimating filtering models (Rothenberg et al. (1991) is a notable exception), which has impeded their use in policy analysis.

3. Housing Allocation and Markets

Throughout history, most consumer goods have been allocated through markets. The twentieth century witnessed a sharp rise in the size and scope of government. The communist countries adopted central planning, which entailed government construction, operation, and direct allocation of housing. Elsewhere too, government intervention in the housing sector increased. The pendulum has started to swing in the opposite direction, but by historical standards government intervention in the housing sector remains high. Even in the United States, with the most free-market of western economies, much low-income housing continues to be allocated outside the market, and the allocation of market housing to be deliberately influenced by a myriad of government policies.

Why have governments everywhere doubted the ability of markets to allocate housing? Textbooks in microeconomic theory give a standard list of (classic) market failures or inefficiencies—public goods, natural monopoly, and technological externalities. None is important in the context of housing. The primary motive for extensive government intervention vis-à-vis housing has rather been to improve equity. Some believe that the income distribution generated by the free market is just but most do not. How then is equity to be achieved within a generally market-oriented economy? There are two broad approaches. One is income redistribution, general egalitarianism; the other is effective redistribution through the taxation/subsidization or nonmarket allocation of specific commodities, specific egalitarianism. Most economists favor general egalitarianism because it respects consumer sovereignty. But most governments evidently favor specific egalitarianism. They—and the person

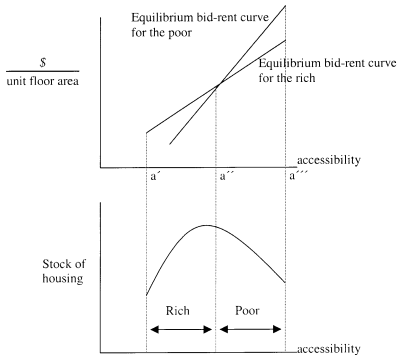

Figure 1
Equilibrium bid-rent curves

on the street—feel that the consumption of some goods (basic housing, clothing, food, education, and medical care) is more meritorious than others (alcohol, cigarettes, drugs, and luxuries). Many governments have favored the nonmarket allocation of housing over subsidization, because it permits more refined targeting and is administratively simpler. There has been remarkably little modern economic analysis of nonmarket housing allocation (Anas and Cho (1988) is a notable exception). The rest of the section will therefore consider market allocation under competition.

The relevant economic theory was developed by Alonso (1964). He asked: How does the market determine the pattern of household location within a city? Suppose that there are two household groups, rich and poor, that each household derives utility from only other goods and housing floor area, and that locations differ only in accessibility. In equilibrium, all rich households achieve the same utility level, U^R. Associated with this utility level is the equilibrium housing bid-rent curve for the rich, which indicates the housing rent (per unit floor area) payable by the rich at different locations consistent with this utility level. Similarly, in equilibrium, all poor households achieve the same utility level, U^P, associated with which is their equilibrium housing bid-rent curve.

Two basic principles apply: (a) A housing unit goes to whomever bids the most for it; and (b) rent equals the bid rent of whomever bids the most for it. The top panel of Fig. 1 plots the equilibrium bid-rent curves for the two groups as a function of accessibility, a. Applying the two principles implies that housing between accessibility levels a′ and a″ goes to the rich and that between a″ and a‴ to the poor, and that housing rent as a function of accessibility is given by the 'upper envelope' of the equilibrium bid-rent curves. The lower panel of the figure shows the stock of housing available at different accessibility levels. The equilibrium utility levels are determined by the conditions that, for each group, the quantity of housing demanded equals the quantity of housing supplied. The slope of a group's housing bid-rent curve is the premium that group is willing to pay for improved accessibility. As drawn, the poor outbid the rich at more accessible locations because they are willing to pay a higher premium per unit floor area for improved accessibility. This result has been used in explaining the post-World-War-II suburbanization of wealthier US households.

The accessibility-rent tradeoff is central in determining the pattern of household location. The above analysis can be extended to treat other factors, such as the relative value different groups attach to amenities, disamenities, public services, and neighborhood composition. Each group has an equilibrium bid-rent surface over all locations within the metropolitan area, and lives where it outbids all other groups. Groups may differ not only in income but also in demographic status, race, job location, etc. Extending the analysis to account for taste heterogeneity within groups yields the probabilities that a particular location is occupied by a household in a particular group. The same competitive allocation mechanism operates with respect to other housing attributes. A household group occupies housing of a particular quality, for example, if it outbids all other household groups.

The asset value of a housing unit is the discounted present value of expected future rents minus expected costs, adjusted for uncertainty. The volume of construction depends on profitability, as measured by asset price minus supply price; similarly, the volume of rehabilitation depends on the gain in asset price through rehabilitation minus the cost; and maintenance expenditure depends on how steeply housing value rises with quality.

The above discussion indicates how a perfectly competitive housing market works. In actual housing markets, housing units and households exhibit multidimensional heterogeneity. To find a unit that suits its tastes, a household undertakes costly search over vacant units. The friction of search confers market power on landlords, who are consequently price-setters rather than price-takers. With entry and exit close to free, actual housing markets are therefore more accurately described as monopolistically competitive, with excess capacity taking the form of vacant units. Furthermore, housing markets are matching markets, with market tightness (the ratio of searching households to vacant units) playing an important role in the market adjustment process (Anglin and Arnott (1999)). Down-payment requirements on mortgages and credit restrictions on developers also significantly affect the market's operation.

Whether the gain in accuracy from treating the housing market as monopolistically rather than perfectly competitive justifies the increased analytical complexity is an empirical issue and depends on context. Normally little is sacrificed in using perfectly competitive models to analyze the positive effects of

specific housing policies. But imperfectly competitive models should be used to provide microfoundations for housing macroeconomics since nonprice adjustment is important, and to evaluate welfare changes from housing policies.

In the frictionless world of perfect competition, and without taxes, a household would be indifferent between renting and owning. That tenure choice is in fact important indicates that it should be considered from the perspective of imperfect competition. Though often treated in the context of housing demand, tenure choice is an aspect of market equilibrium since it concerns the contractual form through which housing services are exchanged. Rental and single-family owner occupancy are the most common contractual forms; others include condominium ownership (under which common areas are jointly owned) and cooperative ownership (under which a resident is essentially a shareholder in the building).

4. *Real Estate Economics*

Real estate economics is superbly surveyed in DiPasquale and Wheaton (1996). How does residential real estate economics differ from traditional housing economics? Both are concerned with the housing sector, but real estate economics regards it from an investment perspective, housing economics from a policy perspective. At the beginning of the twenty-first century much of the best research on the economics of housing is being done under the rubric of real estate economics.

One important development has been the extension of real estate investment theory to treat uncertainty satisfactorily. Because real estate investment decisions are made under uncertainty and are largely irreversible, they are usefully viewed as the exercise of real options (Titman (1985), Capozza and Li (1994)), to which the theories of options valuation and irreversible investment can be applied.

Most housing economics is microeconomics applied to the housing sector. But housing macroeconomics is important too. Housing macroeconomics is concerned with the relationships between housing macroeconomic variables (stock, starts, completions, vacancy rates, average rents, etc.), and with how fluctuations in economy-wide variables affect the housing sector, and vice versa. Most large-scale macroeconomic forecasting models of the 1960s contained a housing module. After a period of inactivity, housing macroeconomics has reappeared on the research agenda—in real estate economics, as the study of residential real estate cycles. Current work (e.g. Stein (1995)), and DiPasquale and Wheaton (1996)) suggests that the amplitude of fluctuations in residential real estate values and investment is sensitive to the severity of borrowing constraints and the form of expectations.

Real estate economists have also explored the interlinkages between the real and financial sides of residential real estate markets, and the design of optimal contracts, considering incentive, selection, and risk-sharing effects, between banks and borrowers, developers and contractors, landlord and tenants, and between the sellers and buyers of houses and real estate agents.

5. *Housing Policy*

Housing policy debates tend to pit economists against planners. The divisions between the two groups are social philosophical. Since economists value individual freedom to choose, they hesitate more than planners in imposing their vision of the good life on others. Also, economists generally emphasize efficiency and believe it is best achieved via markets, with government enforcing property rights and intervening minimally to mitigate market failure, while planners tend to be suspicious of markets and to object to the inequities and apparent disorder they generate. For all these reasons, planners almost invariably favor more government intervention than do economists.

The planner's view dominated housing policy in western Europe from the 1920s through to the 1980s, especially in the Scandinavian countries where housing markets were largely supplanted by government allocation. The objective of decent and affordable housing for all was largely achieved, but at the cost of considerable inefficiency. During the last decade, these countries have been dismantling the Byzantine allocative and regulatory structures which had evolved, and moving to increasingly free housing markets. Thus, the economist's view is currently in the ascendancy.

Economists' belief in the efficiency of markets is at least well-reasoned. Two central results in microeconomic theory are the First Theorem of Welfare Economics, which states that under perfect competition a market equilibrium is efficient, and the Second Theorem, which states that (under convexity conditions) any efficient allocation can be achieved as a competitive equilibrium with appropriate lump-sum redistribution. If, therefore, an economy satisfies the Theorems' conditions, efficiency and equity can be simultaneously achieved if the government only enforces property rights and redistributes income appropriately, leaving everything else for markets to do, and almost all government intervention in the housing sector is harmful. Whether real world economies come close to satisfying the Theorems' conditions is clearly crucial in determining the appropriate role of government in housing provision, and therefore needs to be considered very carefully.

In the 1960s most neoclassical economists believed that, since buyers and sellers are many and entry and exit free, housing markets are almost perfectly competitive. Most housing economists in the late twentieth century would disagree. Housing markets have many

characteristics which violate the assumptions of perfect competition: housing is a heterogeneous good; transactions costs are important; asymmetric information is endemic; and insurance and futures markets in housing are highly incomplete. As well, because the government cannot observe an individual's 'need' but must instead infer it from a signal (e.g., income) which she can manipulate, all redistributive policies are distortionary. Equity is then most efficiently achieved through a portfolio of policies, including perhaps housing assistance for the poor. Since the optimal portfolio is sensitive to details of the economic environment, housing economists nowadays tend to base their policy advice as much on the results of controlled experiments and empirical studies as on *a priori* reasoning. While consideration of market imperfections expands the scope of potentially ameliorative government intervention, the political economy literature warns that government, too, is highly imperfect.

Government withdrawal from the housing sector has been especially pronounced in direct housing provision. The current view is that low-income housing projects were a mistake. They led to the alienation and ghettoization of the poor, and were excessively costly (since it is efficient for the poor to live in filtered-down housing). Low-income housing policy in the United States today encourages household mixing by race, ethnicity, and income, through subsidizing the nonpoor to live in the projects and the poor to move into middle-income neighborhoods. Other countries are experimenting with selling project units for owner occupancy.

A classic debate in housing economic policy is whether intervention is more effective on the demand or supply side of the market. Generally speaking, housing demand policies can achieve better targeting, but may drive up rents in the short run until the lagged supply response kicks in.

Virtually all housing policies apply to either rental or owner-occupied housing but not both. Most policy makers encourage homeownership, believing it to foster social behavior, civic responsibility, and commitment to neighborhood. The empirical evidence is, however, weak since controlling for the effect that responsible and well-functioning households are more likely to choose homeownership is difficult. Homeownership is encouraged through subsidized and guaranteed mortgages, mortgage interest deductibility under the income tax, tax inducements for first-time homeownership, and subsidies for home improvements.

There are several other elements common to the housing policy debate in most countries.

(a) *Rent control*: (reviewed in Arnott (1995)) Economists have been virtually unanimous in their opposition to traditional (now termed 'first-generation') rent controls, which impose ceilings on rents below market-clearing levels. Almost all jurisdictions which have had first-generation rent controls have either dismantled them or replaced them with second-generation programs which stress tenant security rather than rent reduction.

(b) *Tax policy*: Most countries use tax policies to achieve housing policy goals.

(c) *Eviction rights vs. security of tenure*: The nature of the landlord-tenant relationship bears the seeds of conflict. Policy concerning eviction for nonpayment of rent entails a classic tradeoff between risk sharing and incentives. Predictably, relative to the United States, European jurisdictions favor the tenant.

(d) *Zoning, building codes, and density and height restrictions*: These significantly affect the spatial pattern of rents and hence household location by income-demographic group. In many cities, height and redevelopment restrictions have forced up downtown property values, driving many poor households previously resident there to surrounding locations. Also, affluent North American suburbs have used minimum lot size zoning and other regulations to exclude the poor.

(e) *Antidiscrimination*: Housing audits have been used in the United States to identify discrimination. A pair of individuals, one of whom belongs to a potentially discriminated-against group, are paid to pose as prospective tenants or prospective homebuyers and are given the same resumé; differences in their treatment are then attributed to discrimination. Also, recent empirical research has attempted to identify discrimination by banks in mortgage lending.

(f) *Homelessness*: In most western countries, homelessness is now a very visible housing problem. While many causes are sociological—the weakened bonds of family and community, higher mobility, deinstitutionalization of the mentally ill, and increased drug use—economic causes are important too (O'Flaherty (1996)). The withdrawal of governments from the housing sector has caused low-quality housing rents in many cities to rise so much that homelessness is a rational response for some very low-income individuals and households.

Bibliography

Alonso W 1964 *Location and Land Use: Towards a General Theory of Rent*. Harvard University Press, Cambridge, MA

Anas A, Arnott R 1991 Dynamic housing market equilibrium with taste heterogeneity, idiosyncratic perfect foresight and stock conversions. *Journal of Housing Economics* **1**: 2–32

Anas A, Cho J R 1988 A dynamic, policy-oriented model of the regulated housing market: the Swedish proto-type. *The Review of Economic Studies with Regional Science and Urban Economics* **18**: 201–231

Anglin P, Arnott R 1999 Are brokers' commission rates on home sales too high? A conceptual analysis. *Regional Science and Urban Economics* **27**: 719–749

Arnott R 1995 Time for revisionism on rent control? *Journal of Economic Perspectives* **9**: 99–120

Arnott R, Davidson R, Pines D 1982 Housing quality, maintenance, and rehabilitation. *The Review of Economic Studies* **50**: 467–494

Brueckner J K 1986 The structure of urban equilibria: A unified treatment of the Muth–Mills model. In: Mills E S (ed.) *Handbook of Regional and Urban Economics*, Vol. 2, Chapter 20. North-Holland, Amsterdam, The Netherlands
Brueckner J K 2000 Urban growth models with durable housing: An overview. In: Huriot J M, Thisse J F (eds.) *Economics of Cities: Theoretical Perspectives*. Cambridge University Press, Cambridge
Capozza D, Li Y 1994 The intensity and timing of investment: the case of land. *American Economic Review* **84**: 889–940
DiPasquale D, Wheaton W C 1996 *Urban Economics and Real Estate Markets*. Prentice Hall, Englewood Cliffs, NJ
Goodman A C 1996 Topics in empirical urban housing research. In: Arnott R (ed.) *Encyclopedia of Economics*, Vol. 1, Harwood, Amsterdam, The Netherlands
Grigsby W G 1963 *Housing Markets and Public Policy*. University of Pennsylvania Press, Philadelphia, PA
Henderson J V, Ioannides Y 1983 A model of housing tenure choice. *American Economic Review* **73**: 98–113
Jones L 1993 The demand for home mortgage debt. *Journal of Urban Economics* **33**: 10–28
Mills E S 1972 *Studies in the Structure of the Urban Economy*. The Johns Hopkins University Press, Baltimore, MD
Muth R 1961 Economic change and rural—urban land conversion. *Econometrica* **29**: 1–23
Muth R F 1969 *Cities and Housing*. University of Chicago Press, Chicago
Oates W E 1972 *Fiscal Federalism*. Harcourt Brace Jovanovich, New York
O'Flaherty B 1996 *Making Room*. Harvard University Press, Cambridge, MA
Olsen E O 1986 The demand and supply of housing service: A critical survey of the empirical literature. *Handbook of Regional and Urban Economics*, Vol. 2, Chapter 25
Ortalo-Magné F, Rady S 1999 Boom in, bust out: Young households and the housing price cycle. *European Economic Review* **43**: 755–66
Rosen S 1974 Hedonic prices and implicit markets. *Journal of Political Economy* **82**: 35–55
Rothenberg J, Galster G C, Butler R V, Pitkin J 1991 *The Maze of Urban Housing Markets*. University of Chicago Press, Chicago
Stein J C 1995 Prices and trading volume in the housing market: a model with down-payment effects. *Quarterly Journal of Economics* **110**: 379–406
Sweeney J L 1974 A commodity hierarchy model of the rental housing market. *Journal of Urban Economics* **1**: 288–323
Titman S 1985 Urban land prices under uncertainty. *American Economic Review* **75**: 505–14

R. Arnott

Human Behavioral Ecology

Human behavioral ecology (sometimes referred to as evolutionary ecology) is the application of evolutionary biological models to the study of behavioral variation in humans. Evolutionary approaches to the study of human behavior fall under a variety of names such as sociobiology, biosociology, biocultural or biosocial science, human ethology, socioecology, and evolutionary psychology (Cronk 1991, p. 25). While some of these names are synonyms others represent different emphases that will be described below. Human behavioral ecology attempts to explain behavioral diversity as a consequence of environmentally contingent responses made by individuals in their attempts to maximize their inclusive fitness. Critical to this goal is the specification of environmental factors that constrain an individual's attempt to maximize fitness. As a consequence, individuals develop behavioral strategies designed to solve a variety of adaptive problems set by nature such as producing food, mating, investing in offspring, and managing social interactions between offspring, mates, kin, and unrelated individuals.

1. *Origins and Early History of the Field*

Behavioral ecology is a relatively new discipline and, for the most part, it is practiced by anthropologists. The publication of E. O. Wilson's *Sociobiology* (1975) along with the earlier work of biosocial anthropologists such as Robin Fox and Lionel Tiger (Cronk 1991) and evolutionary biologist Richard Alexander (1974) were critical in stimulating interest in applying Darwinian approaches to the study of human behavioral diversity. Very quickly Wilson's sociobiological approach was found to be limiting because it was largely based on older ethological approaches which posited relatively simple, general, and rigid models of the relationship between genes and human behavior (Winterhalder and Smith 1992, p. 9). This was especially apparent in his final chapter of *Sociobiology* (1975, pp. 547–75) where he considered humans. Here he suggested that certain behavioral traits such as territoriality among foraging peoples were universal (1975, pp. 564–5). This claim and the approach that generated it was shown to be limited in a paper by Dyson-Hudson and Smith (1978), a paper which represented perhaps one of the earliest applications of behavioral ecology to the study of human behavioral variation. Using a cost-benefit model developed by the avian behavioral ecologist J. Brown, they showed that territoriality was variable among tribal populations and was determined by specific environmental factors consistent with evolutionary theory.

As will be more fully described below, environmental cost-benefit approaches in the context optimization models is at the core of much of behavioral ecology. It is vital to understand that behavioral ecology is a branch of evolutionary biology and the core of evolutionary biology is about changes in the frequencies of genes through time. However, behavioral ecology is agnostic about the causative role of genes in the study of human behavioral variation. This position was clarified by Alan Grafen (1984) with his notion of the phenotypic gambit. Phenotypic variation

is shaped by an individual's genotype interacting with a variety of cultural, developmental, environmental, and other non-genetic factors. The degree to which variation in any trait is a consequence of genetic variation is not directly relevant to the formulation or evaluation of behavioral ecological models. This approach is consistent with analyses in non-human behavioral ecology (Alcock 1989).

1.1 Relationship to Other Fields Using Evolutionary Approaches to the Study of Human Behavioral Variation

The new field of evolutionary psychology (Barkow et al. 1992), and, to some extent, its congener human ethology, is often inappropriately confused with behavioral ecology and sociobiology, and it is useful to distinguish between these complementary Darwinian approaches to the study of human behavioral variation. Behavioral ecology attempts to develop hypotheses regarding variation in behavioral strategies that individuals employ to maximize their inclusive fitness. In many studies the central focus is on the reproductive consequences of behavior. In contrast, evolutionary psychologists are interested in elucidating the cognitive mechanisms (or mental modules) that evolved in humans in the 'environment of evolutionary adaptedness.' This concept stresses the fact that all adaptations are to historic environments and whether traits developed in the past are adaptive in current environments is problematical. As a consequence, evolutionary psychologists seek to uncover those mental modules that were designed by natural selection to solve problems such as: What criteria do people use to select mates? How does mate selection criteria differ between the sexes? What factors induce individuals to continue or abandon mating relationships? From this perspective, the reproductive consequences of behavior are not of concern. Other differences distinguishing these fields include behavioral ecology's reliance on naturalistic fieldwork in contrast to questionnaire survey and laboratory experiment, and the relationship between behavior and fertility. Winterhalder and Smith (2000) should be consulted for a more detailed examination of the assumptions, methods, and goals of these fields.

Dual inheritance theory (Boyd and Richerson 1985) is founded on the realization that cultural traits or memes are in some sense identical to genes because both are replicators that use vehicles (bodies) to transmit themselves across generations and are subject to selective forces. Cultural traits differ from genetic traits in their mode of transmission: they may be transmitted from mind to mind through natural selection of carriers of the trait (the trait has a positive effect on fitness), through emulation of high status people who exhibit the trait, by coercion, or by psychological mechanisms that make the trait desirable or habit forming (e.g., various addictions). In addition, cultural traits may be transmitted laterally within generations instead of lineally through generations allowing them to spread more rapidly. Compared to behavioral ecology and evolutionary psychology little empirical work has been done employing dual inheritance models.

1.2 Relationship to Traditional Social and Behavioral Sciences

Behavioral ecology is distinguished from traditional anthropological and social science theory in its emphasis on the individual as the unit of selection and the idea that much of the content of culture is a consequence of decisions by individuals to enhance their inclusive fitness. Ever since its founding as a discipline, anthropologists have argued that individuals sacrifice themselves to enhance the integration or survival of the social systems in which they participate or are much like puppets or robots manipulated by culture (Cronk 1991). Furthermore, any attempt to reduce or link cultural behavior to lower levels of analysis was specifically condemned. Human behavior was believed to have emergent properties irreducible to the facts of psychology and biology. While it is clear that some dimension of human behavior are not usefully reduced to underlying psychological or biological processes simple observation tells us that humans are biological and psychological entities and some of the behavioral diversity can be usefully linked to underlying levels of analysis.

1.3 Theoretical Armamentarium

Behavioral ecology is strongly informed by a series of theoretical breakthroughs that developed in the early 1960s and 1970s in the evolutionary analysis of social and foraging behavior. These include inclusive fitness (or kin selection, Hamilton 1964), reciprocal altruism (Trivers 1971), sexual selection and parental investment (Trivers 1972), sex biasing in parental investment (Trivers and Willard 1973), optimal foraging theory (Winterhalder and Smith 2000), and, most recently, life history theory (Stearns 1992). These papers clearly demonstrated that behavioral variation was an apt subject for investigation from an evolutionary perspective consistent with the modern synthesis and a gene centered view of evolution. Early on in the development of behavioral ecology there was considerable expectation that kin selection would inspire considerable research especially in the areas of food exchange and other forms of cooperative behavior. As will be evident below, few studies have made kin selection a central focus of investigation.

2. Areas of Application

2.1 Foraging Theory

Human dietary selectivity has been a fundamental area of interest for anthropologists because of its strong implication for the structuring of social relations among hunter-gatherers, group size, mobility patterns, and our understanding of economic dimensions of human evolution. One of the earliest applications of behavioral ecological research came with the application of optimal foraging theory (OFT) in the study of hunting, fishing, and gathering (or foraging behavior) among modern hunter-gathers and foraging horticulturalists. OFT allows researchers to develop a large set of fundamental hypotheses that predict which food resources foragers will pursue when encountered during a search (diet breadth), or where foragers will travel to search for resources (patch choice models) and how long they will stay in these places before moving to other areas (marginal value theorem). These models assume that foragers have a goal, which is to maximize their net rate of return while foraging because by doing so they are ultimately able to maximize their fitness. The rate of return is indexed by a currency such as calories or another attribute of food value (e.g., Hill 1988). This assumption is reasonable because food resources are fundamental to survival and reproduction for one of two reasons. When the environment is poor, humans can be said to be 'resources limited.' Therefore, maximizing one's net rate of return allows foragers to acquire the greatest amount of food while foraging. On the other hand, when the environment is relatively rich, humans may be 'time limited' such that efficient acquisition of food resources leaves foragers with more time to engage in alternative fitness enhancing activities such as child care (Winterhalder and Smith 2000).

There have been a number of qualitative and quantitative tests of optimal foraging models applied to the problem of diet breadth (Hill and Kaplan 1992). Qualitative tests predict directional tendencies in prey choice by showing that diet breadth (the number of different species that will be pursued upon encounter), for example, will expand or contract depending on the encounter rates of highly profitable species (Winterhalder and Smith 2000). Quantitative tests predict the precise number of species a forager will pursue under a variety of environmental conditions (Hill and Kaplan 1992), which species will enter or drop out of the optimal diet breadth, and the exact rank-ordering of species in the diet breadth. Most studies of diet breath show a division of labor between men and women, and children and adults (Hill and Kaplan 1992) which seem to be based on game versus vegetable products for the former and skill and strength levels for the latter.

Although OFT was designed to account for foraging behavior in animals and then later extended to human foraging, recent applications have been made in the areas of the origins of agricultural and pastoral food production with considerable success (Winterhalder and Smith 2000). It has rekindled anthropological interest in the problem of conservation among native peoples. Thus far, research indicates that true conservation is rare among native peoples but even more importantly researchers are attempting to specify the conditions under which conservation is likely to emerge and an adaptive strategy.

2.2 Group and Resource Transfer

The movement of goods, and, to some extent, services, between families and individuals has received considerable attention by behavioral ecologists. Traditionally such research is subsumed under the terms of exchange or reciprocity. Behavioral ecologists term such activities transfers because it more neutrally characterizes resource movement between groups and/or individuals compared to traditional terms such as exchange or reciprocity (which assume that the transfer of good or services between individuals will be returned in some material form at some point in time). In the analysis of transfers, one assumes that giving or yielding a resource that one has acquired will be done such that acquirer of the resource gains something in return or avoids a cost in defending that resource. As a consequence, models of transfer assume that the benefits of giving are greater than the costs of monopolizing the resource. The models of transfer developed thus far include (a) scrounging (or tolerated theft); (b) risk minimization; (c) trade or exchange; and (d) the show-off hypothesis.

At one extreme, coercive models of scrounging or tolerated theft assume that the resource acquired cannot be easily or effectively defended against those who lack resources and that the benefit of giving or yielding a portion of the resource presents a higher payoff than the cost of defending it against another who are in greater need and thus more willing to contest the resource. At the other extreme are models of sharing, which assume that individuals give resources to others contingent upon an expectation that something of value will be returned to the giver in the future. For example, when resources cannot be predictably acquired (especially true in hunting) risk minimization or pooling strategies are employed through the mechanism of reciprocal altruism. Individuals who acquire resources share with those who have been unlucky in the hunt with the expectation that in the future should they return home without game others who have acquired game will share with them. Kin selection models have been used to explain why resource transfers tend to be biased toward close kin; sexual selection models (e.g., 'show off hypothesis') argue that actors attempt to acquire large but

high-variance resource packets (e.g., large game animals) to demonstrate their desirability as a mate; and trade models suggest that transfer occurs through specialization via the mechanism of reciprocal altruism.

2.3 Mating Strategies

Models of mating or marital strategies, in one form or another, are largely founded on the polygyny threshold model (Winterhalder and Smith 2000, p. 64). This model is relevant when males differentially control resources necessary for successful reproduction. The polygyny threshold model predicts that females will select mates according the quantity of resources they control. In this situation, polygyny will emerge among those males who can offer more resources to females than unmarried or monogamous males. In short, wealthy males are much more likely to mate polygynously than poor males. In most cases, ownership of material resources (e.g., agricultural land) or high productive productivity (e.g., ownership of large herds) is what differentiates poor from wealthy men. However, in some societies where economic stratification is absent, some men are nevertheless exceptionally productive (e.g., excellent hunters) or have achieved political power or status (e.g., distinguished warriors, leaders, or shamans) and this may be the differentiator between polygynous and monogamous men.

The polygyny threshold model assumes female choice in marital affairs. In some cases this assumption is clearly suspect and male coercion may determine a woman's marital fate. At the same time, some researchers have identified counter strategies that females use to deal with coercive male mating tactics (Winterhalder and Smith 2000, pp. 60–1). Equally troubling is the fact that in some cases female fertility declines under polygyny (this is especially true for junior wives). At this point it is clear that polygyny will nearly always enhance a man's fertility but it frequently impairs a woman's. This latter issue is leading researchers to seek factors that could compensate for lowered fertility. For example, it may be the case that although a woman may bear fewer offspring under polygyny she may produce more grandchildren. This would be true if her sons were polygynous, married earlier than average, or married women with high reproductive value.

Although polygyny is permitted in nearly three-quarters of all anthropological populations, monogamy and the rare case of polyandry (a woman having more than one husband simultaneously) are being investigated through modified versions of the polygyny threshold model. In the case of monogamy it is becoming clear that extramarital mating (concubinage, affairs) and remarriage following divorce (serial monogamy) is more common among wealthy males. Although such systems are officially monogamous, from a reproductive point of view they produce the high male reproductive variance characteristic of polygynous systems.

2.4 Reproduction: Mating, Parenting, and Life History Strategies

Given that the costs of reproduction vary greatly between males and females, they face fundamentally different trade-offs between the allocation of mating and parental effort in reproduction. The higher costs of reproduction for females limit the number of offspring they can produce and pursuit of additional males through mating effort will have little impact on the quantity or quality of offspring they can produce. As a consequence, they are concerned with the genetic quality of their mates (e.g., ability to deal with pathogens) since these qualities will be passed to offspring, and their ability and willingness to invest in offspring given the fact that humans require some degree of biparental care.

Studies of parental effort have focused on the issues of birth spacing, differential investment in offspring, and the interaction between mating and parenting. The ornithologist Lack was the first to recognize the trade-off between offspring number and offspring quality (or the ability of offspring to thrive and eventually reproduce). Since parental resources are limited, more offspring, through a shortening of the birth interval, will mean less investment per offspring and the potential for reducing each offspring's fitness. This trade-off suggests that there will be an optimal number of offspring that maximizes a parent's fitness and this number will vary with environmental circumstances. In general, one would predict that where parental resources are abundant birth intervals should be short and where parental resources are scarce birth intervals should increase. The crosscultural evidence from natural fertility traditional societies strongly supports the relationship between parental resources and reproductive success: wealthy parents or culturally successful parents have higher fertility than poor parents.

Paradoxically, the association between cultural success and fertility is absent in modern contracepting societies. The situation has led behavioral ecologists to reexamine the demographic transition that began in Western countries during the eighteenth and early nineteenth centuries, which has now spread to developing nations. During this transition fertility fell faster than mortality rates fell yielding smaller families while family wealth increased. One solution to this paradox has been to investigate whether parents are investing more in children (e.g., education) thereby creating 'higher quality' children who will be able to compete successfully in highly competitive labor markets. At the same time, these children are delaying reproduction as they pursue higher education and gain entry into high status occupations, which require

further delays in reproduction as they work hard to achieve secure occupational niches. This theoretical perspective derives from life history theory described in Sect. 2.4.3.

In a seminal article, Irons (1979) suggested that there ought to be a correlation between cultural and reproductive success. Cultural systems are structured so as to encourage individuals to achieve certain kinds of culturally valued statuses. These achievements vary culturally depending on local environmental constraints. Both in traditional and in a number of recent societies, which have not entered the most recent demographic transition, culturally valued statuses are achieved through hunting prowess, religious rank, herd ownership, political leadership, occupational status, or land holdings. At this point there are nearly 20 studies showing in nearly all cases a positive correlation between cultural achievement and fitness (either fertility or survivorship of offspring) or factors strongly associated with fitness (number of wives, frequency of extra-pair copulations). However, it is very important to note that this relationship is demonstrated for male but not female reproductive success. Significantly, attributes that lead to high male fitness such as polygyny decrease female fitness at least over the short-term but perhaps not the long term.

2.4.1 Parental investment. In general the amount of investment an offspring receives from parents depends on (a) the relatedness between parent and offspring; (b) the effect of investment on the offspring; and (c) the effect of investment on parents' future reproduction and survival. These hypotheses have been positively evaluated using crosscultural and ethnographic cases studies. For example, the probability of infanticide, abuse, and neglect has been shown to be strongly sensitive to these factors. Infants who are unrelated to a woman's current husband ((a) above), or who have congenital defects ((b) above), or who are born in times of economic privation ((c) above) receive little or no investment. The thinking explicit in this approach is that parents adjust their investment in children according to their perceptions of a child's quality, paternity, current environmental conditions, and the needs of other offspring. Further investigations into differential parental investment have been inspired by the Trivers and Willard (1973) model where sex of offspring may influence investment. This model predicts that wealthy parents will invest more in the sex that has greatest variance in reproductive success (usually males) while poor parents will invest more in offspring with least reproductive variance (usually females). Research on societies as diverse as hunter-gatherers, pastoralists, to eighteenth-century Germany and nineteenth-century India has demonstrated the predicted differential investment by sex and distortions in sex ratios.

2.4.2 Parenting as mating effort. Providing care and giving food to children are conventionally seen as kinds of parental effort by males. However, recent research reveals that care and food provisioning are sometimes allocated to children who are unrelated to males. Consequently, such behavior may be fruitfully viewed as attempts to create or maintain a mating relationship. Essentially males are attempting to demonstrate their ability to invest in offspring and therefore convince a female that they would be a valuable partner.

2.4.3 Life history. Life history strategies deal with age specific schedules of mortality and fecundity and the traits that are directly a result of these schedules (Hill and Kaplan 1999, p. 397). It deals with important developmental events such as age at first and last reproduction, growth rates, birth intervals, and senescence. The theory is based on the assumption that effort allocated to somatic effort (growth, development, and maintaining the body) cannot be allocated to reproductive effort (e.g., mating and parenting). Formal research in this area is relatively new and has focused on the peculiar human trait of menopause, inter-birth intervals among foragers, and demographic transition theory.

The grandmother hypothesis attempts to explain the evolution of menopause, a life history trait that is exceptionally rare among mammals and well documented in humans. The model predicts that women will cease reproduction prior to death in order to insure required investment in current dependent offspring and to assist their children in the rearing of grand offspring. Although attractive, the model does not, as yet, show that continued reproduction until death is not a superior alternative.

Throughout history, humans have passed through a number of demographic transitions beginning with a high mortality and high fertility regime to today's worldwide transition to a low mortality and low fertility regime. A central issue behavioral ecologists are confronting is the apparent paradox between fertility and wealth. In tribal and in many state-level historical populations there is a consistent positive correlation between fertility and wealth. This means that the traditional human pattern was to convert wealth into reproduction. However, at the end of the twentieth century there is a global tendency for parents to reduce reproduction where wealth increases. Several explanations are offered for this trend. An initial set of explanations suggested that parents are now concerned with maximizing the number of offspring over more than one generation by reducing risk through an increase in heritable wealth. The 'human capital' perspective notes that in the current environment there is an increased requirement of parental investment needed to make children competitive in a modern

labor market where earning capacity is determined by parental investment in such things as education. Finally, more recent resolutions of the paradox focus on risk sensitive strategies that seek to lower the variance in expected completed family size. An important point evolutionary psychologists make in this problem is that, for the first time in human history, we possess safe and effective means of contraception, an alternative that humans did not previously have as they made their reproductive decisions. This has led to research demonstrating that there is a kind of correlation between wealth and reproductive success, at least for men. This research shows that wealthy men have greater numbers of premarital and extramarital unions than poor men. This lends some credence to the perspective that some of the adaptations we currently possess cannot be easily understood by looking at the reproductive consequences of behavior.

3. *Summary*

Human behavioral ecology is a young and explicitly Darwinian approach to the study of human behavioral variation. It redresses an absence of evolutionary considerations in the analysis of human behavior and social organization in the social and behavioral sciences. Since its inception in the late 1970s it has spawned the development of new journals and academic specialties explicitly devoted to training students to query the evolutionary bases of behavior. Its penetration into academic life has been uneven. It is most prominent in those disciplines such as anthropology and psychology where human evolution is considered relevant for a holistic understanding of human behavior or in economics where self-interest, methodological individualism, and maximization are key elements in modeling behavior. It has been resisted in departments such as sociology and political science where the findings on human evolution are either ignored or regarded as irrelevant.

See also: Adaptation, Fitness, and Evolution; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Ecology, Cultural; Environmental Psychology: Overview; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolution: Optimization; Evolutionary Social Psychology; Evolutionism, Including Social Darwinism; Human Ecology; Lifespan Development: Evolutionary Perspectives; Natural Environmental Psychology; Psychological Development: Ethological and Evolutionary Approaches

Bibliography

Alcock J 1989 *Animal Behavior: An Evolutionary Approach*. Sinauer, Sunderland, MA

Alexander R 1974 The evolution of social behavior. *Annual Review of Ecology and Systematics* **5**: 325–83

Barkow J, Cosmides L, Tooby J (eds.) 1992 *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, Oxford, UK

Boyd R, Richerson P 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Cronk L 1991 Human behavioral ecology. *Annual Review of Anthropology* **20**: 25–53

Dyson-Hudson R, Smith E A 1978 Human territoriality: An ecological reassessment. *American Anthropologist* **80**: 21–41

Grafen A 1984 Natural selection, kin selection, and group selection. In: Krebs J R, Davies N B (eds.) *Behavioural Ecology: An Evolutionary Approach*. Sinauer, Sunderland, MA, pp. 62–84

Hamilton W 1964 The genetical evolution of social behavior. *Journal of Theoretical Biology* **7**: 1–52

Hill K 1988 Macronutrient modifications of optimal foraging theory: An approach using indifference curves applied to some modern foragers. *Human Ecology* **16**: 157–97

Hill K, Kaplan H 1992 The evolutionary ecology of food acquisition. In: Smith E, Winterhalder B (eds.) *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, New York, pp. 167–202

Hill K, Kaplan H 1999 Life history traits in humans: Theory and empirical studies. *Annual Review of Anthropology* **28**: 397–430

Irons W 1979 Cultural and biological success. In: Chagnon N, Irons W (eds.) *Evolutionary Biology and Human Social Behavior: An Anthropological Perspective*. Duxbury Press, North Scituate, MA, pp. 257–72

Smith E, Winshie in press Conservation and subsistence in small-scale societies. *Annual Review of Anthropology*

Stearns S 1992 *The Evolution of Life Histories*. Oxford University Press, Oxford, UK

Trivers R 1971 The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–57

Trivers R 1972 Parental investment and sexual selection. In: Campbell B (ed.) *Sexual Selection and the Descent of Man: 1871–1971*. Aldine, Chicago, pp. 136–79

Trivers R, Willard D 1973 Natural selection of parental ability to vary the sex ratio of offspring. *Science* **179**: 90–2

Wilson E O 1975 *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA

Winterhalder B, Smith E 1992 Evolutionary ecology and the social sciences. In: Winterhalder B, Smith E (eds.) *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, Hawthorne, NY, pp. 3–23

Winterhalder B, Smith E 2000 Analyzing adaptive strategies: Human behavioral ecology at twenty-five. *Evolutionary Anthropology* **7**: 51–72

R. Hames

Human Capital: Educational Aspects

The notion of 'human capital' is an economic concept, based on the distinction between two types of economic activity: investment and consumption. Economists use the term 'investment' to refer to an activity or use of resources which will generate income in the future, in contrast with 'consumption,' which produces immediate satisfaction or benefits, but does not create capacity to earn future monetary

income. Traditionally, economic analysis of investment and capital tended to concentrate on measurement of physical capital, namely buildings, machinery, or other equipment, which generate income by creating or enhancing productive capacity. More than 200 years ago classical economists, notably Adam Smith in *The Wealth of Nations* (1776), pointed out that education could be regarded as a form of investment, since it is likely to improve the productive capacity of workers, and thus increase their lifetime earnings, just as investment in physical capital, such as building a factory, or buying a new machine, will generate monetary income in the future.

This analogy between investment in physical capital and investment in human capital was recognized by several eighteenth- and nineteenth-century economists, but not fully developed until the 1960s when US economists such as Schultz and Becker published a series of influential works, notably *Investment in Human Capital* (Schultz 1961) and *Human Capital* (Becker 1964), which explored the idea and implications of investment in human capital by attempting to measure, for the first time, the rate of return to investment in education and training. This new emphasis on empirical analysis of human capital, described by another US economist, Mary-Jean Bowman (1966), as 'the human investment revolution in economic thought' proved to be a powerful influence, not only giving rise to a whole new branch of economics—the economics of education—but also affecting education policy in both industrialized and developing countries in a number of ways. This contribution summarizes research and literature on educational aspects of human capital, discusses ways in which estimates of rates of return and the links between education, employment, and economic growth have influenced education policy, particularly on the financing of education, and examines a number of criticisms of the concept of human capital, including the idea that education often represents a 'screening device,' rather than a genuinely productive form of investment, and the more recently developed concept of 'social capital,' which challenges narrow definitions of human capital (see *Education and Economic Growth*; *Education and Employment*; *Education and Income Distribution*).

1. *Measuring the Rate of Return to Investment in Education*

The notion that education is investment in human capital represents more than a simple analogy between education and investment in physical capital. It implies that it is possible to measure the returns to investment in education, and to apply cost–benefit analysis to decisions about education expenditure, in the same way as rates of return are used to analyze the profitability of investment in conventional physical capital. Cost–benefit analysis is designed to express the costs and benefits of a project in monetary terms, and to calculate its expected profitability, or rate of return, by comparing all future benefits with the total costs of the investment. As noted above, there were several early references to the idea of human capital in eighteenth- and nineteenth-century economic literature, but the first serious attempts to measure the economic costs and benefits of education, and compute rates of return, were in the USA in the 1960s. (Schultz 1961, Becker 1964). These were followed quickly by estimates of the rate of return to education in many other countries. Psacharopoulos (1973) presented international comparisons of rates of return to education, both as a social and as a private investment, in 32 countries, and subsequently published several attempts to update and summarize estimates of rates of return around the world, including more than 50 industrialized and developing countries (see, for example, Psacharopoulos and Woodhall 1985, Psacharopoulos 1994).

Such estimates of the rate of return to educational investment for the individual (the private rate of return) or to society as a whole (the social rate of return) rest on several crucial assumptions. Cost–benefit analysis of education assumes that the economic benefits of education can be measured in monetary terms by using the average earnings differentials between more and less educated workers as a measure of their superior productivity. Workers with higher levels of education have better employment opportunities, and therefore higher average lifetime earnings, than those with less education, and these additional earnings, after various statistical adjustments, are used by economists as a proxy measure of the direct economic benefits of education, and compared with the private and social costs of education (including costs of tuition and earnings forgone by students in school or university). Most rate of return estimates concentrate on these direct monetary benefits, and either ignore or sidestep the problem of measuring the indirect benefits to society (sometimes called 'spillover' benefits or 'externalities') of investing in education. More attention has been paid recently to indirect benefits, but most rate of return estimates still rely mainly on the direct monetary benefits of investment in human capital.

International comparisons of rates of return to education (for example Psacharopoulos and Woodhall 1985) reveal certain general patterns:

(a) Social returns are consistently lower than the private rate of return.

(b) Social and private rates of return to primary education tend to be higher than rates of return to secondary or higher education.

(c) The rate of return to education is higher in developing countries than in developed countries.

(d) The rate of return to investment in education is higher than the average rate of return to physical capital in developing countries, though not necessarily in developed countries.

2. *Implications of Rates of Return for Education Policy*

Analysis of costs, benefits, and rates of return on investment in human capital raises important implications for both economic and education policies. The idea that education contributes to economic growth by developing and disseminating knowledge, and by improving the quality and productivity of the labor force, is now widely acknowledged and has influenced the economic strategies and educational planning of many individual governments, as well as international agencies and multilateral agencies such as the World Bank. For example, a study for the World Bank (Psacharopoulos and Woodhall 1985, p. 3) notes that:

> Education, like other forms of investment in human capital, can contribute to economic development and raise the incomes of the poor just as much as investments in physical capital, such as transport, communications, power, or irrigation. The World Bank … has long recognized the importance of investment in education and has been active in this field since 1962. When the Bank was first established in 1944, however, education was not counted among the productive purposes for which it was authorized to provide investment capital. This attitude prevailed right through the 1950s, when official statements of Bank policy took the view that the Bank should concentrate its lending on projects designed to make a direct contribution to the productive capacity of its members, rather than finance projects such as the construction and equipment of schools, colleges and universities. Even though such projects were considered essential to the development of a country, the Bank's proper role, as a bank, was said to be confined to providing loans for the more directly productive sectors of the economy. Gradually, this attitude changed, however, both within the World Bank and in the world at large.

More recently, the 1998–9 World Development Report (World Bank 1999), devoted to *Knowledge for Development*, emphasized that 'knowledge is crucial for development,' arguing that education as a form of investment in human capital was a 'key ingredient' in the economic growth and success of East Asian economies such as Hong Kong (China) the Republic of Korea and Singapore (sometimes called the East Asian Miracle economies). (See also *Education and Economic Growth* for a more extended discussion of this topic.)

Policies determining investment by governments, lending by agencies such as the World Bank, and aid policies of bilateral and multilateral agencies have been influenced not only by the assumed and often demonstrated relationship between education and economic growth, but also by the consistent finding that the rate of return to primary education is the highest among all educational levels, particularly in developing countries. Studies of the economic benefits of primary education were influential in building support for the World Conference on Education for All at Jomtien in 1990, and for the subsequent emphasis on achieving universal primary education as an economic as well as an educational goal (see Colclough and Lewin 1993).

Rate of return estimates also have important implications for policies on education finance. The fact that private rates of return to educational investment are higher than the social rate of return, particularly for higher education, has been used to justify the introduction of tuition fees or student loans at the university level in several countries. For example, when the UK government first announced the introduction of student loans in 1988, the Department for Education and Science (1988, p. 10) presented estimates of social and private rates of return and argued that:

> the return on the individual graduate's personal investment in higher education … in recent years has been in the region of 25 percent, taking all subjects together. The same analysis shows that the return to society on its investment in the graduate's higher education has been much lower: between 5 percent and 8 percent. The individual graduate benefits more than the community as a whole from the latter's investment in the former's higher education. The division of costs, on the other hand, goes the other way. That is why there is scope for the individual student to bear more of the cost of the investment.

Critics argue that such arguments ignore, or at least tend to underestimate, the indirect benefits of education, which accrue to society as a whole as much as (or even more than) the individual. These include the inter-relationships between education, family health, fertility and child mortality, which are now recognized as crucial components of the benefits of education for girls and women in developing countries. An early attempt to analyze differences between rates of return for males and females, (Woodhall 1973) emphasized that 'human capital' should not be interpreted as 'investment in man,' being just as relevant to women as to men, but there was little progress in developing measures of social or indirect economic benefits of education until the 1990s, when a study of *Women's Education in Developing Countries* (King and Hill 1993) included estimates of the returns to women's education which demonstrated clearly the importance of social as well as monetary returns to education for women and girls. This led Larry Summers, then chief economist of the World Bank, to declare, in a Foreword to his book 'I have become convinced that once all the benefits are recognized, investment in the education of girls may well be the highest return investment available in the developing world.'

(Summers 1993, p. v). This has crucial implications for policies to improve equity by reducing barriers to female participation in education.

3. Objections to Human Capital Theories: Education as a 'Screening Device'

Although the concept of human capital has a long history in economic theory it remains a controversial idea, when applied to education, and there have been many objections and criticisms to the notion of measuring the rate of return to educational investment. The earliest discussions of human capital suggested that education or training raised the productivity of workers, and hence increased their lifetime earnings, by imparting useful knowledge and skills. However, this assumption was challenged by critics who argued that the earnings of educated workers reflected their superior ability, rather than the specific knowledge and skills acquired during their education. Moreover, critics argued that highly educated workers were more likely to come from higher social or income groups, to work in favored occupations and in urban rather than rural areas. Many estimates of rates of return to education now adjust the observed earnings differentials of educated people to allow for the observed influence of other factors, such as ability or social class background, on earnings, or in the absence of suitable data, they simply assume that 60 percent, rather than the full 100 percent of observed earnings differentials, are due to education, rather than ability or other factors (an adjustment that is often called the 'alpha coefficient' in the literature on the economics of education).

Many critics go further, however, and suggest that education does not improve productivity by imparting knowledge and skills, but simply acts as a 'screening device,' which enables employers to identify individuals of high innate ability or with certain personal characteristics, attitudes, or motivation, which employers value and which they therefore reward by means of higher earnings. This objection to theories of human capital, often called the 'screening hypothesis,' has attracted considerable controversy. Dore for example, in *The Diploma Disease*, an influential critique of the effects of education in developing countries, (Dore 1997, p. 93), presents many criticisms of what he calls the 'rate of return argument,' culminating in the assertion:

> We observe people who have certain kinds of education subsequently (on average) getting higher levels of income than other people with less education. But how do we know that this enhanced earning power is *because* of their education (implying on a broader scale, that more education in a society would make that society more productive)? Is the 'investment' metaphor genuinely justified? Is it true that the education *changes* a man, *enhances* his capacity to work and in this way alters the price at which he can sell his labours?

Dore gives numerous examples of educational qualifications being used as a 'screening device' or 'filter' in selecting job applicants—an aspect of what he calls the 'diploma disease'—and contrasts the 'screening hypothesis' with 'traditional "human capital theorists" who persist in believing that people earn more because their education has *changed* them and that the "investment" assumptions of their rate of return analysis are correct' concluding that 'different institutional practices, different social structures, different educational traditions and degrees of equality of opportunity make the intervening mechanisms between schooling and earnings levels very different in different countries.' (Dore 1997, pp. 94–5)

Such criticism was challenged, however, on the grounds that while a 'weak' version of the screening hypothesis is undoubtedly true, since employers may well use educational qualifications as a proxy for other characteristics when selecting their work force, there is no evidence to support the 'strong' version of the hypothesis, that education has no direct effect on productivity. The fact that employers continue to pay educated workers more than the uneducated throughout their working lives refutes this.

Even so, the idea of education as a screen or filter has been important in influencing theories of human capital. Blaug (1976, p. 850) in a review which he describes as a 'slightly jaundiced survey of the empirical status of human capital theory' predicts that:

> in time, the screening hypothesis will be seen to have marked a turning point in the 'human investment revolution in economic thought,' a turning point to a richer, still more comprehensive view of the sequential life cycle choices of individuals.

4. Concepts of Social and Cultural Capital

Just as economists now recognize that the concept of human capital must be enlarged to take account of the effect of education on individual attitudes and motivation, as well as knowledge and skills, and that the indirect benefits (externalities) of education may be even more important than the direct economic and monetary benefits, sociologists and scholars from other disciplines now emphasize the complementarities between human capital, social capital and other forms of investment. Bourdieu (1986) argued that 'capital can present itself in three fundamental guises: as *economic capital*, which is immediately and directly convertible into money and may be institutionalized in the form of property rights; as *cultural capital* which is convertible, on certain conditions, into economic capital and may be institutionalized in the form of

educational qualifications, and as *social capital*.' (Bourdieu 1986, p. 47). At about the same time Coleman (1988, pp. 14–15) argued that analysis of the purely economic effects of human capital is deficient and needs to take into account social structures and relationships in the family and community as well as effects of investment on the individual: 'norms, interpersonal trust, social networks, and social organization are important in the functioning not only of society but also of the economy.' (Coleman 1988, reprinted in Dasgupta and Serageldin 2000, p. 14).

More recently the concept of 'social capital' has attracted considerable attention, including a new study for the World Bank (Dasgupta and Serageldin 2000) that emphasizes the complementarities between human capital and social capital, and between concepts from sociology, anthropology, and political science as well as economics, in analysis of the effects of investment and social action. The overwhelming conclusion of this review is that educational aspects of human capital must also be included in such a 'multifaceted perspective.'

See also: Education and Economic Growth; Education and Employment; Education and Income Distribution; Education, Economics of; Educational Institutions and Society; Educational Productivity

Bibliography

Becker G S 1964 *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Princeton University Press, Princeton, NJ

Blaug M 1976 The empirical status of human capital theory: A slightly jaundiced survey. *Journal of Economic Literature* **14**: 827–55

Bourdieu P 1986 The forms of capital (translated by Nice R). In: Richardson J E (ed.) *Handbook of Theory of Research for the Sociology of Education*. Greenwood Press, New Haven, CT, pp. 241–58 [reprinted in Halsey A H, Lauder H, Brown P and Wells A S (eds.) 1997 *Education: Culture, Economy, Society*. Oxford University Press, Oxford, UK and New York, pp. 46–58]

Bowman M J 1966 The human investment revolution in economic thought. *Sociology of Education* **39**: 111–38

Colclough C, Lewin K 1993 *Educating All the Children*. Clarendon Press, Oxford, UK

Coleman J S 1988 Social capital in the creation of human capital. *American Journal of Sociology* **94** (suppl.) [reprinted in Dasgupta P, Serageldin I (eds.) 2000 *Social Capital: A Multifaceted Perspective*. The World Bank, Washington, DC, pp. 13–39]

Dasgupta P, Serageldin I (eds.) 2000 *Social Capital: A Multifaceted Perspective*. The World Bank, Washington, DC

Department of Education and Science (UK) 1988 *Top-up Loans for Students*. Her Majesty's Stationery Office, London

Dore R 1997 *The Diploma Disease: Education, Qualification and Development,* 2nd edn. Institute of Education, University of London, UK

King E M, Hill M A (eds.) 1993 *Women's Education in Developing Countries: Barriers, Benefits and Policies*. John Hopkins University Press, Baltimore, MD and London

Psacharopoulos G 1973 *Returns to Education: An International Comparison*. Elsevier, Amsterdam

Psacharopoulos G 1994 Returns to investment in education: A global update. *World Development* **22**: 1325–43

Psacharopoulos G, Woodhall M 1985 *Education for Development: An Analysis of Investment Choices*. Oxford University Press, Oxford, UK and New York

Schultz T W (ed.) 1961 Investment in human capital. *American Economic Review* **51**: 1–17

Smith A 1776 *The Wealth of Nations*

Summers L 1993 Foreword. In: King E M, Hill M A (eds.) *Women's Education in Developing Countries: Barriers, Benefits and Policies*. John Hopkins University Press, Baltimore, MD and London

Woodhall M 1973 Investment in education: A reappraisal of the concept of human capital. *International Review of Education* **XIX**: 9–29

World Bank 1999 *Knowledge for Development: World Development Report 1998/9*. Oxford University Press for the World Bank, Oxford, UK and New York

M. Woodhall

Human Cognition, Evolution of

Human cognition is the way it is, in large part, because it evolved the way it evolved. Unfortunately, for science, however, the key evolutionary processes took place long ago and cognition does not fossilize. This means that scientific investigations of human cognitive evolution must rely on a variety of more-or-less indirect sources of data. Among the most important of these are: (a) the cognition of nonhuman primates; (b) the fossils and material artifacts left behind by humans at various periods of their evolution; and (c) the cultural-cognitive processes at work in modern human populations.

Following common practice in the various branches of comparative biology and psychology, it may be assumed that if all the species of a given evolutionary taxon possess a certain physical trait or behavioral propensity, then the common ancestor for that taxon also possessed that trait or propensity. In the case of humans, the taxon of direct relevance is primates, with special attention to *Pan troglodytes* and *Pan paniscus* (referred to collectively as chimpanzees), because human beings shared a common ancestor with them a mere 6 million years ago. The cognitive processes common to chimpanzees and human beings thus represent the best evidence we have for the cognitive processes of the very first hominids.

For most of the day, every day, primates forage for food, quite often for patchy resources such as fruits and new leaves and buds (which only grow on some

plants at some times of the year), and sometimes for insects or other animal prey who can move freely on their own. In order to forage efficiently, therefore, primates must:

(a) remember 'what' is 'where' in their local environments, for example, which fruits are in which trees at what times;

(b) take creative detours and shortcuts to get to the food as quickly as possible;

(c) track the visible and sometimes invisible movements of prey and other objects;

(d) recognize and categorize different foods and objects on the basis of perceptual features;

(e) quantify relative amounts of foods on the basis of perceptual features; and

(f) use insight in problem solving, for example, in extracting food from substrates.

A variety of empirical studies—both observational in the wild and experimental in captivity—demonstrate that virtually all anthropoid primates have all of these cognitive skills (see Tomasello and Call 1997 for a review). Primates also must use sophisticated cognitive skills to deal effectively with their social environments. Among the most important of these (again, see Tomasello and Call 1997, for a review) are the abilities to:

(a) recognize individuals within the social group;

(b) form social relationships with group mates based on such things as kinship, friendship, and dominance rank;

(c) recognize the 'third-party social relationships' of other group mates based on such things as kinship, friendship, and dominance rank;

(d) predict the behavior of individuals based on such things as their emotional state and their direction of locomotion;

(e) use many types of social and communicative strategies to outcompete group mates for valued resources;

(f) cooperate with conspecifics in problem-solving situations and in forming social coalitions and alliances; and

(g) engage in various forms of social learning from conspecifics.

Indeed, the fact that the primate social environment presents such a wide array of cognitive challenges has led some theorists to posit that the truly unique features of primate cognition—involving such things as understanding and dealing with relational information—have come from their adaptations not to the physical world but to the social world (Humphrey 1976, Tomasello and Call 1997). Empirical work on primate cognition—coupled with the working assumptions of comparative biology—has thus established that by the time hominids set out on their own evolutionary pathway some 6 million years ago, they already possessed a plethora of cognitive skills, derived mainly from adaptations in the context of foraging for food and dealing effectively with group mates. In brief, they already conceptualized a sensory-motor world of permanent objects—arrayed in a representational space and categorized and quantified in some fundamental ways—and a social world of animate conspecifics with whom they could cooperate and compete in various complex ways.

However, if we compare contemporary chimpanzees and human beings there are still some cognitive skills that are unique to humans and that therefore must have arisen after the human–chimpanzee divergence 6 million years ago. Most important, in the physical world human beings, and only human beings, understand the causality of events; that is, they understand events in terms of the unobservable 'forces' that explain why they happen in the way they do (Visalberghi and Tomasello 1998). Understanding causality enables individuals to interact with their physical words in much more flexible ways, for example, by creating or blocking key antecedent events in order to produce a desired consequent. Analogously, in the social world human beings, and only human beings, understand the intentionality of behavior; that is, they understand behavior in terms of the unobservable mental 'forces' (thoughts, beliefs, desires) that explain why individuals behave the way they do (Visalberghi and Tomasello 1998). Understanding intentionality enables individuals to interact with conspecifics in much more flexible ways, for example, by attempting to manipulate their knowledge or motivational states so that they will act in desired ways. But perhaps even more importantly, understanding conspecifics as intentional agents constitutes the social–cognitive foundation for human culture, which gives human cognition its species-unique social-collective dimension (see Sect. 3).

1. Hominid Material Artifacts

For the first 4 million of its 6 million years of existence, the human lineage consisted exclusively of one or another species of australopithecine. Australopithecines were small in stature, bipedal, with ape-sized brains, and no stone tools; many modern paleoanthropologists thus refer to them as 'bipedal apes.' Although we know little of the behavior of these earliest hominids, there are no clear indications that they cognized the world in anything other than the typical great ape manner.

With the birth of the genus *Homo* some 2.5 to 2 million years ago, the human lineage began to show signs of some unique cognitive adaptations. Most important, humans began to make stone tools in ways that required at least some degree of planning and causal understanding of the way rocks could be modified and used instrumentally (Schick and Toth 1993). Thus, early in this period *Homo habilis* made Oldowan (pebble) tools, which required some understanding of how rocks break when struck in certain

ways, and which were often made in one location and then carried to another for subsequent use. Later, *Homo erectus* made Acheulean hand axes (bifaced and symmetrical), which showed a clear ability to hold in mind a definite end product during the process of manufacture and in addition an ability to understand some of the basic causal forces at work in stone tool making. During the early parts of this period (especially from 2 to 1.5 million years ago), the brains of hominids expanded in size by about one third, perhaps because of the cognitive demands of making and using these complex technological products.

A second major brain expansion about 500,000 to 200,000 years ago increased hominid brain size by a factor of one third once again, basically to its modern size. This took place in a set of hominids known collectively as archaic *Homo sapiens*, followed soon by modern *Homo sapiens*. These hominids invented some new techniques for making stone tools, initially the Levallois technique in which a large stone core is specially prepared so that smaller flakes of a special type may be systematically struck off, and later the technique of making blade tools (in much the same way) whose great length to width ratio made them much more multipurpose. Following this in many populations were specialized microlith tool industries, with all kinds of highly specialized tools for highly specialized purposes (each made in a specialized way). Interestingly, some of these flakes, blades, and micoliths were hafted onto wooden or bone handles, which made them even more efficient cutting tools. Of special importance from a cognitive point of view, manufacturing these hafted tools required their makers to understand, before the manufacturing process began, that different parts of the tool had different instrumental functions. These so-called composite tools may thus indicate the ability of archaic and early modern *Homo sapiens* to understand and plan for the different causal roles of the different components of a tool.

Another feature of hominid tool use suggests something of how individuals learned to make and use their tools. Both Oldowan and Acheulean types of tools showed great uniformity of use at particular historical periods across all groups of hominids who used them, no matter how far apart from one another they may have lived. In the case of the Oldowan tools of *Homo habilis* (and relatives), this uniformity was confined to Africa and lasted for hundreds of thousands of years. In the case of the Acheulean tools of *Homo erectus*, this uniformity extended all over Africa, Western Europe, and the near East and India, and lasted for over one million years. But with the rise of the more modern techniques of tool making and use, especially the blade and microlith technologies uniquely associated with modern humans, different populations of human beings began to make and use their tools in population specific ways. This new pattern of distribution among social groups suggests the possibility of the cultural transmission of tool making and using techniques in the different populations, in much the manner of contemporary human populations.

We cannot know much about the social lives of the earliest modern humans because we have so few fossils and artifacts. But if indeed truly cultural organization emerged only with modern humans, then it is reasonable to suppose that there was some change in social cognition that led to this new form of social organization. The main reason to believe this is that before nine months of age human infants are 'acultural' creatures in the sense that they are not yet able to appropriate the cultural artifacts and practices of those around them actively. Soon a major change in their social cognition occurs and they begin to understand other persons as intentional beings like themselves, with mental states like their own. This is indicated by the emergence a whole suite of new behaviors such as:

(a) following the gaze direction and pointing gestures of adults;

(b) engaging in shared attentional interactions with adults;

(c) imitating adult actions on external objects; and

(d) directing adult attention through pointing and other gestures, all of which indicate a new understanding that other persons have intentional relations to the world (Carpenter et al. 1998).

Because of this new understanding, infants begin to acquire in earnest the tools, linguistic symbols, and other cultural artifacts and activities of their social groups (Tomasello et al. 1993). (Note that children with severe autism never develop this social cognitive ability and they, in a sense, remain acultural individuals; Hobson 1993.) If we assume this same relationship between social cognition and cultural learning in human evolution, we can use the artifactual evidence for the emergence of human cultures to infer that a change in social cognition occurred with modern humans.

2. *Culture and Cognition*

With the advent of modern humans and their cultural organization, human cognition took on a radically different character. First, with the advent of skills of cultural learning and teaching (resulting from the social-cognitive ability to understand others as intentional and mental agents), human beings began to create artifacts and social practices that accumulated modifications over time. Thus, the so-called 'cultural explosion'—which took place in some, but not all, human populations at about 40,000 years ago—was characterized by many new technological inventions and very rapid changes and improvements in existing technologies. Tomasello et al. (1993) hypothesized that this kind of cumulative cultural evolution, in which the complexity of material and symbolic artifacts ratchets up very quickly over historical time, is

only possible with the human form of cultural learning in which the learner understands the intentions of another individual in the sense that she understands what the behavior is 'for.' The need for this kind of understanding is clearest in the case of tools and symbols since both of these artifacts clearly point to things beyond themselves: the tool to its instrumental function in modifying the physical world and the symbol to its communicative function in modifying other people's attention and behavior.

Once this kind of cultural learning was operative, human populations could, in effect, pool their cognitive resources in a way unique in the animal kingdom. Chimpanzees and other nonhuman primates benefit from observing the activities of their conspecifics, but because they do not understand the intentions of others, what they learn from these observations is limited (Tomasello 1996, see also Kummer and Goodall 1985). Human beings, on the other hand, are able to acquire relatively faithfully many of the tools, symbols, and social practices of others, so that inventions spread relatively rapidly and are preserved in the group effortlessly until some new and improved innovation comes along, which is then acquired faithfully, and perhaps modified further, by another generation. An important result of this ratcheting up of cultural inventions over historical time, is that human children are born into a world of artifacts and social practices that embody something resembling the accumulated wisdom of their entire social group throughout its entire cultural history (Donald 1991).

When children acquire the use of a material or symbolic artifact, or learn to engage in a conventional social practice in their culture, they internalize the intentional states of the adult inventors and users of the artifact or social practice (Vygotsky 1978). This internalization process is of special importance when children acquire linguistic symbols, because a human language embodies all of the myriad ways that the people of a social group have found it useful to categorize and construe their worlds for purposes of communicating with one another. Because the same entity may be construed in many different ways linguistically—this object is a dog, an animal, a pet, or a pest—in acquiring a natural language children are induced to take multiple simultaneous perspectives on entities in the world, with some of these perspectives being common to all languages and some of them being unique to particular languages. When internalized by the child, then, the process of linguistic communication creates in each individual a species-unique form of perspectivally based cognitive representation, commonly known as symbolic representation (Tomasello in press).

As a way of highlighting the cognitive importance of processes of cumulative cultural evolution over historical time and their interaction with processes of cultural learning and internalization over ontogenetic time, we may imagine a human child raised on a desert island in isolation from all other human beings, with a normal brain, body, and sense organs, but with no access to tools, other material artifacts, language, graphic symbols, writing, Arabic numerals, pictures, people who could teach her things, people whose behavior she could observe and imitate, or people with whom she could collaborate. It is likely that as an adult this child's cognitive skills would look very little like those of contemporary human beings (Tomasello in press). The point is thus that modern human beings have the cognitive skills they do because they inherit things from their progenitors both biologically and culturally (cf. the Dual Inheritance Model of Boyd and Richerson 1985, Durham 1991). They inherit biologically the ability to understand conspecifics as intentional agents—and the cultural learning processes that this enables—and these abilities then in turn open the way to the cultural line of inheritance and children's internalization of the ways of understanding the world embodied in their culture's accumulated artifacts and social practices.

3. *Conclusion*

Much of this account, in common with other accounts of human evolution, is necessarily speculative. But it is based on a solid body of research demonstrating precisely how human cognition is similar to and different from that of nonhuman primates, and on a relatively straightforward search for the two most important species-unique aspects of human cognition in the human artifactual record. First, the earliest humans some 2 million years ago began to demonstrate some new skills of causal understanding in the context of complex tool manufacture and use. Second, and even more important, the earliest modern humans some 150,000 years ago began to demonstrate some new skills of social cognition—understanding conspecifics as intentional and mental agents—that fundamentally changed the process of human cognitive evolution because they enabled uniquely powerful processes of cultural learning and cumulative cultural evolution. These cultural processes, in turn, transformed the ontogenetic niche in which human cognition developed ontogenetically from a social into a truly cultural niche including artifacts, symbols, and social practices with accumulated histories. The ability to pool cognitive resources in these ways may even have been the competitive advantage that modern humans had over other contemporaneous hominids. The ultimate outcome of these evolutionary and cultural processes in humans today is a form of cognition that is culturally and symbolically constituted in ways unprecedented in the animal kingdom.

See also: Animal Cognition; Brain, Evolution of; Cognitive Psychology: History; Cognitive Psychology:

Overview; Emotions, Evolution of; Evolution of Cognition: An Adaptationist Perspective; Intelligence, Evolution of; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Lifespan Development: Evolutionary Perspectives

Bibliography

Boyd R, Richerson P 1985 *Culture and the Evolutionary Process*. The University of Chicago Press, Chicago

Carpenter M, Nagell K, Tomasello M 1998 Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development* **255**

Donald M 1991 *Origins of the modern mind*. Harvard University Press, Cambridge, MA

Durham W 1991 *Coevolution: Genes, Culture, and Human Diversity*. Stanford University Press, Stanford, CA

Hobson P 1993 *Autism and the Development of Mind*. Erlbaum, Hillsdale, NJ

Humphrey N K 1976 The social function of intellect. In: Bateson P, Hinde R A (eds.) *Growing Points in Ethology*. Cambridge University Press, Cambridge, UK, pp. 303–21

Kummer H, Goodall J 1985 Conditions of innovative behaviour in primates. *Philosophical Transactions of the Royal Society of London Series B* **308**: 203–14

Schick K, Toth N 1993 *Making Silent Stones Speak*. Simon and Schuster, New York

Tomasello M 1996 Do apes ape? In: Galef J, Heyes C (eds.) *Social Learning in Animals: The Roots of Culture*. Academic Press, New York

Tomasello M in press *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA

Tomasello M, Call J 1997 *Primate cognition*. Oxford University Press, Oxford, UK

Tomasello M, Kruger A, Ratner H 1993 Cultural learning. *Behavioral and Brain Sciences* **16**: 495–552

Visalberghi E, Tomasello M 1998 Primates causal understanding in the physical and psychological domains. *Behavioural Processes* **42**: 189–203

Vygotsky L 1978 Mind in society. In: Cole M (ed.) *The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

M. Tomasello

Human Development and Health

Both 'health' and 'human development' are broad terms, covering a lot of conceptual and empirical ground. Both will be defined quite broadly here. Health will encompass the broad range of indicators of physical and mental illness, as well as disability and mortality. Human development is associated with two relevant bodies of scholarship (see Lerner 1998 for extensive discussions). One view, seen most often in the early history of growth-oriented child development, defines human development as the unfolding of characteristics and capacities as one ages. A second view, used more often in current-day developmental psychology and by social scientists, defines human development as the patterns of skills, attitudes, and behaviors that emerge over the life course. These definitions are not discordant; what distinguishes them most are the assumptions and theoretical contexts on which the definitions rest.

Except for the field of developmental psychopathology (Cicchetti and Cohen 1995), the dominant psychological view is a developmental one that focuses on the patterns typically observed across individuals who are not subjected to unusual traumas or disruptions. Generally it is assumed that developmental processes and outcomes are characteristic of the species; hence, variability is expected to be modest, although both personal experiences and the broader social and cultural contexts can affect the nature and timing of developmental phenomena to a limited extent. The strongest and most generalizable developmental patterns are observed in the early years of life when it is possible to predict with considerable accuracy developmental milestones such as walking, talking, and the capacity for abstract thinking. There has been less interest in and study of developmental changes during adulthood, which are fewer in number, less dramatic, and less consistent across individuals. In fact, it was the fields of lifespan psychology and psychological gerontology that made a major contribution to reorienting developmental psychology to a fully contextual view (see *Lifespan Development, Theory of*, Schaie 1996).

In line with a contextual view of development, social scientists expect substantial variation in the ways in which skills, attitudes, and behaviors are patterned across individuals and the life course. Although they do not discount the possibility of universal or nearly universal laws of development, life course scholars assume that social and cultural forces are the primary determinants of human behavior—and that those forces will exert strong influences on the timing and nature of even those behaviors that are characteristic of the species. Because of dramatic differences in social systems and cultural traditions, wide individual and group variations in patterns of behavior over time are expected. Even when confronted with patterns of behavior that are strongly associated with age, social scientists will tend to view those patterns as the result of strong cultural proscriptions and/or the influence of powerful social forces (Elder 1992).

1. The Link Between Health and Human Development

Scholars studying developmental processes first became interested in illness as a threat to normal

development. Serious illnesses and/or disabilities can constrain the individual's ability to master developmental tasks and/or the timing of mastering them (e.g., Cicchetti and Cohen 1995, Newacheck and Taylor 1992). Moreover, failure to master early developmental tasks can disrupt the achievement of later ones, resulting in a snowballing pattern of developmental disadvantage. Although early research focused largely on disruptions to highly specific developmental tasks (e.g., cognitive development, motor skills), recent research casts a wider net, including, for example, studies of how severe childhood illness can delay or disrupt the development of social skills, self-esteem, and academic achievement (e.g., Gortmaker et al. 1990).

Another research tradition examines the relationships between health and human development from the opposite point of view: how patterns of development influence subsequent health. In this tradition, successful accomplishment of developmental tasks (be they biologically driven or socially/culturally prescribed) is viewed as a form of resource accumulation (human and social capital) that affects health later in the life course. A broad body of research now documents the links between early achievements—including intelligence, educational achievement, self-esteem, feelings of self-efficacy, and the social skills required for successful social relationships—and health in middle and later life (e.g., Clausen 1993, Elder 1999).

Although these are distinct research traditions, both are part of a broader and deeper intellectual transformation of how health and illness are understood by both scientists and the larger public. From the middle of the nineteenth century until the middle of the twentieth century, the *medical model* dominated our understanding of the nature and causes of illness. The medical model views illness as biological dysfunction caused by the failure of the organism's biological defenses to prevent damage or disturbance. It is a complex model, recognizing that threats to the body's integrity can be generated from either external forces (injuries, viral and bacterial antagonists) or internal dysfunctions (congenital deficits, cellular dysregulation). But it is a purely biological model that ignores potentially health-relevant social, psychological, and cultural factors.

Since the 1950s a multidisciplinary model, sometimes called the *biopsychosocial model* of illness, has become the dominant perspective on illness among both scientists and the larger public. The major tenet of this model is that illness is a product of the complex and interacting effects of biological, psychological (including behavioral), social, and cultural factors. An important corollary of this model is that biological intervention alone typically cannot prevent, cure, or manage illness. Both the causes of and appropriate interventions for illness and premature mortality occur at multiple levels. Research on the reciprocal and complex links between human development and health has been one important illustration of the utility of the biopsychosocial model of illness.

2. *Current State of Knowledge*

2.1 *The Effects of Health on Development*

A broad body of research documents the fact that childhood illness, especially chronic illness, is associated with increased risk of developmental problems. This has been demonstrated for a variety of physical illnesses including epilepsy (Hermann et al. 1989), juvenile onset diabetes (Hoare 1984), and cystic fibrosis (Thompson et al. 1994), to name but a few. Problems and delays are observed most frequently for cognitive development and learning, emotional control and maturity, the ability to form successful attachments with peers, and behavioral problems at school and at home. Mental illness during childhood is associated with similar developmental problems (Rutter and Rutter 1993). Although the number of studies that examine the consequences of childhood chronic illness on long-term developmental outcomes is small, research to date suggests that developmental problems tend to multiply over time, supporting psychologists' predictions that early developmental disruptions bode poorly for mastery of later developmental tasks.

There also is strong evidence that childhood mental illness is associated with poor *adult* outcomes. Turnbull et al. (1990), for example, report that the onset of major psychiatric illness before the age of 16 is associated with lower levels of educational attainment, lower occupational attainment and income, and higher rates of separation and divorce. Family formation is also affected by childhood mental illness. Forthofer et al. (1996) report that childhood and adolescent mental illness is associated with earlier, but less stable marriages.

The effects of chronic childhood physical illness on adult achievements are less clear and appear to vary across illnesses. Children suffering from illnesses that can be managed effectively typically attain levels of education, occupation, and income similar to those of their healthy peers (Newacheck and Taylor 1992). In contrast, illnesses that result in frequent incapacity and/or threats to survival (such as cystic fibrosis) and those that involve the brain (such as epilepsy) are associated with poorer adult socioeconomic outcomes.

Although childhood physical and mental illnesses are associated with developmental problems in a number of domains, cognitive problems, especially learning disabilities, have the most severe consequences for both short- and long-term outcomes, as demonstrated in studies comparing children who did

and did not experience learning disabilities as a result of epilepsy (Hermann et al. 1989). Children with epilepsy who did not have learning disabilities resembled healthy children with regard to social competence, depression, and behavior problems at school and at home. In contrast, children who experienced learning disabilities as a result of epilepsy exhibited developmental problems in a wide range of areas. Relative to other developmental problems, learning disabilities also are most strongly related to low levels of socioeconomic achievement during adulthood.

There is also an intergenerational component to the effects of health on development during childhood and adolescence. A broad body of research consistently reports that children who live with a parent who suffers from chronic physical illness, chronic or intermittent mental illness, and/or disability are at increased risk of developmental problems or delays (e.g., Hirsch et al. 1985). Indeed, the effect sizes for parental illness are as great as, and sometimes greater than, those for childhood illness. Again, a variety of developmental problems are observed, including cognitive and learning problems, behavior problems at school and at home, general social competence, and the ability to make successful attachments to peers.

Given the strong evidence that chronic illness (of self or parent) places children and adolescents at risk of developmental problems, recent research appropriately focuses on the processes and mechanisms that account for these relationships. This is a highly complex issue because there are multiple pathways by which illness affects development. For children who are chronically ill, characteristics of the illness, its *sequellae*, and the nature and efficacy of medical management all play a role in helping to protect against or, alternatively, increasing the risk of developmental problems. As noted above, some illnesses have direct effects on developmental outcomes (e.g., when diseases cause cognitive problems or constrict the child's ability to form bonds with peers). Illness severity has an impact on development beyond the symptoms and impairments generated by the disease. For example, the greater the level of absence from school as a result of illness, the more severe the developmental problems (McCubbin 1988). Medical treatments also play a role in developmental outcomes. Some medications have side effects that place the child at risk of developmental obstacles (e.g., inability to concentrate); other treatments (e.g., chemotherapy) can impair functioning in multiple domains. Finally, the stigma associated with chronic illness (Hermann et al. 1989) can lead to rejection from peers and discomfort in public places (e.g., diabetic children require special diets and cannot participate in many 'normal' childhood activities).

Beyond the constraints of the illness and its treatment, the family plays a strong role in increasing or lessening the ill child's risk of developmental problems. Quality of parenting and the parents' methods of coping with a chronically ill child are especially important (McCubbin 1988). The entire dynamic of family life is changed by the presence of a chronically ill child. Parenting requires more time, special skills, and maintaining a delicate balance between 'normalizing' the child's life as much as possible, while also providing high levels of supervision and medical management. Demands change over time as the illness waxes or wanes and as the child ages. Providing care to a chronically ill child over long periods of time is draining and affects all areas of parent's lives (e.g., the marital relationship, relationships with other children, labor force participation, financial security) (Avison et al. 1991).

Family dynamics also appear to explain the conditions under which children experience developmental problems when a parent is chronically ill. Two processes seem to be especially important in this regard. First, if parental illness is accompanied by high levels of marital discord, the likelihood of developmental problems in the children is greatly increased (e.g., Mann and MacKenzie 1996). Second, the risk of developmental problems increases exponentially if the parent's illness is associated with being 'emotionally unavailable' to the child (e.g., Wootton et al. 1997). Indeed, developmental delays can be observed by the age of 2 when the parent is emotionally estranged or distant from the child (Radke-Yarrow et al. 1985).

3. *The Effects of Development on Health*

There also is impressive evidence that developmental outcomes during childhood and adolescence have strong effects on health in middle and late life. There are three primary pathways by which early developmental outcomes exert persisting effects on health: socioeconomic achievements, social integration, and social support.

Of all the social factors related to health, socioeconomic status (SES) is the most consistent and powerful (House et al. 1994, Link and Phelan 1995). SES typically is operationalized as a combination of income and education, and sometimes occupation as well. All three SES components are related to health and longevity, although they seem to operate in somewhat different ways. The relationship between income and health is multifaceted, for example, income is related to access to and quality of healthcare, and to goods and services that promote health (e.g., safe housing, adequate nutrition). Occupation is, of course, a primary determinant of income, but also affects health via prestige and greater autonomy and control in occupational settings (e.g., Link et al. 1993). Education typically exhibits the strongest relationships with health outcomes, in part via its effects on occupation and income, and more directly via better health habits and increased knowledge and use of

information about health and healthcare (Ross and Wu 1995). SES during adulthood is strongly related to early development. As noted earlier, school performance and educational attainment are major developmental outcomes during early life.

Social integration refers to the attachments individuals sustain with the larger society and are typically measured in terms of occupational, organizational, and community roles. A broad body of research, beginning with Durkheim's classic work on suicide, has documented strong links between social integration and health and longevity (House et al. 1988). Again, early developmental outcomes are linked directly to social integration during adulthood. Sustaining social integration requires social competence, reasonable intelligence, and the self-discipline required for adequate role performance—all of which are developmental outcomes that are established, in large part, during early life.

There is impressive evidence that health is related robustly to social relationships. The effects of social ties on health operate in a variety of ways. First, high-quality relationships with family and friends are directly related to better health and longevity (e.g., House et al. 1988). Second, close relationships are the foundation from which social support emerges during times of stress. Social support takes a variety of forms ranging from consolation and reassurance to the provision of information and tangible assistance. High levels of social support help to protect against the onset or exacerbation of illness and facilitate recovery from illness (de Leon et al. 1999, George et al. 1989). The ability to form appropriate attachments and the social skills needed to sustain social bonds are established early in life.

Thus, early development sets the stage for subsequent health in a variety of ways. Moreover, the benefits of successful development tend to increase over time, whereas the disadvantages that result from unsuccessful development also spiral over time. This pattern, known as cumulative advantage/disadvantage (Ross and Wu 1996), results in diverging levels of health across adulthood, until very late life when biological factors may act as 'levelers,' narrowing the differences between the advantaged and disadvantaged.

Health in middle and late life is affected by both proximal and distal factors. Unquestionably, proximal and contemporaneous factors are typically more powerful predictors of morbidity and mortality than are distal patterns. Nonetheless, life course research has added to the knowledge base in two ways. First, sometimes, early events and experiences have persisting demonstrable effects on health and longevity, even when more proximal factors are taken into account. Second, the indirect effects of early events and experiences often help to elucidate the selection processes that result in more proximal antecedents of morbidity and mortality.

4. *Methodological Challenges*

A variety of methodological issues complicate efforts to better understand the links between development and health; two especially important ones are considered here. First, adequate depiction of the complex and time-dependent relationships between development and health requires longitudinal data collected over a significant portion of the life span. Such studies are especially important for identifying the effects of development on health because it is not until middle age that a significant proportion of the population exhibits health problems. Few studies to date assess developmental and health outcomes over the appropriate age range (i.e., from childhood/adolescence until at least late middle age). Such studies are very expensive in both time and money, making a substantial increase in the future unlikely.

Selective mortality also poses a serious threat to our understanding of the links between development and health, both of which are related strongly to mortality. Extant research is based on survivors, which has two important implications. First, it is likely that research findings *underestimate* the effects of developmental outcomes on subsequent health because the most unhealthy are likely to have died. Second, selective mortality may account for some of the observed patterns among survivors. For example, the narrowing of SES differences in health during late life may be due to earlier mortality among low SES persons rather than a 'biological leveling' at the end of life (Bartley and Plewis 1997). There is no way to overcome this problem. However, given longitudinal data covering large segments of the life course, the effects of selective survival can be estimated. The possible effects of selective survival should be considered in the interpretation of all empirical findings.

5. *Future Directions*

Future research on the links between health and human development is likely to focus on two major topics. First, efforts will continue to trace the specific pathways and mechanisms by which development and health affect each other. Although extant research has made important contributions to this issue, substantial components of these processes remain unknown and replication of findings to date is especially important because of reliance on cross-sectional and short-term longitudinal research designs. Second, future research will address the extent to which and ways in which developmental and health problems in early life can be compensated for or reversed. Some contributions to this issue will result from naturalistic research that identifies compensatory factors that permit early disadvantage, in either development or health, to be overcome or minimized. Other contributions will result from research in which interventions designed to

prevent developmental or health problems or to compensate for them are evaluated in terms of feasibility, cost, and effectiveness. Given the complex web of factors associated with early developmental and health problems, the challenges of arresting or reversing trajectories of disadvantage will be great, but the benefits will be great as well.

See also: Adolescent Health and Health Behaviors; Adult Development, Psychology of; Chronic Illness, Psychosocial Coping with; Coping across the Lifespan; Coronary Heart Disease (CHD), Coping with; Developmental Psychology; Health Behavior: Psychosocial Theories; Health Behaviors; Health Behaviors, Assessment of; Human Development, Successful: Psychological Conceptions; Illness: Dyadic and Collective Coping; Lifespan Development, Theory of; Midlife Psychological Development; Well-being and Health: Proactive Coping

Bibliography

Avison W R, Noh S, Speechley K N 1991 Parents as caregivers: Caring for children with health problems. In: Albrecht G A, Levy J L (eds.) *Advances in Medical Sociology*. JAI Press, Greenwich, CT, Vol. 2, pp. 65–94

Bartley M, Plewis L 1997 Does health-selective mortality account for socioeconomic differences in health? Evidence from England and Wales, 1971–1991. *Journal of Health and Social Behavior* **38**: 376–86

Cicchetti D, Cohen D J (eds.) 1995 *Developmental Psychopathology*. Wiley, New York, 2 Vols

Clausen J A 1993 *American Lives*. University of California Press, Berkeley, CA

Costa P T, McCrae R R 1996 Mood and personality in adulthood. In: Magai C, McFadden S H (eds.) *Handbook of Emotion, Adult Development, and Aging*. Academic Press, San Diego, pp. 369–83

de Leon C F M, Glass T A, Beckett L A, Seeman T E, Evans D A, Berkman L F 1999 Social networks and disability transitions across eight intervals of yearly data in the New Haven EPESE. *Journal of Gerontology: Social Sciences* **54B**: S173–89

Elder G H 1992 Life course. In: Borgatta E F, Borgatta M L (eds.) *Encyclopedia of Sociology* pp. 1120–30

Elder G H 1999 *Children of the Great Depression*, 25th Anniversary edn. Westview Press, Boulder, CO

Forthofer M S, Kessler R C, Story A L, Gotlib I H 1996 The effects of psychiatric disorders on the probability and timing of first marriage. *Journal of Health and Social Behavior* **37**: 121–32

George L K, Blazer D G, Hughes D C, Fowler N 1989 Social support and the outcome of major depression. *British Journal of Psychiatry* **154**: 478–85

Gortmaker S L, Walker D K, Weitzman M, Sobal A M 1990 Chronic conditions, socioeconomic risks, and behavioral problems in children and adolescents. *Pediatrics* **85**: 267–76

Hermann B P, Whitman S, Dell J 1989 Correlates of behavior problems in children with epilepsy. In: Hermann B P, Seidenberg M (eds.) *Childhood Epilepsies: Neuro-psychological, Psychosocial, and Intervention Aspects*. Wiley, New York, pp. 143–57

Hirsch B J, Moos R H, Reischi T M 1985 Psychosocial adjustment of adolescent children of a depressed, arthritic, or normal parent. *Journal of Abnormal Psychology* **94**: 154–64

Hoare P 1984 Does illness foster dependency? A study of epileptic and diabetic children. *Developmental Medicine and Child Neurology* **26**: 14–19

House J S, Landis K, Umberson D 1988 Social relationships and health. *Science* **241**: 540–5

House J S, Lepowski J M, Kinney A M, Mero R P, Kessler R C, Herzog A R 1994 The social stratification of aging and health. *Journal of Health and Social Behavior* **35**: 213–34

Lerner R M (ed.) 1998 *Handbook of Child Psychology. Vol. 1: Theoretical Models of Human Development*. Wiley, New York

Link B, Lennon M C, Dohrenwend B 1993 Socioeconomic status and depression: The role of occupations involving direction, control, and planning. *American Journal of Sociology* **98**: 1351–87

Link B G, Phelan J 1995 Social conditions as fundamental causes of disease. *Journal of Health and Social Behavior* **extra issue**: 80–94

Mann B J, MacKenzie E P 1996 Pathways among marital functioning, parental behaviors, and child behavior problems in school-age boys. *Journal of Clinical Child Psychology* **25**: 183–91

McCubbin M A 1988 Family stress, resources, and family types: Chronic illness in children. *Family Relations* **37**: 203–9

Newacheck P, Taylor W R 1992 Childhood chronic illness: Prevalence, severity, and impact. *American Journal of Public Health* **82**: 364–71

Radke-Yarrow M, Cummings M, Kuczynski L, Chapman M 1985 Patterns of attachment in two- and three-year olds in normal families and families with parental depression. *Child Development* **56**: 884–93

Ross C E, Wu C 1995 The links between education and health. *American Sociological Review* **60**: 719–45

Ross C E, Wu C 1996 Education, age, and the cumulative advantage in health. *Journal of Health and Social Behavior* **37**: 104–20

Rutter M, Rutter M 1993 *Developing Minds: Challenge and Continuity Across the Life Span*. Basic Books, New York

Schaie K W 1996 Adult Intellectual Development: The Seattle Longitudinal Study. Cambridge University Press, New York

Thompson R J, Gustafson K E, George L K, Spock A 1994 Change over a 12-month period in the psychological adjustment of children and adolescents with cystic fibrosis. *Journal of Pediatric Psychology* **19**: 189–203

Turnbull J E, George L K, Landerman R, Swartz M S, Blazer D G 1990 Social outcomes related to age on onset among psychiatric disorders. *Journal of Consulting and Clinical Psychology* **58**: 832–9

Wootton J M, Frick P J, Shelton K K, Silverthorn P 1997 Ineffective parenting and childhood conduct problems: The moderating role of callous-unemotional traits. *Journal of Consulting and Clinical Psychology* **65**: 301–8

L. K. George

Human Development, Bioecological Theory of

The bioecological model, together with its corresponding research designs, is an evolving theoretical system for the scientific study of human development

over time. Within the bioecological theory, development is defined as 'the phenomenon of continuity and change in the biopsychological characteristics of human beings, both as individuals and as groups. The phenomenon extends over the life course, across successive generations, and through historical time, both past and future.' The term 'future' raises a question: How is it possible to investigate scientifically phenomena that have not yet taken place? The question, of course, is hardly new; indeed it pervades every field of scientific endeavor. However, we are the only species that has, over historical time, developed the capacity to engage successfully in scientific inquiry, and thereby, in many respects, has been able to change the nature of the world in which we live. As a result, within certain limits, we humans have altered the nature and course of our own development as a species (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998).

To place the bioecological theory of human development into a larger context, it is important to recognize that many of the general perspectives advanced and elaborated in this theory, are also parts of other related lines of theoretical and empirical inquiry into human development. Examples are lifespan psychology (Baltes et al. 1999), cultural psychology (Cole 1995), Magnusson's developmental theory of contextual-interactive holism (Magnusson and Stattin 1998), and, especially, the work of Robert Cairns, who through communications and publications extending over three decades, has played a major role in the evolution of the four defining properties of bioecological model: process-person-context-time. Cairns is of course best known as the founder and principal protagonist of developmental science. Examples of his books and articles that have been most relevant to the evolution of the ecological model are cited in the bibliography (Bergman et al. 2000, Cairns 1970, Cairns and Cairns 1994). The specific profile of the theory of Bronfenbrenner and his colleagues is its comprehensive interdisciplinary and integrative focus on the age periods of childhood and adolescence and its explicit interest in applications to policies and programs pertinent to enhancing youth and family development.

1. *Defining Properties of the Bioecological Model*

The term 'evolving' highlights the fact that the model, along with its corresponding research designs, has itself undergone a process of development over its own 'life course.' Another defining property of the bioecological model specifies that it deals with two closely-related but nevertheless fundamentally different developmental processes, each taking place over time. The first defines the phenomenon under investigation: namely, that of continuity and change in the biopsychological characteristics of human beings. The second focuses on the development of the scientific tools—the theoretical models and corresponding research designs required for assessing the continuity and change.

These two tasks cannot be carried out independently, for they are the joint product of emerging and converging ideas, based on both theoretical and empirical grounds—a process called 'developmental science in the discovery mode' (Bronfenbrenner and Evans 2000, pp. 999–1000). In the more familiar 'verification mode' the aim is to replicate previous findings in other settings to make sure that the findings still apply. By contrast, in the discovery mode, the aim is to fulfill two broader but interrelated objectives:

(a) Devising new alternative hypotheses and corresponding research designs that not only call existing results into question, but also stand a chance of yielding new, more differentiated, more precise, replicable research findings and, thereby, producing more valid scientific knowledge.

(b) Providing scientific bases for the design of effective social policies and programs that can counteract newly emerging developmentally disruptive influences. This has been an explicit objective of the bioecological model from its earliest beginnings.

A major challenge to today's bioecological model is to discover how such new working hypotheses and corresponding research designs can be developed for the future. One answer lies in the possibility that, despite historical change, some elements of the model, and their inter-relationships, may remain constant across both time and space. From this perspective, today's model has several distinctive defining properties that become the foundation for the rest. Some are of relatively recent origin; others date back to the model's earliest formal beginnings (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998, for reviews). Each is expressed here in the form of a Proposition.

1.1 *Proposition I*

An early critical element in the definition of the ecological model is experience. The term is used to indicate that the scientifically relevant features of any environment for human development include not only its objective properties but also the way in which these properties are subjectively experienced by the persons living in that environment. This equal emphasis on an 'experiential' as well as an 'objective' view springs neither from any antipathy to behaviorist concepts, nor from a predilection for existential philosophical foundations. It is dictated simply by a hard fact. Very few of the external influences significantly affecting human behavior and development can be described solely in terms of objective physical conditions and

events (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998).

A critical term in the foregoing formulation is the word 'solely.' In the bioecological model, both objective and subjective elements are posited as driving the course of human development; neither alone is presumed sufficient. Moreover, these elements do not always operate in the same direction. It is therefore important to understand the nature of each of these two dynamic forces, beginning on the phenomenological or experiential side. Both of the underlined terms are relevant because, while related to each other, they are typically applied to somewhat different spheres. The former is more often used in relation to cognitive development and pertains mainly to changes in how the environment is perceived at successive stages of the life course, beginning in early infancy and proceeding through childhood, adolescence, adulthood, and, ultimately, old age.

By contrast, 'experience' pertains more to the realm of feelings; for example, anticipations, forebodings, hopes, doubts, or personal beliefs. These too, emerge in early childhood, continue through life, and are characterized by both stability and change. They can relate to self or to others, and especially to family, friends, and other close associates. They can also apply to the activities in which one engages; for example, those that one most or least likes to do. But the most distinctive feature of such experiential qualities is that they are 'emotionally- and motivationally-loaded', encompassing both love and hate, joy and sorrow, curiosity and boredom, desire and revulsion, often with both polarities existing at the same time but usually in differing degrees. A significant body of research evidence indicates that such positive and negative subjective forces, evolving in the past, can also contribute in powerful ways to shaping the course of development in the future (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998).

But these are not the only powerful forces at work. There are others that are more objective in nature. This does not mean, however, that they are necessarily either more or less influential, mainly because the two sets of forces are interdependent and affect each other. Like their subjective counterparts, these more objective factors also rely for their assessment on corresponding theoretical models and associated research designs that have evolved over time. These more objective relationships are documented below in the form of two propositions (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998). The first specifies the theoretical model, and provides concrete examples; the second foreshadows the corresponding research designs for their assessment.

1.2 Proposition II

Over the life course, human development takes place through processes of progressively more complex reciprocal interaction between an active, evolving biopsychological human organism and the persons, objects, and symbols in its immediate external environment. To be effective, the interaction must occur on a fairly regular basis over extended periods of time. Such enduring forms of interaction in the immediate environment are referred to as proximal processes. Examples of such processes include feeding or comforting a baby; playing with a young child; child-child activities; group or solitary play; reading, learning new skills; athletic activities; problem solving; caring for others; making plans; performing complex tasks; and acquiring new knowledge and know-how.

For the younger generation, participation in such interactive processes over time generates the ability, motivation, knowledge, and skill to engage in such activities both with others and on one's own. For example, through progressively more complex interaction with their parents, children increasingly become agents of their own development, to be sure only in part.

In sum, proximal processes are posited as the primary engines of development. The next defining property speaks to the corresponding research designs.

1.3 Proposition III

The form, power, content, and direction of the proximal processes producing development vary systematically as a joint function of the characteristics of the developing person (including genetic inheritance); the environment—both immediate and more remote—in which the processes are taking place; the nature of the developmental outcomes under consideration; and the continuities and changes occurring in the environment over time, through the life course, and during the historical period in which the person has lived.

Propositions II and III are theoretically interdependent and subject to empirical test. An operational research design that permits their simultaneous investigation is referred to as a Process-Person-Context-Time model (PPCT).

In the corresponding research designs for the bioecological model, the element of time has special importance. To show that development has actually occurred, the research design must demonstrate, or at least make plausible, that the elements in the design, and their dynamic relationships to each other, have influenced the 'biopsychological characteristics of the developing person over an extended period of time.'

For example, a rich data archive generously made available by Small and Luster (1990) from their statewide studies of youth at risk in Wisconsin has made possible some reanalyses of working hypotheses derived from the newly-emerging formulations.

Figure 1 depicts the results from a reanalysis of the effects of parental monitoring on the academic

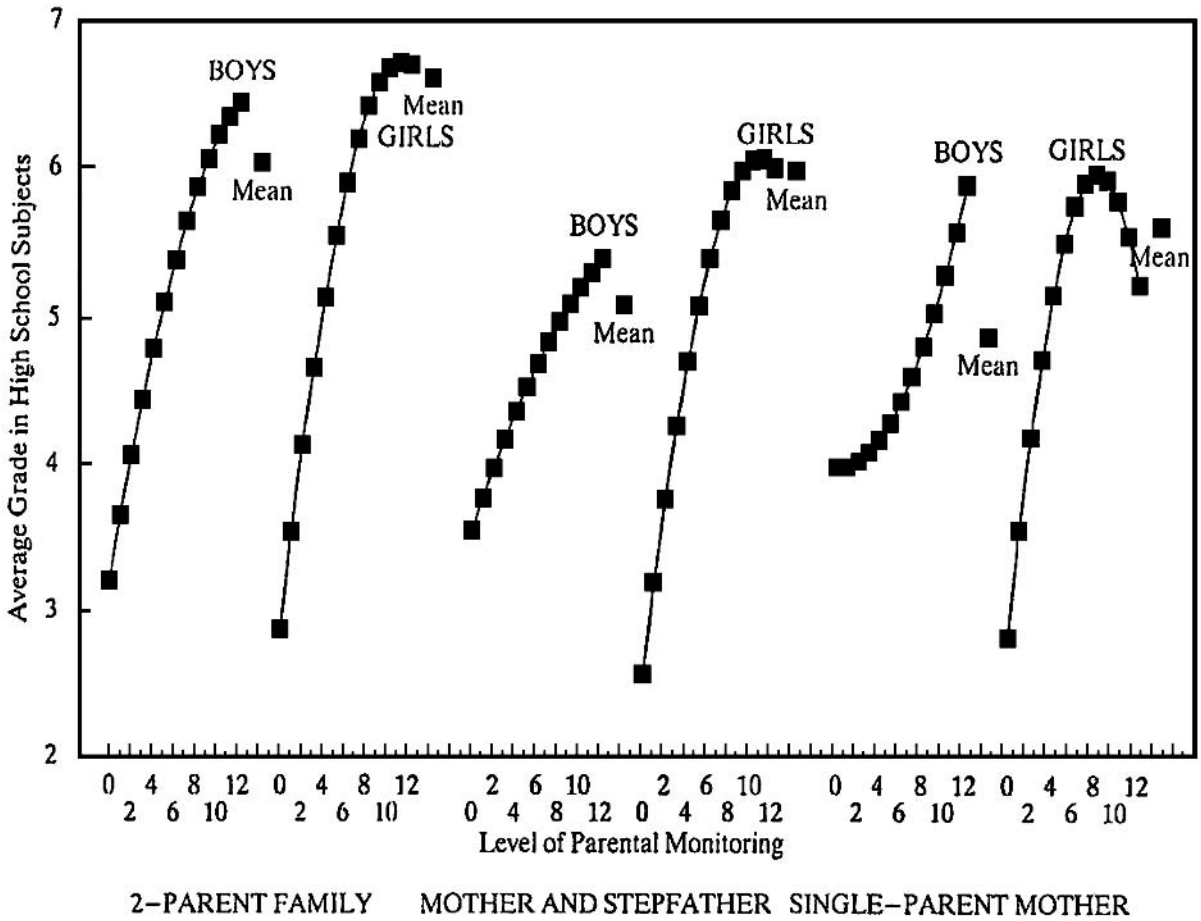

Figure 1
Parental monitoring and high school grades by gender: mothers with more than a high school education

achievement of high school students living in one of the three most common family structures found in the total sample of over 2500 cases. The students were between 14 and 16 years of age. The participants were also differentiated by gender, and stratified by two levels of mother's education, with completion of high school as the dividing line. The figure shows the results only for the high school graduates.

Parental monitoring was defined as 'the effort by parents to keep informed about and set limits on their children's activities outside the home.' Levels of parental monitoring, ranging from 0 to 12, are shown on the horizontal axis, and grade point average (GPA) appears on the vertical scale. The markers to the right of each curve record the average GPA for each of the six groups.

The data were fitted to a curve with a declining slope, rather than simply with a straight line, based on the following considerations derived from the bioecological model: Higher levels of academic performance require mastery of progressively more complex tasks, and hence are more difficult to achieve. As can be seen in Fig. 1, the relation between parental monitoring and school grades does indeed show a declining curvilinear trend. This effect, however, is far stronger for girls than for boys, particularly in the most advantaged ecological niche—families with two biological parents in which the mother had some education beyond high school.

Both of these results are consistent with two working hypotheses derived from the bioecological model. The first stipulates that proximal processes (in this instance, parental monitoring) are likely to have greater impact in two-parent families than in those in which the mother is a single-parent or the father is a stepparent. The second hypothesis posits a stronger and longer-lasting influence of the family on the development of females than of males.

In addition, a distinctive feature of the pattern for girls is a marked flattening of the curve at higher levels of monitoring, and, in the case of daughters of single-parent mothers, even becomes a turnaround. This finding suggests that under such circumstances, the demands on the girls may become so great that the existing proximal processes are not equal to the task, and as a result, bring fewer educational returns. Finally, an analysis (not shown) of data on students whose mothers had no more than a high-school education revealed a similar pattern, but the constructive influence of monitoring was appreciably weaker, and its greater benefit to girls was reduced. Nevertheless, daughters of mothers with less than a high-school education both in single parent and in stepfather families still had higher GPA scores than did sons. Moreover, within each level of mothers' education statistically significant differences by family structure were found in school achievement, with students growing up in two-parent families receiving the highest grades and those from single-parent families the lowest.

To turn to a related issue: although proximal processes function as the engines of development, the energy that drives them comes from deeper sources that take us back to the experiential world of Proposition I (Bronfenbrenner and Evans 2000, Bronfenbrenner et al. 1996, Bronfenbrenner and Morris 1998). Both the subjective and the objective forces exert an especially strong influence on development during the formative years (from early infancy to young adulthood). A substantial body of research over the past century indicates that, two or three decades ago, these forces lay mainly within the family, with parents acting as the principal caregivers and sources of emotional support for their children, and with other adult family members living in the home being next in line. To a lesser extent, other relatives, family friends, and neighbors also functioned in this role.

However, there has been a marked change in this pattern over the past three decades. Parents, and other adult family members as well, have been spending increasing amounts of time commuting to and working at full-time jobs (in which overtime is increasingly often required or expected).

The nature of this trend and its relevance for human development are conveyed in the Propositions that follow. (For brevity's sake, the term 'child' is used below to encompass the entire period from infancy through young adulthood.)

1.4 Proposition IV

In order to develop—intellectually, emotionally, socially, and morally—a child requires, for all of these, the same thing: participation in progressively more complex activities, on a regular basis over an extended period of time in the child's life, with one or more persons with whom the child develops a strong, mutual emotional attachment, and who are committed to the child's well-being and development, preferably for life (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998). The prerequisites stipulated in Proposition IV then lead to the developmental consequences described in the next Proposition.

1.5 Proposition V

The establishment of a strong mutual emotional attachment leads to internalization of the parent's activities and expressed feelings of affection. Such mutual ties, in turn, motivate the child's interest and engagement in related activities in the immediate physical, social, and—in due course—symbolic environment that invite exploration, manipulation, elaboration, and imagination. The next Proposition broadens the family circle.

1.6 Proposition VI

The establishment and maintenance of patterns of progressively more complex interaction and emotional attachment between parent and child depend to a substantial degree, on the availability and involvement of another adult, a third party, who assists, encourages, spells off, gives status to, and expresses admiration and affection for the person caring for and engaging in joint activity with the child. It also helps, but is not absolutely essential, that the third party be of the opposite sex from that of the other person caring for the child, because this is likely to expose and involve the child in a greater variety of developmentally-instigative activities and experiences (Bronfenbrenner et al. 1996). Where this is an attachment to two or more parent figures, each can serve as a third party to the other.

The research evidence for this Proposition came mainly by default. It was produced by demographic data documenting a rapid rise in the proportion of single-parent households. The trend began in the 1980s, and then continued at an even faster rate through most of the 1990s. The overwhelming majority of such homes were those in which the father was absent and the mother bore primary responsibility for the upbringing of the children.

A large number of investigations of developmental processes and outcomes in families of this kind have since been conducted across a range of cultural and social class groups. In general, the findings lead to two complementary conclusions:

(a) Even in families living in favorable socioeconomic circumstances, children of single-parent mothers, or of fathers, for whom no other person is

acting reliably in a 'third party' role are at greater risk for experiencing one or more of the following developmental problems such as: hyperactivity or withdrawal; lack of attentiveness; difficulty in deferring gratification; poor academic achievement; school misbehavior; and frequent absenteeism.

(b) At a more serious level, there is the so-called 'teenage syndrome' of behaviors that tend to be associated together: dropping out of school; involvement in socially-alienated or destructive peer groups; smoking, drinking, frequent sexual experience, adolescent pregnancy, a cynical attitude toward work, and—in the more extreme cases—drugs, suicide, vandalism, violence, and criminal acts. Most of these effects are more pronounced for boys than for girls (Bronfenbrenner et al. 1996).

Not all single-parent families, however, exhibited these disturbed relationships and their disruptive effects on children's development. Systematic studies of the exceptions have identified what may be described as a general 'immunizing' factor. For example, children of single parents were less likely to experience developmental problems especially in families in which the mother (or father) received strong support from other adults living in the home. Also helpful were nearby relatives, friends, neighbors, members of religious groups, and, when available, staff members of family support and child care programs. What mattered most was not only the attention given to the child—important as this was—but also the assistance provided to the single parent or by others serving in the supportive roles cited in Proposition VI. It would seem that, in the family dance, 'it takes three to tango.'

But dancing is not the whole story. By the 1980s, theory and research in the ecology of human development had documented an accelerating trend toward greater permissiveness in styles of child rearing in American families. At the same time, successive scientific investigations had revealed progressively greater developmental advantage for strategies that placed increased emphasis on parental discipline and demand. The interpretation that emerged from analyses of the available data suggested that widespread application of these research findings would serve as an effective response to the developmentally disruptive changes taking place in contemporary society.

At this point, it is important to mention two other bodies of research that contributed significantly to the development of bioecological theory and its corresponding research designs. The first is now of long standing. In 1974 the sociologist Glen H. Elder, Jr., in his classic volume Children of the Great Depression (1974, 1999), extended the concept of development beyond the formative years to encompass the entire life course.

The second addition has yet to be fully exploited. Bronfenbrenner and Ceci (1994), taking the bioecological model as their point of departure, suggested an empirically testable alternative for the established scientific paradigm used in behavior genetics. The proposed alternative model: (a) allows for nonadditive synergistic effects; (b) employs direct measures of the environment; and (c) proposes proximal processes as mechanisms of person–environment interaction through which genotypes for developmental competence are transformed into phenotypes. The model predicts that: (a) estimates of heritability (h^2) for developmental competence increase markedly with the magnitude of proximal processes; (b) heritability measures the proportion of variation in individual differences that are attributable only to 'actualized' genetic potential, with the degree of 'unactualized' potential remaining unknown; (c) actualized genetic potential (h^2) will vary with the quality of the environment, and will increase as that quality is improved (for example, through providing job opportunities, health services, and intervention programs in low-income neighborhoods).

The authors also suggested that high levels of such patterns of parental behavior as 'neglect, abuse, or domination' could serve as powerful mechanisms for actualizing genetic potentials for developmentally maladaptive behaviors that both disrupt proximal processes and produce developmental disarray.

1.6.1 The bioecological model in the discovery mode: future perspectives. This section is based on Propositions and working hypotheses derived from the bioecological model for which, as yet, there are few empirical data. The section begins with future prospects for addressing the second stated goal of the bioecological model: that of 'providing needed scientific bases for the design of effective social policies and programs that can counteract newly-emerging developmentally-disruptive influences.'

In accord with the latter objective, the section begins with an unorthodox proposal. By and large, thus far, theory and research on human development have been concerned with the influence of the older generation on the development of the younger. In the Proposition that follows, the direction is reversed. (It should be noted that the basic idea underlying this Proposition is not new, and is foreshadowed both in the theory of Vygotsky (1978), and in the contemporary 'action' theory of Brandtstädter (1998, 1999).)

1.7 Proposition VII

The psychological development of parents is powerfully influenced by the behavior and development of their children. This phenomenon occurs through the life course; is more evident during the formative years when most children are living at home in the care of their parents; and often becomes especially pronounced during adolescence, when the young begin to

strive for independence both as individuals and as members of peer-groups. Such behavior is particularly likely to occur amongst those adolescents or youth who have comparatively little contact with their parents or other caring adults earlier in life. Although many studies have focused on the development of such alienated young people, the impact of the latter's behavior on the subsequent development of their parents has yet to receive the systematic investigation hat it deserves. The converse of the foregoing Proposition—the influence of the successful transition of children through adolescence and young adulthood on the constructive development of their parents—has regrettably received even less scientific attention.

1.8 Proposition VIII

Over the life course, the process of attachment exhibits a turnaround. In the beginning, it is the children who are the beneficiaries of the parents' irrational commitment, whereas toward the end the roles are reversed. Then it is the elderly parents who receive the love and care of their now middle-aged children. If, however, there was no attachment at the beginning, there may be no attachment at the end.

In this regard, developmental science has yet to address a curious omission both with respect to theory and research design. Resort to search engines in psychology and related fields has thus far failed to identify any investigations of the influence of parent-child attachment on the future development of the parent in contrast to that of the child.

This is not quite the case, however, for the next proposition in the discovery mode. The theoretical model, the corresponding research design, as well as half of the necessary empirical data are already available. The only problem is to find or conduct a study that meets the requirements of this model.

1.9 Proposition IX

If an investigation conducted in the past meets the requirements of the bioecological model, including assessment of developmental outcomes 'over an extended period of time,' then replication of the study at a later point in time would reveal whether the processes under investigation were still valid or had been nullified or superseded by subsequent historical changes. When the latter occurs, the investigator is confronted with the challenge of proposing alternative working formulations for explaining the observed phenomena.

At the conclusion of this article, we move from the domain of theory and research design to the world of reality and action. In the bioecological model, these two worlds have never been far apart. Especially from 1970 to 2000, these two worlds have become ever closer to each other. At a more general level, the findings from both domains reveal what has been referred to as 'growing chaos' in the lives of children, youth, families, schools, the world of work, and the ever-greater commuting in between (Bronfenbrenner and Evans 2000, p. 1022). The most recent report of this phenomenon appears in an article by Bronfenbrenner and Evans (2000). It contains the following summary about the nature of 'chaos' and its developmental consequences.

'Chaos integrates the various elements involved, and foreshadows the role [of chaos] in the bioecological model in terms of what is called 'chaotic systems.' Such systems are characterized by frenetic activity, lack of structure, unpredictability in everyday activities, and high levels of ambient stimulation. Background stimulation is high, and there is a general lack of routinization and structure in daily life. The environment is also a major source of interruption of proximal processes in the form of residential noise, crowding, and classroom design' (p. 121).

At the turn of the century, we are left with a troubling question: From the perspective of the bioecological model, what is the prospect for the future development of our species? The answer to that question lies with the willingness of the USA and other economically developed countries to heed the emerging lessons of developmental science. At the moment, it is difficult to know what the answer will be. The future could go either way.

Given this alternative, then surely it becomes the responsibility of developmental science to communicate such knowledge as we possess, and to do so in words that can still find an echo. Here is a first draft.

In the USA it is now possible for a youth, female as well as male, to graduate from high school, or university, without ever caring for a baby; without ever looking after someone who was ill, old, or lonely; or without comforting or assisting another human being who really needed help. The developmental consequences of such a deprivation of human experience have not as yet been scientifically researched. But the possible social implications are obvious, for—sooner or later, and usually sooner—all of us suffer illness, loneliness, and the need for help, comfort, and companionship. No society can long sustain itself unless its members have learned the sensitivities, motivations, and skills involved in assisting and caring for other human beings.

See also: Adulthood: Developmental Tasks and Critical Life Events; Biopsychology and Health; Development: Social-anthropological Aspects; Developmental Sciences, History of; Genetics and Development; Human Development and Health; Human Evolutionary Genetics; Life Course in History; Life Course: Sociological Aspects; Lifespan

Development, Theory of; Lifespan Theories of Cognitive Development; Observational Methods in Human Development Research; Plasticity in Human Behavior across the Lifespan; Psychological Development: Ethological and Evolutionary Approaches; Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Baltes P B, Staudinger U M, Lindenberger U 1999 Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology*. **50**: 471–507

Bergman L R, Cairns R B, Nilsson L G, Nysted L (eds.) 2000 *Developmental Science and the Holistic Approach*. Lawrence Erlbaum Associates, Mahwah, NJ

Brandtstädter J 1998 Action perspectives on human development. In: Lerner R M (eds.) *Theoretical Models of Human Development. Handbook of Child Psychology*, 5th edn. Wiley, New York, Vol. 1, pp. 807–63

Brandtstädter J 1999 The self in action and development: Cultural, biosocial, and ontogenetic bases of intentional self-development. In: Brandtstädter J, Lerner R M (eds.) *Action and Self-Development: Theory and Research Through the Life Span*. Sage Publications, Thousand Oaks, CA, pp. 37–65

Bronfenbrenner U, Ceci S J 1994 Nature-nurture reconceptualized in developmental perspective: A bio-ecological model. *Psychological Review* **101**(4): 568–86

Bronfenbrenner U, Evans G W 2000 Developmental science in the 21st century: Emerging questions, theoretical models, research designs, and empirical findings. *Social Development* **9**(1): 115–25

Bronfenbrenner U, Morris P A 1998 The ecology of developmental processes. In: Lerner R M (ed.) *Handbook of Child Psychology Theory*, 5th edn. Wiley, New York, Vol. 1, pp. 993–1028

Bronfenbrenner U, McClelland P, Wethington E, Moen P, Ceci S J 1996 *The State of Americans: This Generation and the Next*. Free Press, New York, pp. 150–88

Cairns R B 1970 Towards a unified science of development. *Contemporary Psychology* **15**: 214–15

Cairns R B, Cairns B D 1994 *Lifelines and Risks: Pathways of Youth in our Time*. Cambridge University Press, Cambridge, UK

Cole M 1995 Culture and cognitive development: From cross-cultural research to creating systems of cultural mediation. *Culture and Psychology* **1**: 25–54

Elder G H, Jr 1974 *Children of the Great Depression*. University of Chicago Press, Chicago

Elder G H, Jr 1998 The life course and human development. In: Lerner R M (ed.) *Handbook of Child Psychology. Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 939–91

Elder G H 1999 *Children of the Great Depression* (25th Anniversary edn.). University of Chicago Press, Chicago

Magnusson D, Stattin H 1998 Person-context interaction theories. In: Lerner R M (ed.) *Handbook of Child Psychology. Theoretical models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 685–759

Small S, Luster T 1990 Youth at risk for parenthood. Paper presented at the *Creating Caring Community Conference*. Michigan State University, East Lansing, MI

Vygotsky L S 1978 *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

U. Bronfenbrenner

Human Development, Successful: Psychological Conceptions

From its outset human development is dependent on the presence of essential and beneficial conditions in the physical and social environment. Some of these conditions are fundamental and the same for all human beings (oxygen, water, food), while others are more complex and necessary to varying degrees (e.g., support of caregivers, language education, culturally specific competencies). Human beings also differ from one another in their psychological and physical needs: congenital and acquired vulnerability and deficiencies need special support and encouragement, some of them may even require specialist intervention or preventive treatment (e.g., specialized diets in the case of inherited metabolic disturbances such as phenylketonuria). At any event, there needs to be a minimum degree of fit between the individual and his or her environment before development can actually take place. Within these bounds, however, a very wide range of human development can unfold: An individual's predispositions and qualities encounter an immeasurably vast range of diverse environmental influences (socialization, critical and non-normative life events, historical influences, etc.). These influential factors operate throughout an individual's ontogenesis (Baltes et al. 1998). Human development never reaches completion, is multidirectional and seldom irreversible, while at the same time features a high degree of interindividual variability and intraindividual plasticity (Baltes 1987, Baltes and Baltes 1990). Moreover, there is now a broad consensus that human development does not take place uniformly in an irreversible sequence of different stages and phases, but is a process that individuals themselves actively influence in substantial ways (Brandtstädter 1998).

In view of this broad range of possibilities, the question of how and under what circumstances human development can succeed proves to be complex in several respects. The purpose of this chapter is to deal with two questions. The more important, of course, is the question of what makes development successful: which are the prerequisites of an 'art of life' (Staudinger 1999)? To answer this question, however, the term 'successful' needs to be clarified.

1. *What is 'Successful' Development? The Problem of Criteria*

Since the work of David Hume (1978), many have argued that while empirical sciences like psychology may, for reasons of principle, apply criteria which are value judgments, they cannot justify or legitimize the use of such criteria. According to this view, no matter how good science may be, it can only describe and explain the world, but it cannot judge it. Is the question of 'successful' development hence unanswerable from the start, at least from the viewpoint of an empirical science?

Actually, the relationship between empirical description and explanation on the one hand and value judgments or normative standards on the other is in fact more complex than that. Even if empirical facts do not constitute sufficient conditions for making evaluations, as a rule they are nevertheless necessary ones (Brandtstädter and Schneewind 1977). Any decision or action is predicated not only upon preferences and value judgments but also upon expectations which take account of empirical facts. For instance, the demand for a juvenile who commit crimes to be 'educated' by a punishment assumes that punishment actually serves to influence his future actions. This is an assumption that can, and indeed must, be empirically tested. In this respect science plays a key role in answering the question of successful development, even if the criterion for measuring 'success' has had to be determined on a prescientific basis.

Yet it is open to question whether assessments of human action and biographies really are as completely arbitrary as is suggested by a Humean perspective. Without wishing to become involved in the finer points of the philosophical debates surrounding the 'deducibility' of normative statements from descriptive ones, it can nevertheless be said that all positions purporting to offer proof of the inherent distinction between 'is' and 'ought' in fact tend to presuppose this 'gap' of principle rather than to prove its existence (Mavrodes 1968). In search for firm foundations for value judgments as well as for empirical descriptions, the last line of defense is the practice of simply not calling certain statements into question because the grounds for doubting them appear more dubious than the statements themselves.

So although we can forget any prospect of finding 'the' best biography *per se* in the foreseeable future, indisputably negative examples can easily be identified within the variations realized. For instance, there is no doubt that a person who persistently and repeatedly violates others for no reason, leads a less 'productive' life than one who shows reasonable consideration for others' interests and needs when taking his or her actions and decisions. Likewise, one can hardly describe an old age marred by illness and loneliness, causing distress particularly to the aged person but also to his or her relatives and friends, as 'successful' without arousing the suspicion that this is an artificial, academic discussion or simply cynicism. The details of examples such as these will be controversial, and here too particular constellations will lead to changes in our assessments. Nevertheless, such examples show that in many cases value judgments ('better'–'worse') are scarcely more doubtful than the 'brute' facts.

However, the search for universal criteria of successful development which can be applied to all people irrespective of their individual, cultural and historical circumstances, is likely to be fruitless, because it follows from the actional perspective of human development that people's 'success' depends to no small degree on what they themselves consider to be successful. Within the development-regulating normative rules and developmental tasks of a culture, which incidentally are themselves subject to cultural development processes, the individual organizes ways of shaping his or her own life, looking for and construing orientations which give meaning to his or her personal development (Brandtstädter 1998).

These considerations point toward a solution to the question of criteria. Beyond the fulfillment of the fundamental conditions essential for life, the prerequisites for successful development are clearly those meta-competencies and meta-resources which ensure that individuals can make use of opportunities and encouragement, can compensate for deficiencies and vulnerabilities, can choose and realize their own developmental goals, and thus can actively shape their development within the framework defined by immediate and potential possibilities (Brandtstädter and Schneewind 1977). This metaperspective implies a multicriteria approach on the one hand, and the integration of subjective and objective criteria of 'success' on the other (Baltes and Baltes 1990).

2. *What Makes Development Successful?*

If human development is basically the product of culture and, at the same time, the result of individual action, an essential prerequisite for successful development is the capacity to act, i.e., to choose developmental goals, to make decisions, to form intentions and to put them into practice, and also to overcome resistance and to revise strategies where necessary. Acting successfully entails first the precise competencies needed to carry out a particular action in practice. Additionally, an adequate representation of these competencies and potentials in the person's self-concept is a prerequisite for an appropriate and autonomous choice to be made from among the possible developmental options. Thus, processes of self-regulation perform an integrating, 'orchestrating' function for human development (Staudinger et al. 1995).

Self-regulating processes can essentially be classified into two modes (Brandtstädter and Greve 1994). In the assimilative mode, goals are maintained even

against resistance, and strategies are pursued tenaciously until they are achieved. Obviously, development can only succeed if people are sufficiently willing and capable; anyone who always gives up immediately will achieve neither essential tasks nor those of his or her choosing. In the accommodative mode, when faced with goals that are truly or are perceived to be unachievable, or with irreversible losses and deterioration, preferences and orientations are adapted in a way that avoids meaninglessness and depressive tendencies, and preserves subjective well-being, but at the same time brings new, achievable goals into view, which revive the capacity and motivation to act. These aspects are illustrated with examples from the research fields of juvenile delinquency and successful aging.

2.1 Autonomy and Social Integration: The Sample Case of Juvenile Delinquency

Although difficulties of coping with developmental tasks occur throughout the lifecourse, juvenile delinquency serves as a good example with which to illustrate such difficulties. Challenges such as physical changes and sexual maturation, the need to develop interpersonal skills (e.g., for relationships with the opposite sex), acquiring education and, in particular, forming a personal and social identity are not just the storms to be weathered during adolescence; these challenges are vitally bound up with laying the groundwork for successfully solving the tasks and meeting the demands of adult life (Crockett and Crouter 1995). In particular, social developmental tasks play a particularly prominent role during the years of youth (Petersen and Leffert 1995). Especially at a juvenile age, the dualism involved in social integration is patently evident: at one and the same time, social integration is an aim in itself and a prerequisite for successfully fulfilling the developmental task of preparing adulthood. However, both the necessity and the opportunity to change and challenge social contexts are characteristic features of adolescence too. Since the dividing line between creatively transgressing the social rules and a deviant or pathological disregard for social limits can be a fuzzy one, these dilemmas of individuation vs. integration and of autonomy vs. adaptation may serve as a background for the explanation of juvenile delinquency (Coie and Dodge 1998). Delinquent juveniles, in particular those whose deviance is 'adolescence-limited' (Moffitt 1993), have evidently not managed to cope successfully with these tasks (Petersen and Leffert 1995). Thus, personal resources of coping with discrepancies between the social demands on the one hand and the personal goals, desires and competencies on the other gain a key function in explaining juvenile delinquency and, as a consequence, successful juvenile development. In particular, the goodness of fit between the individual and his or her social environment thus becomes an important precondition and, at the same time, a central criterion for developmental success of adolescence.

2.2 Resources of Resilience: The Sample Case of Successful Aging

Going yet another step further, human development in general can be regarded as a hierarchical, interactive sequence of micro- and macroscopic coping processes. While current coping processes and resources are shaped by developmental conditions, coping reactions can also act as forces in creating future development. That leads on to the investigation of developmental constellations of individual and social resources that make an individual either resilient against or vulnerable to new strains as they arise. The discussion of the conditions under which people cope successfully with aging illuminates this aspect very well. Numerous studies have shown that people's ability to function and their sense of well-being are not greatly impaired as they grow old, if at all. Indicators of overall psychological status such as self-esteem, general sense of well-being, or being convinced of having control over one's own life are variables which, taken together, do not show any substantial covariance with age (e.g., Brandtstädter et al. 1993, Staudinger et al. 1995). As a consequence, the stereotype of aging being a process strongly characterized by deficiencies and loss has gradually gone out of favor as being too one-sided and incomplete, to be superseded by a more multifaceted view of 'productive' and 'successful' aging (Baltes and Baltes 1990). These findings are most surprising when set against the well-documented fact that, whether in terms of cognitive development, changing sensory functions or morbidity, aging does entail a real increase in the losses and impairment people suffer (Bond et al. 1995).

This apparent contradiction between the growing strains and losses on the one hand and the stability of the self and well-being on the other is resolved once the personal constellations of 'resilience' (Rutter 1990) are made the focus of the theoretical perspective (Baltes and Baltes 1990, Brandtstädter 1998, Staudinger et al. 1995). In addition to the model of developmental self-regulation already mentioned (Brandstädter et al. 1993), the main model providing important insights that help to explain resilience in old age is that of 'selective optimization with compensation' (Baltes and Baltes 1990). As losses occur, they can be partly cushioned or alleviated by way of selective improvement (training, expertise), and compensated for as people reorient their lives or change their strategies. Against the background of the actional conception of development, it is important to point out that resilience is not just a reactive quality, but is often brought to bear proactively, before the troubling event

actually occurs or the degenerative tendency sets in. For instance, the fact that people have less chance of becoming crime victims as they grow older is partially explained by an increasing tendency to behave more cautiously (Greve 1998).

3. *Conclusion: Courses and Conditions of Successful Development*

In terms of cognitive functions, social integration, and self-regulation, human development features a considerable amount of plasticity (Baltes et al. 1998), which not only makes up the core prerequisite for the dynamics of resilience and the coping processes discussed here, but indeed is also a general precondition for the very possibility of lifespan development. Self-regulation processes are both a central criterion for adaptation processes on the one hand and they substantially control them on the other.

Personal well-being and psychological health, and social integration can only be maintained and improved if the individual can be protected against unfavorable conditions by resilient resource constellations, if personal and social coping resources can help to cushion and compensate for the crises and losses that will inevitably occur, and if the individual capacity to act is largely maintained, whether by stabilizing the required conditions or by finding a new orientation toward achievable objectives. Actually, all of these aspects can be seen both as the products and the producers of successful development.

Development essentially consists in upholding and implementing the individual's capacity to adapt to the discrepancies in new situations and developmental tasks as he or she seeks (via assimilation or accommodation) to attain a solution to, or dissolution of, the discrepancies that are generating a crisis or creating a burden, if these are unavoidable in the long run. Thus, successful development implies a progressive adaptation which simultaneously maintains or indeed expands this plasticity and adaptivity. As far as we know today, this is possible—and usual—at least up to the latest years of very old age (Baltes 1997). That leads to the somewhat tautological-sounding statement that successful development essentially means ensuring that development, i.e., progressive adaptation, will always be possible and really will occur.

The emerging perspective of adaptivity and activity as the roots of successful human development suggests essential new research perspectives focusing on sources of resiliency and vulnerability at any stage of human development. Moreover, even if the preconditions and processes are known which provide successful development, the ways to improve the individual's capacity to acquire or to use them are still to be investigated. Thus, studies in applied developmental science are needed in order to discover techniques and strategies of individual and social improvement of both of these core conditions, with respect to special interventions, e.g., in the case of juvenile delinquency, as well as with respect to primary prevention of developmental crises, e.g., in the case of successful aging in very old age.

Bibliography

Baltes P B 1987 Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* **23**: 611–26

Baltes P B 1997 On the incomplete architecture of human ontogeny. *American Psychologist* **52**: 366–80

Baltes P B, Baltes M M 1990 Psychological perspectives on successful aging: The model of selective optimization with compensation. In: Baltes P B, Baltes M M (eds.) *Successful Aging. Perspectives from the Behavioral Sciences*. Cambridge University Press, New York

Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 1029–143

Bond L A, Cutler S J, Grams J (eds.) 1995 *Promoting Successful and Productive Aging*. Sage, Thousand Oaks, CA

Brandtstädter J 1998 Action perspectives on human development. In: Lerner R M (ed.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 807–63

Brandtstädter J, Greve W 1994 The aging self: Stabilizing and protective processes. *Developmental Review* **14**: 52–80

Brandtstädter J, Schneewind K A 1977 Optimal human development: Some implications for psychology. *Human Development* **20**: 48–64

Brandtstädter J, Wentura D, Greve W 1993 Adaptive resources of the aging self: Outlines of an emergent perspective. *International Journal of Behavioral Development* **16**: 323–49

Coie J D, Dodge K A 1998 Aggression and antisocial behavior. In: Eisenberg N (ed.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*. Wiley, New York, Vol. 3, pp. 779–862

Crockett L J, Crouter A C 1995 Pathways through adolescence: An overview. In: Crockett L J, Crouter A C (eds.) *Pathways Through Adolescence*. L Erlbaum Associates, Mahwah, NJ, pp. 1–12

Greve W 1998 Fear of crime among the elderly: Foresight, not fright. *International Review of Victimology* **5**: 277–309

Hume D 1978 [1739–40] *A Treatise of Human Nature*. Clarendon Press, Oxford, UK

Mavrodes G I 1968 On deriving the normative from the nonnormative. *Papers of the Michigan Academy of Science, Arts, and Letters* **53**: 353–65

Moffitt T E 1993 Adolescence—limited and life-course—persistent antisocial behavior: A developing taxonomy. *Psychological Review* **100**: 674–701

Petersen A C, Leffert N 1995 What is special about adolescence? In: Rutter M (ed.) *Psychosocial Disturbances in Young People*. Cambridge University Press, Cambridge, UK, pp. 3–36

Rutter M 1990 Psychosocial resilience and protective mechanisms. In: Rolf J, Masten A S, Cicchetti D, Nuechterlein K H, Weintraub S (eds.) *Risk and Protective Factors in the Development of Psychopathology*. Cambridge University Press, Cambridge, UK, pp. 181–214

Staudinger U M 1999 Social cognition and a psychological approach to an art of life. In: Blanchard-Fields F, Hess T (eds.) *Social Cognition and Aging*. Academic Press, New York, pp. 343–75

Staudinger U M, Marsiske M, Baltes P B 1995 Resilience and reserve capacity in later adulthood: Potentials and limits of development across the life span. In: Cicchetti D, Cohen D J (eds.) *Developmental Psychopathology*. Wiley, New York, Vol. 2, pp. 801–47

W. Greve

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Human Ecology

Human ecology addresses the relationships of humans to their environments. This demarcation concept unites an array of otherwise unconnected scholarly traditions. The traditions are distinguished by their different interpretations of ecology itself, by the different theory frames they employ for the study of human ecology, and by their respective objects of explanation or problematics. This article begins with the different interpretations as a preview for sorting and examining the different theory frames and their problematics.

1. *Interpretations of Ecology*

Interpretations of the concept and substance of ecology differ according to the unique vantage points provided by independent scholarly traditions.

1.1 *The Biological Interpretation*

Ecology was originally given a biological interpretation in 1866 when Ernst Haeckel defined the science in accord with the work of Darwin, who had attempted to explain evolution in terms of the adaptations of organisms to the biotic and abiotic features of their environments. A generation later another foundation concept appeared: since evolutionary studies indicated there was an ecology of living organisms before humans evolved, and humans themselves are living organisms, there must therefore be a human ecology. On the biological interpretation, humans are part and parcel of the natural world, and are subject to the same biophysical dynamics that govern the organization of other living communities. For most of the twentieth century, however, although gradually shifting in the last couple of decades, biologists took only a passing interest in the scientific study of the ecology of humans *per se*.

1.2 *The Metaphorical Interpretation*

Social scientists, by contrast, were quick to seize on the idea of a human ecology with independent investigations begun in the 1920s, in India, by Radhakamal Mukerjee, who followed the biological interpretation, and in the USA by Robert Park, whose students did not. Aware of theories of community succession that their biological colleagues were then developing for natural communities, in which society was used as a metaphor to understand natural organization, some early US sociologists came to use nature as a metaphor to understand social organization. Setting aside the rhetorical uses of 'ecology' as a synonym for 'contextual' or 'environmental,' two variants of metaphorical interpretation can now be found in social science. The first applies bioecological concepts to the study of human communities *per se*, reasoning by analogy. Here, institutional or organizational settings are taken as contexts or environments for humans within their groups, populations, or societies, which are supposed to be organized parallel to natural communities. This variant generally overlooks the fact that natural communities contain multiple species and humans are a single species, preferring to finesse this by elaborating biological species and evolutionary analogies for societal divisions of labor. The second variant, discussed in Sect. 2.3, treats ecological knowledge as a social construct. Left unaccounted for in the metaphorical interpretation are biological connections between human societies and other natural systems. The metaphorical interpretation is a blueprint for the study of the internal workings of society through the adaptations of its populations or groups by means of their social organizations and shared cultural knowledge (see *Biology's Influence on Sociology: Human Sociobiology*).

1.3 *The Ideological Interpretation*

With the reawakening of the modern environmental movement worldwide and its growth in the last decades of the twentieth century, an ideological interpretation of ecology has appeared. In this, the biological and metaphorical interpretations of ecology are freely interwoven and supplemented with value judgments gauged against standards supposedly congruent with ecological science. Some versions of the ideological interpretation use nature as an ideal to evaluate and reform social and political organization, economic practice, and personal conduct, while others try to frame environmental ethics so as to integrate ecological science with politics, economics, philosophy, education and various religions and cultures. The ideological interpretation is neither monolithic nor uniform, but its advocates generally fall to the left of ideological spectra. The common thread is that ecology is more than science or metaphor: It is a world view with ethical implications for personal, cultural,

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

and institutional change (see, for example, Evernden 1992, Oelshlaeger 1994, Orr 1992).

2. Theory Frames for Human Ecology

The multifarious research traditions of human ecology may be understood in terms of the different theory frames they have employed, which reflect their respective interpretations of ecology. Different theory frames address different problematics.

2.1 The Evolutionary Theory Frame

The evolutionary theory frame assumes that ecological forces act on genetic and cultural variation among individuals and populations so as to effect a selective survival and reproduction. Using this assumption, neo-Darwinian principles of natural selection are applied to understand how individuals adapt to their natural environments, how different social species reflect common principles of social organization in response to ecological contingencies, and how populations and cultures are affected by pathogenic disease.

Individual behavior and social organization are classic problematics in social science. The application of evolutionary theory attempts to provide a fuller explanation of these. One strand of theory develops a human evolutionary ecology patterned on the behavioral and evolutionary ecology of the biological sciences. In this, an adaptive design is assumed to be the product of the elemental behavior of individuals, such as foraging, territorial defense, time allocation and mate selection, all of which are presumed to satisfy economic cost–benefit principles (Smith and Winterhalder 1992). It has yet to be demonstrated whether neo-Darwinian evolutionary–economic models have application beyond the case of hunter-gatherers. Because organic traits (such as brain size, bipedalism, and vocal chords) underlie cultural capabilities, and culture is understood as a human adaptation, some evolutionary models assume the coevolution of biological and cultural traits in ecological context (Boyd and Richerson 1985, Durham 1991). Another strand of theory focuses on problematics of social organization among different species, with a view to discovering common principles but without presuming that the properties of social organization can be reduced to the summed behavior of individuals. In addition to comparisons with social primates, a research tradition of long standing, Machalek's (1999) transpecific sociology enables the comparison with humans of the demographic constraints, division of labor, and homeostatic mechanisms that appear in colonies of social insects, and of common strategies of environmental exploitation across social species (see *Cultural Evolution: Overview*; *Human Ecology*).

Problematics of epidemiology, ancient, and contemporary, fall within an evolutionary theory frame inasmuch as most scholars consider pathogens to be the principle agents of natural selection operating on humans since sedentary lifestyles were adopted 10,000–12,000 years ago. The most startling and horrible example of the human ecology of disease is the cultural and demographic extirpation of untold millions of native Americans in times of post-Columbian exploration, due to epidemic pathogens imported by Europeans and their African slaves, and from which the tribes had no inherited or acquired immunities. In modern times, medical authorities are again sounding alarms about the development and spread of epidemic pathogens, old and new, because normal genetic mutations eventually produce drug-resistant variants of old as well as new, and now dangerous, variants of formerly benign organisms. Since microbic pathogens are part of the living environment of all humans, human culture and society have to adapt to ever-changing conditions of pathogenic disease just as pathogens adapt to humans—an inherently ecological condition. The incidence of pathogenic diseases is known to be affected by agricultural practices, dietary customs, migration patterns, population densities, climate alteration, cultural hygiene practices, lack of potable water, sexual mores, modern transportation that transcends normal environmental barriers, and technological methods of intervention (see Ewald 1994, Kipple 1997).

Intervention methods typically disrupt population equilibria that might normally exist between pathogens and humans, thus preventing natural environmental controls from exerting themselves. Instead of stabilizing natural environments, technology implemented through the social organization of medicine and public health rather attempts to control pathogens so as to stabilize their relationship to humans. This method in itself, by intervening to change natural biota, provides a stimulus for microbic evolutionary adaptation. The general principle is that changed environments bring changed frequencies of microbic pathogens, human intentions notwithstanding. Some scholars (Mascie-Taylor 1993) have tended to conceive human ecology entirely in these epidemiological terms, however, this has not been predominant.

2.2 The Environmental Theory Frame

The environmental theory frame dominates modern research traditions of human ecology. It too presumes a biological interpretation of ecology, but in some cases substantially supplemented by considerations of social science targeted to problematics of subsistence and sustainability.

Anthropologists and geographers have a long tradition of investigating how human behavior, culture, traditions, settlements, built environments, spatial

location, territory, and humans' sense of place are affected by efforts to procure subsistence from natural—typically regional—environments. Research programs investigating cultural ecology, ecological anthropology, environmental geography, and environmental history consistently demonstrate how the development of culture and society vary with settlement patterns, technology and means of material resource acquisition, and how natural environments are modified by sociocultural variations in subsistence practices. A cultural materialism, which viewed culture and its symbolism as epiphenomenal, was implicit or explicit in many of these classic approaches, as was a dualistic construal of nature and culture. Modern approaches, especially political ecology, are different (see Biersack 1999, Scoones 1999). They focus on the interaction of spatial and temporal environmental dynamics with people situated in political, social, institutional, and cultural context, and emphasize the effects of human agency within the push–pull parameters of social, economic and political inequalities, on the one hand, and local or regional environmental transformations, on the other. Political ecology conceives human–environment interactions as mutual, thus linking human ecology to political economy and crossing from problematics of subsistence to those of sustainability (see *Feminist Political Ecology*).

No problematic has so dominated modern environmental discourse as the question of limits to growth, specifically, whether industrialism is sustainable into the indefinite future as it is adopted round the world. Biologists began the debate in the 1960s by importing from population ecology concepts of carrying capacity limits, and they averred that cascading world population growth augured for environmental catastrophe. Some social scientists amplified this neo-Malthusian argument, noting that industrialism is not a population but a way of life, one which produces too much consumption not just too many consumers.

Attentive to anthropogenic environmental degradation (aquifer depletion, biodiversity loss, deforestation, ozone depletion, global warming, etc.), the new social science tradition of examining environmental interfaces with human production and reproduction is now secure with the appearance of numerous textbooks and treatises on environmental sociology, environmental psychology, and ecological economics (see Costanza et al. 1997, Harper 1996). These specialties tend to conceive human–environment interactions within the existing theory traditions of their established disciplines. Recently, ominous neo-Malthusian analyses have been challenged by more optimistic theories of ecological modernization (Spaargaren and Mol 1992).

Theory aside, problematics of sustainability are routinely encountered as matters of practice in the concerted effort to promote sustainable development in economically less secure nations of the world. Specifically promoted by the United Nations, sustainable development is taken to mean eliminating world poverty while not compromising local environmental stabilities (Brundtland Commission 1987). In fact, sustainable development as practiced has promoted the transfer of industrial technology to promote economic growth. Where the ideal has succeeded, and by consensus it has not much, in effect sustainable development programs have promoted further industrialization—thus rekindling the question of the long term sustainability of human–environment interactions as local, regional, and global environmental degradation compounds, interactively cascading from industrialism's increased energy capture and waste disposal. The concern over environmental degradation directs the problematic of sustainability, which currently is targeted more towards local than toward global applications (see *Human–Environment Relationships*).

2.3 *The Analogic–Symbolic Theory Frame*

The analogic–symbolic theory frame works with a metaphorical interpretation of ecology, with different variants selectively emphasizing analogies or symbolic constructs as central to theory development. Generally speaking, the theory frame takes biological ecology just seriously enough to borrow some of its content (sometimes merely its words) so that nature can serve as a metaphor to understand society, thus advancing social science. In the classic sociological tradition, 'environment' usually had no biological referent, and neither does 'community' or 'ecosystem,' which are favorite analogs in modern sociological human ecology. As noted earlier, US sociology provided the first sustained research tradition of human ecology, something of an urban geography, concerned as it was with spatial relations in cities, that beginning in mid-twentieth century the influential Amos Hawley gradually transformed into an ecological demography. In this, territorially circumscribed human populations are characterized by technologically based organizational adaptations in the sustenance producing activities of populations within their social, political, and economic environments. These orienting concepts have not changed in more than a generation (see Micklin and Posten 1998). They reflect a social and demographic, but not a biological, materialism in the conception of purely sociological problematics. Direct descendants, such as organizational ecology, and distant cousins, such as family ecology, likewise depend on theoretical analogies to biological phenomena in order to relate social entities to social environments (see *Population Ecology*).

A wholly independent and more recent set of problematics relating society and environment applies a social constructionist interpretation to environmental social science, counterweighting its biological interpretation. In this, natural environments become

material for the analysis of symbolic meanings and understandings by which social groups perceive and define environmental events as real and respond to them collectively. One-to-one theory analogs with bioecology are not sought, but the metaphorical interpretation of ecology runs deep because symbolized realities are accorded primacy over biophysical realities. Natural environments are taken to have significance only as people socially construct, perceive, define, and negotiate them as real, and make contestable claims about them. The problematics address social responses to natural environments insofar as environmental issues are defined through shared meanings (Hannigan 1995).

Theory variants that are exclusively analogic or symbolic usually seek to separate not integrate human ecology and bioecology, so as to strengthen social science *per se*. Implicitly, sometimes explicitly, many variants reflect the species prejudice that nonhuman organisms represent 'lower' forms of life and the cultural prejudice that Western industrialism represents the 'highest' form of social progress. These two prejudices are explicitly criticized in many variants of the interactive theory frame.

2.4 *The Interactive Theory Frame*

The interactive theory frame for human ecology is the newest, most diverse and most numerous in its scholarly problematics. Its growing traditions may only be listed here. They include assorted environmental ethics, such as deep ecology, the land ethic, and animal rights; assorted spiritual ecologies, including native US land wisdom, Christian-based ecological process theology, and strains of Buddhist, Taoist, and Jainist thought; some applications of complexity theory for evolving interactive systems; Gaian ecology; politically grounded approaches, such as social ecology and green Marxism; ecofeminism; and more (see, for example, Bookchin 1995, Benton 1996, Callicott 1994, Tucker and Grim 1993, Warren 1996).

All of these assume some variant of the ideological interpretation of ecology, reflecting a wish to integrate ecological knowledge with selected priorities in human affairs. Some are based in science, while others are based in philosophy, politics, religion, or tradition, but almost none maintains a hard and fast distinction between objective and subjective knowledge. Almost all take a holistic rather than reductionistic approach to their problematics, in which the traditional boundaries of scientific disciplines, and even the classic boundaries between science and the humanities, are regarded as anachronistic. Most reject any nature–culture dualism, and many disavow anthropocentrism. Multiple distinct problematics are defined, depending on the point of view. The overriding theme construes human ecology as knowledge of the complex interactive connections that integrate human activities with environmental contexts. The overall result is an eclectic combination of biological, environmental, moral, and even esthetic, approaches to human ecology (see *Nature, Concepts of: Environmental and Ecological*; *Ecology, Deep*; *Feminist Political Ecology*).

3. *Conclusion*

Human ecological scholarship on the biological, environmental, and ideological interpretations, with their derivative problematics, is now being increasingly writ into the curricula of educational institutions. Calls for 'green education' are being implemented internationally, programs in environmental science that include a human component are expanding, and one US undergraduate institution of higher learning, the College of the Atlantic, devotes its curriculum entirely to human ecology. While still stymied by the traditional boundaries of academic disciplines, and sometimes by parochial scholarship, human ecology—the subject, not the name—is experiencing dynamic growth, as befits the coming century of the environment.

See also: Darwinism: Social; Ecological Economics; Ecological Imperialism; Ecology: Organizations; Ecology, Political; Environmental Challenges in Organizations; Sustainability Transition: Human–Environment Relationship; Sustainable Development

Bibliography

Benton T (ed.) 1996 *The Greening of Marxism*. Guilford Press, New York

Biersack A 1999 Introduction: from the 'new ecology' to the new ecologies. *American Anthropologist* **101**: 5–18

Bookchin M 1995 *The Philosophy of Social Ecology*. Black Rose Books, Montreal, Canada

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Brundtland Commission (World Commission on Environment and Development) 1987 *Our Common Future*. Oxford University Press, Oxford, UK

Callicott J B 1994 *Earth's Insights*. University of California Press, Berkeley, CA

Costanza R, Cumberland J, Daly H, Goodland R, Norgaard R 1997 *An Introduction to Ecological Economics*. St. Lucie Press, Boca Raton, FL

Durham W H 1991 *Coevolution: Genes, Culture, and Diversity*. Stanford University Press, Stanford, CA

Evernden N 1992 *The Social Creation of Nature*. Johns Hopkins University Press, Baltimore, MD

Ewald P W 1994 *Evolution of Infectious Disease*. Oxford University Press, Oxford, UK

Hannigan J A 1995 *Environmental Sociology: A Social Constructionist Perspective*. Routledge, London

Harper C L 1996 *Environment and Society: Human Perspectives on Environmental Issues*. Prentice-Hall, Upper Saddle River, NJ

Kipple K F 1997 *Plague, Pox, and Pestilence*. Weidenfeld and Nicholson, London
Machalek R 1999 Elementary social facts: emergence in non-human societies. In: Freese L *Advances in Human Ecology*. JAI Press, Greenwich, CT, Vol. 8, pp. 33–64
Mascie-Taylor C G N (ed.) 1993 *The Anthropology of Disease*. Oxford University Press, Oxford, UK
Micklin M, Poston D L Jr. (eds.) 1998 *Continuities in Sociological Human Ecology*. Plenum, New York
Oelshlaeger M 1994 *Caring for Creation*. Yale University Press, New Haven, CT
Orr D W 1992 *Ecological Literacy: Education and the Transition to a Postmodern World*. State University of New York Press, Albany, NY
Scoones I 1999 New ecology and the social sciences: what prospects for a fruitful engagement? *Annual Review of Anthropology* **28**: 479–507
Smith E A, Winterhalder B (eds.) 1992 *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, New York
Spaargaren G, Mol A P J 1992 Sociology, environment, and modernity: ecological modernization as a theory of social change. *Society and Natural Resources* **5**: 323–44
Tucker M E, Grim J A (eds.) 1993 *Worldviews and Ecology*. Bucknell University Press, Lewisburg, PA
Warren K J (ed.) 1996 *Ecological Feminist Philosophies*. Indiana University Press, Bloomington, IN

L. Freese

Human Ecology: Insights on Demographic Behavior

1. Introduction

Human ecology is the 'the study of the form and the development of the community in human populations' (Hawley 1950, p. 68, Hawley 1971) for which the unit of analysis 'is not the individual but the aggregate which is either organized or in the process of being organized' (Hawley 1950, p. 67) and where the concept of 'community' is defined as a 'territorially based system of functional interdependencies that results from the collective adaptation of a population to its environment' (Berry and Kasarda 1977, p. 12). Human ecology proposes that populations adapt to social, economic, political, and biophysical environments through organizational innovations, and that systems grow 'to the maximum size and complexity afforded by the technology for transportation and communication' in reducing the time cost of maintaining the accessibility and exchange required by interdependent populations (Hawley 1986, p. 7). It is important to recognize that it is this theoretical framework, not the mere fact that aggregate data are used, that places research in the human ecological tradition.

2. Contemporary Human Ecology

The intellectual lineage of human ecology traces from the emergence of plant and animal ecology in the late nineteenth and early twentieth centuries and its application to human populations in the work of Park and Burgess (1921), McKenzie (1933), and other scholars at the University of Chicago. The approach of these scholars, while innovative, was ultimately limited in scope. Perhaps most limiting was the initial insistence that cultural variables be excluded from the purview of human ecology in favor of a focus on spatial patterns resulting from a process of unplanned competition. By contrast, the insights into demographic phenomena described herein depend almost entirely on *contemporary human ecology* (hereafter, *human ecology*) which denotes the body of theory and research that began with the publication in 1950 of Amos H. Hawley's *Human Ecology*, a landmark volume which led to a major reorientation of the field. Hawley's work dramatically shifted the focus from the description of physical features of the city and mapping of social phenomena to 'the ways in which human populations organize in order to maintain themselves in given environments, thus relegating spatial analysis to a minor though still useful position in the discipline' (Hawley 1986, p. 3). It is not spatial arrangements themselves, but rather the social relations often reflected in spatial patterns and processes, that are of interest to human ecologists (Frisbie and Kasarda 1988 for further discussion).

Albeit something of an oversimplification, the human ecological paradigm can be thought of as encompassing four broad categories of variables: population, organization, environment, and technology—codified by Duncan (1959, 1964) as the ecological complex bearing the acronym POET. It is crucial to recognize that, in human ecology, population is defined generically to mean 'a number of units sharing one or more given characteristics…(which acquires) operational significance only when clothed with a set of institutions and located in a space-time context' (Hawley 1998, pp. 11–12). At least since Hawley's redirection of the field (1950), the primary unit of analysis for human ecology has been the organized population, not the individual units comprising populations. In addition, human ecology places a much heavier emphasis on analysis of the social and economic environment, as compared to the biophysical environment (Hawley 1998). This distinguishes (but does not isolate) human ecological concerns from those which characterize bioecology and the environmental movement.

The term 'demographic behavior' is used broadly here to refer, not only to the basic processes of fertility, mortality, and migration, but also to *social demographic* phenomena, with social demography defined as 'the analysis of how general social and cultural factors are related to population structure and pro-

cess' (Ford and De Jong 1970, p. 4). In this context, an impressive range of human ecological insights into a large number of demographic issues might be identified. This article primarily emphasizes contributions to the study of the systemic attributes and processes involved in population growth and societal development, the distribution and redistribution of population, and intra- and interurban morphology. While this selectivity is logical in light of human ecology's emphasis on systems, it warrants mention that the well-known contributions of human ecology to research on social networks and residential segregation are given little attention inasmuch as these are the subjects of other articles included in this volume (see *Social Networks and Fertility* and *Residential Concentration/Segregation*, *Demographic Effects of*, respectively). Further, while much of the present discussion concentrates on research conducted in the US, human ecology has been drawn upon rather extensively in cross-national research (Frisbie and Al-Khalifah 1998).

3. *Human Ecology's General Theoretical Contribution to Macrolevel Demography*

It is clear that human ecology has contributed substantially to the preservation and advance of macrolevel theory in the social and behavioral sciences, generally, and in demography, in particular (Poston and Frisbie 1998). However, in the mid-twentieth century, two events combined to diminish the interest of demographers and other social scientists in macrolevel analysis. The first was the publication of work dealing with what came to be known (erroneously) as the 'ecological fallacy.' Robinson (1950) argued that results of research in which aggregates are the units of analysis cannot be used to infer relationships among individuals and that a large proportion of research that employs aggregate data does so only because information on individuals is not available. The second was actually a sequence of events involving the emergence of modern techniques of sampling and survey analysis which greatly enhanced the opportunity for rigorous microlevel studies (Linz 1969). For a time, demography listed far toward reliance on individual-level research, even in the face of warnings by renowned scholars that the discipline was ill-served by a fixation at the microlevel (e.g., Ryder 1980).

A more balanced view prevails, partially due to the efforts of 'ecological demographers' both early on (Schnore 1961) and later (Namboodiri 1988). As it turned out, Robinson's thesis was only partially correct. The problem identified by Robinson is not 'ecological' in nature, but rather an issue that arises whenever cross-level generalizations of any kind are contemplated. Moreover, many of the central tenets of human ecology are inherently macrolevel constructs devoid of any counterpart at the individual level. For example, the concept of the division of labor makes sense only when applied to groups. Analogous reasoning applies to many demographic variables. Individuals are born, they die, and in between they may migrate, but only groups, not individuals, have birth, mortality or migration *rates*.

A signal contribution of human ecology has thus been the development of a theoretical framework for a wide range of demographic studies of organized aggregates. Examples include investigations of the relationship between commodity exchange and position of cities in the US metropolitan hierarchy (Eberstein and Frisbie 1982), as well as in the worldwide system of cities (Meyer 1986), the effects on the volume and rate of migration, of functional niche and extent of the division of labor (Murdock et al. 1992, Poston and Mao 1996), changes in the social stratification of suburbs (Guest 1978, Krivo and Frisbie 1982), segregation of racial and ethnic populations (Massey and Denton 1989), and the 'birth' and 'death' rates of formal organizations (Carroll 1984, Hannan and Freeman 1977). Such work, in turn, offers a substantive grounding for mutlilevel analyses, in which, e.g., the behavior of individuals is partially explained in terms of the social or ecological context in which they exist.

4. *Specific Insights on Demographic Phenomena*

4.1 *Population Growth*

Perhaps no demographic issue has seized and held the attention of scholars, policy makers, the media, and the general public as that of rapid population growth. Interest in the 'population explosion' intensified *circa* 1970 in the wake of publications which predicted exhaustion of resources and irremediable environmental pollution if population and economic growth continued unchecked (Erlich 1968, Meadows et al. 1974). A wide range of less apocalyptic, but still highly disturbing, publications focused attention on less developed countries (LDCs) and the actual or anticipated deleterious consequences of rapid population growth for development and increases in the standard of living in these nations—effects epitomized by the piling up of excess rural populations in peripheral slums of huge cities (Davis 1975).

Early on, some scholars challenged the view that continued rapid population growth spelled worldwide collapse (see Davis 1990 for a summary of the position of 'alarmists' and 'skeptics'). Regardless of whether one tends toward pessimism or optimism, it is evident that demography has to come to grips with the Malthusian contention that population growth will inevitably be limited by exhaustion of natural and/or nonrenewable resources. Before one can meaningfully discuss its utility for better understanding population

growth, it is necessary to recognize that, in human ecological theory, population is regarded as a necessary, but not sufficient, condition. An important analytical implication is that population

> cannot operate independent of organizational circumstances. In and of itself it cannot cause war, resource exhaustion, or environmental pollution, as Malthusians have argued. Such outcomes are explained as due to maladaptations or malfunctioning of organization (Hawley 1986, p. 26).

Human ecologists have also called attention to certain errant assumptions that are perhaps more apt to be found in popular or 'vulgar' interpretations than in Malthus' work itself (Keyfitz 1982). Among the most troublesome of these is the naïve perspective that the concept of manageable population size can be operationalized as a 'simple ratio of organisms to units of a resource or resources'—naïve because it fails to take into account 'the manner in which the population is organized to use the resources' (Hawley 1950, p. 151). This flaw apparently underlies the conclusion that concepts such as 'carrying capacity' and 'limits to growth,' have failed to provide either 'analytic power, or foresight, instead of hindsight' (Davis 1990, p. 13).

Constructive insights are found in the human ecological model of system change, growth, and equilibrium. In *system* terms, growth refers to 'maturation … through the maximization of the potential for complexity and integration implicit in the technology for movement and communication possessed at a given point time' (Hawley 1986, p. 52). This perspective has direct roots in Boulding's (1953) theory of nonproportional change which suggests that a limit to growth occurs, not when some ratio of organisms to resources is reached, but when the communications and transportation costs involved in maintaining the coherence of a system become prohibitively high. Growth of systems can be expected to take the form of a logistic curve which describes a period of rapid growth which gives way to slow or no growth. Equilibrium is this view is a *temporary* state, and population increase will resume in combination with system expansion as additional organizational and technological inputs allow breaking through the upper asymptote imposed by the costs of movement and coordination of activities (Hawley 1986, 1998).

Applied specifically to demography, this insight might be translated as follows:

> As to demographic growth, its initial acceleration is a natural byproduct of … economic transformation since, predictably, the fruits of economic improvement are in part used for obtaining better health and lowered mortality. Malthusian outcomes, however, need not be the inevitable result of rapid population growth. Such growth is transitory: given economic success, the spontaneous onset of "demographic transition" can be confidently expected. Pressure built into the reward mechanisms of modern industrial society induce behavioral changes that eventually lead to low fertility, hence to low or no population growth (Demeny 1990, p. 417).

Recent worldwide fertility declines are consistent with this perspective. A 1999 United Nations publication shows that the world total fertility rate (TFR = the number of children a female surviving to the end of her reproductive life would bear given prevailing age-specific fertility rates (Keyfitz and Flieger 1990, p. 16)), after hovering around 5.0 during the period 1950–70 had fallen precipitously to 2.93 in 1990–95, with an expected TFR = 2.71 for 1995–2000 (UN 1999). By 1975–80, fertility in developed nations was already below replacement level (TFR = 1.91; for replacement, TFR is approximately 2.1, allowing for mortality of children). In LDCs, the rate for 1990–95 was 3.27, but this represented a rather astonishing decline from TFR = 6.00 in 1965–70. Based on its medium variant, the UN (1999, pp. 520–1) projects that average fertility among today's LDCs will fall to bare replacement level within the first four to five decades of the twenty-first century.

Of course, even with this major slowdown in population growth, 50 years into the twenty-first century, the world will still have a population much larger than today's roughly six billion (UN 1999, US Agency for International Development 1999). Some scholars (e.g., Bongaarts and Feeney 1998) have argued that current low fertility rates are, at least in some contexts, temporarily, rather than permanently, depressed to below replacement, while others believe the contemporary trend is likely to be permanent or at least rather persistent (Lesthaeghe and Willens 1999).

5. *Urbanization*

The concentration, deconcentration, and organization of urban populations have been central to ecological theory and research. As noted, Hawley defined human ecology as the study of the form and development of the community, of which urban places are a prime example. Although neither the *highly stylized spatial* models of urban form and growth developed by Burgess and Park, nor those of other scholars, have remained central to urban studies, these heuristic representations had a basis in analytical constructs that are important building blocks for ecological theories of the distribution and interaction of populations in both intraurban and interurban systems.

5.1 *Intraurban Systems: Land Use and Population Distribution*

Among the most directly relevant of human ecological insights for understanding *intrametropolitan* systems are those having to do with land use and population concentration/deconcentration. Of special interest is

Burgess' notion of geographical gradient as a basis for the hypothesis predicting an inverse relationship between intensity of land use by a population and distance from center to the periphery of cities (cf. Hawley 1971). This relationship, formalized by Colin Clark (1951), can be expressed in the form of natural logarithms as

$$ln\, d_x = ln\, d_o - bx \qquad (1)$$

where d_x is the density of population at distance x from the center, d_o is the central density, and b is an empirically derived density gradient. Reflected in this equation is a theory of land use which begins with the premise that functional differentiation and interdependence of populations and organizations in a city require mutual accessibility. Since, *ceteris paribus*, the center of a city is the point most accessible to all other parts of the city, there will be competition for core location, thereby driving up the cost of land at or near the center which, in turn, requires that more intensive use must be made of space. As ease of access declines with distance from the core, so too does the cost of land and the intensity of its use (Berry and Kasarda 1977, Hawley 1971, Turner 1994). The magnitude of the insight afforded by the human ecological theory of urban land use for demographers and others interested in the distribution and redistribution of population is seen in the fact that the validity of this model has been demonstrated by research using data covering a period of over a century and a half and for more than 100 cities (Frisbie and Kasarda 1988, p. 634).

Ecological theory also predicts that reduction in the friction of movement of population, materials, and ideas over time and space will cause the slope of the gradient to become less and less steep. As modern means of transportation developed and dramatic advances were achieved in accumulating, processing, and disseminating information, simple models based on a core-to-periphery gradient in a monocentered city have become progressively less relevant as metropolitan areas have increasingly become multinodal. But, because it is the systemic relationships reflecting the need for mutual accessibility of functionally interdependent populations, not spatial patterns, *per se*, that are central to the theory, the relevance of the ecological perspective remains in the face of rapidly changing urban systems. To illustrate, in the 1970s, as documented by numerous scholars and summarized by Edmonston and Guterbock (1984), several factors seemed inimical to the continuation of the rapid pace of suburbanization. Transportation costs had risen dramatically due to the reduction of petroleum supplies by major producing nations, while income growth in constant dollars slowed notably and housing prices soared—all of which raised the cost of moving to, and commuting from, the urban periphery. In addition, demographic composition had shifted such that the types of families that had traditionally fueled movement to suburbs, i.e., larger households headed by couples of prime child-bearing age, became proportionately fewer. Taking the coefficients of Clark's equation as dependent variables for more than 200 metropolitan areas, Edmonston and Guterbock demonstrated that mean central densities of cities continued to decline over a 25-year time span and that 'the density gradient dropped consistently' (1984, p. 918). That is, the initial concentration at the core of cities occurred due to the need for accessibility. The large negative sign of the density gradient reflected the less developed transportation and communications technology available during the early stage of city building, but the slope flattened as technological innovations emerged. When the cost of movement over time and space rose sharply in the 1970s, the deconcentration of population nevertheless continued because, by that time, jobs and market outlets (industrial parks and shopping malls), recreational facilities, and utilities had also expanded well into the urban periphery (Berry and Kasarda 1977, Edmonston and Guterbock 1984). Thus, in the face of altered social and economic conditions, the human ecological paradigm continues to be a reliable guide for research on population distribution and redistribution. The reason, in part, is due to the conceptualization of space as a time-cost variable. What organized populations seek to minimize is not *distance* between interdependent enitities, but the cost in time, effort, and other resources required for regular interaction (Stephan and McMullin 1981).

Another crucial aspect of intraurban population distribution in the US is residential segregation by race and ethnicity. As mentioned above, this topic is addressed in detail later in this volume. However, a brief discussion is in order here because it is demographic analysis that has most fully documented segregation, and a substantial portion of the explanation for these findings has been provided by human ecological theory. A prime example is seen in the work of Massey and colleagues who, over many years, have traced the sharp separation of African Americans from whites (and the lower levels of separation of Latinos and Asians from whites), taking into account variation in socioeconomic status, nativity, and other factors (Massey 1979, Massey and Denton 1989, 1993, Farley and Frey 1994). Based on an examination of segregation in the US and five other nations, Massey concluded that 'the operation of common processes of *ethnic* concentration and dispersal are well-summarized by the ecological model' (Massey 1985, p. 315; emphasis added), especially as the model pertains to chain migration, invasion, and succession of populations, the urban economy, and the housing market. Although human ecological theory is much less helpful in accounting for black–white segregation, Massey and Eggers (1990) have usefully addressed this phenomenon in research on 'the ecology of inequality.'

5.2 *Interurban Systems*: *National and World Patterns*

A useful point of departure for many demographers interested in *intermetropolitan* organization has been the classification of cities according to horizontal and vertical dimensions. The horizontal dimension refers to functional specialization of cities in various goods and services at various stages of production for the market, as operationalized in terms of types of industrial activities (Duncan et al. 1960, Frisbie and Kasarda 1988). The vertical dimension refers to the level of dominance exercised by cities in a metropolitan hierarchy as measured by performance of coordinative functions, including financial services and organization of the market (Duncan et al. 1960, Eberstein and Frisbie 1982, Wanner 1977) on a scale that may be regional, national, or international in scope (Frisbie and Kasarda 1988, Wilson 1984).

Moving down the ladder of abstraction leads to a crucial hypothesis, viz., that a positive association will exist between the performance of higher-order functions and extensiveness of involvement in the urban web of interdependence. An empirical test (Eberstein and Frisbie 1982), using data on 171 US urban centers, with metropolitan function defined as performance of financial, commercial, administrative, and distributive activities and extensiveness of interdependence measured in terms of volume and variety of trade flows and number of exchange linkages among urban centers, provided considerable support for this hypothesis. In a study of the world system of cities, Meyer concluded that '(t)he hypotheses of ecological dominance were confirmed' (1986, p. 553). Specifically, Meyer found that, with a focus on the key functions of banking and finance, world core cities (e.g., New York, London, Tokyo) 'dominate the peripheral metropolises of South America, and within the core the upper level metropolises exert even greater dominance than the lower level metropolises' (1986, p. 553). In addition to the general support provided for ecological theory, such findings reinforce the social demographic point that 'while all metropolises are large cities, not all large cities are metropolises. Population size is a concomitant; function is the keynote' (Vance and Sutker 1954, p. 115).

6. *Migration*

'Change without movement is impossible' (Hawley 1950, p. 324). Based on this tenet, human ecology views migration (whether within or across national boundaries) as a means through which something approaching a balance between population size and life chances may be *temporarily* achieved. Many of the most fundamental concepts employed in migration research are rooted in human ecological theory, including the recognition that, because migration involves a destination as well as a place of origin, it is necessary to consider both 'push' and 'pull' factors. Recently, human ecological models, which also include population size and distances between origin and destination as predictors, have been able to account for well over half the variation in migration streams of interstate migration in the US (Poston and Mao 1996) and interprovincial migration in China (Poston and Mao 1998). The ecological view of the push of 'overpopulation' as an impetus to migration refers not only to an excess of population to land, food, or some other natural resource, but more generally to availability of ecological niches in the form of job opportunities, especially as these are generated by the movement of capital (Hawley 1950).

Another contribution is seen in the notion of 'chain migration,' a phenomenon to which classical ecologists called attention. Chain migration means that 'migrants do not select destinations randomly, but follow social networks … to specific jobs in particular neighborhoods of selected cities' (Massey 1985, p. 318). The concept has enhanced our understanding of how residential segregation occurs and of how migrants adapt at destination through drawing on 'social capital' provided by co-ethnics (Massey 1985, 1991).

Clearly, the human ecological perspective must be supplemented by other theories of migration (Portes and Fernandez-Kelly 1989). One problem with the 'push—pull' model is that it cannot explain, '(g)iven the same set of expelling forces and factors of attraction, why is it that some individuals leave while others stay' (Portes 1991, p. 76). This weakness is inherent in the macrolevel approach of human ecology. Microlevel studies essentially seek an answer to the question 'Who moves and why?' By contrast, for human ecological studies of migration, the central question is 'Where do people go and why?'

While ecological theories of migration involve much more than a simple 'push–pull' hypothesis, there is, nevertheless, a certain tendency to neglect population movements that are forced or 'impelled' by political agents at origin (Petersen 1958). An analogous gap is evident at destination due to the lack of attention paid by ecologists to the impact of immigration policies instituted by receiving countries seeking to restrict, and/or influence the composition of, immigration flows (Freeman and Bean 1997). Further, despite the importance attributed by Hawley to capital flows as a determinant of immigration (1950, p. 330), research in human ecology seems to have failed to appreciate fully the importance of the penetration of capital for the creation of economic imbalances that stimulate immigration (Portes and Fernandez-Kelly 1989).

The theoretical contributions of human ecology to this line of research did not cease with the appearance of Human Ecology. Hawley has extended his own ideas concerning the spatial movement and redistribution of population, particularly with respect to urban-oriented migration (1971, 1986). More recently,

as part of a comprehensive theory of 'assembling of human populations' that involves geopolitical as well as ecological processes, Turner has set forth a set of propositions that envisions the rate of net immigration as 'a positive and additive function of: (a) the level of transportation and communication technologies, (b) the rate and extent of contact with external populations, (c) the differentiation as well as volume and velocity of markets, (d) the diversity of the population, and (e) the size of population settlements' (Turner 1994, p. 88). Coupled with the notions of chain migration and population flows preceded by capital flows, Turner's formulation seems highly pertinent to many of the international population flows that have characterized the twentieth century.

7. *Summary*

Contemporary human ecology has sustained and helped extend the theoretical foundation of macrolevel demography. It has also contributed in numerous ways to the understanding of specific demographic phenomena, as illustrated by reference to research on population growth, urban form and process, and migration, with brief attention given to residential segregation by race and ethnicity. While other examples could have been employed, it is palpable that human ecology provides a paradigm that, in spite of certain weaknesses and oversights, greatly facilitates comprehension of how populations grow and distribute themselves within, and adapt to, the natural and social environments.

See also: Demography, History of; Ecological Economics; Ecological Fallacy, Statistics of; Human Ecology; Migration, Economics of; Migration, Theory of; Population Ecology; Population Pressure, Resources, and the Environment: Industrialized World; Residential Segregation: Sociological Aspects; Urban Geography

Bibliography

Berry B J L, Kasarda J D 1977 *Contemporary Urban Ecology*. Macmillan, New York

Bongaarts J, Feeney G 1998 On the quantum and tempo of fertility. *Population and Development Review* **24**: 271–91

Boulding K E 1953 Toward a general theory of growth. *Canadian Journal of Economics and Political Science* **19**: 326–40

Carroll G R 1984 Organizational ecology. *Annual Review of Sociology* **10**: 71–93

Clark C 1951 Urban population densities. *Journal of the Royal Statistical Society Series A* **114**: 490–6

Davis K 1975 Asia's cities: problems and options. *Population and Development Review* **1**: 71–86

Davis K 1990 Population and resources: fact and interpretation. *Population and Development Review* **16**(suppl.): 1–21

Demeny P 1990 Tradeoffs between human numbers and material standards of living. *Population and Development Review* **16**(suppl.): 408–21

Duncan O D 1959 Human ecology and population studies. In: Hauser P M, Duncan O D (eds.) *The Study of Population*. University of Chicago Press, Chicago

Duncan O D 1964 Social organization and the ecosystem. In: Faris R E L (ed.) *Handbook of Modern Sociology*. Rand-McNally, Chicago

Duncan O D, Scott W R, Lieberson S, Duncan B 1960 *Metropolis and Region*. Johns Hopkins University Press, Baltimore, MD

Eberstein I W, Frisbie W P 1982 Metropolitan function and interdependence in the US urban system. *Social Forces* **60**: 676–700

Edmonston B, Guterbock T M 1984 Is suburbanization slowing down?: Recent trends in population deconcentration in US metropolitan areas. *Social Forces* **62**: 905–25

Erlich P 1968 *The Population Bomb*. Ballentine, New York

Farley R, Frey W H 1994 Changes in segregation of Whites from Blacks during the 1980s: Small steps towards a more integrated society. *American Sociological Review* **59**: 23–45

Ford T F, De Jong G F 1970 *Social Demography*. Prentice-Hall, Englewood Cliffs, NJ

Freeman G P, Bean F D 1997 Mexico and US worldwide immigration policy. In: Bean F D, de la Garza R O, Roberts B R, Weintraub S (eds.) *At the Crossroads: Mexico and US Immigration Policy*. Lanham, Boulder, CO

Frisbie W P, Al-Khalifah H M 1998 Ecology's contribution to cross-national theory and research. In: Micklin M, Poston Jr D L (eds.) *Continuities in Sociological Human Ecology*. Plenum, New York

Frisbie W P, Kasarda J D 1988 Spatial processes. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA

Guest A M 1978 Suburban social status: Persistence or evolution. *American Sociological Review* **43**: 251–64

Hannan M T, Freeman J 1977 The population ecology of organizations. *American Journal of Sociology* **82**: 929–64

Hawley A H 1950 *Human Ecology: A Theory of Community Structure*. Ronald, New York

Hawley A H 1971 *Urban Society: An Ecological Approach*. Ronald, New York

Hawley A H 1986 *Human Ecology: A Theoretical Essay*. University of Chicago Press, Chicago

Hawley A H 1998 Human ecology, population and development. In: Micklin M, Poston Jr D L (eds.) *Continuities in Sociological Human Ecology*. Plenum, New York

Keyfitz N 1982 The evolution of Malthus's thought. In: Keyfitz N (ed.) *Population Change and Social Policy*. Abt Books, Cambridge, MA

Keyfitz N, Flieger W 1990 *World Population Growth and Aging*. University of Chicago Press, Chicago

Krivo L J, Frisbie W P 1982 Measuring change: the case of suburban status. *Urban Affairs Quarterly* **17**: 419–44

Lesthaeghe R, Willens P 1999 Is low fertility only a temporary phenomenon in the European Union. *Population and Development Review* **25**: 211–28

Linz J J 1969 Ecological analysis and survey research. In: Dogan M, Rokkan S (eds.) *Quantitative Ecological Analysis in the Social Sciences*. MIT Press, Cambridge MA

Massey D S 1979 Residential segregation of Spanish Americans in United States urbanized areas. *Demography* **16**: 553–63

Massey D S 1985 Ethnic residential segregation: A theoretical synthesis and empirical review. *Sociology and Social Research* **69**: 315–50

Massey D S 1991 Economic development and international migration in comparative perspective. In: Diaz-Briquets S, Weintraub S (eds.) *Determinants of Emigration From Mexico, Central America and the Caribbean*. Westview Press, Boulder, CO

Massey D S, Denton N A 1989 Hypersegregation in US metropolitan areas: Black and Hispanic segregation along five dimensions. *Demography* **26**: 373–91

Massey D S, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA

Massey D S, Eggers M L 1990 The ecology of inequality: Minorities and the concentration of poverty, 1970–1980. *American Journal of Sociology* **95**: 1153–88

McKenzie R D 1933 *The Metropolitan Community*. McGraw-Hill, New York

Meadows D H, Meadows D L, Randers J, Behrens III W W 1974 *Limits to Growth: A Report for the Club of Rome Project on the Predicament of Mankind*, 2nd edn. Universe Books, New York

Meyer D R 1986 The world system of cities: Relations between international financial metropolises and South American cities. *Social Forces* **64**: 553–81

Murdock S H, Backman K, Hwang S, Hamm R R 1992 Sustenance specialization and dominance in international and national ecosystems: Implications for post-1980 migration in counties in the United States. In: Jobes P C, Stinner W F, Wardwell J M (eds.) *Community, Society and Migration: Noneconomic Migration in America*. University Press of America, Lanham, MD

Namboodiri K 1988 Ecological demography: Its place in sociology. *American Sociological Review* **53**: 619–33

Park R E, Burgess E W 1921 *Introduction to the Science of Sociology*. University of Chicago Press, Chicago

Petersen W 1958 A general typology of migration. *American Sociological Review* **23**: 256–66

Portes A 1991 Unauthorized immigration and immigration reform: Present trends and prospects. In: Diaz-Briquets S, Weintraub S (eds.) *Determinants of Emigration from Mexico, Central America, and the Caribbean*. Westview Press, Boulder, CO

Portes A, Fernandez-Kelly M P 1989 Images of movement in a changing world: A review of current theories of international migration. *International Review of Comparative Public Policy* **1**: 15–33

Poston Jr D L, Frisbie W P 1998 Human ecology, sociology, and demography. In: Micklin M, Poston Jr D L (eds.) *Continuities in Sociological Human Ecology*. Plenum, New York

Poston Jr D L, Mao M X 1996 An ecological investigation of interstate migration in the United States. *Advances in Human Ecology* **5**: 303–42

Poston Jr D L, Mao M X 1998 Interprovincial migration in China, 1985–1990. *Research in Rural Sociology and Development* **7**: 227–50

Robinson W S 1950 Ecological correlations and the behavior of individuals. *American Sociological Review* **15**: 351–7

Ryder N B 1980 Where do babies come from? In: Blalock Jr H M (ed.) *Sociological Theory and Research: A Critical Appraisal*. Free Press, New York

Schnore L F 1961 The myth of human ecology. *Sociological Inquiry* **31**: 128–39

Stephan G E, McMullin D R 1981 The historical distribution of county seats in the United States: A review, critique, and test of time-minimization theory. *American Sociological Review* **46**: 907–17

Turner J H 1994 The assembling of human populations: Toward a synthesis of ecological and geopolitical theories. *Advances in Human Ecology* **3**: 65–91

United Nations 1999 *World Population Prospects*. United Nations, New York, Vol. 1

United States Agency for International Development 1999 *Popbriefs*. USAID, Washington, DC

Vance R B, Sutker S S 1954 Metropolitan dominance and integration in the urban South. In: Vance R B, Demerath N J (eds.) *The Urban South*. University of North Carolina Press, Chapel Hill, NC

Wanner R A 1977 The dimensionality of the urban functional system. *Demography* **14**: 519–37

Wilson F D 1984 Urban ecology: Urbanization and systems of cities. *Annual Review of Sociology* **10**: 283–307

P. Frisbie

Human Evolutionary Genetics

Human evolutionary genetics is the inference of human history and evolution through the study of genetic variation. While much biological variation among humans is readily observable, variation at the DNA level typically is hidden. Only in the last decades of the twentieth century, with the development of molecular biological techniques, has this variation become detectable. The hidden patterns of human genetic variation have been shaped by events and processes of the past. For instance, the differing long-term census sizes of the human and chimpanzee species have led to dramatically different levels of genetic diversity within the two species. Chimpanzees are more genetically diverse, indicating that the species has undergone less severe bottlenecks than has the human species. While genetic data may not indicate why the human species went through a bottleneck, they provide at least a partial solution to the puzzle that is human evolutionary history.

The field of human evolutionary genetics is built on the premise that one can infer evolutionary history through the analysis of human genetic variation in the context of relevant theory and nongenetic data. Patterns of genetic variation are considered to be the outcome of a set of processes that can be modeled, primarily according to population genetics and molecular evolutionary theory. The goal is to 'undo' the process, in a sense, in order to infer the past. The extent to which nongenetic data are incorporated, either explicitly or implicitly, cannot be overstated. Geographic, linguistic, archaeological, demographic, historical, and climatological data, along with findings from other areas in biology, are incorporated to varying degrees as human evolutionary geneticists draw conclusions from the genetic data.

1. Two Directions within Human Evolutionary Genetics

The study of genetic variation provides two categories of insight into human history and evolution. Each category reflects an emphasis on one set of evolutionary processes. While the pattern of variation of some DNA segments may reflect both sets of processes, one set is often presumed to have predominated.

1.1 Phenotypic Evolution

A longstanding focus of human evolutionary genetics is phenotypic evolution, including changes in morphology, physiology, and behavior. By comparing patterns of genetic variation of humans and closely related species, for example, researchers can infer the processes that led to the current human phenotype (to the extent that this phenotype has an inherited genetic basis). Also included in this category is the study of the evolution of susceptibility to disease. To date, most findings regarding phenotypic evolution have been in this area. The more general question of what, at the DNA level, distinguishes humans from other species has only recently reached the forefront of research in human evolutionary genetics but promises to be an exciting research direction. Critical in this area of research is an understanding of the relationship between the variation at the DNA level and the variation at the phenotypic level.

1.2 Population Processes and Events

A second category, exemplified by the human–chimpanzee comparison above, includes population processes and events. Researchers attempt to infer, for example, population sizes of the past, rates of migration between groups, and the timing of population movements. This was a primary focus of the field during the latter half of the twentieth century, and remains so into this century. A central question has been the timing and location of the origin of anatomically modern humans.

2. 'Population' and 'Group'

Essential to the practice of human evolutionary genetics are definitions of the terms 'population' and 'group.' The less precisely defined term, 'group,' is used here to mean any collection of individuals. In a theoretical framework, the term 'population' is defined very precisely, as a set of individuals constituting a mating pool. All individuals of the appropriate sex in the population are considered to be equally available as potential mates. Groups of humans rarely, if ever, fit this definition of a population. The boundary between one population and another is obscure. In practice, therefore, human evolutionary geneticists delineate populations along linguistic, geographic, sociopolitical, and/or cultural boundaries. A population might include, for example, all speakers of a particular Bantu language, all inhabitants of a river valley in Italy, or all members of a caste group in India.

3. 'DNA Segment'

The term 'DNA segment' is defined here as any length of DNA that is passed intact (without recombination of paternal and maternal DNA) from parent to child. Under this definition, any single nucleotide constitutes a DNA segment, as does the entire non-recombining portion of the Y chromosome. The entire mitochondrial (mt) DNA genome can also be considered an independently evolving DNA segment. Essential to this definition is the consistency of history across the nucleotides of a DNA segment. That is, without recombination, all nucleotides of a given individual's DNA segment have been inherited via the same set of ancestors (e.g., father, paternal grandfather, etc., for the non-recombining portion of the Y chromosome). Success in inferring human evolutionary history depends in part on the extent to which we can infer the history of a particular DNA segment. It also depends on the number of independently-evolving DNA segments studied, as discussed below.

While there are, as exemplified above, stretches of the human genomes (nuclear and mitochondrial) that fit the above definition of a DNA segment, most stretches of the nuclear genome have the potential to undergo recombination. When human evolutionary geneticists analyze DNA sequences such as the Human Leukocyte Antigen (HLA) region on chromosome 6, they must either assume that recombination has occurred rarely enough to be ignored, or bring recombination into their analytical framework.

4. Summarizing Genetic Variation

In summarizing genetic variation, initially researchers count the number of copies of each genetic variant in each group under study. For instance, frequencies of the variants A, B, and O (also called alleles) of the ABO blood group system, one of the earliest polymorphic (variable) genetic systems studied, have been estimated for thousands of human groups around the world.

4.1 Within Group Genetic Diversity

Gene diversity (heterozygosity), defined as the chance that two DNA segments drawn randomly from a

group are different, summarizes within-group variation and is estimated from allele frequencies. Other measures of diversity take into account additional information, such as the number of differences between two DNA segments (nucleotide diversity), or the total number of variable positions along a DNA segment (number of segregating sites), within a group. Molecular evolution and population genetic models allow the interpretation of such diversity estimates in terms of mutation rates, population size changes, natural selection, and migration.

4.2 Among and Between Group Diversity

Human evolutionary geneticists often condense information regarding the genetic variation among a set of human groups into a single measure, typically called F_{ST} (Hartl and Clark 2000). Estimates of F_{ST} and similar statistics lie behind the oft-mentioned finding that the greatest proportion of the total genetic variation in the human species lies within human groups rather than among them. This has been interpreted as evidence for a relative lack of isolation among human groups. While F_{ST} summarizes variation among groups, genetic distances summarize the genetic difference between a pair of populations (Hartl and Clark 2000). Relatively small genetic distances are considered to be evidence that two groups have been less isolated from one another than have other pairs of groups. Differences in population sizes, however, strongly influence estimates of genetic distance.

5. Patterns of Variation for Different Segments of the Human Genome Evolve Independently

At each segment of the human genome, variation is introduced through mutation and subsequently modified through natural selection, random genetic drift, and gene flow between groups. The current pattern of variation of each such segment tells a story. Closely linked segments (those near one another on a chromosome that undergoes recombination, or on the same non-recombining chromosome) are likely to tell a similar story, at least going back a short time. In general, however, segments of the human genome evolve independently and therefore have different histories. The most recent common ancestor (MRCA) of all contemporary representatives of a particular DNA segment (e.g., the non-recombining portion of the Y chromosome) existed in a particular place at a particular point in time. The MRCA of another, independently evolving DNA segment (e.g., the mitochondrial genome) is likely to have existed in quite a different place, at a different point in time.

In the absence of natural selection, the pattern of variation at a particular segment reflects population sizes and movements of people. Even if each DNA segment evolves independently and therefore has its own history, population size and migration influence the variation at all segments of the biparentally inherited parts of the genome similarly. Human evolutionary geneticists, therefore, combine the information from multiple DNA segments to draw conclusions regarding population processes. This consideration of multiple gene histories increases the robustness of conclusions (Mountain 1998).

An important consideration is the rate of mutation at each genomic location. DNA repair mechanisms that influence the rate of mutation differ between nuclear and mitochondrial DNA. Furthermore, DNA segments that include repeated elements (e.g., CACACACACACA) tend to mutate more rapidly than do nonrepetitive segments. Indeed, even within a short DNA segment of one hundred nucleotides, mutation rates may vary by an order of magnitude. The rate of mutation strongly influences the time depth over which a DNA segment is informative. A slowly mutating segment, for instance, will be uninformative over a short timescale. Mutation rates are very difficult to measure or estimate, primarily because mutation typically is rare (usually far fewer than 1 mutation per 10,000 nucleotide sites per generation).

With respect to gene flow, it is important to distinguish between interspecific and intraspecific questions regarding history and evolution. For questions regarding multiple species (e.g., the Hominoids), it is typically assumed that gene flow between species is essentially nonexistent. One consequence of a lack of exchange between groups is a greater chance of observing fixed genetic differences. That is, all individuals of one group are of one genetic type, while all individuals of another group are of a different type. Intraspecific questions must be approached differently in that gene flow may have occurred between groups below the species level. The current paucity of fixed differences between human populations reflects, among other factors, a great deal of genetic exchange between populations. There is a great deal of debate regarding the precise nature of genetic exchange between human populations. Migration models typically include a constant rate of gene flow over time. While such a model is likely to apply to human history at some level, more appropriate models include changing rates of gene flow and the possibility of rare but influential major migration events (Fix 1999).

When natural selection has a stronger influence on genetic variation than random genetic drift, the pattern of variation within and among groups is affected. Natural selection has clearly played a strong role in generating human genetic variation, especially in constraining the observable variation. While an environmental factor such as average temperature may influence variation at a number of loci simultaneously, and in a similar direction, other factors may influence only one or two loci. This provides one

way of distinguishing between a model of natural selection and a model involving population processes alone; the latter are expected to influence all loci, while environmental factors influence only a subset. Incorporating natural selection in addition to population processes into evolutionary models greatly increases the total number of models under consideration, so that demonstrating the impact of positive natural selection (i.e., new alleles arise within an individual and provide that individual with a selective advantage over other individuals) is a serious challenge. One thoroughly investigated case of natural selection influencing human genetic variation is that of the rise of lactose tolerance in adults (Durham 1991).

6. *Development of the Field*

As currently practiced, research in human evolutionary genetics builds upon over a century of the development of a theoretical framework, of technical tools (data analysis and biological), and of collections of data. Prior to 1900 there was essentially no research in the realm of human evolutionary genetics. The latter half of the nineteenth century was critical, however, in that C. Darwin, G. Mendel, and their contemporaries provided the basis for a theoretical framework with an emphasis on variation and natural selection. This continued into the twentieth century with the classic work of G. H. Hardy and W. Weinberg in 1908 (Provine 1971). Independently, they described the simple but elegant relationship between the frequencies of an allele and the frequencies of resulting genotypes (in a population, with individuals mating at random with respect to genotype).

6.1 *Development of Tools and Theory: 1900–50*

The first half of the twentieth century saw the development of some of the essential elements in the field—technical tools and population genetic theory. Serological (blood typing) techniques, developed by 1900, made possible the genotyping of the ABO locus and led to a decrease in the risk associated with blood transfusions. Initial studies of the ABO locus focused on the variation among United States' soldiers. The resulting data are recognized as the first population data for what are now called 'classical genetic markers': Blood groups, and protein and enzyme polymorphisms. The discovery of blood groups and protein polymorphisms continued slowly during this period.

Other dramatic developments relevant to the field of human evolutionary genetics prior to 1950 lay in the area of population genetic theory, 'the quantitative synthesis of Mendelism and Darwinism' (Provine 1971, p. 129). As early as the 1920s Sewall Wright, Ronald A. Fisher, and J. B. S. Haldane were publishing in the field. Fisher's *The Genetical Theory of Natural Selection* (1930), Wright's 'Evolution in Mendelian Populations' (1931) and Haldane's *The Causes of Evolution* (1932) are part of a proliferation of theoretical work of this period. They were joined by Theodosius Dobzhansky, whose *Genetics and the Origins of Species* (1937) was just one of several major contributions. Together, these authors outlined the theory behind the roles of natural selection, mutation, migration, and random genetic drift in the genetic evolution of populations (see Provine 1971, for an accessible history of the origins of theoretical population genetics). These theoreticians continued contributing well into the second half of the century.

By 1950, a number of classical genetic markers had been studied across human populations (Cavalli-Sforza and Bodmer 1971). Variation within the Rh genetic system, discovered in 1939, was considered in light of population genetic theory, as researchers attempted to explain the existence of alleles that are clearly deleterious to individuals under certain conditions. While studies often focused on the health impact of genetic variation, there was also an anthropological component: Researchers sought to measure and explain the genetic differences among human groups. The elevated frequency of Rh − individuals among Basques, for instance, was initially seen as evidence that they were descendents of the earliest modern humans in Europe. In 1948, Linus Pauling and colleagues described the role of the different forms of hemoglobin in sickle cell anemia, developing a revolutionary technique called electrophoresis to distinguish the different forms. At about the same time, J. B. S. Haldane put forth a hypothesis to explain the high frequency of another hemoglobin disorder, alpha thalassemia. He suggested that the blood cells of individuals heterozygous (having two different alleles) at this locus were more resistant to attack by malarial protozoa. This hypothesis was subsequently extended to sickle-cell disease. The story of the advantage of the sickle cell allele in malarial regions remains one of the most convincing examples of the population genetic phenomenon known as heterozygote advantage. For a succinct summary of the research see Durham (1991).

6.2 *Development of the Field: 1950–75*

The third quarter of the twentieth century saw rapid development in several dimensions of human evolutionary genetics. In the mid-1950s, Mourant published his first *Distributions of Human Blood Groups* (1954) with tables of population frequencies for nine blood groups and the sickle cell trait. In the mid 1960s, Harris, Hubby, and R. Lewontin (see Lewontin 1974, for references) discovered that a large fraction of

proteins show variation within a species. This discovery led to a dramatic increase in the number of informative polymorphic systems typed across human populations: by the early 1970s, Cavalli-Sforza and Bodmer (1971) list 27 such systems. In 1972 Lewontin (1972) published a seminal article concluding that most human genetic variation is found within local human populations. Remarkably, today's more sophisticated techniques and larger data sets consistently confirm this original finding.

A number of research groups carried out detailed studies of nonWestern populations of particular interest to anthropologists at the time. J. Neel and colleagues focused on South American populations, particularly the Yanomama of southern Venezuela, beginning in the 1960s. They were surprised by the extent of genetic differentiation between Yanomama villages, attributing this to the small size of groups founding the villages. For details of this and other regional studies conducted during the 1960s, see Crawford and Workman (1973).

During this period new numerical methods, phylogenetic methods, and computer simulation were introduced and were immediately applied to the burgeoning set of data on classical genetic markers. In the early 1960s, L. Cavalli-Sforza and A. Edwards developed several methods for generating trees from population genetic data, and applied these to data for five blood groups for 15 human populations from around the world. While their initial results are not entirely consistent with results from larger, subsequent data sets, many of the features of that initial tree can be found in recent analyses as well. One surprising finding was that genetic and anthropometric traits (physical dimensions) gave very different pictures of differentiation among human groups. While anthropometric traits indicated a similarity between Africans and Australians, genetic traits (blood groups) indicated then, and still do, that these geographically distant populations are genetically distant as well.

Meanwhile theoretical population geneticists, particularly P. A. P. Moran and his students, as well as M. Kimura, continued to extend and develop models relevant to the field of human evolutionary genetics (Ewens 1979). In the context of new understanding about DNA and mutation, Kimura (1969) argued that most mutations, or variants, found in living individuals provide neither a survival/reproductive disadvantage nor advantage to the individuals whose genomes include these mutations. Instead, he claimed, these mutations are neutral, or nearly so. If most variation is neutral, then patterns of variation are to be interpreted in terms of population history rather than in terms of natural selection. During this period, statistical methods were developed to test the null hypothesis of no natural selection. The subsequent finding that the null hypothesis was difficult to reject (Lewontin 1974) had a direct impact on the field of human evolutionary genetics. It reinforced the idea held by some geneticists that random genetic drift (random changes in allele frequencies over time) plays an important role in microevolution (see Kimura 1983, for discussion of controversy). While drift was certainly recognized as a powerful evolutionary force prior to Kimura's work, his 'Neutral Theory' of molecular evolution combined with the results of tests for natural selection led to a pendulum swing, with increased emphasis on drift (Weiss and Buchanan 2000). This was particularly the case for recently derived variation such as that found within the human species.

A number of research developments in biology, such as the discovery of the structure of DNA by J. Watson and F. Crick in 1953, had a significant impact on the field only much later, after DNA sequencing became sufficiently routine for population studies. The development of the clarifying concepts of cladistics by W. Hennig during the 1950s and 1960s had a much more immediate impact within systematics than within human evolutionary genetics. Nonetheless, Hennig's work has clearly had an influence in recent years, as phylogenetic methods using parsimony criteria have been applied to human DNA sequence data. In 1960, L. Pauling and E. Zuckerkandl introduced the concept of mutations occurring regularly enough at a particular nucleotide site to be considered 'clock-like.' To the extent that evolution at the molecular level is clock-like, researchers can infer dates, at least relative dates, given the number of mutational differences between DNA sequences.

6.3 Development of the Field: 1975–90

By the early 1980s, new techniques for assessing genetic variation supplemented serological and protein typing. Restriction enzymes were used to cut DNA segments at very precise locations. The sizes of the resulting DNA fragments varied depending on the underlying DNA sequence of an individual. Gel electrophoresis, introduced much earlier (1948), enabled researchers to compare and estimate fragment sizes. This approach led to an explosion in the number of known human polymorphisms during the 1980s. These were initially ascertained for medical genetic research (in the search for genes associated with disease), but were immediately recognized as valuable for human evolutionary genetics research. While thousands of polymorphisms were discovered, only a few hundred were actually studied in a geographically broad set of human populations.

During the 1980s DNA sequencing began replacing restriction analysis as the tool of choice within human evolutionary genetics. The initial focus was on the most highly variable regions of the mitochondrial genome, with its large number of copies per cell. The comparison of mtDNA sequences from numerous individuals throughout the world brought public

attention to the field. The public was captivated by the notion of 'the mother of us all,' as the mtDNA ancestor was labeled. Many people were also surprised by the interpretations of the patterns of mtDNA variation: That the common ancestors of living members of the human species lived relatively recently (less than 200,000 years ago rather than 1–2 million years ago (Klein 1999)) and that these ancestors probably lived in Africa.

7. *Human Evolutionary Genetics at the Turn of the Millennium: Population Processes and Events*

7.1 *Human Genome Project*

The rapid development of human evolutionary genetics over the final decade of the twentieth century owes a great deal to the Human Genome Project. Much of the technology developed under the auspices of that project (e.g., improved speed and accuracy of DNA sequencing methods) is currently applied in the field. In 1991, a group of human evolutionary geneticists led by L. Cavalli-Sforza proposed that the Human Genome Project be extended to include consideration of variation among individuals. Almost immediately, the proposal encountered criticism from groups concerned that geneticists might exploit individuals contributing DNA samples. Nonetheless, a subsequent Human Genome Project plan did include consideration of variation across individuals as a major goal. New methods of detecting variation among individuals were developed. The result is hundreds of thousands of nucleotide sites known to be polymorphic in humans across the entire human genome. Very few, however, have been studied in a broad set of human populations.

7.2 *DNA of Ancient Peoples*

Under some conditions (e.g., low temperature, aridity) DNA persists relatively intact within bones and teeth for thousands, if not hundreds of thousands, of years (Wayne et al. 1999). During the 1980s human evolutionary geneticists began to take advantage of this preservation of genetic information of the past. DNA from the bones of Neanderthal specimens, for example, has been compared with that of living humans in order to address the question of whether Neanderthals and anatomically modern humans interbred on a large scale within the last 100,000 years. The DNA from early anatomically modern humans promises to provide even further insight into human evolution during this time period (Relethford 2001). With respect to more recent human evolution, ancient DNA of Native Americans living prior to 500 years ago has demonstrated no evidence of a dramatic loss of genetic variation in the intervening period, despite a dramatic reduction in the number of Native Americans (O'Rourke et al. 2000).

7.3 *DNA of Living Peoples: Inferences Across a Range of Time Scales*

While the DNA of ancient peoples has great potential to provide insight into human evolution, the DNA of extant peoples has been studied much more extensively (Cavalli-Sforza et al. 1994). The DNA of living people retains the signatures of events and processes at various time depths ranging from hundreds of thousands of years down to tens of years. Autosomal, Y chromosomal, and mtDNA of living people has shed light, for instance, on the peopling of Europe over the last 40,000 years. Researchers have estimated the relative contributions of Paleolithic and Neolithic Europeans to the present-day European gene pool. Independently evolving segments of the human genome are consistent in supporting the hypothesis of a relatively large contribution by Paleolithic Europeans to today's European gene pool, while Neolithic peoples from the Near East appear to have contributed roughly a quarter of that gene pool (Lell and Wallace 2000).

Two examples illustrate the potential for making inferences on a shorter time scale (a few thousand years at most). The social structure of the Hindu caste system, for instance, has been investigated using a combination of Y chromosome and mtDNA. Human evolutionary geneticists have compared the relative rates of male and female movement between castes, concluding that the caste system is driven by the social mobility of women (Bamshad et al. 1998). As a second example, oral traditions have been examined in light of DNA variation. According to oral history, the Lemba, a Bantu-speaking people of southern Africa, descend from Jews who moved from the Middle East to Yemen and then to southern Africa within the last 2,700 years. The finding that Lemba Y chromosomes are similar to those common amongst Jews and rare amongst the Lemba's African neighbors supports this tradition (Owens and King 1999).

7.4 *New Theoretical Tools*

While much of the population genetic theory and many of the data analysis tools relevant to human evolutionary genetics were developed prior to the 1990s, valuable new tools have been introduced more recently. Among these are Markov Chain Monte Carlo (MCMC) methods (see, for example, Griffiths and Tavare 1995), and coalescent-based approaches (Kingman 1982). MCMC methods make possible

estimation of effective population size, population growth rate, migration rate, and mutation rate. The coalescent approach to inference involves considering the history, or genealogy, of a sample of present-day genes. The approach has been most valuable in efforts to date events of gene histories such as the age of the MRCA of all human Y chromosomes (Thomson et al. 2000). Most applications of the genealogical approach have been in the context of lengthy, non-recombining segments of the genome (i.e., mtDNA and the non-recombining portion of the Y chromosome). A major goal in the field today is to apply the same approach to a large number of segments of the autosomes and X chromosome.

8. *Human Evolutionary Genetics at the Turn of the Millennium: Phenotypic Evolution*

For an unknown number of segments of the human genome, current variation among human beings or among primates reflects the impact of natural selection. In these cases, while population processes and events have played a role, differences in fitness between individuals with different alleles have had greater influence. The most readily detected evidence of selection is in the context of disease, as exemplified above in the case of sickle cell anemia, malaria, and the hemoglobin genes. The field of human evolutionary genetics overlaps with medical genetics in that researchers in both fields seek to understand the origins of any genetic variation that appears to influence disease. The history of the gene associated with cystic fibrosis, for example, has been explored extensively.

A current focus of human evolutionary genetics is phenotypic traits other than disease, such as locomotion, dentition, and skin color (Weiss and Buchanan 2000). While little is known about the genetic diversity underlying skin color, variation at one gene, the melanocortin-stimulating hormone receptor gene, is known to be associated with red hair and fair skin that burns easily. Genetic diversity at this locus is significantly higher than the average nucleotide diversity in human populations and is likely to be an adaptive response to variation in environmental factors such as the availability of sunlight (Owens and King 1999).

Recently, researchers within human evolutionary genetics have begun focusing their attention upon the question, 'What makes us human?' Primate evolutionary geneticist Morris Goodman (1999) and others have argued for a 'Human Genome Evolution Project.' The goal is to identify features unique to the human genome that influence such traits as cognition, the form of the hind limb, and susceptibility to disease. There are approximately 40 million nucleotide differences between the human and chimpanzee genomes amongst the roughly 2×10^9 base pairs of non-repetitive DNA. While most of these DNA level differences appear to confer no difference at the phenotypic level, a subset clearly does. The challenge is to identify that subset.

Known differences between humans and chimpanzees include chromosomal inversions and fusions, gene duplications, and repetitive element insertions (Varki 2000). Changes in the regulation of gene expression, however, are expected to be a more common source of phenotypic change (Varki 2000, Weiss and Buchanan 2000). Goodman and colleagues have described one example, wherein γ globin genes went from being expressed only within the embryo to being expressed primarily during the fetal stage. This change correlates with an evolutionary trend towards prolonged fetal life in the anthropoid primates (Goodman 1999). The search for other such changes is likely to occupy human evolutionary geneticists for the foreseeable future.

9. *The Human Evolution Puzzle: Genetics Provides Some of the Pieces*

The study of human evolution takes place within a broad range of academic disciplines. These include the fields of paleoanthropology and archaeology (Klein 1999), as well as such seemingly disparate fields as linguistics, developmental biology (Weiss and Buchanan 2000) and paleoclimatology (Hewitt 2000). Only within the context of findings in these fields can the study of genetic variation provide insight into some of the many details that constitute the evolutionary history of the human species. Human evolutionary geneticists face an exciting challenge as they attempt to integrate their findings regarding human evolution with those of counterparts in these linked disciplines.

See also: Cultural Evolution: Overview; Evolution, History of; Evolutionary Selection, Levels of: Group versus Individual; Genotype and Phenotype; Human Genome Diversity and its Value; Human Genome Project: German Perspective; Human Genome Project: Japanese Perspective; Population Dynamics: Theory of Stable Populations; Primates, Evolution of

Bibliography

Bamshad M J, Watkins W S, Dixon M E, Jorde L B, Rao B B, Naidu J M, Ravi Prasad B V, Rasanayagam A, Hammer M F 1998 Female gene flow stratifies Hindu castes. *Nature* **395**: 651–52

Cavalli-Sforza L L, Bodmer W 1971 *The Genetics of Human Populations*. W H Freeman, San Francisco, CA

Cavalli-Sforza L L, Piazza A, Menozzi P 1994 *History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

Crawford M H, Workman P L (eds.) 1973 *Methods and Theories of Anthropological Genetics*. University of New Mexico Press, Albuquerque, NM
Dobzhansky T 1937 *Genetics and the Origins of Species*. Columbia University Press, New York
Durham W H 1991 *Coevolution*. Stanford University Press, Stanford, CA
Ewens W J 1979 *Mathematical Population Genetics*. Springer-Verlag, New York
Fisher R A 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK
Fix A 1999 *Migration and Colonization in Human Microevolution*. Cambridge University Press, Cambridge, UK
Goodman M 1999 The genomic record of humankind's evolutionary roots. *American Journal of Human Genetics* **64**: 31–9
Griffiths R C, Tavare S 1994 Simulating probability-distributions in the coalescent. *Theoretical Population Biology* **46**: 131–59
Haldane J B S 1932 *The Causes of Evolution*. Harper & Row, New York
Hartl D L, Clark A G 2000 *Principles of Population Genetics*. Sinauer Associates, Sunderland, MA
Hewitt G 2000 The genetic legacy of the quaternary ice ages. *Nature* **405**: 907–13
Kimura M 1969 The rate of molecular evolution considered from the standpoint of population genetics. *Proceedings of the National Academy of Sciences, USA* **63**: 1181–8
Kimura M 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK
Kingman J F C 1982 The coalescent. *Stochastic Processes and their Applications* **13**: 235–48
Klein R G 1999 *The Human Career: Human Biological and Cultural Origins*. University of Chicago Press, Chicago, IL
Lell J T, Wallace D C 2000 The peopling of Europe from maternal and paternal perspectives. *American Journal of Human Genetics* **67**: 1376–81
Lewontin R 1972 The apportionment of human diversity. *Evolutionary Biology* **6**: 381–98
Lewontin R 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York
Mountain J 1998 Molecular evolution and modern human origins. *Evolutionary Anthropology* **7**: 21–37
Mourant A E 1954 *The Distribution of Human Blood Groups*. Blackwell, Oxford, UK
O'Rourke D H, Hayes M G, Carlyle S W 2000 Spatial and temporal stability of mtDNA haplogroup frequencies in native North America. *Human Biology* **72**: 15–34
Owens K, King M-C 2000 Genomic views of human history. *Science* **286**: 451–3
Provine W B 1971 *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago, IL
Relethford J H 2001 Ancient DNA and the origin of modern humans. *Proceedings of the National Academy of Sciences, USA* **98**: 390–1
Simpson G G 1944 *Tempo and Mode of Evolution*. Columbia University Press, New York
Thomson R, Pritchard J K, Shen P, Oefner P, Feldman M W 2000 Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences, USA* **97**: 7360–5
Varki A 2000 A chimpanzee genome project is a biomedical imperative. *Genome Research* **10**: 1065–70
Wayne R K, Leonard J A, Cooper A 1999 Full of sound and fury: The recent history of ancient DNA. *Annual Reviews of Ecology and Systematics* **30**: 457–77
Weiss K M, Buchanan A V 2000 Rediscovering Darwin after a Darwinian century. *Evolutionary Anthropology* **9**: 187–200
Wright S 1931 Evolution in Mendelian Populations. *Genetics* **16**: 97–159

J. L. Mountain

Human Genome Diversity and its Value

1. Introduction

The era of modern genetics began with the rediscovery of Mendel's laws in 1900. However, long before this rediscovery, Charles Darwin (1859) recognized that variation in characteristics among individuals provided the basis of origin of species. Evolutionary biologists and population geneticists have long been interested in studying such variations, especially those that have genetic bases, and their origins and maintenance. Genomic diversity of a population is the raw material for its evolution by natural selection. Genomic diversity of a population can be measured in several ways. The most commonly used measure is the probability that two genes at any particular position of the chromosomes chosen randomly from the population are different. Of course, it is not practically possible to choose genes, since at any position of the chromosomes, two genes remain in combination in an individual. The above probability is equivalent to the probability that if an individual is drawn at random, this individual will possess a combination of two different genes at the chromosomal location under consideration. The different gene variants at any chromosomal location (locus) are called alleles. The combination of two alleles at a locus in an individual is called the genotype of the individual. When the genotype comprises the same two identical alleles, it is said to be homozygous; otherwise, heterozygous. The higher this probability, the greater is the genomic diversity. When all individuals of a population have identical genes, this probability attains the value of zero indicating that there is no genomic diversity in the population at that locus.

1.1 The Beginning of Human Genome Diversity Studies

The fact that individuals can differ in inherited traits was first clearly established with the identification of the ABO blood group system by Landsteiner in 1900. Soon after World War I, the Hirschfelds (1919) first discovered that human populations differ in frequencies of ABO blood group types. This finding marked the beginning of human genome diversity

studies. Subsequently, the development of biochemical- and molecular-genetic techniques have provided greater understanding and more refined estimates of human genome diversity.

1.2 The Importance of Genetic Diversity

It is crucial for any population to be genetically diverse. Humans, like many other species, are susceptible to various infections, many life-threatening. There is an increasing realization from recent research that genetic factors are involved in conferring susceptibility/resistance to most such infections. If all individuals in a population were to be genetically the same, then in the event of an outbreak of a new infection, it is possible for the entire population to be easily wiped out by the infectious agent if no individual carries genes conferring resistance to this agent.

2. Human Genomic Diversity at Various Levels

The development of starch gel electrophoresis by Oliver Smithies in 1955, permitted genomic diversity to be estimated at the level of proteins. Electrophoretic surveys of a variety of serum proteins and red-cell enzymes indicated extensive polymorphism among humans (Harris 1966). He showed that about 30 percent of human proteins are polymorphic, however, the genomic diversity estimated on the basis of such polymorphisms averaged between 10 percent and 20 percent. What this means is that at loci that determine various proteins and enzymes, there is more than one allele present in roughly 30 percent of such loci, however, at these loci, only about 10–20 percent of individuals are heterozygous on the average. Since electrophoresis can only detect such genomic variation that result in electrical charge change, and since only about 25 percent of genomic changes result in charge changes of the corresponding protein molecules (Nei 1975), this technique can only detect a small fraction of genomic changes. Hence, estimates based on data of such polymorphisms necessarily provide an underestimate of the true genomic diversity. There are, however, large differences in estimated diversities using different human proteins. These differences are, in part, due to differences in the degree of functional constraint of various proteins (Nei 1987). Not all proteins are functionally equally important. Changes in genes determining such proteins that are functionally important are generally more detrimental to the individual than changes in genes determining functionally less important proteins. Random changes, known as mutations, in genes create new alleles and enhance genomic diversity. Since mutations in genes of functionally important proteins are more detrimental, the new alleles created by such mutations are quickly removed from the population by natural selection. Hence, genomic diversity is related to functional constraint.

The development of the Southern blotting technique in 1975, allowed polymorphisms to be studied directly at the DNA level. The first DNA variants using this technique were detected in the late 1970s as restriction fragment length polymorphisms (RFLPs). Such length polymorphisms are due to cleavage or noncleavage of DNA at particular sites, commonly caused by single nucleotide polymorphisms (SNPs), that are detectable using specific restriction endonucleases. RFLP data also showed significant genomic diversity in human populations. Nucleotide sequencing, however, is the most direct method to investigate genomic changes. While systematic nucleotide sequencing studies involving large segments of the human genome and a large number of individuals are just beginning, from the limited data that are already available it has been estimated that the average level of diversity per nucleotide varies between 0.0001–0.005 (see Chakravarti 1999 for a summary of recent results). Thus, in every 10,000 nucleotide positions along the human chromosomes, between 1 and 50 positions are estimated to differ between any two copies of that chromosome. This remarkable 50-fold difference in nucleotide diversity is again in part due to the fact not all regions of the genome are functionally equally important. Accumulation of mutations is tolerated in the functionally less important parts of the genome, but strongly selected against in the more important parts of the genome.

3. The Value of Human Genome Diversity in Understanding Human Evolution

With the passage of time, a population splits into subpopulations because of various cultural, economic, demographic, and ecological reasons. For example, after a period of demographic expansion of a population, subgroups of individuals may be compelled to move to another locality because of depletion of resources in the original habitat. Subsequent to this move, the original group and the newly founded subgroup may cease to have any genetic contact that is realized through exchange of mates. Generally, such subgroups are founded by a limited number of individuals, which impacts upon the genomic diversity of the subpopulation. When several such subpopulations are formed and remain isolated from one another so that individuals of one subpopulation do not exchange mates with other subpopulations, genomic diversification of the subpopulations is hastened. (Fusion of populations also occurs, but is possibly much less common that fission, especially in expanding populations.) Admixture, or the exchange of genes, increases the genomic similarity between two

populations. The primary barriers to admixture are cultural and linguistic differences and geographical distance. Thus, one expects genes, cultures, and languages to evolve in tandem. Indeed, genomic diversity depends on demographic size and history of a population. Recent advances in molecular and statistical techniques have shown, for example, that pairwise comparisons of sequence differences among people within a population can reveal past demographic patterns of the population (Slatkin and Hudson 1991, Rogers and Harpending 1992, Thomson et al. 2000).

3.1 Some Important Evolutionary Inferences

The above facts have been used to make important inferences on human evolution. By examination of worldwide human genetic variation, Cavalli-Sforza et al. (1994) have found a major concordance between language similarities and genetic affinities. Their studies have also revealed that the spread of agriculture was by human migration, rather than through the spread of customs and knowledge. The most interesting, and perhaps the most controversial, inference derived from genome diversity data has been that all anatomically modern humans evolved in a single geographical region, Africa. If one considers a segment of the genome on which the possibility of any strong effect of natural selection may be negligible, and if by comparing the nucleotide sequences of this segment in two individuals, one finds them to differ by a certain number of nucleotides, then it is possible to easily estimate the date of their common ancestor if the mutation rate (the rate at which nucleotide changes take place per year) is known and can be assumed to be constant over time. Based on nucleotide sequence data and several other lines of genetic evidence, it has been inferred that (a) modern humans are a relatively young population, (b) Africans are genetically the most diverse group of humans, (c) the largest genetic divergence among Old World populations is between Africans and non-Africans, and (d) modern humans arose in Africa between 200,000 and 100,000 years ago and dispersed throughout the Old World to replace archaic human species completely (Stoneking 1993, Nei 1995), although some geneticists (Ayala 1995) do not believe in complete replacement or even in the single-origin scenario for anatomically modern humans (Wolpoff 1999, Relethford 1995).

3.2 Human Genome Diversity Studies Can Answer Some Central Biological Questions

The nature and extent of genome diversity depend on various evolutionary forces, such as, mutation, natural selection, random genetic drift, admixture, and recombination. The relative importance of the various evolutionary forces in determining and maintaining diversities within and between human populations remains poorly understood. However, from past studies of human populations with diverse structures, it is clear that random genetic drift (genetic events dictated primarily by chance that result from the small size of local groups) and the flow of genes from one population to another through migration or mate exchange are responsible for much of the patterning within human populations. No reliable estimate exists of the fraction of the human genome which is under the direct effect of natural selection. Put differently, how much of the observed variation in the human genome has a direct bearing on the fitness of individuals is largely unknown.

With the clearer understanding of the features of the human genome that is now emerging, thanks to the Human Genome Project, it will soon become possible to answer some of these central biological questions. For this, it will be necessary to carefully partition the genome into different equivalence classes of segments based on particular features, such as function, tissue specificity, gene density, distance from chromosome centromere or telomeres, etc., and then estimating within- and between-population diversities separately for each of these classes.

4. Human Genome Diversity and Molecular Medicine

The practice of medicine today focuses on 'treatment after disease onset'; the promise of molecular medicine is 'prevention before disease onset.' The hope for this promise is that genetic bases for most diseases with be discovered as a result of the spectacular advances in molecular genetics. Genes for diseases will be mapped on chromosomes; mutations precipitating clinical conditions will be identified. These discoveries will permit delineation of genotypes that are predisposed to specific diseases. Knowledge of the genotype of an individual will allow prediction of risk to disease. Preventive measures, that will not only include lifestyle modification but also genotype-specific therapy, will be undertaken for high-risk individuals.

4.1 Genome Diversity Studies Can Lead to Gene Discovery

Genomic diversity data on human populations can be used successfully in localizing on chromosomes the genes that cause inherited disorders, identifying predisposing/susceptibility factors to various diseases, and estimating relative roles of genetic and environmental factors in disease causation. In a large population, alleles at a pair of loci generally combine randomly into possibly genetic types. Thus, in a large population, there is usually no association between loci, unless they are physically very close on the same chromosome. However, in small isolated populations

or in recently admixed populations, strong associations between alleles at different loci are expected. Put differently, in such cases genetic loci behave non-independently in a statistical sense. This phenomenon is known as linkage disequilibrium. Disequilibrium decays every generation at a rate dependent on how far apart the genes are on the chromosome, which is measured by the rate of recombination between the genes. By examining patterns of linkage disequilibrium between diseases and genomic markers, especially in small isolated populations, it is possible to map loci controlling diseases. For example, in recently admixed populations, disequilibrium can be observed over chromosomal areas as large as 5–10 cM, which is about 2–5 percent of the estimated linkage length of an average human chromosome (Morton 1991). These estimates, in turn, help identify mutations that cause various genetic disorders. Thus, genome diversity studies can lead to gene discovery.

Variations at the DNA level among individuals that occur, on an average, once in every 500–1000 base pairs (bp) along the 3 billion bp human genome can result in a wide range of susceptibility or resistance to diseases. Polymorphisms in different regions of the genome can interact among themselves and/or with the environment in conferring susceptibility/resistance. By DNA sequencing of appropriate genomic regions (e.g., those that are involved in detoxification pathway, signal transduction, etc.) in individuals exposed to various levels of pathogens and environmental toxins and correlating the information thus obtained with clinical outcomes, it is potentially possible to identify specific alleles as well as environmental exposures that precipitate disease. Genome diversity research will enable more precise identification of the interaction of genomic polymorphisms with environmental agents in disease causation. Such research can lead to implementation of public health programs for protecting susceptible populations and for targeting genetic screening to groups at greater risk to disease. Indeed, the focus is shifting from gene discovery to gene diversity.

4.2 *Risk Assessment via Genome Diversity*

One of the primary values of genome diversity research is that it will help define genetic polymorphisms and gene products that can predict development, progression, and severity of disease. While for rare diseases, single mutations with large effects are generally involved in the precipitation of disease, for common diseases, perhaps common variations in several genes are involved. These common variations singly do not significantly enhance the risk to a common disease, but jointly precipitate the disease condition. So the estimation of frequencies and effects of common genetic polymorphisms, particularly single nucleotide polymorphisms (SNPs), in defined populations, are expected to play an integral role in molecular medicine. These estimates may, one day, enable high-risk patients to be diagnosed at a very early stage of disease and be treated with effective therapies that reverse the disease process or even prevent disease from occurring. Drugs can be targeted specifically to those genotypically defined patients who will respond well to the drug with no adverse side effects.

5. *Debunking the Myth of Race*

Humankind is a reproductive community, all members of which are connected by ties of mating and parentage. Yet, any two individuals, identical twins excepted, carry different sets of genes. To systematically study this enormous genomic diversity, biologists and anthropologists have described and named complexes of organisms to identify for themselves, and to let others know, what they are talking and writing about. Thus, classifying and giving names to classes began as a convenience-tool of science. Human biologists too have named classes of populations as 'races.' For example, Garn (1965) recognized nine 'geographical races' and 32 'local races' of man. Most such racial classifications were based on visible characteristics, such as skin color, nose height, etc. Such characters are generally adaptive to environmental conditions. Further, such characters also influence mate choice. Therefore, racial differences that are recognized by such characters are really the products of natural and sexual selection. Darwin (1871) noted 'Although the existing races of man differ in many respects … yet if their whole organization be taken into consideration they are found to resemble each other closely in a multitude of points.' Studies on human genome diversity have gradually made the ineptitude of typological approaches evident to increasing majorities of anthropologists. Yet, racial classifications are often used in sociopolitical debates in which certain races are claimed to be 'genetically superior' to others. Past studies on genetic diversity have shown that there is usually much greater intra-populational genetic variation than interpopulational variation (Lewontin 1972, Nei 1975, Barbujani et al. 1997). This means that a pair of individuals from a so-called 'genetically superior race' may be genetically more different than when one of them is compared with another from a 'genetically inferior race.' This is clearly in contradiction to the notion of genetic 'superiority.' Indeed, genome diversity studies have provided further evidence that has helped debunk the myth of races. Through their analyses of data compiled from genetic studies in global populations, Cavalli-Sforza et al. (1994) have shown that frequencies of genes often form geographical gradients, which make the task of drawing lines on a map to divide the regions of different races nearly impossible. Thus, race boundaries are never sharp. Worse still, gradients of frequencies of different genes often criss-cross on a

geographical map. When boundaries of gene frequency gradients are steep, they often correspond with linguistic boundaries or other similar cultural barriers to gene mixing. Thus, observed genetic differences among populations have no relation with genetic 'superiority' or 'inferiority.' Such differences are primarily the results of natural selection, barriers to genetic intermixture and also chance effects (random genetic drift). Human genome diversity studies have been valuable in debunking the myth of race.

See also: Evolution, History of; Evolutionary Social Psychology; Human Evolutionary Genetics; Human Genome Project: German Perspective; Human Genome Project: Japanese Perspective

Bibliography

Ayala F J 1995 The myth of Eve: Molecular biology and human origins. *Science* **270**: 1930–6

Barbujani G, Magagni A, Minch E, Cavalli-Sforza L L 1997 An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences USA* **94**: 4516–9

Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

Chakravarti A 1999 Population genetics—making sense out of sequence. *Nature Genetics Supplement* **21**: 56–9

Darwin C 1859 *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Murray, London

Darwin C 1871 *The Descent of Man, and Selection in Relation to Sex*. John Murray, London

Garn S M 1965 *Human Races*. Charles Thomas, Springfield

Harris H 1966 Enzyme polymorphisms in man. *Proceedings of the Royal Society, London B* **164**: 298–310

Hirschfeld L, Hirschfeld H 1919 Serological differences between the blood of different races. *Lancet* **ii**: 675–8

Lewontin R C 1972 The apportionment of human diversity. *Evolutionary Biology* **6**: 381–98

Morton N E 1991 Parameters of the human genome. *Proceedings of the National Academy of Sciences USA* **88**: 7474–6

Nei M 1975 *Molecular Population Genetics and Evolution*. North Holland, Amsterdam

Nei M 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York

Nei M 1995 Genetic support for the out-of-Africa theory of human evolution. *Proceedings of the National Academy of Sciences USA* **92**: 6720–2

Relethford J H 1995 Genetics and modern human origins. *Evolutionary Anthropology* **4**: 53–63

Rogers A, Harpending H C 1992 Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* **9**: 552–69

Slatkin M, Hudson R R 1991 Pairwise comparisions of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555–62

Stoneking M 1993 DNA and human evolution. *Evolutionary Anthropology* **2**: 60–73

Thomson R, Pritchard J K, Shen P, Oefner P J, Feldman M W 2000 Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proeedings of the National Academy of Sciences USA* **97**: 7360–5

Wolpoff M H 1999 The sytematics of *Homo*. *Science* **284**: 1774–5

P. P. Majumder

Human Genome Project: German Perspective

1. The Human Genome Project

The Human Genome Project differs from any previous biological or medical project in size and cost. Its ambitious goal is the deciphering (sequencing) of all 3 billion building blocks of our genetic make-up—the so-called DNA—by 2005, the identification of all genes encoded in that DNA, and the understanding of the role of these genes in health and disease. The knowledge of these genes and their function is crucial for basic biological research as well as for the improvement of prevention, diagnostics, and therapy of disease. It is the prerequisite for a targeted design of pharmaceuticals and for novel approaches like gene therapy. This knowledge poses hope to millions of affected people and contains also an immense economic potential.

The Human Genome Project was started in the USA in 1990, James Watson, the codiscoverer of the DNA structure being its first co-ordinator. It was clear from the beginning, due to the estimated cost of US $ 3 billion and the immense amount of work involved, that the Human Genome Project had to include many countries. The Human Genome Organization (HUGO), an independent international organization of genome scientists, was established to co-ordinate the duties. The US Human Genome Project started off with considerable public funding (US $ 87 million in 1990) and was soon followed by the UK and France. Between 1991 and 1996 France contributed a comprehensive genetic map of the human genome. It is noteworthy that this work was predominantly financed by private money from a patients' association. The British Wellcome Trust for its part set up the world's largest sequencing facility, the Sanger Centre. Smaller initiatives later emerged, for instance in Japan and Canada. No such activity, however, was seen in Germany until 1995. By 2001 a nearly complete 'working draft' of the human genome has been presented by the publicly funded Human Genome Project, including a significant German contribution. The complete sequence will be available in public databases some time ahead of schedule, probably by 2003. In 1998 an emerging competition to the Human Genome Project by private companies, namely by

Craig Venter's Celera, had speeded up the deciphering of the human genetic code tremendously. But the knowledge of our comprehensive genetic make-up is not considered as a benefit for mankind by everybody. Profound ethical issues are raised by the possibilities inherent in this knowledge. This had been realized by the founders of the project and therefore, about 3 percent of the budget was dedicated to exploring the ethical, legal, and social implications of the Human Genome Project. In most countries participating in the Human Genome Project, an extensive discussion took place about the opportunities and risks attached to these novel technologies. In Germany a parliamentary committee was established to elucidate the topic. But the discussion was much more polarized in Germany than in other countries.

2. *Cutting Edge Technology is Considered a Necessity for Germany's Industrialized Society*

Germany is a strongly industrialized country. Due to its lack of raw materials, its economy is heavily dependent on export-oriented manufacturing industries. To reach and maintain its high standard of living, and to survive on the world market, Germany is in permanent need of cutting-edge technologies. Traditionally Germany's strength lay in the engineering and chemical industries. Other fields like agriculture and coal mining have undergone deep structural crises and are nowadays only viable on the basis of high subsidies. This is particularly true for rural areas or areas that have been focused on one industrial branch for generations. German labor law and the strong position of German trade unions protect employees from unemployment and provide broad social protection, but on the other hand slow down structural changes. Revolutionary technical developments can give a well-developed industry an opportunity to solve structural problems and to gain new markets. During the 1970s and 1980s, progress in microelectronics and computer technology constituted such a revolution. But neither industry nor academia realized this development in time so that it was subsequently missed in Germany. During this time, future prospects for highly qualified young scientists became increasingly worse. Germany had to face a severe 'brain drain,' predominantly to the United States. The structural problems dramatically deteriorated following German reunification—the entry of the Five New *Länder* (states) of the former German Democratic Republic to the Federal Republic of Germany, which became known as 'German Disease' in the 1990s. The economic recovery—the so-called *Wirtschaftswunder*—of West Germany after World War II, often promised to East Germans after reunification by West German politicians, was not easy to repeat. The resulting increasing economic pressure led to a resumption of the debate about the pros and cons of biotechnology in general. After the end of the Cold War, other ethically controversial economic sectors such as the arms industry became less significant. Up to this time there had been no co-ordinated funding for biotechnology research, let alone any participation in the Human Genome Project, although the scientific basis was established and many technical developments pointed the way ahead. Though among a few scientists discussions on the need for genome-wide analysis had started in Germany around 1980, leading up to a proposal to sequence the entire genome of the bacteria Escherichia coli or even human chromosome 21 at the European Molecular Biology Laboratory in Heidelberg where much of the early technology had been developed, both the German scientific community, as well as funding agencies were at this point by and large not able to recognize the immense potential of genome-wide analysis.

While a large number of new powerful biotechnology companies emerged in the United States during the 1980s, recombinant pharmaceuticals had to be imported to Germany. The first version of the German Genetic Engineering Law was very restrictive and bureaucratic. It took the Frankfurt-based company Hoechst (now Aventis) years to get the approval to produce recombinant human insulin in their brand-new facilities. As a consequence, most large German pharmaceutical companies transferred their research departments to locations in the United States. Politicians realized very late the immense economic damage caused by this development. To take countermeasures two large co-ordinated research programs were started by the Federal Ministry of Research—the German Human Genome Project (DHGP) in 1995 and the BioRegio Competition in 1996.

3. *Genomic Research Against the Background of German History*

To fully understand the assessment of and the reservation about genetics in German society, it is helpful to take a closer look not only at economic aspects but also at intellectual history and the National Socialist period.

3.1 *The Idea of Eugenics*

At the beginning of the twentieth century various political and intellectual movements in search of a new order existed in Germany. As in many other countries at that time many of these were characterized by strong nationalism and/or antisemitism. Some of them had their roots in German romanticism or naturalism: movements like the *Wandervögel*—a ramblers' association and precursor of a new German youth movement expressing a strong love of nature.

This milieu, in combination with vague Darwinian ideas, ushered in the idea of improvement of the human race and prevention of its degeneration under the overwhelming influences of a modern and industrialized world: the idea of eugenics originally introduced by Galton in the 1870s. Many of these movements were infiltrated by National Socialists and forced to take the party line during the *Gleichschaltung* after the Nazis had seized power in 1933. Fragments of these simplistic ideas were soon adopted by the Nazis. When the philosopher Nietzsche talked about the 'overman,' a superior human, it was interpreted that an inferior human was the conclusive consequence. Eugenics was not a purely German affair. Many countries, including the United States, had a strong eugenics movement during that time. In none of these countries, however, was eugenics used as a justification for profound discrimination, leading to the Nuremberg Laws of 1935 and eventually the extermination of some 6 million Jews by 1945.

3.2 The Role of German Physicians and Geneticists During the National Socialist Period

The role of German physicians and geneticists during the Nazi period is complex and exceeds by far the scope of this article. It should be sufficient to illustrate their influence by some examples. After the rediscovery of Mendel's Rules many scientists were worried that mental diseases like schizophrenia or idiocy, or the mixing of the white 'race' with individuals considered to be inferior (e.g., members of ethnic groups from the former colonies) would lead to a decline of European culture. Already during the first two decades of the twentieth century, the superiority of the white race to other humans and the destruction of so-called 'life unworthy to be lived' was discussed openly even in international scientific circles. Eugen Fischer, a scientist at the University of Freiburg, published a book in 1913 in which he addressed 'the problem' of mixed marriages between white men and Hottentot women and considered the progeny of such marriages as nonviable without protection by a superior race. Many of the racist ideas of Fischer and others were ingested in the emerging National Socialist ideology at that time. In 1927 Fischer became the director of the newly founded Kaiser Wilhelm Institute (KWI) for Anthropology, Human Heredity, and Eugenics in Berlin; in 1939 he was elected president of the University of Berlin. In this latter position he was responsible for signing the dismissal notices of his Jewish colleagues. A motivation for most of the scientists was an almost paranoid anxiety of the decline of their own culture and 'race.' In many cases, personal advantages that derived from the dismissal of their Jewish colleagues might have encouraged this development. Scientists supporting the National Socialist idea of racial hygiene like Fischer, Rüdin of the KWI of Psychiatry, Hallervorden of the KWI of Brain Research, or von Verschuer, Fischer's successor as the director of the KWI for Anthropology, were welcomed by the leaders of the Nazi party. Most of the scientists involved, however, considered themselves as being apolitical. From their point of view they supported the system for the sake of science. Eugen Fischer wrote in a German newspaper: 'It is a rare and special good fortune for theoretical science to flourish at a time when the prevailing ideology welcomes it and its findings can immediately serve the policy and the state' (*Deutsche Allgemeine Zeitung*, March 28, 1943). Since most of these scientists did not have an official post in the Nazi administration and were not actively involved in the killing of people, only a few had to flee from Germany after the end of World War II. Josef Mengele, known as the 'Angel of Death' of Auschwitz, managed to escape to Brazil where he lived unmolested probably until 1979. Those who had deliberately worked with human experimental material like eyes, brains, etc., gained from concentration camps or psychiatric hospitals or had asked for this material from their coworkers, mostly avoided prosecution.

3.3 Germany Dealing with its Difficult Past

After World War II the faith of society in a detached and neutral science was profoundly disturbed. Genetics came under special suspicion because its advocates had provided 'scientific' basis for the National Socialist racial hatred. Scientific terminology had been adopted and occupied by the Nazis. As the Nobel Prize laureate Christiane Nüsslein-Volhard puts it quoting the term 'selection,' 'Scientists think of Darwin's fundamental theory of the origin of species while German society even today is reminded of the gas chambers of Auschwitz.' The reasons for this are diverse and many. Most geneticists did not differentiate themselves properly from the National Socialists. The borderline between fanatic followers and murderers like Josef Mengele, opportunists who profited from the system like Eugen Fischer and Otmar von Verschuer (who was Mengele's superior in Berlin), and those who had remained in their positions with tacit permission, remained hazy. After 1945 no genuine reappraisal of the role of physicians and geneticists during the Nazi period occurred. Directly after the war, Germany was too busy working on its economic recovery to ask questions about guilt. Many members of the intellectual elite had either emigrated or had been killed during the Nazi period. But many key proponents of the genocide reappeared in important academic positions. Otmar von Verschuer became professor of genetics at the University of Münster where in 1954 he even rose to the position of Dean of the Medical Faculty. The Kaiser Wilhelm Society was renamed Max Planck Society (MPS), and Adolf Butenandt, the Nobel Prize laureate of 1939 who was sometimes criticized for not differentiating himself clearly from the Nazi regime, became its first president.

Many of the experiments that were committed in Auschwitz and elsewhere were financially supported by the Kaiser Wilhelm Society or the Deutsche Forschungsgemeinschaft (DFG), the national German Research Council. It was, however, not before the late 1990s that DFG and MPS initiated an independent review of their own role during the NS period.

Because many natural sciences had fallen into disrepute after 1945, public discussion was completely dominated by intellectuals of the social sciences. This was in sharp contrast to the period from the turn of the century until 1933 when, for instance, physicists like Einstein, Heisenberg, and Planck were actively involved in the social and political debate, and disputes among artists, philosophers, theologians, and scientists were common. After 1945, genetics had not become common knowledge in Germany even in intellectual circles. Instead, a vague fear of genetic reductionism and determinism remained. This led Robert N. Proctor, a US historian of science, to state: 'The legacy of Nazism influences the attitudes many Germans show towards genetics today and is probably responsible for the absence of a prominent science-oriented intellectual elite. Where are the Goulds, Hawkings and Dawkins of Germany?' The polarization between social scientists and natural scientists who were mainly seen as technocrats still persists. The seizure of power by man over himself created the fear that humanity could be the first victim.

3.4 The German Human Genome Project

During the 1980s the tremendous successes of genetic engineering boosted the biotechnology industry in the USA and provided the prerequisite for the idea of the Human Genome Project. While in the US, companies like Amgen turned over more than US $ 1 billion per year, the production of recombinant pharmaceuticals was still hampered in Germany. At the same time, recombinant pharmaceuticals and vaccines had to be imported into Germany. One of the first dedicated centers for genome research in Germany, the *Genzentrum* in Munich had bulletproof windows installed in fear of an attack. The international Human Genome Project was started in 1990 without any German participation. But increasing economic pressure forced Germany to a rethinking of genome research. After a thorough investigation of the issue by a parliamentary commission, two programs of the Christian Democratic-Liberal coalition started in the mid-nineties. The BioRegio Competition aimed at exploiting the synergy of self-organization of whole regions as centers of excellence for biotechnology and was unexpectedly successful. The German Human Genome Project (DHGP), integrated into the international effort to decipher the human genome, was thematically much more focused. It was strongly industry-oriented and a special infrastructure was to guarantee the effective transfer of the knowledge gained to industry. Initial efforts to raise half of the budget from the German pharmaceutical industry failed. Subsequently, the whole budget of DM 40 million per year was granted by the Federal Ministry of Education and Research. To safeguard the interests of industry, a Patent and Licensing Agency (PLA) was established whose task was to screen the results of the projects granted within the DHGP and to guarantee a patent application if applicable. The patent is granted to the inventor at no cost. In turn, the Industry Association gets the right of first refusal on the patent.

The agreement also gave the Industry Association the right to screen the data produced in the German DNA sequencing centers—a clear infringement of the so-called Bermuda Principles—an agreement between all academic sequencing centers worldwide making all data immediately publicly available without any restrictions. The Industry Association waived their claims only after massive pressure from the international scientific community.

The DHGP has had a profound effect on the German biotechnological industry. Contrary to the expectation that the big pharmaceutical companies would exploit the scientific results, licensing of patents was sparse. Instead, many researchers preferred to start their own companies, a tendency up to then only rarely found in German academia. The success of the DHGP initiative led to Germany exceeding the United Kingdom in the number of biotech companies by the year 2000.

Initial concerns among scientists that funding of genome research would be cut after the change of government to a coalition of Social Democrats and the Green Party were groundless. Not only has the DHGP been maintained, but also a plant genome project—initiated by the prior administration—as well as a microorganism genome project were started, assisted by additional projects on bioinformatics and proteomics. In 2000, a long-expected concept document was released by the Federal Ministry of Education and Research. This paper draws up a long-term perspective for genomic research in Germany and safeguards the existence of central infrastructure units (resource centers) for genomic research. At the same time, boosted by the announcement of the draft version of the human genome, a tremendous increase in funding was approved that eventually brought Germany in line with genomic funding in countries like Japan, UK, and France.

The economic success as well as the prospects of new diagnostic and therapeutic possibilities has very much changed the German attitude towards human genetics. Medical genetics is now much more accepted than it was a few years ago. In contrast, due to concerns about health as well as to the aggressive marketing strategies of certain agrocultural biotech companies, plant genetic engineering has met with

deep disapproval in Germany. However, the fatalistic attitude in Germany towards fundamental natural sciences has somewhat changed. The former lack of discussion in society that permitted simplistic arguments to gain predominance has been overcome. The majority of the German intellectual elite has realized that an informed discussion within society is needed to solve the problems stimulated by genome sciences.

See also: Ethical Issues in the 'New' Genetics; Eugenics, History of; Eugenics in Europe; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease-related Characteristics; Genetic Testing and Counseling: Educational and Professional Aspects; Human Evolutionary Genetics; Human Genome Diversity and its Value; Human Genome Project: Japanese Perspective

Bibliography

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J 1996 A comprehensive genetic map of the Human Genome based on 5,264 microsatellites. *Nature* **380**: 152–4

Ernst & Young 2000 *Gründerzeit; Zweiter Deutscher Biotechnologie-Report 2000*. Ernst & Young, Stuttgart, Germany

Hammerstein N 1999 *Die Deutsche Forschungsgemeinschaft in der Weimarer Republik und im Dritten Reich; Wissenschaftpolitik in Republik und Diktatur*. Beck, Munich, Germany

International Human Genome Sequencing Consortium 2001 Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921

Kaufmann D (ed.) 2000 *Geschichte der Kaiser-Wilhelm-Gesellschaft im Nationalsozialismus; Bestandsaufnahme und Perspektiven der Forschung*. Wallstein-Verlag, Göttingen, Germany

Klee E 1997 *Auschwitz, die NS-Medizin und ihre Opfer*. Fischer Verlag, Frankfurt, Germany

Laubichler M D 1999 Frankenstein in the land of Dichter and Denker. *Science* **286**: 1859–60

Müller-Hill B 1998 *Murderous Science: Elimination by Scientific Selection of Jews, Gypsies, and Others in Germany, 1933–1945*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Proctor R N 2000 *Adolf Butenandt (1903–1995). Nobelpreisträger, Nationalsozialist und MPG-Präsident; Ein erster Blick in den Nachlaß. Ergebnisse 2*. Präsidentenkomission (Max Planck Society) Geschichte der Kaiser-Wilhelm-Gesellschaft im Nationalsozialismus, Berlin

Sachse C, Massin B 2000 *Biowissenschaftliche Forschung in Kaiser-Wilhelm-Instituten und die Verbrechen des NS-Regimes; Informationen über den gegenwärtigen Wissensstand. Ergebnisse 3*. Präsidentenkomission (Max Planck Society) Geschichte der Kaiser-Wilhelm-Gesellschaft im Nationalsozialismus, Berlin

Schiermeier Q 1997 German truce on genome data access. *Nature* **388**: 409

Schmuhl H-W 2000 *Hirnforschung und Krankenmord; Das Kaiser-Wilhelm-Institut für Hirnforschung 1937–1945. Ergebnisse 1*. Präsidentenkomission (Max Planck Society) Geschichte der Kaiser-Wilhelm-Gesellschaft im Nationalsozialismus, Berlin

Unterhuber R 1997 Germany gets warning on access to sequence data. *Nature* **387**: 111

1997 Germany may relax genome data policy. *Science* **276**: 1323

2000 Human Genome Project makes its mark in Germany. *Nature* **406**: 821

J. Maurer and H. Lehrach

Human Genome Project: Japanese Perspective

The Japanese Human Genome Project (JHGP) was initiated in 1988. The project experienced some difficulties in the initial stages, such as scarcity of resources and the perception that genome research was not hard science. As a result the JHGP was unable to follow the specific aims of sequencing DNA of the human genome. This article explains how the JHGP adopted strategies to overcome these problems and how it became an effective contributor to the international human genome project.

1. History of the Human Genome Project in Japan

The drive toward an international approach to the Human Genome Project began in 1985. In 1988, as a response to this development and the progress of the human genome project activities in the United States and Europe (see *Human Genome Project: German Perspective*), the Japanese human genome project (JHGP) conducted a feasibility study. The study was supported by a general research grant from the Ministry of Education, Science, Sports and Culture (ESSC: MONBUSHO). As a result of this study, the research council of MONBUSHO proposed the formulation of a special human genome research grant and a five-year plan for the promotion of genomic research.

The genome research supported by this special grant had a project-like characteristic, and because of this was confronted by criticism from various sources. There was a strong belief among certain groups that this kind of project was not 'hard research' but rather mere routine work. Moreover, difficulties were experienced in relation to the scarcity of technicians skillful in this area who could support the genome researchers and also in the reluctance shown by the government and private institutions to invest heavily in a single project. As a result of these difficulties

Table 1
Mass DNA data submission to DDBJ (submitted and distributed January 1998–December 1998)

Organism	Accession number (number of entries)	Submitter (institution)
BCT (bacteria)		
Escherichia coli	ABO11548-ABO11549 (2)	Makino, K. (Osaka University)
Unidentified alpha proteobacterium	ABO11682-ABO11791 (110)	Shikano, S. (Tohoku University)
Unidentified nitrogen-fixing bacteria	ABO11841-ABO11964 (124)	Ohkuma, M. (RIKEN)[a]
Bacillus stearothermophilus	ABO20064-ABO26165 (102) ABO20234-ABO20235 (2)	Matsuura, T. (Osaka University)
Plasmid colIb-p9	ABO21078 (1)	Sampei, G. (The University of Electro-communication)
HUM (human)		
Homo sapiens	ABO14511-ABO14610 (100) ABO18254-ABO18353 (100) ABO20626-ABO20725 (100)	Ohara, O. (Kazusa DNA Research Institute)
Homo sapiens	ABO19437-ABO19441 (5)	Matsuda, F. (Kyoto University Graduate School of Medicine)
Homo sapiens	ABO20858-ABO20878 (21)	Hirakawa, M. (JST)[b]
HTG (high throughput genomic sequencing)		
Arabidopsis thaliana	ABO10068-ABO10077 (10) ABO10692-ABO10700 (9) ABO11474-ABO11485 (12) ABO12239-ABO12248 (10) ABO13387-ABO13396 (10) ABO15468-ABO15479 (12) ABO16870-ABO16893 (24) ABO17059-ABO17071 (13) ABO18107-ABO18121 (15) ABO19223-ABO19236 (14) ABO20742-ABO20755 (14)	Nakamura, Y. (Kazusa DNA Research Institute)
VRL (viruses)		
Hepatitis G virus	ABO10155-ABO10202 (48)	Tanaka, Y. (Nagoya city University Medical School)
Human immunodeficiency virus type 1	ABO14775-ABO14874 (100)	Sato, H. (National Institute of Infectious Diseases)

a The Institute of Physical and Chemical Research Japan. b STA, Japan.

JHGP could not follow the specific aim of sequencing the DNA of the human genome.

Instead, JHGP adopted the strategy of developing excellent genome-related research. By this strategy, JHGP successfully contributed to the construction of human genome physical maps, the development of full-length cDNA construction techniques, and the making of cDNA expression maps, leading to an enhancement of human genome research in Japan. However, JHGP was far behind the United States and Europe, particularly in the field of genetic disease analysis based on human genome mapping. More specifically, as most of the major human genome projects were still in their infancy, JHGP was behind in terms of planning rather than the research itself.

Although, for Japan, this predicament was partly due to some researchers failing to recognize the importance of the human genome project, the main reason was the failure to establish a single human genome research center coordinated by the relevant ministries and institutions. For instance, JHGP was planned independently by two ministries and a government agency. These bodies included the MONBUSHO, comprising many University researchers, the Ministry of Health and Welfare (MHW) supervising medical activities, and the Science and Technology Agency (STA) supporting various large-scale projects. As a result of the lack of collaboration between these government organizations, talented researchers, useful materials, and research money were unnecessarily dispersed. Moreover, because of the slow response of bioindustries to the human genome project, government agencies were unable to formulate a collaborative liaison with them in Japan.

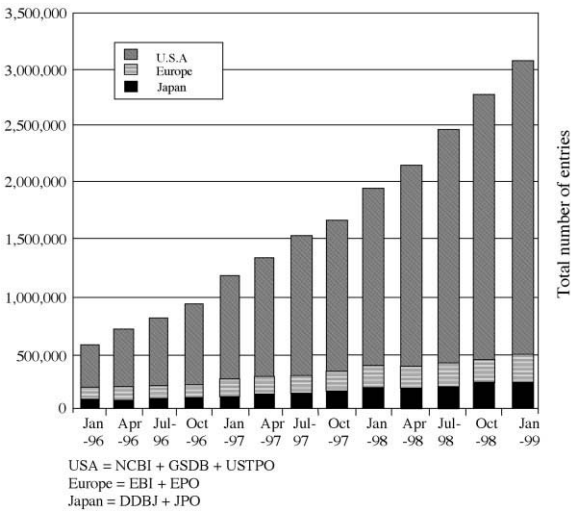

Figure 1
Contribution of the three geographical areas to the DDBJ/EMBL/Genbank International Nucleotides Database (January 1996–January 1999)

These problems were essentially based upon the political and social infrastructures within Japan; however, in spite of these difficult situations, several success stories emerged from JHGP. For example, seven research groups in Japan successfully contributed to the completion of the *Bacillus subtilis* (*B. subtilis*) genome sequence by determining the nucleotide sequences of its shared region of 1.3 Mb. The complete genome sequence of *Synechocystis sp.* was determined by the Kazusa DNA Research Institute, which was established by Chiba prefecture and industries. The Ministry of Agriculture, Forestry and Fisheries of Japan (MAFF) launched the rice cDNA project, which is currently leading international research on the rice genome. In 1998, the *Escherichia coli* (*E. coli*) genome project team in Japan successfully completed three quarters of the *E. coli* genome sequence, although the American team completed the whole *E. coli* sequence in the same year.

Genome Informatics was initiated from scratch because there were no major research activities which could act as a focal point, except the DNA databank of Japan (DDBJ) within the National Institute of Genetics (NIG). In particular, the recruitment of talented researchers, education and training of research students, installment of a computer network for genome research, and the establishment of a DNA database had to be done within a short period of time. Although these were successfully dealt with to some extent, there were still some problems to be solved such as intimate collaboration among the different institutions in charge of various bioinformatics activities.

2. *The Present Situation of the Human Genome Project in Japan*

As genome research advances internationally, there has been increasing recognition that genome research constitutes the necessary foundation for the future development of biology, medical research, and biotechnology. In particular, the necessity of establishing a large-scale genome-sequencing center has been accepted widely. However, the realization of this plan has been difficult because of budget limitations.

Since the basic law of science and technology development was passed in 1997, the budget limitations have, to some extent, disappeared under the national goal that Japan should become a country of creative science and technology. Because of this basic law, the government invested approximately $US 200 million in the genome research project in the 1998 fiscal year alone. Presently, various activities in JHGP are conducted under the initiative of governmental organizations and private industries. MONBUSHO successfully established the Center for Information Biology in NIG to support DDBJ and the newly established Human Genome Research Center in Tokyo University. The Ministry of International Trade and Industry (MITI) and the MHW coordinated a promotion business venture of genome-related enterprise and the STA created the genome science center in Yokohama. Moreover, large research grants were made to many genome research groups through various funding mechanisms. However, not all of these activities are along the lines of well-planned strategies of genome research.

The general focus of JHGP is toward the determination of large-scale genome sequences and a functional analysis of the genes that will be discovered by sequence determination. JHGP is placing more emphasis on the areas of genomic structure, function, and information. As for genome sequence determination, Kitazato University, Japan Marine Science and Technology Center, and STA and the Product Evaluation Center under MITI are pushing the genome sequencing of microorganisms by producing 122 Mb each year. In 1995, the Japanese Science and Technology Corporation initiated the human genome-sequencing project. The subgenomic segments such as HLA, immunogloblin variable, Down's syndrome, and oncogene-related regions, have been determined for about 15 Mb by Kitazato University, Cancer Institute, Keio University and Tokai University.

In total, the sequence production in Japan is about 5 Mb per year, which is one order of magnitude smaller than that of the Sanger Center in the United Kingdom and the Sequence Center at Washington University in the United States. Therefore, JHGP has great expectations of the role of the newly established Genome Science Center, which is planned to produce 30–50 Mb per year. Kazusa Research Institute is producing several Mb genome sequences of *Arabidopsis* per year.

Moreover, the genome project team supported by MAFF has a ten-year plan of sequence determination of the complete rice genome of 400 Mb, producing 20 Mb per year for the initial several years. In summary, the sequencing capacity of genomes in Japan has so far been 15–20 Mb per year. Judging from recent increases of the number of genome research groups focusing on bacteria, fungi, lower organisms, insects, and fish and the enhancement of genome research on humans and the mouse, 200 Mb sequence production per year is expected in the near future.

JHGP has a slight edge over other countries in the field of functional genomics. In particular, the Osaka University genome team was the first to analyze a cDNA expression profile of human genetic data (called 'body mapping') and other genome teams have quickly developed a full-length cDNA determination technique. Although these activities were not directly connected to a bioindustry as they are in the United States, MITI recently pushed the utilization of these technical advances to industry. Moreover, human genome diversity has become recognized as important for disease gene analysis and functional analysis. In particular, the so-called 'complex diseases' such as multi component diseases are becoming important targets of the human genome diversity project (see *Human Genome Diversity and its Value*). This project will require a lot of genetic data concerning the Japanese population, in particular a large amount of polymorphism data from several thousand people (see *Race: Genetic Aspects*).

Several preliminary research projects have already started in Japan. The cDNA project of *Caenorhabditis elegans* (*C. elegans*) genome has been conducted by NIG in cooperation with the Sanger Center and Washington University. The *B. subtilius* and *E. coli* genome teams have already begun functional analysis of all genes contained in their respective organisms. Many researchers are participating in these genome projects by conducting bioinformatics research and constructing databases. Thus, JHGP has resumed its original goals in a more integrated and coordinated way.

3. The Future Challenge of the Human Genome Project in Japan

In Japan, the aim of the human genome project was redirected to not only sequence the human genome but also to include the functional analysis of genes as well as the elucidation of protein tertiary structures. Although this redirection has successfully eased various criticisms, the main focus of the human genome project has been diffused to a considerable extent, leading to an ambiguity of the Japanese contribution to the international effort of human genome sequencing. Given the expectation that human genome sequencing has been achieved mainly by the United States and Europe, the Japanese human genome project will be successful if it orients its major effort towards the promotion of functional genomics, comparative genomics, evolutionary genomics, and genomic diversity analysis, along with the development of DNA chips or microarrays.

Conducting functional genomics requires analytic technologies that are capable of dealing with a large number of genes and proteins. Toward meeting these requirements, genome researchers in Japan have started contributing to the development of differential display, DNA chips and the identification of variant DNA. In the near future, these efforts should be connected to advancement of industrial technologies to produce unique and significant methodologies useful to functional genomics.

For genome informatic research, DDBJ at CIB/NIG, the Human Genome Research Center at Tokyo University, the Chemistry Institute at Kyoto University, and other research groups are making efforts to incorporate information science into genome research, in the belief that 'genome science is virtually information science.' There is no doubt that informatic aspects are extremely important to genome research. In particular, the elucidation of gene networks will be crucial for understanding the development and evolution of organisms, and in this sense, bioinformatics will play a crucial role. Another aim of JHGP is the allocation of talented bioinformatitions and the education of students who understand both biological and information sciences. This will be an important component of the future activities of the JHPG.

See also: Bioethics: Examples from the Life Sciences; Bioethics: Philosophical Aspects; Brain, Evolution of; Evolution, History of; Genes and Culture, Coevolution of; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Studies of Behavior: Methodology; Genetic Testing and Counseling: Educational and Professional Aspects; Human Evolutionary Genetics; Human Genome Project: German Perspective

Bibliography

Bellgard M, Itoh T, Watanabe H, Imanashi T, Gojobori T 1999 Dynamic evolution of genomes and the concept of genome space. *Annals of the New York Academy of Science* **870**: 293–300

Gojobori T 1999 Evolution. In: Creighton T E (ed.) *The Encyclopedia of Molecular Biology*. Wiley, New York

Gojobori T 1999 DNA sequence analysis. In: *Encyclopedia Life Sciences*. Macmillan Reference Ltd

Normile D 1997 Japan centers in on genome work. *Science* **278**: 1700–2

Sugawara H, Miyazaki S, Gojobori T, Tateno Y 1999 DNA Data Bank of Japan dealing with large-scale data submission. *Nucleic Acids Research* **27**: 25–8

T. Gojobori

Human Resource Management, Psychology of

Human resource management (HRM), as practiced in work organizations, includes five broad activities that together constitute the HRM system: staffing, retention, development, adjustment, and managing change (Cascio 1998). *Staffing* comprises the activities of (a) identifying work requirements within an organization, (b) determining the numbers of people and the skills mix necessary to do the work, and (c) recruiting, selecting, and promoting qualified candidates. *Retention* comprises the activities of (a) rewarding employees for performing their jobs effectively, (b) ensuring harmonious working relations between employees and managers, and (c) maintaining a safe, healthy work environment. *Development* is a function whose objective is to preserve and enhance employees' competence in their jobs by improving their knowledge, skills, abilities, and other characteristics; HR specialists use the term 'competencies' to refer to these items. *Adjustment* comprises activities intended to maintain compliance with the organization's HR policies (e.g., compensation, discipline), and business strategies (e.g., cost leadership). *Managing change* is an ongoing process whose objective is to enhance the ability of an organization to anticipate and respond to developments in its external and internal environments, and to enable employees at all levels to cope with the changes.

1. Responsibilities and Objectives

These activities can be carried out at the individual, work team, or larger organizational unit (e.g., department) level. Sometimes they are initiated by the organization (e.g., recruitment efforts or management development programs), and sometimes they are initiated by the individual or work team (e.g., voluntary retirement, safety improvements). It might appear that the activities of staffing, retention, development, adjustment, and managing change are the special responsibilities of the HR department. However, these responsibilities also lie within the core of every manager's job throughout any organization—and because line managers have authority (the organizationally granted right to influence the actions and behavior of the workers they manage), they have considerable impact on the ways workers actually behave. This implies two things: (a) a broad objective of HRM is to optimize the performance and productivity of all workers in an organization, and (b) a special objective of the HR department is to help line managers manage those workers more effectively. The HR department accomplishes this special objective through policy initiation and formulation, advice, service, and control in close communication and coordination with line managers.

Generally speaking, HR specialists provide the technical expertise in each area, while line managers (or, in some cases, self-directed work teams) use this expertise in order to manage people effectively. In a small business, however, line managers are responsible for both the technical and managerial aspects of HRM.

For example, in the area of retention, line managers are responsible for treating employees fairly, resolving conflicts, promoting teamwork, and providing pay increases based on merit. To help line managers do these things effectively, the HR department must devise a compensation and benefits system that employees will perceive as attractive and fair, establish merit increase guidelines that will apply across departments, and provide training and consultation to line managers on all employee relations issues, such as conflict resolution and team building.

2. The Evolution of Human Resource Management

To appreciate where the field of HRM is going, it is important to appreciate where it has come from. As it is practiced in most developed countries around the world, modern HRM has emerged from nine interrelated sources.

(a) Rapid technological change that increased the specialization of labor associated with the industrial revolution.

(b) The emergence of free collective bargaining, with constraints established for unions and for employers.

(c) The scientific management movement.

(d) Early industrial psychology.

(e) Governmental personnel practices growing out of the establishment of the Civil Service Commission in the USA.

(f) The emergence of personnel specialists and the grouping of these specialists into personnel departments.

(g) The human relations movement.

(h) The behavioral sciences.

(i) Social legislation and court decisions.

2.1 The Industrial Revolution

Three characteristics of the industrial revolution were the development of machinery, the linking of human power to the machines, and the establishment of

factories in which a large number of people were employed. The result was a tremendous increase in job specialization as well as in the amount of goods that workers could produce. 'Division of labor' became a rallying cry of this revolution.

Clearly, the industrial revolution greatly accelerated the development of business and commerce. Owners and entrepreneurs generally did quite well for themselves, but the average citizen fared poorly in comp-arison to today's workers in terms of purchasing power and working conditions. Labor was considered a commodity to be bought and sold, and the prevailing political philosophy of *laissez-faire* resulted in little action by governments to protect the lot of workers (French 1994).

2.2 The Emergence of Free Collective Bargaining

Because of the way they were abused, it was inevitable that workers would organize to protect themselves and to improve their lot in life. From the perspective of workers, the industrial revolution fostered specialization, along with the need for workers within each specialization to organize themselves against its abuses. Trade unions, also called 'labor unions,' spread rapidly, and so did the incidence of strikes (see also *Labor Unions*). For example, in 1886 employees at the McCormick Reaper Works in Chicago went on strike for an eight-hour work day (Cohen 1960). In the United States, courts generally adopted an antiunion stance until 1935. In that year the National Labor Relations Act was passed. It focused largely on labor's right to organize. It became an unfair labor practice for an employer to coerce or to restrain employees in the exercise of their rights, to dominate or interfere with a labor organization, or to refuse to bargain collectively with a legal representative of the employees. Subsequent legislation refined, broadened, and set legal limits on the scope of management and union activities. From the perspective of labor, the development of free collective bargaining and the USA labor movement created the need for what is now recognized as effective human resource management.

2.3 Scientific Management

From the perspective of management, the scientific management movement also created the need for effective HRM. Frederick Winslow Taylor was the prophet of scientific management, and the stopwatch was his 'bible' (Bell 1972). Taylor began his experiments in the steel industry in 1885. His objective was the systematic analysis and breakdown of work into its smallest mechanical elements, and then their rearrangement into the most efficient combination. In addition to the scientific study of the task itself ('time and motion' study), Taylor argued that individuals selected to do the work should be as perfectly matched physically and mentally to the demands of the job as possible, and that overqualified individuals should be excluded.

Employees should be trained carefully by supervisors (whose own work was also divided into specialties) to ensure that they perform the work exactly as specified by prior scientific analysis. In no case, however, should employees ever be called upon to work at a pace that would be detrimental to their health.

Finally, to provide an incentive for the employee to follow the detailed procedures specified (and closely supervised by line supervisors on the shop floor), Taylor felt that employees should receive an addition of between 30 and 100 percent of their ordinary wages when the task was done right and within the time limits specified. The first piece-rate incentive system had arrived.

Taylor also was interested in the social aspects of work, although he saw little good emerging from the social interaction within work groups. He felt that work groups fostered a level of individual efficiency equal to the level of the least productive worker in the group. In other words, he believed that the efficiency of the group would not be any greater than the efficiency of the least productive member. Thus Taylor tried to institute a kind of social physics into the workplace. While many of his ideas are controversial, there is no doubt that application of the principles of scientific management resulted in much higher productivity than would otherwise have been possible.

2.4 Early Industrial Psychology

In 1913, Hugo Munsterberg's book, *Psychology and Industrial Efficiency*, described experiments in selecting streetcar operators, ship's officers, and telephone switchboard operators. His contributions to human resource (or personnel) management are noteworthy for his emphasis on the analysis of jobs in terms of (a) the abilities required to do them, and (b) the development of testing devices—aptitude as well as work sample tests—that were able to forecast subsequent job performance (Munsterberg 1913).

Munsterberg's 1917 book, *Business Psychology*, included the results of questionnaires sent to hundreds of laborers. The questionnaires requested anonymous data on attitudes toward their work, their involvement in the labor movement, and their leisure activities.

Both Munsterberg and Taylor were concerned with how workers were selected for jobs, as well as with the individual worker's feelings and aspirations. Although they differed on several issues (notably, on how social structure affected productivity), Munsterberg's writings clearly suggest that he felt that the two emerging fields of personnel selection and social psychology (the study of how people affect and are affected by one another) were complementary (Landy 1992, Moskowitz 1977).

Other developments at about the same time included the use of rating sheets for interviewers, advances in checking references given by workers, and in statistical methods for estimating validity (the extent to which selection devices accurately forecast job performance). World War I accelerated the development of intelligence tests, such as the Army Alpha and the Army Beta, so that each individual could be matched more effectively with job requirements. Other kinds of psychological measures also appeared during and after World War I, such as measures of aptitude, interest, and personality.

2.5 The US Civil Service Commission

The Pendleton Act of 1883 established the US Civil Service Commission (today known as the US Office of Personnel Management). A forerunner in progressive personnel policies, the act provided that competitive examinations be administered as a basis for employment in the public service. It also provided a measure of employment security for those selected, prohibited discharge for refusing to engage in political activity, encouraged a nonpartisan approach to appointments, and mandated that a commissioner administer the act (Van Riper 1958). Perhaps the major impact of this act was to foster employment and promotion policies in the federal government based on merit. Over the years, these policies have influenced personnel practices in state and local governments as well as in private industry.

2.6 Private Industry's Approach to Personnel Management

Historians consider 1912 the approximate date of the emergence of the modern personnel department (Eilbert 1959). The term 'personnel,' with its connotation of managing people in organizations, began to appear about 1909. Between 1900 and 1920, while advances were being made in scientific management, industrial psychology, and the federal civil service, a number of personnel specialists emerged in companies such as B. F. Goodrich, National Cash Register, and Standard Oil of California. The specialists managed such areas as employment, employee welfare (financial, housing, medical, and educational), wage setting, safety, training, and health. This kind of specialization formed the basis for the organization of the modern human resources department.

2.7 The Human Relations Movement

In Chicago, in 1923, the Hawthorne Works of the Western Electric Company provided the setting for one of the most famous behavioral research efforts of all time. The purpose of the research was to study the effects of illumination on worker productivity. In one experiment, production went up when the illumination increased. In another experiment, production went up when the illumination decreased. After three years of continually inconsistent results, the researchers concluded that in a situation involving people, it was impossible to change one condition (the lighting) without affecting other variables at the same time. Employee motivation and work group morale seemed to be the really crucial factors.

The results of subsequent experiments at the Hawthorne Works by Professors Elton Mayo, Fritz Roethlisberger, and T. North Whitehead of the Harvard Business School indicated that productivity was directly related to the degree of group teamwork and cooperation. The level of teamwork and cooperation, in turn, seemed to be related to the interest of the supervisor and the researchers in the work group, the lack of coercive approaches to productivity improvement, and the participation afforded the workers in changes affecting them (Pennock 1930, Roethlisberger and Dickson 1939). In short, the researchers came to view the organization of workers as a *social system*, in contrast to Taylor's view of that organization as a *technical–economic system*.

The conviction that group behavior and workers' feelings were associated with morale and productivity characterized much of the research and theorizing in the human relations movement for the next two decades. In practice, unfortunately, many managers seemed to use human relations for the short-term purpose of manipulating workers to increase output, rather than for the long-term goal of satisfying worker needs while meeting organizational needs. Failure to find evidence that these programs made a difference in workers' satisfaction or that happy workers were productive workers helped kill this approach to HRM (Brayfield and Crockett 1955).

2.8 The Behavioral Sciences

The behavioral science approach to managing people is an outgrowth of the human relations studies, although it embraces a wider base of academic and applied disciplines and is concerned with a wider range of problems. The term *behavioral sciences* refers to the social and biological sciences concerned with the study of human behavior. Much of the knowledge of HRM and many of its practical applications have come from such behavioral science disciplines as the following:

(a) *Industrial/organizational psychology*: the study of the behavior of people at work (see *Industrial–Organizational Psychology: Science and Practice*).

(b) *Social psychology*: the study of how people affect and are affected by one another.

(c) *Organization theory*: basic philosophies about why organizations exist, how they function, how they should be designed, and how they can be effective.

(d) *Organizational behavior*: the study of the causes of individual and group behavior, and of how this knowledge can be used to make people more productive and satisfied in organizational settings.

(e) *Sociology*: the study of society, social institutions, and social relationships.

We now know that the ways people behave in organizations cannot be explained simply by human relationships. The organization itself, through its unique 'culture,' molds, constrains, and modifies human performance. The way the organization is structured, the authority attached to different positions, and job and technology requirements clearly affect behavior. Even though the present understanding of the determinants and effects of behavior is incomplete, there is a better sense of the ways in which separate influences interact with each other to affect individuals.

Another area that has exerted a considerable influence on HRM theory and practice is *general systems theory*. Systems theory has developed from such fields as biology, physics, and cybernetics (the study of mechanical–electrical communications systems and devices). It has been applied and elaborated in every behavioral science field.

2.9 Social Legislation and Court Decisions

Governments in most developed countries around the world have passed social legislation and courts have issued decisions on unfair discrimination in employment. While some countries are more progressive on these issues than others, the laws and court decisions frame national policy on equal employment opportunity (EEO). In the USA, discrimination in employment is prohibited on the basis of race, religion, national origin, gender, age, and disability. Aggrieved individuals have the right to pursue their claims in court. At the level of the organization, HRM specialists work with top managers to develop and implement EEO policies to ensure that individuals are treated fairly in all aspects of the employment relationship.

3. Growth Stages of Human Resource Management

The foundation for modern HRM rests on the factors described earlier. Beyond that, HRM as practiced in the USA developed in four stages.

3.1 File Maintenance

The first stage may be called the 'file maintenance' stage, for it typified HRM activities through the mid-1960s and the degree of emphasis on employee concerns. 'Personnel' was the responsibility of a special department. Its responsibilities included screening applicants, conducting orientation for new employees, collecting and storing personal data on each employee (e.g., date of birth, years of company service, education), planning the company picnic, and circulating memos 'whose impertinence was exceeded only by their irrelevance' (Meyer 1976).

3.2 Government Accountability

The second growth stage of HRM, the 'government accountability' stage, began soon after the Civil Rights Act of 1964 was passed. Laws governing antidiscrimination, pensions, and health and safety were enacted, federal agencies to enforce the laws were created, and court rulings affecting almost every aspect of the employment relationship were issued. All of these events accelerated the rise and importance of the HRM function. Class action lawsuits and the large financial settlements for the winning suits illustrated the costs of mismanaging people. Managers outside the HRM function began to take notice because top management let them know that ineptitude in this area would not be tolerated. Staying out of federal court became a top organizational priority. These trends also signaled the need for specialized competence in each aspect of HRM. Thus specialties began to appear in compensation and benefits, affirmative action (the promotion of minority concerns in all aspects of employment), labor relations, and training and development.

3.3 Organizational Accountability

In the late 1970s and 1980s, when many firms were struggling simply to survive, a combination of economic and political factors (high interest rates, growing international competition, shrinking US productivity growth) led to a demand for greater accountability in dollar terms of all the functional areas of business. HRM activities were no exception. Thus this third stage may be termed 'organizational accountability.' Although methods of assessing the costs and benefits of HR programs are available, they are not widely known (Cascio 2000). In addition, social trends, such as more women in the workforce, more minorities, immigrants, older workers, and poorly educated workers, accelerated the demands for improving the quality of work life, for managing cultural and ethnic diversity, and for continual training and retraining.

3.4 Strategic Partnership

In the 1990s and in the new millennium, HRM has evolved to a fourth growth stage as it assumes a central

role in the struggle of both large and small firms to gain and sustain a competitive advantage in the worldwide marketplace. This stage may be termed 'strategic partnership.' Top management looks to the HR department, as it does to line managers, to control costs, to attract and retain top talent, to enhance competitiveness, and to add value to the firm in everything it does.

4. Learning From the Past, Looking Toward the Future

As noted earlier, the field of human resource management is composed of five broad activities that together constitute the HRM system: staffing, retention, development, adjustment, and managing change. The field as it is known today has emerged from nine interrelated sources: the industrial revolution, the emergence of free collective bargaining, scientific management, early industrial psychology, governmental personnel practices growing out of the establishment of the Civil Service Commission in the USA, the emergence of personnel specialists and the grouping of these specialists into personnel departments, the human relations movement, the behavioral sciences, and social legislation and court decisions.

Since the 1950s, HRM has evolved through four growth stages: file maintenance, government accountability, organizational accountability, and strategic partnership. However, to be business partners who make genuine contributions to the success of an organization, HR professionals need to be proactive, not reactive. They must anticipate changes in the social, legal, political, and economic environments, and then facilitate effective organizational responses to them. Some of these issues include finding and keeping top talent in tight labor markets; capitalizing on the information revolution in ways that will benefit organizations, such as Internet-based recruitment and the development of interactive, company-based intranets to facilitate employee inquiries and inputs; managing the performance of geographically dispersed employees and teams (virtual teams); and providing training and development opportunities on demand. Indeed, this is the new paradigm of work: any time, anywhere, in real space or in cyberspace. HRM will contribute to the bottom-line success of organizations and to the betterment of employee welfare by developing proactive responses to the workplace challenges wrought by this new paradigm.

See also: Industrial–Organizational Psychology: Science and Practice; Job Analysis and Work Roles, Psychology of; Job Design and Evaluation: Organizational Aspects; Labor Unions; Performance Evaluation in Work Settings; Psychological Climate in the Work Setting; Strategic Human Resources Management; Well-being and Burnout in the Workplace, Psychology of

Bibliography

Bell D 1972 Three technologies: Size, measurement, hierarchy. In: Davis L E, Taylor J C (eds.) *Design of Jobs*. Penguin, London

Brayfield A H, Crockett W H 1955 Employee attitudes and employee performance. *Psychological Bulletin* **52**: 396–424

Cascio W F 1998 *Managing Human Resources: Productivity, Quality of Work Life, Profits*, 5th edn. Irwin/McGraw-Hill, Burr Ridge, IL

Cascio W F 2000 *Costing Human Resources: The Financial Impact of Behavior in Organizations*, 5th edn. Southwestern, Cincinnati, OH

Cohen S 1960 *Labor in the United States*. Charles E. Merrill, Columbus, OH

Eilbert H 1959 The development of personnel management in the United States. *Business History Review* **33**: 345–64

French W L 1994 *Human Resources Management*, 3rd edn. Houghton Mifflin, Boston, MA

Landy F J 1992 Hugo Munsterberg: Victim or visionary? *Journal of Applied Psychology* **77**: 787–802

Meyer H E 1976 Personnel directors are the new corporate heroes. *Fortune* **93**: 84–8

Moskowitz M J 1977 Hugo Munsterberg: A study in the history of applied psychology. *American Psychologist* **32**: 824–42

Munsterberg H 1913 *Psychology and Industrial Efficiency*. Houghton Mifflin, Boston, MA

Pennock G A 1930 Industrial research at Hawthorne and experimental investigation of rest periods, working condi tions, and other influences. *Personnel Journal* **8**: 296–309

Roethlisberger F J, Dickson W J 1939 *Management and the Worker*. Harvard University Press, Cambridge, MA

Van Riper P P 1958 *History of the United States Civil Service*. Row-Peterson, Evanston, IL

W. F. Cascio

Human Rights, Anthropology of

Human rights constitute a field of increasing anthropological concern, not only for its substance but also for its inherent tensions. Since 1948, when the Universal Declaration of Human Rights (UDHR) was internationally accepted, anthropology has moved from an explicit opposition to the idea of a universal standard of human rights, over a more routinely skeptical attitude to a growing positive interest of both an empirical and a theoretical nature. This trend is related both to changing notions of anthropology and to the historical process of globalization. (See

Downing and Kushner 1988, and Messer 1993, and Hastrup 2001a, 2001b for overviews of the relationship between anthropology and human rights.)

1. *Precursors*

Although human rights, in the sense used here, were not on the international agenda before 1948, it was a latent idea for centuries, especially as far as civil rights were concerned. The American 'Bill of Rights' (1776) and the French declaration on '*Les droits de l'homme et de citoyens*' (1789) are often cited as ancestors of the UDHR (along with others), and from there the roots of human rights are traced back to classical Athens or early Christianity. In the nineteenth century, when a wave of humanism swept over Europe, one of the results was a new idea of an essential 'humanity,' deriving from seventeenth- and eighteenth-century treatises in political philosophy as well as from sixteenth-century renaissance humanism. Political philosophy in many respects was a forerunner of anthropology that was to take shape in the late nineteenth century. The idea of a universal humanity was a precondition of the nascent discipline, destined to give the proposition about a shared humanity empirical substance by way of qualifying cultural differences as variations upon the common theme of humanity.

In the early ages of anthropology the conjunction of the two concerns of equal rights and local difference is passionately testified to by Franz Boas, who, in his letter-diary from Baffin Island (1883–4), writes about his sincere wish to live in the USA:

> I should much prefer to live in America in order to be able to further those ideas for which I live. … What I want to live and die for, is equal rights for all, equal possibilities to learn and work for poor and rich alike! Don't you believe that to have done even the smallest bit for this, is more than all of science taken together? (quoted in Cole 1983, p. 37)

We know that later Boas was to fight fervently against racism, and on the whole most anthropologists see their discipline as genuinely respectful of all peoples. This collective sentiment makes no simple entry into the discussion of a universal standard of human rights, however, as this is conceptualized and articulated in those declarations, conventions, charters, and protocols that have been passed in the decades following the UDHR.

2. *Anthropological Reactions to the Universal Declaration of Human Rights*

Human rights in the modern sense were born in 1948 with the adoption of the UDHR. In the drafting process, advice and comments were sought from various bodies, including the American Anthropological Association (AAA). A statement was produced by Melville Herskovits (1947) in the name of the AAA which carefully considers the idea of a universal standard and notes its local cultural implications, each of which is firmly anchored in contemporary anthropological knowledge. Noting that human rights appertain to the individual, the argument proceeds through the following successive conclusions:

(a) 'The individual realizes his personality through his culture, hence respect for individual differences entail a respect for cultural differences' (p. 541);

(b) 'Respect for differences between cultures is validated by the scientific fact that no technique of qualitatively evaluating cultures has been discovered' (p. 542);

(c) 'Standards and values are relative to the culture from which they derive so that any attempt to formulate postulates that grow out of the beliefs or moral codes of one culture must to that extent detract from the applicability of any Declaration of Human Rights to mankind as a whole' (p. 542).

This is clearly a stand against what was perceived as a Western bias in the UDHR, and it leads to the final conclusion: 'Only when a statement of the right of men to live in terms of their own traditions is incorporated into the proposed Declaration, then, can the next step of defining the rights and duties of human groups as regards each other be set upon the firm foundation of the present-day scientific knowledge of Man' (p. 543).

What transpires is a profound skepticism, grounded in the state of the anthropological art, including the general professional sentiment that culture is the sole legitimate source of moral values. Herskovits expresses it thus: 'Even where political systems exist that deny citizens the right of participation in their government, or seek to conquer weaker peoples, underlying cultural values may be called on to bring peoples of such states to a realization of the consequences of the acts of their governments, and thus to enforce a brake upon discrimination and conquest' (p. 543).

As pointed out by Richard Wilson (1997, p. 2), this is an attempt to play anthropology's trump card: its unique ability to elicit and assert any culture's underlying values. In addition it could be said that it also features the comparative knowledge inherent in anthropology, a knowledge both of cultural difference and of its translatability into valid, and more or less shared, human standards of good governance and respectable political action. In spite of the disclaimer, which was militantly underscored by Julian Steward (1948, p. 251), when he said 'we are prepared to take a stand against the values in our own culture which underlie … such imperialism,' the UDHR was passed in 1948.

Together, the two attitudes exemplified by Boas and Herskovits, both of which are understandable and laudable, take us right to the core of anthropology's

comparative consciousness and its ambiguous implications. Since 1948, the discussion of the concept of human rights in anthropology has centered on the unsolved question of universalism or relativism (Dembour 1996). Another target has been the cultural translation of human rights concepts, allegedly deriving from a Western, and Christian, liberal democratic system, into 'like' cultural concepts within other moral and political systems. This is neither an easy task (as shown by e.g., Wilson 1996, and Asad 1997), nor is it an obvious way to go (Hastrup in press).

In actual practice the anthropological concern with human rights has taken various forms since 1948, and has been named accordingly as advocacy, activism, antiracism, and anti-imperialism. It is characteristic, though, that most of these trends have continued the relativistic attitude embedded in the AAA statement (and the implied skepticism toward human rights), while also advocating equal rights on the basis of a vision of a shared humanity. At the beginning of the twenty-first century, when human rights are appealed to by the international community under the banner of a 'culture of human rights' with increasing force, anthropologists have also taken a renewed interest in this culture.

3. *The Culture of Human Rights*

The intense talk of human rights since the early 1990s is not necessarily a symptom of a high moral profile worldwide, rather it is a symptom of the malaises of the times and of the will to confront them by way of new international laws. The choice of cure reflects a general process of legalization of culture in a postmodern era fraught with ontological insecurity, if we are to believe Anthony Giddens (1991). Where, earlier on, local cultures provided people with a sense of belonging and identity, the present globalization and fragmentation of identities have allegedly left people at a loss. They then look toward the law as the glue of community, whether local or global; the law is the new means of inclusion or exclusion *par excellence*.

Whether it is possible to single out causal links in this complex process or not, we do experience an increasingly strongly articulated 'culture of human rights.' We encounter this notion in most recent statements made by the UN high commissioner for human rights, Mary Robinson, as well as by academics of various kinds (e.g., Rorty 1993). From an anthropological perspective it is a peculiar culture in the sense that it is declared rather than lived, and that it is future-oriented rather than based in tradition. As such it parts company from those cultures that were the traditional object of anthropology, but so does, indeed, anthropology itself, becoming increasingly concerned with connections between the local and the global, and their mutual implications. This provides a theoretically significant conjuncture between anthropology and the human rights discourse.

The culture of human rights is distinct also in its explicit acknowledgement of multiple sources. Whereas, traditionally, nations and other self-declared ethnic groups have referred to a particular genesis, an ultimate ancestor, or a culture-hero, the human rights culture draws upon all kinds of ancestry. As Mary Robinson (1998) puts it:

> Today the Universal Declaration of Human Rights stands as ... one of the great documents in world history. The *travaux préparatoires* are there to remind us that the authors sought to reflect in their work the differing cultural traditions in the world. The result is a distillation of many of the values inherent in the world's major legal systems and religious beliefs including the Buddhist, Christian, Hindu, Islamic, and Jewish traditions.

Not surprisingly, after the demise of colonialism, and with the increasing critique of Eurocentrism and the hegemony of Western values, the search for roots has changed from the attempt at establishing a one-stranded genealogy to suggesting multiple sources. Mary Robinson's statement may be empirically doubtful, but it goes to show that there is a will to shift whatever Western legacy one might like to trace in the human rights instruments to a truly global commitment.

The declaration of a global culture of human rights is thus closely related to the historical process of globalization. In one sense, the notion of globalization simply states the obvious; due to increasingly rapid information and money exchange systems, we experience a time–space compression, that accounts for a new global consciousness. This global consciousness has important implications, one apparently being the appeal to a transcendent 'culture of human rights.'

To invoke a culture is also to invoke a shared language; as far as the human rights culture is concerned, the language is decidedly legal. As 'declarations' the international charters are exemplars both of language and of law. Whether hard or soft law, it is inevitably an expression of particular ideas about society, while not necessarily embracing its total reality. Far from being simply rules for social behavior, law is part of a distinctive manner of imagining the real (Geertz 1983, p. 173).

In the case of the international discourse on human rights, there is a concept of the modern nation-state as the responsible agent vis-à-vis the citizens, and a vision of an international community that transcends the more narrow interests of the states. Clearly, this community is no less 'imagined' than the nation (cf. Anderson 1983). It is imagined in terms of precisely those international charters that are its main expression. As a literally *declared* culture, its language comes to carry a much heavier burden of representation than any 'natural' language, being just one among other media for storing cultural knowledge.

Any culture in some ways rests on representation; for it to be conceivable as a whole, it must be demonstrated as such. In a 'legal culture,' such as the international culture of human rights, the representation becomes of a particularly reductive kind, because the experiences, by being represented as if on permanent trial in a courtroom, become subjected to an unavoidable process of reduction, or as Clifford Geertz has it about law in general, a 'skeletonization of fact, the reduction of it to the genre capacities of the law note' (Geertz 1983, p. 172). Legal thought and vocabularies, then, are constructive of social realities rather than merely reflective of them; this also applies to the legal vocabulary of human rights, and anthropologists naturally take an increasing interest in this particular social construction of a (global) community.

According to some anthropologists, globalization and human rights conjoin in a social transformation of unprecedented magnitude. Ellen Messer, for one, writes like this:

> Since the close of World War II, the United Nations has been assembling declarations, legislation, and enforcement mechanisms to promote human rights. Both the ongoing efforts to establish a global community and to base membership on a universal, but evolving standard of values constitute perhaps the greatest social transformation of this century. (Messer 1996, p. 1)

This transformation is part of a larger trend that has often been seen as the end of modernity, implying an end also to ontological security. This could be one of the reasons why human rights have become the new standard of hope to many. The historical changes are matched by new scholarly efforts. We have witnessed important changes in the conception of the anthropological object, that can be summed up under the rubrics of a de-essentialization of the object (culture, society) on the one hand, and a humanization of the subject, seen now to be an active agent of history, on the other. This facilitates a truly anthropological investigation into issues of human rights.

Intellectually, one of the greatest advances of the twentieth century was a willingness to learn from history and anthropology that humans are extremely malleable (e.g., Rorty 1993, p. 115). Thanks to comparative knowledge of humanity in time and space we have realized the flexibility and the self-shaping character of humans, and given up on the idea of a naturally rational animal.

The historical and intellectual climate is thus very much in favor of achieving a 'human rights culture.' While the dominant trends in present day anthropology—the stress upon globalization, on process, and on agency among other things—make a new approach to universal human rights possible, there is no theoretical reason why we should not at the same time maintain a distinct, more or less bounded object of analysis: the local worlds of people. This has fueled a new interest in what are known as cultural rights.

4. *Cultural Rights*

In view of the traditional anthropological ambivalence toward human rights, generally based in view of culture as an inviolable and inherently valuable entity, the notion of cultural rights might ease the tension. This is not self-evident, however.

In the early instruments, such as the UDHR and the 1996 International Covenant on Social, Economic, and Cultural Rights (ICSECR), 'cultural rights' are individual rights to 'enjoy culture,' in compliance with the general emphasis on individual rights and the obligation of the state to facilitate them. In addition to this focus on the individual's cultural rights, a more challenging (politically and theoretically) perspective on cultural rights is opening up with increasing force, however. This is a notion of collective cultural rights, in relation to which the individual enjoyment of culture gives way to, or is at least supplemented by, a right of a particular group to protect and develop their culture. It was notable already in the *International Covenant on Civil and Political Rights* (ICCPR 1966, Article 27): 'In those States in which ethnic, religious, or linguistic minorities exist, persons belonging to such minorities shall not be denied the right, in community with the other members of their group, to enjoy their own culture, to profess and practice their own religion, or to use their own language.' It is surprising, perhaps, that the protection of minorities is placed within the civil and political rights, and not the social, economic, and cultural rights; if nothing else, this confirms the lack of clarity with which the concept of culture is hampered in the international instruments.

In spite of the weak beginnings and the unclear focus within the field of cultural rights, there is an increasing interest in them at the beginning of the twenty-first century. It is related to the historical fact that most conflicts today are so-called ethnic conflicts, at closer inspection often rooted in economic discrimination or scarce resources. Ethnic difference often covers other inequalities, at least when they are expressed in conflict. The ethnic conflicts take place within states, that are the traditional duty-holders in the human rights system, and they raise endless new questions of how to define those peoples, minorities, or indigenous populations that are the new beneficiaries of collective rights. In spite of the definitional problems, the UN system proceeds toward an increasing proliferation of instruments dealing with collective rights.

One such instrument is the UN *Declaration on the Rights of Persons Belonging to National or Ethnic, Religious or Linguistic Minorities* (DM) of 1992. As a Declaration it is not legally binding in the strict sense of the term, but state signatories nevertheless commit

themselves to the spirit of the Declaration. In many ways the DM can be seen as an extension of Article 27 of the ICCPR quoted above:

> States shall protect the existence and the national or ethnic, cultural, religious and linguistic identity of minorities within their respective territories and shall encourage conditions for the promotion of that identity. (DM, Article 1)

> Persons belonging to national or ethnic, religious and linguistic minorities … have the right to enjoy their own culture, to profess and practice their own religion, and to use their own language, in private and in public, freely and without interference or any form of discrimination. (DM, Article 2)

From an anthropological perspective, what seems problematic is the recurrent emphasis on people's own culture as a natural property. The embedded concept of culture is an (outdated) essentialist one that goes counter to more recent anthropological notions of culture as a more or less temporary framework of understanding.

In the 1993 *Draft Declaration on Indigenous Peoples Rights* (IPR) the notion of culture as essence is further cemented; to give just one example:

> Indigenous peoples have the right to practice and revitalize their cultural traditions and customs. This includes the right to maintain, protect, and develop the past, present, and future manifestations of their cultures, such as archeological and historical sites, artifacts, designs, ceremonies, technologies, and visual and performing arts and literature, as well as their right to the restitution of cultural, intellectual, religious, and spiritual property taken without their free and informed consent or in violation of their laws, traditions, and customs. (IPR, Article 12)

While one cannot deny that the indigenous peoples have suffered the more sinister effects of colonialism, it is unlikely that such collective rights that are propounded in the IPR will undo history. It is still only a draft declaration but seen in conjunction with the previously cited instrument there is evidence of a gradual transformation of cultural rights from individual to collective rights, that may potentially threaten the very foundation of human rights: their claim to universality. It seems, then, that at a time when anthropologists have left behind the notion of culture that prompted the AAA to take a stand against the idea of universal human rights, the very same notion hits back from within the system of human rights itself, potentially undermining it.

5. *New Anthropological Avenues*

The brief discussion of cultural rights shows that there is a need to reconsider the issue of human rights in the light of recent theoretical developments. If the notion of culture as essence is obsolete and cultural relativism generally self-refuting, human rights can no longer be invalidated as a matter of course. The comparative perspective of anthropology itself is based on an idea of a shared humanity and a cross-cultural comprehensibility that tallies with the idea of equality and dignity in all humans. One of the things that anthropology has helped realize is the imaginability of 'others' as somehow like 'us' in spite of diversity.

Ethnographies of violence and suffering are still needed to add a dimension of experience to the (legal) language of rights, and thus contribute to a shared understanding of those sufferings that a universally accepted standard of human rights may help alleviate. Not only ethnography, but also our regained confidence in anthropological theory will contribute to the achievement of human rights, because anthropology's theoretical language can embrace cultural differences within a single (scholarly) discourse. Theory expresses what is not otherwise said; as such it is a social practice and contributes to the making of the world—not unlike the representation of society in legal terms, as discussed above. This imaginative contribution makes room for the renewed rapprochement between anthropology and human rights that we are witnessing.

See also: Boas, Franz (1858–1942); Cultural Rights and Culture Defense: Cultural Concerns; Human Rights, History of; Human Rights in Intercultural Discourse: Cultural Concerns; Human Rights: Political Aspects; Offenses against the Laws of Humanity: International Action

Bibliography

Anderson B R O'G 1983 *Imagined Communities. Reflections on the Origin and Spread of Nationalism*. Verso, London

Asad T 1997 On torture, or cruel, inhuman and degrading treatment. In: Wilson R A (ed.) *Human Rights, Culture and Context. Anthropological Perspectives*. Pluto Press, London, pp. 111–33

Cole D 1983 'The value of a person lies in his *Herzensbildung*': Franz Boas' Baffin Island letter-diary, 1883–1884. In: Stocking G W Jr (ed.) *Observers Observed: Essays on Ethnographic Fieldwork*. University of Wisconsin Press, Madison, WI, pp. 13–52

Dembour M-B 1996 Human rights talk and anthropological ambivalence. The particular contexts of universal claims. In: Harris O (ed.) *Inside and Outside the Law*. Routledge, London, pp. 19–40

Downing T, Kushner G (eds.) 1988 *Human Rights and Anthropology* (Cultural Survival Report 24). Cultural Survival, Cambridge, MA

Geertz C 1983 Local knowledge: Fact and law in comparative perspective. In: *Local Knowledge*. Basic Books, New York, pp. 167–234

Giddens A 1991 *Identity and Modernity*. Polity Press, Cambridge, UK

Hastrup K (ed.) 2001a *Human Rights on Common Grounds. The Quest for Universality*. Kluwer Law International, The Hague

Hastrup K (ed.) 2001b *Legal Cultures and Human Rights. The Challenge of Diversity*. Kluwer Law International, The Hague

Hastrup K in press Anthropology's comparative consciousness: The case of human rights. In: Fox R, Gingrich A (eds.) *Comparison in Anthropology*. Routledge, London
Herskovits M 1947 AAA statement on human rights. *American Anthropologist* **49**: 539–43
Messer E 1993 Anthropology and human rights. *Annual Review of Anthropology* **22**: 221–49
Messer E 1996 Anthropology, human rights, and social transformation. In: Moran E F (ed.) *Transforming Societies, Transforming Anthropology*. University of Michigan Press, Ann Arbor, MI
Robinson M 1998 *The Universal Declaration of Human Rights: A Living Document*. United Nations Press Release 27 January 1998 (UNHCHR)
Rorty R 1993 Human rights, rationality, and sentimentality. In: Shute S, Hurley S (eds.) *On Human Rights. The Oxford Amnesty Lectures 1993*. Basic Books, New York, pp. 111–34
Steward J 1948 Comments on the Statement of Human Rights. *American Anthropologist* **50**: 351–2
Taylor C 1985a *Human Agency and Language. Philosophical Papers I*. Cambridge University Press, Cambridge, UK
Wilson R A 1996 The Sizwe will not go away. The Truth and Reconciliation Commission, human rights and nation-building in South Africa. *African Studies* **55**(2): 1–20
Wilson R A 1997 Human rights, culture and context: An introduction. In: Wilson R A (ed.) *Human Rights, Culture and Context. Anthropological Perspectives*. Pluto Press, London, pp. 1–27

K. Hastrup

Human Rights, History of

The expression 'History of Human Rights' has two different meanings. It can be understood (as it usually is) as the history of the idea and the concept of human rights, or (less frequently) as the historical occurrence and study of human rights and their abuses. Therefore, the basic debates about the idea and concept of human rights are first clarified here from a historical perspective. Then, a range of methodological and ethical problems in the historical study of human rights are presented.

1. History of the Idea and Concept of Human Rights

1.1 Moral vs. Positive Rights

In 1689–90, just after the Glorious Revolution of 1688, the English philosopher John Locke (1632–1704) wrote his *Two Treatises on Government*, which contained the first classical formulation of human rights (then still called 'natural rights'). In retrospect, this work stood at the juncture of a set of amazingly diverse ideas, some of which went back to antiquity and some of which will survive the twenty-first century. Locke's point of departure was the doctrine of natural law as it had developed since antiquity and had been transformed, in certain respects almost beyond recognition, by medieval Christianity and early-modern philosophers. According to Locke, the doctrine maintained that nature endowed all humans with certain basic moral entitlements (such as the right to life, liberty, and property), to which the man-made laws of society, both positive and customary, had to conform.

These natural rights were thought to possess five distinct characteristics: they were discoverable by reason, itself a faculty of human nature; they existed before humans entered society; they applied to all humans by virtue of their being human; they were inalienable; and they were restricted only by the recognition of the rights of others. As such, these rights limited the power of the society—generally in its predominant historical-political form, the state—and its laws, and respect for them was the first test for the legitimacy of political power. Ancient ideas on rationalism, equality, and the restrained exercise of power formed the substratum for this natural law doctrine, which proved to be very influential. Despite its claim of eternity, however, many of the doctrinal components, such as the exact catalog of moral principles or the source of authority for these principles, turned out to be rather flexible.

Although temporarily in eclipse because of the ascendancy of the sixteenth-century absolutist state, the natural law doctrine would be fostered by three developments initiated in that very period. With the Renaissance came an emphasis on individualism and secularism, the outcome of the religious struggles following the Reformation taught the necessity of tolerance, and the seventeenth-century scientific revolution marked the triumph of rationalism. These developments reversed the philosophical priorities of the hierarchical medieval societies: natural law gradually became associated with rights rather than with duties and with individuals rather than with groups.

The perception of the source of these natural rights also changed. Whereas Christianity had maintained that God was the ultimate source and political absolutism had invested the monarch with divine rights, the consent of the governed came to be seen as the final authority. Individuals gave this consent in a social contract, which presupposed that they possessed rights before entering society. Governments owed their origin to that contract, by which free and equal citizens agreed to entrust to a neutral sovereign body some of their rights, to accept its authority and rules as long as it protected them, and to rebel against it if it did not. This formed the basis for the rule of law and the separation of, and balance between, the legislative, executive, and judicial functions of government.

Although previous theories (such as Hugo Grotius') paved the way and later ones expressed similar ideas, it was Locke's political theory which was to prosper, especially outside England. In the age of Enlightenment it exerted profound influence upon the politicians behind the great revolutions at the end of the

eighteenth century. In the United States and France, natural rights, mainly of a political and civic character, were incorporated into the declarations and constitutions of the revolutionaries. In 1789 natural rights transformed into the 'rights of man' (*droits de l'homme*). This marked the beginning of a process of codification of rights within the constitutions of most states and in international law that still has not ended.

From this brief historical overview, it is clear that, although many aspects of the human rights conception have an ancient pedigree, other crucial ones have not. The concept was the result of a unique historical combination of a series of events and philosophical thought reflecting upon them, and developed relatively late in Western history. The natural law doctrine itself knew a curious fate: with unprovable tenets and with the possibility of abusive claims 'in the name of nature,' it became overshadowed by the powerful state-centered thinking of the nineteenth century (as it had been in the sixteenth), but it survived. It was rehabilitated in a modest form after the two world wars. The twentieth-century approach did not derive human rights from nature any longer, but from (historically determined) rational ideas on human dignity.

1.2 Absolute vs. Relative Rights

The debate on moral vs. positive rights was continued by liberal, conservative, and socialist critics of human rights. Typical of this criticism were the works of Jeremy Bentham (1748–1832), Edmund Burke (1729–97) and Karl Marx (1818–83). These critics argued against the abstract and absolute character of human rights. They maintained that rights had to be related to the society in which they were to be exercised. The liberal utilitarian Bentham believed only in the force of positive legislation. For him, natural rights were imaginary, 'nonsense upon stilts.' He feared that they were powerful rhetoric in the hands of rulers and a substitute for effective legislation.

Unlike Bentham, the conservative Burke did not deny the existence of natural rights. However, he feared them for reasons contrary to Bentham's: in the hands of the common people, natural rights would stimulate revolutionary sentiments and cause social upheaval leading to terror and to the destruction of the social fabric. Rights (and the duties linked with them) ought to be derived from the principles and values emanating from the particular traditions of a society.

For the socialist Marx the human rights doctrine was not radical enough. It left the inequalities generated by capitalism untouched (particularly private property) and therefore served the interests of the ruling bourgeois class. Because of these inequalities, the workers were unable to enjoy their rights fully, which therefore remained largely formal. As a class, the workers had a duty to create the conditions for social change by claiming collective social and economic rights with which to overturn the economic system. The new, egalitarian society would be one without a state, laws, or rights.

These thinkers did not stand alone in their criticism, but their arguments gained a classic character and inspired powerful currents. Bentham's ideas gave impetus to the tradition of legal positivism. Ideas similar to Burke's would feed the romanticism and historicism of the nineteenth century to the extent that they too emphasized cultural identity and diversity. Marx's concerns heightened awareness of the need for social legislation and influenced the 1917 Russian Revolution.

1.3 Men vs. All Humans and Individuals vs. Groups as Rights-holders

Clearly, until long after the era of revolutions, the concept of 'rights of man' covered only very restricted categories of humans, usually male tax-paying citizens. The large majority of women, slaves, and foreigners were excluded. The struggle to include these categories in the ranks of rights-holders would lead to the final transformation of the concept into 'human rights.' As early as 1791, Olympe de Gouges struggled for women's rights. With the movement against slavery, another group followed. Categories of individuals deemed particularly vulnerable (women, children, aliens, refugees, prisoners, the disabled) were eventually protected by special regulations.

Likewise, the difficult problem of how to realize group rights was introduced at an early stage in the human rights debate. Groups deemed vulnerable (minorities, colonized peoples, indigenous peoples) increasingly received special attention. Specific treaties for national, ethnic, or religious minorities were drafted long before general treaties came into effect. The idea of self-determination appealed to peoples under colonial rule or comparable alien subjugation and proved to be a crucial concept. It was a dangerous concept too, because the narrowing of self-determination to only one of its possible interpretations—secession—was potentially explosive and because consensus was lacking as to whom the rights-bearer ('a people') exactly was. Self-determination was promoted to a group right only in the 1966 covenants (see below).

From an acute attention to making the concept 'human' all-embracing, two further questions on the extension of rights can be derived. First, do past and future generations have human rights? As to past generations, one could think of the right to a decent burial or the right to be treated with respect in historical works. As to future generations, the preservation of the cultural heritage of humanity and of the

natural environment are chief concerns as is the relationship between human rights and the human genome. As to future generations in relation to past generations, the accountability of successor governments for human rights abuses committed by their precursors and the obligation to investigate them are matters of legitimate debate. Second, what consequences does the anthropocentric character of human rights have, as it excludes other sentient beings and is often based on restricted and exploitative conceptions of nature?

1.4 Domestic vs. International Jurisdiction Over Human Rights

Since 1800, two tendencies, one international, the other national, partly opposite, partly complementary, have dominated the debate over who was to control the protection of human rights. A process of internationalization started when human rights were incorporated into the constitutions of growing numbers of countries. At the same time, human rights became part of the agenda of intergovernmental meetings, leading to the humanitarian interventions of the nineteenth century (continuing a tradition dating from the seventeenth century) and the gradual incorporation of human rights into international law. That human rights abuses often created conflicts that threatened international peace was a lesson learned slowly. Treaties were drafted to protect selected vulnerable categories such as slaves, war prisoners, and minorities. The League of Nations, founded in 1919, did not formally recognize human rights but added pioneering treaties for minorities, refugees, and mandate territories.

The decisive historical moment for internationalization, however, came during World War II, when the link between human rights violations and war (hence the link between human rights protection and international peace and security) became clearly visible and, shortly after, when the Nazi and Japanese atrocities led to worldwide condemnation embodied in the international war crimes tribunals of Nuremberg and Tokyo. Although the United Nations only paid secondary attention to human rights in their 1945 Charter, this did not prevent work on an International Bill of Human Rights, consisting of a morally compelling but not legally binding Universal Declaration of Human Rights, adopted in 1948, and two International Covenants, one on economic, social, and cultural rights, and one on civil and political rights, both approved in 1966 and entered into force in 1976.

In the efforts to transform human rights into positive rights by making them enforceable in positive law, usually three stages are distinguished: standard-setting (codification of standards), ratification (making standards binding), and implementation (enforcing standards via sanctions). Seminal steps for enforcing standards were the elaboration of complaint mechanisms for state and nonstate actors (mainly since the 1970s), the appointment of a High Commissioner for Human Rights (1993), and the adoption of the Statute of an International Criminal Court with universal jurisdiction to prosecute alleged perpetrators (including heads of state and government) of genocide, crimes against humanity, war crimes committed in international and internal conflicts, and crimes of aggression (1998). Important preparatory work for these new devices was done by a coalition of nongovernmental organizations, the representatives of civil society. The elaboration of human rights institutions and mechanisms on a world scale was parallelled, and often preceded, by similar developments on a regional scale in Europe, America, Africa, and the Arab world.

In the midst of this evolution, the state was the key player. The cherished principles of the Westphalian world order, established in 1648, were national sovereignty and interstate equality. Their corollary, domestic jurisdiction, excluded interference from abroad in the internal affairs of the state. The vigorous doctrine of nationalism and successive waves of decolonization since the late eighteenth century strengthened these principles. Paradoxically, recognition of national sovereignty made possible progress in international law with its core principle of interference in the internal affairs of states. In the first place, states slowly realized that when they adopted human rights as the basis of the rule of law, they implicitly created a duty to protect them everywhere, including beyond their borders. In the second place, it was painfully obvious that the states themselves were the main perpetrators of human rights abuses. Therefore, states reluctantly but voluntarily began to accept an international monitoring element that restricted their sovereign domestic power. The number of humanitarian interventions increased drastically in the 1990s, but their enabling conditions, forms (diplomatic, political, economic, or military), potential abuses, and effects were hotly disputed.

1.5 First- vs. Second- vs. Third-generation Human Rights

As the success and the criticism of the human rights doctrine expanded the list of human rights, analysts grouped them into generations. The first generation, tied to the idea of liberty, were the traditional or individual rights of a political and civic nature. They were characterized as negative rights ('freedoms from') because they favored governmental abstention. The second generation, tied to the idea of equality, were the modern or collective rights of a social, economic, and cultural nature. They were characterized as

positive rights ('rights to') because they favored governmental intervention. The third generation, tied to the idea of fraternity, were modern or collective rights of a global nature. They were characterized as solidarity rights and usually included the rights to development, human security, peace, and self-determination. The rights of indigenous peoples and the right to a life-sustaining environment are frequently added, but these are also listed as fourth-generation rights by some.

Each generation emerged in a particular historical context. The first generation was the result of the late-eighteenth-century revolutions led by the bourgeoisie that abolished feudalism (and colonialism in the case of the United States). Increasing political emancipation and democratic participation in the next two centuries gave further impetus. The second generation developed in the nineteenth century as the socialist tradition with its reformist and revolutionary strategies, and introduced ideas of welfare and social legislation to counter the exploitation of working classes and colonial peoples and to create the necessary conditions for the full exercise of rights. The third generation is usually associated with the Third World demand for a worldwide redistribution of power and wealth; another factor is the tension generated by the staggering increase of interdependency and global problems on the one hand and the need to safeguard cultural identities and diversity on the other. The three generations of rights can be placed on a continuum with justice at the one end and solidarity at the other.

This bewildering diversity of human rights occasioned controversies on their compatibility, priority, and indivisibility. Because all human rights were conceived as inalienable, they sometimes came into mutual conflict. A classic example of a conflict within the same generation of rights is the tension between expression and religion: what is called 'an opinion' by one person, might be labeled 'blasphemy' by the next. But most conflicts occurred between, rather than within, generations. Whereas first-generation rights assumed the character of universally enforceable standards, second- and third-generation rights were rather aspirations whose fulfilment was utterly dependent on the creation of specific conditions. Faced with all these demands, some states, mainly in the First World, gave priority to first-generation rights, while others, mainly in the Second and Third World, struggled, if not for the primacy, then at least for the equivalence of second- and third-generation rights. Acrimonious debates were held on whether political freedom or economic development was to be accorded priority. Those defending the priority of political freedom ('free minds') were accused of perpetuating inequality, those defending the priority of economic development ('full bellies') of perpetuating dictatorial government. In the Cold War era (and beyond), these arguments were used as political propaganda. The hypothesis that political freedom or democracy should be suspended until certain levels of economic development are attained is contradicted by the bitter experience that in such cases freedom is usually trampled underfoot, while no economic development is attained after all. Unaccountable governments generally exacerbate economic and environmental crises. Moreover, how can a present suspension of freedom be conceived as preparing for its future protection? The reverse hypothesis seems to have firmer ground: with their political rights respected, citizens will participate more fully in the economic development of their society. The correlation between economic development and political freedom seems to be accepted, but under which conditions, for which precise political and economic areas, and in which causal direction the correlation is valid remains largely unresolved.

At the 1993 World Conference on Human Rights in Vienna, the second such conference ever and the first after the Cold War, the indivisibility of human rights was strongly reaffirmed. Even so, seven rights acquired a special status as they were the only ones stipulated by the 1966 Covenant to be nonderogable, even in times of public emergency. They are the rights to life, to be recognized as a person before the law, and to freedom of thought, conscience, and religion; and the prohibitions against torture, slavery, imprisonment for nonfulfillment of contractual obligations, and retroactive application of criminal laws.

1.6 Universality vs. Relativity of Human Rights

Seen from a historical perspective, most major religions and ethical systems have been concerned with human dignity. Opinions differ, however, on whether they were also concerned with human rights (viewed as one specific way to realize human dignity). Developing the human rights concept took two millennia in Western history and was so full of unforeseen links, ruptures, reversals, and coincidences, that it may convincingly be called a unique product of that history (see above). The quasi-universal emphasis on the community in which individuals had mainly duties switched in Europe and its American colonies to an emphasis on individuals and rights. Elements of the human rights concept, such as some natural law principles, were available outside the West, but the combination of factors that led to its formulation was not. As the West conquered the rest of the world, it introduced its dominance and violence. At the same time, its notions of freedom, nationalism, and cultural diversity, developed initially for home consumption, fell on fertile ground overseas. These notions, including human rights, proved able to transcend their particular roots and context. In the hands of colonized peoples, they echoed similar indigenous values and served to justly criticize Western exploitation in the language of the West itself. Thus, whereas human

dignity is a universal value with venerable traditions everywhere, the claim of the universality of human rights has an anthropological basis, not a historical one, and is rooted in the universal appeal of a specific system of thought.

The anthropological debate was no less vivid than the historical debate. During the preparatory work on the 1948 Universal Declaration, leading American anthropologists emphasized the cultural relativity of values and denounced the universalist pretensions of the intended declaration as cultural imperialism. Moreover, the historical critics of the universality of human rights (presented above) were succeeded by neo-Marxists, communitarians, and postmodernists. But the most influential group of critics were the vociferous ruling elites of some non-Western countries, especially in Asia and Africa. They challenged specific formulations of private rights such as religion or marriage, and advocated, instead of universality, the principle of domestic jurisdiction or the primacy of second- and third-generation rights over first-generation rights. Remarkably, the surviving victims of imprisonment, torture, and other abuses in non-Western countries (among them many future elites of these countries) found such criticism unconvincing. These victims argued that denying universal human rights to citizens outside the West was itself a sign of cultural arrogance. A new consensus seemed to emerge when 171 states unanimously reaffirmed the universality claim at the 1993 Vienna Conference and thereby made the claim far more representative than at the time of the 1948 Universal Declaration which was approved by a vote of 48 to zero with eight abstentions. At the same time, care for second- and third-generation rights is growing. In 1998 a draft of a Universal Declaration of Human Responsibilities was submitted to the United Nations General Assembly, as an attempt to achieve a better balance between freedom and responsibility and between justice and solidarity.

2. *History and Historiography of Human Rights and Their Abuses*

The second interpretation of the expression 'History of Human Rights' points to historical research on human rights. When looking at this area, we are confronted with a paradox. On the one hand, as human rights encompass so many spheres of life, large parts of the historiographical production inevitably deal with concrete aspects of human rights and their abuses (e.g. the history of education, the history of slavery, the history of censorship). Equally, historians endeavored to write histories of emancipation movements or of freedom. Grand theories of history and world histories sometimes had implicit views on human rights, and the more recent they were, the more they covered the whole of humanity in an equitable manner. Indeed, with little exaggeration one could say that almost all history touches on some present human rights concern. On the other hand, it is hard to find research explicitly designed to study a particular right in historical perspective. Some of the methodological and ethical problems involved in this type of research are outlined below.

2.1 *Methodological Problems*

In the following example, a general question is analyzed to demonstrate the methodological complexities of human rights research. That question is: 'Do human rights abuses increase or decrease over time?' A historical question because of its time dimension, it is also an ethical question to the extent that the answer gives clues about the moral progress of humanity (philosophers of all times have been in dire need of such clues). Questions as broad as this one should be rendered operational. This may be done by narrowing the 'human rights' of the question to one of its categories, 'civil and political rights,' by selecting from this category one important right, 'the right to life,' and, finally, by selecting from this right its most serious abuse, 'genocide.' Reformulated in operational terms, the question thus becomes: 'Do genocides increase or decrease over time?' This question can be approached logically and historically.

A priori logical considerations distinguish between absolute and relative levels of abuse. In absolute terms, the case of an overall increase of genocides is defensible. The population growth on earth, the increasing sophistication of repressive bureaucracies and technologies, and the anonymity often physically and morally shielding the perpetrator are all contributing factors in making this statistically probable. Our perception, however, may be deficient in two contradictory directions: we may overestimate the present absolute level of cruelty because of our lack of knowledge of past cases; alternatively, we may underestimate the actual level because we are too impressed by the countervailing forces, especially since the recent expanding human rights awareness has led to a higher state of general alert and perhaps to more deterrence. In relative terms, the question is very different: given that an equal amount of cruelty probably leads to greater destruction in modern times, were societies more repressive or more human towards their citizens in the past than they are today? In tackling this question, one ought to apply criteria for measuring the degrees of cruelty, but which ones? The number of victims in relation to the total population? The presence of especially vulnerable categories of victims? The occurrence of particularly cruel types of repression? In addition to this difficulty, there is the danger of anachronistically projecting back contemporary conceptions and values into the past. In the

face of the pain suffered by each individual victim, however, such considerations on the moral economy of death may sound superficial or scornful. A probable absolute increase of cruelty or repression in history, then, does not imply that humanity is now more barbarous than in the past, nor that it is less.

Does a historical approach shed more light on the problem? The 1948 United Nations Genocide Convention called genocide 'an odious scourge' that 'at all periods of history (…) has inflicted great losses on humanity.' Thus the term 'genocide'—coined in 1944 by Raphael Lemkin—seems to cover phenomena in all historical periods. Which phenomena? The Convention provides a definition (a series of 'acts committed with intent to destroy, in whole or in part, a national, ethnical, racial, or religious group') that genocide scholars reluctantly accept or decline all together. These scholars argue that many more group types than the ones from the definition have been genocide victims (notably political and social groups) and debate the boundaries between genocides and comparable crimes. They also report that many researchers suffer from psychological barriers in studying the almost unbelievable cruelty of genocide. The body of genocide research only began to grow after 1945, and with the exception of the relatively well-studied Holocaust, it is still lacking in many instances.

Not only do definitions and psychological resistance raise problems, so do heuristics and epistemology. It is often excruciatingly difficult to know how many died and what exactly happened during each specific genocide. Drawing statistics on the number of victims that are more than 'guesstimates' is a skill in its own right. Rudolf Rummel's elaborate statistics (1994) show that the democide rate (a concept including genocide, politicide, and mass murder but excluding war dead) was almost certainly higher in the twentieth century than in any previous century: he presents a calculated total of 169,202,000 dead for the period 1900–87 (including 38,566,000 genocide victims) vs. a hypothetical total of 625,716,000 dead for the period from the thirtieth century BCE to 1900. However, these figures do not allow conclusions about a possible increase of democide, and by implication of genocide, because Rummel derived his hypothetical total of pre-twentieth-century democide from the twentieth-century democide rate and the population for each century since the thirtieth century BCE. Linked to the statistical problem is the source problem: few sources normally survive, at least for the earlier genocides, in history. The sources were either not kept, hidden, lost, or destroyed. When they are available, they come from either perpetrators, victims, or bystanders. When there are sources from two or all parties, they frequently conflict with each other. Finally, their authenticity and reliability have to be tested. In short, the usual techniques of historical criticism have to be applied with unusual sharpness. Even when sources allow the historian to distill a picture of the killings, proof of the intention of the perpetrator to destroy a group as such (an essential component of all genocide definitions) is difficult. Openly stated motives are exceptional and notoriously unreliable and, therefore, intent generally has to be ascribed. This is precarious, especially where the genocide is unanticipated.

Once we have reliable descriptions of the events, the question remains whether they catch the essence of the horror. The limits of representing the unrepresentable have been much debated in recent decades. The analysis and interpretation of the data pose their own problems. First, genocide is a complex interdisciplinary phenomenon with historical, psychological, sociological, political, legal, philosophical, and religious aspects. Second, the systemic and human factors conditioning genocides and the extent to which they are unique and comparable are open to much controversy. Third, it is important to know whether the nature of genocide has changed over time. Among the dozen or so typologies of genocide, some address this problem, however inconclusively. Frank Chalk and Kurt Jonassohn, for example, distinguish between retributive genocide (to eliminate a real or potential threat), despotic genocide (to spread terror among real or potential enemies), developmental genocide (to acquire economic wealth), and ideological genocide (to implement an ideology). They then classify 22 cases from history under the most relevant type. The distribution of cases reveals that the fourth type is a predominantly twentieth-century phenomenon with few historical antecedents. Even if this suggests that the nature of genocide has (partially) changed, the question remains why. With all its methodological difficulties summarily depicted above, the historical approach is obviously far from conclusively answering the initial question either.

2.2 Ethical Problems

The historical study of human rights and their abuses also involves ethical problems. Some of these are related to the researching historians, others to the human beings they study. Most are not unique for this type of research, but they present themselves cogently here. Human rights research can be dangerous, especially when the historians are working on contemporary material in their own society or on cases of the past which suggest easy analogies with the present or endanger the legitimacy of present rulers. This may lead to the persecution of the historians and entail reflection on their professional ethics: what are the rights and obligations of historians working under such restrictive conditions and what should their colleagues working under freer conditions do?

What rights do the human beings under scrutiny have, apart from the right to be treated with respect in historical works (see above)? With regard to the perpetrators of crimes, the question is whether moral

judgments of their acts are possible, and if possible, permissible, and if permissible, obligatory. Opinions of historians widely differ on these points. Few believe that they are judges for eternity, but nevertheless most will admit that there is something like a right to historical truth, for which it is never too late. They would probably subscribe to the professional obligation not to deny or forget the fate of the dead and surviving victims. This justified claim, however, can be at odds with impartial research. It is a task with many risks. Without the passion of the survivors, historians may 'normalize' the cruel abuses of the past by inserting them into the stream of history. They may omit crucial findings for fear of breathing new divisive fever into the collective memory. But still, the distinctive professional contribution of historians to the protection of human rights, if any, is to see that the dead do not die twice.

See also: Cultural Property and Ancestral Remains, Repatriation of; Ethnic Cleansing, History of; Ethnic Conflicts; Genocide: Anthropological Aspects; Genocide: Historical Aspects; Holocaust, The; Human Rights, Anthropology of; Human Rights in Intercultural Discourse: Cultural Concerns; Human Rights: Political Aspects; War Crimes Tribunals

Bibliography

American Anthropological Association Executive Board 1947 Statement on human rights. *American Anthropologist* 539–43

Amnesty International 1961–*Report*. Amnesty International, London

Andreopoulos G J (ed.) 1994 *Genocide: Conceptual and Historical Dimensions*. University of Pennsylvania Press, Philadelphia, PA

Castberg F 1968 Natural law and human rights: an idea-historical survey. In: Eide A, Schou A (eds.) *International Protection of Human Rights*. Almqvist & Wiksell, Stockholm

Chalk F, Jonassohn K 1988 The history and sociology of genocidal killings. In: Charny I W (ed.) *Genocide: A Critical Bibliographic Review*. Mansell, London

Chalk F, Jonassohn K 1990 *The History and Sociology of Genocide: Analyses and Case Studies*. Yale University Press, New Haven, CT

Donnelly J 1982 Human rights and human dignity: an analytic critique of non-Western conceptions of human rights. *The American Political Science Review* **76** (whole issue)

Hill Ch R 1969 Some philosophical problems about rights. In: Hill Ch R (ed.) *Rights and Wrongs: Some Essays on Human Rights*. Penguin, Harmondsworth, UK

Human Rights Watch 1991–*World Report*. Human Rights Watch, New York, Washington, London, Brussels

Johnson M G 1998 A Magna Carta for mankind: writing the Universal Declaration of Human Rights. In: Johnson M G, Symonides J (eds.) *The Universal Declaration of Human Rights: A History of its Creation and Implementation, 1948–1998*. UNESCO, Paris

Jones P 1994 *Rights*. Macmillan, Basingstoke, UK

Jongman A J (ed.) 1996 *Contemporary Genocides: Causes, Cases, Consequences*. PIOOM, Leiden, The Netherlands

Kuitenbrouwer M, Leenders M. (eds.) 1996 *Geschiedenis van de Mensenrechten: Bouwstenen Voor een Interdisciplinaire Benadering* [*History of Human Rights: Materials for an Interdisciplinary Approach*]. Verloren, Hilversum

Langley W E 1999 *Encyclopedia of Human Rights Issues Since 1945*. Greenwood Press, Westport, CT

Laqueur W, Rubin B (eds.) 1990 *The Human Rights Reader*. New American Library, New York

Lauterpacht H 1950 *International Law and Human Rights*. Stevens, London

Maris C W 1992 Een Natuurlijke Historie van de Mensenrechten [A natural history of human rights]. In: Cliteur P B, Deelder A, Rutgers M R (eds.) *Mensenrechten* [*Human Rights*]. Gouda Quint, Arnhem, The Netherlands

Rosenbaum A S (ed.) 1980 *The Philosophy of Human Rights: International Perspectives*. Aldwych Press, London

Rummel R J 1994 *Death by Government*. Transaction Publishers, New Brunswick/London

Rüsen J 1993 Human rights from the perspective of a universal history. In: Schmale W (ed.) *Human Rights and Cultural Diversity: Europe, Arabic-Islamic World, Africa, China*. Keip Verlag, Goldbach, Germany

Schwelb E 1968 Human rights. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences VI*. Macmillan, New York

Szabo I 1978 Fondements historiques et développement des droits de l'homme. In: Vasak K (ed.) *Les Dimensions Internationales des Droits de l'Homme*. UNESCO, Paris

United Nations Development Program 1992 Political freedom and human development. *Human Development Report 1992*. Oxford University Press, New York and Oxford, UK

Weston B H 1994 Human rights. In: McHenry R, Hori Y C (eds.) *The New Encyclopaedia Britannica XX*. New Encyclopaedia Britannica, Chicago

A. De Baets

Human Rights in Intercultural Discourse: Cultural Concerns

Though human rights are generally recommended, there are still some strong criticisms. One criticism refers to human rights, somewhat derogatorily, as the civil religion of modernity and, by this, takes them to be only a historical phenomenon of a particular culture. Other critics fear that the spread of human rights is nothing but a new and subtle version of Western colonialism or imperialism, namely a supremacy of a its legal culture, which forces values on other cultures. But actually, human rights belong to a universal legal morality, which is common to all mankind although it was thought of in the most radical manner by Western modernity.

Within Western legal culture, human rights are taken as a matter of course. But as far as their foundations are concerned, there are difficulties which are partly of a political, partly of an intercultural, and partly of an anthropological nature. A specific philosophical justification, a transcendental foundation of human rights might help to overcome

these difficulties. This foundation, at the same time, might permit one to show the universal validity of human rights beyond liberal legal cultures.

1. Human Rights and Basic Rights

The first, political difficulties arise when one considers the possibility that a declaration of human rights might simply remain a declaration. For instance, the People's Republic of China used tanks against demonstrating students, although it recognizes freedom of procession and demonstration as stated in Article 87 of its constitution. The answer to this may simply be compliance with positive law. In order to give human rights an actual power in the world, instead of remaining just word on the page, they have to become solidly embedded in human practice. To this purpose human rights have to make their way into the sometimes written, sometimes unwritten constitution, thereby doing both: obliging the branches of government to human rights and restricting their power, among others rendering possible proceedings before the court of law against any abuse of governmental power and of the nonrecognition of rights. Wherever this occurs, human rights, which were previously just a component of universal legal morality, become elements of positive law. They are transformed from mere human rights into basic rights of a particular legal community.

Not only does this differentiation of human rights and basic rights direct itself against simple declarations, it also draws attention to those origins of basic rights which lie before and beyond the state. In particular, history's first declarations are aware of this elementary meaning. In this respect, the French *Déclaration* (1789) follows the Virginia Bill of Rights (1776). Their beginning establishes valid rights at a level which lies before the state. Subsequently, the basis for legitimation is laid for the exercise of political power; Virginia's Article 2 and France's Article 3 state that the people is the origin of all government. It is not until the third step that the function of the state is touched upon; here the limitation of the exercise of political power as well as the positive tasks of the state is mentioned. In that sense, from a philosophical point of view, we have to reverse Benjamin Barber's title: not 'Democracy first, then Rights,' but 'Human Rights first, then Democracy.'

This sequence contains my first thesis: human rights are rights of all human beings raised against all other human beings. From the standpoint of legitimation the claims in question are primarily raised by all men against one another and they are at the same time granted to one another. With regards to human rights, people recognize themselves as beings with equal rights. Positive law and the branches of government are legitimized in relation to human rights only at a second and, more specifically, subsidiary level: they are incumbent upon transforming innate rights in the process of the establishment of a constitution and further legislation into positive law. The scholar of law says, 'rights are guaranteed'; with this he means a secondary accomplishment that is strongly differentiated from a primary accord. In German, we distinguish *gewähren* which is done by men themselves, and *gewährleisten* which is the task of the branches of government: the legislator, the executive power, and the courts.

2. An Intercultural Legal Discourse

The second difficulty with human rights is constituent of the fact that we expect the recognition of human rights from all cultures, but that the discourse on human rights only takes place within one of them, namely the Western culture. For this reason there is a risk that human rights, instead of constituting the core of a universal legal morality, will give away their essence and will thereby decline to the level of an 'export-article of Western culture.' In this sense, the objection of 'ethnocentric thinking' can be raised, for a legal institution like human rights, Western culture claims to be valid for the whole of mankind.

Although itself a product of the West, philosophy lends its voice to non-Western cultures. It does not do so in a now current form of a mutation, an act which turns philosophers into a kind of fundamental anthropologist, into theoreticians of a world society who, like Lyotard (1983), Feyerabend (1987), and Rorty (1989), place themselves in the tradition of American anthropology—one thinks, for example, of Herskovits' 'Statement on Human Rights' (1947)—and who, in their fear of destroying the characteristics of non-Western cultures, support a pure relativism. Instead of getting lost in the byways of fundamental anthropology, and its exaggerated relativism, philosophy engages in an intercultural legal discourse.

This discourse begins, in passing, by invalidating the mentioned fear. What Lyotard describes, somewhat drastically, as the 'tyranny of the universal' is based upon a misunderstanding, namely, the confusion of universality with uniformity as though universal legal principles should be equated with the removal of social and cultural differences. Let us take an example: the equal rights of different religions or that of different languages, from the standpoint of human rights, do not undermine the existence of one of them; by contrast, the principle of equality guarantees coexistence of the different parties.

In the next step, our discourse turns the reproach of ethnocentric thinking into a philosophical challenge, in other words, into a question whose answer is still open: how can a regionally developed institution which is not always taken seriously—we recall that the Virginia Bill of Rights was proclaimed in a slave state—lay claim to an absolute universality? Obviously, such a claim is permitted only on a condition which, from the standpoint of legitimation, is self-

evident but which is all too readily overlooked in the case of human rights. Legitimation must be separated from the historical genesis and evolution. My *second thesis* is this: whoever wants to impute the idea of human rights to other cultures must for once divorce himself from their intellectual history, also from their legal and social history, for instance from the question regarding economic prerequisites. If human rights are truly to exist, then they should do so not for the reason that capitalism, the modern form of the economy, requires a global protection of economic rights. Besides, there are well-known practices which contradict the idea of human rights and are nonetheless compatible which a global protection of economic rights, namely, slavery and human bondage, inequality of man and woman, the lack of rights for children and, last but not least, colonization. Furthermore, the legitimation of human rights also needs to be uncoupled from a certain phase of the modern state's development, namely from Reformation and Absolutism. If innate rights truly exist, then the Western experience can have a thoroughly heuristic significance; but in the end, only culturally invariant arguments have the power of justification.

3. *A Copernican Revolution of Anthropology*

A third difficulty with human rights is expressed in the question: what remains when one considers the separation of legitimation from the historical genesis and abstracts from the particular context of European modernity? The view that the idea of human rights is connected with a certain image of mankind is widespread. To this there is an annexation which not only profiles the matter as far as its content is concerned, but also limits its validity; namely that an image of mankind presents itself as Jewish, another as Christian, a third as Muslim, a fourth as Buddhist, and finally a fifth—a definitive neither/nor!—as atheist. Moreover, the modern European picture of mankind is individualistic and relatively ahistorical, whereas the African image of mankind is determined by a characteristic solidarity of two dimensions: a solidarity with fellowmen and with the ancestors. We can generalize: all specific images of mankind, all views of human nature are therefore culturally dependent.

The consequence for our justification: we even have to rid ourselves of Western images of mankind and evolve a *logos tou anthropou*. In other words, we have to evolve a concept of mankind which, valid for the whole of humanity, is independent of different histories and cultures.

A concept of mankind which is independent of cultural particularities, and therefore interculturally valid, has for a long time be no stranger to Western thought. Since Greek Antiquity, we know two anthropological definitions: men as *zoon logon echon* and as *physei politikon zoon*. Recently, however, skepticism is the rule. If one lends an ear to philosophers like Rorty (1989, p. 44), then skepticism towards anthropology becomes such a matter of course that one may content oneself with only a short aside: 'fruitless by way of anthropology.' But scepticism is not new. Since Rousseau's concept of *perfectibilité* (*Discours sur l'inégalité*, part 1) and Nietzsche's dictum that man is 'more indeterminate than any other animal' ('*unfestgestellter als irgendein Tier sonst*': *On the Genealogy of Morals*, 3rd essay, section 13), philosophy has doubted whether it is possible to fix an essence of man that is independent of history and culture. In the 1920s it was Georg Lukàcs who supports this skepticism which was reinforced later in the 1930s by Max Horkheimer (1935). Their critique was not directed specifically at human rights, but it was partially to blame for the fact that Western Marxism has had such difficulties with this legal institution up to now. Lukàcs (1923, p. 204) sees it as 'the greater danger of every anthropological standpoint,' that it 'causes man to ossify as fixed objectivity and thus pushes dialectics and history to one side.'

The idea of human rights counters skepticism by weakening its claim. According to Aristotle's classical topos, *physei politikon zoon* (*Politics* I, 2), the specifically Greek form of a community, the *polis*, is the only place where man can lead a really meaningful life. Even if Aristotle is right and human potentialities can be fully realized only in a political community, the form of this community remains open. The above mentioned difference between universality and uniformity first becomes important here: human rights favor neither the Greek city state nor the modern form of territorial state nor any other form of political communities.

Anthropological relativization goes even further. In contrast to Aristotle's *Politics*, human rights allow a largely instrumental understanding of the political community. The state does not have to represent what the great neo-Aristotelian philosopher, Hegel, termed '*die Wirklichkeit der sittlichen Idee*' ('the actuality of the ethical Idea': *Philosophy of Right*, para. 257), at least not its full actuality. What a sociologist, a 'pupil' of Luhmann, Willke (1983) terms a 'demystification of the state' remains possible. The idea of human rights admits that the 'political system' establishes no more than premises and framing conditions, whereas true humanity can be sought elsewhere, and then in different ways. (Contemporary philosophy speaks of the difference between the right and the good.)

Where this openness for different ways is lacking and where what counts is a normative concept of man, a concept of the humane man alone, one runs the risk of denying elementary rights to people who are not so human. According to Aristotle, anyone who suffers a fundamental deficiency in terms of rational endowment—in short, anyone who is mentally handicapped—is *physei doulos* and rightly viewed as having lesser rights, in other words, he is taken as a slave. To

prevent this consequence, the idea of human rights is not allowed to favor particular forms of self-realization; with regard to the humane, we need an intentional nonspecification.

The institution of human rights intentionally contains what I shall call a fragmentary anthropology and for that reason alone can be imputed to other cultures. Although *prima facie* a drawback, nonspecification is in fact an advantage: far from being indifferent to more demanding definitions of man, it is, by contrast, open, ensuring that not only different cultures, but, within those cultures, different subcultures and individuals, enjoy the right to be unique, different, and even eccentric.

Nonetheless, this right is not unlimited. Human rights resist a radical pluralization. They alone make possible the very thing that critical intellectuals such as the 'fundamental anthropologists' mentioned earlier are so keen to achieve but which they are unable to situate adequately within their intellectual framework. The plurality of cultures and subcultures can exist only under conditions which are no longer pluralizable. Thus, in the idea of human rights modesty and immodesty enter into a characteristic union. To limit all plurality and to assume that conditions exist which are equally valid throughout history is an immodest aim. On the other hand these conditions make a modest claim, not demanding to present the form of a human life but to be only a condition of a plurality of forms of life and thus abandoning any normative concept of what is humane.

This connection is possible only if we readjust the anthropological perspective; philosophical anthropology has to undergo a conversion which fully deserves the ambitious title of a Copernican revolution. Instead of defining man from the standpoint that has normally been adopted since classical antiquity, that which allows him *eudaimonia*: happiness, self-realization, or a meaningful existence, we dispense with any teleological or normative concept. My *third thesis* is this: human rights are not concerned with perfecting what is humane but with its initial conditions, with elements which deserve the qualification of the relevant rhetoric: as elements inalienable to each person they are his inherent birthright. Here, we find my initial thesis: since we are dealing with conditions of the possibility to be a man, and not a man in a biological specification but as a responsible agent, we can speak of transcendental elements or o(relatively) transcendental interests, relative to (responsible) agency.

Transcendental interests alone leave the task of legitimation underdefined. This much is clear from even a cursory glance at legal history. Since man was regarded by Judaeo-Christianity as having been made in the image of God or as *zoon logon echon* by the Greeks, the West has been familiar with a universal concept of man. Yet, no legal consequences have been drawn from this; both the Jewish-Christian tradition and the Greek one allow slavery and the inequality of women. A correlative phenomenon, familiar since the thirteenth century, is the sort of bill of rights typified, for example, by the Magna Charta Liberatum and the Golden Bull of Andrew II of Hungary. Although these charters were drawn up in states in which both ideas—'the image of God' and 'rational endowment'—were taken for granted, freedoms were not given to man as man. This deficiency remained—not, it is true, in the legal text, but in legal reality—until the time of the earliest declarations of human rights (we remember the case of a slave state proclaiming a bill of human rights).

Indirectly we stumble across a further difficulty here, which, from the point of view of legitimation, is even greater. Man should have a subjective claim to certain transcendental or anthropological interests. A claim can be made where another person has to meet that claim; subjective rights are bound by their very concept to correlative obligations, human rights to corresponding human obligations.

Three factors deserve emphasis here. First, the claims do not exist in an absolute sense; the right to exist does not mean, for example, that one does not have to die at all. Second, the right which, although only relative is not conditional, refers to ones fellow men and demands that they refrain from using force.

In an intercultural legal discourse, an objection arises at this point which is familiar from the liberalism/socialism debate: particular attention is given to civil rights and liberties, which, on the African side, are accused of favoring the individualist picture of man familiar to the West. To the extent that our fellow men are required to perform a service, not even civil rights are based on any individualistic anthropology. A social nature is assumed in an entirely traditional way, only its manner is no longer traditional. According to Aristotle, the *polis* is natural because man is invested with certain social impulses which ultimately produce the *polis* through a series of certain intermediary stages. The anthropology of human rights does not question these social impulses; it extends its horizons and discovers a complication which is the third factor deserving emphasis.

The complication has two consequences. In the first place, what has hitherto been called simply the 'social being' is given a qualifying rider: it becomes a *positive* sociality providing mutual help, with a *negative* counterpart in the form of a reciprocal threat. From an extended social anthropology—let us say 'co-operation plus conflict'—our legal institution singles out only one aspect, the conflict, namely, a threat from one of our equals, a threat, moreover, which affects the anthropological interests; these are threats of truly existential weight. Since these threats come from our fellow men, each individual is both a potential agent and, at the same time, a potential victim.

The anthropology bound up with being both victim and agent recalls Hobbes' dictum, *homo homini lupus*,

and, like that dictum, sees itself exposed to the reproach that it presents a pessimistic image of man. But, at least as far as the idea of human rights is concerned, this reproach is unjustified. Being modest once more, negative social anthropology is not claimed to be exclusively valid but, once again, to be fragmentary anthropology—not a basic feature of the individual but a danger which cannot be excluded.

The complication caused by threat has a second consequence, inasmuch as the evolution of the *polis* can no longer be regarded as the mere development of pre-existent social impulses, a far greater contribution of ones own is necessary; although he is a social being, man has to produce himself as a social being. Society arises through mutual recognition, which can assume a variety of forms; human rights emphasize only a small part, and that part concerns the elementary level of anthropological interests.

The anthropology of human rights relates to existential threats but refrains from passing judgement on the nature of the conflict, whether that judgement be positive or negative. Therefore, a *fourth thesis* reads as follows: the danger of conflict is accepted as an anthropological fact which is favored in relation to man's cooperational nature only to the extent that it is his conflictual nature alone which makes human rights necessary. Where social relations consist only of mutual help and where help is not only always spontaneous but can be relied on, in a word, where man's positive social nature exclusively is involved, human rights lose their task.

4. *Transcendental Interests and Transcendental Reciprocity*

It is the fact that elementary interests are at stake that constitutes the seriousness, even existential seriousness of human rights and explains the feelings and the pathos bound up with this institution. An indispensable element—the character of a claim—is not yet clear, however. In many arguments there is a hiatus at this juncture, a legitimate leap which fails to explain why a subjective claim insists on the recognition of transcendental interests. I do not dare to suggest that I can justify this claim in the face of every doubt, but I am making a proposal based on a generally recognized principle.

A moral claim insists on the recognition of a service where the service is rendered on condition that a corresponding service be rendered in return. This generally recognized principle regards the obligations of legal morality (and not of all morality) as a giving that is made on condition that a countergiving is made; it operates with the reciprocity of giving and countergiving, you could say, with an exchange. Anyone who lays claim to those services which are rendered only under the condition of reciprocity now stands under an *obligation*. Conversely, a (subjective) *right* is involved as long as one actually renders the service which is performed under the condition of a service in return.

In terms of human rights, this general principle must now be applied to the particular case of anthropological interests. For them, there is a correlation between rights and duties where certain interests not only have an anthropological status but also fulfil the condition that they are realized only in and through reciprocity. Here, and only here, in the case of an 'inherent reciprocity,' or an inherent sociality, is one no longer free to say whether to accept an exchange or not. To the extent that, in keeping with their qualification as anthropological, the interests are indispensable and to the extent that they are also bound up with the idea of reciprocity, the indispensability is transferred to the reciprocity; the corresponding exchange is similarly indispensable.

The proposal obtains a moral, albeit undemanding moral element: at the basis of human rights is a modest morality that is withdrawn from the usual controversies. This morality—fairness in exchange—applies in those cases where it is not a question of one person giving and the other taking, but, rather, where there is a reciprocity of giving and taking and where, moreover, the giving and its recompense are more or less equal in value.

Since a moral element is present, the legitimation of human rights obeys the structure 'anthropology plus morals.' In this legitimation the moral part is as good as uncontroversial, but the same cannot be said of the anthropological part. Human rights represent a problem that is very much bound up with legal *ethics*, and yet it is their other aspect, that of legal *anthropology*, which contains the greater difficulties. At least on *their* account what was formerly a great discipline, philosophical anthropology, merits a minor revival.

It is the both-victim-and-agent anthropology which allows us to describe human rights as an exchange which anyone can enter into with anyone else. Each of us exchanges his own potential for committing an act of violence for the interest of not falling victim to another's violence. Characteristic of this exchange is its inherent reciprocity: one can realize one's own interest only by virtue of a service which the others—*all* the others—perform. In addition, we rediscover a difference to traditional social philosophy. Mutuality consists not of positive services but of a negative reciprocity, in an all-round rejection of violence. Nonetheless, the exchange has a positive significance as such, and not only as consequence, otherwise there would be no reason to agree to it. When the corresponding human *obligations* are fulfilled, human *rights* should be granted. From the reciprocal refusal to kill one's fellow man, to rob them, or to prevent them from practicing their religion issue the right of life and limb, the right to property, and the right to religious freedom.

The question which now arises is whether, for the sake of the positive purpose, each of us should agree to this negative exchange. Does the condition 'innate interest' apply just as much as the condition 'inherent reciprocity'? The question has the character of a preferential choice which is decided on the question of what people prefer. Do they prefer to have the right to kill, but coupled with the danger of being killed themselves, or do they prefer the integrity of life and limb, but at the price of having to accept the corresponding obligation and no longer kill any person? Or, put another way: does the individual prefer to be both the victim and the perpetrator of violence or does he prefer a neither/nor situation?

According to a first attempt of philosophical legitimation—that of Hobbes—every man has an overriding desire: not to die at the hands of a fellow man. The English Civil Wars of that time make it clear enough that many people see religious freedom as more important, while others prefer political or cultural ideas, and a third group prefers honor.

The counterexamples can be subsumed under the heading of self-preservation. What is preserved, however, is not the physical self but a 'moral self.' Is it here, therefore, that the predominant interest is to be found? In the conflict of 'physical vs. moral self,' one can certainly decide in favor of mere survival. If there is really an innate interest, it must be sought on a logically higher level, beyond the conflict of 'physical or moral self.'

A third attempt remains with the idea of an overriding desire, but takes that desire not as something negative but, rather, positive, something, moreover, which is no longer a dominant but an all inclusive interest. The interest that includes all other interests within it we call happiness. As it fulfils all needs and interests, it forms the natural guiding principle of every individual. With so formal a definition, however, it is not possible to answer the question, whether a person would rather perpetrate a violent act and at the same time be its victim or whether he would prefer neither one nor the other. If we introduce a further definition, therefore, and insist on defining the content, the individual is left with the possibility of sacrificing his happiness to a fellow man or to a religious, cultural, or political ideal. As a formal definition, happiness is unrelinquishable, but it forms too weak a premise; conversely, a definition which takes account of content reflects a predominant desire which can be seen in relative terms and, like this, represents too powerful a premise.

Anyone who concludes from this that there is no interest in the corresponding exchange fails to see that many people want nothing more than their own survival, while others want only religious freedom, others again only political freedom, and so on. On each occasion, the corresponding rights are advantageous, but they are not the same rights for each person. In other words, people are happy to vote for human rights, but they interpret that vote very differently.

We need to look for a fourth answer which preserves the strengths of the first three attempts while overcoming their weaknesses. On the one hand, we need a stronger premise than a merely formal definition of happiness; on the other, the undue assumption of an absolutely overriding desire must be abandoned; what we are looking for is an interest that is both substantial and nonsubstitutable. My *fifth thesis* is as follows: this double condition is met by interests which one can neither abandon in favor of higher interests nor weigh against interests of equal importance. Truly indispensable—and, therefore, anthropological—interests consist of elements which, logically, are on a higher level and which, *faute de mieux*, are called (relatively) transcendental interests. To the extent that, in our abstaining from the use of force, our interests are realizable only in reciprocity, in other words, in an exchange, one can speak—with the same proviso—of a transcendental exchange or a transcendental reciprocity. Human rights refer to conditions of agency based on reciprocity.

The idea of human rights presupposes that there are indeed interests on this higher level and that there is, therefore, an exchange on a logically higher level. Let us take an intentionally uncontroversial example—the protection of life and limb. Even someone not especially attached to life has an interest—conscious or unconscious—in that life, since he can otherwise neither desire anything nor seek to satisfy his desire. Independently of what one strives to achieve or seeks to avoid in terms of life's content and, therefore, a condition of our ability to act, a condition of agency, life is the precondition of a practical desire. It is at this point that the singular nature of human rights emerges most clearly, without the corresponding acts of renunciation, the very thing that distinguishes man—his agency—is not capable of existing within a social perspective. This is true even of those who are prepared to sacrifice their own lives: the martyr wants to decide in what cause he lays down his life, in other words, he acts out of religious or political conviction and not, say, because of some robber's act of violence.

The natural law of the Enlightenment appealed to the principle of self-preservation. If I consider the principle of agency to be more convincing, it is because it is more formal and, therefore, has fewer preconditions attached to it. When I speak of agency, it is not, of course, an ability to act *sans phrase* but simply that part which involves reciprocity.

5. *Against a Liberalist Restriction*

Man is not merely a living being. More than life and limb are involved in the indispensable conditions behind his agency; in consequence, there is more than

just this one particular human right. We can argue directly on the basis of our agency. Since the concept of action involves intentionality, the conditions for agency also include the conditions of intentionality, in other words, the ability to speak and think. Furthermore, in accordance with the claim that man cannot become a man but within a society of fellow men, agency is bound up with positive social relations and cooperative conditions, thereby opening up an additional field of transcendental interests. (We find the traditional anthropological definition of man: he is *zoon*: an animal, *logon echon*: endowed with reason and language, and *zoon politikon*, a social being.)

My proposal: the majority of human rights can be reconstructed on the basis of these three aspects, which, taken together, make the ability to act or agency, a real possibility, in other words, the living being, the being who is able to talk and to reason, and the cooperational being. The insight contained herein, namely, that there are further transcendental interests, is not of course sufficient to proclaim corresponding human rights. In order to do that, we also need to prove a transcendental reciprocity.

The politics of recent decades have grown sensitive to other aspects, including questions of personal, social, and cultural identity, of economic and social prerequisites and, not least, ecological conditions. Here, too, there are forms of reciprocal threat for which our legitimation model remains applicable. At all events, we are dealing with new forms, with threats which are both indirect and which involve a collective responsibility, so that the legitimation model needs to be developed further. But even here it can be stressed—and this is my *sixth thesis*—that the last word has not yet been spoken on human rights within the classic, liberalist canon.

According to classical liberalism each should know his interests best; but since Nietzsche and Freud this view has prevailed as naive. The concept of transcendental interests reaches far beyond this controversy. Within elementary interests, transcendental refers to a stratem that we have not only independent of attitude but also independent of being man or woman, Christian, Muslim, or nonbeliever. The singular character is due to the fact that one does not want anything else but the following: to be an entity capable of action.

According to Rorty (1989, p. 28), the thought that the belonging to a biological species implies certain rights is a type of metaphysical consolation, which the pragmatist renounces. The preceding legitimation shows on the one hand that it is not the biological belonging that establishes rights, but rather the interaction of transcendental and moral arguments; and on the other hand, that for a metaphysical consolation not only the pragmatist but also the theoretician of human rights can renounce.

A legal order cannot safeguard human rights without governmental powers. My *seventh thesis*—a thesis which is itself subdivided—argues against anarchism. According to this thesis, a public power is needed first to provide a more detailed definition of human rights; this is the task of the legislator. But even with more detailed definitions there are controversial interpretations, so that an unpartial and authoritative interpretation, i.e. an independent judicial system has to settle the relevant controversies. Finally, there is another complication, and once more a complication in the form of a danger: one can vote for the recognition of rights and still attempt to enjoy the advantage of a mutual renunciation of freedom without paying the price of renouncing of one's freedom. It is above all in those instances where countergivings do not occur at the same time as the giving that one can enter into an exchange of freedoms in the manner of a fare dodger. This possibility reveals a further element of negative social anthropology, namely, the danger of parasitically exploiting a universal advantage. Public powers are demanded a third time, therefore being needed here as an implementational force—a kind of 'sword of justice.'

Even if more detailed considerations are needed for a convincing legitimation (see Höffe 1987, Chaps. 13–14), this hypothesis of differentiation appears nonetheless plausible. Looked at from the standpoint of the institution of human rights, the classical—specifically Aristotelian—idea of a political nature develops as two partial natures, one of which is concerned with the law, the other with the state. Man is a being that needs, on the one hand, elementary rights and, on the other, the public definition and implementation of those rights, and in this context, he needs democracy. By consequence: rights first, but in any case, democracy, too. Against an overrating of the state, this aspect is secondary, however, being subsidiary to the primary task, which is that, in order to be a human being among his fellow men, man needs elementary rights and correlatively elementary obligations.

See also: Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations

Bibliography

Feyerabend P K 1987 *Farewell to Reason*. London [dt. *Irrwege der Vernunft*, Frankfurt/Main 1989]

Höffe O 1987 *Politische Gerechtigkeit. Grundlegung einer kritischen Philosophie von Recht und Staat*. Frankfurt/Main [Political Justice, Cambridge/Oxford 1995]

Herskovits M J 1947 Statement on human rights. *American Anthropologist* **49**: 539–43

Horkheimer M 1935/1988 *Bemerkungen zur philosophischen Anthropologiein: Gesammelte Schriften 3*. Frankfurt/Main, pp. 249–76

Lukács G 1923 *Geschichte und Klassenbewußtsein Malik-verlag, Berlin*. [*History and Class Consciousness: Studies in Marxist Dialectics*. London 1971]

Lyotard J F 1983 *Le différend*. Editions de Minuit, Paris [*The Differend: Phrases in Dispute*, Manchester 1989]
Rorty R 1989 *Contingency, Irony and Solidarity*. Cambridge University Press, Cambridge and New York
Willke H 1983 *Entzauberung des Staates: Überlegungen zu einer sozialen Steuerungstheorie*. Athenäum, Königstein, Germany

O. Höffe

Human Rights: Political Aspects

Human rights are, literally, the rights that one has simply as a human being. Although the fate of human rights is largely determined by national political struggles and practices, the focus of this article is international, because only in international studies has human rights become a well-established 'mainstream' concern of a social science discipline.

Prior to World War II, human rights was an ignored, even suppressed, issue in international relations. Even the notoriously 'idealistic' Covenant of the League of Nations does not mention human rights. The turning point was the creation of the United Nations, in the shadow of the Holocaust. The UN Charter lists promoting human rights as a principal objective of the Organization. Even more importantly, the Universal Declaration of Human Rights, adopted by the General Assembly on December 10, 1948, established substantive international human-rights norms that have come to be accepted, at least in word, by almost all states and most leading movements of political opposition in all regions of the world.

In the first two decades after the Universal Declaration, however, both superpowers used international human rights largely as an instrument of partisan political abuse. Although decolonization helped to spur relatively nonpartisan international attention to self-determination and racial discrimination, through the early 1970s no prominent state pressed international human-rights concerns except in the presence of strong political interests.

Rights that are given, either by 'benevolent' dictators or foreign powers, usually end up being rescinded. Only when right-holders are able to, and at least periodically do, act to vindicate their rights are they likely to be enjoyed securely. Thus foreign states and international organizations, as well as transnational nongovernmental organizations (NGOs), have increasingly focused their efforts on supporting the development of autonomous, intermediate associations in civil society, in recognition of the fact that international action usually can only be a supplement to national action.

The other major change of the post-Cold War era has been the growing integration of human rights into other areas of the international action. For example, at the UN during the Cold War, human rights were rigorously segregated from peace and security, and development activities. Today, a human-rights mission is a standard part of peace and security operations. Furthermore, the threats to peace and security to which the UN responds are most likely to arise from systematic human-rights violations.

The penetration of human rights into other domains of international action is far from complete. Most notably, although the United Nations Development Programme stresses the links between human rights and development, human rights have made almost no headway in the work of the International Monetary Fund, which regularly imposes structural adjustment programs that have significant negative impacts on economic and social rights. Nonetheless, at the start of the twenty-first century human rights are increasingly seen as not merely a regular part of international relations but a concern with vital rights enumerated in the Universal Declaration. Thus the Vienna Declaration, adopted by consensus by 171 participating states, in its first operative paragraph 'reaffirms the solemn commitment of all States to fulfill their obligations to promote universal respect for, and observance and protection of, all human rights and fundamental freedoms for all.'

As human-rights values have filtered into the consciousness and practice of states, bureaucrats, citizens, and civil societies across the globe, the bases of national and international political legitimacy have been transformed. The Vienna Declaration, again in its first operative paragraph, proclaims protecting and promoting human rights 'the first responsibility of Governments.' And in sharp contrast to half a century ago, governments today can expect to be reminded, often loudly, of their human-rights responsibilities by foreign states, international and regional organization, the transnational human-rights movement, and, to the extent that they permit it, their own citizens.

The global human rights regime, however, is based on *national* implementation of international human-rights norms. Norm creation has been thoroughly internationalized: the UN Commission on Human Rights drafts treaties and declarations through a process that emphasizes consensus. Implementation, however, lies almost exclusively with sovereign states.

Although the UN Security Council has asserted its enforcement powers in a handful of cases, multilateral human-rights institutions have monitoring rather than enforcement powers. The UN Commission on Human Rights, in addition to its vital work in developing international norms, investigates violations in particular states and substantive issues of broad international concern. The Human Rights Committee (HRC) and the Committee on Economic, Social, and Cultural Rights, which supervise state compliance with the 1966 International Human Rights Covenants, are restricted principally to reviewing periodic reports submitted by states. (The HRC also investigates

individual complaints, but these procedures have been used infrequently—less than 250 decisions in more than 20 years—and its findings have no legal force.) Supervisory committees created under other international human-rights treaties—dealing, most prominently, with racial discrimination, women, torture, and children—are also largely restricted to reviewing and commenting on periodic state reports.

Such reporting schemes have considerable value. In conscientiously preparing a report, states often identify areas where improved performance is possible. At minimum, states know that they are subject to periodic scrutiny by an impartial international body. But these supervisory committees cannot even order, let alone, compel states to alter their practices. Multilateral action thus is most likely to bring change not where violations are most extreme, and thus 'most needed,' but in countries where violations are less severe and where governments have a significant internal motivation for change.

The principal exception is genocide, where a system of international enforcement (although not prevention) does seem to be emerging. The Statute of the International Criminal Court, adopted at Rome in 1998, establishes individual criminal responsibility for genocide, crimes against humanity, and war crimes. Furthermore, antigenocide norms are increasingly enforced through armed humanitarian intervention (e.g., Bosnia, Rwanda, East Timor). But for all other human-rights violations—that is, all but a tiny (if important) handful of violations—multilateral institutions monitor state practices rather than enforce international norms.

The powers of regional human-rights organizations vary immensely. At one extreme, the European Court of Human Rights provides authoritative judicial enforcement of the 1950 European Convention on Human Rights for the members of the Council of Europe. At the other extreme, there are no regional human-rights institutions for Asia and the Middle East. The Americas lie about halfway in between. Members of the Organization of American States (OAS) are subject to investigation by the Inter-American Commission on Human Rights, but the Inter-American Court of Human Rights has been much less active, and considerably less effective, than its European counterpart. The African (Banjul) Charter of Human and Peoples' Rights has created only a very weak commission and no court.

In looking at international implementation, we also must recognize that states, for all the talk of globalization, remain the central international human-rights actor, not merely through their control of international and regional organizations but in their bilateral relations. But they still typically apply only the weaker of their available means—praise and protests, symbolic actions, and occasional financial incentives and penalties—in the pursuit of human-rights objectives. No state, weak or strong, 'nasty' or 'nice,' has been willing to sacrifice vital national interests, or even very much in the way of economic, security, or political interests, to human rights goals.

But states also lack the authority to use force unilaterally against international human-rights violators. And whether undertaken by states or regional international organizations, sanctions short of direct military intervention, as the embargos against Cuba, Iraq, and Serbia suggest, typically inflict far more harm on innocent victims than the targeted regimes. Furthermore, even where external coercion ends abuses or topples a repressive regime, this usually is only a first step toward establishing a rights-protective regime, as Somalia and Bosnia suggest. Thus even where external actors are willing to use coercive means to pursue their human-rights objectives, it is not clear that they have the capabilities to have a decisive positive impact.

This suggests an unusually important role for transnational human rights NGOs. They have been, and remain, important transmitters of international human-rights norms. They have a single-minded focus on human rights, in sharp contrast to states and international organizations, which must balance a range of often competing interests. And they increasingly understand their role as supporting the local human rights advocates and progressive elements in 'civil society' that, along with local governments, hold the key to the future of human rights.

Four events of the mid-1970s signaled the maturing of human rights as an international issue. The 1973 coup in Chile provoked a major and sustained international reaction that for the first time cut across Cold War divisions, with many European states joining the criticism of the US-backed military dictatorship. The Helsinki Final Act of 1975 introduced human rights into the mainstream of East–West relations. The award of the Nobel Peace Prize to Amnesty International in 1977 marked the emergence of transnational NGOs as significant international human rights actors. And when Jimmy Carter became president of the United States in 1977, one of the world's two leading powers began to raise human rights issues with allies as well as enemies.

Human rights violators continued to meet criticism with vigorous appeals to sovereignty and nonintervention. States rarely went beyond verbal and symbolic action in support of their human rights objectives. The moral power of international and nongovernmental organizations to mobilize shame often ran up against shameless governments. Nonetheless, by the early 1990s the international human-rights landscape had been significantly altered.

At the 1993 World Human Rights Conference in Vienna, the Cold War ideological wranglings over the relative priorities of economic, social, civil, and political rights, and parallel North–South debates over the relationship between human rights and development, were largely discredited. Although national

or regional culture continues to be used to justify substantial derivations from international norms, such arguments are generally rejected for the linkages to a wide range of other national and international interests.

We can see a comparable maturing of human rights as an area of scholarly inquiry. Just twenty-five years ago, one could count on one's fingers the number of social scientists who specialized in human rights. Through the end of the Cold war, the scholarly literature on international human rights was largely legal, institutional, and policy oriented, as well as poorly integrated within the leading disciplines. In recent years, human rights has moved into the mainstream of political science and international studies, at least in the English-speaking world. And we are beginning to see a new generation of scholarship that increasingly adopts mainstream disciplinary methods. Thus this encyclopedia appropriately includes an entry on human rights, whereas a comparable project just a quarter century ago almost certainly would not have.

Bibliography

Alston P (ed.) 1995 *The United Nations and Human Rights: A Critical Appraisal*. Clarendon Press, Oxford, UK

Cortright D, Lopez G 2000 *The Sanctions Decade: Assessing UN Strategies in the 1990s*. Lynne Rienner, Boulder, CO

Donnelly J 1989 *Universal Human Rights in Theory and Practice*. Cornell University Press, Ithaca, NY

Donnelly J 1998 *International Human Rights*, 2nd edn. Westview Press, Boulder, CO

Dunne T, Wheeler N J (eds.) 1999 *Human Rights in Global Politics*. Cambridge University Press, Cambridge, UK

Forsythe D P 2000 *Human Rights in International Relations*. Cambridge University Press, Cambridge, UK

Hannum H 1992 *Guide to International Human Rights Practice*, 2nd edn. University of Pennsylvania Press, Philadelphia, PA

Risse T, Ropp S C, Sikkink K (eds.) 1999 *The Power of Human Rights: International Norms and Domestic Change*. Cambridge University Press, Cambridge, UK

Shue H 1980 *Basic Rights: Subsistence, Affluence, and US Foreign Policy*. Princeton University Press, Princeton, NJ

United Nations Development Programme 2000 *Human Development Report 2000: Human Development and Human Rights*. Oxford University Press, New York

United Nations Human Rights Website, http://www.unhchr.ch/hchr.un.htm

Wheeler N 2000 *Saving Strangers: Humanitarian Intervention in International Society*. Oxford University Press, Oxford, UK

J. Donnelly

Human Sciences: History and Sociology

Reference to 'the human sciences' is loose and variable. With few exceptions, there is no correspondence between the term and existing disciplines. Indeed, the term has different connotations for different academics. For some, it refers to what they hope will be a field constituted by debate about what sort of systematic knowledge of human beings is possible, not by a particular body of knowledge. By contrast, others think the field is simply the biology of human beings. Most commonly, the term is one of convenience: 'the human sciences' covers the past and current pursuit of knowledge of human activity, without specifying precisely what is included. The term may cover psychology, linguistics, sociology, anthropology, economics, human geography, and political science, and also parts of the organizational sciences, history, biomedicine, cognitive science, neuroscience, and other areas as well. Nevertheless, it is increasingly common, and significant, for the term to signify a cluster of debates about knowledge and the achievements—or limitations—of existing disciplines.

1. Comparative Meaning and Usage

The difficulty of providing precise definition comes from the subject. 'Man' is not an empirically delimited subject. Theoretical argument about how to characterize the subject, rather than empirical investigation of a given domain, is central. At the same time, empirical research from areas of activity as divided from each other as evolutionary biology and cultural studies have considerable influence. The content of the field is variously said to include the biological human species, the 'Being' of 'Man,' language itself, socially and historically constructed culture, practices of ordering human life, and so on.

When the term refers in an open-ended way to knowledge of humans, usage relates to the '*sciences de l'homme*,' a phrase denoting a conventional division of Francophone academic life covering the psychological and social sciences. There is also overlap between the human sciences and the '*Geisteswissenschaften*,' the sciences concerned with the activity of the human mind or spirit. As German-language usage indicates, the meanings of the word 'science' itself enter argument about the human sciences. In Continental Europe, a disciplined, rational body of knowledge forms a 'science' (theology, for example, may be referred to as a human science), while in the Anglophone world, 'science' is often shorthand for 'natural science.' Except in the usage of some natural scientists, reference to the human sciences follows Continental European usage and includes humanities disciplines as potential 'sciences.' Looked at historically, the human sciences have roots in the eighteenth- and nineteenth-century 'science of man' and 'the moral sciences.' These English-language expressions referred to areas of knowledge thought to be distinctive by virtue of human reason, language, morality, and society.

In the social sciences, it is becoming common for 'the human sciences' to include disciplines that are not subsumed by the natural sciences and that share in

common the goal of understanding human activity and its products. Institutionally, these disciplines often form social science faculties, though they may also spread into natural science faculties (e.g., psychology) or humanities faculties (e.g., history). Reference to the collective term is more than a matter of convenience. Deep and complex issues are at stake. One view holds that debate about the nature of knowledge, or 'science,' of the human subject is intrinsic to being human; 'human science' is therefore an essentially contestable area. The long history of disagreement about the identity, content, and methods of all aspects of the human sciences supports this view. This disagreement continued in the late twentieth century, especially in controversy about the respective contribution of the natural sciences and humanities.

But sometimes scholars have held the view that one particular science provides the foundation for a unified human science. Around 1800, French intellectuals argued on behalf of '*idéologie*,' a science of the origin of action and ideas in sensation and organization. In the USA in the 1950s, there was an attempt to fashion 'behavioral science' as unified human science; more recently, some scientists have promoted evolutionary biology as the basic science of human existence.

2. *The Evaluative Connotations of Usage*

Earlier encyclopedias of the social sciences contain no reference to the human sciences. Recent usage reflects changing arguments and values in the academic community. First, a number of scholars welcome the interplay of different approaches, and do not think disagreement is a failing of science; according to some, disagreement is intrinsic to knowledge of humans. Second, they argue that the natural sciences do not, and cannot, provide explanations for much of the human sphere. This point is made because natural science knowledge does not address certain topics (e.g., symbolization, and meaning), not because natural science is thought undeveloped.

Third, many scholars hope to establish new relations between the psychological and social sciences and the humanities (that is, disciplines like art history and literary theory). Fourth, an interest in the human sciences often indicates sympathy with the view that history of belief in human realities and their representation is necessary to knowledge. Finally, usage expresses conviction that the existing social science and humanities disciplines are historical, social constructs, which structure knowledge in contingent and possibly restrictive ways. Deeper knowledge may need an overarching conception, signaled by reference to 'the human sciences.'

The spread of the term therefore correlates with the critique of positivist views of the social sciences and of 'behavior' as a unifying concept, and interest in rethinking the relations of history and scientific knowledge and overcoming disciplinary specialization in the psychological and social sciences. Some observers see correspondence between these values and the spread of a postmodern ethos. The welcome given to open-ended argument, skepticism about the unification of knowledge, and turn to the examination of 'science' using theories and methods developed in other areas of academic culture, does suggest that spreading interest in the category, 'the human sciences,' is part of a wider change.

Yet, there are some scientists for whom the term 'the human sciences' simply describes the extension of the natural sciences to encompass the human sphere. This usage is rejected by most academics. They do not deny that human beings are part of nature (and hence a subject for natural science), but they maintain that human activity is self-created through historical processes of reflection and language (and hence that there are non-natural science forms of objective knowledge). There are major differences of emphasis among those scholars who have no pretension of creating a unified human science. Some, such as the philosopher Charles Taylor, may be called 'humanist.' They have argued for the distinctiveness of the human sciences because of values thought intrinsic to human activity. This orientation has been critically dissected in 'the linguistic turn,' a cluster of arguments about language or representation—rather than 'nature,' 'Being,' or 'Man'—as the subject matter of knowledge. Scholars such as Michel Foucault may be called 'antihumanist,' as their projects deny reference to anything 'essentially human,' and instead consider only the conditions of possibility for knowledge at a particular time. Yet others, such as the philosopher of science, Ian Hacking, find, ways to hold in tension this polarity while opening new views of the human subject.

3. *History of the Human Sciences*

The modern disciplines of psychology and the social sciences are creations, for the most part, of the late nineteenth and early twentieth centuries. Their establishment is part of the modernizing process, the recreation of human activity as subject to rational discipline, evident at least from the early modern period in Europe. During what historians call 'the Enlightenment,' there was a deeply influential attempt to extend knowledge of laws of nature to include individual and social activity. Scholars worked in a variety of settings, outside as well as within the academy, and according to many intellectual agendas. Giambattista Vico, whose '*scienza nuova*' encompassed myth, poetry, and history as human creations, was a professor of rhetoric at the University of Naples. David Hume wrote 'of human nature' as a literary essayist, while his friend Adam Smith taught about 'the wealth of nations' in courses on moral philosophy at the University of Glasgow. Montesquieu was a French aristocrat whose comparative study of legal systems, government, and custom enhanced notions of

the relativity of social life, while J. G. Herder, a civil servant in Saxe-Weimar, linked the fate of peoples and the fate of culture and language.

There was, however, no reference to 'the human sciences' but rather to 'the science of man.' Historians use the former term to encompass the background for all the intellectual activity divided since the late nineteenth century among the modern disciplines. The philosopher of the social sciences, Peter T. Manicas, for example, claimed that political economy was the first 'scientific' human science. Historians of 'the human sciences' distinguish themselves from historians of psychology, economics, or other disciplines, by stressing the historically relative classification of knowledge and practice.

In the nineteenth century, G. W. F. Hegel, Auguste Comte, Karl Marx, Herbert Spencer, and Charles Darwin each brought human existence within the scope of a unified 'science.' Each attempted to make human existence intelligible through historical forms of explanation. The very generality and scope of their work, however, contributed to the major revisions associated with discipline formation in the social sciences. Speculative historical reconstruction integrated poorly with the disciplined search for precise, empirical knowledge, and with the knowledge generated in the context of social problems.

The outcome was a commitment to differentiating professional and amateur, to institutionalizing research and training within universities as disciplines, and to separating what was held to be science from application. There was debate about the foundations of knowledge and relations between knowledge of nature and knowledge of culture. J. S. Mill, under the heading of 'the moral sciences,' stated the case for explanatory continuity, though he thought that an enormously complex task lay ahead to explain social life by laws of human nature. In the German- and French-language universities, there were arguments, still cited at the end of the twentieth century, about the need for hermeneutic, or interpretative, understanding of human actions; about the distinction between nomothetic (or law-like) and idiographic (or particular) explanations; about structural versus historical explanation (especially in economics and linguistics); and about how values relate to science. The participating scholars—such as Wilhelm Dilthey, Gustav Schmoller, Emile Durkheim, Ferdinand de Saussure, and Max Weber—give the human sciences many of their primary reference points.

In the English-speaking world, especially in the USA, the outcome was the establishment of empirically based disciplines, often accompanied by belief that this followed natural science example. The US model of professional research became dominant almost everywhere after 1945. For many researchers, this carried forward the enlightenment vision of improving human life on the basis of objective knowledge. But only when there were hopes for a more unified and more effective social science was there discussion of a global project comparable to 'the science of man.' In the interwar years, the Rockefeller Foundation supported a science of personality and society. The Ford Foundation's entry into the social science field in 1952 fostered 'the behavioral sciences.' A commitment to unification was matched at the level of national politics by a perceived 'moral lag,' belief that natural science and technology had outstripped the political and moral capacity to order human affairs.

Reference to 'the human sciences' dates mainly from the 1980s. Weaknesses in both Marxist and empiricist philosophies of social science led to an historical as well as philosophical re-examination of theory. Developments in the humanities and social sciences directed empirical research to the 'representation' of knowledge and power rather than to questions, customary in the empiricist mainstream, about objective truth and material reality. By contrast, some biologists concurrently claimed that their field was uniquely competent to understand human beings. Thus, different groups, with different interests, found it useful to promote the same generic term. The only English-language term available was 'the human sciences,' a term preferred to 'the science of man' in the English-speaking world, as it appeared neutral in relation to gender.

The social circumstances were not propitious for new large-scale institutional arrangements, and the number of scholars who identified themselves as primarily occupied with the human sciences, as opposed to occupied in one of the existing disciplines, remained small. Nevertheless, some organizations, particularly emphasizing history, appeared: a journal, *History of the Human Sciences* (founded 1988); the European Society for the History of the Human Sciences; the *Société Française pour l'Histoire des Sciences de l'Homme*; and a Forum for the History of the Human Sciences within the History of Science Society in North America. When institutions or conferences use the term 'the human sciences,' it tends to be a utilitarian response to local organizational needs, bringing under a common administrative, and sometimes research, umbrella such disciplines as anthropology, applied social science, and social psychology.

4. *Theoretical Content*

If we take the view that the content of the human sciences is open-ended, the field has the following character. First and foremost, the field does not presuppose what the human subject *is*, but seeks to analyze and understand the historical and epistemic conditions of human activity at any particular time and place. There is therefore a literature about the representation of human action and human nature—

in practice as well as in written, visual, or other sources—and about such basic categories as 'economy,' 'modernity,' and 'the self.' The field, understood in this way, does not presuppose evolutionary biology or, in common language, 'human nature,' as the subject of objective knowledge, but rather seeks to understand how what is accepted to be objective knowledge is created through historically or humanly constituted processes. The humanities disciplines have a marked influence, encouraging interest in such topics as symbolic representation, rhetoric, and identity formation, whether of individuals, gender, ethnicity, or culture.

The field overlaps with 'the history and philosophy of the social sciences.' But whereas earlier English-language philosophy tended to take for granted the empiricist agenda of social science, and examined, for example, the nature of law-like explanation, realism, and reason and agency, the new field does not. It is more concerned with the constitution of agendas in the first place, and hence examines topics such as the influence of disciplinarity on knowledge, power and representation, and the 'given' quality of such basic concepts as 'society' and 'emotion.' Empirical studies of particulars, rather than abstract theorizing, often suggest the form of analysis.

Recent history and sociology of science (i.e., of natural science) developed independently, for the most part, of questions specific to the psychological and social sciences. But, of course, the sociology of scientific knowledge and technology is itself a major contribution to, and part of the history of, the social sciences. Some scholars, like Kurt Danziger, have extended the sociology of knowledge to the psychological and social sciences. More commonly, however, historians have informally employed the new sociology of science to explain activity in the social sciences in terms of specific factors of time and place—that is, to provide 'contextual' explanations.

Much research in the human sciences expresses general shifts, like the increased stress on 'reflexivity' and 'the linguistic turn,' in intellectual culture. The former term has multiple uses, but in the most general sense it denotes the activity of reflecting—'reflecting' in the double sense of making representations of existence and re-creating existence through the process. The most obvious instance is language: language both represents and constitutes the world as it is in human activity. Argument about the impossibility of going 'beyond' language to the world, society, or human nature encouraged research focused on language (or other signs, understood as forming a language) as the content of the human sciences. This is 'the linguistic turn.' What such claims and shifts imply is subject to intense, and often esoteric, debate. But it is an implication that a field built around such debate is open-ended in its scope and prospects.

Influential discussions derive from Foucault. Foucault, at one stage, claimed that the human sciences are part of a modern 'episteme,' or framework establishing the conditions of truth. This, he held, came into existence around 1800 through practices creating reflection on human beings as subjects, especially in educational and reformatory institutions. He described reflexivity as a special condition of the modern age, and he contrasted this with the manner in which knowledge became possible through language, classification, and political economy in the earlier 'Classical' age. While many scholars controvert the specifics of Foucault's disparate statements, his general exposure of the conditions of knowledge remains significant. His search for the sources of the human sciences in practical arrangements for discipline and government, his exploration of knowledge as power, and his use of discourse bridging the social sciences and humanities, are also conspicuous features of modern research.

Thus, the human sciences exist as an unstable and variable cluster of social science and humanities disciplines and topics, sometimes grouped merely for convenience, but also grouped because shared questions about knowledge and values are at stake. How usage of the term will develop is unclear. But the future will take account of the way the 'historical turn' has expanded and rethought historical knowledge about the sources and content of representations of human activity. It will continue to work through the implications of 'the linguistic turn' and of representation itself as human activity.

The human sciences add up to a critical reflection on the constitution of disciplines, and indeed of science, in the century and a half since the modern system of legitimating knowledge became established in universities. At the beginning of the twenty-first century, it would be surprising if past prescriptions remained unquestioned. New social forces, such as appear in feminist and postcolonial perspectives, have considerable impact. Different views of how the field is constituted obtain—from an extreme opinion that the human sciences are a branch of biology to the opposite extreme that ignores biology in toto. Within the social sciences, a conception of theory as the unifying and causally explanatory framework for knowledge is diminishing in appeal. Rather, scholars working under the rubric of the human sciences offer empirically researched accounts of matters that, through much of twentieth-century psychology and social science, were taken for granted. If earlier social scientists thought it relatively unproblematic to develop knowledge on the basis of existing categories and concepts, the human sciences open new perspectives by making that unproblematic quality itself the subject of research.

See also: Anthropology, History of; Demography, History of; Developmental Sciences, History of; Economics, History of; History and the Social Sciences; History of Science; History of Science: Constructivist Perspectives; History: Overview; Psychiatry, History

of; Psychology: Historical and Cultural Perspectives; Scientific Disciplines, History of; Sociology, History of

Bibliography

Blanckaert C, Blondiaux L, Loty L, Renneville M, Richard N (eds.) 1999 *L'histoire des sciences de l'homme: Trajectoire, enjeux et questions vives*. L'Harmattan, Paris

Danziger K 1990 *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge University Press, Cambridge, UK

Felt U, Nowotny H, Taschwer K 1995 *Wissenschaftsforschung: eine Einführung*. Campus, Frankfurt, Germany

Foucault, M 1970 *The Order of Things: An Archeology of the Human Sciences*. Random House, New York (French edition 1966)

Fox C, Porter R, Wokler R (eds.) 1995 *Inventing Human Science: Eighteenth-century Domains*. University of California Press, Berkeley, CA

Hacking I 1995 *Rewriting the Soul: Multiple Personality and the Science of Memory*. Princeton University Press, Princeton, NJ

Manicas P T 1987 *A History and Philosophy of the Social Sciences*. Basil Blackwell, Oxford and New York

Nelson J S, Megill A, McCloskey D N (eds.) 1987 *The Rhetoric of the Human Sciences: Language and Argument in Scholarship in Public Affairs*. University of Wisconsin Press, Madison, WI

Porter T M, Ross D (eds.) 2002 *The Cambridge History of Sciences, Volume 7: Modern Social Science*. Cambridge University Press, Cambridge, UK

Ricoeur P 1981 *Hermeneutics and the Human Sciences: Essays on Language, Action and Interpretation*. Cambridge University Press, Cambridge, UK

Smith, R 1997 *The Fontana History of the Human Sciences*. Fontana, London; same as *The Norton History of the Human Sciences*. Norton, New York

Taylor C 1985 Interpretation and the sciences of man. In: *Philosophy and the Human Sciences: Philosophical Papers*. Cambridge University Press, Cambridge, Vol. 2, pp. 15–57 (first published 1971)

R. Smith

Human–Computer Interaction

The increasing contribution of other disciplines is making human–computer interaction focus less on the individual, with the consequence that a great deal can be gained from other fields within the social sciences, and perhaps, provide more insight into human behavior in return.

Human–computer interaction, more widely known by its acronym HCI, is primarily the study of how people use the complex technological artifacts that now pervade their lives. A secondary, but arguably more important focus, is on how these artifacts can be designed to facilitate this use. As an applied discipline, HCI encompasses theoretical and academic researchers from psychological and computer science backgrounds, who may be interested in the area as a domain within which to apply and evaluate their theories of human or system behavior, as well as designers and engineers from industrial and commercial backgrounds, for whom the pay-off is a saleable product, built faster and more economically than their competitors'. In recent years, with the rapid technological development of computing devices and their increasing ubiquity within telecommunications, the field has expanded to include social and organizational researchers, who examine the impact that the technologies have upon human–human interaction. This itself feeds back to the identification of novel applications and the development of further devices. HCI is now an active, multifaceted, and methodologically diverse component of the domain that it studies, both following and directing technical innovation.

1. Origins in Human Factors and Cognitive Psychology

The roots of HCI predate the invention of computers, and the term itself has only been used since the 1970s. At their simplest, computers are just machines, built to help people carry out tasks, and the dominant paradigm for studying HCI has been that of ergonomics or human factors, with study of the *interaction* between the human and computer subjugated to that of the human's *use* of the computer. An earlier name for the field was man–machine interaction, and within the field, the words person and human are still generally replaced by the term 'user.' In the same way that early time and motion researchers would analyze the tasks that a factory or office worker performed, breaking their job down into its constituent elements, measuring the time each part took and assessing their relationships to each other to identify any 'wasted,' redundant, or unnecessary actions, so too have HCI researchers attempted to decompose the user's task, and to identify task steps that are difficult to perform, hard to learn, or confusing.

This emphasis on task analytic decomposition reflects the early predominance within HCI of cognitive psychologists, who brought with them information processing models of human thought. These models were themselves based on a metaphor of the mind as a computer: one that received sensory input, processed it with access to memory and inferential resources, and produced motor output to act upon the world. By far the most influential of these within HCI has been the GOMS model (Card et al. 1983), which stands for Goals, Operators, Methods, and Selection. Based upon an intentionally simplistic conception of a model human processor (MHP), a GOMS analysis would identify the goals that the user has to reach, and the operators or mental actions available to the user to

carry out the task of reaching the goal. Frequently used series of operators become chunked as easily accessible 'methods' that are linked to a subgoal or specific part of the task, and selection routines enable the user to 'decide' which alternative methods or operators to apply if there is more than one possible route to the goal. The MHP provides estimates of the length of time that executing each basic sensory, cognitive, and motor operator takes (these values being empirically derived from cognitive psychology experiments), and so once an accurate GOMS analysis of the task has been produced, it is possible to sum the durations of each component to produce a prediction of the time taken for the whole task to be performed. Subsequent elaboration of the approach to allow the perceptual, cognitive, and motor parts of the architecture to operate in parallel, and to model the processing of concurrent activities using critical path analysis, has had notable success in modeling the efficiency of telephone toll and assistance workstations (Gray et al. 1993).

Despite successes such as this, a limitation on the application of GOMS in practice is that it is intended to model only the optimal performance of a skilled user in ideal circumstances, and so it is best suited to industrial-style settings where a particular task may be executed hundreds or thousands of times a day. An interface design that is a second or two quicker to use in these situations will thus result in a large saving in time, and presumably money, for the organization that is paying for it. GOMS is unable to say anything about another crucial aspect of interface use, that is, how long it might take someone to become an expert, how many errors a nonexpert might make, or how pleasant or unpleasant the experience of executing the operations might be. Indeed, the critical assumption within GOMS that operators can be chunked as methods is not actually addressed by the model itself. It remains the job of the analyst to recognize which sequences of operators can be represented as methods, and how finely these methods need to be related hierarchically. Recognizing that this development of knowledge about the task is the key to learning about an interface, Newell (1990) proposed what he termed a unified cognitive architecture specifically to allow problem solving and learning to be modeled as a process of goal decomposition, such that goals are iteratively broken down into subgoals. By specifying the operation of this architecture within a 'programming language' called Soar, Newell provided a technique for predicting the way that users would integrate their postulated initial knowledge with information derived from the computer interface through their interaction with it. In effect, a Soar analyst writes a program in the form of 'If, … Then …' rules, and this is then 'run' to simulate a user's learning behavior.

The GOMS and Soar research communities have had a great deal of influence within the cognitive psychology side of HCI, but as simulation models, their attention to detailed aspects of the interface is both their strength and weakness. While it allows them to identify small but critical aspects of a design as crucial problems for a user, they both require complete specifications of an interface, and for Soar, the task context and information available to a user. They therefore come into their own for the contrastive evaluation of extant alternatives, but are hard to apply systematically at earlier stages of design, when the question is 'how?' rather than 'which?' The same is arguably true of newer, more detailed techniques such as LICAI, which is based on Kintch's construction–integration model of goal formation in text comprehension (Kitajima and Polson 1997), and EPIC (Kieras and Meyer 1997). EPIC builds on the model human processor by incorporating many theoretical concepts such as the distinction between declarative and procedural knowledge drawn from Anderson's ACT class of models, and limited capacity visual and auditory working memories coordinated by executive processes. However, as Kieras and Meyer themselves note, 'EPIC was not developed primarily for addressing HCI problems but is a larger scientific endeavor to represent important theoretical concepts of human intelligence or abilities' (Kieras and Meyer 1997, p. 394).

2. *Development of Usability Evaluation Methods*

The apparent concentration of cognitive psychologists on the development of cognitive architectures that would ideally explain all of human behavior, as well as the human use of computers, can justifiably be seen as a problem from the point of view of their colleagues who are involved in the day-to-day design of computer interfaces. It was accepted fairly quickly that full GOMS models were too laborious to be practical, and a simpler (but inevitably less reliable) 'keystroke level model' was developed that avoided the need for a full task analysis. Soar models also require a programming and analysis commitment that can seem disproportionate in the context of design, especially when an apparently small modification in the design requires the whole model to be reprogrammed and rerun. It can also be difficult to identify reasons for the simulated behavior that results from an analysis. These cognitive architectures, then, may prove to be more useful in informing their core disciplines than in practical application.

In reaction to these perceptions of too much cost for too little benefit, a number of 'cut down' usability evaluation methods have been developed. All are to some extent derived from well-founded and principled psychological theories, but by encapsulating key components of the modeling techniques within a process of some form, seek to minimize the overhead costs of learning the theory itself, and to provide a positive cost–benefit trade-off by reducing the need for user

trials to be conducted with early prototypes or mock-ups of interfaces (Nielsen and Mack 1994). Cognitive walkthroughs, for example (Polson et al. 1992), provide a checklist by which aspects of the interface can be compared with the user's notional task, step by step, and screen by screen. The steps of the checklist are in fact derived from cognitive complexity theory, and the method focuses on the semantic appropriateness of the verbal labels or commands used throughout the interface, although the practitioners applying it do not need to know this.

Other approaches seek to avoid any role for theory at all, and base themselves in the codification and reuse of practical experience. From the earliest days of HCI, sets of guidelines have been compiled, both within design teams to provide a 'house style' and in a wider sense as a shared set of 'best practices,' so that designers encountering a problem for the first time have access to the solutions found or attempted by others. Some, such as those presented by Gardiner and Christie (1987), have an empirical basis, and specify specific design features such as acceptable screen color combinations and choices for typefaces. Others are less formal, and direct the designer towards desirable attributes or practices such as aiming for consistency, compatibility and coherence. The rapidly evolving nature of the domain conspires against such stored knowledge, however, for anything designed before the mid-1990s is almost certain to be based on expired technology, whose constraints no longer apply. Concrete prescriptions about design solutions become outdated too quickly for their promulgation, while the more abstract (such as 'be consistent') have proven too abstract to be of much real use.

3. *HCI and the Design Process*

This focus upon evaluation reveals a fundamental problem in early notions of the relationship between design and modeling. An implicit assumption of many HCI approaches was that design could be thought of as a sequential or 'waterfall' process, with discrete stages progressing from a requirements analysis, through specification layers, to implementation, with a final evaluation of the finished product. By locating evaluation at this final stage, HCI staff found themselves limited to what became known as a 'fire-fighting' role, being asked to repair problematic designs at a point when any redesign would result in a product falling behind schedule, further worsening the cost implications of their involvement. Organizationally, it is very difficult to convince people that the design they have been committed to for months has some basic usability flaws.

As Landauer (1995) has pointed out in his econometric analysis of the computer industry, design flaws have to be identified much earlier in design if their identification is to be helpful and cost effective to repair. Fortunately, few designs actually proceed in such a strictly formal manner, with a more serendipitous, parallel, and iterative process being more typical. Such cyclical processes allow HCI contributions to be made at more appropriate junctures, and have lead to an emphasis upon user-centered design (e.g., Norman and Draper 1986), in which user testing becomes a central aspect of design throughout the process, design itself is directed towards prototyping, and designers are encouraged to 'plan to throw one away,' following Brooks' seminal advice (Brooks 1975).

A very different approach to providing support to designers in these less structured situations comes in the form of a notational decision support tool called QOC, or questions, options, and criteria (MacLean et al. 1991). Aimed at recording the 'design rationale,' QOC represents the design space through which a product has traveled, listing the questions that the design team were trying to answer at various points, the options that were considered, and the criteria that they used to choose one option over another. By making decisions and trade-offs explicit, and recording the reasons that were cited for making them, QOC provides a mechanism both for managing the decision process and for recording it for later review. It has been found to be particularly helpful in providing background information for people who were not part of the decision at the time, either because they only become involved at a later stage or because they were working on some other part of a large project. QOC is important in that it represents a wider concern with design as a process, rather than simply with evaluating its final products.

Involvement in design at a point before the final product has been created means, inevitably, that HCI methods need to provide advice to designers rather than evaluating finished designs. While 'cut-down' usability inspection methods and heuristic evaluations can be conducted on partial designs and mock-ups, it has been argued that their pragmatic separation from theory weakens their value. While they may well identify problems in a design, the fact that they are not associated with an explanatory theoretical underpinning means that the causes of the problems can remain unexplained. The benefit to the designer of a rapid identification of problems is accompanied by the cost of working out why the problem has occurred and what is to be done about it. May and Barnard (1995) recommended that HCI researchers should work towards providing 'supportive evaluation methods' that assist the designers by accounting for usability problems in terms of some theoretical concepts that are helpful in suggesting redesigns.

They referred to a distinction made by Long (1987) between the 'discovery representation' of a scientific theory and its 'application representations.' The discovery representation allows academics and researchers to explore phenomena, such as typical error patterns in HCI, or the way that users learn about interfaces

and solve problems. The application representations, on the other hand, are encapsulations of the theory specifically targeted towards practical application. These may take several different forms, depending upon the intended practitioners and the expected purpose. Unlike cut-down methods, the application representations retain a direct link to the parent theory and, in order to use them, information has to be expressed in theoretically important ways. The advantage for practitioners is that the results of their use is also couched in theoretically relevant forms, and so is linked to an explanatory body of ideas that can be recruited to assist in redesign.

May and Barnard (1995) present the example of a cognitive architecture called interacting cognitive subsystems, or ICS. As a discovery representation, it is unlike GOMS, Soar or EPIC in that it does not seek to simulate cognitive processing, but provides a framework for identifying bottlenecks in the flow of information through a series of cognitive subsystems, each of which is specialized for processing a different level of mental representation (see *Cognitive Science: Overview*). Like the simulation models, these levels include sensory, cognitive, and motor representations. The modeling process involves describing the likely pattern of exchanges between the various levels, the degree of skill-based behavior of relevant aspects of a task, the involvement of the user's recent and longer-term memories, and, crucially, the structure of the user's conceptual task or of perceptual representations of the interface. In this discovery form, the modeling is too complex to be carried out by anyone other than an expert in the theory itself, and the results are couched in terms of the processes and their configuration, rather than specific design consequences.

To support the use of ICS within design, two alternative application representations have been developed: one is a production rule-based 'expert system,' which partly replicates the role of the ICS theorist in constructing a principled model of an interaction; the other is a pencil-and-paper technique for reasoning about the changes over time in the structure of the mental representations encoded at the various ICS subsystems (Barnard and May 1999). Both of these application representations require some familiarity with the core concepts of ICS, but neither requires practitioners to construct a full ICS model themselves. Their value lies in providing problem-solving tools for designers to apply to partially specified or speculative ideas about interfaces, before any commitment has been made.

4. *Influx of Novel Approaches*

In moving away from a simulation of a user interacting with a computer, ICS is part of a trend away from the concentration upon a dyadic, turn-taking relationship between a single user and a single computer. Classical HCI models can be caricatured as having a human issuing commands to a computer, waiting for it to carry them out, and then inspecting the screen while the computer patiently waits for them to form some new goals. While this might have been an appropriate model for HCI in the early 1980s, the advent of distributed computing architectures and the massive interconnection of computing systems over the Internet means that it is inadequate. Even modern personal computers will be carrying out multiple tasks, doing background processing almost in-between each keystroke. The humans are not always isolated users pursuing solitary goals, either, but are using computers as part of their everyday activities, often in social or work settings that involve interacting with other humans with the support of their computers. An analysis of interaction that does not allow for such factors has inevitable flaws.

These developments have led to a new involvement within HCI of social and organizational psychologists, as well as researchers from other social sciences such as ethnomethodology, anthropology, and sociology (Dourish and Button 1998). This has led to an influx of novel theoretical ideas, in particular Vygotsky's activity theory (e.g., Nardi 1995) and Suchman's situated action (Suchman 1987). These ideas re-emphasize the relevance of the context within which human actions are formed, and as Carroll (1997) noted, have brought a continental European perspective into what had previously been a largely North American and British endeavor. Instead of the user and their system, activity theory deals with an 'activity system' composed of social aspects such as attitudes, beliefs, and experiences as well as technological factors.

Another development which moves away from the traditional HCI dyad has come from a computer science perspective, where the idea of a system as a single entity has long been outdated. With the introduction of 'object oriented' ideas, computer scientists have moved towards seeing complex systems as built up from a large number of quasi-independent 'interactors' that exchange information and tasks, with no constraint for these interactors to be part of a single machine. A large body of theory has evolved to model such distributed and concurrent activity, and a logical step has been to see a user's cognitive processes as just another set of interactors. The unit of analysis for such 'syndetic models' (Duke et al. 1999) is the overall system composed of human and computer interactors, and their conjoint activity can be modeled using mathematical formalisms to prove literally that certain interactions can or cannot occur, or that certain assumptions must be met for them to be possible.

5. *Prospects for HCI as a Discipline*

Both the social and the computational contributions to HCI currently fall into the discovery mode of analysis, in Dowell and Long's terms, rather than the

application mode. It is not yet clear whether these novel directions will crystallize into methods or representations that are directly applicable by design practitioners, or whether they will serve more to direct the commercial development of products away from the classical PC in an office towards other niches. They may also represent the evolution of HCI as a discipline in its own right, with its own methods and theories, rather than as an interdisciplinary area colonized by theories developed in cognitive psychology and artificial intelligence. The tension between academic theorists and practical developers will undoubtedly continue to be a major dynamic upon the field, with its predominance of conference papers enabling non-empirical presentations of novel technologies or design tools to share the limelight with hard-nosed experimental evaluations. In recent years, the increasing theoretical quality of the few specialized HCI journals may provide justification for those who have argued for universities to establish HCI departments in their own right, with their own undergraduate courses. On the other hand, the progressive involvement of computing artifacts within our everyday lives is making their use an ordinary part of human behavior. The increasing contribution of other disciplines is making HCI focus less on the individual, with the consequence that a great deal can be gained from other fields within the social sciences and, perhaps, provide more insight into human behavior in return.

See also: Cognitive Theory: SOAR; Computers and Society; Human–Computer Interface; Science and Technology, Social Study of: Computers and Information Technology; User Interface Design

Bibliography

Barnard P J, May J 1999 Representing cognitive activity in complex tasks. *HCI* **14**: 93–158

Brooks F P 1975 *The Mythical Man Month; Essays on Software Engineering*. Addison-Wesley, Reading, MA

Card S K, Moran T P, Newell A 1983 *The Psychology of Human–Computer Interaction*. Erlbaum, Hillsdale, NJ

Carroll J M 1997 Human–computer interaction: psychology as a science of design. *International Journal of Human Computer Studies* **46**: 501–22

Dourish P, Button G 1998 On 'technomethodology': foundational relationships between ethnomethodology and system design. *HCI* **13**: 395–432

Long J B 1987 Cognitive engineering and HCI. In: Warr P (ed.) *Psychology at Work*. Pelican, Harmondsworth, UK, pp. 73–93

Duke D J, Barnard P J, Duce D A, May J 1999 Syndetic modeling. *HCI* **13**: 337–93

Gardiner M M, Christie B (eds.) 1987 *Applying Cognitive Psychology to User–Interface Designs*. Wiley, Chichester

Gray W D, John B E, Atwood M E 1993 Project Ernestine: validating a GOMS analysis for predicting and explaining real-world task performance. *HCI* **8**: 237–309

Kieras D E, Meyer D E 1997 An overview of the EPIC architecture for cognition and performance with application to human–computer interaction. *HCI* **12**: 391–438

Kitajima M, Polson P G 1997 A comprehension-based model of exploration. *HCI* **12**: 345–89

Landauer T 1995 *The Trouble with Computers*. MIT Press, Cambridge, MA

MacLean A, Young R, Bellotti V, Moran T 1991 Questions, options and criteria: elements of design space analysis. *HCI* **6**: 201–50

May J, Barnard P J 1995 Towards supportive evaluation during design. *Interacting with Computers* **7**: 115–43

Nardi B A (ed.) 1995 *Context and Consciousness: Activity Theory and Human–Computer Interaction*. MIT Press, Cambridge, MA

Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

Nielsen J, Mack R L (eds.) 1994 *Usability Inspection Methods*. Wiley, New York

Norman D A, Draper S W 1986 *User Centered System Design*. Erlbaum, Hillsdale, NJ

Polson P G, Lewis C H, Rieman J, Wharton C 1992 Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man Machine Studies* **36**: 741–73

Suchman L A 1987 *Plans and Situated Actions: The Problem of Human–Machine Communication*. Cambridge University Press, New York

J. May

Human–Computer Interface

1. Definitions: The Interface as a Carrier of Epistemologies

The computer/human interface refers to the modalities through which people interact with computational technologies. These include different data input devices: you can give the computer information through punched cards or through typing numbers on to a screen. And these include different ways of manipulating software: one can type symbols at a 'command line' or can 'navigate' a data field that appears as a three dimensional space. Specialized applications—software designed for the physically challenged, software designed to simulate airplane flight—have increased the number of available interface technologies. Computer users may employ joysticks, headsticks, trackballs, voice recognition, and touch sensitive screens to communicate with the software that governs the machine's behavior.

The breadth of the topic of computer/human interfaces and how to improve them for usability and ease is too vast to review comprehensively here (Norman 1998, Raskin 2000). This essay examines the most common shifts in interface use over the last half century up to 2000, and centers on the social and psychological implications of interface experience rather than its technical details. Interface design has not only changed the instrumental computer: what the

computer can do for us; it has changed the subjective computer: what the computer does to us, to our ways of relating to others, thinking about ourselves, seeing the world. As the dominant knowledge technology of our era, computational machines offer us images of what it means to know (Penrose 1989, Dennett 1991, Turkle 1995).

Over the last half century to 2000 the trend in interfaces has been from favoring a style of interaction in which the computer is approached as a mechanism towards interactions in which the computer is approached as a behaving and aware organism. Otherwise put, the trend has been from people acting directly on the machine to interacting indirectly with its increasingly elaborate 'self presentation' in which many layers of programming stand between the underlying mechanism and what is presented to the user. The movement has been from interface as an environment for commands (Weinberg 1971) to interface as environment for conversation or theater (Laurel 1990).Whether the computational object presents an interface that demands small or large bodily gestures or particular words or sounds, this direction (increased layering; increased opacity) has had significant psychological and epistemological implications. The computer has moved from being the primary technological model of a culture of calculation towards modeling a way of knowing in a culture of simulation. In a culture of simulation, users are encouraged to take computers 'at interface value.'

1.1 The Recent History of Personal Computing Interfaces

The first personal computers of the 1970s and early 1980s, like the mainframes and minicomputers, required users to know how to issue exact instructions. Being comfortable with their command-line required users to put themselves into the frame of mind of those who had programmed the computer, who had given it its fundamental rules. Those who were most comfortable with this kind of information interface were those who knew something about programming and its intellectual aesthetic (Schneiderman 1987).

Through the late 1970s, the computer interface presented the machine as a giant calculator. The interface carried the message that no matter how complicated a computer might seem, what happened inside it could be mechanically unpacked. The computer interface was a set of command lines that would get you close to those rules that lay beneath the surface of the machine.

The presentation of the machine through its interface captured the essence of the computer in a culture of calculation. This essence corresponded to a modernist or enlightenment model of what it means to understand: it means to have knowledge of what makes the 'clockwork' tick, and the programmed computer was conceived of very much as mechanical clockwork. When computer users talked about 'transparency' what they meant is that the interface was a window for looking beyond the 'magic' of the machine's extraordinary powers to the mechanism that made them possible.

What was true of the language that surrounded the large mainframes and mini-computers was even more true of the language that surrounded the first personal computers. These machines, first appearing in the late 1970s and early 1980s, required users to know how to issue exact instructions and they were encouraged to optimize their computers' performance by learning assembly language, by going deeper and deeper into the lower levels of a machine's structural organization. By the mid-1980s, increased processing power made it possible to build graphical user interfaces, commonly known by the acronym GUI, which hid the 'bare machine' from its user. The new opaque interfaces—the first popular one on the mass market was the 1984 Macintosh—represented more than a technical change. The Macintosh 'desktop' introduced a way of thinking about the computer that put a premium on the manipulation of a surface simulation. The desktop's interactive objects, its dialogue boxes in which the computer 'spoke' to its user, pointed toward new kinds of experience in which people did not so much command machines as enter into conversations with them (Apple 1985). In personal relationships, we often interact without understanding what is going on within the other person; similarly, when we take things at (inter)face value in the culture of simulation, if a system performs for us, it has all the reality it needs.

In 1980, most computer users who spoke of transparency were referring to a transparency analogous to that of traditional machines, an ability to 'open the hood' and poke around. But when users of the Macintosh talked about its transparency, they were talking about seeing their documents and programs represented by attractive and easy-to-interpret icons. They were referring to an ability to make things work without needing to go below the screen surface. Today, the word 'transparency' has taken on its Macintosh meaning in both computer talk and colloquial language. In a culture of simulation, resonant with a post-enlightenment or postmodern aesthetic, when people say that something is transparent, they mean that they can see how to make it work, not that they know how it works.

2. Simulation and Its Discontents

Suppose you receive this message on a computer screen when playing a simulation computer game: 'Your orgot is being eaten up.' (Turkle 1995). You do not need to know what an orgot is to play the game

although you can infer from the context of the game that it is a kind of organism. Players interact with the interface of the computer game in the same way they interact with the interface of a video game. In computer games, as in video games, one learns from the process of play. You do not first read a rulebook or get your terms straight. The game allows you to manipulate its variables directly; it enables you to avoid commands, but experiment with trial balloons. You are able to act on an intuitive sense of what will work without understanding the rules that underlie the game's behavior. Your game will teach, for example, that you that you must give your 'trilobytes' algae or they will become extinct. Thus, you can keep playing even without any clear idea of what is driving events. Clarity and a grasp of the rules are not essential to the process.

People use their contact with objects to keep in touch with their times, to discover 'what's important.' They use objects to work through powerful cultural images, to help arrange these images into new and clearer patterns. From this point of view, the opaque interfaces of simulation games and simulation tools help to help us think through current challenges to traditional epistemologies where the manifest refers back to the latent, the signifier to the signified. The personal computer culture began with small machines that captured a utopian modernist vision of transparent understanding, but the scintillating, pulsating surfaces of simulation computer games are dynamic, seductive, and elusive, a world without depth, a world of surface. *Players can stay on the surface, taking things at (inter)face value*. The computer interface serves as a carrier object which, in a significant sense, brings postmodernism down to earth.

The culture of simulation presented by the computational interface is our everyday experience of the postmodernist aesthetic. And since it is common to experience the computer as a 'mind machine' this aesthetic influences computational models of mind. For example, 15 years ago the computational models of mind that dominated academic psychology were modernist in spirit: nearly all tried to describe the mind in terms of centralized structures and programmed rules. In contrast, today's models often embrace a postmodern aesthetic of complexity and decentering. Mainstream computer researchers no longer aspire to program intelligence into computers but expect intelligence to emerge from the interactions of small sub-programs. If these emergent simulations are 'opaque,' that is, too complex to be completely analyzed, this is not necessarily a problem. After all, these theorists say, our brains are opaque to us, but this has never prevented them from functioning perfectly well as minds.

Beyond this, the development of the windows metaphor for computer interfaces was a technical innovation motivated by the desire to get people working more efficiently by 'cycling through' different applications much as time-sharing computers cycle through the computing needs of different people (Microsoft 1995). In practice, however, windows have also become a potent metaphor for thinking about the self as a multiple, distributed, 'time-sharing' system. The self is no longer simply playing different roles in different settings, something that people experience when, for example, one wakes up as a lover, makes breakfast as a mother, and drives to work as a lawyer. The life practice of windows is of a distributed self that exists in many worlds and plays many roles at the same time.

2.1 Conversation, Cuddles, and Creatures

Computers have been thought of as 'artificial brains' for generations. What is new in the current analogies of computers to the brains of behaving systems is the development of interfaces through which we can interact with computers as though they were creatures, either human or animal. An example is the conversational interface used to communicate with software programs known as 'agents' or 'bots.' The computer users types out their side of a conversation in colloquial language; the software agent responds more or less in kind. Of course, such conversations work best if the domain of discussion is highly proscribed, but the experience of such interfaces is quite dramatically that of interacting with an 'other.' Computer agents that literally present 'faces' as their interfaces dramatically increase the degree of anthropomorphism of the computational object. Voice recognition software enhances the conversational effect and the experience of the machine as a being.

Computational creatures are being designed as toys for children. Today, children are asked to interact with these creatures through gesture and touch; in the near future, the interactions will be through language and tone. In the case of the Furby, a furry, cuddly, robotic toy that seems to learn and love, the computational interface presents a creature successfully 'learning' language and being nourished by food; the child is able to respond with gestures expressing affection. In the case of another computational creature, Sony's pet dog robot, AIBO, the computational interface presents a 'trainable' dog whose personality can be shaped by the quality of the user's ministrations, communicated by speech and sound. In the case of My Real Baby, a robotic doll, the interface enables the doll to speak increasingly complex language and show emotional expression through facial musculature while affording the user the possibility of showing nurturance through feeding, caressing, diaper changing, burping, and similarly anthropomorphic 'creature comforts.' Here, perhaps most dramatically, computational design has created an interface that through its emulation of the human explicitly diverts attention from the underlying mechanism and focuses it ex-

clusively on the meaning of the 'baby's' expressions, sounds, and words. Almost no children express curiosity about how the object 'works' in the traditional sense of understanding anything about its 'innards.' They accept its behavior as that of a creature, a creature that while not human, is frequently categorized as 'sort of alive.'

3. Future Trends

Traditional approaches to studying user interfaces have focused on usability, human factors (including level of stress), and psychological studies that have examined sensorimotor adaptation, problems of information overload, ergonomics, and the resonance of interfaces with varying intellectual styles. The projected future of computation and its interfaces suggest that philosophical and even theological concerns might well be added to this mix since the interface is promised to become not only more transparent but more intimately engaged and ultimately fused with its users.

With the computer/human interfaces that researchers project in the near-future the user will experience the interface as a creature (as in increasingly advanced robots) as prosthetics (a pair of glasses that also serve as a computer screen for mobile net browsing), a piece of clothing (with embedded processors that for example use the temperature of the human body as the input to warm or cool the skin), or as a virtual world that people can fully immerse in and navigate by moving their bodies as one would move in the real physical world. Finally, there are proposals for the interface to become part of the self. In this scenario the interface becomes a digital sensor embedded in human flesh. Creatures, prosthetics, immersive environments, and in the end, the interface disappears and people reappear as cyborg.

See also: Cognitive Modeling: Research Logic in Cognitive Science; Cognitive Science: Philosophical Aspects; Computational Imagery; Computers and Society; Discrete State Models of Information Processing; International Migration by Ethnic South and Southeast Asians; Science and Technology, Social Study of: Computers and Information Technology; User Interface Design; Virtual Reality and Spatial Cognition

Bibliography

Apple Computer 1985 *Inside Macintosh*. Apple Computer, Cupertino, CA, Vol. 1

Dennett D 1991 *Consciousness Explained*. Little, Brown, Boston, MA

Laurel B 1990 *The Art of Human-Computer Interface Design*. Addison-Wesley, Boston, MA

Microsoft 1995 *The Windows Interface Guidelines for Software Design*. Microsoft Press, Redmond

Norman D 1998 *The Invisible Computer*. MIT Press, Cambridge, MA

Penrose R 1989 *The Emperor's New Mind*. Oxford University Press, London

Raskin J 2000 *The Humane Interface: New Directions for Designing Interactive Systems*. Addison-Wesley, Boston

Schneiderman B 1987 *Designing the User Interface*. Addison-Wesley, Boston

Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Basic Books, New York

Weinnberg, G 1971 *The Psychology of Computer Programming*. Van Nostrand Rienhold, New York

S. Turkle

Human–Environment Relationship: Carrying Capacity

Carrying capacity is the maximum population that a specified environment can support indefinitely without degradation of that environment (Ellen 1982, p. 41). Ecological carrying capacity links population of biota to a given habitat in which biophysical elements of the habitat determine the maximum population. In human–environment relationships, however, carrying capacity involves the added dimension of technology and management strategies; as these human-based elements change, so does the carrying capacity (Brush 1975 and Street 1969). Thus, humankind maintains the ability to expand and reduce carrying capacity, and vastly different assessments of the actual trends underway follow from the spatial scale of analysis (i.e., landscape, region or biosphere) and perspective (i.e., environment as resource commodities or as environmental services) taken (cf. Ehrlich et al. 1977 and Simon and Kahn 1984).

Contemporary uses of carrying capacity suggest that the concept was borrowed from the ecological sciences by the human sciences. In reality, however, the ecological sciences took a cue from the ideas of Robert Malthus and subsequent developments of Malthusian principles about human population–resource relationships. The carrying capacity idea can be found in various kinds of works, even though that label was or is not applied. For example, the geographer–geomorphologist John Wesley Powell used the concept in his assessment of the Homestead Act of 1862 (USA), which allowed people to claim 160 acres of public land for farming. He warned that 160 acres were insufficient in most cases because the land in question, that west of the 100th meridian, was arid and much of it lacked access to water. He maintained that these lands, absent the potential for irrigation, had value primarily for livestock and thus 'a farm unit should not be less than 2560 acres' in the arid West of the USA (Darrah 1951, p. 233).

Such examples notwithstanding, the nineteenth century development of the carrying capacity concept

in the human science is often credited to geographical and anthropological research among indigenous peoples in the less developed world, who typically maintained a fairly direct relationship between subsistence and the immediate biophysical environment (Brush 1975 and Street 1969). The British geographer William Allan (1965), for example, used the concept of carrying capacity to evaluate the Native Reserve resettlement process in southern Africa. The essential data required for the calculation were: '(a) a soil or land map based on a suitable form of land classification; (b) an estimate, for each of the land classes employed, of the proportion of land which is practicable for cultivation under the prevailing system of agriculture; (c) a survey of the customary land-usage of the people concerned, with particular reference to the duration of cultivation and subsequent rest on the soils of each land class; and (d) an average for the acreage under cultivation per head of population at any one time' (Allan 1965, p. 8). He found that the carrying capacity of the Native Reserves, registered as persons per unit area, was exceeded at their founding, based on the customary land uses of the inhabitants. Other versions of carrying capacity assessments have been called for and employed for planned occupation throughout the tropical world, including the colonization of the Amazon basin (Fearnside 1986).

As noted, technology and land resource–use strategies play an essential role in the use of carrying capacity. As these change, so does the capacity assessment. The 'Green Revolution' is a case in point. This massive change and transfer of technology and land-management strategies greatly expanded the production and hence 'carrying capacity' of many areas throughout the world by making possible the intensification of agriculture through hybrid seeds, fertilizers, pesticides, herbicides and irrigation schemes. Yet, this revolution and its land-use impacts also expanded the reach of environmental drawdown, from increased downstream flooding and sedimentation to large-scale releases of trace gases to the atmosphere, raising serious questions about the spatial and hierarchical scales at which carrying capacity should be addressed. While earlier studies could concentrate on a village or small area as a well-defined population–resource entity, it is now recognized that, particularly in industrialized countries, such entities no longer conform to such well-defined boundaries. Technology permits a place-based population to substitute for its environment by importing resources and exporting waste. Thus, carrying capacity now involves an assessment of the worldwide 'ecological footprint' of a place-based population (Rees 1992), or the full extent of the use and environmental impacts of that population. Through this lens, what may appear to be a sustainable carrying capacity in an area, registered by arrested or slowed environmental degradation, or even its reversal, may be obtained by transferring elsewhere the environmental consequences of the area's occupation and the population's consumption. This situation necessitates an examination of carrying capacity on a global as well as a local basis, as the recognition of the 'silent' services provided by environment, such as pollinators for crops, necessitates an examination that includes more than immediate population–resource relationships (Vitousek et al. 1997).

See also: Agricultural Change Theory; Ecological Economics; Ecology, Cultural; Environment and Development; Environmental and Resource Management; Environmental Economics; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; Feminist Political Ecology; Green Revolution; Greening of Technology and Ecotechnology; International and Transboundary Accords, Environmental; IPAT (Impact, Population, Affluence, and Technology); Planning Issues and Sustainable Development; Precautionary Principle; Sustainable Development

Bibliography

Allan W 1965 *The African Husbandman*. Barnes and Noble, New York

Brush S B 1975 The concept of carrying capacity for systems of shifting cultivation. *American Anthropologist* **77**: 799–811

Darrah W C 1951 *Powell of the Colorado*. Princeton University Press, Princeton, NJ

Ellen R 1982 *Environment, Subsistence and System*. Cambridge University Press, Cambridge, UK

Ehrlich P R, Ehrlich A H, Holdren J P 1977 *Ecoscience: Population, Resources, Environment*. W. H. Freeman, San Francisco, CA

Fearnside P M 1986 *Human Carrying Capacity of the Brazilian Rain Forest*. Columbia University Press, New York

Rees W E 1992 Ecological footprints and appropriated carrying capacity: what urban economics leaves out. *Environment and Urbanization* **4**: 121–30

Simon J L, Kahn H (eds.) 1984 *The Resourceful Earth: A Response to Global 2000*. Blackwell, Oxford, UK

Street J 1969 An evaluation of the concept of carrying capacity. *Professional Geographer* **21**: 104–7

Vitousek P M, Mooney H A, Lubchenco J, Melillo J M 1997 Human domination of earth's ecosystems. *Science* **277**: 494–99

M. E. Geores

Human–Environment Relationship: Comparative Case Studies

A comparative case study is a research approach to formulate or assess generalizations that extend across multiple cases. The fundamental intent of a compara-

tive case study—development of knowledge—applies equally in the natural and social sciences, with the latter incorporating qualitative analysis in a discursive mode as well as the quantitative and positivist modes of the sciences. Comparative case studies are particularly important to questions of human–environment interaction, combining case study and comparative traditions and thus extending the depth of the individual case study across several instances in time and/or space.

1. Case Study and Comparative Method

A case study is a rigorous, detailed, spatially bounded, temporally limited description and analysis of an event, process, place, or community intended by the researcher to exemplify a more general insight, principle, or theory. It 'is an exploration of a "bounded system" or case (or multiple cases) over time through detailed, in-depth data collection involving multiple sources of information rich in context' (Creswell 1997, p. 61). Case studies of particular communities or places are a common and classical mode of analysis in anthropology, human geography, and sociology.

Comparative analysis, in contrast, explores similarities revealed in different situations or cases sharing some common element(s) while differing in others. Comparative analyses are also common among social sciences, varying in the depth to which components are explicated, from statistical abstractions through vignettes to a level of detail common in individual case studies. Often, a case study has unique elements whose effects on the validity of generalizations may be ambiguous or uncertain. Thus case studies addressed through a conjunction of insights from alternative perspectives will reveal commonalties of particular interest. Comparison between cases can help to delineate such relationships and potentially to clarify ambiguity in a singular case.

1.1 Qualitative and Quantitative Aspects

Comparative analyses may proceed via quantitative analysis of numerical abstractions from cases, through detailed narrative description and qualitative analysis. There is no necessary relationship between qualitative and quantitative approaches in case studies and comparative analyses. Qualitative case studies are frequent in anthropology, some schools of sociology, and contemporary cultural geography. Case studies can also be internally as well as comparatively quantitative, particularly in human and economic geography. Whereas comparative studies are often qualitative, they may be quantitative when there are a large number of cases with a small number of variables, as in social survey research.

1.2 Explanatory Status of Comparative Case Studies

Both case studies and comparative analyses can be *inductive* or *deductive*. However, case studies and qualitative comparative analyses share a common explanatory limitation with models—they are fundamentally *heuristic* rather than *definitive* in that they cannot prove the validity of generalizations beyond providing persuasive evidence in particular contexts (Oreskes et al. 1994). Conversely, a case study illuminating and supporting a general principle does not preclude another case casting doubt. As Leach (1968, p. 344) stated about the comparative method in anthropology,

> …comparison…is not, and cannot be, a disguised form of scientific experiment leading to an explanation.' Moreover, neither misinterpretation of the case nor failure of the case to illustrate a generalization necessarily means that a hypothesis is not valid. However, a comparative case study approach strengthens the heuristic value of individual cases by bringing in-depth understanding in different circumstances, with greater confidence in the conclusions drawn from them. 'The generality of the case study…is relative to the methodological value of the test process involved in the experiment itself and "holds until proved otherwise" (Hamel 1993, p. 40).

Multiple cases strengthen the confidence of the researcher that the generality is sound.

2. Comparative Case Study Strategies

Researchers investigating human–environment issues can avail themselves of many alternative strategies for developing comparative case studies. Strategies included here are certainly not exhaustive nor mutually exclusive, but they suggest ways in which comparative case studies can be structured creatively for investigation, presentation, and analysis. Often, the power of the comparative analysis is strongly related to the selected strategy. These strategies include research on the basis of community, place, perspective, structure, system, scale, transects, gradients, time, events, analogs, or theory as organizing principles. In practicality, the major organizing principle of a comparative case study often includes others that are subsidiary, but inherent in the nature of the cases selected. For example, selection of case communities along a transect incorporates both community and place perspectives as well, even though the transect is the organizing theme. Examples of each strategy are included in Table 1.

2.1 Community

Comparative case studies in disciplines such as sociology and anthropology often focus on studying

Table 1
Comparative case studies: a literature guide

Community	Bennett J W 1969 *Northern Plainsmen: Adaptive Strategy and Agrarian Life*. Aldine, Chicago
	Mortimore M 1999 *Working the Sahel: Environment and Society in Northern Nigeria*. Routledge, London
Place	Moran E F 1993 *Through Amazonian Eyes: The Human Ecology of Amazonian Populations*. University of Iowa Press, Iowa City, IA
	Turner B L, Hyden G, Kates R W 1993 *Population Growth and Agricultural Change in Africa*. University of Florida Press, Gainesville, FL
Perspective	Knight C G 1974 *Ecology and Change: Rural Modernization in an African Community*. Academic Press, New York
	Meinig D W 1979 *The Interpretation of Ordinary Landscapes: Geographical Essays*. Oxford University Press, New York
Structure	Goldschmidt W R 1966 *Comparative Functionalism: An Essay in Anthropological Theory*. University of California Press, Berkeley, CA
	Moran E F 1995 *The Comparative Analysis of Human Societies: Toward Common Standards for Data Collection and Reporting*. L Rienner, Boulder, CO
	Turner B L, Brush S B (eds.) 1987 *Comparative Farming Systems*. Guilford, New York
System–subsystem	DeWalt B R, Pelto P J (eds.) 1985 *Micro and Macro Levels of Analysis in Anthropology: Issues in Theory and Research*. Westview Press, Boulder, CO
	Mitchell J K 1974 *Community Response to Coastal Erosion: Individual and Collective Adjustments to Hazard on the Atlantic Shore*. University of Chicago, Department of Geography, Research Series no. 156
Scale	Wilbanks T J, Kates R W 1999 Global change in local places: how scale matters. *Climatic Change* **43**: 601–28
	Easterling W E et al. 1998 Spatial scales of climate information for simulating wheat and maize productivity: the case of the US Great Plains. *Agricultural and Forest Meteorology* **90**: 51–63
Transects	Kromm D E, White S E (eds.) 1992 *Groundwater Exploitation In The High Plains*. University Press of Kansas, Lawrence, KS
	Moran E, Brondízio E 1998 Land use change after deforestation in Amazonia. In: Liverman D et al. (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC, pp. 94–120
Gradients	Allan W 1965 *The African Husbandman*. Barnes and Noble, New York
	Porter P W 1965 Environmental potentials and economic opportunities—a background for cultural adaptation. *American Anthropologist* **67**: 409–20
Time	Broek J O M 1932 *The Santa Clara Valley, California: A Study in Landscape Changes*. Oosthoek, Utrecht, The Netherlands
	Carlstein T 1982 *Time Resources, Ecology and Society*. George Allen and Unwin, London
Events	Blaikie P M 1985 *The Political Economy of Soil Erosion in Developing Countries*. Longman, London
	Burton I, Kates R W, White G F 1993 *The Environment as Hazard*. Guilford Press, New York
Analogs	Glantz M H (ed.) 1988 *Societal Responses to Regional Climatic Change: Forecasting by Analogy*. Westview Press, Boulder, CO
	Wood J S, Knight C G 1974 Cross-cultural perspectives on human adjustment to arid margins. In: Newman J L (ed.) *Drought, Famine and Population Movements in Africa*. Syracuse University Foreign and Comparative Studies/Eastern Africa Series no. 17, Syracuse NY, pp. 75–126
Theory	Boserup E 1966 *The Conditions of Agricultural Growth: The Economics of Agrarian Change Under Population Pressure*. Aldine, Chicago
	Worster D 1985 *Rivers of Empire: Water, Aridity, and the Growth of the American West*. Pantheon, New York

different social communities to ascertain, for example, generalities surrounding a particular cultural pattern (relations between matrilineal systems and matrilocal residence); social experience (immigrants in the new world city); or community structure (comparative political systems). In human–environment studies, comparison between communities is also important. Examples of such comparative case studies include investigation of traditional or vernacular architecture; studies of pastoral life-styles and range management strategies; cross-cultural studies of local environmental relations; analysis of ethnic farming populations in

the same locale; and studies of traditional navigation technologies and knowledge systems used by island inhabitants.

2.2 Place

Places which provide examples of a shared geographical setting or similar human–environmental process may be comparatively studied. Examples include studies of human–environment relations in specific kinds of environment, such as mountains, flood plains, and island or coastal areas; research on societal responses to extreme population density; and investigation of places affected by technological transformation, such as impacts of transportation change (super highways, dieselization of railroads), agricultural mechanization leading to rural depopulation, management of water resources, and transitions in household energy use in developing countries.

2.3 Perspective

An alternative to selection of different cases analyzed from a similar perspective is to bring multiple perspectives to a single case. Such an approach explicitly invokes a multioperationism desirable in all case studies—the convergence of different kinds of evidence toward a more complete and rich understanding. One common method for an approach based on perspective is the *emic* vs. *etic* familiar from anthropology—the understanding held by the individuals involved in contrast to the analytical viewpoint of the researcher. This distinction is now an important element of postmodern qualitative inquiry across the social sciences. The use of perspective in comparative study includes landscape analysis, analysis of environmental controversy, and multiperspective community or place studies.

2.4 Structure

Comparative case studies may be developed on the proposition that selected cases reflect the functioning of a common underlying structure, somewhat analogous to comparative linguistics. For cross-cultural translation of color terms, a formal Munsell color chart can provide a common structure on which to map various language terms for color. Less formally, comparative analysis may be based on typologies which provide the framework for comparison. In the human–environment realm, structural frameworks used as the basis for comparative case studies include Linnean plant classification compared to indigenous ethnobotanical plant classification; fundamental requisites of a society; comparative farming systems; and human response systems in the face of natural hazards.

2.5 System–Subsystem

A comparative case study may be organized to look at various levels of a complex system to establish relationships, mechanisms, and emergent properties that vary by the level under investigation. Examples include studies of the public role in environmental management at local, state, and national levels; studies of agricultural change viewed at the family, village, society, and regional levels; and studies of the driving forces of deforestation, from family- and farm-level decisions, through regional and national forces to the global economy.

2.6 Scale Analysis

Case studies of the occurrence of phenomena at different spatial scales may be undertaken to ascertain variance or invariance of form or process. Scale analysis has become an important element in the study of the natural environment, including hydrological, climatological, and ecological systems. In human–environment investigations, scale analysis is undertaken in studies of the causes and potential impact of climate change to suggest local and larger-scale determinants of greenhouse gas emissions; appropriate levels of generalization of impacts, such as climate change on farming; and initiatives for mitigation.

2.7 Transects

Geographical transects provide one kind of spatial sampling that researchers can use to develop information and insight that can be generalized across a society or region. A transect provides one means of minimizing sample bias in an attempt to develop generalizations and analysis of differences within a society or region that might reasonably be assumed to be homogeneous. Such studies include investigation of market towns along a transect in an agricultural region; research on the dynamics of colonist settlement along new roads in rainforest areas; or exploratory archeological investigation of settlement remains in a region. In transect studies, the area being investigated may be found not to be homogeneous, suggesting comparative study of social differences (such as ethnic enclaves in an agricultural region) or environmental

differences (such as local availability of ground water in the High Plains of the USA). One special kind of transect selected in recognition of difference rather than homogeneity is the gradient.

2.8 Gradients

Cultural ecologists in geography and anthropology often undertake comparative case study analyses using sample cases selected across environmental, spatial, or social gradients rather than in areas representative of a homogeneous area. One principle underlying such studies is that society adapts to environmental differences; thus the purpose of study is to understand the nature of adaptations and how they function in the face of environmental change. Examples include studies of agricultural adaptation along an altitudinal or climatic gradient; investigation of the impact of proximity on market involvement in a developing country; or research on the human–environment interactions in different societies occupying similar environments, either across a social gradient within a similar environmental region, such as farmers in Africa, or living interspersed in the same location, such as pastoralist and agriculturalist land use management.

2.9 Time

A temporal approach to comparative case study examines the same place at different times, different places at the same time which reflect temporal differences, or different places at different times. In the first approach, a single place is examined at different times. Such a study might show the development of successive human–environment relations as population or other variables change through time. When a geographical location was occupied by different societies through time, such studies in geography were offered as evidence of 'sequent occupance (sic),' one of the strategies used to refute environmental determinism. Investigation of different places at the same time could be seen as a quasi-temporal process when diffusion of an innovation or process of change is spreading geographically and some areas represent an older core location with earlier adaptation while others represent preadaptation or emerging stages. An example might be the diffusion of a cash-cropping complex in a traditional rural society where various locations along a space–time gradient represent stages of adaptation of a new technology.

2.10 Events

In the evolution of comparative case study in the human–environment realm, one salient mode of analysis focused on human response to natural hazards. Natural hazards research sought commonalties, first, in human response to floods, followed by attention to other natural hazards, to the hazardousness of places, and eventually to technological hazards. In all of these realms, comparative case studies provided the core strategy for the development of general frameworks for natural hazard analysis. Seen more generally, selection of hazard case studies for comparative research may be organized around extreme geophysical events, such as tornadoes, hurricanes, and floods, or longer term environmental threats such as drought, deforestation, or soil depletion and erosion. Comparison was also based on spatial dimensionality and temporal occurrence of hazards, as well as on hazard geometry. The event-based approach can be extended to other kinds of events: human and ecological impacts of dams; local resistance to imposed political or technological change; or suburban sprawl.

2.11 Analogs

An analog is based on the premise that objects or events having some similarity may also have other similarities. Thinking by analogy in fact underlies the comparative case study approach, in that the similarity among cases makes them in fact analogous in some respect. As a formal organizing principle, analogs inherently involve two or more cases that share sufficient commonalty that insight can be gained by comparison. There is no formal procedure for developing analogs; rather, good analogies depend on 'pragmatic considerations regarding contexts, interests and purposes' (Jamieson 1988, p. 81) and represent *a priori* insight before the full development of the comparative cases. In comparative case study using analogs, there is no necessary juxtaposition of the cases in time or space; rather, comparison originated in a conceptual linkage that appeared promising to the researcher. Analogous case studies are strengthened when they are selected with explicit similarity criteria.

Comparative case studies developed within an analog framework include studies of the impact of modern technology in developing countries; society's response to acid rain, ozone depletion, and future greenhouse gas stabilization; responses to past climate change; and adaptation to drought. Often, comparative case studies built upon analogs are future-oriented, asking whether insights gained from the analogs can be applied to future decisions.

2.12 Theory

Comparative case studies can address theory, wherein the cases are selected to test hypotheses from the theory. Examples include comparative case studies of the process of agricultural change with respect to population growth; modeling society's progress through economic, demographic, disease, or mobility

transitions; testing Wittfogel's concept of the hydraulic civilization; or explicating a framework for understanding the processes of environmental adaptation. In such studies, individual cases are not only compared with other empirical cases, but compared also with the expectations from *a priori* theory.

3. The Comparative Case Study Method

The comparative case study approach draws upon the same kinds of method that are common to the case study and comparative methods. The researcher links a hypothesis to certain cases using one or more of the strategies described previously, gathers information about the cases, develops evidence from the cases, presents the cases, and constructs explanation or generalization from the evidence. Insight, creativity, intuition, and even serendipity cannot be dismissed as important elements that the researcher uses to formulate hypotheses and to develop the strategy of research.

The depth of analysis of the constituent cases in a comparative study selected for *instrumental* purposes may not be equivalent to that of an individual case examined for its *intrinsic* interest (Creswell 1997). Focus on an issue or hypothesis draws attention to specific relevant issues or variables in the cases. Yin (1989) referred to this difference as *holistic* versus *embedded* analysis, the latter connoting the selection of cases within a wider research question.

3.1 Tools for Analysis

Researchers developing new case studies for comparative analysis have a wide variety of available tools and data sources, including interviews, surveys, participant observation, ethnographic research, field observation, officially collected information, archival research, environmental analysis, remote sensing, qualitative analysis, quantitative analysis, systems analysis, structural analysis, and various modes of modeling. In contrast, researchers undertaking meta-analysis of previous studies may be confined to information developed through other researchers' selection of tools and information; the researcher may undertake further study to amplify information on the case. As with individual case studies, the results of comparative analyses are usually narrative, although normative, probabilistic, statistical, simulation, or graphic models are possible modes of conceptualizing and communicating research outcomes.

4. Evaluation of Comparative Case Studies

Although there are no formal rules for evaluating comparative case studies, there exist reasonable criteria against which studies can be assessed. These include establishing the relevance of the cases; thorough identification and consultation of information sources; the adequacy and accuracy of reflecting information from sources; correct interpretation of sources; objectivity in reporting and interpretation; and internal consistency as evidenced by triangulation or multioperationism. If the cases purport to be a 'natural' experiment, controls provide an important strengthening of the argument. Additional criteria include the possibility of replication; reporting of contrary evidence and/or interpretation; specification of exceptions and special circumstances; recognition of and explanation for contrary cases; rigor versus superficiality; and value-added insight and understanding. Primary case studies must also meet the canons of field or archival work in the relevant discipline; comparative case studies built from meta-analysis of previously published work have potential for review by original researchers.

5. Conclusion

Comparative case studies offer a richness of insight that may be lost in the abstractions of statistical analysis, while maintaining the tradition of the case study as a means to explore generalities in the complexity of the real world. The intriguing variety of strategies for organizing comparison means that this approach to understanding human relationships to the environment will continue to be a vital part of human–environment scholarship.

See also: Comparative History; Comparative Method, in Evolutionary Studies; Comparative Studies: Method and Design; Global Environmental Change: Human Dimensions; Human Ecology; Human–Environment Relationship: Indicators; Human–Environment Relationships

Bibliography

Creswell J W 1997 *Qualitative Inquiry and Research Design, Choosing among Five Traditions*. Sage, Thousand Oaks, CA

Eisenstadt S N 1968 Comparative study. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 14, pp. 421–9

Hamel J 1993 *Case Study Methods*. Sage, Newbury Park, CA

Jamieson D 1988 Grappling for a glimpse of the future. In: Glantz M H (ed.) *Societal Responses to Regional Climatic Change*. Westview Press, Boulder, CO, pp. 73–93

Leach E 1968 The comparative method in anthropology. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 1, pp. 339–44

Oreskes N, Shrader-Frechette K, Belitz K 1994 Verification, validation, and confirmation of numerical models in the Earth sciences. *Science* **263**: 641–6

Ragin C C 1987 *The Comparative Method: Moving Beyond Qualitative and Quantitative Analysis*. University of California Press, Berkeley, CA

Ragin C C, Becker H S 1992 *What is a Case? Exploring the Foundations of Social Inquiry*. Cambridge University Press, Cambridge, UK

Stake R 1995 *The Art of Case Study Research*. Sage, Thousand Oaks, CA
Yin R K 1989 *Case Study Research: Design and Methods*. Sage, Newbury Park, CA

C. G. Knight

Human–Environment Relationship: Indicators

In 1987 the World Commission on Environment and Development introduced a resonant idea: 'Sustainable development,' the commission wrote (1987, p. 43) 'is development that meets the needs of the present without compromising the ability of future generations to meet their own needs.' Sustainable development names a hope—that human prosperity can be reconciled with the natural world upon which all material economies ultimately rely. It is a notion that has exercised a broad influence in public discussions, notably in the United Nations Conference on Environment and Development in 1992 and the document *Agenda 21* (see Sitarz 1993) adopted at that conference in Rio de Janeiro.

Finding an operational definition of sustainable definition has proved difficult, however. In a dynamic global economy where substitutes for a wide range of goods and services are readily available—at a price—it is difficult to define 'needs' in a way that makes clear what must be conserved or preserved so that future generations can meet their needs (as they define them). In addition, we have yet to develop a generally accepted accounting system that can render authoritative judgments when an ecosystem verges on environmental bankruptcy. But there is no question that biodiversity is being lost at a historic pace, that fossil fuels cannot be replaced on the timescales on which they are now being burned, or that soil erosion diminishes the agricultural capability of some desperately poor people.

It is in this context that indicators play a significant role as people and societies search for ways to meet the requirements of nature and humans. Sustainable development is a long-term, large-scale phenomenon, if it is feasible at all, and it is likely to emerge slowly and unevenly in the world economy. Like other historical phenomena, sustainability may be hard or impossible to see at the level of everyday experience until it is well underway. Indicators rooted in scientific measurement provide an alternative way to perceive and to assess progress toward a sustainable future.

Indicators are quantitative measures of human well-being, economic activity, and natural processes and conditions; they are needed to sense the degree to which human activity may be continued or expanded in the future. The most useful indicators are repeated, precise measurements of natural and social phenomena. Such data series illuminate trends and identify changes; accordingly, indicators can sound alarms, define challenges, and measure progress. The measure-

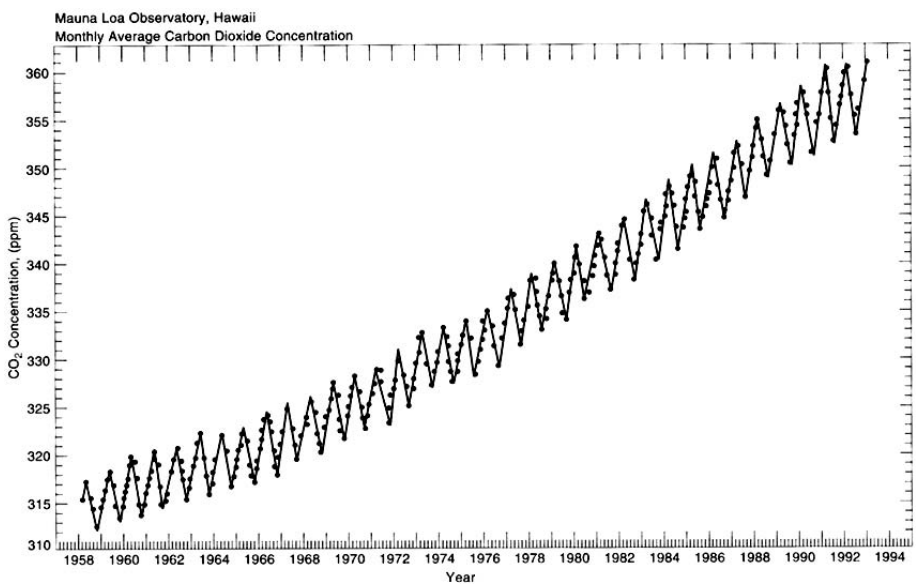

Figure 1
Atmospheric carbon dioxide concentrations at Mauna Loa, Hawaii

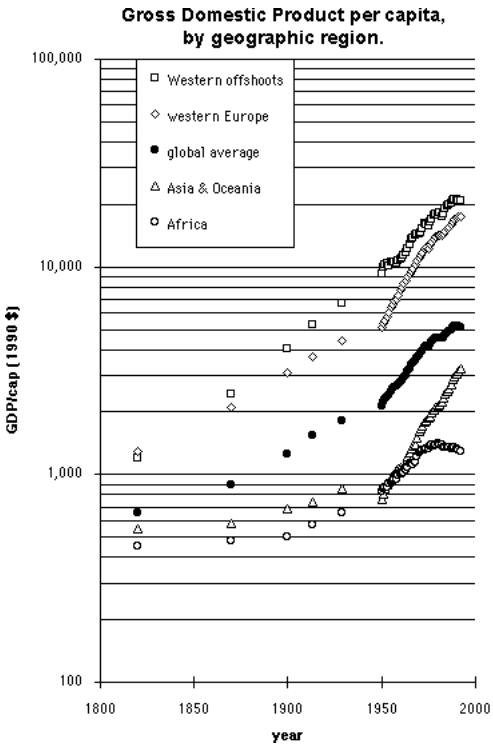

Figure 2
Estimated per capita gross domestic product, by region (source: Maddison 1995)

ment and interpretation of indicators requires sustained efforts and resources from governments, firms, nongovernmental organizations, and the scientific community. Indicators communicate, sometimes evocatively and often across cultural and national boundaries. The indicators described below illustrate how both social and natural realms have been affected by human action over the long term and on large scales.

Measurements of carbon dioxide levels in the atmosphere warn of climate change, but the warning is ambiguous in policy terms (see Fig. 1). Like other gases, carbon dioxide mixes and spreads in the atmosphere, so that its impacts are widespread even when its sources are localized. (Note the seasonal oscillations, however: blending within the northern and southern hemispheres takes place faster than global mixing.) Like other changes in climate, responses are delayed in time and reflect multiple perturbations, many not of human origin. Nonetheless, the steady long-term trend reflected in this data series has triggered worldwide scientific attention and international agreements. The attention commanded by this indicator is a measure of the influence of environmental concerns in formal decision-making, the arenas in which quantitative indicators are indispensable for communicating credible information.

Estimates of economic output demonstrate that economic growth has been substantial in all parts of the world over the past two centuries. Although even the poorest have seen their incomes increase, the gap between rich and poor has also widened (see Fig. 2). Over the next several decades, indicators will be needed to monitor and report on human welfare and life support systems, with measurements on different spatial and temporal scales to inform social choice and debate. Some of these indicators may come to define whether human populations are heading toward sustainability, in the way that infant mortality and life expectancy symbolize progress in public health. Some indicators can also aid navigation, helping individuals and societies to steer a course toward sustainability.

This article sketches current work on indicators and outlines an approach grounded on a widely shared set of normative objectives.

1. *Indicators of Sustainable Development*

Humans have made repeated, precise measurements of some phenomena since ancient times; astronomical and meteorological signs have guided agriculture for millenia. The use of indicators expanded as industrial societies sought to measure and manage a widening variety of environmental and societal parameters (see Porter 1995).

Today, many people routinely scan prices, news and weather reports, and other sources to inform action and expectation in the short run. Indicators can perform similar functions with respect to long-term, large-scale changes. A striking example of the influence of indicators is provided by the adoption of a common European currency at the turn of the century. In preparation for monetary union, the governments of the European Union agreed to harmonize economic policies by meeting and maintaining budget deficit targets, defined as specified fractions of gross domestic product.

For complicated social objectives such as sustainable development, no single indicator provides a definitive benchmark. Instead sets of indicators have been proposed, identifying multiple parameters that together provide assessments of three basic dimensions—prosperity, environmental quality, and human well-being. Indicators have usually been interpreted using the Pressure-State-Response (PSR) framework, which posits linkages between human action and environmental consequences (Adriaanse 1993). Human activities exert *pressures*, such as burning gasoline in cars, that alter the *state* of environmental variables, such as the quality of air in a city. Impaired states, in turn, elicit *responses*, such as regulations governing pollution-control technology. PSR makes use of data that often are already collected for administrative purposes. The approach captures an important element of sustainable development:

humans can impair the life-support systems of the natural world, calling forth responses aimed at protecting environmental quality. It is straightforward to expand PSR to include the impact of economic development on human welfare.

There is widespread agreement on the utility of PSR but no convergence on a single set of indicators for the measurement of sustainability. There have been numerous efforts to assemble indicators of sustainable development, however, ranging in jurisdictional scope from municipal to international and spanning ecological scales from watersheds to the planet as a whole. (For a compilation, see International Institute for Sustainable Development 2001.) Many indicators in these sets do not have long time series, so the ability to discern trends is limited. Coverage is uneven. Although environmental indicators outnumber the rest, there is greater depth in economic and some social dimensions. Many variables are available for nation-states but few at smaller spatial scales. Information on land cover and land use is weak, especially in developing countries, and institutional indicators are sparse. Some proposed indicators reflect considerable imagination, notably those proposed by the international Balaton Group (Meadows 1998) and in many community-scale projects. For instance, gardening activity was used as a measure of community vigor in Seattle, Washington.

Indicator sets assembled by the UN Commission for Sustainable Development (CSD) and the World Bank complement one another in an illuminating fashion. The CSD indicators (1996) were assembled using the PSR framework. Selected through a consensus process without an agreed operational definition of sustainable development, the CSD indicators are numerous, diverse in the methods used to measure development or sustainability, and include a large number of indicators where reliable measurements do not exist.

The World Bank (1997), in contrast, estimated three national capital accounts. Each attempts to capture the value of an aspect of national economies. 'Produced' capital includes physical capital and financial claims, both familiar in standard accounting practice. A second account measures natural capital, the resources and services provided by the natural world. In principle this includes standing timber and soil fertility. Estimates of natural capital are as yet primitive. A third component of wealth, quantitatively the largest, is human resources, the economic value of labor, knowledge, and social institutions.

The Bank's indicators are highly aggregated and have been estimated using drastic assumptions, but they are conceptually clear. The wealth of nations should be considered in three parts. At least at the margins, wealth can be transferred from one account to another in ways advantageous to people—as when productive farmland is fenced off to create a nature reserve. By contrast, the United Nations CSD indicators do not warn unambiguously of imminent hazards in any ecosystem or society, nor do they provide guidance on how to pursue sustainable development.

The United Nations indicator set does include variables that are not now being measured, however. This is a notable strength of its process, reflecting a realization that much of what needs to be sensed about sustainable development remains unclear. The UN also has sponsored a wide-ranging research effort under the aegis of the International Council of Scientific Unions (Moldan et al. 1997).

In sum, the existing indicators of sustainable development are limited in their effectiveness by the lack of agreement on what to develop, what to sustain, and for how long—that is, a lack of agreement on the meaning of sustainable development. Proposed indicators of this politically popular idea have drawn upon work done in past decades on the measurement of human welfare and the condition of the environment. These efforts bring together many sources of illumination but have yet to produce a set of measurements of social and natural conditions that could plausibly lead to prosperity for all while conserving the life-support systems on which human economies rest.

2. *Trends and Transitions*

That we do not have a clear path to a sustainable future is hardly surprising: identifying indicators that point to a sustainable future is a normative task, one that is inescapably difficult in a fractious world of multiple cultures. One approach, used recently by the US National Research Council (1999), suggests directions in which indicators of sustainability might evolve over the next several decades. In a report entitled *Our Common Journey*, the Research Council's board on sustainable development (BSD) focused attention on a sustainability *transition*, a decades-long process in which human economies might move toward a durable relationship between meeting needs and conserving nature. Instead of the elusive endpoint of sustainable development, the BSD sought to highlight the scientific, technological, and societal innovations required to 'feed, house, nurture, educate, and employ many more people than are alive today, while preserving the basic life support systems of the planet and reducing hunger and poverty.'

The underlying metaphor is the demographic transition, the long-term slowing of human population growth which is projected to lead to a stable population size in the latter part of the twenty-first century. The transition in population, already underway for more than two centuries, has brought the human species from a situation of relative stability with high birth rates and high death rates to a situation of stability at low birth rates and low death rates. Because death rates were lowered before birth rates, there has

been a large net increase in population size. A sustainability transition would correspondingly carry human economies from an agro-ecological past, through a rapidly growing industrial phase, into a future in which materials would be used far more efficiently, generating less pollution. As population sizes stabilize, moreover, wide disparities in material standards of living may be moderated by addressing hunger and poverty. Like the demographic transition, a sustainability transition would evolve over decades.

The sustainability transition is an optimistic scenario; signs that it may be in progress are mixed. The indicators above show rapid, perhaps exponential growth rather than stabilization. Yet forests have regrown in advanced industrial nations; famine has been significantly reduced in poor countries, even as the number of hungry people in the world continues to expand; and environmental concerns arose, perhaps surprisingly, in societies that suffered from some of the worst pollution. None of these indications points to a turning away from unbounded expansion, and there is nothing so clear as the peak in the rate of growth in human numbers which occurred in the late 1960s.

Calling for research to make a sustainability transition possible is an unsurprising message from groups of natural and social scientists. What is less apparent at first sight is the intellectual strategy and its implications for indicators. The sustainability transition eschews prediction in favor of a *normative* definition of the task ahead, to meet human needs while preserving the life-support systems of the planet. The BSD found a basis for defining human needs in treaties and the declarations of international conferences devoted to nutrition, housing, social welfare, education, and employment. While such statements articulate aspirations with few resources or implementation plans to back them up, on the decades-long timescale of a sustainability transition, one can imagine quantitative changes large enough to make a qualitative difference with respect to both human needs and the conservation of the natural world.

3. Navigating Transitions

Like the carbon dioxide trend above, a sustainability transition may well require quantitative indicators to be seen at all. The BSD proposed two families of indicators, one to perceive transitions and one to guide social choice in pursuit of deliberate transitions.

Seeing a sustainability transition using the BSD definition implies sensing the degree to which human needs are met and the degree to which life-support systems are put under stress. Human welfare has been monitored since the depression of the 1930s drew attention to the role of governments in meeting basic needs. Appropriate quantitative indicators have been developed, but at the turn of the century they were not yet being reported regularly in a forum, such as the annual report of the United Nations Development Program, where crises could be highlighted and emergent transitions discerned.

Environmental indicators are numerous and need conceptual organization. The BSD's proposal was geographic, focusing attention on four scales of significance to a long-term transition:

(a) *Global circulatory systems* are affected by human activity, including the atmosphere, climate and ocean circulation, trade and travel, and the spread of exotic species and diseases. Trouble in the circulatory systems is important because the scale of circulation can implicate the entire planet more rapidly or persistently than governments can regulate. Although greenhouse gases and ozone-depleting substances were carefully studied by the end of the twentieth century, biological and information-intensive flows were not yet being assessed as components of a global circulatory system.

(b) *Regions vulnerable to critical environmental damage* are being identified (Kasperson et al. 1995). For instance, the Aral Sea and the Valley of Mexico are experiencing severe, possibly irreversible damage. Because the governing factors—including a vulnerable natural setting, lack of feasible alternatives, and governing institutions unable to take effective action—vary among regions, there is not a single set of indicators that can monitor the combination of social and natural factors that lead to irreversible damage.

(c) *Ecosystem services for urban areas* are only beginning to be monitored systematically (e.g., Baltimore Ecosystem Study 2001), even though urbanization is continuing to accelerate in developing countries. Because many ecosystem services are unpriced (e.g., flood control) or allocated by monopolies (e.g., water supply), market signals are absent or misleading. Yet ecosystem services tend to be irreplaceable when damaged and inflexible because they are supplied by costly infrastructure.

(d) *Parks and protected areas*, managed to enable their biota to persist indefinitely, have been identified on a place by place basis, rather than through a consistent set of appraisals of their long-term sustainability. Attempts to conserve biodiversity on larger scales (Olson and Dinerstein 1998) are just getting started.

In each of these settings, indicators form an indispensable but incomplete part of the intelligence needed to discern a sustainability transition in the decades to come.

A complementary function of indicators is to inform public policy and other decisions, large and small. Some navigational aids are available, although it seems likely that others may also come into wide use.

(a) *National capital accounts*, emerging from the World Bank's work, could provide a simple but crude measure of sustainability: each of the three kinds of capital might be transformed by human activity, but so long as a nation's total capital increases over time,

its trajectory may lie in roughly the direction to contribute to a sustainability transition. The word 'roughly' is important: a utilitarian metric is inadequate to assess sustainability. But the capital accounts draw attention to transformations among forms of wealth—transformations that will continue through, and beyond, any long-term search for sustainability.

(b) *Assessing policies* using indicators assembled under the PSR framework can lead to adaptive management. Adaptive management treats policies as experiments, designing them so that lessons may be learned reliably from the implementation of policies, even those that fail (Lee 1993). In the PSR context, this means using pressure and state indicators to test the effectiveness of responses.

(c) *Monitoring ongoing transitions* provides analysis of trends, particularly those that carry large future commitments, such as population growth. Even if the population continues to stabilize, it will be essential to improve technology and energy efficiency so that humans can accomplish more, economically, with less impact on the natural world.

(d) *Surprise diagnosis* will be needed because a sustainability transition is an improbable development. Reflecting on environmental surprises over the past three decades, Kates and Clark (1996) concluded that we can learn from surprises, so that we can better anticipate, avoid, or mitigate their consequences. But knowing how to improve management is also a temptation to operate closer to the edge. Surprises could therefore become *more* frequent as humans gain better knowledge of the world; this possibility qualifies the conventional notion that science is valuable because it improves our ability to control or at least to predict danger. From this perspective, surprises are valuable indicators in themselves. It is well accepted that surprise should produce humility. Surprise should also produce curiosity. On the timescale of the sustainability transition, curiosity and the learning it prompts are likely to be important, whether or not control can be extended in the short term.

4. Indicators and Social Learning

There is no agreement about which indicators to use to measure the movement of economies toward or away from sustainability. One should accordingly expect spirited debates over the value, biases, and meanings of indicators. In the related sphere of economic policy one can observe over the past half-century sharp disputes in the US over economic growth, the incidence of poverty, unemployment, and inflation. All these have been indicators that American politicians think will influence voter behavior (Tufte 1978). Remarkably, the independence of the data-gathering and analytical organizations has survived, despite their location within government agencies (see Duncan and Shelton 1978).

This is an important lesson: the independence of science is central to the social value of scientific information. Another lesson is that surprise is itself a valuable indicator. Governments and societies should anticipate unexpected things to happen. In a policy context, this idea is a kind of precautionary principle: because surprise is likely, action should be undertaken with thought, humility, and caution.

Indicators used to report on a sustainability transition are likely to be biased, incorrect, inadequate, and indispensable. Getting the indicators right is likely to be impossible in the short term. But not trying to get the indicators right will surely compound the difficulty of discovering, inventing, and achieving sustainable development.

See also: Development: Sustainable Agriculture; Ecology, Political; Environment and Development; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Change and State Response; Environmental Cognition, Perception, and Attitudes; Environmental Justice; Environmental Planning; Environmental Policy: Protection and Regulation; Human–Environment Relationships; Sustainable Development; Sustainable Transportation

Bibliography

Adriaanse A 1993 Environmental Policy Performance Indicators. A study on the development of indicators for environmental policy in the Netherlands. April. Netherlands: Sdu Uitgeverij Konintginnegracht

Baltimore Ecosystem Study 2001 Web site: http://www.ecostudies.org/bes/ Visited 8/20/01

Duncan J W, Shelton W C 1978 *Revolution in United States Government Statistics, 1926–1976*. US Government Printing Office, Washington, DC.

International Institute for Sustainable Development 2001 *Compendium of Sustainable Development Indicator Initiatives and Publications*. Web site: http://iisd1.org/measure/compindex.asp. Visited 8/20/01

Kasperson J X, Kasperson R E, Turner B L II (eds.) 1995 *Regions at Risk: Comparisons of Threatened Environments*. United Nations University Press, Tokyo

Kates R W, Clark W C 1996 Environmental surprise: Expecting the unexpected. *Environment* **38** (2): 6–11, 28–34

Keeling C D, Whorf T P 2001 Atmospheric CO_2 records from sites in the SIO air sampling networks. In: *Trends: A Compendium of Data on Global Change*. US Department of Energy, Oak Ridge, TN

Lee K N 1993 *Compass and Gyroscope. Integrating Science and Politics for the Environment*. Island Press, Washington, DC

Maddison A 1995 *Monitoring the World Economy, 1820–1992*. Development Centre of the Organisation for Economic Co-operation and Development, Paris

Meadows D 1998 *Indicators and Information Systems for Sustainable Development*. A report to the Balaton Group. The Sustainability Institute, Hartland Four Corners, VT, pp. 78. Web site: http://www.sustainer.org/pubs/. Visited 8/20/01

Moldan B, Billharz S, Matravers R 1997 *Sustainability In-*

dicators. A Report on the Project on Indicators of Sustainable Development. Wiley, New York
National Research Council (USA) 1999 *Our Common Journey: A Transition Toward Sustainability. Report of the Board on Sustainable Development*. National Academy Press, Washington, DC. Web site: http://lab.nap.edu/catalog/9690.html. Visited 8/20/01
Olson D M, Dinerstein E 1998 The Global 200: A representation approach to conserving the earth's most biologically valuable ecoregions. *Conservation Biology* **12**: 502–15
Porter T M 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ
Sitarz D (ed.) 1993 *Agenda 21: The Earth Summit Strategy to Save our Planet*. EarthPress, Boulder, CO
Tufte E R 1978 *Political Control of the Economy*. Princeton University Press, Princeton, NJ
United Nations Commission for Sustainable Development 1996 *Indicators of Sustainable Development. Framework and Methodologies*, United Nations, New York
World Bank 1997 *Expanding the Measure of Wealth: Indicators of Environmentally Sustainable Development*. World Bank, Washington, DC
World Commission on Environment and Development 1987 *Our Common Future*. Oxford University Press, New York

K. N. Lee

Human–Environment Relationship: Prehistoric

Human–environmental relationships comprise a web of culturally imposed constraints in a complex environment, within which people adopt a variety of subsistence pursuits. These may or may not have long-term effects on that environment. During the course of Pleistocene prehistory, humans were few in numbers and constrained by their technology and social organization from unlimited spatial expansion or assuming the trophic role of a super predator. But they may have influenced megafaunal extinctions or vegetation change through manipulation of fire. By the beginning of the Holocene, greater subsistence success allowed exponential population growth, coinciding with agricultural origins in several world regions. Domesticated plants and animals became artifacts, allowing humans to achieve ecological dominance and incrementally transform their environments. But the management strategies followed and the landscapes created were constructed culturally.

1. *Inter-relationships*

Human–environmental inter-relationships are complex and nuanced. On the one hand, resources vary over time and are unevenly distributed in geographical space, so that strategies for utilization will differ for the short and long term, and at small or large scales (Butzer 1982). On the other, human communities subscribe to a wide array of subsistence and cultural modes, and attempt to reconcile new information with traditional values. Alternative pathways for possible change are evaluated through economic and cultural choices, which presume a major role for cognition, information, and decision-making. Environmental factors set 'boundary conditions' to human behavior at a particular technological and organizational level only in so far as they pose socially unacceptable risk. Short- or long-term environmental crises may be met by routine accommodation, active mitigation, or shifts in subsistence pursuits or social organization. Outcomes therefore are unpredictable.

2. *A Pleistocene Prelude to Change*

The roughly two million years of Pleistocene prehistory—the Paleolithic—were critical for human evolution and presumably influenced the 'hard-wiring' of basic human behavior. Emerging human–environmental interrelationships not only suggest a datum to examine such patterns in more recent times, but may have set a trajectory for later environmental behavior. Given the rudimentary archaeological record and in the absence of direct 'observation,' it is next to impossible to grasp the cognitive assessment of environmental problems, or even to infer a response or to understand attenuated processes of systemic change. A very real danger in Pleistocene prehistory is to posit a cause and effect relationship, based on little more than coincidence of several categories of proxy data. There are no simple answers to the broad, interdisciplinary questions about Pleistocene human–environmental relationships.

2.1 *Human Vulnerability to Pleistocene Environments*

The expansion of early human groups to higher latitudes and new continents during the course of the Pleistocene gives the impression that they learned progressively to cope with cold or arid environments, where resources were different than in tropical Africa. But the Paleolithic record is strikingly discontinuous in space and time, raising doubts as to the ability or choice of Paleolithic peoples to colonize 'marginal' environments with widely dispersed but clustered resources (Butzer 1982, Chaps. 12–14, 1991).

In what is now the semiarid interior of southern Africa, settlement was limited to periods of substantially moister climate and closely linked to water sources. Even some valleys with dependable rivers were unutilized for repeated spans of many tens of

millennia, and not all periods of moister climate were accompanied by an automatic territorial expansion of settlement. Effective human occupation across the dry interior of southern Africa was delayed until well into Holocene times, after 10,000 before present (BP).

Paleolithic people probably avoided this region because food resources would have been few and thinly spread during the dry seasons of poor years. Plant foods would have had low predictability and low net productivity, while the available game would have been difficult to hunt without bow and arrow, sophisticated traps, or complex game drives. If bands of hunters did penetrate such dry regions, they would have been so widely dispersed as to encounter neighboring bands only sporadically, leaving them with incomplete information as to the status of resources, e.g., animal movements or the condition of plant foods (Butzer 1991). During periods of severe drought stress, starving animals also lack fat while excessively lean meat poses severe metabolic problems for starving people (Speth 1987).

The true capacity of drought-prone, low productivity, and low predictability environments (see Butzer 1982) to sustain human life is a great deal less than its overall resources, averaged over time and space, would suggest. A similar 'marginality' model is offered by Bocquet-Appel and Demars (2000) for western Europe between 40,000 and 11,000 BP In glacial-age France settlement was concentrated in a 'refuge zone' located in northern Aquitaine. During less harsh, subarctic conditions prior to 22,000 BP small temporary sites extended northeastward to the Rhine, but then retracted to southwestern France during the glacial maximum, c. 22,000–16,000 BP Thereafter, as the climate ameliorated, an increasing number of larger sites are found in areas previously not occupied, as far afield as Britain and Germany. Across these 30 millennia, regional population may have increased from 4,000 to as many as 40,000 people, in part facilitated by improved hunting skills.

There was a vast 'marginal' environment in western Europe where not drought, but the low productivity and predictability of a subarctic environment, with dangerous turns of weather and hard, snowy winters, made glacial-age occupation hazardous except for summer-season hunting forays. Risk-prone environments such as semiarid thorn bush or arctic tundra seem to have been avoided, in favor of demographic clustering in dependable, productive environments. The persistent and extensive spatial gaps of the Paleolithic archaeological record in better-investigated areas appear to be real, rather than simply a function of site visibility and preservation.

Until the close of the Pleistocene, human groups did not demonstrably 'push the envelope' by venturing far or long into marginal environments. Incremental human expansion into higher latitudes, especially during the late Pleistocene, probably targeted environmental mosaics capable of supporting sizeable population clusters on a long-term basis. That implies a cognitive ability to explore new regions, so as to assess viability and risk, in the context of available technology and subsistence skills.

Changing hunting proficiencies are demonstrated by quantitative differences in the animals slaughtered in South African or western European sites of different periods. Before about 40,000 BP hunting focused on relatively docile animals, whereas in Late Palaeolithic times more dangerous animals or those from difficult montane environments were hunted regularly, and aquatic resources were systematically exploited on defined, seasonal rounds (Klein 1999). In part this reflected the invention of the bow and arrow, more versatile throwing javelins, and a variety of harpooning and fishing gear. But better tactical and collaborative organization of hunting and fishing activities will have been just as important.

The exponential increase in hunting efficiency evident across Eurasia between 25,000 and 10,000 BP implies a revolutionary growth in food-energy exploited, making possible rapid demographic expansion during the closing millennia of the Pleistocene. In both Cantabrian Spain and interior South Africa site sizes or artifact concentrations increase exponentially during the Late Pleistocene, and there were some 30 to 70 times as many sites (per millennium) about 10,000 years ago than there were about 100,000 BP. The 'take-off' in site numbers and implicit population growth begins at different times in different regions, namely 22,000 BP in Cantabrian Spain, 16,000 BP in France, 14,000 BP in northwestern Siberia, 10,500 BP in North America, and 8000 BP in temperate Australia (Butzer 1991, Bocquet-Appel and Demars 2000).

This carries implications for the relative proficiency of prehistoric peoples to utilize environmental resources. Paleolithic populations remained very small and vulnerable across nearly two million years, precisely because they lacked the technology and cognitive ability to move higher up the trophic pyramid. They only achieved the role of a major predator during the closing millennia of the Pleistocene.

2.2 *Late Pleistocene Megafaunal Extinctions*

The disappearance of mammoth, mastodon, woolly rhino, great sloths, or sabre-tooth 'tigers' from the Holarctic fauna at the end of the Pleistocene continues to intrigue. Giant kangaroos and cow-sized diprotodons from Australia also belong on this list of 'megafauna' (Martin and Klein 1984). But unlike the end-Cretaceous extinction of the dinosaurs, those of the terminal Pleistocene did not eliminate the dominant orders of larger animals, but fell within the scope of evolutionary 'turnovers' incompletely documented for earlier phases of the Pleistocene.

Species extinctions are a grim reality of recent centuries, and can be attributed to chemical impacts,

habitat destruction and harassment, predation by overhunting, or deliberate extermination. Some such extinctions or near-extinctions were rapid, others followed many centuries of sustained attrition. But for Late Pleistocene megafaunal extinctions there is neither direct testimony nor a promising analytical method to identify cause and effect, such as for the demise of the carrier pigeon, or the near-extinction of bison, bald eagle, or whooping crane.

The scope of this problem is illustrated by the catastrophic extinctions on New Zealand after its first colonization by the Polynesian ancestors of the Maori, less than a millennium ago (Anderson and McGlone 1992). Two families of large wingless birds (moas) became extinct, as well as an array of songbirds and marine taxa, all prior to European intrusion during the early 1800s. Although moa extinction within 4 centuries is well-enough documented, together with butchery, woodland fragmentation, and introduced dogs and rats, we will never know the human dimensions of this event, on the very threshold of oral history. Yet catastrophic extinctions on isolated islands of the Pacific Ocean during the last 2 millennia involved fewer variables than the response of more complex, continental ecosystems to a familiar human predator before 10,000 BP. These recent examples do not therefore provide a model applicable to the Late Pleistocene.

The case of extinctions in Australia makes the point. Between perhaps 30,000 and 14,000 BP a total of 42 larger vertebrates, mainly marsupials, and several taxa were subject to size reduction. Two of the vertebrate families that disappeared were huge, clumsy and slow-moving browsers, reminiscent of large mammals of the early Tertiary. The kangaroo family (Macropodidae) lost six genera (or 30 species), but it could be argued that the surviving kangaroos and wallaby (10 species) are 'modernized,' efficient, and smaller forms of the same lineages (phyletic replacement). In the big picture, this suggests the consequence of intergeneric and interspecific competition.

There also were several major environmental discontinuities during the same time span, including filling and drying of lakes, expansion and 'downsizing' of river systems, and extensive activation of eolian sands (Ross et al. 1992). Pollen profiles suggest re-expansion of browsing habitats about 10,000 BP (Dodson et al. 1992). In combination with the sudden, drastic climatic perturbations for which Australia is noted today, that implies repeated episodes of fragmentation and reduction of habitats during the Late Pleistocene, rather than a single and unidirectional change at its close. Environmental flux of sufficient amplitude would selectively disadvantage the demographic viability of 'archaic' forms, accelerating competitive exclusion among genera with overlapping niches.

People settled parts of Australia before 40,000 BP, but they were few in number, their distribution limited to more productive environments, and they introduced no animals other than the dingo, at a much later date (Kohen 1995). Artifacts have been found at a handful of faunal sites that lack patent evidence of butchering activity. How does one link the coexistence of human hunters or scavengers with extinctions over a time range of 25 millennia? Alternatively, how does one distinguish an evolutionary turn over from human predation? Temporal correlations do not provide process-links.

The case of North American extinctions is equally puzzling. Seven common genera disappeared 12,000–10,000 BP, at a time of demonstrated human predation, but of the 31 other extinct genera or species (all poorly dated), at least several phased out while even a human presence on the continent is contested (Grayson 1991). Twenty-seven of the extinct genera and six extinct species were found only in North America. In effect, this great faunal turnover primarily impacted New World forms, many of which were 'archaic,' to select for a smaller group of expansive ruminants (bison, deer, pronghorn), and carnivores (bear, wolf, coyote). But not all niches were refilled, and Spanish equids successfully recolonized grassy shrublands of the American West and Argentina.

Late Pleistocene environments in North America have no modern analogs: both grasslands and woodlands were 'patchy,' with fine-grained mosaics, greater local plant diversity, and unfamiliar seasonal productivity patterns (Graham et al. 1996). Both the extinct grazers and browsers were unspecialized, filling their energy requirements with a broad range of plants, as is typical for monogastric ungulates. At the close of the Pleistocene, such 'patchy' biomes were replaced by monotonous grasslands or closed-canopy forests, more suitable for specialized ruminants (bison, deer, caribou), whose stomachs can degrade plant toxins and more rapidly assimilate energy. All the large, monogastric grazers (mammoth, horse) and mixed browsers (mastodon, ground sloths, camelids), which had long gestation periods, became extinct. Also symptomatic of environmental stress, but inadequately understood, is the progressive reduction of size ('dwarfing') of the disappearing taxa during the Late Pleistocene, as is the 'dwarfing' of phyletic replacements among surviving genera worldwide.

Unlike Australia or North America, the contiguous continents of the Old World were little affected by extinctions (Martin and Klein 1984): a few phyletic replacements here and there, and the more dramatic disappearance of elephant and rhinoceros in Palearctic Eurasia for the first time since the Tertiary. Unimpeded contact and faunal interchange in the Old World implies a steady intergeneric competition for overlapping niches, which may have driven Old World faunas to modernize earlier. As on the other continents, that does not exclude human predation as a triggering mechanism for extinction of some vulnerable taxa. But neither does the present evidence single

out humans as super-predators ten millennia ago, whatever the environmental inputs and biological feedbacks of end-Pleistocene extinctions.

2.3 Fire Ecology and Vegetation Change

The human use of fire is a potent tool of habitat alteration or destruction. That is supported by both ethnohistorical evidence and quantification of charcoal particles from pollen profiles. Fire depends on the availability of fuel, climate, and ignition. 'Natural' fires, ignited by lightning strikes, have facilitated expansion of grassland and savanna since mid-Tertiary times (Kershaw et al. 1997). Less clear is the degree to which prehistoric peoples manipulated fire so as to alter vegetation ecology, because it is difficult to separate the evidence for artificial vs. 'natural' fires.

In Australia, indigenous people continue to use fire for many purposes: to drive game, to clear undergrowth for visibility and stimulate fresh growth, and to serve a range of defensive, domestic, and social purposes. The popular truism that indigenous fires were 'controlled' habitually is not supported by the facts, nor is it consonant with the experience of wildfires on other continents. Fire was an instrument of environmental management in that it once kept most Australian eucalyptus woodlands grassy and free of shrubby undergrowth, although such woodlands had already evolved across five million years in an interactive relationship with fire.

Pollen profiles from south-eastern Australia that cover most of Holocene time show charcoal particles increasing exponentially during the last 2 millennia or so, to culminate shortly before or after AD 1800, even though there is no consistent climatic trend during that period. It is possible that the elaborate burning culture recorded after 1788 only evolved during the late Holocene, after changes in cultural practices. The density of archaeological sites in south-eastern Australia doubled during the last two millennia, compared with the previous three (Dodson et al. 1992), suggesting that increasing use of fire primarily reflected demographic expansion or greater mobility. Given the much smaller and disjunct Australian populations of the Late Pleistocene, it seems improbable that human manipulation of fire could have produced a striking charcoal peak 38,000 BP in northern Queensland (see Kershaw et al. 1997), since the 'event' saw rainforest replaced by eucalyptus woodland, which instead implies climatic change. To argue that early use of fire was a co-agency in megafaunal extinctions will require a comprehensive, ecological case and an archaeological context.

Unlike Australia, which has a desertic core, sub-Saharan Africa is dominated by grasslands with scattered trees (savannas), mosaics of savanna and open woodland, or open woodland with patches of grass. Although extensive tracts of such open vegetation reflect either sandy substrates or extensive seasonal flooding, others have been modified and maintained by recurrent fire. It was once believed that Africa's savannas were created and expanded progressively by deliberate burning through much of human prehistory, but that is no longer tenable. The great diversity of cursorial ungulates for which Africa is noted began to evolve during the Miocene, while Pliocene deposits reveal abundant pollen of grasses as well as evidence of charred wood. During glacial-age intervals of drying, such as 30,000–10,000 years ago, much of sub-Saharan Africa was an almost treeless grassland, with woodland or forest reconstituted at the beginning of the Holocene.

Pollen profiles from the highlands of Malawi show that a mosaic of grassland and forest remained unchanged from the Pleistocene to the present (Meadows 1984). In interior South Africa, the distribution of grass and semidesert shrub shifted according to climatic trends and was little affected by the first appearance of indigenous herders. Until agricultural activities began to create patches of disturbed vegetation in tropical Africa, mainly after 3000 BP, increases in grass pollen and charcoal currently are attributed to episodes of drier climate, thus the opening of the Cameroon 'Grass Fields,' formerly believed to be anthropogenic (Zogning et al. 1997). In Africa and elsewhere it is important to distinguish slash-and-burn agriculture, as a focused form of micromanagement within forest environments, from the modification of savannas or open woodlands to attract game or improve pasturage for livestock. Prehistoric fire ecology primarily concerns the latter, and assumes that deliberate burning reinforces natural processes.

3. Changing Holocene Relationships with the Environment

Regional population growth at the end of the Pleistocene sooner or later set in train a variety of adaptive responses to new or changing environmental opportunities. In some areas, increasingly skilful hunter–foragers turned to a more specialized exploitation of coastal resources, while others remained more conservative generalists, despite technological innovation. Elsewhere hunter–foragers chose to diversify their ecological niche by investing more time in the harvesting of large numbers of small mammals, at least on a seasonal basis. Tapping into a more predictable food supply, such strategies repaid the effort by allowing a more sedentary residential pattern, even if small hunting parties fanned out in pursuit of bigger game at opportune times. Longer-term residence at central encampments, in caves or out in the open, in turn allowed a more persistent exploitation and manipulation of diversified plant foods.

In several world regions this strategic shift led to the domestication of plants or animals, or both. This was

not a socioeconomic revolution, but a long trajectory of co-evolution, during which people experimented with and eventually 'managed' a selection of plant and animal resources. They favored and tended more desirable strains until new domesticated species emerged, to be maintained through deliberate planting and breeding. This inaugurated a novel set of relationships between people and the environment in that both plants and animals became artifacts with which people fundamentally changed their habitats, involuntarily or deliberately replacing native biota with cultivated plants and domesticated livestock.

3.1 The Impacts of Early Agriculture

Agricultural origins have been amply documented by multidisciplinary archaeological research in the Near East and in southern Mexico; similar processes are apparent, but imperfectly understood in Andean Peru, sub-Saharan Africa, and south-eastern Asia. Morphological ('hard') evidence for domesticated grasses and legumes is available for sites along the Near Eastern steppe-woodland ecotone by 10,500 BP, implying that the process of experimentation and manipulation began several millennia earlier. Agricultural origins in the New World may have been even more attenuated, but that database is spotty.

As long as communities practicing complementary agriculture remained few, small, and confined to the natural habitats of their cultivated plants and livestock, environmental impact would be minor. But these new pursuits were contagious across the Near East and much of Europe, introducing new plants or animals to environments where indigenous biota were displaced. Agricultural villages appear on the coasts of Greece, Crete, and Cyprus by 8,000 BP, and a distinctive new agrosystem had emerged around the coastal plains of the western Mediterranean about 7,000 BP In some areas, resident hunter–foragers domesticated local wild cattle prior to agricultural penetration (Yugoslavia), while elsewhere expanding agriculturalists allowed outbreeding of livestock or cultivars (Hungary), to produce new domesticated variants. Dispersal across western Europe before 5,000 BP proceeded in stages, each with a probable sequence of ecological readjustment in new settings, indigenous innovations, interchange between resident hunter–foragers and intrusive farmers, and eventual indigenization of agricultural traits.

This early, Neolithic agriculture was localized and often ephemeral, and most pollen evidence shows that woodlands soon regenerated. But strong counter-examples verify repetitive fire clearance, cultivation, and stockgrazing that converted extensive areas into grass, weeds, or secondary vegetation 6,000–5,000 BP (Poland) (van Zeist et al. 1991). In parts of the Mediterranean Basin, early Neolithic pastoralists probably used fire clearance to open up pasturage in native woodlands as early as 7,000 BP (Greece, Dalmatia, eastern Spain), with reforestation delayed up to five centuries. Such impacts were local rather than general, but in some areas they may have initiated or reinforced a trend to more xeric forest types. Yet settlement was spatially discontinuous, even in 'preferred' environments, until about 4,000 BP, judging by repeated episodes of woodland recovery and presumed local abandonment (Butzer 1996).

3.2 Plow Agriculture, Pastoralism, and Ecological Change

Towards 3,500 BC (5,400 BP) the invention of the plow (Mesopotamia) marked a significant transformation in the complexity and productivity of agriculture. The plow required a draft animal and opened larger fields to cultivation, while also exposing soil to potential erosion. Animals no longer served only as meat on the hoof. Cows, sheep, and goats were milked, and the milk converted to butter or cheese. Wool was skeined and woven into cloth as new breeds of woolly sheep were selected. The fertilizing properties of animal manure were probably recognized, as animals and their products became interlinked systemically with the range of subsistence activities.

More work was invested in a plot of land, including preparation, planting, weeding, and harvesting, to increase productivity. Plowing along hillsides deliberately or inadvertently created stepped slope profiles that may have suggested the principle of terracing. Swamps were drained and heathlands ameliorated with transported soils. In arid environments, opportunistic local manipulation of runoff developed into more complex methods of water control, including canal irrigation. Orchard crops, olive oil, and wine emerged as commercial products. Exchange networks and growing urban markets further reduced risk and, spurred by metals, metal products, and pottery, began to favor economic integration over wider areas.

Particularly in the Mediterranean Basin, much land was unarable, but animals were mobile and could be moved to rough upland pastures when the snows melted or lowlands became subject to summer drought. Short and long distance mobility patterns (transhumance) emerged that brought stony uplands into productivity. Controlled burning began to be used to manage such pastures on a regular basis, keeping woodlands open (but not deforested) and burning back woody or thorny shrubs. By the end of the Bronze Age (1,200 BC), much of the Mediterranean world was used and modified by farmers and herders, even though vast tracts of montane forest remained. A familiar cultural landscape was in place (Levant, Greece), expanding to the far west by Roman times, and bringing a simplified version to higher latitudes during the Middle Ages Ages (van Zeist et al. 1991).

Intensified agriculture represented a quantum jump in production, demography, and interregional

networking, in managed environments that were increasingly artificial. Native biota were replaced progressively, with cultural controls over natural selection, succession, and diversity. To maintain such managed ecosystems required constant or increasing investment of labor, in order to maintain artificial equilibrium conditions and avoid ecological degradation (simplification). This was the case in the Bronze Age Mediterranean Basin and Near East, as well as China. A millennium later, similar transformations are apparent in parts of Mesoamerica, the Andean world, and several nuclei in southeastern Asia.

Changing human–environmental relations were not limited to agricultural regions. In the central USA sedentary utilization of rich, riparian resources incorporated controlled use of a variety of minor plant foods well before the appearance of standard domestication. In arid lands and the sub-Arctic, or in the high Andes, mobile herding patterns emerged that were based on new domesticated genera such as camel, reindeer, and llama. Even in the Arctic, groups of hunter–foragers devised new collaborative methods to effectively exploit coastal resources, by mobile strategies such as those of the evolving 'Eskimo' cultures (Alaska to Greenland). A mastery of the environment is broadly evident that was qualitatively and quantitatively different from that of the Late Pleistocene.

With increasing environmental exploitation by expanding populations, the potential for ecological impairment or degradation was rapidly enhanced. But the fact that discernible damage did not keep pace suggests that farmers built on the cumulative experience of trial and error in familiar environments, to articulate more conservationist strategies (Butzer 1996). Ideally, land use seeks to minimize both long-term environmental damage and short-term subsistence risk. Pollen and geomorphologic records show that for the most part, Mediterranean farmers and herders were successful in doing so during the last four millennia, despite sporadic 'accidents' of mismanagement in some areas. Only during Late Medieval times, when population was substantially greater than in Classical times, was supplementary food imported from outside the Mediterranean region. Until the early 1900s, productivity was maintained or enhanced, which implies sustainability across 7000 years of agricultural land use (Butzer 1996).

An anomalous episode of disturbance is dated to the transition of the Bronze and Iron Ages, c. 1,000 BC. It saw intense and often prolonged vegetation disruption (Spain, northern Greece, Northwest Turkey) by pastoralists moving into the Mediterranean environment from temperate Europe or the Balkans. Presumably they misapplied management methods better suited to humid environments. Much the same happened during the Middle Ages, after desert nomads moved into the Levant and North Africa, or Slav herders occupied northern Greece. In other cases initial agricultural occupance led to a flush of soil erosion (Greece), and soil loss was locally problematic after pastoral expansion or agricultural decline, thus during the later first-millennium BC (Palestine, western Turkey, Greece, Italy, Spain) and again in Byzantine or Medieval times (Palestine, Greece, Italy, North Africa, western and central Europe). Although the cultural landscape and its humanized woodlands were repeatedly rehabilitated, soil thinning had to be compensated for with manure, but cumulatively this does not compare with watershed destruction by mechanized agriculture or industrialization since the 1950s.

In the New World, where the focus of recent discussion has been on the impacts of European-style land use, more pragmatic work shows extensive forest disturbance or removal, and even soil erosion, during indigenous demographic expansion many centuries before 1492 (Central America, Mexico, eastern United States) (Pohl et al. 1996, Butzer and Butzer 1997, Peacock 1998).

The human imprint on the landscape at the threshold of history was profound. In combination with the devastation of island ecosystems, such as New Zealand or Madagascar, it includes episodes of biotic impairment and even extinction, with enduring imprints. Yet most environments today, even 'wilderness' areas, are humanized to some extent and many if not most agricultural landscapes provide esthetic comfort and psychological anchoring in their particular cultural contexts. The human ascent to unprecedented demographic success and ecological dominance has impoverished floras and faunas the world over, often replacing them with vast monocultures or a handful of livestock species—even before the side-effects of industrialization could pollute the waters or turn machinery loose on the soil mantle. But until the dawn of a global consciousness, the transformations of the last 10 millennia were guided by community decisions, made through culture-specific lenses, with the goal of subsistence success across the generations, guided by imperfect information, despite conservative and conservationist strategies. That is the *reality* of human-environmental relationships, however they are judged today in the face of global ecological crisis.

See also: Ecology, Deep; Environmental and Resource Management; Environmental Cognition, Perception, and Attitudes; Human–Environment Relationships; Hunter–Gatherer Societies, Archaeology of; Hunting and Gathering Societies in Anthropology; Land Degradation; Sustainable Development

Bibliography

Anderson A, McGlone M 1992 Living on the edge: Prehistoric land and people in New Zealand. In: Dodson J (ed.) *The Naive Lands: Prehistory and Environmental Change in Australia and the South-west Pacific*. Longman Cheshire, Melbourne

Bocquet-Appel J P, Demars P Y 2000 Population kinetics in the upper palaeolithic in Western Europe. *Journal of Archaeological Science* **27**: 551–70
Butzer K W 1982 *Archaeology as Human Ecology: Method and Theory for a Contextual Approach*. Cambridge University Press, Cambridge, UK
Butzer K W 1991 An old world perspective on potential mid-Wisconsinan settlement of the Americas. In: Dillehay J D, Meltzer D J (eds.) *The First Americans: Search and Research*. CRC Press, Boca Raton, FL
Butzer K W 1996 Ecology in the long view: Settlement histories, agrosystemic strategies, and ecological performance. *Journal of Field Archaeology* **23**: 141–50
Butzer K W, Butzer E K 1997 The 'natural' vegetation of the Mexican Bajío: Archival documentation of a 16th-century savanna environment. *Quaternary International* **43**(44): 161–72
Dodson J, Fullagar R, Head L 1992 Dynamics of environment and people in the rested crescents of temperate Australia. In: Dodson J (ed.) *The Naive Lands: Prehistory and Environmental Change in Australia and the South-west Pacific*. Longman Cheshire, Melbourne, Victoria
Graham R W, Lundelius E L, FAUNMAP 1996 Spatial response of mammals to late quaternary environmental fluctuations. *Science* **272**: 1601–6
Grayson D K 1991 Late pleistocene mammalian extinction in North America: Taxonomy, chronology, and explanantions. *Journal of World Prehistory* **5**(3): 193–231
Kershaw A P, Bush M B, Hope G S, Weiss K F 1997 The contributions of humans to past biomass burning in the tropics. In: Clark J S, Cachier H, Goldammer J G, Stocks B (eds.) *Sediment Records of Biomass Burning and Global Change*. Springer, Berlin
Klein R G 1999 *The Human Career: Human Biological and Cultural Origins*, 2nd edn. University of Chicago Press, Chicago
Kohen J 1995 *Aboriginal Environmental Impacts*. University of New South Wales Press, Sydney
Martin P S, Klein R G (eds.) 1984 *Quaternary Extinctions: A Prehistoric Revolution*. University of Arizona Press, Tucson, AZ
Meadows M E 1984 Past and present environments of the Nyika Plateau, Malawi. *Palaeoecology of Africa* **16**: 353–90
Peacock E 1998 Historical and applied perspectives on prehistoric land use in eastern North America. *Environment and History* **4**: 1–29
Pohl M D, Pope K O, Jones J G, Jacob J S, Piperno D R, de France S D, Lentz D L, Gifford J A, Danforth M E, Josserand J K 1996 Early agriculture in the Maya lowlands. *Latin American Antiquity* **7**: 355–72
Ross A, Donnelly T, Wasson R 1992 The peopling of the arid zone: Human-environment interactions. In: Dodson J (ed.) *The Naive Lands: Prehistory and Environmental Change in Australia and the South-west Pacific*. Longman Cheshire, Melbourne, Victoria
Speth J D 1987 Early hominid subsistence strategies in seasonal habitats. *Journal of Archaeological Science* **14**: 13–29
van Zeist W, Wasylikowa K, Behre K E (eds.) 1991 *Progress in Old World Paleoethnobotany*. Balkema, Rotterdam
Zogning A, Giresse P, Maley J, Gadel F 1997 The late holocene palaeoenvironment in the Lake Njupi area, West Cameroon: Implications regarding the history of Lake Nyos. *Journal of African Earth Sciences* **24**: 285–300

K. W. Butzer

Human–Environment Relationships

The relations between human activities and their biophysical milieu have never been a subject of deep and sustained interest to the Western social and behavioral sciences. Such attention as they have been given has been fitful and partial, and it has arguably always been much less than their importance merits. Environment—in the sense of the biophysical world—has not been a topic that the orthodox human sciences have felt comfortable in addressing. This article traces the general scope of human–environment relations and of the past and present study of them in the human sciences.

1. *The Term 'Environment'*

'Environment' is a troublesome word in that it can mean quite different things in different contexts. In some fields, it has long been used to refer to the social surroundings of human existence and behavior. In that sense, it encompasses much of the standard subject matter of the social and behavioral sciences. Elsewhere, it denotes the biophysical surroundings of human activity. In that sense, it encompasses much of what the social and behavioral sciences have deliberately chosen not to study. An emphasis on 'environment' is associated confusingly in the one case with the rejection of 'nature' as an important factor (as in the 'nature–nurture debate') and in the other with its affirmation (as in 'nature–society relations'). An 'environmentalist,' depending on the context, can be one who believes that social influences outweigh hereditary ones in human development (a frequent meaning of the term today in psychology and sociology); that human beings are molded by climate, terrain, and other elements of their biophysical surroundings (the meaning given the term in the 1968 *Encyclopedia of the Social Sciences*); or that vigorous action should be taken to slow or reverse human-induced change in earth systems (the usual meaning of the term in ordinary parlance today). Only the second and third of these meanings correspond to that of 'environment' taken, as here, in the biophysical rather than the social sense. Both of them, moreover, exemplify a second problem: the frequent usurpation of the term by a narrow and particular sense of it that obscures its full scope.

Even if 'environment' denotes not the totality of the biophysical universe, but only those portions of it that are relevant in some way to human activity, its boundaries are not easy to fix. They can and often do change as human activities change or as understanding of their relations with their surroundings does. At or near the boundaries, the environment can include the mineral deposits of the deep ocean floor (now that the technology for their exploitation exists), the strato-

spheric ozone layer (now that its role in protecting life on the earth's surface is recognized), the stars when they were used for navigation, and a comet or an asteroid threatening to hit the earth. If the outer limits are always uncertain and always in flux, however, the core elements are easily defined: climate, soil, terrain, bedrock, mineral and fossil fuel deposits, atmosphere, fresh and salt water, and biota. They are the basic elements that compose what the Russian geochemist Vernadsky (1929) called the biosphere: the zone in which life as we know it can exist and which life, in turn, can affect and transform.

To speak of nature or the biophysical world as the environment of human existence is at least to some degree to separate the two realms. Such a separation is often challenged and resisted. The question has long been debated whether humans are fundamentally a part of, or apart from, the natural world. The grounds on which they are asserted to, or not to, belong to nature are of the most varied sorts. Social arrangements can be described as given by nature in order to justify and 'naturalize' them and rebut criticism of their injustices. Nature, as in the concept of natural rights, can at the same time be used as a standard by which existing social arrangements are measured and found wanting. During the nineteenth century, Darwinian and Lamarckian theories of human origins tended to dissolve certain boundaries between humanity and nature by emphasizing the importance of the organism's evolution from its surroundings and its continuous shaping by the demands that these impose. They were and have continued to be adopted and employed for widely varying ends by both conservatives and radicals.

At the same time, the American geographer Marsh (1864), a contemporary of Darwin's, tended rather to emphasize the distinctions between the human and the natural. He insisted forcefully, and with a wealth of evidence, that human actions modifying the earth were distinct in kind rather than merely degree from natural processes of change. They were, Marsh pointed out, far more rapid and abrupt in occurrence and far more radical and destabilizing in their character. But the very scale and scope of human impact, which have both grown enormously even since Marsh's time, tend in another way to blur the boundaries between society and environment: by making the latter less and less of a distinct entity. Undoubtedly human impact has altered and even transformed many aspects of the earth. Human actions, both deliberately and incidentally, have reshaped them into what is sometimes called a second nature that can no longer be regarded as something quite separate from society. Many people now spend their lives in climates that, at the scale of the built environment, are controlled closely and deliberately; that locally and regionally have in many cases been altered by human actions though rarely by conscious human intent; and that globally may or may not already bear the marks of greenhouse gas emissions from fossil-fuel burning and land-cover change. Little of the environment can be considered wholly 'natural' if that term means pristine and unaltered.

Much of the environment has not been transformed, however, and all of it, however much altered, remains distinct from the human realm in another sense. Human beings possess certain characteristics of reasoning and reflexivity different in kind or at least in very significant degree from the rest of the known world. No other species or biophysical agent has equivalent attributes allowing it to carry out equally elaborate projects affecting the environment, nor have the actions of any changed the globe nearly so drastically in so short a time as have those of the one species possessing them. That humanity and environment are in some important ways separate is implied by the very existence of the social and behavioral sciences (and the humanities) as core fields of study distinct from the natural sciences.

Yet their separation very often is challenged from the margins by two broad programs for reducing one to a branch of the other and asserting their unity in terms of both substance and appropriate methods of study. Extreme forms of biological determinism propose to account for human and social action as the product of heredity and evolutionary imperatives, environmental determinism as the product of the shaping forces of climate and other elements of the human habitat. Both seek to annex the social realm to the natural. Annexation of another kind is envisioned by the more radical approaches found in the social studies of science, which propose to treat the knowledge generated by the natural sciences not as a means of explaining social action and processes but as itself one of their products. Science is seen not as a privileged source of fact but as a social activity like any other and governed by the same rules and to be analyzed in the same terms. These perspectives offer valuable correctives to an unthinking separation between the human and the natural. Weaker forms of them—evolutionary psychology, for example, or the study of the social correlates of environmental attitudes and activism—have been fertile sources of challenging new findings and explanations. Thus far, at any rate, some basic distinction between humanity and the biophysical world has continued to be accepted as unavoidable in practice.

2. *Human–Environment Relations*

But even if environment and society are distinct realms, not so simply and unproblematically related as reductionists on either side would argue, they are closely interdependent and interacting ones. If the autonomy of the social from the natural sciences (or of the natural from the social) is justifiable, what is less so is the thoroughness with which the social sciences, by

and large, have cleansed themselves of contact with the biophysical world. They have, in general, ignored the fact that all human activities, though not reducible to the biophysical plane alone, nonetheless have a biophysical side to them. All human actions require energy and matter from the environment and release both back into it. All have certain prerequisites in the form of the basic necessities of human existence—air, food, water, a tolerable temperature range, and so on—that the environment must provide; and these only begin the list of significant relations between society and its biophysical milieu.

Broadly speaking, the environment can represent a source of services (a term now coming into use to replace the narrower one of 'resources'). Such services include background benefits from the environment that are enjoyed even though they are not consciously and deliberately exploited as natural resources. Among these are the key life-support systems of the biosphere. Other environmental services range from the absorption of wastes to the provision of spiritual and esthetic satisfaction and natural amenities such as scenery and agreeable weather. The environment can equally serve as a source of disservices, ranging from the hazards of violent natural events such as storms, earthquakes, and epidemic and endemic diseases to disamenities and constraints on human activity. To the extent that the environment acts as a determinant of human qualities, its effects can again be classified into services (the formation of valuable qualities) and disservices.

But there is no such thing as a service or a disservice without a recipient thereof. If some aspects of the environment are essential to the existence of any kind of human life as we know it—though not necessarily as it might exist in the future—whether and how any other plays either role depends crucially on the variables of what it is that people do and wish to do. Environmental features that help them do what they wish to do represent services, those that interfere represent disservices. Actors differentiated by location, occupation, class, status, gender, and any number of other variables will benefit or suffer differentially. The significance of the environment, then, is never determined solely by the characteristics of the environment; it is determined also by the ways of life that people lead within it.

As a result, a change on either side may be sufficient to change the relationship. It is by now clear that the environment on its own cannot be presumed to be stable or equilibrium seeking. Natural processes of change and fluctuation on many different scales of time and space are ubiquitous. Over time, they have increasingly been supplemented and, in some cases, overshadowed by human-induced changes, both those made consciously and those resulting unintentionally from actions taken. These may enhance or lessen the biophysical services and disservices received by different groups and individuals. Possible environmental instability must be considered in the study of past environment–society interactions and in planning for future ones.

But environmental change, while it can be sufficient to alter environment–society relations, is not necessary for an alteration to take place. Equally sufficient can be shifts occurring on the social side. Changes in technology may mean that some aspects of the environment will cease to provide services or disservices, while others will begin to do so. Falling water was a vital source of industrial energy through direct-drive waterpower in the era before widespread fossil fuel use, and it became one again with the development of the means of hydroelectric power production. Uranium became an important resource only through the development of nuclear power generation. Some technologies are developed to alleviate existing problems in environment-society relations: those of weather monitoring and forecasting, for example. Others incidentally create new problems: the new role of clouds, fog, and downdrafts as hazards created by the development of aviation. Technological changes themselves are entangled closely with changes in social organization, which themselves may also alter society–environment relations. Such variables as the structure of ownership and exchange, household and gender relations, and political system, for example, all play an important mediating role between natural events and their possible human consequences. Famine, for example, may follow drought that drastically reduces the crop harvest, but a severe drought is neither a necessary nor a sufficient condition for famine. One of the few well-established regularities in the social sciences, documented by the economist Amartya Sen, is that no modern famine has occurred in a country, however poor, with well-functioning democratic institutions. They act, it would seem, to prompt effective government interventions preventing disaster. A certain environment, environmental feature, or environmental event, even if unchanged, will thus bear different relations to society at different times. Changes in the environment itself, of course, whether themselves natural or human-induced, can also alter alter the role of the environment in the lives of its inhabitants.

But an account of human–environment relations in these largely utilitarian terms remains only a partial one. It excludes, most notably, questions about environmental ethics or the proper relations between humanity and environment. The social sciences have found it difficult enough to address issues of values where only human actions are involved. Their particular focus has almost disabled them from assimilating or even recognizing arguments about the rights of other species and other features of the environment: the American biologist Leopold's (1949) 'land ethic,' to take one example. Their usual tendency is to treat such beliefs—and understanding of the environment more generally—as expressions of social position

rather than to grapple with their possible validity. But the rise of Green parties and ecocentric groups and movements and of expressed support for notions of the rights of nature pose external challenges to the use of purely human-centered criteria, ones that have grown in urgency as human power to alter the environment, with likely effects including the widespread extinction of other species, has also grown.

3. Human–Environment Relations in the Social Sciences

These observations suggest the wide range that human–environment interactions cover. The actual range of interactions recognized and studied by the mainstream social sciences at any one time has always been much narrower. The most important early form in which environment impinged on the fledgling social sciences was that of environmental determinism. This doctrine offered one promising and plausible way of solving the formative puzzle of the social sciences: explaining why human behavior and institutions varied so widely from one place and people to another. The French political and social theorist Charles-Louis de Secondat, Baron de Montesquieu, in the eighteenth century treated climate and soil in *L'esprit des lois* as important factors in social life. He asserted, for example, that the people of cold or barren or mountainous areas tended to be vigorous and freedom-loving and those of warm, fertile plains to be more indolent and passive. Political and social institutions suited to one environment, he maintained, would be inappropriate to another. Many thinkers of the French and the Scottish Enlightenment echoed Montesquieu in seeking to explain differences between societies as at least in part growing out of differences in climate or other aspects of the physical environment.

But the concept of environmental influences soon changed from a useful part of the social theorist's conceptual baggage into an irksome burden. The shift proceeded in step with the organization of the social sciences into separate disciplines of sociology, anthropology, history, geography, economics, law, and so on. A few nineteenth-century historians—Henry T. Buckle in England, John W. Draper in the US, and V. O. Klyuchevsky in Russia among them—continued to assert the fundamental importance of environment in human affairs, but the trend of studies ran strongly in the opposite direction. Far more typical was the stance taken by Emile Durkheim in sociology. In his chief empirical work, he rejected meteorological and climatic explanations of patterns of suicide in favor of ones involving the character of social organization. At a more general level, he proposed that social facts be explained solely in terms of social causes. The approach excluded not only psychological or individual-level explanations in sociology, but also biological and physical-environmental ones. Not all of sociology was Durkheimian, but the rejection of environment as a significant factor in social life was not one of its claims to be contested widely within the discipline. It was paralleled by the work of Franz Boas and others in anthropology who likewise insisted on the liberation of social inquiry from environmentalist as well as biological theories of cultural differentiation. Environmental determinism lasted longest in geography, but even there it faded under strong criticism from within and without the discipline.

Classical political economy from the outset recognized environmental conditions as determinants of scarcity and cost. It attributed the key phenomenon of land rent to differences in the inherent quality of the soil. Malthusian population theory presented the disparity between limited natural resources (the ability of the land to produce food) and potentially unlimited human wants as a basic condition of social life. The classical labor theory of value, however, tended to promote human activity over nature as the key source of wealth. Environment was demoted to a secondary role in another and more decisive way in neoclassical economics. Natural services were absorbed into the wider stock of capital and deprived of any distinctive qualities vis-à-vis human-produced goods. Natural capital and services were presumed to be perfectly replaceable by non-natural and the possibility of absolute scarcity of vital elements of nature neglected. Market mechanisms would ensure that nonrenewable resources would be depleted at the socially optimal rate and would also ensure that substitutes for them would be devised as necessary. These assumptions were questioned only by fringe schools of thought that had the standing of heresies vis-à-vis the orthodox consensus: theorists who proposed energy rather than monetary measures of wealth and efficiency, conservationists who warned of the exhaustion by waste of key resources not readily replaced by others.

In orthodox Marxist thought, not only the labor theory of value, but also the violent rejection of Malthusian notions of absolute scarcity as ideological fictions and the emphasis on class conflict and on a universal sequence of stages of evolution left environment little room to play a serious role. The chief exception was the much debated concept of distinctive 'hydraulic societies' expounded in the mid-twentieth century by Wittfogel (1957). Developing the earlier notion of the 'Asiatic mode of production,' he proposed that the formation of stable centralized despotisms was favored by the demands of large irrigation and water-control systems in dry climates—the characteristics of the environment thus playing an important role in social outcomes. But the idea only represented such a heresy within Marxism as energy economics did within the neoclassical discipline.

Elsewhere in the social sciences, determinism was for a time replaced by the opposed doctrine of possibilism, prefigured in the work of late nineteenth and early twentieth century French regional geo-

graphers led by Paul Vidal de la Blache and influentially expounded in the 1920s by the historian Febvre (1922). Possibilism pointed to the fact that different societies maintained quite different relations with similar environments as proof that the environment itself never determined any particular way of life. It offered, rather, a wide range of possibilities for habitation and use, some of which any particular society would be inclined to exploit and many of which it would not. In the postwar years, a possibilist approach was one of the many novelties informing the work of a school of historical research centered on the journal *Annales: Economies, Sociétés, Civilisations*. Close study of nature–society relations remained an important endeavor not only in *Annales* history, as it came to be known, but also in the subfield of anthropology and geography known as cultural or human ecology. It too was possibilistic in showing how society–environment relations, especially those of subsistence in nonindustrial settings, could be understood only by study of the characteristics of culture and habitat both. The very strength of environmental research in these precincts, however, only tended to reinforce the notion that environmental differences mattered and continue to matter chiefly for less advanced societies employing primitive technology and engaged in primary production.

That was not the only belief that tended to discourage wider attention to the environment. Another was the fear that taking the environment seriously led easily to environmental determinism and in turn to political implications of a necessarily reactionary sort. The fear, though excessive, gained plausibility both from some of environmentalism's earlier associations and from an easy analogy with biological determinism. According the natural world any significant role in human affairs seemed to some, given the unevenness of natural patterns across the earth's surface, a step toward legitimating inequalities of power and position. Environmental determinism was most reliably expelled by expelling environment altogether.

It is understandable and by no means unfortunate that there should have been a strong reaction against a doctrine that had shown itself so intellectually barren and empirically untenable as environmental determinism. It is more regrettable that environmental relations of nondeterministic sorts should have come under the same ban. But their neglect was also a predictable result of the way in which academic research and training increasingly came to be organized. Environmental studies of any sort suffered from the deepening institutional and intellectual division of territory between the social and the natural sciences. Part of the price at which the hard-won autonomy of the social sciences came was the disenfranchisement or marginalization of subjects located in their new boundary zones. Part of the price of the increasing specialization of work within them was the increasing difficulty of acquiring an adequate training in both the natural and the social elements of any research problem. The way in which the social sciences tended to be defined in opposition to the natural split even geography into human and physical and anthropology into cultural/social and physical branches. It was thereby made most difficult for any topic relating the two realms to be studied with any degree of sophistication.

The 1968 *Encyclopedia of the Social Sciences* appeared at a particular low point in social science attention to the biophysical environment, and the entries reflected it. The entry on 'Environmentalism' was a short account, largely historical, of environmental determinism, and the entry on 'Pollution' discussed not chemical contamination of air and water but ritual defilement in the purely cultural sense. Also included were a short piece on 'environment,' chiefly an exercise in definition (albeit a very useful one), and essays on 'conservation,' cultural ecology, and land and water resources. Biographical sketches of a few important figures such as Marsh and Malthus rounded out its treatment of the theme. Taken together, the entries conveyed no sense of the biophysical environment as an important topic for the social sciences.

But events were already taking place outside of the academic world that would again force environmental questions, at least in some form, on their attention. The environmental movement beginning in the 1960s and 1970s expressed a widespread public sense of concern over the impact of human activities on the earth. The list of new fears included ones as varied as forest damage from acid rain, desertification in arid and semi-arid lands, water supply shortages and water pollution, air pollution at various scales, the loss of biodiversity, and climate change at local, regional, and global scales. Activists called for intervention to protect the environment and combat the social forces—be they overpopulation, overconsumption, capitalist greed, socialist mismanagement, or any other—that they saw as responsible for the degradation and fraying of the biosphere.

Closely following the rise of public concern, attention to human-environment relations has reemerged in the social sciences, but only in the narrow form of study of the human dimensions of human-induced environmental change. For some social scientists, it has been a characteristic and congenial task to examine environmental concern itself as a social movement having social causes, without necessary reference to any truth that the claims that it makes might possess. Such can be a fruitful path of analysis, one that asks how environmental problems are selected by certain groups and actors for attention among many others that are ignored. It is not difficult to show that public concern about many environmental issues substantially outruns understanding of the scientific basis for concern. Thus, the roots of concern may lie less in the objective situation (even if the scientific account of it is accepted as objective) than in other

areas that many researchers, particularly in Western Europe, have sought to identify. A second line of work has been much more visible and on the whole more generously supported by governments and other funding agencies. This form of attention to human–environment relations has largely accepted the idea that a global environmental crisis genuinely exists or threatens and has sought to shed light on its human dimensions: its causes, its consequences, and the ways by which it might be managed or mitigated. This approach, perhaps the dominant one in the US, evidently stems less from the traditions of social science research than from the pressures of external concerns.

Unfortunately, such research has still more often than not been conducted with the tools and within the assumptions and frameworks of the existing social science disciplines and subfields whether appropriate to environment or not. Orthodox social theory developed with little or no reference to environment is, thus, brought to environmental questions and used to recommend policy as if it required no modification to meet them. Yet its inadequacy to deal with many questions without being modified is plain. The long-standing neglect of biophysical issues of scale, scarcity, and irreversibility in economics has been challenged by the emergence of ecological economics, but to date with little apparent impact on the larger field. Sociological, anthropological, and political paradigms developed in other contexts tend to be applied indiscriminately to environmental questions without careful consideration of how applicable they are. The anthropocentrism of the social sciences as a whole makes it difficult if not impossible for them to deal with notions of the rights of nature and inherent value seriously. Most recently, many adherents of postmodernism in the social sciences have had to struggle with the labor of trying to reconcile their reluctance to acknowledge anything as objectively true with their belief that a genuine global environmental crisis exists and must be addressed.

The long separation between the study of society and the study of the earth has hindered the growth of theory indigenous to and appropriate to nature–society issues that does not oversimplify either side of the relation. The natural sciences have been no less reluctant than the social sciences to modify their favored theories and approaches in order to address the full range of links between society and environment. Where they have not ignored the human role, many have tended to fall back on a narrow and reductionist set of social theories to make sense of it. A characteristic example is the currency in natural-science discussions of environmental problems of the concept of 'carrying capacity,' the supposed maximum population that an area can support. It is a concept largely rejected as too mechanistic by social scientists who emphasize the culture-bound nature of the assumptions that must be made to define capacity in particular cases. The view that human overpopulation and overconsumption are the key driving forces of environmental degradation is more widely held in the natural than in the social sciences, whose practitioners, for their part, are sometimes prone to suppose that the capacity of the environment to support life and activities is infinitely flexible or that the only serious limits are social. Many natural scientists also show an affinity for technical solutions to environmental problems—such as carbon sequestration, albedo modification, or a shift to nuclear power as a response to the threats of global climatic change—without adequate consideration of the crucial problems tied up with their social feasibility or public acceptance.

Human–environment relations may now be far more urgently important a research topic than ever, given the possible dimensions of rapid global environmental change and the measures that coping with it might demand. Yet a long period of neglect means that the knowledge base in this area is weak. Little is known about what environmental change might mean for society because little is known about what environment means for society. The present atmosphere of crisis does offer some hope that at last a detailed understanding of human-environment relations will begin to be constructed. Yet an adequate assessment of many environmental issues cannot be achieved from a background only in the natural or social sciences, and it remains to be seen whether the organization of education and research will prove capable of providing the necessary conditions to nurture such work.

See also: Cultural Geography; Environment and Development; Environmental Determinism; Environmental Policy; Environmentalism: Philosophical Aspects; Nature, Concepts of: Environmental and Ecological; Environmentalism: Preservation and Conservation; Ecology, Cultural; Ecology, Deep; Ecology, Political; Global Environmental Change: Human Dimensions; Cultural Landscape in Environmental Studies

Bibliography

Arnold D 1996 *The Problem of Nature: Environment, Culture and European Expansion*. Blackwell, Oxford, UK

Bates M 1968 Environment. In: Sills D L (ed.) *The International Encyclopedia of the Social Sciences*.

Bramwell A 1989 *Ecology in the Twentieth Century: A History*. Yale University Press, New Haven, CT

Clark S (ed.) 1999 *The Annales School* (4 Vols.). Routledge, New York

Coates P 1998 *Nature in Western Attitudes since Ancient Times*. University of California Press, Berkeley and Los Angeles, CA

Costanza R (ed.) 1991 *Ecological Economics: The Science and Management of Sustainability*. Columbia University Press, New York

Cotgrove S 1982 *Catastrophe or Cornucopia*. Wiley, Chichester, UK
Daily G C (ed.) 1997 *Nature's Services*. Island Press, Washington, DC
Febvre L 1922 *La terre et l'evolution humaine*. Alcan, Paris
Forde C D 1934 *Habitat, Economy and Society*. Methuen, London
Glacken C J 1967 *Traces on the Rhodian Shore: Nature and Culture in Western Thought to the End of the Eighteenth Century*. University of California Press, Berkeley and Los Angeles, CA
Guha R, Martinez-Alier J (eds.) 1997 *Varieties of Environmentalism: Essays North and South*. Earthscan, London
Leopold A 1949 *A Sand County Almanac*. Oxford University Press, New York
Marsh G P (1864) 1965 *Man and Nature; or, Physical Geography as Modified by Human Action*. Harvard University Press, Cambridge, MA
Murphy R 1997 *Sociology and Nature: Social Action in Context*. Westview Press, Boulder, CO
Nash R 1989 *The Rights of Nature: A History of Environmental Ethics*. University of Wisconsin Press, Madison, WI
Rayner S, Malone E (eds.) 1998 *Human Choice and Climate Change* (4 Vols.). Battelle Institute Press, Columbus, OH
Redclift M, Benton T (eds.) 1994 *Social Theory and the Global Environment*. Routledge, London
Thomas F 1925 *The Environmental Basis of Society: A Study in the History of Sociological Thought*. Century, New York
Thomas W L Jr (ed.) 1956 *Man's Role in Changing the Face of the Earth*. University of Chicago Press, Chicago
Turner II B L et al. (eds.) 1990 *The Earth as Transformed by Human Action*. Cambridge University Press, New York
Vernadsky V I 1929 *La biosphère*. Alcan, Paris (English trans. *The Biosphere* 1998)
Wittfogel K 1957 *Oriental Despotism*. Yale University Press, New Haven, CT
Worster D 1994 *Nature's Economy: A History of Ecological Ideas*, 2nd edn. Cambridge University Press, New York

W. B. Meyer

Humanistic Geography

Humanistic geography studies people's relationships to nature, space and time, their behavior, feelings and perspectives on environment. Sharing enquiry with other branches of the discipline, its domains of interest overlap strongly with the humanities and social sciences. Its scholarly value and main contribution to society consists in its ability to raise people's levels of awareness about taken-for-granted ways of life and thought, inviting critical reflection on humanity's relationships to the earth.

'Diversity is at the heart of geography—from the varieties in continents and climates to the interrelationships between natural resources and how people live. Diversity is also central to human rights.' Thus, Mary Robinson, UN High Commissioner for Human Rights, addressed delegates at the 29th International Geographical Congress in Seoul, August 14–18, 2000, upon receipt of the *Earth and Humanity Medal 2000* from the International Geographical Union. She was honored for not only her championing of civil liberties and economic opportunity, but also for proclaiming access to intact and healthy environments as basic human rights. A few weeks later, scholars from all over the world assembled in Rome to explore the potential role of universities in the quest for a 'new humanism.' Renaissance humanism, one speaker recalled, 'set up a new idea of truth as a dynamic statement, not previously defined or constructed, but something to be discovered and then applied … the joint venture of new scientific discovery with the humanistic approach to mankind's problems enabled the university to nourish innovation and offer a rigorous critique of institutions and social relations' (Bricall 2000). Such 'airing' of university life, he noted, was sadly lacking today. 'Humanistic geography' may have been a twentieth-century invention, but many of its central themes at the dawn of the twenty-first century resonate with centuries-old melodies.

1. Geography and Humanism

Humanists down the centuries have explored the nature of humanity, its passions and powers, while geographers have studied the earth where humans, among many other life forms, make a terrestrial home. For each facet of humanness—rationality or irrationality, faith, emotion, artistic genius, or political prowess—there is a geography: for each geographical interpretation of the inhabited earth there are implicit assumptions about the nature of humanness. To slogans such as Protagoras' 'man is the measure of all things' and Alexander Pope's 'the proper study of mankind is man,' geographers would note that *Homo sapiens* is a terrestrial species, with projects and plans played out in particular environments, even if diverse cultures have created and been inspired by myths and symbols. *Humanus*, literally, means 'earth dweller.'

The term humanism evokes at least four distinct connotations. First, it implies questions regarding the nature of humanness (*humanitas*) with concerns about individuality and sociality, freedom and responsibility, rationality and hedonism, conservatism and creativity. Second, it refers to humanist modes of knowing which oppose scientific reductionism, aiming to elucidate rather than to explain, emphasizing the subjectivity of consciousness and the need for intersubjective dialog. Third, it is expressed in the humanities—fields of learning in the arts, classical literature, and the cultivation of civic virtue. Fourth, it is evident in an enduring concern about the human condition, e.g. in humanitarianism, which has sought to encourage social responsibility and liberal politics. Historically, from whatsoever ideological stance it has emerged,

humanism has included a call for critical reflection on conventional ways of interpreting reality and a plea for integrity and authenticity in everyday life (von Wright 1978, Buttimer 1990).

Humanistic geography also reveals these four lines of concern. What is subsumed under this rubric varies from one country or language tradition to another, the terms 'social,' 'radical,' 'critical' and 'humanistic' often being seen as virtually interchangeable (Buttimer 1974, 1976, Tuan 1976, Relph 1981, Rowntree 1988, Pellegrini 1992, Claval 1999). For some the main goal is still to restore human subjectivity to a field where scientific objectivism has been so dominant, some emphasizing human attitudes and values, others cultural patrimony; some focusing on the aesthetics of landscape and architecture, others on the emotional significance of place in human identity (Ley and Samuels 1978, Buttimer and Seamon 1980). Several also advocate humanitarian concern and engagement in the resolution of social or environmental problems. Virtually all would agree with the formulation of Yi-fu Tuan that 'the humanist's competence lies in interpreting human experience in its ambiguity, ambivalence, and complexity' (Tuan 1976, p. 275). For over three decades humanistic geographers have explored the aesthetics of landscape, issues of memory and myth, moralities and imaginations, and the multifaceted textures of place (Adams et al. 2001). This 'humanist turn' has thus evoked sharper sensitivity to the inextricable connections between thought and context, experience and expertise, and the adequacy of geography to elucidate contemporary reality.

Some have looked askance on these developments as a retreat from problems or as simply a critique of the status quo (Entrikin 1976, Relph 1981). Anti-humanist strains have also been heard from doctrinaire structuralists, and staunch advocates of positivist spatial analysis. In the 1980s many, disillusioned about contradictions in Western legacies, condemned humanism as the arch culprit, a source of *hybris* and vaunting ambition, a myth overdue for dismissal. Postmodernist scenarios of texts reproducing texts in a labyrinth of self-reflecting mirrors precluded assumptions about human intentionality or meaning in lived reality. By the late 1980s there was a growing awareness of the extent to which the entire intellectual heritage of the West had shaped, and been shaped by, the peculiar social history of Western humanity. By century's end there was a notably increased interest in critical emancipatory elements which had previously remained marginal within the profession. These diverse currents pointed to weaknesses in the inherited lore and have opened up new worlds of exploration. And linked to this, there has also been a re-affirmation of geography as fundamental education for mature citizenship, for cross-cultural understanding and for environmentally sustainable ways of life. The 'new humanism' promises to be far more 'geographical' than its predecessors.

An enduring concern of humanistic geography has been the increased level of specialization among knowledge fields. While the narrowing of analytical focus has yielded sharper insight into particular processes, the overall fragmentation of research horizons has seriously diminished the geographer's capacity to deliver comprehensive perspectives on humanity and environment. In a so-called *Magna Carta* for the 'Humanities,' William Caxton (1422–91) insisted that attention be confined to mankind as an object of study in itself, not as part of 'nature,' which was the concern of biology, nor as an object of divine grace, which was the concern of theology. Hence emerged a clear separation of scientific and humanist modes of knowing. The former was to provide analytical guidelines for the analysis and representation of research results, while the latter was to invite discovery and understanding, processes which might not only vary from one individual to another, but also imply elements of emotion, aesthetics, and moral judgment which, strictly speaking, were to be eliminated from the scientist's agenda. At the dawn of the twenty-first century, there are resounding cries for liberation from such institutional dogmatism and inherited meta-discourses which obfuscate and distract attention from the far more tyrannizing meta-discourses of commercially-driven globalization.

An essential quality of humanist discourses in Western scholarly traditions, it could be argued, is a plea for freedom: freedom from bonds of ignorance, fear, oppression or oblivion, freedom to create and to explore new horizons in thought and life (Buttimer 1990). The emergence of a 'humanist' turn within geography during the latter half of the twentieth century might be described as a liberation cry for aspects of thought and experience which had been ignored or suppressed within the practices of previous generations. At the heart of humanist concern at the dawn of a new millennium is the patent failure of well-intentioned scientifically-based attempts to alleviate poverty, inequalities and injustices worldwide. Ironic indeed is the apparent paradox that, as volumes of environmental research have grown, so too has the magnitude of environmental problems. Questions now arise as to whether our taken-for-granted sectorally specialized kinds of science are adequate for elucidating these problems, or whether, in fact, taken-for-granted application of scientific results is part of the problem (Escobar 1997).

The twentieth century ended with a litany of 'post-' and 'anti-'movements in the scholarly world—emphasis placed on elements from which many sought to be free; horizons toward which a humanist spirit would aspire remained unclear. And yet there emerged a resounding chorus proclaiming plurality and diversity. Far from one orthodox brand of geographical knowledge, there are varieties of geographical knowledges—those based on everyday lived experiences, those of women, children, minorities, immigrants and

emigrants, professionals and clients, rich and poor. In fact, many of the major challenges facing humanity and environment today could be redefined in terms of competing, contested or conflicting geographical knowledges.

2. Humanistic Geography

Geography today is recognized as not only a professional field of expertise, an academic discipline, but also a reality of everyday lived experience. Humanistic geographers have consistently promoted attention to everyday geographical experience, not only for individual human persons, but also for the diverse cultural worlds whose home territories are increasingly invaded by globally organized technological and commercial systems. They advocate coresponsible engagement in matters of territorial and social discretion in humanity's modes of dwelling. Illustrations of contemporary themes for research and reflection in humanistic geography include (a) Dwelling, (b) Nature and Culture, (c) Mindscape and Landscape, and (d) Text, Image and Construction of Regional Knowledges.

2.1 Dwelling: Essential Quality of Humanness

Geography is the stuff of everyday experience for all forms of terrestrial life. Mountain and plain, river and lake, woodland and wildlife may be 'explainable' in the categories of natural science, but in reality each cultural group understands nature, space, and time through its own special filters. To survive on planet earth, all creatures develop a geographical sense, a sense of place, space, time, and movement. How to negotiate our diverse geographies is surely one of the enduring challenges of existence, from the politics of empire to the arrangement of one's kitchen, office, and garden. A geographer tells stories of dwelling, and the interplay of mindscapes and landscapes, in terrestrial environments.

The progressive 'humanization' of the earth over the past half-million years, offers fascinating ground for study and reflection (Glacken 1967). Humanity has assumed a dominant role in shaping landscapes, harvesting resources, tipping the natural checks and balances among other life forms, in culturally and technologically diverse ways of life. Each cultural group has devised its own variety of ways in which the bio-physical environment was understood and how humans should relate to it. Some, as in the Judaeo-Christian and Hellenic traditions, have placed humanity above all other life forms and have regarded intellect as the highest manifestation of human spirit. In Hindu and Buddhist traditions, humanity has been regarded as an integral part of cosmic order, recycling, like other life forms, in a continuous process of reincarnation. Symbolic transformations are quintessentially human projects: the representation of reality in art and artifact, monument and metaphor, science, technology, and national flags. The human body itself has been a universal catalyst for symbol making in many civilizations, most common of which has been the organic analogy of the human person as a whole (body, soul, personality) as a microcosm of the universe.

Human modes of dwelling, with central themes of place and space, home and horizon, memory and identity, remain a primary focus of attention for humanistic geographers (Buttimer and Seamon 1980, Seamon and Mugerauer 1985). Emphasis is placed on the aesthetics and appropriateness of human settlements and the dramatic differences in built environments throughout the world. Humanistic geographers are skeptical of universal generalizations such as those which glorify or demonize humanity. Latecomers anyhow to earth history, some have argued, humans may now be en route toward self-destruction. While acknowledging some grounds for such gloomy prognoses, humanistic geographers continue to probe diverse landscapes, documenting evidence of human ingenuity in adapting lifeways to environment throughout history, and reiterating the perennial challenge of environment and ecology for global humanity today.

2.2 Nature and Culture

Throughout human history nature has constituted the source of inspiration and foundation for cosmology and human identity. Australian Aboriginals revere rocks as dreaming centers and song-lines evoke a sense of on-going creation. The *Upanishads* celebrate the eternal Asvatta (fig tree) and banyan tree, while North American Sioux build their entire cosmology around the sacred cottonwood tree. For East Asian societies, life was seen to emerge from the sacred lotus, and devout Hindus bathe in the sacred waters of the Ganges. Artefactual and cartographic evidence suggests that most cultures have imaged their own place as the center of the world, with horizons varying in shape and extent depending on their livelihoods and travel. The cargo cults of Polynesian islanders re-enact the stories of colonial occupation, Obijwe and Cree have detailed mental maps of trapping lines across the North American Plains—territory viewed by Euro-settlers as a vast undifferentiated arena for conquest by railway, lumber company, or settler (Brody 1981).

Humans everywhere have looked to nature for insight into the secrets of creation, birth and death, meaning and purpose in everyday life. Human inventions such as the domestication of animals and plants, technological innovations in agriculture, transport, and commerce have been construed and defended in

terms of cosmology and value system. The enduring salience of symbolism in shaping contemporary landscapes constitutes a central platform for 'new cultural geography' (Cosgrove and Martins 2000).

For the 350 million so-called indigenous peoples of the earth today, nature and human identity are inseparable. Multinational companies, often in league with national governments, find themselves confronted with a resounding *cri du coeur* from autochthonous populations eager to protect the integrity of their home environments and maintain discretion over minerals, forest, fish, and wildlife in their ancestral lands. The dramatic clash of human interests which characterizes the use of earth resources has at last found its way to public consciousness. The languages within which such clashes are overtly negotiated are characteristically those of technical expertise in the service of governments, corporations or transnational lobbies; the voices of *dwelling*, by people in place, are often ignored. Humanist geographers seek alternative bases on which human discourse about earth resources might be negotiated. Acutely aware of the extent to which scientific and popular notions of 'nature' have been socially constructed and applied in sometimes contradictory ways, they provide catalysts for critical reflection on taken-for-granted policy and programmes.

2.3 Mindscape and Landscape

The visible forms of landscape are the texts that geographers read and seek to interpret. To understand the humanized landscape one needs to understand cultural diversity among societies: not only their economies and traditional skills, but also the sense of sacred and profane in their perceptions of nature, space, and time. Successive groups have migrated over time from one area to another bringing their talents, images, hopes and prejudices with them. Characteristically, they try to recreate the new landscape in the image of the past, usually an idealized version of it. Contemporary landscapes have all been formed through the sedimentation of many layers of human occupance.

Among the strongest advocates of the 'humanistic' turn in geography were scholars who already acknowledged the hermeneutic value of landscape interpretation (Meinig 1979). 'Our human landscape,' Pierce Lewis remarked, 'is our unwitting autobiography, and all our cultural warts and blemishes, our ordinary day-to-day qualities are there for anybody who knows how to look for them' (Lewis 1985, p. 13). Landscape interpretation, adopting some of the models of literary criticism, 'read' landscape texts and signs as products in their own right, while also inviting enquiry into culturally varying modes of symbolic transformation in history. Debates over agency versus structure in the social construction of landscape features became the *sine qua non* preamble to many research ventures of the 1980s and 1990s, as reflected in new scholarly journals such as *Ecumene* and *Géographie et Cultures*.

To communicate insight on the interactions of mindscape and landscape naturally poses critical issues of language, a challenge which geographers share with colleagues in literature, philosophy and art (Pocock 1981). It has surely been on questions of language and power, semiotics and symbolism, that humanistic geographers found challenging moments of interaction with structuralists. Slogans of 'death to the human subject,' short-lived but militant, scoffed at notions of potential human creativity to transcend structural imperatives. The deconstruction of inherited meanings, as suggested in the work of Derrida, could reveal the multiple and conflicting readings that could be made on particular texts. In the 1980's, post-structuralist discourses emphasized the multiplicity of potential interpretations and the undecidability of scientific propositions. In sharp contrast with the previous generation's deeply held convictions about epistemological certainty and images of the world as a mechanical system, many voices of the late 1980s were casting doubt on all scientific claims to objectivity.

2.4 Text, Image and Construction of Regional Knowledges

Humanistic geography is concerned not only with ways in which humans, in their everyday lifeworlds, have interpreted their surroundings. It is also concerned with images of the earth projected in academic (scientific) products (Dematteis 1985, Buttimer et al. 1999). As with the earth's landscapes, the mindscapes of professional geographers have reflected the cultures and world-views of their sponsors and audiences. Geography's status as an academic discipline has mirrored the fluctuating fortunes of nations and empires, flowering at moments of societal self-confidence, altruism, or expansionary challenge; relaxing in routine operational tasks of housekeeping and inventory, text-book writing, and functionally specialized research in times of stability or depression. And as active participants in a global community of scholars, geographers have also negotiated their practices within the reigning scientific paradigms of particular periods.

Once national ministries of education established geography in university departments, the discipline's agenda became harmonized with that of its major sponsor, the state. To this day, with some notable exceptions, the practice of geography has borne a distinctively national character. From its beginnings as an academic discipline, geography has been valued as training ground for the exploration and conquest of space and resources; for the imposition of order

deemed rational by the conqueror; for the delivery of information on areas, distances, flora and fauna, peoples and cultures, in language categories and narrative frames which could be understandable and useable 'back home.' It is indeed ironic that while most places are subject today to influences which transcend territorial boundaries, the practice of geography is so tightly shaped by national and institutional contexts.

Humanistic geography is interested primarily in examining the geographical knowledges which have been promulgated in various settings. Starting from the accounts of geographers themselves, it examines the story of ideas in practice, the interplay of creativity and context in the practices of various periods, seeking to evaluate the implications of such practices for everyday life and landscape. Such approaches have not been popular, as Yi-fu Tuan warned in the mid-1970s: 'self-knowledge, the prime reward of the humanistic enterprise, has always been suspect in Western culture' (Tuan 1976, p. 276). While little evidence of its popularity in other cultures is available in geographical literature, it is certainly laudable that Yi-fu Tuan, indubitably the most influential writer of humanistic geography, has now shared his own autobiographical reflections on a fascinating, and possibly unique career journey (Tuan 1999). This account should prod readers to think critically about their own career journeys and taken-for-granted perspectives on life and world.

3. *Conclusions*

Each discovery in humanity's geographical understanding of the world has heralded changes in its self-images, hopes, and fears. The dawn of a new millennium reveals radically revised perspectives on mindscape and landscape. Geographers have dared to look inwards, exploring cultural differences in environmental perceptions and experience, while colleagues in the humanities dare to look outwards, freely exploring themes such as landscape, identity and place, space and time. Today's world map of power and politics has taken on radically new dimensions due to transformations in technology and trade. The explosion of information and the transcontinental circulation of people, commodities, and ideas raise exciting new possibilities for thought and life. The spatial juxtaposition of culturally diverse peoples should become the springboard for new creativity in politics and social life, rather than problems to be solved by social engineering. Humanistic geography has much to offer by way of cross-cultural and historical evidence of creativity in discovering environmentally appropriate solutions. Gaia's human envelope, the anthroposphere, needs to be understood more as a complex drama than simply as battleground of ecological versus economic rationality; more as *oecumene*, potential home for humanity, a species which urgently needs to rediscover the art of dwelling.

Much rhetoric has been aired on the Western world's anthropocentric perspectives, critics usually implying alternatives such as the deification of nature, of the social collective, or the reification of ideologies playing out their purposes with humanity and nature. Perhaps the West, far from overestimating *humanus*, has grossly underestimated it. Recovery of the human subject, and recognition of human agency as an integral part of the lived world, the creativity again apparent on peripheral and previously marginalized regions, all betoken fresh potential for humanistic geography today. As a stance on life, humanism welcomes the challenge of discerning the creative potential of individuals and groups to deal with the surface of the earth in responsible and co-responsible ways (von Wright 1978). Nor is human creativity confined to the intellectual sphere; it involves emotion, aesthetics, memory, faith and will. The humanist turn in geography has delivered inspiration to practitioners of physical, economic, cultural and social geography. Its aims are not well served by investing energy on staking claims for becoming a special branch of the field. Humanism should thus more appropriately be considered as a leaven in the dough rather than as a separate loaf in the *smörgåsbord* of geographic endeavor. The emancipatory *élan* recuperable even from our Western traditions could enable geography itself to perform as leaven in the mass of contemporary science and humanities. The Renaissance of humanism calls for an ecumenical rather than a separatist spirit; it calls for excellence in special fields as well as a concern for the whole picture. From our Western corner, perhaps the greatest challenge is that of bridging worlds through dialog and mutual understanding.

Bibliography

Adams P, Hoelscher S, Till K (eds.) 2001 *Textures of Place: Exploring Humanist Geographies*. University of Minnesota Press, Minneapolis, MN

Bricall J 2000 Some recent trends in the universities. *Giubileo delle Università. Meeting with His Holiness John Paul II with the Academic Community* 9 September, 13–18

Brody H 1981 *Maps and Dreams. Indians and the British Columbia Frontier*. Douglas and McIntyre, Toronto, Canada

Buttimer A 1974 *Values in Geography*. Commission on College Geography Research Report No. 24. Washington, DC

Buttimer A 1976 Grasping the dynamics of lifeworld. *Annals of the Association of American Geographers*. **66**: 277–92

Buttimer A 1990 Geography, humanism and global change. *Annals of the Association of American Geographers* **80**: 1–33

Buttimer A 1993 *Geography and the Human Spirit*. Johns Hopkins University Press, Baltimore, MD

Buttimer A, Seamon D (eds.) 1980 *The Human Experience of Space and Place*. Croom Helm, London

Buttimer A, Brunn S D, Wardenga U (eds.) 1999 *Text and*

Image. Social Construction of Regional Knowledges. Institut für Länderkunde, Leipzig, Germany
Claval P 1999 Qu'apporte l'approche culturelle à la géographie? *Géographie et Cultures* **31**: 5–24
Cosgrove D, Martins L M 2000 Millenial geographics. *Annals of the Association of American Geographers* **90**: 97–103
Dematteis G 1985 *Le metafore della Terra. La geografia umana tra miro e scienza*. Gianfracomo Feltrinelli, Genoa, Italy
Entrikin N 1976 Contemporary humanism in geography. *Annals of the Association of American Geographers* **66**: 615–32
Escobar A 1997 Anthropology and development. *International Social Science Journal* **154**: 97–115
Glacken C J 1967 *Traces on the Rhodean Shore*. University of California Press, Berkeley, CA
Lewis P 1985 Beyond description. *Annals of the Association of American Geographers* **75**: 465–77
Ley D, Samuels M (eds.) 1978 *Humanistic Geography: Prospects and Problems*. Maroufa Press, Chicago
Meinig D 1979 *The Interpretation of Ordinary Landscapes*. Oxford University Press, New York
Pellegrini G C (ed.) 1992 *Humanistic and Behavioural Geography in Italy*. Pacini Editore, Pisa, Italy
Pocock D C D (ed.) 1981 *Humanistic Geography and Literature. Essays on the Experience of Place*. Croom Helm, London
Relph E 1981 *Rational Landscapes and Humanistic Geography*. Croom Helm, London
Rowntree L 1988 Orthodoxy and new directions: Cultural/humanistic geography. *Progress in Human Geography* **12**: 575–86
Seamon D, Mugerauer R (eds.) 1985 *Dwelling, Place and Environment*. Kluwer Academic Publishers, Dordrecht, The Netherlands
Tuan Y 1976 Humanistic geography. *Annals of the Association of American Geographers* **66**: 266–76
Tuan Y 1999 *Who am I? An Autobiography of Emotion, Mind and Spirit*. University of Wisconsin Press, Madison, WI
von Wright J H 1978 *Humanismen som livshållning* [*Humanism as Stance on Life*]. Rabén & Sjögren, Stockholm, Sweden

A. Buttimer

Humboldt, Wilhelm von (1767–1835)

1. Politics

When Wilhelm von Humboldt, born June 22, 1767 at Potsdam, resigned from his office as the Prussian Minister of Constitutional Affairs at the end of 1819, this was a major catastrophe for the political future of Germany. For Humboldt himself it was an act of liberation and a step towards what he regarded as his real vocation, the study of languages, to which he dedicated the rest of his life. Chateaubriand, the French poet and politician, considered this strange scientific activity as the spleen of somebody who had to kill his time as a consequence of the loss of political influence. But it was in fact the achievement of a philosophical and anthropological quest which had started in Humboldt's youth. Studying the languages of the world was not at all something completely extraneous to his political work. The same motivation which would guide his linguistic work had also governed his political activities: it was a quest for the creative center of the human mind and for the achievement of individual felicity and autonomy within the totality of other individualities.

This is exactly what had created his political difficulties. At the end of his political career, Humboldt had elaborated a project for the constitutional transformation of Prussia: a proposal to establish a constitutional monarchy with a gradually developing representational system. The French Revolution, which he had welcomed as a decisive step in the history of humanity, had failed in Humboldt's eyes because of its rationalistic one-sidedness: the transformations of France had been governed too radically by abstract rational principles and had not taken into account historically grown entities. For Humboldt, the creation of the new should always be the result of a 'synthesis,' of a sympathetic union of two entities which, creating a new entity by their union, maintain their identity (the alternative—and lesser—methods of uniting two entities are 'isolation' and 'incorporation,' the former being only the juxtaposition of the two, the latter the brutal destruction of their integrity through their unity). Humboldt proposed a 'synthesis' of the traditional prerogatives of the monarch and the interests of the economically active classes (with a consideration also for the traditional rights of the nobility), a 'marriage' which would have transformed Prussia eventually into something new, a constitutional monarchy—and which would have had a model function for the rest of Germany. Not only for the inner organization of Prussia, but also for the whole of the area called 'Germany,' Humboldt projected a political structure according to the same principle of a 'synthetical' union: a confederation which would have left quite a lot of autonomy to its members but nevertheless have given important political functions to the whole (like warfare, freedom of movement and free trade within the union—very much like the USA).

But, in 1819, the times were not propitious anymore for reforms and for the creation of new political structures as after the defeat of Prussia by Napoleon which was the beginning of so many happy reforms in Prussia, like the reform of the educational system, the renewal of the municipal administration and the emancipation of the Jews. Now Napoleon was defeated and the old classes were striking back and were not willing to make any concessions to the bourgeoisie, to the memory of the 'people's war' or to the experience of freedom and civil rights imported by the Napoleonic armies. Humboldt was the representative of liberal political reforms, and, as such, he had to disappear—in the time of Austria's Metternich, of Hardenberg and a reactionary king in Prussia.

2. *University and* Bildung

But if Humboldt did not succeed in realizing his constitutional dreams, he seems to have had an overwhelming success with the establishment of the University of Berlin some years earlier. Humboldt is best and generally known for this activity, which in fact occupied him for little more than a year (from February 1809 to June 1810) and which was only a part of the reorganization of the whole Prussian educational system. Humboldt indeed made the decisive administrative step for the establishment of the University of Berlin. But what was established was only partially Humboldt's University. He planned the university as a state-independent, economically autonomous institution, with an independence that would guarantee the freedom of research and instruction. Therefore, Humboldt requested that the university should be given a donation of land as the economic basis of that autonomy. Instead, the Berlin University was from the very beginning a state establishment which—it is true—set unprecedented standards by the quality of its new faculty, by new forms of teaching, by the unity of research and teaching for professors as well as for students. The concept of the research university was successfully realized. As a state-dependent institution it would forge a well-trained loyal elite for the King of Prussia rather than free citizens of the realm of the spirit.

Humboldt's proposals for the establishment of the Berlin University and the other educational institutions became, however, seminal texts for a pedagogical theory whose central concept was that of *Bildung* (educational and cultural development). Following the convictions of German idealistic philosophy and the classical German poets (Schiller and Goethe), Humboldt wanted human individuals to be given every possible means to develop their physical and intellectual capacities. The educational institutions have to grant access to that formation of the human individual. The main aim of *Bildung* is not the acquisition of professional skills but the unfolding of the highest intellectual and artistic aspirations. Ancient Greece as the absolute summit of human development is the cultural reference of that educational program. The concept of *Bildung* certainly gave an ambitious aim to the educational institutions of Prussia and of Germany as a whole. This concept even served as a justification of otherwise unexplainable investments of the state in intellectual enterprises like the humanities, which have no practical ends. *Bildung* was an enormously successful concept as one of the central values of German identity, giving education a very high position in the value system of the nation. But just as Humboldt's University project was only partially realized in the Prussian University, the idea of *Bildung*—a generous and grandiose idea—did not so much create the (also politically) autonomous and free individuals Humboldt was thinking of, but had more of a compensatory function: consoling the German bourgeoisie for its nonparticipation in political power. Hence, all that wonderful *Bildung* did not prevent the German catastrophe.

3. *Anthropology*

The establishment of the University of Berlin and *Bildung* secured Humboldt's place in the pantheon of the German nation. But his main activity and—in the long run—also his most successful intellectual achievement, was his linguistic and philosophical work. In this domain, however, Humboldt is not—as some historiographers tend to think—a founding father of the scientific project of historicocomparative linguistics that is linked to the names of Franz Bopp and Jacob Grimm. Bopp and Grimm invented the principles for a historical reconstruction of the unity of Indo-European, Germanic, Romance, Slavic, and other languages and for the comparison of historically related languages. This was to be the successful—romantic—project of the mainstream linguistics of the nineteenth century. Humboldt is, on the contrary, the founder of an alternative linguistic project, which is more the summary of the linguistic ideas of the Enlightenment: the study of the diversity of the languages of the world. *Verschiedenheit*, or diversity, is the key term of Humboldtian linguistics which was to flourish—at least in part—in the twentieth century, when the historicocomparative paradigm was relieved by synchronical descriptive linguistics.

Humboldt's linguistic project was integrated into a larger 'anthropological' project. 'Anthropology' in the eighteenth century was a term which designated the empirical study of human beings, as opposed to philosophy, which investigates the universal nature of the human race. Humboldt sketched a vast research program for the investigation of the cultural manifestations of humankind. Understanding the creativity of the human mind was the philosophical motive that gave Humboldt's anthropological project its special dynamics. Two events which occurred in his lifetime were decisive for this research into creativity: first, as for any young intellectual of his generation, the French Revolution, the creation of a completely new, enlightened, political organization. Hence, one of his first publications was an article on the first French constitution (1791) and for his friends he wrote a booklength political treatise 'On the limits of the activity of the state' (1792), which was printed only after his death and which became a founding document of political liberalism. The other revolutionary event in Humboldt's lifetime was Goethe. Here again, what fascinated the young intellectuals was the 'genius,' the mystery of an unprecedented creativity. Humboldt lived in Jena, as a friend of Schiller's, in close contact with Goethe. He studied Immanuel

Kant, whose philosophy was to become the framework of his own theoretical developments. Humboldt's work is centered around the Kantian concept of imagination, *Einbildungskraft*, the faculty of the human mind that is responsible for the creation of the new.

Since imagination as the mediation between the senses and the intellect had remained rather mysterious in Kant's philosophy, Humboldt set out to find its roots. He discovered them in the difference between the sexes. The opposition of the male and the female in nature is the very basis of every new creation, not only of physical generation (*Erzeugung*) but also of the most 'sublime creature' of the imagination: thought. The loving embrace of the female and the male generates new living beings in the physical as well as in the intellectual world. Kant was shocked by Humboldt's somewhat pre-Freudian interpretation of his philosophy. But despite Kant's criticism, sexual union, 'matrimony' (*Vermählung*), love, would be the structural model for everything Humboldt set out to investigate: art, politics, education, and language. Just as dialectics is the fundamental figure of thought in Hegel, the erotic encounter is the basic structure of Humboldt's thought.

After his reflections on the anthropological roots of creativity, Humboldt wrote a book on Goethe (1798) which contains a very modern theory of aesthetics: art is not only a creation of the artist's imagination but also a (language-like) communicative event between the producer of the work of art and the observers, the readers, the listeners who recreate the work of art in their act of reception and thereby accomplish the creative process.

4. *Language*

While still in Jena, Humboldt wrote his first reflections on language: 'On thinking and speaking' (1795–6), a short sketch which already contained the main features of his language theory. Humboldt's sojourn in France and in Spain (1797–1801) would be decisive for his making language the very center of his intellectual endeavours. In Paris, he discussed philosophical issues with the *Idéologues*, in whose philosophy language played a crucial role, and here and in his journey to the Basque Country, he encountered the Basque language whose profoundly divergent structural properties intrigued Humboldt. Later, his brother Alexander would bring linguistic material from America, and in Rome Humboldt had access to the linguistic materials of the Jesuits (Lorenzo Hervás). For years, the Amerindian languages were Humboldt's empirical linguistic objects before he turned to the Malayo-polynesian languages to which his main—posthumous—work *On the Kavi Language on the Island of Java* (1836–9) is devoted. The introductory volume of that book, *The Diversity of the Human Language-Structure*, completed before his death on April 8, 1835 in Tegel (near Berlin), is Humboldt's major work. With language, Humboldt arrived at the very heart of his general quest for the creation of the new by that mysterious Kantian faculty of *Einbildungskraft*.

Language is not just a means for the communication of thought, it is the very creation of thought. According to the Kantian system, the senses and the intellect create the concept through their synthesis. But for Humboldt, the concept is not created independently of the word: word and concept form an undissoluble synthetical unity; thought is created as sound. Language is the 'formative organ of thought' ('*das bildende Organ des Gedanken*,' 1903–36, VII, p. 53), 'labor of the mind' ('*Arbeit des Geistes*,' 1903–36, VII, p. 46). Like art, language (thought-sound) also needs the dimension of the other: it must be heard and understood. Language as a cognitive activity reckons with the other, language is thought in the dimension of the other: *Mitdenken*, or cocognition. But surpassing art, language is a reciprocal creative activity: the hearer becomes speaker, 'the word formed by me resounds from the mouth of the other' (1903–36, VII, p. 56). This ultimate reciprocity is the completion of the linguistic syntheses. The duality of I and Thou is the 'fundamental type' of language.

Humboldt discovered that language is actually at the center of the Kantian system. Humboldt's language philosophy was a linguistic turn of transcendental philosophy. Since Bacon's discovery of the *idola fori*, European philosophy had become gradually aware of the fact that language is not just a device for the communication of thought but first and foremost a cognitive activity. This insight finds a coherent post-Kantian formulation in Humboldt.

The second important moment of that philosophy of language is the insight that the formation of thought, the linguistic 'transformation of the world into the property of the mind' (1903–36, IV, p. 420), always proceeds according to the special ways of historically particular communities, that is, according to particular languages. There is no such thing as language in general, language comes in languages. Thought generated by language, hence, is not uniform and universal but historically and culturally fragmented in different 'worldviews' (*Weltansichten*). The task of language, the generation of thought, is universal, but that task is realized in thousands of different ways in the different languages of humankind.

Now the new—or Leibnizian—moment of that insight is its positive interpretation: from Bacon onwards, philosophy has lamented over the particularity of worldviews in the different languages and wanted to get rid of those 'prejudices'; the semantics of natural languages is like 'a mist before our eyes' (Locke) which philosophical enlightenment must discard. Only Leibniz gave an optimistic twist to the discovery of the linguistically generated diversity of thought when he envisaged a study of all the languages

of the world in order to know the 'marvellous variety of the operations of our mind.' Herder took that idea from Leibniz, and Humboldt not only integrated it into the Kantian framework but also developed a huge program of descriptive linguistics.

5. *Comparative Study of Languages*

All the languages of the world were to be described according to what Humboldt calls their 'inner coherence' (*innerer Zusammenhang*), their particular structure. The languages of the world had mostly been grasped according to the traditional categories of Greek and Latin grammar. But the structural properties of a language can only be captured through its own categories. Humboldt envisaged an encyclopedia of the languages of the world, a new *Mithridates*. He started a whole series of descriptive works on different languages of the world of which only the *Mexican Grammar* (Humboldt 1994) has been published up to now. A second kind of structural studies aimed at a revised form of general grammar (which was also based on the Greco-Latin grammar): comparative investigations of linguistic categories throughout the languages of the world. Humboldt's article on the dual (1827) is an example of that research. After the completion of these structural descriptions, research into the parental relationships of languages (genealogy) or into structural similarities among languages (typology, classification) can be undertaken.

What is often considered as Humboldt's main contribution to linguistics, the so-called Humboldtian 'typology,' is neither central to his linguistic aims nor did Humboldt propose typology as an alternative to the dominant linguistic school, that is, to the historicocomparative studies of languages. On the contrary, in his last book, his unfinished work on Kavi, Humboldt eventually envisaged a comparative study of the Malayo-Polynesian languages which would follow Grimm's and Bopp's work on the Germanic and Indo-European languages. Humboldtian typology is a linguistic myth created by superficial readings of his texts. Humboldt was very sceptical about the classifications of the Schlegels, and eventually he explicitly refuted the legitimation of classifying languages, since languages are individuals and should be described as such.

The summit or keystone of Humboldtian linguistics is therefore the investigation of the 'character' of languages. This is just the opposite of classification and typology, of grouping similar entities together. It is a tentative attempt to grasp the very individuality of each individual language. Humboldt was not sure whether 'characterization' was still a scientific task. But whatever it is, it has to be tackled since it is the ultimate aim of linguistics to see how the human mind uses the instruments it has created, to investigate the very end of language: speech and discourse. The—necessary—structural description of a language is only 'comparable to its dead skeleton' (1903–36, VI, p. 147), the life of a language is discourse (*verbundene Rede*), a life giving 'character' to that language. Therefore, only languages that have a multifaceted literature yield the possibility of being described as having characters.

Nearly two centuries after Humboldt's sketch of a huge linguistic program, it can be said that the structural description of human languages he envisaged was the main task and achievement of linguistics in the twentieth century. It is not by chance that Humboldt is present in the seminal theoreticians of modern linguistics (Saussure (see *Saussure, Ferdinand de (1857–1913)*), Bloomfield (see *Bloomfield, Léonard (1887–1949)*), Hjelmslev). The nineteenth century was not very Humboldtian, with its emphasis on diachronical research as well as in its naturalistic research methods. Humboldt's 'keystone' of linguistics, the 'characterization' of languages, was attempted by some literary-minded linguists (e.g., Vossler), but was discredited by nationalistic and ideological interpretations of linguistic structures. Since the structural descriptions of the languages of the world have now virtually been completed, it is not astonishing that in actual linguistics typological attempts are again very important. The so-called Humboldtian typology is often quoted in that context. But it is—even more than the other famous Humboldtian invention, 'Humboldt's University'—rather a myth than a creature of the author. Without any foundation in Humboldt's work is also the widespread conviction that Humboldt discovered fundamental principles of Chomskyan linguistics, like the 'infinite use of finite means.' Humboldtian linguistics as a whole is in direct opposition to that kind of linguistic research: it is a hermeneutical research into the cultural diversity of human languages in historical communities, preferentially through (literary) discourse, and not a naturalistic investigation of the structure of the universal human mind considered as a universal grammar.

See also: Creativity and Cognition; Education, History of; Educational Institutions, History of; Educational Philosophy: Historical Perspectives; Instructional Psychology; Language and Literature; Language and Thought: The Modern Whorfian Hypothesis; Language, Continental Philosophy of; Liberalism; Liberalism: Historical Aspects; Linguistic Anthropology; Linguistic Turn and Discourse Analysis in History; Linguistic Typology; Linguistics: Overview.

Bibliography

Borsche T 1981 *Sprachansichten. Der Begriff der menschlichen Rede in der Sprachphilosophie Wilhelm von Humboldts*. Klett-Cotta, Stuttgart, Germany

Borsche T 1990 *Wilhelm von Humboldt*. Beck, Munich, Germany

Di Cesare D 1998 Einleitung. In: *Humboldt W von. Über die*

Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts. Schöningh, Paderborn, Germany, pp. 9–141
Haym R 1856 *Wilhelm von Humboldt. Lebensbild und Charakteristik*. Gaertner, Berlin
Humboldt W von 1836–9 *Über die Kawi-Sprache auf der Insel Java*, 3 Vols. Akademie der Wissenschaften, Berlin
Humboldt W von 1903–36. In: Leitzmann A et al. (eds.) *Gesammelte Schriften*. Behr, Berlin, 17 Vols
Humboldt W von 1988 (trans. Heath P) *On Language. The Diversity of Human Language-Structure and its Influence on the Mental Development of Mankind*. Cambridge University Press, Cambridge, UK
Humboldt W von 1994 *Mexicanische Grammatik*, Ringmacher M (ed.). Schönigh, Paderborn, Germany
Menze C 1965 *Wilhelm von Humboldts Lehre und Bild vom Menschen*. Henn, Ratingen, Germany
Meschonnic H (ed.) 1995 *La pensée dans la langue. Humboldt et après*. Presses Universitaires de Vincennes, Saint-Denis, France
Müller-Vollmer K 1967 *Poesie und Einbildungskraft. Zur Dichtungstheorie Wilhelm von Humboldts*. Metzler, Stuttgart, Germany
Müller-Vollmer K 1993 *Wilhelm von Humboldts Sprachwissenschaft. Ein kommentiertes Verzeichnis des sprachwissenschaftlichen Nachlasses*. Schöningh, Paderborn, Germany
Quillien J 1991 *L'anthropologie philosophique de G. de Humboldt*. Presses Universitaires de Lille, Lille, France
Scharf H W 1994 *Das Verfahren der Sprache. Humboldt gegen Chomsky*. Schöningh, Paderborn, Germany
Sweet P R 1978/80 *Wilhelm von Humboldt. A Biography*, 2 vols. Ohio University Press, Columbus, OH
Trabant J 1986 *Apeliotes oder Der Sinn der Sprache. Wilhelm von Humboldts Sprach-Bild*. Fink, Munich, Germany
Trabant J 1990 *Traditionen Humboldts*. Suhrkamp, Frankfurt am Main, Germany

J. Trabant

Hume, David (1711–76)

David Hume, a leading figure in the Scottish Enlightenment, contributed significantly to the disciplines of psychology, history, political science, and economics—as well as to more theoretical discussions in ethics, aesthetics, epistemology, and philosophy of religion. He was a controversial figure in his own day because of his criticisms of both natural and revealed religion and because of the political opinions he was thought to express in his *History*. His skeptical arguments concerning induction are still discussed by philosophers interested in probability theory, and his account of the foundations of morality enters into current disputes about the source of moral obligation.

1. *Life and Writings*

Hume was born in Edinburgh on April 26, 1711. The family was of genteel but of modest means: His father, Joseph Home, an advocate and member of the minor gentry of Scotland, died when he was 2 years old and his mother Katherine, who never remarried, brought up him, his sister, and older brother on their small estate of Ninewells in the Scottish Borders. Hume probably began taking classes at Edinburgh University in 1721 where he studied the Latin classics, moral philosophy, mathematics, natural philosophy, and possibly history. There is no record of his ever having taken a degree (which was not unusual at this period) and, in a short autobiography written just before his death, he notes that, while his family intended that he study law, his own interests were in 'Philosophy and general Learning.' He probably began planning his first philosophical work, *A Treatise of Human Nature*, in 1729 when he was about 18 years old.

Suffering from complaints which we would now consider to be largely psychosomatic, Hume decided in 1734 that he needed to pursue a more active life; he left Scotland for Bristol where he worked briefly for a merchant. After a few months he went to France, where he spent the next 3 years writing and studying. He went first to Reims, but soon settled at La Fleche in Anjou, in the vicinity of the Jesuit College where, more than 100 years earlier, René Descartes had been a student. The first two books of his *Treatise of Human Nature* ('Of the Understanding' and 'Of the Passions') were published in 1739, about a year after he returned to England, and the third ('Of Morals') was published in 1740 after he returned to Scotland. The *Treatise*, from which he hoped to gain literary fame, 'fell dead-born from press,' though it is taken by modern scholars as an important and seminal work. It was followed in 1741–2 by his very successful *Essays, Moral and Political* which were written to reach a more popular audience.

Hume unsuccessfully applied for university chairs at Edinburgh University in 1744–5 and at the University of Glasgow in 1751–2, where he had strong support from his friends (among them William Cullen) but was opposed by many of the clergy, who accused him of irreligion. While his *Treatise* contained no explicit discussions of religion, this was changed in the rewriting of its first book, which was published in 1748 under the name of *Philosophical Essays Concerning Human Understanding*. It contained his famous argument against rational belief in testimony for miracles, as well as the basis for his later attack on the design argument for the existence of God. Continuing to search for a career, Hume spent some time during the 1740s as secretary to General St. Clair and participated in a rather futile attack in 1746 on the coast of Brittany during the War of Austrian Succession. In 1747 he traveled with the General as his aide-de-camp on a Military embassy to the capitals of Britain's allies, Austria and Sardinia.

In 1749, Hume returned once more to Scotland and began an immensely productive decade. In 1751 he moved to Edinburgh from Ninewells; he was to live

there much of the rest of his life as a leading figure of the Scottish Enlightenment. Hume was elected secretary of the Philosophical Society of Edinburgh—a group devoted to the study of science and technology. (He was later, in 1762, to present to the Society Benjamin Franklin's paper describing the use of lightning rods.) In 1751 he also published his *Enquiry Concerning the Principles of Morals*, a reworking of the third book of his *Treatise of Human Nature*. His correspondence at this time also indicates that he had already written substantial portions of his posthumously published *Dialogues Concerning Natural Religion*. In 1752 Hume published his popular *Political Discourses* and began work on what was to become his most successful piece of writing, namely his multivolume *History of England*. He became Keeper of the Advocates' Library (later incorporated into the National Library of Scotland) in January, which gave him direct access to a large collection of historical sources. The first volume of the *History*, concerned with the reigns of James I and Charles I, was published in 1754 and the remaining volumes between 1756 and 1761. In 1758, Hume published *Four Dissertations*, which included his *Natural History of Religion*, *Of the Passions*, *Of Tragedy*, and *Of the Standard of Taste*.

In 1763, after the end of Seven Years War, Hume was invited to go to France as Secretary to the Ambassador, Lord Hertford. In France, where his writings were admired, he was lionized both by the mighty and intellectuals. He developed close relations with many of the philosophes, especially d'Alembert. Hume was *Charge d'affaires* for about 4 months in 1765 after Hertford left Paris. He returned to England at the beginning of 1766 bringing with him Jean-Jacques Rousseau, on the advice of Parisian friends; Rousseau sought refuge in England from persecution on the Continent. Hume and other British friends procured a pension for Rousseau from the King. However, this whole affair turned into a disaster as a result of Rousseau's paranoia. Rousseau denounced Hume as his enemy. Hume's defense of his own conduct was published by d'Alembert and Suard in Paris in October of 1766. Hume's last public appointment was as an Under Secretary of State from February 1767 to January 1768. He retired to Edinburgh where he continued to publish new editions of his works and died on August 25, 1776 as Britain's foremost man of letters.

2. *General Philosophical and Psychological Principles*

Philosophers still disagree about the nature of Hume's basic philosophical principles. It is agreed that he was a skeptic who argued throughout his writings that our basic commerce with the natural and social worlds is not founded on reason, but on the non-rational processes of the imagination and passions. What is in dispute is the nature of his scepticism, and the degree to which he sought to radically alter the conceptions of the natural and social worlds as conceived by his contemporaries (Read and Richmond 2000). While early twentieth-century philosophers saw Hume as an early forerunner of positivism, this interpretation has been challenged by recent scholars (Wright 1983). The question concerning Hume's adoption of conservative or liberal principles has also been a subject of dispute—though he claimed neutrality in discussing the political debates of his own day (Livingston 1998, Stewart 1992).

In his *Treatise of Human Nature* Hume set out to found all the other sciences on the science of human nature, and stressed that this science itself must be based on experiment and observation. Hume's foundational science of human nature was a science of the human mind. The basic psychological principle that he sought to establish in his *Treatise* was that of the association of ideas and impressions through resemblance, contiguity in space and time, and causality. He followed Locke in rejecting innatism and holding that the elemental building blocks of our ideas must be based on experience—the impressions of external and internal experience; however, he argued against Locke that these sources cannot supply us with our common sense or scientific conceptions of reality. Rather, the latter arise through irrational processes of the imagination and habits, based on the principles of association.

In a famous analysis, Hume argued that the principle of scientific induction by which we infer from past instances to future ones is based on custom and habit, and is not susceptible to rational justification. He denied that we have any understanding of the principles by which the natural world operates; at the same time, he argued that we naturally believe in causation between successive events that we observe to be constantly conjoined.

Hume held that we have as much reason to believe in necessary causes in the human as in the physical world. He argued that 'the conjunction between motives and voluntary actions is as regular and uniform, as that between the cause and effect in any part of nature' (1999, p. 154). At the beginning of his *Treatise of Human Nature* he set out to apply the experimental method to man, as it had been applied by Galileo, Newton, and others in the previous century to physics. At the same time, he believed that there is a problem in carrying out controlled experiments in the human sciences, because 'reflection and premeditation' would disturb the natural processes of the mind (2000, p. 6). He proposed to base his experiments on an observation of human behavior as it manifests itself 'in the common course of the world, by men's behaviour in company, in affairs, and in their pleasures.' While much of his data throughout his writings is taken from history, in Book 2 of the

Treatise, Hume made a series of observations of contemporary social behavior, testing his hypotheses concerning the causes of the passions of pride, shame, love, and hatred. He asked his reader to do thought experiments in which the antecedent circumstances were varied and the resulting passions were predicted. He regarded the subjects which we today call sociology and psychology as closely intertwined.

3. *Historiography, Economics, and Politics*

Hume was both a philosopher of history and a philosophical historian. He both articulated and applied principles for the assessment of historical sources. These principles are carefully laid out in his essay *Of Miracles* where he discusses the criteria we ought to use to judge the reliability of witnesses, as well as the need to balance those results against the intrinsic likelihood of the occurrence of an event (1999, p. 169ff). Following discussions of probability in John Locke and other early modern writers (Stewart in Stewart and Wright 1994, p. 171ff), he formulated criteria for the reliability of written or oral testimony: these included considerations such as consistency with other reports of the event, the determination of the character and expertise of the witnesses, the number of witnesses, and the self-interest of the witnesses in convincing others of the truth of their accounts. Hume held that we need to balance these considerations of the reliability of testimony against the absolute likelihood of the occurrence of a miracle. He held that in identifying an event as a miracle (i.e., a violation of a law of nature) we are already committed to the view that it has a 'proof' against its very occurrence. This claim is based on his argument that the only evidence we have for laws of nature is past regularity. He concludes that in the case of miracle reports the witnesses would have to be absolutely unimpeachable—a state of affairs that has never obtained. More generally, Hume argues that we must balance the testimony given by historical witnesses against the probability of the occurrence of any event. Thus, in his essay *Of the Populousness of Ancient Nations*, in *Essays Moral, Political and Literary* (1987, p. 377ff), he denies the claims of his ancient sources concerning enormous populations and armies of the ancient world. He argues, mainly on economic grounds, that the ancient nations, unlike modern ones, could not have supported such high populations as ancient writers claimed (D. Wootten, in Norton 1993, p. 281ff). In his *History of England*, Hume frequently appealed to general principles of human nature—those which he believed to be established by the newly emerging human sciences—to evaluate the credibility of his historical sources.

Hume is an important forerunner of classical economic theory. He argued that economic forces influence political structures, and that the development of modern liberty and government by law had been due to the development of trade and industry. Indeed, in the *Wealth of Nations*, Adam Smith states that Hume was the first writer to reverse the accepted view that political liberty precedes economic development. Hume saw in the rise of the merchants and tradesmen a check to the despotism that existed in premodern society. He argued that it was in the interest of these classes to live by equal laws and to limit the power of the aristocracy and the monarch. His analysis was based not only on his study of the rise of parliament in the early seventeenth century that led to the English civil war, but also in comparisons of modern European society generally with that of the ancient and medieval worlds. While Hume accepted James Harrington's social determinist view of the development of modern society, his originality lay in his postulation of the commercialism as a key factor in limiting the absolute power of the monarch (D. Wootten, in Norton 1993, p. 281ff).

Hume regarded man's economic behavior as deeply rooted in his psychological makeup. He argued that there are three factors that constitute human happiness—action, pleasure, and indolence (*Of the Refinement of the Arts*, in *Essays Moral, Political and Literary* 1987, p. 268ff). He regarded industry as a key to the leading of a balanced life in which pleasure and indolence follow naturally after a period of engaging activity. But, above all, he claimed that there is no pleasure that feeds on itself so much as the desire of the merchant to see his fortune grow. The habit of industry stimulated by this desire creates a frugality that does not naturally exist in an agrarian society based on inherited wealth. The love of gain which marks the activity of the tradesman and merchant results in large sums of money which can be lent at interest to stimulate further economic growth (*Of Interest*, in *Essays Moral, Political and Literary* 1987, p. 301). Low interest rates will be maintained by continuing increase in wealth.

Hume also argued that 'the encrease and consumption of all the commodities, which serve the ornament and pleasure of life, are advantageous to society' (*Of the Refinement of the Arts*, in *Essays Moral, Political and Literary* 1987, p. 272) and that a consumer society encourages the development of hard-working persons whose labor can be turned to the service of the state in times of national need. Key to his analysis is the idea that habits of hard work motivated by self-interest strengthen the nation–state. Developing the ideas of Bernard Mandeville, Hume argued against those in his day who complained that wealth and self-indulgence of the individual enfeeble both the individual and the society as a whole.

While Hume argued that 'Politics may be reduced to a Science' (1987, p. 14ff), he was skeptical about the ability of theorists in his own day to arrive at more than a few general principles of political behavior that would always apply (*Of Civil Liberty*, in *Essays Moral*,

Political and Literary 1987, p. 87ff). Certainly, he thought that history has taught us some general political principles concerning the different political behaviors of men under the different forms of government such as monarchy, aristocracy, and democracy. For example, from the ancient world we have learned that nonrepresentative democracy results in constant 'tumult and sedition.' However, he also argued that, given the relative newness of civilization, and the potential effects of 'education, customs, or principles' on human beings, it is impossible to predict how political behavior might evolve under circumstances which have not yet existed. He noted that among a number of other false generalizations, Machiavelli had concluded that if a prince relies on a single advisor that advisor would overthrow his master. Hume acknowledged that the evidence of ancient history, as well as that of 'the little disorderly principalities of Italy' of Machiavelli's own day, encouraged such a generalization; but he pointed out that it had been proved false by the experience of European hereditary monarchies of the seventeenth and eighteenth centuries. In general, while Hume thought that while there were a few generalizations that could be relied on in the science of politics, a cautious scholar needed to put principles forward as mere hypotheses to be modified by future circumstances.

Hume argued that political institutions evolve slowly, and that a necessary condition for civil liberty, namely the balance of power between parliament and the monarch, is the result of complex historical factors; liberty, he believed, is always under the threat of disappearing if one side or the other develops too much power. This caution, and his fear of the mob, made him a profoundly conservative thinker, especially toward the end of his life.

See also: Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Smith, Adam (1723–90)

Bibliography

Hume D 1947 *Dialogues Concerning Natural Religion*, N K Smith (ed.) Bobbs-Merrill, Indianapolis, IN

Hume D 1983 *The History of England from the Invasion of Julius Caesar to the Revolution in 1688*. Liberty Classics, Indianapolis, IN

Hume D 1987 *Essays Moral, Political and Literary*. Liberty Press, Indianapolis, IN

Hume D 1995 *Four Dissertations*. Thoemmes Press, Bristol, UK

Hume D 1998 *An Enquiry Concerning the Principles of Morals*. Oxford University Press, Oxford, UK

Hume D 1999 *An Enquiry Concerning Human Understanding*. Oxford University Press, Oxford, UK

Hume D 2000 *A Treatise of Human Nature*, D F Norton, M J Norton (eds.) Oxford University Press, Oxford, UK

Livingston D W 1998 *Philosophical Melancholy and Delirium: Hume's Pathology of Philosophy*. Chicago University Press, Chicago

Mossner E C 1980 *The Life of David Hume*, 2nd edn. Clarendon Press, Oxford, UK

Norton D F (ed.) 1993 *The Cambridge Companion to Hume*. Cambridge University Press, Cambridge, UK

Read R, Richmond K A (eds.) 2000 *The New Hume Debate*. Routledge, London

Rotwein E 1955 Introduction. In: Rotwein E (ed.) *David Hume, Writings on Economics*. Nelson, London and University of Wisconsin Press, Madison, WI

Stewart J B 1992 *Opinion and Reform in Hume's Political Philosophy*. Princeton University Press, Princeton, NJ

Stewart M A, Wright J P (eds.) 1995 *Hume and Hume's Connexions*. Edinburgh University Press, Edinburgh, UK and Penn State University Press, University Park, PA

Wright J P 1983 *The Sceptical Realism of David Hume*. University of Minnesota Press, Minneapolis, MI

J. P. Wright

Hunger and Eating, Neural Basis of

Eating is a familiar behavior. In humans, it is strongly influenced by cultural, social, and experiential factors. Consequently, eating has been a subject of considerable interest to social and behavioral scientists. In mammals, including humans, food intake is also a regulatory behavior with the primary function of supporting the continuous energy demands of body tissues. When considered from this perspective, food intake is influenced by hunger, satiety, and the biologic mechanisms that couple eating with internal caloric supplies and a stable body weight. This article discusses the signals that are important in enabling the central nervous system to control food intake, and also describes mechanisms thought to integrate these signals.

1. Historical Perspective

Biologic and behavioral studies on the control of appetite and satiety have shifted their focus markedly during the past three decades. The earlier preoccupation with the 'dual center' model, whereby food intake was controlled by a lateral hypothalamic feeding center and a medial hypothalamic satiety center, was based largely on investigations in which the putative centers were destroyed by stereotaxic electrolytic lesions (Stellar 1954). This once-dominant idea has now been discarded by most investigators in light of subsequent observations that have led to reinterpretation of the initial results. Specifically, the prolonged period in which animals do not eat after lateral hypothalamic lesions has been reproduced by destroying the dopaminergic neurons that ascend through the diencephalon; this aphagia and other related behavioral impairments appear to reflect a

decrease in cerebral arousal that precludes all voluntary behaviors, not just feeding (Stricker and Zigmond 1976). Moreover, the syndrome of overeating and obesity after medial hypothalamic lesions seems to derive primarily from the greatly increased level of insulin in the circulation associated with a decrease in sympathetic neural tone, as a result of which ingested calories are rapidly and preferentially stored via lipogenesis rather than being used gradually for ongoing metabolic demands (Friedman and Stricker 1976, Powley 1977).

More recent studies have emphasized metabolic and humoral factors in peripheral organs, in addition to the brain. An increasing number of investigations have related food intake to the functions of the stomach, liver, and adipose tissue, and have suggested an important role of peptide hormones, especially insulin, cholecystokinin (CCK), and leptin, in the control of food intake. In considering these issues, most contemporary investigators have abandoned the traditional depletion–repletion model, whereby a deficiency of caloric nutrients somehow triggers an appetite for food, the ingestion of which then leads to nutrient repletion and satiation. It is now clear that there are no true shortages of metabolic fuels under normal physiologic conditions. Instead, the key variable in controlling food intake seems to be the alteration in the source of the metabolic fuels used during each of the two phases of caloric homeostasis: during the absorptive phase metabolic fuels are supplied from recently consumed food that still is in the alimentary canal and being assimilated, whereas in the postabsorptive phase metabolic fuels are generated from accumulated triglyceride, glycogen, and protein stores derived from food consumed more remotely in time. Thus, according to this perspective the shift from satiety to hunger is not associated with an absolute metabolic deficit, but instead occurs when animals enter the postabsorptive state and must mobilize calories from tissue stores in order to maintain caloric homeostasis. Moreover, eating does not correct a true deficit or restore some aspect of tissue function to normal; rather, it simply re-establishes the pattern of an ample supply of usable calories from the gastrointestinal tract instead of from stores in adipose tissue, liver, or muscle (Woods and Stricker 1999).

2. *Satiety Signals*

Because the cells of animals cannot long survive depletion of calories, hunger cannot result from cellular nutrient shortages. Instead, the sensation of hunger apparently does not need to be stimulated but is always present, and animals are constantly stimulated to eat unless they have just eaten, in which case the inhibitory effects of satiety generated from the previous meal are still operative. In support of this perspective are laboratory findings that (a) experimenters cannot easily precipitate eating in satiated rats but can do many things to stop eating by hungry rats, (b) the size of spontaneous meals eaten by laboratory rats is unpredictable, but however much the animals eat in a given meal, meal size predicts how long they wait until they eat again, and (c) food deprivation does not cause a shortage of caloric nutrients in the circulation. The point here is that food intake appears to be initiated by the gradual disappearance of inhibitory satiety stimuli that were generated by the previous meal, not by the gradual increase in an excitatory hunger stimulus (LeMagnen 1985).

2.1 *Gastric Distension*

Several signals appear to stimulate satiety after a meal. One results from gastric distension, detected by stretch receptors on the walls of the stomach; the greater the distension, the greater the inhibition of eating (McHugh and Moran 1985). This afferent signal projects to the caudal brainstem via the gastric branch of the vagus nerve; as expected, animals eat unusually large meals when the gastric vagus is cut or its projection sites in the brain are destroyed. Experiments in rats have revealed that an intestinal hormone secreted during the meal, CCK, potentiates the gastric vagal signal and produces a heightened sensation of satiety related to gastric distension. This signal is related to the volume of the gastric contents, not its nutrient contents, because there are no known detectors of calories in the stomach. Drugs that acutely block CCK receptors increase food intake during a meal, as if the animal were less aware of gastric distension signals (Smith and Gibbs 1992).

2.2 *Ingestion of Calories*

The ingestion of calories, detected in the liver and possibly in the intestines, provides a second satiety signal, which is sent to the brainstem by another branch of the vagus nerve. This stimulus appears to develop only in the presence of insulin, a pancreatic hormone secreted when food is consumed. Thus, satiety normally is not lost as the stomach empties because the gastric signal related to distension is replaced by the postgastric signal related to caloric delivery (except in animals with insulin-deficient diabetes mellitus).

2.3 *Thirst and Nausea*

Two other meal-related signals also are known to be potent inhibitors of food intake. One is thirst,

resulting, for example, from the presence of relatively large amounts of salt in the food. The other is nausea, which might result from the inadvertent intake of toxins in food. However, these two variables are not relevant to most meals. Receptors that detect dietary salt have been located in the hepatic portal vein, which receives ingested food soon after it has been absorbed from the intestines. Receptors for circulating toxins are located in the area postrema in the brainstem. These two signals, together with gastric distension, provide three noncaloric stimuli that have marked inhibitory effects on food intake (Stricker and Verbalis 1991).

2.4 Body Weight

The influence of body weight on food intake must also be included in any consideration of satiety signals because food intake and body weight have been shown to be closely related. Thus, following a period of food deprivation, food intake is increased until the lost body weight is restored. Similarly, following a period in which food has been consumed in excess, food intake is diminished until the body weight that has been gained is lost.

Changes in body weight largely reflect changes in the storage of calories in adipose tissue. These stores are diminished during food deprivation to provide the calories needed for cellular function, and they are replenished when food intake resumes. Thus, as ingested calories are siphoned off to the adipose tissue, their effects on the liver are diminished, thereby reducing the duration of this postgastric satiety signal. In addition, food empties from the stomach more rapidly, thereby reducing the duration of gastric satiety. The opposite effects occur when calorie stores in adipose tissue have been augmented during a period of excessive food intake; both gastric and postgastric satiety signals are prolonged because the rate at which calories are stored in the swollen adipocytes is decreased (Stricker 1984).

2.5 Leptin

Aside from this indirect influence of body weight on food intake, in recent years a direct inhibitory signal to the brain from adipose tissue has been discovered. This satiety stimulus is a blood-borne protein, leptin, that is secreted from adipose tissue in proportion to adipocyte size; the fatter the animal, the more leptin is secreted and the less inclination the animal has to eat. Leptin receptors have been found in the basal hypothalamus, and when they are stimulated food intake is reduced (Woods et al. 1998).

The discovery of leptin illuminates genetic obesity in two strains of mice. Mice of one strain are born with leptin receptors in their brain but without the capacity to synthesize leptin; mice of the other strain are born without leptin receptors in their brain but with the capacity to synthesize leptin. As might be expected, an intravenous infusion of leptin reduces food intake in the first strain but not in the second.

3. Central Neural Systems

Food intake is a simple behavior influenced by a very complex array of stimuli and situational variables, the details of which remain to be fully illuminated. However, it is already clear that there are many levels at which the central nervous system exerts control over eating. For example, the spinal cord and brainstem influence all aspects of caloric homeostasis via the autonomic nervous system. The hypothalamus and limbic forebrain receive diverse signals related to the food being eaten and the degree of body adiposity, and integrate them with information about the taste of the food, the memory of that food and previous eating opportunities, competing motivations, and other aspects of the environment. Among the numerous neural controls of food intake, those located in the brainstem and hypothalamus are perhaps the best known, so this section will focus on these two brain areas.

3.1 Brainstem

Considerable evidence has accumulated that some control of food intake can be exerted by the brainstem. For example, it is well established that the nucleus of the solitary tract and the adjacent area postrema in the brainstem receive sensory fibers from gustatory receptors in the mouth and throat as well as afferent information from the stomach, intestines, and liver. Thus, these linked brainstem sites likely are the first locus at which relevant sensory input from the viscera is integrated with taste sensations in the central control of food intake. Other evidence derives from investigations of the decerebrate rat, in which all axonal connections between the caudal brainstem and the forebrain are severed at the midcollicular level in the midbrain (Grill and Kaplan 1990). The animal cannot seek food or initiate spontaneous meals but, like intact animals, it will reflexively swallow liquid food put directly into its mouth. Thus, when investigators deliver sucrose solution through an oral fistula, the decerebrate rat will swallow the sweet fluid, provided its stomach is empty, as after a period of food deprivation. However, it will not swallow water or saline when deprived of food and it will no longer swallow sucrose solution (but instead lets the administered fluid passively drip from its mouth) when its stomach is full, or after having received a systemic injection of CCK. Hence, sufficient control of food

intake for such appropriate responses exists entirely within the caudal brainstem, although decerebrate animals clearly are incapable of receiving input from forebrain sites that could modify their intake based upon learning or other experience.

Recent experimental observations provide additional support for an important role of the brainstem in the control of meal size. When the area postrema is ablated, rats eat much larger meals than normal, although total daily food intake is normal; that is, they compensate for their large meals by eating less frequently, indicating that long-term controls of food intake are not impaired. Similar effects are obtained in rats pretreated with capsaicin, a neurotoxin that destroys (among other things) most gastric afferent vagal axons and thereby eliminates the food intake-reducing effects of CCK, and therefore is assumed to damage the signal of gastric distension.

The food-related sensory information that enters the brainstem is relayed along prominent neural projections to the hypothalamus, amygdala, and other portions of the limbic system, as well as to the thalamus and gustatory cortex. Reciprocal neural connections to the brainstem from these rostral sites allow cognitive functions and emotions to influence the control of eating. Linked elements participate in the central control of the gastrointestinal tract, the liver, and autonomic function generally.

3.2 Hypothalamus

Complementing these controls of meal size are the influence on meal frequency exerted by body weight. As mentioned, changes in body fat stores are reflected in the levels of leptin in the circulation. Receptors for leptin are concentrated in the *arcuate nucleus*, which is located in the ventrobasal portion of the hypothalamus. Receptors actually have been associated with two different types of neurons in the arcuate nucleus. One, located in the medial arcuate nucleus, uses *neuropeptide Y* (NPY) as a neurotransmitter. These cells are of interest because it is well-established that injection of NPY into the brain of rats evokes eating, and that chronic infusion of NPY into the brain leads to obesity. Furthermore, axons of NPY-synthesizing neurons project from the hyopothalamic arcuate nucleus to the same focal site in the lateral hypothalamus at which eating is elicited by direct administration of NPY. Food deprivation stimulates activity in these NPY-containing neurons, whereas obesity or systemic administration of leptin depress their activity (Woods et al. 1998).

The other group of cells, located in the lateral arcuate nucleus, contain pro-opiomelanocortin (POMC). Those cells also have leptin receptors but, unlike the NPY-containing cells, the POMC-containing cells increase their activity when stimulated by leptin. The POMC-containing neurons use α-melanocyte stimulating hormone (α-MSH) as a neurotransmitter, and activate melanocortin receptors on lateral hypothalamic cells to decrease their activity and thereby inhibit eating. As might be expected, food deprivation decreases activity in the POMC-containing neurons, whereas their activity is elevated when rats are obese and circulating leptin levels are high. In short, there is a dual control by the arcuate nucleus over lateral hypothalamic activity and its control of food intake, one neural input (NPY) stimulating feeding and the other (α-MSH) inhibiting it; this influence is normally stimulated by alterations in blood-borne leptin reflecting changes in body adiposity (Elmquist et al. 1999).

The noncaloric signals (i.e., gastric distension, thirst, nausea) all have been shown to stimulate oxytocinergic neurons in the paraventricular nucleus of the hypothalamus (Stricker and Verbalis 1991). Projections from NPY- and POMC-containing cells in the arcuate nucleus also terminate in the paraventricular nucleus, allowing the integration of these various stimuli so as to influence output systems that control autonomic function and, indirectly, food intake. The central neural circuits that integrate the peripheral signals that control meal size (i.e., signals associated with the gastric, intestinal, and hepatic stimuli generated by a meal) with those that control meal frequency (i.e., signals associated with blood-borne leptin and perhaps other signals reflecting adiposity) are also currently under investigation.

See also: Body, History of; Body Image and Gender; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Eating Disorders, Determinants of: Genetic Aspects; Health and Mortality, Body Stature as a Measure of; Obesity and Eating Disorders: Psychiatric; Obesity, Behavioral Treatment of; Obesity, Determinants of: Genetic Aspects; Weight Regulation: Psychological Aspects

Bibliography

Elmquist J K, Elias C F, Saper C B 1999 From lesions to leptin: Hypothalamic control of food intake and body weight. *Neuron* **22**: 221–32

Friedman M I, Stricker E M 1976 The physiological psychology of hunger: Physiological perspective. *Psychological Review* **83**: 409–31

Grill H J, Kaplan J M 1990 Caudal brainstem participates in the distributed neural control of feeding. In: Stricker E M (ed.) *Handbook of Behavioral Neurobiology*. Plenum Press, New York, Vol. 10, Chap. 6, pp. 125–49

LeMagnen J 1985 *Hunger*. Cambridge University Press, London

McHugh P R, Moran T H 1985 The stomach: A conception of its dynamic role in satiety. *Progress in Psychobiology and Physiological Psychology* **11**: 197–232

Powley T L 1977 The ventromedial hypothalamic syndrome, satiety, and a cephalic phase hypothesis. *Psychological Review* **84**: 89–126

Smith G P, Gibbs J 1992 The development and proof of the CCK hypothesis of satiety. In: Dourish C T, Cooper J S, Iversen S D, Iversen L L (eds.) *Multiple Cholecystokinin Receptors in the CNS*. Oxford University Press, London, pp. 166–82

Stellar E 1954 The physiology of motivation. *Psychological Review* **61**: 5–22

Stricker E M 1984 Biological bases of hunger and satiety: Therapeutic implications. *Nutrition Reviews* **242**: 333–40

Stricker E M, Verbalis J G 1991 Caloric and noncaloric controls of food intake. *Brain Research Bulletin* **27**: 299–303

Stricker E M, Zigmond M J 1976 Recovery of function following damage to central catecholamine-containing neurons: A neurochemical model for the lateral hypothalamic syndrome. *Progress in Psychobiology and Physiological Psychology* **6**: 121–88

Woods S C, Seeley R J, Porte D Jr, Schwartz M W 1998 Signals that regulate food intake and energy homeostasis. *Science* **280**: 1378–82

Woods S C, Stricker E M 1999 Central control of food intake. In: Zigmond M J, Bloom F E, Landis S C, Roberts J L, Squire L R (eds.) *Fundamental Neuroscience*. Academic Press, San Diego, CA, Chap. 41, pp. 1091–109

E. M. Stricker

Hunter–Gatherer Societies, Archaeology of

Hunter–gatherer societies traditionally have been regarded as representing the 'original' human society—that is, preagricultural peoples reliant upon 'wild' foods and interacting with 'natural' landscapes. In this way, they have been differentiated from other, agricultural, societies. From a sociological perspective, hunter–gatherers emerged as the 'precapitalist' baseline in Marx's evolutionary framework—functioning as the original 'communal' society, practicing an economy different from those of others (Lee 1988). From a biological perspective, Darwin, among others, envisaged them as the human prototype, a position they still hold in more recent biological–ecological frameworks (Bettinger 1991). Anthropology and archaeology have absorbed these viewpoints, but more recently there has developed a significant change in emphasis—a redefinition of hunter–gatherers incorporating a much wider range in societal variation (Ingold et al. 1988).

Narrower, more traditional definitions had emphasized the following features. Hunter–gatherers had been perceived as small-scale, essentially egalitarian societies, of low population size and density and generally rather mobile. These societies were seen largely as *avoiding* accruing an economic surplus, practicing storage and impacting significantly upon land and resources. In contrast, recent revisions consider that hunter–gatherers also incorporated more complex societal formations within their ranks. These included extensive social networks and signs of social hierarchy—as well as chiefdoms based on hereditary ranked lineages. Demographics included relatively high population sizes and densities, heightened territorial behaviour, and levels of sedentism. Hunter–gatherer economies can be composed of significantly high levels of intensification achieved through economic surpluses and storage. It is envisaged that they had considerable impact upon land and resources via firing and other practices now recognized by recent ecological impact assessments (Williams and Hunn 1982).

Ethnographic studies of the northwest coast peoples of North America (from northern California to British Columbia) included many of the characteristics described above, some living in permanent wooden-housed villages of up to 1,000 people. These hunter–gatherer peoples placed a strong emphasis upon fishing and practiced no agriculture (Suttles 1968, Ames 1985). The Calusa of Florida, another North American ethnographic example, reflect all these features and more—including a centralized state formation which extracted tribute from subservient villages, a royal family, and a standing army. The Calusa also were without agriculture (Marquardt 1988). Recently the ethnographies of more classic hunter–gatherers have also been considerably revised (Schrire 1984, Lourandos 1997). In all, this evidence points to the possibility of a similarly broad range of pre- and nonagricultural peoples having existed in the past, when most of the habitable environments of the world were occupied by hunter–gatherers. Archaeology can provide answers to these questions, as recent investigations have unearthed.

1. *Archaeological Examples*

Archaeological examples of the historical evolution of these more varied features of hunter–gatherer society have been studied recently in a number of time periods and regions of the world. We now discuss five of these case studies in turn.

1.1 *European Upper Palaeolithic*

The Upper Palaeolithic in Europe took place during the final stages of the last ice age, from about 35,000 to 10,000 years ago. Modern people, *Homo sapiens*, had colonized the continent by this time. While the cultures of this lengthy period have much in common, clear trends towards more complex hunter–gatherer patterns are discernible in the last half of this period. For

example, after the coldest, driest phase which occurred around 20,000–18,000 years ago (the glacial maximum), climates were milder, especially in well-located southern latitudes such as southwestern France. At this time, known as the Magdalenian period, spectacular painted rock art, including images of extinct herd animals such as mammoth and bison, appear in limestone caves across the region. Occupation sites at this time are more numerous and much larger than previously. The suggestion is of increased numbers and densities of people living a more sedentary lifestyle. Hunting strategies were to target large numbers of herd animals that migrated seasonally to the highlands and back again, following the bloom of pasture grasses. The cave art which occurs in varied, regionalized styles, suggests the development of more territorial social units, each with their own set of artistic repertoires. Cave art, site numbers, and the economy, therefore, among other things, reflect a more populous, sedentary hunter–gatherer society. Burials, which are more common during this period, indicate that certain individuals were afforded more attention in death, suggesting the appearance of aspects of social differentiation (Conkey 1985, Gamble 1986, Jochim 1983, Mellars 1985).

At this time, out on the frigid and largely treeless southern Russian Plain, hunters resided in permanent hamlets of well-constructed mammoth bone dwellings. These bases were located in the lengthy riverine corridors, such as the Don and Dnepr, through which myriads of large herd animals migrated annually. Meat was stored in frozen subterranean pits. Hamlets were part of an intricate network of long-distance exchange—in exotic stone, shell, and 'Venus' figurines—which united peoples resident in a harsh, vast arctic desert. Social inequality is suggested by unequal numbers of storage pits around individual dwellings, and perhaps 'status' burials, in a prior phase (Soffer 1985). More intensive economic strategies are evident during this period, as was the case also in southwestern France.

1.2 Eurasian Mesolithic

Traditionally, the Mesolithic of Eurasia was considered a depauperate period when contrasted with the more spectacular Upper Palaeolithic. Recent reevaluations, however, point to complex trends, culminating in patterns of land and resource management, and ultimately, in some areas, to agriculture. This was a time of major climatic change, as warmer, moister conditions, similar to those of today, followed upon the colder glacial Pleistocene period around 10,000 years ago. The Mesolithic, therefore, is a time of adaptation by hunter–gatherers to changing environments both natural and cultural. The temperate zones of central and northern Eurasia were transformed from arctic deserts to mosaics of deciduous forests, lakes, and marshes. Following upon melting glaciers, the seas rose, creating, for example, the vast Baltic Sea and its resource-rich coastline in northwestern Europe. Populous, semi-sedentary hunter–gatherer societies appeared on these coastlines, of for example present-day Scandinavia, as sea levels stabilized about 6,000–5,000 years ago. They established large shell middens and practiced a wide range of fishing, hunting, and gathering (including storage of nuts). Comparable societies appeared in estuaries and river valleys throughout Europe (Price 1987, Zvelebil 1986).

At the same time in Japan, beginning around 12,500 years ago, the Jomon hunter–gatherers on the rich environments of the main island of Honshu lived in semi-sedentary subterranean pit houses along the coast and inland. They fished intensively and stored quantities of nuts, acorns, and seeds in underground storage pits. It has been suggested that nut groves may have been an economically managed resource as perhaps also were seeds, including grass seeds. These populous, territorially based and ritually complex hunter–gatherers intensified still further their fishing/storage economies through time. In the final stages of their history they resisted for some time the introduction of agriculture (from the Asian mainland) (Akazawa 1986, Crawford 1992), as did also the Scandinavian hunter–gatherers.

It was in drier regions of the Middle East (southwest Asia), however, that Mesolithic hunter–gatherers developed complex societies and economies that also gave rise to agriculture. The Natufian hunter–gatherers of present-day Israel and Palestine between 12,500 and 10,500 years ago practiced intensive collection and storage of wild grass seeds and nuts. They resided in hamlets of small, round huts, and practiced the long-distance exchange of obsidian and shells. Shrines are found within their hamlets where they also buried their dead. Around 10,000 years ago, the domestication of wild grasses—including wheat and barley—took place, signaling the shift to agriculture. However, all other cultural features were essentially similar to those existing before. Agriculture, therefore, was but an extension of the complex hunting–gathering out of which it had sprung. And later agricultural villages, including the walled town of Jericho, were centered upon previously existing Natufian hamlets (Bar-Yosef 1987, Henry 1985). The dynamics of complex hunting–gathering, therefore, gave rise to agriculture as we know it today.

1.3 Archaic North America

The Archaic period of North America, like the Mesolithic, takes place as climate shifted from glacial to modern after about 8,000 years ago. Long-term trends, taking thousands of years, can be recognized in the rich riverine regions of the American mid-continent, where the densest hunter–gatherer populations resided. In riverine corridors such as those of the Mississippi and Ohio, population grew slowly over

thousands of years, while in the Illinois valley, semi-sedentism in the form of cold-weather bases appeared in time. Intensification of the hunter–gatherer–fisher economy accompanied these developments. It took some 4,000 years for long-term trends towards more complex social formations to develop in these regions during the Middle Woodland period. By about 2,100 years ago complex ceremonialism in the form of extensive mounded burial precincts and exchange and trade in a wide range of exotic material (copper, obsidian, pottery) united most of the regions of mid and eastern North America. The richness of the natural environment cannot be the only factor controlling the development of these complex hunter–gatherer societies, as prime environmental conditions had been established as far back as 6,000–5,000 years ago, long before the above cultural features appeared and flowered. Some of the most complex societies—such as those of the Ohio Hopewell—were located outside the riverine corridor itself, on the more arid plains and were relatively mobile rather than semi-sedentary. Their most common unifying feature lay in the ceremonial realm (Bender 1985, Brown 1985). Although agriculture was known to these peoples, it was not adopted by them until more recent times—held at bay, as we have already observed, for some Mesolithic hunter–gatherer societies.

1.4 Preagricultural Mesoamerica

Less archaeological evidence is available for observing trends towards hunter–gatherer cultural complexity in preagricultural Mesoamerica, the region which lies between the two continents of North and South America. Hunter–gatherers of this region also developed agriculture independently. Very long-term trends, however, spanning some 4,000 years, suggest a gradual shift towards higher levels of 'domesticated' produce since about 7,000 years ago. Even by about 4,000 years ago less than one third of the diet was agricultural. By this time some of the main domesticated plants included the staples of later times, maize and beans. Sedentism, in the form of villages, occurred *after* this time, when agriculture had been more firmly established. Aspects of complex exchange and trade had accompanied these trends, and agricultural villages (and later cities) were constructed around ritual precincts, which strongly suggests that one of the key driving forces behind communal aggregations (that is, sedentism) was inter-group ritual itself (Flannery 1968, 1976, MacNeish 1970).

1.5 Holocene Australia

Recent evidence from the continent of Australia suggests that comparable trends towards more complex hunter–gatherer societies, similar to those of the Mesolithic and Archaic periods discussed above, are evident in Australia, especially from about 4,000–3,000 years ago. Aspects of these trends are also evident long before, for example after the glacial maximum or from about 16,000 years ago. In more recent times, however, long-term trends towards greater numbers of archaeological sites and use of sites occurs in many regions of the continent, including most major environmental zones—arid, tropical, and temperate. Demographic changes, including increasing populations and greater levels of sedentism, or reduced mobility, are suggested here, and are accompanied by other sociocultural changes (Lourandos 1997). In the tropical Cape York peninsula located in the northeast, shifts to more regionalized patterns of painted rock art occur during the last 2,000 years or so, suggesting more territorially based societies (David and Lourandos 1998, 1999). In arid Australia during the same period, a shift occurs towards more intensive seed-based (including grass seed) economies. The ethnographic descriptions of seed use (including grass seeds) by Australian Aborigines indicated they were managed via a wide range of practices including firing, irrigation, and the planting and broadcasting of seeds (Smith 1989, Veth 1989). As in all world regions studied, climatic change—in this case towards drier climates—accompanies many of these trends, but cannot on its own explain them.

2. Discussion

Trends can be observed, therefore, in many areas of the world which incorporate features of broader societal variation among nonagricultural peoples. These have included: increased population sizes and densities; sedentism storage and surplus; intensified and specialized economic practices; the suggestion of land and resource management; and broader social, including more hierarchical, formations. To what extent these prehistoric societies reflect ranking, including chiefdoms and the like, is more difficult to discern clearly. However, aspects of this, and the social context in which these systems are embedded, has been established in a number of case studies. The trends themselves took place over huge spans of time covering, in some cases, thousands of years.

Explanations of the trends in the development of hunter–gatherer society are many but tend to be somewhat polarized. As seen in the above case studies, climate, among other things, strongly influenced many of the general patterns and trends. But the dynamics within hunter–gatherer society itself also need to be considered. For example, incentives to intensify production (economy) may be embedded within the competitive social arena itself. In all, however, such

large-scale processes must have involved a complex set of factors—among them, ecological, demographic, and sociocultural.

Methodological problems also limit our investigations. Difficulties lie in gaining access to finer-grained, middle-range archaeological details. Issues of scale, for example, tend to blur distinctions between macro and micro issues in both time and space. Enormous spans of time and vast regions have been investigated in the general models offered. Access to archaeological data is a slow, meticulous process and answers to many of the questions raised here will take time to resolve.

In all, future directions lie with a much wider range of questions and approaches than were posed in more traditional times. The history of our hunter–gatherer past, quite clearly, has begun to be written anew.

See also: Chiefdoms, Archaeology of; Evolutionary Approaches in Archaeology; Fishing Societies; Hunting and Gathering Societies in Anthropology; Kinship in Anthropology

Bibliography

Akazawa T 1986 Hunter–gatherer adaptations and the transition to food production in Japan. In: Zvelebil M (ed.) *Hunters in Transition: Mesolithic Societies of Temperate Eurasia and Their Transition to Farming*. Cambridge University Press, Cambridge, UK, pp. 151–66

Ames K M 1985 Hierarchies, stress, and logistical strategies among hunter–gatherers in northwestern North America. In: Price T D, Brown J A (eds.) *Prehistoric Hunter–gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 155–80

Bar-Yosef O 1987 Late Pleistocene adaptations in the Levant. In: Soffer O (ed.) *The Pleistocene Old World: Regional Perspectives*. Plenum, New York, pp. 219–36

Bender B 1985 Prehistoric developments in the American midcontinent and in Brittany, northwest France. In: Price T D, Brown J A (eds.) *Prehistoric Hunter–gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 21–57

Bettinger R L 1991 *Hunter–gatherers: Archaeological and Evolutionary Theory*. Plenum Press, New York

Brown J A 1985 Long-term trends to sedentism and the emergence of complexity in the American Midwest. In: Price T D, Brown J A (eds.) *Prehistoric Hunter–gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 201–31

Byers D S (ed.) 1970 *The Prehistory of the Tehuacan Valley*. University of Texas Press, Austin, TX

Conkey M W 1985 Ritual communication, social elaboration, and the variable trajectories of Paleolithic material culture. In: Price T D, Brown J A (eds.) *Prehistoric Hunter–gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 299–323

Crawford G 1992 Prehistoric plant domestication in East Asia In: Cowan W, Watson P J (eds.) *The Origins of Agriculture*. Smithsonian Institute Press, Washington, DC, pp. 7–38

David B, Lourandos H 1998 Rock art and socio-demography in northeastern Australian prehistory. *World Archaeology* **30**(2): 193–219

David B, Lourandos H 1999 Landscape as mind: Land use, cultural space and change in north Queensland prehistory. *Quaternary International* **59**: 107–23

Flannery K V 1968 Archaeological systems theory and early Mesoamerica. In: Meggers B J (ed.) *Anthropological Archaeology in the Americas*. Anthropological Society of Washington, Washington, DC, pp. 67–87

Flannery K V (ed.) 1976 *The Early Mesoamerican Village*. Academic Press, New York

Gamble C 1986 *The Palaeolithic Settlement of Europe*. Cambridge University Press, Cambridge, UK

Henry D O 1985 Preagricultural sedentism: The Natufian example. In: Price T D, Brown J A (eds.) *Prehistoric Hunter–gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 365–84

Ingold T, Riches D, Woodburn J (eds.) 1988 *Hunters and Gatherers 1: History, Evolution and Social Change*. Berg, Oxford, UK

Jochim M 1983 Palaeolithic cave art in ecological perspective. In: Bailey G N (ed.) *Hunter–gatherer Economy in Prehistory*. Cambridge University Press, Cambridge, pp. 212–9

Lee R B 1988 Reflections on primitive communism. In: Ingold T, Riches D, Woodburn J (eds.) *Hunters and Gatherers: History, Evolution and Social Change*. Berg, Oxford, UK, pp. 252–68

Lourandos H 1997 *Continent of Hunter–gatherers: New Perspectives in Australian Prehistory*. Cambridge University Press, Cambridge, UK

Marquardt W H 1988 Politics and production among the Calusa of South Florida. In: Ingold T, Riches D, Woodburn J (eds.) *Hunters and Gatherers: History, Evolution and Social Change*. Berg, Oxford, UK, pp. 161–88

Mellars P A 1985 The ecological basis of social complexity in the Upper Palaeolithic of Southwestern France. In: Price T D, Brown J A (eds.) *Prehistoric Hunter–gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 271–97

Price T D 1987 The Mesolithic of Western Europe. *Journal of World Prehistory* **1**(3): 225–305

Schrire C (ed.) 1984 *Past and Present in Hunter–gatherer Studies*. Academic Press, Orlando, FL

Smith M A 1989 Seed gathering in inland Australia: Current evidence from seed-grinders on the antiquity of the ethnohistorical pattern of exploitation. In: Harris D R, Hillman G C (eds.) *Foraging and Farming: The Evolution of Plant Domestication*. Unwin Hyman, London, pp. 305–17

Soffer O 1985 Patterns of intensification as seen from the Upper Paleolithic of the Central Russian Plain. In: Price T D, Brown J A (eds.) *Prehistoric Hunter–gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 235–70

Suttles W 1968 Coping with abundance: Subsistence on the Northwest Coast. In: Lee R B, DeVore I (eds.) *Man the Hunter*. Aldine, Chicago, pp. 56–68

Veth P 1989 Islands of the interior: A model for the colonization of Australia's arid zone. *Archaeology in Oceania* **24**: 81–92

Williams N W, Hunn E S (eds.) 1982 *Resource Managers: North American and Australian Hunter–gatherers*. Westview Press, Colorado, CO

Zvelebil M 1986 *Hunters in Transition: Mesolithic Societies of Temperate Eurasia and Their Transition to Farming*. Cambridge University Press, Cambridge, UK

H. Lourandos

Hunting and Gathering Societies in Anthropology

'Hunting and gathering' is a conventional term for subsisting on the 'natural produce' of the earth by hunting and trapping wild animals, gathering wild roots, fruits, and berries, fishing, and collecting honey of wild bees, as opposed to producing food by agriculture and herding. By hunting and gathering societies (HGS) ethnographers mean contemporary indigenous people who fully or partly engage in these activities, now or formerly, or who see themselves as hunters. (For brevity, the terms 'hunter-gatherers,' 'hunters,' and 'foragers' are sometimes used synonymously.) The term is also used in general reference to prehistoric populations, living before the agricultural revolution (around 10,000 years BC). In common usage, it also stands for an alien 'Other,' diversely constructed by modernity in order to define itself. In this article, the primary focus is placed on the first meaning, while note is taken of its brush with the other two. The term is first examined historically, with an outline of its origins and changes of usage over time. Next, several clusters of features that are partially shared by these societies are presented in model terms, after which the diverse populations concerned are described by region, with the focus on their historical circumstances.

1. *Contested Meanings*

An understanding of the expanding meanings of HGS necessitates a return to the origin of the field and perusal of its historical development. In the seventeenth century, writers invoked 'primitive peoples,' who hunt and gather for their subsistence, in discussing the 'natural state' of humankind. By the eighteenth century these were typed 'hunter-gatherers,' in distinction from 'herders,' within a broad 'savage' class, in writings on the origins of property relations. In the nineteenth century evolutionist work continued to draw on these specific peoples, often without referring to the class as a whole, in studies of the origins of a variety of other social institutions, from kinship through politics to religion.

The notion consolidated as a central term in early twentieth-century cultural-ecology theory, most influentially in Julian Steward's work (1936). Concerned with the relation between society and environment, Steward argued that for any given mode of subsistence, successful adaptation determines a certain level of socio-cultural integration. In the case of hunting-gathering, he argued, this is a 'band,' which is universally patrilineal, exogamous, and territorial. HGS was associated with bands, and hunting peoples were comparatively studied as band-societies (see, for example, Damas 1969), until the 1966 symposium 'Man the Hunter' (Lee and DeVore 1968). This interdisciplinary symposium is commonly considered the beginning of modern hunter-gatherer studies.

The symposium was inspired by the assertion that human beings had subsisted by hunting for over 99 percent of the time of human life on earth. Linking cultural-ecological and (neo)behavioral-evolutionary perspectives, its major objective was to learn about human origins and nature by comparing 'living hunters.' Uniform patterns found among them, it was posited, can be regarded as hunter-gatherers' successful adaptations, present and past.

Serious reservations were raised about its agenda. One objection, for example, was that no contemporary group could be seen as 'living relics,' or as direct descendants, of the Pleistocene. Even if a group had been found to subsist purely on hunting, its people would have had a history as long as any other, and they would not have lived any longer in a world populated by hunters as in the Pleistocene. Another objection was that over the span of 2,000,000 years humans had changed considerably, biologically, and psychologically, enough to render homogenizing them inappropiate.

The ethnographic comparison, at the same time, shook accepted wisdom to the core. Contrary to Steward's persistent thesis, research revealed that most bands are actually socially fluid and nomadic. In many cases, gathering makes no less important a contribution to subsistence than hunting. In contrast to contemporary stereotypes and bourgeois intuition, the comparison showed that these peoples meet their needs with surprisingly little labor. (Marshall Sahlins provocatively dubbed them 'the original affluent society.')

Outweighing the reservations, this successful outcome entangled HGS with the 'Man the Hunter' project. HGS came to mean a type of people, rather than diverse peoples engaged in a type of action, encompassing populations over a vast evolutionary span of time. The term now embodied the hypothesis that, at their core, these populations have similar socio-cultural forms, determined by adaptation to the environment. From a vague term of reference to 'very simple' peoples, and then to 'band societies,' it connoted a firm, homogeneous, and exclusive social category, a category of humanity defined techno-economically.

Subsequent ethnography research, which burgeoned in the wake of 'Man the Hunter,' grew diverse in focus and approach, as seen for example in the series of International Conferences on Hunting and Gathering Societies (1978 to the present, e.g., Ingold et al.

1988, Vols. 1–2, Burch and Ellana 1994). Issues arising from late twentieth-century indigenous circumstances and cultures were addressed, including political struggles and self-administration in modern states, and indigenous peoples' own perceptions of their environment, economy, and self. Historical, interpretative, regional, and applied perspectives commanded attention; evolutionary-ecological studies of foraging behavior increasingly took a separate course. Indigenous people started contributing to academic research, as anthropologists and interested parties. This growing ethnography showed a diversity of social forms among hunting-gathering peoples, from egalitarian to hierarchical, and even slave-owning societies. The 1970s meaning of HGS, as a firm, homogeneous, and exclusive social category, was increasingly challenged. In an attempt to save it, the category was diversely divided into subtypes (e.g., see Woodburn 1982); or it was recharacterized, for example, by a collective appropriation of nature (see Ingold 1987) or by a distinct cultural perception of the environment (see Bird-David 1990).

The 1980s and early 1990s saw many controversies (see review by Myers 1988), sparked partially by inflexible readings of the term HGS. Scholars disputed the category as pertaining to a class of societies that are both singularized by hunting-gathering and distinguished from agricultural and pastoral peoples, because the ethnography revealed continuities between hunter-gatherers and other societies, as well as great diversity among the hunter-gatherers themselves. Other scholars questioned the categorization of societies by mode of subsistence, as research showed heterogeneous and changing economies. Hunting-gathering was found to be articulated with other pursuits, going back in some cases centuries and even millennia, and, at any rate, undergoing rapid changes since the mid-twentieth century. Students debated whether any living group sufficiently complies with the subsistence criterion and whether this criterion generates a meaningful category, one whose members share more than the selecting mode of subsistence. Even the field at large was critiqued as spurious, by researchers seeing it in terms of the 1960s objective to learn about prehistory and evolution. Their criticism drew on ethno-historical and archaeological work suggesting arguably that these societies have not escaped world and regional forces. HGS is sometimes used nowadays simply to denote the assemblage of indigenous peoples discussed by practitioners within the field of hunter-gatherer studies (see Kelly 1995).

HGS has acquired multiple layers of meanings. It is a circumlocution for a historically-situated field of research that provides a stimulating comparative context and a scholarly communal framework within which to explore a broad array of distinctive issues. Ethnographers construct 'partial connections' (in Strathern's (1991) sense) between these societies by studying their respective groups while keeping the others 'in mind.' They call to attention partially shared features, which determine common issues in the field, and contribute distinctive perspectives to anthropology generally. Hunting and gathering activities are a focus of attention, rather than a criterion of demarcation, for scholars as for the indigenous peoples themselves.

2. *Issues*

This section presents in model terms some of the partially shared features that came to light in the course of ethnographers' comparative discussion of their field experiences and material.

2.1 *Hunting and Gathering*

In many of these societies, people combine these activities with other economic pursuits (e.g., trade in minor forest produce, horticulture, work for neighbors, wage-labor, social benefits, etc.). However, hunting and gathering themselves carry special significance; they are a total phenomenon encompassing economic, social, and symbolic aspects. Hunting–gathering, moreover, often influences the way in which people carry out such other activities. They often return to hunt and gather at least periodically, when not prevented from doing so. They regard what they collect as 'real food,' and while they do pursue other activities, in some cases hunting–gathering still provides a major part of their diet.

2.2 *Band Society*

In many of these societies people live in small groups (bands) consisting of several families; the groups are flexible. Making visits, and even shifting residence, from place to place, are common. The total population numbers a few thousand, and may be as low as several hundred. Men and women often know each other personally: they all are (or at least are classified as) family and relatives. Their sense of themselves as family often survives growing integration into nation-states.

2.3 *Sharing*

The land and its resources in many of these societies are often collectively appropriated (see Ingold 1987). Game is publicly shared; other resources, whether procured from the environment or purchased, are shared in response to requests directed by those who lack something to those whom they perceive have it ('demand-sharing': Peterson 1993). Sharing and the expectation of it is part of broader nonaggrandizing

economic reasoning and conduct. While hard to express in modern urbanite language, sharing extends not only to distributive but also to social experience. People in the group share 'each other' (Ingold 1987): company, dwelling spaces and work, stories, memories, and personal histories. They experience their selves as coalescent, rather than as a group of separate individuals. Relationships between people, rather than being regenerated by coercing external structures, spring from the vivid experience of sharing. Sharing, which is normatively promoted, fosters an egalitarian society in many cases (see Woodburn 1982). While closely connected, people usually enjoy considerable personal autonomy.

2.4 Giving Environment

Nature is locally experienced neither as a mechanical system nor as objects for human use and ownership, but as an assembly of sentient beings related to and shared with (Bird-David 1990). These beings (animals, forest, winds, hills, etc.) frequently are approached as relatives, often as grandparents. People care for them, and this helps sustain their collective identity. People engage with these beings through Shamanistic practices and by pursuing hunting–gathering. To hunt and to gather are perceived not just as subsistence pursuits but also as means of communication with the 'natural' relatives. Even when forced to reside away, and to adopt modern life-styles, people commonly practice their rituals, and return to their places to hunt-gather and 'keep in touch' with these relatives. This animistic perception is integrated with deep ecological understanding of their neighborhood. It often propagates responsiveness to and sometimes responsibility for the environment, along with its utilization.

3. Populations

The indigenous populations considered as hunter-gatherers are described in this section, by geographical cluster, in keeping with the *Cambridge Encyclopaedia of Hunters and Gatherers* (Lee and Daly 1999). The figures provided should be regarded as suggestive, as politics of identity, especially in First World situations, and dispersal and inaccessibility, especially in Third World countries, severely limit enumeration.

3.1 Australia

About 266,000 people identify themselves as of Aboriginal or Torres Strait Islander descent, constituting 1.6 percent of the national population (1991 census). This is the case on a continent inhabited solely by hunter-gatherers until the arrival of Europeans in 1788. Compared with an estimated 200 languages and 600 dialects in that year, they now divide into groups speaking about 50 languages. Among the best known are Aranda, the nineteenth-century study of whom was pivotal in anthropology, and Walripi- and Pintupi-speakers, desert-dwellers and the last to make contact with Europeans. A disproportionate number of Aboriginal people are now largely dependent on social security payments, and do poorly by any standard social indicator. At the same time, a revival of tradition is on the increase, in part due to the growing national and international markets for Aboriginal art (especially the desert 'dot-painting'). Politically, significant advances have been recently made. First was the establishment in 1990 of the Aboriginal and Torres Strait Islander Commission (an indigenous body formed to make policy and distribute federal funding for Aboriginal people). This was followed by the High Court recognition in 1992 of native land title (the 'Mabo case').

3.2 North America

Commonly called American Indian Nations in the US, and First Nations in Canada, there are about 500 indigenous groups in this subcontinent. The stratified Tlingit and Haida of the northwest coast, and the egalitarian Arctic Eskimo–Aleut speakers, are renowned for their role in early anthropology. The sub-Arctic Innu, Cree, and Dene are known for their opposition to contemporary environment-threatening development schemes. Prior to European colonization, two-thirds of the subcontinent had been occupied by hunting and gathering people. Devastated by the colonization, some groups perished, and some survived in fragmented forms in reservations. To date, a disproportionate number experience severe social problems. However, recent decades have seen expansions of indigenous organizations, rapid growth of arts, and revival of traditional spirituality (attracting New Age followers). Hunting and gathering economies are being renewed, especially on the north and the northwest coast, as a result of some judicial breakthroughs that brought a degree of recognition of aboriginal and treaty rights. A major landmark was the 1999 establishment of Nunavut, an Inuit majority territory within Canada's federal structure.

3.3 South America

Foraging people in the subcontinent's southern cone were almost extinct by the 1920s, and only a handful of each group remains today, assimilated into other communities. A wide spectrum of forest communities, cultivators and hunters by trade, live scattered in the Amazonian hinterlands. In the 1960s they were viewed predominantly as devolved agriculturists, and only later did their egalitarian and nomadic life-styles earn them the attention of the discipline. The Paraguayan Ache is one of the best known groups. Most of these indigenous people have been converted and restricted to reservations; since the 1960s large-scale economic

development has threatened their environment and life-ways even more. As a result of joint campaigns by indigenous activists and international conservation and human rights movements, a slow but significant movement towards recognition of indigenous territories has been taking place since the 1980s.

3.4 North Eurasia

Latecomers into the hunter-gatherer discourse, there are 26 Arctic and sub-Arctic groups in the Russian north. They are divided into at least six language families, and number approximately 184,500, constituting 0.06 percent of the entire Russian Federation population (in 1990). In the seventeenth and eighteenth centuries the majority moved to reindeer pastoralism as their main subsistence occupation, supplementing it by hunting and fishing. Autonomous districts were established for some of them in the 1930s, although political and economic decisions were still made in Moscow. A substantial majority is now modernized, urbanized, and Russified, but activists (mainly writers) campaign for cultural renewal. Under the mounting economic difficulties in Russia, many are attempting to restore some traditional practices.

3.5 Africa

Bushmen (San) and Pygmy groups caught the 1960s' imagination, not least as their habitat was regarded the 'cradle of humankind.' They number 100,000 and 285,000 respectively; each comprises diverse linguistic groups spread across more than half a dozen countries. A further 55,000 people, largely in East Africa, bring the total number of African foragers, part-foragers, and former-foragers to 440,000. Only a tiny fraction of them subsist intensively on foraging. Most keep close ties with cultivators and pastoralists; the Pygmy groups are known for what is described as symbiotic relationships with their neighbors. With worsening circumstances, some groups sought help from development agencies and church groups and some joined environmental activists in protest against ill-advised large development schemes. The Ju/'hoansi (alias !Kung), the best known hunter-gatherer group owing to Lee's extensive work (e.g., 1979), have initiated self-help schemes intended to develop awareness of nation-politics, and leadership and management styles adapted to it.

3.6 South Asia

Less than two dozen indigenous groups live in the South Asian subcontinent. These total less than 20,000 out of over 1,000,000,000 people (wider estimates, including former foragers, put their number at 1,300,000). They include the exceptional Jarwas and Sentinelese of the Andaman Islands, who by hostile behavior have warded off outside contact up to the present, and South Indian forest-groups who, going back 2,000 years, have played an important part in the economies of the subcontinent's ancient civilizations. The latter groups have provided the main society with exotic spices, medicinal herbs, honey, and meat, occasionally guiding Hindu pilgrims and officiating in their forest temples. Some groups are being pressed to 'develop' into settled agriculturalists by a national modernist policy that sees them as 'backward people' whose existence shames the country; some combine hunting–gathering with trade in minor forest produce and wage-work. Sri Lanka's Wanniyala-aetto (formerly Veddahs) raised strong opposition to their forced relocation, going as far as the United Nations.

3.7 Southeast Asia

Foraging and swidden farming are closely linked in this region, with historical shifts back and forth between them, frequent combinations, and close trade-links between foragers and farmers. In the Philippines there are about 25 distinct linguistic groups, totaling 15,000; in Malaysia, eight linguistic groups total 2,500 people; in Indonesia the total is 15,000, in Borneo it is 12,000, in China it is 4,500, and in Sumatra and Thailand it is 800 and a few hundred, respectively. Governments in this region encourage commercial exploitation of these people's forest habitat, and they rapidly become poor farmers and laborers.

4. Conclusions

Stimulated by the hope to learn about human evolution and prehistory, the study of 'living hunters' has developed into a fruitful area of study in and for itself. These living peoples still afford a study of behavioral-ecological parameters of hunting–gathering, which are important for them, and which they pursue when they can. Over and above this, however, they open a window to socioenvironmental ethics and praxis, conceptions of society and nature, human relationships and economy, which are markedly different from familiar models. In broadening scholars' imagination, that cannot but be biased by urban life-experiences and intuitions, hunting and gathering peoples can help scholars think about other forms of life: human life in the past as well as possible futures. Indigenous peoples for their part increasingly explore new modernities that integrate their traditional practices with life in twenty-first-century nation-states.

See also: Economic Anthropology; Hunter–Gatherer Societies, Archaeology of; Social Evolution: Overview

Bibliography

Barnard A 1983 Contemporary hunter-gatherers: Current theoretical issues in ecology and social organization. *Annual Review of Anthropology* **12**: 193–214

Bird-David N 1990 The giving environment: Another perspective on the economic system of gatherer-hunters. *Current Anthropology* **31**: 189–96
Burch Jr E S, Ellana L J (eds.) 1994 *Key Issues in Hunter-gatherer Research*. Berg, Oxford, UK
Damas D (ed.) 1969 *Contributions to Anthropology: Band Societies*. National Museum of Canada, Ottawa, ON
Ingold T 1987 *The Appropriation of Nature: Essays on Human Ecology and Social Relations*, 1st edn. University of Iowa Press, Iowa City, IA
Ingold T, Riches D, Woodburn J (eds.) 1988 *Hunters and Gatherers*. Berg, Oxford, UK, 2 vols
Kelly R L 1995 *The Foraging Spectrum: Diversity in Hunter-gatherer Lifeways*. Smithsonian Institution Press, Washington, DC
Lee R B 1979 *The Kung San: Men, Women and Work in a Foraging Society*. Cambridge University Press, Cambridge, UK
Lee R B, Daly R (eds.) 1999 *Cambridge Encyclopaedia of Hunter-Gatherers*. Cambridge University Press, Cambridge, UK
Lee R B, DeVore I (eds.) 1968 *Man the Hunter*. Aldine, Chicago
Myers F 1988 Critical trends in the study of hunter-gatherers. *Annual Review of Anthropology* **17**: 261–82
Peterson N 1993 Demand-sharing: Reciprocity and the pressure for generosity among foragers. *American Anthropologist* **95**: 860–74
Steward J H 1936 The economic and social basis of primitive bands. In: Lowie R H (ed.) *Essays in Anthropology in Honor of A. L. Kroeber*. University of California Press, Berkeley, CA
Strathern M 1991 *Partial Connections*. Rowman & Littlefield, Savage, MD
Woodburn J 1982 Egalitarian societies. *Man* **17**: 431–51

N. Bird-David

Hurst, James Willard (1910–97)

It is hardly controversial that J. Willard Hurst of the University of Wisconsin was the most influential scholar ever within the realm of US legal history. By emphasizing factors beyond legal doctrine and by altering the way people think about time, place, and change itself, Hurst enlivened and transformed an entire academic field. As leading legal historian Lawrence Friedman said: 'Willard Hurst, as a scholar, did what few scholars have ever succeeded in doing: he created a whole new field of study, and virtually out of nothing.'

1. *Biographical Background*

James Willard Hurst was born on October 6, 1910 in Rockford, Illinois, where his father ran a movie theater and had other business interests. When Hurst attended Williams College from 1928 to 1932, he majored in economics, became editor in chief of the student newspaper, and president of the International Affairs Club. He also wrote a prize-winning student essay on international implications of the English Gold Clause crisis of 1931, published in the *Journal of Political Economy* the year he graduated. Hurst then attended Harvard Law School, and reported that he was disgruntled about the emphasis on abstractions and the failure to pay attention to the Depression and to the ideas of the legal realists. A note editor of the *Harvard Law Review*, Hurst graduated in 1934 and spent the 1935 to 1936 academic year as a research fellow for Felix Frankfurter, working primarily on a lecture series which Frankfurter published as *The Commerce Clause under Marshall, Taney, and Waite* (1937).

Hurst next served as law clerk for 80-year-old US Supreme Court Justice Louis D. Brandeis during the 1936 term, one of the court's most contentious ever. Frankfurter reported that Brandeis said that Hurst ranked as his best clerk ever, 'professionally speaking.' Brandeis thought the states to be invaluable laboratories for progressive regulation and lawmaking, and he urged young men who worked for him to return to their home regions. Hurst became one of the very few to heed Brandeis's advice. He began his lifetime association with the University of Wisconsin in 1937, when he joined the law faculty in Madison, only a short drive north of Hurst's hometown.

By 1941, Hurst had compiled several books of teaching materials and had coauthored *Law in Society* with his dean, Lloyd Garrison. This pioneering casebook focused on the myriad ways that the plight of a widow—whose husband, a carnival worker, was crushed under a company tractor—intersected with legal institutions beyond the courts. After Pearl Harbor, Hurst and his wife, Frances Wilson, moved to Washington, DC. Hurst first worked for the general counsel of the Board of Economic Warfare and then joined the navy, where his assignments included work on a highly publicized treason case, leading eventually to his *The Law of Treason in the United States* (1971).

Even before returning to the Wisconsin law faculty after World War II, Hurst began to articulate in print the crucial link he perceived between legal history, done in a new way, and 'opportunities to apply informed, humane reason to influence the course of events.' The 'Wisconsin idea' of enlightened political leadership had lasting appeal for Hurst—and he repeatedly rejected job offers to teach at other prestigious institutions. In the course of rejecting an offer from Yale Law School in 1938, for example, Hurst wrote to his mentor Felix Frankfurter that Wisconsin afforded 'just about an ideal "laboratory" situation.' Despite Yale's offer of higher rank and twice the salary, Hurst advised Frankfurter that he preferred Madison 'particularly because of this extraordinarily interesting state government machinery here on the doorstep.' Although Hurst was extremely reluctant to travel from his Wisconsin base throughout his life, he nonetheless was directly involved in launching the

international Law and Society movement, as well as in pioneering a new approach to legal history that has had worldwide impact. He taught full time at Wisconsin until 1980 and continued to publish after retiring from teaching. Hurst died in Madison on June 18, 1997.

2. *Hurst's Scholarship*

Hurst's scholarly career repeatedly underscored a core idea. In 1971, he summarized the enormous project he had undertaken: 'A realistic history of law in the United States will be a social history of law, taking law as man-made, and as the product of both deliberation and drift, but not of any immanent superior order of reality.' Doing legal history from the outside—from the perspective of sociological jurisprudence—led Hurst to turn his back on the 'great man' tradition and to reject the previously dominant focus on formal judicial opinions. He firmly opposed teleological accounts that tended to celebrate the gradual improvement of law over time. Instead, Hurst focused on economic matters—as shaped by the general culture—and on the ordinary processes of statutes, administrators, low-level cases, and the everyday work of lawyers.

In particular, Hurst concentrated on analyzing what he termed 'our prime inheritance,' which he defined as 'middle-class ways of thinking.' With exceptional tenacity, he set about tracking 'the expansion of middle-class energy.' He sought to describe and to analyze how most people in the USA responded to a central problem. That basic issue, according to Hurst, confronts 'whether men will be able to use the power of the politically organized community to secure the basic conditions of a decent life for the individual, without thereby destroying the very values they are trying to realize.'

Throughout his decades of scholarship and teaching, Hurst combined unremitting and careful attention to massive low-level details with sweeping descriptions of general patterns. He sought to develop a new integration of the social-science ideas of his era with careful historical consideration of the largely unexamined mass of law as it functioned on the ground, far removed from the doctrines promulgated by selected appellate judges that dominated—and still dominate—US legal education.

Robert Gordon described the great reach of Hurst's innovative academic project as follows:

> After expanding the relevant record beyond prior appellate cases and their facts to all prior legal acts and the social influences on them, and the pending proceedings to the overall task of designing workable legal-regulatory arrangements for the present age, Hurst wrote his books as background briefs for present-day lawmakers.

To Hurst, social patterns mattered greatly. But he perceived that generalized thoughtlessness and a strong tendency to accept inertia and drift often served to counterbalance, or even to disable, other strong, competing US commitments.

Through his prodigious output of scholarship on markets, legislation, monetary policy, and administrative actions large and small, Hurst pointedly emphasized the frequent unthinking clash between core US values. He seemed to delight in clearly describing basic and sobering paradoxes. Thus, for example, Hurst observed: 'Individuals realize their humanity only in society, but they realize their individuality only in self-awareness, which consists in some sense of separateness.' Moreover, he noted that such tensions afford protective coloration for 'shrewd men' who 'warp individualistic symbols of the constitutional ideal to sanction unchecked organized power.'

As Hurst challenged traditional legal distinctions between public and private activity, he pushed himself and those who followed him to develop both extensive local data and expansive theories to capture and explain overarching patterns. In *The Growth of American Law* (1950), Hurst argued that 'law has been more the creature than the creator of events.' This famous book emphasized the impact of physical, technological, and social facts on nineteenth-century law and paid particular attention to the often inglorious role of the legal profession. Some have criticized Hurst for being too much a functionalist, and for inadequately attending to issues of power. Yet Hurst's 946-page *Law and Economic Growth: The Legal History of the Lumber Industry in Wisconsin, 1836–1915* (1964), for example, highlighted the failures of laissez-faire. It also extensively documented Wisconsin's broad and detailed state regulation of a single industry—during a period generally thought to have been at or near the apex of free-market ideology.

3. *Hurst's Impact*

William Novak asserted in 2000 that Hurst surely ranks as 'a broad-gauged socio-legal thinker of the first order, writing in the tradition of Tocqueville, Weber, Holmes, and Pound.' Hurst was original in his particular emphasis on place and time, particularly through his focus on his native Midwest. He also innovated through his unusual illustration of the difficulty in discerning any clear patterns of progress in US legal developments. Indeed, Hurst concentrated on exploring the intricate network of constraints—some perceived, some never understood, some formally legal, some not legal at all—that repeatedly confined and constrained the freedom of everyone.

In Hurst's view, Americans competed fervently, if often blindly, to manipulate whatever system they found in place in their particular context. He generally found it necessary to bemoan the power of deep habits that often produced irrational development and sub-

stantial depredations over time. Yet Hurst also provided crisp accounts of the constitutional impulse among Pike Creek squatters, for example, as well as broad studies of such issues as the fierce, largely irrational conflict over gold and silver currency in the late nineteenth century.

Hurst sometimes has been inaccurately caricatured as part of the celebratory framework of some forms of 'consensus history.' Yet the very broad-gauged quality of his work (an ongoing quest, in Novak's words, 'to try to twist the tail of the cosmos') emphasized the mainstream at the cost of excluding, or at least understating, the importance of minorities and of dissenting viewpoints throughout US legal history. Hurst's faith in fact-gathering led him to produce not merely a new way of looking at law, but also an extended historical critique of what he called 'bastard pragmatism'—the propensity of people throughout US history to allow themselves to be dominated by the short term and the bottom line.

Through his entire career, Hurst emphasized the difficult, if not impossible, goal of increasing both public and private accountability. In *Law and the Conditions of Freedom* (1956) he noted a clear historic preference among Americans for 'the promise of the future over commitments to the past, but also for the ambitions of the present over the claims of the future.' Somewhat ironically, Hurst nevertheless remained deeply committed to the progressive 'Wisconsin idea.' To Hurst facts were crucial. Only through facts could a doubting optimist, such as Hurst himself, imagine enlightened leaders actually enhancing the general pursuit of 'lives of good conscience and fulfillment.' If Hurst sometimes seemed unabashed in adopting what Daniel Ernst called 'a presentist approach to the past,' he remained first and foremost an innovative, exceptionally influential historian.

The unusual number and breadth of symposia about Hurst's work only begin to suggest how significantly his emphasis on the impact of markets, legislation, monetary policy, social habits, and administrative actions transcended earlier boundaries. Through his enormous output of scholarship, his legendary generosity in giving scholarly assistance to others, and his decades of noteworthy teaching, Willard Hurst changed the way US legal history is written and taught. By now, as Lawrence Friedman put it, 'You're either a Hurstian or a reviser of Hurst.' Hurst not only redefined an entire academic field, but he did much to open the borders of legal history to other disciplines. Hurst's career challenges his successors to consider carefully the intricate, crucial, often paradoxical intertwining of legal ideas and social realities within the many realms in which people actually live.

See also: Law: History of its Relation to the Social Sciences; Law, Sociology of; Legal Education; Legal Scholarship

Bibliography

Diamond S 1979 Law and social order in the United States–Hurst, JW. *Michigan Law Review* **77**: 784–94
Erlanger H (ed.) 1985 Review symposium: The work of J. Willard Hurst. *American Bar Foundation Research Journal*: 113–44
Gordon R (ed.) 1975 *Law & Society Review* **10**: 5–333
Hartog H 1994 Snakes in Ireland: A conversation with Willard Hurst. *Law and History Review* **12**: 370–90
Scheiber H 1970 At the borderland of law and economic history: The contributions of Willard Hurst. *American Historical Review* **75**: 744–56
Soifer A 1992 In retrospect: Willard Hurst, consensus history, and the growth of American law. *Review of American History* **20**: 124–44
Tomlins C (ed.) 2000 Engaging Willard Hurst: A symposium. *Law and History Review* **18**: 7–221
Tushnet M 1972 Commentary: Lumber and the legal process. *Wisconsin Law Review* 114–32
1980 *Wisconsin Law Review*: 1090–442
1997 *Wisconsin Law Review*: 1122–210, Hurst bibliography: 1205–10

A. Soifer

Husserl, Edmund (1859–1938)

Edmund Husserl was one of the most influential philosophers of the twentieth century. Inventor of the 'phenomenological method' for the analysis of physical and mental phenomena, initiator of the 'phenomenological movement' of contemporary human sciences, he became the great critic of naive naturalism in philosophy and psychology ('psychologism'), and of what he regarded as the loss of orientation, certainty, and thereby of credibility, of modern science. His appeal for 'going back to the things themselves,' and for a foundation of all science and philosophy in a clearly understood 'life-world' (*Lebenswelt*), inspired the work of a great number of philosophers (e.g., Gadamer, Heidegger, Ingarden, Levinas, Merleau-Ponty, Ricoeur, Sartre) and social scientists (e.g., Buytendijk, Lewin, Schütz).

1. Life and Academic Career

Edmund Gustav Albrecht Husserl was born in Prossnitz, Moravia, on April 8, 1859. He studied mathematics, physics, astronomy and philosophy at the universities of Leipzig (1876–1878), Berlin (1878–1881), and Vienna (1881–1882), where he obtained his Dr.Phil degree in mathematics with Leo Königsberger in 1883. After a short period as assistant

to the mathematician Karl Weierstrass in Berlin, and a year of military service, Husserl continued his philosophical studies with Franz Brentano in Vienna from 1884 to 1886, and with Carl Stumpf in Halle between 1886 and 1887, where he qualified as lecturer in 1887 with a postdoctoral thesis entitled 'Ueber den Begriff der Zahl. Psychologische Analysen' ('On the Concept of Number. Psychological Analyses').

Husserl became a professor in 1894, and moved to Göttingen in 1901. In 1916, he accepted a call to Freiburg to succeed Heinrich Rickert in the chair of philosophy there. Husserl retired in 1929. After the seizure of power in Germany by the National Socialists in 1933, his status of emeritus professor was suspended because of his Jewish descent, and, after a short reinstatement, definitively revoked in 1936 after the Nuremberg race laws. Husserl died in Freiburg on April 27, 1938.

2. *Writings, Subjects, Approaches, and Contexts*

Husserl's significance for philosophy, which is certainly greater than that for the social and behavioral sciences, rests on his establishment of a new way of dealing with philosophical problems, which he called 'phenomenology,' turning this expression into a 'technical' term by giving it a clearer delimitation and a weightier role than it had enjoyed with earlier philosophers such as Lambert or Hegel. The methods and content of Husserl's phenomenology, however, changed considerably during their development, which may be summarized in terms of three 'turns' of direction and aims: (a) the turn from early foundational work to anti-psychologism, (b) the shift from descriptive to 'constitutive' phenomenology, and (c) coping with the 'crisis' of modern science by a new focus on the life-world.

2.1 *From Early Foundational Work to Anti-psychologism*

In *Philosophie der Arithmetik* (*Philosophy of Arithmetic*, 1891, Husserl 1950, Vol. XII, pp. 1–283), Husserl made an attempt at a descriptive—psychological foundation of arithmetic, presumably having in mind an extension of Weierstrass's program of basing the vast number of mathematical concepts on a small number of fundamental ones, by providing an even deeper foundation for the latter by grounding them in the psychical acts underlying the mathematician's work. In the course of these investigations, Husserl introduced (independently of C. von Ehrenfels, who is normally credited with this discovery) 'gestalt qualities,' which he treated under the name *figurale Momente*.

Less successful (with hindsight) was Husserl's criticism of G. Frege's definition of number. It invoked an objective but sharp counter-critique by Frege, which seems to have persuaded Husserl of the impossibility of taking the special science of psychology (which proceeds inductively, falling short of universally valid statements), as a foundation for mathematics or even for all special sciences ('psychologism'). Therefore, in *Logische Untersuchungen* (1901), aiming at a foundation of logic which at that time was usually set up on a psychological basis, Husserl advocated a consequent anti-psychologism which influenced contemporary philosophy profoundly. This concept was extended in his work 'Philosophie als strenge Wissenschaft' ('Philosophy as a Rigorous Science') (Husserl 1950, Vol. XXV, pp. 3–62) to a critique of historicism and (under the label of 'naturalism') scientism (1910–1911).

The positive foundation of logic was to be laid by a new foundational discipline to which Husserl, oddly enough, gave the name 'descriptive psychology,' despite his claim that it differed from empirical psychology as well as from Brentano's descriptive classification of psychical (mental) phenomena; it was to be 'descriptive' in the sense of presenting the immediately given mental facts without presuppositions (*voraussetzungslos*). According to this earliest form of phenomenology, that which is intended in a mental act (perception, judgment, idea, etc.) is transcendent to the act itself and given only in 'adumbrated' (*abgeschattet*) manner, through changing phenomena, and therefore inadequately. By contrast, everything immanent to the act is free of adumbration and can therefore be regarded as given in an absolutely adequate manner. Moreover, intuition of what is immanently given appears as free of structural complications like those attributed to judgment in contradistinction to sensation. The contents of the sphere of immediate experience (*Erlebnisse*) are beyond every doubt; with 'intentionality' at its center, it is the secure ground of a foundational discipline which is capable of justifying statements about essences (*Wesen*) and their relations, by a method which Husserl first called 'eidetics,' but later called 'intuition of essences' (*Wesensschau*).

Particular relevance to a philosophy of the sciences may be claimed for Husserl's remarks on 'pure logic.' Pure logic is an *a priori* theory of science dealing with two kinds of fundamental categories: categories of meaning (concept, proposition, truth, inference, etc.) on the one hand, and categories of objects (object, state of affairs, connection, number, etc.) on the other. Husserl takes the unity of the propositions of a science (or a discipline) to consist in the unity of its explanatory principles, and finds his next task in the uncovering of the conditions of the very possibility of theories. The route is to proceed from preliminary studies, e.g., a 'purely logical grammar' investigating the components and connections of propositions, via the statement of laws obtaining between such categories, to an *a priori* 'theory of possible theory forms.'

These investigations, continued only in *Formale und transzendentale Logik* (*Formal and Transcendental Logic*, 1929) and in *Erfahrung und Urteil* (Husserl 1939), share a methodical aim: the 'fixation of unambiguous, sharply distinct verbal meanings' (Husserl 1901, Vol. I, p. 244; Husserl 1973 (English translation 1973), Vol. I, p. 238).

This search for 'verbal meanings' is also one of the foremost goals of contemporary methodical philosophy and philosophy of science. Husserl, however, did not approach the joint methodical aim by actually constructing a language of science, but recommended an 'intuitive presentation (*Vergegenwärtigung*) of essence in adequate ideation,' obviously conceived as independent of and antecedent to any linguistic construction.The ideation to be performed is not an idealization (as in geometry), but pure description, aiming at the elevation of that which is inexact and essentially vague to something typical, a 'type.' By setting out in this way, phenomenology increasingly moved away from the other methodical currents of modern philosophy and came closer to the philosophy of life (*Lebensphilosophie*), from which, however, Husserl differs sharply methodically by ascribing a foundational value to experience, after its precise rendering as 'pure experience' (*reines Erlebnis*) in which meaning, anterior to everything real, is believed to constitute itself (philosophical intuitionism).

2.2 The Step from Descriptive to 'Constitutive' Phenomenology

Object and method of phenomenology were extended and partly revised in *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie* Husserl (1913), and pure logic was also characterized in a new way. In Husserl 1901, the predicate 'ontological' had been restricted (diverging from tradition) to sciences such as geography, astronomy, and anatomy, which deal with concrete objects, and not with abstract objects (like the theoretical or nomological sciences). Now, the old term 'ontology' is taken up again in a new sense. Husserl takes being to be subdivided into highest kinds or 'regions,' so the sciences which deal with them yield 'regional ontologies.' Above these (material) ontologies we find the formal sciences, mathematics and logic, which are aimed at 'form essences' (*Formwesen*) and give rise to a *mathesis universalis*. Treating the 'empty forms of region' as such, they comprise the forms of all possible material ontologies, and are nothing else but a 'formal ontology.'

In *Formale und transzendentale Logik*, formal logic was subdivided into (a) a pure doctrine of forms of meaning, or logical grammar, (b) the logic of consequence, or logic of non-contradiction, and (c) the logic of truth. The purely formal ontology was characterized more clearly as the logic of objects by severing it from the 'apophantic' logic of judgment, while the nomological sciences are deductive, formal–analytic sciences describing 'definite manifolds.' Fundamentally new is Husserl's assessment of ontology as a naive–dogmatic discipline that is turned into a science only by its founding on a 'transcendental logic.' The term 'transcendental,' as explained in 1936, refers to 'the motif of inquiring back into the ultimate source of all the formations of knowledge, the motif of the knower's reflecting upon himself and his knowing life' (Husserl 1954, p. 100). Husserl's reflective 'pure consciousness' of 1913 is now determined further by its function as the 'transcendental ego,' having everything as its object, while never becoming an object itself, and granting access only through 'eidetic intuition.'

The absolute sphere, which is the foundation of ontology and may be distinguished from the ontological sphere as 'semasiological' (R. Zocher), is the object of a 'critical' transcendental philosophy, which Husserl can only conceive of as a fully developed phenomenology. It makes use of the grasping of essences as a tool for analyses which approach the problem of foundation not as a methodological task, but as a problem of constitution in the classical sense: in this sphere, according to Husserl, valid statements may be obtained only by a sequence of increasingly radical reduction steps, each of which is to consist in a certain 'abstention' (*epoché*). The first two steps, i.e., 'phenomenological reduction,' leading to the phenomena, and 'eidetic reduction,' leading to the essential structure of consciousness, carry us on to the transcendental sphere. Access to it, however, is only given by 'transcendental reduction,' which is the 'absolutely universal reduction,' opening up the transcendental subjectivity which constitutes the meaning of the world and its contents, and thereby all objecthood.

2.3 The Crisis of European Science and the Lifeworld

The period between World War I and the seizure of power by National Socialism in Germany abounds with diagnoses of crises, from the foundational crises in most sciences (claimed for psychology, sociology, economics, etc., following the foundational crisis of classical mathematics so dramatically exposed by H. Weyl in the 1920s) to a general crisis of the times, of intellectual and moral orientation, and of European ('rational') culture generally.

Husserl gave much thought to this phenomenon and made out a growing loss of confidence in the sciences hitherto regarded as the firm foundation of Europe's rational culture. He analyzed this situation comprehensively in an article 'Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie' published in 1936 in a new and little-

known Serbian journal and later forming the first two parts of the monograph bearing the same title (Husserl 1954). The sciences, it seemed to him, while 'functioning' as well as ever technically, had lost sight of their origin as well as of their aims, and thereby lost their former role and status as a guide and a standard of reliable knowledge and moral responsibility. Husserl hoped for a renewal by recognition of the fact that all scientific knowledge and activity is grounded in and grows out of everyday life. This is a realm usually not reflected upon but capable of thorough analysis as our 'life-world,' the domain not only of the physically 'given,' but primarily of our social life. Our immediate experience is not that of objects given to our consciousness, but that of situations in which we find ourselves with other people, in social groups and institutions, participating in common meaningful action, handling (linguistic and other) symbols, and approaching with understanding even 'the great symbolic systems such as language, myth, religion, art, etc.; all of which are essential elements of the *Lebenswelt* and therewith of highest interest to the social sciences' (Schütz 1959, p. 96). The phenomenological analysis of these largely 'pre-predicative' constituents of the life-world was envisaged by Husserl in the *Crisis* treatise and begun in his late and posthumously published work *Erfahrung und Urteil* (1939), where he has recourse to the life-world and to the primacy of praxis over mere contemplation. But in front of them, he places the 'pre-predicative' evidence of perception and of judgment based on perception, and thereby once more a prelinguistic foundational domain. The new approach remained to be developed by others.

3. *Contexts of Discussion and of Influence*

Notwithstanding the fact that 'Husserl literally "thought through his pen"' (Embree et al. 1997, p. 327), entrusting an incredibly large part of his ideas to shorthand memos that remained unpublished during his lifetime (and part of them until now), his thinking did not develop monologically, without context. During Husserl's student years, the careful study of Descartes, Leibniz, the British empiricists and Brentano was recommended to him by T. Masaryk (today less famous for his *Versuch einer concreten Logik* of 1887 than for his later career as the first president of Czechoslovakia), and Husserl himself testifies to the influence of William James's *Principles of Psychology* of 1890 on the shaping of his phenomenological thought. Having acquired some fame as a creative philosopher, Husserl corresponded with many experts in science and philosophy, and discussed his ideas at invited lectures at other German universities, and abroad in Vienna, Prague, Paris, Amsterdam, London and elsewhere—events certainly prepared by the translation of the *Logische Untersuchungen* and the *Ideen* into several foreign languages.

The beginning of the 'phenomenological movement' may be seen in the exchange of ideas with students and collaborators of T. Lipps in Munich, who learned about Husserl's investigations through their fellow student J. Daubert after he met Husserl in Göttingen in 1902. In spite of having nothing in print, Daubert had considerable influence on the 'Munich phenomenologists' (among them M. Geiger, A. Pfänder, and A. Reinach), some of whom (like H. Conrad-Martius, A. Koyré, and E. Stein, later one of Husserl's private assistants) came to Göttingen to work with Husserl. They became the leading members of the 'Göttinger philosophische Gesellschaft' and main authors of the *Jahrbuch für Philosophie und phänomenologische Forschung*, founded in 1913, which published some of the great contributions (phenomenological and other) to contemporary philosophy, such as Husserl's *Ideen* in 1913, M. Scheler's *Der Formalismus in der Ethik und die materiale Wertethik* (*Formalism in Ethics and the Material Ethics of Value*) in 1913–16, and M. Heidegger's *Sein und Zeit* and O. Becker's *Mathematische Existenz* (*Mathematical Existence*), both in 1927. After its discontinuation in 1930, the publications' goals and role were taken up by the American journal *Philosophy and Phenomenological Research*, founded in 1940.

In Husserl's most active period, private assistants helped organize his literary production, defended his positions (e.g. E. Fink against neo-Kantian critics), and prepared lectures and notes for publication (e.g., the *Zur Phänomenologie des inneren Zeitbewußtseins*, prepared by E. Stein and edited by M. Heidegger in 1928, and, posthumously in 1939, *Erfahrung und Urteil* by L. Landgrebe). In 1940, H. L. Van Breda saved most of Husserl's *Nachlass* (including 40,000 pages of manuscripts, Husserl's private library and much of his correspondence), which today forms the core of the Husserl Archives at the Catholic University at Louvain, Belgium. Philosophical and related work in the phenomenological spirit is, beyond this, upheld by 'Phenomenological Societies' in several countries.

4. *Impact on the Social and Behavioral Sciences*

Husserl's own writings did not take up any topics directly relevant to the social and behavioral sciences and therefore had little influence on them during his lifetime. Yet, they greatly stimulated interest in the phenomenological method in other authors, who explored its applicability and fruitfulness within the social sciences, broadening and strengthening the phenomenological movement. Whereas the investigations of A. Reinach into the *a priori* foundations of civil law, and of M. Scheler into the origins of cultural development, are contributions to the philosophy of

the social sciences rather than to social sciences proper, the potentiality of a phenomenological approach was recognized early in the psychology of personality (especially in the psychology of the subconscious and in psychiatry), and in experimental psychology, where it opened up new vistas for social psychology and the psychology of perception (D. Katz, A. Metzger, M. Wertheimer): the development of gestalt psychology (K. Köhler, K. Koffka) and of K. Lewin's field theory are unthinkable without the impact of Husserl's phenomenology (see also MacLeod 1947). In the field of sociology, the *formal sociology* (or sociological formalism) of A. Vierkandt and G. Gurvitch have drawn explicitly on phenomenological ideas, and more recently, sociological systems theory, social constructionism and ethnomethodology have been clearly influenced by phenomenological thought (P. Berger, H. Garfinkel, E. Goffman, T. Luckmann, Niklas Luhmann, I. Srubar).

According to A. Schütz, 'Husserl's signal contribution to the social sciences consists ... in the wealth of his analyses pertinent to problems of the *Lebenswelt* and designed to be developed into a philosophical anthropology' (Schütz 1959, p. 97). This development was brought about by Schütz himself, who met Husserl personally only in 1932, after publication of his own pioneering work *Der sinnhafte Aufbau der sozialen Welt* (*The Phenomenology of the Social World*, 1972), but established close intellectual contacts for the six remaining years of Husserl's life. Schütz became the towering figure of phenomenological sociology, and it is due to the considerable impact of his work that Husserl's phenomenology, beyond its historical merit of exhibiting the limits of naturalistic and atomistic thinking, has gained a lasting influence on the social and behavioral sciences of today.

See also: Hermeneutics, History of; Phenomenology in Human Science; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Philosophy, Psychiatry, and Psychology; Psychologism, History of; Reduction, Varieties of; Religion, Phenomenology of; Schütz, Alfred (1899–1959)

Bibliography

Bell D 1990 *Husserl*. Routledge, London

Embree L et al. 1997 *Encyclopedia of Phenomenology*. Kluwer, Dordrecht, The Netherlands

Herzog M, Graumann C F (eds.) 1991 *Sinn und Erfahrung. Phänomenologische Methoden in den Humanwissenschaften*. Asanger, Heidelberg, Germany

Husserl E 1901 *Logische Untersuchungen*. Niemeyer, Halle, Germany (trans. into English, 1970 *Logical Investigations*. Routledge & Kegan Paul, London, and Humanities Press, New York)

Husserl E 1913 (rev. 1976), 1950, 1952 *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*. Vol. 1 Niemeyer, Halle, Germany; Vol. 2, Biemel W (ed.), Nijhoff, The Hague, The Netherlands; Vol. 3, Biemel M (ed.) Nijhoff, The Hague, The Netherlands (trans. into English 1980–89 *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy*. Vol. 1: Nijhoff, Holland, Vols. 2 and 3: Kluwer, Dordrecht, The Netherlands)

Husserl E 1939 *Erfahrung und Urteil. Untersuchungen zur Genealogie der Logik*. Academia, Prague (trans. into English 1973 *Experience and Judgment. Investigations in a Genealogy of Logic*. Northwestern University Press, Evanston, IL, and Routledge & Kegan Paul, London)

Husserl E 1954 *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: Eine Einleitung in die phänomenologische Philosophie*. Nijhoff, The Netherlands (trans. into English 1970 *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*. Northwestern University Press, Evanston, IL.) (Pts. 1 and 2 had already been published in the journal *Philosophia* (Belgrade) **1** (1936), 77–176)

Husserl E 1950 ff. *Husserliana: Edmund Husserl, Gesammelte Werke*. Nijhoff, The Netherlands, and Kluwer, Dordrecht, The Netherlands

MacLeod R B 1947 The phenomenological approach to social psychology. *Psychological Review* **54**: 193–210

Schütz A 1959 Husserl's importance for the social sciences. In: Van Breda H L, Taminiaux J (eds.) *Edmund Husserl 1859–1959: Recueil Commémoratif publié à l'Occasion du Centenaire de la Naissance du Philosophe*. La Haye, Nijhoff, The Netherlands, pp. 86–98

Schütz A 1962–97 *Collected Papers*. Nijhoff, The Netherlands (Vol. I: *The Problem of Social Reality*, ed. Natanson M 1962; Vol. II: Studies in Social Theory, ed. Brodersen A 1967; Vol. III: *Studies in Phenomenological Theory*, ed. Schütz I 1966)

Smith B, Smith W D (eds.) 1995 *The Cambridge Companion to Husserl*. Cambridge University Press, Cambridge, UK

Wolf K H 1979 Phenomenology and sociology. In: Bottomore T, Nisbet R (eds.) *A History of Sociological Analysis*. Heinemann, London, pp. 499–556

C. Thiel

Hutterite Demography

The unique demography of the Hutterites—a contemporary Mennonite sect of Germanic origin who live in the Great Plains of the USA and Canada—has played an important role in the development of demographic theory and methods. They practice no birth control for religious reasons and their high fertility levels give reason to believe they do not intentionally control their fertility (Eaton and Mayer 1953). Their communally shared life-style, including the sharing of material resources, is also presumed to limit social or cultural variations in Hutterite fertility (Larsen and Vaupel 1993). Because of their uniformly high and apparently uncontrolled fertility, studies of the patterns and determinants of fertility among the Hutterites have helped to shape our theoretical expectations for populations that do not make any conscious efforts to limit childbearing. And, as a result,

the Hutterites have also been widely used in demographic methods as a fertility standard to which other populations are compared.

1. Hutterite Fertility

Many different estimates of fertility and fertility determinants among the Hutterites have been published in different studies (see Larsen and Vaupel 1993). All of these studies agree, however, that Hutterite fertility is exceptionally high. Eaton and Mayer (1953) reported a median completed family size of 10.4 children among married women over the age of 45 in 1950. Recalculations using the data published by Eaton and Mayer report a lower completed family size with a median of 9.4 children, and a mean of 8.97 among the same women (Lang and Göhlen 1985). These estimates are consistent with the modal completed family size of 10 children, and the maximum of 15 children, reported for the Hutterite sample analyzed by Larsen and Vaupel (1993). Even in this very high fertility population, however, fecundability and the probability of conception decline substantially with age. The monthly probability of conception is estimated to decline by about two-thirds between ages 20 and 35 (Larsen and Vaupel 1993), likely reflecting age-related influences such as higher probabilities of fetal loss, lower coital frequency, longer postpartum amenorrhea, etc., among older age women. Nonetheless, the age-specific influences on Hutterite fertility resulting in a decline of fecundability across ages appear to be largely biological in nature. Compared with other populations, the Hutterite population has exceptionally high fertility levels and the age-related decline in fecundability offers no apparent evidence of intentional fertility limitation.

2. The Hutterites as a Natural Fertility Standard

Because of their doctrines and behavior, the Hutterites are presumed to represent a *natural fertility* population. That is, as Henri (1961) argued, their childbearing patterns reflect no efforts to limit fertility bound to the number of children already born, i.e., parity-*specific* fertility limitation. In fact, the Hutterites were one of the populations Henri (1961) used to define and illustrate the concept of natural fertility. He also ensured a special standing for the Hutterites as a high fertility standard when he found that they had the highest age-specific fertility levels of 13 historical populations he studied, all of which appeared to be natural fertility populations.

2.1 The Hutterites as a 'Standard Natural Fertility Population'—Coale's 'I' Indices

Hutterite fertility levels and age-specific fertility patterns have been used as standards in some of the most widely used fertility measures of recent years. The Princeton European Fertility Project (Coale and Watkins 1986) popularized the use of Coale's (1969) 'I' indices which compare fertility in a population with that of the Hutterites through indirect standardization. Coale's index of overall fertility (I_f) compares the number of births to women in a population with the number of births these women would have had if they had experienced the same high age-specific fertility rates as married Hutterite women (i.e., 0.550 births per year among women aged 20–24, 0.502 to those aged 25–39, 0.447 to those aged 30–34, 0.406 to those aged 35–39, 0.222 to those aged 40–44 and 0.061 to those aged 45–59, see Henri 1961). Coale's index of marital fertility (I_g) similarly compares births to married women with the fertility they would have had if they had experienced the Hutterite age-specific marital fertility rates. And, Coale's index of illegitimate fertility (I_h) draws the same comparison for nonmarried women, again using Hutterite age-specific marital fertility rates as a standard. A final fourth index is intended to capture effects of the proportion married upon fertility (I_m) through comparing the fertility married women would have with that which all women would have, if both experienced Hutterite age-specific marital fertility rates.

2.2 Age Patterns of Hutterite Fertility as a Standard—Coale and Trussell's 'M' and 'm' Indices

Coale and Trussell's (1978) more formal mathematical modeling of age-specific marital fertility control also incorporates the fertility pattern of the Hutterites and Henri's (1961) other natural fertility populations as a standard for comparison. Coale and Trussell used 10 of Henri's natural fertility populations, including the Hutterites, to specify a standard age-specific pattern for natural fertility (see *Demographic Models*). The Coale–Trussell model attempts to parameterize the differences in overall levels and age patterns of fertility from these natural fertility expectations in populations with varying levels of fertility control. The level of fertility is then captured by two parameters of the model, Coale and Trussell's '*M*' index of fertility levels and their '*m*' index indicating departures from a natural fertility age pattern (1978).

2.3 Hutterite Fertility Determinants in Micro- and Macrosimulations

Not surprisingly, demographic simulation models have also made use of Hutterite fertility standards. Ridley and Sheps' (1966) initial (i.e., Cohort A) microsimulation of fertility, for example, used values of fecundability and postpartum nonsusceptibility which were derived from the Hutterite population. Bongaarts' more recent (1976) macrosimulation of fertility also incorporated Henri's (1961) natural fertility estimates as standards for age-specific patterns of

fecundity and for infecundity associated with sterility. In these simulations the Hutterite fertility standards are used to represent biological fertility parameters in the absence of fertility control.

3. Limitations of the Hutterite Fertility Standard

As the theoretical definition of natural fertility has been challenged, so have the measures of fertility control which have utilized the Hutterite fertility standard. Comparisons with a Hutterite standard population, such as Coale's I_g index, may reflect differences in fertility which do not arise from parity specific control (Santow 1995). And, as in the case of Coale's '*M*' and '*m*' indices, such comparisons may also fail to capture the heterogeneity of fertility behavior in populations being compared with the natural fertility standards (Okun 1994), including the possibility of fertility control which is not closely tied to parity such as birth spacing. Although comparative indices and methods face many of the same limitations as any indirect standardization, incorporating the Hutterites as a constant standard population for empirically-based comparisons has nonetheless been of considerable advantage to demographers. Use of the Hutterite standard has allowed meaningful comparisons of fertility levels and patterns across widely different historical populations. The Hutterite standard has also provided indices with a clearly interpretable and unambiguous reference to an empirically observed high fertility standard. And most criticisms of the measures incorporating the Hutterite standard concern their interpretation in terms of underlying fertility behavior (e.g., diffuse birth spacing or parity-specific stopping) rather than their ability to index the levels and patterns of age-specific fertility across populations.

4. The Hutterites and Bio-demography

Underlying their use as a high-fertility standard in many different methods, the Hutterites in and of themselves offer a unique opportunity for demographers to explore biological aspects of fertility and fecundity. An absence of any evident parity-specific fertility control, along with very high fertility levels, makes the Hutterite population one which may be used to study biological determinants of fertility without the confounding influence of fertility limiting behaviors. However, a troubling concern in the study of historical populations that lack widespread evidence of fertility control is the extent to which there may be many subgroups within the population with different fertility levels and patterns arising from different behavior and sociocultural circumstances. The Hutterites' communal lifestyle and uniformly high fertility give no reason to suspect wide differences in fecundability and fertility from social or cultural forces. And the uniformity of the population in many demographic behaviors (e.g., universal marriage and early ages at marriage, short durations of breast-feeding and early weaning, short birth intervals, absence of induced abortion, and low levels of premarital conception, etc.) is likely to minimize many sources of variation in fertility compared with other populations. This relative homogeneity, or absence of subgroup variation, is especially important to the recent use (and assumptions) of event history methods to estimate the underlying biological patterns of fecundability among the Hutterites. This allows researchers estimating fecundability with Hutterite data to assume the underlying patterns and variations they discover arise from biological variation (see Larsen and Vaupel 1993 for a review and example of such studies).

As further studies provide more refined estimates of Hutterite fertility and fecundity, these will likely continue to serve as empirically grounded standards for our demographic expectations in the absence of intentional fertility limitation. The theoretical importance of these expectations to demography is evident in the significant role which studies of the Hutterites have had in shaping the concept of natural fertility and debates over the nature of fertility limitation. Their methodological import is clear in the wide variety of methods already grounded in the Hutterite fertility standard.

See also: Demographic Data Regimes; Demographic Models; Demography, History of; Fertility and Culture: Anthropological Insights; Fertility Control: Overview; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations

Bibliography

Bongaarts J 1976 A dynamic model of the reproduction process. *Population Studies* **30**(1): 59–73

Coale A J 1969 The decline of fertility in Europe from the French Revolution to World War II. In: Behrman S J, et al. (eds.) *Fertility and Family Planning: A World View*. University of Michigan Press, Ann Arbor, MI

Coale A J, Trussell T J 1978 Technical note: Finding the two parameters that specify a model schedule of marital fertility. *Population Index* **44**(2): 203–13

Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ

Eaton J W, Mayer A J 1953 The social biology of very high fertility among the Hutterites: The demography of a unique population. *Human Biology* **25**: 206–64

Henri L 1961 Some data on natural fertility. *Eugenics Quarterly* **8**: 81–91

Lang H, Göhlen R 1985 Completed fertility of the Hutterites: A revision. *Current Anthropology* **26**(3): 395

Larsen U, Vaupel J W 1993 Hutterite fecundability by age and parity: Strategies for frailty modeling of event histories. *Demography* **30**(1): 81–102

Okun B 1994 Evaluating methods for detecting fertility control: Coale and Trussell's model and cohort parity analysis. *Population Studies* **48**(1): 193–222
Ridley J C, Sheps M C 1966 An analytic simulation model of human reproduction with demographic and biological components. *Population Studies* **19**(3): 297–310
Santow G 1995 Coitus interruptus and the control of natural fertility. *Population Studies* **49**(1): 19–43

D. L. Anderton

Hybridity

'Hybridity,' originally a biological term, has been used by authors in the social sciences, literary, artistic, and cultural studies to designate processes in which discrete social practices or structures, that existed in separate ways, combine to generate new structures, objects, and practices in which the preceding elements mix. During the 1980s and 1990s its use has increased, particularly in English and Spanish, whereas in other languages, such as French, the term coexists with adjacent concepts like *mestizaje* and syncretism, favored in French literature. Some authors include these classic terms in 'hybridization,' conveying greater flexibility to the term to include processes of intercultural confluence.

1. *The Term Hybridity in Biology and in Social Sciences*

The use of hybridity in humanistic and social science studies has been questioned due to the biological origin of the concept. The risk of transferring the sterility commonly associated with the term to society and culture is put forth. Even though this criticism can still be found in recent writings, it is the extension of a nineteenth-century belief, when hybridization was looked upon with suspicion because it was believed to harm social development. However, this opinion has changed since 1870 with Mendel's theory of the enhancement produced by genetic mixing and its twentieth-century examination (Olby 1984). Strictly considered, biology, particularly botany, teems with fertile hybridizations which take advantage of characteristics of cells from different plants to enrich the growth, resistance, quality, and economic and nutritional worth of food stuffs made from these combinations (Callender 1988). Hybridization of human DNA with bacteria to produce protein is also common; this is, in fact, the main means to generate insulin.

Adoption of the term by the social sciences is not legitimized by its biological background. Similarly to other concepts taken from biology, for example, *reproduction*, for which a new elaboration had to take place in order to talk about social, economic, and cultural reproduction, the theoretical solidity and explanatory power of hybridity in social sciences depends upon the manner in which the term is relocated within networks of specific concepts referring to society and culture.

On the one hand, hybridity is linked with more or less classical notions of social thought that allude to the separation and mix among cultures and strata: race, ethnicity, *mestizaje*, miscegenation, syncretism, identity, multiculturalism, heterogeneity, creolization, transculturation, bricolage. On the other, its treatment and epistemological consistency depends upon the manner in which different authors link the concept with terms and theories, such as tradition, modernity, postmodernity, border, contradiction, globalization, that condition the understanding of hybridity.

The sociocultural construction of hybridity has taken place in several disciplines. As a linguistic process it was elaborated, among others, by Mikhail Baktin, who used the term to characterize the coexistence of learned and popular speech in the same goup, or even in a single speaker, and by Homi Bhabha, who defines hybridity as a metonymy of presence and places it amidst power relations, noting differences between colonial and resistance hybridization, depending on whether the linguistic and literary mixes serve domination or subversion (Bhabha 1994). Some historians of the Americas speak of hybridity to explain the interweaving of pre-Columbian imagery with the colonizers' Spanish one and with the culture developed later, during modernity (Gruzinski 1999).

2. Mestizaje, *Syncretism, Transculturation, Creolization, and Hybridization*

Mestizaje and hybridization are as old as exchanges among peoples, migrations, and wars of conquest wars. The Tower of Babel is a story of *mestizaje*. Pliny the Elder spoke of hybridization when referring to the migrants that arrived in Rome in his day. Several historians and anthropologists have illustrated the decisive role of *mestizaje* in certain areas, such as the Mediterranean, since the days of Classical Greece (Laplantine and Nouss 1997). But the more intense—or researched—periods of interethnic and intercultural mixing happened with Europe's expansion into America.

The mix of Spanish and Portuguese colonizers first, then of English and French, with American Indians, to which the African slaves were added, made *mestizaje* the prevailing process within the societies of the so-called New World. Nowadays, less than 10 per cent of the population of Latin America is descended from

indigenous peoples alone (this percentage decreases in the US and Canada). Communities of European descent that have not mixed with native peoples are also a minority. Thus, the makeup of all of the Americas requires the notion of *mestizaje*, both biologically—production of phenotypes from genetic crossbreeding—and culturally: blending of European and native American habits, beliefs, and ways of thinking.

For a very long time, studies on physiognomic and chromatic aspects of *mestizaje* were privileged. Skin color and physical features are still critical in the ideological construction of subordination and discrimination; however, in the social sciences and in more democratic political thought, *mestizaje* presently centers on the cultural dimension of the combination of identities. In anthropology, cultural studies, and cultural policy making, the issue is presented as devising ways to share multicultural life styles (Kymlicka 1995).

Of the diverse crossings that take place among cultures, the religious ones have received the most attention. They are usually referred to by a specialized term: syncretism. European expansion in Africa, Latin America, and Asia gave way to multiple mixed forms where heterogeneous beliefs fused and new divinities emerged (the Virgin of Guadalupe in Mexico; the combination of African ancestors, indigenous figures, and Catholic saints in Brazilian Umbanda). Practically every religion has fundamentalist currents that seek to defend pure forms, but the magnitude of migration and the transcontinental spread of rituals and beliefs during the eighteenth and nineteenth centuries stressed these hybridizations and increased tolerance towards them to such a point that in countries such as Brazil, Cuba, Haiti, and the US, it was frequent to find a double or triple religious affiliation. An example may be a Catholic person participating in an Afro-American ritual or in a New Age ceremony. If we consider syncretism in a wider sense, such as the simultaneous adhesion to several belief systems, not only religious ones, the phenomenon grows remarkably, particularly among the many people who resort to indigenous or Asian medicine for certain ailments, to allopathic medicine for others, and even to Catholic or Pentecostal rituals. The syncretistic use of these health resources usually is combined with musical fusion and multicultural social organization systems, as is the case in Cuban *Santeria*, Haitian Voodoo, and Brazilian *Candomblé* (Rowe and Schelling 1991).

Some of these phenomena were what Fernando Ortiz, Cuban historian and anthropologist, thought about when he proposed the term *transculturation* (Ortiz 1995/1947) to designate 'processes in which features are not mechanically aggregated, but give way to a new phenomenon, independent and original.' Bronislav Malinowski's endorsement of that utterance in an introduction to Ortiz's book, and its future development by literary critics Angel Rama and Mary Pratt in their study of travel writing, are still appealing; particularly in literature and in Anglo-Saxon and Latin-American cultural studies.

The word creolization has also been used to refer to intercultural combinations, especially in anthropology and linguistics. Rigourously, it designates the language and culture created by variations of the basic tongue with other languages in the context of the slave trade. It includes the mixtures French has undergone in America and the Caribbean (Louisiana, Haiti, Guadeloupe, Martinique) and in the Indian Ocean (Reunion, Mauritius), or Portuguese in Africa (Guinea, Cape Verde), in the Caribbean (Curaçao) and Asia (India, Sri Lanka). But whereas it presents paradigmatic tensions between orality and writing, learned and popular sectors, center and periphery, in a diversity *continuum*, Hannerz suggests its usage could be extended to the transnational field to name 'processes of cultural confluence' characterized 'by power, prestige, and material resources inequality' (Hannerz 1997). Even though he is not the only author to stress the inequality and discontinuity present in hybridizations, the emphasis he places on examining increasing flows between center and periphery, together with the asymmetries among markets, states, and educational levels, helps avoid considering *mestizaje* as simple intercultural homogenization and reconciliation (Hannerz 1996).

3. *Hybridization, Modernity, and Globalization*

3.1 *Hybridization*

Whereas the terms mentioned above are nearly used as synonyms by several authors, *mestizaje* takes precedence in studies devoted to the mix of races or ethnic groups, and syncretism to refer to religious fusion or traditional symbolic movements. Besides the fact that *mestizo* has no English equivalent (some authors include the Spanish word in their English writing), research carried out during the 1980s in Spanish, English, and occasionally in French and Portuguese adopted 'hybridity' to include diverse intercultural mixes: ethnic *mestizaje*, syncretism regarding beliefs, and other modern fusion in which arts and crafts integrate with industry, the cultured with the popular and traditional and with mass media messages (Bhabha 1994, García-Canclini 1995, Bernd and de Grandis 1995, Papastergiadis 1997, Webner and Modood 1997).

3.2 *Modernity*

Modernity, an expression of which are nation states, has coherently organized—sometimes in reality, others in appearance—the diversity of cultures. There

are fewer than 200 nation states which comprise more than 600 languages and some 5,000 ethnic groups. These states have had to organize multiculturality through diverse policies, at the same time that they facilitated communication among ethnic groups and the different languages included in their territory. Even in the cases in which a single language was imposed over the 20, 50, or 100 groups that formed the nation, that common language adopts diverse modulations in the same English- or Spanish-speaking country when they combine or hybridize with the words, syntactic structures, and semantic references of the speakers of other tongues.

In the meantime, two other characteristics of modern development, transnational expansion of capitalism and of mass communication technologies, advance relationships and exchanges between local traditions and cultures of modernity. The crafts and celebrations of peasant groups travel, together with their producers, to cities and often to other countries, or they enter into contact with foreign tourists. Ethnic music, linked to specific rituals of a certain group, is not only taped in CDs, movies, and videos to be distributed worldwide; it is also mixed with classical and contemporary music, with jazz, rock, salsa, and other local and regional kinds of music. This is how the chicha, a mix of Andean and Caribbean rhythms was produced. Other examples include the jazzy reinterpretation of Mozart by the Afro-Cuban group Irakere, the re-creation of English, Indian, and other music by the Beatles, Peter Gabriel, and many other musicians.

The rigid borders established by nation states became porous. Few cultures can be defined today as stable units, with precise boundaries based on the occupation of a confined territory. But the multiple opportunities to hybridize neither imply indetermination nor unrestricted freedom. Hybridization happens under specific historical and social conditions, in the midst of production and consumption systems that sometimes operate as constraints, as can be witnessed by the lives of many migrants. One of the social entities that support but also condition hybridization are the cities. The multilingual, multicultural megalopolis, for example, Paris in the nineteenth and twentieth centuries, Vienna at the beginning of the twentieth century, London, Berlin, New York, Los Angeles, Buenos Aires, Sao Paulo, Mexico, and Hong Kong are studied as centers where hybridization fosters greater conflict as well as greater cultural creativity (Appadurai 1996, Hannerz 1996).

3.3 Globalization

Globalization stresses these trends of modernity by creating world markets of material goods and money, messages, and migrants (workers, tourists, exiles). These constant streams and interactions have weakened borders and frontier posts; also the autonomy of local traditions. At the same time, they foster many more forms of hybridization than they ever did in the past: productive, commercial, communicational, and in consumption habits. To the characteristic modalities of hybridization, which came into existence due to migrations and trips, as well as to the educational integration policies by nation states, the mixtures generated by cultural industries have to be added. Industrialization encompasses not only the mass media (radio, movies, television, video), but also industrial production for massive audiences in international exhibitions, folkloric groups' tours, and globalized theme parks such as Disney's (Harvey 1996).

However, the main globalization analysts note that it not only integrates and generates *mestizajes*; it also segregates, produces new inequalities, and stimulates differentialist reactions. Identity alliances emerge (Latins or Arabs in the US or in Europe) as do organizations of regional integration (European Union, NAFTA, Mercosur), which in part aim at strengthening partial groupings to counteract the indiscriminate interaction of globalism (Beck 2000). Some of the actors find resources to resist both globalization and hybridization at once in these alliances.

3.4 The Future

Several anthropologists and cultural studies specialists maintain that the generalized relevance of hybridization requires changes in the ways social sciences are formulated. It is impossible to analyze large units, that is, nations, states, transnational enterprises, without noticing their internal diversities and modulations, as often happens in disciplines devoted to analyzing at the macro level, such as sociology and economics. It is not acceptable either that anthropology, traditionally dedicated to small communities, continues to see solely their homogeneity, and overlooks the heterogeneity and hybridization caused by the transnational goods and messages that intersect these communities. To be able to politically negotiate diversity and hybridizations, we need a '*mestizo* way of thinking' (Laplantine and Nouss 1997, Gruzinski 1999).

Fundamentalisms are not the only opposition to religious syncretism and intercultural *mestizaje*. There is an extended resistance to accepting hybridization because it generates insecurities within cultures and conspires against their ethnocentric self-esteem. It also presents a challenge for modern analytical thought, used to the binary separation: civilized, wild; national, foreign. This schematism leaves aside frequent ways in which culture is shared nowadays, for example, take

someone who is Brazilian by nationality, Portuguese by tongue, Russian or Japanese by origin, and Catholic or related to Afro-American religions.

Hybrid thought will not look forward to working with radically different, homogenous, compact units as much as with intersections, transitions, and transactions; less with identity than with heterogeneity and multiple senses of belonging. To understand the contemporary world it is indispensable to study how cultures mix, to what rhythms and modalities, what they lose and gain in the process, how they agree to cohabit, and how they negotiate to keep the difference alive. To comprehend hybridity and to learn to be cosmopolitan, 'multicultural gourmets', capable of traveling in diverse cultures and savoring the differences, has become necessary to fathom globalization and the ethnic and national societies' crisis. In brief, to understand society.

See also: Creolization: Sociocultural Aspects; Globalization and World Culture; Multiculturalism, Anthropology of; Multiculturalism: Sociological Aspects; Paradigms in the Social Sciences; Religion: Culture Contact; Specialization and Recombination of Specialties in the Social Sciences; Syncretism

Bibliography

Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MI

Barbero J M 1987 *De los medios a las mediaciones*. Ediciones G., Gili, México

Beck U 2000 *What is Globalization?* Polity Press, Oxford, UK

Bernand C 1993 Altérités et métissages hispano-américains. In: Descamps C (ed.) *Amériques Latines: une altérité*. Editions du Centre Pompidou, Paris

Bernd S, de Grandis R 1995 *Imprivesíveis Américas: Questóes de Hibrida (ao Cultural nas Américas)*. Sagra-Luzzatto, Porto Alegre, Brazil

Bhabha H K 1994 *The Location of Culture*. Routledge, London and New York

Callender L A 1988 Gregor Mendel: An Opponent of Descent with Modification. *History of Science* **XVI**

Cornejo Polar A 1997 Mestizaje e hibridez: los riesgos de las metáforas. Apuntes. [Miscegenation and Hybridity: The Dangers of Metaphor. Notes] *Revista Iberoamericana* **LXIII** (180): 341–4

García-Canclini N 1995 *Hybrid Cultures*. University of Minnesota Press, Minneapolis, MI

Gruzinski S 1999 *La pensée métisse*. Fayard, Paris

Hannerz U 1996 *Transnational Connections*. Routledge, London

Hannerz U 1997 Fluxos, fronteiras, híbridos: Palavras-chave da antropología transnacional. *Maná, Rio de Janeiro* **3**(1): 7–39

Harvey P 1996 *Hybrids of Modernity*. Routledge, London

Kymlicka W 1995 *Multicultural Citizenship*. Clarendon Press, Oxford, UK

Laplantine F, Nouss A 1997 *Le Méttissage*. Dominos, Paris

Olby R C 1984 Historiographical problems in the history of genetics. *Riv. Sror. Sci.* **1**(1): 25–38

Ortiz F 1995/1947 *Cuban Counterpoint*. Duke University Press, Durham, NC

Papastergiadis N 1997 Tracing hybridity in theory. In: Webner P, Modood T (eds.) *Debating Cultural Hybridity*. Zed Books, Atlantic Highlands, NJ

Rowe W, Schelling V 1991 *Memory and Modernity*. Verso, London

Webner P, Modood T (eds.) 1997 *Debating Cultural Hybridity*. Zed Books, Atlantic Highlands, NJ

N. García-Canclini

Hypertension: Psychosocial Aspects

Hypertension is a complex disorder of physiological regulation involving many interacting biological and environmental factors. Although the prevalence of hypertension has decreased over recent years, the latest US survey placed the age-adjusted prevalence ($> 140/90$ mmHg) at 20.4 percent. Interest in the role of psychosocial factors in hypertension has a long history beginning with theories about social stress, affective processes, and individual behavioral dispositions in its etiology. Noteworthy are the early psychosomatic writings about the significance of anger expression and suppression for high blood pressure.

1. Direct and Indirect Evidence

Numerous studies provide evidence of a definite role for psychosocial factors in hypertension. Direct evidence is derived from cross-sectional studies of individuals with hypertension compared to control samples and longitudinal studies to determine what factors are associated with the development of hypertension. Indirect evidence comes primarily from studies of psychosocial factors affecting blood pressure in healthy individuals with or without various predisposing characteristics. Psychosocial factors are also relevant to the degree of adherence to medications, engagement in health-related behaviors affecting risk for hypertension, effectiveness of antihypertensive medications, and nondrug methods of managing hypertension based on psychosocial concepts and methods.

2. Blood Pressure Reactivity and the Sympathetic Nervous System

Elevations in blood pressure can be elicited by many different stimuli and circumstances involving psychological and social factors. Examples are exposure to noxious, painful, threatening, or otherwise stressful

stimuli and events, participation in demanding tasks requiring repeated behavioral and/or mental adjustments, continuous striving and active coping, frustration in achieving goals, mental effort especially in solving difficult problems, momentary moods and behaviors involving fear, anger, and other emotional states, inhibition or suppression of emotional states and reactions, conflicts about the expression of angry feelings or other emotions, and interpersonal and group conflict. Socialcultural norms define what are appropriate and desired reactions and behaviors. The extent to which blood pressure is affected is determined by the balance between demands placed and the individual's resources and ability to cope with the demands.

A common theme in the literature is that individual differences in cardiovascular reactivity to various stimuli and events serve as a risk factor for hypertension and cardiovascular disorders in general. Increased sympathetic nervous system activity and associated hormonal changes is considered the neural source of excessive cardiovascular reactivity (Folkow 1989). In established hypertension, sympathetic activity is increased, particularly in individuals in whom the elevated blood pressure appears to have a basis in the nervous system ('neurogenic' hypertension) (Esler et al. 1977). Moreover, most antihypertensive medications appear to regulate blood pressure by interfering with mechanisms of sympathetic activation. Low parasympathetic activity (vagal tone) may also be a factor in hypertension and may contribute by slowing cardiovascular recovery from stress.

The large body of experimental evidence linking excessive cardiovascular reactivity in humans to established hypertension is compelling on empirical as well as theoretical grounds (Light 1987). Moreover, psychosocial factors are also associated with higher casual or ambulatory levels of blood pressure, suggesting that the transformation from excessive reactivity to a chronic elevation may take place with repeated instances of blood pressure increases. Thus, in some samples of subjects, during waking hours and even during sleep, average levels of blood pressure have been found positively associated with such personality traits as hostility and defensiveness (Shapiro et al. 1993). Physiological models have been developed to describe the cardiovascular dynamics underlying the transformation from early heightened reactivity to later established higher levels of blood pressure.

3. *Methodological Complications in Human Studies*

Despite the vast literature, debates continue about the conclusiveness of the evidence supporting the cardiovascular reactivity hypothesis based on responses to laboratory stressors. The findings are not always consistent or easily integrated. Individuals vary widely in their blood pressure responses to various circumstances. Variations in the kinds of individuals and methods chosen for study may result in different findings and conclusions. Laboratory settings may impose constraints on the participants so that they may not respond to challenges and threats with their preferred method of coping. Individuals also vary in the degree to which their responses persist over time for any one provocation or generalize across provocations. The association between perceptions of stress and the blood pressure response to such stress is not consistent because some individuals tend to report not feeling stressed, anxious, or angry while at the same time showing increased physiological reactivity to provocations or elevated blood pressure levels in general. These individuals either generally do not feel negative emotions or wish to make a good impression on other people. This characteristic has been labeled 'defensiveness.' Finally, many individuals learn to cope effectively with stress or to keep their emotional responses under control. For example, experienced air traffic controllers or other professionals dealing with emergencies do not necessarily show elevated levels of blood pressure (Sega et al. 1998). By taking these various factors into account, further sense may be made out of some of the apparently inconsistent findings, thus adding to confidence in the soundness of the reactivity model.

Thus, the ability to predict individual differences in reactivity and in the development of sustained high levels of blood pressure and established hypertension has been improved by considering other psychological factors such as various personality traits and coping styles (anger/hostility, defensiveness and emotional suppression, denial and repression, anxiety, lack of control and helplessness, Type A behavior pattern, depression, pessimism) (Johnson et al. 1992, Jorgensen et al. 1996, Suls et al. 1995). In the social domain, contributing factors are low socioeconomic status, social disintegration and family disruption (poverty, high crime rates, high divorce rates), occupational stress (job strain, job dissatisfaction, high job demands, little independence in one's work duties or latitude in the ability to make decisions), low levels of family and social support, social isolation, taxing responsibilities such as providing care to disabled family members, marital conflict, minority social status, certain ethnicities, and certain gender roles (Colhoun et al. 1998, Pickering et al. 1998). For example, anger and hostility do not necessarily have the same implications for men and women.

4. *Psychobiological Models*

The ability to predict whether repeated elicitations of increases in blood pressure will lead eventually to chronically elevated blood pressure and to cardiovascular disease is probably best improved when

certain biological factors are included in the test models, such as obesity, physical inactivity, poor physical conditioning, salt sensitivity, excitable autonomic nervous system, deficiencies in certain vitamins and minerals, intake of sodium and potassium, presence of other disorders such as diabetes, high glucose, insulin, or lipid levels, certain racial characteristics, and other genetic or constitutional predispositions such as family history of hypertension and cardiovascular disease. For example, in a study of occupational stress in nurses, those nurses for whom both parents had hypertension and who were hostile, defensive, or anxious had elevated ambulatory blood pressure, which was highest during work days (Goldstein and Shapiro 2000). Other psychosocial, life style, and habitual behavior patterns need to be considered as they are relevant to one's diet, activity level, and intake of caffeine, alcohol, and other substances. For example, high levels of social stress or certain personality traits like a high level of trait hostility may be associated with unhealthy health habits such as smoking, low levels of exercise, poor physical conditioning, and poor nutrition.

5. *Diagnostic Issues*

Psychological factors are relevant to the diagnosis of hypertension. Some patients may show elevations in blood pressure during a doctor's visit and evaluation but not at other times ('white coat effect'). This effect has been related to an alerting response to the examination or anxiety about one's health. In one study, trait hostility was found relevant to this effect. Those individuals who were highly hostile were consistently high in blood pressure during the examination and over the course of the day, whereas low hostile individuals were more likely to show elevated pressure only during the evaluation and not during the day (Jamner et al. 1993). By having the patient record blood pressure at home, useful information is obtained on blood pressure variability at different times and occasions that may guide decisions about treatment and increase the patient's adherence to medications or otherwise facilitate pressure reduction. Reliable devices are now available for this purpose.

6. *The Problem of Blood Pressure Variability*

Blood pressure can be quite variable both in the short term with a range of 5 mmHg or more over the course of a min and 30–40 mmHg or more over the course of waking hours during the day. These facts provided the impetus for the development of portable automatic 24-hour blood pressure recorders, initially used in the diagnosis of hypertension and the evaluation of drug effects and more recently in the investigation of blood pressure in different settings (Shapiro et al. 1993). Ambulatory blood pressure recording methods buttressed by information obtained from diaries of self reports of moods and the situations encountered in everyday life offer many advantages, including a large number of blood pressure measurements under different conditions (work vs. nonwork or off days, sleep vs. waking periods) and in a variety of repeatedly occurring situations, locations, and activities. Valuable data have been obtained from studies of special populations: paramedics, traffic control agents, secretaries, laboratory technicians, emergency dispatchers, nurses, bus drivers, college students. The research is challenging because of the complexities and uncontrolled nature of the observations, but novel analytical tools show great promise.

7. *Psychosocial Factors Affecting Treatment*

Psychosocial factors play a role in the treatment and management of hypertension. The response to antihypertensive drugs may partly depend on psychosocial factors, although research is sparse on this topic and these factors are not usually considered in the choice of medications used to treat patients. In one example, individuals with high levels of hostility were found less responsive to diuretics than nonhostile individuals, suggesting that the former might be more responsive to agents directly affecting the sympathetic nervous system. Ethnicity can affect the response to drug and non-pharmacologic treatments of hypertension (Brownley et al. 1999). Psychological factors influence degree of adherence to medications and may affect the degree to which side effects of medications occur and affect quality of life.

A great deal of research has been devoted to the evaluation of nondrug treatments of hypertension derived from psychological principles (Silverberg 1990). These procedures seem most suitable as alternatives in patients with mild hypertension, especially those who are free of significant cardiovascular disease risk factors (family history, high cholesterol level, smoking) and in other patients as a means of reducing the amount of medication needed for blood pressure control or completely eliminating the need for such medication. Reducing drug requirements may be of special value in dealing with undesired side effects of medications in some patients. The methods include progressive relaxation, simple relaxation, biofeedback-guided relaxation, biofeedback training involving blood pressure, heart rate, and other cardiovascular measures, biofeedback training using measures associated with sympathetic nervous system functioning, cognitive behavioral methods of facilitating coping with stress and dealing with anger ('stress management training'), hypnosis, and autogenic training. As yet there is no consensus about the effectiveness of these procedures, and major trials remain to be done. Other traditional methods such as yoga and meditation may be useful as a means of increasing awareness and control of mental and bodily functioning. Regular physical exercise can

facilitate blood pressure reduction and improve the coronary circulation. Whether or not people exercise, avoid or take medications, or use herbal, homeopathic, or various nondrug methods is determined by many different psychological, social, and cultural factors.

8. Summary and Conclusions

From the standpoint of the social and behavioral sciences, hypertension has attracted a great deal of attention. It forms the backbone of the growing fields of health psychology and behavioral medicine, and benefits greatly from the integration of methods, concepts, and findings of psychology, psychophysiology, social psychology, physiology, and biological and medical sciences. The evidence is supportive of the role and significance of psychosocial factors in the etiology, course, and treatment of hypertension. These factors include personality and coping styles, social and occupational stress, social status, family and social support, ethnicity, and gender roles. This knowledge has the potential for deepening our understanding of the mechanisms involved in hypertension, improving the effectiveness of both drug and non-drug approaches to treatment, and facilitating efforts to prevent the disorder. There is great need for further systematic research and especially for theoretical and empirical integration of psychosocial and biological knowledge of the disorder.

See also: Biopsychology and Health; Cardiovascular Conditioning: Neural Substrates; Coronary Heart Disease (CHD), Coping with; Coronary Heart Disease (CHD): Psychosocial Aspects; Coronary-prone Behavior, Type A; Gender and Cardiovascular Health; Stress and Cardiac Response; Stress and Health Research

Bibliography

Brownley K A, Hurwitz B E, Schneiderman N 1999 Ethnic variations in the pharmacological and non-pharmacological treatment of hypertension. *Human Biology* **71**: 607–39

Colhoun H M, Hemingway H, Poulter N R 1998 Socioeconomic status and blood pressure: An overview analysis. *Journal of Human Hypertension* **12**: 91–110

Esler M, Julius S, Zweifler A, Randall O, Harburg E, Gardiner H, De Quattro V 1977 Mild high-renin essential hypertension: Neurogenic-human hypertension. *New England Journal of Medicine* **296**: 405–11

Folkow B 1989 Sympathetic nervous system control of blood pressure: Role in primary hypertension. *American Journal of Hypertension* **2**: 103S–111S

Goldstein I B, Shapiro D 2000 Ambulatory blood pressure in women: Family history of hypertension and personality. *Psychology, Health, & Medicine* **5**: 227–40

Jamner L D, Shapiro D, Hui K, Oakley M E, Lovett M 1993 Hostility and difference between clinic, self-determined and ambulatory blood pressure. *Psychosomatic Medicine* **55**: 203–11

Johnson E H, Gentry W D, Julius S (eds.) 1992 *Personality, Elevated Blood Pressure, and Essential Hypertension*. Hemisphere, Washington, DC

Jorgensen R S, Houston B K, Kolodziej M E, Schreer G E 1996 Elevated blood pressure and personality: A meta-analytic review. *Psychological Bulletin* **120**: 293–320

Light K C 1987 Psychosocial precursors of hypertension: Experimental evidence. *Circulation* **76**: 167–76

Pasic J, Shapiro D, Jamner L D, Hui K K 1994 Hostility and the response to diuretic in mild-to-moderate hypertension. *American Journal of Hypertension* **7**: 503–8

Pickering T G, Devereux R B, James G D, Gerin W, Landsbergis P, Schnall P L, Schwartz J E 1997 Environmental influences on blood pressure and the role of job strain. *Journal of Hypertension* **14**(suppl.): S179–S185

Sega R, Cesama G, Costa G, Fennanio M, Bombelli M, Mancia G 1998 Ambulatory blood pressure in air traffic controllers. *American Journal of Hypertension* **11**: 208–12

Shapiro D, Jamner L D, Goldstein I B 1993 Ambulatory stress psychophysiology: The study of 'compensatory and defensive counterforces' and conflict in a natural setting. *Psychosomatic Medicine* **55**: 309–23

Silverberg D S 1990 Non-pharmacological treatment of hypertension. *Journal of Hypertension* **8**(suppl.): S21–S26

Suls J, Wan C K, Costa P T Jr. 1995 Relationship of trait anger to resting blood pressure: A meta-analysis. *Health Psychology* **14**: 44–56

D. Shapiro

Hypnosis, Neural Basis of

Modern conceptions of hypnosis can be traced directly to the work of Franz Anton Mesmer, a Viennese physician working in Paris just prior to the French Revolution. Within 30 years of Mesmer's work, most of the basic hypnotic phenomena from amnesia to pain control to hallucinations had been reported (see Ellenberger 1970). James Braid, a Scottish surgeon, coined the term hypnosis in the 1840s. Nowadays, hypnosis refers to the acceptance of a social role, usually involving deep relaxation, and engagement in fantasy. During hypnosis a lowering of critical judgment permits a wide range of behavioral, cognitive, and affective responses to suggestion (see, for example, Orne 1970). This article addresses the question of whether response to hypnosis can be related to specific neural mechanisms or events.

During the last two hundred years, a variety of exaggerated and mistaken claims have been made for hypnosis. Claims have ranged from enhanced psychic abilities to past life regression, to revived memories of alien abduction. In the past, such claims have resulted in periods of scientific disrepute for hypnosis. Nevertheless, during the last century fine scientists (e.g., Hull, Hilgard, Orne, Barber, Sarbin and their students) have been drawn to study hypnosis (see Sheehan and Perry 1976). As well as elucidating the domain of hypnosis, their careful investigations have increased understanding of the potential pitfalls that appear in

experimentation with human research participants (see Orne 1959, 1970).

An observer, watching a demonstration of hypnosis with a highly hypnotizable research participant, sees a series of simple verbal suggestions that result in relatively spectacular alterations of behavior, thought, emotion, and perception. For example, the hypnotized burn patient lies quietly and reports no pain, seemingly relaxing while his dressings are changed: a highly painful procedure (Hilgard and Hilgard 1975). The laboratory research participant asked to re-experience the distant past seems to become childlike, entirely captured by the delusion that he or she is many years earlier. Asked to hallucinate the absence of an obstacle between where the participant sits and some other place, the participant claims to see nothing in the intervening space. But when asked to walk to that place (and thus, through the obstacle whose absence is being hallucinated), the participant walks around the obstacle without seeming to notice doing so. Asked to look back and see that there is nothing there, there is easy agreement with no seeming sense that walking around the obstacle has just contradicted this statement. Asked to forget all that has happened until a pencil is tapped twice, the participant will later 'awaken' with amnesia that is (almost always) temporary and reversible.

Thus, in response to brief verbal suggestions, the hypnotized individual seems to see, hear, feel, smell, and taste in apparent contradiction to the stimuli actually present (see Orne and Hammer 1974). Memory, the sense of volition, mood, and even awareness of self may be altered. With appropriate suggestions, such effects may be extended into the posthypnotic period.

Hypnotic phenomena are easy to elicit and robust. A graduate student, given two or three hours training and a standardized script, such as one of the *Stanford Hypnotic Susceptibility Scales*, can elicit all the major hypnotic phenomena in a psychology laboratory (Weitzenhoffer 1959, 1963; see also Shor and Orne 1962). There they can be studied under controlled conditions. Note that for all practical purposes, hypnotic suggestions are likely to be carried out only in settings seen as culturally appropriate. These include therapeutic, classroom, entertainment, research, and forensic settings (Orne 1970).

1. Problems with Seeking Psychophysiological Concomitants of Hypnosis

People commonly speak of being 'hypnotized' or in a hypnotic 'trance.' The underlying assumption is that hypnosis is a distinct, altered state of consciousness. Most early theorizing (and much current clinical thought) about the hypnotic state employed circular reasoning that goes something like this: Q: Why do you say that person is hypnotized? A: Because she is responding to hypnotic suggestions. Q: Why does she respond to hypnotic suggestions? A: Because she is hypnotized.

As Orne (1959) showed us, the hypnotized participant acts as shared cultural expectations suggest. In fact, the deeply relaxed hypnotic participant lying or sitting quietly while listening attentively to the hypnotist's voice manifests only one of many forms hypnosis has assumed. For example, in the late eighteenth century, Mesmer's Parisian patients sometimes had pseudo-epileptic seizures during hypnosis. In the late nineteenth century, Charcot's Parisian patients demonstrated first catalepsy, then lethargy, and finally somnambulism (but generally did not have seizures) when deeply hypnotized. At the same time in the south of France, Bernheim's hypnotized patients simply restfully reclined (see Ellenberger 1970). Most modern research participants and patients sit quietly and look very relaxed during hypnosis, although they may also be hypnotized with an 'active-alert' induction, involving, for example, pedaling a stationary bicycle vigorously (see Banyai and Hilgard 1976).

Obviously, peripheral neural activity during these varying hypnotic activities differs widely, reflecting the activity, not hypnosis. Nor has there been any evidence of brain activity unique to hypnosis *per se*. Nor is there evidence of specific brain activity that accompanies the entire hypnotic experience. After 50 years of modern research, the lack of evidence for any unique psychophysiological marker of a hypnotic trance state is one of the factors that have led researchers largely to abandon the concept of a hypnotic trance as a causal explanation for response to hypnotic suggestions (see Kirsch and Lynn 1995).

Concepts such as response expectancies, role involvement, self-delusion, and cognitive abilities are currently used to explain hypnotic responses. This does not mean that hypnotizing people does not have specifiable effects or that the experiences of highly hypnotizable subjects are consciously faked or otherwise theoretically unimportant. Rather, hypnotic effects result from making suggestions to willing participants of differing abilities and from the expectations and cognitive models that a hypnotic induction engenders both for the hypnotist and the participant (see Sheehan and Perry 1976).

At present, it is clear that the ability to respond to hypnotic suggestions lies largely in the subject (and not in the skill of the hypnotist as was previously held). Reports of being hypnotized or in a trance are seen as reflecting a subjective experience that is one of the many effects of cognitive and emotional set, social and historical setting, and suggestion (see Kirsch and Lynn 1995). And while all subjective experiences must have some neurophysiological substrate, psychology as a field of study has not reached the point where many can be pinpointed. Nonetheless, if phrased properly, reasonable questions about the psychophysiologic concomitants of hypnosis can be asked.

2. *Psychophysiological Differences in Response to Hypnotic Induction*

A discussion of the neural basis of hypnosis may be phrased in terms of answers to three questions. First, are there measurable differences in brain activity in response to hypnotic induction among those who subsequently respond strongly to hypnosis in comparison to those who do not? Second, when measured outside hypnosis, are there overall psychophysiologic differences between participants who usually respond to hypnotic suggestions for major changes in perception, memory, and/or cognition and those who do not? Third, are there patterns of brain activity associated with response (vs. lack of response) to specific hypnotic suggestions? While there are interrelated issues among these three questions, it is probably least confusing to look at these questions sequentially.

First, does hypnotic induction result in measurably different neural patterns among those who subsequently respond to the more spectacular hypnotic suggestions (those calling for hallucinatory or major cognitive changes) and those who do not? The evidence, while mixed, favors a tentative 'Yes, but for reasons that render the difference unimportant.' Brain function studies are divided between those that find highly hypnotizable, well lateralized, right-handed research participants shifting to greater relative right-hemisphere activation after hypnotic induction and studies that find no differences. The lack of consistent findings may be methodological: studies finding differences tend to use ratio measures that are meant to be sensitive to small differences (see, for example, LaBriola et al. 1987). Studies that do not find differences after induction tend to look at absolute values of measures of brain activation, a less sensitive measure (Graffin 1995). So there is some evidence that successful induction of hypnosis often involves a 'shift to the right' a finding in line with task performance measures of the effects of induction.

Why is this unimportant? Activation changes in response to induction are relatively small and disappear once other suggestions are made or tasks required. Relative hemispheric activation, with or without hypnosis, depends on the mental task participants are asked to perform. So the 'shift to the right' seemingly reflects a response specific to the 'task' of induction. And that response might well change were a different type of induction used (see Crawford and Gruzelier 1992).

3. *Psychophysiological Differences between Highly Hypnotizable Participants and their Less Hypnotizable Counterparts?*

Many studies distinguish between highly hypnotizable subjects and their less hypnotizable peers. Most modern scales of hypnotic responsiveness comprise standardized, relaxation oriented inductions followed by 5 to 12 hypnotic suggestions that are each scored pass/fail based on an objective criterion. Entire hypnotic scales are often administered by tape recorder and both experimenter-scored individual formats and self-scored group formats are available. Total score, the measure of each participant's hypnotizability, is simply number of suggestions passed (see Kihlstrom 1985).

All social scientists use Pearson's reliability coefficient, a coefficient that varies from $+1.00$ (relative scores on two measures are identical) to 0.00 (no linear relationship between two sets of scores) to -1.00 (relative scores are identical in size but opposite in sign). Hilgard and his colleagues have found that test-retest reliability of hypnotizability scales usually averages about $+.90$ over periods of one week and about $+.70$ over periods from 10 to 25 years (Piccione et al. 1989). These data indicate that hypnotizability is a highly stable difference among individuals. Among personality traits, the reliability of hypnotizability measurement is equaled or exceeded only by multi-task measures of intelligence.

Hypnotizability is not easily modifiable and governs the large majority of the variance in most experimental studies of hypnotic phenomena. Moreover, what has come to be called 'hypnotizability' is observable in response to suggested alterations in perception, memory, and/or cognition with or without formal hypnosis (see Kihlstrom 1985).

There are two fairly well established psychophysiologic correlates of hypnotizability. First, highly hypnotizable research participants (highs) show higher levels of resting theta activity than their less hypnotizable counterparts. Spontaneous EEG activity may be viewed as constituted of waveforms of different frequencies (delta -0 to 3 Hz, theta -4 to 7 Hz, alpha -8 to13 Hz, beta -14 to 30 Hz). Waking EEG records are largely made up of alpha and beta activity, with a small percentage of theta activity. When measured during rest, this small proportion of theta activity is slightly higher among the more highly hypnotizable participants, especially in recordings from the more anterior brain regions. While no convincing explanations for this phenomenon have yet been advanced, the notion that the theta differences between highly hypnotizable participants and their less hypnotizable counterparts are related to attentional abilities is the most popular hypothesis (see Graffin et al. 1995, Sabourin et al. 1990).

Second, if a variety of other factors are controlled (e.g., task difficulty, familial history of right handedness), well lateralized, right-handed research participants will show greater relative left cerebral hemispheric activation during verbal and mathematical tasks and greater right hemisphere activation during mental imagery. This pattern of task specific hemispheric asymmetry is seen more clearly among highly hypnotizable participants than among their less

hypnotizable counterparts. Highs show greater relative left cerebral hemisphere activation during verbal and mathematical tasks than do less hypnotizable participants. Similarly, highs show greater relative right hemisphere activation during imagery than do low or moderately hypnotizable participants. (The inclusion of moderately hypnotizable subjects provides some control for compliance. Moderates show the same level of compliance to behavioral suggestions as the highs, but are largely unresponsive to hallucinatory suggestions.) These findings have been interpreted as showing greater cognitive flexibility for highs and are in line with some findings in studies using task performance measures (see Crawford and Gruzelier 1992).

4. Psychophysiological Correlates of Response to Specific Hypnotic Suggestions

The third question is whether psychophysiologic differences can be found between those responding to specific hypnotic suggestions and those who fail to respond. This question can be asked in respect to hypnotic phenomena of central interest to experimental psychologists (e.g., hypnotic hallucinations), clinicians (e.g., healing of skin and immune system related disorders), or both (e.g., hypnotic analgesia).

Until the recent advent of functional magnetic resonance imaging (fMRI), activity in the normal waking human brain has been studied, for the most part, using measures of spontaneous and evoked electrical activity in the brain. Researchers usually utilize one of three electroencephalographic (EEG) technologies in which brain activity is recorded from the surface of the skull using standardized electrode placements. Obtaining EEG signals is similar to making an electrocardiogram except that EEG signals are measured in microvolts rather than millivolts signals studied with an EKG. (Since much greater amplification is required, EEG recording requires great care in avoiding electrical, movement, and other artifacts.)

Once having obtained an artifact free EEG signal, researchers digitize and transform it. Most frequently they use signal processing, mathematical (Fourier) transformations to determine the proportion of time in which alpha, beta, and theta activity occur. Alternatively, one may integrate the total area under the curve of an EEG record and compare total area at various sites in the brain during different tasks. In both methods one examines the signals from many recording periods (usually of 1 second to 10 minute duration) over minutes or hours of total recording time (see Karlin et al. 1980, Perlini and Spanos 1991, Sabourin et al. 1990). Almost all the research underlying the conclusions in the earlier sections of this essay has used one of these two techniques.

A third EEG procedure involves the presentation of brief, repeated auditory, visual, or tactile stimuli to elicit acute changes in electrical brain activity called event-related evoked potentials. Evoked potentials can be seen when the spontaneous EEG, used as data in the first two methods, is mathematically removed and only short-term changes in EEG signals remain. The precise shape of evoked potentials differs in relationship to a variety of factors including the duration, intensity, predictability, meaning, and sensory modality of the stimulus. However, when such factors are controlled, one expects a sequence of alternating positive and negative potentials during the second or so after stimulus presentation.

Several studies seemingly indicate that hypnotic suggestions to hallucinate alter evoked potentials among highly hypnotizable participants in ways consistent with their reported hallucinatory experiences. In these studies, highly hypnotizable participants who claim to experience either a particular hallucination or amnesia are compared to less hypnotizable counterparts who do not report similar experiences. Across visual and auditory stimuli and across experiments, highly hypnotizable participants seem to show augmented or faster positive evoked potentials and/or decreased or delayed negative potentials when asked to hallucinate the presence of stimuli that are not there. Again consistent with suggested hallucinations, when asked to hallucinate the absence of stimuli that are present, highly hypnotizable subjects show decreased or delayed positive evoked potentials, and/or augmented or faster negative evoked potentials compared to their less hypnotizable counterparts. Parallel findings to those found in negative hallucinations have emerged after amnesia instructions. In sum, highly hypnotizable participants show brain activity consistent with their reports of hypnotic hallucinations (see Crawford 2001).

Perhaps because it is the most clinically relevant negative hallucination, the effects of hypnotic analgesia instructions, instructions not to experience pain when exposed to a painful stimulus, have been studied with a variety of measures of brain activity. These have included fMRI, cerebral blood flow using 133-xenon, SPECT, and integrated EEG. Such studies confirm the pattern seen in the evoked potential studies: instructions to experience analgesia leads to measurable changes in brain function indicating central amelioration of the painful stimuli among highly hypnotizable participants, but not in lows or moderates (see Crawford 2001, Karlin et al. 1980). It is also noteworthy that other studies indicate that hypnotic analgesia is not mediated by endogenous opiate systems (Hilgard and Hilgard 1975).

As to immune system and skin disorders, the positive data remain largely anecdotal. There is a long history of treating skin disorders with hypnosis. (For example, hypnosis is still treatment of choice for warts that do not respond to conventional treatment.)

Similarly, asthma may be treated with hypnosis (although there is some evidence that hypnotic effects are limited to relaxation of the upper airways, rather than changes at the level of the lung where oxygen exchange occurs). Finally, there have been a number of studies on the ability of suggestion and imagery to alter immune responding directly. However, findings have been mixed and no clear conclusions can yet be drawn other than that large, robust effects seem unlikely (see Rhue et al. 1993).

5. Conclusions

Hypnotic responses remain intriguing, somewhat spectacular effects produced by simple verbal suggestion. Since the 1950s, careful rigorous work has been done to elucidate underlying mechanisms and any psychophysiologic concomitants of these phenomena. Inducing hypnosis *per se* has no clear physiologic effects. However, when compared to their less hypnotizable counterparts, highly hypnotizable subjects show slightly higher levels of theta activity during rest and greater task consistent asymmetry on strongly lateralized cognitive tasks. This is consistent with the notion that highs are capable of more flexible attention than are their less hypnotizable counterparts. Moreover, when successfully engaged in positive and negative hallucinations, including hypnotic analgesia and amnesia, highly hypnotizable subjects show cortical activity consistent with their reported subjective experiences across a number of measures of brain function.

See also: Electroencephalography: Basic Principles and Applications; Electroencephalography: Clinical Applications; Event-related/Evoked Potentials; Hallucinations, Neural Basis of; Hypnosis, Psychiatry of; Neglect, Neural Basis of; Pain, Management of; Sleep: Neural Systems

Bibliography

Banyai E, Hilgard E 1976 A comparison of active-alert hypnotic induction with traditional relaxation induction 1976. *Journal of Abnormal Psychology* **85**: 218–24

Chaves J, Perry C, Frankel F (eds.) 1998 Special Issue: Hypnosis in the relief of pain: Parts 1 & 2. *International Journal of Clinical and Experimental Hypnosis* **45**(4): 354–3355 & **46**(1): 4–5

Crawford H 2001 Neuropsychophysiology of hypnosis: Toward an understanding of how hypnotic interventions work. In: Burrows G, Stanley R, Bloom P (eds.), *Handbook of Clinical Hypnosis*. Wiley, New York

Crawford H, Gruzelier J 1992 A midstream view of the neuropshysiology of hypnosis: Recent research and future direction. In: Fromm E, Nash M (eds.) *Contemporary Hypnosis Research*. Guilford, New York, pp. 227–66

Ellenberger H 1970 *The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry*. Basic Books, New York

Graffin N, Ray W, Lundy R 1995 EEG concomitants of hypnosis and hypnotic susceptibility. *Journal of Abnormal Psychology* **104**: 123–31

Hilgard E, Hilgard J 1975 *Hypnosis in the Relief of Pain*. Kaufman, Los Altos, CA

Karlin R, Morgan D, Goldstein L 1980 Hypnotic analgesia: A preliminary investigation of quantitated hemispheric electroencephalographic and attentional correlates. *Journal of Abnormal Psychology* **89**: 591–94

Kihlstrom J 1985 Hypnosis. *Annual Review of Psychology* **36**: 385–418

Kirsch I, Lynn S 1995 The altered state of hypnosis: Changes in the theoretical landscape. *American Psychologist* **50**: 846–58

LaBriola F, Karlin R, Goldstein L 1987 EEG laterality changes from prehypnotic to hypnotic periods. *Archives of Biological Psychiatry* **16**: 1–5

Orne M 1959 The nature of hypnosis: Artifact and essence. *Journal of Abnormal and Social Psychology* **58**: 277–99

Orne M 1970 Hypnosis, motivation, and the ecological validity of the psychological experiment. In: Arnold W, Page M (eds.) *Nebraska Symposium on Motivation*. U. of Nebraska Press, Lincoln, NE, 187–265

Orne M, Hammer A 1974 Hypnosis. *Encyclopedia Britannica*, 15th edn. Helen Hemingway Benton, Chicago, pp. 133–40

Perlini A, Spanos N 1991 EEG, alpha methodologies and hypnotizability: A critical review. *Psychophysiology* **28**: 511–30

Piccione C, Hilgard E, Zimbardo P 1989 On the degree of stability of measured hypnotizability over a 25-year period. *Journal of Personality & Social Psychology* **56**: 289–95

Rhue J, Lynn S, Kirsch I 1993 *Handbook of Clinical Hypnosis*. American Psychological Association, Washington, DC

Sabourin M, Cutcomb S, Crawford H, Pribram K 1990 EEG correlates of hypnotic susceptibility and hypnotic trance: Spectral analysis and coherence. *International Journal of Psychophysiology* **10**: 125–42

Sheehan P, Perry C 1976 *Methodologies of Hypnosis*. Erlbaum, Hillsdale, NJ

Shor R, Orne E 1962 *The Harvard Group Scale of Hypnotic Susceptibility: Form A*. Consulting Psychologists Press, Palo Alto, CA

Weitzenhoffer A 1959 *The Stanford Hypnotic Susceptibility Scale for use in research investigations in the field of hypnotic phenomena Forms A & B*. Consulting Psychologists Press, Palo Alto, CA

Weitzenhoffer A, Hilgard E 1963 *The Stanford Profile Scales of Hypnotic Susceptibility Scales: Forms I and II*. Consulting Psychologists Press, Palo Alto, CA

R. A. Karlin and E. Orne

Hypnosis, Psychiatry of

1. Definition

This treatment technique is thought to have received its name from the English surgeon James Braid (1795–1860) whose clinical observations likened the physiological condition of those who were responsive to the intervention to the state of sleep. It is the direct descendent of the seventeenth and eighteenth century

practices then called 'Magnetism,' 'Somnambulism' and 'Mesmerism,' although its roots can be traced to the religious and healing arts of antiquity. Today practitioners of hypnosis would subscribe to the American Psychological Association's (1993) definition of hypnosis as a 'procedure wherein changes in sensations, perceptions, thoughts, feelings or behaviors are suggested.'

2. Relevant History

Chroniclers of hypnosis (see Ellenberger 1970, Gauld 1992) point to the work of Franz Anton Mesmer (1734–1815) as the cornerstone for the establishment of this technique in modern medicine. Mesmer's belief that an imbalance in magnetic fluid led to disease states in humans gave rise to the ritualized application of magnets to individuals with a variety of medical and psychological problems. Not only the metal apparatus, but also certain people were thought to possess the 'animal magnetism' to restore the magnetic equilibrium and health. French government and scientific investigators who tested his work were able to dispute Mesmer's disease theory, but not the results. People continued to recover from a variety of ailments by means of suggestions (i.e., ideas) provided by charismatic individuals, and the technique soon spread throughout Europe and the USA.

Sigmund Freud (1856–1939) made his living as a young psychiatrist by treating patients, almost exclusively, with hypnosis. As a student of the famous neurologist Jean Martin Charcot (1825–1893) and the physician Hippolyte Bernheim (1840–1919), he learned about the efficacy of an idea to facilitate an action or symptom relief. His early technique employed suggestions to deny his patients' complaints energetically and to provide assurance and motivation for the performance of more adaptive behaviors. Freud also observed hypnotized subjects seeming to possess knowledge without realization, carrying out post-hypnotic suggestions without awareness of the source, and blindly following the commands of the hypnotist. From this clinical data and as a translator of Charcot's and Bernheim's writings on suggestion, hysteria and hypnosis, he became convinced that there were powerful, unconscious mental forces extant in humans that influenced behavior.

In Charcot's Salpetriere, Freud learned that traumatic events create trance states in hysterics, similar to hypnosis, wherein individuals are vulnerable to autosuggestion, and painful emotions become pathologically trapped in the nervous system. Hysterics, then, suffered painful reminiscences of these traumata that became expressed cognitively, affectively, somatically and/or behaviorally. Since they possessed similar characteristics, hypnotic trance was employed to assist the individual to revisit the scene of the original trauma, re-experience the original affects for cathartic discharge and recover from their symptoms through catharsis and insight. This technique was the basis for Freud's uncovering treatment in the *Studies on Hysteria* (Breuer and Freud 1893–1895) and hypnosis became one of the underpinnings of psychoanalytic theory and technique.

The power of suggestion, the influence of unconscious mental processes in humans and the use of hypnosis to access memories were not the only contributions of hypnosis to psychoanalysis. Freud was fascinated by the question of 'influence' and what led people to follow the directions of others. He reasoned that group activity and hypnotic treatments encourage people to minimize individual interests, personal control, discernment, personality style, etc., and to develop intense emotional relationships with people for reasons that are difficult to account for objectively. Freud saw the rapport created between doctor and patient in hypnosis to be comparable to that of tribal members for the chief and used it as a paradigm for his understanding of therapeutic influence and transference.

Where Mesmer might be credited with taking hypnosis away from the exorcists and placing it at least on the fringe of traditional medicine, Freud clearly explored hypnosis for what it could offer as a psychotherapeutic technique, and for what it might contribute to his model of the mind. It is clear from a close reading of his work that Freud continued throughout his career to use hypnosis as an archetype with which to compare and study his incipient psychoanalytic theory. However, his report that psychoanalysis began with his rejection of hypnosis relegated this technique to an inferior place among methods for mental health treatment. Very few of his followers would risk alloying the 'copper of suggestion' with the 'pure gold of psychoanalysis.' Freud's efforts to distance himself and psychoanalysis from hypnosis had something of the same dampening effect as Mesmer's failure to validate magnetic theory scientifically.

The birth of psychology and social science research techniques during the twentieth century brought new methods of inquiry into the essence and effects of hypnosis. The rise of interest in mind–body medicine also fueled clinical interest and investment in this area. Both of these developments, along with the creation of several professional journals and international societies for the study and practice of hypnosis, contributed to the continued development of this field.

3. Therapeutic Applications

3.1 Medicine

A technique that can alter sensations, perceptions, thoughts, feelings and/or behavior can be especially helpful in today's medical setting (Covino and Frankel 1999). Patients and practitioners have long experi-

enced anecdotal reports of symptom relief from something that a physician said, did or provided. Researchers in the laboratory and clinical setting have been laboring to establish empirical support for the use of hypnosis with a variety of symptomatic complaints and to manage common emotions in the medical setting to facilitate treatment.

Most clinicians have witnessed the influence of a patient's perception on the subjective experience of pain. Where an objective physiological event almost always serves as the cause of an injury, the meaning that the person makes of the event determines the amount of suffering. In experimental laboratories, hypnotized subjects have demonstrated an ability to tolerate electric shock, ischemic and cold pressor pain better than their nonhypnotized peers or when they employed other pain management strategies. Hypnotic suggestions for analgesia, distraction, and dissociation have been found to be superior to placebo, relaxation, and cognitive interventions in the management of a variety of painful conditions. Clinical research finds success for hypnosis in the management of pain associated with surgery, cancer, reflex sympathetic dystrophy, burns, interventional radiology, angioplasty, bone marrow transplant, and colonoscopy.

A number of experimenters have been successful at producing raucous breathing and decreased pulmonary function among a subset of asthma patients by means of suggestion alone. Whether these patients were invited to believe erroneously that they were encountering an allergen or breathing a bronchoconstricting drug, the idea alone was enough to affect their breathing adversely. Several well-designed studies have found that hypnosis training has a better impact on asthma symptoms, 'as needed' (p.r.n.) medication use, and pulmonary function tests than education, relaxation, guided imagery, and placebo inhalers. Furthermore, asthma patients who were trained to use hypnosis to manage their breathing have demonstrated a superior ability to tolerate methacholine challenge testing.

Allergists, dermatologists, and others have reported successful treatment of warts, eczema, psoriasis and other skin conditions with hypnosis. In clinical studies, suggestions to imagine bathing in salty water, increasing warm and tingling feelings on the skin, reducing blood flow to infected areas and general healing have assisted patients to relieve itchiness, reduce inflammation and decrease the number and size of warts. Likewise, patients with severe burns have been able to refrain from scratching, tolerate dressing changes and reduce their use of pain medications with the help of hypnosis.

Many patients with irritable bowel syndrome experience emotional distress in the form of anxiety and anger that adversely impacts gastric secretion, motility, and pain. A considerable number of these patients find relief from placebo interventions and several studies report alterations in gut physiology with hypnosis. Gastric acid secretion, gut motility, bowel habits, abdominal distention and abdominal pain have been reported to improve with very brief hypnotic interventions aimed at increasing relaxation, controlling smooth muscle and gut activity and bowel function.

Hypnotic training has been found to diminish the severity and frequency of bleeding episodes among patients with hemophilia. Training a small sample of patients to use hypnosis for relaxation and anxiety reduction has controlled factor usage, spontaneous bleeding and the need for transfusion. In addition, preoperative training in self-hypnosis for the control of anxiety and bleeding has been found to reduce the number of blood transfusions required by surgical patients.

One of the most common causes of noncompliance with chemotherapy among cancer patients is the significant distress that they experience from the associated nausea and vomiting. Anti-emetic drugs often have their own side effects and are typically ineffective with the anticipatory emesis that is often a conditioned response. General anesthesia, also, commonly causes postoperative nausea and vomiting among surgical patients. Well-designed clinical interventions have found support for hypnosis training over relaxation techniques, distraction or individual counseling in the management of nausea and the frequency and intensity of emesis. This has been the case for pediatric as well as adult cancer patients.

Although a number of factors including maternal age, number of previous deliveries and medication use can influence the duration of labor, women trained in hypnosis have experienced speedier and more satisfactory deliveries. Suggestions for relaxation, analgesia and comfort have been successful in influencing analgesic medication use and pain ratings as well as the length of time spent in the first and second stages of labor. Hypnosis trained primigravid women performed more like parous mothers than first time mothers without hypnosis on the above variables.

Numerous methodological difficulties limit confidence and widespread application of hypnotic techniques in medicine. As with most clinical research, the heterogeneity of the patients in terms of illness onset, type, course, previous treatments, co-morbidity etc., limits what can be said with confidence regarding the use of hypnosis in medicine. Also, many experimenters fail to include an adequate description of hypnosis, or might include hypnotic interventions as a component of a more complex psychosocial interventions, or fail to ascertain whether the subject was indeed hypnotized during the test. Nonetheless, qualified empirical evidence exists to support the efficacy of hypnotic techniques in the treatment of pain, asthma, dermatological disorders, irritable bowel syndrome, hemophilia, hyperemesis and obstetrics (Pinnell and Covino 2000).

3.2 Psychotherapy

The addition of hypnosis to cognitive–behavioral treatments for anxiety and habit disorders (smoking, weight control) improves the effect of traditional psychotherapeutic interventions. The ability to alter perception, adjust expectations, increase motivation and imagine more adaptive behaviors facilitates change, especially among those suffering from anxiety disorders.

Where it was once thought that autonomic nervous system hyperarousal was the result of so called stressful life events, recent research highlights the importance of cognitive processes such as perception. One person's strain might well be another's relief, depending upon the meaning that is made of the situation. An invitation from the hypnotist to perceive an event in a more benign fashion creates the opportunity for emotional relief and more adaptive behavior. Imaginative individuals can, also, use their capacity to picture frightening situations vividly to expose themselves to feared stimuli successfully. These patients are able to achieve some of the same mastery as '*in vivo* exposure' without the time and travel. Behaviors can be mentally rehearsed, problems anticipated and resolved with the help of suggestion, focused concentration and imagery. Likewise, unique motivational ideas can be the object of attention without distraction or criticism, and used to provide significant force for behavioral change.

Most efforts to use hypnosis for smoking cessation and weight control have seen modest results. Patients often report achieving relaxation and satisfaction with the technique, but pounds lost have been few and smoking abstinence is rarely achieved by more than 20 percent of program participants. This might well be due to the manner in which suggestions are employed to motivate patients and reduce appetite and desire. Anxiety symptoms related to surgeries and phobias have been especially responsive to hypnotic training, especially when combined with other psychological techniques. Much more is known these days about the complex manner in which patients cycle through processes of habit change as well as the benefits of medication support. Further work in this area would do well to integrate differing hypnotic approaches into the diverse phases and processes that people use in making habit change happen.

A number of clinicians have advocated the use of hypnosis for patients in psychotherapy. The hope in this area is to assist people to recover memories of difficult experiences as well as to teach a coping mechanism for managing extreme anxiety and dissociative symptoms. These therapists propose that the absorption, dissociation and intensity of emotions that often accompanies hypnotic trance are similar to the experiences of many victims of trauma. In a manner similar to state-dependent learning, and in much the same way that Charcot and Freud employed this technique, patients are invited to experience hypnosis, revisit the initial trauma and have an opportunity to express, integrate and control their associated emotions. In addition, those who spontaneously experience dissociative states are taught to enter into a hypnotic trance to reduce anxiety and to achieve some ability to terminate or control hyperarousal.

In a fashion that mirrors Freud's later work, the use of hypnosis as a psychotherapeutic technique is the focus of considerable controversy at the moment. In general, as was the case with Freud, the main idea of memory recovery sits in the context of credibility and veridicality. For some patients and practitioners it is painful and angering to raise the issue of truth in memory, since so many of those who were physically and sexually abused in their youth were frightened into secrecy. Yet, Freud seemed to see the inevitable intertwining of suggestion with dissociation in hypnosis as limiting the utility of this technique in psychotherapy.

Laboratory research finds that some modest gains in memory can be achieved from efforts to recall, especially when these are made proximate to the experience. Cues, reviews, relaxation, and other motivational techniques also seem to aid the recovery of memory. Nonetheless, hypnotic techniques have generally failed to improve recall of word lists and eyewitness reports. On the other hand, considerable evidence is mounting to suggest that people in clinical and experimental situations can be lead to the erroneous belief that an event or experience did indeed occur. The fact that these circumstances are more likely to take place in situations of heightened emotionality, uncertainty and unfamiliarity makes many clinicians cautious about using hypnosis as an 'uncovering' technique.

4. Essential Elements and Putative Mechanisms

At the end of such a review, it would be nice to be able to say clearly how and why hypnosis is and has been so helpful to patients in a variety of circumstances. Despite its age as a psychological intervention, this is not yet possible. A good deal of research has, thus far, failed to identify a specific neurological marker or neurophysiological substrate for hypnosis. There are, however, several ways that the beneficial effects of hypnosis have been understood.

Suggestions are ideas presented to a subject for uncritical acceptance. In hypnosis, the subject is asked to concentrate on these ideas in a manner that diminishes distraction, lessens logical analysis and maximizes acceptance and acquiesce. Suggestions are not commands, but invitations that can be rejected or accepted. However, like the response of patients to placebo medications, there is often a sense of

involuntariness to hypnotic responses and an attribution of causation to the hypnotist. The physical context, psychological expectations or the pressure to fit in well with a group can also provide direction for a person's thoughts, feelings or behaviors. Some theorists maintain that these psychosocial factors alone or in combination result in successful hypnotic response.

The ability to become absorbed in an idea as if it is real and to use fantasy and imagery to alter cognition and/or behavior is also seen as a determinant of successful hypnotic response. Such special concentration is thought to invite a state of dissociation where consciousness is divided. The ability to separate awareness of sensation, memory, time, behavior etc., is not only part of daily experience (e.g., shifting the gears of a car without awareness, daydreaming), but it can be created by hypnotists and imaginative patients. Such an ability is thought to facilitate the use of hypnosis in management of painful conditions and anxiety. Since the higher cognitive faculties are blocked from elaborating pain or anxiety signals, emotional or behavioral responses are muted or stopped.

There are a number of clinical, and perhaps nonspecific, factors that facilitate or hinder the response of patients to hypnosis. Patients must be motivated to follow the suggestions of the hypnotist. A fearful, overly defensive, unwilling person will rarely succeed with hypnosis and one who feels safe, valued, and cared for by an 'expert' most likely will. Research indicates that hypnotizability appears to be a rather normally distributed, stable, trait. Although there are some that claim that anyone can achieve some benefit from hypnotic interventions, many hypnosis experts find that those with heightened ability are the most helped. Certain conditions favor suggestion such as ambiguity, uncertainty, novelty and hyperemotionality. Good hypnotic technique endeavors to recruit or develop some of these elements.

See also: Conscious and Unconscious Processes in Cognition; Freud, Sigmund (1856–1939); Hypnosis, Neural Basis of; Hysteria; Psychoanalysis, History of; Psychological Treatment, Effectiveness of

Bibliography

Breuer J, Freud S 1893–1895 Studies on hysteria. In: Strachey J (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London, Vol. 2, pp. 1–312

Covino N A, Frankel F H 1999 Clinical hypnosis. In: Nicoli A M Jr. (ed.) *The Harvard Guide to Psychiatry*. The Harvard University Press, Cambridge, MA, pp. 552–6

Ellenberger H F 1970 *The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry*. Basic Books, New York

Gauld A 1992 *A History of Hypnotism*. Cambridge University Press, New York

Pinnell C M, Covino N A 2000 Empirical findings on the use of hypnosis in medicine: A critical review. *The International Journal of Clinical and Experimental Hypnosis* **48**: 170–94

N. A. Covino

Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of

This article will examine the role of psychological factors on the regulation of the hypothalamic–pituitary–adrenal (HPA) axis. The HPA axis regulates the primary endocrine responses to stress, which in turn regulate major metabolic and neural processes. The major principle that will be elaborated is that both activation and inhibition of the HPA axis are in part modulated by psychological factors. We will examine first specific interventions that occur during development that permanently alters the responses to stress. Second, we will discuss some of the mechanisms whereby responses to stress can be modulated behaviorally.

1. The HPA Axis

The secretion of glucocorticoids (GC) by the adrenal cortex under conditions of stress is an important mammalian mechanism of adaptation. The response to stress involves complex interactions between the neurosecretory hypothalamic neurons, the corticotrophs of the anterior pituitary, and the adrenal glands. Various neural afferents mediating the stress response converge at the parvocellular neurons of the hypothalamic parvocellular nucleus (PVN). The cell bodies of the PVN neurons are one of the primary sites of the synthesis of corticotropin-releasing factor (CRF) and project to the median eminence. Activation of the PVN neurons by a variety of different stimuli induces release of CRF into the portal capillaries through which CRF is transported to the anterior pituitary.

The binding of CRF to receptors in the anterior pituitary in turn induces the release of adrenocorticotropin (ACTH). Subsequently, ACTH in the general circulation stimulates the secretion of GC, primarily corticosterone in the rodent and cortisol in primates and humans. GC secreted by the adrenal serve a variety of regulatory and permissive functions which recruit the organism's endocrine, metabolic, immune, and behavioral responses to stress (Munck et al. 1984). Further circulating GCs participate in their own regulation by exerting a negative feedback on stress-induced activity. There are numerous negative feedback sites, which include the PVN, anterior pituitary, and possibly the hippocampus.

The HPA axis is highly responsive to changes in both the interior and exterior environment. One of the more interesting developments in this field is that there appear to be different neural pathways that respond to different stimuli that activate the PVN neurons. Currently, five different pathways have been described (Sawchenko 1991). Certain pathways are activated when visceral and interoceptive signals are involved in the activation of the HPA axis. One of the major pathways involves the structures of the principle components of the limbic region of the telencephalon, which includes the hippocampal formation, the septal, and the amygdaloid nuclear complexes.

Not only have the limbic structure been implicated as possible sites for GC feedback but they are also implicated in the processing of emotional responses (LeDoux et al. 1988). The amygdala plays a pivotal role in the 'fear circuit' described by Davis (1992). The hippocampus also appears to play an important role in the processing of affective information. Given the importance of affective stimuli in the activation of the HPA axis it is not surprising that there should exist neural pathways responsible for processing this class of stimuli.

2. *Early Experience*

After more than four decades of research using a variety of animal species it has been well established that the effects of early experience are diverse and pervasive. One early experience manipulation, early handling, modifies the HPA axis so as to permanently alter the endocrine (ACTH and GC) response to stress and to modify the neural components of the HPA axis that are involved in regulating the stress response. The early handling paradigm involves an apparently innocuous manipulation in which the experimenter removes the rat pup from its mother once daily, placing it either alone or as a group in a novel container for a few minutes and then reuniting it with the dam in the home cage. Early handling is most effective if initiated between birth and 14 days postnatally.

The early studies demonstrated that early handling profoundly altered the GC response to several different challenges. In general, as adults, the magnitude of the GC response was reduced and further, the GC response in the early handled rats returned to basal levels more rapidly indicating a more efficient negative feedback mechanism (Batnagar et al. 1996). The initial studies were limited by the technologies available at that time. With technological advances it became possible to study other components of the HPA axis. Thus, early handled animals showed a reduction in ACTH secretion following stress when compared to non-handled rats. Early handled rats also have significantly lower basal levels of CRH gene expression (Plotsky and Meaney 1993).

Another well-documented effect of early handling is to permanently modify the GC receptors in the hippocampus. In the central nervous system (CNS), two types of GC receptors are found and each functions to regulate different aspects of HPA activity (DeKloet 1991). The mineralocorticoids (MR) are involved in the tonic regulation of the HPA axis, that is, circadian rhythms. In contrast, the glucocorticoid receptor (GR) appears to be an important extrahypothalamic site of an inhibitory feedback effect on HPA activity although it is not clear that the hippocampus is a negative feedback site for GC. Occupancy of available GRs seems to be required to restrain stress induced ACTH secretion.

In an extensive series of studies Meaney and coworkers (see Batnagar et al. 1996) have established that early handled animals have an increase in hippocampal GR sites as well as increased GR mRNA. Early handling has little effect on the MRs. It has been postulated that the increase in GR density increases the sensitivity of the hippocampus to circulating GCs, enhancing the efficiency of the negative feedback inhibition of HPA activity, thus permitting early handled animals to reduce the poststress levels of ACTH and GC.

All the current information on the effects of early handling on the HPA axis indicates that manipulation permanently modifies several of the important components of neural circuitry that regulate the HPA axis. Thus, the effects of early handling can be attributed largely to alterations in the CNS mechanisms that regulate ACTH and GC secretion.

One hypothesis proposed to account for the effects of early handling is that handling affects the pup indirectly by producing changes in the behavior of the dam (the maternal mediation hypothesis). Maternal behavior parallels and compliments the need of the developing organism and the parent–offspring relationship is most intense during the preweaning period when the dam's role in regulating vital neonatal functions is fundamental. Disturbances of mother–infant relationships by brief periods of maternal separation alters the pup's behavior in a manner that would elicit increased pup-directed behavior by the dam, such as licking, grooming, retrieving, etc. This increase in pup-directed behavior could bring about pervasive changes in regulation of the HPA axis. Using a more naturalistic approach (Caldji et al. 1998) demonstrated that pups who received high levels of licking and arched-back nursing from their dams showed the characteristic pattern of HPA activity that closely resembled the early handled pups.

Plotsky and Meaney (1993) have used relatively long postnatal maternal separations as a model of adverse early experience. Thus, pups were separated from the dams for 3 hours daily beginning on day 2 and continuing through day 14. When compared to pups that had experienced short periods on maternal separation GC levels were significantly higher fol-

lowing restraint and further, basal levels of CRH gene expression in the PVN were elevated. Other investigators have used more extended periods of separation, that is, 8 hours every other day, and have also found that HPA activity is altered in the direction of a more exaggerated response.

Thus, it appears as though the HPA axis can be modified bidirectionally depending upon the length of the separation. It is possible that the signals emitted by the pup are different following brief vs. a more prolonged separation and that the effects on maternal behavior are manifested in very different ways resulting in seemingly paradoxical outcomes. The data do clearly indicate that early experiences do result in a permanent and apparently irreversible influence on the adult HPA axis regardless of the direction of these changes.

3. *Coping*

Based primarily on the initial observations of Selye (1956), it became clear that the endocrine system and more specifically increased secretion of the so-called stress hormones ACTH and GC was a critical component of the response to stress. Selye also described these responses as nonspecific. Thus, regardless of the stimulus that provoked these responses they were general and occurred nonspecifically to any or all challenges.

Mason (1968) presented a point of view which differed radically from the widely-held beliefs of the time. Mason emphasized a crucial role for 'the psychological approaches involved in the emotional or arousal reactions to threatening or unpleasant factors in the life situation as a whole.' He argued that the nonspecific response to stress is primarily behavioral or psychological in nature. However, although Mason recognized the importance of psychological factors associated with the endocrine activation following stress he did not elaborate on what some of these dimensions could be.

In an attempt to further specify the psychological aspects of the endocrine stress response Levine and Ursin (1990) invoked a cognitive approach in which the lack of information (uncertainty) resulted in the activation of the HPA axis. Uncertainty could be a function of the absence of information and/or the failure to fulfill an expected outcome or to avoid a negative one. The absence or loss of control and the lack of predictability also creates uncertainty which when processed by the brain activates the neurobiological events that result in increased HPA activation.

3.1 *Control*

One of the most critical psychological dimensions that contribute to the regulation of the HPA axis is control. In a remarkable series of experiments Weiss (1972) convincingly demonstrated the importance of control in reducing the physiological consequences of exposure to stress. Using gastric ulceration as the measure of stress it was shown that rats that could make a response that terminated an electric shock showed a marked reduction in the number of gastric ulcers when compared to a 'yoked' group that received identical shocks but were not permitted to respond. This initial experiment clearly demonstrated the importance of control. In further studies two other psychological operations also ameliorated the extent of gastric ulceration. These were predictability and feedback. These three elements, control, predictability, and feedback, all serve to reduce arousal and thus HPA activity and are postulated to contribute to the process of coping.

There are many examples of the importance of control in modifying the response of the HPA axis. Rats permitted to escape from shock showed a reduction in GC levels although they received shock on every trial (Davis 1977). In contrast repeated exposure to inescapable shock resulted in increased secretion of GC (sensitization). Hanson et al. (1976) exposed Rhesus monkeys to a loud noxious noise. One group of monkeys was allowed to control the duration of the noxious stimulus by making a lever press response to terminate the noise. A comparable group of monkeys were exposed to the identical duration of the noise, but were not permitted to regulate the duration. The monkeys that exercised control showed levels of GC that were similar to nonnoise exposed monkeys. In contrast, the 'yoked' control animals exhibited extremely high levels of GC. In this experiment the monkeys that could control the duration of the noise were re-exposed to the noise on a subsequent day but on this occasion they were not permitted to exercise control. Their GC levels were now elevated and exceeded those of the noise alone group. This finding represents an example of yet another psychological dimension involved in the regulation of HPA activity, the failure of expectancies.

3.2 *Expectancies*

Physiological activation in response to uncertainty is a central component of the model presented by Sokolov (1963). Although this model was an attempt to account for the neural mechanisms involved in the process of habituation, a broader interpretation of this model could also account for why uncertainty would result in the activation of the HPA axis. When a subject is presented with a stimulus that not been previously experienced or is unexpected an orienting response occurs. Both behavior and physiological responses are elicited by the stimulus. Amongst the number of physiological components of the orienting response is HPA activation (Hennessy and Levine 1979).

If the stimulus is repeated frequently the response tends to diminish and eventually disappears. Sokolov's

model is based on a matching system in which new stimuli or situations are compared with a representation in the CNS of prior events. This matching process results in the development of expectancies. Uncertainty can, therefore, be viewed as a violation of expectancies. From this perspective it is the evaluation of the event or events by the individual that is crucial in determining whether that event will elicit a stress response resulting in HPA activation.

(Bolles 1972) identified two distinct types of expectancies, stimulus and response outcome expectancies. Stimulus expectancies refers to the storage of information that one stimulus precedes another (classical conditioning). Response outcome expectancies results from the learning that a particular response results in a positive outcome, that is, obtaining reinforcement or avoiding a negative outcome. The development of positive outcome responses is dependent on some of the coping mechanisms previously described. Thus, if an organism can exercise control by making an active response that leads to a predictable outcome this information will be stored in the CNS as a positive outcome expectancy.

Thus, when rats can obtain reinforcement on a highly predictable schedule removal of the reinforcement (extinction) produces a marked elevation in GC. Further when rats are shifted from a highly predictable schedule to an unpredictable schedule of reinforcement GC are again elevated. In contrast when the shift in reinforcement schedules is from low predictability to high predictability GC levels significantly decrease. More recently a similar outcome has been reported in primates. Squirrel monkeys were required to obtain food in a laboratory model of foraging. When the foraging demand was high and food was available unpredictably cortisol levels were persistently elevated. When these monkeys were shifted to a low demand foraging schedule cortisol levels decreased rapidly. The effects of violations of expectancies are also observed in aversive situations. Rats that successfully learned to avoid shock show a progressive decrease in GC as their performance reached asymptote. However, if the animals are prevented from making the avoidance response, although shock is not administered, GC levels are once again elevated.

3.3 Social Factors

There is a large literature that has documented the importance of social behavior and social relationships on the HPA axis. At first glance these may appear to be a separate class of events. They can be, however, subsumed within the more general theories of stress and coping. An extensive series of studies using nonhuman primates (Levine and Coe 1988) demonstrated that at almost every point in development social variables can modulate the HPA axis, both in terms of increased activity and the inhibition of the response to environmental perturbations. Disruptions of social relationships invariably result in increased HPA activity. Conversely, the presence of familiar social partners can reduce, and at times eliminate, the normal increase in HPA activity in squirrel monkeys. One of the most robust increases in GC levels is observed when infant squirrel monkey are separated from their mothers and other members of their natal social groups. These responses occur even when the infants are separated repeatedly and show no signs of habituation. If the infant is separated from the mother and permitted to remain with their familiar social partners the response is markedly reduced. This occurs only when mother infant dyads are group housed. When the dyad is housed individually the magnitude of the GC response is similar regardless of whether the infant or the mother was removed from the home cage. This phenomenon is often called social buffering, and has been demonstrated in the adult monkeys. Thus, when adult squirrel monkeys were exposed to a natural predator (a live Boa Constrictor) placed in a wire mesh enclosure above their cage, individually housed animals show a striking elevation of plasma cortisol, whereas if the monkeys are exposed to the snake when group housed cortisol levels do not differ from basal levels. Other experiments using different paradigms have demonstrated that the presence of social partners ameliorates the HPA response to stress.

These findings are consistent with the proposed effects of control and feedback on the HPA axis. For the infant, the mother is the primary source of emotional security and serves to reduce the infant's arousal levels. Thus, the infant's capacity to achieve contact and maintain proximity to the mother is perhaps the infant's earliest experience with the use of control. For the adult, the availability of stable and familiar social relationships provides a set of predictable outcomes due, in part, to a prior history of social interactions. Therefore, the predictability of these social interactions would serve to reduce uncertainty.

4. Summary

The contemporary view of stress largely is characterized as an imposition model. The voluminous number of studies in which severely traumatic stimuli were used to elicit HPA activation supports this view. However, there are numerous examples of HPA activation being elicited by the omission of a predictable outcome (frustration) or the absence of a salient feature in the environment such as maternal loss. In both classes of events the common element appears to be the lack of information (uncertainty) related to obtaining positive outcomes or the avoidance of negative outcomes.

The view presented in this article focuses on the capacity of the brain to process information pertaining to obtaining of positive outcomes and/or the avoidance

of negative outcomes. Stimuli that induce HPA activation do so as a consequence of the lack of information (uncertainty) concerning the obtaining of positive outcomes or the avoidance of negative outcomes.

See also: Anxiety and Fear, Neural Basis of; Motivation, Neural Basis of; Neuroendocrinology; Posttraumatic Stress Disorder; Psychobiology of Stress and Early Child Development; Sexual Behavior and Maternal Functions, Neurobiology of; Stress and Cardiac Response; Stress and Coping Theories; Stress: Measurement by Self-report and Interview; Stress, Neural Basis of; Stress: Psychological Perspectives

Bibliography

Batnagar S, Shanks N, Plotsky P M, Meaney M J 1996 Hypothalamic–pituitary–adrenal response in neonatally handled and nonhandled rats: Differences in facilitatory and inhibitory neural pathways. In: McCarty R, Aguilera G, Sabba E, Kventnansky R (eds.) *Stress: Molecular Genetic and Neurobiological Advances*. Gordon and Breach, New York

Bolles R C 1972 Reinforcement, expectancy and learning. *Psychological Review* **79**: 394–409

Caldji C, Tannenbaum B, Sharma S, Francis D, Plotsky P M, Meaney M J 1998 Maternal care during infancy regulates the development of neural systems mediating the expression of fearfulness in the rat. *Proceedings National Academy of Science* **95**: 5335–40

Davis M 1992 The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience* **15**: 353–75

DeKloet E R 1991 Brain corticosteroid receptor balance and homeostatic control. *Frontiers in Neuroendocrinology* **13**(2): 95–164

Hennessy J W, Levine S 1979 Stress, arousal and the pituitary–adrenal system: A psychoendocrine model. In: Sprague J M, Epstein A N (eds.) *Progress in Psychobiology and Physiological Psychology*. Academic, New York

LeDoux J 1995 Emotions: Clues from the brain. *Annual Review of Psychology* **46**: 209–35

Levine S, Coe C L 1988 Psychosocial modulation of neuroendocrine activity. In: Pancheri P, Zichella A (eds.) *Biorhythms and Stress in the Physiopathology of Reproduction*. Hemisphere, New York

Levine S, Ursin H 1991 What is stress? In: Brown M R, Koob G F, Rivier C (eds.) *Stress: Neurobiology and Neuroendocrinology*. Marcel Decker, New York

Mason J W 1968 A review of psychoendocrine research on the pituitary–adrenocortical system. *Psychosomatic Research* **30**: 576–607

Munck A, Guyre P M, Holbrook N J 1984 Physiological functions of glucocorticoids on stress and their relation to pharmacological actions. *Endocrinological Reviews* **5**: 25–44

Plotsky P M, Meaney M J 1993 Early, postnatal experiences alters hypothalamic corticotropin-releasing factor (CRF) mRNA, medium eminence CRF content and stress induced release in adult rats. *Molecular Brain Research* **18**: 195–200

Sawchenko P E 1991 The final common path: Issues concerning the organization of central mechanisms controlling corticotropin secretion. In: Brown M R, Koob G F, River C (eds.) *Stress: Neurobiology and Neuroendocrinology*. Marcel Dekker, New York

Selye H 1956 *The Stress of Life*. McGraw-Hill, New York

Sokolov E N 1963 *Perception and the Conditioned Reflex*. Pergamon, Oxford, UK

Weiss J M 1972 Influence of psychological factors on stress induced pathology. In: Porter R, Knight J (eds.) *Physiology and Emotions in Psychosomatic Illness*. Elsevier, Amsterdam

S. Levine

Hypothalamus

1. Introduction

The hypothalamus is a small brain structure (0.4 percent of total brain volume) primarily of gray matter providing important functions which can be summarized to maintain homeostasis. From an evolutionary perspective, the hypothalamus can be considered a very old component of the central nervous system.

Historically, the first observations concerning hypothalamic functions were described for the control of the pituitary gland since endocrinological abnormalities (e.g., due to tumors) were more obvious than disturbances of autonomic functions, which are also under hypothalamic control. In 1930 Ernst and Berta Scharrer developed the concept of 'neurosecretory' hypothalamic cells derived from their morphological structure, and this dramatically stimulated subsequent research. It became obvious that the hypothalamus combines characteristics of glands and neuronal tissue. However, the focus of interest remains in investigations of hypothalamic control via the pituitary gland, which led to the discovery and chemical characterization of vasopressin and oxytocin in the early 1950s. The finding by Schally and Guillemin that the hypothalamus controls the secretion of other fundamentally important hormones such as TSH, FSH, ACTH, and LH was honored by the Nobel prize in 1977.

2. Neuroanatomical Localization and Structure

The hypothalamus is located in the diencephalon. It is laterally bounded by the subthalamus, medially by the third ventricle, rostrally by the lamina terminalis, caudally by the mesencephalon, and dorsally by the thalamus.

Although these boundaries are well defined, the anatomical structure of the hypothalamus is much more complex. From stimulation and lesion studies it is known that the hypothalamus consists of several

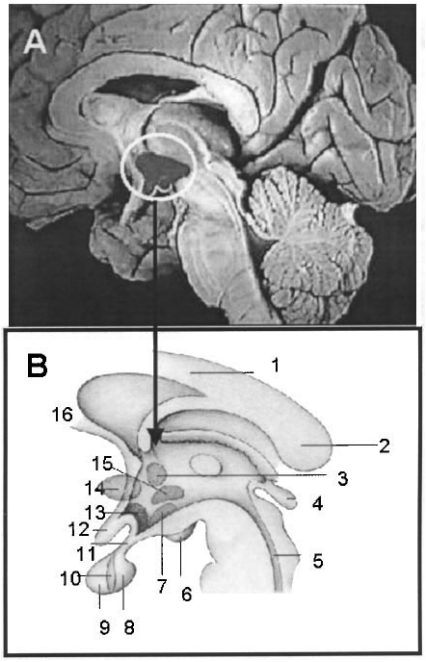

Figure 1
Localization of the hypothalamus in the brain (A) and hypothalamic nuclei (B). 1 Corpus callosum. 2 Splenium of the corpus callosum. 3 Paraventricular nuclei. 4 Pineal gland. 5 Aqueduct of mesencephalon. 6 Mammillary body. 7 Tuber cinererum. 8 Neurohypophysis. 9 Adenohypophysis. 10 Intermediate part. 11 Infundibulum. 12 Optic chiasma. 13 Supraoptic nucleus. 14 Preoptic nuclei. 15 Posterior hypothalamic nucleus. 16 Fornix.

neuronal clusters referred to as nuclei (see Fig. 1). Previous attempts at describing the anatomical relationships between these nuclei made use of a hexagon (see Fig. 2).

Based on the systematic topography of the hypothalamus given in Fig. 1, several areas can be distinguished. Most of the nuclei are located in the midline, medial, and lateral areas, which are themselves divided into supraoptic, tuberal, and mammillary sections. The preoptic nuclei are located in the preoptic region of the medial zone. The preoptic region is bounded by the septum in both directions. The anterior (supraoptic) region contains the anterior nucleus, the supraoptic nucleus, and the paraventricular nucleus. The intermediate (tuberal) region consists of the dorsomedial nucleus, the ventromedial nucleus, and the infundibular or arcuate nucleus. In the posterior region of the medial zone, the posterior nucleus and the mammillary nucleus can be identified. Finally, the lateral zone consists of the lateral hypothalamic nucleus. In the area of the periventricular zone many neurons have been identified as having

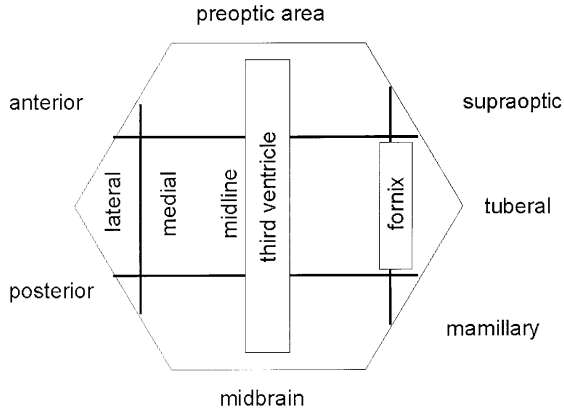

Figure 2
Schematic diagram showing the anatomical relationships between different nuclei of the hypothalamus as a hexagon (after Sladek and Sladek 1990)

fundamental importance for the secretion of releasing or inhibiting neurohormones. However, a clear distinction of the different nuclei has not been given for this area.

3. *Functions*

Although the hypothalamus can be considered a small brain structure, its functions are widespread and of fundamental importance. It is known that the hypothalamus is responsible for at least seven categories of functions. These are: the regulation of the anterior pituitary; the control of water balance; the regulation of autonomic responses; the regulation of feeding; the regulation of drinking; the regulation of reproductive functions; and the reinforcement and control of circadian functions. For an understanding of these functions, the characterization of hypothalamic inputs from other systems is essential (for a review, see Risold et al. 1997).

Retinal ganglion cell axons project through the retinohypothalamic tracts to the ventral regions of the suprachiasmatic nuclei (SCN). It is known that these nuclei couple the *biological rhythms* to the light/dark cycle. These rhythms can be divided into several categories depending on their frequency. Events repeated nearly every 24 hours are defined as circadian rhythms (e.g., changes in concentration of ACTH and cortisol). Periods with lower frequencies are considered circa-annual or infradian (e.g., season-dependent behavioral or metabolic changes or the menstrual cycle). Those with shorter frequencies (faster occurrence of periods) are defined as ultradian rhythms (e.g., respiration). The hypothalamus controls infradian and circadian rhythms. The important role of the SCN can be demonstrated by lesion studies in

which animals lost rhythmicity. The SCN projects to the paraventricular nucleus, which is connected to the spinal cord by efferent fibers. Noradrenergic fibers project to the pineal gland and affect the production and release of melatonin, concentrations of which markedly depend on the light/dark cycle. Moreover, recent studies have demonstrated that the SCN is involved in the regulation of autonomic responses (changes in heart rate) after brief stimulation by a flash of light.

Interestingly, the lateral hypothalamus seems to be a target for pheromonal stimuli derived from the main olfactory inputs. Moreover, projections from the medial amygdalar nucleus, as well as from the bed nuclei of the stria terminalis, provide accessory olfactory signals to some regions of the medial zone of the hypothalamus. This close connection to the vomeronasal organ is not fully understood with respect to its physiological function. However, two different pathways including amygdalar projections seem to be involved in the modulation of sexual and/or aggressive-defensive behavior.

Another source of 'information' is derived from viscerosensory and somatosensory signals. Most of these projections into the lateral hypothalamus are derived from the solitary tract and the parabrachial nuclei, which both receive visceral information form cranial nerves. However, there is evidence that the periventricular zone is an important target for projections providing 'visceral' signals derived from the brain stem involving catecholaminergic fibers. It is a likely assumption that these projections may provide information inducing the regulation of complex processes such as ingestive behavior. The lateral hypothalamus has been identified as a main target for spinal cord fibers which send information concerning sensations of touch or pain. The presence of fibers presenting auditory information has been demonstrated for the ventromedial nucleus of the hypothalamus. It has been suggested that these projections are important for regulation of lordosis and milk-ejection in lactating females. Somatosensory signals including nociceptive stimuli seem to reach the hypothalamus by activating neurons of the spinohypothalamic tract or involving the catecholamine system located in the brain stem. As a consequence, norepinephrine is released from the paraventricular nucleus (PVN) of the hypothalamus. In general, catecholaminergic projections to the PVN may play a pivotal role in stress responses by activating neurosecretory functions of the hypothalamus, as indicated by a reduced or blocked activation of the hypothalamus–pituitary–adrenal (HPA) axis after lesions of catecholaminergic fibers of the brain stem. There is agreement that the hypothalamus is one of the most important brain structures preparing the individual for appropriate reactions when faced with threat or other types of stress. One set of reactions concerns the cardiovascular system, which is controlled by the hypothalamus via the autonomic nervous system, especially the vagus nerve and/or the initiation of hormone secretion. The fundamental role of the hypothalamus in orchestrating multiple influences on the pituitary will be discussed later.

Hypothalamic control of thermoregulation can be considered an excellent example for maintaining homeostasis. A center responsible for heat dissipation has been identified in the preoptic area and the anterior hypothalamus. The main physiological consequences of an activation of these neurons is suppression of shivering, panting, and vasodilation of cutaneous vessels with the common goal of reducing temperature. On the other hand, conservation of temperature seems to be mediated by the posterior hypothalamic region. Reactions to maintaining or increasing temperature are shivering, vasoconstriction, and suppression of panting. In his early work Bligh (1966) suggested the presence of two regulatory systems. One is mainly responsible for maintaining the normal variation of body temperature, whereas the other becomes activated as a consequence of more serious increases or decreases in body temperature. It is known that the neurons involved are sensitive to higher (37 °C–40 °C) and lower (32 °C–36 °C) than normal temperatures. Afferent pathways transport peripheral information (e.g., cooling or heating of the skin) to these neurons, initiating the regulatory responses. However, one has to realize that the hypothalamus is not the only structure involved in thermoregulatory responses. It is known that in 'spinal' animals (brain disconnected from spinal cord) some basic thermoregulatory responses are still intact, indicating the presence of lower centers for regulation. There is agreement that thermoregulation involves several regulatory circuits acting on different levels, with the hypothalamus the most sensitive one (Satinoff 1978).

Early lesion studies had revealed that the hypothalamus mediates ingestive behavior. Distinct areas responsible for feeding have been identified in the ventromedial nucleus as a kind of satiety center, and in the lateral hypothalamus as a feeding center. Lesion of the ventromedial nucleus results in hyperphagia and polydipsia, whereas stimulation leads to the opposite behavior. There is some evidence that the paraventricular nucleus affects the vagus nerve via its projections to the brain stem, and that the lesions result in a disturbed parasympathetic control of the liver and the pancreas. A closer look reveals that after lesion of the ventromedial nucleus, hyperphagia occurs with temporary dramatic weight gain. However, after a state of obesity has been reached, ingestive behavior becomes normalized and no further weight gain can be observed, although the hypothalamic function is still disturbed. This clearly demonstrates that further brain structures are involved in the regulation of ingestive behavior.

For the lateral hypothalamus, stimulation leads to hyperphagia, and lesions lead to aphagia. However, the mechanisms involved are much more complex

since animals with lesions of the lateral hypothalamus start to eat after an appropriate period of time. Again, other brain areas such as the amygdaloid nuclei, the frontal cortex, and the substantia nigra are involved in regulation of feeding. More recently it became clear that leptin produced by fat cells is important for the regulation of ingestive behavior. Leptin receptors have been identified in the lateral hypothalamus (Hakansson et al. 1998), the supraoptic arcuate, and the paraventricular nuclei. A low quantity of leptin can be considered a signal for increased fat intake. Peripheral (derived from the gut in response to food intake), as well as central (probably produced within the hypothalamus), cholecystokinin seems to play an important role as a satiety hormone. However, neuropeptide Y, known to stimulate appetite, is found in the arcuate nucleus, the medial preoptic area, and the paraventricular nuclei of the hypothalamus. Application of neurotransmitters or neuropeptides into the paraventricular nucleus leads to preferences for different food components. While norepinephrine is accompanied by intake of fat and carbohydrates, galanin stimulates the intake of fat. Leptin suppresses both stimulations of ingestive behavior induced by NPY and by galanin. Orexin has been discovered as a protein that is exclusively produced in the lateral hypothalamus. This neuropeptide markedly increases eating behavior after injection into the hypothalamus.

The most common reason for the induction of thirst is loss of water due to sweating, urination, or respiration. This osmotic thirst, rather than the thirst caused by hypovolemia, is regulated by the hypothalamus. The basis for this is the presence of osmoreceptors in hypothalamic cells. In fact, injection of salty fluid into the hypothalamus induces drinking. Neurons responding to changes in osmotic pressure have been identified in the preoptic area, the anterior hypothalamus, the supraoptic nucleus, and the circumventricular organ. As a consequence of stimulation, AVP (vasopressin, antidiuretic hormone) is released to stop further loss of water. Moreover, increased concentrations of angiotensin II and stimulation of baroreceptors are sensitive signals to the preoptic area resulting in drinking behavior.

Another important function of the hypothalamus is coordination and control of sexual behavior. For female behavior, it is known that estradiol, a sex steroid produced during the ovulatory cycle, specifically stimulates neurons of the ventromedial hypothalamus. These neurons exhibit receptors for estrogen and progesterone, and they are mainly responsible for inducing so-called lordosis behavior. This signals the receptivity of the female to the male population by several postural and physiological signs. It can be concluded that the ventromedial hypothalamus is important for controlling changes in steroid hormone concentrations and, furthermore, that it activates several pathways at the appropriate time during the ovulatory cycle to induce the lordosis response. It is believed that estrogens can activate the hypothalamic receptors for progesterone. During pregnancy, when progesterone levels are very high, a down-regulation of these receptors appears with a decrease in sexual receptivity. The medial preoptic area of the hypothalamus is enriched with steroid-sensitive neurons, and evidence has been published that lesions in this area completely abolish male copulatory behavior in many species. The medial preoptic area is a target for pheromonal stimulation as well. Vomeronasal information reaches the amygdala, which sends signals to the medial preoptic area, indicating that this specific hypothalamic area plays a salient role in coordinating sensory and motor responses important for copulation. Both aspects are initiated by levels of testosterone which, interestingly, affect neuronal targets after aromatization into estrogen. The preoptic area of the hypothalamus is characterized by a sexual dimorphism, which may explain the differences in male vs. female sexual behavior. Taken together, the preoptic area of the hypothalamus is of fundamental importance for controlling sexual behavior, moreover, for sexual orientation as well. Interestingly, a clear-cut sexual dimorphic nucleus of the preoptic area could be identified as being larger in males than in females (Gorski et al. 1978). The reason for this dimorphism is brought about by testicular androgens (testosterone), which obviously influence the development of this specific nucleus within the preoptic area of the hypothalamus.

From animal studies, it is known that the medial preoptic area plays a key role in mediating maternal behavior (Numan 1978). Lesions in that area abolished nest-building and pup care in animals. Important pathways mediating these effects are the connections between the medial preoptic area and the ventral tegmental area. Disruptions of these connections lead to the same effects as described for lesions of the medial preoptic area. Estrogen receptors in this hypothalamic area are of major importance for regulating maternal behavior, as demonstrated by direct injections of estrogen or by suppression of this behavior after treatment with specific antagonists. Moreover, the stimulation of maternal behavior has been demonstrated for other hormones such as progesterone, oxytocin, and prolactin as well. However, more recent reports suggest the involvement of additional hypothalamic areas such as the ventromedial part for mediating maternal behavior.

A last behavioral component mediated in part by the hypothalamus is aggressive behavior. The first reports demonstrating that stimulation of the hypothalamus induces aggressive behavior were published about 70 years ago (Hess 1928). However, the precise mechanisms of these responses are still unknown. There is evidence that the lateral and frontal area of the ventromedial nucleus is of major importance. Hypothalamic aggression can be facilitated by testosterone, which indicates the initiation of offensive

Table 1
Hypothalamic-releasing hormones stimulating tropic hormones of the pituitary

Hypothalamus	Pituitary
Gonadotropin-releasing hormone (GnRH)	Follicle-stimulating hormone Luteinizing hormone
Thyrotropin-releasing hormone (TRH)	Thyroid-stimulating hormone (TSH)
Corticotropin-releasing hormone (CRH)	Adrenocorticotropic hormone (ACTH)
Somatocrinin Somatostatin	Growth hormone (GH)
Prolactin-releasing peptide Prolactin-inhibiting factor	Prolactin

rather than defensive behavioral patterns, since androgens are ineffective in defensive behavior. Other data from the literature support the view that the hypothalamic type of aggression is offensive because behavior is not suppressed by benzodiazepines or alcohol (which is not true for defensive aggression) but by serotonin agonists, which are more potent in reducing offensive behavior. Interestingly, many peripheral hormones seem to stimulate offensive aggression by affecting hypothalamic neurons. Of special interest is the finding that injections of cortisol into the hypothalamus of golden hamsters induce aggressive behavior. This may explain why stress, which activates the HPA-axis, resulting in increased levels of glucocorticoids, is frequently associated with aggression.

One of the most important functions of the hypothalamus is the orchestration of hormone regulation. Nuclei from the hypothalamus, such as the supraoptic nucleus and the paraventricular nucleus, are sources for the two hormones oxytocin and arginine-vasopressin (often called antidiuretic hormone, ADH). These hypothalamic areas elicit nerval impulses to the posterior pituitary in order to release the hormones into the circulation. The important function of vasopressin for water balance has already been discussed. Oxytocin has several functions concerning reproductive and parental behavior (see above). Furthermore, oxytocin stimulates contractions of the uterus and is involved in the milk ejection reflex. However, this hormone is present in males as well, and it has been suggested that oxytocin is of importance for social contacts and attachment.

Another fundamental influence on hormone secretion can be seen in the hypothalamic control of the anterior pituitary. The mechanism of regulating anterior pituitary hormone release is quite different from the hypothalamic influence on the posterior pituitary. While neurons from various hypothalamic nuclei transport the hormones along the axons into the posterior pituitary and release them after specific stimulation, neuroendocrine or neurosecretory cells of the hypothalamus influence the anterior pituitary. They release hormones and transport them into the medium eminence; after stimulation by action potentials, these hormones are transported into the hypothalamic pituitary portal system. These releasing hormones reach the anterior pituitary and stimulate other cells to release tropic hormones into the blood stream; these are different from the releasing hormones provided by the hypothalamus. Table 1 shows the releasing hormones from the hypothalamus, and corresponding tropic hormones derived from the pituitary.

The stimulating effects of releasing hormones derived from the hypothalamus on secretory cells of the pituitary are of fundamental importance since they regulate energy metabolism (e.g., TRH), sexual and reproductive functions (e.g., GnRH), and growth, as well as behaviors related to adaptation (e.g., CRH) and homeostasis. The precise functions of these tropic hormones secreted in the pituitary, as well as their endocrine effects at the levels of target organs, are not within the scope of this entry. However, it should be noted, as described above, that the hypothalamus is the target for many of these peripheral hormones, leading to the modulation of behavior on the one hand and providing complex negative feedback loops for controlling further secretion of hormones on the other.

4. *Summary*

The hypothalamus can be considered to be a brain structure which integrates the complex interactions between hormonal and behavioral processes with the purpose of maintaining homeostasis and survival of the individual and the species (Freeman 1995). It is involved in many different fundamental processes—not only in the regulation of metabolism and hormones, but also in feeding, drinking, reproductive, maternal, and aggressive behavior. It has therefore been suggested that one of its major functions is organization of behavior.

See also: Hunger and Eating, Neural Basis of; Hypothalamic–Pituitary–Adrenal Axis, Psychobiology

of; Neuroendocrinology; Olfactory System; Sex Hormones and their Brain Receptors; Sexual Behavior and Maternal Functions, Neurobiology of; Thirst and Drinking

Bibliography

Bligh J 1966 The thermosensitivity of the hypothalamus and thermoregulation in mammals. *Biological Reviews* **41**: 317–67

Freeman M E 1995 The hypothalamus. In: Conn P M (ed.) *Neuroscience in Medicine*. J. B. Lippincott Company, New York

Gorski R A, Gordon J H, Shryne J E, Southam A M 1978 Evidence for a morphological sex difference within the medial preoptic area of the rat brain. *Brain Research* **148**: 333–46

Hakansson M L, Brown H, Ghilardi N, Skoda R C, Meister B 1998 Leptin receptor immunoreactivity in chemical defined target neurons of the hypothalamus. *Journal of Neuroscience* **18**: 559–72

Hess W R 1928 Stammganglien-Reizversuche. *Berichte der gesamten Physiologie* **42**: 554–5

Numan M 1978 Medial preoptic area and maternal behavior in the female rat. *Journal of Comparative and Physiological Psychology* **87**: 746–59

Risold P Y, Thompson R H, Swanson L W 1997 The structural organization of connections between the hypothalamus and cerebral cortex. *Brain Research Reviews* **24**: 197–254

Satinoff E 1978 Neural organization and evolution of thermal regulation in mammals. *Science* **201**: 16–22

Sladek J R, Sladek C D 1990 Morphology of the endocrine brain, hypothalamus, and neurohypophysis. In: Becker K L (ed.) *Principles and Practice of Endocrinology and Metabolism*. J. B. Lippincott Company, New York

J. Hennig

Hypothesis Testing in Statistics

A *statistical hypothesis* is a statement about a population parameter, and the two complementary hypotheses in a hypothesis testing problem are called the *null hypothesis* and the *alternative hypothesis*. They are denoted by H_0 and H_1, respectively.

If θ denotes a population parameter, the general format of the null and alternative hypotheses is $H_0{:}\,\theta \in \Theta_0$ and $H_1{:}\,\theta \in \Theta_0^c$ where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. Typically, a hypothesis test is specified in terms of a test statistic $W(X_1, \ldots, X_n) = W(\mathbf{X})$, a function of the sample. The sample space for W is partitioned into the *rejection region*, R, and its complement, the acceptance region. If $W \in R$ is observed, the null hypothesis H_0 is rejected and the decision is made that H_1 is true. If $W \notin R$, H_0 is accepted as true.

For example, in a study to examine the factors that influence a student's success at completing a four-year college, one factor of interest might be socioeconomic status (SES). If, for simplicity, we just have two groups (high SES and low SES), and the success probabilities of the respective groups are p_h and p_l, then a hypothesis test of interest may be $H_0{:}\,p_h = p_l$ vs. $H_1{:}\,p_h > p_l$.

This is an example of a one-sided test, in which the alternative hypothesis specifies a direction. In contrast, we may consider the two-sided test $H_0{:}\,p_h = p_l$ vs. $H_1{:}\,p_h \neq p_l$ in which no direction is specified in the alternative. If we let $\hat{p}_h$ and $\hat{p}_l$ denote the observed success rates in the two SES classes, the one-sided test would reject the null hypothesis if $\hat{p}_h - \hat{p}_l$ is big, while the two-sided test would reject the null hypothesis if $\hat{p}_h - \hat{p}_l$ is either big or small.

A hypothesis test of $H_0{:}\,\theta \in \Theta_0$ versus $H_1{:}\,\theta \in \Theta_0^c$ might make one of two types of errors. If $\theta \in \Theta_0$ but the hypothesis test incorrectly decides to reject H_0, then the test has made a *Type I error*. If, on the other hand, $\theta \in \Theta_0^c$ but the test decides to accept H_0, a *Type II error* has been made. In our example, in testing $H_0{:}\,p_h = p_l$ versus $H_1{:}\,p_h > p_l$, if we conclude that $p_h > p_l$ but, in fact, $p_h \leqslant p_l$ we have made a Type I error.

1. Constructing Tests

There are many methods of deriving test statistics for a hypothesis test, a few of which follow.

1.1 Likelihood Ratio Tests

The likelihood ratio method of hypothesis testing is related to maximum likelihood estimators (MLE) (discussed in *Estimation: Point and Interval*). Suppose that we have a sample $X_1, \ldots, X_n$, where the X_i are independent, each with distribution $f(x \mid \theta)$. The observed values of these random variables are denoted by $\mathbf{x} = (x_1, \ldots, x_n)$, with likelihood function $L(\theta \mid \mathbf{x})$ given by

$$L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{1}$$

The *likelihood ratio test statistic* for testing $H_0{:}\,\theta \in \Theta_0$ vs. $H_1{:}\,\theta \in \Theta_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta \mid \mathbf{x})}{\sup_{\Theta_0} L(\theta \mid \mathbf{x})} \tag{2}$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{\mathbf{x}{:}\,\lambda(\mathbf{x}) \leqslant k\}$, where k is any number satisfying $0 \leqslant k \leqslant 1$.

If we interpret the likelihood function as measuring how likely the values of θ are, then we see that the LRT

is comparing the plausibility of the θ values in the null hypothesis with those in the alternative. Small values of the LRT statistic are then interpreted as being evidence against H_0, and lead to rejection.

If the null hypothesis consists of a single value θ_0, and the alternative is everything else, then the LRT statistic is simply $\lambda = L(\theta_0 | \mathbf{x})/L(\hat{\theta} | \mathbf{x})$, where $\hat{\theta}$ is the MLE of θ.

Example. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, 1)$ population. The LRT statistic for testing H_0: $\theta = \theta_0$ versus $H_1: \theta \neq \theta_0$ is

$$\lambda(\mathbf{x}) = \frac{L(\theta_0 | \mathbf{x})}{L(\bar{x} | \mathbf{x})} = \frac{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n} (x_i - \theta_0)^2/2]}{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n} (x_i - \bar{x})^2/2]} \tag{3}$$

If $T(\mathbf{X})$ is a sufficient statistic for θ then, as with maximum likelihood estimators, the LRT statistic is a function of T. That is, $\lambda(\mathbf{x})$ depends on $\mathbf{x}$ only through $T(\mathbf{x})$.

1.2 Bayesian Tests

The Bayesian paradigm prescribes that the sample information be combined with the prior information using Bayes' Theorem to obtain the posterior distribution $\pi(\theta | \mathbf{x})$. All inferences about θ are now based on the posterior distribution. In a hypothesis-testing problem, the posterior distribution may be used to calculate the probabilities that H_0 and H_1 are true.

One way a Bayesian hypothesis tester may choose to use the posterior distribution is to decide to accept H_0 as true if

$$\frac{P(\theta \in \Theta_0 | \mathbf{X})}{P(\theta \in \Theta_0^c | \mathbf{X})} \geqslant k \tag{4}$$

for some constant k, and to reject H_0 otherwise. Equivalently, we can reject H_0 if $P(\theta \in \Theta_0^c | \mathbf{X})$ is greater than a specified number.

Example. Let $X_1, \ldots, X_n$ be independent identically distributed (i.i.d.) $\mathrm{n}(\theta, \sigma^2)$ and let the prior distribution on θ be $\mathrm{n}(\mu, \tau^2)$ where σ^2, μ, and τ^2 are known. Consider testing $H_0: \theta \leqslant \theta_0$ vs. $H_1: \theta > \theta_0$ where we decide to accept H_0 if $P(\theta \in \Theta_0 | \mathbf{X}) \geqslant P(\theta \in \Theta_0^c | \mathbf{X})$. After some calculation, we find that H_0 will be accepted as true if

$$\bar{X} \leqslant \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2} \tag{5}$$

1.3 Union–Intersection and Intersection–Union Tests

In some situations, tests for complicated null hypotheses can be developed from tests for simpler null hypotheses. The union–intersection method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection, say $H_0: \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma$, where Γ is an arbitrary index set. If tests are available for each of the problems of testing $H_{0\gamma}: \theta \in \Theta_\gamma$ vs. $H_{1\gamma}: \theta \in \Theta_\gamma^c$ where the rejection region for the test of $H_{0\gamma}$ is $\{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\}$, then the rejection region for the union–intersection test is

$$\bigcup_{\gamma \in \Gamma} \{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\} \tag{6}$$

The rationale is that if any one of the hypotheses $H_{0\gamma}$ is rejected, then H_0 must also be rejected.

A complementary method, the intersection–union method, may be useful if the null hypothesis is conveniently expressed as a union. Suppose we wish to test the null hypothesis $H_0: \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma$, and $\{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\}$ is the rejection region for a test of $H_{0\gamma}: \theta \in \Theta_\gamma$ vs. $H_{1\gamma}: \theta \in \Theta_\gamma^c$. Then the rejection region for the intersection–union test of H_0 versus H_1 is

$$\bigcap_{\gamma \in \Gamma} \{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\} \tag{7}$$

and H_0 is false if and only if all of the $H_{0\gamma}$ are false, so H_0 can be rejected if and only if each of the individual hypotheses $H_{0\gamma}$ can be rejected.

Example. Returning to our example about success in four-year schools, we might argue that there are measures of success other than finishing. For example, yearly income could be used to measure success, and denoting mean incomes in the high and low SES groups by θ_h and θ_l respectively results in the hypothesis test

$$H_0: \theta_\mathrm{h} \leqslant \theta_\mathrm{l} \text{ or } p_\mathrm{h} \leqslant p_\mathrm{l} \quad \text{vs.} \quad H_1: \theta_\mathrm{h} > \theta_\mathrm{l} \text{ and } p_\mathrm{h} > p_\mathrm{l} \tag{8}$$

where the groups are considered different only if H_1 is accepted.

If we have estimators $\hat{\theta}_\mathrm{h}$, $\hat{\theta}_\mathrm{l}$, $\hat{p}_\mathrm{h}$ and $\hat{p}_\mathrm{l}$, a rejection region for the intersection–union test could be given by

$$\{(\hat{\theta}_\mathrm{h}, \hat{\theta}_\mathrm{l}, \hat{p}_\mathrm{h}, \hat{p}_\mathrm{l}): \hat{\theta}_\mathrm{h} - \hat{\theta}_\mathrm{l} > k_1 \quad \text{and} \quad \hat{p}_\mathrm{h} - \hat{p}_\mathrm{l} > k_2\} \tag{9}$$

Thus the intersection–union test decides that the high and low SES groups have different levels of success, that is, H_1 is true, if and only if it decides that each of the individual parameters are different.

The topic of acceptance sampling, in which one tests whether to accept a product based on a collection of

measurements, provides an extremely useful application of an intersection–union test (see Berger 1982).

There are many other methods available for constructing hypothesis tests, methods based on invariance, pivots, robust, or large sample arguments, to name a few. For more on hypothesis testing see Lehmann (1986).

2. Evaluating Tests

As mentioned previously, a hypothesis test might make one of two different kinds of errors. The probabilities of making these errors can be calculated using the power function. If R denotes the rejection region for a test, the power function is

$$P_\theta(\mathbf{X} \in R) = \begin{cases} \text{probability of a Type I error} & \text{if } \theta \in \Theta_0 \\ 1 - \text{the probability of a Type II error} & \text{if } \theta \in \Theta_0^c \end{cases} \tag{10}$$

A good test has power function near one for most $\theta \in \Theta_0^c$ and near zero for most $\theta \in \Theta_0$.

Example. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, σ^2 known. The likelihood ratio test of $H_0: \theta \leqslant \theta_0$ vs. $H_1: \theta > \theta_0$ rejects H_0 if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > k$ and has power function

$$P_\theta(\mathbf{X} \in R) = P\left(Z > k + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} \right) \tag{11}$$

where Z is a standard normal random variable.

After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way. One method of reporting the results of a hypothesis test is to report the size $(\sup_{\theta \in \Theta_0} P_\theta(X \in R))$, α, of the test used and the decision to reject H_0 or accept H_0. The size of the test carries important information. If α is small, the decision to reject H_0 is fairly convincing, but if α is large, the decision to reject H_0 is not very convincing because the test has a large probability of incorrectly making that decision.

Another way of reporting the results of a hypothesis test, one that is data dependent, is to report the p-value. Typically, not one but an entire class of tests are constructed, a different test being defined for each value of α. The p-value for the sample point $\mathbf{x}$ is the smallest value of α for which this sample point will lead to rejection of H_0 (see *Significance, Tests of*).

Because rejection of H_0 using a test with small size is more convincing evidence that H_1 is true than rejection of H_0 with a test with large size, the interpretation of p-values goes in the same way. The smaller the p-value the stronger the sample evidence that H_1 is true.

Many other types of evaluations of test can be done. The theory of most powerful tests shows how to construct best tests under a variety of conditions (see Lehmann 1986 or Casella and Berger 2001, Chap. 8). Hypothesis tests can also be evaluated using risk functions (see *Decision Theory: Classical*), as in Hwang et al. (1992).

3. Asymptotics

For the likelihood ratio test statistic (2), the following general theorem allow us to construct a large sample test.

Theorem 1. *Let $X_1, \ldots, X_n$ be a random sample from a probability distribution function or probability mean function $f(x \mid \theta)$. Under some regularity conditions on the model $f(x \mid \theta)$, if $\theta \in \Theta_0$ then the distribution of the statistic $-2 \log \lambda(\mathbf{X})$ converges to a chi-squared distribution as the sample size $n \to \infty$. The degrees of freedom of the limiting distribution is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\theta \in \Theta$.*

(The 'regularity conditions' are concerned mainly with the existence and behavior of the derivatives (with respect to the parameter) of the likelihood function, and the support of the distribution (it cannot depend on the parameter). See Casella and Berger (2001, Sect. 10.6) or Lehmann (1986, Sect. 8.8) for precise conditions.)

Rejection of $H_0: \theta \in \Theta_0$ for small values of $\lambda(\mathbf{X})$ is equivalent to rejection for large values of $-2 \log \lambda(\mathbf{X})$. Thus,

$$H_0 \text{ is rejected if and only if } -2 \log \lambda(\mathbf{X}) \geqslant \chi^2_{\nu,\alpha}$$

where ν is the degrees of freedom specified in Theorem 1.

As might be expected, Theorem 1 has wide applicability. In particular, it is extremely useful in categorical data analysis (see *Multivariate Analysis: Discrete Variables (Loglinear Models)*).

Other large-sample test constructions are based on asymptotic normality of a point estimator (see *Estimation: Point and Interval*). Suppose we wish to test a hypothesis about a real-valued parameter θ, and $W_n = W(X_1, \ldots, X_n)$ is a point estimator of θ, based on a sample of size n, that satisfies

$$\frac{W_n - \theta}{\sqrt{\mathrm{Var}\, W_n}} \to Z$$

where Var W_n is the variance of W_n and Z is a standard normal random variable. We now have the basis for an approximate test, for example, we would reject $H_0: \theta \leqslant \theta_0$ at level 0.05 if $(W_n - \theta_0)/\sqrt{\text{Var}\, W_n} > 1.645$. Note that Var W_n could depend on θ_0 and we can still use it in the test statistic. This type of test, where we use the actual variance of W_n, is called a *score test*.

If Var W_n also depends on unknown parameters we could look for an estimate S_n^2 of Var W_n with the property that $(\text{Var}\, W_n)/S_n^2$ converges in probability to one. Then, using Slutsky's Theorem (see Casella and Berger 2001, Sect. 5.5), we can deduce that $(W_n - \theta)/S_n$ also converges in distribution to a standard normal distribution. The large-sample test based on this statistic is called a *Wald test*.

4. Conclusions

Hypothesis testing is one of the most widely used, and some may say abused, methodologies in statistics. Formally, the hypotheses are specified, an α-level is chosen, a test statistic is calculated, and it is reported whether H_0 or H_1 is accepted. In practice, it may happen that hypotheses are suggested by the data, the choice of α-level may be ignored, more than one test statistic is calculated, and many modifications to the formal procedure may be made. Most of these modifications cause bias and can invalidate the method. For example, a hypothesis suggested by the data is likely to be one that has 'stood out' for some reason, and hence H_1 is likely to be accepted unless the bias is corrected for (using something like Scheffe's method—see Hsu 1996).

Perhaps the most serious criticism of hypothesis testing is the fact that, formally, it can only be reported that either H_0 or H_1 is accepted at the prechosen α-level. Thus, the same conclusion is reached if the test statistic only barely rejects H_0 and if it rejects H_0 resoundingly. Many feel that this is important information that should be reported, and thus it is almost required to also report the p-value of the hypothesis test.

For further details on hypothesis testing see the classic book by Lehmann (1986). Introductions are also provided by Casella and Berger (2001) or Schervish (1995), and a good introduction to multiple comparisons is Hsu (1996); see also *Hypothesis Tests, Multiplicity of*.

See also: Explanation: Conceptions in the Social Sciences; Hypothesis Testing: Methodology and Limitations

Bibliography

Berger R L 1982 Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**: 295–300

Casella G, Berger R L 2001 *Statistical Inference*, 2nd edn. Wordsworth/Brooks Cole, Pacific Grove, CA

Hsu J C 1996 *Multiple Comparisons, Theory and Methods*. Chapman & Hall, London

Hwang J T, Casella G, Robert C, Wells M T, Farrell R H 1992 Estimation of accuracy in testing. *Annals of Statistics* **20**: 490–509

Lehmann E L 1986 *Testing Statistical Hypotheses*, 2nd edn. Springer, New York

Schervish M 1995 *Theory of Statistics*. Springer, New York

G. Casella and R. L. Berger

Hypothesis Testing: Methodology and Limitations

Hypothesis tests are part of the basic methodological toolkit of social and behavioral scientists. The philosophical and practical debates underlying their application are, however, often neglected. The fruitful application of hypothesis testing can benefit from a clear insight into, the underlying concepts and their limitations.

1. The Idea of Hypothesis Testing

A test is a statistical procedure to obtain a statement on the truth of falsity of a proposition, on the basis of empirical evidence. This is done within the context of a model, in which the fallibility or variability of this empirical evidence is represented by probability. In this model, the evidence is summarized in observed data, which is assumed to be the outcome of a stochastic, i.e., probabilistic, process; the tested proposition is represented as a property of the probability distribution of the observed data.

1.1 Some History

The first published statistical test was by John Arbuthnot in 1710, who wondered about the fact that in human births, the fraction of boys born year after year appears to be slightly larger than the fraction of girls (cf. Hacking 1965). He calculated that this empirical fact would be exceedingly unlikely (he obtained a probability of 1/483 600 000 000 000 000 000 000) if the probability of a male birth were exactly 0.5, and argued that this was a proof of divine providence, since boys—some of whom will be soldiers—have a higher risk of an early death, so that a higher ratio of male births is needed to obtain an equal ratio of males among young adults. We see here the basic elements of a test: the proposition

that the male birth ratio is 0.5, related to data by regarding these as outcomes of stochastic variables, the calculation that the data would be unlikely if the proposition were true, and a further conclusion interpreting the falsity of the proposition.

One of the first statistical procedures that comes close to a test in the modern sense was proposed by Karl Pearson in 1900. This was the famous chi-squared test for comparing an observed frequency distribution to a theoretically assumed distribution. Pearson derived the now well-known statistic to test the proposition, or *hypothesis*, that the probabilities of the various possible (finitely many) outcomes of some random variable are given by certain preassigned numbers. Pearson proved that this statistic has (in a large-sample approximation) the chi-squared distribution; this distribution can therefore be used to calculate the probability that, if the hypothesis holds, the test statistic will assume a value equal to or larger than the value actually observed.

The idea of testing was further codified and elaborated in the first decades of the twentieth century, mainly by R. A. Fisher (e.g., 1925). In his significance tests the data are regarded as the outcome of a random variable X (usually a vector or matrix), which has a probability distribution which is a member of some family of distributions; the tested hypothesis, also called the null hypothesis, is an assertion which defines a subset of this family; a test statistic $T = t(X)$, which is a function of X, is used to indicate the degree to which the data deviate from the null hypothesis; and the significance of the given outcome of the test statistic is calculated as the probability, if the null hypothesis is true, to obtain a value of T which is at least as high as the given outcome. If the probability distribution of T is not uniquely determined by the null hypothesis, then the significance is the maximal value of this probability, for all distributions of T compatible with the null hypothesis. The significance probability is now often called the p-value (the letter p referring to probability). With Fisher originates the convention to consider a statistical testing result as 'significant' if the significance probability is 0.05 or less—but Fisher was the first to recognize the arbitrary nature of this threshold.

A competing approach was proposed in 1928 by J. Neyman and Egon Pearson (the son of Karl). They criticized the arbitrariness in Fisher's choice of the test statistic and asserted that for a rational choice of test statistic one needs not only a null hypothesis but also an alternative hypothesis, which embodies a proposition that competes with the proposition represented by the null hypothesis. They formalized the testing problem as a two decision problem. Denoting the null hypothesis by H_0 and the alternative H_1, the two decisions were represented as 'reject H_0' and 'do not reject H_0.' (The second decision was also called 'accept H_0,' but it will be discussed below that this is an unfortunate term.) Two errors are possible: rejecting a true H_0, and failing to reject a false H_0. Neyman and Pearson conceived of the null hypothesis as a standard situation, the burden of proof residing with the researcher to demonstrate (if possible) the untenability of this proposition.

Correspondingly, they called the error of rejecting a true H_0 an error of the first kind and the error of failing to reject a false H_0 an error of the second kind. Errors of the first kind are considered more serious than errors of the second kind. The probability of—correctly—rejecting H_0 if H_1 is true, which is 1 minus the probability of an error of the second kind, given that the alternative hypothesis is true, they called the power of the test. Neyman and Pearson proposed the requirement that the probability of an error of the first kind, given that the null hypothesis is indeed true, do not exceed some threshold value called the significance level usually denoted by α. Further they proposed to determine the test so that, under this essential condition, the power will be maximal.

In the Neyman–Pearson formulation, we obtain richer results (namely, specific precepts for constructing good tests) at the cost of a more demanding model. In addition to Fisher's null hypothesis, we also need to specify an alternative hypothesis; and we must conceive the testing problem as a two-decision situation. This led to vehement debate between Fisher on the one hand, and Neyman and E. Pearson on the other. This debate and the different philosophical positions are summarized by Hacking (1965) and Gigerenzer et al. (1989), who also give a further historical account. The latter study also discusses how this controversy was resolved in the teaching and practice of statistics in the social sciences by a hybrid theory which combines elements of both approaches to testing, and which has been treated often as an objective precept for the scientific method, glossing over the philosophical controversies.

Examples of this hybrid character are that, in accordance with the Neyman–Pearson approach, the theory is explained by making references to both the null and the alternative hypotheses, and to errors of the first and second kind (although power tends to be treated in a limited and often merely theoretical way), whereas—in the spirit of Fisher—statistical tests are regarded as procedures to give evidence about the particular hypothesis tested and not merely as rules of behavior that will in the long run have certain (perhaps optimal) error rates when applied to large numbers of hypotheses and data sets. Lehmann (1993) argues that indeed a unified formulation is possible, combining the best features of both approaches.

Instead of implementing the hypothesis test as a 'reject/don't reject' decision with a predetermined significance level, another approach often is followed: to report the p-value or significance probability, defined as the smallest value of the significance level at which the observed outcome would lead to rejection of the null hypothesis. Equivalently, this can be defined

as the probability, calculated under the null hypothesis, of observing a result deviating from the null hypothesis at least as much as the actually observed result. This deviation is measured by the test statistic, and the p-value is just the tail probability of the test statistic. For a given significance level α, the null hypothesis is rejected if and only if $p \leqslant \alpha$.

2. Hypothesis Tests in Empirical Research

2.1 The Role of the Null Hypothesis

In the social and behavioral sciences, the standard use of hypothesis testing is directed at single research questions, practically always fragments of larger investigations, about the existence of some or other effect. This effect could be a difference between two group means, the effect of some explanatory variable on some dependent variable as expressed by a regression coefficient in multiple linear regression analysis, etc. Expressed schematically, the researcher would like to demonstrate a research hypothesis which states that the effect in question, indeed, exists. The negation of this hypothesis, then, is understood as the proposition that the effect is absent, and this proposition is put up as the null hypothesis. The research hypothesis, stating that the effect exists, is the alternative hypothesis. Rejecting the null hypothesis is interpreted as support for the existence of the hypothesized effect. In this way, the burden of proof rests with the researcher in the sense that an error of the first kind is to state that there is evidence supporting the existence of the effect if, in reality, the effect is absent. The usual welcoming phrase for the rejection of the null hypothesis is the statement that a significant result has been obtained.

2.2 Example: The t-test

Consider, as a major example, the proposition that two groups, arbitrarily labeled A and B, are different with respect to some numerical characteristic X. Available data are measurements X_{Ai} of this characteristic for some individuals i in group A and measurements X_{Bi} of the characteristics for other individuals i in group B. The first step in modeling this is to consider the observed values X as outcomes of random variables, usually with the assumption that they are stochastically independent and have a probability distribution not depending on the individual i. The next step is, usually, to focus on the expected values, i.e., population means in the two groups, denoted here by μ_A and μ_B. The tested proposition is formalized as the statement that μ_A and μ_B are different (two-sided alternative hypothesis) or as the statement that one (e.g., μ_A) is bigger than the other (e.g., μ_B) (one-sided alternative hypothesis). The null hypothesis H_0 is defined as the statement that μ_A and μ_B are equal.

The most commonly used test for this testing problem is *Student's t-test*, called after the pseudonym of W. Gosset, who laid the mathematical basis for this test in 1908. The test statistic is

$$T = \frac{M_A - M_B}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) S^2}}$$

where M_A and M_B are the two sample means, S^2 is the pooled within-group variance, and n_A and n_B are the two sample sizes. (Definitions of these quantities can be found in any statistics textbook.) This formula illustrates the property of many test statistics that an observed effect (here the difference between the two sample means) is compared to a measure of variability (based here on the within-group variance). Student/Gosset showed that, if the variable X has a normal distribution with the same variance in the two groups, then T has, if the null hypothesis is true, the so-called t distribution on $df = n_A + n_B - 2$ degrees of freedom. For the two-sided alternative hypothesis, H_0 is rejected for large values of the absolute value of T, for the one-sided hypothesis H_0 is rejected for large values of T itself. The threshold beyond which H_0 is rejected is called critical value, and is determined from the t-distribution in such a way that the significance level has the pre-assigned value. The one-sided threshold at significance level α is denoted $t_{df;\alpha}$, so that the 'rejection region' for the one-sided t-test is given by $\{T > t_{df;\alpha}\}$. The power of the one-sided test is larger than that of the two-sided test for $\mu_A > \mu_B$ (corresponding to the one-sided alternative hypothesis) and smaller for $\mu_A < \mu_B$. This is, in the Neyman–Pearson formulation, the reason for using the one-sided test if the alternative hypothesis is one-sided (in which case parameter values $\mu_A < \mu_B$ are not taken into consideration).

The conventional value of the significance level in the social and behavioral sciences is 0.05. For large values of the combined sample size $n_A + n_B$, the critical value of the t-test approximates the critical value that can be computed from the standard normal distribution. This is because the fact that the variance is estimated and not known beforehand becomes less and less important as the combined sample size gets larger; if the population variance was known beforehand and substituted for S^2, the test statistic would have the standard normal distribution, and the test would be the so-called z-test. For this test, the critical value is 1.645 for the one-sided test and 1.960 for the two-sided test. The power depends on many quantities: the sample sizes, the population means and variances, and the significance level. As examples, the power of the one-sided z-test for $\alpha = 0.05$ is equal to 0.50 for

$$\frac{\mu_A - \mu_B}{\sigma} = 1.645\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

where σ is the within-group standard deviation. The power is equal to 0.95 if $(\mu_A - \mu_B)/\sigma$ is equal to twice this value.

2.3 The Role of Assumptions

The probability statements that are required for statistical tests do not come for free, but are based on certain assumptions about the observations used for the test. In the two-sample t-test, the assumptions are that the observations of different individuals are outcomes of statistically independent, normally distributed, random variables, with the same expected value for all individuals within the same group, and the same variance for all individuals in both groups. Such assumptions are not automatically satisfied, and for some assumptions it may be doubted whether they are ever satisfied exactly. The null hypothesis H_0 and alternative hypothesis H_1 are statements which, strictly speaking, imply these assumptions, and which therefore are not each other's complement. There is a third possibility: the assumptions are invalid, and neither H_0 nor H_1 is true. The sensitivity of the probabilistic properties of a test to these assumptions is referred to as the lack of robustness of the test. The focus of robustness studies has been on the assumptions of the null hypothesis and the sensitivity of the probability of an error of the first kind to these assumptions, but studies on robustness for deviations from assumptions of the alternative hypothesis have also been done, cf. Wilcox (1998).

One general conclusion from robustness studies is that tests are extremely sensitive to the independence assumptions made. Fortunately, those assumptions are often under control of the researcher through the choice of the experimental or observational design. Traditional departures from independent observations are multivariate observations and within-subject repeated measures designs, and the statistical literature abounds with methods for such kinds of dependent observations. More recently, methods have been developed for clustered observations (e.g., individual respondents clustered within groups) under the names of multilevel analysis and hierarchical linear modeling.

Another general conclusion is that properties of tests derived under the assumption of normal distributions, such as the t-test, can be quite sensitive to outliers, i.e., single, or a few, observations that deviate strongly from the bulk of the observations. Since the occurrence of outliers has a very low probability under normal distributions, they are 'assumed away' by the normality assumption. The lack of robustness and sensitivity to outliers have led to three main developments.

First, there are non-parametric tests, which do not assume parametric families of distributions (such as the normal). Most of these are based on the ranks of the observations rather than their numerical values. They are standard fare in statistical textbooks. Second, robust tests have been developed, based on numerical values of the observations, which are less sensitive to outliers or heavy-tailed distributions, e.g., by some kind of automatic downweighting of outlying observations (e.g., Wilcox, 1998). The purpose of such tests is to retain a high power as much as possible while decreasing the sensitivity to deviations from the normal distribution or to departures from other assumptions.

Third, diagnostic means have been developed to spot single observations, or groups of observations, with an undue influence on the result of the statistical procedure (e.g., Cook and Weisberg 1999, Fox 1991). The idea behind most of these diagnostics is that most of the observations come from a distribution which corresponds well enough to the assumptions of the statistical model, but that the observations could be contaminated by a small number of poorly fitting observations. Ideally, these observations should also be recognizable by close inspection of the data or data collection process. After deleting such poorly fitting observations one can proceed with the more traditional statistical procedures, assuming that the remaining data do conform to the model assumptions.

3. Confidence Intervals

Confidence intervals are closely related to hypothesis tests. They focus on a parameter in the statistical model. Examples of such parameters are, in the two-sample situation described above, the difference of the two population means, $\mu_A - \mu_B$, or the within-group standard deviation, σ. To frame this in general terms, consider a one-dimensional statistical parameter denoted by θ. A null hypothesis test is about the question whether the data is compatible with the hypothesis that the parameter θ (in this case $\mu_A - \mu_B$) has the particular value 0. Instead, one could put forward the question: what are the values of θ with which the data is compatible? This question can be related to hypothesis testing by considering the auxiliary problem of testing the null hypothesis $H_1: \theta = \theta_0$ against the alternative hypothesis $H_1: \theta \neq \theta_0$, for an arbitrary but fixed value θ_0. The data may be said to be compatible with any value of θ_0 for which this null hypothesis would not be rejected at the given level of significance. Thus, the confidence interval is the interval of non-rejected null hypotheses.

Another way of viewing the confidence interval is as follows. A confidence coefficient $1-\alpha$ is determined by the researcher (a conventional value is 0.95). The confidence interval is an interval with lower boundary L and upper boundary U, calculated from the data and therefore being random variables, with the property that the probability that $L \leqslant \theta \leqslant U$, i.e., the

interval contains the true parameter value, is at least $1-\alpha$. It can be proved mathematically that the interval of non-rejected null hypotheses has precisely this property.

The confidence interval can be inverted to yield a hypothesis test for any null hypothesis of the form H_0: $\theta = \theta_0$ in the following way. For the most usual null hypothesis that $\theta = 0$, the hypothesis is rejected if the value 0 is not included in the confidence interval. More generally, the null hypothesis $H_0: \theta = \theta_0$ is rejected whenever θ_0 is outside the confidence interval. This shows that a confidence interval is strictly more informative than a single hypothesis test.

4. Problems in the Interpretation and Use of Hypothesis Tests

While hypothesis tests have been applied routinely by hosts of social and behavioral scientists, there have been continuing debates, sometimes vehement and polemic, about their use, among those inclined to philosophical or foundational discussion. Many contentious issues can be found in Harlow et al. (1997), Nickerson (2000), and Wilkinson and TFSI (1999). Within the context of the present contribution I can only give a personally colored discussion of some main points which might contribute to a better understanding and a more judicious application of hypothesis tests.

The reason that there has been so much criticism of the nature and use of hypothesis tests, is in my view the difficulty of reasoning with uncertain evidence and the natural human tendency to take recourse to behavioral rules instead of cogitate afresh, again and again, about how to proceed. The empirical researcher would like to conclude unequivocally whether a theory is true or false, whether an effect is present or absent. However, experimental and observational data on human behavior are bound to be so variable that the evidence produced by these data is uncertain. Our human tendency is to reason—if only provisionally—as if our conclusions are certain rather than tentative; to be concise rather than accurate in reporting our arguments and findings; and to use simple rules for steering how we make inference from data to theories. We are tempted to reason as if '$p > \alpha$' implies that the investigated effect is nil, and the presumed theory is false, while '$p \leqslant \alpha$' proves that the effect indeed exists and the theory is true.

However, such cognitive shortcuts amount to serious misuse of hypothesis testing. Much of the criticism of the use of hypothesis testing therefore is well-founded. In my view this implies that better interpretations of hypothesis testing should be promoted, and that tests should be used less mechanically and combined with other argumentations and other statistical procedures. In any case, the user of tests (or other statistical procedures) should realize that these procedures are based on the assumption that there is random variability in the data which cannot be completely filtered out of the results. Whether a test result is significant or not depends partly on chance, and each researcher obtaining a significant result should be aware that this could be an error of the first kind, just as a non-significant result could be an error of the second kind.

4.1 Some Misinterpretations

The following list tries to correct misinterpretations that have haunted applications of statistical tests. A more extensive discussion can be found, e.g., in Nickerson (2000).

(a) A common misinterpretation is that non-rejection implies support for the null hypothesis. Nonrejection should be interpreted, however, as an undecided outcome: there is not enough evidence against the null hypothesis, but this does not imply that there is evidence for the null hypothesis. Maybe the sample size is small, or error variability is large, so that the data does not contain much information anyway. Usually the alternative hypothesis contains probability distributions that approximate the null distribution; e.g. when testing $\mu_A - \mu_B = 0$ against $\mu_A - \mu_B \neq 0$, the difference between the two population means can be tiny, while the alternative hypothesis is still true. In other words, the power of practically all tests is quite low in a sizeable part of the alternative hypothesis, so if one of the distributions in this part would prevail, chances would be low of rejecting the null hypothesis. Therefore, nonrejection provides support for those parts of the alternative practically as strongly as for the null hypothesis itself, and non-rejection may not be interpreted as support for the null hypothesis and against the alternative hypothesis.

(b) If one wishes to get more information about whether a nonsignificant result provides support for the null hypothesis, a power study is not the answer. Statistical power is the probability to reject the null hypothesis, if a given effect size obtains. Its interpretation cannot be inverted as being the degree of support of the null hypothesis in the case of nonsignificance. Power studies are important in the stage of planning a study. Once the study has been conducted, and one wishes to see in more detail to what extent the null and alternative hypotheses are supported, a confidence interval is the appropriate procedure (also see Wilkinson and TFSI, 1999, p. 596).

(c) Test results tell us nothing about the probabilities of null or alternative hypotheses. The probabilities known as the significance level or the power of the test are probabilities of certain sets of outcomes given the condition that the null or, respectively, the alternative hypothesis is true.

(d) A significant result is not necessarily important. Very low p-values do not in themselves imply large effect sizes. A small estimated effect size still can yield a low p-value, e.g., if the corresponding standard error is very small, which can be caused by a large sample or by good experimental control. Investigation of effect sizes is important, and is a different thing than hypothesis testing. The importance of reporting effect sizes is stressed in Schmidt (1996) and Wilkinson and TFSI (1999).

(e) The alternative hypothesis is not the same as a scientific theory. If a researcher is investigating some theory empirically, this can be based on tests of which the alternative hypotheses are deduced from the theory, but always under the assumption that the study was well designed and usually under additional technical assumptions such as, e.g., normality of the distributions. Since the alternative hypothesis is only a consequence and not a sufficient condition for the theory, it is possible that the alternative hypothesis is true but the theory is not. To find support for theories, one has to find corroborations of many consequences of the theory under various circumstances, not just of one of its implications. On the other hand, it is possible that the theory is true but the alternative hypothesis deduced from it is not, e.g., because the experiment or observation was ill-designed or because the auxiliary technical assumptions are not satisfied.

(f) The null hypothesis is just what it says: a hypothesis. The test is not invalidated by the mere fact that there are other reasons for thinking, or knowing, that the null hypothesis is wrong. Even if there may be other reasons for which the null hypothesis is wrong, still it can be sensible to check whether the data at hand are, or are not, compatible with it.

4.2 Limitations of the Neyman–Pearson Approach to Hypothesis Testing

The Neyman–Pearson (1928) formulation changed the concept of hypothesis testing because it provided a rational basis for the choice of a test statistic. This was obtained at the cost of a quite narrow frame for the notion of a hypothesis test: a two-decision problem in which one error ('type I') is considered much more serious than the other error ('type II'). This model for decision-making is often inappropriate and, when it is useful, it is more commonly a useful approximation than a faithful representation of the actual decision situation. Some of its limitations are discussed briefly in this section.

(a) Even when a dichotomous decision must be taken as a result of a scientific investigation, it is possible that both types of error should be considered equally serious. This is the case, e.g., when two therapies are compared and the most promising one must be chosen for further study or for practical application. If the costs of both kinds of error (choose therapy B, while A is better; or choose therapy A, while B is better) are about the same, then there is no asymmetry between the two competing hypotheses, and a significance test is not in order.

(b) Often there are more than two outcomes of the decision situation. For example, in the two-sample situation discussed above, where the population means are μ_A and μ_B, and where the researcher entertains a theory implying that $\mu_A > \mu_B$, it is natural to define three decision outcomes defined as '$\mu_A > \mu_B$ (support for the theory),' 'undecided,' and '$\mu_A < \mu_B$ (evidence against the theory).' Usually, however, such a situation is represented by a Neyman–Pearson testing problem with null hypothesis H_0: '$\mu_A = \mu_B$' and alternative H_1: '$\mu_A > \mu_B$.' If the t-statistic for $\mu_A - \mu_B$ yields a strongly negative outcome ('significance in the wrong direction') then the researcher only concludes that the null hypothesis is not rejected, whereas the data conveys the message that the alternative hypothesis should be rejected. The data then have to be considered as evidence against the theory embodied in H_1. Such a three-decision formulation is closer to the actual purpose of hypothesis tests in many situations where a one-dimensional parameter is being tested, cf. Leventhal (1999).

(c) The Neyman–Pearson approach is formulated as if the data at hand is the only evidence available about the null and alternative hypotheses: it is an *in vacuo* formulation. This is almost never realistic. Often the propositions under study are theoretically plausible or implausible, there is earlier empirical evidence, there may be converging evidence (as in the case of triangulation studies) in the same investigation, and other evidence may become available in the future. In the light of this surrounding evidence, it is unrealistic—perhaps arrogant—to presume that the researcher has to take either of two decisions: 'reject H_0' or 'do not reject H_0.' It is more realistic to see the data at hand as part of the evidence that was and will be accumulated about these hypotheses. This is a major reason why presenting the result of the test as a p-value is more helpful to other researchers than presenting only the 'significant–nonsignificant' dichotomy (also see Wilkinson and TFSI 1999).

There are several ways in which the results obtained from the data at hand, perhaps summarized in the p-value, can be combined with other evidence. In the first place, an informal combination can be and often will be made by a sensible, nonformalized weighing of the various available pieces of evidence. One formal way of combining evidence is provided by the Bayesian paradigm (e.g., Gelman et al. 1995). This paradigm presupposes, however, that the prior and current ideas of the researcher about the plausibility of all possible values of the statistical parameter under study be represented in a probability distribution for this parameter, a very strong requirement indeed. Other formal ways of combining evidence have been developed and are known collectively by the name of

meta-analysis (e.g., Cook et al. 1992, Schmidt 1996). These methods require that the several to-be-combined hypothesis tests address the same substantive hypothesis, or that they can be regarded as tests of the same substantive parameter.

5. Conclusion

The statistical hypothesis test is one of the basic elements of the toolbox of the empirical researcher in the social and behavioral sciences. The difficulty of reasoning on the basis of uncertain data has led, however, to many misunderstandings and unfortunate applications of this tool. The Neyman–Pearson approach is a mathematically elegant formulation of what a hypothesis test could be. Its precise formulation rarely applies literally to the daily work of social and behavioral scientists, very often it is a useful approximation to a part of the research question, sometimes it is inappropriate—but sacrosanct it is not. A hypothesis test evaluates the data using a test statistic set up to contrast the null hypothesis with the alternative hypothesis, and the p-value is the probability to obtain, *if* the null hypothesis is true, outcomes of the test statistic that are at least as high as the outcome actually observed. Low values therefore are evidence against the null hypothesis, contrasted with the alternative hypothesis. It is more conducive to the advance of science to report p-values than merely whether the hypothesis was rejected at the conventional 0.05 level of significance.

In the interpretation of test outcomes one should be aware that these are subject to random variability; and that the probability calculations are based on assumptions which may be in error. Incorrect assumptions can invalidate the conclusions seriously. Nonparametric and robust statistical procedures have been developed which depend less on these kinds of assumption, and diagnostics have been developed for checking some of the assumptions. When reporting empirical results, it is usually helpful to report not only tests but also estimated effect sizes and/or confidence intervals for important parameters. To sum up: 'It seems there is no escaping the use of judgment in the use and interpretation of statistical significance tests' (Nickerson 2000, p. 256).

See also: Hypothesis Testing in Statistics; Hypothesis Tests, Multiplicity of; Model Testing and Selection, Theory of; Significance, Tests of

Bibliography

Cook R D, Weisberg S 1999 *Applied Regression Including Computing and Graphics*. Wiley, New York

Cook T D, Cooper H, Cordray D S, Hartmann H, Hedges L V, Light R J, Louis T A, Mosteller F 1992 *Meta-analysis for Explanation. A Casebook*. Russell Sage Foundation, New York

Fisher R A 1925 *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, UK

Fox J 1991 *Regression Diagnostics*. Sage, Newbury Park, CA

Gelman A, Carlin J B, Stern H S, Rubin D B 1995 *Bayesian Data Analysis*. Chapman & Hall, London

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L 1989 *The Empire of Chance*. Cambridge University Press, New York

Hacking I 1965 *Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK

Harlow L L, Mulaik S A, Steiger J H (eds.) 1997 *What if there were no Significance Tests?* Erlbaum, Mahwah, NJ

Lehmann E L 1993 The Fisher, Neymann–Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association* **88**: 1242–9

Leventhal L 1999 Updating the debate on one- versus two-tailed tests with the directional two-tailed test. *Psychological Reports* **84**: 707–18

Neyman J, Pearson E S 1928 On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20A**: 175–240, 263–94

Nickerson R S 2000 Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* **5**: 241–301

Schmidt F L 1996 Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* **1**: 115–29

Wilcox R R 1998 The goals and strategies of robust methods (with discussion). *British Journal of Mathematical and Statistical Psychology* **51**: 1–64

Wilkinson L, Task force on statistical inference 1999 Statistical methods in psychology journals. *American Psychologist* **54**: 594–604

T. A. B. Snijders

Hypothesis Tests, Multiplicity of

Standard procedures for controlling the probability of erroneous statistical inferences apply only for a single comparison or determination from a given set of data. Multiplicity is encountered whenever a data analyst draws more than one inference from a data set. With multiplicity, standard procedures must be adjusted for the number of comparisons or determinations that actually or potentially are performed. If some adjustment is not made, the probability of erroneous inferences can be expected to increase with the degree of multiplicity. In other words, the data analyst loses control of a critical feature of the analysis.

Questions arising from problems of multiplicity raise a diversity of issues, issues that are important, difficult, and not fully resolved. This account empha-

sizes the issues, tries to avoid being trapped in undue detail, and provides examples of the application of a recommended procedure.

It is helpful to recognize several general guidelines:

(a) Two approaches to the same data, with different purposes or with different questions being posed, need not yield compatible results.

(b) Exact equality of long-run values is very rare and should not play an appreciable role either in one's thinking or one's practice.

(c) When posing questions about data, greater realism will be very valuable.

(d) The situations faced must not be characterized either as completely 'fixed' or completely 'random.'

(e) In selecting a bouquet of techniques for the next decade or the next millennium, the selection must not be cast in stone; further evolution is to be expected (in later months, years, or centuries).

Issues of multiplicity appear in a wide variety of settings, including

(a) outcomes, each measured with variability, in several situations:

(i) comparison of all pairs of situations ('multiple comparisons'); and

(ii) comparison between each situation and a standard ('multiple determination');

(b) selection of one or more elements from a set of alternatives:

(i) selection of how variables are to be expressed for analysis (i.e., what re-expression, if any, should be used?);

(ii) selection of which of the available variables are to participate in an analysis (subset selection); and

(iii) reduction in size of a set of alternatives in an attempt to reduce their number without losing the best.

1. Illustrative Settings, Error Rates, and Procedures

Some important issues arise for all of these alternatives. These issues are considered and illustrated for one example of multiple determination and one example of multiple comparisons.

Consider k independent estimates $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ of, respectively, $\mu_1, \mu_2, \ldots, \mu_k$ and an estimate of variability applicable to all these estimates. This is typical enough of multiplicity problems to render conclusions that are broadly helpful.

1.1 Multiple Determination

Defining a standard, λ, and for each μ_i, drawing conclusions about μ_i–λ based on $\bar{y}_i$–λ is, by definition, making a multiple determination. (In some applications, μ_i is a mean change from one time to another. Often in that case, $\lambda = 0$.) Experience and common sense suggest strongly that μ_i and λ are certain to differ in some decimal place, so that a focus on a null hypothesis that $\mu_i = \lambda$ is quite unrealistic and misleading.

With an interest in μ_i–λ for each value of i, it is reasonable to ask for one of these conclusions: (1) to act as if $\mu_i < \lambda$; (2) to act as if the sign of μ_i–λ is not (yet) determined; or (3) to act as if $\mu_i > \lambda$. Here (1) or (3) is a 'definite' result, while (2) is 'indefinite.'

1.2 Multiple Comparisons

Consider the $k(k-1)/2$ individual pair-wise comparisons y_i–y_j. Using these to draw conclusions about the set of μ_i–μ_j is, by definition, making multiple comparisons, although the definition is more general, and applies to any set of linear contrasts among the μ_i. Because μ_i and μ_j are certain to differ in some decimal place, there is no good reason to focus on the null hypothesis that $\mu_i = \mu_j$. Instead, it should be asked, for each (i, j) pair, whether one should (1) act as if $\mu_i < \mu_j$; (2) act as if the sign of μ_i–μ_j were not (yet) determined; or (3) act as if $\mu_i > \mu_j$.

1.3 In General

Notice the use of '$>$' or '$<$' and the avoidance of '$=$' and '$\neq$.' 'Accepting the null hypothesis' is usually at best misleading. Confounding '$>$' and '$<$' as '$\neq$' is unresponsive and confusing. Thus, focusing on direction is important when dealing with only one comparison (or determination), and becomes vital when dealing with several or many comparisons.

In these formulations, an 'error' is asserting one direction when the (unknown) truth is the other direction. (One should forget about the 'null case' of equality.) Asserting that the direction is not yet established may be a wasted opportunity, but is not an 'error.' The task is to control error while minimizing wasted opportunity, i.e., minimizing indefinite results.

Two classical choices for controlling errors consider expected error rates over replications of the study. (See, for example, Hochberg and Tamane 1987, p. 7.) They are:

$$E\left\{\frac{\text{number of errors}}{\text{number of comparisons}}\right\} \leqslant .025$$

('individual,' 'unadjusted,' or 'per comparison')

and

$$E\left\{\frac{\text{number of errors}}{\text{number of families}}\right\} \leqslant .025$$

('simultaneous,' 'Bonferroni,' or 'per family').

Here, 0.025 could easily be replaced by another choice. (For consistency with commonly used notation, the chosen value will be denoted by $\alpha/2$.) 'Family' refers to an identified set of comparisons whose simultaneous error rate is to be controlled. Normally, 'number of families' would be taken equal to unity. These two choices for controlling errors differ by a factor of k for multiple determinations and of $k(k-1)/2$ for multiple comparisons, so the difference between the two can be quite large.

There has been a feeling for decades (e.g., Tukey 1994/1953) that 'individual' or 'unadjusted' is too soft while 'simultaneous' is too severe. More recently, Benjamini and Hochberg (1995) introduced a compromise that is receiving increasing acceptance.

The two parts of their innovation are a definition of an alternative error rate, the false discovery rate (*FDR*), and a moderately simple procedure that, under the simplest circumstances, controls the *FDR*. The definition is the expectation of the ratio of the number of erroneous assertions of direction to the total number of assertions of direction, i.e.,

false discovery rate =

$$FDR = E\left\{\frac{\text{number of errors}}{\max\{1, \text{number of definite results}\}}\right\}$$

Let p_{crit} be the tail area (usually for each of two tails) of the null sampling distribution of the test statistic for any single comparison being judged by a multiplicity-respecting procedure. Each such procedure will stipulate a probability of error or average fraction of error that is bounded by $\alpha/2$ when judging confidence of direction. The value of p_{crit} will depend on the sort of confidence to be attained. Let K be the number of determinations or comparisons in the family under consideration, where $K = k$ for multiple determinations and $K = k(k-1)/2$ for all pair-wise multiple comparisons. Let $h = 1, \ldots, K$ be the rank of the p-value associated with the test statistic for the comparison concerned when ordered from smallest to largest, so that the observed p-values—$p_{(h)}$ for the hth comparison—are weakly increasing from $h = 1$ to $h = K$. Three distinct procedures, two classical choices and one that controls the *FDR*, are defined as follows:

Unadjusted: be confident of the observed direction for comparison h if

$$p_{(h)} \leqslant p_{\text{crit}} = p_{\text{UNA}} = \alpha/2$$

FDR (Benjamini and Hochberg 1995): be confident of the observed direction of the hth comparison when $p_{(h)}$ is the largest p-value such that

$$p_{(h)} \leqslant p_{crit} = p_{B\text{–}H}(h) = (h/K)\alpha/2.$$

Moreover, be confident of the observed direction for all comparisons g with

$$p_{(g)} \leqslant p_{(h)}, \textit{ i.e.}, g = 1, \ldots, h.$$

Bonferroni (per family): be confident of the observed direction for comparison h if

$$p_{(h)} \leqslant p_{\text{crit}} = p_{\text{BON}} = (1/K)\alpha/2.$$

Because

$$(1/K) \leqslant (h/K) \leqslant 1$$

it must be true that

$$p_{\text{BON}} \leqslant p_{\text{B–H}}(h) \leqslant p_{\text{UNA}}$$

Specifically, $p_{\text{B–H}}(h) = (h)p_{\text{BON}}$ and $p_{\text{UNA}} = (K)p_{\text{BON}}$. The three collections of confident directions corresponding to these three approaches are nested, with the collection for the unadjusted procedure the largest and that for the Bonferroni the smallest.

2. *Examples: The NAEP Trial State Assessment*

Data from the National Assessment of Educational Progress (NAEP) Trial State Assessment were subjected to two separate analyses by Williams et al. (1999). The data for the first analysis consist of average eighth-grade mathematics proficiency scores for the 34 states that participated in both the 1990 and 1992 NAEP assessments; data for the second analysis consist of average eighth-grade mathematics proficiency scores for the 41 states that participated in 1992 (Johnson et al. 1993).

2.1 *Multiple Determination*

One way of assessing year-to-year change in state mathematics achievement involves directly testing the significance of change in each of the 34 states. An ordinary Bonferroni adjustment for control of Type I error per family, as was used by NAEP (Johnson et al. 1993), results in an effective critical p-value of $p_{\text{BON}} = 0.000735$ for $\alpha/2 = 0.025$. Table 1 contains the computed difference between 1990 and 1992 mathematics achievement means for each state and the computed pooled standard error. It also includes the statistical decision about confidence in direction (with confident direction indicated by 'b' in each column) for each state under the three multiplicity treatments. For this example, critical p values are used as a basis for comparison of the adjustment procedures. The cutoff

Table 1
Mean eighth-grade mathematics achievement change, 1990 to 1992 (and standard error), by state, and t, p value, and p_{crit} values for three multiple comparison adjustments, $k = 34$[a]

State	$X_{92} - X_{90}$	(s)	t	p value	p_{UNA}	p_{B-H} (h)	p_{BON}
GA	−0.323	(1.77571)	−0.182	0.42814	0.025	0.025000	0.000735
AR	−0.777	(1.48529)	−0.523	0.30141	0.025	0.024265	0.000735
AL	−1.568	(2.01745)	−0.777	0.22004	0.025	0.023529	0.000735
NJ	1.565	(1.92728)	0.812	0.20999	0.025	0.022794	0.000735
NE	1.334	(1.52772)	0.873	0.19320	0.025	0.022059	0.000735
ND	1.526	(1.68552)	0.905	0.18445	0.025	0.021324	0.000735
DE	1.374	(1.34651)	1.020	0.15581	0.025	0.020588	0.000735
MI	2.215	(1.84727)	1.199	0.11761	0.025	0.019853	0.000735
LA	2.637	(2.07943)	1.268	0.10482	0.025	0.019118	0.000735
IN	2.149	(1.63556)	1.314	0.09694	0.025	0.018382	0.000735
WI	2.801	(1.96269)	1.427	0.07936	0.025	0.017647	0.000735
VA	2.859	(1.92992)	1.481	0.07187	0.025	0.016912	0.000735
WV	2.331	(1.39639)	1.669	0.05013	0.025	0.016176	0.000735
MD	3.399	(1.92320)	1.767	0.04113	0.025	0.015441	0.000735
CA	3.777	(2.11460)	1.786	0.03956	0.025	0.014706	0.000735
OH	3.466	(1.85022)	1.873	0.03295	0.025	0.013971	0.000735
NY	4.893	(2.53195)	1.933	0.02901	0.025	0.013235	0.000735
PA	4.303	(2.20545)	1.951	0.02786	0.025	0.012500	0.000735
FL	3.784	(1.93266)	1.958	0.02745	0.025	0.011765	0.000735
WY	2.226	(1.09641)	2.030	0.02339	0.025[b]	0.011029	0.000735
NM	2.334	(1.14816)	2.033	0.02325	0.025[b]	0.010294	0.000735
CT	3.204	(1.53443)	2.088	0.02052	0.025[b]	0.009559	0.000735
OK	4.181	(1.75467)	2.383	0.01018	0.025[b]	0.008824	0.000735
KY	4.327	(1.61804)	2.674	0.00482	0.025[b]	0.008088[b]	0.000735
AZ	4.994	(1.85110)	2.698	0.00452	0.025[b]	0.007353[b]	0.000735
ID	2.956	(1.06775)	2.769	0.00374	0.025[b]	0.006618[b]	0.000735
TX	5.645	(1.88770)	2.990	0.00202	0.025[b]	0.005882[b]	0.000735
CO	4.326	(1.38868)	3.115	0.00141	0.025[b]	0.005147[b]	0.000735
IA	4.811	(1.48805)	3.233	0.00100	0.025[b]	0.004412[b]	0.000735
NH	4.422	(1.35399)	3.266	0.00090	0.025[b]	0.003676[b]	0.000735
NC	7.265	(1.58701)	4.578	0.00001	0.025[b]	0.002941[b]	0.000735[b]
HI	5.550	(1.17134)	4.738	0.00001	0.025[b]	0.002206[b]	0.000735[b]
MN	6.421	(1.35226)	4.748	0.00001	0.025[b]	0.001471[b]	0.000735[b]
RI	5.097	(0.94844)	5.374	0.00000	0.025[b]	0.000735[b]	0.000735[b]
t_{crit}					2.00	2.47	3.33

a Source: Williams et al. 1999. Copyright 1999 by the American Educational Research Association and the American Statistical Association; reproduced with permission from the publisher. b Confident direction of change.

critical t value for each procedure is also displayed in the bottom line of Table 1.

The unadjusted per comparison approach is confident about the direction of 15 differences, whereas the Benjamini–Hochberg (B–H) *FDR* procedure is confident about 11 of these, and the Bonferroni technique is confident about the direction of only four of the differences.

2.2 Multiple Comparisons

All pair-wise mean differences between the states' 1992 eighth-grade mathematics achievement scores were considered. Because 41 states participated in the 1992 assessment, the family size is $K = 41 \times 40/2 = 820$ comparisons. Using the Bonferroni adjustment, there are 480 confident directions between states, and the use of the B–H technique results in an additional 159. The unadjusted analysis increases the number of confident directions beyond the B–H technique by only six.

2.3 For Both Examples

For each example above, the B–H *FDR* procedure has led to a greater number of confident directions than the Bonferroni procedure. For the examples, we may expect that no more than 0.025 of the confident

directions identified by the B–H procedure are in error. (This is what it means to control the *FDR*.) As is generally the case, the relative advantage of the *FDR* procedure becomes greater with increasing numbers of comparisons (Kesselman et al. 1999).

3. General Considerations

3.1 Qualitative Status

In a situation where only a very few of the results are definite, the B–H false discovery procedure performs much like the Bonferroni simultaneous approach. Because these situations include those where 'definite' is purely accidental, as in extreme instances of 'data mining,' it is important that the approach selected resembles a severe procedure, and the *FDR* approach does so.

In a situation where the total number of definite results is a large fraction of the total number of comparisons an algorithm that controls the *FDR* behaves like the unadjusted procedure. If many comparisons deserve a definite result, it is reasonable that the selected approach should not lead to excessively many errors, and again, the *FDR* approach satisfies the criterion.

3.2 Family Size

One of the many problems associated with multiplicity is a frequent strong dependence of result on the number of comparisons in the family. Applying 'simultaneous' to Vermont vs. Maine as New England states can give results fairly different from applying it to Vermont vs. Maine as two of the 'lower 48,' with a family more than 75 times as large in the latter case. Controlling *FDR* turns out to be much less dependent on family size than the use of simultaneous control.

If the individual elements, e.g., the individual determinations, are stochastically independent, then it is known (Benjamini and Hochberg 1995) that using the *FDR* algorithm does control the rate of false discoveries, and may be unnecessarily conservative. It is reasonable to believe that the result of applying the *FDR* algorithm to elements lightly correlated in the sort of way arising from multiple comparisons will be even more conservative.

For making multiple comparisons in the special case where the true values follow an idealized Gaussian distribution, simulations have shown (Lewis and Tukey 2000) that the algorithm can be made less conservative while still protecting the rate of false discoveries, at least when the number of values being compared is no greater than 25. The modified algorithm follows the B–H approach, but uses $(h/K)\sqrt{(h+1)/2}$ in place of (h/K) in the formula for $p_{\text{B–H}}(h)$, with a caveat that enlargement should stop when the raw modification would otherwise be more liberal than the unadjusted analysis. This modification, here denoted by 'TSR,' results in the following expression for the series of critical values:

$$p_{\text{TSR}}(h) = \min\{1, (h/K)\sqrt{(h+1)/2}\}(\alpha/2)$$

Simulations indicate that a substantial improvement in power may be attained with this modification to the B–H procedure in the situations for which it is appropriate.

For additional recent work on procedures that control the *FDR*, see Shaffer (1999) and Benjamini and Hochberg (2000).

3.3 'Fixed' and 'Variable'

A major source of debate associated with multiplicity and factorial design has to do with 'fixed' vs. 'variable,' i.e., fixed versus random. A significant part of the confusion and conflict can be traced to an unwarranted belief, in tests of hypotheses, that mean squares themselves (or what mean squares estimate) are important. With interest just in directionality, each of the individual questions and answers in multiple comparisons involve $\bar{y}_i - \bar{y}_j$, where i and j are specified, and there is *no sense in asking* whether i and j are fixed or variable. (It may make sense, as shown below, to ask if something else is fixed or variable.)

3.4 A Prototype

The simplest multiple-comparison situation that illustrates the issues is a two-way table of responses with some replication in the cells and questions directed to differences between rows. Starting with k rows and m columns, each comparison involves only two rows. Replacing two rows of measurements by one row of estimated differences yields a 1-by-m table with some replication. (Readers may choose their favorite identities for rows and for columns; in particular, columns = blocks and rows = varieties or treatments gives appropriate realizations.)

If the columns and calculations are fixed, then the uncertainty about the answer is a consequence of whatever makes replicates within a cell different. An estimate of the corresponding variance is, say, an f-error term.

If the columns or calculations would be different were the data to be collected again, in particular, if different applications would require different weights, then the uncertainty about the overall results would reflect two sources of variability: whatever makes replicates differ, and the long-run differences (for a selected comparison) between one column and another.

(The second would not appear if the columns, etc., would be the same.) The resulting error term will be larger than the f-error term; just how much larger will depend on the extent to which differences in columns and calculations are reflected in the overall comparison.

In one crisply specified case where the columns are a random sample from a specifiable infinite population of columns, (and the calculations are fixed), there is a well-determined error term, say, the r-error term. Clearly

$$\text{r-error term} > \text{f-error term}$$

In practice, it is not likely to have exactly either columns and calculations fixed, or columns that constitute a simple random sample from a population.

Two alternatives, 'fixed' and 'variable,' are not enough.

A good way to provide a reasonable amount of realism is to define 'c' by

$$\begin{aligned}&\text{appropriate error term}\\&\quad = (c)[\text{r-error term}]\\&\qquad +(1-C)[\text{f-error term}]\end{aligned}$$

so that 'everything fixed' corresponds to $c = 0$ and 'random' is a particular case of $c = 1$. It pays then to learn as much as possible about values of 'c' in the real world.

3.5 Mathematization vs. Realism

Those who believe that data analysis can and should be mathematized are likely to be neither happy about the introduction of 'c' nor highly tolerant of a major role for such a quantity that is not estimable from data. When they recognize, however, that 'c' is about the data analyst or the data analyst's client—about just what question is being asked or about what is the most realistic question that can reasonably be asked—then they may be more willing to think about 'c' and even to work with its default values. Those who recognize the need for more realism today (and even more tomorrow), however, must be prepared for a growing seriousness about 'c' and a growing impact of nonzero values of 'c' on the conclusions to be drawn from the analysis of data. Some may find it useful to give 'c' a Bayesian interpretation, but this is not a necessity.

The only context in which something like being serious about 'c' has passed into a subject-matter field seems to be climatology. The usefulness of a climatological result is clearly measured by its performance over the set of years from its publication until the publication of its next revision. It will be based on what happened, quantitatively, in a set of years before its construction. Weather is a nontrivial time series (surely not a white noise), so neither $c = 0$ nor $c = 1$ can be expected to do well. Empirical study has led to using a base period for climatology of roughly the last 25 years. This is not directly related to a 'c' but is of a similar flavor.

Most data analysis intended to illuminate the real world is rather like climatology. The data, whether experimental or historical, has been collected under some circumstances and in the past, while its analysis has been carried out to estimate what will or what would happen in the future. The strength of the link from past to future is important; judging it effectively is not likely to be easy. Difficult issues need to be faced. If analyses are to be increasingly realistic, decisions about those issues will be required.

3.6 Plausible Defaults for 'c'

To begin to face the tasks, some values of 'c' can be set forth for some crudely described situations. Using such a list as a basis for choosing a 'c' in real data analyses is likely to be better than doing what has become customary, i.e., to assume either that $c = 0$ or $c = 1$. Values on this list will be likely to change over time as more is learned. (Here 'columns' refers to whatever versions are involved.)

$c = 0.0$ fixed columns and calculation

$c = 0.1$ fixed columns and somewhat variable calculation

$c = 0.5$ fixed columns and illustrative weights

$c = 1.0$ random columns and fixed weights

$c = 1.4$ somewhat illustrative columns

$c = 1.6$ illustrative columns

$c = 2.0$ very illustrative columns

For simple random sampling without replacement from a finite population, the value of 'c' is $1-$(sample size)/(population size).

3.7 More Quantitative Conclusions

The focus so far has been on directionality, an appropriate focus when not dealing with crisp measurements, when the hope is to be convinced by the analysis about the definiteness of at least some directions. Especially in some branches of physical science and engineering, the appropriate interest is in the sizes of differences, sizes that may be measured to a small fraction of their nominal value. Instead of directions, the focus will then be upon confidence intervals.

If the concern really is with all the pair-wise differences among k varieties, treatments, or circumstances, or if the pair-wise differences we are concerned about are selected independently of the numbers to be differenced, the Benjamini and Hochberg approach has to regard the results for all differences as definite so that the *FDR* approach could be taken to coincide with the individual unadjusted approach.

If it really is desired to make a simultaneous statement about 'all differences' or to validate a process of focusing on the single difference or the few differences that appear to be most different from a reasonable anticipation, the only viable alternative is

to take a simultaneous approach, presumably using percentage points for a Studentized-range distribution.

These two procedures need not yield compatible results. A difference may have a definite direction by *FDR* but a simultaneous confidence interval that includes long-run differences of both signs. Such incompatibility should be regarded only as evidence that the analyst needs to be clear which question is being asked. Is the analyst looking for a few definite directions (or for all the directions that can reasonably be believed to be definite) or is the interest in providing intervals covering the worst that is likely to have happened for any difference?

3.8 Other Kinds of Multiplicity

The broad issues presented here in discussing multiple determination and multiple comparisons appear in most of the other situations where multiplicity matters. Details may vary, but the broad issues are parallel. For more comprehensive treatments of multiplicity, see, for example, Miller (1981), Hochberg and Tamane (1987), and Hsu (1996).

Bibliography

Benjamini Y, Hochberg Y 1995 Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**: 289–300

Benjamini Y, Hochberg Y 2000 On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**: 60–83

Hochberg Y, Tamane A C 1987 *Multiple Comparison Procedures*. John Wiley & Sons, New York

Hsu J C 1996 *Multiple Comparisons: Theory and Methods*. Chapman & Hall, New York

Johnson E G, Mazzeo J, Kline D L 1993 *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics*. Report 23-ST05. National Center for Education Statistics, Washington, DC

Keselman H J, Cribbie R, Holland B 1999 The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise Type I error control. *Psychological Methods* **4**: 58–69

Lewis C, Tukey J W 2000 Improved multiple comparison procedures for controlling the false discovery rate. Manuscript in preparation

Miller R G 1981 *Simultaneous Statistical Inference*, 2nd edn. Springer-Verlag, New York

Shaffer J P 1999 A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. *Journal of Statistical Planning and Inference* **82**: 197–213

Tukey J W 1994/1953 The problem of multiple comparisons. In: Braun H I (ed.) *The Collected Works of John W. Tukey, Multiple Comparisons, 1948–1983*. Chapman & Hall, New York, Vol. VIII, pp. 1–300

Williams V S L, Jones L V, Tukey J W 1999 Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics* **24**: 42–69

L. V. Jones, C. Lewis, and J. W. Tukey

Hysteria

The concept of hysteria is almost as old as medicine itself. Explanations of the complaints and behavior of the predominantly female patients suffering from this enigmatic condition have changed many times in response to changing ideas about the causes of disease, but none of these explanations has ever been entirely convincing. For at least the last 100 years physicians have been puzzled by the varied manifestations of the condition, dismayed by its ambiguities, and often exasperated by the behavior associated with it. As a result the term is now fading into history, though the various syndromes it embraced still exist and still await a satisfactory explanation.

1. The History of Hysteria

The concept of hysteria (from the Greek word *hystera* for womb) originated in the second millenium BC. Egyptian priest physicians attributed the paroxysmal symptoms of their female patients to wandering of the womb, and treated them either by the local application of fragrant potions to entice the uterus back to its proper station, or by the oral administration of nauseous draughts to drive it back. These assumptions were inherited by Hippocrates and Galen, and so was born the tenacious idea of a disease of women manifested by recurrent fits and other dramatic symptoms of sudden onset.

The European literature of the last 400 years is summarized by Merskey (1979). Thomas Sydenham was probably the first to recognize explicitly the role of emotion in the genesis of hysteria, for in 1697 he said in his *Discourse* that 'The *External Causes* of this Disease are either violent motion of the Body, or more frequently vehement Agitations of the Mind by the Assaults of Grief, Anger, Fear and such like Passions.' Sydenham was, of course, allowing for 'internal causes' as well, but the identification of these remained elusive. Sir Benjamin Brodie established in the 1830s that the muscles and peripheral nerves of patients with hysterical paralyses or anaesthesias appeared quite normal when examined postmortem, and by the 1870s Sir James Paget was convinced that the problem was one of volition rather that of neurological pathology. As he observed succinctly, 'They say "I cannot"; it looks like "I will not"; but it is "I cannot will."' Even

so, the great French neurologists Charcot and Janet still assumed at the end of that century that the fundamental cause of the suggestibility and hypnotizability of hysterics was a hereditary nervous degeneration. The assumption that hysteria occurred only in women also persisted well into that era. Landouzy stated firmly in 1846 that it was '*exclusive au sexe feminin*' and Charcot regarded ovarian tenderness as one of the stigmata of the condition and attached great significance to the precipitation of hysterical symptoms by pressure on the ovary. Only in World War I, when large numbers of soldiers in all the combatant armies developed florid hysterical symptoms in response to the appalling stresses of trench warfare, was the assumption of an exclusively female illness finally abandoned.

The idea that hysteria was a purely psychological disorder was introduced by Breuer and Freud (1896) in their famous *Studies on Hysteria* and contemporary opposition to their views owed almost as much to the abandonment of neurological pathology as it did to their novel and disturbing ideas about infantile sexuality. According to the corpus of psychoanalytic theory that developed from this revolutionary treatise, hysteria was a specific clinical entity arising from sexual conflicts originating in the oedipal phase of psychosexual development. Failure to relinquish the incestuous tie to the opposite sexed parent that develops during infancy leads to sexual conflicts in adult life because the sexual drive, or libido, retains its forbidden incestuous quality. The drive is, therefore, repressed and its energy converted into a hysterical symptom, which not only protects the patient from conscious awareness of the incestuous wishes, and hence from the anxiety that would otherwise result (primary gain), but simultaneously provides a symbolic expression of it (the rhythmic movements of a hysterical convulsion, for example, expressing the patient's coital fears or phantasies). In addition, the symptom itself often provided other benefits (secondary gain) by rendering the subject visibly ill, or making crucial activities impossible.

Despite the lack of supporting empirical evidence, and the wartime evidence that widespread and florid hysterical symptoms were precipitated by overwhelming stress rather than by infantile sexual experiences or phantasies, this psychoanalytic concept of hysteria remained dominant until the 1960s.

2. *Criticisms of the Concept and its Multiple Meanings*

The English physician Skey commented as long ago as 1867 that 'One good reason that may be assigned to the persistent employment of the term Hysteria—a term we all know to be objectionable—is the difficulty of finding a substitute for it.' Those objections and that difficulty have persisted ever since. Indeed, the objections have been expressed forcefully almost continuously from the early years of the twentieth century by German, English, and even French psychiatrists. The most consistent criticisms of hysteria have centered on the multiple meanings of the term and the wide variety of symptoms and behaviors to which it has been applied. At least seven different meanings can be distinguished.

2.1 *Neurological Conversion Symptoms*

Muscular weakness or paralysis, affecting one, two, or all four limbs, or resulting in loss of speech; convulsions simulating epilepsy; or a loss of sensation (commonly blindness or tunnel vision, deafness, or a circumscribed loss of cutaneous sensation) assumed to depend on the psychoanalytic mechanism of conversion. As the extent of the paralysis or sensory loss reflects the patient's ideas rather than anatomical facts, and as it can often be demonstrated neurologically that the relevant motor or sensory pathways are undamaged, the diagnosis of hysteria can usually be made with some confidence.

2.2 *Hysterical Pain and other Psychophysiological Reactions*

Visceral symptoms like vomiting or otherwise unexplained pain assumed to be produced by the same conversion mechanism. Unlike neurological conversion symptoms, there is usually no easy way of demonstrating that such symptoms must be psychogenic and as a result the diagnosis of hysteria remains uncertain.

2.3 *Dissociative Reactions*

Episodes of amnesia, fugue states, and rare cases of multiple personality assumed, after Janet, to be based on a restriction or dissociation of consciousness. Like conversion symptoms, hysterical amnesias and fugues usually develop in highly stressful situations and have the effect of providing at least a temporary respite from that stress.

2.4 *Anxiety Hysteria*

An obsolete psychoanalytic term for phobic anxiety, implying that repressed anxiety is displaced onto a neutral object or situation instead of being converted into a bodily symptom.

2.5 *Epidemic or Mass Hysteria*

Since the 'dancing manias' of fourteenth-century Europe, over 100 outbreaks of mass hysteria have been recorded, usually in schools, convents, or other communities of girls or young women, and usually at times of obvious stress (Sirois 1974). Shared appre-

hensions and suggestion are clearly important as those affected are always in close contact with one another and develop almost identical symptoms. Influential members of the community are generally also the first to be affected.

2.6 Briquet's Syndrome

A well-defined pattern of behavior, first described by the French physician Briquet in 1859 and seen almost exclusively in young women. They repeatedly consult doctors with multiple and constantly changing somatic symptoms for which no adequate cause can be found. They also have unsatisfactory sexual lives and are prone to develop mood disturbances and conversion symptoms. This behavior persists into middle age or beyond, and as a result such women acquire voluminous medical files, have multiple abdominal and pelvic operations, and often become dependent on analgesics or tranquillisers (Guze and Perley 1963).

2.7 The Hysterical Personality

A personality type, first clearly described by Janet at a time when hysteria was still regarded as an almost exclusively feminine condition, which is assumed to be particularly susceptible to develop conversion and other hysterical symptoms. As described, the characteristic features are attention-seeking behavior, shallow, labile emotions, and self-centered, demanding, interpersonal behavior. A variety of tactics are used to attract and retain the attention of others, including striking or sexually provocative clothing, flirtatious behavior, flattery, dramatized accounts of their past lives or current circumstances, and extravagant gestures.

Despite the linkages between these seven meanings the inevitable consequence of this diversity of usage was confusion and serious ambiguity. The situation was summarized trenchantly by the American psychiatrist Chodoff (1974): 'Entities that are clinically quite different are being held together artificially by little more than the authority of an ancient name' he asserted; 'the diagnosis (of hysteria) has become a balloon filled with air rather than substance … it is a fossil encrusted with and obscured by successive layers of meaning.' For good measure he reminded his audience that the term 'hysteric' was used widely by young male psychiatrists to describe almost any attractive female patient.

3. Hysteria Discarded

Although similar criticisms had been voiced many times before, Chodoff's strictures struck home, and when the American Psychiatric Association produced the revolutionary 3rd edition of its *Diagnostic and Statistical Manual* in 1980 the term hysteria had been discarded (see *Nosology in Psychiatry*). Most of its constituent syndromes survived, however, though with revised titles. Conversion symptoms became conversion disorder and anxiety hysteria became phobic disorder, and both were shorn of their original psychodynamic explanations. Dissociative reactions likewise became dissociative disorders. Briquet's syndrome was renamed somatization disorder, and the hysterical personality was renamed histrionic personality disorder. Mass or epidemic hysteria was not mentioned, presumably because it was assumed that each individual would be categorized separately under the appropriate syndrome. When the World Health Organization produced the next (10th) revision of the International Classification in 1992 its treatment of hysteria was almost identical. The only significant difference was that dissociative and conversion disorders were combined in a single category. Quite suddenly, therefore, one of the oldest, most evocative, and most reviled diagnoses in medicine's nomenclature had been expunged, at least for all official purposes.

4. The Limitations of the 'Medical Model'

Discarding the term hysteria has not, of course, solved the underlying problem of why the patients previously so labeled behaved in the way they did. A large part of that problem is that physicians have traditionally been conditioned by a set of mainly unwritten assumptions—what social scientists call the 'medical model'—to assume that every *bona fide* patient must have a disease of some kind. If the patient's complaints can be related to a pathology of some sort no problem arises. But if they cannot, most doctors tend to oscillate uneasily between two alternative attitudes: either there is nothing wrong with them or they are psychiatrically ill; and both alternatives are often unsatisfactory. The former may be true in organic disease model terms, but it provides no explanation for the patient's behavior and no relief of their suffering. The latter is easily taken to imply that the patient must have some kind of mental illness with a discrete cause of its own. Traditionally, hysteria has been one of these mental illnesses, but only one of them. For physicians have also invoked the concepts of hypochondriasis, neurasthenia, psychophysiological reaction, psychosomatic reaction, 'functional overlay,' and malingering in an attempt to explain their patients' behavior, and all these concepts, except possibly the last, are inadequately defined and overlap with one another.

5. The Sick Role and Illness Behavior

It is increasingly well recognized that the question to be asked when patients consult doctors with complaints that appear to be symptoms of bodily illness but in whom no evidence of illness can be found is often not 'What is the matter with them?' but 'Why are

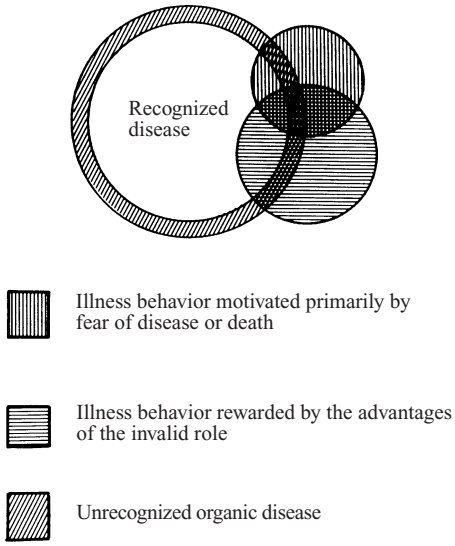

Figure 1
A tentative classification of illness behavior (reproduced by permission of Cambridge University Press from Kendell R E 1983 Hysteria and the hysterical (histrionic) personality. In: Russell G F M, Hersov L A (eds.) *The Neuroses and Personality Disorders*. Cambridge University Press, Cambridge, UK)

they behaving in this way?' If the issue is formulated in this way the concepts of the 'sick role' and 'illness behavior' originally formulated by Parsons (1951) and Mechanic (1962) provide the starting point for a better understanding of the problem (Pilowsky 1969). As such writers have emphasized, the sick role in contemporary society involves many privileges. Invalids are exempted from many normal social obligations; children don't have to go to school and adults don't have to go to work, and often continue to be paid nonetheless. They are also exempted from responsibility for their incapacity, and so cannot simply be told to pull themselves together. Indeed, others are under an obligation to be kind and sympathetic, and sometimes to take over their responsibilities as well. The only reciprocal obligation on the patient is to seek and accept appropriate treatment, so that they utilize the privileges of the sick role for as briefly as possible.

Patterns of behavior that are rewarded tend to increase in frequency at the expense of those that are not, and the rewards of the sick role are so substantial, and are often experienced so frequently in childhood, that it is hardly surprising that not all those who consult doctors with bodily symptoms have objective evidence of disease. Against this conceptual background, Kendell (1983) proposed a tentative classification of invalid behavior in the form of a Venn diagram (Fig. 1). Although reality is undoubtedly more complex, the relationships between the three overlapping populations in this diagram illustrate some basic principles. Most illness behavior is generated by recognized disease (including well-defined psychiatric syndromes like schizophrenia and phobic anxiety). But substantial parts of the whole are generated either by fear of disease or by the 'positive reinforcement' provided by the rewards of the invalid role. There are large and important overlaps between these three populations, and the main recognized disease population is surrounded by a penumbra of unrecognised but genuine disease (mainly people with accepted but so far undiagnosed illnesses, like occult cancers, but perhaps also including subpopulations with as yet unrecognized syndromes, like people with narcolepsy 100 years ago). The management of individual patients depends on which of the 11 populations formed by the intersections of these four circles they come from, rather than on which of a few mutually exclusive diagnostic categories they belong to. Broadly speaking, illness behavior generated by fear of disease corresponds to hypochondriasis and illness behavior motivated by the attractions of the invalid role to what used to be called hysteria.

Essentially, the sick role is attractive, and so liable to be adopted, whenever the balance of its advantages and disadvantages outweighs those of health. To some people the demands of everyday life are so onerous—either because they lack the energy or ability to cope successfully, or because only when ill do they receive sufficient affection, sympathy or attention from other people—that the sick role is preferable most of the time. To many more, the sick role becomes attractive only in situations in which their responsibilities are abnormally heavy (soldiers in battle, for example), or when they are in trouble (impending exams, debt, the threat of blackmail or prosecution, and so on), or when the prospect of financial gain increases the attractions of illness still further (so-called compensation neurosis). The very high incidence of hysterical phenomena in exhausted combat troops suggests that everyone is potentially capable of developing hysterical symptoms, just as they are capable of developing epilepsy, but vary in their liability to do so. This variation presumably depends partly on constitutional variables like suggestibility and capacity to dissociate, partly on past experience, and partly on the subject's current social situation.

6. *Contemporary Understanding of Hysterical Behaviors*

The social and behavioral perspective described above is now accepted widely and implies that hysteria should be regarded not as an illness, or even as a group of related illnesses, but as a set of ways of reacting to stressful situations. It also provides a plausible ex-

planation of what is known of the epidemiology of hysterical symptoms and of the choice and outcome of individual symptoms, and has important implications for diagnosis and treatment.

6.1 Implications for the Epidemiology of Hysteria

Few community surveys of the incidence or prevalence of psychiatric syndromes have generated satisfactory measures of hysteria, because the range of phenomena embraced by the term is so vague and variable, and because most hysterical phenomena are short-lived. Consequently, most of the epidemiological 'facts' referred to below are derived, *faute de mieux*, from the opinions and experience of clinicians. Subject to this proviso, the reason the incidence of conversion hysteria is falling both in the industrial and the developing world (in India there is epidemiological evidence of this) is probably that in both settings levels of education and sophistication are rising and life becoming less stressful. The reason hysterical symptoms are seen mainly in adolescents and young adults is probably because the adoption of a dependent role comes most readily to those who have only recently forgone the privileges of childhood and are least experienced at coping with the demands of the adult world. The predominance of conversion symptoms in women almost certainly has a similar explanation. In most cultures women have always been dependent on their husbands or fathers. As a result, they have been forced to utilize the few opportunities provided by that dependent role to get their own way in a male dominated world. Only in the last two or three generations, and only in industrial countries, has this slowly ceased to be so, which is probably why the number of young women presenting to doctors with conversion symptoms has fallen steadily since Charcot's day and why the female predominance has now fallen below 2:1. The frequent coexistence of hysteria and organic disease is also readily explained, for those who are genuinely ill or handicapped—by epilepsy or multiple sclerosis, for example—are the very people likely to have the most difficulty coping with the demands of everyday life, and with the most vivid and recent awareness of the rewards of the sick role. Finally, the predominance of left-sided limb paralyses and anesthesias, first remarked upon by Briquet, is explained, for in a right-handed person this minimizes the handicap without reducing the social impact.

6.2 Implications for Choice of Symptom and Outcome

According to psychoanalytic theory the choice of symptom is determined by its symbolic meaning for the patient, and doubtless this is sometimes so. Usually, though, it is determined by cultural expectations and the subject's own experience of illness, in themselves or others. The nurse who develops hysterical blindness, for example, had severe conjunctivitis as a child, and the young man who loses the use of his legs had a grandfather who was housebound for years after a stroke.

Just as the development of hysterical symptoms is encouraged strongly by situations in which the balance of the advantages and disadvantages of sickness outweigh those of health, so recovery is encouraged by a reversal of that balance of advantage. This is why the symptom usually fades away if the situation that provoked its appearance resolves itself. If that situation persists, however, or if the inadequacies of the subject are paramount, the advantages and disadvantages of the chosen symptom become important. This is why disabling symptoms, like blindness or inability to stand or walk, have a comparatively good prognosis, whereas those which can be switched on or off as occasion demands, like fits or vomiting, or cause little disability, like tremors, are more intractable.

6.3 Implications for Investigation and Treatment

There are two implications for the investigation of patients with suspected hysterical symptoms, both well known. The first is that medical attention, admission to hospital, and glamorous investigations may easily potentiate the attractions of invalidism. (Indeed, many psychiatrists suspect that the fascinating phenomenon of multiple personality is largely the product of excessive medical interest.) For this reason it is important that the investigations needed to make sure that the patient's symptoms are indeed hysterical are completed as quickly as possible, and not repeated without good reason. The second is that it is immaterial whether or not the patient is aware that he or she is in some sense feigning illness, because the treatment is the same either way. Although the dilemma usually was avoided by assuming unconscious motivation, the traditional medical or psychoanalytic approach should require a distinction between conscious motivation (which is malingering, and so subject to strong disapproval) and unconscious motivation (which is an illness—hysteria—and so worthy of sympathy and treatment) in spite of the fact that it is generally impossible to distinguish the two. Indeed, the patient's degree of awareness often varies from time to time, so that what is initially unconscious becomes conscious, and vice versa.

There are always two components to the management of hysterical symptoms: an attempt to form a psychotherapeutic relationship with the patient, and an assessment of the elements of their social environment which are encouraging or discouraging

invalid behavior. The former is never straightforward, and inexperienced therapists often allow themselves to be manipulated by flattery or coquetry, or end up angrily rejecting the patient because they resent being manipulated, or indeed do both these things in succession. Occasionally it is possible by gentle degrees to lead the patient into a discussion of the emotional predicament from which their symptoms arose, and of the relationship between that predicament and their symptoms. More often there is a sustained denial of emotional difficulties of any kind, and it becomes increasingly clear that the subject is more interested in being a patient than in being cured, and curiously uninterested in the origins of their symptoms.

Arriving at an adequate understanding of the conflicting emotional and social pressures on the patient, both those encouraging and discouraging invalidism, and assessing the possibilities for strengthening the latter, usually involves talking to other key figures in their life as well as the patient. If the patient no longer needs to be ill, they either get better spontaneously, at least in the short term, or do so in response to any treatment that allows them to recover without loss of face. If, on the other hand, there are cogent reasons why invalidism is still more attractive than health, attempts have to be made to modify the attitudes and behavior of those with most influence on the balance of advantage and disadvantage referred to above, and at times this involves devoting more time to a key relative than to the patient. In the past hypnosis was used widely in the treatment of hysteria, and there is no doubt that, thanks to the suggestibility of people with hysterical symptoms, hypnosis is often dramatically successful in the short run. But unless the attractions of invalidism are also tackled, or the underlying predicament has resolved itself, success is usually short-lived.

6.4 The Limitations of Contemporary Understanding

Although this contemporary social and behavioral view of hysterical behaviors explains many otherwise puzzling clinical observations, it provides no explanation of the mechanisms of conversion and dissociation. The explanatory theories proposed originally by Freud and Janet over 100 years ago have been abandoned, but nothing has yet been offered in their place and the neurophysiology of both phenomena remains a mystery. Indeed, shorn of the original psychoanalytic assumption that unconscious anxiety was being converted into a neurological deficit, the term conversion disorder has lost all meaning. And hysteria itself has been discarded. But as Lewis observed a generation ago, 'a tough old word like hysteria dies very hard.' It may yet come back.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Culture as a Determinant of Mental Health; Freud, Sigmund (1856–1939); Mental and Behavioral Disorders, Diagnosis and Classification of; Psychoanalysis, History of; Somatoform Disorders

Bibliography

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. (*DSM-III*). American Psychiatric Association, Washington, DC

Breuer J, Freud S 1896 Studies on hysteria. In: *Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. 2. Hogarth Press, London

Chodoff P 1974 The diagnosis of hysteria: An overview. *American Journal of Psychiatry* **131**: 1073–8

Guze S B, Perley M J 1963 Observations on the natural history of hysteria. *American Journal of Psychiatry* **119**: 960–5

Kendell R E 1983 Hysteria and the hysterical (histrionic) personality. In: Russell G F M, Hersov L A (eds.) *The Neuroses and Personality Disorders*. Cambridge University Press, Cambridge, UK

Lewis A 1975 Survival of hysteria. *Psychological Medicine* **5**: 9–13

Mechanic D 1962 The concept of illness behavior. *Journal of Chronic Diseases* **15**: 189–94

Merskey H 1979 *The Analysis of Hysteria*. Bailliere Tindall, London

Parsons T 1951 *The Social System*. Free Press, New York

Pilowsky I 1969 Abnormal illness behavior. *British Journal of Medical Psychology* **42**: 347–51

Sirois F 1974 Epidemic hysteria. *Acta Psychiatrica Scandinavica*, **51**: 7–44

R. E. Kendell

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

I

Ideal Type: Conceptions in the Social Sciences

The term 'ideal type' (*Idealtyp*) was chosen by Max Weber (1864–1920) to denote the methodological status of certain theoretical constructions central to many social sciences, e.g., the economists' 'law of demand' or any of the concepts defined in his *Economy and Society* ('social relationship,' 'sultanism,' etc.). Common to all is their reference to the subjective action-orientations of individuals, for in Weber's opinion, socio-cultural phenomena become objects of social-scientific interest precisely under their description as 'subjectively meaningful' (see *Problem Selection in the Social Sciences: Methodology*). His point in labeling such constructions 'ideal types' is to emphasize that they cannot be rightfully interpreted as 'laws.' This tenet has never ceased to be controversial.

1. The Ideal Type as Summary of Empirical Findings (Weber)

Most of Weber's methodological reflections bear on the defense of historical science against the Positivist doctrine that to be worthy of the name, a science must aim to discover laws. His reasoning follows the philosopher Heinrich Rickert's (1863–1936), who submits that a discipline's scientific legitimacy is a function of the cognitive value of the selective representation of concrete reality constructed by it, and that to qualify as worth knowing, a representation has either to encompass every phenomenon or to accommodate every significant detail of a single properly selected phenomenon. On this view, scientific 'knowledge' is obtained through the methodical application of a suitable principle of abstraction (or 'concept formation'). There are two such principles: generalizing abstraction—defining natural science—proceeds by disregarding all properties that make things dissimilar, so as to arrive at a set of generalizations universal in scope ('laws of nature') under which every phenomenon can be subsumed and in this sense explained. Individualizing concept formation—defining history—singles out a culturally significant social arrangement and aims to depict it as a configuration of component features that make it unlike any other. The explanation of this conjunction can only be historical, as the combined effect of a unique sequence of unique constellations of causal antecedents. The prototypical individual concept, then, consists in an entire causal narrative in which the development of some unique aspect of a people's civilization is traced from its first beginnings to its (temporarily) final condition (see *Narrative, Sociology of*).

As regards its implications for Weber's conception of the ideal type, Rickert's account of abstraction/concept formation has two noteworthy features. First, concepts are viewed as assertions that something is/was empirically the case. Individual concepts (historical narratives) state the occurrence of complex cause-effect sequences. General concepts explicate their claims in their definitions (which are thus interpreted as summations of numerous findings, depicting the common features of the phenomena classed together). Classification and generalization amount to the same thing. Second, since generalizing abstraction aims to encompass ever wider domains of phenomena, it must necessarily discard their subjectively meaningful aspects. A science of socio-cultural phenomena that preserves them as such can therefore only be historical (see *Generalization: Conceptions in the Social Sciences*; *Classification: Conceptions in the Social Sciences*; *Explanation: Conceptions in the Social Sciences*).

Rickert's ideas ground Weber's position in the controversy following Carl Menger's (1840–1921) criticism of the German economists' rejection of 'abstract theory' in favor of economic history. Predisposed to defend the historical approach, yet convinced of the value of theory, Weber seeks to accommodate both within Rickert's framework. His problem is that of squaring the historical narrative's unavoidable reliance on general concepts (including economic ones) with the incongruity of these concepts' general form with their reference to individuals' subjectively meaningful orientations (which is a requirement of their serviceability in historical narratives). The methodologically crucial manifestation of this incongruity is the disparity between the conceptual definitions' claim to the uniform, explicit, and unalloyed presence of these meanings in the minds of the relevant individuals, and the observed variations in clarity and purity. Given his understanding of general concepts as summarizations of singular instances, Weber tries to make methodological sense out of the conceptual representation of empirical spectra of meaning as strictly general uniformities. His solution is straightforward: their general form notwithstanding, these are not truly 'generalizing' concepts, but 'idealizing' ones;

they depict not those aspects with regard to which phenomena are exactly like others (of a class), but those that by degrees render them 'culturally significant' or 'typical.' Hence, their methodological status is that of 'ideal types' (see *Causal Counterfactuals in Social Science Research*).

Weber considered the formulation of ideal types to be the task of a specialized generalizing discipline—sociology—that would supply the individualizing science of history with its apparatus of general concepts. Methodologically speaking, ideal-type formation is therefore constitutive of a Weberian sociology; it also denies it autonomous scientific value (see *Theory: Conceptions in the Social Sciences*). Whatever cognitive value ideal types may have relates to their utility for historical science. This utility is first of all heuristic: As descriptions of characteristic patterns ideal types can furnish initial guiding assumptions for the scrutiny of sources, especially regarding complex and non-obvious causal relationships and the unobtrusive presence of specific meaning elements in the opaque tangle of concrete action-orientations. Second, they can serve as a basis for educated guesses to bridge gaps in the available evidence. Finally, they are irreplaceable as descriptive devices. To historians who desire precision they provide a clear terminology and an efficient way of pinpointing a phenomenon's peculiarity by specifying its deviation from an ideal-typical definition functioning as a benchmark.

2. *The Ideal Type as Mode of Experience (Schutz)*

While Weber's position harnessed theory and history together, its ultimate vindication of the historical approach could hardly satisfy Menger's intellectual heirs. Ludwig von Mises (1881–1973) objected that the most general propositions of social science could not possibly be derived inductively from atheoretical historical data since the sources are incomprehensible unless they are apprehended in terms of 'the principle of action.' The grasp of human life-manifestations as instrumental and preference ordering underwrites all experience of them as 'actions' and therefore has *a priori* status, like the principle of causality. In von Mises' view, the theorems of economics are nothing but logical implications of the axiomatic category of action and therefore are universally valid independent of all empirical evidence, their grasp made possible through (categorical) 'conception' (*Begreifen*), which is the special mode of cognition that reveals the universal aspects of meaningful behavior. For, so von Mises, human conduct is reason in action, and the universality of reason is the basis of the universality of conduct's structures.

Von Mises' critique of Weber on the grounds of an epistemology of comprehension is shared by Alfred Schutz (1899–1959), who also proclaims the universal validity of 'pure' economic theory. That he simultaneously considers its concepts to be ideal types indicates a conception fundamentally different from Weber's; certainly, his analysis of *Verstehen* as the ideal-typifying grasp of subjective meanings is at least in part conceived to offer a phenomenological–empiricist alternative to von Mises' rationalism (see *Phenomenology in Human Science*; *Interpretive Methods: Micromethods*; *Interpretive Methods: Macromethods*). Accordingly, Schutz formulates a theory of the experiencing of meanings, and of the differences between its everyday and scientific versions. His argument proceeds through a discussion of four scenarios:

(a) An agent's (ego's) understanding of his or her own action's subjective meaning is described as his or her direct awareness of its specific identity as the entailment of his or her action's wherefore ('in-order-to motive' or 'project').

(b) Ego's understanding of someone else's (*alter's*) subjective action meaning is analyzed as his or her experience of what alter is/was experiencing, through the medium of *alter's* inner experiences' outward manifestations (such as acts or pronouncements). Manifestations can be perceived 'objectively' as self-contained entities with identities of their own (as specific things, performances, states); they can also be taken as purveyors of information about their producers' 'inner durations.' In 'objectifying' perception, a bundle of experiences is disembedded from its original setting—the cognizing subject's ever-changing stream of consciousness—and, as a unit, held still over and against it so as to appear as a de-subjectified, bounded, and stable entity classifiable in a schema of 'objective' meaning. Any agent's perceptual bestowal of such objectivity, unity, and permanence (with its attendant assimilation into an objective-meaning schema) on elements lifted from the primordial flow of his or her lived experiences, for Schutz constitutes (ideal-)typification. Even after objectification, however, the reference of action-products to the processes in human consciousness underlying their generation is not obliterated, and in subjectifying or symptomatic perception, it can be made topical. Ego's grasp of *alter's* inner life (understanding) is always based on the subjectifying exegesis of his or her own experiences of such manifestations, aiming to recover their constitutive in-order-to motives. Understanding thus presupposes a particular 'intentional' attitude, i.e., the 'directing' of the 'attention' toward the phenomena not as objective entities but as embodiments of their creators' projects.

In everyday face-to-face relationships a most important role is played by the 'signitive' apprehension of *alter's* body as an expressive field. This apprehension—neither inferential nor hypothetical, but a *sui generis* grasp of the other person's lived experiences at the very moment of their occurrence—is of such immediacy that *alter's* consciousness appears to ego

without any opaqueness and as operating in a manner exactly like his or her own. This prereflective conviction supports the procedure by which ego grasps *alter's* inner life through an explication of his or her own experience of its manifestations. Fantasizing them as generated by himself or herself, he or she imagines what for him or her would function as their in-order-to motives, and treats them as *alter's* own.

(c) When *alter* is not physically present, ego must do with a set of manifestations much reduced in number and symptomatic reach. *Alter* loses his or her quality as a palpable, many dimensional, and multilayered dynamic presence and becomes a more or less anonymous and schematic 'contemporary.' As a function of this anonymity, ego's subjectifying intentional attitude—his or her 'Other-orientation'—undergoes a profound change in structure. In the face-to-face situation, the 'Thou-orientation' founding signitive apprehension captures the other person as a concrete, spontaneous, live self. Beyond this sphere, direct apprehension becomes impossible, and ego-*alter* interaction is no longer structured through dynamic mutual attunement; instead, the participants now treat each other as objects of predicative orientation. The corresponding 'They-orientation' directs ego's attention to those of his or her experiences that can function as substitutes for the unavailable direct experience of *alter's* duration. These substitutes are found in ego's store of objectified motivational constellations accumulated in past encounters with numerous individuals' consciousnesses. They can serve this function because, lifted from their original context, they have become de-subjectified and anonymous. The result of their deployment is the experience of *alter* as a type with a typical project, indistinguishable from other individuations of the same type.

(d) Sociologists adopt the same orientation as contemporaries, but use scientifically tenable (clear and precise) ideal-typifications to grasp their subjects' typifying action orientations. Since these structure the interactions sociologists are interested in, sociological ideal types must be compatible with the agents' types. In addition to exhibiting this 'adequacy at the level of meaning,' they must also be 'causally adequate,' singling out the meanings likely to be the operative ones. Both criteria are satisfied through the formulation of types of rational action orientation, as these are clearly understandable and pervasively used by contemporaries.

The ideal-typical character of social-scientific formulations does not preclude their universal validity, as Weber would have it, confounding the types' 'ideality' with their 'anonymity.' As long as the elements synthesized in them have been so abstracted (idealized) as to exhibit affinity to any particular (kinds of) inner durations no longer, they depict totally anonymous, fully generic processes of consciousness. Such are the propositions of theoretical economics, capturing the actions of a faceless 'pure anyone,' of whom nothing is presumed except that he or she is exchanging goods to satisfy preferences.

In making the universal validity of economic theory a function of the empirical vacuousness of the project attributed to the anonymous 'one,' Schütz adopts von Mises' position. Both authors also agree that substantive assumptions about goals of action are properly regarded as a matter of economic history, where Weberian ideal types reign. Schütz's considerations here are articulated by Felix Kaufmann (1895–1949), who charges theory with providing the 'basic concepts' ('forms') that serve as a field's *a priori* foundation by defining its phenomena, and distinguishes this task from empirical 'application,' where questions of empirical evidence become relevant. One must not confuse, *à la* Weber, the universality of a field's categorical forms with the empirical confirmation of generalizations in its domain of application.

3. *The Ideal Type as Definition of a Model*

The more recent proponents of nomological social science do not propose a theory of abstraction but highlight the role of general empirical hypotheses in nomological-deductive explanations; accordingly, they focus on the explanatory potential of ideal-typical generalizations (see *Explanation: Conceptions in the Social Sciences*). Hempel, for instance, regards ideal types as theoretical systems akin to idealizations in natural science (e.g., the theory of ideal gases); these postulate that under 'ideal' conditions the systems' behaviors exhibit specific regularities. Prototypically these conditions describe empirically never encountered extremes of actual system properties or entities, which raises the question of the explanatory value of 'laws' that either have no empirical domain or lead to false predictions when applied to existing nonideal constellations. This problem is resolved when the law can demonstrably be derived from an independently established, more general theory, but Hempel notes the nonavailability of such a theory even in economics, the most advanced social science, and ends up regarding its idealizations as 'intuitive' initial simplifications awaiting further refinement. However, economists have not been moving to improve their central axioms, and Rosenberg considers them essentially unrefinable generalizations of folk psychology: their fundamental categories ('belief,' 'preference,' etc.) fail to classify phenomena in terms of natural kinds and therefore cannot yield genuine laws.

Hausman undertakes to defend the discipline's methodological peculiarities by viewing theories as general hypotheses that predicate the assumptions of specific models of particular portions of the world. In this perspective, models are definitions of more or less complex predicates (concepts, structures, systems), such as 'classical particle system' (defined by Newton's laws of motion and gravitation), 'simple consumption

system' (Hausman 1992, p. 35), or 'exchange' (Weber 1968, p. 72), and are constituted by sets of assumptions which *per se* assert nothing about the world but merely describe a structure. They may, however, be declared true of some empirical domain, as attributes of actual systems, which are thus asserted to exhibit the structure defined by the model. A predication of this sort is a theoretical hypothesis.

The point behind model construction is to simplify reality's complications. In natural science it is often possible to create experimental situations approaching the model's simplifications and thus to test its empirical adequacy as a predicate. The unavailability of this opportunity in economics has led, on the one hand, to a preoccupation with model construction *per se*, disregarding empirical issues, and, on the other hand, to a conception of economics as 'inexact,' only approximately true (see *Systems Modeling*).

Economic models usually are definitions of (hypothetical) economies or markets, suggesting to Hausman that the discipline is less concerned with the establishment of genuine nomological propositions than with the implications of certain behavioral generalizations—presumed to be roughly correct—for the actions of individuals, markets, and institutions operating in particular circumstances. Compared with the corresponding empirical domains, they represent radical simplifications and idealizations, achieved through the deduction of a basic system structure from a set of theoretically favored principles of economic behavior, and by treating whatever is not so deducible as the confounding effects of 'external,' (noneconomic) factors (which for this reason are excluded from the model).

The highly 'abstract' character of economic models may well be unavoidable. Yet in view of the 'inexactness' of the underlying behavioral premises, conjoined to the practical impossibility of empirically controlling for the effects of the 'external' variables, the models' lack of 'realism' raises the question of their usefulness as predicates in assertions about the real world. In Hausman's view, to make scientific sense, statements employing such predicates must be understood as (implicitly) qualified by vague *ceteris paribus* clauses, which make their truth dependent on the satisfaction of a number of unspecified conditions. Ideally, a *ceteris paribus* clause serves as a shorthand for a list of specific and separately analyzable causal factors whose explicit inclusion in a model renders true otherwise false statements predicating it of some empirical structure(s). Economists, however, frequently invoke the clause without having any precise idea of what specifically is covered by it. This sort of employment amounts to the claim that there are some unidentified factors whose inclusion in the model predicated of some empirical system(s) will make the otherwise false assertion true. Hausman is very much aware that the line between an assertion's 'vague qualification' and its permanent immunization against falsification is easy to cross. How much one may want to stake on vaguely qualified assertions will vary with the context and must depend on one's assessment of the extent of the basic behavioral postulates' 'inexactness.' Hausman believes that a case can be made in their favor but also acknowledges economists' anti-sociological commitment to economics as a 'separate' science, whose enduring appeal seems to involve more than matters of scientific fruitfulness.

See also: Causal Counterfactuals in Social Science Research; Classification: Conceptions in the Social Sciences; Generalization: Conceptions in the Social Sciences; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Narrative, Sociology of; Phenomenology in Human Science; Problem Selection in the Social Sciences: Methodology; Schtz, Alfred (1899–1959); Systems Modeling; Theory: Conceptions in the Social Sciences; Weber, Max (1864–1920)

Bibliography

Burger T 1976 *Max Weber's Theory of Concept Formation*. Duke University Press, Durham, NC

Hausman D M 1992 *The Inexact and Separate Science of Economics*. Cambridge University Press, Cambridge, UK

Hempel C G 1965 *Aspects of Scientific Explanation*. Free Press, New York

Kaufmann F 1925 Logik und Wirtschaftswissenschaft. *Archiv für Sozialwissenschaft und Sozialpolitik* **54**: 614–56

Mises L von 1933 *Grundprobleme der Nationalökonomie*. G. Fischer, Jena, Germany

Mises L von 1960 *Epistemological Problems of Economics*. D. van Nostrand, Princeton, NJ

Rosenberg A 1992 *Economics—Mathematical Politics or Science of Diminishing Returns?* University of Chicago Press, Chicago

Schütz A 1932 *Der sinnhafte Aufbau der sozialen Welt*. Julius Springer, Wien, Germany

Schütz A 1967 *The Phenomenology of the Social World*. Northwestern University Press, Evanston, IL

Weber M 1922 *Gesammelte Aufsätze zur Wissenschaftslehre*. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany

Weber M 1949 *Max Weber on the Methodology of the Social Sciences*. Shils E A, Finch H A (eds. trans). Free Press, New York

Weber M 1956 *Wirtschaft und Gesellschaft*, 4th edn. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany

Weber M 1968 *Economy and Society*. Bedminster Press, New York

T. Burger

Idealization, Abstraction, and Ideal Types

1. Abstraction

Abstraction is a psychological action in which usually one, but possibly more, aspects of a thing or process,

or a group of things or processes, are mentally separated out from that thing or process, or group of things or processes. All intelligent thought about the sensible empirical world can be understood as being abstractive. The empirical world throws up an ever-changing melange of details which invades the senses. To render this datum intelligible, inconvenient details and transitory changes need to be abstracted away from and the simple, stable, and general located. The successful identification of natural and social kinds would seem to require the ability to perform abstractions. Individual police officers, who instantiate the social, kind *police officer*, vary considerably in personal appearance and in other forms of possible identification. In order to reliably identify an individual as a police officer, the stable and enduring underlying properties that are common to the kind police officer must be identified.

Because the process of abstraction involves the disregarding of particular features of things or processes in an attempt to develop more general concepts, abstractions are sometimes conflated with universals. However, there is no requirement that a successful abstraction need lead to the development of a concept which identifies a universal. A successful abstraction may lead to the identification of relatively invariant generalizations about a kind of thing or process which fall well short of being universally applicable. Whereas the universal is conventionally contrasted with the particular, the opposite of the abstract is usually held to be the concrete.

Not everyone would accept that abstraction is as pervasive in the mental lives of individuals as has been suggested here. Plato can be interpreted as denying that the intelligibility of the sensible requires abstraction. It may appear that the mind is abstracting from the experience of concrete entities to the identification of underlying general kinds. However, Plato would argue that it is actually noticing resemblances between particular entities and a world of perfect forms, which it was familiar with before this life. Berkeley and Hume also downplay the role of abstractions in everyday thought. They object to an account of abstraction provided by Locke, holding that what cannot be separated in reality cannot be separated in the mind either (Baxter 1997). Berkeley and Hume see no need for abstraction from the empirical, in order to render the world intelligible, because they do not allow that the sensible is distinct from the intelligible.

Whether or not Berkeley and Hume are right to reject portrayals of abstraction as being all pervasive in thought, it is clear that the process of abstraction plays a central role in the mental lives of individuals. The ability to apply the sophisticated apparatus of modern mathematics and logic to actual situations depends, *inter alia*, on the capacity to reason abstractly. Also, the capacity for abstract thinking enables previously unfamiliar experiences to be rendered intelligible, by relating these to the familiar. The nineteenth century empiricist John Stuart Mill presents the example of 'South-Sea islanders,' whose only experience of four-legged animals is of hogs. The islanders are able to make sense of European talk of other four-legged animals by abstracting from the concrete example of hogs to the general concept of the quadruped (Mill 1950, p. 299).

2. *Idealization*

Idealization, which is enabled by abstraction, is also a psychological act which, while not as pervasive as abstraction in everyday thought, has been extremely important in modern scientific thinking. While abstraction only involves the mental separation of aspects of things or processes, from those things and processes, when these may not be so separable in reality, idealization involves the mental rearrangement of features of reality so as to assist the development of explanations. A familiar idealization is the frictionless plane, used to exemplify mechanical explanations. To conceptualize the frictionless plane, real planes which have friction-causing surfaces are abstracted away from. However, frictionless planes are not merely abstractions. It is convenient to imagine things which have surfaces that are frictionless, even though it is generally accepted that all surfaces exert some friction on objects which they come into contact with, however slight. For the purposes of thought experiments, disbelief is suspended and the convenient idealized fiction of a rearranged and simplified reality, which suits explanatory needs, is allowed; a reality in which, for example, a surface can come into contact with an object and exert no force of friction on that object whatsoever.

The application of scientific laws and models to real complex situations often requires considerable idealization. In order to practically apply Newtonian mechanics to predict the trajectory of a driven golf ball, the golf ball is idealized as a perfect sphere which is struck with a constant force in the direction in which it is intended to travel. Details of the shape and surface of the actual ball which, in reality will affect its trajectory, are conveniently ignored. Other, potentially relevant causal factors such as wind resistance and humidity are also ignored. The resulting idealized calculation of the golf ball's trajectory will, at best, be approximately true some of the time. The trajectory of a golf ball which is hooked or sliced will not be accurately predicted, nor will the trajectory of an irregularly shaped golf ball be accurately predicted, and nor will the trajectory of a golf ball which is struck into a strong wind be accurately predicted. The idealizations made are useful ones. The utilization of a simple model which exactly describes the motion of an ideal golf ball enables an approximate description of the motion of actual golf balls in a useful range of cases. An explanation of the golf ball's trajectory which did

not employ idealizations and which involved an attempt to include all causally relevant factors would be horrendously complex, if it was possible to calculate at all.

Galileo initiated the scientific revolution, in part, by legitimizing use of the abstract language of mathematics to enable the application of idealized explanatory laws and models in the natural sciences. Galileo's *Dialogue Concerning Two Chief World Systems, Ptolemaic and Copernican* (Galilei 1953) is a polemical work in which a proponent of the new approach to science, which endorses idealization, argues against an advocate of the Aristotelianism of Galileo's day that was opposed to idealization in the sciences. In response to the major Aristotelian objection to idealization, that the physical world is not a world of perfect geometrical shapes and therefore cannot be accurately described in geometrical terms, Salviati, Galileo's proponent of idealization, draws an analogy between the 'geometrical philosopher' and a merchant who allows for the weight of packaging materials when calculating the weight of his wares.

> … when he wants to recognize in the concrete the effects which he has proved in the abstract, [he] must deduct the material hindrances, and if he is able to do so, I assure you that things are in no less agreement than arithmetical computations. The errors, then, lie not in the abstractness or concreteness, not in geometry or physics, but in a calculator who does not know how to make a true accounting. (Galilei 1953, pp. 207–8)

The recognition that idealization is endemic to scientific explanation is problematic for the confirmation of theories. An empirical test is rarely, if ever, a decisive test of a theory. Instead it is a test of a conjunction of the theory plus a series of idealizing assumptions. A falsification of this conjunction implies that either the theory is false or that one or more of the idealizing assumptions are to be rejected. Unless the magnitude of the combined effect of all of the idealizing assumptions which stand between a theory and empirical reality can be identified, or unless a theory which accurately describes reality without the aid of idealizing assumptions can be constructed, an exact experiment which decisively confirms or disconfirms a theory cannot be performed. This problem is better known in the philosophy of science as the problem of auxiliary hypotheses (Lakatos 1970).

3. *Idealization in Economics*

The social science in which idealization is most obviously employed, and most explicitly defended, is neoclassical economics. This is because neoclassical economists make idealizing assumptions which are plainly false and therefore in need of defense from criticism. Most if not all of the generalizations which form the corpus of neoclassical economics are true only in idealized situations where fully informed, fully rational agents, who are concerned only with maximizing expected utility, trade in conditions of unfettered competition. Economic reality is very different from this idealized situation. The prices of actual goods usually do not vary in the ways in which the neoclassical laws of supply and demand suggest that they should, and real humans do not appear to act so as to always maximize expected utility. Not only do many of their actions have no apparent relation to the maximization of expected utility, they often appear to act in ways which actually result in a decrease in their expected utility.

The most influential justifications of the use of idealization in economics have been conducted on Galilean lines and appear as part of a defense of economics as being methodologically similar to the natural sciences. A historically influential defense of idealization in economics is presented in John Stuart Mill's *On the Definition of Political Economy; and on the Method of Investigation Proper to It*. According to Mill:

> That which is true in the abstract, is always true in the concrete with proper *allowances*. When a certain cause really exists, and if left to itself would infallibly produce a certain effect, that same effect, *modified* by all the other concurrent causes, will correctly correspond to the result really produced (Mill 1967, pp. 326–7).

Mill notes the unreality of Euclidean geometry for '… lines, angles, and figures, as human hands can construct' (Mill 1967, p. 327), and draws an analogy to the situation of the 'political economist':

> No mathematician ever thought that his definition of a line corresponded to an actual line. As little did any political economist ever imagine that real men had no object of desire but wealth, or none which would not give way to the slightest motive of a pecuniary kind. But they were justified in assuming this, for the purposes of their argument; because they had to do only with those parts of human conduct which have pecuniary advantage for their direct and principal object; and because, as no two individual cases are exactly alike, no *general* maxims could ever be laid down unless *some* of the circumstances of the particular case were left out of consideration (Mill 1967, p. 327).

Mill's defense of idealization ties its use to a metaphysical picture of the world as being composed of causal tendencies which continue to be efficacious when they are not being interfered with by the activity of other causal tendencies (Hausman 1992, pp. 33–53). This is a metaphysical picture which has been argued for recently, and for similar reasons, by Cartwright (1989), whose Millian heritage is quite explicit.

While Mill argued for the general legitimacy of idealization, Marx went further and attempted to

demonstrate the ways in which idealizational explanations in the social sciences could be rigorously tested, or at least this is how he has been interpreted by Leszek Nowak. According to Nowak, Marx should be recognized as 'the Galileo of the social sciences' because his application of Galilean idealization to the social sciences is the first instance of the use of rigorous scientific reasoning in the social sciences (Nowak 1980, p. 38). Nowak's Marx relates idealizational generalizations to real situations in which these can be tested, by identifying each interfering cause which affects the expression of an underlying factor represented by an idealizational law, and attempting to calculate the magnitude of the effect of each such interfering cause.

Nowak interprets Marx's presentation of the law of value—a law relating the value of a commodity to the amount of labor devoted to its production—as an attempt to show how this idealized law could be used to explain the actual price of that commodity in particular economic circumstances. Nowak holds that Marx's law of value is strictly true only when adjustments are made for eight factors which can affect actual economic situations. For example, the law of value applies when the supply and demand of a commodity are exactly balanced, but ordinarily this is not the case. If the magnitude of all factors which distort the value of a commodity are identified, then actual market prices can be explained and decisive tests of the idealized law of value produced, or so Nowak's Marx would argue. Marx and Mill are only some of the better known defenders of idealizations in economics. More recent defences have been provided by Lionel Robbins and Milton Friedman, among others (Hamminga and De Marchi 1994).

Opposition to the use of idealized explanatory models and laws, in economics and other social sciences, is often tied to broader concerns. In particular, some opponents of the use of Galilean idealization in the social sciences attempt to draw a sharp distinction between the methods appropriate to the social sciences and those appropriate to the natural sciences. One influential way of characterizing this distinction is to hold that naturalistic explanations are inappropriate in the social sciences where interpretive understanding is the appropriate methodological framework. On some such views, explanatory laws and models are out of place in the social sciences, and the fact that these are idealized only serves to demonstrate how unrealistic and inappropriate they are in the social sciences. The possibility of a sharp methodological divide between the natural and the social sciences has been under attack, in recent times, by methodological pluralists who want to recognize the potential legitimacy of a range of methodologies in both the natural and the social sciences (Bohman 1991, Clarke 1998). If methodological pluralists are right, then there is no reason to think that idealized nomic explanations are generally inappropriate in the social sciences, although they may be inappropriate in some social sciences, or in some aspects of social science. Recently there has been a renewed effort by naturalists to defend the legitimacy of nomic explanations in the social sciences, and this has included explicit defenses of the utility of idealized explanatory laws in the social sciences (Kincaid 1996, McIntyre 1996).

A focused attack on the use of idealized explanations in neoclassical economics has been put forward by Hollis and Nell (1975, pp. 43–6). Hollis and Nell draw a distinction between the practical application of idealizing explanatory assumptions, such as the idealizing assumption of frictionless motion in physics and the practical application of idealizing explanatory assumptions in neoclassical economics, such as the idealizing assumption that adequate information about the market is available to all participants. In the case of frictionless motion, a clearly defined variable, the coefficient of friction, which can be accurately adjusted for, is assumed away. Actual participants in actual market places often have inadequate information about the market and as—in contrast to the case of frictionless motion—the effects of inadequate information are not well understood, it is practically impossible to adjust real data to precisely test explanatory laws and models which make the idealizing assumption of adequate information.

Hollis and Nell are right that the effects of inadequate information cannot be easily measured and adjusted for, whereas the effect of friction on the motion of objects traveling along a plane can. However, the simplicity of the frictionless plane is atypical of idealizations in the natural sciences where approximations and adjustments for factors which are not fully understood are frequently made (Cartwright 1983, 1989, Laymon 1995). If it is accepted that idealizations are being inappropriately used whenever the effect of interfering factors cannot be quantified—an assumption which Hollis and Nell appear to have made—then considerations of consistency suggest that the results of a very large proportion of the natural sciences ought to be rejected on the very same methodological grounds.

4. *Ideal Types*

Although it did not originate with him, the notion of an ideal type has come to be associated closely with Max Weber, and it occupies a central place in his thinking about social scientific methodology (Weber 1949, 1968). There is a large and growing literature devoted to the interpretation and exposition of Weber's methodology, which one of the leading contributors to it describes as being in a 'poor state' (Burger 1987, p. xiv). Unfortunately this is a comment which strikes this author as an accurate assessment. No doubt the poor state of commentary on Weber's

writings on ideal types is due, in part, to the difficulty of combining the skills of a historian of ideas, a sociologist, and a philosopher of science, which are all needed to properly assess Weber's methodological thought.

Weberian ideal types are a form of idealization which Weber thought to be important in, and distinctive to, the social sciences. Ideal types in the social sciences play many of the roles that idealizations in the natural sciences play. They are used to identify and separate underlying causal factors that are obscured by the interplay of causes which contribute to social environments. When messy, complex data about social environments is appropriately simplified by abstraction, they enable explanatory generalizations and models to be developed.

Weber is at pains to ensure that ideal types not be understood as averages, or as direct attempts to form empirical social generalizations:

> It is thus the "idea" of the *historically* given modern society, based on an exchange economy, which is developed for us by quite the same logical principles as are used in constructing the idea of the medieval "city economy" as a "genetic" concept. When we do this, we construct the concept "city economy" not as an average of the economic structures actually existing in all the cities observed but as an *ideal type*. An ideal type is formed by the one-sided *accentuation* of one or more points of view and by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent *concrete individual* phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified *analytical* construct. In its conceptual purity, this mental construct cannot be found empirically anywhere in reality (Weber 1949, p. 90).

Like idealizations in the natural sciences, ideal types can perform explanatory work at a number of different levels of generality, and do so in a number of different ways. Weber employs the concept of the ideal type in several different explanatory roles (Burger 1987, pp. 130–5). First there are 'individual ideal types' which can play a role in clarifying conceptions of specific historical phenomena. Even if there is no actual instance of the pure individual ideal type of feudalism, it aids historians in understanding feudal societies, and the era in which feudalism was prevalent, if they can see how and where particular societies approximated to the ideal type feudalism. Second, there are 'generic ideal types' describing the underlying causal factors which influence many or all societies. Here Weber would include the idealizations which constitute the 'fundamental concepts' of economics (Weber 1949, p. 100). A third role for ideal types is to describe developmental sequences relating historical epochs. Marx's postulated historical sequence leading to the development of pure socialist societies is an example of this role. In all of these roles ideal types are not obviously distinct from idealizations in the natural sciences, which can also be specific, generic, or developmental. However, there is a fourth way in which ideal types are used in an explanation which appears to distinguish them from idealizations in the natural sciences, which is when they are used as 'essential' ideal types.

Essential ideal types describe the ideas which are characteristic of the thinking of a particular group of people, and which imbue their thought with meaning. If the pre-Raphaelite art movement of the nineteenth century is to be understood, for example, it will not be enough only to examine actual instances of pre-Raphaelite art. The ideal type of art which the pre-Raphaelites were striving to perfect will have to be understood. It may be that no single pre-Raphaelite artwork completely succeeded in capturing the artistic ideals which the pre-Raphaelites aimed to achieve. Nevertheless, there is no hope of understanding the pre-Raphaelite art movement, and its causal interactions with other art movements and with the wider world, unless the expressive aims of its participants are understood. The same is true of broader and more complicated intellectual movements such as Christianity and Communism.

Because the ideal type is a form of idealization which legitimizes causal explanations in the social sciences and can also be an explanatory form which gives a distinctive place to human meaningfulness, its use by Weber can be understood as a means to find middle ground between methodological dualist interpretivists such as Dilthey, who draw a sharp divide between natural scientific explanation and social scientific understanding, and methodological monist positivists and naturalists, who typically seek to subsume social scientific concepts under natural scientific categories. The actual role that Weber's discussion of ideal types played in the history of the intense and polemical methodological debates in Germany early in the twentieth century—the 'Methodenstreit'—is much more complex than this story allows, however.

Although ideal types continue to be a subject of considerable historical and sociological interest (e.g., Burger 1987, Ringer 1997, Runciman 1972), they have rarely been discussed in recent philosophy of science and social science (exceptions include Papineau 1976 and Weinert 1996). In part this is because philosophical discussion of ideal types has mostly been subsumed into discussion of idealizations and in part it is because idealization has not been a focus of a great deal of discussion in twentieth century philosophy.

5. *Contemporary Problems*

While it will be conceded by most that there are particular situations where idealizations can be useful in the provision of approximately accurate descriptions, there remains much dispute about the range of

the utility of idealizational explanations. If idealized laws and models are to be reliably used in scientific explanation, the ways in which actual situations deviate from ideal ones will have to be accounted for. A temptingly simple way of dealing with this problem, which is explicit in Nowak's Marxist treatment of idealization, is to understand each idealization as a limit of a series of real cases and make adjustments accordingly (Nowak 1980). Sometimes this approach is appropriate. The frictionless plane is a limit case of a series of actual planes which are successively less affected by friction. So it is a simple matter to correct for the effect of friction to precisely explain actual cases. However, scientists often have little idea of how to go about correcting for idealizations in order to accurately describe real situations (Cartwright 1983, 1989, Laymon 1995).

In the experimental sciences, the gap between idealizations and the reality they are intended to help describe can be reduced by designing more complex experiments which involve fewer idealizations. Also, experimental situations which eliminate or minimize the effect of interfering causes can be set up, thereby reducing the need to idealize (Laymon 1995). While these approaches will typically reduce the magnitude of the deviation of idealized explanatory laws and models from reality, there remain a number of problems with their application. First, scientists do not appear to construct complete lists of causal factors which might interfere in experiments. Indeed it is not clear that they could do this. Second, causal factors do not always interact in an additive way, so their combined effect resists simple calculation. A third problem is that there is an explanatory cost involved in the use of highly specific experiments. The explanations arrived at are applicable to a narrower band of cases. The cost of accuracy is the loss of generality.

In the social sciences there are moral and practical barriers to experimentation. However idealizations, especially explanatory appeals to ideal types, remain omnipresent in social science explanations. While the gap between idealized explanations, and the real situations that they can be applied to, can be reduced in the experimental sciences, there is little hope of bridging the gap in the social sciences. Social scientists appear to be stuck with the choice of either admitting that their preferred idealizations may be wrong-headed or insisting, without apparent justification, that these are acceptable *a priori*. But idealizations can go badly wrong (Cartwright 1995) and it seems a mistake to disallow this possibility by appealing to *a priori* justification. Nowak's Marx attempted to abstract from the actual circumstances of nineteenth century capitalist economies to the law of value. Neoclassical economists, who deny that economics should concern itself with the concept of value at all, would say that this is a paradigmatic example of idealization gone wrong. But what guarantee do neoclassical economists have, or could they have, to think that their preferred idealizations stand on a surer footing?

A further subject of dispute centers around a metaphysical consequence of the prevalence of idealization in science which has been raised against scientific realists by Nancy Cartwright (1983, 1989), reviving and developing the antirealist thinking of Pierre Duhem, the French physicist and philosopher who wrote at the end of the nineteenth century. Cartwright (1983) agrees with scientific realists who hold that that the fundamental laws of science are explanatorily powerful generalizations, but denies that this in itself is a good argument for their truth. The fact that fundamental laws are idealized means that they are protected from direct empirical testing by loose approximations, by the use of open-ended *ceteris paribus* clauses, and by the unrealistic assumption that causal factors are straightforwardly additive. Scientific realists generally concede that these problems exist, but argue that nonfundamental laws in science, which are special instances of more fundamental laws, are more amenable to testing and can be better confirmed. However, Cartwright denies that actual nonfundamental laws in the natural sciences are literally to be understood as logical derivations of fundamental laws. In a careful study of the construction of amplifiers, Cartwright (1983, pp. 107–12) demonstrates how engineers approximate away from general theory in *ad hoc* ways so as to find descriptively accurate laws which have only the appearance of being logically derived from general laws.

Scientific realists can accede to all of Cartwright's points and hold out the hope of improving experimental techniques, and thereby improving scientific explanations to the point where there is less room for doubt as to the veracity of fundamental laws. However, because of a lack of experimentation in the social sciences, the social scientist who is a realist about particular grand theories in the social sciences, is faced with the same problem without apparent hope of a solution. No progress is made towards demonstrating that a grand sociological or economic theory is true by showing how it is applied in a particular situation to provide a more specific theory, which is empirically well established, if the application of the grand theory to the particular situation turns out to depend on *ad hoc* modifications, and is not an instance of logical derivation. However few, if any, specific social scientific theories appear to be logically derived from grand theories.

See also: Causes and Laws: Philosophical Aspects; Ideal Type: Conceptions in the Social Sciences; Mathematical Models in Philosophy of Science; Mental Models, Psychology of; Mental Representations, Psychology of; Problem Selection in the Social Sciences: Methodology; Scientific Reasoning and Discovery, Cognitive Psychology of; Social Science, the Idea of; Weber, Max (1864–1920)

Bibliography

Baxter D L 1997 Abstraction, inseparability, and identity. *Philosophy and Phenomenological Research* **57**: 307–30

Bohman J 1991 *New Philosophy of Social Science: Problems of Indeterminacy*. MIT Press, Cambridge, MA

Burger T 1987 *Max Weber's Theory of Concept Formation*, expanded edn. Duke University Press, Durham, NC

Cartwright N 1983 *How the Laws of Physics Lie*. Oxford University Press, Oxford, UK

Cartwright N 1989 *Nature's Capacities and their Measurement*. Oxford University Press, Oxford, UK

Cartwright N 1995 False idealization: a philosophical threat to scientific method. *Philosophical Studies* **77**: 339–52

Clarke S 1998 *Metaphysics and the Disunity of Scientific Knowledge*. Ashgate, Aldershot, UK

Galilei G 1953 *Dialogue Concerning the Two Chief World Systems, Ptolemaic and Copernican*, trans. Drake S. University of California Press, Berkeley, CA

Hamminga B, De Marchi N 1994 Idealization and the defence of economics: notes toward a history. *Poznan Studies in the Philosophy of the Sciences and the Humanities* **38**: 11–40

Hausman D M 1992 *Essays on Philosophy and Economic Methodology*. Cambridge University Press, Cambridge, UK

Hollis M, Nell E J 1975 *Rational Economic Man: A Philosophical Critique of Neo-Classical Economics*. Cambridge University Press, Cambridge, UK

Kincaid H 1996 *Philosophical Foundations of the Social Sciences*. Cambridge University Press, Cambridge, UK

Lakatos I 1970 Falsification and the methodology of scientific research problems. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, UK, pp. 91–196

Laymon R 1995 Experimentation and the legitimacy of idealization. *Philosophical Studies*. **77**: 353–75

McIntyre L C 1996 *Laws and Explanation in the Social Sciences: Defending a Science of Human Behavior*. Westview Press, Boulder, CO

Marx K 1968 *Capital*. Progress Publishers, Moscow, Vols. I–III

Mill J S 1950 *Philosophy of Scientific Method*. Hafner, New York

Mill J S 1967 On the definition of political economy; and on the method of investigation proper to it. In: Mill J S (ed.) *Essays on Economics and Society*. University of Toronto Press, Toronto, ON, pp. 309–39

Nowak L 1980 *The Structure of Idealization*. D. Reidel, Dordrecht, The Netherlands

Papineau D 1976 Ideal types and empirical theories. *Br. J. Philos. Sci.* **27**: 137–46

Ringer F 1997 *Max Weber's Methodology: The Unification of the Cultural and Social Sciences*. Harvard University Press, Cambridge, MA

Runciman W G 1972 *A Critique of Max Weber's Philosophy of Social Science*. Cambridge University Press, Cambridge, UK

Weber M 1949 *The Methodology of the Social Sciences*. Free Press, Glencoe, IL

Weber M 1968 *Economy and Society: An Outline of Interpretive Sociology*. Bedminster Press, New York, Vols. I–III

Weinert F 1996 Weber's ideal types as models in the social sciences. In: O'Hear A (ed.) *Verstehen and Humane Understanding*. Cambridge University Press, Cambridge, UK, pp. 73–93

S. Clarke

Identity and Identification: Philosophical Aspects

It is doubtful that the concept of identity is susceptible to a substantial philosophical treatment at a high level of generality. This is not so much because there are too many disparate theories of identity, but more because the sort of things of which the identity is taken up by philosophers are too disparate for uniform treatment. Broadly speaking, two conspicuously different sets of interests make such a treatment particularly difficult. The concept of 'identity,' when applied to such very basic categories as objects, properties (or universals), events, and persons, forms a cluster of themes in metaphysics, and these receive a kind of analysis far removed from such themes as national, ethnic, racial, or sexual identity that are usually discussed in political philosophy and moral psychology. No obviously common notion of identity that is either tractable or interesting spans both sets of interests.

One point of intersection between the more metaphysical and the more social themes might be this. The concept of a person has long been thought of as an ethical category, but more recently the most interesting accounts of the metaphysical issue of personal *identity* have begun to stress these normative and ethical considerations in a rigorous way, thereby infusing metaphysics with value. Although this development is often resisted, it should not really come as a surprise since the metaphysical question of personal identity has always been one of the criteria of identity of a person over time, and if that question was answered wholly in non-normative terms, then the answer would be disappointing in having no relevance to evaluative questions such as why a person should accept responsibility for actions of a past self that is identical to them, or why should a person particularly care—as they do—for the wellbeing of a future self that is identical to them. To insist that any answer to the metaphysical question should have such a relevance would be to begin to unify the two sets of disparate interests in the notion of identity mentioned earlier. Having said that, this article will not take up the metaphysical issues surrounding identity any further, but will focus entirely on the concept of identity that is of more direct interest to the social sciences.

Here too the issues are diverse. There is the familiar question as to the identity of collective social phenomena: of classes, of nations, and of society itself, and whether or not their identity is in some sense dependent ('supervenient') on the identity of individual persons who belong to them. Much has been written on this subject but it is less current than another subject, which is: What is it for an individual person to be a social type, a Muslim, say, or a white male, a Quebecois, a gay or lesbian … ? The currency of the subject is a result of the importance it has come to have

in politics. The rest of this article will restrict its focus to this question.

Although the intensity of identity politics in many parts of the globe in the last few decades of the twentieth century has given rise to the constant use of the term 'identity' as well as to a glamorous theoretical interest in the concept it expresses, there has been little clarity or rigor in its theoretical deployment. One initial step towards imposing some theoretical order on the concept is to distinguish at the outset between its subjective and objective aspects. The term 'identification' may be usefully introduced to cover the former aspects. That is, when a person is said to have a certain identity owing to some characteristics they have, and with which they identify, then identity is being thought of in its subjective aspects. If their identity is owed to characteristics they possess but with which they do not necessarily identify, then we are thinking of identity in its objective aspects. This distinction is of the utmost importance, because it eventually reveals the conceptual links that the concept of identity has with a range of other notions in moral and political philosophy, such as autonomy and reason (to mention just two of the most fundamental), and with the significant debates about them.

1. *Subjective Aspects of Identity*

What is meant by identifying with some characteristic, thereby making it an identity-imparting characteristic in the subjective sense? This is actually a more complicated question than it might seem.

A first attempt to answer it might be to say that someone identifies with a certain characteristic, if they value it. Thus one may value the fact that one belongs to a certain nationality or ethnic community or even to a certain profession, if one can be said to have the identity (in the subjective sense) of an Indian or Korean–American, say, or a teacher and writer. This seems to be a minimal initial condition for identifying with it; but it is clearly not sufficient since one may have values from which one is oneself alienated. This can be a fairly common phenomenon. To be alienated from one's values is structurally akin to being alienated from one's desires. Just as an alcoholic may be disgusted with his own addictive desire for alcohol, so also someone may disapprove of his own patriotism, or find his pride in his profession intolerably smug.

A further condition has consequently to be added to our minimal condition before a characteristic imparts subjective identity: one has to endorse one's valuing of that characteristic. That would presumably ensure that one is not alienated from its valuing. There is further reason why endorsement of this kind is necessary apart from the attempt to solve the problem of values one is alienated from: the topic of identity *qua* identification is not merely about what one is but also about what one conceives oneself to be. This idea, therefore, brings with it the reflective endorsement of the relevant valued characteristics.

To endorse a value, it is often said, is to have a second-order value. Someone must value the fact that they value the characteristic of belonging to a certain nationality or profession, before they can be said to have the identity of an Indian or a teacher, in this subjective sense. (We should add that something like a second-order valuing may not be necessary in order to ensure that there is no alienation from one's own first-order values. All that may be needed is something negative: that there is no second-order disapproval of one's first-order value. However, as already noted, the second-order level comes in a more positive form, because identification is unavoidably about reflective matters such as what one takes oneself to be.)

With this second-order valuing in place, have we said something sufficient about identification? Not yet, since one's second-order values can be highly neurotic, and when they are they can be values that one is also alienated from. For example, someone may feel their second-order value which disapproves of their first-order valuing of their role in their profession as being too smug, the profession itself being too prim, too censorious, or too much of a super-ego phenomenon.

What further condition must now be added to give a sufficient account of identification? One possibility is to conceive of identification as requiring a receding hierarchy of orders of value. At each order, one is not alienated from a value if one endorses it at a higher order. So, in our example, any neurosis regarding a person's second-order disapproval of his pride in his profession is ruled out if he has a third-order approval of his second-order disapproval.

This solution, however, raises worries about an infinite regress. In order to avoid such a regress, a second possibility is to conceive of identification as not emerging in a receding hierarchy but as requiring a coherence among one's values, no matter what order of value is being considered. So, in our example, the second-order disapproval is neurotic, not because there is a lack of third-order approval of the second-order disapproval, but rather because the latter does not cohere with one's other values at all levels, including the first-order values. The second-order disapproval is something that one is not alienated from, i.e., something that one identifies with, only if it coheres well with the first-order values and other second-order values one has. Here no infinite regress threatens, but some philosophical account of coherence among values must be worked out to match the coherentist accounts of belief and knowledge that we have already available to some degree of sophistication. It is worth noting that if this coherentist way of thinking of identification is right, identification is very closely tied to rationality in values, since the point of a coherentist account of value is presumably that it is an account of when one's values are rational. In a word, one identifies with one's values to the extent that the

value is rational, in the sense of being fortified by coherence with our other values. The conclusion is attractive, as is any conclusion which allows two seemingly separate themes (rationality and identification) to be united.

The trouble, however, is that if identification is given by a rationality-imparting coherence among our values, then we will be identified with all the values we have which are rational in this sense. But that is of not much help with respect to the idea of identity since it follows from this that our identity (in the subjective sense) is never going to be anything very distinctive. It will pick out nothing very special or identity-imparting among all our coherent values such as the ones we have been discussing, one's valuing the fact that that one is an Indian, a teacher, a writer, a Muslim … no one of these will be more importantly relevant to one's identity than other characteristics of ours which we value coherently, such as one's weight, one's love of cricket or of dessert. Precisely what seemed attractive about this view of identification is what makes it of no particular help on the subject of identity, with which we are concerned. (However, on this point it should be noted that the hierarchical conception does no better since it too gives no indication as to why we should think that higher-order valuings of first-order values regarding such things as one's nationality or profession should be more relevant to one's identity than those which endorse first-order valuings of one's weight or of dessert.)

It is not of much help to say that in a relatively coherent set of values, some will still be more coherent than others, and it is these which are relevant to identity. These will still be far too many to be distinctive in the way we want the values to be which give one one's identity. There will be nothing very interesting about the notion of identity, and no particular reason to take it up as a theme in social and political philosophy, if our identity is always given by a wide range of characteristics which we value and whose valuing we endorse.

Perhaps this difficulty teaches a deflationary lesson. Perhaps it is a sign that there is something inflated about the very idea of identity, that our thinking that there is something specially distinctive about some characteristics is misguided, that it is not something we should expect, and that we have come to expect it only because of the recent rise of identity politics which has elevated some characteristics—nationality, ethnicity, linguistic, and religious allegiances—beyond anything warranted by or echoed in the actual moral–psychological economies of ordinary citizens. There is some point to this qualm, a point that needs much nuanced development in longer discussions of the subject. But it can also be a point that is too glibly made. For it does seem, at first sight anyway, to be quite accurately descriptive of at least a small, vocal, and influential body of citizens in many polities who display a strong identification with these very characteristics. And of them at least, some notion of identification must be given that shows why these characteristics are valued more distinctively than the many others that are also coherently and rationally valued by them.

We therefore still lack a sufficient account of identification which will make the notion relevant to our subject of identity. To repeat: what we need is not merely endorsement (of one's own valuing of some characteristic of oneself) by some higher-order values or by coherence with other values, but some further element that makes the value endorsed more central and distinctive in our psychological economies.

An obvious thought here might be to say that the values which are more central are those that are more intense than other values, especially since it does seem as if Muslims, Quebecois, Serbians, etc., who seem most visible in identity politics, value their Muslimness, etc. very intensely, more intensely than other things they value. It is a question, however, of how theoretically useful it is to assert 'intensity' as a primitive and unanalyzable property that values have; but even apart from this problem, the proposition is wrong, for in some cases of weakness of will we act on values that are very intense but which we do not endorse. Nor does it improve things to make the requirement of intensity move up one level by requiring the endorsement of identity-imparting values at the second-order level to be more intense than the endorsement of other values. This is because the problem created for the requirement of intensity by cases of weakness of will carries upward to higher orders as well.

A better way of getting at the required further element is to say that the endorsement of the value must be such that it makes the value concerned more *unrevisable* (rather than more intense) relative to the other values one holds. For it is surely intuitive that Muslims or Quebecois, whose identity is caught up with their valuing of their characteristics, are less likely to give up valuing them than the other values they hold, or perhaps—a more subtle variation of the intuition, one that will be developed briefly below—are less likely to conceive of themselves as giving up the value.

For this thought to be genuinely promising, we need (a) to show what sort of endorsement of a value makes the value relatively unrevisable, and we need (b) to ground the idea of unrevisability in something which would not show the reluctance to revise to be irrational by the lights of the agent herself (in the way cases of weakness of will show that acting on or even holding some of our most intense values is irrational by the agent's own lights).

The first is the most significant task: which among all the values that equally cohere with one another are the more unrevisable? As we said, it is only if we answer this question that we capture what is distinctively identity-imparting about valuing being a

Muslim or Quebecois for the sort of agent we are concerned with. Here is a way of bringing out why this should be so. Any answer to the question would have an analogous effect to what Quine (1953) and Putnam (1975) intended when they argued that some beliefs or propositions are at the very center rather than further out toward the periphery in our physical theory. The idea behind this metaphor was intended to replace misguided traditional ideas of 'analytic' or 'synthetic' *a priori* propositions. Their point was that for the theory to be the theory it is, is given by the beliefs or propositions which are at the center of our physical theory (e.g., our belief in the transitivity of length), since they are more immune to revision than the ones further out on the periphery which are more exposed to what Quine called the 'tribunal of experience.' If those central beliefs do get revised, then it is not clear whether the theory (the very same theory, that is) has been refuted, or whether we have changed the subject because the meanings have changed. So also, analogously, we might say that a person's identity is given by his or her relatively unrevisable values—however we characterize them—and if those are given up, then it is not clear whether it is a change in the ordinary sense where the overall identity remains constant whilst a change in value has taken place, or whether the overall identity itself is changed. This analogy, though inexact, is all the same roughly intuitive and reflects our ordinary talk, when in a fit of nationalist sentiment, we say things such as 'I will lose my sense of self if I betray my country' (or, as in Forster's British schoolboy morality, 'I will lose my sense of self if I betray my friend'); the sort of thing we are not likely to say of other things we value. Compare 'I will lose my sense of self if I give up my love of dessert,' which shows that these latter first-order values are analogous to beliefs or propositions at the periphery for Quine. (These are mere examples intended to convey the structural point of the analogy intuitively. Of course it is possible, though perhaps not routine, that someone may value his sweet tooth in the way a stereotypical nationalist or British public schoolboy values country or friendship, that is, analogously to Quine's center of the web rather than the periphery. The possibility in no way spoils the analogy, it merely shows that identities might be eccentric or bizarre on occasion.)

So much for the significance of unrevisability. But (a) was the task of defining it; it sought to define a way of endorsing one's valuing of some characteristic one possesses that shows that value to be relatively unrevisable compared to other values one holds. Since we have already seen that coherence cannot provide such a special way of endorsement, let us return to the idea of second-order valuing of a first-order value in order to explore the sort of endorsement needed. What do we need to add to a second-order value to make the first-order value endorsed as (relatively) unrevisable in one's psychological economy? An example may help to make the question and an answer to it less abstract.

Let us take some of the more absolutist Muslims in Iran over the last two decades of the twentieth century. They have often urged something that approximates unrevisability of their Islamic values. One way they have done so is to argue that Iran needs to protect itself; not just others in Iran, but even the absolutists themselves should protect themselves against their own 'moral' weakening and corruption in the face of the inevitable spread of the pernicious values of modernity in general and of the West in particular. And they have argued (like Ulysses did, anticipating the Sirens) that this protection should be ensured by entrenching Islamic values so deeply now that, were Muslims even to be so weakened, the social, political and legal institutions would not make it easy for them to shed their Islamic way of life. Such a form of endorsement of one's Islamic values vividly shows it to be more unrevisable (in a very special sense) than other values one has and endorses having. Consider that the endorsement takes a counterfactual form: we value something in such a way that we want ourselves to be such that we live by the value, even if we (counter to present fact) do not value it, anymore. At the time of valuing it, then, such a value stands out as very distinctive. The sort of unrevisability here is quite special because it is not so much that the value (at the first-order level) itself is permanent or immutable, but even though we may revise things later and cease to value it, the fact is that at the time of valuing it in this way, one (at the second-order level) yearns for the value to be unrevisable and relatively permanent, unlike all the other values which we endorse in the more ordinary way. That surely makes it part of one's deepest self-conception since one would disapprove of oneself if one did not have the value at some other point, enough to now make sure that one's self, at that other point, would live according to the value. 'Identity-imparting,' 'self-constituting,' etc., seem apt descriptions for values held and endorsed in this way.

Task (b) remains. Not all values which are unrevisable in this way are rational. What needs to be added is that these unrevisable values must also cohere with our other values. Coherence of an unrevisable value with one's other values allows for the rationality of the unrevisability since reluctance to revise could only be irrational if the value one does not revise is one that does not have support from one's other values.

Before leaving the subjective aspects of identity and turning to the objective aspects, we should make one important cautionary remark to prevent a misunderstanding of the notion of identity as we have just defined. One should not be put off by the specific example given above, and think that the sort of endorsement that generates identity in this way is a sign of fanaticism or illiberalism, just because we have become used to thinking of Islamic absolutism to be illiberal. That would be to allow substantive opinions to blind us to the merits of the theoretical conclusion we have come to; that is to say, blind us to the

structural feature of the endorsement that generates identity. After all, the special way we value our own basic constitutional rights reflects just such an endorsement of our own identity-generating values as liberals, since all it does is reflect the same structure as the Iranian absolutists in the earlier example: we elevate a very few of the many values we more or less coherently hold to fundamental constitutional 'rights,' for example that of free speech, precisely because we want to protect ourselves from giving in to any weakening of those values, for example in the case of so strong a dislike of another's substantive views, that we might find ourselves wanting to censor him. What our elevation of these values into fundamental rights does is express the fact that what we now want is that we cannot later censor someone we strongly disagree with, if indeed later we have weakened enough to wish to censor him. A liberal's deepest self-conception, i.e., a liberal's identity, and a Muslim's identity, therefore, whatever the differences between them on other matters may be, is given by the same counterfactual structures of endorsing and identifying with their cherished values that our theoretical conclusion rests with. Such a conclusion therefore should be a satisfying point of conclusion for the subjective aspect of identity.

2. *Objective Aspects of Identity*

When we turn to the objective aspects of identity, identification on the part of the subject in question with the identity-imparting characteristic(s) is not a necessary condition. Thus, for instance, identities when they are thought to be given by characteristics of descent, such as race, are objective in this sense. Chromosomatic ways of defining gender identity are similarly objective. But biological criteria are not the only criteria that are routinely invoked. Inter-subjective and social criteria are also much favored. Thus, for instance, Marxists often claim that one's identity is given by one's role in a particular economic formation in a given period of history, that is to say, one's class identity as 'class' is defined by Marx (1978).

Many oppose the purely biological ways of thinking about various kinds of identity, such as racial and gender identity, claiming that these identities are 'socially constructed' by the perceptions and attitudes of one's fellows, by the zeitgeist of a particular period, or by the conceptual categories and social institutions at a given time. Foucault (1972) and those influenced by him have made much of this, and Foucault himself gave detailed historical and social accounts of particular concepts and institutions in Europe as being determining of identities. In fact it is interesting that Foucault and his followers claim that it is not only the biological and other scientific criteria that are caught up in social factors of this kind, but the subjective ones we discussed in the last section as well. These too are shaped by conceptual and institutional formations far removed and hidden from the exercise of our reflective self-understanding, thereby showing the ideals of individual autonomy that we assume to reside in the idea of identification to be illusory. This article will not take up these issues raised by Foucault's influence. It will look instead briefly at the motivations for looking at objective factors of identity at all, over and above the subjective ones.

Many subjects may identify with some characteristic that is not to others what is most salient about them and it is thought important by many political philosophers that it is these latter which may often define their identity, no matter what subjects may conceive themselves to be. A good example of this can be seen in Stalin's (1953) well-known definition of a 'nation,' which stresses the importance of historical and economic criteria for national identity, with a view to providing a corrective to what were seen as somewhat premature and ungrounded subjective identifications of 'nationality' found in many secessionist demands in different parts of the world. Here the motivation for objectivist criteria of identity is (at least implicitly) political.

But underlying this is a more interesting theoretical rationale which points to important issues of a more philosophical nature. The claim that agents may have a certain identity even if they do not consider themselves to do so implies that what one takes oneself to be can be mistaken, a kind of self-deception, or at least a self-myopia (which may not always involve the motivated element often associated with self-deception, but involves at least the idea that one may sometimes simply be too deep for oneself—where 'deep' is not intended as a bit of eulogy).

It would be philosophically clarifying to make a distinction between two different sorts of appeal to objectivist identities which are said to be (possibly) hidden from a subject's own self-conception. One claim—the weaker one—is that subjects often betray signs of a certain identity in much of their behavior, even if they do not endorse and identify with what is reflected in their behavior. The other, stronger, claim does not even require that something in the subjects' behavior reflects the identity given by the unendorsed characteristic; rather that the characteristic and the identity is given by the deliverances of some (social, political, economic, or biological) theory regarding these subjects.

(a) The weaker claim not surprisingly, is less controversial since it requires that the characteristics of a subject which are going to define his identity are something that he at least reveals in his behavior. The subject may not endorse them, he may not even acknowledge them, but if the only good explanation of his behavior is that he has those characteristics, and if those characteristics are salient compared to others, then some claim can be made regarding how they

impart his identity. Within this view, the more extreme cases will be where the subject does not even acknowledge the characteristics as being revealed in their behavior. Many of the identities that surface in Freudian (Freud 1963) and psychoanalytic theories make much of this sort of case (Oedipal, narcissistic identities). The less extreme cases will be those where there is acknowledgement of the characteristics, but no endorsement of them on the part of the subject. These are likely to be more common. What may be called 'silent' identities, as in 'silent majorities' often consist of subjects who are not self-identified with a certain pattern of behavior, but will not be in any particular state of denial (as they are in the more extreme cases) about whether their behavior reveals the characteristics they are seen to have. It is very likely, for example, that many moderate Muslims in Iran and other Islamic countries, who do not identify with absolutist or fundamentalist Islam, may all the same admit that much in their behavior offers mute assent to the Islamist elements in their societies.

(b) The stronger claim very often appeals to biological criteria, but is most interesting when it does not. Since the biological criteria are in any case usually caught up with social factors (see the point made about them above during the brief discussion of 'social construction' of identity), they will be ignored here. Perhaps the most well known, well worked out, and widely discussed of the stronger objectivist version of identity, which is not biologically based, is due to Marx and those influenced by him. What makes for having a class identity, say, a proletarian identity, is not any kind of self-identification with the working class, not even any behaviour that suggests certain unacknowledged or unendorsed allegiances to that class, but simply the objective fact of having a certain place and function in the relations of production during the modern capitalist period of economic history. What is remarkable and controversial about this view, more so than anything found in (a) above, is that something regarding the self and its identity is being attributed, without any basis or manifestation being required in the conscious or unconscious behavior of the selves or agents concerned. A working-class person who exhibits no proletarian consciousness nor any of the solidarity and forms of behavior appropriate to the class, and none of whose behavior reflects an unconscious betrayal of such solidarity or consciousness, is nevertheless said to have proletarian class identity, albeit with a 'false consciousness.' It is only because he has this identity that there can be cause to call such a subject's consciousness 'false.' It is false precisely because he fails to conceive himself aright, fails to see his deepest self as it is determined by objective historical and material relations.

It is such a view of self and identity (where self and self-conception can fall so radically apart) which filled Isaiah Berlin (1969) with anxiety in his discussion of 'positive liberty,' since what it encourages is the idea that the achievement of self-realization of individual citizens, that is, the achievement of their own autonomy and liberty (in the positive sense), is now left to states or to the 'vanguards' of political parties, which lay claim to greater understanding of what some subject's self really and objectively is. In such a view, according to those alarmed by it, there is no paradox in the expression 'forcing someone to be free.' (Lenin's (1975) 'vanguardism' was frank about denying any such paradox.)

Underlying political anxieties of this kind is the more philosophical issue of external as opposed to internal reasons, which is much discussed in contemporary moral psychology. An internal reason is a reason for one to do or believe or value something, which appeals to some other evaluative element in one's moral–psychological economy. An external reason makes no such appeal to an internal element; it requires only some objective fact that need not even be recognized by the subject for whom it provides a reason. Thus in the orthodox Marxist tradition, a proletarian, given his historically determined identity, has (an external) reason to be a revolutionary even if there is no element in his moral–psychological economy which values it. Berlin's (1969) anxieties about statist tyranny, carried out in the name of self-realization, autonomy and positive liberty, were thus implicitly and more deeply about the very idea of external reasons, even though he never quite articulated them as having that underlying target; however, it becomes very explicit in a denial of the cogency of the very idea of external reasons in a brilliant essay by Bernard Williams (1981) (a philosopher much-influenced by Berlin), though the point is marred in that essay by a somewhat confused equation of internal reasons with a Humean notion of value and motivation.

This last set of points provides a good closing point for this article on identity, which has distinguished fundamentally between the subjective and objective aspects of the concept. To a considerable extent, which of these two aspects we emphasize in our study of the concept will be a matter of theoretical decision, a decision which, in turn, depends on non-arbitrary philosophical considerations having to do, as we have just seen, with themes at some distance from identity, such as autonomy and moral reasons. In itself, this is to be expected since self, freedom, and reason have been closely connected themes in philosophy ever since Kant, both in the analytical and the European traditions of the discipline. Though much more needs to be said in detail to make the links between these themes explicit, it is safe to say that that the more inclined we are to be uneasy about the idea of 'external reasons,' the more likely we are to stress the subjective rather than the objective aspects of identity.

See also: Collective Identity and Expressive Forms; Ethnic Identity/Ethnicity and Archaeology; Ethnic

Identity, Psychology of; Identity in Anthropology; Mass Media and Cultural Identity; Multiculturalism and Identity Politics: Cultural Concerns; Personal Identity: Philosophical Aspects; Religion: Nationalism and Identity; Social Identity, Psychology of; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns

Bibliography

Berlin I 1969 Two concepts of liberty. In: Berlin I (ed.) *Four Essays on Liberty*. Oxford University Press, London, pp. 118–72

Bilgrami A 1992 What is a Muslim? In: Appiah A, Gates H L (eds.) *Identity*. University of Chicago Press, Chicago, pp. 198–219

Elster J 1979 *Ulysses and the Sirens*. Cambridge University Press, Cambridge, UK

Foucault M 1972 *Archaeology of Knowledge*. Pantheon Books, New York

Frankfurt H 1988 *The Importance of What We Care About*. Cambridge University Press, Cambridge, UK

Freud S 1963 *An Outline of Psychoanalysis*. Norton, New York

Hacking I 1999 *The Social Construction of What?* Harvard University Press, Cambridge, MA

Lenin V I 1975 What is to be done? In: Tucker R (ed.) *The Leninist Anthology*. Doubleday, New York, pp. 156–92

Marx K 1978 The Communist Manifesto. In: Tucker R C (ed.) *The Marx–Engels Reader*, 2nd edn. Norton, New York, pp. 469–500

Putnam H 1975 The meaning of meaning. In: Putnam H (ed.) *Mind, Language and Reality*. Cambridge University Press, Cambridge, UK

Quine W V 1953 Two dogmas of empiricism. In: Quine W V (ed.) *From a Logical Point of View*. Harvard University Press, Cambridge, MA

Rovane C 1998 *The Bounds of Agency*. Princeton University Press, Princeton, NJ

Smith A 1991 *National Identity*. University of Nevada Press, Reno, NV

Stalin J 1953 *Marxism and The National Question*. Progress Publishers, Moscow

Taylor C 1989 *Sources of the Self*. Harvard University Press, Cambridge, MA

White S L 1990 Rationality, responsibility, and pathological indifference. In: Flanagan O, Rorty A (eds.) *Identity, Character and Morality*. MIT Press, Cambridge, MA, pp. 450–87

Williams B 1981 Internal and external reasons. In: Williams B, *Moral Luck*. Cambridge University Press, Cambridge, UK, pp. 101–13

A. Bilgrami

Identity in Anthropology

'Identity' is a vexed topic in anthropology. Simply described, a person's identity is his or her psychologically salient individuality, the way he or she feels different from or similar to other people. But what that means to anthropologists has shifted over time.

At one time anthropologists, and perhaps all social scientists, tended to think of identity as a clear and unambiguous assertion of selfhood, which directed behavior and motivated actions. Self, person, and identity were more or less interchangeable. When Erik Homburger Erikson (for many years the central theorist of identity) wrote about identity, he described it as a kind of consolidation of self, so that when someone acquired her identity, the way she interacted with the world—her ability to trust, to work, and to play—was recognized externally by others in a way that was consonant with her own internal understandings. For Erikson, then, 'identity' was something that had to be achieved, and its achievement was a developmental milestone in the human life cycle. The Eriksonian self stood facing the world as the private awareness of that person which, when mutually recognized and consciously accepted, became her identity. 'Identity formation … arises from the selective repudiation and mutual assimilation of childhood identifications and their absorption into a new configuration, which, in turn, is dependent on the process by which a society (often through sub-societies) identifies the young individual, recognizing him as somebody who had to become the way he is and who, being the way he is, is taken for granted' (1963, p. 159). By definition, identity was unitary and more or less coherent.

In the closing decades of the twentieth century, that simplicity has been lost. There are at least three reasons for this: the postmodern turn in anthropology and elsewhere in the social sciences; the recent focus on power and agency; and the emergence of what is called 'identity politics.' Each suggests a somewhat different perspective. Nonetheless, one can say that 'identity' is no longer conceived as necessarily unitary, no longer perceived as a match between inner experience and outer acknowledgment, and no longer understood as a developmental achievement. But before we discuss these changes, some discussion of terms is in order.

1. Definitions

At least four terms are inextricably intertwined in these discussions: identity, self, agency, and person. Were the intellectual world to be sensibly organized, anthropologists would use these four words to refer to different concepts, but that is not always the case. As a result, the discussions can seem a little murky. It is helpful in this context to turn to discussions in the psychological literature that have been read by anthropologists and to use them to clarify the central concepts.

To begin with the self, more complicated psychoanalytic approaches to the self began to emerge with the psychoanalyst Heinz Kohut's work. In a famous article (1959), he pointed out that the tools of psychoanalysis are introspection and empathy, and he

argued that psychological phenomena like the self thus necessarily include introspection and empathy as essential elements. The 'I-experience,' as he called it, captures the way an individual introspects; thus evidence for the self is fundamentally different from evidence for the observable body. Kohut went on to establish his own vocabulary about the 'I-experience,' the most useful phrase probably being the 'self-object,' the other person that an individual may use to soothe himself and restore inner harmony, as a mother can reassure a child that an imagined dragon is not real. His most fundamental contribution, however (at least for the present purpose), was to distinguish between the introspective nature of the self and the differently observable nature of the person.

This is the distinction that the psychoanalyst Roy Schafer later develops as the distinction between the person, as the agent who acts, and the content of the ideas they have about personhood, the person's understanding of his or her 'I.' He calls these 'self-narratives' (but often refers to them as selves). Schafer also argues that from this perspective, there is clearly more than one self, a position that some anthropologists and psychologists also now adopt (Kondo 1990, Gergen 1991). There may be the self who is confident, assertive, and effective and the one who is weak, embarrassed, and reactive (it is that 'second' self which the analysand will present in the analytic hour). There is the self we become when speaking French, freer, more sophisticated, we feel, than our English-speaking self (cf. Kaplan 1993). There is the work self, the play self, the parenting self, the self in different roles, and the self under different circumstances. Schafer (1992, p. 25) quotes a male analysand speaking to his analyst: 'I told my friend that whenever I catch myself exaggerating, I bombard myself with reproaches that I never tell the truth about myself, so that I end up feeling rotten inside, and even though I tell myself to cut it out, that there is more to me than that, that it is important to me to be truthful, I keep dumping on myself.' There are, Schafer suggests, eight selves of five types in that remark. The types are: actual self, ideal self, self as place, self as agent or subject, self as object. The selves are: analysand self, social self (talking to friend), bombarding self, derogated self, exaggerating self, conciliatory advisor self, advisory self, and the defended self with redeeming features. Schafer is perhaps being excessive (it is his exceeding self) but the point is powerful. To call all of what his analysand describes the product of a single self is to lose a good deal of helpful detail.

From this perspective, a self—or more precisely, a self-narrative or self-representation—is a cognitive schema, a bounded collection of conceptions and images that an individual uses to perceive, to categorize, and to experience his world: in short, to think with. Clearly, people are motivated by more than one simple cognitive packet. They are Americans, professionals, parents, neighbors; they are shy in some settings, assertive in others; skilled and aggressive in some games, cautious and awkward in others. Some theorists conceptualize these 'packets' in terms of roles. Mardi Horowitz et al. (1996), more complexly, write of 'role-relationships.' We establish, they argue, particular ways of relating to people that are driven by our concepts of who they are, who we are in relation to them, and how we are to act in that relationship. They see these role relationships as integrated in a sense of personal identity. Claude Steele (1988) provides evidence for that integration in an experiment in which women in Salt Lake City were called by a man posing as a pollster. While the call was ostensibly about conducting a future poll on women's issues, some women were told that it was commonly known that they were uncooperative with community projects. Some were told that it was commonly known that they were cooperative. Some were told that they were bad drivers. Two days later, a woman called them back and asked them to help with a community project. Salt Lake City is a heavily Mormon community, with a strong ethic of community cooperation. Perhaps unsurprisingly, those who were insulted as uncooperative were twice as likely to agree to help out as those who were praised, as if to clear their good names to themselves. What was more remarkable, however, is that those who were called bad drivers were also twice as helpful as those who had been praised. Steele went on to argue that people often respond to damage to one self-representation by affirming success in another. I hang back with embarrassment as a lousy athlete when friends at a beach party decide that we should all play volleyball, but I can make myself feel better by reminding myself, when I drop the ball, that I am competent in the classroom. Those of us who drop volleyballs can enhance our self-esteem by bringing a terrific homemade pie to the next seaside event.

At the same time, the term 'self' seems to imply something more than this tale of loosely integrated but shifting schemas suggests. Cultural psychologists Markus and Kitayama (1994, p. 569) argue that 'the model which underlies virtually all current social science views the self as an entity that a) comprises a unique, bounded configuration of internal attributes (e.g., preferences, traits, abilities, motives, values, and rights) and b) behaves primarily as a consequence of those internal attributes.' They go on to argue that in the West, for the most part people assume that the major normative task of the self is to maintain its independence, and in other, particularly Asian, settings, the normative task becomes maintaining interdependence with others. (This proposition has been a subject of active debate in anthropology.) They use the concept of the 'self,' then, to indicate not only schemas of the 'I,' but processes of acting in the world.

The cognitive psychologist Ulric Neisser (1988) makes sense of this confusion in an article widely read by psychologically-minded anthropologists. He argues

that there are five kinds of self-knowledge, five ways in which we have knowledge of our 'I.' There is the ecological self, the 'I' perceived with respect to the physical environment; the interpersonal self, experienced in emotional communication, the way 'I' behave in a particular human interchange; the extended self, based on memory and anticipation, the 'I' who has done certain things in the past and will likely engage in certain routines in the future; the private self, experienced as the awareness that other people cannot see your thoughts, that (in some ways) only I can know me; and the conceptual self, the wealth of concepts that I believe comprise my self: my roles (professor, friend, dog-owner), my features (I have a liver, a nose, and a mind), and my specific traits (I am tall, short, smart, stupid, brunette, blond). It is important, Neisser argues, to distinguish between kinds of self-knowledge, because otherwise the self is full of apparent contradictions. It is physical yet mental, public yet private, directly perceived yet incorrectly imagined, universal yet culture-specific. 'Although there is nothing with which we are more familiar, we are often enjoined to know ourselves better than we do' (1988, p. 35). He resolves the contradictions, however, by speaking of different 'selves.' 'They differ in their origins and developmental histories, in what we know about them, in the pathologies to which they are subject, and in the manner in which they contribute to human experience' (1988, p. 35).

The major distinction here is between 'self' and 'person.' 'Self' is used to refer to the introspective experience of the 'I,' and while there are at least two dominant ways of theorizing this experience—the first, as a collection of cognitive schemas and the second, as a more complex integration of those schemas with other psychological processes—the term focuses in on a subject-centered internal experience of the subject. 'Person,' by contrast, is an other-centered term. It evokes an individual's understanding of the shared qualities of individuals like him or herself—the common qualities of those who have selves. Most anthropological discussion of the person takes its starting point from Mauss' (1938) famous essay. Mauss pointed out that the concept of the person changed considerably over time and across space. It was only in the Roman Empire, for example, that the concept of the person came to refer to a locus of rights, which has remained the legal meaning of person in English and American law today. As Geertz (1983, p. 59) points out, the Western concept of the person is 'a rather peculiar idea within the context of world cultures.'

The distinction between 'identity' and 'agency' redescribes that internal/external opposition in more specific terms. 'Agency' is the volitional capacity of all persons: their ability to choose, to intend, to act. When anthropologists talk about 'agency,' they contrast it to 'structure,' by which they tend to mean the external constraints on human freedom institutionalized within a society (the kinship structure, political structure, economic structure, and so forth; see Ortner's (1996) useful discussion).

'Identity' tends to be used for a particular kind of self: the internal, subject-centered experience of unique agency. It would (or should) seem strange to say, 'his whole self is wrapped up in being an anthropologist.' That would seem to imply that even the way he located his body in a room and his memories of childhood were somehow centered on being an anthropologist. But to say, 'his whole identity is wrapped up in being an anthropologist' means that he thinks that the way he currently chooses, intends, and acts has a lot to do with the way he conceives of himself as an anthropologist and not as a chemist. The concept 'self' necessarily entails a sense of not being another person. 'Identity' indexes a much sharper, more conscious sense of difference: having a different kind of agency, making different kinds of choices. The term 'identity,' then, should be used for a subject-centered sense of personal agency. It answers the question, 'who am I?' by asking, 'what can I do in the world?' (See Holland et al. 1998 for a detailed discussion of identity and agency through a dialogic, practice theory orientation.)

2. *Contemporary Influences*

There are three dominant influences on contemporary identity theory within anthropology. First is the literary, postmodern turn in the social sciences. This has brought an emphasis on narrative and performance. Identity becomes something which individuals perform, enact, and present; often, the medium of performance is understood as narrative (see Steedly 1993, Stewart 1996). Lavie's *Poetics of Military Occupation* (1990) is a good, relatively early, example of this kind of work. The goal of Lavie's ethnography is to describe the way Bedouin have come to understand themselves as actors in the world. She presents this through presenting the performance of different Bedouin roles—the sheikh, the fool, the old woman—and through the stories Bedouin tell about themselves. And she uses photographs which, unlike the photographs of most ethnographies, enact the thesis of the book.

Second is the emphasis on power and agency. Like the literary, postmodern turn, this influence permeates much of contemporary anthropology. Its relevance to identity theory lies in the way agency is understood to act within the constraints of external power. Ethnographies with these concerns focus on those who would seem deprived of power or importance, and then demonstrate how even within the margins of the nation state, or in the lowest ranks of the factory, people identify themselves as effective actors in the world. For example, Tsing (*In the Realm of the Diamond Queen*, 1993) organizes her ethnography around a woman who is not only marginal in the eyes of most Wester-

ners (she lives in the Indonesian hills) but in the eyes of her own people: she is intermittently psychotic and is the leader of a small, unpopular cult. Tsing uses her speeches as the vehicle to illustrate how those on such margins nevertheless conceive of themselves as effective actors in relationship to the center, and to some extent actually are.

The third and most powerful influence comes from cultural studies, and more specifically from postcolonial studies and identity politics. Postcolonial studies focuses on relationships between the colonizer and the colonized, and identity politics on relationships between different races and/or genders, but for both the central issue is an asymmetrical power relationship and its consequences.

Memmi's *The Colonizer and the Colonized* (1965) is a foundational text in postcolonial studies. Memmi argues that he who is colonized models himself on the colonizer, aping his customs and yearning for his appearance. He loses his past; he learns the history of Cromwell but nothing of his own progenitors. He grows embarrassed by his mother tongue. 'The first ambition of the colonized is to become equal to that splendid model [of the colonizer] and to resemble him to the point of disappearing in him' (1965, p. 120). The colonized agrees, in Memmi's words, to destroy himself and become what he is not. But then he learns that in the end the colonizer will reject him anyway, and he is filled with shame and self-hatred (1965, pp. 121–2). Ultimately, he has no choice but to recognize the anger in his ambivalence and to rebel. 'Revolt is the only way out of the colonial situation, and the colonizer realizes it sooner or later' (1965, p. 127). And yet even if he is successful he remains still tortured, still living in a psyche defined by the colonizer, defined by what he is not. 'So goes the drama of the man who is a product and victim of colonialism. He almost never succeeds in corresponding with himself.'

Frantz Fanon, in *Black Skin, White Masks* (1967) wrote even more starkly about colonialism and race: 'for the black man there is only one destiny. And it is white' (1967, p. 10). The terrible cost of colonialism in Fanon's eyes is that, in yearning to be like the white colonizer, the black colonized man comes to accept the white man's degrading vision of the black man and so to lose himself. He is alienated from himself when he feels most white, and humiliated when he feels most black. And so, perpetually, the black man is torn, rejecting himself to become white, only then to grasp a more terrible vision of his never-to-be-scrubbed-clean skin. 'A Negro is forever in combat with his own image' (1967, p. 194).

Identity politics takes this central relationship between a dominant group and a subordinate group, an us and a them, and makes theories out of the attempts to change it. For those confronted by the inferiority of their skin, gender, or accent, should they adopt the mainstream style and hope to be respected, or reject the mainstream in order to build a world in which its values will not matter? Those who write within this arena tend to see the route to political resolution of their problems through the conceptual deconstruction of this basic opposition. West, for example, writes of the 'double consciousness' of the rejection of the mainstream (and its association of blackness with inferiority) while assimilating into it. He writes admiringly of 'cultural workers who simultaneously position themselves within (or alongside) the mainstream while clearly aligned with groups who vow to keep alive potent traditions of critique and resistance' (1990, p. 33).

Bhabha, another central voice in identity politics, similarly diagnoses the central problems of the colonized subject as ambivalence and hybridity. The colonial subject identifies with the colonizer and yet cannot be the colonizer. If he identifies with his nativeness, he confronts his own condemnation; if he identifies with the Westernization he has adopted, he confronts his own alienation. 'The taking up of any one position, within a specific discursive form in a particular historical conjecture, is then always problematic' (1990, p. 81). Bhabha uses the word 'hybridity' to describe the solution, or condition, of the multiply defined self which both reiterates binary colonial categories and ultimately subverts them, but through echoes and not direct opposition. Bhabha suggests that the postcolonial subject, the marginal subject, is someone who feels that her capacity for self-declaration has been taken from her repeatedly through a long history in which she has been forced to conform to the dominant culture's prejudices and biases.

As a result, the politics in identity politics become fragmented and complex. Another theorist, Lata Mani, argues that to understand these different perspectives as a unity 'sets up problematic chains of equivalences between, say, people of color in the United States, people from the third world, lesbians, gays … It is not as though difference is not acknowledged, for an inventory of difference is crucial to this narrative; rather it is that difference is insufficiently engaged' (1992, p. 393). Political affiliation becomes possible only though acknowledging difference, because these politics are as much about selfhood as about political economy.

Bhabha sees this complexity of identifications as a condition of modernity. This does tend to be a view about elites (or the middle class) by those who are elites (and have money to spend at Tower Records) but, as elites, many theorists agree. The anthropological critic James Clifford (1988, p. 11) diagnoses contemporary identities thus: 'intervening in an interconnected world, one is always, to varying degrees, "inauthentic": caught between cultures, implicated in others. Because discourse in a global system is elaborated vis-à-vis, a sense of difference or distinctness can never be located solely in the continuity of a culture or tradition …'

3. Contemporary Identity

One could argue that the mastery of multiple narratives is the core project of the postmodern politics of identity. This perspective on identity is quite different from Erikson's naturalistic vision of an identity which can be intuited by an objective observer over a limited period of time. The most insistent message of these identity theorists is the refusal to be characterized by someone else's narrative, and to assert their own authority in laying claim to a narrative of agency that ultimately may not be so different from those they have rejected. When Gayatri Spivak (1987), for example, denounces her characterization as a Marxist, a feminist, a deconstructionist (she is famous for being all three), she does so, one suspects, not because she rejects the personal resonance of those categories for her, but because she resists having them imposed. When members of an academic culture are complicit in the dicta that only women can write on feminism, only men may write about the men's movement, only African-Americans can write on black politics, we acknowledge the importance of allowing those who have been disempowered to experience empowerment through the casting of their narratives of self-characterization.

From this perspective on the politics of identity, successful (healthy, appropriate) identity in a modern postcolonial context is less a self-characterizing narrative with a mirroring world than a sense of command over narrative complexity: narratives which one has to some extent chosen, whose mastery gives one a sense of direction, adequacy and goodness. Rather than locating identity in single narratives, or in narratives matched to external approval, we capture more of the emphasis of these accounts in their experience of comfort and control in slipping from one narrative to another (see Luhrmann 1996). This comfort is what Schafer evokes in his account of the freeing of the 'imprisoned' analysand: not that the analysand is unconstrained by narrative, but that the analysand comes to feel that her narratives of personhood, complex and contradictory as they might be, are those that she has chosen—that rather than being controlled by the brute oppression of a narrative externally imposed, she is in mastery of the narratives which structure her presentations of personhood to the world.

But, as the decades pass, that formulation will change again.

See also: Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Identity and Identification (Philosophical Aspects); Identity Movements; Identity, Social; Erikson Erik H.; Politics of Tribalism/Race/Ethnicity; Postmodernism (Philosophical Aspects); Postcolonialism; Ethnic Identity, Psychology of; Ethnic Identity/Ethnicity and Archaeology; Race Identity

Bibliography

Bhabha H 1990 The other question: Difference, discrimination & the discourse of colonialism. In: Ferguson et al. (eds.) *Out There*, MIT, Cambridge, pp. 71–88

Clifford J 1988 *The Predicament of Culture: Twentieth Century Ethnography, Literature, and Art*. Harvard University Press, Cambridge, MA

Erikson E H 1963 *Childhood and Society*. Norton, New York

Fanon F 1967 *Black Skin, White Masks*. Grove, New York

Ferguson R et al. (eds.) 1990 *Out There: Marginalization and Contemporary Cultures*. MIT Press, Cambridge, MA

Geertz C 1983 *Local Knowledge: Further Essays in Interpretative Anthropology*. Basic Books, New York

Gergen K J 1991 *The Saturated Self: Dilemmas of Identity in Contemporary Life*. Basic Books, New York

Grossberg L et al. (eds.) 1992 *Cultural Studies*. Routledge, New York

Holland D et al. 1998 *Identity and Agency in Cultural Worlds*. Harvard University Press, Cambridge, MA

Horowitz M, Stinson C, Milbrath C 1996 Role-relationship models: A person-schematic method for inferring beliefs about identity and social action. In: Jessor R, Colby A, Shweder R (eds.) *Ethnography and Human Development*. University of Chicago Press, Chicago

Kaplan A 1993 *French Lessons: A Memoir*. University of Chicago Press, Chicago

Kohut H 1959 Introspection, empathy and psychoanalysis. *Journal of the American Psychoanalytic Association* **7**: 459–83

Kondo D K 1990 *Crafting Selves: Power, Gender and Discourses of Identity in a Japanese Workplace*. University of Chicago Press, Chicago

Lavie S 1990 *The Poetics of Military Occupation*. University of California Press, Berkeley, CA

Luhrmann T M 1996 *The Good Parsi: The Fate of a Colonial Elite in a Postcolonial Society*. Harvard University Press, Cambridge, MA

Mani L 1992 Cultural theory, colonial texts: Reading eyewitness accounts of widow burning. In: Grossberg L et al. (eds.) *Cultural Studies*, Routledge, New York, pp. 392–404

Markus H R, Kitayama S 1994 A collective fear of the collective: Implications for selves and theories of selves. *Personality and Social Psychology Bulletin* **20**(5): 568–79

Mauss M 1938 A category of the human mind, the notion of person, the notion of self. In: Carrithers M, Collins S, Lukes S (eds.) 1985 *The Category of the Person: Anthropology, Philosophy, History*. Cambridge University Press, Cambridge, UK, pp. 1–26

Memmi A 1965 *The Colonizer and the Colonized*. (Trans. Greenfield H) Orian Press, New York

Neisser U 1988 Five kinds of self-knowledge. *Philosophical Psychology* **1**(1): 35–59

Ortner S B 1996 *Making Gender: The Politics and Erotics of Culture*. Beacon, Boston

Schafer R 1992 *Retelling a Life: Narration and Dialogue in Psychoanalysis*. Basic Books, New York

Spivak G C 1987 *In Other Worlds: Essays in Cultural Politics*. Methuen, New York

Steedly M 1993 *Hanging Without a Rope: Narrative Experience in Colonial and Postcolonial Karoland*. Princeton University Press, Princeton, NJ

Steele C 1988 The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Social Psychology* **21**: 261–301

Stewart K 1996 *A Space on the Side of the Road: Cultural Politics in an 'Other' America*. Princeton University Press, Princeton, NJ

Tsing A L 1993 *In the Realm of the Diamond Queen: Marginality in an out-of-the-way Place*. Princeton University Press, Princeton, NJ

T. M. Luhrmann

Identity in Childhood and Adolescence

Identity will be discussed as an individual psychological characteristic. In its broadest sense, identity refers to a person's sense of continuity with the past that constitutes the foundation for a meaningful personal and social present, which, in turn, suggests directions for their future. The theorist in the latter part of the twentieth century who is most closely associated with the concept of identity is Erik Erikson (1980). In fact, a recent biography of Erikson (Friedman 1999) is entitled *Identity's Architect*. The most widely used research paradigm developed to study Erikson 's ideas empirically is the identity status model developed by Marcia et al. (1993). Both Erikson's psychosocial developmental theory and Marcia's identity statuses are located within the general context of psychodynamic (ego psychoanalytic) theory. This contextual placement is especially useful because it establishes identity as a developmental structural concept having implications for later adult psychosocial development as well as historical roots in earlier psychosocial periods. The theorists, in addition to Erikson, who speak most directly to the childhood conditions necessary to the establishment of a sense of identity are John Bowlby, Heinz Kohut, and Margaret Mahler. This discussion of identity will be presented within the context of these theories as well as the empirical work of Marcia and colleagues.

1. Identity Described from an Eriksonian Position

Erikson's psychosocial developmental theory is based upon the epigenetic principle that certain ego abilities will develop in a predictable stage sequence given an 'average expectable environment.' He views identity as the stage-specific task of adolescence. Although he has offered different definitions of identity at different times, a comprehensive one is:

> The integration now taking place in the form of the ego identity is more than the sum of the childhood identifications. It is the inner capital accrued from all of those experiences of each successive stage, when meaningful identification led to a successful alignment of the individual's *basic drives* with his *endowment* and his *opportunities*. In psychoanalysis we ascribe such successful alignments to 'ego synthesis' … (Erikson 1959, p. 94, italics in original)

and identity refers to 'a conscious *sense of individual identity*; … an unconscious striving for a continuity of personal character; … as a criterion for the silent doings of *ego synthesis*; and, finally, as a maintenance of an inner *solidarity* with a group's ideals and identity (Erikson 1980, p. 109; italics in original).

The period following puberty and before young adulthood is seen as a time when individual needs (e.g., to be independent, to be sexually expressive) and abilities (e.g., physiological maturity, cognitive sophistication) articulate with societal demands (e.g., to begin to put away childhood, to prepare for an occupation), and rewards (e.g., granting of increasing autonomy, offering rewarding occupational niches and relevant education/training). It is at this point in the life cycle that the individual is expected to begin to leave the childhood position of 'one who is given to and who takes' to 'one who is to give to, and be responsible for, others.' A crucial shift in ideological perspective is required to navigate this transition. This is one of the reasons that Erikson emphasized the importance to identity of an ideology, a kind of theory of oneself and one's social outlook: a *weltanschauung*. Childhood worldviews are not serviceable in adult life, and adolescence constitutes the period of transition between the two. Besides ideology, another area of importance in identity formation in adolescence is occupational choice. Making decisions about this area involves adolescents' inner self-examination of their competencies, wishes, and goals and their outer exploration of available rewarding social contexts. Therefore, one might say that an adolescent is constructing an identity when he or she is exploring alternatives and establishing commitments in the life areas of ideology and occupation. Some of the childhood constituents whose development is necessary to identity formation are described next.

2. Childhood Antecedents to Identity

2.1 Eriksonian View

From an Eriksonian psychosocial developmental perspective (Erikson 1980), the childhood precursors to the development of identity are the more or less successful resolution of preceding psychosocial stages. These involve the formation of predominant senses of basic trust (from infancy), autonomy (from early childhood), initiative (from play age), and industry (from school age). Basic trust (vs. mistrust) refers to the infant's sense that the world (and, hence, oneself) is trustworthy and largely benevolent and predictable. A sense of autonomy (vs. shame, doubt) is formed as one modifies one's will (in areas such as toilet-training and locomotion) to coincide with the demands of significant socializing figures in such a way that an eventual feeling of justice is attained. During the

periods of trust and autonomy, a self-structure (described in Sect. 2.2) is also being formed. Initiative involves the harnessing of previously developed autonomy in early childhood to the planning and completion of projects, both physical and interpersonal (viz., oedipal). Initiative is a crucial component in the ability to 'play,' both by oneself and with others. The immediate psychosocial precursor to identity is a sense of industry which springs from the child's experience of self as a competent worker with peers and teachers in school. As each psychosocial stage's challenge is met and resolved in an essentially positive manner, these accrued ego strengths furnish the adolescent with the psychological foundations to resolve successfully the identity issue of adolescence. Failure to optimally resolve earlier stages jeopardizes, but does not obviate, identity development, because every stage is assumed to occur in some form at every other stage. A particular psychosocial stage is not just resolved or unresolved once and for all time. Hence, all of the stage issues preceding identity recur at adolescence and there is an opportunity for their reresolution. However, because the adolescent must now face not only the present issue of identity but past ones as well, it is necessary to provide the person with a better-than-average expectable environment in the form of psychotherapy, counseling, or special programs aimed at enhancing psychosocial development.

2.2 Object Relational Perspective

Object relations theory is that branch of psychodynamic thought that focuses on relationships being more crucial to personality development than are individual drives and abilities (see Greenberg and Mitchell 1983). Here, the important identity-preceding structure is the self, a personality structure formed out of interpersonal interactions. The view of the development of the self presented here combines the thought of Bowlby (1982), Mahler et al. (1975), and Kohut (1977). The self is formed in infancy and early childhood (up to four years of age) out of the internalized and 'metabolized' interactions between the child and significant other persons (sometimes called 'self-objects'). Although the origins of the self lie in self-other interchanges, the self is experienced as one's own, and one comes to sense one's existence as both a separate and interdependent being. The conditions necessary for the establishment of a self include an initial period of undifferentiated symbiosis (Winnicott's 'dual unity') with a mothering/caregiving figure, differentiation from that figure, attachment to the caregiver and other significant objects, exploratory back and forth movements from the attachment figure, and eventual individuation as the self–other interactions become internalized and take the form of a secure self. Concurrent with, and following, the establishment of a self are conditions, described by Kohut (1977), that promote the maintenance and enhancement of that self. These are the presence of several kinds of 'self-objects,' namely 'mirroring' (an object who reflects the infant's ambitious grandiosity), 'idealizing' (an object who can be admired, identified with, and whose strength can be shared), and twinship' (a 'best friend,' a companion of one's heart). These self-objects continue to be important to self-maintenance and enhancement throughout the life cycle.

2.3 Classical Psychoanalytic Approach

The important identity-preceding structure described by classical psychoanalytic theory is the superego (Freud [1933] 1964). This personality structure is assumed to be formed at about five to seven years of age and is based upon internalizing the values of parental figures in order to defend against experiencing negative feelings occasioned by fear of punishment or of loss of love. The conscience aspect of the superego reflects parental prohibitions: what one ought not to feel, be, or do. The ego ideal portion of the superego is composed of parental aspirations: what one should feel, be, or do. Hence, the superego provides an internalized 'parent,' experienced as oneself, that is a source of guilt and shame, as well as a source of self-esteem. It is especially the ego ideal portion of the superego that undergoes modification during the identity formation process of adolescence. During this period, which may be seen as a second differentiation phase (Blos 1967), adolescents may substitute contemporary popular figures (e.g., sports or music stars) or literary or scientific heroes for original ego ideal parental introjects in efforts to become their own persons, rather than, or in addition to, being only their parents' child.

2.4 Summary of Perspectives

In summary, there are three important personality structures posited by different branches of psychodynamic theory, each taking form at different points in the individual life cycle. The self (an object relational concept) is formed in infancy and early childhood out of the internalized and transmuted interactions between the individual and important self-objects. The superego (a classical psychoanalytic concept) is developed during play age and is based upon the internalization of parental prohibitions and aspirations. An identity (an ego psychoanalytic concept) is formed during adolescence and is indicated by self- and environmental exploration and eventual commitment in significant life areas. A self establishes *that* one is; a superego determines *how* one should be; an identity is a statement of *who* one is. A self and a superego are necessary, but not sufficient, conditions for the formation of an identity.

3. Empirical Work: The Identity Statuses

The four identity statuses (Marcia et al. 1993) are ways in which a late adolescent might be found to be dealing with the identity issue. They are intended to reflect, on a behavioral level, the nature of the identity structure described by Erikson. Criteria for determining the identity statuses are the presence or absence of exploration and commitment in the areas of occupation, ideology, and interpersonal values. Identity status is typically assessed by means of a structured interview scored with a standardized manual, although sometimes a questionnaire is used. 'Identity achievement' persons have undergone a period of exploration and have made ideological, occupational, and interpersonal value commitments. 'Moratorium' individuals are currently in the exploration period, actively searching among alternatives. When this exploration process becomes emotionally fraught, these kinds of moratoriums are described as being in an 'identity crisis.' 'Foreclosure' persons are those who are committed in the interview content areas, but who came to those commitments with little or no exploration; usually, they have adopted directions laid down for them by parents or other early authority figures. 'Identity diffusions' persons are uncommitted in important life directions and are not currently engaged in a process leading to commitment (as are the moratoriums).

Over 35 years of research has established empirically some of the following characteristics of the four identity statuses. Identity achievement individuals have been found to be resistant to experimental attempts to raise or lower their self-esteem, to be nonconforming to group pressure, to think effectively under stressful conditions, to employ the higher levels of moral reasoning, to be advanced in intimacy development, to be complex in thought processes, and to come from families where differences among members are acknowledged and accepted. In addition, they appear to have a strong and autonomous sense of self, are capable of secure attachment relationships, and are realistically high in self-esteem and relatively unimpeded by rigid superego strictures.

Moratorium persons, like identity achievements, are relatively stable in self-esteem and not easily moved to conform. Sometimes their level of cognitive performance exceeds that of achievements and they tend to function at the highest levels of moral thought. Occasionally, they vacillate between harsh self-judgment and lax self-permission. They are the lowest among the statuses in their endorsement of authoritarian values, probably reflective of their attempts to differentiate themselves from their parents. Their relationships with their families are ambivalent, and they tend to be somewhat insecure in their attachment, at least for the duration of their moratorium period.

Foreclosure individuals are the least cognitively flexible of the statuses and the most highly endorsing of authoritarian values, suggesting a relatively unmodified ego ideal. They employ conventional reasoning about moral issues and either acquiesce superficially to, or stubbornly resist, positions discrepant from their own. In relationships, they give the appearance, but not the substance, of intimacy. They describe their families in unrealistically ideal terms. In terms of attachment, they are about evenly divided between secure and insecure attachment patterns.

Identity diffusion persons are vulnerable to self-esteem manipulation and become disorganized in their thinking when under stress. They are at the lower levels of both moral reasoning and capacity for intimacy. Diffusions experience a marked distance between themselves and their parents, especially the parent of the same sex, whom they feel that they can neither emulate nor please. It is not surprising, then, that they are the most insecure in attachment among the statuses.

It should be kept in mind that most of the studies that have furnished the above findings have involved the use of the identity status interview, have been done in Western, technologically advanced countries, and most of them have involved college students. However, there have recently been a fairly large number of studies being conducted in developing countries and with other social classes and ethnic group, especially by Jean Phinney and her colleagues (see Roberts et al. 1999). Whether or not the foregoing descriptions of the identity statuses will obtain with these noncollege groups remains to be seen. One might expect, for example, a foreclosure individual to appear differently in a culture that prizes foreclosure than in one valuing exploration and change. In any case, the interview method would provide more flexibility in such investigations than the more static questionnaire measure.

4. Identity Development in Adulthood

While identity is first consolidated at late adolescence, this is only the initial identity. An identity can be expected to undergo successive transformations at least with each life cycle stage (see Kroger 2000). Hence, the identity serviceable to the late adolescent will probably be reformulated in early adulthood when intimacy is the primary psychosocial concern. Another change can be expected in middle adulthood around issues of generativity. Finally, integrity at old age poses perhaps the greatest challenge to the reconstruction and maintenance of an identity in the face of necessary physical, mental, and interpersonal losses. The processes by which identity changes in adulthood appear similar to those described by Piaget (1972) to account for changes in cognitive structure: assimilation, disequilibration, and accommodation (Whitbourne and Connolly 1998). An identity assimilates experiences so long as they are comprehensible within the identity structure. When they can no longer

be sufficiently accounted for by the individual's identity 'theory,' a state of disequilibration exists, and a subsequent accommodation, or change, in identity structure is called for in order to handle the new experiences. In a sense, an achieved identity contains the seeds of its own destruction. Because the identity was formulated to deal with certain experiences, it affords some efficiency in experiential processing and permits one to be aware of new experiences previously not within awareness. The growing awareness of experiences not assimilable in the current identity leads eventually to the disequilibration of that identity (a moratorium period) and to the necessity of formulating a new one (achievement). As life experiences accumulate in adulthood, prior identities are modified and made more serviceable. These achievement-moratorium-achievement (MAMA) cycles (Stephen et al. 1992) are not undertaken by all persons. There are some who foreclose at adolescence and, by means of employing defensive processes, or by remaining in foreclosing settings, never experience disequilibration and identity reformulation. Also, identity diffusions may drift through life taking on, superficially, whatever identities are provided to them by their environments. The assumption in identity theory is that MAMA cycles are probably the most developmentally adaptive trajectories.

Research has shown some relationships between identity development and the resolution of adult psychosocial stages following adolescence. Achievements and moratoriums are more advanced in intimacy than are foreclosures and diffusions (Marcia et. al. 1993). Persons who are resolving positively the generativity issues of middle age look very much like identity achievements; and those who are having difficulty here appear much like foreclosures and diffusions (Bradley and Marcia 1998). Persons characterized as integrated at old age are more likely to be identity achieved than those not so integrated. (Hearn et al.) Thus far, however, all studies involving linkages between identity and preceding or succeeding psychosocial stages have employed simultaneous measurement. Except for research within the identity development period (approximately ages 13–23), there is no longitudinal research linking earlier stages with identity for the same individuals.

5. *Criticism and Suggestions for Further Research*

Both the Eriksonian approach to identity, as an individual personality variable, and the identity status approach to the study of Eriksonian ego identity have not been without their critics. Some theorists see society constructing individuals' identities, rather than individuals' constructing their own identities out of social givens (e.g., Gergen and Gergen 1997). These critics, though, have difficulty in explaining individual differences in identity content and process within the same culture. Some writers have questioned the degree to which the identity statuses fully represent Erikson's ideas (e.g., Cote and Levine 1988, but see also Waterman 1988). However, these critics overlook the substantial nomological network of studies that have actually yielded accrued meaning to both the identity statuses and to Erikson's identity construct. Similar to any scientific construct, the identity status paradigm is a 'work in progress.' Yet, it remains the predominant method for studying the development of ego identity in adolescence.

There is a clear need for longitudinal studies on a developmental theory such as Erikson's. This includes tracing, empirically, the childhood antecedents of adolescent identity as well as the transformations of identity throughout the life cycle. What are the necessary childhood conditions for identity formation at adolescence? What does identity look like at adulthood, at old age? In addition, now that we have some validated measures of the adult psychosocial stages we can be more specific about identity development's impact on their resolution. Is identity formation at adolescence a necessary condition for, say, generativity at middle adulthood? Additional areas of research concern include the impact of social conditions on identity formation. What kinds of social contexts foster or restrict identity development at adolescence? How does identity differ cross-culturally?

Who one is, to oneself and to others, is a lifetime project and a singularly cherished accomplishment and possession. Human history is replete with stories of persons and nations willing to go to any lengths to preserve or establish their own identities. What is significant about Erikson's approach to identity, in addition to the richness of his description, is that he locates its initial formulation within the individual lifecycle as the specific developmental task of adolescence, to be continued by means of successive reformulations throughout adulthood.

See also: Europe: Sociocultural Aspects; Everyday Life, Anthropology of; Food in Anthropology; Genealogy in Anthropology; Identity and Identification: Philosophical Aspects; Personal Identity: Philosophical Aspects; Self-development in Childhood; Self-evaluative Process, Psychology of; Self-knowledge: Philosophical Aspects; Self-monitoring, Psychology of; Self-regulation in Childhood

Bibliography

Blos P 1967 The second individuation process of adolescence, *Psychoanalytic Study of the Child* **23**: 245–63

Bowlby J 1982 *Attachment and Loss. Vol. I: Attachment*, 2nd edn. Basic Books, New York

Bradley C L, Marcia J E 1998 Generativity-stagnation: A five-category model. *Journal of Personality* **66**(1): 39–64
Cote J E, Levine C 1988 A critical examination of the ego identity status paradigm. *Developmental Review* **8**(2): 147–84
Erikson E H 1980 *Identity and the Life Cycle*. Norton, New York
Freud S [1933] 1964 New introductory lectures on psychoanalysis. In: J. Strachey (ed. and trans.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London, Vol. 22
Friedman L J 1999 *Identity's Architect: A Biography of Erik. H. Erikson*. Scribner, New York
Gergen K J, Gergen M M 1997 Narratives of the self. In: Hinchman L P, Hinchman S K (eds.) *The Idea of Narrative in the Human Sciences*. State University of New York Press, Albany, NY, pp. 161–84
Greenberg J R, Mitchell S A 1983 *Object Relations in Psychoanalytic Theory*. Harvard University Press, Cambridge, MA
Hearn S, Glenham M, Strayer J, Koopman R, Marcia J E 2001 *Between integrity and despair: Toward construct validation of Erikson's eighth stage*. Under review. Simon Fraser University, Burnaby, BC, Canada
Kohut H 1977 *The Restoration of the self*. International Universities Press, New York
Kroger J 2000 *Identity Development: Adolescence Through Adulthood*. Sage, London
Mahler M, Pine F, Bergman A 1975 *The Psychological Birth of the Infant*. Basic Books, New York
Marcia J E, Waterman A S, Matteson D R, Archer S A, Orlofsky J L 1993 *Ego Identity: A Handbook for Psychosocial Research*. Springer-Verlag, New York
Piaget J 1972 Intellectual evolution from adolescence to adulthood. *Human Development*. **15**: 1–12
Roberts R E, Phinney J S, Masse L C, Chen Y R, Roberts C R, Romero A 1999 The structure of ethnic identity of young adolescents from diverse ethnocultural groups. *Journal of Early Adolescence*. **19**(3): 301–22
Stephen J, Fraser E, Marcia J E 1992 Moratorium-achievement (MAMA) cycles in value orientations and reasoning system correlates Lifespan identity development. *Journal of Adolescence*. **15**: 283–300
Waterman A S 1988 Identity status theory and Erikson's theory: Communalities and differences. *Developmental Review* **8**(2): 185–208
Whitbourne S K, Connolly L A 1998 The developing self in midlife. In: Willis S L, Reid J E (eds.) *Life in the Middle: Psychological and Social Development in Middle Age*. Academic Press, San Diego, CA, pp. 25–45

J. E. Marcia

Identity Movements

The main objective of social movements is to transform society so as to reflect the social actors' interests and view of the world. Such was the case with the labor movement in the early twentieth century and the decolonization movement in the 1960s (see *Social Movements, History of: General* and *Social Movements: Resource Mobilization Theory*; Chazel 1992, Oberschall 1983). The identity movements that came to the fore in the last third of the twentieth century are of a different nature, because they express two complementary types of collective demands: (a) the defense of interests and the promotion of rights of certain groups of individuals who feel discriminated against, and (b) the search for symbolic recognition by a significant other.

Identity movements defend the interests, world visions, and values of groups of individuals or communities defined by such characteristics as phenotype (or *race*) and ethnicity; sex; language; sexual orientation; mythical origins and ancestral territory, in the case of First Nations in the Americas and the South Pacific; and religion, in the case of certain societies. Melucci (1989) even posits that *all* social movements have an identity dimension (on the link between social movements and identity, see Jasper 1997, Larana et al. 1994).

Identity movements have three main and very distinct objectives. First, they denounce injustice toward minorities. Second, they convey the idea that specific cultures must be taken into consideration when public policies are elaborated so that they meet the specific needs of minorities. Third, they demand greater control of their institutions—a demand that sometimes goes as far as self-government.

Individuals involved in identity movements promote their own interests (Olson 1965) based on their way of seeing things and their personal knowledge and values. Their action is therefore marked by instrumental, cognitive, and axiological rationales. Boudon and Bourricaud (1982/2000) have shown that the two dimensions of collective behavior described by Smelser (1962), relation with norms and relation with values, are so closely linked as to become interdependent. The shared values of the members of a group or community play a central role in emerging identity movements, at least as important as the pursuit of individual interests.

This article will specify what should be understood by identity, define the characteristics of different types of identity movements, and discuss the relationship between the universal and particularistic norms which are specific to each movement.

1. *Identity*

In the early twentieth century, identity was often defined according to 'objective' traits, such as English tenacity, German discipline, and French ingenuity. What we now call identity was then known as the 'soul of peoples' (see Siegfried 1950). Today, the problem of identity is no longer seen in this light (see *Social Identity, Psychology of*).

It is no longer possible to separate personal and collective identities. Individuals now belong to several social networks and groups, which, in turn, shape their identity, and they draw from society the elements that they need for self-definition. Today, they are faced

with a great number of choices, many more than those of preceeding generations. Education, individual and collective fulfillment, social mobility, career mobility, geographical mobility, and lifestyle are some of the social processes that help to build identity.

Individuals play a key role in the building of collective identities especially in countries such as Australia, Canada, and the United States, where the mass arrival of immigrants shapes the basis on which identity is redefined. Canada, for example, which welcomed more than eight million immigrants between 1945 and 2000, defines itself no longer as a binational society but as a bilingual and multicultural one. Australians are questioning the relevance of maintaining the royal symbols that are remnants of their colonial past. The same can be said of older countries such as England or France, whose present collective identity is marked by changes in individual lifestyles and moral values and the intermingling of peoples of diverse ethnic origins. Aspects of Ancien Régime France and Victorian England coexist with their more modern counterparts. Such contrasts can now be found in various degrees in every society. Immigration, urbanization, international trade, unequal regional development, differing interpretations of the past, social conflicts, and social stratification affect the construction of collective identities.

Taylor (1989) defines the concept of identity three ways. First, identity defines the moral outlook of each individual, in the sense that individuals are able to determine what is important to them. This corresponds to Erikson's (1968) conception that individual identity implies self-knowledge and a successful and accepted synthesis of diverse influences, especially those of the individual's family.

The second definition came into being with modernity and has enabled individuals to have a better grip on their own destiny. Theoretically, within the limits of their capacities, people can become what they want to be; the future is no longer dictated solely by ascribed status in society. Political freedom, education, and economic development have enabled individuals to innovate, and, to a certain point, be responsible for their future. Each individual may have diverging or even conflicting interests, but each will eventually define values that will be expressed through various social movements and collective conduct.

Taylor insists that self-affirmation and recognition by a significant other are the two conditions for achieving a successful identity. Social actors involved in identity movements are motivated both by what they see as a deficiency in democracy, and by the search for recognition which involves negotiation and sometimes struggle and which is probably the main feature of such movements.

Because individuals are also part of a people or nation and therefore also define themselves by their collective past, national identity is the third definition that Taylor gives to the concept of identity. A nation or historical community offers people another perspective made up of a shared culture, lifestyle, and language (Todorov 1995). Nations are also imagined communities (Anderson 1991). However, most countries are not homogeneous; in the second half of the twentieth century, they have had to deal with the problem of recognition of various nations within their own midst. One can therefore say that the relationship with the other, which is at the heart of identity assertion, is an internal problem in most countries.

Dumont (1974) distinguishes three types of human communities, each with a particular way of building self-identity: primary groups (small groups based on face-to-face interactions), integrated groupings (in which individuals play interchangeable roles), and reference groups (defined by a sense of belonging, an awareness of forming a community). Diversity, plurality, conflicts, and differences become inevitable as social groups grow in size, and the need to adopt a common reference then becomes essential to ensuring cohesion. Evidence of adhesion must eventually be transposed into a series of collective signs and symbols in order to become conscious references for each member within the grouping. Dumont goes on to describe the constantly repeated mechanisms that contribute to the building of identities: ideologies, literature, and historiography (to which, today, we could add the media) play a key role in the shaping of a common reference. He speaks of conflicting interpretations between various categories of actors involved in the process of building identities, and defines society as a series of interpretive practices.

The Scottish example is relevant here. Language is not a central issue in defining Scottish national identity, despite a recent interest in Gaelic; more important is identification with a mythic territory, a common history, and recollection of ancestral wars (Macdonald 1997). The situation is quite different in Quebec and Catalonia, where promotion of the national language (French and Catalan, respectively) plays a central role in the building of the respective identities.

2. *Identity Movements*

Four types of identity movements will be characterized by the cognitive and axiological systems that determine their action (see Boudon 1999), the means and field in which the action takes place, and the meaningful others involved in the search for recognition.

2.1 *National Identity Movements*

Movements of national affirmation are undoubtedly the best-known types of identity movements. Gellner (1997) estimates that there are 15,000 national cultures demanding recognition in a world where there are only

a few hundred nation-states. These movements challenge the dominant national reference in their respective countries, and they are determined to achieve official recognition. They are composed of national minorities—not to be confused with ethnic minorities resulting from immigration, who make demands of another nature.

National minorities fight social inequalities of which they see themselves as victims, and many demand that official institutions be modified. These types of protests generally give way to a search for a new constitutional framework within a nation-state. Situations and demands, in accordance with the historical and national contexts, are extremely diversified (Calhoun 1994).

2.2 *Aboriginal Identity Movements*

The self-assertion of aboriginal nations in countries shaped by influxes of immigrants must be considered separately. There are approximately 250 million aboriginals living in these immigrant societies throughout the world, mainly in the Americas and the South Pacific. This enormous population (equivalent to half of that of Europe or two-thirds of that of the United States) is divided into several thousand nations or distinct groups. In Canada, for example, there are 633 different bands and aboriginal nations—some groups numbering only a few hundred people—speaking 57 different languages.

Conquest and colonization by Europeans marginalized aboriginal nations, sometimes moving them to reservations, as is still the case in Canada and the United States. These peoples have been the victims of serious historical injustices, and some of them were nearly extinguished by warfare. Today, they are demanding control of a number of institutions that will help them maintain their specific collective identities. Some aboriginal nations in the United States, the Sioux for example, and a large number in Canada and Australia are striving for self-government within their respective countries, but their small populations prevent most of them from claiming complete sovereignty.

Aboriginal communities build their action on three main aspects of their value system. First, there is the feeling of a common mythical origin. Second, and probably at the heart of their culture, is their relationship with nature, the land, and ancestral territory. Third, is the idea of kinship and blood: aboriginal communities generally do not integrate strangers, for one is either born Native or one is not.

The search for symbolic recognition is particularly important to aboriginal communities, since, traditionally, their culture has been denigrated or denied. The leaders of these communities are now striving to build a new sense of identity in order to counteract the feeling of self-contempt. This explains why, for many of them, interpretation and self-definition have become so important.

2.3 *Particularistic Identity Movements*

Other identity movements do not question existing institutions but aim for the rightful participation of specific groups of individuals, sometimes even imposing various modes of participation suited to their particular characteristics. Such is the case with linguistic or ethnic minorities that insist on having better access to important positions in the civil service and business organizations.

This type of movement includes the great majority of the feminist movements that demand an end to sex discrimination in public institutions and private enterprises. Although the feminist movement has very diverse components, it presents itself as an identity movement because it defends the interests of a portion of the population who feel wronged simply because they are women, and because of their opposition to a meaningful other who, in turn, becomes the main point of reference: men. Groups of women now insist on empowerment and their share of all higher appointments in various occupations. The most radical feminist movements go so far as to claim that all social relations are gender relations.

The civil rights movement for black Americans, which climaxed under the leadership of Martin Luther King, is also an identity movement (Morris 1984). Even though this movement is typically American, in the second half of the twentieth century, it has taken on a very important symbolic value in most Western countries; it also gave rise to programs for affirmative action in favor of visible minorities in the United States.

2.4 *Normative Identity Movements*

Other identity movements set out to modify social norms that affect the private, civil, and professional lives of individuals. Actors in these movements aim for symbolic acceptance of their differences and want to change social stereotypes. These differences are of a sexual, intellectual, or physical nature, such as the visible characteristics of certain immigrants (Bauman 1996). Movements based on sexual orientation and in defense of the handicapped are good examples of this type, whose participants feel that they are the victims of prejudice. Their complaints have been legally validated by many countries and have helped to change public attitudes. In most developed countries, homosexuality is no longer seen as a criminal offence and is officially accepted in some institutions such as the armed forces. Today, some countries even acknowledge same-sex unions.

3. *Universalism vs. Particularism*

Identity movements and the associated communitarian philosophy see values as being interwoven with specific cultures (Touraine 1997, Elbaz and Helly 2000).

Some analysts worry about the negative consequences of cultural relativism underlying certain demands put forth by identity movements (see Boudon 1999, Hardin 1995). These critics emphasize that communitarian norms easily converge with the interests of the actors involved in identity movements, whereas universal norms seem to be weaker and take more time to surface because they do not serve the immediate interests of individuals. In this sense, Hardin (1995) has demonstrated that mobilization by spontaneous coordination may even bring about conflicts between different groups.

For other analysts, countries and modern constitutions today have no option but to accept and make the most of cultural diversity (Tully 1995). Multiculturalism—which has three different meanings and refers to a sociological reality, a philosophy, or a state policy—must be held up as an example. Kymlicka (1989) claims that multiculturalism does not split society up into isolated subgroups but, on the contrary, offers a common framework within which differences may blossom in a context that respects diversity.

See also: Action, Collective; Communitarianism, Sociology of; Ethnic and Racial Social Movements; Ethnic Conflicts; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity, Sociology of; Feminist Movements; Identity in Anthropology; Minorities; Multiculturalism: Sociological Aspects; Race Identity; Race Relations in the United States, Politics of; Racial Relations; Social Movements and Gender; Social Movements, Sociology of

Bibliography

Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Bauman G 1996 *Contesting Culture: Identity in Multiethnic London*. Cambridge University Press, Cambridge, UK

Boudon R 1999 *Le sens des valeurs*. Presses Universitaires de France, Paris

Boudon R, Bourricaud F 1982/2000 *Dictionnaire de sociologie*. Presses Universitaires de France, Paris

Calhoun C (ed.) 1994 *Social Theory and the Politics of Identity*. Blackwell, Oxford, UK

Chazel F 1992 Mouvements sociaux. In: Raymond B (ed.) *Traité de sociologie*. Presses Universitaires de France, Paris, pp. 263–312

Dumont F 1974 *Les idéologies*. 1st edn. Presses Universitaires de France, Paris

Elbaz M, Helly D (eds.) 2000 *Mondialisation, citoyenneté et multiculturalisme*. Les presses de l'Université Laval, Québec and L'Harmattan, Paris

Erikson E H 1968 *Identity: Youth and Crisis*, 1st edn. W. W. Norton, New York

Gellner E 1997 *Nationalism*. New York University Press, Washington Square, New York

Hardin R 1995 *One for All: The Logic of Group Conflict*. Princeton University Press, Princeton, NJ

Jasper J M 1997 *The Art of Moral Protest: Culture, Biography, and Creativity in Social Movements*. University of Chicago Press, Chicago

Kymlicka W 1989 *Liberalism, Community, and Culture*. Oxford University Press, Oxford, UK

Larana E, Johnston H, Gusfield J R (eds.) 1994 *New Social Movements: From Ideology to Identity*. Temple University Press, Philadelphia, PA

Macdonald S 1997 *Reimagining Culture: Histories, Identities and the Gaelic Renaissance*. Berg, Oxford, UK

Melucci A 1989 *Nomads of the Present: Social Movements and Individual Needs in Contemporary Society*. Temple University Press, Philadelphia, PA

Morris A D 1984 *The Origis of the Civil Rights Movement: Black Communities Organizing for Change*. Free Press, New York

Oberschall A 1983 *Social Conflict and Social Movements*. Prentice-Hall, Englewood Cliffs, NJ

Olson Jr. M 1965 *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge, MA

Siegfried A 1950 *L'âme des peuples*. Hachette, Paris

Smelser N J 1962 *Theory of Collective Behavior*. Free Press of Glencoe, New York

Taylor C 1989 *Sources of the Self. The Making of the Modern Identity*. Harvard University Press, Cambridge, MA

Todorov T 1995 *La vie commune. Essai d'anthropologie générale*. Éditions du Seuil, Paris (coll. La couleur des idées)

Touraine A 1997 *Pourrons-nous vivre ensemble? Égaux et différents*. Fayard, Paris

Tully J 1995 *Strange Multiplicity. Constitutionalism in an Age of Diversity*. Cambridge University Press, Cambridge, UK

S. Langlois

Identity: Social

Problems of social, group, and collective identity have become a major focus of contemporary research. Such identity supposedly is determined by ethnic or other important characteristics. Interest in these issues presumably has been sparked by two factors. First, there are concerns of and concerns for minority ethnic groups in many nations, either from immigration, as in the US, or from historical intermingling, as in Belgium and most nations of the world. Second, the assertion of ethnic and gender identity has been a central element in recent politics in many nations. Much of the research on social identity has an inescapably normative tone, and much of it is written as though by advocates of particular identities or even, implausibly, by advocates of identity *per se*.

A social scientist must view identity as largely a social construction that is malleable to some degree, especially in younger ages, and as plural rather than singular, at least in advanced societies. People who

declare their identity commonly view it as objective and normatively commanding. Hence, identity and identity politics pose a sharp example of the common phenomenon that social science explanations are rejected by the very people whose behavior they ostensibly explain. This may be part of the reason that, especially, postmodern advocates of the importance of identity reject explanations as ideologically corrupted.

There are three central issues in social identity that will be discussed here. What does it mean? What are its moral implications? And what motivates people to act on their identities? The first question raises the issue of what it means both to a social scientist and also what it means to an individual who has some putative identity. The second involves claims about what actions or policies are justified by aspects of social identity, especially when these characterize substantial groups. The last question shifts the focus from supposedly objective facts of identity to subjective issues of motivation and, hence, as discussed below, from identity to identification. This question arises independently of whether identity is socially constructed or objective. For example, philosophers once supposed the possibility of an ardent Nazi, so ardent in his commitments as even to think it right that, once his Jewish identity was discovered, he should be expelled from Nazi society.

These three questions have answers that are not entirely separable, because they inform each other. The third question involves issues of explanation. How do we explain an individual's identification with a particular group? If meanings are theory-laden, then the first question also involves issues of explanation.

The most common categories of social identity are racial, gender, religious, national, communal, and ethnic, all of which are taken to be chiefly or at least partially ascriptive. We might add familial, professional, and other identities to this list for at least many people, but for many of the theorists of social identity these categories suggest the breakdown of the central idea, as the plural ties of modern societies also commonly do. Indeed, some of the discussions of identity take the form of advocacy of a restoration of some of the earlier forms of social life, without the corrosive effects of modern liberalism and pluralism on community.

There is a long history of the analysis of individual or personal identity by psychological theorists, from James (1981) forward. And there is a rich history of the effort to socialize the idea of a person or the self, from at least Mead (1934; see Joas 1997). An implication of much of the latter literature is that social identity is important in large part because it determines personal identity so heavily. Advocates of social identity sometimes are accused of having an oversocialized conception of the person. Some of the strongest contemporary advocates of social identity share the postmodern and communitarian rejection of any strong notion of individual autonomy, such as has descended from the universalist stances of the utilitarians and Immanuel Kant. They therefore think it makes no sense to speak of your identity without essentially definitive recourse to your social context and background. Bridges between the psychological study of personal identity and the sociological study of social identity are, however, few and often narrow (but see Taylor 1989). Most of the discussion is not empirical and it often has the strong scent of ideological pleading. This is a topic that is sorely in need of explanatory accounts.

One of the great strengths of the psychological approach to personal identity is that it at least struggles with and has some chance of explaining the formation of such identities. For example, Sigmund Freud and many others have argued for the 'narcissism of minor differences' (Blok 1998). Bourdieu (1984) supposes that 'Social identity lies in difference, and difference is asserted against what is closest, which represents the greatest threat.'

If we are to understand social identity, we probably need to have accounts and explanations of its formation (Hardin 1995, Laitin 1998). Commonly, its genesis simply is assumed and accounts start from a present, assumed identity. Indeed, many writers even give credit to some notion of collective memory that connects, say, contemporary Serbs with the dreadfully lost war with Turks in 1389 in the Field of the Blackbirds. Somehow, the English seem to be exempt from the effects of the dreadful memories of Hastings Field and Americans do not even remember, much less resent the burning of the White House by the British. The assertion of historical memory contributes to mystification rather than explanation or understanding. Such mystification is, in a famous remark of Ernest Renan, required for nationalism, because one can be committed to the rightness and purity of a nation only by forgetting its brutal and messy past. Mystification is also required for many claims of group identity.

1. *The Meaning of Social Identity*

Many descriptive accounts of social identity virtually essentialize it to suppose that, if we know relevant descriptive facts about you, we know your social identity. Against any such move, others note that even a descriptive account could conclude that your identity is quite varied. You are an Italian, Catholic, New York accountant with particular social activities in specific groups. There may be no one you know who shares all of these characteristics that are, together, central to your own life. Hence, it would make sense to say that you have multiple identities, which likely are evoked in different contexts or on different issues. But maybe you dislike your Italian neighbors and their vocal style, you have rejected Catholicism and have stopped going to mass, and you wish you could move

out of New York and change your profession. All of the things you now reject might nevertheless have effects on you and might help others to understand parts of your behavior and attitude, so that in some explanatory sense they are part of your identity. Still, in your case, there is an odd difference between the use of your characteristic identity to explain things you do or think and your actual commitments. Much of the literature that explicitly or implicitly justifies actions from identity takes relevant commitments for granted once identity is established. But your case suggests that this correlation of ascriptive characteristics and commitments can be entirely wrong. The idea of social identity in a pluralist world in which individuals themselves have plural commitments seems to break down, and the recent efflorescence of communitarian thought is largely a response to this breakdown (see e.g., Etzioni 1993). In a plural society, the philosophical concern with identity and its fundamental meaning for even a mere physical object becomes acute. The traditional philosophical example is the boat that is rebuilt over many years, one rotted plank at a time. Suppose no piece of the original boat remains but that 'the' boat continues to sail. Is it the same boat? Are you in any meaningful sense the same person who was born and lived for many years in a different society before you migrated, went to university, developed radically different tastes, rejected your prior religious views, and lost contact with virtually everyone you knew earlier, and so forth? And if personal identity is in question, can social identity be any more stable?

Identities that are asserted forcefully are often virtually invented by leaders with agendas (on the deliberate creation and manipulation of national identities, see Anderson 1991, Hardin 1995, Hobsbawm 1990, Lebovics 1992). Many of the communities with which individuals identify are hierarchical or otherwise diverse in ways that suggest there is not a unique identity or kind of person at issue. The ascriptive variables on which identity might be based are, as noted, numerous and varied. It is therefore almost impossible to talk about social identity without making the term denotative, without virtually pointing and saying, '*That* is an identity.' What is missing is some standard set of characteristics that defines a social identity beyond the fact that some significant numbers of people assert that they have such an identity. Social identity theory is woefully undertheorized.

2. *The Ethics of Social Identity*

Much of the contemporary literature on identity is normative, not causal, and is therefore unlike the early Chicago school of sociological work on ethnic minority groups. It is driven partly by the post-modern rejection of explanation and its focus on description and normative judgment that seems to flow merely from description, contrary to the famous principle of David Hume that one cannot deduce a moral conclusion from a mere fact, of an 'ought' from an 'is.' Of course, specific claims of identity in actual contexts typically are normatively loaded. They commonly are claims for political rights or redress of wrongs (Kymlicka 1989).

How should the social theorist or policy maker respond to identity claims? We generally do not honor relativistic moral claims, such as that you or your group has a special status that others must honor. Rather, we expect moral claims to be universalistic in some sense. For you to make a moral claim, you must grant implicitly that all who are in similar circumstances, somehow defined, can make the same moral claim. If you insist that your claim should motivate others, the claims of those in a position analogous to yours should motivate you to some extent. Assertions of the rightness or goodness of one's own group and its claims often are little more than group solipsism. Adding the appellation 'group' to solipsism does not rescue it from the moral opprobrium that solipsism normally has.

When an identity is enforced on an individual (or group) by others, as in contexts of racism, we might agree easily that there is a moral aspect to the individual's claims from identity. Indeed, racist treatment seems to be a simple case of universalistic concern. It is perhaps this quality, when it is present, of much of the assertion of identity by suppressed groups in their own defense that has motivated concern with group identity and that has elevated it to the status of a moral issue.

Actual supporters of group identity in many contexts would blanch at the thought of support for, say, Caucasian American, Anglo-Saxon, or German identity, although each of these historically has been asserted against immigrant groups and, in the US, former slaves. For contemporary identity theorists, the only apparent candidates for claims of rights of identity are minorities, those in secondary status, and third world populations. But these are not moral statuses *per se*. Hence, we would have to be articulate about what characteristics of these groups entitle them to special concern. A particular social identity cannot be good or right except contingently. Otherwise, it would be wrong of me to reject mine—or, more typically, some of mine.

While we could have compelling explanatory discussions of social identity, the contemporary literature to date has few such discussions. Instead, authors commonly take identities for granted and they focus on the normative qualities of social identity and, typically therefore, of communal life. They often pose such life as good and as hostile to traditional liberal principles of universality and autonomy, which would seem to entail opening the doors of community to free entry by outsiders, much to the detriment of the

community's ur-members. To establish that something is good, we might show that it follows from some theory of the good or that it fits human nature or that it was brought about by a relevant procedure. Communitarianism is essentially proceduralist about the good. If the community generates a practice, that practice is good for the community. This is a peculiar move in standard ethics, because proceduralism typically has been the path to the right, not to the good.

The question for a communitarian definition of the good is what the procedure is about and how it works. In law, procedures often are supposed to work for discovery of the truth or for protection of people from particular kinds of official abuse. Hence, they make right. Communal creation of values commonly is supposed by communitarians to make those values good, at least for the community that creates them. Because we assert our identity, that identity has moral weight, it is part of our good. Unfortunately, although practicing communitarians are almost always partial to their own group and hostile and even murderous to certain other groups, this implication has a universal quality that suggests that any group that asserts its identity has thereby defined the content of much of its good and the good of its members. It is partly therefore that communitarianism has taken a dramatically conservative turn and that defenders of the goodness of stultified, narrowly focused communities that often entail severe restrictions on the opportunities of their children have become communitarians. The quasi-universal quality of communitarian thought is about the universality of the way a group's good gets defined, not about the content of that good. Some communitarians seem to espouse a return to religion (for others) not because of the rightness of the content of the beliefs but because of the discipline and control of the self that religiosity brings.

3. *Motivations To Act*

Perhaps the least addressed of the three questions central to understanding social identity is what motivates individuals to act on their identities. In a sense, this question changes our focus from identity to identification, or 'a sense of inner identity,' which was a central concern of Erikson (1964). Such identification has not often been addressed in the contemporary claims for identity. Indeed, there is often not even a clear distinction between personal and social identity. There have been three major approaches to the analysis of identity beyond mere assertion or description of it. These are psychological, moral, and rational choice accounts.

The most influential psychological account has been that of Erikson, although Erikson's focus is not on social identity but rather on the individual's creation of an identity or a self, all, of course, under massive social influence. Despite the initial promise and compelling qualities of Erickson's work on identity, however, psychological accounts have often had an *ad hoc* quality that has perhaps kept them from being followed up extensively by anyone other than their original authors. There is, therefore, not a coherent body of work that gives clear insight into why one's objective circumstances of language, religion, skin color, gender, prior history, and so forth should motivate one to energetic and even risky action on behalf of whatever one might identify with.

It is not difficult to give a moral account of the motivations of many individuals who act on behalf of a group, as, for example, Frederick Douglass did for American blacks or as Mahatma Gandhi did for the lower castes of India. But it is difficult to construct a moral story that justifies Serbian ethnic cleansing or to see scurrilous leaders such as Slobodan Milosevic or Franjo Tudjman as moral. Many members of various ethnic groups do claim that they have been wronged morally and that their actions against others are morally grounded. But it is hard to moralize the ethnic cleansing in Yugoslavia, most of which has been carried out by soldiers under orders, or the spontaneous violence by uncontrolled mobs that have wreaked havoc and death on Sikhs in India after Indira Gandhi's assassination, on blacks during hundreds of lynchings in the US, and on Tutsis in Rwanda during the incomparably bloody rampage of the Hutu extremist *interahamwe*. Interestingly, these are all instances of majority ethnic groups attacking minority groups, and the ethnic groups committing these atrocities were in political power at the time (although the Hutus soon lost power).

One way of ostensibly explaining any specific motivation is to reduce it to other motivations that are well understood and that have much broader application. Some explanations of social identity implicitly posit it as a motivation of its own kind, a kind that has only relatively recently been discovered. This is a very unsatisfactory move because supposed social identities do not correlate with actions following from the motivation. Psychological accounts are relatively fragmented because they do not reduce the relevant motivations to some widely accepted core.

Rational choice explanations gain what power they have from the attempt to reduce particular motivations to some narrower set of motivations, often simply to interests, which can be moderately well defined and measured across a large variety of contexts and actions. In such an account of individual identification with a group, we look to see how certain interests of the individual might be satisfied by participation in the life of the group. And we look to see what interests members of the group might have to impose constraints on each other to enforce their commitments to the group. In essence, groups commonly produce norms of exclusion, whose content is rules for belonging to the group and whose chief means of sanction is the possibility of excluding

miscreants from the group. If membership has its pleasures, the sanction of exclusion has bite. It might be trumped by inducements from outside the group, and therefore many people leave groups with which they have previously been strongly associated. Indeed, they leave nations and home communities, even at great cost, for the benefits they might expect elsewhere. It is a striking feature of norms of exclusion that their sanctions are often easily motivated, while the sanctions against violating universal norms are very hard to motivate unless they can be brought to bear in essentially dyadic or small number contexts (Hardin 1995).

Such an account of their commitments is unlikely to satisfy members of a group, who are more likely, with Renan, to justify their commitments by mystifying the value and reason of the group. Unfortunately, the only account that would not have this problem of rejection by the group is the account the members themselves would give. That account is unlikely to lead to understanding and is typically likely to be almost entirely normative. The social scientist who tries to explain communal commitments and social identities is apt to lead an existence somewhat barren of such community. Gellner (1998) argues that moderns, with their individualistic life and thought, may suffer a romantic wish for, or even insistence on, community and a particular social identity. Wishing will not make it happen, and most moderns will be able to speak of their own social identity only in the plural, often the extremely plural, and, as James argued long ago, in a plural that is in frequent flux. An irony of the lives of the theorists of social identity and community is that they too are inherently plural in their own identities and that they typically lead their lives in the cold comfort of anti-communal, universalistic universities.

See also: Collective Beliefs: Sociological Explanation; Collective Identity and Expressive Forms; Communitarianism, Sociology of; Cultural Assimilation; Cultural Diversity, Human Development, and Education; Cultural Relativism, Anthropology of; Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Identity and Identification: Philosophical Aspects; Identity Movements; Integration: Social; Party Identification; Race Identity; Regional Identity; Transportation Planning

Bibliography

Anderson B R O'G 1991, 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London, rev. edn.

Blok A 1998 The narcissism of minor differences. *European Journal of Social Theory* **1**: 33–56

Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA

Erikson E H 1959 *Identity and the Life Cycle*. International Universities Press, New York

Erikson E H 1964 *Insight and Responsibility*. Norton, New York

Etzioni A 1993 *The Spirit of Community: Rights, Responsibilities, and the Communitarian Agenda*. Crown, New York

Gellner E 1998 *Language and Solitude*. Cambridge University Press, New York

Hardin R 1995 *One for All: The Logic of Group Conflict*. Princeton University Press, Princeton, NJ

Hobsbawm E J 1990 *Nations and Nationalism since* 1780: *Programme, Myth, Reality*. Cambridge University Press, Cambridge

Honneth A, Joas H 1988 *Social Action and Human Nature*. Cambridge University Press, Cambridge

James W 1981 *The Principles of Psychology*. Harvard University Press, Cambridge, MA

Joas H 1997 *Die Entstehung der Werte*. Suhrkamp, Frankfurt am Main (Chap. 9 rev. as The Autonomy of the Self: The Meadian Heritage and Its Postmodern Challenge.) *European J. Social Theory* **1**(1998): 7–18

Kymlicka W 1989 *Liberalism, Community, and Culture*. Clarendon Press, Oxford

Laitin D D 1998 *Identity in Formation: The Russian-Speaking Populations in the Near Abroad*. Cornell University Press, Ithaca, NY

Lebovics H 1992 *True France: The Wars over Cultural Identity, 1900–1945*. Cornell University Press, Ithaca, NY

Mead G H 1934 *Mind, Self, & Society from the Standpoint of a Social Behaviorist*. University of Chicago Press, Chicago

Taylor C 1989 *Sources of the Self: The Making of the Modern Identity*. Harvard University Press, Cambridge

R. Hardin

Ideology: History of the Concept

The concept of ideology is a highly contested notion and there is no general consensus today about how best to define the term. In the work of some authors, 'ideology' is used in a general way to describe sets of values and beliefs which guide political action; one could describe the values and beliefs which underpin political parties as 'ideologies' in this sense. In the work of other authors, however, 'ideology' is used in a more critical way to refer to ideas or symbolic forms which in some way distort or misrepresent the social world, or which serve the interests of some individuals or groups at the expense of others. Used in this second sense, the concept of ideology is often linked to a type of social or political analysis which adopts a critical stance towards the social world. So to study ideology, in this second sense, is also to look critically at ideology; it is not simply to describe values or beliefs but also to unmask them, to show how they misrepresent the world or how they serve dominant interests.

1. Brief History of the Concept

The term 'ideology' was first used in the late eighteenth century by the French philosopher Destutt de Tracy (1754–1836). An educated nobleman who was strongly

influenced by the European Enlightenment, Destutt de Tracy wanted to develop a new discipline which would be concerned with the systematic analysis of ideas and sensations. He described this proposed discipline as 'ideology'—literally, the science of ideas. Since all scientific knowledge involved the combination of ideas, ideology would be the 'first science'; it would also provide a basis for education, morality, and politics. Ideology would enable us to understand human nature, and hence would enable us to regulate social and political life in accordance with the needs and aspirations of human beings. Destutt de Tracy's vision for this infant discipline was thus a natural development of certain themes of the Enlightenment, such as the emphasis on progress and on the capacity of human beings to understand and control the world through systematic, scientific analysis.

In the early nineteenth century, the meaning of 'ideology' was transformed in various ways. This transformation began when the French emperor, Napoleon Bonaparte, turned against Destutt de Tracy and ridiculed the pretensions of his work. Napoleon described Destutt de Tracy and his associates as 'idéologues,' and he condemned ideology as an abstract, metaphysical doctrine which was divorced from the realities of political power. As his military campaigns ran into difficulties and his position weakened at home and abroad, Napoleon's attacks on the idéologues became more sweeping and vehement; increasingly he viewed ideology as the source of France's difficulties, since it sought to base laws on the abstract analysis of ideas rather than on the lessons of history. The meaning of the term 'ideology' began to shift: it was no longer used primarily to refer to the science of ideas but was used increasingly to refer to the ideas themselves, and especially to ideas which were alleged to be erroneous, misleading, and divorced from the practical realities of social and political life.

The term was taken up and further transformed by Marx and Engels in the 1840s and 1850s. Marx and Engels were familiar with Destutt de Tracy's work and with Napoleon's attack on it, and in some of their early writings they used 'ideology' in a way that was informed by this controversy. In *The German Ideology*, written in 1845–6 but left unpublished, Marx and Engels criticized the views of the so-called 'Young Hegelians' such as Ludwig Feuerbach, Bruno Bauer, and Max Stirner (Marx and Engels 1970). In characterizing the views of these thinkers as 'the German ideology,' Marx and Engels were following, very broadly, Napoleon's use of the term: the work of the Young Hegelians was the equivalent, in the relatively backward social conditions of nineteenth-century Germany, of the doctrines of Destutt de Tracy and his associates. Just as Napoleon had poured scorn on these doctrines, giving the term 'ideology' a negative connotation, so too Marx and Engels derided the views of their compatriots. The views of the Young Hegelians were 'ideological' in the sense that they were abstract and metaphysical; these thinkers were preoccupied with ideas and the analysis of ideas, and they failed to grasp the real material conditions of social and political life.

In later writings, however, Marx laid the foundations for a somewhat different and more innovative account of ideology. He suggested that, in societies divided into classes, ideas may play an important role in articulating the conflicting aims and interests of different social classes. In his 1859 Preface to *A Contribution to the Critique of Political Economy*, Marx distinguished between the material conditions of production, on the one hand, and what he called 'ideological forms,' on the other (Marx 1968). While the material conditions of production were the real basis of social and political life and the driving force of social change, he suggested that it was at the level of ideological forms of consciousness—including legal, political, religious, aesthetic, and philosophical forms—that individuals become aware of the conflicts stemming from the material base and fight them out. Moreover, 'just as our opinion of an individual is not based on what he thinks of himself,' so too we cannot take these ideological forms of consciousness at their face value: we must show that they are rooted in, and shaped by, the material conditions of social life. By explaining ideological forms of consciousness in this way, we can also *unmask* them—that is, we can show that they are expressions of the interests of particular groups and classes. Hence an analysis based on the scientific study of the material conditions of social life can at the same time be developed, Marx suggested, as a critique of ideological forms of consciousness which it both explains and unmasks.

These and other contributions by Marx and Engels provided a rich source of inspiration for subsequent authors, and the concept of ideology became a central topic of debate in Marxist circles in the twentieth century (for a full discussion of ideology in relation to Marxism, see Larrain 1983). But the concept was also taken up by sociologists, political scientists and others who wanted to use 'ideology' in a way that would be more neutral and less concerned with questions of normative judgment and critique. This shift of emphasis was evident in the work of Karl Mannheim, whose *Ideology and Utopia*, first published in 1929, became a key text in the complex history of the concept (Mannheim 1936).

In Mannheim's view, most earlier conceptions of ideology were what he called 'particular' conceptions, by which he meant conceptions which remained at the level of more-or-less conscious disguises, deceptions and lies. When we characterize the ideas and views of our opponents as 'ideological' and regard them as misrepresentations or distortions of the real nature of the situation, we are presupposing the particular conception of ideology in this sense. Mannheim contrasted this with what he called the 'total' conception of ideology. We presuppose the total con-

ception when we try to grasp the modes of thought and experience, the *Weltanschauung* or 'world-view,' of an epoch or group and view it as an outgrowth of a collective life-situation. The particular conception remains at the level of individuals engaged in accusation and deception, whereas the total conception has to do with collective thought systems which are related to social contexts.

In Mannheim's view, Marx was the first thinker to make the transition from the particular to the total conception of ideology, but he did so only in a partial and limited way. Marx moved beyond the particular conception in so far as he sought to trace ideas back to the class positions of their exponents. But Marx's approach retained elements of the particular conception in so far as he sought to discredit bourgeois thought and to criticize the ideas of his adversaries, while refraining from applying the same approach to his own thought. In other words, Marx adopted a rather partial version of the total conception; what is necessary, argued Mannheim, is to make the full transition to the total conception. When one fully adopts the total conception, the analysis of ideology ceases to be the intellectual weapon of a party and becomes instead a method of research in social and intellectual history, a method Mannheim described as 'the sociology of knowledge.' For Mannheim, therefore, 'ideology' in the sense of the total conception could be regarded as the interwoven systems of thought and modes of experience which are conditioned by social circumstances and shared by groups of individuals, and which can be analyzed by means of a discipline—the sociology of knowledge—that moved beyond the partisan character of earlier approaches.

Marx and Mannheim are only two of the many social thinkers who, in the nineteenth and twentieth centuries, used the concept of ideology and tailored it to their own specific concerns (for accounts of the contributions of other thinkers, see Lichtheim 1967, Barth 1976, Larrain 1979, Thompson 1984, Eagleton 1991). The literature abounds with different conceptions of ideology; it also abounds with differing assessments of the significance of ideology (or ideologies) and the role it plays in modern societies. One assessment of the significance of ideology which was influential in the 1950s and 1960s was the so-called 'end of ideology' thesis.

2. *The End of Ideology Thesis*

The end of ideology thesis was originally put forward by a number of liberal and conservative thinkers in Europe and the United States in the 1950s, including Raymond Aron, Edward Shils, and Daniel Bell (see Aron 1957, Shils 1958, Bell 1960, Waxman 1968). In its original formulation, the end of ideology thesis was an argument about the alleged decline of radical and revolutionary political doctrines in developed industrial societies. It was argued that in the wake of World War II, the defeat of Nazism and Fascism, the Moscow trials, and the denunciation of Stalinism among other things, the old ideologies stemming from the eighteenth and nineteenth centuries had lost much of their persuasive power. These ideologies had taken hold primarily among groups of intellectuals who had become disaffected with existing social and political institutions, and who had expressed their disaffection by calling for radical change. But the political events of the early twentieth century had exposed the naivety and danger of such calls. It was becoming increasingly clear that the problems confronting developed industrial societies could not be solved by the kind of radical social change espoused by Marxism and communism, since this kind of change gave rise to its own problems and to new forms of violence and repression.

The end of ideology theorists argued that the old 'ideological politics' were a spent force, and were giving way to a new sense of pragmatism in developed industrial societies. Revolutionary passion was waning and was being replaced—in the West at least—by a pragmatic, piecemeal approach to social change within the framework of a mixed economy and a redistributive welfare state. The end of ideology theorists generally recognized that ideologies would continue to flourish in less developed societies, and they did not altogether rule out the possibility that revolutionary passions might occasionally reappear as isolated outbursts in the developed industrial societies. But they maintained that, as a general situation in which the political arena is animated by radical and revolutionary doctrines, which arouse passion and heated conflict, the age of ideologies is over and ideology has ceased to be a significant feature of modern industrial societies.

Of course, the end of ideology theorists were using 'ideology' in a very special sense. Ideologies, in their view, were a particular kind of secular belief system: they were comprehensive, totalizing doctrines which offer a coherent vision of the social-historical world and which demand a high degree of emotional attachment. For most of these theorists, Marxism was the epitome of ideology in this sense. Marxism offered a systematic, totalizing vision of the social-historical world; it predicted a future which would be radically different from the present, and which could only be realized through the dedicated action of individuals who believed unflinchingly in their cause. These were the characteristics of ideology: totalizing, utopian, impassioned, dogmatic. The end of ideology in this sense was not the end of political debate and conflict, but these debates and conflicts would no longer be animated by totalizing, utopian visions which incited individuals to revolutionary action. Political processes would be increasingly institutionalized within a pluralistic framework in which political parties or

groups competed for power and implemented pragmatic policies of social reform.

The end of ideology thesis was put forward at a time—the late 1950s—when economic growth in many Western countries had helped to create conditions of relative stability, and when the idea of combining a mixed economy with a welfare state had been accepted by many intellectuals and politicians as a viable institutional framework for developed industrial societies. The political turmoil of the late 1960s and the economic recessions of the 1970s and 1980s undermined these conditions to some extent, and it could be argued that, contrary to the expectations of the end of ideology theorists, ideology has remained an important aspect of politics and political struggle in Western industrial societies. One could plausibly argue, for example, that Margaret Thatcher's reorientation of the Conservative Party in Britain in the 1980s, and the rise of neoliberalism in many Western countries in the 1980s and early 1990s, were instances when ideology had a significant impact in the political realm. Part of the difficulty with the end of ideology thesis is that it used the term 'ideology' in a very special sense (with Marxism as the prime example), and this undoubtedly limited the usefulness of the thesis as a way of understanding the changing character of politics.

With the benefit of hindsight, the end of ideology thesis in its original form may well appear to be a product of its time. Nevertheless, there may be some substance in the view that ideological differences between the major political parties in many Western countries have declined to some extent since the 1960s, partly as a result of broader changes in social structure. During the second half of the twentieth century, the changing character of work in Western industrial societies has transformed the social context of politics. Parties could no longer rely on the old social classes which once provided the core of their support. Traditional doctrinal divisions were attenuated, and increasingly parties and their leaders had to struggle to win the support of a growing pool of uncommitted voters. The traditional class-based party politics, with its strongly opposed belief systems and its strong contrast between left and right, has not disappeared, but it has been weakened by these social transformations. Ideology has not been eliminated from party politics but its role has undoubtedly changed, as parties and their leaders know that they can no longer count on a stable and committed electoral base and must try to win support by means of specific policy pledges and carefully devised electoral campaigns.

3. *The Concept of Ideology Today*

Given its long and complex history, there are some social scientists today who believe that it would be best to dispense altogether with the concept of ideology. They argue that the concept is simply too ambiguous, too controversial and contested, to be retained as a useful tool for social and political analysis. But this view is not universally shared. There are others who believe that the concept of ideology still has a useful role to play, although differences remain in terms of how best to define the concept. While the term 'ideology' is still used in a variety of different ways, one can distinguish between two broad conceptions of ideology which are prevalent in the literature today.

Many social and political analysts today use the term 'ideology' in a largely descriptive way to refer to any system of thought or belief which animates social or political action: this is what could be described as a *neutral conception* of ideology. Thus ideologies can be thought of as discrete and relatively coherent systems of thought or belief which inform social and political action, and the task confronting the analyst of ideology is to delineate these systems and describe their main features. (This is the view adopted, for example, by Martin Seliger; see Seliger 1976.) This conception is exemplified by the tendency to think of ideologies in terms of 'isms'—Marxism, communism, socialism, Fascism, Nazism, liberalism, conservatism, etc. Most modern political parties, as well as social and political movements like the women's movement and the environmental movement, have ideologies in this sense. The capacity of such organizations or movements to mobilize support depends on their ability to produce and diffuse belief systems which have some degree of appeal among activists and followers.

Other social and political analysts use the term 'ideology' in a more normative way to refer to ideas, beliefs or symbolic forms which are in some way misleading, illusory, or one-sided, or which serve the interests of powerful groups. We can describe these uses as variants of a critical conception of ideology. The analysts who adhere to this conception argue that the very characterization of something as 'ideological' unavoidably conveys a negative connotation: it implies that the ideas or symbolic forms are partial or are harnessed to particular interests. There are differing views in the literature about how exactly this critical conception of ideology should be understood. Some theorists regard ideologies as sets of ideas about the social world which are in some sense 'false' or illusory, whereas other theorists prefer to think in terms of the inter-relations between symbolism and power. An example of the latter approach is the view that ideology should be conceptualized in terms of the ways in which symbolic forms serve, in specific circumstances, to establish and sustain relations of domination (see Thompson 1990). If one adopts a critical conception, then the analysis of ideology is not simply a matter of describing values and beliefs: it also has a normative dimension. In other words, the analysis of ideology is linked to questions of social and political criticism. Another implication is that the study of ideology has to be extended beyond the sphere of organized

politics. While the institutions of political power are clearly an important site of ideology, it is also important to consider the ways in which meaning is mobilized by other institutions, such as the media and the educational system.

While there may be no consensus about how best to use the term 'ideology,' many commentators would agree that the study of ideology is an indispensable part of social and political analysis. Political systems, social and political movements, and relations of power and domination are always interwoven with ideas, beliefs, and symbolic forms of various kinds; power is rarely exercised without some kind of symbolic basis or support. It is this aspect of power, and the ways in which relations of power are interwoven with ideas, beliefs, and symbolic forms, which define the domain of study for those who are concerned with problems of ideology.

See also: Critical Theory: Contemporary; Critical Theory: Frankfurt School; Culture, Sociology of; Ideology: Political Aspects; Ideology, Sociology of; Knowledge, Sociology of; Marx, Karl (1818–89); Marxist Social Thought, History of; Utopias: Social; Work and Labor: History of the Concept; Work, Sociology of

Bibliography

Aron R 1957 *The Opium of the Intellectuals*. [trans. Terence Kilmartin]. Secker and Warburg, London

Barth H 1976 *Truth and Ideology* [trans. Frederick Lilge]. University of California Press, Berkeley, CA

Bell D 1960 *The End of Ideology*: *On the Exhaustion of Political Ideas in the Fifties*. Free Press, Glencoe, IL

Eagleton T 1991 *Ideology: An Introduction*. Verso, London

Larrain J 1979 *The Concept of Ideology*. Hutchinson, London

Larrain J 1983 *Marxism and Ideology*. Macmillan, London

Lichtheim G 1967 The concept of ideology. In: Lichtheim G *The Concept of Ideology and Other Essays*. Random House, New York, pp. 3–46

Mannheim K 1936 *Ideology and Utopia: An Introduction to the Sociology of Knowledge* [trans. Louis Wirth and Edward Shils]. Routledge and Kegan Paul, London

Marx K 1968 Preface to *A Contribution to the Critique of Political Economy*. In: Marx K, Engels F (eds.) *Selected Works in One Volume*. Lawrence & Wishart, London, pp. 180–4

Marx K, Engels F 1970 *The German Ideology*, pt 1. Arthur C J (ed.). Lawrence and Wishart, London

Seliger M 1976 *Ideology and Politics*. George Allen and Unwin, London

Shils E 1958 Ideology and civility: On the politics of the intellectual. *The Sewanee Review*. **66**: 450–80

Thompson J B 1984 *Studies in the Theory of Ideology*. Polity Press, Cambridge, UK

Thompson J B 1990 *Ideology and Modern Culture: Critical Social Theory in the Era of Mass Communication*. Polity Press, Cambridge, UK

Waxman C I (ed.) 1968 *The End of Ideology Debate*. Funk & Wagnalls, New York

J. B. Thompson

Ideology: Political Aspects

A political ideology is a set of ideas, beliefs, values, and opinions, exhibiting a recurring pattern, that competes deliberately as well as unintentionally over providing plans of action for public policy making, in an attempt to justify, explain, contest, or change the social and political arrangements and processes of a political community.

The concept of ideology is subject to a range of different, partly incompatible, interpretations. In the Marxist tradition it relates to an illusory distortion of material reality, supported by exploitative power relations, and it is employed critically to unmask those dissimulations and to eliminate ideological thinking as false consciousness. In non-Marxist approaches it is split among three trends. The first is a largely pejorative reference to abstract and closed idea systems, of a totalitarian or doctrinaire nature, which are contrasted with looser belief systems or with pragmatism. The second is an ostensibly value-free perspective in which ideologies are empirically ascertainable attitudes towards political issues to be explored using behavioral methods. The third recognizes the inevitability of ideologies as ubiquitous mapping devices of cultural symbols and political concepts that constitute a crucial resource for understanding sociopolitical life and enable collective choices to be made concerning the shaping of that life. Thus, what constitutes a political ideology, and describes its main features, is partly influenced by the conception utilized.

1. Genesis of Modern Ideologies and their Study

Although ideologies are as old as human thought and organization themselves, they emerged in a differentiated political form in the nineteenth century. This reflected the Enlightenment-inspired exploration of new thinking, the gradual spread of education and broader political articulateness, and the rise of mass politics, within which context ideologies served as important mobilizing tools for wider bands of active citizenry. Initially the term 'ideology' was coined by Antoine Destutt de Tracy to refer to the scientific study of ideas. Soon, however, the term was diverted to address some of the mounting epistemological and political concerns of the age. In particular, the search for social truths became a central bequest of the positivist tradition, and the discovery of power relationships among groups engendered innovative social theorizing. Ideology was now employed by Karl

Marx and Friedrich Engels to signify two main features. The one was the absence of true consciousness, replaced by imagined and metaphysical phantoms of the human mind. The ending of alienation through scientific socialism would rectify this defect and eradicate ideology. The other was to characterize ideology as an instrument of oppression by the ruling class in order to establish its dominance over the dehumanized majority. Ideology was located firmly within a set of socioeconomic capitalist practices and interests. It reflected in an inverted form the social contradictions of material conditions, so that repressive practices—such as the 'free' market—appeared as liberating ones.

Within the Marxist tradition Antonio Gramsci (1971) strengthened the notion of ideology-cum-power through his theory of the hegemony of a consensual historical bloc, but he also detached ideology from its ephemeral and dogmatic nature. He acknowledged the role of mass political activity in shaping popular culture through secular faiths, and recognized ideology in its highest sense as a conception of the world consolidated by intellectuals, directing social and political practices, and present in all manifestations of collective life, including art, religion, literature and law. Louis Athusser (1984) continued this reconciliation with ideology by emphasizing its integrative, 'interpellating' function. Ideology was both an imaginary representation of the real and a permanent lived relation between individuals and their conditions of existence. He too, though, still saw the role of central political institutions as producing 'ideological state apparatuses' that ensured their control over the established order.

From the late nineteenth century onwards, the European political arena became increasingly dominated by a struggle among competing political ideologies, in particular varieties of conservatism, liberalism, and socialism, all undergoing considerable growth. While Marxists could dismiss most of these as bourgeois illusions, the power of ideas as a means to organize political action, through political parties and their programs, as well as in extraparliamentary forms, became evident. The family of socialisms was subject to multifold internal divisions which underscored the subtlety of ideological argument. Attack and counterattack assisted in raising the profiles of the major ideologies and in accepting them as instruments of political conflict as well as transformation. To begin with, studies of ideology were mainly of historical and concrete ideologies, expository and frequently exhortatory. The threat of the totalitarian ideologies of communism and fascism changed that outlook, and a different pejorative connotation from the Marxist one imbued both popular and scholarly understandings. Ideology as a phenomenon was now characterized not only by its dogmatism, but by the superimposition of alien thought systems on a subjugated population, as part of a struggle for world domination. By contrast, liberalism and conservatism presented themselves as respectively open and flexible, or concrete and empirically based, sets of political ideas. Their claims to universalism were packaged in turn as rationally persuasive or invisibly natural, rather than manipulative.

Karl Mannheim (1936) assisted in changing the conceptual interpretation of ideology and alerted observers of politics to its ubiquity. Mannheim accepted the Marxist focus on the historical and social genesis of thought, but expanded ideology to cover the multiple social understandings of diverse social groups, each undergoing a common experience. Ideology was thus subject to a new epistemology that regarded all historical knowledge as relative to the thinker's sociohistorical location. Mannheim applied the term ideology solely to the interests of ruling groups intent on conserving existing social arrangements. It was divided between a particular conception of ideology as a lie or error and hence a function of individual psychology; and a total conception that referred to the entire *Weltanschauung* of a group. Ideology was contrasted with utopia, which on Mannheim's understanding was a transformative set of ideas, though more akin to what would now be termed a progressive ideology. Mannheim proffered ideology as a normal feature of sociopolitical thinking, and Marxism was itself unmasked as ideological. However, his attempt to transcend relativism through a sociology of knowledge was less convincing. He accorded free-floating intellectuals the capacity to incorporate the perspectives of different groups into a holistic relationism and to attain a critical approximation of truth. This tortuous quest only served to underline the problem of attaining objectivity in the social sciences.

2. *Analytical Perspectives*

2.1 *Empirical Approaches*

After World War II, US political scientists injected an empiricist dimension into identifying ideology. Employing behaviorist procedures to explore political attitudes, opinions, and evaluations, they concentrated on mass publics in democratic regimes. This method has applied interviewing and statistical techniques to aggregating and analyzing research findings. The regularities and constraints in the belief systems of social groups are related to the political thinking of elites and to public policy. These broadly functionalist approaches are innovative in their social-psychological focus on integration, and in recognizing ideology as unreflective and semiarticulate as well as reasoned and coherent. But they have also assimilated the study of ideology into a methodological individualism and a predilection for cognitive expression. Political ideology was now far removed from Marxist under-

standings: the question of its truth was bypassed in favor of regarding it as a scientifically neutral input into political systems. However, a reaction to that neutrality has been proffered by those who see ideology as possessed of a monolithic logic, closed to information that might challenge its coherence. Such interpretations have supported the contentions of many ethicists who consider ideology to be a betrayal of the moral expectations attached to political philosophy, rather than a different, but equivalently significant, form of political thought. Ideology may thus be relegated again to the category of bad theory.

2.2 Interpretation and Mapping

Anthropological and linguistic insights radically transformed interpretations of the functions and forms of ideology. Clifford Geertz (1964) portrayed ideology not as a mask, a weapon, or an expression of material interests, but as an ordered system of interdependent cultural symbols that imparted meaning to political systems. Ideologies provided patterned and holistic maps that both reflected and shaped a problematic social and political reality, and without which individuals would be unavoidably disorientated. This shift to the semantics of ideology was also evident in linguistics, with its emphasis on the interconnectedness of words as the basis of grammar and meaning. The inevitability of ideology had become ever more salient. If concepts were signified by words, then polysemic political concepts—the building blocks of ideologies—could form disparate semantic fields. Moreover, symbols and practices also combined to form systems of meaning. These theories allowed for an interpretation of ideology not only as sets of ideas having an impact on political action, but as pervading patterned social practices, often bereft of social transparency, yet decodable as unconscious incarnations of ideological stances.

2.3 The Unconscious, Emotion, and Reason

The introduction of the unconscious as a valid source of ideological data has itself been a site of scholarly dissent. The nonreflective manifestations of ideology may also be conflated with irrationality. Some political scientists asserted that ideologies were not only total but infused with passion: as Daniel Bell (1962) put it, secular religions. This led to the conclusion that, as extreme ideologies were apparently burning themselves out, the end of ideologies was in sight. Such views underestimated the emotive power of political ideas. They also incorporated a definitional constraint that ignored additional features of ideology that render the political 'center' just as ideological as the 'right' and the 'left,' or that cannot account for the rise of social movements such as feminism and environmentalism—political and ideational expressions of new social concerns. An alternative perspective identifies all political thought as proffering a mixture of rational and arational argument—both being a legitimate part of the ideological arsenal. All ideologies require some, albeit variable, degree of coherence, but they also incorporate emotion in the shape of fervor and rhetoric, vital to the recruiting of political support, and in the shape of commitment to non-negotiable core values. That, it is claimed, distinguishes them from 'best practice' political philosophy, where logic, consistency, intentionality and moral persuasiveness constitute the criteria by which political ideas are assessed. More recent hermeneutical arguments recast the role of the unconscious, as in Paul Ricoeur's notion of 'surplus of meaning' beyond the intentions of the agent. Here unlimited readings of the same texts and ideas account for differentiated ideological interpretations. Ideologies are diversely consumed, as well as produced, by groups. This offers one reason for the indeterminacy of ideological meaning and for the richness of idea resources they supply. The range of social understandings can never be captured by any political utterance and the study of ideology needs to incorporate this inescapable multiplicity. Among others this has led to a break with a sole focus on the great 'isms' and to developing more subtle microanalyses of ideological variability.

2.4 Poststructuralism and Microstructuralism

At the beginning of the twenty-first century, developments in the analysis of ideology display two tendencies. Post-Marxist and poststructuralist theory regard ideology as a modernist construct through which narratives necessary to maintaining the social order are produced, or that paper over the void that renders the social order itself a fiction. Ideology therefore assumes the features of permanence, while remaining nebulous and 'unreal.' These standpoints have also expanded the source material for political ideology to include cultural goods, literary and visual. The second tendency builds on semantics and hermeneutics in order to present ideologies as differing configurations of the major political concepts they contain. For instance, the concept of liberty may be present in most ideologies, but its meaning is determined from case to case by the proximate concepts that surround it, private property or social welfare pulling it in different directions. Ideologies are hence forcefully assertive, as well as unwitting, competition over the 'correct' and legitimate meanings of political words and ideas, and by means of that control, over the high ground of politics. The essential contestability of concepts requires that the 'decontestations' that ideologies offer are logically arbitrary but culturally significant, and reflect fundamental, if diverse, social understandings. Liberalism or conservatism have so far displayed stable, though moderately varying, conceptual patterns; totalitarian ideologies exhibit greater inflexibility in their

internal configurations; and other ideological clusters, such as the New Right, may exhibit fleeting or hybrid structures, or be less comprehensively systemic, as in nationalism. Political ideologies possess loose family resemblances and vary by dint of the different internal conceptual morphology each family exhibits. Continuities of time, space, political reasoning, and psychological dispositions hold together what could otherwise be, from a postmodernist viewpoint, a fragmentary and elusive form. In parallel, discourse analysis has located political ideology in ordinary language and communication—text and conversation—illuminating forms of human interaction. Such approaches have partly replaced the classification of political ideologies on a left–right continuum, which is too simplistically based on dichotomies such as individualism versus collectivism. They regard ideology not as coterminous with political thought, but as one of its dimensions.

See also: Ideology: History of the Concept; Ideology, Sociology of; Marx, Karl (1818–89); Marxism and Law; Marxism in Contemporary Sociology; Marxism/Leninism

Bibliography

Athusser L 1984 *Essays on Ideology*. Verso, London

Bell D 1962 *The End of Ideology*. Collier Books, New York

Freeden M 1996 *Ideologies and Political Theory: A Conceptual Approach*. Clarendon Press, Oxford

Geertz C 1964 Ideology as a cultural system. In: Apter D E (ed.) *Ideology and Discontent*. The Free Press, London, pp. 47–76

Gramsci A 1971 *Selections from the Prison Notebooks of Antonio Gramsci*. In: Hoare Q, Nowell-Smith (eds.) Lawrence and Wishart, London

Hamilton M B 1987 The elements of the concept of ideology. *Political Studies* **35**: 18–38

Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy*. Verso, London

Mannheim K 1936 *Ideology and Utopia*. Paul, Trench, Trubner London

Marx K, Engels F 1974 *The German Ideology*. In: Arthur C J (ed.) Lawrence and Wishart, London

Van Dijk T A 1998 *Ideology: A Multidisciplinary Approach*. Sage, London

M. Freeden

Ideology, Sociology of

The concept of ideology constitutes a significant chapter of social sciences in general and sociology in particular. The exact localization of this chapter remains, however, uncertain. Linked from the point of view of its emergence and its development to the Marxist tradition, the analysis of ideological phenomena overflows very quickly and very largely this tradition. Producing, under certain conditions, positive knowledge about the social order, the analysis of ideologies intends sometimes to fulfill a 'performative' function, i.e., contribute to the transformation of this social order. Finally, constitutive of various considerations concerning the social determinants of beliefs, the sociology of the ideologies is quickly led to define its own condition and, through this reflexive movement, to appear on a level quite as empirical as epistemological.

This pluridimensionality of the sociological analysis generates the feeling of confusion characteristic of many debates relating to ideology, its causes and consequences. This feeling is reinforced since more than 15 possible definitions of the very concept of ideology correspond to the diversity of the manifestations of the ideological phenomena through contemporary sociological writings (Eagleton 1991). After having briefly made sensitive the nature of this diversity, we will approach the way in which sociologists have tried to characterize and to explain the ideological phenomena. Three interdependent aspects will be successively considered: the referent, the structure, and the value of truth of ideologies. We will be interested finally in revival of the sociological analysis on the relationship between science and ideology.

1. The Diversity of Ideological Phenomenon

The origin of the sociological use of the concept of ideology can be found in Marx's writings. Admittedly, as reminded by many commentators, the concept of ideology existed before Marx. One thus finds a particular form of this concept at the end of the eighteenth century with the French tradition of the *idéologues* who, led by Destutt de Tracy, saw in the *Ideologie* a new discipline: the science of the ideas. Further still the theory of the *idola* developed by the English philosopher Francis Bacon in his *Novum Organum* anticipates certain aspects of both science of the ideas of the French *idéologues* and the modern theory of the ideologies. But it is with Marx that a certain use of this concept appears. Its well-known analysis of the *coup d'etat* of Louis Bonaparte is from this point of view very instructive (Marx 1852). The important part of its argument consists of showing how, once the riots of June 1848 were over, dissensions within the *aparti de l'ordre*—the right wing of the French parliament—helped make this *coup d'etat* unavoidable. This party was composed of many fractions which all, Marx affirms, have a 'superstructure of impressions, illusions, ways of thinking, and some philosophical conceptions.' To understand the origin of these 'ways of thinking' it is essential to identify the 'competition of interest,' the will of each fraction 'to restore its own supremacy.' The land aristocracy and the industrial bourgeoisie

develop different political representations thus: the former theorizes the benefits of the absolute monarchy, the latter those of parliamentary monarchy. In both cases, Marx suggests, they tend to give a general value to 'ways of thinking' which actually constitute only the formulation of objective conditions which ensure the domination of a social group over another. The analysis of Marx does not end with this comparison between particular 'interests' and 'ways of thinking,' it also underlines the way in which certain individuals embody ideologies. Describing the 'ideologists' consists mainly for Marx of studying the 'writers,' the 'press,' and more largely 'representatives' of the *bourgeoisie*. In the present case, the inability of these ideologists to express the real opinion of those they should represent: little before the *coup d'etat*, Marx points out, 'the representatives and represented had become strangers to each other.'

The argument is simple and easily acceptable: when an institution—here a political mode—appears in conformity with the particular interest of a social actor, individual, or collective, this actor generates, directly or indirectly, a whole set of ideas, representations, and theories intended to make the value of this institution acceptable by the greatest number. Also, even if the value objectively makes real sense only for the social actor considered. When K. Mannheim looks back to this argument in *Ideology and Utopia* (1929/1991), he presents it as a fundamental stage towards the formation of the sociology of knowledge: the first attempt to stress 'the role of the position of class and the interests of class in the thought.' This attempt rests, however, on a confusion between two conceptions of the ideology which is necessary to identify. In the first—particular conception—the 'ways of thinking' are 'mystification more or less conscious of the real nature of a situation whose exact recognition would not be in agreement with the interests of the ideologist'; they are reducible to a psychology of the interests, concern primarily the individuals, and are connected with lies on the moral level or errors on the epistemological level. In the second—total conception—the same 'ways of thinking' are not 'illusions' deliberately maintained by the individuals according to their more or less immediate interests, but elements of a total 'system of significance,' 'knowledge grown out of our experience in the real situations of the life.' One of the rare examples of ideological 'knowledge' in this 'total' conception suggested by Mannheim is the moral judgement related to the loan of money with interest. Formulated and assimilated by the Church like an ethical standard, the morally reprehensible character of the loan with interest becomes ideological, affirms Mannheim, since the social actors try to maintain it out of the social framework to which it is 'adapted': a precapitalist society based, economically and socially, on personal relations of intimacy. Through its attachment to this judgement, the Church expresses its difficulty in thinking of social reality independently of a 'system of significances' adapted to a disappeared social framework from now on.

This total character of the ideology described by Mannheim acquires a new meaning with the compared analysis by Raymond Aron (1964) devoted to the contemporary political modes. In its *nouvelles leçons sur les sociétés industrielles*, Aron identifies in an ideal typical way each political mode by an 'official ideology,' i.e. a 'complex doctrinal system,' made up on the one hand of general principles resulting 'from theoretical analysis of what really occurs' and on the other hand of proposals qualified as 'misleading,' 'distorted,' or even 'absurd.' More than the simple critical census of these proposals, Aron underlines the specificity of all 'official ideology': the will of its promoters to withdraw it from any critical discussion. Why, Aron asks, do the partisans of the socialist ideology generally refuse to talk on a reasonable basis about the advantages and disadvantages of their conception of the individual and society? The answer is to be found in the 'universal' vocation of the studied ideology: it is given for its partisans as for its detractors as a global system of interpretation of the historico-political world.

2. *Sociological Characterization of Ideologies*

As suggested by these examples, the sociologists use the concept of ideology to describe a phenomenon of 'belief': belief in the value of the absolute monarchy for the land aristocracy (Marx), belief in the morally reprehensible character of the loan with interest for the Church (Mannheim), belief in the universal vocation of the collectivist mode for the Socialists (Aron). To speak about belief is to describe an ethical or cognitive reality—principle, idea, theory, doctrines, etc.—which not only makes 'sense' for an individual or collective actor, but a reality about which this actor may develop a deep feeling of conviction. It also describes a factor likely to influence the behavior of this actor. Innumerable sociological and anthropological work made it possible to identify the diversity of the beliefs and their consequences according to the social systems considered. *The difficulty which arises consequently is that of the specificity of the ideological phenomenon as belief compared to the various possible registers of the general phenomenon of the belief.* Sociologists generally approach this specificity in terms of degree and not of nature under three interdependent aspects: (a) referent of the belief, (b) forms of the belief, and (c) value of truth of the belief.

2.1 *Referent*

The ideology as belief is singular by the centrality which it accords to the nature and becoming of

politico-social organization. The referent is actually double. The ideologist aims by his discourse not only at describing the legitimate social order but also and especially at directing the consequently collective action. He proposes more than one simple explanation of observable social reality: a program to transform it. The direction of this transformation can be multiple. When Mannheim defines the ideology in opposition to utopia, by making the former an exclusive vector of conservatism (a barrier to the rise of capitalism in the case of the taboo of the loan with interest) where the latter would be progressist, he only preserves the initial views of Marx. This vision of the ideology is, however, too restrictive. Not only are the ideologies not the prerogative of the conservatives—T. Parsons (1959) distinguishes four types of ideologies: conservative, counter-, reform, and revolutionary—but one ideology can, according to social and historical circumstances, serve various social purposes. Liberalism, Socialism, or various contemporary forms of Nationalism (E. Gellner 1983, P. Birnbaum 1997) appear, at a given time of their history, as vectors of social emancipation as well as general principles of immobilism or conservatism.

This double referent—sociopolitical and practical—makes it possible to differentiate ideology from other types of beliefs, religious ones in particular. If the majority of the religions contain a representation of the social organization, they define only rarely a collective action plan. This representation is generally the consequence of a reality considered to be more central: a transcendent realm revealed by the faith. *The ideologist speaks in the name of reason*. Its truth claims positive values of science and is diffused by suggestion, persuasion, but much more rarely by revelation. Many sociologists (R. Bendix 1964, A. Gouldner 1976) thus establish a strong parallelism between the development of the ideologies at the end of the eighteenth century and the emergence of Modernity defined as a 'rational' social order in opposition to a 'traditional' social order. This by no means implies that the religious, mythical, or more largely traditional elements disappear from the ideological discourse, but that these elements occupy a peripheral place in this discourse.

2.2 Forms

When Aron analyzes the contemporary political modes, he considers ideology as an explicit system of beliefs developed around a core of descriptive and prescriptive proposals. The dominant features of the system are unity, coherence, and totalizing character—the capacity to potentially bring a response to any interrogation.

This traditional representation of ideology as a 'monolithic' block needs to be tempered: all ideologies contain irreducible tensions and contradictions. On the one hand, because any ideology contains irreducible tensions and contradictions. The goal of any ideologist is certainly to give an overall coherence to what he perceives as real and necessary, but one cannot judge the nature of a system of beliefs only on the base of how it wishes to appear. Sociologists insist thus on the 'compromises' and the 'adjustments' necessary to hold in the same unit theoretical logic and the principle of action (M. Seliger 1976). Other sociologists proposed to consider the production of ideology as a 'tinkering' process during which the ideologist tries to satisfy the demand of meaning of his public without ever being able to exert a true control on his resources (F. Bourricaud 1980).

Moreover, the contemporary ideologies no longer have the width of the all-encompassing ideological systems of the end of the nineteenth and the beginning of the twentieth centuries. If it still appears illusory to regularly announce the 'end of the ideologies' (D. Bell 1960, M. Lipset Seymour 1960), it is undeniable that the form under which they appear evolves. For R. Boudon (1999), modern ideologies evoke the image of archipelago, while the old ideologies evoked rather the idea of a continent. 'We have ideologies,' writes Boudon, 'as to what should be done about unemployment, educational opportunities, fighting against crime or drug addiction and on a myriad of subjects, as well as how it should be done. But these theories are weakly related to one another. We have ceased to believe that they could be derived from an all-encompassing theory. We have all kinds of local ideologies; we no longer believe in general ideologies.'

2.3 Value of Truth

The link established by the Marxist tradition between the concepts of ideology, alienation, and 'false conscience' led many sociologists, Marxist and non-Marxist, to identify the ideological phenomenon with false beliefs. The ideologist produces a representation of the social order he wishes to be socially considered as the unique depository of a true meaning, but both representation and meaning are mainly illusory. The ideology, according to the formula of T. Parsons, should be regarded as a 'deviation compared to scientific objectivity.' The relative disaffection of the category of ideology in the poststructuralist and postmodernist currents of thought is thus principally due to the fact that this category seems to imply a naive conception of truth as correspondence, but also a rigid demarcation between the science and the other discursive registers.

The reality of the ideological phenomenon is, however, more complex. From a theoretical point of view, some sociologists propose to define it inde-

pendently of any reference to science and objective knowledge. C. Geertz (1964) for example conceives the discourse of the ideologist as a 'symbolic act' intended to produce an effect of mobilization: the 'cognitive' function of ideology is consequently second behind its 'pratico-social' function. This approach is not however without difficulties: it gives of ideology a hypertrophied representation, and especially misses the fact that even when conceived as a symbolic act, ideology remains open to judgements of truth and falsehood. From a more empirical point of view, many studies show that ideologies grant a universal validity to proposals whose validity is only limited; or more simply that they mix doubtful or fragile proposals with more solid, quasiscientific proposals. General ideologies as Marxism, liberalism, or certain forms of conservatism or developmentalism are based explicitly on scientific argumentations. More local ideologies such as the widespread belief that mechanization or globalization are the main cause of unemployment can be seen as grounded on factual evidence, but evidence of limited validity. If ideology cannot be conceived as a delirious thought, it appears from now on quite as imprudent to accept it in block as to reject it in block.

3. *Sociological Explanation of Ideologies as Collective Beliefs*

Ideology being only one particular case of the general phenomenon of belief, its sociological explanation relies on the same principals as the explanation of other beliefs. When a sociologist explains an enigmatic phenomenon, he tries not only to infer the existence of relations between a whole set of variables (dependent an independent) but also to test, according to multiple modalities, the validity of his inference. There is not, however, a single way to conceive the factors to consider nor even the type of relations likely to be tested. Sociologists considered the production and the diffusion of the ideological phenomenon in three different ways: first as the consequence of a 'causal' process; second as a response to a 'functional' requirement, and finally as the product of a 'subjective rationality.'

The first general type of explanation considers the production and diffusion of ideological belief as the consequence of 'forces' on which the social actor has no control. These forces can be of various types. The Marxist tradition closely associates the production of ideology with the existence of class 'interests' which dominates the conscience. According to this tradition, the fact for an individual of belonging to a class is enough to mechanically impose a set of collective beliefs on his conscience. The general thesis is that if the dominated class accepts blindly social representations which contribute to reproduce its subordination, the dominant class generates social representations which maintain its domination. The example of the study of the 18 brumaire shows, however, that Marx develops, in certain empirical writings, a less simplistic vision of the production process of ideologies. Those ideologies are described as more than strategical theories imposed on social actors: representations of limited validity inspired by reality such as can be observed from a given social position. From a different point of view, the forces which drive the actors to produce or adhere to ideologies are also sometimes conceived as 'passions.' In its classical *Traité de sociologie générale*, Pareto considers thus that the ideologies—he speaks of 'derivations'—are primarily the effect of feelings remaining inaccessible to the conscience. More precisely, Pareto affirms that the social actor wants to be convinced of the value of his passion, and produces derivations for this purpose only. The ideologies would thus be the epiphenomenal expression of dominant passions. Such a representation, however, remains not easily generalizable insofar as on the one hand the ideologies are seldom lived by those which produce them or diffuse them like simple feelings, and on the other hand the rationalization process that characterizes ideologies is hardly reducible to a simple 'logical varnish.'

The second general type of sociological explanation considers the production of ideologies as a response to a double demand: demand for 'security at an individual level, demand for 'equilibrium at a more general level. The insecurity is regularly described by sociologists as a permanent aspect of the social systems. The individuals continuously feel the need to be able to think that 'everything will be all right.' Precisely, observes Parsons (1964), 'the strains involved in such expectations may be lessened by displacing the fulfillment outside the immediate field of action, either into the future as in the case of the leader of a movement for reform, or into a transcendental sphere (…).' At a collective level, the absence of integration is considered a permanent risk for the reproduction of the social system: principles, standards, goals differ according to the social sector considered (political, family, economical, etc.). Ideological belief is considered, from a functional point of view, as an ambivalent phenomenon: it produces disintegration because any ideological expression is generally directed against a pre-existent belief system, but it can also produce 'solidarity' insofar as the ideological orientation makes it possible to maintain the collective identity of a given group. The functional approach shows thus how a belief makes it possible to solve an individual and collective problem. Interpreted however in too extensive a manner, this approach generates a certain number of skids, among which are the identification of the society to an 'organism' and of ideology to a 'disease.' Another contemporary skid of the functional approach consists of seeing behind any ideology a social function, but moreover to define this function as

exclusively directed towards the maintenance of the domination of a social group on another.

The third general type of sociological explanation consists mainly of adopting a similar attitude as M. Weber towards the ideological beliefs for the magic beliefs or, to a certain extent, Durkheim for the religious beliefs: reconstruct the 'meaning' of the ideological belief for the social actor by seeking the system of reasons or preferences which this actor produces to justify to himself his adhesion. The rationality considered by the sociologist is not reducible to an instrumental one but is also quite different from the 'derivations' described by Pareto. The sociological analysis opens on a set of 'subjective reasons' that represents the necessary conditions of the ordinary thought. These reasons are of general but nonuniversal validity and manifest themselves on different levels: utilitarian, axiological, traditional, or epistemic. Inspired by studies of A. Downs (1957) and H. Simon (1982), the restricted theory of the ideology elaborated by R. Boudon (1989) thus stresses the necessary 'comprehensive' dimension of the sociological analysis. When some people adopt statements or ideas that appear to the sociologist as false or doubtful, the sociologist has to consider that this may derive from the fact that, given the situation, these beliefs are to them the conclusions of reason they perceive as 'strong.' This explicative strategy is directly linked to an active theory of ordinary knowledge: the social actor produces knowledge, not simply by contemplating reality, but by interpreting it from its social position, and more generally from questions which are only seldom directly provided by reality itself. This last approach has a certain number of objective advantages. In particular, it makes it possible to give an account of an important subjective data: the feeling of conviction and not of interiorization which generally characterizes the relation of any individual to his own ideological beliefs.

4. Science and Ideology: Extrascientific 'Interest' and Relativism

The analysis of the relationship between science and ideology is traditionally focused on the ideological dimension of social sciences and sociology in particular. The ideologist shares in fact with the sociologist the will to produce a discourse on society. It is not rare that certain sociologists consider their analysis as the methodical prolongation of a preexistent ideological point of view. The historical examples are numerous and well known: openly conservative, Frédéric Le Play theorizes the traditional form of family—*famille souche*—as 'the social unit per excellence'; convinced liberal, Herbert Spencer proposes an evolutionary model closely associated with the diffusion of 'social Darwinism,' etc. Inversely a sociological theory or paradigm conceived independently of any ideological commitment can, when it is exploited out of its immediate context of validity, generate an ideological skid. The mechanism is once again well known: it consists mainly of projecting on the theory or sociological paradigm the totalizing character of ideologies and by doing so asking those theories and paradigms to answer any possible interrogation.

The significant development since the beginning of the 1970s of the social studies of natural science made it possible to highlight the importance of economic, cultural, and political factors on the course of scientific research. These studies show on the one hand how the scientific institutions satisfy a social demand, and on the other hand how the scientists integrate in the course of their practices different cultural elements, in particular extrascientific 'interests' (Barnes 1974). The reality of these facts should not however be misinterpreted. It leads certain radical sociologists to affirm the ideological character of all sciences and to consequently develop a relativistic conception for which the superiority of the scientific approach of reality compared to other 'representations of the world' is nothing but an occidental myth.

This extreme relativism rests on confusion between the intrinsic quality of a knowledge (objectivity or absence of objectivity) and its referential dimension (the focus of science). Nobody rejects the fact that the scientific practice cannot develop without a certain number of resources. The unequal allocation of these resources exerts a considerable influence on the choice of the research topics, on the level of development of the disciplines, sometimes even indirectly on the degree of requirement of the experimental control. This social conditioning of science does not however invalidate the fact that a scientific knowledge has an intrinsic validity, independent of its immediate context of production. The sociocultural determination and the rational determination of scientific knowledge are by no means necessarily mutually exclusive.

See also: Alienation, Sociology of; Belief, Anthropology of; Collective Beliefs: Sociological Explanation; Collective Memory, Anthropology of; Culture, Sociology of; Ideology: History of the Concept; Ideology: Political Aspects; Knowledge, Sociology of; Pareto, Vilfredo (1848–1923); Science, Sociology of; Social Movements, Sociology of

Bibliography

Aron R 1964 *La lutte de classes: Nouvelles leçons sur les sociétés industrielles*. Gallimard, Paris

Barnes B 1974 *Scientific Knowledge and Sociological Theory*. Routledge & K Paul, London

Bell D 1960 *The End of Ideology*. Free Press, Glencoe, IL

Bendix R 1964 The age of ideology: Persistent and change. In: Apter D E (ed.) *Ideology and Discontent*. The Free Press of Glencoe, London

Birnbaum P 1997 *Sociologie des nationalismes*, 1st edn. Presses Universitaires de France, Paris
Boudon R 1989 *The Analysis of Ideology*. University of Chicago Press, Chicago
Boudon R 1999 Local vs general ideologies: A normal ingredient of modern political life. *Journal of Political Ideologies* **4**(2): 141–61
Bourricaud F 1980 *Le bricolage idéologique*. Presses Universitaires de France, Paris
Downs A 1957 *An Economic Theory of Democracy*. Harper, New York
Eagleton T 1991 *Ideology: An Introduction*. Verso, London
Geertz C 1964 Ideology as a cultural system. In: Apter D E (ed.) *Ideology and Discontent*. The Free Press of Glencoe, London
Gellner E 1983 *Nations and Nationalism*. Cornell University Press, Ithaca, NY
Gouldner A W 1976 *The Dialectic of Ideology and Technology*. Seabury Press, New York
Lipset Seymour M 1960 *Political Man: The Social Bases of Politics*. Doubleday, Garden City, NY
Mannheim K 1929/1991 *Ideology and Utopia*. Routledge & K Paul, London
Marx K 1852 The Eighteenth Brumaire of Louis Bonaparte. In Marx K and Engels F. *Collected works* Vol. 11. Lawrence & Wishart, London
Parsons T 1959 An Approach to the Sociology of Knowledge: Transactions of the Fourth Congress of Sociology, Milan and Sressa. In: Curtis J, Petras J (eds) 1970 *The Sociology of Knowledge: A Reader*. Gerald Duckworth, London, pp. 25–49
Parsons T 1964 *The Social System*. Free Press, Glencoe, IL
Seliger M 1976 *Ideology and Politics*. Allen & Unwin, London
Simon H A 1982 *Models of Bounded Rationality*. MIT Press, Cambridge, MA

M. Dubois

Illiteracy, Sociology of

Between complete *illiteracy* and good *literacy*, different levels of abilities may be distinguished, including *quasi-illiteracy*. Research from the last decades shows that the majority of adults in highly developed countries, as well as many men and women in the Third World, rank somewhere among these in-between levels. This situation seems to depend mainly on the fact that the way of life and work for large numbers of the population, in both less developed and, paradoxically, developed countries, does not require more than rudimentary or mediocre literacy skills.

1. *A Look at Illiteracy Statistics*

In most Western and Eastern European countries including Russia, as well as in Australia, Canada, Israel, New Zealand, the USA or, for instance, Argentina, Cuba, Mongolia, the Philippines, Puerto Rico, South Korea and Tajikistan, official statistics show that adult complete illiteracy is practically non-existent or very rare. Almost every other country lags somewhat or further behind these.

Table 1
Percentage of illiterate adults from 1900 to 1995 in six countries (UNESCO estimates and national statistics)

	1900	1950	1970	1995
USA	11	3	1	—
Italy	50	14	6	1.8
Mexico	78	43	25	11
Brazil	65	51	32	17
India	94	81	66	48
Mali	—	—	94	68

During the twentieth century, the absolute number of completely illiterate adults increased in the Third World due to population growth. In terms of percentages, however, adult illiteracy declined everywhere. Table 1 illustrates this point as regards six countries typical of different degrees of economic development.

The official statistics of illiteracy among adults in many countries are notoriously imprecise. They have another great deficiency, which is that they give no information about the real literacy capacities of the adults who are not counted as completely illiterate.

2. *Functional Illiteracy*

Third World countries are well aware of the extent of functional illiteracy, that is, of the gravely insufficient literacy level of many adults who are not completely illiterate. On the other hand, in developed countries, significantly high rates of functional illiteracy were almost unimaginable up until recent times, following decades of universal compulsory free education and, in particular, since the development of postcompulsory schools following World War II. This conviction has been steadily waning after converging results of various research showing that functional illiteracy was frequent in modern societies. The first international survey was conducted in 1994–5 under the auspices of the Organization of Economic Co-operation and Development (OECD 1995). The reports published by the OECD in 1995 and 1997 covered 12 countries, seven in Western Europe, plus Australia, Canada, New Zealand, Poland and the USA. The same, extensive battery of reading, writing, and arithmetic exercises typical of current tasks of daily life and work was used in evaluating the literacy level of adults aged 16 to 65. This international survey gave an interesting picture of the distribution of the adult population of typical developed countries by literacy levels at the end of the twentieth century.

Table 2
Percentages of adult population aged 16–65 (born in the country) by literary levels. Three examples: Germany, Sweden and the USA in 1994 (OECD 1995)

Literacy levels	Germany	Sweden	USA
Good	19	36	23
Mediocre	41	40	34
Low	31	19	27
Quasi-illiteracy	9	5	16
Total (percent)	100	100	100

The four categories of Table 2, which excludes immigrants, correspond to the following levels of literacy.

(a) Good: ability to at least understand newspapers and nontechnical articles written for the general public, write a simple letter, perform elementary arithmetic, read a road map, etc. These skills correspond roughly to the kind of reading, writing, and arithmetic abilities that compulsory schooling tries to instill.

(b) Mediocre: Between nearly good and nearly low levels.

(c) Low: at most can understand parts of short, very easy articles, complete a basic application for employment, interpret a pay stub, etc.

(d) Quasi-illiteracy: at most can understand or write isolated phrases, add or subtract small numbers; often can only understand or write isolated words and numbers. Exceptionally, complete illiteracy.

Sweden was the top-ranked country of the OECD survey. Yet even in Sweden, the adults found to have good levels of literacy did not make up the majority, but rather only one-third of the total adult population. In every other country surveyed by the OECD, the percentage of cases of good levels of literacy is lower. In Sweden, approximately one-quarter of the adults are clearly functionally illiterate (categories c and d). In the 11 other countries, the rate of functional illiteracy is even higher.

Landmarks of change are extremely scarce. A good short-term comparison of the literacy levels of younger Americans aged 21–5 is possible. Their average literacy abilities in 1992 were even more mediocre than in 1985 (Kirsch et al. 1993).

In Switzerland, from 1879 to 1913, the literacy skills of all young men aged 19–20 years old were evaluated each year by means of military reading, writing, and arithmetic exercises which were conducted at a level as constant as possible from year to year. Great progress was made. The exams ceased to be carried out after 1913. Some literacy evaluations of recruits were again conducted during the last quarter of the twentieth century, particularly in 1984. The results suggested a decline compared to pre-World War I statistics. This decline has since been confirmed by the results of young Swiss respondents aged 16–25 in the 1994–5 OECD survey.

3. *Social Contexts and Illiteracy*

The reasons for the extent of complete or functional adult illiteracy in less developed countries are well known. First, a fraction of the children do not attend primary school at all. The proportion of such cases is decreasing due to the increase in the number of schools and also, probably, because more families are becoming interested in the possibility of sending their children to school. Often, however, this interest is weak, as is the motivation of the children. Dropping out of school before the end of the first year of primary school is frequent. In sub-Saharan African countries, as well as in Latin America and Southeast Asia, approximately one-fifth of the children on average drop out of the first year of primary school and still many before the end of the fourth year. This high precocious dropout rate is generally attributed to the dissatisfaction of rural and urban poor families, who do not see the connection between the skills necessary for their milieu and what the school is teaching. These families tend to think that their children are wasting time at school. In many instances, they prefer to have the children around for helping with their own job or working outside of the family. One

> should also note the importance of the considerable increase in the number of persons who have not only attended primary school but have also taken more advanced courses … which enables employers to require higher levels of education than formerly, whether in the sectors of production, services or the civil service. For some families, whose incomes do not enable them to educate their children beyond the primary level, the later loses the importance it used to have as a certain means for social advancement and for improving their material situation. (Tanguiane 1990)

Many Third World adults live and work in social strata where illiteracy is the rule and almost all communication is oral. In such cases, the skills acquired at school or through adult education courses and alphabetization campaigns are never or rarely used. Generally, they deteriorate or disappear altogether. Thus, people in such situations tend to return to complete illiteracy or close to it. During the nineteenth century in Europe, the culture and economic conditions of large numbers of the traditional rural classes and the proletariat had similar effects (Weber 1976). How is functional illiteracy possible today, in the Netherlands, Australia and other highly developed countries? In Sweden, where there are slightly more than five million Swedish-born adults aged 16–65, the percentages of Table 2 correspond roughly to 268,000 quasi-illiterates and one million with low literacy levels. In Germany, according to the same table, there would be about five million native-born illiterate

adults and more than 10 million with low literacy levels. In the USA, the corresponding numbers would be 26 and 44 million, respectively. Yet the presence of so many quasi and semi-illiterates went unnoticed until the revelations of the surveys at the end of the twentieth century. The strong shock produced by the results of these surveys underscores the invisibility of functional illiteracy in most modern nations.

The quasi-illiterate and low-literate adults living and working in these societies are generally not even aware of having reading, writing, and computing problems themselves. Less than 10 percent to one-third of the quasi-illiterates (depending on the criteria and country), admit to having poor literacy levels. As for adults with low literacy, the percentage of respondents who consider themselves as having poor literacy skills is insignificant, ranging from less than 1 percent up to 5 percent.

The public's long-held unawareness and the previously mentioned self-evaluations may help to move forward in clarifying the problem of widespread quasi-illiteracy and low literacy in the adult populations of developed countries.

If the limited literacy abilities of a significant portion of the adult population of such countries would prevent them from coping normally at work, home, in stores, on the road, etc., then they would constantly make mistakes and have difficulties in the world around them. Yet this is not the case for several reasons.

(a) Except for a small percentage of the lower levels of the group, quasi-illiterate adults are not completely incapable of reading, writing, recognizing numbers, making out fragments of sentences, and performing very simple arithmetic problems. The literacy skills of low-literate adults are even a little less limited. In other words, quasi-illiterates and low-literates are capable of performing very simple or at least rudimentary literacy tasks.

(b) It is true that advanced societies depend more and more on the variety and quality of highly specialized knowledge. However, that does not mean that every member of society needs to possess sophisticated reading, writing, and arithmetic skills. As with housework, many jobs and leisure activities do not require literacy skills more advanced than those of quasi-illiterates or low-literates. Technical advances have led to an explosion of push-button and similar devices, from remote controls to bar codes. In advanced countries, about one-half of the jobs—many highly skilled in other respects—can be done with basic abilities: read road signs, enter numerical codes, read or write routine sentences or computer messages with the same few words, etc.

(c) Reading for pleasure is by no means commonplace for most adults, and writing even less so. A large percentage of adults never or almost never read books. Television and radio furnish an abundance of their information and entertainment. Paper and pencil computation has practically become a lost art form due to the availability of all sorts of electronic calculators.

(d) When faced with a situation where reading, writing, or computing would be useful, it is relatively easy for people to find a way out. The most common solution is to communicate orally with family members, colleagues, friends, employees of shops and public services, etc. Personal and administrative documents can usually be filled in by one, more or less designated member of the family. In the workplace, there are usually one or two employees of the group who handle most of the reading and writing tasks, if any. These ways of getting out are used to a great extent by people of all social classes and literacy levels and do not attract much attention or ridicule.

(e) At the turn of the nineteenth and twentieth centuries, and during some decades thereafter, literacy among children was highly prized, particularly for parts of the peasantry and working classes. It was regarded as a sign of personal dignity, autonomous judgment, and as a factor of material improvement and social mobility. Today, literacy skills have lost most of their prestige.

Insofar as home and neighborhood life influences children, these circumstances often contribute to the insufficiencies and fragility of the literacy abilities acquired at school. These same circumstances tend to cause adults with limited literacy to stay at poor levels, and the literacy skills of many others to deteriorate. The most deprived and other lower social classes are highly susceptible to the risk of quasi-illiteracy or low literacy. But the other social strata are not completely immune to this risk either. As they constitute the large majority of the population, they account for an important proportion of the quasi- and low-literate adults. In the USA, three-quarters of the adults living below the poverty level were either quasi-illiterate or low-literate in 1992. Yet 57 percent of the quasi-illiterate adults, as well as more than three-quarters of the low-literates belonged to the vast and heterogeneous nonpoor classes.

In spite of the great differences that separate the less and most developed worlds, the same basic mechanisms are at work in both with respect to literacy and illiteracy. They dominate the action of the multiple, more contingent factors influencing individual motivations and achievements at a young age and later in life. These mechanisms on the whole form rather clear tendencies towards an approximate equilibrium between, on the one hand, the common subjective and objective needs of the individual's social and occupational milieu and, on the other hand, their reading, writing, and computing capacities. Similar equilibrating tendencies develop with all other kinds of practical knowledge, from occupational and computer skills to foreign languages.

As long as they cope with the limited demands inherent in their roles at work and in society, with

respect to reading, writing, and computing, functional illiterates are generally not inclined to improve their literacy skills. Therein lies the greatest obstacle to the success of basic adult education programs.

See also: Cognitive Styles and Learning Styles; Educational Institutions, History of; Educational Learning Theory; Literacy and Illiteracy, History of; Literacy, Cognitive Psychology of; Literacy Education; Oral and Literate Culture; Printing as a Medium; Reading Skills; Socialization, Sociology of

Bibliography

Abadzi H 1994 *What We Know About Acquisition of Adult Literacy: Is There Hope?* The World Bank, Washington, DC

Cipolla C M 1969 *Literacy and Development in the West*. Penguin Books, Harmondsworth, UK

Girod R 1998 *L'illettrisme* (*Functional Illiteracy*), 2nd edn. Presses Universitaires de France, Paris

Hunter C St. John, Harman D 1979 *Adult Illiteracy in the United States. A Report to the Ford Foundation*. McGraw-Hill, New York

Irwin S, Kirsch I, Jungeblut A 1986 *Literacy: Profiles of America's Young Adults*. National Assessment of Educational Progress, Princeton, NJ

Kirsch I, Jungeblut A, Jenkins L, Kolstad A 1993 *Adult Literacy in America*. National Center for Educational Statistics, Washington, DC

Organization of Economic Co-operation and Development 1995 *Literacy, Economy and Society*. OECD, Paris

Organization of Economic Co-operation and Development 1997 *Literacy Skills for the Knowledge Society*. OECD, Paris

Organization of Economic Co-operation and Development 2000 *Literacy in the Information Age – Final Report of the International Adult Literacy Survey*. OECD, Paris

Rassekh S 1990 *Perspectives on Literacy: A Selected World Bibliography*. International Bureau of Education and UNESCO, Paris

Tanguiane S 1990 *Literacy and Illiteracy in the World: Situation, Trends and Prospects*. International Bureau of Education and Unesco, Paris

United Nations Educational, Scientific and Cultural Organization 1957 *World Literacy at Mid-Century: A Statistical Survey*. Monographs on Fundamental Education. UNESCO, Paris, Vol. XI

United Nations Educational, Scientific and Cultural Organization 1995 and 1998 *Compendium of Statistics on Illiteracy*. UNESCO, Paris

Weber E 1976 *Peasants into Frenchmen: The Modernization of Rural France, 1870–1914*. Stanford University Press, Stanford, CA

Wagner D A, Puchnez L D (eds.) 1992 *Literacy – Developing of the Future*. The Annuals of the American Academy of Political and Social Sciences, Philadelphia

R. Girod

Illness Behavior and Care Seeking

1. *What is Illness Behavior?*

The phrase 'illness behavior' conveys two, important, meanings. It differentiates disease, a biological phenomenon, e.g., cancer, congestive heart failure, diabetes, etc., from the behaviors associated with disease. These 'behaviors' range from the private experiences of disease symptoms through the subjective meaning of the disease and the emotions accompanying it, to the overt expression of 'sickness' such as seeking medical care, being excused from work and household duties, and other behaviors or 'privileges' involved in the sick role (Charmaz 1999). Investigators have also differentiated between illness behaviors and 'health behaviors,' the latter referring to actions taken in the absence of symptoms of physical illness (Kasl and Cobb 1966). This second differentiation is based on the assumption that it is important to identify whether behavior is instigated and sustained by the presence of symptoms and disease, or motivated either by concerns about the possibility of illness, or by motives totally unrelated to health and illness, e.g., physical appearance and social conformity.

The distinction is of questionable value, however, since behaviors to avoid and control illness can be performed in the presence of disease and the absence of symptoms as not all disease is symptomatic; and in the absence of disease and the presence of symptoms, as symptoms are experienced in the absence of detectable disease. It seems more reasonable to define 'illness behavior' as any action taken to cope with and manage the experience of symptoms and the presence of disease, or the anticipated occurrence or threat of symptoms and disease. A behavior that effects health, positively or negatively, but it is not seen to be health relevant, would not be regarded as a health behavior.

2. *Why is Illness Behavior Important?*

The use of health care resources is an important aspect of illness behavior. Using care has monetary and nonmonetary (mortality; poor quality of life) costs for the health care user, the user's immediate social network—family—and to the social institutions sharing the costs. Reducing monetary costs is cited frequently as a reason for reducing overuse, i.e., the use of care in the absence of diagnosable disease. The conditions or syndromes generating these 'unnecessary financial costs' include: somatization, chemical sensitivity, chronic fatigue syndrome, and most recently, Gulf War syndrome. Each involves the use of care, the absence of diagnosable disease, symptoms that the 'sick' person attributes to underlying disease, and each is presumed to be produced by the interplay of severe life stresses and personal vulnerabilities. The 'worried

well,' a label often applied to elderly women who are frequent participants in blood pressure screening programs, form another class of resource user, whose worrying drives behavior in the absence of disease and symptoms. The magnitude of the costs generated by these behaviors has been exaggerated, however, as they mainly involve the use of primary care services, which consume little more than 20 percent of the total health care budget, including both visits for unverifiable as well as verifiable conditions.

Life threatening conditions, e.g., coronary disease and cancers, and end-of-life treatments are the major source of financial costs to the health care system. Illness behaviors associated with the underuse of diagnostic and treatment regimens can add to their already considerable financial expense, and add to what may be more important costs of mortality, morbidity and reductions in quality of life. For example, risks to life, extensive morbidity, and the need for more invasive treatments, e.g., complex surgery and extended rehabilitative services, may result from delay in seeking care for cancers and coronary disease, as well as for less dramatic life threatening conditions such as appendicitis and ectopic pregnancies. These delays may be due to incorrect self or medical diagnosis, the absence of symptoms, and/or denial. Morbidity and social dysfunction can result from resistance to participation in post-treatment rehabilitative regimens following episodes of myocardial infarction, stroke and cancer. Resistance may emerge due to individual, socially supported but mistaken beliefs such as 'complete rest is essential to avoid death from heart attacks,' or to misinterpretation of the pain from the healing of the surgical wounds as pain from failure to repair the underlying disease, e.g., the pain attributed to the effect of physical activity on a damaged heart.

3. *Determinants of Illness Behavior and Care Seeking*

Given the number of syndromes used to describe illness behavior and the many ways in which it can inhibit effective self management, it appears wise to focus on those features of illness behavior that are common to multiple conditions. Describing just a few syndromes, e.g., somatization, hypochondriasis, or body dsymorphic disorder, would be of less interest and would overlap with their description in other entries in this volume. We will begin, therefore, by focusing on the factors affecting care seeking, and then address the effects of these factors on self-management during chronic illness.

3.1 *Symptoms and Care Seeing*

Each of us knows that changes in somatic experience, or symptoms, are powerful stimuli for seeking medical care. But symptoms alone are clearly not sufficient to motivate care seeking. Survey respondents have reported experiencing at least one new symptom each week of the year (Demers et al. 1980), but it is the rare individual indeed who seeks medical care 52 times a year, after excluding medical visits demanded for employment or insurance. Thus, symptoms are clearly a necessary but not sufficient condition for care seeking (Cameron et al. 1995). Somatic changes become symptoms when they are elaborated upon or represented by the individual as indicators of underlying illness. They may be represented as indicators of minor, self-limited and self-treatable events, of possibly serious disease in need of further evaluation, or as distinctive signs of disease threat that can be painful and which demand immediate attention from a qualified practitioner (Chrisman and Kleinman 1983, Levenhthal et al. 1980). Care seeking follows the interpretation that one is 'ill,' unable to manage the condition on one's own, and therefore, in need of medical attention.

3.2 *Symptom Interpretation: A Social-psychological Process*

Psychological theories model the cognitive and emotional processes that are internal to the person experiencing symptoms (see *Symptom Awareness and Interpretation*). When somatic sensations are given meaning, they are represented as symptoms of a suspected, specific disease, with a perceived cause (virus; stress), time-frame (to develop, to resolve, etc.), consequences (missed work, hospitalization, etc.), and responsiveness to treatment (control). As the individual lives in a social environment, these processes are sensitive to and affected by context. Chrisman and Kleinman (1983) posit three contexts that are involved in the interpretation of somatic changes: the popular sector, which comprises the individual's immediate social contacts, family and community, who interpret the person's somatic changes in light of a shared belief system; the folk sector, consisting of nonprofessional specialists, ethnic healers, religious or faith healers; and the professional sector, or culturally defined specialists such as physicians, osteopaths, chiropractors, etc., in Western countries such as Europe and the USA. The initial hypotheses respecting the meaning of the change is usually supplied by the individual experiencing the change, and members of the popular sector. The professional and folk sectors come into play when the initial popular evaluations suggest a need for additional expert help.

Participation by these three levels of social influence occurs in at least two different ways. First, the symptomatic individual, i.e., the prospective care-seeker, has internalized schemata or expectations as to what his or her symptoms represent, and how others

would interpret them. These expectations give meaning to what would otherwise be unelaborated sensations and signs. Second, when others in the popular sector are told about or observe these somatic changes, they will suggest interpretations of them and may also recommend specific ways of managing them. Studies show that great majority of symptomatic adults—often 90 percent or more—will talk to someone, usually a family member, about their symptoms. The response may be supportive ('What can I do to help?'), directive ('You should see your doctor right away!'), or simply grant permission to seek medical care ('I think it's something most anyone would get checked over').

3.3 Factors Affecting the Interpretive Process

Painful and disabling symptoms call for expert assistance as they are perceived as being beyond individual control. But symptom severity, e.g., excruciating pain, is not the factor most often associated with care seeking. Factors that appear to predict care seeking in the majority of illness episodes include variables such as duration (symptoms exceeding the expected time-line for minor and ignorable events), familiarity (do not fit into the pattern of familiar, manageable illnesses), and worry (worry about the symptoms). The properties of symptoms that are critical for care seeking are those that imply specific meanings or hypotheses regarding their identity; that is, whether the symptoms are signs of a common chest or head cold, temporary gastric upset, or indicators of a possibly serious illness such as a stroke, stomach ulcer or cancer (Bishop 1991). Thus, symptoms which last longer than expected, and symptoms that are unfamiliar, may be interpreted as a sign of a serious, life-threatening event. Once individuals believe they are indeed sick, they are more likely to be fearful and worried if they are uncertain about the meaning of the symptoms, and believe they may indicate a serious threat. The consequences of fear and worry, i.e., whether they provoke or delay care seeking, appears to depend upon their perceived source. Thus, worry that a disease is a threat to life is likely to encourage care-seeking, but worry or fear about treatment will delay it.

A variety of contextual factors also affect care seeking. For example, it is often proposed that life stresses increase the likelihood of care seeking (Mechanic 1979). Though some studies support their proposition, many others are inconsistent with it. Simple, main effect hypotheses, i.e., that an increase in a single factor such as stress is related to an increase in a behavior such as care seeking, fail to capture the complexities of decision making. An example of these complexities can be seen in a study of mothers' use of medical care for their sick children. Stressful life events pose barriers to seeking care on the day these events occur. A day later, when the event is resolved and the barrier it created is removed, there is an increase in care seeking above the level that would be seen had the event not occurred (Gortmaker et al. 1982). Personal interpretations as to how life stresses affect health further complicate the process of decisions to seek care. For example, if ambiguous symptoms such as fatigue and muscle aches appear in the context of a recent life stressor, i.e., a stressor that has lasted for but a few days, the symptoms are likely to be evaluated as signs of stress rather than as signs of illness. Symptoms of stress in this case are not seen as meriting medical care, and do not increase the level of care seeking. But if ambiguous symptoms appear in the context of prolonged life stress, i.e., stresses that have lasted four or more weeks, these symptoms are more likely to be interpreted as signs of stress-induced illness (Cameron et al. 1995). Adults will seek medical care under these conditions with the same frequency as for symptoms that are distinct signs of non-stress related illness.

It is worth noting that the duration required for a stressor to influence a change in decision from not seeking care to seeking care for ambiguous symptoms is roughly the same duration of life stresses that Cohen and colleagues (Cohen et al. 1998) discovered to be needed for stress to increase susceptibility to cold viruses. They assessed the presence and duration of life stress in participants, who were then randomized into five groups, all of whom live in an experimental unit. The members of four of these groups were subjected to controlled exposure to different cold viruses; the members of the fifth were exposed to neutral material. The persons who became sick after exposure to the viruses had been under stress for four or more weeks: shorter stressors did not increase susceptibility to colds. Thus, in at least this instance, people's appraisals of ambiguous symptoms, in combination with their life situations, correspond with data on actual effects of stress on susceptibility to colds.

A wide range of procedures are used in the evaluative processes underlying self-diagnosis. These include social comparison, e.g., if many other persons have the same symptoms it is likely not serious (Croyle and Jemmott 1991), or if we shared exposure to a particular pathogen, e.g., tainted food, we are likely to have the same illness; and historical self-comparisons, e.g., is this a new symptom or one experienced before, or if it is a change in function, is the trajectory of change worsening or improving. Symptom response to treatment, e.g., to home remedies, are also diagnostic. The headache that is cleared by aspirin is less likely to be seen to indicate serious, underlying disease than the headache that is recalcitrant to self-management. Treatments that control symptoms are valued over those that do not, though the evidence for efficacy may be based on little more than strongly held subjective convictions arising from a placebo effect that is

unrelated to any direct physiological benefit from treatment.

3.4 The Interpretive Process in Abnormal Illness Behavior

'Somatoform disorders' is the overarching label encompassing those syndromes defined by the reporting of symptoms and the conviction that one is ill in the absence of medically definable disease, e.g., somatization, hypochondriasis, etc. The process underlying this 'disorder' involves the amplification and conversion of benign somatic sensations into symptoms attributed to illness (Kellner 1990, Kirmayer and Robbins 1991). Repeated care seeking, doctor shopping, etc., are the consequences of this cognitive process. Many investigators have concluded that neurotic tendencies or trait negative affect, may underlie excessive care seeking in the absence of verifiable disease (Watson and Pennebaker 1989). Trait negative affect refers to a relatively stable feature of personality characterized by self reports that one's mood or emotional state is usually worried, anxious, blue or depressed; individuals with high scores on scales measuring this trait report a larger number of somatic complaints than individuals with low scores. There is little evidence, however, that persons high on trait negative affect make greater use of health care (Watson and Pennebakar 1989), as it is unlikely that many of these individuals have the additional cognitions that are necessary for such overuse, i.e., the beliefs that they are sick and that their somatic sensations are symptoms of illness. These beliefs distinguish those who suffer from somatoform disorders, less than 1 percent of the population, from the substantial number of persons who score high on trait negative affect (Robins et al. 1984). Thus, while individuals in both categories may report worry about health, worry is not equivalent to or sufficient for interpreting and attributing sensations to illness. Individuals with relatively high scores on trait negative affect appear able to distinguish between their life-long symptoms of emotional distress and symptoms of illness.

4. Chronic Illness, the Self and Illness Behavior

Analyses of illness behavior and care seeking often ignore the difference between acute and chronic illness. A focus on chronic disease raises at least two additional sets of questions for the study of illness behavior: those concerning adaptation to or living with illness, and those concerning changes in the self system. Chronic illnesses such as osteoarthritis, adult onset (type 2) diabetes, cancers (e.g., breast, prostate, colon), hypertension and cardiovascular disease, appear with increasing frequency in the later years of life. The average 70-year-old may have four or more such illnesses. These diseases bring reduction in physical function that restrict participation in daily activities, and these in turn are associated with increases in negative, i.e., depressed, affect, and declines in positive affect (Benyamini et al. 2000, Zeiss et al. 1996). They also lower people's assessments of their health, and these reductions are associated with reduced expectations respecting functional ability and life expectancy (Idler and Benyamini 1997). Managing unfamiliar disease can contribute to the sense of hopelessness that is an integral component of depression, and lower quality of life. The evidence indicates, therefore, that functional decline associated with both advanced age and chronic illness carries additional, psychological hazards, and these psychological changes reinforce the very functional deficits that initiated them (see *Chronic Illness, Psychosocial Coping with*).

4.1 Illness Behavior and the Management of Chronic Disease

Both cognitive and affective processes can be identified as sources of illness behaviors that are inappropriate for disease management. Cognitive factors are visible in individuals with congestive heart failure, i.e., their hearts ejecting less than 40 percent of the blood in the left ventricle to the rest of the body. These patients face episodes of heart failure and death due to the gradual decompensation of heart muscle and accumulation of fluid in the lungs and throughout the body. As their symptoms (fatigue, swollen legs, etc.) are slow to change and are omnipresent, they do not recognize them as indicators of their underlying heart condition. They fail to monitor changes in their symptoms and do not make effective use of available treatments in response to change: they act when facing crises, e.g., inability to breathe.

Another cognitive factor, conservation of resources (Leventhal and Crouch 1997), a strategy that encourages inactivity, appears to be maladaptive for living with many chronic illness. For example, it is associated with increases in pain over a three-year time span in rheumatoid arthritis patients (van Lankveld et al. 2000), and among coronary patients, it encourages inactivity and avoidance of exercise needed for improved cardiac function. A contrasting cognitive factor, belief that a strategy of vigorous engagement is necessary for good health, can lead cardiac and cancer patients to adopt behaviors such as jogging and the use of herbal supplements in order to strengthen the disease fighting capacity of their immune system. Although exercise is effective in reducing risk of mortality (Bijnen et al. 1999) few herbals deliver on their promises for health gains and some generate risk as they may interact adversely with chemotherapies of proven value, and many patients using supplements fail to communicate their use to their physicians.

A wide range of illness behaviors can be instigated by the emotional reactions elicited by the dysfunction and dysphoria of chronic illness. Depression is associated with nonadherence to medical treatment for angina, end-stage renal disease and cancer (DiMatteo et al. 2000), and both high and low levels of anxiety are related to inadequate self-management in patients with severe asthma. Asthmatics low on anxiety fail to take protective action (use of inhalers) to early signs of asthmatic attacks, and asthmatics high on anxiety over-use protective measures and need hospitalization due to the exacerbation of symptoms by fear.

4.2 Illness Behavior and the Changing Self

Chronic illness can change how individuals define themselves, and how they act in relation to others, to work and to the tasks of daily life. The functional and emotional impact of chronic illness and its permanence, presents a challenge to the very definition of the self. Thus, chronic illness represents a challenge both from the bottom up—its symptoms, signs, dysfunctions and effects—and from the top-down—the meaning of these signs and symptoms (Leventhal et al. 1999). Prior self identities and competencies and the social context (the popular (family), folk, and professional sectors), both directly influence and moderate these changes in self. Intellectual, emotional and social resources can support illness behaviors which are effective for the management of disease even when disease is not curable. Successful illness behaviors and effective disease management can strengthen the individual's sense of competence, create new and strengthen existent identities, and enhance the individual's ability to participate in life despite the drastic impact of illness on the physical self. Illness behavior at its best, i.e., the performance of new actions in managing the self with disease, can be the source of new, stable identities that define the value of life.

See also: Care and Gender; Caregiver Burden; Caregiving in Old Age; Health and Illness: Mental Representations in Different Cultures; Health Behaviors; Health Care Delivery Services; Stress: Psychological Perspectives

Bibliography

Benyamini Y, Idler E L, Leventhal H, Leventhal E A 2000 Positive healthiness: An important influence on self-assessment of health: Expanding our view beyond illness and disability. *The Journal of Gerontology* **55B**(2): 107–16

Bijnen F C H, Feskens E J M, Caspersen C J, Nagelkerke N, Mosterd W L, Kromhout D 1999 Baseline and previous physical activity in relation to mortality in elderly men: the Zutphen elderly study. *American Journal of Epidemiology*. **150**: 1289–99

Bishop G D 1991 Understanding the understanding of illness: Lay disease representation. In: Skelton J A, Croyle R T (eds.) *Mental Representation in Health and Illness*. Springer-Verlag, New York, pp. 32–59

Cameron L, Leventhal E A, Leventhal H 1993 Symptom representations and affect as determinants of care seeking in a community dwelling adult sample population. *Health Psychology* **12**: 171–9

Cameron L, Leventhal E A, Leventhal H 1995 Seeking medical care in response to symptoms and life stress. *Psychosomatic Medicine* **57**: 37–47

Charmaz K 1999 From the 'sick role' to stories of self: Understanding the self in illness. In: Contrada R J, Ashmore R D (eds.) *Self, Social Identity, and Physical Health*. Oxford, New York, pp. 209–39

Chrisman N J, Kleinman A 1983 Popular health care, social networks, and cultural meanings: The orientation of medical anthropology. In: Mechanic D (ed.) *Handbook of Health, Health Care, and Health Professions*. Free Press, New York, pp. 569–90

Cohen S, Frank E, Doyle W J, Skoner D P, Rabin B S, Gwaltney J M 1998 Types of stressors that increase susceptibility to the common cold in healthy adults. *Health Psychology* **17**: 214–23

Croyle R T, Jemmott III J B 1991 Psychological reaction of risk factor testing. In: Skelton J A, Croyle R T (eds.) *Mental Representation in Health and Illness*. Springer-Verlag, New York, pp. 85–107

Demers R Y et al. 1980 An exploration of the dimensions of illness behavior. *Journal of Family Practice* **11**: 1085–92

DiMatteo M R, Lepper H S, Croghan T W 2000 Depression is a risk factor for noncompliance with medical treatment: Meta-analysis of the effects of anxiety and depression on patient adherence. *Archives of Internal Medicine* **160**: 2101–7

Gortmaker S L, Eckenrode J, Gore S 1982 Stress and the utilization of health services: A time series and cross-sectional analysis. *Journal of Health and Social Behavior* **23**: 25–38

Idler E L, Benyamini Y 1997 Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior* **38**: 21–37

Kasl S V, Cobb S 1996 Health behavior, illness behavior and sick-role behavior. *Archives of Environmental Health* **12**: 531–41

Kellner R 1990 Somatization: Theories and research. *Journal of Nervous and Mental Disease* **178**: 150–60

Kirmayer L J, Robbins J M (eds.) 1991 *Current Concepts of Somatization: Research and Clinical Perspectives*. American Psychiatric Press, Washington DC

Leventhal E A, Crouch M 1997 Are there differences in perceptions of illness across the lifespan? In: Petrie K J, Wienman J A (eds.) *Perceptions of Health and Illness: Current Research and Applications*. Harwood Academic Press, London, pp. 77–102

Leventhal H, Idler E L, Leventhal E A 1999 The impact of chronic illness on the self system. In: Contrada R J, Ashmore R D (eds.) *Self, Social Identity, and Physical Health: Interdisciplinary Explorations*. Oxford University Press, New York, pp. 185–208

Leventhal H, Meyer D, Nerenz D 1980 The common sense representation of illness danger. In: Rachman S (ed.) *Contributions to Medical Psychology*. Pergamon, New York, pp. 7–30

Mechanic D 1979 The stability of health and illness behavior: Results from a 16-year follow up. *American Journal of Public Health* **69**: 1142–45

Robins L N, Helzer J E, Weissman M M, Orvaschel H, Gruenberg E, Burke J D, Regier D A 1984 Lifetime prevalence of specific psychiatric disorders in three sites. *Archives of General Psychiatry* **41**: 949–58

van Lankveld W, Naring G, van't Pad Bosch P, van de Putte L 2000 The negative effect of decreasing the level of activity in coping with pain in rheumatoid arthritis: An increase in psychological distress and disease impact. *Journal of Behavioral Medicine* **23**: 377–91

Watson D, Pennebaker J W 1989 Health complaints, stress, and distress: Exploring the central role of negative affectivity. *Psychological Review* **96**: 234–54

Zeiss A M, Lewinsohn P M, Rohde P, Seeley J R 1996 Relationship of physical disease and functional impairment to depression in older people. *Psychology and Aging* **11**: 572–81

H. Leventhal

Illness: Dyadic and Collective Coping

Within the last few decades, the individual perspective of viewing illness as a personal issue has been enlarged to include the social context (e.g., social support, family resources, marital functioning).

In regard to coping with illness in marriages, this enlarged perspective has been suggested by a number of theorists and researchers who have proposed different conceptualizations of dyadic coping, such as (a) an effort to establish and maintain close relationships, (b) a congruence of individual coping efforts by both partners, and (c) an interpersonal coping process involving both partners in order to reduce problem- and emotion-focused stress (see *Stress and Coping Theories*).

1. Dyadic Coping

1.1 Dyadic Coping as Relationship-focused Coping

The theorists of the first approach assume that partners must not only engage in problem- or emotion-focused coping when facing an important chronic or life-threatening illness, but also have to address the emotional needs of both partners and attempt to deal with the impact of the illness on the relationship. Consequently, Coyne and Smith (1991) have proposed relationship-focused coping as a third coping category, which includes one partner's efforts to reduce stress in the other partner and to guarantee his or her physical and psychological wellbeing. The authors suggest two main strategies: (a) active engagement (involving the partner in constructive problem-solving, problem-oriented discussions, and helping to explore the feelings of the partner) and (b) protective buffering (hiding concerns, denying worries, avoiding disagreements). DeLongis and O'Brien (1990) and O'Brien and DeLongis (1997) describe a similar aspect, namely focusing on dyadic efforts in order to maintain the quality of a relationship in situations of high stress (e.g., the dementia of one partner). They assume that relationship-focused coping preserves and enhances social relationships (or disrupts dysfunctional relationships). Thus, they emphasize the importance of empathic coping (affective awareness of the partner's problem and the ability to communicate this understanding to them in an appropriate manner) (see *Social Support and Stress*).

1.2 Dyadic Coping as Coping Congruence between Partners

Theorists referring to dyadic coping in terms of congruence or discrepancy between partners argue that the compatibility of problem- and emotion-focused coping strategies by both partners may be predictive of marital adjustment. This approach emphasizes (a) the usefulness of similar (congruent) or complementary coping strategies by partners facing a common stress event (Revenson 1994), (b) the assumption that it is optimal for couples to have similar emotion-focused coping strategies but dissimilar problem-focused coping strategies (Barbarin et al. 1985), or (c) that it is the average level of coping among partners that may be what is of particular importance (Pakenham 1998, Terry 1989)

1.3 Dyadic Coping as an Interpersonal Process

In the conceptualization of dyadic coping as an interpersonal process (Bodenmann 1995, 1997), a systemic-process oriented view of stress and coping in close relationships is suggested. Here, dyadic coping is understood as the interplay between the stress signals of one partner and the (dyadic) coping reactions of the other partner to these signals. Depending upon the stress event at stake, in dyadic coping the stress management resources of both partners are activated in order to maintain or restore a state of homeostasis among both partners as individuals, within the couple as a unit and in regard to other social partners, since it is assumed that one partner's wellbeing and satisfaction depend highly upon the other's wellbeing and satisfaction as well as social integration of both partners. According to the profile of the stressful situation and the individual and dyadic appraisals and goals, common (e.g., joint problem solving, mutual solidarity), supportive (e.g., practical advice, emotional comfort, belief in partner), or delegated dyadic coping (e.g., one partner takes over duties and tasks of

the other) are differentiated in this approach. These different forms of dyadic coping can be problem-centered or emotion-centered and may be either positive or negative (hostile, ambivalent, superficial) in nature.

1.4 Dyadic Coping and Physical Illness

A number of studies have investigated stress and dyadic coping in physical illness, such as cancer (Baider et al. 1995, Barbarin et al. 1985, Hannum et al. 1991, Kayser and Barbarin 1992), myocardial infarction (Coyne and Smith 1991, 1994, Suls et al. 1997), multiple sclerosis (Pakenham 1998), rheumatoid arthritis (Manne and Zautra 1990, Revenson and Majerowitz 1991), and physical disabilities among the elderly (Wright and Aquilino 1998) (see *Coronary Heart Disease (CHD), Coping with*).

Researchers using the congruence concept of dyadic coping predominantly reported that a high degree of similarity or high mean scores (average of the coping skills by both partners) were associated positively with marital adjustment, reduced burden for the spouse, and overall improved functioning in couples coping with severe illness (e.g., Pakenham 1998, Revenson 1994).

Others, such as Barbarin et al. (1985), have found that congruence between the wife and the husband is advantageous with regard to emotion-focused coping, whereas complementarity in problem-focused coping is correlated with better marital functioning. Kayser and Barbarin (1992), on the other hand, reported higher marital satisfaction and dyadic cohesion when both partners were congruent in problem-focused coping. They found lower satisfaction scores among wives when the husband practiced more emotion-focused coping strategies (independent of the wife's use of coping) in dealing with the cancer of the child.

A study by Baider et al. (1995) examining couples coping with the malignant melanoma of one partner suggested that dyadic coping (support by the spouse) has a different meaning for husbands and wives. Whereas husbands refer primarily to the dyadic coping of their spouse, wives tend to mobilize their extra-marital support network as well. In addition, Wright and Aquilino (1998) have pointed out that reciprocity in emotional support between the caregiving partner and the impaired spouse plays a crucial role in regard to marital adjustment and to the perception of the situation as burdensome.

Researchers referring to the relationship-focused coping construct reported no consistent findings concerning the usefulness of the protective buffering strategy. Coyne and Smith (1991) and Suls et al. (1997) found higher distress scores in wives and husbands as a function of the use of this strategy by the husband, the wife, or both. Thus far, relationship-focused coping proved to have little impact on illness recovery. However, Coyne and Smith (1994) showed that the patient's self-efficacy depended upon their spouse's practice of active engagement. Further research is necessary in order to highlight the usefulness of relationship-focused coping.

2. Dyadic Coping and Mental Disorders

2.1 Dyadic Coping in Depressed Couples

Mitchell et al. (1983) were among the first researchers to present a study integrating social aspects into the framework of stress regulation in depressed couples (see also *Depression, Hopelessness, Optimism, and Health*). They illustrated that clinically depressed persons coped less efficiently than their partner or controls, and that they experienced less social support from significant others. Partners of depressed individuals taken from a community sample, however, fell between the scores of depressed and nondepressed participants. The important role of support from the partner was also pointed out by Moos et al. (1998). A study by Gotlib and Whiffen (1989) assessing stress and coping in couples in which the wife was depressed revealed that depressed persons and their partner displayed more often congruent dysfunctional coping (confrontation, self-control, and escape-avoidance).

In a study conducted with couples where one partner was depressed ($N = 42$ couples), it was found that depressed couples, when compared with nonclinical controls, communicated their emotional stress less often and practiced dyadic coping (emotion-focused supportive and common dyadic coping) significantly less. On the other hand, both the depressed individual and their partner displayed ambivalent dyadic coping more frequently (Bodenmann 2000).

2.2 Dyadic Coping in Couples with an Anxiety Disorder

Studies concerning dyadic coping in couples in which one partner suffers from an anxiety disorder are rare (see *Anxiety and Anxiety Disorders*). In a study ($N = 70$ couples), it was found that there are significant differences between those couples suffering from anxiety disorder and control couples from a community sample. Both the patient with the anxiety disorder and their partner reported less problem- and emotion-focused stress communication and a higher rate of ambivalent and hostile dyadic coping. The partners of the patients, however, displayed more problem- and emotion-focused supportive dyadic coping (Bodenmann 2000).

2.3 Dyadic Coping in Couples with Sexual Dysfunction

In a study conducted with couples suffering from a sexual dysfunction disorder ($N = 120$ couples), empirical evidence was found indicating that those couples in which both partners experienced the sexual dysfunction disorder relied more heavily on ambivalent dyadic coping (see *Sexual Attitudes and Behavior*). Husbands suffering from a sexual dysfunction disorder perceived less emotion-focused supportive dyadic coping and less delegated dyadic coping from their spouse and a higher rate of ambivalent dyadic coping. It is interesting to note that in women, disorders of sexual desire and sexual aversion were associated with lower dyadic coping, whereas in men lower dyadic coping scores were related to sexual aversion and anorgasmy.

2.4 Dyadic Coping of Partners of Patients with Dementia

A study conducted by Tuor and Stadler (1999) using our concept of dyadic coping revealed that partners of AD patients ($N = 65$ couples) reported significantly less problem- and emotion-focused stress communication and less problem- and emotion-focused supportive dyadic coping from their partner, whereas their own problem- and emotion-focused dyadic coping scores were elevated (see *Dementia: Overview*). When the severity of the illness was included, all dyadic coping scales were reduced in the caregiving partner. It was interesting to note that negative dyadic coping (hostile, ambivalent) was not higher than among control persons. However, the partners of dementia patients seemed to practice their supportive dyadic coping much more superficially.

3. Discussion

The integration of dyadic coping in the study of coping with severe and chronic illness seems to call for attention. However, as this article suggests, there is currently still a lack of broad research activity including dyadic coping. Whereas social support from network partners outside the marriage are often examined, the relevance of the partner's support or dyadic coping are mostly ignored. As the first section has shown, there are several interesting approaches defining dyadic coping which have already proven their usefulness and clinical significance. The assumption that couples in which both partners contribute to dyadic coping and in which the partner is a substantial source of support in cases of illness-related needs has been supported by a number of studies. However, broader research activity is necessary in order to highlight the direct and moderating effects of dyadic coping in addition to individual coping efforts. It is a field that represents a promising challenge for new and innovative studies.

See also: Coping across the Lifespan; Social Support and Health; Social Support and Recovery from Disease and Medical Procedures; Social Support and Stress

Bibliography

Baider L, Perry S, Holland J C, Sison A, Kaplan DeNour A 1995 Couples and gender relationship: A sample of melanoma patients and their spouses. *Family Systems Medicine* **13**: 69–77

Barbarin O A, Hughes D, Chesler M A 1985 Stress, coping, and marital functioning among parents of children with cancer. *Journal of Marriage and the Family* **47**: 473–80

Bodenmann G 1995 A systemic-transactional view of stress and coping in couples. *Swiss Journal of Psychology* **54**: 34–49

Bodenmann G 1997 Dyadic coping—a systemic-transactional view of stress and coping among couples: Theory and empirical findings. *European Review of Applied Psychology* **47**: 137–40

Bodenmann G 2000 *Stress und Coping bei Paaren* (*Stress and Coping in Couples*). Hogrefe, Göttingen, Germany

Coyne J C, Smith D A F 1991 Couples coping with a myocardial infarction: A contextual perspective on wives' distress. *Journal of Personality and Social Psychology* **61**: 404–12

Coyne J C, Smith D A F 1994 Couples coping with myocardial infarction: Contextual perspective on patient self-efficacy. *Journal of Personality and Social Psychology* **8**: 1–13

DeLongis A, O'Brien T 1990 An interpersonal framework for stress and coping: An application to the families of Alzheimer's patients. In: Stephens M A P, Crowther J H, Hobfoll S E, Tennenbaum D L (eds.) *Stress and Coping in Later-life Families*. Hemisphere, New York pp. 221–40

Gotlib I H, Whiffen V E 1989 Stress, coping and marital satisfaction in couples with a depressed wife. *Canadian Journal of Behavioural Sciences* **21**: 401–18

Hannum J W, Giese-Davis J, Harding K, Hatfield A K 1991 Effects of individual and marital variables on coping with cancer. *Journal of Psychological Oncology* **9**: 1–20

Kayser K, Barbarin O A 1992 Effects of coping congruence on marital functioning among parents of children with cancer. In: (eds.) *54th Annual Conference of the National Council on Family Relations*, Orlando, FL

Manne S L, Zautra A J 1990 Couples coping with chronic illness: Women with rheumatoid arthritis and their healthy husbands. *Journal of Behavioral Medicine* **13**: 327–42

Mitchell R E, Cronkite R C, Moos R H 1983 Stress, coping, and depression among married couples. *Journal of Abnormal Psychology* **92**: 433–48

Moos R H, Cronkite R C, Moos B S 1998 The long-term interplay between family and extrafamily resources and depression. *Journal of Family Psychology* **12**: 326–43

O'Brien T, DeLongis A 1997 Coping with chronic stress: An interpersonal perspective. In: Gottlieb B (ed.) *Coping with Chronic Stress*. Plenum Press, New York pp. 162–90

Pakenham K I 1998 Couple coping and adjustment to multiple sclerosis in care receiver–carer dyads. *Family Relations* **47**: 269–77

Revenson T A 1994 Social support and marital coping with chronic illness. *Annals of Behavioural Medicine* **16**: 122–30
Revenson T A, Majerovitz S D 1991 Spouses' support provision to chronically ill patients. *Journal of Social and Personal Relationships* **7**: 575–86
Suls J, Green P, Rose G, Lounsbury P, Gordon E 1997 Hiding worries from one's spouse: Associations between coping via protective buffering and distress in male post-myocardial infarction patients and their wives. *Journal of Behavioral Medicine* **20**: 333–49
Terry D J 1989 *Stress, Coping, and Adaptation in Married Couples*. Unpublished doctoral dissertation. The University of Queensland, Australia
Tuor S, Stadler T 1999 *Individuelles und dyadisches Coping bei Ehepartnern von Demenzerkrankten* (Individual and dyadic coping in persons whose spouse suffers from dementia). Unpublished thesis. Basel University, Switzerland
Wright D L, Aquilino W S 1998 Influence of emotional support exchange in marriage on caregiving wives' burden and marital satisfaction. *Family Relations* **47**: 195–204

G. Bodenmann

Imagery versus Propositional Reasoning

1. Introduction

Probably the first theories of thinking, developed by the ancient Greeks, relied on the idea that thought relies on series of mental images. However, as Berkeley (1710) pointed out, images cannot represent categories easily and are inherently ambiguous—and yet our thoughts often rely on categories and are rarely (if ever) ambiguous. Thus, thinking must rely, at least in part, on some other type of mental representation. The obvious alternative candidate is language, but words are often ambiguous. Frege (1879) argued that thought relies on internal representations that are more abstract than language. These representations have come to be known as 'propositional representations,' and are like unambiguous statements of the gist of an idea (e.g., see Anderson and Bower 1973). Sachs (1967) provided empirical evidence for such representations in her demonstration that people remember not the specific words in a sentence but rather the idea expressed by it (see also Bransford and Franks 1971). Most cognitive scientists concur that propositional representations require an abstract and unambiguous syntax, but there is debate about the precise structure of this syntax, its similarity to natural language, and how this code is translated into natural language (Stillings et al. 1987).

Some cognitive scientists maintain that all thought relies solely on propositional representations, to the exclusion of other possible forms of internal representation (e.g., Pylyshyn 1973). In contrast, other theorists have maintained that the mind makes use of multiple forms of internal representation. These theorists have collected much evidence that mental imagery is a separate medium of thought, which preserves the spatial structure of pictures (e.g., Kosslyn 1980, 1994, Shepard and Cooper 1982). Although thought does not rely exclusively on imagery, imagery does play a role. In this article we consider the interplay between propositional representations and mental images during reasoning.

2. Propositional Reasoning

2.1 The Structure of Propositions

Propositions capture relations between concepts or objects. Propositions are characterized by a syntax that includes both (a) elementary symbols and (b) formal rules specifying how the distinct symbols may be combined. The meaning of a proposition depends upon the content, that is, the meaning each symbol and combination of symbols conveys. For example, the proposition 'Hans likes Maria' represents a belief a person may hold. This belief might be coded as the proposition '$x(h, m)$,' where x represents being in the state of 'liking,' and h and m represent Hans and Maria, respectively.

2.2 Reasoning with Propositions

Consider the following conditional syllogism:

If Hans likes Maria, then Hans also likes Susie.
Hans does not like Susie.
∴ Hans does not like Maria.

Does this conclusion follow from the premises? Most people find this problem easy, but psychologists have not yet determined how people reach the proper conclusion in even simple situations like this. Aristotle considered syllogistic reasoning to be the highest form of human thought, and William James (1890) dedicated nearly an entire chapter to the topic in his *Principles of Psychology*.

Until the late twentieth century, the dominant psychological view of reasoning was that people are naturally endowed with (or somehow acquire) the formal rules of logic. For example, the logical rule of *modus tollens* will always apply when terms are substituted for these variables:

If p, then q,
not q,
∴ not p.

These theorists maintain that people reason by constructing propositional representations of each of

the premises. They then apply formal rules of inference one by one to these representations, in steps that lead from the first premise to the final conclusion (e.g., see Braine et al. 1984, Rips 1994). The main rules are based on each of the logical connectives, such as *if*, *and*, *or*, and *not*.

2.3 Reasoning Errors

If the second premise above is changed to 'Hans does not like Maria', how will the conclusion change? Marcus and Rips (1979) report that 21 percent of respondents concluded that 'Hans does not like Susie.' This is a possible answer according to the rules of formal logic, but not a necessary one; Hans could like Susie even if he does not like Maria. Marcus and Rips provide support that an error in *encoding* produced this outcome. Specifically, that people encode the first premise as a biconditional, 'If p, then q' equals 'If q, then p,' instead of, as the laws of logic dictate, as a unidirectional.

Errors are very common in reasoning about such categorical syllogisms. Using the quantifiers *all*, *some*, and *no*, Mayer (1992) counted 64 possible pairs of premises, which in turn imply only 19 necessary, unambiguous conclusions. This leaves plenty of room to generate false conclusions or mistake possible conclusions for necessary ones. Can formal rules theories continue to account for these mistakes? Two common errors are as follows:
Undistributed middle:

All A are B; all C are B; therefore, all C are A.
Bostonians like cream pie.
Jack likes cream pie.
∴ Jack is a Bostonian.

Premise errors:

Some A are C; some C are A; therefore, some C are B.
Some Bostonians are college educated.
Some mechanics are Bostonians.
∴ Some mechanics are college educated.

According to formal rule theories, the contents of a syllogism should have no bearing on whether the conclusion follows from the premises. Nevertheless, several lines of research have shown that one's beliefs about and the contents of the premises can bias one's conclusion, that syllogisms with false premises and/or abstract symbols are more difficult than syllogisms with true premises and concrete symbols, and that people are more likely to come to a valid conclusion if they tend to agree with the premise's conclusion (Janis and Frick 1943, Evans et al. 1983).

2.4 Accounting for the Errors

Psychological theories of reasoning should be able to account for human reasoning errors and predict under what circumstances they will occur. There are currently several psychological theories that hold promise of satisfying these requirements (Begg and Denny 1969, Johnson-Laird 1998, Cheng and Holyoak 1985). These theories are based on semantic rules of reasoning instead of the syntactic method of formal rules (see Mayer 1992 for a good review). Formal rules are, by definition, independent of context or knowledge. However, our capacity to evaluate the various instantiations of formal rules and contrast the truth value of every possible conclusion based on the rules of formal logic is constrained by our limited working-memory capacity. It also seems that formal rules are not very useful for solving the practical reasoning problems that humans face. Semantic rules are derived from situation-specific (context-dependent) heuristic processes and the pre-existing knowledge they activate. For this reason, semantic theories postulate that the most optimal reasoning strategies select only highly relevant (and available) information for retrieval from long-term memory (Stillings et al. 1987). In other words, in everyday reasoning people tend to employ reasoning strategies that usually work effectively, even if they at times cause errors.

3. Imagery in Reasoning

The revival of interest in imagery representation was bolstered by anecdotal accounts of how visual mental imagery has facilitated understanding of scientific problems. For example, Albert Einstein reported visualizing himself traveling at the speed of light alongside a beam of light, and observing the appearance of the beam. This 'thought experiment' helped him to see that light is not an electromagnet field with no velocity, and to realize that such a thing did not exist. Nikola Tesla reported using visual mental imagery when working on his inventions. He used a visual image to conceive the details of the machine he was working on, and noted that his 'mental pictures' were more vivid than any blueprint. He also would run mental 'dry runs' to examine the machine for wear and tear over weeks of time (see Kosslyn 1983).

3.1 The Structure of Images

Experiments that have investigated image transformation (Shepard and Cooper 1982), and image zooming and scanning (Kosslyn 1980) have provided empirical support for the claim that mental imagery is a distinct type of internal representation. In these studies, subjects' response times increase as a function of the amount that an object must be transformed, zoomed in on, or scanned. Such evidence led Kosslyn (1975) to propose a distinct analogue subsystem, the visual

buffer, as the repository of internal representations that do not describe shape (as do propositional representations), but rather depict shape. That is, each part of a depictive representation corresponds to part of the object or scene being depicted, and the distances between the representations of the parts preserve the corresponding distances between the parts themselves. The visual buffer is a spatial structure that supports such representations. Various processes access perceptual memory and reconstruct the appearance of an object in the visual buffer; once reconstructed, the shape can be reinterpreted and manipulated in various ways. These depictive representations, unlike propositional representations, do not have an inherent truth content, are modality specific (propositions are amodal), are not formed according to a syntax, and are inherently ambiguous.

3.2 Reasoning with Images

Visual reasoning has two major components: first, a particular image must be formed in the service of solving a particular problem; second, once formed, the image must be manipulated and interpreted appropriately. Imagery can be used in two distinct types of visual reasoning, symbolically or as a type of 'mental simulation.'

3.2.1 Symbolic reasoning with images. Some types of imagery use relatively abstract images to stand in for concepts. For example, Huttenlocher (1968) provided empirical support that depictive representations can help people to reason about ordinal relations. She presented people with problems such as the following.

> John is taller than Sam, and Rick is shorter than Sam, who is tallest?

Subjects report visualizing a vertical axis to represent height, and visualizing a dot for each person—with each dot's height representing the relative height of the corresponding person. The same type of representation can be used even with abstract relations, such as 'smarter than.' Shaver et al. (1974) provide evidence that spatial arrangements (e.g., above–below) or metaphors for spatial arrangements (e.g., better–worse) are processed easiest when people construct such images, and that subjects make fewer errors when they use images. They also report that a subject's spatial ability predicts how well a person can use imagery to aid such reasoning (see also Hegarty and Sims 1994). Kosslyn and James (summarized in Kosslyn 1983) had subjects memorize the relative beauty of five people. They asked their subjects to visualize either five dots (one standing for each person) on a vertical axis that was positioned in the center of their visual field, or that was positioned off to one side. The subjects required less time to use the image in reasoning about the relative relations between pairs of people when they visualized the five-dot pattern in the center, which has higher resolution than the periphery (Finke and Kosslyn 1980). Furthermore, when a different group of subjects overlearned the relations (i.e., memorized them extremely well), this effect disappeared. It seems that overlearning fostered propositional coding of the relations so that imagery no longer was helpful for performing the task.

All students of logic quickly learn how imaging Venn diagrams helps one to avoid making the syllogistic fallacies mentioned above. However, it is not always clear how to form a model of abstract relations, and some types of problems that involve negation (see Johnson-Laird 1998) and contradiction (see Schwartz and Black 1996) cannot be visualized without propositional support.

3.2.2 Simulation reasoning. Many studies have shown that mental images can serve as a surrogate for perception (e.g., for reviews see Kosslyn 1980, 1994, Shepard and Cooper 1982). Thus, images can allow one to reason about the physical properties of objects in their absence, as was evident in the anecdotes we related about Einstein and Tesla. Instead of lugging furniture around, one can simply visualize the multiple ways of arranging furniture in one's living room. Such visualized patterns do not *arbitrarily* stand for objects or concepts as in symbolic imagery, but instead represent objects as they might actually appear. When using such an image in reasoning, one may let go of those features that are not relevant to the problem at hand, while honing in on pertinent ones. Hegarty (1992) showed that the static depiction of a pulley system can be mentally 'animated' using a piecemeal imagery strategy. That is, subjects animate local chunks stepwise (usually in the direction of causality) followed by other local chunks in order to infer something about the entire system.

There is good evidence that imagery simulations draw on some of the same mechanisms used in actual perception. For example, concurrent visuospatial tasks interfere more with mental simulation than with concurrent verbal tasks (Sims and Hegarty 1997). Moreover, spatial ability is correlated with a subject's accuracy in such tasks (Hegarty and Sims 1994), and hand gestures assist and promote spatial thinking (Athavankar 1999). Indeed, the time to infer the motion of interlocking gears is proportional to the angle through which the gears must move (Schwartz and Black 1996). When Schwartz and Black (1999) asked subjects which of two glasses (of equal height, but different diameter) almost full of water would pour at the shallower tilt, subjects believed the glass

with the smaller diameter would pour first. When then asked to simulate mentally using their hand and an empty glass (while marking the imagined water level with their thumb), however, they correctly changed their answers. Such findings suggest not only that imagery is dissociated from our descriptive beliefs (propositions), but also that there might be a haptic component of mental imagery that is distinct from the visual component.

4. Conclusion

The human information processing system is remarkably flexible, which is one source of its enormous power; indeed, no machine yet approaches the problem-solving prowess of a human being. Such flexibility arises in part because we not only can make use of different types of representations, but also can use them in different ways. Some reasoning may be most effectively accomplished using only propositional representations, and other reasoning may be most effective using a combination of types of representations. Images play a key role in some forms of reasoning, if only as 'mental notepads,' helping one to keep track of the things being reasoned about and the relations among them. However, it is clear that images alone cannot support reasoning. Images must be 'under description': any given image could have an infinite number of meanings (e.g., an apple could stand for an apple, fruit, food for worms, motherhood, etc.), and yet we are not confused about what we are thinking about when we have an image. The image must be interpreted by processes that themselves are not ambiguous. Hence, the role of imagery in reasoning must be understood within the context of a larger processing system, which also makes use of propositional representations.

See also: Feature Representations in Cognitive Psychology; Mental Imagery, Psychology of; Mental Representations, Psychology of; Problem Solving and Reasoning, Psychology of; Propositional Representations in Psychology; Visual Imagery, Neural Basis of

Bibliography

Anderson J R, Bower G H 1973 *Human Associative Memory*. V. H. Winston, Washington, DC

Athavankar U A 1999 Gestures, mental images and spatial reasoning. In: Gero J S, Tversky B (eds.) *Visual and Spatial Reasoning in Design*. International Conference on Visual and Spatial Reasoning in Design. Key Centre of Design Computing and Cognition, Sydney, Australia, pp. 103–27

Begg I, Denny P 1969 Empiricial reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology* **81**: 351–54

Berkeley G 1710 *A Treatise Concerning the Principles of Human Knowledge*. Open Court Edition, LaSalle, IL, 1963

Braine M D S, Reiser B J, Rumain B 1984 Some empirical justification for a theory of natural propositional logic. In: Bower G H (ed.) *The Psychology of Learning and Motivation*. Academic Press, New York

Bransford J, Franks J 1971 The abstraction of linguistic ideas. *Cognitive Psychology* **2**: 330–50

Cheng P W, Holyoak K J 1985 Pragmatic reasoning schemas. *Cognitive Psychology* **17**(4): 391–416

Evans J St B, Barston J L, Pollard P 1983 On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition* **11**(3): 295–306

Finke R A, Kosslyn S M 1980 Mental imagery acuity in the peripheral visual field. *Journal of Experimental Psychology Human* **6**(1): 126–39

Frege G 1879 Begriffschrift. In: van Heijenoort J (ed.) *From Frege to Goedel: a Source Book in Mathematical Logic 1879–1931*. Harvard University Press, Cambridge, MA, 1967

Hegarty M 1992 Mental animation: inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**(5): 1084–102

Hegarty M, Sims V K 1994 Individual differences in mental animation during mechanical reasoning. *Memory and Cognition* **22**(4): 411–30

Huttenlocher J 1968 Constructing spatial images: a strategy in reasoning. *Psychological Review* **75**: 550–60

James W 1890 *Principles of Psychology*, 2 Vols. Henry Holt, New York

Janis I L, Frick F 1943 The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms. *Journal of Experimental Psychology* **33**: 73–7

Johnson-Laird P N 1998 Imagery, visualization, and thinking. In: Hochberg J (ed.) *Perception and Cognition at Century's End: Handbook of Perception and Cognition*. Academic Press, San Diego, CA, pp. 441–67

Kosslyn S M 1975 Information representation in visual images. *Cognitive Psychology* **7**: 341–70

Kosslyn S M 1980 *Image and Mind*. Harvard University Press, Cambridge, MA

Kosslyn S M 1983 *Ghosts in the Mind's Machine: Creating and Using Images in the Brain*. W. W. Norton, New York

Kosslyn S M 1994 *Image and Brain*. MIT Press, Cambridge, MA

Marcus S L, Rips L J 1979 Conditional reasoning. *Journal of Verbal Learning and Verbal Behavior* **19**(2): 199–223

Mayer R E 1992 *Thinking, Problem Solving, Cognition*. W. H. Freeman, San Francisco

Pylyshyn Z W 1973 What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin* **80**: 1–24

Revlin R, Leirer V O 1980 Understanding quantified categorical expressions. *Memory and Cognition* **8**: 447–58

Rips L J 1994 *The Psychology of Proof: Deductive Reasoning in Human Thinking*. MIT Press, Cambridge, MA

Sachs J D S 1967 Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics* **2**: 437–442

Schwartz D L, Black J B 1996 Analog imagery in mental model reasoning: depictive models. *Cognitive Psychology* **30**: 154–219

Schwartz D L, Black T 1999 Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **25**(1): 116–36

Shaver P, Pierson L, Lang S 1974 Converging evidence for the functional significance of imagery in problem solving. *Cognition* **3**: 359–75

Shepard R N, Cooper L A 1982 Mental Images and Their Transformations. MIT Press, Cambridge, MA
Sims V K, Hegarty M 1997 Mental animation in the visuospatial sketchpad: evidence from dual-task studies. *Memory and Cognition* **25**(3): 321–32
Stillings N A, Feinstein M H, Garfield J L, Rissland E L, Rosenbaum D A, Weisler S E, Baker-Ward L 1987 *Cognitive Science: an Introduction*. MIT Press, Cambridge, MA

S. J. Stose and S. M. Kosslyn

Imagination: History of the Concept

In the course of the recent emergence of constructivism, which regards the world less as objectively given than as constructed and reproduced through (symbolic) practices as a web of social representations, and after the so-called 'linguistic turn' (see *Linguistic Turn*), the discursive construction and narrative constitution of the social and historical world has been in the center of interest in social theory and research. Despite their variety, reaching from semiological to analytical to hermeneutic accounts, these 'linguistic' approaches share a common perspective: social life and its history is seen as representations of the structure of its textualization and/or narration (White 1973, Clifford 1988, Ricoeur 1983). History and society, however, do not constitute themselves merely by narration, but also—and significantly—by nonlinguistic practices and images. Most recently, Castoriadis (1998) has underlined the 'imaginary institution of society' as the basic component of a comprehensive social theory. As a starting point to overcome both narrow textualism as well as the older neopositivism, the role of images and the imagination (*Einbildungskraft*) in social practices, which is at the core of any representation, be it narrative, iconographic, or aesthetic, has been focused upon. By this move, one has had to return to the classical philosophical topic—reaching from Plato and Aristotle to Kant and Husserl—of the imagination as the source and foundation of thought. Cognition is unthinkable without imagination. Even if one argues that cognition unfolds through the analytical operations that make phenomena distinguishable, these phenomena are shaped in and by imagination.

The semantic reach of the term 'imagination' includes as various connotations as mental imaginery (*Abbild*), invention of the mind (*Einbildung*, *Erfindung*) or illusion, which links imagination to fantasy. The substantive meaning of imagination or *Einbildungskraft* is rather similar to the Latin *imaginatio* and is distinguished—as in Kant—from fantasy, invention, and chimera. The semantic reach of the term fantasy, however, was clearly moved to the realm of the unreal during the eighteenth and nineteenth centuries. Although the term imagination had long been linked to various doctrines about what could be considered as 'true,' it is only from then onwards that the various connotations of different traditions collapsed into the one major distinction between the true and the untrue—for Ludwig Feuerbach, for instance, imagination is a key concept of the critique of religion, since religion presents imagined beings as real beings.

Across all varieties of meanings, however, there is a common reference of the term to the human ability to form—more or less intentionally—images of phenomena, be these absent or present. This faculty of mental imagery has been given central place in philosophical discourse with regard to (a) knowledge, concept formation, and language, (b) trace and memory, and (c) aesthetics.

1. *Knowledge, Concept Formation, and Language*

Within the philosophical legacy, 'the image' has been conceptualized as that which connects cognition to perception, the subject to the object. With the onset of modern times, empirical verification becomes the cornerstone of all knowledge. *Empería*, as that which the senses take note of, includes not only that which appears to the senses, but also the preservation of a sense impression in memory as well as the logical connection between such sense impressions in judgment (cf. Bohn 1990, p. 7).

The conjunction between thought (knowledge), image (as original image, *Urbild* or *eidos*) and memory goes back, in intricate ways, to Greek reason. In the ancient Greek tradition—and with Kant as well—the faculty of imagination is the source and foundation of thinking (to be distinguished, however, from both concepts and fantasy). For Aristotle, thinking occurs through mental images. Thinking is possible because we can place the things 'in front of our eyes'; then, the imaginative faculty can work upon the perception by the senses. Without the faculty of imagination, thinking would not be possible. Imagination is either a function of memory, of the retention of that which has made an imprint on memory, or it is productive, namely when it belabors the *phantasmata* that present themselves—that is why in ancient terminology imagination was called fantasy.

The Platonic term *phantasia* refers to both the truthful representation and the illusory appearance. *Phantasis*, vision, and *phantasma* are not linked primarily to the treacherous appearance, but they are true and untrue at the same time. *Phantasma* is that 'which appears,' thus not a representation. It is not primarily an intentional image, but that which appears in dream or when awake. For Aristotle, *imaginatio* is the faculty to receive the sense impressions, whereas *phantasia* is the faculty that bundles the *phantasmata* of the *imaginatio*. Thus, for Aristotle (*De memoria*),

every image becomes *theorema* or *mnemoneuma*, that is, the realistic representation of a thing or the reference towards another thing that shows no connection to the former one.

These constellations of thinking, which bring knowledge, image, and memory together, link imagination to sense perception. In one version, imagination reproduces—passively—the perceptions of the senses, whereas fantasy creates them—actively. In another, Kant's version, imagination is seen as twofold. It is passive as the receiver of sense perceptions, but it is active by idealizing them, that is, by holding them in an immaterial way.

For Kant, human cognitive capacity, on the one hand, is subdivided into two separate faculties, sensibility (*Sinnlichkeit*) and understanding (*Verstand*), to which imagination is then added as a third faculty. In analogy to this triple distinction, on the other hand, understanding is also distinguished from reason (*Vernunft*), a distinction to which the faculty of judgment is added. Every intuition (*Anschauung*) contains within itself a manifoldness that would be represented as unity, if 'the mind (*Gemüt*) did not distinguish time in the sequence of impressions' (Kant 1974, A 98). To turn the manifold into the unity of intuition, what is necessary is 'to run through this multiplicity and to hold it together' (A 98). Kant calls this operation the 'action of synthesis of apprehension.' It is followed by the synthesis of reproduction in imagination, that is, the connection of the representations according to a constant rule (to which the appearances are subjected as well). The final operation is the synthesis of the recognition in the concept, in which that which was gradually seen and that which was reproduced are united. 'There is therefore present in us an active capacity for synthesis of the manifold, a capacity which we call imagination and whose direct action on perceptions I call apprehension. For the imagination serves to bring the manifold of intuition onto an image, hence it must, prior to this, gather up impressions into its activity, that is to say it must apprehend them' (Kant 1974, A 119). The productive synthesis of imagination is—in contrast to the reproductive faculty of imagination, which rests on experience—thus the condition of possibility of knowledge. Imagination mediates between understanding and sensibility, by presenting the concept with the perception to which it corresponds. This operation of presentation is a scheme, in which the making of an image is no simple copying or reproducing, but a process of bringing into relation. Imagination here is linked simultaneously to intuition, image, and concept, in a way which is not only reproductive, but active. In *Anthropologie in pragmatischer Hinsicht*, imagination then is conceived as the faculty of perceiving an object 'even without its presence,' and a distinction between 'poetic' and productive or 'merely recalling and reproductive' aspects of imagination is introduced (1977, § 25, p. 466).

Hegel (1986, § 452–5) similarly brought memory in conceptual proximity to imagination by developing a three-stage model of recollection, imagination, and memory (*Zusatz*, § 451). He, too, made a distinction between reproductive and productive imagination. In this model, imagination is the faculty that determines the images, since imagination moves them into their being. Images that are no longer determined by the perception are, without consciousness (*bewußtlos*), stored independently of time and space; as infinite arrays of images, slumbering in the inner being, in the 'simple night,' (§ 454) they are unthought and not yet brought into the 'form of reason' and therefore belong only formally to the human being. To turn into conceptions (*Vorstellungen*) such abstractly stored images require the synthesis of the interior image with the remembered being; thus they will be brought to intelligence. Hegel calls *reproductive imagination* the intelligence that is active in this faculty and ties it to the associating relation of images. Imagination then is, first, the activity of associating images and subsuming them under general conceptions. In a second step, intelligence identifies its general conceptions with the particularities of the image and gives them existence as images. This image assumes the double form of symbol and of sign; such symbolizing an sign-creating fantasy forms the point of transition to memory (*Zusatz*, § 455). This sign-creating fantasy is—as symbolizing, allegorizing and poetic imagination—linked to language. As *productive* imagination, the formal aspect of art, the imagining activity arrives at its point of completion in itself and forms the general or the idea in the form of sensible being, of the image. Recollection, thus, turns a perception into an image that will be stored as something general in intelligence. As a reproductive faculty, imagination can repeat conceptions and connect them in new ways, but even though this connecting occurs rather freely, it is still linked to perception. As sign-creating fantasy, in contrast, it endows a sign with meaning in a sovereign way—such theory of signs forms the basis of Hegel's theory of language (Homann 1970, p. 273).

For Kant, consciousness is constituted in the moment of intuition in a movement in which activity and passivity join together. And for Husserl, consciousness is intentional, thus not restricted to the image as a phenomenon, but moving towards that which is represented in the image. Even though Husserl, thus, sees the object as always transcending consciousness, he nevertheless remains close to Kant's thinking, aiming to think inscription as that which constitutes sense impression as well as intellect. For Sartre (1986, 1994) similarly, following Husserl, the image is not the contents of consciousness, but consciousness itself and thus the possibility of knowledge.

Heidegger (1951) shows how Kant arrives at the concept of transcendental imagination, uniting sensibility and understanding and being rooted in time determined as pure self-affection—time and ima-

gination thus belong together. It was also Heidegger who directed the attention to a further connection between image and cognition. The modern era is characterized by its conquest of the *world as image*. It thus brings something into being that it puts in front of the 'presenting-producing' subject (*vorstellend-herstellend*; Heidegger 1980, p. 92). In precisely this sense, the experiment becomes this mode of producing by presenting (*vorstellendes Herstellen*) that characterizes modern science (see Shapin and Schaffer 1985). That which is to be researched is constituted as an object; the results of research are considered as a representation of reality; the world becomes readable like a book, and philosophy turns into 'the mirror of nature' (Blumenberg 1981, Rorty 1979, see *Representation: History of the Problem*). With the entry of a modern scientific conception into the historical and social sciences, subjective and objective accounts of the social and historical world are separated from each other and different genres and modes of writing established. By this move, however, the affinities between literary writing and scientific formation of concepts and knowledge, affinities that reside in their common ground in the power of imagination, are denied.

Even the words that are to provide evidence of thinking in concepts, namely, are always already connected to innumerable metaphors and figurative word-images (Blumenberg 1960). As Nietzsche remarks: 'We have not seen anything but the pictures of "causes and effects"! And it is precisely this pictoriality that makes the insight into the connection that is more essential than the sequence impossible' (1993, 2nd book, § 121). Human beings do not encounter the truth of the things and of the causal relations in these images, but only ever themselves: 'How is explanation to be possible if we turn everything into a picture, into our picture!' (Nietzsche 1988a, § 112, 1910, pp. 157–8). Word and image form an infinite movement that cannot be interrupted or destroyed: 'When I understand the pictoriality in language, then I do not transcend the picture-world in which we live. The images have not been removed when we have recognized that they are images. Truth and reality are grasped in the image. Reason that no longer wants the image retains only nothingness' (Jaspers 1990, p. 19). The figurative character of concepts cannot be dissolved, since even in the process of destruction only one thing happens, changes, and is created: images. Concepts are unable to provide any absolute notion, any real word of truth or any thinking that is purified of images, but they always speak of the forgetting of this notion and this word. As Nietzsche remarks 'truths are illusions about which it has been forgotten that they are illusions, worn-out metaphors without sensory impact, coins that have lost their image and now can be used only as metal, and no longer as coins' (Nietzsche, 1988b, pp. 880–1). There are no 'proper' words, they emerge by forgetting the word to which always already image on image have attached themselves, the one 'continuing the other, elucidating, modifying, correcting, effacing it' (Jaspers 1990, p. 19).

2. *Trace and Memory*

The relation between knowledge, image, and memory, thus, has a long intellectual tradition with a variety of different constellations. Already in the ancient 'art of memory' of Simonides or Cicero, remembrances of things as well as of words and concepts are seen as being inscribed in memory through sites and images (*notae*) (Yates 1966). Kant revives this connection when he links cognition to the *notae*, and relates intuition (*Anschauung*) and knowledge to such inscription and annotation in the memory. Freud (2000) similarly conceptualizes the psychical apparatus as a mystic writing-pad which—like a *tabula rasa* (which can be seen as passive as in Locke, or as active as in Hegel)—receives impressions and perceptions as well as stores them like memory.

This long, often rewritten and interrupted tradition reaches from Plato's 'wax tablets,' into which the things we remember imprint themselves (*Theiatet*), to Freud's 'magic pad' and to Derrida's concept of the trace: imprints, traces left in memory by the things and images released into narration by language. Remembrance and trace, concept and knowledge form the key components of this long and familiar story. St Augustine already referred to the pictoriality of this constellation, since for him the representing—in the sense of 'making present'—of the past occurs when the remembrance that is stored in memory appears to view, and—as he adds—if these images were missing, nothing could be named at all (Saint Augustine 1912). Obviously, it is not the things themselves that have made their imprint on memory, but their traces. When things past are remembered, what is retrieved from memory are not things themselves, but only words, which represent the images of these things which are imprinted on the mind like traces. It was already Aristotle who noted that it is not the thing which is imprinted into the *anima* to be conserved in the archive of memory, but its *trace*. Traces mark the *possibility* of remembrance of the disparate connections between them and, thus, of cognition and knowledge (Gawoll 1986/87, 1989).

Those traces are neither image nor absence of image, they are the possibility of image. Ferraris (1996) develops—with the concept of trace—a perspective that undermines the conventional dualism of aesthetics and image, on the one hand, and logos and concept, on the other, while maintaining the possibility of thinking their difference. Imagination touches upon remembrance and temporality through retention and protention, but it does so not as the product of time, but as its possibility instead. This view implies that perception and reception become the same: They open

towards the senses and to the intellect in a single, though divided moment. The opposition between intuition, image, logos, and concept is thus dissolved without denying the differences between these aspects.

3. Aesthetics

Christian Wolff and Gottlieb Baumgarten had already pointed to the connection between sensation, aesthetics, and knowledge. With the turn away from the principle of mimesis, of the imitation of nature, and with the radicalization of the principle of autonomy of the arts during the eighteenth century, then, the role of (productive) imagination for knowledge and for aesthetics is being revaluated.

Imagination—which Kant cast as the faculty of perception even without the presence of the object—is, when it is productive, a faculty of original representation. Aesthetic judgment judges in the presence of the object as much as under the condition of the absence of its objectness. The work of imagination when enjoying the beautiful shows itself independently of interest or cognition. Aesthetic grasping is not possible by mere conceptual framing and synthesis; aesthetic judgment, though, is nevertheless a judgment of reflection and not of the senses—otherwise it would merge with mere pleasure. Kant's third critique—complementing Kant's overall system—establishes the specificity of aesthetic judgment, or of the judgment of taste.

While the philosophical tradition assumed the intelligibility of the world by philosophical means, and thus saw philosophy as providing the arts with the means of interpreting the world, romantic thought, in particular early romanticism, reverts this relation and denies the primacy of philosophy for knowledge. It is precisely imagination that is to guarantee the autonomy of the aesthetic sphere, since imagination goes beyond sense perception. 'Imagination is this wonderful sense that can *replace* all senses for us—and that stands that much already under the power of our will. The outward-oriented senses appear to be entirely governed by mechanical laws—imagination, in contrast, is evidently not tied to the presence of, and touch by, external stimuli' (Novalis 1977, II, p. 650). On the basis of Kant's critique, the task of the arts is reformulated such that aesthetic imagination and aesthetic communication can bring together that which had been separated. At the same time, the active pole of imagination is emphasized, its faculty to create images independent of empirical perception.

Fichte's theory of productive imagination (*Grundlage der gesamten Wissenschaftslehre*, 1794), in particular, had a strong impact on the aesthetic conceptions of romanticism. Fichte understands by imagination a faculty that is suspended in between determination and non-determination, between the finite and the infinite, a suspending, and back and forward movement, of transcendental imagination however, refers to transitions between opposites and not their mediation and 'sublation' (*Aufhebung*) in a Hegelian dialectics. Schlegel 1988 draws upon such idea of a movement of productive imagination in his concept of 'progressive universal poetics' and in his view of infinite irony as poetic reflection, which explicates and presents the hidden transcendental mobility of self-consciousness (1988, II, pp. 114–6, see *Romanticism: Impact on Social Thought*). The conception of infinite irony resorts to the idea of an infinite reflexivity, which is not limited by any kind of fixation, and thus also rejects the rhetoric distinction between that which is meant and inauthentic utterances. In infinite irony, poetics is not only seen as the 'image of an era' (1988, p. 114), but imagery is also linked to ever-incomplete becoming and, thus, to time. The opposition between concept and image, between aesthetics and logic, between the sensible and the intelligible can no longer be upheld in such nondialectical indecidability.

According to Schlegel, imagination, always striving towards the pictorial, receives its highest form of freedom by free invention, thought, and poetics. For Schlegel, the image is the counter-thing that has been liberated of the domination by the thing when the spectator's faculty of imagination meets with the one of the artist. It is linked to a kind of love that is inclined towards the 'counter-I,' aiming at the overcoming of the separation of theoretical and practical philosophy and having to achieve the anticipation, 'in a leap,' of a concept that yet remains to be formed (see Homann 1970, p. 282).

Even though the concept of imagination appears no longer to hold a central place in philosophical discourse as a category of artistic activity and as an instance of mediation, it has remained vibrant as a liberating power that restores the lost human freedom against technology and reification—such as, for instance, in Herbert Marcuse (1964) (or similarly in Castoriadis, see above), who thinks freedom in aesthetic categories and demands the liberation of fantasy through political praxis.

See also: Aristotle (384–322 BC); Hegel, Georg Wilhelm Friedrich (1770–1831); Imagery versus Propositional Reasoning; Kant, Immanuel (1724–1804); Nietzsche, Friedrich (1844–1900)

Bibliography

Augustine 1912 *The Confessions*. Harvard University Press, Cambridge, MA

Blumenberg H 1960 Paradigmen zu einer Metaphorologie. *Archiv für Begriffsgeschichte* **6**: 7–142

Blumenberg H 1981 *Die Lesbarkeit der Welt*. Suhrkamp, Frankfurt am Main, Germany

Bohn V (ed.) 1990 *Bildlichkeit Internationale Beitrage zur Poetik*. Suhrkamp, Frankfurt/M, Germany
Castoriadis C 1998 *The Imaginary Constitution of Society* trans. by Blamey K, MIT Press, Cambridge, MA
Clifford J 1988 *The Predicament of Culture. Twentieth-century Ethnography, Literature, and Art*. Harvard University Press, Cambridge MA and London
Ferraris M 1996 *L'immaginazione*. il Mulino, Bologna, Italy
Fichte J G 1956/1974 Grundlage der gesamten Wissenschaftslehre. Meiner, Hamburg, Germany
Freud S 2000 1925 Notiz uber den 'Wunderblock'. Studienausgabe. In: Mitscherlich A, Richards A, Strachey J (eds.) *Psychologie des Unbewupten*. Fischer, Frankfurt/M, Germany, Vol. 3 pp. 363–69
Gawoll H-J 1986/87 Spur: Gedächtnis und Andersheit. Teil I: Geschichte des Aufbewahrens. *Archiv für Begriffsgeschichte* **30**: 44–69
Gawoll H-J 1989 Spur: Gedächtnis und Andersheit. Teil II: Das Sein und die Differenzen—Heidegger, Levinas und Derrida, *Archiv für Begriffsgeschichte* **32**: 269–96
Hegel G W F 1986/1830 *Enzyklopädie der philosophischen Wissenschaften*, Dritter Teil. Die Philosophie des Geistes. In: Moldenhauer E, Michel K M (eds.) *Werke*, Vol. X. Suhrkamp, Frankfurt am Main, Germany
Heidegger M 1951/1929 *Kant und das Problem der Metaphysik*. Vittorio Klostermann, Frankfurt am Main, Germany
Heidegger M 1980/1938 *Die Zeit des Weltbildes*. Vittorio Klostermann, Frankfurt am Main, Germany, pp. 73–110
Homann K 1970 Zum Begriff Einbildungskraft nach Kant. *Archiv für Begriffsgeschichte* **14**: 266–302
Jaspers K 1990 *Die Sprache. Über Das Tragische*. Piper, München
Kant I 1974/1781 Kritik der reinen Vernunft. In: Weischedel W (ed.) *Werkausgabe, Vols. III–IV*. Suhrkamp, Frankfurt am Main, Germany (*Critique of Pure Reason*, Politis (ed.) trans. Based on Meiklejohn, Everyman, London 1993)
Kant I 1977/1798 Anthropologie in pragmatischer Hinsicht. In: Weischedel W (ed.) *Werkausgabe, Vol. XII*. Suhrkamp, Frankfurt am Main, Germany, pp. 399–690
Marcuse H 1964 The One-dimensional Man. *Studies in the Ideology of Advanced Industrial Society*. Beacon Press, Boston
Nietzsche F 1988a [1882/1887] *Die fröhliche Wissenschaft*. In: Colli G, Montinari M (eds.) *Kritische Studienausgabe, Vol. III*. DTV/de Gruyter, München, pp. 343–651 (*The Joyful Wisdom* [trans. Common T], 1910, Foulis, London)
Nietzsche F 1988b/1873 Über Wahrheit und Lüge im aussermoralischen Sinne. In: Colli G, Montinari M (eds.) *Kritische Studienausgabe*, Vol. I. DTV/de Gruyter, München, pp. 873–90 (On truth and lying in an extra moral sense. In: Sander L, Gilman, C B, Parent D J (eds.) *Friedrich Nietzsche on Rhetoric and Language*. 1989, Oxford University Press, New York)
Nietzsche F 1993/1881/1887 *Morgenröthe. Gedanken über die moralischen Vorurtheile*. In: Colli G, Montinari M (eds.) *Kritische Studienausgabe, Vol. III*. DTV/de Gruyter, München, pp. 9–331
Novalis 1977 Schriften. In: Kluckhohn P, Samuel R (eds.) *Die Werke Friedrich von Hardenbergs*, Kohlhammer, Stuttgart, Germany, Vol. I–IV
Ricoeur P 1983 *Temps et recit*. Seuil, Paris, Vol. 1–3
Rorty R 1979 *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ
Sartre J-P 1986/1940 *L'imagination*. Gallimard, Paris
Sartre J-P 1994/1936 *L'imagination*. PUF, Paris
Schlegel F 1988/1798 Athenäums-Fragmente. In: Behler E, Eichner H (eds.) *Kritische Schriften und Fragmente* 1794–1828, *Studienausgabe, Vol. II*. Ferdinand Schöningh, Paderborn, München, pp. 105–56
Shapin S, Schaffer S 1985 *Leviathan and the Air-pump*. Princeton University Press, Princeton, NJ
White H 1973 *Metahistory. The Historical Imagination in Nineteenth-century Europe*. Johns Hopkins University Press, Baltimore, MD
Yates F A 1966 *The Art of Memory*. Routledge and Kegan Paul, London

H. Friese

Immigrant Fertility

The phrase immigrant fertility is fraught with ambiguity. Does it refer only to the fertility of immigrants? Does it also refer to the fertility of international migrants who are not immigrants in the technical sense of the term? Does it refer to the fertility of minority group members who are descendants of immigrants? In its legal sense, the term immigrant refers only to persons who, with the authorization of their countries of destination if not always their countries of origin, cross international borders in order to establish residences in new countries. Immigrants include neither unauthorized international migrants nor authorized nonimmigrant crossers (tourists, students, and business travelers). Whether or not to include unauthorized migrants in research on immigrant fertility depends on the purpose of the study. Such decisions will not only be the subject of debate, they may also affect the results of research. In countries with birthright citizenship, such as the United States, the offspring of unauthorized migrants are citizens perforce and may constitute a nontrivial component of the fertility of a given birth cohort or time period. Because of space restrictions, the present discussion cannot address or resolve all such definitional issues. So we simply note here that by immigrant fertility we mean the fertility behavior in a given country of destination of immigrants, unauthorized international migrants, and their relatively recent descendants (the second and third-or-later generations).

Fertility refers to the childbearing patterns of women or couples in their new countries of residence. Because the levels and patterns of immigrant fertility hold significant demographic, sociological, political, and economic implications for destination societies, they are of considerable practical as well as theoretical interest. Analyses of factors affecting fertility levels and trends among immigrants and nonimmigrants may involve a focus on the number of births, the timing of these births, and associated reproductive behavior such as union formation (including marriage

and co-habitation) and contraceptive behavior. Among the most important factors affecting the levels and trends in the fertility behavior of immigrant groups are variables relating to the processes of incorporation, disruption, selectivity, and ethnic and minority group formation that may be experienced by the members of immigrant groups.

1. Major Substantive Domains

Demographically, the childbearing patterns of immigrants directly affect the future size and composition of the overall population. Thus, any distinctive levels or patterns of childbearing among immigrants should be factored into the production of population estimates and projections (Edmonston and Passel 1992, Smith and Edmonston 1997). These effects are most consequential in situations where immigration streams substantially increment the size of socially demarcated ethnic (minority) groups, return migration is comparatively low, and the childbearing schedules of the immigrant women are distinct from those of the native-born (majority group). Such circumstances have recently characterized the demographic regimes in many countries (e.g., the United States, Israel, Australia, Belgium). The demographic impact of immigrant fertility may be especially pronounced at the subnational level when immigrant populations are heavily concentrated in certain regional locations.

At various times and places, the dynamics of intergroup relations have been influenced by a politicization of immigrant family formation patterns. Members of majority groups may see higher immigrant fertility, real or perceived, as threatening to the majority group and as reinforcing ethnic or racial stereotypes of associated immigrant groups. The history of the United States reveals a recurring interest in the relative reproductive levels of the latest round of immigrants (and their children) vis-à-vis those of the native-born population. King and Ruggles (1990), for example, point to the inclusion of natality stereotypes into turn-of-the-century nativist ideology, including predictions of 'racial suicide,' even though the *actual* fertility of immigrants and their offspring was on par with or below that of third-generation Americans. A century later, unflattering stereotypes of large immigrant family sizes remain alive in the political discourse of many countries.

Sociological concerns, however, dominate social science research on immigrant fertility. Such research often focuses on the extent to which the childbearing patterns of immigrants vary from those of natives and on the pace with which convergence between the two groups occurs. Such convergence is what we mean when we refer to incorporation. We do not use the term assimilation here because we wish to emphasize that convergence can involve the movement of two groups toward a single point, rather than merely the movement of immigrants toward the level of the native-born group. The persistent and pervasive influence of incorporation perspectives in the sociological literature on immigration has led to the general hypothesis that with the passage of time (either individual or generational), immigrant groups and their descendants will become increasingly less and less dissimilar from natives along a variety of social and economic dimensions. As a consequence, such perspectives predict immigrant and native childbearing patterns will come to resemble those of one another (Bean and Swicegood 1985). The extent of convergence of fertility behavior has often been viewed as a measure of social integration between immigrants and natives (e.g., Coleman 1994).

2. Incorporation Processes

The preponderance of research on immigrant fertility has focused on the childbearing patterns of women who have migrated from less-developed countries with relatively higher fertility levels to more developed countries where demographic transitions to lower mortality and fertility are further advanced. Historically, empirical research has concentrated on assessing levels and changes in the gap between the fertility levels of immigrant and native-born women. Theories of incorporation processes have provided the theoretical backdrop for such investigations. Incorporation perspectives on immigrant childbearing implicitly if not explicitly invoke the operation of social influences that are grounded in both origin and destination societies. Because acculturation, or the process of cultural incorporation, has been assumed not to occur quickly, the hypothesis has often been advanced that childbearing norms acquired prior to immigration (or from immigrant parents) will for some period of time continue to influence fertility after migrants arrive in the destination society. The significance of cultural norms will depend on the degree of difference between reproductive norms at origin and destination and the strength with which they continue to be salient for immigrants in their new homes. Little research has attempted directly to assess the influence of fertility norms. Instead, most research has focused on structural incorporation by examining key elements of the structural position of immigrants in destination societies, such as education and labor market position, and how these affect immigrant and native fertility behavior.

Both conceptually and empirically, incorporation may be viewed as occurring among individual immigrants and across generations. In a given destination, either or both of these may be in operation. However, empirical research has tended to address one or the other. In the first case, which involves only the examination of immigrant women, the term adap-

tation is sometimes used to refer to changes in the fertility behavior of immigrant women that occur with increasing time in the destination country. Studies of adaptation are likely to focus on correlations between measures of length of time spent at destination and changes in the gap between immigrant and native-born fertility. In the second case, which involves the examination of both immigrants and their descendants, the fertility of immigrant women is compared to that of the second and subsequent generations. This approach is frequently confined to a single country of origin group compared to later generational co-ethnics and natives. The existing literature has not yielded empirical generalizations about the factors conditioning the pace of either of these kinds of incorporation.

3. *Disruption Processes*

Over the period when international migration occurs, the complex interplay of forces deriving from planning and carrying out the move, changing living arrangements, undergoing spousal and family separations, and experiencing unanticipated crises may result in the disruption of childbearing schedules. The reallocation of resources may be required to finance the move, spouses may undergo lengthy periods of separation, and relocation into a new society can invoke a variety of stresses both economic and cultural. Empirical analyses provide indirect evidence of the operation of disruption effects on the fertility of immigrants to the United States (Blau 1992, Kahn 1994, Stephen and Bean 1992). However, the absence of longitudinal fertility data makes these effects difficult to detect, in part because some or all of the fertility postponed by disruption may be compensated for later by accelerated childbearing during the postmigration years. Kahn's (1994) analysis of US data produced patterns consistent with general disruption effects, but the degree of compensation varied by national origin group. Among women who were recent immigrants in 1980, Mexican-born women increased fertility relative to the native-born population over the remainder of the decade, thus indicating substantial compensation, whereas the fertility of Asian-born women fell further behind that of the native born. In another case, Aussiedler immigrants to Germany from the former Soviet Union showed fertility levels prior to settlement in Germany that were 50 percent greater than those of comparable German women. In the several years following immigration, their fertility dropped to only 40 percent of that of native German women, whose fertility rates are comparatively low by international standards (Dinkel and Lebok 1997). In another case, Indochinese refugees to the United States in the 1980s exhibited very high fertility during the period immediately following their resettlement (Weeks et al. 1989), demonstrating that disruption can sometimes be so short-lived as to be almost nonexistent.

4. *Selectivity Processes*

On average, immigrants are seldom representative of the general populations in their former countries of residence. Moreover, the characteristic profile of recent immigrants tends to change over time. Immigrant streams can differ with respect to factors like the percentage of risk takers, the level and distribution of human capital resources, and the percentage married or already having children at the time of migration. The selectivity of the immigrant streams with respect to such fertility-related characteristics must be considered when examining (a) variation in fertility patterns across immigrant groups from different origins or (b) the change in fertility patterns across successive immigration cohorts. Several studies have provided indirect evidence of selection effects. In the United States, postwar changes in the selectivity of immigrants both within and across origin countries have resulted in relatively higher fertility among the immigrant population than would otherwise have been the case (Blau 1992). Selectivity effects are also important because they can condition the strength and duration with which factors in the country of origin continue to influence childbearing in destination societies. For example, Kahn (1988) found that childbearing patterns more typical of the country of origin are most pronounced among the least educated, least incorporated immigrant women in the United States.

5. *Analytical Design and Methodological Issues*

Research on incorporation has clearly shaped the analytical design of immigrant fertility studies. A common feature is the comparison of the behaviors of immigrant women with those of some reference group. The appropriate referent group is defined by the substantive focus of the study (see Singley and Landale 1998). Measures of fertility behavior are modeled as a function of immigrant status and demographic factors such as age and marital or union status. Fertility differences between immigrants and the reference group are determined and then statistical controls for other factors such as female education, labor force participation, income, and individual and household characteristics are included in the models. The specific set of control variables is dictated by what aspect of fertility behavior is under consideration and by the theoretical and disciplinary orientations of researchers.

In most places immigrants are a relatively small portion of the population. Thus, large sample sizes are necessary for sufficient numbers of immigrants to allow for generalization. This problem is exacerbated in situations where the immigrant population is spread across multiple national origins. Information on date of immigration is not always collected and in some countries immigrant status is not included in official

statistics once naturalization has occurred. When only cross-sectional data on number of children are available, researchers cannot distinguish between *exposure effects*, or the amount of time spent at destination, and *cohort effects*, which pertain to the group of immigrants arriving at their new country at a particular time. Exposure effects provide evidence on incorporation processes, while cohort effects may be grounded in changing selectivity of the immigrant stream. Similar confounding of conceptually distinct influences may occur when examining a time series of period fertility rates for evidence of immigrant/native-born convergence. Coleman (1994) for example, reports that since the 1960s the annual birth rates for immigrants in many European countries have moved closer to those of the native-born population. This convergence, while consistent with the idea of behavior change within the life course of immigrant women, could also reflect different childbearing patterns among more recent immigrants compared to those migrating in earlier periods. Again processes of incorporation cannot be clearly distinguished from cohort change. Processes of adaptation similarly may be empirically confounded with disruption effects. A pattern of lowered fertility around the time of immigration is consistent with either type of effect.

The simultaneous and most complete assessment of all of the factors involved in differential immigrant fertility would require detailed longitudinal fertility data for (a) immigrant women and their descendants in countries of destination, (b) native-born women in countries of destination, and (c) nonmigrant women in countries of origin. For destination countries with multiple-origin streams, the data would need to be collected for women in each country of origin. Questions regarding the aggregate effects of selectivity over time would further require data from multiple age cohorts. Data limitations have hindered the possibility of directly addressing intergenerational changes in childbearing. Analyses that examine differences across generation groups typically compare immigrant women with second or later-generational women, all of whom are the same age. Data have not been available to compare the fertility of immigrant women directly with that of their daughters.

6. *Current Approaches and Emerging Lines of Inquiry*

Recent research reflects a trend toward more direct measurement of factors associated with incorporation. Okun (1997) finds, for example, that national military service is strongly associated with adoption of modern forms of birth control among North African and Middle Eastern Jewish immigrants to Israel. Military service is argued to rapidly break down traditional fertility norms prevailing in the country of origin. Other current research more explicitly considers discontinuities between the social context of origin and destination societies. Hwang and Saenz (1997) compare the fertility levels of immigrant women from the People's Republic of China with those of foreign-born Chinese women from other source countries. The immigrant women from the People's Republic reveal lower cumulative fertility overall, but higher levels of childbearing once in the United States. This heightened fertility increases with each additional year within the childbearing ages spent in the United States. The authors' view this as consistent with an 'emancipation' hypothesis whereby freedom from the restrictions of the Chinese one-child population policy allow pronatalist and son preference values to be realized. Still other current work involves an effort to anchor the study of immigrant fertility more squarely into a life course perspective. Singley and Landale's (1998) research on migration and fertility among Puerto Rican women illustrates the unfolding relationships between the timing of migration, marriage, and childbearing, viewing migration more as a series of events, each of which may reciprocally interact with decisions related to other life course transitions. Such frameworks are very worthwhile but demand detailed longitudinal data to implement, a luxury unavailable in the case of many countries, especially less developed countries.

The amount of immigration and the nature of immigrant settlement patterns are both factors that contribute to establishing the social and economic contexts in which immigrants live, work, and make decisions relating to reproductive behavior. These contexts may shape the extent to which old norms are reinforced or new ones encountered, and they condition the way in which personal characteristics such as education or language proficiencies end up influencing family-building strategies of individuals or couples. The new patterns of international migration to the United States offer a challenge to the incorporation perspective that is anchored in a focus on the context of the immigrant experience. Massey (1995) argues that the new immigration will have a different meaning for ethnicity and race in American life than early European waves because it is grounded in fundamentally different structural circumstances. Flows of new immigrants are expected to be more continuous across a longer duration of time, creating a state of more or less 'permanent immigration.' Geographic concentration of the new immigration creates and sustains distinctive language and cultural communities on an unprecedented scale.

To the extent that sustained streams of new immigrants reinforce ethnicity, distinctive norms about childbearing may also be strengthened within the entire ethnic community, including native-born generations. This may be especially true for a group like Mexican Americans in the United States. Their immigration occurred from a higher fertility context.

Areas of residential concentration in the United States are relatively close to Mexico, their migration patterns frequently involve return episodes, and numerous large Spanish-language communities are available as destination points. Evidence of fundamental shifts in fertility regime has been found in a recent study of Mexican origin generational differences (Bean et al. 2000). The fertility of Mexican-born women was observed to substantially exceed that of non-Hispanic white women both before and after the introduction of standard sociodemographic controls. In accordance with the predictions of incorporation perspectives, the fertility of the second generation was considerably lower than that of immigrants. However, third- and later generation women showed levels of childbearing that were not only higher than those of non-Hispanic white women but also higher than those of second-generation women. Such higher fertility suggests that incorporation perspectives may not be fully adequate for explaining many extant ethnic patterns of childbearing.

See also: Assimilation of Immigrants; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants; Immigrants: Economic Performance; Migration and Health; Migration: Sociological Aspects; Migration to the United States: Gender Aspects

Bibliography

Bean F D, Swicegood C G 1985 *Mexican American Fertility Patterns*. University of Texas Press, Austin, TX

Bean F D, Swicegood C G, Berg R 2000 Mexican origin fertility: New patterns and interpretations. *Social Science Quarterly* **81**: 404–20

Blau F D 1992 The fertility of immigrant women: Evidence from high fertility source countries. In: Borjas G J, Freeman R B (eds.) *Immigration and the Work Force: Economic Consequences for the United States and Source Areas*. University of Chicago Press, Chicago

Coleman D A 1994 Trends in fertility and intermarriage among immigrant populations in Western Europe as measures of integration. *Journal of Biosocial Science* **26**: 107–36

Dinkel R H, Lebok U H 1997 The fertility of migrants before and after crossing the border: The ethnic German population from the former Soviet Union as a case study. *International Migration* **35(2)**: 253–70

Edmonston B, Passel J S 1992 Immigration and immigrant generations in population projections. *International Journal of Forecasting* **8**: 459–76

Hwang S-S, Saenz R 1997 Fertility of Chinese immigrants in the U.S.: Testing a fertility emancipation hypothesis. *Journal of Marriage and the Family* **59**: 50–61

Kahn J R 1988 Immigrant selectivity and fertility adaptation in the United States. *Social Forces* **67**: 108–28

Kahn J R 1994 Immigrant and native fertility during the 1980s: Adaptation and expectations for the future. *International Migration Review* **28**(3): 501–19

King M, Ruggles S 1990 American immigration, fertility, and race suicide at the turn of the century. *Journal of Interdisciplinary History* **20**: 347–69

Massey D S 1995 The new immigration and ethnicity in the United States. *Population and Development Review* **21**: 631–52

Okun B S 1997 Innovation and adaptation in fertility transition: Jewish immigrants to Israel from Muslim North Africa and the Middle East. *Population Studies* **51**: 317–35

Singley S G, Landale N S 1998 Incorporating origin and process in migration-fertility frameworks: The case of Puerto Rican women. *Social Forces* **76**(4): 437–64

Smith J, Edmonston B (eds.) 1997 *The New Americans: Economic, Demographic, and Fiscal Effects of Immigration*. National Academy Press, Washington, DC

Stephen E H, Bean F D 1992 Assimilation, disruption and the fertility of Mexican origin women in the United States. *International Migration Review* **26**: 67–88

Weeks J R, Ruben G, Rumbaut C B, Korenbrot C C, Minkler D 1989 High fertility among Indochinese Refugees. *Public Health Reports* **104**: 143–50

C. G. Swicegood and F. D. Bean

Immigrants: Economic Performance

Five issues have dominated labor market research about immigrants: (a) the determinants of immigrants' economic gains, (b) whether native-born residents gain or lose, (c) the changing labor market wages and skills of immigrants, (d) life-cycle assimilation of immigrants, and (e) generational progress. Despite extensive research, they remain controversial. The vast bulk of existing empirical research describes the American experience, but according to Zimmerman (1995) the less plentiful research on European migration does suggest that the wage and employment effects of immigration on European natives are similar to the USA. Consequently, while the conceptual issues that I discuss transcend national boundaries, I will rely in this article on empirical research findings from the USA.

1. The Gains of Immigrants

Ex ante, immigrants gain from immigration, or else they would not have moved. Yet, some basic questions about the selectivity of immigrants, the magnitude and determinants of economic gains from immigration, and the transferability of skills across national borders remain unresolved. The number of migrants and the gains from immigration depend on economic and political conditions in sending and destination countries, the costs of migration, and legal limits placed on numbers and types of migrants.

Individuals move if the gain is positive and stay at home if it is negative. The benefits from immigration

are related to differences in skill prices and moving costs across countries and the set of legal and administrative rules governing immigration. If the costs of moving are low (say geographically adjacent countries or more similar cultures), migration flows will be larger and less skilled migrants will find the move profitable. Consequently, less proximate countries will send the fewest immigrants but immigrants with the highest skills. Similarly, a place such as the USA, a country with a high skill price, should attract immigrants from countries with low skill prices (such as Mexico and the Philippines). But those immigrants who do come to the USA from countries with high skill prices (such as Western Europe and Canada) must be relatively high skilled to justify the move. Finally, migration flows will be smaller when skills are less transferable across borders.

Legal immigrants to the USA are highly selective. In 1996, male legal immigrants in their last job in the country from which they came earned about $20,000, almost twice the average wage in the sending country. The economic gains from legal immigration to the USA are also large and highly variable (Jasso et al. 2000a). The average gain exceeded $10,000 in the first year and was much higher among employment visa immigrants than those who came for marriage or family reunification. Immigrants with high incomes in the USA relative to their skills were more likely to be from countries with high gross domestic product (GDP), low income inequality, and politically competitive systems (Borjas 1987). Finally, labor market skills are imperfectly transferable across borders, with skill transferability lowest for those with a different language and among those who migrate for family-related reasons.

2. *Gains to Domestic Residences*

Do native-born residents gain from immigration? Economic theory points to possible effects of immigration on employment and wages of native-born workers, trade in products with other countries, the growth rate of the economy, and prices people pay for goods and services. Under reasonably general assumptions, a receiving nation, overall, gains economically from immigration. By increasing the supply of labor in receiving countries, immigrants produce new goods and services, but are only paid the marginal value of the last good they produce. Marginal values are less than the total value of the new goods produced; the difference goes to receiving-country residents who must gain as a group. Immigration allows domestic labor to specialize in producing goods at which they are relatively more efficient, thereby breaking the rigid link between domestic consumption and domestic production. From this perspective, the effects of immigration are comparable to those of international trade.

Even when the economy as a whole gains, there are losers and winners among native residents. The gainers are the owners of productive factors complementary with immigrants and those who buy goods and services produced by immigrant labor. The losers are substitute native-born workers who compete with immigrants and whose wages will fall. Another group that suffers from new waves of immigrants are immigrants from earlier waves who are close substitutes in the labor market.

In the longer run, immigration can affect rates of economic growth only to the extent that immigrants arrive with a different skill mix than the native-born. To affect growth rates, this difference must persist over each new generation. If the children or grandchildren of immigrants come to look just like the native-born, all that immigration does is augment the population and the scale of the economy; it does not change per capita income at all. There is evidence that the skills of the descendants of immigrants converge over a few generations toward the national norm, suggesting that long-run impacts on economic growth are modest.

The literature summarized in Smith and Edmonston (1997) places the domestic gain to the US economy for all immigration since 1980 between one and ten billion dollars a year—a small but positive number relative to total GDP. The negative distributional impacts appear to be concentrated on domestic workers with less than a high school education. Perhaps a third of the 10 percentage point decline in the relative wage of those with less than high school education that was observed between 1980 and 1988 was due to immigration. This effect does not appear to be concentrated in the areas where immigrants live, but instead is dispersed across the USA. This dispersal comes about because competing native workers migrate out of or do not migrate to areas to which immigrants move.

3. *The Changing Labor Market Performance of Immigrants*

In the USA, the foreign-born on average have earned less than native-born workers, a wage gap that has been significantly widening in recent decades (Smith and Edmonston 1997). There also exists considerable economic heterogeneity among immigrants. Among men and women, 'recent immigrants' and those from Latin America have received the lowest wages. In 1970, the wage gap compared with native-born workers for immigrant men was 48 percent; by 1990 it had almost doubled to 83 percent. Similarly, the gap for Mexican immigrants rose from 65 percent in 1970 to 94 percent by 1990. This large and increasing immigrant wage gap can largely be attributed to expanding skill differences between immigrant and native-born workers. On average, immigrants come from countries where the average education, wages, and

skill are far below those in the USA. A shift in national origins of the foreign-born toward countries where education levels are lower, such as Mexico and Latin America, contributed to the expanding skill and wage differential. As this skill gap widened, the wage gap widened even more owing to the rising wage inequality in the USA. Rising wage inequality implies larger wage differences holding constant skill differences.

Since the mid-1980s, these trends have reversed as relative incomes and schooling of immigrants rose during the 1990s (Jasso et al. 2000b). This economic improvement is especially pronounced when comparisons are limited to legal immigrants. Many foreign-born respondents in surveys are not legal immigrants. For example, only 20 percent of Mexicans who reported that they immigrated since 1990 in the 1995 and 1996 Current Population Survey were legal immigrants (Passell 1999); the remainder were either nonimmigrants or illegal immigrants. Data on legal immigrants alone indicate that during most of the last 25 years their labor market quality has been as high as that of native-born workers and has been rising since the mid-1980s. Given the vast differences between legal and illegal immigrants, research must more clearly distinguish among types of foreign-born residents.

Much less research has been conducted on immigrant employment. Employment rates of immigrants are initially below those of native-born workers, and employment rates of recent immigrants have fallen relative to those of natives. However, immigrants catch up to natives relatively quickly, so that after some years in the USA their employment rates are similar to those of natives.

4. *Life-cycle Assimilation*

Economic assimilation over an immigrant's lifetime has been a source of considerable controversy, but some consensus is now being reached. Chiswick's (1978) original study using cross-sectional data from the 1970 Census indicated considerable catching up in the sense that immigrants who had been in the USA a longer time had lower wage gaps with native-born workers. The Borjas (1985) critique argued that this was an inappropriate comparison since more recent immigrants were members of different entry cohorts who may have been of lower labor-market quality. To address this question appropriately, it is necessary to follow immigrants over their lives in the receiving country. Since no large panel surveys of immigrants exist, researchers instead have tracked synthetic cohorts (defined by their year of immigration) across successive cross-sections such as the decennial US Censuses. Using this analytical technique, on average, male and female immigrants both appear to narrow their percentage wage gaps with native-born workers as their stay in the USA lengthens (Smith and Edmonston 1997). Over time, the percentage wage gap closes significantly for immigrants from Europe and Asia and modestly for others. But initial wage gaps at time of entry has been growing so the time it takes to reach wage parity with native-born workers will take longer. Mexican immigrants are the principal exception to this general rule as they essentially maintain their initial percentage wage gaps with native-born white workers. To date, there is no convincing answer to why Mexican immigrants do less well. One can be speculative about the role of language or the implications of the geographic closeness to country-of-origin, but there is little concrete evidence to document any explanation.

A difficulty common in all existing research is that the analytical technique of tracking across successive cross-sections is problematic if immigrant cohorts are not closed. An immigrant cohort can be depleted if immigrants subsequently emigrate. If emigrés are highly selective, wage trajectories obtained from cohort tracking will be biased. The problems caused by out-migration may not be trivial as about one-third of Mexican immigrants who said in 1970 that they arrived after 1965 had emigrated by 1980. Given the repeated back-and-forth movements between origin and destination countries, a second problem involves the ambiguity in the Census question on year of immigration. It is unclear whether respondents have in mind their first, last, or most salient entry when answering this question. Combined, these methodological problems imply that conclusions about economic assimilation of immigrants must be qualified. These problems will only be resolved by surveys that actually follow immigrants (including those who emigrate) over their full careers.

5. *Generational Assimilation*

The long-run economic achievements of immigrants have more to do with generational progress than life-cycle assimilation. There has been some concern expressed that generational assimilation has slowed, at least compared with earlier European immigrant experiences. However, recent research indicates that no decline exists in the degree of intergenerational assimilation in education or earnings since the 1950s, and that changes in rates of exogamy (marriage outside one's ethnic group) across generations appear stable over time (Card et al. 2000). Generational increases in exogamy are also related to the economic status of the immigrant father.

The conventional wisdom for Hispanics is more pessimistic. Cross-sectional comparisons between first-, second-, and third-generation Hispanics typically show a narrowing of schooling and income gaps between the first and second generations, but either retrogression or little progress between the second and

third generations. However, these patterns do not relate to generational issues because, in any cross-section, members of the second generation are not the children of current immigrants, nor are the third generation the direct descendants of the contemporary second generation. The demographically appropriate analysis aligns generations so that each generation was born 25–30 years later than the previous one. When this is done, Hispanic schooling and wage deficits with the native-born are uniformly smaller in the second generation than in the first and are lower still in the third generation (Smith 1999). Based on the historical record, second- and third-generation Hispanics have narrowed their economic and education disparities with native-born whites.

See also: Assimilation of Immigrants; Generations, Sociology of; Immigration: Consequences for Fiscal Developments in the Receiving Population; Immigration: Public Policy; Life Course: Sociological Aspects; Migration, Economics of; Migration to the United States: Gender Aspects

Bibliography

Borjas G J 1985 Assimilation, changes in cohort quality, and the earnings of immigrants. *Journal of Labor Economics* **3**(4): 463–89

Borjas G J 1987 Self-selection and the earnings of immigrants. *American Economic Review* **77**: 531–53

Card D, Di Nardo J, Estes E 2000 The more things change: immigrants and the children of immigrants in the 1940s, 1970s, and the 1990s. In: Borjas G J (ed.) *Issues in the Economics of Immigration*. University of Chicago Press, Chicago, pp. 227–70

Chiswick B 1978 The effect of Americanization on the earnings of foreign-born men. *Journal of Political Economy* **86**: 897–921

Jasso G, Rosenzweig M R, Smith J P 2000a. Determinants of immigrants, economic gains from immigration. Unpublished manuscript

Jasso G, Rosenzweig M R, Smith J P 2000b The New Immigrant Survey Pilot (NIS-P): overview and new findings about U.S. legal immigrants at admission. *Demography* **37**: 127–38

Passel J 1999 Undocumented immigration to the United States: numbers, trends, and characteristics. In: Haines D W, Rosenblum K E (eds.) *Illegal Immigration to America: a Reference Handbook*. Greenwood Publishing Group, New Haven, CT, pp. 27–111

Smith J P 1999 Progress across the generations. Unpublished manuscript

Smith J P, Edmonston B 1997 *The New Americans: Economic, Demographic, and Fiscal Effects of Immigration*. Report of the National Research Council. National Academy Press, Washington, DC

Zimmermann K 1995 Tackling the European migration problem. *Journal of Economic Perspectives* **9**(2): 45–62

J. P. Smith

Immigration

Immigration is the permanent movement of people across states, seen from the perspective of the receiving (rather than sending) states. While a 'transfer of jurisdictions' (Zolberg 1981, p. 5) is thus constitutive of immigration, traditional theories of immigration have surprisingly ignored the state, focusing instead on the economic motivations and social ties that initiate and perpetuate migration flows. The traditional view can be found, for instance, in this encyclopedia's 1968 predecessor, which discussed 'migration' only under its 'social' and 'economic,' but not political aspects (Petersen 1968, Thomas 1968). Still a recent state-of-the-art review of the world's major 'migration systems' indirectly defended the traditionally non-political migration theories, in arguing that states were in principle incapable of controlling social-network based migration flows (Massey et al. 1998).

While the study of immigration continues to be a domain of economic sociologists and demographers, a major development in the past two decades has been a growing interest by political scientists and political sociologists in the central role of the state in soliciting, channeling, and absorbing immigration. This reflects not only a disciplinary evolution and renewed interest in the state across the social sciences, but some changing features of contemporary immigration itself, for which the state has become increasingly central. The classical notion of immigration was coined in the 'stateless' period before World War I, when some 48 million Europeans moved relatively unhindered by entry controls to the Americas and to Oceania. After a brief, restrictionist interlude in the inter-war period, the image of 'stateless' immigration was revived after World War II, when the classic settler nations abolished racially motivated restrictions on immigration, and when some northern European countries underwent large-scale labor migrations from the south. States already figured prominently for these flows, because a conscious change of policy (as in North America or Australia), bilateral agreements with the governments of sending countries (as in Germany or Switzerland), or a geopolitically motivated absention from controlling postcolonial flows (as in Britain or France) were necessary to initiate them. However, the involvement of states was invisible, because this immigration was wanted and considered unproblematic. A typical example of this period is France, where until the late 1960s immigration from the south proceeded largely spontaneous and outside the official channels of the *Office Nationale d'Immigration*, but was legitimized *ex post* by the generous distribution of residence and work permits (see Viet 1998, pp. 233–54).

The picture of 'stateless' immigration—though always a fiction—changed radically in the mid-1970s, when under the impact of a serious economic recession

European states shifted to zero-immigration policies, and when seemingly 'uncontrolled' illegal immigration across its southern land border galvanized a restrictionist turn in the USA. Now, 'states' began to matter, either as bulwarks against unwanted immigration (as in the rhetoric of political elites), or as sieves that chronically failed to live up to their restrictionist intentions (as in the account by some academics, like Cornelius et al. 1994). In this historical moment, when the central role of the state in international migration became problematic and—therefore—visible, political sociologists and political scientists discovered 'immigration,' suggesting models to explain states' immigration policies and how the latter derived from—or violated—their sovereignty (Sect. 1), and investigating the impact of immigration on the membership component of nation-states, which is the topic of a new literature on 'immigration and citizenship' (Sect. 2). Some demographers and economic sociologists, the traditional 'owners' of the immigration domain, still think that the relationship between state policies and international migration, while considered increasingly important, has remained 'under-theorized and little studied' (Massey et al. 1998, p. 286). As the remainder of this entry shall show, this view is mistaken.

1. *Explaining Immigration Policy*

A state's decision to admit immigrants, without which there is either no immigration or at best 'illegal' immigration, is formally a function of its sovereignty, as codified in international law, according to which states are free to reject or to expel non-citizens. This linkage between immigration policy and sovereignty is expressed in the standard notion, propagated by governments around the world, that immigration policy is 'national interest' policy (see US Commission on Immigration Reform 1997, p. 1). The problem is that, particularly in liberal states, the 'national interest' is no fixed thing, standing above group interests, and notoriously short of organizational manifestation. In a path-breaking essay, Gary Freeman (1995) explained in rational-choice terms why 'special' instead of 'national' interests shape the immigration policies of liberal states.

Because the costs of immigration are diffusely spread over the population, while its benefits are concentrated (for employers and co-ethnics), it follows from the logic of collective action that the beneficiaries of concentrated benefits win over the bearers of diffuse costs, in a constellation that Freeman calls (following Wilson 1980) 'client politics.' This simple model explains the paradox that liberal states, despite their restriction-minded elite rhetoric and public opinion, have remained largely open to (certain kinds of) immigrants. A number of authors have quickly picked up and revised or amended Freeman's model. In a cogent analysis, Jeannette Money (1999) has argued that because of the spatial concentration of immigration also the costs of immigration are concentrated, thus remedying the free-rider problem in these geographic areas, and giving rise to a constellation that J. Q. Wilson (1980) has called 'interest group politics.' The restrictionist impulse, first picked up by local politicians, can go nationwide if the respective district is crucial for maintaining or gaining a national electoral majority. Demonstrating this swing-vote dynamic in the cases of France, Great Britain, and Australia, Money's revision of Freeman's model helps explain why some states remain open to new immigration, while in others immigration control is put on the national political agenda. Joppke (1998b, 1999b) has argued that the logic of client politics is more firmly established in classic immigrant nations than in European states, where pro-immigrant lobbies are much weaker, and where nominal 'zero-immigration policies' fit more easily with the rhetoric of 'national interest' policy. In addition, he stressed (much like Hollifield 1992) that liberal norms, which invest immigrants with rights, are a second constraint on state policy, next to interest-group pressure. Reference to liberal norms helps explain why European states, despite the absence of an ethnic lobby there, have remained open to certain kinds of immigrants, most notably the family members of labor migrants.

The analysis of immigration policy is often linked to the larger question of the resilience or transformation of state sovereignty in an era of 'globalization' (see the contributions in Joppke 1998a, Part I). This debate has suffered from a loose definition of sovereignty. Stemming from International Relations theory, the notion of sovereignty refers to the formal last-instance authority of the state within 'its' territory. In this formal sense, even ineffective immigration controls say nothing about the presence or absence of state sovereignty over immigration. Instead, the crucial test is if actors other than the state have acquired formal last-instance authority over immigration matters.

With some important exceptions, such as the free movement regime within the European Union (which is unlikely to be emulated elsewhere), this test will almost always be negative. An altogether different question is whether the presence of unwanted migrants within their territories (such as illegal migrants or asylum-seekers) indicates a declining empirical capacity (as against formal authority) of states to control immigration. The gruesome effectiveness of western states to bring down the numbers of unwanted asylum-seekers in the 1990s through 'safe-third-country' rules, tightened legal procedures, and social deterrence measures, suggests that there is little evidence for this (see Guiraudon and Joppke 2001)

2. *Immigration and Citizenship*

A second line of new and innovative scholarship on immigration has opened up over the impact of

immigration on states as 'membership associations' (Brubaker 1992). A variety of authors have concluded from the upgrading of immigrant rights across liberal states that 'citizenship' as the unitary membership form and focus of collective identity in modern states has been devalued (Schuck 1984), if not set on a path of decline (Soysal 1994). At the same time, migrants' 'transnational' practices and 'diasporic' forms of association are said to increasingly bypass the nation-state, thus rendering irrelevant old models of one-directional integration and assimilation (Glick Schiller et al. 1992).

'Post-' or 'trans-national' approaches have underestimated the resilience and adaptability of the national citizenship construct (see Joppke 1999a). The judicially nonreviewable 'plenary power' of the American federal state over immigration matters has allowed the drastic 1990s curtailment of social welfare rights and benefits even for long-settled legal immigrants, thus showing the inherent vulnerability of non-citizens (see Schuck 1998). In Europe, immigration has triggered not a decline, but a transformation of citizenship. The continent that two centuries ago had been the pacesetter of the postfeudal, 'modern' principle of blood-based citizenship (*jus sanguinis*) is now undergoing a reverse trend toward territorial citizenship (*jus soli*), which had been precisely found wanting in the dawning era of nation-states. In the 1990s, all member states of the European Union (except Greece, Austria, and Luxembourg) have facilitated the access to citizenship, especially for second- and third-generation immigrants, granting them a right to citizenship either at birth or majority age (see Hansen 1998).

The de-ethnicization of citizenship in response to immigration goes along with an intransitive mode of immigrant integration in liberal states. According to this intransitive mode, immigrants are not the object of forced assimilation, as in the 'Americanization' campaign in the early twentieth century, but are to be allowed to be the subjects of freely willed integration. Curiously, immigration scholars (as well as policy makers) often stick to a naive and unquestioned notion of integration with regard to immigrants, while it is notoriously unclear what actually 'integrates' the vast non-immigrant 'rest' of modern societies. Nevertheless, against the backdrop of weakening nationalist semantics in western states, all that immigrants are expected of is adherence to the procedural values that constitute liberal-democratic states; they are not expected to become 'assimilated,' that is, to take on the substantive life forms of a particular nation-state. Nathan Glazer (1997) has expressed this new consensus, which sets a sharp counterpoint to the assimilation request in the early twentieth century, in the happy notion that 'we are all multiculturalists now.' This change of attitude on the part of liberal states is visible, for example, in Germany's new naturalization rules, issued in the early 1990s, according to which cultural assimilation was no longer an individually examined prerequisite for the discretionary granting of citizenship. Now, the applicant had a right to citizenship, after having legally resided in the territory for a certain period. The *de facto* multiculturalism of liberal states is also visible in the new 'Americanization' program suggested by the 1990s US Commission on Immigration Reform, according to which 'ethnic and religious diversity based on personal freedom is compatible with national unity' (1997, p. 25).

Ernest Gellner (1983) famously defined the principle of modern nationhood as the congruence of political and cultural boundaries. As carriers of legitimate cultural difference, immigrants are one (but only one) destroyer of nationhood thus defined, contributing to the general process of cultural pluralization in western states. However, this *de facto* multiculturalism is to be distinguished from the public recognition and the granting of special group rights to oppressed minorities, which is the content of stronger programs of *de jure* multiculturalism. Its perhaps most concise philosophical formulation (Kymlicka 1995) concedes that immigrants, by voluntarily leaving their society of origin, have 'waived' the right to have their homeland resurrected in the receiving society. Despite some exceptions (such as Australia or the Netherlands), liberal states have implicitly followed this normative reasoning, allowing immigrants as individuals to pursue their religious and cultural ways of life, but denying them special rights of representation or self-government as ethnic groups.

There is one interesting exception to the cultural laissez-faire of liberal states vis a vis immigrants: language. Most states continue to make competence in their official language(s) an individually tested condition of voluntary citizenship acquisition (naturalization). In fact, in some states—such as Germany—the removal of a cultural assimilation test has gone hand in hand with the introduction of tougher language requirements. It is not far-fetched to assume that old-style assimilationists have found in language an eagerly embraced substitute. However, the status of language as a prerequisite of citizenship is ambivalent. On the one hand, language may be related to the self-conception of a political community, and thus be elevated into a quasi-ethnic marker. This is clearly the case in the Canadian province of Quebec, which prefers francophone over all other immigrants, and forces the children of immigrants to attend French-language schools. On the other hand, language is also a functional necessity for states—the exercise of authority and the democratic process depend on agreeing on a common language. It is thus unlikely that a de-ethnicized immigrant integration will ever drop language as a key concern. Whether language plays more the role of ethnic marker or of functional necessity will depend on the context. But the cases of Quebec and Germany (concerned more than most

other states about the nation-building implications of immigration) suggest that in liberal states language obligations are the rather thin residuals of the obsolete project of cultural assimilation.

See also: Cultural Assimilation; Emigration: Consequences for Social Structure in the Sending Population; Ethnonationalism: Cultural Concerns; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Immigration and Migration: Cultural Concerns; Immigration: Consequences for Fiscal Developments in the Receiving Population; Immigration: Public Policy; Migration into the Americas and between American Countries; Multiculturalism; Pluralism; Transnational Citizenship: Cultural Concerns

Bibliography

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Cornelius W, Martin P L, Hollifield S (eds.) 1994 *Controlling Immigration*. Stanford University Press, Stanford, CA

Gellner E 1983 *Nations and Nationalism*. Cornell University Press, Ithaca, NY

Glazer N 1997 *We are all Multiculturalists Now*. Harvard University Press, Cambridge, MA

Glick Schiller N, Basch N, Blanc-Szanton L 1992 Transnationalism. *Annals of the New York Academy of Sciences* **645**: 1–24

Guiraudon V, Joppke C (eds.) 2001 *Controlling a New Migration World*. Routledge, London

Freeman G 1995 Modes of immigration politics in liberal democratic states. *International Migration Review* **29**(4): 881–913

Hansen R 1998 A European citizenship or a Europe of citizens? *Journal of Ethnic and Migration Studies* **24**(4): 751–68

Hollifield J 1992 *Immigrants, Markets, and States*. Harvard University Press, Cambridge, MA

Joppke C (ed.) 1998a *Challenge to the Nation-State: Immigration in Western Europe and the United States*. Oxford University Press, Oxford, UK

Joppke C 1998b Why liberal states accept unwanted immigration. *World Politics* **50**(2): 266–93

Joppke C 1999a How immigration is changing citizenship: A comparative view. *Ethnic and Racial Studies* **22**(4): 629–52

Joppke C 1999b *Immigration and the Nation-State: The United States, Germany, and Great Britain*. Oxford University Press, Oxford, UK

Kymlicka W 1995 *Multicultural Citizenship*. Oxford University Press, Oxford, UK

Massey D et al 1998 *Worlds in Motion*. Cornell University Press, Ithaca, NY

Money J 1999 *Fences and Neighbors: The Political Geography of Immigration Control*. Cornell University Press, Ithaca, NY

Petersen W 1968 Migration, social aspects. In: Sill D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York

Schuck P 1984 Membership in the liberal polity: The devaluation of American citizenship. In: Brubaker R (ed.) *Immigration and the Politics of Citizenship in Europe and North America*. University Press of America, Lanham, MD

Schuck P 1998 The re-evaluation of American citizenship. In: Joppke C (ed.) *Challenge to the Nation-State: Immigration in Western Europe and the United States*. Oxford University Press, Oxford, UK

Soysal Y 1994 *Limits of Citizenship: Migrants and Postnational Membership in Europe*. University of Chicago Press, Chicago

Thomas B 1968 Migration, economic aspects. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York

US Commission on Immigration Reform. 1997 *Becoming an American: Immigration and Immigrant Policy*. Government Printing Office, Washington, DC

Viet V 1998 *La France immigrée*. Fayard, Paris, France

Wilson J Q 1980 *The Politics of Regulation*. Basic Books, New York

Zolberg A 1991 International migrations in political perspective. In: Mary Kritz, Li L L, Zlotnik H (eds.) *International Migration Systems: A Global Approach*. Clarendon, Oxford, UK

C. Joppke

Immigration and Migration: Cultural Concerns

In recent years, there has been a renewed interest in the study of human migration. During the last decades of the twentieth century, most major nation-states have seen the topic of immigration emerge as a significant issue with important public opinion, policy, and research implications. Indeed, with well over 100 million international migrants worldwide, plus an estimated 30 million refugees, immigration has become a topic of global concern (Suárez-Orozco 1996).

Migration, in Latin, *migrare*, meaning to 'change residence,' has been a defining feature in the making of humanity from our very emergence as a species in the African savanna. Social scientists have traditionally defined migration as the more or less permanent movement of people across space (Petersen 1968). In the language of the social sciences people 'emigrate' out of one location and become 'immigrants' in a new setting.

1. Immigration and Time

The definition of migration as the more or less permanent movement of people across space suggests a number of important cultural concerns. First is the matter of the relative permanence of immigrants in a new setting. For many, perhaps most, immigration represents a permanent move, for others it is a temporary state before eventually returning 'home.' A central feature of the great transatlantic immigration that took place between Europe and North and South America from the 1890s until the 1910s was the high proportion of people who returned to Europe. By

some accounts, well over a third of all the Europeans who came to the Americas went back 'home' (Moya 1998).

'Sojourners' represent yet another pattern where temporality defines immigration. They are the large numbers of immigrants who move for well-defined periods of time, often following a seasonal cycle to eventually return home. Large numbers of Mexican agricultural workers in California have historically followed the sojourner pattern (Cornelius 1992).

A third type is the constant shuttling back-and-forth that seems to define the lives of many new immigrants. In recent years, some scholars of immigration have argued that new transnational and global forces structure the journeys of immigrants in more complex ways than was previously seen (Basch et al. 1995). This research suggests that many immigrants are players (economically, politically, and culturally) both in their newly adopted lands *and* in their communities of origin—moving 'back and forth' in ways seldom seen in previous eras of large-scale immigration (Suárez-Orozco 1998).

1.1 Immigration and Space

The idea of immigration as movement across space also requires some elaboration. Immigration, at a minimum, involves a change in residency and a change in community. Over the years, scholars have concentrated on two major types of large-scale migration: 'internal migration' (within the confines of a nation-state) and 'international migration' (across international borders). While many scholars would argue that the large-scale movement of people *within* a nation-state is a phenomenon of a different order than the large-scale movement of people *across* international borders, the differences between these two broad types of migration are often quite blurred.

Internal migrants might share many characteristics with international migrants: many move from rural villages to urban centers, many experience linguistic and cultural discontinuities, and many face the same bureaucratic and legal restrictions international migrants do.

While much attention has been focused on international migration, most immigrants today are internal migrants staying within the confines of their nation-states—China, Egypt, and Brazil are countries that have experienced high levels of internal migration. Indeed, *contra* the impression that most international migrants head to the developed world (i.e., Europe and North America), most immigration today is an intracontinental (i.e., within Asia, within Africa, etc.) phenomenon.

2. The Causes of Immigration

Why do people migrate? Scholars of immigration have generally theorized patterns of migration flows in terms of economic forces, social factors, and cultural practices. Social scientists who study the economic causes of immigration have long-ago examined 'push' factors including unemployment, underemployment, and wage differentials (Dussel 1998). Changing cultural models about social standards and economic expectations have also been implicated in why people migrate (Moya 1998). 'Pull' factors include demand for foreign workers in various sectors of a host country's economy. In many settings, transnational labor recruiting networks deliver low-skilled foreign workers into jobs that native workers have come to culturally construct as unappealing. In Japan, for example, immigrant workers are sometimes called '3k workers' for the Japanese words for 'dirty, demanding, and dangerous' jobs (Tsuda 1996).

In other cases, immigrants become a vital part of the well-remunerated knowledge intensive sector of the economy. Nearly 40 percent of all engineers in the Silicon Valley of Northern California are immigrants. Indeed, in recent years, countries of immigration have witnessed an unprecedented pattern of economic mobility among new arrivals (Suárez-Orozco and Suárez-Orozco 2001).

Immigrants today tend to be highly heterogeneous—in terms of sociocultural and religious background, race and ethnicity, and educational background and skill. They defy easy categorizations. In the United States, for example, immigrants are at once among the most educated and skilled *and* least educated and skilled workers. Immigrants tend to be over-represented in the category of people with doctorates or, indeed, winners of the Nobel Prize—just as they are over-represented in the category of people without a high school diploma. Worldwide immigrants now tend to be much more 'bimodal' in their socioeconomic profile than ever before.

2.1 Transnational Networks

Scholars of immigration have long maintained that cultural and social practices can generate—and sustain—substantial migratory flows. People migrate because others—relatives, friends, and friends of friends—migrated before them. Indeed, the best predictor of who will migrate is who migrated before. Transnational family reunification continues to be a critical vector in immigration today. In the year 1996, 915,900 immigrants were formally admitted in the US. Among them, 594,604 were family-sponsored immigrants (Suárez-Orozco 1999). Since the early 1970s family reunification is one of the few formal ways to migrate into Europe (Suárez-Orozco 1994). A number of studies have examined how transnational migratory social chains can generate a powerful momentum of their own. Roger Waldinger (1997) has explored the sociocultural dynamics of transnational immigrant labor recruiting networks.

Immigrants find their kin and kith jobs. Waldinger's research in immigrant-rich Southern California suggests that immigrant social networks provide employers with a nearly unlimited supply of workers with little or no outlay. Transnational social networks provide connections to people able to fill jobs, at times even before the jobs are vacant. The social networks function optimally when they efficiently and economically generate quality information—about jobs, about applicants, about bosses.

2.2 Immigration and War

We have a reasonable understanding of how 'love' (family reunification) and 'work' drive immigration. On the other hand, the role of war and its relations to large-scale migratory flows has been generally neglected. Yet throughout history war and migration have been closely linked. The threat of labor shortages during World War II led to the creation of temporary labor recruiting efforts to attract much needed immigrant workers to the United States (Calavita 1992). The resultant 'bracero' program became a powerful force in building—alas, via family reunification—a Mexican migration momentum that eventually turned into the largest and most powerful immigration flow into the United States this century (Suárez-Orozco 1998).

In the aftermath of World War II, many of the major Northwestern European democracies such as Germany and Belgium developed 'guest worker programs' to recruit foreign workers—initially in southern Europe, and subsequently in the Maghreb region of North Africa and in Turkey (Suárez-Orozco 1994). These programs came to an end in the early 1970s. Yet family reunification and chain migration continued to bring immigrants from North Africa into Europe for years.

The Cold War both deterred immigration—because of strict Iron Curtain controls—and generated large population displacements. The robust Cuban diaspora in the United States can be traced more or less directly to the Cold War (Molyneux 1999). The low-intensity warfare in Central America during the 1980s generated a new wave of large-scale immigration. As a result, there are now well over a million Central American immigrants in the United States (Suárez-Orozco 1989). In the 1990s, the ongoing conflicts in Zimbabwe and Angola have generated large-scale migratory flows especially into South Africa.

Natural disasters have also displaced populations and started new migratory flows. The 1999 hurricanes, which devastated much of Central America, initiated significant flows of emigrants into North America.

3. Immigration and the State

Whether social scientists examine the case of internal or international migrants, there is a consensus that the apparatus of the nation-state is decidedly implicated in migratory processes: by both what the state does and by what it cannot do. States are, *inter alia*, in the business of regulating the movement of people—both internally and internationally. The right to leave a country, that is the right to emigrate, is a relatively recent phenomenon (Moya 1998).

Nation-states regulate, and carefully police, the inflow of international immigrants across borders. Large-scale international immigration is in significant ways the product of nation building—Argentina, Australia, and Israel come to mind as archetypal examples. Likewise, the reconfigurations of national boundaries have historically and contemporaneously generated large-scale migratory flows. The partition of British India into Pakistan and India was accompanied by one of the 'largest migrations in human history' (Petersen 1968, p. 290). More recently, the disintegration of the former Yugoslavia led to massive, mostly involuntary migratory movements.

In the area of international migration, nation-states generate policies designed to establish who is a 'legal' or 'illegal' immigrant, who is an 'asylum seeker,' a 'refugee,' and a 'temporary guest worker.' States regulate how many immigrants are legally admitted every year. In the US, for example, since 1990, an average of about a million legal immigrants have been admitted annually. On the other hand, legal immigration into Northwestern Europe was greatly curbed following the oil crisis of the early 1970s (Cornelius et al. 1994).

3.1 Asylum Seekers

States also regulate the flows of asylum seekers—those escaping a country due to a well-founded fear of persecution. Agents of the states decide who is formally admitted as a refugee. In the post-Cold War era, there has been an explosive growth in the numbers of asylum seekers worldwide. For example, some 369,000 foreigners requested asylum in Europe during the year 1998. Only a small portion of those seeking asylum are eventually granted formal refugee status.

In recent years, many postindustrial democracies—including the United States and throughout Northwestern Europe—have developed new strategies to deal with increasing numbers of asylum seekers (Suárez-Orozco 1994). For example, the 13,000 Kosovars that arrived in Germany in mid-1999 were given a three-month renewable 'Temporary Protective Status' on the condition that they not apply for refugee status—in effect, forfeiting all the rights and entitlements that come with formal refugee status. Similar arrangements were made for asylum seekers from Bosnia.

In the face of growing numbers of asylum seekers, and a widespread public concern that many of them

are 'economic refugees' in search of a better life in wealthier countries, new formal and informal strategies have been put into place in varied countries. Many of these new strategies seem to be designed to prevent asylum seekers from accessing 'safe countries'—where, under Geneva Convention agreements, they would have the right to a 'fair hearing.'

The high-seas interdiction program put into effect in the United States in the early 1990s is an example. The strategy was conceived to prevent large numbers of Caribbean (especially Haitian) asylum seekers form arriving in US territory—or even within its territorial waters—where they could establish certain legal protections. Apprehending them in international waters, and returning them to Haiti, leave asylum seekers with little practical recourse under international law. In Europe, a similar strategy has been to deem certain areas in international airports not part of the national territory—for example, parts of Zaventem airport are not technically Belgian territory but are considered to be international territory. Asylum seekers entering such airports have been turned back because they are said to remain in international territory, and hence not come under the jurisprudence of the Geneva Convention (Suárez-Orozco 1994).

While advanced postindustrial democracies are likely to continue to face significant numbers of asylum seekers, the greatest numbers of asylum seekers are in the developing world. For example, at the close of the millenium there are over three million asylum seekers in the African continent.

3.2 Illegal Immigration

While the state wields substantial power—both in regard to internal and international migration—there are nevertheless areas where it faces strict limitations in the management of human migratory flows. Nowhere are the limitations of state power more obvious than in its inability to control illegal immigration. In many parts of the world, undocumented or illegal immigration has become a more or less permanent problem, which periodically emerges as an unsettling political issue.

In the United States, for example, it is estimated that there are some 5–10 million illegal immigrants constituting roughly 20 percent of the total foreign-born (immigrant) population (National Research Council 1997). In Europe, the ratio of illegal immigrants to legal immigrants is a more carefully guarded secret because of its dangerous political connotations. Most hard-core right wing political parties in Europe, including France's *Front National*, Belgium's *Vlams Bloc*, and Austria's *Freedom Party* revolve around anti- (illegal) immigration platforms. In the 1990s, these once-marginal parties have made substantial gains with electorates quite concerned with the problem of undocumented immigration.

The enduring problem of illegal immigration in many parts of the world suggests that immigration is now structured by powerful economic factors, social forces, and cultural practices that seem impervious to state actions such as controls of international borders (Cornelius et al. 1994). Transnational labor recruiting networks, enduring wages differentials between nation states, and family reunification generate a powerful migratory momentum that is not easily contained by unilateral, or even multilateral, state interventions to curb it.

4. Economics of Immigration

Few topics are more controversial than the economic consequences of large-scale immigration. Do immigrants help or hurt the economies of host countries? Do immigrants carry their own weight or do they represent a burden to citizens and other established residents? Do complex postindustrial economies need low-skilled immigrant workers or have they become redundant? Much of the recent scholarship on immigration and the economy has tended to focus on such concerns as the fiscal implications of immigration, the issue of immigrant competition with native workers, and the related issue of immigration and wages. Another important theme has been the economic integration and progress of immigrants over time.

The research findings on the economic consequences of immigration are somewhat contradictory—some economists claiming that immigrants are a burden to tax payers and an overall negative influence especially on postindustrial economies (Huddle 1993) and others suggesting that they continue to be an important asset.

A recent study on the economic, demographic, and fiscal effects of immigration by the US National Research Council (NRC) concludes that in the American setting 'immigration produces net economic gains for domestic residents' (NRC 1997, p. 3). Not only do immigrants 'increase the supply of labor and help produce new goods and services' but their presence also 'allows domestic workers to be used more productively, specializing in producing goods at which they are relatively more efficient. Specialization in consumption also yields a gain' (NRC 1997, pp. 3–4). The NRC estimates that the immigration-related 'domestic gain may run on the order of $1 billion to $10 billion a year' (NRC 1997, p. 5). Given the size of the US economy (about 7 trillion dollars), it is clear that immigrants will neither 'make it' nor 'break it.'

In fiscal terms the NRC data suggest, 'Immigrants receive more in services than they pay in taxes' (NRC 1997, p. 7). The panel estimates that 'if the net fiscal impact of all U.S. immigrant-headed households were averaged across all native households the burden

would be … on the order of $166 to $226 per native household.' The NRC study an other studies conclude that while immigration is a plus in overall economic terms, low-skilled new immigrants have contributed to a modest drop in the minimum wage of low skilled workers. They found that a 5 percent drop in wages since 1980 among high school dropouts could be attributed to the new immigrants. There is, however, no evidence to suggest that new immigration has 'hurt' the economic condition of native minority workers such as African-Americans (NRC 1997, p. 5).

4.1 Immigrant Workers in the Global Economy

Other recent research engages the theoretical debate over the role of immigrant workers in the global, postindustrial economy. In the context of the increasingly advanced knowledge-intensive economies of today, are low-skilled immigrant workers simply anachronistic? Are immigrant workers a left over from an earlier era of production? Wayne Cornelius' comparative research of the use of immigrant labor in two paradigmatic postindustrial economic settings, San Diego County, California, USA, and Hamamatzu, Japan, suggests a remarkable convergence in patterns of growing reliance on immigrant labor—in spite of rather marked differences in national context (Cornelius 1998). Cornelius' data reveal a pattern of enduring, indeed voracious, postindustrial demand for immigrant labor. Cornelius concludes 'As immigrants become a preferred labor force, employers do more to retain them, even in a recessionary economy' (Cornelius 1998, p. 128). Cornelius argues that immigrant workers are desirable to a wide variety of employers for three basic reasons. First, immigrants are willing to do low-pay work that is boring, dirty, or dangerous with little or no prospects for upward mobility and that even in firms involving highly advanced technologies such work is critical. Second, employers perceive them quite favorably—as reliable, flexible, punctual, and willing to work overtime. Indeed, employers often prefer them to native-born workers. Third, immigrant transnational labor recruiting networks are a powerful method for 'delivering eager new recruits to the employer's doorstep with little or no effort on his part' (Cornelius 1998).

Other studies examine the issue of the socioeconomic progress made by immigrant workers. The research of Dowell Myers tracks, over time and across generations, various dimensions of the economic adaptations of immigrant-origin men in the state of California. His work explores three sequential outcomes: educational attainment, occupational mobility, and earnings. In some fundamental ways, the recent Mexican immigrant experience in Southern California seems to replicate earlier patterns of immigrant adaptation. Yet in other ways, Myers findings suggest new—and disturbing—patterns.

Myers' research reveals that upon arrival Mexican immigrant men tend to be poorly educated, work in low-skilled occupations, and earn low incomes. Myers finds that over time immigrant men make modest improvements in their economic condition. However, he also suggests that important changes occur across younger cohorts within the first generation. These changes, according to Myers, are strongly related to the much higher educational attainment of immigrant children. In other words, Myers finds an old story with a new set of characters: poorly educated immigrant men make modest gains over time but their children are able to attain more education in the new country. Still, Myers data reveal a disturbing new pattern: among the children of immigrants higher education 'does not appear to fully convert into higher occupational status or earnings; and higher occupational status translates even less well into higher earnings. These under-returns are most pronounced for the more recent arrivals from Mexico and for young cohorts, including native-born, both of whom newly entered the labor market in the 1970s and 1980s.' Myers concludes 'The social implications of these falling returns to education and occupation are regrettable, because the declining reward system may discourage other immigrant children from investing in schooling as the route for status mobility' (Myers 1998, p. 188).

5. Immigration and Cultural Transformation

In all countries of immigration there have been major debates surrounding the cultural consequences of large-scale population movements. In the US a palpable concern, not always fully articulated, relates to how the new immigrants, the majority of whom are non-English speaking, non-European, people of color migrating in large numbers from the developing world, will culturally adapt and, indeed, transform their new country. In Western Europe, there are similar concerns about the cultural adaptations of large numbers of immigrants coming into the Judeo-Christian continent from the Islamic world (Suárez-Orozco 1991). Likewise in Japan, a country where racial and cultural homogeneity is deeply implicated in the construction of cultural identity, there are questions about the long-term consequences of increasing migration from Thailand, Korea, China, the Philippines, and South America (Tsuda 1996). Large-scale immigration is both the cause and consequence of important cultural transformations. Immigration inevitably leads to cultural changes and accommodations among both new arrivals and native citizens (Ainslie 1998). Immigration can be said to be the consequence of cultural change in that new cultural tastes and changing

cultural conceptions of what is an acceptable standard of living have been implicated in large-scale migratory flows (Sassen 1988). Culturally immigrants not only significantly reshape the ethos of their new communities (Ainslie 1998, Gutierrez 1998) but they are also responsible for significant cultural transformations 'back home' (Durand 1998). As immigration researcher Peggy Levitt has argued, in many settings, immigrant 'social remittances' profoundly affect the values, cultural models, and social practices of those left behind (Levitt 1998). Because of a new ease of mass transportation and because of new communication technologies, it seems that immigration is no longer structured around the 'sharp break' with the country of origin that once characterized the trans-oceanic experience (Ainslie 1998). Immigrants today are more likely to be at once 'here' and 'there'—bridging increasingly unbounded national spaces (Basch et al. 1994) and in the process transform both home and host countries.

Studies of an earlier era of large-scale immigration in the US argued that those who abandoned their cultural practices in favor of quickly 'assimilating' to the uniform norms of the mainstream tended to display successful outcomes in terms of health, schooling of children, and socioeconomic mobility. More recent research suggests a somewhat more complex picture—where the long-term benefits of unambivalent assimilation may not be so obvious.

Various contemporary researchers stress that we can no longer assume that upon arrival immigrants face a uniform mainstream society. Rather, immigrants are most likely to interact in complex ways with many *other* immigrant and native minority groups. In the United States, increasing residential, labor, and school segregation means that many immigrants, especially poorer immigrants of color, are much more likely to interact with other ethnic minorities rather than with the white mainstream population (Orfield 1998).

For large numbers of immigrant children today, the social point of reference is not 'white,' mainstream, culture but rather the culture and social structure of ethnic minorities. In the words of the eminent scholar of immigration, Alejandro Portes:

> This puzzle is whether today's children of immigrants will follow their European predecessors and move steadily into the middle-class mainstream or whether, on the contrary, their ascent will be blocked and they will join children of earlier black and Puerto Rican migrants as part of an expanded multiethnic underclass (Portes 1994, p. 634).

Increasingly immigration will need to be framed in the context of powerful—and as of yet little understood—transnational formations. New patterns of capital flows, new information technologies, new patterns of communication, changing cultural expectations, and the ease and affordability of mass transportation are generating dynamics that transverse the traditional boundaries of the nation state. Global capitalism is increasingly characterized by 'borderless' economies predicated on transnational capital flows, newly opened markets, and immigrant-dependent economic niches. All of these factors would suggest that immigration is certain to remain a vital social phenomenon in the new millenium.

See also: Assimilation of Immigrants; Cultural Assimilation; Cultural Evolution: Overview; Cultural Geography; Immigration: Public Policy; Migration, Economics of; Migration History; Migration, Theory of

Bibliography

Ainslie R 1998 Cultural mourning, immigration, and engagement: vignettes from the Mexican experience. In: Suárez-Orozco M M (ed.) *Crossings: Mexican Immigration in Interdisciplinary Perspectives*. David Rockefeller Center for Latin American Studies and Harvard University Press, Cambridge MA

Basch L, Schiller N G, Blanc C S 1995 *Nations Unbound: Transnational Projects, Postcolonial Predicaments and Deterritorialized Nation-States*. Gordon and Breach Science Publishers, Basel, Switzerland

Borgas G 1999 *Heaven's Door: Immigration Policy and the American Economy*. Princeton University Press, Princeton, NJ

Calavita K 1992 *Inside the State: The Bracero Program, Immigration and the INS*. Routledge, New York

Cornelius W A 1992 From soujourners to settlers: The changing profile of Mexican immigration in the United States. In: Bustamante J A, Reynolds C W, Hinojosa R A Ojeda (eds.) *U.S.–Mexico Relations: Labor Market Interdependence*. Stanford University Press, Stanford, CA

Cornelius W A, Martin P L, Hollifield J F (eds.) 1994 *Controlling Immigration: A Global Perspective*. Stanford University Press, Stanford, CA

Durand J 1998 Migration and integration. In: Suárez-Orozco M M (ed.) *Crossings: Mexican Immigration in Interdisciplinary Perspectives*. David Rockefeller Center for Latin American Studies, Harvard University Press, Cambridge, MA

Dussel E P 1998 Recent structural changes in Mexico's economy: a preliminary analysis of some sources of Mexican migration to the United States. In: Suárez-Orozco M M (ed.) *Crossings: Mexican Immigration in Interdisciplinary Perspectives*. David Rockefeller Center for Latin American Studies and Harvard University Press, Cambridge, MA

Espenshade T (ed.) 1997 *Keys to Successful Immigration: Implications of the New Jersey Experience*. The Urban Institute Press, Washington, DC

Gutierrez D G 1998 Ethnic Mexicans and the transformation of 'American' social space: reflections on recent history. In: Suárez-Orozco M M (ed.) *Crossings: Mexican Immigration in Interdisciplinary Perspectives*. David Rockefeller Center for Latin American Studies, Harvard University Press, Cambridge, MA

Huddle D 1993 *The Costs of Immigration*. Released by Carrying Capacity Network, July, Dallas, TX
Levitt P 1998 Commentary. In: Suárez-Orozco M M (ed.) *Crossings: Mexican Immigration in Interdisciplinary Perspectives*. David Rockefeller Center for Latin American Studies, Harvard University Press, Cambridge, MA
Molyneux M 1999 The politics of the Cuban diaspora in the United States. In: Bulmer-Thomas V, Dunkerley J (eds.) *The United States & Latin America: The New Agenda*. David Rockefeller Center for Latin American Studies and Harvard University Press, Cambridge, MA
Moya J C 1998 *Cousins and Strangers: Spanish Immigrants in Buenos Aires, 1850–1930*. University of California Press, Berkeley, CA
Myers D 1998 Dimensions of economic adaption by Mexican-origin men. In: Súarez-Orozco M M (ed.) *Comings: Mexican Immigration in Interdisciplinary Perspectives*. David Rookeffer Center for Latin American Studies and Harvard University Press, Cambridge, MA
National Research Council 1997 *The New Americans: Economic, Demographic, and Fiscal Effects of Immigration*. National Academy Press, Washington, DC
Orfield G 1998 Commentary. In: Suárez-Orozco M M (ed.) *Crossings: Mexican Immigration in Interdisciplinary Perspectives*. David Rockefeller Center for Latin American Studies, Harvard University Press, Cambridge, MA
Petersen W 1968 Migration. In: Sills D L, Merton R K (eds.) *International Encyclopedia of the Social Sciences*. Macmillan, New York
Portes A 1994 Language and the second generation. *International Migration Review* **28**(Winter): Special issue
Sassen S 1988 *The Mobility of Labor and Capital*. Cambridge University Press, New York
Suárez-Orozco M M 1989 *Central American Refugees and U.S. High Schools: A Psychosocial Study of Motivation and Achievement*. Stanford University Press, Stanford, CA
Suárez-Orozco M M (ed.) 1991 Migration, minority status, and education: european dilemmas and responses in the 1990s. *Anthropology and Education Quarterly* **22**(2): Theme issue
Suárez-Orozco M M 1994 Anxious neighbors: Belgium and its immigrant minorities. In: Cornelius W A, Martin P L, Hollifield J F (eds.) *Controlling Immigration: A Global Perspective*. Stanford University Press, Stanford, CA
Suárez-Orozco M M 1996 Unwelcome mats. *Harvard Magazine* July-August, pp. 32–5
Suárez-Orozco M M (ed.) 1998 *Crossings: Mexican Immigration in Interdisciplinary Perspectives*. David Rockefeller Center for Latin American Studies and Harvard University Press, Cambridge MA
Suárez-Orozco M M 1999 Latin American immigration to the United States. In: Bulmer-Thomas V, Dunkerley J (eds.) *The United States & Latin America: The New Agenda*. David Rockefeller Center for Latin American Studies, Harvard University Press, Cambridge, MA
Suárez-Orozco C, Suárez-Orozco M M 2001 *Children of Immigration*. Harvard University Press, Cambridge, MA
Tsuda T 1996 Strangers in the ethnic homeland. Ph.D. dissertation, University of California at Berkeley
Waldinger R 1997 *Social Capital or Social Closures?—Immigrant Networks in the Labor Market*. Lewis Center for Regional Policy Studies, University of California, Los Angeles, Working Paper Series, 26, Los Angeles

M. M. Suárez-Orozco

Immigration: Consequences for Fiscal Developments in the Receiving Population

Immigration, particularly when it is from poorer to richer countries, often gives rise to concerns about economic impacts on the receiving country. One leading concern is that immigrants will impose a net fiscal burden on the nonimmigrant taxpayers by disproportionately drawing on public assistance, and, through their children, the educational system. At the same time, other people believe that immigration will alleviate population aging of industrial nations and bring them fiscal relief from the costs of providing health care and pensions to their growing populations of elderly. Most immigrants pay taxes as well as receive benefits, so it is the net effect that matters. In modern industrial states, immigrants are seen as imposing net costs through public education, poverty programs, and some health programs; and potential net benefits through programs for providing health and pensions for the elderly, with ambiguous effects on net health costs in general. Do immigrants, on net, impose a burden on taxpayers, or provide relief? The answer may influence policy decisions on the numbers and characteristics of immigrants to be admitted.

1. *Sources of Fiscal Impacts*

The history of these concerns extends at least into the nineteenth century in the US, and perhaps farther in other countries. However, the potential fiscal impact of immigration grew larger in industrial nations over the course of the twentieth century, as government spending grew in relation to the sizes of their economies, and as new welfare programs were initiated and grew to take a larger share of these expanding government budgets.

Some of the fiscal effects of immigrants would arise in a similar way from adding any new member to the population, for example a newborn native (see *Childbearing, Externalities of*). Thus an immigrant increases the size of the population and increases the number of tax payers. The increase in the size of the population will increase the demand for social infrastructure and public capital of many kinds, such as roads, water supplies, sewer systems, police stations, fire departments, airports, and court systems. Either these must be expanded, requiring both capital outlays and increased current expenditures, or the quality of services delivered to the pre-existing population must decline. Given the nature of the tax system, these increased costs will be born by everybody, constituting a negative fiscal impact. We call these congestible services. For other services provided by the government, the quality of the service received per person is not degraded nor costs increased when the population

grows. The goods generating such services are called public goods, and national defense (and related expenses such as services for veterans) the leading example. A larger tax base reduces the cost to each individual taxpayer of paying taxes to provide such public goods, so immigration has a positive fiscal impact through this route. An additional member of the population also dilutes common ownership rights in any collectively owned goods such as public lands and forests, fisheries, or oil reserves, but also dilutes fixed financial responsibilities such as national debt and its servicing. The net effect can be positive or negative.

Other fiscal impacts arise because immigrants have different characteristics than natives. For example, immigrants have different demographic behavior: relatively high probability of returning to the source country, different fertility (in industrial countries, typically higher), and different mortality (controlling for ethnicity, typically lower). They have different cultural backgrounds (and may be less likely to put their elders in nursing homes, for example), and speak different languages (which may impede their absorption in the labor force or require special educational treatment). Immigrants tend to be in their young adult ages on arrival. Their educational attainment may be higher or lower on average than that of natives. Their incomes after arrival may be quite different than those of natives. These characteristics affect the taxes they pay, and the cost of the benefits that they and their children receive. The task is to assess these various impacts, and arrive at a quantitative evaluation.

Surveys or censuses often provide information on tax payments and benefits used by individuals or households, that can be identified as immigrants or immigrant-headed households. This information can be supplemented from the national income and product accounts and other sources to estimate public debt, defense expenditures, educational expenditures per pupil, and so on. From these building blocks, an estimate of fiscal impact can be constructed.

Although this may appear to be a straightforward accounting exercise, results depend heavily on just how the question is framed. For example, what is the net fiscal impact of: an immigrant in the current population? A currently arriving immigrant? A household headed by immigrants? An immigrant, over the life cycle? An immigrant, over the life cycle, together with all descendants? The biggest differences in approach are cross-sectional analysis vs. longitudinal, and analysis of immigrants only vs. immigrant households.

2. *Cross-sectional Analysis*

Immigrants tend to be in the younger working years at the time of arrival; relatively fewer children and elderly immigrate. People of working age typically pay far more in taxes than they receive in benefits. Consequently, if the focus is on more recent immigrants still in the working ages, or if the pace of immigration has been increasing so that most immigrants are recent arrivals, then the finding will generally be that immigrants are a net fiscal benefit, simply as a result of their particular age distribution. For example, in the US in 1994, the average immigrant in the population had a beneficial net fiscal impact of $1,800 per year (Lee and Miller 1997, p. 316–17). Many problems arise because this is a cross-sectional calculation based on a single year. Some of these immigrants arrived early in the twentieth century and still survive, while others arrived within the most recent year studied. Such immigrants may have characteristics that are very different from those arriving today, yet all are lumped together in the calculation. Furthermore, the immigrants will impose different costs or benefits as they age, and the representation of ages among the current immigrant stock may be misleading, particularly if immigration has fluctuated in the past. After arrival, immigrants will themselves age in the future, at which time taxes and benefits may be quite different. In addition, children pose special problems. Foreign born children are counted as immigrants, but they are relatively few. Often the children of immigrants born in the country of destination are not themselves considered to be immigrants, in which case their costs and benefits do not enter into this calculation. However, the presence of these children is a consequence of their parents' arrival as immigrants, and therefore is a part of their fiscal impact. Omitting these children (by counting only immigrants) imparts a positive (beneficial) bias to the estimated impact.

In an effort to take these children into account, some analysts focus on the net fiscal impact of households headed by immigrants (Garvey and Espenshade (1998) is an excellent example, and Akbari (1991) an interesting variation). These will include foreign-born children as well as those who are born in the destination country. Changing the unit of observation from the individual immigrant to the immigrant household has a profound effect on the results. In the US case, the +$1,800 impact per immigrant now becomes a −$400 per immigrant per year (Lee and Miller 1997, pp. 316–17). This estimate is still cross-sectional and has all the associated problems. In addition, although now the children are counted, there are new problems, as follows: In many countries, including the US, the children born in the destination country are considered to be citizens, not immigrants. Such children enter the calculation only so long as they remain in their immigrant parents' households, which is mainly during ages when they are receiving costly public education. Once they become financially independent and move out of their parents' households, they no longer enter into the calculation. This approach therefore misses their working stage, in which they pay more in taxes than they pay in benefits.

Consequently, the immigrant-household method is biased toward finding an adverse impact because it counts these children only when they are young and costly and not when they are older and net contributors. If we were to include the costs and benefits of the whole second generation for the US, the average impact per immigrant would rise from −$400 when only children under 20 are counted, to +$900. Even within the cross-sectional approach, these variations in definitions lead to wide variation in the results. This is not satisfactory.

Until a few years ago, almost all analyses of the fiscal impact of immigration were based on one of these two methods: the individual immigrant or the immigrant household. Rothman and Espenshade (1992, p. 411) review 17 US studies; of these, nine take the individual immigrants approach, and the remaining eight take the immigrant household (or family) approach. Vernez and McCarthy (1996, pp. 7–8) update this survey with a look at nine studies since 1992, most of which use the immigrants-only approach. All these studies are biased one way or the other due to the method used.

3. *Longitudinal Analysis*

Consider the purpose of this analysis. It is intended to address the concern that additional immigrants admitted today will impose net costs on the pre-existing population. From this perspective, it is clearly new immigrants that matter, not immigrants that arrived decades ago and that may have very different characteristics. It is clear that the fiscal impacts of admitting an incremental immigrant will not all be exerted immediately, but rather will unfold over the immigrant's lifetime, conditional on survival and on remaining in the country. It is also clear that children born to the immigrant, and their taxes and benefits throughout their own, and their children's lives, are all part of the fiscal impact unfolding over time. In sum, the approach must be longitudinal. In principle, a household-based approach could be used, provided that all individuals were tracked after leaving one household and entering a new household, for many decades or even centuries. However, in practice this is not feasible, and an individual approach is therefore preferable.

The first thoroughgoing and detailed application of the longitudinal approach in this context was Lee and Miller (1997), followed by Auerbach and Oreopoulos (1999) and Storesletten (2000). The first two take a partial equilibrium approach, in which many potential feedbacks through the economy are ignored. Storesletten takes a general equilibrium approach, incorporating the effects of immigrants on wages and interest rates, in a dynamic perfect foresight setting; however, he includes less demographic detail and does not include public goods. Auerbach and Oreopoulos use a generational accounting framework. The approach of Lee and Miller will now be briefly described. Although the results from these three studies may appear to differ, once comparable assumptions are made they are all roughly consistent.

Lee and Miller (1997) estimate taxes paid and the cost of benefits received for immigrants by age, educational attainment, and time since arrival, as well as by age and educational attainment for the children of immigrants, and separately for all others in the population. These age profiles are then projected forward in time for a long time, based on assumptions about future rates of labor productivity growth, interest rates, trajectories of other government expenditures, and the assumption that taxes are adjusted so that the government debt to GDP ratio does not exceed 0.8. This requires tax increases when population aging accelerates. They also project the child-bearing, survival and re-emigration of all immigrants and their descendants. Estimated impacts take into account effects through public goods, congestible goods, and collective wealth or debt, in addition to transfers through the welfare state. What initially appeared to be a straightforward accounting problem based on hard numbers has thus become a highly speculative and uncertain exercise in very long-term projection.

Results can be summarized in various ways. For example, the net present value (NPV) of the long stream of taxes minus cost of benefits due to an incremental immigrant can be calculated, perhaps separately for each age and level of education at arrival by country of origin, sex, or other attributes. For the US, the NPV averaged across the actual age–sex–education composition of the immigrant stream around 1994, and using a real interest rate of 3 percent was +$80,000. This comes to $2,400 per year. Multiplying by 1.4 million gross immigrants per year, and dividing by the US population of 260 million in 1994, this implies an impact per destination resident of only +$13 for all the immigrants arriving in a given year, which is small. It is useful to examine the results more closely by level of government, by characteristics of the immigrants, and by time horizon.

In the US, the federal government has the main responsibility for providing pensions and health care for the elderly. The state and local governments have the main responsibility for providing public education and some other services for children. Consequently, immigrants are highly beneficial to the federal budget, with an NPV of +$105,000 per immigrant in 1994, while they impact state and local governments negatively with an NPV of −$25,000, combining to the net value of $80,000. In other countries, with a different division of responsibilities, the situation would, of course, be different. But in the US, certain states and local areas that receive large numbers of immigrants suffer a substantial negative fiscal impact. The federal gains from immigration are spread evenly

across the residents of all states, whether those states receive immigrants or not. This leads to political tensions between local, state, and federal governments. Some states have sued the federal government for compensation, particularly for the costs of undocumented immigrants.

Results also vary by characteristics of the immigrants. As suggested by Storesletten (2000), governments wishing to achieve a positive fiscal impact could use estimates like these to inform their policy decisions. The NPV for immigrants and their descendants is highest for immigrants arriving between ages 15 and 25, peaking earlier for those with little education at arrival, and later for those with more. At any age, more highly educated immigrants have higher NPVs. For example, for arrivals at age 20, those with less than a high school education have an NPV of +$33,000; of exactly high school education, +$146,000; and for more than high school, +$288,000. Much of this is due to the contributions of descendants for arrivals before age 30 or so. Arrivals later in life who do generate descendants have low or negative NPVs, and elderly immigrants are particularly costly. For example, an immigrant arriving at age 70 has an NPV of between −$149,000 and −$225,000, depending on education.

Finally, results vary strongly by horizon. The NPVs abstract from this dependence on time. Immigrants are initially quite costly while they assimilate into the job market and while their US born children are young. It takes about 20 years before the federal benefits per year begin to outweigh the costs at the state and local level. Thereafter, the net effect is positive.

Suppose there are an additional 100,000 immigrants per year in perpetuity, with the same distribution of characteristics as in the current flow. The fiscal consequences of this scenario can be projected for each year in the future, with no need to calculate present values. This avoids many of the uncertainties of the projections many generations into the future. For the US this incremental perpetual stream of immigrants would initially impose a cost per US resident of about $10 per year for the first 20 years, and thereafter there would be a growing net benefit, reaching about $80 per person per year by 2050.

4. Final Reflections

The topic of fiscal impacts of immigration has generated heated political controversy in many countries. It is true that depending on governmental structures and the degree of decentralization, immigration may cause serious fiscal pressures in certain years in certain localities. However, the net effect for the average immigrant in a typical industrial country might well be positive, since most industrial countries currently have tax-based public pension systems and face severe budgetary stresses from population aging. However, the net effect would also depend on the level of national debt and on the proportion of the national budget spent on defense. Perhaps the most important message from the studies reported above, though, is that the net effects of immigrants are surprisingly small at the national level. They hardly appear sufficient to be an important consideration in forming immigration policy.

See also: Assimilation of Immigrants; Childbearing, Externalities of; Immigration; Immigration and Migration: Cultural Concerns; Internal Migration: Developing Countries; Internal Migration (Further Types): Industrialized Countries; Internal Migration (Rural–Urban): Industrialized Countries; Migration and Health; Migration, Economics of; Migration out of Europe; Migration: Sociological Aspects; Migration, Theory of; Population Pressure, Resources, and the Environment: Industrialized World; Welfare State

Bibliography

Akbari A H 1991 The public finance impact of immigrant population on host nations: Some Canadian evidence. *Social Science Quarterly* **72**: 334–46

Auerbach A J, Oreopoulos P 1999 Analyzing the fiscal impact of US immigration. *American Economic Review* **89**: 176–80

Bonin H, Raffelhuschen B, Walliser J 1999 *Can Immigration Alleviate the Demographic Burden*? Working paper of the Economic Research Policy Unit, Institute of Economics, University of Copenhagen

Garvey D L, Espenshade T J 1998 Fiscal impacts of New Jersey's immigrant and native households on state and local governments: A new approach and new estimates. In: Smith J P, Edmonston B (eds.) *The Immigration Debate*. National Academy Press, Washington, DC

Lee R D, Miller T W 1997 The life time fiscal impacts of immigrants and their descendants. Project on the economic demography of interage income reallocation, Demography, University of California, Berkeley. Draft of Chapter 7 for Smith J P, Edmonston B (eds.) *The New Americans. A Report of the National Academy of Sciences Panel on Economic and Demographic Impacts of Immigration*. National Academy Press, Washington DC

Rothman E S, Espenshade T J 1992 Fiscal impacts of immigration to the United States. *Population Index* **58**: 381–415

Smith J P, Edmonston B (eds.) 1997 *The New Americans. A Report of the National Academy of Sciences Panel on Economic and Demographic Impacts of Immigration*. National Academy Press, Washington DC

Storesletten K 2000 Sustaining fiscal policy through immigration. *Journal of Political Economy* **108**: 300–23

Vernez G, McCarthy K F 1996 *The Costs of Immigration to Taxpayers: Analytical and Policy Issues*. RAND, Santa Monica

R. D. Lee

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Immigration: Public Policy

A common feature of twentieth century life was the immigration line. It is considered quite normal for states to regulate movement across international borders and sometimes within them. Anyone expecting to cross an international border is expected to have a passport with all the necessary documentation. In many countries, an internal passport (or some sort of national registration document) is mandatory for everyday business and official transactions.

From a historical perspective, the issuance of passports and the regulation of international migration by states are modern phenomena (Torpey 2000, p. 7). Prior to the nineteenth century, international borders were only approximately known and rarely policed. The strict regulation of international migration followed from the creation of 'nation–states'—the very peculiar form of state organization that has proliferated around the globe over the past two centuries. In the twenty-first century, there is very likely to be a gradual crumbling of the rigid laws and restrictions about who can cross international boundaries and settle in other countries. In the closing decades of the twentieth century, quite a few countries liberalized their immigration laws to allow more immigration and more diversity among the immigrants admitted. The European Common Market allows citizens from any country in the association to move to (and work in) another. Although it is unlikely that there will be a complete end to immigration restrictions, the radical limitations on international movements that characterized much of the world in the mid-twentieth century are likely to fade.

1. Long Distance Migration in Historical Perspective

Long-distance migration is as old as humankind (Davis 1974, McNeill 1984). Although some views of human history paint a portrait of immobile communities, wedded to the land and local villages over the millennia only to be disrupted by modern civilizations and global capitalism, this is a very partial view. Local communities did persist over long stretches of human history, but the peoples who inhabited them were connected to frontier regions and cities through periodic waves of out-migration in times of population surplus. Moreover, local villages and communities were periodically subjected to conquest that led to displacement, enslavement, or absorption into dominant populations. Patterns of uneven population growth, environmental changes, and technological differences between populations inevitably made migration, voluntary and involuntary, a recurrent feature throughout history. These patterns of population movement, however, were rarely regulated by state policies.

Prior to the rise of nationalism in the eighteenth and nineteenth centuries, the primary forms of state organization in the world were empires and kingdoms ruled by absolutist monarchs. Empires were defined by their centers—the city where the monarch ruled. The boundaries of empires were, however, ill defined and probably oscillated with the power of the center to extract taxes and conscript labor from the periphery. The cities of empires were invariably multiethnic and drew peoples from a variety of distant locations. The primary imperative for the openness of cities to migrations was survival. With the exception of Japan, concentrations of populations inevitably led to high levels of mortality far in excess of natural increase. Cities required a continuous influx of labor from rural areas just to maintain their population. Moreover, workers for various urban roles, including artisans, traders, slaves, and soldiers, were often imported or recruited from distant, culturally distinct peoples. These multiethnic cities were not necessarily harmonious or even tolerant, but social and cultural antagonisms were usually held in check because of common interests or by political force.

During the nineteenth and twentieth centuries, the political map of the world was transformed with the dissolution of empires, the rise and fall of imperialism, and the emergence of many new states. In many cases, state formation was based on a claim of nationalism, which usually implied an ethnic homeland or a sense of belonging to a common people. Nationalism has proven to be an extremely potent ideology of state building in the modern era. Indeed, the boundaries of Europe were redrawn after World War One to give many nations their own state, a policy legitimized by the ideology of national self-determination.

Among the many problems created by nationalism was the uncertain political status of domestic minorities in modern 'nation states.' McNeill (1984, p. 17) notes that the 'barbarian ideal of an ethnically homogeneous nation is incompatible with the normal population dynamics of civilization.' According to McNeill, the European nineteenth-century ideology of nation building (based on a single people in one country) was only realized by the coincidence of rapid population growth and the incorporation of regional peoples into a national myth of a common language and culture. If other ethnic groups could not be absorbed in the national population, the alternatives were expulsion or the removal of the rights of citizenship. The logic of nationalism was expressed by Hobsbawm (1992, p. 134) as 'The homogeneous territorial nation could now be seen as a programme that could only be realized by barbarians, or at least by barbarian means.' Although there is no simple one-to-one correlation between nationalism and the tightening up of international boundaries to migration, the logic was similar. A country that wanted to preserve a homogeneous ethnic homeland would not want to open its borders to other peoples.

Just as nationalism was taking hold in Europe in the nineteenth century and beginning to spread around the globe, there were other forces that were expanding the numbers of the potential migrants and the feasibility of large scale, long distance population movements. The most important factors were the acceleration of population growth in many parts of the globe after 1750, opportunities for settlement in frontier societies, and the demand for labor in the emerging industrial economies. Population pressures, with growing numbers of people living at the margins of subsistence, were exacerbated with the commercialization of the agricultural economy and displacement of peasants from the land. In different countries, famines, persecution, and pogroms added to the reasons for exodus in the nineteenth century. These pressures combined with the cheapening costs of long distance travel increased the attractions of sparsely settled frontier areas, especially in the New World. All of these conditions led to massive waves of migrants, numbering in the millions, crossing the Atlantic (and the Pacific) over the later half of the nineteenth century.

In most areas of the New World, the initial reaction to the increasing numbers of immigrants in the nineteenth century was positive. Immigrants were thought to contribute to economic growth and the general concern was of too few, not too many newcomers. But as the decades passed, the numbers increased, and alarm bells began to ring.

2. *The Closing of the Borders of the United States*

The most open borders of the nineteenth century where those in the New World and Australasia. With the depopulation of the indigenous peoples through conquest and the spread of Old World diseases, the New World became the demographic and economic frontier that attracted long distance migrants from around the world, especially from Europe. The migration to the New World was monumental, both in its demographic size and the diversity of its origins. For the 75-year period from the mid-nineteenth century to the end of the first quarter of the twentieth century, almost 50 million Europeans went to the United States alone (Massey 1988). In spite of some frictions, immigrants were generally welcomed in the New World. Labor was scarce and the endless frontier needed to be settled. Land grants, ksubsidized passage, and labor recruitment were among the strategies used to induce migrants at various times during the eighteenth and nineteenth centuries.

Not everyone came voluntarily. For decades, Great Britain would send English and Irish 'convicts' to its colonies in North America, and after the American Revolution, to Australia. It is estimated that more than 10 million Africans were transported as slaves to the Caribbean and to North and South America (Klein 1999, pp. 208–11).

By the middle of the nineteenth century, some of the occasional fears and antagonisms against immigrants were beginning to coalesce into social movements. In the decades prior to the American Civil War, there was a substantial movement against Catholic immigrants, particularly from Ireland, in several American states. These antipathies crystallized in the 'Know Nothing Party' which in 1855 elected six governors and sent a number of representatives to Congress (Jones 1992, p. 134). The term preferred by party members was the American party, and their philosophy was simply that of 'Americanism,' which seemed to represent a fear of internal minorities because of their foreign connections (Higham 1988, p. 4). The Know Nothing movement collapsed before the Civil War, and did not have any lasting impact, except as a harbinger of the nativist trend that emerged in subsequent decades.

The first American effort to close the door to immigration was directed against Chinese on the West Coast in the 1870s (Saxton 1971). By 1882, the anti-Chinese coalition had become so strong that Congress passed, and the President Chester A. Arthur signed, a bill that was popularly known as the 'Chinese Exclusion Act' (Hutchinson 1981, pp. 77–84). Although the facts of immigration restriction are clear, the motivations for it are still debated. There is no doubt that the anti-Chinese sentiments (and later anti-Japanese) were thoroughly infused with racial ideology. Popular prejudices against Asians were openly expressed in newspapers and by most political leaders (Saxton 1971, Daniels 1977). The question is whether racism was the primary reason or just a convenient ideology for those who had genuine fears of economic competition with the new immigrants. In her theory of the 'split labor market,' Bonacich (1972, 1984) argued that much of the antagonism and discrimination against Asian immigrants by working class whites, who led the movement for immigration bars, was based on fears that Asian immigrants' willingness to work for very low wages undercut the incomes of white workers.

The movement to exclude Chinese (and other Asian) immigration to the United States was not a singular event. In his book, *The Great White Walls are Built*, Price (1974) describes how similar restrictive immigration laws were passed in Australia and Canada. Moreover, the restrictions on Asian immigration foreshadowed the movement to exclude immigration from Southern and Eastern Europe to the United States, which culminated with the 'national origins' quotas in the 1920s.

The movement to restrict European migration to the United States was a complex phenomenon that extended for several decades with strange political, economic, and ethnic alliances on both sides. The standard economic account posits capital and labor as the main protagonists with business and employer

groups advocating free immigration and workers arguing the opposite. There were, however, many other sides to the debate. Immigrant communities, particularly in big cities, were a strong political force against immigration restriction, while Congressional representatives from rural areas were generally opposed to open immigration.

Perhaps the most important force moving the United States toward limits on immigration was the rising tide of nativism—the fear of foreigners, which gradually became intertwined with racial ideology in the first two decades of the twentieth century. American nativism had deep roots in anti-Catholicism and a fear of foreign radicals, but the belief in the inherent superiority of the Anglo-Saxon 'race' became the dominant element of the ideology in the late nineteenth century (Higham 1988, Chap. 1). These beliefs and the link to immigration restriction had widespread support among many well-educated elites. The Immigration Restriction League, founded by young Harvard-educated Boston Brahmins in 1894, advocated a literacy test to slow the tide of immigration (Bernard 1980, p. 492). It was thought that a literacy test would reduce immigration from Southern and Eastern Europe, which was sending an 'alarming number of illiterates, paupers, criminals, and madmen who endangered American character and citizenship' (Higham 1988, p. 103).

For three decades, the battle over immigration restriction was waged in the courts of public opinion and in Congress. In 1910, the Dillingham Commission (a congressionally appointed commission named after Senator William P. Dillingham of Vermont) issued a 42-volume report, which assumed the racial inferiority of the new immigrants from Eastern and Southern Europe relative to the old stock immigrants from Northwestern Europe (Bernard 1980, p. 492). Social Darwinism and scientific racism were in full flower with many leading scholars warning against allowing further immigration of 'beaten members of beaten breeds' (Jones 1992, pp. 228–30).

When the passage of a literacy test in 1917 did not have the intended impact of slowing immigration from Southern and Eastern Europe, Congress passed the Quota Act in 1921 to limit the number of annual immigrants from each country to three percent of the foreign-born of that nationality in the 1910 Census (Bernard 1980, pp. 492–3). These provisions were not strong enough for some restrictionists, who passed another immigration law in 1924 that pushed the quotas back to two percent of each nationality counted in the 1890 census, a date before the bulk of the new immigrants had arrived. The policy then led to the 'national origins quotas,' which were based on very dubious means of estimating the national origins of the American population (Higham 1988, pp. 316–24, Anderson 1988, pp. 140–9). There were no quotas allocated for Asian countries and no mention of any possible immigration from Africa.

Timmer and Williamson (1998) argue that the immigration restrictions that took hold in the United States and many other countries about the same time in the early decades of the twentieth century were primarily motivated by economicconsiderations and not by xenophobia or racism. There is some evidence that immigration from poorer areas might have slowed gains of domestic workers (Hatton and Williamson 1998; however, see Carter and Sutch 1998), and that the political alliances that did finally lead to immigration restriction were shaped, at least in part, by fears of competition with immigrant workers (Goldin 1994). Although restrictive policies were clearly formulated to address the fears of wage competition of American workers, the conclusion that this was the only causal variable seems to ignore the highly charged ideological climate of the era of immigration restriction.

Nationalism and racism were the reigning ideologies of the late nineteenth and early twentieth centuries. Although not every political outcome was determined by these ideologies, they did shape the immigration policy agenda in fundamental ways. In the United States, racial arguments were used by politicians, scholars, and the mass media to convince the American public and the government that the historically open door of immigration should be closed (Higham 1988). This was not an easy case to make in a country whose identity was that of a 'nation of immigrants.' It took several decades of overt anti-immigrant rhetoric and a broad political coalition before the US Congress was able to pass restrictive immigration legislation that closed the door to mass immigration.

Another factor that probably helped to turn the tide against an open system of international migration was the increasing rate of population growth. For most of the nineteenth century, migration had been a necessity because of the high mortality in cities and a general shortage of labor to settle the frontier and to work in the factories of the new industrial age. With declining levels of mortality in the early twentieth century, most countries were generally able to meet their labor needs from natural increase. In such circumstances, the nationalist and racial impulses were, perhaps, given a freer hand to regulate immigration policies over the middle decades of the twentieth century.

3. *The Immigration Door Begins to Open*

The imposition of the national origins quotas in the 1920s, followed by the Great Depression and World War Two, lowered immigration to its lowest levels since the early decades of the nineteenth century. The 'racial' character of the national origins quotas was exemplified by the very limited numbers of Jewish refugees fleeing Nazi Germany who were allowed to enter the United States while the quotas for Great Britain were not utilized. After World War Two, US immigration policies came into conflict with America's

new leadership role in the international system. If American political ideals were to influence other countries, the discriminatory character of the national origins quotas could be held up as an example of hypocrisy. In vetoing the McCarran–Walter Immigration Act of 1952 (which reaffirmed the national origins quota system), President Truman stated:

> The quota system—always based upon assumptions at variance with our American ideals—is long since out of date… The greatest vice of the present system, however, is it discriminates, deliberately and intentionally, against many of the peoples of the world… It is incredible to me that, in this year of 1952, we should be enacting into law such a slur on the patriotism, the capacity, and the decency of a large part of our citizenry (quoted in Keely 1979, pp. 17–18).

Congress overrode Truman's veto, and the national origins quota remained the law of the land for another 13 years. The domestic and international pressures for immigration reform continued to grow with each passing year. The dam finally broke with the landslide 1964 election, which brought a reform minded Congress into office. Among the major pieces of Great Society legislation passed was the 1965 Immigration Act, which was championed by a number of senior members of Congress who were the children of immigrants from Southern and Eastern Europe.

The 1965 Immigration Act replaced the national origins quota system with a new preference system based on the principles of family reunification and skills. In the decades following the 1965 Immigration Act, there have been a series of new laws that have modified the numerical limits and procedures of immigration and the admission of refugees (Smith and Edmonston 1997, pp. 22–30). In general, these reforms have liberalized immigration to expand the numbers and to create more possibilities for admission.

The sponsors of the 1965 Immigration Act were primarily interested in allowing a freer flow of immigration from Southern and Eastern Europe, the countries hit hardest by the national origins quotas. The primary response, however, has been a major immigration flow from Asia. The first wave of Asian immigrants in the late 1960s and early 1970s were able to utilize the provisions allowing for those in skilled occupations of high demand (nurses, engineers, doctors, etc.) to enter. These early arrivals were then able to use the family reunification criterion to sponsor their relatives. At about the same time, there was a major new wave of immigration from Latin America, especially from Mexico and Cuba.

The movement to a less restrictive policy of immigration to the United States has been paralleled by comparable reforms in other countries. In the early 1970s, Australia ended its 'White Australia' policy and allowed significant numbers of Asians to immigrate. In the early 1990s, the countries of the European Common Market loosened restrictions on interstate migration. Citizens of any country in the European Common Market can move to any other country and are free to seek employment or attend schooling on equal terms with natives of the country. The appearance of these common patterns in a number of countries and regions suggests that the nationalist impulse, which sought to limit and control international migration, was waning during the last few decades of the twentieth century.

Underlying the change in immigration policies were broad economic and demographic forces. Population growth in advanced industrial countries had slowed, and there was a shortage of native-born persons who were willing to work in lowly paid positions in the economy, including seasonal agricultural labor. For the second half of the twentieth century, there was virtually an unlimited supply of immigrant labor, sometimes skilled and always highly motivated, from developing countries. The prejudices against immigrants and nativist fears have not entirely disappeared, but their open expression has been sharply reduced in most modern postindustrial societies. These changes in economics, demography, labor demand, and ideology have contributed to a much freer flow of international labor migration in the late twentieth century (Castles and Miller 1998, Massey et al. 1998).

In terms of policies of international migration, two distinct patterns emerged in Europe and the United States. Many European countries adopted 'guest worker' programs, which were intended to be temporary stays by workers from the labor surplus countries of Southern Europe and North Africa. Guest workers were generally not eligible for citizenship and were expected to return home when their contracts were completed. The United States maintained a formal immigration framework, whereby legal immigrants could apply for citizenship after five years, but also tolerated a parallel system of 'illegal immigration.' Both policies were reluctant to acknowledge the realities of international migration. Most temporary workers did not return home, and they often brought their families and became permanent residents of the host society. The relative openness of the US society to illegal or 'undocumented immigration,' has led to the conclusion that efforts to regulate immigration are primarily symbolic. Understanding the 'failure' of immigration policies requires looking beyond mistaken assumptions and inadequate enforcement to see how deeply embedded international migration has become in the modern world economy.

4. *International Migration in the Twenty-first Century*

The contradictions between tightly regulated international borders and the modern world economy are becoming increasing clear. Most immigration policies, of whatever type, are residues of the first half of the twentieth century, when regulated borders were a

hallmark of modern statecraft. If the role of a state was to promote the welfare of the national population (a group defined by descent or membership), then a clear objective was to keep others (nonmembers of the nation) out, or to keep the numbers of immigrants to very modest levels. Policies of tightly regulated borders developed in nineteenth-century nation-building states, initially in Europe, and then spread around the globe in the twentieth century, including the traditional immigrant receiving societies in the New World and Oceania.

These policies 'worked' because domestic population growth in most countries was sufficient to meet labor demand. Indeed, population growth reached record levels everywhere in the twentieth century. Although rapid population growth created immense pressures in many labor surplus countries, there were few places that needed additional labor or allowed open migration. Passport controls were expensive and irksome to many, but they became accepted as normal features of modern states. Over the last few decades of the twentieth century, however, strains in the system of tight immigration policies were beginning to show.

The first sign was 'labor demand' in industrial countries that could not be met by domestic supply, at least not at the wages offered. Employers found it more desirable to import labor from abroad than to raise wages or to mechanize production. If this pattern were found in only one country or in only a few sectors, then it might be possible to consider a fairly narrow explanation in terms of political cultures or market rigidities. The demand for 'cheaper immigrant labor,' however, spans many sectors (agriculture, manufacturing, construction, repair services, restaurants, and childcare) in most industrial countries, including a growing number of rapidly growing developing countries. The increasingly global international economy seems to create recurrent needs for labor greater than that available from domestic population growth.

The demand for immigrant labor is not restricted to unskilled manual labor. The United States and other industrial countries have encountered a shortage of scientific and engineering workers, particularly in the high tech sector. This demand has been met, in part, by allowing many talented foreign students in American universities to convert their student visas to immigrant status. There has also been a gradual shift over the last few decades to more open immigration policies for a variety of reasons—refugees, agricultural workers, 'illegal' immigrants with long residences in the country, peoples in countries that have too few American citizen relatives to sponsor them, and workers in high demand by US employers.

These moves toward more liberal immigration policies in the United States are part of a broader international context with comparable patterns emerging in other countries. The policy of free movement of citizens in the European Common Market is the most striking example, but there are trends toward generous policies of admitting refugees and temporary workers in many parts of the world. There are even a few examples of more generous citizenship policies, but these are halting, often facing a domestic backlash from nationalist sentiments that have been weakened, but not disappeared.

More liberal immigration policies appear to be highly functional in modern industrial and post-industrial societies. Standard economic theory posits that domestic migration is a functional response to wage differentials between areas. Migration allows for workers to benefit from higher wages in growing areas and stimulates the economy to operate more efficiently by creating larger and more porous labor and consumer markets. Indeed the logic for lessening barriers to migration is similar to that of international free trade. Economic theory suggests that all countries benefit from the free flows of capital, goods, and technology across international borders. International migration is often excluded from discussions about expanding international trade (such as in the NAFTA debate), largely because of political considerations rather than economic theory.

Globalization is the most powerful trend in the world today. There are few places on Earth that are not exposed to the presence of the international forces of the mass media, multinational corporations, and Hollywood images. Every commodity from fresh food to electronic products moves around the globe in such profusion that most persons are unaware of the nationality of the producers of the goods (and services) they consume. Although international trade has always created competition between businesses and workers in different countries, the current era with instantaneous communication and cheap transportation has created a qualitatively new international community. In this setting, barriers to international labor mobility are an anachronism of the earlier era. Just as most countries, regardless of political ideology, have striven to make passport lines more efficient in recent times to encourage the very profitable tourist sector, it seems that most twenty-first century societies will ease immigration barriers in order to profit from the increasingly globalized world economy.

See also: Immigrants: Economic Performance; Immigration and Migration: Cultural Concerns; Immigration: Consequences for Fiscal Developments in the Receiving Population; Migration, Economics of; Migration into the Americas and between American Countries; Migration: Sociological Aspects; Racism, History of; Racism, Sociology of

Bibliography

Anderson M 1988 *The American Census: A Social History*. Yale University Press, New Haven, CT

Bernard W S 1980 Immigration: history of US policy. In: Thernstrom S (ed.) *Harvard Encyclopedia of American Ethnic Groups*. Harvard University Press, Cambridge, MA, pp. 486–95
Bonacich E 1972 A theory of ethnic antagonism: The split labor market. *American Sociological Review* **37**: 547–59
Bonacich E 1984 Asian labor in the development of California and Hawaii. In: Cheng L, Bonacich E (eds.) *Labor Immigration Under Capitalism*. University of California Press, Berkeley, CA, pp. 130–85
Carter S, Sutch R 1998 Historical background to current immigration issues. In: Smith J P, Edmonston B (eds.) *The Immigration Debate: Studies on the Economic, Demographic, and Fiscal Effects of Immigration*. National Research Council, Washington, DC, pp. 289–366
Castles S, Miller M J 1998 *The Age of Migration: International Population Movements in the Modern World*, 2nd edn. Guilford Press, New York
Davis K 1974 The migrations of human populations. *The Human Population* (A Scientific American Book). W. H. Freeman, San Francisco, pp. 53–65
Daniels R 1977 *The Politics of Prejudice: The Anti-Japanese Movement in California and the Struggle for Japanese Exclusion*. Atheneum, New York
Goldin C 1994 The political economy of immigration restriction in the U.S., 1890 to 1921. In: Goldin C, Libecap G (eds.) *The Regulated Economy: A Historical Approach to Political Economy*. University of Chicago Press, Chicago
Hatton T J, Williamson J G 1998 *The Age of Mass Migration: Causes and Economic Impact*. Oxford University Press, New York
Higham J 1988 *Strangers in the Land: Patterns of American Nativism 1860–1925*, 2nd edn. Rutgers University Press, New Brunswick, NJ
Hobsbawm E J 1992 *Nations and Nationalism Since 1780: Programme, Myth and Reality*, 2nd edn. Cambridge University Press, Cambridge, UK
Hutchinson E P 1981 *Legislative History of American Immigration Policy: 1798–1965*. University of Pennsylvania Press, Philadelphia, PA
Jones M A 1992 *American Immigration*, 2nd edn. University of Chicago Press, Chicago
Keely C 1979 *U.S. Immigration: A Policy Analysis*. The Population Council, New York
Klein H S 1999 *The Atlantic Slave Trade*. Cambridge University Press, Cambridge, UK
Massey D S 1988 Economic development and international migration in comparative perspective. *Population and Development Review* **14**: 383–413
Massey D S, Arnago J, Hugo G, Kouaouci A, Pellegrino A, Taylor E J 1998 *Worlds in Motion: Understanding International Migration at the End of the Millennium*. Clarendon Press, Oxford, UK
McNeil W H 1984 Human migration in historical perspective. *Population and Development Review* **10**: 1–18
Price C A 1974 *The Great White Walls Are Built: Restrictive Immigration to North America and Australasia 1836–1888*. Australian National University Press, Canberra, Australia
Saxton A 1971 *The Indispensable Enemy: Labor and the Anti-Chinese Movement in California*. University of California Press, Berkeley, CA
Smith J P, Edmonston B (eds.) 1997 *The New Americans: Economic, Demographic, and Fiscal Effects of Immigration*. National Academy Press, Washington, DC
Timmer A S, Williamson J G 1998 Immigration policy prior to the 1930s: Labor markets, policy interactions, and globalization backlash. *Population and Development Review* **24**(4): 739–71
Torpey J 2000 *The Invention of the Passport: Surveillance, Citizenship, and the State*. Cambridge University Press, Cambridge, UK

C. Hirschman

Imperialism, History of

1. Introduction: The Problem of a Definition

'Imperialism is not a word for scholars,' Sir Keith Hancock remarked a long time ago, and he was right (see Wesseling 1997, p. 74). Scholars have to make clear what they mean when they use certain concepts or terms, and therefore have to give definitions. This, however, is impossible with the word 'imperialism.' The problem is not that there are no definitions of imperialism, rather the contrary. There are about as many definitions of imperialism as there are authors who have written on the subject. They vary from those that refer to one specific form of imperialism, mostly Europe's nineteenth century colonial expansion, to others which give a very general meaning to the word, such as the one in Webster's Dictionary: 'any extension of power or authority or an advocacy of such extension.' Clearly, such a definition can cover almost any situation. Not surprisingly therefore, the word has often simply been used as an invective in order to criticize the policy of another country.

So defined, imperialism is useless as a scholarly concept. However, in serious studies, the word has always had a more limited meaning. The problem is exactly how limited its meaning should be. Sometimes the word is used in a universal historical way in order to characterize the politics of a dominant power. Thus, some historians have spoken of Roman or even Assyrian imperialism, but this is highly exceptional. In historical studies, imperialism generally refers to the policy of European countries, and primarily of the UK during the nineteenth and twentieth centuries, aiming at the expansion of their power and influence over other continents. It is in this context that the term imperialism originated and began to be used as a political and historical concept. Historically speaking, the word imperialism is therefore obviously closely associated with colonialism. While colonialism was only used to refer to one specific form of alien rule, namely the colonial one, imperialism acquired a wider meaning and included various other forms of influence over alien nations and states. For example, the financial influence of France and Germany in the Russian and Ottoman Empires, or such things as British 'gunboat policy' and American 'dollar diplomacy.'

After the end of the colonial empires the word 'colonialism' could only be used to refer to a phenomenon from the past and thus fell out of use. 'Imperialism' however continued to be used, and from then on also indicated those forms of domination that were formally different from, but factually comparable to, those formerly practiced by the colonial powers. For a while the word 'neocolonialism' was also used for this purpose, but somehow that term was less successful. By the end of the Second World War, America had become the new superpower. Accordingly, imperialism was now mainly applied to describe the foreign policy of the USA vis-à-vis other countries, in particular in Latin America, Asia, and Africa. There was also an attempt to make the concept applicable to the policy of the Soviet Union with regard to the Central and Eastern European countries that came under its influence after 1945 (Seton-Watson 1961), but this was not very successful. The reason for this is that historically speaking, imperialism has connotations with capitalism and not with communism, and with overseas possessions, and not with adjacent countries. Although there clearly was a Soviet Empire, it was not considered to be an example of imperialism but of traditional power politics. Only in its very general meaning as another word for all forms of power policies or simply as an invective, was it also used to describe communist countries such as the Soviet Union and China. After the end of the Cold War this use of the word imperialism lost much of its earlier attraction.

In this article imperialism is used in the sense of its initial meaning, that is to say as a term to indicate the extension of formal or informal, mostly European, rule over Asian and African countries in the late nineteenth and early twentieth centuries as well as, more generally, for some other forms of Western predominance during and after the colonial period.

2. *Imperialism: The History of a Concept*

Like 'colonialism,' which was probably first used in the title of a book of a French socialist critic of the phenomenon, Paul Louis' *Le Colonialisme* from 1905, 'imperialism' was originally a French word. It was from the 1830s onwards that the terms *impérialiste* and *impérialisme* came into use in France. They referred to the empire of Napoleon and to the imperial pretentions of his nephew Louis Napoleon, later known as Napoleon III. The colonial connotation only came after the word had begun to be used in the UK in the 1860s. Then, of course, the empire it referred to was no longer the continental one of France but the overseas empire of Great Britain (Koebner and Schmidt 1964).

Although the word imperialism was already used in the UK in the 1860s, the historical concept only appeared in 1902 with the publication of J. A. Hobson's *Imperialism. A Study* (Hobson 1938). Hobson, a radical but not a socialist, was deeply impressed by the South African War (1899–1902). In 1900 he published a book on this subject, *The War in South Africa. Its Causes and Effects*, in which he argued that power in South Africa had fallen into the hands of a small group of financiers 'chiefly German in origin and Jewish in race' (Hobson 1900, p. 189). In his famous book *Imperialism. A Study* he elaborated this vision into a general theory of imperialism, and used the term imperialism to indicate the 'expansion of Great Britain and of the chief continental Powers' (Hobson 1938, p. 27). The word expansion referred to the fact that over the previous thirty years a number of European nations, Great Britain first and foremost, had 'annexed or otherwise asserted political sway over vast portions of Africa and Asia, and over numerous islands in the Pacific and elsewhere' (Hobson 1938, p. 15). For Hobson the meaning of the word imperialism was very clear: it was the establishment of political control. He was also explicit about the forces behind it. Various people such as an 'ambitious statesman, a frontier soldier and an overzealous missionary' might play some role in it, 'but the final determination rests with the financial power' (Hobson 1938, p. 59). Thus Hobson offered us a definition (imperialism is the expansion of political power of European countries over the non-European world), a periodization (im perialism took place over the previous thirty years, thus between 1870 and 1900), and an explanation: it was the result of the workings of the financial powers. In order to explain their behavior, Hobson argued that, as a consequence of the capitalist system, the British economy suffered from underconsumption. As a result of this, surplus capital could no longer be invested profitably in England itself. Therefore, the capitalists were 'seeking foreign markets and foreign investments to take off the goods and capital they cannot sell or use at home' (Hobson 1938, p. 85).

As Hobson's theory implied a criticism of capitalism, it had a certain attraction for Marxist thinkers. As a result of this, a new Marxist theory of imperialism was born. While originally Marx and Engels had considered colonialism as an 'objective' progressive force (Avineri 1968), now Marxist theorists such as Karl Hilferding and Rosa Luxemburg scorned late nineteenth century imperialism as a form of exploitation and suppression. The Marxist theory of imperialism became very influential when it was appropriated by a man who was not only a theorist but also a practical politician, Lenin. In 1916 he published his famous brochure *Imperialism. The Highest Stage of Capitalism* (Lenin 1916).

Lenin's ideas were mostly based on the work of the previous-mentioned Marxist authors, who in turn had been inspired by Hobson's theory. It was therefore understandable that a direct link was seen between Hobson's and Lenin's theories, so much so that it became fashionable to speak of the 'Hobson–Lenin thesis.' There are, however, two important differences

between Hobson and Lenin. First, for Hobson the flight of capital from the metropolis to the overseas world was a consequence of the development of capitalism, but not a necessary consequence. The origin of the problem was underconsumption. Therefore, theoretically, it should also be possible to solve the problem by increasing the purchasing power of the working classes. Indeed, Hobson remarked: 'If the consuming public in this country [Great Britain] raised its standard of consumption to keep pace with every rise of productive powers, there could be no excess of goods or capital clamorous to use Imperialism in order to find markets' (Hobson 1938, p. 81).

Second, and more importantly, Hobson and Lenin tried to explain two different things. Hobson, who wrote his book during the South African War, wanted to explain the division of the world, and more specifically of Africa, in the late nineteenth century. Lenin, who wrote in 1916, tried to explain the redivision of the world of which the First World War was the most spectacular outcome. The word Africa hardly appears at all in Lenin's brochure. The period he referred to was also different from the one dealt with by Hobson: not 1870–1900 but thereafter. He wrote explicitly about this: 'I have tried to show in my pamphlet that it [imperialism] was born in 1898–1900, not earlier' (see Stokes 1969, p. 289). Thus Lenin parted ways with Kautsky and Luxemburg, for whom imperialism was little more than another word for colonialism (Stokes 1969, p. 297). For Lenin it was something else: not the highest stage of colonialism but of capitalism.

Although the capitalist theory of imperialism was not generally accepted, and alternative interpretations were launched and had some influence, some form of economic interpretation became the standard explanation of imperialism during the 1920s and 1930s. Imperialism was considered as having originated from economic problems in Europe that were characteristic of the late nineteenth century, in particular the need to guarantee the flow of raw materials to the industrialized countries, and the protection of overseas markets for the sale of their industrial products. This consensus broke down after the Second World War under the influence of decolonization and the rise of the American empire. The new world political situation also had an impact on the theory of imperialism. In a famous article, 'The Imperialism of Free Trade,' two Cambridge historians, Jack Gallagher and Ronald Robinson, developed the concept of 'informal empire' (Gallagher and Robinson 1953). They argued that the real zenith of the British Empire was not to be found in the late nineteenth century but rather in the mid-Victorian period of informal British economic hegemony. For Britain, the entire nineteenth century was one of expansion. It was an imperial century. Britain's imperial expansion manifested itself in various forms: emigration, trade, overseas investments, the establishment of naval bases, etc. The extension of political authority over foreign people was only one form of imperialism, and not even the most important one. The mid-Victorian empire was comparable to the informal American empire that came into being after 1945. It worked with informal means because that was the best way of doing things. The maxim of British policy makers was: informal empire if possible, formal only if necessary. Due to foreign competition and rivalry, however, the late Victorians were forced to formalize their Empire, and they did so, willy-nilly.

While Gallagher and Robinson discovered imperialism before empire, other theorists also discovered imperialism after empire. This resulted not so much from a reflection on the rise of the American empire, but from a reassessment of decolonization. While after the First World War the European powers had increased their territorial possessions—for example, by the division of parts of the Ottoman empire—and stabilized their colonial rule, the situation was very different after the Second World War. In Asia, the process of decolonization started immediately after the war and was followed later in Africa. Thus, in the 1960s, most of the former colonies had become independent politically. But political independence did not automatically bring an end to the social problems, nor to the economic dependency of the ex-colonies. Some of the new states became even more dependent on the Western-dominated world system than they had been before. For many observers it was clear that the end of empire was not at the same time also the end of imperialism. Some theorists worked this out in the theory of dependency. According to the *dependencianistas*, imperialism was not only the extension of political control, but it also included the dependency of less developed parts of the world on the industrial powers. Empire was only one form of imperialism, one stage in the history of Western dominance.

However, why one form of imperialism was replaced by another remained a question, an answer to which was also given by Gallagher and Robinson in their famous book on the partition of Africa: *Africa and the Victorians* (Gallagher and Robinson 1961). Here they argued that changes in the periphery, that is in the overseas world rather than in the mother countries, were responsible for the changes in the ways and means of imperialist control. Although *Africa and the Victorians* dealt primarily with British policy, the theories developed here had a wider meaning. While the theory of the imperialism of free trade was typically a theory about British imperialism, the peripheral theory was applicable to the imperialist activities of other nations as well. In many cases, changes in the non-Western world were decisive in determining imperialist action. Egypt's financial problems, for example, led to increasing foreign interference, and this, in turn, to a 'nationalistic,' or rather proto-nationalistic, reaction which plunged Egypt into an internal political crisis led again to foreign intervention and occupation. The discovery of minerals in South

Africa, to give another example, led to a complete change in the balance of power in that part of the continent.

Ronald Robinson later elaborated this interpretation into a more general theory based on the observation of the important role of the African and Asian partners of the imperialist rulers. In this so-called 'collaborationist theory,' imperialism is conceived of as a system of collaboration between European and non-European forces before, during, and after colonial rule. The changing forms of imperialism are considered as changing forms of collaboration that resulted from changes in the bargaining positions of the various parties (Robinson, in Owen and Sutcliffe 1972, pp. 117–42).

The Gallagher and Robinson theories were followed by a greater number of studies on the economic significance of the British Empire and the role of economic factors in British imperialism (Davis and Huttenback 1986, Cain and Hopkins 1993a, 1993b). The important place of Britain in the debate on imperialism is understandable, because Britain was the imperial power *par excellence*. But for that very reason Britain was not the most typical imperial power. Rather, it was atypical and therefore the discussions in other European countries on imperialism have followed different lines and focused on different questions. Chronologically speaking, however, the European revisionist theories were developed in the same years as the British: the debate started in the 1960s and continued well into the 1980s.

3. *National Articulations*

In France, Henri Brunschwig's *Mythes et Réalités de l'Impérialisme Colonial Français, 1871–1914*, which appeared in 1960, set the tone for the debate on French imperialism (Brunschwig 1960). According to Brunschwig, the causes of French imperialism were not to be found in economic demands but in the development of French nationalism after the defeat in the Franco-Prussian war of 1870. The protectionist factor was a myth, political factors were decisive. Given the specific intellectual climate that existed in France after the Second World War, and in which Marxism played such an important role, it was to be expected that Brunschwig's book would lead to great controversy, as it did. But the Marxists could hardly deny the fact that the French colonial empire had been of little economic importance in France. In order to rescue the Marxist theory of imperialism, they therefore argued that French imperialism was not to be found in the French colonies but elsewhere, in the Russian and Ottoman empires. They argued that French colonialism was not imperialist, and French imperialism not colonial (Bouvier and Girault 1976). In 1984, an important study by Jacques Marseille, based on an extensive data bank on French colonial trade, threw new light on the question of economic interest. His conclusion was that in the beginning the colonies were useful to French industry from an economic point of view, but subsequently they became a burden (Marseille 1984).

There was also a strong connection between imperialism and nationalism, but it is not altogether clear to what extent imperialism was a result of nationalism. This is because the decision to found a German colonial empire was very much the decision of one man, Chancellor Bismarck. Therefore, in Germany, the discussion on imperialism has always been concentrated on Bismarck and his motives. There were two main interpretations, a foreign political one (imperialism as a move in Germany's international relations) and one in terms of domestic policy, like electoral success, financial pressure groups, etc. The discussion was reopened when H.-U. Wehler (1969) added new elements to this debate. Although he stressed the economic background of imperialism, he agreed that, as in the case of France, the German colonial empire had not been very profitable. In his view, the link between economics and empire must be sought on a different level. He emphasized the social problems of the *Reich* (its lack of legitimation because of its creation *von oben*, by force) and considered Bismarck's bid for colonies as a shrewd political move intended both as part of a general, more-or-less anti-cyclical, economic policy, and of a social policy seeking to unite the Germans around issues of foreign policy, and thus to overcome internal tensions. Thus Wehler's emphasis was more on the domestic than on the diplomatic motives of German imperialism under Bismarck (Wehler 1969). Here, the debate on German imperialism touched upon a wider discussion, the one on the problem of continuity and discontinuity in German foreign policy, the so-called German *Sonderweg*, and the place of national socialism in German history.

Italian imperialism was also studied from a special perspective. It was not very successful during the classical period of imperialism but it continued during the interwar years, under the influence of fascism. The French historian Jean-Louis Miège has emphasized not only the demographic factor in Italian imperialism, but also its political and ideological dimensions—the nationalistic reaction to the loss of population as a consequence of emigration—comparing it in this respect to Spanish imperialism (Miège 1968). The interpretation of Portuguese imperialism was long dominated by Hammond's theory of an 'uneconomic,' that is to say a primarily nationalistic, form of imperialism. Gervase Clarence-Smith later challenged this view by making a strong case for an economic interpretation of Portuguese imperialism. He argued that economic motives went hand in hand with other ones such as missionary zeal, nationalism, and others (Clarence-Smith 1985).

The case of Belgium is very special, because in the nineteenth century, Belgium was an anticolonialist country, but in spite of this, it was eventually to acquire one of the biggest European colonies in Africa, the Belgian Congo (later called Zaire, now Congo again). That this happened was due to the extraordinary zeal, tenacity, ruse, and ruthlessness of one man, King Leopold II. Jean Stengers has analyzed the singular nature of the king's imperialism which was one of old-fashioned economic exploitation and in this respect inspired by the example of The Netherlands (Stengers, in Owen and Sutcliffe 1972, pp. 248–76).

In the Netherlands the historical discussion on imperialism began rather late. The most important contribution to the debate came from a book by Maarten Kuitenbrouwer (Kuitenbrouwer 1991). In this he argued that the Dutch case was roughly analogous to others, and that the Netherlands followed more or less the general pattern. It has also been noted, however, that Dutch imperialism was defensive rather than offensive, reluctant rather than enthusiastic. In this respect it was comparable to that of Britain. In both cases there was more continuity than discontinuity, and what discontinuity there was, derived from a change in circumstances, not in policy (Wesseling 1997).

The historical debate on imperialism was mainly about the traditional colonial powers of Western Europe but observations have also been made about other countries. Russian imperialism poses interesting questions and offers paradoxical aspects because, on the one hand, Russia was an object of Western European financial imperialism, but on the other it was itself also acting as an expansionist power by extending its empire to the East and eventually to the shores of the Pacific (Geyer 1977, Le Donne 1997). The case of Japan is particularly interesting, because it is the only Asian nation which became an imperial power. Like other Asian countries it was first confronted with Western influence but it reacted in a very different way to this challenge. After having been forced to 'open' the country in 1853, it accepted Western notions and techniques very rapidly, so much so that already by the 1890s it had started its expansion into China. Japanese imperialism was continued in the 1930s and of course during the Second World War. Some analysts have also considered Japan's economic expansion after 1945 as a form of informal imperialism (Mommsen and Osterhammel, 1986, pp. 53–82).

The concept of American imperialism is a very complicated one. Of course America has been considered as the main imperialist power since 1945, but how this was related to its earlier expansion is unclear. While some authors consider the Russian expansion to the East as a form of imperialism, the American conquest of the West has hardly ever been interpreted in this way. Traditionally, one has observed that American imperialism only came about at the turn of the twentieth century, with the Spanish–American war of 1898 and the American take-over of Cuba and the Philippines from Spain.

3.1 *Explanation: Motives and Means*

The rich literature on imperialism that has been published from the 1960s onwards has led to a revision of the traditional views on the origins and meaning of late nineteenth century imperialism. Transformations in Europe, but also in the overseas world, have received attention as factors that can explain the new imperialist attitude. The main distinction is between European interpretations on the one hand, which underline economic, political, strategic, and ideological motives, and peripheral interpretations which give special attention to activities and developments in the overseas world and in particular to the 'frontiers' of European influence. The new research has also given attention to such topics as the ecological aspects of imperialism (Crosby 1986), cultural imperialism (Said 1978, 1993), the impact of imperialism on the sciences (Petitjean et al. 1992), etc.

Much of the debate on imperialism concerned the motives of the imperialists. In order to understand the origins of imperialism, however, attention has also to be given to another aspect, not the motives but the means. The development of imperialism cannot be understood by looking only at transformations in Europe and the overseas world, and the incentives for imperialist actions that were created by these. What also was necessary for such action was the disposal of the necessary means (Headrick 1981, 1988). It had always been virtually impossible for Europeans to survive in the environmental conditions of tropical Africa. New developments in the medical sciences, such as the prophylactic use of quinine (as from the 1840s), made it possible for Europeans not only to live, but also to work and even to fight under such conditions. The development of new means of transportation (steamships, railways), the opening of new sea routes (like the one via the Suez Canal), the revolution in the means of communication (the telegraph, and later on the telephone and wireless communication) made the extension of imperial rule possible. Finally, but perhaps most importantly, the development of new weapons, and in particular of the machine gun, gave the Europeans an enormous advantage in their battles with non-European nations. Colonial wars became successful almost by definition and the European colonial armies became 'ever victorious' armies. Entire continents could be conquered at very small cost to the conquerors.

Thus the great technological superiority of the Europeans came into existence during the latter part of the nineteenth century due to the so-called Second Industrial Revolution which took place in Western

Europe and created rivals for British trade. These technological transformations not only offered the means for imperial expansion, but also led to new demands in the European societies which had their effects on foreign and colonial policy. Social and economic questions assumed increasing importance. State welfare provisions expanded. The import of tropical products at affordable prices was considered as a matter of public concern. This called for sustained economic exploitation which presupposed the existence of peace and order; in other words, effective authority.

The balance of power in Europe also changed dramatically in the 1870s. In the early nineteenth century, from Napoleon to Bismarck, Europe had found itself in an exceptional political situation. Germany and Italy did not yet exist. Britain had eliminated France as a maritime and colonial rival. Spain, Portugal, and the Netherlands, the old colonial powers, had had their day. Hence, Britain enjoyed *de facto* world supremacy, much as the USA was to do after the Second World War. All these factors were swept away in the 1870s. After its defeat in 1870, France sought compensation for its loss as a continental power by strengthening its overseas role. Germany and Italy, newcomers both, claimed a place under the sun.

Internal political factors also played a part. European governments were faced with a new phenomenon: they had to take the wishes of their electorate into account. Economic growth and social harmony became declared objectives of government policy. The Paris Commune of 1871 accentuated the danger of social revolution and hence the importance of social issues. The state was asked to do more things for more people. Conversely, technical progress, economic growth, and growing political involvement by the citizens of Europe created the conditions for a strong state. The military might of the European powers reached unprecedented heights. At the same time in the overseas world, transformations took place which often proceeded from previous contacts with Europe or European settlers, and which changed the existing internal balance of power. Thus in many respects a new situation came into being after 1870, in Europe as well as in the overseas world, and therefore, after all forms of revisionism, it is still justified to speak of the period of 1870–1914 as an age of imperialism, as Hobson did when he introduced the concept about a century ago.

See also: African Studies: History; Antiquity, History of; Capitalism; Colonialism, Anthropology of; Colonialism: Political Aspects; Colonization and Colonialism, History of; International Relations, History of; Spirit Possession, Anthropology of; Structuralism

Bibliography

Avineri S (ed.) 1968 *Karl Marx on Colonialism and Modernization* Doubleday, Garden City, New York

Bouvier J, Girault R 1976 *L'Impérialisme Français d'Avant 1914*. Mouton, Paris

Brunschwig H 1960 *Mythes et Réalités de l'Impérialisme Colonial Français, 1871–1914*. Colin, Paris [*The Myth of French Imperialism, 1870–1914*. Pall Mall Press, London, 1966]

Cain P J, Hopkins A G 1993a *British Imperialism. Innovation and Expansion, 1688–1914*. Longman, London

Cain P J, Hopkins A G 1993b *British Imperialism. Crisis and Deconstruction, 1914–1990*. Longman, London

Clarence-Smith G 1985 *The Third Portuguese Empire, 1825–1975. A Study in Economic Imperialism*. Manchester University Press, Manchester, UK

Crosby A W 1986 *Ecological Imperialism. The Biological Expansion of Europe, 900–1900*. Cambridge University Press, Cambridge, UK

Davis L E, Huttenback R A 1986 *Mammon and the Pursuit of Empire. The Political Economy of British Imperialism, 1860–1912*. Cambridge University Press, Cambridge, UK

Gallagher J, Robinson R E 1953 The imperialism of free Trade. *Economic History Review*, Second Series **6**: 1–15

Gallagher J, Robinson R E 1961 *Africa and the Victorians. The Official Mind of Imperialism*. Macmillan, London

Geyer D 1977 *Der russische Imperialismus. Studien über den Zusammenhang von innerer und auswärtiger Politik, 1860–1914*. Vandenhoeck & Ruprecht, Göttingen, Germany [*Russian Imperialism. The Interaction of Domestic and Foreign Policy, 1860–1914*. Berg, Leamington Spa, 1987]

Headrick D R 1981 *The Tools of Empire. Technology and European Imperialism in the Nineteenth Century*. Oxford University Press, New York

Headrick D R 1988 *The Tentacles of Progress. Technology Transfer in the Age of Imperialism, 1850–1914*. Oxford University Press, New York

Hobson J A 1900 *The War in South Africa. Its Causes and Effects*. Nisbet, London

Hobson J A 1938 *Imperialism. A Study*, 3rd rev. edn. Allen & Unwin, London

Koebner R, Schmidt H D 1964 *Imperialism. The Story and Significance of a Political Word, 1840–1960*. Cambridge University Press, Cambridge, UK

Kuitenbrouwer M 1991 *The Netherlands and the Rise of Modern Imperialism. Colonies and Foreign Policy, 1870–1902*. Berg, Oxford, UK

Le Donne J P 1997 *Russian Empire and the World, 1700–1917*. Oxford University Press, New York

Lenin V I 1916 *Imperialism. The Highest Stage of Capitalism*. Progress, Moscow

Marseille J 1984 *Empire Colonial et Capitalisme Français, Histoire d'un Divorce*. Michel, Paris

Miège J-L 1968 *L'Impérialisme Colonial Italien de 1870 à nos Jours*. Société d'édition d'enseignement supérieur, Paris

Mommsen W J, Osterhammel J (eds.) 1986 *Imperialism and After. Continuities and Discontinuities*. Allen & Unwin, London

Owen R, Sutcliffe B (eds.) 1972 *Studies in the Theory of Imperialism*. Longman, London

Petitjean P, Jani C, Moulin A M (eds.) 1992 *Science and Empires. Historical Studies about Scientific Developments and European Expansion*. Kluwer, Dordrecht, The Netherlands

Said E W 1978 *Orientalism*. Routledge & Kegan Paul, London
Said E W 1993 *Culture and Imperialism*. Chatto & Windus, London
Seton-Watson H 1961 *The New Imperialism*. Bodley Head, London
Stokes E 1969 Late nineteenth-century colonial expansion and the attack on the theory of economic imperialism: A case of mistaken identity. *Historical Journal* **12**: 285–301
Wehler H-U 1969 *Bismarck und der Imperialismus*. Kiepenheuer & Witsch, Cologne, Germany
Wesseling H L 1997 *Imperialism and Colonialism. Essays on the History of European Expansion*. Greenwood Press, Westport CT

H. L. Wesseling

Imperialism: Political Aspects

Imperialism is a form of international hierarchy in which one political unit, or polity, effectively governs or controls another polity. It is one of the oldest known political institutions, characterizing relations between peoples in ancient Mesopotamia, China, and Rome through modern Europe. It includes both rule within relatively contiguous areas—as in the Habsburg and Ottoman Empires—and the overseas colonies held by various European states after the age of discovery.

The term has a long and tortured history. It was apparently first used as an invective against the expansionist policies of Napoleon I, and has been employed most frequently to refer to the colonial practices of the European states in the late nineteenth and early twentieth centuries, attempts to revise the international territorial status quo, and the economic domination of one country by another (also referred to as 'neo-colonialism,' see below). Used as a tool of political rhetoric, the term is highly malleable and often devoid of any general meaning.

As an analytic concept, imperialism refers to the effective domination of one political community by another. According to Michael Doyle (1986, p. 19), '(e)mpires are relationships of political control imposed by some political societies over the effective sovereignty of other political societies … . Imperialism is the process of establishing and maintaining an empire.' By domination we mean the ability of the dominant polity, the metropole, to decide policy for the subordinate policy, the colony. This ability may entail substantial delegation of decision-making authority to elite members of the colony, but the metropole retains the power to decide what gets delegated and how, and when this authority is revoked.

Three corollaries are important. First, imperialism occurs only where distinct political communities exist. Subordinate communities may have a prior history of independence or a new political consciousness may emerge that creates a relationship of imperialism. Without the possession of a distinct political identity, however, class conflict or unequal political opportunities may exist but not imperialism. Second, the colony lacks an international political 'personality'; that is, while it possesses an identity as a distinct polity, it does not interact with other states as a sovereign equal. Finally, exploitation of the weak by the strong is not essential to imperialism, but it is an often natural outgrowth of effective domination. The affinity between domination and exploitation explains the typically pejorative status of the term.

Imperialism is an extreme form of international hierarchy in which the colony is, in principle, a subject of the dominant state. If imperialism forms one end of a continuum of international hierarchies, and 'anarchic' relations between sovereign equals forms the other, we can identity a range of increasingly hierarchical relationships. In spheres-of-influence, the subordinate members remain independent but are constrained by dominant powers from forming relationships such as alliances with other great powers. Latin America under the Monroe doctrine is a classic example. In protectorates, subordinate states yield control over their foreign and defense policies to dominant powers; although subordinates remain independent, they transfer control over specific areas of policy to other states. With continuing responsibility for their defense, the United States today retains protectorates over the Federated States of Micronesia and the Republic of the Marshall Islands. In informal empires, subordinates are subject to imperial states across wide ranges of policy but retain their international personalities and interact with third parties on the basis of sovereign equality. Eastern Europe under the Soviet Union is a particularly clear example. Protectorates and informal empires are often grouped with formal empires as forms of imperialism.

Neocolonialism is a hierarchy produced through the functioning of an impersonal international market. In this variant, the dominant state need not intend to control the subordinate but the latter is sufficiently dependent upon the former economically that it has little choice other than to comply with (and even anticipate) the metropole's desires. In this case, economic dependence produces political dominance—the core of imperialism—but the mechanism of control is indirect. For some, this is a virulent form of modern imperialism. For others, the absence of intent negates the political relationship. Neocolonialism remains a contested concept.

1. *Theories of Imperialism*

The major explanations for imperialism can be grouped into three general categories. Metrocentric theories focus on the dispositions or internal charac-

teristics of imperial states. Writing in 1902 (1965), for instance, John Hobson grounded the motivation for overseas expansion in the necessity for advanced capitalist states to export their surplus capital. This theme was later the foundation for V. I. Lenin's famous monograph, *Imperialism: The Highest Stage of Capitalism*, first published in 1917. Neo-Marxists later argued that the military–industrial complex and other features of capitalist states actually created a need for capital, leading states to create colonial and neocolonial relations with developing regions in order to extract wealth (see Magdoff 1969).

Pericentric theories emphasize conditions within the colonial polities. Where metrocentric theories focus on the push behind expansion, pericentric theories draw attention to the forces that pull imperialists into hierarchical relationships. John Gallagher and Ronald Robinson's *The Imperialism of Free Trade* (1953) set the direction for much research in this tradition. Positing a constant drive for domination and a preference for indirect rule wherever possible, Gallagher and Robinson explained variations in imperialist outcomes by conditions in the periphery. In particular, where peripheral polities possessed stable regimes and effective collaborators, they argued, imperialists could govern indirectly through informal empires. Only where the peripheral societies were unstable politically or lacked elites willing to protect their interests would metropoles be forced to create formal empires and govern directly. As it was effectively ruled by landed interests tied to the British market, in their view, Argentina escaped the need for recolonization but was nonetheless caught within Britain's free trade web. Later work on neocolonialism further developed these insights.

Systemic theories of imperialism, typically part of larger realist theories of international relations, highlight competition between the great powers (see Cohen 1973). The struggle for survival and influence between great powers creates an ever widening gyre of competition, in this perspective, that both leads metropoles to seize territories to augment their resources and allows them to compensate one another using peripheral territories to maintain an effective balance of power. The classic case of imperialism driven by systemic competition was the so-called race for Africa in the late nineteenth century.

The best explanations of imperialism have always combined ideas from more than one of these traditions. Even Hobson and Gallagher and Robinson, for instance, augmented their theories with a focus on systemic competition, which they both saw as accentuating the metropolitan or peripheral causes of imperialism. More recently, synthetic works integrating all three approaches have appeared that provide relatively complete explanations of imperialism (see Doyle 1986 Smith 1981). Although different authors emphasize different dimensions and different episodes, any single instance of imperialism carries traces of all three sets of factors. The 'state-of-the-art,' therefore, recognizes and builds upon metrocentric, pericentric, and systemic insights.

A fourth, nascent explanation is also being developed that draws upon neo-institutionalist theory, and especially theories of relational contracting as developed in economics. In this approach, imperialism is understood as simply another form of organizational hierarchy. Emphasis is placed on explaining why hierarchy, in general, and this form, in particular, is superior for obtaining the goals of the actors rather than alternative institutions, such as confederations, protectorates, and state-to-state relations. Three considerations are central. First, the larger the benefits from pooling resources and efforts between the units, in this case the metropole and colony, the more risks or costs the units are willing to accept in building a relationship. Second, the greater the expected costs of opportunistic behavior by the subordinate polity, the more important it is to the dominant state to control the actions of its partner. Especially significant here are assets that are specific to the two parties and which would be costly to the metropole if they were withdrawn. In the absence of some specific assets, we would expect an 'arms-length' relationship between independent units, as the benefits of cooperation could then be captured through market exchange. Only where there are substantial assets that are specific to the two parties is imperialism necessary. Finally, there are costs to governing any relationship, and these are likely to escalate with greater hierarchy. Imperialism is typically costly both for the colony, which gives up valued freedom, and for the metropole, which must either bind its own hands to limit its ability to exploit its partner or use coercion to impose its rule. Thus, states avoid empire if control is either unnecessary or can be achieved in some less hierarchical fashion. In this approach, empire is most likely when there are large benefits from pooling resources between the units, highly specific assets at risk, and the costs of governance do not rise sharply with greater hierarchy.

Employing this approach in the case of classic imperialism, Jeffry Frieden (1994) argues that formal empires arose in the age of plantation agriculture and raw materials extraction as a means of preventing local elites from appropriating site specific assets in production: once foreign investors opened a copper mine, for instance, this site specific asset was then subject to expropriation by the local government. Knowing this, foreigners would not invest in such assets without the control made possible by the formal empire. As multinational corporations have increasingly taken the place of agricultural or raw materials investors, Frieden continues, formal empires have become obsolete. Multinational corporations possess firm specific assets such as technology or brand names that cannot be easily appropriated by host governments; a government can seize an axle assembly plant, for example, but it is of little value without access to

the multinational's worldwide production and marketing network. Thus, formal empire is less necessary to protect corporate assets. David A. Lake (1996) has applied similar arguments to the case of hierarchies motivated by national security needs. In its emphasis on comparative institutions, this approach brings an additional dimension to the study of empires.

2. *The End of Empire?*

Decolonization was one of the most significant events in modern world history. At the dawn of the twentieth century, most of the globe was ruled from Europe. By 1963, the overseas empires had seemingly evaporated. The strong norm against imperialism as affirmed in numerous United Nations declarations strongly suggests that the age of empire is over. Its imprint endures, however. In the periphery, imperial rule destroyed local structures of governance. In some areas, such as India, the foundations of parliamentary rule were erected in their place. In other regions, especially Africa, only weak state structures existed when independence arrived, producing political instability and continuing impediments to economic progress (see Jackson 1990). In the imperial states themselves, political elites underwent a wrenching process of accommodation to their now diminished international political and economic status (see Kahler 1984).

Although scholarly interest waned with decolonization, imperialism nonetheless remains a vibrant issue in the contemporary world. Weakly institutionalized groups within multinational states, such as the Chechens in Russia or the Tibetans in China, are pressing demands for independence, declaring in words and actions that the states within which they are embedded are empires. Other groups are rediscovering nationalist identities, like the Scots in Great Britain, and asserting new political rights—transforming intrastate politics into imperial politics. Even as the overseas empires have ended, a new age of imperial struggle is emerging.

This new age will be politically more complex than the old. The principle of national self-determination, articulated by President Woodrow Wilson at the Versailles peace conference, could be easily employed in the fight against overseas empires. Today, this principle directly contradicts contemporary notions of sovereignty as an inviolable whole. As demonstrated in the recent NATO intervention in Kosovo, Western nations abhor the political domination of the local Albanians by their Serb rulers but they cannot condone and even less promote the breakup of a sovereign state. As both old and new imperial relationships within multinational states become contested, politicians and analysts alike will need to rethink issues of political identity, statehood, and empire—and perhaps experiment with new forms of political hierarchy.

See also: Balance of Power, History of; Balance of Power: Political; Colonialism, Anthropology of; Colonialism: Political Aspects; Colonization and Colonialism, History of; Dependency Theory; Geopolitics; Imperialism, History of; International Relations, History of; Multinational Corporations; Nation-states, Nationalism, and Gender; Postcoloniality

Bibliography

Cohen B J 1973 *The Question of Imperialism: The Political Economy of Dominance and Dependence*. Basic Books, New York

Doyle M W 1986 *Empires*. Cornell University Press, Ithaca, NY

Frieden J A 1994 International investment and colonial control: a new interpretation. *International Organization* **48**: 559–93

Gallagher J, Robinson R 1953 The Imperialism of free trade. *Economic History Review* **6**: 1–15

Hobson J A 1902/1965 *Imperialism: A Study*. University of Michigan Press, Ann Arbor, MI

Jackson R H 1990 *Quasi-States: Sovereignty, International Relations and the Third World*. Cambridge University Press, New York

Kahler M 1984 *Decolonization in Britain and France: The Domestic Consequences of International Relations*. Princeton University Press, Princeton, NJ

Lake D A 1996 Anarchy, hierarchy, and the variety of international relations. *International Organization* **50**: 1–33

Magdoff H 1969 *The Age of Imperialism: The Economics of U.S. Foreign Policy*. 1st Modern Reader Edition. Monthly Review Press, New York

Smith T 1981 *The Pattern of Imperialism: The United States, Great Britain, and the Late-Industrializing World Since 1815*. Cambridge University Press, New York

D. A. Lake

Implementation: Political

Implementation studies are inevitably normative. In 1973, the field was created virtually overnight by Jeffery Pressman, and Aaron Wildavsky, in whose book 'Implementation' the guiding research questions were: 'how well was this authoritative mandate (law, regulation, program, official policy pronouncement) implemented?' and 'how might it have been better implemented?' Later researchers redefined the normative standard as achieving the values implicit in some mandate rather than in executing its prescriptive details. Then the guiding question became: 'how can a complex implementation process bring out the values implicit in, or improve upon the raw materials provided by, some authoritative mandate?' Such normative questions have provided an intellectual focus for a field unified by little else.

1. 'Implementation' Defined

For simplicity, I restrict discussion to legislative mandates. The raw materials are typically: (a) a legislative mandate; (b) some new budgetary resources and; (c) a designated responsible agency. In most circumstances these three assets are not enough. Other resources are needed. Mandates defining regulatory programs will often require compliance resources from businesses or ordinary citizens. Service-delivery programs will often need participation from citizens who perceive the availability of the program, are willing to make the effort to establish eligibility, and can manage to sustain continued participation. The responsible agency may also wish to involve relevant nonprofit agencies. Other program agencies with overlapping jurisdictions or responsibilities must sometimes be asked to cooperate. Overhead agencies must supply permission to recruit personnel, lease office space, enter into contracts, and so on. 'Implementation' is therefore a process of constructing an operational program out of both newly mandated and extant resources. 'Successful' implementation means that the result is at least minimally cost-effective and is produced without unacceptable delay.

2. Explaining Success and Failure

The early implementation literature, inspired by the troubles of Great Society programs, was focused on documenting and explaining implementation failure. The (partial) subtitle of 'Implementation' was '*How Great Expectations in Washington Are Dashed in Oakland; Or, Why It's Amazing that Federal Programs Work at All* …' The principal explanatory framework in the early literature highlighted the institutions and practices of American pluralist and federalist democracy that were unfriendly to implementation, e.g., the susceptibility of programs to continuing political conflict over their goals and resources, weak or nonexistent interagency and interjurisdictional coordinating systems, and cumbersome accountability procedures and heavy-handed overhead agencies. In this world of fragmented political power, the many resources required for program construction are controlled by different actors—from front-line regulatory enforcers and service professionals to personnel officials writing civil service job descriptions—with only partially overlapping interests and with highly variable abilities and motivations to induce cooperation among one another (Nakamura and Smallwood 1980). Furthermore, these actors normally—and often quite properly—have different priorities from those of the advocates for the new mandate. They have ongoing responsibilities, after all, some of which are poorly fulfilled as it is, given normal governmental resource constraints.

Consider these institutional problems in system terms. Constructing a complex operational system out of many, varied, and interdependent elements is vulnerable to failure from many quarters. Bureaucratic resistance, prolonged bargaining, political sabotage, weak technical capacity, and a host of other conditions could affect many of these elements (Bardach 1977). Failure seems overdetermined. Yet failure appears to be much less frequent than this model suggests. The model omits the possibility of adaptation, i.e., strategically sacrificing controversial or overly ambitious goals and creatively improvising with respect to means. It also omits momentum, the fact that success with respect to securing one system element helps to beget success with respect to others.

The capacity to plan and execute effective adaptation and momentum-building strategies often takes the form of committed political leaders, entrepreneurial program managers, clever legislative staff and their principals, and the like. This, at least, is the impression conveyed by many of the case-study reports of successful implementation. Although lacking a methodological and theoretical rationale, such case studies may nevertheless have illuminated an important phenomenon. The current theoretical challenge of greatest significance is to explicate the process of adaptation and to identify factors that increase the likelihood of success or failure. The corresponding empirical challenge is to identify the conditions under which skilled adaptive leadership is drawn into certain implementation processes but not into others.

3. Design Failures and Implementation Failures

The early literature usefully distinguished failures of implementation from three common design-stage mistakes that simply show up during the course of implementation and which might have been aggravated (or occasionally ameliorated) by the implementation process. First, design-stage expectations are usually too high. To survive the legislative process, a policy or program mandate will often have had its merits oversold and its limitations understated. Second, designers sometimes commit fundamental conceptual errors, e.g., supposing that price controls in competitive markets can be implemented without causing shortages. Third, they may choose an inappropriate policy instrument. Arguably, the programs studied by Pressman and Widavsky, concerning the subsidized creation of low-skill jobs in a poor city, represented failures of instrument selection. The chosen instrument was a subsidy to firms promising to hire from a general target population. With no easy way to enforce the promises, however, the implementing agency probed the prospective employer firm for months on end, trying to clarify what might count as a legitimate excuse for underperformance and trying to test the firm's good faith and resolve. A much better choice of instruments would have been an *ex-post-facto* wage subsidy for workers actually hired.

4. The Characteristics of Generic Governmental Instruments

In the literature 'implementation problems' also include those that beset most operational governmental programs. Although the early implementation literature focused on 'control' problems, the focus is now broader, encompassing the standard problems of public administration and public management, e.g., how to recruit good personnel and motivate quality performance, how to keep a focus on results rather than on personal or careerist objectives or simply on following procedures and staying out of trouble, how to protect the program from ideological opponents in the legislature or within the cadre of higher-level political appointees, how to build trust and communications channels within and across agencies, and how to create a culture of continuous improvement. The 'reinventing government' movement has been a modest source of ideas and inspiration about how to cope with general management problems.

But the distinctive contribution of implementation research regarding such problems has been to analyze the forms they take when expressed in generic governmental instruments (Salomon 1989). Two of the most important of these are regulation (e.g., sanitary code enforcement and electric utility rate-setting) and pure service-delivery (e.g., employment and training) either directly or through provider subsidies.

4.1 Regulation

Command-and-control (or 'social') regulation is typically a response to information failures—regarding unsafe bus maintenance, say, or filth in restaurant kitchens—which prevent consumers or employees from protecting themselves through either market action or the threat of lawsuits. Command-and-control programs typically have three main components: a set of standards, a cadre of inspectors and auditors to detect noncompliance, and a set of sanctions with which to threaten noncompliant firms or citizens. This sort of regulation is vulnerable to errors of under-regulation or over-regulation or to both simultaneously. Political pressures and scientific uncertainty can lead to overly stringent or overly lax standards. Standards can also be too prescriptive or too ambiguous. The enforcement process is vulnerable to budgetary strictures, over-concentration on symbolic targets, go-by-the-book rigidity, and in some cases corruption.

Economic regulation includes natural monopolies such as local telephone service and (former) oligopolies such as long-distance freight hauling. Most scholars observe that the regulators tend to get 'captured' by the more powerful firms within the regulated industry. Arguably, the regulators are as much captured by a sense of responsibility for the economic well-being of the industry they regulate, however, and allow 'well-being' to be interpreted through the self-interested lenses of the major firms.

Technological change and more realistic economic analysis have undermined the rationale for much traditional economic regulation, and many regulatory activities, along with a few agencies, have been abolished. A new addition, however, has been environmental regulation. Although it often uses command-and-control methods, it also counts as 'economic' regulation because implementers are concerned with how their decisions on standards and enforcement actions affect interfirm, interindustry, and inter-regional competitive balances.

4.2 Service Delivery

The chief implementation challenge in service delivery programs is preventing inequitable distribution to constituencies favored by politics or by the preferences of service professionals. Middle-class neighborhoods, for instance, tend to attract better teachers to their schools for both reasons. A second challenge is keeping service quality at a high level while being unable to rely on traditional market-based incentives such as the threat of consumer exit.

When government uses subsidies to enlist private providers, such as hospitals treating Medicaid patients or nonprofit organizations to provide vocational training, it typically relies on regulations and audits to attempt to ensure that the providers are targeting their services correctly, maintaining a minimal acceptable quality, and not engaging in fraud or abuse. Many of the problems of command-and-control regulation are thereby imported into service delivery programs.

Service programs aimed at 'deviant' populations such as drug abusers, criminal offenders, the mentally ill, and welfare recipients typically have a regulatory, paternalistic component as well as a pure service component. Implementing personnel usually have considerable discretion over how to balance the two components. Large-scale ideological trends, and more immediate political pressures, also affect the balance.

5. The Practitioner's View of Implementation

Because implementation studies are normative, their results can be helpful in guiding practice. The literature advises policy designers to: (a) 'map' the functional requirements of a projected operational system and how they are each to be obtained (Elmore 1979); (b) write 'scenarios' regarding possible implementation pitfalls en route to such a system; and (c) estimate the probability of 'compliance from target groups' whose cooperation is needed for critical tasks (Mazmanian and Sabatier 1983). Designers can also consider

institutional solutions, e.g., contracting with an organization in the private sector to take on the responsibility of starting and running a new AIDS clinic or building and operating a new jail.

Implementers can also view the experience of operational implementation, whether successful or otherwise, as a way to improve long-run policy design and to create a mandate of sorts for what they are doing. Consider some program domain such as welfare-to-work, and construe all the 'implementers,' from the Secretary of Health and Human Services, through the state and county senior managers, down to the lowest level case workers as a sort of network through which ideas and effort can flow and can become organized. These ideas concern both policy design and the choice of implementation instruments. From the mid-1980s through the mid-1990s, this network produced a good deal of lore, and a bit of experimental evidence, that a more work-oriented welfare policy might be effective and would not be unduly harmful to the poor. Counter-claims were also produced. But in the end, the experience of many implementers combined with conservative Republican political leadership and very widespread support in public opinion created the federal welfare reform legislation of 1996. The preferred implementation instruments were a high degree of devolution to the states with regard to means but rather sharp financial incentives to compel adherence to federal government objectives. Deficiencies in the legislation will no doubt be brought to light in the course of implementation and further change will occur.

In some sense, then, 'the policy system is learning,' or 'evolving,' and the implementation process is a part of that system (Goggin et al. 1990, Majone and Wildavsky 1979). To be sure, 'learning' and 'evolving' are implicitly favorable ways to characterize a policy change. Many observers would object to this characterization of welfare reform. But eschewing value-laden language altogether would denature a phenomenon like implementation, which is so intimately connected to democratic policy choice and to the challenges of creating public value for the citizenry.

See also: Administration in Organizations; Executive Branch, Government; Organizational Decision Making; Organizations, Sociology of; Planning, Administrative Organization of; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Regulation and Administration; Regulation: Empirical Analysis; Regulatory Agencies

Bibliography

Bardach E 1977 *The Implementation Game: What Happens After a Bill Becomes a Law*. MIT Press, Cambridge, MA

Elmore R 1979 Backward mapping: Implementation research and policy decisions. *Political Science Quarterly* **94**(4): 601–16

Goggin M L, Bowman A O, Lester J P, O'Toole Jr L J. 1990 *Implementation Theory and Practice: Toward a Third Generation*. Scott, Foresman, Glenview, IL

Majone G, Wildavsky A 1979 Implementation as evolution. In: Pressman J L, Wildavsky A (eds.) *Implementation*. University of California Press, Berkeley, CA

Mazmanian D A, Sabatier P A 1983 *Implementation and Public Policy*. Scott, Foresman, Glenview, IL

Nakamura R T, Smallwood F 1980 *The Politics of Policy Implementation*. St. Martins Press, New York

Salomon L M (ed.) 1989 *Beyond Privatization: The Tools of Government Action*. Urban Institute Press, Washington, DC

E. Bardach

Implicit Learning and Memory: Psychological and Neural Aspects

Learning is the acquisition of new information that leads to lasting changes in behavior, and memory reflects the storage and retrieval of learned information. Implicit learning and implicit memory refer to instances in which acquisition, storage, and retrieval of learned information proceed without the person's awareness that they are learning or remembering. For example, how many times must you turn an automobile's steering wheel to smoothly execute a right turn? This is a difficult question to answer explicitly, yet we skillfully make many such turns each day. At some level we have learned the information and we use it each time we drive. However, skills, such as driving, rely primarily on implicit forms of learning and memory. Few of us intentionally learn the requisite number of steering wheel turns—we simply acquire the skill through practice. Rarely do we consciously reflect upon the information that is controlling our steering movements—we perform them automatically without awareness. The present review focuses on laboratory studies of implicit learning and memory that have provided a foundation for understanding the nature of the underlying psychological and neural processes.

1. Implicit Memory

'Implicit memory is revealed when previous experiences facilitate performance on a task that does not require conscious or intentional recollection of those experiences,' Schacter (1987, p. 501). By contrast, explicit memory 'is revealed when performance on a task requires conscious recollection of previous experiences.' Implicit memory subsumes various different types of phenomena, and we do not attempt to comprehensively review all of them here (for review and discussion, see Roediger and McDermott 1993, Schacter and Buckner 1998). Instead, we focus on a type of implicit memory known as priming. Priming

refers to a change in the ability to identify or produce an item as a result of a particular prior encounter with the item. Priming is assessed with tests in which subjects attempt to identify briefly flashed stimuli, complete word stems or fragments, with the first word that comes to mind, or produce items from a category in response to a category cue. In stem-completion priming, for example, after recently reading a list of words that included 'COFFEE,' people are more likely to complete 'COF—' with 'COFFEE' than with an alternative (e.g., COFFIN). Priming can be implicit because it can occur without the ability to remember explicitly that 'COFFEE' was on the original list.

Scientific interest in priming was piqued by the discovery of fundamental differences between priming and explicit memory. Psychological research established that explicit memory typically depends upon learning and remembering semantic information (i.e., information about meaning). If you forget to bring your grocery list to the market, you would be likely to remember 'COFFEE' in terms of its semantic associations: a morning drink, helps you stay awake, and so forth. You would be much less likely to remember 'COFFEE' from superficial perceptual characteristics, such as the spelling or sound of the word, the appearance of the word on the written list, and the like. Unlike explicit memory, many instances of priming depend greatly upon memory for such perceptual characteristics (Curran and Schacter 1996). In addition to such psychological differences between priming and explicit memory, differences in the underlying brain systems were indicated from studies of people with brain injuries (Squire and Knowlton 2000). A variety of studies have shown that amnesic patients with damage to medial temporal and diencephalic brain regions, and severe deficits in explicit memory, nonetheless often show normal priming effects. Intact priming in amnesia has been shown for a variety of stimuli, including visual and auditory words, nonsense words (e.g., TARB), pictures, and novel objects (Schacter and Buckner 1998, Squire and Knowlton 2000).

Recognition that priming relied on perceptual information and brain systems distinct from those supporting explicit memory lead to the hypothesis that priming depends upon brain regions that normally support perceptual functions. Several neuroimaging studies, using positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), have shown that perceptually based visual priming is accompanied by reduced activity in posterior cortical brain regions that have previously been implicated in visual word and object processing (Schacter and Buckner 1998). Complementary studies have shown that perceptual priming is absent in patients with damage to the visual cortex (Gabrieli 1998). Thus, the neuroimaging and neuropsychological data converge on the conclusion that visual cortical regions play an important role in perceptually based priming effects.

Even before recent neuroimaging and neuropsychological research began delineating the brain systems underlying perceptual priming, psychologists began to appreciate that not all forms of implicit memory are perceptually dependent (reviewed by Roediger and McDermott 1993). Whereas perceptual priming tests typically provide subjects with information that is perceptually related to the to-be-remembered item (e.g., word stems), 'conceptual priming' can be demonstrated when conceptually (i.e., semantically) related cues are provided (e.g., 'DRINK —?,' 'TEA—?,' or 'Grown in South America?' for 'COFFEE'). Delineating psychological differences between conceptual priming and explicit memory has been difficult because they share many features, but the underlying neural substrates have been dissociated. As with perceptual priming, amnesic patients with explicit memory deficits can show normal conceptual priming effects (Schacter and Buckner 1998, Squire and Knowlton 2000). Neuroimaging studies indicate that conceptually based priming effects are accompanied by reduced activity in the prefrontal cortex (Schacter and Buckner 1998). These priming-related brain regions are also implicated in a wider variety of semantic information processing functions (e.g., understand the meaning of words). Thus, the same brain areas that normally allow us to process semantic information appear to support conceptual implicit memory.

In addition to providing neuroanatomical localization, neuroimaging studies of priming can potentially elucidate the nature of the underlying processes. Most priming studies indicate that activity levels are higher when relevant brain regions process unprimed rather than primed stimuli. Such 'priming attenuation' may indicate that primed information is processed more efficiently than unprimed information (analogously, priming an engine makes it start more easily). Conversely, other studies have found higher levels of brain activity for primed than unprimed information. Such 'priming enhancement' may reflect the formation of new perceptual representations for primed information. Recent work has indicated that attenuation is observed for stimuli that are previously known to subjects (e.g., famous faces, familiar symbols, words), but enhancement is observed for previously unknown stimuli (e.g., unfamiliar faces, unfamiliar symbols, novel objects). Thus, the mechanisms underlying priming effects may fundamentally differ according to the nature of the to-be-remembered information. Efficient processing may prime familiar stimuli, whereas the establishment of new perceptual representations may prime unfamiliar stimuli (Henson et al. 2000).

2. *Implicit Learning*

Reber first coined the term 'implicit learning' to describe the ability to learn complex information in the absence of awareness (reviewed by Reber 1993). In

contrast to implicit memory, which primarily emphasizes limited awareness and/or intention during retrieval, implicit learning primarily emphasizes limited awareness/intention during encoding (Stadler and Frensch 1997). Reber developed an artificial grammar-learning task in which subjects are exposed to seemingly random consonant strings that, unbeknownst to them, were generated from a complex set of rules (an 'artificial grammar'). After viewing a list of grammatical stimuli (e.g., XRNRTV, FRNRTZ, etc.), subjects are able to discriminate between grammatical and nongrammatical strings despite their limited ability to articulate the information upon which their judgments are based.

Reber originally believed that subjects in an artificial grammar-learning experiment develop an abstract mental representation of the rules that comprise the grammar (e.g., the second and fourth letters are always the same). In contrast, episodic theories maintain that subjects learn by storing individual exemplars (e.g., XRNRTV, FRNRTZ, etc.). Episodic theories can account for subjects' abilities to correctly classify grammatical strings that they have never seen before by positing that judgments are based on the global similarity between the test item and stored exemplars. For example, one could categorize *F*RNRT*V* and *X*RNRT*Z* as grammatical because of their similarity to the previously stored exemplars (*X*RNRT*V*, *F*RNRT*Z*). Grammatical test items will tend to be more similar to encoded exemplars than nongrammatical items, so subjects will be more likely to classify grammatical items accordingly.

The debate over abstractionist and episodic theories will not be settled here (reviewed by Neal and Hesketh 1997), but one possible resolution is that implicit learning leads to abstract representation whereas explicit learning results in exemplar storage. This possibility has been addressed by studying implicit learning in amnesic patients. Amnesic patients can learn to make grammaticality judgments as accurately as control subjects, despite their impaired explicit memory for studied exemplars (Squire and Knowlton 2000). Exemplar-specific implicit learning was inferred from amnesic patients' normal sensitivity to the frequency of letter bi-grams (e.g., NR) and tri-grams (e.g., NRT) that were in the original training set (e.g., XRNRTV, FRNRTZ). Abstract implicit learning was inferred from amnesic patients' normal ability to transfer learning across physical changes in the letters entered into the grammar (though letter-set transfer is not universally agreed to demand an abstract representation, Neal and Hesketh 1997). For example, the grammar could be modified to shift each letter two positions back within the alphabet (e.g., compare VPLPRT and DPLPRX with the original examples). As a whole, these results suggest that implicit learning can result in both abstract and episodic representations.

Currently, there is no evidence to directly link artificial grammar learning to particular neural mechanisms beyond the data from amnesic patients indicating that medial temporal and diencephalic regions are not necessary for such learning to occur. Other evidence for priming-like effects has been documented with categorization tasks that are conceptually similar to artificial grammar learning. Artificial grammar learning can be considered to be a special case of category learning in that subjects are required to judge whether individual stimuli (e.g., consonant strings) are members of a larger class (e.g., a grammar). Patients with amnesia show normal category learning in tasks requiring subjects to decide if novel dot-patterns belong to the same category as a set of training patterns (Squire and Knowlton 2000). A recent fMRI experiment showed that processing of categorical dot-patterns elicits less visual cortex activation than does processing of noncategorical patterns (Reber et al. 1998). Thus, perceptual categorization may involve mechanisms similar to those that support perceptual priming.

Another form of implicit category learning has been demonstrated with probabilistic classification tasks (Squire and Knowlton 2000). These classification tasks require subjects to predict which of two outcomes (e.g., sunshine or rain) will follow particular cue combinations (e.g., combinations of one to four geometric symbols). The cues are probabilistically predictive of outcomes, so that subjects' predictive accuracy increases with training. As observed for the other forms of category learning, amnesic patients show normal probabilistic classification learning. Probabilistic classification learning is impaired in patients with Parkinson's disease (PD) and Huntington's disease (HD). PD and HD are neurological diseases that cause severe movement impairments because of damage to a subcortical brain region know as the basal ganglia. Thus, the basal ganglia may make a critical contribution to probabilistic classification. However, basal ganglia damage has not been shown to interfere with category learning of novel dot-patterns or artificial grammars (Reber and Squire 1999).

Implicit sequence learning is another task that, like probabilistic classification learning, requires subjects to learn associations between multiple cues and outcomes, and it has been more clearly linked to activity within the basal ganglia (reviewed by Clegg et al. 1998, Curran 1997). Implicit sequence learning has been primarily studied with a serial reaction time (SRT) task. Subjects typically view a stimulus moving across different visual locations and press location-corresponding response keys as quickly as possible. Normal subjects and patients with amnesia show improved performance (i.e., respond faster and more accurately) when stimulus locations follow a repeating pattern, compared with conditions in which stimulus locations are randomly determined. This form of learning does not depend on explicit knowledge. Implicit sequence learning is impaired in patients with PD or HD.

Impaired sequence learning in PD and HD patients suggests a basal ganglia contribution that has been confirmed by sequence-learning studies using PET and fMRI (Clegg et al. 1998, Curran 1997). These studies additionally have identified learning-related changes in primary motor cortex, supplementary motor areas, premotor cortex, and the thalamus. The distribution of learning-related brain activity suggests that SRT learning involves changes in motor areas that subserve visually guided movement. Cognitive psychologists have similarly concluded that implicit sequence learning involves response-based information processing by showing that learning depends more upon response consistency than stimulus consistency (Willingham et al. 2000). Neuroimaging studies have primarily found that sequence learning leads to increased activity in implicated areas—as would be expected from the view that implicit learning and memory for novel information should enhance relevant brain activity. However, learning-related changes in brain activity are not so simply described when one considers the multiplicity of implicated brain areas and the complete time-course of learning. Across the temporal course of sequence learning, thalamic activity initially decreases then later increases (Rauch et al. 1998). Rauch et al. hypothesize that these effects reflect input from distinct basal ganglia circuits that respectively suppress and enhance thalamic activity. Further understanding of the functional significance of learning-related increases and decreases in activity will depend on more detailed temporal information as well as a better understanding of the neurophysiology of implicated brain areas.

Why would basal ganglia function be related to probabilistic classification and sequence learning, but not categorization of artificial grammars or random dot-patterns? The categorization tasks typically include a passive learning phase in which subjects are exposed only to category members, followed by a distinct discrimination test designed to tap categorical knowledge. Learning and test phases are procedurally identical in probabilistic classification and sequence-learning experiments, and feedback is given to differentiate incorrect from correct responses. Task consistency or the provision of feedback may be relevant to understanding basal ganglia involvement. Behavioral studies have demonstrated the importance of response consistency for sequence learning (Willingham et al. 2000), but this issue has yet to be investigated with probabilistic classification learning. Others have emphasized the importance of feedback for types of learning that depend upon the basal ganglia (Reber and Squire 1999).

3. Conclusions

The discovery of implicit learning and memory has advanced the view that learning and memory are not the product of a unitary neurocognitive system that is dedicated to learning and memory. Dissociations between implicit and explicit forms of learning/memory exemplify the perspective that learning and memory depend upon multiple brain systems. Further investigations have indicated that implicit learning and memory are attributable to diverse processes. The over-riding principle appears to be that many, if not all, brain systems can support implicit learning and memory. Implicit learning and memory involve changes in information-processing systems that normally contribute to specific cognitive or behavioral functions, such as perception, movement, and the like. Visual perceptual processes support visual priming and visual category learning. Semantic processing regions support conceptual priming. Brain systems related to movement control underlie implicit sequence learning. Ongoing research is aimed toward understanding the relationships among these forms of implicit learning and memory, and specifying the characteristics of each.

See also: Conscious and Unconscious Processes in Cognition; Implicit Memory, Cognitive Psychology of; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Priming, Cognitive Psychology of

Bibliography

Clegg B A, DiGirolamo G J, Keele S W 1998 Sequence learning. *Trends in Cognitive Sciences* **2**: 275–81

Curran T 1998 Implicit sequence learning from a cognitive neuroscience perspective: What, how, and where? In: Stadler M A, Frensch P A (eds.) *Handbook of Implicit Learning*. Sage Publications, Thousand Oaks, CA, Chap 11, pp. 365–400

Curran T, Schacter D L 1996 Implicit memory and perceptual brain mechanisms. In: Herrmann D, McEvoy C, Hertzog C, Hertel P, Johnson M (eds.) *Basic and Applied Memory Research: Theory in Context*. Erlbaum, Mahweh, NJ, Vol. 1, Chap. 11, pp. 221–40

Gabrieli J D 1998 Cognitive neuroscience of human memory. *Annual Review of Psychology* **49**: 87–115

Henson R, Shallice T, Dolan R 2000 Neuroimaging evidence for dissociable forms of repetition priming. *Science* **287**: 1269–72

Neal A, Hesketh B 1997 Episodic knowledge and implicit learning. *Psychological Bulletin and Review* **4**: 24–37

Rauch S L, Whalen P J, Curran T, McInerney S, Heckers S, Savage C R 1998 Early thalamic deactivation during early implicit sequence learning: A functional MRI study. *Neuroreport* **9**: 865–70

Reber A S 1993 *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. Oxford University Press, New York, Vol. 19

Reber P J, Squire L R 1999 Intact learning of artificial grammars and intact category learning by patients with Parkinson's disease. *Behavioral Neuroscience* **113**: 235–42

Reber P J, Stark C E-L, Squire L R 1998 Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences, USA* **95**: 747–50

Roediger H L I, McDermott K B 1993 Implicit memory in normal human subjects. In: Spinnler H, Boller F (eds.) *Handbook of Neuropsychology*. Elsevier, Amsterdam, Vol. 8, Chap. 2, pp. 63–131
Schacter D L 1987 Implicit memory: History and current status. *Journal of Experimental Psychology: Learning* **13**: 501–18
Schacter D L, Buckner R L 1998 Priming and the brain. *Neuron* **20**: 185–95
Squire L R, Knowlton B J 2000 The medial temporal lobe, the hippocampus, and the memory systems of the brain. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 765–79
Stadler M A, Frensch P A (eds.) 1997 *Handbook of Implicit Learning*. Sage Publications, Thousand Oaks, CA, Chap. 53
Willingham D B, Wells L A, Farrell J M, Stemwedel M E 2000 Implicit motor sequence learning is represented in response locations. *Memory and Cognition* **28**: 366–75

T. Curran and D. L. Schacter

Implicit Memory, Cognitive Psychology of

1. Phenomena, Methods of Investigation, and Explanations of Implicit Memory

Does experience influence our thoughts and actions without being aware of it? A growing body of scientific literature on implicit memory deals with this question. In the databases for psychological literature we can find e.g., 19 references to 'implicit memory' for the five year period between 1984 and 1988, and more than 200 titles between 1999 and September 2001. What makes this concept so attractive, what is its definition, and what phenomena does it explain?

'Implicit memory is revealed when previous experience facilitates performances on a task that do not require conscious or intentional recollection of those experiences' (Schacter 1987, p. 501). This can be illustrated by the use of a picture-naming task: subjects are more successful in terms of speed or correctness when naming pictures or words that have been shown previously compared to pictures or words that have not been presented before. The assumption that such facilitation effects (called priming) do 'not require conscious or intentional recollection' of previous experience, is taken for sure from the point of view of the experimenter, who in fact does not instruct the subjects to use any kind of recollection when naming the pictures. Thus, 'implicit memory' refers to memory effects that can be shown by implicit tasks or indirect tasks that, in contrast to explicit tasks, do not instruct the subjects to remember what happened in the past. Here, the terms 'implicit' and 'explicit' refer to different tasks, distinguished operationally by the instructions given to subjects at test. The terms 'indirect memory tasks' and 'direct memory tasks' are also frequently used in the literature to make this distinction in instruction.

The most commonly used explicit memory tasks are free recall, cued recall, and recognition. In free recall, subjects are instructed to report a prior episode without any additional help, while in cued recall, additional cues are presented that might prompt the answer (e.g., the first word of a studied word pair is given in test while the second has to be recalled). In recognition, a previously presented item or episode is presented again (together with new ones), and subjects are asked whether they have seen the item before or not.

The most frequently used implicit memory tasks are identification, fragment completion, word stem completion, and lexical decision. In identification, subjects are given brief exposure (some milliseconds) to a stimulus, which they have to identify. A variation of this task is the clarification procedure. In this version the subjects are presented repeatedly with degraded pictures or words that gradually become more complete in each presentation until the subjects can identify the items. In completion tests, subjects are either given a word stem (g a r _ _ _) or a word fragment (g _ r _ e _) together with the instruction to complete it with the first word that comes to mind (e.g., g a r d e n). In a lexical decision task, subjects are required to state whether or not a particular letter string (words and nonwords are presented) is a word. In all of these tasks, implicit memory is revealed if items of a prior study phase receive more frequent, more correct, or faster responses compared to control items not presented in the study phase.

The notion of implicit memory has become particularly attractive because of the assumption that measures of implicit memory tap an unaware or unconscious form of retention. In this sense, implicit memory refers to a special memory system or to memory processes that are different from the memory system or memory processes underlying explicit memory. In this rationale, implicit memory refers to the saving and use of experience that lies outside of conscious retrieval or recollection of the past. Thus, implicit memory stands in contrast with explicit memory, where retrieval is used with recollective experience of the past.

The main evidence for the distinction between implicit memory and explicit memory comes from the stochastic independence of memory performance on implicit and explicit memory tests and from functional dissociations between explicit and implicit measures of retention. Stochastic independence refers to the finding that memory for a single item in one test cannot be predicted by the memory performance for the same item on the other test. Functional dissociations denote the fact that certain variables affect explicit memory tests but not implicit memory test, or do so in the opposite way.

It was demonstrated that the manipulation of levels of processing (structural or shallow encoding vs. semantic or deep encoding), attention and reading vs.

generating a word during study had strong effects on explicit measures of memory, but no effect on priming. On the other hand, manipulations of surface features of the stimuli, such as typography or study-test changes in modality of presentation reduced or eliminated priming in implicit tasks, whereas performance in explicit measures was not significantly influenced (for a review, see e.g., Roediger and McDermott 1993).

Functional dissociations have also been found in quasi-experimental research. Amnesic patients with bilateral damage to the medial temporal lobes and to related limbic and diencephalic structures are severely impaired on explicit tests of memory while showing normal memory on implicit tests (Squire et al. 1987). The same pattern of functional dissociation between explicit and implicit measures is revealed in young children and mentally retarded subjects (Perrig-Chiello and Perrig 1993) and elderly people (La Voie and Light 1994).

Given this data, it seems perfectly reasonable to infer the existence of a cognitive system with separate mechanisms that refer to the apparent divide in terms of implicit and explicit memory. Many individual theories have evolved to explain this data, and why they differ. They can be classified in one of two general approaches: memory systems theories and processing theories. The memory systems theories assume that there are distinct memory systems in the brain, that these systems provide the neural underpinnings for performances on various memory tests, and that the systems can be selectively impaired. The most prominent distinctions refer to episodic memory, semantic memory, procedural memory, and perceptual representation systems (Tulving and Schacter 1990). In processing theories, the principle of Transfer-Appropriate-Processing (TAP) and the distinction between perceptual and conceptual processing are held responsible for explaining the empirical findings of implicit memory research. In this view, transfer between learning and test—and thus performance—is best if the task and processes involved share a maximum of features (Roediger 1990). Conscious recollection of past events, as assessed by explicit tests of memory, is attributed to the episodic memory in the system approach, and to dominant conceptual processes in the process view. Priming in the implicit tests of memory is related to semantic memory, procedural memory, and perceptual representation systems in the system theory, and to perceptual processes in the processing account.

2. *Refinement of Tasks and Theories*

In the last decade of the twentieth century, more than two dozen tasks have become standard to investigate implicit memory performance (Roediger and McDermott 1993). Different types of implicit memory tests have been established. Distinctions are made on different dimensions: the contrast of perceptual and conceptual tests, the degree of resemblance of a target item between learning and test, the reference to different modalities involved to process the stimuli (verbal, nonverbal, visual, acoustic), and the information contained in the stimuli (item-specific, relational or associative).

More subtle functional dissociations between memory tasks challenge simple views of both the system approach and the TAP approach. The TAP framework has great success accounting for experimental dissociations of implicit and explicit memory tests that are rather perceptual or conceptual in nature. It comes to its limit, however, when dissociations among conceptual implicit and conceptual explicit tests or perceptual implicit and perceptual explicit tests are obtained. Such is the case, when, e.g., large age-related differences are found in two explicit memory tasks, regardless of the perceptual–conceptual difference, while age-related differences in implicit memory are revealed in a conceptual category-production task, but not in a perceptual word-fragment completion task (Jelicic et al. 1996). The authors explain these different realizations of perceptual and conceptual information in free recall and priming with selective age-related impairments in different cortical regions (see *Implicit Learning and Memory: Psychological and Neural Aspects*).

Further demonstrations on how implicit memory research comes to continuously more precise task analysis, and how finer grained theories relate to effects of levels of processing (LOP) on priming. It could be shown that the effects of levels of processing on a category-exemplar production test were greater when study lists were organized by category than when they were presented randomly (Mulligan et al. 1999). The explanation of these results by the authors might indicate the direction of development of the processing view to explain dissociations within classes of implicit tests (either conceptual or perceptual). They refer to the distinction between item-specific and relational processing. Relational information is shared by to-be-retrieved items, whereas item-specific information is unique to an item. LOP is viewed as a manipulation of item-specific encoding, and list organization is considered to affect relational encoding. Their data show that LOP effects in a category-exemplar production task are greater when relational information is highly salient.

3. *Towards the Measurement of Individual Differences in Implicit Memory*

It would be a superb achievement of memory research if theoretical distinctions of experimental psychology could be used to diagnose a special aptitude or ability

of one specific subject on the basis of population norms. Implicit memory research seems to be aimed at a successful integration of experimental research and an individual difference approach. For example, quasi-experimental research will possibly succeed in discriminating amnesic patients with different etiologies by the use of different implicit memory tests (see Moscivitch et al. 1993 for a review). At this point, however, it is mandatory that memory tasks have to be evaluated not only on the basis of experimental criteria but also according to the psychometric standards of the individual-difference approach. The critical question concerns the reliability of an individual's characterization, or the reliability of the assignment to one group or another. In the light of these considerations, it is essential to assess the reliability of tests and validity of constructs. This issue is just beginning to come into the discussion, and the first empirical findings are somewhat disappointing (Meier and Perrig 2000, Buchner and Wippich 2000). In a series of experiments the test–retest reliability, split-half reliability, and parallel-test reliability of explicit measures (recall, cued recall, and recognition) varied between $r = 0.43$ and $r = 0.65$. With the implicit measures (perceptual clarification and word stem completion) the highest correlation was $r = 0.39$, and the others ranged between $r = 0.08$ and $r = 0.29$ (Meier and Perrig 2000). Such a lack of reliability, of course, is unacceptable for diagnostic purposes; at the same time it is problematic for experimental research. Lack of reliability could even be responsible for functional dissociations, when a variable has an effect on explicit memory performance but not on implicit tasks. Lower reliability leads to a higher proportion of error variance in the observed variance. If the influence of an independent variable on two different kinds of tests is investigated, then the test with low reliability must show a greater effect than the more reliable test to reach significance. Given the fact that the expected maximum correlation of a test with any other test is limited by its reliability, it is clear that low correlations of implicit memory measures with any other variable are at least partly due to the low reliabilities of these measures. Thus, there is clearly a problem that has to be put on the top of the research agenda of implicit memory research.

4. *Confounded Factors and the Phenomenological Status of Implicit Memory*

The special attractiveness of implicit memory lies in the assumption that measures of implicit memory tap a nonconscious, automatic form of retention, or incidental retrieval. However, the issue of awareness is a thorny one that will keep the field busy for some time to come. The distinction between implicit memory tests and explicit memory tests are well established at the operational level. On the other hand, despite neat dissociations between explicit and implicit memory tests and stochastic independence between them, we do not know for sure whether and to what degree subjects use intentional retrieval and conscious recollection in implicit memory tasks. Furthermore, we do not know to what degree automatic and nonconscious processes are involved in intentional retrieval. In order to speak of automatic or unconscious effects of memory, criteria have to be satisfied that exclude contamination with intentional retrieval (Perrig 2000). This challenge is not unique to implicit memory but to any approach that tries to separate conscious from unconscious processes as in implicit learning, implicit expertise, skill acquisition, automatic detection, or nonconscious perception.

There are different techniques to disentangle conscious recollection from automatic or unintentional retrieval. (a) A strict retrieval intentionality criterion asks for holding constant all overt conditions affecting subjects at study and test except for the instructions given at time of the test (Schacter et al. 1989). (b) The subjects can be queried about the strategies they used in completing the task to separate 'test aware' from 'test unaware' subjects. Test aware subjects are those who realize at one point during the implicit memory test that some items are from the study phase. (c) The process dissociation procedure developed by Jacoby (1991) estimates the amount of automatic and controlled processing within one and the same task. In recognition e.g., the measurement model assumes that either recollection or automatic retrieval facilitates a yes-response. The basic logic to separate controlled and automatic processing includes an experimental condition where the subjects have to say 'no' to old items if they remember that the items belong to a specified set of items from the study list. If the subjects nevertheless say 'yes' to some of those items, it is assumed that this is based on an automatic response because conscious recollection works in opposition to the automatic response and would produce a 'no' answer. (d) The sensitivity criterion of Merikle and Reingold (1991) states that one should be able to show greater effects of a variable on an implicit test than on an explicit test in order to speak of unconscious or automatic effects. This must be so, because the information would be used in the explicit measure if it were available for conscious retrieval. Empirical data fit this assumption when priming can be produced in absence of recognition or recall.

5. *Implicit Memory in the Broad Range of Implicit Mental Processes*

An important question is whether implicit memory represents processes that are unique or at least distinct from processes that are studied in implicit learning,

implicit expertise, skill acquisition, automatic detection, or unconscious perception. It could well be that the very same implicit processing principles underlie phenomena that are studied in other research areas using different experimental paradigms (Graf and Masson 1993, Kirsner et al. 1998, Reder 1996). A broader view that integrates procedures and data of different research domains could open the researcher's mind for valuable progress in explaining implicit processes and their differences to explicit information processing.

Is there a common psychological explanation for automatic and controlled processing, and for the behavior of a patient with brain stroke, who spontaneously switches between English and French in answering questions but cannot translate words and sentences when asked to do so? Are the same principles involved when people gradually improve in solving complex perceptual or conceptual problems despite the fact that they cannot identify the feature or principle on which their new expertise is based? Would those principles even explain the fact that information can be extracted from words during brief presentation under conditions where the words cannot be identified?

In fact, the terminology to describe what seems to control behavior outside of intentional control in other research areas comes very close to the definition of implicit memory. Cohen (1984, p. 96) speaks of 'procedural knowledge,' '… when experience serves to influence the organization of processes that guide performance without access to the knowledge that underlies the performance.' Reber et al. (1991, p. 888) speak of 'implicit learning,' '… when a complex, rule-governed knowledge base is acquired largely independently of awareness of both the process and the product of the acquisition.' The fluent switch between notions like 'implicit knowledge,' 'implicit learning,' and 'implicit memory' either confronts us with the fact that psychology has—despite precision at the operational level—weaknesses at the terminological level, or it asks for synthesis in the explanation of implicit phenomena or implicit cognition. This endeavor is not a question of semantics, but the need for definitions that point out problems rather than hide them. Future work will have to show whether different definitions are needed for skill and habit learning, simple classical conditioning, and priming, all phenomena that for the time being can be considered as instances where memory is expressed through performance rather than by recollection.

See also: Elaboration in Memory; Encoding Specificity in Memory; Implicit Learning and Memory: Psychological and Neural Aspects; Incidental versus Intentional Memory; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Memory: Levels of Processing; Memory Models: Quantitative; Memory: Organization and Recall; Memory Retrieval; Mood-dependent Memory; Short-term Memory, Cognitive Psychology of

Bibliography

Buchner A, Wippich W 2000 On the reliability of implicit and explicit memory measures. *Cognitive Psychology* **40**(3): 227–9

Cohen N J 1984 Amnesia and the distinction between procedural and declarative knowledge. In: Squire L R, Butters N (eds.) *The Neuropsychology of Memory*. Guilford Press, New York, pp. 83–103

Graf P, Massan M E (eds.) 1993 *Implicit Memory: New Directions in Cognition, Development, and Neuropsychology*. Laurence Erlbaum Associates, Hillsdale, NJ

Jacoby L L 1991 A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language* **30**(5): 513–41

Jelicic M, Craik F, Moscovitch M 1996 Effects of ageing on different explicit and implicit memory tasks. *European-Journal of Cognitive Psychology* **8**(3): 225–34

Kirsner K, Speelman C, Maybery M, O'Brien-Malone A, Anderson M, MacLeod C (eds.) 1998 *Implicit and Explicit Mental Processes*. Laurence Erlbaum Associates, Mahwah, NJ

La Voie D, Light L L 1994 Adult age differences in repetition priming: A meta-analysis. *Psychology and Aging* **9**(4): 39–53

Meier B, Perrig W J 2000 Low reliability of perceptual priming: Consequences for the interpretation of functional dissociations between explicit and implicit memory. *Quarterly Journal of Expermental Psychology* **53** A (1): 211–33

Merikle P M, Reingold E M 1991 Comparing direct (explicit) and indirect (implicit) measures to study unconscious memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **17**: 224–33

Moscovitch M, Vriezen E, Goshen-Gottstein Y 1993 Implicit tests of memory in patients with focal lesions or degenerative brain disorders. *Handbook of Neuropsychology* **8**: 133–73

Mulligan N W, Guyer P S, Beland A 1999 The effects of levels-of-processing and organization on conceptual implicit memory in the category exemplar production test. *Memory and Cognition* **27**(4): 633–47

Perrig W J 2000 Intuition and levels of control: The non-rational way of reacting, adapting and creating. In: Perrig W J, Grob A (eds.) *Control of Human Behavior, Mental Processes, and Consciousness*. Laurence Erlbaum Associates, Mahwah, NJ

Perrig-Chiello P, Perrig W J 1993 Implicit and explicit memory in mentally retarded, language/learning disabled and normal kindergarten children. *Swiss Journal of Psychology* **54**(2): 77–86

Reber A S, Walkenfeld F F, Hernstadt R 1991 Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology*: *Learning*, *Memory*, *and Cogniton* **17**: 888–96

Reder L M (ed.) 1996 *Implicit Memory and Metacognition*. Laurence Erlbaum Associates, Hillsdale, NJ

Roediger H L 1990 Implicit memory: Retention without remembering. *American Psychologist* **39**: 373–80

Roediger H L, McDermott K B 1993 Implicit memory in normal human subjects. In: Spinnler H, Boller F (eds.) *Handbook of Neuropsychology*. Elsevier, Amsterdam, pp. 63–131

Schacter D L 1987 Implicit memory: History and current status. *Journal of Experimental Psychology*: *Language*, *Memory*, *and Cognition* **13**(3): 501–18

Schacter D L, Bowers J, Booker J 1989 Intention, awareness and implicit memory: The retrieval intentional criterion. In: Lewandowsky S, Dunn J C, Kirsner K (eds.) *Implicit Memory: Theoretical Issues.* Erlbaum, Hillsdale, NJ, pp. 47–65

Squire L R, Shimamura A P, Graf P 1987 Strength and duration of priming effects in normal subjects and amnesic patients. *Neuropsychologia* **25**(1-B): 195–210

Tulving E, Schacter D L 1990 Priming and human memory systems. *Science* **247**: 301–6

W. J. Perrig

Impression Management, Psychology of

Impression management (also called self-presentation) refers to the processes by which individuals attempt to control how they are perceived by other people. By conveying particular impressions of their abilities, attitudes, motives, status, emotional reactions, and other personal characteristics, people can influence others to respond to them in desired ways. Impression management is a common means by which people influence one another and obtain various goals (Jones and Pittman 1982).

The concept of impression management was popularized by Erving Goffman (1959), a sociologist and ethnologist, who viewed impression management not only as a means of influencing how one is treated by other people but also as an essential aspect of social interaction. According to Goffman (1959), social encounters are mediated by the impressions that people form of one another. As a result, smooth and effective interpersonal interaction requires that each interactant projects a public identity to guide others' behaviors. The study of impression management was introduced into social psychology by Edward Jones (1964), whose early research on ingratiation focused on the ways in which people influence others by making themselves appear likeable and socially attractive. During the 1960s and 1970s, research on impression management developed along two distinct paths. One line of research focused on the situational and dispositional factors that affect the kinds of impressions people try to convey of themselves, and the other line applied self-presentational perspectives to the study of other psychological phenomena, such as attitude change, conformity, leadership, psychological disorder, and health risks.

1. *The Nature of Impression Management*

People use many different behaviors in the service of impression management. Any behavior that conveys information about an individual may be used as a self-presentational tactic (Leary 1995, Schlenker 1980). This is not to say that all behavior is necessarily self-presentational—only that virtually any behavior may be used for self-presentational purposes. Most empirical research has focused on people's descriptions of their attributes and attitudes to foster impressions in others' minds. However, people also impression-manage nonverbally. For example, people can regulate how others perceive them through their facial expressions, gestures, ways of standing and sitting, and choices of seating location (around a table or in a classroom, for example). They also use props—physical objects such as clothing, furnishings, automobiles, and other personal possessions—to promote certain images of themselves. Behaviors as mundane as eating, grooming, and driving also have a self-presentational component. Furthermore, many instances of helpfulness, as well as aggression, may reflect people's efforts to be viewed as helpful versus threatening, respectively.

Self-presentation has sometimes been viewed as a deceitful and manipulative means of influencing other people, and sometimes it is. However, impression management is not necessarily deceptive. People often engage in strategic impression management to assure that other people perceive them accurately. Because personal characteristics and inner states are not always obvious to other people, people must often take steps to insure that others perceive them accurately. Furthermore, although people sometimes present images of themselves that they know are untrue, more often they veridically present those aspects of themselves that they believe will help to achieve their interpersonal goals in a particular situation. Although self-presentation is tactical, it is not necessarily deceptive.

Most self-presentation involves efforts to convey a socially desirable impression in order to be regarded positively. Generally, people are more likely to achieve their goals if they are perceived as likeable, competent, ethical, or otherwise socially desirable than if others regard them negatively, so desirable self-presentations predominate. However, people sometimes wish to convey socially undesirable impressions of themselves. For example, people may want to be seen as hostile because doing so helps them to intimidate others to do as they wish, or as weak or helpless to elicit support from other people. For this reason, it is important to distinguish between desirable self-presentations (i.e., those that are likely to result in favorable evaluations) and desired self-presentations (i.e., those that the individual believes will accomplish his or her goals).

Although little research has examined self-presentational behaviors in children, evidence suggests that children begin to show a concern with other people's evaluations of them soon after the emergence of self-awareness, typically by age two. By age three, children show evidence of trying to influence other people's perceptions of them, for example, by denying that wrongdoing was their fault. Children become increasing skilled at controlling others' impressions of them

with age, although they do not appear to attain an adult level of self-presentational ability until late adolescence.

2. The Determinants of Impression Management

People have many reasons to monitor and regulate how they are perceived by others. Most important outcomes in social life—including friendships, group memberships, romantic relationships, desirable jobs, status, and influence—partly depend on other people perceiving the individual as being a particular kind of person or possessing certain kinds of attributes. Because most of people's goals depend on making particular desired impressions and not making undesired impressions, people are pervasively concerned with the impressions that other people form of them. People appear to monitor how they are coming across on an ongoing basis, but the degree to which they are motivated to impression-manage and the nature of the impressions they try to foster vary across situations and among individuals.

2.1 Impression Motivation

Three primary factors affect the degree to which people are motivated to impression-manage in a particular social encounter (Leary and Kowalski 1990). First, people are more highly motivated to control how they are perceived by others when they believe that their public images are relevant to the attainment of desired goals. Conversely, when the impressions that other people form have few implications for one's outcomes, the motivation to impression-manage will be low. This fact is one reason that people are more likely to impression-manage when interacting with powerful, high status people than with less powerful, low status individuals.

Second, people are more highly motivated to impression-manage the more valuable the goals for which their public impressions are relevant. When achieving important, highly desirable goals depends on the kinds of impressions others form of them, people are more motivated to monitor and control how they are perceived than when the goals are less valuable.

Third, people are more highly motivated to impression-manage when a discrepancy exists between how they want to be perceived and how they believe other people currently perceive them. Thus, public failures and embarrassing events that convey impressions that are discrepant from the images that people wish others to form cause them to engage in self-presentational efforts to repair their damaged image.

2.2 Impression Construction

When people are motivated to manage their impressions, features of the social context and their own personalities determine the specific nature of the impressions that they try to convey. Five sets of factors are particularly important determinants of impression construction (Leary and Kowalski 1990). First, the content of people's self-presentations is affected by the roles that they occupy and by the norms in the immediate social context. Typically people want to convey impressions that are consistent with their roles and the prevailing norms. Many roles—such as the roles of leader, military officer, and member of the clergy—carry self-presentational requirements regarding the kinds of impressions that people who occupy those roles should and should not convey. Furthermore, social norms may prescribe or proscribe how people in a particular social context should appear.

Second, people tailor their self-presentations to the values of the individuals whose perceptions are of concern. Sometimes people fabricate identities that they think others will value. More commonly, however, they selectively present truthful aspects of themselves that they believe coincide with the values of a particular target and withhold information about themselves that they think others will value negatively. In this way, people can present accurate impressions of themselves that are nonetheless tailored to the preferences and values of other people.

Third, people's choices of public images are influenced by how they think they are perceived currently. Often, self-presentational behaviors are aimed at dispelling undesired impressions that other people currently hold of the individual. When people believe that others have, or are likely to develop, an undesired impression of them, they often try to refute the undesired impression by showing themselves to be different from what others believe. When refutation of the negative impression is not possible, they may compensate by projecting particularly desired impressions on other, unrelated dimensions.

Fourth, the kinds of impressions people try to create are influenced not only by the social context but also by their own self-concepts. People often want other people to see them as they really are, but must deliberately manage their impressions in order to be viewed accurately. In addition, people's self-concepts constrain the images they try to convey. Not only do many people believe that it is unethical to present images of themselves that are inconsistent with how they really are, but they usually doubt that they will successfully sustain public images that are inconsistent with their actual characteristics. The risk of self-presentational failure, and the social sanctions that accompany it, deter people from presenting impressions that are vastly discrepant from how they see themselves.

Finally, the images that people try to project are also influenced by their desired and undesired selves—how they wish to be and not to be. People tend to manage their impressions in ways that coincide with their desired selves and that steer away from their undesired selves.

3. *Self-presentation and Behavior*

When people are in the presence of others, few of their behaviors are unaffected by their desire to maintain certain impressions in others' eyes. Even when people are not expressly trying to create a particular impression, prevailing concerns with their public images constrain their behavioral options. Whatever else people may be doing, they typically do it in ways that don't create undesired impressions. Concerns with others' impressions impact virtually all areas of social life.

3.1 *Interpersonal Behavior and Emotion*

In an early effort to offer a self-presentational explanation of interpersonal behavior, Tedeschi et al. (1971) argued that phenomena that had previously been attributed to the need to maintain cognitive consistency actually reflected people's efforts to maintain an impression of consistency in other people's eyes. Several studies subsequently supported their suggestion that phenomena that had been attributed previously to cognitive dissonance were affected by self-presentational processes (Schlenker 1980).

Self-presentational perspectives have been applied to the study of a variety of social psychological phenomena, including conformity, aggression, prosocial behavior, leadership, negotiation, social influence, gender, stigmatization, and close relationships (Baumeister 1982, Leary 1995, Schlenker 1980 Tedeschi 1981). In each instance, research has shown that these phenomena are affected, in part, by people's efforts to make particular impressions on others. In group settings, for example, people in positions of leadership have been shown to tailor their public images to the current demands of the group's situation. Research also shows that prosocial behavior is affected by self-presentational motives; people are more likely to be helpful when their helpfulness is publically known rather than anonymous, and they behave more prosocially when they desire to repair a damaged social image by appearing helpful. Likewise, many instances of aggression are self-presentational efforts to show that the individual is willing to hurt others to get his or her way. Many differences in the behavior of men and women can be traced to gender-biased norms regarding the kinds of impressions that are appropriate for men versus women. In brief, self-presentational concerns pervade social life.

Self-presentational processes also affect emotional experience. People become socially anxious when they are motivated to make a desired impression on others but doubt that they will do so successfully. Whenever the self-presentational stakes are high, doubts in one's self-presentational efficacy evoke anxiety. The nervousness that people experience on first dates, in job interviews, and while speaking in public can be traced to self-presentational concerns. Furthermore, when events occur that have undesirable implications for people's public images—showing them to be less competent, likeable, ethical, or otherwise socially desirable than they desire—people experience embarrassment. The face-saving behaviors that accompany episodes of embarrassment are efforts to undo whatever self-presentational damage has occurred (Leary 1995, Schlenker 1980).

3.2 *Mental and Physical Health*

Impression management processes have also been studied in the context of mental and physical health. The self-presentational use of psychological symptoms has received considerable attention. Braginsky et al. (1969) showed that hospitalized schizophrenics modify the severity of their disordered behavior depending on whether making a disturbed or nondisturbed impression will be most beneficial to them. Research on samples of university students likewise shows that people will exaggerate, if not fabricate, reports of psychological distress when doing so serves their social goals. Similarly, hypochondriasis appears to have self-presentational features: people convey impressions of illness, injury, and infirmity when doing so helps to achieve desired outcomes, such as excusing poor performance, eliciting support, or avoiding onerous responsibilities.

Self-presentation is also a risk factor for physical injury and illness. People engage in a number of dangerous behaviors for self-presentation reasons. Risky behaviors as diverse as suntanning, steroid use, excessive dieting, practicing unsafe sex, and fast driving can be traced to efforts to make certain impressions on other people. Furthermore, concerns with one's impressions may deter people from seeking medical attention when needed because of embarrassment about the presenting condition or about the medical examination itself. People may even refuse to undergo medical treatments if doing so compromises their public images (Leary et al. 1994).

See also: Control Behavior: Psychological Perspectives; Person Perception, Accuracy of; Personality and Conceptions of the Self; Self-conscious Emotions, Psychology of; Self-evaluative Process, Psychology of; Self-monitoring, Psychology of

Bibliography

Baumeister R F 1982 A self-presentational view of social phenomena. *Psychological Bulletin* **91**: 3–26

Bragnisky B M, Braginsky D D, Ring K 1969 *Methods of Madness: The Mental Hospital as a Last Resort*. Holt, Rinehart, Winston, New York

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, New York
Jones E E 1964 *Ingratiation*. Appleton-Century-Crofts, New York
Jones E E, Pittman T S 1982 Toward a general theory of strategic self-presentation. In: Suls J (ed.) *Psychological Perspectives on the Self*. Erlbaum, Hillsdale, NJ, Vol. 1, pp. 231–62
Leary M R 1995 *Self-presentation: Impression Management and Interpersonal Behavior*. Westview Press, Boulder, CO
Leary M R, Kowalski R M 1990 Impression management: A literature review and two-component model. *Psychological Bulletin* **107**: 34–47
Leary M R, Tchvidjian L R, Kraxberger B E 1994 Self-presentation may be hazardous to your health. *Health Psychology* **13**: 461–70
Schlenker B R 1980 *Impression Management: The Self-concept, Social Identity, and Interpersonal Relations*. Brooks/Cole, Monterey, CA
Tedeschi J T (ed.) 1981 *Impression Management Theory and Social Psychological Research*. Academic Press, New York
Tedeschi J T, Schlenker B R, Bonoma T V 1971 Cognitive dissonance: Private ratiocination or public spectacle? *American Psychologist* **26**: 685–95

M. R. Leary

Imprisonment: Sociological Aspects

Most people have an opinion about imprisonment: prison is a place for criminals, a room behind solid walls and bars, a nasty place for nasty people. But, as so often, complexities of meaning emerge with reflections and studies. To start with, an examination of the roots of the concept introduces new associations. In Old French, two meanings of the term exist. The first is associated with being apprehended, or held captive. The second meaning has to do with considering a person as being a capture, a prize, a person with a value for other persons.

The first sense, that of being captured, might also have several meanings. It might be understood as being captured in the small room behind walls and bars. But a single parent living on the tenth floor with three sick children at home might also experience a sense of being captive there. People in a caste or rigid class society might also have opinions on what it is to be captured in one's position in life, as might remote villagers watching life in London on the BBC. And so might the electronically tagged factory worker whose bracelet would trigger an alarm in the police or probation service if he did not go straight home after work. Imprisonment is not only in small rooms. Whole societies, or segments of them, can experience the meaning of being in prison. Attempts to grasp the phenomenon of imprisonment open up renewed perspectives on what freedom is. A prisoner on hunger strike, naked in a punishment cell, was reported as saying that he had never felt so free, because there was nothing more to lose.

Prisons can also contain prizes, valuable to those who keep them. Rich or important people are of value to those who capture them. They can be exchanged for money or political privileges. Whole segments of populations might be valuable to power holders. Slaves are often useful for labor-intensive tasks. They were also useful in the war industry both in the USSR and Germany. Ordinary prisoners can also be seen as precious. Large segments of societies make a living from keeping them captured.

1. The Penal Picture

Moving back to the phenomenon as most people in the Western world experience it, prison is a place for those waiting to receive a sentence, or having received such a sentence. A useful measure of prison populations in a country is the number of prisoners per 100,000 inhabitants. There are many alternative indicators, e.g., the flow through the system, but it is difficult to find reliable comparative data for this flow. The figures presented here include both people waiting for trial and prisoners who have received their sentence. Most figures are for year 2000. Sources are Walmsley (2000) and Christie (2000).

Table 1 gives the information for Western Europe. As a general tendency, the Scandinavian countries have low prison populations, none of them with more than 70 prisoners per 100,000 inhabitants. At the top in Western Europe is Portugal, with the UK in second place.

Table 2 demonstrates the dramatic differences between Western Europe and Central and Eastern Europe. The Russian Federation is at the very top in

Table 1
Prisoners per 100,000 inhabitants in Western Europe

Portugal	131
UK	124
Spain	116
Germany	97
Italy	93
Switzerland	89
Austria	85
The Netherlands	87
Ireland	78
France	75
Sweden	64
Denmark	62
Norway	59
Greece	55
Finland	52
Iceland	29

Table 2
Prisoners per 100,000 inhabitants in Central and Eastern Europe

Russian Federation	640
Belarus	550
Ukraine	436
Latvia	410
Lithuania	355
Estonia	320
Moldavia	266
Romania	221
Czech Republic	216
Hungary	156
Slovakia	135
Bulgaria	115
Turkey	108
Slovenia	57

Table 3
Prisoners per 100,000 inhabitants in the American continent

USA	709
Belize	459
Chile	375
Panama	303
Costa Rica	229
Honduras	172
Colombia	153
Mexico	148
Nicaragua	143
Brazil	141
Uruguay	123
El Salvador	109
Argentina	107
Peru	107
Canada	103
Bolivia	102
Venezuela	98
Paraguay	75
Guatemala	69
Ecuador	69

Europe with no less than 640 prisoners per 100,000 inhabitants. Close to Russia are the former members of the USSR, also with extremely high prison figures compared to Western Europe. Particularly surprising is the high number of prisoners in the Baltic countries, Latvia, Lithuania, and Estonia. But some of the countries from the former Eastern block are by now abandoning the tradition of having a large prison population. This is particularly the case for Poland, Hungary, and Slovakia, and most of all for little Slovenia, with only 57 prisoners per 100,000, similar to the Scandinavian countries.

Table 3 gives the figures for the American continent. The USA is at the top, in the table as in the world. For the other countries, there is great variations in the use of prisons. Particularly in South America there are great similarities to the figures for Western Europe.

2. *Behind the Variance*

Confronted with the differences between countries in these tables, the need for an explanation becomes urgent. The conventional view is that the relative differences in the number of prisoners are related to differences in crime level. But this explanation has two major flaws. First, countries otherwise rather similar might differ when it comes to the relative number of prisoners, while other very different ones might have similar numbers of prisoners. The difference between Canada and the USA is a point in case. Two countries, so close, with a joint border, the same language for most people, the same type of media and economic system, and to a considerable extent even common views about money—but still this enormous difference in the number of prisoners. While Russia and the USA, with all their differences, have nearly exactly the same relative number of prisoners. A second element which weakens the 'crime' explanation is the historical development behind many of these figures. Russia decreased its prison population from 1,400 per 100,000 in 1950 to 353 in 1989 under Gorbachev. But since then they have again increased, and since Glastnost have more than doubled their relative number of prisoners. They now have one million prisoners. It was not crime that changed, but politics. This was also the case with Poland and Hungary after the collapse of the USSR. Finland had a similar development in the 1960s. The US figures departed in 1980 from a level much closer to the European, but later exploded, an explosion not reflected in indicators of their crime situation. This growth shows no indication of flattening out. In 1999, the increase was 110,000 new prisoners. In 2000 the USA passed the two million mark.

3. *Prison Life*

On a global scale, Russia and the USA stand out as the big incarcerators. They are similar in their large numbers of prisoners, but extremely different in living conditions within the prisons. In both cases, prisons are dangerous places to live in, but more so in Russia. Of the one million prisoners there, 92,000 have tuberculosis—30,000 having a multiresistant type. Those waiting for trial live under particularly harsh conditions. Overcrowding is extreme. In some prisons there is not sufficient space for all prisoners to lie down at the same time. After trial, prisoners are sent to the so-called 'colonies,' the former Gulags. Here, material conditions are better, but social life is often filled with strains. The internal life in these colonies is often

characterized by extreme internal differentiation, with a caste system developing in many camps. Some prisoners gain a large amount of power; they become the internal commanding force with privileges and comfort. At the bottom, there is a pariah caste living under the most miserable of conditions, eating leftovers and serving the power holders in all ways, including sexually.

While overcrowding and disease are major problems in the Russian situation, extreme forms of isolation are a problem in a number of the prisons in the USA. This is particularly the case in the so-called maxi-maxi prisons, where inmates are often isolated in windowless cells, built from solid blocks of concrete and stainless steel. Metal doors of each cell open to the inside corridor and do not allow prisoners to communicate between cells. Smaller doors within each door allow officers to pass in food and to shackle prisoners before opening the cell door. In some of these prisons, each cell has its own toilet, sink, and shower. They also have doors leading outside to very small enclosed balconies where inmates can take their daily one hour of exercise. The doors to the balconies can be controlled electronically from a central control room. There is very limited possibility for the prisoner to meet any other human beings.

These descriptions do not do justice to all prisons or colonies, either in Russia or in the USA. In both countries well-ordered prisons and colonies can also be found, with considerable contact both between inmates and between inmates and staff, and where self-government among prisoners seems to function without abuse.

4. *The Goal of the Operation*

With the enormous increase in the prison population in major industrial countries, attempts to understand this whole area gain a new focus. Penal law and sociology has an old, learned tradition in studies of the reasons for punishment. Theories of deterrence, treatment, or incapacitation were compared and evaluated. A great number of studies tried to compare the effects of various measures. But a shift in research interests seems to be under way. Theories on reasons for punishment and the treatment effects of various measures are not quite as interesting when it becomes clear that the political system puts people into prison for reasons far outside any of the main reasons for punishment in the old theories. Studies such as the one on 'administrative criminology' by Feeley and Simon (1994) become more relevant.

From the 1950s to the 1980s, much energy was also spent on studies of the inside life of relatively well-ordered prisons. Sykes (1958) studied the society of captives, the roles that appeared within prison, and the power structure between prisoners and their guards. Cohen and Taylor (1972) described the survival mechanism of long-term prisoners, and conveyed not only their misery and despair, but also the great resources and the will to survive found among those contained for a life without future. Goffman (1961) helped us to see the general character of various types of total institutions. But sophisticated studies of life within well-ordered prisons are not in themselves able to explain the recent situation where prison figures in leading world powers are exploding, and the suffering involved is so overwhelming. Instead, new types of studies appear, such as Stern (1998) on the global situation and Walmsley (1998) on the World Prison Population. Highly relevant links between the old penal situation and the new can be found in Foucault (1977), Cohen (1985), and Mathiesen (1990). Imprisonment is a reflection of the social structure. Prisons are there to capture segments of the population, particularly those seen as superfluous, those not able to participate in the consumer society (Bauman 1998). In their captured position, they are also a prize—they are here given a use function as raw material for the control industry (Christie 2000). Studies of the crime control industry provide an indicator of the importance of the protection of civil society.

See also: Control: Social; Crime and Class; Crime and Ethnicity (Including Race); Crime, Sociology of; Criminal Justice, Sociology of; Foucault, Michel (1926–84); Law: Economics of its Public Enforcement; Prison Life, Sociology of; Prisons and Imprisonment; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Rehabilitation of the Offender

Bibliography

Bauman Z 1998 *Globalization. The Human Consequence*. Polity Press, Cambridge, UK

Christie N 2000 *Crime Control as Industry. Towards GULAG's Modern Type*. Routledge, London

Cohen S 1985 *Visions of Social Control. Crime, Punishment and Classification*. Polity Press, Cambridge, UK

Cohen S, Taylor L 1972 *Psychological Survival. The Experience of Long-Term Imprisonment*. Penguin Books, Harmondsworth, UK

Feeley M, Simon J 1994 Actuarial justice: The emerging new criminal law. In: Nelken D (ed.) *The Future of Criminology*. Sage, London, pp. 173–201

Foucault M 1977 *Discipline and Punish. The Birth of the Prison*.

Goffman E 1961 *Asylums. Essays on the Social Situation of Mental Patients and Other Inmates*. Anchor Books, New York

Mathiesen T 1990 *Prisons on Trial. A Critical Assessment*. Sage, London

Sykes G M 1958 *The Society of Captives. A Study of a Maximum Security Prison*. Princeton University Press, Princeton, NJ

Stern V 1998 *A Sin Against the Future. Imprisonment in the World*. Penguin Books, Harmondsworth

Walmsley R 2000 World Prison Population List. *Home Office Research, Development and Statistics Directorate. Research Findings No 116*. London (and also continuously updated at International Centre for Prisons Studies: http://www.kcl.ac.uk//depsta/rel/icps/home.html)

N. Christie

In Vitro Neurophysiology

1. *Background*

Neurophysiology is the study of the functional properties of neurons, glia, and networks. Historically it has been dominated by electrophysiology—the electrical recording of neuronal events ranging from the molar (the electroencephalogram, EEG) to the cellular (intracellular recording of the properties of single neurons). However, as the neuron is an electrochemical machine, it is impossible to separate electrical events from the biochemical and molecular processes that bring them about. Thus neurophysiologists today use techniques from chemistry (calcium imaging), physics (functional magnetic resonance imaging, fMRI), and molecular biology (site directed mutations) to study brain function. As we shall see, the use of *in vitro* techniques has facilitated all aspects of modern neurophysiology and can be fairly stated to have permitted many of them.

In vitro literally means 'in glass' and here refers to keeping isolated small pieces of brain alive in a special chamber (usually made of plastic, not glass) that provides life support and permits electrophysiological access (stimulation and recording). *In vitro* neurophysiology has its origins in the biochemical studies on energy metabolism by Henry McIlwain and others in the 1950s (McIlwain 1984). They cut brain and other tissue into thin (< 0.5 mm) slices, which were kept alive in vials of artificial cerebrospinal fluid (ACSF) from which was extracted substances of interest. It was a small step to immobilize the slices on a net and place recording and stimulating electrodes into them to determine if they exhibited 'normal' activity. To the great surprise of some, the neurophysiological responses they exhibited were remarkably normal in spite of the trauma of extracting and slicing the delicate tissue.

In vitro approaches are not appropriate for all neurophysiological studies. As the slice is removed from the network of which it is normally a part, it follows that studies of network properties are generally inappropriate for *in vitro* approaches. Most slice preparations preserve monosynaptic junctions well and allow for the controlled neurophysiological study of pre- and/or postsynaptic properties of the preserved synapse. As a consequence slices are best suited for cellular neurophysiology, as opposed to systems neurophysiology.

2. *The* In Vitro *Technique*

Two aspects of preparing brain slices (Alger et al. 1984) are of utmost importance: being gentle in handling the tissue (to minimize trauma) and doing so rapidly (to avoid ischaemic damage). Brain slices are prepared from living animals either under general anesthesia or following rapid cervical dislocation. The whole brain is removed and the area of interest blocked and transferred to a cutting device. Brain slices are typically cut 400 μm thick with a manual slicer or a Vibratome. Thicker slices will experience difficulty with O_2 diffusion into the center of the slice, whereas thinner slices will behave abnormally due to damage to inhibitory local circuit neurons. Once cut, slices are either directly transferred to a slice chamber, or are placed in a holding chamber for use later. In both cases the slice is exposed to warmed (usually) and oxygenated ACSF. Investigators have used a wide variety of conditions to maintain brain slices, differing principally in temperature, ACSF composition, and slice chamber design.

A basic brain slice chamber (Fig. 1) consists of a well holding one or more slices, a means of maintaining a constant temperature, a source of oxygen and ACSF, along with a binocular dissecting microscope to view the transilluminated slices, and recording/stimulating electrodes, their manipulators, and associated electronics. Chambers maintain slices either on a net at the interface of warmed, oxygenated ACSF and humidified 95 percent O_2/5 percent CO_2, or totally submerged in warmed and oxygenated ACSF and held in position between two nets. Which slice chamber is superior is a

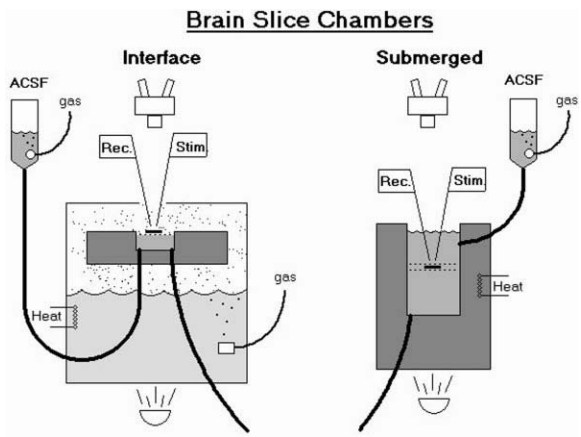

Figure 1
Brain slice chambers place thin (400 μm) brain slices either at the interface of bathing ACSF and oxygenated atmosphere or submerge the slices completely in oxygenated ACSF. Both varieties have their adherents and supply an appropriate physical and chemical environment to allow the tissue to remain relatively normal for periods ranging up to 10 hours

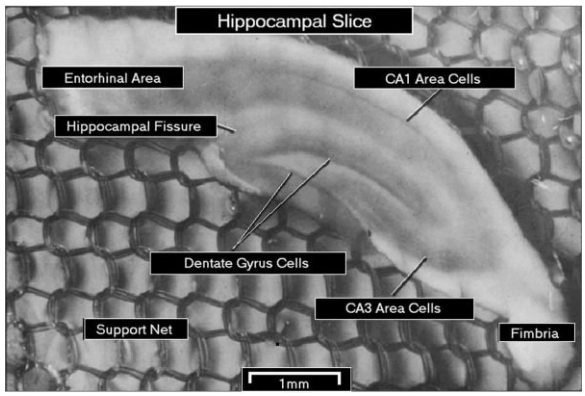

Figure 2
The hippocampus has been often employed for *in vitro* studies of cellular and synaptic physiology and plasticity due to its unique anatomy. The *in vitro* hippocampus in the photograph is a slice cut transverse to the long axis of the hippocampus and preserves a three synapse network (dentate gyrus > CA3 > CA1)

subject of debate and controversy. Studies exposing slices to drugs often prefer the submerged design as it offers fast drug access to both surfaces of the slice. However ACSF flow rates are generally higher in submerged designs (interface slices obtain O_2 from the atmosphere), resulting in greater washout and, it is claimed, less healthy slices. The temperature at which slices are maintained varies from room temperature ($\sim$ 22°C) to physiological body temperature (37°C), with most studies being done at 30 to 35°C. Given that most mammalian systems are optimized to body temperature, surprisingly little difference is noted across the temperatures used. In general, slices are viable longer at lower temperatures. ACSF composition also varies across laboratories, principally in the concentration of K^+ and Ca^{2+} and whether or not drugs are added to attenuate inhibitory influences. The effect of these ACSF differences is reasonably well understood.

Between the time the blood supply is interrupted and the slice is exposed to oxygenated ACSF it experiences anoxia. To prevent ischaemic damage during this period (particularly important for neocortical tissue) investigators have employed several tactics. The use of general anesthetic agents offers considerable protection from ischaemia. To further reduce metabolism during the slice preparation some investigators use whole-body cooling prior to extracting the brain, often in conjunction with performing the dissection and transfer to the slice chamber over ice.

Slices have been made from a wide variety of brain structures. The first and by far the most common tissue studied *in vitro* is the hippocampus (Fig. 2, Teyler 1980). The hippocampus is particularly well suited because of its unique anatomy. It is a primitive, three-layered cortical tissue wherein the principle cell types are found in dense sheets and whose dendritic and axonal processed are highly ordered. The experimenter can visually guide electrodes into specific cell body or synaptic areas with a high degree of confidence and with an understanding of the origins and properties of the fibers being stimulated and cells from which recordings are made. In less ordered tissue, like neocortex and subcortical nuclei, visual cues are largely absent and their heterogeneous anatomy makes the precise identification of what is being stimulated less certain. The hippocampus is also unique in that its slices preserve a three-synapse circuit, due to the laminar anatomy of this structure. Most other brain areas do not lend themselves to preserving a polysynaptic circuit within the two-dimensional brain slice. While this can sometimes be addressed by preparing slices with curved cutting blades or by manipulating the tissue on the cutter, as a rule most other regions are limited to the study of single synapses. While this is a limitation to the study of networks, it is an advantage in the detailed cellular study of specific synaptic nodes in a network.

3. *Advantages*

Neurophysiological studies of neurons and circuits *in vivo* must deal with the inaccessibility of the tissue (protected by the cranium) and its complexity (which renders interpretation difficult). Neuroscientists have approached these two issues by adopting model systems—animals or preparations displaying less complexity and greater access. One approach is to use organisms with simpler nervous systems, such as *Aplysia* and *Hermissenda*, or surgically reduced preparations, such as isolated spinal cord, sub-pial cortical islands, and isolated cells maintained in culture. Both approaches have been extremely useful in the study of cellular neurophysiology because they have reduced the complexity of the tissue to a more manageable level while greatly improving experimental access. The former, however are nonmammalian organisms whose evolution may favor different cellular processes making extension to mammalian systems problematic, while the latter are cells that have either undergone development in a unique environment or, in the case of primary cultures, have undergone the trauma of isolation and plating into a culture system where they generally do not form patterns of connectivity as *in vivo* and may be exposed to a foreign chemical environment which results in an altered phenotype.

The *in vitro*, or brain slice, technique is also an example of a surgically simplified preparation, but an acute one made from forebrain tissues in most cases. Such a preparation can be made to preserve some of

the normal connections of the neurons under study and renders them exquisitely accessible for experimentation. The removal of a synapse from the network in which it is found greatly reduces the confounding influence of the network and facilitates the detailed study of synaptic/membrane function. In addition to simplifying the tissue, it is made two-dimensional, thus obviating the need for stereotaxic devices and their inherent inaccuracies (requiring histological verification of electrode placements) to place recording and stimulating electrodes accurately. In slice preparations the tissue is typically directly observed via microscope and electrodes are directed to structures or areas of interest under visual control. *In vivo* cellular neurophysiological studies seek to record from single neurons (extracellularly or intracellularly) or small groups of nearby neurons (field recordings) and are plagued by problems of movement derived from normal cardiac and respiratory activity. While ingenious solutions have been devised to permit such studies *in vivo*, these problems are totally absent *in vitro*, which represents a major advantage. While these practical matters are experimentally advantageous, the slice approach potentially offers more significant advantages in terms of the degree of experimental control the investigator possesses. As the investigator determines the composition of the bathing ACSF, its composition can be manipulated at will. *In vivo* studies do not have the experimental luxury of, for example, changing the extracellular composition of K^+ or Ca^{2+}, as can be done readily *in vitro*. Similarly, the introduction of drugs or compounds *in vivo* involves first breaching the blood–brain barrier and then dealing with uncertainties of drug distribution, concentration, liver degradation, and actual locus of drug action. In slice preparations drugs can be applied in known concentration with guaranteed access by superfusion (adding the drug to ACSF) or by pressure ejection or iontophoresis through a pipette. Drugs toxic to whole animal preparations (e.g., enzyme inhibitors, metabolic blockers) can be safely delivered to isolated brain slices. Similar experimental advantages are seen with respect to temperature and pH.

Often neurophysiological studies are combined with subsequent histological studies, for instance, to identify cells expressing particular markers. In slice preparations such experiments can be done online by exposing the slice to appropriate illumination revealing the cells possessing that feature and allowing physiological manipulations to be done only to those cells. Similarly, intracellular (whole cell or sharp electrode) recordings can be directed to individual cells possessing particular morphologies by utilizing infrared microscopy (Jaffe and Brown 1997). The combination of electrophysiological and optical technologies is particularly well suited to slice preparations. The spatial and temporal dynamics of $[Ca^{2+}]_i$ are important for understanding many aspects of synaptic transmission, synaptic plasticity, and cell death; and their study in hippocampal and neocortical pyramidal cell dendrites has relied heavily on *in vitro* approaches (Perkel et al. 1993).

4. *Limitations*

Like all simplified preparations, brain slices possess limitations. A fundamental limitation is that slices have a limited lifespan—generally on the order of 8 to 10 hours, although longer survival times are possible. The limited viability of brain slices is advantageous in that they do not have the time to alter their structure and function in response to their changed environment, as gene expression, protein synthesis, and transport generally require longer times (but see Kirov et al. 1999). Slices do not suddenly die; rather their behavior gradually deteriorates, suggesting the extreme view that slice neurophysiologists are studying slowly dying tissue. A window of relative normalcy has been documented, but its limited duration means that studies that require extended observation (gene expression, protein synthesis, growth, repair) cannot generally be done *in vitro*. To address this temporal limitation the *ex vivo* slice preparation has been developed. In this approach a manipulation is performed *in vivo* and brain slices made after an appropriate period of time (days to weeks). This allows the process under study to go to completion before the slice is prepared (or slices can be obtained at intervals to track the process). A further variant places standard brain slices in a tissue culture environment where the slice undergoes simplification while retaining connectivity, thus allowing the *in vitro* study of long timecourse events (Gahwiler et al. 1997). These long-term organotypic slice cultures are subject to the same kinds of alterations in gene expression (and thus structure and function) experienced by any cells maintained in culture and deprived of their normal *in vivo* environment. Brain slices are maintained in a chamber (described in Sect. 2) that is usually continually bathed with fresh ACSF. The flowing ACSF, however, can wash out extracellular modulatory agents (like neurotrophins) normally present and released by the tissue.

Several of the advantages of brain slices are also limitations. While surgical simplification reduces complexity it also removes the tissue from normal synaptic and extracellular influences. Thus the tissue may not behave as *in vivo*. As the investigator controls the ACSF environment, an incomplete understanding of the roles of compounds normally present *in vivo* and thus generally absent in *in vitro*, coupled with the washout of compounds released from slices, has unknown effects on cellular function.

A caution in the use of brain slices is that certain phenomena are readily observed *in vitro* but difficult to duplicate *in vivo*, and vice versa. Examples include the establishment of a seizure focus (facilitated *in vivo*), the

expression of long term-potentiation (facilitated *in vitro*), the expression of long-term depression (facilitated *in vitro*), and the presence of spontaneous activity (facilitated *in vivo*). While many of these differences are understood, they serve to caution that the *in vitro* preparation is not the same as an *in vivo* preparation. A recent example is the observation that pyramidal cells from hippocampal slices possess almost twice as many dendritic spines as their *in vivo* counterparts—a difference attributed to decreased spontaneous synaptic activity *in vitro* (Kirov et al. 1999).

5. *Future Uses of* In Vitro *Preparations*

In vitro preparations have been employed for a wide variety of investigations and will continue to be in the twenty-first century as new brain areas are studied using this technology. Three areas of emerging research using *in vitro* techniques seem destined to become more widespread. First, as greater understanding emerges regarding specific neural networks from functional imaging studies of intact organisms, *in vitro* methods will be ideally suited to characterize and analyze the cellular and synaptic properties of individual elements of the network. Second, brain slices will be increasingly employed for approaches that utilize optical techniques in which a thin, relatively normal, two-dimensional slice are ideal. As optical methods become more precise (e.g., more sensitive detection methods, smaller pixel sizes to improve resolution) and as new dyes are developed to visualize new molecules of interest, *in vitro* approaches will be adopted more and more for purely morphological experiments. A third area for *in vitro* approaches lies in molecular neurobiology. As molecular approaches allow for the elimination or over-expression of a wide variety of specific proteins, brain slices will be employed to ferret out the neurophysiological functions of these molecules.

See also: Cortical Activity: Differential Optical Imaging; Hippocampus and Related Structures; Ion Channels and Molecular Events in Neuronal Activity; Lesion and Behavior Approaches in Neuroscience; Long-term Potentiation (Hippocampus); Long-term Depression (Cerebellum); Long-term Depression (Hippocampus); Neurotransmitters; Perception and Cognition, Single-/Multi-neuronal Recording Studies of; Synapse Formation; Synaptic Efficacy, Regulation of; Synaptic Transmission

Bibliography

Aitken P G et al. 1995 Preparative methods for brain slices: A discussion. *Journal of Neuroscience Method* **59**: 139–50

Alger B E, Dhanjal S S, Dingledine R, Garthwaite J, Henrerson G, King G L, Lipton P, North A, Schwartzkroin P A, Sears T A, Segal M, Whittingham T S, Williams J 1984 Brain slice methods. In: Dingledine R (ed.) *Brain Slices*. Plenum Press, New York, pp. 381–438

Dingledine R 1984 *Brain Slices*. Plenum Press, New York

Gahwiler B H, Capogna M, Debanne D, McKinney R A, Thompson S M 1997 Organotypic slice cultures: A technique has come of age. *Trends in Neuroscience* **20**: 471–7

Jaffe D B, Brown T H 1997 Calcium dynamics in thorny excrescences of CA3 pyramidal neurons. *Journal of Neurophysiology* **78**: 10–18

Kirov S A, Sorra K E, Harris K M 1999 Slices have more synapses than perfusion-fixed hippocampus from both young and mature rats. *Journal of Neuroscience* **19**: 2876–86

Lipton P et al. 1995 Making the best of brain slices: Comparing preparative methods. *Journal of Neuroscience Method* **59**: 151–6

McIlwain H 1984 Introduction: Cerebral subsystems as biological entities. In: Dingledine R (ed.) *Brain Slices*. Plenum Press, New York, pp. 1–5

Perkel D J, Petrozzino J J, Nicoll R A, Conner J A 1993 The role of Ca^{2+} entry via synaptically activated NMDA receptors in the induction of long-term potentiation. *Neuron* **11**: 817–23

Schurr A 1999 Preparation and use of brain slices. *Methods: A Companion to Methods in Enzymology* **18**: 109–16

Schurr A, Rigor B M 1995 *Brain Slices in Basic and Clinical Research*. CRC Press, Boca Raton, FL

Teyler T J 1980 The brain slice preparation: Hippocampus. *Brain Research Bulletin* **5**: 391–403

T. J. Teyler

Incest, Inbreeding, and their Consequences

The terms inbreeding and consanguinity are used to describe unions between couples known to share at least one common ancestor. While now rare in Western societies, marriages between close biological kin are preferential in many parts of the world, including North and sub-Saharan Africa, the Middle East, the Turkic- and Farsi-speaking countries of Central and South Asia, and much of the Indian subcontinent. Although the rates and types of inbred union may vary according to religious and societal norms, marriage between first cousins is especially common and in a country such as Pakistan accounts for over 50 percent of all marital unions. An incestuous relationship is a union between biological relatives that is closer than permissible under prevailing legislation. Most commonly, incest is defined as mating between first-degree relatives, i.e., father–daughter, mother–son, or brother–sister, who have 50 percent of their genes in common. However, in some countries the definition may be widened to include half-sib and uncle–niece unions.

1. *Human Inbreeding*

Many examples of successful consanguineous unions are cited in the biblical texts, including the marriage of Abraham with Sarah his half-sister (Genesis 20:12), and Amram and Jochebed, the parents of Aaron and

Moses who were nephew and aunt (Exodus 6:20). At a later date the permitted degrees of marital relationships were extensively defined, with marriages up to and including uncle–niece permitted (Leviticus 18:7–18). These guidelines were generally adopted by the Protestant churches after the Reformation, although in most denominations the closest form of approved union has been between first cousins. For Roman Catholics, diocesan dispensation has been and remains a prerequisite for marriage between first cousins, and the Christian Orthodox church adopts a similarly proscriptive stance even with more remote relationships. Among the other major world religions, attitudes to close kin marriage are generally favorable or neutral in Buddhism, Confucianism, Islam, and Judaism. However, for the Aryan Hindus of North India, pedigrees are examined over the preceding seven generations for males and five generations for females to ensure avoidance of a consanguineous union.

1.1 Attitudinal Changes in the Response to Consanguineous Marriages

Prior to the mid-nineteenth century, first-cousin marriages were commonly contracted in many Western societies, including countries such as Spain and Italy where a large majority of the population was Roman Catholic. From the 1850s onward an often acrimonious debate commenced within the scientific and medical communities of North America and Western Europe into the biological effects of close kin marriage, with Charles Darwin, who had married his first cousin Emma Wedgwood, a leading protagonist (Bittles 1994). In Europe, and in most countries that had been colonies of the major European powers, this debate resulted in few if any legislative changes. However, within the USA the eventual outcome was the passage of laws at state level to control, and in many cases to ban, first-cousin unions, even though the information on which these laws were based frequently appeared to be more emotive than factual in nature. Michigan, New Jersey, and Ohio subsequently rescinded their prohibition on first-cousin marriages, but they remain illegal in twenty-two states and are a criminal offense in eight others (Bratt 1984, Ottenheimer 1990). Interestingly, the legislation passed by some individual states recognized the traditions of specific constituent communities and so, for example, uncle–niece marriages (but not aunt–nephew) are permissible within the Jewish community of Rhode Island, while Native Americans in Colorado are free to marry their stepchildren.

1.2 Contemporary Societies in which Consanguineous Marriage is Preferential

As indicated in Sect. 1.1, the current legal situation in the USA differs markedly from that in many other parts of the world where consanguineous unions, a term derived from the Latin *consanguinitas* and indicating 'shared blood,' are strongly preferential. The reasons for this preference are mainly social and economic in origin, and include the strengthening of family relationships and the maintenance of family property and land-holdings. Prenuptial arrangements also are greatly simplified, and the security of marrying a partner whose entire family background is known is perceived as a major benefit in ensuring both the initial success and the lasting stability of the union. Variation in the specific types of marriage contracted, e.g., with father's brother's daughter first-cousin unions preferred in Arab Muslim communities as opposed to mother's brother's daughter in the Dravidian Hindu states of South India and among the Han Chinese, does however indicate additional customary influences.

1.3 The Genetic Effects of Inbreeding

In consanguineous unions the partners share genes inherited from one or more common ancestor and, for example, in first-cousin marriages the spouses are predicted to have 12.5 percent of their genes in common. This means that on average their progeny will be homozygous at 6.25 percent of gene loci, i.e., they will have received identical gene copies from each parent at these sites in their genome. If the same mutant gene is inherited from both parents the individual will express the disorder, either at birth or later in life depending on the nature and site of the mutation, thus contributing to the phenomenon of inbreeding depression.

The coefficient of inbreeding (F) is a numerical estimate of the degree of inbreeding of an individual, and so for first-cousin offspring $F = 0.0625$. Similarly, for the progeny of uncle–niece unions $F = 0.125$, while for second-cousin offspring $F = 0.0156$. In many communities there is a long and continuous history of consanguineous unions, and so the cumulative level of inbreeding may be significantly higher than the value calculated for a single generation. Under these circumstances, a correction can be applied to account for the effects of ancestral inbreeding using the formula:

$$F = \Sigma(1/2)^n\,(1 + F_A)$$

where F_A is the ancestor's inbreeding coefficient, n is the number of individuals in the path connecting the parents of the individual, and the summation (Σ) is taken over each path in the pedigree that goes through a common ancestor.

2. Incest

With a few notable exceptions, incest avoidance is, and appears to have been, a practice common to virtually all societies. Various explanations have been

offered for incest avoidance including Freudian guilt, and theories ascribed to Lévi-Strauss and others that it serves both to maintain the existence of the family and to encourage the establishment of affinal relations with other kin groups. According to the Westermarck hypothesis, incest avoidance arises principally because of negative imprinting against close associates of early childhood (van den Berghe 1983). Studies conducted on individuals raised under mixed-sex child-rearing regimes in Israeli *kibbutzim* have been cited in support of the negative imprinting theory, since they rarely marry or enter into sexual relationships (Talmon 1965, Spiro 1965). Perhaps more convincingly, a detailed investigation into the practice of *sim-pua* marriage in Taiwan, whereby a girl is adopted with the expectation that she will marry a son of the adoptive family, demonstrated aversion on the part of some intended spouses. Where a marriage did occur, *sim-pua* unions often were characterized by lower mean fertility and lesser marital stability (Wolf and Huang 1980).

2.1 Historical Examples of Dynastic and Non-dynastic Incest

There have been a number of societies in which dynastic incest was practised over multiple generations, including the Eighteenth and Nineteenth Dynasties and the Ptolemaic and Roman periods in Egypt, Zoroastrian Iran, the Incas, and the royal families of Hawaii (Middleton 1962, Shaw 1992). In Pharaonic Egypt, brother–sister or half-sister unions were regarded as a potent means of maintaining and strengthening the royal house and bloodline, although the Pharaoh also had ready access to other females who were non-relatives. Information has emerged on incestuous marriages outside these ruling classes. During the period of Roman Egypt, from the first to the third centuries AD: the records indicate that full brother–sister unions accounted for 19.6 percent of marriages in the city of Arsinoe, with a further 3.9 percent between half-siblings (Scheidel 1997). However, it has been argued that these unions may have been atypical, having been contracted in response to the restricted numbers of potential marriage partners in the minority Greek settler community. Among the Zoroastrians in Persia, the underlying rationale for incestuous unions appears to have been based on specific religious beliefs, and to have emerged just prior to the Arab invasion in 636 AD. All three types of incestuous union, father–daughter, brother–sister, and mother–son were advocated. In each case they were described as being of especial religious merit, with their practice a means of expiating mortal sin.

2.2 Incest in Contemporary Societies

There is the conundrum that, despite the claimed near-universality of an incest taboo, an extensive and growing literature exists on the occurrence of incest in contemporary Western societies. Most commonly these reports have emanated from case studies conducted either on persons examined because of intellectual handicap or in psychiatric clinics, thus leading to the general conclusion that incestuous relationships are highly detrimental to those classified as victims. In some cases, however, it has been suggested that the outcome need not necessarily be psychologically destructive, with women in stable marriages acknowledging earlier father–daughter sexual relationships which had commenced prior to puberty in the female. As will be discussed in Sect. 3.2, given the fact that incest is both illegal and widely regarded as morally reprehensible, it would be difficult to determine how common this latter response might be without exposing respondents to the risk of self-incrimination.

3. Biological Outcomes of Consanguinity and Incest

The first major study into the biological effects of inbreeding was conducted by Dr Samuel Bemiss of Louisville, Kentucky in 1858, using reports forwarded by medical colleagues on the outcomes of unions ranging from incest to third-cousin marriages. Since that time hundreds of additional studies have been undertaken based on a variety of sampling techniques, including pedigree analysis, household surveys, questionnaires administered to hospital in- and out-patients, Roman Catholic dispensation records, and surname analysis. In populations such as Dravidian South India, where uncle–niece ($F = 0.125$) and first-cousin unions ($F = 0.0625$) are preferential and jointly account for some 30 percent of marriages, unions beyond second cousins ($F < 0.0156$) are of limited biological significance (Bittles et al. 1991). Conversely, in a country such as Finland where consanguineous unions are rare, or in endogamous communities in which close-cousin marriage has been proscribed on religious grounds, biologically remote relationships including third-cousin unions ($F = 0.0039$) may result in moderate levels of cumulative inbreeding through time and so prove to be of clinical relevance.

3.1 Measuring the Biological Outcomes of Inbred Marriages

In an attempt to rationalize the results of inbreeding surveys into a biologically meaningful and reproducible format, a method was devised to calculate the numbers of lethal gene equivalents in a community or population, by comparing death rates in the progeny of consanguineous and unrelated couples. Lethal gene equivalents are the numbers of detrimental recessive genes carried by an individual in the heterozygous state, which, if homozygous, would result in

Table 1
The outcomes of first-degree incestuous unions

Country of origin	No. studied	Follow-up (year)	Death and severe defect	Other defect/ mild intellectual handicap	Normal	Source
USA	18	0.5	6	5	7	Adams and Neel (1967)
UK	13	4–6	4	4	5	Carter (1967)
Czechoslovakia	161	1–37	65	18	78	Seemanová (1967)
Canada	21	0.5–19	9	4	8	Baird and McGillivray (1982)
Totals	213		84 (39.4%)	31 (14.6%)	98 (46.0%)	

death and so add to the genetic load of the population. The number of lethal gene equivalents in a population can be calculated according to the formula:

$$-\log_e S = A + BF$$

where S is the proportion of survivors in the study population, A measures deaths that occur under random mating, B represents deaths caused by the expression of recessive genes via inbreeding, and F is the coefficient of inbreeding (Morton et al. 1956). By plotting a weighted regression of the log proportion of survivors (S) at different levels of inbreeding (F), A can be determined from the intercept on the Y-axis at zero inbreeding ($F = 0$), and B (the number of lethal gene equivalents) is given by the slope of the regression.

Despite a number of limitations, this regression technique offers a simple and convenient method of assessing and comparing data on the effects of inbreeding in different populations. Through time, estimates of the mean number of lethal gene equivalents in the genome have been revised downwards, in part resulting from improved sampling techniques and the recognition that earlier surveys may have produced spuriously high values because of inadequate control for non-genetic variables. Most recently, an excess mortality rate of 4.4 percent was calculated for deaths among first-cousin offspring from approximately 6 months gestation to 10 years of age, equivalent to 1.4 lethal genes per person (Bittles and Neel 1994). The measurement of postnatal morbidity is less precise since it is reliant on the diagnostic criteria employed. As a result the estimates obtained can vary quite widely, e.g., from 1.4 percent to 4.1 percent for total morbidity prior to the age of reproduction (Schull and Neel 1965).

3.2 The Biological Outcomes of Incest

As first degree relatives have 50 percent of their genes in common, the progeny of an incestuous union will be homozygous at 25 percent of loci and so $F = 0.25$. Given the strongly antagonistic legal and social attitudes to incest, direct assessment of associated biological outcomes may be subject to significant ascertainment bias, since a detailed examination including determination of paternity may only be initiated if a child shows symptoms of physical and/or intellectual handicap. The net result is a paucity of data, which may or may not reflect the incidence of incestuous pregnancies that successfully proceed to term. Problems can also arise in attempting to control for the potentially adverse effects of non-genetic variables, which include maternal and paternal age, parental disease, and unsuccessful attempted interruption of the pregnancy.

3.3 Direct Examination of the Progeny of Incest

Table 1 comprises data from the four most comprehensive published studies of incest. The first and obvious observation is that very few cases have been systematically investigated, surprisingly so since the topic of incest features prominently in the social and behavioral literature. Three of the studies were prospective, and in one investigation strenuous efforts were made to recruit matched non-incestuous controls (Adams and Neel 1967). There were no control subjects in two of the other studies (Carter 1967, Baird and McGillivray 1982), and the brief report by Carter (1967) was further restricted by a lack of information on the physical and mental status of the parents or their socioeconomic background. In several studies the period of follow-up was brief, and so disorders with a later age of onset may not have been detected.

By far the largest survey was that of Seemanová (1971), conducted retrospectively in the former Czechoslovakia on incestuous births reported between 1933 and 1970. An ingenious approach to the problem of controls was employed, with the outcomes of incestuous matings compared to those of pregnancies undertaken by a subset of the same women with unrelated partners. Unfortunately, the methodology could not overcome the fact that the physical and mental status of these women, 33 percent of the

original total, was significantly superior to the remainder of the study group. In addition, many of the incestuous pregnancies had taken place when the women had just completed menarche, with mean and modal ages of 18.9 and 16 years in the father–daughter matings and 19.9 and 14 years in the brother–sister matings, versus 24.9 and 21 years in the married pregnancy group. As very young maternal age is associated with an adverse pregnancy outcome, these discrepancies while unavoidable nevertheless cause significant problems in comparing the test group and their half-sibs.

How then should these four sets of data be assessed? Clearly, the mean rate of physical and mental abnormality of 39.4 percent is very high. But, as discussed in Sects. 2.2 and 3.2, the collection of data on this subject is extremely difficult and, as acknowledged by several authors, control for non-genetic variables in the incestuous unions may have been incomplete. In the two studies for which unrelated reference groups were available, 9 of the 113 control children died or had a serious defect (8.0 percent). Thus the mean excess level of death and severe defect in the offspring of incestuous unions was 31.4 percent, not all of which was necessarily genetic in origin.

3.4 Comparison of Incest with Other Categories of Inbred Union

The second major method of estimating the probable levels of defect in the progeny of incestuous matings is based on information gained from legal consanguineous unions. For example, it is assumed that as the progeny of an incestuous mating have a coefficient of inbreeding of $F = 0.25$, their levels of disease and disability will be four times higher than in first-cousin offspring ($F = 0.0625$). The main advantage of this approach is that information on first-cousin unions is more plentiful, and it is less likely to be subject to bias during data collection. A potential disadvantage is that the relationship between the level of inbreeding and rates of disease and disability may be nonlinear, with disproportionately larger increases in the prevalence of deaths and defects at closer degrees of inbreeding. This situation could arise if incest was associated with increased rates of conditional lethals, genes that are only expressed under especially stressful circumstances, as might be encountered in an incestuous pregnancy. To date there has been no evidence that conditional lethals are expressed in the progeny of close kin unions. Therefore the excess mortality rate of 4.4 percent derived for deaths among first-cousin offspring (Bittles and Neel 1994) can be applied to estimate the prevalence of lethal defects in an incestuous pregnancy, i.e., 16 percent to 20 percent excess deaths. Likewise, from Schull and Neel (1965) excess morbidity among incestuous offspring would predictably range from 6 percent to 16 percent, suggesting a total rate for excess death and major disability in children born to first-degree relatives of between 22 percent and 36 percent.

4. Conclusions

The data currently available on the outcomes of both legal consanguineous marriages and incestuous relationships merit further examination and updating. Since the major study undertaken by Schull and Neel (1965) in Japan no comparably detailed investigation has been attempted, despite the remarkable epidemiological transition which has occurred during the last two generations. Across Asia, genetic disorders are becoming increasingly obvious as deaths due to infectious diseases decline, and early childhood mortality is replaced by extended morbidity. This change is especially important among couples in consanguineous unions, where higher rates of genetic disorders are to be expected, involving both autosomal recessives and polygenic disorders. In fact, because of better access to diagnostic and health care facilities, the burden of genetic disease will probably be most obvious and of greatest immediate significance among migrants from Africa and Asia now resident in North America, Western Europe, and Australasia, who have chosen to continue their tradition of close kin marriage. With respect to incest, it seems remarkable that the voluminous social and behavioral literature is still so inadequately matched in biological terms. Studies to identify the precise nature, extent, and timing of any genetic defects that may be expressed are needed, especially since in a significant proportion of cases the children of incestuous unions are offered for adoption. The advent of DNA-based analytical methods will greatly facilitate this process, and provide information of previously unparalleled insight and sophistication. In the interim, appropriate caution should be exercised in judging the biological outcomes of all categories of close kin union.

See also: Eugenics, History of; Familial Studies: Genetic Inferences; Families and Households, Behavioral Demography of; Family, Anthropology of; Family as Institution; Family Processes; Genetics and Mate Choice; Incest Prohibition, Origin and Evolution of; Partner Selection across Culture, Psychology of

Bibliography

Adams M S, Neel J V 1967 Children of incest. *Pediatrics* **40**: 55–62

Baird P A, McGillivray B 1982 Children of incest. *Journal of Pediatrics* **101**: 854–57

Bemiss S M 1858 Report on influence of marriages of consanguinity upon offspring. *Transactions of the American Medical Association* **11**: 319–425

Bittles A H 1994 The role and significance of consanguinity as a demographic variable. *Population and Development Review* **20**: 561–84
Bittles A H, Mason W M, Greene J, Appaji Rao N A 1991 Reproductive behavior and health in consanguineous marriages. *Science* **252**: 789–94
Bittles A H, Neel J V 1994 The costs of human inbreeding and their implications for variations at the DNA level. *Nature Genetics* **8**: 117–21
Bratt, C S 1984 Incest statutes and the fundamental right of marriage: Is Oedipus free to marry? *Family Law Quarterly* **18**: 257–309
Carter C O 1967 Risk to offspring of incest. *Lancet* **1**: 436
Middleton R 1962 Brother–sister and father–daughter marriage in Ancient Egypt. *American Sociological Review* **27**: 603–11
Morton N E, Crow J F, Muller H J 1956 An estimate of the mutational damage in man from data on consanguineous marriages. *Proceedings of the National Academy of Sciences of the United States of America* **42**: 855–63
Ottenheimer M 1990 Morgan, Lewis, Henry and the prohibition of cousin marriage in the United States. *Journal of Family History* **15**: 325–34
Scheidel W 1997 Brother–sister marriage in Roman Egypt. *Journal of Biosocial Science* **29**: 361–71
Schull W J, Neel J V 1965 *The Effects of Inbreeding on Japanese Children*. Harper and Row, New York
Seemanová E 1971 A study of children of incestuous matings. *Human Heredity* **21**: 108–28
Shaw B D 1992 Explaining incest: Brother–sister marriage in Graeco-Roman Egypt. *Man* (N.S.) **27**: 267–99
Spiro M E 1965 *Children of the Kibbutz*. Schocken Books, New York, pp. 326–35, 347–50
Talmon Y 1965 The family in a revolutionary movement—the case of the kibbutz in Israel. In: Nimkoff N F (ed.) *Comparative Family Systems*. Houghton Mifflin, Boston, pp. 259–86
van den Berghe P L 1983 Human inbreeding avoidance: Culture in nature. *Behavioral and Brain Sciences* **6**: 91–102
Wolf A P, Huang C 1980 *Marriage and Adoption in China, 1845–1945*. Stanford University Press, Stanford, CA

A. H. Bittles

Incest Prohibition, Origin and Evolution of

Most anthropologists now agree that there is at least one universal institution—the incest taboo, which is usually defined as socially sanctioned disapproval of sexual relations between certain classes of kin. There is also general agreement that while the scope of the taboo and the sanctions invoked by its breach vary widely, it applies universally to sexual relations between mother and son, and almost universally to father/daughter and brother/sister relations. The exceptions are the Greek-speaking population of Roman Egypt and the ruling families of several old, long since defunct tributary states (see Hopkins 1980).

The incest taboo, like Berlin before German reunification, is located on the border between two great empires and, like Berlin in the 1950s, has symbolic as well as strategic significance. The taboo is a rule that is commonly codified as law and thus, in one respect at least, clearly belongs to the domain of inquiry ruled by the social sciences. But the taboo is about sex, and specifically about sex between people, most of whom are related genetically and thus, in another respect, clearly belongs to the domain of inquiry ruled by the biological sciences. Put another way, the taboo may be said to stand at the long-disputed boundary between nature and culture.

The great French anthropologist, Claude Levi-Strauss, recognized this, and tried to use the taboo to define that boundary and defend social science's claim to an exclusive right to everything on the cultural side. The taboo was, in his memorable phrase, 'a passage from nature to culture, from animal to human life' (Levi-Strauss 1960, p. 278). By accepting Sigmund Freud's claim that 'the first choice of object in mankind is regularly an incestuous one' (1923, p. 343), he located the taboo on the cultural side of the frontier and turned it into a frontier fortress. The taboo was social not only in form. Its origins were social and could only be divined by social scientists. Far from expressing an innate inclination, the primary purpose of the taboo was to repress an innate inclination.

The passage effected by the invention of the incest taboo was, in Levi-Strauss's account, a heroic journey. At some early stage in its history, mankind was forced to recognize that 'in order to free itself from a wild struggle for existence,' it had to forswear its preferred sexual objects. Instead of marrying the women born into their natal group, men had to exchange them for women born into other groups. The choice was between

> biological families living in juxtaposition and endeavoring to remain closed, self-perpetuating units, over-ridden by their fears, hatreds, and ignorances, and the systematic establishment, through the incest prohibition, of links of intermarriage between them, thus succeeding to build, out of the artificial bonds of affinity, a true human society, despite, and even in contradiction with, the isolating influence of consanguinity (Levi-Strauss 1960, p. 278).

Levi-Strauss's account of the incest taboo is best read as a sociological version of the story told by Freud in *Totem and Taboo* (Freud 1950). Another version is Bronislaw Malinowski's (1930) account of the taboo as the means by which the family is saved from the sexual competition that would inevitably erupt if men were allowed sexual access to their mothers and sisters, and women to their fathers and brothers. A third version, that still attracts attention, argues that the incest taboo exists to protect human beings from the biologically deleterious effects of

inbreeding. The common element in all of these stories is that the incest taboo exists to protect human beings from human nature. In its absence, society would dissolve into small, isolated groups; or the family would be torn asunder by sexual competition among its members; or the population would degenerate as a result of close inbreeding.

On the topic of the incest taboo, anthropologists act like a segmentary lineage. No matter how bitterly they oppose one another on other issues, Marxists such as Claude Meillassoux and Maurice Godelier, cultural materialists such as Leslie White and Marvin Harris, structuralists such as Claude Levi-Strauss and Rodney Needham, and radical social constructionists such as Marshall Sahlins and David Schneider unite to oppose vigorously anyone who suggests that instead of restraining human nature, the incest taboo may be rooted in human nature. Faced with an argument that says the incest taboo is not entirely conventional, they return to Sir James Frazer's assurance that 'we may always safely assume that crimes forbidden by law are crimes which many men have a natural propensity to commit.' On this assumption it is easy to argue that 'if the law represses [incest], as it represses other natural instincts, it does so because civilised men have come to the conclusion that the satisfaction of these natural instincts is detrimental to the general interests of society' (Frazer 1910, p. 97).

Larry Arnhart (1998) has called the tradition these arguments represent 'ethical transcendentalism.' It is a tradition that began in Greece with the sophists and found expression in the writings of Thomas Hobbes, Bernard Mandeville, and Immanuel Kant as well as in work of such modern social scientists as Emile Durkheim and Max Weber. Its counterpart is the 'ethical naturalism' that originated among the Socratic philosophers and was developed by Thomas Acquinas, Francis Hutcheson, Adam Smith, David Hume, and Charles Darwin. Its most important representative in the modern debate over the origins of the incest taboo was the Finnish anthropologist/sociologist/philosopher, Edward Westermarck. Frazer's purpose in insisting that 'the law only forbids men to do what their instincts incline them to do' (1910, p. 97) was to counter Westermarck's suggestion that 'there is a remarkable absence of erotic feelings between persons living closely together from childhood' and that this sexual indifference, 'combined with a positive aversion when the act is thought of,' is the cause of the incest taboo (1922, pp. 192–3). The suggestion was an anathema to a transcendentalist like Frazer because it located the source of the moral disapproval that forms the core of the taboo on the natural side of the great nature/culture divide.

When Westermarck published his proposal in 1891 it was greeted by the leading scientists of the time, including Anthony Wallace and E. B. Tylor, as 'the solution' to the problem of the incest taboo (see Wolf 1995, pp. 7–9). But by the time the fifth edition appeared in 1922, criticisms initiated by Frazer and developed by Freud had reversed this judgment. In his review, Bronislaw Malinowski, then Westermarck's colleague at the London School of Economics, admitted that 'Prof. Westermarck's theory of exogamy … seems to find favour with no one.' The list of Westermarck's critics reads like an honor role of twentieth-century social scientists—Robertson Smith, Andrew Lang, Sir James Frazer, Havelock Ellis, Sigmund Freud, William MacDougall, Emile Durkheim, Brenda Seligman, Bronislaw Malinowski, George Peter Murdock, and Claude Levi-Strauss. When a distinguished consortium of social and biological scientists gathered in 1960 to evaluate research on the incest taboo they did not even discuss what by then was known as 'the Westermarck hypothesis.' It is mentioned in their report 'only for the sake of completeness' (Aberle et al. 1963, p. 258).

Westermarck sank beneath a rising tide of transcendentalism initiated by the institutionalization of the social sciences in the early years of the twentieth century. To establish sociology and anthropology as specialized fields of study with buildings and budgets, their founding fathers had to identify independent subject matters. They had to make society and culture autonomous phenomena. The Westermarck hypothesis threatened this effort by breaching the barriers set up to defend these newly created domains from the claims of psychologists and biologists. Westermarck was mentionable only for the sake of completeness, because his hypothesis surrendered to the competition the only institution that everyone accepted as universal.

The evidence social scientists cited in rejecting the Westermarck hypothesis consisted exclusively of Freud's claim that 'psychoanalytic investigations have shown beyond the possibility of doubt that *an incestuous love-choice* is in fact the first and regular one' (Freud 1923, pp. 220–1). That this was not their real reason for rejecting the hypothesis is clearly evident in the fact that they did not cite Freud's views on any other important subject. The real reason social scientists cited Freud on incest was to capture the taboo for their newly founded disciplines. Freud served their purpose admirably. He was the most famous psychologist of the century, yet concluded that the causes of the incest taboo 'are not to be sought in the psychology of the individual' (1923, p. 221).

The Westermarck hypothesis was hailed initially as the solution to the problem of the incest taboo, because it appeared before the social sciences had erected transcendalist barriers to define their domains. It was resurrected 50 years after its announced demise by a generation of social scientists trained in the 1950s and early 1960s when, for a brief period, these barriers were left ungarrisoned. The Vietnam War did not mark the beginning of a revolutionary period in the social sciences. It marked the end of a revolutionary period. In the 1950s the rapid growth of higher

education created such an ample supply of money and jobs that, for a few years, competition between the disciplines gave way to an intellectual euphoria in which interdisciplinary training was encouraged. The result was a small cohort of social scientists who were relatively undisciplined.

The members of this cohort who did the most to effect a Westermarck revival were Robin Fox and Arthur Wolf. Although Fox (1952) was reluctant to reject Freud's account of the incest taboo, he initiated the Westermarck revival by treating his hypothesis as a respectable alternative. He also encouraged Joseph Shepher to undertake what turned out to be an influential study of the effects of the early association experienced by children raised together in Israeli kibbutzim. The problem with testing the Westermarck hypothesis had always been that the only children who are commonly reared together are brothers and sisters. Evidence that they are not attracted to one another could not be used to support Westermarck, because this is exactly the effect predicted by social scientists such as Levi-Strauss, who argued that the incest taboo 'remodels the biological conditions of mating and procreation' (1960, p. 278). The kibbutzim practice of communal rearing offered a rare opportunity to disentangle the effects of the confounded variables. Though the children of some kibbutzim were reared together from birth as intimately as any brothers and sisters, they were not brothers and sisters, and were not subject to the strictures of the incest taboo. Yet Shepher's survey of 2,769 marriages turned up only 13 cases in which the couple had been reared together, and in 12 of these cases the couple had not been brought together until after age three. The one exception was a couple brought together at birth but separated between ages two and six (Shepher 1971, pp. 296–7).

Westermarck's papers at the Åbo Academy in Turku include a letter from a former student inviting him to teach in China. He accepted the invitation, but was unable to make the trip, unfortunately, for if he had he might have discovered the natural experiment that would have proved his hypothesis. For reasons pointed out by John Hartung (1985), the data from Israel were not conclusive. A better test was available in China, where women commonly gave away their daughters and reared in their place wives for their sons. Arthur Wolf (Wolf and Huang 1985) calls the marriages that result 'minor marriages,' and contrasts them with what he terms 'major marriages.' Where couples married in the minor fashion were reared together from an early age, couples married in the major fashion did not meet until the day of their wedding.

Although the Taiwanese legal scholar Tai Yen-hui had noted as early as 1943 (Tai 1943) the relevance of minor marriages to the Westermarck hypothesis, the opportunity offered was not fully appreciated until 1958, when Wolf undertook a study of children's play in a small village in the Taipei Basin. He found that more than half of his subject's parents had been reared together and were not happily married. Village gossip indicated that they were far more likely to engage in extramarital relations than their neighbors who were married in the major fashion (Wolf 1968). This unanticipated discovery led Wolf into what became a 40-year effort to vindicate Westermarck. With evidence drawn from household registers compiled by the Japanese colonial government, he reconstructed the marital histories of 14,200 women, demonstrating that minor marriages produced 40 percent fewer children than major marriages and three times as many divorces. What made these results particularly compelling was the further discovery that they only held among women who had joined their future husband's family before age three (Wolf 1995). The critical variable was, as Westermarck had always maintained, *early* association.

It is puzzling that, despite spending much of his adult life in Morocco, Westermarck did not attempt to gather systematic information about the effects of *bint'amm* marriage, which was the preferred form of marriage in many Arab communities, and united cousins raised in the same household. Fortunately, Westermarck's failure in this regard was made good by Justin McCabe and Alex Walter. McCabe's 1974–5 study of 93 marriages in Lebanon, and Walter's 1991 study of 275 marriages in Morocco demonstrated that, among Arabs, as among Israelis and Taiwanese, 'there is a remarkable absence of erotic feelings between persons living closely together from childhood.' The only question these studies raised is why Westermarck did not pit *bint'amm* marriages against Freud's hysterics.

In most discussion of the orgins of the incest taboo, 'the Westermarck hypothesis' refers to Westermarck's claim that early childhood association inhibits sexual attraction. In fact, this is only one of three linked propositions that together constitute his explanation of the incest taboo. The first, highly controversial when proposed in 1891, is now accepted as dogma by biologists. It is simply the claim that inbreeding depresses reproductive fitness. Westermarck took this to be the reason childhood association inhibits sexual attraction. He argued that 'in this case, as in other cases, natural selection has operated, and by eliminating destructive tendencies and preserving useful variations, has moulded the sexual instinct so as to meet the requirements of the species' (1934, p. 159).

In his introductory lectures in 1917, Freud suggested that his audience 'would hardly recall without smiling' the view that 'sexual attraction is diverted from members of the opposite sex in one family owing to their living together from childhood' (1923, pp. 343–4). As we enter the twenty-first century, it is Freud's view that elicits smiles. The testimony of kibbutz marriage, minor marriage, and *bint'amm* marriage all agree that this, Westermarck's second proposition,

was right all along. Attention is now turning to the third of Westermarck's three propositions. It addresses what Bernard Williams (1983) has termed usefully 'the representation problem,' the problem of how a behavioral tendency comes to be represented in language and conceptual thought as a rule. Westermarck's (1906 and 1908) solution was to argue that incest is one of a class of acts that causes us to feel hostile toward another person even though that person's behavior does us no harm. This is because what the person does arouses painful emotions by forcing us to contemplate an emotionally unacceptable possibility. Thus the taboo is simply the reaction of the community to acts that arouse painful emotions in the majority of its members.

At the time of writing, the most widely accepted alternative to Westermarck's solution to the representation problem is an argument advanced by Richard Burton (1973) and William H. Durham (1991). They claim that human beings have long recognized the deleterious effects of inbreeding, and that the taboo was instituted to avoid these effects. Thus the debate is turning from the question of whether or not early association inhibits sexual attraction to the question of whether or not early humans would have noticed that the children of incest are less fit than other children. It is still far too early to say whether or not Westermarck will win a third round, but no one can now doubt that he will be credited with the solution to the problem of the incest taboo. In an obituary that reads like an indictment, Levi-Strauss characterized Westermarck as 'less a continuator than a survivor' (Levi-Strauss 1982, p. 181). His judgment now says more about him than it does about Westermarck.

See also: Childhood Sexual Abuse and Risk for Adult Psychopathology; Family, Anthropology of; Incest, Inbreeding, and their Consequences; Kinship in Anthropology; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Taboo

Bibliography

Aberle D F, Bronfenbrenner U, Hess E H, Miller D R, Schneider D M, Spuhler J M 1963 The incest taboo and the mating patterns of animals. *American Anthropologist* **65**: 253–65

Arnhart L 1998 *Darwinian Natural Right: The Biological Ethics of Human Nature*. State University of New York Press, Albany, NY

Burton R V 1973 Folk theory and the incest taboo. *Ethos* **1**: 504–16

Coggin J M, Sturtevant W C 1964 The Calusa: A stratified, non-agricultural society (with notes on sibling marriage). In: Goodenough W H (ed.) *Explorations in Cultural Anthropology*. McGraw-Hill, New York, pp. 202–7

Durham W H 1991 *Coevolution*. Stanford University Press, Stanford, CA

Fox R 1962 Sibling incest. *British Journal of Sociology* **13**: 128–50

Fox R 1980 *The Red Lamp of Incest*. Dutton, New York

Frazer J G 1910 *Totemism and Exogamy*, 4 Vols. Macmillan, London

Freud S 1920 *A General Introduction to Psychoanalysis* [trans. J Riviere]. Pocket Books, New York

Freud S 1950 *Totem and Taboo* [trans. J Strachey]. Norton, New York

Hartung J 1985 Review of *Incest: A Biosocial View* by J Shepher. *American Journal of Physical Anthropology* **67**: 169–71

Hopkins K 1980 Brother–sister marriage in Roman Egypt. *Comparative Studies in Society and History* **22**

Levi-Strauss C 1960 The family. In: Shapiro H L (ed.) *Man, Culture, and Society*. Oxford University Press, New York, pp. 261–85

Levi-Strauss 1969 *The Elementary Structure of Kinship*, trans. J H Bell, J R von Sturmer, R Needham. Beacon Press, Boston, MA

Levi-Strauss C 1982 The work of Edward Westermarck. In: Stroup T (ed.) *Edward Westermarck: Essay on His Life and Works*. Philosophical Society of Finland, Helsinkl, Finland

Malinowski B 1992 Sexual life and marriage among primitive mankind. *Nature* **109**: 502–4

Malinowski B 1930 Culture. *Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 4, pp. 621–45

McCabe J 1983 FBD marriage: Further support for the Westermarck hypothesis of the incest taboo? *American Anthropologist* **85**: 50–69

Shepher J 1971 Mate selection among second-generation kibbutz adolescents and adults: Incest avoidance and negative imprinting. *Archives of Sexual Behavior* **1**: 293–307

Tai Y H 1943 *Sim-pua* zakkō (Miscellaneous notes on *sim-pua*). *Minzoku Taiwan* **3**: 2–4

Walter A 1997 The evolutionary psychology of mate selection in Morocco. *Human Nature* **8**: 113–37

Westermarck E 1906 *The Origin and Development of Moral Ideas*, Vol. 1. Macmillan, London

Westermarck E 1908 *The Origin and Development of Moral Ideas*. Vol. 2. Macmillan, London

Westermarck E 1922 *The History of Human Marriage*, 5th edn. Macmillan, London

Westermarck E 1934 *Three Essays on Sex and Marriage*. Macmillan, London

Williams B 1983 Evolution, ethics, and the representation problem. In: Bandall D S (ed.) *Evolution from Molecules to Men*. Cambridge University Press, Cambridge, UK

Wolf A P 1966 Childhood association, sexual attraction, and the incest taboo: A Chinese case. *American Anthropologist* **68**: 883–98

Wolf A P 1995 *Sexual Attraction and Childhood Association: A Chinese Brief for Edward Westermarck*. Stanford University Press, Stanford, CA

Wolf A P, Huang C S 1980 *Marriage and Adoption in China, 1845–1945*. Stanford University Press, Stanford, CA

A. P. Wolf

Incidental versus Intentional Memory

The distinction between incidental and intentional memory deals with differences in people's expectations and strategies while encoding episodic information. In a standard memory experiment, intentional memory refers to study conditions in which participants are

forewarned that they will be tested on material to which they are exposed. By contrast, learning is said to be incidental when participants are not expecting a memory test. Although learning in everyday life is both incidental and intentional, the former type of activity is more prevalent in most everyday situations in that people observe events and activities without having explicit instructions about a forthcoming memory test.

Other things being equal, intentional learning instructions typically produce better recall and recognition performance than incidental learning instructions. A central issue in this context is, however, whether differences in performance following incidental vs. intentional leaning instructions are due to the nature of intent *per se* (e.g., through increased motivation, arousal, and activation). An alternative possibility is that these effects can be attributed to processing differences following incidental vs. intentional learning instructions (e.g., due to differences in study time, elaboration, organization and subsequent rehearsal).

Concerning the latter notion, although intentional learning instructions produce better memory performance than incidental learning instructions in a variety of task conditions, these differences can be understood without making specific assumptions about the nature of intent *per se*. Instead, past research indicates that intent to learn is not of crucial importance, and that the effect of intentional learning can be attributed to the nature of processing activity during encoding and subsequent retention interval. In other words, memory performance following incidental learning instructions is comparable to that of intentional learning instructions given that similar encoding processes are used in both conditions.

As a support for this notion, Hyde and Jenkins (1973) reported an experiment in which they attempted to equate participants' encoding activities, while manipulating learning instructions (i.e., incidental vs. intentional learning). Participants studied lists of words under five different conditions. These *orienting tasks* differed in terms of *level of processing* with some tasks reflecting relatively shallow processing level, whereas the other orienting tasks involved deeper, more semantic processing. Hyde and Jenkins (1973) found that recall performance increased as a function of processing level and, more important in the present context, that the same pattern of results was observed for the incidental and intentional learning conditions. In other words, participants' orientation toward a stimulus, rather than explicit intention to remember, was the important determinant of retrieval success. These findings indicate that intentional learning instructions have indirect, rather than direct, effects on subsequent memory performance in that intention to learn induces more effective encoding strategies, including rehearsal and recoding.

Similar findings have been observed when participants' motivation has been manipulated by using different forms of incentives. For example, Nilsson (1987) reported a study in which three groups of students were instructed to memorize lists of words for subsequent free recall and recognition tests. To manipulate motivation, one group of participants was informed of a monetary incentive prior to study. A second group was informed of these incentives after study but prior to the test, whereas no mention of incentives was made to the third group of participants. The results of the study indicated a dissociation between subjective reports and objective memory performance. Subjective reports indicated that the incentives had affected the amount of effort put into the memory task, but these subjective reports did not reflect differences in actual memory performance. Instead, the three groups showed comparable levels of recall and recognition performance.

As noted by Roediger and McDermott (1993), the distinction between incidental and intentional learning during study parallels the distinction between implicit and explicit memory instructions at test. In the latter case, *explicit memory* tests direct participants to recall past experience (i.e., intentional retrieval), whereas in *implicit memory* tests the influence of past experience is revealed under task conditions that reflect incidental retrieval (see also Schacter et al. 1989). Consistent with Hyde and Jenkins' (1973) findings with explicit memory tests, intention to learn has no effect on implicit memory performance. For example, Roediger et al. (1992) used a standard levels-of-processing manipulation with a graphemic and semantic orienting task. Participants were either informed or not about the forthcoming (explicit and implicit) memory test, in which word stems and word fragments were presented as cues. Roediger el al. (1992) found no effect of intention either on implicit or explicit memory performance (see also Bowers and Schacter 1990, Greene 1986).

Although intention to learn has minimal effects on explicit and implicit memory, research on delayed intentions, or prospective memory (see Brandimonte et al. 1996 for an overview) indicates that representation of intentions may have specific dynamic properties. One basis for this hypothesis was Lewin's (1961) notion that unfulfilled intentions produce a persistent 'goal tension.' According to Lewin, the after-effect of intention is a force that produces a 'quasi-need' to carry out planned action. Lewin stated that the clearest subjective experience of this force occurs in the resumption of interrupted tasks, when a general pressure—that 'that there is something I should do'—appears (Lewin 1961, p. 1251).

Zeigarnik, Lewin's student in his Berlin laboratory, examined the memorial consequences of interrupting actions. Zeigarnik (1927) (see also Ovsiankina 1928) presented participants with a series of concrete tasks, such as threading beads or drawing a vase. Participants were allowed to complete one half of the tasks, whereas

the experimenter interrupted the remaining tasks, interspersed throughout the series, before participants could complete them. Immediately after the completion of the series, participants were given a free recall test in which they were asked to recall the names of the tasks. Zeigarnik demonstrated superior recall of activities that were interrupted prior to completion as compared to completed actions. Although these findings were consistent with the assumption that incomplete intentions produce a persisting task tension, task interruption research has been criticized for conceptual and methodological shortcomings (see Butterfield 1964, van Bergen 1968, Ryan 1970 for reviews).

Goschke and Kuhl (1993, 1996) provided more recent support for the notion that representation of intentions may have specific dynamic properties. Consistent with Lewin's notion of goal tension, Goschke and Kuhl reasoned that intended actions are represented at a heightened level of activation as compared with material which is not intended for future action. Goschke and Kuhl (1996) proposed that the retrieval of an (nonroutine) intention depends upon both the quality of available retrieval cues and the level of activation at which the intention is represented at any given moment.

To examine whether intentions are represented at a higher or more sustained level of activation, Goschke and Kuhl (1993) developed a paradigm in which participants first memorized verbal descriptions of two different activities (e.g., setting the dinner table and clearing a messy desk). After the study phase, participants were informed that they should later either execute one of the scripts or observe the experimenter executing the action and register possible mistakes. The second script in each condition was neither executed nor observed, but the participants were informed that their memory for both scripts of a pair would be tested. Immediately after the execution or observation instructions, the participants received a recognition test for words for both scripts. Goschke and Kuhl reasoned that if intention-related concepts are characterized by a higher level of activation than nonintention-related concepts, the recognition latencies for words from the to-be-executed script should be faster than words from the to-be-remembered script. The results of Goschke and Kuhl's (1993) study supported this prediction by showing an 'intention-superiority' effect even when the use of strategies was limited. However, in a control condition in which the participants had to observe someone else executing a script, latencies for words from the to-be-observed script did not differ from neutral words.

Although these findings indicate that intention-related concepts are characterized by a heightened and more sustained level of activation than other memory contents, they do not necessarily justify the conclusion that the intention-superiority effect underlies successful (prospective) memory performance. In other words, although the intention-related concepts were recognized faster than nonintention-related concepts in Goschke and Kuhl's study (see also Koriat et al. 1990, Mäntylä 1996, Marsh et al. 1998 for similar findings), the question remains open as to whether an increased activation of intention representation also facilitates subsequent memory performance.

See also: Implicit Learning and Memory: Psychological and Neural Aspects; Implicit Memory, Cognitive Psychology of; Memory: Levels of Processing; Memory: Organization and Recall

Bibliography

Bowers J S, Schacter D L 1990 Implicit memory and test awareness. *Journal of Experimental Psychology: Learning, Memory and Cognition* **16**: 404–16

Brandimonte M, Einstein G O, McDaniel M A (eds.) 1996 *Prospective Memory: Theory and Applications*. Erlbaum, Hillsdale, NJ

Butterfield E C 1964 The interruption of tasks. Methodological, factual, and theoretical issues. *Psychological Bulletin* **62**: 309–22

Goschke T, Kuhl J 1993 Representation of intentions: Persisting activation in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**: 1211–26

Goschke T, Kuhl J 1996 Remembering what to do explicit and implicit memory for intentions. In: Brandimonte M, Einstein G O, McDaniel M A (eds.) *Prospective Memory: Theory and Applications*. Erlbaum, Hillsdale, NJ, pp. 53–91

Greene R L 1986 Word stems as cues in recall and completions tasks. *Quarterly Journal of Experimental Psychology A* **38**: 663–73

Hyde T S, Jenkins J J 1973 Recall of words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior* **12**: 471–80

Koriat A, Ben-Zur H, Nussbaum A 1990 Encoding information for future actions: memory for to be-performed tasks versus memory for to-be-recalled tasks. *Memory and Cognition* **18**: 568–78

Lewin K 1961 Intention, will, and need. In: Shipley T (ed., trans.) *Classics in Psychology*. Philosophical Library, New York, pp. 1234–88 [Original work published 1926, *Psychologische Forschung* **7**: 330–85]

Mäntylä T 1996 Activating actions and interrupting intentions: Mechanisms of retrieval sensitization in prospective memory. In: Brandimonte M, Einstein G O, McDaniel M A (eds.) *Prospective Memory: Theory and Applications*. Erlbaum, Hillsdale, NJ, pp. 93–113

Marsh R L, Hicks J L, Blink M L 1998 The activation of completed, uncompleted and partially completed intentions. *Journal of Experimental Psychology: Learning, Memory and Cognition* **24**: 350–61

Nilsson L-G 1987 Motivated memory: Dissociations between performance data and subjective reports. *Psychological Research* **49**: 183–8

Ovsiankina M 1928 Die Wiederaufnahme unterbrochener Handlungen. *Psychologische Forschung* **11**: 302–79

Roediger H L, McDermott K B 1993 Implicit memory in normal human subjects. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*. Elsevier, Amsterdam, Vol. 8, pp. 63–131

Roediger H L, Weldon M S, Stadler M A, Riegler G H 1992 Direct comparison of two implicit memory tests: Word fragment and word stem completion. *Experimental Psychology: Learning, Memory and Cognition* **18**: 1251–69
Ryan T. A 1970 *Intentional Behavior: An Approach to Human Motivation*. Ronald Press, New York
Schacter D L, Bowers J, Booker J 1989 Intention, awareness, and implicit memory. The retrieval intentionality criterion. In: Lewandowsky S, Dunn J C, Kirsner K (eds.) *Implicit Memory: Theoretical Issues*. Erlbaum, Hillsdale, NJ, pp. 47–65
van Bergen A 1968 *Task Interruption*. North-Holland, Amsterdam
Zeigarnik B 1927 Über das Behalten von erledigten und unerledigten Handlungen. *Psychologische Forschung* **9**: 1–85

T. Mäntylä

Income Distribution

The subject of income distribution is concerned with explaining and evaluating the way in which the total national income is divided. The concern of classical economists with the shares of factors of production (capital, labor, and land) in national income has evolved into an analysis of the distribution among persons. Distribution by factors is still relevant, but largely as a means to understanding the personal distribution. Statistics are regularly published in most countries showing the shares of, for example, the top 20 percent in total personal income, or summary measures of inequality such as the Gini coefficient. But several issues arise when interpreting them in the light of positive and normative economic theory.

1. *Income Distribution and Economics*

For David Ricardo, writing in 1817, 'to determine the laws which regulate [the distribution of income] is the principal problem in Political Economy', and for many non-economists today, the issue of how national income is divided must appear one of the central questions of economics. Why are airline pilots paid more than railroad drivers are? Why are CEOs paid in a week what others take a year to earn? Do large fortunes arise mainly from inheritance?

Classical economists, such as Ricardo, were concerned with the determinants of payments to factors of production, and much of what can be found in today's textbooks under the heading of the 'Theory of Distribution' is concerned with the factor distribution of income. But it is the personal distribution of income—how much is received by individuals—which is more obviously of interest in public debate. The subject of the personal distribution has not however always been regarded as important by the economics profession. In the first part of the twentieth century, there were important contributions, but the topic was never central to the concerns of economists. In the second half of the century, there were indeed times when interest in the distribution of income was at a low ebb, economists appearing to believe that differences in distributive outcomes were of second order importance compared with changes in overall economic performance. People justified a lack of distribution concern by the argument that those at the bottom would gain more from employment policies and the promotion of economic growth than from redistribution.

At the start of the twenty-first century, the position is different. Europe has seen prolonged unemployment. There has been widening wage dispersion in a number of OECD countries, notably in the USA. Rising affluence in rich countries coexists, in a number of such countries, with the persistence of poverty. Economists have responded quickly to the renewed policy interest in distribution, and there is an extensive recent literature, both theoretical and empirical (see Atkinson and Bourguignon 2000). What can we say about the determinants of the distribution of income, its evolution over time, and the way it is affected by policy? There are many competing theoretical answers to these questions. Theoretical inquiry has an empirical counterpart. Statistical institutes publish evidence on the distribution of income on the basis of various statistical concepts and indices: dispersion of earnings, inequality of household's net income after taxes and transfers, inequality of wealth, etc. We have to ask whether we have developed the concepts necessary to interpret this evidence, and whether the data are of adequate quality.

Studying income distribution is important not only in its own right but also because it assists our understanding of various fields of economics. Even if there were no ethical reason for studying distribution, it would still be required as a conditioning parameter in other economic phenomena. A good example is the relationship between growth and distribution. Growth may affect distribution, and conversely changes in distribution may lead to faster growth. Aggregation is the methodological bridge between distribution issues and mainstream economic analysis. For example, aggregate consumer demand is the sum of the demands of individual households. The treatment of differences between households has long been standard in consumer theory, but other examples are emerging, such as the introduction of heterogeneous agents in macroeconomics. Distribution has always been prominent in development economics. It was clearly central to Marxian economics.

2. *Seeking to Explain the Distribution*

No unified theory of income distribution exists; instead the literature offers a series of building blocks

which provide part of an explanation. This section reviews these various blocks and the links between them.

2.1 Factor Incomes

A natural starting point is the theory of competitive general equilibrium, which in its standard Walrasian formulation takes as given the distribution of all productive factors in the economy and focuses on the rate at which they are paid. It is a simple model, but it underlies much of applied and theoretical literature on income distribution.

In a static framework, each individual is endowed with productive factors, where these may include different types of capital or labor skills. Firms produce various goods using these factors of production. Individuals receive income from the sale of their productive factors and from their ownership of firms. The competitive equilibrium model is closed by determining the prices and factor rewards that equilibrate demand and supply of the various goods and factors. The distribution of income thus results from the combination of the multi-dimensional distribution of endowments, the per unit returns to these endowments, and of the distribution of the ownership of firms within the population. In such a framework, a theory of income distribution is essentially a theory of factor rewards, and this explains the location of the subject in many economic textbooks as part of the theory of pricing.

This general equilibrium formulation has been applied both in empirical detail, as in Computable General Equilibrium models of income distribution, and in relatively simple theoretical versions. An example of the latter is where there are two factors of production: labor supplied by workers who make up a fixed fraction of the population, and capital owned by capitalists, who do not work and make up the rest of the population. The relative distribution of incomes in the population depends only on the income and population shares. This two-class economy, reminiscent of Ricardo, is the simplest justification for reducing the issue of income distribution to that of factor shares. At the beginning of the twenty-first century, however, this is scarcely adequate for several reasons. We need to explain the distribution of factor incomes within classes, such as the size distribution of wages. Capital takes the form not only of material assets, but also of human capital: the investment which people make in themselves in the form of education and training. People cannot be identified with a single source of income: one individual may be in receipt of wages, interest income, and rent (for example through owning a house). We cannot therefore draw any direct implications for the personal distribution from observations of changes in factor prices.

2.2 Wage Differentials

There has been a long-standing historical interest in the evolution of the differential between the wages of skilled and unskilled workers, with evidence going back to 1300 in England, when building craftsmen were paid 3d a day, compared with 2d a day for laborers. At the end of the twentieth century in the USA, there has been considerable debate about the causes of the observed increase in wage dispersion, this being interpreted in terms of the supply and demand for skill.

One set of theories of the distribution of wages has emphasized the role of random shocks, whereby individual earnings are governed by a stochastic process. If the logarithm of earnings follows a random walk, then, after some time, earnings are distributed lognormally among individuals with an increasing variance over time. This is the well-known Gibrat's law. With a Galtonian mean reversion process, the distribution of wages tends toward some limit that depends only on the characteristics of the distribution of the stochastic shocks and the reversion parameter. It is via the latter that 'economics' may enter, determining the degree of magnification of the inequality due to the stochastic term, and the convergence or otherwise of the process.

Stochastic models have been repeatedly criticized on the grounds that they do not incorporate individual optimizing behavior and market clearing. In contrast, in a supply and demand framework, changes in wage differentials, say between skilled and unskilled workers, depend on shifts in demand or in supply. A model of this kind was the basis for the 'race' between technological development and education described by Tinbergen (1975) and is very much behind today's debate on increasing wage inequality. Tinbergen referred to people being educated to graduate or high-school level. (The analysis refers to an advanced country, in that it assumes that everyone receives at least a high school education.) Total output is produced using graduate labor, high school labor, and capital, and from the production function we can obtain the profit-maximizing choice of labor by competitive firms. If the relative supplies are fixed in the short-run, then we can solve for the market clearing wage differential. It has been argued that this differential increased because demand has shifted away from unskilled labor: for example on account of technological factors biased in favor of skilled labor, or, in a multisectored model, biased in favor of sectors intensive in the use of skilled labor. An alternative explanation is that the expansion of world trade has led to a drop in the relative price of goods that are relatively intensive in unskilled labor.

Over time, the supply of factors will respond to shifts in demand possibly canceling the initial increase in the wage differential. For Tinbergen, the outcome depended on a race between the growth in the relative

numbers with graduate education and the demand shifts increasing the importance of graduate labor in production. Behind the supply adjustment lies a theory of investment in human capital, of the kind developed by Mincer (1958) and Becker (1967). Investment in human capital by an individual is governed by the equalization of marginal return and marginal cost. If the return on investment in education, or its marginal cost is individual-specific, then accumulation differs across persons. In such a model, the long-run cause of income differences, apart from stochastic terms, becomes the difference in abilities; we have therefore simply pushed the problem back one stage to the explanation of these ability differences.

Applied to the skilled/unskilled wage differential, the human capital theory has two important implications. First, to the extent that there are no differences in ability, and no stochastic terms, then no lifetime inequality is introduced: the wage premium for skill is simply a compensating differential. Second, the compensating wage differential depends on the rate of interest. A rise in real interest rates could explain observed widening in the wage distribution.

Present-day theories of wage differentials are essentially framed in terms of supply and demand. They contrast markedly with earlier writing, where there has been a creative tension between market force and alternative explanations of wage differentials. The competitive model has to be modified to allow for trade union/employer negotiations and for the impact of government, both on collective bargaining and directly on wage determination. If it is the case that unions lead to lower earnings dispersion, then one possible cause of observed increases in dispersion is the decline in union power and coverage. The direct impact of government has included minimum wage protection, income policies, and 'fair wages' requirements for government contractors. Trade union bargaining and statutory wage determination may be incorporated into the supply and demand framework, as is actually done in modern theories of wage setting; and the same applies to monopsonistic behavior by employers. More of a challenge is the view that supply and demand only place limits on the possible wage differentials, with other factors such as notions of fairness or equity determining where between these limits wages actually lies. Such a 'range theory' of wage differentials has long been implicit in much institutional writing on labor economics.

Institutional models have also emphasized labor market segmentation. Wages are fixed in some firms or sectors at a level above the competitive rate for a given type of labor. Workers outside this privileged segment of the labor market would like to enter, but the number of openings is limited and rationing occurs according to some scheme that is arbitrary or partly dependent on the characteristics of the workers. Several mechanisms have been invoked to explain this segmentation of the labor market and the wage dispersion that it creates: efficiency wages, the existence or the uneven strength of labor unions, and more generally the distinction between insiders and outsiders in the internal labor market. An extreme case of segmentation is labor market discrimination by which some individuals in the labor force are simply prevented from being hired in some jobs or at some wage levels on the basis of their ethnic origin or their gender.

2.3 *Wealth Accumulation*

Income from financial capital and real assets forms a smaller part of total income for persons than labor income, but it is typically more unequally distributed in market economies. The distribution of capital income depends on that of personal endowments and assets. To understand this, the theory must tackle the issue of the accumulation of productive factors.

The standard assumption is that accumulation decisions are based on maximizing behavior, and hence depend on the sequence of current and expected future factor rewards. The most elementary version of the accumulation equation refers to a single asset, possibly including human capital, being accumulated by individuals who are identical apart from their level of asset-holding and who operate in markets with constant prices. The dynastic model offers a benchmark where the altruism of a person extends to his/her descendants, their descendants, and so on for an infinite future. This is equivalent to assuming that the person lives forever and optimizes over an infinite horizon. Under certain further assumptions, including that stochastic shocks on wealth are persistent, the optimal consumption at each point of time is simply the income flow from wealth. Wealth, income, and consumption then all follow a random walk, and their variance in the population increases linearly with time. There is some evidence that the dispersion of consumption expenditures tends to increase continuously with age in a given cohort, so that the preceding model could be satisfactory for an intragenerational theory of income and wealth distribution. However, it does not seem to fit the most obvious stylized facts of the intergeneration transmission of inequality, and in particular the apparently nonincreasing variance of wealth.

The model of accumulation just described relies on a rather extreme form of altruism. If agents were selfish and indifferent to the fate of their descendants, wealth should decrease at the end of one's lifetime and bequests should essentially be involuntary, explained by the natural uncertainty of life duration, coupled with the absence of a good annuity market. But a host of intermediate cases may be envisaged where bequests enter personal utility and transfers to children are made at death or during one's lifetime under the form of human capital. Fertility behavior must be taken

into consideration. Large families mean that, with equal division, wealth is divided more rapidly; where families die out, on the other hand, wealth passes into other hands. There is a rich literature on the simultaneous determination of fertility and intergeneration transfers of wealth and/or human capita, and on how estates are divided among heirs.

In the models just described, differences across individuals arise on account of random shocks on income and their magnification over time, and of their initial endowments. Individuals are also likely to differ in their preferences for saving, for number of children, and in the degree to which they wish to pass on wealth. But the heterogeneity in their initial endowments of wealth is likely to play a key role and to leave a long shadow on the distribution. Imperfections in the capital market may mean that some people become increasingly trapped in low wealth. Because of information asymmetries between lenders and borrowers leading to moral hazard and risk selection problems, people with a low level of wealth cannot borrow against future incomes or can do so at a rate of interest which depends negatively on their current wealth used as a collateral. In the field of human capital, and financing entrepreneurship, these imperfections can exacerbate initial inequalities. Two families with similar initial wealth may end up with different long-term wealth.

All the preceding dynamic theories of the distribution of productive assets assume that the price system, as summarized by the rate of return on wealth, is independent of the distribution. In a dynamic general equilibrium framework, the personal distribution of assets, the distribution of fixed factor among firms and the price system all depend on each other. This relationship has been analyzed within the framework of Solow's aggregate neoclassical growth model by Stiglitz (1969). With an identical linear savings relationship, the behavior of aggregate capital converges to a steady state level of capital, and this implies in turn that ultimately the individual wealth-holdings also converge. But very different results may be obtained by modifying the assumptions about the model of growth and factor shares, and in particular by introducing the capital market imperfections described earlier.

2.4 Public Policy and Redistribution

The state can affect the distribution of income in many ways apart from the taxes, transfers, and spending (on education, health, etc.) of which people naturally think. Minimum wage legislation is an obvious example, as are other forms of government intervention in the labor market, including trade union legislation, income policies, legislation governing hiring and firing, immigration controls, occupational licensing, and public employment. In the capital market, there may be distribution consequences of government monetary and debt policies, but in addition the government regulates financial institutions, legislates for consumer credit protection, and may operate savings and other banks. Industry is affected by competition policy, company law, and rules governing public procurement. In recent years, there has been privatization of nationalized enterprises or the sale of other state assets. The impact on the distribution of income of all these policy areas needs to be studied in any assessment of the overall impact of the government.

Turning to the impact of the government budget, we see that the gross incomes generated by production are typically reduced by taxation, used to finance public spending, including transfers which constitute a source of personal incomes which has no counterpart on the production side. The state may in part finance spending through borrowing, and the national debt adds to the range of assets held by the personal sector.

The impact of taxes and transfers on the distribution of income depends crucially on their incidence, in that the existence of the taxes and transfers may change market incomes. If, for example, top pay rates are set in net of tax terms (to attract executives on an international market), a tax increase will lead to higher gross pay and not to lower disposable income. The redistributive effect of taxation cannot be measured by comparing market and disposable incomes; the comparison needs to be with a counterfactual market income in the absence of taxes. In order to determine the latter, we have to make assumptions about the working of the economy, and this is one of the standard issues in public finance.

The importance of taxes and transfers in the distribution of personal income means that the behavior of the government has to be part of our explanation. The government's actions cannot be treated as purely exogenous, and we need to look to the theory of public choice, or 'political economy.' There has indeed been a resurgence of interest amongst economists in the politics of income redistribution. Most models share a common approach to the determination of political equilibrium, which is some version of the median-voter model. It is however important to ask how far the findings depend on whether the outcome is governed by the preferences of the median voter, or by the ideology or preferences of political parties, or by political pressure from different interest groups, or by bureaucratic control of civil servants or agencies. There has been relatively little research by economists, which has set side by side different possible explanations of income redistribution, and examined the sensitivity of the conclusions drawn to the choice of model.

3. Measuring Inequality

The preceding section reviewed existing economic theories seeking to explain the distribution of income.

The empirical counterpart of these theories consists of studies, which compare income distribution data across various societies, or at different points of time for the same society and which try to relate observed differences to a set of exogenous characteristics of the societies being analyzed. Considerable work has been developed along these lines since the comparisons undertaken by Pareto (1897), who fitted a particular mathematical distribution to the upper part of the income distribution and used the resulting parameters as the basis for an explanation of differences. As has been evident even from the days of such pioneering studies, the measurement of inequality raises both conceptual problems and questions concerning the quality of the underlying evidence.

3.1 Conceptual Issues: Measuring Inequality

Traditionally, individual observations were arranged into a vector indicating the proportion of people falling in selected income bands. Comparing two distributions of income then consisted of comparing these frequency histograms. At the beginning of the twenty-first century, modern computing possibilities allow one to work directly with the individual observations rather than grouping them and to obtain more flexible estimates of the income frequency function through Kernel techniques (Silverman 1986). Other representations of the distribution of income include the 'distribution curve' and the Lorenz curve. The former simply cumulates people below a given income level. The Lorenz curve cumulates the population in increasing order of income, and shows on the vertical axis their cumulative share in total income. The Lorenz curve is equivalent, up to the division by mean income, to other representations of the data, but it allows a ready comparison of two or more distributions. A situation of particular interest is where one Lorenz curve lies everywhere above, or at least not below, another, which means that the bottom X percent of the population always have a larger share of total income, for all values of X. This situation is referred to as 'Lorenz dominance.' Where distributions differ in their mean incomes, as where comparing different countries, we may use the generalized Lorenz curve. This replaces the relative percentage of total income on the vertical axis by the absolute total income per head, so that it is now denominated in currency.

Comparisons of Lorenz curves may however be unclear or ambiguous because curves are close to each other or cross several times. This is one reason why so many authors prefer to rely on scalar inequality measures that summarize the departure of the distribution from equality and satisfy various basic properties. The Gini coefficient has for long been the most popular such measure. The reasons for its popularity are not entirely clear, but may be due to its graphical interpretation as the area between the Lorenz curve and the diagonal, relative to the whole triangle. Other inequality measures used include the relative absolute or logarithmic mean deviation, the coefficient of variation and entropy measures. Using different scalar inequality measures to compare distributions may lead to contradictory conclusions, one distribution appearing more unequal than another with respect to one measure, but the opposite being true with another measure. A condition for such a contradiction not to occur, for a wide class of inequality measures, with distributions with the same mean and total population, is Lorenz dominance.

3.2 Conceptual Issues: Definition of Income

A second conceptual issue is the definition of the variable under consideration: inequality of *what*? Income is typically studied on the basis that it is a reasonable indicator of individual welfare, but this may be criticized on at least two grounds. The first is that we need to enrich the informational basis for welfare judgments, including for instance life expectancy or quality of life; this has been emphasized by Sen (see Sen 1992), who has contributed to the development of a nonwelfarist approach.

The second argument is that income provides an unsatisfactory measure of individual welfare. Income may for example be subject to transitory fluctuations that can be smoothed out by the individual; this may provide a rationale for using observed consumption expenditures rather than income to estimate inequality. According to the permanent income hypothesis, differences in consumption expenditures over time and between agents should reflect differences in permanent disposable household income rather than transitory shocks to income. As a matter of fact, it is typically true that the distribution of consumption expenditures is less unequal than that of current income. This being said, the use of consumption rather than income data raises problems of definition and observation, notably the treatment of durables and the necessity of imputing a value for their services.

If income is chosen as the variable to be considered, then a comprehensive definition embraces all receipts that increase a person's command over resources in a specified period. As such, it includes all income in kind, such as production for home consumption or fringe benefits; it includes all capital gains or losses on all assets (not just those that are realized); it includes imputed rent from home ownership and other physical assets. The definition is in principle far-reaching and in practice hard to achieve.

3.3 Conceptual Issues: Unit of Analysis

A third conceptual issue concerns the definition of the recipient unit: inequality among whom? Empirical

analysts generally have to choose between a range of alternatives, depending on the data at hand and the issue to be addressed:

(a) individuals, whether they have an income or not,

(b) individual income recipients,

(c) families of related adults and dependent children,

(d) spending units, that is individuals pooling their income together and sharing the same consumption budget,

(e) households, that is, people living at the same address.

These definitions may lead to different evaluations of the degree of inequality of a distribution, and possibly to different representations of its evolution over time. The main difference between distribution data defined with reference to these various recipient units has to do with the 'matching' of individual earners into households and the size of these households. The choice of recipient unit depends on the issue that is addressed. Individual earnings data are better adapted to positive studies focusing on the labor market, whereas household income may be more appropriate in normative studies addressing the issue of inequality in living standards, although this is influenced by the degree to which resources are shared within the household.

Whatever unit of analysis is chosen, an adjustment has to be made for the differing needs of units of different size and composition. The most meaningful concept from that point of view is the distribution of 'equivalized incomes' where total income is expressed per adult equivalent. One such adjustment is to take per capita income, but this makes no allowance for economies of scale, and a commonly used alternative adjustment is to divide total household income by the square root of the household size. There is a voluminous literature on equivalence scales: see, for example, Deaton and Muellbauer (1980). A further choice to be made is the weighting to be given to each unit: between treating each household as one and treating each person as one, so that we impute to every member in the household the per adult equivalent income.

3.4 Income Distribution Data

There has been considerable improvement in the availability of data about the distribution of income. Advances have been made at the national level, where in many countries a significant investment has been made in carrying out new household surveys, in linking administrative data, and in refining methods of analysis. Administrative data, of which the most important are income tax records, have historically provided long runs of continuous data; today they may be linked with other sources such as social security and labor market agency records. One major step in the case of surveys has been the establishment of panel studies such as the Michigan Panel Study of Income Dynamics (see Brown et al. 1996) and the German Socio-Economic Panel (see Burkhauser and Wagner 1994), which provide data on the same individuals or households over a span of years. Of particular significance has been the assembly of data sets that can be compared across countries. Here the way has been pioneered by the Luxembourg Income Study (LIS), which brings together microdata on households; this provided the basis, for example, for the comparative study of income inequality in OECD countries published by the OECD in 1995 (Atkinson et al. 1995).

Different sources of data have different strengths and weaknesses. Household surveys are perhaps the most common source. They suffer from non-response that is differential across groups of the population, for which only approximate corrections can be made. Among respondents, there is the problem of people failing to disclose income sources or understating income. Comparisons of total household income reported in surveys with National Accounts data suggest an average rate of underestimation between 10 and 20 percent of total disposable income, reflecting both differential nonresponse and under-reporting by respondents. Administrative data may be more complete, although declarations to fiscal authorities may not be more reliable than those of household surveys. Administrative data may also be less tailored to the purpose of studying income distribution, for example containing less contextual data and employing income definitions that reflect tax law rather than the comprehensive ideal described above.

Income data generally refer to a well-defined observation period: the week, the month, or the year. The shorter the observation period, the less subject the earnings information may be to inaccuracy but the more contaminated are the data by transitory income components—premium payments, sickness leave, delay in transfer payments, etc. In the case of investment incomes, payments may only be made annually. On the other hand, where the data refer to an extended period, such as a year, they may be less accurate if based on recall questions. Also account has to be taken of what to do with people who are present in the sample for only part of the period.

In some income distribution estimates, information is combined from several sources to yield 'synthetic' estimates. For instance, income tax data on higher incomes may be merged with household survey data for the rest of the distribution, drawing on their relative strengths. The estimates may be adjusted using national accounts or administrative data. Taxes and transfers may be calculated using a simulation model and added to a survey data set. Such a procedure may be required where the original survey does not contain the information, or where the tax information in the survey relates to a different time period from the income information.

All data sources are imperfect, and any empirical study should include information that can be used to assess the reliability of the conclusions.

See also: Class: Social; Equality and Inequality: Legal Aspects; Generations, Sociology of; Inequality; Inequality: Comparative Aspects; Ricardo, David (1772–1823); Social Stratification; Wage Differentials and Structure; Wealth Distribution; Welfare Programs, Economics of

Bibliography

Atkinson A B, Bourguignon F 2000 (eds.) *Handbook of Income Distribution*. Elsevier, Amsterdam

Atkinson A B, Rainwater L, Smeeding T M 1995 *Income Distribution in OECD Countries*. OECD, Paris

Becker G S 1967 *Human Capital and the Personal Distribution of Income*. W S Woytinsky Lecture No. 1, Institute of Public Administration, University of Michigan, Ann Arbor, MI

Brown C, Duncan G, Stafford F 1996 Data watch: The panel study of income dynamics. *Journal of Economic Perspectives* **10**(2): 155–68

Burkhauser R V, Wagner G 1994 The socio-economic panel after ten years. In: Burkhauser R V, Wagner G (eds.) Proceedings of the 1993 International Conference of German Socio-Economic Panel Study Users. *Vierteljahrsheft zur Wirtschaftsforschung* **1/2**: 7–9

Deaton A, Muellbauer J 1980 *Economics and Consumer Behavior*. Cambridge University Press, Cambridge, UK

Mincer J 1958 Investment in human capital and personal income distribution. *Journal of Political Economy* **66**: 281–302

Pareto V 1897 *Cours d'Economie Politique*. Pichon, Paris, Vol. 2

Ricardo D 1911 *Principles of Political Economy*. Dent, London

Sen A 1992 *Inequality Reexamined*. Harvard University Press, Cambridge, MA

Silverman B W 1986 *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London

Stiglitz J E 1969 Distribution of income and wealth among individuals. *Econometrica* **37**: 382–97

Tinbergen J 1975 *Income Distribution*. North-Holland, Amsterdam

A. B. Atkinson and F. Bourguignon

Income Distribution: Demographic Aspects

Demographic variables are often cited as important determinants of the distribution of income. The following sections discuss the effects on the distribution of income of three demographic variables: (a) age structure; (b) marriage and household composition; and (c) differential fertility by income class. These are topics that have received considerable attention and demonstrate important issues in analyzing demographic aspects of inequality. For a more complete review of these issues and other important demographic aspects of income distribution that are not treated here, see the surveys by Pestieau (1989) and Lam (1997).

1. Age Structure and Income Inequality

One of the most important demographic variables differentiating populations is their age composition. To the extent that changes in age structure affect economic variables such as saving rates and the age profile of wages, there are likely to be effects of age structure on the observed distribution of income. Even if the direct economic effects of age structure are small, age structure may affect standard measures of income inequality for purely compositional reasons. We may want to consider, for example, whether a population dominated by young workers tends to have higher income inequality than a population dominated by older workers, simply because of compositional effects of age structure.

The concave relationship between age and earnings observed in virtually all populations will tend to create a relationship between the age structure of the population and overall income inequality. There is no obvious prediction about whether a younger or older population will have higher inequality, however, based on this component alone. An additional factor is that age-specific income inequality tends to vary with age. One theoretical argument about this relationship was advanced by Mincer (1974) based on a model of life-cycle investments in human capital. Mincer argued that at young ages some individuals trade off current earnings for schooling and training, creating relatively high intracohort inequality. As better-trained individuals reach the 'overtaking' age at which their steeper age-income profiles cross the profiles of individuals with less training, there is relatively low inequality. Beyond the overtaking age the income profiles for better-trained individuals rise above the profiles for lesser-trained individuals, causing inequality to rise at older ages. Taking a different theoretical approach, Deaton and Paxson (1997) note that the life cycle theory of savings implies that within-cohort inequality in income and consumption will increase steadily over the life cycle as a result of the accumulation of transitory shocks experienced over individual lifetimes. They provide empirical evidence of just this sort of 'fanning out' of within-cohort incomes over the life cycle using data from Taiwan, the US, Great Britain, and Thailand.

The fact that both the mean and variance of intracohort earnings vary as a function of age makes it difficult to predict the overall effect of population age structure on income inequality. As noted by Lam (1984) and Deaton and Paxson (1997), the effects of age structure can be productively modeled as a decomposition into between- and within-cohort components of overall inequality, exploiting the features

of stable population models. Changes in age structure affect both within- and between-cohort components, with changes potentially in opposite directions. Deaton and Paxson find important interactions between population growth and economic growth in analyzing the effects of age structure on inequality. With low economic growth, an older age structure tends to increase within-cohort inequality, but it tends to decrease between-cohort inequality by an offsetting amount. With more rapid economic growth, however, an older age structure implies increases in both within-cohort and between-cohort inequality, leading to potentially sizeable increases in overall inequality associated with population aging.

2. *The Role of Marriage and Household Composition*

A standard issue in analysis of income inequality is whether the unit of analysis should be individuals, families, or households. There are many good arguments for using families or households for analyzing income inequality, but moving beyond individual incomes introduces a host of demographic issues relating to marriage, fertility, and household structure.

One frequently analyzed link between demographic behavior and the distribution of income works through marriage markets and the correlation in spouses' incomes. A large literature has developed analyzing the effect of increasing women's labor force participation rates on the distribution of family income. The issue is directly related to the nature of assortative mating on wages and the relationship between one spouse's wage and the other spouse's labor supply. An early important contribution to this literature by Mincer (1962) noted the empirical regularity of a negative relationship between a husband's income and his wife's labor supply. Mincer argued that this negative relationship will tend to have an equalizing effect on family income because the labor market earnings of married women will have a bigger effect on families with poor husbands than those with rich husbands. As with many other issues discussed here, the relationship between women's labor supply and income inequality turns out to be less predictable than Mincer's conjecture suggested, although Mincer's analysis and the decomposition approach he introduced had a major impact on subsequent research.

A large literature has continued to analyze this issue in recent decades, a period characterized by rapid increases in female labor force participation and substantial changes in the distribution of both individual and family incomes in many countries. Decompositions of family income inequality into components attributable to different family members have been a staple of this literature. These decompositions produce conflicting results, sometimes implying an equalizing effect of wives' earnings and sometimes suggesting a disequalizing effect. Cancian and Reed (1999) provide a recent update on the US experience. Their decompositions show rising correlations in husbands' and wives' earnings, a disequalizing trend that is offset by large declines in women's earnings inequality. They estimate that the net contribution of wives' earnings in explaining recent increases in US family income inequality is relatively unimportant, although other authors estimate larger effects using different time periods and different methodologies.

A related issue is the role that household size and composition play in analyzing income inequality. Kuznets (1976) explores the relationship between household income and household size and resulting implications for the distribution of household income. A consistent cross-national empirical regularity demonstrated by Kuznets is a tendency for household size to vary directly with total household income but inversely with per capita household income. This empirical pattern does not imply a general prediction about whether total household income or per capita household income will be more equal, however, and numerous empirical investigations indicate no consistent pattern across countries. A useful approach is provided by Schultz (1982), who decomposes per capita household income into components related to total household income, the number of adults, and the number of children. Schultz finds that total household income is more equal than per capita household income in Colombia, while the opposite is the case in rural India, a result he attributes in part to a stronger negative relationship between income and fertility in Colombia.

3. *Differential Fertility, Intergenerational Mobility, and Inequality*

An argument that is often made connecting demographic change and income inequality is that differential fertility rates across income groups will lead to increasing income inequality across generations. The dynamics of differential fertility and the distribution of income turn out to be quite complex, even in relatively simple models. A considerable literature has developed in economics, demography, and sociology analyzing these dynamics, with several important points emerging.

Consider the question of whether higher fertility rates for the poor will lead to increases in the proportion of poor across generations, and the impact such changes would have on the distribution of income. Analyses of the effects of differential fertility on the distribution of income are more realistic if they recognize that children will not necessarily inherit the income class of their parents. Preston (1974) provides an important demonstration of the role of inter-

generational mobility in analyzing the relationship between differential fertility and the distribution of occupations. This analysis builds on basic principles of Markov processes, noting that intergenerational mobility introduces stability into intergenerational dynamics, even in the presence of differential fertility. Lam (1986) applies similar models to the distribution of income, analyzing the dynamics of a system in which differential fertility across income classes is combined with intergenerational mobility. In such a system it is not the case that higher fertility rates by the poor cause the population to eventually become 100 percent poor. As long as there is sufficient intergenerational mobility there will be a steady state, with differential fertility affecting the nature, but not the existence, of that steady state.

Lam (1986) shows that these models can in general produce surprising results, with no general predictions about the effect that higher fertility for the poor will have on either the proportion poor or the level of inequality in the steady state. Some additional reasonable restrictions on the models can lead to plausible predictions, however. Chu and Koo (1990), for example, show that if fertility falls with income, and if children of poor parents have lower probabilities of being rich than the children of rich parents, then a decrease in the fertility of the poorest income class will cause a reduction in steady-state income inequality.

A further important refinement on these models is to relax the Markovian assumptions, making fertility, investments in education, and wages endogenous with respect to the skill distribution in each period. Kremer and Chen (1999) present a model in which multiple equilibria are generated. One equilibrium is characterized by high inequality and high income differentials in fertility, while the second equilibrium is characterized by low inequality and low income differentials in fertility. The model implies that exogenous investments in education of poor children could induce a movement from the high-inequality to the low-inequality regime.

4. Conclusion

The topics discussed above demonstrate some common themes in research on the demographic aspects of inequality. Working with measures of dispersion tends inherently to complicate analysis of the links between demographic change and inequality. In each of the literatures discussed above it is easy to find examples of plausible-sounding conjectures that turned out to ignore important complications in the demography or the economics. Significant advances have been made in each of these areas in recent years by creatively combining modeling of the population dynamics with careful attention to economic–demographic interactions, and by incorporating these models with the wealth of household survey data that is now available.

See also: Age Structure; Aging and Health in Old Age; Economic Development and Women; Feminist Economics; Fertility: Political and Political–Economic Perspectives; Fertility: Proximate Determinants; Income Distribution; Inequality; Labor Supply; Mobility: Social; Socioeconomic Status and Health; Wealth Distribution

Bibliography

Cancian M, Reed D 1999 The impact of wives' earnings on income inequality: Issues and estimates. *Demography* **36**: 173–84

Chu C C, Koo H W 1990 Intergenerational income-group mobility and differential fertility. *American Economic Review* **80**: 1125–38

Deaton A S, Paxson C H 1997 The effects of economic and population growth on national saving and inequality. *Demography* **34**: 97–114

Kremer M, Chen D 1999 Income-distribution dynamics with endogenous fertility. *American Economic Review* **89**: 156–60

Kuznets S 1976 Demographic aspects of the size distribution of income—exploratory essay. *Economic Development and Cultural Change* **25**: 1–94

Lam D 1984 The variance of population characteristics in stable populations, with applications to the distribution of income. *Population Studies* **38**: 117–27

Lam D 1986 The dynamics of population growth, differential fertility, and inequality. *American Economic Review* **76**: 1103–16

Lam D 1997 Demographic variables and income inequality. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam

Mincer J 1962 Labor force participation of married women: A study of labor supply. National Bureau of Economic Research, *Aspects of Labor Economics*. Princeton University Press, Princeton, NJ

Mincer J 1974 *Schooling, Experience, and Earnings*. Columbia University Press, New York

Pestieau P 1989 The demographics of inequality. *Journal of Population Economics* **2**: 3–24

Preston S H 1974 Differential fertility, unwanted fertility, and racial trends in occupational achievement. *American Sociological Review* **39**: 492–506

Schultz T P 1982 Family composition and income inequality. *Population and Development Review* **8** (Suppl.): 137–50

D. Lam

Income Taxes

In the eyes of the taxpayer, income tax is the most visible form of taxation and therefore also politically the most controversial. Controversies surround the issues of how income tax affects individual behavior in areas such as labor supply, saving, home production, and tax evasion, as well as how it affects the overall distribution of income and the standard of living.

Even more controversial, obviously, is the use of income tax to redistribute income, which raises the more normative question of how progressive income tax ought to be.

1. The Nature of Income Taxation

Income tax is the prime example of a *direct* tax; it is levied directly on the incomes of the individual consumers in the economy, and it can be designed to take account of the individual circumstances of the taxpayer. In contrast, indirect taxes are levied on anonymous market transactions and are paid at the same rate irrespective of the taxpayer's income, number of children, age, health—to mention just a few of the taxpayer characteristics that have over time been taken into account in the design of the personal income tax. It is this tailoring of the tax to the individual's *ability to pay* which makes it such a potentially important instrument for the redistribution of income.

The definition of taxable income varies in the legal systems of different countries and over time. Labor income—wages and salaries—forms the central element of the tax base. To this is added income from capital in the form of interest, dividends, etc. However, income from capital is a more elusive concept than income from labor. Thus, countries differ in the extent to which capital gains are subject to taxation, and they differ in the extent to which non-cash income, such as the imputed income from owner-occupied housing, is subject to tax. Perhaps even more important, rules differ regarding the deductibility of negative capital income, such as interest payment on mortgage loans. Tax laws also differ in the extent to which they allow the deduction of particular expenses, such as the cost of travel to work, the cost of child care, trade union fees, etc.

These differences in the tax base may reflect different perceptions among policymakers of what is a just and fair tax system; they may also reflect differences in the balance of political power among various special interests. What is clear is that different legal definitions of the tax base create different economic incentives for the taxpayers as well as different consequences for the after-tax distribution of income.

Income tax may be progressive, proportional or regressive. A proportional tax is one where the *average tax rate* is constant; the proportion of income paid in tax is the same for all levels of income. Note that this implies that the *marginal tax rate*—the percentage tax paid out of a small increase in income—is also constant. In a progressive tax system the average tax rate increases with income, and this implies that the marginal rate is higher than the average rate. Finally, in a regressive tax system the average tax rate is decreasing, with the marginal tax rate being lower than the average rate.

Countries differ considerably both in their legal definitions of the tax base and in the degree of progressivity of their income tax schedules. There is also significant variation between countries as regards the importance of income tax as a source of revenue. To illustrate, in 1993, on average for the EU as a whole, the personal income tax (inclusive of social security taxes) accounted for 36.4 percent of total tax revenue, whereas in France it was 27.5 percent and in Denmark 54.7 percent (Sørensen 1998).

2. The Income Tax and Economic Behavior: Labor Supply

How do variations in income tax rate affect the labor supply of individuals and households? The economic theory of labor supply takes its point of departure from a model of a single individual who faces certain market opportunities. In the simplest version of this model, the wage rate per hour worked is constant, so that earnings are proportional to hours worked. In deciding how many hours to work, the individual takes account both of the fact that higher earnings make it possible to consume more, but that it also costs something in the form of less leisure. At the optimal number of hours worked, it must be the case that the subjective marginal value of leisure equals the real aftertax wage rate. For longer hours of work the subjective value of leisure is higher than the wage rate; at shorter hours it is lower.

Now consider what happens if the marginal tax rate increases. Two effects can be identified. First, since the reward to labor at the margin falls, there is an incentive to work less; this is called the substitution effect. Second, since the tax increase means that the taxpayer is less well off than before, his or her marginal valuation of leisure falls; this is what is known as the income effect, which in itself tends to increase work effort. Clearly, the two effects pull in opposite directions, so that it is not possible to formulate any clear hypothesis about the connection between taxes and labor supply.

A common objection to this theory is that there are in fact few individuals who are able to choose their hours of work freely, and that the theory therefore tends to focus attention on a problem of little importance. In the short run this is certainly true for a large number of employees. However, the theory can be interpreted in a number of different ways to make it applicable to many more aspects of labor supply. Consider, for example, a young worker who is entering the labor force and considering different job opportunities. Some jobs are offering long hours and high earnings, while others promise more leisure time at the cost of lower earnings. The occupational choice between job alternatives can in fact be modeled in a way that is formally similar to the hours of work model, and the substitution and income effects of

taxation emerge in this setting as well. The same is true of models of the choice of retirement age; a tax increase decreases the reward to a further extension of one's working life and creates an incentive to earlier retirement (the substitution effect), but it also makes it necessary to work longer than before to achieve a given standard of living during retirement (the income effect).

A number of attempts have been made to estimate models of labor supply and taxation using econometric methods. Since the nature of the data sets employed in these studies differ considerably, one might expect that the results would be highly specific with respect to the country, the time, and the nature of the industry to which the data pertain. This is true to some extent, but there are also some general results that characterize a large number of studies. One is that the labor supply of women tends to be more elastic with respect to tax changes than that of men. When women's marginal tax rates go up, they tend to work significantly less, whereas for men one typically finds very small effects. Attempts have also been made to follow the theoretical model and estimate the substitution and income effects separately. One then frequently finds that the reason that the total effect for men is small is not that both effects are insignificant, but rather that although each is of substantial magnitude, they tend to cancel each other.

How is one to interpret the result that the labor supply of women is more elastic than that of men? The most natural explanation is that in most data sets, women tend to be the secondary breadwinners in the family; they tend to work shorter hours and have lower earnings than their male partners. Variations in family labor supply in response to tax changes will therefore most naturally take the form of adjustments in the degree of labor force participation by women.

In the theoretical framework sketched above, 'leisure' was implicitly defined as 'the amount of time not spent working in the labor market.' But in fact people spend a lot of time on household work; studies of time use tend to show that household work takes up about as much time on the average as work in the labor market. The substitution effect of higher taxes in favor of leisure is accordingly also in favor of household work; in other words, income tax 'subsidizes' labor used in household production.

The tax rates that are relevant for the theoretical and empirical study of labor supply are most naturally considered to be simply the statutory rates of the tax law. However, in some contexts it is necessary to take a broader view. Suppose that an individual or household receives a transfer, in cash or in kind, which is income contingent; in order to qualify for it, one's income must not exceed a certain limit. Imagine now that the income of a particular household increases from somewhere below this limit to somewhere above. This increase is subject to the statutory rate of tax, which, in the case of a low-income household, can be supposed to be relatively low. However, in addition, the household loses its transfer, and this loss is in economic terms exactly like a tax increase. The combined effect of the statutory rate and the loss of the transfer payment results in an *effective tax rate* which could be much higher and might even exceed 100 percent, thus creating a powerful disincentive to work among low-income families, often referred to as a 'poverty trap.' This is not just a theoretical possibility but has been identified as an empirical fact in the tax and transfer systems in a number of countries. Clearly, it is very important to consider the interaction between tax and transfer systems in order to avoid unintended and unfortunate consequences of their combined effects.

3. *Income Tax and Economic Behavior: Saving and Portfolio Choice*

The base on which income tax is levied includes income from capital as well as from labor. A simple extension of the theoretical framework so far is to imagine that the individual lives in two periods, and that he or she saves some of his or her labor income in the first period. This will then become available, with interest added, in the second period. Interest income is subject to tax. What happens to saving when the tax rate increases?

As in the labor supply case, we can identify a substitution and an income effect. As the reward for saving (the aftertax interest rate) falls, there is an incentive to save less. On the other hand, the lower the net-of-tax interest is, the more it is necessary to save to achieve a given level of future consumption. The substitution and income effects clearly work in opposite directions. However, we also need to take account of the fact that some people are not net savers in the first part of their life. Instead, they borrow, with the loan to be repaid with interest in the later part of their life cycle. If the interest on the loan is tax deductible, a higher rate of tax provides an incentive to borrow more, i.e., to save less. This, again, is the substitution effect. But now the lower interest rate makes it possible to borrow more than before without having to reduce future consumption; thus the income effect on saving for a borrower is negative, not positive. Since some people are borrowers and some lenders, the income effect for all consumers together might not be very significant and the substitution effect would tend to dominate. Empirical studies also tend to show that there is in fact a negative tax effect on saving, although the magnitude is not always very strong.

This picture of tax effects on saving seems to assume that all saving takes the effect of lending or borrowing in some kind of homogeneous asset. In reality, of course, people hold a portfolio of assets (housing, bank deposits, mortgage debt, shares, etc.) and income tax can therefore be expected to influence not only the

aggregate amount of saving but also the composition of the savings portfolio. This is particularly so since income tax law in many countries does not provide for the taxation of different asset returns at a uniform rate. While interest on bank deposits is often taxed at the statutory rate, interest on some types of bonds may be tax exempt, capital gains may be tax free provided one has held the stock for a sufficiently long time, and the returns from collector's assets such as art, antique furniture, and stamps are usually not taxed at all. Empirical studies also seem to indicate that portfolio composition is indeed very sensitive to tax effects.

In the academic literature there is one type of tax effect which has received special attention by economic theorists. Some assets are relatively safe in the sense that their return can be predicted with a high degree of certainty; others are more risky. Could income tax be expected to change the balance between safe and risky assets in savers' portfolios? Suppose that there is full loss offset, so that any negative returns on risky assets can be deducted against the remaining income. Then one can deduce, basically from the assumption that the individual saver is averse to risk, that an increase in the rate of income tax will induce the saver to invest *more* in the risky asset than before. Although the government takes a share of the expected return, it also shares in the risk, thereby protecting the saver from the full consequences of low or negative returns. This theoretical insight has had a considerable influence on the way that economists view the bias of investment incentives related to income tax, but the hypothesis is not an easy one to test on empirical data. The major reason for this is, as argued above, that there are so many other tax incentive effects on portfolio composition that the risk substitution effect becomes very hard to trace.

4. *Corporate Income Tax*

All income ultimately accrues to private individuals, so that the tax on corporate income or profits is in the last instance also a tax on individual consumers. In fact, some economists have suggested that it might be a good idea to abolish the tax on corporations and instead collect all income from corporations at the hands of the individual stockholders. However, abolishing the tax on corporations might not be a politically popular issue, and in addition there are some real economic arguments why it should be preserved. One is that collecting the tax at the corporate level helps to reduce incentives for tax avoidance and evasion; another is that it is an efficient way to collect some tax revenue from foreign owners of domestic companies who are not subject to the domestic tax on personal income.

In the theoretical literature there is a further argument in favor of the corporate income tax. Assume that corporations maximize profits and that markets are efficient. Then it would be desirable to have taxes that do not lead to any distortions of production and investment decisions in corporations. However, if corporations are subject to a (say) 40 percent tax on profits, the decisions that would maximize profits in the absence of tax are obviously the same as would maximize 40 percent of profits. Hence the corporate income tax is *neutral*; it collects revenue for the government without leading to any distortions of decisions at the firm level and leaves the efficient market outcome unaffected.

However, this picture is very much simplified. The main reason for this is that the tax definition of corporate income is in several respects different from the concept of profits. One important source of deviation between the two is the treatment of real capital depreciation. The real depreciation of a piece of machinery or a building or another type of capital equipment during a year is part of the costs of the corporation, and in order for the corporate income tax to be neutral, real capital depreciation should count as part of the costs in arriving at a definition of taxable income that corresponds to true profits. However, a corporation may employ thousands of different types of real capital, and it is not possible for the tax authorities to allow for the deductibility of true depreciation for all of these. Instead, they typically allow standard depreciation rates for large groups of capital assets such as machines, buildings, transportation equipment, and so on. But this means that for some types of capital the corporation is allowed to deduct less than the true depreciation, for others more. In its interaction with the tax depreciation rules the corporate income tax in fact works as a selective tax on some types of capital equipment and as a subsidy on others. Similar issues may arise regarding interest deductibility, stock valuation, and capital gains on both real and financial assets. It is therefore not surprising that empirical studies tend to show that there are substantial tax effects on production and investment decisions in corporations (Jorgenson and Yun 1991).

5. *Tax Evasion*

Tax evasion is an illegal activity which aims to hide taxable income from the view of the tax authorities. It should be distinguished from tax *avoidance*, which consists in trying to reduce one's taxable income by exploiting the tax law while staying inside its boundaries. While the distinction between the two may sometimes be unclear, there are at least a large number of cases that definitely belong in the evasion category; these are sometimes referred to as the hidden or black economy.

Tax evasion is a risky activity. By holding back information from the tax authorities, there is always a risk of being discovered, in which case one typically

faces a penalty rate on the unreported income. The gain on a dollar of unreported income, i.e., the regular rate of income tax, must be balanced against the probable loss, which is the probability of detection times the penalty rate of tax. When the regular income tax rate goes up, the gain increases, and it thus tends to encourage evasion, e.g., in the form of working in the hidden economy. Theory seems to support the common concern that a high tax level encourages cheating and dishonesty. However, this should not lead us to expect that high-tax countries necessarily have more tax evasion than countries with lower tax levels. The extent of evasion depends not only on the level of income taxation, but also on a number of other social and economic factors, and these are likely to vary considerably among countries.

Theoretical hypotheses in this area are particularly hard to confront with empirical data; in the nature of things, there are no official statistics for the hidden economy, and survey research faces some special difficulties in making individuals answer questions truthfully. A number of the empirical studies that have been made use very indirect methods, such as deducing the size of the hidden economy from the public's holding of cash, and the reliability of this kind of approach is highly disputed (Cowell 1990). More standard approaches tend to yield estimates which indicate that the hidden economy has a volume of 2–10 percent of the GNP of the official economy. Many studies also indicate that evasion is higher for self-employed than for salaried employees. This is in good accordance with the theory, for wages and salaries are usually reported directly by employers, so the probability of evasion being detected is much higher for employees than for the self-employed, implying that evasion is a less risky gamble for the latter.

It should be added that the emphasis on rational risk-taking behavior as an explanation of tax evasion does not imply that in the economist's view of the world all taxpayers are completely amoral, reporting their income only when a cost–benefit calculation has shown them that evasion does not pay. There are strong indications that many taxpayers report their true income even when under-reporting could easily go undetected. The chief aim of the theory is to explain the behavior of those who do evade taxes, and although they too may take moral considerations into account, they are also likely to consider the risk-taking aspects of their activities in a rational manner.

6. *Income Tax and the Redistribution of Income*

It is widely accepted in most countries that income tax should take account of the taxpayer's ability to pay, and the most common interpretation of this principle is that the tax should be progressive. This raises the question of whether income tax in fact has led to a situation where post-tax or disposable income is more evenly distributed than pretax or factor income. In most empirical studies the answer to this question is an affirmative one, although the estimated magnitude of the redistributive effect varies substantially. A general feature of empirical studies is that the redistributive effect is smaller than the formal progressivity of income tax schedule would lead one to expect, and the main reason for this is the role played by exemptions or tax credits. As an example, suppose that mortgage interest payments are fully deductible and that their share of gross income tends to increase with income. Then, as gross income increases, taxable income increases less than disposable income in percentage terms, and this mitigates the increase in the effective average tax rate. The importance of this effect has been well documented in a number of studies (Wagstaff et al. 1999).

From a theoretical point of view, one would expect that the progressivity of the tax system would have an effect on pretax incomes. The markets at least for some occupations would react to high marginal tax rates by increasing gross incomes; this is the problem of *tax incidence*. Since the gross incomes that would have existed had there been no income tax or a purely proportional tax are unobservable, this observation makes it much more difficult to study the redistributive impact of income tax. In this connection it is interesting to observe that countries with a high degree of pretax inequality also tend to be the ones with high post-tax inequality, and moreover that the redistributive impact of income tax is high in some countries which have a low degree of inequality in pretax incomes (Wagstaff et al. 1999). However, it is not clear whether this means that the incidence problem is in fact of little importance or whether there is some other explanation. If a country is dominated by egalitarian social and political values, these may influence not only the progressivity of the tax system, but also the structure of pretax earnings.

7. *Income Tax, Economic Efficiency, and the Equality–Efficiency Tradeoff*

Income tax has always been under attack from critics for leading to distortions of economic incentives and consequently to inefficient social outcomes. The criticism has been especially focused on the distortion of work incentives, which leads to incentives to work less than would be desirable from a social point of view. As pointed out in Sect. 2, there is clear evidence for such an effect in the case of women, whereas for men the net effect tends to be very small. However, in considering the efficiency loss from income taxation, it is the substitution effects, and not the total effects, that should concern us. The reason is that the income effects would always be present whatever the form of the tax system (and given the amount of revenue to be

collected). The substitution effects are the direct consequences of the distortion of wages as efficient market prices, and if our sole concern is with efficiency, we should attempt to design the tax system in such a way as to make the substitution effects as small as possible. Similar concerns are relevant in judging the effects of income taxation on saving and investment behavior. The taxation of interest income might lead consumers to save less and to invest in the assets or industries with the most favorable tax rules rather than in those with the highest rates of return.

However, efficiency is not the only standard by which to judge income tax. As demonstrated in the progressive structure of income tax in most countries, one of the goals of income taxation is to redistribute income from the rich to the poor, and the gains on the side of redistribution must be set against the cost of a less efficient allocation of resources. This is a recurrent problem in the design of economic policy; there is a need to choose a tradeoff between equality and efficiency (Okun 1975). However, if more equality is a real goal of economic policy, it should presumably be worth some cost in terms of less economic efficiency in order to achieve it. These ideas have been formalized in the theory of optimal income taxation, which attempts to show how the optimal marginal rate of income tax depends both on the efficiency cost, as measured by the elasticity of labor supply, and on the social attitude to inequality. Clearly, what economic theory can contribute to this issue is not to produce exact numbers, but to help establish a logically consistent way of thinking about a complex issue that involves both economics and ethics (Sandmo 1999).

8. *Proposals for Tax Reform*

A number of proposals have been advanced for basic reform of the income tax system, frequently motivated by the desire to improve the terms of the equality–efficiency tradeoff. For example, is it possible to achieve the same degree of redistribution at a lower efficiency cost? Tax reforms in a number of Western countries since the 1980s have tried to move in this direction by decreasing the formal degree of progressivity in the tax system, while at the same time eliminating many of the deductions and exemptions that had contributed to a lower degree of effective progressivity. Studies of the American (Auerbach and Slemrod 1997) and Swedish (Agell et al. 1996) reforms of 1986 and 1991 do in fact conclude that their overall effects have been to leave the redistributive effects of the tax system virtually unaffected, while removing many of the tax distortions both in labor and capital markets.

A more radical proposal for reform is the adoption of a so-called flat tax. There are several versions of this, but one that was proposed by Hall and Rabushka (1995) and has been much discussed in the USA involves imposing a constant marginal tax rate on all income above a certain exemption level. Although the name of the tax indicates that the major point in its favor is the simplification of the tax structure, probably its more important features are the elimination of a large number of deductions and a more uniform taxation of income from capital. In this it is in line with the general trends in the reforms of the 1980s and 1990s.

Another reform, the intellectual history of which goes back a long time and which was much discussed during the late 1970s and early 1980s, was the possible adoption of an expenditure tax (Meade Committee 1978, Kay and King 1990). Although the expenditure tax is often seen as an alternative to income tax, it may also be regarded as simply a particular form of income taxation. Briefly, an expenditure tax is an income tax with full deduction for all saving, but with negative saving being added to income. What is taxed is accordingly expenditure or consumption, not income. There are two main claims for the expenditure tax. First, since differences in the standard of living between persons are primarily related to their consumption, not income, consumption is the more logical base on which to level a progressive tax. Second, the treatment of saving and dissaving leads to a more uniform and therefore more efficient taxation of capital income.

Although neither the flat tax nor the expenditure tax have so far achieved any direct political success, the political and economic debates over these proposals may indirectly have had a considerable influence on the reforms which were actually adopted in the following years.

See also: Income Distribution; Labor Supply; Social Justice

Bibliography

Agell J, Englund P, Södersten J 1996 Tax reform of the century—the Swedish experiment. *National Tax Journal* **49**: 643–64

Auerbach A J, Slemrod J 1997 The economic effects of the Tax Reform Act of 1986. *Journal of Economic Literature* **35**: 589–632

Cowell F A 1990 *Cheating the Government*. MIT Press, Cambridge, MA

Hall R E, Rabushka A 1995 *The Flat Tax*, 2nd edn. Hoover Institution Press, Stanford, CA

Jorgenson D W, Yun K-Y 1991 *Tax Reform and the Cost of Capital*. Oxford University Press, Oxford, UK

Kay J A, King M A 1990 *The British Tax System*, 5th edn. Oxford University Press, Oxford, UK

Meade Committee 1978 *The Structure and Reform of Direct Taxation. Report of a Committee Chaired by Professor J E Meade*. Allen & Unwin, London

Okun A M 1975 *Equality and Efficiency: The Big Tradeoff*. The Brookings Institution, Washington, DC

Sandmo A 1999 Asymmetric information and public economics: the Mirrlees–Vickrey Nobel Prize. *Journal of Economic Perspectives* **13**: 165–80

Sørensen P B (ed.) 1998 *Tax Policy in the Nordic Countries*. Macmillan, London

Wagstaff A, van Doorslaer E, van den Burg H, Calonge S, Christiansen T, Citoni G, Gerdtham U-G, Gerfin M, Gross L, Häkinnen U, John J, Johnson P, Klavus J, Lachaud C, Lauridsen J, Leu R E, Nolan B, Peran E, Propper C, Puffer F, Rochaix L, Rodriguez M, Schellhorn M, Sundberg G, Winkelhake O 1999 Redistributive effect, progressivity and differential tax treatment: personal income taxes in twelve OECD countries. *Journal of Public Economics* **72**: 73–98

A. Sandmo

India, Religions of

Indian civilization contains a great number of religious traditions, such as Hinduism, Buddhism, Jainism, Islam, Christianity, and Sikhism. All these traditions have influenced each other and have become part and parcel of Indian civilization in one way or another. A large number of Indian traditions refer, positively or negatively, to a set of religious books, called the Vedas. These scriptures were, originally, the intellectual property of a priestly caste of Brahmans and the basis of their rituals and philosophies. Brahmans as a caste form the apex of a hierarchy of status categories which can be seen as an early form of the later much more elaborate Indian caste system. Buddhism and Jainism emerged around 500 BC in protest against the Brahmanical monopoly of Vedic ritual. They are essentially ascetic religions which have much in common with other speculative philosophies which continued to accept the authority of the Vedas. Buddhism spread from India to the rest of Asia, but became marginal in India, while Jainism did not expand outside of India, but continued in India as the religion of, primarily, businessmen. Christianity came to Kerala in the first centuries of the Common Era and in other varieties with the Portuguese to Goa in the sixteenth century and with other Europeans in the centuries thereafter. Islam came to India at the end of the first millennium, along the coast with traders and in the northern plains as an element of the military and religious expansion of a variety of groups from West and Central Asia.

1. *Hinduism*

The majority of the Indian population today calls itself Hindu. Hinduism is a term that only in the eighteenth century came to be used for a variety of religious traditions which dominate the Indian religious landscape. When they are Brahmanical, these traditions refer to the Vedas, but when they are not they can refer to a number of religious texts, oral or scriptural. Gods like Vishnu and his reincarnations Rama and Krishna are very popular. Likewise, a god like Shiva, often represented in an aniconic, phallic form, is worshipped all over India. In many localities one finds local manifestations of gods like Vishnu or Shiva. A special and important category of gods are Mother Goddesses which often protect against evil and diseases. The typical form of worship of images of gods is the offering of gifts like flowers or sometimes goats or buffaloes. These temples are serviced by Brahman or ascetic priests. Temples are everywhere, but especially in sacred places, which are somehow connected to the story of the god. Hindus make pilgrimages to these places, especially at auspicious times in the lunar calendar. There is also an elaborate complex of rituals which are performed in the domestic context, often related to moments in the life cycle. Finally, there are groups of priests, specialized in warding off evil and witchcraft and performing mortuary rites.

In terms of the social organization of religion there are five elements in Hindu traditions. First, there is caste. Brahman groups have continued from the earliest times to dominate a large part of religious beliefs and practices. Ritual purity and sacred knowledge are the foundations of their supremacy. Other castes have also created cults of their own, but often they are seen of lower status than the Brahmanical cults. Second, there are rituals of kingship. There is the conception of a generic relationship between royal deity and human king which is reflected in ritual practice. In a large number of regions in India one finds a network of temples and palaces which are mutually interdependent. Third, there is the level of the village. The local community is often ritually represented in a number of festivals. Fourth, there is the notion of world renunciation. Ascetics have for a long time up to the twenty-first century built monastic orders and drawn lay support for their religious and worldly activities. They are innovators of religious traditions, but also as free agents in a village society important long-distance traders, money brokers, and military entrepreneurs until the coming of the colonial state. Fifth, there is the concept of bhakti (devotion). One cult after another sprang up in India during the second millennium, channeling waves of devotional enthusiasm. Often ascetic leaders were involved in this and in a number of cases they revolted against the Brahmans and the caste system. By that token they were sometimes able to find a distinct religious tradition, such as the religion founded by Guru Nanak, which evolved gradually into Sikhism and defined itself as different from Hinduism.

Hinduism is a term which has developed from 'Hindu,' a term used by Muslims who came to the river Indus or Sindh (from Sanskrit Sindhu) in the second half of the first millennium. 'Hindu' designates the

inhabitants of this area of Al-Hind or Hindustan and their customs and beliefs. The term Hindu was made into an 'ism' by the British in the eighteenth century. The conceptualization of Hinduism as one particular religion among other religions and theoretically comparable to them can be understood in relation to a number of philosophical arguments about the universality of the category of 'religion' in the European Enlightenment. The conceptualization of their religious traditions by Hindus themselves had evolved along lines of philosophical argumentation about dharma (the sacred order of things) by religious scholars who are easiest divided in followers of the God Shiva and followers of the God Vishnu. Their philosophical arguments were often accompanied by ritual injunctions, especially for temple worship. Below this level of high philosophical argumentation there were a number of devotional movements which can broadly be divided into those which focus on the worship of images and those which have an aniconic image of God. The British introduced the essentialized and unified concept of Hinduism as a religion in an already existing arena of lively polemics about religious difference.

2. *Modern Period*

In the modern period the expansion of the colonial state is crucial for defining the location of religion in India. In the eighteenth century the British East-India Company followed a traditional policy of religious patronage benefiting the religious establishments in the areas which came under their growing influence. This policy was further supported by a strong interest in Hindu scripture from a viewpoint of comparative linguistics and comparative religion. This so-called 'orientalism' of the late-eighteenth century was vehemently attacked at home and in the colony by utilitarian administrators and evangelical missionaries who were united in their rejection of anything Hindu. In education and religious worship the colonial administration was forced to withdraw from giving its patronage to religious establishments and a policy of 'religious neutrality' was established gradually in the nineteenth century.

The response of Indian religious thinkers, institutions and movements to colonial interference under the aegis of religious neutrality has been central to the formation of modern religions in India during the nineteenth century. Especially attacks by evangelical missionaries on religious practices and social customs led to both reformist and conservative responses. In Bengal, Rammohan Roy (1772–1833) studied Christianity and felt great affinity with the rational critique of religious orthodoxy, launched in Britain by Unitarians. In 1827 he founded the British Indian Unitarian Association, but his study of the Vedas and Upanishads quickly led him to explore the limits of Unitarian universalism. In 1828 he decided that a universal, rational religion had to be based upon the Vedas and Upanishads and created the Brahmo Samaj, a small, but influential elite movement, propagating a deist and universalist kind of religion, based upon Hindu sources. The movement is particularly opposed to 'superstitious customs' of 'ignorant people,' 'deceived' by their Brahman leaders.

Rammohan was in the forefront of public action. In January 1830 he, together with 300 residents of Calcutta, presented a petition to Governor-General William Bentinck in support of government prohibition of widow immolation, a practice found in some parts of India when after the demise of a husband the widow throws herself on the pyre of her husband to be burned with him. Rammohan rejected this practice on the basis of his reading of Hindu scripture. He distinguished authoritative sources, such as the Vedas, from other sources. It is significant that he did not refer to authoritative interpretations by learned Brahmans, but relied entirely on his own judgement. This is an important step towards the laicization of Hinduism, that is to say, the formation of religious movements, carried by lay believers and relatively independent of traditional, sacred interpreters of the Hindu canon. The Brahmo Samaj was the cradle of an intellectual Hinduism which was at the end of the century personified by Swami Vivekananda (1863–1902), the founder of the Ramakrishna movement.

Movements like the Brahmo Samaj are sometimes called Neo-Hindu because they have been clearly influenced by European thought and Christianity. On the other hand, however, they are also in direct interaction with more traditional thinkers and movements. A good example is the illiterate saint Ramakrishna (1836–88), a priest of a Kali temple near Calcutta whose devotional practices, ecstasies, and unconventional behavior was very attractive to the intellectual Brahmo Samaj adherents, because he seemed to signify a certain untainted Hindu authenticity. The great reformer Vivekananda took Ramakrishna as his spiritual guide.

Vivekananda became famous after his visit to the World Parliament of Religions which was organized in Chicago in 1893. He offered a packaged Hinduism, in which spirituality and yoga (mental and physical discipline) were keywords, first to a world audience and later to an audience at home. Central to Vivekananda's religious message is the notion of 'practical Vedanta,' in which he combined the Vedanta (= philosophical commentaries on the Vedas) with participation in social activism rather than with ascetic withdrawal from the world, as had been common in Hindu traditions. It is this message which could be tied up with nationalism and political activism as was done by political leaders such as Mohandas ('Mahatma') Gandhi (1869–1948).

Also outside of Bengal we see a number of movements which interact both with Hindu tradition and

with modernity, as brought by colonial rule. The Arya Samaj is the most important among them. Founded by Dayananda Saraswati (1824–83), an ascetic who had been initiated in one of the most austere monastic orders in Hinduism, it is an iconoclastic movement which identifies image worship with the decline of an earlier pure Vedic religion. It attacked also the caste system and its system of marriage and the traditional Brahmanical priesthood. The Arya Samaj became and remains one of the central religious pillars of Hindu nationalism.

3. *Public Sphere*

It is not only Hinduism which is developing modern movements in the new public sphere created under colonial rule. The general pattern is an attack on traditional priestly leadership, on 'backward,' popular religion and an emphasis on free, lay access to scripture and to religious debate. In this way the foundation of religious authority changes and becomes mediated by education and new forms of communication, such as print, and later radio and television. Muslim movements such as the Deobandis, the Ahmadiyyas, and the modernist Aligarh movement share with the Hindu movements of the period the defense of their religion against the onslaught of missionary criticism and the attempt to formulate an answer to modern intellectual challenges. Clearly, they are all deeply political since they mobilize people against certain aspects of colonial rule as well as against aspects of their religious traditions and traditional leadership. They are the basis for the formation of the Muslim political party, the Muslim League, and for Muslim participation in the Congress Party. Also the Sikhs have to respond to colonial modernity. The British gave the Sikhs preferential treatment in their recruitment for the army, which made the rather fluid Sikh identity both more valuable and more fixed. The Singh Sabha movement is especially important, since it challenges the control over the centers of Sikhism, the temples or gurdwaras. These centers were owned by hereditary priests, but following the general pattern of laicization the Singh Sabha movement demanded that they came under the control of the lay community of Sikhs. Control over the committee which governs these temples is still one of the biggest prizes in Sikh politics. It is also the basis of the power of the Sikh political party, the Akali Dal.

The twentieth century witnessed a further politicization of religious identity. The greatest competition was between Hindus and Muslims. The colonial state reinforced this competition by creating separate job opportunities, separate civil laws, and separate electorates. The advent of mass politics brought with it the mass mobilization of people on the basis of reified identities. Mohandas Gandhi, the great leader of the Congress party, was brilliant in his strategies to dismantle British power, but these strategies were based explicitly upon Hindu practices. Both he and, to a lesser extent, Jawaharlal Nehru were perceived by Muslims as Hindu despite their open secular politics. Radical movements emerged which challenge the idea of India as a plural civilization. 'To be an Indian is to be a Hindu' was a slogan of V. D. Savarkar (1883–1966), one of the founding ideologies of Hindu nationalism. Movements such as the Rashtriya Swayamsevak Sangh (RSS) and the Hindu Mahasabha wanted to establish Hindu hegemony in independent India. The assassins of Gandhi in 1947 emerged from such organizations. On the Muslim side, the Muslim League became gradually more and more radical until it demanded Pakistan, a homeland for Muslims, in 1942 and ultimately forced the Partition of India. The tragedy of the violently enforced migration of Hindus to India and Muslims to Pakistan, the division of East and West Pakistan which led in 1971 to the foundation of Bangladesh, the territorial dispute about Kashmir, and the threat of warfare between these two nuclear powers, were all results of the political mobilization of religious difference on the Indian subcontinent.

4. *Nationalism*

Hindu nationalism needs to construct a Hindu majority among the Indian population to realize its claims to political hegemony. It not only runs against the resistance of non-Hindu religious communities, such as Muslims or Sikhs, but also against that of so-called 'untouchables' and 'tribals' who form a large minority in that section of the population which has to be constructed as a Hindu majority, but who are within Hindu practices severely discriminated. There are a number of anti-Brahman, and sometimes anti-Hindu, movements among these underprivileged groups which demand special protection and positive discrimination from the state. Most of these demands concern education and employment, but participation in Hindu rituals and access to temples are also important in the mobility of these groups. One of the most important movements among 'untouchables' was led by Bhim Rao Ambedkar (1891–1956), a Mahar untouchable from West India who has a doctorate in law from Columbia University. It was Ambedkar, one of the principal authors of the Indian constitution, who late in his life concluded that the untouchables would not be equal citizens in India as long as they continued to be second-rate Hindus. He led a mass conversion to Buddhism, primarily by his own community of Mahars, but also followed by other untouchable groups in India. The conversion to Buddhism, however, has not solved the discrimination of these groups in India and more recently a unifying tactic has been to claim a more secular, militant identity in terms of race rather than religion.

In Indian Islam, one of the major developments in the twentieth century was the founding of the Tablighi

Jama'at in 1927 by Maulana Muhammad Ilyas (1886–1944). In 1934 Ilyas started a great revivalist campaign among Muslims of Mewat, an area near Delhi. Going from person to person, he called upon them to follow their Muslim duties, such as praying in the mosque. Gradually he formed groups of laymen who went for fixed periods on tours to invite their fellow Muslims to perform their duties. This created a snowball effect whereby more and more people became involved in this activity. The groups went around first in the area of Mewat, then in other places in India and Pakistan, and, finally, in the 1950s abroad to countries such as Nigeria, Morocco, Britain, Canada, and France. The movement thus spread from South Asians to other Muslims and is now one of the most successful Islamic movements in the world. The Tablighi Jama'at is professedly apolitical and in that way very different from the major Islamist movement of the subcontinent, the Jama'at-i-Islami. This movement was founded in 1941 by a former journalist, Maulana Maududi (1903–79), one of the most influential Islamist thinkers of the twentieth century. It is a much smaller movement than the Tablighi Jama'at, but highly political and organized as a tightly and hierarchically run revolutionary force which aims to capture the state in order to transform society according to Islamic principles. In Pakistan it has a considerable influence in politics and also in the rest of the Muslim world revolutionary movements of the Islamist persuasion have taken it as an inspiring example.

The 1980s saw the development of some ultra-nationalist religious movements in India. The Vishwa Hindu Parishad, founded in Bombay in 1964 by some religious leaders with the help of the Rashtriya Swayamsevak Sangh, organized a country-wide protest against the conversion to Islam of a group of 'untouchables' in the South Indian village of Meenakshipuram in 1981. The success of this protest led to the staging of another All-India campaign in 1984 to demand the demolition of a sixteenth century mosque in Ayodhya, in North India, which had allegedly been built on the site of a temple commemorating the birthplace of the God Rama. This campaign became one of the most important issues in Indian politics in the decade following its start. In 1992, the mosque was demolished with the tacit acceptance of the Indian state.

The 1980s saw also the development of a radical Sikh separatism in the Punjab, led by Sant Jarnail Singh Bhindranwale (1947–84), a young religious leader. His militant career started with intersectarian strife among the Sikhs, but developed into straightforward confrontation with the Indian state. In 1984 the most important Sikh shrine, the Akal Takht in Amritsar, was captured by the militants who demanded Khalistan, a separate Sikh state, imitating the earlier demand for Pakistan by the Muslims of South Asia. In the same year the Indian army decided to attack the Golden temple and chase the militants. The action caused great disaffection with the state among Sikhs worldwide and led to the assassination of Mrs Indira Gandhi (1917–84), the then prime minister of India.

During the twentieth century there was a growing nationalization of Indian religions, which makes it difficult to separate them from processes of state formation. The migration of South Asians outside India has led to these forms of religious nationalism being spread to the rest of the world.

See also: Area and International Studies: Development in Southeast Asia; Buddhism; Conflict: Anthropological Aspects; Conflict Sociology; Hindu Law; Internal Warfare: Civil War, Insurgency, and Regional Conflict; Marx, Karl (1818–89); Revolution; Revolutions, Sociology of; Revolutions, Theories of; Social Movements, Sociology of; South Asia: Socio-cultural Aspects; South Asian Studies: Culture

Bibliography

Appadurai A 1981 *Worship and Conflict under Colonial Rule: A South Indian Case*. Cambridge University Press, Cambridge, UK

Dirks N B 1987 *The Hollow Crown. Ethnohistory of an Indian Kingdom*. Cambridge University Press, Cambridge, UK

Dumont L 1970 *Homo Hierarchicus: The Caste System and Its Implications*. Weidenfeld and Nicolson, London

Fuller C J 1992 *The Camphor Flame: Popular Hinduism and Society in India*. Princeton University Press, Princeton, NJ

Metcalf B D 1993 Living Hadith in the Tablighi Jama'at. *Journal of Asian Studies* **52**(3): 584–608

Thapar R 1997 Syndicated Hinduism. In: Sonntheimer G D, Kulke H (eds.) *Hinduism Reconsidered*. Manohar, New Delhi, India

van der Veer P 1994 *Religious Nationalism: Hindus and Muslims in India*. University of California Press, Berkeley, CA

Weber M 1967 [1917] *The Religion of India: the Sociology of Hinduism and Buddhism*. Free Press, Glencoe, IL

P. van der Veer

Indicator: Methodology

Concepts in the social sciences are often abstract. Current knowledge or data allows only imperfect empirical representations of them. But abstract ideas such as intelligence, the values of goods, individual attitudes, or social class are so important to the social sciences that we attempt to empirically track them even if our ways of doing so are limited. 'Indicator' is a common term to refer to the variables that we use to detect these concepts empirically. Other terms are *proxy variables*, *items*, *measures*, *scores*, *scales*, *indices*, or *observed variables*. Occasionally some of the terms have a 'part-whole' relation where, for instance, a

scale or an index might refer to the sum of items or the sum of indicators. But there is not complete consistency in these uses so that it is common for researchers to refer to a Likert scale even though it is a single item or indicator. Furthermore, the frequency of the use of these terms varies by discipline. Economists, for example, regularly refer to proxy variables while educational testing psychologists frequently refer to items or scales. Regardless of the name, they all refer to *a variable that is directly associated with a latent variable such that differences in the values of the latent variable mirror differences in the values of the indicator*.

Latent variables are variables that represent concepts. There should be a separate latent variable for each dimension of a concept and a minimum of one indicator per latent variable. The indicator's importance derives from what it reveals about the latent variable. The indicator informs us about the magnitude of the latent variable across cases or over time for the same cases. The key idea is that the indicator provides information on the latent variable. We might, for example, be interested in socioeconomic status (SES) of individuals where we define SES as the overall standing of an individual in the social and economic stratification systems of a community and use a single latent variable to represent this overall standing. Income and education are two indicators of SES, but they do not fully capture SES. Or we could have a test designed to tap the anxiety levels of individuals, but the item or overall test score would not perfectly measure anxiety. So the use of the term indicator and other related terms gives recognition to the fact that the variables we have to work with are not identical to their respective latent variables.

1. Measurement Strategies

When faced with latent variables for which there is no ideal measure, researchers typically follow one of three strategies: (a) ignore the latent variable, (b) include indicators, but ignore their measurement errors, or (c) include indicators, but take account of their measurement error. An example of strategy (a) is a study that is interested in the educational attainment of individuals. Suppose that the researcher recognizes that an individual's motivation or drive plays a role in educational attainment, but given the difficulty of measuring motivation she or he ignores it. The danger of this all too common decision is that the omitted latent variable is likely to bias the estimates of the impact of the other included explanatory variables. So in the educational attainment empirical analysis, part of the effect attributed to say, parental educational attainment might actually be due to the unmeasured motivation variable.

In strategy (b) we could include an indicator of motivation, but ignore its measurement error. Here we have a partial control for and an estimate of the impact of motivation, but bias is likely in both the estimate for motivation and those estimates for the other variables associated with motivation. Contrary to conventional wisdom disregarding measurement error is not a conservative approach. In general, random measurement errors in two or more explanatory variables will not always lead to underestimates of effects of motivation or the effects of the other variables in the analysis (Bollen 1989). A natural question is whether we are better off to include the indicator and ignoring the measurement error as in strategy (b) or ignoring the variable altogether as in strategy (a). Those researchers who have explored this question generally find that it is better to include an imperfect indicator than to ignore the latent variable altogether (Aigner 1974, McCallum 1972, Wickens 1972).

Strategy (c), including indicators and taking account of their measurement errors, is generally the best choice. Here we could include indicators of motivation and control for the influence of the measurement error in these variables. This approach would halt the spread of bias due to omitting the latent motivation variable or treating the motivation indicators as if they were perfect measures. A requirement for taking account of the measurement error is the formulation of a measurement model.

2. Measurement Models

Measurement models refer to the implicit or explicit models that relate the latent variable to its indicators. Several basic questions require consideration prior to building a measurement model: (a) Are there *causal* or *effect indicators*? (b) Are there *multiple* or *single indicators*of the latent variable? (c) Is the latent variable continuous or noncontinuous? (d) Are the indicators continuous or noncontinuous? (e) Does more than one latent variable influence the indicator?

The first question concerns whether the indicators influence the latent variable (causal indicators) or whether the latent variable drives the indicators (effect indicators). The overwhelming majority of social science work on measurement has focused on effect indicators where the indicators depend on the latent variable. Sometimes these are called 'reflective indicators.' Differences in verbal ability, for instance, determine the differences in values for a test of verbal ability. Or an antitax attitude should determine the responses that an individual gives to a question on attitudes toward taxes. Effect indicators are implicitly assumed in factor analysis and in much of measurement theory in the social sciences.

Blalock (1964) called attention to causal indicators, in which the indicator influences the latent variable rather than the reverse. Some refer to these as 'formative indicators.' Education and income of an

individual, for instance, can influence SES making these variables likely to be causal indicators. The amount that a child watches violent television shows, plays violent video games, or views violent movies are causal indicators of the latent variable of exposure to media violence. Often researchers do not consider that an indicator may be causal and treat all variables as effect indicators. This is not an inconsequential decision since causal and effect indicators have different properties and their incorrect classification can bias estimates from a model (Bollen and Lennox 1991). So a researcher needs to decide whether indicators are causal or effect indicators or some mixture of types.

The primary method for making this determination is performing mental experiments where the researcher imagines whether shifts in an indicator are likely to result in shifts in a latent variable as would occur with causal indicators or vice versa as would be the case for effect indicators. For instances, increases in time spent watching violent television shows increases the overall exposure to media violence. Increased overall exposure to media violence does not necessarily indicate increased watching of violent television. In some cases empirical tests can be brought to bear to test for causal versus effect indicators (Bollen 1989, Bollen and Ting 2000). A complication in dealing with causal and effect indicators is that the same set of indicators could be causal indicators with respect to one latent variable and effect indicators with respect to a different latent variable. As an illustration, the watching of violent television shows would be a causal indicator with respect to the latent variable of overall exposure to media violence, but the same television-viewing variable would be an effect indicator if the latent variable were propensity to seek violent entertainment.

The second characteristic of measurement models is whether there are multiple indicators of a latent variable or just one. Multiple indicators are generally preferable to single measures. In the case of effect indicators, multiple indicators are often essential in 'identifying' (permitting unique estimates of parameters in) the measurement model, are helpful in testing the number of latent variables underlying the measures, and enable assessments of the quality of the indicators. Multiple causal indicators can lead to a fuller description of the latent variable's main components. For instance, using education, occupational prestige, and income as causal indicators of SES provides a more comprehensive picture of SES than would income alone. In addition, leaving out causal indicators usually will bias the estimates of the effects of the remaining causal indicators on SES. So for both types of indicators it is better to have more rather than less.

The third and fourth distinctions from above are whether the latent variables and indicators are continuous or not. Continuous refers to whether the latent or indicator variable is best thought of as approximating a variable that has infinite gradations of magnitude or not. Indicators at best can approximate continuous variables since the limitation of measuring instruments will not permit infinite gradations. However, indicators approximate continuity in measurement to varying degrees. Population size, number of crimes, or counts of correct answers on a test are strictly speaking noncontinuous since the measures have interval units. However, in many situations, little harm is done in treating such indicators as if they were continuous. Alternatively, if we use a three-point scale to measure an attitude, it is difficult to consider the indicator continuous even if the corresponding latent variable is continuous.

One source of confusion when considering whether a latent variable is continuous or not is that researchers sometimes confound the indicators' properties with those of the latent variable. Consider the situation of a survey question on abortion that asks whether the respondent supports or opposes a woman's right to an abortion. Though the response categories for the question might permit only two answers, this does not mean that the abortion attitude *latent variable* is a dichotomy. It seems far more likely that the latent attitude is continuous such that not everyone who agrees to the right to abortion has the same degree of support. Some might be just above a threshold on the continuous abortion attitude and this leads them to choose the support category whereas others might be way above the threshold when they give their support response. In this and related situations, the latent variable is continuous even though the design of the measure forces the responses into categories. The researcher needs to consider whether it makes sense to consider the latent variable as continuous or categorical without confounding the nature of the indicator with that of the latent variable.

Theoretically, it is possible to have any combination such as a continuous latent variable and noncontinuous indicators or a noncontinuous latent variable and noncontinuous indicators. In practice some combinations are far more frequent than others. Consider effect indicators first. Treating both the latent variable and its effect indicators as continuous is particularly common. A typical equation for a measurement model for this case is

$$y_i = \alpha_i + \lambda_{i1}\eta_1 + \varepsilon_i \tag{1}$$

where y_i is the ith effect indicator that is approximately continuous and the right hand side of the equation includes an intercept (α_i), a factor loading coefficient (λ_{i1}) that gives the expected difference in the indicator for a one unit difference in the continuous latent variable (η_{1i}), and the continuous disturbance (ε_i). The disturbance has a mean of zero and is uncorrelated with the latent variable. This model captures many factor analysis (Mulaik 1972) and structural equation model formulations (Bollen 1998), though factor

analyses often deviate the observed variables from their mean so as to eliminate the intercept term.

An increasingly popular but less common formulation is to have a continuous latent variable with one or more noncontinuous indicators. With two exceptions, the same equation captures this situation. One is that a y_i^* replaces y_i where the y_i^* is a *latent*continuous indicator of the latent variable. The second difference is that there is an auxiliary measurement model that links the continuous y_i^* latent indicator to the noncontinuous observed indicator, y_i. For instance, if y_i is a dichotomous 0,1 variable the auxiliary equation might represent it as

$$y_i = \begin{Bmatrix} 1 \ if \ y_i^* \geqslant 0 \\ 0 \ if \ y_i^* < 0 \end{Bmatrix}$$

An exact nonlinear function would link the continuous latent indicator to the observed noncontinuous effect indicator. This is easily generalized to ordinal and censored effect indicators. This formulation is common in structural equation models (e.g., Muthen 1984). Item Response Theory (IRT) provides a related approach to having continuous latent variables and noncontinuous indicators (Lord 1980, Hambleton and Swaminathan 1985).

Less frequent are instances where both the latent variable and its effect indicators are categorical. For instance, we might have several informants who report whether the head of the household is receiving welfare or not. The latent variable is welfare status with only two categories: receiving or not. The informant indicators have the same two categories though each informant might make a classification error. Latent structure or latent class analysis (Lazarsfeld and Henry 1968) provides a means to handle this situation. Other combinations such as a categorical latent variable in combination with continuous effect indicators or a mixture of continuous and noncontinuous effect indicators are also possible, though they are rare in practice.

Causal indicators are far less studied than effect indicators. However, incorporation of continuous and noncontinuous causal indicators in a model with a continuous latent variable is possible provided the latent variable has sufficient outcomes to permit model identification (e.g., Joreskog and Goldberger 1975). Identification is too complicated to cover here, however, suffice it to say that continuous or categorical causal indicators do not create special difficulties. Perhaps the least studied case is that of having causal indicators (continuous or not) combined with a categorical latent variable.

A final consideration is whether more than one latent variable influences an effect indicator. It could be that two indicators derived from a questionnaire are from similarly worded questions such that the wording is a method latent variable that influences both items in addition to the substantive variable that they tap. Or indicators deriving from the same judge might share a common bias beyond their common dependence on the same latent substantive variable. It also is possible that two or more substantive latent variables help to determine the values of an indicator. For instance, a measure of the democracy of a political system might measure both democracy and the stability of the system. An indicator of depression might reflect both negative affect and sleep disorders. In these situations, the measurement model should allow two or more latent variables to affect the indicator. Generally researchers try to design indicators so that they primarily represent a single latent variable, but this is not always feasible.

3. *Properties of Indicators*

Properties of the indicators are useful to both current and future researchers who plan to use them. Among the two most important properties are the validity and the reliability of the indicators. Validity and reliability are properties that have received their greatest attention in the case of measurement models with continuous latent variables and approximately continuous effect indicators. Indicator *validity* concerns whether the indicator really measures the latent variable it is supposed to measure. Construct validity, criterion validity, and content validity are types of validity that researchers sometimes examine. Construct validity, for instance, assesses whether the indicator is associated with other constructs that it is supposed to relate to and not associated with those that it should not. Criterion validity compares the indicator to some standard variable that it should be associated with if it is valid. Content validity examines whether the indicators are capturing the concept for which the latent variable stands. See Nunnally and Bernstein (1994) for further discussion.

Reliability focuses on the consistency or 'stability' of an indicator in its ability to capture the latent variable. It is distinct from validity in that you can have a reliable indicator that does not really measure the latent variable. A general definition of the reliability of an indicator is the variance of the 'true' (latent variable) variance divided by the total indicator variance. A variety of statistics to estimate reliability exist. One of the most popular to measure the reliability of several combined effect indicators is Cronbach's (1951) alpha. It also makes a number of assumptions that might be difficult to satisfy in practice. Alternative measures of reliability built from less restrictive assumptions also are available (Bollen 1989).

IRT assumes a continuous latent trait and a categorical effect indicator, usually dichotomous or ordinal. IRT focuses on other properties of categorical items or indicators such as item discrimination and

item difficulty (Hambleton and Swaminathan 1985). IRT is similar to the traditional treatments of reliability and validity in that it too focuses on effect indicators. Latent class or latent structure analysis (Lazarsfeld and Henry 1968) also deals with effect indicators. It applies when we have latent categorical variables with categorical indicators. The degree of classification error of the observed categorical variables provides information on the accuracy of the indicator. The combination of a latent categorical variable with continuous effect indicators are less extensively developed than are the cases of continuous latent variables with continuous or categorical effect indicators.

The measurement properties of causal indicators are less discussed. An important point is that use of the causal indicator assumes that it is the causal indicator that directly influences the latent variable. If the causal indicator itself contains measurement error, then this needs to be part of the measurement model.

4. Indicator Development and Selection

A salient component of indicator methodology is the development or the selection of indicators. The most critical and most neglected aspect of this task is providing a clear *theoretical definition* of the concept that a researcher seeks to measure. The theoretical definition assigns the meaning to the concept. It should explain in as simple of terms as possible what it is that is to be tracked by the indicators. A clear theoretical definition should identify the number of distinct aspects or *dimensions* of the concept. Each dimension requires a separate latent variable. Furthermore the theoretical definition should clarify whether the latent variable is continuous or not. Each latent variable is ideally measured with several indicators.

Once a clear theoretical definition is in hand, the researcher collecting original data can design indicators that correspond to this definition. If only secondary data are available, the theoretical definition provides guidance in the selection of indicators. With the indicators selected to represent each latent variable, the next step is to decide whether each indicator is a causal or effect indicator. That is, an investigator needs to decide whether the indicator influences or is influenced by the latent variable to which it corresponds.

Next, the researcher should construct a model summarizing the relation between the indicators and the latent variable(s). The model should take account of the continuous or noncontinuous nature of both the latent variable(s) and the indicators. Once the measurement model is constructed, a researcher can then explore ways of estimating it. If, for instance, both the latent variables and effect indicators are essentially continuous, then confirmatory factor analysis applies. After estimation the researcher can assess the overall fit of the model and the properties of the indicators. This check might reveal that modifications to the model are required. It might be the case that some indicators prove faulty or the overall fit of the model might be poor, raising questions about the original specification. The analyst might then implement minor or major changes ranging from the introduction of additional parameters to the model to the formulation of quite a different model. The researcher would then estimate the new model and again assess its adequacy.

5. Summary

Indicators and related terms give recognition to the gap that exists between social science concepts and measures. During the last half of the twentieth century a variety of techniques and statistical models have been developed to enable researchers to study and to take account of the measurement error in our indicators. We also have the means to estimate relations between different latent variables while controlling for measurement error. Though the measurement models cover many situations, we still require further development of techniques to handle some combinations of categorical and continuous latent variables and indicators. However, a far more serious problem is the prevalent practice by social scientists of ignoring measurement error in their indicators or ignoring altogether latent variables that are difficult to measure.

See also: Classical (Psychometric) Test Theory; Computerized Test Construction; Data Collection: Interviewing; Dimensionality of Tests: Methodology; Indicator: Methodology; Instrumental Variables in Statistics and Econometrics; Measurement, Representational Theory of; Measurement Theory: History and Philosophy; Mediating Variable; Psychometrics; Reliability: Measurement; Test Bias

Bibliography

Aigner D J 1974 MSE dominance of least squares with errors of observation. *Journal of Econometrics* **2**: 365–72

Blalock Jr H M 1964 *Causal Inferences in Nonexperimental Research*. University of North Carolina Press, Chapel Hill, NC

Bollen K A 1989 *Structural Equations with Latent Variables*. Wiley, New York

Bollen K A 1998 Structural equation models. In: Armitage P, Colton T (eds.) *Encyclopedia of Biostatistics*. Wiley, Chichester, UK

Bollen K A, Kwok-fai Ting 2000 A tetrad test for causal indicators. *Psychological Methods* **5**: 3–22

Bollen K A, Lennox R 1991 Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin* **110**: 305–14

Cronbach L J 1951 Coefficient Alpha and the internal structure of tests. *Psychometrika* **16**: 297–334

Hambleton R K, Swaminathan H 1985 *Item Response Theory: Principles and Application*. Kluwer-Nijhoff Publishers, Boston
Joreskog K G, Goldberger A 1975 Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* **70**: 631–39
Lazarsfeld P F, Henry N W 1968 *Latent Structure Analysis*. Houghton Mifflin, New York
Lord F M 1980 *Applications of Item Response to Theory to Practical Testing Problems*. Erlbaum, Hillsdale, NJ
McCallum B T 1972 Relative asymptotic bias from errors of omission and measurement. *Econometrica* **40**: 757–58
Mulaik S 1972 *The Foundations of Factor Analysis*. McGraw-Hill, New York
Nunnally J C, Bernstein I H 1994 *Psychometric Theory*, 3rd edn. McGraw-Hill, New York
Wickens M R 1972 Use of proxy variables. *Econometrica* **40**: 759–61

K. A. Bollen

Indigenous Conceptions of Aging

The phenomenon of aging is encountered by every individual in every society. Infancy, childhood and youth are periods of development of an individual, with one reaching one's maximum capacity during middle age. As the last stage of human life, old age acquires special significance.

1. *What is Aging?*

It is important to understand aging from different perspectives, e.g., biological, psychological, and social aspects. According to Birren and Renner (1997, p. 4), behavioral scientists define aging as the regular changes that occur in mature, genetically representative organisms living under representative environmental conditions as they advance in chronological age. Further they explain three aspects of human age and aging. The biological age of an individual can be defined as an estimate of the individual's present position with respect to his potential life span. Psychological age refers to the adaptive capacities of individuals, that is, how well they can adapt to changing environmental demands in comparison with average demands. Social age refers to the rules and social habits of an individual with respect to other members of a society.

2. *Markers of Old Age*

There is no single criterion for defining old age. Though chronological age is not a good criterion to identify a person as 'old' or 'aged,' it is the one most commonly used. In Western countries, 65 years is considered as the cutting point between middle age and old age, in developing countries it is between 55 to 60 years. The basis for this is the average life span of an individual, which is linked with the retirement age. Apart from chronological age, there are other indicators/markers of old age. Biologically, menopause is considered the beginning of old age in the case of women. In Indian society, like many other traditional societies, marriage of one's own children and becoming a grandparent herald the beginning of gradual withdrawal from active occupational and family life and turning towards spiritual growth.

3. *Societal Views on Aging*

The concept of aging varies from society to society and has undergone modifications over the passage of time. In many societies, aging is determined by existing cultural norms with regard to progressive stages of family life of an individual. The ancient Chinese scholars delineated seven phases in a man's life, while Pythagoras in the sixth century BC compared human life to seasons. The Hindu concept of Ashramdharm divided life into four stages, namely, Brahmacharyashram (apprenticeship), Gruhasthashram (householder), Vanprasthashram (withdrawal), and Sanyasthashram (renunciation).

Kakar (1982) compares Hindu concept with Erikson's theory of eight stages of man's life. In his view, the individual is not viewed in isolation, but in interaction with the sequence of generations and in shared mutuality. Referring to the last two stages, which are of importance to the present theme, he points out that care and wisdom are virtues in both these concepts. The difference between these two theories is their origin. The Hindu view is a religious or ideal one, while Erikson's approach is clinical and developmental.

In African society as in many other traditional societies, old age is regarded as sacred or pious since the older person is nearer to the state of ancestors than the younger generations. Old age is considered to be dignified, honorable, and respectful. To neglect one's parents is an unpardonable act of ingratitude. There is a saying that means, 'when someone looked after you to grow your teeth, you should look after that person when they lose teeth.'

Old age is depicted negatively in terms of physical decline and decrepitude. Older persons are often characterized or stereotyped as ill, tired, mentally slow, self-pitying, unhappy, and unproductive. They are often depressed by feelings of loneliness and alienation. The negative concept of old age derives from the relationship of individual and society—the absence of a role, the isolation from the significant social life, the marginality in family relationships, and the lack of commitment, partly induced, partly forced on old people.

3.1 Myths about Aging

The attitudes towards older persons are often based on myths rather than reality. Among these myths, we find: the basic needs of individuals change as they age, and most old people are pretty much alike. The older people are not viewed as separate entities with their own needs, past experiences, circumstances, and health status. In a study of women from a matriarchal society in Khasis, it was found by Chhagte (1992) that widowhood was perceived as a means for developing one's self to earn dignity and respect from the society. The aging women played important roles in decision making and provided emotional support.

4. Images of Aging

Many studies have indicated that chronological age and physical appearance are not important indicators of aging according to older persons. Positive images of aging across culture depict older persons as knowledgeable, wise, experienced guides, and caring, whereas negative images depict older persons as insecure, forgetful, interfering, and dependent. How the society or individual views aging depends on many factors.

5. Factors Affecting Conceptions of Aging

5.1 Demography of Aging

The aging of a population is an inevitable consequence of demographic transition and can create an imbalance in the age structure over a period of time. The world's population continues to age with increasing rapidity. According to the United Nations Populations Projections (UN 1991), the world had an aged population (60 and above) of around 488.8 million in 1990, which will increase to 1,205.3 million by 2025.

On a global level, the increase in the proportion of the elderly started only after 1956 from a level of around or below 8.1 percent and is expected to reach around 14 percent by 2025. Until recently, the problem of aging was viewed as a problem of developed countries. However, due to remarkable reductions in fertility levels and increasing life expectancy in developing countries, it is believed that the future demographic transition will proceed more rapidly in the developing regions than it did in the developed ones and will involve large absolute numbers. In the year 2025, there is a projection of 1.2 billion old people in the world, 71 percent of whom are likely to be living in developing regions. More than half of the world's old people by 2025 will be just in Asia (UN 1991).

5.2 Industrialization and Family Structure

Several studies have suggested that the effect of urbanization, industrialization, and educational background may influence the attitude towards the elderly. According to many sociologists, the higher position enjoyed by the elderly was due to the multi-generation family system and the common land holdings by the family. The older persons owned the land and managed the family affairs by virtue of their age and by holding the economic power. However, the breakdown of the joint family system due to urbanization and migration and the partition of common land holdings and property, has depleted the economic power in the hands of the elderly. It is also contended that the status of the aged declined with technological and economic development. Status in industrial societies is based on individuals' acquired formal skills and education. As the industrial society is oriented towards the individual rather than the group, the aged are bound to be rendered useless, generating a feeling of meaninglessness in them. (Mahajan 1992).

5.3 Residential Patterns

The shift from joint to extended and then to nuclear family is a product of industrialization and urbanization. This has altered the residential patterns of older people. Due to the migration of younger generations, older persons tend to live as a couple or as a single person, even in developing societies. Some villages in Asia have predominantly older people living on their own. This has a direct bearing on the familial roles and responsibilities.

5.4 Availability of Roles

By and large, society determines the status and roles assigned to an older person, which makes a person either a contributing member or a dependent member. What alternative roles are available to the older men and women, if the traditional roles become redundant?

5.5 Financial Dependence/Available Resources

In many societies, investments in children's education, career building, and marriage are heavily valued. The parents assume that children, especially sons, will provide for them during their old age. Hence, there are no savings. Due to materialistic attitudes, children may not feel obliged to reciprocate. The result is inadequate financial resources for maintaining one's lifestyle.

5.6 Changing Values

A shift is observed from collectivism to individualism as a by-product of materialism. The economic power has shifted from older generation to middle generation. The need for privacy has increased on part of the younger generations. Certain basic human values such as honesty and freedom are not defined and interpreted by young and old in the same way.

6. Research Priorities

Considering aging as an ongoing process, which has implications for various disciplines, there is a need for longitudinal interdisciplinary research which focuses on old age as one of the important stages of human life. Since aging has become a global issue, many research projects have been undertaken in this field. It would be worth consolidating cross-cultural research to enhance the understanding of this field. To enhance the quality of life of older persons, it is important to review the plans and implementations of global, regional and national policies.

See also: Age: Anthropological Aspects; Age, Sociology of; Aging and Education; Aging and Health in Old Age; Aging, Theories of; Cognitive Aging; Differential Aging; Ecology of Aging; Indigenous Knowledge and Technology; Indigenous View on Human Development: West African Perspective; Life Course in History; Old Age and Centenarians; Population Aging: Economic and Social Consequences; Retirement and Health; Social Cognition and Aging; Stereotypes, Social Psychology of; Wisdom, Psychology of

Bibliography

Birren J E, Renner V J 1977 Research on the psychology of aging: principles and experimentation. In: Birren J E, Schaire K W (eds.) *Handbook of the Psychology of Aging*. Van Nostrand Reinhold, New York

Chhagte Z 1992 *Self-others Perception of Aging Women in Meghalaya*. Masters thesis, M S University of Baroda, India

Kakar S 1982 Setting the stage: the traditional Hindu view and the psychology of Erik H. Erikson. In: Kakar S (ed.) *Identity and Adulthood*. Oxford University Press, New Delhi

Krishnan P, Madadevan K (eds.) 1992 *The Elderly Population in Developed and Developing World: Policies, Problems, and Perspectives*. B R Publishings, New Delhi

Vatuk S 1980 The aging women in India: self-perception and changing roles. In: Desouza A (ed.) *Women in Contemporary India and South Asia*. Manohar, New Delhi

P. Dave

Indigenous Knowledge and Technology

Social science discussions revolving around the term indigenous technical knowledge have a relatively recent pedigree. This is not to say that research, especially in geography, anthropology, and archaeology, was not interested in documenting peasant resource management technologies—indeed there is a long tradition of such work, above all in Third World settings (Wilken 1987; e.g., see *Ecology, Cultural*; *Agricultural Change Theory*). However, the popularization of 'indigenous technical knowledge' as a term, and as a point of reference in debates about technology, is a feature of the last three decades of the twentieth century, and owes much to discussions in international agricultural and rural development. Indeed it derives from a concern to influence development programs in order that they become more cognizant of the many values of indigenous technologies.

Even though some colonial officials and writers recognized the qualities and intricacies of small-holder farming systems, much colonial policy ultimately aimed to promote commercial cash crop farming, using technologies and practices modeled on principles of European farming and promoting quite profound changes in local agriculture. This belief in the importance of 'transforming traditional agriculture' (Schultz 1964) was subsequently enshrined in the period following the Second World War and particularly following the independence of former colonies. This was—and indeed remains—the era of development in which international (and national) organizations have aimed to foster the modernization of Third World societies and economies through a range of external interventions. Until recently one of the most important of these interventions was rural development—and an important component of rural development strategy was the modernization of agriculture through the introduction of packages of new 'improved' technologies intended to increase agricultural yields. The packages incorporated innovations such as high yielding and other improved crop varieties, agrochemical fertilizer, pesticide and herbicide inputs, irrigation (in some cases), etc. (e.g., see *Green Revolution*).

A range of institutions were created to foster such agricultural development, but internationally the most important have been the International Agricultural Research Centers. These centers, each with a particular focus (maize and wheat; potato; rice; livestock, etc.) are located in different parts of the world, but have regional and generally global mandates to support the improvement of production systems for their particular specialty commodities. These centers are supported financially by government, multilateral, and foundation funds. They in turn collaborate with national agricultural research and technology transfer programs. These programs—though often quite weak in practical terms—are in theory the axial national institution for the modernization and enhancement of agriculture. They similarly work in generating and adapting new crop technologies (many received from the international centers) which are then released to farmers through a range of mechanisms.

While in the early years this commitment to the modernization of agricultural technology brought some notable increases in yields and output, some observers began to argue that such approaches tended

to benefit wealthier rather than poorer farmers, often had deleterious environmental impacts, and were inappropriate for the ecological and socioeconomic conditions under which poorer farmers produced. These critical observations inspired a series of studies which concluded that this was the wrong way for external interventions to attempt to improve peasant agriculture. Instead, it was argued, rural people already possessed a wealth of indigenous technical knowledge—i.e., a knowledge of the qualities of and relationships among crops, soils, climate, pest control, water management, etc. According to such critiques, development ought to begin by respecting and understanding this knowledge, and then find ways of improving it 'from within.' This was a call for a bottom-up form of agricultural (and rural) development that would build upon the resources that farmers already had. Indeed, for this reason much of this work has been characterized, both by its authors as well as by critics, as 'populist' in intent, laying more faith in the capacities of the popular sectors than in those of modern science and bureaucracy.

This critique of the so-called 'top-down' development model has since inspired a large body of research into different aspects of indigenous technologies in many locations. In many regards this work has argued that people's natural resource management practices are generally adapted to local circumstances. In this sense, work on indigenous technology is part of a far longer tradition of work in cultural ecology and human ecology interested in the ways in which local societies adapt to their environments as part of a process of transforming them (e.g., see *Ecology, Cultural*). It is also related to work on agricultural intensification and decision making that aims to understand the conditions under which resource users change their practices in order to intensify (or in some cases disintensify) production (e.g., see *Agricultural Change Theory*). These different traditions share an interest in the ways in which technology mediates the relationship between society, population, economy, and the environment. This work has assembled an important body of knowledge on indigenous agricultural practice and change, some more encyclopedic, some more analytical (of the many examples, see for instance, Wilken 1987, Denevan and Padoch 1987).

A principle difference between these different approaches though is the extent to which research directs itself to issues of policy. While cultural and human ecological research has tended to keep its distance from development programs, indigenous technology research has been explicitly concerned with influencing policy debates and rural development programs. It has aimed to change the ways in which crop research is conducted, to press for far greater participation of people in technology development and agricultural policy formulation more generally, and to craft a different vision for the future of agriculture in Third (and increasingly First) World settings.

This concern for the policies and approaches of development reflects the origins of these discussions and also the people who have participated in them. Indeed the debate on indigenous technology owes at least as much to applied social (and biological) scientists working outside academia in the international agricultural research centers and other development institutions as it does to the work of academics (though several important studies on indigenous resource management by academic social scientists with applied concerns in the 1950s and 1960s also helped pave the way for this work). Though often hired by the international centers to address why it was that crop breeding and other forms of technology development were not having the positive effects that had been hoped for them following early successes, these social scientists often became critical commentators on the larger model of agricultural modernization being fostered through the work of the international centers.

Of the many examples of work done by social scientists in the international centers, Robert Rhoades' book, *Breaking New Ground* (1984), written from the author's vantage point as an anthropologist at the International Potato Center in Peru, constituted an important early statement. It addressed the limitations of the modernization model, and argued for the importance of better understanding indigenous technical knowledge in order to build on it, and to foster more constructive engagements between indigenous expertise and scientists' expertise through processes of participatory research and technology development. The book suggested that this had implications both for crop research as well as for the conduct of anthropology, and academic anthropologist Paul Sillitoe (1998) has recently reiterated much of this argument for a change in the way much anthropology is conducted.

Early, more academic collections (Brokensha et al. 1980, IDS 1979) were also important in arguing the case for indigenous technical knowledge, but it was the publication of Paul Richards' book *Indigenous Agricultural Revolution* in 1985 that marked a real turning point in the debate. Although it was an academic study involving careful, detailed cultural ecological research, this book spoke deliberately to the agricultural development community. On the basis of ethnographic, agronomic, and archival research, Richards demonstrated the ways in which farmers exploit a range of ecological opportunities, develop production systems, and actively manage varietal diversity in crops. In some sense, the book argued that the processes of active crop selection, creation of new niches for production, etc., emphasized so much by the literature on the origins of agriculture and crop domestication, applied equally to contemporary peasant agriculture, and that these everyday, grassroots processes could and should once again constitute the basis for another agricultural revolution on the basis of so-called

'peoples' science.' The breadth and academic rigor of the argument gave it particular weight.

While Richards' argued this case as a social scientist, a group of biological scientists have been similarly important in promoting the case for indigenous technology on the basis of its agro-ecological merits. These authors—working in a range of ecological contexts—have assembled more scientifically validated bodies of evidence to demonstrate that many indigenous technologies are effective in controlling pests, managing (and creating) soil fertility, managing water efficiency, etc. An important early synthetic statement of this position was Miguel Altieri's *Agroecology: The Science of Alternative Agriculture* (1983). Altieri's work, as a biological scientist, also shows what can be achieved when the researcher becomes more actively involved in initiatives to link indigenous technology to wider agricultural concerns. In particular he and colleagues have pointed to ways in which indigenous modes of managing biodiversity might become the basis for far larger programs of biodiversity maintenance—programs that differ substantially from the formal centralized germplasm bank approach.

Of the many other initiatives of note, the approach of the recently deceased Mike Warren merits comment. A researcher activist like Altieri (and perhaps like many committed to this topic), Warren dedicated a large part of his professional career to advocating for indigenous technical knowledge (see Brokensha et al. 1980 and Warren et al. 1995). He created programs, networks, and research initiatives documenting technical practices as a means of conserving them, validating them, and indeed arguing forcefully for the role that they might play in rural development.

The effect of this work has indeed been to change some—if by no means all—thinking underlying crop breeding and the development of new technologies in the agricultural research community. It has, however, also been subject to a series of critical appraisals that have themselves opened up other avenues of inquiry around the topic of indigenous technology. While the theoretical basis of indigenous technology research is generally implicit and often somewhat unclear, it is actor (rather than structure) focused. Others however have argued that there are serious structural constraints on agriculture deriving from the political economy and that to detach actor from structure is conceptually inadmissible and also leads to inappropriate policy recommendations (Watts 1983 is an important early statement of this position). Thus, critics suggest, to explain agricultural practice, productivity, and income in terms of choice, innovation, and knowledge without referring to this political economy is to give a deficient explanation. Any resulting policy recommendations will therefore fail because they are based on bad explanation: both actor and structure need to be part of any explanations of indigenous agricultural dynamics. Critics have also argued that the focus on 'farmer' knowledge is too generic: it does not elucidate sufficiently who knows what within local populations, nor the ways in which agricultural knowledge is revealed and hidden depending on the social relationship at stake (see some of the essays in Scoones and Thompson 1994). According to the analytical frameworks used in these critiques, knowledge cannot be understood nor talked about except in relation to the social position of the knower. Explanations which allow this separation will lead to policies that favor some types of knowledge (and therefore some knowers) and not others. These critiques begin to link discussions of indigenous technical knowledge to wider debates on the sociology of knowledge.

A third set of criticisms revolve around the very term indigenous technical knowledge. It is often unclear what makes this knowledge 'indigenous.' Many of the farmers ostensibly having indigenous knowledge are not in fact indigenous to a location: many have moved in from elsewhere, and been part of wider processes of cultural contact and hybridization. Thus, it is suggested, to use the term 'indigenous' is to make exaggerated claims to the authenticity and legitimacy of this knowledge, and that what is being discussed is instead 'local' knowledge—a more modest and more accurate label, although also oversimplified because much knowledge and practice is syncretistic, a product of cultural diffusions, interactions, and integrations. Such an interpretation may not lead to conclusions that such technical knowledge is any the less adapted to local conditions, but it does shift how such knowledge should be understood. Conceptually, such an interpretation re-opens the case for external introduction of technology, if it is in fact the case that local technology is already a product of the local and nonlocally.

Similarly, it is not clear why this is 'technical' knowledge. Technology and agricultural practices convey various meanings, which may be interpreted in different ways according to the age, gender, ethnicity, origin, and other characteristics of the interpreter. Thus, to discuss such knowledge only in technical terms is to give a partial interpretation. To understand the possibilities of recovering and building on this knowledge, it is essential—so such a reflection goes—to appreciate the fuller significance of this knowledge in a local context. A related reflection on the 'technical' dimension is that which—though sympathetic to the concerns and motivations of those who have worked on indigenous technology—suggests that a focus on technology is insufficient. Indeed many researchers who began with an interest in indigenous technologies have broadened their concerns to understand the relationships between technology, identity, local organization, and political economy as part of an effort to understand the conditions under which a locally grounded form of development might be possible (Bebbington 1993). Though consistent with

the initial motivation of those concerned with indigenous technical knowledge—i.e., to rethink rural development strategies and legitimate the ordinary practices of resource constrained farmers—this project works with a broader canvas. It has the aim of understanding what an alternative form of local development might look like, how it might occur, and, *inter alia*, what its technological basis might be.

Finally, it ought be noted that while work on indigenous technical knowledge has done much to improve understanding of farmer resource management and to change approaches to rural and agricultural development, small farm agriculture is still threatened in many ways. Contemporary threats, however, appear to come less from the modernizing ideologies of development programs that students of indigenous technical knowledge criticized; rather they derive from the evolving nature of a global food system in which control of seed stocks and supply chains is increasingly vested in a relatively few, powerful transnational corporations.

See also: Agricultural Change Theory; Ecology, Cultural; Ecology, Political; Environment and Development; Ethnobiology; Green Revolution; Indigenous Knowledge: Science and Technology Studies

Bibliography

Altieri M A 1983 *Agroecology: The Science of Alternative Agriculture*. Westview Press, Boulder, CO

Bebbington A 1993 Modernization from below. An alternative indigenous development? *Economic Geography* **69**(3): 274–92

Brokensha D, Warren D M, Werner O (eds.) 1980 *Indigenous Knowledge Systems and Development*. University Press of America, Washington, DC

Chambers R, Pacey A, Thrupp L A (eds.) 1989 *Farmer First. Farmer Innovation and Agricultural Research*. Intermediate Technology Publication, London

Denevan W M, Padoch C (eds.) 1987 *Swidden-Fallow Agroforestry in the Peruvian Amazon*. New York Botanical Garden, Bronx, NY

IDS 1979 Rural development: Whose knowledge counts? *IDS Bulletin* **10**(2): Special issue

Rhoades R E 1984 *Breaking New Ground: Agricultural Anthropology*. International Potato Center, Lima, Peru

Richards P 1985 *Indigenous Agricultural Revolution. Ecology and Food Production in West Africa*. Hutchinson, London

Schultz T W 1964 *Transforming Traditional Agriculture*. Yale University Press, New Haven, CT

Scoones I, Thompson J (eds.) 1994 *Beyond Farmer First. Rural Peoples' Knowledge, Agricultural Research and Extension Practice*. Intermediate Technology Publication, London

Sillitoe P 1998 The development of indigenous knowledge. A new applied anthropology. *Current Anthropology* **39**(2): 223–52

Warren D, Brokensha D, Slikkerveer L (eds.) 1995 *Indigenous Knowledge Systems: The Cultural Dimension of Development*. Intermediate Technology Publications, London

Watts M 1983 Populism and the politics of African land use. *African Studies Review* **26**: 73–83

Wilken G C 1987 *Good Farmers*. University of California Press, Berkeley, CA

A. Bebbington

Indigenous Knowledge: Science and Technology Studies

1. Convergence in Definitions

Science and technology studies (STS) and the study of indigenous knowledge are both interdisciplinary areas of investigation rather than singular disciplines. Their social and institutional origins were distinct, and early development in these areas occurred largely in isolation from each other. STS has disciplinary origins in history, sociology, policy, and philosophy, while the study of indigenous knowledge originated in anthropology and grew to overlap with agrarian studies and environmental history. Recent scholarship in both areas, however, suggests a convergence of their definitions of knowledge and their methodological assumptions.

'Indigenous knowledge' has long been associated with the terms 'local knowledge' or 'ethnoscience,' indicating knowledge systems that are specific to cultures or groups in particular historical or social contexts (see Richards et al. 1989). Adding the prefix 'indigenous' or 'ethno' to the terms 'knowledge' and 'science' seemed to mark the embeddedness and context dependence of indigenous truths, in implicit contrast to the context-free, singular truth of science. The distinct meanings and uses of the terms 'indigenous knowledge' and 'science' both depended, thus, on a dichotomy that lay at the heart of modern definitions of knowledge, separating universal, value-free, static truth from situated, value-laden, changing cultural beliefs. This dichotomy has been radically challenged by STS, which suggests that all knowledge, including scientific knowledge, is specific to its particular cultural context. All knowledge, then, might be considered 'indigenous.' The methods of study of the two fields should converge, according to this argument, since all knowledge should be studied with attention to its cultural and historical contexts. As a result, the distinct and separate foci of the two fields (science being conventionally studied in First World locations, and indigenous knowledge in the Third World) would lose the logic of their original delineation.

The status of the separation between 'indigenous' and 'universal' knowledge, the conditions of its construction, and the consequences of its erasure are issues of critical debate in STS. In a polemical and influential essay, *We Have Never Been Modern* (1993), Bruno

Latour argued that if 'modernity' signifies the policing of boundaries between nature and culture, belief and reason, humans and nonhumans, then anthropological study reveals that we have never in fact been modern, for our practices have always involved hybrids of these categories. Latour and others, revisiting Western knowledge from a distanced perspective, adopted and popularized ethnomethodological research. STS practitioners pictured the communities they studied as 'tribes' with characteristic beliefs and rituals. Western knowledge is, in this method of analysis, 'estranged' from its context, or 'othered' (Douglas and Wildavsky 1982). Conversely, non-Western knowledges are presumed no less 'rational' or no more 'context-dependent' than their Western counterparts (Watson-Verran and Turnbull 1995, Turnbull 2000).

Studying the cultural and political role of scientific knowledge in different societies, scholars note that nature and culture are produced by networks of contingent histories; we encounter them not in their 'pure' states, but as hybrids—'socionatures' (Jasanoff et al. 1995, p. 58) or 'natures-cultures' (Latour 1993, p. 104). Anthropological observers of indigenous knowledge have always been interested in the ways in which boundaries were (or failed to be) drawn between nature and culture, and in the political weight of certain constructions of nature (see, for example, Lévi-Strauss 1969). Recent work in disciplines such as peasant studies, cultural geography and anthropology, and development also reveals how ideas of nature have historically shaped and been shaped by social, political, and economic relationships (Peet and Watts 1996, Bryant and Bailey 1997). These disciplines have supported interlinked analyses of the changing representations of nature, the production of ideologies of progress and modernity, and the functional relationship between the domination of indigenous groups and the production of scientific knowledge about them.

The study of indigenous knowledge has increasingly shifted away from the province of ethnocentric ethnographers who tended to value it in opposition to science, assuming it to embody primitive expressions of folklore that defy rationality and mythologize social relations. Since the 1980s, anthropologists and interdisciplinary social scientists have viewed indigenous knowledge rather as coherent systems of belief and explanation that are homologous with scientific thought and that can often offer better accounts of local phenomena than those theorized by universalist science. This sort of investigation has prompted the reassessment of operative assumptions in medicine, pharmacology, resource management, and science education, as well as in nonscientific fields such as international law, policy, and development studies.

Although definitions of 'science' and 'indigenous knowledge,' and the modes of their investigation, exhibited a convergence over the last three decades of the twentieth century, studies of these fields remain separated, institutionally and, at times, ideologically. Anthropologists, in the wake of postcolonial critiques of the imperial construction of ethnographic power, have become increasingly self-reflexive and critical of their own representations of indigenous knowledge. Anthropological studies exposing ideological constructions of indigenous knowledge have become inseparable from efforts (in theoretical, pedagogical, and political realms) to alter power differentials in favor of marginalized groups. STS, on the other hand, has privileged studies of canonical European sciences such as physics, placing relatively less importance on the study of non-Western sciences and the ideological practices of science in the context of global political differentials.

In the political sphere, the study of indigenous knowledge has generated a host of policy implications, ranging from support for collective cultural rights and intellectual property rights for indigenous communities to the re-examination of political assumptions about identity, modernity, and global citizenship. The politics of the nineteenth and twentieth centuries, especially in the contexts of colonialism, nationalism, postcolonialism, and globalization, have influenced, and in turn been influenced by, scholarship and activism on the subject of indigenous knowledge. The study of indigenous knowledge has facilitated, and in turn been enhanced by, the granting of an influential international voice to indigenous rights groups and environmental nongovernmental organizations. These dynamics continue to play themselves out in such local and global issues as the politics of multiculturalism, the ownership of knowledge in the information economy, and the role of nationalism and group identity in a globalized world. These are large, interdisciplinary topics in themselves, but even as they grow wider in scope and diffuse in their disciplinary underpinnings, the issues raised by the study of indigenous knowledge continue to intersect with questions about the authority of science and its representations of nature, about the politics of science and technology as interpreted by international institutions and policymakers, and about the role of technology and development in the forging of a more equitable world. Creative work in these overlapping regions of STS and indigenous knowledge studies has yielded both theoretical insights and practical results.

2. *Theoretical Results: Indigenous Knowledge and Epistemology*

Epistemological concerns are necessarily central to the study of indigenous knowledge. Scholars of indigenous knowledge argue that if we take non-Western knowledges seriously, we are quickly led to abandon a naive realism with respect to scientific truth. That is, if we investigate indigenous knowledges for their inter-

nal coherence, social role, and explanatory power, rather than as instances of nonrational narratives, then we accumulate strong support for the argument that explanation, evidence, and truth are socially constructed and context-dependent. This suggests, one argument continues, that what we see in the world, within and among cultures, are competing rationalities, rather than a singular mode of rationality and reason (commonly held to be the province of the west, and its sciences) vs. nonrational or mystical modes of thinking that are antithetical to science (commonly held to be found outside the west). Such an argument requires a complex and situated definition of rationality rather than a simple one assessing syllogistic consistency or the efficient choice of technical means for particular instrumentalist ends (see Wilson 1970, Sen 2000).

Disagreement persists about how best to characterize indigenous rationalities with respect to Western science, as illustrated in a well-known debate between anthropologists Marshall Sahlins (1995) and Gananath Obeyesekere (1992). Sahlins, holding that different cultures possess different rationalities, suggested that anthropologists ought to cultivate a 'regard for cultural difference.' He called on anthropologists to respect the worldwide resurgence of particular cultures as 'distinct forms of life,' and to value their opposition to a dominant, utilitarian, Western modernity. In his own work, Sahlins suggested that Hawai'ian thought, for example, does not distinguish the practical from the mythical, nor the observable from the fictional, although these oppositions characterize European epistemologies. Gananath Obeyesekere challenged this view, maintaining that all cultures possess a 'practical rationality'—that is, a flexibility and creativity expressed through manipulating and interpreting the world.

Obeyesekere's interpretation of Hawai'ian history via the attribution of a 'practical rationality' to the Hawai'ians was, in Sahlin's view, an example of 'symbolic violence' to the culture, and amounted to a neo-colonialist model that reduced all cultural practices to Western logic and rationality. Obeyesekere contended that the cross-cultural interpretation of practical rationality, although complex and difficult, is 'necessary if one is to talk of the other culture in human terms.'

The categories of this theoretical debate recur in the realm of activism. For example, many activists in developing countries have found it useful to employ a relativist argument that rejects the subordination of indigenous knowledge to science, while retaining the postulate of incommensurable difference. That is, they value indigenous knowledge over scientific reason, arguing that narratives of re-enchantment are needed to restore the integrity of indigenous cultures, lost or rendered inauthentic through violent encounters with European Enlightenment science. This position sees an incommensurability between tradition (seen as authentically indigenous) and modernity (seen as ethnocentrically European). Although phrased in the vocabulary of anticolonial struggle, it is reminiscent of eighteenth- and nineteenth-century Euro-American romanticist and primitivist discourses in its desire to value a nature-centered spirituality over a technicist instrumentality. In both industrialized and developing countries, however, romantic and biocentric attempts to preserve the authenticity of indigenous knowledge have been critiqued from perspectives suspicious of epistemological relativism and biocentric romanticism. For example, Meera Nanda, supporting practical, rational 'people's science' movements in India, argued that the nostalgia for authentic indigeneity has oppressive overtones of nativism and cultural nationalism (Nanda 1997).

3. *Practical Results: Indigenous Knowledge and Politics*

The notion of indigenous knowledge holds practical allure for biotechnologists who believe that local knowledge of nature might hold the key to decades of future research into medicine and pharmaceuticals. It represents an ethical dilemma for the architects of globalization, heavily criticized for seeming to obliterate diversity in the interest of a homogeneous and efficient global marketplace. This section surveys the practical results of research into indigenous knowledge, and the political gains made by the movement to preserve indigenous knowledge.

The term 'indigenous people' has a somewhat diffuse reference, and can range from small peasants, linked to domestic and global markets, to tribes, isolated from other groups. In the last two decades of the twentieth century, there was a surge in support, from many different interest groups, for indigenous peoples' rights to their knowledge, resources, and cultural practices. Advocates for indigenous knowledge often use the term synonymously with 'eco-friendly' or environmentally correct knowledge.

Historical changes have often altered indigenous communities' relationship with their environment, or altered the environment itself, so that methods once sustainable might not be so in a changed context. Nor are all indigenous communities necessarily committed to preserving natural environments. For example, the Kissia of the Republic of Guinea, who use forest resources, tend to conserve forest growth, while other tribes in the same region, who are not dependent on trees for their survival, try to reduce forest growth wherever possible (Fairhead and Leach 1997).

Practical struggles facing indigenous communities have often involved rights of access to water, fuel, and other elements of their natural surroundings that are crucial to their survival. Activists have defended the rights of indigenous people to manage their natural surroundings independently from 'top-down' projects

that seek to impose universalist models of productivity and resource management. In these arguments, indigenous knowledge is seen as having both cultural and pragmatic significance. The state and the global economy, representing hegemonic modes of resource extraction and cultural homogenization, are often seen as antagonistic to the interests of indigenous communities.

Antagonism between the state and indigenous communities dates back to the colonial era, when the expansion of European territories was in large part an effort to control not only more markets and labor power, but also natural resources. Colonial environmental management often met with strong resistance from indigenous communities whose control over their environment was weakened or destroyed by colonial policies (Ranger 1989, Peet and Watts 1996). Soon after the mid-twentieth century's wave of decolonization, in the 'development decades' of the 1960s, postcolonial nations struggled to catch up with the productivity of the industrialized west. Postindependence development involved, in many countries, shifts in economic policy to maximize resource extraction for exports, large conversions from food crop to cash crop cultivation, and increasing industrialization. The implementation of these policies (advocated by large international development agencies such as the World Bank) often caused growing pollution problems, and the depletion of soil, water, and other natural resources. This most dramatically affected small farmers and indigenous groups, displacing them from their livelihood and threatening their modes of survival. Indigenous environmental movements, such as the antidam movement in India's Narmada valley, have often characterized the World Bank and the postcolonial nation-state as perpetuating colonial exploitation by subordinating local knowledge and experience to the rationalizing forces of science and economics (Rich 1995, Scott 1998).

The rapid growth, through the 1970s, of environmental ethics and economics did not always bring support for indigenous communities (see Harvey 1996). Neo-Malthusian and neo-Hobbesian ecological scholarship identified overpopulation, poverty, and weakly governed communities as responsible for much environmental degradation, and led to policies promoting the privatization of common property, or advocating a global Leviathan. Influential theorists such as Garrett Hardin and William Ophuls predicted that the combination of resource scarcity with social complexities would lead to an ecological crisis (Hardin 1968, Ophuls 1977). Many saw these looming crises as legitimating more authoritarian state-led solutions, giving primacy to private property and modern resource management techniques. Since most indigenous communities are not affluent, and are dependent on jointly governed common resources, these theories often pitted them against the prevailing expert wisdom.

Hardin's 'tragedy of the commons' thesis spurred detailed studies of indigenous environmental knowledge, increasingly critical of Hardin's assumptions. New scholarship on the commons, combining historical and political economic analyses of ecology, discovered empirical evidence that commons use did not inevitably lead to tragedy (Ostrom 1990, Carney 1996). Indigenous use of commons land was found to be more complex than simply opening access to all users, including irresponsible 'free riders.' Research in diverse communities found that successful commons use was invariably tied to complex, differentiated sets of communities with separate but overlapping rights and responsibilities, facilitating individual as well as group uses of lands and resources. In order to recognize and maintain these delicate social relations of resource use, both indigenous participants and outside administrators needed to be aware of the complexities of social stratification, and historical changes in social relationships among indigenous communities, and between them and the state. In many developing countries, however, colonial and postcolonial administrations altered the meanings of landholding and resource use traditions through successions of new legal definitions and administrative hierarchies, over time squeezing indigenous systems into models that privileged private over communitarian interests. Histories of disparate indigenous groups reveal a continual flux of local, regional, and global knowledges and practices, rendering unviable a nostalgia for a pure localism, or notions of pristine authenticity (McCay and Acheson 1987). Third World political ecology argues that support for rights of indigenous peoples should be inseparable from the development of more nuanced historical analyses of environmental and social relations.

Since the 1980s, the importance of indigenous knowledge has been widely acknowledged by influential international institutions. In 1982, recognizing that disparate indigenous movements all over the developing world were articulating concerns similar to each other, the United Nations established the Working Group on Indigenous Populations, annually bringing together participants from indigenous communities around the world.

Indigenous groups, working in solidarity with each other and developing an international alliance facilitated by the annual working group meetings in Geneva, prepared a joint statement for the 1992 Earth Summit at Rio, asserting indigenous rights as a fundamental requisite to the future of sustainable development. Recognising the growing importance of biotechnology and its links with traditional knowledge of medicinal plants, it stated: '[W]e demand guaranteed rights to our intellectual property, and control over the development and manipulation of this knowledge' (Article 44, Charter of the Indigenous–Tribal Peoples of the Tropical Forests, reprinted in International Alliance of Indigenous–Tribal Peoples of the

Tropical Forest and the International Working Group for Indigenous Affairs 1996, p. 9).

The 1993 UN Declaration on the Rights of Indigenous Peoples recognizes indigenous peoples' rights to the preservation of all forms of their knowledge, including cultural, spiritual, medical, scientific, and technological. It also recognizes the right of indigenous peoples to protect their intellectual property, and to the restitution of any such property that was taken from them in violation of their laws or customs. In 1998, the World Intellectual Property Organization established the Global Intellectual Property Issues Division, recognizing indigenous peoples as the 'most important' beneficiaries of intellectual property (Roulet 1999, p. 129).

The successes of indigenous peoples' movements in the late twentieth century were accompanied by a popularization, at both expert and lay levels, of the terms and categories of 'indigeneity.' These successes were linked, paradoxically, to a tendency critiqued by political ecologists. Critics pointed out that increased funding for the protection of indigenous groups often went along with romanticized descriptions of such groups. In a study of South Asian environments and ethnicities, Sumit Guha suggested that the 'poetic vagueness' in the definitions of indigenous groups feeds into a tendency to regard such groups as static remnants of original racial groups, supporting policies that seek to preserve them as museumized 'endangered species' (Guha 1999, pp. 4–5). Although the discourse of authentic indigeneity preserves historically inaccurate, nativist assumptions, many movements have used it strategically for political gains.

International recognition and support for indigenous knowledge grew dramatically through the 1990s. Nevertheless, as conflicts between indigenous rights and international trade grew rather than diminished, it remained to be seen how far courts would go to enforce 'soft law' elements such as international conventions to impose regulations on transnational corporations and states protecting indigenous communities. The future of indigenous knowledge in this age of globalization may well be worked out in policy debates and legal battles rather than in philosophical texts. Interpretations of the various assumptions in environmental law, however, will depend on the application of scholarship to issues such as the construction of scientific certainty, the historical contingency of systems of knowledge, and the social construction of nature and culture. In this way, the theory and practice of STS and indigenous knowledge may become more closely interlinked, and have significant consequences for the shape of the globalized world.

See also: Folk, Indigenous, and Customary Law; Historiography and Historical Thought: Indigenous Cultures in the Americas; Indigenous Knowledge and Technology; Knowledge, Anthropology of; Lex Mercatoria; Science and Technology, Anthropology of; Science and Technology Studies: Ethnomethodology

Bibliography

Bryant R, Bailey S 1997 *Third World Political Ecology*. Routledge, London

Carney J A 1996 Converting the wetlands, engendering the environment: the intersection of gender with agrarian change in Gambia. In: Peet R, Watts M (eds.) *Liberation Ecologies: Environment, Development, Social Movements*. Routledge, London

Douglas M, Wildavsky A 1982 *Risk and Culture. An Essay on the Selection of Technical and Environmental Dangers*. University of California Press, Berkeley, CA

Fairhead J, Leach M 1997 Culturing trees: socialized knowledge in the political ecology of Kissia and Kuranko forest islands of Guinea. In: Seeland K (ed.) *Nature is Culture: Indigenous Knowledge and Sociocultural Aspects of Trees and Forests in Non-European Cultures*. Intermediate Technology Publications, Southampton, UK, pp. 7–18

Guha S 1999 *Environment and Ethnicity in India, 1200–1991*. Cambridge University Press, Cambridge, UK

Hardin G 1968 The tragedy of the Commons. *Science* **162**: 1243–8

Harvey D 1996 *Justice, Nature, and the Geography of Difference*. Blackwell, Oxford, UK

International Alliance of Indigenous–Tribal Peoples of the Tropical Forest and the International Working Group for Indigenous Affairs 1996 *Indigenous Peoples, Forests, and Biodiversity: Indigenous Peoples and the Global Environmental Agenda*. IWGIA Document No. 82, Denmark

Jasanoff S, Markle G, Petersen P, Pinch T (eds.) 1995 *Handbook of Science and Technology Studies*. Sage Publishers, London

Latour B 1993 *We Have Never Been Modern* [trans. Porter C]. Harvard University Press, Boston

Lévi-Strauss C 1969 *The Raw and the Cooked* [trans. Weightman J, Weightman D]. Harper & Row, New York

McCay B, Acheson J (eds.) 1987 *The Question of the Commons: The Culture and Ecology of Communal Resources*. University of Arizona Press, Tucson, AZ

Mohanty S 1998 *Literary Theory and the Claims of History: Postmodernism, Objectivity, Multicultural Politics*. Oxford University Press, Delhi

Nanda M 1997 Against social de(con)struction of science: cautionary tales from the Third World. In: Meiksins Wood E, Bellamy Foster J (eds.) *In Defense of History: Marxism and the Postmodern Agenda*. Monthly Review Press, New York

Obeyesekere G 1992 *The Apotheosis of Captain Cook: European Mythmaking in the Pacific*. Princeton University Press, Princeton, NJ

Ophuls W 1977 *Ecology and the Politics of Scarcity: Prologue to a Political Theory of the Steady State*. W. H. Freeman, San Francisco

Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK

Peet R, Watts M (eds.) 1996 *Liberation Ecologies: Environment, Development, Social Movements*. Routledge, London

Ranger T 1989 Whose heritage? The case of the Matobo National Park. *Journal of Southern African Studies* **15**(2): 217–49

Rich B 1995 *Mortgaging the Earth: The World Bank, Environmental Impoverishment, & the Crisis of Development*. Beacon Press, Boston
Richards P L, Slikkerveer J, Phillips A O 1989 Indigenous Knowledge Systems for Agriculture and Rural Development: The CIKARD Inaugural Lectures. Studies in Technology and Social Change, No. 13, Iowa State University, Ames, IA
Roulet F 1999 Human Rights and Indigenous Peoples: A Handbook on the UN System. IWGIA Document No. 92, Copenhagen, Denmark
Sahlins M 1995 *How 'Natives' Think: About Captain Cook, for Example*. University of Chicago Press, Chicago
Scott J C 1998 *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT
Sen A 2000 East and west: the reach of reason. *New York Review of Books* **47**(12): 33–8
Turnbull D 2000 *Masons, Tricksters and Cartographers: Comparative Studies in the Sociology of Scientific and Indigenous Knowledge*. Harwood Academic Publishers, Amsterdam
Watson-Verran H, Turnbull D 1995 Science and other indigenous knowledge systems. In: Jasanoff S, Markle G, Petersen P, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage Publishers, London, pp. 115–39
Wilson B (ed.) 1970 *Rationality*. Harper and Row, Evanston, IL

K. S. Philip

Indigenous View on Human Development: West African Perspective

Development has four different meanings: (a) the act of developing or state of being developed, (b) a product or result of developing, (c) a significant event, and (d) a group of buildings, such as dwellings. This article focuses on the meanings in (a) and (b) to explore West African views on the human life course in the light of scientific psychology. The term 'indigenous' refers to that which is native or natural in an area or environment.

An indigenous viewpoint deals with what is salient and fosters understanding of a particular setting. Although most people cannot articulate an understanding of the world, their knowledge is organized generally into a worldview or outlook on the universe, and the place of human beings in it. It is social representation with which a given culture makes sense of human life. A set of social, cultural, and existential realities marks out an indigenous West African worldview, which differs remarkably from that informing scientific psychology.

1. *A Fragmented, Exclusive Science Inspired the Emergence of Indigenous Psychology*

Centuries before the rise of modern science and academic psychology, human cultures shared a folk psychology—a set of beliefs and practices about psychological issues in a particular culture. Folk developmental psychology is part of indigenous psychology derived from ethnotheories of development. It emerged from dissatisfaction with psychology's insensitivity to, and failure to connect with, the local context of rural and low class people and minority cultures around the world. For instance, a great deal of African knowledge exists in proverbs, folklore, and practices that are neither easily translatable into psychological jargon nor accessible to current research tools. In fact, many questions on human development in Africa are unanswered or, worse still, unasked, because the grids of scientific psychology were fashioned to suit and reveal primarily Western children or, more accurately, the (post)modern child. As a result, the field is contested terrain, not only because its tools have been derived from a tiny segment of European-driven cultures, but also because of the considerable diversity in how its mission manifests itself in theory, method, and practice. Diversity is not a problem, but the nature of the human condition *is*, and psychologists must come to terms with what the wide variety of childhood circumstances portend for the global image of human development.

Developmental psychology is a science of exclusion in that it purports universality but fails to attend to '95 percent of the world's children' (Zukow 1989). In fact, the extant developmental literature is mainly on the development of a segment of Western children. Other patterns of development exist, and deserve to be incorporated into the discipline's database. The need is to open up to the developmental paths of all people. Compelling logic for this necessity derives from the fact that between a universal humanity and specific individuality lies a huge cultural zone in every person created by a particular worldview, which inspires a specific course of life.

2. *West African Ontogeny: An Indigenous View on Human Development*

Indigenous psychology assumes that every culture conceives of human nature in its own terms, and organizes development according to that image. A West African social ontogeny draws on African life-journeys and conception of personhood in terms of its 'becoming.' It recognizes three phases of the life course: spiritual selfhood, social selfhood, and ancestral selfhood. While newborns are entering the world of the living, the dying are at the threshold of a spiritual world. This viewpoint extends the human life course to an afterlife. Developmental psychology has focused almost exclusively on social selfhood, the experiential self, itself divided into seven stages, the first and the last, respectively, linking spiritual and ancestral selfhoods. As in Erikson's psychosocial

theory, each stage of social ontogeny is marked by a distinct developmental task, conceptualized in terms of important transitions between patterns of social participation that define the culture's perceptions of the family, children, and their welfare. The stages include spiritual selfhood, period of the newborn, social priming, social apprenticeship, social entrée, social internment, adulthood, and old age, in juxtaposition with the conventional stages of prenatal development, neonatal period, infancy, childhood, puberty, adolescence, adulthood, and old age (see Nsamenang 1992). Children are socialized gradually and systematically to play more mature roles in valued social and economic activities at each successive stage. African (and other) cultures also grapple with the nonexistential phases of human selfhood in their own terms.

This perspective differs in theoretical concern from the more individualistic accounts by Freud, Erikson, and Piaget, but goes beyond them to specify that environments and cultures write specific texts on to human development, an essentially biological process. This introduces differences in the rate and purpose of development, thereby providing an innovative framework for examining developmental change as a biocultural phenomenon at the gene–environment interface.

3. *West African Conception of the Child*

The West African metaphor for child is *seed*, cultivated in a sociological garden in which child care is shared among adults and older siblings. In the garden, children learn systematically a cultural curriculum, from which they abstract the social, affective, cognitive, and other rule systems of the culture through observation, social participation, and anticipatory socialization. In so doing, children start to develop a shared identity from the moment they begin to distinguish 'that which is me' from 'that which is not me.' The practice of talking to children and projecting the personality that should emerge promotes the separation of 'self' from 'nonself.' This is a cognitive process that allows children progressively to develop an appropriate gender identity and personality.

3.1 *Socialization of Responsibility*

West African children perform a much wider range of economically productive and maintenance activities than is the case in contemporary, urbanized Western societies. The goal of training is to socialize responsible intelligence. The expectation that older siblings will care for infants and perform some chores is a priming process for such socialization. Priming is an ontogenetic strategy to prepare children progressively for their future adulthood. Through it, children make progress toward adulthood as they 'graduate' from one role to another. The Nso of Cameroon, for example, assign pivotal roles to children at different 'stations' of the life cycle, depending on the perceived level of the child's maturity and competence. Sending children on errands is another form of responsibility training that permits children to develop listening skills, ability to follow instructions, widen perspective, and follow goals. Yoruba children who went on errands demonstrated superior cognitive performance on an adapted Bayley Mental Scale (Yoruba Mental Subscale) compared to peers who did not (Ogunnaike 1997). Role assignment underscores parental awareness of children's cognitive maturity and readiness for specific tasks.

3.2 *Social Rather than Cognitive Competence as a Desirable Endstate of Development*

The West African social field is a dense network in which relatedness and sense of community are virtues. A central idea for developmental psychology, then, is the fact of human embeddedness and interdependence. As a 'situated' phenomenon, individual behavior and mentality are better understood within a network of behaving cognitive agents interacting with each other and the setting. It is perhaps in this light that West Africans value cognition not as an endstate of development, but as a means to social competence. Thus, whereas the Western child is socialized primarily through instructed learning by adults, with considerable cognitive stimulation, the West African child is trained primarily through participatory learning in ongoing cultural activities which highlight sharing and responsibility, particularly within the peer culture.

4. *Toward Internationalizing the Discipline: Theoretical and Methodological Challenges*

Although Lev Vygotsky ventured a sociocultural theory of psychological development, his framework viewed development cognitively, in terms of language and conceptual thought, rather than socially. Because theorizing has tended to dichotomize the social, cognitive, affective, and other components of development, an ideal situation is to weave them, and the interactions among them, into a comprehensive theory. Social ontogeny is a model that holds the potential to thread these facets together. The conceptual requirement is to open up to human diversity; the methodological challenge is to access and assess all facets of psychological functioning. This requires a multicultural mindset that accepts and grapples with all sources of developmental knowledge.

The conceptualization demands a new understanding of theory in close proximity to actual psychological phenomena, thereby necessitating a paradigm shift. It requires the researcher and researched to enter a

critical discourse within the hermeneutic circle of academic and indigenous psychologies in order to ground theory and method in sociocultural reality. It acknowledges that data *by* the people differ from data *about* the people. This approach is likely to advance a universalistic psychology of understanding, say, the development of thought and emotion in culture.

Evidence that the criteria psychologists and parents employ to judge children's competence do not always overlap reinforces the approach. The heart of the paradigm is collaborative interdisciplinary networking that permits several disciplines to cross-fertilize and enrich the research process. This approach calls for a learning posture that permits innovation to attend to phenomena in context rather than to focus stubbornly on predetermined samples of behavior. Because many aspects of developmental phenomena are inaccessible to current research techniques, it is absurd to adopt an either/or mentality to quantitative and qualitative methods. An ethnographic approach that utilizes both methods, especially with social learning, may reveal certain dimensions of psychological being, which are excluded by an either/or ideology, thus offering insights for theoretic innovation and creativity in methodology. The way forward is to borrow wisely and gain usefully and creatively from the perspectives of several disciplines rather than to hold obsessively to a single tradition. Furthermore, because physical, cultural, demographic, and other imperatives sometimes differ across cultures and contexts, indigenous psychology, being more sensitive to phenomena that might be ignored elsewhere, can enrich international psychology. It can inform the field of the extent to which some elements of human development portrayed by a Eurocentric discipline are inappropriate for non-Western peoples, thus highlighting what are, and are not, universal aspects of psychological functioning. Two contributions in this direction are the image of self and cognitive competence within the family among the majority of humanity (see Kagitcibasi 1996), and the introduction of social ontogeny as an alternative developmental theory (see Nsamenang 1992).

Indigenous psychology is not a project of cultural essentialism; neither is it a feature of a specific culture. Rather, it is an account of the dialectic between psychological attributes and the environmental conditions that shape them. In this sense, indigenous psychology in West Africa becomes an account of how the West African ecology primes human development. Because it resonates with African conceptions of family and childhood, it can inform psychologists and development agents about how West African communities will relate to interventions for children, a link that has been missing from development work in Africa.

If psychology was fashioned, say, within the African or Indian worldview, what difference might that make to its theory, method, and practice?

See also: Cultural Diversity, Human Development, and Education; Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Indigenous Conceptions of Aging; Indigenous Knowledge and Technology

Bibliography

Cruts A A N 1994 *Folk Developmental Psychology: An Empirical Inquiry into Social Constructionism*. CIP-Gegenens, Koninklijke Bibliotheck, Den Haaglemark, Denmark

Kim U, Berry J W (eds.) 1993 *Indigenous Psychologies: Research and Experience in Cultural Context*. Sage, Newbury Park, CA

Kagitcibasi C 1996 *Family and Human Development Across Cultures: A View from the Other Side*. Erlbaum, Mahwah, NJ

Nsamenang A B 1992 *Human Development in Cultural Context: A Third World Perspective*. Sage, Newbury Park, CA

Nsamenang A B 1999 Eurocentric image of childhood in the context of the world's cultures. An essay review of Images of Childhood. In: Hwang P C, Lamb E M, Sigel I E (eds.) *Human Development*. **42**: 159–68

Ogbu J U 1981 Origins of human competence: A cultural–ecological perspective. *Child Development* **52**: 413–29

Ogunnaike O A 1997 Yoruba toddlers: Relating cognitive performance to family sociodemographics and mediating factors in the child's environment. Unpublished doctoral dissertation, Tufts University, Medford, MA

Serpell R 1993 *The Significance of Schooling*. Cambridge University Press, Cambridge, UK

Valsiner J 1997 *Culture and the Development of Children's Action: A Theory of Human Development*. Wiley, New York

Walsh D J, Tobin J J, Graue M E 1993 The interpretive voice: Qualitative research in early childhood education. In: Spodek B (ed.) *Handbook of Research on the Education of Young Children*. Macmillan, New York, pp. 464–76

Zukow G P 1989 Communicating across disciplines: On integrating psychological and ethnographic approaches to sibling research. In: Zukow G P (ed.) *Sibling Interaction Across Cultures: Theoretical and Methodological Issues*. Springer, New York

A. B. Nsamenang

Individual Identities in Organizations

1. Trends in the Study of Individuals

Traditionally in Western (particularly American) studies, social scientists have treated the individual as a primary entity to be investigated. This pursuit entailed an essentialist view of the individual—assuming that the individual had '… an "authentic self" with knowable characteristics' (Gergen 1991, p. 7). Recently, these essentialist assumptions about individuals have been challenged as scholars have come to attach increased importance to the contexts in which people exist.

Although common sense, everyday experience, and a long history of psychological research in the Western world seem to suggest that organization studies should treat the individual person's self as a fundamental unit of analysis. These assumptions have been challenged at least both implicitly and explicitly. The implicit challenge was captured by Nord and Fox (1996) and Gergen (1991) advanced perhaps the most explicit challenge.

1.1 Implicit Challenge

Nord and Fox's (1996) implicit challenge was a serendipitous outcome of their attempt to summarize knowledge about the individual in organization studies using traditional headings such as personality, motivation, and learning. The difficulty they had in doing so led them to a surprise realization—the individual as previously assumed by organizational psychologists may have gradually vanished. However, to that point this disappearance (if it was one) was primarily a latent trend in American organizational behavior—a discipline that had traditionally rested on an assumption of the essentialist individual. Historically this discipline had grown out of psychology and had sought to uncover the essential properties of individuals, assess variations among individuals, and capitalize on this understanding to help to make organizations function more efficiently and more effectively.

However, the change was subtle. On the one hand, Nord and Fox reported considerable counter evidence that the essentialistic perspective was far from entirely dead. For example, the strong interest in the so called 'big five' personality characteristics and the increase in research on personality traits during the period 1987–1992 in comparison to the period 1974–1986 noted by Pfeffer (1997) revealed that many investigators continued to employ the essentialist view. On the other hand, both in applied, and basic psychology, many other researchers were finding it increasingly difficult to sustain it, for example, many psychologists and organizational behaviorists who were publishing in mainstream journals even though not confronting the essentialist view *per se*, were increasingly emphasizing the role of context. This development was evident over time as a number of psychological inquiries conducted in traditional terms that when viewed over time reflected a pattern of a partially conscious skepticism about the essentialist perspective.

Nord and Fox (1996) pointed to the existence of this implicit overall pattern cautiously ending the title *The Individual in Organization Studies: The Great Disappearing Act?* with a question mark. Despite the caution they noted that in contrast to the traditional essentialistic approach of a few decades earlier '… emphasis has shifted from viewing individuals independently of context to consideration of the interplay between individuals and their context' (p. 148). Defining context as '… attributes of the physical and social systems in which individuals exist' (p. 149), Nord and Fox found that the progression of chapters in the acclaimed *Annual Review of Psychology* as well as in the highly regarded *Handbook of Industrial and Organizational Psychology* revealed that from the late 1970s on, research on the individual in organizations had become increasingly contextualized. In addition, they pointed to the growing role of contextualized understandings in leading mainstream psychological journals such as work by Sampson (1985, 1988, 1989) in *American Psychologist*, Baumeister (1987) in the *Journal of Personality and Social Psychology*, and Markus and Kitayama (1991) in the *Psychological Review*. A key theme of Nord and Fox was that this disappearance of the individual in the organization studies and related literature was an emerging and implicit development over the last several decades.

For present purposes, Nord and Fox's suggestion that the individual may be disappearing in organization studies is best viewed as an inductive product from a review of work by researchers who were operating under the traditional positivist traditions of general and organizational psychology. For the most part these scholars were not even aware of an altered mindset about the position of the individual. In short, the trend from the traditional essentialist view of the individual to a much more contextualized stance was implicit, appearing to be a latent outcome of the mainstream literature where the contributors were probably operating within the relatively essentialistic traditional paradigm.

1.2 Explicit Trend

By contrast Gergen (1991) offered a much more explicit and less cautious recognition of a similar but longer term trend. In *The Saturated Self*, Gergen argued persuasively that an altered view of the individual had occurred. In contrast to Nord and Fox's inductive use of research from organizational studies and traditional social science, Gergen applied an avowedly postmodern perspective to a broad array of historical and academic knowledge. Despite this difference, the congruence between his notion of the saturated self and Nord and Fox's disappearence thesis is strong.

Gergen argued that '… traditional assumptions about the nature of identity are now in jeopardy … The very idea of individual selves—in possession of mental qualities—is now threatened with eradication' (p. x). The complexity of Gergen's book makes it difficult to summarize in a short space, so only a few highlights can be considered here.

According to Gergen the dominant intellectual trends of the twentieth century featured two vocabu-

laries, a first inherited largely from the nineteenth century he termed as the romanticist view of self (a view that attributed to each person characteristics of personal depth such as passion, soul, creativity, and moral fiber). Beginning in the twentieth century, a second vocabulary—the modernist one (the view in which the chief characteristics for the self did not reside in a domain of depth, but rather in the ability to reason, including beliefs, opinions, and countless intentions)—challenged the romantic vocabulary. Despite their differences both vocabularies shared essentialist views of the self.

However, Gergen asserted that both vocabularies were becoming outdated as a result of a set of forces that produce social saturation. As he put it: '… both the romantic and modern beliefs about the self are falling into disuse and the social range in which they support are eroding' (p. 6). Social saturation entails a multiplicity of incoherent and unrelated languages of the self. Everything that people know to be true about themselves is challenged by other voices within them, producing fragmentation due to the large number of incoherent and disconnected relationships a person has. Gergen wrote that these relationships: '… pull us in myriad directions, inviting us to play with such a variety of roles that the very concept of authentic self with knowable characteristics recedes from view. The fully saturated self becomes no self at all' (p. 7). As the saturated self develops under postmodern conditions, it throws the very concept of personal essences into doubt. Drawing on the work of anthropologist Clifford Geertz, Gergen suggested that some cultures contrast sharply with the American culture by not sharing the American emphasis on the individual; in these other cultures a person is often considered as a representative of some general category of social significance in the culture (e.g., 'first born' or 'grandfather of …'). Moreover, members of these other cultures do not hold the notion of biological givens as strongly as Americans tend to and in some of these cultures investigators have found it difficult to locate any terms referring to 'inner states.' Gergen observed that in America the essentialist view of the individual led people to view the human being more or less as a machine and further that holding the image of the individual as a machine but that recently, this essentialist self has given way to a new form of self-consciousness in the postmodern period, because the technologies of social saturation contribute to '… erasure of the individual self' (p. 49). Gergen suggested that what we take to be the individual self is increasingly indistinguishable from the pattern of social relationships with which the physical individual interacts. As people come to feel and express themselves more fully in a growing number of relationships they become populated with the character of others. They contain multitudes, even though they '… appear to one another as single identities, unified, of whole cloth. However, with social saturation, each of us comes to harbor a vast population of hidden potentials …' (p. 71).

Keeping in mind Nord and Fox's (1996) and Gergen's (1991) suggestions of the disappearance of or at least radical transformation of the individual as an analytical category and the growing emphasis on context, some other recent developments in sociology and psychology add support to the disappearance hypothesis. Here, two of these developments one from sociology and the other from social psychology are considered.

Recently, the famed sociologist Sennett (1998) coined the phrase 'corrosion of character' to describe the impact of recent changes in the nature of work upon people. Sennett began with a conversation he had with the son of one of his working-class respondents from Sennett and Cobb's (1972) famed study decades ago published as a book *Hidden Injuries of Class*. The changes were enormous. Nearly a half century ago the stable relationships people experienced at work contributed to them developing a rather stable sense of self. However, now the stable work relationships have vanished as recent changes in technology and life in economic organizations as exemplified by the motto now guiding many organizations 'no long term' have disappeared. As a result, people now encounter rapidly changing and temporary attachments to their work and organizations. As a result, individuals have experienced a 'corrosion of character.' As Sennett put it, the experience of disjointed time threatens '… the ability of people to form their characters into sustained narratives' (1998, p. 31).

Yet another challenge to the essentialist self can be found in recent social-psychological literature. Consider, for example, the October 1994 special issue of *Personality and Social Psychology Bulletin* focusing on the self and the collective. In their introductory essay, the issue's editors Miller and Prentice (1994) observed that traditionally social psychologists viewed the group as something external to people and focused on understanding how individuals behaved in a group. But then, roughly in the late 1960s and early 1970s, Tajfel (1969) and his co-workers pioneered a very different view—one that located groups inside the heads of individuals, instead of asking how individuals behaved in groups and demonstrated '… that the social categories with which a person identifies have a profound impact on his or her psychological functioning' (p. 451). Further, Miller and Prentice observed that this shift resembled a trend in anthropology that has led to conceiving culture in cognitive terms, that is, a conceptual structure in people's heads' (p. 451). They asserted that all the writers in the special issue assumed '… that the study of the collective is a necessary companion to a study of the self' (p. 451). In other words, they rejected an individuated view of the self in favor of a more socially embedded one that '… anchors the self-concept in the broader world'

(p. 451) and views the content of one's self-concept as derivative of the social groups a person identifies with. Thus, the collective is part of the self not merely a context for making judgments about oneself. This development helped to breakdown some barriers between psychology and other disciplines especially with contemporary cultural psychology in which the self is closely tied to the collective. As Miller and Prentice (1994) put it: '… cultural psychologists argue that collectives are important because a person's construal of the relation of the self to relevant social categories determines much of the *structure* of the self' (p. 452). Again, contemporary conceptions of the self tend to emphasize a social more than an individualistic nature.

Further, in the first paper in the special issue itself, Turner et al. (1994) advanced 'self-categorization theory.' Their view tends to de-personalize self-perception: '… individuals tend to define and see themselves less as differing individual persons and more as interchangeable representatives of some shared social category membership' (p. 455). Thus, the self varies as a function of social situations and the concept of the self as a separate mental structure is not necessary. Rather all a person's cognitive resources, including long term knowledge, implicit theories, cultural beliefs, and so on, are available and are deployed when necessary to create the needed self-category. In short, people deploy their cognitive resources reflexively to categorize themselves when appropriate. Turner et al. (1994) added 'we can therefore reject the notion of self-concepts as … stored, invariant structures and the associated model of category use as a process of *activating an already existing representation* …' (p. 459). Clearly, another example of the abandonment of the essentialist view of the person.

Having considered how the essentialist individual has in many respects been transformed into social identity or as Gergen (1991) put it the saturated self, we turn to a second recent trend—this one at a more macrolevel. The development of the concept of organization identity and an associated change in thinking about organizations.

2. *Conceptual Developments Concerning Organization Identity*

2.1 *Identity and the 'Disappearing' Organization*

In the 1980s, the concept of organization identity began to become popular. Interestingly in light of what we have seen about the fate of the individual, some students of organizational identity seem to have suggested that in a way the traditional organization may be disappearing too. But when we view the two trends together, we see more of a transformation of each into the other. For example, whereas the individual appeared to be moving into the collective, the collectivities we call organizations seem to be disappearing into the individual. For example, recently, leading students of organization identity (Albert et al. 2000) reasoned that environmental turbulence has driven organizations to become increasingly organic, therefore traditional bureaucratic structures are less prevalent. They wrote: '… increasingly, an organization must reside in the heads and hearts of its members.' Of course, we must be careful here. Their statement does not say the organization has disappeared but merely suggests, in developing an intellectual view, that its focus is changing. This shift reveals the potential organizations have to serve as forms of the collective relationships that Turner et al. (1994) and Miller and Prentice (1994) suggested define individual identity.

Knowledge from the growing literature on organization identity can be summarized under three headings: (a) conceptual work on the identity concept, (b) the concept's practical relevance for managing today's organic structures, and (c) identifying, that is, the process through which individuals become identified with organizations. Recent developments on these matters are reviewed below.

2.1.1 Conceptual work on organization identity. Two things can safely be said about efforts to conceptualize organization identity: (a) interest in the topic is relatively new and (b) knowledge about it is somewhat chaotic. First, while the topic is not totally new to organization studies it would seem that topics from the past such as organizational culture and Etzioni's (1961) classic typology of types of organization involvement share related concerns. However, recently 'identity' *per se* has attracted widespread interest. Albert and Whetten's paper (1985) paper in *Research in Organizational Behavior* sparked this interest. Albert and Whetten began by acknowledging that historically identity had been treated as a loosely coupled set of ideas; but it would be better if the term 'organizational identity' was made more scientifically tractable (i.e., a conceptualized term in a way to make multiple empirical questions and hypotheses become visible).

To this end, they suggested three criteria for an adequate statement of organizational identity: (a) central character, (b) distinctiveness, and (c) temporal continuity. The first criterion, a statement of central character, distinguishes the organization on a basis of something that is important and essential. In defining the second criterion, distinctiveness, they relied on parallels here between the organizational identity and the psychologist Erickson's (1980) discussion of individual identity. Following Erickson, they advanced identity as a classification that locates the individual as recognizably different from some others, but simultaneously similar to those who are members of one's

own class. Their third criterion, identity over time, treated the temporal aspect of identity. Again drawing on Erickson's work on individual identity they pointed to the negative consequences of the lack of fundamental continuity over time and formulated an intriguing model of paths of identity change an organization often follows over its life cycle. Further, they noted that oftentimes organizations have dual rather than just single identities, for example, universities have many features that are similar to those of churches and features that are similar to those of businesses.

Most writers who have drawn on Albert and Whetten's (1985) concept of organization identity emphasize some variants of these criteria and define identity as what members of an organization believe is central, enduring, and distinctive about their organization. However, although there is considerable agreement on these features of a definition, there also is substantial controversy about what the term means.

The controversy is great enough to justify saying that conceptualization of organization identity is somewhat chaotic. Information from two recent publications—Whetten's (1998) book *Identity in Organizations* and a special topic forum in the January 2000 issue of the *Academy of Management Review*—illustrates the chaos.

Scott and Lane (2000), commenting on the controversy about definition, observed that in contrast to Albert and Whetten's definition, some scholars have defined organizational identity at the individual level '… as a cognitive image held by a member of an organization' (p. 43). This definition is associated with efforts to link organization identity to an organization's image which following Gioia et al. (2000) '… includes notions involving the ways organization members believe others see the organization' (p. 63).

The definition controversy stimulated Gioia et al. (2000) to reconceptualize organization identity as '… a potentially precocious and unstable notion, frequently up for redefinition and revision by organization members' (p. 64). However, elsewhere Gioia (1998) had provided deeper insight into the sources of the problem. Writing on the conceptualization problem Gioia (1998) examined three perspectives: (a) functionalist, (b) interpretive, and (c) postmodern.

Social scientists who use the functionalist perspective, according to Gioia, believe it is appropriate to view phenomena through natural science like models, make realist and objectivist ontological assumptions, and to treat identity as something that exists and is available for study. Summarizing this view Gioia (1998) observed: 'The study of identity is often treated as the study of something someone or some organization has: it is thus rendered as understandable to an interested observer and usually is conceptualized in fairly dispassionate, theoretical terms' (p. 26).

In contrast, when one uses the interpretive lens things are more subjective and the distinction between the researcher and the researched is blurred. The project of the researcher under interpretivism is to render the constructions of the informants faithfully and to accurately represent their interpretations. Using the third (i.e., postmodern) lens introduces additional contrasts. The postmodern perspective encourages challenging or at least suspending judgment on most current ontological, epistemological, and methodological assumptions. It questions the basis for all beliefs and study, including its own. When organizational identity is studied under this lens, frequently it is construed as at best a fiction perpetuated by dominant parties.

Clearly, if all three of these perspectives are currently in use, and they seem to be, it is not surprising that there is a lack of agreement about what organization identity is. Gioia concluded '… organizational identity is still an embryonic area although of some significant potential for understanding how and why organization members think and act as they do' (p. 29).

Among others commenting on the conceptual problems in defining identity (whimsically one might say about the identity of identity) have been Bouchikhi et al. (1998) who concluded their published conversation about identity by observing: '… one's understanding of the concept of organizational identity depends upon the perspective one takes in approaching it' (p. 76).

The confusion about what 'identity' is was perhaps best summarized by Ashforth (1998) when writing the epilogue to the book *Identity in Organizations: Building Theory through Conversations*, reflecting back on the conversations that had produced the book, concluded: 'The identity question (Who Am I?) may be a question with which both individuals and organizations must struggle in order to create a sense of meaning, and a sense of direction, for their activities. This book attempts to convey a sense that identity is a concept (or construct) in its infancy. I invite you, the reader, to join us in the marvelous discovery process centered around organizational identity …' (p. 293).

Ashforth's statement is instructive in two ways. First, it captures the ambiguity surrounding the concept. Second, its upbeat tone captures the excitement that students (albeit at this point a relatively small number) of the subject have for the concept.

2.1.2 Sources of excitement. What exciting potential do these people see? For one thing they see the concept's potential for bridging across topics and levels of analysis. Albert (1998), for example, observed that a major advantage of the concept of identity is that it is a bridging concept—it facilitates asking cross-level questions that link micro- and macrostructures and processes.

In that same volume, Gioia (1998) suggested other reasons for excitement, its fit with how people see

things, its flexibility, and practical value to managers, and its openness. Reflecting on Albert and Whetten's original paper, he suggested that their view of identity '… captures the essential (in the phenomenological sense) features of an organization' (p. 21) and that the dimensions they outlined parallel those noted for individuals, although differing mainly in their collectively-shared character. Another reason for excitement is that the concept of organizational identity is highly flexible. This flexibility is especially valuable to modern organizations because their environments are changing so rapidly and to adapt, organizations need to be able to change their identities rapidly and they need to display identities of both stability and instability simultaneously. They need to manage the appearance of stability and order so that change can be managed, while still retaining central features of core identity. Thus, not only is identity useful as a conceptual bridge, it also is a potential aid to managers seeking to cope with opposing demands to maintain some sense of coherence while simultaneously adopting to the demands of rapidly changing environments. Further reason for excitement Gioia concluded is that the area of study is a '… wide open domain' (p. 29).

The role of openness in generating intellectual excitement becomes clear when we examine the identity concept in light of recent accounts of the scientific process such as the work of the physicist/philosopher Giere (1988). Giere observed that science proceeds through a process of individual scientists bringing their cognitive resources (special competencies) to bear on intellectual topics. Thus, one might expect a wide-open domain to attract the attention of a variety of scholars. Currently, the study of identity invites people of all stripes from many areas of study to apply their cognitive resources. Recall, for example, Albert and Whetten (1985) sought to shape a diverse network of ideas into a scientifically tractable concept. Then too note that the topic of identity is related to other popular organization concepts including culture and legitimacy, and also by its nature invites input from a host of diverse literatures and disciplines. In short, 'identity' offers great opportunities for people having diverse cognitive resources to become involved. The variety of contributors to the recent *AMR* special forum is a case in point; it brought writers with a host of diverse sets of cognitive resources, including stakeholder analysis, psychodynamics, postmodernism, and demographic diversity, to the table. All of this follows a paper seeking to make the concept scientifically tractable.

We have already seen that the concept of organizational identity offers practical as well as intellectual benefits. This fact deserves special attention.

2.2 Value of Identity to Contemporary Managers

If identity is a response to the particular question Albert and Whetten (1985) observed 'What kind of organization is this?' (p. 292). Answers to it refer to features that are '… core, distinctive, and enduring' (p. 292). Among other things this orientation leads to the practical value of the concept of contemporary managers.

Throughout the discussion so far, there have been hints that 'organization identity' is useful for managers of today's organizations because they need some way to achieve some modicum of internal stability while simultaneously responding flexibly to turbulent environments. Albert et al. (2000) showed that the concept of identity can serve in exactly this way. They wrote: '… it becomes more important to have an internalized cognitive structure of what the organization stands for and where it intends to go—in short, a clear sense of the organization's identity. A sense of identity serves as a rudder for navigating difficult waters' (p. 13). Similarly, Pratt and Foreman (2000) argued that a firm's identity can be a source of sustainable competitive advantage and that since organizations often have multiple identities, the concept may be helpful in conceptualizing ways to manage the pluralism. Scott and Lane (2000) extended the managerial value of the concept by relating it to stakeholder theory. They observed: '… Managers choose organizational images for presentation to stakeholders for strategic reasons. Corporate reputation building is principally concerned with promoting attractive organizational images … for purposes of goal attainment, and it is the primary job of leadership to mange organizational identity toward that end … Managers, as organizational leaders, are expected to assume the role of speaking for the organization. Making public commitments on the part of an organization influences people to chance their self-perceptions and act consistently with the presented self-image … Although all organizational participants have perceptions of organizational identity and can promote these publicly, they are not charged with doing so—this is not an expectation of their organizational role' (p. 47). In short, the concept of identity may help managers deal with external as well as internal constituents.

This leaves us with one final topic for completing our overview of current work on conceptualizing organization identity—the process by which people come to identify with the organization. In a word—identifying or the identification process.

2.3 Identifying

We have seen that students of identity believe that a major value of the concept of organization identity has been its ability to help bridge between the micro and macro levels of analysis. One path onto this bridge is study of the process by which individuals come to identify with collectivities such as organizations. To do this, it is useful to de-emphasize the noun 'identity' in favor of more process oriented terms such as

identifying and the 'identification process.' This shift draws attention to areas of study that shed light on related processes but were not originally concerned with identity. Here, two are considered: first some writings of Karl Weick and, second, the long-standing interest in organizational socialization.

Weick's (1995) work on sensemaking, especially his discussion of the individual as 'sensemaker,' using reactions of others to come to an understanding of self, is helpful. Weick's clever question 'How can I know who I am until I see what they do?' (p. 23) summarizes to this perspective neatly, revealing that individuals develop identities by observing how their actions and beliefs about themselves affect the reactions of others. These reactions are the substance for a sensemaking process that produces an individual's view of him/herself. This view helps us to see how (i.e., the process through which) what we have said about the individual becoming an expression of the social context may occur. However, Weick provided even greater insight into the process in this, the sense of self is highly influenced by a dynamic interaction with the organizational context in which people are seen engaging in a process of sensemaking through interaction in his study of the Mann Gulch disaster.

Weick's (1993) study of the Mann Gulch disaster (a forest fire that killed 13 firefighters) informs our understanding of identification process mainly by a negative instance—the loss of identity. Weick maintained that a major contributor to the breakdown in social structure that led to the disaster was a foreman's command to the trapped crew to drop their tools. How could such a command make for disaster? To many readers not facing an onrushing fire, the command to people facing one to drop the heavy equipment they were carrying seems sensible since the reduced weight should have made it easier for the men to run away (especially since some of the running was up a steep incline) from the fire. However, Weick argued it turned out to be a source of disaster because the tools defined the men's identity of themselves as firefighters. Dropping their tools removed their identity—the individuals could no longer answer the question 'Who am I?' Further, since an authority figure told them to discard this source of identity, the legitimacy of their whole social order came into question. Thus, Weick's analysis, although not intentionally directed to studying identity, revealed how everyday elements such as one's tools impact individuals' identification processes (another example of the socially created individual discussed earlier) and how people's identities are part of the collective order.

Insight into the identifying process can be gained from research on another topic—organizational socialization. Ashforth and Mael (1989) using social identity theory (SIT) argued that research on socialization reveals that in organizations, social identification occurs as situational definitions and self-definitions emerge through symbolic interactions and people develop self-definitions through interpreting responses of others. In addition to these insights into identifying from sources not grounded in the identity concepts, the identity literature itself provides helpful insights.

In the identity literature, Ashforth (1998) suggested four metaphors for thinking about the process: imprinting, love, play, and theater. Perhaps the most valuable study concerning identifying contained in the identity literature itself is Dutton and Dukerich's (1991) path-breaking case study of the Port Authority of New York and New Jersey. This study is especially useful in revealing the dynamic relationship between individual and organizational identities. By revealing the interplay between an organization's identity and the identity of its members Dutton and Dukerich called attention to the role of an organization's image (i.e., the way that members believe that outsiders view the members' organization). Because an organization's image is an indicator of each member's identity, image is directly related to the self-esteem one can derive from belonging to the organization. Dutton and Dukerich considered six attributes that their informants believed distinguished the Port Authority as an organization and discovered how these interpretations and the organization's responses to a troubling environment change coevolved. They concluded that the organization's identity acted as an important reference point that members used to assess importance of presenting issues and to guide how they responded. In sum, Dutton and Dukerich showed how individual and organization identity interact to influence how individuals come to identify with the organization.

3. *Conclusion*

In drawing together what we have seen so far, several key themes are prominent.

(a) Recently, strong challenges to the essentialist view of the individual have appeared. In a sense, as an analytical tool, the traditional, essentialist individual seems to be disappearing. As this is happening, social entities are increasingly recognized as a major source of data for people trying to answer an identity question such as 'Who am I?'

(b) In a similar sense, the entities called organizations, as we knew them in the twentieth century, seem to be disappearing and, as Albert et al. (2000) put it: 'Increasingly, an organization must reside in the heads and hearts of its members' (p. 13).

(c) In organization studies, the micro and macro sides of the field have not been fully integrated.

(d) In recent years, identity has been advanced as a concept with potential to provide a bridge between the micro (i.e., individual) abutment of the field and the macro (i.e., organizational) abutment.

We have seen evidence to support the validity of all four points. However, when the four are taken

together they reveal a somewhat bizarre result. Certainly, the idea of identity is a potential bridge between the micro and the macro. However, as we have seen, the abutments at both ends that are to be bridged seem to be in a sense disappearing with the two disappearing ends becoming enmeshed in each other as shown by points (a) and (b). If so, at the extreme we might be left with a bridge, that is, 'identity' from nowhere to nowhere.

If we extend this metaphor, we could envision being left with the bridge called organization identity. As the study of identity seems to be moving as summarized by combining points (a) and (b) above, for some purposes individual and organization identity could merge into one concept.

Although such an extreme outcome seems bizarre, the stimulus to even imagine it, derived as it was here from exploration of trends in the study of individual and organization identity, does suggest identity is a high potential concept for organization studies in the future. The fact that the idea already has practical as well as theoretical value further attests to its potential in the postmodern world where the hard substance of traditional entities seems to be softening.

See also: Identity and Identification: Philosophical Aspects; Organization: Overview; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture; Organizational Culture, Anthropology of; Organizations, Sociology of; Personal Identity: Philosophical Aspects

Bibliography

Albert S 1998 The definition and metadefinition of identity. In: Whetten D A, Godfrey P C (eds.) *Identity in Organizations*. Sage, Thousand Oaks, CA, pp. 1–13

Albert S, Ashforth B E, Dutton J E 2000 Organizational identity and identification: Charting new waters and building new bridges. *Academy of Management Review* **15**(1): 13–17

Albert S, Whetten D A 1985 Organizational identity. In: Cummings L L, Staw B M (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 7, pp. 263–95

Ashforth B E 1998 Epilogue: What have we learned, and where do we go from here? In: Whetten D A, Godrey P C (eds.) *Identity in Organizations*. Foundations for Organizational Science, Thousand Oaks, CA, pp. 268–72

Ashforth B E, Mael F 1989 Social identity theory and the organization. *Academy of Management Review* **14**(1): 20–39

Baumeister R R 1987 How the self became a problem: A psychological review of historical research. *Journal of Personality and Social Psychology* **52**(1): 162–76

Bouchikhi H C, Fiol M, Gioia D A, Golden-Biddle K, Hatch M J 1998 The identity of organizations. In: Whetten D A, Godfrey P C (eds.) *Identity in Organizations*. Foundations for Organizational Science, Sage, Thousand Oaks, CA, pp. 33–80

Dutton J E, Dukerich J M 1991 Keeping an eye on the mirror: Image and identity in organizational adaptation. *Academy of Management Journal* **34**(3): 517–54

Erickson E H 1980 *Identity and the Life Cycle*. Norton, New York

Etzioni A A 1961 *A Comparative Analysis of Complex Organizations*. Free Press, New York

Gergen K J 1991 *The Saturated Self*. Basic Books, New York

Giere R N 1988 *Explaining Science*. University of Chicago Press, Chicago

Gioia D A 1998 From individual to organizational identity. In: Whetten D A, Godfrey P C (eds.) *Identity in Organizations*. Foundations for Organizational Science, Sage, Thousand Oaks, CA, pp. 17–31

Gioia D A, Schultz M, Corley K G 2000 Organizational identity, image, and adaptive instability. *Academy of Management Review* **25**(1): 63–81

Markus H P, Kitayama S 1991 Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review* **98**: 224–53

Miller D T, Prentice D A 1994 The self and the collective. *Personality and Social Psychology Bulletin* **20**(5): 451–3

Nord W R, Fox S 1996 The individual in organizational studies: The great disappearing act? In: Clegg S R, Hardy C, Nord W R (eds.) *Handbook of Organization Studies*. Sage, Thousand Oaks, CA, pp. 148–74

Pfeffer J 1997 *New Directions for Organization Theory*. Oxford University Press, Oxford, UK

Pratt M G, Foreman P O 2000 Classifying managerial responses to multiple organizational identities. *Academy of Management Review* **25**(1): 18–42

Sampson E E 1985 The decentralization of identity: Toward a revised concept of personal and social order. *American Psychologist* **40**: 1203–11

Sampson E E 1988 The debate on individualism: Indigenous psychologies of the individual. *American Psychologist* **43**(1): 15–22

Sampson E E 1989 The challenge of social change for psychology. *American Psychologist* **44**(6): 914–21

Scott S G, Lane V R 2000 A stakeholder approach to organizational identity. *Academy of Management Review* **25**(1): 43–62

Sennett R 1998 *The Corrosion of Character*. Norton, New York

Sennett R, Cobb J 1972 *The Hidden Injuries of Class*. Vintage Books, New York

Tajfel H 1969 Cognitive aspects of prejudice. *Journal of Social Issues* **25**: 79–97

Turner J C, Oakes P J, Haslam S A, McGarty C 1994 Self and collective: Cognition and social context. *Personality and Social Psychology Bulletin* **20**(5): 454–63

Weick K E 1993 The collapse of sensemaking in organizations: The Mann Gulch disaster. *Administrative Science Quarterly* **38**: 628–52

Weick K E 1995 *Sensemaking in Organizations*. Sage, Thousand Oaks, CA

Whetten D A 1998 Preface: Why organizational identity and why conversations? In: Whetten D A, Godfrey P C (eds.) *Identity in Organizations*. Sage, Thousand Oaks, CA, pp. vii–xi

W. R. Nord

Individual/Society: History of the Concept

The concepts 'individual' and 'society' are intimately related. Societies, whatever else they may have been thought to be, are collections or groups of individuals. Many philosophers and psychologists have thought

that individuals only took on the characters they did by virtue of the influence of the society in which they grew up. But which had priority? Which was the cause and which the effect? The problem of how the concepts of individual and society are related has been part of a more extensive philosophical debate on the nature of classes and their relations with their own individual members. Some have argued that we can dispense with class concepts in favor of ones referring only to individuals, others that any such attempt will necessarily fail.

1. Philosophical Aspects of the Individual/Society Contrast

1.1 Kinds of Groups

The problem of how to interpret the concepts of individual person and social collective is embedded in a more ample philosophical issue, the general question of how to understand the contrast between individuals and groups and the inter-relations between them. There are two main kinds of collectives or groups that are of relevance to social science. There are taxonomic groups in which the members share some set of qualifying properties but may have no real relations with one another. Two people may be members of an ethnic minority solely on the basis of their mother tongue, or their religious affiliation. This concept of a collective is particularly appropriate to the study of groupings like social classes. Then there are structured groups like families or institutions, the members of which may be biologically or socially related to one another. In structured groups members are usually different from one another with respect to group membership. Mum and Dad are members of the nuclear family just insofar as one is female and the other male. In recent sociology and social psychology such groups have been analyzed by the use of the concept of role. It seems particularly appropriate to the study of institutions and teams.

1.2 The Related Problem of Universals

Philosophical problems associated with taxonomic groups have a history that reaches back through Medieval attempts to deal with the problem of universals into antiquity and the Platonic forms. How could one account for the fact that many individual entities were the same color? Or the same shape? According to the Ancient Greek philosophers this fact had to be explained by some version of the form and matter distinction. Clearly similar things shared a common form. How did that form exist? Plato built his account on the principle that the property displayed by each individual entity was a realization in the material world of a form that transcended it. The Platonic account has often been rejected. In the later Middle Ages, William of Ockham argued that the only common element in a group of things of the same color, for example, is the use of the same *word* to describe each of them.

1.3 Emergent Properties

The idea that collectives are somehow more than the sum of their members has persisted, albeit in rather different ways, down to the present day. The notion of an 'emergent property' has been coined to describe cases in which a collective of parts, say a brain composed of neurons, has a property, a capacity for thought, that is not possessed by any neuron taken singly. Emergent properties contrast with aggregative properties such as weight. Both brains and their component cells have weight. An important debate in social science centers on the issue of whether among the emergent properties of social structures are causal powers. Some have argued that agency is an attribute only of the human beings who create these structures.

1.4 Internal and External Relations

The third main issue concerning the relation between individuals and collectives is tied in with the distinction between internal and external relations. Two beings are internally related if their natures are partly or wholly determined by their standing in those relations. Marriage is an internal relation in that a person becomes a husband or a wife by virtue of being related by marriage to someone else. Spatial relations are commonly cited as examples of external relations between people, though they may not be for material things in general. The general question is whether people are the products of the societies in which they have a place or are the societies' products of the activities of people? Contemporary structuration theory takes a middle way, seeing people as shaped by their social relations and in turn, in their subsequent ways of living, recreating those social relations, albeit imperfectly.

2. Hobbes and Locke: the Genesis of the Sovereign

2.1 Leviathan: Individuals Covenant to Maintain a Sovereign

In Protestant Europe the individualism of the renaissance soon evolved into the idea of the state as a corporation of citizens. What sort of collective could

such a corporation be? Where is the sovereign to be found in such a picture of the state? From what source does the state derive authority, when the Divine Right of monarchs is no longer accepted? The overt philosophical problem, which is to ask in what way a corporation exists other than as an aggregate of its members, was displaced by the more pressing question of how a group could come to have an identity to which a citizen owed a duty of obedience. The social contract, an agreement to give up some of one's individual freedoms for the benefits of civil society, in particular personal security, was invoked by Hobbes, Locke, and later Rousseau as a necessary precondition for a social order to arise out of individual aspirations and intentions.

This was the moment at which the collective emerged from the aggregate of individuals. It was also the ground of the state's authority. Both aspects of the social contract appear very clearly in Hobbes's *Leviathan* of 1651.

> [the agreement that creates civil government] is more than Consent or Concord; it is a real Unity of them all, in one and the same Person [the Sovereign], made by Covenant of every man with every man … [This creates a Commonwealth, the essence of which is] One Person, of whose Acts a great Multitude, by mutual Covenants one with another, have made themselves everyone the Author, to the end he may use the strength and means of them all, as he shall think expedient, for their Peace and Common Defence (Hobbes 1651, Part 2, Chap. 17).

Yet it is not the collective that is sovereign, since the purpose of the creation of the unity is the delegation of powers to an individual. Each person is the 'author' of the acts of whosoever is the sovereign.

2.2 *The General Will*

Rousseau, writing a century later, dispensed with the sovereign as such. Here we have the submergence of the individual in the collective.

> Each one of us puts into the community his person and all his powers under the supreme direction of the general will; and as a body we incorporate every member as an indivisible part of the whole (Rousseau 1764 [1968], Book I, Chap. 6).

Nevertheless he insisted that the legislative role of the social group was necessarily distinct from the power to execute the general will.

The metaphysical question of the kind of reality that the collective or group might have, by way of contrast with that of the members taken individually, is not directly addressed by these authors. For instance no one in the high era of the social contract paused to ask for the criteria of identity and individuality of a society, nation, or institution. Nor is the issue of the status of emergent properties raised so far as I know. It is certainly implicit in Rousseau's version of the social contract in that the 'general will' of the collective is explicitly distinguished from the aggregate of personal wishes and intentions, the 'will of all.'

3. *Vico and the Role of Unintended Consequences*

The question of how a group of individuals can come to have a collective identity troubled Giambattista Vico (1668–1774). Unlike his distinguished predecessor, Machiavelli, he did not see the collective as the product of the power and will of the sovereign. He remarks that

> It is true that men have themselves made this world of nations … But this world, without doubt, has issued from a mind often diverse, at times quite contrary, and always superior to the particular ends that men had proposed to themselves; which narrow ends, made means to serve wider ends, it has always employed to preserve the human race upon this earth (Vico 1725, p. 1108).

How is this possible? How can one account for the transition from anarchy to the institutions of the state? It cannot be by the accidental coordination of individual motivated acts. The superior 'mind' Providence employs individual intentions to bring about social order.

Providence is not God, but should be interpreted, I believe, as a metaphor for the flux of unintended consequences of individual acts. There is no need for a social contract to bring the conditions of order into being in Vico's view. Social order comes to be because the conditions for its emergence are already in existence.

> Yet chance did not divert them nor fate force them out of this natural order. For at the point when the commonwealths were to spring forth, the matter was all prepared and ready to receive the form, and there came forth the commonwealths composed of body and mind. The prepared materials were these men's own religions, their own languages, their own lands, their own nuptuals, their own names, their own arms … and because these things were their own they were completely free and therefore constitutive of true commonwealths. … The fathers at this juncture, by uniting themselves in an order, came to produce the sovereign civil power, just as in the state of nature they had held the family powers … (Vico 1725, p. 630).

By bringing about the birth of families, each with its own language and rules of life, Providence had at the same time brought into being the natural law to be used in the management of commonwealths.

The identity of a nation was identical with the identity of a family, only writ large. Vico's collectives were structural not taxonomic.

3.1 The Problematic Status of Collectives

Vico's account of the genesis of the state neatly avoids the issue of how individuals can jointly constitute a collective entity the properties of which are not just aggregates of the properties of individuals. He denies individual intentions an explicit role and at the same time traces the means of maintaining institutions back to simpler institutions in which individual intentions directly create social order. In essence Vico assumes the native character of institutions. They are just there.

4. Utilitarians and Collectivists

4.1 Individuals and Freedom

The Utilitarians, Jeremy Bentham (1789), and the Mills, father and son, had no difficulty with the concept of collectives of human beings, for rather the opposite reason from Vico. For them the properties of groups were just aggregates of the attributes of the individual members. The happiness of all was the arithmetical sum of the happinesses of each. The Utilitarian account of the relation between individual and collective is sometimes used to illustrate the fallacy of composition. To infer from the fact that each of the 20 soldiers in a platoon is six feet in height that the platoon is 120 feet high is an obvious fallacy. Is it not equally fallacious to give a measure of the happiness of a society as the sum of the happinesses of individual members?

John Stuart Mill's (1859) *On Liberty* presupposes a sharp distinction between individual and society, at least when 'society' takes the form of government, public opinion and so on. He argued for the fundamental moral propriety of resisting those who claim to be the voice of the mass. Freedom and individuality are more or less the same thing for Mill. There is a contrast for him between the self-subsisting power of the individual to resist and the power of the state to coerce.

We would now tackle the problem of picking out which properties could be aggregated and which could not by invoking the idea of an emergent property, and using examples of the difference between the properties of constituent cells and the organs of which they are components to illustrate the point. What could one make of the idea of a group happiness as an emergent property of the individual activities of the members of a society?

4.2 Collectives and Mass Societies

Running more or less parallel to Mill's individualistic utilitarianism, Hegelian idealism presented a very different picture. History is a collective progress since the world spirit, *Geist*, is embodied in the human race as material beings, more or less as individual minds are embodied in individual bodies. All relations between all beings whatsoever are internal and constitutive. As the realization of the one World Spirit human beings have a unity as a kind of supra-individual.

4.3 Marx and Durkheim

Marx's sociology, the structure of which is drawn more or less directly from Hegel's metaphysics, pitted class against class in a dialectic of contradictory interests (Marx 1867 [1965]). The this-ness of individuals gave way to the individual class member as relevant to sociology only as the vehicle for the necessary attributes of class membership. Borrowing from Hegelian Idealism, Marxian 'Man' was the nexus of internal relations, the relations of the social necessities of certain means of the production of the material necessities of life. However, Marx's two main social classes were taxonomic, in the sense I have defined above. To be a member of the working class or of the bourgeoisie was to exemplify a certain set of attributes, though those attributes were possessed by virtue of standing in certain internal relations to other people.

Durkheim's plea for sociology as the science of 'social facts' comes close to Marx's unpsychological sociology insofar as Durkheim's social facts are attributes of groups. In his *Elementary Forms of the Religious Life* (Durkheim 1895 [1982]), the ceremonies of the Native Australians are to be interpreted as presentations of collective facts about the local social order through participation in which loyalty and commitment to that order are renewed. Again the collective is thought of in aggregational rather than structural terms.

5. What Sort of Science is Sociology?

5.1 Windelband and the Nomothetic/Idiographic Distinction

The individual/collective issue appears not only in metaphysics and social theory, but in methodology as well. That there are two kinds of human sciences was the claim of Windelband (1894 [1998]). There are studies of individuals as such, an idiographic science. There are studies of people in groups or collectives, a nomothetic science. The analogue of a law of nature had no place in idiographic science while, perhaps with the help of statistical analysis, it could have a place in nomothetic science. This distinction is related to but not identical to that between two kinds of empirical enquiry. There is the intensive design, the study of a group by the study of a typical member and there is the extensive design, the aggregation of information about the members of a population into a statement about the population as a whole. The results of studies based on the extensive design will tend to ignore emergent properties, since they are not aggregates of the properties of individuals.

5.2 The Reduction of Collectives to Individuals

In the twentieth century the issue has surfaced again, this time in the guise of the problem of explanation. Are the collective aspects of human societies to be accounted for as just patterns of the unintended consequences of individual actions according to individual intentions? Or are they mass social effects of mass social causes? If we take the former view, explanations of mass phenomena must be sought not in other mass phenomena, but in the intentions of individual actors as they ramify through a population and interact with the intentions of other actors. This proposal has been called 'methodological individualism.' It has been revived by Popper (1961) and others. The idea is an old one. Adam Smith (1776) thought that individual choices in the marketplace 'added up' to the illusion of the 'hidden hand.' It had an important place in Menger's (1883 [1981]) advocacy of microeconomics, the study of the grounds of individual acts of choice as opposed to statistical analyses of vast masses of data. It cannot be a wholly individualistic story since some reference to structure seems to be essential to understand how individual actions could aggregate into something that could be perceived as orderly.

Adam Smith (1776) and Carl Menger (1883 [1981]) did not feel the need to offer arguments for methodological individualism, since each thought he had shown its methodological advantages. In the 1950s a more recognizably philosophical debate began, forcing the positions to be more clearly differentiated and the arguments pro and con more carefully formulated. Three lines of argument in favor of collectivism can be discerned. A metaphysical argument emerged to the effect that individuals are simply products of social processes (Lukes 1973). Then there was an epistemological argument that the tactic of studying only individuals will not allow the recovery of properties of groups. Finally there was a conceptual argument that collectivist concepts such as 'nation' or 'army' cannot be replaced by individualistic ones like 'citizen' and 'soldier' with preservation of meaning.

I think it is fair to say that the issue has not been resolved to anyone's entire satisfaction. The degree to which one sees the ideal form for a society along the polar contrast between collectivist and individualist is not just a question of one's political or religious affiliations. It is also a matter of how far one sees oneself as a product of ready-made social orders, and how far one thinks of oneself as a crusader against the tyranny of the majority.

See also: Bentham, Jeremy (1748–1832); Consequentialism Including Utilitarianism; Control: Social; Durkheim, Emile (1858–1917); Freedom: Political; Hobbes, Thomas (1588–1679); Individualism versus Collectivism: Philosophical Aspects; Influence: Social; Locke, John (1632–1704); Marx, Karl (1818–89); Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects; Mill, John Stuart (1806–73); Norms; Reduction, Varieties of; Rousseau, Jean-Jacques (1712–78); Sovereignty: Political; State and Society

Bibliography

Bentham J 1789 [1948] *Introduction to the Principles of Morals and Legislation*. Blackwell, Oxford, UK

Durkheim E 1895 [1982] *The Rules of Sociological Method*. Macmillan, Basingstoke, UK

Giddens A 1984 *The Constitution of Society*. Polity Press, Cambridge, UK

Hobbes T 1651 [1946] *Leviathan*. Blackwell, Oxford, UK

Lukes S 1973 *Individualism*. Blackwell, Oxford, UK

Marx K 1867 [1965] *Capital*. Lawrence & Wishart, London

Menger C 1883 [1981] *Problems of Economics and Sociology*. New York University Press, New York

Mill J S 1959 *On Liberty*. Longman Green, London

Ockham W 1990 *Ockham: Philosophical Writings* [ed. Boehner P] Hackett, Indianapolis, IN

Popper K R 1961 *The Poverty of Historicism*. Routledge and Kegan Paul, London

Rousseau J-J 1764 [1968] *The Social Contract*. Trans. M. Cranston. Penguin Books, Baltimore, MD

Smith A 1776 [1976] *Inquiry into the Nature and Causes of the Wealth of Nations*. Clarendon Press, Oxford, UK

Vico G 1725 [1948] *Scientia Nova* [*The New Science*]. Trans. T. G. Bergin and M. H. Fisch. Cornell University Press, Ithaca, NY

Windelband W 1894 [1998] Geschichte und Naturwissenschaft (History and natural science). Trans. J. T. Lamiell. *Theory and Psychology* **8**(1): 5–22

R. Harré

Individualism versus Collectivism: Philosophical Aspects

The issue between individualism and collectivism, as it is taken in this article, is the ontological issue as to whether the regularities of social life, in particular the regularities that have come to light in social science, cast doubt on the status of human beings as more or less autonomous intentional subjects. This issue should be distinguished from the methodological issue as to whether the explanation of social phenomena ideally should be in terms of the properties of individuals, as it should be distinguished from the moral issue as to whether social institutions should always be assessed from the point of view of their impact on individuals. Emile Durkheim, the great French sociologist, was the most influential defender of ontological collectivism, according to which individuals, as seen from the perspective of social science,

lack the autonomy which they ascribe to themselves in common sense. Just as astronomy taught us that the sun does not literally cross the sky, and physics that solid objects are not solid in any literal sense, so the collectivist idea is that the social sciences teach us that individuals are not what they seem; they are not the self-determining centers of decision-making that they take themselves to be. But hard collectivism in this sense looks implausible, as we shall see; only some soft versions of the doctrine—versions that sideline or downgrade intentional psychology without undermining it—look to have any serious support (Pettit 1993).

1. *The Emergence of Collectivism*

The sociological way of thinking that brought Durkheim to this conclusion had its origins in the rise of social statistics in the early part of the nineteenth century. Responsive to electoral and utilitarian pressures, the governments of the newly industrializing countries, sponsored the statistical mapping of a great variety of phenomena. 'It was necessary to count men and women and to measure not so much their happiness as their unhappiness: their morality; their criminality; their prostitution; their divorces; their hygiene; their rate of conviction in the courts. With the advent of laws of statistics one would find the laws of love or, if not that, at least the regularities about deviancy' (Hacking 1981, 1990).

This statistical mapping of large-scale society, this avalanche of numbers, had an important intellectual impact, because it revealed that social aggregates display surprising constancies, constancies that are often invariant over changes in the individuals, and in the mentalities of the individuals, involved. The statisticians constructed indices of divorce, suicide, crime, sickness, prostitution, and a variety of other phenomena. It soon became clear from their work that there were constancies over time in many such indices, and there were constancies in the correlation of different indices with one another. At the beginning of the nineteenth century it would have been common to regard such statistical patterns as epiphenomena of intentional action which we should expect to be as variable as we seem to find one another. But from the earliest days of statistical mapping, it was obvious that the patterns were not as variable as that; it began to seem that they had a life of their own.

This revelation of unexpected pattern encouraged a sort of statistical determinism, one associated in the beginning with French writers like A. M. Guerry and Adolphe Quetelet. Already by 1832, Guerry could write: 'We are forced to recognise that the facts of the moral order are subject, like those of the physical order, to invariable laws' (Hacking 1991). Such determinism came to be described as an astronomical conception of society, a conception in which people were driven by forces akin to a cosmic force like gravity; the fact that Quetelet was an astronomer lent credence to this characterization of the view (Hacking 1991). The astronomical conception soon spread abroad, receiving a powerful impetus in the English-speaking world, and indeed more broadly, with the publication in 1857 of T. H. Buckle's *History of Civilisation in England.*

The astronomical conception of society received perhaps its most dramatic expression in Tolstoy's *War and Peace*, which was written between 1863 and 1869. One passage will serve to give the tenor of the view 'Ever since the first person said and proved that the number of births or crimes is subject to mathematical laws, that certain geographical and politico-economical laws determine this or that form of government, that certain relations of the population to the soil lead to migrations of peoples—from that moment the foundations on which history was built were destroyed in their essence. By disproving those new laws, the old view of history might have been retained. But without disproving them, it would seem impossible to continue studying historical events, merely as the arbitrary product of the freewill of individual men. For if a certain type of government is established, or a certain movement of peoples takes place in consequence of certain geographical, ethnographical, or economic conditions, the freewill of those persons who are described to us as setting up that type of government or leading that movement cannot be regarded as the cause' (Tolstoy 1972).

The avalanche of numbers and the statistical determinism associated with it led to the development of the new science of sociology. True, the person who invented the term 'sociology', Auguste Comte, was averse to seeing the discipline as a merely statistical enterprise. But the rise of statistical mapping had a crucial impact on Emile Durkheim, who was responsible for establishing sociology as a respectable science. The first empirically based study that he undertook was an investigation of suicide—published in 1897—and this reveals both how deeply he was impressed by the statistical constancies of suicide and how naturally he moved towards the conclusion that those constancies had a life of their own, a life independent of the whirl and flow of individual choice.'Each society is predisposed', he remarked, 'to contribute a definite quota of voluntary deaths' (Durkheim 1951).

Durkheim accepted that aggregate constancies have a life of their own, remaining surprisingly invariant over relevant changes at the individual, intentional level. He postulated aggregate, social regularities to explain those constancies. An aggregate regularity may explain a constancy by subsuming it, as when a constancy in the correlation between two indices—say, the divorce rate and the suicide rate—is taken to reveal a regularity. Or an aggregate regularity may explain a distinct constancy, revealing a relationship between it and some independent factor: say, a relationship

between an aggregate index or correlation of indices on the one side and a background factor like increased population density, an established current of opinion, growing urbanization, or even the fact that the factor in question—say, the correlation—is socially functional in some way. Durkheim gives expression to the sort of regularities that he saw at the origin of statistical constancies, when he writes in the study on suicide: 'the relations of suicide to certain states of social environment are as direct and constant as its relations to facts of a biological and physical character were seen to be uncertain and ambiguous. Here at last we are face to face with real laws' (Durkheim 1951).

In order to understand the sorts of regularities that Durkheim envisaged, here are some examples. They are not chosen on the grounds that Durkheim actually endorsed them or even considered them—he did not—but because they rPepresent a fair spread of the sort of sociological doctrine with which he was associated.

(a) Increased unemployment leads to a rise in crime.

(b) Urbanization leads to a decline in religious practice.

(c) Policies for increasing employment cause inflation.

(d) Economic stratification comes about in order to secure personnel for crucial positions.

(e) Capitalism is stable when it is optimal for developing productive power.

(f) States act internationally in their best economic interest.

(g) States decline in political influence as they fall behind in economic capacity.

(h) Companies maximize expected returns.

(i) The protestant ethic facilitates the rise of capitalism.

(j) Capitalism generates a breakdown in community values.

Regularities of these kinds are social-structural in character. They are social regularities, because they each demand that a number of individuals display certain intentional responses: they require that individuals commit crime in a certain proportion, give up attending church, and so on. And they are structural, because they are discontinuous with intentional regularities. It is not supposed that unemployment leads to crime, or urbanization to religious decline, because it causally impacts in some intentionally intelligible way on the antecedents of intentional regularities; it is not supposed, for example, that it is because people become aware of unemployment or urbanization that they produce the behavior required for the consequent.

Durkheim and most of those who followed him took the sorts of regularities illustrated to give the lie to intentional psychology. He saw them as pointing us towards constraints which undermine or limit human autonomy. He saw them as vindicating, in essence, something like the so-called astronomical conception of society. The constraining factors to which social-structural regularities pointed, in Durkheim's view, range from the morphological features of a society, like the density of its population, to the norms or rules institutionalized there, to the currents of opinion and the enthusiasms which take over from time to time (Durkheim 1938, Lukes 1973). He described all of these as social facts and argued that social facts constrain intentional agents from without. 'Currents of opinion, with an intensity varying according to time and place, impel certain groups either to more marriages, for example, or to more suicides, or to a higher or lower birth rate, etc … A social fact is to be recognised by the power of external coercion which it exercises or is capable of exercising over individuals' (Durkheim 1938).

In giving social facts this constraining role, Durkheim was explicitly challenging the individualist view—a view he naturally presents as an extreme position—that individuals enjoy intentional autonomy. Thus, he wrote of social facts that: 'when we define them with this word 'constraint,' we risk shocking the zealous partisans of absolute individualism. For those who profess the complete autonomy of the individual, man's dignity is diminished whenever he is made to feel that he is not completely self-determinant' (Durkheim 1938).

2. *The Implausibility of Hard Collectivism*

There are two different ways in which social-structural regularities might prove to compromise the intentional autonomy of individuals in the sort of measure that Durkheim apparently had in mind: in the measure, as we may take it, that would give the lie to individualism and support a position truly deserving of the title of 'collectivism.' Individuals will be autonomous in the relevant sense so far as they form their beliefs and desires—their intentional attitudes—and act on those beliefs and desires in a manner that satisfies the general expectations that we hold in relation to a rational or approximately rational agent. One way in which social regularities would compromise individual autonomy is by overriding those intentional regularities, the other by outflanking them (Pettit 1993).

The social regularities would override the intentional if they sometimes conflicted with them, predicting different outcomes for the same situation, and if they proved in those cases to be the more reliable laws. The social regularities would outflank the intentional if it transpired that things are organised at the intentional level—intentional agents and attitudes are selected—so that the social regularities are assured of being satisfied. In the overriding case, the social regularities would predetermine what happened; in the outflanking case, they would predestine it.

The social sciences have not produced results robust enough to support collectivism in the literal or hard sense associated with the overriding or outflanking

thesis. Hume argued that the testimonial evidence that a natural miracle occurred—that a law of nature was suspended—is almost always going to be outweighed by the day-to-day evidence of the reliability of the laws of nature. We might argue in parallel that the social scientific evidence in favor of an intentional miracle—in favor of the overriding of intentional regularities in otherwise normal agents—is almost always going to be outweighed by the day-to-day evidence of the reliability of intentional laws.

As it is with the overriding thesis, so it is with the outflanking one. This thesis would hold that in cases where human beings act so as to ensure that a social regularity obtains, that is, no accident: were they going to have attitudes that would lead them not to support that regularity, then they would not have been there; they would not have been selected. The case envisaged parallels that of natural selection. Assume that mammals have been selected so as to favor the regularity whereby they can digest certain foods. That means that were certain creatures going to have a system that did not support that regularity, then they and their ancestors would not have been there; they would not have been selected. But while there is evidence of certain forms of social selection, there is no evidence for the powerful sort of selection—parallel to natural selection—that would be needed to support the outflanking alleged.

It ought not to cause any embarrassment for social scientists, however, that the overriding and outflanking theses—and the hard collectivism they would entail—are not borne out empirically. For if we recur to the examples given earlier, we can see that social laws of the sort that impressed Durkheim can be intereting and significant, quite consistently with not entailing hard collectivism. That those laws obtain is fixed in part by the fact that human beings are intentional systems who obey standard intentional laws, and in part by contingent facts about the attitudes of those individuals. But that does not mean that just knowing about individuals and their intentional psychology will give us knowledge of those laws. Not at all. It may take considerable data-collection and considerable data-analysis before it begins to become plausible, say, that increased unemployment leads to a rise in crime or increased urbanization to a decline in religious practice. Such a regularity will be plausible in good part, because we can see how it fits with our knowledge of intentional psychology: we can see, for example, how unemployment might lead to crime by increasing both the opportunity and the incentive for criminal acts. But that does not mean that we can derive the law from armchair reflection on our psychology.

Not only will social laws be interesting in this sense, even if they do not entail hard collectivism. They can also be of explanatory importance. Explanatory or methodological individualists hold that explanations which go to the level of individual psychology are to be preferred in social science to higher-level, structural or aggregate explanations. Elster assumes that just the fact of getting nearer the detail of production makes such an explanation intrinsically superior: 'a more detailed explanation is also an end in itself' (Elster 1985). Taylor (1988) agrees: 'A good explanation should be, amongst other things, as fine-grained as possible.' But the rejection of hard collectivism does not necessarily entail the adoption of such explanatory individualism.

Assume that explanation usually involves giving information on causal process. The reason that social laws are explanatorily important is that reference to such a law may give information on the causal process at its origin that is not available just from the story about the detailed psychological profiles of the individuals involved. Thus, we might be told of the individual psychological aetiologies that produced the level of crime in a given year, and be told that that level is higher than previously, without being given the information provided by reference to the law, as we may assume it to be, relating increased unemployment to increased crime. Indeed, we might not even have noticed that most of the new or more frequent offenders were unemployed. The reference to that law provides the information that the increased crime would materialize, not only in the presence of the precise psychological profiles of actual offenders, but under any condition where there is an increase of unemployment. It serves an explanatory purpose that is neglected in the more detailed, individual-level stories (Jackson and Pettit 1992, Pettit 1993).

3. *The Plausibility of Soft Collectivism*

But there is also another line of thought that ought to reconcile many social scientists to the failure of hard collectivism. This is that the argument just given is consistent with the truth of some claims that have a collectivistic ring but that do not involve either the overriding or the outflanking of intentional law. Any one of these claims might be thought to support a soft collectivism: that is, a position that is collectivistic so far as it makes intentional psychology somewhat less important than it may have seemed but that is consistent, nevertheless, with the soundness of intentional psychology. I shall mention five such claims, three or four of which are quite plausible theses. The claims each address a somewhat different question from that to which hard collectivism provides an answer.

The first question which it is important to distinguish from the hard collectivism issue is the question that divides those who are well described as atomists and holists: the question as to whether the individual agent noncausally depends on relations with other people for the appearance of distinctive human capacities (Pettit 1993). Holists maintain that individual

people are dependent on one another for the possession of distinctive human capacities: say, for the capacity to think. But this belief need not lead them to hold that people cannot therefore have the autonomy of subjects who act more or less rationally in the light of more or less rationally held beliefs and desires. Intentional psychology is silent on the question of whether a capacity like that of thought requires communal resources. The holist assertion that thought does require those resources cannot, therefore, undermine that psychology.

The second question, that must be distinguished from the hard collectivism issue, has come to the fore in recent thought. Let us agree that people intentionally or nonintentionally produce—and reproduce—a variety of social entities by the patterns manifested in their attitudes and actions. They constitute social groups and instrumentalities and they constitute the social properties that attach to certain individuals: the status enjoyed by this person, the power enjoyed by that. But here now is a question raised by such entities. Do the entities exercise a reciprocal influence on individual agents? Do they serve to determine, and indeed rationally determine, the opportunities, the motives and the perceptions of those agents, and thereby to shape their further intentional responses? There are some who think, or who think certain individualists have to think, that the answer to this question is, no (Ruben 1985). Those who defend an affirmative answer, therefore, will appear in the guise of collectivists. And this means, in effect, that most contemporary social theorists will be given a collectivist cast (Giddens 1976, Coleman 1990).

The affirmative answer to this question is surely the correct one; social entities exercise a reciprocal influence on agents, at least so far as agents come to form beliefs and desires that involve those entities in their contents. But admitting the reciprocal influence of social entities in this manner does not in any way threaten the standing of our intentional psychology and does not involve a commitment to hard collectivism. I may be influenced by the objects in my social world, as I am influenced by the objects in my physical milieu, coming to form beliefs and desires in respect of them, but this does nothing to compromise my capacity as an intentional subject.

A third proposition that is collectivistic in spirit but not damaging to intentional psychology is a thesis of revisability. It is that thesis that the regularities identified in social science often challenge intentional psychology in the sense of pointing us towards unsuspected disturbing factors. Like the holistic and reciprocity doctrines, this revisability thesis is perfectly plausible. Consider the sort of regularity which much of Goffman's work is given to supporting: that people, in his language, are extremely loath to break frame; that they readily establish a set of shared assumptions about what is say-able and do-able in any interaction and that they generally conform to those assumptions, even at the cost of considerable frustration of their other desires or ideals (Goffman 1975). Anyone who has read the work of Goffman is likely to conclude, quite reasonably, that here we have been alerted to a factor that is often going to disturb people's intentional performance. The work of such social scientists bears out the truth of the revisability thesis ascribed to collectivists, that the identification of social regularities is likely to point us to conditions that disturb the operation or revelation of people's intentional processes and in that sense that it is likely to affect our intentional psychology.

But though the revisability thesis is plausible, it does not serve to displace intentional psychology in a collectivist manner. The frame law used in illustration leaves intentional psychology fundamentally intact. Intentional psychology is an open-ended practice; it remains possible at any stage for its practitioners—who are of course also its subjects—to revise their conception of favorable conditions. If social science can interact with intentional psychology to force such revisions, in particular revisions that can be internalised by participants in their practice, then that does not involve any displacement of intentional psychology; on the contrary, it serves as a way of enriching the practice. Consider the effect of the discovery in optics that a rotating disc with a certain black-white pattern looks to be colored in different ways at different speeds. This does not undermine the intentional-psychological story of color perception but simply points us to a perturbing factor—rotation—that generally goes unnoticed but that is of a kind with factors—say, to do with lighting conditions—that are commonly recognized. A similar gloss will apply to the deliverances of social science, under the revisability thesis.

So much for the holism, the reciprocity, and the revisability claims which should be distinguished from hard collectivism. The fourth claim that should equally be distinguished from it is a thesis about the dispensability of individual contributions to social life, the fifth a claim about the inevitability of those contributions.

The dispensability thesis would say that while any socially significant event is going to come about because of the actions of this or that individual or set of individuals, still no individual makes an indispensable contribution. Had that individual not existed or not acted appropriately, still the phenomenon in question would have occurred. It is either the case (a) that the individual's contribution was not a necessary part of the total cause of the phenomenon or (b) that in the absence of that individual some other agent would have taken their place.

Durkheim was motivated by the existence of certain social constancies—the suicide rate, the divorce rate, and the like—to defend a thesis of type (a). The fact that these constancies survive variation in the individuals and the mentalities of the individuals involved

shows that no one individual contribution is a necessary part of the relevant total cause. But in defending this sort of dispensability claim, Durkheim was doing nothing to displace intentional psychology, so that the claim could be defended equally well by an individualist. The social constancies in which he was interested are characterized at such a high level that the absence of one individual contribution would not make a difference to the obtaining of the constancy. An artefact of aggregation and abstraction accounts for the dispensability of individual contributions, not any radical questioning of intentional psychology.

What of the other sort of dispensability thesis: the type (b) thesis that even with social phenomena which do actually depend on particular individuals—say, something like a political revolution—that phenomenon would still have occurred even in the absence of a relevant contributing party; it would still have occurred, because there would have been someone else available to take the vacated place? This type of claim is not associated with Durkheim particularly but is to be found in other, more historically oriented traditions of social science. The main source is the Marxist tradition and here Engels is quoted as making a telling remark about Napoleon Bonaparte 'That such a man, and precisely this man, arises at a determined period and in a given country is naturally pure chance. But, lacking Napoleon, another man would have filled his place' (Sartre 1963). Is this sort of dispensability thesis inimical to our intentional psychology? Again, I think not, though I doubt that it is generally true (James 1984). All that is required for it to be the case that an individual's contribution is dispensable in this sense is that a number of individuals have the capacity and motivation required to make the sort of contribution involved but that at most one individual can have the opportunity. To think that this was true of all individually necessary contributions to socially significant phenomena might be dogmatic, but it would not involve rejecting intentional psychology.

The fifth and last sort of thesis that should be distinguished from hard collectivism fits better with less empirical parts of Durkheim's work than his study of suicide: for example, with his analysis of the division of labor characteristic of our sort of society. I describe it as an inevitability thesis, as distinct from a dispensability thesis. What it says is that various features of social life limit the opportunities available to individual agents—they are structural constraints on feasible options—in such a way that it is inevitable that agents will act so as to sustain certain social constancies: 'In an extreme version this would mean that the constraints jointly have the effect of cutting down the feasible set to a single point; in a weaker and more plausible version that the constraints define a set which is so small that the formal freedom of choice within the set does not amount to much' (Elster 1979). This could be what collectivists have in mind, at least some of the time, when they denigrate the place of the intentional subject. Certainly it is a thesis that many social scientists defend.

The inevitablility thesis, even in the weaker version, is generally unsound. But that is beside the point in our present discussion. More important, it does no more than the dispensability thesis to undermine intentional psychology. It just might be the case, as the inevitability thesis has it, that our opportunities are tailored to the production of certain social constancies: that we are so limited in our range of options that it is inevitable we will act so as to sustain such constancies. But even if it were true, that would cast no serious doubt on the soundness of intentional psychology or on our status as autonomous subjects. It would not mean that we individual subjects were compromised by social laws in a manner that undermined our view of ourselves as intentional subjects. Our intentional psychology represents us as more or less rational subjects but it says nothing on the range of opportunities that are actually available to us as agents.

It may come as a surprise that hard collectivism is not entailed by the inevitability thesis or the dispensability thesis. Many self-described individualists would certainly want to deny the general truth of such propositions. But even if that is so, the matters in question do not go to the heart of the issue between individualism and collectivism: they do not have an intimate bearing on the status of our intentional psychology. They support a soft sort of collectivism but not the hard kind that would provide a really significant challenge to our self-image.

See also: Atomism and Holism: Philosophical Aspects; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Collective Beliefs: Sociological Explanation; Collectivism: Cultural Concerns; Colonialism, Anthropology of; Economics, Philosophy of; Individual/Society: History of the Concept; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects

Bibliography

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Durkheim E 1938 *The Rules of Sociological Method*. Free Press, New York

Durkheim É 1951 *Suicide*. Free Press of Glencoe, New York

Elster J 1979 *Ulysses and The Sirens*. Cambridge University Press, Cambridge, UK

Elster J 1985 *Making Sense of Marx*. Cambridge University Press, Cambridge, UK

Giddens A 1976 *New Rules of Sociological Method*. Hutchinson, London

Goffman E 1975 *Frame Analysis: An Essay on the Organization of Experience*. Penguin, Harmondsworth, UK

Hacking I 1981 The taming of chance by an army of statistics. *The Times Higher Education Supplement*: 12–13
Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK
Jackson F, Pettit P 1992 In defence of explanatory ecumenism. *Economics and Philosophy* **8**: 1–21
James S 1984 *The Content of Social Explanation*. Cambridge University Press, Cambridge, MA
Lukes S 1973 *Émile Durkheim: His Life and Work: A Historical and Critical Study*. Penguin, Harmondsworth, UK
Pettit P 1993 *The Common Mind: An Essay on Psychology, Society and Politics*. Oxford University Press, New York
Ruben D-H 1985 *The Metaphysics of the Social World*. Routledge & Kegan Paul, London
Sartre J-P 1963 *The Problem of Method*. Methuen, London
Taylor M 1988 *Rationality and Collective Action. Rationality and Revolution*. Cambridge University Press, Cambridge, UK
Tolstoy L 1972 *War and Peace*. Heinemann, London

P. Pettit

Industrial and Organizational Psychology: Cross-cultural

Cross-cultural industrial and organizational psychology is the systematic study of behavior and experience in work and organizational situations that occur in different cultures, that are influenced by culture, or result in changes in cultures.

1. Domain

Specifically, cross-cultural industrial and organizational psychology examines work-related phenomena occurring in different kinds of organizations in more than one culture. It attempts to measure aspects of culture and relate them to each work-related, managerial, or organizational phenomenon. It examines cultural influence on work and health, job satisfaction, productivity, decision making, turnover, motivation, labor–management relations, multicultural teams, participation and industrial democracy, preparation of individuals for work in another culture, cultural diversity, and the fit of organizational and national cultures.

2. History

2.1 Early History of Industrial and Organizational Psychology

This field emerged in 1920. In that year two significant events took place. First, Garth (1920) published a paper that measured cultural differences in work behavior. It was based on the number of items of addition that were completed by samples of European, African, and Native Americans (of course, at that time they were called whites, Negroes, and Indians!). The point was to determine the 'work curves' of these samples. Fatigue reduced the number of additions completed at later points in time. The Native Americans were least susceptible to fatigue, while the African Americans were the most susceptible to fatigue, with the European Americans in between. The cultural differences were minor, and no controls for obvious rival hypotheses were presented. However, the author did mention that perhaps the way the samples were selected did not make them equivalent. Second, the Swiss psychologist Claparede established the International Association of Applied Psychology which focused on international industrial and organizational topics, and now holds an international congress every four years. At these congresses applied psychology data collected in different parts of the world are presented and discussed.

In addition, Frost (1920) published a paper which suggested that psychology can be valuable in studying the adjustment of immigrants to the American industrial environment. 'If the Swede, the Pole, the Italian, the Greek, are to be taught the advantages of industrial democracy, the lesson must be built upon a background of proper racial psychology' (p. 21). Soon after, Garth et al. (1933) presented a study of work behaviors that included Japanese and American data, and reported no cultural differences.

Wilpert (2000) traced the past, present, and future of applied psychology, and identified four periods. The first (1922–30) focused on health and satisfaction from work and on epistemological perspectives (e.g., should we use only experiments or also systematic observations of real life events?). The second (1931–7) reflected concern for international solidarity in view of imminent political threats. The third (1949–55) widened the concern for social and political problems and focused on the problem of living together as individuals, families, communities, and nations. The last (1964–) reflected global concerns, spiritual outlooks, and included some methodological backtracking. He noted that societal changes are reflected in the way applied psychology is evolving. Thus, the increased female labor market participation, increased life expectancy, increased labor migration (especially from less economically developed to more economically developed countries), and changes in values are inevitably changing applied psychology.

2.2 The Field Becomes Established

Cross-cultural psychology did not emerge as a self-conscious discipline separate from anthropology and

psychology and yet linked to both until about 1967. An important turning point was a conference held in Ibadan, Nigeria, which exposed psychologists to the need to avoid 'intellectual colonialism' (collecting data in different countries without sensitivity to the local culture, i.e. without the help from or giving credit to local psychologists). By the early 1970s, the use of translated Western instruments to collect data in non-Western cultures was discredited among sophisticated researchers, yet widely done because of the impracticalities of careful cross-cultural work.

The leading cross-cultural researchers argued that ideally, the instruments that are used in this field should include two aspects (or factors after factor analysis): one aspect should go to the heart of the way the culture operates in the specific domain, and consist of emic, i.e., culture-specific items. The other aspect should include etic items (i.e., items that have universal meaning). The first aspect can provide an optimal description of a phenomenon in a specific culture; the second aspect can provide comparisons across cultures.

Indeed, one of the purposes of this field is to discover what aspects of work and organizational behavior are universal (etic) and what aspects are culture-specific (emic). Humans have a lot in common. We share 99 percent of our DNA with some of the higher primates, so in many ways we are the same. Yet different cultures emerge because of different ecologies (resources, geography) and different intergroup influences (cultural diffusion, historical influences, such as wars). For example, if we were to study which aspects of different kinds of work are most valued by humans, we are likely to find more enjoyment of fishing in an ecology that has fish than in an ecology that does not. In short, liking fish has meaning only in some cultures, and is not a universal value. Other concepts (e.g., susceptibility to monotony) may be universal, and thus may permit cross-cultural comparisons.

A metaphor can make the point clearer. Suppose we are comparing apples and oranges. We can compare them on etic dimensions such as price, weight, and size, but a full description of the fruit based on such dimensions is inadequate. To capture the meaning of 'apple' one needs to measure apple flavor, an emic dimension. In short, we need to use both emic and etic concepts and dimensions for a complete description and cross-cultural examination of a phenomenon.

Triandis (1972, pp. 35–58) presented an extensive discussion of both the method and the ethics of cross-cultural research, and examples (pp. 89–263) of the use of the various methods. He advocated the inclusion of both emic and etic items, and the local standardization of scales in cross-cultural investigations. He examined the sampling of cultures, individuals, and items. He urged checks of the reliability and construct validity of the scales in each culture. He advocated controls of response sets and other rival hypotheses such as the cross-cultural equivalence of (a) the understanding of the experimental task, (b) familiarity with the method, (c) motivation, (d) response style, (e) test conditions, (f) perception of the experimenter, and (g) social desirability of the responses. He examined the role of anonymity, experimenter bias, and problems of interpretation of the data from different cultural perspectives. The chapter also discussed how cross-cultural research should be organized to avoid intellectual imperialism. Tapp, Kelman, Triandis, Wrightsman, and Coelho (1974) presented a code of ethics for such research. It advocated leaving something of value in each of the cultures studied. However, the cross-cultural community considered this code too demanding and restrictive, thus it was not adopted by the International Association of Cross-Cultural Psychology.

The avoidance of ethnocentrism, by including the perspectives of each culture under study in the research, the use of multimethod approaches (to control for the interaction between culture and method), the test of hypotheses with laboratory, survey, and ethnographic data, the testing of rival hypotheses, the checks of the reliability and validity of each method in each culture were advocated by several researchers in the 1970s, and gradually began to be used in cross-cultural research.

Another product of the Ibadan conference was the establishment of the *Cross-Cultural Newsletter* (1967), first edited by Triandis, which eventually became the *Cross-Cultural Bulletin*. In the late 1960s several journals (*International Journal of Psychology, Journal of Social Psychology*) announced that they welcomed cross-cultural data. The Directory of individuals interested in cross-cultural research was published (Berry 1968). The Human Relations Area Files established a journal which was eventually (1992) called *Cross-Cultural Research*. It is now the official organ of the Society for Cross-Cultural Research. The *Journal of Cross-Cultural Psychology* was established by Walter Lonner in 1970. It became the official journal of the International Association of Cross-Cultural Psychology (IACCP). The *Annual Review of Psychology* started chapters reviewing cross-cultural work in 1973. *Culture and Psychology* was first published in 1995.

3. Content

3.1 Handbooks Summarize the Field

The first handbook of industrial psychology was by May Smith (1944), and included no chapters with cross-cultural material. Dunnette in the *Handbook of Industrial and Organizational Psychology* (1976, reprinted in 1983) included an excellent chapter by Barrett and Bass (1976) that covered theoretical and methodological problems. This chapter examined

motivation, values, job satisfaction, managerial goals, cultural differences in attitudes toward compensation, motives and economic development. It also explored the match of organizational and cultural values. An important focus was on management and supervision (including management styles, authoritarianism), and preferences for different kinds of supervisors in different cultures. An important topic was participation of workers in job design (J. R. P. French). As an illustration of a finding of cross-cultural industrial psychology, French and others found that in the United States participation of employees in work design increases their productivity. But this was not the case in Puerto Rico or Norway. Puerto Rico is a more hierarchical culture than the United States, so that workers found it embarrassing when their supervisors asked for their ideas about how the work should be done. Norway had norms about how decisions are to be divided between labor unions and management, and asking the workers how the work should be done did not conform to those norms.

Barrett and Bass also reviewed various studies that included clusterings of countries (e.g., European countries are closer to each other in managerial attitudes than is Japan). It did an excellent job of examining factors in employee selection, and factors that increase the success of expatriates such as cross-cultural training.

Dubin's handbook (1976) included a section on work in different social systems, with chapters on labor movements (Landsberger), industrial relations in Japan (Karsh), work and management in Poland (Matejko), industrialization in Israel (Weinshall), and the self-management of Yugoslav enterprises (Jerovsek).

Drenth, Thierry, Willems, and de Wolff in the *Handbook of Work and Organizational Psychology* (1984) included one chapter on cross-cultural perspectives (Drenth and Groenendijk 1984). Bass and Drenth in *Advances in Organizational Psychology* (1987) edited chapters that included, among other parts, reports with cross-cultural data from India and Spain (Filella), decision making (Heller and Misumi), participation and industrial democracy (Wilpert), and comparative management (Ronen and Kumar).

The second edition of the Dunnette *Handbook of Industrial and Organizational Psychology* included a whole volume (Triandis, Dunnette, and Hough 1994) containing international material. It included a chapter by Triandis that provided an overview of the cross-cultural industrial and organizational psychology field, and discussed dimensions of cultural variation and the definitions of culture. This volume examined employee selection across cultures (also discussed by Levy-Leboyer), organizational cultures, motivation in different cultures (also discussed by Ronen), employee control of the organization, job design (also discussed by Hesketh and Bochner), organizational development, and conflict resolution.

This volume of the handbook also had chapters on action theory (Frese and Zapf), time and work behavior (Thierry), leadership (Graen), and aging and work (Forteza and Prieto, as well as Warr). In that volume also, Erez presented a model of industrial and organizational cross-cultural psychology. Culture-specific chapters included chapters by Kashima and Callan (the Japanese work group), Redding, Norman, and Schlander (the nature of individual attachment to the organization: a review of East Asian variations), Wang (culture, economic reform, and the role of industrial and organizational psychology in China), and Sinha (culture embeddedness and the developmental role of industrial organizations in India). Finally, there was a chapter on dealing with cultural diversity in the USA (Triandis, Kurowski, and Gelfand).

The *Handbook of Cross-Cultural Psychology* (H. C. Triandis, General Editor) was published in 1980–1. Volume 5 of that handbook included an organizational psychology chapter by A. S. Tannenbaum. The three-volume second edition of that handbook (edited by J. W. Berry) appeared in 1997. The third volume of this edition examined applications and included a chapter by Hui and Luk (industrial/organizational psychology). That volume also included some other chapters that are relevant to cross-cultural industrial and organizational psychology, such as those of Kagitcibasi (individualism and collectivism), Gudykunst and Bond (intergroup relations), Best and Williams (sex, gender, and culture), Berry and Sam (acculturation and adaptation), and Brislin and Horvath (cross-cultural training and multicultural education).

3.2 Milestones in the History of the Field

Harbison and Myers in *Management in the Industrial World* (1959) and Haire, Ghiselli, and Porter in *Managerial Thinking: An International Study* (1966) are viewed by many industrial psychologists as the first milestones of cross-cultural industrial and organizational psychology. Listed sequentially across time, Roberts in the *Psychological Bulletin* (1970) presented a critical evaluation of cross-cultural industrial psychological research. Fiedler, Mitchell, and Triandis in the *Journal of Applied Psychology* (1971) presented a method for cross-cultural training that was later shown (Black and Mendelhall 1990) to be effective in improving the performance of persons from one culture working in another culture. Inkeles and Smith in *Becoming Modern* (1974) showed that the level of modernity of people in six cultures was correlated with working in factories. Bass and Burger in *Assessment of Managers: An International Comparison* (1979) contributed a study of managers in several countries. England, Nagandhi, and Wilpert in *Organizational Functioning in Cross-Cultural Perspective* (1979) ex-

amined whether culture or other factors account for most of the variance of organizational phenomena. Hofstede in *Culture's Consequences* (1980) supplied a major dataset showing cultural differences in the values of IBM employees across a large number of countries. Sinha in the *Nurturant Task Leader* (1980) published a book that described effective leadership in India. Bhagat and McQaid in a *Journal of Applied Psychology Monograph* (1982) reviewed cross-cultural studies that used the subjective culture perspective.

3.3 The Recent Quarter Century

Nancy Adler in the *Journal of Applied Behavior Science* (1983) examined organizational development in a cross-cultural perspective. Ayman and Chemers in the *Journal of Applied Psychology* (1983) showed that Iranian workers want managers who are benevolent paternalists (warm and stern). Graham in the *Journal of International Business Studies* (1983) examined negotiation behavior in three cultures. Wilpert and Sorge (1984) in *International Perspectives on Organizational Democracy* summarized a major cross-cultural study of industrial democracy. Misumi in *The Behavioral Science of Leadership: An Interdisciplinary Japanese Research Program* (1985) presented a Japanese theory of leadership that became influential in cross-cultural studies of leadership, in many cultures. Adler in *International Dimensions of Organizational Behavior* (1986, 2nd edition 1990) published a widely used text that covered the field. A team of scholars studied the meaning of work in several cultures (MOW 1987). Smith and Peterson in *Leadership, Organizations and Culture: An Event Management Model* (1988) published a study of leadership that focused on event management. Black and Mendelhall (1990) reviewed the effectiveness of cross-cultural training programs. Hofstede in *Cultures and Organizations* (1991) contrasted national and organizational cultures, and included a methodological appendix concerned with cross-cultural research. Schwartz (1992) presented a theory of value structure and data from 20 countries. Later he published data from 50 countries. Many cross-cultural psychologists use both the Hofstede and Schwartz datasets as predictors of some organizational phenomena they happen to be working on. Earley (1993) showed that in collectivist cultures there is less social loafing than in individualist cultures when employees work with in-group members. Earley and Erez (1997) in *New Perspectives on International Industrial and Organizational Psychology* edited a collection of papers on industrial and organizational psychology that covered broad topics such as the use of theory in that field, approaches to the definition of culture, motivation, diversity in an international context, intercultural communication, decision making in multicultural teams, individual–union relationships across cultures, human resource management, leadership, negotiation and reward allocation, as well as critical commentaries by the editors. Robert House and a team of more than 100 social scientists will publish their work on the implicit meaning of leadership in over 100 cultures during the early part of the twenty-first century.

4. The Future

Predictions are hazardous, but it is likely that the major developments in this field will be in the area of dealing with diversity, especially when the diversity includes differences in culture, social class, and gender.

See also: Collectivism: Cultural Concerns; Cross-cultural Psychology; Cross-cultural Research Methods; Cultural Psychology; Development: Organizational; Industrial Geography; Industrial–Organizational Psychology: Science and Practice

Bibliography

Barrett G, Bass B M 1976 Cross-cultural issues in industrial and organizational psychology. In: Dunnette M D (ed.) *Handbook of Industrial and Organizational Psychology*. Rand McNally, Chicago, pp. 1639–86

Berry J W 1968 *Directory of Cross-Cultural Psychologists*. Department of Psychology, Queen's University, Kingston, ON

Bhagat R S, McQaid S J 1982 Role of subjective culture in organizations: A review and directions for future research. *Journal of Applied Psychology Monograph* **67**: 653–85

Black J S, Mendelhall M 1990 Cross-cultural training effectiveness: A review and a theoretical framework. *Academy of Management Review* **15**: 113–36

Dubin R 1976 *Handbook of Work, Organization, and Society*. Rand McNally, Chicago

Earley P C 1993 Social loafing and collectivism: A comparison of the United States and the People's Republic of China. *Administrative Science Quarterly* **34**: 565–81

Frost E 1920 What industry wants and does not want from the psychologist. *Journal of Applied Psychology* **4**: 18–24

Garth T R 1920 Racial differences in mental fatigue. *Journal of Applied Psychology* **4**: 235–44

Garth T R, Ikeda K, Gardner D A 1933 Japanese work curves. *Journal of Applied Psychology* **17**: 331–6

MOW International Research Team 1987 *The Meaning of Work*. Academic Press, New York

Schwartz S 1992 Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: Zanna M (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 25

Smith M 1944 *Handbook of Industrial Psychology*. Philosophical Library, New York

Triandis H C 1972 *The Analysis of Subjective Culture*. Wiley, New York

Wilpert B 2000 Presidential Address: 24th International Congress of Applied Psychology, San Francisco, August 9–14, 1998. *Applied Psychology: An International Review* **49**: 3–22

H. C. Triandis

Industrial Ecology

1. Introduction

The words 'industry' and 'ecology' were combined to the phrasing of 'industrial ecology' in the late 1980s, when two notions gained a solid foothold in the minds and perspectives of those working for a sustainable development (see *Sustainable Development*). One notion was that the minimization of waste and emissions in individual process steps or products need not result in an overall reduced environmental impact. The other was that money could be made through a reduction of waste, emissions, or resource utilization. These two notions gave birth to a variety of concepts that renewed the debate on sustainable industrial development. Whereas some of these concepts focus on products (e.g., life cycle assessment, design for the environment) or facilities and production processes (e.g., pollution prevention), industrial ecology focuses on industrial systems. In this contribution industrial ecology is discussed as a vision for a sustainable world, as an emerging interdisciplinary research field, and how it can be made operational at different systems levels.

2. Vision

Industrial ecology finds its legitimization in the question of what the environmental impact would be when, within a period of three or four decades, with a growing world population, each individual person lives at, ideally, a standard of living common in Western industrialized countries during the 1990s (see *IPAT (Impact, Population, Affluence, and Technology)*). The amounts of natural resources required to sustain the accompanying level of consumption would probably not be available in this planet and the levels of pollution and waste generated would probably exceed the world's regenerative capacity (Frosch and Gallopoulos 1989; see *Human–Environment Relationship: Carrying Capacity*).

Such arguments have generated widespread support for the view of approaching sustainability through increased eco-efficiency (Schmidheiny 1992) and break-through innovations in technology and consumption (Von Weizsäcker et al. 1997). Industrial ecology is a perspective which helps to find 'imaginative solutions to these resource and environmental problems,' by adopting the notion that 'the industrial system ought to be modified so as to mimic the natural ecosystem in its overall operation' (Frosch and Gallopoulos 1992, p. 271). In this view, industrial ecology is the basis for a radical transformation of industrial society (Ehrenfeld 1997).

The metaphor from biology is the most important conceptual contribution of industrial ecology. In an industrial ecosystem, 'the consumption of energy and materials is optimized and effluents of one process [..] serve as the raw materials for another process' (Frosch and Gallopoulos 1989, p. 94), much like nutrients flow in biological ecosystems. The industrial network is seen as a system of mutual dependent transformation processes which is an interrelated part of a larger whole, 'analogous in its functioning to a community of biological organisms and their environment' (Frosch and Gallopoulos 1992, p. 272). Hence the fundamental importance of industrial metabolism to industrial ecology (Ayres 1989; see *Industrial Metabolism*).

Industrial ecology aims at a reduction of the environmental impact of materials and energy flows through increased resource efficiency at the systems level. Two dimensions characterize the systems thinking in industrial ecology: 'space' and 'time.' Regarding 'time,' industrial ecology has adopted life-cycle thinking. Just like concepts such as product stewardship and integrated substance chain management it takes into account the potential environmental impacts in all the stages in the life cycle of specific materials and products including design, production, use and disposal, i.e., 'from cradle to grave.' Regarding 'space,' industrial ecology has adopted different entities to delineate materials and energy flows, including the entire world, river basins such as the Rhine, the economy of a specific nation or region, industries or even companies. End-of-pipe pollution control and pollution prevention have traditionally been the guiding concepts for waste reduction in individual companies and facilities; industrial ecology extends to these concepts by stressing inter- rather than intra-firm linkages, dependencies and opportunities for waste reduction.

3. Research Field

Industrial ecology constitutes an emerging, multidisciplinary research field, whose researchers share the broad vision sketched above. They take a systems view in studying material and energy flows and their transformations into products, byproducts and wastes through industrial and natural systems (industrial metabolism), and they attempt to reduce the environmental impact of industrial systems on ecological systems by reducing the materials intensity of production and consumption. However, there is not (yet) a clear demarcation of the field. Some restrict industrial ecology to the study of biological, chemical, and

physical processes, whereas others include the study of informational, legal and economic incentives into industrial ecology, or even the study of inter- and intra-organizational structures, arrangements and coordination, and consumption (see *Consumption, Environmentally Significant*). Despite a shared unit of analysis—the flow and transformation of materials and energy—different research approaches into industrial ecology can be distinguished. The first aims at describing materials and energy flows. The second aims at their control through the traditional governance instruments of hierarchy, the market and the law. To be effective, this approach requires environmentally relevant information to be collected and processed for use in decision making. To that end, various tools are being developed. Finally, observing that the control of materials and energy flows through traditional governance instruments is not the entire story, new governance instruments are being developed.

3.1 Description of Flows and Transformations

Descriptive research into industrial ecology aims at the understanding of the flows of materials and energy both within industrial and between industrial and ecological systems, as well as their transformations in products, byproducts and effluents. For example, the industrial metabolisms of elements such as chlorine and various metals, of materials such as paper and of products such as chloro-fluoro carbons (CFCs) and polyvinyl chloride (PVC) have extensively been described (see *Industrial Metabolism*).

Such studies are criticized on various grounds. First, the choice of elements, materials, and products is biased towards those that are released as nonfugitive, nondilute, point source industrial wastes or used as nondissipative consumer products (O'Rourke et al. 1996). There is little consideration of how dissipative products such as coatings, detergents, food and pesticides would fit into an industrial ecology. To some extent, 'green chemistry' (Anastas and Breen 1997) may help reduce the environmental impact of dissipative products, but it leaves unquestioned the open-ended character of these flows. Second, the metabolism of elements, materials and products within industry requires the input of energy, as does the closing of currently open-ended flows by recycling. However, the energy component in both the current, open-ended and the future, closed flows remains largely unquestioned. Finally, in some articles systems boundaries have implicitly been defined, in most they are defined in terms of administrative boundaries (country, region). This choice of systems boundaries probably reflects the interests of the data collecting bodies which are often national or supra-national public bodies or industry associations. Cross-border material flows is one of the challenges to industrial metabolism (Rejeski 1997), yet from the industrial ecology perspective it would be more relevant to look for boundaries as defined by the area within which a material is transformed and transported by economic, physical, chemical, and biological forces. To choose natural systems boundaries does not resolve the problem since the economically motivated flow of materials and energy tends not to be restricted by oceans, watersheds, and mountain ridges.

3.2 Governance of Materials Flows: Traditional Approaches

Industrial ecology is aimed at the deliberate and rational control of materials and energy flows. Many studies share the general assumptions that decision makers—whether product developers, corporate executives, public administrators, or consumers—would take different decisions if they had the 'right information' enabling them to evaluate the environmental consequences of their decisions, if they were exposed to the 'right prices' reflecting the monetary value of the externalities of materials and energy use (O'Rourke et al. 1996), and if they were not told to do otherwise by regulators.

The collection and sharing of information on environmentally relevant aspects of production and consumption are critical components of industrial ecology. Industrial ecology requires firms to more closely cooperate in order to make its promise of increased environmental benefits come true. Therefore, firms need to collect, process, and share data on resource utilization, waste-stream monitoring, and intra and inter-firm component flows. In this respect, existing information systems can be modified to support life-cycle oriented environmental management (Shaft et al. 1997). Thus, many corporate functions will have to deal with environmental issues whereas before they had not; this needs careful planning, top management commitment, and dedicated programs to involve those functions (Dillon 1994).

Consumers, too, need to know about the environmental impact of products and services if they are to value reduced environmental impact in their purchases. Third-party voluntary environmental product labels were introduced as a means of informing consumers accurately on the environmental qualities of products. According to Salzman (1997), such labels have proven not to be effective in influencing consumers' buying behavior, but they may have resulted in reduced environmental impact in different ways, e.g., by stimulating competition among producers for the label and by enabling public and institutional purchasing programs to select labeled products over nonlabeled products.

Not only the information, also the prices need to be right (see *Environmental and Resource Management*). Various studies deal with the questions of whether and how recycling can be economically viable and environmentally beneficial. One of the lessons learnt is the vulnerability of recycling systems to changing costs for collection, transportation, and final disposal of unprocessable wastes. However, recycling schemes can be made more robust if clever solutions are found for collecting goods to be recycled, if those goods are designed such as to meet critical specifications of the recycling system and, if the goods are recycled into something that is marketable (Isaacs and Gupta 1997, Lave et al. 1998).

Taking the industrial ecology perspective may enhance a firm's competitive position through increasing resource efficiency at three levels: within the firm's value chain, within the wider value system, and beyond the value system in a situation of industrial symbiosis (Esty and Porter 1998). The waste exchange network in Kalundborg (Denmark) is one example of industrial symbiosis, the one that developed in the Austrian province of Stryria is another. Schwarz and Steininger (1997) studied why companies participate in the Styrian network. Cost calculations constitute the main reason, both for companies offering wastes as raw materials (increased revenues, reduction of disposal costs) and for companies procuring these wastes (cost reductions). They also studied how the exchanges within the network are governed. Most of the participating companies have entered into long-term contractual agreements or even joint-ownership subsidiaries because of the need to secure the quantity, quality, and price of the wastes to be traded. Alternative governance structures such as integrating waste valorizing activities within the company, or trading waste through spot market arrangements, are considered less efficient, essentially because of higher transactions costs to both the vendor and the seller. There may also exist situations in which taking the industrial ecology perspective does not result in enhanced firm competitiveness, e.g., if the associated cost exceeds the benefits, or if following the suggestions of industrial ecology would lead to undertaking activities that are not valued or permitted in the prevailing economic and institutional frameworks (Esty and Porter 1998).

An example of the latter is the study of Heaton and Banks (1997). They detail how the US statutory system of environmental regulation inhibits technological innovation. 'Perverse incentives' are in place that stimulate firms to use older, more polluting equipment because newer technologies need to meet stricter pollution control levels and hence are more expensive. Additionally, the use of 'best available technology' standards is an impediment to innovation because technology tends to develop at a quicker rate than the adaptation of the definition of what is the 'best' available technology. In such situations, legislative reform is required in order for companies to adopt environmentally more beneficial technologies and practices (see *Environmental Policy: Protection and Regulation*).

3.3 Tools

Most of the tools for industrial ecology are developed to deal with the information problem. Each of these tools has its own body of literature, journals and practitioners, which complicates a comprehensive discussion. Suffices a short characterization of the relationships between the various tools.

Materials flow analysis (MFA) and substance flow analysis (SFA) are basic analytical tools for industrial ecology derived from the principle that matter cannot disappear—what goes in must come out. The life-cycle assessment of products (LCA) extends to these analyses by attempting to quantify the environmental impact of the use of materials and substances in particular product designs. The resulting environmental profile of a product can be used for comparison against competing products or for suggesting ways to improve that particular product design, i.e., design for the environment. The LCA methodology can also be used to justify the eco-labeling of product designs. When applied to production processes, MFA/SFA is helpful to identifying inefficiencies or leaks, which in turn be remedied by process innovations, i.e., cleaner production or pollution prevention. For knowing whether or not any progress has been made one needs environmental performance measures that heavily rely on process data. Both managerial and environmental information systems are likely to contain relevant data. Each of these tools in the 'industrial ecology toolbox' has one or more of four basic functions (Van Berkel et al. 1997):

(a) to enable the identification, quantification and allocation of environmental impacts to (parts of) production processes and product life cycles;

(b) to facilitate the generation of improvement options for products designs, product life cycles, and production processes;

(c) to provide a structured approach for the evaluation of, and priority setting among, environmental impacts and/or environmental improvement options; and

(d) to specify procedures and routines for the development of industrial ecology projects.

An obvious problem with information based tools is that they require considerable amounts of data which are costly to collect, while it remains doubtful whether information would in itself be helpful given the complex trade-offs, ambiguities, and uncertainties in product development. Moreover, it appears that the environmental impact of products is determined to a greater extent by the question of how consumers use a product, than by the product design (Udo de Haes

1997). Finally, to focus on the environmental benefits of one design alternative over another may entirely miss the more fundamental question of whether the functionality or service of that product cannot be delivered by different means (O'Rourke et al. 1996).

3.4 Governance of Materials Flows: New Approaches

Several authors are less convinced that material and energy flows can actually be managed through information, regulation, and prices. For example Harrison (1998) points out that co-operative approaches to environmental protection, such as negotiated agreements, flexible enforcement, voluntary codes and agreements, have been devised because of the difficulty of implementing traditional 'command and control' regulation and market-based incentives. Little is known, however, about their efficacy and efficiency. It might be the case that a rigorous evaluation of such co-operative approaches reveals as much about the specific policy instrument design as about companies' motives and ability to join them.

For some companies in some industries, trends of globalization of trade and competition and flexible specialization shape the scope for strategic decision making on materials use (Bragt et al. 1998, Boons and Baas 1997). Other companies may deliberately choose to follow industrial ecology strategies without regulatory or market incentives commanding them to do so. One example is an outdoor apparel manufacturer that decided to substitute organically for conventionally grown cotton in its garments. Prices and information constituted significant barriers to overcome, but of greater importance was a paradigm shift in the company's management team, which made them accept the risk of a far less competitive cost structure. Because of this paradigm shift, the company was motivated to help existing suppliers adopt, and to look for new suppliers meeting, new quality standards (Chouinard and Brown 1997).

At a more fundamental level, the traditional governance approaches to, and tools of, industrial ecology can be criticized on the underlying assumptions of how firms make decisions (Taylor 1996). Industrial ecology tends to view materials and energy flows as the combined result of natural forces and the aggregate productive and consumptive activities of firms and consumers, whose decisions are primarily guided by direct economic and legal constraints. Within this perspective, Rejeski (1997) points out that too little attention has been given to the service sector and household consumption as mobilizers of materials flows. Beyond this perspective, industrial ecology has hardly begun to consider strategic decision making, inter-firm collusion, the exercise of power, mimetic behavior, cultural preferences, socio-economic institutions, etc., both as additional explanatory factors to understand why materials and energy flows have materialized as they have, and, in some instances, as reasons for the difficulty of changing those materials and energy flows. Systems of production and consumption are not invented, but have developed in a complex way, in which coincidences and unanticipated consequences intertwine with the actions that follow upon the intentions of individuals, organizations, and institutions.

4. Putting Theory into Practice

The practical value of industrial ecology is being explored in two directions. One direction is the application of the principles of industrial ecology at a local level in the development of eco-industrial parks. The other is the exploration of reduced resource intensity of production and consumption at the societal level through dematerialization—which is understood to include product life extension, decarbonization and the substitution of services for products.

4.1 Eco-industrial Parks

An eco-industrial park (EIP) is: 'A community of manufacturing and service businesses seeking enhanced environmental and economic performance through collaboration in managing environmental and resources issues [...]. By working together, the community of businesses seeks a collective benefit that is greater than the sum of the individual benefits each company would realize if it optimized its individual performance only. The goal of an EIP is to improve the economic performance of the participating companies while minimizing their environmental impact' (Lowe 1997, p. 58). The paradigmatic exemplar of an eco-industrial park is the 'industrial symbiosis' between a number of companies in and around the town of Kalundborg (Denmark) that exchange wastes and heat (Fig. 1). The Kalundborg symbiosis was not planned, but has gradually developed from a co-operation between the neighboring industries and the municipality of Kalundborg since the early 1960s (Ehrenfeld and Gertler 1997). Eco-industrial parks can be developed at green-field sites or through the renewal and restructuring of existing industrial parks.

During the 1990s, the concept of eco-industrial park has quickly developed into a planning tool for industrial development in both industrialized and developing countries. Although the example of Kalundborg shows that self-organization is an important factor in the development of industrial ecosystems, several examples exist where university researchers, specialized consultants, and person-

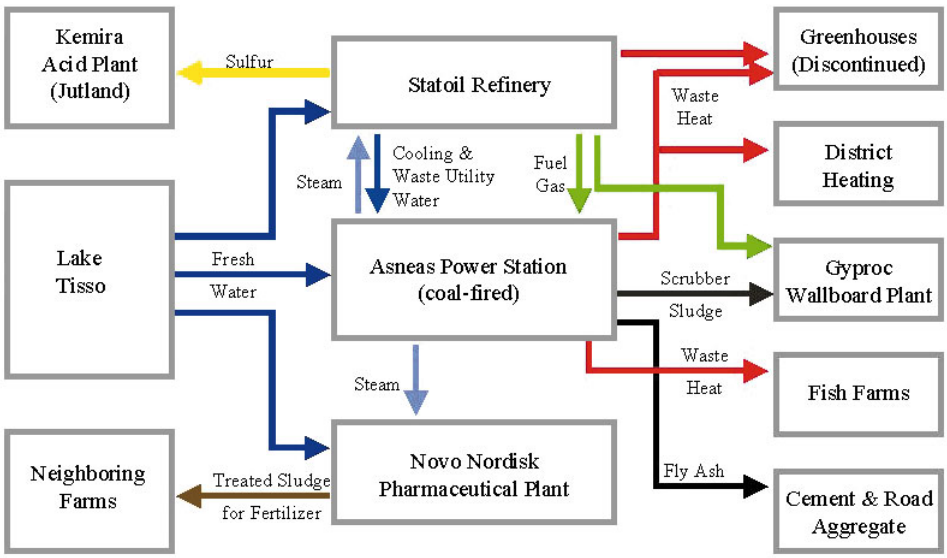

Figure 1
The industrial ecosystem at Kalundborg, Denmark (Source: Ehrenfeld and Gertler 1997, copyright Massachusetts Institute of Technology and Yale University. Reproduced with permission)

nel from economic and environmental development agencies have contributed in helping to develop eco-industrial parks. Several conditions for success have been identified: a clear vision of park's values and performance objectives, careful screening of new companies, methods and information to support companies in seeking by-product trades, flexibility in recruitment strategy, and assurance of continuing support. Challenges and risks that need to be faced include undesired resource dependencies, technological lock-in, leaking of proprietary information, and uneven quality (Lowe 1997)—the same principal-agent problems stemming from the risk of misuse of private information that pertain in any collaborative venture.

4.2 Dematerialization

Dematerialization refers to the 'absolute or relative reduction in the quantity of materials used and/or the quantity of waste generated in the production of a unit of economic output' (Cleveland and Ruth 1998, p. 16). It is considered a crucial strategy for advancing industrial ecology at the societal level. However, the suggestion to develop a service economy that is based on the delivery of function rather than products, intermingles with other debates on dematerialization. For several decades, it has been a central issue in concerns over society's dependency of, and impact on, natural resources. Society's intensity of materials use has strategic, geo-political, and security implications, in addition to environmental significance. Countries strive for some basic level of materials autarchy, especially the US, which may be realized by increasing direct control over, or reducing the level of dependency on, critical resources.

There is a lively debate on the topic of dematerialization. It has even been argued that dematerialization is a natural event in the development of economies —the dematerialization hypothesis. The relevance of this debate for industrial ecology is twofold. First, if it were the case that economies at a certain point in their development do dematerialize, then the need for a radical change in society's metabolism—as advanced by industrial ecology—would be less pressing. Second, the research methods developed for empirical research into the dematerialization thesis can also be used to evaluate to what extent industrial ecology's efforts to make society less material intensive are effective.

Dematerialization is considered the consequence of changing demand for end products in different stages of economic development, because increased demand for services and products with higher added value, results in per capita reduction of materials intensity (Malenbaum 1978). When the market for a particular base material saturates, its use rate as measured in kilograms per unit of gross domestic product declines and at a later stage even its per capita consumption may start to decline (Larson et al. 1986). Sources of per capita dematerialization may originate from process or product innovations, i.e., a more efficient

materials use, from materials substitution and from changing consumer preferences due to higher income levels. The additional dollar is spent on services, higher-value and knowledge-intensive goods, thus reducing the materials intensity of consumption. The resulting declining growth in materials intensity may eventually lead to a decrease of a society's materials intensity.

Several authors have contested the dematerialization thesis. Labys and Waddell (1989) argue that society 'transmaterializes'—through a process of materials substitution—rather than dematerializes. The intensity of materials use in society does follow the typical S-shaped curve of growth, maturity, and decline, but because the collection of data on the use of materials tends to lag behind in the introduction and growth stages, there is the suggestion of, but no real, dematerialization. De Bruyn and Opschoor (1997) propose that periods of dematerialization are followed by periods of 'rematerialization.' This pattern is explained by the evolutionary view of economic growth as a 'punctuated equilibrium' in which periods of stable economic growth are interrupted by clusters of innovations. Whether or not there is dematerialization depends on the relative importance of two rates over time: the rate of economic growth and the rate of efficiency gains in materials use. For dematerialization to occur, it is at least needed that rate of efficiency gains is greater than the rate of economic growth. Since it is likely that during periods of stable economic growth any efficiency gains in productivity are used to increase production (Bunker 1996), an increase of materials intensity will be observed, i.e., rematerialization. Thus, dematerialization may only be observed when there is a major structural or technological change in a society. Still others contest the dematerialization thesis by pointing out that the materials intensity of industrialized countries tends to converge in a narrow range between 16–20 tons of material per capita per year, and that despite structural changes and growth of services in advanced economies, the intensity of materials flows in economies is still growing in absolute terms.

Not only the dematerialization thesis itself has been criticized, so has the proposition that dematerialization would have environmental benefits. If all material eventually dissipates into the environment, then a reduction of materials use need not be beneficial if it is not considered how (in what form, what concentration) that material dissipates into the environment. It might be the case that a more intense use of a material enables a better containment of that material and a reduction of its dissipation rate to the environment

All in all, it must be concluded that at present there is no compelling evidence that society (i.e., the US) is dematerializing. Moreover, knowledge of the extent of, and mechanisms behind, patterns of materials use is scarce and limited to individual materials and specific industries, mostly related to metals (Cleveland and Ruth 1998).

See also: Human–Environmental Relationships; Industrial Geography

Bibliography

Anastas P T, Breen J J 1997 Design for the environment and Green Chemistry: The heart and soul of industrial ecology. *Journal of Cleaner Production* **5**(1–2): 97–103

Ayres R U 1989 Industrial metabolism. In: Ausubel J H, Sladovich H E (eds.) *Technology and Environment*. National Academy Press, Washington, DC, pp. 23–49

Boons F A A, Baas L W 1997 Types of industrial ecology: The problem of coordination. *Journal of Cleaner Production* **5**(1–2): 79–86

Bragt A, Bridge G, Den Hond F, Jose P D 1998 Beyond greening: New dialogue and new approaches for developing sustainability. *Business Strategy and the Environment* **7**(4): 179–92

Bunker S G 1996 Raw material and the global economy: Oversights and distortions in industrial ecology. *Society & Natural Resources* **9**: 419–29

Chouinard Y, Brown M 1997 Going organic: Converting Patagonia's cotton product line. *Journal of Industrial Ecology* **1**(1): 117–29

Cleveland C J, Ruth M 1998 Indicators of dematerialization and the materials intensity of use. *Journal of Industrial Ecology* **2**(3): 15–50

De Bruyn S M, Opschoor J B 1997 Developments in the throughput—income relationship: Theoretical and empirical observations. *Ecological Economics* **20**: 255–68

Dillon P S 1994 Implications of industrial ecology for firms. In: Allenby B R, Richards D J (eds.) *The Greening of Industrial Ecosystems*. National Academy Press, Washington, DC, pp. 201–7

Ehrenfeld J R, Gertler N 1997 Industrial ecology in practice: The evolution of interdependence at Kalundborg. *Journal of Industrial Ecology* **1**(1): 67–79

Ehrenfeld J R 1997 Industrial ecology: A framework for product and process design. *Journal of Cleaner Production* **5**(1–2): 87–95

Esty D C, Porter M E 1998 Industrial ecology and competitiveness: Strategic implications for the firm. *Journal of Industrial Ecology* **2**(1): 35–43

Frosch R A, Gallopoulos N E 1989 Strategies for manufacturing. *Scientific American* **261**(3): 94–102

Frosch R A, Gallopoulos N E 1992 Towards an industrial ecology. In: Bradshaw A D, Southwood R, Warner R (eds.) *The Treatment and Handling of Wastes*. Chapman & Hall, London

Harrison K 1998 Talking with the donkey: Cooperative approaches to environmental protection. *Journal of Industrial Ecology* **2**(3): 51–72

Heaton G R, Banks R D 1997 Toward a new generation of environmental technology: The need for legislative reform. *Journal of Industrial Ecology* **1**(2): 23–32

Isaacs J A, Gupta S M 1997 Economic consequences of increasing polymer content for the US automobile recycling infrastructure. *Journal of Industrial Ecology* **1**(4): 19–23

Labys W C, Waddell L M 1989 Commodity lifecycles in US materials demand. *Resources Policy* **15**: 238–51

Larson E D, Ross M H, Williams R H 1986 Beyond the era of materials. *Scientific American* **254**(6): 24–31
Lave L, Conway-Schempf N, Harvey J, Hart D, Bee T, MacCracken C 1998 Recycling postconsumer nylon carpet. *Journal of Industrial Ecology* **2**(1): 117–42
Lowe E 1997 Creating by-product resource exchanges: Strategies for eco-industrial parks. *Journal of Cleaner Production* **5**(1–2): 57–65
Malenbaum W 1978 *World Demand for Raw Materials in 1985 and 2000*. McGraw-Hill, New York
O'Rourke D, Connelly L, Koshland C P 1996 Industrial ecology: A critical review. *International Journal of Environment and Pollution* **6**(2–3): 89–112
Rejeski D 1997 Mars, materials, and three morality plays: Materials flows and environmental policy. *Journal of Industrial Ecology* **1**(4): 13–18
Salzman J 1997 Informing the green consumer: The debate over the use and abuse of environmental labels. *Journal of Industrial Ecology* **1**(2): 11–21
Schmidheiny S (with the Business Council for Sustainable Development) 1992 *Changing Course*. MIT Press, Cambridge, MA
Schwarz E J, Steininger K W 1997 Implementing nature's lesson: The industrial recycling network enhancing regional development. *Journal of Cleaner Production* **5**(1–2): 47–56
Shaft T M, Ellington R T, Meo M, Sharfman M P 1997 A framework for information systems in life-cycle-oriented environmental management. *Journal of Industrial Ecology* **1**(2): 135–148
Taylor M 1996 Industrialization, enterprise power, and the environmental change: A exploration of concepts. *Environment and Planning A* **28**: 1035–51
Udo de Haes H 1997 LCA can be very relaxed. *Journal of Industrial Ecology* **1**(4): 3–5
Van Berkel R, Willems E, Lafleur M 1997 Development of an industrial ecology toolbox for the introduction of industrial ecology in enterprises—I. *Journal of Cleaner Production* **5**(1–2): 11–25
Von Weizsäcker E, Lovins A B, Lovins L H 1997 *Factor Four: Doubling Wealth—Halving Resource Use*. Earthscan, London

F. den Hond

Industrial Geography

Industrial geography is related to but distinct from industrial location. Industrial location is the study of the location decision-making processes of firms and reflects the trade-off between access to natural resources and access to markets. Early work in the field focused on the single location, pricing taking firm (see *Location Theory*) (Weber 1971). Subsequent work incorporated concerns including the nature of the market (see *Spatial Interaction Models*) (Lösch 1954), the economic spill-over effects of firms in the same location (Hoover 1948), and the importance of and variability in pools of labor (see *Spatial Labor Markets*) (Storper and Walker 1989). In contrast with the study of industrial location, this chapter focuses on the historical development of the industrial landscape in the Western world.

1. *The Industrial Geographic Landscape*

What most people see today in the industrial landscape is the sum of multiple periods of economic development. In American and Western European settings, the old mill town is a physical representation of a type of industrial landscape that evolved prior to or in conjunction with the Industrial Revolution (see *Historical Geography*). By beginning with the late eighteenth and early nineteenth centuries, we necessarily leave out a considerable period of the economic history that forms the bedrock of the contemporary industrial landscape. Under feudalism, for example, industrial activity was focused on the manor and the creation of goods for local consumption. Occasional surpluses were produced that were traded in market towns (Rosenberg and Birdzell 1986). With the decline of feudalism and the ascendancy of urban centers of trade and production, the industrial landscape began to emerge in the form of urban workshops. Independent manufacturers were only one source of manufactured goods, however. Merchants were key to early industrialization; they organized not only the markets for manufactured goods, but also production outside the influence of urban guilds. The 'putting out' system of rural manufacturing extended the influence of the industrial landscape significantly beyond urban centers. Rural manufacturing provided a check on urban wages but eventually declined when merchants could no longer guarantee or control output from rural manufacturers (see *Rural Geography*).

Prior to the Industrial Revolution, the production of manufactured goods unfolded within a system of handicraft manufacture under the purview of a master craftsman. One or more journeymen or apprentices assisted an exceedingly clever and dexterous person, the master.

The regulation of this form of production rested with guilds, production organizations that regulated skills and access to skills, and to a much lesser extent access to markets (Landes 1969). By restricting access to skills, craftsmen were able to preserve their markets and ensure public belief in the credibility of their products. Guilds acted as brakes on the expansion of output and thus overproduction. In later periods, guilds also thwarted technological change, a detail that led ultimately to their own demise.

The movement from craft production to the putting out system was an outgrowth of the merchants' increasing control over producers. The expansion of the market and the lack of access to and knowledge of its extensiveness placed artisans at a decided disadvantage. Market reach and knowledge of the tastes and preferences of distant customers was a clear

advantage of merchants who could accumulate demand and periodically fill it (see *Economic Geography*) (Glasmeier 1990).

Putting out, while highly efficient in an era with few economic alternatives, grew less and less viable as an organizing system over time. As workers grew wealthier and additional avenues of income creation became available to them, merchants found it increasingly difficult to control the output and quality of cottage industry goods. During periods of market growth, it became harder to meet new demand as workers began to assert more control over their lives and over their labor process. The need to find an alternative, more controllable system of manufacture became overwhelming as markets grew and demand burgeoned. In Britain, geographic range had essentially been reached and in some cases exceeded. Transport costs became a real impediment.

As control of workers' efforts lessened, and the limits of geographic expansion were reached, a slow but steady transformation took place as machines began to replace human labor, and the organization of labor within specific locales and structures provided the control employers sought. The movement to a factory-based system of manufacture was neither rapid nor complete. The slow demise of the putting out system provides one explanation for the emergence of factories: the need to control labor and the production process. Volumes have been written on this subject, from a wide variety of perspectives and with an equally broad number of emphases. What is important here is the way the movement from a more decentralized to a more centralized system removed control of production from workers laboring in the manufacture of a good, and turned it over to the capitalist. Of additional importance is the recognition that the process of change did not occur in the same way simultaneously in all relevant locations.

The ways in which mechanization unfolded across industrial sectors also are relevant to understanding industrial geography. The reasons for early and rapid mechanization in some sectors and resistance of machine support in others can be reduced to several intervening factors. The nature of the good largely determined its susceptibility to mechanization. This attribute went hand-in-hand with the extent of the market for the good. If the good was high value/low volume and of a luxurious nature, labor costs were relatively insignificant and thus a minimal stimulus to change, unless of course the transition to machines was twinned with a movement toward a broader market and a more widely disseminated good. In this case, market shifts could simply require change in order to hold a longer-term market position. Institutional factors also could serve as impediments to change. The number of possible explanations supporting the lack of investment are certainly sufficient to see that the prospect of a smooth functioning factor market acting as a stimulus for change was inherently contingent on elements other than just the short-term costs of production.

Early centers of industrialization were highly specialized—a result partly of natural resource location as well as of built-up competencies (see *Technology Districts*), the initial impetus of which was often accidental. Trade was essential to the growth of industrialization in place. The formation of institutions and social relationships combined with key developments in technology facilitated and ultimately encouraged trade among regions. Fundamental to the success of trade (see *International Trade: Geographic Aspects*) across long distances was the development of shipping technology that allowed the movement of large volumes of goods in reasonable time. Simultaneously, in order for trade to occur, mechanisms were needed to reduce risk and to finance the time between the manufacture and sale of a good. With the creation of a number of institutional innovations such as the banking, credit, and insurance sectors, industrialization in the late eighteenth and early nineteenth centuries moved beyond local production for local consumption and toward production for trade.

The Industrial Revolution (see *Economic Geography*; *Historical Geography*) marks a critical moment in the formation of the industrial landscape. Essential to the pervasive diffusion of the Industrial Revolution was the ability to harness power sources in order to concentrate economic activities in space. Simultaneously, the development of new technologies led both to an enhanced capacity to organize industrial activity and to the creation of new manufactured goods themselves. Finally, the organization of production within factories allowed the achievement of economies of scale and task specialization across the production process.

Before the beginning of the eighteenth century, the primary source of power used in production was human and animal (Rosenberg and Birdzell 1986). This dependence significantly restricted the size, scale, and location of industrial activity. With the ability to harness inanimate power sources, industrial activity moved toward a larger scale and a greater degree of specialization. First water, then charcoal, and finally coal became the defining power sources of the early Industrial Revolution. The industrial landscape began to spread out as manufacturing became unshackled from water power sources. Timber and charcoal were an intermediate solution; both, however, quickly declined in importance when most of the trees were cut down and charcoal became uneconomic to manufacture. With the discovery of large supplies of coal, production spread further and moved quickly beyond a relatively few industrial centers.

Technological change (see *Technology Districts*) is a critical marker of the evolving industrial landscape. In addition to the ability to utilize inanimate energy sources, a hallmark of the Industrial Revolution was the development of machines that replaced human

action. Metal-cutting and -shaping machines were vital to the formation of an industrial equipment industry. New materials and the understanding of the chemical basis of new materials made possible more complex manufacturing processes and new products. This in turn led to specific spatial concentrations of firms in the same industry, as divisions of labor formed and fanned out in independent enterprises.

2. *The Organization of Industry in the Factory Era*

The organization of capitalist enterprises into factories occurred as both merchants and manufacturers sought to control and coordinate an entire production process (see *Economic Geography*; *Regulation Theory in Geography*). During the craft era, workers exhibited some control over the labor process (declining through time). This had important implications for the volume of output and therefore the size of markets that could be served, the level and source of innovations, and the sources and control of power. The need to harness these capabilities into a more centralized system of command and control had major implications for how industrial systems evolved. According to Lazonick, 'The transition to the era of proprietary capitalism brought the integration of asset ownership with managerial control and with it an internalization of aspects of the craft system. In particular, in order for firms to generate (what were by the standards of the time) high-quality products at low unit costs.' Further, 'firms relied on an ample supply of highly skilled and well-disciplined labor. Senior workers—who eventually came to be known collectively as the aristocracy of labor—not only provided their own skills to the building and operating of machinery, but also recruited junior workers whom they trained and supervised on the shop floor' (1991, p. 25).

The heavy reliance on skilled labor had obvious spatial implications. Firms were more or less bound to a particular place. This dependence on labor is one of the prime determinants of spatially bound production systems (see *Location Theory*; *Economic Geography*), aptly captured by such well-known observers as Marshall (1997). The external economies within such a complex were substantial. Materials, labor, and services (distribution) were sufficiently honed that spatial proximity enabled work in progress to flow through its vertically specialized branches, and to market its output (Lazonick 1991, p. 26). These regional concentrations facilitated high levels of specialization. Consequently, employers' ongoing reliance on skilled labor to organize work on the shop floor diminished the need for business firms to invest in the development of managerial structure and organizational capabilities (Lazonick 1991, p. 26). It also meant that pressures to cut costs were resolved through market transactions rather than through internal restructuring.

The system was self-perpetuating. Because proprietors had access to this complementary system of transactions it was unnecessary to invest in or develop complex managerial structures. Instead, in the early phases of factory industrialization, proprietors tended to own single establishments and rely on the market for the provision of necessary inputs. Firms mostly remained small and new establishments were formed to fill niches. The manufacturing process was inherently labor-intensive and reliant on a cadre of skilled and unskilled workers to manipulate the raw material into final, finished form. The advantages of this supply system, combined with the benefits of a labor force that was self-reproducing, resulted in low fixed costs. Although problems of coordination would eventually become disabling, nonetheless Britain, and other countries including Switzerland, built a globally competitive industry in the early nineteenth century based on just such a system of organization.

What unseated this remarkably flexible system was the emergence of mass production and the manufacture of volumes of products, all of more or less the same quality and capability (see *Regulation Theory in Geography*). In the early era of mass production, the American system became the model to emulate. Unfolding in fits and starts, this system has been characterized as 'the sequential series of operations carried out on successive special-purpose machines that produce interchangeable parts' (Hounshell 1984, p. 15). Implicit in this description was the move away from a dependence on skilled labor, a system based around fragmented and decentralized production, and highly variable, relatively high-cost goods (Hounshell 1984). Costs were not the only concern. Equally important were similarity of parts, interchangeability, and rapid production.

3. *The Emergence of Managerial Capitalism and Factories*

The transition to machine-based, vertically integrated, large-scale industrial manufacturing occurred over a long period of time. There are examples of metal-cutting machines assisting in the rough shaping of parts for final finishing as early as the middle of the eighteenth century. However, it was the middle of the nineteenth century before widespread implementation of this technological and organizational innovation occurred. The movement toward this form of production was neither pervasive nor persistent. There were many obstacles to overcome, not least the construction of the necessary machinery and the resolution of the inherent conflicts presented by machine-based output to established labor relations and hence to the production system itself (Lazonick 1991). Ultimately, the movement toward the American system caused turmoil throughout the entire system as

machine-made parts and large-scale assembly yielded an unprecedented volume of goods. Many problems were associated with the flow of goods, and spilled over into the acquisition of consistent inputs and, equally important, the development of an effective means of distribution. Goods were no longer produced with a proximate market in mind. Indeed, the international supply of foreign markets became an increasingly important part of firm strategy. To produce at large scale necessarily meant distribution at large scale, ushering in the move toward managerial capitalism.

The evolution toward machine-based manufacturing and the ensuing complementary innovations in scale and scope of production, distribution, and management of the modern industrial enterprise arose rather gradually, starting in the 1870s (Chandler 1991). The rise of the multiunit, capital-intensive corporation was the result of three investments. First, production levels had to be large enough to utilize economies of scale and scope of new technological innovations. Second, marketing and distribution capabilities had to be of a scale large enough to sell the goods produced by the new processes of production in the volume in which they were made. Third, a managerial hierarchy was needed to manage and coordinate the operations of large-scale production and distribution and to allocate resources for future production and distribution (Chandler 1991, p. 432). That this new paradigm first took hold primarily in the USA and Germany has much to do with the relatively late industrialization of the two countries compared with Britain, and hence the absence of a pre-existing labor process and managerial system to overcome. In the USA, the shortage of skilled labor placed a premium on developing machine-based alternatives to industrial production. Preceding the Civil War and strongly afterwards, American firms in a wide range of industries pursued mechanical solutions to the expansion of production to achieve economies of scale and serve burgeoning domestic and international markets.

The rise of the managerial corporation brought with it the integration of multiple functions, multiple units, multiple products, and all of the attendant physical and human resources. Although each unit had independent capabilities to execute and manage the enterprise, coordination was offered and accomplished by middle managers who were themselves managed by top officials of the corporation. Quite strikingly, the emergence of this mode of organization and coordination arose rather quickly at the end of the nineteenth and beginning of the twentieth centuries and was most observed in a subset of all industries comprising the US economy. These industries tended to be sectorally concentrated, to consist of firms that were vertically integrated within the same industry group, and to be characterized by high-volume, capital-intensive production processes. In other words, what linked these firms were product characteristics that yielded the need for a uniform managerial structure designed to confront and control problems of production and distribution while benefiting from the potential associated with specific sectoral characteristics.

The sectors and firms most successful in executing the full complement of attributes of the modern corporation were those that benefited from scale and scope linked to standardization of the production process. At the same time, however, there were other prerequisites to effective implementation, particularly relating to market control. To support the huge capital costs required to achieve economies of scale, a relatively few players had to exercise market power. The emergence of oligopolies in such sectors as oil, steel, and heavy machinery went hand-in-hand with the rise of the modern corporation. Control was both seized and achieved through investments in up- and downstream linkages, a capability that most smaller, more consumer-oriented goods producers could not make.

With the emergence of the managerial corporation, firms were now in a position to own and control resources and markets distributed across space (see *Regulation Theory in Geography*). The spatial organization of the national economy evolved from production for local consumption, to regional markets of production and consumption, and finally to national markets and corporate structures of spatially decentralized production systems. The ability to standardize products and produce at mass scale, twinned with the development of national brands and brand recognition, resulted in the creation of huge corporations that controlled vast business assets.

4. *The Rise of Collective Capitalism*

The model of market coordination embodied in managerial capitalism remained an effective organizing system until the 1960s. The lack of cut-throat international competition, and the presence therefore of stable markets, buoyant growth, and rising standards of living, provided hospitable conditions for the cultivation and persistence of the large, vertically integrated corporation (see *Economic Restructuring: Geographic Aspects*; *Regulation Theory in Geography*). The seeds of turmoil that would come to destabilize this system of production, consumption, and market regulation were sown much earlier in the century, however, and lay dormant until after World War II. The challenge to this dominant system arose, ironically, as a new competitor—in this case Japan—coupled long-standing and under-appreciated competency in the production of goods with state-supported (yet private-sector-led) manufacture of new products capable of competing directly in the world's most sophisticated industries and its most developed markets (Conroy 1995). The national pursuit of both import substitution and export-led development policies

provided infant industries with the protection required to achieve economies of scale and hence low overall costs. Accustomed to developing technological competency through technology transfer, the Japanese successfully adopted and adapted major innovations, primarily from the USA, to take the lead in key consumer and producer products over the postwar period.

The emergence of the Japanese as effective competitors in many consumer and producer product markets dates back to before World War II. In the last three decades of the nineteenth century, Japan had begun to develop the competency to produce tradable goods, primarily through the purchase of foreign know-how and technology. By the 1920s, in several key sectors, Japan was second if not first in the world in terms of exports (Glasmeier et al. 1993). Although cut off from international trade in the 1930s, Japan continued to build up its productive capacity until just before the declaration of war. This time the emphasis was placed on heavy and chemical industries.

The emergence of collective capitalism as a distinct mode of organization is both a creature of history and a more modern institutional form. At the end of World War II, Japan had little choice but to pursue an export-led development strategy (McCraw 1986, Samuels 1994). The country lacked resources and had a small land base with a large population. In order to stave off poverty and economic despair, the nation had to develop export competencies in a variety of sectors. For reasons that can only be understood retrospectively, the Japanese did not choose to pursue products that were labor-intensive despite its large labor force; rather, it moved rapidly into capital-intensive production as part of economic recovery in the 1950s.

The second major stimulant in the formation of collective capitalism was in part a holdover of the structure of the national economy that had developed since the beginning of the twentieth century. Japan's economy is comprised of interlocking business organizations that enjoy both vertical and horizontal affiliations. Furthermore, the system of integration reaches down to the level of select blue-collar workers in first-tier firms who are offered lifetime employment in return for a life of dedication to the firm. In this way, guaranteed their jobs, workers willingly accept technological change as a part of labor–management relations.

Although many distinctions are attributable to the Japanese system, of particular importance is the role of the enterprise groups—networks of firms—that form the modern-day extensions of the *zaibatsu*, the family-controlled businesses that led the development of the Japanese economy from the end of the nineteenth century until the occupation after World War II (see *Technology Districts*). These enterprise groups consist of a central business firm around which reside vast numbers of both affiliate and non-affiliate subcontractors. This disintegrated system allows the core firm to enjoy the benefits of vertical integration, especially in terms of financing, technology transfer, and internal markets, without the disadvantages of ownership and bureaucratic control. The core firm is able to ensure quality, quantity, and timely and effective delivery of parts and services due to long-term contracts and investments made by the core firm into subcontracted operations (Lazonick 1991).

The emergence of collective capitalism is essential in understanding the industrial geography of the late twentieth century. Along with recognition of differences in the organization of industry, today there is an increasing interest in how culture helps to shape economic space (see *Economic Geography*). Throughout the life of an industry, different conventions and rules of behavior form the basis for collective action. Organizational behavior and prevailing institutional norms structure the range of possibilities of firms and regions. This interaction forms the industrial landscape. Today's industrial landscape is thus the sum of long-evolving processes, the residue of which is reflected in the structure and formation of business enterprises, industrial clusters, and systems of industrial activity. Contemporary concerns increasingly focus on elucidating the importance of culture as a determinant of corporate behavior (Schoenberger 1997). In this context, culture is defined as the rules and practices, identities and aspirations of individuals at the most intimate level. There is increasing interest in the ways in which culture manifests within people, which in turn shapes firms within regions. In this way, industrial geography is not just the study of a physical place *per se*, but is also an examination of the reservoir of experiences and actions that come to represent actors in a locality and its economy.

See also: Human–Environment Relationships; Industrial Ecology

Bibliography

Chandler A 1991 *Big Firms and the Wealth of Nations*. Harvard University Press, Cambridge, MA

Conroy M 1995 Industrial strategies, the newly industrializing economies, and new international trade theory in Latin America. *Environment and Planning A: Commentary* **27**(1): 1–10

Glasmeier A K 1990 A missing link: The relationship between distribution and industrial complex formation. *Entrepreneurship and Regional Development* **2**: 315–33

Glasmeier A K, Thompson J, Kays A 1993 The geography of trade policy: Trade regimes and location decisions in the textile and apparel complex. *Transactions of British Geographers* **18**: 19–35

Harrington J W, Warf B 1995 *Industrial Location: Principles, Practice and Policy*. Routledge, New York

Hoover E 1948 *The Location of Economic Activity*. McGraw Hill, New York

Hounshell D A 1984 *From the American System to Mass Production, 1800–1932. The Development of Manufacturing*

Technology in the United States. Johns Hopkins University Press, Baltimore
Landes D 1969 *The Unbound Prometheus*. Cambridge University Press, London
Lazonick W 1991 *Business Organization and the Myth of the Market Economy*. Cambridge University Press, New York
Lösch A 1954 *The Economics of Location*, 2nd edn. Yale University Press, New Haven, CT
Marshall A 1997 *Principles of Economics*. Prometheus Books, New York
McCraw T 1986 *America Versus Japan*. Harvard Business School, Boston
Rosenberg N, Birdzell L 1986 *How the West Grew Rich*. Basic Books, New York
Samuels R 1994 *Rich Nation, Strong Army, National Security and the Technological Transformation of Japan*, 1st edn. Cornell University Press, Ithaca, NY
Schoenberger E 1997 *The Cultural Crisis of the Firm*. Blackwell, Cambridge, MA
Storper M, Walker R 1989 *The Capitalist Imperative*. Blackwell, Oxford UK
Weber A 1971 *The Theory of the Location of Industries*. Atheneum, New York

A. K. Glasmeier

Industrial Metabolism

The word 'metabolism' comes from the Greek, meaning to change or transform. The concept in biology describes the combined effect of numerous biochemical reactions that convert the materials in a living cell to provide for its growth and maintenance. While the concept of metabolism is uniformly accepted at the cellular level and for individual organisms, its application to larger systems of living organisms or human societies remains underdeveloped. The term 'industrial metabolism' derives from the notion that industrial economies, like biological organisms or natural ecosystems, can be understood as systems for material transformation with distinct metabolic pathways that evolve over time. Industrial metabolism offers a useful perspective for the study of collective human behavior and its effect on the materials cycles that comprise the workings of the natural environment.

Social analyses have considered industrial production and resource-use statistics as social indicators since the dawn of the Industrial Revolution. In the decades since the 1960s, industrial metabolism has been used as a paradigm to describe the exchange of materials among industrial operations in a way analogous to the description of material and energy balances in natural ecological systems. This metabolic analysis contributes to the formulation of technological, economic, and policy instruments aimed at addressing social environmental concerns. In the context of modern social and political awareness about environmental quality the concept offers an analytical framework that is practically motivated to improve the quality of the natural environment through changes in industrial and social norms.

Industrial metabolism examines the exchange of materials and energy between human society and nature at all spatial scales, applying the same analytical parameters to describe material fluxes within and between natural and human systems. At the largest scale this includes study of anthropogenic contributions to atmospheric concentrations of trace gases and the flow of excess nutrients from agricultural activities to water bodies. At smaller scales, industrial metabolism addresses the use of resources in the human economy through the development of resource-accounting frameworks for political and economic entities as well as life-cycle analyses of the materials used in the manufacture of industrial and consumer products.

Emphasis is placed on following the 'material flow' from its origins (e.g., mined ore) through the industrial processes used for its transformation, to the products that contain it, and finally to its disposal or re-entry into the industrial system. At its most comprehensive, industrial metabolism links to human and ecosystem metabolism, for instance by considering the life cycle of metals and organic chemicals in plant tissue and the human diet. The information is used to construct a 'materials balance' of a system and draw conclusions regarding the actions needed to improve the environmental character of its metabolism. Possible indicated actions might include industrial process change, product substitution, consumer education, government regulation, development of enabling technologies, installation of new resource recovery capacity, and economic incentives by government (see *Global Environmental Change: Human Dimensions*).

To determine existing and potential metabolic pathways, industrial metabolism considers the social dynamics that influence the integrity of the natural environment. To establish quantitative relationships between material flows and social dynamics, industrial metabolism compares indices of industrial production and social consumption to social and demographic variables such as population, gender, age, land use, and economic activity. To aid in the formulation of public policy, industrial metabolism bases its findings in the context of government institutions, economic incentives, and cultural biases that influence the flow of physical resources and their impact on the natural environment.

1. *Natural History of Chemistry*

One main thrust of research in industrial metabolism compares the natural history of the basic biogeochemical reactions observed in planetary evolution and contemporary ecosystems with those employed by

industrial society. By studying the basic chemical processes observed in nature for achieving beneficial transformations of materials, this inquiry seeks to identify industrial processes that minimally perturb global biogeochemical cycles and least interfere with the services provided to humans by nature.

From a historical perspective, the Earth represents an evolved system for recycling materials. Over geological periods the planet has allowed for the biogeochemical feedback necessary to arrive at a state hospitable to human life over the majority of the globe. Characteristic of this evolution is the natural system's ability to find uses for discarded by-products. For instance, at the planetary scale we find the development of organisms using photosynthesis to produce energy. The life-forms employing this process came to dominate a planet rich in atmospheric oxygen, the waste product of previous life. This system attribute, a product of evolution, is also evident over shorter timescales in natural ecosystems. For these systems, scientists describe a complex web of life supported by the wastes from animals and plants, identifying the significance of those wastes as providing the basic nutrients for regeneration.

Industrial metabolism notes that in selected instances the need to find uses for waste products drives the evolution of processes used in human industry. For example, the search for uses for unwanted coal tar led to the birth of the synthetic dye industry at the beginning of the twentieth century in Western Europe. Natural gas, a leading energy source at the end of the twentieth century, was once 'flared' at wellheads before its value was recognized for energy and commercial products from nitrogen fertilizer to synthetic rubber. Chlorine entered the human environment initially as a by-product from the manufacture of sodium chemicals and eventually became the far more marketable product. Following the evolutionary principle evident in natural, as well as human, systems, industrial metabolism seeks to reveal opportunities for using waste materials from contemporary industrial systems (see *Sustainability Transition: Human–Environment Relationship*).

Industrial metabolism stipulates that resource use and waste generation can be avoided if human industry can learn from the design models found in nature and overcome the barriers to replicating natural systems for material transformation in human industry (see *Industrial Ecology*). One of the most basic differences between natural and human-made systems derives from the fact that natural systems are driven by solar energy in its different forms, which contrasts with human energy production as a relatively diffuse spatial source. The intensive use of energy to generate conditions of high pressure and temperature for human production processes further distinguishes human from natural metabolic processes that typically occur under conditions of ambient pressure and temperature. Human-made material transformations also generally follow a series of linear discrete transformations (i.e., unit operations) in contrast to the more continuous processes found in nature. In addition to learning from the design of natural systems, industrial metabolism seeks to identify, and engineer, natural processes to replace commercial ones and thus reduce human resource use and waste generation (see *Greening of Technology and Ecotechnology*).

2. *Resource Accounting*

A complementary thrust of industrial metabolism research explores the actual physical flow of materials within industries and regions. Materials-flow analyses rely on the law of conservation of mass (i.e., no new mass is created or destroyed in a closed system) to provide a complete materials balance equating the quantities identified as inputs with those labeled as outputs. Employing the materials balance principle allows for estimation of material flows in industrial and consumer products as well as in by-products and wastes even when data are missing. The analysis may include material flows apparent to the public, for example in household wastes, as well as the hidden material flows occurring in mines, quarries, power plants, factories, farms, and disposal sites.

Because both large- and small-scale flows have an impact on the environment, industrial metabolism examines bulk materials, typically measured in millions of tons, as well as toxic materials measured in micrograms. Large-scale flows such as hydrocarbon fuels used to supply energy, and fixed nitrogen used for food production, have an impact on global biogeochemical cycles that govern climate and ecosystems. Smaller volume flows that enter the human food chain may pose chronic or acute threats to human health and threaten the biological balance found in natural ecosystems.

The industrial metabolism of any material must be determined within the boundaries of a defined system. Once within system boundaries, flows must also cross an internal boundary in moving from 'nature' to the 'human' domain in order to be included in the account. Geographically based analyses examine the industrial metabolism of material sources and sinks formed by industrial, agricultural, and urban activity, and the surrounding environment. Analysts can also map material flows for individual facilities or entire sectors, provided that they consistently define how to evaluate indirect material flows occurring outside the sector under study. Finally, the material itself may provide the boundary, if the effort calls for improved understanding of its metabolism independent of other associated flows.

Criteria for including materials in the metabolic analysis range from narrowly restrictive to broadly inclusive. Some account only for those material flows associated directly with economic transactions. Other

frameworks include hidden material flows such as mine tailings from metal production, overburden removed to access coal seams, and eroded soil from agricultural land. Still other accounting frameworks tabulate the translocation of dirt and rock for infrastructure such as road building and harbor dredging.

3. Future Development

Several factors motivate contemporary interest in industrial metabolism as a framework for analyzing the environmental impact of human economies. Research aimed at developing operating industrial economies that do not damage the natural environment, benefit from the study of how natural systems are able to operate without permanently depleting natural resources or damaging life-support systems. Resource accounting offers a compact and transparent method to convey environmental information to government officials, resource managers, and industrial operators whose decisions influence environmental quality.

Challenges to industrial metabolism include the need to refine methods for materials-flow analysis to counter limits in the ability of government and industry to monitor all environmentally important materials streams. Far more work needs to be done in the realm of environmental science to establish causative relationships between human emissions and damage to the environment. Finally, the inescapable influence of specific geographical and cultural features on human–environment interactions requires models of individual and collective behavior to describe and improve environmental quality at the local level where humans reside and interact with one another.

See also: Ecology, Political; Environment and Development; Environment Regulation: Legal Aspects; Environmental Change and State Response; Environmental Policy; Human–Environment Relationship: Indicators; Sustainable Development

Bibliography

Ayres R U, Simonis U E (eds.) 1992 *Industrial Metabolism*. United Nations University Press, Tokyo

Ayres R U, Schlesinger W H, Socolow R 1994 Human impacts on the carbon and nitrogen cycles. In: Socolow R, Andrews C, Berkhout F, Thomas V (eds.) *Industrial Ecology and Global Change*. Cambridge University Press, New York, pp. 121–56

Fischer-Kowalski M 1998a Society's metabolism: The intellectual history of materials flow analysis, part 1: 1860–1970. *Journal of Industrial Ecology* **2**: 61–78

Fisher-Kowalski M 1998b Society's metabolism: The intellectual history of materials flow analysis, part 2: 1970–1988. *Journal of Industrial Ecology* **2**: 107–36

Frosch R A 1996 Toward the end of waste: Reflections on a new ecology for industry. *Daedalus* **125**: 199–212

Stigliani W M, Anderberg S 1992 *Industrial Metabolism at the Regional Level: The Rhine Basin*. International Institute for Applied Systems Analysis, WP-92-10, Laxenburg, Austria

I. K. Wernick

Industrial Policy

Industrial policy is an approach to economic policy in which the state gives direct support to particular firms and industries. Since the modern state has a pervasive influence on economic activity, virtually every industry is influenced by its actions. Even general codes typically have a differential impact upon particular firms and industries. There is thus a sense in which every industry is shaped by state action and when the state acts in a deliberate and purposive way to affect the fate of an industry, either on its own behalf or in response to pressures from the industry itself, one can speak of industrial policy in the generic sense of the term. The term might be extended still further to refer to cases where a lead enterprise, and/or an employer association, pursues a strategy for the industry as a whole. But we generally reserve the term for what might be called industrial policy in the large. This is an approach in which particular firms and industrials are supported in pursuit of *national* economic policy goals. Such policy is generally conceived in terms of a broader plan, or strategy, of national economic development. Instruments of support range from tariff protection and import restrictions, tax concessions, subsidies, and targeted services to direct intervention through institutions of the state itself, including the state's own research and development laboratories, state schools and universities and state-owned enterprises. Indeed, industrial policy is a major justification for the state to undertake these activities directly on its own account.

Industrial policy in this larger sense was practiced widely in the early postwar decades, although the term itself seems to date from the 1970s, when this approach to national economic development became the subject of a widespread debate. In more recent decades industrial policy has fallen out of favor, so much so that the term has taken on a pejorative connotation. Its proponents have tended to use various euphemisms and circumlocutions, although one is hard pressed to point to a particular word or phrase that has taken its place. These changes reflect a major shift in the intellectual and political climate of the late postwar period. In some ways, the climate has now shifted back once again in favor of an industrial policy, although one which is understood and practiced very differently from that of the early postwar decades.

In the first few decades after World War II, among the advanced developed countries the leading practitioners of industrial policy were France and Japan. In France, the postwar policy was associated with the Commissariat du Plan (The Planning Commission)

and a process that became known as indicative planning. The term 'indicative' contrasted with the directive planning of Soviet-bloc countries in Eastern Europe and involved a process of consensus-building through discussion and debate within committees organized by the Commissariat. A number of these were industry committees and included representatives of government and administrative agencies as well as business and labor organizations. The idea was to reach a broad-based but also fairly specific consensus on the direction in which each industry should be moving and the kinds of adjustments in industrial organization and state policy that would facilitate that movement. The French state has also been very active in promoting new industries. It invested heavily in information technology, for example, through Honeywell–Bull and, along with other European countries, in Airbus (Cohen 1977, Stoffaes, 1983–84).

In Japan, industrial policy dates back to the nineteenth century and the earliest stages of the country's industrialization. Indeed, it was the state that initiated the development process, itself creating and operating particular production facilities and then subsidizing, but also directing, private enterprise to do so. In the postwar period, policy has operated through the Ministry of International Trade and Industry (MITI). MITI has controlled the flow of all foreign technology and capital into the country and has sought to direct it so as to maximize its contribution not only to economic growth but also to national autonomy and economic independence. It has also sought to direct the evolution of strategically important industrial sectors internally in a detailed way, establishing specific production targets and imposing them upon the industry, dictating output and market share for particular firms, and sometimes even the production methods and organizational structures through which those targets are pursued. Toward this end, it has utilized virtually every instrument available to the state, including direct intervention in the economy to organize industrial cartels (Johnson 1982).

The classic work on industrial policy in the immediate postwar period is Andrew Shonfield's *Modern Capitalism: The Changing Balance of Public and Private Power* (Shonfield 1965). Shonfield saw developments in France and Japan as prototypical of the approach to policy emergent in a more or less explicit and self-conscious form in all advanced industrial countries. Shonfield traces these developments in the major countries of Europe and the USA.

The impetus for—and practice of—industrial policy in the developing world has been a little different. It is generally associated with the process of economic development itself and the conviction that development will not occur spontaneously on its own. In the immediate postwar period, industrial policy was given particular impetus by *dependency theory* (Prebisch 1963). That theory argued that without a conscious and deliberate effort to protect and develop its own industries, the third world would be relegated to a permanent position as supplier of primary products to the industrial world. It led to very high protective tariffs and quantitative restrictions designed to promote import substitution. In automobiles, countries with very large domestic markets, such as Brazil, Mexico, and India, have also imposed local content requirements upon foreign investors. Countries with significant military purchases—China, Turkey, and Korea, for example—have made foreign direct investment in local parts production a condition for purchases of military aircraft. In recent years, the case for industrial policy as a tool for development has been reformulated in terms of theories of learning.

Three strands of thought worked together to create an intellectual climate favorable to industrial policy in the early postwar decades. First, an extensive literature composed of contributions from different scholarly disciplines converged upon the nature of industrial society and the direction in which it was evolving over time. Undoubtedly, the dominant figure in this literature was Karl Marx, but a number of other scholars who were politically antagonistic to Marx and disagreed with him on a variety of specific points nonetheless shared and propagated a similar vision. These included Max Weber, Joseph Schumpeter, Alexander Gerschenkron, John Kenneth Galbraith, and Walter Rostow.

The essential element of this vision was that early capitalism might have been constituted by small entrepreneurial firms which composed an economic system that could be coordinated and directed by the market, but that in late capitalism the economy was coming to be dominated by very large, extremely capital-intensive enterprises. The canonical industries for most of these observers were automobiles, consumer durables, steel, rubber, and plate glass. There were, to be sure, important differences between these various analysts, but the differences in some ways served to underscore the similarities and created the sense that the broad contours of industrial society were unproblematic. This made industrial planning seem unproblematic as well.

Second, these theories of industrial society and late capitalism had the definite implication that markets would be increasingly less effective instruments for managing the economy over time and that some other form of management involving planning and the direct allocation of resources by individual enterprises would take its place. This created ample space for arguments that enterprises themselves would take too narrow a view and that specific government intervention in the national interest would be necessary to achieve an optimal result. It was a small step from this to industrial policy.

The third element of the intellectual climate conducive to industrial policy was Keynesian economics. John Maynard Keynes' *General Theory*, published just before World War II, came to dominate economic

thinking in the postwar period. Essential to the new analytical structure that Keynes created was the assertion that the market would not stabilize the macroeconomy and sustain full employment automatically on its own. Keynes' own explanation for this was that certain key prices, particularly the prices of labor and capital, were rigid and did not respond to competitive forces. This suggested the possibility that prices *in general* were unable to direct the economy and that government action was required in their place.

Keynesianism was absorbed into economics in the USA in a way that sidestepped this particular implication. In what was known as the neoclassical synthesis, the leading American Keynesians combined the new macroeconomics with a view in which markets continued to function effectively to govern the economy at the microeconomic level and to give individuals and industries their direction. They did this, by and large, by accepting the basic macroeconomic conclusion that price variation would not function automatically to maintain full employment, but they did not explore too closely the particular mechanisms that prevented this from happening. Most of their scientific work—for which the leading Keynesian economists later received Nobel prizes—operated within the traditional assumptions of a competitive market economy. Hence this created a predisposition within the discipline to find a way of reconciling the contradictions between macro- and microeconomics along neoclassical lines. But Keynes definitely broke new ground and increased the plausibility of approaches to understanding and managing the economy predicated on the notion that markets could not be counted on even for the management of individual firms and industries.

The disenchantment with industrial policy as an approach to national economic management in recent decades has been part of a broader process of government withdrawal from active involvement in microeconomic activity. It is associated with the political ascendancy of Margaret Thatcher in Great Britain and Ronald Reagan in the USA. But a turn toward the market was already apparent in the USA under the Carter Administration in the 1970s, when the process of deregulation of transport and telecommunications was begun. Among the most aggressive practitioners of industrial policy, the retreat came later and was less complete. In France, for example, policy shifted from promoting 'national champions' in the early postwar decades to a focus on 'lame ducks' in the 1980s, and only retreated further as the budgetary implications of protecting declining industries became overwhelming (Berger 1981). Japan in some sense has never abandoned its commitment to industrial policy.

The demise of industrial policy was, however, not simply political. It also reflected a radical change in those elements of the intellectual climate that had made it attractive and plausible in the earlier period. An important part of that change took place within the discipline of economics and centered on the Keynesian heritage. Keynes was not exactly abandoned: the basic apparatus of macroeconomic analysis that he had created was retained. But the microeconomic foundations were substantially rebuilt; the basic understanding of why the macroeconomy might not be able to maintain a full employment equilibrium was revised in a way that made it less dependent on rigidities in wages and prices (Mankiw 1990). As a consequence, arguments about the ability of prices to regulate and direct the development of individual industries and particular firms became more plausible. The economics profession increasingly came to see government intervention as an interference with the market rather than a remedy for market failure.

At the same time, in the 1980s and 1990s, the texture of the economy shifted dramatically in way that called into question the prevailing assumptions about the underlying nature of industrial society and the direction in which it was moving. The service sector became increasingly important and manufacturing less central. Some very large companies that had seemed permanent fixtures on the industrial landscape experienced serious difficulties and came close to bankruptcy. A few actually disappeared from the industrial scene. Virtually all large firms began to experiment with new organizational forms and managerial approaches that called into question the notion of bureaucratic management. Large companies also began to limit the range of activities in which they were directly engaged: they tended to focus on their 'core competencies' and to purchase ancillary services from subcontractors and external collaborators. New technologies emerged in communication, information processing, and biotechnology whose implications were often surprising and difficult to predict.

These new leading sectors were populated with many small, entrepreneurial companies. As a result, small enterprises in competitive industries suddenly became plausible agents of future development. In the past, small firms had been associated with traditional industries such as garments and shoes, and had thus been viewed as vestiges of the pre-industrial era (Piore 1984, 1994). A similar reversal in the expected pattern of development took place with respect to the role of education. In the early postwar decades, the wage premium for education had declined; thus seeming to confirm an expectation that originated with Marx that the skill requirements of jobs would decrease as industrial development proceeded. In the 1980s and 1990s, the education premium increased, and it began to look like the skill requirements of jobs were actually rising as the economy evolved over time.

It is symptomatic of the changed climate that 'industrial society' as a term used to characterize contemporary civilization has fallen into disuse. It has been replaced in our vocabulary by the older notion of

'capitalism' or by the term 'postindustrial society.' The second term is particularly telling of the emergent intellectual climate: it defines the new era by what it is not, but says virtually nothing about what it is. It thus seems to deny the notion that we have any idea where we are headed. In this sense, it seems to foreclose the possibility of planning upon which the older notion of industrial policy was premised.

The other term replacing 'industrial society,' i.e., 'capitalism,' has a much more definite implication: it points toward the market as the guiding economic institution. Given the dichotomy between markets and hierarchies in which the debate about industrial policy, and industrial organization more broadly, has come to be conducted, the changes in the institutional structures of advanced industrial nations certainly seemed to point in this direction. The market seemed to be the governing institution in coordinating the small entrepreneurial firms which were emerging in new industries as well as coordinating the relationships between the newly streamlined large companies and their outside contractors and collaborators in older sectors of the economy. Central direction and control through industrial policy no longer seemed necessary.

Even as these intellectual currents came to dominate the political and scholarly landscape, a new set of understandings was developing which provided the foundations for industrial policy, albeit of a somewhat different kind than that which prevailed in the early postwar period. Here again one can distinguish between developments within economics as a discipline and trends within social and policy science more broadly.

Within economics itself, the new climate has been shaped by endogenous growth theory and the 'new theory of international trade' which emerged together in the 1980s (Krugman 1992, Romer 1989a, Romer and Rivera-Batiz 1991). Whereas the older theories of growth and trade emphasized comparative advantage and 'natural' or 'innate' differences among countries, the new theories focused on the processes through which competitive advantages were created. They gave a central role to economies of scale and to multiple equilibria, i.e., alternative paths of development all of which are consistent with a country's initial endowment of resources. The new theories thus raised questions about how industries grew to scale and countries shifted from lower level equilibrium growth paths to paths consistent with high incomes and rapid development. The literature focused initially on how this happened 'naturally' in the absence of particular public policies. But it opened an intellectual space for policy. In many ways, it was reminiscent of the older infant industry arguments for protection from international competition.

But the new theories have not led to a revival of interest in policies of this kind, at least within the economics profession itself. Instead, economists have focused on the economies of agglomeration, why firms within an industry tend to locate together in a particular place and how this fosters economies of scale. Another strand in the endogenous growth theory is concerned with the role of research and development activities (Romer 1989b). Why economics has moved in this direction and why the new theories have not led to a reappraisal of the case for free trade is not altogether clear. But one factor has been the politics of the profession itself. Economists as a group have constituted themselves as a political force against protectionism. Paul Krugman, one of the leading figures in both endogenous growth theory and trade, has been particularly ferocious in attacking those who would use the new theories to argue against free trade—whether they operate from within the economics profession or the wider political arena (Krugman 1992).

The second factor that has created a renewed space for industrial policy is the process of globalization itself. In the increasingly large and open international marketplace, individual countries become bit players, more like a single firm than a sovereign entity. It makes less and less sense to think of a national economy as reproducing within itself a complete industrial structure or even as being composed of particular industries. Just as an enterprise is located in a particular part of a particular industry, a country can now think of its industrial locations in this way as well. The institutions governing the emergent trade system have reinforced this analogy. They prohibit protectionism. They also prohibit government subsidies, both direct and in the form of targeted tax breaks. The range of instruments available to governments in promoting particular industries is thus restricted to education, research and development, and internal institutional structures of property rights, incentives, and the like. The menu of possible policies is thus increasingly similar to that of a business enterprise as opposed to a sovereign nation.

The field of management science which is focused on how firms choose their location within the marketplace is called *business strategy*, and as a country, or a region within a country, looks more and more like a firm, industrial policy is readily reconceptualized as a branch of that field. Michael Porter, who is the leading figure in the development of this business strategy both as a field of scholarship and as a domain of management education, has himself developed this analogy in a text, *The Competitive Advantage of Nations*, which is increasingly influential in regional planning and development studies (Porter 1990).

Finally, given the fact that governments are now limited by international treaties in the kinds of policies they can pursue to support particular industries, and forced to think instead in terms of longer range structural developments, the institutional environment in which industries operate must be counted as a central component of the 'new' industrial policy. The interest in, and concern about, institutions as a

determining factor in the fate of particular national economies has been reinforced by the experience in the countries of the former Soviet Bloc and the difficulties they have had in the transition from socialism. This experience has heightened the sense that the market is itself a social creation and does not automatically emerge on its own, simply because it is no longer restricted by hostile government policies and regulations. But it has also brought back to the fore the question of what the limits of the market are.

In thinking about this last question, two analytical constructs have become increasingly central to planners and policy makers: *supply chains* and *industrial districts*. Both constitute structures in which relations among firms involve direct interactions, as opposed to arms-length relationships mediated by the market. The nature of those interactions and the role they play are, however, very different in the two cases.

Supply chains envisage firms as being linked together in a vertical chain of input–output relationships stretching from raw materials at one end to the distribution of the final product to the consumer at the other end (Gereffi and Korseniewicz 1994). But the relationships are so tightly coupled that they require administrative control and interaction and cannot be coordinated by the prices and trade relationships generated by a competitive marketplace. In the past, the relationships along a supply chain required close physical proximity and were often incorporated within a single enterprise. They are now increasingly spread out geographically and involve a series of distinct firms which are independently owned and operated. Such supply chains nonetheless tend to be controlled and directed by firms at a particular, strategic position: assemblers in the automobile industry, for example, or retailers in the clothing industry. Industrial policy, for developing countries, can then be viewed as a strategy for attracting particular parts of the supply chain. The process of development comes to be conceived as one of moving along the chain over time and of gaining, or retaining, control over strategic positions along it. It is a way of thinking about how countries must move to maintain economic prosperity when they no longer have direct control over their costs through protection and subsidies. A particular position on the chain comes to be viewed as a learning opportunity that will enable the country to move upward as its costs rise and it loses its initial advantage to other parts of the world. Policy instruments over which governments retain control, such as education, training and research, and development, are deployed in a strategy conceived in these terms.

The industrial districts constitute agglomerations of independent enterprises that, like the links in a supply chain, interact directly with each other (rather than indirectly in response to price signals generated by a competitive market). But the particular relationships are fluid, the firms are on a more equal footing with respect to power and responsibility, are generally in the same industry and stand at essentially the same place in terms of input–output relation (Schmitz and Nadvi 1999). Examples include the financial districts of New York, London, or Tokyo or the garment districts of Paris, Milan, and New York. The concept of an industrial district originates with Alfred Marshall who studied such districts in late nineteenth/early twentieth century England, and attributed them to external economies of collocation. This industrial form is particularly widespread in central Italy, where it can be found in a range of industries from high fashion clothing and shoes to specialized machinery and motorbikes (Pyke et al. 1990). It has become a focus of both the endogenous growth theory in economics and the strategic planning literature growing out of business strategy. As a fulcrum for industrial policy, it tends to shift the focus radically from the role of the government in shaping the economic environment in which firms operate to the social climate which enables firms to cooperate with each other and to the potential role of government in fostering that climate. In this sense, it opens industrial policy to insights drawn from sociology and anthropology (Putnam 1993).

See also: Corporatism; Policy History: State and Economy; Policy Process: Business Participation; Regulation and Administration; Regulation, Economic Theory of; Regulation Theory in Geography.

Bibliography

Amsden A H 1989 *Asia's Next Giant: South Korea and Late Industrialization*. Oxford University Press, New York

Berger S 1981 Lame ducks and national champions: Industrial policy in the Fifth Republic. In: Andrews W G, Hoffmann S (eds.) *The Fifth Republic at Twenty*. State University of New York Press, Albany, NY, pp. 292–310

Cohen S S 1977 *Modern Capitalist Planning: The French Model*. University of California Press, Berkeley, CA

Galbraith J K 1952 *American Capitalism: The Concept of Countervailing Power*. Houghton Mifflin, Boston

Galbraith J K 1972 *The New Industrial State*. New American Library, New York

Gereffi G, Korseniewicz M 1994 *Commodity Chains and Global Capitalism*. Praeger, Westport, CT

Gerschenkron A 1962 *Economic Backwardness in Historical Perspective: A Book of Essays*. Belknap Press of Harvard University Press, Cambridge, MA

Johnson C A 1982 *MITI and the Japanese Miracle : The Growth of Industrial Policy, 1925–1975*. Stanford University Press, Stanford, CA

Krugman P R, Obstfeld M 1987 *International Economics: Theory and Policy*. Little, Brown, Glenview, IL

Krugman P R (ed.) 1992 *Strategic Trade Policy and the New International Economics*. MIT Press, Cambridge, MA

Mankiw G 1990 A quick refresher course in macroeconomics. *Journal of Economic Literature* **28**(4): 1645–60

Organization for Economic Co-operation and Development (OECD) 1971 *The Industrial Policies of 14 Member Countries*. OECD, Paris

Piore M J 1984 *The Second Industrial Divide: Possibilities for Prosperity*. Basic Books, New York
Piore M J 1994 Corporate reform in American manufacturing and the challenge to economic theory. In: Allen T J, Scott Morton M S (eds.) *Information Technology and the Corporation of the 1990s: Research Studies*. Oxford University Press, New York
Porter M E 1980 *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. Free Press, New York
Porter M E 1990 *The Competitive Advantage of Nations*. Free Press, New York
Prebisch R 1963 *Towards a Dynamic Development Policy for Latin America*. United Nations, New York
Putnam R D 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ
Pyke F, Becattini G, Sengenberger W (eds.) 1990 *Industrial Districts and Interfirm Cooperation in Italy*. International Institute for Labour Studies, Geneva, Switzerland
Romer P M 1989a Increasing returns and new developments in the theory of growth. Working Paper 3098. National Bureau of Economic Research, Cambridge, MA
Romer P M 1989b Endogenous technological change. Working Paper 3210. National Bureau of Economic Research, Cambridge, MA
Romer P M, Rivera-Batiz L 1991 International trade with endogenous technological change. Working Paper 3594. National Bureau of Economic Research, Cambridge, MA
Rostow W W 1990 *The Stages of Economic Growth: A Non-communist Manifesto*. Cambridge University Press, Cambridge, UK
Samuels R 1987 *The Business of the Japanese State: Energy Markets in Comparative and Historical Perspective*. Cornell University Press, Ithaca, NY
Schmitz H, Nadvi K 1999 Clustering and industrialization: Introduction. *World Development* **27**(9): 1503–14
Schumpeter J A 1950 *Capitalism, Socialism, and Democracy*. Harper & Row, New York
Shonfield A 1965 *Modern Capitalism: The Changing Balance of Public and Private Power*. Oxford University Press, London
Stoffaes C 1983–84 Politique Industrielle. Les Cours de Droit, Institute d'Etudes Politiques de Paris, Paris

M. J. Piore

Industrial Relations and Collective Bargaining

The field of industrial relations generally is defined as the employment relationship and the rules that govern it. In practice, the field tended until the 1980s to be equated with collective means of governing the employment relationship, in particular collective bargaining, CB (which may be defined as the process of negotiation of the terms and conditions of employment between an employing organization and its employees, organized as a group in a trade union or similar body) (see *Employers' Associations*; *Labor Unions*). The emergence of human resource management (HRM) challenged this easy assumption and led to a debate, particularly in the US and the UK, as to whether industrial relations was not only a mere subset of HRM but also one of declining importance as the role of unions was weakening. In the present writer's view, 'industrial relations' is a sufficiently flexible concept to embrace HRM, but for the purposes of this article the term will be delimited to the collective aspects of employment. Even with this limited definition, the field remains a key aspect of the ways in which working life is organized, not least because CB remains well-established in many countries; moreover, different modes of organizing industrial relations have important implications for equality, social justice, and economic performance.

1. *Origins and Intellectual Focus*

The term CB was coined by the British scholar Beatrice Webb in 1891 and developed in her joint work with Sidney Webb, *Industrial Democracy* (first published in 1897). They explained CB in terms of an evolution from a purely individual and temporary agreement between an employer and a worker (what economists now call a spot price), through a 'common rule' established with a group of workers on an informal basis, to a genuine collective bargain which represented a fixed, agreed set of rules between an employer (or possibly a group of employers) and a recognized body, usually a trade union, representing employees. At about the same time, CB was endorsed by the British government as the preferred way of managing the 'labor problem,' a public policy which survived largely intact until 1979. The other major country to see a similar focus on CB was the United States. During the early years of the twentieth century liberal and progressive reformers responded to the severe and often violent labor disputes of the time by arguing that CB provided a more mature way of resolving differences. The term 'industrial relations' first received wide currency through the Commission on Industrial Relations (1912), with the first academic departments using the name being established during the 1920s, when there were several experiments, notably in sectors such as clothing, to develop mutual acceptance between employers and trade unions (Kaufman 1993). The key public endorsement of CB came in the National Labor Relations (Wagner) Act of 1935 which famously declared in its preamble that CB was the best means not only of reducing strikes but also of handling all aspects of industrial conflict.

The key components of this Anglo–American view of CB were the following. First, it was voluntary: employers and unions engaged in it because they recognized the benefits, not because they were required to do so. Second, it assumed that the two parties were adversaries in that they represented the conflicting

interests of profits and wages. This did not imply that they were totally opposed, and indeed the third principle was that CB could provide a system of industrial governance. It would develop agreed rules for settling the key issues of employment such as pay, promotion, and discipline. But ultimately the authors of the rules reflected opposing positions. The fourth principle was that the parties accepted the legitimacy of the other. Finally, the resort to sanctions such as strikes and lockouts was not forbidden, but these were to be weapons of last resort after agreed procedures had been exhausted. It was hoped and expected that CB would develop as the established means of resolving disputes throughout the economy.

Academic analysis developed through the work of such scholars as John T. Dunlop in the US and Allan Flanders in the UK. In his book *Industrial Relations Systems* (1958), Dunlop developed the concept of a web of rules and argued that an industrial relations system comprised actors who establish sets of rules; the system develops a 'ideology,' or common set of ideas defining the role of the actors; and the system is shaped by its economic and social context. Flanders (1970) developed the idea of rules by identifying three weaknesses with the Webbs' model of CB. First, a collective bargain was not an individual bargain writ large but was different in character: it was not actually a bargain at all but was a process for the making of rules. Second, CB is a power relationship between organizations. Third, the Webbs saw bargaining as an economic matter whereas in fact it had a strong political component: it was about the management of the firm and the whole way in which the use of labor was regulated, not about a simple exchange of wages. CB had a 'managerial' as well as a 'market' aspect.

Such work helped to establish industrial relations as a distinct academic field, if not a discipline. A focus on the creation of rules distinguished it from economics, which concentrated simply on the price of labor. Interest in the institutions of work and their inner functioning demarcated it from sociology, which tended to see such institutions as reflections of social class. Industrial relations opened the black boxes of how the price of labor was established and how class relationships were actually organized.

The approach of Dunlop and Flanders also had weaknesses. As British critics argued, the focus on rule-making and formal institutions assumed a natural progression towards 'order' and neglected continuing relationships of power and inequality; there was a permanent conflict over the organization and control of work which insitutionalized CB could manage but not totally regulate (Hyman 1989). Such criticism merged with growing interest from outside conventional industrial relations in the labor process and conflict over the performance of the work task. The absence of such a critique in the US was arguably one reason why industrial relations as a discipline lost some of its vitality in that country and why it was relatively poorly equipped to face the challenge of HRM.

2. *CB Systems*

When seen in international perspective, Anglo–American theory also needs to be extended, for CB has very different 'meanings and functions' in different countries (Bean 1994). In this theory, CB is a private, voluntary process conducted largely within individual firms with the outcome being a collective agreement. Yet in France, for example, the state has been an active agent, key agreements are at national and sectoral rather than company level, the practice of 'extending' agreements to nonsignatories is well-established, and even the idea of a specific agreement signed by all parties has never been established and instead some unions have refused to sign agreements so that CB becomes an arm's length collective relationship rather than a tightly defined process of rule-making and rule-administration. The discussion below focuses mainly on the three countries just mentioned, and also Sweden and Germany, and draws on the accounts in Adams (1995), Bean (1994), Crouch (1993), Ferner and Hyman (1998), and Locke et al. (1995) without further specific reference.

2.1 *Level and Scope of CB*

Clegg (1976) identified key different dimensions of CB. The most important are the level of bargaining (whether it occurs at national or sectoral level or at the level of a whole company or at a lower level within different parts of a company), its scope (the range of issues which it covers), and its extent (the proportion of relevant employees covered by it).

The work of Clegg and of later writers (Sisson 1987) established the main international variations in the conduct of CB. In most European countries the level is national or sectoral. Sweden is the clearest example of a nationally-oriented system. From the 1930s to the 1980s the key negotiation was that between the national employers' organization and the trade union federation of manual workers. Germany after 1945 represents the most developed sector-level arrangements, with unions being organized on clear industrial lines and with collective agreements being made for each industry. In addition, the terms of these agreements are extended to firms in each industry that are not directly parties to them. In countries such as France and Italy, there has been a more variable pattern, with the national level coming to prominence, notably in times of crisis, but with the sectoral level that at which agreements often are negotiated formally.

The scope of agreements in these countries is generally broad, with agreements specifying many of

Table 1
CB coverage (percent of workforce covered)

	1980	1990
Australia	88	80
France	85	n.a.
Germany (West)	82	82
Japan	28	23
Sweden	n.a.	83
USA	26	18
UK	70	47

Source: Traxler (1996)

the key terms and conditions of employment such as wages and holiday entitlements. Often, they are underpinned by laws on minimum holiday entitlements and dismissal.

In the United States, collective agreements generally have an even broader scope, one reason being the limited extent of legal provision. For example, unemployment insurance, which in Europe commonly is covered by statute, has been weak or absent, and during the 1950s contracts began to introduce unemployment protection. A typical American collective agreement often is extremely detailed, with anything not covered explicitly in the contract being left to managerial discretion. As for level, America is largely a country of company agreements. From the 1950s to 1970s some sectors had, in effect, sectoral-level substantive terms, with one company agreement setting the pace for others.

Britain provides a distinctive picture. The scope of agreements has been much narrower than in other countries, with only minimal terms and conditions often being specified. Collective agreements were often procedural, that is, setting out procedures to be followed in cases of dispute, rather than substantive. In contrast to most countries, agreements are not legally enforceable but are merely 'gentlemen's agreements,' binding in honor only. The level of agreements was characterized by the official Donovan Commission of 1968 in terms of 'two systems': the formal system of industry-level agreements, and the effective and informal system in which bargains were struck at the level of the plant or even the individual work group. Britain has long had a uniquely uncoordinated system.

According to Sisson (1987), this reflected the failure of British employers to 'neutralize the work place.' Britain's long industrial history meant that by the early nineteenth century many trades of skilled workers had entrenched customs governing wages and hours of work. Reliance on 'custom and practice' was never rooted out by employers, even when they won apparent victories in open confrontations. This failure has in turn been related to other factors including the unwillingness of the state to support a sustained policy of suppressing unions, employers' own divisions, and the deeply entrenched commitment of employees to customary standards.

The sharpest contrast generally is made with Sweden. The main confrontations in Swedish labor history were at national level, and the key Basic Agreement of 1938 laid down the terms of future bargains, which continued to be at national level. A strong and inclusive employers' organization, underpinned by a relatively homogeneous industrial structure, was able to make a firm agreement with an equally inclusive union organization. Other European countries also faced key historical turning points; naturally, the specific relationships reflected the balance of forces between employers, unions, and the state, but the outcome was generally a multi-employer system that neutralized the work place.

American employers achieved neutralization by means of company-level agreements. These were established rapidly in the wake of the Wagner Act as comprehensive, usually three-year, deals. A key component was their legal enforceability. In most European countries, too, collective agreements are legally enforceable, but this fact often has limited practical salience. It has not prevented French or Italian unions from conducting frequent strikes, and in countries such as Sweden and Germany it is more a reflection of a wider belief in contractual obligations than an independent cause of these countries' generally low strike rates. In the US, by contrast, tight legal rules on legitimate strike action, backed up by employers' willingness in the key years of the late 1930s and the 1940s to put legal principle into practice, rapidly created what labor historians identified as a 'work place rule of law': a system of private justice enshrined in the CB contract, underpinned by public law.

2.2 Extent of Bargaining

The extent of CB requires separate treatment, for it is harder to establish than might appear. Such coverage is not coterminous with union membership, since a contract is likely to cover all workers in a given unit whether or not they are members of the union that negotiated it. In Britain and the US, however, it is reasonably clear who is covered by a collective agreement. America, in particular, uses the language of 'exempt' and 'nonexempt' employees (which was introduced by the Fair Labor Standards Act 1938), with the former (broadly, managerial and supervisory employees) not being covered by CB. Where a collective agreement exists, it will specify clearly the sites and personnel affected. In some other countries, the situation is also clear. In countries such as Sweden, where union membership is very high and embraces many white-collar groups who in America might be classed as exempt, coverage is also extremely high. But in France, for example, a collective agreement is less

clearly defined. Many French unions refused formally to sign contracts since to do so conflicted with communist ideology; they were in practice covered even if they were not signatories. Similarly, it is well known that French union membership is a poor guide to the extent of workers' union affiliation. France, along with Germany, is also a case where collective agreements are commonly extended across a whole sector to cover firms and workers not directly party to them (Traxler 1996).

Traxler (1996) summarizes the best estimates that are available. An extract of his data is in Table 1. As explained above, the gap between union membership and bargaining coverage is very wide in France and virtually zero in the US. It is also worth noting that coverage and membership are much higher in Canada than in the US, despite great similarities between these countries. As Adams (1995) explains, this difference reflects two forces: the intensity of employer opposition to CB and the degree of activity of public policy in containing this opposition.

CB coverage, thus, remains high in many countries. Yet it should also be noted that in countries such as the US and UK it was never complete. There are several reasons. First, managerial and other employees were never within its scope. In America this exclusion was particularly clear. In Britain, it was less so; in particular, managers in the public sector were quite likely to be covered. But in both countries private sector managers were generally excluded. Second, there is the small-firm sector where CB was often seen as inappropriate. Third, many sectors did not develop bargaining institutions. A good example comes from Britain, where Wages Councils were set up in sectors such as clothing and catering, where firms were small and unions weak. These councils were established by law and set wages and conditions. They were seen widely as precursors of CB but only very rarely did bargaining grow out of them. When they were abolished in 1993, they still covered over 10 percent of the country's work force. CB did not grow naturally, and flourished in countries where it was sustained by wider systems of interest representation.

3. *Corporatism*

Established industrial relations concerns with the structure of bargaining were complemented from the 1970s with interest deriving from political science in corporatism (Goldthorpe 1984). Corporatism was defined as a system of interest representation in which peak associations of capital and labor represented their members in dealings with the state, which licensed their activities. The scope and level of CB were taken as an important indicator of the degree of corporatism. Early work saw corporatism as a continuum, with Sweden and also Austria being highly corporatist, countries such as France and the UK being weakly corporatist, and North America and Japan having no meaningful corporatism at all.

The concept of corporatism made two contributions. First, it placed narrowly defined industrial relations systems in a broader social context, and encouraged an explanation of these systems in terms of the role of the state and key periods in the regulation of the labor problem. In the US, for example, state intervention through the Wagner Act reflected employer and union preferences for company-level CB but also reinforced this emphasis and helped to preclude corporatist arrangements. Second, it focused attention on outcomes. Researchers increasingly argued that corporatism shifted labor problems from the industrial to the political sphere: issues of distribution between capital and labor could be handled in the context of welfare transfers and social security benefits, and the need to use the strike as a bargaining weapon was reduced. In addition, corporatism reduced self-seeking behavior: under atomized bargaining, every group seeks its own benefits, even though this will produce inflation that wipes out these benefits. Centralized bargaining can eliminate such tendencies, and lead to a virtuous circle in which low wage pressure produces low inflation which leads to low unemployment, which in turn yields returns to unions so that they moderate wage demands.

This analysis produced academic criticism. Some countries, notably Switzerland and Japan, had many of the alleged benefits of corporatism without being meaningfully corporatist. One response was to widen the concept, although it was also weakened as a result. A second was to argue that economic performance was not linearly related to corporatism. Instead, noncorporatist countries such as the US were argued to perform well, because labor markets functioned in a neoclassical fashion. There was, thus, a U-shaped relationship, with countries with middling degrees of corporatism performing the worst. Yet perhaps the most sweeping criticism was that corporatism was static: it might help to resolve labour problems during low unemployment periods, but it tended to set up structures that were resistant to change and unable to cope with more turbulent times (see Therborn 1992, Crouch 1993). Such criticism was not purely academic but also reflected the agendas of government of the Right elected on programs of market-led reform. This leads into issues of the decentralization of CB, but corporatism was not dead and it retains significant appeal.

4. *Decentralization and Decollectivization*

It was never the case that CB was firmly established in such countries as Britain and America. Coverage was patchy, and in America in particular it came under a

number of challenges, even during its heyday. From the 1980s, however, much more extensive challenges emerged, and these embraced many countries. The challenges embraced all the main dimensions of bargaining. Its level has been lowered, with moves from industry level agreements to the plant level. The scope of bargaining has been reduced, as employers stress direct participation rather than bargaining and aim to move away from what they see as the adversarialism of bargaining towards a more team-oriented approach. And the extent of bargaining has been reduced, as employers retreat completely from collective agreements.

Katz (1993) argues that these trends can be identified in many advanced countries. Sweden is a particularly clear case. The 'Swedish model' was consolidated in the 1950s with centralized national-level bargaining as a key component. This structure, it is important to note, was established at the insistence of employers and not, as the Webbs might have expected, the trade unions; the reason was that employers wished to ensure co-ordination and discipline. The model exhibited tensions as early as the 1960s, including problems of including the growing public sector within it and shop-floor discontent on nonwage issues. The pressures increased during the 1980s as private sector employers, particularly those in the export sector, saw a wage system dominated by principles of equality as too inflexible to reflect their needs. The end of the long period of Social Democrat governments in 1976 also marked the end of unquestioned state support for the structure of bargaining.

Trends were equally clear in Britain. Some large employers began in the 1980s to remove recognition for bargaining purposes from trade unions. Structural change was also important, as the decline of sectors where bargaining was strong, notably manufacturing and coal, was matched by the growth of service industries with little or no tradition of bargaining. The scope of bargaining was also reduced, as employers moved away from traditional custom and practice on issues such as staffing levels to more unilateral regulation. Trends towards the decentralization of bargaining to plant level, which had been in train from the 1970s, continued.

According to Katz, the decentralization of bargaining reflects two main forces. First, there is the changing balance of power between managers and unions, which has allowed management more scope to choose bargaining levels. This is not, however, a sufficient explanation for, in an echo of Flanders' criticism of the Webbs, it is not necessarily the case that unions seek the most centralized level; as Flanders pointed out, CB performs roles for managements as well as workers, and managements may have an interest in neutralizing the work place. Moreover, says Katz, shifts in the balance of power cannot explain trends in Germany, where union power has not greatly weakened and yet where decentralization can be observed. The second factor is the need to handle new forms of work organization such as teamwork. These are implemented at the workplace level and they require agreements to handle them. Thus even in a country like Sweden it has been necessary to develop enterprise and plant bargaining.

Katz's explanation calls for two qualifications. First, it gives insufficient attention to the extent of co-ordination practised within large companies. UK studies have shown that large companies monitor pay rates and staffing levels in their subsidiaries, so that effective control has not been returned to local level. They also assess the overall financial performance of subsidiaries, so that local autonomy can exist within tight confines (Brown et al. 1995). Second, power may have rather more of a role than Katz suggests. Most of the agenda of teamwork is driven by managerial concerns about efficiency, and many of the leading exemplars of teamwork have no collective agreements or ones with limited union influence. It is true that, as many studies on both sides of the Atlantic have shown, HRM initiatives are as common in unionized as in nonunion firms. However, case studies generally indicate that the bargaining power of unions has been reduced. Changes in the level and the scope of bargaining have been closely related. It is also true that managements do not have an inherent interest in decentralized bargaining. However, in the particular context of the 1980s and 1990s, with high unemployment and rapid changes in work practices, managerial power was enhanced and it was possible to use this to press through changes in work practices on a plant by plant basis, with each plant's workers fighting for survival against other plants.

Most importantly, the factors identified by Katz may be offset by other tendencies. Traxler (1996) has questioned how far the extent of bargaining has changed. Echoing many other writers, he suggests that two kinds of system can be identified: the 'exclusive,' where bargaining lacks coordination and usually is at company level, as in Japan, the US, and the UK; and the 'inclusive,' where national or industry levels allow coordination, as in Australia and many European countries. He argues that decentralization is a major feature only in the exclusive cases. As discussed below, this account has been developed by studies of neo-corporatism which stress the continuing value of regulatory institutions. Katz has perhaps over-stressed one tendency, which can be counteracted by others. Yet Traxler may downplay the extent to which the scope, as opposed to the extent, of bargaining has changed. The model of inclusive systems may require development. Crouch (1994), for example, identifies various forms of corporatism. A good example is Australia which, although having an 'inclusive' system, has faced significant market-led pressures. As Hampson (1999) argues, the country's corporatism has always been 'enigmatic' in never fully including employers and in being subject to major pressures to

decentralize bargaining. Although the country's unique structures remain in place, how they work has changed dramatically.

5. *Potential for and Effects of CB*

The debate on decentralization increasingly recognized that bargaining structures cannot be placed on a single continuum. The Nordic countries, for example, are commonly described as being both centralized, in that bargaining is conducted at national and/or industry level, and decentralized, in that the legal regulation of industrial relations is relatively light. Countries' industrial relations systems are best characterized in terms of the articulation between levels (Crouch 1994). The UK and the US have relatively weak articulation, with industry and national level bargaining being limited in scope and with employers' organizations never having had a strong focus at these levels. Several countries have, by contrast, developed linkages between levels. In Germany, collective agreements have covered non-wage issues such as job security and training and have, thus, sought links with changes in work organization which have developed at company level. In countries including Italy, The Netherlands, Spain, and Ireland, variants of centralized co-ordination have re-emerged. Two trends can, thus, be discerned. The first is towards decentralization, but here there is a difference between unplanned decentralization and the co-ordinated approaches in some countries: decentralization does not necessarily mean atomization. Second, there are trends towards national co-ordination. As studies of The Netherlands and Italy in particular suggest, such co-ordination can bring substantive benefits including wage moderation and procedural ones in that processes of wage setting have a predictability and order. Such points echo Therborn's (1992) critique of the concept of corporatism: it had the right idea but tended to apply it through attempts to categorize countries on simple scales whereas corporatism is in fact a multifaceted process, rather than a set of institutions, and is best understood in terms of the historical interactions between employers, unions, and the state.

The concept of neocorporatism has, thus, emerged. It is used to argue, for example, that contrasts between North America (with rapid job generation) and Europe (with rigid corporatist structures giving 'excessive' protection to employees) are too simple: The Netherlands, for example, has created jobs successfully through a mixture of market reforms and revived corporatist arrangements.

Such arguments have taken up implicitly one of Flanders's main points: that CB is not simply a means of setting the price as labor but is also concerned with securing dignity for the worker. CB was a means of establishing an industrial rule of law. This was reflected in fairer systems of discipline and discharge, for example. Studies during the decline of CB have revealed the power of this observation.

First, it has been shown that income inequality is shaped by CB. Inequality is greatest in countries where bargaining is weakest and most decentralized. And the rise in inequality in the countries where this has been sharpest, the US and the UK, reflects the decline of CB institutions. Second, similar results are reported for the gender inequality in pay: this is highest where CB is most decentralized, and some writers argue that the most effective route to gender equality is not specific policies on sex discrimination but effective more general bargaining institutions. Third, there is long-standing evidence that where unions are most established in a workplace, and hence where CB is most developed, there are clear outcomes. These include relatively low rates of absenteeism and quitting, low uses of disciplinary sanctions by employers, and wages that are higher than they otherwise would be (Blanchflower and Freeman 1992).

In short, within a single nation state neocorporatism can bring benefits to capital and labor, while even where it is absent CB has effects on income distribution and the quality of working life. Yet the decline of CB and the absence of any effective alternative has led, particularly in the UK and the US, to arguments about a 'representation gap': without CB, how are employee interests to be identified, organized, and expressed? A challenge for CB is whether it can be renewed sufficiently to offer a meaningful model of industrial government, particularly in white-collar and service-sector contexts where collective representation has been rare. A second is to promote the articulation of levels of CB. In Europe, there are some positive developments in terms of legally mandated employee representation systems, while even in Japan and Korea some quasicorporatist arrangements can be discerned. Fewer such developments are discernible in North America. CB has shown the power to adapt in the past and it may continue to do so in the future.

See also: Corporatism; Industrial Ecology; Industrial Policy; Industrial Relations, History of; Industrial Sociology; Labor History; Labor Movements, History of; Labor Unions; Law and Economics; Law and Economics: Empirical Dimensions; Organizational Decision Making; Organizations and the Law; Regulation, Economic Theory of; Regulation: Working Conditions; Working Classes, History of

Bibliography

Adams R J 1995 *Industrial Relations under Liberal Democracy*. University of South Carolina Press, Columbia, SC

Bean R 1994 *Comparative Industrial Relations*, 2nd edn. Routledge, London

Blanchflower D G, Freeman R B 1992 Unionism in the United States and other advanced OECD countries. *Industrial Relations* **31**: 56–79
Brown W, Marginson P, Walsh J 1995 Management: Pay determination and CB. In: Edwards P (ed.) *Industrial Relations*. Blackwell, Oxford, UK, pp. 123–50
Clegg H A 1976 *Trade Unionism under Collective Bargaining*. Blackwell, Oxford, UK
Crouch C 1993 *Industrial Relations and European State Traditions*. Clarendon Press, Oxford, UK
Crouch C 1994 Beyond corporatism. In: Hyman R, Ferner A (eds.) *New Frontiers in European Industrial Relations*. Blackwell, Oxford, UK
Dunlop J T 1958 *Industrial Relations Systems*. Holt, New York
Ferner A, Hyman R (eds.) 1998 *Changing Industrial Relations in Europe*. Blackwell, Malden, MA
Flanders A D 1970 *Management and Unions*. Faber, London
Goldthorpe J H (ed.) 1984 *Order and Conflict in Contemporary Capitalism*. Clarendon Press, Oxford, UK
Hampson I 1999 Between control and consensus: 'Globalisation' and Australia's enigmatic corporatism. In: Edwards P, Elger T (eds.) *The Global Economy, National States and the Regulation of Labour*. Mansell, London, pp. 138–59
Hyman R 1989 *The Political Economy of Industrial Relations*. Macmillan, Basingstoke, UK
Katz H C 1993 The decentralization of Collective Bargaining: A literature review. *Industrial and Labor Relations Review* **47**: 3–22
Kaufman B E 1993 *Origins and Evolution of the Field of Industrial Relations in the United States*. ILR Press, Ithaca, NY
Locke R, Kochan T, Piore M (eds.) 1995 *Employment Relations in a Changing World Economy*. MIT Press, Cambridge, MA
Sisson K 1987 *The Management of Collective Bargaining*. Blackwell, Oxford, UK
Therborn G 1992 Lessons from corporatist theorizations. In: Pekkarinen J, Pohjola M, Rowthorn B (eds.) *Social Corporatism*. Clarendon Press, Oxford, UK
Traxler F 1996 Collective Bargaining and industrial change. *European Sociological Review* **12**: 271–87
Windmuller J P, Gladstone A (eds) 1984 *Employers Associations and Industrial Relations*. Clarendon Press, Oxford, UK

P. K. Edwards

Industrial Relations, History of

In most developed countries, wages and working conditions, forms of employment, bonuses, and allowances are regulated, either by law or by collective agreements. The latter are themselves bordered by juridical or customary rules which define their beginning, process, and results. Their coherence does not merely reflect the logic of the national institutions and of the economy. It takes into account the social actors themselves and the aggregation of their decisions, which constitute industrial relations systems. They vary throughout the course of history. They also differ amongst North America, Japan, and Europe and within Europe itself. But everywhere these systems have been deeply altered or even challenged since the 1980s.

1. *The Concept of Industrial Relations*

The term 'industrial relations' was born in the USA. It is the product of both changing relations between companies and wage-earners and of academic attempts to instil some order into the turmoil of social change.

1.1 *Commons*

John R. Commons, an economist at the University of Wisconsin (Madison), produced, during the first third of the twentieth century, the first major interpretation of social relations in 'American industrial society.' He argued that US workers were more wage-conscious than class-conscious. If companies became aware of these fundamental tendencies, there was room for contracts between management and trade unions. Conflicts could be avoided, minimized, or solved. Government was welcome to promote economic growth by supporting the contractualization of industrial relations. Commons and his followers stressed the importance of institutions for organizing society. Accordingly, they considered trade unions to be a source of social progress. They privileged collective bargaining as fitted to US exceptionalism and adapted to unionization by trade, not by industry.

Commons and his followers were not pure and simple academics. They actively contributed to the liberalization of US social legislation which developed between 1918 and 1940.

1.2 *Dunlop*

John T. Dunlop, a sociologist at the University of Wisconsin, working between the 1950s and the 1970s, went further in the same direction. He systematized the idea of systems of industrial relations at local, regional, or national levels. Thus he stressed the interdependence between the key elements of industrial relations: for instance, between the definition of a trade union, its representativity, and the nature of a negotiation. This interdependence is characterized not only by a common juridical doctrine, but also by reciprocal strategies carried out by managers and unions. Moreover, rules in industrial relations need to be interpreted by reference to the policies of unions, companies, and trade associations. Social actors create rules and bind each other by such regulations, at least for a specific period of time. Like Commons and his followers, Dunlop and his colleagues worked as experts for public administrations and played a significant part in the evolution of collective bargaining and of legislation till the first oil crisis. Unlike Commons, they moved from the idea of US exceptionalism to its apparent opposite: a possible universal convergence between the national systems of industrial relations. 'Industrialism' would bring about a civilization of

social relations everywhere by generating stable processes of collective bargaining. Meanwhile, Dunlop recognized the plurality of national and continental systems.

2. Three Models in Europe

At the end of the nineteenth century and during the twentieth, Europe developed specific features in its industrial relations systems, which are still in force. Collective bargaining is first and foremost the responsibility of employers' associations and trade unions by trade or sector, coordinated by confederations. Bargaining at company level came later, is growing, but remains under the aegis of federations. A distinction is made between collective bargaining, which deals mostly with wages and working time, and participation of wage-earners, which includes work conditions, welfare, and the adaptation of a broader collective agreement to a firm; also, more recently, information on the firm's strategy and its impact on employment. In Germany, participation extends to codetermination, introduced under the Weimar Republic and generalized in the aftermath of World War II, by laws passed in 1951–2. Governments are involved in the national regulation of industrial relations. Intra-European variety may be reduced to three types of patterns. The patterns are quite different on other continents.

2.1 The UK and the Republic of Ireland

This is the oldest industrial relations system. Trade unions are closely related to trades and collective bargaining is basically at the level of the firm or the workshop. Conversely, strikes are frequent at this level, yet they do not foster political radicalization. Labour legislation is light. Politics and society tend to privilege the freedom of action of employers and unions.

2.2 Central and Northern Europe

This group comprises Germany, Austria, Switzerland, and the Scandinavian nations. It is characterized by a trend towards uniqueness in several forms: uniqueness of employers' associations—powerful, structured, obeyed; uniqueness of trade unions—strong and connected to political power; uniqueness of social conflicts—abiding in written or implicit rules. Collective bargaining is centralized by branch. So are strikes, limited to the period of negotiation. In these countries government intervenes by legislating on industrial relations. However, employers and unions have moved to restrict its interference. The Netherlands and Belgium may be included in this group, although both have a plurality of trade unions.

2.3 The Latin Countries

This group covers France, Italy, Spain, Portugal. The keyword here is pluralism: pluralism of conflicts, with a significant percentage of preventive strikes and of wildcat strikes; pluralism of trade unions, with a pole of socialist origin, another of Christian origin, and one which calls itself autonomous or independent; pluralism of political influences among wage-earners. In addition, there is limited authority and even, for many years, pluralism of employers' organizations themselves. Collective bargaining long proceeded only at the national level (for one branch or for the general economy), except at the end of local strikes. Bargaining at company level became a more regular practice in the mid-1950s or even later. Government regulates a variety of issues, except in Italy.

3. North America and Japan

3.1 North America

The US labor movement long featured a pluralism of organisations (in fact till 1955), and employers were as reluctant as elsewhere to negotiate with unions. Things changed with the New Deal. The 1935 Wagner Act, however deeply modified by the 1947 Taft–Hartley Act, made collective bargaining a right for employees, but a union had to be accredited by the vote of a majority of them as the sole counterpart of management in collective bargaining. Law deemed collective bargaining to progress 'in good faith,' i.e., to aim at an agreement. Thanks to the sitdown strikes of the mid-1930s, bargaining spread all over the USA and its contents broadened. The areas of bargaining are now wages, working hours, and work conditions. They may include other matters which have consequences for these elements. The closed-shop clause has been declared illegal. As for Canada, it generally followed the US trends.

3.2 Japan

Up to Japan's defeat at the end of World War II, collective bargaining was exceptional in Japanese companies and would materialize only at the end of strikes. Practices and legislation changed in the late 1940s, under the influence of Japanese liberals and US occupying forces. Collective bargaining became a constitutional right for wage-earners. A company has to bargain with any union which asks for it, and even with any group of employees which does not constitute a union, provided it can express a common will. The employer must present counterproposals to the demands of the employees and negotiate 'loyally.' The contents of collective bargaining are extremely broad. A wave of strikes enabled both the spread of unions and of collective bargaining in postwar Japan. How-

ever, radical unions were finally defeated; worker activism and protest gave way to management victory in the 1960s and 1970s. Company unions took the upper hand in pursuit of economic affluence. Collective bargaining became a type of cooperation with management, and often quite an influential one. This evolution laid the foundations for a corporate-centered society.

In North America and in Japan, when there is no union in the workplace or no agreement is reached, the rights of the employee may be in fact minimal and management keeps its unilateral powers.

4. Recent Changes

4.1 The End of an Era

Many authors consider the stages reached by national industrial relations systems after World War II to have been parts of social settlements built after the strains of the international economic depression of the 1930s and of the wars of the early 1940s, paving the way for a new era of growth. When growth decelerated in the time of oil crises in the 1970s, when national tariff barriers were lowered, and neoliberal ideas made a spectacular headway in the West, industrial relations systems came under attack. At the same period in a number of countries, though not all, membership of unions started to decline. The breakdown of the population changed from a majority of blue-collar workers to an increasing proportion of white-collar employees, in keeping with a growing proportion of women at work and a shift from industry to services. Unemployment simultaneously rose.

4.2 A Shift in Industrial Relations

On a world scale, trade unions entered an era of concessions to employers in terms of wages. The flexibility of work conditions grew. Mechanisms of social protection weakened. Companies moved to a greater individualization of wages, careers, even contracts, and to a more intense involvement of each employee in the workplace. Integration quite often became their motto. Japan was the last country to undergo this new trend and to try and adjust to the challenge. In many countries militancy and the yearly number of strikes and strikers declined. Yet both large companies and governments stressed the importance of collective bargaining to accompany such far-reaching changes and to stabilize the emerging global competitive order. In some countries ways were found to integrate small and medium-sized enterprises in industrial relations systems. Outside the USA the reduction of working time became a target for collective bargaining and legislation. In each country some of the trade unions began to adapt and modernize their strategies and structures. Some were able to regain members in new working groups. But altogether their societal position has not yet been stabilized and the balance between individual and collective has been altered in significant aspects, in favor of the former.

5. Current Debates

5.1 The Impact of New Practices on the Paradigm

These changes in industrial relations themselves have caused new debates among scholars about Dunlop's paradigm. The development of framework agreements, of agreements on discussion methods, of agreements setting a policy and targets rather than rights, in short of agreements which are not contracts, is quite different from the explicit forms and the strict obligations characterizing most of collective bargaining in the USA during the 1950s and the 1960s. Collective actors appear less neatly defined, more unstable, and the representativity of unions is under redefinition in some countries. In short, actors may become more defined by action itself. A number of scholars argue that such developments do not invalidate Dunlop's paradigm, as it is based on autonomous actors able to create rules and to keep their word.

5.2 Recurrent Critiques

Simultaneously older critiques of the paradigm find a greater audience. A number of authors have stressed the importance of unorganized conflict, expressed by absenteeism and turnover. They also stress that the US model of the social conflict as a simple form of adjusting industrial rules may reduce its dimensions and that trade unions are not everywhere limited to the provision of personal services to wage-earners. These authors doubt the possibility of a convergence of national industrial relations systems, despite the development of multinational companies and of continental economic unions. Is it still so easy to say that these systems are functional to industrialization? Are the frontiers of these systems clearly defined, as they indeed produce part of their economic or technological 'context' themselves? Are they not less stable than their earliest analysts implied? What are the multiple sources of their dynamics? Therefore, research on industrial relations is coming back on the agenda of social sciences, as well as of organizations and of individuals.

See also: Business History; Class: Social; Economic History; Industrial Policy; Industrial Relations and Collective Bargaining; Industrial Sociology; Industrialization, Typologies and History of; Labor Movements, History of; Labor Supply; Labor

Unions; Trade Unions, Economic Behavior of; Trade Unions: Empirical Analyses; Work, History of; Working Classes, History of

Bibliography

Dunlop J T 1970 *Industrial Relations Systems*, 2nd edn. Southern Illinois Press, Carbondale, IL

Giraud O, Lallement M 1997 L'institutionnalisation des relations professionnelles en RFA. *Entreprises et Histoire* **6**(October): 36–47

Gordon A 1998 *The Wages of Affluence: Labor and Management in Postwar Japan*. Harvard University Press, Cambridge, MA

Morin M L 1994 *Le Droit des Salariés à la Négociation Collective, Principe Général du Droit*. Librairie générale de droit et de jurisprudence, Paris

Reynaud J D, Eyraud F, Paradeis C, Saglio J (eds.) 1990 *Les Systèmes de Relations Professionnelles. Examen Critique d'une Théorie*. Editions du CNRS, Lyon, France

Slomp H 1995 *Between Bargaining and Politics. An Introduction to European Labor Relations*. Praeger, Westport, CT

P. Fridenson

Industrial Society/Post-industrial Society: History of the Concept

As the proportion of occupations engaged in manufacturing industry declines, to be replaced by work in those sectors of the economy called generally 'services,' social observers have tried to understand the wider changes involved. While much has thereby been learned about the character of society after the decline of industry (i.e., 'post-industrial society'), continuing extensive change makes it difficult to reach firm conclusions.

1. The Theory of Economic Sectors

An awareness that the rise of mechanized manufacturing would have implications for social life going far beyond the production of goods itself had been one of the central ideas stimulating the emergence of sociology as an intellectual discipline during the mid-nineteenth century. Before the growth of industry, social life had been governed primarily by the routines and social relationships of agriculture and landholding. Industry brought a vast range of new occupations; new concentrations of population changed the character of cities; and relations between the owners of capital and those who worked for them replaced relationships based on land tenure as the basic structure of power and inequality.

Central to the change had been the gradual shift of population from agricultural to industrial work. By the mid-twentieth century, however, it was becoming clear that, in those areas which had industrialized first, particularly Great Britain, this process would peak and be replaced by a new change. A number of observers, notably the British economic geographer Colin Clark (1940), had noted a rise in various occupations which 'were not concerned with the production of goods.' They called these 'service' occupations and developed a model of three sectors of the economy, which also embodied evolutionary assumptions about the order of change: a primary sector, in which people extracted useful material from nature (agriculture and mining), was succeeded by a secondary one, in which people fashioned the material so extracted into goods (manufacturing and construction activities); which was in turn to be succeeded by a tertiary sector, in which people made use of the goods produced (services). That distinction of three sectors entered common parlance and became one of the main means by which economies were analyzed and change over time predicted.

For a number of years the most industrialized countries (UK, USA, and some cases on continental Europe) continued to experience growth of both industrial and services sectors (at the expense of agriculture) but from the 1970s it became clear, first in the USA, then in the whole of Western Europe and later Japan, that employment in the secondary sector had joined the primary in a secular decline and that virtually all net employment growth was coming from the tertiary sector. This led to predictions that, as the proportion of the workforce employed in industry declined, there would be extensive social change across many areas of life, possibly as marked as that involved in industrialization itself. The idea of post-industrial society—not just post-industrial employment—was born.

1.1 Subdivisions of the Services Sector

At the same time, various economists, geographers, and sociologists were becoming aware of the inadequacy of the tripartite division of sectors. The services sector was not really defined at all by the idea of using the products of industry, but was just a residuum, made up of everything that could not be allocated to the other two sectors. A number of writers attempted further analysis of the activities that were being included in the services sector. Katzovien (1970) suggested distinguishing between services complementary to industry, 'new' services such as the welfare state and leisure, and 'old,' mainly domestic, services. Singer (1971) made a similar proposal, but replaced the distinction between old and new by one between 'collective' and 'personal'. Most contributions sought

to continue the evolutionary idea of a gradual move away from nature, though Katzovien's reference to 'old' services indicates some doubts about the simplicity of this.

The German sociologist Joachim Singelmann (1978) produced a useful synthesis of the debate with a six-sector model. The primary and secondary sectors remained as in Clark's model, but the tertiary was restricted to those activities concerned with the movement of manufactured products: transport, distribution, and all sales activities. A fourth sector identified services organizing and financing productive activities (banking, other financial services, and all other business services). The fifth sector followed Singer in distinguishing 'collective' services, those activities which had little at all to do with goods, but comprised direct inter-personal services of a social and community kind, such as health and education. This left for a sixth sector Singer's 'personal' services, including leisure and domestic services.

Since Singelmann's contribution at the end of the 1970s, there has been little further progress in the construction of models of sectors, though both the theory of and empirical support for the model were considerably advanced by the Spanish sociologist Manuel Castells (2000). Most popular and political, as well as much economic, discussion still uses the basic three-sector model. However, most national and international statistical services now collect data on occupations which permit identification of five and often six sectors. Also, theories of post-industrial society draw on the idea of the different sectors in order to complete their models of the character of that society.

1.2 Problems in the Analysis of Sectors

This task is made difficult by considerable differences among these sectors. Observers who assume that the characteristics of some extend to all, will make mistakes, as in the frequent assumption (discussed in Sect. 2.1) that services employ predominantly educated labor forces. By far the largest single component, in many countries now larger than manufacturing as an employer, is the social and community services sector. This certainly does employ a particularly high proportion of educated, scientifically trained workers, and is often taken as paradigmatic of post-industrial society (e.g., Daniel Bell's 1974 particularly influential account). However, in the majority of countries, this sector depends rather heavily on public investment and employment. Today this seems to have peaked, as public spending and the role of the public sector are undergoing some decline in most parts of the world.

A second sector which also employs highly educated labor, business services, is growing extremely rapidly. However, it remains very small and, as Singelmann pointed out, often comprises just an out-sourcing of activities which could be part of manufacturing enterprises. This raises an important methodological problem in the study of post-industrial society. Occupations are usually classified according to the predominant activity of the organization for which a person works. Thus, an accountant working for an aircraft-manufacturing firm will be classified as employed in the manufacturing sector; an accountant working for an accountancy firm will be classified as employed in the business services sector. Indeed, if the aircraft-manufacturing firm stops employing its own accountants and subcontracts to specialized accountancy firms, the occupational statistics will show a shift from manufacturing to business services—a move towards post-industrialism—though in fact no change at all will have taken place in work activities.

Should some firms be making decisions to out-source ancillary services while others are doing the opposite and in-housing previously subcontracted work, there would be no overall effect on employment patterns. However, since at least the 1980s, there has been in a number of advanced economies a major trend for large firms to out-source ancillary activities and concentrate on what they define as their core business. There is also evidence that this tendency has been stronger in some countries (e.g., USA) than others (e.g., Germany). While these changes and differences are sociologically important in a number of ways, they have the effect of exaggerating the wider significance for the pattern of human activity of the trend to post-industrial society.

The two remaining services sectors are arguably not really post-industrial at all, as Katzovien's (1970) distinction between new and old services suggested. Although the distributive sector includes some advanced activities in telecommunications, much of it, and much of its recent job growth, involves ancillary services not only to industrial, but also to pre-industrial, society: transporting goods and persons, selling things in shops. The personal services sector constitutes mainly 'servant' activities, a large number of which are certainly preindustrial.

More recently, the growth of information technology and economic globalization have further complicated the question of identifying and distinguishing sectors (see the detailed account in Castells 1996). The use of computers in very many areas of production has confused the once strong distinction between manufacturing and many services employments. Further, the most advanced firms use the internet to fragment and recombine production and distribution activities using disaggregated subcontracted production units. This changes completely the idea of an employing organization located primarily in a certain range of activities, which is central to the idea of occupations and even sectors. In globalized firms, these activities will be located across different parts of the world, with manufacturing components of the task being based in some countries and various associated services in

others. The analysis of economic sectors and change within them needs to develop further. Nevertheless, the theory of post-industrial society has made considerable progress on the basis of the knowledge so far available.

2. *Theories of Post-industrial Society*

The strength of the concept of post-industrial society depends on its being able to identify broader social changes propelled by the initial shift in occupational structure. The leading theorists have tried to do this.

2.1 *Education and Class in Post-industrial Society*

In particular, both the American Daniel Bell (1974) and the Frenchman Alain Touraine (1969), the leading initial theorists, expected a rise in the educational levels of the work force; for Bell, the university-based scientist working for public service would be the emblematic figure of post-industrial society, in the same way that the production-line worker had been the emblematic figure of industrial society and the peasant digging the land that of preindustrial. In this they were partly right; the typical formal educational levels of persons working in many services activities are considerably higher than those in manufacturing. However, as noted, this is true of only some sectors: business, and personal and community services; not distribution and personal services (Castells 1996, Crouch 1999).

Bell and Touraine entertained opposite expectations of class structure. For the former, the decline of the industrial proletariat, a declining importance of material as opposed to intellectual capital, and the overall rise in education standards would see the emergence of a more classless society. Touraine, however, did not expect capitalist management to respond creatively to the new intellectual capacities of the work force. He therefore predicted intensified class conflict as the intellectual capacities of the work force grew. Both have been wrong. Bell was right in expecting that the decline of manual manufacturing work would produce a decline in class organization. However, he was unable in 1974 to anticipate the extraordinary resurgence of financial capital and shareholder power which would follow the global liberalization of capital markets and their association with new information technology to acquire an historically unique flexibility and mobility.

Bell also exaggerated the extent to which the new occupations would be educated ones and did not predict the growth of the many new low-paid jobs, particularly in distributive and personal services, which have been an intrinsic part of the rise of post-industrial society, and the widening gap that would exist between the skilled and unskilled parts of its diversified work force. To the extent that the USA is the country which most embodies trends towards postindustrialism, it shows its society to be one of sharper inequalities than the industrial society of the midtwentieth century (Crouch 1999, Chaps. 5 and 13). Touranaie was mistaken in that the new highly educated section of the work force did not in general become an alienated, disaffected class, but were absorbed alongside or within the ranks of management. However, professional workers in public services—a significant proportion of the highly educated work force—have in many countries often acted as Touraine predicted. This became particularly important from the 1990s onwards, as governments, business interests, and taxpayers began to resent the fiscal cost of public services, and governments began to privatize them.

Not only models of class, but concepts of occupational hierarchy in general, change with post-industrialism and yet sociologists, government statisticians, and political debate still use the images that derive from the factory as the paradigmatic work place: a small group of managers and professional experts; a number of more or less routine office workers; and large numbers of manual workers with various grades of skill. Hardly any services enterprises have that appearance. On the other hand, the different services sectors all produce very different structures. For example, a typical business services or social and community services organization will have a small group of managers, a large staff of professionals, and a small number of routine nonmanual and possibly manual staff; while a typical distributive or personal services firm might have a small core of managers and a large work force engaged in relatively low-skilled tasks that cannot be identified easily as manual or nonmanual.

To date, there have been few attempts to provide a viable paradigm for the occupational hierarchies that would be typical of post-industrial society, nor successful class models based on these. The most successful so far has been that of the Danish sociologist Esping-Andersen (1993), who does so by relating the transition from industrial to post-industrial society to that from the Fordist to post-Fordist economy. This contrasts the familiar industrial and Fordist class model based on bureaucratic rank and the distinction between manual and nonmanual work with a post-Fordist, service model based on differences of scientific and professional skills. However, this involves certain oversimplifications. In particular, it ignores the fact that several sectors of manufacturing have also been identified as 'post-Fordist,' while many organizations providing services exhibit the bureaucracies seen as typically Fordist.

2.2 *Gender in Post-industrial Society*

Strangely, neither Bell nor Touraine anticipated what has been perhaps the most significant social change

involved with the passage to post-industrial society. All service sectors, however they differ among themselves, usually have far larger proportions of female employees than manufacturing industry (Crouch 1999 Chap. 4). There are some exceptions to this, in eastern Europe and parts of Asia, but in the industrial heartland of North America, Western Europe and Japan, manufacturing industry employs very few women. This is not the case of services sectors, especially social and community services, where the majority of workers are usually female. Usually women occupy the junior roles within these organizations. More recent contributions, particularly those of Esping-Andersen (1993, 1999) have brought this to the centre of their analysis.

The major increase in female labor-force participation, which has been experienced in the most advanced economies since the 1970s, is almost entirely due to the shift from industrial to services employment; very rarely has there been an increase in women working in manufacturing. A central characteristic of post-industrial society, justifying the argument that post-industrialism marks a fundamental social change, is therefore a major shift in gender relations.

2.3 Post-industrialism and Post-materialism

It is tempting to relate the idea of post-industrialism to that of postmaterialism. This concept originated in the work of certain US psychologists (Herzberg 1966, Maslow 1970), who believed they could identify a hierarchy of human needs. People had certain basic needs—primarily for food and shelter—which had to be satisfied before they could move on to more sophisticated ones, ending with the so-called non-material needs, such as culture and leisure pursuits. This was taken up by the political scientist Roland Inglehart (1977, 1990, 1997) in a number of works which have tried to identify differences in political aspirations and voting behavior among successive generations in Western societies as these moved into high levels of affluence and presumed material satiety.

This thesis can be criticized at a number of points beyond current concerns, but there is a clear parallel between it and that of the move from industrial to post-industrial society. This is, however, at least in part a misleading parallel. The shift away from employment in production industries towards services does not necessarily imply a similar shift in consumption. The decline in manufacturing employment partly reflects the major improvements that have taken place in productivity in manufacturing and the greater difficult of securing productivity improvements in service delivery. To some extent, it also represents a shift in goods production towards the newly industrializing countries, which goods are then imported into the post-industrial countries. Finally, Singelmann's (1978) arguments that the sectoral shift is sometimes only an artifact of changing corporate strategies on out-sourcing further limits the extent to which changing occupational patterns can be related to consumption.

3. The Future of Post-industrialism

The changes involved in the emergence of post-industrial society have not yet reached a point of rest. As noted above, globalization and the growth of information technology are continuing to have important effects on occupational and economic structures. Globalization involves the growth of extremely large corporations which have activities ranging across several sectors and which locate different parts of these activities in different parts of the world. These moves are an element of what has become the overwhelming demand of firms: flexibility. This has been made their central operational priority by a combination of the uncertainty of markets in the post-Keynesian economy and the new centrality of stock exchanges. Maximizing shareholder value has become the overriding objective of firms and this means a capacity to switch activities rapidly. If a firm's managers see its main chance for maximizing profits in excelling at a particular kind of activity (say, running hotels), they bend all their efforts to acquiring expertise and reputation in that activity. If, however, their main aim is to maximize shareholder value, they must be prepared rapidly to move out of running hotels into something different if share values are advancing more rapidly in the latter.

Having flexibility of this kind goes beyond the process mentioned above, whereby corporations retain a core business but sub-contract ancillary activities. Having the idea of a core business itself becomes a rigidity. The most advanced firms are trying to subcontract more or less everything except a strategic headquarters decision-making capacity. Virtually all other tasks are out-sourced to small firms, self-employed workers or franchise holders. As noted, information technology can be of great assistance to them in the complex organization tasks which this involves. This development is only beginning and it may find its own limitations in problems of ensuring quality control, integrating activities, and motivating work forces. However, to the extent that it does develop, it implies such a degree of organizational fragmentation that the allocation of work activities to anything as static as economic sectors becomes extremely difficult. Economic and occupational changes are continuing at such a rate that social science is finding it very hard to conceptualize the kind of society they are producing. The last word has certainly not been written on what social form follows industrial society.

See also: Differentiation: Social; Economic Globalization and Gender; Industrialization; Industrialization, Typologies and History of; Labor, Division of; Modernity; Modernity: Anthropological Aspects; Modernity: History of the Concept; Modernization and Modernity in History; Modernization, Sociological Theories of

Bibliography

Bell D 1974 *The Coming of Post-industrial Society: A Venture in Social Forecasting*. Heinemann, London

Castells M 2000 *The Rise of Network Society*. 2nd edn. Blackwell, Oxford

Clark C 1940 *The Conditions of Economic Progress*. Macmillan, London

Crouch C 1999 *Social Change in Western Europe*. Oxford University Press, Oxford

Esping-Andersen G 1993 Post-industrial class structures: An analytical framework. In: Esping-Andersen G (ed.) *Changing Classes: Stratification and Mobility in Post-industrial Societies*. Sage, London

Esping-Andersen G 1999 *Social Foundations of Post-industrial Economies*. Oxford University Press, Oxford

Herzberg F 1966 *Work and the Nature of Man*. World Publishing Co, Cleveland, OH

Inglehart R 1977 *The Silent Revolution: Changing Values and Political Styles among Western Publics*. Princeton University Press, Princeton, NJ

Inglehart R 1990 *Modernization and Post-modernization: Cultural, Economic and Political Change in 43 Countries*. Princeton University Press, Princeton, NJ

Inglehart R 1997 *Culture Shift in Advanced Industrial Society*. Princeton University Press, Princeton, NJ

Katzovien M A 1970 The development of the services sector: A new approach. *Oxford Economic Papers* **22**: 362–82

Maslow A H 1970 *Motivation and Personality*, 2nd edn. Harper & Row, New York

Singelmann J 1978 *From Agriculture to Services: The Transformation of Industrial Employment*. Sage, Beverly Hills, CA

Singer P 1971 *Força de Trabalho e Emprego Brasil 1920–1969*. Cadernos CEBRAPS, Vol. 3

Touraine A 1969 *La Société post-industrielle*. Denoël, Paris; 1974, *The Post-Industrial Society*. Wildwood, London

C. Crouch

Industrial Sociology

Industrial sociology denotes the systematic study of the cultural and historical forces that foster the development of manufactories, and the diverse organizational structures, infrastructures, and market structures with which this development is associated. Researchers also analyze the correlates of and consequences, for a society, (a) of the systems of distribution, stratification, administration, and education and (b) of cultural, legal, political, and social institutions that evolve to accompany industrialization. Finally, investigators tend to specialize in studies of microscopic phenomena (like work settings), mesophenomena (like industries, unions, and interest groups), and macrophenomena (like national public policies, business cycles, and international trade relationships).

Industrial sociology has its intellectual headwaters in both the Enlightenment and in the Enlightenment's own origins in fifteenth and sixteenth century natural law. The most relevant legacy of these bodies of thought was the Western tradition according to which 'liberty,' 'individualism,' and 'egalitarianism' could be balanced by 'justice,' and widespread beliefs that liberty could be equated with private property in the enumerations of individuals' rights. The efforts, in England and in America's constitutional convention, to reconcile individualism and liberty with equality, justice, and the common good became, in their sum, the most fundamental of the conundra to which industrial sociologists have addressed themselves in both historical and comparative terms in modern and then in contemporary times (Kerr et al. 1994, Polanyi 1957, Taylor 1960).

The first intellectual effort to reduce these complex ideas into a coherent statement about a society's trade, commerce, and sociopolitical organization came in the writings of Adam Smith, a Scottish professor of moral philosophy. Smith's Enlightenment comprehended: (a) the shift from mercantilism to free trade, and to what soon would be called capitalism; (b) the increasingly higher degrees of specialized labor mobilized by property owners free to deploy their property without government fetters; and (c) the belief that an economy's markets would ultimately serve consumers, though the term, 'economy' in the national sense, was barely used until the 1890s.

These theses have been among the contested subjects of the continuing investigations of industrial sociologists in their efforts to determine the characters, in macrocosmic terms, of the interdependencies of capitalism with democracy, although micro-(organizational phenomena) and mesolevel studies (stratification and public policies) have become more dominant.

Smith and other eighteenth century liberals believed that *laissez-faire* policies, i.e., market-driven economies, would be productive of: (a) equitable returns to investors, workers, and managers; and (b) large sums of returns paid to the factors of production, and to public and private agents who serve them, a sum that would constitute a nation's wealth, later to be measured as a nation's gross national and gross domestic products.

Note here that market forces were admired as producers of both equity and efficiency; in more general terms, the 'common good,' would be the sum of the values generated by self-serving men who

pursued their individual interests (in returns on their capital) and of self-serving persons, otherwise, who are free and well advised to gather and assess information on their opportunities, pursue relevant training avenues, and possess traits of temperament and such commitments to 'virtue,' as would make them attractive to employers.

It has been industrial sociologists' tasks to conduct research into every part of every choice regarding public and private activities strongly implied in the foregoing text and to study the evolving correlates and consequences of these choices by employers, workers, and consumers from the late eighteenth century until industrial sociology began to fragment into specialized segments in the 1960s.

Smith, himself, offered two insightful observations supporting some doubts, at least about the efficiencies he imputed to market forces. First, in a telling discussion of the workings of a pin factory, he extolled the virtues of 'mass production' over hand-crafted pins: a highly rational-specialized (and oppressively tedious) division of labor that afforded what we later came to call 'economies of scale.' This argument became and remains part of the key defense of hugh and complex corporations, but friendly critics point out that 'scales' may be pursued beyond diminishing returns and will simultaneously raise the costly barriers to entry of would-be competitors to challenge dominant, i.e., 'oligopolistic,' firms jealous of their market shares and thus aggressive in their use of their pricing power.

Smith's second reservation was that he doubted whether his 'unseen hand'—price competition—would consistently discipline producers: 'whenever two or more men of commerce meet,' he fretted, 'there is born a conspiracy' to set prices, multilaterally, at the expense of customers; these doubts have been honored by most social scientists who study industry structures.

Finally, in his discussion of the deadening effect of endlessly repetitive and highly specialized tasks in the pin factory, he offered a notion that Marx three-quarters of a century later, elaborated famously into his theory of 'alienated labor,' a theory of workers' interests that sharply flavored both microanalytic studies of organizations and mesostudies of social stability in the US, especially from the late 1920s to the mid-1970s.

As we will see, Smith's paradigm, or 'model' still admired (if not closely read) by many of today's 'Liberals' (i.e., conservatives), has been the subject of sufficient additional research on markets' operations ('industry structures,' opportunity structures, pricing strategies, worker turnover, 'informal' work group behavior, and 'mal-distribution' of income) to lead many industrial sociologists to offer alternative scenarios of the impacts of industrial capitalism; the results show clearly that Smith's and others' *laissez-faire* model can fail to describe capitalism accurately, most conspicuously when 'market imperfections' (e.g., 'rigged markets') together with 'market failures' (recessions and depressions) generally tend to stimulate popular demands for ameliorative regulatory activities, often displacing organized 'class warfare.' Limitations on faith in markets' disciplinary capacities have led to differences in the structures among capitalist economies and in the amount of regulatory activity that is popularly demanded of central governments.

With this background we can move (a) to the conundrum—that dedication to equality and individual liberty (cum property) will very likely lead to inequalities, as some free persons prosper, far more than others, and thus lead economic winners to search for rationalizations that will foster sufficiently widespread sentiments, in a democracy, that inequalities are 'legitimate,' and beliefs that a social system can thus be reasonably stable if citizens are at least equal in political and social, but not necessarily so in economic terms; and (b) to the correlates of the resulting tensions.

1. *The Evolution and Progress of Industrial Sociology*

Smith's thesis was fleshed out by Smith's contemporaries, Henri St. Simon and the latter's one-time student, Auguste Comte, in France; their disagreements have left industrial sociologists with somewhat contradictory perspectives, still with us, for studying macro- and microcosmic industrial phenomena.

St. Simon celebrated the French Revolution by extolling the application of science to all social organizations, a 'rational system' approach, in which merit-by-examination would replace aristocratic birthright in claims to the legitimate exercise of authority over others. Comte, however, pursued a 'natural system' perspective, according to which applications of reason (read science) should be significantly tempered by sensitivity to popular participation, to 'spontaneous' voluntarism—e.g., religious devotions and family formation—and to society's traditions; St. Simon's work has strongly flavored the treatment of bureaux-bureaucracies from Max Weber at the turn of the twentieth century, to Blau and Meyer (1987).

Comte's work was recaptured, 1930–70, in innumerable studies of workers' 'informal groups,' by the founders of the long enduring 'human relations' movement, founded by Elton Mayo and Chester Barnard (both business practitioners) with support from social psychologists, during the 1930s, all at the Harvard Business School; both were critical of F. W. Taylor's 'scientific management'—'one best way of designing jobs'—and of what they called economists' 'rabble hypothesis' regarding workers' putatively individualistic and exclusive interests in wages (Barnard 1956, 1958, Roethlisberger and Dickson 1937). The differentiations of modern macro- and

microsociologists of industry were accordingly most clearly marked, in the early 1960s, in extraordinarily lively debates (Machlup 1967) over the virtues of 'managerialism' (human relations) over 'marginalism' (the more atomistic-individualistic model of workers' material interests) in economists' theory of the firm.

When rationalistic business leaders and their social science colleagues move from factory-level studies to macropolicy issues, however, they admire markets as prototypical natural systems; contrariwise, modern liberal observers shift from their natural system view of organizations (favoring collective bargaining, for example) to the rationalist position (favoring government initiatives to shape economic forces) at the macroanalytical level.

The modern social sciences came into their own late in the nineteenth and early twentieth centuries with the institutional, cultural, and psychological analyses, in France by Emile Durkheim, in the German-speaking world by Max Weber and Sigmund Freud, and in Italy by Vilfredo Pareto and Roberto Michels; these figures influenced industrial sociologists just as they did other social scientists. A little earlier Karl Marx offered an admiring assessment of capitalists' systems, for which Smith had become a founding theoretician, but saw only class conflicts inherent in capitalism' 'production relationships.' Observant of the Marxian contribution, the early theoreticians offered systematic studies (a) of the role of labor's traditions in shaping economic achievements; (b) of the self-seeking roles of elites regarding public policy; (c) of the character of modern economic organizations—*bureaux* or bureaucracies —and; (d) in Freud's case, a full-blown theory of, among other things, personality developments, modal personality structures, and a theory of differential responses to authority among different persons that he identified with child rearing customs, a theory that inspired a number of analysts to identify industrial systems with deferred gratification and even with the repression of natural impulses.

Collectively, these writers contributed to both St. Simons 'rational system' and Comte's 'natural system.' In their midst, Herbert Spencer (in England) and Charles Sumner (at Yale University) applied Charles Darwin's theory of natural selection to whole human societies—'Social Darwinism' (Hofstadter 1944). The results, grist for macro- and mesoscosmic-industrial sociologists, were complexes of ideas about (a) justice and income distribution; (b) the legitimacy of the authority of employers; and (c) the ways in which labor-management conflicts could be defused (and unions avoided) by therapeutic counseling, urged by the Harvard Business School researchers who were inspired by their studies of work groups at the General Electric Company's Hawthorne works; (d) and finally, about the ways in which bureaucratic organizations could be fine-tuned by efforts to expand workers' 'zones of indifference' toward managers' initiatives, i.e., in ways that would enhance leaders' capacities to safeguard their claims to legitimacy, while minimizing what Chester Barnard (1938) termed 'the pathologies [or dysfunctions] of systems of status [differentiations].'

This intellectual ferment encouraged macroindustrial sociologists to look more closely at the sources of stabilities of institutions in and the durability of the 'partnership' of capitalism with democracy; those interested in those macro- and mesoquestions turned increasingly to the work of legal historians with interests in the conundrum, noted earlier, of the problematical Enlightenment linkages of liberty, property ownership, and individualism, on the one hand, to equality, and justice on the other.

These legal and institutional historians, led by John R. Commons and the so-called Wisconsin School, traced the 'legal foundations of capitalism' and focused on the corporate person, and the full array of rights, as bundles of private property, that these *persona ficta* were accorded by the Supreme Court in the Dartmouth College case in 1819 (Commons 1968) and many later decisions.

The more micro-oriented industrial sociologists embraced the concept of corporate persons and their legal attributes, but focused on natural persons, workers especially, in their informal associations with each other in work groups that were often seen to be subversive of employers' goals by their inclination to set production quotas, as at the Hawthorne Works. Managers' own restrictions of output, to their embarrassment, were documented by the writers in the Commons School, M. Josephson, among them.

For most of the period since the early 1930s, industrial sociologists have thus been segmented; one group has been interested in macrosocial structures, stratification, 'webs of rules' (Dunlop), collective bargaining, and ameliorative public policies to deal with market failures, and other relevant ameliorative programs. These programs include 'welfare state'-type interventions on behalf of citizens that these analysts conceived *not* to be 'unfit,' as in Charles Sumner's Darwinian perspective, but victims of the chill winds of markets that were sometimes far more tempered, on their own behalf by large, sector-dominant corporations than by market forces.

The micro-oriented industrial sociologists, human relationists, came into their own, however, during World War II—while unions were neutralized by wage and price controls and by rationing; these investigators offered help in motivating workers to be productive in pursuit of military victory overseas, while union leaders and critics of business leaders' 'manipulation of workers' backed off as they were neutralized by restrictive public policies.

The last of the prewar influences on macroindustrial sociologists were the findings of A. A. Berle and Gardner Means (1938) and by the Congress's Temporary National Economic Committee that the 'fittest' corporate persons in America's markets, like those in

West Europe (Brady 1943) gained their powerful places by taming their markets with the help of the price-setting leaders of the dominant firms in steel, aluminum, glass, rubber, heavy electrical equipment, autos, and appliances—the core, at the time, of the American economy—and by rugged nonprice competition rather than 'destructive' cost-cutting price competition.

2. *Industrial Sociology in the Post-World War II Era*

During the post-World War II period, industrial sociologists, long attentive to the oligopolistic character of the 'core economy,' became interested in the 'peripheral economy' and the differences between the favorable circumstances of blue collarites whose unions capitalized on the postwar economic booms, assisted by Cold War defense expenditures in the manufacturing sector, while the economy's service sector, dominated by retailing, finance (unorganized and with lower wage jobs) and their managers, began to swell. Not a small part of this growing population was linked to the expansion of 'staff' workers, called 'white collar overhead,' in large manufacturing corporations, whose earnings and benefits reflected the gains of their unionized peers.

The first investigations indicated an intensification of the 'bureaucratization' of corporations with consequent concerns about a new, 'industrial serfdom' (growing numbers of blue and white collarites tied to well-heeled employers who had used benefit programs—e.g., healthcare—to attract workers during wartime wage controls) making for a well-situated 'core' population vis-a-vis a peripheral and expanding low wage sector. Sensing they were indeed better off, the first group weathered stagnant incomes, 1970–93, with little protest, while Lyndon Johnson's 'war against poverty' in the late 1960s had already served to focus the attention of 'have nots' more on their educational deficiencies than on their 'relative deprivations.'

In the early 1960s economists urged that better educated persons accounted for a significant portion of the rates of economic growth in North America and Western Europe and that the returns to education were attributable to their putative productivity, the productivity having been measured, tautologically, from their earnings. This human capital theory (Becker 1964) gained no support in a less crude study, i.e., of direct measures of actual productivity and education (Berg 1970), but the theory squared well with some studies of 'status attainment,' in which the dollar returns to occupations are combined with their prestige value to create indexes and then scales of achievement. These status attainment studies took on a life of their own, outside industrial sociologists' purview and, with the aid of emerging computing capacities to process masses of data, helped the field of stratification to become and continue to be one of the most active self-standing areas of sociological research.

Of the major lines of postwar investigation, during the late 1950s and into the 1970s, the work of an interuniversity consortium, 'Industrialization and Industrial Man,' was extraordinarily influential, in scholarly terms, in seeking to sort out the elements of the conundrum adumbrated at the beginning of this article.

The consortium was immensely productive—well over 50 monographs and volumes alone in studies over two decades—in identifying, in comparative terms, the convergence and divergences in the industrialization processes in North America, West Europe, and the USSR, touching on the organization, education, and management of human resources, on industrial relations, and on roles of the cultures, traditions, political arrangements, welfare policies, and infrastructures of the nations under review. The work of the consortium was a high water mark in industrial sociology; its contributors addressed virtually all the permutations and combinations of institutional, cultural, legal, political and managerial arrangements and their individually distinctive contributions to democratic and capitalist ways of governance and production, respectively. A final report (Dunlop et al. 1975) and afterstatements (Kerr et al. 1973) are invaluable resources. Among the most seminal of the individual contributors, in a study of authority in industrial corporations, was R. Bendix (1956).

The increase in price competition, 1970–99, meantime, has had the effect of tempering the proportion of corporations in all industrialized economies once able to manage their competitive environments and, with that, a celebrated balance of power—big business, big labor, and big government, a 'pax Americana,' began to become ineffective (Galbraith 1952). Industrial sociologists, who recognized the stabilizing effects of an 'unstable triad,' also recognized the decline in unions' influence with substantial reductions in unions' membership ranks, starting in the 1960s and, by the 1970s, the raise of single interest groups of sundry types, as well, with the conspicuous loss of the political and economic commerce among big unions, big government, and big corporations, i.e., the tempering of corporations' industrial market power by 'countervailing powers.' Industrial sociologists literally lost a subset of phenomena—both power plays and policy confrontations that were built into the shifting 'two-on-one' *detentes*, on critical issues, that countervailing powers facilitated. The Pax American augmented America's vaunted pluralism, a structure now undermined by proliferations of interest groups.

With the waning of community power studies, well developed before and immediately after World War II, like stratification research and the 'politics of late stage capitalism,' industrial sociologists saw their field

further narrowed as price competition forced a change of interests away from workers' satisfaction and morale among microresearchers and the emergence, instead, of opportunities for increasing productivity and profit by 'downsizing,' 'outsourcing,' and 'off shoring,' as methods for dealing with costs.

Productivity, meantime, has indeed increased, as foreign imports threatened the jobs of some and layoffs have disciplined others; the human relations tradition among microindustrial sociologists has given way to studies of 're-engineering' and other reforms. Business leaders have been made increasingly aware, since the 1980s, that merger movements, for example, feed on corporations' stock values which can be enhanced by selling off segments of the acquired 'targets' and 'downsizing' what remains in order to increase productivity which is measured by a formulation in which the dollar values of output are divided by the number of persons employed; stock values reflect this narrowly bizarre mensurational technique, as corporations prepare themselves to become, themselves, commodities, alongside their products. The most recent study of this new order of phenomena has recently been offered by Cappelli (1999), though the 'externalizing' of work to 'temps' and contract workers is basically an extension of cycles of 'centralization' and 'decentralization' in big industries, 1950–90.

It should be noted that the essential collapse of the human relations school, except for 'sensitivity programs' focused on sexual harassment and race relations, did not even change very much the character of American industry. The reform movements, inspired initially by the human relationists, rarely had enduring effects in most industries, as the data fairly clearly show in a mid-1970s empirical assessment (Berg 1978).

While the linkages of capitalism to democracy are no longer the focus of the attentions of other than industrial sociologists, in the meantime, their other interests have been productively pursued under the headings 'world systems,' political sociology; African American studies; gender studies; stratification (as noted); organizations, work and occupations; and, with renewed interest, economic sociology. The latter was imaginatively limned, early on, by Smelser (1959, 1963).

Other concerns, still pursued essentially as industrial-sociological topics, include labor markets (Berg 1979), industrial relations (Locke et al. 1995, Kerr et al. 1994), and industrial democracy in many publications by W. G. Runciman, in England. The apparent contradictions between the savings and consumption propensities in America, favoring consumption, have been revisited by D. Bell in a new afterward to an5 earlier masterpiece (1978, 1996). Finally, the shift toward capitalist systems and away from 'command economies,' after the Cold War has been helpfully explored by Yergin and Stanislaw (1998). These notable efforts suggest that industrial sociology will be quickened by the so-called global economy, which is not an economy in any useful sense of the term, but a semiorganized congeries of national economies. The emergence of a new world system could well be the occasion for the mobilization of the macro-, meso-, and microanalytical 'triple threat' strategies of industrial sociologists' timely offerings.

Two significant circumstances, beyond those hinted, suggest a revitalization of industrial sociological interests. First, the rapid recoveries (sparked in some measure, by American, German, and by British and other corporate bargain hunters) of the economies that were severely troubled in 1997–99; these developments had helped their highest and their middle classes, but less than half of the other populations (Japan excepted) of the 'Asian Tigers.' The rapid recent recoveries have thus masked the plights of millions upon millions of workers in each of the many Asian economies; the question, now, of a balance of capitalism with democracy is thus not a meaningful one. The situation is not unlike that of the great powers, in the 1920s and the 1950s, vis-à-vis the rest of the world; the 'global economy's' national members' stability, based on growth, cannot be so easily assured, and the resulting dynamics may well be new grist for industrial sociologists' traditional mills.

See also: Capitalism: Global; Comte, Auguste (1798–1857); Economic Growth: Theory; Economic Sociology; Industrial and Organizational Psychology: Cross-cultural; Industrial Ecology; Industrial Geography; Industrial Relations and Collective Bargaining; Industrial Relations, History of; Industrialization; Industrialization, Typologies and History of; Industrial–Organizational Psychology: Science and Practice; Labor History; Labor Unions; Labor, Division of; Marx, Karl (1818–89); Rural Industrialization in Developing Nations and Gender; Schumpeter, Joseph A (1883–1950); Weber, Max (1864–1920); Work, History of; Work, Sociology of

Bibliography

Barnard C I 1956 *Organization and Management: Selected Pages*. Harvard University Press, Cambridge, MA

Barnard C I 1938 *The Functions of the Executive*. Harvard University Press, Cambridge, MA

Becker G 1964 *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*, 1st edn. National Bureau of Economic Research, New York

Bell D 1996 *The Cultural Contradictions of Capitalism*. Basic Books, New York

Bendix R 1956 *Work and Authority in Industry: Ideologies of Management in the Course of Industrialization*. Wiley, New York

Berg I E 1970 *Education and Jobs: The Great Training Robbery*. Praeger, New York

Berg I E, Freedman M, Freeman M 1978 *Managers and Work Reform: A Limited Engagement*. Free Press, New York
Berg I E 1979 *Industrial Sociology*. Prentice-Hall, Englewood Cliffs, NJ
Blau P M, Meyer M W 1987 *Bureaucracy in Modern Society*, 3rd edn. Random House, New York
Brady R A 1943 *Business as a System of Power*. Columbia University Press, New York
Capelli P 1999 *The New Deal at Work: Managing The Market-Driven Workforce*. Harvard Business School Press, Boston
Commons J R 1968 *Legal Foundations of Capitalism*. University of Wisconsin, Madison, WI
Dunlop J T 1955 *Industrial Relations Systems*. Holt, New York
Dunlop J, Harbison F H, Kerr C, Meyers C H 1975 *Industrialism and Industrial Man Reconsidered: Some Perspectives on a Study over Two Decades of the Problems of Labor and Management in Economic Growth*. Inter-University Study of Human Resources in Economic Development, Princeton, NJ
Galbraith J K 1952 *American Capitalism: The Concept of Countervailing Powers*. Houghton Mifflin, Boston
Hofstadter R 1944 *Social Darwinism in American Thought*. University of Pennsylvania Press, Philadelphia
Kerr C, Dunlop J T, Harbison F H, Myers C A 1973 *Industrialism and Industrial Man*. Penguin Press, London
Kerr C, Staudohar P D (eds.) 1994 *Labor Economics and Industrial Relations: Markets and Institutions*. Harvard University Press, Cambridge, MA
Locke R, Kochan T, Piore M 1995 *Employment Relations in a Changing World Economy*. MIT Press, Cambridge, MA
Machlup F 1967 Theories of the firm: marginalist, behavioral, managerial. *The American Economic Review* **57**(1): 1–33
Polanyi K 1957 *The Great Transformation*. Rinehart Co., New York
Roethlisberger F J, Dickson W J 1937 *Management and the Worker*. Harvard University Press, Cambridge, MA
Smelser N J 1959 *Social Change in the Industrial Revolution*. University of Chicago Press, Chicago
Smelser N J 1963 *The Sociology of Economic Life*. Prentice-Hall, Englewood Cliffs, NJ
Taylor O H 1960 *The Classical Liberalism, Marxism and the Twentieth Century*. Harvard University Press, Cambridge, MA
Yergin D, Stanislaw J 1998 *The Commanding Heights*. Simon and Schuster, New York

I. Berg

Industrialization

Industrialization is the process of applying mechanical, chemical, and electrical sciences to reorganize production with inanimate sources of energy. A technological criterion of industrialization by no means entails technological determinism. On the contrary, comparative studies of industrialization have served as a key field for social investigators to highlight the influence of culture, political contingency, and the timing of local development in relation to world history. Cross-national comparisons of industrialization have undermined unilinear models that had predicted that the international sharing of technology would lead to convergence in the social institutions of manufacture.

1. Key Features

Processes of industrialization across diverse kinds of capitalist societies as well as across state-dominated societies with central planning, such as the former Soviet Union, have shared essential similarities. Initially industrialization is marked by massive transfers of labor out of agriculture and into factories that have concentrations of capital equipment. Increases in the productivity of the labor devoted to manufacture come to balance increases in demand for goods, however, and employment in the service sector increases more rapidly than in manufacture after initial industrialization. In the leading industrial societies of Europe, East Asia, and North America, consequently, employment growth during the second half of the twentieth century was concentrated in services, professions, and finance. Even in industrialized countries that are net exporters of manufactures, including Germany and France, the absolute number of workers in manufacturing has been in decline since the 1970s. Some of the most dynamic national economies, including those of the USA and the UK, have become net importers of manufactured goods and net exporters of know-how and of services derived from the use of technology. Insofar as industry is identified by the application of scientific know-how and of inanimate sources of energy, however, it is misleading to speak of a process of 'deindustrialization' in these expanding and comparatively prosperous economies. Concentrations of smoke-stack industry are likely to continue shifting their location around the globe. There is little to suggest that a particular kind of industry comprises a transhistorical requisite for economic preeminence.

Processes of industrialization have universally produced key changes in culture. As the sociologist Reinhard Bendix emphasized, the concentration of workers in factories has everywhere triggered the elaboration of novel ideologies to legitimate the employers' exercise of authority. The mobilization of resources for industrialization also entails innovative definitions of responsibilities and rights over the factors of production, including guidelines for finance or capital allocation and for monitoring workers' health in some measure. Ownership of resources by the state alters only the details of these regulatory challenges. In all settings industrialization has demanded new statutory frameworks for governing the flow of labor across nations, cities, and firms. Industrialization triggers the development of procedures for identifying intellectual discoveries and for singling out their creators for reward, whether in the form of patents and copyrights or through other kinds of social privilege.

Despite these profound similarities in industrialization, scholars guard against exaggerating its concomitants and against identifying industrialization with modernity. The mechanization of some forms of production has called forward more of its opposite, traditional hand work. In nineteenth-century UK, for example, production in centralized facilities with steam power created great demand for the inputs of artisanal trades, including tools and equipment. Mechanization of the simpler branches of enterprise, such as spinning, stimulated home crafts that proved more difficult to mechanize, such as weaving. Factories in Europe, East Asia, and elsewhere have relied upon and invigorated traditional kinship networks to secure better supplies of reliable labor. The growth of factory systems that employed whole families in relatively closed communities sometimes cut down on geographical mobility of labor and reinforced the influence of male family heads who supervised the labor process and recruitment. Especially in contrast to craft work by independent producers for putting-out merchants, therefore, industrialization can intensify the authority of family heads or the subordination of women.

In late nineteenth-century Europe, when employers located factory work in rural areas to reduce wage costs, industrialization also stemmed migration to cities. Rural factories generated supplemental income for families of small land-holders who would have been unable to sustain themselves in the countryside with only agriculture or handiwork. Industrialization in some settings is compatible with the growth of domestic employment using electric-powered looms, grinders, and other home equipment. Most recently, headway in telecommunications and computer linkages has enabled professional workers in finance and science to work from home or to reside in rural communities. Depending on context, therefore, industrialization may run counter to such putatively modern patterns as geographical mobility, centralization of production, personal independence from family authority, and urbanization.

2. *Historical Origins*

The first historical instance of industrialization, sometimes termed an 'industrial revolution,' occurred in the UK from approximately 1760 to 1830. The country's coal output rose tenfold in this period, due in part to the installation of steam-driven water pumps in mines. Pig iron production increased sixteenfold, and many enterprises, including the celebrated cotton mills, introduced steam-powered equipment. No consensus has emerged about the primary causes of the UK industrial revolution. Some historians, adapting Marx's suggestions, have focused on how property relations in the English countryside accelerated improvements in agricultural productivity and thereby released workers for manufacturing. Imperial UK's location in the world economy is also a plausible key, particularly because the country's cotton firms relied on demand from captive colonial markets to cushion the risks of investing in costly machinery. The development of a broadly shared middle-class culture of consumption in the UK increased domestic demand for industrial products. Apart from these explanations centered on macroeconomic conditions, some investigators spotlight the most immediate factors favoring technical innovation. A disproportionate number of the technological pioneers in business belonged to Protestant denominations that sacralized the application of science in everyday life. The UK was unique for having a style of practical education that diffused knowledge of mechanics and engineering among the common people. A widespread popular science and motivation to apply it contributed to the incessant simple improvements in machine design that characterized early mechanization.

Yet explanations of industrialization that highlight the supply of the inputs of labor, technology, and know-how on one side, as well as demand for the product outputs on the other side, can prove inattentive to what lies between: the social institutions of industrial work itself. What led entrepreneurs to centralize work in mechanized factories? Unlike merchants who gave work out to domestic workers, factory employers could directly accelerate throughput to meet shorter delivery deadlines, increase the obligatory hours of work, regulate and reorganize work procedures to ensure quality or flexibility, and prevent competitors from gaining knowledge of innovative changes in products or technology hidden behind their factory walls. Early industrialization was driven perhaps as much by employers' need to control the process of labor as by a need to save on labor in the market. On the eve of industrialization the commercial development of domestic putting-out industries across many regions of Europe supported demographic expansion by making early marriage and new family formation economically more feasible for young adults. In these environments of rapid population growth, employers lacked compelling incentives to invest in centralized machine production purely for the sake of economizing on labor.

Indeed, manufacturers and engineers in Britain's initial lead sectors of industrialization rarely justified new technology by the savings in labor effort. They claimed that such classic inventions as the Crompton spinning mule and the use of chlorine in bleaching works increased turnover and thereby reduced the outlays on supplies and land per unit produced. The management goal, as a textile mill proprietor put it, was to 'produce the goods on time.' Not unexpectedly, therefore, variation in regional economic variables—such as the level of commercialization of household crafts, artisanal specialization, and labor supplies—correlated very inexactly with variation in

the transition to factory systems. Many regions at the forefront of dynamic networks of commerce and putting-out systems by 1750, such as East Anglia, were largely bypassed in the early Industrial Revolution. They show that industrialization did not represent the logical culmination of trade, capital accumulation, the separation of labor from independent means of subsistence, and formation of a nonagricultural work force. Entrepreneurial interest in commanding the production process played a role of its own.

3. *Profit and Investment in Technology*

Industrial development highlights the key economic tensions of capitalist societies. The inputs of technology and labor at the worksite have a quantifiable market value, but the conduct of manufacture itself depends only on their qualitative use values. Put another way, investors and economic planners follow the exchange values of the resources committed to manufacture and of the outputs distributed in the market, but commercial success depends on the separate moment of utilizing the concrete resources and employees at the worksite. Capitalist societies vary in how much their systems of finance and management focus on the need to upgrade technology at the worksite from only a short-term pecuniary perspective, and how much the technical drive to improve productivity steers investment in its own right. National economies that have experienced unusual industrial expansion, such as Germany's in the late nineteenth century, have sometimes pumped more funding into research and technology than prudent financial logic alone might have justified. In the classic case of the successful development of nineteenth-century German cotton textiles, for instance, investment surged at moments when earnings were ebbing and uncertain, but when the interest in upgrading equipment and the use of labor was keen. The example of Japan's slow growth in the last decade of the twentieth century suggests that in some settings, however, it is possible to overspend on research or on capital-intensive technology.

When producers invest in new industrial methods, they call upon their cultural conceptions of the capitalist economy to navigate the uncertain waters. In particular, their orientation to profit intertwines with popular conceptions of labor as an abstract, quantifiable resource in the workplace. Nineteenth-century manufacturers in the UK comprise a classic example of employers who believed the quantifiable increments of labor they appropriated from workers were delivered as if embodied in the workers' produce. British manufacturers accordingly viewed profit as the gain that resulted from buying labor products and other tangible articles at low prices and from reselling them higher. In the British cotton trade, for instance, manufacturers focused on cornering cheap raw materials, even if inferior grade inputs retarded improvements in labor productivity. Given their focus on profit from the sphere of exchange, British manufacturers purchased innovative equipment in the nineteenth century if they could foresee relatively immediate payoffs in trading in the current environment. Their contemporaries in Germany imagined that employers appropriated labor in the form of *Arbeitskraft*, 'labor power.' German manufacturers accordingly viewed profit as the result of the efficient conversion of this labor capacity into an output at the point of production. Industrial employers in Germany invested more heavily in new technology and in reorganizing work than did their counterparts in similar economic settings in Britain.

Practices for upholding relations of trust among firms have offered advantages in industrial development even when they contradict short-term profit seeking. In Japan, for example, corporations in the late twentieth century purchased assembly parts from suppliers through long-term relations that carried diffuse mutual responsibilities. So long as a customary supplier strode to boost its performance, corporations bought its components even when the price or quality fell below that offered by an emerging competitor. Models of individually rational economic behavior would expect corporations to switch to the best or least expensive provider in short order. Yet the longer-term purchasing commitment may paradoxically foster technological improvements in industries as wholes by increasing confidence that products issuing from renewed investment find secure buyers. The habitual Japanese purchasing system created effective organization for industrial innovation by ensuring an appropriate time horizon for carrying out research and design, by putting incremental but not lethal pressure on suppliers to upgrade their results, and above all by generating diffuse obligations to uphold quality and service beyond the letter of a formal contract.

4. *Contingencies in Industrial Organization*

Karl Marx, perhaps the most famed theorist of industrialization, inferred a single path of factory organization based on cumulative investments in machine technology. He expected that advanced firms would mass produce standard wares and simplify tasks expected of workers to reduce continually labor inputs per product manufactured. The assembly line made famous by the US auto maker Henry Ford, as well as the state-run firms of socialist regimes, attempted to embody a similarly rigid logic. Yet examinations of historical diversity in manufacture have undermined the assumption that industrial development necessarily depends on maximizing the production of uniform goods. Industrial districts such as Solingen and Neuchâtel in the nineteenth century, and Bologna and Palo Alto in the twentieth century, have prospered

by exporting short runs of custom products. The hundreds of small cutlery firms in Solingen, for example, obtained crude blades from mechanical forges and on purchasers' demand custom finished them with steam-powered grindstones into literally thousands of types of cutlery. By 1914 Solingen had outstripped foreign competitors by focusing as much on variety in design as on cost minimization. This alternate course of development, called flexible specialization, comprises a salient strategy today for manufacturers in the mature economies. Producers who bear high wage costs can compete in the world market by rapidly adapting technology to add value to products through innovative or specialized design, often with the aid of computers.

Whether flexible specialization works effectively as an industrial strategy under capitalism depends on political events and on the infrastructure of public institutions, not purely on economic conditions. Firms that remain modest in size seem more adept at adjusting technology to changes in the needs of market niches. Much of the employment growth and dynamism in mature economies is consequently focused in small-scale enterprises. Yet these smaller organizations typically lack sizable pools of funding for research into fundamentally new kinds of technology, and they risk stagnation if they vary products only minutely. For enterprises to share the costs of basic advances in technology, flexible specialization typically calls for a complement of cooperative research centers, universities with ties to business, or other institutes for promoting local technological improvement. The creation of these supporting institutions can grow out of municipal campaigns for local entrepreneurs, out of guild programs for promotion of the trade, and out of scientific provisions by central states. Flexible specialization also requires safeguards against sweatshop wage reductions, so workers have reason to invest in their own skills and to contribute their ingenuity to the adaptation of equipment in changing markets. Labor relations supporting this path of industrialization may emerge unintentionally out of struggle between labor and capital at key political junctures. In the cutlery trade of Solingen, for instance, employers were 'forced into their luck' by powerful unions in the 1870s that successfully pushed for codified procedures to divide productivity gains between workers and employers.

An accelerating need for firms to deploy fundamentally new scientific discoveries has further undermined the accustomed models of factory work. Frederick Winslow Taylor and other managers at the beginning of the twentieth century believed that the division of work into fixed elementary tasks should take primacy in industrial organization. By contrast, corporations embedded in fast-changing scientific communities today find their primary challenge is to decide on responses to events and to speed collective learning of sophisticated skills. The tasks in nineteenth-century manufacture could be carried out separately and their outputs added up to assemble a product. Contemporary industry, by contrast, often finds that work processes are more interactional than additive and that the challenge is to encourage mutual adjustments among work groups as they address qualitatively new assignments. Firms of relatively modest size dependent on continual scientific change have therefore encouraged collegial and participatory work styles rather than hierarchical and bureaucratic forms. Yet no single method for organizing industry is becoming dominant. Large multinational corporations that lean more heavily toward bureaucratic organization tie much of the global economy together, and the top 10 dozen of them account for perhaps one-quarter of the world's production. The intensification of global exchange has opened new opportunities for mass production, because it widens economies of scale for making durable goods as well as for offering marketing and services. The diverse kinds of industrial organization are not simple historical alternatives, but call on each other in mutual exchange.

5. *The State and Industrial Change*

With the relative decline of state barriers to trade, and with the increasing salience of an international division of labor, processes of industrialization scarcely represent local or national trajectories of change. They are formed by the position of local production in the changing industrial networks of the globe. Entry into an industrial subsector such as elementary computer chips may bring high rewards and global leadership in one global setting yet represent a tributary, dead end path in the next. In some eras of international development the state has emerged as a critical actor for identifying prospective lead sectors and for facilitating their penetration by local firms. In the late twentieth century the governments of Japan, Korea, and Taiwan were prominent among those that guided local industry into more rewarding stations in the global hierarchy of labor. States successful at sponsoring relative industrial advance in the late twentieth century maintained their autonomy from narrowly self-interested business groups by building an organization of career bureaucrats who were both recruited and remunerated by meritocratic standards. Yet developmental states also benefited from the close-knit social networks between these state officials and entrepreneurs. Unofficial social ties, based in part on comradery among elite university graduates, encouraged the circulation of information about the business and technical environment as well as effective negotiation and implementation of strategic goals.

In Japan, for instance, the state's amalgam of bureaucratic autonomy and informal social connections to economic decision-makers was key to its triumphal stewardship of the domestic computer

industry. The central state not only offered tax incentives for domestic computer design while discouraging imports. It also used the in-depth knowledge of the industrial players it acquired to allocate patent information and funds for shrewdly chosen research projects. It set up a sound cluster of competing firms within each segment of the computer market. Japan's office for economic policy distributed the findings of research cooperatives to individual firms for competitive product development, and it defended firms against excessive price-cutting so they could devote resources to further research. Sponsorship of technological risk-taking among competing firms, as in Japan, is currently more successful than public investment in a single champion firm. As part of global historical development, however, comparative industrial advance cannot be linked to a particular kind of state policy across all eras of development. Indeed, in eras when the most promising pathways toward technological development in an industrial sector may be unchartable, or when their pursuit depends on transnational scientific collaboration, the nation state may become ineffective at organizing development. The very success of local industries at one stage of development creates subsequent challenges for government policy makers. For success is apt to vault leading local firms into multinational corporations whose decision makers have less interest in national development as such because they take their investments across the globe as the foundation for strategy.

6. *Diversity and Discontinuity*

The concept of industrialization, an inheritance from nineteenth-century social science, originally carried evolutionary connotations of long-term, cumulative progress. But processes of industrialization vary fundamentally by the kind of control over labor that entrepreneurs seek in a particular epoch, by the location of an enterprise in the swiftly changing global division of labor, and by the variety of technology exploited. In neither the mature nor the newly developing regions of the world economy do industrial achievements today necessarily descend from the same economic and technological logic as that of the first Industrial Revolution. Investigators can appreciate the diversity in processes of industrialization if they discard the assumptions of sequential advance or of continuity in development that once governed modern social thought.

See also: Capitalism; Capitalism: Global; Economic Growth: Theory; Industrial Society/Post-industrial Society: History of the Concept; Industrial Sociology; Labor, Division of; Marx, Karl (1818–89); Modernization, Sociological Theories of; Technological Innovation; Technology and Organization; Work, History of; Work, Sociology of

Bibliography

Bendix R 1974 *Work and Authority in Industry. Ideologies of Management in the Course of Industrialization*, 1st edn. University of California Press, Berkeley, CA

Biernacki R 1995 *The Fabrication of Labor. Germany and Britain, 1640–1914*. University of California Press, Berkeley, CA

Dore R 1983 Goodwill and the spirit of market capitalism. *British Journal of Sociology* **34**(4): 459–82

Evans P 1995 *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press, Princeton, NJ

Jacob M C 1997 *Scientific Culture and the Making of the Industrial West*. Oxford University Press, New York

Sabel C F, Zeitlin J (eds.) 1997 *World of Possibilities. Flexibility and Mass Production in Western Industrialization*. Cambridge University Press, Cambridge, UK

von Tunzelmann G N 1995 *Technology and Industrial Progress. The Foundations of Economic Growth*. Elgar, Aldershot, UK

R. Biernacki

Industrialization, Typologies and History of

1. *The Process of Industrialization*

Industry is distinct from agriculture, forestry, fishing, mining, quarrying, transportation, finance, trade, and the myriad of services that make up separable but classifiable tasks performed by the workforces of modern economies. Throughout history men and women manufactured commodities for use or for trade and sale. No society has operated without producing some range and levels of industrial output. Industrialization refers, however, to economic change that is recent and different in scale and scope from the manufacture of artifacts. As a socioeconomic process, industrialization includes a rapid transformation in the significance of manufacturing activity in relation to all other forms of production and work undertaken within national (or local) economies.

Economists, historians, and sociologists have measured and compared industrialization in statistical form as it appeared in national accounts and evolved historically for a large number of countries. Their data shows that, as industrialization proceeds, shares of workforces employed in and national outputs emanating from agriculture and other primary forms of production decline, and the shares of employment and output that are classified as industrial increase. Output and employment emanating from the third macro

sector of national production, services, goes up or down in relative terms. Services include all forms of non-commodity output that are sold (and/or supplied) either to consumers (for instance as health care) or utilized as 'inputs' (e.g., distribution, legal advice, accountancy, etc.) in order to sustain both manufacturing and primary forms of production. Clearly as and when industry grows more rapidly than other forms of commodity output then the allocation of services changes towards manufacturing and away from farming, fishing, forestry, and mining (Kuznets 1965).

For sustained development there is no substitute for industrialization, which can also be measured as the reallocation of a nation's stock of capital (embodied in the form of buildings, machines, equipment, tools, infrastructure, communications, and distribution networks) away from primary and towards industrial production. But macro data is more likely to be available for the foreign trade of nations which allows observers to measure the progress of industrialization over the long run in the form of predictable shifts in the composition of a country's exports and imports. Sales of domestically produced manufactured exports normally grow in significance and purchases of foreign manufactures diminishes as a share of total imports (Jones 1988).

Economic data has been classified in heuristic ways and has been disaggregated into numerous activities and functions in order to expose the extent, pattern, and pace of industrialization over time across regions and among countries. These essentially taxonomic exercises help to define and make concrete a process that has proceeded on a global scale for centuries. They expose national variations from more general or regional patterns and contribute to the understanding of major economic variables that historically have fostered or restrained industrialization in various parts of the world (Kuznets 1965).

Industrialization has been an important process for the welfare of humankind because the reallocation of labor, capital, and other national resources towards industry has usually been accompanied by technological and organizational change, which led to higher levels of output per person hour; rising living standards; population growth; urbanization; cultural changes; and shifts in the balance of power among nations. Thus, industrialization can also be defined in social, cultural, and political terms, and sociology depicts the history of industrial societies in terms of a set of interconnected characteristics, hegemonic values, and legal systems, represented as functional for the development of modern industry.

How, when, and why particular societies moved from pre-industrial to industrial norms, motivations, status, family systems, and modern institutions that characterize industrial society remains difficult to explain. Sociological theories elaborate upon the type of *changes* required in individual behavior and social institutions for the modern industry to succeed. They contrast 'traditionalistic patterns of action' that are defined as ascriptive, multidimensional, communitarian, familial, and authoritarian with the types of individualistic, achievement-orientated, mobile, entrepreneurial attitudes and behavior that became dominant in cultures as industrialization took hold. This approach depends upon the vocabularies and concepts drawn from sociology, psychology, and cultural anthropology, and on analyses inspired by Max Weber that continue to be preoccupied with value systems (derived ultimately from religions) that have 'motivated' the 'drive to industrialize' in different national and cultural settings (Parsons 1971).

Unfortunately, historical records are not clear on whether social changes precede or accompany industrialization. Until the late twentieth century, sociological approaches to industrialization have, moreover, been more concerned with its disruptive, dislocative, and potentially negative consequences for families, communities, villages, and regions, than with its nature, origins, and positive effects on living standards. Read as a social process, industrialization often leads to differentiation flowing from the division of labor, class formation, and uneven regional development. As industry diffuses from country to country, it becomes associated with diminishing returns, de-industrialization, unemployment, and the economic decline of some nations. Pessimistic writings about industrialization are often derived from Marx. Fortunately modern sociological understanding of industrialization combines several traditions and schools of theory with historical enquiry and a more process-centred global perspective which has exposed how complex, multifaceted, and variable the process of industrialization has become since Marx, Comte, Durkheim, and other canonical social scientists wrote their critiques.

2. *The British Industrial Revolution and the Diffusion of Modern Industry*

Modern industry, concentrated in particular regions and towns, housed in factories, organized in corporations and firms, and using machinery and inanimate forms of energy evolved gradually. That evolution has preoccupied generations of economic and social historians who have explained why industry became dominant in some national economies before others and analyzed major forces that carried industrial sectors forward from one stage to another. Their writings are heavily concentrated upon the epoch which opens with the beginnings of the British Industrial Revolution in the mid-eighteenth century and closes with World War I. Latterly, historical literatures have emerged dealing with proto-industrialization in Europe and Asia before the First Industrial Revolution and with late industrialization characteristic

of Russia, Eastern Europe, East Asia, and Latin America. The bibliography of literature on the histories of firms, industries, manufacturing sectors located in villages, towns, regions, countries, continents includes entire libraries of monographs, essays, and articles. The volume of research covering millennia of time and every conceivable set of spatial arrangements defies summary. Thus the task for historians (and social scientists) and the aim of this essay is to review the theories, patterns, and typologies that have been imposed upon this mass of otherwise incomprehensible local and historical detail (O'Brien 1998).

Shortly after 1945, when the causes of the wealth of nations emerged to dominate the agenda for the social sciences, historians began to formulate models to make sense of Europe's precocious industrial development. At that time the most accessible heuristic device was to refer continental industrialization to a paradigm case, the British Industrial Revolution, as it supposedly proceeded from the 1750s to the 1840s. On the assumption that Britain's lead could be ascribed to the early diffusion of superior industrial technology, it followed that the economic history of Europe could be understood as the spread of techniques of production and new forms of industrial and commercial organization from the first industrial nation to more backward economies on the mainland. In this perspective, British methods constituted not merely initial but best practice with universal application. Diffusion then became a major concept and the central problem for European economic history, at least for the nineteenth century, was to explain Britain's early start and the rate at which other European economies adopted British ways (Landes 1969).

Yet European statesmen, businessmen, and other contemporary observers rarely perceived of their national goals in terms of a race to catch up with Britain. Indeed many expressed doubts about the 'British way' and historians of European economies have not allowed their accounts of the industrialization of economies on the mainland to be dominated by a model inspired largely by Anglo-American perceptions and experience. Diffusion is now perceived as a complex process of creative adaptation rather than a mere emulation. The transfer of technology only succeeds when a matrix of related capacities into which machinery can fit has been built up and is receptive to change (Berg and Bruland 1998). Once in place, European firms and technologies spawned their own problems, solutions, and national trajectories for future innovation (Mokyr 1990).

Furthermore, late twentieth century historical research has restored perceptions of the British Industrial Revolution as a drawn-out process in which the upswings in national income and industrial production of the eighteenth century are no longer represented as 'discontinuities' let alone 'take offs.' Britain's technological lead has emerged as less novel, diffused slowly, and eventually succeeded because it rested upon the evolution of a commercialized economy throughout the British Empire over several generations before the American and French Revolutions (Crafts 1985). Comparative historical research which now brings Asian as well as European economies into the frame has also elaborated upon foundations of Britain's early start which were specific to that well-endowed offshore island. For example, the relatively high level of energy consumed per capita and per worker, obviously reflected abundant and accessible supplies of cheap coal (Wrigley 1988). Located at the hub of a rapidly growing Atlantic economy, when waterborne transport was the cheapest way to conduct commerce, a small island was more likely to reap larger gains from oceanic trade than continental rivals. That locational advantage was, moreover, safeguarded and extended by extraordinarily high levels of public investment in sea power which over the long run lowered transaction costs for British merchants, provided favored access to imperial and to foreign markets, and weakened the economies of Britain's Iberian, Dutch, and French rivals. Writing long after Trafalgar (where the Royal Navy finally won the centuries-old battle for the Atlantic) and sanctified by an ideology of liberalism, British historians tended to overlook the role of the Royal Navy in securing for British business 'disproportionate' shares of world trade, shipping, shipbuilding, and commercial services (O'Brien and Quinnault 1993).

Thus the First Industrial revolution can be represented as something of a special and less of a paradigm case for the economic history of Europe (Mokyr 1985).

3. *Regions, Towns and Proto-industrialization*

Industrialization was a regional process. Modern industry certainly did not spread randomly over national maps but 'clustered' within defined geographical boundaries. Insights are gained by observing the factors and circumstances which prompted industry to locate, grow, and decay in some regions rather than others; and in some places before others (Pollard 1981). Clearly the geographical matrix for the analysis of economic change must be specified in terms of what the historian is trying to explain, and the rise and decline of regions forms an important component of national economic histories. Where and when markets for commodities and factors of production are bounded in space, regional, rather than national or multinational, economies constitute the obvious units for study.

Research into European regions of proto-industrialization over three centuries before 1800 promises to create the evidence required for a typology which might help historians explain the long-term evolution of major branches of manufacturing (particularly

textiles, but also leather goods, finished metal wares, clocks and watches, and furniture) into mechanized factory industries over the nineteenth and twentieth centuries (Musgrave 1999). However, no single model has emerged to account for the complex range of cases that historians have uncovered of tendencies for industries to concentrate geographically, to specialize, and to engage in long-distance external and internal trade. Histories of proto-industrialization usually elaborate upon a set of initial conditions which contributed to the availability and willingness of local workers (males, females, and children) to turn from dependence upon agriculture to become involved (either part- or full-time) in manufacturing. Their accounts include: infertile land, inheritance systems, size of farms, communal rights to land, the cost of alternative supplies of food, the productivity of local agricultures, regional specialization, and competition within agriculture. Populations living close to subsistence levels, on relatively small holdings, and afflicted by high densities of labor to cultivable land were often among the first to embrace opportunities to supplement family incomes provided by industrial work, but so did landless and seasonally underemployed laborers from richer and more commercialized agricultural areas (Kriedte et al. 1981).

But mercantile organization and capital were also required and they usually emanated from towns. Towns housed the majority of educated and innovative people. Productive knowledge was more readily diffused across economies where populations became spatially concentrated into 'networks' or 'hierarchies' of cities and towns. The development of towns improved the efficiency of labor markets, promoted the commercialization of agriculture, and replaced barter with monetary transactions. Europe's maritime cities established imperial outposts overseas which fostered international trade and specialization (de Vries 1984).

By the eighteenth century, the map of industrial Europe could be colored into fairly well defined regions of proto-industrialization, specialized upon the production of textiles, clothing, boots and shoes, clocks and watches, tools, weapons, furniture, household utensils, and pottery. Several of these urbanized regions had survived and adapted to competition over long periods of time. Others had recently emerged as loci of concentrated production. All these industries faced limits to improvements in labor productivity through further refinements to the division of labor or from organizational improvements to the putting-out system.

Over the nineteenth century and as mechanical innovations came on stream and markets widened through innovations in transport, the long-term development of these regions had endowed them with several advantages for transitions to factory systems of production. For example, their sales outlets had been long established; they contained supplies of labor experienced in manufacturing for distance markets; capital and managerial skills were available from urban merchants and all kinds of other external economies flowed from the prior concentrations of industry in the hinterlands of towns. Local agricultures had long demonstrated their capacity to respond to the demands of industrial populations for food and raw materials (Braudel 1981–4).

Yet despite, and sometimes because of, these favorable historical traditions, many advanced regions of proto-industrialization did not move rapidly forward to embrace new techniques of production and modes of organization. In some cases traditions proved to be more of a hindrance than help or several regions lacked some essential ingredients for mechanization, such as supplies of cheap energy or engineering skills.

Research into proto-industrialization constitutes a program with explanatory potential, but it only covers certain sectors of modern industry. Producers' goods (iron, steel, ships, transport equipment, bricks, glass, and minerals) as well as some important consumer goods (such as flour, beer, and soap) were not manufactured under conditions that even approximate to a putting-out system. Furthermore, if the program is to fulfill its promise we still need to know rather more about the varieties and adaptabilities of organizational forms that are encompassed under the heading of proto-industrialization and far more about the causes of success and failure, both before and, *above all*, during the transition to the factory system (Cipolla 1973–4).

4. *Patterns and Typologies of European Industrialization*

Largely because he conceived of European industrialization as a process of diversity rather than uniformity, Gerschenkron produced a heuristic model which focused research upon variations in the *way* European economies passed through a decisive stage of accelerated industrial growth, or the 'great spurt.' In his view, Britain forged ahead because conditions for 'spontaneous industrialization' developed early—and the problem for 'relatively backward economies' on the mainland was to find effective substitutes for missing (British) prerequisites. Thus delays exhibited by continental economies in their endeavors to catch up with Britain are imputed to a range of 'prerequisites' for successful industrialization available to those who transformed the British economy (but in shorter supply elsewhere in Europe), and also to the time they took to develop substitutes for their cultural, institutional, and technological deficiencies (Gerschenkron 1962).

Gerschenkron (like Rostow before him) neglected natural endowments and the role of the state for the British Industrial Revolution, and he concentrated

upon other elements of industrialization which he hypothesized would vary systematically with the timing and degree of backwardness exhibited by 'follower' countries, as they passed through a clear discontinuity in the growth rate of industrial production (Rostow 1965).

Gerschenkron predicted that when historians examined the build-up of modern industry in Europe, country by country, they would observe that the more backward or late the economy when it began that process, the more rapid would be the rate of acceleration in the growth of industrial production, and the more pronounced would be its reliance on capital as distinct from consumer goods. He anticipated that among 'latecomers' new industrial enterprises would emerge as technically sophisticated and capital intensive, larger in scale and more prone to monopoly. Such enterprises would rely to a great degree upon external sources (banks and the state) for both investible funds and entrepreneurial initiatives. For their markets and supplies of cheap inputs they would depend far less than their British counterparts had upon rising productivity in domestic agricultures. Historical research (inspired by his hypothesis) has not merely widened the sample to include potentially deviant cases from the Austrian Empire, the Balkans, and Scandinavia, but has deepened knowledge of British, Dutch, French, Italian, and even Russian industrialization in ways that appear inimical to the survival of Gerschenkron's typology as a usable heuristic device to generalize about European, let alone Asian, industrialization (Sylla and Toniolo 1991).

The theoretical and statistical problems involved in delineating phases of trend acceleration seem to be insurmountable and as historians turned away from imperfect and partial indices of industrial and national production to investigate the regional and urban proto-industrial foundations of modern industry, the drama of 'take-offs,' 'spurts,' and 'discontinuities' in macroeconomic history has faded from view.

Expectations that historical records might reveal distinctively national patterns of industrial organization cannot be tested because the data required to quantify and systematically compare the prevalence of larger-scale, more capital intensive, and monopolistic firms across countries cannot be constructed. Furthermore, typologies which juxtapose continental deviations against a 'stylized' version of British industrialization now look anachronistic. Late twentieth century research on British industrial organization reveals variance in scale, capital intensity, and departures from competitive norms across industries throughout the First Industrial Revolution (Floud and McCloskey 1994). Variations in plant size, competitiveness, and factor intensity are not systematically correlated to backwardness, but are related to legal conditions for incorporation, the development of capital markets, changing technology, and above all to product mix. Long-term changes in the composition of manufacturing output do not expose any systematic or predictable shifts from the production of consumer to capital goods or vice versa (Pollard 1990).

Nevertheless, transitions from simple to complex processes of manufacturing activity are normally orderly. Thus the initial mechanization and standardization of several consumer-goods industries tended to be located within well-defined regions of specialization, utilizing the investible funds, skilled labor, and entrepreneurial talents available in established proto-industrial regions. Factories emerged on a small scale with relatively low capital–labor ratios to engage in the linear transformation of raw materials into familiar products sold on well-established markets. In many cases resources could be mobilized from within the traditional economy in order to replace imported manufactures or the products of handicraft industry. As manufacturing expanded, as the agricultural sector embarked on mechanization and farmed with chemicals, and as transport networks began to cover national terrains, demand and supply constraints on the expansion of domestic capital goods sectors gradually diminished. At that stage a 'second wave' of industrialization carried economies forward (Mathias and Davis 1990–8).

Deviations from a path dependent process of diversification from simple to complex systems of manufacturing occurred, but are explicable in terms of country size, natural endowments, comparative advantages within an evolving world economy, direct foreign investment, and state policies. Typologies of European industrialization which pay insufficient attention to geography and foreign trade miss essential elements of the process. Europe developed over the nineteenth and twentieth centuries as a continent of large and small political units, with a skewed distribution of natural resources. As possibilities for foreign trade expanded, that promoted different patterns of industrial growth from country to country (Aldcroft and Sutcliffe 1999). For example, at comparable levels of per capita income very few European economies achieved anything like Britain's comparative advantage in the export of manufactures. As late as 1913 the shares of 'manufactured commodities' to total exports varied from under 10 percent for Greece, Denmark, Norway, and Portugal to over 75 percent for Britain, Belgium, Switzerland, and Germany, with the rest of Europe somewhere in between. Systematic comparisons of trade data across countries and studies of long-term trends in the composition of net imports (exports minus imports) for major manufactured commodities offers historians prospects for the statistical delineation and comprehension of the patterns of industrialization undertaken by different European economies before and after 1914 (Bairoch 1976).

General propositions about the role of banks have also proved difficult to force into the form of generalizations. Modern research on financial intermediation

has fruitfully shifted attention towards the problem of analyzing responses by banks to industrialization as it proceeded in particular national and regional contexts. Cameron summarized the results of international comparisons by admitting he lacked criteria to measure their success or efficiency. Although he was prepared to laud the contribution of banks in Scotland and Sweden and to deplore their sparseness in France—an economy which apparently managed to achieve respectable levels of per capita income and industrial output without much help from its banks (Cameron 1967). For Germany, (once the paradigm example for the positive role of banks) there is now a debated hypothesis that the overpraised Kreditbanken misallocated capital and slowed up overall growth (Fremdling and Tilly 1976). In the Czech Crownlands (Bohemia, Moravia, and Upper Silesia) risk-averse banks did little except supply a modicum of working capital. After 1873 Austrian banks 'lost most of their interest in industrial enterprises for years to come.' For Hungary it was not banks but 'unencumbered access to the Austrian capital market' that appears crucial to the beginning of industrialization (Good 1984). Norwegian commercial banks showed little initiative and failed to play any leading role in the process of economic growth. In general it is difficult to find cases where banks clearly played the progressive and entrepreneurial role in national industrialization (Teichova et al. 1997). In analyzing connexions between the state and industrialization Gerschenkron suggested that before 1914 several European governments supplied substitutes for deficiencies in private enterprise, private capital markets, and household demand. Although his observation seems plausible and is based upon historical examples (drawn largely from Russia and Eastern Europe), statistically it has remained untestable because the scale and intensity of governmental intrusion into an otherwise market-led spontaneous process cannot be measured (Gerschenkron 1962). One macroeconomic indicator—the ratio of public consumption to gross national expenditure—shows no tendency to vary in any systematic way across countries (O'Brien 1994).

Instead economic historians have moved on to classify and to attempt to evaluate potential effects on private industrial investment (domestic and foreign) flowing from the different fiscal, tariff, monetary, and exchange rate policies pursued by European governments over the nineteenth and twentieth centuries, and for this topic there is a bibliography of sophisticated studies which apply macroeconomic theory to history (cited in O'Brien 1994).

Nevertheless to confine investigations into the direct and indirect role of governments to their macroeconomic policies would be to neglect the core functions performed by states to promote industrialization, namely the creation and reform of laws and institutions which underpinned the operation and spread of markets for commodities and factors of production. Such markets operated within a framework of rules which regulated cooperation and competition among individuals and enterprises engaged in economic activities. One of the central traditions of 'old' economic history was to study the laws, regulations, property rights, and enforcement procedures which molded and constrained economic behavior over those long spans of history before the era of the French Revolution, when 'inefficient' sets of institutions are perceived to have held European economies inside their production possibility boundaries. That program (associated with the German historical school) has been revived with the aid of theories and a vocabulary borrowed from modern economics (Hoselitz 1961). Neo-liberal institutional economics has been applied to the history of early modern Europe and suggests that frameworks for the operation of commodity and factor markets emerged in some societies (Holland and England) long before others (Spain and Russia) (North and Thomas 1973).

Clearly it remains important to understand how governments constrained the progress of industrialization by defining and enforcing inefficient systems of rules and property rights. The general problem resides in comparing legal frameworks for economic competition and cooperation across countries, in analyzing alterations to laws and property rights over time, and in systematically linking specified variations in law to economic growth. Although the research programme is concentrated on the early modern period when states established (or failed to establish) conditions for subsequent specialization and the widening of markets in the eighteenth and nineteenth centuries, its relevance to a latter era of the 'Second Industrial Revolution' (which witnessed the rise of new industries, such as chemicals, electricity, and motorized transportation, and the concomitant appearance of 'big government' to regulate the operations of corporations, unions, and financial intermediaries) is also clear (Mathias and Pollard 1989). Historians have work to do and are still some way from theorizing or generalizing about connexions between states and industrialization (Tilly 1990).

5. *Conclusions*

A long line of distinguished social scientists and historians (including St. Simon, Schmoller, Weber, Bloch, and Braudel) have offered students of European economic history, theories, concepts, approaches, and typologies designed to help them organize, clarify, and generalize from the vast accumulation of knowledge now in print about the history of European industries. At present when the British Industrial Revolution has re-emerged as a special case, the potential for histories, of European industrialization constructed around the concept of technological diffusion has been degraded (Landes 1969).

From philosophers of the Enlightenment, through Marx to the German historical school, scholars of European industrial development have found it difficult to organize their analyses without recourse to stages of growth. Rostow's stage theory certainly provided a memorable vocabulary but his schema is not considered to be a particularly illuminating way to write industrial history (Rostow 1965). Conceptually, the whole notion of a sharp discontinuity no longer attracts historians of economic growth, inclined to analyze that process with biological rather than aeronautical metaphors (like take-offs) and to explain accelerations (if, when, and where they might be measured) as the product of fortuitous political events, or basically as the final outcome of a protracted process of institutional change and the prior accumulation of physical and human capital (Braudel 1981–4).

Gerschenkron's typology inspired generations of scholars and survives as the best frame of reference for economic historians concerned to place their research within wider and potentially deeper perspectives that emanate from comparisons across Europe. But more often than not, the heuristic value of typologies seems to reside in the organization of evidence that simply fails to fit the country or case into a pattern (Sylla and Toniolo 1991). Meanwhile, and as research upon economies, regions, towns, industries, firms, and factors of production proliferates, no scholar with Gerschenkron's awesome learning and capacity for generalization has emerged to reconstruct an alternative typology which might serve to draw the long history of industrialization into a European frame of reference (Senghaas 1982).

That is why the majority of economic historians continue to pursue research within the framework of nation states and to offer their scholarship as a basis for comparative history. Indeed that program is exemplified by most textbooks in the field, which tend to be surveys of foreign scholarship arranged country by country, industry by industry, input by input. Readers of these analytically sophisticated books in economic history are implicitly asked to make intellectual leaps from national histories moving through long centuries of industrialization to comparative history unfolding across Europe. They will certainly find scholarly and detailed elaborations on 'connexions' between techniques of production, transport facilities, types of financial intermediation and state policies, foreign trade, urbanization, institutions, legal systems, cultures, on the one hand; and the pace and pattern of industrialization, on the other. Alas, their connexions can rarely be specified, investigated, and quantified in ways that might suggest how observed variations across regions and countries in the manifold factors promoting or restraining industrialization through time might be 'tested' for significance.

Meanwhile (and after two generations of postwar research on the economic histories of Africa, the Near East, South East and South East Asia) another metanarrative concerned with the 'Great Divergence' or why Europe industrialized ahead of economies located on other continents is beginning to generate syntheses, generalizations, and exciting controversies for historians to pursue (Landes 1998, Frank 1998, Pomeranz 2000).

See also: Development, Economics of; Development: Social; Development: Socioeconomic Aspects; Economic History; Financial Institutions in Economic Development; Industrial Society/Post-industrial Society: History of the Concept; Industrialization; Modernity; Modernity: History of the Concept; Modernization and Modernity in History; Modernization, Sociological Theories of

Bibliography

Aldcroft D, Sutcliffe A (eds.) 1999 *Europe in the International Economy 1500–2000*. Elgar, Cheltenham, UK

Bairoch P 1976 *Commerce Exterieur et Developpement Economique de l'Europe an XIXe Siecle*. Mouton, Paris

Berg M, Bruland K (eds.) 1998 *Technological Revolutions in Europe*. Elgar, Cheltenham, UK

Braudel F 1981–4 *Civilization and Capitalism 15th–18th Century*. Harper & Row, New York

Cameron R 1967 *Banking in the Early Stages of Industrialization*. Oxford University Press, New York

Cipolla C H (ed.) 1973–4 *The Fontana Economic History of Europe, Vols 2, 3, 4(I), 4(II)* Collins, London

Crafts N F R 1985 *British Economic Growth during the Industrial Revolution*. Oxford University Press, Oxford, UK

de Vries J 1984 *European Urbanization 1500–1800*. Methuen, London

Floud R, McCloskey D (eds.) 1994 *The Economic History of Britain since 1700, Vol. 1 1700–1860*. Cambridge University Press, Cambridge, UK

Frank G 1998 *ReOrient: Global Economy in the Asian Age*. University of California Press, London

Fremdling R, Tilly R 1976 German banks and German growth and economic history. *Journal of Economic History* **36**: 416–24

Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Harvard University Press, Cambridge, MA

Good D 1984 *The Economic Rise of the Habsburg Empire*. University of California Press, Berkeley, CA

Hicks J 1969 *A Theory of Economic History*. Clarendon Press, Oxford, UK

Hoselitz B 1961 *Theories of Economic Growth*. Free Press, New York

Jones E 1988 *Growth Recurring. Economic Change in World History*. Oxford University Press, Oxford, UK

Kriedte P et al. 1981 *Industrialization before Industrialization*. Cambridge University Press, Cambridge, UK

Kuznets S 1965 *Economic Growth and Structure*. Norton, New York

Landes D S 1969 *The Unbound Prometheus. Technological Change and Industrial Development in Western Europe*. Cambridge University Press, Cambridge, UK

Landes D S 1998 *The Wealth and Poverty of Nations*. W. W. Norton, New York

Mathias P, Davis J (eds.) 1990–8 *The Nature of Industrialization, Vols. 1–6*. Blackwells, Oxford, UK

Mathias P, Pollard S (eds.) 1989 *Cambridge Economic History of Europe*, Vol. 8. Cambridge University Press, Cambridge, UK
Mokyr J (ed.) 1985 *The Economics of the Industrial Revolution*. Rowan & Allanheld, Totowa, NJ
Mokyr J 1990 *Lever of Riches: Technological Creativity and Economic Progress*. Oxford University Press, New York
Musgrave P 1999 *The Early Modern European Economy*. St. Martin's Press, New York
North D C, Thomas R P 1973 *The Rise of the Western World*. Cambridge University Press, Cambridge, UK
O'Brien P K (ed.) 1994 *The Industrial Revolutions in Europe* (2 vols). Blackwell, Oxford, UK
O'Brien P K (ed.) 1998 *Industrialization. Critical Perspectives on the World Economy* (4 vols). Routledge, London
O'Brien P K, Quinnault R (eds.) 1993 *The Industrial Revolution and British Society*. Cambridge University Press, Cambridge, UK
Parsons T 1971 *The System of Modern Societies*. Prentice-Hall, Englewood Cliffs, NJ
Pollard S 1981 *Peaceful Conquest. The Industrialization of Europe 1760–1970*. Oxford University Press, Oxford, UK
Pollard S 1990 *Typology of Industrialization Processes in the Nineteenth Century*. Harwood Academics, New York
Pomeranz K 2000 *The Great Divergence: China, Europe and the Making of the Modern Economy*. Princeton University Press, Princeton, NJ
Rostow W W 1965 *The Stages of Economic Growth*. Cambridge University Press, Cambridge, UK
Senghaas 1982 *The European Experience: A Historical Critique of Development Theory*. Leamington Spa, UK
Stearns P 1993 *The Industrial Revolution in World History*. Westview Press, Boulder, CO
Sylla R, Toniolo G (eds.) 1991 *Patterns of European Industrialization*. Routledge, London
Teich M, Porter R (eds.) 1996 *The Industrial Revolution in National Context*. Cambridge University Press, Cambridge, UK
Teichova A, Van Hentenryk G K, Ziegler D 1997 *Banking, Trade and Industry*. Cambridge University Press, Cambridge, UK
Tilly C 1990 *Coercion, Capital and European States, AD 990–1990*. Blackwell, Oxford, UK
Weber M 1961 *General Economic History*. Collier Books, New York
Wrigley A 1988 *Continuity, Chance and Change: The Character of the Industrial Revolution in England*. Cambridge University Press, Cambridge, UK

P. K. O'Brien

Industrial–Organizational Psychology: Science and Practice

1. *What is I–O Psychology?*

The objective of this article is to describe the breadth of topics studied by industrial–organizational (I–O) psychologists, with an emphasis on what I–O psychologists do to improve organizational effectiveness. I–O psychology is an applied branch of psychology dealing with people's attitudes, behavior, cognition, and emotions at work. More specifically, it is the science of collecting, analyzing, and using data to help organizations make better decisions about the selection, utilization, and management of workers. Non-academic I–O psychologists may work for private consulting firms, large private firms, or government agencies (including the military). Regardless of the setting, the I–O psychologist most frequently operates in an advisory capacity, conducting research that leads to the implementation of a new human resource technology or organizational strategy, or evaluating the impact of an ongoing one. For example, an I–O psychologist may design a new work simulation for selecting customer service representatives, conduct an employee attitude survey to determine the primary determinants of turnover, or evaluate the effectiveness of a management development training program. In the USA, a graduate degree (master's or PhD) is typically required to be considered an I–O psychologist. In other countries, such as those in Western Europe, practitioners often hold master's degrees, while researchers and professors hold PhDs. There are over 100 graduate programs in the USA that offer specialized degrees in I–O, and approximately 30 more in Australia, Canada, and Europe.

2. *A Brief History of I–O Psychology*

The notion that organizational efficiency can be improved by classifying workers and matching tasks to worker attributes can be traced in time back through the writings of Adam Smith, John Locke, Thomas Hobbes to Plato and Aristotle (see Katzell and Austin 1992). Yet, at the beginning of the twentieth century, there was no formal discipline resembling what is now known as industrial–organizational psychology. In 1903, W. L. Bryan gave his presidential address to the American Psychological Association, in which he encouraged psychologists to examine the activities and functions of everyday life. Given the centrality of work to everyday life, this can be construed as a formal call for the application of psychology to industry. Interestingly, in the published version of this address, Bryan referenced an earlier speech in which he called for more research on *individual* psychology. However, Bryan mistakenly wrote *industrial* psychology and did not catch his mistake. Thus, the first instance of the term industrial psychology is as a typographical error (Muchinsky 1997).

The German Psychologist Hugo Munsterberg is considered by many to be the father of industrial psychology (Katzell and Austin 1992). His 1913 book addressed the application of psychological methods to practical problems in industry, including selecting workers, improving worker safety, designing work

environments, and using psychology in sales. Concurrently, industrial engineers Frederick Taylor and Frank Gilbreth advocated job design, training, and selection programs to improve the productivity and efficiency of employees. By the early 1920s, principles for the selection of employees developed to a point that remains constructive to this day. For example, Freyd (1923) presented the following five assumptions of personnel selection: (a) people have measurable abilities, skills, and habits that are relatively stable; (b) people differ in these abilities and these differences should be the basis for decisions about jobs—all other things being equal, those with higher levels of abilities relevant to a job can be expected to perform better than those with less ability; (c) relative differences in ability remain relatively constant even after training or experience; (d) different jobs require different ability—effective selection requires matching abilities required by a job to abilities held by individuals; (e) abilities can be measured and used to rank-order job candidates. Freyd also spoke of the need to conduct job analyses and validate instruments used for selection.

Beginning in the 1930s, the field of industrial psychology broadened with the advent of the so-called human relations movement. Charges of the exploitation of workers by Taylor gave rise to the need to understand and support workers' psychological needs. Instrumental in this shift was a series of studies conducted at the Western Electric plant in Hawthorne, Illinois between 1924 and 1932. Although much of the original data were lost, never reported, or, in fact, ran contrary to the phenomenon that bore the town's name, the studies gave rise to the so-called 'Hawthorne Effect' (Rice 1982). The Hawthorne Effect, that research participants will perform better simply because of the attention paid to their behavior, probably never occurred, and has never been intentionally replicated. Nonetheless, through the 1940s and 1950s, it provided a strong impetus to the human relations movement in industry by convincing organizations that improving job conditions and attending to worker needs would increase worker motivation and productivity.

Through the 1950s and 1960s, greater numbers of researchers were being trained in I–O. Psychology as a whole was dominated by principles of behaviorism, and stimulus–response paradigms affected theory development in I–O. Given these two influences, there arose a number of behaviorally based taxonomies or theories of leadership (Stogdill 1963), motivation (e.g., Herzberg 1968, McGregor 1960), group behavior (Lewin 1951, Cartwright and Zander 1968), worker attributes (Fleishman 1965, 1972, McCormick et al. 1969), performance appraisal (Smith and Kendall 1963), and training (Gagne 1962). While cognitive influences on behavior and performance have slowly begun to influence theory and practice since the late 1980s, a number of these theories have held up reasonably well over the years.

3. *What Does an I–O Psychologist Do?*

As may be inferred from the prior discussion, the field traditionally has been divided into two broad subdisciplines: industrial and organizational psychology.

3.1 *Industrial Psychology*

Topics traditionally considered with the domain of industrial psychology include (a) the evaluation of representative tasks and requirements of jobs, (b) the assessment of individual attributes and skills, and (c) the process of matching individuals to jobs, given (a) and (b). Specific topics generally classified under industrial psychology include: job analysis and job evaluation, recruitment, test development and validation, personnel selection and placement, training and development, and performance appraisal and performance management. Through these processes, an organization comes to understand the core attributes of jobs, identify potential applicants inside and outside the organization, build selection and screening methods that optimize its ability to find the best available applicant, train new workers to proficiency, and monitor and evaluate performance to ensure employee development and satisfactory performance on the job.

Since 1980, the field of industrial psychology has undergone a period of unification driven by the assumption that successful hiring and placement decisions can be made based on a fairly narrow set of worker characteristics (called knowledge, skills, abilities, and other personnel characteristics, or KSAOs). From 1965 to 1980, much of the research and practice surrounding personnel selection was influenced by the Civil Rights Act of 1964 in the USA. The Civil Rights Act made it illegal to discriminate on the basis of race, gender, religion, or national origin, unless done as a business necessity. Through a series of later court decisions, business necessity was interpreted as job relevant which in turn meant that the organization charged with discrimination must provide logical or statistical evidence (through test validation) that performance on the challenged test was related to job performance. The Civil Rights Act has been humorously referred to as the full employment act for I–O psychologists, but it ultimately led to the reinforcement of a long-held tenet that the determinants of job performance are situationally (i.e., organizationally) specific and that job analysis, test development, and test validation are required each time a new selection system is implemented.

Unification began in the mid to late 1970s following the popularization of meta-analysis as a tool for aggregating the results from multiple studies (see *Meta-analysis: Overview*). Meta-analysis, or validity generalization in the vernacular of selection researchers, applies common statistics (e.g., mean and

standard deviations) to the outcomes of multiple studies, rather than to the responses of subjects in single studies. Further, by estimating and removing the impact of extraneous sources of error such as sampling error or differences across studies in test reliability, a meta-analysis or validity generalization study may lead to the conclusion that a particular type of test predicts job performance regardless of the setting or type of job (Schmidt and Hunter 1977). This indeed was the conclusion reached by two I–O psychologists, Schmidt and Hunter, who developed modern validity generalization techniques and conducted multiple large scale meta-analyses on common employment tests (e.g., Schmidt et al. 1986). While debate continues on both their methodology and sweeping conclusions, they had a significant impact on moving the field away from a doctrine of situational specificity and toward one that is beginning to recognize common attributes or worker requirements across jobs as well as a common model of job performance across jobs.

I–O psychologists at the beginning of the twenty-first century might intervene in an organization much as they would have 30, 50, or even 70 years ago: using an analysis of common or critical tasks and requisite KSAOs to construct a selection test battery, performance appraisal, or training program. Popular selection methods include interviews, paper-and-pencil tests, or work samples. However, the movement towards unification has also resulted in emerging models or taxonomies of job requirements, job performance, and worker characteristics that offer promise for streamlining these processes and enabling I–O psychologists to focus on other strategies for improving organizational effectiveness. For example, John Campbell (1990) has proposed a multidimensional model of job performance applicable to most, if not all, jobs. Campbell suggests that performance in any job is a function of three determinants: declarative (or factual) knowledge, procedural knowledge (operative skill or know-how), and motivation. Further, job performance can be described by all or a subset of eight higher-order factors: job specific task proficiency (e.g., a carpenter must be able to join two pieces of wood), non job-specific task proficiency (e.g., cleaning a work site), written and oral communication, demonstrating effort, maintaining personal discipline, facilitating peer and team performance, supervision/leadership, and management/administration. Borman and Motowidlo(1993) further categorized successful job performance into two behavioral classes: task-related activities and what they term contextual behavior—activities such as regularly coming to work on time, helping others, or demonstrating organizational citizenship. To evaluate performance in most jobs, I–O psychologists can derive assessment instruments rooted in these eight factors and involving multiple technologies such as objective performance metrics or subjective performance evaluations (e.g., ratings by supervisors or customers; see *Performance Evaluation in Work Settings*).

Arising from the meta-analytic work of Schmidt and Hunter and others, emerging evidence suggests that adequate (though not optimal) selection systems can be built using a simple combination of a general intelligence test and a measure of personality (Behling 1998). Cognitive ability is an excellent predictor of performance in most, if not all jobs. More intelligent people learn to perform the job better than do less intelligent persons and typically perform better after training (Ree and Earles 1992). Certain fundamental personality traits, particularly conscientiousness and emotional stability, have been consistently reported to add significant incremental variance in the prediction of job performance, both in the USA (Barrick and Mount 1991) and Europe (Salgado 1997). Considering the taxonomies of Campbell (1990) and Borman and Motowidlo (1993), it is not difficult to imagine a generalized selection system in which cognitive ability measures are used to predict the acquisition of the declarative knowledge and procedural knowledge necessary to perform task-related activities, and personality measures are used to assess motivational determinants of citizenship and certain nontask-specific performance dimensions (e.g., facilitating team performance).

Even the process of job analysis—the foundation for decisions about test development, performance appraisal, and training design—is becoming streamlined and driven by a single taxonomy (see *Job Analysis and Work Roles, Psychology of*). With the aid of a team of I–O psychologists (Peterson et al. 1995), the US government has developed O*NET, a comprehensive source of occupational information available to employers, job seekers, trainers, and analysts. For a database of over 1,000 job titles. O*Net contains data that include training and experience requirements, worker characteristics (abilities, values, and work styles), worker requirements (e.g., technical or problem-solving skills), and occupational-specific information (e.g., labor market forecasts). Though not yet fully developed or publicly accessible, O*NET promises to be a valuable resource for identifying selection criteria, planning staffing requirements, and educating individuals seeking employment opportunities.

3.2 Organizational Psychology

Organizational psychology is the second broad subdiscipline recognized with I–O psychology. Topics traditionally considered within the domain of organizational psychology include: (a) internal states of individuals related to jobs (e.g., work motivation, job satisfaction, and organizational commitment); (b) determinants of effective interpersonal interactions

within organizations (e.g., leadership, group behavior, and team performance); and (c) the design of work, organizational structures, and organizational systems that facilitate individual and organizational effectiveness. As can be seen from this topic list, the potential impact of organizational psychology is far broader than industrial psychology, ranging from values, emotions, and cognitions within a person through the dynamics of interpersonal communication (in groups, teams, or leader/member dyads) to systemwide interventions such as the creation of a continuous learning environment (see *Organizational Behavior, Psychology of*).

As a field of study, it is easy to relate the field of organizational psychology to the experiences of most persons. Why does someone love or hate his or her job? Are satisfied workers better job performers? How does job stress undermine job performance and does job stress spill over into home life? What makes someone an effective leader? Why are some tasks intrinsically interesting and others boring? How can some organizations consistently sustain employee growth and firm performance? While the questions asked by organizational psychologists feel familiar, it is harder to capture or describe the methods used by organizational psychologists to improve organizational effectiveness. Very often, interventions are systemwide, their casual impact on measurable outcomes are not readily apparent. For example an I–O psychologist may conduct an attitude survey to study reasons for high turnover among entry-level workers, then recommend an array of system changes including employee flextime, new compensation plans, self-managed work teams, and additional communication training for managers. Given seemingly random fluctuations in product demand, the overall economy, competitors' strategies, and available labor pools, the relationship between the I–O intervention and reduced turnover may be difficult to isolate.

One area of organizational research that has produced consistently positive findings (and successful interventions) is the goal-setting theory of work motivation. While goal setting cannot account for all motivated behavior, as first articulated by Locke (1968), goal-setting theory simply states that employees work harder and perform better with a goal than without. Moreover, more difficult and more specific goals result in higher motivation (and better performance) than easier or less specific goals; thus social workers who set a goal of helping 50 clients per-month will perform better than if they had set a goal of 25 clients or simply 'trying their best.' Goal-setting theory has been supported by considerable field and lab research, and holds valuable practical implications for managers. Managers can motivate employees by helping them set specific goals and by gaining commitment to goal attainment.

While I–O psychologists have been able to provide valuable tools to managers and leaders such as goal setting and group decision-making strategies, the determination of what makes someone an effective leader has proven elusive. Researchers in the 1940s sought to define physical, intellectual, or personality traits of successful leaders. After failing to come to consensus on leader traits, researchers in the 1950s focused on what leaders did, seeking to find a set of effective leader behaviors (see Stogdill 1963). Fiedler (1967) was the first researcher to advocate a contingency approach—what leader behaviors were the most effective depended on the situation (often defined in terms of task difficulty, characteristics of followers, and/or forms of environmental uncertainty or strain). While most modern leadership theories can be characterized as contingency models, there is little agreement as to what is the best approach.

As organizations move into the twenty-first century, the nature of work and characteristics of the workforce appear to be changing rapidly and already are greatly affecting the research of organizational psychologists. Consider the implementation of computers and technology. Not only does the automation of work change worker requirements (e.g., telephone workers who once carried bundles of wire up poles now program fiber optic relay changes on the computer), but hold implications for where work is done (e.g., telecommuting), how information is shared (e.g., the use of video conferencing and e-mail), and even who or what piece of equipment is considered part of a team (e.g., should a robot be invited to an autoworker's retirement party?).

Demographic population shifts have resulted in a workforce that, compared to 30 years ago, is older, more ethnically diverse, and more likely to be female. Not surprisingly, distinctions between work and nonwork settings are blurring, and I–O psychologists are being asked to help organizations manage the impact of diversity, work–life balance, and the introduction of new technology on member satisfaction and productivity (Igbaria and Tan 1998, Lobel 1999). Examples of this type of work include evaluating the cost effectiveness of providing an onsite daycare or exercise facility, training managers to interact effectively with members of diverse ethnic groups, or surveying members for indicators of lower organizational commitment given perceived violations of psychological contracts by the organization (Morrison and Robinson 1997).

It is fair to criticize the overall impact of I–O research on the selection and management of workers. For example, organizations in the USA use unstructured interviews for selection, and organizations in some European countries routinely use handwriting analysis, despite considerable evidence by I–O researchers that neither method is reliable or valid. However, I–O psychologists are continually seeking new ways to reach and educate organizations regarding effective management practices. As we have seen in this article, as new methods are found for

bridging the gap between science and practice, I–O psychologists are prepared with generalized principles of human performance, and ever-expanding research on the dynamics of the attitudes, cognitions, and behaviors of individuals in organizations.

See also: Ethics in Organizations, Psychology of; Group Processes in Organizations; Industrial and Organizational Psychology: Cross-cultural; Industrial Sociology; Job Analysis and Work Roles, Psychology of; Job Design and Evaluation: Organizational Aspects; Job Design, Psychology of; Leadership in Organizations, Psychology of; Organizational Behavior, Psychology of; Organizational Climate; Personnel Selection, Psychology of; Psychological Climate in the Work Setting; Stress in Organizations, Psychology of; Work, Sociology of

Bibliography

Barrick M R, Mount M K 1991 The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology* **44**: 1–26

Behling O 1998 Employee selection: Will intelligence and conscientiousness do the job? *Academy of Management Executive* **12**: 77–86

Borman W C, Motowidlo S J 1993 Expanding the criterion domain to include elements of contextual performance. In: Schmidt N, Borman W C (eds.) *Personnel Selection*. Jossey-Bass, San Francisco, pp. 71–98

Campbell J P 1990 Modeling the performance prediction problem in industrial and organizational psychology. In: Dunnette M D, Hough L M (eds.) *Handbook of Industrial and Organizational Psychology*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA, Vol.1, pp. 687–732

Cartwright D, Zander A (eds.) 1968 *Group Dynamics*: *Research and Theory*, 3rd edn. Harper and Row, New York

Fiedler F 1967 *A Theory of Leadership Effectiveness*. McGraw-Hill, New York

Fleishman E A 1965 The description and prediction of perceptual-motor skill learning. In: Glaser R (ed.) *Training Research and Education*. University of Pittsburgh Press, Pittsburgh, PA, pp. 137–75

Fleishman E A 1972 On the relation between abilities, learning, and human performance. *American Psychologist* **27**: 1017–32

Gagne R M 1962 Military training and principles of learning. *American Psychologist* **17**: 83–91

Freyd M 1923 Measurement in vocational selection: An outline of research procedure. *Journal of Personnel Research*. **2**: 215–49, 268–84, 377–85

Herzberg F 1968 One more time: How do you motivate employees? *Harvard Business Review* **46**(1): 53–62

Igbaria M, Tan M 1998 *The Virtual Workplace*. Idea Group Publishing, Hershey, PA

Katzell R A, Austin J T 1992 From then to now: The development of industrial–organizational psychology in the United States. *Journal of Applied Psychology* **77**: 803–25

Lewin K 1951 *Field Theory in Social Science*. Harper and Row, New York

Lobel S A 1999 Impacts of diversity and work-life initiatives in organizations. In: Powell G (ed.) *Handbook of Gender and Work*. Thousand Oaks, CA, Sage, pp. 453–74

Locke E A 1968 Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance* **3**: 157–89

McCormick E J, Jeanneret P R, Mecham R C 1969 *A Study of Job Characteristic and Job Dimensions as Based on the Position Analysis Questionnaire*. Occupational Research Center, Department of Psychological Sciences, Purdue University, Report No. 1, Lafayette, IN

McGregor D 1960 *The Human Side of Enterprise*. McGraw-Hill, New York

Morrison E W, Robinson W L 1997 When employees feel betrayed: A model of how psychological contract violation develops. *Academy of Management Review* **32**: 226–56

Munsterberg H 1913 *Psychology and Industrial Efficiency*. Houghton-Mifflin, Boston

Peterson N G, Mumford M D, Borman W C, Jeanneret P R, Fleishman E A (eds.) 1995 *Development of a Prototype Occupational Information Network* (*O*NET*) Utah Department of Employment Security, Salt Lake City, UT, Vols.1–2.

Ree M J, Earles J A 1992 Intelligence is the best predictor of job performance. *Psychological Science* **1**: 86–9

Rice B 1982 The Hawthorne effect: Persistence of a flawed theory. *Psychology Today* **31**(2): 70–4

Salgado J F 1997 The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology* **82**: 30–43

Schmidt F L, Hunter J E 1977 Development of a generalized solution to the problem of validity generalization. *Journal of Applied Psychology* **43**: 529–40

Schmidt F L, Hunter J E, Outerbridge A N 1986 Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology* **71**: 432–9

Smith P C, Kendall L M 1963 Retranslation of expectations. *Journal of Applied Psychology* **47**: 149–55

Stodgill R M 1963 *Manuals for the Leader Behavior Description Questionnaire–Form XII*. Ohio State University, Columbus, OH

K. Kraiger

Inequality

The most straightforward definition of inequality is 'differences among people in their command over social and economic resources.' To be useful, however, one must go further and specify inequality: (a) of what, (b) among whom, and (c) how summarized. Furthermore, interest in 'inequality' is often prompted by a concern with 'inequity,' which can be defined as 'morally or ethically unjustifiable differences among people in their command over resources.' Although measurement of inequality often cannot avoid the problem of choosing which differences among people are of greatest concern, and that choice inevitably involves some values, in principle the discussion of inequality concerns issues of fact. In general, it is useful to distinguish as clearly as possible between analytic or descriptive statements about social in-

equality—i.e., how society *is*—and statements about inequity which involve moral evaluations of how society *should be*.

1. Inequality of What?

Although analysts of inequality would often like to discuss inequality of well-being, utility, or happiness, such differences among people in their internal psychological satisfactions must typically be inferred from observable inequalities in social and economic resources. Academic disciplines also tend to specialize somewhat in their focus. While political scientists emphasize the inequality of political power, social inequality in prestige and status is the domain of sociologists. Economists typically limit their analysis to the inequality of income or wealth or consumption. However, there has been a notable broadening of economic perspectives to the inequality of capabilities or economic functionings (see Sen 1985, 1992).

Inequalities in power, status, income, or wealth clearly interact, but these are also analytically distinguishable dimensions of inequality, which are not necessarily perfectly correlated. Furthermore, whatever the dimensions of inequality considered, many analysts would think it important to distinguish between inequality of opportunity (i.e., inequality in the sets of potential choices open to individuals) and inequality of result (i.e., inequality in the specific outcomes actually observed).

The appropriateness of a distinction between inequality of opportunity and inequality of outcome often depends on the time frame of analysis. In access to housing, for example, at any given point of time there will be some persons who are homeless and some who are adequately, or perhaps munificently, housed. These outcomes will depend partially on the choices that were available in earlier periods of time to each individual and on the events they have experienced. In the housing example, an analyst who stressed inequality of opportunity would want to know how many people ever had the opportunity to buy a home, and the income to keep up payments, while an emphasis on inequality of outcome would simply focus on the percentage of current homeowners.

The distinction between inequality of opportunity and of outcome is particularly important for intergenerational issues. In a multigenerational context the inequality of outcomes experienced by a given generation of parents will largely determine the opportunities which their children's generation face. By unconscious socialization and imitation, or by conscious choice of child-rearing practices, time allocation, schooling, and material bequest, parents pass social, cultural, human, and financial capital on to their children—but they can only do so to the extent that they have acquired such assets themselves. Inequality of outcome among parents, and the choices made by them, therefore condition the inequality of opportunity experienced by their children.

In the analysis of inequality, there is a great deal of complexity (and potential for ambiguity) in the empirical operationalization of broad concepts like power, status, income, or wealth. In general, it is important to distinguish between stocks (the total resources available at a point in time) and flows (the change in resources over time). Although the case can also be comparably made for social and political resources like power or status, inequalities in economic resources offer a particularly clear example.

Wealth differs from income precisely in the stock/flow distinction. Wealth is the total stock of economic resources—total potential consumption—that an individual has at a point in time. The classic Hicksian definition of income, on the other hand, is the maximum amount that a person could consume during a period of time and still be as well off at the end of the period as at the beginning. Measurement of wealth therefore requires specification of the date of measurement while income measurement is always for a specific period of time.

In practice, the accounting period over which one measures income flows is far from innocuous. Analysts have to choose whether it is lifetime, annual, or shorter period (e.g., monthly) income that should be measured. The underlying reason for studying inequality should determine the appropriate period over which to measure income flows, but the choice will influence measures of inequality. Over any individual's lifetime, some changes in income will follow a predictable pattern (for example, when workers get raises based on seniority), while others (such as the income losses caused by illness or unemployment) may be less predictable. If income is measured over a longer interval of time, periods of low and high income will be averaged, implying less measured inequality among income recipients.

The most common measure of economic resources used in analysis of economic inequality is annual money income, which includes cash earnings, rental interest and dividend income, and cash transfers from government. Money income can be calculated before or after tax, and in countries with progressive income tax systems the choice makes a significant difference to measured inequality. However, in a complex modern economy, it is also common to find complex cases in which the calculation of annual money income is not straightforward. For example, the cash flow of self-employed individuals or entrepreneurs typically has to be adjusted to reflect the depreciation of the capital they use in production. Since stock options and capital gains are very important for upper income groups, calculation of the level of income inequality is sensitive to whether income is calculated to include only the value of stock options actually exercised in the current period, or to include the value of unrealized capital gains.

Furthermore, a full definition of 'income' would include the value of nonmarketed goods and services received, as well as money income. In countries in which much of the population is agricultural, the value of food produced and consumed within the household may be a large fraction of total income. Individuals also receive a benefit from leisure time enjoyed, the goods and services which (like child care) are produced in the household, and the services of owner-occupied houses. They may also benefit from services (like medical care or education) provided by government agencies. Arguably, the value of these benefits should be added to cash income.

In addition, individuals who live in larger households benefit from a greater ability to share resources (e.g., due to household 'economies of scale,' a couple with a household income of $40,000 is better off than two individuals who each have $20,000 and must live separately). If the objective is to measure inequality in economic well-being, some account should be taken of living arrangements. Economists have therefore argued that 'equivalent income,' which adjusts income for household size, is the best measure to use in the analysis of inequality in economic well-being.

The complexities in measuring 'income' are a reminder that in measuring inequality, 'the devil is in the details.' Although 'income' is a complex concept, income inequality is only part of economic inequality more broadly conceived. Moreover, since institutions change over time and differ across countries, comparisons of inequality can often be sensitive to these measurement issues—and in the analysis of inequality, comparative statements on time trends or differences among societies are the major issue of interest.

2. *Inequality Among Whom?*

Because assessing the extent of inequality involves a comparison of resources, it is crucial to be clear about whose resources are being compared—and which resources. An analyst who looks only at inequality in family or household income is ignoring the fact that families and households differ significantly in size. On the other hand, if an analyst were to look only at the distribution of money income directly received by individual persons, the fact that many people (e.g., young children, unpaid spouses) have zero cash income of their own would tend to exaggerate measured inequality.

Because individuals without income of their own typically reside in households in which other family members do have cash income, the effective degree of sharing of income within households is crucial to inequality measurement. However, this is difficult to observe in practice, and it is often convenient for researchers to assume that all persons within a household have the same level of effective income (Phipps and Burton 1995 demonstrate that alternative assumptions can make a big difference to perceived poverty). Similarly, it is common to assume that all members of a given household have the same socioeconomic status. Effectively, such procedures amount to assuming away inequality within groups of people (in this example, coresident household members) and concentrating attention on inequality between groups.

Table 1
A comparison of inequality in two hypothetical societies

	Men			Women		
Arcadia	4	6	8	2	4	6
Bucolia	0	6	12	6	6	6

Assuming equality within groups can be problematic (e.g., much feminist scholarship has challenged the presumption of within-family equality in consumption, status, and power). Furthermore, the larger the group is, the harder it generally becomes to maintain the presumption of within-group equality. There is much social interest in the extent of inequality between broad groups of people (e.g., occupational groups, regions, racial or ethnic groups, age cohorts, etc.). In some instances, between-group comparisons are driven by an underlying theoretical framework which defines categories of interest (e.g., bourgeoisie/proletariat). However, when within-group inequality is large relative to that between groups (as in income comparisons between age groups, where inequality among individuals of the same age is far greater than the differences in average income between people of different ages), comparisons of average group characteristics may suggest unwarranted stereotyping.

Furthermore, the term 'inequality' can be used in two subtly different conceptual senses. In some discussions—discrimination is an example—'inequality' is used to refer to differences between particular individuals or groups (e.g., person A is not equal to person B). In other instances, 'inequality' refers to differences among people in general (e.g., in country A there is more inequality than in country B). These usages need to be distinguished, since pairwise and group inequalities do not necessarily imply aggregate inequality differences. Table 1 presents a numerical example of men and women in 'Arcadia' and 'Bucolia' to make the point.

Is there more inequality in Arcadia than in Bucolia, or less? Would eliminating between group inequality imply an end to inequality? Inequality among groups of people is illustrated by comparison of men and women in Arcadia and Bucolia. In Arcadia, women have a lower average income (4) than men (6), while in Bucolia the average is the same (6). However, the equality of male and female average income in Bucolia does not imply equality of individual incomes. Inequalities in average income between groups also do not imply individual inequalities (in Arcadia, for

example, two-thirds of men can be matched with women with exactly the same income while in Bucolia only one match can be found).

To answer the seemingly simple question of whether, overall, there is more inequality in Arcadia than in Bucolia, or less, one must compare the entire distribution of outcomes. Although there is a large middle class in Bucolia (two-thirds of the inhabitants have the same outcome (6)), there are also greater extremes (from 0 to 12). In Arcadia, middle-class incomes are more differentiated, but the extremes are not as widely separated. Hence, answering the seemingly simple question whether there is more or less inequality requires finding a way to summarize the number and size of differences in outcomes among all individuals.

3. *Inequality—How Summarized?*

The example of Arcadia and Bucolia has been simplified by the presentation of only one outcome, among a small number of individuals. In actuality, modern societies have millions of inhabitants who differ on a wide number of dimensions. There are, therefore, a very large number of potential comparisons. Even if one restricts attention to a single variable, like income, it is not a simple matter to summarize the extent of inequality.

One approach to inequality measurement is to order the population from worst off to best off, divide them into groups of equal size and calculate the share of total resources flowing to each group. (If five groups are used, the result is the calculation of five 'quintile' shares of total resources.) If, for all low income groups, this calculation always produces the answer that the less well off get a larger share of total income in country A, then it is easy to say that country A has less inequality. However, in the Arcadia/Bucolia comparison, the poorest sixth of Bucolia get a smaller share (zero compared to 6.6 percent), at the same time as the second poorest sixth of Bucolia get a larger share (16.6 percent compared to 13.2 percent). Which country has more inequality? If one wishes to know whether there is more inequality in country A than in country B, or whether there is more inequality at time A than at time B, inequality must be summarized in a single number.

The most popular index summarizing the extent of inequality within society is probably the Gini index. Mathematically, the index is calculated as the average differential (in absolute value) between any two people in society, as in Eqn. (1).

$$G = \frac{1}{2\,\bar{y}\cdot n\cdot(n-1)}\cdot\sum_{i}^{n}\cdot\sum_{j}^{n}|y_i - y_j| \qquad (1)$$

The Gini index also has a convenient graphical interpretation. When individuals are ordered from poorest to richest, Fig. 1 graphs the relationship between the cumulative percentage of income received on the horizontal axis, and the cumulative percentage of population on the vertical axis—the result is known as a 'Lorenz curve.' If everyone in society had the same income—perfect equality—then the Lorenz curve would be the straight line OY. If one person in society had all the income and the rest had nothing—perfect inequality—the Lorenz curve would lie along the horizontal axis for almost all its length.

Actual societies are somewhere between perfect equality and perfect inequality and as an example Fig. 1 graphs the Lorenz curve for Sweden and the USA. A convenient feature of the Gini index is that it is equal to the ratio of the area between the Lorenz curve and the line of perfect equality (area A in Fig. 1) and the triangle OPY. (In the early 1990s, the Gini index of equivalent after-tax money income of Sweden was 0.222, substantially less than the Gini for the USA of 0.371; see Osberg and Xu 2000.)

As Fig. 1 indicates, the Lorenz curve for Sweden lies entirely inside the Lorenz curve for the USA. The verbal interpretation of such a case is that the poorest x percent of individuals in Sweden always have a larger share of total income than the poorest x percent in the USA—whatever the value of x. If this is true, it is straightforward to say that there is less inequality in Sweden than in the USA. However, when Lorenz curves cross, comparisons between societies become sensitive to which part of the income distribution is of greater interest.

The dashed line in Fig. 1 is drawn to represent 'Bucolia,' in which the bottom sixth of individuals get nothing, the middle two-thirds share equally two thirds of total income, and the top sixth get the remaining third of total income. The Lorenz curve for this hypothetical society crosses the Lorenz curve for both Sweden and the USA. The poorest sixth of individuals in this hypothetical society get a smaller share of national income than they would in actual Sweden or USA, while the second and third poorest sixths get substantially more. In hypothetical 'Bucolia' the bottom half of the income distribution gets 33 percent of total income. In the real world of 1994, in the USA the bottom half of the income distribution received, after taxes and transfers, is about 24.8 percent of total equivalent income (for full details of methodology and data see Osberg 2000). Which is the more unequal?

When Lorenz curves cross, it matters which part of the income distribution one thinks to be most important. If one thought that the most important aspect of inequality was the share of income received by the very poorest, then Bucolia should be counted more unequal than Sweden or the USA. However, the Gini index for Bucolia is 0.139—well below that in either Sweden or the USA. If inequalities among 'most people'—i.e., those in the middle part of the income distribution—are thought to be the most important

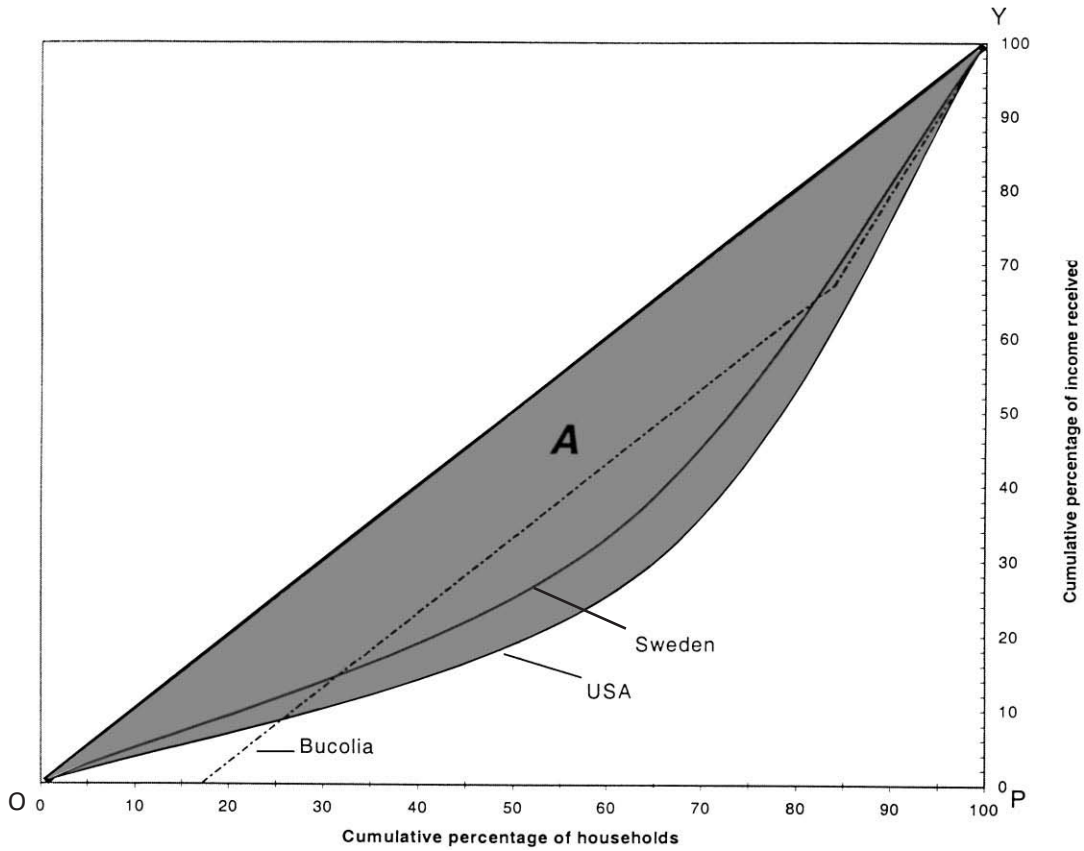

Figure 1
Relationship between cumulative percentages of income and population

issue, then the low Gini index in Bucolia is the appropriate measure. In Bucolia, two-thirds of people have exactly the same income, and by this criterion one would count Bucolia as considerably more equal than either the USA or Sweden. However, if inequality between the poorest and everyone else is thought to be the important issue, one would disagree with the Gini ranking. When Lorenz curves cross, a different ranking of inequality may be obtained, depending on which part of the distribution is of greater concern. In this case, a single index number for inequality may be insufficient.

The Gini index is not the only possible single index in widespread use, and the choice of index to use depends upon which aspect of inequality is of greatest importance. Theil's index of inequality can be calculated as in Eqn. (2) and has the attractive property that it can be decomposed, so that an analyst can say how much of total inequality is due to inequality within social groups and how much is due to inequality between groups. Also, because it is based on the average proportion of mean income received by individuals, it responds strongly to changes in the dollar incomes received by low income individuals.

$$R = \frac{1}{n}\sum_i \frac{Y_i}{\bar{Y}} \log \frac{Y_i}{\bar{Y}} \tag{2}$$

The Coefficient of Variation is another decomposable index of aggregate inequality, which is derived from the standard deviation of income, as in Eqns. (3) and (4). It also has the property of decomposability (unlike the Gini index, in general) but since it is based on the average squared deviation, it is most responsive to changes in income among the more affluent.

$$Var(Y) = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n} \tag{3}$$

$$CV = \frac{\sqrt{Var(Y)}}{\bar{Y}} \tag{4}$$

The literature on inequality (ably summarized in Silber 1999) has broadly agreed on a set of criteria that an acceptable index of inequality should satisfy. The principle of transfers argues that if a transfer of income is made from a richer person to a poorer person, an acceptable inequality index should always decline. The anonymity principle states that an index of inequality should depend only on the relative sizes of incomes and be invariant with respect to the identities of those occupying each position in the income hierarchy. The properties of scale invariance and population invariance require an index to be the same if all incomes are measured in different units or if identical populations are added together. These properties are not possessed by some possible measures (summarizing inequality by the 90:10 ratio of the top 10 percent's average income to that of the bottom 10 percent would, for example, not satisfy the principle of transfers, since transfers from richer to poorer in the middle part of the distribution would not affect the index).

In some cases, measurement of inequality is unambiguous—if the Lorenz curve for one country lies entirely inside that of another, all acceptable inequality indices will give the same ranking. However, as Atkinson (1970) observed many years ago, it is common in the real world to observe crossing of the Lorenz curve of two populations. In this case, when choosing an index of inequality the analyst is implicitly also choosing to emphasize inequality in a particular part of the distribution.

In some instances, it is possible to reduce inequality in several dimensions of life to a common denominator—for example, a good deal of work has been done in establishing the monetary value of access to different public services (such as education or health care). However, it is also reasonable to suppose that individuals differ in dimensions (e.g., income and life expectancy) that cannot easily be made equivalent. In this case, unless the dimensions of inequality are perfectly correlated, the assessment of aggregate inequality will depend on the relative weight to be assigned to each aspect of inequality and the correlation and shape of each variable's distribution.

4. Summary

If there were only two types of people, who differed only on one easily measured dimension, then inequality would be a fairly simple concept—the difference between 'the rich' and 'the poor.' In practice, however, there are multiple dimensions to inequality, and multiple differences between people. As soon, for example, as one adds a 'middle class,' inequality could be seen as either the differences between the rich and the poor, or the differences between the poor and the middle class or those between the middle class and the affluent. Key concepts like income, status, or power are also ambiguous and difficult to measure, in practice.

Furthermore, as Atkinson (1998), Brandolini (1998), and Osberg (2000) have demonstrated, although there have been clear trends in inequality in some countries in recent decades, there is no universal trend in all countries. Household income depends on household composition (i.e., on the processes determining formation, dissolution, and size of households); on the total market income of all household members from capital (which is influenced by inheritance patterns and taxation); as well as on individual incomes from labor (which depends on trends in both hourly wages and the distribution of work hours across employees, and on the correlation of male and female earnings within households); on the transfer incomes for which household members are eligible; and on the household's treatment by the tax system. Trends in the inequality of household income depend on changes in all these variables, some of which are heavily influenced by the public policies of governments, but some of which depend more on shifts in culture and society (both of which vary by country). National differences in culture and politics, and the vagaries of the evolution of such differences, thus have many channels of influence—and no segment of the income distribution is insulated from the impact of political decisions and public policy. It is not, therefore, really surprising to find a diversity of national experiences—as Brandolini (1998, p. 38) puts it: 'Neatly defined and unambiguous trends are unlikely to result from this multiplicity of factors.'

There is, therefore, a good deal of complexity in the concept of 'inequality.' This complexity does not imply either that measurement of inequality is impossible or that a measure can be found that will justify any hypothesis. However, recognition of the complexity of inequality measurement does imply that analysts should test explicitly whether, in a particular instance, all plausible measures of inequality agree. If there is ambiguity in inequality rankings, analysts should be prepared to be clear about which aspect of inequality is of greatest importance.

See also: Class: Social; Equality and Inequality: Legal Aspects; Inequality: Comparative Aspects; Mobility: Social; Social Demography: Role of Inequality and Social Differentiation; Social Inequality and Schooling; Social Inequality in History (Stratification and Classes); Social Stratification

Bibliography

Atkinson A B 1970 On the measurement of inequality. *Journal of Economic Theory* **2**: 244–63

Atkinson A B 1998 The distribution of income in industrialized countries. In: *Income Inequality: Issues and Policy Options*, Symposium Sponsored by the Federal Reserve Bank of Kansas City, Jackson Hole, WY, August 25–27

Brandolini A 1998 A bird's-eye view of long-run changes in income inequality. Mimeo, Research Department, Bank of Italy, August
Osberg L 2000 Long run trends in economic inequality in five countries—a birth cohort view. Working Paper No. 222, Luxembourg Income Study
Osberg L, Xu K 2000 International comparisons of poverty intensity: Index decomposition and bootstrap inference. *Journal of Human Resources* **35**(1): 51–81
Phipps S, Burton P 1995 Sharing within families: Implications for the measurement of poverty among individuals in Canada. *Canadian Journal of Economics* **28**(1): 177–204
Sen A K 1985 *Commodities and Capabilities*. North Holland, Amsterdam
Sen A K 1992 *Inequality Re-examined*. Russell Sage Foundation, Harvard University Press, New York
Silber J (ed.) 1999 *Handbook on Income Inequality Measurement*. Kluwer, Boston

L. Osberg

Inequality: Comparative Aspects

Inequality is defined here as sanctioned differential access to valued things, experiences, and conditions in human societies. Comparative aspects are similarities and differences found among systems of inequality as a result of comparing societies in all of their variety: through time (developmentally) and space (cross-culturally), in societies small and large, homogeneous and complex, prehistoric, historic and contemporary, throughout the world.

1. *Social Inequality*

Social inequality has been debated from the time of Aristotle and Plato to Rousseau, Karl Marx, Max Weber and innumerable contemporary authors (Beteille 1987). The classic triumvirate, 'class' (economic order), 'status' (honor) and 'party' (power) are widely cited as its essential dimensions, with power—the ability to prevail over others—as the virtual independent variable. Powerful people are privileged; powerless people are rarely so, and then only under sponsorship of the powerful. Their perspectives on inequality reflect their places in the system: those who look from below see and experience a different world than those whose gaze and experience is from above, and the latter are the ones in a position to define and explain the system, as well as enforce it.

This entry draws on understandings of social inequality obtained, primarily by anthropologists, through cross-cultural and cross-temporal research focused on peoples who have been otherwise largely ignored: those commonly labeled, if noticed at all, as primitive, aboriginal, tribal, native, indigenous, heathen, nomadic, peasants, untouchables, colonials, minorities, ethnics, immigrants, refugees—in short, those rendered vulnerable for whatever reason, or unreason.

2. *Egalitarian (Unranked) Societies*

Dominance and differential evaluation exist in every society, but in some—those often loosely described as 'egalitarian'—are limited to elemental roles found in all societies: those of family, sex/gender, age/seniority, specific activities or skills, and cherished or stigmatized personal characteristics. If additional inequalities occur they are more symbolic than material or political. Inequalities in such societies have received increasing attention from ethnographers and archaeologists in recent years (Flanagan 1989).

These are small, mobile bands of kinsmen with households as the basic social and economic units. Theirs are subsistence economies derived by foraging wild species, without privileged or curtailed access to resources or the means to obtain them. They share their economic pursuits and rewards, without expectation of specific returns, in what has been termed 'generalized reciprocity' (Sahlins 1968). Requisite knowledge and skills are present in every household, extra-familial authority or occupational roles are rare or absent, and sexual division of labor is complementary with little or no differential evaluation.

2.1 *Assertively Egalitarian Societies*

Even in 'egalitarian' societies, equality of power, possessions and prestige is in jeopardy. The *Hadza* of Tanzania, are foragers who consume the products of a successful foray before embarking on another. This, Woodburn (1982) describes as an 'immediate return system.' He characterizes societies with such economies as 'assertively egalitarian' in that their members characteristically resist and sanction attempts to exert authority or consume resources without sharing.

2.2 *Relatively Egalitarian Societies*

By contrast, foraging economies in which long-term collective investments of resources are employed (e.g. construction and maintenance of corrals for game drives), Woodburn terms 'delayed-return systems': the return on investment of labor is delayed rather than immediate. Although they are egalitarian to an extent found among few non-foragers, and are therefore 'relatively egalitarian,' they entail authority and accumulation untenable in immediate-return societies as a result of the need for coordination of delayed-return

activities and preservation of products for use during periods of delay.

All societies whose economies and organization are more complex than those of 'assertively egalitarian' foragers have delayed-return systems. Those with economies and social systems more complex than those of 'relatively egalitarian' foragers, are inegalitarian or 'rank societies' (Fried 1967).

3. *Inegalitarian (Rank) Societies*

Inegalitarian societies are those wherein members are individually, or more often collectively, hierarchically ordered in access to valued things, conditions and experiences. All societies discussed below are in this sense inegalitarian. This presentation here moves from those of relatively 'small scale' (small in numbers and relatively homogenous), to increasingly 'large scale' (large in population and socially complex). This sequence reflects and extends the strong correlation which exists between scale and inequality.

3.1 *Affluent Foragers and Farmers: Transition to Assertive Inequality*

Some cultural anthropologists and many archaeologists have turned to supplementing largely static extant typologies, with analyses of transitions, especially those from unranked to, and within, rank societies (Price and Gebauer 1995, Kirch and Marshall 1992).

Rank society emerges when circumstances permit some members to acquire privileged or, in the extreme, exclusive access to essential resources. This has traditionally been attributed to the invention of domestication—primarily of plant species, hence agriculture—and the resulting opportunity for some to control land and command the labor of landless others to till and guard it. Similar conditions obtain for pastoralists' control of their herds.

It has long been recognized that some foraging societies share with early farmers the inequalities characteristic of agricultural and pastoral peoples (Koyama and Thomas 1981). The most widely cited example is the peoples of the US northwest coast, who developed maritime economies devoid of domesticated species, combined with complex and highly inegalitarian social structure (Lightfoot 1992). This seeming anomaly is explained by the annual runs of salmon which, accompanied by technologies of preservation and storage, provides a food supply as predictable and productive as any domestic crop or herd. Inequality results from opportunities for some to control the best fishing sites and equipment, and to mobilize personnel to exploit them. This system entails ranked kin groups (clans) and their chiefs. The latter accumulate goods provided by their clansmen for subsequent ceremonial, status-enhancing distributions to rival chiefs and clans—the storied potlatch.

Other examples of such systems are found primarily in horticultural societies, which employ two distinct kinds of leadership role and social organization: big man and chief (Sahlins 1968).

3.2 *Big Men, Big Women*

The simpler of the two structures is 'big man' (rarely 'big woman') society. The big man's role is an informal, acquired, or achieved status. Through personal charisma and persuasion he recruits a volunteer following whose number contribute valued goods to a fund which he will ceremonially distribute to a competing big man and his followers (as in a potlatch). Recipients thereby incur a debt to be paid off, and if possible exceeded, in a reciprocal distribution. The magnitude of the distribution is the measure of a big man's stature, while his followers receive credit (and when next he is on the receiving end, goods) proportional to their contributions.

3.3 *Chiefs*

A chief differs from a big man in that his is an office, ascribed to him as a result of his paramount position in his clan. Fellow clansmen are his obligatory supporters and he their hereditary headman. As chief, he must behave in a chiefly manner, ever responsible and generous to his clan. His behavior follows from his role; the opposite of the big man, whose role follows from his behavior.

Big man organization may evolve into chieftainship when big men find that kinsmen are more reliable supporters than others and therefore preferentially, and eventually exclusively, recruit them. As kinship becomes the *de facto* principle of organization, the followers of a charismatic big man may become in effect, and ultimately in fact, clansmen of their paramount kinsman, and he their chief.

3.4 *Chiefs to Kings, Kinship to Class: Transition to Complex Society*

The offices of both chiefs and kings are ascribed on the basis of kinship with their predecessors. But they differ fundamentally: the relevance of kinship for the king stops with his coronation; for the chief, investiture is only the beginning. The king becomes sovereign over all within his domain, beholden to no one; the chief is paramount only among his kinsmen, and as kinsman

is beholden to them as they are to him. The king rules, accumulates, and consumes; the chief proposes, coordinates, accumulates, and redistributes.

Kingship thus represents the demise of kinship as the basis for social structure, replaced by economic stratification: ownership of productive resources including land, territory as the dimension of governance, authority and force as its method, and depersonalization as a result of dramatic increases in scale. There appear for the first time, ranked economic classes and cultural categories.

As reciprocity is the economic hallmark of unranked society, and accumulation for conspicuous redistribution is the hallmark of big-man and chiefly organization, accumulation for conspicuous consumption is a hallmark of kingship and the state.

4. *Stratification*

Contemporary nation states are socioeconomically stratified in a variety of complex ways, their populations distributed along continuums of differential access to necessities and rewards. These distributions are often perceived as comprising distinctive, ranked categories—social strata—of people similarly positioned.

An issue to which comparative analysis has made major contributions, is the nature and function of stratification: does it primarily serve to select the best qualified for the most important roles in society by offering them the highest rewards, or does it serve mainly to withhold opportunities from vulnerable segments of society, thus assuring that there will be people with no choice but to perform cheaply and tractably the necessary tasks shunned by others? Caste in India, and US race/ethnic relations, are examples which support the latter interpretation (Berreman 1960), as is the status of Nepalese and Brazilian immigrant labor in Japan (Yamanaka 2000).

4.1 *Dimensions of Stratification*

Power is key to stratification, but class and status are closely linked to it; the three vary together. Stratification systems differ in significant ways, the analysis of which is facilitated by comparison along three axes of inequality. One defines the nature of the ranking, contrasting status strata, ('honor and privilege') at one pole, and class strata (the 'economic order'), at the other. A second axis describes the locus of characteristics which determine an individual's status, contrasting those regarded as intrinsic to individuals (ascribed), with those regarded as extrinsic (acquired or achieved). A third addresses the composition of strata: whether they comprise named, organized, acting groups, or analytically discernible but unorganized categories. Variations along these axes, together with the cultural explanations that rationalize each system, constitute the significant differences among them.

4.2 *Acquired Economic Status: Categories as 'Classes,' Groups as 'Sodalities'*

Acquired status is based on possessions and privileges, its inequalities defined by economic factors (income, net worth) and their accompanying lifestyles (patterns of consumption, education, occupation). These are classes, of which it is useful to contrast: (a) those comprising objectively discernible categories, useful to an analytic observer, but amorphous, unorganized and not necessarily recognized even by those who comprise them; and (b) those which comprise groups, publicly recognized, members aware of their shared situation and its sources, organized, acting in their common interests. Marx recognized this distinction when he described the first as 'groups in themselves,' and the second as 'groups for themselves.'

Ranked economic *categories* are here designated simply classes. Ranked, economically based *groups*, such as labor unions and professional associations, are designated associations or sodalities.

4.3 *Feudalism and Estates: Groups with Ascribed Economic Status*

Feudal systems are complex agricultural societies comprising hierarchical relations among landed nobility, various levels of commoners and rural peasants. Because patron–client relationships with respect to land, are a defining characteristic, such systems fall well within the 'economic order.' They therefore constitute systems of class stratification (Stavenhagen 1975). But these are classes with a difference, for they are explicitly ranked, with specified rights and obligations. So long as a person belongs to a given category, he or she shares its rank. However, mechanisms exist—usually economic—by which an individual may, with difficulty, move to another category thereby acquiring its rank. Status is thus intrinsic to the category but not to the individual. Such named, bounded and jurally defined categories can be labeled estates, to distinguish them from classes.

4.4 *Ascribed Honor: Caste, 'Race,' and Ethnicity*

Ascribed statuses of honor are distinguished by the fact that rank is assigned on the basis of membership in differentially privileged and esteemed groups or categories rather than on the basis of activities or possessions. They are also distinguished by the fact that, as a result of ascription, those who share them constitute publicly recognized, self-acknowledged

groups. Often they are culturally distinct, cohesive—sometimes corporate—groups *for* themselves. Weber called them status groups. Four types will be contrasted below: 'caste/race,' 'ethnicity,' 'sex/gender,' and age/seniority.

4.4.1 Caste/race. Status assigned to culturally distinct, socially segregated categories of people on the basis of shared and differentially evaluated characteristics claimed to be innate, is unalterable. 'Race,' as it is socially constructed in the US and 'caste' (*jati*), as it is constructed in India, are classic examples. Castes are localized, internally organized, endogamous, and often corporate, groups ranked according to conceptions of ritual purity. Most low ranking castes lay claim to higher status than that accorded them—claims that are routinely ignored or brutally quashed by high castes. Attempts to validate their claims entail status emulation: adoption of high caste rituals and behaviors, a process familiar in Indian social science as *Sanskritization* (Srinivas 1966). They are exceedingly rarely successful, and then only if backed by unprecedented power.

Stigmatized 'races' in the US differ from low castes of India in that they are nationwide social categories rather than localized groups, but they share with low castes their birth-ascribed subjection to disparagement, discrimination, and exploitation, rationalized by accusations of intrinsic worthlessness. Similarly disparaged 'outcaste' groups have included the *Burakumin* of Japan and *Hutu* of Rwanda and Burundi. All respond to their oppression with resentment and resistance. The only consensus on inequality in these systems is the pragmatic one of who has the power, under what circumstances and with what effect it will be used. When this consensus is breached, the systems are often vigorously challenged. These fundamental similarities are here deemed to be sufficiently significant to justify the combined term 'caste/race,' except where their differences are at issue.

4.4.2 Ethnicity. Ethnicity, or ethnic group, are terms for ascribed identity bestowed on groups or categories defined by their distinctive cultural attributes such as language and religion. They are frequently, but not necessarily, ranked within their societies. If ranked, their status is not wholly ascriptive for, while the ethnic group's ranking is ascribed, the defining attributes of its individual members are acquired through shared socialization—learning—and what is learned can be rejected or forgotten. Therefore, as in the case of estates, rank is intrinsic to the group but extrinsic to individuals, so that individual mobility within the system is possible. If circumstances permit, people can acquire or construct other ethnicities, other statuses (Barth 1969).

4.4.3 Sex/gender. Sex is a usually birth ascribed and immutable category of difference, which in all societies serves as a basis for division of labor, and in most complex societies is a basis for inequalities. Gender is the social construction of its nature and significance. Many societies define other genders as well, among them the *berdache* 'men–women' of several American Indian societies and the 'manly hearted women' of some of them, Polynesian *fa'a fahini* ('in the manner of women'), and India's 'third sex,' the *hijara*. These, and gay, lesbian, bisexual, and trans-sexual genders, of Euro–US and other complex societies, are but some of the diverse gender roles worldwide. Some are stigmatized, but as many are not and occupy conventional or exceptional roles in their societies.

The relative status of the sexes in human societies has been long debated. Universal female subordination is often taken for granted. To challenge this one has only to look across the full range of human societies, including especially those which resemble the ones in which 99.9 percent of our ancestors, through time, have lived (Leacock 1978). Among more complex and inegalitarian societies, one must consider those whose systems of descent, inheritance and post-marital residence follow the female line. The highest incidence of such systems are found in Southeast Asia and Melanesia, among horticultural peoples in which women own the land, do the lion's share of cultivation of staple crops, and enjoy a high-to-superior degree of autonomy, authority, and respect.

Gender status correlates positively not only with the prominence of its role in descent, residence, and horticulture, but its contribution to the household or kin group economy, whether foraging, horticulture, agriculture, pastoralism, wage labor, or professions. In complex societies it correlates most highly with the economic contribution made from outside of the household or 'domestic sphere' (Caulfield 1981).

In discussing gender inequalities it must not be overlooked that, while every society has distinctive female and male cultures, the genders are largely socialized in the intimacy of family and kin of both sexes, sharing class, ethnic, caste/race statuses, and environments. They are therefore familiar with one another to an extent that people of other ranked identities are not. However, the extent to which such experiences contribute to respect and empathy cannot be predicted from the fact of contact. As studies of US 'race' relations have demonstrated, contact does not necessarily reduce prejudice, and may exacerbate it by seeming to confirm stereotypes, unless it occurs in status-equal circumstances.

US women have long been discriminated against, while sharing class and ethnic status with male family members. In India, women share the ritual and economic status of their male caste fellows, while being subordinated to males. Circumscription of women in activities outside the household is an

important symbol of high caste rank. *Sanskritization*, caste status striving, entails increasing subordination of women (Berreman 1993).

4.4.4 Age/seniority. Age categories and cultures, like those of gender, are ascribed, socially constructed, familiar to all, and highly variable. As a basis for inequality, age and seniority are at their minimum in unranked societies. Societies organized in lineal kin groups or clans often place great importance on seniority, as the first born in a senior (most often male) lineage becomes the paramount.

A number of societies, the most thoroughly reported of which are East African cattle herders, are organized into ranked, age-based groups, more often male than female, whose members share distinctive ritual, economic, military, or other privileges and responsibilities.

Comparative studies of age in other kinds of societies include Mead's classic *Coming of Age in Samoa* (1928), which challenged the universality of adolescence as a period of trauma, and Henry's *Personality and Ageing* (1973), which argued that individuals are regarded as persons to the degree that they possess symbols of attachment to their social system, and that in the US such symbols (gainful employment, license to drive) are often withdrawn from the elderly leading to the experience of diminished personhood, resulting in social suffering, hostility and depression. This is precisely the kind and consequence of denial of full humanity inflicted by racism, casteism, ethnic and gender discrimination.

5. *Colonialism: International Hegemony*

Relationships between colonizer and colonized are among the most powerful of the world's structures of domination and exploitation (Asad 1973). In its classic context, colonialism has referred to nations and peoples annexed for their natural, manufactured or human resources by one of the 'great,' usually European, maritime, powers. Subsequently the concept has been directed to other contexts, as in 'third world colonialism' (hegemony over developing nations, often by their erstwhile colonizers), and 'internal colonialism' (domination of minorities within developed nations, e.g. African Americans, Palestinians in Israel).

By analogy, 'academic colonialism' refers to the dominant role well-funded researchers from 'advanced' nations play in 'developing' nations, while local scholars are scantily funded if at all. It is most damaging and resented when indigenous scholars are evaluated by the standards and fashions of colonial countries, ratified by foreign degrees, and by publication in foreign journals, in colonial languages. Social and behavioral scientists, perhaps more often than others, may find themselves strangers in their own academies and virtual pariahs in world scholarship (Saberwal 1967).

6. *Full Circle: Modern Egalitarian Societies?*

Egalitarian social systems require egalitarian economies. Both have been on the agenda of utopian communities and ideal societies studied by social and behavioral scientists: Soviet and Chinese communes, Israeli Kibbitzum, socialist societies, and anarchist communities (Barclay 1982), American 'counter-cultural' communities of the1960s–1970s and various egalitarian cults. Many live in small, self-contained communities, combining interdependence and individual/familial autonomy, bound by ties of kinship (usually fictive), sharing ideology and communal activities, while avoiding in so far as possible participation in the complex society, economy, and politics of the world system. They often pursue subsistence economies through cooperation, sharing and reciprocity, their production for use rather than profit, abjuring private property, accumulation of wealth, hierarchies of authority or privilege, and status symbols. That is, they reproduce the social and economic practices of small-scale, 'assertively' or 'relatively' egalitarian societies, with which human societies began and for so very long survived. Prehistoric, recent, and current exemplars have virtually all succumbed to appetites, blandishments, and requirements of stratified, state organized, producer–consumer society.

Bodley closed *Victims of Progress* (1975) with the observation that small-scale, egalitarian culture 'is gone—after a very long and successful tenure,' and it seems that the presently dominant, stratified 'culture of consumption … too, will disappear … but after a very brief and preposterous career.' He was probably right, but hope, courage, and determination do spring eternal and could prevail.

See also: Discrimination; Discrimination, Economics of; Discrimination: Racial; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Inequality; Social Inequality and Schooling; Social Inequality in History (Stratification and Classes); Social Stratification

Bibliography

Asad T (ed.) 1973 *Anthropology and the Colonial Encounter*. Ithaca Press, London

Barclay H 1982 *People without Government: An Anthropology of Anarchy*. Kahn and Averill, London

Barth F 1969 Introduction. In: Barth F (ed.) *Ethnic Groups and Boundaries, The Social Organization of Culture Difference*. Little Brown, Boston, pp. 9–38

Berreman G D 1960 Caste in India and the United States. *American Journal of Sociology* **66**: 120–27
Berreman G D 1993 Sanskritization as female oppression in India. In: Miller B D (ed.) *Sex and Gender Hierarchies*. Cambridge University Press, Cambridge, UK, pp. 366–92
Beteille A 1987 *The Idea of Natural Inequality and Other Essays*. Oxford University Press, New Delhi, India
Bodley J H 1975 *Victims of Progress*, 1st edn. Mayfield, Mountain View, CA, p. 174
Caulfield M D 1981 Equality, sex and mode of production. In: Berreman G D (ed.) *Social Inequality: Comparative and Developmental Approaches*. Academic Press, New York, pp. 201–19
Flanagan J G 1989 Hierarchy in simple egalitarian societies. *Annual Review of Anthropology* **18**: 245–66
Fried M 1967 *The Evolution of Political Society*. Columbia University Press, New York
Henry J 1973 Personality and ageing with special reference to hospitals for the aged poor. In: *On Sham, Vulnerability and Other Forms of Self-Destruction*. Vintage Books, New York, pp. 16–37
Kirch P V, Marshall S 1992 *Anahulu: The Anthropology of History in the Kingdom of Hawaii*. University of Chicago Press, Chicago
Koyama S, Thomas D H (eds.) 1981 *Affluent Foragers (Senri Ethnological Studies 9)*. National Museum of Ethnology, Osaka, Japan
Leacock E 1978 Women's status in egalitarian society, implications for social evolution. *Current Anthropology* **19**: 247–75
Lightfoot K 1993 Long-term developments in hunter–gatherer societies: Recent perspectives from the coast of North America. *Journal of Archaeological Research* **1**: 167–201
Mead M 1928 *Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilization*. Morrow, New York
Price T D, Gebauer A B 1995 *Last Hunters, First Farmers: New Perspectives on the Prehistoric Transition to Agriculture*. School of American Research, Santa Fe, NM
Saberwal S et al. 1967 Academic colonialism: A symposium on the influences which destroy intellectual independence. *Seminar* **112**: 9–40
Sahlins M 1968 *Tribesmen*. Prentice-Hall, Englewood Cliffs, NJ
Srinivas M N 1966 *Social Change in Modern India*. University of California Press, Berkeley, CA
Stavenhagen R 1975 *Social Classes in Agrarian Societies*. Anchor Books, Garden City, NY
Woodburn J 1982 Egalitarian societies. *Man* **17**: 431–51
Yamanaka K 2000 Nepalese labour migration to Japan: from global warriors to global workers. *Ethnic and Racial Studies* **23**: 62–93

G. D. Berreman

Infancy and Childhood: Emotional Development

Emotion is a familiar but psychologically complex human experience. Consequently, the growth of emotion integrates diverse developmental processes in infancy and childhood, including psychobiological growth, perceptual processes, emotional and social understanding, self-awareness and self-understanding, and the growth of self-regulation. The study of emotional development has also been stimulated by basic questions concerning the nature of emotion itself, and its role in behavior and development. This article explores the issues that have contributed to considerable research interest in emotional growth early in life.

1. *The Nature of Emotional Development*

Emotions color the life experience of people of all ages. For infants and children, emotions have especially powerful motivational consequences because emotional understanding is rudimentary, capacities for emotion self-regulation are limited, and the young child's engagement or withdrawal from events is primarily emotional in quality. Emotional development is central to understanding the growth of social competence, intellectual achievement, self-understanding, and the earliest origins of mental health or psychopathology because of the importance of emotion to the organization and regulation of early behavior (Lewis and Haviland-Jones 2000). Emotional development provides a window into the psychological growth of the child as, from infancy to late childhood, emotions become integrated into the broader fabric of behavioral competence, become objects of thought and analysis, can be strategically expressed or concealed in social situations, are shaped by the child's experience of significant relationships, motivate prosocial as well as aggressive tendencies, become part of self-understanding (especially in feelings of guilt and pride), and color the individuality of personality growth.

1.1 *Emotions and Development*

Emotional development was vigorously studied earlier in the twentieth century, and has recently received renewed attention after a period of dormant interest. There are several ways that the contemporary study of emotional development differs from traditional approaches rooted in psychoanalytic or Piagetian theories. First, emotions are regarded as centrally important to understanding development in infancy and childhood, rather than as a peripheral or disruptive facet of early growth. Second, emotional development offers a uniquely integrative perspective on human development. Understanding emotional growth requires synthesizing biological perspectives on development (including neurobiological growth, temperamental individuality, and neuroendocrine functioning) with social perspectives (including relational and family influences, and the child's developing understanding of cultural norms). Third, emotions are

viewed as having constructive as well as disorganizing features. Although emotions are capable of undermining effective functioning (which accounts for current scientific interest in the growth of emotion regulation), developmental researchers also emphasize how emotions motivate and organize social interaction, underlie the quality of early social relationships, shape self-understanding, and provide incentives for developing new skills and capabilities. This is consistent with a new appreciation for the constructive, organizational role of emotions in other areas of psychological study.

Contemporary study of emotional development in infancy and childhood is also motivated by clinical concerns—most notably, a growing realization of the early emergence of emotionally related disorders in young children, including depression, anxiety disorders, and behavioral and conduct problems. Thus the study of emotional development is informed by developmental psychopathological studies of the origins of these disorders in troubled parent–child relationships, temperamental vulnerability, environmental and genetic risk, and individual capacities for emotional self-regulation. This research further underscores the complex, integrative quality of emotional growth, and the various risks that exist for emotional dysfunction as well as alternative avenues to prevention and remediation.

Developmental research on emotions has required a broad range of methodologies. These include fine-grained analysis of videotaped facial expressions, naturalistic observations of social interaction at home or school, conversations with young children about their comprehension of emotional events, detailed study of the responses of children to emotionally evocative laboratory vignettes, parent-report approaches, and psychobiological measures of heart rate, cortisol, or cerebral asymmetry during emotionally arousing situations. Developmental scientists are increasingly integrating these methodological strategies in broadly based longitudinal studies that enable them to study the emergence of emotional capacities over time and in the context of family processes, psychobiological maturation, and other influences.

1.2 What is Emotion?

Contemporary study of the functions of emotion in behavior and development is accompanied by debate over the nature of emotion itself. Two broad perspectives can be distinguished. The structuralist perspective defines emotion in terms of discrete emotional states, each with unique patterns of subjective feeling, cognitive appraisal, physiological arousal, and facial expression (Izard 1991). These emotion components are tightly interconnected and are deeply rooted in human evolution because of their relevance to reproductive success (e.g., feelings of anger involve visceral, perceptual, subjective, and expressive components relevant to self-defense). This view is consistent with the everyday tendency to distinguish basic emotions like anger, fear, sadness, and joy as readily dissociable states.

More recently, however, functionalist perspectives on emotional development have questioned whether emotional life is so discretely packaged and have offered an alternative view (Saarni et al. 1998). In this perspective, emotion is defined in terms of the quality of the transactions between the person and the environment on matters of significance to the individual. In this regard, what defines an emotion is not a network of internal subjective, physiological, and other components, but rather the constellation of a person's goals, appraisals, action tendencies, and other capacities in relation to environmental inventives, obstacles, and opportunities. As a result, emotional experience is not confined to a predefined set of emotion categories, but can encompass the rich variety of subtly nuanced emotional blends characteristic of everyday experience, as well as the breadth of how emotion is conceptualized in diverse sociocultural groups based on how specific person–environment transactions are conceived.

A functionalist approach is also well suited to developmental analysis. One way that emotions develop, for example, is in how the generalized arousal conditions of early infancy mature into the more specific, organized, and refined emotional states apparent in older children and adults. Young infants do not necessarily experience or express emotions in ways that fit into the discrete emotion categories of English speakers, but appear instead to experience more diffuse affective systems that become progressively better differentiated as a result of early social experiences and socialization, cognitive growth, and psychobiological maturation. Studying this developmental process requires methodological creativity and breadth, as functionalist theorists suggest. Rather than relying exclusively on facial expressions as primary indicators of underlying subjectivity, for example, developmental researchers increasingly index emotion multimodally, integrating facial, vocal, and other behavioral indicators into an organizational interpretation in relation to the child's interaction with environmental opportunities and challenges.

2. What is Emotional Development the Development of?

The primary challenge of developmental analysis is understanding why developmental changes in emotion occur, and how emotional growth is associated with concurrent advances in social, cognitive, linguistic, and physical development. Exploring the various components of emotional development reveals how

emotional experience integrates significant aspects of psychological development in infancy and childhood.

2.1 Psychobiological Growth

Emotions are biologically basic features of human functioning. They are predicated on richly interconnected brain structures and hormonal influences that organize the interplay of activation and regulation that constitutes emotional reactions. The young infant's capacity to exhibit distress, fear, surprise, happiness, and rage reflects the early functioning of these biologically deep-rooted psychobiological systems. Emotional development is based, therefore, on neurophysiological, neuroendocrine, and other biological processes that change rapidly in infancy and childhood (LeDoux 1996, Panksepp 1998). For example, the developing stability of sympathetic and parasympathetic nervous system functioning in infancy helps to account for the growing predictability, manageability, and environmental responsiveness of the infant's emotional responding. Later in childhood, as higher brain regions (especially in the frontal neocortex) progressively mature and become integrated with early developing regions of the limbic system and other subcortical structures, emotional life becomes increasingly characterized by more acute emotion appraisals, complex emotional blends, and growing capacities for emotional self-regulation. Because an apparently simple emotional reaction is actually a surprisingly complex psychobiological event, developmental changes in multiple brain regions, hormonal functioning, and neurotransmitter regulation are involved in emotional growth.

2.2 Perceiving and Responding to Emotions in Others

Throughout life, emotions are evoked by one's perception of emotion in others. Infants are capable of discerning the emotional meaning in an adult's facial and vocal expressions before the end of the first year, and there is evidence that they respond resonantly to the sound of another infant's crying even earlier. Moreover, by the end of the first year, infants not only attend to the emotional expressions of adults but also incorporate the meaning of these expressions into their understanding of the adult's behavior. This phenomenon is called 'social referencing,' and it is analogous to how people enlist the emotional expressions of others into their assessment of their circumstances (Feinman 1992, Saarni et al. 1998). In a comparable manner, when a one year old encounters an unfamiliar person or object, the sight of a trusted adult's reassuring smile or terrified look (especially if it is accompanied by appropriate vocalizations and other behavior) significantly influences the child's approach or withdrawal tendencies. It appears that the infant 'reads' appropriately the meaning of the adult's emotional expression and its referential relevance to the unfamiliar event.

In the years that follow, advances in emotional understanding significantly expand the young child's capacity to respond appropriately and insightfully to the perception of emotion in others. By the end of the second year, for example, toddlers are more consciously aware of the subjectivity of emotional experience (i.e., another person can feel differently than oneself) and of the causes for others' emotions. This is often manifested when young children are observed comforting a distressed adult or peer, teasing a sibling, during pretend play, or even in humor (Denham 1998).

In early childhood, moreover, genuinely empathic responses to another's distress can be observed. Initially, young children respond with concerned attention to the sight and sound of another's distress, but empathic responses increase in sophistication and scope in the years following. Early empathy is not reliably accompanied by prosocial initiatives toward the distressed person because empathic arousal is a motivationally complex experience for young children (Zahn-Waxler and Radke-Yarrow 1990). Young preschoolers have been observed to laugh, tease, become distressed, and even hit a distressed parent, for example, as well as offering comfort. With increasing age and greater understanding of their own efficacy as prosocial agents, empathy becomes somewhat more regularly associated with helping others.

2.3 Emotional Understanding

The advances in emotional understanding accompanying conceptual growth enable children to better understand others' emotions, and also their own. Consistent with growth in their naive 'theory of mind,' for example, two to three year olds appreciate how emotion is associated with the satisfaction or frustration of desires (which vary for different people), and four to five year olds understand the more complex associations between emotion and thoughts, beliefs, and expectations (Bartsch and Wellman 1995). The latter is important because it contributes to the growing realization that emotions can be based on false beliefs or appearances (a peer anticipating a delicious dessert that is actually missing from her lunch bag), and that outward emotional appearance does not necessarily reveal underlying emotional reality.

In middle childhood, children begin to understand the dynamics of emotional experience, such as how emotional intensity gradually dissipates over time, how specific emotions are related to specific antecedent causes, and how a person's unique background, experiences, and personality can yield unique emotional reactions. By late childhood, children begin to

appreciate how multiple emotions can be simultaneously evoked by a single event, such as feeling happy and afraid when performing before a group (Saarni et al. 1998). Despite these conceptual advances, important features of emotional understanding do not emerge significantly until adolescence, such as an awareness of the psychological causes of emotion in relational experience, the experience of ambivalence or emotional conflict, and a grasp of complex emotional states (e.g., poignancy).

2.4 Emotion, Self-Awareness, and Self-Understanding

Emotional development is complexly tied to the growth of self-awareness and self-understanding in infancy and childhood. Prior to a dawning awareness of the self as a psychological entity, for example, a young child's emotional repertoire does not include self-referent emotions such as pride, guilt, shame, and embarrassment. Late in the second and during the third years of life, however, as young preschoolers reveal increasing psychological self-awareness (e.g., calling self by name, insisting on 'doing it myself'), these self-referent emotions begin to appear (Tangney and Fischer 1995). As young children begin to see the self as an object of analysis by others, and also by oneself, the simple joy of success becomes pride in achievement, and in response to conspicuous attention young children increasingly respond with gaze aversion, smiling, and other signs of embarrassment.

Emotional reactions to the self are significantly shaped by the reactions of others to the child. The development of self-referent emotions is advanced not only by growing psychological self-awareness but also by the young child's increasing understanding of behavioral standards and the ability to apply those standards to an evaluation of personal behavior (Kochanska and Thompson 1997). Guilt derives, for example, from an awareness of how personal behavior has departed from an acceptable standard. Parental reactions are crucial to the young child's dawning awareness and application of these standards of conduct, and indeed the earliest expressions of self-referent emotions commonly occur in contexts where young children can explicitly perceive, or anticipate, parental reactions to their accomplishments or misconduct. With increasing age, children appropriate parental standards that become the child's own internalized expectations for conduct. Thus the manner in which parents convey, reinforce, and attach emotional and relational significance to a young child's achievements or misbehavior is an important influence on the child's emotional life, and developing self-esteem and self-concept.

This underscores the complexity of the association between self-understanding and the growth of self-referent emotions. A dawning psychological self-awareness is necessary before self-referent emotions become part of a young child's emotional repertoire. Once these emotions are possible, however, the circumstances in which they are elicited and their relational consequences are significant for further growth in self-concept and self-understanding.

2.5 Emotion and Emotional Displays

One consequence of young children's developing understanding of the difference between belief and reality is the growing realization that emotional expressions may not accurately reveal underlying emotional reality. This awareness emerges at the same time that children discover the social expectations governing emotional expressions. A person is supposed to show delight when opening a gift in the presence of the gift-giver, for example, even if the gift is undesirable. This and other emotional 'display rules' governing the expression of emotions in social situations proscribe how emotions are used to avoid hurting others' feelings, protect self-esteem, and preserve relationships.

Young children do not understand display rules, but they can be observed using them as young as age four. Not until middle childhood do children grasp the nature and purposes of emotional display rules, but once they do so they can enlist them in their interactions with family members, peers, and others (Saarni 1999). As they do so, children also begin to appreciate the privacy of emotional experience: emotions need not necessarily be conveyed to others if one can manage emotional experience and its expression.

2.6 The Development of Emotion Regulation

One of the important differences between infants and adolescents is in their capacities for emotion regulation. Emotion regulation is distinct from emotional display rules because although the latter regulate the *expression* of emotion, the former manages emotional *experience*. People have many reasons for managing their emotions: to feel better under stress, elicit support from others, act courageously, resist bullying, enhance motivation, think better, and for many other reasons. Although the goals underlying emotion regulatory efforts are easily understood in very young children, these goals become increasingly complex and nuanced in the more complex social situations of older children. Competent emotion regulation may be observed, for example, when a child loudly protests a peer's provocations when adults are nearby, but quietly tolerates the bully when adults are absent. Moreover, in conditions of risk for developmental psychopathology, the goals underlying a child's emotion regulatory efforts may be complex, mutually inconsistent, and thus, in some senses, self-defeating (Thompson 1994).

In infancy, emotions are primarily regulated by others, especially parents. Parents manage a baby's

emotions by directly intervening to soothe the child, and also by regulating the emotional demands of familiar settings like home or child care (often in accord with their perceptions of the child's temperament), altering how the child construes an emotional arousing situation (such as by smiling reassuringly when a friendly, but unfamiliar, adult approaches the child), and later by actively coaching young children on expectations or strategies of emotion management. More broadly, the security that young children derive from their emotional attachments to caregivers is an important resource for emotion regulation because of the confidence it inspires that events will not be uncontrollable and that the adult will intervene helpfully when necessary (Cassidy 1994). Emotion regulation is fostered by others, in other words, not only by their immediate interventions but also by the reassurance provided by the child's relationship with a caring adult.

From an early age, however, young children are also developing rudimentary skills for managing their own emotional experience (Thompson 1994). This can be observed initially in the comfort-seeking of a distressed infant or toddler, but skills of emotion self-regulation expand considerably in early childhood. Young children can be observed making active efforts to avoid or ignore emotionally arousing events, changing or substituting goals that have been frustrated, redirecting activity or attention, using distraction, engaging in encouraging or reassuring self-talk, obtaining further information about the situation, or enlisting other simple behavioral strategies. In the years that follow, moreover, older children become increasingly proficient at enlisting psychologically more sophisticated strategies of emotion regulation, such as using emotion-blunting ideation, redefining the situation (or attributions for behavior), or acting in a manner inconsistent with felt emotion. The growth of strategies for emotion self-regulation occurs in concert with the development of emotion understanding and children's increasing psychological self-awareness, and contributes to developing social competence with peers and adults.

3. *Relational Influences on Emotional Development*

Emotions are not just the outward expression of an internal experience, but are deeply influenced by the social contexts in which they occur. Emotional development illustrates how early in life, intrinsically human and biologically fundamental response processes take shape in a social context.

Social influences on emotional development are especially apparent in children's close relationships with caregivers and other adults. Many things happen in their relationships with adults that guide the course of children's emotional growth (Thompson 1998). Early in life, the responsiveness of caregivers to a baby's emotions influences the frequency, persistence, and intensity of the infant's emotional signaling, as well as the baby's developing confidence in the adult's helpful assistance. The adult's emotional expressions influence the child's vicarious emotional and behavioral responses to events, in a phenomenon earlier described as social referencing. Parental behavior is often also a catalyst for empathic responding, especially when parents fix the attention of offspring on another's distress and carefully explain the causes of that person's feelings. As earlier noted, the behavioral standards that parents convey in their responses to good behavior and misbehavior shape the earliest experiences of pride, guilt, shame, and embarrassment, as well as developing self-concept and self-esteem. More broadly, the simple conversations shared by parents and their young offspring about the day's events help to color and socialize emotional understanding by how the parent describes and explains emotional experiences, attributes motives and emotions to people, and interprets the emotional events they have shared (Dunn 1994). Among the many things children learn from these conversations are the sociocultural rules governing emotional expressions in social situations. Finally, adults in close relationships are influential as they directly manage the emotional arousal of young children, and as they coach (and model) the child's use of personal strategies for emotion self-regulation.

Beyond these specific influences, the broader quality of the parent–child relationship provides a context of security and confidence, or of insecurity and uncertainty, that colors emotional growth and the development of emotion self-regulation (Cassidy 1994). In the security or insecurity of the parent–child attachment relationship that is inaugurated in infancy and maintained throughout childhood, young children experience reassurance that emotions are constructive, enlivening, and manageable facets of life experience through the support of others, or instead learn that emotions are potentially overwhelming and undermining events that are faced alone.

Because of the importance of these relational influences, the 'emotional climate' of the home is a significant influence on early emotional development (Gottman et al. 1997). The emotional climate is significantly shaped by sociocultural values, for example, that guide how emotions are socialized and how children learn to interpret their emotional experiences. In more extreme conditions, unfortunately, the home climate is characterized by the emotional turmoil created by marital dysfunction, parent–child interactions defined by coercion or abuse, or the adult's affective problems. In these circumstances, caregivers not only fail to be sources of emotional support for the child but they contribute to emotional demands that most young children find difficult to manage and which are, for many, potentially overwhelming (e.g.,

Cummings and Davies 1994). As a consequence, children are themselves at risk for emotional disturbance resulting from their inability to cope with the emotional demands they face, their enmeshment in the emotional turmoil of their caregivers, their heightened sensitivity to cues of emotional arousal around them, and their exposure to a compelling model of emotional dysfunction in the adults they live with (Garber and Dodge 1991). Sadly, studies of such children illustrate further the importance of close relationships to healthy emotional growth.

In more constructive family environments, by contrast, parents contribute to the child's growing ability to understand and enlist emotions constructively in social relationships, and enjoy emotional experience personally. In other words, they contribute to the growth of emotional competence (Saarni 1999).

4. *Future Directions*

This discussion reflects two themes that will guide future research on emotional development. The first concerns children's developing representations of emotional experience, especially as they predict children's responses to emotionally arousing events. How do children understand emotion and its influence upon them, and how does this affect their efforts to manage emotion and respond adaptively to situations that evoke emotion? How do these representations change with age and in concert with changes in other representational abilities? The second, and related theme, concerns the emotional dimensions of developmental psychopathology. How are relational influences, self-awareness, and other developmental processes relevant to emotional growth pertinent to how children cope with anxiety disorders, depression, conduct problems, and other emotion-related forms of psychopathology? Future research on both questions will advance normative understanding of emotional development, and its applications to clinical and other practical issues.

See also: Adulthood: Emotional Development; Attachment Theory: Psychological; Communication and Social Psychology; Early Childhood: Socioemotional Risks; Emotions, Children's Understanding of; Emotions, Evolution of; Emotions, Psychological Structure of; Harlow, Harry Frederick (1905–81); Self-regulation in Adulthood; Self-regulation in Childhood

Bibliography

Bartsch K, Wellman H M 1995 *Children Talk About the Mind*. Oxford University Press, New York

Cassidy J 1994 Emotion regulation: Influences of attachment relationships. In: Fox N A (ed.) *The Development of Emotion Regulation: Biological and Behavioral Considerations. Monographs of the Society for Research in Child Development* **59**: 228–49

Cummings E M, Davies P 1994 *Children and Marital Conflict*. Guilford, New York

Denham S A 1998 *Emotional Development in Young Children*. Guilford, New York

Dunn J 1994 Changing minds and changing relationships. In: Lewis C, Mitchell P (eds.) *Children's Early Understanding of Mind*. Erlbaum, Hove, UK, pp. 297–310

Feinman S (ed.) 1992 *Social Referencing and the Social Construction of Reality in Infancy*. Plenum, New York

Garber J, Dodge K A (eds.) 1991 *The Development of Emotion Regulation and Dysregulation*. Cambridge University Press, New York

Gottman J M, Katz L F, Hooven C 1997 *Meta-emotion: How Families Communicate Emotionally*. Lawrence Erlbaum Associates, Mahwah, NJ

Izard C E 1991 *The Psychology of Emotions*. Plenum, New York

Kochanska G, Thompson R A 1997 The emergence and development of conscience in toddlerhood and early childhood. In: Grusec J E, Kuczynski L (eds.) *Parenting and Children's Internalization of Values*. Wiley, New York, pp. 53–77

LeDoux J 1996 *The Emotional Brain*. Simon & Schuster, New York

Lewis M, Haviland-Jones J M 2000 *Handbook of Emotions*, 2nd edn. Guilford, New York

Panksepp J 1998 *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, New York

Saarni C 1999 *The Development of Emotional Competence*. Guilford, New York

Saarni C, Mumme D, Campos J J 1998 Emotional development: Action, communication, and understanding. In: Damon W(ed.), Eisenberg N (vol. ed.) *Handbook of Child Psychology Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, Vol. 3, pp. 237–309

Tangney J P, Fischer K W (eds.) 1995 *Self-conscious Emotions*. Guilford, New York

Thompson R A 1994 Emotion regulation: A theme in search of definition. In: Fox N A (ed.) *The Development of Emotion Regulation: Biological and Behavioral Considerations. Monographs of the Society for Research in Child Development* **59**: 25–52

Thompson R A 1998 Early sociopersonality development. In: Damon W (ed.), Eisenberg N (vol. ed.) *Handbook of Child Psychology Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, Vol. 3, pp. 25–104

Zahn-Waxler C, Radke-Yarrow M 1990 The origins of empathic concern. *Motivation and Emotion* **14**: 107–130

R. A. Thompson

Infant and Child Development, Theories of

From infancy through childhood, and across the adult and aged years, individuals develop, showing systematic and successive changes over time (Baltes 1987). These changes are interdependent with variations in

people's families, communities, societies, cultures, and physical ecologies. Changes within one level of organization, for example, developmental changes in personality or cognition within the child, are reciprocally related to developmental changes within other levels, for example, involving changes in caregiving patterns or spousal relationships within the family (e.g., Lerner 1998).

Moreover, the reciprocal changes among levels of organization are both products and producers of the reciprocal changes within levels. For example, over time, parents' 'styles' of behavior and of rearing influence children's personality and cognitive functioning and development; in turn, the interactions between personality and cognition constitute an emergent 'characteristic' of human individuality that affects parental behaviors and styles and the quality of family life (e.g., Lewis 1997).

These interrelations illustrate the integration of changes within and among the multiple levels of organization comprising the ecology of human life (Bronfenbrenner 1979). Human development within this ecology involves organized and successive changes —that is, systematic changes—in the structure and function of interlevel relations over time (Ford and Lerner 1992). In other words, the human development system involves the integration, or 'fusion' (Tobach and Greenberg 1984), of changing relations among the multiple levels of organization that comprise the ecology of human behavior and development. These levels range from biology through culture and history (Bronfenbrenner 1979, Elder et al. 1993, Gottlieb 1997, Riegel 1975). Indeed, the embeddedness of all levels of the system within history provides a temporal component to human development (Elder et al. 1993); makes the potential for change a defining feature of human development (Baltes 1987); and as such assures that relative plasticity (i.e., the potential for systematic change across ontogeny) characterizes development across the human life span (Lerner 1998).

Given that human development is the outcome of changes in this developmental system, then, for individual ontogeny, the essential process of development involves changing relations between developing people and their changing context (Lerner 1998). Similarly, for any unit of analysis with the system (e.g., for the family, studied over its life cycle, or the classroom, studied over the course of a school year), the same developmental process exists. That is, development involves changing relations between that unit and variables from the other levels of organization within the human development system. Accordingly, the concept of development is a relational one: Development is a concept denoting systemic changes —that is, organized, successive, multilevel, and integrated changes—across the course of life of an individual (or other unit of analysis).

A focus on process, particularly on the process involved in the changing relations between individuals and their contexts, is at the cutting edge of contemporary developmental theory and, as such, is the predominant conceptual frame for research in the study of child development (Lerner 1998). These conceptual and empirical orientations represent the essential approaches within the preponderant majority of theoretically derived scholarship in the contemporary study of human development.

1. *The Evolution of Theories of Infant and Child Development*

The integrative, relational emphasis in current theoretical models of child development stands in contrast to the focuses of theories that predominated in developmental psychology in earlier periods (Cairns 1998). These theories often involved causal splits between nature and nurture (Overton 1998). On the one hand, these theories emphasized predetermined and biological (organismic) bases of development (that is, either genetic or maturational sources of change). Examples are attachment theory (e.g., Bowlby 1969), which stressed the importance for infant and child development of evolutionarily based, early-life interactions between the newborn and the mother; ethological theory (e.g., Lorenz 1965), which emphasized the role of purported instincts in establishing motor and emotional behavior patterns in development (e.g., as occurs when a precocial bird, such as a duck, follows—and forms an alleged immutable social bond, termed 'imprinting'—with the first moving object it sees after its birth); and behavioral genetics (e.g., Plomin 2000), which stresses that, to the extent that genetic differences between people account for the variance in trait distributions (e.g., in regard to interindividual differences in scores for characteristics such as temperament or intelligence), that trait is caused, within the person, by heredity (see *Behavioral Genetics, Psychological Perspectives on*).

On the other hand, some theories that subscribed to a causal split between nature and nurture looked to the environment as the source of behavior and development. Through a mechanistic reductionism (e.g., to the purported stimulus–response connections that comprised all behavior), these theories attempted to account for all changes in people's functioning across life. Examples include the theory of Bijou and Baer (1961), which stressed the application of classical and operant conditioning principles to child development; and the early social learning theories of Bandura and Walters (1959), which emphasized the observation and imitation of models as a basis for the social learning of behaviors displayed by developing children.

Other theories stressed more of an interaction between organismic and environmental sources of development. For instance, Piaget (1970) proposed a developmental theory of cognition, one wherein the

child's knowledge progressed through a series of universal, invariantly ordered, and qualitatively distinct stages that, progressively, involved the elaboration of internal and reversible cognitive operations (or actions). Cognitive development was thought to be propelled by an attempt by individuals to create a balance (an 'equilibrium') between (a) their efforts to change the environment to fit existing knowledge, and (b) their making changes in cognition (accommodating) to match the environment.

In turn, psychoanalytic theorists, such as Sigmund Freud (1954), stressed the importance of instinctual forces or energies (i.e., 'libido') and inborn (nativistic) personality structures (the 'id') in the stages of emotional (or 'psychosexual') development that he proposed as universally applicable to all people. Similarly, neopsychoanalytic theorists, such as Anna Freud (1969) and Erikson (1968), emphasized the role of emergent personality structures (i.e., the 'ego') in the maturationally based stages of emotional development they proposed. While both psychoanalytic and neopsychoanalytic theorists stressed the role of instinctual or maturational contributions to development, they believed that the environment could either change the rate of development across stages or, in extreme cases, arrest emotional development (a situation termed a 'fixation'; Freud 1954).

Nevertheless, despite these differences in emphases on nature, nurture, or the interaction between the two, there remained in the discipline a presupposition that there were two distinct sources of development, that is, that there was a theoretical split between variables associated with heredity and those associated with the context. In other words, there was a belief that the individual and the environment were able to be separated, not only in a conceptual sense but also in the sense of empirical reality. As such, it was the role of theory to explain the contributions of these two separate domains of reality to human development (Overton 1998).

2. *Dimensions of Contemporary Theories*

The stress in contemporary theories, however, is on a 'healing' of the nature–nurture split (Gottlieb 1997), and on accounting for how the integrated developmental system functions, that is for understanding probabilistic epigenesis. Gottlieb (1997 p. 90) defined this process as being 'characterized by an increase of complexity or organization—that is, the emergence of new structural and functional properties and competencies—at all levels of analysis (molecular, subcellular, cellular, organismic) as a consequence of horizontal and vertical coactions among its parts, including organism–environment coactions.'

As such, the forefront of contemporary developmental theory and research is represented by theories of process, of how structures function, and how functions are structured over time (Lerner 1998). For example, most contemporary research about child development is associated with theoretical ideas stressing that the dynamics of individual-context relations provide the bases of behavior and developmental change (Ford and Lerner 1992, Gottlieb 1997). Examples from the literatures of infant and child development include the study of infant motor development (Thelen and Smith 1994), skill development in childhood (Fischer and Bidell 1998), and parent–child relations and parenting during infancy and childhood (Collins et al. 2000).

Thus, in emphasizing that systematic and successive change (i.e., development) is associated with alterations in the dynamic relations among structures from multiple levels of organization, the scope of contemporary developmental theory and research is not limited by (or, perhaps better, confounded by an inextricable association with) a unidimensional portrayal of the developing person (e.g., the person seen from the vantage point of only cognitions, or emotions, or stimulus–response connections, or genetic imperatives; e.g., see Piaget 1970, Freud 1954, Bijou and Baer 1961, Plomin 2000, respectively). Rather, the power of the contemporary stress on processes of dynamic person–context relations is the 'design criteria' imposed on research, method, and application pertinent to the study of any content area or dimension of the developing person. This power is constituted by four interrelated assumptive dimensions of contemporary theories of child development (Lerner 1998). Accordingly, it is useful to discuss these dimensions in order to illuminate the key theoretical and methodological (e.g., research design and measurement) issues pertinent to understanding how biological, psychological, and contextual processes combine to promote behavior and development across the life span.

3. *Change and Relative Plasticity*

Contemporary theories stress that the focus of developmental understanding must be on systemic change (Ford and Lerner 1992). This focus is required because of the belief that the potential for change exists across the life span (e.g., Baltes 1987). Although it is also assumed that systemic change is not limitless (e.g., it is constrained by both past developments and by contemporary contextual conditions), contemporary theories stress that relative plasticity exists across life—although the magnitude of this plasticity may vary across ontogeny (Lerner 1998).

There are important implications of relative plasticity for the application of developmental science. For instance, the presence of relative plasticity legitimates a proactive search across the life span for charac-

teristics of people and of their contexts that, together, can influence the design of policies and programs promoting positive development (Birkel et al. 1989).

4. *Relationism and the Integration of Levels of Organization*

Contemporary theories stress that the bases for change, and for both plasticity and constraints in development, lie in the relations that exist among the multiple levels of organization that comprise the substance of human life (Ford and Lerner 1992, Schneirla 1957). Different models of these levels have been proposed. For instance, Riegel (1975) argued that these levels ranged from the inner biological level, through the individual/psychological level and the proximal social relational level (e.g., involving dyads, peer groups, and nuclear families), to the sociocultural level (including key macro-institutions such as educational, public policy, governmental, and economic systems) and the natural and designed physical ecologies of human development. In turn, Bronfenbrenner (1979) noted that there were instances of the social ecology of human development (such as the home or day care) within which infants and children interacted directly (he termed such settings 'microsystem'). In addition, Bronfenbrenner (1979) noted that there were connections among these microsystems (an ecological level he termed the 'mesosystem') and there were portions of the ecology (such as parents' workplaces) within which children did not interact but that, nevertheless, affected their behavior and development. Finally, he indicated that there was a higher order level of the ecology (the 'macrosystem') that, by being the 'seat' of broad cultural institutions and social policies, influenced all other levels of the ecology of human development.

Despite the terminological differences among these theorists, all were interested in the role of the context of human development and how fused, dynamic, or systemic relations between these levels and the child provided the basis of human development. All these theorists believed that these levels are structurally and functionally integrated with the developing individual, thus requiring a systems view of the levels involved in child development (Ford and Lerner 1992, Sameroff 1983, Thelen and Smith, 1994).

Developmental contextualism (Lerner 1998) is one instance of such a developmental systems perspective. Developmental contextualism promotes a relational unit of analysis as a requisite for developmental study: Variables associated with any level of organization exist (are structured) in relation to variables from other levels; the qualitative and quantitative dimensions of the function of any variable are shaped as well by the relations that variable has with ones from other levels. Unilevel units of analysis (or the components of, or elements in, a relation) are not an adequate target of developmental analysis; rather, the relation itself—the interlevel linkage—should be the focus of such analysis (Lerner 1998, Riegel 1975).

Relationism and integration have a clear implication for unilevel theories of development. At best, such theories are severely limited, and inevitably provide a nonveridical depiction of development, due to their focus on what are essentially main effects embedded in higher order interactions (e.g., see Wahlsten 1990). At worst, such theories are neither valid nor useful. Accordingly neither biogenic theories (e.g., genetic reductionistic conceptions such as behavioral genetics or sociobiology; Freedman 1979, Plomin 2000), psychogenic theories (e.g., behavioristic or functional analysis models; Bijou and Baer 1961), nor sociogenic theories (e.g., 'social mold' conceptions of socialization; e.g., Homans 1961) provide adequate theoretical frames for understanding human development). Simply, neither nature nor nurture theories provide adequate conceptualizations of human development (cf. Hirsch 1970). For instance, theories that stress critical periods of development (e.g., Bowlby 1969, Erikson 1968, Lorenz 1965), that is, periods of ontogeny constrained by biology (e.g., by genetics or by maturation), are seen from the perspective of theories that stress relationism and integration as conceptually flawed (and empirically counterfactual).

Moreover, many nature–nurture interaction theories also fall short in this regard; theories of this type often treat nature and nurture variables as separable entities, and view their connection in manners analogous to the interaction term in an analysis of variance (e.g., Bijou and Baer 1961, Erikson 1968, Plomin 2000, Wahlsten 1990). The cutting edge of contemporary theory moves beyond the simplistic division of sources of development into nature-related and nurture-related variables or processes; instead the multiple levels of organization that exist within the ecology of human development are seen as part of an inextricably fused developmental system.

5. *Historical Embeddedness and Temporality*

The relational units of analysis of concern in contemporary theories are understood as change units. The change component of these units derives from the ideas that all of the above-noted levels of organization involved in human development are embedded in history, that is, they are integrated with historical change (Elder et al. 1993). Relationism and integration mean that no level of organization functions as a consequence of its own, isolated activity. Each level functions as a consequence of its fusion (its structural integration) with other levels (Tobach and Greenberg 1984). History—change over time—is incessant and continuous, and it is a level of organization that is fused with all other levels. This linkage means that change is a necessary, an inevitable, feature of

variables from all levels of organization (Baltes 1987); in addition, this linkage means that the structure, as well as the function, of variables changes over time.

Indeed, at the biological level of organization one prime set of structural changes across history is subsumed under the concept of evolution (Gould 1977, Lewontin et al. 1984); of course, the concept of evolution can be applied also to functional changes (Darwin 1872, Gottlieb 1997). In turn, at more macro levels of organization many of the historically linked changes in social and cultural institutions or products are evaluated in the context of discussions of the concept of progress (Nisbet 1980). The continuity of change that constitutes history can lead to both intraindividual (or, more generally, intralevel) continuity or discontinuity in development—depending on the rate, scope, and particular substantive component of the developmental system at which change is measured (Brim and Kagan 1980). Thus, continuity at one level of analysis may be coupled with discontinuity at another level; quantitative continuity or discontinuity may be coupled with qualitative continuity or discontinuity within and across levels; and continuity or discontinuity can exist in regard to both the processes involved in (or the 'explanations' of) developmental change and in the features, depictions, or outcomes (i.e., the 'descriptions') of these processes (Cairns and Hood 1983).

Of course, not all people will change in the same way. One person may show rapid change, whereas another may change more slowly. Still another person may not change at all. Such differences in patterns of within-person change can occur in regard to any particular substantive area of development (e.g., intelligence, personality, or emotions). As a consequence, between any two or more times in life, and for any domain of development, there may be differences between people in how they change. In other words, there may be between-person differences (i.e., interindividual differences) in within-person change (i.e., in intraindividual change). When such between-person differences arise they are termed 'instability.' If, however, there are not such between-person differences present between two or more points in time, then stability is present. Baltes and Nesselroade (1973) have explained that in the study of human development across the life span one may find any combination of continuity–discontinuity (within person constancy or change) and stability–instability (between-people differences—in their respective patterns of intraindividual change. In other words, continuity–discontinuity is a dimension of intraindividual change and is distinct from, and independent of, stability–instability—which involves between-person change, and is, therefore, a group, and not an individual, concept (Baltes and Nesselroade 1973).

In sum, since historical change is continuous, temporality is infused in all levels of organization. This infusion may be associated with different patterns of continuity and discontinuity across individuals. The potential array of such patterns has implications for understanding the importance of human diversity.

6. *The Limits of Generalizability, Diversity, and Individual Differences*

The temporality of the changing relations among levels of organization means that changes that are seen within one historical period (or time of measurement), and/or with one set of instances of variables from the multiple levels of the ecology of human development, may not be seen at other points in time (Baltes et al. 1977, Bronfenbrenner 1979). What is seen in one data set is only an instance of what does or what could exist. Accordingly, contemporary theories focus on diversity —of people, of relations, of settings, and of times of measurement (Lerner 1998).

Individual differences within and across all levels of organization are seen as having core, substantive significance in the understanding of human development (Baltes 1987). Diversity is the exemplary illustration of the presence of relative plasticity in human development (Lerner 1998). Diversity is also the best evidence that exists of the potential for change in the states and conditions of human life (Brim and Kagan 1980).

Moreover, the individual structural and functional characteristics of children constitute an important source of their development (Lerner 1998). The individuality of each child promotes variation in the fusions he or she has with the levels of organization within which the child is embedded. For instance, the distinct actions or physical features of children promote differential actions (or reactions) in others toward them (Lerner 1998). These differential actions, which constitute feedback to the child, shape at least in part further change in his or her characteristics of individuality (Schneirla 1957). For example, the changing match, congruence, or goodness of fit between the developmental characteristics of children and of their context provide a basis for consonance or dissonance in their ecological milieu; the dynamic nature of this interaction constitutes a source of variation in positive and negative outcomes of developmental change (Thomas and Chess 1977).

The major assumptive dimensions of contemporary theories of child development—systematic change and relative plasticity, relationism and integration, embeddedness and temporality, generalizability limits and diverstiy—are very much intertwined facts of a common paradigmatic core. And, as is also the case with the levels of organization that are integrated to form the substance of developmental change, the assumptive dimensions form the corpus of superordinate developmental systems views of human development (Ford and Lerner 1992), for example,

developmental contextualism. As is the case with the several defining features of the life span developmental perspective, which—according to Baltes (1987)—need to be considered as an integrated whole, the assumptive dimensions of contemporary developmental theories need to be appreciated simultaneously. Such appreciation is required to understand the breadth, scope, and implications for research and application of this 'family' of conceptual frameworks.

7. *Conclusions*

Contemporary theories of child development and the research associated with it take an integrative approach to the multiple levels of organization presumed to comprise the nature of human life; that is, 'fused' (Tobach and Greenberg 1984) relations among biological, psychological, and social and physical contextual levels comprise the process of developmental change in human life. Rather than approach variables from these levels of analysis in either a reductionistic or in a parallel-processing approach, theories, such as developmental contextualism (Lerner 1998), rest on the idea that variables from these levels of analysis are dynamically interactive—they are reciprocally influential over the course of human ontogeny.

It is such ideas that shape an increasing proportion of the empirical work in the study of child development, especially those instances that reflect longitudinal, change-sensitive, multilevel, integrated, and dynamic approaches to the study of child–context relations. Only through such research will adequate understanding be developed of the bases and import of the multiple pathways that comprise the life span.

In turn, theoretical views such as developmental contextualism do not just provide an agenda for a developmental, dynamic, and systems approach to research about child development. Such theory also allows us to envision the possibility of promoting positive developmental trajectories in children (Lerner 1998). We may actualize this vision if we remain assiduously committed to a developmental systems orientation; if we recognize the 'double-edged sword' nature of plasticity that derives from the functioning of this system; and if we therefore create, through policies and programs, a social context that promotes positive development across the life course. Such a context would provide a social network encompassing the familial, community, institutional, and cultural components of the ecology that impacts people's behavior and development across their life.

Ultimately, then, there is great potential 'practical' (applied) value for developmental theories (Lewin 1943). Indeed, at the beginning of the twenty-first century there is growing recognition that traditional and artificial distinctions between science and service, and between knowledge generation and knowledge application and practice, need to be reconceptualized if society is to successfully address the harrowing developmental sequels of the social, economic, and geopolitical legacies of the twentieth century (Lerner et al. 2000). Scholars, practitioners, and policy makers are increasingly recognizing the important role that developmental science can play in stemming the tide of insults to the quality of life caused by poverty, premature births, school failure, child abuse, crime, adolescent pregnancy, substance abuse, unemployment, welfare dependency, discrimination, ethnic conflict, and inadequate health and social resources.

Successful application of the theory, substantive knowledge, and methodology of developmental science to the promotion of human capital will be determined by the extent to which research findings can be effectively related to problems of daily living (Lerner et al. 2000). Research designs that examine topics of immediate social concern, that consider both normative and atypical developmental pathways as means of promoting and enhancing human welfare, that take into account the contextual nature of development and employ ecologically valid means of assessing functioning, and that are sensitive to the ethical dimensions of action research are required if science is to make a difference in the lives of community members, young and old. Without such research, the knowledge produced by developmental scientists risks being ignored or misused by practitioners, educators, the public, policy makers, and other potential consumers. With such theory-derived research, developed by combining the assets of our scholarly and research traditions with the strengths of educators, practitioners, and policy makers serving the community, such collaborations can demonstrate that a science devoted to using its scholarship to improve the life chances of all people is of inestimable value to civil society.

See also: Adolescent Development, Theories of; Adult Development, Psychology of; Aging, Theories of; Attachment Theory: Psychological; Behavioral Genetics: Psychological Perspectives; Cognitive Development in Childhood and Adolescence; Developmental Psychology; Lifespan Development, Theory of; Piaget's Theory of Child Development; Prenatal and Infant Development: Overview; Psychoanalysis: Adolescence (Clinical–Developmental Approach); Socialization in Adolescence; Socialization in Infancy and Childhood

Bibliography

Baltes P B 1987 Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* **23**: 611–26

Baltes P B, Nesselroade J R 1973 The developmental analysis of individual differences on multiple measures. In: Nesselroade J R, Reese H W (eds.) *Life-span Developmental Psychology: Introduction to Research Methodological Issues*. Academic Press, New York, pp. 219–51

Baltes P B, Reese H W, Nesselroade J R 1977 *Life-span Developmental Psychology: Introduction to Research Methods*. Brooks/Cole, Monterey, CA
Bandura A, Walters R H 1959 *Adolescent Aggression*. Ronald Press, New York
Bijou S W, Baer D M (eds.) 1961 *Child Development*. Appleton-Century-Crofts, New York
Birkel R, Lerner R M, Smyer M A 1989 Applied developmental psychology as an implementation of a life-span view of human development. *Journal of Applied Developmental Psychology* **10**: 425–45
Bowlby J 1969 *Attachment and Loss: Vol.1 Attachment*. Basic Books, New York
Brim O G Jr., Kagan J (eds.) 1980 *Constancy and Change in Human Development*. Harvard University Press, Cambridge, MA
Bronfenbrenner U 1979 *The Ecology of Human Development*. Harvard University Press, Cambridge, MA
Cairns R B 1998 The making of developmental psychology. In: Damon W (Series ed.), Lerner R M (Vol. ed.) *The Handbook of Child Psychology: Theoretical Models of Human Development*. Wiley, New York, 5th edn./Vol. 1, pp. 25–106
Cairns R B, Hood K E 1983 Continuity in social development: A comparative perspective on individual difference prediction. In: Baltes P B, Brim O G Jr. (eds.) *Life-span Development and Behavior*. Academic Press, New York Vol. 5, pp. 301–58
Collins W A, Maccoby E E, Steinberg L, Hetherington E M, Bornstein M H 2000 Contemporary research on parenting: The case for nature and nurture. *American Psychologist* **55**: 218–32
Darwin C 1872 *The Expression of Emotion in Men and Animals*. J. Murray, London
Elder G H Jr., Modell J, Parke R D 1993 Studying children in a changing world. In: Elder G H, Modell J, Parke R D (eds.) *Children in Time and Place: Developmental and Historical Insights*. Cambridge University Press, New York, pp. 3–21
Erikson E H 1968 *Identity, Youth and Crisis*. 1st edn. Norton, New York
Fischer K W, Bidell T R 1998 Dynamic development of psychological structures in action and thought. In: Damon W (Series ed.), Lerner R M (Vol. ed.) *The Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York; Vol. 1, pp. 467–561
Ford D L, Lerner R M 1992 *Developmental Systems Theory: An integrative Approach*. Sage, Newbury Park, CA
Freedman D G 1979 *Human Sociobiology: A Holistic Approach*. The Free Press, New York
Freud A 1969 Adolescence as a developmental disturbance. In: Caplan G, Lebovici S (eds.) *Adolescence*. Basic Books, New York, pp. 5–10
Freud S 1954 *Collected Works*. Hogarth, London
Gottlieb G 1997 *Synthesizing Nature–Nurture: Prenatal Roots of Instinctive Behavior*. Erlbaum, Mahwah, NJ
Gould S J 1977 *Ontogeny and Phylogeny*. Belknap Press of Harvard University Press, Cambridge, MA
Hirsch 1970 Behavior-genetic analysis and its biosocial consequences. *Seminars in Psychiatry* **2**: 89–105
Homans G C 1961 *Social Behavior: Its Elementary Forms*. Harcourt, Brace & World, New York
Lerner R M 1998 Theories of human development: Contemporary perspectives. In: Damon W (Series ed.), Lerner R M (Vol. ed.) *The Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York Vol. 1, pp. 1–24
Lerner R M, Fisher C B, Weinberg R A 2000 Applying developmental science in the 21st: International scholarship for our times. *International Journal of Behavioral Development* **24**: 24–9
Lewin K 1943 Psychology and the process of group living. *Journal of Social Psychology* **17**: 113–31
Lewis M 1997 *Altering Fate*. Guilford Press, New York
Lewontin R C, Rose S, Kamin L J 1984 *Not in Our Genes: Biology, Ideology, and Human Nature*. Pantheon, New York
Lorenz K 1965 *Evolution and Modification of Behavior*. University of Chicago, Chicago
Nisbet R A 1980 *History of the Idea of Progress*. Basic Books, New York
Overton W F 1998 Developmental psychology: Philosophy, concepts, and methodology. In: Damon W (Series ed.), Lerner R M (Vol. ed.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 107–89
Piaget J 1970 Piaget's theory. In: Mussen P H (ed.) *Carmichael's Manual of Child Psychology*. Wiley, New York, Vol. 1, pp. 703–32
Plomin R 2000 Behavioural genetics in the 21st century. *International Journal of Behavioral Development* **24**: 30–4
Riegel K F 1975 Toward a dialectical theory of development. *Human Development* **18**: 50–64
Sameroff A J 1983 Developmental systems: Contexts and evolution. In: Kessen W (ed.) *Handbook of Child Psychology: History, Theory, and Methods*. Wiley, New York, Vol. 1, pp. 237–94
Schneirla T C 1957 The concept of development in comparative psychology. In: Harris D B (ed.) *The Concept of Development*. University of Minnesota, Minneapolis, MN, pp. 78–108
Thelen E, Smith L B 1994 *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA
Thomas A, Chess S 1977 *Temperament and Development*. Brunner/Mazel, New York
Tobach E, Greenberg G 1984 The significance of T. C. Schneirla's contribution to the concept of levels of integration. In: Greenberg G, Tobach E (eds.) *Behavioral Evolution and Integrative Levels*. Erlbaum, Hillsdale, NJ, pp. 1–7
Wahlsten D 1990 Insensitivity of the analysis of variance to heredity–environment interaction. *Behavioral and Brain Sciences* **13**: 109–20

R. M. Lerner

Infant and Child Mortality: Central and Eastern Europe

1. Historical Outlook

At the beginning of the twentieth century, Central and Eastern Europe was handicapped compared with the West in terms of delayed socioeconomic development. Higher mortality rates for infants and children reflected this situation. While infant deaths per 1,000 live births in western and northern European countries ranged from 80 (Norway) to 150 (Belgium), Central Europe had levels exceeding 200: Czech lands, comprising the current Czech Republic, 225, Slovakia 207, Hungary 213, and Austria 216. The infant mortality

rate was equally high in Germany (199). In addition to delayed socioeconomic development, additional factors contributed to this unfavorable pattern—notably lifestyles, techniques of childrearing (breastfeeding, weaning, wet-nursing), and the status of children (Chesnais 1992). Eastern European countries had similarly elevated levels. In Romania, one-fifth of the children died before their first birthday. A slightly lower infant mortality rate was reported in Bulgaria (161) and Poland (163). In Russia, the areas with a predominantly ethnic Russian population had a rate of 298 infant deaths per 1,000 live births, in contrast to 188 in the other ethnic groups. A desperate situation in infant mortality was primarily attributed to the general socioeconomic underdevelopment of the country, habits and traditions associated with a predominantly rural culture, and an overall lack of knowledge about sanitary and health care. Infant mortality rates in urban and rural areas were virtually the same. The infant mortality rate in Russia tended to be much higher in the Orthodox population (263) than in the Catholic (151), Lutheran (161), and Judaic (116) populations. The observed differentials according to religious affiliation have been explained by different patterns of infant feeding (Zakharov 1994).

A long-term trend of decline in infant mortality in Europe started at the end of the nineteenth century, but Romania, Bulgaria, Albania, and Russia lagged behind. The decline was interrupted by peaks due to the two world wars. In Russia, the evolution was, in addition, disturbed by upheavals related to civil war (1918–22), famines (1921, 1933, and 1946), and forced collectivization. This unfavorable situation in infant mortality patterns of Central and Eastern European countries remained unchanged until after the end of World War II.

After World War II, Eastern Europe (now called Central and Eastern Europe) appeared as a new political region. This socialist bloc displayed strong inequalities in demographic behavior, economic indicators and social structures. In 1950, the infant mortality rate varied from 64 per 1,000 live births in the current Czech Republic to more than 100 in Romania, Bulgaria, and the former Soviet Union. In Western Europe the risk of infant death ranged from 21 (Sweden) to 55 (Federal Republic of Germany) at that time. During the communist regime in Eastern Europe, the effort toward massive schooling and rapid coverage of the entire population with basic health services reduced the greatest contrasts among the socialist countries as well as within individual countries. The downward shift in infant mortality rate was more pronounced for less economically advanced countries, and this led toward greater homogeneity. The progress was particularly strong during the 1950s when predominantly organizational measures led to decreases in infant mortality, as well as in maternal mortality (Štembera 1993). Since the 1960s, Western countries experienced a significant drop in infant mortality accompanied by an even stronger decrease in neonatal mortality and mortality of low-birth-weight babies. In contrast to Western countries, where special care policies were developed for high-risk pregnancies, uniform prenatal care was stressed in former socialist countries on the presumption that social risk differentials were nonexistent. The health-extensive approach based on a large-size medical staff with only basic training could not compete with a health-intensive system that involved specialized training and sophisticated equipment. As a consequence, infant mortality rates stagnated or deteriorated in all Eastern European countries. Data on infant mortality were not published in the USSR from 1974 to 1987 because of this increase.

2. *Current Trends*

During the postsocialist social and economic transition, such uniformity has been broken and previous inequalities have emerged again. In most former republics of the Soviet Union (FSU), the recorded infant mortality rate has increased (Blum and Avdeev 1994), partly due to the introduction (primarily in the Baltic States) of the standard WHO definition of a live birth. Infant mortality has decreased only slightly in the transition period in the Czech Republic, Slovenia, Slovakia, Hungary, and Poland. Recently (1996, 1997), an improvement has been observed in Estonia and Lithuania. With a 1999 infant mortality rate of less than five per 1,000, the Czech Republic and Slovenia are the closest to Western European standards.

2.1 *Age-specific Patterns*

The risk of infant death currently depends primarily on intrauterine development, and the role played by post-neonatal mortality is minimal (for children older than 27 completed days). The reduction and frequency of high-risk pregnancies determine the level of infant mortality. There are particularly strong differentials across countries in mortality in the early neonatal period (less than seven completed days), particularly for very low birth-weight babies (less than 1,500 g). At about 12 deaths per 1000 live-born children, the infant mortality rate in 1996 was twice as high in Poland as in the Czech Republic, due primarily to extremely high risks in the early neonatal period and/or low birth weight.

2.2 *Medical Causes of Infant Deaths*

Congenital anomalies and perinatal conditions are the most frequent causes of death in infants under the age of one year. The variation of infant mortality rate between these countries varies by a factor of three for congenital anomalies and by a factor of five for perinatal conditions. The lowest rates are in Slovenia and the highest in Latvia. Romania and Moldova are

Table 1
Czech Republic 1986–92. Interaction of mother's age and birth order (relative risks adjusted for birth weight, gestation, marital status, and education)

Mother's age	Birth order 1	2	3	4+
−19	1.038	1.563	1.006	
20–4	1	1.269	1.818	1.357
25–9	0.951	1.065	1.394	1.508
30–4	0.986	0.969	1.111	1.249
35+	1.262	1.102	1.105	1.384

Table 2
Infant mortality rate according to mother's marital status per 1,000 live births (singletons) combined with birth weight (g)

	Single			
Birth Weight	Czech Rep. 1988	Hungary 1988	Belgium 1988	Czech Rep. 1992
−2499	93.7	90.4	85.0	87.4
2500+	7.1	11.2	5.8	4.4
Total	15.5	24.0	11.9	12.8
	Married			
Birth Weight	Czech Rep. 1988	Hungary 1988	Belgium 1988	Czech Rep. 1992
−2499	115.3	112.8	86.0	99.7
2500+	4.9	5.3	4.3	3.8
Total	9.6	13.1	7.9	7.9
	Divorced/Widowed			
Birth Weight	Czech Rep. 1988	Hungary 1988	Belgium 1988	Czech Rep. 1992
−2499	130.9	121.15	7.7	96.3
2500+	6.3	8.8	3.4	6.7
Total	17.7	26.7	8.1	14.5
	Total			
Birth Weight	Czech Rep. 1988	Hungary 1988	Belgium 1988	Czech Rep. 1992
−2499	113.7	109.6	84.8	97.6
2500+	5.1	5.9	4.4	3.9
Total	10.1	14.5	8.3	8.5

in a rather exceptional situation in that infections and respiratory diseases are significant causes of infant mortality in the post neonatal period in addition to the previously mentioned groups.

2.3 Demographic and Social Factors in Infant Mortality

Demographic variables such as the mother's age and birth order have J-shaped profiles in conventional studies of infant mortality patterns. Differentials according to age and birth order are more pronounced in the Czech Republic, Hungary, and Poland than in Belgium, for example. The increase in the infant mortality rate is particularly striking after the age of 35 and for birth orders four and more. According to a special study for the Czech Republic from the period of 1986–92 (Rychtaříková 1999), the negative impact of higher birth order can also be seen in the interaction with mother's age (Table 1). The results concerning

Table 3
Infant mortality rate according to mother's education per 1,000 live births (singletons) combined with birth weight (g)

	Basic		
Birth Weight	Czech Rep. 1988	Hungary 1988	Czech Rep. 1992
−2499	102.2	105.0	86.0
2500+	7.0	7.7	6.5
Total	15.0	20.7	14.5
	Vocational		
Birth Weight	Czech Rep. 1988	Hungary 1988	Czech Rep. 1992
−2499	110.3	94.4	106.4
2500+	5.1	8.8	4.2
Total	9.9	15.4	9.1
	Secondary and Higher		
Birth Weight	Czech Rep. 1988	Hungary 1988	Czech Rep. 1992
−2499	126.4	115.7	96.2
2500+	4.5	4.9	2.9
Total	8.7	11.2	6.2

relative risks (controlled for birth weight, duration of gestation, sex, mother's education, and marital status) suggest that to have an additional child at an 'inappropriate' age is risky. Thus, to have a second child at an age less than 20, to have a third child in the age group 20–24, or to have a third or fourth child at age 25–29 implies an elevated death risk for the child. Birth order also has social implications, because a higher relative risk of infant death is observed for children of a higher birth order even when the child has a normal birth weight (2,500g or more).

Two Central European countries (the Czech Republic and Hungary) have a gradient of age-specific infant mortality rates according to marital status similar to that of Belgium (Table 2). Lower values are experienced by children born to married women than to single, divorced, or widowed mothers (Rychtaříková 1995). In Belgium, differences in pregnancy outcomes are smaller between married and single mothers than in both these Central European countries. However, the pattern is inverted when the outcomes according to marital status are combined with low birth weight. Infants weighing less than 2,500 g have a better chance of survival when born to a single mother than when the mother is married (Syrovátkar and Rychtaříková 1984). A higher frequency of low birth weight in single mothers results in a higher crude (nonadjusted) infant mortality rate.

The impact of the education of mothers as a determinant of infant mortality could be questionable in an egalitarian society with weak social-class differentiation, and where health care and social opportunities are accessible to everyone. However, education is also a proxy for socioeconomic status and lifestyle. When birth weight is disregarded, the infant mortality rate decreases when the mother's educational level increases, as illustrated in Table 3. This pattern is also valid for children born in normal birth-weight categories. For lower birth weight (less than 2,500 g), a similar inversion occurred in the Czech Republic in 1988 and 1992. In this weight class, children born to mothers with no more than a basic education had the lowest infant risk of death. This phenomenon has also been documented in the US Afro-American population (Wilcox and Russel 1990).

During the period of postsocialist transition, children in formerly socialist countries have been more affected by the increase in poverty than other traditionally vulnerable population groups, such as the elderly. A number of countries also show some evidence of increasing child maltreatment, including the detrimental use of child labor. In Central and Southeastern Europe, unemployment and losses in family benefits have penalized families with children; in the successor republics of the FSU countries, an additional factor has been the dramatic rise in wage inequality. Also reflecting the effect of unfavorable social settings is the excess postneonatal mortality in Romania, Moldova, Ukraine, and Bulgaria (Fig. 1(a)). When child mortality rates at ages 1–4 and 15–19 years are correlated, most FSU countries show increased risks for teenagers. With the weakening of

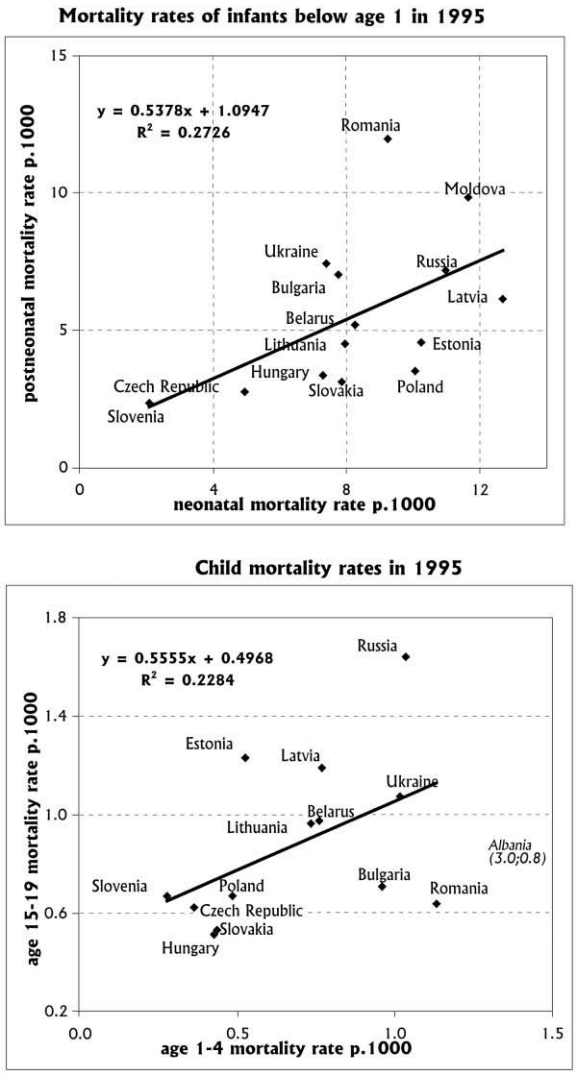

Figure 1
(a) Mortality rates of infants below age 1 year in 1995;
(b) child mortality rates in 1995

social controls, new threats have arisen—drinking, smoking, drug abuse, sexually transmitted diseases, accidents, poisonings, and violence. The reappearance of diseases of poverty, including tuberculosis, has been reported among children. Historical inequalities observed at the beginning of this century are re-emerging today because of currently growing contrasts between former socialist countries in infant and child survival, including mother's health. In 1995, the maternal mortality rate ranged from about five maternal deaths per 1,000,000 live births in Central Europe (except Hungary) to more than 40 in Russia, Moldova, and Romania. Factors known as risky for pregnancy or infant/child survival have been significant in FSU countries, as well as in Romania and Bulgaria, where the ratio of legal abortions to live births—as well as the share of births to mothers below age 20—is high as compared with Central Europe. All countries in transition have poor access to health services and screening. The responsibility for monitoring children has been shifting from schools to parents. The system of family support has been eroded considerably during the period of transition. The observed deterioration of public health and family support systems has been disproportional, as has the development of infant/child mortality. Currently, four different regions are emerging in child health and infant mortality, namely, Central Europe, Southeastern Europe, the Baltic countries, and the other successor countries to the former Soviet Union.

See also: Eastern European Studies: Economics; Eastern European Studies: Society; Infant and Child Mortality in Industrialized Countries; Infant and Child Mortality in the Less Developed World; Mortality, Biodemography of; Mortality: The Great Historical Decline

Bibliography

Blum A, Avdeev A 1994 *La Mortalité Infantile dans les États Issus de l'URSS*. Santé et Mortalité des Enfants en Europe, Inégalités Sociales d'Hier et d'Aujourdhui, Chaire Quetelet. Louvain-la-Neuve, Belgium

Chesnais J C 1992 *The Demographic Transition*. Clarendon Press, Oxford, UK

Gardos E, Rychtaříková J 1994 *Recent Trends in Infant Health and Mortality in Central and Eastern Europe*. Santé et Mortalité des Enfants en Europe, Inégalités Sociales d'Hier et d'Aujourdhui, Chaire Quetelet. Louvain-la-Neuve, Belgium

Rychtaříková J 1995 *Diferenční Aspekty Kojenecké Úmrtnosti v České Republice, Belgii a Maďarsku* (*Differentials of Infant Mortality in the Czech Republic, Belgium and Hungary*). Neonatalogické Listy 1/1995 4/ÚPMD, Praha, Czech Republic

Rychtaříková J 1999 Sociální a biologické faktory kojenecké úmrtnosti (Social and biological factors of infant mortality). *Demografie* (Czech Republic) **2**

Štembera Z 1993 Influence of political changes on Eastern European perinatology: Trends in Czech perinatology after disintegration of communist rule. *Journal of Maternal-Fetal Investigation* **3**

Syrovátka A, Rychtaříková J 1984 Naissances vivantes et décès de moins d'un an selon le poids à la naissance en République Socialiste Tchèque entre 1950 et 1980. *Population* (France)**3**

UNICEF 1997 *Children at Risk in Central and Eastern Europe: Perils and Promises*. Monee Project Regional Monitoring Report 4.

Wilcox A J, Russel I T 1990 Why small black infants have a lower mortality rate than small white infants: The case for population-specific standards for birth weight. *The Journal of Pediatrics* **116**: 7–10

Zakharov S V 1994 *The Second World War as a Turning Point of Infant Mortality Decline in Russia*. Santé et Mortalité des

Enfants en Europe, Inégalités Sociales d'Hier et d'Aujourdhui, Chaire Quetelet. Louvain-la-Neuve, Belgium

J. Rychtaříková

Infant and Child Mortality in Industrialized Countries

In most industrialized countries, about five out of 1,000 newborn children will die before their first birthday. A similar risk level will not be reached again before age 60. The first year of life has a very high risk of death, especially when compared to the 10 or 15 years that follow. At those ages mortality is very low, namely less than 0.5 deaths per 1,000 children between the first and fifth birthday, and below 0.2 deaths between the ages of five and 15. Infant mortality (during the first year) and child mortality (during the four following years) are quite different from each other. Infant mortality essentially results from circumstances during pregnancy and delivery, as well as from congenital anomalies, while child mortality depends on several causes, among which accidents become increasingly prominent.

1. A Century of Continuous Improvement in the Struggle Against Infant Mortality

Infant mortality is measured by the infant mortality rate, which is usually computed by dividing the number of infant deaths (i.e., deaths before the first birthday) observed during a calendar year by the total number of live births during the same year. If the total number of births fluctuates widely, it may be necessary to distinguish infant deaths among births of the index year from those of births from the previous year, and to divide each one by the corresponding total number of live births.

At the beginning of the twentieth century, only a few countries had infant mortality rates below 150 deaths per 1,000 live births. Present rates are often between four and five deaths per 1,000. This is 25 to 30 times lower, and is due to important improvements during the late twentieth century. At the beginning of the twenty-first century we may hope for, not the disappearance of infant mortality, but at least a continuing further decrease to an even lower level of perhaps 3 per 1,000, something which would have been unimaginable in 1990.

Figure 1 shows infant mortality rates observed in England and Wales since 1840, and illustrates the scale and tempo of the changes in infant mortality. There was no progress in the nineteenth century, and the years 1880 and 1890 even showed an increase in mortality. It was only during the first years of the twentieth century that infant mortality started to decrease. It reached levels below 100 infant deaths per 1,000 live births by the end of World War I, and below 50 per 1,000 births by the beginning of World War II. In this 40-year period infant mortality was reduced largely to a third of its level at the turn of the twentieth century. This improvement was the result of general

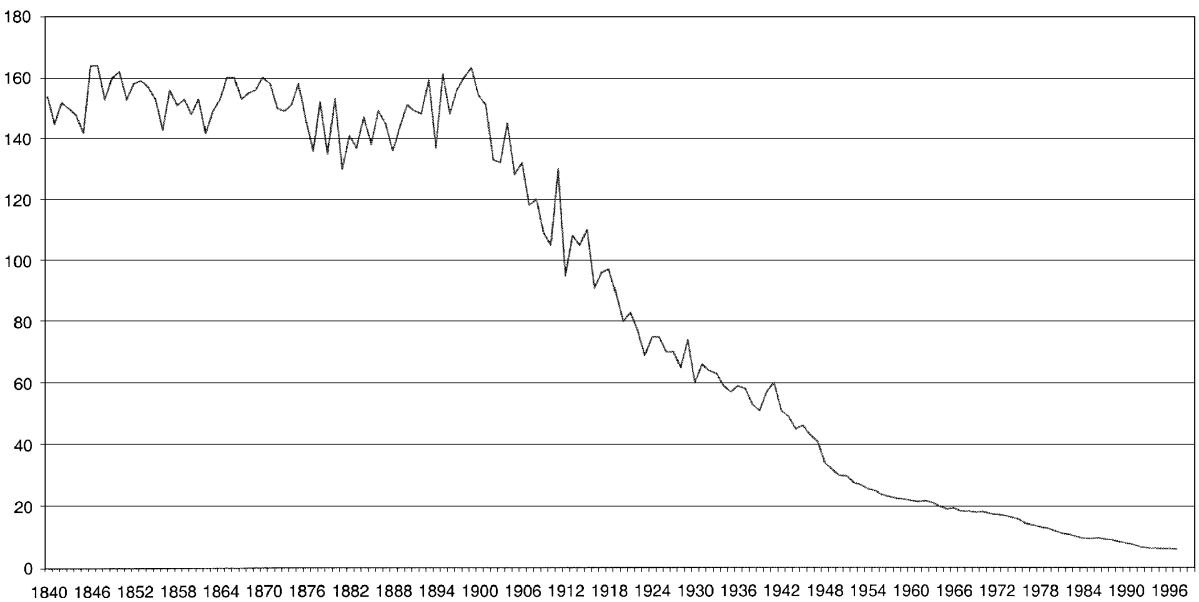

Figure 1
Infant mortality rates: England and Wales (1840–1997) (deaths before one year for 1,000 living births)

progress in medicine, and also improvements in hygiene, especially concerning feeding practices. Since World War II, that is during a period roughly as long as the preceding one, the infant mortality rate in England and Wales was reduced to a seventh of what it had been. In particular, during the years immediately following the war, the diffusion of antibiotics helped reduce the infant mortality rate to less than 30 per 1,000 at the beginning of the 1950s. Since the middle of the 1980s, infant mortality has been under 10 per 1,000, and one might think that progress would have slowed down since then. Nothing of the kind has occurred, however. On the contrary, since the beginning of the 1990s even more important progress has taken place. Due to the increase in antibiotics the infant mortality crisis has practically disappeared. Until the beginning of the 1940s mortality risks were still acute at this age. England and Wales were among many countries who saw such a development. Comparable trends have been observed in all industrialized countries, although some started earlier and many others a little later.

2. *Persistent Differences Between Countries*

A few years after the end of World War II, there were important differences in infant mortality between industrialized countries (Table 1). In Europe, infant mortality levels can be organized approximately along two geographical axes: mortality increases from north to south, and mortality increases from west to east. A map of infant mortality can be drawn based on a 'center' from which mortality of newborn children goes up as one moves towards the 'periphery.' At the beginning of the 1950s, the situation was relatively satisfactory in only a small group of countries that included Scandinavia, the UK, The Netherlands, and Switzerland, which had rates roughly situated between 20 and 30 per 1,000. It was less acceptable in a second group, composed of Germany, France, Belgium, Spain, and Italy, which had rates between 40 and 70 per 1,000; and it was comparatively disastrous in other countries, namely in Central and Eastern Europe, where rates varied from 80 to 140 per 1,000.

During the subsequent years, all countries benefited from the diffusion of antibiotics and the resulting decrease of infant mortality, but improvements varied. Between 1950 and 1960 they remained under 30 percent for countries already more advanced in the struggle against infectious mortality, and they were also modest in the most isolated of Southern European countries (Portugal, the Balkans), in which infectious diseases are still deadly today. On the other hand, during the same period infant mortality rates were roughly halved in a whole group of Western European (Germany, France, Belgium, etc.) and Central European countries (Bulgaria, Hungary, Poland, Czechoslovakia, etc.). The average infant mortality rate for Europe as a whole decreased from 71 to 42 per 1,000, so there was an improvement of 41 percent. This does not include Eastern Europe, for which data for 1950 are unknown.

Between 1960 and 1970, the improvements were less impressive (31 percent for Europe as a whole), and the average infant mortality rate in 1970 was 29 per 1,000. No country had obtained a rate of less than 10 per 1,000, and the countries still showing rates of more than 50 per 1,000 were the same as before. These were still countries such as Portugal and, in particular, Albania and some Yugoslav republics.

The following twenty years showed stagnation in Eastern Europe and increasing improvement in Western Europe, with a demarcation line which reproduced the division between the two main geopolitical groups at the time. This development was especially clear during the 1970s: the average infant mortality rate was halved in the West, while it was only reduced by a third in the East.

In 1990, infant mortality in Europe was a picture of contrasts. In the West, no country except Portugal had an infant mortality rate over 10 per 1,000. In the East, all countries except Slovenia had rates above this threshold, and some even had 20 per 1,000 (Albania, Macedonia, Romania, Yugoslavia). The most recent data show the permanence of this division; except that a few countries of Central and Eastern Europe (Slovenia and the Czech Republic in particular) now record rates which are equivalent to those observed in the West.

Since 1950, the spectacular decrease of mortality in Europe has not been able to mask the stability of the respective positions of the various countries. Northern European countries, Switzerland, and The Netherlands have always held places in the highest rank, while the lowest ranks have always been occupied by Albania, Bosnia-Herzegovina, Macedonia, Romania, and Yugoslavia. Another clear phenomenon is the relative gap between the most advanced countries and the most backward ones. This increased until about 1980. At the end of the 1990s, it decreased but remained comparable with the differences of 50 years previously. The infant mortality rates of the three most advanced countries in 1999 were a fifth of those of the three most backward countries, just as in 1950.

In non-European industrialized countries, progress was similar to that in Western Europe. The case of Japan deserves special notice. In Japan, a country of 126 million inhabitants, the infant mortality rate (4 per 1,000) has been the same as in Sweden or in Switzerland, countries of only a few million inhabitants.

On the whole, the last years of the twentieth century were important years of improvement. Infant mortality has dropped to 5 per 1,000 in a growing number of countries, and this pace of improvement remains unbroken. There is surely no possibility of eliminating

Table 1
Infant mortality rates in industrialized countries (1950–1998): number of deaths during the first year of life per 1,000 live births

	1950	1960	1970	1980	1990	1999
Northern Europe						
Denmark	31	22	14	8	8	4
Finland	44	21	13	8	6	4
Iceland	22	13	13	8	6	2
Norway	28	19	13	8	7	4
UK	31	23	19	12	8	6
Sweden	21	17	11	7	6	3
Western Europe						
Austria	66	38	26	14	8	4
Belgium	53	31	21	12	8	5
France	52	27	18	10	7	5
Germany	60	35	23	12	7	5[a]
Ireland	46	29	20	11	8	6
Luxembourg	48	32	25	12	7	5
The Netherlands	27	18	13	9	7	5
Switzerland	31	21	15	9	7	5
Southern Europe						
Greece	35	40	30	18	10	6
Italy	64	44	30	15	8	5
Portugal	94	78	56	24	11	6
Spain	64	44	28	12	8	5
Albania	121	83	98	50	28	12
Bosnia and Herzegovina	126	107	69	32	15	11[a]
Croatia	118	70	34	21	11	8
Macedonia	137	115	88	54	32	15[a]
Slovenia	81	35	25	15	8	4
Yugoslavia	117	85	54	33	23	13[a]
Central Europe						
Bulgaria	95	45	27	20	15	15
Czech Republic	64	20	20	17	11	5
Hungary	86	48	36	23	15	8
Poland	108	56	36	25	19	9
Romania	117	76	49	29	27	19
Slovakia	103	29	26	21	12	8
Eastern Europe						
Russia	–	37	23	22	18	17
Belarus	–	35	19	16	12	11
Estonia	–	31	18	17	12	10
Latvia	–	27	18	15	14	12
Lithuania	–	38	19	15	10	9
Moldova	–	49	24	35	19	19
Ukraine	–	30	17	17	13	13
Outside Europe						
Canada	42	27	19	10	7	6[a]
USA	29	26	20	13	9	7
Australia		20	18	11	8	5[a]
New Zealand	28	23	17	13	8	5[a]
Japan	60	31	13	7	5	4

Sources: National statistics, UN, WHO
a 1998 – not available.

every risk of death in this age group, as newborn babies are especially fragile, but a rate of roughly 3 per 1,000 might be reached in the early twenty-first century.

3. Components of Infant Mortality

Infant mortality is a complex and heterogeneous phenomenon. It depends on the circumstances of pregnancy and delivery, as well as on genetic factors. In addition it reflects the newborn child's fragility towards attacks from its external environment, especially infectious disease.

We can separate two large groups of causes, namely 'endogenous mortality,' which results from congenital anomalies, circumstances of pregnancy, or accidents during delivery, and 'exogenous mortality,' which results from external causes. As detailed statistical data about the causes of death are often missing, it was thought to be possible to distinguish these two components by an examination of deaths during the first year of life. An example is provided by J. Bourgeois-Pichat, who designed a graphic method for a biometric analysis of infant mortality at the beginning of the 1950s (Bourgeois-Pichat 1952). His method was based on the following hypothesis: every death which occurs in the last eleven months is exogenous. In any case it may be observed that the share of infant mortality assigned to months one to 11 is practically invariable both in time and in space; it does not depend on the level of infant mortality. Later observations showed that this hypothesis was not always correct. In fact, an approach based strictly on the length of life leads to a more operational result. We may distinguish between:

neonatal mortality, meaning mortality occurring during the first 27 days after birth, which is divided into *early* neonatal mortality for the first week, and *late* neonatal mortality for the three following weeks; and

postneonatal mortality, i.e. mortality occurring between the 28th day and the first birthday.

Moreover, in order to take into account the crucial characteristics of mortality at the time of birth and the difficulty in distinguishing between a child who died just before delivery and a child who died during delivery, the category *perinatal mortality* includes stillbirths and neonatal deaths (or even sometimes only the early neonatal deaths). Stillbirths are outcomes of conceptions presumed viable based on the length of the pregnancy or on the weight or height of the child.

Along with the various components of infant mortality just indicated, there are corresponding rates calculated either only from live births (neonatal and postneonatal mortality rates) or from the total number of births including stillbirths (the perinatal mortality rate).

Table 2 contains the components of infant mortality in most industrialized countries around 1995. (The countries are listed according to an increasing rate of infant mortality.) Early neonatal mortality, i.e., mortality during the first week, is higher than the other components except in Belgium, Moldova, and Romania. In fact, the mortality risks of the first year are mostly concentrated in the first early days of life. More accurate risk calculation, done on a day-per-day basis, shows that the first day in particular has a very high risk. Nearly 30 percent of the average total first-year deaths occur on this single day in countries where the infant mortality is under 10 per 1,000.

Differences between countries are more apparent with postneonatal mortality (a difference by a factor of 10 between Sweden and Romania) than with late neonatal mortality (a factor of 7 between Japan and Latvia). Late neonatal mortality is, in turn, more diverse than early neonatal mortality (from 1 to 5 between Japan or Sweden and Macedonia). However, these differences are smaller among the richest industrialized countries. For instance, between the 15 countries of the European Union, the ratio between neonatal mortality rates is as 1 to 2 between Finland and Greece, and a little more than 1 to 3 between Sweden and Belgium for postneonatal mortality.

The respective shares of neonatal and postneonatal mortality of total infant mortality have changed appreciably during the last few decades. Decrease of infectious mortality, again a result of increased use of antibiotics, brought about a rapid decrease of postneonatal mortality, thereby increasing the share of deaths occurring during the first four weeks of life. Progress in neonatal medicine has led to important success in detection and treatment of diseases of the newborn. That is why it has been possible to save more of these infants or at least keep them alive longer. More and more deaths due to endogenous causes occur after the first four weeks i.e., during the postneonatal period. But at the start of the twenty-first century the share of neonatal deaths is starting to increase again: it is at about 70 percent in the UK, Sweden, and The Netherlands, and it is over 60 percent in France. A mark of new progress can be seen in this increase: children whose lives were only extended by a few weeks or months are now being saved, or at least they increasingly reach the decisive mark of their first birthday.

4. Causes of Infant Deaths

It is difficult to compare causes of infant deaths in industrialized countries even though most of them have detailed statistical data at their disposal. Despite standardization efforts on the part of the World Health Organization, there are still national peculiarities in diagnosis, codification, and tabulation. However, some characteristics do emerge. The overwhelming majority of modern infant deaths are caused by conditions that originated in the perinatal period and by congenital anomalies. These two cause groups account for more than 70 percent of deaths and also have a striking age pattern. Congenital

Table 2
Infant, neonatal, and postneonatal mortality rates around 1995 in industrialized countries (rates per 1,000 live births)

	Mortality rates ……					Mortality rates ……			
	Early neonatal	Late neonatal	Postneonatal	Infant		Early neonatal	Late neonatal	Postneonatal	Infant
Japan	1.5	0.6	1.8	3.9	Portugal	3.6	1.1	2.8	7.5
Finland	2.1	0.6	1.3	4.0	Czech Republic	3.2	1.7	2.8	7.7
Sweden	2.1	0.7	1.2	4.0	USA	4.2	0.9	2.9	8.0
France	2.2	0.7	2.0	4.9	Greece	4.1	1.7	2.4	8.1
Denmark	2.9	0.8	1.3	5.1	Belgium	3.2	1.0	4.0	8.2
Norway	3.0	0.6	1.6	5.2	Hungary	5.5	1.8	3.4	10.7
Germany	2.4	0.8	2.1	5.3	Lithuania	5.6	2.4	4.5	12.5
Austria	2.5	0.9	2.1	5.4	Poland	8.0	2.1	3.5	13.6
Slovenia	2.5	0.6	2.5	5.6	Estonia	7.9	2.3	4.6	14.8
Australia	3.2	0.7	2.0	5.9	Russia	9.3	2.5	6.9	18.6
Spain	2.8	1.1	2.2	6.1	Latvia	8.3	4.1	6.4	18.8
Ireland	3.2	0.8	2.1	6.1	Romania	6.3	3.0	12.0	21.2
Canada	3.4	0.8	1.9	6.1	Moldova	8.4	3.2	9.9	21.5
UK	3.3	0.9	2.0	6.2	Macedonia	10.7	2.8	9.2	22.7
Italy	4.1	1.2	1.7	7.1					

Source: WHO and national statistics
Early neonatal mortality = mortality during the first seven days. Late neonatal mortality = mortality between the seventh and the end of the 27th day. Postneonatal mortality = mortality between the 28th day and the end of the first year. Infant mortality = mortality during the first year.

Table 3
Congenital anomalies and perinatal diseases in France in 1950 and 1994 (rates for 100,000 children less than one year old, both sexes taken together)

	1950	1994
Congenital anomalies	303	134
Of this: congenital anomalies of the circulatory system	127	68
Perinatal diseases including:		
Delivery complications, obstetrical trauma, hypoxy and asphyxy	212	111
Newborn hemolytic disease	58	1
Prematurity, slow development and immaturity	890	20

anomalies and conditions originating in the perinatal period show up in the extreme frailty of the newborn children who die just after birth. Infectious diseases, notably those of the digestion system or the lungs, are only marginal, often under 5 percent. The rate of accidents is also very low, rarely over 3 percent of the total number of deaths, which is evidence of the protection and care afforded newborn children. Finally, sudden infant death syndrome is often responsible for anywhere between 10 and 20 percent of infant deaths, making it the primary cause of postneonatal mortality.

These remarks mainly pertain to countries that have mortality rates of less than 10 per 1,000. In other countries, especially in Central and Eastern Europe, where infant mortality rates are still over 20 per 1,000, infectious diseases have a more prominent place. A more accurate analysis of congenital abnormalities and of conditions originating in the perinatal period can be done with the help of the data for France in Table 3. From 1950 to 1994 congenital anomalies decreased by a little more than half, and conditions originating in the perinatal period decreased by nearly 90 percent. Within this group, mortality linked to complications during labor and delivery or to an obstetric trauma was not even cut in half, while other causes, particularly those linked to prematurity, decreased by 98 percent. Because of this differential development, the hierarchy of the various causes of deaths was modified appreciably, as follows:

(a) The difference between the frequency of deaths resulting from congenital anomalies and deaths due to conditions originating in the perinatal period is decreasing. If this continues, congenital anomalies could become the leading cause of infant mortality.

(b) Of the conditions originating in the perinatal period, the share of deaths linked to circumstances during delivery was raised to about two thirds (compared with only 15 percent in 1950).

These facts suggest that new improvements in the struggle against infant mortality are essentially derived from better detection of congenital anomalies, either permitting treatment 'in utero' or resulting in an abortion, and through new improvements in delivery conditions.

5. *Child Mortality*

Child mortality, that is death between the first and the fifth birthday, is measured by a rate equal to the ratio of the deaths of this age and the average population in the same age range. (This is different from the infant mortality rate, which is obtained by dividing the number of deaths by births.) In most industrialized countries this rate is under 0.5 deaths per 1,000 children of this age; the lowest rate in 1995 was 0.17 per 1,000 in Sweden. Only a few Central and Eastern European countries have a rate above 0.5 per 1,000, particularly Estonia, Moldova, Romania, and Russia, where it is over 1 per 1,000. Child mortality is much lower than infant mortality, as the rate of death between one and five years on average is one twentieth of rates during the first year of life.

Child mortality has dropped everywhere since the end of World War II, but the rate of improvement varies according to the country in the same way as in the case of infant mortality. In Northern Europe the child mortality rates, which had been below 1.5 per 1,000, were reduced between five- and eight-fold; whereas in the Southern European country Portugal child mortality is about one-twenty-fifth of what it was at the beginning of the 1950s.

These improvements reflect a decrease in the total causes of death at this age. For instance, in France all diseases causing childhood deaths were reduced by 90 percent between 1950 and the mid-1990s, and accidents decreased by nearly 70 percent. When one takes this differentiated development into account, the weight of accidents is increasing. In 1950 one in every six childhood deaths was due to an accident; at the end of the twentieth century it was more than one in three.

6. *Factors in Infant Mortality*

There are many factors that influence infant mortality, and they are apparent at every level. The following, simplified categories can be distinguished:

(a) political and socioeconomic factors of a country, which reflect its development, and its financial and organizational ability to supervise sanitation, to promote actions preventing infant mortality, and to create general access to health care, education, employment, and housing;

(b) factors connected with the child's mother, in particular those with a direct effect on embryo development and on children's health: life conditions, general state of health, and potentially harmful

behavior (consumption of tobacco or alcohol, and so on); and

(c) factors connected with the individual child, notably gender, length of gestation, and weight at birth.

In these complex categories, comprised of closely linked elements, it is difficult to isolate decisive factors. Obviously infant mortality depends on the interaction of many elements, such as the wealth of a country, the quality of its health-care system, the woman's marital status, the mother's conditions of life, her state of health, her genetic history, her level of education, her knowledge of child care, the health of the newborn, and so on.

Since we do not have a model for grading these various factors, we look to risk aggravating factors. Generally, infant mortality in industrialized countries is aggravated when the mother is poorly educated, single, inactive, or belongs to a disadvantaged ethnic minority. Therefore infant mortality follows a social gradient, and a newborn's probability of survival may vary according to the social environment in which it was born (Botting and Cooper 1993, Cnattingius and Haglund 1992, Macfarlane 1996).

Infant mortality is higher for boys than for girls. As both sexes are submitted to the same hazards at this age, higher male mortality is essentially biologically determined (Tabutin 1996). Infant mortality also varies considerably according to the weight of the newborn. The mortality of low-weight children (under 2,500 g) and of very low-weight children (under 1,500 g) is considerably higher than the mortality of heavier children. In Sweden, for instance, six per 1,000 newborn children died before their first birthday in 1990, but of the very low-birth-weight babies, 210 per 1,000 died during the first week of life. The relationship between birth weight and death risk deserves more attention, as the proportion of very low-weight children is now increasing in industrialized countries. This is due to two factors. First, the development of fertility treatment has resulted in an increase in the number of multiple births. Second, improvements in prenatal medicine enable children to be born who previously would have died during gestation or delivery.

Offsetting these risk aggravating factors, we may point out some risk attenuating factors. In particular, the number of pregnancies with a high risk factor is becoming less important because of changes in reproductive behavior. Women are postponing having their first children, and there are now very few families with a large number of children. This results in a lower number of pregnancies among very young and relatively older women. In addition, there have been improvements of detection of various malformations or pathologies in utero, and there is the possibility of either carrying out a therapeutic abortion or of treating such problems. Generally speaking, there have been improvements in techniques that allow for the birth and survival of the frailest children, and these tend to reduce the number of infant deaths, especially during the first week of life.

In the final analysis, it seems as if even lower infant mortality is possible in industrialized countries, especially if social inequalities can be reduced. In the more advanced countries expected progress is very restricted, but there is still the possibility of important improvement in Central and Eastern Europe.

See also: Demographic Data Regimes; Infant and Child Mortality in the Less Developed World; Life Expectancy and Adult Mortality in Industrialized Countries; Mortality, Biodemography of; Mortality: The Great Historical Decline

Bibliography

Bourgeois-Pichat J 1952 An analysis of infant mortality. *Population Bulletin of the United Nations* **2**: 1–14

Botting B J, Coooper J 1993 Analysing fertility and infant mortality by mother's social class as defined by occupation. *Population Trends* **74**: 27–33

Cnattingius S, Haglund B 1992 Effects of socioeconomic factors on late fœtal and infant mortality in Sweden. In: *International Collaborative Effort on Perinatal and Infant Mortality*. US Centers for Disease Control/National Center for Health Statistics, Bethesda, MD, Vol. 3, pp. IV38–IV42

Macfarlane A 1996 Inégalités et santé des enfants en Europe: Une perspective méthodologique. In: Masuy-Stroobant G, Gowbin C, Buekens P (eds.) *Santé et Mortalité des Enfants en Europe–Inégalités Sociales d'Hier et d'Aujourd'hui*. Academia Brylant, Louvain-la-Neuve, Belgium

Masuy-Stroobant G 1996 *Santé et Mortalité des Enfants en Europe–Inégalités Sociales d'Hier et d'Aujourd'hui*. Academia Brylant, Louvain-la-Neuve, Belgium

Masuy-Stroobant G 1997 Infant health and infant mortality in Europe: lessons from the past and challenges for the future. In: Corsini A, Viazzo P (eds.) *The Decline of Infant and Child Mortality–The European Experience: 1750–1990*. Martinus Nijhoff, The Hague, The Netherlands

Tabutin D, Willems M 1996 La surmortalité des filles jusqu'en 1940: un bel exemple des inégalités sexuelles dans l'histoire occidentale. In: Masuy-Stroobant et al. (eds.) *Santé et mortalité des enfants en Europe–Inégalités sociales d'hier et d'aujourd'hui*. Academia Brylant, Louvain-la-Neuve, Belgium

A. Monnier

Infant and Child Mortality in the Less Developed World

This year, more than 10 million children below five years of age have died, or will die around the world, over 90 percent of them in developing countries. This death toll represents about one child death every three seconds. The vast majority of these children are dying

from diseases preventable or treatable with simple and low-cost medical techniques. Tremendous progress has, however, been accomplished since the 1950s. The proportion of children deceased below age five declined from about 200 per thousand births around 1950 to 120 per thousand in 1980–5 and 80 in 1990–5. Improvement in health care and sanitation has resulted in the survival of 15 million children who would have died under the conditions of the 1950s.

1. *Mortality Measures*

Numbers of child deaths are rough estimates rather than exact measures because accurate data on child health and mortality in developing countries are not easy to come by. The most obvious measure is infant mortality, but only one-quarter of the world's population lives in countries where death recording is complete, and these are countries with the lowest mortality. However, major data collection efforts have been undertaken since the 1980s and techniques for estimating child mortality have improved so remarkably that it is now possible to make reliable comparisons among all the world's regions and to better understand disease processes in developing countries.

Two measures are traditionally used by demographers to characterize child mortality: the infant mortality rate and the under-five mortality rate. The infant mortality rate is the ratio of deaths below age one during a given calendar year to the number of live births during that same calendar year. The under-five mortality rate is the proportion of children who die before reaching their fifth birthday in a one-year cohort of live births. Both rates are usually provided per thousand births.

2. *Geographic Disparities*

The mortality of young children has declined everywhere around the world. The favorable trend observed at the regional level conceals, however, a wide diversity of situations (Fig. 1). The 50 least developed countries still experience an under-five mortality rate above 150 per thousand and differences between regions and among countries in each continent are increasing continuously.

Latin America has been the first developing region to benefit from medical progress. The decline started in the early decades of the twentieth century and has been continuous since then. Latin America is also the most homogeneous region of all. The overall probability of dying below five is around 50 per thousand. The two poorest countries, namely Haiti and Bolivia, still experience rates of 100 per thousand, but in many countries children survival chances are now close to those observed in Europe or North America (see *Infant and Child Mortality in Industrialized Countries*). This is particularly true in the Caribbean where under-five mortality rates around 10 per thousand are not uncommon.

Asia exhibits the largest internal diversity. The average probability of dying below five is about 75 per thousand but it ranges from 45 per thousand in Eastern Asia to 100 in South-Central Asia, a ratio of more than two to one. Variations are even wider among countries within the same subregion on the continent. To give but a few examples, the infant mortality rate reaches 130 per thousand in Iraq and 100 per thousand in Yemen, more than 10 times the level recorded in Cyprus and Israel, where it is now below 10. Similarly, in Southeast Asia, it is about 115 per thousand in Lao and Cambodia, compared to 10 per thousand in Brunei and only 6 in Singapore. But the overall level of mortality in Asia is largely driven by the two most populated countries of the planet, namely China and India, which experience rates of infant mortality of 45 and 80 per thousand, respectively, and an under-five mortality of 50 and 100 per thousand.

Africa remains by far the region with the highest level of infant and child mortality. 10 percent of all children die during their first year and nearly 15 percent before reaching their fifth birthday. Wide disparities are observed between Northern and Southern Africa on the one hand, where children below age five experience a probability of dying around 80 per thousand, and other regions on the other hand, where the probability reaches over 150 per thousand. Several African countries exhibit extremely high levels of under-five mortality: in Guinea, Guinea-Bissau, Malawi, Mali, Niger, Rwanda, Sierra Leone, and The Gambia, more than 20 percent of all children die before their fifth birthday. However, intraregional variations are smaller than in Asia because few countries have reached a level of infant and child mortality below 50 per thousand.

Intraregional disparities are the result of diverging trends much more than variations in the initial levels of mortality. Indeed, historical research suggests that the infant mortality rate was well above 200 per thousand in preindustrial Europe as well as in developing countries prior to the twentieth century (see *Mortality: The Great Historical Decline*). By 1950, Latin America already presented an average level of mortality significantly below the pretransitional level, by contrast with Asia and Africa. During the following decades, while mortality continued to decline in the former region, the other two main continents of the developing world initiated a significant improvement in the level of infant and child mortality (Fig. 2). In Asia, however, the decline was much faster than in Africa during the period 1955–75. While both regions exhibited similar levels of infant mortality around 1950 (about 180 per thousand), the proportion of children dying during their first year fell by 60 percent

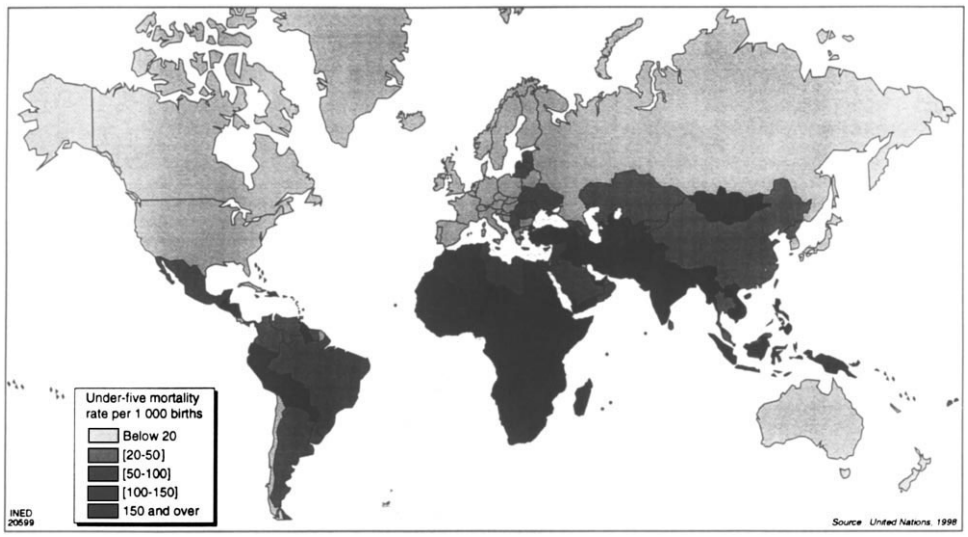

Figure 1
Under-five mortality rate in 1990–5

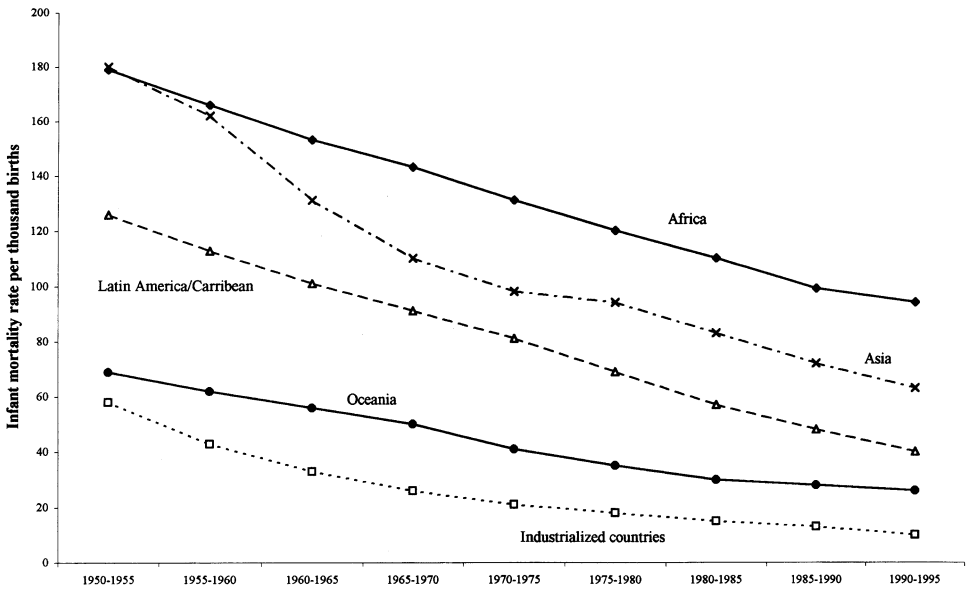

Source: United Nations, World Population Prospects. The 1998 Revision, Volume I, Table A.27.

Figure 2
Infant mortality rate by region: 1950–5 to 1990–5

in Asia, from 160 per thousand to 100 per thousand in 15 years. During the same period, the proportional decline reached 30 percent in Latin America and only 25 percent in Africa.

3. *Causes of Infant and Child Deaths*

As socioeconomic development proceeds, countries move through the epidemiological transition. The

concept was invented by Omran (1971) to describe the shift in disease composition, from infectious and parasitic diseases to congenital and degenerative diseases (see *Mortality, Epidemiological, and Health Transitions*). This trend parallels a change in the age structure of mortality, with an increase in the average age at death. More than 60 percent of all deaths are experienced by children below five in high mortality populations compared to 3 percent in industrialized nations.

3.1 The Predominance of Infectious and Parasitic Diseases

Most of these deaths are attributable to a small number of pathologies. Five diseases, or disease categories, are responsible for over 70 percent of all deaths under age five. The highest toll is due to acute respiratory infections, including pneumonia, the most common cause of illness and death among the children of the Third World. These diseases cause more than three million deaths a year among children below age five. Diarrheal diseases are a close second, with almost one in four of all child deaths. Measles continue to kill more than one million children annually and tetanus, another preventable disease, about half a million, about as much as malaria, to which more than 600,000 deaths are attributed. Most of these diseases also represent a high proportion of morbidity in industrialized countries. They are, however, much less lethal there because of the natural and induced resistance of the immune system of children born in these countries. Malnutrition is a major factor in the weakening of natural resistance, while immunizations are fundamental to the building of artificial resistance.

3.2 The Role of Malnutrition

The major role of nutrition in child mortality processes is now well documented (see *Health and Mortality, Body Stature as a Measure of*). The effect of malnutrition on mortality was overlooked for a long time because malnutrition hardly ever kills directly. Less than 200,000 deaths of children below age five are attributed to either marasmus or kwashiorkor, the most acute forms of malnutrition. However, recent research has demonstrated that malnutrition is an initial or aggravating factor in about 55 percent of all child deaths. Malnutrition is usually the result of a combination of inadequate dietary intake and infection. It affects over one-third of all the children under age five in the developing world.

There exist different kinds of malnutrition. Malnutrition can not only result from a deficiency of protein and energy but also from a lack of vital minerals and vitamins in children's diets. The 1990 World Summit for Children, which was organized under the auspices of the UNICED, singled out deficiencies of three micronutrients (iron, iodine, and vitamin A) as being particularly common and of special concern for children in developing countries. Zinc has been recently added to the list. The general effect of malnutrition is to lower the defenses in children's immune systems, resulting in a greater incidence, severity, and duration of common childhood diseases. Vitamin A deficiency makes children especially vulnerable to infections and worsens the course of many infectious diseases. Death rates in children lacking vitamin A are 20–30 percent higher than in other children. Vitamin A deficiency has been shown more particularly to reduce children's resistance to diarrhea and measles. Not only does malnutrition increase the probability of infectious and parasitic diseases, but diseases also speed nutrient loss and suppress appetite, provoking a vicious circle ultimately leading to death. For instance, in parts of Africa where malaria is common, studies show that about one-third of the cases of child malnutrition are caused by malaria.

In an attempt to reduce the incidence of malnutrition in the earlier months of life, international organizations such as the World Health Organization (WHO) and UNICEF have been actively promoting breastfeeding. Though breastfeeding durations have declined significantly since the 1980s due to the combined effects of commercial pressure and the increased participation of women in the labor force, such efforts have proved useful in encouraging breastfeeding in the poorest countries. Breastmilk contains all the nutrients, antibodies, hormones, and antioxidants an infant needs to thrive during the first six months of life. Breastfed children not only show better immune responses to immunizations, but they also exhibit natural resistance against a number of infectious diseases, gastrointestinal and respiratory in particular. By comparison, commercial infant formulas are an expensive and inferior substitute for breastfeeding. They are frequently overdiluted in order to save money and mixed with unsafe water before being fed to the child in unsterilized bottles capped with an unclean teat. Research indicates that in countries where infant mortality rates are high or moderately high, a bottle-fed baby in a poor community is 14 times more likely to die from diarrheal diseases and four times more likely to die from pneumonia than a baby who is exclusively breastfed.

3.3 Old Curses and New Threats

Old curses have recently intensified. Epidemics of plague and upsurges in cholera have demonstrated a surprising vitality of these diseases in parts of Africa, Asia, and Latin America. Malaria has also showed frustrating resistance to the many approaches tried to combat the disease. The parasite is no longer vul-

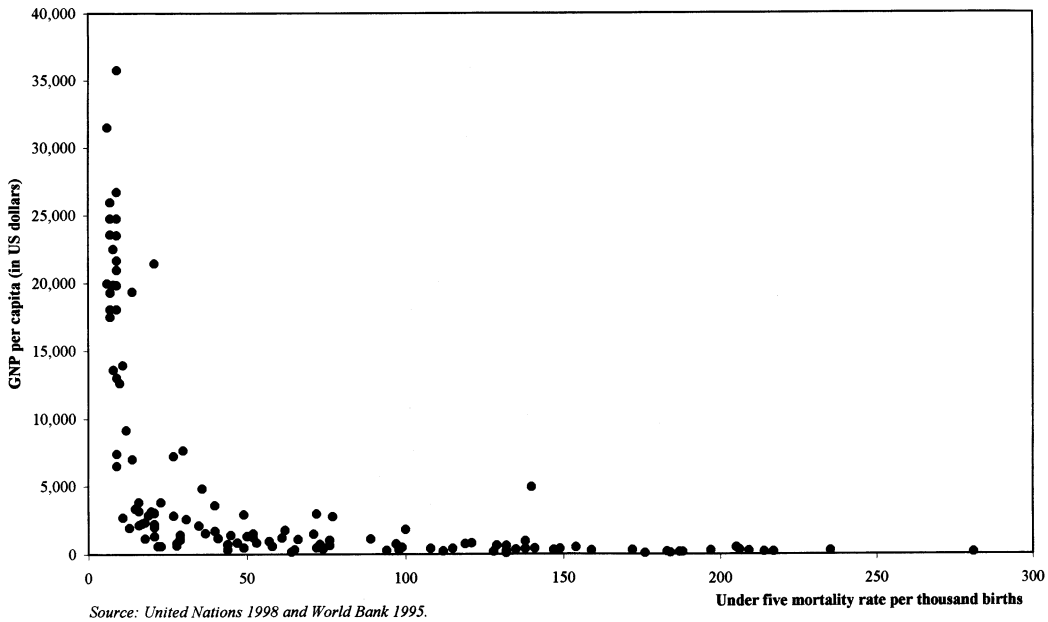

Figure 3
Relationship between the probability of dying below five and national income per capita

nerable to some of the most powerful antimalarial compounds available. For a time, the widespread use of pesticides reduced the incidence of the anopheles mosquito transmitting the parasite, but it, too, has developed resistance. In addition to such historically deadly diseases, new threats to the health of children have emerged, the most dramatic of which is undoubtedly AIDS.

AIDS is becoming a major killer of young children in parts of the developing world (see *Mortality and the HIV/AIDS Epidemic*). The AIDS epidemic is now a main public health issue in Africa and in parts of Asia, threatening millions of children with premature death. Its real impact is very difficult to measure because of a lack of statistical data and problems of diagnosis in the countries most affected by the epidemic. The burden of the disease has been steadily increasing for children since the early 1980s. According to recent estimates from the Joint United Nations Program on HIV/AIDS (UNAIDS) and the WHO, about 500,000 children below age 15 are now dying each year from the disease, and nearly 600,000 are newly infected. The majority of them acquire the infection directly from their mothers. Mother-to-child transmission of HIV, also known as vertical transmission, occurs during three segments of life: during pregnancy, during labor and delivery, and through breastfeeding. Recent research tends to indicate that a child stands the greatest risk of vertical transmission (about 20 percent) during the time of late pregnancy and childbirth. There is an additional 15 percent risk that an infant will become infected through breastmilk. Consequently, more than one in three of the babies born to mothers infected with HIV will themselves be infected. Most will die before the age of five. The cumulative number of children infected with HIV/AIDS reached more than six million by the end of the twentieth century, and the mortality of young children has been increasing up to 50 percent in countries most affected by the epidemic, mainly located in Eastern and Southern Africa such as Namibia or South Africa. In Zimbabwe, AIDS has already become the biggest single killer of the nation's under-fives.

In addition to its direct impact, AIDS also indirectly affects the health of children in developing countries. AIDS is an expensive disease. In communities with limited resources, the epidemic diverts a tremendous proportion of the available budget away from basic health care towards intensive care. In countries with mature epidemics, such as Uganda, Zambia, and Zimbabwe, some doctors report that three-quarters of the beds in hospital pediatric wards are occupied by children ill from AIDS. Another indirect effect of the epidemic is the dramatic increase in orphanhood. Nearly 10 million children have been orphaned in the 1990s because of the disease, more than 95 percent of them in sub-Saharan Africa. Because of the very slow changes in sexual behavior, the prohibitive cost of existing treatments, and the difficulty of developing an effective vaccine, the deadly impact of the epidemic is

expected to increase during the next decade in the most affected countries.

4. *Economic Determinants*

Poverty is viewed as the underlying factor of high child mortality in developing countries. Obviously, economic deprivation at the family level affects the health of children. A low purchasing power is related to poorer nutritional diets, the lack of ready access to a safe water supply and proper sanitation, as well as unhygienic conditions in and around homes. Such conditions are typically associated with a high prevalence of childhood diarrhea and they have significant implications for the spread of many infectious and parasitic diseases.

However, there is an unclear relationship between national income and the level of infant and child mortality. The wealthier countries are clearly those in which children are the least likely to die. As indicated by Fig. 3, countries with a GNP per capita above $5,000 have a level of infant mortality below 25 per 1,000. Yet, the relationship does not appear to be systematic to the extent that those countries faring the worse in terms of health and mortality are not always the poorest. There is a wide variety of situations among poor countries with similar levels of national income per capita. To provide but a couple of examples, while GNP per capita in the Lao PDR is significantly higher than in Vietnam ($370 vs. $240), infant mortality is half as high in the latter, reaching 50 per 1,000 compared to more than 100 per 1,000 in Lao. The contrast is even more striking in sub-Saharan Africa where the ratio of infant mortality reaches three to one in countries with the same level of GNP per capita ($500–550), such as Zimbabwe and Guinea, where the infant mortality rate reaches 50 and 140 per 1,000, respectively. Poverty *per se* is thus insufficient to explain geographic differentials in infant and child mortality across the world and other factors must be accounted for.

5. *Political Commitment*

Experience and research demonstrate that without a commitment at the government level, the health of children improves very slowly. Poor countries experiencing a low level of mortality relative to their level of socioeconomic development are those where a primary health care strategy has been implemented early and steadily at the national level, following recommendations first developed at the 1978 World Health Organization conference in Alma-Ata. The strategy associates direct medical intervention, such as immunizations and oral rehydration therapy, to nutrition programs, and to activities indirectly affecting health, such as hygiene, education, and family planning campaigns. China, Cuba, Vietnam, and Costa Rica are well-known examples of poor countries where the primary health care strategy has been successfully implemented. Even countries that have not been able to implement the strategy as a whole but only in part have benefited from a decline in child mortality.

In many regions of the world, the international community has played an important role in promoting primary health care services. WHO, UNICEF, and many other public or private organizations have worked with governments towards the goal of immunizing the world's children against six preventable diseases. Due to very active campaigns carried out during the 1980s, rates of child immunization against measles, tetanus, poliomyelitis, tuberculosis, pertussis (whooping cough), and diphtheria have risen from about 10 percent in the mid-1970s to 30 percent 10 years later and 80 percent around 1990. This achievement has meant the prevention of an estimated three million child deaths a year. However, the proportion of children vaccinated has leveled out since the early 1990s. In some of the poorest countries of Africa, immunization rates have even declined by about 10 percent.

In conclusion, dramatic progresses have been accomplished since the 1970s, so that children in many Asian and Latin American countries do not experience a higher risk of dying than those born in industrialized regions. There remain sources of serious concern for the future of child health in the most deprived areas of the Third World, however, and inequalities continue to increase among subgroups of national populations and among countries. The challenge is now to circumvent economic and political difficulties in order to implement the large-scale health interventions known to effectively reduce infant and child mortality.

See also: Adult Mortality in the Less Developed World

Bibliography

Omran A R 1971 The epidemiologic transition : a theory of the epidemiology of population change. *Milbank Memorial Fund Quarterly* **49**(4): 509–38

UNAIDS Joint United Nations Programme on HIV/AIDS 1998 *AIDS epidemic update: December 1998*. UNAIDS, New York

UNICEF 1998 *The State of the World's Children 1998*. Oxford University Press, Oxford, UK

United Nations 1998 *World Population Prospects: The 1998 Revision*, Volume I. Comprehensive Tables. Department of Economic and Social Affairs, Population Division. United Nations, New York

World Bank 1995 *World Development Report 1995*. Oxford University Press, Oxford, UK

World Health Organization 1999 *The World Health Report 1999. Making a Difference*. World Health Organization, Geneva, Switzerland

M. Barbieri

Infant Development: Physical and Social Cognition

At the beginning of the twentieth century the received wisdom among psychologists was that human infants have virtually none of the cognitive skills they will later have as adults. Sigmund Freud had infants dreaming away their days in primary process (autistic) thinking, and William James described their phenomenal world as 'a blooming, buzzing confusion.' But since that time dramatic progress has been made in our understanding of infants' minds, and we now know that many of the basic structures of human cognition develop very early in infancy.

The modern era was ushered in primarily by the work of Jean Piaget. In a series of exquisitely detailed observations on his own three infants' ontogenies—first made in the 1930s but only made available to the international community in the 1950s—Piaget documented that from birth infants are actively and continuously exploring their environments, and that their cognition is already organized. On the basis of these observations, Piaget proposed a sequence of six cognitive stages, from birth to 18 months of age, that structured infants' interactions with the world. These stages were wholly general and governed the way infants could physically interact with the environment. He posited six levels of sensorimotor schemes (Piaget 1952), and for the kinds of knowledge infants could construct from these interactions, he posited six corresponding levels of knowledge of the physical world of space, time, causality, and objects (Piaget 1954).

Piaget's most general claim was that human infants are cognitive beings, it is just that their cognition is confined to the here-and-now of sensorimotor interactions with the world. Thus, for example, for very young infants objects only exist when they are interacting with them, and space is only experienced as self and other movement. The engine of Piaget's theory—the mechanism that takes infants beyond the sensorimotor level and on to more adult-like ways of conceptualizing the world—was action in the concrete sense of manipulation of objects. In Piaget's theory, neonates begin with their five senses totally uncoordinated and with objects as ephemeral perceptions only. But as they act on objects and perceive the effects of particular actions through the various different senses simultaneously, the senses become coordinated and so objects become objectified. Only by the end of the infancy period—after a vast number of sensorimotor interactions with the world—can infants cognitively and symbolically represent external entities in the absence of their own immediate perceptual contact with them.

1. *Beyond Piaget*

Although it gave infants credit for much more complex cognitive processes than its predecessors, Piaget's theory turned out to be still too conservative. The main problem was that Piaget depended almost exclusively on evidence provided by infants' spontaneously produced overt actions, and human infants are indeed very slow to develop motorically. But in the 1970s new methods were developed that enabled researchers to probe and test infants' cognitive competencies without relying on the production of complex motor sequences. These methods were much more experimentally rigorous than Piaget's, and exploited those few actions that infants are very skilled at and engage in frequently; for example, looking, sucking, kicking, and reaching.

The first of Piaget's proposals to fall was the proposal that infants possess five distinct and uncoordinated sensory modalities. In a seminal study, Spelke (1976) had young infants view two different films on adjacent monitors and at the same time listen to an audio recording that matched one of the two visual events (e.g., there was a noise every time an object banged the ground on one of the films). Infants as young as four months of age looked longer at the film that matched the audiorecording, suggesting at least some form of intermodal integration. At around the same time, Meltzoff and Moore (1977) discovered that newborn infants (less than 48 hours old) reliably imitated adult behaviors directed to them; such things as tongue protrusions and mouth openings. Given that infants perceive the adult face visually but perceive their own facial movements only proprioceptively, skills of intermodal integration were once again implied.

These two pioneering studies demonstrated that Piaget was wrong about early intermodal integration, but they did much more than that. First, if infants coordinate information from the different sensory modalities practically from birth, it is possible that this cross-modal information gives them a more objectified understanding of objects at an earlier age as well. These studies thus initiated a plethora of new research into infants' early understanding of the world of space and objects. Second, these studies and others that followed them demonstrated that infants possess considerable knowledge about the world before they can actively manipulate objects or anything else in their environments. Thus, Piaget's focus on action as the mechanism of development was called into serious question. Third, Meltzoff and Moore's study of

neonatal imitation in particular opened up exciting new questions about infants' early emerging skills for dealing with the social world—an arena of development about which Piaget had very little to say. Finally, and perhaps most importantly, Spelke's study demonstrated that infants' looking preferences could be used to ask very specific and very important questions about their cognitive skills. This new, child-friendly technique (along with the related technique of visual habituation–dishabituation) enabled researchers to answer questions about infant minds that Piaget never thought it possible to ask.

Buoyed by these spectacular new findings and powerful new methodologies, researchers in the 1980s and 1990s opened up a number of new and exciting windows on infants' cognitive development. These may be grouped, roughly, into findings about: (a) perception and learning, (b) understanding of the physical world, and (c) understanding of the social world.

2. *Early Perception and Learning*

Infants' perception of the world is surprisingly adult-like from a surprisingly early age. This general fact has been established mostly with habituation and paired-comparison techniques. In the typical habituation paradigm, infants are repeatedly exposed to a stimulus until their visual attention to the stimulus decreases to a pre-established level, essentially until they become bored with the stimulus. During a test phase, the infant is then shown a novel stimulus and the familiar stimulus simultaneously (i.e., paired comparison method) or one after the other (habituation method) and visual attention is again measured. If infants look longer to the novel stimulus, but continue to look away from the familiar stimulus, it is inferred that they can discriminate between the two stimuli. In such habituation studies, young infants discriminate among various properties of stimuli such as color, size, and shape (see Johnson 1998), perceive voices and sounds (e.g., Kuhl et al. 1992, Pegg et al. 1992), and even particular odors (Marlier et al. 1998). Other research has shown that much of this performance is based on cross-modal information. For instance, newborns manifest a visual preference for a novel image that corresponds to the shape or texture of an object that was previously orally explored (i.e., Meltzoff and Borton 1979, Gibson and Walker 1984). In addition, they reliably turn their head and eyes in the direction of a sound source, meaning that spatial location from this early age is specified by both auditory and visual information (Muir and Clifton 1985). By four months of age infants reach for an object that they see in the light, but will also reach when they only hear the object making a sound in the dark (Clifton et al. 1993).

Infants also are able to learn new things quite rapidly from soon after birth. Using instrumental learning paradigms, infants are reinforced for producing a particular type of behavioral response such as kicking their legs or sucking at a particular rhythm or speed. Infants readily learn such associations and are generally eager to modulate their actions to produce interesting outcomes. In one of the first studies to use such a paradigm, 4–12-month olds sucked on a dummy nipple to turn on a visual light display. With only three minutes of experience with such response–stimulus conjugation, infants significantly increased their rates of sucking compared to infants whose sucking produced no contingent outcome (e.g., Siqueland and DeLucia 1969, Kalnins and Bruner 1973, Lewis et al. 1985). In addition, newborn infants learn to modulate their actions to see an image of their mother as opposed to a stranger (Walton et al. 1992), to hear their mother's familiar voice or a story that was repeatedly read to them while in utero (DeCasper and Fifer 1980), and even to hear speech sounds from their native language (Eimas et al. 1971). Indeed, eight-month-old infants are so skilled at learning that they are able to pick up statistical patterns in their perceptual input with two minutes or less of exposure to a disembodied voice, played while the infant is playing on the floor and seemingly paying little attention (e.g., Saffran et al. 1996). Infants can extract these same kinds of patterns from other kinds of perceptual inputs as well; for example, sequences of arbitrary tones or even lights (Saffran et al. 1999).

3. *Understanding the Physical World*

Going a step beyond the perceptual abilities of infants, researchers have also designed paradigms to assess infants' physical knowledge at an age before they can even reach for or manipulate objects. Using one or another variation of the habituation technique, infants typically are familiarized with some event as they watch it several times; for instance, seeing one object pass behind a screen. Then during a testing phase, the infant sees the same event, but this time the screen is removed and the infant sees one of several possibilities. For example, they might see the original object in an improbable position, two objects, perhaps a different object in the position, or no object at all. As young as three months, infants look reliably longer at improbable compared to probable outcomes. From these kinds of studies, researchers have found that infants as young as two or three months possess much physical knowledge of the world and readily engage in physical reasoning about objects, relying upon principles such as object solidity, object continuity, the notion that two objects cannot be in the same location at one time, and the notion that objects exist continuously in space (see Spelke 1994, Spelke et al. 1992).

Similar techniques have also been used to assess infants' object concept more directly. In one classic study, commonly referred to as the Draw Bridge Study

(Baillargeon et al. 1985), infants were familiarized to a screen rotating alternately from upright to flat on a table. Following this familiarization phase, infants saw a solid object that was placed behind the screen, making it impossible for the screen to rotate flat onto the table. During a test phase, they saw either a possible event—the screen stopped moving at the position where the solid object would have obstructed its motion—or an impossible event—the screen rotated flat to the table, seemingly passing through the solid object. Infants as young as three months of age—before they are reaching for, grasping, or manipulating objects—looked reliably longer at the impossible event. These findings suggest that infants have a rudimentary concept of objects that enables them to reason about and predict their basic physical transformations. In a related series of studies, Wynn (1992) has shown that infants are surprised when they see one object disappear behind a screen and then, when the screen is removed, two objects are there (or they see two objects behind a screen but later only one is there). These studies evidence some basic understandings of quantities of objects as well.

Habituation paradigms have more recently been extended to assess the ability of young infants to relate objects to one another, to categorize them, and to recall information about them. In a study of infants' understanding of basic causal relations, six-month olds were habituated to either a noncausal event in which two objects moved independently or to a causal event in which one object collided into the other. Following habituation, infants then observed a reversal of the event. Infants who were habituated to the causal event looked longer when the causal event was reversed. In contrast, those infants who had watched the noncausal event did not seem to notice the role reversal (Leslie 1984). Dishabituation to role reversal only in the causal sequence suggests that infants detected the relations among these objects in the original events. Of course, the question of whether infants are endowed with an understanding of causal relations from birth or whether such understanding develops in a more progressive manner remains open (e.g., Oakes and Cohen 1990). In any case, such findings point to the precocious inclination of infants to perceive the relations and potential meaning among objects and events in the world.

In studies of categorization, infants are pre-exposed to one or more members of a particular category. They are habituated to these members and then shown a novel stimulus that does not belong to the category. As with traditional habituation paradigms, it is assumed that infants classify the stimulus as different from the familiarized category if they look longer to the new test stimulus as compared with a within category exemplar. Impressively, at least by the middle of the first year, infants come to represent a variety of categories such as spatial relations, geometric patterns, faces, and animals (e.g., Eimas and Quinn 1994, Quinn and Eimas 1987, Roberts 1988, Sherman 1985). Other common paradigms used to investigate infants' ability to categorize information have relied upon sequential touching or object manipulation. The sequential touching paradigm exploits infants' inclination to touch and explore objects, and is considered analogous to an adult-like object sorting behaviour. By 9–12 months infants simultaneously touch the objects from one category before touching objects from the other category and differentially categorize objects from global and basic-level categories (Mandler and McDonough 1993, Sugarman 1981).

In terms of early memory, using conjugate reinforcement paradigms in which they are reinforced to kick their legs in order to learn about the features of objects hanging from a mobile, young infants readily learn and recall experiences and retain this information over several days and sometimes weeks (i.e., Rovee-Collier 1995).

4. *Understanding the Social World*

Modern research has found that young infants also exercise impressive cognitive skills in understanding their social world. Soon after Meltzoff and Moore (1977) found that young infants would match some adult behaviors, Trevarthen (1979) and others discovered that they also interact with adults reciprocally in patterned ways. The first research of this type simply noted that young infants and their mothers interact with one another in a kind of turn-taking sequence, sometimes called a protoconversation (see Bullowa 1979 for a review). This discovery was followed by more rigorously controlled investigations.

Most important is the so-called still-face paradigm. In the traditional paradigm, infants engage for several minutes in a normal face-to-face interaction with an adult social partner. Then dyadic interplay is halted when the adult suddenly adopts and holds a neutral still face for about one to two minutes. Infants as young as two months respond to such perturbation with increased negative affect and gaze aversion (Tronick et al. 1978). This reaction to the still face is interpreted as the expression of social expectations by the infant and the sense of a disruption of positive coregulation (Muir and Hains 1993). Other evidence derives from the interpersonal contingency paradigm designed by Murray and Trevarthen (1985), in which infants interact with their mothers over a television monitor. In this study, infants sometimes watched their mothers interacting with them live, and at other times watched a replay of their mothers' behavior. At two months of age, infants reacted more positively to their mother interacting with them 'live' compared to watching a replay video.

The development of social expectations and the social behavioral repertoire of human infants develops rapidly over the course of the first year. It is also

important to note that subsequent work has established a number of compelling phenomena concerning neonatal imitation; for example, the fact that six-week-old infants will seek to reproduce even extremely novel behaviors (Meltzoff and Moore 1994). Nevertheless, there are still some investigators who are not convinced that what infants are doing in these studies is truly imitation, rather than something much less cognitively sophisticated such as preparing for object exploration (e.g., Anisfeld 1991, Jones 1996).

At around 9 to 12 months of age infants begin to engage in a new type of social interaction. Specifically, they begin to engage in interactions that are triadic in the sense that they involve the referential triangle of child, adult, and some outside entity to which they share attention. Thus, infants at this age begin to flexibly and reliably look where adults are looking (gaze following; Scaife and Bruner 1975), use adults as social reference points (social referencing; Sorce et al. 1985), and act on objects in the way adults are acting on them (imitative learning; Meltzoff 1988)—in short, to 'tune in' to the attention and behavior of adults toward outside entities. At this same age, infants also begin to use communicative gestures to direct adult attention and behavior to outside entities in which they are interested—in short, to get the adult to 'tune in' to them (Bates 1979). This revolution in the way infants relate to their worlds begins when infants understand other persons as intentional agents like the self who have a perspective on the world that can be followed into, directed, and shared (Tomasello 1995, Carpenter et al. 1998)—and it presages the emergence of symbolic language, the ultimate tool of joint attention (Tomasello 1999).

5. *Conclusion and Prospects*

The overall picture is thus of an infant much more cognitively competent than James or Piaget ever imagined. From the first few months of life, in many cases before they even reach for or manipulate objects, human infants perceive a world of stable objects, and they soon can categorize, quantify, and perceive causal relations among these objects in some fundamental ways. They also are especially attuned to other persons and interact with, and even imitate, them in special ways.

It is not the case, however, that all researchers agree totally on this picture. Of most immediate concern, some researchers have questioned the interpretation of studies based on preferential looking or habituation. The question is not about the findings themselves, but rather about whether they might be more parsimoniously interpreted as evidence of perceptual strategies and expectations, not high-level cognition (Bogartz et al. 1997, Haith 1998). Thus, while looking time measures indicate that four-month-olds show sensitivity to the solidity of physical objects (i.e., Spelke et al. 1992), if children are asked to physically interact with the exact same experimental arrangement they do not evidence the same sensitivity to solidity until two to three years of age (Hood et al. 2000). A major challenge for future research in infant cognition is thus to determine whether such findings simply reflect different extraneous task demands placed on relatively fragile cognitive beings or whether, instead, they reflect different processes of human cognition. Perhaps relevant to this question is recent research into the cognitive processes of nonhuman primates, who show many—but not all—of the same cognitive skills as human infants (Tomasello and Call 1997).

In all, recent research in infant cognition has been among the most exciting in all of the behavioral and cognitive sciences, unearthing on a regular basis infant competencies that few would have expected just a few decades ago. Research in the coming decades will be aimed both at making more such discoveries and also at determining what these discoveries mean in the larger context of human evolution and ontogeny.

See also: Brain Development, Ontogenetic Neurobiology of; Cognitive Development: Child Education; Cognitive Development in Childhood and Adolescence; Cognitive Development in Infancy: Neural Mechanisms; Infancy and Childhood: Emotional Development; Infant and Child Development, Theories of; Infant Education; Memory Development in Children; Piaget's Theory of Child Development; Prenatal and Infant Development: Overview; Social Cognition in Childhood; Socialization in Infancy and Childhood; Visual Development: Infant

Bibliography

Anisfeld M 1991 Neonatal imitation. *Developmental Review* **11**: 60–97

Baillargeon R, Spelke E S, Wasserman S 1985 Object permanence in five-month-old infants. *Cognition* **20**: 191–208

Bates E 1979 *The Emergence of Symbols: Cognition and Communication in Infancy*. Academic Press, New York

Bogartz R S, Shinskey J L, Speaker C J 1997 Interpreting infant looking: The event set by event set design. *Developmental Psychology* **33**(3): 408–22

Bullowa M 1979 *Before Speech*. Cambridge University Press, New York

Carpenter M, Nagell K, Tomasello M 1998 Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development* **63**(4, Serial No. 255)

Clifton R K, Muir D W, Ashmead D H, Clarkson M G 1993 Is visually guided reaching in early infancy a myth? *Child Development* **64**: 1099–110

DeCasper A J, Fifer W P 1980 Of human bonding: Newborns prefer their mothers' voices. *Science* **208**: 1174–6

Eimas P D, Quinn P C 1994 Studies on the formation of perceptually based basic-level categories in young infants. *Child Development* **65**: 903–17

Eimas P D, Siqueland E R, Jusczyk P, Vigorito J 1971 Speech perception in infants. *Science* **171**: 303–6

Gibson E J, Walker A S 1984 Development of knowledge of visual–tactual affordances of substance. *Child Development* **55**: 453–60
Haith M M 1998 Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior & Development* **21**: 167–79
Hood B, Carey S, Prasada S 2000 Predicting the outcome of physical events: Two year olds fail to reveal knowledge of solidity and support. *Child Development* **71**: 1540–54
Johnson S 1998 Object perception and object knowledge in young infants: A view from studies of visual development. In: Slater A (ed.) *Perceptual Development: Visual, Auditory, and Speech Perception in Infancy*. Psychology Press, Hove, UK, pp. 211–41
Jones S S 1996 Imitation or exploration? Young infants' matching of adults' oral gestures. *Child Development* **67**: 1952–69
Kalnins I V, Bruner J S 1973 The coordination of visual observation and instrumental behavior in early infancy. *Perception* **2**: 307–14
Kuhl P K, Williams K A, Lacerda F, Stevens K N, Lindblom B 1992 Linguistic experiences alter phonetic perception in infants by 6 months of age. *Science* **255**: 606–8
Leslie A M 1984 Spatiotemporal continuity and the perception of causality in infants. *Perception* **13**: 287–305
Lewis M, Sullivan M W, Brooks-Gunn J 1985 Emotional behaviour during the learning of a contingency in early infancy. *British Journal of Developmental Psychology* **3**: 307–16
Mandler J M, McDonough L 1993 Concept formation in infancy. *Cognitive Development* **8**: 291–318
Marlier L, Schaal B, Soussignan R 1998 Neonatal responsiveness to the odor of amniotic and lacteal fluids: A test of perinatal chemosensory continuity. *Child Development* **69**: 611–23
Meltzoff A N 1988 Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology* **24**: 470–6
Meltzoff A N, Borton R W 1979 Intermodal matching by human neonates. *Nature* **282**: 403–4
Meltzoff A N, Moore M K 1977 Imitation of facial and manual gestures by human neonates. *Science* **198**: 75–8
Meltzoff A N, Moore M K 1994 Imitation, memory, and the representations of persons. *Infant Behavior and Development* **17**: 83–99
Muir D, Clifton R 1985 Infants' orientation to the location of sound sources. In: Gottlieb G, Krasnegor N (eds.) *Measurement of Audition and Vision in the First Year of Postnatal Life: A Methodological Overview*. Ablex, Norwood, NJ, pp. 171–94
Muir D, Hains S M 1993 Infant sensitivity to perturbations in adult facial, vocal, tactile, and contingent stimulation during face-to-face interactions. In: de Boysson-Bardies B (ed.) *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*. Elsevier, Amsterdam, pp. 171–83
Murray L, Trevarthen C 1985 Emotional regulation of interactions between two-month-olds and their mothers. In: Field T M, Fox N A (eds.) *Social Perception in Infants*. Ablex, Norwood, NJ, pp. 177–97
Oakes L M, Cohen L B 1990 Infant perception of a causal event. *Cognitive Development* **5**: 193–207
Pegg J E, Werker J F, McLeod P J 1992 Preference for infant-directed over adult-directed speech: Evidence from 7 week-old infants. *Infant Behavior and Development* **15**: 325–45
Piaget J 1952 *The Origins of Intelligence in Children*. International Universities Press, New York
Piaget J 1954 *The Construction of Reality in the Child*. Norton, New York
Quinn P C, Eimas P D 1987 On categorization in early infancy. In: Oates J, Sheldon S (eds.) *Cognitive Development in Infancy*. Lawrence Erlbaum, Hove, UK, pp. 131–61
Roberts K 1988 Retrieval of a basic level category in prelinguistic infants. *Developmental Psychology* **24**: 21–7
Rovee-Collier C 1995 Time windows in cognitive development. *Developmental Psychology* **31**: 147–69
Saffran J, Aslin R N, Newport E L 1996 Statistical learning by 8-month-old infants. *Science* **274**: 1926–8
Saffran J R, Johnson E K, Aslin R N, Newport E L 1999 Statistical learning of tone sequences by human infants and adults. *Cognition* **70**: 27–52
Scaife M, Bruner J S 1975 The capacity for visual joint attention in the infant. *Nature* **253**: 265–6
Sherman T 1985 Categorization skills in infants. *Child Development* **56**: 1561–73
Siqueland E R, DeLucia C A 1969 Visual reinforcement of nonnutritive sucking in human infants. *Science* **165**: 1144–6
Sorce J, Emde R N, Campos J J, Klinnert M D 1985 Maternal emotional signaling: Its effect on the visual cliff behavior of 1-year-olds. *Developmental Psychology* **21**: 185–200
Spelke E S 1976 Infants' intermodal perception of events. *Cognitive Psychology* **8**: 553–60
Spelke E S 1994 Initial knowledge: Six suggestions. *Cognition* **50**: 431–45
Spelke E S, Breinlinger K, Macomber J, Jacobson K 1992 Origins of knowledge. *Psychological Review* **99**: 605–32
Sugarman S 1981 The cognitive basis of classification in very young children: An analysis of object-ordering trends. *Child Development* **52**: 1172–8
Tomasello M 1995 Joint attention as social cognition. In: Moore C D P J (ed.) *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum, Hillsdale, NJ, pp. 103–30
Tomasello M 1999 *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA
Tomasello M, Call J 1997 *Primate Cognition*. Oxford University Press, New York
Trevarthen C 1979 Communication and cooperation in early infancy: A description of primary intersubjectivity. In: Bullowa M M (ed.) *Before Speech. The Beginning of Interpersonal Communication*. Cambridge University Press, New York, pp. 321–47
Tronick E Z, Als H, Adamson L, Wise S, Brazelton T B 1978 The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child Psychiatry* **17**: 1–13
Walton G E, Bower N J, Bower T G R 1992 Recognition of familiar faces by newborns. *Infant Behavior and Development* **15**: 265–9
Wynn K 1992 Addition and subtraction by human infants. *Nature* **358**: 749–50

T. Striano and M. Tomasello

Infant Education

Infant education in the family context, without any intervention program, is largely understood as a nondirective and caring activity for a baby to secure survival, to meet physiological as well as psychological needs, and to provide both emotional warmth and

cognitive stimulation well-tuned to the growing competencies of the infant. More than 'education' concepts like 'socialization' or 'enculturation' highlight the major targets of early child care and instruction under nonintervention conditions. The concept 'infant education' is used mostly to describe programs that have been developed by public institutions (state, community, church) and aim at children's competence enhancement at a very early age by intervention. Programs try to compensate early deprivations which encompass wider environmental and economical hardship in living conditions, as well as parental incompetence and disengagement to raise a healthy and competent child. Infant education addresses children directly by focusing on the infants as developing individuals and it is also directed towards parents and communities for establishing educational facilities which foster the infant, contribute appropriately to the maintenance of physical health, and provide stimulation and emotional attention for developing children's social and cognitive skills.

1. Definition of Concept

1.1 Addressees

Programs aim both at risk and nonrisk children. Risk children are children who grow up in poverty, with unmarried or divorced mothers, in a violent environment or with health problems. Nonrisk children do not show these characteristics, but they still may lack cognitive and social competence, good emotion regulation, and a feeling of well-being. The dual approach to reach both risk and nonrisk children emphasizes a group or even mass orientation rather than a focus on individual help. 'Education' in infancy is often associated with a too formal approach to teaching, but most programs do not only aim at reaching the children but include parents and address the quality of interaction between caretakers and infants. Thus, infant education can be defined as an institutionalized preventive intervention for families with infants.

1.2 Contents of Programs

Participants in these programs are parents who intend to maximize their child's potential at an early age, with the help of a mediating person (e.g. community support), or by giving the child into an institution such as special day care center. Infant education focuses on basic aspects of early development such as mutuality in mother–child interaction, emotion regulation, rhythm of attention, support for exploring, and aid that a child can learn to soothe him- or herself when being distressed or in need for comfort. Aside from stimulation and affectivity provided by the caretakers, infant education encompasses also the enhancement of developmental sensitivity in parents for age-appropriate forms of stimulation and attention. Some projects strongly concentrate their efforts on parents and teach them to implement the program during communication with children in order to compensate infants' possible deviations from a normative course of development.

2. Intellectual Context of Infant Education Programs

An old dream of mankind is to enlighten the entire next generation by appropriate education, to invent the 'superbaby,' that is, a new generation of human beings superior to the present generation. After longer periods of intellectual stagnation such as, for example, after World War II in the USA, a time characterized by prosperity and the Cold War in the 1950s, the so called 'Sputnik shock' mobilized public activities in the USA and Western Europe to identify unexploited intellectual capacities in the population. At the same time, new ideas and fantasies flourished about more equality and more liberty in a better society.

2.1 Programs Focusing on Finding More Intellectual Capacity and on Creating Increased Equality

Early intervention programs in the 1960s and 1970s concentrated mostly on cognitive stimulation. The main target group were African-American children from families living in poverty. The basic assumption underlying these programs was the belief that early intervention could help find intellectual reserve capacities, equalize chances at school entry, and support children in acquiring basic skills. Moreover, the intention was also to prevent educational, physical, social–emotional, and cognitive problems during later development. The curriculum movement during the 1960s believed that, in the sense of Amos Comenius, all things could be taught to all. However, evaluations of programs brought sobering results: After some immediate gain in IQ points during the program inside the institutions, children rapidly lost most of their gains outside the institutions (Schafer 1970).

Starting from these results, new projects were designed which focused more on a family, rather than in a child-oriented approach (e.g. Levenstein 1970). Mothers were included in the programs as teachers of their own children and this time results confirmed the expectations. Inclusion of mothers or even the entire family showed an increase in program effectivity compared with interventions where only infants were included.

2.2 Programs Focusing on Early Optimization of Learning Capacities

Stimulation at a very early age was believed to guarantee the highest degree of fostering the child's

available capacity. Infant education was driven by the philosophy that the more stimulation a baby receives, the better off he or she will be in later life. Early teaching was applied by parents who felt responsible for the children's chances to successfully enter college and university and later make it to the top. In expensive infant schools, toddlers were taught letters and numbers, they had to listen to the music of Beethoven, Mozart, or Bach, and they were confronted with paintings of Van Gogh, Matisse, or Toulouse-Lautrec. However, the intense teaching and emphasis on intellectual excellence was during a time when children need parental attention, warmth, and love to a high degree, and many experts in early development warned parents not to exaggerate their pressure for high achievement on the kids to become bright intellectuals and neglect their needs for social and emotional attention.

3. *History of Infant Education*

To enculturate a child as early as possible has been a salient issue already during the times of early humanism in Europe. After the long period of scholasticism new ideas began to emerge about the potential of infants and new education programs were formulated aiming at a general liberalization of human thinking. Contrary to Augustin (AD 354–430) who saw the infant as a two-edged subject, a rude, culture-free, and wild being on the one hand, and a hope in the future, a possible multiplier of Christian faith on the other. Beyond that, Anselm of Canterbury (1033–1109) was one of the first few scholars who formulated concrete guidelines for early education. Children should be supported from infancy on, given enough time for play, and instructed without physical punishment. Erasmus Rotterdamus (1466–1536) went a step further when he emphasized in his work 'on early scientific education of children' the eminent role of the mother for infant education. His friend and colleague Juan Vives (1492–1540) instructed mothers to talk very carefully to their children, particularly during the first years. Both Erasmus and Vives underlined the high responsibility of mothers for the quality of development in early childhood and stressed the importance of the dialogue between mother and infant.

3.1 *Theoretical Background: Evolution and Behaviorism*

Over time, tremendous changes occurred in the conceptualization of infant education. Beliefs about possible influences on infants' development run between the two extremes of 'nature' and 'nurture'; beliefs about what should be transmitted in education and enculturation can be found between 'conservation of tradition' on one end and 'creation of the new human being' on the other. After Darwin and his publication *Descent of Man* (1871), human development was seen by many as mirroring the evolution process of the human species. According to Ernst Haeckel, infant development was considered to 'replicate' human evolution. Intervention during the early years appeared to be rather ineffective or useless, the growth of skills and abilities was believed to be a kind of natural unfolding process (Hall 1883). However, the rise of animal psychology and learning theory at the turn of the century changed the view dramatically. Now, the infant was believed to be totally malleable by environmental manipulation (Watson 1928).

3.2 *Extension of Program Aims*

Behavioral roots and the belief that learning processes can be manipulated by changes in environmental conditions led to the early intervention programs in the 1960s which concentrated on children from poor and underprivileged families. When during the 1970s activities and competencies of infants were further investigated, early stimulation was believed to be a necessary prerequisite to optimize development. With the intention to maximize their children's potential at a very early age, middle-class parents began to teach their children academic topics by applying techniques of programs developed to boost underprivileged children. In the 1980, infant education changed its basic character from a compensatory to a competitive philosophy. The war on poverty in the 1960s turned to become a combat for enhancing the chances of middle-class children. 'Everyone wants to raise the smartest kid in America rather than the best adjusted and happiest kid' (Brazelton, cited in Langway et al. 1983, p. 44).

4. *Current Theory and Research*

Home environments had not been systematically observed for a long period of time. Since the 1930s, interest in the observation of children's natural environment had decreased dramatically, and there was no conceptualization about the transmission of deprivation from the larger environmental conditions to the children's proximal environments. In reaction to the nonlasting effects of cognitive improvements, new programs were started using mothers as teachers. Thus, children could keep a continuing relationship with the mother as the programs were addressing the mother–child dyad directly. Results showed that cognitive and emotional components seemed to reinforce each other.

4.1 Increasing Evidence of Infants' Early Social Competencies

Experimental studies have amply shown infants' early social competence to imitate their mothers' facial movements or even to restart a 'frozen' interaction with their mothers by producing new facial expressions to stimulate their communication partners (Tronick 1989). This research brought forward the idea that babies are capable to acquire skills—mostly intellectual ones—even at a very early age but that they also need a stable and warm relationship with a primary caregiver.

During the 1970s, studies comparing children with specific language skills with control groups showed the importance of the mother–child relationship for the acquisition of language (Baumrind 1971). The main difference consisted in the mode of verbal stimulation and interaction styles. Mothers of early competent children let their children continue their activities and did not intervene or disrupt, they changed their instructions according to the growing cognitive competence of their children, and gave more and more elaborate advices if required. Children began to ask for help in a more and more sophisticated manner. Those who were classified low in their competence had mothers who intervened, interrupted activities, and demanded to follow instructions. These children had less chances to explore and to understand. The best summary for these differences in learning conditions was given by Hess and Shipman (1965): 'The meaning of deprivation is the deprivation of meaning' (p. 885).

4.2 Emotional Implications of Early Day Care

From the 1980s on, the general use of day care increased considerably. Programs were now developed not to compensate children's growing up in poor environments but rather to meet the needs of working middle-class mothers to push their children to high cognitive abilities. Long debates began whether day care could be harmful to children's emotional development, particularly when Bowlby's (1973) concept of a 'secure base' as formulated in his attachment theory was considered. Detrimental effects for children under three years were expected when separated too long from their mothers. However, research on day care exhibited balanced results, taking day care as a given and collecting descriptive evidence about children's experiences in different kinds of day care programs. Quality day care was created as a demand for communities to consider the wealth of children as the highest criterium (Belsky et al. 1982). Belsky and Steinberg (1978) reviewed the main studies dealing with day care and pointed to the following results: day care has neither beneficial nor deleterious effects, children growing up in high-risk environments do have benefits, and early day care is not disruptive of the child's emotional bond with the mother.

5. Evaluation of Programs, Cognitive Effects in Longitudinal Studies

The first evaluation studies which tried to measure long-term effects of early intervention programs did not produce results that could easily encourage evaluators to implement more similar intervention programs. There had been quite a number of intervention programs for disadvantaged African-American children from lower social classes during the 1960s and early 1970s (Hess and Shipman 1965, Karnes et al. 1970, Schaefer 1970, Lambie et al. 1974). Most of these programs led to some gains in children's IQ, but gains did not persist for long after school entry. Later, a debate started about the usefulness of home-based versus center-based programs. In the end, consensus was reached that there are some long-term effects, when children's proximal environments, that is, their mothers or even the families, were included in the program.

5.1 General Factors Contributing to Long-term Effects

In a review evaluation describing long-term effects of early intervention programs on later academic achievement, Rutter (1987) mentions four aspects for continuous effects into middle childhood: '(1) IQ gains do not persist long after school entry unless the intervention also continues; (2) there are some modest lasting increments in scholastic achievements (in some studies only); (3) intervention is followed by some positive changes in the attitudes of children and parents, and of teachers toward the children; and (4) most especially, intervention results in a lower proportion of children requiring special educational treatment' (p. 1270). Quite similarly, the 'Consortium,' a group of independent researchers who studied long-range effects of early intervention (Lazar and Darlington 1982), found four overall effects of early infant education programs: (a) positive effect on meeting later school requirements, (b) significant gain in IQ points which lasted over time, (c) increase in self-esteem associated with higher performance in school, and (d) effect on positive maternal attitude toward the child's school performance. Rutter (1987) interpreted these results as the possibility of both a direct and an indirect impact on children's development such as improvement of cognitive performances and a positive influence on self-esteem on the one hand, and on attitudes toward learning in general on the other. Increased parental aspirations may trigger a positive cycle of achievement.

5.2 Specific Factors Contributing to Long-term Effects

More recent evaluations in the 1990s underlined the importance of family support in programs for infant education. For example, a program which first promoted family support and then offered preschool education in an institution could promote better preparation for school readiness and higher achievement in school. In the Carolina Abecedarian Project (CAP) and its replication, the CARE project, where the beginning of intervention was very early, at 6 to 12 weeks of age, and lasted longer with intensive care, up to five years of age, a whole scale of different activities and supports was offered to the participating families: full-day developmental day care, occasional home visits, parent–group meetings, on-site medical care and a toy-lending library. During school age, intervention lasted three years, and consisted of a home–school resource teacher who provided an average of 15 home visits per year to assist parents in supplementing school instruction (after Weissberg and Greenberg 1998, p. 904). These programs try to change parents' behaviors and attitudes toward both the child and the school (Schweinhart et al. 1993, Sameroff 1991). Such programs can also prevent delinquent behavior and maladjustment as newer studies have shown evaluating programs with frequent home visits which seem to enhance parents' competence and improve family functioning when this is also combined with a supporting community as a complementary component (Yoshikawa 1994).

6. Probable Future Directions of Theory and Research

After the experience of continuous disappointments of high hopes in changing children's environmental disadvantages by infant education programs during the 1960s and 1970s, and after a dramatic change in public interest from compensation to competition in early childhood, now the thinking about children and their healthy development in a changing society should find a new balance. In the late 1990s, the impact of new knowledge about infants' very early competencies on the one hand, and on dispositional as well as environmental effects on the individual course of development on the other has opened a more holistic view on children and their development. Physical, emotional, and cognitive growth are seen as being strongly intertwined. As evaluation studies have shown, intervention on only cognitive aspects did not lead to expected compensatory or optimization effects. It is hoped that early education will recover as a discipline which not only addresses disengaged parents of disadvantaged children on the one hand, and highly motivated and already highly privileged parents on the other, but as an area of research producing advice for all young parents. As dual earner careers in young couples having children need early instructions about the needs of infants and their developmental changes, the early cognitive stimulation is only a single component and not the entire gamut of necessary conditions which a child needs to grow up healthily. After a period of focusing on cognitive stimulation in children, perhaps now the time has come to emphasize other aspects such as relationship quality, emotion regulation, and early rhythm in mutuality between child and mother as prerequisites for the successful development of skills and abilities and the growth of life competence.

See also: Child Care and Child Development; Cognitive Development in Childhood and Adolescence; Early Education and Care in the World Regions; Infancy and Childhood: Emotional Development; Infant and Child Development, Theories of; Infant Development: Physical and Social Cognition; Play and Development in Children; Prenatal and Infant Development: Overview; Socialization in Infancy and Childhood

Bibliography

Baumrind D 1971 Harmonious parents and their preschool children. *Developmental Psychology* **4**: 99–102

Belsky J, Steinberg L D, Walker A 1982 The ecology of day care. In: Lamb M E (ed.) *Nontraditional Families: Parenting and Child Development*. Erlbaum, Hillsdale, NJ, pp. 77–116

Belsky J, Steinberg L D 1978 The effects of day care: A critical review. *Child Development* **49**: 929–49

Bowlby J 1973 *Attachment and Loss: Separation*. The Hogarth Press, London, Vol. II

Darwin C 1871 *Descent of Man*. John Murray, London

Hall G S 1883 The contents of childrens' minds. *Princeton Review* **2**: 249–72

Hess R D, Shipman V C 1965 Early experience and the socialization of cognitive modes in children. *Child Development* **36**: 869–86

Karnes M B, Teska J A, Hodgins A S, Badger E D 1970 Educational intervention at home by mothers of disadvantaged infants. *Child Development* **41**: 925–35

Lambie D Z, Bond J T, Weikart T B 1974 *Home Teaching of Mothers and Infants*. High Scope Educational, Ypsilanti, MI

Langway L, Jackson T A, Zabarsky M, Shirley D, Whitmore J 1983 Bringing up superbaby. *Newsweek* **13**: 44–9

Lazar I, Darlington R 1982 Lasting effects of early education: A report from the consortium for longitudinal studies. *Monographs of the Society for Research in Child Development* **47**: Nos. 2–3, Serial No. 195

Levenstein P 1970 Cognitive growth in preschoolers through verbal interaction with mothers. *American Journal of Orthopsychiatry* **40**: 426–32

Rutter M 1987 Continuities and discontinuities from infancy. In: Osofsky J D (ed.) *Handbood of Infant Development*, 2nd edn. Wiley, New York, pp. 1256–96

Sameroff A J 1991 Prevention of developmental psychopathology using the transactional model: Perspectives on host, risk agent and environmental interactions. In: *The Prevention of Mental Disorders: Progress, Problems, and Prospect*. National Institute of Mental Health, Washington, DC

Schaefer E S 1970 Need for early and continuing education. In: Denenberg V H (ed.) *Education of the Infant and Young Child*. Academic Press, New York, pp. 61–82
Schweinhart L J, Barnes H V, Weikart D P 1993 *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27*. High/Scope Press, Ypsilanti, MI
Tronick E Z 1989 Emotions and emotional communication in infants. *American Psychologist* **44**: 112–19
Watson J B 1928 *Psychological Care of Infant and Child*. Norton, New York
Weissberg R P, Greenberg M T 1998 School and community competence—enhancement and prevention programs. In: Damon W, Sigel I E, Renninger A (eds.) *Handbook of Child Psychology: Child Psychology in Practice*, 5th edn. Wiley, New York, Vol. 4, pp. 877–954
Yoshikawa H 1994 Prevention of cumulative protection: Effects of early family support and education on chronic delinquency and its risks. *Psychological Bulletin* **115**: 1–27

K. Kreppner

Infanticide and Induced Abortion

Abortion and infanticide are among the methods human societies and the individuals within them use to affect net fertility. Deliberate actions to prevent conception despite sexual intercourse are referred to as contraception, while actions intended to prevent a pregnancy from yielding a live birth are referred to as induced (as opposed to spontaneous) abortion, and actions intended to prevent a newborn from growing to a viable adult are referred to as infanticide.

1. Methods and Prevalence of Infanticide and Abortion

Infanticide has been observed among most mammal species and is recorded in nearly all human cultures. Infanticide is usually indirect, in that children are abandoned rather than being actively put to death (but see Boswell 1988, who argues that abandoned children were often raised by a third party, if only to be sold as slaves). In addition, some children die from other forms of extreme neglect verging on infanticide (e.g., by being left at a foundling hospital, by being accidentally smothered while sleeping in the bed with adults, or by being inadequately fed by a wet nurse).

In many premodern societies, infanticide appears to have been common and demographically important. Survivors of attempted infanticide figure prominently in Western literature (Moses, Oedipus). Reports for European cities in the eighteenth century suggest that a quarter of all children were abandoned to foundling hospitals, where the probability of surviving to adulthood appears to have been well below a third. Some forms of abortion were known in premodern societies (e.g., herbal remedies, massage, or insertion of foreign objects into the uterus), but they had substantially less than perfect effectiveness, often involved considerable risk to the woman, and do not appear to have been widely used (though the prevalence of abortifacients in the premodern period is the subject of some scholarly debate; see Van de Walle 1994). In Western Europe and the US, abortion appears to have become more effective and demographically important in the mid-nineteenth century.

In the modern period, most abortions are performed surgically, using vacuum aspiration (in which, often under local anesthesia, the uterus is emptied through suction induced by a syringe or a pump) or dilation and surgical curettage (in which, usually under general anesthesia, the cervix is dilated and the lining of the uterus is scraped). Legal surgical abortion performed during the first three months of pregnancy is much safer to the pregnant woman than a live birth (death rates of about 1:100,000 vs. 10:100,000); rates of complication and death are much higher for illegal abortions performed outside of regular medical facilities.

More recently, a medical approach to inducing abortion within three weeks after a missed period (7 weeks after the last period) has become widely available. This method, called RU-486/miferpristone/Mifeprex, has been widely used in Europe since 1988 and was approved for use in the US in 2000.

In addition, medical approaches on the boundary between contraception and abortion have become better publicized and more widely available. Specifically, special purpose 'morning after pills' or even a high dose of some oral contraceptive within three days after intercourse cuts the probability of pregnancy by about 75 percent, apparently by inhibiting ovulation or preventing a fertilized egg from implanting in the uterus (a process that usually takes six to nine days after intercourse).

Because of its ambiguous legal and moral status, information on abortions is difficult to collect. Not even all countries in which abortion is legal collect information on abortions as part of their vital statistics system. Furthermore, even when such reports are collected they often appear to be too low (in the US perhaps by 10 percent). Survey-based prevalence estimates appear to be even lower (in the US perhaps by 40 percent) and the degree of under-reporting is not uniform in population subgroups. These problems with survey data are likely to badly bias survey-based estimates of pregnancy rates and contraceptive failure rates. Thus, our knowledge of the prevalence of abortion is poor.

The available evidence suggests that the prevalence of surgical abortion varies widely (see Henshaw et al. 1999). Worldwide there were estimated to be about 46 million abortions per year in the mid-1990s (slightly less than half of them illegal). This figure implies about 35 abortions per 1000 women age 15 to 44 (the 'abortion rate'), about 26 abortions per 100 known pregnancies (the 'abortion ratio'; though note that the

abortion ratio is sometimes defined as abortions per live birth), and on average about one abortion over each woman's lifetime (the 'total abortion rate'). In Russia and several other countries in Eastern Europe, the total abortion rate is above 2; in most of Western Europe and the English-speaking countries the rate is below, often well below, one-half. In the United States, the rate is 0.69. Asia, Africa, and South America are between these two extremes, but with considerable intra-regional variation.

In the US, for which detailed information is available (Koonin et al. 2000), the abortion ratio peaked in the mid-1980s and has since fallen by about 15 percent. (The patterns for the total number of abortions and the abortion rate are similar.) About 43 percent of all women are projected to have an abortion during their lifetime. Finally, 58 percent of women having abortions report that they had used a contraceptive method during the month they became pregnant, though usually the less effective methods (over half were using the condom) and the available evidence suggests that the methods was not always used consistently or correctly.

Pregnancies are most likely to be terminated by abortion for relatively young and old women (about a third of all pregnancies for women under 25 or over 40), for blacks (compared to whites; a third vs. a sixth), and for unmarried women (40 percent vs. 7 percent). About half of all abortions are the woman's second or higher-order abortion. These and other characteristics of those having abortions suggest that abortions are more common among those who lack financial and social resources and perhaps have too little control over their lives. Finally, while women sometimes choose to abort because of rape, incest, fetal abnormalities, and threats to their physical health, the overwhelming share of abortions are due to conventional motivations for family planning—not wanting another child now (or ever) because of financial considerations, relationship problems (not married, not wanting a child with this partner, lack of support from current partner), and life course issues (still in school, age).

2. *Attitudes Towards Infanticide and Abortion*

Evolutionary biology suggests that infanticide should be rare. Natural selection acts to maximize reproductive success. Infanticide appears to lower the number of offspring. Thus, we would expect a taboo on infanticide. Infanticide is, however, known in many mammal species and in most human societies.

More general models suggest, however, that infanticide can increase the net number of surviving offspring and thus is likely to be the result of natural selection. Specifically, given the very long period from birth until independence for humans and given the hunter-gatherer ecology in which the human species is understood to have evolved, attempting to raise two closely spaced children was likely to lead to the death of both of them. Anthropological evidence from modern hunter-gatherer populations suggests that prolonged breast-feeding led to interbirth intervals of three to four years. If a birth occurred despite breast-feeding, infanticide would increase the net number of surviving children. Similar analyses apply to the birth of twins and unhealthy infants (Scrimshaw 1984).

Moreover, modern Western discussion of infanticide, abortion, and even contraception has been strongly affected by Judaeo-Christian perspectives on the value of life and the need to treat human beings as ends, not means (see Posner 1992). Depending on when—for the purposes of defining murder—'life' is considered to begin, infanticide or abortion may be 'murder' or at least morally problematic (though perhaps balanceable against other considerations; e.g., the health of the mother, rape, and genetic defects). Most modern legal systems consider 'life' to begin at birth, so under this definition infanticide is murder, but abortion is not. It is, however, important to note that in modern Western societies, even the taboo on infanticide is not absolute; many individuals would allow a profoundly deformed or severely mentally retarded newborn to die. In contrast, the Catholic Church treats a fetus as a full-person (until 1869 from 'quickening,' thereafter at conception) and thus views abortion as murder.

Even for those who do not consider 'life' to begin until birth, abortion is morally ambiguous since the fetus represents a potential life. Polling data for the US suggests considerable ambivalence about abortion and its appropriate legal status and strong opposition in some subpopulations. On the other side, proponents of abortion advocate it as a form of fertility control, as a way of saving lives of women who would have illegal abortions, and in terms of the right of a woman to make informed and voluntary decisions about childbearing.

The combination of the moral ambiguity of abortion and strong institutional opposition from the Catholic Church and conservative Protestant groups has led to legal regulation of abortion in many countries. Abortion was almost universally illegal in the early twentieth century. This first changed in the early years of the Soviet Union, where from 1920 to 1936 abortion was legal, widely available, and encouraged as the primary method of fertility control. In the period after World War II, abortion was first legalized in the Scandinavian countries, much of Eastern Europe, and Japan; and later in most of Western Europe. In the US, several states legalized abortion beginning in the mid-1960s, and in 1973 abortion became legal nationally with the controversial Roe vs. Wade decision of the US Supreme Court finding a constitutional right to abortion before viability (at that time, 28 weeks; with improving medical technology, several weeks earlier).

At the beginning of the twenty-first century, abortion is legal in almost all of Europe (though restricted in Germany) and most of Asia (sometimes, as in the case of Bangladesh, in the form of 'menstrual regulation'; see Dixon-Mueller 1988), but illegal in most of Africa and South America. In many of the places where abortion is illegal, the abortion rate is nevertheless substantial (e.g., much of Africa and Central and South America). Furthermore, the strictness with which such laws are interpreted and the frequency of prosecution vary widely.

Even in places where abortion is legal, in particular the US, and Southern and Eastern Europe, the specific conditions under which abortion is allowed (e.g., permissible reasons, gestational age, facilities and practitioners, parental notification or consent for minors, counseling and waiting periods) are often restricted by law. In addition, the details of those regulations and whether government health programs fund abortions are politically salient (e.g., they appear in public referenda, are issues in campaigns for public office, cause problems for the funding of foreign aid) and fluid. Similarly, abortion has been prominent at UN conferences and policy debates about family planning.

3. *Behavioral Models and Empirical Analyses*

Standard models for the use of contraception suggest that the prevalence of abortion should be affected by the perceived costs and benefits of children (and their determinants), the broadly defined cost of abortion (i.e., monetary costs, time costs, and psychic costs), and the broadly defined cost of other methods of fertility control.

Beyond these insights from the general theory of fertility control, theoretical approaches to abortion focus on the temporal position of abortion between conception and birth. This approach suggests two paths to abortion. First, especially in places where contraceptives are difficult to obtain, some women do not contracept, planning that if they become pregnant, they will abort. Second, some women do contracept but plan (sometimes only implicitly) to abort if they experience a contraceptive failure. The limited survey evidence suggests that both paths are important. In the US, slightly more than half of women having abortions report that they were contracepting at the time of conception; the rest were not using any method.

Abortions following contraceptive failure appear to result from two different phenomena. First, the standard modern life-course—early initiation and high prevalence of sexual activity and a desired family size of about two—implies that many women must contracept for over 200 months. Thus, even given consistent use of contraceptive methods with low failure rates (e.g., the oral contraceptive), a sizeable fraction of all women will experience an unintended pregnancy during their lifetime; many of them will choose abortion. Second, among those nominally contracepting (i.e., using any method at some time in the previous month), some subpopulations have high pregnancy rates. Such rates appear to be due to a combination of selection of a method with a high use failure rate (for example, the condom, chosen partially because it protects against sexually transmitted diseases) and inconsistent and ineffective usage of the method (e.g., not taking every pill every day). These two phenomena lead both to abortions and to high rates of voluntary sterilization.

Social psychological considerations also appear to be important. Some women choose not to contracept because of ambivalence about their sexuality (e.g., they deny that they are sexually active, or that they plan sex). Other women choose not to contracept (or not to contracept effectively) because they underestimate the probability of conception (e.g., they underestimate their own fecundity or they overestimate the effectiveness of their method as used) or highly discount the future. For these women, abortion is a second chance that allows them to remedy the effects of their earlier denial or incorrect perceptions.

Finally, some women choose abortion rather than contraception because they learn something because of the pregnancy. Particularly in parts of Asia, it appears that abortion is used as a tool for sex selection and yields an imbalanced sex ratio. Similarly, some women become pregnant in order to learn the true intentions of a partner, and abort if the revealed intentions are unfavorable. Some empirical studies find an increase in nonmarital births with lower costs of abortion, which is consistent with this motive for abortion. These ideas have been extended to models of equilibrium in the labor market, where they imply that the availability of abortion lowers women's bargaining position in the marriage market and reduces the marriage rate. (Men need no longer promise marriage in order to get sex and need no longer marry a woman if she becomes pregnant; Akerloff et al. 1996.)

Statistical analyses of the determinants of abortion and the effects of abortion regulation have been hampered by the difficulty of measuring abortion rates. Early work considered country-level time-series for contraception, abortion, and fertility. More recent work has applied difference-of-difference methods to time-series of cross-sections data to explore the effect of specific public policies and abortion access on abortion and fertility (e.g., Joyce and Kaestner 1996, Klerman 1999, Levine et al. 1999).

Rational-choice analyses using dynamic programming methods suggest that raising the cost of abortion should be expected to increase both contraceptive use (and other methods for fertility control such as decreasing coital frequency and delaying marriage) and realized fertility. While abortions are often common even when abortion is illegal, legalization of abortion in Europe and the US sharply increased

abortion rates and decreased fertility. In addition, public funding of abortions in the US through Medicaid appears to raise abortion rates and lower fertility rates for blacks. Given poor measurement of contraceptive prevalence, it has been difficult to estimate the effect of changes in the cost of abortion on the prevalence of contraception. In principle, the effect could be detected as the difference between the change in the number of abortions and the number of births. The clearest case appears to be Romania's banning of abortion after a period when it was legal. The number of births increased much less than the previous number of abortions (even after accounting for illegal abortions) and there were reports sharp decreases in coital frequency. Similar effects appear to have been present when abortion became legal in other countries of Eastern Europe, especially when experience with contraceptives was of short duration (Frejka 1983).

Similarly, decreasing the cost of contraception should be expected to increase contraceptive prevalence. This should prevent unintended pregnancies, some of which would have been terminated by abortion; thus lowering both the abortion rate and the fertility rate (see Rahman et al. in press for evidence from Bangladesh). Such substitution has been used to explain the simultaneous increase in contraceptive prevalence, decrease in abortion, and decrease in fertility in Russia in the early 1990s, in Hungary in the late 1960s and early 1970s, in Japan in the 1960s; and in Kazakhstan in the 1990s with the introduction of a USAID contraceptive distribution program. Similarly, the (broadly defined) low costs for contraception, easily available without stigma, appear to explain high rates of sexual activity, high rates of contraception, but low rates of abortion and low fertility in Western Europe.

See also: Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility Control: Overview; Fertility: Proximate Determinants; Fertility Transition: China; Fertility Transition: East Asia; Infant and Child Mortality in Industrialized Countries; Infant and Child Mortality in the Less Developed World; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations

Bibliography

Akerloff G A, Yellen J L, Katz M L 1996 An analysis of out-of-wedlock childbearing in the United States. *Quarterly Journal of Economics* **CXI:2**: 277–318

Boswell J 1988 *Kindness of Strangers: The Abandonment of Children in Western Europe from Late Antiquity to the Renaissance*. University of Chicago Press, Chicago

Critchlow D T 1996 *The Politics of Abortion and Birth Control in Historical Perspective*. The Pennsylvania State University Press, University Park, PA

Dixon-Mueller R 1988 Innovations in reproductive health care: menstrual regulation policies and programs in Bangladesh. *Studies in Family Planning* **19**(3): 129–40

Frejka T 1983 Induced abortion and fertility: a quarter century of experience in Eastern Europe. *Population and Development Review* **9**: 494–520

Henshaw S K, Singh S, Haas T 1999 The incidence of abortion worldwide. *International Family Planning Perspectives*. **25(supp.)**: 530–8

Joyce T, Kaestner R 1996 The effect of expansions in Medicaid income eligibility on abortion. *Demography* **33**: 181–92

Klerman J A 1999 US abortion policy and fertility. *American Economic Review* **89**(2): 261–4

Koonin L M, Strauss L T, Chrisman C E, Parker W Y 2000 Abortion surveillance—United States, 1997. CDC Surveillance Summaries, December 8, 2000 *Morbidity and Mortality Weekly Report* **49**(No. 55–11)

Levine P B, Staiger D, Kane T J, Zimmerman D J 1999 Roe v. Wade and American fertility. *American Journal of Public Health* **89**(2): 199–203

Posner R A 1992 *Sex and Reason*. Harvard University Press, Cambridge, MA

Rahman M, DaVanzo J, Razzaque A in press Do better family planning services reduce abortion? Evidence from Matlab, Bangladesh. *Lancet*

Scrimshaw S 1984 Infanticide in human populations: societal and individual concerns. In: Hausfater G, Hardy S B (eds.) *Infanticide: Comparative and Evolutionary Perspectives*. Aldine, pp. 439–62

Van de Walle E 1994 Review of John M. Riddle. Contraception and Abortion from the Ancient World to the Renaissance. *Population and Development Review* **20**: 221–4

J. A. Klerman

Infectious Diseases: Psychosocial Aspects

Infectious diseases are illness caused by invasive pathogens, such as viruses, bacteria, fungi and parasitic infestations. When the germ theory of disease was established by Pasteur and Koch in the nineteenth century, a mechanistic view of infectious disease became dominant, with illness depending on exposure to the agent and the virulence of the infection. However, the ascent of immunology in the mid-twentieth century shifted the balance towards a recognition that the strength of bodily defenses plays a significant role in infectious disease.

Psychosocial factors are relevant to the experience of infectious disease in two principal ways.

(a) Psychosocial factors may influence the acquisition of infection or the reactivation of latent infection. Much recent research on upper respiratory illnesses such as the common cold and on herpes virus infections (genital and oral herpes) has been directed at this possibility.

(b) Psychosocial factors may affect the course of infectious illness once it is developed, influencing the duration and severity of symptoms.

Readers interested in more detailed overviews of psychosocial aspects of infectious disease are referred to recent reviews (Cohen and Williamson 1991, Zorrilla et al. 1996, Steptoe and Cohen 1999).

1. Mediating Processes

There are four general processes through which psychosocial factors such as acute and chronic life stress, social support, and coping may influence the acquisitional course of infectious disease.

1.1 Exposure to Pathogens

Since infectious diseases are caused by pathogens, any changes in exposure that are stimulated by psychosocial factors may influence the chances of becoming ill. For example, some people respond to distressing events in their lives by seeking social support and mobilizing social networks. This might have the unwanted effect of exposing the individual to more respiratory infections. Alternatively, people may respond by social withdrawal, reducing contact with airborne pathogens. Differences in exposure to infections resulting from crowding and other factors probably account for much of the socioeconomic gradient in conditions such as infectious mononucleosis (glandular fever). Exposure to pathogens is a key issue in the prevention of outbreaks of serious waterborne, airborne and sexually transmitted infections in the developing world, but little psychosocial research has yet touched on this topic (Aboud 1998).

1.2 Changes in Behavior

Health behaviors and lifestyle have an influence on infectious disease. Smoking, alcohol consumption, and physical exercise all influence vulnerability to infection and the duration of symptoms. Malnutrition is probably the principal cause of diminished immunity and susceptibility to infection across the world. Although severe protein energy malnutrition is rare in developed countries, undernutrition is common among the elderly and impairs immune resistance. If the pattern of health behaviors is altered in relation to psychosocial factors such as life stress, then the acquisition and course of infectious disease may be modified. Behaviors can have direct effects on infectious disease, such as when sexual behavior stimulates the reactivation of latent herpes virus, or the acquisition of human papilloma virus and risk of cervical cancer. Behaviors related to health service utilization are associated with socioeconomic status in the developing world, and play an important role in determining risk of diseases such as malaria (Castilla and Sawyer 1993).

1.3 Symptom Reporting

People vary substantially in the extent to which they attend to physical sensations and recognize them as symptoms of illness. Research on illness behavior has shown that the same infection may elicit quite different levels of complaint, work absence, and health service utilization, depending on levels of psychosocial stress. Negative affect also influences symptom reporting and illness behavior, with more symptoms being reported in distressed individuals independently of objective illness (Feldman et al. 1999).

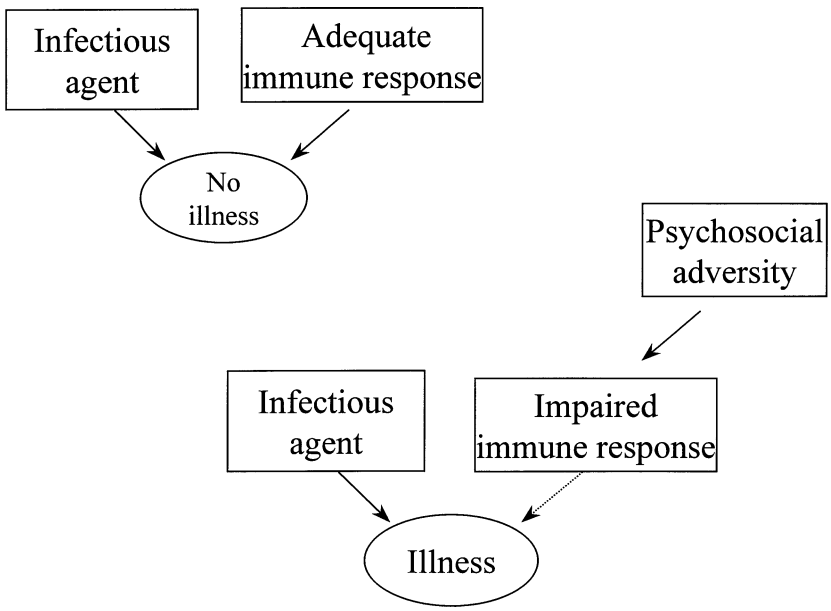

Figure 1
Schematic outline of the link between infection, immune defenses, psychosocial factors, and illness

1.4 Suppression of Bodily Defenses

The final possibility is that psychosocial factors influence susceptibility and the course of infectious disease by modifying physiological defenses. Depending on the infectious agent, skin abrasions, immune processes, or mucosal barriers may be involved. One of the principal interests in psychoneuroimmunology research has been the possibility that psychosocial factors reduce immune defense, allowing infections to become established that would otherwise be resisted. These hypothetical associations are illustrated schematically in Fig. 1 (Steptoe 1998). The upper panel illustrates the pattern of healthy normal response to minor infections, where the individual mounts an adequate defense that excludes the infection and prevents illness. In the lower panel, exposure to the infectious agent takes place against a backdrop of impaired immune function due to psychosocial stimulation that is not sufficient to prevent illness from developing.

These mediating processes are not mutually exclusive, and one of the difficulties in research on infectious disease lies in teasing out the various pathways through which the acquisition or duration of infection are affected.

2. Methodological Aspects

The processes mediating between psychosocial factors in infectious disease outlined in Sect. 1 have important implications for how research on this topic is carried out.

2.1 Verification of Infection

Interpretation of associations is enhanced greatly by objective verification of infection. If infection is assessed merely by the subject or a family member, then it is not possible to rule out symptom reporting biases, and spurious associations between psychosocial factors and unfounded symptoms may be identified. Objective verification allows the timing of infection in relation to psychosocial stimuli to be established, so that causal sequences can be described. The most precise verification is possible in experimental studies of infection, where volunteers are infected deliberately with metered doses of virus or bacteria. The acquisition of infection then is assessed by culturing samples obtained some days later. Unfortunately, biochemical verification of infection is difficult and invasive in naturalistic studies, requiring repeated sampling of blood or nasal washes (for upper respiratory infection) from the individual under investigation. As an alternative, clinical examination by health professionals or trained observers is valuable for objectively assessing the occurrence of infection.

2.2 Standardized Measures of Symptoms

Since symptom reporting can be influenced by psychosocial factors independently of objective infection, it is important to reduce variability by employing standardized measures. These measures detail the number, type, and duration of symptoms. Work on upper respiratory infection, for example, typically assesses 10–14 standard symptoms of colds. Such indices are particularly useful in assessing the duration and severity of symptomatic episodes.

2.3 Measures of Health Behavior and Biological Factors

One of the most important implications of the discussion of mediators in Sect. 1 is that adequate measures are taken of social activity, smoking, alcohol consumption, sleep patterns, diet, and other health behaviors. Analysis of these variables will establish whether changes in health behaviors account for variations in vulnerability to infections. Measurement of hormonal and immunological factors is of course extremely interesting in developing an understanding of mediating biological processes.

2.4 Longitudinal Designs

Many studies of infectious disease have utilized cross-sectional or retrospective designs, with assessments of the frequency of infection in the past, along with measures of stress and other factors. Such studies have been used to find out whether the number of infections over the past month or year was greater among individuals who experience more life stress or have poorer social supports than others. Cross-sectional and retrospective studies are limited in the inferences that can be drawn, since selective recall biases may operate. For example, it is possible that individuals with depressed mood will recall more episodes of infection, and also report more negative life experiences or poorer social supports than others. As with many topics in health psychology, the most convincing data derive from longitudinal studies in which psychosocial factors are assessed prior to illness, so that prospective links can be established. Studies involving repeated measures of psychosocial, behavioral, and biological factors are particularly strong, since they allow the time course of vulnerability to be established.

3. Experimental Studies

Many studies have investigated factors such as stress in relation to infection in animal models. Animal studies do not suffer from many of the methodological problems inherent in human research, but are limited

in terms of the subtlety of psychosocial investigation. This literature has been reviewed recently (McCann et al. 1998, Rabin 1999).

Experimental studies of infection in humans are of course rare, and are only justified for mild forms of disease. But an important program of work on colds and other upper respiratory tract infections has been conducted by Sheldon Cohen and colleagues in the UK and USA that throws considerable light on the role of psychosocial factors. In these studies, volunteers free of infection are inoculated with standard doses of a virus. If an appropriate dose is administered, not everyone will become ill. Differences in the subsequent rates of infection and illness must, therefore, be due to differences in vulnerability. Social activities are limited by quarantine, and health behaviors such as smoking are controlled, so psychosocial impairment of bodily defenses can be evaluated.

Cohen et al. (1991) demonstrated with this paradigm that susceptibility to common cold viruses was associated directly with a psychological stress index created through an amalgam of measures of major life events over the past 12 months, perceived stress and negative affect. The dose–response association was independent of age, allergic status, and other risk factors. Later work has established that chronic stressors such as long-term work problems and relationship difficulties are particularly relevant. By contrast, social networks have a protective effect, with lower susceptibility to experimental infection among volunteers with more extensive and diverse social networks. Cohen et al. (1997) found that the rates of infection and illness were more than four times greater in volunteers with limited social networks, compared with high social networks, after adjustment of control variables.

These studies have established clearly that psychosocial factors can influence vulnerability to infectious disease. They have been less successful in identifying psychobiological mediators, since susceptibility is not associated with catecholamines, cortisol, or various markers of immune function. However, more recent work from Cohen's group has shown that associations between psychological stress and the severity of symptoms were related to the levels of the inflammatory cytokine interleukin-6 (IL-6) production in the nasal passages following influenza administration. More severe symptoms were experienced by participants who reported higher perceived stress, and were accompanied by raised IL-6. This suggests that inflammatory mediators may be involved in symptom production in stressed individuals.

4. *Colds and Upper Respiratory Illness*

Upper respiratory illnesses are caused by many different types of microorganism. The common cold is a viral infection, and more than 200 different viruses have been implicated. Acute sore throats are commonly caused by streptococcal bacterial infection, while influenza is a viral disease. Naturalistic studies of psychosocial factors in upper respiratory illness involve assessing spontaneous illnesses in people living their everyday lives. They have shown similar effects to those identified in experimental studies, albeit with less control over confounding factors (Cohen and Williamson 1991, Steptoe and Cohen 1999). For example, a number of studies have followed volunteer populations over several months with periodic measures of life stress, coping, health behavior, and illness. It has been demonstrated in both adults and children that susceptibility to colds is greater in people who experience high rather than low life stress. Interactions between life-events stress and social support have also been observed, with lower incidence of colds among individuals reporting high support. In addition, episodes of upper respiratory infection have been shown to be preceded for one or two weeks by high levels of perceived stress or a high frequency of minor life events (hassles). Since the acquisition of an infection is followed by several days of incubation, studies of this kind suggest the pathways outlined in Fig. 1 (right panel) may be operating. However, it is not yet proved feasible to obtain detailed neuroendocrine or immunological data in naturalistic studies. The mediating mechanisms therefore remain unclear, and it is possible that some of the associations observed are due to changes in exposure to pathogens rather than psychobiological influences on bodily resistance.

5. *Psychosocial Influences on Symptoms*

The interaction between the brain and the immune system is two-way, with behavioral function being affected by immune responses as well as psychosocial factors influencing the immune system. It has been recognized over recent years that the behavioral concomitants of infectious disease such as lethargy, loss of appetite, increased sleep, and loss of interest in the outside world are stimulated in part by immune products such as interleukin-1 (McCann et al. 1998). These cytokines are released as part of the biological response to infection, and help to coordinate the pattern of behavior that conserves and protects the organism, and may thereby be adaptive in promoting survival.

In humans, psychosocial factors not only affect susceptibility, but also the duration and severity of symptoms. This has been studied in relation to upper respiratory illness. In experimental inoculation studies it has been found that symptom reporting is associated positively with perceived stress and with certain personality profiles. Patterns of psychological coping are also relevant. One study assessed the duration of

symptoms of naturally occurring colds in a sample of children (Turner Cobb and Steptoe 1998). Prior to the study, patterns of coping had been assessed with standardized measures. It was found that children who typically dealt with stress using problem-focussed coping strategies tended to suffer from colds for a relatively short time. In contrast, children who stated that they used avoidant coping strategies such as distraction and social withdrawal experienced colds of longer duration. These effects remained significant after statistical adjustment for gender, age, family composition, social class, parental smoking, and parental alcohol consumption. It is possible that those who used distraction and avoidant coping strategies failed to respond directly to their illnesses, and did not take the necessary steps in terms of self-care and activity restriction that might prevent lengthy episodes of symptomatic disease.

6. *Recurrence of Herpes Virus Infections*

Herpes simplex viruses cause a range of diseases, the most common of which are oral herpes (cold sores) and genital herpes. Different types of herpes virus cause the two diseases. Both forms of herpes infection are common and typically benign, though they may cause considerable distress. In more than half of cases, the virus is not eliminated following first infection, but becomes permanently established in sensory ganglia. In some individuals the virus is reactivated periodically, and this leads to a recurrence of infection. Reactivation is partly under immunological control, and may take place when immune surveillance is impaired by psychosocial factors.

Studies of both types of herpes infection have been carried out which suggest that, in some cases, recurrence of infection is preceded by periods of negative mood or stress. Colds also seem to be important in the triggering of recurrent oral herpes. A meta-analysis of 16 studies indicated that depressed mood was associated with recurrence of herpes simplex virus infection, while stress was not (Zorrilla et al. 1996). However, the models of stress that have been employed in this area have been relatively simple. They have evaluated the experience of stressors alone, and not the interaction between stressors and psychosocial coping resources.

Another type of herpes virus is Epstein–Barr virus (EBV), responsible for infectious mononucleosis (which may present as a sore throat). A few studies have shown that stress and negative moods are associated with this type of infection. Interestingly, there is also psychoneuroimmunological work that throws light on the role of EBV. It has been shown that antibody titers for EBV are raised by life stress, for example, during the period prior to important academic examinations, or in people experiencing marital conflict (Kiecolt-Glaser et al. 1988). A raised antibody titer suggests that EBV has been reactivated, and allowed to replicate due to failure of immune control. If unchecked, the reactivation may lead to fresh episodes of illness.

7. *Other Infections*

One important type of infection with extensive psychosocial ramifications is human immunodeficiency virus (HIV); this will not be discussed here, since it is examined elsewhere in the Encyclopedia. Other types of infectious illness such as urticaria, streptococcal throat infection, and tuberculosis have also been studied, but data are too limited to draw general conclusions. A large range of more severe infections such as poliomyelitis, rabies, and salmonella have been investigated in animals, with evidence for increased vulnerability with exposure to psychological stressors (Peterson et al. 1991). However, there are important variations with species, age, concomitant disease and prior exposure, and these make it difficult to draw simple inferences for humans.

8. *Conclusions and Future Directions*

It is likely that research on psychosocial influences on infectious disease will become increasingly important over the next decade. One reason is that it is now recognized that infection and inflammatory processes play an important role in 'noninfectious' diseases such as coronary atherosclerosis, peptic ulcer, and gastric cancer (Lorber 1996). Limited psychosocial research on these conditions from the perspective of infection has been carried out. Indeed, most research on psychosocial aspects of infection has been carried out on a limited group of infections, and much remains to be learnt about resistance to fungi, parasites, and other microorganisms. The recognition of the importance of psychosocial factors in vulnerability and adaptation to infectious disease may lead to new and more effective methods of helping people cope with these illnesses.

See also: AIDS, Geography of; Illness Behavior and Care Seeking; Mortality and the HIV/AIDS Epidemic; Psychoneuroimmunology; Sexually Transmitted Diseases: Psychosocial Aspects

Bibliography

Aboud F E 1998 *Health Psychology in Global Perspective*. Sage, Thousand Oaks, CA

Castilla R E, Sawyer D O 1993 Malarial rates and fate: A socioeconomic study of malaria in Brazil. *Social Science and Medicine* **37**: 1137–45

Cohen S, Doyle W J, Skoner D P, Rabin B S, Gwaltney J M 1997 Social ties and susceptibility to the common cold. *Journal of the American Medical Association* **277**: 1940–4

Cohen S, Tyrrell D A J, Smith A P 1991 Psychosocial stress and susceptibility to the common cold. *New England Journal of Medicine* **325**: 606–12

Cohen S, Williamson G M 1991 Stress and infectious disease in humans. *Psychological Bulletin* **109**: 5–24

Feldman P J, Cohen S, Doyle W J, Skoner D P, Gwaltney J M 1999 The impact of personality on the reporting of unfounded symptoms and illness. *Journal of Personality and Social Psychology* **77**: 370–8

Kiecolt-Glaser J K, Kennedy S, Malkoff S, Fisher L, Speicher C E, Glaser R 1988 Marital discord and immunity in males. *Psychosomatic Medicine* **50**: 213–29

Lorber B 1996 Are all diseases infectious? *Annals of Internal Medicine* **125**: 844–51

McCann S M, Sternberg E M, Lipton J M, Chrousos G P, Gold P W, Smith C C 1998 *Neuroimmunomodulation: Molecular Aspects, Integrative Systems, and Clinical Advances*. New York Academy of Sciences, New York

Peterson P K, Chao C C, Molitor T, Murtaugh M, Strgar F, Sharp B M 1991 Stress and pathogenesis of infectious disease. *Reviews of Infectious Diseases* **13**: 710–20

Rabin B 1999 *Stress, Immune Function, and Health*. Wiley-Liss, New York

Steptoe A 1998 Psychophysiological bases of disease. In: Johnston M, Johnston D (eds.) *Comprehensive Clinical Psychology, Volume 8: Health Psychology*. Elsevier Science, New York

Steptoe A, Cohen S 1999 Psychoneuroimmunology, stress and the common cold. In: Eccles R (ed.) *Mind, Diet, and the Common Cold*. International Conference and Symposium Series 259. Royal Society of Medicine, London

Turner Cobb J M, Steptoe A 1998 Psychological influences on upper respiratory infectious illness in children. *Journal of Psychosomatic Research* **45**: 319–30

Zorrilla E, McKay J, Luborsky L, Schmidt K 1996 Relation of stressors and depressive symptoms to clinical progression of viral illness. *American Journal of Psychiatry* **153**: 626–35

A. Steptoe

Inferences in Discourse, Psychology of

If listeners or readers understand discourse, they understand much more than what is stated explicitly in the sentences. On the basis of their knowledge of the language and their knowledge of the state of affairs in the world, they are able to understand what has been left implicit in the text but is intended to be communicated. In text (a),

> (a) There were municipal elections yesterday. Because the majority of the lower town voted for the local party, there was a shift toward the left in the city council. But the right-wing party was not completely disappointed. It had expected to lose much more.

readers can infer that the local party is a left-wing party; that there is a causal relation between many people voting for a particular party and that party getting a stronger representation; that there are height differences in the town and that the town is probably located on a river bank; that the city council used to be more right-oriented; that there is a contrast between the shift toward the left and not being completely disappointed, and consequently that the shift is a reason for being disappointed and that there are other reasons (worse expectations) why the right-wing party was not so disappointed; and that 'it' refers to the right-wing party and not, for example, to the city council. All these pieces of information are not stated explicitly in the text, but readers do understand them. They are called inferences.

An inference is defined as information that is not expressed explicitly by the text but that can be derived from the text on the basis of the comprehender's knowledge and that is encoded into the representation the comprehender constructs of the text. In this article, the notion of inference is restricted to the derivation of implicit information that occurs in spontaneous language processing. These inferences are distinguished from logical deductions in reasoning and problem solving (Johnson-Laird 1983, see also *Reasoning with Mental Models*), such as the logical deduction 'Some artists are chemists' from the premises 'Some beekeepers are artists. All beekeepers are chemists.' Logic is concerned with the assessment of the validity of arguments in reasoning. Although listeners and readers certainly make logical deductions, and some inferences can be considered as logical inferences, everyday language behavior is not characterized by the evaluation of the validity of the arguments in reasoning.

Two kinds of inference in everyday language behavior can be distinguished. The first kind is an inference that is the derivation of new information. This is what in ordinary language is called an inference. For example, from the sentence 'John selected his stranston shoes because there was much mud,' the reader can infer that apparently stranston is a material or brand that has advantages if there is much mud. Given that the reader is not familiar with stranston, the reader cannot know that the conjunction 'because' is correctly used. However, by assuming that the sentence makes sense, the reader can derive the inference as new information. The second kind of inference is an activation of available knowledge. Examples from text (a) are that 'it' refers to the right-wing party, and that, given the contrastive relation indicated by 'but,' a shift toward an opponent party is in general a reason for disappointment. This kind of inference in general is not called an inference in everyday language, but most of the psycholinguistic research focuses on this kind of inference.

1. *Issues in Inference Research*

There are two issues that make inferences an important topic of research in language understanding. In under-

standing a discourse, the number and variety of inferences that may be derived from the discourse seem to be almost unlimited. This may lead to a computational explosion. Yet, the human processing system has limited capacity, and comprehension is accomplished too quickly for many time-consuming inferences to be made. Therefore, the first issue is how to account for the control of inferences: Which inferences are made and which ones are not made? This issue will be discussed in Sect. 3. The second issue concerns the inference as a process. What constitutes the process of making an inference; how does the inferred information get activated and deactivated; how does the information in the text interact with the reader's knowledge? Some aspects of the process of inferring will be discussed in Sect. 4. This article focuses on inferences in reading because most of the psycholinguistic research on inference addresses reading.

2. Methods in Inference Research

The questions whether an inference is made and when it is made can be investigated by measuring specific aspects of reading behavior. The measurements are made during the reading of the words or sentences that trigger the inference (*on-line methods*) or after reading those words or sentences (*off-line methods*).

2.1 Off-line Methods

The assumption underlying off-line methods is that the inferences are encoded into the mental text representation, as is information that was expressed explicitly by the text, and therefore that they can be investigated in tasks such as reproduction, recognition, and verification. In a *reproduction* task, the reader is required to reproduce the text after having read it. Information that was not stated explicitly in the text but that is reproduced by the reader is supposed to be inferred by the reader. In a *recognition* task the reader has to judge whether particular words or sentences occurred in the text. The rationale of the recognition task is similar to that of the reproduction task. Information that is incorrectly judged as having appeared in the text is assumed to be inferred. Many people who had studied the sentences, 'John was trying to fix the birdhouse. He was pounding the nail when his father came out to watch him and to help him do the work,' incorrectly recognized the sentence 'John was using the hammer to fix the birdhouse when his father came out to watch him and to help him do the work.' This indicates that an *instrumental inference* ('hammer') was made (Bransford and Johnson 1973). Reproduction and recognition tasks can indicate whether the inferences are made but in general cannot distinguish whether the inferences are made during reading (at encoding) or during the measurement (at retrieval). In a *verification* task, readers judge whether the content of a sentence is true or false with respect to the content of the text. Not only the accuracy, but also the reaction time for the verification or the recognition can be important. If inferred information is recognized or verified as quickly as explicit information, this is an indication that the inference is made during reading, and not at the moment of the off-line task. Off-line methods are used frequently in combination with on-line methods to answer the question of when inferences are made.

2.2 On-line Methods

On-line methods are employed during reading and aim to detect the ongoing inference process immediately. The assumption is that inference processes require time. Therefore, they are detected by a long reading time at the moment the inference is made, relative to a control condition in which no inference is made. Reading times are measured in a self-paced reading task in which the reader by pressing a button exposes successive units of text (words or clauses) in a window on a computer screen. The interval between button presses is defined as the reading time for the unit of text. For example, for understanding the second sentence in 'Mary got the picnic supplies out of the car. The beer was warm,' the inference is required that there was beer among the picnic supplies. Therefore, the reading time for the second sentence in this sequence should be longer than in the sequence 'Mary got the beer out of the car. The beer was warm' (Haviland and Clark 1974). Reading times are also measured in eye-tracking recordings that reveal what the reader looks at and how long. This makes it possible to measure fairly exactly when the inferences are made. Just and Carpenter (1978) found that the gaze duration on 'killer' in 'The killer left no clues for the police to trace' is longer if this sentence is preceded by the sentence 'The millionaire died on a dark and stormy night' than if the sentence beginning 'The killer … ' comes after the sentence 'The millionaire was murdered on a dark and stormy night.' Reading the word 'killer' requires the information that the person was murdered; this information has to be inferred after 'die' and is explicit in 'murdered.'

Other methods are administered immediately after the reading of the word or sentence that triggers the inference and can be considered on-line methods. In these tasks, a word is presented as a probe during reading, at or shortly after the moment that the inference is supposedly made. The probe is related to the information that is presumably inferred. An example is the probe word 'break' after the sentence 'No longer able to control his anger, he threw the

delicate porcelain vase against the wall.' The task may be to decide whether or not the string of letters of the probe forms a word (*lexical decision*), to decide whether or not the probe word had been presented in the previous sentence(s) (*probe recognition*), or to pronounce the probe word (*naming*). The assumption is that if the inference is encoded into the text representation, the lexical decision time for the probe and its naming time are shorter and the time needed to indicate that 'break' did not occur in the text is longer than in a control condition (see Sect. 4.1) in which the inference is not made. The status of the different lexical techniques is not quite undisputed. In some investigations, it is claimed that these lexical tasks are sensitive to transient activations (i.e., inferences that are activated only for a short time) rather than to inferences. In particular, naming is sometimes considered a task that does not reflect the nature of the text representation and, accordingly, does not reflect inferences that are encoded in the text representation. In other studies it is suggested that lexical decision and probe recognition, in contrast to naming, register the fit of the probe with the preceding context rather than the encoding of inferences.

3. *Some Established Findings in Inference Research*

The issue that has stimulated much psycholinguistic research on inferences since the early 1970s is the following question: Which inferences are made and which ones are not made? This issue was investigated by looking for classifications of inferences, on the assumption that some kinds of inference are made during reading and some other kinds are not.

Most common is the distinction between *necessary* and *elaborative inferences*. Inferences are called necessary if without them the text representation is not coherent. Inferences that are not necessary for coherence are called elaborative, optional, or embellishing. Two aspects of coherence can be distinguished. *Referential coherence* is achieved by the fact that a sentence in a text deals with entities that are expressed earlier in the text. Linguistic devices to indicate referential coherence are referential (or anaphoric) expressions, such as pronouns and definite noun phrases that refer to an antecedent in the text (see *Anaphora Resolution*). *Relational coherence* is achieved by the fact that the content of a sentence has a conceptual relation with other sentences, such as a causal, contrastive, or concessive relation. Relational coherence can be expressed by conjunctions. If, in text (a), the inference about the co-referentiality between 'it' and 'party' or the inference about causality between 'voting' and 'shift' is not made, there is no coherence between the sentences. The pronoun 'it' is considered as an instruction to find a particular referent; the conjunction 'because' is considered as an instruction to find a causal relation in the context. If such instructions are not executed, the representation is not coherent. In this way, inferences triggered by referential expressions and conjunctions contribute to the coherence of the representation.

The notion of *necessary inference* is sometimes defined in a different way and contrasted with *pragmatic inference*. An inference is necessary if it follows logically from the propositions in a text. From 'John is taller than Pete and Pete is taller than Bill', it follows that John is taller than Bill. The causal inference between the propositions with 'voting' and 'shift' in (a) is not only necessary for the coherence, but also necessarily follows from the text. It can be deduced from the conjunction 'because' and is called a conventional implicature.

Inferences that do not follow logically from the text are called pragmatic inferences.

Another distinction that is made frequently is that between *backward* and *forward inferences*. Backward inferences relate the current part of the text to an earlier part. Forward inferences anticipate information that might be expressed in the subsequent text. Backward inferences contribute in general to the coherence of the text and are in that sense necessary inferences, while forward inferences are not. Backward inferences are much more restricted than forward inferences: the reader can anticipate a great number of things.

Inferences can also be classified with respect to their content. This yields a list of inferences that can be extended *ad libitum*. It includes inferences about instruments, causes, consequences, goals, time, place, and protagonists.

Established findings in inference research (Singer 1994, Van den Broek 1994, Garrod and Sanford 1994) are that necessary inferences—i.e., inferences that achieve coherence between sentences—are made during reading. These inferences are in general backward inferences. Elaborative, embellishing inferences are not made during reading; they are not required for comprehension. They are in general forward inferences. There are many possible forward inferences that can be made, and, since they do not achieve coherence, there is no motivation for making them. Inferences that can necessarily be derived from the text are far from always made; in general, it depends on whether they contribute to the coherence or not.

4. *Trends in Inference Research: Toward a More Differentiated View on Inferences*

The research discussed in the previous section can be characterized by two views on inferences. First, inferences were considered as dichotomous entities: they are made or not made. Second, inferences were described with respect to their function in the text:

whether they are made or not depends on whether they contribute to the coherence of the text. Since the late 1980s and early 1990s, there has been a gradual change in inference research. The recent trends will be discussed in two subsections. First, the dichotomous view on inferences is being replaced by a process view on inferences: What happens if an inference is made? Second, attention has shifted from the function of inferences in the text to the availability of information and knowledge as determinants of inferences, and in this way to the role of the reader.

4.1 The Process Character of an Inference

The shift in attention toward a process view on inferences was stimulated by research on forward inferences. Presumably, the reason was that backward inferences were well established and that the idea that forward inferences are not made was not quite convincing. Several studies (McKoon and Ratcliff 1986) demonstrated that forward inferences about very predictable events can be made, but are only partially or minimally encoded. After presentation of several sentences, among them, 'The director and the cameraman were ready to start shooting when suddenly the actress fell from the 14th floor,' the word 'dead' was presented as a probe for recognition. If an inference about dying were encoded in memory, it would interfere with the correct answer 'no.' When the target word was preceded by another word from the same sentence, an interfering effect did indeed occur, but when the target word was preceded by a neutral word (e.g., 'ready'), no interference occurred. This result was interpreted as evidence for a minimal encoding of the inference, such as 'something bad happened,' because a prime word from the same sentence was necessary to strengthen the match between the probe word and the memory representation.

Other research demonstrated that inferences are built up and decay over time. First, forward inferences are not generated until some time after the presentation of the words the inference is based on. Experiments in which the time between the context that primes the inference and the presentation of a target word to detect the inference is varied indicate that a delay of about 1 s is necessary for the inference to be built up (Calvo et al. 1999). This observation may explain why no evidence for forward inferences was obtained in most of the earlier studies. If, in a reading experiment, subsequent information that does not support the inference is presented very quickly, the inference is not built up. However, if readers can read at their own pace or if subsequent information supports the inference, forward inferences can be built up. Second, an inference can be made for a brief period of time. It need not remain active. Keefe and McDaniel (1993) found that the naming latency time for the word 'break' was shorter after 'One day, no longer able to control his anger, he threw a delicate porcelain vase against the wall' than after the control sentence, 'One day, unable to control his impulses, he went out and purchased a delicate porcelain vase,' when the probe word was presented with a short delay after the target sentence, but there was no facilitation when there was an intervening sentence between target sentence and probe. They argued that forward inferences are drawn but are quickly deactivated. But the likelihood that forward inferences are encoded seems to increase if the information to be inferred is foregrounded and salient, and if the inference concerns a causal relation (Klin et al. 1999).

4.2 Availability of Information and Reader's Knowledge as Determinants of Inferences

Whether inferences are made or not depends not only on the function of the inference in the *text*, but also on whether the *reader* has information and knowledge available. This has been acknowledged in the currently most pervasive theories of inferences: the *minimalist theory* and the *constructionist theories*. In the minimalist theory, language processing and inferencing are described largely in terms of bottom-up processing of the information in the text. In the constructionist theories, the reader's search for meaning and the reader's knowledge play an important role, so that understanding entails a considerable amount of top-down processing.

In their minimalist theory on inferences, McKoon and Ratcliff (1992) argue that only two kinds of inference are drawn spontaneously during reading: inferences that serve to achieve *local coherence* and inferences that are based on *easily available information*. Local coherence refers to coherence between propositions that are not farther apart in the text than one or two sentences. Information is easily available if it is expressed in the current two or three sentences or if it is well-known general knowledge, such as instances of categories (collie—dog). In this way, the availability of information in the text is defined in terms of the linear structure of the text. But, as constructionist theories point out, information that is globally relevant, such as information about the goal of the protagonist, is not only available at the moment of reading the sentence that expresses that information, but remains available throughout longer stretches of the discourse. It is higher-order information in the hierarchical text structure. This information is shown to be available in the absence of local incoherence (Myers and O'Brien 1998).

In constructionist theories, the knowledge of the reader plays a much more important role than in minimalist theory. According to constructionist theories, readers construct a *mental model* or a *situational model*. Such a model contains not only the information in the propositions of the text but also

information that is constructed by the reader, including elaborative inferences and global inferences (Johnson-Laird 1983, Garnham and Oakhill 1996, Zwaan and Radvansky 1998). Many participants (Bransford and Johnson 1973) who heard a sentence such as 'Three turtles rested on a floating log and a fish swam beneath them' recognized, incorrectly, the sentence, 'Three turtles rested on a floating log and a fish swam beneath it,' but participants who had heard the sentence, 'Three turtles rested beside a floating log and a fish swam beneath them,' did not often recognize incorrectly the corresponding sentence with 'it.' Apparently, comprehension requires the activation of knowledge about spatial relations.

The central idea in the constructionist theory of Graesser et al. (1994) is that reading is considered as a search for meaning. The readers' goal and knowledge guide the reading process. Readers try to construct a meaning representation that is coherent both at the local and at the global level. They try to explain the actions, events, and states mentioned in the text; i.e., readers try to answer 'why' questions with respect to the text. According to this theory, inferences concerning causal antecedents and superordinate goals are made because they explain actions in the text, whereas inferences about causal consequence and subordinate goals and states are not made because they do not contribute to the explanation. Currently, there is some debate on these claims of constructionist theory. What is undisputed is that the reader's knowledge is an important factor in controlling inferences. Even local inferences are not made if they do not address information that is part of the reader's knowledge (Noordman and Vonk 1992).

5. *Outlook*

Considering the state of the art in inference research, we can expect some developments. First, attention will be paid increasingly to the role of knowledge in inferences. This concerns knowledge from long-term memory, but also knowledge from previous parts of the text. Second, inference research will move in the direction of reading models. In fact, the incorporation of knowledge in inference research will emphasize inferences not so much as an isolated research topic, but as a component of the reading process. In reading research, some bottom-up models, such as construction-integration (Kintsch 1998) and resonance (Myers and O'Brien 1998), and top-down models (based on constructionism) have been developed recently. Progress in inference research will be made if inference processes can be integrated in such reading models.

See also: Comprehension, Cognitive Psychology of; Inferences, Psychology of; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Literary Texts: Comprehension and Memory; Macrostructure in Discourse Comprehension, Psychology of; Narrative Comprehension, Psychology of; Sentence Comprehension, Psychology of; Text Comprehension: Models in Psychology

Bibliography

Bransford J D, Johnson M K 1973 Considerations of some problems of comprehension. In: Chase W G (ed.) *Visual Information Processing*. Academic Press, New York

Calvo M G, Castillo M D, Estevez A 1999 On-line predictive inferences in reading: Processing time during versus after the priming context. *Memory and Cognition* **27**: 834–43

Garnham A, Oakhill J 1996 The mental models theory of language comprehension. In: Britton B K, Graesser A C (eds.) *Models of Understanding Text*. Erlbaum, Mahwah, NJ

Garrod S C, Sanford A J 1994 Resolving sentences in a discourse context: How discourse representation affects language understanding. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA

Graesser A C, Singer M, Trabasso T 1994 Constructing inferences during narrative text comprehension. *Psychological Review* **101**: 371–95

Haviland S E, Clark H H 1974 What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior* **13**: 512–21

Johnson-Laird P N 1983 *Mental Models*. Cambridge University Press, Cambridge, UK

Just M A, Carpenter P A 1978 Inference processes during reading: Reflections from eye fixations. In: Senders J W, Fisher D F, Monty R A (eds.) *Eye Movements and the Higher Psychological Functions*. Erlbaum, Hillsdale, NJ

Keefe D E, McDaniel M A 1993 The time course and durability of predictive inferences. *Journal of Memory and Language* **32**: 446–63

Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK

Klin C M, Guzmán A E, Levine W H 1999 Prevalence and persistence of predictive inferences. *Journal of Memory and Language* **40**: 593–604

McKoon G, Ratcliff R 1986 Inferences about predictable events. *Journal of Experimental Psychology-Learning, Memory, and Cognition* **12**: 82–91

McKoon G, Ratcliff R 1992 Inference during reading. *Psychological Review* **99**: 440–66

Myers J L, O'Brien E J 1998 Accessing the discourse representation during reading. *Discourse Processes* **26**: 131–57

Noordman L G M, Vonk W 1992 Reader's knowledge and the control of inferences in reading. *Language and Cognitive Processes* **7**: 373–91

Singer M 1994 Discourse inference processes. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA

Van den Broek P 1994 Question asking and answering. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA

Zwaan R A, Radvansky G A 1998 Situation models in language comprehension and memory. *Psychological Bulletin* **123**: 162–85

W. Vonk and L. G. M. Noordman

Inferences, Psychology of

From a legal perspective, inferences are easily defined by contrasting them with direct assertions. Assertions have physical representations, e.g., they exist on paper, whereas inferences do not; and asserting a falsehood can lead to a criminal conviction, whereas implying a falsehood is typically not a crime. One of the important findings of inference research has been to show that this black and white distinction between inferences and assertions does not hold for the cognitive processes involved in inferencing. This insight led to a reconceptualization of the goal of comprehension and attempts to define the computational processes involved in inferencing. The main approach has been to determine which classes of inferences are typically made in comprehension. This research has been characterized by controversy and has uncovered a number of methodological problems in assessing the occurrence of inferences. Nonetheless, it has convincingly demonstrated that inferencing is not a single cognitive skill that individuals may be more or less good at because it is very dependent on knowledge, a finding confirmed with brain imaging.

1. Is Explicit Always Easier?

Although intuition might suggest that explicit information should always be easier to process than implicit information, one of the great insights of inference research has been to show that this notion of processing ease rests on a faulty assumption about the primary goal of language comprehension. It was initially thought that the goal was to construct a representation of the meaning of the text. In this view, explicit information is easy because it is direct, whereas implicit information requires an effortful process of determining what information is missing and then supplying it. Early studies of inferencing reinforced this view that inferring implicit information is more difficult than processing explicit information. For example, Haviland and Clark (1974) found it took longer to read a target sentence when the preceding sentence only implicitly mentioned the referent of the target sentence rather than explicitly mentioned it, as in:

Implict: *Herb took the picnic supplies out of the car.*
Explicit: *Herb took the beer out of the car.*
Target: *The beer was warm.*

However, many early studies also showed that, as far as the memory representation was concerned, implicit information was functionally no different than explicit. For example, Bransford et al. (1972) found that people could not even recognize whether statements had been implicit or explicit when they could be integrated to form a story. Perhaps more remarkable was the finding that implicit information could be more accessible in memory and thus easier to verify than explicit information. Potts (1972) found that when the statements referred to a linear ordering, e.g., A is bigger than B, B is bigger than C, explicit statements were more difficult to verify than implicit statements; this is because the explicit referred to adjacent terms whereas the implicit referred to referents more distantly spaced in the ordering. These memory studies were soon supported by reading time studies that also found that explicit was not always easier than implicit. For example, Sanford and Garrod (1981) found that it took no longer to read *The clothes were made of pink wool* when it was preceded by the implicit *Mary dressed the baby* than by the explicit *Mary put the clothes on the baby*.

In order to accommodate findings such as these, the field underwent a change in its meta-assumption about the goal of comprehension. The goal for the comprehender was no longer representing the text but rather representing what the speaker intended the text to be about. This was referred to as the mental model view of discourse (Johnson-Laird 1983). In this view, the difference between what is given and what must be filled in with an inference is defined relative to the comprehender's knowledge structures, not the text. But since knowledge structures cannot be observed, the distinction is not so black and white. The advantage, however, is that this perspective allows explicit information not always to be easier. It also accommodates the fact that making information explicit is not even necessary for effective communication; for example, when two people know each other well, just a few words can function to activate the appropriate knowledge. It also accommodates the fact that being explicit can actually be objectionable; try making all the thoughts behind even a simple sentence explicit. The result is typically the kind of incomprehensible verbiage found in legal documents; incomprehensible because the words serve to detour one from the intended knowledge structures.

The mental model view better accommodated the data—both data showing that explicit is not always easier, and data showing that the meaning of a text and the ease with which it can be processed vary with the comprehender's knowledge and the context in which it occurs. However, it left the notion of what an inference is ambiguous. Because the comprehender's knowledge cannot be observed, inferences have continued to be defined as information not explicit in the text. Ease of processing, however, is now defined not by whether or not information is in the text, but rather by how closely it corresponds to the comprehender's knowledge. Depending on the par-

ticular comprehender, the computational processes of inferencing can range from simply activating existing knowledge to generating new knowledge with extensive computational effort.

2. *Computational Processes in Inferencing*

Computational models are specified by their inputs, the processes that operate on these inputs, and their outputs. There is wide agreement on the inputs to inferencing, namely textual elements and the comprehender's existing knowledge. There is less agreement on the processes and outputs. One issue concerns whether simply activating a concept that was not explicit in the text is a legitimate case of drawing an inference. For example, does activating *spoon* when comprehending *stirred the soup* constitute an inference? At issue is whether inferencing must involve constructing new knowledge. If merely activating information counts as inferencing, then one needs to distinguish between at least three levels of activation; these vary in how long the information stays activated and how easily it can be referred to by subsequent text. At the simplest level, there is information that is merely activated in long-term working memory, then there is information selected as focal in working memory, and finally information that actually gets incorporated into the long-term representation of the text (Kintsch 1998). Related questions concern how other structures in the text serve to enhance and suppress activation (Gernsbacher 1990) and whether the activation of the inferential information occurs as a function of automatic resonance or whether an effortful search of memory is required (Myers and O'Brien 1998). Finally, there is the issue of size of the output. It can range from an activated concept, such as *spoon* in the above example, to a proposition, e.g., *that one event caused another*, to a set of propositions constituting a higher-order knowledge structure, e.g., *one's schema for romantic relationships*. The problem for the field has been using a single term, *inferences*, to refer to such a broad range of knowledge structures and computational processes. As the field makes progress in specifying the computations, it is likely to begin classifying inferences according to the type of computations required (Kintsch 1998).

3. *Types of Inferences*

The past two decades of inference research have been dominated by a quest to determine which types of inferences (as defined by type of information, not type of computation) are typically drawn during comprehension. This question is central to several enterprises. One is the design of artificial intelligence language understanding systems; designers need to know which types of inferences are typically drawn so that they can get their programs to function as typical language users. Another is the study of individual differences in language and cognition. Deficits in inferencing can only be assessed by knowing which inferences nonimpaired individuals typically draw. Finally, those developing computational models of inferencing need to know which types of inferences have behavioral consequences.

Early research identified *bridging inferences*—inferences that bridge the gap between two textual elements—as typically drawn (Haviland and Clark 1974). They are also called *backward inferences* and *coherence inferences* because when a gap in coherence is detected, the comprehender works backward to build a bridge to the preceding text; to illustrate with the example from Haviland and Clark given earlier, connecting *beer* to *picnic supplies* involves making the bridging inference that the picnic supplies must have included beer. What constitutes a gap and what constitutes coherence presupposes a theory of knowledge and a theory of how the language maps on to that knowledge; although the field is far from such a theory, some of the dimensions of coherence have been specified. Research shows that backward inferences are made to maintain referential, causal, and temporal coherence (Graesser, Millis and Zwaan 1997).

Another class of inferences is *forward inferences*, also called *elaborative* or *predictive* inferences. Whereas backward inferences are made to establish coherence and are assumed to occur at coherence breaks in the text, forward inferences merely elaborate the developing representation and can occur at any point. Compelling reasons have been proposed for making forward inferences, and it is often said that more skilled readers make more forward inferences. However, there are also costs associated with making such inferences. For example, if *spoon* is inferred from *stirred the soup*, reading a subsequent sentence about *the spoon* might be facilitated by having spoon already in an active state. However, if the text proceeds to identify some other term as the instrument of the stirring, e.g., *stick*, having inferred spoon could be costly. Hence it is an interesting empirical question as to whether such forward inferences are drawn. The types of forward inferences thought most likely to occur and thus the ones that have been studied the most are instrument inferences (e.g., inferring *spoon* from *stir the soup*), inferences about the likely consequences of events (e.g., inferring *cut* from *stepped on jagged glass*), and instantiations of general terms (e.g., inferring *cow* from *milk the animal*).

In a review of this research, McKoon and Ratcliff (1992) concluded that few forward inferences seem to be drawn. They characterized the comprehender as operating according to the *minimalist hypothesis*. It states that the only inferences typically drawn are those required for local coherence and those involving activations of readily available pre-existing associa-

tions. This started quite a controversy. To many researchers, minimalism represented a return to the earlier text-centered view of comprehension. They argued for the more knowledge-based, *constructivist* view (Graesser et al. 1994), and produced several studies arguing against minimalism by showing that even when texts are locally coherent, inferences related to global coherence are still typically drawn (e.g., Myers et al. 1994). Yet to be determined, however, is the role of computational effort in making global coherence inferences. To the extent that these inferences are only made using minimal processing, such as resonance with activated knowledge structures (Myers and O'Brien 1998), the notion of the comprehender as a minimalist processor may still be valid.

4. *Methodological Issues*

The main paradigms used to assess inferencing can be divided into two classes: *activation measures* and *memory measures*. Activation measures detect inferences by determining whether the inference concept is in an activated state relative to some control condition where the inference would not be expected to be drawn. They include reading times for a sentence where the inference might be drawn, and also reaction times to make a lexical decision or name a word related to the inference. Memory measures require subjects to access their mental representations of the text and determine if the inference is part of that representation. They include cued recall, sentence verification, question answering, and recognition tests.

Much has been learned about the methodological problems associated with these measures (Keenan et al. 1990). The main problem with memory measures is that they do not allow one to determine whether the inference was made during comprehension or if it was made as a result of the test. Because memory tasks require a person to go back to their representation of the text to determine if information is present, they provide the person with another opportunity to draw the inference. This is a real problem if one seeks to know whether inferences are being drawn on-line as the text is being understood.

Activation measures have two problems. First, not everyone counts activating a concept as making an inference. Second, inference concepts may be activated spuriously by virtue of their association with other words in the text, a phenomenon referred to as word-based priming. For example, if the concept *dead* is activated by the sentence *the fashion model fell from the roof during the photo shoot*, it may either indicate that a forward inference was drawn about the model being dead from her fall or it may be that *dead* is activated because it is related to *shoot*. Thus, activation measures of inference require assessing activation relative to a sentence that controls for word-based priming.

The need to control for word-based priming points to the general problem of what the proper control condition is for assessing inferences. The commonly used procedure of comparing reading times on explicit to implicit versions may be contaminated by surface form memory facilitating performance on the explicit. Thus, implicit may be slower than explicit, not because a time-consuming inference needs to be made, but because it does not match the target sentence on surface form.

The point in time when an inference is assessed may also affect whether or not it appears to have been drawn. Activation assessed immediately following the target text segment may suggest an inference was drawn, whereas a test occurring at the end of the passage may not detect the same inference because the activation dissipates if the inference is not incorporated into the long-term representation of the text.

Because the different types of tasks used to assess inferencing can yield different results, it is important that research aimed at determining the types of inferences made during comprehension use more than one measure of inferencing, with multiple control conditions, and assessments at various points in time. Converging operations such as these can help dispel the distorted picture of inferencing that can emerge when classes of inferences are confounded with the methods used to assess them.

5. *Individual Differences in Inferencing*

The literature on individual differences in cognition typically refers to inferencing as a unitary skill. It is thought to be a skill that young children or poor readers need to develop (Yuill and Oakhill 1991), and it is thought to be a single skill that can be selectively impaired with certain types of brain damage, e.g., to the right hemisphere (Beeman and Chiarello 1998). While it is recognized that there exists a range of skill in inferencing, this range is typically attributed to other factors, such as working memory capacity. Thus, as one's working memory capacity increases, so does one's inferencing skill, because with greater working memory capacity the inference process can connect more widely separated segments of text. This view of inferencing as a single process does not fit with the range of computational processes in inferencing described above. Nor does it recognize the fundamental reliance of inferences on specific knowledge structures. People cannot be categorized as generally good or poor at inferencing because the occurrence of inferences depends on having appropriate knowledge (Noordman and Vonk 1992). Furthermore, brain damage may not impair a general skill of inferencing so much as impair select knowledge structures, resulting in an inference deficit only in a certain knowledge domain. For example, autistic patients can

make inferences about the physical causes of events but not social causes (Happe 1994).

Brain imaging studies of discourse processing are currently in their infancy. These studies contrast connected discourses to single sentences and have shown increased activation in the temporal lobes and the posterior cingulate cortex for discourses. Most interesting, however, is the finding that the brain areas involved in discourse processing vary as a function of the knowledge domain tapped by the inference processes (Fletcher et al. 1995). Such brain imaging results support the richly complex view of inferencing that has emerged from behavioral studies.

See also: Causal Inference and Statistical Fallacies; Inferences in Discourse, Psychology of; Language and Social Inferences

Bibliography

Beeman M, Chiarello C (eds.) 1998 *Right Hemisphere Language Comprehension*. L. Erlbaum, Mahwah, NJ

Bransford J D, Barclay J R, Franks J J 1972 Sentence memory: a constructive vs. interpretative approach. *Cognitive Psychology* **3**: 193–209

Fletcher P C, Happe F, Frith U, Baker S C, Dolan R J, Frackowiak R S J, Frith C D 1995 Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition* **57**: 109–28

Gernsbacher M A 1990 *Language Comprehension as Structure Building*. Erlbaum, Hillsdale, NJ

Graesser A C, Millis K K, Zwaan R A 1997 Discourse comprehension. *Annual Review of Psychology* **48**: 163–89

Graesser A C, Singer M, Trabasso T 1994 Constructing inferences during narrative text comprehension. *Psychological Review* **101**: 371–95

Happe F G E 1994 Communicative competence and theory of mind in autism: a test of relevance theory. *Cognition* **48**: 101–19

Haviland S E, Clark H H 1974 What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior* **13**: 512–21

Johnson-Laird P N 1983 *Mental Models*. Cambridge University Press, Cambridge, UK

Keenan J M, Potts G R, Golding J M, Jennings T M 1990 Which elaborative inferences are drawn during reading?: a question of methodologies. In: Balota D A, Flores d'Arcais G, Rayner K (eds.) *Comprehension Processes in Reading*. L. Erlbaum, Hillsdale, NJ

Kintsch W 1998 *Comprehension*. Cambridge University Press, Cambridge, UK

McKoon G, Ratcliff R 1992 Inference during reading. *Psychological Review* **99**: 440–66

Myers J L, O'Brien E J 1998 Accessing the discourse representation during reading. *Discourse Processes* **26**: 131–57

Myers J L, O'Brien E J, Albrecht J E, Mason R A 1994 Maintaining global coherence during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**: 876–86

Noordman L G M, Vonk W 1992 Reader's knowledge and the control of inferences in reading. *Language and Cognitive Processes* **7**: 373–91

Potts G R 1972 Information processing strategies used in encoding of linear orderings. *Journal of Verbal Learning and Verbal Behavior* **11**: 727–40

Sanford A J, Garrod S C 1981 *Understanding Written Language: Explorations of Comprehension Beyond the Sentence*. Wiley, New York

Yuill N, Oakhill J 1991 *Children's Problems in Text Comprehension: an Experimental Investigation*. Cambridge University Press, Cambridge, UK

J. M. Keenan

Infertility: Demographic Aspects

Infertility was, until recently, the neglected problem of reproduction, overshadowed by concerns about unwanted or unplanned pregnancies and by attention to rapid population growth. Yet for women, men, and couples who are not able to have the children they want, the problem looms large. In 1994, the United Nations International Conference on Population and Development adopted a Program of Action that included the following definition: 'Reproductive health is a state of complete physical, mental and social wellbeing, and not merely the absence of disease or infirmity, in all matters relating to the reproductive system and its processes. Reproductive health therefore implies that people are able to have a satisfying and safe sex life and that they have the capability to reproduce and the freedom to decide if, when, and how often to do so' (United Nations 1994). For the first time, policy emphasis was given to the capability to reproduce. This article first defines infertility and discusses its measurement, and then offers an overview of the extent of infertility and its consequences and treatment. Noncontracepting and contracepting populations are dealt with separately.

1. Definitions and Measurement

For many years, couples who had 12 months of unprotected intercourse without conception were considered infertile. Yet, it is well known that significant proportions even of young, fecund couples do not conceive in that period of time and do so later without any treatment. The World Health Organization (1975, 1993) recommends that a two-year period without conception be used to define infertility. These definitions are appropriate for identifying couples who may want to seek medical treatment.

Social scientists have adopted a definition based on the inability of a noncontracepting sexually active woman to bear a liveborn child (Pressat and Wilson 1985). The clinical definition is based on inability to conceive. The demographic definition is broader, including both inability to conceive and inability to carry a pregnancy to term. Primary infertility is defined

as the inability to have one child, while secondary infertility occurs when a woman who has borne at least one child is unable to have another. As in most fertility analyses, the woman is emphasized for two reasons: it is far easier to identify the mother of a child than the father, and reproductive capacity diminishes with age far more rapidly for women than for men. It should be noted that it is only rarely possible to distinguish between infertility of the *woman* and infertility of the *couple*.

The level of infertility in a population is usually measured by the proportion of reproductive age women who are infertile and the age-specific proportions infertile (e.g., the proportion infertile at age 20, 25, etc.). Some analyses have been based on asking women in a survey whether or not they have experienced trouble having a wanted birth, and whether they have tried for at least two or three years to bear a child. This method has the advantage of collecting direct information from women about the problems they have encountered.

Another approach uses records of a woman's age at marriage and at each successive birth and can only base infertility estimates on the demographic definition, since no information on difficulties in conceiving is available. The proportion of women who do not have a birth within two years of marriage or of the previous birth is much higher than the proportion who do not have another birth (Trussell and Wilson 1985, Larsen and Menken 1989). Lengthy birth intervals are common in populations in which babies are breastfed, especially without supplementation, for many months. Larsen and Menken (1989) recommend counting a woman as infertile only after five years without a birth. They refer to a woman who has a child and does not have another birth within five years as *subsequently infertile*, beginning the month after the last birth. The proportion of women subsequently infertile at age 15, 20, 25, etc. offers a good estimate of the age pattern of infertility in the population. Note that although we refer to *women*, we actually have in mind *couples* classified by the woman's age.

These methods are most appropriate for populations in which there is little use of contraception or other means of fertility control. Demographic data give us no opportunity to demonstrate whether or not those who control fertility through contraception or abstinence are infertile. If a contraceptive user goes five years without a birth, it may be because she is an effective user; alternatively, she may be infertile. Estimates of infertility that assume all contraceptive users are fertile will be too low. Some analysts, therefore, have based their estimates only on those who are *not* using contraception. But couples who control their fertility may be doing so because they have all the children they want. They may be the most fertile group in the population.

If infertility is measured only from those who are *not* contracepting, the estimate will be too high, since non-users are likely to include those who have not yet achieved the number of children they want, perhaps due to their infertility. Simulation studies offer the possibility of assessing how sensitive the infertility estimates are to the proportion of contraceptive users in the population. Larsen (1994) found that estimates were robust to contraceptive use by about five percent of women.

2. *Noncontracepting Populations*

Primary infertility affects only a very small proportion of women. Bongaarts and Potter (1983) suggest that fewer than 3 percent are biologically incapable throughout their lives of ever having a child. Among the Hutterites, a noncontracepting, high-fertility American community, infertility rises only modestly until age 35 and more rapidly thereafter. Table 1 shows the percent subsequently infertile among Hutterite women in the middle of the twentieth century. Infertility was quite low for young women, e.g., just 3 percent for 20–24-year-olds. Larsen (2000) estimated infertility for 28 countries that conducted World Fertility Surveys or Demographic and Health Surveys between 1978 and 1997 in Africa, the region with the highest known infertility in the world. Even the low-infertility countries, exemplified by Rwanda and Togo in Table 1, have more infertility than the Hutterites. The African increase in infertility with age is in part due to biological aging, but most must be attributed to disease.

The most striking aspect of infertility in Africa is its wide variation, which can occur within as well as between countries. High rates of secondary infertility occur in a band of countries across sub-Saharan Africa. In six countries that Larsen studied, over 30 percent of women aged 30–34 are infertile. She estimates that women in these countries average 1.7 fewer children than they would in the absence of reproductive problems. High infertility in this region was first noted by Retel-Laurentin nearly 40 years ago in parts of the Central African Republic. These high rates are attributed to very early initiation of sexual intercourse, and sexually transmitted diseases (STDs). The primary infertility-causing diseases appear to be gonorrhea and chlamydia, which can, if untreated or treated improperly, lead to occlusion of the Fallopian tubes, thereby preventing ova from descending from the ovary to the uterus.

Evidence is accumulating that infertility today may be related to HIV infection. HIV-infected women, even if they are asymptomatic, have lower fertility rates than non-infected women. This finding raises great concern about the accuracy of estimates of HIV infection, since one basis for such estimates is HIV testing of pregnant women who come to clinics for prenatal care. If HIV-positive women are less likely to come, the estimates of proportions infected are too low.

Table 1
Levels of infertility in selected populations

	Age						
	15–19	20–24	25–29	30–34	35–39	40–44	20–44
Hutterites							
mid-20th century[1]		3	4	6	10	31	
Contemporary Africa—High infertility[2]							
Cameroon 1991		15	23	35	53(50)	75	8
Central African Republic 1994–5		17	26	38	55	76	29
Contemporary Africa—Moderate infertility[2]							
Niger 1992		9	15	25	42	63	17
Contemporary Africa—Low infertility[2]							
Rwanda 1992		2	5	9	21	49(46)	7
Togo 1988		3	6	9	20	47	7
Contemporary industrialized countries							
USA 1982[3]	4		10		12		
USA 1995[3]	6		11		13		
France 1988[4]			6	6	6	6	

1 Percent subsequently infertile. Source: Larsen 1994 2 Percent subsequently infertile. Source: Larsen 2000. Numbers in parentheses indicate estimated percent infertile when all contraceptive users are assumed to be fertile and the difference from the original estimate is at least three percentage points 3 Percent with impaired fertility, age groups 15–24, 25–34, 35–44. Source: Chandra and Stephen 1998 4 Percent not managing to conceive, women aged 25+. Source: Leridon 1992

One explanation proposed for this high infertility is that Arab slave traders spread STDs to indigenous populations of the areas in which they worked. Another explanation has to do with migrant labor, which was concentrated in urban areas and, in many places, involved primarily men. Long family separations, it is suggested, led to a rapid rise in prostitution and STDs, which were then carried back to wives in rural villages. This theory has also been applied to the spread of HIV/AIDS. In addition, the multiple sex partners entailed by the practice of polygamy in much of Africa promotes the spread of any STD that is brought into the marriage.

Rightly or wrongly, infertility in Africa, as in most places, is usually attributed to the woman, and consequences are greater for women than men. Infertile women may garner less respect in their communities, face higher rates of divorce and polygamy (as husbands take additional wives in the hope of having more children), and higher risks in old age of having few or no children to help support them.

Much of infertility in contemporary developing countries is preventable through postponement of sexual initiation until full female maturity, preventive measures such as condom use and few sex partners, care during pregnancy, and prompt treatment for STDs before pelvic inflammatory disease (PID) and scarring occur. Yet even without the burden of HIV/AIDS, health systems are rarely able to provide treatment, and the HIV/AIDS pandemic has reversed all the gains in life expectancy that were achieved in many African countries in the latter part of the twentieth century.

3. *Contracepting Populations*

As mentioned above, infertility is difficult to measure in countries in which fertility control is used extensively. In 1995, 27 percent of all women and 41 percent of married women of reproductive age in the USA had chosen sterilization, either of the woman or the man, to terminate childbearing. Fertility surveys showed impaired fecundity (defined as three years without bearing a wanted child and trouble reported in childbearing) affected 6 percent of women aged 15–24, 11 percent of 25–34-year-olds, and 13 percent of 35–44-year-olds. Especially in the oldest age group, the proportion is lower than the actual proportion infertile because of extensive contraceptive use and sterilization, which mask infertility (Chandra and Stephen 1998). A French survey conducted in 1988 showed that only just over 6 percent of women who wanted another child did not manage to become pregnant after at least several years of trying to do so (Leridon 1992).

Yet there are persistent fears that infertility in industrialized countries is higher than previously believed and even is on the rise. In part, such fears were catalyzed by a French study from 1982 of conception rates among women coming for artificial insemination.

The study showed a sharp drop-off after age 30, and that was widely interpreted as indicating a similar drop-off for all women. An accompanying editorial suggested that women who were delaying childbearing to establish a career would have to alter their plans and have their babies early if they were to have them at all (Menken et al. 1986). The US data for the 1970s and 1980s did not show any change, but the most recent fertility survey (Chandra and Stephen 1998), which is from 1995, indicates a higher proportion of women who report impaired fecundity. However, over that time there was a much greater increase in fertility among older women. In 1982, there were 21.2 births per 1,000 women aged 35–39; the figure increased by over 60 percent in 1995, to 34.3 births per 1,000 women. For women aged 30–34, the birth-rate increased almost 30 percent. Clearly, a higher proportion of women were trying to have a child in 1995 compared to 1982. Therefore a much higher proportion was not using contraception, which can mask infertility problems. Yet, the rise in the proportion reporting impaired fecundity was not much higher—13 percent for the oldest age group in Table 1 and 11 percent for women aged 25–34.

Several potential causes of a real increase in infertility include increases in abortions and STDs, and a decline in semen quality. A number of studies have found no confirmation of the notion that women who have induced abortions in modern clinics or hospitals suffer increased infertility. In about 20–40 percent of STD cases, women who have gonorrhea or chlamydia develop PID and about 20 percent of women with PID become infertile. The proportion of women who reported they had PID declined from 14 percent in 1982 to 8 percent in 1995, but some fear that this decline is not real but is due to more 'silent' infections that women do not recognize. For over 25 years, there have been suggestions that the quality of human sperm has been declining, perhaps as the result of environmental changes. An analysis of over 60 studies published between 1938 and 1990 (Carlsen et al. 1992) concludes that the decline is genuine and speculates that it may have reduced male fertility. As yet, there is no confirmation of this result, but concern remains, both about the effects of STDs and of environmental effects on semen.

But the fears appear to be greater than warranted by these concerns. Several explanations why infertility may be overestimated in the media and by the public have been suggested (Aral and Cates 1983, Menken et al. 1986, Bongaarts 1982). First, as suggested earlier, infertility was previously overshadowed by problems of controlling *unwanted* fertility. More women and couples who want no more children are using more effective contraception and sterilization, thus reducing the numbers of unwanted births. A secondary effect of this change may be that fertility appears to be under individual control. There is an expectation that conception should occur quickly once individuals decide they want to have a child. Leridon (1992) refers to this phenomenon as 'impatience to conceive.' A second reason may be that the medical profession now sees a higher proportion of its clientele made up of those who *are* having infertility problems. But the explanation for this change more likely lies in the fact that those who have had the number of children they want, have chosen methods of fertility control that remove the necessity for visits to specialists in childbearing, leaving the doctors' offices and clinics to those in need of infertility treatment. A third reason is the delay in childbearing that means that more women are hoping to have a child at older ages when, indeed, their fecundity may have declined, without any increase in infertility among women their age. Finally, the silence concerning infertility has been broken and the ways to handle infertility have changed, drawing more attention to infertility issues than before. Adoption of children who clearly are not the biological child of one of their adoptive parents has increased, making visible the choices their parents have made. High-tech treatment of infertility is routinely discussed and used by parents in the hope of producing a wanted child, and the media publicizes the successful results. The increased attention in no way means that the incidence of infertility has changed.

4. *Summary*

Infertility from senescence increases gradually with age. Variation is quite local and is influenced by disease patterns, especially of STDs. In contracepting populations, the perception of infertility varies with the proportion of people who want only a few children and stop after having them. Those who are infertile are, therefore, more likely to be among those attempting to have their first or additional children, thus increasing the perception of the prevalence of infertility. Clearly those who postpone childbearing until their late 30s are at an increased risk of infertility. In much of the world the most pressing need is for prevention and treatment of STDs that can prevent people from becoming parents of the number of children they desire.

See also: Fertility: Institutional and Political Approaches; Fertility: Political and Political–Economic Perspectives; Gender and Reproductive Health; Marriage; Reproductive Medicine: Ethical Aspects; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations

Bibliography

Aral S, Cates W Jr 1983 The increasing concern with infertility: Why now? *Journal of the American Medical Association* **250**(17): 2327–31

Bongaarts J 1982 Infertility after age 30: A false alarm. *Family Planning Perspectives* **14**(2): 75–8
Bongaarts J, Potter R G 1983 *Fertility, Biology and Behavior: An Analysis of the Proximate Determinants*. Academic Press, New York
Carlsen E, Giwercman A, Keiding N, Skakkebaek N E 1992 Evidence for decreasing quality of semen during past 50 years. *British Medical Journal* **305**: 609–13
Chandra A, Stephen E H 1998 Impaired fecundity in the United States: 1982–1995. *Family Planning Perspectives* **30**(1): 34–42
Larsen U 1994 Sterility in sub-Saharan Africa. *Population Studies* **48**: 459–74
Larsen U 2000 Primary and secondary infertility in sub-Saharan Africa. *International Journal of Epidemiology* **29**: 285–91
Larsen U, Menken J 1989 Measuring sterility from incomplete birth histories. *Demography* **26**: 185–202
Leridon H 1992 Sterility and subfecundity: From silence to impatience? *Population: An English Selection* **4**: 35–54
Menken J, Trussell J, Larsen U 1986 Age and infertility. *Science* **233**: 1389–94
Pressat R, Wilson C 1985 *The Dictionary of Demography*. Blackwell, New York
Trussell J, Wilson C 1985 Sterility in a population with natural fertility. *Population Studies* **39**: 269–86
United Nations 1994 Programme of action of the 1994 International Conference on Population and Development. (A/CONF.171/13). Reprinted in *Population and Development Review* **21**(1): 187–213 and **21**(2): 437–61
WHO 1975 The epidemiology of infertility. Report of a scientific working group. *WHO Technical Report Series*, no. 582, Geneva
WHO 1993 *WHO Manual for the Standard Investigation and Diagnosis of the Infertile Couple*. Cambridge University Press, Cambridge, UK

J. Menken and U. Larsen

Influence: Social

When an individual's attitude, opinion, cognition, or behavior is forcefully developed or changed through physical sanctions or punishment, the process at play is called 'power.' However, when the creation or change is 'voluntary' or 'unconscious' the process is called 'influence.' The term 'social influence' in the broadest sense refers to all influences that individuals receive from all types of formal institutions, other individuals, and informal groups.

Formal institutions include schools, mass media, and business corporations. This article focuses on the influence of mass media as an example of the influence of formal institutions, before going on to consider personal influence and informal group influence.

1. *Mass Media Influence*

While the influences of home and school are dominant during childhood, the influences of the mass media become increasingly more pervasive after individuals reach adolescence. When radio surged in popularity among the masses during the 1930s it was feared that mass communications, especially radio, had a powerful and instant effect on a population's attitude and behavior. This fear was reinforced by the success of the Nazis' propaganda machine in Germany and the large-scale panic in the USA in 1938 caused by a single radio program called 'The Invasion from Mars' (Cantril et al. [1940] 1966). This radio play, broadcast without any prior warning by Orson Wells's Mercury Theater on CBS, was so realistic and caused such widespread panic that some people actually thought Martians had landed on Earth and sadly committed suicide. As a result, the apparent immediate influence of mass media came to be compared to a fast-acting hypodermic injection or an atomic bomb.

Empirical research on the influence of mass media, however, gradually revealed that it could not be that overwhelmingly influential. It was found that although mass media might reinforce people's existing attitudes and opinions, or even create new ones, they could hardly convince people to change already strongly held beliefs (Klapper 1960). It was also found that even though mass media information might reach each individual directly, its influence did not. Influence, if any, was thought to reach individuals through so called 'opinion leaders' (Katz and Lazarsfeld 1955). This view of the mass media's weak influence, however, was later challenged by several new 'powerful media' theories.

One of these theories deals with the 'agenda-setting effect' of mass media (McCombs and Shaw 1972). Even if the mass media's persuasive power is limited, especially in the short run, they exert a strong influence on people's perception of the salience of each issue. There are 'good news' and 'bad news' stories for everybody and for every organization. If the mass media persistently play up the 'bad news' events surrounding some political candidate or party it will affect the readers' or viewers' attitudes, and thereby the result of an election. In such a case, the mass media are not actually persuading their readers or viewers to change their minds; they are merely taking advantage of conventional social norms and reinforcing existing attitudes. Nevertheless, mass media can influence individual behavior (such as voting) by setting the priority of people's attention in regards to the issues that are being debated.

Another important role of mass media is that they tell their readers or viewers what other people are thinking. In other words, mass media inform their audience what the prevailing 'public opinion' is regarding each issue. In doing so, mass media tend to amplify the opinions they support. In an extreme case the mass media might treat a minority opinion as if it were that of the majority or some new ineluctable trend. As a result, the holders of the minority opinion, supported by the mass media, become louder and more strident while those of the majority opinion lose

their voice and fall increasingly more silent. Thus, through the 'spiral of silence' mechanism orchestrated by mass media, the majority opinion comes to be replaced by that of the minority (Noelle-Neumann 1984).

2. *Personal Influence*

The 'opinion leader' mentioned above refers not only to professional opinion makers such as journalists, critics, and university professors but to all men and women who influence others around them on a daily basis, in some specific area, such as local politics, fashion, sports, travel, medical care, or even cooking. These opinion leaders are often asked for advice by people who look up to them. To meet such expectations, they expose themselves to the mass media more extensively than their followers. Opinion leaders compare and evaluate opinions and information that is first presented by the mass media. As a result, the influence of mass media first takes place among opinion leaders who in turn regurgitate their latest batch of news to their followers in the form of guidance or advice. Even if a person's opinion is shaped by the mass media, that opinion may be quickly replaced by that of an opinion leader because such people's opinions are trusted more by their followers than are those of the mass media. This process is called the 'two step flow' of social influence (Lazarsfeld et al. [1944] 1968).

Another explanation for this 'personal influence' is provided by the 'cognitive consistency theory.' According to this theory, people are affected by those whom they like because the cognition that one likes someone and the cognition that one's opinion is different from theirs is psychologically inconsistent and therefore uncomfortable. In such a case, it is predicted that people would change their opinion of, or attitudes toward, the person in question (Festinger 1957, Heider 1958, Newcomb 1953).

3. *Group Influence*

It is known that people tend to conform to the group that they belong to, participate in, or even just happen temporarily to be part of. In other words, the group tends to influence its members. There are several reasons for this conformity. One of them is that other people's behavior in the same group often functions as a stimulus or reminder. For example, if one sees other group members starting to eat, one feels hungry. Another reason is 'fear of isolation.' Not only humans but also many animals know almost instinctively that it is safer to be with the group than to be isolated. The atavistic 'fear of isolation' may be primitive and instinctive but makes sense. It is highly probable that to be isolated costs much more in terms of survival than it would to behave as a group member because, generally speaking, the group has far more information, experience, knowledge, and wisdom than each individual member. There really is safety in numbers.

It is for this reason that each individual usually has several 'reference groups.' 'Reference groups' are similar to 'opinion leaders.' Individuals are loyal to the norms, opinions, and inclinations of their reference groups. These reference groups may be their families, companies, local communities, or governments. Sumner's (1906) notion of an 'ingroup' or 'we-group' is a typical example of the 'reference group.' While individuals are loyal to the 'ingroup' or 'we-group' they tend to be indifferent or sometimes hostile to the 'outgroup,' 'they-group,' or 'othersgroup.'

Some Japanese sociologists and social psychologists think that what functions as social pressure or constraint on each individual is caused by an abstract social construct called *seken* rather than the influence of specific groups. According to them, individuals are conscious of three different kinds of 'societies.' The first is *miuchi* or *uchi*, which originally meant immediate family but can be applied to close relatives and other immediate-familylike small groups. The second is *tanin* or *soto*, meaning 'others,' 'outsiders,' or 'strangers.' This is not the same as Sumner's 'outgroup' because *tanin* is not by any means recognized as a group or a collection of groups. People may be indifferent to *tanin* but are not hostile to it (unlike Sumner's 'outgroup'). Finally, *seken* literally means 'world' or 'society.' Specifically, this is the area between the 'immediate family' and the 'others.' This area consists of people whom you know and who know you (other than immediate family members). Friends, neighbors, acquaintances, and relatives are examples.

An important concept used to explain social pressure and constraint is that of shame. Before whom do people usually feel ashamed? People do not feel ashamed before strangers. People do feel ashamed before their immediate family members, but they know that family members will eventually understand and forgive them. It is of *seken* that people are most afraid. *Seken* will remember failure, scandal, or misconduct for a long time and will not forgive as easily as immediate family members might. *Seken* may seem to refer to 'reference groups' but these are different because *seken* is not a group or a collection of groups and people are not always loyal to *seken*.

The same is true of 'keeping face.' Where do people 'lose face'? It is in *seken*. Therefore, individuals are always concerned with *sekentei*. There is no equivalent single word in English, but *sekentei* refers to how your friends, neighbors, acquaintances, and relatives think of you. Thus, according to Japanese sociologists and social psychologists, *sekentei* is a powerful source of social pressure and a ubiquitous constraint on individual attitudes and behavior (Inoue 1977). This theory may be more applicable to collectivistic soci-

eties (where human relations are more important) than to individualistic societies. But even in individualistic societies people do have a sense of 'shame' and 'face,' so this Japanese theory should be helpful in explaining, at least in part, the causes and effects of social influences in individualistic societies too.

See also: Cognitive Dissonance; Mass Communication: Empirical Research; Networks: Social; Opinion Formation

Bibliography

Cantril H, Gaudet H, Herzog H [1940] 1966 *The Invasion from Mars: A Study in the Psychology of Panic*. Princeton University Press, Princeton, NJ

Festinger L 1957 *A Theory of Cognitive Dissonance*. Row, Peterson, Evanston, IL

Heider F 1958 *The Psychology of Interpersonal Relations*. Wiley, New York

Inoue T 1977 *'Sekentei' no Kozo* [*The Structure of 'Sekentei'*]. Nihon Hoso Shuppan Kyokai, Tokyo

Katz E, Lazarsfeld P F 1955 *Personal Influence*. Free Press, Glencoe, IL

Klapper J T 1960 *The Effects of Mass Communication*. Free Press, Glencoe, IL

Lazarsfeld P F, Berelson B, Gaudet H [1944] 1968 *The People's Choice: How the Voter Makes up his Mind in a Presidential Compaign*. Columbia University Press, New York

McCombs M E, Shaw D L 1972 The agenda-setting function of mass media. *Public Opinion Quarterly* **36**(2): 176–187

Newcomb T M 1953 An approach to the study of communicative acts. *Psychological Review* **60**: 393–404

Noelle-Neumann E 1984 *The Spiral of Silence: Public Opinion—Our Social Skin*. University of Chicago Press, Chicago

Sumner W G 1906 *Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores and Morals*. Ginn, Boston

Y. Ito

Informal and Underground Economics

There is a widespread feeling that a substantial and increasing share of value-creating activities take place outside the official economy. This holds, in particular, for developing and transition economies, but also for high-income economies. Such activities are not recorded by the system of national income accounting, which has become the accepted standard in all countries of the world.

The existence and increase of an underground economy gives rise to four major sets of concerns. The economic and social conditions of individuals, households, and countries are evaluated in a biased way if one relies on the official statistics. Thus, the official number of unemployed persons may hide that an (unknown) share of them actually work and receive wage income. As a consequence, the macroeconomic policies are likely to be too expansionary and social policy too excessive. A second concern is the loss of tax revenue, as underground activities escape taxation. A third concern interprets the underground economy as an indicator of an unhealthy relationship between citizens and government. The taxpayers are dissatisfied with the public services they get for their contributions and seek to redress the balance by escaping to the underground economy. It is feared that such a reaction makes a government unable to finance the public goods necessary for an economy and society. A fourth concern goes even further, and fears that an underground economy undermines the cement of society because it favors everyone acting in his or her own interests, while solidarity is renounced. In contrast to these concerns, the existence of an underground economy has also been seen in a positive light. Opponents of extensive state activity welcome the limits imposed on taxation and regulation, and applaud the dynamism and flexibility in the underground sector.

1. Denominations and Definitions

The phenomenon is known, and has been discussed in literature, under many different names: informal, unofficial, irregular, parallel second underground, subterranean, hidden, invisible, unrecorded, and shadow economy or moonlighting. In several languages, the term most often used is 'black economy' ('le travail au noir' in French, 'Schwarzarbeit' in German, and 'svarta sektor' in Swedish).

No single definition exists, but it depends on the purpose. The most precise and predominantly used definition seeks to relate the underground economy to officially measured national income: it comprises all presently not recorded productive (i.e., value-adding) activities which should be in the gross national product (GNP). This definition allows us to compare and to add the underground economy to GNP.

This definition excludes two major activities:

(a) Production that by convention is not part of GNP, in particular private household activities. The evaluation of its size has engendered a research area of its own. Depending on the approach and measurement technique in developed economies, the household sector comprises between 30 percent and 50 percent of GNP.

(b) Tax evasion is not value-adding but redistributional and is therefore not included as such in the above definition (e.g., when taxes on interest payments are evaded). However, in general, no taxes are paid on underground activities (such as moonlighting for house building) which are value-adding. Thus, underground activities and tax evasion are related, but certainly not identical. Tax evasion has also become a research area of its own. For the United States,

between 1973 and 1992 for example, it has been estimated that 17 percent of total taxes owned have been evaded.

The underground economy should not be identified with illegality. Some activities are perfectly legal but are not subject to taxes (e.g., because of their small size), and therefore escape measurement in official statistics. Other activities are legal as such, but taxes are evaded. Finally, on illegal activities (such as drug production and distribution), no taxes are paid.

Broader definitions of the informal and underground economy, depending on their purpose, include private household production and redistributional activities. In the following, the more narrow definition considering unrecorded productive activities will be focused on.

2. *Measurement Approaches*

The major effort of economists has been to measure the size of the underground economy relative to official GNP; its size compared to other countries, and its development over time. The other social sciences, especially sociology, have put more emphasis on the structure: who participates where? The respective studies often provide highly detailed analyses of particular regions, industries, and types of underground activities. A considerable part (it varies strongly from country to country) of informal activity is undertaken as a second job and part-time. This allows underground workers to be eligible for social security benefits and makes detection more difficult. Not surprisingly, casual work is dominant (if not the only kind of work) for illegal immigrants. In developed economies, casual work is most prevalent in the following sections: in agriculture (especially seasonal labor), in construction, and in the services sector, particularly in repairs (automobiles and other) and cleaning, in hotels and restaurants, as well as in wholesale and retail trade. Accordingly, among the occupations with most illegal work are cleaning persons, waiters, people working at home, construction and mechanical workers.

In the 1980s and 1990s, the quest to measure the size of the underground economy led to interesting methodological innovations, i.e., the challenge to 'measure the invisible' resulted in new measurement methods. Three general sets of measurement approaches can be distinguished.

2.1 *Direct Approaches*

An obvious way to analyze the underground economy is to undertake surveys among (supposed) suppliers and demanders of casual services. This method allows us to gather a detailed picture of the structure of this sector. However, because a substantial part of these activities is illegal, the interviewees may not be prepared to disclose their involvement. Somewhat surprisingly, research suggests that disclosure is not a serious problem. Typically, men are more involved in the underground economy than women, the young more than the old, people without work more than employed people, and the most prominent sectors are construction and all kinds of services. In contrast, where substantial capital assets (which are visible) are needed, less underground activity is observed.

A second direct approach is based on auditing of tax returns undertaken by tax collection and social security administrations. A sample of tax-payers is scrutinized in depth and under threats of sanctions for failing to collaborate. This approach provides detailed information on the biggest evaders, in particular the self-employed, who have better opportunities for concealment. Capital income also lends itself easier to evasion than labor income. Tax audits have proved unable to reveal all tax evasion, and are limited to taxable activities. As a large part of tax evasion is redistributional, that part is not relevant to the narrow definition of the underground economy.

The two direct methods (surveys and tax auditing) have the disadvantage of permitting point estimates only. Moreover, it is unlikely that they capture all 'shadow' activities, so they can be seen as providing lower bound estimates. They are unable (at least at present) to provide estimates of the development and growth of the shadow economy over an extended period of time. They have, however, one considerable advantage—they can provide detailed information about shadow economy activities and the structure and composition of those who work in the shadow economy.

2.2 *Indirect or Discrepancy Approaches*

The underground economy is reflected in discrepancies showing up in various markets. Persons working in the unofficial sector are able to *spend* more than their officially recorded income. The discrepancy between the two may be observed at the level of individual households, as well as in the aggregate national accounts. This approach is questionable, as this expenditure–income discrepancy may either be due to measurement errors (which is indeed often the case) or to reasons unrelated to the underground economy (e.g., the use of credits, or reductions in wealth).

Another discrepancy may be observable in the *labor market*. A decline in official participation rates, or a low participation rate compared to other countries, may be an indication of unofficial work. But again, the discrepancy may be related to other factors. Moreover, this approach is unable to isolate those persons who are at the same time active in the official and the unofficial economy (which seems often to be the case, according to other approaches).

A third discrepancy may be visible in the *monetary market*. The dominant approach starts with the

assumption that underground transactions are paid in cash in order to make detection more unlikely. The size of the underground economy is reflected in the amount of cash used in a country beyond that used for official transactions. This approach is elegant and easily applicable because the amount of currency is well documented. The assumption that unofficial activities are transacted in cash is, however, questionable. Empirical research suggests that between 20 percent and 30 percent of the unofficial activities are not paid in cash, i.e., either by payment in kind or via a bank. This fact is especially bothersome for the currency demand approach when the share of cash payments changes over time and differs between countries. A significant portion of some currencies is held outside the country issuing it; thus the US dollar is widely used in South America and Asia. Again, the discrepancy is influenced by many factors unrelated to the underground economy, such as the use of credit cards. Finally, it is problematic to infer the size of the unofficial sector from currency transactions, because the velocity of cash circulation may differ between the official and the unofficial sector. Some of the difficulties just mentioned have been successfully addressed by more recent research. Thus, the use of credit cards and the amount of currencies outside a country have been taken into account. Most importantly, instead of comparing the actual use of cash to the one deemed necessary for the official economy, a cash demand function is empirically measured, i.e., econometrically estimated. This allows us to check for influences (such as changes in the interest rate or the increasing use of cash substitutes) unrelated to the underground economy. Moreover, the extra use of cash has been directly attributed to causal factors, in particular to an increase in the tax and social security burden. Figures for the size and development of the shadow economy can be calculated by comparing the difference between the development of currency when the direct and indirect tax burden and government regulations are held at their lowest value, and the development of currency with the current (higher) burden of taxation and government regulations. The currency demand approach is one of the most commonly used approaches.

The most recent discrepancy approach looks at *physical inputs*, in particular the use of electricity. How much electricity would normally be used to produce the official national income is calculated. The excess use can be attributed to the underground economy. This approach again has the great advantage of relying on easily available data, which is a distinct advantage for developing and transition economies. However, not all underground activities use much, if any, electricity, and the relationship between production and electricity used may change over time, or differ between countries, due to substitution and technical progress.

A general problem of all discrepancy approaches is that one has to assume a base year without underground economy. Only then is it possible to attribute the existence and rise of a discrepancy to the underground economy.

2.3 The Model Approach

This method focuses on the causes and effects of the underground economy. By constructing a model, it seeks to identify the unobserved sector in between.

Three sets of factors are taken to be the main motivators of whether to engage in the unofficial economy:

(a) The burden of taxation and social security contributions as well as government regulations provide incentives to exit the official economy. The latter hinder or even prohibit activities in the official economy (e.g., because of lacking work permits, especially for foreigners), which provides an incentive to undertake them unofficially. This holds, of course, also for illegal activities such as the production and distribution of drugs. An important reason for exiting the official sector is the restrictions imposed on working time. Low work hours, long holidays, early retirement and, above all, unemployment, are important reasons for moonlighting.

(b) The expected punishment of working in the underground economy provides a disincentive for doing so. It consists of the probability of being caught and the size of punishment. For persons outside the established society (for example, illegal immigrants), or for self-employed persons, expected punishment is lower than for other persons, and therefore a higher rate of participation in the underground sector is likely.

(c) The moral costs are another disincentive to be active in the underground economy. A good citizen has moral qualms about undertaking a forbidden activity. These moral costs are closely related to 'tax morale' which motivates citizens to pay their dues to the state.

The effects of the underground economy are reflected in the traces visible in the labor, money, and product markets discussed above. A specific econometric technique called 'unobservable variables approach' allows us to estimate the size of the underground economy. This technique seeks to identify the size of the underground sector, which cannot be observed itself, but whose causes (e.g., tax and social security burden, unemployment etc.) and whose effects (e.g., fall in the participation ratio) can be measured. This approach is the most comprehensive and builds on a well-structured behavioral model, but it requires a large amount of data. As these are often not available (in particular not for developing and transition economies), this approach is not generally applicable. The estimation technique moreover tends to be statistically fragile (i.e., small changes in specification and values of variables strongly affect the estimates).

3. Empirical Estimates

The following tables serve to indicate approximate magnitudes of the size and development of the underground economy, defined as productive, value-adding activities, which should be part of GNP. Table 1 presents a rough comparison of the size of the underground economies relative to GNP for a selection of Western European countries, Japan, and the United States for the mid-1990s, using the currency demand approach.

The South European countries (Greece and Italy) have an underground economy almost one-third as large as the officially measured GNP: followed by Spain, Portugal, and Belgium, having a shadow economy between 20 and 24 percent of (official) GNP. According to these estimates, the Scandinavian countries also have a sizeable unofficial economy (between 18 and 20 percent of GNP), which is attributed mainly to the high fiscal burden. The 'central' European countries (Ireland, The Netherlands, France, Germany, and Great Britain) have a smaller underground economy (between 13 and 16 percent of GNP), probably due to a lower fiscal burden and moderate regulatory restrictions. Smaller underground economies are estimated to exist in countries with relatively low public sectors (Japan, the United States, and Switzerland), and comparatively high tax morale (United States and Switzerland).

Table 2 provides a rough comparison of the size of the underground economy relative to official GNP for a selection of developing and transition economies for the beginning of the 1990s, using the physical input (electricity) demand approach. Some of these countries (Nigeria, Egypt, and Thailand) are estimated to have an underground sector nearly three-quarters

Table 1
Size of the underground economy relative to GNP in various European countries, mid-1990s. Estimation based on the currency demand approach

Greece Italy	27–30 percent
Spain Portugal Belgium	20–24 percent
Sweden Norway Denmark	18–23 percent
Ireland France Netherlands Germany Great Britain	13–16 percent
Japan United States Austria Switzerland	8–10 percent

Source: Compiled from Schneider and Enste (2000)

Table 2
Size of the underground economy relative to GNP in various developing and transition countries, at the beginning of the 1990s. Estimates based on the physical input (electricity) demand approach

Developing Countries	
Africa	
Nigeria Egypt	68–76 percent
Tunisia Morocco	39–45 percent
Central and South America	
Guatemala Mexico Peru Panama	40–60 percent
Chile Costa Rica Venezuela Brazil Paraguay Columbia	25–35 percent
Asia	
Thailand	70 percent
Philippines Sri Lanka Malaysia South Korea	38–50 percent
Hong Kong Singapore	13 percent
Transition Economies	
Central Europe	
Hungary Bulgaria	24–28 percent
Poland Romania	16–20 percent
Slovakia Czech Republic	7–11 percent
Former Soviet Union Countries	
Georgia Azerbaijan Ukraine Belarus	28–43 percent
Russia Lithuania Latvia Estonia	20–27 percent

Source: Compiled from Schneider and Enste (2000)

Table 3
Growth of the underground economy relative to GNP for selected West European countries and the United States, 1960–95. Estimates based on the currency demand approach (rounded figures)

	1960	1995	Percentage point increase
Sweden	2 percent	16 percent	14.5 percent
Denmark	4.5 percent	17.5 percent	13.0 percent
Norway	1.5 percent	18.0 percent	16.5 percent
Germany	2 percent	13.2 percent	11.2 percent
United States	3.5 percent	9.5 percent	6 percent
Austria	0.5 percent	7 percent	6.5 percent
Switzerland	1 percent	6.7 percent	5.7 percent

Source: Compiled from Schneider and Enste (2000)

the size of officially recorded GNP. In many countries, the size is one-quarter to one-third of GNP. In Asian countries, with a comparatively low public sector, high tax morale, or high expected punishment (Hong Kong, Singapore), the underground economy is estimated to be similar to that in many 'Northern' European countries.

Transition economies are estimated to often have substantial unofficial activities, many around one-quarter of GNP. An exception is ex-Czechoslovakia where, according to these estimates, the underground sector is clearly around 10 percent of GNP.

Table 3 reports estimates of the growth of the underground economy (relative to GNP) for selected Western countries and the United States, using the currency demand approach.

The Scandinavian countries (Sweden, Norway, and Denmark) and the German-speaking countries (Germany and Austria) exhibit a sizeable increase of the underground economy within the 35 years covered (1960–95), and tend to understate the respective effects on unemployment. Also the countries with a low share in the beginning (Switzerland and the United States) show a significant increase, with the US share more than doubled. Sizeable increases have been estimated, with few exceptions, for all types of countries and all kinds of approaches: the increasing importance of the underground relative to the official economy is a robust phenomenon. It has been attributed to many different causes, but the most important and most often cited ones are the rise in the burden of taxes and social security contributions, the increase in public regulations, especially on labor markets, as well as a long-term fall in civic virtue in the population.

4. *The Effects of the Underground Economy on the Official Economy*

In order to study the effects of the underground on the official economy, the underground economies have been integrated into macroeconomic models; this yields an extended macromodel of the business cycle, as well as tax and monetary policy linkages with the shadow economy. As a result, it becomes clear that these effects should be taken into account for tax and regulatory policies. The presence of a shadow economy tends to overstate the inflationary effects of a fiscal or monetary stimulus, and tends to understate the respective effects on unemployment. When the growth of the shadow economy and the 'official' economy are positively related (which is likely to be the case when entry costs into the shadow economy are low due to a low probability of enforcement), an expansionary fiscal policy has a positive stimulus for both the formal and informal economies. It has also been found that the US productivity slowdown over the period 1970–89 was vastly overstated, as the under-reporting of income, due to the more rapid growth of the US shadow economy during this period, was disregarded.

The underground economy is beneficiary insofar as it responds to the economic environment's demand for urban services and small-scale manufacturing. The casual sector provides the economy with dynamic and entrepreneurial spirit and can strengthen competition, increase efficiency, and put effective limits on government activities. The casual sector contributes to the creation of markets, increases financial resources, and transforms the legal, social, and economic institutions necessary for accumulation. Moreover, a substantial part (up to 70 percent) of the earnings gained in the shadow economy is quickly spent in the official sector and thus boosts the official economy. These expenditures tend to raise consumer expenditures as well as (mostly indirect) tax revenues.

5. *Policy Consequences*

The growth of the underground economy over the few last decades and its effect on the official economy, both in general perception and scholarly research, has prevailingly been evaluated as a negative development,

which should be counteracted. In particular, politicians and public officials have pointed out that the state's capacity to provide the desired public services is undermined because tax revenue is lost. But not all commentators share this view. Some see it as a welcome and effective limit on the tendency of governments to continually raise the fiscal burden, and to impose more and more bureaucratic restrictions on the economy and society. More generally, the rise of the underground economy is interpreted to indicate that the relationship between the state and the citizens is fragile and needs improvement. It has also been argued that many developing and transition economies would break down or function at a much lower level of production and welfare if the underground economy did not exist.

The fight against the underground economy is a recurrent theme in many countries. The dominant method is to increase deterrence. The probability of being caught is raised by more regular and intensive controls, often by the police. Punishment is raised by imposing higher fines and, in severe cases, prison sentences. Both consumers and suppliers of goods and services produced in the underground sector are targeted.

The success of such deterrence policies is rather doubtful. In the case of the underground sector related to drugs, prostitution, and alcohol prohibition, the effect has turned out to be at best weak, if not counterproductive. Many persons active in the underground economy move more deeply into illegality in order to make detection more difficult. As a result, the underground activities become more criminalized, and more difficult to observe and to influence. Well-organized and ruthless, organized crime (often called the Mafia) becomes more dominant.

A deterrence policy need not always be a failure, but the successes tend to be short run. Provided the demand for underground goods and services remains intact, the profit opportunities in the underground economy become so large that supply reappears and the underground economy recuperates. A less oppressive policy is to lure people into the official economy by legalizing parts of the underground economy, and by facilitating the move into the official economy (e.g., by granting an amnesty). Such measures have proved to be only moderately successful.

A positive approach to raise the motivation to stay in the official economy by improving the efficiency of public services, reducing the tax and social security burden imposed on labor, and/or by raising civic virtue, has been used only rarely. Many policy makers doubt whether such measures work at all. In any case, they are effective only in the longer run. One possibility is to adjust public supply more closely to what the citizens desire, and to lower cost for a given quantity and quality of public supply. This can be attempted by streamlining public sector activities (e.g., by New Public Management). A more fundamental way is to improve the political process by opening the political arena to contending interests, and by strengthening the democratic participation rights of the citizens.

See also: Immigrants: Economic Performance; Work, History of; Work, Sociology of

Bibliography

Andreoni J, Erard B, Feinstein J 1998 Tax compliance. *Journal of Economic Literature* **36**: 818–60

Cowell F A 1990 *Cheating the Goverment. The Economics of Evasion*. MIT Press, Cambridge, MA

De Soto H 1989 *The Other Path: The Invisible Revolution in the Third World*. Harper and Row, New York

Feige E L (ed.) 1989 *The Underground Economies. Tax Evasion and Information Distortion*. Cambridge University Press, Cambridge, UK

Frey B S, Werner W, Pommerehne 1984 The hidden economy: state and prospect for measurement. *Review of Income and Wealth* **30**(March): 1–23

Gaertner W, Wenig A (eds.) 1985 *The Economics of the Shadow Economy*. Springer-Verlag, Berlin

Mars G 1994 *Cheats at Work. An Anthropology of Workplace Crime*. Dartmouth, Brookfield, CT

Pozo S (ed.) 1996 *Exploring the Underground Economy: Studies of Illegal and Unreported Activity*. W. E. Upjohn Institute for Employment Research, Kalamazoo, MI

Pyle D J 1990 *The Informal Economy*. Kluwer Academic Publishers, Utrecht

Rogoff K 1998 Blessing or curse? Foreign and underground demand for the Euro notes. *Economic Policy: The European Forum*. **26**: 261–304

Schneider F, Enste D 2000 Shadow economies: size, causes and consequences. *Journal of Economic Literature* **38**: 77–114

Simon C P, Witte A D 1982 *Beating the System—The Underground Economy*. Auburn House Pub. Co., Boston

Tanzi V (ed.) 1982 *The Underground Economy in the United States and Abroad*. Lexington Books, Lexington, MA

Thomas J J 1992 *Informal Economic Activity*. Wheatsheaf, London

United Nations 1993 *Inventory of National Practices in Estimating Hidden and Informal Economic Activities for National Accounts*. United Nations, Geneva

Weck-Hannemann H, Pommerehne W W, Bruno S F 1984 *Schattenwirtschaft*. f. Vahlen, Munich

B. S. Frey and F. Schneider

Informal Reasoning, Psychology of

Informal reasoning occurs when a given task, its components, and/or its outcomes vary in degree of certainty or acceptance. This is in contrast to reasoning found in tasks based upon logic and mathematics, which involve certainty and soundness. This article begins with discussion of terminology related to

informal reasoning, followed by a selective review of informal reasoning research, and concluding with consideration of theoretical issues.

1. Terminology

In psychology, reasoning typically refers to the process by which a person makes an appropriate inference that is based upon some particular given information. Inferences may be generated and evaluated, the latter based upon the inference's appropriateness. The inference may be a conclusion or it may fill in an information gap. The reasoning also may consist of a complex series of steps in which particular information is organized and/or used with inferences, thereby enabling the reasoner to reach a goal. Reasoning in designing an experiment or buying a car would be examples of a larger reasoning context.

The term informal reasoning suggests that there are two types of reasoning, formal and informal. This terminology is unfortunate because the distinction is primarily one of task and not of reasoning. Formal tasks are generally those of deductive reasoning as found in logic that has its origins in Aristotle's development of logic. Such tasks, for example, employ the categorical syllogism, structures of conditional logic, or mathematical problems, as proving a theorem. Furthermore, until relatively late in the twentieth century, psychological research on reasoning was dominated by the use of formal tasks. Such research had as a central question whether a person could evaluate a syllogism or some other structure accurately with respect to its validity. If not, the question became why the error was made. Also, a related issue was the question of whether humans are rational, with rationality defined in terms of correct performance on a given formal task.

Examples of tasks involving informal reasoning are making a decision, determining routes and distances to a destination, arguing with a son or daughter, or writing a political speech. Evaluation of performance on such tasks occurs via use of two criteria, the acceptability of the reasoning and the reasoner's effectiveness in reaching the goal.

Informal reasoning tasks differ from formal tasks in at least three ways: first, formal tasks may be transformed symbolically and evaluated, as 'All A are B,' etc., which makes these tasks content-free. Indeed, it is primarily when meaningful content is used that individuals stray from logical performance. In informal reasoning tasks, understanding the content is always critical. Second, evaluation in the formal task requires determining whether the reasoning is consistent with the rules of validity for the particular logical structure under study. In the informal case, evaluation is performed in relation to the acceptability of the contents of the reasoning and effectiveness in reaching the goal. Third, statements found in formal structures are regarded as true or not true; in the informal situation, statements may be regarded as true, not true, or at some particular level of certainty or acceptance between these extremes.

The language structure most common to informal reasoning is the argument. In particular, it is the form of an argument referred to by Aristotle as the enthymeme. This structure is a claim supported by a reason (or premise supporting a conclusion). Indeed, most informal reasoning research has been conducted in the context of argumentation.

2. Informal Reasoning Research

2.1 Argumentation Skill

A number of studies have demonstrated that young children are able to state claims and justify them. Children, however, need to have knowledge of and perhaps interest in, the subject matter of the argument. Young children, for example, will argue about the possession of a toy. If child A has a toy and child B takes it, child A may not only claim ownership of it but also may provide reasons why child B should return it.

The reasons offered by children have been classified. For example, children of ages approximately three to five were observed in a playroom situation (Eisenberg and Garvey 1981) in which one child would make an assertion such as 'Give me the toy' and the child with the toy would say 'No,' sometimes adding a reason. The first child then indicated why he or she should have the toy, and the conflict would sooner or later reach a conclusion. The strategies used, in ascending order with respect to adaptiveness and resolving the conflict, were insistence, mitigation or aggravation, conditional directive (if you will do X, I will do Y), and compromise. Although compromise was the best resolution strategy, a number of studies have shown compromise is found infrequently in conflict situations.

Similarly, the persuasion skills of preschool children have been studied. Children were asked to convince 'mother' or a friend to buy or share a toy. Occasionally there was a negative response to the child and the question was how did the children respond to these negative replies. Bargaining and guarantees were the most frequently used strategies, with boys tending to invoke norms and appeal to a sense of fairness more than girls, and girls making more requests (Weiss and Sachs 1991).

Children at least as young as age five are also able to make and justify moral decisions. Children were asked if they would pull hairs from a tiger, something that would hurt the tiger, if the contents of the hair were going to make some people well. Younger children tended to say no while older children were likely to

answer yes. Both age groups provided supporting reasons. Similarly, children promised to help each other finish tasks in order to be able to attend a baseball game. When the game was called off because of rain, younger children felt the promise should be kept while older children did not. In both of these studies younger and older children differed in what reasons and norms they invoked in making their decisions; older children and adults understand more of the contractual nature of agreements whereas younger children use less complicated norms such as that you do not want to hurt animals. In addition, the argument has been made that young children argue in much the same way as adults, but their world knowledge and social knowledge are less developed than older people (see Stein and Miller 1991).

As children become older, they improve in their knowledge of, and use of, argument skills. Indeed, older children are more skillful in generating counterarguments and refutations. The issue, however, is complicated by the role of subject matter knowledge and intellectual ability. In one study (Means and Voss 1996), the improvement in argument skill performance with age was demonstrated for the one-third of the students scoring highest on intellectual tests. Students in the middle third improved some, but those in the lowest third did not improve over grades 5–11. Furthermore, not only did knowledge of the subject matter (drugs) increase with intellectual ability, but intellectual ability was related to how well the students were able to use their own knowledge.

While children as young as three years of age demonstrate the ability to state arguments, generating written arguments does not occur until a later age. This probably occurs because writing includes mastery of a number of interrelated skills. Following a classroom debate on the topic of pollution, children of ages seven or eight were able to write a positioned argument about the pollution while children over age 14 were able to write a more developed argument involving, for example, counterarguments and refutations. Student involvement in the topic also enhanced performance. Also, another task demonstrated that student knowledge of the structural components was related to their essay performance, thereby suggesting a general learning of argumentation skills (Golder and Coirier 1994).

2.2 *Argumentation in Adults*

The generation and especially evaluation of arguments by adults has received surprisingly little research. In one study (Kuhn 1991), people in their teens, twenties, forties, and sixties were asked three questions: why students failed in school, why so many prisoners return to prison, and what produces unemployment. The investigator analyzed the protocols in relation to statements of possible causes (theories) and the related justification (or evidence) thereof.

Kuhn classified the stated theories as single cause, multiple parallel cause, or multiple alternative causes. While no differences were obtained for gender or age, more educated people tended to state more complex arguments. Kuhn also distinguished between evidence and pseudoevidence. Evidence consisted of demonstrated covariation, causal relations, or was indirect. Pseudoevidence, essentially being illustration, occurred when individuals provided a scenario such as 'My classmate failed in eighth grade. He would go home, his parents were working until late, and he had to take care of his brothers.' Some people provided no evidence, taking their theory as evidence. While on less than half of the opportunities individuals provided genuine evidence, education level again yielded higher performance. Pseudoevidence was quite prevalent, with many participants feeling it was sufficient as a justification. The idea that a story provides evidence even though it does not provide a test of the theory is used in many everyday situations, including political speeches and advertising. Kuhn also found that people who worked in one of the domains of the three questions showed more knowledge of the topic, but reasoned in much the same way as people of their corresponding educational level. Philosophers were able to discuss evidence appropriately, a finding from which Kuhn concluded that knowledge of topic and argumentative skills are two different things, in that you may learn one without learning the other. College education, however, teaches or should teach what constitutes knowledge within a domain and the nature of evidence and arguments within that domain.

When individuals are asked to generate arguments for and against a given position, a bias occurs in which the person states more arguments supporting his or her own position than the other position. This my-side bias has been found in children as young as five years and with adults. The role of sociocultural context in argument generation has been shown by Zammuner (1997) who, conducting the study in Italy, found results consistent with the idea that more arguments had to be generated by those people in favor of abortion rather than those opposed to it because the illegality of abortion in Italy required that a stronger case had to be made for it.

2.3 *Reasoning as a Dialectical Process*

While the discussion of informal reasoning thus far has focused on the individual, informal reasoning is also found in dyadic and larger social settings. Vygotsky (1978) maintained that reasoning is acquired via a dialectical process in which a person, when interacting with others, corroborates, contradicts, or puts in question what he or she already knows.

Reasoning then develops via this interaction. This view suggests beliefs may be modified via such interactions.

In one study dyads discussed a controversial issue, with dyad members congruent or incongruent in their respective positions. The experience tended to move people away from extreme positions to more central positions. Their dialogs also showed incorporation of information and reasoning from the partner.

As with the children, informal reasoning has been observed in adult conflict situations. Hofer and Pikowsky (1993) studied the reasoning in the dialogs of mother and daughter arguments. Daughters, in their effort toward independence, tried to weaken or reject the mother's arguments while stating their own preferences. Mothers, in their effort of control, supported and explained their own arguments and asked questions with the daughters responding to them. Arguments were seldom resolved.

Informal reasoning has also been studied in major negotiation contexts. In an analysis of the transcripts of three high level political meetings, including the Hitler–Chamberlain Munich dialogs, Axelrod (1977) found that the data did not support a debate model. Instead, progress in the conference negotiations was made by the insertion of new ideas that frequently were derived from an examination of the two positions.

How contradictory information is processed has been studied in an educational context. Sometimes individuals ignore contradictions, assimilating them into their existing knowledge base. Other times they do 'knowledge building,' that is, they examine the contradictory statements, consider the implications of them, and modify their knowledge to take them into account. The contradictory statements could be rejected, but if they are, it is for reasons considered during the process (Chan et al. 1997).

Being able to reason well has long been regarded as an important educational goal. The goal is most frequently found in statements related to critical thinking. A number of investigators cited have noted that acquiring argumentative skills and subject matter are not the same thing. Yet the two are related. If one is going to argue about a historical issue, knowledge of that issue is critical. At the same time, argumentative skills are also needed, and these are probably acquired via subject matter learning, especially but not exclusively in history. Students are likely to need more experience in using the knowledge they have acquired in argumentative and other contexts in order to acquire argumentation skills.

3. *Theoretical Issues*

In her criticism of the deduction–induction distinction, Govier (1987) noted that the definition of induction is indefinite; it could even mean everything that is not deduction. Moreover, a strict definition omits some important forms of reasoning such as causal, by analogy, counterfactual, casuistry, case-based reasoning, pro and contra evaluation in decision making, conflict and negotiation, and reasoning in persuasion. The scope of informal reasoning also is indefinite; all of the forms of reasoning mentioned and others could be regarded as types of informal reasoning. Indeed, defining it in a more limited way would be extremely difficult.

With respect to theory, there is no theory of informal reasoning *per se*. Instead, there are theories pertaining to the types of informal reasoning such as those of analogical reasoning, argumentation, problem solving of ill-structured problems, causal reasoning, conflict resolution, and various approaches to decision making. These theories are relatively specific and have not been integrated. The present article focuses on argumentation both because of its cross-discipline pervasiveness and because it is basic to informal reasoning. The focus however, suggests the need to consider three related issues.

The scope of informal reasoning goes beyond the examples of this chapter. In academia, it is readily observed in history and the social sciences, but it also may be found in science. Dunbar (1995), for example, studied the thinking of scientists in microbiology and cell biology laboratories. The thinking involved evidence evaluation, analogies, and problem finding, generally not formal logic structures, but obtaining information, using it, and making judgments about data interpretation and subsequent research.

As previously noted, deductive reasoning has had a privileged position in the study of reasoning. But the arguments found in informal reasoning are of the claim-support type, which it has been argued consists of a syllogism with a missing premise, thereby making it deductive. But Govier (1987) makes the point that any claim-supporting argument can be made logically sound by adding one or more premises. This makes logical soundness of an enthymeme trivial and points to the importance of the enthymeme as a means of justification and/or persuasion.

Informal reasoning is also closely related to rhetoric. In the last half of the twentieth century two works, by Toulmin (1958) and Perelman and Olbrechts-Tyteca ([1958] 1968), have been instrumental in renewing interest in rhetoric. Each emphasizes a jurisprudence model to reasoning rather than a model based upon logic. Toulmin's model describes components of a single argument, with the other work emphasizing argument schemes and the role of the audience. While these works have been criticized, it seems reasonable to assume that future work in informal reasoning will involve the rhetorical function of informal reasoning.

See also: Critical Thinking, Cognitive Psychology of; Dialectical Thinking, Psychology of; Logic and Cognition, Psychology of; Nonstandard Reasoning;

Practical Reasoning: Philosophical Aspects; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning; Problem Solving (Everyday), Psychology of; Reasoning with Mental Models; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Axelrod R 1977 Argumentation in foreign policy settings. *Journal of Conflict Resolution* **21**: 727–44

Chan C, Burtis J, Bereiter C 1997 Knowledge building as a mediator of conflict in conceptual change. *Cognition and Instruction* **15**(1): 1–40

Dunbar K 1995 How scientists really reason: Scientific reasoning in real-world laboratories. In: Sternberg R J, Davidson J E (eds.) *The Nature of Insight*. pp. 365–95

Eisenberg A R, Garvey C 1981 Children's use of verbal strategies in resolving conflicts. *Discourse Processes* **4**: 149–70

Golder C, Coirier P 1994 Argumentative text writing: Developmental trends. *Discourse Processes* **18**: 187–210

Govier T 1987 *Problems in Argument Analysis and Evaluation*. Foris, Providence, RI

Hofer M, Pikowsky B 1993 Validation of a category system for arguments in conflict discourse. *Argumentation* **7**: 135–48

Kuhn D 1991 *The Skills of Argument*. Cambridge University Press, Cambridge, UK

Means M L, Voss J F 1996 Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction* **14**(2): 139–78

Perelman C, Olbrechts-Tyteca L [1958] 1968 *The New Rhetoric: A Treatise on Argument*. University of Notre Dame Press, Notre Dame, IN

Stein N L, Miller C A 1991 I win-you lose: The development of argumentative thinking. In: Voss J F, Perkins D N, Segal J W (eds.) *Informal Reasoning and Education*. Erlbaum, Hillsdale, NJ, pp. 265–90

Toulmin S 1958 *The Uses of Argument*. Cambridge University Press, New York

Vygotsky L S 1978 *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

Weiss D M, Sachs J 1991 Persuasive strategies used by preschool children. *Discourse Processes* **14**: 55–72

Zammuner V L 1997 For or against: The expression of attitudes in discourse. *Text* **7**(4): 411–34

J. F. Voss

Informal Sector

The notion of the informal sector emerged from international research on Third World urban development in the early 1970s as a name for activities and enterprises escaping formal economic regulation. It and other terms for nonregulated economic activities have since been applied to research in the West, yet the bulk of informal sector research continues to be undertaken in the Third World. The focus has shifted from detailed descriptions of informal sector activities in different settings, to questions about their absorbtive capacity and potential for growth, to dynamics of informalization in general, including its gendered nature. This shift entails conceptual clarifications of the nature of informal sector activities and their implications for understandings of economic growth dynamics in general.

1. Meanings of 'Informal'

The term 'informal' is more of a residual notion than a concept, defined in relationship to something else, that is, to what it is not. The term 'sector' implies linkages to the rest of the economy. In social science scholarship since the early 1970s, including policy research and development economics, and in applied work, the term 'informal' has been used variously for a segment of a labor market, a set of activities, a nonregulatory status, a form of insertion in labor markets, a type or size of firm, and a general characterization of an economy. The phenomena to which the 'informal sector' refers are not new. Before this notion became current, social scientists were well aware that persons who do not perform labor for wages are not necessarily without work, although they are often counted as such in official statistics.

The 'informal sector' gave a name to economic activities that occupied large portions of urban populations in Third World countries. The scholarship that has arisen around it has characterized the linkage to the wider economy in two main ways. Perhaps the most common characterization of this relationship stems from Hart's work in Accra, Ghana, in the early 1970s. Hart coined the term 'informal economy' for a wide range of income generating activities with varying degrees of legality. He termed them informal so as to contrast them with formal, contractual wage labor (1973). The second characterization is influenced by the work of de Soto (1989) in Peru that explained informal sector activities as a response to ineffective state intervention, a result of the failure of both the state and the market to provide employment, housing, and services.

2. Informality: Sector or Process

Since the early 1970s, the informal sector has been a subject of extensive debate and research in which the principal focus has been on Western policy interventions aimed at increasing economic growth in developing countries. Some scholars also began to explore what was variously labeled irregular, black, underground, shadow, or hidden economies in the US and Europe, and some illuminated the operation of the second or parallel economies in socialist planned econ-

omies. While criminal activity *per se* usually is not considered part of the informal sector, petty scams and workplace pilfering are (Henry 1978). And the informal sector provides organizational practices and networks that facilitate drug trafficking and smuggling.

The chief questions raised in this scholarship concern the content, structure, and dynamics of the informal sector, that is, what to include and exclude, how precisely to differentiate it from the formal sector, and what growth potential it may have. Some scholars have suggested alternative terms and frameworks, such as casual work, petty commodity production, or householding and provisioning (Bromley and Gerry 1979, Mingione 1991). Still, the informal sector notion continues to be widely used for lack of any better descriptive term. Despite their conceptual misgivings, most scholars agree that discussions about economic life must take account of informal economic developments in their relationship to the wider economy. For example, MacGaffey (1991) found informal activities to be so central to efforts at making a living in Zaire during the last half of the 1980s that she characterized them as the Real Economy.

The idea of 'popular' or 'informal economies' is another way of highlighting the pervasiveness of nonregulated economic practices, recognizing their overall importance. Thinking processually about 'informalization' avoids the arbitrariness of sectoral dichotomies. It also helps to bridge the conceptual gap between production and reproduction, directing attention to household strategies and the ways in which household composition, stage of development cycle, and the gender and age of members introduce cultural assumptions into the division of labor. Yet although women and children are a highly visible segment of many informal markets, development advocates have only recently begun to pay attention to the relationships of age and gender to informal sector developments. For further discussion, see *Apprenticeship: Anthropological Aspects*; *Center–Periphery Relationships*; *Economic Anthropology*; *Urban Anthropology*.

2.1 The 'Informal Sector' and Third World Development

Research into informal sector developments in the Third World has clustered around two particular Western policy interventions aimed at promoting economic growth. The first cluster of work was influenced by Hart's research among migrants in Ghana that identified an 'informal economy' of self-employment, legitimate and illegal as opposed to regularized wage labor (1973). International development agencies were at this time concerned about the failure of industrialization efforts to improve economic growth in much of the Third World. Hart's observations were absorbed quickly into development policy directed at satisfying basic needs by harnessing informal income opportunities. In the International Labour Office's Kenya mission, for example, such activities were explained as 'a way of doing things' (ILO 1972), characterized by ease of entry, reliance on indigenous resources and family ownership, smallness of scale, labor intensive with adapted technology, skills obtained outside the formal school system, and unregulated and competitive markets. Here, as elsewhere in the developing world, governments were encouraged to take a positive role, ceasing to demolish unauthorized housing, changing trade and licensing procedures, and providing technical and outreach support to informal sector activities.

The second cluster of work on the informal sector appeared in the wake of the International Monetary Fund's (IMF) structural adjustment programs, particularly from the last half of the 1980s on. Many of the policy measures (e.g., employment retrenchment both in public and private work, introduction of value added taxes, removal of subsidies on food, fertilizers, and housing, and introduction of user fees in schools and hospitals) advocated by the IMF and the World Bank to transform state-controlled economies into market economies had adverse effects on already vulnerable groups. In fact, the volume and differentiation of informal sector activities increased across much of the Third World during the 1980s. Strapped for revenue and burdened by accelerating international debt, many governments began encouraging interventions by local and external nongovernmental organizations (NGOs). The centerpiece of this renewed policy attention is 'micro-enterprises' and their enhancement through credit arrangements and the provision of basic training in technical skills. Many observers who were disenchanted with the lackluster performance of Third World governments over the previous decades viewed informal sector activity as examples of free market initiatives. Perhaps more than any other scholar King (1996) has documented, with rare longitudinal insights, the local effects of these shifting policy interventions on the activities and achievements of informal sector artisans in Kenya.

2.2 The 'Informal Sector' and the West

Informal economic activities take place in advanced industrial countries as well as in the Third World (Ferman et al. 1987, Portes et al. 1989). Small-scale enterprises that bypass official regulation and off-the-book labor practices provide products and services that are traded and consumed throughout most of the West today. These activities are in parts of the US associated with immigrant communities, whereas in parts of Europe, for example Italy, they are based on artisanal enterprises and putting out systems. In the second or parallel economies of the former socialist countries in Eastern Europe, producers and consumers supplied, swapped, and bartered goods and services to

alleviate market shortages, avoiding the bureaucratic bottlenecks of centrally planned economies. Studies such as these are demonstrating that informal sector activities in the West are not exclusively associated with a bygone era of economic history, but keep reappearing in different places, operating in a variety of politico-economic contexts.

3. Policy Concerns

Definitions of the informal sector have important implications for the identification of target groups and types of intervention. If the informal sector provides an economic safety net, it does so differently, depending on type and level of activity, the resources brought to bear on its operation, and the local cultural context that shapes the participation of actors (Rakowski 1994). Whether a growth potential in fact exists requires understanding of the nature both of the activities in question and of the social relations that structure them. Some researchers avoid such discussions entirely by using descriptive terms like 'micro-enterprises.' But micro-enterprises encompass a range of activities that draw on widely diverse access, require different types of resources and inputs, and above all, they put the young and the old and women and men to work on unlike terms. Unless the economic and social nature of such activities is accounted for in more specific terms, development assistance risks being restrictive, thus jeopardizing the very goals of expanding employment and income generation that it seeks to promote. Anticipating that persons with few means may forfeit loans, some credit schemes, for example, target better-off participants, in this way accentuating already existing socioeconomic differentiation rather than reducing it.

4. The Market Meets the 'Informal Sector'

Since the 'informal sector' term was coined in the early 1970s, the economic context in which research on this topic first emerged has changed. Shifts in the global economy are altering production processes in an international division of labor that is affecting both the industrial West and the developing world, but differently so. Recent scholarship is reckoning with the linkages between local conditions and global forces that are shaping informal sector developments both in the West and the Third World (Dilley 1992). The introduction of 'free market' policies in the former centrally planned economies both in Eastern Europe and postcolonial Africa, for example, is giving rise to new research questions concerning the experiences of ongoing economic transformations. In much of Eastern Europe and Africa, an informal sector culture had been taken so much for granted that its operational principles have become pervasive. Relevant questions in such settings concern the effects of free market operations on members of the many households whose livelihoods were made from informal sector activity.

While the term the 'informal sector' may remain part of our descriptive vocabulary, new scholarship is likely to continue to disentangle the complexities of informal activities that need to be understood in their own right if we are to develop policy programs to support them. The enthusiastic reception in some circles of de Soto's praising of the resourcefulness of the urban poor must be tempered by a recognition of the heterogeneity and change of informal sector activities as well as of the exploitative conditions under which much of this sector's work is carried out, particularly by women and children. Lessening this state of affairs on a global scale requires more than self-help by small-scale entrepreneurs and intermittent external support by NGOs. It invites government action on many fronts to reshape the institutional structures that continue to reproduce livelihoods in the interstices of the economy.

See also: Household Production; Informal and Underground Economics; Organization: Informal

Bibliography

Bromley R, Gerry C 1979 *Casual Work and Poverty in Third World Cities*. Wiley, New York

de Soto H 1989 *The Other Path: The Invisible Revolution in the Third World* [1986 *El Otro Sendero: La Revolucion Informal*. Editorial El Barranco, Lima, Peru]. Harper and Row, New York

Dilley R 1992 *Contesting Markets: Analyses of Ideology, Discourse and Practice*. Edinburgh University Press, Edinburgh, UK

Ferman L, Henry S, Hoyman M 1987 The informal economy. *The Annals of the American Academy of Political and Social Science* **493**

Hart K 1973 Informal income opportunities and urban employment in Ghana. *Journal of Modern African Studies* **11**: 61–89

Henry S 1978 *The Hidden Economy: The Context and Control of Borderline Crime*. Robertson, Oxford, UK

International Labour Office (ILO) 1972 *Employment, Incomes and Equality: A Strategy for Increasing Productive Employment in Kenya*. ILO, Geneva, Switzerland

King K 1996 *Jua Kali Kenya: Change & Development in an Informal Economy 1970–95*. Currey, London

MacGaffey J 1991 *The Real Economy of Zaire: The Contribution of Smuggling & Other Unofficial Activities to National Wealth*. University of Pennsylvania Press, Philadelphia

Mingione E 1991 *Fragmented Societies: A Sociology of Economic Life beyond the Market Paradigm*. Blackwell, London

Portes A, Castells M, Benton L 1989 *The Informal Economy: Studies in Advanced and Less Developed Countries*. The Johns Hopkins University Press, Baltimore, MD

Rakowski C 1994 *Contrapunto: The Informal Sector Debate in Latin America*. State University of New York Press, Albany, NY

K. T. Hansen

Information and Knowledge: Organizational

Information is data that have meaning. Meaning arises as data are interpreted by sense-making frameworks (Bruner et al. 1956). This sense making depends on what is already known as well as what data are available.

Organizational information derives its meaning from the sense-making frameworks that characterize specific organizations. Many organizational members need information in order to fulfill their responsibilities, so other members gather data and convert it into information. Organizations can acquire data by scanning environmental data sources, by ingesting new members, or through internal activities, such as accounting and research. Most organizations include personnel who specialize in various forms of data acquisition or information processing, and large organizations have departments that focus on such specialties. Organizations convert data into information by discussing the implications of data, by adding to databases, or by feeding data into decision processes. They process information by filtering it, integrating it, disseminating it internally, and interpreting its implications. Nearly all organizations incorporate rules and procedures that cause personnel to initiate actions or to halt actions when they receive information that satisfies certain conditions (March and Simon 1958, Starbuck 1983).

Organizations emit data to mislead or undermine their competitors, and many organizations seek to gain competitive advantages by gathering data that their competitors lack. For example, data gathering and analysis enabled a large cosmetics company to avoid making a costly pricing blunder. This company buys data that are recorded by the checkout scanners in chain stores, and it gets daily reports from sales personnel who record data from retail outlets and transmit them to headquarters nightly. When a major competing firm introduced a new product, this cosmetics firm planned a defensive price reduction on its high-priced product X to keep the new competitor from taking too much business away from X. However, data analyses revealed that the new product was not actually competing with product X but with a much lower-priced product.

On the other hand, data gathering or information processing may be carried out so poorly that they undercut organizational performance. For instance, a rapidly growing health-care insurance firm processed receipts from clients much faster than claims from doctors and hospitals. As one result, this firm greatly overestimated its revenues, underestimated its costs, and made unrealistic predictions about its profits ···until the manager in charge of information processing reported that bill payments had fallen in arrears by nearly a year.

Some research shows that communications in organizations reflect the statuses and aspirations of organizations' members. People in hierarchies talk upward and listen upward (Porter and Roberts 1976). They send more messages upward than downward, they pay more attention to messages from their supervisors than to messages from their subordinates, and they try harder to establish rapport with supervisors than with subordinates. People also bias their upward messages to enhance good news and to suppress bad news.

Other research indicates that formal communications in organizations are generally problematic. Organizations bias their formal reports to win the support of employees, customers, investors, and the general public; and the members of organizations use formal reports to promote their careers or other interests. As a result, formal reports are permeated with misrepresentations and biases, and the organizations that take formal reports seriously are prone to run into trouble (Altheide and Johnson 1980, Hopwood 1972). Such issues make informal communication important. For instance, Grinyer and Norburn (1975) found that more profitable firms pay less attention to the information transmitted via formal reports and rely more strongly on informal communication channels, and that more profitable firms base their analyses on information drawn from diverse sources.

Knowledge is an accumulation of information. Whereas information comprises a flow that changes rapidly, knowledge is a stock that changes only incrementally. Some knowledge is explicit; other knowledge is tacit (Baumard 1999, Polanyi 1966). Organizations can attempt to share and organize explicit knowledge, whereas tacit knowledge remains private and disordered. Nonaka (1994) argued that individuals often acquire knowledge tacitly and this tacit knowledge does not become organizational knowledge until it becomes explicit. Many business firms are attempting to foster the conversion of tacit knowledge to explicit knowledge. One tactic for doing this has been to organize employees into on-line communities that identify and discuss shared issues on intranets.

Activities such as teaching, consulting, and practicing law draw on extensive knowledge without processing large amounts of current information. Other activities, such as bookkeeping and data entry, process much information without using much knowledge. However, distinctions between data, informa-

tion, and knowledge are often difficult to draw, as the conversion of data into information depends on knowledge, which derives from data and information (Dretske 1999, Machlup 1980). Since the meaning of knowledge arises in part from its structure, knowledge reflects the intentions of the people who compile or organize it, and different people often extract different knowledge from a shared flow of information.

Not only is it difficult to distinguish information and knowledge from each other, the labels that people apply to transformations of information and knowledge are often unclear. People who perform very similar tasks may describe their activities as applying old knowledge to new problems, as creating new knowledge, or as preserving knowledge that already exists. For example, it can be difficult to distinguish between creating knowledge and applying it; new knowledge research may have direct applicability, efforts to apply knowledge may yield fundamentally new insights. Indeed, to create valid knowledge about systems as complex as a human body or an economy, people have to apply their current knowledge. Similarly, applying and preserving knowledge complement each other. Merely storing knowledge does not preserve it for long periods, as people must relate it to their current problems and activities. They have to translate it into contemporary language and frame it within current issues.

1. Recognizing the Importance of Information and Knowledge

Strangely, both managers and scholars paid little attention to the centrality of information in organizations' activities until the last half of the twentieth century. Not until the 1950s did articles and books begin to portray managers and work groups as decision makers. It was the 1970s, however, before there were many studies of organizations as decision-making and information-processing systems.

One factor driving recognition of the centrality of information was the rising economic importance of service industries and knowledge-based work. In industrialized countries, the fractions of workers who process information have risen gradually throughout the twentieth century. The fastest growing occupations have been clerical, professional, and technical workers, including managers and administrators. In the latter part of the century, the fastest growing industries included telecommunications, software, and computing services. The United States provides the most extreme case. By 1990, the information sector accounted for three-quarters of the US gross domestic product, and over half of US workers were doing some type of information work.

A second, and possibly more important, factor that has focused attention on information has been development of computer and telecommunication technology. Computers have not only facilitated the processing of information, they have provided metaphors for information processing by both individual people and social systems. Thus, equipped with an appropriate sense-making framework, managers and scholars began to notice the importance of information processing.

In 1958, March and Simon proposed the then novel idea that organizations can be viewed as information-processing systems. March and Simon pointed out that information sources and channels affect organizations' perceptions, intraorganizational conflict, and goal coherence. They emphasized the human limitations that prevent people and organizations from acquiring and processing unlimited amounts of information. They also characterized people and organizations as using programs to process information, a direct analogy to computers. Later, Cyert and March (1963) described some programs used by business firms to make decisions about prices and quantities.

Galbraith (1973) proposed an approach to organization design that focuses on the communication needed to coordinate tasks. He pointed out that programs can substitute for communication. When tasks can be adequately described with programs and personnel follow these programs, personnel can coordinate their actions without much communication. Programs are effective, however, only when organizations face familiar situations. If organizations continually encounter new situations, their old programs will not be effective and personnel will have to coordinate by communicating with each other. Galbraith pointed out that organizations can make communication less necessary by separating tasks into self-contained clusters and by lowering their performance goals, thus leaving some resources underutilized. Organizations can also substitute direct horizontal communication across departmental boundaries for vertical communication up and down hierarchies; this moves decision making down hierarchies to the people who have the needed information rather than bringing information up hierarchies to higher-level decision makers.

Studies of organizational knowledge did not appear in numbers until the 1990s. One theme has been the increasing importance and distinctive characteristics of professional-service organizations, such as health-care agencies, law firms, and public-accounting firms. Knowledge-intensive firms make esoteric expertise the focus of their competitive strategies (Starbuck 1992). As well as becoming more important economically, professional-service organizations have grown much larger and expanded globally (Brock et al. 1999). A second theme has been the importance of information-processing resources and knowledge resources to the profitability of all sorts of organizations. Current thinking about strategic management emphasizes the

usefulness of valuable resources that are rare and difficult to imitate (Barney 1991), and many organizations seek to accumulate valuable and proprietary knowledge that they can exploit strategically. For instance, chemical and pharmaceutical companies employ skilled scientists and spend heavily on research aimed at developing patentable, proprietary products. Some manufacturing firms assert that cadres of experienced and highly trained managers comprise their key assets. Many firms have appointed 'learning officers' who are charged with identifying useful knowledge and disseminating it to personnel who can use it.

It seems likely that information and knowledge will grow more and more important in industrialized societies. This rising emphasis involves several components.

(a) The industrialized societies are shifting away from manufacturing and toward services. For instance, in Europe, telecommunications activities have been growing about 10 percent annually and software and computing services have been growing nearly 20 percent annually.

(b) Knowledge workers and information workers are replacing manual workers. For example, machine-tool operators are giving way to technicians who monitor computer-controlled machines.

(c) Workers of all kinds are gaining more education and more information-processing skills.

(d) Business is investing heavily in equipment for information work. In the US in 1986, information-related equipment accounted for 40 percent of capital investment.

(e) Knowledge-intense and information-intense organizations are emerging and employing millions of people.

(f) Corporations are depending on communications and information processing to maintain their coherence as they globalize. Transoceanic communications have been growing about 35 percent annually, with the faster growth being in the Pacific region.

(g) Firms are forming alliances and coalitions that rely on shared databases and electronic communications instead of face-to-face meetings (Hedberg et al. 1997). Financial analysts have forecast that business-to-business (i.e., B2B) commerce via the Internet will be the fastest growing business activity during the early years of the twenty-first century.

Despite the fact that computers and electronic communications devices have existed in some form for decades and they have become prevalent in Europe and North America, only a small fraction of the world population now has use of computers, e-mail, or the Internet. At the same time, technological development has actually been accelerating. Thus, the effects of these technologies on people, organizations, and societies are only beginning. Some of the current trends appear to be as follows: Organizations are gaining access to more data more quickly. This creates opportunities for surveillance from afar, so 'competitive intelligence' is becoming easier and more productive. In addition, supervisors can more easily monitor the activities of distant subordinates, and organizations can more easily monitor the activities of their partners in alliances. Organizations have more freedom to spread geographically, and they can collaborate effectively without actually merging. Finally, as communications replace physical travel and direct supervision, some employees are engaging in telework and workgroups are replacing hierarchies.

See also: Bureaucracy and Bureaucratization; Conflict: Organizational; Corporate Culture; Design and Form: Organizational; Innovation: Organizational; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Organizational Behavior, Psychology of; Organizational Change; Organizational Decision Making; Organizations, Metaphors and Paradigms in; Organizations, Sociology of; Organizations: Unintended Consequences; Rational Choice and Organization Theory; Strategic Intelligence; Strategy: Organizational; Technology and Organization

Bibliography

Altheide D L, Johnson J M 1980 *Bureaucratic Propaganda*. Allyn and Bacon, Boston

Barney J B 1991 Firm resources and sustained competitive advantage. *Journal of Management* **17**: 99–120

Baumard P 1999 *Tacit Knowledge in Organizations*. Sage, London

Brock D M, Powell M J, Hinings C R 1999 The restructured professional organization: corporates, cobwebs and cowboys. In: Brock D M, Powell M J, Hinings C R (eds.) *Restructuring the Professional Organization: Accounting, Health and Law*. Routledge, London, Chap. 11, pp. 215–29

Bruner J S, Goodnow J J, Austin G A 1956 *A Study of Thinking*. Wiley, New York

Cyert R M, March J G 1963 *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ

Dretske F I 1999 *Knowledge and the Flow of Information*. CLSI Publications, Stanford, CA

Galbraith J R 1973 *Designing Complex Organizations*. Addison Wesley, Reading, MA

Grinyer P H, Norburn D 1975 Planning for existing markets: Perceptions of executives and financial performance. *Journal of the Royal Statistical Society, Series A* **138**: 70–97

Hedberg B, Dahlgren G, Hansson J, Olve N-G 1997 *Virtual Organizations and Beyond: Discover Imaginary Systems*. Wiley, Chichester, UK

Hopwood A G 1972 An empirical study of the role of accounting data in performance evaluation. *Empirical Research in Accounting: Selected Studies* [suppl. to the *Journal of Accounting Research*] **10**: 156–82

Machlup F 1980 *Knowledge, Its Creation, Distribution and Economic Significance*. Princeton University Press, Princeton, NJ

March J G, Simon H A 1958 *Organizations*. Wiley, New York

Nonaka I 1994 A dynamic theory of organizational knowledge creation. *Organization Science* **5**: 14–37
Polanyi M 1966 *The Tacit Dimension*. Doubleday, Garden City, NY
Porter L W, Roberts K H 1976 Communication in organizations. In: Dunnette M D (ed.) *Handbook of Industrial and Organizational Psychology*. Rand McNally, Chicago, pp. 1553–89
Starbuck W H 1983 Organizations as action generators. *American Sociological Review* **48**: 91–102
Starbuck W H 1992 Learning by knowledge-intensive firms. *Journal of Management Studies* **29**: 713–40

W. H. Starbuck and P. Porrini

Information Processing and Psychopathology

How do we interpret and impart meaning to information from our surroundings? Does the manner in which we process and interpret information affect our psychological functioning? Peoples' perceptions of the world do not reflect a fixed reality. Rather, perceptions involve a dynamic interplay between the external stimuli that people observe and their highly subjective (and often unconscious) attentional and interpretative processes that allow them to impute meaning to experience. The scientific construct that we now refer to as 'information processing' dates back to Wilhelm Wundt (the father of experimental psychology), who initiated the systematic study of the ways in which people perceive external stimuli.

Although information processing has a long tradition in experimental psychology, applications of information processing models to the field of psychopathology are relatively recent. This is particularly surprising given that several forms of psychopathology clearly involve aberrations in the accurate processing of information. Nevertheless, advances in recent years suggest that the manner in which people process and interpret information does significantly influence their emotional functioning and well-being. Moreover, an information-processing approach to psychopathology promises to refine our understanding of a number of psychological disorders.

1. *Background*

Although the specific emphases have changed, for much of the twentieth century most forms of psychopathology were conceptualized either as disorders of biological origin or as problematic behaviors that resulted from faulty learning. For example, the disorder we know today as schizophrenia was originally called *dementia praecox*, or early dementia, by Kraepelin (1883), who believed that the brains of individuals who developed schizophrenia had begun to deteriorate prematurely. Similarly, scientists ascribed the etiology of depression and anxiety to various neurotransmitters and other biological factors. For instance, depression was hypothesized to result from too little serotonin and norepinephrine (e.g., Schildkraut 1965), and some forms of anxiety were postulated to be linked to a biological 'readiness' to develop fear to specific stimuli (e.g., Garcia and Koelling 1966).

These biological views of psychopathology were refined and strengthened in the 1950s and 1960s with the advent of psychotropic medications as a major treatment modality for schizophrenia, depression, and anxiety. At the same time, psychologists were beginning to develop behavioral conceptualizations of specific forms of psychopathology. Depression, for example, was hypothesized to be caused by the individual exhibiting a reduced repertoire of behaviors. Lewinsohn (1974) contended that the individual's reduced behavioral repertoire led to a decreased rate of positive reinforcement from the environment, which resulted in further reductions in behavior. Similarly, anxiety disorders were conceptualized from a behavioral perspective as a constellation of avoidance behaviors. Phobic individuals avoid the feared object, and people diagnosed with generalized anxiety disorder do not engage in adaptive behaviors through which they would learn that their avoidance of feared or conflictful situations is unwarranted. Moreover, the relief that came from avoiding a feared situation served to reinforce the avoidance behaviors.

These behavioral conceptualizations of psychopathology, and particularly of depression and anxiety, led to the development of effective behaviorally-focused treatments. As was the case with pharmacotherapies, the success of behavior therapy strengthened behavioral conceptualizations of psychopathology. In the 1970s, however, there was a 'cognitive revolution' in psychology.

2. *The Cognitive Revolution in Psychology*

In 1974, W. N. Dember described what he referred to as a 'cognitive revolution' in psychology in general, and in clinical psychology in particular (Dember 1974). Clinical psychologists became concerned with cognitive processes. There was growing dissatisfaction with the simple behaviorist stimulus–response (S-R) learning model of psychopathology, and increasing attention was being given to information-processing models of cognition. The cognitive revolution led psychologists to revisit intrapsychic functioning as a possible cause of psychopathology, opening doors that would reconsider the role of thought in human behavior.

Cognitive scientists became increasingly interested in two areas of human functioning: emotion and automatic processing. Psychologists offered cognitive theories of emotional functioning that emphasized the interaction of cognition and emotion in understanding human behavior (e.g., Bower 1981, Oatley and Johnson-Laird 1987). Psychologists also began to use methodologies developed by cognitive scientists to examine information processing that occurred outside of conscious awareness or control. This line of research focused on such areas as selective attention, implicit memory, and schema functioning (e.g., Kihlstrom 1987). The research activity in these new cognitive approaches paved the way for the study of cognitive functioning in psychopathology, and particularly in those psychological disorders that are characterized by difficulties in emotion regulation.

3. *Information Processing Perspectives on Psychopathology*

Psychological disorders began to be conceptualized and studied from an information-processing perspective. For example, some researchers suggested that schizophrenia is characterized by a breakdown or dysfunction in the filter mechanism that is responsible for screening out stimuli that are not related to the individual's current thought (e.g., Venables 1964). In the absence of that filtering, the schizophrenic is flooded with stimuli. Cohen et al. (1992) refined this filtering hypothesis by suggesting that schizophrenic individuals have a decreased ability to use contextual information to select appropriate behavior. Thus, from an information-processing perspective, persons diagnosed with schizophrenia are hypothesized to be unable to process and filter stimuli effectively, or to use contextual information to guide their thoughts and behaviors.

The growing interest in information-processing approaches to the study of psychopathology had an even greater impact on theory and research in the clinical disorders of depression and anxiety. Cognitive theorists implicated specific patterns of information processing as playing a causal role in the development of depression and anxiety. For example, Beck (1976) ascribes central importance to *negative schemas* in rendering an individual vulnerable to experiencing depression or anxiety. More specifically, Beck postulates that negative schemas, or expectancies, play a causal role in depression and anxiety by influencing negatively the selection, encoding, categorization, evaluation, and recall of stimuli in the environment. That is, people who are prone to developing depression or anxiety are characterized by negative schemas that influence their perceptions of, and interactions with, their environments and, consequently, lead to these disorders.

Beck hypothesized that these negative schemas develop from early adverse experiences in childhood. For people who are vulnerable to depression, these experiences concern loss and failure; for anxiety-prone individuals, they involve themes of danger and threat. Once they are activated, the schemas act as 'filters' through which stimuli and events in the environment are perceived, evaluated, attended to, and remembered. Thus, individuals interpret the information from the environment so that it is consistent with the schemas, selectively attending to particular stimuli or distorting the information to achieve congruence; positive stimuli may be selectively filtered out, and negative or neutral information may be perceived as being more negative than is actually the case.

These negative information-processing biases are postulated to lead to depression or anxiety, depending on the content of the schemas. Thus, from an information-processing perspective, schemas are postulated to play a critical etiological role in predisposing individuals to experience clinically significant episodes of depression or anxiety, particularly in the face of specific events that activate schema-driven information-processing biases. A similar conceptualization of the relation between cognition and emotion has been formulated by Bower (1981). Bower articulated a semantic associative network model in which concepts, emotions, and experiences are represented by distinct nodes that are interconnected within a large network. Within this network, activation of a particular node spreads automatically to adjacent nodes. Thus, the activation of emotion nodes spreads to adjacent cognitive nodes to produce mood-congruent memory. Bower posits that it is this activation of concepts and experiences which is related to the specific emotion that makes it easier to recall events that are congruent with the affective state.

Considered broadly, these cognitive models hypothesize that individuals who are experiencing emotional disorders should be characterized by negative biases in attention and memory. At the time these formulations were proposed, however, there were few reliable methods available for assessing biases in information processing. Recently, however, in large part because of the increasing interest in cognitive science, considerable progress has been made in the assessment of information-processing styles.

4. *How Are Information-processing Biases Studied?*

These information-processing formulations of the causes of various forms of psychopathology stimulated an extraordinary number of empirical studies. Early efforts to examine schematic functioning and information processing in psychopathology relied on self-report questionnaires to assess negative cognitive functioning. Although these were important studies,

the self-report methodology has two significant limitations. First, the information-processing formulations that are being tested with these measures are, in large part, theories of cognitive *processes* as well as cognitive *content*. That is, Beck and Bower formulated hypotheses concerning not only the content of biases in the information processing of depressed and anxious individuals, which concerns themes of failure and loss, and danger and threat, respectively, but more important, *how* stimuli and information involving these themes and this content are attended to, interpreted, and remembered, and how these cognitive processes are related to the course of the disorder.

Unfortunately, although questionnaire methodologies are adequate to examine cognitive content, they are not well suited to an assessment of cognitive processes. Second, because self-report questionnaire measures typically require participants to make conscious, deliberate, and thoughtful responses, they are simply unable to assess the existence and functioning of schemas or associative networks, which are hypothesized to be activated automatically and to operate largely outside of individuals' awareness.

To address these limitations, investigators developed and adapted methodologies derived from research in experimental cognitive psychology to examine the functional role of cognitive biases in a number of forms of psychopathology. In contrast to self-report questionnaire measures, these methodologies permit the assessment of automatic cognitive processes and schematic functioning. For example, several tasks have been developed to examine attentional biases in depression and anxiety. In the *Emotion Stroop Task*, participants are presented with emotionally valenced stimuli printed in different colors. Typically, the stimuli are selected to be relevant to the disorder under investigation: depressed-content words in the study of depression, anxious-content words in the study of anxiety, etc. As in the original Stroop task, participants are instructed to ignore the content, or meaning, of the stimuli and name the color in which each word is presented as quickly as they can. Information-processing formulations posit that the cognitive schemas of depressed or anxious individuals should lead them to exhibit biased attention to loss/failure or threat/danger stimuli, respectively. If these formulations are valid, the participants should be relatively unable to disattend from the content of the emotionally relevant words and consequently, should exhibit longer latencies (i.e., greater interference) to naming the ink colors of these stimuli.

The *Dichotic Listening Task* is conceptually similar to the Emotion Stroop Task. In this task, different information is presented simultaneously through headphones to each ear. Participants are instructed to attend to and repeat (shadow) neutral 'target' stimuli that are presented to one ear, and to ignore distractor stimuli that are presented simultaneously to the other ear. Thus, participants attempt to direct their attention to the ear in which the target stimuli are presented, and ignore stimuli that are presented in the 'unattended' ear. Again, the negative schemas of depressed or anxious persons are hypothesized to make it more difficult for them to ignore stimuli presented to the unattended ear when the stimuli are negatively valenced, than when the stimuli are neutral or positively valenced. This diversion of attention to negative stimuli presented in the unattended ear means that fewer attentional processing resources are available for tracking the 'target' information presented in the attended ear. Consequently, depressed and anxious participants are hypothesized to exhibit a greater number of shadowing errors of the neutral target stimuli on those trials in which the distractor stimuli are negatively valenced.

The *Dot-Probe Task* was developed to examine selective attention to emotional stimuli. In this task pairs of words are presented briefly (generally 500 to 750 msecs) and simultaneously on a computer screen. Both words then disappear from the screen and a small dot appears in the spatial location of one of the words. Participants are required to press one of two buttons to indicate, as quickly as they can, whether the dot replaces the word on the left or on the right. If, as information-processing formulations of psychopathology hypothesize, depressed or anxious individuals are biased to attend to emotionally relevant stimuli, they should be faster to detect the dot probe when it appears in the spatial location of a relevant emotion word (to which they were attending when it disappeared from the screen), than when the dot appears in the location of the neutral word.

Information-processing tasks have also been developed to examine memory biases in psychopathology. Perhaps the most frequently used paradigm involves an assessment of participants' memory for emotional stimuli presented during an experimental session. Participants are instructed first to encode emotionally valenced adjectives, often by rating how well each word describes them. Participants are then required to complete a follow-up task. In some studies, participants are asked to recall as many stimuli as possible. In other investigations, participants are asked to make perceptual judgments as quickly as possible of words that were or were not presented in the encoding task (e.g., lexical decision task), or to complete word-stems of old and new words with the first word that comes to mind (e.g., word completion task).

All of these information-processing tasks permit an assessment of cognitive processes that, until recently, could not be examined empirically. Findings from investigations utilizing these tasks in the study of various emotional disorders have led to refinements and extensions of cognitive theories of psychopathology, and have provided the impetus for studies attempting to integrate the examination of cognitive and biological factors in psychopathology.

5. Results of Studies of Information Processing and Psychopathology

The results of investigations using these new methodologies have clearly expanded our knowledge of information processing and psychopathology. For example, supporting a schema-based theory of anxiety, there is little question now that individuals diagnosed with an anxiety disorder automatically deploy their attention to stimuli involving themes of danger and threat. Thus, anxious persons have been found to show attentional interference effects for danger/threat words on the Emotion Stroop Task and the Dichotic Listening Task, and have also been found to exhibit selective attention to threat words on the Dot-Probe task (see Gotlib and MacLeod 1997, for a review of this literature). There are also some data to suggest that depressed individuals are characterized by a similar attentional bias to loss/failure stimuli on these tasks. Interestingly, however, there is stronger evidence to support the operation of a memory bias for negative material in depressed persons, a specific bias that does not appear to characterize anxious individuals.

A number of theorists have interpreted these differences in the information processing of depressed and anxious persons as evidence of fundamental differences in the nature of these two disorders. Whereas the attentional bias that characterizes anxious individuals is postulated to reflect the chronically elevated vigilance and heightened anticipation and arousal of this disorder, the memory bias observed in depressed persons is evidence of the rumination and regret often associated with this disorder. Indeed, this 'forward-looking' versus 'backward-looking' difference in the information-processing styles between anxious and depressed individuals has led psychologists to formulate distinct evolutionary theories of these two disorders. These differences in cognitive functioning have also led cognitive scientists to revise existing etiologic theories of depression and anxiety and to make new predictions regarding information processing and psychopathology (e.g., Gotlib and Krasnoperova 1998, Williams et al. 1996).

6. Conclusions and Future Directions

The 1990s witnessed a remarkable surge of interest in information-processing approaches to our understanding of psychopathology. Investigations examining negative cognitive biases in the attention and memory functioning of individuals diagnosed with depression and anxiety have yielded findings that contributed significantly to our understanding of these debilitating disorders. New and innovative methodologies continue to be developed to assess the relation between information-processing characteristics and psychopathology. One important recent direction taken by scientists in this field involves the study of the information-processing styles of individuals who are not currently experiencing psychopathology, but who are at risk for the development of emotional disorders (cf. MacLeod and Hagan 1992). In this context, investigators are beginning to assess biases in attention and memory in individuals who have had previous episodes of depression, and in offspring of depressed or anxious parents. The results of these investigations will undoubtedly help us to gain a better understanding of the factors that increase individuals' vulnerability, or susceptibility, to experience significant emotional disturbance.

Findings from studies examining information processing and psychopathology are also beginning to provide the impetus for investigators to assess the interaction of cognitive and biological functioning in emotional disorders. For example, using functional magnetic resonance imaging (fMRI), scientists are examining patterns of brain activation in response to various types of positive and negative stimuli in patients suffering from depression and anxiety. And other researchers are studying physiological arousal of depressed and anxious patients as they process positive and negative information. The results of such integrative investigations will lead to broader theoretical conceptualizations of different forms of psychopathology.

Finally, although it is still early in the process, it is likely that findings from studies of the information processing of individuals diagnosed with different forms of psychopathology will lead to innovations in the treatment of these disorders. Already, cognitive therapies have a strong foundation of empirical support. As results of information-processing studies continue to influence theories of psychopathology, approaches to psychotherapy based on these findings promise to develop in tandem.

See also: Anxiety and Anxiety Disorders; Attention: Models; Depression; Depression, Clinical Psychology of; Differential Diagnosis in Psychiatry; Perception without Awareness, Psychology of

Bibliography

Beck A T 1976 *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York

Bower G H 1981 Mood and memory. *American Psychologist* **36**: 129–48

Cohen J D, Servan-Schreiber D, Targ E, Spiegel D 1992 The fabric of thought disorder: A cognitive-neuroscience approach to disturbances in the processing of context in schizophrenia. In: Stein D J, Young J E (eds.) *Cognitive Science and Clinical Disorders*. Academic Press, San Diego, CA, pp. 99–127

Dember W N 1974 Motivation and the cognitive revolution. *American Psychologist* **29**: 161–8

Garcia J, Koelling R A 1966 Relation of cue to consequence in avoidance learning. *Psychonomic Science* **3**: 123–4

Gotlib I H, Krasnoperova E 1998 Biased information processing as a vulnerability factor for depression. *Behavior Therapy* **29**: 603–17
Gotlib I H, MacLeod C 1997 Information processing in anxiety and depression: A cognitive developmental perspective. In: Burack J, Enns J (eds.) *Attention, Development, and Psychopathology*. Guilford Press, New York, pp. 350–78
Kihlstrom J F 1987 The cognitive unconscious. *Science* **237**: 1445–52
Kraepelin E 1883 *Lehrbuch der Psychiatrie* [1883; 8th edn. 1915]. Clinical psychiatry. Translated from 7th German edn. by A R Diefendorf: Scholars' Facsimiles and Reprints, Delmar, NY
Lewinsohn P M 1974 A behavioral approach to depression. In: Friedman R J, Katz M M (eds.) *The Psychology of Depression: Contemporary Theory and Research*. Wiley, New York, pp. 157–85
MacLeod C, Hagan R 1992 Individual differences in the selective processing of threatening information, and emotional responses to a stressful life event. *Behaviour Research & Therapy* **30**: 151–61
Oatley K, Johnson-Laird P N 1987 Toward a cognitive theory of emotions. *Cognition and Emotion* **1**: 29–50
Schildkraut J J 1965 The catecholamine hypothesis of affective disorders: A review of supporting evidence. *American Journal of Psychiatry* **122**: 509–22
Venables P H 1964 Input dysfunction in schizophrenia. In: Maher B A (ed.) *Progress in Experimental Personality Research*, Vol. 1. Academic Press, Orlando, FL
Williams J M G, Watts F N, MacLeod C, Mathews A 1996 *Cognitive Psychology and Emotional Disorders*, 2nd edn. Wiley, Chichester, UK

I. H. Gotlib

Information Processing Architectures: Fundamental Issues

The term 'information processing architecture' refers broadly to the arrangement of mental subsystems that are hypothesized to be active in the performance of one or more psychological tasks. For instance, the simplest, most prototypical and opposed types of architectures are serial (one-at-a-time) vs. parallel (simultaneous) arrangement of two or more separate subsystems or processes. More complex arrangements are mentioned below (see *Additive Factor Models*). Further, there are a number of other aspects of perceptual and cognitive processing that are often included under the 'architecture' label including the questions of the basis on which cognitive processes will cease, various kinds of independence and dependence, and processing capacity. These will be described and discussed.

An early researcher especially pertinent to this review was Donders, a nineteenth-century Dutch physiologist. Donders believed that he could uncover the durations taken by various thought processes through his method of subtraction. The method of subtraction was based on the idea that complicated mental activities were compounded in a simple sequential fashion from less complex parts. Let mean response time be written as $\overline{\mathrm{RT}}$ and response time in general as simply RT. Then, for instance, the scientist might engage the subject in a task requiring both perception and decision and compare $\overline{\mathrm{RT}}$ from that condition with $\overline{\mathrm{RT}}$ from a task requiring only perception. The difference in $\overline{\mathrm{RT}}$s was interpreted as the mean duration of the decision operation.

The issues selected for review here seem elemental in the following sense. The construction of almost any system intended to carry out the processing (e.g., detection, search, comparison, recognition, recall, analysis of various kinds, and so on) of a finite number of tasks or items would have to make decisions on each of the studied issues. In addition, they are somewhat unique in having been subjected to quite general theoretical, quantitative, and methodological analysis; perhaps more than any other such concepts in the field. The issues will be introduced in the context of a popular experimental paradigm and then discussed in more detail.

1. An Experimental Paradigm and the Issues

The concepts to be defined below have figured prominently in studies on short-term memory and visual display search and we shall employ that type of paradigm for illustration. Considerable impetus was given to the information processing approach by several pioneering studies in the 1960s focusing on short-term memory and visual processing, using RT as the dependent variable (Atkinson et al. 1969, Egeth 1966, Sternberg 1966). We focus on Sternberg's short-term memory search paradigm (1966) for illustration. Short-term memory is specified operationally by the tasks requiring the retention of a small number of items for anywhere from a few seconds to several minutes. In this paradigm, a varying number (less than or equal to the number that can be maintained in short-term memory without error) of items, for instance, randomly arranged letters, is presented to the experimental participant. This is called the memory set. Then, a few seconds later, that participant is presented a so-called probe item and the task is to indicate very quickly, but at a very high level of accuracy, whether or not the probe was contained within the memory set. As introduced above, RT is the primary dependent variable of interest here, although a substantial history of work with accuracy also exists.

The first, and in some ways the most critical question, referred to the temporal structure of comparison of the probe item with the memory items. A hallmark of many early information processing models was seriality of operation, that is, each subsystem or component could operate only when preceding mechanisms were completed and only one could operate at any given time. Thus, *seriality* has the

connotation that no overlap of processing times occurs, in addition to the sequentiality of starting times. This latter type of constraint is referred to as *discrete flow*. The alternative possibility arose early on that all the items might be searched simultaneously, that is, *in parallel*. This issue of processing architecture can refer to the manner of treatment with regard to items within a single stage of processing, for instance memory search, as in the present case, or to the arrangement of larger components of the system, such as stimulus encoding vs. memory storage.

Of course, arrangements other than serial vs. parallel are possible, for instance, a system being serial part of the time and parallel part of the time. When all operations under focus begin at the same time but can finish at different times, the system (or mode of processing) is of course, simply parallel. However, when operations can feed into other components, but with overlapping processing times, the system is said to have continuous flow, rather than discrete flow. A great deal has been learned about discrete-flow architectures of considerable complexity over the past two decades (see *Additive Factor Models*). We will confine ourselves to serial and parallel systems and other processing characteristics that these can involve.

The primary RT durations under study in the memory search task are those associated with comparison of the probe to the memory items. Nonetheless, other time intervals, such as those associated with early sensory processing and late motor components must also be taken into consideration. It is a usual, but by no means uncontested assumption, that these so-called residual times are serially arranged with the process under study and are also stochastically independent of the latter (Dzhafarov 1992, Luce 1986, Townsend and Ashby 1983).

Besides the architecture issue, *independence vs. dependence* among the processing times of the items is also an important question. Another critical concept is that of *capacity*, which refers to how processing times are affected by the number of things to be worked on (Townsend and Ashby 1983). This is most easily illustrated with a version of parallel processing where the system slows down when the number of items that is being processed increases (that is, *limited capacity*). However, limitations in capacity that are indirect, even with serial processing, can be conceived. For instance, a serial processor might speed up as it goes through the items, due to warm-up effects, or slow down due to inertial factors or fatiguing of the processor. Even though capacity and independence are logically separate notions, they can interact. For instance, an important type of parallel system, one that can mimic serial processing, assumes that as the processing of each item is completed, its processing capacity is reallocated to remaining items (Townsend and Ashby 1983). This obviously affects the overall RT, but also creates a positive dependence among the item processing times.

Yet another important notion is that of *stopping rule*. Depending on the task, it may or may not be necessary for the participant to process all of the items in order to make a correct response. In the memory search paradigm, if the probe item is present and located in the current stimulus set, the processing can cease at that instant, without finishing the remaining items. This possibility is known as *self-termination*. However, since short-term memory search consumes only a few hundred milliseconds it is possible that the system nevertheless completes all items. This event is called *exhaustive* processing. On probe-absent trials, it is necessary to process all of the memory items in order to be sure of correctly making a 'no' response, that is, exhaustive processing must occur. In some experimental designs, all the items are probes. These need not be physically identical. For instance, the task might be to say 'yes' if any of the items is a vowel and a target-present trial could contain all vowels. This latter case permits the possibility of *first-terminating* or *minimum time* processing. Of course, it is an empirical question as to whether any kind of self-termination can actually take place in high-speed perceptual or cognitive operations, and must be addressed experimentally in each case.

Although the serial and parallel class of models both contain an infinite class of possibilities, the serial notion has traditionally been attached to a particular serial model that assumes identical processing time random variables on each item, independent of the number of items in memory (i.e., the load) and of the order of processing the items. It is also often assumed that the individual processing times are themselves mutually stochastically independent. We call this the *standard serial model*. Let the density of processing time for each item be designated $f(t)$ and that for the independent residual processing time be $g(r)$, where T and R are the respective random variables for t and r. Then the density on a probe-absent trial, $p(\mathrm{RT} = t)$, or $p(t)$ for short, is just the mathematical convolution of the n processing densities and the residual time density: $p(t \mid \mathrm{EX}) = f_1 * f_2 * \ldots * f_n$ where each f_i, $i = 1, \ldots, n$, is a replica of. The expected or average theoretical RT is $\mathrm{E(RT)} = n \cdot \mathrm{E(T)} + \mathrm{E(R)}$. Notice that processing is exhaustive, as designated by EX in the left-hand side, in this case. In the case of self-terminating (ST) processing on a probe-present trial, the number of items finished before the probe is found is itself a random variable. Under usual conditions, the probability that the probe will be found in the ith processing position is just $1/n$, so the average density for this case, is just $\mathrm{E}\ [p(\mathrm{T} \mid \mathrm{ST})] = (1/n)\, \Sigma_{j=1}^{n} f_1 * f_2 * \ldots * f_{i-1} * f_i$ times. Similarly, the expected RT, with $\overline{\mathrm{RT}}$ as the sample statistic, is just $\mathrm{E(RT \mid ST)} = (1/n) \cdot \Sigma_{i=1}^{n}$ $\mathrm{E(T)} + \mathrm{E(R)} = [(n+1)/2] \cdot \mathrm{E(T)} + \mathrm{E(R)}$, which yields the time honored result that this line has half the slope of that for the exhaustive serial RT.

The concept of parallel processing, like serial processing, is usually confined in the experimental litera-

ture to a very special kind of parallel processing. Processing is often said to be parallel if average RT is invariant across increases in load. When exhaustive processing has to occur, as in the probe-absent trials of memory search, this stipulation actually implies a very unusual kind of parallel processing, unlike the situation with the standard serial model. For instance, consider a set of parallel models with independent processing times, with the additional provision of unlimited capacity, in the sense that the probability distribution on completion time for each item does not change as n is increased. This class of models serves as a prototype of parallel processing in an analogous way to the standard serial model class just presented. Let us call this class the standard parallel model.

Let $g(t)$ be the fixed density function on processing time for each of the n items. Then $E[RT \mid EX) = E\{MAX[T_1, T_2, \ldots T_{n-1}, T_n]\} + E(R)$, where R is as above and T_i is the processing time of the item in position i. We drop reference to R which plays no role in the function shape, letting PT stand for processing time, that is, $PT = RT - R$. Then $E[PT \mid EX] = \int_{t'=0}^{\infty} [1 - G(t')]^n \, dt'$, where $G(t)$ is the cumulative distribution function associated with g. It is straightforward to show from this formula that mean RT is indeed an increasing, concave function of n (Townsend and Ashby 1983, p. 92). Note that $g(t)$ is invariant across values of n. Hence, many investigators use a very restrictive and typically unrealistic criterion for parallelity when they demand flat rather than increasing functions $\overline{RT}$ of n, even under exhaustive processing conditions.

It can also be observed that if mean exhaustive $\overline{RT}$s are flat functions of n, then mean self-terminating (again, a single position exists in the memory list that contains the probe) times would actually be decreasing, within the same model—a strong prediction which apparently has never been checked in studies using this logic. Thus, in the rare cases where flat exhaustive $\overline{RT}$ functions are actually found with exhaustive processing, the implications are quite strong. Models that can make such predictions are considered in Townsend and Ashby (1983, Chap. 4).

In contrast, in the case of ordinary self-terminating processing when a single position contains the probe, standard parallel models do predict a flat $\overline{RT}$ function. This can be readily intuited, since only the channel which processes the probe matters in self-terminating processing, and it is independent of all the rest (and is of unlimited capacity into the bargain). Finally in this model, mean first-terminating times should decrease with n.

2. *Mimicking Dilemmas*

As mentioned earlier, one of the driving forces behind mathematical meta-modeling in this area was the discovery that mathematical representations of diametrically opposed psychological principles could nevertheless sometimes mimic one another, even to the point of mathematical equivalence of their defining probability laws (Murdock 1971, Townsend and Ashby 1983). Hence, we will take some time here to outline the state of the art with regard to such questions within the present set of issues.

Consider a model for the processing times of n items under fixed experimental conditions. With regard to the parallel–serial issue, suppose no assumptions are made other than the probability mixture of generalized convolutions in the case of seriality and joint distributions on processing times in the case of parallelity. Then the parallel and serial classes of models are equivalent, in the sense that mappings can be provided that homeomorphically carry one joint density into the other (Townsend and Ashby 1983).

Nevertheless, over time, accumulating theoretical results have demonstrated that if the scientist is willing to make further restrictive, but sometimes still very general (and reasonable) assumptions about the models, and/or more complex experimental designs are utilized, the parallel–serial issue can be decided. For instance, certain rather fundamental differences between serial and parallel processing can be explored in experimental methods designed to exploit those differences (Townsend and Ashby 1983).

3. *Factorial Methodologies for Architectures*

One of the most promising and general approaches to identifying mental architecture is that based on the notion of *selective influence* of experimental factors, a notion first employed in tests of strict seriality by Sternberg in his well-known *additive factors method* (1969) (see also *Additive Factors Models*). All factorial methodologies, like the original Sternberg strategy depend on the *selective influence assumption*, namely that distinct experimental factors affect distinct processing components (i.e., subsystems), the assumption of selective influence. It can be assumed that $RT(X + \Delta X, Y)$ refers to the case where the X factor has prolonged RT but Y is at base level, and so on. Basically, the fundamental statistic for the original method and for most extensions, was the *mean interaction contrast*, MIC. The MCI is defined as

$$MCI = \overline{RT}\,(X + \Delta X, Y + \Delta Y) - \overline{RT}\,(X + \Delta X, Y) - [(\overline{RT}\,(X, Y + \Delta Y) - \overline{RT}\,(X, Y)].$$

Schweickert (1978) in his *latent mental network theory* contributed the first major extension of the additive factors method involving more complex architectures under the assumption of selective influence. Townsend and Ashby (1983) found that the

mean interaction contrast distinguished parallel and serial stochastic models when selective influence, was assumed and Schweickert and Townsend (1989) produced general theorems for Schweickert's latent networks, within a stochastic setting, and assuming exhaustive processing. (See *Additive Factor Models*; *Network Models of Tasks*.)

Although the early theorems in all this work were accomplished in the context of exhaustive processing, analogous results can be found in the case of self-terminating and first-terminating processing times (e.g., Townsend and Nozawa 1995). Because Sternberg's original ideas inherent in his additive factors method have been extended in so many new directions, it has been suggested that the general approach be referred to as *systems factorial technology* (Townsend and Thomas 1994). For instance, one novel strategy has been to enlist entire RT distributions in providing more powerful tests of parallel vs. serial processing or other related architectures (Dzhafarov and Schweickert 1995, Roberts and Sternberg 1993, Townsend 1990, Townsend and Ashby 1983; see also Balakrishnan 1994). For instance, in analogy to the mean interaction contrast, the scientist can form a contrast function composed of the double difference (corresponding to the double difference in mean RTs in the usual case) of cumulative distribution functions. This new statistical function turns out to be very helpful in assaying mental architecture (Townsend and Nozawa 1995). Another example of the use of entire distributions will be reviewed below.

The assumption of selective influence is critical to the legitimacy of systems factorial technology. Much has been learned in recent years about its foundational underpinnings, what may go awry if it is violated, and about certain experimental indications of its failure (Dzhafarov 1997, Townsend and Thomas 1994).

4. *Discriminability Results on Other Issues*

The question may be raised as to whether the other processing issues defined earlier also suffer great problems in identification within the search paradigm. With regard to the stopping rule, mathematical investigations have shown, somewhat ironically, that the same types of memory search data that were incapable of deciding the parallel–serial question could in many instances prove that processing was self-terminating rather than exhaustive, although it is more difficult to prove that processing is exhaustive (Townsend and Colonius 1997).

Nevertheless, recent theoretical and empirical discoveries utilizing the entire RT distributions rather than means provide much strengthened tests of architecture (parallel vs. serial) and in addition allow one to firmly distinguish stopping rules at the same time. In fact, assuming selective influence at the distributional level, Townsend and Nozawa (1995; see also Townsend 1990) proposed an experimental design they called the *double factorial paradigm* in which investigators can test architecture and stopping rule as well as capacity within the same block of experimental trials.

The most difficult issue of the ones under discussion is that of stochastic dependence, even though it plays a vital role in processing systems. Interestingly, at this point in time, dependence is more readily and more directly assessed within experimental designs based on accuracy rather than RT (Ashby and Townsend 1986).

5. *Challenges for the Future*

The bringing together of mathematics and experimental methodology has lead to a useful stratagem that we call meta-modeling. It facilitates the development of experimental methodologies that are capable of testing broad classes of models, rather than highly specific models, against one another. This approach is especially helpful when attempting to settle well-defined issues, for example diametrically opposed concepts such as parallel vs. serial processing. It may be more difficult to use this strategy with highly complex and detailed models of phenomena (e.g., Van Zandt and Ratcliff 1995).

Meta-modeling has arguably led to striking advances in the ability to decide experimentally a number of elementary, but major, issues in the purview of human information processing over the last several decades of the twentieth century. Some of these were outlined above. However, many challenges remain. For instance, long-standing challenges relate to the somewhat pesky residual time component introduced earlier. It is difficult to know in many cases how successful our various strategies can be until its nature is firmly tied down. As noted earlier the residual time is usually assumed to be in series with, and stochastically independent of, the other processing components (Luce 1986, Smith 1995, Townsend and Ashby 1983). However, this assumption may be and has been challenged. Thus, it has been contrasted with the extreme opposite assumption of perfect correlation, but still within a series arrangement of the residual component with the other stages of processing (Dzhafarov 1992).

The most time-honored approach has been to attempt to separate the residual time component from those under study employing Fourier analysis, assuming stochastic independence of these components. This has proven to be a tricky and arduous strategy. Luce (1986), and Sheu and Ratcliff (1995) provide useful reviews and commentaries on this approach. Another kind of powerful, if often difficult, approach regarding the residual time component has been to

prove that experimental predictions hold true regardless of its presence (Dzhafarov and Schweickert 1995, Roberts and Sternberg 1993; Townsend and Ashby 1983).

The residual time problem also abuts a more general possibility, namely that processes in the same forward sequence may not satisfy discrete flow, but rather obey continuous flow properties. For instance, almost all systems based on differential and integral equations would fall into the continuous flow class. Some intriguing progress, both theoretical as well as empirical, has been made on certain subclasses of such systems (McClelland 1979, Miller 1993, Schweickert and Mounts 1998), but much more remains to be done.

See also: Additive Factor Models; Discrete State Models of Information Processing; Network Models of Tasks

Bibliography

Ashby F G, Townsend J T 1986 Varieties of perceptual independence. *Psychological Review* **93**: 154–79

Atkinson R C, Holmgren J E, Juola J F 1969 Processing time as influenced by the number of elements in a visual display. *Perception & Psychophysics* **6**: 321–6

Balakrishnan J D 1994 Simple additivity of stochastic psychological processes: tests and measures. *Psychometrika* **59**: 217–40

Dzhafarov E N 1992 The structure of simple reaction time to step-function signals. *Journal of Mathematical Psychology* **36**: 235–68

Dzhafarov E N 1997 Process representations and decompositions of response times. In: Marley A A J (ed.) *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*. Erlbaum Associates, Mahwah, NJ, pp. 255–78

Dzhafarov E N, Schweickert R 1995 Decompositions of response times: an almost general theory. *Journal of Mathematical Psychology* **39**: 285–314

Egeth H 1966 Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics* **1**: 245–52

Luce R D 1986 *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press, New York

McClelland J L 1979 On the time relations of mental processes: an examination of systems of processes in cascade. *Psychological Review* **86**: 287–330

Miller J 1993 A queue-series model for reaction time, with discrete-stage and continuous-flow models as special cases. *Psychological Review* **100**: 702–15

Murdock B B Jr 1971 A parallel-processing model for scanning. *Perception & Psychophysics* **10**: 289–91

Roberts S, Sternberg S 1993 The meaning of additive reaction time effects: Tests of three alternatives. In: Meyer D E, Kornblum S (eds.) *Attention and Performance XIV*. MIT Press, Cambridge, MA, pp. 611–54

Schweickert R 1978 A critical path generalization of the additive factor method: analysis of a stroop task. *Journal of Mathematical Psychology* **18**: 105–39

Schweickert R, Mounts J 1998 Additive effects of factors on reaction time and evoked potentials in continuous flow models. In: Roberts F S, Dowling C, Theuns P (eds.) *Progress in Mathematical Psychology*. Erlbaum Associates, Mahwah, NJ, pp. 311–28

Schweickert R, Townsend J T 1989 A trichotomy method: interactions of factors prolonging sequential and concurrent mental processes in stochastic PERT networks. *Journal of Mathematical Psychology* **33**: 328–47

Sheu C F, Ratcliff R 1995 Fourier deconvolution of reaction time data: a cautionary note. *Psychological Bulletin* **116**: 285–99

Smith P L 1995 Psychophysically principled models of visual simple reaction time. *Psychological Review* **102**: 567–93

Sternberg S 1966 High-speed scanning in human memory. *Science* **153**: 652–4

Sternberg S 1969 The discovery of processing stages: extensions of Donders' method. In: Koster W B (ed.) *Attention and Performance*, Vol. II. North Holland Press, Amsterdam

Townsend J T 1990 The truth and consequences of ordinal differences in statistical distributions: toward a theory of hierarchical inference. *Psychological Bulletin* **108**: 551–67

Townsend J T, Ashby F G (eds.) 1983 *Stochastic Modeling of Elementary Psychological Processes*. Cambridge University Press, Cambridge, UK

Townsend J T, Colonius H 1997 Parallel processing response times and experimental determination of the stopping rule. *Journal of Mathematical Psychology* **41**: 392–7

Townsend J T, Nozawa G 1995 On the spatio-temporal properties of elementary perception: an investigation of parallel, serial and coactive theories. *Journal of Mathematical Psychology* **39**: 321–60

Townsend J T, Thomas R D 1994 Stochastic dependencies in parallel and serial models: effects in systems factorial interactions. *Journal of Mathematical Psychology* **38**: 1–34

Van Zandt T, Ratcliff R 1995 Statistical mimicking of reaction time distributions: Mixtures and parameter variability. *Psychonomic Bulletin and Review* **2**: 20–54

J. T. Townsend

Information Society

It is commonplace to regard 'information' as a defining feature of the contemporary world. Commentators routinely refer to us living in 'information societies,' to our operating in a 'global information economy.' These sorts of observations respond to the indisputable growth of information in recent years, something manifest in new media and media products and in the growth of educational provision, and more generally evidenced in the widespread permeation of information and communications technologies. Though it has become popular to describe societies such as the United States, Germany, and Japan as information societies, as a concept the term is open to criticism.This centers on the claim that the information society represents a novel social system, one which is characterized by differences from what has gone before. Against this, there are other thinkers who insist that activities have become more informationally

saturated, but this has not in any major way changed their basic patterns.

1. Definitions

It is possible to identify six ways of defining an information society, each analytically separate. Five of these may be considered as belonging to a single category, one which seeks to establish quantitative measures of the information society.

1.1 Technological

Technological conceptions center on an array of innovations which have appeared chiefly since the late 1960s. The suggestion is that their impact is so profound as to lead to a reconstitution of the social world. All commentators in this category echo themes previously articulated by Marshall McLuhan (1964), and technological indicators are the most widespread in discussions of the information society (e.g., Feather 1994).

For example, Alvin Toffler (1980) suggests that the world has been impacted by three 'waves' of technological innovation, each as unstoppable as the mightiest tidal force. The first was the agricultural revolution and the second the industrial revolution.The third is the information revolution that is now engulfing us and which presages a new way of living.

More recently, futurism has been stimulated by computing's capacity to transform telecommunications, to merge the two technologies. It is this spread of computer communications technologies, to create what is now often thought of as an 'information superhighway,' that currently inspires most speculation about a new society in the making (Negroponte 1995).

1.2 Economics

This approach charts the proportional growth in economic worth of informational activities (Machlup 1962, Porat 1978). If one is able to plot an increase in the proportion of gross national product accounted for by information businesses then logically there comes a point at which one may declare the achievement of an 'information economy.' Once the greater part of economic activity is taken up by informational activity rather than say agriculture or manufacture, then it follows that we may speak of an information society.

1.3 Occupational

This is the approach that is most favored by sociologists. Here the occupational structure is examined over time and patterns of change observed. The suggestion is that we have achieved an information society when the preponderance of occupations is found in information work: thus the information society is said to have arrived when clerks, teachers, lawyers, and comedians outnumber colliers, steelworkers, dockers, and builders.

In Daniel Bell's (1973) writing, the decline of manufacturing employment and the rise of service sector employment is interpreted as the loss of manual jobs and its replacement with white-collar—and hence informational—work. More recent studies suggest that work is becoming more 'knowledge intensive' (Lash and Urry 1994), activities such as design, media and management requiring more intelligence and education than hitherto (Reich 1992, Leadbeater 1999).

1.4 Spatial

Here the major emphasis is on information networks which connect locations and in consequence can have profound effects on the organization of time and space. A key idea is that electronic highways result in a new emphasis on the flows of information, something which leads to radical revisions of time–space relations. In a 'network society' constraints of the clock and of distance are relieved, corporations and even individuals being capable of managing their affairs effectively on a global scale, with remarkable consequences. This approach has been most authoritatively developed by Manuel Castells in his trilogy, *The Information Age* (1996–8), though Castells (2000) himself refuses to use the concept 'information society.'

1.5 Cultural

The cultural conception of an information society is the most easily acknowledged, yet the least amenable to measurement. Its starting point is that there has been an unprecedented explosion of symbols and signs over recent years. There is simply a great deal more information about than ever before, a point easily appreciated when one reflects on the extraordinary growth of symbols in today's society (media, clothing, even body shape all contribute to this environment). Today we have round-the-clock television, continuous music supply, advertisements at every corner, fashions and styles displayed everywhere—so much information that it can seem perverse not to conclude that this is indeed an 'information age' (Poster 1990).

2. Problems of Definition

Criticisms of attempts to define the information society return us to thinkers who, while acknowledging

the growth of information, refuse to endorse the idea that we have in consequence entered into a new type of world, an information society.

Objections revolve around discontent with quantitative measures when they are used to designate qualitative change. The central criticism is that quantitative indexes of the spread of information and information technologies cannot be interpreted as evidence of deep-seated social change. On the contrary, they can be regarded as the consolidation of established patterns of interest and control.

Definitions of the information society offer a quantitative measure (numbers of white-collar workers, percentage of GNP devoted to information etc), and assume that, at some unspecified point, we enter an information society when this begins to predominate. But there are no clear grounds for designating as a new type of society one in which all we witness is greater quantities of information in circulation and storage.

Furthermore, it is at least intellectually possible to imagine a radically different sort of society coming into being, one that may even merit the title information society, though this transformation may be manifested in only small quantitative increases of information. That is, it may be feasible to describe as a new sort of society one in which it is possible to locate information of a qualitatively different order and function, though quantitatively information changes may appear small. This does not require that we discover that a majority of the workforce is engaged in information occupations or that the economy generates a specified sum from informational activity. For example, it is possible to imagine an information society where only a small minority of information experts hold decisive power (science fiction writers, such as H. G. Wells in *The Time Machine* and Kurt Vonnegut in *Player Piano*, do this where they portray the domination of an intellectual élite).

3. *Theoretical Knowledge*

Indeed, there is a sixth conception of the information society which posits something like this. It suggests that an information society is one in which theoretical knowledge takes precedence over the practical. Humans throughout history responded to conditions as they found them, while today things are organized, harnessed, and controlled by starting from known theoretical principles such as laws of motion, chemical reactions, and the stress capacities of metals. Daniel Bell (1973) first suggested this 'primacy of theoretical knowledge' (though the emphasis of his book was on the numerical growth of white-collar employment). It has been extended by Nico Stehr (1994) who defines a 'knowledge society' as one in which theoretical knowledge is constitutive of virtually everything that we do, from designing new technologies to making sense of our own lives. Such a conception echoes the work of Anthony Giddens (1990) who argues that heightened reflexivity is essential to the conduct of contemporary affairs. Accordingly, there is a heavy reliance on theoretical knowledge, conceived here as that which is abstract and generalizable. For example, sewerage treatment plants, air traffic control, and even dating agencies closely and routinely monitor their (and other) activities, results of which are continuously fed into established but malleable 'expert systems' that are designed around principles aimed at producing optimum results (such as potable water, safety margins for aircraft, success rates of relationships).

There are problems here in clarifying what constitutes theoretical knowledge (Kumar 1978, pp. 219–30), but it is noticeable that this conception evokes a qualitative rather than quantitative change in information/knowledge to designate a new form of society. This emphasis sets it apart from the five other conceptions of the information society reviewed above. These identify an information society as a new type of society by presuming that a qualitative change can be defined by calculating how much information is in circulation, how many people work in information jobs, how much more media output there is, and so on. What we have here is the assumption that quantitative increases transform into qualitative changes in the social system.

4. *Continuity and Change*

Most of the criticisms of information society conceptions come from those who reject the contention that there has been a systemic social change. They refuse quantitative measures of information's growth which exclude analysis of the content of that information while yet alleging that more information will transform society. As a rule, these critics argue that increases in information serve specific interests and thereby function to perpetuate the status quo. It may be helpful if we compare these opposed theories of continuity and change.

There are several theories emphasizing change and information's centrality to this process. Not surprisingly, these endorse a conception of the information society. Of those available Daniel Bell's theory of postindustrial society remains the most robust.

4.1 *Postindustrial Society*

Bell's theory argues that we may trace the transition from a preindustrial, through an industrial, to a postindustrial society by charting shifts in employment. In preindustrial times most people found work in agriculture; in the industrial epoch, employment was chiefly in factories engaged in manufacture; and in the postindustrial society most work is in services. Since most service work is nontangible person-to-

person activity, then the material with which people operate in postindustrialism is information. This contrasts with work in agricultural societies where people struggle against nature, and it is different to industrial society where work was chiefly with machinery. Accordingly, postindustrialism is synonymous with an information society.
Criticisms include:

> Postindustrial theory is historically dubious. For instance, the great expansion of the service sector, evident in all advanced societies during the late twentieth century, has been fueled by dramatic declines of employment in agriculture rather than in manufacture. Such empirical patterns cast doubts on Bell's claim that societies have shifted employment from agriculture, through industry, to a service society (Kumar 1978, pp. 200–4).

> The theory is neoevolutionist, implying a directionality to history that is questionable. Critics regard this vision of change as unwarranted in view of events such as world wars, genocide, and the re-emergence of ethnically based nationalisms in recent decades, not to mention the threats posed to the environment by phenomena such as global warming, nuclear energy, and acid rain.

> Bell argues that the motor of change is increased productivity, which allows first a transfer of workers from the land to industry and later a shift from industry to services without loss of wealth. Continually increased productivity in agriculture and industry is required to fund the uninterrupted expansion of services. To critics of Bell this is a deterministic explanation of change.

> Bell misunderstands the expansion of the service sector, seeing it as a consumer of resources supplied by industry and agriculture. Bell 's theory posits that agriculture and industry are regarded as the wealth-creating sectors of society and that gains achieved there allow society to spend that wealth on services that use it up. However, this perception ignores the intimate ties of the service, agricultural, and manufacturing sectors in an integrated economy, in particular being blind to the central function of producer services (e.g., banking, finance) in the day-to-day operation of an advanced economy. The real economy is not divisible into neat halves, but is rather integrally connected into an organic whole. As such, Bell's entire theory is cast into doubt (Gershuny and Miles 1983).

> The features Bell highlights as heralding the coming of postindustrialism are all long-term aspects of industrial development (e.g., rationalisation, growth of services, expansion of white-collar work), so how can he claim that these announce a new postindustrial age? (Kumar 1995).

4.2 Informationalized Capitalism

Turning from this account, we encounter the work of Herbert Schiller, whose argument is that the information explosion is part and parcel of the advance of capitalism in and beyond the late twentieth century.

There are three key refrains in the writing of Schiller. The first draws attention to the pertinence of market criteria in informational developments. In this view it is essential to recognize that information and communications innovations are decisively influenced by the market pressures of buying, selling, and trading in order to make profit. To Schiller the centrality of market principles is a powerful impulse towards the commodification of information, which means that it is, increasingly, made available only on condition that it is saleable.

The second insists that class inequalities are a major factor in the distribution, access to, and capacity to generate information. Bluntly, class shapes who gets what information and what kind of information they may get. Thereby, depending on one's location in the stratification hierarchy, one may be a beneficiary or a loser in the 'information revolution.' The decisive factor in access to information is the ability to pay, something which ensures that within advanced societies like the UK and America there is an information rich that can purchase virtuoso services, while there is an information poor (Schiller 1996) that must be satisfied with what Schiller calls 'garbage information' (mass entertainment, advertisement-supported television, junk mail, etc.), while the same process is manifest globally with information rich societies such as the United States and information poor who are reliant on Hollywood leftovers to fill their impoverished television schedules and on Western news agencies to report what is happening around the world (and even in neighboring countries).

The third key contention of Herbert Schiller is that the society which is undergoing such momentous changes in the information areas is one of corporate capitalism. That is, contemporary capitalism is one dominated by corporate institutions which are concentrated organizations that command a national, and generally international, reach. For this reason, in Schiller's view, corporate capitalism's priorities are especially telling in the informational realm. At the top of its list of priorities is the principle that information will be developed for private rather than for public ends.

Clearly these are established features of capitalism. Market criteria and class inequalities have been important elements of capitalism since its early days, and even corporate capitalism has a history extending over a century, though many of its most distinctive forms have appeared in the late twentieth century. But to Herbert Schiller this is precisely the point: the capitalist system's long-established features are the key architectural elements of the so-called 'information society.' From this perspective those who consider that informational trends signify a break with the past are not credible since, asks Schiller, how can one expect the very forces that have generated information to be superseded by what they have created?

The 'information society' therefore reflects capitalist imperatives and, simultaneously, these informational developments sustain and support capitalism. In this

way Schiller accounts for the importance of information in ways which at once identify how the history of capitalist development has affected the informational domain and, at the same time, how information has become an essential foundation of that historical development (cf. Sussman 1997).

Several criticisms may be leveled at Schiller's analysis of the information domain. These include:

> The possibility that social change may be examined on at least two different axes: on the one hand, along an axis of feudalism to capitalism, while on the other, along an industrial to a postindustrial/information society axis. One might consider the argument that Schiller and Bell talk across one another because they are looking at different aspects of social change. Schiller's account paints an unrelentingly negative picture of the spread of information under the aegis of corporate capitalism, one which suggests a nostalgic vision of information before the intrusion of capitalism. But just how valid is this depiction in view of our knowledge of illiteracy rates even in the nineteenth century, lack of education, and the demonstrable inadequacy of reading material for working people?
>
> Schiller's focus on information inequalities may be a useful remainder of class differences, but his binary division into the information rich and poor is sociologically crude.

5. *Conclusion*

Though the concept of the information society has achieved considerable popularity in and outside academia, it appears to have limited value for social scientists. The most acclaimed analyst of information today, Manuel Castells (2000) has even recommended that 'we should abandon the notion' (p. 10). As we have seen, the defining criteria are vague and imprecise. The concept has some heuristic value (Lyon 1988), in so far as it encourages interest in and debate about developments that are of inarguable significance. However, the most serious objection to the concept is that the claim that it represents a profound social transformation cannot be supported on the basis of the quantitative indices that are usually advanced. Clearly, there are more information and communications technologies, information is of greater economic weight, and there are more information occupations. But the idea that these signal a shift towards a new society is mistaken. As Krishan Kumar (1995) insists, the information explosion

> has not produced a radical shift in the way industrial societies are organized, or in the direction in which they have been moving. The imperatives of profit, power and control seem as predominant now as they have ever been … The difference lies in the greater range and intensity of their applications … not in any change in the principles themselves (p. 154).

This being so, it is ironic that the most persuasive conception of an information society, that which centers on the role of theoretical knowledge, is the least examined by information society adherents.

See also: Cultural Studies: Cultural Concerns; Culture, Sociology of; Globalization and World Culture; Globalization, Anthropology of; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Information Society, Geography of; Information Technology; International Communication: History; International Communication: Regulation; Internet: Psychological Perspectives; Mass Media, Political Economy of; Science and Technology: Internationalization

Bibliography

Bell D 1973 *The Coming of Post-Industrial Society: A Venture in Social Forecasting*. Basic Books, New York

Castells M 1996 *The Information Age. Vol. 1, The Rise of the Network Society*. Blackwell, Oxford, UK

Castells M 1997 *The Information Age. Vol. 2, The Power of Identity*. Blackwell, Oxford, UK

Castells M 1998 *The Information Age. Vol. 3, End of Millennium*. Blackwell, Oxford, UK

Castells M 2000 Materials for an explanatory theory of the network society. *British Journal of Sociology* **51**(1): 5–24

Feather J 1994 *The Information Society: A Study of Continuity and Change*, 2nd edn. Library Association, London

Gershuny J I, Miles I D 1983 *The New Service Economy: The Transformation of Employment in Industrial Societies*. Pinter, London

Giddens A 1990 *The Consequences of Modernity*. Polity Press, Cambridge, UK

Kumar K 1978 *Prophecy and Progress: The Sociology of Industrial and Post-Industrial Society*. Allen Lane, London

Kumar K 1995 *From Post-Industrial to Post-Modern Society: New Theories of the Contemporary World*. Blackwell, Oxford, UK

Lash S, Urry J 1994 *Economies of Signs and Space*. Sage, London

Leadbeater C 1999 *Living on Thin Air: The New Economy*. Viking, London

Lyon D 1988 *The Information Society*. Polity Press, Cambridge, UK

Machlup F 1962 *The Production and Distribution of Knowledge in the United States*. Princeton University Press, Princeton, NJ

McLuhan M 1964 *Understanding Media: The Extensions of Man*, 1st edn. Abacus, London

Negroponte N 1995 *Being Digital*, 1st edn. Stoughton, London

Porat M U 1978 Communication policy in an information society. In: Robinson G O (ed.) *Communications for Tomorrow*. Praeger, New York pp. 3–60

Poster M 1990 *The Mode of Information: Poststructuralism and Social Context*. Polity Press, Cambridge, UK

Reich R B 1992 *The Work of Nations: Preparing Ourselves for 21st Century Capitalism*. Vintage, New York

Schiller H I 1996 *Information Inequality*. Routledge, New York

Stehr N 1994 *Knowledge Societies*. Sage, London

Sussman G 1997 *Communication, Technology and Politics in the Information Age*. Sage, Thousand Oaks, CA

Toffler A 1980 *The Third Wave*, 1st edn. William Morrow, New York

F. Webster

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Information Society, Geography of

The 'information society' is a historically specific term, referring to contemporary societies in which the production and exchange of information, dramatically accelerated by the growth of information technologies (IT), has become a defining feature of social activity. Information exchange has always been central to the functioning of society, but only since the 1950s has it become broadly separable from face-to-face communication and identified as a distinct sector of economic production. The 'geography of the information society' explores how IT, telecommunications structures, and the geography of human activity are co-evolving in such societies, paying attention to the two-way relationship between technological developments and sociospatial processes, and to the associated growth of geographic information and geographic information technologies.

The geography of the information society is related to the somewhat older subdiscipline of communications geography, but itself is a recent and rapidly evolving area of study. The first major publications on this topic date from the late 1980s (Hepworth 1990, Robins 1992, Kellerman 1993). Earlier work focused mostly on economic aspects of the geography of the information society, but increased attention is being paid to cultural and political aspects. This article focuses on geographical processes driving the development of IT, the geographical organization of the production and use of these technologies, and their impacts on space, place, and the geographic scale of influential societal processes.

1. *Geography and the Information Revolution*

The digital computer drives the explosion of IT that are qualitatively increasing humankind's ability to communicate remotely. Telegraphs first enabled spatial flows of information to be separated from face-to-face communication, and thus from geographical restrictions on human movement. The telephone and the telefax extended this capacity but virtually all telecommunications are now integrated into digital computer networks.

Information technologies are the latest in a very long history of communications technologies which have been developed to satisfy three processes fundamental to a spatially extensive society: (a) The need for the production and exchange of commodities to occur increasingly rapidly over greater and greater distances, 'the annihilation of space with time' (Harvey 1985 p. 35); (b) the need to exert control over larger and larger social systems of increasing geographic scope; and (c) the human desire for liberation from the limitations of daily life in particular places, by learning about and experiencing other places. The information revolution can be seen as transcending the shipbuilding yards, ships, and ports of the nineteenth century, and the automobile factories, highways, and garages of the twentieth century, with the Silicon Valleys, satellite systems, fiber optic cables, and teleports of the twenty-first century.

The widespread availability of geographic information and geographic IT is increasingly important as this revolution proceeds. Computing and transmission networks can now readily handle the complexity of spatial data and the digital maps that can be produced with them. Geographic information systems (GIS) are increasingly available for manipulating and displaying such data. A wide range of other geographic IT producing or analyzing geographically referenced information, such as satellite reconnaissance systems and geographical positioning systems (GPS), are increasing the capacity of certain social actors to access geographic information immediately almost anywhere around the globe.

2. *The Spatial Organization of the Information Economy*

As the production, storage, and exchange of information have become key sectors of economic activity, a new economic geography has been emerging. In part this reflects the location of producers of IT, and partly it is a result of the impact of IT on the locational patterns of other economic activities. Particular attention has been paid to the mushrooming of new clusters of innovative IT firms, in such places as Silicon Valley, Cambridge, UK, and Tokyo-Yokohama, Japan. These are paradigmatic examples of technology districts, which have grown as a result of dynamic external economies created by the elevated pressure to both collaborate and compete with nearby rivals. This is contrasted with a different locational logic for the mass production activities associated with producing IT, assembling chips and computers, and reproducing software. Branch plants performing these activities can be found where lower wage workers are available, both within the advanced capitalist countries and in newly industrializing countries and export processing zones in the third world.

An emerging spatial division of labor, between dynamic industrial districts where innovative activities and intellectual labor cluster, and branch plant regions dominated by assembly activities employing unskilled labor, is itself a result of the impact of IT on the geography of production. The clusters represent 'new industrial spaces' that benefit from the ability of small firms using computer-assisted manufacturing technologies to operate as efficiently as large factories, from the advantages of agglomerating close to other such small and dynamic firms, and from 'just in time' methods of minimizing inventories by using IT to ensure that supplies are delivered as they are needed, as in Toyota City, Japan. Information technologies

also enable the operation of subsidiaries to be controlled from distant headquarters. Different components of the production process can then be located worldwide, wherever labor costs, environmental regulations, and agglomeration economies are best suited to that activity, creating commodity chains of global scope.

Such a spatial division of labor is restructuring geographical differences in the location of economic activities, thereby contributing to uneven development, even as it links distant places more closely together with the help of IT. Knowledge-based industries cluster in the new industrial spaces described above, but also in the national and global cities that attract corporate headquarters seeking access to information and expertise. These cities also host clusters of producer service industries providing the information and expertise sought. There is an extraordinary clustering of such activities in a handful of cities worldwide, reflecting the symbiotic nature of knowledge-based activities (Castells 1989).

Attention has only recently turned to the location of the infrastructure through which information flows. Notwithstanding popular perceptions that distance no longer matters in the information society, it has been pointed out that accessibility to the means of transmitting information varies greatly geographically (Graham and Marvin 1996). A number of cities are actively seeking to develop smart office districts and teleports, in the belief that the absence of these is hindering their attractiveness for knowledge-based industries. The availability of computers varies significantly at intra-urban and international geographical scales: Internet addresses cluster in the largest cities, and in the business districts of those cities; teleports are found in relatively few cities; and fiber optic lines entwine the developed world, barely reaching, for example, the African continent. Spatial and social inequality of access to the means of telecommunication has significant implications for the impact of the information society on space and place. This is because investments in telecommunications infrastructure are durable investments that shape the location of activities relying on them, in turn creating demand for new infrastructure investments in the same areas.

Geographic IT are beginning to have a distinct impact on the location of economic activities. Transportation behavior is altering as a result of intelligent highway monitoring systems, on-board navigation systems, and the use of GPS to monitor commercial transportation operators and cargoes. Military surveillance is more effective, and bombing campaigns more spatially selective, as a result of the development and use of such technologies. This affects locations chosen for strategic economic activities (such as weapons manufacture) and is changing the spatial impact of bombing campaigns on the economic infrastructure of a country.

3. *Impacts on Space*

The information society is widely depicted as instrumental in the decline of geographical distance as a significant impediment to societal processes, as well as the growth of globalization. In this view, the capacity of information networks to transfer information or money as easily across the globe as across the street has two interrelated effects. First, there is no reason for human activities to agglomerate together in geographic space; cities and industrial districts are no longer necessary, because distance does not inhibit communication (Mitchell 1995). Second, is the integration of distant events through information networks: the global integration of finance markets, the transnationalization of firms' geo-economic strategies, and the decline of the nation state—harbingers of a global village where social processes operate at the global scale, and thus of the end of geography (Martin 1994). Such claims are far from novel. At the end of the nineteenth century similar claims were made for the telegraph, and they reappear with each new leap forward in communications (Marvin 1988). Yet cities continue to grow and uneven development persists.

In fact, the spread of information networks has contradictory impacts on space, simultaneously reducing and increasing the distance between social actors. Distant territories and markets are connected together in new ways, but not in ways that eliminate inequalities in accessibility. The emerging spatial metric is better captured by the metaphor of spatial wormholes than that of the global village. While certain social actors have access to the channels of cyberspace, through which they are instantaneously connected with related actors in distant places, their immediate neighbors may not be connected at all. This means that the geographical propinquity of social actors does not imply that they are similarly situated with respect to other social actors and places. '[S]pace is not "shrinking," but rather must be perpetually recast' (Kirsch 1995, p.544). Information networks are restructuring the social distance between actors, and thus their situation with respect to one another, in ways that loosen, but do not eliminate, the relation between social distance and physical distance. The ways in which this is happening depend on how and whether different types of social actors, in different places, gain access to particular IT and information.

Information technologies have reduced spatial barriers between different aspects of working life. Elimination of distance barriers has made it easier to relocate routine production and information processing activities to distant lower wage locations, with US corporations placing such back-office activities as telemarketing in San Antonio, Texas, insurance claims in Ireland, or processing the information on consumer warranty cards in the Caribbean (Goss 1995, Warf 2000). This diffusion of routine information-processing activities is counterbalanced, however, by the

agglomeration of nonroutine knowledge-based activities in large national or global cities which become centers of control. The reduction of distance barriers is, thus, also associated with increased inequality.

Processes of suburbanization and exurbanization have also been accelerated by reducing these barriers. Back-office activities are relocated from downtown to suburban areas, and away from inner city work forces, to tap a white female labor force willing to work in such jobs. The growth of telework and telecommuting, as information technology makes it possible to separate work from workplaces, is contributing to the spatial dispersion of urban activities into the countryside. This is widely predicted to reduce differences between urban and rural life, and also personal travel and its social and environmental costs.

In discussing back-office relocation and the effect of telecommuting on urban form, a great deal of attention is given to the degree to which telecommunication can replace face-to-face communication. The need for nonroutine face-to-face interaction has long been recognized as an important determinant of the agglomeration of offices as places for nonroutine information exchange. Analysis of technology districts has similarly emphasized the importance of face-to-face contact in developing the shared knowledge, rivalries, and governance systems that sustain local economic dynamism (Storper 1997). Notwithstanding the continuing evolution of telecommunications technologies, it is argued that they are likely to remain an inadequate substitute for face-to-face communication for much critical decision-making. The kinds of activities that are likely to be remotely located with the help of IT are those that can be completed successfully using relatively routinized communications, and can be coordinated, and monitored, at a distance.

New IT also continually alter the need for, and skills associated with, particular tasks, altering the locations where those tasks are best performed. Telecommunication made it easier to decentralize routine information processing into back offices, only for that process to be reversed as new information systems eliminate the need for manual data entry and expert systems replace telephone operators. Less skilled jobs are replaced with fewer but more skilled software and computer operators and these activities relocate to centers of control (Warf 2000). Similarly, the decentralization of IT associated with the growth of desktop computing and the Internet has led to the recentralization of information storage in clusters of mainframe computers known as 'server farms.' The particular tasks that can be decentralized alter, but a spatial division of labor endures between well-paid knowledge-based activities, based on face-to-face communication, and routine activities in more peripheral and lower-wage locations.

Questions about the degree to which IT will undermine the spatial cohesion of urban areas are paralleled by questions about the desirability of the counter-urbanization which is occurring. Lower density lifestyles may in fact place a greater burden on the physical environment than higher density lifestyles, because high density settlements can economize in transportation and energy consumption. While journey-to-work trips are reduced by telecommuting, other kinds of trips will increase as workers have more freedom during the working day, and it is not clear that the net effect will be reduced travel (Graham and Marvin 1996).

Information technologies are affecting the cultural geography of society, facilitating both personal mobility and the spread of information about remote places as a result of the growth of 'personable extensibility'; that is, the ability of individuals or groups to extend their lifeworlds to embrace distant places (Adams 1995). Telecommunications accelerate the diffusion of cultural norms and trends between cultures, and the homogenization and globalization of cultural norms, particularly the worldwide popularity of US movies, TV programs, popular music, consumption habits, and ideas. Yet the same processes have differentiating effects: US culture is becoming more multicultural as it absorbs cultural influences from elsewhere; French culture is revived in defense against Anglophone influence; and local cultures are learning to use IT to enhance their visibility and coherence.

4. *Impacts on Place and Identity*

Reflecting claims about the impact of the information society on space, it is commonly argued that place as an organizing theme for social life is withering away. Two aspects of place are stressed: the economic coherence and vitality of places; and the relationship between personal identity and place. Arguments about the former reflect those about space. 'The dominant tendency is toward a horizon of networked, ahistorical space of flows, aiming its logic over scattered segmented places…' (Castells 1996, p. 428). In this view the hypermobility of investment capital, facilitated by IT networks and by related processes of economic political, and cultural globalization, undermine the coherence of places and the ability of their inhabitants to control the future of those places. Arguments about the latter discuss how IT undermine the ways in which everyday life is tied into, and shaped by, the places where this occurs; that is, the relationship between identifying with a place and individual identity.

The city of London, a global financial center whose economic livelihood depends on instantaneous calculations of risks and returns drawing on worldwide information flows, has been studied intensively to examine whether place matters in the information society. In this case, place matters greatly. The city 'is a system… dependent on electronic communications but… the increasingly intricate patterns of connection that are able to be produced… trigger off demands for more rather than less face-to-face interaction'

(Leyshon and Thrift 1997, p. 292). The very complexity and quantity of information arriving through the spaces of flows must be interpreted. Thus, a defining feature of economic life in the city is personal contact, familiar cultural rituals of communication, and reliance on personal experience and visual clues to assess the expertise and reliability of informants. As a consequence, key decisions are still taken within the compact space of the city, reinforcing both a sense of and an identification with this place, and enhancing its growth and power nationally and globally. A similar symbiosis between information technology and place exists in other global financial centers.

The possibility that place can be reinforced in an information society has been noted for other clusters of economic activities. Places able to benefit from dynamic external economies are capable of developing and maintaining competitive advantages which draw in and retain investments even in an increasingly interdependent information society. It is even argued that any place can use this strategy to anchor its economic future, although others argue that inherited uneven development means that the search for competitive advantage can reinforce spatial inequality (Leitner and Sheppard 1998).

The commercial deployment of geographic IT may also enhance the importance of place. By classifying consumers into a large number of 'lifestyles' and using GIS and census data to attach one of these categories to each census tract, firms have developed cost-effective ways of geographically targeted marketing—techniques also adopted to target voters in elections. One consequence of such strategies is to shape the choices, consumer purchases, voting, and identity of people in each place, and the meaning of the places in which these practices occur (Goss 1995, Curry 1998).

On the other hand, by reducing the degree to which social practices are associated with particular places, the information society reduces attachment to place and the cohesion of place. Cyberspace is pointed to as exemplifying this. Use of the World Wide Web has made it easier to join electronic communities of geographically scattered individuals sharing particular interests, substituting interaction and identification with such communities for interacting and identifying with geographical neighbors. In addition, the increasing ability of IT to create immersive environments, or 'mirror worlds,' may make experiences with real places less attractive than those with virtual ones. New senses of space and time are being created; the Internet creates a capacity to be in two places at once, and distance is measured by communications time instead of geographical separation. Yet the kinds of identity that result, as identity construction is detached from place, may require less interpersonal responsibility and may never be fully separable from those formed through living in real places (Robins 1995).

Notwithstanding the 'spaceless' nature of cyberspace, the cognitive principles used to make sense of it and the metaphors used to describe it draw heavily on those developed as a result of human experience with material space. Immersive environments are organized by conventional spatial principles; their immersive character stems from their ability to seem indistinguishable from real places; and cyberworlds tend to emulate the spatiality of material worlds. The differentiated geography of communications infrastructures and of access to cyberspace means that there are close relationships between cyberspace and material space (Kitchen 1998, Janelle and Hodge 2000).

The Internet has become a medium which local place-based groups have been able to utilize to promote and gain global attention for their particular agendas, such as the Zapatista movement in Chiapas (Castells 1997). At the same time, the information society is promoting supranational economic and political processes, ranging from global finance markets, to multinational information systems for monitoring migrants and refugees, and the homogenization of practices governing the exchange of geographic information and intellectual property rights (Curry 1998). The information society is contributing to processes of 'glocalization,' whereby both local scale and global scale processes are growing simultaneously in importance (Swyngedouw 1997), but is erasing neither distance nor place.

See also: Computers and Society; Globalization and World Culture; Globalization: Geographical Aspects; Information Society; Information Technology; Information, Economics of; Place in Geography; Spatial Pattern, Analysis of; Spatial Thinking in the Social Sciences, History of; Telecommunications and Information Policy; Telecommunications: Policy

Bibliography

Adams P C 1995 A reconsideration of personal boundaries in space-time. *Annals of the Association of American Geographers* **85**: 267–85

Castells M 1989 *The Informational City*. Blackwell, Oxford, UK

Castells M 1996 *The Rise of the Network Society*. Blackwell, Oxford, UK

Castells M 1997 *The Power of Identity*. Blackwell, Oxford, UK

Curry M R 1998 *Digital Places: Living with Geographic Information Technologies*. Routledge, London

Goss J 1995 We know who you are and we know where you live: The instrumental rationality of geo-marketing information systems. *Economic Geography* **71**: 171–88

Graham S, Marvin S 1996 *Telecommunications and the City: Electronic Spaces Urban Places*. Routledge, London

Harvey D 1985 *The Urbanization of Capital*. Blackwell, Oxford, UK

Hepworth M E 1990 *Geography of the Information Economy*. Guilford Press, New York

Janelle D, Hodge D (eds.) 2000 *Accessibility in the Information Age*. Springer Verlag, Berlin

Kellerman A 1993 *Telecommunications and Geography*. Belhaven Press, London

Kirsch S 1995 The incredible shrinking world? Technology and the production of space. *Environment and Planning D: Society and Space* **13**: 529–55
Kitchen R M 1998 *Cyberspace: The World in Wires*. Wiley, New York
Leitner H, Sheppard E 1998 Economic uncertainty, inter-urban competition and the efficacy of entrepreneurialism. In: Hall T, Hubbard P (eds.) *The Entrepreneurial City*. Wiley, London
Leyshon A, Thrift N 1997 *Money/Space: Geographies of Monetary Transformation*. Routledge, London
Martin R 1994 Stateless monies, global financial integration and national economic autonomy: The end of geography? In: Corbridge S, Martin R, Thrifit N (eds.) *Money Power and Space*. Blackwell, Oxford, UK
Marvin C 1988 *When Old Technologies were New*. Oxford University Press, New York
Mitchell W J 1995 *City of Bits: Space, Place and the Infobahn*. MIT Press, Cambridge, MA
Robins K (ed.) 1992 *Understanding Information: Business, Technology and Geography*. Belhaven Press, London
Robins K 1995 Cyberspace and the world we live in. In: Featherstone M, Burrows R (eds.) *Cyberpunk/Cyberspace/Cyberbodies*. Sage, London
Storper M 1997 *The Regional World: Territorial Development in a Global Economy*. Guilford Press, New York
Swyngedouw E 1997 Neither global nor local: "Glocalization" and the politics of scale. In: Cox K R (ed.) *Spaces of Globalization: Reasserting the lower of the Local*. Guilford Press, New York
Warf B 2000 Telecommunications and economic space. In: Sheppard E, Barnes T J (eds.) *Companion to Economic Geography*. Blackwell, Oxford, UK

E. Sheppard

Information Technology

Information technology (IT) covers any form of technology, that is, any equipment or technique used by a company, institution, or any other organization which handles information. The term information technology was coined probably in the late 1970s to refer to this nexus of computer-based technologies for handling information. It incorporates computing, telecommunication technologies, and includes consumer electronics and broadcasting as it is getting more and more digitized. Spearheaded by the computer, the decades since the mid-1960s have been characterized by extreme development. Since the late 1970s, cheap microelectronics have permitted the diffusion of these technologies into almost all aspects of daily life and have furthermore almost inextricably cross-fertilized and intermingled their multiple application branches, which include industry, commerce, administration, education, medicine, scientific and professional work, entertainment, and domestic work.

Nations with advanced IT-industries have realized that developing competence in information technologies including new media is essential, expensive, and difficult; large-scale information technology systems are gaining economic feasibility and various national research and education programs for stimulating these developments have been founded. The fundamental capabilities that usually are perceived as being essential are: VLSI (Very Large-Scale Integration) chip design, production facilities, and a common infrastructure like the Internet for the storage and transmission of digital multimedia information (including video, audio, and other continuous media data in addition to conventional data and text).

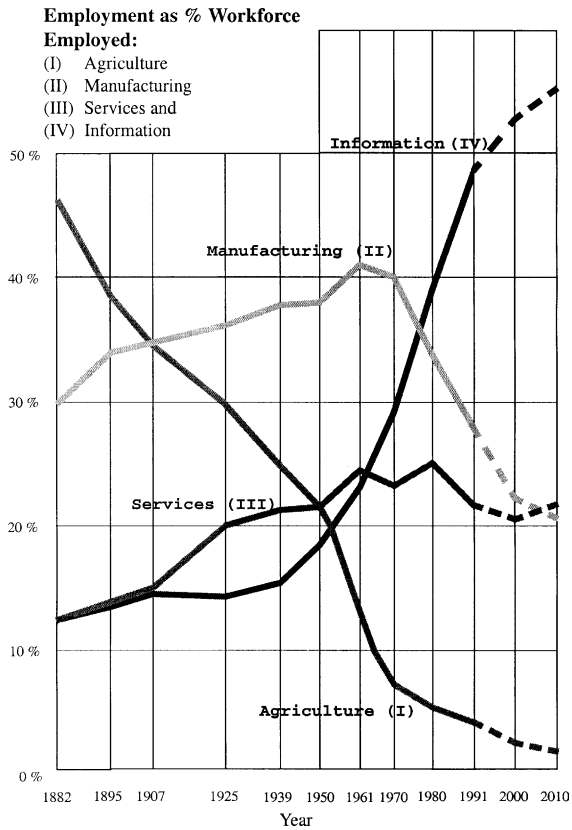

Figure 1
Structural change in the workforce employed in the sectors (I) agriculture, (II) manufacturing, (III) services and (IV) information in the German economy for about the last 100 years (see Grauer and Merten 1997)

For a detailed treatment of the above topics and for further references see *Organization*: *Overview*; *Network Analysis*; *Intelligence*: *Organizational*; *Learning*: *Organizational*; *Research and Development in Organizations*.

1. *Information Technology and Impact on Society*

Computers are transforming the way people communicate, do business, like to entertain, and educate

themselves. This technological evolution has not only had a profound impact on the business community, but also on private and professional lives. The computer evolution, however, has given rise also to serious social issues including the vulnerability of personal information and job security. This evolution has an impact on the whole society. In Fig. 1 the effect of structural change in the employment according to four sectors of the German economy is shown. This pattern of structural change can be observed as similarly in all national economies only with a time shift. The evolution of IT-systems is the basis for the rise of employment in the information sector and the decline in manufacturing by increasing quality and quantity of production. The employment in manufacturing follows the pattern of agriculture in the mid-twetieth century.

The reason that probably all parts of the society are touched by IT can be seen in the fact that the price of computing has dropped by half every 2–3 years (see Fig. 2) in opposite to the increasing costs of producers' durable equipment.

There have been increasing levels of business investment in IT equipment. These investments account for over 10 percent of new investments in capital equipment in developed countries. On the other side the revenues for the information and communication systems industry have increased tremendously since their beginning in the early 1950s. Since the 1990s their importance dominates that of the automobile industry.

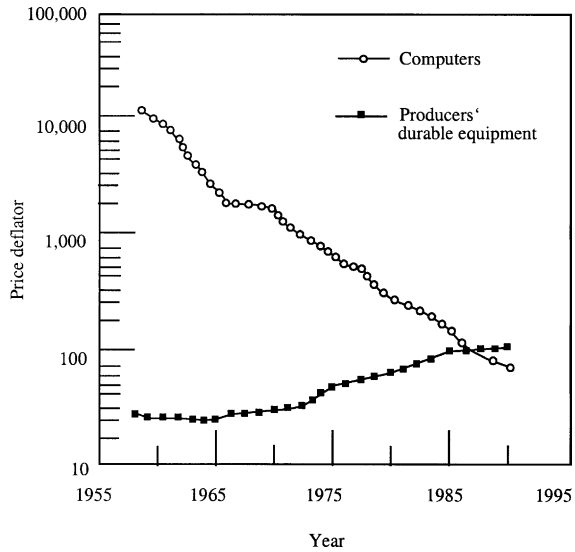

Figure 2
The decline of the cost of computing relative to other capital purchases (see Brynjolfsson 1993)

2. *Historical Background*

The abacus is most likely one of the first mechanical counting devices. It has been traced back 5,000 years. Another important step towards computers as we know them today was the counting-wheel. The French mathematician and philosopher Blaise Pascal (1623–62) finally invented and built the *Pascaline* in 1642 to help his father, a tax collector. This counting-wheel design was used by all mechanical calculators until the mid-1960s.

Everyone, from bankers to navigators, used mathematical tables during the industrial revolution. However, these hand-calculated tables were usually full of errors. Charles Babbage (1793–1871) envisioned a steam-powered 'differential engine' and then an 'analytical engine' that would perform calculations accurately. Although Babbage never perfected his devices, they are forerunners for many concepts used in the modern general-purpose computer. A further important step in the development of IT was the introduction of punched-card tabulating machines by Herman Hollerith (1860–1929). The machine was devised for the US Bureau of the Census in 1890 and had an incredible impact on the necessitated time for completing that year's census, reducing it from the usual 8 to a mere two and a half years. In 1896, Hollerith founded the Tabulating Machine Company and in 1924 the company's general manager, Thomas Watson, changed its name to International Business Machine Corporation, a company which has dominated the development of computers for several decades.

The beginning of modern computer history can be marked by the invention of the Z3 computer (electric relays controlled by a program) by the German engineer Konrad Zuse in 1941 and the ABC-computer by Dr. John V Atanasoff, a professor at Iowa State University, a year later. Both systems used the essential principles of modern computers and were based on binary systems.

The key component of these first computers was the electronic tube. These vacuum tubes acted as electronic switches, replacing earlier electromechanical relays. These machines took up a lot of space and generated considerable heat, so that special air conditioning was required; their reliability was relatively poor. Around 1960, vacuum tubes were replaced by solid-state transistors, which were smaller in size, more reliable, generated less heat and operated at much higher speeds. Instructions would be executed significantly faster by these new generation of computers, but due to high expenses, only the largest companies could afford them. In 1963, the Digital Equipment Corporation introduced the PDP-8, which is generally considered to be the first successful minicomputer. It was the answer to an increasing demand for small computers for businesses and science.

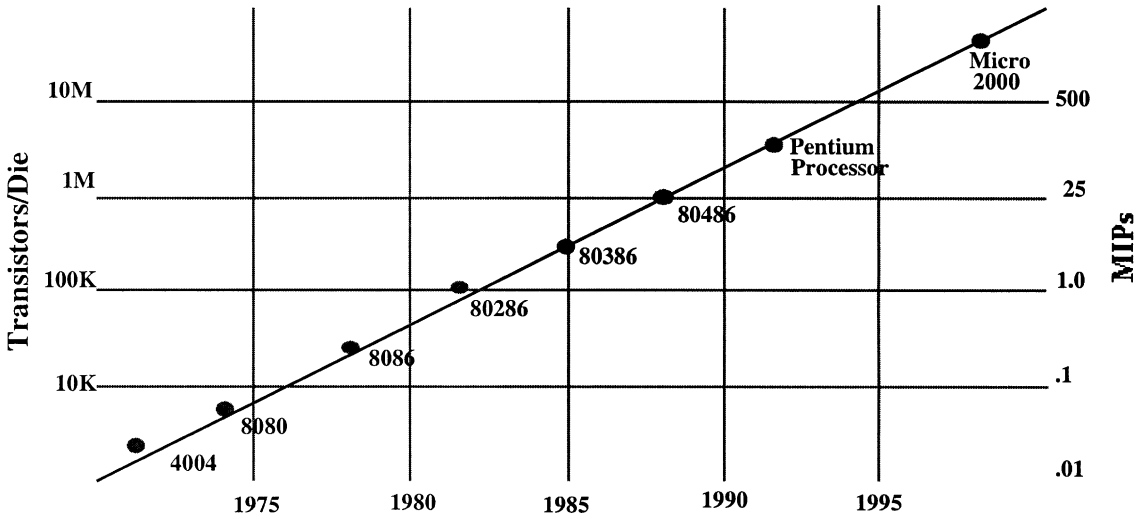

Figure 3
The development of the numbers of transistors on a chip and the performance (MIPs; million instructioned per second) known as Moore's Law

The next generation of computers, in the 1970s, were characterized by miniaturized integrated circuits. The IBM System 360 line of computers, introduced in 1964, is seen as a leading system of this time. The System 360 line was designed as a family of computers with upward compatibility to meet the different and growing demands in the companies. Up to the end of the 1970s analysts describe the main force behind IT evolution by the so called Grosch's Law, which explains the economies of scale in computer design (Zerdick et al. 1999). Grosch's Law says that computer performance is a function of costs squared—twice the cost gives four times the performance.

The next period of IT, in the 1980s and the beginning of the 1990s, is characterized by very large-scale integration of circuits, the introduction of the personal computer, and multiple processors in one machine. The driving force behind IT in this period of time is described by the so-called Moore's Law, which means that the number of transistors that can fit on a chip doubles every 18 months (see Fig. 3).

Since the mid-1990s the IT development is dominated by a rapid worldwide growth Internet users. This global communication system is the basis for a new type of business—the Network Economy. The key feature of this period is described by the so called Metcalfe's Law: The value of a network goes up as the square of the number of its users (Shapiro and Varian 1999).

3. *Information Technology and Management*

Apart from the principal economic resources of business—capital, material, and labor—information is relevant increasingly as an input factor. Management decisions are intuitive and expedient without information but not necessarily economical. Access to information and the computer's ability to answer questions and solve problems with high speed and accuracy improves decision-making and helps management to realize objectives, which was not possible in the past. The application of information systems is a key resource in gaining a competitive position. There are several advantages:

(a) customer service has been improved and ex-tended;

(b) the level of efficiency has been raised. Sales and office staff save both time and effort recording information necessary for operations and control;

(c) load and capacity potential has increased;

(d) information accuracy has improved, enhancing the firm's image and reducing loss as a result of errors;

(e) the query and interactive capabilities provide the right information at the right time to the right person; and

(f) management receives important information for planning, control, and decision making.

Certain applications of IT are universal and equally appropriate for many business environments. These so-called functional applications can be found in most organizations and include functions like payroll, accounts receivable, accounts payable, general ledger, inventory management and control, human resource development, budgeting, project management, customer relationship management, supply chain management, electronic business and workflow management. Workflow management refers here to computer-based applications associated with general office work and

involve word processing, spreadsheets, computer graphics, database management software, electronic mail, image processing (facsimile and image scanner), voice processing (voice message switching and teleconferencing as a group productivity tool, call center operations), and office information systems.

Information systems that are designed to process routine business activities across multiple functional areas of large enterprises are called Enterprise Resources Planning (ERP) systems. ERP systems provide such organizations with highly solutions that rely on the use of common database systems. Linkages between different functional activities within an enterprise and across is organizational boundaries are achieved by maintaining all relevant data in a single database. Thus data corresponding to a business transaction are entered into the system only once, and the effect of this transaction on different business processes and corresponding reports can be determined immediately.

Some computer applications are unique to a particular type of business, such as production scheduling (manufacturing, typesetting (publishing), actuarial accounting (insurance), and special effects (multimedia systems) in marketing, advertisement, and entertainment (movies, online theatre scripts). Robotics, computer-integrated manufacturing (CIM), and computer-aided design (CAD) are prominent examples for product development and manufacturing.

Information systems designed especially for top-level managers are called decision support systems (DSS). In contrast, transaction processing systems (TPS) are tools in supervisory management and executive information systems (EIS) refer to information systems utilized by middle management for tactical planning. There are no generally accepted definitions to the above-named information systems.

Higher-level information systems serve a different purpose than those designed for managers at operational levels. The latter systems specify actual operations. Systems for top managers process feedback data and data from external sources in order to forecast trends and put figures into context—information used by middle management to decide what must be done to improve the firm's competitive position.

4. Conclusions

The development of high-performance, high-reliability, comprehensive communication networks, both intra-organizationally and interorganizationally, is emphatically not stationary. Concurrently, both hardware and software technologies are evolving in ways that make it possible to maintain large amounts of information on-line and to have access to this information in conjunction with the communication networks from almost any location with the available distributed databases. Furthermore, an improvement in user-friendliness and the increased capability of advanced personal computers and workstations enables users to work with these systems with little formal training. The improvements include the interconnection of computers in local area networks as well as in regional, national and international networks. In addition, operations on personal computers, workstations, or mainframes are becoming increasingly conferment. The tendency towards open systems, which allow common sets of software to be transferred from one computer make to another, is the basis for this development.

Current development includes the problem of information overload of business managers. What managers need is support in evaluating and selecting data which represents important information for a given problem situation and to transform this data into knowledge for improved decision-making. IT is only beginning to deal with these issues.

See also: Information Society; Information Society, Geography of; Information, Economics of

Bibliography

Brynjolfsson E 1993 The productivity paradox of information technology. *Communications of the Association for Computing Machinery* **36**(12): 67–77

Elliot G, Starkings S 1998 *Business Information Technology: Systems, Theory and Practice*. Addison Wesley Longman, Reading, MA

Grauer M, Merten U 1997 *Multimedia: Entwurf, Entwicklung und Einsatzin betrieblichen Informationssysteme*. Springer-Verlag, Berlin

Long L, Long N 2000 *Computers*, 8th edn. Prentice-Hall Business Publishing, New York

Shapiro C, Varian H R 1999 *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, Boston

Senn J A 1998 *Information Technology in Business: Principles, Practices, and Opportunities*, 2nd edn. Prentice-Hall Business Publishing, New York

Zerdick A, Picot A, Schrape K, Artope A, Goldhammer K, Lange U T, Vierkant E, Lopez-Escobar E, Silverstone R 1999 *Die Internet-Ökonomie: Strategien für die digitale Wirtschaft*. Springer-Verlag, Berlin

M. Grauer

Information Theory

1. Introduction

Radios, telephones, televisions, and, last but not least, computers are devices that store, process, and transmit information; today they belong to the basic equipment of the average citizen in the industrialized world. It is

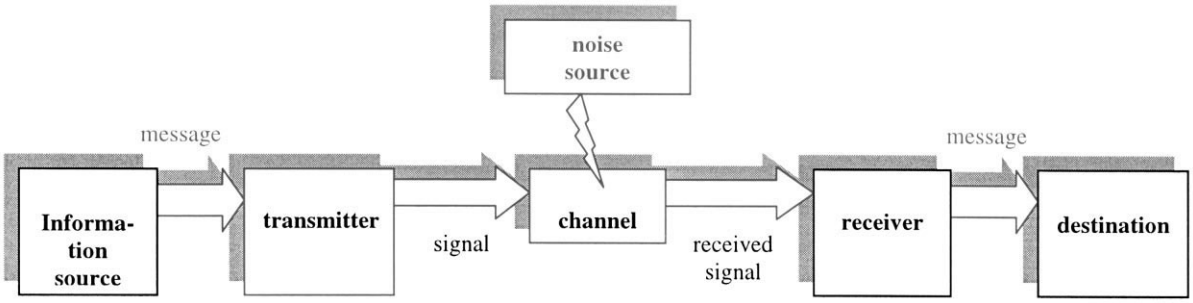

Figure 1
A typical channel for information transmission

no exaggeration to claim that the twentieth century witnessed the greatest revolution of communication and information technology since the invention of the printed book. This vigorous development created the need for a well-founded theory and a rigorous terminology that would allow discussion about information with scientific precision. In 1948, the US electrical engineer Claude Shannon provided a synthesis of the fundamental ideas in classical information theory (1948). For the sake of historical precision it should be noted that some of these ideas had already been put forward, although in less rigorous terms, by Hartley (1928). Yet it was Shannon's achievement to set information theory on firm theoretical grounds. He also discovered the basic mathematical laws of systems that communicate information. Separating relevant information (signal) from irrelevant information (noise) was one of the problems Shannon also tackled and solved. Another problem that Shannon essentially solved was the setting of upper bounds on what can be achieved with a given information-carrying medium or channel. For these seminal contributions to information theory, Shannon is considered its founder.

In telephony for instance, there is a source of information, namely a person speaking into a telephone. The message is the sound that is produced, while the transmitter is composed by the microphone and the equipment that changes this sound into an electromagnetic wave, which becomes the signal. The channel is the space between the transmitting and the receiving telephone, i.e., the receiver. Any static or other phenomenon distorting the signal is called noise. An unambiguous specification of the particular message chosen by the information source has to be transmitted so that the original message can be reconstructed at the receiving point. The receiver reconverts the received signal into an audible sound. Information theory deals with the problem of specifying a quantitative relation between a given message and the set of all possible messages from which it is chosen (see Fig. 1).

2. *Quantifying Information*

Imagine a simple game for two players: one player thinks of a number between 1 and 64 and the other has to guess what the number is by asking only questions that can be answered by 'yes' or 'no.' The strategy followed by a player could be, for instance, to begin by asking the question: 'Is the number less than or equal to 32?'. If the answer is, say, yes, then the player could ask whether it is less than or equal to 16. If the answer is, say, no, then the remaining set of candidate numbers is 17, 18, 19, …, 32. One possible optimal question at this moment—clearly not the only optimal question—would be: 'Is the number less than or equal to 24?'. Clearly, the least number of yes-no questions required to determine the number is 6. Roughly speaking, the amount of information contained in the message describing an event out of a set of possible events corresponds to the number of yes-no questions needed to isolate that event from all others in the set, following an optimal strategy. If, in the game above, the number to be guessed were, for instance, 23 then the sequence yes-no-no–yes-no-yes, beginning from right to left, is a code for the number 23 that

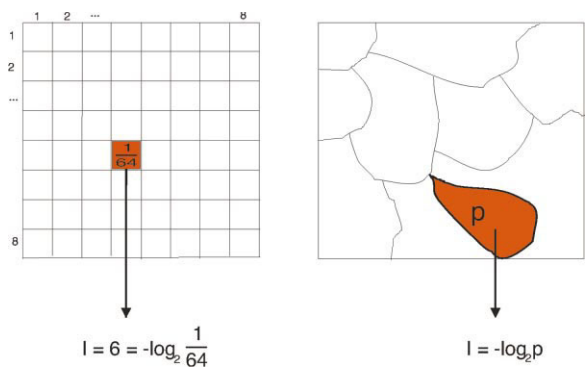

Figure 2
Quantifying information

corresponds to the strategy described above and the specification of the message would then be this ordered sequence of 'yes' and 'no' obtained as answers. The unit for measuring information is the bit. One bit of information corresponds to a message describing one event out of two equally likely events. Two bits correspond to a message describing an event out of four equally likely events. One event out of 2^N equally likely events has N bits of information, or, in a somewhat more suggestive form, it has $-\log_2 1/N$ bits of information (see Fig. 2).

If an event with probability $1/N$ has $-\log_2 1/N$ bits of information, it is natural to extend this measure to an event with probability p by establishing that it must have $-\log_2 p$ bits of information. It was shown by Shannon that the negative logarithm of the probability is the only function satisfying all the properties one expects from information (see also Luce 1960). Characterizing the properties of this measure for information and the concepts related with it was one of Shannon's breakthroughs. Given a set of N events with probabilities $p_1, p_2, \ldots, p_N$, the amount of information associated with the ith message is $-\log_2 p_i$ so that the expected value of information per message is given by

$$\mathrm{H} = -\{p_1 \log p_1 + p_2 \log p_2 + \cdots + p_N \log p_N\}$$

This quantity represents the uncertainty contained in the source. Shannon had originally called this expression the average information of the message set. It was John von Neumann who convinced Shannon to introduce the term 'entropy' for average information, seeing the natural connection between average information and entropy in thermodynamics. The term entropy stems from thermodynamics and was introduced by Rudolf Clausius, who formulated the second law of thermodynamics in 1850 (the entropy of an isolated system is always increasing until it reaches its maximum). He composed the term entropy with the two Greek roots 'tropein' (turn) and 'en' (into) in analogy with the term 'energy'. Entropy takes on its smallest value, zero, when there is one message that is certain to occur (i.e., the described event has a probability 1) and all other messages never occur. Intuitively, there is no information in a message that was *a priori* certain. Maximal entropy occurs when all messages are equally likely, which corresponds to the situation of greatest uncertainty about outcomes. The ratio of the uncertainty of a source to the maximal entropy possible on the same set of symbols is the relative entropy of the source, and

$$\text{“}1 - \text{relative entropy”}$$

is the redundancy of the source. The specification of a message as a (finite) sequence of binary symbols (say, for instance, 0 for 'no' and 1 for 'yes') is called a 'code word' for the message. An important feature of the average information, H, is that it determines the saving in transmission time that is possible, by proper encoding, because of the statistics of the message source.

Based on the concept of information Shannon introduced measures for information rate as well as for channel capacity. He stressed the fact that information has all the properties of matter: it can be transported, compressed, and segmented into small pieces. The diagram in Fig. 1 can be viewed as a transportation system. The channel can be seen as transporting matter from a source to a receiving site. In this situation, there are two important quantities: the rate at which information is produced at the source and the capacity of the channel. Obviously if the rate of production is greater than the capacity of the channel, it will be impossible to transport the full output of the source. Suppose, however, that there is a device—the encoder—at the source that, just like a mill, cuts information down into smaller pieces, then 100 percent efficiency can be achieved. Shannon proved that, if the information rate at the source is less than the channel capacity, then encoders and decoders can be constructed such that the source output is recreated at the destination with as small an error probability as desired. Conversely, if the information rate at the source is greater than the capacity, then the error probability cannot be made arbitrarily small.

Information and related concepts have led to a unified and theoretically sound treatment of uncertainty and reduction of uncertainty in statistics and decision theory. In the set of all distributions on a finite sample space the distribution that maximizes entropy is the uniform distribution, for which all single elements are equally probable. Laplace's *uncertainty principle* (1812) advised to assume that a distribution in uniform, whenever no further knowledge is at hand. For far more than one century probabilists and statisticians searched for an adequate extension of the uncertainty principle to the case when partial information on a distribution is available. Information theory provided the ideal extension via the concept of entropy: it was called the *maximum entropy principle*, or MaxEnt. It states that whenever some of the marginals of the distribution are known, one should assume that the distribution is the one—and only one—that maximizes entropy in the set of all distributions with those marginals. By doing this one guarantees that no additional information is added to the distribution. A family of important algorithms has been developed for the construction of the MaxEnt distribution (Bishop et al. 1975). The MaxEnt principle and algorithms for constructing the MaxEnt distributions form the fundamental ingredients for the designation of *a priori* distributions when events are not equiprobable.

An excellent, modern survey of information theory and its applications was provided by Cover and Thomas (1991).

3. Conclusions

Information theory has had a major impact on behavioral science and on cognitive neuroscience. Psychologists have discovered important relationships between the amount of information in a stimulus and reaction time to the stimulus. An important finding concerns the adaptation of participants to the statistics of information conveyed by presented stimuli and their ability to recognize redundancy of information in stimuli. This leads to shorter reaction times. Another important finding concerns human short-term memory: Pollack (1953) and Miller (1956) noted that basic information is coded in 'chunks.' The ability to form chunks enables organisms to make better use of their limited capacity short-term memory. Because short-term memory appears to be capable of holding only five to seven chunks, efficient information processing depends on the ability to pack more information into a single chunk. Grouping into doublets (two-bit-chunks) and recoding (e.g., 00 = A , 01 = B, 10 = C, 11 = D) can double the number of bits that can be held in short-term memory at a given time.

Forming triplets or larger groupings further increases the number of memorized bits in a nonlinear way. There are constraints to this procedure. Chunking is only effective with sufficiently long presentation-times. With presentation-times under 0.5 seconds, no successful grouping takes place (Kleinberg and Kaufman 1971). Yet with presentation-times of 1 s and longer there is a clear advantage in using chunks. An excellent, by now classical, survey of applications of information theory in psychology was provided by Attneave (1959).

Information theory has allowed a treatment of redundancy in language. For example, the sentence "Sm mn in ths rm us glsss" can be understood as 'Some men in this room use glasses', although several vowels have been deleted. Shannon studied the redundancy of English language and discovered that English texts can be reconstructed when 50 percent of the letters, chosen at random, are deleted. Other important applications of information theory deal with word perception. If the source vocabulary is small, humans are able to perceive words nearly perfectly, whereas perception becomes increasingly difficult as the source vocabulary becomes larger (Miller and Selfridge 1950).

In cognitive neuroscience, information theory has become an essential tool for understanding brain function. Modeling the brain as an information processing machine has led to important advances in both theory and practice. The production of action potentials (spiking) of neurons can be discretized and interpreted as a binary process. In this way, a temporal sequence of spikes can be represented as a binary vector of 0 and 1. All information coming from sensory inputs is processed by neurons using this simple binary language. Hebb conjectured (1949) that neural information processing cannot be understood just in terms of the behavior of individual neurons. Rather, the brain must be viewed as a system of so-called cell assemblies, wherein information processing occurs as an emergent property resulting from the interaction of these cell assemblies. Hebb's conjecture has received a great deal of experimental confirmation in the last decades of the twentieth century (Gray et al. 1989). The famous binding problem, or problem of identifying an object by recognizing individually and 'binding together' features that characterize the object, appears to be solved by means of synchronization of spiking activity (Singer 1994). Theoretical neuroscientists have developed many different approaches to modeling how information is stored in the brain. The most convincing and persistent model has been the associative memory (e.g., Palm 1980) that stores similar patterns close to each other, where the closeness between patterns is measured in terms of information overlap.

Bibliography

Attneave F 1959 *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods and Results*. Holt, New York

Bishop Y, Fienberg S, Holland P 1975 *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA

Cover T, Thomas J 1991 *Elements of Information Theory*. Wiley Series in Communications, New York

Gray C M, König P, Engel A K, Singer W 1989 Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **388**: 334–7

Hartley R V L 1928 Transmission of information. *Bell System Technical Journal* **7**: 535–63

Hebb D 1949 *The Organization of Behavior*. Wiley, New York

Kleinberg J, Kaufmann H 1971 Constancy in short-term memory: bits and chunks. *Journal of Experimental Psychology* **89**: 326–33

de Laplace P 1812 *Théorie analytique des Probabilités*. Courcier, Paris

Luce R D 1960 The theory of selective information and some of its behavioral applications. In: Luce R D (ed.) *Developments in Mathematical Psychology*. Free Press, Glencoe, IL, pp. 1–119

Miller G A, Selfridge J A 1950 Verbal context and the recall of meaningful material. *American Journal of Psychology* **63**: 176–85

Miller G A 1956 The magical number seven, plus minus two: some limits on our capacity for processing information. *Psychological Review* **63**: 81–97

Palm G 1980 On associative memory, *Biological Cybernetics* **36**: 19–31

Pollack R B 1953 The assimilation of sequentially encoded information. *American Journal of Experimental Psychology* **66**: 421–35

Shannon C 1948/1962 The mathematical theory of communication. In: Shannon C, Weaver W (eds.) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana IL, pp. 3–91

Singer W 1994 Time as coding space in neocortical processing: A hypothesis. In: Buzsaki G, et al. (eds.) *Temporal Coding in the Brain*. Springer Verlag, Berlin

L. Martignon

Information, Economics of

1. Introduction

Prior to the 1980s, the 'economics of information' usually meant the economics of search. Search began in the 1960s with the seminal research by George Stigler, which spawned an enormous and growing literature. Both search and noncooperative game theory are considered in this article. As we will see, they are closely related and much can be gained by fostering a closer relation.

Stigler modified competitive price theory by noting that price is a random variable with a given cumulative distribution function. He thereby enlarged price theory to include the search activity necessary to gather information about prices. He defined search as follows: 'A buyer (or seller) who wishes to ascertain the most favorable price must canvas various sellers (or buyers)—a phenomenon I shall term "search".' The author of the famous text *The Theory of Price* knew that search theory would receive widespread attention because of its intimate link with price. That is, the primary reason for the almost immediate influence of Stigler's work was his focus on price uncertainty. Any such major change in price theory eventually would reverberate throughout all of economics. This is what has occurred since the 1960s.

The game theory meaning of the 'economics of information' centers around the formation, continuation, and dissolution of contractual relationships. The incentives of parties to a contract are determined by the different information possessed by each. To exemplify these asymmetric information problems, consider an auction of a valuable object with the following information structure. There are two buyers and one seller. One buyer is willing to pay more for the object, while the other has more information about the object's relative price. If the buyers can collude, 'the Nash program directs us to view this collusion as the outcome of a process of communication where each player has choices about what to say. In this process, the more-informed buyer may try to mislead the less-informed buyer about the value of the object, and the less-informed buyer should rationally take this possibility into account' (Myerson 1999, p. 1075). All problems of incentive-based 'economics of information' can be formulated as Nash noncooperative games.

While Stigler's information model fits snugly into standard price theory, Nash's noncooperative game theory complements price theory, but is not reducible to it. Stigler's fit is snug because he takes the distribution of prices as exogenously determined. This corresponds to the basic assumption of price theory—consumers accept prices as exogenous. Thus, the equilibrium problem is not addressed. Nash enlarged the scope of price theory so interactive behavior can be studied in almost any social institution. This was apparent only after the contributions by Harsanyi, Selten, Kreps et al., thereby explaining the long period between Nash's discovery and its adoption by the economics profession.

In recent years these two research programs have started to merge. Search is embedded in Abraham Wald's sequential analysis in that it determines when the process of information collection should terminate. Gideon Schwarz (1994) notes that Wald used the theory of two-person zero-sum games to unify statistical inference. Statistical inference was viewed as a game played by the statistician against nature. Statistics from this game theory perspective was called decision theory. One of the most important topics in decision theory is optimal stopping. Thus, game theory is as applicable to search as it is to other informational problems in economics.

2. Search Theory

2.1 Introduction

Modern search theory solves the problem of when to terminate the collection of information and take an action. Search is the confluence of two seemingly unrelated intellectual streams: the formal evolution of optimal stopping rules beginning with Cayley (1875) and the verbal, common sense literature found in Knight (1921) and implicit in Adam Smith. The first stage of this confluence occurred when Stigler (1961) blended these two streams with his fixed-sized-sampling rule. The second stage replaced the fixed-sized sampling rule with an optimal stopping rule, thereby combining the insights of Stigler, Bellman, and Wald.

The third stage seeks the source of an endogenous distribution of prices (wages) to replace the exogenous distribution in the Stigler and sequential models. This large and complex research effort was initiated by Rothschild (1973). The fourth stage replaced the one-sided sequential model with a two-sided equilibrium problem, where there is competition and interaction among optimizing agents. This ongoing research is based on the matching and bargaining models devised by Diamond, Jovanovic, Maskin, Mortensen, and Pissarides.

In the early work on search, many economists placed a much higher value on the verbal insights of common sense, when they sought a solution to a search perplex. Mathematical formulations of these insights were undervalued. Today, the situation is somewhat reversed in that formal analysis is ap-

plauded, while common sense is depreciated. Both attitudes diminish economics. The first does not encapsulate insights in a formal model. Hence, testing and extensions are difficult if not impossible. The second closes off economics from the imaginative intuitions of common sense.

2.2 Intuitive, Common Sense Precursors to the Economics of Search

See Lippman and McCall (1993) for further detail.

In 1930, Cannan clearly saw the *moral hazard* problem associated with unemployment insurance. He also noted that the best strategy for a job-seeker was to 'Take what you can get now or hold out another week, when something better may turn up.' With the advantage of hindsight this looks very much like the reservation wage policy.

In 1932, Hicks recognized the unanticipated unemployment effects of the search which is connected to the presence of imperfect information, when he claims that firms' 'knowledge of opportunities is imperfect [whence] it is not surprising that an interval of time elapses between dismissal and re-engagement during which the workman is unemployed.' His statement that 'the unemployment of the man who gives up his job in order to look for a better one' recognizes that voluntary unemployment is joined with search as a productive activity.

An early explicit statement viewing search unemployment as a productive activity was made by Hutt in 1939.

> A worker … may refuse immediately available work … because he feels that to accept it will prevent him from seeking for better openings. … When actively searching for work … he is really investing in himself. … He is doing what he would pay an efficient employment agency to do—the search for a better opening is worth the risk of immediately foregone income (Hutt 1977, p. 83).

In addition to the labor literature anticipating the job search model, other prominent precursors in the economics of search include Knight, Schumpeter, and Alchian who were concerned with the search for natural resources, the search for inventions, and evolutionary economics.

Knight (1921) was keenly aware of the ancient activity of prospecting for natural resources 'where the possibility of securing wealth by the discovery of natural resources is known … resources will be attracted into the field of searching for them in accordance with men's estimates of the chances of success in relation to the outlays to be incurred.'

Schumpeter (1942) continued the advance toward Stigler's introduction of search in his concern for uncertainty and the opportunistic contractual behavior which uncertainty fosters.

> Practically any investment entails, as a necessary complement of entrepreneurial action, certain safeguarding activities such as insuring and hedging. Long range investing under rapidly changing conditions … is like shooting at a target that is not only indistinct but moving—and moving jerkily at that. Hence it becomes necessary to resort to such protecting devices as patents or temporary secrecy of processes or, in some cases, long-period contracts secured in advance.

In Alchian's (1951) evolutionary view, innovations may result from bungled imitation as well as from conscious efforts to improve. He considers innovation, like prospecting for natural resources, as a form of search.

It should also be noted that Simon (1951) suggested a stochastic theory of the employment relation resembling Wald's sequential analysis rule. In this remarkable paper, Simon also studied credible threats and promises, reputation and cooperation in a pre-game theory environment. He thereby suggested a connection between the two meanings of the 'economics of information.'

No one contributed more to the informal literature on the importance of information to economic decisions than Hayek. While not specifying an explicit search rule, Hayek (1978) clearly saw the importance of the concrete knowledge and special decisions made by millions of dispersed individuals. The key which facilitated the transition from *particular* knowledge to a *specific* decision was the price system:

> The recognition that a man's efforts will benefit more people and on the whole satisfy greater needs, when he lets himself be guided by the abstract signals of prices rather than by perceived needs, and that by this method we can best overcome our constitutional ignorance of most of the *particular* facts, and can make the fullest use of the knowledge of *concrete* circumstances widely dispersed among millions of individuals, is the great achievement of Adam Smith.

Thus for Hayek the crucial economic problem is the search for and use of dispersed knowledge whose common source is the price system.

The evolution of job search reveals how useful economic models are generated by *both* formal mathematics and common sense. The origin of the mathematical strand may be quite remote from practical applications, while the original intuitive insight may have little if any formal content. Nevertheless, as both strands evolve a threshold emerges. When this threshold is hit, a moment of insight occurs which unifies the two strands into a new economic model, illuminating a whole set of previously invisible phenomena. The success of the new approach is gauged by empirical data analyzed by novel econometric procedures that are frequently novel.

2.3 The Formal Development of Search Theory

See Lippman and McCall (1993) for further detail.

Stopping rules emerged from a very practical problem: Johannes Kepler's search for a wife in 1613. His procedure was a version of the secretary problem,

which was formally designed by Arthur Cayley in 1875 and rejuvenated more precisely by Leo Moser in 1956. (See Ferguson (1989) for a scholarly treatment of the secretary problem's evolution.)

In 1961, D.V. Lindley solved the following version of the secretary problem. The decision maker selects from n candidates who are presented in random order, with no recall of rejected candidates. The ability of the candidate is revealed at the presentation. This information tells the decision maker nothing about the abilities of the remaining candidates. The payoff function is 1, if the best secretary is chosen, and 0, otherwise. The optimal policy for large n is to interview and reject n/e of the candidates, and then choose the next candidate with the highest rank observed. The probability of choosing the best candidate is $1/e$, where e equals 2.718.

In 1960, MacQueen and Miller posed and solved the house-hunting problem. Independent, identically distributed offers $X_1, X_2, \ldots$ arrive one at a time. The return from stopping after n observations is $R_n = \max(X_1, \ldots, X_n) - nc$. MacQueen and Miller found the stopping rule N which maximized ER_N, where E is the expectation operator.

In 1961, Stigler formulated his fixed sample size model. The buyer obtains n price quotations for the item he wishes to purchase. A cost c is incurred each time a quote is made. The mathematical formulation of Stigler search is: choose the sample size n to minimize $E\min\{X_1, \ldots, X_n\} - nc$.

In 1970, the house-hunting problem was applied to job search in a simple model by McCall. The fixed sample size of Stigler's model is replaced by a stopping time N and the sequential decision model of search is given by: select a stopping time N to maximize $E\max\{X_1, \ldots, X_n\} - Nc$.

The optimal stopping time is: accept the first offer which exceeds the reservation wage ξ. The reservation wage ξ is the solution to the first-order condition

$$c = \int_x^\infty (y-x)\mathrm{d}F(x) \equiv H(x) \tag{1}$$

where $H(x)$, the marginal gain from one more offer, is strictly decreasing and convex. The first-order condition has ξ as its unique solution since $H' < 0$. The economic interpretation of the first-order condition is: equate the marginal cost c of taking an additional offer with the expected increase in the return from taking one more search. It has been shown that ξ is the expected return from search (see Lippman and McCall 1976). The time until an acceptable offer is obtained, a period of frictional unemployment, is a geometric random variable with mean $[1-F(\xi)]^{-1}$.

2.4 Equilibrium Price (Wage) Dispersion

Rothschild (1973) raised the important question of the existence of an equilibrium price dispersion, namely, what gives rise to the nondegenerate price distribution assumed in simple search models. This question provoked an enormous literature on equilibrium search models which sought a positive answer to Rothschild, thereby negating Diamond's (1971) finding that all sellers charge the monopoly price. In Diamond's model, buyers are identical, sellers are identical, and buyers have the same search cost. If the rules are reversed—buyers instead of sellers make take-it-or-leave-it offers and sellers incur search costs—then markets clear at the competitive price. The preferred way out of this dilemma is to introduce heterogeneity into the model. In McMillan and Rothschild (1993), there are several sources of heterogeneity among the 31 papers surveyed. They include differences in sellers' production costs, buyers' search costs, buyers' beliefs about the offer distribution, sellers' time varying prices, buyers' inventory holdings, and advertising which yields differential buyer information. Burdett and Mortensen (1998) obtain a unique equilibrium solution to a game where a continuum of employers establish wage offers and a continuum of workers search sequentially. If workers search when unemployed *and* employed, wage dispersion is achieved.

2.5 Two-sided Search and Bargaining

Since the early 1980s, the early one-sided optimization search models have been generalized to two-sided equilibrium search models. This research combines game theory with search theory by introducing bargaining into the formation of matches. The steady-state equilibrium is the intersection of best reply functions for workers and employers, i.e., a Nash equilibrium, *when utility is nontransferable*. If utility is perfectly transferable, the matched pair bargains until either an agreement or an impasse is reached. The Nash solution is the equilibrium of this cooperative game. See Burdett and Wright (1998).

McMillan and Rothschild (1993) note that the joint study of bargaining and search fortifies each. In the search theory used by Diamond (1971) and in the price dispersion models described above, after a seller is located by the searching buyer, the seller presents a take-it-or-leave-it price offer. But in many circumstances these offers lack credibility, i.e., the seller has an incentive to bargain if his offer is rejected by the buyer. In bargaining theory, it is usually assumed that a single buyer and a single seller are locked in their negotiations. In fact, each often has the option of searching for another trading partner.

2.6 Econometric Methods

The sequential search model gave rise to a sequence of novel econometric methods beginning with the work

of Heckman and Singer; and Kiefer and Neumann. Structural estimation has been developed to estimate sophisticated models that have grown from the simple sequential model.

The impetus for structural estimation was Lucas' insight that reliance on reduced-form estimation in policy analysis gives rise to systematic forecasting errors. This insight is a consequence of the rational expectations hypothesis, which is a version of the Nash equilibrium. Thus, the estimation of search models relies on the pillar of contract and incentive models. Indeed, most structural estimation has been conducted within Markov decision processes. These processes are natural for modeling stochastic sequential decision making. Control variables d_t and state variables s_t, $t = 1, \ldots, T$ characterize the decision process. The individual is represented by three primitives: a utility function $u(s_t, d_t)$ denoting preferences at time t, a Markov transition probability $p(s_{t+1}|s_t, d_t)$ denoting the agent's beliefs about future states given current states and decisions, and the discount factor β. The optimal decision rule δ is obtained via dynamic programming. See Gotz and McCall (1980) for an early application.

An error term ε_t is usually imbedded in the discrete decision process. This insures that the econometric model is nondegenerate. The structure of the error term is determined by the specifics of the problem. The error term is a fundamental element of the overall optimization, a remarkable rise in its stature compared to earlier econometric models. It is now considered an unobserved state variable, the usual state variable s_t being decomposed into (x_t, ε_t), where x_t is a subvector observed by the econometrician and ε is observed only by the agent.

Rust (1994) identifies practical and theoretical reasons for the recent interest in structural estimation. Theoretically, 'structural estimation is the most direct way to assess the empirical validity of a specific Markov decision process model.' On a practical plane, 'structural models can generate more accurate predictions of the impact of policy changes than reduced-form models.' This practical reason lies at the heart of the Lucas critique: reduced-form estimates are unreliable forecasts of policy changes because they ignore the impact of policy on the 'deep parameters' (u, p, β). Wolpin (1995) is an excellent survey of this literature by one of its founders.

2.7 *The Effects of Labor Market Policies in Equilibrium Search Models*

Over the past decade there has been an increasing interest in economic models that study the flow of jobs and workers among diverse states. Most of this research relies on equilibrium models and emphasizes the frictions in both search and recruiting. Using a rich set of data on flows in labor markets, this research studies job creation, job closure, unemployment, and labor mobility. Matching and equilibrium are crucial to the theoretical approach. Decisions made by employers on the demand side of the market are highlighted. A comprehensive review of this prodigious literature is Mortensen and Pissarides (1999).

Ljungqvist and Sargent (1995a, 1995b) have devised sophisticated extensions of the simple sequential search model to study equilibrium unemployment behavior in European countries. In Sweden, they recommend a smoothing of the distorted incentives in the labor market by adopting less generous unemployment benefits. In European welfare states, their equilibrium search model finds that:

> when the generosity of unemployment insurance increases, the upward pressure on the unemployment rate eventually dominates the reduction in labor mobility due to high marginal income taxes. The higher unemployment rate is caused by unemployed workers reducing their search intensity and increasing their reservation rates. Higher unemployment pushes taxes upward which in turn increases the unemployment rate further, and so on.

3. *Contracts, Incentives, and Asymmetric Information*

3.1 *Introduction*

The past two decades of the twentieth century have witnessed dramatic changes in economic theory. Several years ago most graduate texts in economic theory were founded on price theory. Today, in the major economic departments, economic theory revolves around noncooperative game theory. Price theory has been relegated to the back seat. Before the ascendance of game theory, the emphasis in economics was on the usefulness of economic models in explaining empirical phenomena. While this emphasis persists, it has been overshadowed by the quest to apply strategic thinking to resolve economic problems.

The key ingredients of this remarkable transformation are: the Nash equilibrium (NE) concept, Harsanyi's characterization of contracts with asymmetric information as noncooperative games, and Selten's explicit consideration of time and the elimination of equilibria associated with noncredible threats and promises.

Nash's equilibrium is a strategy for each player where each strategy is a best response to the strategies of the other players in the n-person noncooperative game. This concept pervades economics. As a simple example, consider the traveler who drives a car first in the US and then in the UK. He drives on the right-hand side of the road in the US as does everyone else. When he visits the UK, his best response to British drivers is to drive on the left side. These Nash equilibria are akin to rational expectation's equilibria.

The NE is not immaculate. Jack Hirshleifer (private communication), sees

> two major problems with NE: (1) Each player's decision is supposed to be a 'best reply' to the opponent's corresponding choice. But the Nash protocol requires *simultaneity*. Thus, neither side *knows* the opponents' strategy to which it is supposed to be replying optimally. Without that knowledge, how can someone make a 'best reply'? ... (2) ... the 'best reply' has to be to the opponent's *strategy*, and not just to his observed *actions or moves*. In all but the very simplest cases, only a tiny fraction of a player's full strategy will ever be visible. ... So in general the player can never know whether his current strategy is or is not a 'best reply' to what the opponent has in mind. ... NE is fine for 'toy worlds' of our textbooks, but I remain skeptical of its general predictive validity.

van Damme and Weibull (1995) observe that 'John Harsanyi showed that games with incomplete information could be remodeled as games with complete but imperfect information, thereby enabling analysis of this important class of games and providing a theoretical foundation for "the economics of information".' Signaling, moral hazard, and adverse selection are prominent members of this class. A game is said to be one of *complete information* if Nature makes the first move and this move is observed by all players. A game has *perfect information* if each *information set*—the set of nodes in the tree such that one is known to be the actual node—is a singleton. See Rasmusen (1989). Harsanyi adopts a Bayesian approach assuming that each player may be of several types—a type specifies the information a player possesses regarding the game. The resulting refinement of Nash's equilibrium is called a Bayes–Nash equilibrium.

Reinhard Selten was the first to refine the Nash equilibrium for analysis of dynamic strategic interactions. Such refinement is necessary since many equilibria entail noncredible threats and do not make economic sense. Selten's formalization of the requirement that only credible threats should be considered, the concept of *subgame perfect equilibrium*, is used widely in the industrial organization literature. It has generated significant insights there and in other fields of economics.

At first it was not clear how the problems of asymmetric information could be formulated as noncooperative games. Indeed, much significant research in signaling and insurance was performed in the 1970s with little reference to game theory. These models generated important economic implications, but were stymied by problems of equilibrium and credibility. This was changed suddenly when fundamental papers by Kreps and Wilson (1982) and others showed that this research could use the deep insights of Harsanyi and Selten to illuminate credibility and equilibrium. The ensuing unification of industrial organization was analogous to the epiphany that occurred in search theory when optimal stopping, dynamic programming, and matching were applied to search problems.

3.2 Principal-agent Models

This section is a brief discussion of three of the most important asymmetric information models: moral hazard, adverse selection, and signaling. All three belong to the class of principal-agent models. In these models, the principal P designs the contract. The agent A either accepts or rejects the contract depending on the expected utility of the contract vis-à-vis the utility from other alternatives. Finally, the agent performs the task specified by the principal. The two parties are opposed in that the revenue for the agent is a cost for the principal and the effort of the agent is costly to him and beneficial to the agent. In determining the optimal contract between principal and agent, this opposition must be resolved. Macho-Stadler and Perez-Castrillo (1997) see this as 'one of the most important objectives of the economics of information.'

There are two parties to a P–A model. One is informed while the other does not know a crucial piece of information. For example, in the insurance P–A model, the insurance company is the principal and the insuree is the agent. It is usually assumed that the agent knows his health status, whereas the principal is uncertain. The P–A model is a bilateral monopoly. Hence, the nature of the bargaining process must be specified. For simplicity, it is assumed that either P or A receives all of the surplus. For example, P does this by stipulating a 'take-it-or-leave-it' contract, which the agent either accepts or rejects. Salanie (1997) notes that bargaining under asymmetric information is very complicated: 'There is presently no consensus among theorists on what equilibrium concept should be used.' Salanie observes that: 'the P–A game is a Stackelberg game in which the leader (who proposes the contract) is called the principal and the follower (the party who just has to accept or reject the contract) is called the agent.'

The actual bargaining in a concrete setting is unlikely to have this 0–1 structure, where the game terminates if either an agreement or disagreement happens. Instead, the bargaining may continue or, more likely, the disgruntled agent will search until he finds an acceptable contract. Such a model would combine the P–A analysis with search theory.

3.3 Three Principal-agent Models with Asymmetric Information

Moral hazard occurs when the agent's actions are not observed by the principal. More specifically, moral hazard is present when the subject's actions are: (a) influenced by the terms of the contract, and (b) in a way that is not fully specified in the contract. These hidden actions are evident in automobile insurance contracts, where the driving behavior of the insuree is not known by the insurance company. The contract usually has a deductible to mitigate the risk to the insurance company.

Adverse selection occurs when the agent is privy to pertinent information before the contract is finalized. This information is hidden from the principal. Adverse selection is evident in health insurance contracts where certain characteristics (health status) of the agent are imperfectly known by the insurance company. An insurance contract is designed for each group with a particular health characteristic (sickly or healthy) so that each member of the group has an incentive to buy it.

Signaling occurs when one of the parties (the agent) to the contract has pertinent information regarding his type. Before entering the contract, the agent can signal the principal that he has this type, i.e., the behavior of the informed party conveys (signals) the type information to the uninformed party.

A key observation is that each of these asymmetric information models can be represented as a noncooperative game. The crucial equilibrium concept is a refinement of the Nash equilibrium. For example, the signaling game has a perfect Bayesian equilibrium. Excellent discussions are given in Gibbons (1992) and Macho-Stadler and Perez-Castrillo (1997).

One of the most important topics in the economics of information is the optimal design of contracts under symmetric and asymmetric information. In the symmetric case, the principal designs the contract so that the expected marginal payoff equals the marginal cost of effort. If the principal is risk neutral, then the optimal contract is one in which the principal accepts all the risk and the agent is, therefore, fully insured, receiving a payoff that is independent of the outcome. In the asymmetric case, the situation is much more complicated. The optimal contract balances the conflict between two opposing goals: efficiency in the allocation of risk between principal and agent and maintaining the incentives of the agent. For lucid presentations of the analysis, see Macho-Stadler and Perez-Castrillo (1997) and Salanie (1997).

4. Future Developments

It seems clear that the merger between search and game theory will continue. The benefits from this include: strategic thinking will become even more apparent in search models and the econometric testing of game theory models should develop into a cottage industry alongside experimental testing. Search models will continue to be used to evaluate labor policies like cross-country unemployment and other macro issues. Their use in other economic areas should also flourish.

In his evaluation of the applications of Nash equilibrium, Selten claims that we still know little about actual behavior and that empirical research and theories of bounded rationality are needed. Rubinstein (1995) is more emphatic. 'Economic theory should deal with the real world. It is not a branch of abstract mathematics even though it utilizes mathematical tools. Since it is about the real world, people expect the theory to prove useful in achieving practical goals. But economic theory has not delivered the goods.' Many would dispute the final sentence, but most would agree that game-theoretic models require extensive econometric testing. This has already begun, with structural estimation being the favored method, but much remains. This is an exciting time for the economics of information.

Bibliography

Alchian A A 1951 Uncertainty, evolution, and economic theory. *Journal of Political Economy* **58**: 211–21

Burdett K, Mortensen D T 1998 Wage differentials, employer size, and unemployment. *International Economic Review* **39**: 257–73

Burdett K, Wright R 1998 Two-sided search with nontransferable utility. *Review of Economic Dynamics* **1**: 220–45

Cannan E 1930 Review of: The post-war employment problem by H. Clay. *Economic Journal* **40**: 45–53

Cayley A 1875 Mathematical questions and their solutions. *Educational Times* **22**: 18–19

Diamond P A 1971 A model of price adjustment. *Journal of Economic Theory* **3**: 156–68

Ferguson T S 1989 Who solved the secretary problem? *Statistical Science* **4**: 282–96

Gibbons R 1992 *Game Theory for Applied Economists*. Princeton University Press, Princeton, NJ

Gotz G A, McCall J J 1980 Estimation in sequential decision making models. *Economic Letters* **6**: 131–6

Hayek F 1978 *New Studies*. University of Chicago Press, Chicago, p. 269

Hicks J R 1932 *The Theory of Wages*. MacMillan, London

Hutt W H 1977 *The Theory of Idle Resources*. Liberty Press, Indianapolis, IN

Knight F H 1921 *Uncertainty and Profits*. Hart, Schaffner and Marx, New York

Kreps D 1990 *Game Theory and Economic Modeling*. Oxford University Press, Oxford, UK

Kreps D M, Wilson R 1982 Sequential equilibria. *Econometrica* **50**: 863–94

Lindley D V 1961 Dynamic programming and decision theory. *Applied Statistics* **10**: 39–51

Lippman S A, McCall J J 1976 The economics of job search: A survey. *Economic Inquiry* **14**: 155–89

Lippman S A, McCall J J 1993 Search and the development of the economics of information. *Estudios de Economia* **20**: 223–49

Ljungqvist L, Sargent T J 1995a The Swedish unemployment experience. *European Economic Review* **39**: 1043–70

Ljungqvist L, Sargent T J 1995b Welfare states and unemployment. *Economic Theory* **6**: 143–60

Macho-Stadler I, Perez-Castrillo D 1997 *An Introduction to the Economics of Information*. Oxford University Press, Oxford, UK

MacQueen H, Miller R G 1960 Optimal persistence policies. *Operations Research* **8**: 362–80

McCall J J 1970 Economics of information and job search. *Quarterly Journal of Economics* **84**: 113–26

McMillan J, Rothschild M 1993 Search. In: Aumann R, Hart S (eds.) *Handbook of Game Theory with Economic Applications*. North-Holland, Amsterdam

Mortensen D T, Pissarides C A 1999 New developments in models of search in the labor market. In: Ashenfelter O, Layand R (eds.) *Handbook of Labor Economics*. North-Holland, Amsterdam, Vol. 3
Myerson R B 1999 Nash equilibrium and the history of economic theory. *Journal of Economic Literature* **37**: 1067–82
Rasmusen E 1989 *Games and Information*. Basil Blackwell, New York
Rothschild M 1973 Models of market organization with imperfect information. *Journal of Political Economy* **81**: 1283–-308
Rubinstein A 1995 John Nash: The master of economic modeling. *Scandinavian Journal of Economics* **97**: 9–13
Rust J 1994 Structural estimation of Markov decision processes. In: Engel R F, McFadden D L (eds.) *Handbook of Econometrics*. North-Holland, Amsterdam, Vol. 4
Salanie B 1997 *The Economics of Contracts: A Primer*. MIT Press, Cambridge, MA
Sargent T J 1999 *The Conquest of American Inflation*. Princeton University Press, Princeton, NJ
Schumpeter J A 1942 *Capitalism, Socialism and Democracy*. Harper and Row, New York
Schwarz G 1994 Game theory and statistics. In: Aumann R J, Hart S (eds.) *Handbook of Game Theory*. Elsevier Science B.V., Amsterdam, Vol. 2
Simon H A 1951 A formal theory of the employment relationship. *Econometrica* **19**: 293–305
Stigler G J 1961 The economics of information. *Journal of Political Economy* **69**: 213–25
van Damme E, Weibull J W 1995 Equilibrium in strategic interaction: The contributions of John C. Harsanyi, John F. Nash and Reinhard Selten. *Scandinavian Journal of Economics* **97**: 15–40
Wolpin K I 1995 *Empirical Methods for the Study of Labor Force Dynamics*. Harwood Academic Publishers

S. A. Lippman and J. J. McCall

Infrastructure Investment

The city building process is a product of infrastructure investment. At the most general level, the term infrastructure describes the underlying foundation of the physical and social facilities of the city (Seeley 1993, p. 20, Konvitz 1985, p. 130), both the physical systems and technologies, and the health, education, and welfare facilities that knit the city into a collective fabric. From this perspective, Herbert Muschamp (Perry 1995, p. 1) observes that, 'infrastructure is shorthand for the structural underpinnings of the public realm.'

While such an encompassing approach to infrastructure has conceptual resonance, most studies carve a more particular focus (OECD 1991, Neutze 1997, Graham 2000), defining infrastructure as the underlying physical foundation of the city: the material systems, networks, and technologies (roads, water and sewer systems, power grids, and telematic networks) that serve key urban relations; the social flows, economic exchanges and political relations of the city. This foundational primacy of physical infrastructure affords it a rather unique status in the literature: namely, the bridges, roads, and water systems of the city are not described simply as important physical artifacts but as investments in the city building process. The returns on such investments are measured in terms of a society's economy, health, and social well-being.

Hence infrastructure scholarship and policy often begin with the normative position that the more an urban population is able or willing to invest in physical infrastructure, the more apt the urban settlement is to prosper economically and benefit socially from the resulting conditions of efficiency, health, and human development. This is perceived to be as true in new or emerging urban centers as it is in the established cities of the developed world, thereby making issues of definition, public or private provision of such systems, finance, and technology equally important to every modern city

1. Infrastructure: Evolution from Private to Public Function

The term 'infrastructure' as it is employed here, is relatively new: a conception of the late twentieth century that is used to describe the systemic applications of technology to the needs of urban society, producing what historian Joel Tarr (1984, p. 4) calls the 'technological sinews' of the city: its transportation networks, power grids, utility systems, and telecommunicative linkages. For much of the nineteenth and twentieth centuries, urban physical facilities were not viewed systemically as 'physical 'networked' services' (OECD 1991, p. 19), but as either privately produced, geographically individuated 'internal improvements' or equally decentralized, governmentally generated 'public works.'

For, example, the seventeenth-century invention of the British toll road, the late eighteenth-century canal systems, and the early railways were initially proprietary responses to the requirements of the Industrial Revolution. Technological advances in water treatment, electricity, gas distribution, and modes of communication were virtually all developed by private inventors and investors. Only later, as the impact of such inventions on the dramatic and uneven shifts in urban populations, levels of disease, congestion, and fire became apparent, did they come to be viewed as a public responsibilities or goods. Even then, cities' 'different…business leaders, politicians and professionals' (Tarr 1984, p. 5) would take years to determine which technology or proprietary service was the right one and more years to decide how to plan, pay for, and ultimately build the service (Konvitz 1985, Anderson 1988). For example, the benefits of water and sewage treatment to health and economic development were

well known by the middle nineteenth century, but it took US cities until 1940 to build such systems. World wide, hundreds of millions of urbanites live in political-economic environments that still do not offer reliable water and healthy sewage systems, transportation systems, or power

As such, the history of successful infrastructure investment is less about the direct impact of technology and more about the interaction between technology and society (Tarr 1984). Put another way, recent scholars posit that infrastructure networks are both social and technical: they are always contested and their impacts on cities are never neutral (Graham 2000). No matter how 'public' the good is perceived to be, the impacts of telecommunicative grids, transportation roads, and energy will always privilege some places and people over others (Swyngedouw 1993).

1.1 Public Infrastructure Investment

However convoluted its investment process and uneven its impacts, infrastructure investment became essentially public practice in the twentieth century, carried out either by governments or heavily regulated private firms. Several reasons are offered for this (Neutze 1997, Perry 1995). First urban infrastructure often serves the essential health and well-being of the citizenry, requiring full citizen access to such services at reasonable cost. Second, this implies the provision of goods at a scale and price (given the elasticity of demand of several of these services) that is beyond the capacity or interest of private providers. Third, urban infrastructure goods often constitute natural monopolies, whereby it is less costly to have a single provider than several service delivery systems. For example, it would be very costly to have competing water and sewer companies putting parallel networks of mains and pipes under every street in a city to give consumers a choice of systems. Fourth, some forms of urban infrastructure produce almost exclusively public goods or externalities (or services which can not be used by one consumer without also being made available to all), such as public roads, sidewalks, and parks. Fifth, equity can require that certain key benefits of urban infrastructure should not be withheld from any citizen. (Neutze 1997, pp. 29–31)

1.2 Special Purpose Governments

Just as the conditions of scale, cost, and equity have led to public provision of physical infrastructure, the sheer size and the accompanying obligations of debt and delivery of infrastructure systems has often led one level of government (especially national governments) to offload many of these services on to lower levels of government. Municipal governments in turn have further devolved the delivery of infrastructure, one by one, to special purpose governments or public authorities (Sbragia 1996, Perry 1995). In the United States, for example, such governments issue the majority of all bonds used to finance infrastructure investment. It is argued that such decentralized structures work well to:

(a) preserve and circumvent local and state debt limits, maintaining the legal requirements of general municipal fiscal balance while allowing special purpose, extra-municipal public agencies to generate debt, raise revenue, and build physical infrastructure networks; and

(b) create more direct relationships between the government and the capital markets, thereby better satisfying market requirements for enhanced credit risk management (through municipal debt guarantees, or bond insurance, or policies of securitization, such as the provision of bank letters to offset credit exposure, etc.) in order to pay for infrastructure.

Critics of such public corporations point to their privatistic practices as evidence of the increasing transformation of government into 'business' (Walsh 1978). Adherents suggest that the privatization of public infrastructure services in general and through special purpose governments and public authorities in particular is part of the century-long move away from the inefficiencies of the state and back to the private sector as the best, most stable source of infrastructure investment (Smith 1999).

2. Infrastructure Investment as Public/Private Practice

In many ways, the dominant normative consideration of public infrastructure investment has come full circle, with a twist. Where conditions of scale and cost, and developmental importance made infrastructure investment essentially a public or governmental function at the beginning of the twentieth century, the very same issues are challenging the capacities of the state and forcing new mixes of public and private service delivery at the beginning of the twenty-first century. The government, at national and local levels, is finding it difficult to fully conceive, finance, build, or efficiently maintain modern systems of urban infrastructure. This is held to be equally true for basic infrastructure systems in the emerging urban centers of the developing world (Smith 1999) and for advanced telematic systems of telecommunication and digital media that are radically transforming the space of urban flows in general (Graham and Marvin 1996).

2.1 Emerging Cites in an Urban World

Nowhere are these arguments more forcefully made than in the literature on infrastructure development in emerging cities. Increasing patterns of economic globalization and migration into cities are central features of global geographic and social development.

Sometime in the last decade of the twentieth century, the world became 'urban,' more than 50 percent of the global population could be found living in towns and cities. The emerging cities of this new urban world face tremendous infrastructural challenges with limited technological and fiscal resources, and yet they represent stable and viable places to ever more people (Clark 1996).

Just as issues of health, transportation, and scale challenged the United States and Europe at the beginning of the twentieth century, similar patterns, at an even greater scale, challenge the emerging urban centers of Latin American, Asia, and Africa in this century. Millions continue to move to the cities without access to basic infrastructure facilities such as safe water, sanitation, and garbage disposal.

2.2 *Financial Practices: A Public-Private Mix*

While national governments remain the single most important source of infrastructure delivery in the emerging urban world, the costs, which can be as high as 10 percent of GNP, suggest that they cannot do it alone. They must attract domestic and international equity capital for private investments in the infrastructure. This requires the move to Western-style financing schemes and patterns of governmental devolution of infrastructure investment, making it a more 'public-private' activity worldwide.

In the past, municipalities (especially in developing countries) were able to combine their own taxes with central government transfers to build public works, but now infrastructure demands make them just as fiscally burdened as their national governments. In addition, they are required by both the national government and international banks to 'forge partnerships with financiers, private operators and constituents' (El Daher 1997, p. 1). The sources of revenue available to governments to pay for such services generally fall into three categories: (a) taxes, (b) user fees, and (c) other forms of leases (and purchase agreements) and revenue schemes. Depending on their success in using these sources of revenue, cities' infrastructure investments require the organization and constitution of government to accommodate new debt structures, create relationships with the capital markets, and regulate public-private infrastructure agreements. In certain key areas, infrastructure investment is becoming even more privatized or transferred to the private sector through deregulation, concessions, contracting out, and private management and control agreements.

2.3 *Access to Private Investment Markets*

Considering all of this, a most important signifier of infrastructure investment in cities is their access to the credit markets, that is providing the institutional transparency, strong government regulatory power, debt guarantees, and revenues needed to secure an attractive private investment climate in which to develop infrastructure. While these are common issues of municipal infrastructure finance everywhere in the urban world, in no places are they any more evident than in developing countries where not only can the fiscal strength of the municipality be open to question, but also the very existence of a stable viable domestic credit market can be equally problematic. Cities in developing countries often have more success generating direct private investment in fully privatized public works than in creating the stability of process required of a local municipal bond market.

In a rapidly urbanizing world and uneven investment climate, nongovernmental intermediaries like the World Bank, the IMF, the Inter-American Development Bank, and the Asian Development Bank have become important sources of infrastructure investment policy. They attempt to provide debt guarantee instruments and policy strictures designed to influence the creation of stable domestic capital entities, with the larger purposes of not only institutionalizing long-term markets but also enhancing governmental legitimacy across regimes.

3. *Infrastructure Investment and Urban Development*

Infrastructure investment is an important way to view the urban. It represents networks of technology and urban life, binding them together in patterns that are clearly important elements of cities and their development. It comprises a substantial share of the sunk physical capital of cities; materially delivering the networks of heat, light, water, telematics, and power that are central to urban life; without them modern cities do not exist (Graham 2000, p. 115). The material grids of such components of the urban landscape and their accessibility go a long way toward explaining what is meant by the 'developed' and 'developing' urban world.

The study of infrastructure investment provides an understanding of the state as a site of debt and public regulation and the market as a site of capital and risk-generated requirement. The restructuring of the market into risk niches and intermediaries and the state into devolving agencies of delivery and transfer in both developed and developing urban centers are central characteristics of infrastructure investment as a public-private practice of the new urban world. More generally the full implementation of infrastructure investment is a key element of the dominant practices of urban political economy: the structuring of government, the creation and regulation of capital markets, the conditions of uneven urban economic

development, and the socio-technical networks of service and underservice that represent the distributions of power in the city and globally (Tarr 1984, Swyngedouw 1993). In sum, the study of infrastructure investment is the study of urban development.

See also: Communication: Geographic Aspects; Public Goods: International; Telecommunications: Policy; Telegraph; Telephone; Transportation Planning; Transportation: Supply and Congestion; Urban Geography

Bibliography

Anderson L 1988 Fire and Disease: The development of water supply systems in New England, 1970–1990. In: Tarr J A, Dupuy G (eds.) *Technology and the Rise of the Networked City in Europe and America*. Temple University Press, Philadelphia, PA

Clark D 1996 *Urban World/Global City*. Routledge, London

El Daher S 1997 *Municipal Bond Markets: Prospects for Developing Countries*. Infrastructure Notes, The World Bank, Urban No. FM-8b, Washington, DC

Graham S 2000 Introduction: Cities and infrastructure networks. *International Journal of Urban and Regional Research* **24**(1): 114–19

Graham S, Marvin S 1996 *Telecommunications and the City: Electronic Spaces, Urban Places*. Routledge, London

Konvitz J 1985 *The Urban Millennium: The City-building Process from the Early Middle Ages to the Present*. Southern Illinois University Press, Carbondale, IL

Neutze M 1997 *Funding Urban Services: Options for physical Infrastructure*. Allen & Unwin, St Leonards, NSW, Australia

Organization for Economic Co-operation and Development (OECD) 1991 *Urban Infrastructure: Finance and Management*. OECD, Paris

Perry D 1995 Building the public city: An introduction. In: Perry D C (ed.) *Building the Public City: The Politics, Governance and Finance of Public Infrastructure*. Sage, Thousand Oaks, CA, pp. 1–20

Sbragia A 1996 *Debt Wish: Entrepreneurial Cities, U.S. Federalism, and Economic Development*. University of Pittsburgh, Press, Pittsburgh, PA

Seeley B 1993 The Saga of American Infrastructure. *Wilson Quarterly* Winter

Smith A J 1999 *Privatized Infrastructure: The Role of Government*. Thomas Telford, London

Swyngedouw E 1993 Communication, mobility, and the struggle for power over space. In: Giammopoulos G, Gillespie A (eds.) *Transport and Communications in the New Europe*. Belhaven, London

Tarr J A 1984 The evolution of the urban infrastructure in the nineteenth and twentieth centuries. In: Hanson R (ed.) *Perspectives on Urban Infrastructure*. National Academy Press, Washington, DC, pp. 4–66

Walsh A H 1978 *The Public's Business: The Politics and Practices of Government Corporations*. MIT Press, Cambridge, MA

D. C. Perry

Infrastructure: Social/Behavioral Research (Japan and Korea)

1. Phases of Development After 1945

Modern social science originated outside Asia (Inoguchi 1995). In Japan and Korea, which was under Japanese colonial rule for much of the first half of the twentieth century, modern social science meant, initially, European-born Marxism planted onto Asian soil. Hence, the social sciences meant the opposition science, *Oppositionswisssenschaft*, the science those opposing the regime adopted to reveal the oppressive and exploitative nature of the system. Marxist influence permeated the social sciences deeply. Coupled with the overtly nationalistic overtones characteristic of Asia, Marxism was not necessarily conducive to the early adoption and diffusion of behavioral and social sciences that characterized the American scene of the 1950s and 1960s. In other words, the dominance of Marxism in the social sciences meant higher barriers against the adoption of American-style behavioral and social sciences even in the 1950s and 1960s.

Furthermore, the earlier dominance of state science, *Staatslehre*, the science that purported to be useful to the state's governance, had a lingering influence on the fledgling social sciences in the first half of the twentieth century and beyond. Instead of trying to come up with law-like generalizations, state science was interested in supplying specific, concrete, and context-based knowledge. It encouraged historical and institutional description. State science was interested in applied subjects like law and economics, but not political science, sociology, or social psychology. State science was interested in history and geography, as far as they could be utilized in running the country and the colonies. Therefore, in Japan and Korea in the 1940s and 1950s, the traditions of the opposition science and state science, two most overtly politically conceived traditions, slowly adopted the American-style social sciences of the 1940s and 1950s in a mutually reinforcing direction.

Nevertheless, American-style social sciences started to permeate steadily into the rest of the world after 1945. The remarkable development of the American model of the professional academic market in the first half of the twentieth century prepared the United States to excel itself in the social sciences in the latter half of the century. As carefully portrayed by Zunz (1998), it is characterized by competition, based on a certain set of criteria within the academic community, and the loosely orchestrated concertation of government, industry, and the mass media to nurture and make use of academic resources and products. Furthermore, the influx of many European refugees stimulated the American scene dramatically. The number of Nobel Prize winners from the United States was modest each year between 1901 and 1937, some-

what comparable to the figure of Japan after 1945. Thereafter, it rose dramatically, achieving a dominant share (Inoguchi 1989).

Three major features of American-style behavioral and social sciences are (a) a strong drive to come up with law-like generalizations; (b) a strong commitment to systematic, empirical hypothesis testing; and (c) shared beliefs in anonymous referees producing good judgments and improved manuscripts. These features were not necessarily in good harmony with the prevailing academic cultures of Japan and Korea in the 1940s and 1950s. Their interests lay in more historical and contextual descriptions, not necessarily in law-like generalizations. Their efforts to advance the argument were not necessarily made in a sustained fashion of systematic empirical hypothesis testing. They were extremely averse to the idea of their work being anonymously refereed for publications. Thus in the 1940s and 1950s, American-style behavioral and social sciences were not able to make a striking advance in Japan and Korea. Only in the 1960s and 1970s, were they to ride on the tide of behavioral and social sciences as understood and practiced in the United States. The 1960s and 1970s were also the period of the extraordinary economic growth of Japan and Korea. In tandem with economic development, the academic infrastructure was consolidated, and academic orientations came of age in their national environments.

2. *National Environments and Institutional Structures*

The higher educational legacy of imperial Japan, which ended in 1945, contributed most directly to the shape of the academic infrastructure of the behavioral and social sciences in Japan and Korea after 1945. Its major features were:

(a) an elitist structure of imperial universities with the number of professors and enrolled students being highly limited;

(b) a curriculum serving the needs of the government in terms of supplying the elite cadres in applied areas of study such as medicine, agriculture, engineering, law, and economics; and

(c) two European languages, German and French, equally emphasized with English, for students to learn.

After 1945, the number of colleges and universities grew rapidly in US-occupied Japan, where each of the 47 prefectures built a national university. Furthermore, private universities grew in number more dramatically in the 1960s and 1970s. The elitist nature of Japanese universities was played down considerably throughout those years. Rather, Japan, and to a lesser extent Korea, registered a high admission rate of high-school graduates comparable to that of the United States, achieved through a very inclusive enrolment scheme. The nature of the curriculum has also significantly changed in tandem with the extraordinary growth of student enrolment and staff appointment from the 1960s to the 1990s. In the case of the University of Tokyo, for instance, the relative priority of staff appointments remained the same. It expanded in applied science areas, especially engineering, the life sciences, medicine, pharmacology, and agriculture. As of 2000, 80 percent of the annual budgetary allocation goes to those applied science areas. Only 20 percent of it goes to law, economics, literature, education, and other nonscience areas of study (University of Tokyo 2000). However, in most private universities, those applied science areas of study have not been well prioritized in terms of faculty development and student enrollment, thanks to the general paucity of budgetary resources as a whole.

However, this does not mean that a favorable national environment has been created for the development of the behavioral and social sciences. Within the non-applied science areas of study, roughly the traditionally conceived arts and sciences, autonomous departmental status has not been given to the behavioral and social scientific disciplines. The political science discipline has been an appendix to schools of law (a legacy of state science); the school of economics has a curriculum giving very high importance to applied economics and history; and the school of literature has departments of sociology, social psychology, and anthropology without their autonomous status in terms of staff appointment and budgetary allocation. Rather, they remain parts of the schools of law, economics, and literature and sometimes serve as their mere appendices.

In this regard, Korea, unlike Japan, thanks to its nationalism and much stronger American influences in higher education, has developed its institutional structure far ahead of the Japanese imperial legacy. The emphasis on European languages was replaced by the predominant use of English after 1945. This has facilitated the diffusion of American-style behavioral and social sciences.

Turning to the topic of institutional and financial setups, there are three major ways in which behavioral and social science research is conducted:

(a) within the university system;

(b) by extrauniversity institutes; and

(c) the system of project funding through national organizations and foundations.

Behavioral and social science research is predominantly conducted within the university system. Small-scale research is normally conducted by making best use of a small amount of research budget made available to each professor. At some universities, schemes for facilitating larger and more interdisciplinary research exist. However, the size of budget tends to be small. Therefore, once research needs get large in terms of funding needs, one tends to rely on other schemes (the Japan Society for the Promotion of Sciences 2000).

In addition to the university system, there are some research institutes primarily devoted to the planning, implementation, and dissemination of research on national and sometimes international bases. They are either publicly funded or run by the mix of private and public money. In Japan, for example, the National Museum for Ethnographic Research in Osaka, the largest such institute run by public money, focuses on anthropology. The Institute organizes various projects incorporating professors working in universities. The Institute of Statistics and Mathematics in Tokyo, also run by public money, is a similar organization that focuses on public opinion research on national characters and life styles. In Korea, a similar institute, the Sejong Institute, is publicly and privately funded and works on a variety of subjects in sociology, political science, and urban studies.

Some institutes work on more direct, public policy related research topics like 'public policy priorities in aging societies,' while others work on more purely academic subjects like 'life styles, postmodern, post-Confucian, and post-Marxist.' They vary enormously (NIRA 2000, Yamamoto, 2000).

The system of project funding through national organizations and foundations is important in Japan and Korea. In Japan, the Japan Society for the Promotion of Sciences plays roles similar to those of the National Science Foundation of the USA, giving financial support to various academic projects (the Japan Society for the Promotion of Sciences 2000). The National Institute for Research Advancement in Tokyo is an organization funding more public policy focused projects run by various think tanks in Japan and elsewhere (NIRA 2000, Yamamoto, 2000). The Japan Foundation in Tokyo is a foundation run by public money to assist research financially and to disseminate academic, artistic, linguistic, and other professional projects globally (the Japan Foundation 2000). In Korea, too, similar organizations and foundations exist. Many private foundations exist in both countries in addition to these public organizations and foundations. Some focus on gender studies, academic and professional exchanges, human security, and international financial needs, while others give preference to more public policy related projects.

In terms of infrastructure development for the behavioral and social sciences, the university system is central and most helpful after all. Most project focused funding does not support infrastructure development. In order to develop an academic infrastructure, such as building joint centers for survey research or for experimental laboratories, interuniversity consortiums for databases, electronic networks, and publishing houses, one needs to rely mostly on public money.

A joint university center for survey research does not exist in Japan. What exists is the three to five year coalition based on the project funding of scientific research. With it, you build a research team and hire an opinion polling company for the survey. With the ending of the project, everything except the database and published volumes disappears. My own scientific research project is a recent example of this. It is funded by the Ministry of Education over four years (1999–2002). It deals with 'globalization and the cultural dimensions of democracy in 18 societies of Asia and Europe' and has 1,000 in each sample, done by face-to-face interviewing (Blondel and Inoguchi 2000). A number of universities such as the University of Tokyo and Keio University in Tokyo have departments of social psychology and sociology, and research centers/institutes for research by mass communication. However, none of them undertakes such coalition building and coordination roles institutionally.

It is the same with respect to experimental laboratories. Again, they must be built with project funding for scientific research and can be maintained and expanded if, luckily, they continue for several years. One recent example is Toshio Yamagishi's experimental laboratory built to examine modes of trust experimentally and cross-nationally (Yamagishi 2000). With regard to social scientific databases, the University of Tokyo's Institute of Social Sciences (http://www.iss.u-tokyo.ac.jp/) has set up an inter-university consortium for databases in a similar fashion as the University of Michigan's Inter-university Consortium for Social and Political Research. Focusing on sociological, political, historical, and economic Japanese data, the consortium has been quite a success, albeit on a rather small scale, in organizing available databases and making them academically totally open and accessible to its members.

In terms of election networks, again project funding of scientific research by the Japan Society of Promotion of Sciences and the Ministry of Education is vital. Ikuo Kabashima has done so on seven Japanese election panel surveys through the 1990s, producing six published volumes dealing with various aspects of Japanese electoral behavior (Kabashima 1998–2000). Kabashima's database has largely expanded the data bank of the journal called *Leviathan*: *the Japanese Journal of Political Science*. The data bank has been more narrowly focused on election data and has run on a much smaller scale.

Publishing houses dealing with behavioral and social science books do not necessarily thrive in business. Given the fact that academic publishers cannot anticipate the regular purchase of academic books by major university libraries in the order of 300 to 500, the publishing infrastructure needs to be further consolidated. The publishing infrastructure for English language volumes has not been built in any meaningful way. Kodansha International is the only commercially self-sustainable publisher in the English language. It is not primarily oriented to academic books. The United Nations University Press in Tokyo has been quite a success in the late 1990s in continuously publishing good academic books on such topics as democracy, humanitarian interventions, and

the global environment, albeit on a modest scale. It has been distributed in North America by the Brookings Institution Press and has been included in the Columbia International Affairs Online (CIAO) Internet service on new publications. University presses such as the University of Tokyo Press and Kyoto University Press have English-language publishing departments. However, they only publish sporadically and their published books are mostly based on translations of books originally written in Japanese.

3. *Strengthening the Asian Research Base*

Social science research in Asia in general, and in Japan and Korea in particular, in the second half of the twentieth century, has become well institutionalized in its respective national settings and organizational structures. Japan and Korea, having diluted the two strong legacies of state science and opposition science that were prevalent in the first half of the twentieth century, and brought in a new set of American-style behavioral and social sciences, developed a fairly solid foundation for social science research in the second half of that century. Although its infrastructure remains to be much more vigorously improved in the future, the stage is clearly set for its further consolidation, especially in the light of the two pronounced trends of international cooperation and cross-national comparisons.

In Japan and Korea, as well as in Asia in general, international cooperation started with the United States. The United States has been a senior partner for Asia in general and Japan and Korea in particular for the whole time. Through bilateral research and teaching with the United States, by 2000, both latter countries had enhanced their national foundations of social science research. Toward the end of the twentieth century, it was clearly discerned that the trend in international cooperation had come to encompass the United States as well as Asia. Politics and economics matter here. In much of the latter half of the twentieth century, regional cooperation did not flourish much in part because of economics and politics. However, in the last decade of the twentieth century, Asian countries (here meaning the Asia-Pacific countries: Japan, Korea, China, and the 10 Association of Southeast Asian Nations members) came to interact among themselves.

It is not surprising that the new trend of enhanced intraregional research cooperation is in broad harmony with the increase in intraregional trade and the frequency of intraregional summit meetings. To a considerable extent, it looks as if the then newly found regional (i.e., Asian) identity pushed such a trend. However, this trend was primarily due to the very strong interest in cross-national comparisons. Once national achievements have been completed more or less in terms of per capita national income and high levels of literacy, a sort of looking around and probing the cross-national propositions in social science research has become a norm, replacing the traditional approach by focusing on national patterns and developments. It is still far from claiming the beginning of an Asian social science.

Yet, a number of such seeds have already been sown and the prospect is not so bad. Let me give a couple of examples. Demography and the sociology of social mobility have been one of those areas of research where cross-nationally standardized observations are relatively institutionalized. The Japanese Sociological Association's panel study of Japanese social mobility, which started in the 1960s, has been further enhanced by incorporating comparative measurement and analysis in the 1990s (Seiyama and Hara 1999). The societies chosen for comparison are industrialized democracies of the West and fledgling democracies of the Asia-Pacific region. Another example is a new project comparing values, norms, and life styles of nine Asia-Pacific societies. It plans to utilize the Gallup Millennium Survey done in January 2000 in 82 countries (Tomiie and Mano 2000, Emiko and Satoko 2000). The co-directors, Ahn Chung-Si of Seoul National University and Takashi Inoguchi, have been running this project.

Up to the late 1990s, the leadership for social science research cooperation in Asia-Pacific used to come primarily from the United States. Increasingly, since then, the trend is for intraregional cooperation by local initiatives, whether from Korea, Japan, Singapore, or Australia. Since there has not been an all-inclusive regional social science research council taking conceptual leadership and funding facilitation like the United States Social Science Research Council has long done, one cannot claim too much at this stage. However, one can easily discern budding initiatives and publications on such subjects as national identity, globalization, democratization, human rights, and civil society. Universities such as the University of Tokyo, Yonsei University, Chularonkorn University, the National University of Singapore, and the University of Malaya as well as think tanks like the Institute of Southeast Asian Studies (Singapore), the Japan Center for International Exchange (Tokyo), and the United Nations University (Tokyo) are major agents for more academically oriented undertakings. With the resultant, cumulatively enhanced networks within the Asia-Pacific, one can hope with some cautious optimism that one can start discussing some budding features of social science in the region. These would surely synthesize a diversity and openness of views and angles that would come up with empirically testable comparative propositions on major trends and patterns of democratization, digitalization, and globalization, and their differentiated impacts on local societies.

See also: China: Sociocultural Aspects; Japan: Sociocultural Aspects; Korea: Sociocultural Aspects; Science and the State; Science Funding: Asia

Bibliography

Blondel J, Inoguchi T 2000 Globalization and the cultural dimensions of democracy in 18 societies in Asia and Europe. *Paper prepared for presentation at the World Congress of the International Political Science Association*. IPSA 2000 Program, University of Quebec, Quebec, Canada

Emiko T, Satoko M 2000 *Young Women Looking for Greater Rights*. Nippon Research Center, Tokyo

Inoguchi T 1989 Four Japanese scenarios of the future. *International Affairs* **65**(1): 15–27

Inoguchi T 1995 Democracy and the development of political science in Japan. In: Easton D, Gunnell J, Stein M (eds.) *Regime and Discipline: Democracy and the Development of Political Science*. University of Michigan Press, Ann Arbor, MI

Kabashima I (ed.) 1998–2000 *Studies on Japanese Electoral Behavior*. Bokutakusha, Tokyo, 6 Vols.

National Institute for Research Advancement 2000 *Think Tank Yoran* (*The overview of think tanks*). NIRA, Tokyo, http://www.nira.go.jp

Seiyama K, Hara J 1999 *Shakai ryudosei* (*Social Mobility*). University of Tokyo Press, Tokyo

The Japan Foundation 2000 *The Catalogue 2000*. JF, Tokyo

The Japan Society for the Promotion of Sciences 2000 *The Catalogue* JF, Tokyo

The University of Tokyo 2000 *The University Catalogue*. UT, Tokyo

Tomiie E, Mano S 2000 *Kojinshugi no kakudai suru Nihon* (*Individualism permeating in Japan*). Nippon Research Center, Tokyo

Yamagishi T 2000 *Shinrai no kozo* (*The structure of trust*). Chuo koron shinsha, Tokyo

Yamamoto T (ed.) 2000 *Imagining a Civil Society in the Asia-Pacific Community*. Japan Center for International Exchange, Tokyo, and the Institute for Southeast Asian Studies, Singapore

Zunz O 1998 *Why the American Century*? University of Chicago Press, Chicago

T. Inoguchi

Infrastructure: Social/Behavioral Research (Latin America)

The social sciences in Latin America present an array of manifestations, differing widely across disciplines and types of institution. There is also a great deal of variation in the nature of graduate and postgraduate training, research experience, and professional employment. At a time when significant developments take place in the international agenda of research methodology and interdisciplinary approaches, the regional intellectual map reveals great dynamism at the same time that it unveils the presence of persistent vulnerability and stagnation.

1. Graduate Training

The Latin American tradition of tertiary studies opens the door to regulated professions after five years of study and, despite the fact that courses with nonprofessional titles have arisen in the various social fields, the association with university education historically has been quite clear. The social, legal, communication, and behavioral sciences are clearly in the majority (29.2 percent of enrolments). If we include in them economics and administration, their proportion rises to 41.3 percent. The group that results from adding together educational disciplines and the humanities represents 18 percent of the total enrolment which, added to the above-mentioned disciplines, gives a figure of 60 percent of higher-education student population devoted to sociocultural and educational disciplines.

The range of courses in the social field is not homogeneous either between countries in the region or within one and the same country. Within a motley and variegated picture of the social science field, the main systems producing the greatest numbers of graduates with first degrees are those of Mexico, Brazil, Argentina, Peru, Colombia, and Venezuela. The social and human sciences take the lion's share of enrolment in all higher-education systems, including those of Latin America, for the simple reason that professional activities requiring language proficiency and general knowledge about society and the contemporary world are much more numerous than those requiring specialized and technical knowledge. Adding all the disciplines related to the social field they total half a million graduates, or 62 percent of the total graduate population from higher education.

Two main groups of social science courses may be distinguished. The first one is oriented to the employment market (basically for people who already work or who seek a secondary-school position and/or general training). The second, much smaller, comprises the traditional more academically oriented disciplines.

1.1 Market-oriented Courses

Expansion of higher education, especially from the 1970s on, swelled the number of students in the social sciences and the humanities, who were recruited largely from those who had been unsuccessful in applying for more prestigious courses. Such students seek an ill-defined professionalization that the university is unable to supply in satisfactory measure, since it requires much more structured supervision than the university is prepared to offer. The majority

of students enroll in courses with high drop-out rates and conferring dubious professional status and are from social backgrounds manifestly less privileged than those seeking to enter more competitive social professions. Thousands of students enter higher education inadequately educated and enroll in private institutions which pay the teachers badly or do not provide the minimum working conditions needed, or in the 'difficult' mass departments of public universities. The climate of lack of motivation prevailing in such departments and schools explains why many teachers who wish to do research seek refuge in postgraduate work.

Curricula for the market-oriented courses—administration, social service, law, teaching, communication, and others—have often been weak at both undergraduate and postgraduate level because, being based on extremely pragmatic educational schemes, they have virtually no academic or disciplinary traditions of their own. They are like hybrids unable to reproduce, despite the efforts they often make to copy the academic rituals of the more established disciplines (congresses, specialist journals, research projects, postgraduate programs, etc.). Professional training in the social sciences is a difficult and controversial issue. New social demands imply changes in the definition of the problems to be considered. One way to proceed might be to include the subjects of applied courses in the central work agenda of the most highly qualified social scientists, thus expanding their range of subjects and their involvement in topics normally considered to lie outside their specialty. This is not an exclusively Latin American problem, though. The available evidence shows that two very different types of teacher co-exist in the most successful market-oriented courses: persons from the world of production, business, or services proper who, as part of consultancy activities, move between the academic sector and those other sectors and who convey to the students the practical and applied meaning of their occupations, and academic social scientists responsible for research and for ensuring the academic and intellectual quality of the courses (Schwartzman 1997).

In general, the field of the social professions is being taken over by administration, economics, and law, and by various administrative courses peculiar to the technical sector, such as industrial engineering, with little remaining for the traditional social science faculties. Taking into account that social science courses are among the most inefficient in terms of numbers of graduates compared with enrolled students, and concentrating here only on enrolment figures to give an idea of the size of the population involved, it may be observed that among the 10 most popular degree-level courses in Mexico, that of chartered accountant took first place in 1996 (34,653 enrolled students), followed by law (20,983 students) and administration (20,523 students). Medicine came fourth with 8,609 students. In absolute terms, administration, law, teaching, and arts courses in Brazil in 1997 each have more than 100,000 registered students, constituting a third of the total number of enrolments for higher courses (administrative courses have the highest enrolment of all, with 160,000 students, followed by law; engineering occupies third place with 130,000, and accountancy and economics come sixth and seventh with 89,000 and 68,000 students, respectively). In Argentina, the courses with the greatest number of registered students in the public universities in 1994 were law, with 82,896 registered students, psychology with 23,446, social communication with 11,741, and economics with 4,524. The private universities accounted for the highest number of students registered for courses in administration and economic sciences: 38,801 students; law, political science, and diplomacy 26,386, and other social sciences 12,624, with a grand total of 77,811 enrolled students (62.4 percent of total enrolments in the private universities, not counting another 12.3 percent for the humanities).

1.2 The Traditional Academic Courses

The social sciences in the stricter or more 'academic' sense, including sociology, anthropology, and political science, are taught in smaller departments or schools and involve small groups. As they are related more closely to research and postgraduate activity, they are referred to under these headings. The more traditional social disciplines, history and geography, expanded from the 1940s onwards in countries like Brazil, in order to meet the needs of secondary teaching, in accordance with a French tradition that proved impossible to revive before it was overtaken by the dramatic deterioration suffered by secondary education in the region. The profession of secondary-school teacher gradually lost its luster, in the overall crisis that overwhelmed that level of study. However, history and geography renewed themselves in several countries and produced valuable modern professionals and research schools.

The most recent quantitative surveys of the social sciences report certain common symptoms in the region which indicate a certain stagnation on the part of the traditional social disciplines: a drop in enrolments, cancellation of research programs and units in the public sector, a greater uncertainty among students of these subjects about their professional future, a growing reduction in suitably attractive economic opportunities for research and teaching, and a wide variety of problems at educational establishments, *inter alia*, form part of the fluctuating fortunes of the social disciplines. Paradoxically, however, most study plans, regardless of the level concerned, place the emphasis on a professional profile with a bias towards research. While there are greater differences within the public and private sectors than between individual

institutions belonging to each of the two sectors, public institutions appear to place greater emphasis on research-oriented training, although this is invariably a very small part of academic activity.

1.3 Teachers' Recruitment

The recruitment of academics in higher-education institutions in the social and administrative sciences increased most between 1970 and 1992, when it acquired its greatest ascendancy. The mass explosion in tertiary-level enrolments was such that these academics were for the most part absorbed into university life. The absorption rate in Mexico was as follows: 56.5 percent were recruited into teaching positions as soon as they finished their first-degree studies or even without getting a degree (35.4 percent), only 2 percent having a doctorate and 4.3 percent a master's degree, and without teaching experience (60.5 percent), without having practiced their profession (30.4 percent) or without any research experience in research (91.8 percent) (Gil et al. 1994). In the early 1990s 60.8 percent of academics still only had a first degree and 13.5 percent a specialization diploma. At the other extreme were the natural and exact sciences, with 54.4 percent doctor's and master's degree holders. The academic training levels of university personnel rose strongly in most countries during the late 1990s, as a result of conscious efforts by national authorities. However, not always the postgraduate studies in which they engaged themselves to improve their qualifications culminated successfully.

Data on developments in the teaching profession in Argentina suggest that its composition there and in certain other countries (like Mexico) suffers from structural problems, such as the very small proportion of full-time contracts compared with the total. Only 4.5 percent were full-time and 19 percent half-time by mid-1990s. The rest (76.5 percent) were part-time, contrasting with the basic and technological sciences where the picture was mainly one of full-time and half-time teaching. In other countries, such as Brazil and Venezuela, efforts have been made to introduce full-time teaching in universities, although in the latter, deteriorated work conditions in public universities in the 1980s and 1990s have resulted in a lower number of full-time positions and an insufficient replenishment rate of posts vacated by retiring staff.

2. Postgraduate Courses

Demands for credentials of formal academic training have risen significantly. People often go in for postgraduate courses in order to improve their skills and/or obtain higher qualifications that will confer formal institutional recognition on their ongoing academic careers. Yet it is clear that the postgraduate phase is not exclusively synonymous with a high level of research but can also mean greater professionalization. Despite the stated expectations and aspirations of master's and doctorate programs as channels contributing to raise the research capabilities, the training of researchers is limited because these programs appear in many cases to remain fundamentally divorced from research activity. Researchers and the institutions where they work often experience great difficulty in linking research and teaching, owing to the existence of rigid institutional thinking. In many institutional contexts in Latin America, research and education are juxtaposed rather than integrated activities.

In Brazil, for some decades, postgraduate courses in social sciences were developed around the master's degree, which assumed great importance and became a sort of mini-doctorate until full doctorates began to be introduced much later (Durham 1991). Government policy was already recommending in 1965 that a research doctorate should be distinguished from a professional doctorate, with the same distinction being applied to master's degrees. Failure to observe that recommendation caused serious difficulties in the different fields. Despite the regulatory intention behind initial government planning, the growth in postgraduate courses in the 1970s went out of control. There were too few qualified lecturers, postgraduate courses were introduced in already saturated regions, programs requiring little investment proliferated and an expansion exceeding the human and financial resources available occurred. More recently, CAPES (Coordination of Advanced Training for Higher-level Personnel) has been performing a useful task of monitoring and evaluating postgraduate courses, thus helping to raise and homogenize quality standards of performance. In 1997 it accredited the following courses in the social field: in the applied social sciences, including law, administration, economics, architecture and town planning, land-use planning, demography, information science, communication, social service, and industrial design, a total of 143 courses between 104 master's degrees and 39 doctorates; in the human sciences 178 master's degrees and 87 doctorates, which include the traditional humanities (philosophy, history, education, theology, and geography) and the social sciences (sociology, anthropology, psychology, political science). A third area (with 69 master's and 41 doctorate courses) comprises linguistics, the arts, and the fine arts. Master's courses in the social field altogether account for 27.7 percent of all master's courses validated by CAPES and for 24.37 percent of all doctorate courses. The greatest number of courses per discipline is found at master's level in education (47 master's degrees and 20 doctorates), the arts (44 master's degrees and 29 doctorates), and psychology (28 master's degrees and 16 doctorates), with high figures also being observed for economics (26 and 11), administration (23 and 8), and sociology (21 and 12).

In Mexico, the other Latin American country with a high proportion of postgraduate courses, it has not been possible to identify the total number of courses for the social sciences, although information has been obtained on enrolments and graduates. Registered postgraduate numbers in the social sciences in 1997 were 37,160 (42.3 percent). If the education and humanities population is added to this, a figure of 53,550 (61 percent) is obtained.

The growth in the number of postgraduate students in the various social disciplines and the humanities is uneven and significant differences exist between courses at the various levels. The specialization level is dominated by courses leading to professional qualifications, for example, administration (1,083 enrolments), law (1,359), taxation and finance (2,231), psychology (558), and sales/marketing (435). The master's degree level also has many students enrolled in professional qualifying courses in administration (16,923), law (2,851), taxation and finance (2,425), and psychology (2,248) but it already shows the presence of some courses directed at training for research, for example, in economics and development (2,104), social sciences (603), and communication sciences (518). At doctorate level, with much lower numbers, law enjoys the greatest number of students (478), followed by the social sciences (342), anthropology and archaeology (246), and economics and development (158). With regard to the most relevant disciplines for promoting a core of social researchers, only a modest expansion is noted at master's level, which can be attributed to the absence of diversified options on the labor market and to the fact that the academic market does not appear large enough to absorb future new researchers. A particularly noticeable drop of 44.5 percent in sociology enrolments has occurred since the mid-1980s. During the same period the number of economics students underwent a major increase of 270.4 percent, in anthropology the figure grew by 62.8 percent, and in history by 66.3 percent. At doctoral level, although increases are found in the five social and human science research disciplines, there are only 952 students at that level, representing 15.5 percent of total doctorate numbers in 1997.

3. *Social Research and its Institutional Loci*

The bulk of current centers has been established since 1970, especially between 1970 and the early 1980s. The proliferation of centers in the last third of the twentieth century seems to be inspired by needs of various kinds. Part of the reason for this can be found in the impetus given to the whole of higher education throughout those years. Many were set up as an aid to teaching, others because of the need to explain local or regional realities or on account of some particular social or cultural problem. The smallest number were probably established in order to promote theoretical, methodological, and instrumental progress and innovation. It is also important to note, however, that a considerable proportion of the existing centers has been set up since the mid-1980s, with an emphasis on education, economics, anthropology, and sociology centers. Other centers were also set up for history and administration during this period. Certain disciplines, such as education and anthropology, although cultivated in a large number of traditional centers, have recently received a boost that increased their number. Institutional and disciplinary variety is the norm.

Different social science trends and models are found in the region: at one extreme, are the early 'modern' programs, such as the Sociology Institute of Buenos Aires University and, soon afterwards, the Torcuato di Tella Institute, in Argentina; FLACSO in its several branches (Buenos Aires, Brasilia, Santiago, San José, Havana, Quito, El Salvador, Guatemala, México D.F., and Santo Domingo), born as a cooperation initiative between UNESCO and the governments of the region and aimed at promoting education, research, and technical cooperation in the social sciences throughout the subcontinent; the University Research Institute of Rio de Janeiro (IUPERJ), the Postgraduate Program in Social Anthropology (PPGAS) of the National Museum, also in Rio; the Brazilian Center of Analysis and Planning (CEBRAP) in Sao Paulo; the Department of Political Science in the Federal University of Minas Gerais, and the Integrated Master's Degree Program in Economics and Sociology (PIMES), which was taken over by the Federal University of Pernambuco in Brazil; El Colegio de México; the National Institute of Anthropology and History (INAH) of México; the Institute of Sociological Research, which was soon renamed as Sociology Institute of the University of Chile and the Center for Economic Research and the Sociology School in the Catholic University of Chile; and the Development Studies Center (CENDES) in Venezuela.

At the other extreme are the programs created through the efforts of groups of lecturers, who utilized the material, administrative, and personnel resources available in the undergraduate departments and survived by 'stealing' undergraduate facilities. Between these two extremes arose programs at the new, more flexible, more modern, and less bureaucratic universities where the shift in activities from the undergraduate to the postgraduate phase enjoys more institutional support, which in turn helps to attract external financing for research and infrastructure. Los Andes and del Valle Universities in Colombia, the University of Campinas (UNICAMP) and Brasilia University (UNB) in Brazil, the Ibero-American University and the Autonomous Metropolitan University in Mexico are in this situation. The courses with the most satisfactory results seem to have enjoyed a greater relative autonomy in their links with university administrative structures.

Many recent centers have been set up with inexperienced staff who are trained to below doctorate level, while the more experienced researchers remain at the more soundly established centers. The proportion of researchers under the age of 35, however, is noticeably small, which indicates an aging tendency on the part of the researcher population and suggests an uncertain future for some of these disciplines. In Brazil, considerable effort has been made to ensure the highest possible level of training for university lecturers involved in postgraduate activity and in research. Argentina, Chile, and Mexico in recent years have established government programs for providing further and refresher training for the teaching staff covering, among other things, the postgraduate training of academic personnel. Colombia boosted its general postgraduate scholarship program during the 1990s with the support of COLCIENCIAS, but of late seems to have lost its dynamic edge.

The level of geographical concentration of research units in the various countries, particularly the largest, have tended to fall since the 1970s, resulting in a slightly more balanced distribution, although a strong concentration persists in metropolitan areas. In Mexico, all states have research centers in some social discipline, which is not the case in other knowledge areas. There, however, as in Brazil, Venezuela, Colombia, Chile, and Argentina, the persistence of the regional question, despite endeavors to ensure a better distribution of resources and the repeated attempts to establish research groups in the province, suggests that perhaps other causes should also be examined, for example, the possibility that academic institutions in these regions possess less autonomy and are overdependent on the political interests and views of dominant local groups. Situations of this kind can impede the introduction of universal criteria for personnel recruitment and the distribution of resources, thus exacerbating the factionalism inherent in academic structures.

Although the social sciences show a substantial increase in postgraduate courses, no equivalent impact is perceivable on the percentage of recognized researchers in those countries where the formal recognition of researcher status has been introduced. In Mexico, while the creation of the National Researchers System (SNI) led to the formal establishment of a social, institutional, and normative framework for recognition and validation of the researcher's activity and position, it is clear that it also brought with it standard specific requirements with which the academic community is virtually compelled to identify. The social science and humanities field is the one with the smallest concentration of researchers included in the SNI, despite the fact that it has the largest numbers of graduate students and a significant proportion of postgraduate students; and 60 percent (1,053 researchers) are at the lower level, which suggests low productivity according to the program's measuring parameters. Although this is the area of the SNI that has been growing most consistently since 1984, the population willing to participate in this program appears to have reached a ceiling fluctuating around 800 yearly applications.

In Venezuela, a similar program to the Mexican one exists, the Researcher Promotion Program (PPI), with features and trends resembling the Mexican ones. It includes 18.9 percent of researchers in the social science area (270 out of a total of 1,422 researchers in 1997) even though the social sciences account for more than 80 percent of students following degree courses. Nearly 50 percent of the social researchers recognized by the program are situated at the lower level, which implies the minimum productivity acceptable under the program.

The type and rate of production are influenced by disciplinary patterns and traditions: the more 'empirical' or 'pragmatic' disciplines generally are slanted towards production with more immediate implications (e.g., articles); books, on the other hand, appear to be typical expressions of the more traditional disciplines. There are, nevertheless, differences in styles and standards of publishing and, in general, of production of results, even within the same discipline, depending on whether the researcher's approach tends to be local or international. As an illustration of the support and encouragement given to publishing standards, it is worth mentioning the existence of significant publishing industries in Mexico and Brazil, which supply the needs of the educational sector and recently have been responding to pressures upon the academic market to produce a greater number of results in the framework of the evaluation systems laid down.

A large number of social science journals is published in several countries. Given the current initiative of producing the Latin American and Iberian Index of scientific periodical journals (LATINDEX), it is expected that a supplementary regional index to the Science Citation Index, with the scientific production in Spanish and Portuguese will become available in the not too distant future. Meanwhile, several countries have undertaken the task of evaluating the formal and eventually the substantive quality of the journals locally produced and thus several national lists of selected good quality social science journals are available in the National Science Council offices. So far, however, those journals have very limited international visibility (Cetto and Alonso 1999). As an illustration of this, it may be said that the Institute of Scientific Information's Journal Citation Reports (1997 Social Science edition) included only one Argentine journal (*Desarrollo Económico*), two Brazilian ones (*Dados-Revista de Ciencias Sociais* and *Revista de Saúde Pública*), one from Colombia (*Revista Latinoamericana de Psicología*), and four Mexican ones (*Revista Mexicana de Psicología*, *Salud Mental*, *Salud Pública de México*, and *Trimestre Económico*).

In addition to this and turning inwards in the

region, it is observed that despite the existence of a large number of specialist journals, an essay style prevails in many fields. This essay style may be largely the result of an intellectual tradition, but it also relates to the limited market for employment in academic institutions, which means that opportunities for social scientists in the region often have been provided by publishing houses catering for general readership, by the daily and weekly press, by the political parties, by consultancy, and the preparation of technical reports for the governmental and private sectors. An important consequence of such situations frequently has been to restrict the range of topics covered and approaches to working and writing. In many places, issues have concentrated on a limited number of topics and options for discussion around a small number of individuals. This may have hindered the development of a social science involving the use of more complex quantitative techniques or the handling of less fashionable international literature. It is not easy to sustain a debate on any topic in an academic symposium when the results of the discussion bear the imprint, to a greater or lesser extent, of the views of a particular political party or agree with what the journalists of the mass media contrive, or wish, to hear or consider that their readers wish to read, or reproduces discussions and views developed and sustained in the world intellectual centers of Paris or Boston. This has been the price to pay for the role of intelligentsia that many social scientists have assumed in Latin America.

4. *Substantive Issues: Themes and Methods*

At the crossroads of institutional articulation and thematic development in the region is found the Latin American Social Sciences Council (CLACSO). Since its creation in 1966, it has formed the most extensive coordination body for social science research centers in Latin America and the Caribbean, currently including 101 member centers. Its Executive Secretariat has always operated in Buenos Aires. CLACSO has developed a work program which strengthens interchange mechanisms in order to bring about a greater integration of Latin American social sciences and which defends the working conditions of social scientists at member centers and other institutions in the region whose academic activities and/or personnel were marked by years of authoritarian repression in a number of countries. Its postgraduate program was drawn up to deal with two major areas: the Southern Cone Research Program, which, with financial support from the Council, provided aid in the countries of the sub region to researchers experiencing work difficulties because of their political views; and in cooperation with the United Nations Development Program and UNESCO, the Young Researchers Training Program, since it had become apparent that the main problems in the region were a lack of funds for research and the difficulties experienced by young graduates of universities in obtaining funds from international agencies.

In recent years, the Council's academic activity has been directed at its own medium- and long-term planning against a background of institutional reorganization, at rethinking the Working Groups Program in order to counteract the effects of thematic/organizational dispersion, and to continue action in subject-matter areas of particular importance for the analysis of democratization and adjustment processes in the region. Its 20 working groups have a membership of some 3,000 researchers in a program of academic interchange, debates, and publications which is very illustrative of the thematic interests in the region. Among them are youth, international economics; education, work, and social exclusion; memory and human development, poverty and social policy; unions, workers, and social movements; Mercosur and integration; and political parties and electoral systems. In view of the increasing development of various Latin American information networks, the Network of Networks (*Red de Redes*) project was established with the support of the IDRC in Ottawa in order to improve access by the final user to existing information resources by linking up 18 regional information networks. Its web site (URL: http://www.clacso.org) contains extensive information on research institutions in the region, regional databases about researchers and research projects, a virtual library, and other services.

On the cognitive, social, and functional levels, what social scientists do when they claim to be carrying out research is, in fact, extremely varied. Views on what constitutes research not only vary according to fields of academic activity but also depending on ideological and personal approaches and they evolve over time. In Latin America the great diversity that exists in disciplines and research styles mostly has developed in different socio-institutional contexts. Where policies, either general or relating to individual establishments, have existed, however, these have tended towards homogeneity usually patterned on the norms of the physical sciences. New and old themes combine in novel contributions to social understanding. The necessary linkage between the rich accumulated conceptual legacy of the region's social thought and that required by the new heuristics concerns as much the social sciences as the social movements in today's world of change in sociopolitical and technoscientific models.

See also: Infrastructure: Social/Behavioral Research (Japan and Korea); Infrastructure: Social/Behavioral Research (Russia); Infrastructure: Social/Behavioral Research (United States); Infrastructure: Social/Behavioral Research (Western Europe); Professionali-

zation of Social and Behavioral Scientists: United States

Bibliography

Brunner J J 1988 El caso de la sociología en Chile. Formación de una disciplina (The Case of Sociology in Chile. The Formation of a Discipline). FLACSO, Santiago, Chile

Cetto A M, Alonso O 1999 (eds.) Revistas científicas en América Latina (Scientific Journals in Latin America). ICSU/UNAM/CONACYT/FCE, Mexico

Durham E 1991 La política de posgrados en el Brasil (Postgraduate policy in Brazil). In: Cardenas J H (ed.) Doctorados. Reflexiones para la formulación de politicas en América Latina (Doctorates. Thoughts on policy formulation in Latin America). TercerMundo/Editores/Universidad Nacional de Colombia/Centro Internacional de Investigaciones para el Desarrollo, Bogotá, pp. 193–218

Gil A M G et al. 1994 Los rasgos de la diversidad; un estudio sobre los académicos mexicanos (The Characteristics of Diversity: A Study of Mexican Academics). UAM-Azcapotzalco, Mexico

Schwartzman S 1997 A redescoberta da Cultura (The Rediscovery of Culture). EDUSP-FAPESP, São Paulo, Brazil

Sonntag H R 1988 Duda/certeza/crisis: la evolución de las ciencias sociales de América Latina (Doubt/Certainty/Crisis: The Evolution of the Social Sciences in Latin America). UNESCO/Nueva Sociedad, Caracas, Venezuela

Vessuri H 1992 Las ciencias sociales en la Argentina: diagnóstico y perspectivas (The Social Sciences in Argentina: Diagnosis and prospects). In: Oteiza E (ed.) *La política de investigación cientifica y tecnológica argentina. Historia y perspectivas (Argentine Scientific and Technological Research Policy. History and Prospects)*. Centro Editor de América Latina, Buenos Aires, pp. 339–63

H. Vessuri

Infrastructure: Social/Behavioral Research (Russia)

The state of social science is an indicator of the way a nation views, comprehends, and reflects itself. The development of Russian social thought has been continuously deformed by the strict political censorship of different regimes that were able to sever its natural links with the evolving civil society institutions and to enforce protracted periods of academic stalemate. Though it is never possible to suffocate human creativity, originality, and ingenuity, it was possible to trim social science into an awkward but handy configuration by molding its infrastructure and by cutting the requisite feedback communication. The cost of this authoritarian control is a distorted decision-making structure, an extensive type of economic modernization in a largely unstructured, neotraditionalist society doomed to resolve most of its conflicts through discredited, inefficient, yet inevitable state intervention. As the former KGB chief Juri Andropov admitted when he became head of the state: 'We do not know the society we live in.'

Incidentally, in the same year (1983), only 99 empirical social surveys were completed in a country with a population of 283 million. With the collapse of the Soviet Union, the infrastructure of social and behavioral science became spatially dissected along the formerly unimportant state and regional borders. This discontinuity affected the life of the academic community, its research cooperation, the mainstream methodology, and the general perception of reality. More significant was the temporal shift of social gears. The 'frozen' nation with a stagnant social structure suddenly melted down and found itself in a state of unpredictable flux. After a decade of social transformation and market transition, Russian social science faces a challenge of hitherto unstudied empirical disciplinary and interdisciplinary fields, as well as an opportunity for detached comprehension of the overlooked retrospective social realities. Hence comes an objective integrative role for the social science institutions and structures, which they seem now more fit to play, but whether or not that role will be accepted by the power, the market, and society still remains to be seen.

1. *The Rise and Fall of Social Science Instrumentalization*

Many observers note structural, institutional, and research similarities in Western and Eastern European phases of development. However striking they may seem in the past decades, similar trends in the social sciences had been caused by quite different reasons. Their roots go deep into the past. Authoritarian attempts to instrumentalize relations with social thought had started in the age of Enlightenment when, trying to improve her international reputation, Catherine the Great took credit for the first publication of the complete works of Voltaire and awarded a fellowship to Diderot. Her subordinates followed suit by masquerading dire reality behind the faked facades of Potyomkine 'villages' and claiming that every Russian peasant household was getting a hen in their weekly ration. Utopian social experiments in 'forming a new man' also started at that time. But for authoritarian power, encapsulated in a steel frame of ossified bureaucracy, censorship and loyalty were the prime tools of public control and the works of Auguste Comte remained banned until 1889 for ideological reasons. Because it was less strictly controlled, fiction thus gained prominence and acquired a formative role as a reflexive tool of the society. Literature, including F. Dostoyevsky's social and L. Tolstoy's moral writing, became a major source of social thought. This explains an exaggerated and lasting public enchantment with the scientifically versed dogmatic pres-

criptions that were vested with magic power to reshape life along doctrinal lines.

This was precisely what Bolshevik leaders wanted from the social sciences and by the mid-1920s, when revolutionary aspirations petered away, they had gradually usurped this tool, banning the positivist paradigm as well as most Marxist thinking and promoting one of the Marxian strands of thought to the position of the 'only true and all encompassing' social teaching. Russian educational, institutional, and research infrastructures had been trimmed correspondingly as the new doctrine passed itself off as the ultimate solution that could countenance no tolerance of other schools of thought either inside or outside the country. Even the term 'social sciences' fell under suspicion and was replaced by 'societal sciences.'

Sociologists Pitirim Sorokin and Nikolai Berdyaev, together with several hundred other social scientists and thinkers of international renown were expelled from the country. In 1917 St. Petersburg University was closed. After that, no access to the empirical and theoretical work of preceding decades was possible, though university professors continued to share their knowledge with the new generation of students and thus were able to maintain traditionally high standards of education. This dramatic break with the international and indigenous research traditions in the social sciences greatly narrowed their evolving infrastructure. It was soon to be replaced by a new one that was held to be 'socialist in essence and national in form,' and that, in due course, was able to integrate and unify social knowledge (and dogmas) articulation, research, and dissemination across 15 nation states and subsequently in most of Eastern Europe.

For five decades, humanities and the social sciences were divided into the good (Soviet) and bad (bourgeois) traditions. Many disciplines (including sociology, psychoanalysis, genetics, and later cybernetics) were ostracized and banned for three decades until the thaw of the mid-1950s. The unveiled humanitarian story of Russian spiritual resistance bears the names of hundreds of reputed scholars who risked their lives and perished in the Gulag because they dared to think and write in ways that were no longer allowed.

The prohibited, neglected, and suppressed works on a broad range of social science disciplines were rediscovered and made public only in the Gorbachev era. The dramatic change of focus came from the very top and (contrary to the reformist choice made by the Chinese Communist leadership) was well intended rather than well calculated. It had been based on a rather naive assumption that the easing of stringent controls would by itself recuperate the USSR and get it into a common European home. Official Soviet social science had no experience to draw on to predict the effects of transformation on such a grand scale. The impact of the staggering institutional deficiency on the transition had not been anticipated. Though the breaks in tradition could not be recuperated by retrospective publication and bold analogies, the impact of these powerful flashbacks on the society was enormous. Blank pages of national history rediscovered by *glasnost* actually took the entrenched communist moral stands together with the ideologically ossified social institutions off ground and paved a way to the final collapse of socialism.

The instrumental rationale and the general limitations of the social sciences through more than half of the twentieth century were shaped by the supra-nation-state building imperatives and by combat lessons of the two world wars. The Russian Revolution and a devastating civil war had bred a belief in the universal efficiency of the command decision-making system. Totalitarian over-centralization became the most potent tool of governance. The flamboyant behavioral research and social engineering of the early revolutionary years, which ranged from mass blood transfusion to Freudian reactology, were discarded when it became apparent that mere state violence could effectively drain rural resources for accelerated industrialization and militarization. Regular political surveys and extensive social monitoring became the prerogative of the secret police, the territorial party committees, and the state statistical institutions. Most of the social and statistical information had been strictly classified, including the notorious erasure of the complete 1937 All-Union census data. After the reinstitutionalization of sociology in the 1960s and of political science in the 1970s, all politically sensitive academic research was also classified.

In the environment of strict party and KGB institutional control, the only viable dimension of the social sciences infrastructure became predominantly internal. The institutional and administrative traditions of academic-scientific culture depended on such matters as maintaining high academic standards of excellence and disciplinary innovation, principled integrity, nonparticipation in politics. At various times this implied a need for mimicry and a lot of dual thinking, but even with all the lip service to the obligatory official doctrinal tenets, the key elements for professional standards in the social sciences largely remained intact. This may be illustrated by the Academy of Sciences' blunt refusal to remove its academic title from the exiled academician A. Sakharov in spite of the Politburo's stringent demand. In a similar way it declined proposals to establish an academic institute of scientific communism though this dubious social discipline (with corresponding university chairs throughout the country) had been a standard element in higher education since 1963. These distinctions explain a considerable degree of diversity and dissent in social science's orientations under the Communist regime even within a constrained research configuration.

The multi-tiered infrastructure of the social (societal) sciences constructed in the Soviet Union had been conceived as the most monumental in insti-

tutional, spatial, and mass ideological coverage. Its configuration in scope and priorities resembled an inverted pyramid resting on a narrow base of academic research. Academic units had no educational obligations and were more important as institutes dominating in applied and fundamental research. They were relatively independent and had strong institutional linkages with their branches in Siberia, the Urals, and the Far East, as well as with the 15 national academies of science. Two other academies (medicine and pedagogy) oriented toward practical needs had several applied research institutes in public health, demography, and secondary education.

The higher education system, including the university social sciences, had greater regional coverage and was oriented to ideological, knowledge base, and secondary analysis areas of social study conducted mostly in universally spread professorial chairs of Marxism–Leninism or scientific communism and atheism. Several universities had faculties, departments, or laboratories that were science-driven and provided some opportunities for problem-solving work to the graduate and postgraduate students. But both academic and university studies aspiring for an academic qualification, apart from the public defense in corresponding disciplinary councils, had to pass through the rigid screening procedure of the All-Union Attestation Committee.

Over that layer was a plethora of specialized organizational applied research scientific entities that provided special professional services. Set up by various ministries or the state planning committee, they covered themes ranging from cinematography and art criticism to urban planning and healthy food production and to social aspects of military service and combat operations. They would also be loosely related to the extra-university government-directed network of territorial social data monitoring institutions with statistical-quantitative orientation. Overlapping and unnecessary duplication was a specific feature of the party-state system.

Relevant social data collected in every territorial unit would also get to the parallel network of the party committees that could overrule governmental plans and, through *obkoms*, reported directly to the Central Party Committee. The party network designed to boost growth, very soon turned into an institutionalized social impediment to change and development. The greater the discrepancy between normative prescriptions and real life situations, the greater was the need for social indoctrination that, in principle, had to cover the entire adult population of the country. This task was vested in vocational 'institutes' of Marxism–Leninism and in the regular political information gatherings held at all enterprises that were controlled by party committees. A more varied repertoire of social subjects was presented by the lecturers of the All-Union 'Knowledge' society that had its branches in every town.

The well-intended *ad hoc* communist social discretion was practiced as a main form of institutionalized 'scientific' social practice till the early 1960s. By that time, over 100 million peasants had moved out of decaying rural areas and, when extensive exploitation of mass enthusiasm in plowing the virgin lands failed to yield the expected results, the need for a reform finally became apparent to some key decision makers. Externally, the interventionist support of national-liberation movements, the Cuban missile crisis, and growing displeasures in Eastern Europe also spelled out the need for something institutionally more reliable than pure rhetoric and virgin historical materialism. Its proponents in the academy were greatly impressed by the scope and range of coverage of foreign studies at the Amsterdam Sociological Congress (August 1956) and an interdisciplinary social study of technical progress had been started on the UNESCO initiative. The Soviet Sociological Association was established in June 1958—a decision driven more by a desire to put on a better international face rather than by internal necessity. The positive element here was not only an exposure to the international social research agenda but legalization of contacts with foreign, at that time particularly Polish, scholars.

Yet it took the social sciences another decade before there was an institutional breakthrough. The Institute for Concrete Social Research (ICSR RAS) was finally set within the Academy of Sciences in 1968. It launched three major research projects on social structure and social planning, on the management of social processes, and on the history of sociology. Within the first year of its work, 20 social surveys had been made for the Central Committee of the SPSU. But with the military incursion in Czechoslovakia political winds again went in the wrong direction. Economic reforms were curbed and social changes came to a stalemate. The flood of foreign (in Russian translation) and local social science publications turned into a tiny trickle. Ideological blinders were imposed on all social publications including the only academic journal, *Sociological Research* set up in 1974. Institutional constraints once again had turned social studies into the stale waters. In 1972 and in 1976 the staff of the ICSR RAS had been ostracized and the new period of state-imposed stagnation lasted at least until 1988. Reading between the lines had been the favourite pastime of the academic community and innovative research would appear only on the periphery of the social sciences' institutional mainstream.

2. *New Perspectives*

The collapse of socialism liberated the party-state nomenclatura from any social responsibility and during the Yeltsin era much of the democratic reform effort was compromised. In contrast to the 'old deal' pressures, the institutional structure of the social

sciences remains largely unattended, and though its financing still remains below subsistence level, it has gained momentum for development and an impetus for institutional change. 1991 was pivotal in this respect. The Russian Academy of Sciences (RAS) reclaimed its status lost in the 1930s. The Institute of Sociology (www.isras.rssi.ru) split up and the Institute for Social and Political Studies (www.ispr.ras.ru) was formed in Moscow, while the Sociological Institute RAS remains in St. Petersburg. All these institutes together with the other academic research units and related professional associations are subordinated to the RAS branch on philosophy, sociology, psychology, and law. The Institute for Scientific Information in Social Sciences, as well as fundamental science libraries, currently enjoys a special status within the academy.

Other branches in the social sciences include history (institutes of oriental studies, archaeology, culturology, general and Russian history, ethnology and anthropology, history of material culture, military history); economics (institutes of economy, national economic forecasting, market problems, economic problems of the transition, socioeconomic problems of employment, and the Central Economic Statistical Institute); world economy and international relations (institutes of peace, of USA and Canada, of Africa, of Europe, of Latin America, of comparative politology and workers movement, and of world economy and international relations). The next three branches include the literature and language branch (institutes of world literature, Russian literature, Russian language, linguistics, language research, history of arts, problems of architecture, and urban construction); oceanology, atmosphere physics, and geography (institutes of geography, global climate and ecology, water resources, lake research, etc.); and the physiology branch (institutes of man, of human brain, physiology, evolutional physiology and biochemistry, higher nervous activity and neurophysiology, of medical and biological studies, etc.) This incomplete list is intended just to outline the continuity of the maintained range of publicly financed academic research activities throughout the painful transition period.

With the democratic changes, foreign donor foundations have moved in and academic grants were offered to various projects in the national social sciences. Focused on pivotal issues, these grants provided an alternative infrastructure that helped to invigorate research and exchange programs, to write the new textbooks, and to disseminate these publications in the provinces. George Soros was the first and the most effective in these promotional and support activities, starting with his 'cultural initiative,' later transformed into the 'open society institute' and with the Research Support Scheme that promotes cross-cultural academic cooperation. His efforts were followed by a number of Western private philanthropic institutions like Carnegie, MacArthur, and the Ford foundations. Two newly established indigenous funds RFFI and RGNF (the Russian Fund for Fundamental Research and the Russian State Scientific Fund) are also active in this market. Thus, a competitive wedge has been established in social science activities.

A considerable proportion of support funding helped to keep afloat all the Russian academic journals including Sociological Studies (*Sotsiologicheskiie issledovanija*) (http://www.isras.rssi.ru/R_SocIs.htm), Sociological Journal (http://win.www.nir.ru/socio/scipubl/socjour.htm), Journal of Sociology and Social Anthropology (http://www.soc.pu.ru:8101/publications/jssa/), and the Social Sciences Journal. This continuity seems important as it helps to maintain high academic standards. More significant, however, is the free market competition among the emerging private publishers. In the 1990s the stock of the available translated literature and publications in the social sciences has grown considerably, giving ample place for the requisite variety of innovation, yet the public libraries frequently can afford neither expensive new books nor subscriptions to the foreign academic periodicals as the traditional state policy of obligatory purchases and hard currency allocations to the libraries has been stopped. An unexpected side effect of the democratic and market changes has been another structural constraint and challenge to the functional foundations of the social sciences: their traditional sphere has been invaded by a range of recently founded self-appointed 'academies' that find their rationale mostly in trading their own academic titles to those who crave this special sort of prestige. At the same time, various surrogate subcultures have grown up on the fringes of the academic sphere, offering the public their para-science services. Strategic rethinking of the present truce with various new forms of mass occultism may be one of the institutional tasks for the social sciences in the twenty-first century.

Turning financial constraints into a window of opportunities, a great number of freelance and private research structures came into the emerging market for the social sciences. They now range from the abovementioned panacea institutions to the recently founded social monitoring units (also including the 'black PR' foundations for efficient policy) in the President's administration, to private analytical firms (institutes), and subcontracted focus groups. Some of the new nonacademic structures have NGO status but their activities—that need thoroughly checked documen tation—may cover important segments of social life, particularly when they systematically monitor new developments and conduct empiric surveys. The Committee of Soldiers' Mothers and the ecological 'Kedr' movement are good examples of such organizations. On the contrary, such well-established organizations as trade unions that had infinitely more advantageous institutional resources have failed to maintain their social niche in this time of transition.

Private independent institutions that consistently operate within the domain of academic science started to emerge in the end of 1980s. The oldest of them, the public opinion polling center Vtsiom (http://www.wciom.ru) has accumulated in the last decade of the twentieth century a database with over 2,000 polls and, together with the other active monitoring institutions like the Fund for Public Opinion FOM (http://www.fom.ru) and Romir, has gained considerable weight in sociopolitical surveying. Due to decades of political void in Russia, the role of the polling institutes and the media at the beginning of the twenty-first century is more formative than informative. In these circumstances, sociologists are worried that public-opinion polling has been turned into a surrogate instrument for political manipulation. It appears that with the help of social technologies relayed through media channels even political nonentities, like the Edinstvo/Medved party, can win seats at elections after less than six months of campaigning—as evidenced by the last elections for the Duma and the presidency (in 1999 and 2000 respectively). In the first decade of the twenty-first century, the rapidly shifting nexus between the vested interests and social science application will become the most rapidly evolving institutional configuration.

Recent developments in higher education have introduced paid education in both private and state institutions. Major changes in the curriculum have reoriented the students to multiple courses in market-related disciplines, yet behavioral and social studies are also gaining prominence. Sociology has replaced Marxism as an obligatory course in university education, with more than 250 new departments established throughout the country. The state standard for sociology currently lists 14 obligatory themes in the history of sociology, 12 in sociology, and 13 themes in the methods of research. Apart from structural educational changes, empirical social studies have become a regular activity in university science, for instance in the Center for Sociological Research MSU (http://www.opinio.msu.ru) or in the work of the Sociological Research Center of the Federal Education Ministry (http://www.informika.ru/windows/goscom/cinorgan/socio/first_pg.html). Many universities are becoming active in their student research and degree exchange programs viewed as a precondition for futher professional training. But, on the whole, the budgetary and structural limitations in the higher education reform leave little place for radical changes in the traditional forms of university research. Such instruments as Internet resources are rapidly forming the most promising base for strengthening student research and university science. The social science infrastructure of today is still not systematically presented on the Internet, but since about 1997 major efforts have been made to form the aggregate professional platforms and networks that would integrate individual and institutional websites.

This may be illustrated by the Psychological Net of Runet (http://www.nsu.ru/psych/internet/index.htm); the Organizational Management Consultants Net (www.odn.ru); History and computer server (http://kleio.dcn-asu.ru/aik/index.html); by the recently established (1999) Russian Sociological Net (RSN) (www.isn.ru) which has a database of about 100 research institutions and projects (www.isn.ru/sociology.shtml); by a directory of Russian sociological sites (www.isn.ru/sociology/public/sociology.htm) that offers an online periodical COSNET.RU (www.isn.ru/index42.shtml); and by the Russian archive of working documents on economics and sociology (www.ieie.nsc.ru:8101/r-archive/index.htm).

With the massive growth of computerized information gathering and processing one can anticipate the increasing diversification of services rendered through the traditional social sciences' infrastructure, by the NGOs, and by the private freelance commercial agencies. Digital forms of cultural representation are offering new vistas that will both preserve and transcend local cultures. It now seems quite apparent that the transnational social change secures the social sciences a place of prominence in this rapidly evolving cross-cultural discourse.

See also: Communist Parties; Democratic Transitions; National Traditions in the Social Sciences; Socialist Societies: Anthropological Aspects; Soviet Studies: Culture; Soviet Studies: Politics

Bibliography

Jadov V (ed.) 1998 *Sotsiologija v Rossii* [Sociology in Russia]. Moscow

Osipov G (ed.) 1995 *Entsilopedicheskii sotsiologicheskii slovar* [Encyclopedic Dictionary of Sociology]. Moscow

State standard on Curriculum for Sociology Gostandart po kursu Cotsiologija//Sotsiologicheskije isslredivania 1993. No. 7, pp. 151–4

Sociology from A to Z (http://www.glasnet.ru/~asch/-sociology/)

Tchernih A I 1998 *Stanvlenije Rossii Sovetskoi: 20 godi d zerkale sotsiologii*. Moscow

V. Nemchinov

Infrastructure: Social/Behavioral Research (United States)

Progress in science presupposes an infrastructure that sustains research and exchange of new scientific information, that prepares requisite personnel, and that advances cumulative knowledge of the empirical world. This is no less true in the social and behavioral sciences than elsewhere in science and engineering. While science is increasingly international, with its

own sustaining infrastructure, a primary base for advancing science exists within each nation. In the USA, the foundations of the existing infrastructure for the social and behavioral sciences (hereafter, social sciences) were laid in the early decades of the twentieth century and subsequently expanded just after World War II (WWII). At the beginning of the twenty-first century, however, certain limitations of this infrastructure are apparent. Perforce, these factors may constrain progress in social sciences in the USA and may fail to encourage a full participation in or realization of an international science of behavior and society as the new century unfolds.

1. *What is Infrastructure?*

For most of science, 'infrastructure' conjures up images of national laboratories, orbiting telescopes, polar weather stations, or supercolliders; in other words, bricks, mortar, and the research instruments of 'big' science built and maintained at great expense. What this image misses, even for the natural and physical sciences, are the myriad elements of science organizations, such as national academies and scientific associations, funding sources and review systems, libraries and databases, theoretical models and methodologies, journals and other communication media, and advanced training programs that separately and collectively sustain and advance the research enterprise of individuals. In fact, it is easier to say that 'infrastructure is an indispensable adjunct to the efforts of individual scientists' (Prewitt 1995) than it is to specify what comprises the comprehensive and necessary elements of science infrastructure. This may be especially true for the social sciences for which the 'bricks and mortar' components of infrastructure are fewer or less obvious (CBASSE 1998). Thus, any definition is apt to be rather arbitrary if only to the degree that from the myriad enabling elements some selection must be made, lest the list be boundless and the definition, useless.

At a level of abstraction, the terminology, infrastructure of science, emphasizes shared and sustaining resources that not only make scientific research possible for individuals but also add value—often multiplying the value—of individual research activity. 'Shared' refers to what is accessible as a public good for a science, something whose value is not diminished by multiple uses or that produces economies of scale or multiplier effects. For example, a multimillion dollar investment in building an instrument for magnetic imaging of the mind at work is much like a similar dollar investment in a 10-year panel study of a nation's youth as they leave school and enter the workforce. Many investigators can use both 'instruments' for a great variety of analytical studies, each of whose value is not diminished, indeed, perhaps even made greater by virtue of others using the same instrument for different scientific projects. This is especially true when such projects—all conducted at low additional marginal cost because of the same base investment in the 'instrument'—allow for the aggregation of information and therefore newly emergent information not available from the separate studies. For example, multiple time-series of different topical domains might cross-reference brain images over time in cognitive and in emotional activity, or, from the same sampled cohort, might interrelate the economic outcomes of the school to work transition and of the trajectories of social development over the same period. The term 'sustaining' refers to necessary resources without which the advance of science would be slowed or halted, or, would be very expensive to nearly impossible to reconstruct if left to deteriorate. For example, if the US congress were to cut the funding of the National Science Foundation in half for, say, a four-year period because of a change in political or fiscal priorities, research universities might be forced to close some laboratories or institutes, and some might even redirect their missions. A 40-year time-series on the American electorate, broken for a decade, may never retrieve the value it would have had were it not interrupted (much like temporarily 'turning off' a telescope designed to capture rare events, like the birth of a supernova).

2. *Historical Development of Infrastructure*

Most of the foundations of the infrastructure of the social sciences at the end of the twentieth century were laid in the two decades prior to WWII and just after its conclusion (Prewitt 1995, Bulmer 2001). In the earlier period, a remarkable convergence of interests within the leadership of private philanthropy and of the burgeoning social science disciplines and their national associations yielded an extraordinary array of institutions and resources that were to both consolidate and accelerate the growth of social research as a 'science of society.' Elements of this new infrastructure included the University of Chicago and its celebrated Chicago School of community-based social research, the Social Science Research Council in New York City, the National Bureau of Economic Research, and what later was to become the Brookings Institution in Washington, DC. In the later period, with the rapid evolution of the federal administrative state following the great depression and especially after WWII, the National Science Foundation and the National Institutes of Health funded an expansionary phase of basic and mission-related research. Most of this science-building activity occurred within the nation's research universities, themselves a product of the postwar era. Yet the two growth periods for infrastructure resulted from different interests and

goals and emerged from distinctly different historical moments. The course of social sciences in the USA reflects the history of its infrastructure just as much as it does the intellectual biographies of its leaders and evolving corpus of scientific contributions (Deutsch et al. 1971).

2.1 The Foundational Period

At the close of the nineteenth century, the roots of what was to become American social science drew their energy from investigative journalists and reformers in settlement houses and from statisticians seeking to count unemployment and develop statistically reliable representations of labor markets in local areas (Bulmer 2001). It would be incorrect to speak of an infrastructure of social science as such at that time, because the nature of social inquiry was hardly scientific in the sense of being driven by intellectual curiosity and disciplined by interrogating ideas and conclusions with careful measurements of social phenomena and behavior. Statisticians such as Carroll Wright were developing crude labor statistics for the state of Massachusetts after the civil war, mirroring the 'statistical movement' in Victorian Britain and the 'probabilistic revolution' pioneered by Quetelet and others in Europe during the early to mid-nineteenth century (Bulmer et al. 1991) . However, what drove even this descriptive, social statistical enterprise was related to—perhaps even dominated by—what motivated muckraking journalists and also reformers and advocates for the poor like Jane Addams and W. E. B. Du Bois at the beginning of the twentieth century. This was a concern—often embedded within progressive reform ideology—about the apparent fragility of democracy and threats to economic self-sufficiency in an America beset with racial and labor conflicts, massive numbers of immigrants, and unstable economic cycles with impoverishment for many.

This nonscientific concern at the beginning of the twentieth century did lead to systematic social investigation, largely in the hope of pointing the way to reforming society, relieving the misery of the destitute, and preserving democracy. The Settlement House movement, devoted to improving the lot of urban poor, inspired and drew upon a series of mapping projects and urban surveys, e.g., Hull House Maps and Papers; Du Bois' The Philadelphia Negro, the Pittsburgh Survey. These proto-scientific inquiries in the late nineteenth and early twentieth centuries, some commissioned and funded by the Russell Sage Foundation in New York City in the interests of improving social welfare, comprised one cornerstone of the Social Survey Movement. And in turn, they gave impetus decades later to what was to become modern statistical sampling and survey research during and just after WWII (Converse 1987, Bulmer et al. 1991).

Private philanthropy, created by industrial capitalists (or their families) like Andrew Carnegie, the Rockefellers, and Russell Sage in the nineteenth century and administered by charitable foundations bearing their names, provided the initial means and set a scientific course for social science's development in the early decades of the twentieth century (Bulmer and Bulmer 1981, Prewitt 1995). To that extent, these private foundations created the first explicit infrastructure for a scientific social science. Whether this investment was motivated by the self-protective interests of a capitalist social class (i.e., Fisher 1993) or reflected deeper religious and humanitarian concerns in the public interest (i.e., Bulmer 2001, Lagemann 1989) can be debated. In any case, the early infrastructure of social science emerged from a fortuitous convergence of the announced goals of philanthropic foundations and the coalescence, within the USA, of defined social science disciplines—history, political science, psychology, economics, anthropology, and sociology (Ross 1991). An important additional element of that early infrastructure was the creation of graduate schools at a few universities, starting at the Johns Hopkins University in 1876 and then at Clark, the University of Chicago, Columbia, Harvard, and Yale.

As for the foundations, leaders such as psychologist Beardsley Ruml at the Laura Spelman Rockefeller Memorial drew inspiration from the natural and medical sciences as a means of understanding and (it was presumed) controlling the phenomenal (empirical) world and set as a goal the creation of a corresponding science of society. This science of society was not an end in itself; it did not reflect so much a commitment to science *per se* than to use rational analysis and scientific research to reveal the 'root causes' of the massive problems of industrial America and thereby to overcome them. Investment into research and the scientific enterprise, e.g., into the putative causes of urban poverty rather than exclusively into assisting poor families, reflected a broad sea change in the goals of private philanthropy in the early decades of the twentieth century (Prewitt 1995). This change in philanthropy led to the creation of several research institutes in medicine and the natural sciences, e.g., the Rockefeller Institute for Medical Research (1901). But it also coincided with the rapid emergence of several additional foundations and organizations that supported the development of social and behavioral research, e.g., the General Education Board (1903); the Milbank Memorial Fund (1905); the Russell Sage Foundation (1907); and the Carnegie Corporation of New York (1911). Subsequently, these in turn created several core elements of infrastructure: the National Bureau of Economic Research (NBER), established in 1921 under the leadership of economist Wesley Mitchell, was created by the Carnegie Foundation; the Brookings Institution, in 1927 (after consolidation of several other organizations, including the Institute

for Government Research, founded in 1916 with Rockefeller monies); and the Social Science Research Council (SSRC), incorporated in New York City the 1920s with Rockefeller backing.

The early history of the SSRC amply illustrates the evolving close partnership between private philanthropy and university-based research of the social science disciplines after World War I. At the core of this history is a very effective network of leadership within the foundations and the small set of elite university research programs—leadership that circulated between these two worlds and galvanized common cause to promote a positivist science of society. Charles F Merriam, president of the American Political Science Association, spearheaded a cadre of counterpart economists, sociologists, and historians (and joined later by anthropologists and statisticians) with commitments to a scientific approach in their respective fields but also to multidisciplinary research inquiries. They founded the Social Research Council (SRC) in 1923 to use a scientific, interdisciplinary approach in analyzing the problems of society, to develop new methodological tools for this science of society, and to assist university-based academic programs in training a new generation of social scientists for this agenda. Merriam and colleagues considered but ultimately rejected the idea of affiliating with the American Council of Learned Societies (ACLS). (The ACLS was chartered in 1919 to represent the USA within the Union Academique Internationale; it was one of the first councils to span the social sciences and humanities and to represent and develop their scholarship.) Instead, they pursued, unsuccessfully, incorporation within the National Research Council (NRC), the operational arm of the National Academy of Sciences, which at that time incorporated only the fields of psychology and anthropology, the disciplines bearing closest affinity with the natural sciences. The SRC might have failed were it not for Beardsley Ruml, newly appointed director of the Laura Spelman Rockefeller Memorial and recent Ph.D. in psychology from the University of Chicago. By 1922 Ruml had $2 million per year to invest in developing a science of society, but he was uncertain that university-based disciplinary departments, or the associations *per se*, would progress rapidly in that direction. Ruml had endorsed Merriam's unsuccessful bid to be incorporated into the NRC and suggested that the SRC add 'Science' to its name (and thus it became the SSRC). Subsequently, Ruml and the Memorial began a steady flow of financial resources to the SSRC and through it, to university-based faculties to develop their methodological capacity and the volume of science-inspired research into the burgeoning set of social problems of the era. As a Chicago-educated behavioral scientist, Ruml found common cause with Merriam, Chicago faculty member. During this era, the SSRC functioned as the social science counterpart to the NRC and at the same time served as adjunct to the Memorial as a 'pass-through' foundation. Between 1923 and 1929 the Memorial distributed about $40 million for social research and advanced (quantitative) methodological training, much of it distributed through the SSRC (Bulmer and Bulmer 1981, Featherman 1994). Ruml later served briefly as dean of social science at the University of Chicago; and in 1929 the Memorial became the social science division of the Rockefeller Foundation.

As for the social sciences disciplines themselves, their progress as empirically based research fields can hardly be understood without acknowledging the catalytic importance of universities and their gradual transformation as centers for academic research as well as for their original mission of educating undergraduates (Geiger 1993). Here, too, the role of the foundations and private investment in creating this element of infrastructure in the early twentieth century cannot be overlooked, even though their support of universities was quite selective and targeted (Bulmer 2001). For example, the Rockefellers provided foundational financing for the University of Chicago and encouraged its evolution toward a research as well as a teaching university. While the modern research university was not to emerge in the USA until after WWII, universities like Chicago and Columbia were notable incubators of empirical social scientific inquiry much earlier than 1940. At Columbia, for example, a statistical approach in research began in 1890 under the influence of Richmond Mayo-Smith and Franklin Giddings and in the genre of its time was focused on local conditions and problems of urban life. Later, just prior to WWII, the Bureau of Applied Social Research and the leadership of Paul Lazarsfeld elevated Columbia's prominence in empirical social science and embraced national issues and scope as context for refining social theories and quantitative methodologies. Likewise, at the University of Chicago, the seminal empirical studies of W I Thomas on Polish immigrants to Chicago, published just after WWI, were to give rise to a 'Chicago School' of sociological research in 1915–35, under Robert E Park, Ernest Burgess, and William F Ogburn (Bulmer 1984) and to a nascent science of American politics after 1920, with leadership from Charles Merriam and his students Harold Gosnell and Harold Lasswell. However, notwithstanding these and a few other notable examples (e.g., the University of Wisconsin in economics and political economy; the Institute for Research in Social Science at the University of North Carolina, founded in 1924), the presence of social science departments in teachers colleges and teaching universities of the period was quite modest and not given to encouraging research. The notable exceptions usually were tied to investments by foundations and private philanthropy (Bulmer 2001).

The incubation of the social science disciplines within their university base in the three decades prior to WWII yielded several important outcomes. At a

few universities, especially those with graduate schools, research groups provided context and exemplars for a more cumulative development of these young fields as academic disciplines. They increasingly asserted claim to intellectual legitimacy in their own right and not simply as means to the end of social reform. And they evolved as more differentiated (one from the other) disciplines with their own basic concepts and methodological approaches, modeled more or less upon positivist natural science. Prior to the early decades of the twentieth century, and without much infrastructure to sustain it, signal events in the development of the social sciences were episodic and failed to cumulate (Bulmer 2001). For example, the American Social Science Association, created in 1865, was the precursor to the American Economic Association and also the National Conference of Charities and Corrections. Its failure to provide sustenance for a social 'scientism'and its ultimate disappearance doubtless stemmed from the heterogeneous mixing of proto-scientists and social reformers and the related inability to resolve internal tensions between the advancement of social inquiry toward science and the promulgation of a reform agenda (Furner 1975, Haskell 1977). Later, organizations such as the NBER and the Brookings Institution offered a different kind of continuity and cumulation. Their role in effect transformed the origins of proto-social science as a tool in the hands of the reform minded into what, after WWII, would give claim to be objective scientific expertise provided to the world of public administration and public policy (Prewitt 1995). It was the few universities with graduate schools and active research programs, however, that fostered the growth and independence of disciplinary departments of academic social science, intellectual projects worthy in their own right (Shils 1979). This, in turn, was associated with a professionalization of social scientists according to discipline, leading to a disassociation from the reform-minded and social planners.

These university-fostered developments during the foundational period should not be overstated, however. The cadre of social scientists within academe was small and the effects of key leaders of the period—Mitchell in economics, James B Watson in psychology, Merriam in political science, and Ogburn in sociology—were pivotal. And whether within academe or other organizations like the NBER and the Brookings Institution, the practice and refinement of a more scientific social inquiry became a goal in its own right. Toward the end of this pre-WWII period, social science research meant pursuing intellectual curiosity, constructing propositions about the phenomenal social world and behavior, and measuring constructs and collecting data to verify those propositions. 'Social scientists began to construct a naturalistic social science as an end in itself, and under the influence of instrumental positivism, erected positivist scientific method into the chief standard of inquiry' (Ross 1991 pp. 467–8, as cited in Bulmer 2001). Nevertheless, the legacy of the Settlement House Movement, of Hull House itself, was to inspire the development of America's first graduate schools and after WWII, schools of public administration and policy at some of the nation's most eminent research universities with strong positivist social science traditions. One could argue that in the roles American social and behavioral science played and sought to play in the public policy arena during the second half of the twentieth century, academic scientists never fully disassociated themselves from advocacy and the taking of normative stances (Lynn 2001, compare Prewitt 1995).

Throughout the foundational period, however, the government of the USA played a rather insignificant role in creating the early infrastructure of the social sciences. Most government existed at state and local levels; the federal administrative state grew in size and political importance only after the great depression and WWII. While the USA census dates from 1790, the US Bureau of the Census was created as a permanent federal agency in 1902. The first commissioner of the US Bureau of Labor Statistics, the statistician Carroll Wright, was recruited from Massachusetts, where he had created the first such statistical bureau in the USA a decade after the civil war. Other elements of a federal statistical system emerged subsequently (see *Statistical Systems: Censuses of Population*; *Statistical Systems: Labor*; *Statistical Systems: Health*; *Statistical Systems: Education*; *Statistical Systems: Crime*, and *Statistical Systems: Economic*).

Events like wars and economic depressions, and governmental response to them, have channeled the evolution of the social sciences during the foundational period. For example, WWI mobilization of American men and their screening for fitness and assignment gave rise to nonexperimental but nonetheless quantitative psychology, to the psychometric assessment of personality and mental abilities (drawing upon European precursors). This war effort spawned basic research on individual differences but also an 'applied industry' of psychological and educational testing (Carson 2001).

Perhaps the signal contribution of government to the social sciences in the foundational period was President Herbert Hoover's appointment in 1929 of a blue ribbon research committee to assess trends in American society, especially in the aftermath of the stock market crash. Hoover, an engineer by training, believed in using scientific information to make better management decisions and to solve problems. The committee of leading academic social scientists (chaired by NBER economist Wesley Mitchell) was funded by Rockefeller monies, not by the federal government, and the research was staffed at the private SSRC in New York City. Nevertheless, on the eve of what was to be known as the great depression, this committee afforded national visibility and presidential

legitimacy to the nascent science of society that the foundations had sought to create. William Ogburn, Chicago sociologist headed the SSRC research effort, and the committee filed its multivolume report, *Recent Social Trends in the United States* in 1933, after Hoover had left office. The report took the view commonly accepted at the time: if facts and trends about households, towns, and institutions could be discerned clearly and measured precisely, then surely rational decision-makers would use the information in the enlightened public interest. This view contrasted sharply with the orientation of the Social Survey Movement earlier in the century, when data were collected directly in the service of correcting urban problems (Bulmer 2001). However, the report was widely criticized. Within elite academic circles as represented by leadership within the SSRC, deep reservations emerged about the potential distorting effects of association with political agenda setting, especially the compromise of scientific objectivity and of basic intellectual inquiry. In the emerging federal policy world of the New Deal era, the report was dismissed for its lack of politically informed analysis and interpretative prescription: the facts could not and would not speak for themselves. What administrative decision-makers of the emerging federal state sought was knowledge for action in a political arena and tested against the requirements of political decision-making (rather than against the requirements of academic social science professionals seeking legitimacy through objectivity; Lynn 2001).

This dialectic between scientific objectivity and prescription was to become a continuing theme in the post-WWII evolution of social science and its infrastructure. The dialectic represented contrasting points of view of the raison d'être for social science, of its legitimacy. And as the role of federal government in creating additional infrastructure and public funding for social science increased, ultimately displacing much if not all of the role played by the foundations and private philanthropy, the dialectic punctuated the evolution of these fields of science and of public support of them.

2.2 Period of Punctuated Evolution

The great depression and the New Deal and subsequently American participation in WWII contributed to growth in the size and political scope of the federal administrative state. Scientists and engineers, as well as their social and behavioral counterparts, contributed to the war effort in a wide variety of federal agencies. Within the Department of Agriculture, for example, studies of the war's impact on the farm economy and households gave rise to more refined survey research and its technology, especially probability sampling. Because science and technology were so instrumental to the war effort, federal expenditures on research and development (R&D) rose from $100 million in 1940 to $1.5 billion in 1945 (Featherman and Vinovskis 2001). The leadership of Vannevar Bush in the newly created Office of Scientific Research and Development was pivotal in this expansion of federal R&D. Bush also forged links between eminent scientists in the few research universities of the time and the needs of federal agencies and of the war effort overall. Thus, WWII created the funding infrastructure for a lasting partnership between the federal government and the universities and stimulated the fluorescence of the research university.

The effectiveness of this university–government partnership during the war prompted President Truman and Congress to establish the National Science Foundation (NSF) in 1950. The size of a role for federal support of civilian R&D, especially for the support of basic research during peacetime, was debated. And the place of the social sciences—viewed at best as 'applied' and not 'basic' science—within the mission of the NSF was heavily contested. Expenditures for these fields represented a mere 1.2 percent of the NSF budget in 1956 and grew only to 1.6 percent ($890,000) by the end of NSF's first decade (Larsen 1992). Additional support for social and behavioral research emerged in an expanding National Institute for Health (NIH) in the postwar years, although as a federal 'mission' agency, the NIH focused its resources on biomedical and other health-related science. Indeed the pattern of scattering federal support for science across such mission agencies, and the NSF for more basic research, has impeded a consolidated and centralized federal science policy (and corresponding decentralized decisions about science funding) since the close of WWII. Funding for the social sciences came less from federal sources than from the foundations, but on a steadily declining basis after mid-twentieth century. Whereas the Rockefeller and Carnegie Foundations were prime supporters of social science before WWII, the newly created Ford Foundation was prominent in the 1950s. For example, over this period, the Behavioral Sciences Program at the Ford Foundation dispensed nearly $43 million for social research and advanced methodological training in universities (Featherman and Vinovskis 2001).

During the same postwar period, economists (most prominently among social scientists) were recruited into new elements of an expanding Executive Branch. For example, the President's Council of Economic Advisors was established in 1946. But the most rapid expansion of the federal administrative state occurred in the 1960s during the Kennedy and Johnson presidential administrations. The federal government became one of the largest employers of social and behavioral scientists (Featherman and Vinovskis 2001). This was a period of jarring racial turmoil and wide public perception of unsolved and longstanding social problems of crisis proportion, such as poverty

amid plenty. The Great Society initiatives of the Johnson presidency, and large federal social welfare expenditures (over $60 billion in 1970, a 142 percent increase over the decade), drew upon social science research for technical guidance. At the NSF, a Social Science Division was created, expanding somewhat the legitimacy of these fields. Similarly, Congress created a National Institute of Education in 1973, and it promoted a broad agenda of research and of evaluation of school reforms. Correspondingly, support for social research blossomed from $384 million in 1961–2 to $803 million in 1966–8, with half coming from federal mission-oriented agencies (Featherman and Vinovskis 2001). And also in this period, both academic and more policy oriented social scientists joined faculties of programs and schools of public policy at a growing number of research universities. Thus, as with WWII, the so-called War on Poverty and associated aspects of the Great Society added to an expanding infrastructure, its funding, and its personnel base for social and behavioral sciences.

If the 1960s and early 1970s constituted something of a 'golden era' for social science, its infrastructure, and the role of its practitioners in policy circles, the era did not last long and ended for many in disillusionment. Large-scale quasi-experiments such as Project Head Start and the Negative Income Tax experiment, conceived and designed by social scientists in league with public policy officials, failed to demonstrate large or lasting impacts (White and Phillips 2001, Gueron 2001).

Solutions to social problems proved far fewer, more intractable, than promises; and social scientists argued openly among themselves about seemingly arcane (to politicians and the public) reasons for failure (Aaron 1978). Politicians and the public gradually came to the view that public monies invested in social research were irrelevant if not wasteful. Meanwhile, another war, this one in Southeast Asia, deeply divided the American public during the Johnson and Nixon presidencies. Academics in general, and social scientists within universities in particular, tended to oppose this war. From the perspective of the more politically conservative administrations of Presidents Nixon and Reagan, the anti-Vietnam War had so politicized and radicalized the professoriate that research scholars, especially in the social sciences, could no longer be trusted to provide objective research and evaluations (Featherman and Vinovskis 2001). By the mid-1980s federal funding for the social sciences had been cut back draconically.

In the closing two decades of the twentieth century, few signal events affected basic infrastructure of the social sciences. As if as a bellwether for social science in these times, however, two gradual developments are noteworthy. Following the Watergate episode and resignation of President Nixon, Americans grew ever more cynical and distrustful of government in Washington, and government itself, especially during the Nixon and Reagan presidencies. Academic social scientists, perhaps with the exception of economists and lawyers, retreated to universities to undertake more theoretically inspired research. Those with a policy-orientation either joined expanding faculties in schools of public policy or were recruited into a rapidly proliferating set of partisan think tanks (both liberal and conservative) that ringed the Capitol to provide politically inspired analysis and commentary to their respective constituencies. Importantly, the abundant infrastructure of the golden era, especially in the form of repeated cross-sectional surveys (e.g., the Current Population Surveys; the National Election Studies), longitudinal studies (e.g., Panel Study of Income Dynamics; National Longitudinal Surveys of Youth), and quasi-experiments or intervention studies (i.e., Gueron 2001) provided common databases for both academic and policy inspired research and for often fierce partisan debates over issues such as welfare reform (e.g., Danziger 2001) (see *Databases, Core: Anthropology and Human Relations Area Files (HRAF)*; *Databases, Core: Anthropology and Museums*; *Databases, Core: Demography and Registers*; *Databases, Core: Demography and Genealogies*; *Database, Core: Linguistics and CHILDES*; *Databases, Core: Linguistics, Online Grammars, and Dictionaries*; *Databases, Core: Political Science and Political Behavior*; *Databases, Core: Sociology*; *Microdatabases: Economic*). For some social scientists, the century-long search for relevance as honest research brokers telling truth to power (policymakers) ended in confusion or compromise (Lynn 2001), while others withdrew to develop theory and 'basic' research within the academic cloister.

A second development concerned the NSF, to some extent the NIH, and funding for social and behavioral science. Full directorate status for the social, behavioral, and economic sciences was created at the NSF in the early 1990s, after 40 years of contested status as a basic science and of politically vulnerability as either 'social engineering' or 'reformers with socialist politics.' For the first time, this placed the social sciences on an organizational par with physics and engineering, for example. Similarly at the NIH, an Office of Behavioral Research was opened with a direct reporting line to the Director. These achievements owed much to the intellectual and political lobbying of the Consortium of Social Science Associations (COSSA), the political arm of the professional social science associations based in Washington.

Funding for the social sciences once again increased, although not dramatically in constant-dollar terms and always reflecting growth or decline in the budgets of federal mission-agencies. Furthermore, the stability of Congressional appropriations to these social and behavioral units within the NSF, especially, and the NIH, to far lesser extent, could never be assumed. Congressional appropriation and authorization hearings routinely recorded doubts about, if not fun-

damental attacks upon, the worthiness and legitimacy of the social sciences even as others emphasized their indispensability (COSSA 1999).

2.3 'Y2K' and Beyond

In the dawning years of the twenty-first century, American social science and its infrastructure face an international challenge. Once an unchallenged exporter of social science theory and methods to the world, often in the form of graduate students from abroad who received advanced education and doctoral training in the USA, American social science is joined by robust science planning and research proliferation in Europe and other continents. While social science infrastructure in the USA has many international components (see *International Science: Organizations and Associations*; *Microdatabases: Economic*; *Centers for Advanced Study: International/Interdisciplinary*; *Data Archives: International*), it remains arguable whether this infrastructure is as internationalized as within Europe, for example. There the Standing Committee for the Social Sciences within the European Science Foundation, and successive science protocols emanating from the European Union, take explicit account and fund cross-national projects and infrastructure elements, like data archives and advanced computational facilities on a shared, international basis. These approaches mimic procedures in the 'hard' sciences and engineering, which have been internationalized in the USA and throughout Europe for decades.

Social science in the USA remains robust, a legacy of an abundant infrastructure of the twentieth century (i.e., Deutsch et al. 1971, Gerstein et al. 1988). However, as social science production and training grows more multicentered and less heavily dominated by the USA, as opportunities expand for social and behavioral scientists worldwide to engage in collaborative, cross-national projects with biological and physical scientists about environmental and health-related issues, the infrastructure to facilitate this internationalization becomes important to US participation. At the moment, foreign language training for US social scientists is uncommon and suboptimal, even if English is the language of science. The USA lacks an international science conversation, among social and behavioral scientists, that parallels the continental conversation within Europe about infrastructure or about intellectual strategies for context-sensitive comparative research (see *Infrastructure: Social/Behavioral Research (Western Europe)*). In a century that is likely to require the mobilized resources of social scientists worldwide to tackle the new issues of globalization, as well as the abiding problems of human civilization across locales and regions, US social scientists may need to pay greater heed to internationalization and develop new elements of infrastructure.

Bibliography

Aaron H J 1978 *Politics and the Professors: The Great Society in Perspective*. Brookings Institute, Washington, DC

Bulmer M 1984 *The Chicago School of Sociology: Institutionalisation, Diversity and the Rise of Sociological Research*. University of Chicago Press, Chicago

Bulmer M 2001 Knowledge for the public good: The emergence of social sciences and social reform in late nineteenth and twentieth century America, 1880–1940. In: Featherman D L, Vinovskis M A (eds.) *Social Science and Policymaking: A Search for Relevance in the Twentieth Century*. University of Michigan Press, Ann Arbor, MI

Bulmer M, Bales K, Sklar K (eds.) 1991 *The Social Survey in Historical Perspective, 1880–1940*. Cambridge University Press, Cambridge, UK

Bulmer M, Bulmer J 1981 Philanthropy and social science in the 1920s: Beardsley Ruml and the Laura Spelman Rockefeller Memorial, 1922–1929. *Minerva* **19**: 347–407

Carson J 2001 A cultural history of intelligence. In: Porter R, Ross D (eds.) *The Cambridge History of Science: Modern Social and Behavioral Sciences*. Cambridge University Press, New York, Vol. 7

Commission on Behavioral and Social Sciences and Education (CBASSE) 1998 *Investing in Research Infrastructure in the Behavioral and Social Sciences*. National Academy Press, Washington, DC

Consortium of Social Science Associations (COSSA) (1999). Senate colloquy discusses social and behavioral sciences at NSF. *COSSA Washington Update* **18**(18): 1

Converse J 1987 *Survey Research in the United States: Roots and Emergence 1890–1960*. University of California Press, Berkeley, CA

Danziger S 2001 Welfare reform policy from Nixon to Clinton: What role for social science? In: Featherman D L, Vinovskis M A (eds.) *Social Science and Policymaking: A Search for Relevance in the Twentieth Century*. University of Michigan Press, Ann Arbor, MI

Deutsch K W, Platt J, Senghaas D 1971 Conditions favoring major advances in social science. *Science* **171**: 450–59

Featherman D L 1994 SSRC, then and now: A commentary on a recent historical analysis. Items. *Social Science Research Council* **48**(March): 15–22

Featherman D L, Vinovskis M A 2001 Growth and use of social and behavioral sciences in the federal government since World War II. In: Featherman D L, Vinovskis M A (eds.) *Social Science and Policymaking: A Search for Relevance in the Twentieth Century*. University of Michigan Press, Ann Arbor, MI

Fisher D 1993 *Fundamental Development of the Social Sciences: Rockefeller Philanthropy and the United States Social Science Research Council*. University of Michigan Press, Ann Arbor, MI

Furner M O 1975 *Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science 1865–1905*. University of Kentucky Press, Lexington, KY

Geiger R L 1993 *Research and Relevant Knowledge: American Research Universities since World War II*. Oxford University Press, New York

Gerstein D R, Luce R D, Smelser N J, Sperlich S (eds.) 1988 *The Behavioral and Social Sciences: Achievements and Opportunities*. National Academy Press, Washington, DC

Gueron J M 2001 Welfare reform at the state level: The role of social experiments and demonstrations. In: Featherman D L, Vinovskis M A (eds.) *Social Science and Policymaking: A search for Relevance in the Twentieth Century*. University of Michigan Press, Ann Arbor, MI

Haskell T L 1977 *The Emergence of Professional Social Science: The American Social Science Association and the Nineteenth-Century Crisis of Authority*. University of Illinois Press, Urbana, IL

Lagemann E C 1989 *The Politics of Knowledge: The Carnegie Corporation, Philanthropy and Public Policy*. Wesleyen University Press, Middletown, CT

Larsen O 1992 *Milestones and Millstones: Social Science at the National Science Foundation, 1945–1991*. Transaction Press, New Brunswick, Canada

Lynn L E Jr 2001 The making and analysis of public policy: A perspective on the role of social science. In: Featherman D L, Vinovskis M A (eds.) *Social Science and Policymaking: A Search for Relevance in the Twentieth Century*. University of Michigan Press, Ann Arbor, MI

Prewitt K 1995 Social sciences and private philanthropy: The quest for social relevance. *Essays in Philanthropy* (Indiana University Center on Philanthropy), no. 15

Ross D 1991 *The Origin of American Social Science*. Cambridge University Press, Cambridge, UK

Shils E 1979 The order of learning in the United States: The ascendancy of the university. In: Oleson A, Voss J (eds.) *The Organization of Knowledge in Modern America, 1860–1920*. Johns Hopkins University Press, Baltimore, MD, pp. 19–50

White S H, Phillips D A 2001 Designing head start: Roles played by developmental psychologists. In: Featherman D L, Vinovskis M A (eds.) *Social Science and Policymaking: A Search for Relevance in the Twentieth Century*. University of Michigan Press, Ann Arbor, MI

D. L. Featherman

Infrastructure: Social/Behavioral Research (Western Europe)

1. *Development after 1945*

Modern social science, having originated in Europe, emerged as an institutionalized activity in research and teaching only after—and strongly influenced by—its powerful development in the United States. Its evolution in the twentieth century bears the imprint of the authoritarian and totalitarian regimes in continental Europe. While emigrant scholars contributed greatly to the development of the social sciences in the US, the scientific evolution in continental Europe became largely interrupted—more so in Germany and also in Mediterranean countries, and, to a lesser degree in France and Scandinavia. (The summary given in this chapter follows the 'four phases of development in the social sciences' as delineated by Martinotti 1999, pp. 87–91; for France, Italy, and Germany see also Wagner 1990.)

By the end of the Second World War, the social sciences hardly played any role in the European universities. For example, in sociology there were only three chairs in France and none in Italy. In West Germany, political science chairs and a disciplinary association did not exist until 1950. To go to the US and study at an American university for some time, therefore, became an almost necessary precondition for the further education and training of social scientists in Europe. Experiencing 'modern society' and learning the new skills of its systematic observation, made a particular impression on young scholars from countries like Germany, Italy, Spain, and Greece which had been isolated from modern culture for decades. But it is also true of the Scandinavian countries, where the proximity of the Anglo-Saxon world favoured cultural transfer.

As the postwar economic recovery progressed, it became generally understood that the social sciences were part of the modernization of Europe, seen as the expansion of the capitalist mode of production as well as the democratization of society, both closely linked with the American model. While this caused conflicts and resistance from traditional cultural elites and the Marxist-oriented intelligentsia, there was an increasing diffusion of quantitative methods of data collection, particularly survey and primary data. While it is true to say that the European social science tradition had its own deep experience in quantitative analysis, one must also recognize that empirical research had mainly been based on the secondary analysis of 'process-produced data,' particularly statistical data collected by public authorities. What was missing in the continental tradition, and particularly in those countries that had been under despotic regimes, was the experience of field research and the collection of primary data (Martinotti 1999, p. 89). Resistance to the introduction of these new instruments, particularly survey methods and sampling techniques, became expressed from different points of view of traditional European scholarship as well as from administrative agencies with their customary statistical procedures.

Under these circumstances it was not before about 1960 that one can speak of a major expansion of research and acceleration of its academic institutionalization. European societies went through a period of economic 'upswing' that made large masses move to the metropolitan centres of industrial development. The postwar generation of young people, mobilized with the rapid expansion of educational systems, grew up as part of a mass culture, reinforced by the media, that provided a setting for the diffusion of ideas, lifestyles and new organizational patterns without precedent in the European experience. The social sciences, particularly the younger and newly built-up ones, notably sociology, became highly attractive to the student generation for a number of reasons: the relatively young age of university personnel, compared with more established disciplines; the less precise definition of the academic institutionalization of the

social sciences; and the proximity of their subject matter to the intellectual and existential preoccupation of the newly mobilized social groups.

From this background there emerged, in all European countries, a demand for theories of political relevance and of significance for diagnosing and overcoming societal problem constellations (see Dierkes and Wagner 1992, pp. 612–15). This interest, with its commitment to change, stood in opposition to the program of a supposedly ahistorical and value-free explanation of social phenomena. Intellectual controversies and debates within the disciplines pitted the American social science tradition against a variety of Marxist approaches and other 'critical' schools with their supposedly more 'enlightened' responses to the demand for a more encompassing understanding and a more 'committed' interpretation of social facts.

With this orientation, major streams of the social sciences, with Paris and Frankfurt as major centres, acquired a somewhat militant role, linking up with new social movements and entering into debates about the engagement of intellectuals. The traditional social structures and actors, however, managed to retain their overall influence and to uphold traditional explanations for problems of economic and social change. Social science research in much of the 1960s and 1970s can therefore be characterized, on the one hand, by innovative work in many areas, e.g., political participation, labor organization, class structure, the 'capitalist state' and problems of personal and sexual relations. On the other hand, theoretical innovation was limited and research perspectives remained concentrated on the nation-state—despite a growing awareness of conflicts posed by North–South differences, the role of multinational corporations, or global environmental problems.

It was in the later part of the 1970s and in the 1980s that the turmoil of debates, the imbalances between issues and approaches, the dividing lines between theoretical and empirical research gradually became superseded by an ongoing process of professionalization, of methodological consciousness and of new theoretical directions. In retrospect, we can speak of a phase of consolidation in the social sciences in the period of economic recovery after the oil crisis. At a time when liberal social philosophies became dominant, with political sentiments tending to be anti-welfare, and economic policies being influenced by monetarism, the overriding mood seemed to run against the interest in social conditions and the study of society. While this indeed in some countries, notably the UK under Prime Minister Thatcher, made for severe cuts and structural changes, the science policy pendulum did not swing to similar extremes in continental Europe. Debates over the social sciences gradually shifted from theoretical controversies to discussions of their 'uses' in public affairs and policy-making. But by and large institutional growth continued, then slowing down to reach a steady state on advanced levels. (See Martinotti 1999, p. 90, Wagner 1999, pp. 29, 39.)

Towards the end of the 1980s, the social sciences had become well-established academic disciplines, and their presence in policy-making had been well advanced. The academic expansion had very much been related to, and benefited from, expectations with regard to practical uses of the social sciences in the process of societal modernization which had become an important objective, actively pursued by parts of the political and economic elites. In the process, the role of intellectuals became redefined, turning from persons reflecting in general terms about the fate of humankind in history and society to experts for matters of steering society and the economy, and providing knowledge usable as a tool in reformist policy interventions. After the end of the high expectations about the 'scientification of politics' allegedly leading to the 'end of ideology,' social research had become a routine element of many policy processes and, more recently, of management in private companies as well (see Wagner 1990, pp. 441ff.).

The increasing specialization of the applied social sciences, as well as the general trends of the professionalization of the disciplines and their institutionalization in universities and research centres, brought about an increasing fragmentation and specialization which has sometimes been related to a tendency of 'deconstructivism' with critical attacks on large theoretical systems and linked to postmodernist movements. But these tendencies more often are understood as part of a more general search for theories capable of combining macrotrends with microsocial and individual dynamics, and, in so doing, aiming at overcoming the traditional juxtaposition of qualitative and quantitative methods, looking for integration in unified research designs (see Martinotti 1999, p. 90).

Having developed to this advanced stage, the social sciences in the 1990s had to reorient themselves under two dramatic changes of focus: first, the end of the East–West opposition, and second, the process of globalization and internationalization in world regions, particularly the construction of the European entity.

There is an ever more strongly articulated need to gain a better understanding of and to contribute to problem-solving in the process of European integration. Developing a concept of European society has become a major challenge for the social sciences which have traditionally evolved around the concept of a nation-state and of national societies under the stresses imposed by industrialization. Only slowly, but with increasing intensity around the turn of the new century, has it become recognized that the tools and working conditions of social research need to be made adequate for the study of new problem constellations and supra-national processes, and that social science policy-making and research programs have to be

oriented towards cross-national comparison and international cooperation, particularly on the European level. (See Dierkes and Biervert 1992, Kaase et al. 1997, Erikson 1999.)

2. *National Environments and Institutional Structures*

At the beginning of the twenty-first century it is fair to say that in the Western European countries, as a whole, the social sciences have become a standard element in higher education, in public debate and media coverage, in policy-making, administration and corporate management. While, generally speaking, the knowledge of social reality and the diffusion of social facts have become vastly expanded in the previous five decades, the infrastructural basis for generating this knowledge has developed differently in the individual countries. The disciplinary boundaries of the social sciences, research approaches, problem orientations, and structural differentiations in the build-up of research facilities have become shaped by the respective national environments with their historical preoccupations, their academic-scientific 'cultures,' their political and administrative traditions. Expectations regarding the usefulness of social science knowledge for policy-making and the periodically strong demand for social science expertise have contributed to the institutional consolidation and at times—in varying degrees from country to country—to a push for new research structures reflecting characteristic orientations, problem constellations, and modes of public discourse. (See Wagner et al. 1991, Neidhardt and Smelser 1992.)

For example, the social sciences in Sweden may, to a considerable degree, be characterized by pointing, on the one hand, to a traditional demand for exact and reliable methods, going back to early studies of migration and social movements, and by referring, on the other hand, to the interaction with social policies in the Swedish welfare state model and thus to particular forms of contributing to problem-solving as well as, in the process, refining the theoretical outlook on society.

Contemporary social science orientations may often be explained in terms of specific historical evolutions. In the Netherlands, for example, there had been an early involvement of social science research in the practical problems associated with the territorial aspects of social planning, and the social sciences had become shaped in a continuous development by the specifically Dutch version of political conflict resolution, the *pacificatie*, and the corresponding development of institutional networks for interest mediation. While the forms of policy advisory work and research contracting in the Netherlands can thus be characterized in terms of pragmatism, tolerance, and depoliticization, it is quite a different picture that one would paint for Austria. Here, in the prewar period, the constellations of political conflict did not allow for the continuity of similar pragmatic approaches, and the modes of conflict resolution established after the war led to the involvement of social science approaches in the specifically Austrian type of corporatism, making for quite a different variant of a more general continental European phenomenon: the integration of social science research in institutional structures of political representation and in channels of communication between politics, public opinion, policy-making and scientific expertise (see Wagner 1990, pp. 460–3).

These examples, while meant to underline the importance of the diverse national settings, may also serve to illustrate a more general thesis explaining social science orientations with reference to historical paths of social and political development. In these terms, the social sciences emerged in Western European countries in a markedly different way from the way they evolved in England and the US, where the continuity of individualism and civil society made for a more pragmatic style of studying social and political problems empirically and with reformist perspectives. The social sciences on the European Continent, in turn, have been characterized by addressing more basically the foundations of a social and political order which—given the traditional social conflicts, ideological cleavages and propensities for radical change in Europe—have time and again been called into question. They have also shown a greater tendency to ascertain their basic categories and critically discuss their knowledge base.

Institutionally, in all Western European countries, the major part of social science research takes place within the university system. This refers first—and for most countries, foremost—to a multitude of small production units attached to professorial chairs in university departments and loosely organized faculties. Depending on the relative number of students and the weight of the professorial teaching load, university research consists mainly of small projects done with students and often linked to their theses.

At many universities we find, under similar names, institutes for social science research as organizational entities meant to provide better opportunities for cross-disciplinary orientation, for cooperating in empirical research and for contracting and administering larger projects with outside funding. But it is only at very few places that such additional forms of institutionalization, in combination with a considerable number of professorial and other staff positions in the respective departments, have led to larger and more consolidated university research centres. In Germany such larger agglomerations of the social sciences exist at the universities of Mannheim, Cologne, Bielefeld, and Munich. (See Neidhardt 1999.)

By and large, however, it is outside the universities that more extensive research structures have become institutionalized. Extra-university institutes form a

second element in the 'academic research system' in as much as they are to some degree publicly funded and not dependent in their agenda-setting on short-term opportunities offered by the markets of policy analysis and corporate consulting, but rather able to integrate individual projects into larger programs. While indeed at times working on research tasks aimed at contributing to problem-solving, such establishments usually understand themselves as knowledge-oriented and 'science-driven,' and are—at different levels—connected to the universities. (See Wagner 1990, p. 441ff.)

These institutes vary from country to country in number, size, and concept, depending on the respective national environment. In some cases there is only one such extra-university institute as a central establishment, like the Greek National Centre for Social Science Research (EKKE) in Athens. The German case demonstrates a functional differentiation in two directions which, for structural reasons, did not easily develop within the universities and thus called for extra-university institutionalization: on the one hand, the *Wissenschaftszentrum Berlin für Sozialforschung* (*WZB*—Social Science Research Centre, Berlin) and some institutes of the *Max-Planck-Gesellschaft* (*MPG*) and of the *Leibniz-Gemeinschaft* (*WGL*) which undertake larger, long-term, internationally comparative or problem-oriented projects in transdisciplinary research groups; on the other hand *GESIS*, a group of social science infrastructure establishments in Mannheim, Cologne and Bonn, which provide professional services in support of social science research by documenting research projects and their results, by advising researchers on project design, by offering courses in advanced methodologies, and by archiving and giving access to empirical data for the purpose of secondary analysis.

Compared with the grand total of resources and activities in the universities, the 'extra-university' sector of the academic research system is still rather small in Germany and even smaller in Italy with only a few institutions of the *Consiglio Nazionale delle Ricerche* (*CNR*—National Research Council). In France, however, this sector is considerably larger, due to the historic separation of the universities from the centres of academic research. The dominant institution is the *Centre National de la Recherche Scientifique* (*CNRS*) which runs a large number of research institutes separately from or in cooperation with universities or *grandes écoles*; for sociology and ethnology, the *CNRS* with its division on the *Sciences de l'Homme et de la Societé* is far more important as a research institution than the universities. Furthermore, there are the *Maison des Sciences de l'Homme* (*MSH*), the *Ecole des Hautes Etudes en Sciences Sociales* (*EHESS*), the *Fondation Nationale des Sciences Politiques*, the social science sections of the Paris Museums, and the *Collège de France*. (See Machin in Lisle et al. 1984.)

The basic institutional set-up of the academic research system is—in varying forms of composition—being supported and complemented by a third institutional element: the system of project funding through national organizations and foundations. To take the German example again, the *Deutsche Forschungsgemeinschaft* (*DFG*), a publicly funded association of institutions of higher education and research, provides good opportunities for good projects—in whatever kind of research design—to be financed, supplementing the existing institutional resources and using peer review procedures to evaluate project applications. Moreover, the *DFG* and, with greater emphasis, other funding organizations (like the *Volkswagen Stiftung*) make attempts to establish researcher networks and cooperative arrangements within larger, more integrated research programs.

In addition to these institutes and programs that are constituted on scientific grounds with a primary interest in knowledge generation, we find in all the Western European countries other types of social science institutions that are— again in different forms and in varying degrees of development—oriented towards practical needs of government agencies and private organizations. These are, first, publicly financed research institutions directly operated by the state and subject to directives from the respective ministry interested in having a particular need covered by research and analysis. As a second type there are privately organized institutes which may have some basic institutional budget, but receive the main part of their resources from contracts for work on projects commissioned by government agencies as well as corporations, unions, the media etc. The 'state-administrative' part of this sector, often with longer traditions, comprises statistical and agronomic analyses, research on land use and regional development, on demography, public health, on education and vocational training etc., usually with a strong empirical-quantitative orientation. Such 'contractual' research became more developed in the 1960s and 1970s, relating to the policy issues of the day, and sometimes developing a marked expertise through repeated studies over time in a given area.

In some nations and at certain periods in their development, this sector may gain considerable—even dominating—weight in the overall spectrum of social science research. (See Wagner 1990, pp. 443–7.) In the context of this chapter however, there is no room for elaborating, e.g., on the importance of the government-directed non-academic research sector in France, on the effects of administrative traditions and statistical-quantitative orientations, or on the channels for linking social science research to the system of 'planification' through, for example, the *Comité d'Organisation des Recherches Appliquées sur le Développement Economique et Social* (*CORDES*).

These different forms of applied research are indeed part of the institutional environment and the operating

modes of social science research in the respective national contexts. However, while they provide career paths for social scientists and demonstrate the public use of social science knowledge, their infrastructural functions remain rather restricted in comparison with the academic research system with its opportunities for research experiences in larger contexts, for methodological training, for internationally comparative ventures, and for exchanges between approaches and orientations.

3. *Strengthening the European Research Base*

Social science research in Western Europe, in summary, has become institutionalized in a variety of national settings and organizational structures, depending on each country's specific historical background and cultural contexts. International exchange and cooperation, first with the United States, then increasingly across Europe, have served to develop the social sciences to higher degrees of professionalization, with a growing interest in cross-national comparison, but have by far not yet come near to what might be called a 'European' social science. At the beginning of the twenty-first century the ongoing process of European integration had still not become adequately developed as a subject of research aiming at understanding the preconditions, the conduct and the consequences of this process. In addition, the European Union itself, its institutions and major policy areas require the increasing attention of the economic and social sciences, and it has become a matter of 'social science policy' concern to evaluate and redirect the structural conditions of social science research in the (Western) European countries and the perspectives of its development 'towards Europe.'

At the outset of such assessments, it has widely become recognized that the fragmentation of the Western European systems—in terms of social, political and economic problems as well as science and research structures—should not only be understood as obstacle and retardation on the path to a truly European outlook. Turning disadvantage to benefit, one can see European social sciences, with their subject matter differing to some extent from country to country, as providing a densely developed empirical basis for generalizations across time and space. With its richness of constitutional, institutional and cultural variations across (and also within) nation states, Europe has come to be interpreted as a 'natural laboratory' for the social sciences, leading to strong recommendations concerning systematic empirical comparative research. (See Kaase et al. 1997, pp. 3–4.)

While the social science infrastructures in the different European countries have so far been mainly directed towards the study of national developments and problem-solving dimensions, the past decades, with still limited success, have put forward the notion of an increasing need for international cooperation and the corresponding infrastructures that make for exchanges, collaborative work and joint projects. Network linkage has been advanced through the establishment of European journals in different disciplines, through a 'Network Scheme' by the *European Science Foundation* (*ESF*) and through European networks of research units. The most successful of these has been the European *Consortium of Political Research* (*ECPR*) with more than two hundred corporate members across Europe in the early years of the twenty-first century. The *ECPR* organizes annual joint sessions of workshops which take place in different countries and bring together about five hundred European political scientists. A similar organization, the *European Consortium for Sociological Research* (*ECSR*), exists for the neighboring discipline.

Meeting grounds for groups of researchers from different countries and opportunities for collaboration on joint projects over a period of time are being offered by institutes of advanced study, particularly the *Netherlands Institute for Advanced Study in the Humanities and Social Sciences* (*NIAS*) in Wassenaar and the *Swedish Collegium for Advanced Study in the Social Sciences* (*SCASSS*) in Uppsala as well as—with a more general orientation reaching beyond the social sciences—the *Wissenschaftskolleg zu Berlin* or the *Collegium Budapest*. The latter, together with other centers of excellence in Central and Eastern European countries, is meant to advance research capacities within its region by, not least, providing links with researchers from Western countries.

In the framework programmes of the European Union the social sciences have not, until the 1990s, played an explicit role other than having been eligible for the funding of networks of younger researchers as part of general 'Human Capital and Mobility' or 'Training and Mobility of Researchers' programs. The Fourth Framework Programme (1994–1998) did for the first time contain a separate program relating to the social sciences, 'Targeted Socio-Economic Research (TSER).' In addition to one part aimed at the evaluation of science and technology policy options in Europe, it included a part on research on education and training and one on research into social integration and social exclusion in Europe. The latter addressed processes of social exclusion and integration, causes of social exclusion, particularly unemployment, the impact of social integration policies, and research infrastructures. (For the EU research programs see *www.cordis.lu*.)

The EU Fifth Framework Programme (1999–2003) contains, as a major instrument, 'Key Action: Improving the Socio-Economic Knowledge Base,' directed at research tasks on the management of societal change, on individual and collective strategies in a changing society, on employment and unemployment and on social cohesion in Europe, on new perspectives

for learning, on governance, citizenship and the dynamics of European integration, and on the challenge of EU enlargement. In addition, the interest in social science research has become expressed in the proviso that all the other, mostly technology-related, EU programs must include social science research work in a proportion of at least five percent of their respective budgets.

For the researchers, participation in the European research programs has, depending on the background of their experiences in the respective national systems of research funding, meant a reorientation towards new requirements and procedures, sometimes deviating considerably from established modes of competitive research application and evaluation. Following critical discussions and corresponding adjustments of program definition, of evaluation criteria and financing mechanisms, criticism, partly even opposition, centered on the bureaucratic procedures, on the professional status and the scientific excellence brought to bear on program design and project assessment, and on the policy orientation of EU research. (For a general discussion of EU research structures and policies see Max-Planck-Gesellschaft 1994.) It remains to be seen whether the social science research task in the sixth and in subsequent EU framework programs will reflect a convergence between competing principles and complementarity between the research programs of the EU, on the one hand, and the ESF with the different national research councils and scientific organizations, on the other.

One item of discussion that may lead to a clarification of perspectives and differentiation of roles, is the desirable balance between basic and applied research in response to the 'European' dimensions of EU research tasks and their relatedness to EU policy areas. (As one position in the discussion, calling for more basic research, see Erikson 1999, p. 32.) One example is the expression of the overall objective of the program, 'Improving the Socio-Economic Knowledge Base,' namely:

> to improve our understanding of the structural changes taking place in the European society in order to identify ways of managing change and to involve European citizens more actively in shaping their own futures. This will involve the analysis of the main trends giving rise to these changes, the analysis of the relationships between technology, employment and society, the re-appraisal of participation mechanisms for collective action at all levels of governance and the elaboration of new development strategies fostering growth, employment and economic and social cohesion. (European Commission Publications; see www.cordis.lu)

Under this general headline, there has been a pronounced emphasis on the infrastructure for the intended European research as expressed in the preamble to the *TSER* program:

> Research … will require high priority for conceptual and methodological work on constructing and integrating data and indicator systems as a common effort of the European social science research community with the aim of becoming comparable Europe-wide and establishing a common research infrastructure (databases on the parties involved and research results; directories and manual glossaries and thesauruses, etc). This is thus not simply a matter of harmonising statistics but rather, it is creating a scientific system of social reporting which can contribute to the basic understanding of social and economic development. An appropriate proportion of the resources allocated to this field will therefore be earmarked for the definition of a strategy for carrying out this work and developing such infrastructure and to the associated pilot projects. (European Commission Publications; see www.cordis.lu)

The interest in 'data integration' and 'research infrastructure' has become reiterated in later policy statements and, with slow progress, been pursued by additional measures, parallel to the thematic calls of the research programs, and with a view to (and seemingly in competition with) *EUROSTAT*, the European Statistical Office. The same direction is followed by the inclusion of the social sciences in EU actions to support 'large-scale facilities' and by promoting 'access to infrastructures,' recently widening the notion of infrastructures from establishments for data processing and archiving to functions of opening up particular research experiences by coordinated international exchanges.

This interest meets with a core concern of the *ESF*, particularly its Standing Committee for the Social Sciences, which has been active, albeit with limited resources, not only in promoting thematic projects in international cooperation, but also in the discussion and recommendation of social science policies and infrastructural development. In this context it has prepared the proposal for a 'European Social Survey' which in a number of respect overlaps with, or runs parallel to, projects partly financed by the European Commission, to develop a 'European System of Social Reporting and Welfare Measurement,' with about twenty national teams, including some eastern European countries.

With the efforts to create a 'European System of Social Indicators,' emphasizing the scientific context and research dimensions of social reporting going beyond 'mere statistics,' an early interest of the European Community of the 1970s is being taken up again, very much in line with the corresponding activities of the *OECD*, following the strong signal given by the so-called Brooks report on 'Science, Growth, and Society' of 1971 with its explicit inclusion of social science research in science policy concepts.

In a report of 1997 for the *European Science and Technology Assembly (ESTAM)*, an advisory body to the EU, a panel of leading social scientists completed a survey of the strengths, gaps and weaknesses of the economic and social sciences in Europe. In this they recommended measures for the support and advancement of interdisciplinarity for problem-driven re-

search, particularly flexible organizational modes to link scholars from different, even distant, disciplines for a limited period of time; the comparability of data, especially longitudinal microdata, and the systematic linkages necessary for multilevel analyses; the research training of young scientists through graduate colleges, summer schools, etc.; incentives for independent work at earlier stages in academic careers; networks between funding organizations for granting money to comparative projects; national concentrations of methodological resources (similar to *ZUMA, the Zentrum für Umfragen, Methoden und Analysen*, in Mannheim) and links between them across Europe. (See Kaase et al. 1997, pp. 4–10.)

The lack of empirical data collected and analysed on the European level, and the need for methodological advancement and for developing a common European research and training agenda, have also been pronounced by Federico Mayor, the former director-general of UNESCO, following a series of conferences concentrated on European perspectives on the social sciences (see Mayor 1998). Opening the horizon of European integration to include the processes of postcommunist social transition, the notion of a social science 'laboratory' built on the diversity of national traditions may also be fruitful in a wider sense: the challenge would thus not only be to comparatively understand social, political, and economic structures in national contexts; it should also call for the study of intersocietal penetration and of arrangements that work towards international cooperation and integration between countries. With this outlook, a pluralist and transnational social science in Europe might also look for closer relations with social sciences in Asia, Africa, and Latin America, in addition to its profound ties with North America.

See also: Centers for Advanced Study: International/Interdisciplinary; Disciplines, History of, in the Social Sciences; Infrastructure Investment; Infrastructure: Social/Behavioral Research (United States); International Research: Programs and Databases; Quantification in the History of the Social Sciences; Universities, in the History of the Social Sciences

Bibliography

Dierkes M, Biervert B (eds.) 1992 *European Social Science in Transition: Assessment and Outlook*. Campus/Westview, Frankfurt/Boulder, CO

Dierkes M, Wagner P 1992 Conclusions and recommendations. In: Dierkes M, Biervert B (eds.) *European Social Science in Transition: Assessment and Outlook*. Campus/Westview, Frankfurt/Boulder, CO, pp. 611–37

Erikson R 1999 The prospects of European social science. In: *The Social Sciences at a Turning Point?* OECD Proceedings, Paris, pp. 31–5

Kaase M, Baumert J, Erikson R, Hellwig M, Martinotti G, Meidinger C, Ménard C, Newton K, Semin G 1997 *Survey of Strengths, Gaps and Weaknesses in European Science: Economic and Social Sciences*. Report for the European Science and Technology Assembly (ESTA) and for the European Heads of Research Councils (EUROHORCs). Deutsche Forschungsgemeinschaft, Bonn, Germany

Lisle E, Machin H, Yasin S 1984 *Traversing the Crisis: The Social Sciences in Britain and France*. Economic and Social Research Council, London

Martinotti G 1999 The recovery of Western European social sciences since 1945. In: *UNESCO World Social Science Report 1999*. UNESCO Publishing/Elsevier, Paris, pp. 84–91

Max-Planck-Gesellschaft 1994 *European Research Structures—Changes and Challenges: International Aspects of European Research Policy*. Max-Planck-Gesellschaft, München

Mayor F 1998 European perspectives on the social sciences. In: *Proceedings of the Second European Social Science Conference in Bratislava*. UNESCO, Paris

Neidhardt F 1999 Social science organization and policy issues in Germany. In: *The Social Sciences at a Turning Point?* OECD Proceedings, Paris, pp. 59–62

Neidhardt F, Smelser N J 1992 Sociology. In: Dierkes M, Biervert B (eds.) *European Social Science in Transition: Assessment and Outlook*. Campus/Westview, Frankfurt/Boulder, CO, pp. 244–67

OECD 1999 *The Social Sciences at a Turning Point?* OECD Proceedings, Paris

UNESCO 1999 *World Social Science Report*. UNESCO Publishing/Elsevier, Paris

Wagner P 1990 *Sozialwissenschaften und Staat: Frankreich, Italien, Deutschland 1870–1980*. Campus, Frankfurt, Germany

Wagner P 1999 The twentieth century—the century of the social sciences? In: *UNESCO World Social Science Report 1999*. UNESCO Publishing/Elsevier, Paris, pp. 16–41

Wagner P, Weiss C H, Wittrock B, Wollmann H (eds.) 1991 *Social Sciences and Modern States: National Experiences and Theoretical Crossroads*. Cambridge University Press, Cambridge, UK

G. Thurn and F. Neidhardt

Injuries and Accidents: Psychosocial Aspects

Everybody is at risk for incurring an *injury*, although some individuals or populations may have greater risk than others. Injuries are consistently among the contributors to disabilities and death, lost years of productivity and potential life at all ages, resulting in incalculable costs and pain to victims and families around the world (Barss et al. 1998). As the impact of infectious disease on death rates was lessened over time due to vaccines, antibiotics, and improved hygiene, unintentional injuries rose as the leading causes of death and disability, especially for the first four decades of human life. Among the leading causes of injury are the categories of motor vehicle passenger

and pedestrian, drowning, poisoning, falls, firearms, and suffocation (Baker et al. 1992). Numerous commissions and expert panels in the international community since the 1970s have issued reports indicting injuries as the most important public health problem (Her Majesty's Stationery Office 1992, Committee on Trauma Research et al. 1985, National Committee 1989, U.S. Public Health Service 1991, 2000, World Health Organization 1999).

1. Scientific Study of Injuries

Social scientists from disciplines of psychology, public health, and sociology, among others, emphasize the term 'unintentional injuries' because the historical and layperson use of the term 'accidental injury' implies the effect of uncontrollable chance, fate, or other capricious events that cannot be avoided. Professionals now believe that the circumstances resulting in injuries can be subjected to scientific study, and therefore injury outcomes are thought to be explicable and preventable (Roberts and Brooks 1987, Robertson 1998). Additionally, unintentional injuries should be distinguished from intentional harm such as child abuse, homicide, or suicide.

William Haddon, Jr., a physician, was the major contributor to the scientific study into the nature and control of injuries across the lifespan. Haddon's model organized the public health approach to examining the interactive aspects of injuries: the host, the environment, and the agent (Haddon and Baker 1981). For example, Haddon proposed comprehensive strategies for changing the deleterious transfer of injury-causing energy from agent to host. Such strategies include, for example, not creating hazardous producing energy in the first place to reduce injury potential, keeping separate energy sources from the host (i.e., the human victim) in either time or space, or developing the host's resistances to damage. Haddon's influence continues to be felt in this field (National Committee 1989, Robertson 1998). Because human behavior is also involved directly in injuries, science disciplines examining behavior contribute greatly to the understanding and controlling of injuries and death.

2. Methodologies to Study Injuries

Professionals studying injuries utilize several empirical methodologies to gain an understanding of the precursors to an injurious event, of the nature or circumstances of the event itself, and of the aftermath or consequences of an injury.

2.1 Epidemiology

The epidemiological method relies on population-wide statistics about the prevalence and incidence of injuries. These analyses are often specific to certain characteristics such as type and cause (e.g., burns, drowning, firearm wounds, or motor vehicle occupant), demographic variables (e.g., age, sex, socioeconomic class of the injured victim), and injury circumstances (e.g., where the injury occurred: street, house, school; how it occurred: fall, car collision, fire). These epidemiological analyses can convey some aspects of injury development in order to suggest where attention is needed to prevent or control injuries (Baker et al. 1992, Robertson 1998). Epidemiology studies have established that motor vehicles contribute to the most injuries of any cause, indicating that speed, road conditions, passenger seatbelt use, driver's alcohol consumption, for example, are related to these injuries and deaths. Different types of injuries vary in predominance at various ages. For children, a developmental relationship of behavior, ability, and environment has been noted: toddlers are injured in falls and ingestion of poisons; school children are hit by cars as pedestrians; adolescents are injured in car collisions as drivers or passengers, in sports activities, and by firearms. The elderly are injured in falls.

2.2 Behavior Analysis

Psychologists often utilize techniques of behavior analysis to study the antecedent—behavior—consequences sequential chain of events to determine how injuries are produced. For instance, some analyses implicate lack of parental supervision in certain children's injuries and impulsive darting into streets after lost balls. By understanding the behavioral circumstances, prevention focused on behavior change can be undertaken such as providing prompts to engage in a safety behavior or issuing rewards for taking safety actions (Roberts et al. 1987). Other psychologists conduct cognitive development research to determine how children and adults perceive hazards and understand warning labels in order to improve their safety actions or develop educational programs for injury prevention.

2.3 Personality Trait Research

Research into personality traits or relatively stable patterns of behavior leading to injuries has not consistently identified an 'accident prone' type, although some people appear more likely to be injured than others. Some sex and age characteristics of the injured victim often are identified (e.g., younger children, adolescents, males) as being at higher risk for certain types of injuries. Temperament type and activity level have been linked to greater injury risks. Contrary to this line of thinking, according to some scientists, identification of individual differences may

not lead to effective injury control (Robertson 1998). More universal precautions for all people will have greater impact in preventing injuries than selecting individuals with identified markers for an intervention effort.

3. Injury Control

Injury control (prevention of injury) seeks to avoid the occurrence of a potential injury by changing the conditions and the behavior leading to injury, or by reducing the degree of damage when an event occurs. Prevention interventions can target (a) the person who is at risk for injury (child or adult), (b) the authority who makes decisions affecting that person (e.g., a parent, teacher, employer), or (c) the environment to limit possibility of injuries (e.g., through legislation and regulation). Two important approaches to these targets of prevention are structural and behavioral interventions.

3.1 Environmental Modifications to Achieve Safety

Structural interventions are those that modify the environment in order to change potential hazards so that people do not encounter them. For example, fence barriers keep people out of swimming pools, abandoned mines, and quarries. Expandable door gates help keep young children from falling down stairs, while landscaped berms and elevated crosswalks separate pedestrians from street traffic. Covers and guards on moving parts in industrial machinery protect workers. Childproof caps on medicine bottles keep children from potential poisons. Another structural approach is to reduce the degree of injury from an unmodifiable hazard when a person is inadvertently exposed to it, such as automatic seat belts or airbags in cars when in a collision and sprinkler systems in buildings when a fire starts. Environmental change is also conceptualized as passive prevention in that injuries are prevented or reduced in degree without the individual having to take safety actions themselves. Legislation or regulation on consumer products is frequently required to achieve environmental change, such as limiting the size of small objects that could cause choking and suffocation in toys marketed for young children.

In some countries, agencies are given some responsibility to review the safety features of consumer products being marketed to the public. In the USA, the Consumer Product Safety Commission implements a number of regulatory acts with demonstrated effectiveness in improving children's safety. For example, legislative requirements to alter the environment of children to be safer include changing the packaging of poisons and medications such that the number of inadvertent poisonings was reduced; legislation for children's sleepwear to be flame retardant reduced burn injuries; crib and refrigerator safety regulations have reduced suffocation and strangulation deaths. This commission has also issued guidelines for construction of toys and places warnings on toys with small parts that might choke young children (these types of warnings have less evidence of effectiveness than the environmental interventions). Very few products are regulated or reviewed by governmental agencies, but the public erroneously believes only safe products can be sold. Structural changes are sometimes limited in the potential effectiveness when political will or power is not brought to bear on the issue. For example, because of political and car manufacturer resistance, over 30 years passed before automatic air bags were fully implemented in cars in the USA from the date they were first mandated. Other legislation mandating safety have been similarly resisted by those affected (e.g., mandatory motorcycle helmet laws, legal restrictions on handguns, speed limit laws). Indeed, while some countries have little or no commitment to safety as a cultural value or as a governmental concern, others have great acceptance (and lower injury rates across all ages).

3.2 Behavioral Modification to Achieve Safety

Behavioral research focuses primarily on those situations where structural or passive prevention is inapplicable or limited in effect (Finney et al. 1993). Behaviorally-based approaches typically involve active prevention measures in which individuals take action on their own behalf or for others in order to achieve protection. Deficient safety behaviors or excessive safety behaviors can be changed so that the injury-producing sequence is altered. In these cases, for example, safe behaviors are taught systematically to children, their parents, and adults. Empirically supported procedures include rewards and punishments, modeling, prompting, skills training, and rehearsal. For example, in one American community-wide program, school children collected reward stickers and received prize coupons for riding to school properly secured in car safety seats or seat belts. A 'safe playing' program utilizing behavioral principles was implemented in New Zealand to teach children and their parents the safe limits of outdoor activities. Research has demonstrated the effectiveness of behavioral psychology interventions for changing home hazards such as too high temperatures in water heaters and frayed electrical cords, learning fire emergency escape plans, and wearing bicycle helmets (Roberts et al. 1987).

Alexander and Roberts (in press) offered several caveats about orienting injury prevention efforts targeted to the children themselves. All too often, they noted, parents make the assumption that children are protected once they have been trained in a safety

behavior with a resulting decrease in the supervision the parents give to their children in a hazardous environment. A second problem occurs when children are assigned the responsibility for taking safety action when the adults in the children's lives should be the appropriate agents because they have greater control over changing the environment than the children do. Finally, cognitive and motoric development may determine whether children have the skills and capacity to take safety actions and recognize when situations are appropriate or inappropriate for certain behaviors. Research has shown that parents often do not have adequate knowledge to make informed decisions about safety and frequently misperceive that their children are safe in unsupervised situations in the presence of hazards.

3.3 Health Education to Achieve Safety

Much of the injury prevention effort in many countries has been devoted to providing information (see *Health Education and Health Promotion*) to parents to change their child's environment to be safer or for adults to be safer in work or recreation. This information provision may take place through lectures, pamphlets, or brochures about such things as securing storage of poisons above children's reach, safe lifting practices in adults' workplaces, using gates or locks on barriers to keep children away from stairways, buckling children into safety seats and seat belts in cars and in highchairs, and encouraging bicycle riders of all ages to wear helmets. Similarly, public service announcements are often used on radio and television to encourage seat belt use, or to swim safely, for example, or to discourage driving while drunk. There is little evidence to support injury control through providing such information as a sole intervention, although many professionals agree that information about safety is a necessary, but rarely sufficient aspect of effective injury prevention. Health education is not considered a singly effective intervention to prevent injuries, although it is typically the first approach taken. Information about hazards and what a person should do to avoid them can, however, set the stage for implementing both structural and behaviorally-based programs.

3.4 Community Interventions to Achieve Safety

Interventions made to change the attitudes and behavior of whole communities are rarely evaluated in empirical studies, but all too often are maintained because of popular perceptions of effectiveness or political support. Other community-based programs have demonstrated effectiveness when they rely on intense instruction, behavioral rehearsal, and strong community endorsement of safety and reduction of hazards through systematic rewards for safe behavior such as using seat belts and bicycle helmets (Roberts et al. 1987). Action in the community is required to achieve successful implementation of effective programs as established by research, whether directed at changing behaviors, individually or collectively, or instituting environmental modifications to be safer place for all (National Committee 1989). Increasing public attention to injury causes and changing perceptions about acceptable behavior coupled with recognition that such injuries are preventable has led to recent declines in deaths due to drunk driving and the increased use of seat belts.

4. Future of Injury Research and Control

4.1 Research Questions

The National Committee for Injury Prevention and Control in the United States (1989), reflecting international advancements, outlined the significant questions to be answered through scientific research (pp. 35–6): 'Who is being injured? How are these people being injured? Where are these injuries taking place? What are the circumstances under which these injuries occur? How serious are these injuries? How many of the injuries have occurred, and over what period of time? Which of these injuries is most significant in its personal and social consequences? Is the local injury rate from a particular type of injury higher or lower, than the national (or state) rate? What information will be needed to evaluate an intervention?'

4.2 Comprehensive, Integrated Efforts are Needed

Fragmented injury prevention programs, all too often lacking in evidence of effectiveness, are the norm. These critical questions cannot be answered through a single approach or discipline but from a combination of scientific methodologies to gain a comprehensive view of injury causes and possible prevention/control strategies. Thus, there are limitations on the effectiveness of a single approach to injury prevention whether a structural or active/behavioral intervention. The most advanced thinking views an injury control approach as optimally emphasizing comprehensive combinations of structural/passive and behavioral/active prevention.

4.3 Increased Priority for Research and Prevention

Despite the pervasiveness of injuries, professionals frequently note the lack of social outrage to motivate comprehensive injury control efforts in all countries (National Committee 1989, Roberts and Brooks 1987). Although some selected actions have been successful (e.g., in curbing drunk driving in some countries), injuries at all ages and on all continents in the world remain an enormous societal and scientific

issue for research and development of preventive interventions. Injury control needs to be made a higher priority in many countries, especially in the USA, where inadequate attention has been directed to this overall problem than to other health challenges. Enhanced injury surveillance systems or data collection can improve understanding of the nature and characteristics of injuries. Enhanced evaluation of prevention programs can improve modification and implementation of successful interventions. Related to unintentional injuries are interpersonal violence and suicide (see *Violence as a Problem of Health*). These have similar issues of required improved understanding of the causes and development and, most importantly, for prevention and intervention.

As noted by Roberts and Brooks (1987), 'injury prevention involves societal values and tradeoffs of public good versus individual freedom' (p. 8). Personal rights may be abridged when laws or regulations are passed to achieve safety benefit. Speed limits, mandatory motorcycle helmet laws, and requiring smoke detectors in residences, swimming pool fences, controls on gun ownership and use, enforcement of drunk driving laws, are examples of restrictions imposed on individuals in order to advance safer environments. Other regulations may affect economic decisions and profitability for businesses. For example, requiring cars and trucks to have seat belts, airbags, and enhanced rear brake lights restricts the automobile manufacturers. Some countries have a greater acceptance of regulation and governmental control over individuals and business. Culture and traditions play roles in the perception of the 'common good.' Affecting social attitudes are media presentations of hazardous environments and behavior. For example, risky behavior and violence are glorified and use of drugs and alcohol is portrayed as a social norm in music lyrics, videos, television programs, and movies (National Committee 1989).

Social science contributions continue to be made as methodology is applied to gain a better understanding of injury events and causes as well as to empirically evaluate the effectiveness of interventions to prevent or control injuries (Finney et al. 1993).

See also: Accidents, Normal; Environmental Health and Safety: Social Aspects; Environmental Risk and Hazards; Health Behaviors; Health Education and Health Promotion; Occupational Health and Safety, Regulation of; Safety, Economics of; Violence as a Problem of Health; Workplace Safety and Health

Bibliography

Alexander K, Roberts M C in press Unintentional injuries in childhood and adolescence: Epidemiology, assessment and management. In: Hayman L L, Mahon M M, Turner J R (eds.) *Health and Behavior in Childhood and Adolescence: Cross-Disciplinary Perspectives*. Springer, New York

Barss P, Smith G S, Baker S P, Mohan D 1998 *Injury Prevention: An International Perspective: Epidemiology, Surveillance and Policy*. Oxford University Press, New York

Baker S P, O'Neill B, Ginsburg M J, Li G 1992 *The Injury Fact Book*, 2nd edn. Oxford University Press, New York

Committee on Trauma Research, Commission on Life. Sciences, National Research Council and the Institute of Medicine 1985 *Injury in America: A Continuing Public Health Problem*. National Academy Press, Washington, DC

Finney J W, Christophersen E R, Friman P C, Kalnins I V, Maddux J E, Peterson L, Roberts M C, Wolraich M 1993 Society of Pediatric Psychology Task Force Report: Pediatric psychology and injury control. *Journal of Pediatric Psychology* **18**: 499–526

Haddon W H Jr, Baker S P 1981 Injury control. In: Clark D W, MacMahon B (eds.) *Preventive and Community Medicine*, 2nd edn. Little, Brown, Boston, pp. 109–40

Her Majesty's Stationery Office 1991 *The Health of the Nation*. HMSO, London

National Committee for Injury Prevention and Control 1989 *Injury Prevention: Meeting the Challenge*. Oxford University Press, New York

Roberts M C, Brooks P H (eds.) 1987 Children's injuries: Prevention and public policy. *Journal of Social Issues* **43**(2)

Roberts M C, Fanurik D, Layfield D A 1987 Behavioral approaches to prevention of childhood injuries. *Journal of Social Issues* **43**: 105–18

Robertson L 1998 *Injury Epidemiology: Research and Control Strategies*, 2nd edn. Oxford University Press, New York

US Public Health Service 1991 *Healthy People 2000: National Health Promotion and Disease Prevention Objectives*. US Government Printing Office, Washington, DC

US Public Health Service 2000 *Healthy People 2010*. Website: www.health.gov/healthypeople/

World Health Organization 1999 *World Health Report, 1999: Making a Difference*. Website: www.who.org/whr/1999

M. C. Roberts

Injustice: Legal Aspects

What is injustice? Is it simply the absence of justice, the opposite of a utopian achievement? Or does it have a life of its own? Is it possible to disentangle the experience of injustice from the dissatisfactions and disappointments that people experience everyday? How should an understanding of injustice affect our thinking about law? These are the questions to which any student of injustice must ultimately address himself or herself.

The best recent treatment of the subject of injustice and its complex relations to justice and law is found in Judith Shklar's *The Faces of Injustice* (1990). Shklar (1990, p. 19) tries to rehabilitate injustice as a subject of study, arguing that it 'should not be treated intellectually as a hasty preliminary to the analysis of justice.' According to Shklar (1990, p. 19; also Cahn 1949, Folger 1984), 'The real realm of injustice is not in

an amoral and prelegal state of nature. It does not appear only on those rare occasions when a political order wholly collapses. It does not stand outside the gate of even the best known states. Most injustices occur continuously within the framework of an established polity with an operative system of law, in normal times.'

Shklar recognizes that injustice exists at one end of a continuum of human experience and is often confused with other undesirable conditions. To avoid such confusion, the study of injustice must be differentiated from what she calls misfortune. 'When is a disaster a misfortune and when is it an injustice?,' Shklar asks. For her the answer is to be found in the presence or absence of a responsible human agent. 'If the dreadful event is caused by the external forces of nature, it is a misfortune and we must resign ourselves to our suffering. Should, however, some ill-intentioned agent, human or supernatural, have brought it about, then it is an injustice and we may express indignation and outrage' (Shklar 1990, p. 1).

This analysis treats injustice as a matter of malevolent human action, rather than as a matter of structure or of institutional design. It embodies what Gordon (1996, p. 36) calls the perspective of 'narrow agency,' in which injustice is thought of as a wrong 'done by specific perpetrators to specific victims.' But Gordon suggests that the danger of this perspective is that it will lead us to ignore other kinds of injustices. He reminds us that injustices can be done by collective perpetrators to collective victims, such as the Nazi atrocities against the Jews. This kind of injustice involves what he calls 'broad agency.' In addition, injustice can be attributed to bad structures in addition to bad agents. An interest in this kind of injustice 'takes the form of explanations rather than a search for villains and attribution of blame.'

1. *Responses to Injustice*

Whether defined in terms of narrow agency, broad agency, or structure, some believe that in the United States people are too ready and eager to use the language of injustice and to turn to law to redress conditions which in the past would have been regarded as the uncontrollable workings of fate. Typical is Lieberman's (1981, p. xi) contention that ours 'is a law-drenched age … Though litigation has not routed out all other forms of fight, it is gaining public favor as the legitimate and most effective means of seeking and winning one's just deserts. So widespread is the impulse to sue that litigation has become the nation's secular religion.' In this view, our legal culture is today governed by two principles (Friedman 1985, p. 43). 'The first … is the citizen's expectation of fair treatment, everywhere and in every circumstance … . The second … is the general expectation that somebody will pay for any and all calamities that happen to a person … .' Together these two principles create a culture of 'total justice' (Friedman 1985) and what some believe to be a 'litigious society' (Lieberman 1981, Huber 1988).

In an article I co-authored in 1980, my co-authors and I argued that responses to injurious events are more complicated than arguments about 'total justice' or the 'litigious society' admit. These responses, we noted, could be understood as occurring in three stages (Felstiner et al. 1980–81). The first, defining a particular experience as injurious, we called 'naming.' The next step in the emergence of a claim of injustice is the transformation of a perceived injurious experience into a grievance. This occurs when a person attributes an injury to the fault of another individual or social entity' (Felstiner et al. 1980–81, p. 635). It involves 'blaming.' The third step occurs 'when someone with a grievance voices it to the person or entity believed to be responsible and asks for some remedy' (Felstiner et al. 1980–81, p. 635). This step we called 'claiming.'

It was our contention that 'only a small fraction of injurious experiences ever mature into disputes. Furthermore, we know that most of the attrition occurs in the early stages: experiences are not perceived as injurious; perceptions do not ripen into grievances; grievances are voiced to intimates but not to the person deemed responsible' (Felstiner et al. 1980–81, p. 636). Paying attention to the process through which responses to injustice are formulated and to the high rate of attrition in, or resistance to, the making of claims was, we contended, a way of putting the burgeoning argument about the litigation explosion in context. We argued that the processes through which claims emerge, or through which people decide to 'lump it,' are 'subjective, unstable, reactive, complicated, and incomplete' (Felstiner et al. 1980–81, p. 637). Most of what occurs in this domain is cultural, not legal, occurring as it does in the everyday lives of ordinary citizens and in their struggles to make sense of who they are and who they want to be in a world of risk, danger, and injury (Sarat 2000) (see *Legal Culture and Legal Consciousness*).

Since then, some scholars have studied the way in which people actually react to the injurious events of their lives (Merry 1990, Yngvesson 1993, Greenhouse 1986, Ewick and Silbey 1998). That scholarship characterizes litigation as a process of contesting meanings and adjudication as the choice and imposition of one structure of meaning over another (Merry 1990). Contests over meaning in courts or communities thus become occasions for observing the play of power. Meanings which seem natural, or taken for granted, are described as hegemonic, but because the construction of the meaning of injurious events is, in fact, typically contested, scholars show the many ways in which different people respond to injustice. They have found that the tendency to label injurious events in the language of injustice is contingent and

complex, varying from one community to the next, from one kind of injury to another.

In this process, law plays a key role. As Merry sees it, the relationship between injustice and law is not one in which law passively responds to expectations for 'total justice.' Law works, she says (Merry 1990, pp. 8–9), 'not just by the imposition of rules and punishments but also by the capacity to construct authoritative images of social relationships and actions, images that are symbolically powerful. Law provides a set of categories and frameworks through which the world is interpreted. Legal words and practices are cultural constructs which carry powerful meanings not just to those trained in law … but to the ordinary person as well.' Merry's research shows that responses to injustice involve 'a complex repertoire of meanings and categories understood differently by people depending on their experience with and knowledge of the law' (Merry 1990, p. 5). This work connects the study of injustice to the world of law and is continuous with a long history of scholarship about the connections between law and justice and about the ways law is implicated in injustice.

2. *Injustice and Law*

Commentators from Plato to Derrida have called law to account in the name of justice, scolding and chiding it for the injustices which it does, promotes, or tolerates. They also recognize that all too often law, in this culture and elsewhere, is a tool of injustice (see, for example, Cover 1975, Irons 1983). For every step taken toward realizing the good, an equal, if not greater, number have been taken in the name of evil. These commentators have asked that law provide a language for responding to injustice and demanded that it promote, insofar as possible, the attainment of a just society.

They have done so, however, in relentlessly abstract and general language, as if the demands of justice and the problem of injustice could only be apprehended accurately outside of history and context, and as if only philosophers were fit to engage in conversation about these subjects. The justice which is spoken about is, as a result, elusive, if not illusory, and disconnected from the embodied practices of law of the kind that Merry emphasizes. As one writer noted, 'there is no such thing as justice "in general," one can meaningfully discuss only fairly concrete injustices and procedures for their mitigation with existing social machinery or possible ways of improving the overall social organization' (Knight 1963, p. 1).

The turn to injustice is generally associated with an interest in context and particularity, in which attention is given to individuals and institutions and their contingent histories. In one sense, then, when scholars talk about injustice and try to connect it to law they are often reacting against the vexing difficulty that attends to the effort to examine the connection of law and justice.

In earlier days, however, speaking about law and justice was not so vexing or difficult. Justice (*jus*, meaning 'law') was a legal term, pure and simple. At the outset, then, justice was defined and constituted by laws that were 'given' and held to be unchanging and unchangeable. This ineluctable link between justice and law had the virtue of making the boundaries of justice more or less clear, but it had the considerable vice of labeling even heinous, iniquitous laws just. Attention to the problem of injustice could do no critical/reconstructive work in legal scholarship because it was impossible to think of justice as external to law.

Hobbes (1958) to the contrary notwithstanding, most natural law thinkers have resisted this result by insisting that unjust laws are not law (see, for example, Acquinas 1988, George 1992, Finnis 1890, Hittinger 1987). Doing so meant the end of any easy identification of positive or human law with 'real' or binding law. The alternative, embraced by perhaps a majority of those who do legal philosophy (Hart 1961; for an important contrast, see Fuller 1964), is to cut justice and law free from one another, to insist that justice is more than mere conformity to law and to acknowledge that even unjust laws might nonetheless be law. This has freed students of injustice to direct their attention to the practices of law itself, scrutinizing them for their contributions to injustice in society.

More recently, the distance between law and justice, and the potential for injustice in the name of the law, have been recognized in postmodern theorizing about ethics (Cornell 1988, 1992). Thus, as Douzinas and Warrington (1994, p. 23) (see also Keenan 1990) argue: 'Justice has the characteristic of a promissory statement. A promise states now something to be performed in the future. Being just always lies in the future, it is a promise made to the future, a pledge to look into the event and the uniqueness of each situation … . This promise, like all promises, does not have a present time, a time when you can say "there it is, justice is this or that." Suspended between the law and the good … , justice is always still to come or always already performed.'

Severance of the definitional tie between justice and law has left both notions free (if also bound) to acquire new identities (Balkin 1994). It has opened the way for an examination of unjust actions or conditions and for an inquiry into the ways law is implicated in them. However, this definitional separation has by no means put an end to the commonplace conviction that the principal home of justice is law and the special concern of law, perhaps its principal business, is justice. As Clarence Morris (1963, p. 170) notes, 'Though there can be law without justice, justice is realized only through good law.' In fact, law and legal theory continue to be shaped by concerns about justice and injustice, just as understandings of these notions are

shaped by an awareness of law and the concerns of legal theory (Lucas 1980).

3. *From Law to Politics in the Study of Injustice*

By what might fairly be regarded as an accident of intellectual history, concerns about injustice as a motive for studying and reforming law have been somewhat muted in some contemporary discourse which seeks to expand interest in injustice beyond the sphere of law. They have been swamped by a preoccupation with distributive justice, a subject whose contemporary pre-eminence can unquestionably be attributed to John Rawls' *A Theory of Justice* (1971). 'Justice,' Rawls (1971, p. 3) contends, 'is the first virtue of social institutions, as truth is of systems of thought.' But the justice he has in mind is almost exclusively distributive justice, not justice more broadly construed; injustice of a distributive type is not peculiarly the concern of law at all.

Similarly, the philosopher Agnes Heller (1987, p. 54) introduces the notion of a broad, though incomplete, 'ethico-political concept of justice.' In her view, justice and injustice are not simply about principles of distribution; they concern the perspectives, principles and procedures for evaluating institutional norms and rules. Developing themes in Habermas's (1987) communicative ethics, Heller (1987, p. 54) proposes that justice is 'primarily the virtue of citizenship, of persons deliberating about problems and issues that confront them collectively in their institutions and actions, under conditions without domination or oppression, with reciprocity and mutual tolerance of difference' (Young 1990, p. 33). Domination, oppression, and intolerance are thus preconditions for injustice.

Iris Young (1990, pp. 34, 37) expands on even this notion, concluding that 'the concept of justice coincides with the concept of the political,' and while it is 'not identical with the concrete realization of these values in individual lives,' it is attentive principally to 'the degree to which a society contains and supports institutional conditions necessary for the realization' of the values of equal worth as these are promoted or confined in a society's basic institutional arrangements. The study of injustice then moves from law to politics. On this account the subject is as inclusive as the political order itself.

4. *On the Continuing Significance of Procedure*

Despite these broad and encompassing views of justice and injustice, the fact remains that injustice is generally thought to be involved in matters of procedure, punishment, and recompense (Aristotle 1985, Lucas 1980). Thus, while law in its legislative moments might share with other social institutions certain distributional objectives, it is generally thought that law has a separate and distinctive commitment to doing justice, located most conspicuously, though not exclusively, in matters of form and process. This is, of course, symbolized in the universal icon of justice (and invariably associated with law): the statue of Justicia, the blindfolded bearer of sword and balance (see *Rule of Law*; *Due Process in Law*).

Despite certain ambiguities, Justicia presumably embodies law's central link to justice, impartiality (Curtis and Resnik 1987). The instrument of Justicia is a balance, not a scale, so it is only a comparative rather than an absolute measurement that is involved. She is blindfolded to prevent her decision from being inappropriately influenced or determined by forces of fear, bias, or favoritism. Moreover, she is blindfolded—not blind—suggesting a self-willed refusal of vision. If we are to avoid injustice, partial justice, or biased justice, what is required is a denial of sight though not of seeing, a regulation of information though not of knowing, a restriction on what is permissibly attended to though not a deficit of attention (Black 1989).

According to Curtis and Resnik (1987, p. 1727), 'procedure is the blindfold of Justice.' The proposition that procedure is itself an aspect of justice, and that procedural irregularities are in themselves instances of injustice, takes us some distance toward the further result that law and justice are ineluctably linked. Thus, to the extent that one supposes that the Rule of Law and the principle of legality are constituted largely by a commitment to regularized, self-limiting procedures, it is but a short step to the conclusion that procedural justice is inextricably tied to law itself.

Along some such path, it seems possible to argue further that procedures themselves must meet additional normative requirements, sounding in fairness and desert, and other dimensions of justice. As Hart (1961, Chap. VIII) has observed, a rule that is regularly, uniformly applied to prohibit certain minorities from using the town park might, despite its unfailing enforcement, be viciously unjust. Legal justice, then, involves more than general observance of the rules that regulate exercises of legal authority, the rules themselves must be fair, eschewing wholly arbitrary distinctions between and among the persons to whom they apply or on whom they finally have effect. The rules must provide suitable forms of what, in the United States, are known as requirements of due process and equal protection.

Acknowledging the relationship of procedure and justice has spawned a large and important empirical literature which explores what leads people to believe that law is just or unjust (Lind and Tyler 1988, Tyler 1989). This literature suggests that procedure is more important than substance, that the way one is treated is more important than the results one gets in determining whether people believe that they have been treated justly or unjustly in the legal system. As Lind (1998, p. 184) explains. 'If authorities seem to be

making decisions in fair ways, people assume that they can obey orders and follow rules without worrying too much about exploitation. If, on the other hand, authorities seem to be reacting unfairly, exploitation becomes a real possibility and obedience tends to be viewed with more suspicion … . Outcome-oriented theories of justice judgements err because they fail to recognize that people are generally quite ready to accept outcomes that are less than they had hoped for.'

5. *Can Law Overcome Injustice?*

Returning to the questions with which we started, we have seen that injustice is not simply an absence of justice, that it is a fact and an experience different from, though related to, misfortune, and that it may involve the relations of one person to another, of groups of people, and of the attributes of legal, political, or social structure. Though some would focus the attention of persons interested in understanding injustice fairly broadly, law seems to be at the center of any proper understanding of injustice.

Yet having envisaged ways of connecting the study of injustice and of law, it must be granted that not everyone regards the imagined association as an unalloyed good. Hayek (1982), among others, is persuaded that social justice is a mirage, a misguided affection the pursuit of which not only threatens liberty but, eventually, law itself. A somewhat more moderate concern is that the demands of distributive justice and the Rule of Law are inevitably at odds with one another, so that the effort to cure distributive injustices comes at a high price to law itself (Unger 1977, Barnett 1988). A third apprehension is that the meaning of justice is so ambiguous, expansive, and elusive that law invariably will be seen as falling short, as failing to meet justice's demands, as always, then, implicated in injustice. As Balkin (1994, p. 16) argues, laws apportion responsibility, create rights and duties, and provide rules for conduct and social ordering. Law is always, to some extent and to some degree, unjust. At the same time, our notion of justice can only be articulated and enforced through human laws and conventions. We may have a notion of justice that always escapes láw and convention, but the only tools we have to express and enforce our idea are human laws and human conventions. Our conception of the just relies for its articulation and enforcement on the imperfect laws, conventions, and cultural norms from which it must always be distinguished.

Some worry that law's legitimate authority, such as it is, will be dangerously (and unjustly) eroded (Tyler 1990). Alternatively, but with consequences no less grave, the impossibility of satisfying justice's demands may eventually be seen as the fault of justice, not of law.

Still other qualms dampen enthusiasm about the possibility of connecting law to the study of justice and injustice. Some believe that the embrace of legal justice is, at best, a rhetorically appealing device for quieting alarm about the excesses and inequities social and economic arrangements inevitably arouse (Hay 1975). So conceived, legal justice is largely a symbolic response to the incorrigible defects and injustices of liberal capitalist societies, not a positive virtue which these arrangements can promote or make possible (Marx 1978).

The fear that legal justice can be put to work in the name of legitimating social injustice, for example by 'justifying' outcomes that seem patently, outrageously disproportionate has been expressed in a different voice, and on different grounds, by Shklar (1990). She argues that the focus on justice inhibits the development of a community that both recognizes the culturally constructed quality of boundaries between the ideas of 'injustice' and 'misfortune,' and objects to passive injustice as vehemently as it abhors the active violation of rights. But Shklar's dissatisfaction with a narrow interest in legal justice is pressed even further in her contention that contemporary theory tends almost exclusively to regard injustice as 'simply the absence of justice' (Shklar 1990, p. 15, Woozley 1973).

On this view, 'once we know what is just, we know all we need to know,' Shklar (1990, p. 15) writes. Or, as Knight (1963, p. 15) observes, 'One misses a great deal by looking only at justice. The sense of injustice, the difficulties of identifying victims of injustice, and the many ways in which we learn to live with each other's injustices tend to be ignored, as is the relation of private injustice to the public order.' Scholars such as Shklar and Knight contend that we not only lack a full and coherent account of justice; even if we had one, it would not yield an equally full account of injustice, nor would it confront adequately our calloused complacency regarding vast, yet eliminable, misfortunes.

These observations call to mind a variety of uncertainties and ambivalences regarding the relationship of law and justice, but they fall far short of embracing the extreme position that justice be jettisoned from legal discourse. Rather, they remind us of the vastness of the subject of injustice, of the difficulty of constructing a single account capable of holding together its many strands, and of the space that exists to theorize anew about justice and injustice in law and legal theory.

See also: Due Process in Law; Justice and Law; Natural Law; Rights: Legal Aspects; Rule of Law

Bibliography

Aquinas T 1988 *Summa Theologica*. Baumgarth W, Kegan R (eds.). Hackett, Indianapolis, IN

Aristotle 1985 *Nicomachean Ethics*. Irwin T (trans.). Hackett, Indianapolis, IN

Balkin J 1994 Being just with deconstruction. *Social and Legal Studies: an International Journal* **3**: 393–405

Barnett R 1988 Can justice and the rule of law be reconciled foreward? *Harvard Journal of Law and Public Policy* **11**: 597–624
Black D 1989 *Sociological Justice*. Oxford University Press, New York
Cahn E 1949 *The Sense of Injustice: an Anthropocentric View of Law*. New York University Press, New York
Cornell D 1988 Post-structuralism, the ethical relation, and the law. *Cardozo Law Review* **9**: 1587–620
Cornell D 1992 *The Philosophy of the Limit*. Routledge, New York
Cover R M 1975 *Justice Accused: Antislavery and the Judicial Process*. Yale University Press, New Haven, CT
Curtis D E, Resnik J 1987 Images of justice. *Yale Law Journal* **96**: 1727–72
Derrida J 1990 Force of law: 'The mystical foundation of authority.' *Cardozo Law Review* **11**: 919–1046
Douzinas C, Warrington R 1994 The face of justice: A jurisprudence of alterity. *Social and Legal Studies: an International Journal* **3**: 405–26
Ewick P, Silbey S S 1998 *The Common Place of Law: Stories from Everyday Life*. University of Chicago Press, Chicago
Finnis J 1890 *Natural Law and Natural Rights*. Clarendon Press, Oxford, UK
Folger R (ed.) 1984 *The Sense of Injustice: Social Psychological Perspectives*. Plenum Press, New York
Felstiner W L, Abel R L, Sarat A 1980–81 The emergence and transformation of disputes: Naming, blaming, and claiming. *Law and Society Review* **15**: 631–54
Friedman L 1985 *Total Justice*. Russell Sage Foundation, New York
Fuller L L 1964 *The Morality of Law*. Yale University Press, New Haven, CT
George R P 1992 *Natural Law Theory: Contemporary Essays*. Clarendon Press, Oxford, UK
Gordon R 1996 Undoing historical injustice. In: Sarat A, Kearns T R (eds.) *Justice and Injustice in Law and Legal Theory*. University of Michigan Press, Ann Arbor, MI, pp. 35–76
Greenhouse C J 1986 *Praying for Justice: Faith, Order, and Community in an American Town*. Cornell University Press, Ithaca, NY
Gruber G M A (trans.) 1992 *Plato's Republic*. Rev. Reeve C D C. Hackett, Indianapolis, IN
Habermas J 1987 *The Theory of Communicative Competence, Vol. 2: Life World and System*. Beacon Press, Boston
Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK
Hay D 1975 Property, authority and the criminal law. In: Hay D, Linebaugh P, Thompson E P (eds.) *Albion's Fatal Tree*. Lane, London, pp. 17–63
Hayek F A 1982 *Law, Legislation and Liberty, Vol. II: The Mirage of Social Justice*. Routledge, London
Heller A 1987 *Beyond Justice*. Basic Books, New York
Hittinger R 1987 *A Critique of the New Natural Law Theory*. University of Notre Dame Press, Notre Dame, IN
Hobbes T 1958 *Leviathan*. Bobbs-Merrill, Indianapolis, IN
Huber P W 1988 *Liability: The Legal Revolution and its Consequences*. Basic Books, New York
Irons P 1983 *Justice at War: The Story of the Japanese American Internment Cases*. Oxford University Press, New York
Keenan T 1990 Deconstruction and the impossibility of justice. *Cardozo Law Review* **11**: 1675–86
Knight F H 1963 On the meaning of justice. In: Friedrich C J, Chapman J W (eds.) *Justice*. Atherton Press, New York, pp. 1–23
Lieberman J K 1981 *The Litigious Society*. Basic Books, New York
Lind E A 1998 Procedural justice, disputing, and reactions to legal authorities. In: Sarat A, Constable M, Engel D, Hans V, Lawrence S (eds.) *Everyday Practices and Trouble Cases*. Northwestern University Press, Evanston, IL, pp. 177–98
Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum Press, New York
Lucas J R 1980 *On Justice*. Clarendon Press, Oxford, UK
Marx K 1978 Critique of the Gotha program. In: Tucker R (ed.) *The Marx–Engels Reader*, 2nd edn. Norton, New York, pp. 525–42
Merry S 1990 *Getting Justice and Getting Even: Legal Consciousness Among Working-class Americans*. University of Chicago Press, Chicago
Morris C 1963 Law, justice, and the public's aspirations. In: Friedrich C J, Chapman J W (eds.) *Justice*. Atherton Press, New York, pp. 170–90
Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA
Sarat A 2000 Exploring the hidden domains of civil justice: 'Naming, Blaming and Claiming' in popular culture. *DePaul Law Review* **50**: 425–52
Shklar J N 1990 *The Faces of Injustice*. Yale University Press, New Haven, CT
Tyler T 1989 The psychology of procedural justice: A test of the group value model. *Journal of Personality and Social Psychology* **57**: 830–38
Tyler T 1990 *Why People Obey the Law*. Yale University Press, New Haven, CT
Unger R 1977 *Law and Modern Society*. Free Press, New York
Woozley A D 1973 Injustice. *American Philosophical Quarterly* **7**: 109–22
Yngvesson B 1993 *Virtuous Citizens, Disruptive Subjects: Order and Complaint in a New England Court*. Routledge, New York
Young I M 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

A. Sarat

'Inner City,' The: Cultural Concerns

1. Introduction

While the growth of concentrated inner-city joblessness and poverty is, to varying degrees, characteristic of many older industrial cities in Western Europe, it is most pronounced in the central cities of the Northern and Midwestern United States (Summers 1993). As a result of fundamental changes in the national and global economies leading to the decreased relative demand for low-skilled labor, many of the USA's inner-city neighborhoods have experienced unprecedented levels of joblessness since 1970. High joblessness in segregated neighborhoods has contributed to other problems including concentrated poverty, family dissolution, welfare receipt, and declining neighborhood social organization. The net result has been the increasing social isolation of inner-city residents and therefore the crystallization of a set of habits, styles,

skills, orientations, and other ghetto-related cultural traits. Although these traits grow out of conditions of economic and social disadvantage in a highly segregated environment, they also reinforce the economic marginality of inner-city residents. The success of policies to combat inner city woes, including the problems of joblessness, will largely depend upon the extent to which they are integrated into or are a part of broader programs to enhance the economic and social health of the metropolitan region.

2. *The Rise of the Modern Inner-city Ghetto*

One of the legacies of the historic racial and class subjugation in the United States is the growing concentration of poor and minority residents in the nation's impoverished inner-city areas. The demand for low-skill blue collar workers in the North during the early twentieth century, coupled with the rise in southern blacks' economic insecurity brought about by the declining demand for farm labor due to the increasing mechanization of agriculture, helped fuel the migration of blacks to northern industrial cities. In the North, blacks found themselves increasingly segregated in ghettos that previously housed white ethnic immigrants, who themselves were gradually assimilating into US society.

In spite of the tremendous obstacles faced by African-Americans, ghetto neighborhoods, in the early parts of the twentieth century, developed a diversified social and institutional structure, characterized by a mix of poor, working-class, and middle-class blacks and viable institutions (e.g., businesses, schools, churches, and cultural organizations) that fostered a vibrant social and cultural life. This social organization of earlier ghettos helped to buffer the deleterious effects of marginalization and concentrated poverty (Dubois 1899, Drake and Cayton 1962). And despite their disadvantaged status, well into the 1960s most ghetto-dwellers resided in neighborhoods where a majority of adults were employed.

3. *The Social Transformation of the Inner-city Ghetto: Competing Explanations of the Growth of Concentrated Poverty*

Since 1970, conditions in those same inner-city neighborhoods have dramatically declined, reflecting a different social and economic environment than existed in early urban ghettos. The new urban ghetto has a much higher level of joblessness and a greater concentration of poverty (Wilson 1987, 1996, Jargowsky and Bane 1991, Jargowsky 1997). Explanations for the growth of inner-city joblessness and concentrated poverty have centered on four mitigating factors: economic changes, demographic shifts, government programs and policies, and the legacy of racial oppression.

Central-city labor markets have been altered by economic changes affecting the distribution and educational requirements of employment. Beginning in the 1970s, global competitive pressures and declining profits in manufacturing set off a wave of plant closings that decimated the old industrial centers of the USA. Many of the companies that survived moved off-shore or to the South to take advantage of lower production costs. Some employers, attracted by the cheap land and skilled labor force, moved to suburban areas. The decline in central-city manufacturing jobs was a major impetus to the growth of concentrated urban poverty (Wilson 1987, 1996).

Employment in the inner city has also been affected by the sectoral shift from goods-producing to service-producing industries and an associated increase in educational requirements of employment. Whereas most job growth has been in those industries that require higher levels of education, job loss has disproportionately affected the less educated (Kasarda 1989). Not only did the shift to higher-skilled service sector jobs reduce the demand for inner-city workers, new production technologies also eliminated many unskilled manual jobs in manufacturing. The social costs of the loss of low-skill manufacturing jobs and their replacement with more skilled jobs have not been equally shared. In particular, inner-city blacks have disproportionately felt the impact of these changes. The new spatial distribution of employment within metropolitan areas has also disadvantaged inner-city neighborhoods and residents. In the competition for jobs, suburban areas, with their cheaper land, access to highways, and other amenities, have an advantage over central city areas. Thus, the intensification of inner-city poverty is at least partly the result of the increasing suburbanization of employment that has helped to generate a growing mismatch between the suburban location of employment and the inner-city residence of minority populations (Wilson 1987). Inner-city joblessness increases as residents must contend with both fewer employment opportunities nearby and longer commutes to jobs outside the central city.

A demographic trend that has contributed to the spread of concentrated inner-city poverty is the movement of nonpoor families out of inner-city neighborhoods, first by whites and later by blacks. Beginning in the 1950s, nonpoor whites began to move out of many inner-city neighborhoods and into suburban locations. The economic vulnerability of blacks, especially the increased likelihood of joblessness, meant that neighborhoods that became predominantly black were more at risk of becoming high poverty areas. In the face of a less restrictive housing market in the post-civil rights era, many working-class and middle-class blacks also left poor inner-city neighborhoods in search of more favorable living conditions. As the proportion of nonpoor families in the neighborhood declined, the levels

of poverty increased, pushing many previously mixed-income ghetto areas into greater poverty (Wilson 1987, Jargowsky and Bane 1991, Jargowsky 1997).

Some governmental programs and policies have exacerbated the decay of inner-city neighborhoods (McGeary 1990, Wilson 1996). Policies such as the selective underwriting of mortgages, and mortgage interest tax deductions, favored the suburbanization of middle-class whites. This trend was facilitated by the construction of highway networks linking the suburbs and central cities. The construction of massive federal housing projects in inner-city neighborhoods in the 1950s and 1960s isolated poor black families for decades and helped to create some of the most highly concentrated poverty in the country. More recently, the policies of the Reagan and Bush administrations pushed many city governments to near bankruptcy under the New Federalism program that substantially reduced the federal contribution to city budgets, at a time when cities were desperate for federal help in combating escalating urban problems.

Finally, the high levels of racial segregation in American society, combined with higher rates of joblessness and poverty among blacks, helped to spread and intensify the concentration of poverty in predominantly black inner-city areas (Massey and Denton 1993). Segregated ghettos are less conducive to employment preparation and employment than are other urban residential areas. Segregation in ghettos aggravates employment problems because it leads to weak informal employment networks and contributes to the social isolation of families and individuals, thereby reducing their chances of acquiring the human capital skills, including adequate educational training, that facilitate social mobility.

4. *Concentrated Poverty, Social Organization, and Culture in the Inner City Neighborhoods*

During the 1970s and 1980s, social dislocations such as crime, gang activity, joblessness, and welfare receipt soared in many of the USA's segregated inner cities. Attempts by some social scientists to explain these trends centered on the behavior of the urban poor, behavior believed to be rooted in a cultural milieu of deviant norms and values (Murray 1984, Mead 1986).

Alternative explanations emphasized the structural underpinnings of the inner city and the extent to which macroeconomic conditions have undermined the social and cultural life of the ghetto (Wilson 1987). According to this view, ghetto-related behavior and attitudes are best understood as cultural adaptations to the lack of opportunities in the inner-city environment and the larger society (Clark 1965, Hannerz 1969, Rainwater 1970, Anderson 1990).

The social constraints faced by inner-city minorities, and their cultural manifestations, reflect their social isolation from mainstream society, especially the lack of employment opportunities (Wilson 1987, 1996). The outmigration of nonpoor blacks from poor inner-city neighborhoods contributed to this isolation by eroding the social buffer that they provided, and by making it difficult to sustain basic institutions, such as schools, churches, stores, and recreational facilities. By reducing the opportunities for contact with regularly employed adults and the opportunities for access to the social resources that higher status families provide (e.g., job networks), residents in high-poverty neighborhoods are less likely to develop the culturally shaped skills, habits, styles, and orientations so prevalent in neighborhoods organized around work (Wilson 1987, 1996).

Also, concentrated poverty and joblessness undermine the social organization of inner-city neighborhoods, or the ability of residents to maintain effective social control and realize common goals (Wilson 1996). Socially organized neighborhoods have strong and cohesive informal social networks and ample institutional resources for the preservation of neighborhood social order. Strong social networks are characterized by sustained social contacts that are based on trust, reciprocity, and normative consensus, and where there are extensive ties between parents, children, and other institutionally based adults (e.g., police, teachers, and other community leaders) (Coleman 1990). Social control is also aided by such institutional resources as effective policing, quality schools, and adequate community organizations and other locally based associations.

In contrast to socially organized neighborhoods, many segregated inner-city neighborhoods have few institutional resources and weak and less cohesive social networks. As a result residents are less able to enforce norms and control undesirable behavior in the community or to act collectively to preserve the quality of neighborhood life. When the effectiveness of formal and informal social control mechanisms drops below a certain threshold, sanctions against nonnormative behavior become ineffective and opportunities for illegitimate activities proliferate in the community (Sampson and Wilson 1995).

In sum, the cultural milieu of the inner city is embedded in the structural features of concentrated poverty, social isolation, and the lack of neighborhood social organization. Cut off from employment opportunities and trapped in neighborhoods with little access to conventional role models, social networks that contain social capital (e.g., informal job networks), and institutional resources that support families, poor inner-city families face severe obstacles in their efforts to accomplish socially desired ends such as labor force participation and family stability.

5. *The Inner-city Family: A Mediating Variable*

Since 1970, marriage rates have been declining nationally and more children are being raised in single-parent

families. Nowhere are these trends more prevalent than in the inner-city, where mother-only households are the norm in many neighborhoods. This situation has generated much public debate and, while public attention has focused on the marital disincentives of welfare policy, research has shown that a major reason for the collapse of two-parent inner-city families is the growing economic marginality of inner-city males (Wilson 1987, 1996). As male employment prospects receded, so did the economic benefits of marriage. The general weakening of social sanctions against out-of-wedlock childbearing in the larger society and especially in the inner-city ghetto further helped to undermine the foundation for stable relationships.

The high rates of single parenthood are especially troubling because of their association with persistent poverty, welfare receipt, and deleterious effects on children. Children in mother-only households are themselves more likely to be school dropouts, to have lower earnings, and to depend on welfare as adults (Krein and Beller 1988, McLanahan and Garfinkel 1989). Moreover, growing up in communities where prospects for steady employment and stable marriages are perceived as remote, young inner-city adults are more likely to engage in behavior that further jeopardizes their chances for social and economic mobility.

6. *The Inner City and Public Policy*

After years of steady urban decline, the prolonged economy recovery in the USA has recently improved situations in central cities. Unemployment has declined dramatically from 8.6 percent in 1992 to 5.1 percent in 1998. And unemployment in the nation's 50 largest cities fell from 8.6 percent in 1992 to 4.9 percent in 1998, a decline that exceeded the decreases in unemployment in the cities' surrounding suburbs (US Department of Housing and Urban Development 1999).

The positive effects of these changes are seen in even the most depressed parts of the city. A new study by the economists Richard Freeman and William M. Rodgers of low-wage workers in 322 metropolitan areas, reveals that black men aged 16 to 24 with a high school education or less—including many with prison records—are employed in greater numbers, earning larger paychecks and committing fewer crimes than in the early 1990s. Although far too many of these young men are still jobless or in prison, the rise in legitimate employment has accompanied a drop in criminal activity. Indeed, crime has fallen most rapidly in regions with the sharpest declines in joblessness. (Freeman and Rogers 1999, Nasar and Mitchell 1999). Nonetheless, the highly disadvantaged urban areas have not rebounded to the same extent as other areas in the metropolitan regions. Moreover, if the economy slows down and the country is faced with a period of economic stagnation or a new recession the problems of joblessness and poverty will become acute once again.

Any real solution to the economic and social woes of the inner city will have to address the full range of economic and social problems that currently plague them. Policies that guarantee full employment, reduce income inequality, rebuild the physical and social infrastructure of cities, and provide adequate social support for the nation's less fortunate would go a long way towards redressing the USA's urban inequalities. However, in recent years such programs have not been reflected in government policy. Indeed, as discussed above, government action has often unwittingly contributed to urban inequalities. Similarly, governmental agencies often work at cross-purposes, for example, as one seeks suburban economic development, another tries to revitalize low-income urban communities by luring businesses back to the central city. Disparate and divergent policies of this sort have 'worked to weaken the social, economic, and political strength of inner city communities' (Ferguson 1999, p. 587).

Furthermore, cross-national comparisons of urban policy suggest that the fragmented governmental structure of the USA has been, and will continue to be, a major obstacle to halting the decline of inner cities. Few Western European cities witnessed the dramatic urban decay of US cities, in part because central governments there have sought to maintain the livability of cities by controlling urban growth boundaries, implementing more effective urban renewal programs, and building cheap and efficient public transportation systems (Weir 1993, Downs 1993). In contrast, the USA has no comprehensive federal plan to influence urban settlement patterns, and most urban matters are under the jurisdiction of local governments and are heavily influenced by powerful economic actors in the private sector.

However, there is mounting evidence that the problems of jobless ghettos cannot be separated from those of the suburbs, and that the health of metropolitan regions may depend on a viable central city. For example, research indicates that the reduction of central city poverty is associated with increases in metropolitan income growth, and that central city job growth increases the value of suburban properties (Gottlieb 1998).

In the global economy, metropolitan regions continue to compete for jobs. In an era of low transportation and information costs, high mobility and intense global competition, a metropolitan region is at a severe competitive disadvantage if it lacks a healthy urban core. 'In a global economy, firms choose among regions—and the health of the central city is a key factor in deciding which region is best. Even firms that choose to relocate to the suburbs will choose areas surrounding a vibrant central city' (U S Department of Housing and Urban Development 1999, p. 19). In the USA cities and suburbs should look to the creation

of metropolitan and regional governmental bodies or city–suburban collaborative planning mechanisms to help resolve shared metropolitan-wide concerns. And paramount among these concerns should be the revitalization of the most disadvantaged neighborhoods. It would be a serious mistake to assume that efforts to enhance the economic and social health of the metropolitan region can be achieved without combating the problems of poverty, joblessness, and related woes in the inner city.

See also: Population Ecology; Poverty, Sociology of; Urban History; Urban Life: Future Developments; Urban Sociology

Bibliography

Anderson E 1990 *Streetwise: Race, Class, and Change in an Urban Community*. University of Chicago Press, Chicago

Clark K B 1965 *Dark Ghetto: Dilemmas of Social Power*, 1st edn. Harper & Row, New York

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Downs A 1993 Contrasting strategies for the economic development of metropolitan areas in the United States and Western Europe. In: Summers A A, Cheshire P C, Lanfranco S (eds.) *Urban Change in the United States and Western Europe*, 2nd edn. The Urban Institute Press, Washington, DC, pp. 15–37

Drake S C, Cayton H R 1962 *Black Metropolis: A Study of Negro Life in a Northern City*. Harper and Row, New York

Dubois W E B 1899 *The Philadelphia Negro*. University of Pennsylvania Press, Philadelphia, PA

Ferguson R F 1999 Conclusion: Social science research, urban problems, and community development alliances. In: Ferguson R F, Dickens W T (eds.) *Urban Problems and Community Development*. Brookings Institution Press, Washington, DC, pp. 569–610

Freeman R B, Rodgers III W M 1999 Area economic conditions and the labor market outcomes of young men in the 1990s expansion. Unpublished Manuscript, Harvard University, Cambridge MA

Gottlieb P D 1998 *The Effects of Poverty on Metropolitan Area Economic Performance*. A Research Report of the National League of Cities, National League of Cities, Washington, DC

Hannerz U 1969 *Soulside: Inquiries into Ghetto Culture and Community*. Columbia University Press, New York

Jargowsky P A 1997 *Poverty and Place: Ghettos, Barrios, and the American City*. Russell Sage Foundation, New York

Jargowsky P A, Bane M J 1991 Ghetto poverty in the United States, 1970–1980. In: Jencks C, Peterson P E (eds.) *The Urban Underclass*. Brookings Institution, Washington, DC, pp. 235–73

Kasarda J D 1989 Urban industrial transition and the underclass. *Annals of the American Academy of the Political and Social Sciences* **501**: 26–47

Krein S F, Beller A H 1988 Educational-attainment of children from single-parent families: Differences by exposure, gender, and race. *Demography* **25** (May): 221–34

Massey D S, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA

McGeary M G H 1990 Ghetto poverty and federal policies and programs. In: Lynn L E Jr, McGeary M G H (eds.) *Inner City Poverty in the United States*. National Academy Press, Washington, DC, pp. 223–52

McLanahan S, Garfinkel I 1989 single mothers, the underclass, and social policy. *Annals of the American Academy of the Political and Social Sciences* **501**: 92–104

Mead L M 1986 *Beyond Entitlement: The Social Obligations of Citizenship*. Free Press, New York

Murray C 1984 *Losing Ground: American Social Policy 1950–1980*. Basic Books, New York

Nasar S, K B Mitchell 1999 Booming job market draws young black men into fold. *The New York Times* May 23, pp. 1, 21

Rainwater L 1970 *Behind Ghetto Walls*. Aldine, Chicago

Sampson R J, Wilson W J 1995 Toward a theory of race, crime, and urban inequality. In: Hagan J, Peterson R D (eds.) *Crime and Inequality*. Stanford University Press, Stanford, CA, pp. 37–54

Summers A A 1993 Urban America and urban Western Europe: Are they different? In: Summers A A, Cheshire P C, Lanfranco S (eds.) *Urban Change in the United States and Western Europe*. The Urban Institute Press, Washington, DC, pp. 1–14

US Department of Housing and Urban Development 1997 *The State of Cities*. US Department of Housing and Urban Development, Washington, DC

US Department of Housing and Urban Development 1999 *Now is the Time: Places Left Behind in the New Economy*. US Department of Housing and Urban Development, Washington, DC

Weir M 1993 Race and urban poverty: Comparing Europe and America. Occasional Paper no 93-9, Center for American Political Studies, Harvard University, March

Wilson W J 1987 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago University Press, Chicago

Wilson W J 1996 *When Work Disappears: The World of the New Urban Poor*, 1st edn. Knopf, New York

W. J. Wilson, B. H. Rankin, and J. M. Quane

Innovation and Technological Change, Economics of

Technological innovation is the function through which new technologies are introduced into the economic system. It entails recognizing new technological possibilities, organizing the human and financial resources needed to transform ideas into useful products or processes, and carrying out the requisite activities (typically called 'research and development'). It is important because technological advances have played a paramount role in facilitating the radically improved standards of living experienced by the inhabitants of progressive nations over the past several centuries. Innovation is responsive to econ-

omic forces and the lure of profit, but their influence is complicated by uncertainty and the difficulty innovators encounter in capturing a sufficient share of the economic benefits resulting from their contributions.

1. *Acceleration and Systematization*

New technologies have been introduced into production and consumption throughout recorded history. During the eighteenth century, however, there were discernible shifts in the pace and organization of technological change through what is now called the first Industrial Revolution. Landes (1969) and Mokyr (1990). New machines reduced dramatically the amount of labor required to accomplish productive tasks, steam engines provided less costly and more reliable means to drive the machines, the machines themselves were perfected through improvements in machine-building technology, and innovations in agriculture freed workers to produce a widening array of traditional and new manufactured products. One impact of these changes was an increase in the quantity of goods and services that could be produced by the average worker, i.e., an increase in labor productivity. Higher productivity in turn meant enhanced real income or purchasing power, sometimes at first only for some segments of the population but ultimately for virtually all consumers. Landes (1969) estimates that the growth of real income per capita in western Europe increased from a rate of roughly 0.16 percent per annum between AD 1000 and 1700 to 0.4 percent between 1700 and 1750. It then soared into a range of 1.2–1.5 percent per year during the next century, i.e., doubling every 46 to 58 years.

It was always recognized that the acceleration of per capita income growth beginning with the first Industrial Revolution was coupled closely to the implementation of new technologies. Debate continues over the exact contribution of new technology as compared to increases in the amount of capital goods with which work is accomplished, efficiencies associated with increases in the scale of productive operations ('economies of scale'), and the enhancement of 'human capital,' i.e., the education and training of workers (see *Human Capital: Educational Aspects*). Full resolution of the debate is impossible, because those contributors to economic growth are intrinsically complementary. Much new technology is embodied in capital goods and implemented only when new investments are made. Similarly, technological advances and investment typically are needed to realize economies of scale. And neither technological advances *per se* nor their implementation on an industrial scale would be possible without well-educated and well-trained workers. What is clear is that advances in technology play a very important role in driving productivity improvements and hence improvements in real income per capita (Griliches 1995).

The economist Schumpeter (1912, 1934) was the first to stress a systematic functional role for innovation, or the carrying out of new combinations by an entrepreneur, in achieving economically meaningful technological advances. He wrote (1934):

> Economic leadership in particular must be distinguished from 'invention.' As long as they are not carried into practice, inventions are economically irrelevant. And to carry any improvement into effect is a task entirely different from the inventing of it, and a task, moreover, requiring entirely different kinds of aptitudes.

The differentiation of activities directed toward carrying technological innovations into practice was already evident during the late eighteenth century in the famous Boulton & Watt steam engine partnership, which included a group of workers specialized in performing what today would be called research and development (Scherer 1965). In Schumpeter's schema, the 'innovator' recognized the commercial potential of new technologies, raised the capital for and accepted the primary risks faced in implementing them, and organized the effort needed to perfect them and introduce them into the market (today, research, development, and marketing). The spread of this systematized approach to technological innovation was at first gradual, but by 1994, industrial enterprises in 25 nations surveyed by the Organization for Economic Development and Cooperation (1997) spent a combined total of roughly $237 billion (measured at 1990 price levels) on formally organized research and development activities.

2. *Endogenous Innovation*

Most of the mainstream economic writings up to the middle of the twentieth century viewed technological innovation as a function regulated by forces distinct from and independent of the supply and demand conditions through which markets for well-established goods and services reached an equilibrium. Innovation was said to be an 'exogenous' activity, not systematically influenced by economic variables (i.e., 'endogenous'). Innovations came from the efforts of 'hero' inventors who, though not immune to economic motivations, were driven at least as much by the urge to tinker and improve; or as a fortuitous by-product from the advance of scientific knowledge sought for its own sake or in the seventeeenth century, *ad majorem Gloriam Dei*.

Marx (1887) was the first to see innovation as a direct consequence of capitalists' striving to enhance their profits, but his writing had little influence on mainstream economic views. Schumpeter's insights (1912) were more influential, but only with a lag of a half-century. In Schumpeter's schema, the innovator recognized opportunities to achieve supranormal profits by introducing superior alternatives to the

products or processes embraced by the vast horde of competitive enterprises; thus, innovative activity was expressly profit-seeking. The profits realized by the innovator in turn sent a signal to other firms that their accepted technologies were becoming obsolete, but they could defend themselves and gain profits by imitating the innovator's contribution. Thus, an original innovation was followed sooner or later by a 'swarm' of imitators whose competitive efforts drove prices down until only normal returns on investment could be sustained. Further innovation was necessary if supra-normal profits were again to be gained. Schumpeter (1942) called this incessantly repeated chain 'the process of creative destruction,' adding that the organization of innovative functions in modern research and development laboratories imparted a routinization that reduced the heroic role played by Schumpeter's entrepreneur and perhaps (he wrongly feared) jeopardized the continuation of rapid technological progress.

Exactly why profit opportunities opened up for technological innovators was left unsettled by Schumpeter. Conventional wisdom held that the advance of science was the key agenda-setting influence: scientific breakthroughs or, less dramatically, the steady accretion of scientific knowledge, created an opportunity to make economically significant and profitable innovations. This view of the world was challenged by Schmookler (1966), whose qualitative and quantitative historical research pointed toward a crucial innovation-inducing role for demand-side influences. The greater the demand for a set of products, the more profitable improving upon those products was likely to be, and hence the larger was the flow of innovations aimed at satisfying consumers' (or producers') demands for those products (see also Usher 1954).

A formal model of how the economic forces of supply and demand endogenously affect the rate and direction of innovative activity can be synthesized from the contributions of Schmookler (1966), Scherer (1967), and Barzel (1969). Investing in innovation is viewed like other investment activities: a lump sum outlay is required for research, development, and introductory marketing. The innovator chooses to incur that outlay only if the investment is expected to be profitable, i.e., if the anticipated stream of quasi-rents or producer's surplus (cost savings, or new product sales revenues minus production and routine marketing costs), discounted to present value, exceeds the expected lump-sum innovation outlay. Let the lump-sum innovation outlay be RD and the discounted present value of quasirents be V. How much must be spent to realize the innovation depends upon the state of scientific knowledge and related technology. As knowledge advances, what was impossible at one time (i.e., infinitely expensive) becomes less and less costly over time. Smooth and continuous advance in the knowledge base implies that if carrying out the

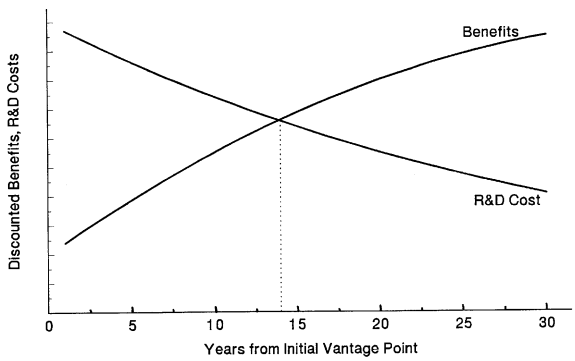

Figure 1
Innovation timing with continuous demand and cost changes

innovation in year zero costs RD_o, implementing it T years later costs $RD_o\, e^{-aT}$,where $100a$ is the percentage rate at which the cost of innovation declines per year as knowledge advances. On the payoff side, let $v(t)$ be the quasi-rent or producer's surplus attainable in year $t$from an innovation. As demand grows, quasi-rents are likely to grow apace, so from an initial quasi-rent level of v_o in year zero, quasirents will grow to $v_o e^{+gT}$ in year T, where g is the annual rate of growth. Assuming for simplicity the innovation to have an infinite revenue-yielding life once implemented, the discounted present value of quasi-rents for an innovation introduced in year T is therefore

$$V(T) = \int_T^{\infty} v_o e^{(g-r)T} dt \qquad (1)$$

where r is the risk-adjusted discount rate. The innovator seeks to maximize its profits from innovation, so it will not plan to innovate if $V(T) < RD_o\, e^{-(a+r)T}$, with discount rate r added to the R&D cost term to reflect the fact that costs incurred in the future are valued less than costs incurred today. Assuming correct expectations about the future (an assumption to be examined critically later), the innovator will innovate at the earliest when those magnitudes come into equality, and perhaps (given recognition lags) not until $V(T) > RD_o e^{-(a+r)T}$.

This perspective on the innovation inducement process and important variants thereof is illustrated in Figs. 1–3. Figure 1 replicates geometrically the model of the preceding paragraph. The solid downward-sloping line labeled 'R&D Cost' shows $RD_o\, e^{-(a+r)T}$ declining exponentially from its initially high value as the time of innovation is delayed from the earliest possible date, year zero. The upward-sloping line labeled 'Benefits' shows the discounted quasi-rents from innovation $V(T)$ rising steadily as the time of innovation is delayed and the innovator taps into revenue-earning possibilities augmented by the

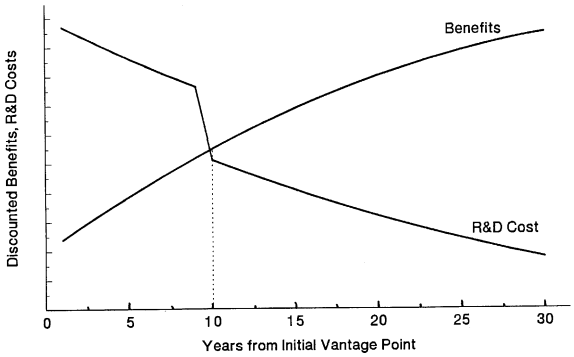

Figure 2
Innovation induced by technology-push

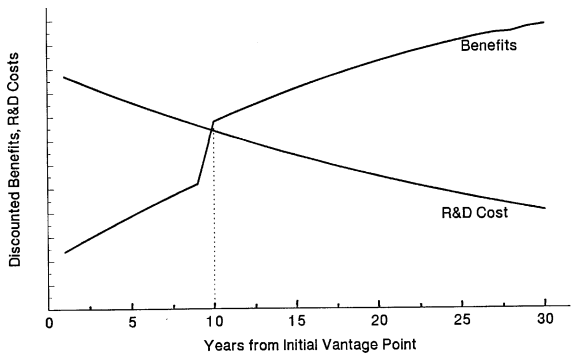

Figure 3
Innovation induced by demand-pull

growth of demand. As discounted R&D costs fall and discounted quasi-rents rise, there comes a time (14 years from the initial vantage point) at which the innovation first becomes profitable. After year 14, the innovation becomes more and more profitable, so the lure of profit becomes increasingly irresistible, making it virtually inevitable that some entrepreneur will rise to the bait and carry out the innovation. Indeed, as time advances and discounted benefits rise relative to discounted R&D cost, increasing numbers of would-be innovators are likely to recognize the opportunity, precipitating more or less simultaneous innovation by two or even more firms—a phenomenon characterized earlier by the sociologist Ogburn (Ogburn and Thomas 1922) as a response to the 'status of the material culture' (see also Merton 1961).

Nevertheless, the road to technological innovations is not always smooth, as portrayed in Fig. 1. Advances in science may occur discontinuously when there are scientific breakthroughs. This case is illustrated in Fig. 2. Up to year 9, the discounted R&D costs of innovating are well in excess of foreseeable discounted quasi-rents, so innovation is not profitable. But in year 9 a scientific breakthrough occurs, substantially reducing the cost of carrying out the developmental work needed to implement the innovation. Innovation is suddenly profitable in year 10 and becomes increasingly profitable as the knowledge base resumes its steady growth thereafter. An innovation induced in this way is called a 'technology-push' or 'science-push' innovation.

The other extreme case is portrayed in Fig. 3. Again, the innovation is clearly unprofitable up to year 9. But in that year demand takes an upward jump—e.g., because an energy shock has suddenly made energy-saving innovations attractive—and the innovation becomes profitable, the more so, the more demand continues to grow if the innovation is not undertaken immediately in year 10. Following Schmookler (1966), innovations induced in this way are called 'demand-pull' innovations.

Reality is some blend of these three polar cases. For Figs. 2 and 3, there is a clear one-sided inducement mechanism—either technology-push or demand-pull. More frequently, both demand-side stimuli and the knowledge base are changing, and it is difficult to tell which has the more powerful influence. Just as both blades of a scissors cut paper, both technology-side and demand-side changes contribute to making technological innovations profitable and thus induce (with some stochastic lag) their appearance on the market.

3. *The Appropriability Problem*

An important complication entails what has come to be called the appropriability problem. Not all of an innovation's incremental value accrues to the innovator as quasi-rents or producer's surplus. Some of the value cannot be appropriated by the innovator, but accrues instead to consumers or other producers. This happens for three main reasons. First, even when the innovator commercializes a new product under monopoly conditions, it is difficult to devise a pricing scheme that does not confer new consumers' surplus upon those who purchase the product. (Only in the implausible case of perfect price discrimination could the innovator appropriate all of the surplus from a product innovation.) Second, if multiple firms innovate more or less simultaneously or if imitation is swift, competition may drive the relevant product's price below the monopoly profit-maximizing level, increasing the amount of surplus accruing to consumers and/or competitors and reducing the amount appropriated by innovators. Third, the very fact that a successful innovation has been achieved conveys, or 'spills over,' valuable information to other economic actors (see *Social Facilitation, Psychology of*; *Information, Economics of*). They can study the innovator's technical solutions and use the insights gleaned thereby to come up with their own innovations, which

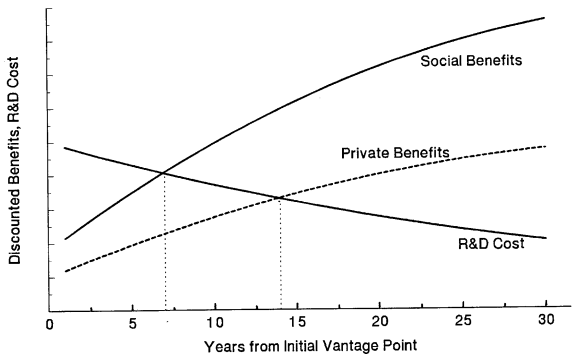

Figure 4
Divergence between social and private benefits

may either improve upon the original innovation's characteristics or provide differentiated new features. Empirical studies (Griliches 1992) have documented extensive spill-overs whose value is not appropriated by innovators.

The most influential quantitative research on the appropriability problem was by Mansfield et al. (1977). For a sample of 17 commercialized innovations, they painstakingly estimated the quasi-rents realized by the original innovator (called the private return) and the surpluses spilling over to consumers, imitators, and subsequent technology users. Converting these estimates to rates of return on the innovator's R&D investment, they found the median private return on R&D investment (i.e., the innovator's return) to be 25 percent, while the median social return (counting in addition surpluses accruing to consumers, rivals, and others) was 56 percent.

The implications of the appropriability problem are illustrated in Fig. 4. The innovator's inability to appropriate all the surplus from their innovation in effect inserts a wedge between the stream of benefits appropriable by the innovator (the private benefits) and the total benefits realized by all actors in the economy (the social benefits). The private benefits function in Fig. 4 is the same as in Figs. 1 and 2; the social benefits function is shifted upward by 80 percent. Whereas the innovation first becomes profitable to a private firm 14 years from the initial vantage point, social benefits begin to exceed R&D cost after year 7. This does not mean that the innovation should be introduced in year 7. An all-knowing decision-maker seeking an ideal solution from the perspective of all participants in the economy would choose an introduction date that maximizes the discounted surplus of social benefits less R&D costs. That date will be the same as the private sector break-even date (14 years) when $k = (a+r)/r$, where k is the ratio of social to private benefits (shown by Mansfield et al. (1977) to have a median value of approximately 2.25), a is the rate of R&D cost decline, as above, and the market growth rate g is assumed to be zero. For a proof, see Scherer (1980). If appropriability is so low that $k > (a+r)/r$, the private break-even date comes later than the socially optimal innovation date; if appropriability is sufficiently high that $k < (a+r)/r$, private break-even may precede the socially optimal innovation date.

Three main classes of remedies have been devised to lessen the innovation market failures believed to come from insufficient appropriability of innovation benefits by innovators.

A solution of particularly long standing, dating back to Venice in the Middle Ages, is the patent system. The first person to make an invention (or under some US interpretations, reduce it to practice) is granted by the government a temporary monopoly on the commercial exploitation of that invention. Patents delay the incursion of imitators and hence help innovators appropriate a larger share of the benefits from their innovations. Surveys of R&D decision-makers (Levin et al. 1987) reveal that the effectiveness of patents in achieving this result varies widely across industries. One reason for this variability is the fact that in some product lines, technological innovators enjoy substantial timing and reputation advantages over imitators even when no patent protection can be secured (Robinson et al. 1994).

Second, insufficient appropriability has been viewed as a rationale for government subsidies to support research and development investments. The relevant policy instruments range from corporate income tax credits for incremental funds spent on R&D (implemented first by Canada in 1962) to the conduct of R&D by government agencies or, more commonly, the issuance by government of contracts reimbursing more or less fully the R&D costs incurred by private firms. The contracts approach suffers from moral hazard problems and possible errors in the choice of innovators, but it is virtually unavoidable when large R&D outlays must be risked to develop high-technology products such as guided missile systems for government use. The vast array of policies chosen by governments to foster innovation in their home markets has been studied by economists under the rubric, 'national innovation systems' (Nelson 1993).

Third, firms have sought to alleviate appropriability problems by entering into joint ventures to conduct R&D and perhaps also to market the innovations achieved thereby. The cooperative R&D approach offers an added advantage of minimizing what might otherwise be costly duplication of more or less identical R&D efforts. Concomitant disadvantages include the lessening of diversity among technical approaches, where diversity enhances the prospects of achieving a successful solution, and reducing the competitive pressures that compel firms to conduct their R&D programs aggressively (see, e.g., Klein 1977).

How much competition is desirable in the pursuit of innovations is a question on which much economic research has been done (see Reinganum 1989). There are conflicting tendencies. On the one hand, competition forces firms to be aggressive and imaginative in their R&D efforts, seizing innovation opportunities at the earliest feasible moment (e.g., at year 14 in Fig. 1). On the other hand, too much competition reduces appropriability and may lead to a market failure under which no firm can anticipate sufficient quasi-rents to make its R&D investments worthwhile. This tension has led to what is called the 'inverted U' hypothesis, holding that the polar extremes of no competition at all (e.g., a secure monopoly) and intense competition are less conducive to innovation than an intermediate degree of competition, e.g., under differentiated oligopoly (see, e.g., Baldwin and Scott 1987).

4. *The Risks and Costs of Innovation*

Innovation is intrinsically an excursion into uncharted waters. The notion that costs and benefits can be predicted with precision, as assumed in the model underlying Figures 1–4, abstracts from the fog of real-world decision-making. Clearly, substantial uncertainty and risk are present. However, in early discussions concerning the economics of innovation, little was known about how much uncertainty actually pervaded innovation decisions. Mansfield made the most important clarifying contributions. In one of several such studies, Mansfield et al. (1977) obtained from the heads of 16 chemicals, pharmaceutical, electronics, and petroleum company R&D laboratories three probability estimates for the portfolio of R&D projects pursued by those laboratories: the probability that an average project would achieve its original technical goals, the probability that, conditional upon technical success, the resulting product or process would be introduced commercially, and the probability that, conditional upon commercialization, the project would yield a return on investment at least as high as the opportunity cost of the firm's capital. For all 16 laboratories combined, the average conditional probabilities were as follows:

Technical success	0.57
Commercialization, given technical success	0.65
Financial success, given commercialization	0.74

Cumulating the three probabilities by multiplication, one finds that on average 27 percent, or roughly one in four, of the projects initially undertaken eventually led to financial success as defined by Mansfield et al. (1977). To make the innovative process broadly worth while, the financial rewards from that minority of successes had to repay not only their own investments but also those made on projects that died at pre-commercialization stages. More recent research (Scherer and Harhoff 1999) has revealed that even among commercialized innovations, the size distribution of rewards (measured in absolute, not rate of return, terms) is highly skew, conforming most closely to a log normal distribution law, and that failure to achieve at least normal returns on commercialized projects may be more common than the survey by Mansfield et al. (1977) implied. The greater uncertainties appear to lie in determining the reaction of consumers and competitors to new products than in the ability of scientists and engineers to achieve success in the laboratory. So skew is the distribution of rewards that it is difficult to make profits converge toward fairly stable averages by supporting feasibly large project portfolios.

The ability to pursue portfolio policies also depends upon the sizes of individual innovation projects. There too, evidence of considerable skew exists. The median case in a large sample of prize-winning technological innovations entailed R&D outlays on the order of $2 million (Scherer 1999). However, the project size distribution has a long thin tail, including hundred-million-dollar outlays to achieve successful new pharmaceutical innovations in the United States and multi-billion outlays for new airliners. In such extreme cases, 'bet the company' risks are difficult to avoid.

How large business enterprises must be to undertake the risks of privately financed innovation has been a subject of continuing debate. Especially in Europe, the dominant belief for many decades following World War II was that large national champion enterprises were better positioned than smaller firms to sustain innovative activities. Evidence from the United States has tended to be inconsistent with this view, in part because there are thousands of companies sizeable enough to support portfolios containing numerous median-outlay innovation projects without risking financial failure. Also, beginning in the United States but spreading during the 1990s to many industrialized nations, there has been a proliferation of venture funds pooling money from wealthy investors who seek the occasional but spectacular returns that can be realized by backing an unusually successful high-technology startup firm.

See also: Innovation: Organizational; Innovation, Theory of; Marx, Karl (1818–89); Schumpeter, Joseph A (1883–1950); Technological Innovation; Technology Assessment

Bibliography

Baldwin W L, Scott J T 1987 *Market Structure and Technological Change*. Harwood Academic Publishers, Chur, Switzerland

Barzel Y 1968 Optimal timing of innovations. *Review of Economics and Statistics* **50**: 348–55

Griliches Z 1992 The search for research-and-development spillovers. *Scandinavian Journal of Economics* **94**: 29–47
Griliches Z 1995 R&D and productivity. In: Stoneman P (ed.), *Handbook of the Economics of Innovations and Technological Change*. Basil Blackwell, Oxford, UK
Klein B H 1977 *Dynamic Economics*. Harvard University Press, Cambridge, MA
Landes D S 1969 *The Unbound Prometheus: Technology Change and Industrial Development in Western Europe from 1750 to the Present*. Cambridge University Press, London
Levin R C, Klevorick A, Nelson R R, Winter S G 1987 Appropriating the returns from industrial research and development. *Brookings Papers on Economic Activity: Microeconomics* 783–820
Marx K 1887 *Capital* (English-language edition by Frederick Engels). Swan Sonnenschein, Lowry & Co., London
Mansfield E, Rapoport J, Schnee J, Wagner S, Hamburger M 1971 *Research and Innovation in the Modern Corporation*, 1st edn. Norton, New York
Mansfield E, Rapoport J, Romeo A, Wagner S, Beardsley G 1977 Social and private rates of return from industrial innovations. *Quarterly Journal of Economics* **91**: 221–40
Merton R K 1961 Singletons and multiples in scientific discovery. *Proceedings of the American Philosophical Society* **105**: 370–86
Mokyr J 1990 *The Lever of Riches*. Oxford University Press, New York
Nelson R R (ed.) 1993 *National Innovation Systems: A Comparative Analysis*. Oxford University Press, Oxford, UK
Ogburn W F, Thomas D S 1922 Are inventions inevitable? *Political Science Quarterly* **37**: 83–98
Organisation for Economic Cooperation and Development 1997 *Basic Science and Technology Statistics*. Paris
Reinganum J F 1989 The timing of innovation: research, development, and diffusion. In: Schmalensee R, Willig R D (eds.) *Handbook of Industrial Organization*, North-Holland, Amsterdam, Vol. 1, pp. 849–908
Robinson W T, Kalyanaram G, Urban G L 1994 First-mover advantages from pioneering new markets. *Review of Industrial Organization* **9**: 1–24
Scherer F M 1965 Invention and innovation in the Watt–Boulton steam engine venture. *Technological and Culture* **6**: 165–87
Scherer F M 1967 R&D resource allocation under rivalry. *Quarterly Journal of Economics* **81**: 359–94
Scherer F M 1980 *Industrial Market Structure and Economic Performance*, 2nd edn. Rand McNally, Chicago
Scherer F M 1999 *New Perspectives on Economic Growth and Technological Innovation*. Brookings Institution, Washington, DC
Scherer F M, Harhoff D 2000 Technology policy for a world of skew-distributed outcomes. *Research Policy* **29**: 559–66
Schmookler J 1966 *Invention and Economic Growth*. Harvard University Press, Cambridge, MA
Schumpeter J A 1912 *Theorie der Wirtschaftlichen Entwicklung*. Duncker & Humblot, Leipzig (trans. Opie R 1934 *The Theory of Economic Development*. Harvard University Press, Cambridge, MA)
Schumpeter J A 1942 *Capitalism, Socialism, and Democracy*. Harper, New York
Usher A P 1954 *A History of Mechanical Inventions* (rev. edn.). Harvard University Press, Cambridge, MA

F. M. Scherer

Innovation: Organizational

Introduction

> There is nothing more difficult to plan, more doubtful of success, nor more dangerous to manage than the creation of a new order of things …
> Niccolo Machiavelli, *The Prince* 1961 [1513]

Despite the attention paid to innovation in the 1990s, which is evident in the increased number of books and articles, the theory itself seems to have changed little since the 1970s. Drazin and Schoonhoven (1996) concluded that research on innovation still has an adaptationist perspective. Its three basic assumptions are '(a) innovation is universally desirable for organizations, (b) once an organization increases its size beyond a critical mass it becomes more inert, less capable of meaningful organizational change, and only haltingly proficient at innovation, and (c) certain structures and practices can overcome inertia and increase the generation rate of innovation.' While this assessment may be accurate overall, it fails to take into account that the development of multilevel models and associated sophisticated statistical modeling and analysis techniques may ultimately lead to incremental or even radical change in innovation theories.

This article reviews the current state of research on innovation and points out pathways for new insights where researchers incorporate findings from learning and institutional theories into multilevel innovation models. In particular, diffusion of innovation models has much to gain from combining traditional trait with context theories and from modeling effects simultaneously. Incorporating more sophisticated theories of power and organizational memory may further revitalize research on generating and managing innovation in organizations. It is important to keep in mind that the importance of innovation for organizational success and survival may be greatly exaggerated or even mythical as von Braun's (1997) analysis of Research and Development (R&D) activities in highly industrialized nations shows. Indeed, he calls the focus on industrial innovation during the post-Cold War the *Arms Race of the '90s*.

1. Definition

Definitions vary in the extent to which they distinguish between that which is objectively or subjectively new. Some authors call only successful commercial exploitation of new ideas an innovation, while others do not distinguish between a novel idea as an invention and commercial innovation (see *Entrepreneurship*). For the latter any 'idea, practice, or object that is perceived as new by an individual or other unit of adoption … is an innovation' (Rogers 1995). Even if one were to

distinguish strictly between invention and innovation, the question to whom something is new or not, still remains. Is the exploitation of a new idea only an innovation once, when it is first discovered? Or is the pacemaker still an innovation in Germany, even though it is already widely being used in Australia or the US? Rogers (1995) prefers a broad understanding of the concept. He terms any idea or practice that is new to an adopter as an innovation. Thus, even where a practice such as activity-based cost accounting (ABC) is used within an industry, for a company that is newly adopting ABC, ABC is an innovation.

2. *Types of Innovations*

Innovations can be classified along a variety of dimensions. The most common typology distinguishes among product, process, and, more recently, also service innovations. Product innovations refer to changes in the offerings of an organization, for example, the introduction of CD and later DVD players. Process innovations refer to changes in the manufacturing or delivery of such offerings, for example, Just-in-Time Manufacturing or Total Quality Management (TQM)/(JIT). The distinctions between product and process innovations, however, seem to become blurred when it comes to service innovations, since product and process go together in the production and delivery of a service. Especially among the process innovations one could also distinguish between technological and administrative procedures. Does the change involve a new technology of transforming input into output or does it involve a new way of administering or organizing the process using existing technologies?

Another typology (e.g., Anderson and Tushman 1990) classifies innovations by the extent to which they rely on existing competencies or change existing products or processes. At one extreme are incremental, at the other extreme are radical innovations. Radical innovations are those that lead to hitherto unheard of products or processes. For example, enlarging the capacity of $5\frac{1}{4}$ inch disks to provide 20 MB of memory involved incremental changes, while developing the $3\frac{1}{2}$ inch disks represented a radical innovation. Radical innovations often involve competency-destroying changes and are associated with organizational reorientation (cf. Tushman and Rosenkopf 1996).

3. *Models of Innovation*

Brown and Eisenhardt's (1995) review distinguishes between rational plan, communication web, and disciplined problem-solving approaches to product development. The main difference among the various models concerns the underlying assumptions about the overall process, that is, whether it is rationally planned or evolves over time.

3.1 *Rational Processes*

Rational models (see *Rational Choice and Organization Theory*) assume that innovation proceeds along a strategic planning process involving information gathering, analysis and evaluation, and action. According to the rational perspective successful innovation is 'the result of (a) careful planning of a superior product for an attractive market and (b) the execution of that plan by a competent and well-coordinated cross-functional team that operates with (c) the blessings of senior management' (Brown and Eisenhardt 1995). Practitioner-oriented prescriptions sometimes list up to 16 steps in the strategically planned, rational product development process ranging from initial screening to detailed market studies, trial production, and ultimately market launch (Cooper 1993). Thus, it seems that the rational process is highly complex, especially if one were to consider turbulent or dynamic environments in the market assessments. Nonetheless, rational models can be found in many organizations and have inspired a large number of research projects. Summarizing research on successful and failed product innovations that had followed a rational plan Brown and Eisenhardt (1995) list product advantages, market attractiveness, and internal organization as the main factors for successful, rational product innovation.

3.2 *Evolutionary Models*

Underlying evolutionary theories is a repeated variation-selection-retention cycle. Evolutionary theories assume that variation—different ways of doing something—exists within any environment and that managers then selectively retain a product or process from such variation and thereby innovate within their organization. For example, organizations may selectively retain an administrative innovation such as idiosyncratic jobs (Miner 1991) and thereby change their processes and products over time.

3.3 *Linking Organizational and Industry Innovation*

Most innovation models work on two levels. At the organizational level, innovation is a means of organizational adaptation to changing environments. However, organizational innovation can also be a means of industry evolution (e.g., Anderson and Tushman 1990; Greve et al. 1995). In fact, the link between organizational innovation and industry evolution follows a well-established research tradition ranging from waves of creative destruction models (Schumpeter 1934) to organizational learning (see *Learning*: *Organizational*) and population level learning (Miner and Haunschild 1995). Even theories, such as those of population ecology, whose focus is on the population rather than individual organization level

entities, are innovation models at heart. Population ecologists argue that some new organizational forms are more likely to survive than others (e.g., Hannan and Freeman 1989) and thereby ultimately lead to innovation within a population of organizations. The linkage between organizational innovation and industry evolution is particularly evident in punctuated equilibrium and technology cycle theories (see *Technology and Organization*).

3.4 Punctuated Equilibrium and Technology Cycle Models

Proponents of punctuated equilibrium theory suggest that organizations 'evolve through alternating periods of convergence and reorientation' (Tushman and Rosenkopf 1996), that is, alternating periods of incremental and radical innovations. Instability and reorientation punctuate stable, convergent periods. Environmental change stimulates technological breakthrough innovations which disrupt stable patterns of interaction and power relations, and thereby lead to organizational or industry-wide reorientation and divergence. Empirical studies have shown that reorientation and turbulence may stem from technological, legislative or wartime jolts, or performance crises and cannot be predicted. After the implementation of radical innovations in the reorientation periods, incremental innovations dominate the convergence periods. These convergence periods can vary dramatically in length.

Technology cycle theories (e.g., Anderson and Tushman 1990) similarly point to alternating periods of radical and incremental change, but mainly operate at the industry level. Tushman and co-workers show that periods of increased discontinuities in an industry alternate with periods of increased fermentation. Periods of discontinuities experience design competition and substitution, while periods of incremental change follow the emergence of a dominant design and are characterized by elaboration of dominant designs. Decisions about a dominant design are not necessarily technology-driven but involve social or political processes which 'abjudicate among multiple technological possibilities' (Anderson and Tushman 1990). Thus, the ultimate dominant design, 'is an outcome of the social or political dynamics of compromise and accommodation between actors or unequal influence.'

Tushman and Anderson (1986) use data from the minicomputer, cement, and airline industries to study technological innovation in those industries. When they differentiate between competence-enhancing and competency-destroying technologies, they notice an interesting pattern. Apparently, competence-destroying technologies stem from new players, that is, influences exogenous to the industry, while competence-enhancing technologies are developed by existing firms, that is endogenous forces. Thus, certain variation may result from environmental change, while other variation grows out of seeds planted within the industry.

4. Diffusion of Innovation

The heterogeneous diffusion model by Strang and his co-workers (Greve et al. 1995) brings together two distinct conceptual models of diffusion processes and allows simultaneous analysis of multilevel effects. While Strang and his co-workers model organizational and network influences, the model itself could also be applicable to reconcile research streams that focus on specific traits of innovations with those that focus on contextual variables, either at the individual adopters, the organizational, or the population level.

4.1 Trait Models

One stream within the diffusion of innovation literature focuses on the traits of particular innovations. Rogers's (1995) five-factor model is a well-known example. The five factors are relative advantage, compatibility, complexity, trialability, and observability. Relative advantage refers to 'the degree to which an innovation is perceived as being better than the idea that it supersedes' (Rogers 1995). Compatibility is the 'degree to which an innovation is perceived as consistent with the existing values, past experiences, and needs of potential adopters' (Rogers 1995). Some innovations are very simple and easy to implement, while others are complex, consist of several inter-related subprocesses, and may be difficult to implement in an organization. Rogers (1995) refers to this as the complexity of an innovation, which he defines as the 'degree to which an innovation is perceived as relatively difficult to understand and use.' According to Rogers, managers are more likely to adopt an innovation after they had a chance to experiment with the innovation. Rogers (1995) refers to this characteristic as trialability and defines it as 'the degree to which an innovation may be experimented with on a limited basis.' Last, but not least, being able to observe the experiences of another organization or individual also has positive effects for the diffusion of an innovation, a concept that Rogers calls observability.

4.2 Institutional Models

Another stream of diffusion research is framed within institutional concepts. This research stream suggests that managers not only consider the traits of an innovation, but also the context. As much of the neo-

institutional research has shown, managers do not only consider the instrumental outcomes of a practice, but also the noninstrumental outcomes such as broad, increased legitimacy and isomorphism with the generally accepted practices of the industry or population of organizations (cf. Powell and DiMaggio 1991).

One of the major claims of institutional theorists is that organizational survival depends on legitimacy defined as 'a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions' (Suchman 1995). Managers have an important role in building, increasing, and maintaining organizational legitimacy. For example, they may only adopt innovations that are legitimate and enhance the organization's legitimacy. Empirical research shows that managers copy innovations from large organizations or from organizations personally known to them. Moreover, managers are likely to implement innovations that have been endorsed by exogenous institutions or have become part of professional knowledge (cf. Powell and DiMaggio 1991).

Adapting Strang and co-workers' multilevel heterogeneous diffusion model may lead to new insights into the diffusion of innovation research by distinguishing the relative impact of traits and organizational or population contexts. Using such a model and the associated methodologies would also allow the incorporation of individual decision-makers' influences into the modeling, for instance, their preferences and personalities such as risk-taking or risk avoiding (see *Organizational Decision Making*).

5. *Enhancing Innovation in Organizations*

The literature on facilitating innovation in organizations falls into two streams. One stream looks at introducing and managing processes that might lead to innovation, while the other examines specific factors that may enhance or impede the innovativeness of an organization. Among those innovativeness factors one could further distinguish between factors that affect the development of particular products and those that operate on the organization as a whole. Lately, the relationship between innovation and strategy (see *Strategy*: *Organizational*) has been emphasized, for instance, in research on the effect of mergers and acquisitions, exiting certain niches or transforming an inert mature organization into an innovative one (see *Organizational Change*).

Rothwell (1994) has written comprehensive reviews on success factors, strategies, and trends for innovativeness. For example, he lists good internal and external communication, effective linkages with external sources of know-how, treating innovation as a company-wide task, high quality production, careful planning and control systems, strong market orientation, top management commitment, and long-term commitment to major projects among other critical elements of successful innovation strategies (see *Entrepreneurship*). Dougherty and Hardy (1996) reinforce Rothwell's (1994) findings, but also add the importance of managing power in product development. Going beyond the effect of powerful individuals such as powerful product champions and powerful team leaders, they introduce multi-dimensional views on power (see *Organizations*: *Authority and Power*).They suggest that generating successful sustainable innovation involves managing power on all dimensions, that is, managing resources, processes, and meaning.

Knowledge and learning theories provide further insights. In particular, applying organizational memory concepts seems fruitful. For example, Moorman and Miner (1998) consider the effect of procedural and declarative memory on improvization where improvization represents one form of creative activity in product development teams. Many novel designs consist of a recombination of existing parts, processes, and routines, that is, they result from improvization based on a recombination of knowledge stored in memory.

6. *Critical Views on Innovation*

Lately, critical views questioning the generally assumed adaptive capacity of innovation have emerged. For instance, von Braun (1997) writes that increased R&D spending in OECD countries has not necessarily led to progress, but simply to change. According to him some ancient wind-powered drainage systems may be characterized as more progressive or advanced than today's diesel or electrically-driven pumps in regard to efficient energy utilization. He especially draws attention to the *Acceleration Trap* and *R&D Spiral*. While his analyses are somewhat general, it is interesting to read that 'it proved difficult, if not impossible to discover a positive relationship between R&D growth and sales growth. This held true even if one assumed a delayed effect of R&D spending increase on sales growth' (von Braun 1997). Contrary to the generally portrayed positive image of accelerating product development and then managing product releases carefully, von Brown points out that the relationship between accelerated product development and shortened product life cycles is highly complex in competitive environments. Von Braun suggests unanticipated negative impacts on overall organizational revenue and survival and finishes his book with recommendations on how to avoid falling into the acceleration trap.

See also: Development: Organizational; Diffusion, Sociology of; Evolution: Diffusion of Innovations; Innovation, Theory of; Technological Innovation; Technology and Organization; Technology Transfer

Bibliography

Anderson P, Tushman M L 1990 Technological discontinuities and dominant designs: A cyclical model of technological change. *Administrative Science Quarterly* **35**: 604–33

Brown S L, Eisenhardt K M 1995 Product development: Past research, present findings, and future directions. *Academy of Management Review* **20**: 343–78

Cooper R G 1993 *Winning at New Products*, 2nd edn. Addison-Wesley, Reading, MA

Dougherty D, Hardy C 1996 Sustained product innovation in large, mature organizations: Overcoming innovation-to-organization problems. *Academy of Management Journal* **39**: 1120–53

Drazin R, Schoonhoven C B 1996 Community, population and organization effects on innovation: A multilevel perspective. Special issue on innovation. *Academy of Management Journal* **39**: 1065–83

Greve H R, Strang D, Tuma N B 1995 Specification and estimation of heterogeneous diffusion models. *Sociological Methodology* **25**: 377–420

Hannan M T, Freeman J 1989 *Organizational Ecology*. Harvard University Press, Cambridge, MA

Machiavelli N 1961 [1513] *The Prince*. Penguin, Baltimore

Miner A S 1991 Organizational evolution and the social ecology of jobs. *American Sociology Review* **56**: 772–85

Miner A S, Haunschild P R 1995 Population level learning. *Research in Organizational Behavior* **17**: 115–66

Moorman C, Miner A S 1998 Organizational improvization and organizational memory. *Academy of Management Review* **23**: 698–723

Powell W W, DiMaggio P J 1991 *The New Institutionalism in Organizational Analysis*. The University of Chicago Press, Chicago

Rogers E M 1995 *The Diffusion of Innovation*, 4th edn. Free Press, New York

Rothwell R 1994 Industrial innovation: Success, strategy, trends. In: Dodgson M, Rothwell R (eds.) *The Handbook of Industrial Innovation*. Elgar, Aldershot, UK, pp. 33–53

Schumpeter J 1934 *The History of Economic Development*. Harvard University Press, Cambridge, MA

Suchman M C 1995 Managing legitimacy: Strategic and institutional approaches. *Acaedmy of Management Review* **20**: 571–610

Tushman M L, Anderson P 1986 Technological discontinuities and organizational environment. *Administrative Science Quarterly* **31**: 439–65

Tushman M L, Rosenkopf L 1996 Executive succession, strategic reorientation and performance growth: A longitudinal study in the US cement industry. *Management Science* **42**: 939–53

Von Braun C-F 1997 *The Innovation War*. Prentice Hall, Upper Saddle River, NJ

T. Rura-Polley

Innovation, Theory of

This article draws on the model of the diffusion of innovations in order to understand how social change occurs at the microlevel. Since the early 1940s, a large set of social scientists have investigated how new ideas, often technological in nature, spread among various types of audiences. Diffusion research provides improved understanding of technological determinism, defined as a belief that technology causes social changes in a system. The diffusion process also illuminates the social process of social determinism, defined as a belief that social forces like government regulations, social values, and human interaction shape technologies through a process of social construction. Research on the diffusion process shows the essentially social nature of the adoption of new ideas. Innovations spread in a population through a process of people talking to others. Such interpersonal communication facilitates a wide sharing of the meanings of a new idea, often leading to its adoption and use by the members of a system, but sometimes resulting in rejection of the innovation.

Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system (Rogers 1995). Diffusion is a special type of communication process concerned with the spread of new ideas, messages that necessarily represent a high degree of uncertainty to the individuals involved. The diffusion of innovations has been studied by sociologists, political scientists, anthropologists, social psychologists, economists, and communication scholars, and by researchers in public health, education, business management, marketing, and other fields.

For instance, anthropologists might study the consequences of a technological innovation on the culture of its users (Pelto 1973). Political scientists investigate the process through which new government programs and policies spread among nations, states, or cities (Walker 1966). Marketing scholars explore how a new product or service is sold to customers by business companies (Moore 1991). Communication scholars study the diffusion of a major news event like the Challenger disaster or a presidential assassination (Deutschmann and Danielson 1960). These varied fields share a common conceptual model of the diffusion of innovations. Few other research topics have received such multidisciplinary attention as has the diffusion of innovations (Rogers 1995).

1. Perceptions of Innovations

An innovation is defined as an idea, practice, or object that is perceived as new by an individual or other unit of adoption. This definition is very broad, in that almost any kind of new idea, ranging from a new medical drug, to a new food, to the Internet, can be perceived by the members of a social system as an innovation. Most innovations that have been investigated in the some 5,000 diffusion studies to date are technological in nature.

The definition of an innovation as an idea perceived as new is consistent with the theoretical perspectives of W. I. Thomas and George Herbert Mead of the

Chicago School, which flourished in the early decades of the twentieth century. Thomas's dictum, that a situation is real to an individual if it is so perceived, suggested that perceptions count in understanding human behavior change. How an innovation is perceived on five dimensions determines its rate of adoption (as discussed later). The symbolic interactionist perspective based on Mead's theory is also involved in the diffusion of innovations model. Initially, individuals perceive an innovation with a high degree of uncertainty; they want to know how it works, whether it is safe or risky to use, where to obtain it, and its advantages and disadvantages. These uncertainties are gradually overcome as an individual obtains technical information, subjective impressions, and social support concerning the new idea.

One or more individuals in a system must venture first in adopting a new idea. These innovators are open to taking risks, are oriented outside of their local system, and have sufficient resources so that the uncertainty of the innovation is not an important barrier to its adoption. When the innovators in a system adopt a new idea, they create widespread awareness–knowledge of it, although most members of the system continue to doubt its validity until the early adopters (who adopt following the innovators), including many opinion leaders in the system, begin using the new idea. At this point, the norms of the system change, so as to support adoption of the innovation. Gradually, once a critical mass is reached in the rate of adoption of an innovation, more and more individuals adopt the new idea. This diffusion process occurs over time as those who have already adopted talk about the innovation with other individuals who adopt, and who then tell yet others about the new idea. Subjective opinions of the innovation are passed via word-of-mouth channels, as individuals who have adopted tell potential adopters about the new idea.

2. *Where Innovations Come From*

A distinction should be made between innovation and invention, defined as the process in which a new idea is discovered or created. In contrast, innovation occurs when a new idea is adopted or rejected. So invention and innovation are two different processes, although both deal with a new idea (Rogers 1995, p. 135).

An innovation, especially a technological innovation, results from a process in which a need or problem is identified, research is conducted to invent a solution, and this invention is then developed and commercialized into a product or service that is sold in the marketplace. Development of an innovation is the process of putting a new idea in a form that is expected to meet the needs of an audience of potential adopters. So the research and development process represents a transformation of an invention into a form that will be more acceptable to potential adopters.

A technology is information that is put into use in order to carry out some task (Eveland 1986). Technology transfer is the exchange of technical information between the R&D workers who create a technological innovation and the users of the new idea. Investigations and experience to date suggest that technology transfer is very difficult; many attempts at such transfer fail. A well-known case of failed technology transfer is the attempted conversion of technological innovations like the mouse, bit-mapped display, icons, and the personal computer (that is, a computer designed to be used by an individual) from the Xerox Corporation's R&D center in California into a marketable product. Instead, Apple Computer marketed these technologies as the Apple Macintosh in 1984. Appropriately, the book reporting this case is entitled *Fumbling the Future* (Smith and Alexander 1988).

Why is technology transfer so difficult? One reason is because the R&D workers who invent or develop the technological innovation perceive that it is so advantageous that it will diffuse with little effort. Such overconfidence is seldom justified. The creators of the technology are usually technical experts who fail to perceive how their customers will regard the innovation (Moore 1991). Seldom do technological innovations diffuse themselves; usually an organized effort is needed to secure the adoption of the new idea. After the critical mass is reached in the diffusion process for the new idea, however, further diffusion may be self-sustaining.

3. *Perceived Attributes of Innovations*

The rate of adoption is the relative speed with which an innovation is adopted by members of a social system. The rate of adoption is usually measured as the number of members of the system that adopt the innovation in a given time period. Certain innovations diffuse relatively slower, and other innovations diffuse faster. Why do certain innovations spread more quickly than others? The main characteristics which determine an innovation's rate of adoption are: (a) relative advantage, (b) compatibility (c) complexity, (d) trialability, and (e) observability.

3.1 *Relative advantage*

Relative advantage is the degree to which an innovation is perceived as better than the idea it supersedes. Relative advantage may be perceived in economic terms, or as social prestige, convenience, and satisfaction. Whether an innovation has a great deal of objective advantage is not so important in determining its rate of adoption as whether an individual perceives the innovation as advantageous. The greater the perceived relative advantage of an innovation, the more rapid its rate of adoption.

3.2 Compatibility

Compatibility is the degree to which an innovation is perceived as consistent with the existing values, past experiences, and needs of potential adopters. If an idea is perceived as being incompatible with the values and norms of a social system, it will not be adopted rapidly.

3.3 Complexity

Complexity is the degree to which an innovation is perceived as difficult to understand and put into use. Certain innovations are easily understood by the members of a social system. Others are more complicated and will be adopted more slowly.

3.4 Trialability

Trialability is the degree to which an innovation can be experimented with on a limited basis. Innovations that are divisible for trial represent less uncertainty to the individual, who can learn by trying out the new idea. Such trial often encourages full-scale adoption of the innovation. An early and important study of the diffusion of hybrid seed corn among Iowa farmers found that almost everyone adopted the seed corn after trying out the innovation on a few acres; several years of trial were required for most individuals before they adopted the hybrid seed on 100 percent of their corn acreage (Ryan and Gross 1943).

3.5 Observability

Observability is the degree to which the results of an innovation are visible to others. The visibility of an innovation stimulates peer discussion of the new idea, and thus more rapid adoption.

Of these five perceived attributes of innovations, relative advantage and compatibility have been found to be most important in determining the rate of adoption of an innovation. However, for certain innovations, like consumer electronics products, perceived complexity may be very important.

The diffusion of the Internet has been exceedingly rapid. From 1990 to 2000, the number of individuals adopting the Internet in North America rose from approximately five million to 200 million (some 70 percent of all US households). During the late 1990s, the rate of adoption of the Internet also reached critical mass in Asia, where the diffusion of the Internet reached about 200 million users by 2000. Perhaps the Internet has diffused more rapidly than any other innovation in the history of humankind.

Why has the Internet diffused so rapidly? The idea of computer networking began in 1969, and required 20 years (until 1989) before reaching critical mass. During this period, many unconnected computer networks arose: ARPANET, BITNET, NSFNET, HPNET, and others. The Internet provides a network of interconnected computer networks which allows an individual to communicate with many others at great distances at rapid speed for a very low cost. This 'faster, better, quicker' quality of the Internet leads to a high degree of perceived relative advantage. This innovation is also compatible with the previous idea of writing letters and using computers, so Internet use did not demand learning many new skills. Where the Internet has been incompatible with a system's values, such as in France, where the English language dominance of the Internet and the World Wide Web is incompatible with French values concerning their language and culture, the rate of adoption of the Internet is relatively slow.

4. Preventive Innovations

Certain types of innovations, even though they might be beneficial for society, diffuse relatively slowly. For instance, preventive innovations are new ideas that are adopted in order to avoid some unwanted future consequences. Because the advantages of preventive innovations are uncertain and delayed in time, their rate of adoption is particularly slow. Consider a preventive innovation like HIV/AIDS prevention. Adoption of a preventive innovation like 'safe sex' may help avoid becoming infected with HIV at some future time. But the individual might not have contracted the virus even without adopting safe sex. The rewards of adoption are delayed in time and uncertain as to whether adoption of the innovation was really needed. The unwanted event to be avoided is a nonevent, the absence of an unwelcome situation that otherwise might have happened. Such nonevents are difficult to count. Examples are noninfection with HIV, births averted by adopting family planning methods, and so forth. These behavior changes are low in visibility and difficult for individuals to perceive.

Taboo innovations, defined as new ideas that are perceived as extremely private and personal in nature (Rogers 1995, p. 237), are also difficult to diffuse. For example, national governments in many countries have launched health campaigns to diffuse condom use for the prevention of unwanted births and for HIV/AIDS prevention. In many cultures, peer discussion of condoms is taboo, and so the rate of adoption of this innovation is relatively slow. The taboo nature of certain innovations may be overcome by the use of humor, or by other strategies.

5. The Social System

A social system is defined as a set of interrelated units that are engaged in joint problem solving to accomplish a common goal. A social system may be

composed of units that are individuals, informal groups, organizations, and/or other subsystems. The social system's norms, the established behavior patterns for the members of a social system, have been found to affect the rate of diffusion of an innovation. When the new idea is introduced in the system, the norms may initially be unfavorable. But after opinion leaders, individuals able to influence informally other individuals' attitudes or overt behavior in a desired way with relative frequency, adopt, a 'tipping point' occurs, and the system's norms then encourage further diffusion.

A critical mass is the point at which enough individuals have adopted an innovation so that the innovation's further rate of adoption becomes self-sustaining. The critical mass is one reason why, after a relatively slow start (20 years in the case of the Internet), the rate of adoption of an innovation in a system takes off to form the familiar S-shaped curve. Eventually the rate of adoption of a new idea grows more slowly, and levels off, as fewer and fewer individuals remain to adopt.

Some social systems are characterized by a rapid rate of adoption of technological innovations. An example is a technopolis (technology city) like Silicon Valley in Northern California; Austin, Texas; Seattle; Tsukuba, Japan; Cambridge, UK; and Bangalore and Hyderabad in India. These technology cities are centers for the R&D to produce technological innovations, and flourish economically because of the vast wealth produced by high-tech spin-off companies. A spin-off is a new company that is formed by individuals who were former employees of a parent organization, and a core technology that is transferred from a parent organization. Spin-offs are a means of technology transfer, particularly in such high-technology industries as computers, biotechnology, and Internet-related companies. Each spin-off company uses a technological innovation as its entry ticket into a high-technology industry. If this technology can be commercialized into an innovative product that diffuses rapidly in the marketplace, the new company may expand rapidly, making its founders very wealthy.

An analysis of the economic growth by metropolitan regions in the USA during the 1990s showed that 70 percent of such growth was explained by the rise of high-tech spin-off companies, which in turn was related to the presence of research universities or other R&D facilities (Devol 1999). Thus technological innovation is the key factor in economic development.

See also: Diffusion: Anthropological Aspects; Diffusion: Geographical Aspects; Diffusion, Sociology of; Evolution: Diffusion of Innovations; Innovation and Technological Change, Economics of; Innovation: Organizational; Internet: Psychological Perspectives; Technological Determinism; Technological Innovation; Technology Transfer

Bibliography

Deutschmann P J, Danielson W A 1960 Diffusion of knowledge of the major news story. *Journalism Quarterly* **37**: 345–55

Devol R C 1999 *America's High-Tech Economy: Growth, Development, and Risk for Metropolitan Areas*. Report, Miliken Institute, Santa Monica, CA

Eveland J D 1986 Diffusion, technology transfer, and implementation. *Knowledge* **8**(2): 303–22

Moore G K 1991 *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers*. Harper-Business, New York

Pelto P J 1973 *The Snowmobile Revolution: Technology and Social Change in the Artic*. Cummings, Menlo Park, NJ

Rogers E M 1995 *Diffusion of Innovations*, 4th edn. Free Press, New York

Ryan B, Gross N C 1943 Diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology* **8**: 15–24

Smith D K, Alexander R C 1988 *Fumbling the Future: How Xerox Invented, then Ignored the First Personal Computer*. William Morrow, New York

Walker J L 1966 The diffusion of innovations among the American states. *American Political Science Review* **63**: 880–99

E. M. Rogers

Institutional Economic Thought

1. Introduction

As the term suggests, 'institutional economics' addresses the nature and role of social institutions in economic life. Such institutions include the firm, the state, money, property, and the family. Institutional economists have been concerned with explaining the nature and origin of these institutions, the reasons for their existence and persistence, and the ways in which they can improve economic efficiency. Other important issues faced by institutional economists include the understanding of the dynamics of institutional change, the relationship between institutions, culture and technology, and the comparative efficiency of different institutions in different cultures or contexts.

In fact, the term 'institutional economics' appeared twice in the twentieth century, each time to cover a broad and distinct stream of economic thought. However, the methodological approaches involved in the two streams are quite different.

The more recent and more familiar usage dates from Oliver Williamson (1975). He coined the term 'new institutional economics' to refer to his seminal transaction cost analysis of the nature and structure of the firm. By the end of the century, the term 'new institutional economics' was in widespread use and was associated with a vast literature. Nobel prizes were given to 'new' institutional economists Ronald Coase in 1991, and Robert Fogel and Douglass North in 1993. Coase had developed the original transaction cost concept in 1937 and was an important influence

upon Williamson. Fogel and North had developed quantitative and other methods to analyze institutional change in economic history. Other leading names associated with the 'new' institutional economics are Masahiko Aoki, Mancur Olson, and Richard Posner. There are now prominent economics textbooks dominated by 'new institutionalist' thinking such as those by Paul Milgrom and John Roberts (1992), Andrew Schotter (1994), and Eirik Furubotn and Rudolph Richter (1997).

The unifying project in the 'new institutional economics' is to explain the existence of political, legal, or social institutions by reference to a model of given, individual behavior, tracing out its consequences in terms of human interactions.

Outside economics, the term 'new institutionalism' is used in different and sometimes confusing ways. In particular there is a approach described as 'new institutionalism' in organizational analysis, associated with names such as James March, Walter Powell, and Paul DiMaggio. The methodology and assumptions involved in their works are quite different from those of the 'new institutional economics.'

The original term 'institutional economics' emanates from Walton Hamilton in an address to the American Economic Association in 1918. The 'old' American institutional economics was much influenced by the nineteenth-century German historical school, which emphasized the role of institutions in the economy and stressed that economic analysis had to take account of historically specific circumstances.

The 'old' institutionalism actually became dominant in economics departments in American academia just after the First World War. Seven out of the 20 presidents of the American Economic Association in the years 1924–43 were declared institutionalists: namely Wesley Mitchell, Thomas S. Adams, John Maurice Clark, Frederick C. Mills, Sumner H. Slichter, Edwin G. Nourse, and Albert B. Wolfe. Furthermore, the influence of the 'old' institutionalism persisted for some time even after the Second World War. At least eight institutionalists have been elected President of the American Economic Association since 1945: Calvin B. Hoover, Simon Kuznets, Edwin E. Witte, Morris Copeland, George W. Stocking, Arthur F. Burns, Joseph Spengler, and John Kenneth Galbraith. Elected sympathizers included Frank Knight, Kenneth Boulding, and Robert A. Gordon. In addition, the 'old' institutionalists Simon Kuznets and Gunnar Myrdal received Nobel Prizes in 1971 and 1974, respectively.

Other schools of thought resemble the 'old' institutional economics. For example, in the 1960s and 1970s much of the economics in Cambridge, UK had a strong institutionalist flavor, particularly in the works of Nicholas Kaldor and Joan Robinson. The more recent work of Robert Boyer, Michel Aglietta, and other members of the French *régulation* school also has strong institutionalist affinities.

A prominent and unifying theme of the 'old' institutionalism is that economic analysis had to be grounded on the specific cultural and economic circumstances of each case. In particular, it is held that individual motives and behavior are often molded by such circumstances, including the institutions involved. This contrast with the 'new' institutionalism will be discussed in greater detail below.

Characterizations of the 'old' institutionalism as purely descriptive or antitheoretical do not bear up to close scrutiny. In the writings of Veblen, Commons, Mitchell, Clark, and others there was an emphasis on the importance of theoretical explanation and theoretical development. Whatever their limitations, the early institutionalists addressed crucial theoretical issues. Some of these are discussed in Sect. 3.

Nevertheless, ever since Veblen there has been a failure of the 'old' institutionalists to agree upon, let alone develop, a systematic theoretical core. The 'old' American institutionalism bequeathed no integrated theoretical system of the stature or scope of that of Karl Marx, Alfred Marshall, Léon Walras, or Vilfredo Pareto. The reasons for this failure cannot be discussed here, save to note that it was not because of a naive and unsustainable belief that economics can proceed with data alone, and without any theory. Although several 'old' institutionalists put their faith in data, they all retained some degree of belief in the importance of an ongoing theoretical project.

The primary reasons for the failure of institutionalism lie elsewhere. In particular, the old institutionalism was partially disabled by a combined result of the profound shifts in social science in the 1910–40 period and of the rise of a mathematical style of neoclassical economics in the depression-stricken 1930s. Behaviorist psychology and positivist philosophy displaced the instinct psychology and pragmatist philosophy upon which the early institutionalism had been built. With their use of formal techniques, mathematical economists caught the imagination of both theorists and policy-makers. In comparison, institutionalism was regarded as technically less rigorous, and thereby inferior (Hodgson 1999).

2. *The Essential Distinction Between the 'Old' and the 'New' Institutional Economics*

What is the essential difference between the 'old' and the 'new' institutional economics? Answering this question is made more difficult because there is no unanimity, even among its adherents, as to what is precisely to be included in the 'new' variety. Furthermore, within both the 'new' and the 'old' institutionalism there are differences of view concerning policy matters, such as the place and role of markets. Nevertheless, an answer to this question is possible if we focus on the common theoretical core of some of

the most prominent and influential 'new' institutionalist writings, such as by Douglass North, Richard Posner, Andrew Schotter, and Oliver Williamson. Despite analytical and policy differences, there are some common presumptions behind all their works.

Examining these, it is clear that the characteristic 'new' institutionalist project is the attempt to explain the emergence of institutions, such as the firm or the state, by reference to a model of rational individual behavior, tracing out the unintended consequences in terms of human interactions. An initial, institution-free, 'state of nature' is assumed. The explanatory movement is from individuals to institutions, taking individuals as given. This explanatory approach is often described as 'methodological individualism' (Hodgson 1988).

A much earlier precedent exists for work along these lines. Towards the end of the nineteenth century, the Austrian economist Carl Menger considered the nature and origin of social institutions such as money. He is thus recognized as one of the early forerunners of the 'new' institutional economics. Menger attacked a notion attributed to the German historical school, that institutions could be explained by examining historical data alone.

Menger saw the institution of money as emanating in an undesigned manner from the communications and interactions of individual agents. Traders look for a convenient and frequently exchanged commodity to use in their exchanges with others. Once such regularities become prominent, a circular process of institutional self-reinforcement takes place. Emerging to overcome the difficulties of barter, money is chosen because it is convenient, and it is convenient because it is chosen.

Apart from the emergence of money, other examples in this literature inspired by Menger include driving on one side of the road, and traffic conventions at road junctions. For example, once the majority of car drivers stick to the right-hand side of the road, it is clearly rational for all drivers to follow the same rule. Accordingly, the emergent convention is reinforced and institutionalized by imitation, and by efficient use of the available information. The role of information in signaling information to actors is emphasized in this approach. This important core theme of an action–information loop is clearly evident, for example, in North's (1981) theory of the development of capitalism, Williamson's (1975, 1985) transaction cost analysis of the firm, and Schotter's (1981) game-theoretic analysis of institutions. This core idea is valuable and important.

However, despite its temporal adjective, the 'new' institutionalism is built upon some very old assumptions concerning the human agent, derived from the individualism of the Enlightenment. In this 300-year tradition, a key idea is the notion that the individual can, in a sense, be 'taken for granted.' Accordingly, the individual is taken as the elemental building block in economic theory. Strictly, it is not a question of whether or not a theorist is found to admit that individuals—or their wants and preferences—are changed by circumstances. Indeed, many economists admit that individuals might so be changed. What is crucial is that the individualistic economist assumes, *for the purposes of economic enquiry*, that individuals and their preference functions should be taken as given. Thus the demarcating criterion is not the matter of individual malleability *per se*, but the willingness, or otherwise, to consider this issue as an important or legitimate matter for economic analysis. The commonplace statement by mainstream economists that tastes and preferences are not the *explananda* of economics thus derives directly from the individualist tradition. Likewise, the conception of economics as 'the science of choice' takes the choosing individual and her preference functions as given. Unlike the 'old' institutionalism, the 'new' institutionalism has also taken such individualistic presuppositions on board.

By contrast, a common thread in the literature of 'old' institutionalism, from Veblen through Commons and Mitchell, to Myrdal and Galbraith, is the idea that in economic analysis the individual should not always be taken as given. The general use of given preference functions to model individuals and their behavior is rejected by these institutionalists. As well as individuals interacting to form institutions, individual purposes or preferences are molded by socioeconomic conditions. The individual is both a producer and a product of her circumstances.

Mitchell (1937) argued that the evolution of money cannot be understood simply in terms of cost reduction and individual convenience. He maintained that money 'stamps its pattern upon wayward human nature, makes us all react in standard ways to the standard stimuli it offers, and affects our very ideals of what is good, beautiful and true' (Mitchell 1937, p. 371). Accordingly, the evolution of money changed the mentality, preferences, and way of thinking of individuals themselves. This does not mean that Menger's core idea of an action–information loop is wrong, but that it has to be supplemented by an account of how institutions can change individual perceptions and preferences.

Such ideas pervade and endure through the 'old' institutional economics. They were emphasized by Veblen and Commons. They persist in the 'old' institutionalist tradition, throughout the twentieth century. For instance, writing as an 'old' institutionalist, John Kenneth Galbraith (1969, p. 152) insisted that individual 'wants can be synthesized by advertising, catalyzed by salesmanship, and shaped by the discreet manipulations of the persuaders.' Again we find an emphasis on the malleability of preferences and an abandonment of the notion of a fixed and given individual.

The 'old' institutional economics is thus distinguished from both mainstream economics and

the 'new institutional economics' precisely for the reason that it does not assume an individual with given purposes or preference functions. Instead of a bedrock of given individuals, presumed by the mainstream and new institutional economics, the old institutionalism holds to the idea of interactive and partially malleable agents, mutually entwined in a web of partially durable and self-reinforcing institutions. This distinction holds despite important theoretical and policy differences within both the new and the old institutionalist camps.

3. *The Theoretical Legacy of the 'Old' Institutional Economics*

Although it never provided a systematic theoretical legacy, the achievements and influence of the 'old' institutional economics have been considerable (Rutherford 1994, Yonay 1998).

For example, Veblen (1899, 1919) was the first social scientist to attempt to develop a theory of economic and institutional evolution along essentially Darwinian lines (Hodgson 1993). Veblen's work shares common features with the much later attempts by economists to use evolutionary metaphors from biology by Armen Alchian, Kenneth Boulding, Friedrich Hayek, and Richard Nelson and Sidney Winter. However, Veblen never regarded evolutionary processes as necessarily optimal or progressive. Although he did not use the term, his writings are replete with examples of path dependent evolution, almost a century before that concept became fashionable (Hodgson 1993).

Veblen's writings are brimming with ideas, many of which have subsequently been taken up by others. In the 1970s the Cambridge economist Joan Robinson recognized the importance of his theoretical contribution to the critique of standard capital theory. Veblen (1919, pp. 185–200) criticized the conflation of 'capital goods' with 'capital.' He rejected the 'factors of production' approach in its entirety, seeing production as much to do with 'the accumulated, habitual knowledge of the ways and means involved … the outcome of long experience and experimentation' (Veblen 1919, pp. 185–6). Veblen's emphasis on the role of knowledge and learning in economic growth is perhaps, simultaneously, his most important and most neglected theoretical contribution.

In his *Theory of Business Enterprise* (1904), Veblen discussed not only the conflict between a pecuniary and an industrial culture but also its manifestation in the separation of ownership and control, made famous later by Adolph Berle and Gardiner Means in their highly influential book on *The Modern Corporation and Private Property* (1932). Furthermore, in the same volume, Veblen (1904, pp. 46–8) hinted at the idea of transaction costs, long before Coase and Williamson.

A number of commentators have perceived similarities between the works of Veblen and John Maynard Keynes, and their joint parallels in President Roosevelt's New Deal policies. Writing after the Second World War, James Duesenberry in his *Income, Saving and the Theory of Consumer Behavior* (1949) accepted Veblen's influence in building a theory of the consumption function upon habitual behavior and Harvey Leibenstein also acknowledged the founder of institutionalism with his 'Veblen effects' in the theory of consumer behavior.

John Commons (1924, 1934) has been acknowledged as a major influence on the behavioral economics of Herbert Simon and the 'new' institutionalism of Oliver Williamson (1975, 1985). Overall, Commons made a major theoretical contribution, including the development of the concept of asymmetric information in contracts (Commons 1924). Indeed, Commons pioneered the whole modern research program in economics and law. Relevant modern ideas, such as the notion that the money supply is endogenous rather than exogenous, have also been credited to Commons. Commons and his students were also highly influential in the development of the subdiscipline of labor economics.

Nobel Laureate Herbert Simon's idea of bounded rationality was also prefigured in the writings of the 'old' institutionalist John Maurice Clark (Rutherford 1994). Clark influenced other leading economic theorists, notably Frank Knight. In fact, Clark supervised the production of Knight's classic work *Risk, Uncertainty and Profit* (1921).

Clark also shared responsibility for another important theoretical development. In a paper published in the *Journal of Political Economy* in 1917, he was one of the first to elaborate the idea of the interaction of the multiplier and the accelerator, seeing it as a driving mechanism in the generation of economic fluctuations. Paul Samuelsonand others later replicated this idea.

The Nobel Laureate Milton Friedman was a student of Mitchell. Friedman acknowledged the influence of both Veblen and Mitchell on business cycle theory, and even translated some of Mitchell's theoretical work into a mathematical model.

There are several important links between institutionalism and the development of Keynesianism. One was the innovation of national income accounting, in which the work of Mitchell played a vital part. Inheriting the German historical school view of the economy as an organic whole, institutionalism thereby developed and sanctioned the conceptualization and measurement of economic aggregates. The theoretical and empirical work involved here was of major importance in the development of twentieth-century economics.

Mitchell was thus one of the fathers of modern macroeconomics. His work is notable for its implicit antireductionist thrust and its consequent contribution to the development of Keynesianism. In the 1920s and 1930s, Mitchell and his colleagues in the National Bureau for Economic Research played a vital role in

the development of national income accounting, suggesting that aggregate, macroeconomic phenomena have an ontological and empirical legitimacy. Through the development of national income accounting, the work of Mitchell and his colleagues influenced and inspired the macroeconomics of Keynes.

The 'old' institutionalists also developed a number of theories of pricing behavior in imperfectly competitive markets (Tool 1991). Traces of the surviving influence of 'old' institutionalist ideas are found in many other areas of theoretical and applied economics.

4. *Development of the 'New' Institutional Economics*

Since the coining of the term 'new institutional economics' in 1975, this school of thought has been responsible for an enormous amount of valuable theoretical and empirical work on the nature and development of economic institutions. Institutions that were previously regarded as mysterious 'black boxes' are now opened up and examined in forensic detail. Not only have the 'new' institutionalists put institutions back on the agenda, they have prompted important and stimulating theoretical debates. With important developments such as transaction cost analysis and the economics of property rights, important insights have been gained. The face of economics as a whole has been transformed.

However, within the 'new' institutional economics there are not only important differences of approach, there are also some common and fundamental problems. In this section it is argued that attempts to explain the origin and sustenance of institutions on the basis of the assumption of given individuals have internal flaws and inconsistencies. Accordingly, attempts to explain institutions in this way may have to be abandoned. The door is opened to a more open-ended and evolutionary approach, redolent of the earlier institutionalists.

In broad terms, two opposite types of error are possible. 'Cultural determinists' place too much stress on the molding of individuals by institutions. Some members of the 'old' institutionalist school, notably Clarence Ayres, committed this error. Such 'over-socialized' views of human behavior have been widely criticized. At the opposite end of the spectrum, the 'new institutional economics' gives no more than weak stress to the processes of institutional conditioning, and focuses primarily on the emergence of institutions out of the interactions of given individuals. This section explores the problems that may arise if exclusive stress is put on the latter direction of causality.

The distinctive 'new' institutionalist project has been identified as the attempt to explain the existence of institutions by reference to a given model of individual behavior, and on the basis of an initial institution-free 'state of nature.' The procedure is to start with given individuals and to move on to explain institutions.

Admittedly, substantial heuristic and other insights have been gained on the basis of this approach. The main problem is the incompleteness of the research program in its attempt to provide a general theory of the emergence and evolution of institutions. Some moves towards an 'evolutionary' approach to solving this problem are addressed at the end of this section.

Alexander Field (1979) has exposed a key problem in the new institutionalist project. In trying to explain the origin of social institutions from given individuals alone, the 'new' institutional economics has to presume a social framework governing their interaction. In any original, hypothetical 'state of nature' from which institutions are seen to have emerged, a number of rules and cultural and social norms are already presumed. No 'thought experiment' involving an institution-free 'state of nature' has yet been postulated without them. For example, in game theoretic presentations of the emergence of institutions or norms (Schotter 1981), take the individual 'for granted' as an agent unambiguously maximizing his or her expected payoff. Yet there can be no games without prior rules, and game theory can never explain the elemental rules themselves. Even in a sequence of repeated games, or of games about other (nested) games, at least one game or metagame, with a structure and payoffs, must be assumed at the outset. Any such attempt to deal with history in terms of sequential or nested games is thus involved in a problem of infinite regress. As another example, Williamson's transaction cost theory of the firm takes its original state of nature as the market. He writes that 'in the beginning there were markets' (Williamson 1975, p. 20). This starting point is characteristic of his approach. From this original context, some individuals go on to create firms and hierarchies. He argues that these endure if they involve lower transaction costs. However, the market itself is a social institution. The market involves social norms, customs, instituted exchange relations, and—sometimes consciously organised—information networks that themselves have to be explained (Hodgson 1988). Market and exchange relations themselves involve complex rules. In particular, the institution of private property itself requires explanation. Markets are not an institution-free beginning. Hence Williamson fails to explain the firm from an institution-free 'state of nature.' In a type of comparative static approach, he assumes one institutional framework and derives another. Accordingly, the 'new' institutionalist project of starting simply from given individuals is abandoned.

There are good reasons why such a starting point is generally misconceived. Before an individual may choose, he or she requires a conceptual framework to

make sense of the world. The reception of information requires a paradigm or cognitive frame to process and make sense of it. Further, our interaction with others requires the use of the institution of language. We cannot understand the world without concepts and we cannot communicate without some form of language. As the 'old' institutionalists argue, the transmission of information from institution to individual is impossible without a coextensive process of *enculturation*, in which the individual learns the meaning and value of the sense-data that is communicated. Any communication of information *always and necessarily* involves such a process of enculturation. Information cannot be received unless the individual has been enculturated to some degree, through prior engagement with institutions. Communication requires an institutionalized individual.

In the 'old' institutional economics, cognition and habit have a central place. Knowledge and learning are stressed. There is also an insistence that the perception of information is not possible without prior habits of thought to endow it with meaning. Without such habits, agents cannot perceive or make use of the data received by their senses. Habits thus have a crucial cognitive role: 'All facts of observation are necessarily seen in the light of the observer's habits of thought' (Veblen 1914, p. 53). Such habits are acquired through involvement in institutions.

The central 'new' institutionalist project of explaining institutions from individuals alone is thus misconceived. The problem of infinite regress identified here undermines any 'new institutionalist' claim that the explanation of the emergence of institutions can start from some kind of original, institution-free ensemble of (rational) individuals, in which there is supposedly no rule or institution to be explained. At the very minimum, 'new' institutionalist stories of the development of institutions depend on interpersonal communication of information. The communication of information itself requires shared concepts, conventions, rules, routines, and norms. These, in turn, have to be explained.

What is being contested here is the possibility of using given individuals as the institution-free starting point in the explanation. It is not possible to understand how institutions are constructed without seeing individuals as embedded in a culture made up of many interacting institutions. Institutions not only constrain but also influence individuals. Accordingly, if there are institutional influences on individuals and their goals, then these are worthy of explanation. In turn, the explanation of those may be in terms of other purposeful individuals. But where should the analysis stop The purposes of an individual could be partly explained by relevant institutions, culture, and so on. These, in their turn, would be partly explained in terms of other individuals. But these individual purposes and actions could then be partly explained by cultural and institutional factors, and so on, indefinitely.

We are involved in an apparently infinite regress, similar to the puzzle 'which came first, the chicken or the egg?' Such an analysis never reaches an end point. It is simply arbitrary to stop at one particular stage in the explanation and say 'it is all reducible to individuals' just as much as to say it is 'all social and institutional.' The key point is that in this infinite regress, neither individual nor institutional factors have complete explanatory primacy. The idea that all explanations have ultimately to be in terms of individuals (or institutions) alone is thus unfounded.

We have seen that the new institutionalist project to explain the emergence of institutions on the basis of given individuals runs into difficulties, particularly with regard to the conceptualization of the initial state from which institutions are supposed to emerge. This does not mean that all 'new' institutionalist research is without value, but it indicates that the starting point of explanations cannot be institution-free: the main project has to be reformulated as just a part of a wider theoretical analysis of institutions.

In a major work, Masahiko Aoki (in press) recognizes the seriousness of the problem of infinite regress within the 'new' institutional economics. He attempts to deal with this problem both by taking individuals as given and assuming a historically-given bedrock of institutions. The focus in this approach is then on the further development of those institutions. This marks an important watershed in the development of the 'new' institutionalism. At least in strict terms, methodological individualism is abandoned. No longer is an attempt made to explain the emergence and evolution of institutions in terms of individuals alone. He adds a given institutional context to the story.

Aoki's work can be seen as a partial step in the direction of a fully-fledged evolutionary account of the growth and development of institutions. This would stress the evolution of institutions, in part from other institutions, rather than from a hypothetical, institution-free 'state of nature.' What is required is a theory of process, evolution, and learning, rather than a theory that proceeds from an original, institution-free, 'state of nature' that is both artificial and untenable.

The 'comparative statics' character of many 'new' institutionalist explanations is obvious. However, one of the reasons for the rise of 'evolutionary' thinking in economics since the early 1980s has been an attempt to break the constraints of this mode of explanation with its two fixed end points. Because there is no answer to the chicken-or-egg question, the question itself has to be changed. The question should no longer be 'which came first?' but 'what processes explain the development of both of them?' Again this implies a movement away from comparative statics and towards a more evolutionary and open-ended framework of analysis. Some moves in this direction by two prominent 'new' institutionalists have already led to a degree of convergence with the evolutionary and open-ended

ideas of the 'old' institutionalists. This is apparent, for example, in the later writings of North (1990).

Such evolutionary explanations involve the search for 'a theory of the process of consecutive change, realized to be self-continuing or self-propagating and to have no final term' (Veblen 1919, p. 37). Emphatically, abandoning the attempt to explain all institutions in terms of given individuals does not mean the abandonment of theoretical explanation. Instead, the origins and development of organizations and institutions are seen as an evolutionary process. Today, there is a substantial amount of work going on in this area, with extensive use being made of evolutionary metaphors taken from biology.

5. *A Research Agenda for Institutional Economics*

Today there are lively and growing research activities within both the 'new' and the 'old' institutional economics (Hodgson et al. 1994). Substantial benefits can be anticipated from an intensive dialog between these two schools. The 'new' institutionalists can learn from the evolutionary, historical, culturally-grounded, and technologically-oriented studies of the 'old' institutionalists. The 'old' institutionalists can usefully devote more attention to the development of a systematic institutional approach to economic theory. In this they can learn from the achievements and mistakes of the 'new' institutional economics.

The philosophical foundations of the social sciences have been shaken in recent years. Debates in philosophy and social theory have questioned many of the standard assumptions and approaches. The future development of institutional economics must address some of these underling philosophical problems, posed by philosophical realists on the one hand and postmodernists on the other. One of the tasks of institutional economics is to address some of the underlying philosophical issues—particularly those relating to the social ontology and nature of institutions (Searle 1995)—and bring useful insights into their work. Evolutionary thinking is once again on the ascendancy in economics (Hodgson 1993). However, the detailed development and explication of a theory of institutional evolution—a task envisaged by Veblen more than 100 years ago—still remains incomplete. One of the impediments to progress in this area is the lack of a consensus on key terms and methodological approaches. The solution to this problem lies through open and extended conversation and debate between the different points of view involved. Only then will it be possible to establish a sufficient consensus on key definitions and terms.

The promise for the future development of institutional economics lies in a fully interdisciplinary approach, using insights from other disciplines such as philosophy, sociology, anthropology, geography, history, psychology, and biology. Such insights have to be woven together in a well-founded theoretical structure, using insights from both the 'old' and the new institutional economics.

See also: Economic Anthropology; Economic Sociology; Economics, History of.

Bibliography

Aoki M in press *Towards a Comparative Institutional Analysis*. MIT Press, Cambridge, MA

Commons J R 1924 *Legal Foundations of Capitalism*. Macmillan, New York (Reprinted 1968 (University of Wisconsin Press, Madison WI); 1974 (Augustus Kelley, New York); and 1995 with a new introduction by Biddle J E, Samuels W J (Transaction, New Brunswick, NJ))

Commons J R 1934 *Institutional Economics—Its Place in Political Economy*. Macmillan, New York (Reprinted 1990 with a new introduction by Rutherford M (Transaction, New Brunswick, NJ))

Field A J 1979 On the explanation of rules using rational choice models. *Journal of Economic Issues* **13**(1): 49–72

Furubotn E G, Richter R 1997 *Institutions in Economic Theory: The Contribution of the New Institutional Economics*. University of Michigan Press, Ann Arbor, MI

Galbraith J K 1969 *The New Industrial State*. Penguin, Harmondsworth, UK

Hodgson G M 1988 *Economics and Institutions: A Manifesto for a Modern Institutional Economics*. Polity Press, Cambridge, UK and University of Pennsylvania Press, Philadelphia

Hodgson G M 1993 *Economics and Evolution: Bringing Life Back Into Economics*. Polity Press, Cambridge, UK and University of Michigan Press, Ann Arbor, MI

Hodgson G M 1999 Evolution and Institutions: On Evolutionary Economics and the Evolution of Economics. Edward Elgar, Cheltenham, UK

Hodgson G M, Samuels W J, Tool M R (eds.) 1994 *The Elgar Companion to Institutional and Evolutionary Economics*. Edward Elgar, Aldershot, UK

Milgrom P R, Roberts J 1992 *Economics, Organization, and Management*. Prentice-Hall, Englewood Cliffs, NJ

Mitchell W C 1937 *The Backward Art of Spending Money and Other Essays*, 1st edn. McGraw-Hill, New York

North D C 1981 *Structure and Change in Economic History*, 1st edn. Norton, New York

North D C 1990 *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, UK

Rutherford M H 1994 *Institutions in Economics: The Old and the New Institutionalism*. Cambridge University Press, Cambridge, UK

Schotter A R 1981 *The Economic Theory of Social Institutions*. Cambridge University Press, Cambridge, UK

Schotter A R 1994 *Microeconomics: A Modern Approach*. Harper Collins, New York

Searle J R 1995 *The Construction of Social Reality*. Free Press, New York

Tool M R 1991 Contributions to an institutionalist theory of price determination. In: Hodgson G M, Screpanti E (eds.) *Rethinking Economics: Markets, Technology and Economic Evolution*. Edward Elgar, Aldershot, UK, pp. 19–39 (Re-

printed in Tool M R 1995 *Pricing, Valuation and Systems: Essays in Neoinstitutional Economics*. Edward Elgar, Aldershot, UK)
Veblen T B 1899 *The Theory of the Leisure Class: An Economic Study in the Evolution of Institutions*. Macmillan, New York (Republished 1961 by Random House, New York)
Veblen T B 1904 *The Theory of Business Enterprise*. Charles Scribner's, New York (Reprinted 1975 by Augustus Kelley)
Veblen T B 1914 *The Instinct of Workmanship on the State of the Industrial Arts*. Macmillian, New York (Reprinted 1990 by Transaction, New Brunswick, NJ).
Veblen T B 1919 *The Place of Science in Modern Civilisation and Other Essays*. B. W. Huebsch, New York (Reprinted 1990 with a new introduction by Samuels W J by Transaction, New Brunswick, NJ)
Williamson O E 1975 *Markets and Hierarchies: Analysis and Anti-Trust Implications: A Study in the Economics of Internal Organization*. Free Press, New York
Williamson O E 1985 *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. Collier Macmillan, London
Yonay Y P 1998 *The Struggle Over the Soul of Economics: Institutionalist and Neoclassical Economists in America Between the Wars*. Princeton University Press, Princeton, NJ

G. M. Hodgson

Institutional Investors

For organizational and management scholars interested in the political dynamics of corporations, the rise of institutional investors in the second half of the twentieth century has provided a key focal point for the analysis of organizational change and shifts in corporate control (e.g., Useem 1996). Institutional investors are organizations such as mutual funds, hedge funds, commercial and investment banks, insurance companies, pension funds, securities companies, trusts, foundations, labor unions, colleges and universities, and corporations that take major positions in publicly traded securities, and often account for a good deal of trading volume on major securities exchanges. As an indication of the growing power of institutional investors, the percentage of US stocks that such actors own has dramatically increased from around 16 percent in 1965 to over 60 percent by the mid-1990s.

In order to shed light on how the emergence of these new, powerful actors has reshaped organizational life, researchers have employed a wide variety of research foci and approaches. Research has highlighted how institutional investors can be viewed as a social movement (Davis and Thompson 1994), corporate action is consequentially shaped by lending and other intercorporate relationships (Mintz and Schwartz 1985, Mizruchi and Stearns 1994), institutional investors may shape organizational phenomena such as research and development expenditures, corporate social performance, and CEO compensation, the hostile takeover of corporations requires a good deal of cultural work by institutional investors and other powerful stockholders (Hirsch 1986), the adoption of poison pills can protect managers from corporate raiders (Davis 1991), and how the creation of investor-relations departments has enabled corporations to better manage the demands of institutional investors (Rao and Sivakumar 1999). While this work has been very revealing, it has tended to focus primarily on issues of corporate control in the US, while neglecting questions regarding how institutional investors have contributed to the reshaping of country-specific financial activities as well as global capital formation processes.

To situate this article on institutional investors, an account is provided first of the rise of pension plans that motivated the growth of the institutional investor movement in the US. Since the majority of organizational research on institutional investors focuses on problems of corporate control, there follows a discussion of how institutional investors have been conceptualized in competing theories about corporate control and argument for more research on how institutional investors are helping to remake capital formation processes around the world.

1. The Emergence of Pensions and the Institutional Investor Movement in the US

The growth of pension plans has provided one of the most significant changes in the structure of US retirement and finance in the twentieth century—providing a foundation for the rapid growth of institutional investing in the second half of the century. The first public employer-sponsored pension plan was established by the City of New York in 1850; the first private employer-sponsored plan was created by the American Express Company in 1875. Up until World War II, however, pension coverage was sparse and the plans that did exist were not guaranteed and often ownership was not vested in the pensioned person.

A confluence of factors before and after World War II fostered the development of pensions in the US. Before World War II, the Social Security Act of 1935 boosted employment-related pension insurance by institutionalizing the idea of retirement (Dobbin 1992). Subsequently, the combination of federally mandated wartime wage controls, as well as a war-related labor shortage, facilitated the development of benefits such as deferred wages as a way for corporations to compete for scarce labor. Also, due to the high corporate tax rate during war time (peaking at over 90 percent), pension plans became more attractive to employers since contributions to such plans lowered taxable income. In addition, labor union-corporation conflict became more intense after

the war. In 1948, the National Labor Relations Board ruled that pensions were to be considered as part of wages, allowing workers to initiate legally a strike over a pension dispute. Pensions became a centerpiece for management–labor negotiations.

Despite worker–manager conflict in major industries such as autos, coal mining, and steel, the emergence of pension benefits came to be associated with a system of management that valued employee commitment to and long career tenures within single organizations. In 1950, fewer than 10 million nonagricultural workers (around 25 percent of the nonagricultural workforce) were covered by private pension plans. By 1970 more than 30 million workers (approximately 50 percent of the nonagricultural workforce) were covered. As pension plans diffused and amassed a considerable amount of assets, however, the management of those assets began to change.

Up until around 1950, most pension plans were constituted as annuities and managed by insurance companies. But in the 1950s, corporate pension plans began to be managed by bank trust departments that were allowed to invest in stocks as well as bonds. At that time, insurance companies were legally limited in the extent to which they could invest in stocks. That migration of pension assets into the stock market represented a major turning point in the development of pension management as an industry. As increasing amounts of money were being directed towards the stock market, there was a corollary growth in financial intermediaries that vigorously competed for the right to manage how other people's money was to be invested in the stock market—an institutional investor movement was born.

Competition among institutional investors was largely based on performance measured by asset growth, leading to an emphasis on riskier portfolio strategies. Between 1965 and 1969, the turnover (a proxy for risk-taking) of bank managed pension portfolios rose from 13.7 percent to 25.7 percent, of internally managed portfolios from 5.5 percent to 9.6 percent, and of investment advisor managed portfolios from 27.4 percent to 55.9 percent. Corporations began to view pension assets as a profit center from which they could directly enhance their bottom line, while previously in the 1950s, pensions were viewed primarily as a way to reduce tax payments.

The shift towards viewing pension assets as a profit center led to a general restructuring of the money management industry. Bank trust departments, that had dominated pension management until the late 1960s, relied on relatively conservative investment strategies, constructing portfolios that consisted mainly of stocks of established companies that paid steady dividends. In the late 1960s, corporate pension plans began to expand beyond bank trusts, using multiple investment advisors. This effectively broke the bank trust monopoly on pension management. As a result, smaller, independent money managers, including those affiliated with mutual fund companies, became able to compete for pension dollars based on their performance record as opposed to institutional reputation and longstanding relationships.

Alongside these developments, the federal government became more involved in the protection of beneficiaries. In 1974, the Employee Retirement Income Security Act (ERISA) was enacted and the Pension Benefit Guarantee Corporation (PBGC) was created to insure qualified plans. ERISA established minimum standards for vesting, proscribed funding formulas, provided insurance for employees whose plans were terminated without adequate assets, required full disclosure of plan benefits and operations to employees, and declared that pension trustees had a fiduciary responsibility to manage pension assets with only beneficiaries in mind, not a corporation's bottom line. This federal government involvement stabilized pension practices, enabling the continued growth of pension plans while also providing more safeguards for pension beneficiaries.

Between 1945 and 1996, pension assets have grown steadily from virtually zero percent to 20 percent of all financial intermediary assets. While the growth of pensions was a critical development that facilitated the institutional investor movement, it was also supported and facilitated by broader changes in US capital formation processes. In particular, the growing legitimation of the stock market after World War II provided important momentum for a major transformation from bank lending dominated to market-based capital formation processes. Commercial banks, which had been the dominant financial actors in the US, increasingly lost their hegemonic position in the postwar era. The percentage of financial intermediary assets managed by commercial banks declined from 60 percent in 1945 to 20 percent by 1996. At the same time, new kinds of institutional investors, such as mutual funds, grew dramatically. Mutual funds began their dramatic rise in the 1980s and now constitute 14 percent of all financial intermediary assets. Hence, while the origins of the institutional investor movement are rooted in the development of pensions, the legitimation of the stock market and the concomitant shift towards market-based capital formation processes, as well as the emergence of new kinds of financial organizational forms importantly shaped the contours of that movement.

2. *The Study of Institutional Investors in Organizational Theory: The Problem of Corporate Control and other Research Possibilities*

Much of the organizational research on institutional investors has focused on the problem of corporate control, grounded in historical analyzes of the emergence and evolution of the corporate form. US Corporations emerged in the mid-nineteenth century

under government sponsorship to perform specific services such as banking and the building of railroads, turnpikes, and canals. Individually chartered public infrastructure corporations were predominant until more generalized state incorporation laws arose in the 1870s. By the turn of the twentieth century, large industrial manufacturing companies and Wall Street financiers joined forces to facilitate the broad diffusion of the corporate form (Roy 1997). This ushered in the era of 'finance capitalism' and the rise of powerful bankers such as J. P. Morgan.

Along with the birth of finance capitalism was the accelerated growth of financial intermediaries and the construction of the US field of finance. The early part of the twentieth century not only witnessed the accumulation of power by money center banks and investment bankers, but also the birth of new financial organizations such as investment and finance companies. Assets controlled by financial intermediaries grew from $19 billion in 1900 to $167 billion in 1929 and to approximately $560 billion by mid-century. It was in the first half of the 1900s that a truly national field of finance began to take shape. In the early part of the twentieth century, the field of finance was highly centralized with the most powerful actors and majority of asset flows located in New York City. This financial concentration came to be pejoratively known as the 'money trust.'

Alongside the growth and concentration of private financial intermediaries, the federal government created a central banking structure and became increasingly involved in legislating and sponsoring the formation of regulatory bodies to monitor and restrict intermediary activities. Governmental regulatory activity increased dramatically in the 1930s amid the Great Depression when explicit efforts were made to dismantle the money trust. Finance capitalism had ended.

The emergence and growth of large industrial corporations was shaped importantly by this rapidly developing capital formation infrastructure (field of finance) that supported and made possible increased corporate activity. As large-scale industrial activity became organized by the corporate form in the early part of the twentieth century, ownership became more diffuse and strategic decision-making power became vested in professional managers who had little financial stake in the firm. The publication of *The Modern Corporation and Private Property* (Berle and Means 1932) proclaimed the rise of managerial capitalism where American corporations came to be theorized as relatively autonomous entities that were unconstrained by outside owners or each other.

The shift from finance to managerial capitalism, however, was not only facilitated by the separation of ownership from control, but also the increased role of the state in capital formation processes, especially after the stock market crash of 1929 and Great Depression. Through the emergence of social redistribution programs and major war mobilization efforts, state-financing activity provided a major impetus for the creation of a stable and reliable field of financiers. In addition, a number of landmark legislative acts and finance regulatory bodies were created. For example, securities and banking legislation between 1933–35 separated commercial and investment banking and created the Securities and Exchange Commission. Also, state regulation and monitoring greatly reduced the scope and power of banks and individual intermediaries.

Up until the mid-1960s, organizational analysis had been dominated by managerial-centered theories which focused on what top executives did given certain constraints having to do with limited time and information as well as intraorganizational political dynamics. Building on Zeitlin's (1974) critique of the Berle and Means (1932) thesis that nonfinancial corporations had been freed from dependence on financial institutions, however, a number of organizational sociologists began to focus directly on corporate governance issues in the 1970s and 1980s, investigating the extent to which managerial action was indeed constrained by financial actors. In their review of the literature on corporate control, Glasberg and Schwartz (1983) argued that the notion of managerial autonomy has been challenged by four main theoretical camps: resource dependence, class cohesion, bank control, and bank hegemony.

Resource dependence theorists maintain the managerialist premise that top managers dominate corporate decision-making but challenge the notion that such autonomy is free from outside constraint. Proponents of this view argue that organizations are involved in an ongoing interdependent struggle for positional advantage, using interlocking directorates, joint ventures, mergers, diversification, and conglomeration to co-opt, internalize, or otherwise neutralize powerful external constituencies. This line of thought has been particularly prevalent in studies of interlocking directorates among financial and nonfinancial firms. The broader social theoretic imagery which emerges is conflictual and pluralistic. Scholars drawing on this theoretical orientation, therefore, tend to view institutional investors as hostile outsiders that managers need to neutralize in order to maintain relative autonomy.

Arguments drawing on class-cohesion theory, that are inspired by Marxist political economic thought, similarly accept the premise that managers are relatively free to make decisions, but claim that corporate executives comprise a unified class of actors that promote coordination and cooperation at the expense of high-powered market incentives (Davis 1991, Domhoff 1974). The orderliness of societal stratification is emphasized to counter the resource dependence conclusion that intraclass struggle best characterizes the elite strata. Recent writings in this tradition, however, have argued that elite institutional investors may be emerging as a class that controls financial flows

and corporate decision-making (Useem 1996). This revised argument seems to move class cohesion arguments closer to the pluralistic imagery of resource dependence arguments though, since institutional investors and managers are viewed as engaged in ongoing conflict over the control of corporations.

Bank-control theory, derived from twentieth-century Marxist financial capital theory, posits that banks and insurance companies, due to their centrality in controlling and allocating capital, are able to use their leverage to dictate corporate policies to serve their own interests. This approach completely rejects all theories that maintain any degree of managerial autonomy. As such, this perspective has been marginalized in organization theory. In addition, recent research has highlighted that the extent to which bank representatives have had formal positions on corporate boards of directors has decreased, indicating a waning bank influence on corporate management (Davis and Mizruchi 1999). This insight dovetails with the more general decline of bank-lending in favor of market-based capital formation processes in the US field of finance.

Bank hegemony theory may be characterized as an effort to try to synthesize bank control and class cohesion theories (Mintz and Schwartz 1985). As opposed to the overt domination proposed by bank-control theorists, power is theorized to be embedded in the relations between financial institutions and corporate producers. Hence, research affiliated with this tradition has looked primarily at direct relations between corporations and financial institutions via interlocks or lending relationships (Mizruchi and Stearns 1994) to analyze the extent to which managers are constrained by financial actors. While an underlying consonance of interest between financial and nonfinancial firms is posited, the financial sector is theorized to develop a kind of class cohesion that ultimately allows for financial institution dominance. The expression of power, however, is much more problematic and diffuse than a bank-control theorist would argue, though more coherent than predicted by class-cohesion theory (Glasberg and Schwartz 1983).

While resource dependence, class cohesion, bank control, and bank hegemony were the main perspectives used to address institutional investors and corporate control in the 1970s and early 1980s, research on these issues has since become more analytical and less overtly engaged in these broader theoretical problematics. For example, debates about the relative veracity of resource dependence approaches that envision a more pluralistic society vs. more Marxist class-cohesion arguments have been backgrounded. Nonetheless, corporate control research has continued to usefully contribute to our further understanding of the relationship of institutional investors to corporate governance and change.

One of the main limitations of this line of research, however, has been its almost sole focus on corporate control as a way to understand financial actors such as institutional investors. This has left the practice of institutional investing as well as related areas of money, banking and finance as components of a broader financial field subject to its own logic of operation woefully understudied. Since the deregulation of banking in the US in 1980, the boundaries between previously segregated financial organizational forms have begun to blur. Now insurance, commercial and investment banking, and mutual fund activities are combined under one roof. These developments call out for more detailed investigations of how financial practices and organizations are being remade as a result of these field-wide transformations.

In addition, we should also be studying how institutional investors are driving or are implicated in broader processes of social change. A first step in this direction would be to study how the activities of financial institutions and institutional investors are reshaping capital formation and corporate governance processes around the globe (Davis and Mizruchi 1999). The deregulatory 'big bang' that occurred in the US in the early 1980s subsequently has diffused to Europe and Japan and other countries around the globe. As a result of these contemporary developments, new kinds of questions have emerged having to do with whether capital formation and corporate governance processes around the globe are being remade in a way that mimics Western capitalistic countries or whether localized state–society relations exhibit more of a path-dependent quality.

As Zysman (1983) showed, the tendency towards market-based capital formation processes in England and in the US are markedly different from the more bank- and state-dominated systems in Japan, France, and Germany. Is this still the case? Recent research on Germany, France, and England indicate that those national capital formation patterns continue to persist (Mayer and Whittington 1999). We must go beyond more simplistic notions of globalization and convergence and investigate how social actors resist such convergences as well as how economy–society relations are informed by cultural logics that make the US model of finance inappropriate or unthinkable (Biggart and Guillén 1999). This may require more detailed ethnographic studies or historical research that can reveal how societal institutions are constructed locally and rooted in geographically situated meaning systems (Ventresca and Porac 2000). While there have been recent efforts to study corporate governance variation cross-nationally, we still know very little about how capital formation processes are changing in the financial systems of particular countries. There is no doubt that institutional investors are key actors in this unfolding drama.

Another important question has to do with whether the old 'money trust' is re-emerging as the boundaries between different financial organizational forms become irrelevant. The consolidation of financial

services into conglomerates has the potential to reinvigorate longstanding debates about elite class control versus pluralism, although with a global dimension. Since financial service conglomerates that are being built through mergers, acquisitions, and alliances in this era of deregulation are transnational in scope, the new, emergent money trust may not be a uniquely US phenomenon. Likewise, institutional investors are active participants in securities markets around the world and play a key role in economic development and global monetary flows. Hence, we may want to begin to study whether a new transnational class of capitalists is in the making, or at least probe how large, global investors are shaping the social organization of countries around the world.

3. Conclusion

The rise of institutional investors both in the US and around the globe is a critically important phenomenon that researchers have only just begun to investigate. While we know a good deal about how the rise of institutional investors has challenged managerial autonomy in the US, we know much less about the relationship between institutional investors and corporate governance in other, especially non-Western, countries. More generally, we know very little about how country-specific capital formation processes are changing amid the continuing globalization of financial markets and the emergence of transnational financial conglomerates. We would benefit greatly from detailed cross-national studies of how economy–society relationships may be changing as a result of these developments. Through such research, we may begin to unearth the heterogeneity of practices lumped under the category 'institutional investors' and reveal the wide variety of influences these actors have around the world.

See also: Capitalism; Capitalism: Global; History of Technology; Information Society; Institutional Investors; International Marketing; Stock Market Predictability; Stockholders' Ownership and Control; Technological Innovation; Venture Capital

Bibliography

Berle A A, Means G C 1932 *Modern Corporation and Private Property*. Macmillan, New York

Biggart N C, Guillén M F 1999 Developing difference: Social organization and the rise of the auto industries in South Korea, Taiwan, Spain and Argentina. *American Sociological Review* **64**: 722–47

Davis G F 1991 Agents without principles? The spread of the poison pill through the intercorporate network. *Administrative Science Quarterly* **36**: 583–613

Davis G F, Mizruchi M S 1999 The money center cannot hold: Commercial banks in the US system of governance. *Administrative Science Quarterly* **44**: 215–39

Davis G F, Thompson T A 1994 A social movement perspective on corporate control. *Administrative Science Quarterly* **39**: 141–73

Dobbin F R 1992 The origins of private social insurance: Public policy and fringe benefits in America, 1920–1950. *American Journal of Sociology* **97**: 1416–50

Domhoff G W 1974 *The Bohemian Grove and Other Retreats*. Harper and Row, New York

Glasberg D S, Schwartz M 1983 Ownership and control of corporations. *Annual Review of Sociology* **9**: 311–32

Hirsch P M 1986 From ambushes to golden parachutes: Corporate takeovers as an instance of cultural framing and institutional integration. *American Journal of Sociology* **91**: 800–83

Mayer M C J, Whittington R 1999 Strategy, structure and 'systemness': National institutions and corporate change in France, Germany and the UK, 1950–1993. *Organization Studies* **20**: 933–59

Mintz B, Schwartz M 1985 *The Power Structure of American Business*. University of Chicago Press, Chicago

Mizruchi M S, Stearns L B 1994 A longitudinal study of borrowing by large American corporations. *Administrative Science Quarterly* **39**: 118–40

Rao H, Sivakumar K 1999 Institutional sources of boundary spanning structures: The establishment of investor relations departments in the Fortune 500 industrials. *Organization Science* **10**: 27–42

Roy W G 1997 *Socializing Capital: The Rise of the Large Industrial Corporation in America*. Princeton University Press, Princeton, NJ

Useem M 1996 *Investor Capitalism*. Basic Books, New York

Ventresca M, Porac J 2000 *The Social Construction of Markets and Industries*. Pergamon, Oxford, UK

Zeitlin M 1974 Corporate ownership and control: The large corporation and the capitalist class. *American Journal of Sociology* **79**: 1073–19

Zysman J 1983 *Governments, Markets, and Growth: Financial Systems and the Politics of Industrial Change*. Cornell University Press, Ithaca, NY

M. Lounsbury

Institutionalism

The study of political institutions has moved back into center stage in political science. Whereas only a few years ago, institutions were mostly casually mentioned in most political science research, today they are a central focus of attention. This change is not simply a change of language or the catching on of a new popular phrase or academic fad. Instead these changes represent an important development in evolution of political science theory and intellectual focus.

1. What are Institutions?

In the broadest sense, institutions are simply rules. As such, they are a foundation for all political behavior. Some are formal (as in constitutional rules) some are informal (as in cultural norms), but without institutions there could be no organized politics. Indeed absent institutions there could be no organization at

all. To understand this point, simply attempt to consider a world in which there were no rules: In this Hobbesian hell, individuals would be forced to 'invent' communication every time they encountered another individual. In this sense, then, if we study social interaction, we study institutions. This does not suggest, however, that all social scientists are 'Institutionalists.'

2. *Institutionalism*

The 'Institutionalism' specifically examines the ways in which institutions structure social and political behavior (North 1990). This burgeoning body of literature argues that policy, politics, and behavior can only be understood in the context of the institutions in which they take place. Thus, for example, Ellen Immergut argues that variations in National Health Insurance (NHI) systems are best explained by variations in national political institutions (Immergut 1992). Similarly, Bo Rothstein shows that Sweden's high union density is best explained by the 'Ghent' unemployment insurance system which gives workers powerful incentives to join Swedish unions (Rothstein 1992). Even more broadly, Douglas North suggests that the very success of Western political economic model is rooted in the peculiar institutions developed in these societies (North 1990).

The central tenet of this New Institutionalism is that institutions are not neutral to policy outcomes. As Peter Hall has suggested, 'On the one hand, the organization of policy-making affects the degree of power that any one set of actors has over policy outcomes.... On the other hand, organizational position also influences an actor's definition of his own interests...in this way, organizational factors affect both the degree of pressure an actor can bring to bear on policy and the likely direction of that pressure' (Hall 1986, p. 12).

In sum, institutions define the rules of the political game and as such they define who can play and how they play. Consequently, they ultimately can shape who wins and who loses. If politics is the study of who gets what, when and why, then institutionalists argue that institutions should be at the heart of that study.

3. *A Brief History of Institutionalism*

Institutionalism has a long established tradition among those interested in politics and political outcomes. Plato's *Republic* is a comparative study of institutions. Similarly, Aristotle's central concern in *Politics* is which kinds of political institutions will produce the best outcomes. James Madison must clearly also be seen as an early American 'Institutionalist' in that he was specifically concerned with which kinds of institutions would produce the best political outcomes and how the specific design of institutions would shape political outcomes.

Political scientists have also long been interested in institutions. Indeed, in its early years political science meant the study of political institutions (Wilson 1891). But, with some important exceptions (cf. Herring 1940, Key 1947), early political science was often more descriptive than analytical. 'Comparative politics,' in particular, consisted mostly of detailed configurative studies of different legal, administrative, and political structures (for a review and critique of this early comparative literature see Verba 1967). In the immediate postwar years a new generation of political scientists attempted to make the study of politics more 'scientific.' For many this effectively meant that political science ought to model itself on the 'hard sciences' which they believed was fundamentally a deductive process. Thus rather than studying the details of political life and inductively uncover the patterns of behavior and action, political 'science' should be a deductive science that seeks to discover of the general laws and fundamental forces that lay behind political action. Focusing on particular institutions, proponents this intellectual agenda implied, was 'a-theoretical.' Concomitant with the push for more abstract laws, political scientists were disillusioned by the failure of parliamentary institutions in Inter-war Weimar Germany (and later in post-colonial Africa) to prevent these polities from devolving into authoritarianism. Clearly, many argued, there were bigger, more important, indeed more fundamental, forces at work in politics and development than political institutions. These forces, they argued, should be the focus of political science.

Thus, behavioralist, functionalist and Marxist perspectives took leading roles in political science theory building through most of the 1960s and 1970s. As a consequence, institutional analysis diminished in prominence...particularly in comparative politics. For Functionalists, political institutions were simply seen as organizations created to fulfill the systemic needs of society. As such, political institutions were neither important nor interesting. For Behavioralists and Marxists, political institutions were simply arenas in which political battles were fought out. While Marxists saw the relevant groups as classes and Pluralists saw them as more narrowly defined, neither theoretical perspective played particular attention to the structure or character of the institutions themselves. The real meat of politics, they argued—and the keys to understanding political outcomes—was found in the articulation of group interests. Though this was rarely explicitly stated, implicit in these theories was the assertion that if politics or policies differed between societies, this difference was surely the result of different constellations of group and class interests or preferences (Almond and Verba 1963, Miliband 1969).

It was not the case, however, that all political scientists had abandoned the study of institutions. Indeed, many of the most widely read scholars in American Politics maintained an explicitly institutional emphasis (cf.

Schattschneider 1960, Greenstein 1963, Fenno 1960, Polsby et al. 1969). In Comparative Politics, as well, several political scientists continued to examine political institutions and their effects on political outcomes (Ekstein 1960, Bendix and Brand 1968, Huntington 1968), even while they sometimes had to defend their 'inductive' approach from the challenge that it was not 'scientific' (Przeworski and Teune 1970, Bill and Hardgrave 1973). For these scholars it was self-evident that if one wanted to understand what government does, one needs to specifically study the institutions through which it acts.

It was probably in Comparative Politics that the search for 'Grand Theory' had its most significant impact. It may be for this reason that a self-conscious return to 'Institutionalism' was most forcefully articulated here. Among the first group of scholars to move in this direction was Peter Katzenstein and his colleagues who sought to explain why several advanced capitalist states responded so differently to the oil shock of the 1970s. The answer, they concluded, was found in the differing institutional structures in these polities and the consequent patterns of economic policy pursued in each nation (Katzenstein 1978). Similarly, Theda Skocpol's study of social revolutions also concluded that one could not explain the course of a country's revolution without examining the nature and structure of the state against at which these revolutions were aimed (Skocpol 1979).

Soon a new 'state centered' approach emerged in comparative political inquiry. These scholars forcefully argued against the behavioralist and Marxist 'grand theory' emphasis then dominant in the study of comparative politics. Instead, they suggested, political outcomes were shaped and structured by the specific actors and their position in the state. One should not treat the state as a neutral 'black box' through which group or class interest was simply translated (Evans et al. 1985). Instead, the state had independent interests and agendas which were separate and different from the interests and preferences of classes and interest groups that made up society.

Of course it did not take long for these scholars to discover that 'the state' is too broad a concept and too varied a set of institutions to be 'taken seriously' without being broken down. To make analytic sense out of the insights, scholars began examining institutions more carefully. If state institutions 'matter,' they argued, then why should this not be equally true of institutions outside the boundaries of the executive state? With these questions 'The New Institutionalism' was born.

4. *Three 'New Institutionalisms'?*

Today three different intellectual approaches lay claim to the term 'Institutionalist' (cf. Hall and Taylor 1996). Each grows out of a different academic discipline and attempts to integrate these different analytic traditions into the understanding of politics. Sociological Institutionalism, as its name implies, grows out of sociology and the study of organizations (e.g., Selznick 1949). These scholars have been centrally interested in understanding culture and norms, as institutions. These scholars emphasize 'folkways,' 'patterns of behavior,' and 'cognitive maps' and argue that these social institutions are critical for understanding the structure of social, political, and economic interactions (March and Olsen 1989, DiMaggio and Powell 1991). Building on upon their analyses of complex organizations, these scholars show the relationship between formal institutions and the structure or patterns of behavior and beliefs. They argue that these informal institutions are core to any understanding of the nonrational aspects human communication and exchange.

'Rational Choice Institutionalism,' taking its lead from economics, emphasizes quite different sets of institutions and their effects. Rational Choice (RC) scholars attempt to apply the formal logic and methods to the study of politics and history and often explicitly eschew the more 'mushy' or less precise variables such as norms and beliefs. Instead, basic assumptions are made about human behavior and motivation in order to uncover and examine the basic laws of political behavior and action. Scholars in this tradition argue that once these laws are discovered, models can be constructed that will help us understand and predict political behavior (Levi 1988). In their deductive model, Rational Choice scholars look to the real world to see if their model is right (test the model). For these scholars, understanding real outcomes is not the first point—creating, elaborating, refining a theory of politics is (Weingast 1996).

The implications of this scientific orientation are substantial. Morris Fiorina, a highly regarded RC scholar at Harvard, put the issue in the following way: 'most PTI scholars are not as interested in a comprehensive understanding of some real institution or historical phenomenon, so much as in a deeper understanding of some theoretical principle or logic… [F]or most PTI scholars, breadth trumps depth; understanding 90 percent of the variance in one case is not as significant an achievement as understanding 10 percent of each of nine cases, especially if the cases vary across time and place' (Fiorina 1995, pp. 110–1).

The third 'New Institutionalist' approach emerges out of what might be considered a more traditional political science. Consequently, it has a quite different aim: Historical Institutionalists are primarily interested in understanding and explaining real world events and outcomes. As with the other approaches noted above, scholars working in this tradition also argue that one cannot explain particular historical outcomes without specifically examining the way in which the political institutions have shaped or structured the political process (Steinmo et al. 1992). But, unlike RC scholars

in particular, Historical Institutionalists came about their 'institutional' arguments inductively after testing a variety of alternative theories (i.e., Marxist, structural functionalist, culturalist, and rationalist) against the outcomes they observed. In other words, Historical Institutionalists are first interested in explaining an outcome (say, for example, why France and Britain have pursued such different styles of industrial Policy (Hall 1986) or why some welfare states generate more popular support than others (Rothstein 1998)); they then proceed to explore alternative explanations for the outcomes they observe.

Historical Institutionalists do not argue that institutions are the only important variables for understanding political outcomes. Quite the contrary, these scholars generally see institutions as intervening variables (or structuring variables) through which battles over interest, ideas, and power are fought. Institutions are thus the focal points in critical junctures in a historical path analysis because political battles are fought inside institutions and over the design of future institutions.

5. *Institutional Change and Evolution*

Institutionalist scholars from a variety of perspectives are currently focusing on one of the most interesting and vexing issues left hanging in the institutionalist literature—explaining institutional change. Institutional stability (or equilibrium) is far easier to explain than change: institutional change means increased uncertainty because any particular set of institutions are embedded in a variety of other institutions; it is difficult to accurately predict the long run consequences of even small rule changes. Once again, every institution favors some interests and ideas over others, and therefore the advantaged are generally loath to change the status quo. Indeed, even current losers may decide that the status quo is preferable to some uncertain future in which they could be even further disadvantaged.

Of course, institutions do change, and sometimes they can change quite dramatically. Still, we as yet have no clear theory of how and why they change (with the exception of the 'punctuated equilibrium' model derived from Steven Gould's evolutionary theory). But clearly this single model of change is inadequate. It amounts to little more than suggesting 'stuff happens.' While it may be true that 'stuff happens,' we need to have a better understanding of why even exogenous shocks get refracted in particular ways. Additionally, we need a better understanding of the role of endogenous variables in the change process.

A variety of scholars have begun to explore the dilemma of institutional change by building more explicitly on evolutionary models suggested in the life sciences (Arthur 1994, Jervis 1997). Indeed, perhaps one of the most interesting outgrowths of the 'institutionalist turn' of political science in recent years may be that political science may begin to move away from its fascination with physics and its emphasis on absolute laws that hold across time and space. An institutionalist political science may instead look more like the life sciences—with their emphasis on contingency, context, and environment. In this event, the goal will be more to explain past (evolutionary) outcomes and less to predict future ones.

See also: Institutionalization; Institutions; Law: New Institutionalism; Norms; Political Science: Overview; Political Sociology; Rational Choice in Politics

Bibliography

Almond G A, Verba S 1963 *The Civic Culture; Political Attitudes and Democracy in Five Nations*. Princeton University Press, Princeton, NJ

Arthur B 1994 *Increasing Returns and Path Dependency in the Economy*. University of Michigan Press, Ann Arbor, MI

Bendix R, Brand C M 1968 *State and Society; a Reader in Comparative Political Sociology*. Little Brown, Boston

Bill J A, Hardgrave R L 1973 *Comparative Politics; the Quest for Theory*. Merrill, Columbus, OH

DiMaggio P J, Powell W W 1991 Introduction. In: DiMaggio P J, Powell W W (eds.) *The New Institutionalism in Organizational Analysis*. The University of Chicago Press, Chicago, pp. 1–38

Ekstein H 1960 *Pressure Group Politics: The Case of the Medical Association*. Stanford University Press, Stanford, CA

Evans P, Rueschemeyer D, Skocpol T 1985 *Bringing The State Back In*. Cambridge University Press, New York

Fenno R F 1966 *The Power of the Purse: Appropriations Politics in Congress*. Little, Brown and Company, Boston

Fiorina M 1995 Rational choice and the new (questionable) Institutionalism. *Polity* **28**(1 (Fall)): 107–115

Greenstein F I 1963 *The American Party System and the American People*. Prentice-Hall, Englewood Cliffs, NJ

Hall P 1986 *Governing the Economy*. Oxford University Press, New York

Hall P A, Taylor R C R 1996 Political science and the three new institutionalisms. *Political Studies* **44**: 936–57

Herring P 1940 *The Politics of Democracy; American Parties in Action*. W. W. Norton & Company, New York

Huntington S 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, CT

Immergut E 1992 *Health Politics: Interests and Institutions in Western Europe*. Cambridge University Press, New York

Jervis R 1997 *System Effects*. Princeton University Press, Princeton, NJ

Katzenstein P 1978 *Between Power and Plenty*. University of Wisconsin Press, Madison, WI

Key V O 1947 *Politics, Parties and Pressure Groups*. Thomas Y. Crowell Company, New York

Levi M 1988 *Of Rule and Revenue*. University of California Press, Berkeley, CA

March J, Olsen J 1989 *Rediscovering Institutions*. Free Press, New York

Miliband R 1969 *The State in Capitalist Society*. Weidenfeld & Nicolson, London

North D C 1990 *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, UK
Polsby N, Gallaher M, Rundquist B S 1969 The growth of the seniority system in the United States house of representatives. *American Political Science Review* **63**: 787–807
Przeworski A, Teune H 1970 *The Logic of Comparative Social Inquiry*. Wiley-Interscience, New York
Rothstein B 1992 Labor market institutions and working class strength. In: Steinmo S, Thelen K, Longstreth F (eds.) *Historical Institutionalism in Comparative Politics:State, Society, and Economy*. Cambridge University Press, New York
Rothstein B 1998 *Just Institutions Matter: The Moral and Political Logic of the Universal Welfare State*. Cambridge University Press, New York
Schattschneider E E 1960 *The Semisovereign People: A Realist's View of Democracy in America*. Holt, New York
Selznick P 1949 *TVA and the Grass Roots: A Study in The Sociology of Formal Organization*. University of California Press, Berkeley, CA
Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge University Press, Cambridge, UK
Steinmo S, Thelen K, Longstreth F (eds.) 1992 *Structuring Politics: Historical Institutionalism in Comparative Analysis*. Cambridge University Press, New York
Verba S 1967 Some dilemmas in comparative research. *World Politics* **20**: 111–27
Weingast B 1996 Political institutions: rational choice perspectives. Goodin R E (ed.) *A New Handbook of Political Science*. Oxford University Press, Oxford, UK, pp. 167–90
Wilson W 1891 *Congressional Government: A Study in American Politics*. Houghton, Boston

S. Steinmo

Institutionalization

Institutionalization as a research topic reached its heyday in the 1970s when spates of articles featuring the 'institutionalization of (fill in the blank)' were published in major journals. The objects institutionalized included the United Nations (Keohane 1969), cultural persistence (Zucker 1977), voting patterns (Przeworski 1975) and political parties (Wellhofer 1974), among others. In the years following this explosion, there appeared other works focusing on a number of clearly political institutions such as the presidency (Ragsdale and Theiss 1997) and the California legislature (Squire 1992). An overview of these articles shows that institutionalization has meant different things to different authors. The common theme, however, was that certain institutional features had persisted and it was the continuity of the pattern which was called institutionalization. An explication of the single most cited article on institutionalization in political science serves as an example of both the concept and its application.

Building on work of Witmer (1964) and Price (1964), Polsby (1968) published 'The institutionalization of the United States House of Representatives' and, with colleagues, published a companion piece (1969) on the rise of seniority in the United States House of Representatives. In the 1968 article, Polsby cites the political development literature including Weber (1947), Durkheim (1938) and Huntington (1965). In examining the development of the US Houses of Representatives Polsby argues from organization theory (e.g., Selznik (1953) that 'the institution' of the US House of Representatives becomes separable from its environment and develops as an institution within the US political system that signals its adaptability and development. The three characteristics Polsby studies are:

(a) boundaries: the differentiation of the institution sufficiently differentiated from its environment so that members need special skills to rise to leadership positions;

(b) internal complexity: the rise of committees and a seniority system which captures the benefits of specialization, and

(c) universalism: the rise of nonpartisanship (or objectivity) in determining whom the House members will be in the case of disputed elections.

The development of the House as an institution necessitating special skills for its leaders is measured by the length of time that it takes a member to become the Speaker. Polsby shows change occurring over the nineteenth century where early in the century Henry Clay could became Speaker in his first term whereas by the end of the century the Speakers on average had served 20 years before becoming Speaker. The 'institutionalization of specialization' in the House was demonstrated by the absence of seniority violations on committees occurring early in the twentieth century. Likewise the rise of fairness is measured by the criteria applied to election disputes—that is, as bi-partisan voting behavior on contested seats rose, Polsby claims that objectivity within the House also rose to the status of a norm.

Each of these patterns of behavior shifted sharply in the 1890–1910 period—Polsby called this period the 'Big Bang period' —in the institutionalization of the House. Many articles on the institutionalization of other legislative bodies followed. The bodies studied ranged from state legislatures to the mother of Parliaments—the House of Commons (Hibbing 1988). Much of the early work treated institutionalization as though it was a full-blown theory. Polsby, however, argued more accurately that he had described a tripartite dependent variable over time to show that discernable patterns had developed. The view of institutionalization as patterned behavior is closer to sociological theories of institutions and it is to this tradition that I turn.

Institutions are at the core of sociology (Durkheim 1938) and in general one can view an institution as organized, standing procedures which taken together constitute 'the rules of the game' (Easton 1981). Such

rules are seen as exogenous to participants. This general view has not led to clear and clean usage. In the analysis that follows I take the definition of Jepperson and use his conceptual scheme:

> Institution represents a social order or pattern that has attained a certain state or property; institutionalization denotes the process of such attainment (1991, p. 145).

On this view, an institution is a social pattern that stands over time and, when there are departures from this pattern, they are sanctioned without collective mobilization. Political institutions like legislatures, bureaucracies and executives are therefore well suited to institutional analysis.

The State Department of the United States government is such an institution. Members in the organization are rewarded or punished for their behavior in regard to the 'rules of the game' and over time rules can change. Precisely because institutions can change institutionalization *per se* does not, as some have thought, necessarily require stability or survival. In fact, shortly after Polsby et al. (1969) and Abram and Cooper (1968) published their work on the institutionalization of seniority, the US House of Representatives essentially ended seniority or, at a minimum, severely constrained it by deposing four senior chairmen.

Institutional change is dependent on context, which can be cultural, environmental or structural. It is this relativity that makes pinning down the meaning of institutionalization hard to do, because the property may change. The best way to think of institutionalization is to see it as a particular state of an institution. Thus, in Polsby's classic article, differentiation and specialization in the House were institutionalized, as attested to by the increasingly lengthy career path of Speakers and the rise of a pure seniority system, both important properties of the House of Representatives at the time of the studies.

The sociological concept of institutionalization began to fall out of favor in political science due largely to two factors: first, most of what could be studied *a la* Polsby was studied and second and more importantly, the 'new' institutionalism driven by economic reasoning came to the fore in Political Science. The analytic tradition of Coase (1937), Williamson (1985) and North (1981) and others (Nelson and Winter 1982) modified the traditional maximizing assumptions of classical economics where individuals maximize behavior over stable preference orderings. Their modifications are that individuals do maximize, though they do so constrained by cognitive limits, incomplete information and difficulties in enforcing agreements. Thus institutions arise to solve a problem and to maintain the solution when the benefits derived from the extant institution exceed the costs. Institutions essentially do this by reducing the uncertainty associated with transactions.

William Riker and his students (especially Shepsle) brought this economic approach to the study of political institutions. According to this approach, the US Congress creates committees and rules in order to solve the recycling or instability problem. In Shepsle's (1986) own words, political institutions are

> *ex ante* agreements about a structure of cooperation … which economize on transaction costs, reduce opportunism and other forms of agency slippage

—thereby increasing cooperation. This view of legislative institutions is disputed by Krehbiel (1991) and Krehbiel and Gilligan (1987, 1988) who argue that Congress established committees not to solve the cycling problem but rather to solve its information problems. Both viewpoints construe the Congress as an institution where policy is dependent upon the agenda-setting powers inherent in legislative rules. In short 'the new institutionalism' sees institutions as choices made by actors for self-interested reasons and the survival of institutions and their features is cost-benefit determined. The initial problem being solved may, among others, be informational, cycling, credible commitment or coordination. Thus the costs of starting the institution are lower than the benefits conferred. Likewise, the continued existence of the institution depends upon benefits exceeding costs. On this view the process of institutionalization over time is planned, chosen and sustained as long as benefits exceed costs.

The sociological response to the inroads made by economic treatments of institutions approach has been work led by March (1981), March and Olson (1984) and Meyer (1977, 1980), which argues that, contrary to economic theory, the institutionalization process is one where organizational structure evolves over time in an unplanned, adaptive, historically determined fashion. In political science, the work of Terry Moe (1987) fits this view, as does work by Ragsdale and Theis (1997) on the American presidency. Moe argues that rational choice institutionalism overemphasizes the formal mechanisms of legislative control at the cost of neglecting more fundamental, indirect, unintentional and systemic societal causes.

The field of international relations in political science also features the conflict between rational choice models of institutionalization and sociological theories, especially in the work of Krasner (1983), Keohane (1984, 1986) and 1988), and Young (1986). The argument is essentially that the economic model cannot account for international institutions because said institutions are clearly a long way from efficient. The sociological view (especially Meyer on systems) is that international institutions do not merely reflect the preferences and power of the units constituting them; rather they shape preferences and power. Recent illustrations of this perspective include Bennett's (1997) work on alliance duration and Martin's (1995) work on firms' preferences. The historical institutionalism of Skowronek (1982), Skocpol (1992) and others also adapts to this argument, holding that the de-

velopment of the American state and its subsequent policies are not the result of a logical, chosen rational process, rather both the state and its policies are the product of exceptional events, individuals, and circumstances that at crucial historical times determine or co-determine the future of institutional development. Moreover, unintended consequences often outweigh the intended and anticipated consequences.

In sum, early work on institutionalization by Polsby (1968), Cooper (1968) and Price (1964) was already embedded in the sociological tradition. It later was overtaken by economics-based theories of institutions and institutionalism illustrated by the work of Shepsle (1986), Weingast (1988, 1995), McCubbins (1987, 1993), Cox (1993), Krehbiel (1991) and others. The sociological approach has begun to revive as March (1981) and March and Olson (1984) and in the study of American politics Terry Moe's (1987) works have come to be widely read and cited. Sociological views of institutionalization seem to dominate the field of international relations with the work of March (1981) and Meyer (1980) being most influential on scholars in this field.

Unsolved puzzles in this field include the following: first, what is an institution? The economic tradition views institutions as the product of human designs where individuals behave instrumentally in order to achieve ends. The sociological tradition views institutions as emerging from human activity, but not necessarily, or even often, the product of conscious design. This conflict will be resolved by careful historical-empirical studies that ascertain the origins of institutions such as committees, bureaucracies and so forth. Second, do institutions reflect the preferences of individual or corporate actions or do they reflect collective outcomes of group interactions? This conflict will be more difficult to resolve because, in economics, preferences are assumed, not studied and in sociology they are not variables in any individual sense. Third, and most relevant for institutionalization, do institutions respond quickly to individual interests and exogenous change as economic theory would have us believe, or do they respond glacially as sociology would have us believe?

Consider the original work on the House of Representatives where change occurs infrequently but when it occurs, it happens over a 20-year period—'the Big Bang 1890–1910'—and then is stable over another 50 years. This appears to be glacial change with punctuated equilibrium in the sociological tradition. The economic perspective leads us to believe that change is more rapid and driven by standard cost-benefit analysis. While there is no seminal work on the institutionalization of an institution in the economic tradition, Cox and McCubbins (1993) and Aldrich's (1995) work on political parties are a good beginning. Parties are created to solve problems, of coordination, distribution, etc. and, over time, parties change as exogenous shocks and members' preferences shift. A study of well-known institutions, such as the House of Representatives, that contrasts the sociological and economic views would make a major contribution by virtue of its being able to pin down the reasons for creating institutions (see Cooper (1970) on the origins of committees for an example) and the rate at which they change. At present, excellent studies of various institutions from one perspective—see Meyer (1977) on junior colleges as institutions—contribute to one perspective but the same institutions are not studied from the other discipline's perspective. Johnson and Libecap's (1994) study of the causes of bureaucratic reform in the 1880s in the USA is a classic example of institutional change studied from the perspective of economic theory which could be easily restudied from a sociological perspective. Excellent studies of separate institutions (often at different times) may enhance discipline specific research but until the same institution is examined from both perspectives over the same time period, knowledge in this area will be partial and fragmented.

See also: Bounded Rationality; Bureaucracy and Bureaucratization; Institutional Economic Thought; Institutionalism; Law: New Institutionalism; Legislatures: United States; Rational Choice in Politics; Sanctions in Political Science

Bibliography

Abram M, Cooper J 1968 The rise of seniority in the House of Representatives. *Polity* **1**: 53–88

Aldrich J 1995 *Why Parties?* University of Chicago Press, Chicago

Coarse R 1937 The nature of the firm. *Economica* **16**: 386–405

Cooper J 1970 *The Origins of the Standing Committees and the Development of the Modern House*. Rice University, Houston, TX

Cox G, McCubbins M D 1993 *Legislative Leviathan*. University of California Press, Berkeley, CA

Durkheim E 1938 *The Rules of Sociological Method*. Free Press, Glencoe, IL

Easton D 1981 *The Political System: An Inquiry into the State of Political Science*. Alfred A. Knopf, New York

Hibbing J R 1988 Legislative change and institutionalization with illustrations from the British House of Commons. *American Journal of Political Science* **32**: 681–712

Huntington S P 1965 Political development and political decay. *World Politics* **17**: 386–430

Jepperson R 1991 Institutions, institutional effects, and institutionalism. In: Powell W W, DiMaggio P J (eds.) *The New Institutionalism in Organizational Analysis*. University of Chicago Press, Chicago

Johnson R, Libecap G 1994 *The Federal Civil Service and the Problem of Bureaucracy: The Economics and Politics of Institutional Change*. University of Chicago Press and NBER, Chicago

Keohane R 1969 Institutionalization in the United Nations General Assembly. *International Organizations* **23**: 859–96

Keohane R 1986 Reciprocity in international relations. *International Organization* **40**: 1–27

Keohane R 1984 *After Hegemony*. Princeton Press, Princeton, NJ
Keohane R 1988 Institutional institutions: Two research programs. *International Studies Quarterly* **32**: 379–96
Krasner S 1983 *International Regimes*. Cornell University Press, Ithaca, NY
Krehbiel K 1991 *Information and Legislative Organization*. University of Michigan Press, Ann Arbor, MI
Krehbiel K, Gilligan T 1987 Collective decision-making and standing committees: An informational rationale for restrictive amendment procedures. *Journal of Law, Economics, and Organization* **3**: 287–335
Krehbiel K, Gilligan T 1988 Complex rules and congressional outcomes: An event study of energy tax legislation. *Journal of Politics* **50**: 625–54
March J 1981 Decisions in organizations and theories of choice. In: Van de Van A H, Joyce W F (eds.) *Perspectives on Organization Design and Behavior*. Wiley, New York, pp. 205–44
March J, Olsen J 1984 The new institutionalism: Organizational factors in political life. *American Political Science Review* **78**: 734–49
Martin C 1995 Nature or nurture? Sources of firm preference for National Health reform. *American Political Science Review* **89**: 898–913
McCubbins M, Sullivan T 1987 *Congress: Structure and Policy*. Cambridge University Press, New York
Meyer J W 1977 The effects of education as an institution. *American Journal of Sociology* **83**: 55–77
Meyer J 1980 The world polity and the authority of the nation state. In: Bergesen A J (ed.) *Studies of the Modern World System*. Academic Press, San Diego, CA
Moe T 1987 Interests, institutions and positive theory. *Studies in American Political Development* **2**: 236–99
Nelson R R, Winter S G 1982 *An Evalutionary Theory of Economic Change*. Belknap Press of Harvard University Press, Cambridge, MA
North D 1981 *Structure and Change in Economic History*. W.W. Norton, New York
Polsby N W 1968 The institutionalization of the US House of Representatives. *American Political Science Review* **1**: 144–69
Polsby N W, Gallaher M, Rundquist B S 1969 The growth of the seniority system in the US House of Representatives. *American Political Science Review* **63**(3): 787–807
Price H D 1964 *The Congressman and the Electoral Arena*. Unpublished manuscript
Przeworski A 1975 Institutionalization of voting patterns, or is mobilization a source of decay? *American Political Science Review* **69**: 49–67
Ragsdale L, Theiss J 1997 Institutionalization of the American Presidency, 1924–1992. *American Journal of Political Science* **41**: 1280–318
Selznick P 1953 *TVA and the Grass Roots*. University of California Press, Berkeley, CA
Shepsle K 1986 Institutional equilibrium and equilibrium institutions. In: Weisberg H (ed.) *Political Science: The Science of Politics*. Agathon Press, New York, pp. 51–82
Skocpol T 1992 *Soldiers and Mothers: The Political Origins of Social Policy in the United States*. Belknap Press of Harvard University Press, Cambridge, MA
Skowronek S 1982 *Building a New American State: The Expansion of National Administrative Capacities, 1877–1920*. Cambridge University Press, Cambridge, UK
Squire P 1992 The theory of legislative institutionalization and the California Assembly. *Journal of Politics* **54**: 1026–54
Weber M 1947 *The Theory of Social and Economic Organization*. Oxford University Press, New York
Weingast B R, Marshall W J 1988 The industrial organization of Congress: Or why legislatures, like firms, are not organized as markets. *Journal of Political Economy* **96**: 1132–63
Weingast B, Shepsle K A (eds.) 1995 *Positive Theories of Congressional Institutions*. University of Michigan Press, Ann Arbor, MI
Wellhofer E S 1974 Political party development: Institutionalization, leadership recruitment and behavior. *American Journal of Political Science* **18**: 135–65
Williamson O E 1985 *The Economic Institutions of Capitalism*. Free Press, London
Witmer R R 1964 The aging of the house. *Political Science Quarterly* **79**: 526–41
Young O R 1986 After hegemony—cooperation and discord in the world political economy—Keohane, R O. *World Politics* **39**: 104–22
Zucker L G 1977 The role of institutionalization in cultural persistence. *American Sociological Review* **42**: 726–43

D. W. Brady

Institutions

In the most general meaning of the term, 'institutions' are defined as constraints or rules that induce stability in human interaction. All human societies have developed a variety of institutions which deal with recurrent basic problems in social life, for example, institutions regulating kinship and family life or economic institutions. In this article, the concept as it is used in contemporary social science and some aspects related to the explanation of institutions will be reviewed.

1. *The Concept 'Institutions'*

In everyday language and in the social sciences, the term 'institutions' is often used in a broad and sometimes amorphous way. It is quite common to denote organizations and other corporate actors such as business firms, political parties, schools, and public administrations as 'institutions.' On the other hand, the term also, more specifically, refers to rules and other constraints of human interaction. 'Institutions' in this sense include conventions (for example, traffic rules), as well as informal and formal social norms. Furthermore the term is often used to describe complex systems of interrelated norms. For example, modern law can be considered as a set of formal norms proscribing or prescribing certain actions. Another aspect of the law is that it employs (mainly negative) sanctions to punish and prevent deviant behavior. In accordance with Max Weber's terms (1978), the law induces stable regularities of nondeviant behavior via

(threats of) sanctions that are employed by a specialized staff of actors. In this sense, the law may be conceived as an organization that enforces certain formal norms.

The concept 'institutions' is thus used in a dual way. First, it denotes rules, constraints, or norms of human interaction. Second, it describes the resulting stable patterns of interaction among a set of agents and the social mechanisms generating this outcome. This latter aspect is sometimes also expressed in the concept 'institutionalization' which plays a central role, for instance, in Talcott Parsons' theory of social systems. 'Institutionalization' means that certain cultural norms become elements of the actors' motivational dispositions, and that norm conformity is rewarded and deviant behavior appropriately punished by the application of sanctions. In Parsons' (1982, pp. 117–18) words, 'institutionalization is an articulation or integration of the actions of a plurality of actors in a specific type of situation in which the various actors accept jointly a set of harmonious rules regarding goals and procedures.'

To allude to the modern choice theoretic framework of game theory, institutions are first represented by the rules of the games actors are playing. These rules define, among other aspects, the set of possible actions (strategies) of the agents. They also specify the consequences of the actions (payoffs) and the available amount of information. Given these rules or constraints, agents choose an action (strategy) that is individually optimal. The optimal choices will, under general conditions, yield an equilibrium. An equilibrium is a combination of actions (strategies) such that no agent has an incentive to deviate from her chosen actions unilaterally, that is under the condition that the other actors choose their equilibrium actions. The stability of an equilibrium is due to the fact that no (rational) actor could gain by shifting to another strategy unilaterally. Consider now a population of actors who recurrently interact in a social situation governed by the rules of the game. Then it may happen that a certain equilibrium will eventually evolve. In other words, a stable pattern of social interactions may emerge under the constraints of the rules of the game.

An 'institution' can thus be conceived to be the set of constraints or rules which determine the opportunities and incentives of the relevant actors in a given situation. Second, interactions within these rules generate equilibria, which may be stable over time if the situation is recurrent. These equilibrium outcomes can be considered an integral component of the institution. To illustrate, consider a coordination situation. Imagine a time with considerable road traffic (e.g., horse-drawn carriages in the seventeenth century) but no universally accepted rules with respect to keeping left or right. If a central authority (the government) were to create the rule that pedestrians use the right side of the road and horse-drawn carriages use the left side, one source of potential accidents could be avoided, namely, inconsistent mutual reciprocal expectations with regard to the usage of public roads. The rule would, so to speak, 'solve' a coordination problem. A coordination problem can roughly be defined as an interaction with multiple equilibria. In the absence of any rules or 'focal points' (Th. C. Schelling), there would be ambiguity with regard to the different courses of action available to road users: should one use the right or the left side of the road? The actors probably do not have intrinsic preferences with regard to the direction of the rule, that is, whether to go right is better than to go left. But they prefer to adopt a rule that gives clear and unique prescriptions and prevents accidents if universally acknowledged, rather than ambiguous rules or rules which are not easily implemented (e.g., the rule that one should determine by chance whether to go right or left). Given that the government has publicly announced a rule, the participants easily form convergent expectations with regard to each other's behavior. In other words, an equilibrium of universal rule-oriented behavior emerges as a consequence of individually adaptive or optimizing behavior under the constraints of the rules of the game and of external constraints created by the government. This gives rise to a 'convention.' Following Lewis (1969), a convention may be defined as a behavioral regularity R in a population of agents who recurrently interact in a coordination situation if R is an equilibrium of the recurrent situation and if almost every member of the population prefers to conform to R on the condition that almost every other member of the population conforms to R too.

Conventions arising in so-called pure coordination situations (with no conflict of interest) are self-enforcing. This is so because any individual who deviates from R would punish himself. There is no incentive to break the rule. In other social situations, for example of the prisoner's dilemma type, there is an incentive to defect. A prisoner's dilemma is a social interaction such that universal cooperation is profitable to everyone, but it is even more profitable to defect unilaterally from cooperation. Cooperation is not an equilibrium situation in the single-shot prisoner's dilemma. In a population of agents who recurrently interact in a prisoner's dilemma, an institution of mutual cooperation may emerge if there is a policing mechanism such that members of the population of agents expect that nonconformity will be punished by (negative) sanctions. The policing or enforcement of cooperation could be achieved under appropriate social conditions due to sanctions that are created by the agents themselves. Of course, intervention by external, third parties (in the sense of Hobbes' *Leviathan*) would be another mechanism to enforce an institution. Generalizing Lewis' concept of convention, a social institution can be defined as a regularity in the behavior of the members of a

population that 'specifies behavior in specific recurrent situations, and is either self-policed or policed by some external authority' (Schotter 1981, p. 11; italics omitted). The self-enforcement of an institution may be due to the fact that the regularity is an equilibrium of a coordination problem. If the recurrent situation has similarities with a prisoner's dilemma, however, an equilibrium of cooperative behavior may result from self-created sanctions (e.g., of the 'tit for tat' type, cf. Axelrod 1984) or from policing by an external authority. The concept 'social institution' in this sense includes conventions as a special case, and also formal and informal social norms (see Lewis 1969, Schotter 1981, Young 1998 for sophisticated elaborations of this line of reasoning).

2. *Some Types of Institutions*

There are many different kinds of institutions. First, adopting the intuitive meaning of institutions as rules or constraints of behavior, it is useful to make the following distinctions. There are rules that regulate the personal life of human agents. These *personal* rules (e.g., the rule to go early to bed every night) resemble individual habits and individual routines. Some of these rules may be interpreted as attempts of individuals to bind themselves. Personal rules may be motivated by individuals' long-term interests to achieve self-control. But they are not social institutions in the proper sense. In contrast to personal rules, *social* rules affect social interactions, that is, recurrent situations of social interdependence. For example, it can be argued that social institutions typically regulate social dilemmas. Social dilemmas are situations where interdependent agents are faced with choices for which individual rationality does not yield a socially optimal outcome. The prisoner's dilemma and public good situations are cases in point. Furthermore, there are *formal* and *informal* institutions. Informal institutions do not rely on an external authority's monitoring and policing of the participants' behavior. These institutions include, for example, social norms of self-help among residents in rural communities (see Ellickson 1991). Other informal institutions are conventions. Though conventions are defined as self-enforcing, they often become codified in formal rules (e.g., time standards or traffic rules). Formal institutions are generally created and arranged by agents who are able to rely on third parties for monitoring and enforcement. Among the most elaborate formal institutions in modern societies one should mention the law. There are also complex formal institutions regulating political committees in political systems and collective decisions in large capitalist firms. However, it is important to notice that formal rules are embedded in informal norms. Committee power in democratic political bodies is determined to a large extent by informal norms that foster vote exchange (logrolling) (Coleman 1973, Shepsle and Weingast 1987) and only to a minor degree depends on formal rules. Thus, formal institutional constraints are often supplemented by informal rules and vice versa.

Second, it is useful to point out the different levels of rules that can be relevant in institutional analysis (see Ostrom 1990, Chap. 2). There is first the *constitutional* level. Constitutional rules define the most general context of the games agents are allowed to play. They determine the set of eligible actors and the principles used in crafting or changing more specific rules of the system. Constitutional rules are not only relevant in the arena of international and national politics but also in regional and local contexts. Second, there is the level of *collective choice*. Collective choice rules specify principles of policy-making. They help to generate unique collective decisions. Third, there is the *operational* level. Operational rules help to implement the decisions of the higher levels. They determine day-to-day decisions, for example decisions related to the monitoring and sanctioning the participants.

Third, institutions may be classified by the spheres of the society to which they belong and contribute. There are *economic* institutions regulating economic exchange, for example, property rights institutions. Other economic institutions include conventions with regard to media of exchange or the specific rules and constraints affecting economic behavior between and within firms (see Williamson 1985). *Political* institutions provide basic rules and procedures to achieve collective decisions. The most important political institutions in contemporary societies are, of course, the constitutions of nation states, for example, the constitution of the United States of America. There are also international political institutions. Besides these basic ground rules of international and national politics there are more specific institutions and laws dealing with the operation of the democratic system. Every human society has developed institutions shaping family life and kinship relations. Modern societies have created a differentiated system of educational institutions. There are also cultural and religious institutions of different kinds.

3. *Explaining Institutions*

The analysis of institutions has always been considered to be one of the most important tasks in social theory. For example, Emile Durkheim (1978, pp. 82–3) emphasized that sociology's subject was 'social facts,' in particular social institutions. A desideratum of sociological analysis, according to Durkheim, is the discovery of the mechanisms, causes, and effects of institutions in the diverse spheres of societal life.

3.1 *Effects of Institutions*

A starting point of many explanations is the study of the effects or consequences of institutions. Not only functionalist analyses in the proper sense, but also

contributions from diverse theoretical perspectives in the social sciences argue that institutions may have beneficial consequences.

Institutions regularize social life. Rule-oriented behavior reduces uncertainty and increases the predictability of others' actions in situations of social interdependence. The institutionalization of social norms in Parsons' sense is a key to the solution of the Hobbesian problem of order. In a 'state of nature' without an enforcing agency, mutual cooperation, for example, the recognition of private property rights, would be profitable, but there are strong incentives to defect from mutual agreements. The institutionalization of social norms of reciprocity and solidarity via their internalization into the actors' motivational systems would be, according to Parsons, a solution to the Hobbesian problem (Parsons 1937). Thus, institutions and the institutionalization of appropriate social norms may foster cooperation and increase efficiency (in the Pareto sense of making agents better off without harming anyone in society).

Institutions such as exclusive private property rights with respect to economic resources are prone to enhance a society's aggregate welfare because they reduce transaction costs (North 1990). If transaction costs within an economic system decrease, the economy will generally become more productive, provided the costs of creating and enforcing these institutions are not too high (see Eggertsson 1990 for a survey of the literature). To allude to a recent institutional innovation, it seems plausible to argue that global communication networks (Internet) constitute an institutional framework that reduces some kinds of transaction costs (in particular costs of acquiring information) in long-distance economic relations. Many institutions, such as property rights, can be said to be part of a society's social capital (Coleman 1990, Chap. 12). In this sense, institutions are a productive asset because they enhance the efficiency, the welfare or, in other words, the well-being of most members of a society. Furthermore, they can have the properties of public or collective goods. That is, they are goods such that it is not feasible easily to exclude agents from their consumption if these goods are available to anyone. In other words, institutions with collective goods properties can be profitably 'consumed' by all those actors who wish to benefit from them, not only by those agents who contributed to the costs of their 'production.'

This public-good characteristic of many institutions makes it clear that the potential beneficial effects of institutions cannot sufficiently explain the emergence and change of institutions. This is so because public goods create incentives to a free ride. If the formation and maintenance of an institution is costly, the amount of voluntary contributions to these costs will generally be too low, that is suboptimal (Olson 1971). At least for this reason, simple functionalist accounts of efficient institutions are incomplete.

3.2 Mechanisms of Institutional Change

There are many different types of processes that are prone to generate or change social institutions. First, there is the extreme type of institutional change by conscious design. In particular, formal institutions such as legal norms are in many cases designed purposively. Parliamentary committees in modern democracies regularly propose such changes. However, many of these changes are accompanied or supplemented by changing informal institutions in the larger society. These informal rules generally change more incrementally.

Second, institutional change may be due to evolutionary forces of an 'invisible hand' (Adam Smith). In this case, the goal-oriented actions of a large number of actors may in the long rule unintentionally produce results that seem to be due to conscious planning by a central authority. But, in fact, the institution is a byproduct of each individual's activity to locally adapt to its circumstances. Institutions of this kind are called 'spontaneous orders' (F. A. Hayek). It can be argued that many conventions have emerged as spontaneous orders (Sugden 1986, Young 1998) in the first place. Once adopted, they may be included in a cultural tradition and will be taught to children by their parents (e.g., table manners).

Invisible-hand processes of institutions and their explanation (Ullmann-Margalit 1978) are attractive because they may yield theoretical arguments with some 'depth.' They explain facts on the collective level of social institutions by principles of purposive action and by a 'generating mechanism' (Boudon 1979). Work in the tradition of evolutionary game theory has expressed these mechanisms more explicitly. It can be shown that certain institutions emerge as special kinds of (stochastic) equilibria in games played by members of a population of agents who recurrently interact in the 'long run' (Young 1998).

Obviously, most empirically observable institutions will not be explicable exclusively by one of these two extreme or ideal types of mechanisms but are probably complex combinations of different types of processes. However, the study of theoretical models representing some aspects of a complex reality in an idealized way may be understood as a theoretical 'thought experiment' (*Gedankenexperiment*). Such thought experiments enhance the clarity of theoretical arguments and provide an antidote against simplistic and empirically dubious 'conspiracy theories' of institutions (K. R. Popper) which argue that most social institutions are explicable in terms of the joint efforts of some group of powerful actors who want to foster their material interests.

3.3 Are Institutions Efficient?

Given the functionalist account, which points out that institutions are beneficial, one may ask whether in fact every institution is efficient. There is indeed some

evidence on efficient rules. Wittman (1982) has investigated various rules of thumb in certain road traffic situations and comes to the conclusion that only the most efficient rules will be used. Rules have evolved such that two types of costs are optimized, namely, costs associated with monitoring the rule and inappropriate incentives created by the rule. A similar view is expressed in Ellickson's (1991) book on informal institutions. However, there is ample evidence on the persistence of inefficient institutions (see Eggertsson 1990, Chap. 8). In the Arab world, inefficient institutions regulating economic exchange have persisted for thousands of years. For example, a form of exchange called the *suq* (C. Geertz) is characterized by high bargaining and measurement costs. These transaction costs could in principle be reduced by major institutional changes. But such changes have not been observed. Moreover, economic historian North (1990, p. 6) points out that the 'central puzzle of human history is to account for the widely divergent paths of historical change. How have societies diverged? What accounts for their widely disparate performance characteristics?' It seems safe to say, according to North, that market processes of competition did not weed out inferior institutions but that inefficient institutions have proved to be highly stable.

There are many different arguments that could explain the persistence of inefficient institutions:

(a) Some social scientists, in particular in anthropology, suggest that institutions are inexplicable by functionalist or rationalist ideas. According to this approach, (cultural) institutions mainly serve 'symbolic' functions (Sahlins 1976). A different approach, however, points out the signaling properties of certain behavioral regularities (social norms) that *prima facie* do not contribute to efficiency. These institutions may be interpreted as consequences of signaling behavior. The conformity to 'ornaments' of social life, such as clothing norms, table manners, and others, may be perceived as a signal that the actor is cooperative in the more fundamental interaction problems of family, political, or business life (Posner 2000).

(b) Inequality in bargaining power can have the effect of selecting institutions that primarily serve the interests of the more powerful agents or groups in a society. Thus, new (formal) institutions need not foster the aggregate welfare of a society or efficiency (North 1990, Chap. 6).

(c) There are many recurrent interaction situations with multiple equilibria, for instance coordination games or iterated prisoner's dilemma games. In cases like these, the outcome may be indeterminate. Furthermore, there is 'path dependence': Small differences in initial conditions and chance circumstances in structurally similar social interactions yield different paths of historical development and therefore different institutional outcomes (see North 1990, Chap. 11, Greif 1994, Young 1998).

(d) Limited rationality and incomplete information of human agents may be obstacles to attaining efficient institutions (North 1990, Chap. 3).

(e) As far as efficiency gains from institutional change are public or collective goods there may be a lack of appropriate incentives to contribute to the costs (see Olson 1971, Ostrom 1990).

4. *The Evolutionary Approach to Institutions*

Recent research on the emergence of institutions emphasizes that not all institutions are necessarily efficient. This body of work uses ideas from evolutionary game theory. It is generally consistent with all of the above mentioned arguments on the inefficiency of institutions. Historically, the roots of evolutionary explanations of institutions are ideas from the tradition of the Scottish Moralists (in particular David Hume) and the Austrian School of Economics (Carl Menger, Friedrich A. von Hayek). In constructing an evolutionary approach, one must keep in mind that there are important differences between biological and social evolution. Theories of social evolution must contain appropriate behavioral assumptions of adaptive behavior or learning processes. The evolutionary approach can be characterized by the following core ideas (see Young 1986). Institutions are the products of interactions within large populations of actors over considerable periods of time. These agents are boundedly rational and act under conditions of limited information. This means that actors may fail to figure out responses to other people's actions which are optimal. However, *under appropriate circumstances*, the *evolutionary process* may in the long run (that is, in the presence of stochastic shocks which correspond roughly with mutations in biological evolution) realize equilibrium outcomes such that efficient institutions persist. But this result may well be an exception rather than the rule. Generally, the evolutionary path will depend on initial conditions. That is, structurally similar social systems may develop qualitatively distinct institutions with considerable differences with regard to the degree of 'efficiency.' A particularly important aspect of initial conditions are 'cultural beliefs' in Greif's (1994) sense. These beliefs are embedded into a society's culture and are a product of history. They are employed as equilibrium selection devices in those interaction situations with multiple equilibria. Therefore, different sets of cultural beliefs may yield different institutional outcomes (at least in the medium run). Greif (1994) illustrates this with the help of a comparison of the cultural beliefs in 'collectivist' and 'individualistic' societies, using the historical data on Maghribini and Genoese traders in the Middle Ages. An evolutionary approach should, in principle, point out the conditions that generate distinct cultural beliefs (and other equilibrium devices such as fairness norms) and the paths of an evolu-

tionary transition from one equilibrium to another (possibly more efficient) equilibrium. The evolutionary approach is a promising research program, but much further work is needed to fully understand and explain concrete cases of empirically observed institutions.

See also: Action, Collective; Collective Behavior, Sociology of; Conventions and Norms: Philosophical Aspects; Dubos, René Jules (1901–82); Durkheim, Emile (1858–1917); Financial Institutions in Economic Development; Functionalism in Sociology; Game Theory; Hobbes, Thomas (1588–1679); Law and Society: Sociolegal Studies; Organizations, Sociology of; Prisoner's Dilemma, One-shot and Iterated; Public Goods: International; Rational Choice Theory in Sociology; Rule of Law; Smith, Adam (1723–90)

Bibliography

Axelrod R 1984 *The Evolution of Cooperation*. Basic Books, New York

Boudon R 1979 Generating models as a research strategy. In: Merton R K, Coleman J S, Rossi P H (eds.) *Qualitative and Quantitative Social Research: Papers in Honor of Paul Lazarsfeld*. Free Press, New York

Coleman J S 1973 *The Mathematics of Collective Action*. Heinemann, London

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Durkheim E 1978 Sociology and the social sciences. In: Traugott M (ed.) *Emile Durkheim on Institutional Analysis*. University of Chicago Press, Chicago

Eggertsson T 1990 *Economic Behavior and Institutions*. Cambridge University Press, Cambridge, UK

Ellickson R C 1991 *Order without Law*. Harvard University Press, Cambridge, MA

Greif A 1994 Cultural beliefs and the organization of society. *Journal of Political Economy* **102**: 912–50

Lewis D K 1969 *Convention*. Harvard University Press, Cambridge, MA

North D C 1990 *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, UK

Olson M 1971 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA

Ostrom E 1990 *Governing the Commons. The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK

Parsons T 1937 *The Structure of Social Action*. Free Press, New York

Parsons T 1982 *On Institutions and Social Evolution* [ed. Mayhew L H] University of Chicago Press, Chicago

Posner E 2000 *Law and Social Norms*. Harvard University Press, Cambridge, MA

Sahlins M 1976 *Culture and Practical Reason*. University of Chicago Press, Chicago

Schotter A 1981 *The Economic Theory of Social Institutions*. Cambridge University Press, Cambridge, UK

Shepsle K A, Weingast B W 1987 The institutional foundations of committee power. *American Political Science Review* **81**: 85–104

Sugden R 1986 *The Economics of Rights, Cooperation and Welfare*. Blackwell, Oxford, UK

Ullmann-Margalit E 1978 Invisible-hand explanations. *Synthese* **39**: 263–91

Weber M 1978 *Wirtschaft und Gesellschaft*, 3rd edn. J. C. B. Mohr, Tübingen, Germany

Williamson O E 1985 *The Economic Institutions of Capitalism*. Free Press, New York

Wittman D 1982 Efficient rules in highway safety and sports activity. *American Economic Review* **72**: 78–90

Young H P 1998 *Individual Strategy and Social Structure*. Princeton University Press, Princeton, NJ

T. R. Voss

Instructional Design

1. Instructional Design and its Relation to Instructional Technology

The term instructional design (ID) refers to the systematic and professional provisions for education or training. Considerations regarding planned instruction have been made at least as long as there have been institutions for instruction and training. The term ID itself appeared for the first time in the USA in the midst of the twentieth century.

ID and instructional (or educational) technology (IT) are related terms with an overlapping meaning. Both expressions are used in a broad as well as in a narrow sense. The term ID is used in a variety of ways with different ranges of meaning:

(a) The narrowest variant concerns only the planning of an instructional treatment.

(b) The somewhat wider usage concerns both planning and constructing of what is needed for a certain instructional solution.

(c) The even wider interpretation adds the implementation of the instructional system or program in a particular context and the evaluation of outcomes. ID is most often used with this interpretation and sometimes referred to as IT with the same broad meaning.

(d) The widest interpretation adds to the previous meaning by including maintenance and management of the instructional system or program over a period of time; this includes conducting ongoing evaluations, planning, and developing upgrades as well as managing all of these processes organizationally and economically. This broad interpretation of ID corresponds with the most comprehensive definition of IT.

ID, under any of these interpretations, can refer to these education or training processes with regard to varying degrees of scope (unit of instruction, lesson, curriculum, program, system, etc.). At the macrolevel, ID concerns instruction of whole curricula, programs, or systems that will cover months or years of instruction; this is called large-scale ID, or, alternatively, curriculum construction. At the middle or mesolevel,

the object of ID are typical lessons that cover hours or weeks of instruction. The microlevel concerns a period of minutes, or short parts of a lesson, such as giving exercises or demonstrating illustrative examples. The most frequent use of theories and models of ID refers to the mesolevel. The limits between the three levels are fluid and not always clearly defined.

Some authors tend to reduce ID to the mere planning of instruction or to restrict IT to hardware and software, but these narrower interpretations do not fulfill the conditions of providing efficient instruction, because efficient instruction has to take into account all relevant factors (e.g., learners, environmental factors) in a systematic and systemic approach.

2. *A Short Historical Overview to Characterize Relevant Aspects of ID*

In the 1930s educational scholars in the USA began to investigate questions like: will students who went through a certain, alternative high school curriculum be successful at the following college? These instructional scientists asked how to state instructional objectives more precisely by taking into account the learner's behavior. Furthermore, it was important to check if the alternative curriculum was implemented in the planned way. For this, the instructional goals were used as a benchmark and critical point of reference. Reaching the intended instructional goals was improved by continual revision of the course during the development and test phase. This cyclic process of goal-oriented checking and improvement of lessons and courses today is called 'formative evaluation.'

The development of ID as a special field of research and development was mainly stimulated by two goals to be achieved:

(a) training as many learners as possible within a short period of time; and

(b) using the most efficient available instructional means (e.g., media, such as written instructions and demonstration films, as well as practical exercises).

The Second World War introduced enormous training problems for the military in the USA. In order to address military training problems, the 'Division of Visual Aids for War Training' within the 'US Office of Education' produced hundreds of films and training handbooks. Military training became an example for the successful connection of research and development with professional practice. It started to prove to be useful that teams for instructional development should include instructional designers as well as subject matter experts and a 'producer.' From that time on, the necessity for professional ID standards and professional ID practitioners has been widely recognized.

In the middle of the twentieth century when learning psychology reached its peak, especially in North America, there was a systematic effort to apply learning research results to practice, including the design of instruction. Thus, important scholars of learning psychology implemented basic learning principles (operant conditioning in particular) in programmed instruction and teaching machines.

In 1965 Gagné published the influential book *The Conditions of Learning* which originally had a behavioristic orientation; later editions provided a cognitive orientation (Gagné 1985). Gagné's new procedure, again based on a theory of learning proposed a systematic choice concerning instructional methods corresponding to a classification of instructional goals by hierarchically ordered kinds of learning. The procedures for evaluation and control of learning success were improved in the 1960s; the term 'criterion-referenced measurement' emerged characterizing goal-oriented tests. One important addition to ID was to give precedence to 'needs assessment' as a necessary precursor to other already existing steps of the ID process. Therefore, ID models should not simply start with the establishment of instructional goals but with an analysis of which instructional goals are useful for anticipated audiences in a particular environment. In the middle of the 1970s a major 'watershed' of professionalism of instructional designers was reached in the USA. For the first time, instructional designers became responsible contract partners for course development ahead of subject matter experts. In Europe this development was delayed.

The turn from behavioral to cognitive psychology produced new approaches in learning theory. In the time of behaviorism the parsimonious theory of operant conditioning encouraged the opinion that the direct transferability of learning theories found in basic research is fairly problem-free. Later, however, criticism of these basic research theories—also of cognitive approaches—included their practical applicability as well as their theoretical coherence. As a reaction, approaches of learning psychology especially created for instructional purpose were developed. In the 1990s this type of learning theory is called 'educational learning theory' (ELT). Steiner (1997) proposed such an ELT by adding European views of learning.

In the last quarter of the twentieth century, learning psychology tended to investigate more and more complex instructional matters. The contributions of modern instructional psychology become increasingly diverse. During the last 25 years, approaches of ID have tried to take more systems thinking to ID (cf. Briggs 1977, Seels and Richey 1994, Jonnassen 1996, Dijkstra et al. 1997, and recently Richey and Fields 2000). This concerns

(a) the view on instructional systems, subsystems and its relationships to the environment;

(b) the development of highly systematic, often linear or branched design procedures; and

(c) more emphasis on systemic approaches of ID, which takes the inter-relation of all relevant factors of instruction into account.

Finally, more attention is given to the philosophical and theoretical grounding of ID principles (cf. Reigeluth 1999, Tennyson et al. 1997).

3. Present Applications of ID

The systematic application of ID generally contains the following subtasks:

(a) performing a needs assessment,
(b) identifying instructional goals,
(c) analyzing goals into subgoals and identifying prerequisite knowledge,
(d) defining what the learners have to learn to bridge the gap between prerequisite knowledge and subgoals,
(e) developing criterion-referenced tests and/or other assessment measures,
(f) developing instructional strategies,
(g) developing materials and preparing media,
(h) conducting formative evaluation, and
(i) conducting summative evaluation (a final evaluation of the whole instructional procedure).

It should be noted that these are typical tasks found in many current ID models; variations among professional practitioner models still exist. Moreover, from a systemic point of view these subtasks are interrelated and are usually not performed in a linear way or only a single time; rather, the practice of ID is interactive and responsive to individual situations (cf. Richey and Fields 2000).

ID approaches can be applied in all fields and situations that deal with education and training, including school settings, technical training, professional training, collaboration with human resource development professionals in industrial and professional organizations, higher education and university settings, as well as in distance and life-long learning situations—using all old and new kinds of media and telecommunications. In short, ID has matured into a broad profession with many connections to other professions and activities, including personnel and project management.

It is especially noteworthy that the practice of ID often involves a variety of experts, such as text designers, media designers, software programmers, subject matter experts, and learning specialists. Such a team is obviously necessary when the development involves technology-based learning or instruction since the required skills and expertise are likely to be distributed among a variety of specialists. When ID teams are involved, it is critical to have recognized standards and an established methodology.

4. Challenges for Future Theory, Research, and Development

Important challenges in the future for theory, research, and development are:

(a) Dealing with integrated sets of learning goals for teaching complex skills and higher-order skills. Such transferable skills become more and more important in our society (cf. van Merriënboer 1997).

(b) An expert can broaden his or her competency by integrating the individual knowledge *and* the resources of getting, storing, and processing information externally by computerized tools. To take this into account we need to develop more powerful, integrated techniques for instructional and task analysis.

(c) Fostering the systematic design of collaborative learning/team training, of new technological developments in distance education, in particular web-based instruction.

(d) We need models for designing media mixes instead of models for media selection.

(e) Dealing with highly heterogeneous target groups (differences in prior knowledge and skills, cultural diversity, older learners) and target groups with low motivation (e.g., 'teens at risk').

(f) Further development of effective forms of assessment. The need for assessment of efficient ID solutions is often underestimated.

(g) Automating ID processes using techniques of artificial intelligence. This can help to lower the cost, e.g., by generating tasks for a given item format and content, or to produce automatically instructional materials for a variety of products of the same type.

(h) Last but not least, a better theoretical foundation for ID, which takes into account the whole problem of providing *useful* knowledge for human beings in our world of nearly unlimited possibilities for getting immense information from everywhere.

See also: Computer-assisted Instruction; Curriculum as a Field of Educational Study; Distance Education; Educational Assessment: Major Developments; Educational Learning Theory; Educational Media; Instructional Psychology; Learning Theories and Educational Paradigms; Learning to Learn; Lifelong Learning and its Support with New Media: Cultural Concerns; Media, Uses of; Self-regulated Learning; Technology-supported Learning Environments

Bibliography

Briggs L (ed.) 1977 *Instructional Design: Principles and Applications*. Educational Technology Publications, Englewood Cliffs, NJ

Dijkstra S, Seel N, Schott F, Tennyson R D (eds.) 1997 *Instructional Design—International Perspectives. Vol. 2: Solving Instructional Design Problems*. Lawrence Erlbaum Associates, Mahwah, NJ

Gagné R M 1985 *The Conditions of Learning and Theory of Instruction*, 4th edn. Holt, Rinehart and Winston, New York

Jonnassen D H (ed.) 1996 *Handbook of Research for Educational Communications and Technology*. Macmillan, New York
Reigeluth C M (ed.) 1999 *Instructional-Design Theories and Models. Vol. II: A New Paradigm of Instructional Theory*. Lawrence Erlbaum Associates, Mahwah, NJ
Richey R C, Fields D F (eds.) 2000 *Instructional Design Competencies: The Standards*, 3rd edn. ERIC Clearinghouse on Information and Technology & The International Board of Standards for Training, Performance & Instruction, Syracuse, NY
Seels B, Richey R 1994 *Instructional Technology: The Definitions and Domains of the Field*. Association for Educational Communications and Technology, Washington, DC
Steiner G 1997 Educational learning theory. In: Tennyson R D, Schott F, Seel N, Dijkstra S (eds.) 1997 *Instructional Design—International Perspectives. Vol. 1: Theory, Research, and Models*. Lawrence Erlbaum Associates, Mahwah, NJ
Tennyson R D, Schott F, Seel N, Dijkstra S (eds.) 1997 *Instructional Design—International Perspectives. Vol. 1: Theory, Research, and Models*. Lawrence Erlbaum Associates, Mahwah, NJ
van Merriënboer J J G 1997 *Training Complex Cognitive Skills: A Four-Component Instructional Design Model for Technical Training*. Educational Technology Publications, Englewood Cliffs, NJ

F. Schott

Instructional Psychology

1. Introduction

Instructional psychology is the study of the processes and outcomes of human learning in a variety of educational and instructional settings, and of the nature and the design of environments that are appropriate to elicit those learning processes aiming at the attainment of competence and of a disposition toward skilled learning, thinking, and problem solving in a given domain. The term 'instructional psychology' was introduced in 1969 by Gagné and Rohwer (1969) as title of their literature review in the *Annual Review of Psychology* of a field of inquiry indicated until then as 'educational psychology.' Today one can say that while educational psychology is somewhat broader in scope than instructional psychology—involving, for instance, psychological aspects of childrearing in the family—both terms cover largely the same domain of research, because most work has been done on studying knowledge and skill acquisition in instructional settings. A major reason for this state of the art is that since the 1960s there has been a growing convergence between the psychology of learning and instruction on the one hand, and cognitive science on the other. As a result the mainstream of theory and research has focused on the analysis of knowledge, performance, and competence in a variety of complex cognitive task domains, and instructional psychology has during the 1980s 'become a vigorous part of the mainstream of research on human cognition and development' (Glaser and Bassok 1989, p. 631).

The question whether the tremendous amount of research in instructional psychology over the past decades has contributed to better educational practices remains controversial. In this respect the distinction between a disciplinary vs. an educational orientation in educational/instructional psychology is relevant. In the first case, instructional psychology is considered as an offshoot of psychology which mainly aims at contributing to the development of theory and methodology within the broader domain of the mother discipline. In the second case, the focus is rather on acquiring a better understanding of education as a basis for the improvement of educational practices. The disciplinary orientation has dominated for a large part of the twentieth century, especially during the associationist and behaviorist era in psychology. The prevailing type of research were laboratory studies characterized by a great concern for internal validity, and, thus, including a high degree of experimental precision. But this way of conducting research easily overlooked educationally important aspects, and has therefore been often criticized for its lack of classroom relevance. Whereas the discipline approach is certainly still alive today, the educational orientation in instructional psychology has gained more and more ground since the 1970s, and has by now resulted in an empirically underpinned knowledge base that can guide the analysis of teaching practices, but also the design of new, powerful learning environments.

A comprehensive review of this extensive field of research is outside the scope of this entry. Snow and Swanson (1992) have distinguished five essential components in a theory of learning from instruction: (a) description of desired end states or goals of instruction in a domain; (b) description of goal-relevant initial states of learners prior to instruction; (c) explanation of the transition processes from initial to desired states; (d) specification of instructional conditions that promote this transition; and (e) assessment of performance and instructional effects. This entry will focus selectively on three of these components, namely desired end states, transition processes, and instructional conditions. For a more thorough overview of the field, especially from an Anglo-Saxon perspective, the reader is referred to the series of review articles in the *Annual Review of Psychology* (especially those from 1981 on), and the first *Handbook of Educational Psychology* (Berliner and Calfee 1996). To get a more comprehensive idea of European research, one can consult the periodicals *European Journal of Psychology of Education* (published since 1986), and *Learning and Instruction. The Journal of the European Association for Research on Learning and Instruction* (published since 1991), and the volumes appearing in the series *Advances in Learning and Instruction*, recently begun to be published by Elsevier Science.

2. Toward a Dispositional View of Competence as Desired End State

During the behaviorist era learning was a prominent issue in psychological research. But with the advent of cognitive science the focus shifted toward the analysis of the structures and processes of human competence in an attempt to answer questions concerning the representation and organization of knowledge, the nature and characteristics of understanding, and the knowledge and cognitive skills involved in competent problem solving. This work, resulting in the identification of critical aspects of expert performance, is relevant from an educational perspective, because it contributes to defining the primary goals of learning in different instructional and training settings. In this respect there is now a broad consensus that becoming skilled in a given domain requires the integrated acquisition of four categories of aptitudes (for a more detailed discussion see, e.g., De Corte 1995):

(a) a well-organized and flexibly accessible domain-specific knowledge base, that constitutes the substance of a subject-matter field;

(b) heuristic methods, i.e., search strategies for problem analysis that increase the probability of finding the right solution, because they induce a systematic approach to the problem;

(c) metacognition which involves knowledge and beliefs about one's own cognitive functioning on the one hand, and skills and strategies for the self-regulation of one's own cognitive processes on the other; and

(d) affective components such as beliefs, attitudes, and emotions relating to a subject-matter field.

In the early days of cognitive science, laboratory investigations using mainly knowledge-lean tasks showed the important role of general cognitive skills in skilled problem solving. But starting in the mid-1970s expert–novice studies in rich content domains demonstrated the significant role of domain-specific knowledge in expert performance. Yet continued research in the 1980s evidenced that domain-specific knowledge could not fully account for human expertise, and showed that metacognitive, self-regulatory skills are a major determinant of successful learning and skilled thinking. Nowadays most scholars endorse the viewpoint that 'we should combine the learning of domain-specific subject matter with the learning of general thinking skills, while also making sure that children learn to monitor and control their thinking and learning' (Bruer 1993, p. 52).

However, the available evidence shows that the integrated acquisition of domain-related knowledge and general cognitive skills is not yet sufficient to overcome the well-known phenomenon of inert knowledge. In other words, expert performance entails more than the sum of the different categories of aptitudes mentioned above; it requires in addition a disposition toward skilled learning, thinking, and problem solving. According to Perkins (1995), a disposition involves, besides ability and motivation, two other crucial components, namely sensitivity and inclination. Sensitivity refers to the feeling for problem situations in which it is relevant and useful to apply acquired knowledge and skills, while inclination is the tendency to do so whenever appropriate. This dispositional conception of expertise accounts for the phenomenon of inert knowledge: students often possess the ability to perform certain tasks or to solve problems, but do not exercise them because of lack of spontaneous inclination and sensitivity.

Since the beginning of the 1990s it has been properly observed that due to the dominance of the information-processing approach, instructional psychology has studied cognition and learning too one-sidedly from a cognitive perspective. Therefore, researchers have begun to pay more attention to building connections between the affective, emotional, and motivational aspects of learning on the one hand, and the cognitive components on the other (see, e.g., Boekaerts 1992). In this respect the dispositional view of skilled learning and thinking offers a useful framework for the integration of both kinds of aptitudes, in the sense that the sensitivity and inclination aspects of the disposition can be conceived of as interfaces or mediators between the cognitive and the affective components.

Being able to transfer knowledge and skills to new learning tasks and problem situations is widely considered as an important goal of education. Therefore, it is not surprising that since the early days of instructional psychology, transfer has repeatedly been, and still is, the main focus of research on learning and instruction. The available literature shows that transfer is a very complex phenomenon, that transfer effects do not occur spontaneously, and that they are even difficult to obtain deliberately (see, e.g., De Corte 1999).

3. Constructive Learning as Lever for the Transition from Initial to Desired End State

Due to the emphasis of cognitive psychological research on the analysis of performance, the study of learning was largely unpopular in the early days of the information-processing approach. But the scene has changed since then, and many researchers working in the cognitive tradition show an active interest in learning processes. Moreover, their work has been enriched by contributions from research on learning in nonschool settings, especially anthropological studies of apprenticeships. From this research a series of characteristics of effective and meaningful learning processes has emerged, which can be summarized in the following definition of learning: it is a constructive, cumulative, self-regulated, goal-directed, situated, collaborative, and individually different process of mean-

ing construction and knowledge building. Only some of these characteristics can be briefly elaborated here (see De Corte 1995 for a more detailed discussion).

3.1 Learning is Constructive

Research has shown that learning is an effortful and mindful process, in which students actively construct knowledge and skills through reorganization of their already acquired mental structures in interaction with the environment. Some scholars—often inspired by the work of Piaget (1955)—take a radical position in this respect, claiming that all knowledge is a subjective and idiosyncratic cognitive construction; others represent a more moderate point of view that allows for the possibility of mediating learning through appropriate guidance. Notwithstanding the large variety in theoretical perspectives along the radical–moderate dimension, the constructivist view certainly implies that acquiring knowledge and skills requires active cognitive processing from the learner.

3.2 Learning is Self-regulated

This feature refers to the metacognitive nature of effective learning, and is implied in the constructive perspective. Indeed, self-regulation means that students manage and monitor their own processes of knowledge building (Schunk and Zimmerman 1998). Skilled self-regulation facilitates appropriate decision making during learning, as well as the monitoring of an ongoing learning process by providing one's own feedback and performance evaluations, and by keeping oneself concentrated and motivated. The more students become self-regulated, the more they assume agency over their learning; consequently they are less dependent on external, instructional support for performing those regulatory activities. An additional argument for promoting students' self-regulation of their learning derives from the finding that high levels of metacognition facilitate transfer: students having good metacognitive awareness and strategies are more able at using what they have learned to approach unfamiliar problems in multiple ways (Brown 1989).

3.3 Learning is Situated and Collaborative

The conception that learning and cognition are situated emerged in the 1980s in reaction to the then prevailing mentalistic, information-processing view of learning and thinking as highly individual and purely cognitive processes occurring in the head, and resulting in the construction of mental representations. In contrast, situativity theory—influenced by the work of Vygotsky (1978)—proposes a contextualized and social conception of learning and thinking: learning is enacted essentially in interaction with the social and cultural context and artifacts, and especially through participation in cultural activities and contexts (Brown et al. 1989). In other words, effective learning is not a purely 'solo' activity, but essentially a distributed one, i.e., the learning effort is distributed over the individual students, their partners in the learning environment, and the resources and tools that are available. Thus, this situated perspective strongly stresses the importance of collaboration in productive learning reflected in such activities as exchanging ideas, comparing solution strategies, and discussing arguments.

4. Powerful Teaching–Learning Environments to Promote Transition to the Desired End State

Powerful teaching–learning environments create the appropriate instructional conditions to evoke in students the learning processes that facilitate transition from their initial state toward the disposition to productive learning, thinking, and problem solving. Therefore, a challenging task for instructional psychology research consists of elaborating and validating a coherent framework of principles for the design of such powerful learning environments. Starting from our current knowledge and understanding of skilled performance in a variety of content domains (as described in Sect. 2) and of the characteristics of effective acquisition processes (see Sect. 3), but also based on observation of the practice of excellent teachers, researchers have begun to address this challenge over the past few years. A promising approach in this respect consists in the creation and evaluation in real classrooms of complex and fundamentally new instructional interventions guided by design principles that embody this present understanding of competence, of effective learning processes, and of successful teaching. Major examples of such orienting principles for the design of powerful learning environments are the following (De Corte 1995):

(a) Learning environments should initiate and support constructive and self-regulated acquisition processes in all students, thus also including the more passive learners. However, from a moderate constructivist perspective, students' construction of knowledge can be mediated through appropriate guidance by teachers, peers, and educational media. In other words, the claim that productive learning requires good teaching still holds true. But, this first principle also implies that systematic interventions should gradually be removed, so that students progressively become agents of their own learning.

(b) Complementary to the first principle, learning environments should allow for the flexible adaptation of instructional support, especially the balance be-

tween self-regulation and external regulation, to take into account individual differences in cognitive as well as in affective and motivational aptitudes. Moreover, the crucial influence on learning of motivational factors points to the necessity of balancing instructional interventions and affective support (Boekaerts 1992).

(c) Taking into account the situated and collaborative nature of effective learning, powerful learning environments should embed students' constructive acquisition processes as much as possible in real-life contexts that have personal meaning for the learners, that offer ample opportunities for distributed learning through social interaction and cooperation, and that are representative of the tasks to which students will have to apply their knowledge and skills in the future. Acquisition of a disposition to skilled learning and thinking, especially the inclination and sensitivity aspects of this disposition, will require extensive experience and practice with the different aptitudes involved, in a large variety of situations.

(d) Because domain-specific knowledge and domain-general heuristic and metacognitive strategies play a complementary role in competent learning and problem solving, learning environments should create opportunities to acquire general cognitive skills embedded in different subject-matter domains.

Cognitive apprenticeship (Collins et al. 1989) is a model for the design of innovative learning environments that is largely in accordance with those principles. The model describes four dimensions that constitute a learning environment: content, teaching methods, sequence of learning tasks, and social context of learning. With respect to content, an ideal learning environment should focus on the acquisition of all categories of knowledge that experts master, namely domain-specific knowledge, heuristic methods, metacognitive skills, and learning strategies. To help students to acquire those different categories of knowledge and skills, the teacher can apply six different teaching methods: modeling, coaching, scaffolding, articulation, reflection, and exploration. The model specifies three principles for sequencing the learning tasks: increasing complexity, increasing diversity, and performing global before local skills. Finally, five guidelines are given for creating a favorable social context for learning: situated learning, opportunities for observation of experts, enhancing intrinsic motivation, fostering cooperative learning, and comparing problem-solving processes and strategies.

Although further research is needed to validate the design principles described above, a series of intervention studies in a variety of content domains that embody those guidelines to some degree have already reported supporting empirical evidence. Often these investigations also integrate educational technology, especially computer programs and, more recently, multimedia systems, in their learning environments. One of the most representative examples of such a research-based attempt at totally changing the classroom environment is the Schools for Thought project in the USA (Lamon et al. 1996).

5. *Conclusion*

This selective, and thus incomplete, review shows that in the 1980s and 1990s research in the field of instructional psychology made substantial progress toward the elaboration of a theory of learning from instruction. Indeed, we now have a fairly good understanding of what skilled performance in a large variety of content domains involves. At the same time the research in which performance of experts and novices was contrasted has contributed to the clarification of one of the components of a theory of learning from instruction that was not explicitly addressed in this article, namely the description of goal-relevant initial states of learners prior to instruction. Considerable advances have also been made in unraveling major characteristics of effective learning processes. And a series of intervention studies has started to address basic questions relating to the design and evaluation of powerful teaching–learning environments to elicit and maintain acquisition processes that are conducive to the attainment of a disposition to competent learning, thinking, and problem solving.

See also: Cognitive Development: Learning and Instruction; Competencies and Key Competencies: Educational Perspective; Education: Phenomena, Concepts, and Theories; Educational Learning Theory; Instructional Design; Learning Theories and Educational Paradigms; Situated Learning: Out of School and in the Classroom

Bibliography

Berliner D C, Calfee R C (eds.) 1996 *Handbook of Educational Psychology*. Macmillan Library Reference, New York

Boekaerts M 1992 The adaptable learning process: Initiating and maintaining behavioral change. *Applied Psychology: An International Journal* **41**: 377–97

Brown A L 1989 Analogical learning and transfer: What develops? In: Vosniadou S, Ortony A (eds.) *Similarity and Analogical Reasoning*. Cambridge University Press, Cambridge, UK, pp. 369–412

Brown J S, Collins A, Duguid P 1989 Situated cognition and the culture of learning. *Educational Researcher* **18**(1): 32–42

Bruer J T 1993 *Schools for Thought: A Science of Learning in the Classroom*. The MIT Press, Cambridge, MA

Collins A, Brown J S, Newman S E 1989 Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. In: Resnick L B (ed.) *Knowing, Learning, and Instruction. Essays in Honor of Robert Glaser*. Lawrence Erlbaum Hillsdale, NJ, pp. 453–94

De Corte E 1995 Fostering cognitive growth: A perspective from research on mathematics learning and instruction. *Educational Psychologist* **30**: 37–46
De Corte E 1999 On the road to transfer: New perspectives on an enduring issue in educational research and practice. *International Journal of Educational Research* **31**: 553–654
Gagné R M, Rohwer Jr W D 1969 Instructional psychology. *Annual Review of Psychology* **20**: 381–418
Glaser R, Bassok M 1989 Learning theory and the study of instruction. *Annual Review of Psychology* **40**: 631–66
Lamon M, Secules T, Petrosino A J, Hackett R, Bransford J D, Goldman S R 1996 Schools for Thought: Overview of the project and lessons learned from one of the sites. In: Schauble L, Glaser R (eds.) *Innovations in Learning: New Environments for Education*. Erlbaum Mahwah, NJ, pp. 243–88
Perkins D 1995 *Outsmarting IQ: The Emerging Science of Learnable Intelligence*. Free Press, New York
Piaget J 1955 *The Child's Construction of Reality*. Routledge, London
Schunk D H, Zimmerman B J (eds.) 1998 *Self-regulated Learning: From Teaching to Self-reflective Practice*. Guilford Press, New York
Snow R E, Swanson J 1992 Instructional psychology: Aptitude, adaptation, and assessment. *Annual Review of Psychology* **43**: 583–626
Vygotsky LS 1978 *Mind in Society. The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

E. De Corte

Instructional Technology: Cognitive Science Perspectives

Intelligent tutoring systems (ITS) are computer-based instructional systems that are founded on the assumption that learning based on individualized instruction by a competent tutor is far superior to classroom-style group instruction. Good human tutors are 'intelligent tutors' because they are able continously to adapt the content and style of their instruction to best meet the needs of the learner. Can we achieve such a functionality using a computer program? This is, in a nutshell, the fundamental goal of ITS.

1. *Traditional Computer-based Instructional Systems*

The first generation of computer tutoring systems were developed in the late 1960s. They were essentially a kind of programmed instructions. They mostly presented pieces of information, e.g., some element of the course content to be learned, asked short-answer questions to check whether the student had learned this element correctly, and branched to further presentations depending on the answer given. The answer provided by the student, simply marked by the program as correct or incorrect, was used to determine the next step of the student's path along the course. These 'traditional' systems are often called computer-aided instruction (CAI) or computer-based instruction (CBI).

This technique already permits first steps towards the individualization of instruction, at least in principle. For example some authors try to anticipate every possible incorrect student response and to prespecify branches depending on the supposed cause of every incorrect answer. The main advantage of these traditional CAI programs is that it allows students to learn or to get trained at their own pace and in their own chosen setting (time and place). Their main drawback was the basically linear progression of the instruction. These systems were unable to explicitly take into account the learners' knowledge, learning style, and preferences, their individual progression, and the underlying cause of their errors.

2. *AI-based Instructional Systems*

The evolution beyond traditional computer-assisted instruction toward ITS has to pass three tests of intelligence. First, the subject matter, or domain, must be 'known' to the system well enough for it to draw inferences or solve problems by itself. Second, the system must be able to diagnose a learner's approximation of the knowledge to be acquired. Third, the tutorial strategy or pedagogical knowledge must be intelligent in the sense that different strategies can be implemented to reduce the difference between the knowledge to be learned and the student's actual knowledge.

The next generation of tutoring systems were based on building knowledge of the subject matter into the program (see Sleeman and Brown 1982 or Wenger 1987 for a detailed documentation and overview). There were different types of these so-called intelligent CAI or ITS. One type coached students working in open-ended learning environments to solve complex problems, such as troubleshooting an electronic circuit, writing a computer program or learning arithmetic skills. Another type attempted to have some kind of a dialog with students, e.g., in the sense of a problem monitor or as a consultant.

The first ITS-type program, Scholar, was developed by Carbonell (1970). Its subject matter was the geography of South America. What was new then was that the system's knowledge of its domain matter was explicitly represented in a semantic network utilizing AI techniques and methods. The nodes of the network stood for geographical objects and concepts. These objects and concepts were organized in a partial hierarchy with relations like superpart, superconcept, and superattribute. Inferences could be made by propagation of inherited properties via these links.

This underlying representation technique gave the system a new quality of competence.

It knew how to solve the tasks given to the student; it no longer needed a data bank of preformulated exercises together with their solutions. Such a system was capable of answering student questions as long as they were domain-related. A further consequence was that the problem-solving process itself could become part of the taught domain. Another new quality of Scholar was its capacity for mixed initiative dialogs based on a simple but robust user interface to correctly interpret the student's questions and responses and to transmit them to the inference engine. Due to the problems of understanding and generation of natural language, only a few systems of this type were built.

2.1 The Basic Architecture of ITS

ITS developed since 1970 usually contain four basic components or modules (Sleeman and Brown 1982, Wenger 1987).

(a) An environment (user interface) in which the student works on complex tasks. For instance, it might be a simulated piece of an electronic circuit that the student tries to troubleshoot.

(b) An expert system (domain knowledge base) that knows enough about the problems that the student works on to solve them by itself.

(c) A student modeling component (student model) that compares the student's behavior with the expert system's behavior in order both to recognize the student's current knowledge for solving the problem and determine what elements of correct as well as incorrect knowledge the student is probably using.

(d) A tutorial module (teaching module) that suggests tasks to be solved, responds to the student's requests for help, and points out mistakes. Such responses and suggestions are based on a library of flexible usable teaching strategies as well as on the system's model of the student's knowledge and plans.

In general, these systems try to simulate good human tutors by the integrated use of different knowledge bases. They are domain experts in the subject matter to be learned by the student, they are teaching experts using different tutorial strategies and interventions for supporting the student's acquisition of knowledge, and they are educational/psychological experts in the assessment and evaluation of the student's domain-related knowledge. Sometimes these ITS are therefore also called knowledge-based tutors (e.g., Murray 1998).

2.2 Student Modeling

The idea of the necessity of some kind of student modeling was around from the beginning. The argumentation is well known. Efficient teaching implies that the communication of the knowledge to be learned is adapted to the cognitive prerequisites and requirements of the learners, especially to their domain knowledge. The representation of domain knowledge implemented in the ITS should reflect the student's mental representation of this knowledge. In its simplest form, there may be only a list of what the student knows or should know. There may also be some techniques to model the learner's errors based on incorrect and/or incomplete knowledge. Such techniques include systematic errors, often modeled as bugs, impasses, misgeneralizations or repairs in procedural skills (Brown and Burton 1978, Young and O'Shea 1981, Sleeman 1984, VanLehn 1990), or as some kind of overlay model (Goldstein 1982). In its most complete state, the student model also contains domain-independent knowledge. In that case, the system also has to know something about knowledge acquisition based on learning processes, therefore referring to different types of students rather than to different domains. Examples of these kinds of student modeling can be found in the work of Anderson and his group (e.g., Anderson et al. 1989, Anderson and Lebiere 1998). With such information, the system is then capable of associating specific tutoring strategies to specific learning strategies.

Thanks to the complex integration of the various system modules, such an ITS is able to guide the learner as evolving to become an expert, at least in principle. Through the presentation of specific questions, tasks, or explanation requests, certain comprehension errors can be detected. Moreover, thanks to the knowledge it might thus have (or even gain) about students and their specific knowledge, problem-solving strategies, and learning mechanisms, it would be able to support a student's learning and knowledge acquisition in quite an individualized manner.

2.3 Applications

In the last half of the 1980s many ITS were developed, despite the immense costs and the complexity involved. But they were not only in use for academic purposes. ITS have moved out of the laboratory and into classrooms and workplaces, where some have proven to be highly effective as learning aides. For example, students working with an Air Force electronics trouble-shooting tutor for only 20 hours gained proficiency equivalent to that of trainees with 48 months of on-the-job experience (Lesgold et al. 1992). In another example, students using the LISP tutor (Anderson 1990) completed programming exercises in 30 percent less time than those receiving traditional classroom instruction and scored 43 percent higher on the final exam.

Although ITS were becoming more common and proving to be increasingly effective, each one had to be built from scratch at a great expense. There was also

no standardization with regard to the software and hardware used. Authoring tools for ITS have not been available until recently (Murray 1998). Although ITS show their potential effectiveness for educational purposes, there are other applications as well (cf. VanLehn 1999). A common practice is to build an open learning environment without the underlying expert system, without any student model, and without any tutorial module. The environment enables student activities that stimulate learning and may be impossible to conduct in the real world. For instance, an environment might allow students to conduct simulated physics experiments on worlds where gravity is reduced, absent, or even negative. Such environments are called interactive learning environments, microworlds, or virtual laboratories. Another new trend is to use networking to allow several students to work together in the same environment.

2.4 *ITS from a Cognitive Science Point of View*

One of the lessons learned in the development of an ITS is that it requires the knowledge and collaboration of different disciplines. But empirical cognitive studies are nevertheless particularly important in developing ITS. Developing the expert module of a tutoring system requires studying experts as they solve problems in order to understand and formalize their knowledge. Developing an effective teaching module requires understanding how students learn so that the tutor's comments will prompt students to construct their own understanding of the subject matter. An overly critical or didactic tutor may do more harm than good. A good first step in developing an application is to study the behavior of expert human tutors in order to see how they increase the motivation and learning of students. Developing the student model one needs to understand how students acquire new knowledge and how they apply their knowlege to solve problems. A good way to achieve this is a detailed analysis of the subject matter, its main concepts, and their relational structure and of important procedural skills to solve problems and tasks within the domain. An empirical study of students' errors and most typical misconceptions might also be helpful.

2.5 *Conclusions*

Many ITS have been developed since the mid-1970s and the requirements to be solved during the development of an ITS generate many fruitful interactions between academic research and application-oriented research with regard to topics like knowledge representation, computerized assessment of knowledge, learning and knowledge acquisition, expertise research. Nevertheless from a cognitive science perspective some difficult problems have not been sufficiently solved until now: First, the requirement of a quick and reliable in-time and on-line diagnosis of knowledge. Even within very complex and large behavioral spaces the system's feedback must be given immediately. Otherwise the user will not accept the system for a long time. The dynamic assessment of valid indicators and their interpretation with respect to the student's knowledge acquisition processes, goals, and problem-solving strategies still includes many open questions. Second, the systematic consideration of motivational and emotional aspects is an important question of research. Third the transitions between the content of the student model and the tutorial knowledge (system feedback, kind of help, strategic-didactical decisions, content and style of information presentation) needs more controlled empirical evidence. ITS are still an active topic of many research activities, not only because of the many problems to be solved but because of the ongoing growing requirement of computerized training and education.

3. *Current Trends*

The present research tends to be more and more interdisciplinary, slowly but surely making more effective the collaboration between course designers, cognitive scientists, computer scientists, and educators/psychologists/teachers. With regard to the uses of new technologies of information and communication in education, the main interest lies in exploiting Internet possibilities, especially the Web. Some important trends are outlined below: distance learning and teletutoring, collaborative learning, and Web-based learning. Finally, an active area surpasses them all: the evaluation of educational systems.

3.1 *Distance Learning and Teletutoring*

Distance learning and teletutoring are extremely active areas of research. Many courses are now being developed on the Web. Distance learning means that the learning material is offered as computer-based training (CBT) or as Web-based training (WBT) that can be used by students in their own learning environment at their place of work or at home. The access and interaction is completely self-organized. Teletutoring means that there is a teacher/tutor who primarily takes the role of a moderator. This teacher supports ongoing learning activities in individuals but also in groups. Communication is the most important concept. One well-known variant of teletutoring is the online seminar in which a classical teaching form is virtually imitated. Some researchers are also investigating ways to use the Web meaningfully and

effectively to develop new types of pedagogical paradigms (e.g., see the recent conferences on AI in education).

3.2 Collaborative Learning

In cognitive science, a whole area of research is striving to enhance computer supported collaborative/cooperative learning (CSCL). The basic goal is to develop learning environments in which groups of students can communicate with each other, interact with each other, and learn and solve problems together. The cooperation might take place at the same time or at different times (synchronous/asynchronous CSCL) and at the same place or at different places (local or distributed CSCL). Probably this field will play a much more prominent role in the future. It integrates the fast growing potential of new information technology (e.g., multimedia, interactivity), new learning theories (situated learning, learning as social activity, distributed learning) and the perspective of knowledge acquisition based on explanation, argumentation, and discourse. A good overview of ongoing research in this field can be found in the various contributions in Koschmann (1996) and Dillenbourg (1999).

3.3 Web-based Learning

Since the Web is the major result of the present technological progress, it can be expected that it will have a steadily growing influence on computerized training and education (e.g., Eisenstadt and Vincent 1998, Forsyth 1998, Bruns and Gajewski 2000). However, since it is only a technological evolution, it will probably act more as a catalyst than as a conceptual or theoretical leading engine. At any rate, this domain will undergo two main positive changes. The first change is the expected gain in the pedagogical capabilities of educational systems. However, this gain will be effective only through a better intercommunication and a better collaboration between educational and psychological researchers, cognitive scientists, and computer scientists. But actually to benefit from this evolution, it will be necessary to overcome several traditional but strong barriers, essentially of a cultural or administrative nature.

The second major change stems from the fast growing communication and multimedia capabilities of information technology. Owing to such capabilities, students or trainees will undoubtedly become more and more autonomous, and will feel more and more responsible for their own learning or training. The responsibility of each subject for his or her own education and training will gain increasing importance. To adapt to this evolution, teachers of all types, including university professors, will have to modify their roles by becoming rather knowledgeable guides than knowledge providers, and by becoming more student-centered. Students will be more and more engaged in their own goals, searching the Web for the information they are interested in. The drawback of these changes is that all participants in this evolution—researchers and developers, but also educators and school administrators—will have to adapt to that evolution adequately, whether they like it or not. Information technology will evolve anyway, and we should actively use the new possibilities of this evolution.

3.4 Evaluation of Educational Systems

Finally, with the variety and number of educational programs and prototypes developed, and with the variety of theories, methods, and techniques put to use in the field of computerized education, the need to better understand the pros and cons of that field gets increasingly stronger. Therefore the task of evaluating educational systems, and the methods and processes used to evaluate them, will receive growing interest.

3.5 Concluding Remarks

At the turn of the twenty-first century there are no up-to-date textbooks on intelligent tutoring. Sleeman and Brown (1982), Polson and Richardson (1988), Self (1988), and Wenger (1987) cover the basic ideas as well as the early systems. Lelouche (1999) gives a survey of the history of ITS and also a general outlook on the present trends of research. Self (1999) reviews the evolution of the ITS research, namely to make computationally precise and explicit the forms of educational, psychological, and social knowledge necessary for the design and implementation of computer-based learning systems. The latest work generally appears first in the proceedings of the conference on Intelligent Tutoring Systems or in the proceedings of the conference on AI in Education. The most relevant journals for this kind of work include the *International Journal of AI in Education* (http://cbl.leeds.ac.uk/ijaied/), the *Journal of the Learning Sciences* (Erlbaum), and *Interactive Learning Environments* (Ablex). There are also many Web-based educational systems available covering quite different subject matters. To get an impression of what is going on in the field, it might be helpful simply to use some of these systems.

See also: Computer Networking for Education; Computer-assisted Instruction; Computers and Society; Digital Computer: Impact on the Social Sciences; Information Technology; Intelligent Tutoring Systems; Science and Technology, Social Study of: Computers and Information Technology

Bibliography

Anderson J R 1990 Analysis of student performance with the LISP tutor. In: Frederiksen N, Glaser R, Lesgold A M, Shafto M (eds.) *Diagnostic Monitoring of Skill and Knowledge Acquisition*. Erlbaum, Hillsdale, NJ, pp. 27–50

Anderson J R, Conrad F G, Corbett A T 1989 Skill acquisition and the LISP tutor. *Cognitive Science* **13**: 467–505

Anderson J R, Lebiere C 1998 *The Atomic Components of Thought*. Erlbaum, Hillsdale, NJ

Brown J S, Burton R R 1978 Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science* **2**: 155–91

Bruns B, Gajewski P 2000 *Multimediales Lernen im Netz*. Springer, Berlin

Carbonell J R 1970 AI in CAI: An artificial intelligence approach to computer assisted instruction. *IEEE Transactions on Man–Machine Systems* **11**(Dec): 190–202

Dillenbourg P 1999 *Collaborative Learning: Cognitive and Computational Approaches*. Pergamon, Amsterdam

Eisenstadt M, Vincent T 1998 *The Knowledge Web: Learning and Collaborating on the Net*. Kogan Page, London

Forsyth I 1998 *Teaching and Learning Materials and the Internet*. Kogan Page, London

Goldstein I P 1982 The genetic graph: A representation for the evolution of procedural knowledge. In: Sleeman D, Brown J S (eds.) *Intelligent Tutoring Systems*. Academic Press, New York, pp. 51–77

Koschmann T 1996 *CSCL: Theory and Practise of an Emerging Paradigm*. Erlbaum, Hillsdale, NJ

Lelouche R 1999 Intelligent tutoring systems from birth to now. *KI - Künstliche Intelligenz* **4/99**: 5–11

Lesgold A, Lajoie S, Bunzo M, Eggan G 1992 Sherlock: A coached practise environment for an electronics trouble-shooting job. In: Larkin J, Chabay R W (eds.) *Computer Assisted Instruction and Intelligent Tutoring Systems*. Erlbaum, Hillsdale, NJ, pp. 20–38

Mandl H, Lesgold A 1988 *Learning Issues for Intelligent Tutoring Systems*. Springer, New York

Murray T 1998 Authoring knowledge-based tutors: Tools for content, instructional strategy, student model, and interface design. *Journal of the Learning Sciences* **7**: 5–64

Polson M C, Richardson J J 1988 *Foundations of Intelligent Tutoring Systems*. Erlbaum, Hillsdale, NJ

Self J 1988 *Artificial Intelligence and Human Learning: Intelligent Computer-aided Instruction*. Chapman and Hall, London

Self J 1999 The defining characteristics of intelligent tutoring systems research: ITSs care, precisely. *International Journal of AI in Education* **10**: 350–64

Sleeman D 1984 An attempt to understand student's understanding of basic algebra. *Cognitive Science* **8**: 387–412

Sleeman D, Brown J S 1982 *Intelligent Tutoring Systems*. Academic Press, New York

VanLehn K 1990 *Mind Bugs: The Origins of Procedural Misconceptions*. MIT Press, Cambridge, MA

VanLehn K 1999 AI and education. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 7–10

Wenger E 1987 *Artificial Intelligence and Tutoring Systems*. Morgan Kaufmann, San Mateo, CA

Young R M, O'Shea T 1981 Errors in children's subtraction. *Cognitive Science* **5**: 153–77

K. Opwis

Instrumental Variables in Statistics and Econometrics

The method of instrumental variables (IVs) is a general approach to the estimation of causal relations using observational data. This method can be used when standard regression estimates of the relation of interest are biased because of reverse causality, selection bias, measurement error, or the presence of unmeasured confounding effects. The central idea is to use a third, 'instrumental' variable to extract variation in the (IV) variable of interest that is unrelated to these problems, and to use this variation to estimate its causal effect on an outcome measure. This article describes IV estimators, discusses the conditions for a valid instrument, and describes some common pitfalls in the application of IV estimators.

1. The Method of Instrumental Variables

1.1 Common Problems with Standard Regression Analysis of Observational Data

In many cases in the social and behavioral sciences, one is interested in a reliable estimate of the causal effect of one variable on another. For example, suppose a mayor is considering increasing the size of the police force; what is the effect of an additional police officer on the crime rate? Or, what is the effect of an additional year of schooling on future earnings? What will happen to economic growth if the central bank raises short-term interest rates by one percentage point? What is the effect of a new medical procedure on health outcomes? These and many other questions require estimates that are causal, in the sense that they are externally valid and can be used to predict the effect of changes in policies or treatments, holding other things constant.

In theory, such causal effects could be estimated by a suitably designed randomized controlled experiment. Very often, however, as in the first three questions, such an experiment could be prohibitively expensive, could be unethical, and/or could have questionable external validity. Even when randomized controlled experiments are available, such as clinical trials of medical procedures, it is of interest to validate the experimental predictions using information on outcomes in the field. Thus, to address such questions empirically typically entails the use of nonexperimental, i.e., observational, data.

Unfortunately, standard regression analysis of observational data can fail to yield reliable estimates of causal effects for many reasons, four of which are particularly salient. First, there could be additional unmeasured effects, leading to 'omitted variable bias'; for example, the educational attainment of parents is

correlated with that of their children, so if parents' education facilitates learning at home but is unobserved then the correlation between years of school and earnings could overstate the true, causal effect of school on earnings. Second, there might be reverse causality, or 'simultaneous equations' bias; for example, more police officers might reduce crime, but cities with higher crime rates might demand more police officers, so standard regression analysis of crime rates on the number of police confounds these two different effects. Third, there could be selection bias, in which those most likely to benefit from a treatment are also most likely to receive it; for example, because ambition is one reason for success both at school and in the labor market, the measured correlation between years of school and earnings could simply reflect the effect of unmeasured ambition. Fourth, standard regression estimates of the causal effect are biased if the regressor is measured with error (see *Measurement Error Models*).

1.2 IV Regression

IV regression provides a way to handle these problems. The main early application of IV methods was for estimating the parameters of a system of linear simultaneous equations, and this remains a good expositional framework. Let y_i denote the outcome variable of interest (say, log future earnings), let Y_i denote the r treatment variables (years of education), let X_i denote K_1 additional observed control variables, and let Z_i denote the K_2 instrumental variables, where these are all observed for observations $i = 1,\ldots, N$. Also let β and γ be unknown parameter vectors, let Φ and Π be matrices of unknown parameters, and let β' denote the transpose of β. Suppose that these variables are linearly related as

$$y_i = \beta' Y_i + \gamma' X_i + u_i \tag{1}$$

$$Y_i = \Phi X_i + \Pi Z_i + V_i \tag{2}$$

where u_i and V_i are 'error terms' that represent additional unobserved effects, measurement error, etc. The coefficient of interest is β; in the schooling example, this is the percentage change in future earnings caused by attending school for an additional year.

In this notation, the four problems listed in the preceding paragraph have a common implication, that the correlation between Y_i and u_i is nonzero, and in consequence the ordinary least squares (OLS) estimator of β will be biased and, in large samples, inconsistent. It is assumed that corr $(X_i, u_i) = 0$; at this level, this assumption can be made without loss of generality, for if this is suspected to be false for some element of X, then it should be listed instead in Y. In the terminology of simultaneous equations theory, y and Y are endogenous variables and X is exogenous.

The key idea of IV methods is that although Y is correlated with u, if Z is uncorrelated with u (that is, if Z is exogenous), then Z can be used to estimate β. Intuitively, part of the variation in Y is endogenous and part is exogenous; IV methods use Z to isolate exogenous variation in Y and thereby to estimate β. More formally, consider the simplest case, in which there is no X in (1) or (2) and which Y and Z are single variables (i.e., $K_1 = 0$ and $r = K_2 = 1$). Then $\text{cov}(y, Z) = \beta\ \text{cov}(Y, Z) + \text{cov}(u, Z) = \beta\ \text{cov}(Y, Z)$. Thus, $\beta = \text{cov}(y, Z)/\text{cov}(Y, Z)$. This leads to the IV estimator, $\hat{\beta}^{\text{IV}} = s_{yZ}/s_{YZ}$, where s_{yZ} is the sample covariance between y and Z. Evidently, if these two sample covariances are consistent and if $\text{cov}(Y, Z) \neq 0$, then $\hat{\beta}^{\text{IV}} \xrightarrow{\text{p}} \beta$. The availability of the instrument Z thus permits consistent estimation of β.

If more than r instruments are available, it makes sense to use the extra instruments to improve precision. There are, however, many ways to do this, since each subset of r instruments will produce its own estimate of β. The most common way to combine instruments is to use two stage least squares (2SLS). In the first stage, Eqn. (2) is estimated by OLS, producing the predicted values $\hat{Y}$. In the second stage, y is regressed against $\hat{Y}$ and X, yielding the 2SLS estimator $\hat{\beta}^{\text{2SLS}}$. With only one instrument, this reduces to $\hat{\beta}^{\text{IV}}$ given above. Under Gaussian disturbances, this provides an asymptotically efficient method for weighting the various instruments that is easy to understand and to implement.

1.3 Conditions for Valid Instruments

The question of whether a candidate set of instruments can be used to estimate β is a special case of the more general problem of identification of parameters in multivariate econometric and statistical models. In the context of (1) and (2), when there is a single Y, the requirements for instrument validity are simple. Then a potential instrument Z must satisfy two conditions: (a) Z is uncorrelated with u; and (b) the partial correlation between Y and Z, given X, is nonzero. The first condition states that the instrument is exogenous. The second condition is that the instrument is relevant. Together, these conditions permit using Z to isolate the exogenous variation in Y and thus to identify β. For example, when $r = K_2 = 1$ and there are no Xs, if Z satisfies the two conditions then β is identified by the moment condition derived above, $\beta = \text{cov}(y, Z)/\text{cov}(Y, Z)$.

When Y is a vector (i.e., $r > 1$), the exogeneity condition (a) remains but the relevance condition (b) is somewhat more complicated. For identification of the independent effects of each Y (all elements of β), a sufficient condition for instrument validity is that the covariance matrix of Z has full rank and that Π has

full row rank. Clearly, a necessary condition for Π to have full row rank is that there are at least as many instruments as Ys ($K_2 \geqslant r$). The number of excess instruments, $K_2 - r$, is referred to as the degree of overidentification of β.

When y is a vector, so that (1) is itself a system of multiple equations, the conditions for identification become intricate (see *Simultaneous Equation Estimation: Overview*; *Statistical Identification and Estimability*). For a rigorous and complete treatment in the linear simultaneous equations framework, see Rothenberg (1971) and Hsiao (1983).

2. IV Estimators and their Distributions

2.1 Linear Models

In addition to two stage least squares, other estimators are available for IV estimation of a single equation that is linear in the unknown parameters. The leading alternative estimator is the so-called limited information maximum likelihood (LIML) estimator. The LIML estimator minimizes the Anderson–Rubin (1949) statistic,

$$F(\beta) = \frac{\tilde{\boldsymbol{y}}'(\boldsymbol{M}_X - \boldsymbol{M}_{[X\,Z]})\tilde{\boldsymbol{y}}/K_2}{\tilde{\boldsymbol{y}}'\boldsymbol{M}_{[X\,Z]}\tilde{\boldsymbol{y}}/(N-K_1-K_2)} \tag{3}$$

$\tilde{\boldsymbol{y}} = (\tilde{y}_1 \ldots, \tilde{y}_N)'$, where $\tilde{y}_i = y_i - \beta'\mathrm{Y}_i$, and $M_X = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)'$. This can be solved analytically as an eigenvalue problem. Many other single-equation IV estimators have been proposed, but these are rarely used in empirical applications; see Hausman (1983) for expressions for the LIML estimator and for a discussion of other estimators.

Under standard assumptions (fixed numbers of regressors and instruments, validity of the instruments, convergence of sample moments to population counterparts, and the ability to apply the central limit theorem), the LIML and 2SLS estimators are asymptotically equivalent and have the same asymptotic normal distribution. That is, $\sqrt{\mathrm{N}}(\hat{\beta}^{\mathrm{LIML}} - \beta)$ and $\sqrt{\mathrm{N}}(\hat{\beta}^{\mathrm{2SLS}} - \beta)$ both have asymptotic normal distributions with the same covariance matrix. However, their finite sample distributions differ. There is a large literature on finite sample distributions of these estimators (see *Simultaneous Equation Estimates (Exact and Approximate), Distribution of*). Because these exact distributions depend on nuisance parameters, such as Π, that are unknown and because few IV statistics are (exactly) pivotal, this literature generally does not provide approximations that are useful in empirical applications. However, some general guidelines do emerge from these studies. Perhaps the most important is that when there are many instruments and/or when Π is small in a suitable sense, LIML tends to exhibit less bias than 2SLS and LIML confidence intervals typically have better coverage rates than 2SLS.

These methods extend naturally to the case that y is a vector and (1) is a system of equations. This entails imposing restrictions on the coefficients in (1) and (2) that are implied by the model, and then using the available instruments to estimate simultaneously all these coefficients with these restrictions imposed. The system analog of 2SLS is three stage least squares, in which all the unknown coefficients are estimated simultaneously using an estimate of the variance-covariance matrix of u obtained from the 2SLS estimates. The system analog of LIML is full information maximum likelihood. For further discussion, see *Simultaneous Equation Estimation: Overview*; *Simultaneous Equation Estimates (Exact and Approximate), Distribution of*.

2.2 Nonlinear Models

Many details of the foregoing discussion hinged on the system being linear in the parameters and variables. In many settings, however, theory suggests models that are nonlinear rather than linear. A leading class of examples is consumption-based models of asset prices, in which the equations contain nonlinearities inherited from posited utility functions of representative agents. In such settings, Eqn. (1) is usefully extended and recast to incorporate these nonlinearities. Specifically, suppose that the model at hand implies,

$$E[h(y_i, Y_i, X_i, \theta) \mid X_i, Z_i] = 0 \tag{4}$$

where as before y_i and Y_i are endogenous variables, X_i and Z_i are exogenous variables, and θ is a vector of parameters. For example, in the linear system (1) and (2), (4) is implied by $\mathrm{E}(u_i \mid X_i, Z_i) = 0$ upon setting $h(y_i, Y_i, X_i, \theta_0) = y_i - \beta' Y_i - \gamma' X_i$ and $\theta = (\beta, \gamma)$.

The orthogonality condition (4) corresponds to the first condition for a valid instrument, that Z is exogenous. If in addition Z is relevant, then (4) can be exploited to estimate the unknown parameters θ. The issue of relevance of Z is equivalent to whether θ is identified. The question of identification in nonlinear models is complex and little can be said about global identification at a general level, although conditions for local identification sometimes yield useful insights (see *Statistical Identification and Estimability*).

Given a model (4) and a set of instruments, in general a variety of estimators of θ are available. Estimation is typically undertaken by minimization of a quadratic form in $h(y_i, Y_i, X_i, \theta)$, times the instruments; 2SLS and LIML estimates obtain as special cases when h is linear. In practice, the choice of the weighting matrix in this quadratic form is im-

portant, and if there is serial correlation and/or heteroskedasticity this weighting matrix (and the standard errors) must be computed in a way that accounts for these complications. Such estimators are referred to as generalized method of moments (GMM) estimators. For discussions of identification and estimation in GMM, see Newey and McFadden (1994).

3. A Brief History of IV Methods

S. Wright (1925) and his father P. Wright (1928) introduced the estimator $\hat{\beta}^{\mathrm{IV}}$ (defined in Sect. 1.2) and used it to estimate supply and demand elasticities for butter and flaxseed; this work, however, was neglected (see Goldberger 1972). The development of IV methods in the 1940s stemmed from the attempts of early statisticians and econometricians to solve what appeared to be two different problems: the problem of measurement error in all the variables, and the problem of reverse causality arising in a system of simultaneous equations that describe the macroeconomy. The term 'instrumental variables' was first used in print by Reiersøl (1945), who developed IV estimation for the errors in variables problem and extended work by Wald (1940) which, in retrospect, also can be seen as having introduced IV estimation. Concurrently, work by Haavelmo (1944) and the Cowles Commission emphasized likelihood-based methods of analyzing simultaneous equation systems, which implicitly (through identification conditions) entailed the use of instrumental variables; Anderson and Rubin (1949) developed LIML as part of this research program.

An important breakthrough came with the development of 2SLS by Basmann (1957) and Theil (1958), which permitted computationally efficient estimation in single equations with multiple instruments. Sargan (1958) introduced IV estimation for multiple equation systems and Zellner and Theil (1962) developed three stage least squares.

The extension of these methods to nonlinear models was undertaken by Amemiya (1974) (nonlinear 2 SLS) and by Jorgenson and Laffont (1974) (nonlinear three stage least squares). The modern formulation of GMM estimation is due to Hansen (1982). GMM constitutes the dominant modern unifying framework for studying the issues of identification, estimation, and inference using IVs.

4. Potential Pitfalls

In practice, inference using IV estimates can be compromised because of failures of the model and/or failures of the asymptotic distribution theory to provide reliable approximations to the finite sample distribution. Much modern work on IV estimation entails finding ways to avoid, or at least to recognize, these pitfalls. This section discusses three specific potentially important pitfalls in linear IV estimation. These problems, as well as additional problems associated with estimation of standard errors, also arise in nonlinear GMM estimation, but their sources and solutions are less well understood in the general nonlinear setting.

4.1 Some Instruments Are Endogenous

If an instrument is endogenous then $\mathrm{corr}(Z_i, u_i) \neq 0$, and reflection upon the derivation $\hat{\beta}^{\mathrm{IV}}$ and its probability limit in Sect. 1.2 reveals that $\hat{\beta}^{\mathrm{IV}}$ is no longer consistent for β. More generally, if at least one instrument is endogenous then the 2SLS and LIML estimators are inconsistent. If at least r instruments are exogenous, however, then it is possible to test the null hypothesis that all instruments are exogenous against the alternative that at least one (but no more than $K_2 - r$) is endogenous. In practice, a researcher might have at least r instruments that he or she firmly believes to be exogenous, but might have some additional instruments which are more questionable.

Testing the null that all the instruments are exogenous entails checking empirically the assumption that the instruments are uncorrelated with u_i. This can be done by regressing the IV residual (estimated by 2SLS or LIML) against Z. The R^2 of this regression should be zero; if NR^2 exceeds the desired $\chi^2_{K_2-r}$ critical value, the null hypothesis of joint exogeneity is rejected.

Although this statistic provides a useful diagnostic and rejection suggests that the full instrument list is (jointly) invalid, failure to reject is not necessarily reassuring since the maintained hypothesis is that at least r of the instruments are valid. Moreover, this test can only be implemented if β is overidentified ($K_2 > r$). The hypothesis that at least r instruments are exogenous is both essential and untestable, and thus must be contemplated carefully in IV applications.

4.2 Weak Instruments

The condition for instrument relevance states that Z and Y must have nonzero partial correlation (given X). In practice, this correlation, while arguably nonzero, is often small, a situation sometimes referred to as the problem of weak instruments. When the instruments are weak, the usual large sample approximations provide a misleading basis for inference: the 2SLS estimator in particular is biased towards the OLS estimator, and asymptotic LIML and 2SLS confidence regions have coverage rates that can differ substantially from the nominal asymptotic confidence level. When $r = 1$, an empirical measure of the strength

of the instruments is the F-statistic testing the hypothesis $\Pi = 0$ in the first stage regression (2). The risk of weak instruments is especially relevant when there are many instruments, for even if some instruments have a large partial correlation with Y, if there are many instruments quite a few of them could be weak so taken together this first-stage F-statistic could be small.

Although no single preferred way to handle weak instruments has yet emerged, one solution is to construct confidence intervals by inverting the Anderson–Rubin statistic (3). This is readily done because $F(\beta_0)$ has an asymptotic $\chi^2_{K_2}$ distribution under the joint null hypothesis that $\beta = \beta_0$ (a fixed vector) and $\text{corr}(X, u) = \text{corr}(Z, u) = 0$ (the instrument relevance condition is not needed for this result).

4.3 Heterogeneity of Treatment Effects in Micro Data

Many modern applications of IV are to microdata sets with cross-sectional or panel data, that is, data sets where the observation units are individuals, firms, etc. An important consideration that arises in this context is the role of heterogeneity in interpreting what it is that IV methods measure. To make this concrete, consider a modification of (1) and (2), where for simplicity it is assumed that $r = K_2 = 1$ and the only X is a constant. Suppose, however, that there is heterogeneity both in the responses of each individual to the 'treatment' Y and in the influence of the instrument on the level of treatment received. In equations, this can be written as,

$$y_i = \gamma + \beta_i Y_i + u_i \tag{5}$$

$$Y_i = \Phi + \Pi_i Z_i + V_i \tag{6}$$

where β_i and Π_i vary randomly across individuals. Suppose that the instrument is distributed independently of $(u_i, V_i, \beta_i, \Pi_i)$, and technical conditions ensuring convergence of sample moments hold. Then $\hat{\beta}^{2\text{SLS}} \xrightarrow{p} E[\Pi_i \beta_i]/E\Pi_i$.

Evidently, with heterogeneity of this form 2SLS can be thought of as estimating $E[\Pi_i \beta_i]/E\Pi_i$, which differs from the usual estimand $\beta = E\beta_i$ if β_i and Π_i are correlated. For example, suppose that $\Pi_i = 0$ for half the population, that $\Pi_i = \Pi$ (which is fixed) for the other half, and that the sample is drawn randomly. If β_i differs systematically across the two halves of the population, then 2SLS estimates the mean of β_i among that half for which $\Pi_i = \Pi$; this differs from the mean of β_i over the full population. More generally, 2SLS can be thought of as estimating a weighted average of the individual treatment effects, where the weights reflect the influence of Z_i on whether the individual receives the treatment. This has been referred to as the 'local average treatment effect'; see Angrist et al. (1996).

5. Where Do Valid Instruments Come From?

In practice the most difficult aspect of IV estimation is finding instruments that are both exogenous and relevant. There are two main approaches, which reflect two different perspectives on econometric and statistical modeling. The first approach is to use *a priori* theoretical reasoning to suggest instruments. This is most compelling if the model being estimated is itself derived from a formal theory. An example of this approach is the estimation of intertemporal consumption-based asset pricing models, mentioned in Sect. 2.2 in which previously observed variables are, under the model, uncorrelated with certain future expectational errors. Thus previously observed variables can be used as instruments.

The second approach to constructing instruments, more commonly found in program evaluation studies (broadly defined), is to look for some exogenous source of variation in Y that either derives from true randomization or, in effect, from pseudo-randomization. In randomized experiments in the social sciences, compliance with experimental protocol is usually unenforceable, so a subject's decision whether to take the treatment introduces selection bias; assignment to treatment, however, can be used as an instrument for receipt of treatment. In nonexperimental settings, this reasoning suggests looking for a variable that plays a role similar to random assignment in a randomized experiment. For example, McClellan, McNeil and Newhouse (1994) investigated the effect of the intensity of treatment (Y) on mortality after four years (y) using observational data on elderly Americans who suffered an acute myocardial infarction (heart attack). Standard regression analysis of these data would be subject to selection bias and omitted variables bias, the former because the decision to pursue an intensive treatment depends in part on the severity of the case, the latter because of additional unobserved health characteristics. To avoid these biases, they used as instruments the distance of the patient from hospitals with various degrees of experience treating heart attack patients, for example, the differential distance to a hospital with experience at cardiac catheterization. If distance to such hospitals is distributed randomly across potential heart attack patients, then it is exogenous; if this distance is a factor in the decision whether to move the patient to the distant hospital for intensive treatment, then it is relevant. If both are plausibly true, then 2SLS provides an estimate of the average treatment effect for the marginal patients, where the marginal patients are those for whom the effect of distance on the decision to treat is most important.

From the humble start of estimating how much less butter people will buy if its price rises, IV methods

have evolved into a general approach for estimating causal relations throughout the social and behavioral sciences. Because it requires valid and relevant instruments, IV regression is not always an option, and even if it is, the practitioner must be aware of its potential pitfalls. Still, when they can be applied, IV methods constitute perhaps our most powerful weapon against omitted variable bias, reverse causality, selection bias, and errors-in-variables in our efforts to estimate causual relations using observational data. For recent textbook treatments of instrumental variables issues see Hayashi (2000, Chaps. 3 and 4) and Ruud (2000, Chaps. 20–2 and 26).

See also: Causal Inference and Statistical Fallacies; Selection Bias, Statistics of

Bibliography

Amemiya T 1974 The nonlinear two-stage least squares estimator. *Journal of Econometrics* **2**: 105–10

Anderson T W, Rubin H 1949 Estimation of the parameters of a single equation in a complete systems of stochastic equations. *Annals of Mathematical Statistics* **20**: 46–63

Angrist J D, Imbens G W, Rubin D B 1996 Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* **91**: 444–72

Basmann, R L 1957 A generalized classical method of linear estimation of coefficients in a structural system of stochastic equations. *Annals of Mathematical Statistics* **20**: 46–63

Goldberger A S 1972 Structural equation methods in the social sciences. *Econometrica* **40**: 979–1001

Haavelmo T 1944 The probability approach to econometrics. *Econometrica* **12** (Suppl): 1–118

Hansen L P 1982 Large sample properties of generalized method of moments estimators. *Econometrica* **50**: 1029–54

Hausman J A 1983 Specification and estimation of simultaneous equation models. In: Grilliches Z, Intrilligator M D (eds.) *Handbook of Econometrics*. North Holland, Amsterdam, Vol. 1, pp. 391–450

Hayashi F 2000 *Econometrics*. Princeton University Press, Princeton, NJ

Hsiao C 1983 Identification. In: Grilliches Z, Intrilligator M D (eds.) *Handbook of Econometrics*. North Holland, Amsterdam, Vol. 1, pp. 223–83

Jorgenson D W, Laffont J 1974 Efficient estimation of nonlinear simultaneous equations with additive disturbances. *Annals of Economic and Social Measurement* **3**: 615–40

McClellan M, McNeil B J, Newhouse J P 1994 Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *Journal of the American Medical Association* **272**: 859–66

Newey W K, McFadden D 1994 Large sample estimation and hypothesis testing. In: Engle R F, McFadden D (eds.) *Handbook of Econometrics*. North Holland, Amsterdam, Vol. 4, pp. 2113–247

Reiersøl O 1945 Confluence analysis by means of instrumental sets of variables. *Arkiv for Mathematik Astronomi och Fysik* **32**: 1–119

Rothenberg T J 1971 Identification in parametric models. *Econometrica* **39**: 577–95

Ruud P A 2000 *An Introduction to Classical Econometric Theory*. Oxford University Press, New York

Sargan J D 1958 On the estimation of economic relationships by means of instrumental variables. *Journal of the Royal Statistical Society, Series B* **21**: 91–105

Theil H 1958 *Economic Forecasts and Policy*. North Holland, Amsterdam

Wald A 1940 The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics* **11**: 284–300

Wright S 1925 *Corn and Hog Correlations*. US Department of Agriculture Bulletin 1300, January 1925, Washington, DC

Wright P G 1928 *The Tariff on Animal and Vegetable Oils*. Macmillan, New York

Zellner A, H Theil 1962 Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica* **30**: 54–78

J. H. Stock

Insurance

There is insurance when one party agrees to pay an indemnity to another party in the event of the occurrence of a predefined random event generating a damage for the latter. The most standard example is the insurance contract where the first party—the insurer— is compensated by being paid a premium by the second party, the policyholder. But many other contracts involve some forms of insurance. For example, in share-cropping contracts, the landlord agrees to reduce the rent of land in case of a bad crop. In many instances, employers agree to pay their employees more than the competitive wage in bad years. In cost-plus contracts, the buyer accepts to pay a larger price if the producer incurred an unexpected increase in cost. In the case of income taxes, the state partially insures the losses of taxpayers by reducing the tax payment when incomes are low.

The shifting of risk is of considerable importance for the functioning of our modern economies. Insurance is desirable for risk-averse agents as a risk-reduction device through the use of the Law of Large Numbers. Moreover, insurance allows for disentangling investment decisions from risk-taking decisions. Without it, there would certainly not have been the experience of the historical economic growth of the twentieth century. Ford, Solvay, Rockefeller, and the others would not have taken the investment risks that they actually took without the possibility to share the risk with shareholders and insurers. Similarly, most consumers would not purchase new expensive cars or houses if they would not be able to insure them.

However, various informational problems (observation of the risk, solvency issues, of the prevention efforts and/or of the loss by the insurer) can make competitive insurance markets inefficient by not providing enough coverage at an acceptable price.

1. The Social Value of Insurance

There is an added value to insurance only because policyholders are risk-averse, that is they dislike zero-mean risk on wealth. Consider an agent facing a random loss X to their wealth. An insurance contract stipulates a premium P and an indemnity schedule I(.) that determines the indemnity I(X) for each possible loss X. There is full coverage if I(.) is the identity function. The actuarial value of the contract is the expected indemnity EI(X). The insurance premium is said to be actuarially fair if it is equal to the actuarial value of the contract: $P = EI(X)$. Suppose that it is the case. Then, the purchase of a full insurance contract at an actuarially fair premium has the effect of replacing a random loss X by its expectation $P = EX$. The private value of such a contract is equal to the risk premium attached to the initial risk by the policyholder. It is increasing with the policyholder's degree of risk aversion and with the riskiness of the loss. As a first approximation, the riskiness of the loss can be measured by its variance.

In order to determine the social value of insurance, the effect of the transfer of the risk on the insurer's welfare should also be measured. Consider an insurer selling n fair full insurance contracts to n policyholders, each of them bearing a random loss X_i, with $X_1, \ldots, X_n$ being identically and independently distributed. The common wisdom is that the insurer does not bear any risk in the aggregate if it is able to sell enough such insurance contracts to cover independent risks, i.e., if n is large enough. This is not true, as explained by Samuelson (1963). It is a fallacious interpretation of the Law of Large Numbers which leads people to believe that accumulating several independent risks generates diversification. Indeed, the aggregate indemnity to be paid by the insurer is the sum of the X_i. Its riskiness measured by its variance equals $n\sigma^2$, where σ^2 is the variance of each individual risk X_i. If the insurer is risk-averse, the risk transfer is not Pareto-improving since it makes the insurer worse off. The classical view of modern finance on this problem is to recognize that insurance companies are not owned by a single person, but rather by a large set of shareholders. If insurable losses are uncorrelated with financial markets returns, the riskiness of insurance companies has no adverse effect on shareholders. In such a situation, the risk transfer to insurance companies is Pareto-improving, and the social value of insurance companies can be measured by the sum of the risk premia associated by policyholders to their random loss.

An alternative insurance scheme is to organize a mutual arrangement among the n risk-averse agents. Consider an arrangement in which the pool guarantees to each of its member a complete coverage of losses against an ex-post contribution equaling the average loss in the pool. In such a scheme, the random loss X_i is replaced by $(X_1 + \ldots + X_n)/n$, whose variance is $1/n$ the variance of X_i. In consequence, this mutual arrangement is Pareto-improving. At the limit, when n tends to infinity, individual risks are completely washed out by diversification.

2. Efficient Risk Sharing with Complete Markets

An important question is to characterize socially efficient allocations of risks in an economy where individuals may have very different initial risk exposures and risk attitudes. Wilson (1968) solved this question when individuals are expected—utility maximizers. Any efficient allocation of risk must satisfy two conditions. The first one is often called the mutuality principle. It states each ex-post individual's consumption only depends upon the ex-post average wealth in the economy. This means that all members raise their consumption in good states, and reduce it in bad states. There is no loser in an improving environment. Notice that this condition is satisfied in the case of the mutual arrangement examined above. The intuition for the mutuality principle is the need for diversification: all individual risks must be gathered in the pool. The second characteristic of a socially efficient risk-sharing rule deals with the sharing of the aggregate risk in the pool. As suggested by the intuition, a socially efficient sharing of the aggregate risk is such that the risk borne by a member of the pool must be inversely proportional to their Arrow-Pratt degree of absolute risk aversion. At the limit, if there is a risk-neutral agent in the economy, this agent should bear all risks, i.e., he should insure everyone else in the economy.

Whether decentralized economies can allocate risks in an efficient way is another question that has been solved by the pioneering works of Arrow (1953), Debreu (1959), and Borch (1962). Suppose that the economic system is such that it is possible to freely insure against any bet, at fixed odds, any amount wished on the occurrence of any event which will not affect the welfare of the insurers in any way. Markets are said to be complete in this case. Suppose also that there is no transaction cost on these markets, and no asymmetric information of any kind. The standard first theorem of welfare economics can then be applied to obtain that competition on these markets yields a Pareto-efficient allocation of risks in this economy. The competitive allocation must in particular satisfy the mutuality principle. At the competitive equilibrium, all households should sell their entire human and physical capital in exchange for a stake of the aggregate economy. A part of this story can be found in the Capital Asset Pricing Model in which all investors have the same composition of the portfolio of risky assets, the 'market portfolio.'

Townsend (1994) tested the mutuality principle in small rural villages in India. He did not reject the hypothesis of efficient risk sharing within each village. But he rejected the hypothesis of efficient risk sharing across villages. More generally, it is quite obvious that

the mutuality principle does not hold in large economies. Some workers may see their income reduced in a booming economy. In case of a laborsaving macroeconomic shock, workers may lose against capital owners. Some classes of retirees may enjoy an increase in their pension benefit even when there is a recession. Some countries may be in recession in a growing world economy. Shiller (1993) recommends that new institutions be created to share these society's large economic risks that are currently inefficiently shared. The welfare impacts of imperfect risk sharing are likely to be very large. In the remainder of this article, the reasons for these market imperfections are examined.

3. *Optimal Insurance with Transaction Costs*

Insurance entails transaction costs. In many lines of casualty insurance, transaction costs may be as large as 30 percent of the premium. Consider an economy with a risk-neutral insurer and consumers with idiosyncratic risks X. suppose that each euro of indemnity yields k euros of deadweight cost. In a competitive insurance market, the premium associated with insurance contract I(.) must be equal to $(1+k)EI(X)$. Mossin (1968) showed that the optimal insurance contract for the policyholder is no longer the full insurance contract. It is always optimal for him to retain part of the risk in order to reduce the insurance cost. This is because risk is a second-order effect relative to transaction costs.

Arrow (1965) examined the optimal form of the risk retention. There are various ways for policyholders to retain a share of the risk. The most standard is to accept a straight deductible, in which the indemnity is either zero if the loss is less than the prespecified deductible, or the loss minus the deductible otherwise. Alternatively, the insurance contract can contain a coinsurance rule in which case the indemnity is a prespecified percentage of the loss. Other clauses can be considered, as upper caps on indemnities, disappearing deductibles, etc. Arrow showed that the optimal insurance contract takes the form of a straight deductible. Any nondeductible insurance contract is dominated by a straight deductible contract with the same actuarial value, but yielding a smaller retained risk for the policyholder. Straight deductibles provide the best compromise between the willingness to cover the risk and the limitation of the insurance deadweight cost.

One can link transaction costs to undiversifiable risks. Obviously, many natural, environmental, or technological risks are in the class of large risks that are difficult to eliminate by using the mutuality principle. The insurer's risk-neutrality for these risks may be questioned. Insurance companies will not provide fair insurance premiums for them. Indeed, shareholders will not be able to diversify the risk associated to the dividends paid by insurance companies that cover these large risks. They will ask for a risk premium, which will increase the cost of capital of these companies. This cost will be passed on to policyholders through a larger premium rate for the component of individual risks that is systematic. It will induce them to retain part of their individual risk. In short, the fact that the risk is systematic induces insurance premiums to contain a positive loading that has an effect equivalent to a transaction cost. This is the logic behind larger deductibles for systematic risks.

4. *Adverse Selection*

Since the seminal paper by Rothschild and Stiglitz (1976), it is recognized that the fact that insurers face a heterogeneous population of consumers is a source of inefficiency on insurance markets. The classical model presented above allows for a heterogeneous population as long as the characteristics of the risk borne by each agent is common knowledge. For example, the fact that women are safer drivers than men is compatible with full insurance of every driver at the competitive equilibrium with a risk-neutral insurance industry. The premium rate for every category of risk will be fair, thereby inducing each individual to purchase full insurance at the optimum.

A problem arises when the population is heterogeneous, but the observable characteristics of the agents are not perfectly correlated to the intensity of their risk. The adverse selection problem originates from the observation that if insurance companies calculate the premium rate on the basis of the average probability distribution in the population, the less risky agents will purchase less insurance than riskier agents. In the extreme case, the low-risk agent will find the premium rate too large with respect to their actual probability of loss. They will prefer not to insure their risk. Insurers will anticipate this reaction, and they will increase the premium rate to break even, only on the population of high-risk policyholders. The literature on adverse selection is devoted to characterizing an equilibrium. Insurers will use the fact that low-risk agents and large-risk agents behave differently in the face of a large set of insurance contracts. In particular, low-risk agents could credibly signal their type by selecting a contract with a large deductible, something that high-risk agents dislike.

The presence of high-risk agents generates a negative externality to lower-risk agents who are unable to find an insurance premium at an acceptable premium rate. To illustrate, this is probably why the proportion of households that purchase life insurance is so small, despite the potential severity of the risk. People have private information about their health status that cannot be observed by insurance companies. Then, only those with the lowest life expectancy purchase life insurance.

The standard policy recommendation for improving risk-sharing efficiency under an adverse selection problem is to make public all relevant information about risks. For example, insurers should be allowed to know whether the potential policyholder has some severe illness. They should also be allowed to use genetic testing. Obviously, redistributional reasons may go against such a policy recommendation if the State is not in a position to compensate poor high-risk agents.

5. *Moral Hazard*

The population of risks can be heterogeneous not only because agents bear intrinsically different risks, but also because they do not invest the same amount of their energy, wealth, or time in risk prevention. In particular, it has long been recognized that individuals that are better covered by insurance invest less in risk prevention if the link between the premium rate and the size of these investments is weak. It will be the case if insurers are not in a position to observe the investment in risk prevention by the policyholder. In that case, the premium rate is not sensitive to the effort made by the policyholder to prevent losses. Obviously, there will be an inverse relationship between risk prevention and insurance coverage: policyholders will not internalize the benefits of their efforts. The level of risk prevention will be inefficient. This is *ex ante* moral hazard, common to all principal-agent problems. Anticipating this low degree of prevention and the higher frequency of losses that it entails, insurers will raise their premium rate. Full insurance will not be optimal for agents. At the limit, no insurance can be an equilibrium. Holmstrom (1979) characterized the equilibrium insurance contract with *ex ante* moral hazard.

To illustrate, this is why it is not possible to insure against the absence of promotion on the workplace, about failure at school or university, about the lack of demand for a new product, or about divorce. To some extent, this is also why it is hard to insure against unemployment, or against environmental and technological risks.

The policy recommendation to fight against *ex ante* moral hazard is the enforcement of norms for risk prevention. This is the case for environmental risks in which ships transporting chemical products have to satisfy various safety requirements that are imposed by regulatory agencies. Automobile driving norms are also standard, as speed limits, alcohol-free driving, etc. Why these norms are mostly organized by a regulatory agency rather than by insurers is not completely clear. One reason is due to the combination of negative externalities and limited liability. If there is more than one principal supervising the implementation of norms, the information among the different principals should be pooled to save on monitoring costs. For example, auto insurers should be allowed to get the information about driver fines by the police.

Another policy recommendation is to allow insurers to discriminate prices among different policyholders that exercise various preventive efforts. Allowing for discrimination is a way to provide incentive to policyholders to invest in risk-reducing activities. In France again, insurers are not allowed to discriminate homeowner premium rates on the basis of natural risks such as earthquake, flood, and storms. The consequences are by now notorious: many households built their houses in areas that were secularly known to be flooded periodically.

6. *Insurance Fraud*

The classical model assumed that the size of the loss is observable. There are many instances in which this is at best a crude approximation of the real world. Contracts can be made contingent only upon observable events. The problem here is to give the good incentives to the policyholder to report their actual loss. The difficulty for insurers to verify claims is at the origin of why it is not possible to insure against loss of happiness, or against some forms of suffering that cannot be measured by physicians.

There exist other types of risk for which outcomes can be observed by the insurer only at a relatively high auditing cost. Townsend (1979), Mookherjee and Png (1989), and others analyzed the optimal risk-sharing scheme in this case. If there is no limit on the penalty that can be imposed to policyholders that do not declare the actual level of their loss, the first-best solution can be attained. Indeed, insurers should announce that they will audit claims with some probability p that is very low. If the policyholder made a fraudulent claim, an unbounded penalty ('death penalty') would be imposed on him. This is enough to give the incentive not to defraud the insurance contract, even if p is very small. In this case, the fact that there is costly claim verification is not detrimental to welfare, and the risk is insurable in full.

But there are several reasons to believe that an unbounded penalty in case of a fraudulent claim is not a realistic assumption: ethics, limited liability, risk of legal errors, etc. *Ex ante*, it is then Pareto-efficient to limit the size of the penalty. In order to report their loss correctly, the insurer will have to audit claims at a high frequency. This entails additional costs on the insurance contract. If the auditing cost is high, or if the frequency of audit necessary to give the good incentive for the policyholder to reveal the truth is too high, consumers would be better off by not insuring the risk. Notice that another way to reduce the willingness to submit a fraudulent claim is to limit the indemnity. The maximal indemnity that is compatible with truth-telling is an increasing function of the penalty and of the probability of audit. Consumers would like to

announce ex ante that they will not submit fraudulent claims *ex post*. That would allow insurers to save the audit cost, thereby reducing the equilibrium premium rate, but the announcement is not credible.

Is *ex post* moral hazard an important problem? It is often suggested that the cost of fraudulent claims may well amount up to 10 percent of premiums paid for such insurance lines as automobile insurance or homeowner insurance. This estimation is the cost of unjustified indemnities to policyholders, not the auditing cost to fight against fraud. This percentage is comparable to the rate of transaction costs, whose effects on insurability have been previously examined.

The policy recommendation is clear from the discussion above: one should impose a larger penalty on policyholders that have been convicted of a fraudulent claim. Several countries in Africa, and to a smaller extent in Europe, have been weak in this area, recognizing fraud as a 'national sport' that should be forgiven. By doing so, the legal system imposes a possibly large cost to Society in terms of a loss of insurability. This weakness has been particularly clear for insurance lines where the indemnity payer does not have the good incentives to be tough on fraud. For example, one may question whether social security organizations are fighting fraudulent claims efficiently. This yields a general distrust of the system, which is detrimental to the unemployed themselves.

7. *Limited Liability*

An individual can cause damage to others, either in the course of his/her profession (medicine, surgery, house building, etc.) or because of other activities (e.g., driving a car). The same kind of external random effect occurs for firms. In most countries, the agent found liable for damage to others must indemnify them accordingly. This is done to force decision makers to internalize all costs generated by their choice. But indemnification is only possible up to the decision maker's financial capacity. Limited liability is a way to protect risk-takers against an excessive financial distress. But it has long been recognized that limited liability distorts the decision of the risk-taker in a way that is socially inefficient. The US Saving and Loans crisis is often explained by the fact that 'zombie' S and Ls adopted in the early 1980s a very risky attitude in an attempt to 'bet for resurrection' after some blows were delivered to their portfolio of (real estate) assets. This is because limited liability gives the agent the equivalent of a free put option. Put it in simpler terms, under limited liability, an insolvent agent can only benefit from taking more risk, because he does not bear the burden of losses. Therefore, if the agent is risk-neutral, he will seek to maximize the expectation of a convex function of his wealth. As a result, he will systematically exhibit a risk-loving behavior, and adopt a very risky attitude. This is a kind of moral hazard problem. Risk aversion mitigates this result, but only for agents who are well capitalized.

The effect of limited liability of the policyholder on their demand for insurance is thus unambiguous: if he is risk-neutral, it is never optimal to cover a risk of loss, even in the most favorable case where the premium rate is fair. Insuring the risk would yield a sure reduction in wealth equaling the expected loss. Not insuring the risk would yield an expected reduction of wealth that is less than it, since the agent bears only part of the risk of loss. Another way of looking at this problem is that the insurance contract creates a 'deep pocket' where victims can find compensation for their losses. This kind of problem is particularly crucial when examining the demand for insurance by firms for catastrophic environmental risks. Limited liability on the part of the insurance also reduces the demand of insurance, since it makes the indemnity dependent on solvency.

Limited liability thus raises several important questions. How to organize compensation for those who bear the negative externalities? How to build an incentive-compatible mechanism that increases loss prevention by decision makers with limited liability? How to solve the market failure of liability insurance markets? How to force firms not to under-capitalize their subsidiaries which are in charge of managing the riskiest part of the business? Two routes have been used. The first one is compulsory insurance. This solves the misallocation of risk in the economy and the organization of a system to compensate the victims. But, most of the time, compulsory insurance has been funded by a flat, non-discriminatory, non-incentive-compatible insurance tariff. The policyholder's investment in loss prevention is not observed by the fund, either because it is difficult to get information on it, or because the fund did not organize an incentive-compatible system.

The second route has been to organize a 'deep pocket' for decision-makers. For example, the hospital that employs an uninsured physician may be made liable in case of the physician's insolvency. Under the US CERCLA, when a bank has been relatively closely involved in the monitoring of a firm's activities, it may be considered by the courts as liable for cleaning up the environmental damages generated by the insolvent firm. The objective of this strategy is to force risk-takers to internalize the full cost of potential losses: the hospital will reduce the income of the careless physician, and banks will increase the loan rate of riskier firms. If there is no asymmetric information between the principal (the hospital, the bank) and the agent (the physician, the firm), the agent will select the socially efficient level of care and insurance. There would then be no insurability problem. But, there is no reason to believe that the principal can monitor the agent at no cost. The CERCLA legislation, for example, introduces more asymmetric information on credit markets. Consequently, there will be more credit

rationing, the cost of capital will be larger, and the structure of banking contracts for firms will be affected. Is insurability worth these costs?

8. *Regulation of the Insurance Sector*

Insurance markets are among the most heavily regulated ones around the world. In the Middle Ages, the Catholic Church was reluctant to recognize insurance as an ethically acceptable activity. The French revolution prohibited insurance companies in 1793. States have been very active to control insurance tariffs, in particular in life insurance. They also imposed compulsory insurance. More recently, they established solvency rules similar to those of the banking sector.

There is a European tradition to use insurance for a redistributive purpose. Insurance implements wealth redistribution if the expected profit of the insurer depends upon the observable characteristics of the policyholders. The expected profit on the low risk class of policyholders is used to cover the expected loss on the high risk class. This scheme is standard in unemployment, health, and disability insurance, but it can also be found in other insurance sectors like automobile (for young and old drivers), homeowner (for natural catastrophes), and life (for the prohibition of using health factors to price life insurance). However, the solidarity among different risk classes is difficult to organize with decentralized insurance markets without strong regulation. The absence of discrimination on the basis of policyholders' observable characteristics that are correlated with the risk introduces adverse selection. It also generates moral hazards if these characteristics can be affected by the preventive effort of the policyholder. In order to alleviate the adverse selection problem, regulators often impose compulsory insurance, as in social security systems. Pricing regulation is also necessary to react to the natural tendency of competitive markets to discriminate. Finally, insurers should be forced to supply coverage to all customers at the undiscriminating premium. Various reforms of the health insurance sector in OECD countries have been proposed to deal with the problems inherent in such heavy regulation.

There is a form of saving in an insurance contract from the policyholder's point of view. Indeed, the indemnity, if any, may be paid a long time after the premium. This is particularly true for life insurance. The lender, i.e., the policyholder, should check the solvency of the borrower-insurer. In perfect competitive markets, insurers facing a higher risk of insolvency should be penalized by collecting smaller premia, to compensate the risk borne by the policyholders. This is enough to provide good incentives for insurance companies to manage their financial reserves efficiently. However, policyholders may not be in a position to evaluate correctly the probability of failure of the insurers. This asymmetric information calls for centralized solvency regulation. It takes two forms. The first type of regulation is about solvency ratios, similar to those imposed on banks. The second type of regulation is about the kind of risks that can be taken by insurers on the investments they make with their financial reserves. Typically, there is an upper limit on the share of reserves invested in stocks and real estate. Many countries prohibit the use of derivative assets, in spite of their potential benefits for risk reduction. These constraints must be strong enough to fight against the moral hazard problem generated by asymmetric information and limited liability. It should not be too strong in order to allow for efficient Asset-Liability Management (ALM) by insurers.

See also: Behavioral Economics; Insurance and the Law; Regulation, Economic Theory of; Regulatory Agencies; Risk, Sociological Study of

Bibliography

Arrow K J 1953 Le Rôle des Valeurs Boursières pour la Répartition la Meilleure des Risques, Edition du CNRS, Paris

Arrow K J 1965 *Aspects of the Theory of Risk Bearing, Yrjo Jahnsson Lectures*. Helsinki, Reprinted in 1971 *Essays in the Theory of Risk Bearing*. Markham Publishing Co., Chicago]

Borch K 1962 Equilibrium in reinsurance market. *Econometrica* **30**: 424–44

Debreu G 1959 *Theory of Value*. Wiley, New York

Holmstrom B 1979 Moral hazard and observability. *Bell Journal of Economics* **10**: 74–91

Mookherjee D, P'ng I 1989 Optimal auditing, insurance, and redistribution. *Quarterly Journal of Economics* **103**: 399–415

Mossin J 1968 Aspects of rational insurance purchasing. *Journal of Political Economy* **76**: 533–68

Rothschild M, Stiglitz J 1976 Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* **80**: 629–49

Samuelson P A 1963 Risk and uncertainty: a fallacy of the law of large numbers. *Scientia* **98**: 108–13

Shiller R J 1993 *Macro Markets: Creating Institutions for Managing Society's Largest Economic Risks*. Oxford University Press, Oxford, UK

Townsend R 1979 Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* **21**: 1–29

Townsend R M 1994 Risk and insurance in village India. *Econometrica* **62**: 539–92

Wilson R 1968 The theory of syndicates. *Econometrica* **36**: 113–32

C. Gollier

Insurance and the Law

Insurance may be uniquely tied up with the law. Social insurance and other public sector insurance arrangements are creatures of statute, governed through administrative law (see *Social Insurance: Legal Aspects*).

Private sector insurance arrangements depend upon a well-functioning contract law and a regulated market (see *Contracts: Legal Perspectives*; *Markets and the Law*). Insurance thoroughly dominates the field of tort law, especially with respect to the liabilities of individuals and small business (see *Torts in Law*). Moreover, because insurance institutions inevitably exert a regulatory force over their subjects, insurance must be understood as a complement to direct state regulation. Finally, some social theorists have argued that insurance has exerted a profound ideological force on law and regulation, as activities increasingly are governed through risk (see *Law and Risk*).

1. *The Term 'Insurance'*

Spencer Kimball, one of the leading insurance law scholars of the twentieth century, once wrote 'There is no good definition of "insurance," for any purpose.' (Kimball 1992, p. 1). Kimball's admonition notwithstanding, it is useful to begin by thinking of insurance as a formal mechanism for sharing the costs of misfortune.

Contemporary insurance arrangements typically involve fixed premiums, paid in advance, and guaranteed benefits in the event that a specified loss or event occurs; but these are not necessary elements of the insurance form. Insurance arrangements typically provide protection against risks that are fortuitous from the perspective of the insured, yet reasonably predictable in the aggregate by an insurer, and which are not so catastrophic that they would overwhelm the financial capacity of insurers (as would, for example, insurance against worldwide depression). Ideas about what is and what is not insurable, however, vary widely over time and place.

Following the French social theorist Francois Ewald, it is helpful to distinguish among four aspects of insurance: technologies, institutions, forms, and visions (Ewald 1991). These categories illustrate the conceptual variety of insurance activities and, correspondingly, some of the deficiencies in analytical definitions.

Insurance, in the sense of insurance 'technology,' refers to a set of procedures for dealing with risk. Examples include the mortality tables and inspection procedures of ordinary life insurance, the incentive-based medical provider contracts and computerized claim processing procedures of managed health care companies, the payroll tax and administrative review procedures of social insurance programs, and the standard-form insurance contracts used in almost all private insurance. As a technology for managing risk, insurance extends beyond what might ordinarily be understood as the insurance field; life insurance companies were pioneers in epidemiology and public health, and fire insurance companies formed the first fire departments. More recently, health insurance companies have been behind many efforts to compare, test, and measure the effectiveness of medical procedures.

Insurance 'institutions' are the various kinds of organizations that provide insurance. Social insurance agencies (such as the German Arbeitsamt) and stock insurance companies (such as the Dutch-based ING Group) are two (very different) insurance institutions. Insurance 'forms' are the various kinds of insurance provided by insurance institutions, as well as the variations in form among them. Unemployment and property insurance are two different forms of insurance. Fire insurance and flood insurance are two different forms of property insurance.

Finally, insurance 'visions' are ideas about and images of (or, alternatively, discursive practices regarding) insurance that animate the development of insurance technologies, institutions, and forms. Three examples follow.

(a) It was once commonly believed that life insurance was immoral, because it interfered with divine providence, equated life and money, or was a form of gambling (Zelizer 1979). This vision of insurance had important consequences for the development of insurance law and institutions in the west. It slowed the growth of life insurance, and it helps to explain the American and French preoccupation during the nineteenth century with establishing the morality of all kinds of insurance. Related ideas affect Islamic insurance institutions today (Vogel and Hayes 1998).

(b) Because the primary benefit of insurance is a sense of security which, for most people, will never be tested by a catastrophic loss, much of the value of insurance rests in the imagination. Recognizing the significance of imagination to insurance, courts in the United States place great emphasis on the 'reasonable expectation of the insured.' In determining the nature of that expectation, courts invoke two distinct visions of insurance. The first mirrors the vision of insurance portrayed in insurance advertising: a promise 'to be there,' conveyed by narratives of family and the need to protect the individual against sudden misfortune. The second mirrors the vision of insurance that companies invoke when denying claims, a complicated amalgam of tough love and protecting the insurance fund, conveyed by narratives of institutional ethics and the need to protect ratepayers against fraud and abuse. In resolving insurance contract disputes, courts first decide which of these visions to employ. Who wins a dispute often depends as much on which vision the court adopts as on how the court applies that vision (Baker 1994).

(c) The actuarial vision of insurance has had great influence over the development of insurance and insurance law. In that vision, the ideal type of insurance involves premiums paid in advance, guaranteed indemnity in the event of a covered loss, and risk-based premiums based on the best available

information regarding the expected losses of the individuals insured (Abraham 1986). This vision of insurance helps explain (i) the decline of fraternal insurance over the nineteenth and early twentieth centuries, as actuarial expectations overcame values of friendship, brotherly love, and charity, (ii) also the decision to model unemployment insurance on private insurance, and the related effort to tightly link benefits to premiums, (iii) the intensity of the popular belief that US Social Security retirement benefits have been earned by the people who collect them, and (iv) the corresponding expert belief that Social Security is not really insurance (because the money to pay today's retirees comes from the contributions made by today's workers and not from the contributions of the retirees themselves). Indeed, the actuarial vision of insurance has been so successful that many well informed people would deny that it is a vision at all and would assert instead that it is the only valid model of insurance (Baker and Simon 2001).

2. *Insurance as a Form of Regulation*

Given the size of insurance institutions relative to western economies, it is surprising that social scientists have not paid more attention to insurance. Indeed, looking at the broad sweep of twentieth century social policy, it is tempting to describe insurance as the sleeping giant of power.

In setting eligibility requirements and benefit levels for social insurance, the state is obviously engaged in the regulation of populations; yet private insurance can also be a crucial form of (delegated) state power. Rather than set their own criteria for access to economic freedoms like operating an automobile or a business, states often mandate that a person obtain some form of insurance. Examples include liability insurance for automobile owners, workers compensation insurance for employers, and (in the United States) surety bonds for companies engaged in business with the state. The state leaves it to the private market to set underwriting criteria that will determine access to these privileges.

Motivated by controlling losses and assessing risk, insurance companies establish norms of conduct which they enforce by contract terms and pricing and, ultimately, through the judicial system. Whether obtained through compulsion or prudence, insurance is a form of regulation. Exclusions and conditions written into coverage for property, life, and health amount to private legislation. Significantly, this 'legislation' often acts with in the home or business, where, traditionally, the sovereignty of the King did not run (O' Malley 1991).

2.1 *Insurance as Tort Regulation*

The field of law most thoroughly dominated by insurance is tort law. Liability insurance pays the costs of claims made against the insured person. Because it is so difficult to collect money damages without insurance, liability insurance determines who is capable of being sued, for what wrongs, and for how much. The result is that tort law in action is shaped to match the liability insurance that is available. Also, liability insurance makes tort law in action more focused on managing aggregate costs and less focused on the fault of individual defendants (Ross 1970).

Other forms of insurance also regulate tort law. For example, liability commonly is limited whenever a statutory insurance regime is established. This not only creates a bar on certain types of tort claims, but also leads plaintiffs to shape their claims to avoid the tort bar, thereby shaping the tort system. In addition, people may be less likely to sue in tort when they have health, disability, or property insurance covering an injury (depressing the aggregate of tort damages). Moreover, in jurisdictions that deduct health or other insurance payments from tort damages, the existence of that insurance also reduces tort damages. Perhaps paradoxically, the existence of first-party insurance can also in some cases increase tort damages, by reducing the immediate financial need that might otherwise force the plaintiff to settle cheaply and quickly.

Insurance has also exerted an ideological effect on tort law. The existence of liability insurance helped make it possible to conceive of tort law as a risk spreading system—indeed, as a form of insurance—rather than simply as a mechanism for determining right and wrong in individual situations (cf. Ewald 1986). This concept of tort law has had a variety of consequences over the last 100 years. These include the simplification of tort doctrine, the elimination of common law exceptions to tort, and the partial replacement of tort law by statutory insurance schemes in jurisdictions such as Quebec, Israel, and New Zealand (Sugarman 2001). More recently, the concept of tort law as insurance has lent intellectual force to political efforts in the US and Britain to cap damages and to eliminate 'coverage' for pain and suffering, on the grounds that the benefits provided by the tort/insurance system should mimic the (lesser) benefits provided by first party insurance (Clarke 1997).

2.2 *Insurance and Governing Through Risk*

The ideological effect of insurance on tort law illustrates a more general phenomenon, namely the role that insurance technologies, institutions, forms, and visions have played in governing through risk (Baker and Simon 2001). Insurance institutions pioneered the use of formal considerations of risk in the direction of organizational strategy and resources. Moreover, the actuarial techniques adopted by a wide range of institutions, from police departments to social service agencies and money managers, all depend upon think-

ing about the world in the probabilistic, demographic manner that insurance has helped make possible (Ewald 1986).

For most of the twentieth century the dominant form of governing through risk was that of spreading risk, as insurance institutions increasingly assumed financial responsibility for risks faced by individuals, families, and organizations. More recently, both public and private insurance institutions may be placing a greater emphasis on individual's responsibility to embrace risk. The governmental rationales inherent in spreading or embracing risk have significant consequences for law. The spreading of risk leads to the simplification and expansion of tort law, to more expansive interpretations of the statutes regulating social insurance and the contracts regulating private insurance, and, in general, to more efforts to soften the hard edges of life. Embracing risk leads to reductions in social insurance benefits, deductibles, and benefit caps in private insurance, and links between pension benefits and market performance.

3. *Insurance Law*

Traditionally, legal scholars have divided the field of insurance law into two parts: (a) the law concerning the relationship between private insurance organizations and their insureds, which is considered a part of contract law; and (b) the law concerning the state regulation of private insurance organizations, which is considered a part of the law of regulated industries (e.g., Clarke 1997, Cousy 1999, Jerry 1996, Schimikowski 1999). (For a broader view of insurance law including the law relating to social insurance, see *Social Insurance: Legal Aspects*.)

3.1 *Insurance as a Species of Contract*

Contemporary insurance institutions grew from two distinct roots: mutual benefit associations dating back to the medieval gilds (or earlier), and marine insurance arrangements dating back to fifteenth century Italian city states (and possibly earlier, e.g., Greek bottomry loans, which were loans to merchants that were repaid only if cargo made it safely to port). Early insurance law treatises report that courts treated both these forms of insurance within the framework of contract law, adjudicating the obligations of the company or society according to the promises made in the insurance contract.

At least by the mid nineteenth century, courts in Europe and elsewhere recognized that insurance contracts differed significantly from what was traditionally understood as the ideal type of contract (a voluntary agreement, with terms that were negotiated between two parties with equal bargaining power). Because of the gatekeeper role of insurance institutions, insurance can hardly be said, in many instances, to be voluntary. Insurance companies almost universally employ standard form contracts with terms that are not subject to negotiation. And, in all but a very few cases, the parties do not have equal bargaining power. Typically, the insurance company is a much larger economic entity; competing insurance companies rarely offer significantly different terms (except sometimes price), and the insurance company has information about the meaning and value of the contract that the applicant for insurance does not. In addition, the money-for-promise nature of insurance gives the insurance company tremendous power once the insured has a claim; at that point, the insured cannot shop for a new insurance policy.

For these reasons, courts have classed insurance contracts within the general category of contracts of adhesion, and have developed somewhat more protective rules regarding the interpretation of insurance contracts (Baker 1994). Insurance contracts have long been regarded as the paradigmatic contract of adhesion (see *Contracts: Legal Perspectives*.)

Recently, the federal legislature and judicial system in the United States has acted to shift an increasingly large portion of the health insurance market away from a private contract regime to a not-yet-stable amalgamation of contract, trust, and administrative law. As private health insurance markets grow in Europe and elsewhere, there are likely to be similar efforts to create a paternalistic, yet market-oriented, legal regime governing the relationship between insurance companies and their members.

3.2 *Insurance as a Regulated Industry*

Contemporary insurance regulation dates to the nineteenth century, when a rash of insurance insolvencies in the US and Europe led to the establishment of state regulatory authorities. States limited the kinds of investments insurance organizations were permitted to make, and mandated the employment of actuaries to calculate rates and reserves, the filing of reports with state agencies, and minimum capital reserves.

In economic terms, the justifications for these and more recent forms of insurance regulation are information problems and the positive externalities of insurance. Most insurance consumers are poorly equipped to evaluate the soundness of insurance companies, to compare insurance contracts, or to evaluate the degree to which insurance companies have lived up to their promises. These information problems justify solvency, contract, and market conduct regulation. The positive externalities of insurance are the benefits that insurance provides to people other than the direct beneficiaries of insurance. Examples include the public health benefits of health insurance and the victim compensation benefits of liability

insurance. These positive externalities justify regulation directed at expanding access to insurance.

Solvency remains the primary focus of insurance regulation worldwide. Regulatory tools have expanded to include risk based capital requirements, electronic auditing of accounts, and a wide variety of limits on the ways that companies can invest the funds held in reserve to pay claims. Not surprisingly, these tools have been used for other than their express purposes. Insurance can be a powerful engine of capital accumulation, and investment regulations can be used to steer capital into preferred fields. French insurance companies, for example, are required to invest some of their funds in French real estate, with the interesting result that French insurance companies have become a major force in the French wine industry. On a larger scale, prohibitions on foreign investment in insurance in countries such as India, China, Brazil, and Argentina were long justified as a way to steer capital to indigenous insurance institutions (typically government owned or authorized monopolies), which would invest the capital locally. Recently, the International Monetary Fund and the World Bank, along with the globalization of the economy, have been significant forces in opening up capital markets—including insurance—to foreign investment (see *Globalization: Legal Aspects*).

Understanding insurance as an institution for storing and accumulating capital, it is no surprise to learn that insurance firms compete with banking and securities firms. Yet, banking, insurance, and securities traditionally have been subject to different regulatory regimes. The contemporary 'convergence' of the insurance, banking, and securities industries in the financial services marketplace places great strain on the existing regulatory institutions, as they struggle with each other and the firms they regulate, both to achieve regulatory ends and to maintain regulatory authority (Jackson 1999). Convergence and the related trend toward globalization are likely to be the primary economic forces driving the evolution of insurance regulation in the foreseeable future. This evolution will address such fundamental issues as whether, and to what extent, there will be democratic control over capital and the proper level of governmental control (local, federal, or international) over regulatory decisions.

See also: Contracts: Legal Perspectives; Health Insurance: Economic and Risk Aspects; Insurance; Legal Insurance; Markets and the Law; National Health Care and Insurance Systems; Regulation, Economic Theory of; Risk, Sociology and Politics of; Social Insurance: Legal Aspects; Torts in Law

Bibliography

Abraham K 1986 *Distributing Risk*. Yale University Press, New Haven, CT

Baker T 1994 Constructing the insurance relationship: Sales stories, claims stories and insurance contract damages. *Texas Law Review* **72**: 1395–433

Baker T, Simon J (eds.) 2001 *Embracing Risk*. University of Chicago Press, Chicago, IL

Clarke M 1997 *Policies and Perceptions of Insurance*. Oxford University Press, New York

Cousy H (ed.) 1999 *International Encyclopedia of Laws, Insurance*. Kluwer Law International, Boston, MA

Ewald F 1986 *L'état providence*. Grasset, Paris

Ewald F 1991 Insurance and risk. In: Burchell et al. (eds.) *The Foucault Effect*. University of Chicago Press, Chicago, IL

Jackson H E 1999 Regulation in a multisectored financial services industry: An exploratory essay. *Washington University Law Quarterly* **77**: 319–97

Jerry R H 1996 *Understanding Insurance Law*. Bender, New York

Kimball S 1992 *Cases and Materials on Insurance Law*. Aspen, Boston, MA

O'Malley P 1991 Legal networks and domestic security. *Studies in Law, Policy and Society* **11**: 171–90

Ross H L 1970 *Settled Out of Court*. Aldine, Chicago, IL

Schimikowski P 1999 *Versicherungsvertragsrecht*. C. H. Beck, München, Germany

Sugarman S D 2000 A century of change in personal injury law. *California Law Review*, **88**: 2403–36

Vogel F E, Hayes S L 1998 *Islamic Law and Finance: Religions, Risk and Return*. Kluwer, Boston, MA

Zelizer V 1979 *Morals and Markets: The Development of Life Insurance in the United States*. Columbia University Press, New York

T. Baker

Integration: Social

Social integration refers, in the first instance, to the extent and intensity of the interlinkages among the constituent parts of a social unit. This definition applies to all types of social units—a small group like a family; a gang or an Internet newsgroup; an organization like a firm, an administration, or a nongovernmental humanitarian organization; a community like a neighborhood, a local church community, a local business community, or a local ethnic community; or a society like a local society, a national society, a supranational society, e.g., European society or even world society (Münch 1998, pp. 27–67).

1. Actor Integration: Cohesion and Breaking Apart, Integration and Disintegration, Harmony and Conflict

What are the constituent parts of social units? On one level, we can say that these are actors as members of social units like groups, organizations, communities, and societies. On a second level, families form constituent parts of clan societies, estates of estate

societies, classes of class societies, functional units of functionally differentiated societies, organizations of organizational societies; groups and divisions are second-level constituent parts of organizations. What is then the extent and intensity of interlinkages between the constituent parts of a social unit? With regard to members of social units as individual actors or corporate actors (groups, organizations, communities, national societies within world society) it is the number of actors involved in recurrent interaction (extent) and the frequency of interactions (intensity). Such interactions can be more or less harmonious or conflictual. Actors can be extensively and intensively linked by conflict. For example, a couple could live in war from marriage to death. Such a case would be covered by our thus far outlined definition of social integration. Forces of attraction which keep them together must be stronger than forces of repulsion (Simmel 1992, pp. 312–15, Coser 1956, pp. 81–85).

Here we enter the field of explaining social integration. One explanation is that they are linked together because both of them have either no alternative at all or alternatives that are too costly so that transaction costs would be too high. For reasons of clarity, we can use the term 'cohesion' in order to denote a type of relationship between the members of a social unit which entails extensive and intensive interlinkage, as there are more forces of mutual attraction than of mutual repulsion. The opposite of social integration is not conflict but social disintegration; both are the extreme poles of a continuum with many gradations in between and a middle zone where it is difficult to determine precisely where integration ends and disintegration begins. We can say that a couple is more integrated the more they share a common life together and the more their action is mutually supportive, and it is more disintegrated the less this is the case. Just as integration has its opposite in disintegration and not in conflict, so conflict has an opposite of its own, namely harmony. Conflict means that with two or more actors, one actor's gain implies the other actor's loss. They may have different goals which cannot be attained at the same time, or they may have the same goal which, however, cannot be reached by every actor at the same time. Harmony, instead, implies that every actor's goal can be realized at the same time or that one and the same goal can be reached by every actor at the same time.

2. *Action Integration: Negative and Positive Integration*

A more ambitious notion of social integration is at stake, when we do not only look at the integration of actors as members of social units, but also at the integration of their actions. This is a second step in defining social integration. The integration of actions can be realized in two ways: negatively and positively. The negative way avoids the actions interfering with each other. The ideal-typical case of this integration of actions is embodied by two farmers who live on their own on their demarcated territory with guarded fences so that there is no chance of intrusion in either's domain of freedom. This kind of coordinated freedom can be called negative freedom. It is a freedom which results from preventing others from intruding in one's domain of free action. The corresponding type of integration can be called 'negative integration.' Durkheim used the term 'negative solidarity' for that (Durkheim 1964). It refers to the mutual acknowledgment of property rights. In the case of negative integration, the actors do not need more than the acknowledgment of their freedom to act by other actors; they do not require their cooperation and support.

The counterparts to negative freedom and negative integration are positive freedom and positive integration, which explicitly rely on mutual cooperation and support (Durkheim 1964). The most elementary form of positive freedom can be found in market exchange. As Adam Smith (1937) demonstrated, the utility maximization of one actor does not occur at the cost of other actors here, but even furthers the other actors' own utility maximization. In the founding years of sociology, it was Herbert Spencer (1972) who promoted this idea. The actors are not engaged in a zero-sum game, but in a game with a growing sum of which potentially every actor can make a profit. Their actions are in a complementary relationship to each other. Both exercising negative freedom on one's own space without interference from others and positive integration through market exchange are only possible if the actors share a minimum of rules of the game; otherwise they would end up in struggles for advantages by unlimited means including force and fraud, the Hobbesian war of all against all (Hobbes 1966). The question is, then, how it is possible that free actors come to share a minimal set of rules of the game. In a first step, cooperation in market exchange might rely on the human individual's sympathy for other people as Adam Smith, as well as David Hume, assumed (Smith 1966, Hume 1980). However, this sympathy mostly does not reach beyond particular communities and does not always guarantee the commitment to the rules of the games. Therefore, there is a need for further arrangements. Here the notion of 'positive freedom' becomes relevant. We move from the Scottish moral philosophers' individualistic theory of market integration to Rousseau's collectivistic theory of integration by the general will based on the republican spirit of citizens (Rousseau 1973). A further tradition of thought, which takes on significance here, is German idealism. It was founded by Fichte, Schelling, and Hegel in response to Kant's critical philosophy and carried on in transformed ways by Marx and his followers right through to critical theory and

Habermas's theory of communicative action (Kant 1956, Hegel 1965, Marx 1962, Habermas 1984, 1989).

3. Social Integration and Social Order

At this stage of our treatment of social integration we realize that social integration comes close to social order as soon as we ask for the integration of free action and of rights to act. Integration and disintegration then correspond to the pair of order and disorder or anomie in Durkheim's terms (Durkheim 1964). From this viewpoint, the actions of individuals are integrated, in as much as everybody can rely on normative regulations that determine the space of their actions so that negative interferences are avoided. Social order of this kind is a normative order, as Talcott Parsons has emphasized in particular (Parsons 1968). It is based on processes of regulation through collectively binding decision making. Its opposite, disorder or anomie, means that individual actors cannot count on such regulations when they encounter other actors and thus have to rely on the power available to them in order to defend their claims for space for action. As everybody has to do so, there is always the danger of the Hobbesian war of all against all.

If actors and groups are free to articulate interests within a wide framework of constitutional rules, and if the actors see themselves dependent on the tolerance and/or cooperation of the other actors, they will engage in bargaining processes in which each actor looks for the maximization of his or her interests. Such rules will have better chances of survival which allow for the growing satisfaction of a growing number of interests. Here the actors conceive of themselves as utility maximizers who learn to cooperate in a way which is beneficial for all. This is how the evolution of cooperation is explained from the point of view of utility maximization. This order is an ongoing process of steady change which serves for one interest some time and another interest at another time, or better, an always changing mixture of them. The evolutionary explanation of social order in economic terms has been particularly demonstrated by Robert Axelrod (1984). Generally speaking, this tradition of thought in sociology goes back to Herbert Spencer (1972) and is represented in contemporary sociology, for example, in the works of George C. Homans (1974), Michael Hechter (1987), and James Coleman (1990). The basic paradigm is rational choice theory. There have been many attempts to transcend the limits of economic thought, beginning with the classics, particularly Adam Smith in his work on moral sentiments (Smith 1966). One of the more prominent recent attempts is Jon Elster's acknowledgment of the noneconomic, in his view nonrational, emotional sources of norm-regulated behavior (Elster 1989).

However, there are also other settings which exert a different shaping influence on the emergence and continuation of social order. There may also be unbridgeable conflicts between the goals of major actors. In this case, social order is an outcome of a power struggle with winners who shape social order by their ideology and impose that order on the defeated through the application of monopolized physical force. This is what Dahrendorf (1958) has emphasized in his conflict theory of social order. A more complex model of conflict theory for the explanation of social order, which includes negotiation and ritual, has been constructed by Randall Collins (1975).

A completely different setting is given, when there are well-established institutions of discursive reasoning and professional legislation in the hands of legal and other types of experts. Both interest negotiation and power struggle have a more limited space for influencing the emergence and continuation of social order. This process takes place much more as the steady building up of an all-embracing, consistent, coherent, and detailed set of laws. The major question, which decides on the success of a certain proposal of legislation, is not so much as to what interests are served and what power can be mobilized, but as to what good reasons can be given and as to how it fits in the existing system of law. This type of social order is a steadily growing legal order which covers almost everything that changes in an ongoing process of adjustments to new problems and challenges. In this perspective, we conceive of social order as a norm-building process. There are different approaches to addressing this problem. The tradition of German idealism sees norm-building as a process of meaning construction. The modern version of this approach is Habermas' (1984, 1989) discourse ethics which explains norm-building as a procedure of rational argumentation. Another perspective of relevance here is phenomenology as developed by Schütz (1964), Berger and Luckmann (1966) and others. Here, norm-building is conceived of as a process of institutionalization through habitualization, sedimentation, and traditionalization as well as legitimation from rudimentary explanations to generalized symbolic. From the point of view of Harold Garfinkel's (1967) ethnomethodology and related conversational analysis, normative order is an ongoing process of ordering through the concerted action of individuals in their respective situation. For symbolic interactionism as promoted by Herbert Blumer (1969)—based on the work of George Herbert Mead (1934)—and by Anselm Strauss (1978), for example, order is negotiated in processes of mutually fitting different perspectives on the situation. In the pragmatist perspective developed in the works of William James, Charles Sanders Peirce, John Dewey, and George Herbert Mead, and revitalized by Hans Joas, order is produced and reproduced by the creative acts of individuals who cooperate in the solution of problems (Joas 1992).

A further factor which exerts an influence on the emergence and continuation of a social order is the building of relationships of solidarity in the sense of mutual support and help. The breakdown of the traditional order and its hierarchical system of privileges for estates was partly caused by lacking the inclusion of the rising bourgeoisie according to its contribution to the working of the society, particularly in terms of tax payments. There was a new class that was inadequately included in the participation in associational, political, and cultural life. The bourgeois revolutions established a new form of inclusion which was much broader. Inclusion was no longer based on inherited status, but simply on achievement (particularly in terms of market income, but increasingly also in terms of education and professional status). This new form of inclusion served as the door for every further movement aiming at participation in economic, political, associational, and cultural life and their organizations. The first one to succeed was the labor movement with the inclusion of the working class; it was followed by different kinds of movements, from the women's movement, the civil rights movement, and the movements for any kind of minority rights, right through to the movements for multiculturalism, prochoice and prolife. The sociological approach that sees the emergence and continuation of social order particularly as a process of solidarity, civic order and citizenship production, reproduction, and transformation, was founded by Emile Durkheim. This approach is carried on, in particular, in the functionalism of Talcott Parsons and in its contemporary revitalization by neofunctionalism. A further major contribution is T. H. Marshall's work on citizenship (Durkheim [1893] 1964, Parsons 1971, Münch 1998, pp. 140–65, Alexander 1998, Marshall 1964).

4. *Social Integration Beyond the Nation State*

A new level of inclusion has now been achieved in the process of transcending the system of nation-states by supranational units like the European Union or even the United Nations. A major driving force behind this transformation of social integration is the international division of labor. In Emile Durkheim's terms, the shrinking of distances through population growth, and, in the highly developed regions, much more through the increasingly faster methods of transport and communication are the major forces which promote the growing division of labor (Durkheim [1893] 1964). In this situation, competition for scarce resources intensifies and leaves specialization as the only effective means of escaping the struggle of all against all, besides suicide or emigration, which are, however, less attractive for people who love their lives and their country. Emigration is also only effective for people in poorer regions yet ever less feasible because of the crowding of the economic centers. The change from nation-state inclusion to European and worldwide inclusion is a further step from mechanical to organic solidarity in Durkheim's view. The high level of equality of results in the European welfare states can be interpreted as the most advanced level of founding the organic solidarity of labor division on much reconstituted mechanical solidarity within the boundaries of the nation-state. It reserves the equality of results for natives and excludes people in the poorer regions from its accumulated wealth. In Max Weber's terms, ingroup morality and outgroup morality are strictly differentiated (Weber 1923, pp. 303–4). Within the boundaries of the European Union, this differentiation is fading and being replaced by one comprehensive morality which is now less substantial and concrete and more formal and abstract justice as fairness within and beyond the boundaries of the nation-state (Münch 1993, 1998).

5. *The Negative Effects of Social Integration*

Social integration is not an overall positive state of affairs. In as much as it is limited to a particular unit below world society, it is always accompanied by at least some demarcation from other social units and corresponding disintegration. In this sense, nation-state disintegration is a precondition for supranational integration. Further on, social integration resides at least to some degree on processes of internal homogenization which may include not only the peaceful merging of different regions, ethnicities, and cultures, but also the imposition of a central culture on peripheral cultures, of the domination of a central ethnicity on peripheral ethnicities, of a central region on peripheral regions. Instruments for this process of 'internal colonization' are the penetration of the center's language, legal system, bureaucracy, school system, and communication system into the peripheral regions (Hechter 1975). In this way, social integration levels down differences and thus the richness of cultures and social life as well as the potential for innovation and change. Thus we can say that a balance of integration and at least some disintegration is a necessary prerequisite for preserving the potential for diversity and innovation. This was particularly emphasized by Emile Durkheim 1982). Georg Simmel 1992) pointed out the interrelation of integration and disintegration, harmony and conflict. Lewis Coser (1956) and Donald Levine (1980) revitalized this contribution by Simmel emphatically.

6. *Social Integration and Systems Integration*

Social integration becomes increasingly complex and increasingly requires a balance with elements of disintegration the more we proceed in the evolution of human societies. Clan societies in the early stages of sociocultural evolution are integrated by blood and rules of intermarriage. Traditional estate societies are

integrated by every estate serving a specific function for the living of the whole and the attribution of privileges according to the inherited status. Cohesion and the integration of actions is promoted by the hierarchical order of inherited ranks (Weber 1976 pp. 133–5, 148–55). Modern societies are integrated in a much more complex way. Their cohesion is based on comprehensive inclusion according to achievement and the constitution of citizenship with civil, political, social, and cultural rights as an abstract community beyond any primordial ties of ethnicity, color, gender, or any other kind of group.

Beyond social integration, modern societies are systemically integrated in as far as they have established functionally specified systems like economy, polity, law, science, education, the media, arts and literature, or religion. Money and democratically legitimated political power are examples of media of communication which serve as means of systemic integration in as much as they allow for the coordination of extremely differentiated economic interests or political goals, irrespective of time, place, and persons involved (Lockwood 1964, Luhmann 1988).

In the process of globalization the functionally specialized systems increasingly reach beyond the nation-state with its relatively strong social integration and produce a world society which is much more integrated systemically than socially through solidarity production and a binding social order (Luhmann 1997). It is disputed whether we approach worldwide systemic integration without social integration or whether there is a necessity and also a chance of a subsequently growing social integration on the supranational and even global levels. If we understand social integration in terms of the relatively strong positive integration of the nation-state, there are little chances for social integration on the supranational and global levels. If we realize, however, that this process involves a transformation of social integration in terms of farther-reaching and internally more differentiated solidarity of networks instead of homogeneous social units, which is accompanied by a corresponding abstraction and formalization of justice, we can imagine that supranational and global systemic integration can still be complemented by social integration.

See also: Anomie; Communities of Practice; Community Sociology; Conflict Sociology; Conformity: Sociological Aspects; Control: Social; Cultural Assimilation; Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Diffusion: Anthropological Aspects; Diffusion: Geographical Aspects; Diffusion, Sociology of; Diversity and Disagreement; Group Decision Making, Social Psychology of; Groups, Sociology of; Identity: Social; Norms; Social Stratification; Structure: Social; System: Social; Values, Sociology of

Bibliography

Alexander J C 1998 *Neofunctionalism and After*. Blackwell, Oxford

Axelrod R 1984 *The Evolution of Cooperation*. Basic Books, New York

Berger P L, Luckmann T 1966 *The Social Construction of Reality*, 1st edn. Penguin, Harmondsworth

Blumer H 1969 *Symbolic Interactionism*. Prentice Hall, Englewood Cliffs, NJ

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press, of Harvard University Press, Cambridge, MA

Collins R 1975 *Conflict Sociology: Toward an Explanatory Science*. Academic Press, New York

Coser L A 1956 *The Functions of Social Conflict*. Free Press, Glencoe, IL

Dahrendorf R 1958 Toward a Theory of Social Conflict. *Journal of Conflict Resolution* **2**: 170–83

Durkheim E 1964 *The Division of Labor in Society*. [trans. G Simpson]. Free Press, New York

Durkheim E 1982 *The Rules of Sociological Method*. [Lukes S (ed.)]. MacMillan, London

Elster J 1989 *The Cement of Society. A Study of Social Order*. Cambridge University Press, New York

Garfinkel H 1967 *Studies in Ethnomethodology*. Prentice Hall, Englewood Cliffs, NJ

Habermas J 1984, 1989 *The Theory of Communicative Action*, 2 vols. [trans. McCarthy T]. Beacon Press, Boston

Hechter M 1975 *Internal Colonialism. The Celtic Fringe in British National Development, 1536–1966*. Routledge and Kegan Paul, London

Hechter M 1987 *Principles of Group Solidarity*. University of California Press, Berkeley, CA

Hegel G F W 1965 *Philosophy of Right*. Clarendon Press, Oxford

Hobbes T 1966 Leviathan. In: Molesworth W (ed.) *Collected English Works of Thomas Hobbes*. Scientia, Aalen, Germany, Vol. 3

Homans G C 1974 *Social Behavior: Its Elementary Forms*, rev edn. Harcourt Brace Jovanovich, New York

Hume D 1980 *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*. Selby-Bigge L A (ed.). Greenwood Press, Westport, CT

Joas H 1992 *Die Kreativität des Handelns*. Suhrkamp, Frankfurt am Main, Germany

Kant I 1956 Die Metaphysik der Sitten. In: Weischedel W (ed.) *Werke in sechs Bänden*. Insel Verlag, Frankfurt am Main, Germany, Vol 4

Levine D 1980 *Simmel and Parsons. Two Approaches to the Study of Society*. Arno Press, New York

Lockwood D 1964 Social integration and systems integration. In: Zollschan G K, Hirsch W (eds.) *Explorations in Social Change*. Routledge, London, pp. 244–57

Luhmann N 1988 *Die Wirtschaft der Gesellschaft*. Suhrkamp, Frankfurt am Main, Germany

Luhmann N 1997 *Die Gesellschaft der Gesellschaft*, 2 vols. Suhrkamp, Frankfurt am Main, Germany

Marshall T H 1964 *Class, Citizenship and Social Development*. Greenwood Press, Westport, CT

Marx K 1962 *Das Kapital*, vol 1. In: *Marx-Engels-Werke*. Dietz, Berlin

Mead G H 1934 *Mind, Self and Society*. [Morris C W (ed.)]. University of Chicago Press, Chicago

Münch R 1993 *Das Project Europa*. Suhrkamp, Frankfurt, Germany

Münch R 1998 *Globale Dynamik, lokale Lebenswelten*. Suhrkamp, Frankfurt, Germany
Parsons T [1937] 1968 *The Structure of Social Action*. Free Press, New York
Parsons T 1971 *The System of Modern Societies*. Prentice Hall, Englewood Cliffs, NJ
Rousseau J J [1762] 1973 *The Social Contract and Discourses*. [trans. Cole G D H]. Dent, London
Schütz A 1964 *Collected Papers II: Studies in Social Theory*. Nijhoff, The Hague, The Netherlands
Simmel G 1992 *Soziologie. Untersuchungen über die Formen der Vergesellschaftung*. Suhrkamp, Frankfurt am Main, Germany
Smith A [1759] 1966 *The Theory of Moral Sentiments*. A. M. Kelley, New York
Smith A 1937 *The Wealth of Nations*. Modern Library, New York
Spencer H 1972 *On Social Evolution. Selected Writings*. Peel J D Y (ed.). University of Chicago Press, Chicago
Strauss A 1978 *Negotiations. Varieties, Contexts, Processes and Social Order*, 1st edn. Jossey-Bass, San Francisco
Weber M 1976 *Wirtschaft und Gesellschaft*. Mohr Siebeck, Tübingen, Germany
Weber M 1923 *Wirtschaftsgeschichte*. Hellman S, Palyi M (eds.). Duncker & Humblot, Munich, Germany

R. Münch

Integrative Assessment in Environmental Studies

In this contribution we will introduce the new research area of Integrated Assessment. To this end, three questions will be addressed:

(a) What is Integrated Assessment?
(b) Why is Integrated Assessment necessary?
(c) How can Integrated Assessment be applied?

Finally, some recommendations are made for the future development of the emerging research area of Integrated Assessment.

1. *What is Integrated Assessment?*

The term Integrated Assessment (IA) was used for the first time in the early 1970s, in particular in the USA and Europe. In Europe, IA has its roots in population–environment research, developing into integrated environmental research, often practiced by doing integrated environmental modeling. During the 1980s and 1990s, many integrated environmental models were developed (Hordijk 1991, Rotmans 1990, Rotmans and de Vries 1997). In the USA, IA was rooted in economic cost–benefit analyses of environmental problems. The impact of supersonic transport on stratospheric ozone and ground level ultraviolet levels increased the awareness that human activities could disrupt the global environment.

Subsequently, there has been increasing recognition and credibility for the rapidly evolving field of IA, illustrated, for example, by the establishment of the European Forum for Integrated Environmental Assessment (EFIEA) (Tol and Vellinga 1998) under the auspices of the European Commission. Multiple definitions of IA circulate (Weyant et al. 1996, Rotmans and Dowlatabadi 1998, Parson 1996), but these definitions have three elements in common: multi- or interdisciplinarity, structuring of knowledge, and decision-support. Using these commonalities, the following definition can be given:

IA is a multi- or interdisciplinary process of structuring knowledge elements from various scientific disciplines in such a manner that all relevant aspects of a social problem are considered in their mutual coherence for the benefit of decision-making.

Metaphorically IA can be viewed as making a puzzle. Everybody sees the separate pieces of the puzzle, but the real art is fitting them together in such a manner that a logical whole arises, which is more than the sum of its parts. The essence of IA is, however, that there is no optimal solution of fitting the scientific pieces of the puzzle together. Depending on the underlying perspective, a multitude of possible solutions exists, and there is no standard-recipe available on how to arrive at this set of possible solutions. This issue explains the heuristic character of IA, being a quest in triplicate: for causalities, coherence, and commonalities.

Although active participation is no prerequisite, more and more people in the IA community are convinced of the vital importance of involvement of stakeholders in the IA process. It is generally acknowledged that involvement of nonscientific and practical knowledge and expertise, valuation and preferences in the form of direct involvement of actors will enrich the process of IA.

This process can be represented in a simplified manner through a sort of demand/supply mechanism, as depicted in Fig. 1. From society's angle there is a need for information about complex social issues, the demand side. Science delivers facts, uncertainties and hypotheses with regard to these complex issues, the supply side. Demand and supply mechanisms come together in an active dialogue, which ultimately leads to integrated visions. In consultation with decision-makers these visions can be translated into decision-strategies. However, the supply side could also be delivered by social actors in the form of nonscientific expertise and judgment, whereas the demand side could be provided by scientists, for instance through prioritizing of research themes.

One of the more successful integrated assessments achieved to date is that of global climate change. Within the framework of the Intergovernmental Panel on Climate Change (IPCC, 1996), the integrated approach has been successful in both scientific and policy terms. In science, it has proven useful in

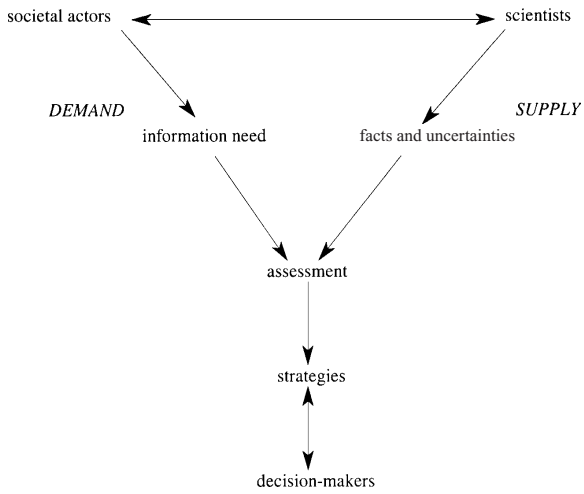

Figure 1
Sketch of an IA process

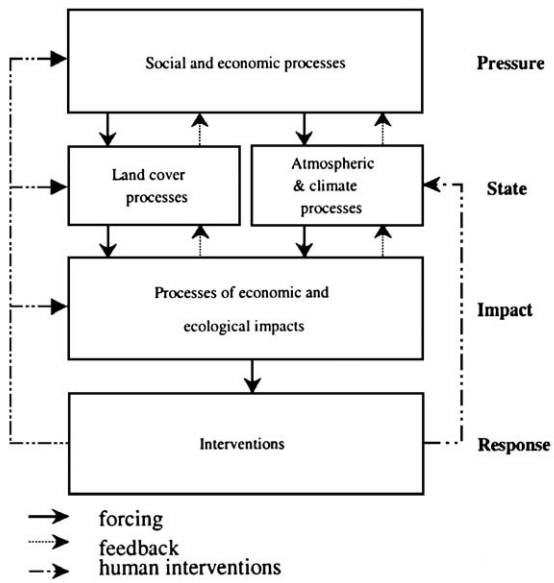

Figure 2
A simple pressure-state–impact-response diagram of the human–climate system

addressing earth system functions and impacts, for example, the imbalance in the global carbon budget, integrated analysis of land use and land cover changes, and the climate impact on sulphate aerosols (Rotmans and Dowlatabadi 1998). In policy terms, the attempt was to underpin the discussion about the question: when do we have to undertake action and with what pace do we have to take what kind of measures?

1.1 Climate Change

An IA of climate change takes into account both the socioeconomic and biogeochemical aspects of the climate problem. The most important cause of global climate change is the large-scale increase in the demand for energy and land/food. The driving forces behind this are economic growth, population growth and changes in life-style. This all leads to changes in the composition of the atmosphere and terrestrial biosphere, resulting in changes in global and regional climate. Major social impacts are the effects on the economic sectors such as agriculture, water management, coastal defense, health care, and on natural ecosystems in terms of land degradation, erosion, decreasing vitality, and change in spatial composition.

An IA model of climate change aims to simulate the intricate dynamics of the human-climate system, in which numerous interactions and feedbacks occur, as shown in Fig. 2. Also, such a model allows for rapid calculation and evaluation of various preventive (e.g., an energy tax regime) or adaptive (e.g., infrastructural adaptations) measures.

2. Why IA?

The world around us is becoming increasingly integrated in its commercial and financial activities, while modern means of travel and communication are breaking down many of the traditional physical and cultural barriers between states. The complex dynamics of strongly interacting short- and long-term processes on various scale levels force us to think and act in a more integrative manner. This increasing complexity means that major social problems can no longer be addressed from one perspective, one country, one state, one culture, one ministry, or one scientific discipline, requiring a new conception of planning. IA tries to provide the new tools for this new way of planning, by getting more grip on the increasing complexity of society: considering problems in the light of other problems, coupling different scale levels in time and space, interrelating different themes, addressing cross-sectoral trends and developments, among others. This integration follows from the growing understanding that the various pieces of the puzzle can no longer be examined in isolation.

Opponents may argue that IA is premature, and can lead to inappropriate confidence in questionable or misleading findings. However, the complexity of the issues demands an integrated approach to ensure that key interactions, feedbacks and effects are not inadvertently omitted from the analysis, leading to surprise consequences. Another argument against IA may be that we have insufficient knowledge of the interactions and level of coherence among the social-

cultural, economic and environmental subsystems. This line of reasoning fails because the uncertainties within the subsystems are also large, and our complex society asks for studies that focus on the interface between the subsystems. Current knowledge indicates that sharp fluctuations in the dynamics of our society can be attributed to changes occurring at the cutting edge of the social subsystems, rather than within the subsystems themselves.

So overall, performing IA has a number of advantages. In general terms, IA can help to:

(a) put a complex problem in the broader context of other problems, by exploring the interrelations of the specific problem with other issues;

(b) assess potential response options to complex problems, be it in the form of cost–benefit and cost-effectiveness or some other analysis;

(c) identify, illuminate, and clarify the different types and sources of uncertainties in the cause–effect chain(s) of a complex problem;

(d) translate the concept of uncertainty into the concept of risk analysis, to assist in decision making under uncertainty; and

(e) set priorities for research topics, also by identifying and prioritizing decision-relevant gaps in knowledge.

3. *What is the Problem?*

IA deals with causally linking those social-cultural, economic and ecological processes that are relevant in the light of the problem under concern. This means integration of physical, monetary, information and policy processes. These processes are unlike, with regard to time scale, spatial scale and dynamics, e.g., many economic processes operate at a relatively short time scale, determined by the financial capital invested. Demographic processes play at a longer term of at least one generation, whereas the effect of many ecological processes becomes visible at a time scale of several generations, if not centuries. IA, therefore, seeks to couple: (a) long-term processes with short-term processes; (b) higher scale processes with lower scale processes; and (c) equilibrium processes with nonequilibrium processes.

The process of up- and down-scaling between different levels in time, space and complexity is called aggregation. The level of aggregation of an IA refers to the spatial and temporal resolution and the level of complexity used in the assessment. In general, the optimal spatial and temporal scale at which an IA operates strongly depends on the nature and scope of the problem.

The real problem of IA is that there is no unifying theory that indicates unequivocally how this analytical process of aggregation should happen. Also, the price of integration is simplification, in order to integrate disciplinary pieces into an overall framework. As a consequence, a manifold of integration routes could lead to an integrated knowledge framework. This explains why IA has been a practical, intuitive and heuristic activity so far. It also underlines the need for procedures and guidelines for testing the quality of the integration process.

4. *Methods for IA*

The methodological approaches used at this time to do IA are relatively immature. In general, two types of methods can be distinguished: analytical methods and participatory methods. While analytical methods are often rooted in natural sciences, participatory methods stem from social sciences. The group of analytical methods is reasonably well defined and basically includes model analysis, scenario analysis, and risk analysis. Their commonality is that they provide analytical frameworks for representing and structuring scientific knowledge in an integrated manner. A plethora of participatory methods, however, exist, varying from expert panels and delphi methods, to gaming, policy exercises and focus groups. They have in common the aim to involve nonscientists as stakeholders in the process, where the assessment effort is driven by stakeholder–scientist interactions. Here, we will only briefly discuss IA models and IA participatory approaches. For a comprehensive and elaborative description of IA methods, refer to Rotmans (1998).

4.1 *IA Models*

IA models are computer simulation frameworks that try to describe quantitatively, as much as possible, the cause and effect relationships of a specific issue and of the inter-linkages and interactions among different issues. Current projects in IA modeling build on a tradition started in the early 1970s, by the Club of Rome (Meadows et al. 1972).

The next generation of IA models explicitly addressed environmental issues, such as acidification (Hordijk 1991) and climate change (Rotmans 1990, Nordhaus 1992). Recent overviews of IA modeling activities in the field of climate change can be found in Weyant et al. (1996) and Rotmans and Dowlatabadi (1998). The latter distinguish between macroeconomic-oriented models, which represent relatively simple, parameterized decision-analytic formulations of complex problems, and biosphere-oriented models, which represent a more comprehensive, process-oriented description of a complex problem. Most macroeconomic-oriented models are neoclassical models based on an equilibrium framework, using traditional economic concepts regarding optimization and capital accumulation, largely ignoring environmental dynamics.

Biosphere-oriented models, however, focus on a systems-based description of the geophysical and biogeochemical processes and feedbacks, but do not adequately represent the socioeconomic system. The Dynamic Integrated Climate Economy (DICE) model is a well-known exponent of the macroeconomic-oriented school, whereas the Integrated Model to Assess the Greenhouse Effect (IMAGE) model (Rotmans 1990, Alcamo 1994) is representative for the biosphere-oriented school. Meanwhile, some attempts are underway to combine the best of both worlds, yielding a hybrid of the two categories above. Examples of such hybrid IA models are ICAM (Dowlatabadi and Morgan 1993), Global Change Assessment Model (GCAM) (Edmonds et al. 1994), and TARGETS (Rotmans and de Vries 1997).

IA models have the advantage that they are flexible and rapid simulation tools, which can easily explore interactions, feedback mechanisms, and uncertainties. Above all, they are tools to communicate complex scientific issues to decision makers, disciplinary scientists, stakeholders, and the general public.

4.2 Participatory Methods for IA

'Participatory methods' is an umbrella term describing approaches for assessment in which nonscientists, such as policy people, stakeholders or even lay people, play an active role. Participatory methods for IA differ with regard to the profile of the participants, the goal of participation and the degree of participation. Three classes of methods exist: dialogue-methods, policy exercises, and mutual learning methods. The dialogue-method is applied in cases where the intended users are considered as a source of information necessary for the analysts to perform the assessment. Within this class, we can discriminate between continuous dialogues, and cases in which the dialogue takes place in a specific phase of the assessment. In the first case, the role of the intended users can be described as that of co-designer. In case the dialogue takes place in the design phase, the role of the intended users is to contribute to the design principles by sketching their wants and needs. This type of dialogue can be described as a user-platform.

Policy exercises build upon the tradition of simulation games. A policy exercise can be described as a flexibly structured process designed as an interface between scientists and policy makers, in which a complex system is represented by a simpler one with relevant behavioral similarity, and from which decision making is part through human participants (Parson and Fisher-Vanden 1997). In general, a game is set up that represents a negotiation process in which the different teams are responsible for a certain country or region. Policy exercises quite often make use of computer support. A computer model can be used as a consulting device or as tool to convert the negotiated agreements into a new 'state of the world.'

The principle behind mutual learning approaches is that participation of stakeholders and citizens enriches the assessment by a multiplicity of perspectives, skills, and competence. The participants are considered as co-producers of knowledge. Two forms of mutual learning can be distingished: the focus group approach, in which scientists play the role of facilitator and observer, and the interactive approach, in which scientists are actively involved as participants. In the first case, special groups are composed typically consisting of citizens, policy makers or other stakeholders who are provided with scientific input. The assessments of these groups are then used in a broader IA process. The ULYSSES project involves IA focus groups with lay people (Kasemir et al. 1997).

A major weakness pertaining to participatory methods is that the use of it in IA is in its infancy, with the result that there are no (as yet) established procedures and work-packages for setting up participatory processes.

5. Quo Vadis *IA?*

In order to meet the high expectations that many people have of IA , a number of steps need to be taken. First, the existing tools and instruments for IA should be improved. Much of the criticism on IA has to do with the opaque structure of the IA methods used, and the fact that these methods are often technocratic abstractions of reality.

Second, new IA research methods have to be developed, which enable qualitative knowledge and quantitative data to be blended; but also methods that could handle different sorts and types of uncertainty; and finally methods that could operate at different aggregation levels in time, space and complexity. Regarding IA models, much attention has to be paid to actor-oriented IA models that describe the behavior of human or institutional actors in time and space by means of a set of dynamic behavior rules. Also, it is important to experiment as much as possible with different participatory methods in an IA context. In this way guidelines and procedures could be developed for the development and application of IA participatory methods.

Third, demand and supply of IA studies should be matched in a better way, as also propagated by Funtowicz and Ravetz (1994). At present, the majority of IA research is supply driven. In order to increase the number of demand-driven IA studies, the best of both analytical and participatory worlds should be combined. In this way the complementarity of both types of IA method could be used. The challenge then is to let nonscientists or stakeholders co-develop analytical IA methods.

Fourth, there is a need for IA-quality criteria to test the quality of IA research (Morgan and Dowlatabadi 1996). We distinguish here three types of quality criteria: analytical criteria, referring to calibration and validation of the used IA research methods, methodological criteria, which concern the quality of the used IA research method, and usability criteria, which refer to the extent to which the IA study is useful for policy. However, because of the rather heuristic and explorative character of many IA studies, these quality criteria should not be measured too stringently.

Finally, the time seems ripe for entering new problem areas from the perspective of IA. So far, it has been applied to only a few environmental problems, of which acidification and climate change are the most important. A real challenge for IA is now to broaden its field of application, such as transport and infrastructure issues, health care problems, technological development, the energy issue, and city planning (Rotmans 1998).

See also: Climate Change, Economics of; Environmental and Resource Management; Environmental Economics; Environmental Planning; Environmental Policy; Environmental Surprise

Bibliography

Alcamo J 1994 IMAGE 2.0: Integrated modelling of global climate change. *Water, Air, and Soil Pollution* **76**: 1–2

Dowlatabadi H, Morgan M G 1993 Integrated assessment of climate change. *Science* **259**: 1813–4

Edmonds J, Wise M, et al. 1994 *Advanced Energy Technologies and Climate Change: An Analysis Using the Global Change Assessment Model* (*GCAM*). Air and Waste Management Meeting, Pittsburgh, PA

Funtowicz S O, Ravetz J R 1994 The worth of a songbird: Ecological economics as a post-normal science. *Ecological Economics* **10**: 197–207

Hordijk L 1991 *An Integrated Assessment Model for Acidification in Europe*. Free University of Amsterdam, The Netherlands

IPCC 1996 *Climate Change 1995: The Science of Climate Change*. Cambridge University Press, Cambridge, UK

Kasemir B, Behringer J, et al. 1997 *Focus Groups in Integrated Assessment*: *The ULYSSES Pilot Experience*. Darmstadt University of Technology, Darmstadt, Germany

Meadows D H, Meadows D L et al. 1972 *The Limits to Growth*. Universe Books, New York

Morgan M G, Dowlatabadi H 1996 Learning from integrated assessment of climate change. *Climatic Change* **34**: 337–68

Nordhaus W D 1992 *The DICE model: Background and Structure of a Dynamic Integrated Climate Economy*. Yale University, New Haven, CT

Parson E A, Fisher-Vanden K 1997 Integrated assessment of global climate change. *Annual Review of Energy and the Environment* **22**

Rotmans J 1990 *IMAGE: An Integrated Model to Assess the Greenhouse Effect*. Kluwer Academic, Dordrecht, The Netherlands

Rotmans J 1998 Methods for IA: The challenges and opportunities ahead. In: Rotmans J, Vellinga P (eds.) *Environmental Modelling and Assessment*, Special issue, pp. 155–79

Rotmans J, de Vries H J M (eds.) 1997 *Perspectives on Global Change: The TARGETS Approach*. Cambridge University Press, Cambridge, UK

Rotmans J, Dowlatabadi H 1998 Integrated assessment of climate change: Evaluation of methods and strategies. In: Rayner S, Malone E (eds.) *Human Choice and Climate Change: An International Social Science Assessment*. Battelle Press, Washington, DC

Tol R S J, Vellinga P 1998 The European forum on integrated environmental assessment. In Rotmans J, Vellinga P (eds.) *Environmental Modelling and Assessment*, Special issue, pp. 181–91

Weyant J, Davidson O et al. 1996 Integrated assessment of climate change: An overview and comparison of approaches and results. In: Bruce J P, Lee H, Haites E F (eds.) *Economic and Social Dimensions of Climate Change*. Cambridge University Press, IPCC, Cambridge, UK

J. Rotmans

Intellectual Functioning, Assessment of

Oversimplifying somewhat, there are three applied approaches to assessing intelligence (Daniel 1997). Each rests on a different conceptualization of what intelligence is and why it should be measured. Each approach, the psychometric, the neuropsychological, and the dynamic, is reviewed below. In addition, several physiological indices—brain wave tracings, reaction and inspection times, nerve conduction velocity, and rate at which glucose is metabolized in the brain, for example—correlate with measures of intelligence. However, physiological indices are not reviewed because they have no applied utility. It is possible, though, that physiological indices will play a stronger assessment role as the twenty-first century progresses (Matarazzo 1992).

1. Psychometric Assessment

Psychometric tests have dominated intelligence testing for a century. The defining feature of this approach is its empirical foundation; 'psychometric' simply refers to the quantitative assessment of psychological states/abilities. While quantitative assessment rests on a massive measurement technology, its theoretical foundations are shallow, as reflected in its origins. The earliest tests that influenced contemporary intellectual measures directly emerged from studies by Alfred Binet and colleagues in France (Cronbach 1984). In 1904, Binet was directed to devise a means of distinguishing educable from noneducable students in the relatively new universal education system. Having

investigated cranial, facial, palmar, and handwriting indices, Binet discovered the direct measure of complex intellectual tasks involving judgment, comprehension, and reasoning most successful in distinguishing among pupils. Based on these pragmatic beginnings, Binet defined intelligence as the capacity to adopt and sustain a direction, make adaptations for the purpose of attaining a desired end, and monitor performance self-correctively. With little elaboration, this definition still directs the psychometric paradigm.

Typically, modern psychometric tests consist of varied subtests that tap diverse aspects of the loosely defined intelligence construct. For example, scales may include subtests that sample a broad range of knowledge (e.g., the names of objects, dates, historical and geographical facts) and require the examinee to assemble colored blocks such that their pattern resembles a prespecified design (Sattler 1992). Again, choice of subtests is not driven by theoretical prescription. Subtests are selected because they work—in combination, they serve to rank individuals according to how much they know and how good they are at solving certain problems. The pragmatic selection of subtests is based on Binet's conception of intelligence as a general or undifferentiated ability (*g*), so that, in principle, the tasks that tap *g* are interchangeable.

At the heart of psychometric testing lies norm referencing (Sattler 1992). Norm referenced tests are developed by administering items in a standardized manner to a representative sample of the population in question. The norm sample is considered 'representative' insofar as it is stratified within age groups for variables that might influence performance differentially, such as sex, geographic region, ethnic status, community size, etc. Scores are scaled such that each individual's derived score represents a relative standing within the norm or standardization group. In this sense, psychometric testing is an empirical endeavor in its purest sense: as a comparative construct, there is little need to theorize about the exact nature of intelligence.

As mentioned, most modern psychometric tests include varied tasks. The original intention was to ensure that *g* was comprehensively surveyed. With time, however, clinicians came to exploit the multitask construction of intelligence tests to make intra-individual distinctions (Kaufman 1990). By looking at the variability among subtests or groups of subtests, assessors hypothesized about relative intellectual strengths and weaknesses. For example, a particular respondent might prove better on tests of memory than on tasks involving conceptualization. It is important to note, though, that the analysis of intra-individual differences developed after the fact; such comparisons are driven by the practicalities of what subtests are available, rather than by a detailed theory about the structure of intelligence.

The empirical base of the psychometric effort implies both weakness and strength. With respect to its limitations, attempts to interpret intra-individual differences based on a selection of subtests that were pragmatically chosen have not been validated empirically (Reschly 1997). Furthermore, the atheoretical approach to task selection has resulted in restricted and incomplete sampling of the intelligence domain (Chen and Gardner 1997). For example, musical and interpersonal abilities are neglected. Instead, there is an emphasis on skills acquired though academic learning, a prized outcome in mainstream Western societies. Therefore, critics object to the fact that psychometric tests measure little more than achievement; they assess what an examinee has learned, not the examinee's potential to learn.

Related to this issue, and magnified by the practice of defining individual intelligence with reference to a norm group, questions have arisen about bias due to (sub)cultural, ethnic, life-experience, and motivational differences. This becomes a social issue when examinees from minority groups are compared to a norm sample whose context, values, and learning experiences are different from their own (Suzuki and Valencia 1997). Testing thereby betrays its original purpose of providing objective data on an individual's intellectual functioning and comes, instead, to discriminate against atypical examinees.

Another difficulty with psychometric tests is that although they usually correlate highly among themselves, this is not always the case (Daniel 1997). Correlations may be influenced by what tasks are included and how they are weighted. Perhaps a greater problem lies in the fact that even where test scores do correlate highly, the same individual may earn discrepant scores on different instruments due to the fact that tests are normed on different standardization groups.

A crucial criticism of psychometric tests is that recommendations derived from these instruments have not been shown to enhance remediation for the examinees (Reschly 1997). Again, this can be attributed to the fact that the content of these scales has not been selected according to any theory of intelligence, brain functioning, or pedagogy.

In other respects, psychometric testing has met with success. Although test tasks are selected pragmatically, they cluster in remarkably similar ways across tests and studies, giving insight into the structure of intelligence. Based on statistical methods that group subtests into clusters according to underlying commonalities (factor analysis), three strata of intelligence have been identified (Carrol 1997). At the highest stratum is a general factor, *g*. This factor subsumes a second stratum of broad factors, including 'fluid' and 'crystallized' intelligence. (Fluid intelligence involves the ability to cope with novelty and think flexibly. Crystallized intelligence involves the storage and use of declarative knowledge such as vocabulary or information.) Subsumed under each broad factor is a set of narrow abilities, such as 'induction' and 'reading

comprehension.' Knowledge of these distinct but interdependent strata can guide construction of new psychometric instrumentation.

Another strength of the psychometric approach derives from its emphasis on quantitative methods; psychometricians strive to ensure that their tests are reliable and valid predictors of performance (Sattler 1992). 'Reliability' refers to consistency of measurement; the more reliable a measure, the less error involved in estimates derived from it. Many psychometric tests boast extremely high internal reliability (the degree to which each component score of the test correlates with the full test score) and short-term 'test-retest' reliability (an index of stability derived by administering the test to the same group of individuals more than once). Furthermore, the long-term stability of IQ has proven impressive, with good predictions over a 20-year time-span. The validity of these tests, too, has proven strong. 'Validity' refers to the extent to which a test measures what it was designed to measure. Intelligence test scores correlate with amount of schooling, quality of work produced in school, occupational status, and performance in the work situation (although the strength of the latter prediction is controversial), both concurrently and predictively. To summarize, although there are serious limitations to psychometric measurement, the approach yields reliable and valid estimates of intellectual functioning. Psychometric tests are accurate classifiers and predictors when used with care in circumscribed contexts.

2. *Neuropsychologically-based Assessment*

In reaction to the pragmatics of psychometric assessment, several theory-based tests have been developed. These are exemplified by the neuropsychological approach to intellectual assessment. A viable neuropsychological approach emerged in the 1980s. This approach reflects an attempt to increase the validity of assessment by basing it on an understanding of brain function and brain–behavior relationships. The emphasis is not on what the examinee knows but on how the examinee reasons; not on observable abilities but on underlying capacities. Empirical considerations such as reliability, validity, and norm referencing are stressed, but they subserve theoretical relevance (Kaufman 2000).

Neuropsychological approaches to intelligence testing have been influenced by the theories of Alexandr Luria, a Russian psychologist. Luria (1980) proposed three levels of brain functioning that mediate distinct types of learning and performance. The lowest functional unit includes attention. Attention involves selectively attending to particular stimuli while inhibiting attention to competing stimuli over time. It is prerequisite to all forms of intelligent behavior. Attention can be assessed with tasks that demand sustained concentration on one stimulus and sustained inhibition of responses to distracting stimuli. The second functional unit involves encoding and processing information. According to Luria, information can be processed simultaneously and/or successively. Simultaneous processing is holistic; it involves organizing stimuli into interrelated groups. Simultaneous processing can be assessed with tasks that require an examinee to recall the spatial locations of stimuli, identify objects based on incomplete drawings, and construct abstract designs from several identical triangles. Successive processing involves integrating stimuli serially, such that each stimulus is related to the one that precedes it in chain-like succession. Successive processing can be tested with tasks requiring, for example, that the respondent repeat numbers spoken by the examiner or repeat an ordered series of hand movements made by the examiner. Planning and monitoring (executive) functions form the third and highest cognitive level, according to Luria. This level involves the integration of attention and simultaneous and successive processing, along with acquired knowledge, so as to permit the formation and execution of plans and the ongoing verification of their efficacy. Neuropsychological tests of intelligence include subtests that assess some or all of these processes. In addition, they may include subtests that are derived from research on information processing, cerebral specialization, and cognitive psychology, and subtests that reflect other theoretical orientations—the distinction between fluid and crystallized forms of intelligence, for example. Scores are derived to reflect ability in each function assessed, as well as overall functioning. These scores can be compared to derive a profile of strengths and weaknesses which, in turn, has direct educational implications.

The neuropsychological approach to intelligence testing has limitations and strengths. A central shortcoming of these theory-based tests lies in the controversy regarding whether they actually measure the theoretical constructs they purport to measure. Factor analyses indicate that subtests do not always cluster so as to confirm the neuropsychological theory on which they are based (Kranzler and Keith 1999). Furthermore, even when the subtests do cluster as expected, their underlying meaning can be ambiguous. For example, the factors that supposedly reflect sequential and simultaneous reasoning in one neuropsychological intelligence test might just as easily reflect verbal and nonverbal reasoning skills. Resolution of issues pertaining to 'construct validity' (whether a test measures what it purports to measure) is crucial; the controversy pertains to whether the theory-based tests in question are actually theory-based.

What of test bias, as discussed in the context of psychometric testing? Bias is not precluded by theoretical considerations pertinent to the neuropsychological approach. Furthermore, the neuropsychological approach relies on norming procedures, such

that the performance of a given examinee is compared to the performance of a standardization sample. While attempts have been made to reduce the discrepancies between whites and minority groups (e.g., by reducing verbal requirements, limiting/changing measures of planning, disproportionately selecting high socioeconomic representatives of minority groups), these strategies work by sacrificing breadth of intellectual sampling and minimizing the differences between whites and minority groups in the normative group artifactually. The issue of bias remains problematic in neuropsychological intelligence testing.

It has been argued that the theoretical underpinnings of neuropsychological tests of intelligence both restrain and augment the breadth of subtest selection (Das et al. 1994). Which perspective one takes depends on the neuropsychological test one is considering and what one considers the purpose of testing to be. It is true that such tests neglect areas of intelligence such as creativity and wit, but the purpose of these new instruments is primarily to aid in pedagogic planning. Hence, it is the perspective these tests afford on a given examinee's strengths and weaknesses with respect to learning that is important. If these tests can successfully divilicate the examinee's ability to plan, attend, and reason successively and simultaneously, such that a successful, deficit-related teaching approach can be developed, the breadth of neuropsychological intelligence testing will be justified. This issue has been debated, however, with some investigators arguing that the adoption of a neuropsychological approach to intelligence testing cannot or has not led to improved pedagogical success (Kranzler and Keith 1999). Others have demonstrated specific gains using carefully controlled trials (Naglieri 1999). For example, students with poor planning improved dramatically in math calculation when instructed in the use of strategies; other students who were good in planning showed more modest improvements when instructed in the use of strategies. At this point, the relative novelty of the neuropsychological approach precludes conclusive comment about treatment validity. If these instruments prove successful in directing effective pedagogy, they will have addressed the major shortcoming of psychometric assessment without sacrificing its empirical strengths.

3. *Dynamic Assessment*

Incorporating a heterogeneous set of evaluation procedures, the defining feature of dynamic assessment is that it evaluates and modifies cognitive functioning simultaneously through mediation and intervention. The product of a dynamic assessment is not an estimate of the examinee's intellectual standing relative to their peers, but an estimate of learning potential. The notion behind the learning potential construct is that individuals with similar starting competencies (e.g., similar IQs) may respond differentially to instruction; what an individual has learned does not necessarily reflect the individual's potential to learn (Grigorenko and Sternberg 1998).

Dynamic assessment represents a reaction against psychometric testing in particular (Haywood et al. 1992), with its purported: over-reliance on previously learned material; failure to consider factors other than ability in measuring IQ, e.g., the (sub)cultural, motivational, personality, and social adequacy differences in examinees; and irrelevance to remediation. The theoretical writings of Lev Vygotsky, a Russian psychologist, form the context of dynamic assessment. Vygotsky (1987) argued that what a child can accomplish when assisted by a mediator ('zone of proximal development') is more indicative of the child's mental development than what the child can do alone ('zone of actual development'). The zone of proximal development is operationalized in dynamic assessment by a collaborative examiner. This examiner provides ongoing feedback, titrated such that it is just beyond the examinee's zone of actual development, until the examinee either solves the problem or gives up. Thus, testing and teaching are joined. This is facilitated by emphasizing tasks that stress fluid, rather than crystallized, reasoning skills. The Vygotskian basis of dynamic assessment is supplemented by information processing concepts in some dynamic approaches. For example, one approach stresses the central role of working memory, the ability to hold old and new information in memory simultaneously, so as to manipulate and transform it. Dynamic instrumentation often includes standardized psychometric tasks, as described above, but they are administered interactively.

Dynamic approaches vary along two inversely related continua: individuality of assessment and psychometric adequacy. At one end of these continua, the assessor works in intensive, individualized interaction, modifying tasks freely, encouraging, prompting, and mediating between child and task at will to derive qualitative (as opposed to quantitative) diagnoses and educational prescriptions (Feuerstein et al. 1979). One commentator described the approach as 'a clinician's dream and a psychometrician's nightmare' (Lidz 1997). At the other end of the continua lies a test-train-retest sequence with standardized mediation and outcome measures. Mediation is delivered through a series of predetermined hints that vary from general to specific. Each hint is offered in response to the youngster's difficulty. Outcome is quantified as number of prompts required to bring the child to correct response and as the ability to transfer to tasks other than those on which training first occurred (Campione and Brown 1987). Some dynamic practitioners doubt the efficacy of standardized dynamic procedures, because clues are not individually modified to the child's needs. However, the method's

standard, quantifiable format renders it amenable to psychometric study.

The goals of dynamic assessment—to assess learning ability, to identify learning processes, and to teach the individual generalizable strategies for problem solving—remain its major strength. No other assessment tradition is so ambitious. Whether or not they succeed in this endeavor, dynamic assessment advocates have sharpened the focus of other assessment approaches, alerting them to the ultimate goal of intellectual evaluation. And in selecting response to remedial efforts as an outcome (as opposed to comparing an examinee with his or her age mates), dynamic assessors propose a powerful means of attenuating experiential bias in intelligence testing. The reasoning is as follows. Suppose there are two children, one biologically disadvantaged with mental retardation, the other who has experienced atypical living/educational opportunities. One might expect relatively little success in teaching the child with mental retardation. However, the educationally atypical child should make great gains, i.e., there should be a large discrepancy between this youngster's pretest and post-test scores (or the child should require relatively few clues to reach problem solution), whether or not the post-test scores approximate the average test score of this child's educationally advantaged age peers. The point is that pretest scores, conceptually comparable to conventional test results, tap a different aspect of intelligence than do difference scores; pretest scores correspond to developed abilities (and, implicitly, past opportunities to develop abilities), while difference scores approximate learning potential. It is by tapping into learning potential that dynamic assessments attenuate the biases of traditional assessment.

However, research support for the claims underlying dynamic assessment is sparse, ambiguous, and often flawed. This is more the case for approaches that emphasize nonstandardized assessment than for those that stress psychometric adequacy, but weak research foundations are common to all approaches (Grigorenko and Sternberg 1998). Little validation of the 'zone of proximal development' exists, for example, and much of what does exist has been conducted at the level of theory, ideology, and description, rather than quantitative experiment. And although change is central to the dynamic philosophy, virtually no work addresses the reliability of change scores. Little has been published on internal consistency, and there is some evidence that change scores derived from different tasks do not correlate. Although there may be reasons to expect incongruence of change scores, these reasons need to be explored and verified so that systematic predictions can be made. Questions have also been raised about variability among examiners in administering, scoring, and interpreting the less standardized assessments, and the possibility of obtaining different results for a given child as a result of this. Few studies address the issue of retest reliability of change scores, and some studies suggest that retest reliability is weak. Factor analytic studies, which are used to demonstrate that subtests intercorrelate according to theoretical expectations, are rarely conducted and do not always confirm hypotheses. Studies have shown that mediation promotes change from pre- to post-test, but it is not always clear how much of the change is attributable to mediation and how much change occurs simply by virtue of practice. Importantly, some research shows that dynamic indices augment prediction of post-test performance beyond data derived from pretest performance (Day et al. 1997). However, other studies have not confirmed the independent value of gain indices over traditional measures; results on this aspect of dynamic testing are ambiguous. From a long-term perspective, some evidence suggests a relation between ability to benefit from mediation and economic and social independence in adulthood among individuals with low IQ. However, there have been no demonstrations that dynamic assessment actually changes school or work performance or leads to recommendations that change performance (Grigorenko and Sternberg 1998). In all, dynamic approaches to assessment represent an ambitious but unproven perspective on human intelligence.

4. *Concluding Comments*

The aforementioned approaches to intellectual assessment have shared and unique strengths. With time, it is reasonable to expect an amalgamation of these strengths such that, with the addition of practical biological measures of intelligence, assessment will become increasingly reliable, valid, fair, and useful. Assuming that this synthesis will reflect current theories regarding the hierarchical structure of intelligence, a reasonable expectation is as follows. In the foreseeable future, *g* will be measured with psychometric instrumentation. The exemplary statistical properties of psychometric tests and the diversity of tasks involved in their structure render them most suitable. It is possible that biological measures will augment or replace the psychometric approach to measuring *g*, although biological paradigms remain primitive. Furthermore, because biological assessment focuses exclusively on brain functioning, rather than on brain–behavior relationships, the remedial implications of this approach will remain limited for some time to come. The neuropsychological approach to intellectual assessment can be used to delineate functional areas (e.g., simultaneous vs. successive reasoning) subsumed by *g* and provide broad remedial recommendations accordingly. Finally, the dynamic assessment approach can be used at the level of the individual by assessment-remediators. This final level addresses issues of bias most directly. However, the complexities of genetic-environmental interactions

will always dictate the need for vigilance with respect to bias. The preceding proposal does not suggest that current tests will be imported into a single paradigm holus-bolus. Nor does it suggest that such a synthesis will resolve the theoretical and empirical inconsistencies integral to each approach. The proposal suggests that the *principles* derived from each approach may be incorporated into an integrated assessment/remediation process, such that the shortcomings of each approach are attenuated. An integrated approach to intellectual assessment will permit successive classification, prediction, and remediation that encompass large groups (e.g., people with average *g*) and the smallest groups (i.e., the individual) alike.

See also: Genetic Factors in Cognition/Intelligence; Intelligence: Central Conceptions and Psychometric Models; Intelligence, Evolution of; Intelligence: Historical and Conceptual Perspectives; Intelligence: History of the Concept; Lifespan Theories of Cognitive Development; Neuropsychological Functioning, Assessment of

Bibliography

Campione J C, Brown A L 1987 Linking dynamic testing with school achievement. In: Lidz C S (ed.) *Dynamic Testing*. Guilford, New York, pp. 82–115

Carrol J B 1997 The three-structure theory of cogntive abilities. In: Flanagan D P, Genshart J L, Harrison P L (eds.) *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. Guilford, New York, pp. 122–30

Chen J-Q, Gardner H 1997 Alternative assessment from a multiple intelligences perspective. In: Flanagan D P, Genshart J L, Harrison P L (eds.) *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. Guilford, New York, pp. 105–21

Cronbach C J 1984 *Essentials of Psychological Testing*, 4th edn. Harper and Row, New York

Daniel M H 1997 Intelligence testing: Status and trends. *American Psychologist* **52**: 1038–45

Das J P, Naglieri J A, Kirby J R 1994 *Assessment of Cognitive Processes: The PASS Theory of Intelligence*. Allyn and Bacon, Needham Heights, MA

Day J D, Engelhardt J L, Maxwell S E, Bolig E E 1997 Comparison of static and dynamic assessment procedures and their relation to independent performance. *Journal of Educational Psychology* **89**: 358–68

Feuerstein R, Rand Y, Hoffman M 1979 *The Dynamic Assessment of Retarded Performers: Learning Potential Assessment Device, Theory, Instruments, and Techniques*. University Park Press, Baltimore

Grigorenko E L, Sternberg R J 1998 Dynamic testing. *Psychological Bulletin*. **124**: 75–111

Haywood H C, Tzuriel D, Vaught S 1992 Psychoeducational assessment from a transactional perspective. In: Haywood C, Tzuriel D (eds.) *Interactive Assessment*. Springer-Verlag, New York, pp. 38–63

Kaufman A S 1990 *Assessing Adolescent and Adult Intelligence*. Allyn and Bacon, Toronto, ON, Canada

Kaufman A S 2000 Tests of intelligence. In: Sternberg R J (ed.) *Handbook of intelligence*. Cambridge Univeristy Press, Cambridge, UK, pp. 445–76

Kranzler J H, Keith T Z 1999 Independent confirmatory factor analysis of the Cognitive Assessment System (CAS: What does the CAS measure?). *School Psychology Review* **28**: 117–44

Lidz C S 1997 Dynamic assessment approaches. In: Flanagan D P, Genshart J L, Harrison P L (eds.) *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. Guilford, New York, pp. 281–96

Luria A R 1980 *Higher Cortical Functions in Man*, 2nd edn. Basic Books, New York

Matarazzo J D 1992 Psychological testing and assessment in the 21st Century. *American Psychologist* **47**: 1007–18

Naglieri J A 1999 How valid is the PASS theory and CAS? *School Psychology Review* **28**: 145–62

Reschly D J 1997 Diagnostic and treatment utility of intelligence tests. In: Flanagan D P, Genshart J L, Harrison P L (eds.) *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. Guilford, New York, pp. 437–56

Sattler J M 1992 *Assessment of Children*, 3rd edn. Sattler, San Diego, CA

Suzuki L A, Valencia R R 1997 Race-ethnicity and measured intelligence. *American Psychologist* **52**: 1103–14

Vygotsky L S 1987 *The Collected Works of Lev Vygotsky*. Plenum, New York, Vol. 1

L. Atkinson

Intellectual History

Intellectual history is the increasingly preferred English name for a discipline that was better known, during most of the twentieth century, as the history of ideas. Both names are currently in use, and the duplicity conveys one of the tensions underlying this branch of historical writing: can scholars usefully define what an 'idea' is? Yet the tension remains largely within the contours of Anglophone scholarship. Other languages have remained loyal to 'ideas' as the basic analytical unit; such terms as *Ideengeschichte*, *la storia delle idee*, *dejiny ideí*, and *l'histoire des idées*, have therefore remained unchallenged.

Intellectual history does not deal with all records of human thinking. The history of ideas has never been concerned with every idea crossing human minds. Rather, it has focused on philosophical ideas discussed by mostly well-known thinkers, primarily belonging to the Western tradition (such as Plato, Locke, and Marx), with occasional glances at non-European thinkers (such as Averroes and Maimonides). Its typical primary sources have been the 'great texts' of the European canon (from Homer and the Bible to Kant's *Critique of Pure Reason* and Mill's *On Liberty*). In recent years, intellectual history has examined a far broader range of authors and texts, reading beyond the best-known thinkers, the 'great books,' and the European canon. But its subject matter—which keeps

it apart from its neighbor disciplines, cultural history and literary studies—remains the written discussions of intellectual questions. It is interested in all records of sustained efforts to debate—within scholarly traditions—aspects of nature and human nature, knowledge and faith, society, and politics.

Historians of ideas identified several key issues that have been part of the literature of classical antiquity, Jewish and Christian thought, medieval and modern European writing: our understanding of the physical world, the applicability of its structure to human situations, faith and reason, morality and government. Broad themes (such as the role of God in the world), long-running ideas (such as 'the great chain of being'), specific concepts (such as love, reason, justice, the good life) and particular terms (such as 'libido,' 'alienation,' 'civil society') all belong to the history of ideas. But historians of ideas, unlike philosophers, do not try to 'explain' these themes and concepts. Nor do they wish, like other social scientists, to 'problematize' them and use them as analytical tools. Instead, they examine how such ideas and terms were used, transformed, transmitted, abandoned, or rediscovered, over time and across cultures.

Such examination can pose questions of continuity and change, gradual or abrupt; it can compare cultures and languages that expressed related ideas or similar terms in different ways; and it can trace the transmission of ideas and concepts. In particular, historians have looked at the passage of ideas between three kinds of interlocutors:

(a) individual thinkers reading and responding to one another, either as contemporaries or across generations;

(b) social groups—such as churches and learned academies—dealing with ideas and transforming them; and

(c) broadly shared ways of approaching ideas—such as philosophical schools, cultural movements, and academic disciplines.

The first category has been the dominant one during most of the twentieth century. Significantly, recent attempts to question the traditional approaches of the history of ideas have often focused on the individual author and reader, suggesting that we may not know as much as we think we know about their minds, intentions, and use of words.

The term 'intellectual history,' running parallel to 'history of ideas' especially in the United States during the second half of the twentieth century, denotes an attempt to depart from the limiting contours of an 'idea,' acknowledging that one may never be able to decide precisely what an 'idea' is. Intellectual history therefore moved on to analyzing texts, the uses of language, and the location of participants within the contours of intellectual discourse. It addresses the complexities of transmission, acknowledging that 'ideas' belong in texts and that texts have contexts. Inspired by neighboring disciplines, especially literary theory, intellectual history has been increasingly aware that 'meaning' and 'understanding' are neither self-evident, nor easy to trace, and can be affected by the historian's own perspective.

New intellectual history has a broader view of the nature of intellectual activity: it is no longer only about traditional scholarly debate, not only about 'great books' by towering thinkers (what the German historian Friedrich Meinecke called 'moving from mountaintop to mountaintop'). Intellectual history has broadened to include the writings of lesser-known, and even obscure, men and women. Its source material encompasses art and fiction, visual and musical ways of conveying ideas, misreception and subversion, emotions and conflicts. The intellectual historian's primary sources can include paintings and private correspondence, musical notes and prayer books aphorisms and epitaphs, poems and cinema footage, landscape descriptions and web sites. The great philosophers have not been abandoned, but they are read in new ways and weighed against new evidence.

1. *The Origins of Intellectual History*

The term 'history of ideas' first appeared in Johann Jakob Brucker's *Historia philosophica de ideis* published in Augsburg in 1723. It was taken up by Giambattista Vico who referred to his monumental *New Science* as—among other things—'una storia delle umane idee' (Kelley 1990). It was the eighteenth century, then, that marked off the history of ideas as a particular field of inquiry that differs from histories of other subjects. Eighteenth-century history of ideas also began to consider its subject matter from what the Enlightenment called a ' critical' perspective. And yet, the Enlightenment did not invent intellectual history single-handed.

1.1 *Parent Disciplines and Inspirations*

Two centuries before Brucker and Vico, the humanist historiography of the Renaissance had reached out to encompass all aspects of human thought and endeavor and aimed at writing 'the history of wisdom' (*historia sapientiae*). Europe's early modern universities developed historical accounts of their own disciplines, doctrines, and schools. The direct parent discipline of the history of ideas was the history of philosophy, which lent a particular philosophical and theological flavor to the historical accounts (Kelley 1990). Consequently, historians of ideas often asked not only the question 'what ideas did past thinkers discuss,' but also, 'what is the meaning of ideas in history,' and 'how does the history of ideas evolved.'

Another important parent discipline was Renaissance rhetoric, which offered a linguistic analysis

of the way human beings think and form their ideas. The humanist tradition of rhetoric was cast aside by major early modern philosophers, like Descartes and Leibniz who claimed to be able to describe the way things are, independent of how they are worded. But the rhetorical approach to ideas was to return, with a vengeance, in the future career of intellectual history.

1.2 Eighteenth- and Nineteenth-century Histories of Philosophy and of the Human Spirit

Enlightenment histories of philosophy, learning, literature, and the arts, which formed the nascent history of ideas, departed from former doctrinal accounts by trying to find a general meaning for the development of thought and learning though the ages. Whether God remained in the picture (Lessing) or not (Voltaire, Kant, Condorce), these histories focused on the 'progress of reason,' or on 'the history of human spirit' (Kelley 1990). The former looked at great philosophers, while the latter followed the broader dispersion of ideas different cultures and historical phases. Both these approaches to the history of ideas offered a variety of evolutionary accounts, which reflected the Enlightenment's fascination with the question of progress, the respective advantages of ancients and moderns, and models of history as a linear advance, as cyclical, or as prone to regressions.

This enlightenment fascination gave rise to two questions still haunting historians today: should intellectual history center on 'great thinkers' or on many minds? And in what sense does the history of ideas 'progress'?

The question of progress opened up a new inquiry about Europe's singularity. It led Enlightenment thinkers and their nineteenth-century heirs to create a new chronology for European history. This chronology was deeply affected by the history of ideas: Rome's grandeur and decline were considered in terms of ideas, religious and political, affecting its fate (Montesquieu, Ferguson, Gibbon). The newly minted 'middle ages' were temporarily cast aside as an intellectually 'dark' era. And, in the hands of Jacob Burckhardt, the Italian Renaissance became a focal point for an account of Europe's intellectual and cultural renewal, its leap to a great future from the springboard of a rediscovered past.

But precisely the same questions led other scholars to look beyond Europe, seeking intellectual traditions in other parts of the world and juxtaposing them against the European narrative. Nineteenth-century scholarly concern with the Jewish scriptures, Chinese religion, Indian mythology, and Islamic philosophy was partly an outcome of the Enlightenment wrestling with questions of linearity in the history of ideas.

Georg Wilhelm Friedrich Hegel saw history as a process of unfolding philosophical patterns, changing with the 'spirit' of the times and tied to the past in intelligible dialectic ways. Hegel also led the way to considering the history of the mind as categorically different from the physical world, inspiring the rise of the 'human sciences' (*Geisteswissenschaften*) in the nineteenth century. The controversial distinction between the physical and the human sciences still haunts scholars today. Intellectual historians are affected by it in their complex relationship with political history: in what sense does the 'world out there' intervene in the history of ideas? Can ideas and reality affect one another? Is their interplay an appropriate subject matter for the intellectual historian?

Another Enlightenment legacy, eclecticism, cast different light on questions of ideas and reality. Eclectics and 'popular philosophers' attempted to pick up the most interesting strands in the history of philosophy, and to compare, combine, and apply them in ways appropriate for the modern age. This approach took up the Enlightenment's key concept of 'criticism' and purported to judge, select, and rearrange contents of earlier philosophical writings. Furthermore, eclectics (notably Victor Cousin) began to suggest that the history of ideas must look to 'external' causes, cultural and environmental. Karl Marx was to follow with a more assertive theory of the dependence of ideas on economic and social factors.

The nineteenth-century views of the role of ideas in history left several unresolved questions: ought intellectual historians approach their subject matter judgementally and selectively, as eclectics and ideological thinkers have done? Or should they—as nineteenth-century historicists and twentieth-century neo-historicists argued—immerse themselves in the inner logic of the texts and the era they are studying, suspend judgement and avoid any attempts to mobilize the past in service of the present?

The Enlightenment's 'critical' approach to the history of ideas was especially interesting when it looked at language. Vico, Condillac, and Herder suggested that many alleged 'truths' of philosophy are anchored in the way thinkers use words. While philosophers like Descartes and Kant claimed to be able to grasp metaphysical truth, either about the world or about human reason, a more historical perspective disclosed that language is an unavoidable and problem-ridden medium of all human thought. Language affects individual thinkers by drawing up the conceptual lines available to their culture and era. Thus, as Christian Garve suggested in 1772, eighteenth-century Englishmen may create a political philosophy based on the concept of 'public spirit,' but their German contemporaries can not; the linguistic and conceptual apparatus is (still) missing.

The question whether language necessarily delimits thought, or may in turn be expanded by thought, is still open. Garve's claim that his German contemporaries could not understand 'public spirit' found an echo in the twentieth-century work of Hans Georg

Gadamet about the 'horizons' of cultures and languages. These horizons change and fuse, but they represent a finite range of possibilities available to any thinker at a given place and time.

And yet, thinkers have always played with language and fought against its conventions. Mary Wollstonecraft, for example, was not content with the prevalent distinction between 'male virtue' (political) and 'female virtue' (matrimonial). She chose to push the concepts around rather than succumb to her era's linguistic horizons. By applying civic virtue to women, she accomplished a theoretical breakthrough (and heralded a social one). Thus, as recent intellectual historians have suggested, language interacts with thought in more ways than one.

2. *Twentieth-century History of Ideas*

The emergence of the new history of ideas in the early twentieth century involved a conscious self-distancing from philosophy, the use of source criticism and analytical methods developed by modern historians from Ranke to the Annales school, and new sensitivity to the scope and limits of its subject matter. During the second half of the twentieth century, the history of ideas became subject to the same transformations that affected historical studies as a whole; and, due to its textual focus, it has been especially prone to critique and innovation inspired by literary theory and by postmodernist thought.

2.1 *Three Founders*

In the American context, Arthur O. Lovejoy is often seen as the founding father of twentieth-century history of ideas, a discipline closely identified with the books he wrote and the journal he edited. There were several other seminal scholars, of whom two were chosen here for representing broad European vistas and unique philosophical perspectives: the German-American Ernst Cassirer and the Russian-British Isaiah Berlin.

Arthur Oncken Lovejoy wished to consolidate the history of ideas and pare off a host of other 'histories' that had become entangled with it. The crux, he suggested, was the study of specific ideas, 'unit-ideas,' in their historical evolving. The most famous example for this approach was Lovejoy's own perusal of the history of one 'unit-idea,' the idea of the great chain of being. Lovejoy offered a historical account of the 'principle of plenitude,' starting with the early Greek notion that all possibilities would eventually be realized, and tracing the trajectory of this idea up to the modern era (Lovejoy 1936). He went on to discuss the histories of such general ideas as Romanticism, primitivism, and evolutionism.

The transformation of ideas over time was for Lovejoy by no means a linear ascent, nor 'an exclusively logical progress in which objective truth progressively unfolds itself in a rational order.' In contrast to linear progressivists, Lovejoy envisaged a perpetual historical 'oscillation' between intellectualism and anti-intellectualism (Kelley 1990). Lovejoy's legacy, especially the *Journal of the History of Ideas* he founded and edited, has affected many American historians. Recent criticism of his approach centres on the primacy of Reason in his work and on the difficulty of clarifying his notion of 'unit-ideas' as independent historical factors.

Like Lovejoy, Ernst Cassirer was a philosopher as well as a historian. Cassirer understood the rise of modern culture in terms of 'symbolic forms,' ways of cultural expressions that gradually acquired autonomous standing, including language, religion, art, and myth. As his seminal work on the Enlightenment (Cassirer 1932) shows, Cassirer's view of the history of ideas was evolutionist, but not rationalist. In the crucial era of the Enlightenment, he suggested, fields of inquiry gradually liberated themselves from the powerful matrix of natural science, creating a modern web of truth-quests leading to Goethe's literature and to Kant's philosophy. Subsequently, in the era of Romanticism, the prevalent focus on Reason transformed itself into a subtler aesthetic concept of the human search for truth.

Cassirer's work has been criticized for reasons similar to critiques of Lovejoy: he did not care to look beyond 'great thinkers,' his work centered on male European philosophers, and his progressivist master-narrative was no longer convincing to late twentieth-century historians. Cassirer, a refugee from Nazi Germany, shared Lovejoy's faith in liberal and rational responses to political evil; late twentieth-century historians did not feel his urge to reflect this moral imperative in scholarly work. On the other hand, Cassirer's approach to the history of ideas helped to broaden its horizons by seriously considering artistic imagination and religious faith as prime sources for the history of ideas, parallel but not reducible to scholarly writings.

Isaiah Berlin, who was reluctant to define himself as a historian, is one of the writers most closely associated with the history of ideas outside the American scholarly tradition. Berlin was deeply engaged with the relationships between ideas and reality. He was convinced that thinkers, especially modern European thinkers, have been able to affect leaders, revolutionaries, and hence the personal fate of millions, for better and for worse.

Berlin's works on Karl Marx, on the 'Counter-Enlightenment,' on English liberalism, and on the Russian intelligentsia convey two fundamental beliefs about the history of ideas. First, Berlin agreed with his protagonists Vico and Herder that the historian can enter the mind and reasoning of past writers, 'feel' them from within and represent their thoughts faithfully and (in Berlin's case) intensely. Second, Berlin's

writing implied that the history of ideas is adjacent to political engagement with ideas. Berlin found two traces running underground in the history of modern European thought: one associated with 'negative liberty,' associated with individual freedom from official intervention; and another associated with 'positive liberty,' a range of yearnings for communal self-fulfilment and meta-individual quests for the good life (Berlin 1958). Similarly, Berlin identified two kinds of thinkers: the 'fox' who is capable of exploring multiple ideas and possibilities, and the 'hedgehog' who creates a monolithic theory based on one type of explanation or one great idea. Berlin's history of ideas was closely associated with his own preference for negative liberty and for thinkers defined as 'foxes' (Montesquieu, Tolstoy, Mill), while exploring with great empathy the arguments for positive liberty and the thinkers who were essentially 'hedgehogs' (Rousseau, Dostoyevsky, Marx).

Criticism of Berlin has often focused on his broader-than-academic description of past writers, and on the idiosyncrasy of his style. Along with his aversion to monolithic explanations, these factors prevented Berlin from creating a historical 'school' of his own. In the last decade of the twentieth century, however, Berlin's work has been taken up by historians of ideas who wish to explore their relationship with political philosophy. Berlin's distinction between negative and positive liberty has been taken up by historians, notably Quentin Skinner, as a reference point for redefining the scope of the history of ideas.

2.2 New Approaches and Problem-setting in the Late Twentieth Century

From the early 1970s intellectual history has evolved in several directions. Interestingly, the new approaches often display 'national' schools of historical writing, banking and disparate scholarly traditions. In addition, they have been informed by innovations in cultural, linguistic, and literary studies.

2.2.1 American developments and critiques of Lovejoy's legacy. In the United States, Lovejoy's legacy has been developed by many disciples and colleagues associated with *The Journal of the History of Ideas*, founded by him in 1940. A great deal of the work has followed in Lovejoy's footsteps, looking at the historical development of themes (no longer necessarily called 'unit ideas'), mainly in European history. Even while abandoning Lovejoy's dualistic, liberal meta-history, the tradition he founded carried on what may be termed a positivist outlook on the history of ideas.

Two important new dimensions were offered by Dominick LaCapra and Hayden White, both of whom have attempted to apply methods derived from literary studies to intellectual history. According to LaCapra, intellectual history ought to benefit from the broader connections made by literary criticism between texts and the contexts of their writing and reading. Literary texts should be taken up as legitimate sources, and the mental environment of authors, both literary and scholarly, should be perused more carefully (LaCapra 1983).

According to Hayden White, structuralism can be applied to intellectual history by taking up the common denominators of realistic and fictional narratives. Texts shed light on their original 'mental climate' whether they are documents or stories, starkly descriptive or ideologically biased (White 1987).

Both LaCapra and White thus expanded both the source pool and the analytical tools of Lovejoy's tradition by appealing to literature as subject matter, historical witness and methodological pointer. Criticism of the structuralist and poststructuralist attempt to collapse the lines between literary and scholarly (or documentary) texts have focused on questions of genre, intention, evidence, and judgement: are novels really akin to philosophical tracts under the historian's gaze? Is intellectual history enriched or impoverished by identifying historical narratives with fictional ones? Are all authors equally reliable, or unreliable? Are all 'stories' equally meaningful? Are traditions irrelevant? Is 'historical truth' obsolete?

2.2.2 The British–American school of linguistic contextualism. The contextualist school, also known as the 'Cambridge school,' has maintained the traditional focus on scholarly and philosophical texts, while introducing a novel emphasis on language into the history of political and social thought.

This emphasis was in some sense heralded by the Welsh historian Raymond Williams, who held that in modern European history many important concepts—'keywords'—have undergone 'conscious changes, consciously different uses; innovation, obsolescence, specialization, extension, overlap, transfer' (Williams 1985). William's pioneering interest in the social evolution of words was, however, overtaken by a broader linguistic approach inspired by discourse theory.

Linguistic contextualism was framed by the works of J. G. A. Pocock and Quentin Skinner. While specializing in early modern Western Europe, its insights can be applicable to other periods and topics of intellectual history. Pocock and Skinner have carried 'context' beyond its traditional social, political, and economic matrices by looking at textual contexts: Sources should not be studied solely on the merits of their own contents, nor just vis-à-vis other 'great works' along the mountain-chain of traditional history of ideas. The most important context is an author's immediate textual environment. A perusal of many

contemporaneous writings, often minor or obscure, will yield a sense of a 'political language,' or what Pocock called a 'paradigm.' Pocock found several early modern paradigms, primarily in English and British history: the languages of the 'ancient constitution,' natural law, classical republicanism, politeness and civility (Pocock 1985). Many authors used more than one 'language'; Montesquieu, for example, combined natural law and classical republican elements. When the coexistence and interplay of such discourses is recognized, an author's conformity, originality and polemics are better understood. Hence the importance of minor figures for shedding new light of the major protagonists of the history of ideas. Thus, John Locke can be examined more closely as responding to his near-contemporary Robert Filmer rather than seen merely as a long-distance interlocutor of Thomas Hobbes on the one side and Jean-Jacques Rousseau on the other.

Skinner placed particular emphasis on the intentions of authors. It is precisely the understanding of contemporary textual context that allows us to identify the use an author makes when choosing particular words. Conventions can be taken up or subverted, shared vocabulary can be accepted, or expanded, or abandoned. But such intentional discursive play, with its subtle layers of allusion, criticism, and irony, can only be unearthed if the language of numerous and variegated contemporaneous texts is studied closely.

Criticism against linguistic contextualism has taken several directions. One direction is that of critical conceptual history, discussed below. From the vantage point of literary studies, Dominick LaCapra has argued that linguistic contextualism over-emphasizes 'documents' at the expense of literature, depriving poetry and fiction of their legitimacy as evidence-givers and of their status as viable primary sources in the history of ideas.

2.2.3 The German school of Begriffsgeschichte (*history of concepts*). Led by Reinhard Koselleck, the multi-volume project of *Geschichtliche Grundbegriffe* ('basic historical concepts,' Brunner et al. 1972–92) analyzes the historical semantics of key political and social concepts in German and other European languages during the 'saddle era' of modernization, 1750–1850. Like the Cambridge School, the *Begriffsgeschichte* attitude requires close contextual examination of primary sources. Authors are seen as working within a conceptual environment, where the major terms are continuously used but their meaning evolves and transforms through intellectual, social, and political changes. Thus, terms such as 'history,' 'civil society,' and 'revolution' acquire new ranges of significance as they move from their Greek or Latin beginnings, via medieval applications, to early modern times. The most crucial transformation, however, took in the decades immediately preceding and following the French Revolution.

Begriffsgeschichte was thus more committed to a strict periodization of turning points in history of ideas than its British and American counterparts. It also remained closer to political chronology. The *Geschichtliche Grundbegriffe* project was intended by Koselleck as a bridge between social history and the history of ideas. It thus maintained a commitment to social, political, and economic contexts alongside the purely linguistic approach.

Furthermore, the *Begriffsgeschichte* focus on concepts—understood as terms, sometimes accompanied by a derivative or a synonym—conveys a more formal understanding of the unit of analysis than either Lovejoy's 'unit-ideas' or Pocock's and Skinner's notions of 'political language.' The adherence to concepts has been criticized for its inflexibility; yet it also allows historians to fine-tune their understanding of conceptual distinctions, shifts of meaning, deliberate changes, neologisms and disagreements about proper usage of concepts (Richter 1990). English-speaking proponents of 'critical conceptual history' have used *Begriffsgeschichte* as a corrective to the overarching generalizations of discourse theory (Ball 1988).

2.2.4 French–American approaches to intellectual history. The intellectual history of France, centered on the French Revolution, its origins, and aftermath, has developed along lines different from the British, American, and German ones. From Daniel Mornet's *Intellectual Origins of the French Revolution* (1933) to the late twentieth-century work of Roger Chartier and the American historians Robert Darnton and Keith Michael Baker, French history of ideas has been less associated with isolating 'unit-ideas,' still lies with political discourse. Partly due to the powerful impact of the Annales school, it has been strongly tied with social and cultural history (Chartier 1982).

Consequently, intellectual history of France often appears in overlap with cultural history, especially where the history of reading, and of the book as an artefact, is concerned. From this perspective, Darnton (1980, p. 337) suggested that intellectual history includes 'the history of ideas (the study of systematic thought usually in philosophical formulations), intellectual history proper (the study of informal thought, climates of opinion and literary movements), the social history of ideas (the study of ideologies and idea diffusion), and cultural history (the study of culture in the anthropological sense, including worldviews and collective mentalités).'

3. *Major Issues, Controversies, Disciplinary Overlaps*

Intriguingly, almost all major issues and controversies associated with intellectual history during the last

third of the twentieth century were associated with its intersections with neighboring disciplines.

Social history represents claims to expand the new contextualism by placing authors and readers, as well as books, within broader social contours. (Hughes 1988). The *Begriffsgeschichte* school, in particular, attempted a social–historical approach to the history of ideas.

Cultural history poses a wide range of problems for intellectual historians, ranging from accusation of narrowness and elitism to suggestions for expanding the range of inspected authors and texts and moving beyond the written word or the individual thinker (Blake 1996).

The intellectual history of women has recently attracted growing attention. Key questions here relate to the past neglect of women thinkers, the effect of gender on the contents of thinkers' ideas and the effect of political thought on women's status and image (Kerber 1997).

The history of science, traditionally part of intellectual history, is recently posing a new set of challenges associated with novel issues such as ecology, genetics, gender differences, and animal experimentation. More fundamentally, the history of science highlights the quest for scientific truth, which provides a context for recent debates about the cultural dependency of historical truth.

Linguistics and literary theory—especially semiology, reception theory, discourse theory, structuralism, and deconstruction—have deeply affected intellectual history. They have opened up questions, as seen above, on the nature of its primary sources, their legibility and reliability, the status of writers and readers and the accessibility (or even the very relevance) of their intentions and thought processes.

Finally, philosophy, the mother discipline of the history of ideas, is still its close accomplice. Lovejoy, Cassirer and Berlin (as well as Alasdair MacIntyre and Quentin Skinner) have all pursued the history of ideas as part of a philosophical work. Richard Rorty saw intellectual history as describing 'what the intellectuals were up to at a given time' and 'their interaction with the rest of society,' therefore exempt from worries about defining philosophy and philosophers, and hence particularly helpful to philosophers in revising their own canon afresh (Rorty 1984).

Rorty's definition—seeing the philosopher as one consumer of intellectual history—incidentally raises one of its most perennial issues: whose intellect, which ideas, belong to (what) history of ideas?

4. New Directions and Future Agendas

Intellectual history is currently responding to external critiques of its narrow, selective traditional subject matter by expanding its temporal and geographical horizons in several ways. In the wake of new interest in women's intellectual history and in minor writers, it is also looking at previously marginalized European cultures and at the non-European world.

Women's intellectual history suggests that women thinkers could be analyzed in different ways than their male counterparts. It is yet unclear whether the history of non-European ideas might similarly transform itself into a non-European history of ideas, raising new conceptual approaches on top of new source materials. However, the question whether gender or ethnic groups require an intellectual history 'of one's own' remains open and is hotly debated.

Historical semantics, which has faced various critical attacks, are being partly rejuvenated by computer-based corpuses and methods of research. Linguistic conceptualism is expanding beyond its Anglophone lair, meeting up with *Begriffsgeschichte* and becoming interested in other (real) languages, political discourses, and questions of reception and interaction. Social and cultural contexts of the production of texts are gaining ever more attention. And clearly, perhaps due to the new attention to female, noncanonical, and non-European thinkers, it appears that the Author, along with her or his opinions and intentions and beliefs, is far from dead.

See also: Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); History and Memory; History: Overview; Intellectuals: History of the Concept; Intellectuals, Sociology of; Knowledge (Explicit and Implicit): Philosophical Aspects; Knowledge, Sociology of; Linguistic Turn and Discourse Analysis in History

Bibliography

Ball T 1988 *Transforming Political Discourse: Political Theory and Critical Conceptual History*. Oxford University Press, Oxford

Berlin I 1958 *Two Concepts of Liberty*. Clarendon Press, Oxford

Berlin I 1997 *The Sense of Reality: Studies in Ideas and Their History*. Farrar, Strauss and Giroux, New York

Blake C N (ed.) 1996 Symposium on 'Intellectual history in the age of cultural studies.' *Intellectual History Newsletter* **18**: 3–69

Boucher D 1985 *Texts in Contexts: Revisionist Methods for Studying the History of Ideas*. Dordrecht, The Netherlands

Brunner O, Conze W, Koselleck R (eds.) 1972–92 *Geschichtliche Grundbegriffe: Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*, 7 Vols. Klett-Cotta, Stuttgart, Germany

Cassirer E 1932 *Die Philosophie der Aufklärung*. Mohr, Tübingen [1951 *The Philosophy of the Enlightenment*. Trans. Koelln F C A, Pettegrove F F]. Princeton University Press, Princeton, NJ

Chartier R 1982 Intellectual history or sociocultural history? The French trajectories. In: Lacapra D, Kaplan S L (eds.) *Modern European Intellectual History. Reappraisals and New Perspectives*. Cornell University Press, Ithaca, NY
Darnton R 1980 Intellectual and cultural history. In: Kammen M (ed.) *The Past Before Us: Contemporary Historical Writing in the United States*. Ithaca, NY
Hughes G 1988 *Words in Time: A Social History of the English Vocabulary*. Oxford University Press, Oxford
Kelley D R 1990 What is happening to the history of ideas? *Journal of the History of Ideas* **51**: 3–25
Kerber L K 1997 *Toward an Intellectual History of Women*. University of North Carolina Press, Chapel Hill, NC
Koselleck R 1979 *Vergangene Zukunft. Zur Semantik geschichtlicher Zeiten*. Suhrkamp, Frankfurt [1979 *Futures Past: On the Semantics of Historical Time*, trans. Tribe K. Cambridge, MA]
LaCapra D 1983 *Rethinking Intellectual History: Texts, Contexts, Languages*. Cornell University Press, Ithaca, NY
Lovejoy A O 1936 *The Great Chain of Being: A Study of the History of an Idea*. Harvard University Press, Cambridge, MA
Pocock J G A 1985 The state of the art. In: Pocock J G A (ed.) *Virtue, Commerce and History: Essays on Political Thought, Chiefly in the Eighteenth Century*. Cambridge University Press, Cambridge
Richter M 1990 Reconstructing the history of political languages: Pocock, Skinner, and the Geschichtliche Grundbegriffe history and theory. *Studies in the Philosophy of History* **24**: 41
Skinner Q 1988 Motives, intentions and interpretation. In: Tully J (ed.) *Meaning and Context: Quentin Skinner and his Critics*. Cambridge
White H 1987 *The Content of the Form: Narrative Discourse and Historical Representation*. Johns Hopkins University Press, Baltimore, MD
Williams R 1985 [1976] *Keywords: A Vocabulary of Culture and Society*, rev. edn. Oxford University Press, Oxford, UK

F. Oz-Salzberger

Intellectual Property, Concepts of

The concept of 'intellectual property' refers to ownership rights relating, for example, to a particular invention (that might be a process or product), a set of scientific research results, the written text, and increasingly so today, electronic versions of all these. Intellectual property (IP) has two forms: industrial property, which encompasses rights over inventions, trademarks, and designs, and copyright covering artistic, literary, musical, audio-visual, and photographic works. The granting of formal monopoly rights for inventions can be traced back to the medieval period in Europe, notably in England and France, though comprehensive legal protection did not appear until the mid-eighteenth century: in the US, for example, it was first enshrined in the 1778 Constitution. The granting of rights is made by IP agencies such as patent offices and patent courts at national and international levels, and the number of rights held is considerable. By 2000, for example, approximately 720,000 patents had been granted by US and European Patent offices.

The field of intellectual property conventionally is regarded as the preserve of the lawyer. However, social science also explores the social and economic processes through which people, organizations, or governmental agencies make claims to IP, and, thereby, to certain legal rights, such as copyright or patents. Social science analysis particularly has been interested in understanding how these rights are negotiated and defined, how they may change over time, their tie-in with changes in the wider economy and innovation system, and the political and ethical issues they raise. These issues have informed, for example, the debate over the patenting of national genetic registers (databases) of whole populations by major pharmaceutical companies in order to facilitate future drug development: such a step has been taken in Iceland. This review of intellectual property focuses in particular on work in the social sciences (e.g., Weil and Snapper 1989, Etzkowitz and Webster 1994, Narin et al. 1997, Etzkowitz et al. 1998) that has explored the links between IP, the innovation system, and the wider 'knowledge society' (see *Knowledge Societies*).

1. *The IP System and Contemporary Innovation*

The conventional rationale for the IP system is that it rewards creativity and innovation by providing monopoly rights to inventors for a specified period of time—typically up to 20 years. In return for disclosing or publishing something new so that others may have access to and use novel ideas, the originator receives some form of payment or recognition (such as acknowledgment in the case of prior copyrighted material). However, the pace and nature of innovation today make this process much more complicated than in the past, especially, for example, in areas such as the 'new genetics.' The question then arises whether IP law keeps up with the pace of knowledge-based innovations (Hoffman and Karny 1988) such that it can properly and coherently determine when monopoly rights are to be granted and when not, when novelty, nonobviousness, and utility (i.e., being capable of industrial application)—the three requirements of any claim to a patent—are in evidence and when not. In this regard, there has been considerable debate over the granting of patents for gene sequences whose function within the context of the human genome is still unknown. Moreover, such discoveries are regarded by many as simply the normal process of scientific investigation which should be made fully available in the public domain.

At the technical level, the laws governing what can and cannot be secured as patentable material, pro-

ducts, or processes are continually subject to reinterpretation to accommodate an ever larger range of life forms and forms of invention. The field of biotechnology has posed particular problems for IP law, not least in inventors being required to make a deposit and full description of their particular inventive step. The organic and complex structures produced through genetic engineering have meant that new arrangements have had to be made to meet this particular legal requirement. Problems, too, are found in the information technology sector where firms developing computer hardware and software promote the spread of their technologies while risking them being cloned or pirated by third parties because the security offered by copyright means very little in real terms.

Apart from the problem of handling the ever-changing demands of innovative technologies, especially those that derive from new interdisciplinary fields (e.g., agro-pharmaceuticals, or bio-informatics), the IP system must also cope with the growth of interface technologies—sometimes called 'handshaking' technologies—that enable independent systems and technical instruments to be compatible with each other. These handshaking technologies depend on competitors collaborating so that each can hook up their particular products to a wider system. Yet the question then arises of how to ensure that those firms whose innovation sets the required standard will enjoy some return on their investment (Shurmer 1996). The regulation of new technologies poses a related but distinct problem for the IP system. The need to ensure that the risks of new drugs or genetically engineered plant organisms are kept to a minimum has often led to regulatory delays on those products coming to the market. As such, one side of the 'contract' with the inventor(s) seems compromised, that is, the real value of the exclusive rights granted to an inventor will be much less because of the reduced period of patent cover once the product has come to market. In the 1990s, all three major markets of the world (the US, Europe, and Japan) extended patent term protection to pharmaceutical products by a maximum of five years to enable drugs firms a longer effective monopoly. This is a good illustration of the IP law having to adjust to meet the demands of both proprietors and regulatory agencies.

It is likely, therefore, that the IP system will encounter ongoing difficulties by the arrival of new technologies in the future as well as by the ways in which these new technologies relate to each other in the contemporary innovation system (Rothwell 1992, Taylor et al. 2000). Not surprisingly, some in the legal profession have gone so far as to argue that the IP system must adopt a technology-specific focus and develop IP law that is customized to particular technology needs (Christie 1994). For others, however, this would be to over-contextualize IP law and lead to a highly contingent interpretation of novelty and innovation (Mellor 1988).

Notwithstanding these debates, the fact that new IP rights have been granted for plant, animal, and human material raises wider socioethical issues about the patenting of life, the control over 'natural' resources, and moral concern over government's own position regarding its regulation of private sector agencies. This debate has been particularly sharp in connection with the Human Genome Project and the storage and use of genetic information (Anderson 1994, Cook-Deegan 1994) and agricultural resources (Shiva 1994). From a purely competitive position it would seem that patenting of gene sequences or indigenous germplasm would promote innovation, yet many have argued that there should be limits on what might be patented even if this constrains the development of some innovative areas of research and development (McNally and Wheale 1998).

At an international level, there has been an attempt—through GATT's (General Agreement on Tariffs and Trade) trade-related intellectual property provisions (TRIPS)—to stabilize and secure intellectual property rights (IPR) in newly industrializing countries and those of postsocialist central and eastern Europe. The GATT enables richer states to impose trade sanctions on those that fail to conform to TRIPS. This indicates that although the harmonization of IP regimes globally may be an important requirement for globalized innovation systems, the tension between global and national interests is likely to prevail. This has been especially apparent in relation to developing countries' desire to protect their indigenous natural and cultural heritage, and to control how both are translated into knowledge resources over which they have collective rights. In effect, this is to challenge the move made by global corporations and organizations to define the meaning of these resources in such a way that they become commodifiable forms of property owned by private interests. This led to the Biodiversity Treaty of 1992 that recognized the rights of countries in the poorer South to conserve and sustain their genetic resources. The impetus behind the Treaty was the growing concern over the way in which countries in the North have been exploiting the natural resources of the South—such as the fauna and flora on which new drugs might depend—and then selling these products protected by patent. The Treaty has, however, had only limited effect in addressing the poorer countries' concerns. The World Intellectual Property Organisation (based in Geneva) has sought to resolve these tensions, but with little or no success.

2. *The IP System at the Institutional Level*

Economic policy in all advanced industrial countries has sought to enhance innovation by encouraging those who produce knowledge to exploit their intellectual property more effectively. This has had an important effect in bringing a wider range of institu-

tions into closer contact with the IP system, notably small to medium sized enterprises (SMEs) and universities.

Government policy in advanced industrial states has been to favor the needs of SMEs but to regard them as poor defenders of their own interests. In part this simply reflects the dependency of many SMEs on larger corporate actors in an economic supply chain. They are unlikely to have the financial resources to defend their claims against infringement by larger companies. Nevertheless, the importance of small firms in general to innovation and employment is well established. In such circumstances, SME firms and their position within the innovation system takes on increasing importance. Evidence shows that they are much more likely to protect their IP through being faster to the market, technical encryption of know-how, strong ties with customers, and so on, rather than deploying formal IP rights (Rappert et al. 1999).

Universities and other public sector institutions engaged in research and development are also central to recent IP-related science policy. A powerful assumption since the early 1980s in both Europe and the US has been that technology transfer and innovation can be enhanced within an economy by raising the incentive within universities to exploit their own research, and to do this by devolving to them intellectual property rights that result from the publicly funded research they conduct. In the US, legislation introduced in 1980 (the Stevenson–Wydler and Bayh–Dole Acts) gave universities the rights to patents derived from federally funded research. This change (mirrored in many other countries since), coupled with increased industrial sponsorship of university research, set in train a growth in patenting such that patents filed by universities quadrupled between 1971 and 1989 in the US and continue to show rapid growth there and across Europe. (Japan has until very recently prevented its universities owning patents.) There is, however, considerable debate over whether this commercialization of university research, while generating some (often only modest) income, works against the idea of universities providing 'public good' knowledge or mistakes the real practical value of university research (Pavitt 1998). Moreover, concern has been expressed that universities are being restricted in their use of patented processes or (in the US) 'discoveries' by firms holding the IP rights to them. For example, those working in genetics research find that, to continue work on the development of new tests, they have to pay high levels of license fee to use genetic material patented by firms.

Social science research (e.g., Packer and Webster 1996, ESRC 1998) has sought to explore the way in which scientists in universities engage with patenting, how they determine the novelty and scope of their results such that they may be most effectively patented, and how this compares with practice in industry. This showed that in order to identify, take out, and market a patent successfully, academic scientists have to develop skills and knowledge that they do not get directly from their training or work experience in the public sector. These include the ability to distinguish legal from scientific novelty and utility. Scientists' judgment about novelty relates to a specific community of researchers within their field, not to the patent law's definition of 'prior art,' which assumes a global, virtual stock of knowledge. Novelty claims in patenting are distinct from those of science inasmuch as they are positioned differently to previous work in the field, by claiming discrete ownership of ideas, rather than, as is normally the case in scientific work, showing how it builds on earlier work. What this research on patenting shows is that the meaning of novelty is never self-evident but the result of a whole range of interpretive steps, negotiations, and reconstructions of knowledge claims about authorship and priority rights (Boyle 1996) by a wide range of social actors. Moreover, in engaging with the IP system to commercialize their work university scientists must also demonstrate an ability to rewrite scientific work in patent style (Myers 1995, Webster and Packer 1995). They must have access to and the capacity to search and use patent literature, an ability and preparedness to work around existing patents, the capacity to delay or alter their academic publications, and the ability to communicate effectively with patent professionals and industrial liaison staff. Patenting, in other words, requires considerable investment in the elaboration, translation, and articulation of claims through various social agents: it is never simply a matter of a purely technical adjudication of what might be deemed to be legally 'novel.' Ultimately, though, the role of law is crucial in determining whose claim to intellectual property will be upheld (Jasanoff 1995). Even so, courts themselves draw on a repertoire of interpretations relating to conventions about novelty, identity, and the boundaries between objects as the product of 'invention' and to be 'found in nature.'

New developments in multimedia technologies—such as the Internet, World Wide Web, digital communication, and broadcasting systems—are now adding even greater complexity to the interpretation of novelty and material rights. These developments are being shaped by global media groups associated with film, video, music, leisure, education, and e-commerce. International conventions that protect the IP rights of artists, composers, authors, and publishers will come under increasing pressure as copyright infringement will be more difficult to detect, or defend against, especially for smaller companies or individuals who have relied on national collecting agencies to secure royalty payments. It is clear that, here again, new technologies operating at a global level can short-circuit the protection the IP system affords.

Some new business sectors, especially the knowledge-intensive business sector covering areas such as design, accountancy, architecture, management con-

sultancy, and environmental services while sometimes using formal IP rights (such as design rights) more usually rely on informal means of controlling their core assets through working with trusted partners or establishing control over a particular area through strong networking relationships with clients and agencies. Protection through reliance on reputation and trust is regarded as equally if not more important for protecting innovation and competitiveness than are legal rights such as trademarks or patents.

The IP system has developed over the past 300 years to protect the rights of the innovator or artist and to achieve two ends: to reward the originator of the idea, yet to make that idea available to everyone else, through, for example, publication of the patent details carrying the design of the product. This system generally works well but is under increasing strain because of the more complex character of technological innovation today, the more complex patterns of research and development that may make priority claims to the inventive step more debatable, and the globalized nature of business that can create major tensions between countries, between multinational companies, and between states and companies. The globalization of interpretive debates surrounding patent claims merely echoes similar processes of the construction of and negotiation over claim and counterclaim made by social actors at more local levels, whether within the laboratory, the patent office, application, or court.

See also: Intellectual Property: Legal Aspects; Intellectual Property Rights: Ethical Aspects; Property, Philosophy of

A. Webster

Bibliography

Anderson C 1994 NIH drops bid for gene patents. *Nature* **263**: 909–10

Boyle J 1996 *Shamans, Software, and Spleens: Law and the Construction of the Information Society*. Harvard University Press, Cambridge, MA

Christie A 1994 Designing appropriate protection for computer programmes. *European Intellectual Property Review* **11**: 486–93

Cook-Deegan R 1994 *The Gene Wars: Science Politics and the Human Genome*. Norton, London

ESRC (Economic and Social Research Council) 1998 *Intellectual Property Research Programme*. Swindon, UK (www.esrc.ac.uk)

Etzkowitz H, Webster A 1994 Science as intellectual property. In: Jasanoff S, Markle G E, Petersen J, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, London

Etzkowtiz H, Webster A, Healey P 1998 (eds.) *Capitalizing Knowledge*. SUNY Press, New York

Hoffman G M, Karny G N 1988 Can justice keep pace with science? *European Intellectual Property Review* **12**: 355–8

Jasanoff S 1995 *Science at the Bar: Law, Science, and Technology in America*. Harvard University Press, Cambridge, MA

McNally R, Wheale P 1998 The consequences of modern genetic engineering: patents, 'nomads,' and the 'bio-industrial complex.' In: Wheale P, von Schomberg R, Glasner P (eds.) *The Social Management of Genetic Engineering*. Ashgate, Aldershot, UK

Mellor J 1988 Patents and genetic engineering—is it a new problem? *European Intellectual Property Review* **6**: 159–62

Myers G 1995 From discovery to invention: The writing and rewriting of two patents. *Social Studies of Science* **25**: 57–106

Narin F, Hamilton K, Olivastro D 1997 The increasing linkage between US technology and public science. *Research Policy* **26**: 317–30

Packer K, Webster A 1996 Patenting culture and the wheel of credibility. *Science, Technology and Human Values* **21**: 427–53

Pavitt K 1998 Do patents reflect the useful research output of universities? *Research Evaluation* **7**: 105–11

Rappert B, Webster A, Charles D 1999 Making sense of reluctance and diversity: Academic–industry relations and intellectual property. *Research Policy* **28**: 873–90

Rothwell R 1992 Successful industrial innovation: Critical factors for the 1990s. *R & D Management* **22**: 221–37

Shiva V 1994 Why we should say 'No' to GATT–TRIP. *Third World Resurgence* **39**: 3–5

Shurmer M 1996 Standardisation: A new challenge for the intellectual property system. In: Webster A, Packer K (eds.) *Innovation and the Intellectual Property System*. Kluwer, Dordrecht, The Netherlands, pp. 47–64

Taylor P, Quintas P, Storey J, Tillsley C, Fowle W 2000 Flexibility, *Networks and the Management of Innovation*. Open University Business School, Milton Keynes, UK

Webster A, Packer K 1995 Patents and technology transfer in public sector research. In: Kirkland J (ed.) *Barriers to International Technology Transfer*. Kluwer, Dordrecht, The Netherlands, pp. 43–64

Weil V, Snapper J W (eds.) 1989 *Owning Scientific and Technical Information*. Rutgers University Press, New Brunswick, NJ

Intellectual Property: Legal Aspects

Intellectual property comprises an interrelated set of legal regimes protecting economic and in some contexts personal interests in inventions, information, works of authorship, images, symbols, and sound recordings. Intellectual property law principally embodies utilitarian precepts—such as promoting technological innovation (patents, trade secrets), spurring creative expression (copyright), and preventing confusion in the marketplace (trademark)—although strands of intellectual property law also reflect natural rights, personhood, and other philosophical perspectives. Economists and policy analysts have substantially enriched our understanding of the relationship between intellectual property protection and innovation and diffusion of new works, but the complexity of this relationship leaves many questions unanswered. The advent of various new technologies and means of distributing intellectual works during

the latter half of the twentieth century, most notably digital technology and the decoding of genetic structure, have raised fundamental questions about the rationale, implications, and appropriate scope of intellectual property protection. As a result, the premises, doctrines, and contours of intellectual property law have increasingly come into question and have sparked growing social, economic, political, and philosophical interest.

1. Origins

The modes of intellectual property law developed and have evolved in response to the rise of commerce, the shift toward industrial means of production and the development of technologies for the reproduction and distribution of works of authorship (Merges et al. 2000). The first use of trademarks can be traced back almost 4000 years to the earliest merchant societies in China, India, Persia, Egypt, Greece, and Rome where craftspeople marked their wares with distinctive symbols in order to identify the source of their goods, develop a distinctive reputation for quality, and assist in resolving ownership disputes. The roots of trade secret protection can also be traced back at least as far as the Roman empire, where courts created a cause of action for corrupting a slave through bribery or intimidation to obtain disclosure of the slave owner's confidential business information. By the time of the Renaissance, most European nation-states had laws that protected businesses (notably, the Guild cartels) from those who used their secret processes and ideas without permission. These early laws were translated during the Industrial Revolution into statutes that protected 'industrial secrets.'

The first formalized patent systems emerged during the early mercantilist period as a means for European nation-states to unify and increase their power and wealth through the development of manufacturers and the establishment of foreign trading monopolies. The term patent, derived from the Latin *patere* (to be open), refers to an open letter of privilege from the government to practice an art. The Venetian Senate enacted the first patent statute in 1474, providing the maker of any 'new and ingenious device … reduced to perfection so that it can be used and operated' an exclusive license of 10 years to practice the invention. Other nations followed suit and the granting of monopolies for inventions became the dominant means of promoting technological advance (Merges 1995).

The need for formal protection against copying did not significantly arise until the invention and diffusion of the printing press in the fifteenth and sixteenth centuries. Out of fear that this new technology would spread religious heresy and foment political upheaval, the British Crown in 1534 prohibited publication of literary works without a license and approval of royal censors. In 1557, it conferred a publishing monopoly upon the Stationer's Company, a group of London printers and booksellers. This monopoly expired in 1694, resulting in competition from new publishers and widespread counterfeiting of published works. The Stationers persuaded Parliament that such economic conditions would not support new works of authorship, leading in 1710 to the first copyright act, the Statute of Anne, 'An Act for the Encouragement of Learning, by vesting the Copies of Printed Books in the Authors or Purchasers of such Copies, during the Times therein mentioned.' By shifting protection from publishers to authors, this statute laid the foundation for a vibrant market for works of authorship and the flourishing of ideas. As technology for recording and distributing new works of authorship developed, copyright law has expanded and adapted to afford protection against copying in these new and evolving media (Goldstein 1994).

The rise of entertainment industries and celebrities in the twentieth century has led to the development of a right of publicity, which protects individuals against the appropriation of their name, image, or likeness for commercial gain. New information technologies have spurred new forms of protection for the design of semiconductor chips and databases.

2. Modern Landscape

The modes of intellectual property law have evolved into highly detailed bodies of statutory law augmented through judicial interpretation. International agreements and the globalization of commerce have harmonized many aspects of national regimes, although variations persist. Furthermore, the development of technologies that cut across the modes of protection, such as computer software, have led to more integration in the practice of intellectual property law (Lemley et al. 2000).

2.1 Patent Law

An inventor of a new, nonobvious, and useful process, machine, manufacture, or composition of matter, or improvement thereof, obtains exclusive rights to make, use, and sell the invention for a term of 20 years from the filing of a patent application. This period, during which the inventor can exclusively exploit the invention, directly or through licensing, provides strong incentives for research and development. The public benefits through the spur to innovation and disclosure of the patented invention. After the patent expires, the innovation becomes part of the public domain, free to all. In addition to utility patents, patents are available in the USA for distinctive plants and ornamental designs for articles of manufacture. Design patents have a 14-year term. A patent can only

be enforced in the nation in which the patent has been awarded.

2.2 Copyright Law

Copyright law promotes creativity in literature and the arts by affording authors and artists lengthy terms of protection (life of the author plus 70 years) against copying. Copyright law protects the expressive elements of a broad range of works—including books, graphical works, dramatic works, choreography, musical compositions, sound recordings, films, sculpture, architectural works, and computer programs—but does not extend to facts, ideas, or utilitarian aspects of such works. Copyright protection inheres upon the creation of original works of authorship fixed in a tangible medium of expression. By contrast with patent protection, copyright law employs a relatively low threshold for protection and substantially longer duration, but the scope of copyright protection is narrower and less absolute. For most categories of works, copyright owners have the exclusive right to make copies, prepare derivative works, and distribute, perform, and display their works during the term of protection. Most continental European nations, and to a lesser extent the USA and other nations, protect attribution and the artistic integrity of expressive works through inalienable moral rights. In response to the increased vulnerability of digital works to widespread piracy, the World Intellectual Property Organization (WIPO) Copyright Treaty requires signatory nations to provide adequate legal protection against the circumvention of copy protection systems and the removal or alteration of copyright management information conveyed along with a copyrighted work.

Unlike patent rights, which are essentially absolute (apart from antitrust limitations), the rights of copyright owners are constrained by the fair use doctrine, which permits limited use of protected works for criticism, news reporting, teaching, scholarship, and research. Copyright law also provides for compulsory licensing of musical compositions and television signals for cable distribution under particular circumstances. Under international conventions, works qualifying for copyright in member nations are protected in other member nations.

2.3 Trademark Law

Trademark law facilitates and enhances consumer decisions and encourages firms to supply quality products and services by protecting names, logos, words, phrases, symbols, sounds, trade dress, product configuration and other means of designating the source of commercial products or services (Economides 1988). A supplier acquires trademark protection by being the first to use a distinctive mark to identify its goods in commerce. Inherently distinctive marks, such as Kodak for photographic supplies, and suggestive marks, such as Coppertone to designate sun tanning products, receive protection immediately, whereas descriptive (or geographic) marks acquire protection only after they have acquired secondary meaning for the consuming public. Trademark protection is perpetual unless a mark is abandoned or becomes associated in the public's mind with a general category of products rather than the source of the product. For example, the terms 'escalator,' and 'thermos,' which initially referred to the products of particular manufacturers, have become generic and hence unprotectable.

Trademark owners have the right to prevent others from using their marks or otherwise confusing consumers as to the source of a product. In some jurisdictions, and now under federal law in the USA, the owner of a famous mark can enjoin and obtain damages for the dilution, blurring, or tarnishment of the owner's mark. Trademark law does not protect functional aspects of a product and allows others to use a mark fairly, such as for comparative advertising or other descriptive purposes, so long as such use does not confuse consumers.

Trademarks are territorial in nature. Unless registered, trademarks are protected only in those geographic regions in which the marked products are sold or advertised. National registration affords nationwide protection of trademarks and the opportunity to obtain foreign registration. Because the USA has not joined the Madrid Agreement Concerning the Prevention of False or Deceptive Indications of Source, the main international trademark convention, US trademark owners must prosecute trademark applications in each country where they seek protection.

2.4 Trade Secret Law

A trade secret is information that derives independent economic value from not being generally known or readily ascertainable and is the subject of reasonable efforts to maintain secrecy. Unlike patent law, information need not be novel in order to qualify for trade secret protection. Most nations protect trade secrets through national or regional regimes. Trade secret protection lasts until the information becomes publicly known. By contrast with patent protection, trade secret law does not protect against independent discovery or reverse engineering by competitors. Those who violate trade secret law are liable for damages and in some cases injunctions against use or further disclosure.

2.5 Other Forms of Intellectual Property Protection

Many jurisdictions recognize some form of unfair competition or misappropriation law prohibiting the

appropriation of some forms of information not protected by traditional modes of intellectual property law. The right of publicity, which is recognized in about half of the USA, protects individuals against the use of their name, image, or likeness for commercial gain.

With the advent of new technologies, some new specialized intellectual property regimes have emerged in recent years. Many industrialized nations, at the urging of the USA, have adopted regimes for the protection of semiconductor chip designs. In 1996, the European Union adopted a directive on the legal protection of databases.

3. *Utilitarian/Economic Theories*

Economic research demonstrates that technological advancement and increased human capital of the labor force account for most of the annual productivity increase in the USA and other industrialized nations (Denison 1985). It is not surprising, therefore, that public policy analysts and social scientists have viewed intellectual property protection principally as a means of promoting intellectual creativity. The precise role of intellectual property protection in promoting technological advancement and the optimal design of such regimes, however, have proven extraordinarily complex and elusive (Menell 1999).

3.1 *Theoretical Perspectives*

The utilitarian theory of intellectual property developed in response to the use of monopoly power to spur innovation. Adam Smith, while generally critical of monopoly power as detrimental to the operation of the 'invisible hand,' nonetheless justified the need for limited monopolies to promote innovation and commerce requiring substantial up-front investments and risk. In the terminology of modern welfare economics, information is a public good: like national defense, it is difficult to exclude those who do not pay for the good from consuming it and additional consumers of the good (e.g., the ideas in a book or improved production process) do not deplete the quantity or quality of the good available to others. Therefore, apart from the costs of distribution and incentives for creation, the optimal cost of information is zero. And in networks markets—such as telecommunications systems, computer operating systems, and standardized parts—there may be positive externalities to more people knowing and using the same information (Shapiro and Varian 1999).

3.2 *Historical, Industry, and Institutional Studies*

Studies across a wide range of industries find that technologies undergo a 'gradual, evolutionary development which is intimately bound up with the course of their diffusion.' (David 1985). Secondary inventions—including essential design improvements, refinements, and adaptations to a variety of uses—are often as crucial to the general of social benefits as the initial discovery. Many studies emphasize the critical commercial importance of linking innovation with understanding consumer needs and astute marketing.

Economic historians and industrial organization economists conclude that patents were rarely the principal means of appropriating returns in most industries (outside of pharmaceuticals and chemicals), and that trade secrecy, lead time, rapid movement down the learning curve, and marketing efforts play substantial roles in the appropriating of returns to investments in research and development (Levin et al. 1987, Mansfield, 1986, Schwartzman, 1976, Taylor and Silbertson, 1973). Saxenian (1994) finds that Silicon Valley has thrived as an incubator of technological innovation because of an extraordinary confluence of factors, including a unique culture of collaboration and sharing of know-how both within and across firms, high mobility of labor, competitive rivalry among many dynamic competitors, high density of complementary specialized enterprises, development of effective trade associations and consortia, creative use of strategic partnering, cross-licensing, a responsive, knowledgeable, and competitive venture capital financing network, close research university-industry relationships, a legal culture emphasizing informal, practical, flexible, and less litigious solutions, a management style emphasizing teamwork, openness, participation, and autonomy of decentralized engineering teams, and the use of stock options to attract, motivate, and reward innovation. Nonetheless, patents continue to play a significant role in the economy as evidenced by the steady rise in the number of patent applications and the importance of patents in the financing of technology companies.

Public funding of research, as well as university research more generally, also play substantial roles in technological advance (National Science Board 1998). Focusing upon social norms motivating scientific research, particularly in university laboratories, Merton (1973) and Eisenberg (1987) suggest that intellectual property rights may undermine scientific progress by supplanting noneconomic reward structures such as publication, reputation, and professional advancement, discouraging the sharing of research and prompt disclosure, and shifting research away from basic research toward projects with more immediate commercial returns. Rosenberg and Nelson (1994) emphasize, however, that intellectual property fosters the diffusion of academic and medical discoveries beyond research institutions into the consumer marketplace. Furthermore, many research universities now protect and license technologies developed on their campuses as a means of funding research, although the implications for scientific progress are complex (Rai 1999).

Content and entertainment industries rely significantly upon copyright protection, although some scholars have questioned whether the incentives for creativity outweigh the economic inefficiency of monopoly exploitation (Breyer 1970). While there is little question that many expressive works would be created even in the absence of copyright protection and that publishers could appropriate some reward streams through first mover advantages and price discrimination of originals, there is little question that the elimination of copyright protection would dramatically alter the incentives motivating many authors and artists and undergirding the burgeoning content industries (Siwek 1999, Tyreman 1971).

3.3 Normative Analysis

The early formal models of intellectual property rights assumed that inventors conducted research in isolation on noncompeting projects (Nordhaus 1969). As such, the models focused on the optimal reward to induce particular innovations by a single inventor. As the historical and institutional literatures reflect, however, the actual environment for innovation is substantially more complex. In particular, different innovators (and firms) often compete to invent first, thus resulting in patent races. Economists developed richer, dynamic models in order to understand the positive and normative implications of rivalrous competition (Barzel 1968, Kamien and Schwartz 1982) as well as models examining design parameters beyond duration of protection such as scope and exclusivity of protection (Gilbert and Shapiro 1990, Klemperer 1990).

Building on Barzel's model of rivalrous competition which emphasizes the inefficient duplication of research, Kitch (1977) argues that the patent system in essence provides the first to claim an invention with a 'prospect' or 'development rights,' authorizing the patent holder to 'mine' the claim exclusively or license those better able to do so. Kitch defends this function of the patent system as promoting efficient (nonrivalrous) exploitation of innovation. Other scholars reach an opposite conclusion. Drawing on a wide range of empirical evidence, theoretical models of bounded rationality, and evidence suggesting high transaction costs in technology licensing, Merges and Nelson (1990) see vibrant competition as a more positive force in spurring invention, innovation, and diffusion of technology than coordinated development by a single prospector.

The modern literature has also developed models to study the implications of cumulative innovation (Scotchmer 1991) and network externalities on the design of intellectual property protection (Menell 1987).

3.4 Positive Analysis

Notwithstanding the serious questions regarding the efficacy and complexity of intellectual property rights, the past two decades has seen a veritable intellectual property 'land grab' as new opportunities to acquire 'intangible' property have emerged. Virtually every field of intellectual property law has seen an expansion of intellectual property rights—software and business method patents (Merges 1999), extension of the term of copyright protection (Patry 1997), anti-dilution law (Port 1994)—and new modes of protection—such as the right of publicity, semiconductor chip protection, and database rights—have emerged. The rapid expansion of intellectual property, often without adequate consideration of the diffuse interests of the public at large, suggests that traditional models of political economy apply reasonably well to the intellectual property arena (Menell 1994, Litman 1989, Ryan 1998, Samuelson 1996).

4. The Expanding Social, Philosophical, and Political Significance of Intellectual Property

Although the utilitarian framework has dominated the analysis of intellectual property, natural rights, personhood, libertarian, distributive justice, and other nonutilitarian theories have played a role in justifying intellectual property, especially with regard to the protection of literary and artistic expression (Gordon 1993, Hughes 1988, Waldron 1993). Continental European nations place greater emphasis than the USA or the UK on the inalienable moral rights of authors. Digital technology and the decoding of genetic structure have raised fundamental questions about the rationale, implications, and appropriate scope of intellectual property protection, pushing the debate further beyond the utilitarian frame of reference.

4.1 The Digital Revolution

As with the development of human language, writing, printing, and industrial methods of production, computer and network technology represent a fundamental advance in the ways in which humans societies function (Robertson 1998). In just a few years, the Internet has reshaped many industries by dramatically reducing information and transaction costs and creating new opportunities and pathways for commerce to flow (Evans and Wurster 1999, Shapiro and Varian 1999). The ease with which expressive works can be created and distributed in the information age pose threats to traditional copyright-based industries as well as great opportunities (National Research Council 2000). More fundamentally, the Internet has begun to transform social and political life by unleashing a tremendous wealth of easily accessible information and enabling social and political communities to form, function, and evolve in new ways.

Intellectual property plays many roles in the new information economy, from providing incentives for new enterprises to regulating the flow of information. As a result, intellectual property protection has increasingly come into conflict with freedom of speech, privacy, contract, and with other fundamental aspects of civil society, leading scholars to reexamine the role and implications of intellectual property (Lessig 1999, Netanel 1996).

4.2 Biotechnology

The decoding of genetic structure has led to the discovery of new ways of identifying and treating human disease, increasing agricultural and aquacultural yields and the nutritional content of foods, and reducing crop vulnerability to pests. These technological advances, however, have raised a host of troubling ethical, religious, social, economic, political, and ecological questions ranging from the morality of cloning humans to the ecological risks posed by the release of genetically modified organisms. At the most basic level, biotechnology research has led some to question the morality of uncritically promoting scientific discovery and the diffusion of new products. For example, the risk-taking and commercialization ethos of the intellectual property system has increasingly come into conflict with the precautionary principle of international and environmental law. The great commercial value of many biotechnology advances has spawned a growing list of quandaries for intellectual property law, including the existence and scope of property rights in human body parts, cell lines, and waste products, whether the human genome or particular sequences of genetic information should be protectable under intellectual property law, the ownership of intellectual property derived from human material, the ownership of intellectual property derived from native plants found in sovereign nations and the rights of indigenous people to use such intellectual property (Sedjo 1992), and the use and dissemination of genetic testing information.

See also: Bioethics: Examples from the Life Sciences; Innovation, Theory of; Intellectual Property, Concepts of; Intellectual Property Rights: Ethical Aspects; Property: Legal Aspects; Property, Philosophy of; Property Rights

Bibliography

Barzel Y 1968 Optimal timing of interventions. *Review of Economics and Statistics* **50**: 348–55

Breyer S G 1970 The uneasy case for copyright. *Harvard Law Review* **85**: 281–351

David P A 1985 *New Technology, Diffusion, Public Policy and Industrial Competitiveness*. Center for Economic Policy Research, Stanford University Press, Stanford, CA

Denison E F 1985 *Trends in American Economic Growth, 1929–1982*. Brookings Institution, Washington, DC

Economides N 1988 The economics of trademarks. *Trademark Reporter* **78**: 523–39

Eisenberg R S 1987 Proprietary rights and the norms of science in biotechnology research. *Yale Law Journal* **97**: 171–231

Evans P, Wurster T S 1999 *Blown to Bits: How the New Economics of Information Transforms Strategy*. Harvard Business School Press, Boston, MA

Goldstein P 1994 *Copyright's Highway: The Law and Lore of Copyright from Gutenberg to the Celestial Jukebox*. Hill and Wang, New York

Gordon W J 1993 A property right in self-expression: Equality and individualism in the natural law of intellectual property. *Yale Law Journal* **102**: 1533–609

Hughes J 1988 The philosophy of intellectual property. *Georgetown Law Journal* **77**: 287–366

Kitch E W 1977 The nature and function of the patent system. *Journal of Law and Economics* **20**: 265–90

Lemley M A, Menell P S, Merges R P, Samuelson P 2000 *Software and Internet Law*. Aspen Law & Business, Gaithersburg, MD

Lessig L 1999 *Code and Other Laws of Cyberspace*. Basic Books, New York

Levin R C, Klevorick, A K, Nelson R R, Winter S G 1987 Appropriating the returns from industrial R & D. *Brookings Papers on Economic Activity: Microeconomics* **3**: 783–820

Litman J D 1989 Copyright legislation and technological change. *Oregon Law Review* **68**: 275–361

Mansfield E 1986 Patents and innovation: An empirical study. *Management Science* **32**: 173–81

Menell P S 1987 Tailoring legal protection for computer software. *Stanford Law Review* **39**: 1329–72

Menell P S 1994 The challenges of reforming intellectual property protection for computer software. *Columbia Law Review* **94**: 2644–54

Menell P S 1999 Intellectual property: General theories. In: *Encyclopedia of Law & Economics*. Edward Elgar, Cheltenham

Merges R P 1995 The economic impact of intellectual property rights: An overview and guide. *Journal of Cultural Economics* **19**: 103–17

Merges R P 1999 As many as six impossible patents before breakfast: Property rights for business concepts and patent system reform. *Berkeley Technology Law Journal* **14**: 577–615

Merges R P, Menell P S, Lemley M A 2000 *Intellectual Property in the New Technological Age*. Aspen Law & Business, New York

Merges R P, Nelson R R 1990 On the complex economics of patent scope. *Columbia Law Review* **90**: 839–916

Merton R 1973 *The Sociology of Science*. University of Chicago Press, Chicago

National Research Council 2000 *The Digital Dilemma: Intellectual Property in the Information Age*. National Academy Press, Washington, DC

National Science Board 1998 *Science and Engineering Indicators—1998*. National Science Foundation, Arlington, VA

Netanel N W 1996 Copyright and a democratic civil society. *Yale Law Journal* **106**: 283–7

Nordhaus W D 1969 *Invention, Growth and Welfare: A Theoretical Treatment of Technological Change*. MIT Press, Cambridge, MA

Patry W 1997 The failure of the American copyright system: Protecting the idle rich. *Notre Dame Law Review* **72**: 907–33

Port K 1994 The 'unnatural' expansion of trademark rights: Is a federal dilution statute necessary? *Seton Hall Legislative Journal* **18**: 433–88
Rai A K 1999 Regulating scientific research: Intellectual property rights and the norms of science. *Northwestern University Law Review* **94**: 77–152
Robertson D S 1998 *The New Renaissance: Computers and the Next Level of Civilization*. Oxford University Press, New York
Rosenberg N, Nelson R R 1994 American universities and technical advance in industry. *Research Policy* **23**: 323–48
Ryan M P 1998 *Knowledge/Diplomacy: Global Competition and the Politics of Intellectual Property*. Brookings Institution, Washington, DC
Samuelson P 1996 The copyright grab. *Wired* **6** (January): 144
Saxenian A 1994 *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press, Cambridge, MA
Schwartzmann D 1976 *Innovation in the Pharmaceutical Industry*. Johns Hopkins University Press, Baltimore MD
Sedjo R A 1992 Property rights, genetic resources and biotechnological change. *Journal of Law and Economics* **35**: 199–213
Shapiro C, Varian H R 1999 *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, Boston, MA
Scotchmer S 1991 'Standing on the shoulders of giants': Cumulative Research and the Patent Law. *Journal of Economic Perspectives* **5**: 29–41
Siwek S E 1999 *Copyright Industries in the US Economy: The 1999 Report*. International Intellectual Property Alliance (IIPA), Washington, DC
Taylor C, Silberston Z A 1973 *The Economic Impact of the Patent System*. Cambridge University Press, London
Tyreman B W 1971 The economic rationale for copyright protection for published books: A reply to Professor Breyer. *UCLA Law Review* **18**: 1100–25
Waldron J 1993 From authors to copiers: Individual rights and social values in intellectual property. *Chicago-Kent Law Review* **68**: 841–87

P. S. Menell

Intellectual Property Rights: Ethical Aspects

'Intellectual property rights' are generally understood as an instrument of economic policy providing incentives for and rewarding a broad range of useful human creativity. Despite this prevailing economic rationale, the roots of, the justification for, and some limitations of exploitation of these rights are reasoned ethically. They will be examined in the context of the notion, economic importance, and implications deriving from scientific and international policy developments.

1. *The Notion of Intellectual Property Rights*

As a generic term, intellectual property rights stand for: patents relating to inventions in all fields of human endeavor; copyright covering literary, artistic, and scientific works; neighboring rights dealing with performances of performing artists, phonograms, and broadcasts; trademarks, service marks, and commercial names concerned with signs distinguishing goods and services as well as economic subjects; industrial design protecting novel and original ornamental or aesthetic aspects of a useful article; or plant breeders' rights relating to new varieties of plants. Conceptually, intellectual property rights can be understood as rule-governed privileges that regulate the ownership and exploitation of objects which are to be characterized as intangible or incorporeal, thus abstract or intellectual. These objects are nonexclusive, since they can be in many places at once, and can be used at the same time without being consumed by their use. The possession or use of such an object by one person thus does not preclude others from possessing or using it as well. Despite their common fundamental character, i.e., being rights in intangible, abstract objects, and being territorially limited, they differ from each other in many legal details and character.

Whereas a patent also gives its owner rights against the independent creator of a similar invention, copyright confers rights against copying but does not affect the independent creation of a similar work. In other words, while patents restrict the actual usage of an idea, copyrights restrict only the copying of an expression of an idea. Necessarily, the period of protection, i.e., during which the rightful owner can prohibit unauthorized use, is tailored according to the specifics of the respective rights—for instance, 20 years from filing the application in case of patents, 50 or 70 years after the death of the author in case of copyright, or renewable at will, i.e., indefinite, in case of trademarks. However, neither patents, nor any other intellectual property right, confer on their owner a license to use the patented invention, copyrighted work, etc. Such use is prohibited whenever it would violate, for instance, laws dealing with public health, safety, environmental protection or certain ethical standards, or would infringe the rights of third parties.

2. *Economic Importance*

Due to enormous advances in science and technology, in particular in communication and information technologies, biotechnology, and life sciences in general, on the one hand, and the global economic policy developments following the conclusion of the GATT-Uruguay Round and the establishment of the World Trade Organization in 1994, on the other, intellectual property rights have gained most remarkable economic importance. They have penetrated, or are about to penetrate, new areas of economy, such as the service sector, life sciences, and agriculture. Patents have been granted for instance, for microorganisms, plants and animals, genes, as well as computer programs and

business methods. Universities and other academic institutions, primarily aimed at generating and disseminating new knowledge, have also not remained unaffected. Consequently, the new, knowledge-based economy has been characterized as 'intellectual property economy' or 'intellectual capitalism.' This trend is reflected, for instance, by the number of patent applications filed and patents issued, which in the USA almost doubled from 1988 to 1998, from 160,000 patents to 260,000 patent applications; by the amount of royalties paid for patent licenses, which increased in the USA from US$3 billion in 1980 to over US$100 billion in 1997; or by the fact that commodities constituted 62 percent of the market value of the manufacturing industry in the USA in 1980, but less than 30 percent in 1998. Also remarkable are the numbers of patents issued to US universities, which went up from 177 in 1974 to 2436 in 1997, the year in which US universities filed some 6,000 patent applications and granted around 3,000 patent licences.

3. Historic Roots and Rationale

The historic roots of patents and copyrights, as the two forms of intellectual property with the longest tradition, go back to the systems of inventors' and printers' privileges at the close of the Middle Ages. Inventors' privileges formed part of the industrial policy of regional sovereigns, being aimed at promoting domestic crafts and industry, and at transferring to their own country the production of goods which had previously been imported from abroad; thus, as expressed in a grant issued by Edward III in 1331, '*instruendi et informandi causa*.' The first statutory regulation, the Venetian Statute of Inventors, of 1474, with its time-limited protection of inventions against imitation, was also intended to stimulate 'deep thinkers and ingenious inventors from other regions to exercise their activity to the advantage of Venice.' Similar economic rationale is to be observed in the English Statute of Monopolies of 1624, under which the grant of patent monopolies of 14 years' duration was restricted to the 'real and first inventor,' and was conditional upon its not being contrary to law or 'mischievous to the state' in any way.

Following the invention of the press, the first printers' privileges were granted in the second half of the fifteenth century in Venice, later followed by Milan, and at the beginning of the sixteenth century by other Italian cities, France, Germany, and England. Whereas in Venice and France some of these privileges were granted to authors, the great majority of privileges were for publishers, or, as in England, for stationers, who acquired a work from its author and organized its printing and sale. In England, where the stationers in 1534 gained protection secured by the Crown against the importation of foreign books, and in 1556 a Charter from Queen Mary, the privileges were combined with censorship over all publications, by mandatory registration with the Stationers' Company. Although the first Copyright Act, issued in 1710 in the reign of Queen Anne (the Statute of Anne), conferred the 'sole right and liberty for printing books' for 14 years to authors and their assigns, it did not change fundamentally the economic rationale of protection.

By naming booksellers and printers as falling among the author's assigns, it demonstrated that the right 'stemmed from commercial exploitation rather than literary creation pure and simple' (Cornish 1999). Despite the remarkable influence which John Locke's theory—that all persons have a natural right to hold property, particularly the fruits of their own labour—played from the very beginning in justifying intellectual property, the economic policy of legislation has remained utilitarian. This is well demonstrated by Art. 1, Sect. 8 of the US Constitution of 1787, providing that 'The Congress shall have power…(8) To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries.' This is equally true for the first modern patent laws. which were passed in 1790 in the United States and 1791 in France, although they strengthened the position of the inventor and awarded him a property right to his ideas. The same rationale controlled also the first US Copyright Statute of 1790.

The doctrine of intellectual property developed in eighteenth-century France, advocated primarily by D'Héricourt, Diderot, and Voltaire as a weapon against extended privileges of publishers. A first true statutory shift of the emphasis from the publisher to the author can be observed in the Decrees of the French revolutionary legislation of 1791 and 1793, which extended the term of protection to the entire life of the author and beyond; first for five, then for 10 years after the author's death. In the course of the nineteenth century France further extended the term of protection to cover also the lifespan of the author's widow, and 20 (later increased to 30) years for the benefit of the children. Eventually, in 1866, the duration of French copyright was extended generally to 50 years after the author's death. Even this shift of emphasis, however, did not add to the economic rights of the author any 'personality right' or *droit moral*, which had been first mentioned by Immanuel Kant. Kant stated that the right of the author 'is not a right in an object (*in re*), i.e., in the copy (for the owner can burn this up before the author's eyes), but an innate right, inherent in his own person, which implies the possibility of opposing the attempt of another person to compel him to speak against his will.' Kant characterized copyright as a supremely personal right (*jus personalissimum*) (Strömholm 1983).

French and German legal doctrine and French case law in the second half of the nineteenth century

eventually paved the way for *droit moral*: first, into the case law, and then into copyright statutes of the principal Continental countries. Although *droit moral* was introduced into the Berne Convention for the Protection of Literary and Artistic Works of 1886, as revised in 1928 in Rome, with a new Article 6 bis—which at present guarantees an author independent economic rights, and even after their assignment, 'during his whole lifetime the right to claim authorship of the work and to object to any distortion, mutilation or other alterations thereof, or any other action in relation to the said work, which would be prejudicial to his honour or reputation'—the common law countries of the United Kingdom and USA remained reserved in this respect. In general, therefore, one distinguishes between 'author's rights' of civil law countries, with their emphasis on the principles of natural rights, and the 'copyright,' rooted in the idea of utilitarianism, of common law countries. However, on balance, it was neither human rights nor copyright doctrine that decisively shaped the law, but economic interests of copyright owners, because: 'In the world of copyright policymaking, a Canadian book publisher has far more in common with a Japanese book publisher than it does with a Canadian librarian who wants to make free copies for library patrons' (Goldstein 2001).

4. *International Legal Developments*

The very nature of the objects to which intellectual property rights relate, and the development of international trade combined with the progress of transportation and communication technologies, already by the nineteenth century revealed the need for international instruments for protecting intellectual property rights. What started in 1883 with the adoption of the Paris Convention for the Protection of Industrial Property, followed by the Berne Convention for the Protection of Literary and Artistic Works in 1886—which established the principles of national treatment or nondiscrimination and minimum rights as binding standards for the contracting states–subsequently developed into a true network of international conventions and treaties covering the entire range of intellectual property rights under the administration of the World Intellectual Property Organization (WIPO).

However, membership in many important conventions long remained limited, and minimum rights incomplete or, as in the case of patents, nearly nonexistent. The turning point came when, during the GATT-Uruguay Round negotiations, issues of international trade, including membership in the World Trade Organization (WTO), were linked directly to an effective protection of intellectual property rights, and the International Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS) was adopted in 1994. Under this new umbrella for international protection of intellectual property rights, all members of WTO accepted unified high standards of protection, including protection requirements, content, and limitations and duration of rights, as well as their enforcement. Developing countries, countries in transition and the least-developed countries were granted transition periods.

5. *Old and New Ethical Aspects in Intellectual Property Rights*

Traditionally, ethical aspects in intellectual property rights have been discussed in the broad context of justification of these rights. Lockean theory of natural rights to the fruits of one's labour, the doctrine of intellectual property, as well as Hegel's personality justification suggesting that the best way of progressing science and the arts is to protect scientists and artists from theft, so as to become the basis for learning by others, are such examples and had some lasting influence on legal developments, reflected, *inter alia*, in broad acceptance of *droit moral*. Modern critics of intellectual property rights also put emphasis on ethical aspects when, for instance, claiming that increasing scope of patentable subject matter or legislatively creating new forms of abstract objects, such as plant variety rights, constitutes effectively the creation of capital, which has the danger that it can act as an enormous power resource for a select few. Furthermore, a global protectionist scheme for intellectual property would help to promote the formation of global factions, resulting in the danger of global rent seeking, and perhaps providing temptation to multinational elites to increase their profits through the simple stratagem of persuading a supranational body to elevate levels of protection for abstract objects already in existence (Drahos 1996). With such reasoning, neither the ethics of the controlling doctrine nor the ethical aspect of attempts to take care of the developing world and the unprivileged in general can be overlooked. The same is true for the debate on intellectual property rights in genetic resources and benefit sharing. Although clear empirical evidence whether intellectual property rights confer a net benefit or a net loss upon society is still missing modern law makers seemingly accept a limited and flexible intellectual property rights scheme as a reasonable and, as revealed by the adoption of TRIPS, internationally necessary effort to maximize benefits while minimizing costs. This attitude is demonstrated by the adoption of statutory rules or acceptance of case law, creating new objects such as DNA sequences, monoclonal antibodies, or for instance, business methods as subject matter of protection. In their overall positive attitude toward exclusive intellectual property rights modern law-makers might have been in part influenced by the negative experience with some alterna-

tive protection found in the former socialist states such as 'investors' certificate' which did not provide for exclusivity and had no beneficial macro-economic effects.

Special attention is paid to ethical aspects when patentability of inventions is at hand, the commercial exploitation of which would be contrary to '*ordre public*' or 'morality.' Under the TRIPS standards, members of the WTO 'may exclude from patentability inventions, the prevention within their territory of the commercial exploitation of which is necessary to protect *ordre public* or *morality*, including to protect human, animal or plant life or health or to avoid serious prejudice to the environment, provided that such exclusion is not made merely because the exploitation is prohibited by their law.' Thus such exclusions may only apply for ethical reasons, first, if commercial exploitation is prohibited, and, second, if they are prohibited for specifically qualified reasons, which need not necessarily be set out in statutes, however. European patent law reflects these ethical considerations by excluding from patentability the human body, at various stages of its formation and development, including germ cells, processes for cloning human beings, processes for modifying the germ line genetic identity of human beings, uses of human embryos for industrial or commercial purposes, processes for modifying the genetic identity of animals, which are likely to cause them suffering without any substantial benefit to man or animal, and in general all therapeutic methods. By declaring, explicitly in principle, that biological material, even that of human origin (provided the prior informed consent of the source person is obtained), as well as animals and plants (except varieties) is patentable, European law makers demonstrated clearly that they do not share respective ethical objections against patentability of life forms. This ethically reasoned approach, which differs in particular from that applied in the USA, where no such exclusions exist, however itself raises ethical concerns. On the one hand, it seems consistent that inventions, which for reasons of ethics may not be commercialized, should also not be patented, although an issue of a patent does not confirm more than that the usual patentability requirements of novelty, inventiveness (nonobviousness), and industrial applicability (utility) were met and the invention adequately disclosed. It does not contain any other value judgment. On the other hand, the invention under such circumstances cannot be commercialized in any case, as long as its exploitation is prohibited. However, if such a ban were to be lifted, anybody could use the invention, and the inventor would remain empty-handed. As history demonstrates—for instance, with contraceptives—this is not a far-fetched scenario and could easily happen again, e.g., with regard to specific uses of stem-cell technology, germ line gene therapy, etc. Empty-handed inventors, and at least to the same extent the lack of incentives for investment in potentially useful research, may be viewed as ethically doubtful results.

See also: Ethics and Values; Ethics in Organizations, Psychology of; Intellectual Property, Concepts of; Intellectual Property: Legal Aspects; Property: Legal Aspects; Property, Philosophy of; Property Rights; Research Conduct: Ethical Codes

Bibliography

Beier F K 1980 The significance of the patent system for technical, economic and social progress. *International Review of Industrial Property and Copyright Law* **11**: 563–84

Cornish W R 1999 *Intellectual Property: Patents, Copyright, Trade Marks and Allied Rights*, 4th edn. Sweet & Maxwell, London

Drahos P 1996 *A Philosophy of Intellectual Property*. Dartmouth, Aldershot, UK

Goldstein P 2001 *International Copyright—Principles, Law and Practice*. Oxford University Press, Oxford, UK

Grandstrand O 1999 *The Economics and Management of Intellectual Property—Towards Intellectual Capitalism*. Edward Elgar, Cheltenham, UK

Landes W M, Posner R A 1989 An economic analysis of copyright law. *Journal of Legal Studies* **18**: 325–53

Kitch E W 1977 The nature and function of the patent system. *Journal of Law and Economics* **20**: 265–90

Machlup F 1959 *An Economic Review of the Patent System*, Study No. 15, Subcommittee on Patents, Trademarks and Copyrights, Senate Committee on the Judiciary, 85th Congress, 2nd Session. U.S. Govt Printing Office, Washington, DC

Moore A D (ed.) 1997 *Intellectual Property—Moral, Legal and International Dilemmas*. Rowman & Littlefield, Lanham, MD

Schricker G (ed.) 1999 *Urheberrecht—Kommentar*, 2nd edn. Beck, Munich, Germany

Strömholm S 1983 Droit Moral—The international and comparative scene from a Scandinavian viewpoint. *International Review of Industrial Property and Copyright Law* **14**: 1–42

Troller A 1983 *Immaterialgüterrecht*, 3rd edn. (2 vols.) Helbing & Liechtenhahn, Basel, Switzerland

Ulmer E 1980 *Urheber- und Verlagsrecht*, 3rd edn. Springer-Verlag, Berlin

J. Straus

Intellectual Transfer in the Social Sciences

1. Intellectual Transfer as a Basic Process

Intellectual transfer is one of the least researched and most common processes in the social sciences. It is defined as the importation into a social science discipline of formal procedures or substantive insights developed in another social science discipline or in the natural sciences, in such a way that those procedures

or insights not only provide suggestive metaphor or imagery, but define a coherent conceptual universe. Such importation may result from a conscious search for more rigorous modes of analysis than those which the practitioners of a discipline currently have available to them; the geographical displacement of a significant body of scholars; an intellectual milieu providing for interdisciplinary cross-fertilization; or the actions of scholars acting as intellectual import–export agents. For much of the twentieth century the first mode of importation was the most common, exemplified by the borrowings made by sociology and anthropology of procedures from biology, physics, cybernetics, and linguistics. The second is best exemplified by the influence of psychoanalysis on a range of disciplines. The last three decades of the twentieth century saw the emergence of a different form of intellectual transfer—interdisciplinary hybridization—the culmination of which is the proliferation of departments of cultural and other 'studies,' in which the 'same' object of inquiry may be addressed by a range of—often epistemologically contradictory—procedures.

2. *Transfer of Procedures*

Whether the classics of modern social science owe their achievements to a spirit of nineteenth century humanism which pervaded them all (Nisbet 1967) or to their location between literary and scientific modes of perception (Lepenies 1988), the subsequent history of sociology, anthropology, economics, political science, and psychology is pervaded by the widespread if intermittent quest for methods and results whose rigor would be comparable to those of the natural sciences. This has entailed a readiness among social scientists to look beyond their own disciplinary boundaries not only for (philosophical) clarification of existing assumptions and procedures, but also for tools which would place social science knowledge on firmer foundations and generate a significant corpus of findings. Within sociology, anthropology and political science, functionalism, structuralism, systems theory, and rational choice theory, are only the most prominent results of intellectual transfer based upon this foundationalist impulse. Functionalism's conception of society as an organic unity is rooted in conceptions of the organism drawn from biology; structuralism's exploration of social codes and messages deploys methods first developed in linguistics; the systems theory of the later Talcott Parsons or Niklas Luhmann is unthinkable without the conceptual apparatus of cybernetics; rational choice theory applies individualist models of utility maximization drawn from economics to all domains of the social and political world; and within mainstream empirical social science the persistent classification and subclassification of social phenomena owes much to the taxonomic procedures of eighteenth century natural history (Foucault 1974). More recent moves to apply chaos theory to society have added a twist to this quest for foundationalist respectability. Here, theories of uncertainty and nonlinear dynamics appear to warrant claims that the natural and social worlds are both marked by the absence of law-like regularity.

Since the intellectual transfer of a formal apparatus entails an initial abstraction of its core components, their uprooting from their original home, and translation into an idiom recognizable to the social science discipline in which they are to operate, the blessings of these efforts have been mixed. All display a tension between the descriptive power of a rigorous theoretical apparatus and the need to adapt it to circumstance, most notably the absence of an isomorphism between the natural and social worlds. A number of consequences of this may be noted:

(a) Techniques and conceptual schemes imported into the social sciences from outside have produced genuine semantic innovation and enhanced the descriptive vocabularies of practicing social scientists.

(b) At the same time, it remains doubtful whether they have added significantly to collective social scientific knowledge of the sources of social order, social change, and human motivation that emerged in the late nineteenth century classics.

(c) While intellectual transfer of this type has increased the theoretical sophistication and rigor with which 'results' or 'findings' are presented, it has produced little if any innovation in empirical research techniques or methodologies.

(d) In cases of a more explicit embrace of natural science models social scientists' understanding of natural science knowledge may be incomplete/inaccurate, as witnessed by the 'Sokal affair' of 1998.

(e) The very success of well-defined approaches such as systems theory or rational choice theory—descriptive power, volume of findings, size of intellectual following—has produced paradigm multiplication—intellectual closure, the progressive internal differentiation of disciplines and disciplinary vocabularies—rather than disciplinary unity.

As a result of this last consequence, histories of disciplines such as sociology or anthropology are forced to combine an in-house relativism/ecumenism with a vaguely felt evolutionism. Here, the more extensive the importation of natural scientific paradigms into social science disciplines, the less any of them taken as a single entity appears describable through Kuhn or Lakatos' accounts of the development of natural science disciplines: scientific revolutions, or the growth of knowledge through criticism.

3. *Transfer of Substantive Insight*

In contrast to the considerable development of formal techniques and conceptual vocabularies in the twentieth century, the history of substantive innovation in

the social sciences is less impressive. Since the turn of the nineteenth century the anthropological claims which grounded the identities of the various disciplines or subdisciplines at their inception have proved remarkably robust. Within social science disciplines with more porous boundaries these models of humankind have continued to define the terms in which the nature of social being is debated and researched. The images of humans as producing animals, as the maximizers of utility, as altruists, as fabricators of a meaningful universe, still dominate mainstream social science at the beginning of the twenty-first century. Some notable additions, however, in which a 'migration of insight' has occurred, may be mentioned. Most marked are the claims of psychoanalysis concerning the basic structure of the psyche, which have exerted widespread influence on a range of disciplines, particularly sociology and anthropology. The work of the archaeologist Uexküll on the relationship between humans and the environment influenced the theory of social institutions developed by a whole generation of conservative German social scientists from the 1930s to the 1950s.

4. *Interdisciplinarity*

If one product of open disciplinary boundaries is a form of intellectual transfer based on a quest for certainty, foundation, and intellectual closure, then another is interdisciplinarity, the mutual influence of two or more disciplines. A major stimulus to this process has been a change in the status of natural science knowledge from being an explicit model or implicit standard for social scientific inquiry, to being a topic of it, subject to the same scrutiny as other realms of human action. This breaking out of the shadows of natural science has left the social science disciplines both on their own and with room to explore their relationship to humanities disciplines such as literary criticism. Three forms of interdisciplinarity may be distinguished:

(a) The combination—common throughout the history of the social sciences—of the major building blocks of two or more disciplines in order to throw light on a clearly defined element of the human condition. Examples of such an approach are the combination of literature and anthropology in Rene Girard's account of violence; of sociology and economics in Talcott Parsons' account of the nature of action; and of sociology and history in Barrington Moore and Skocpol's accounts of democracy and revolution.

(b) The eclectic hybridizing of a variety of discipline-specific conceptual devices and methods as part of an effort to generate a new field of inquiry. Notable examples are 'area studies' which developed in the 1950s and 1960s out of an effort to apply Western social science to Third World and developing societies, and the contemporary phenomenon of cultural studies, which has mushroomed in new universities throughout the Western world.

(c) The utilization of the results and findings of the different social science disciplines in order to develop a comprehensive anthropology (Gehlen 1988).

5. *Mechanisms of Transfer*

The transfer of formal or systematic procedures into a social science discipline depends partly upon the degree of systematization already prevailing within it. Social science in France and Germany—where all-embracing terms such as *Wissenschaft* know no English equivalent—developed at its inception a series of systematic, conceptually rigorous theories of social wholes. The relative immunity of these social sciences to pressure from natural science or the humanities may be attributed, ironically enough, to the fact that many of their founders were trained in jurisprudential modes of reasoning. It is arguable that by contrast, in the English-speaking world the division between the arts and natural sciences has made the social sciences more open, yet also more vulnerable to the rival claims of both a natural science model of good knowledge and a hermeneutic or culturalist inquiry based upon procedures drawn from literary criticism or, after World War II, semiotics. The coexistence of rational choice theory and postmodernism within the same departments of sociology is an index of this.

Yet social science in the English-speaking world is not wholly torn between the importation of scientistic models of explanation and the adoption of literary models of textualist interpretation, sharing much with the social science of the continental European tradition. That this is so owes much to the exile and migration of that tradition's representatives and their presence within Anglo-Saxon higher educational institutions. If Anglo-Saxon sociology, political science, psychology, or anthropology have held their own as *sui generis* and self-confident disciplines, or if political sociology or economic anthropology have later produced systematic interdisciplinary inquiry, it is on the basis of the work of Marx, Weber, Durkheim, Schumpeter, Malinowski, Freud, and Levi-Strauss, much of it mediated by the European intellectual diaspora of the 1920s and 1930s (Anderson 1992). By the same token, the Anglo-Saxon world has become the protective home of a generic 'European' or 'continental' tradition of social science inquiry, based upon and promoting an intellectual transfer between national traditions which contrasts with the entrenchment and mutual suspicion frequently displayed by their native practitioners.

6. *The Future of Intellectual Transfer*

In the 30 years following Lakatos and Musgrave's (1970) *Criticism and the Growth of Knowledge*, the

philosophy of the social sciences has lost much of its appeal to the social sciences themselves. The mirror in which reflection on the identity of the social sciences occurs is no longer that of the epistemological status of social scientific knowledge but, increasingly, that of the social responsibility of both natural and social scientific practitioners (Wallerstein 1996). Natural science models continue to prove attractive, but in doing so contribute only to the internal differentiation of disciplines. The more significant contemporary version of intellectual transfer for the foreseeable future will continue to be that which produces interdisciplinary departments of 'X studies' drawing on a variety of epistemological assumptions and asking few questions concerning the status of the knowledge produced therein. The reflexive apotheosis of this development is the emergence of departments of 'interdisciplinary studies.'

See also: Actor Network Theory; Biology's Influence on Sociology: Human Sociobiology; Diffusion, Sociology of; Disciplines, History of, in the Social Sciences; History and the Social Sciences; Hybridity; Paradigms in the Social Sciences; Science, Sociology of; Scientific Controversies

Bibliography

Anderson P 1992 Components of the national culture. In: *English Questions*, Verso, London

Crebogi C, Yorke J 1997 *The Impact of Chaos on Science and Society*. United Nations University Press, Tokyo and New York

Feyerabend P 1984 *Wissenschaft als Kunst*. Suhrkamp, Frankfurt am Main, Germany

Foucault M 1974 *The Order of Things*. Tavistock, London

Gehlen A 1988 *Man*. Columbia University Press, New York

Lakatos I, Musgrave A (eds.) 1970 *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK

Lepenies W 1988 *Between Literature and Science*. Cambridge University Press, Cambridge, UK

Levine D N 1995 *Visions of the Sociological Tradition*. University of Chicago Press, Chicago, IL

Martins H 1972 The Kuhnian revolution and its implications for sociology. In: Nossiter J (ed.) *Imagination and Precision in the Social Sciences*. Faber, London

Merquior J G 1986 *From Prague to Paris*. Verso, London

Nisbet R 1967 *The Sociological Tradition*. Heinemann, London

Sokal A, Bricmont J 1998 *Intellectual Impostures*. Profile Books, London

Wallerstein I (Gulbenkian Commission on the Restructuring of the Social Sciences.) 1996 *Open the Social Sciences*. Stanford University Press, Stanford, CA

C. Turner

Intellectuals: History of the Concept

The noun 'intellectuels' appeared in France in the 1890s and was largely diffused thanks to the polemics around the Dreyfus Affair (in particular the so-called '*Manifeste des intellectuels*,' published after the famous *J'accuse* paper by Zola issued on January 13, 1898, in *L'Aurore*). This diffusion not only occurred in France but also, sooner or later, in the whole of Europe and even in America (North and South) (Drouin 1994). There are two exceptions to this large diffusion of the new concept: Germany and Russia where there existed former and akin notions, *Intelligenz* and *intelligentsia*, already used as social or political denominations (Müller 1971). Previous uses of the English translation, 'intellectuals,' are attested too before the 1890s, but they seem to have been too rare to have known a broad social circulation (Williams 1976). In the same way, in America, 'intellectuals' became a common notion only in connection with newspaper commentaries about the Dreyfus Affair (Bender 1987).

To understand why this new terminology received a European or even international diffusion, it is necessary to recall the issues of this historical moment. What was at stake in this crisis was not only a political problem but also the affirmation of a new group, defender of universal values against the reason of State (Charle 1990). These values justified the fact that writers, artists, scholars, students, members of liberal professions, and so on, intervened on a collective basis in the political debate, although they were not themselves, for the main part, professional politicians. The other specificity of this moment was that, in other countries, this same cause or other similar ones favored the intervention of intellectuals, but generally in various patterns and with different contents. Since that period, social sciences have debated at a theoretical level, in order to find universal characteristics underlying this new vocabulary (Shils 1972). As a matter of fact, the emergence of intellectuals cannot be assimilated to the apparition of a new permanent social group as some historians or sociologists suggest regularly, wrongly mixing intellectuals, intelligentsia and professions (Bell 1973, Perkin 1989). To know in reality what is meant by the words, intellectuals, *intellectuels, Intellektuelle, intellettuali, intelectuales,* and so on, it is necessary to define them within their specific cultural, social, and historical contexts.

1. The French Case

In France, even before the Dreyfus Affair, the neologism '*intellectuels*' was used by avant-garde circles as a social mark. The *intellectuel* was some kind of mandarin, who despised politics and wanted to distinguish himself from middle class, dominant writers and academics. *Intellectuel* was a sort of superlative of what Flaubert meant by the word '*artiste*' (Bourdieu 1992). But, since the neologism was widely used during the Dreyfus Affair, the initial cultural meaning was replaced by a strong emphasis laid upon the political

acception. In the first phase of the Dreyfus affair, the *intellectuels* were an equivalent of '*dreyfusards*' and, afterwards, when *antidreyfusard* intellectuals also intervened in a collective way, the word *intellectuels* began to define a special category of people who defended political positions based on arguments of social authority, i.e., their competence as thinkers, historians, scientists, professors, writers, or artists.

In France, the birth of *intellectuels* may be first explained by a growing inadequacy of former cultural patterns confronted to the buoyant cultural expansion of the last decades of the nineteenth century. Intellectual professions, far more numerous now, defended their social and symbolic status at once so that collective attitudes appeared, breaking with older individualistic habits. But this defense may be argued on two different grounds: with intellectual and pure values or on professional and pragmatic issues. Against avant-garde writers or academics, generally militating for the first option, professional associations appeared during the same period that were more attached to material interests. The first use of the term *intellectuels* was reserved for the first type of elitist fraternity. It was to be the germ of its political transformation as for the Russian *intelligentsia* diffused a bit earlier in the 1870s. As in Russia too, reformed universities played a major role in the process of emergence and mobilization of nonconformist intellectuals.

The second paradoxical factor of ideological and political change, was the early crisis of parliamentary democracy in France. After the stabilization of the Third Republic, apolitism predominated among intellectuals as if the end of history was already reached. A new politicization occurred with the crisis of official parties and the emergence of extremist factions, which had a large echo among avant-garde writers (in particular anarchism and, to a lesser degree, socialism). This new trend prepared what is the specificity of the Dreyfus Affair, the invention of a new relationship with politics outside the official political scene. Intellectuals in this new terminology pretended to practice politics in a different way. This was possible because the legitimacy of Republican elite laid on the same bases as the legitimacy of intellectuals themselves, i.e., upon merit and individual talent. But in so far as these elite appeared incompetent or corrupted after different crises and scandals (Boulangism, Panama scandal, and so on), intellectuals preserved from these faults might pretend to offer an alternative elite necessary to lead any genuine democracy.

Students, avant-garde writers, and even younger generations of academics, before and during the Dreyfus Affair, expressed these new revendications of being the true representatives of the people against politicians. The State itself, with its growing intolerance towards literary innovations or extremist parties (especially anarchism) contributed to the mobilization of authors against juridical prosecutions through collective manifestos just before the Dreyfus affair.

The Dreyfus affair presented both a true continuity with preceding years and a break due to the scale of mobilization. Afterwards, the main following mobilizations of intellectuals obeyed the same collective rites and values (Ory and Sirinelli 1986). Its founding importance was to prove that this type of mobilization might lead to real political consequences. This twofold mobilization (of *dreyfusards* and *antidreyfusards*) was new and defined, on both sides of the political scene, a general definition of intellectuals, which was not limited to leftist intellectuals. On the contrary in other countries, the equivalents of *intellectuels* are, generally speaking, confined to one side of the political spectrum.

2. *The Peculiarities of English Intellectuals*

In contrast with France, it is generally argued that no intellectuals exist at all, in the continental acception, in the UK. Since around 1980, British historians and sociologists reacted against this strong and vulgarized anti-intellectual bias. Some authors speak of an intelligentsia (Allen 1986, Heyck 1982), i.e., an elitist avant-garde, others of '*public moralists*,' which enhances the role of dominant and academic intellectuals (Collini 1991) or of a '*professional class*' which assimilates intellectual professions to a new class (Perkin 1989). A comparative approach shows that two specific factors may explain the strong difference with the French situation, in spite of the proximity of economic and political conditions in intellectual life, the persistent elitism of English academic life and the relative proximity between intellectual professions and political elite. Established elites, even if they were obliged to reform and enlarge the political system at the end of the nineteenth century, were not contested as an illegitimate power elite, to the degree they were in France. In fact, English dominant intellectuals mainly shared the same values and background as gentlemen and political leaders because they were largely issued from the same public schools and universities.

In front of these dominant intellectuals, appeared, in the last decades of the twentieth century, new types of intellectuals who presented outsider profiles. The best known were the Fabians who could not attend the best colleges and universities and had to find their way through journalism, literature, new academic institutions (e.g., the London School of Economics) or militant politics. But these avant-gardes were very different from contemporary French avant-gardes. They limited their intervention to one main field: for the Fabians, social and political questions, for aesthetes, aesthetic life, and so on. They pretended to create a voluntary structure and to indirectly influence the official sphere indirectly, not at all to destroy or affront it directly.

Even when mobilization occurred on a larger scale, as during the Boers war, English intellectuals used the official means of actions and respected legal frameworks. Finally, the main difference lay in the very different function of the State in England and in Continental Europe. In England, militant intellectuals endeavored to enlarge its role to correct social injustice, whereas, in France and even more in Germany, intellectuals tried first to weaken its authoritarian trends.

3. The German Case

Even if the word *Intellektuelle*, derived from French, continues almost until now to bear a derogatory nuance (Bering 1978), the German genuine equivalents *Intelligenz*, *Gebildete*, and *Geistige*, have been used for a long time but do not imply, as the French term, the same political or social behaviors. This early appearance of the question of 'intellectuals' may be shown through the recurrent discussion about the academic proletariate (first in the *Vormärz* period, then in the 1880s, and finally in the Weimar period: Titze 1990), with the ideological debates within the social-democratic party about the place of *Intelligenz* (Gilcher-Holthey 1986), and also with the *Antisemitimusstreit* in 1879. This last famous polemic about the role of Jews in German society between the conservative historian Treitschke and his liberal colleague Mommsen, a former Forty-eighter (former participant to revolutionary troubles of 1848 in Germany), appears very near, in its arguments, to the debate between opposing French *dreyfusards* and *antidreyfusards*. The rights of minorities and, in particular, of individuals and Jews, lay at the center. Other affairs concerning academic freedom (Arons case or Spahn case), or freedom of creation, like the mobilization against the lex Heinze (1900), show too that the debates about intellectual autonomy were as crucial in Germany as in France and that they succeeded several times to mobilize some groups of intellectuals (Charle 1996).

But, in all cases, mobilization was limited to specific groups and to particular issues which did not put into question the whole structure of State itself as in France. A mere political explanation (an Empire opposed to a Republic) does not suffice. What was specific and new in the Dreyfus case was the convergence of different intellectual groups about common values. In Germany, the corporatist ethos remained stronger even about general issues; free intellectuals and State intellectuals (mainly university professors) despised each other. Academics began at that time to live apart from the political sphere and preferred a general cultural function as State or *Bildung*'s defenders through different associations. Only a small minority of free intellectuals and very few in the Academe put into question dominant elite or national causes.

The *Gebildete* assumed that they represented the true public opinion and that they were the best interpreters of general causes, but they intended to remain in their own field in order to serve their country best. This German intellectual and geographical fragmentation hindered the linkage between local or professional struggles for autonomy (Engelhardt 1986, Hübinger and Mommsen 1993, Ringer 1969, 1992).

4. Southern Europe

In Spain as in Italy, the local equivalent of *intellectuels* seemed also to be in use in the 1890s in connection with, or even before, the Dreyfus case. The French example was very influential for the Spanish and Italian intellectuals because French cultural influence, in the two peninsulas, was already very strong since the French Revolution. Also, the inner social and political situations of Spanish and Italian intellectuals presented some analogies with the French context. As their French homologues, they thought that their countries went through a deep crisis (economic backwardness, military defeat in Spain, emigration, social riots and parliamentary corruption in Italy), which implied some sort of public intervention to find solutions. The strong anticlericalism and antimilitarism, the link between intellectuals and extreme left movements, the emergence of a new nationalism in both countries, also recalled the French debates at the turn of the twentieth century (Serrano and Salaün 1988). Obvious differences also existed: the weaker public audience of intellectuals depending on the cultural backwardness of popular classes (high level of analphabetism) and the persistence of a large sector of opinion among Catholics, hostile to the cultural inheritance of Enlightenment, very influent in both countries, and an overproduction of *laureati* in Italy which could explain a strong commitment of academics to extremist parties (Barbagli 1974, 1982, Michels 1921).

5. Twentieth-century Changes of Meanings

The intertwining of the sociological and the political or ethical viewpoints was perpetually renewed during the history of French intellectuals after the Dreyfus affair. A new sociologization of the word occurred with the attacks against previous *dreyfusards* in the pre-world-war period. The '*parti intellectuel*,' to use Péguy's phrase, was charged by their former allies (for example Georges Sorel) for having used the political struggle in order to conquer eminent positions of power. For their critics, this political party in fact

constituted a social cluster of arrivists, a new elite of mandarins backing upon leftist politicians (Prochasson 1992).

Between the two World Wars, there was a revival of the political and ideological emphasis on the meaning of the word during what has been called the 'Franco-french war' between the right-wing '*parti de l'intelligence*' and the left-wing '*intellectuels de gauche*.' This trend is enhanced because, at the same time, extreme left movements tried to restrict the notion to a sociological sense. The influence of Soviet Marxism may be found here in which *intellectuels* become a mere synonym for the Russian word *intelligentsia* which is nearer a sociological concept after the October revolution than the term *intellectuels* is in French. *Intellectuels* in the phraseology of the Communist Parties are assimilated to a social group in order to deny them any political autonomy and oblige them to define their political attitude within the limits of the Marxist-Leninist orthodoxy and to renounce to their own vision which is, in the French political tradition, far more linked with the French Revolution legacy and the defense of Human Rights (Benda 1927). This intertwining of two traditions after the 1930s explains why leftist intellectuals used social concepts to attack their rightist opponents, while in contrast, rightist intellectuals laid the stress on the idealist conception of the intellectuals' role; they opposed themselves to materialism (i.e., leftist intellectuals influenced by Marxism) instead of only taking the defense of tradition as before (Sapiro 1999).

The decline of Marxism in the *intelligentsia* and the excess of sociological terrorism in intellectual struggles during the 1950s and 1960s (Boschetti 1985) explain why French contemporary intellectual life is dominated by some sort of revival of the primitive meaning of the word *intellectuel*. This complete historical cycle is one of the origins of the renewed interest for the study of intellectuals as political and social actors by French and foreign scholars in the 1980s and 1990s (Bourdieu 1992, Jennings 1993, Judt 1992, Julliard and Winock 1996, Sirinelli 1990, Trebitsch and Granjon 1998).

Anti-intellectualism, which was so frequent in England and Germany before World War I seemed to somewhat decline after the mobilization of all types of intellectuals (scholars as well as writers or journalists) during the *Union sacrée* for propaganda or practical applications devoted to National Defense. It conferred on them a new importance in all political contexts. After World War I, the democratization of politics, in both the UK and Weimar Germany, now placed intellectuals in a political context quite similar to the French one. The growing influence of left parties, of Marxist ideas, of international questions (communism, fascism, pacifism, fear for a new world war) gave birth to international debates among European intellectuals (or even American, if the case of New York intellectuals may be included: Wald 1987). Nevertheless, this did not mean that the specific national traditions were forgotten. Even if notions like *Intellektuelle* or *intelligentsia* were present more than before in the German or English public debate, they never obtained the general influence or consensus which their equivalents enjoyed in France, Italy, or Spain (Bering 1978, Stark 1984). Even innovative sociologists such as Karl Mannheim continued in general to use the older lexis even to propose their new conceptions of a free-floating intelligentsia (*freischwebende Intelligenz*: Mannheim 1929). Debates about the social decline of 'intellectual workers' in Germany or Central Europe used too old phrases such as *geistige Arbeiter* (Jarausch 1990, Titze 1990). The anti-intellectualism of the Nazi movement and the huge migration (after 1933) of progressist or Jewish intellectuals out of Germany and Central Europe stopped, for almost two decades, this timid convergence with the Latin tradition of intellectuals.

The decline and fall of the different fascist or communist regimes in the second half of the twentieth century gave a new actuality to the oppositional figure of intellectuals, in particular in Central and Eastern Europe (the 'dissidents' who recalled the nineteenth century meaning of intelligentsia). In Southern Europe, in Africa, Asia, or Latin America, revolutionary intellectuals or militants for the Human Rights also rejuvenated the nineteenth century European tradition of intellectual youth mobilized for the nation or populist projects. All these convergences or political cycles explain why the historical notion of intellectuals continues to be at the center of many historical, sociological, or philosophical reflexions, trying until today to find general or transhistorical definitions of this term.

See also: Elites: Sociological Aspects; Higher Education; Intellectuals, Sociology of; Mannheim, Karl (1893–1947); Professionalization/Professions in History; Social Change: Types

Bibliography

Allen P 1986 The meanings of 'an intellectual': nineteenth- and twentieth-century English usage. *University of Toronto Quarterly* **55**: 342–58

Barbagli M 1974 *Disoccupazione intellettuale e sistemo scolastico in Italia*. Il Mulino, Bologna [1982 *Educating for Unemployment: Politics, Labor Markets, and the School System Italy, 1859–1973*. Columbia University Press, New York]

Bell D 1973 *The Coming of Post-industrial Society, a Venture in Social Forecasting*. Basic Books, New York

Benda J 1927 *La trahison des clercs*. Grasset, Paris, [1928 *The Great Betrayal*. Routledge, London]

Bender Th 1987 *New York Intellect: A History of Intellectual Life in New York City, from 1750 to the Beginnings of Our Own Time*. Knopf and Johns Hopkins University Press, New York

Bering D 1978 *Die Intellektuellen. Geschichte eines Schimpfwortes*. Klett, Stuttgart, Germany

Boschetti A 1985 *Sartre et les Temps modernes*. Minuit, Paris [1988 *The Intellectual Enterprise*: *Sartre and Les Temps Modernes*. Northwestern University Press, Evanston, IL]
Bourdieu P 1992 *Les règles de l'art. Genèse et structure du champ littéraire*. Éditions du Seuil Paris [1996 *The Rules of Art: Genesis and Structure of the Literary Field*. Polity Press, Cambridge, UK]
Charle C 1990 *Naissance des 'intellectuels' (1880–1900)*. Éditions de Minuit, Paris
Charle C 1996 *Les Intellectuels en Europe au XIXème siècle, essai d'histoire comparée*. Éditions du Seuil, Paris
Collini S 1991 *Public Moralists: Political Thought and Intellectual Life in Britain 1850–1930*. Clarendon Press, Oxford, UK
Collini S 1993 Intellectuals in Britain and France in the twentieth century: confusions, contrasts and convergence? In: Jennings J (ed.) *Intellectuals in Twentieth Century France, Mandarins and Samurais*. St. Martin's Press, London, New York, pp. 199–225
Drouin M 1994 *L'Affaire Dreyfus de A à Z*. Flammarion, Paris
Engelhardt U 1986 *'Bildungsbürgertum.' Begriffs- und Dogmengeschichte eines Etiketts*. Klett-Cotta, Stuttgart, Germany
Gilcher-Holthey I 1986 *Das Mandat des Intellektuellen: Karl Kautsky und die Sozialdemokratie*. Siedler, Berlin, Germany
Granjon M-C, Racine N, Trebitsch M (eds.) 1997 *Histoire comparée des intellectuels*. IHTP, CNRS, Paris
Heyck T W 1982 *The Transformation of Intellectual Life in Victorian England*. Croom Helm, London
Hübinger G, Mommsen W (eds.) 1993 *Intellektuelle im Deutschen Kaiserreich*. Fischer Taschenbuch Verlag, Frankfurt/M, Germany
Jarausch K H 1990 *The Unfree Professions, German Lawyers, Teachers, and Engineers, 1900–1950*. Oxford University Press, New York, Oxford
Jennings J (ed.) 1993 *Intellectuals in Twentieth Century France, Mandarins and Samurais*. St. Martin's Press, London
Judt T 1992 *Past Imperfect: French Intellectuals, 1944–1956*. University of California Press, Berkeley, CA
Julliard J, Winock M (eds.) 1996 *Dictionnaire des intellectuels français*. Éditions du Seuil, Paris
Mannheim K 1929 *Ideologie und Utopie*. F. Cohen, Bonn, Germany [1936 *Ideology and Utopia, an Introduction to the Sociology of Knowledge*, K. Paul, Trench, Trubner and Co., Harcourt, Brace and Co., London, New York]
Michels R 1921 *Le Prolétariat et la bourgeoisie dans le mouvement socialiste italien*. Giard, Paris
Michels R 1932 Intellectuals. In: *Encyclopaedia of the Social Sciences*. Vol. 8, Macmillan, New York, pp. 118–26
Müller O W 1971 *Intelligencija. Untersuchungen zur Geschichte eines politischen Schlagwortes*. Athenäum, Frankfurt/M, Germany
Ory P, Sirinelli J-F 1986 *Les intellectuels en France de l'affaire Dreyfus à nos jours*. Colin, Paris
Perkin H 1989 *The Rise of Professional Society, England since 1880*. Routledge, New York, London
Prochasson C 1992 *Les Intellectuels, le socialisme et la guerre (1900–1938)*. Éditions du Seuil, Paris
Ringer F K 1969 *The Decline of the German Mandarins*. Harvard University Press, Cambridge, MA
Ringer F K 1992 *Fields of Knowledge. French Academic Culture in Comparative Perspective 1890–1920*. Cambridge University Press, Cambridge, UK
Sapiro G 1999 *La guerre des écrivains 1940–1953*. Fayard, Paris
Serrano C, Salaün S (eds.) 1988 *1900 en Espagne, essai d'histoire culturelle*. Presses Universitaires de Bordeaux, Bordeaux, France
Shils E 1972 *The Intellectuals and the Powers, and Other Essays*. University of Chicago Press, Chicago
Sirinelli J-F 1990 *Intellectuels et passions français: Manifestes et pétitions au XXème siècle*. Fayard, Paris
Stark M (ed.) 1984 *Deutsche Intellektuelle 1910–1933, Aufrufe, Pamphlete, Betrachtungen*. Verlag Lambert Schneider, Heidelberg, Germany
Titze H 1990 *Der Akademikerzyklus*. Vandenhoeck and Ruprecht, Göttingen, Germany
Trebitsch M, Granjon M-C. (eds.) 1998 *Pour une histoire comparée des intellectuels*. Complexe, Brussels, Belgium
Wald A M 1987 *The New York Intellectuals. The Rise and Decline of the Antistalinist Left from the 1930s to the 1980s*. University of North Carolina Press, Chapel Hill, NC
Williams R 1976 *Keywords. A Vocabulary of Culture and Society*. Oxford University Press, Oxford, UK

C. Charle

Intellectuals, Sociology of

1. Origins

Intellectuals are people whose main activities involve producing, evaluating, and distributing culture. Their role often involves endorsing or criticizing the cultural objects of their attention.

The intellectual role has been performed in all but the most primitive human societies. However, the Renaissance humanists (circa 1300–1600 CE) were the first influential group resembling today's intellectuals. Renaissance scholars thought they lived in a golden age of virtue and learning. To emphasize the point, they invented the term 'Middle Ages' to refer to the dark gulf separating their era from the glories of the ancient world. Promoting the study of poetry, grammar, rhetoric, history, and moral philosophy, the humanists from Petrarch in the fourteenth century to More, Erasmus, and Rabelais in the sixteenth century influenced the Reformation, the scientific revolution, and the Age of Reason.

Only since about 1600, however, have intellectuals become a large, moderately well-defined, and occasionally self-conscious group. The chief factors that helped to distinguish them were the rise of commerce and industry since the late Middle Ages, the spread of literacy, the growth of markets, and the proliferation of social contexts for intellectual discussion and debate. Flourishing commerce and industry permitted the accumulation of economic surpluses. They were used to expand state and private institutions employing intellectuals. Public education and mass literacy increased demand for intellectual services. In turn, the evolution of markets for these services allowed intellectuals greater freedom to define their political allegiances. The interests of aristocratic and church patrons no longer constrained them.

Finally, a host of modern social settings—ranging from universities to political movements, professional societies to coffee houses, academic journals to mass-circulation newspapers—provided contexts within which intellectuals could formulate their social and political identities.

It was not until the second half of the nineteenth century that 'intellectual' and the kindred 'intelligentsia' entered common parlance. The term 'intellectual' was first employed on a wide scale in France. In 1898, Georges Clemenceau referred to the leaders of the anticlerical and antimilitary camp that opposed the conduct of the Dreyfus trial as *les intellectuels*. Soon the political right was deriding the self-proclaimed conscience of the French nation. Because of the resulting public debate, the term intellectual stuck. The term intelligentsia was popularized in Central and Eastern Europe a few decades earlier. It, too, denoted liberals, socialists, and other critics of authority. Only gradually since the end of the nineteenth century has the term intellectual gained widespread acceptance and succeeded in unwrapping itself from quotation marks.

2. *Theories*

Many normative or moralistic treatments of the problem of intellectuals may be found in the literature. They make a case for what intellectuals *should* do. In contrast, this article focuses on the analytical literature, which seeks to explain *why* intellectuals do what they do. Three main analytical traditions inform the sociological study of intellectuals. Sociologists have regarded intellectuals as (a) members of a class or as a class in their own right; (b) relatively classless; and (c) embedded in a shifting network of class and other group affiliations.

Class theory. Early Marxists asserted that intellectuals are bound to become members of the proletariat under capitalism. Later Marxists, disappointed with the failure of many intellectuals to become radicalized, claimed they are part of the *petite bourgeoisie*. Both Marxists and non-Marxists updated these assertions in the 1970s. The middle of the class structure had expanded enormously over the preceding century. People with university degrees filled many of the middle ranks. Reflecting on this change, some social thinkers argued that the intellectuals form a class. Alvin Gouldner held that the intellectuals form a new 'emancipatory class.' Barbara and John Ehrenreich maintained that intellectuals form part of a new 'professional-managerial class.' Irving Kristol called intellectuals the 'new class of knowledge workers.' These and other scholars disagreed about the political interests of the intellectuals. However, they shared the view that intellectuals form a large, rising, and ideologically quite homogeneous class.

Theories of classlessness. In contrast, Karl Mannheim held that modern intellectuals form neither a class nor part of a class. Instead, they are 'members of a relatively classless stratum which is not too firmly embedded in the social order.' Intellectuals are typically recruited from various classes. However, because they participate in a common educational milieu, their class differences, and the variations in outlook normally associated with them, tend to be suppressed. Others echoed Mannheim's conclusions. Talcott Parsons asserted that intellectuals put 'cultural considerations before social ones.' Everett C. Ladd and Seymour Martin Lipset wrote that the capacity of intellectuals for social criticism, creativity, innovation, and attention to facts enables them to overcome their class socialization—and, for that matter, the influence of many other groups and communities to which they belong.

Theories of shifting social networks. Both class theories and theories of classlessness ignore some crucial issues. For example, intellectuals may be found at all points on the political spectrum. Yet class theories briefly note ideological heterogeneity, dismiss it as an uninteresting side issue, and leave it unexplained. Meanwhile, theories of classlessness minimize the significance of social influences on the shaping of ideas. Thus, they deflect attention from the very issue that the sociology of intellectuals is supposed to analyze.

A third approach seeks to overcome both these problems by focusing on the intellectual's web of shifting group affiliations. According to Pierre Bourdieu, cultural fields tie individual and organizational actors together in various social institutions. These cultural fields are structured as social networks. Different positions in a social network are associated with different endowments of economic, social, and cultural capital. Network position is also associated with cognitive structure. Thus, intellectuals in similar network positions share tastes, ideas, and values. Randall Collins' analysis of philosophical schools takes this argument a step further by showing how intellectuals gain reputations by successfully competing for attention, fame, and influence in intellectual networks.

Bourdieu's and Collins' mesolevel theories are compatible with Robert Brym's macrolevel approach. Developing themes in Antonio Gramsci's essay on intellectuals, Brym argues that the ideologies and political allegiances of intellectuals depend on their social origins and the structure of opportunities for education, employment, and political involvement they face during their careers. These opportunity structures are, in turn, shaped by the relative power of major classes and other groups. Accordingly, to explain intellectuals' ideologies and political allegiances, one must trace their paths of social mobility as they are shaped by the capacity of classes and other groups to expand the institutional milieux through

which intellectuals pass. To the degree these milieux are imprinted with the interests of the classes and other groups that control them, they circumscribe the class and other group interests reflected in intellectuals' ideological and political allegiances. From this point of view, it is an oversimplification to say that intellectuals form a class, are members of a class, or are classless. They are embedded in social networks whose ties to various classes and other groups shift over time and help account for their ideologies and political allegiances. This theme has also been taken up by Jerome Karabel. It is illustrated in the remainder of this article.

3. *Social Origins and Institutional Opportunities*

3.1 *Social Origins*

In Weimar Germany, professors were largely anti-republican and Right-wing. Mostly their fathers were military officers, state bureaucrats, and academics: members or handmaidens of the aristocracy. In contrast, nonacademic intellectuals were inclined to the left. They tended to be children of successful participants in the Industrial Revolution: members of the upstart German bourgeoisie. Similarly, in the middle of the nineteenth century, Russia's conservative and politically moderate intellectuals were recruited overwhelmingly from the aristocracy. On the other hand, the first generation of truly radical Russian intellectuals dates from the 1860s, and it contained in its ranks a large mixture of commoners. Their fathers were merchants, peasants, and petty officials. These examples illustrate that support for the traditional social order is common among intellectuals who have been reared in upper-class families that are threatened by social change. Being born into a less advantaged class has usually been associated with greater potential for intellectual radicalism.

But not always. In North America, rates of upward social mobility are higher than in Europe, and historically class has been a less important social distinction. As a result, class origin has less bearing on the ideologies and political allegiances of North American than European intellectuals. Especially in the United States, ethnic and racial origin, educational background, field of expertise, professional and institutional status, and generational experience matter more. Accordingly, American surveys show that professors in small provincial universities are more likely to be politically conservative than professors in prestigious Ivy League universities. African-American and Jewish intellectuals are usually more liberal than non-Jewish white intellectuals. Young intellectuals who work in the social sciences and cultural fields are more likely to be Left-wing than older intellectuals in industry and business.

3.2 *Economic Opportunities*

As intellectuals reach adulthood, the structure of markets for intellectuals' skills can modify their early political socialization. An abundance of secure jobs that allow free expression can dampen intellectuals' early radical impulses. A lack of such jobs often exerts a radicalizing effect. Some intellectuals may continue to express radical political sentiments after they have taken jobs that suit them professionally. Still, the rate and intensity of radicalism are generally greater when the number of such jobs is fewer.

For example, the Quebec educational system began to produce substantial numbers of highly educated French-speaking graduates in the 1960s. The new graduates found that an English-speaking minority controlled the larger and more efficient businesses in the private sector of the economy. Many highly educated *Québécois* were shut out of good jobs. This, among other factors, encouraged some of them to develop the idea that to become 'masters in their own house' Quebec must become a sovereign state. Surveys show that intellectuals became separatists in disproportionately large numbers and were a driving force leading to the election of separatist governments in 1976 and 1994.

3.3 *Political Opportunities*

The patterns of political opportunities intellectuals face also profoundly affect their ideas and loyalties.

At the level of social organization, political opportunities are structured by the availability of historical agents: workers, peasants, ethnic collectivities, and other groups that intellectuals may demarcate as the chief instruments of social change. Whether intellectuals become, say, socialists, populists, or nationalists is determined partly by which historical agents are mobilized for political action and which are relatively politically dormant at a given time.

At the level of the political system, the capacity of party organizations to absorb new talent is also important in shaping intellectuals' allegiances. For example, before the 1920s, the German Social Democratic Party had attracted most of the country's radical intellectuals. In the 1920s, however, it became (to use the then-current catchwords) 'bossified, ossified, and bourgeoisified.' Old men remained incumbents in the German Social Democratic Party for many years, and young men and women had little hope of rising in its ranks. Lack of opportunity was one reason the young generation of radicals turned in large numbers to the Communist Party of Germany.

If historical agents are largely unmobilized and no party organization is available to sustain the intellectual's beliefs, a process of political disillusionment is likely to set in. The lack of a historical agent combined with the presence of a party organization, however, is likely to lead to intellectual elitism.

4. Intellectual Elitism

Ever since Plato envisaged an ideal society ruled by 'philosopher-kings,' some intellectuals have been politically motivated or subconsciously driven by the conviction that they are better suited than nonintellectuals to create and maintain a just society. For example, in 1816 Hegel referred to Prussia as 'the state of the intelligentsia.' Both he and Comte characterized historical epochs in terms of the accomplishments of the human mind. Moreover, they saw the human mind as the driving force of history and its personification as the intellectual.

Other intellectuals have spurned such elitist views. The degree of elitism expressed by intellectuals varies independently of their position on the left–right dimension of politics. One finds democrats and elitists on both the left and the right. Two sets of circumstances promote intellectual elitism: weak participatory demands by the masses and lack of competition from other elites.

The history of Latin America illustrates how weak mass demand for political participation encourages intellectual elitism. For example, in the 1950s, Cuban revolutionary intellectuals were faced with a small and reformist working class and a politically inert peasantry. Consequently, they took matters into their own hands, forming guerilla bands in the countryside to seize state power. The low level of popular political participation in, and mass control over, the Cuban revolutionary party diminished the likelihood of subsequent democratic development.

The history of socialism and communism in Russia reveals a similar pattern. However, the Russian case is also instructive because it illustrates the opposite tendency. Even apparently unyielding elitists can democratize if they are pressured from below. Thus, when ties between workers and Russian Marxist intellectuals were dense—during the strike wave of the 1890s, the 1905 revolution, and the period of labor militance in 1912–14—most intellectuals wanted their parties to operate according to democratic principles. In contrast, when working class political participation fell—due to labor quiescence caused by troughs in the business cycle, strong government and police reaction to radical activities, and enormous losses in World War I—intellectuals returned to elitist principles of party organization.

A second precondition of democratic practice among intellectuals is vigorous elite competition. Aristotle argued that justice can be maximized not if philosopher-kings rule, but if constitutions are divided against themselves—if, in other words, there exists what is now called a separation of powers between the various branches of government. Following Aristotle, scholars now commonly believe that tyranny and arbitrary rule by an elite can be prevented if the branches of government hold each other in check. One may invoke a broadly similar principle in analyzing intellectual elitism. Nonintellectual elites need to keep intellectuals in check if they are to prevent the political and social dangers that can result from intellectuals taking Plato to heart.

The most compelling case for the benefits of elite competition derives from the history of communism, the epitome of intellectual tyranny in the twentieth century. Yet the dangers of intellectual elitism have hardly subsided now that the communist era has virtually ended. In much of the world, intellectuals are making bolder claims than ever about their ability to forecast and plan social and scientific change. Governments are respecting those claims by seeking the advice of intellectuals, awarding them research contracts, and employing them by the legion. Emboldened by their growing numbers and prestige, some intellectuals, particularly in the United States and Eastern Europe, have proclaimed themselves the ruling class of the future, the real holders of power in an era when knowledge allegedly means more than capital in determining status.

One should not take intellectuals' claims about their accomplishments too seriously. Economists, the 'hardest' of the social scientists, have failed to forecast and regulate economic trends. Natural scientists have arguably helped to create nearly as many social problems as they have solved.

On the other hand, the dangers of too much intellectual influence on political life should be carefully heeded. Individual intellectuals have proved to be exemplary political leaders, but when intellectuals are in a position to impose their blueprints on society the result is often more harm than good. The 'shock therapy' advocated by the academics who formed the core of the Russian government late in 1991 caused tremendous social dislocation and political reaction that undermined reform in that country. This is only the latest example of what can happen when intellectuals rule. It seems reasonable to conclude that the practice of politics is much too serious a matter to be left only to the intellectuals.

Bibliography

Anheier H K, Gerhards J, Romo F P 1995 Forms of capital and social structure in cultural fields: examining Bourdieu's social topography. *American Journal of Sociology* **100**: 859–903

Bourdieu P 1984 [1979] *Distinction: A Social Critique of the Judgement of Taste* [Nice R trans.]. Harvard University Press, Cambridge, MA

Bourdieu P, Loic J D, Wacquant 1992 *An Invitation to Reflexive Sociology*. University of Chicago Press, Chicago

Brint S 1984 New class and cumulative trend explanations of the liberal political attitudes of professionals. *American Journal of Sociology* **90**: 30–71

Brym R J 1978 *The Jewish Intelligentsia and Russian Marxism: A Sociological Study of Intellectual Radicalism and Ideological Divergence*. Macmillan, London

Brym R J 1980 *Intellectuals and Politics*. George Allen and Unwin, London

Collins R 1998 *The Sociology of Philosophies: A Global Theory of Intellectual Change*. Harvard University Press, Cambridge, MA
Ehrenreich B, Ehrenreich J 1979 The professional-managerial class. In: Walker P (ed.) *Between Labor and Capital*. Black Rose Books, Montreal, PQ, pp. 5–45
Etzioni-Halevy E 1985 *The Knowledge Elite and the Failure of Prophecy*. George Allen and Unwin, London
Gagnon A G (ed.) 1987 *Intellectuals in Liberal Democracies: Political Influence and Social Involvement*. Praeger, New York
Gouldner A 1979 *The Future of Intellectuals and the Rise of the New Class: A Frame of Reference, Theses, Conjectures*. Seabury, New York
Gramsci A 1971 [1929–35] The intellectuals. In: Hoare Q, Smith G N (eds. and trans.) *Selections from the Prison Notebooks*. International Publishers, New York, pp. 5–23
Karabel J 1996 Towards a theory of intellectuals and politics. *Theory and Society* **25**: 205–33
Konrad G, Szelenyi I 1979 *The Intellectuals on the Road to Class Power: A Sociological Study of the Role of the Intelligentsia in Socialism*. Harcourt Brace Jovanovich, New York
Kristol I 1972 About equality. *Commentary* **54**: 41–7
Ladd E C, Seymour M L 1975 *The Divided Academy: Professors and Politics*. McGraw-Hill, New York
Lipset S M, Dobson R 1972 The intellectual as critic and rebel: With special reference to the United States and the Soviet Union. *Daedalus* **101**(3): 137–98
Mannheim K 1955 [1929] *Ideology and Utopia* [Wirth L, Shils E trans.]. Harvest, New York
Marx K, Engels F 1972 [1848] Manifesto of the Communist Party. In: Tucker R (ed.) *The Marx-Engels Reader*. Norton, New York, pp. 331–62
Nakhaie R M, Brym R J 1999 The political attitudes of Canadian professors. *Canadian Journal of Sociology* **24**: 329–53
Parsons T 1963 The intellectual: a social role category. In: Rieff P (ed.) *On Intellectuals*. Anchor, Garden City, NJ, pp. 3–24
Spender D 1982 *Women of Ideas and What Men Have Done to Them: From Aphra Behn to Adrienne Rich*. Routledge and Kegan Paul, London

R. J. Brym

Intelligence: Central Conceptions and Psychometric Models

1. Lasting Debates in Intelligence Research: The Elusive Concept of Intelligence

Given its practical importance in individual lives and societies, the topic of 'intelligence' attracts much attention and debates both among the researchers and the general public. The research of intelligence has been pursued by psychometricians, cognitive psychologists, and developmentalists (see Sternberg et al. 1994 for comprehensive reviews; see also *Intelligence: Historical and Conceptual Perspectives* and *Intelligence: History of the Concept*). However, not many other terms in psychology are so elusive as 'intelligence' escaping consensual definitions over more than 100 years of research. Tracking chronologically the reports of three representative forums, each participated by experts in the field, sheds some light on the changing terrain of intelligence research during the twentieth century.

In 1921, a classical symposium convened by the editors of the *Journal of Educational Psychology* was held to discuss three questions: (a) What is intelligence?; (b) How can it be best measured?; (c) What are the most crucial next steps in research? (Thorndike et al. 1921). Among the 17 leading researchers who participated, 14 different answers were given! About one fourth of the participants suggested elementary processes (i.e., perception, sensation, and attention) as primary attributes of intelligence. Another fourth of the experts thought that physiological mechanisms of the brain should be the determining factors, and there were still others who contented that overt adaptive behavioral responses are the key features of intelligence. Sixty-five years later, a succeeding effort with 27 experts was arranged to address the exact same questions (Sternberg and Detterman 1986); however, the degree of lack of consensus remains similar to that was in 1921. Experts of the 1986 forum offered two dozen definitions.

Aside from the diverse conceptions among researchers, two popular books also sparked heated debates in the general public. The views range from renouncing the idea of 'measuring' individuals along the dimension of psychometrically-defined intelligence (S. J. Gould, *The Mismeasure of Man*, 1981) to the belief in a general intelligence factor (R. J. Herrnstein and C. Murray, *The Bell Curve*, 1994). Reacting to these controversial debates, in 1995 the Board of Scientific Affairs of the American Psychological Association called for an another authoritative report from a dozen experts in the field to discuss the 'knowns and unknowns' of intelligence (Neisser et al. 1996). The diversities and controversies surrounding conceptions of intelligence led some researchers to contend that the multiple determinants (and their relations) contributing to individual differences in intelligence could only be understood within integrative frameworks.

2. Some Lasting Trends and an Attempt of Integration

In order to highlight a few continuing trends hidden behind the overwhelming diversities, organizers of the 1986 forum compared the main issues raised in their discussions with those that were discussed in 1921 and showed that there were some general agreements across the two efforts (Sternberg and Detterman 1986). Specifically, conceptions including attributes such as *adaptation to the environment*, *basic cognitive processes*, *and higher-order thinking* (i.e., reasoning, problem

solving, and decision-making) were prominent in both discussions. In addition, the recurring themes of central debates seemed to surround issues about (a) whether intelligence is unitary or manifold, (b) the relative contributions of nature and nurture to individual differences in intelligence, and (c) the relative emphases on the process, content, and product aspects of intelligence.

As an attempt to compile the two dozen different views, an integrative framework (Sternberg and Detterman 1986) with multiple levels was also proposed in the report of the 1986 forum. At the behavioral level, theoretical conceptions that originated within the psychometric tradition focus on measuring the 'products' of information processing and on identifying the structure(s) of intelligence. These theories could be compared mainly in terms of whether the structure of intelligence is, more or less, considered to be unitary or pluralistic. At the processing level, the theories attempt to integrate psychometrically-defined 'products of intelligence' with various concepts of information processing. In this regard, the various conceptions differ in terms of their differential emphases either on elementary cognitive processes or higher-order cognition involving metacognition, planning, and decision-making. At the contextual level, some theories of intelligence stress the dynamics of human-ecology interaction and the concepts of situated intelligence. Another version of the contextualist theories emphasizes the dynamic contexts of lifespan development from birth into old age (see *Lifespan Theories of Cognitive Development*). These contextualist theories evolved primarily in conjunction with a focus on the nature of intelligence as adaptive capacity.

Still there have been other theories focusing on genetic and neurobiological influences on intelligence, which are not within the purviews of this article (see *Intelligence, Genetics of: Cognitive Abilities*; *Intelligence, Genetics of: Heritability and Causation*). In the following sections, we review and compare main theoretical conceptions of intelligence with respect to their differences along the cross-level product-process, the contexts of human-ecology, and the lifespan developmental orientation; as well as their relative emphases on interactions across these dimensions.

3. *Measuring the Products of Intelligence: From Unitary to Pluralistic and Hierarchical Views*

Ever since Alfred Binet's (see *Binet, Alfred (1857–1911)*) success in devising tests to distinguish mentally retarded children from those with behavioral problems, the psychometric approach to quantify individual differences in cognitive competence using standardized intelligence scales (e.g., the Stanford–Binet test or the Wechsler test) played a dominant role in intelligence research. These scales tend to measure scholastic performance with tasks over broad areas (e.g., memory, analytical reasoning, numerical skills, verbal fluency, and practical knowledge). Thus, overall sum scales of such tests predict school performance well. Perhaps not surprisingly because of their emphasis on school-related performances, their predictive validity drops when predicting task performance outside of the classroom, such as everyday problemsolving, vocational competence (Dixon and Baltes 1986), or other more expertise-like outcomes of real-world intellectual functioning (Ericsson and Smith 1991).

With a more general aim to capture a broader spectrum of human cognitive abilities and their interrelations in addition to the applied orientation of predicting academic performance, a number of psychometric theories have been proposed over last century. Scholars such as Burt, Cattell (see *Cattell, Raymond Bernard (1905–98)*), Guilford, Spearman, and Thurstone were the main proponents. At the methodological level, these theories differ with respect to the type of factor analytic methods they adopt and whether these methods are applied on the empirical-exploratory or theoretical-confirmatory basis. At the substantive level, they differ in terms of whether they orient more towards the general factor or multidimensional views of intelligence.

3.1 *Two-factor Theory*

Although subtests of the intelligence scales are often designed to measure specific domains of cognitive performance (like numerical skill and verbal fluency), performances across different tests tend to be positively correlated with each other (i.e., the so-called *positive manifold*). Spearman (1904) first presented a theory of general intelligence which he later expended into what is now known as the two-factor theory. He developed the first statistical model for explaining the positive manifold among task performances. The underlying assumption of this model is that individual differences in a given test can be decomposed into two components. One variance component is common to all other cognitive tasks—a general factor of intelligence (commonly known as *g*), whereas the other component is specific to each task (Spearman 1927). With this method, Spearman showed with several data sets that *g* accounted for all positive correlations among the test performances (see Jensen 1998, for a modern review on *g*). Since then, hundreds of other studies have also found that *g* accounts for rather large portions of the common variance among a broad selection of cognitive tasks.

3.2 *The Primary Mental Ability Theory*

Since the discovery of *g*, a main issue in psychometric intelligence research has been whether the structure of

intelligence is unitary with one general common factor or pluralistic with multiple specific factors reflecting domain-specific abilities. Among the first to question the unitary view was Thurstone, who developed multiple factor analytic techniques and conducted studies that included a broad range of paper-and-pencil items found in most other intelligence scales. In these studies, Thurstone (1938), for instance, was able to identify seven distinct factors reflecting spatial, perceptual speed, numerical, verbal, fluency, memory, and reasoning abilities. However, his assumption of these primary mental abilities being completely orthogonal to each other was not empirically supported. Instead, multiple common factor models later proposed by others showed that the positive manifold emerges again at the level of higher-order factors.

3.3 The Gf-Gc Theory and Related Theories

Building on the multiple primary abilities identified by Thurstone, one influential conceptual model since then has been the *Gf-Gc* theory (Cattell 1971, Horn 1982). Based on results from higher-order factor analyses, this theory distinguishes general fluid (*Gf*) abilities reflecting the individual's basic problem solving, reasoning, learning, and selective attending capabilities from general crystallized (*Gc*) abilities reflecting the acquired knowledge and their utilization. Unlike Thurstone's theory, which assumes independent primary mental abilities, Cattell proposed the ontogenetic *investment theory* to specify a possible mechanism relating *Gf* to *Gc*. He postulated that the culture-, experience-, and knowledge-based crystallized abilities arise, in part, as a result of a process in which the more biology-, maturation- and process-based fluid abilities are been *invested* into suitable learning experiences and contexts as the individual develops. The *Gf* and *Gc* distinction is often compared with Hebb's (1949; see *Hebb, Donald Olding (1904–85)*) differentiation between innate potentials for information processing (Intelligence A) and the acquired level of performance and comprehension (Intelligence B), and Baltes' (1987) juxtaposition of the cognitive mechanics and the cognitive pragmatics (see more details in a later section). Besides these two central broad factors, Cattell and Horn were also able to identify several other factors reflecting visual processing ability (*Gv*), short-term memory (*Gsm*), quantitative abilities (*Gq*), associative storage-retrieval (*Glr*), novel reasoning (*Gr*), and processing speed (*Gs*). These factors have been further verified and captured by comprehensive psychometric test batteries developed by others, such as the Stanford–Binet, Wechsler, and Woodcock–Johnson batteries.

3.4 The Structure of Intellect Model

Moving even further away from the unitary pole of *g* and working towards better delineating the process and product aspects of intelligence, Guilford (1967) proposed the structure of intellect (SOI) model, which specifies the (a) operation, (b) content, and (c) product components of intellectual abilities. The component of mental operation contains five general intellectual processes: cognitive, memory, divergent production, convergent production, and evaluation. The content component encompasses four broad areas (i.e., figural, symbolic, semantic, and behavioral) of information to which operations are applied. The product component entails six possible results (i.e., units, classes, relations, systems transformations, and implications) of applying particular operations in particular contents. Applying factor analytic techniques in a confirmatory manner to test this model deductively, Guilford expected the structure of intellect to result in 120 single independent factors of different operation $\times$ content-$\times$ product ($5 \times 4 \times 6$) combinations. However, in later modifications he acknowledged that the first-order factors could be correlated, therefore giving rise to multiple higher-order factors, but he still denied the existence of a general factor (Guilford 1982). On the one hand, the strength of the SOI theory is its more specific and refined definition of ability structure across dimensions capturing three distinct aspects of information processing. On the other hand, the main critiques of the model targeted at the fact that this model was derived through the use of a certain factor rotation technique (i.e., the procrustean rotation) which later was criticized to be subjective to confirmatory bias and lacking empirical support for most of the factors specified.

Another theoretical model akin to Guilford's facetted conceptualization is the Berlin model of intelligence structure (BIS) by Jäger (1984). This model cross-classifies each cognitive test with respect to an (a) operation and a (b) content facet. Operation is further categorized into four ability groups (reasoning, memory, creativity, and speed), whereas content is divided into three groups (figural, numerical, and verbal). Conceptually, the 12 operation $\times$ content (4×3) combinations taken as a whole represents general intelligence. However, by factoring groups of variables that are homogeneous within one of the two facets but are heterogeneous in the other facet, one can either extract the four operational or the three content factors. Specified as such, methodologically this model implies that any given mental ability is a linear combination of the two factors, and conceptually it simultaneously specifies a mental ability with a process and a content aspect.

3.5 The Three-stratum Theory

Reanalyzing 477 available covariance matrices of cognitive ability tasks, Carroll (1993) integrated the findings from the different theoretical schools into a hierarchical structure with three stratums. At the

lowest stratum, more than 60 narrow first-order abilities could be found. Given these narrow abilities are positively correlated with each other to varying degrees, a second stratum of broad abilities, ich closely resemble the primary mental abilities identified by Thurstone and by Cattell and Horn, could also be extracted. At the highest stratum, a general factor of intelligence captures the positive manifold among the broad ability factors at the second level. The hierarchy with all three stratums then describes the overall structure of intelligence. In terms of empirical breadth, this model comprehensively integrates the empirical basis of human cognitive abilities observed at the behavioral level. In terms of theoretical integration, it shows that diverging psychometric views of intelligence can be understood in terms of their differential focus on one of the stratums within the overall structure, selections of indicator tasks defining the lowest level of the structure, and adoptions of factor-analytical techniques.

Regarding the behavioral psychometric approach to intelligence research as a whole, a key issue that remains is whether factors derived by factor analysis are entities with causal status or organizational concepts capturing a system of the products of intelligence. Defining the structure of intelligence with multiple levels, going from multiple low-level specific abilities to intermediate broad abilities, and, finally, to a general factor at the highest level, has advantages for relating to other theoretical conceptions. Various conceptual connections from psychometric measures of intelligence to other theoretical conceptions that either emphasize inward cross-level integration with information-processing mechanisms and their biological correlates, outward integration with real-world everyday knowledge and problem solving skills, or lifespan developmental integration can be more easily drawn.

4. *Integrating Psychometric Intelligence With Cognitive Processes*

Starting from late 1970s, there has been a development towards specifying the processing aspects of intelligence in hierarchical models within the psychometric tradition and the development of investigating information-processing mechanisms in cognitive science in general. With the parallel developments in both fields, researchers of intelligence started to integrate information-processing concepts into the theoretical conceptions of intelligence. Some theories focus on integrating the measurement and explanation of intelligence by looking inwardly and downwardly within the individual for elementary information-processing correlates (e.g., processing speed and working memory) of intellectual abilities (e.g., Hunt 1980, Resnick 1976) and their neurobiological correlates (Eysenck 1982, see Deary 2000 for recent overview). Other theories focus on relating intelligence with aspects of information processing that are more closely related to real-world competence, such as complex reasoning, metacognition, planning, knowledge, vocational skills, and expertise (e.g., Ackerman 1996, Sternberg 1985, 1999).

4.1 *Mental Speed Theory*

The notion that information-processing speed may underlie individual differences in intelligence goes back all the way to Galton (see *Galton, Sir Francis (1822–1911)*), who tried to measure intelligence by simple sensory and reaction time tasks. Researchers dedicated to the mental speed approach have taken up Galton's initial idea to investigate the relation between intelligence and reaction times measured in a broad range of simple elementary cognitive tasks (see Jensen 1998, Vernon 1987 for reviews). Performances in these tasks tend to be perceptual based and do not require any semantic or practical knowledge. Thus, it is perhaps no surprise that the correlations between the reaction times measured with such simple tasks and intelligence have proven to be only moderate (i.e., hovering around the so called 0.3 barrier, Hunt 1980). Proponents of the elementary information processing theory in general and the mental-speed theory in particular admit that simple reaction time measures are inferior to more complex composite measures derived from mental tests for the measurement and prediction of intelligence. However, they argued that the moderate but consistent correlations between measures derived from simple elementary information-processing tasks and performance on mental tests may be important for unraveling the information processing aspects of intelligence. Variations in the observed correlations across studies and their theoretical interpretations have been the issues of much discussion. The interpretations range from proposing nerve conduction velocity as the basic biological parameter for all cognitive functioning, to several alternative explanations postulating higher-level constructs like attention, motivation, personality factors, or speed-accuracy trade-offs, to be the causes for a spurious relationship between intelligence and mental speed measured with simple de-contextualized tasks.

4.2 *Working Memory Theory*

Another aspect of cognitive function that has been deemed as a likely candidate information-processing correlate of intelligence is working memory (Baddeley 1986). Working memory reflects people's ability to briefly maintain information, while simultaneously carrying out other mental operations on the same or other information. Empirical evidence suggests that tasks requiring greater working memory loads tend to

correlate more highly with intelligence. Furthermore, it has also been found that working memory capacity correlates highly with psychometric reasoning (Kyllonen and Christal 1990). Taken these findings together, proponents of this theory suggest that working memory might be a better candidate process for explaining fluid intelligence than mental speed (Engle et al. 1999, Süß et al. in press).

4.3 The PPIK Theory

Staying with the general sentiment of relating intelligence with information processing, but orienting more towards issues on knowledge representation in artificial intelligence systems rather than the experimental cognitive approach, Ackerman (1996) proposed an integrative framework characterizing adult intellectual development as complex interactions between cognitive processes, personality traits, broad interest and knowledge domains. This framework can be seen as a starting point for disentangling the complex relations between the different aspects of intellectual functioning and personality-motivational modulators (Ackerman and Heggestad 1997). For instance, there is initial evidence indicating that personality traits, such as degrees of openness to experience and typical intellectual engagement, are correlated with the crystallized knowledge, but not with the general fluid abilities. Such findings motivate further intelligence research to explore the roles of broad domains of occupational and avocational knowledge in adult intelligence, which has not been captured sufficiently by the traditional assessment of *Gc*.

4.4 The Triarchic Theory

Sternberg's triarchic theory of intelligence (1985) is another example of an integrative theory that seeks to combine traditional psychometric approach with information-processing perspectives at the level of higher-order cognition, rather than elementary cognitive processes. Three different kind of information-processing components are distinguished in the triarchic theory. *Metacognitive components* are higher-order control processes used for executive planning, monitoring, and evaluating task performance. *Performance components* are lower-order processes that deal with the execution of a task itself. For instance, in an inductive reasoning task, the performance components may involve encoding the test word pairs, comparing the word pairs, and retrieve from memory information about the test items. The *knowledge-acquisition components* are processes involved in learning, retaining and integrating new information with old experiences. More recently, Sternberg (1999) had extended his original theory to encompass a broader spectrum of issues regarding successful everyday real-world competence. For instance, in everyday context, successful intelligent individuals are likely to invoke metacognitive higher-order regulatory thought processes to discern their own strengths and weaknesses, and device strategies to maximize their strengths but compensate for their weaknesses. While this theory is more integrative and ecologically valid in nature, details of many components of the theory await further specification and empirical testing.

5. Bioecological and Lifespan Theories of Intellectual Development

Traditionally, developmentalists have also been among the researchers of intelligence. Piaget, the pioneer of modern developmental psychology (see *Piaget, Jean (1896–1980)*), studied the development of mental processes through which knowledge is constructed during ontogenesis (Chapman 1988). The Piagetian view of intelligence focuses more on the processes of intelligence, rather on the products of intelligence and the individual differences therein. Specifically, central to Piaget's theory of intelligence (1960) is the view that adaptive behavior and knowledge are constructed through an intimate interplay between the child and the environment. Intellectual development results from the child generalizing her current behavior and knowledge to new situations (assimilation) and also adaptively modifying her current behavior and knowledge through new experiences (accommodation). Influenced by Piaget's conceptions, two main variants of later theories of intellectual development share common endorsements of gene-environment (or biology-culture) co-constructivist and contextualist views.

5.1 The Bioecological Theory

With its main focus on understanding the complex interactions between biology and ecology in shaping child development, the bioecological theory (see *Human Development, Bioecological Theory of*; Bronfenbrenner and Ceci 1994) proposes that aside from genetic effects individual differences in intellectual functioning are the result of variations in the bioecological interactions a child experiences during development. The theory has three main tenets. First, there are multiple innate abilities, as opposed to a single ability. Second, innate abilities serve as a range of possibilities that may develop, or fail to develop depending on their interactions with mechanisms and processes in the proximal developmental contexts (e.g., such as parent–child interactions and schooling). Third, motivation plays important modulatory roles in determining how much the environmental context aids the actualization of the individual's innate potential for intellectual development. The notion that

interests and motivational factors play a role in intellectual development has also recently been considered in the PPIK theory (see above).

5.2 The Dual-process Theory of Lifespan Cognitive Development

While also being co-constructive and contextual in nature, the dual-process theory of lifespan cognitive development focuses on understanding the biology–cultural interplay in the gains and losses of intellectual development across the lifespan from early childhood into old age (e.g., Baltes et al. 1999; see also *Lifespan Theories of Cognitive Development*). The lifespan developmental theory has been developed by combining both the psychometric and lifespan developmental approaches.

Combining Hebb's Intelligence A and Intelligence B distinction (see above) and the Cattell–Horn *Gf-Gc* framework with perspectives from cognitive and evolutionary psychology, the dual-process theory of lifespan cognitive development highlights two distinct but interacting aspects of intellectual functioning: the biologically-driven cognitive mechanics of basic information processing, and the contextually-driven cognitive pragmatics of culture-based bodies of acquired factual, procedural, and profession-related skills and knowledge. Furthermore, with an emphasis on the dynamic of lifespan development, the dual-process theory also focuses on how these two aspects of intellectual functioning interact with each other, develop, maintain, and decline throughout the lifespan. In line with Cattell's investment theory, the dual-process theory proposes differential lifespan trajectories for the cognitive mechanics and pragmatics. The biologically-driven cognitive mechanics display an early growth pattern which is then invested into the acquisition of pragmatic skills and knowledge. While most abilities of the cognitive mechanics start to decline in midlife, the efficacy of culture-, experience-, and expertise-based pragmatics remain, well into old age; or, as suggested by Ackerman (1996), lifelong occupational and avocational knowledge could still be maintained or even increase up to late adulthood. However, in old age the role of biology-based cognitive mechanics in regulating the culture-based cognitive pragmatics increases.

6. Concluding Remarks

As reviewed above, there is a continuing effort in intelligence research at evolving its subject matter in terms of which mental representation and phenotypic behavioral expressions of intelligence are part of the theoretical conceptions. A few other themes, although deviating somewhat from the more traditional conventions, have been evolving in parallel with the theoretical conceptions reviewed thus far, and merit acknowledgements. For instance, there have been growing emphases on various forms of expertise, such as professional skills or cognitive skills associated, e.g., with chess, medicine, or computer programming, as part of the core of intelligence (see Ericsson and Lehmann 1996, Newell and Simon 1972 for reviews). Other examples are the various lines of research on extending the predominately cognition-oriented conceptions of intelligence to social intelligence, emotional intelligence (e.g., Davies et al. 1998, Ford 1992, Mayer and Geher 1996), and other presumably separable intelligences, such as musical, spatial, and interpersonal intelligence (i.e., the Multiple Intelligence Theory, Gardner 1993).

As it is evident from the debates and controversies since Galton and Binet, it is not fruitful for intelligence research to continue into the new millennium with questions such as: is intelligence unitary or pluralistic? Is the goal of intelligence research to measure the end product of intellectual functioning and to predict performance, or is it to explain the processes of intelligence? Which plays a greater role in the development of intelligence, nature or nurture? Looking at the terrain of current intelligence research and its related fields, it is probably more fruitful to ask questions that highlight the differential utility of the various conceptions of intelligence in furthering our understanding of the complex dynamic interplay between neurobiology and sociocultural contexts in shaping intellectual development. For instance, given existing data at the different levels, which facets of intelligence maybe more biologically-driven and which are more context- and culture-based? Which aspects of intelligence are primarily implemented through elementary information processing mechanisms and their genetic and neurobiological correlates, which other aspects are open to influences from metacognitive, personality and motivational factors? Which aspects of intelligence are of more direct relevance to real-world problem solving? What are some aspects of intelligence that might be culture-invariant, and which other aspects are culture-dependent?

Integrating most of the central conceptions of intelligence developed over the last century reveals that intelligence as a phenomenon has multiple facets and levels, and is dynamic both in terms of information processing and in terms of the varying contexts in which it is expressed during individual ontogeny. Couched within an integrated framework, intelligence may be viewed as biologically-implemented information-processing capabilities together with knowledge and expertise giving rise to adaptive behaviors in different domains of functioning that are co-constructed through the interplay between socio-cultural contextual influences and the genetic and neurobiological mechanisms across lifespan development. A

growing Zeitgeist is more and more in the orientation of biocultural co-constructivism.

See also: Intelligence, Evolution of; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Intelligence: Historical and Conceptual Perspectives; Problem Solving and Reasoning, Psychology of; Psychometrics; Strategic Intelligence

Bibliography

Ackerman P L 1996 A theory of adult intellectual development: Processes, personality, interests, and knowledge. *Intelligence* **22**: 227–57

Ackerman P L, Heggestad E D 1997 Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin* **121**: 219–45

Baddeley A 1992 Working memory. *Science* **255**: 556–9

Baltes P B 1987 Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* **23**: 611–26

Baltes P B, Staudinger U, Lindenberger U 1999 Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology* **50**: 471–507

Bronfenbrenner U, Ceci S J 1994 Nature–nurture reconceptualized in developmental perspective: A bioecological model. *Psychological Review* **101**: 568–86

Carroll J B 1993 *Human Cognitive Abilities*. Cambridge University Press, New York

Cattell R B 1971 *Abilities: Their Structure, Growth and Action*. Houghton Mifflin, Boston

Ceci S J 1996 *On Intelligence: A Bioecological Treatise on Intellectual Development*. Harvard University Press, Cambridge, MA

Chapman M 1988 *Constructive Evolution: Origins and Development of Piaget's Thought*. Cambridge University Press, New York

Davies M, Stankov L, Roberts R D 1998 Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology* **75**: 989–1015

Deary I J 2000 *Looking Down on Human Intelligence: From Psychometrics to the Brain*. Oxford University Press, New York

Dixon R A, Baltes P B 1986 Toward life-span research on the functions and pragmatics of intelligence. In: Sternberg R J, Wanger R W (eds.) *Practical Intelligence: Nature and Origins of Competence in the Everyday World*. Cambridge University Press, New York, pp. 203–34

Engle R W, Tuholski S W, Laughlin J E, Conway A R A 1999 Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology*: *General* **128**: 309–31

Ericsson K A, Lehmann A C 1996 Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology* **47**: 273–305

Ericsson K A, Smith J (eds.) 1991 *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press, New York

Eysenck H J 1982 *A Model for Intelligence*. Springer, Berlin

Ford M E 1982 Social cognition and social competence in adolescence. *Developmental Psychology* **18**: 323–40

Gardner H 1993 *Multiple Intelligences: The Theory in Practice*. Basic Books, New York

Gould S J 1981 *The Mismeasure of Man*. Norton, New York

Guilford J P 1967 *The Nature of Human Intelligence*. McGraw-Hill, London

Guilford J P 1982 Cognitive psychology's ambiguities: Some suggested remedies. *Psychological Review* **89**: 48–59

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Herrnstein R J, Murray C *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York

Horn J L 1982 The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. In: Craik F I M, Trehub S (eds.) *Aging and Cognitive Processes*. Plenum Press, New York, pp. 237–78

Hunt E 1980 Intelligence as an information processing concept. *British Journal of Psychology* **71**: 449–74

Jäger A O 1984 Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau* **35**: 21–35

Jensen A R 1998 *The G Factor: The Science of Mental Ability*. Praeger, Westport, CT

Kyllonen P C, Christal R E 1990 Reasoning ability is (little more than) working-memory capacity?! *Intelligence* **14**: 389–433

Mayer J D, Geher G 1996 Emotional intelligence and the identification of emotion. *Intelligence* **22**: 89–113

Neisser U, Boudoo G, Bouchard T J, Boykin A W, Brody N, Ceci S J, Halpern D F, Loehlin, J C, Perloff R, Sternberg R J, Urbina S 1996 Intelligence: Knowns and unknowns. *American Psychologist* **51**: 77–101

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Piaget J 1960 The Psychology of Intelligence. Littlefield, Adams, Patterson, NJ

Resnick L B 1976 *The Nature of Intelligence*. Lawrence Erlbaum, Hillsdale, NJ

Spearman C E 1904 General Intelligence, objectively determined and measured. *American Journal of Psychology* **15**: 201–93

Spearman C E 1927 *The Abilities of Man: Their Nature and Measurement*. Macmillan, New York

Sternberg R J 1985 *Beyond IQ: A Triarchic Theory of Human Intelligence*. Cambridge University Press, New York

Sternberg R J 1999 The theory of successful intelligence. *Review of General Psychology* **3**: 292–316

Sternberg R J, Detterman D K (eds.) 1986 *What Is Intelligence? Contemporary Viewpoints on its Nature and Definition*. Ablex, Norwood, NJ

Sternberg R J (ed.) 1994 *Encyclopedia of Human Intelligence*. Macmillan, New York

Süß H-M, Oberauer K, Wittman W W, Wilheim O, Schulze R In press Working memory capacity explains reasoning ability and a little bit more. *Intelligence*

Thorndike E L et al. 1921 Intelligence and its measurement: A symposium. *Journal of Educational Psychology* **12**: 123–47

Thurstone L L 1938 Primary mental abilities. *Psychometric Monographs* **1**

Vernon P A (ed.) 1987 *Speed of Information-Processing and Intelligence*. Ablex, Norwood, NJ

S-C. Li and F. Schmiedek

Intelligence, Evolution of

Intelligence is defined as a special form of biological adaptation (see adaptation and evolution). It is a trait that generates adaptive behavior by altering brain

states, the configuration of which corresponds to memory and is related to some specific feature(s) of the environment. The phylogenetic distribution of intelligence is uncertain, but is likely to have first evolved hundreds of millions of years ago. The intelligence of any species is characterized by the intelligence of the species from which it evolved, and possibly also by special features of intelligence specific to that species. This applies to *Homo sapiens* as to any other species. Intense interest has centered about the notion of intelligence from the time Darwin contrasted intelligence with instinct, thus setting the stage for the nature–nurture controversy, which has created a schism in the behavioral and social sciences since the 1870s. The relatively recent notion that intelligence itself is an adaptation, closely interwoven with instincts, has gone a long way to resolving this long-running dispute.

1. *Defining Intelligence*

Intelligence is notoriously difficult to define in any brief and simple way. One reason for this is that whilst intelligence gives rise to adaptive behaviors, adaptive behavior is a ubiquitous phenomenon, most of which is not a consequence of intelligence. The first requirement, then, is to distinguish between behaviors that are a consequence of the action of intelligence and those that are not. A second reason for the difficulty in defining intelligence is that it has generic meaning in all animals that have the trait of intelligence; but it also has specific meaning in that intelligence may, and often does, manifest itself differently in different species. The problem is made more complicated by the commonly held view on intelligence in our own species, which is partly a manifestation of the problem of extending the generic definition to its species-specific manifestation in *Homo sapiens*. Given the fundamental tenet of evolutionary theory (see *Evolution, History of*) that living forms have a common origin based upon descent with modification, at least some aspects of human intelligence must be related to intelligence in non-human species.

1.1 *A Generic Definition*

Unlike plants that can incorporate and utilize the energy of the sun to fuel their needs, animals must indirectly gain access to the sun's energy by consuming plants, or other living forms, including other animals that eat plants. In order to do this they must move about and act on the world, that is they must behave. Most behavior is the result of inherited patterns of central nervous system (CNS) functions that have evolved as efficient ways of exploiting such energy resources and avoiding damage. Such adaptive behaviors may be described as innate or instinctive in that they are solely caused by genes expressed by appropriate and necessary conditions of development to result in relatively invariant species-typical behavior patterns. Such behavior neither is generated by, nor results in, long-term changes in CNS states. However, in some animals, adaptive behavior is the result of enduring changes to the state of their CNS that arise from experience. This capacity to generate adaptive behavior resulting from within-lifetime experiences leading to changes in CNS states is the generic definition of intelligence. Thus defined, intelligence is identified with psychological mechanisms and processes such as learning, memory, and forms of internally manipulated CNS state changes referred to as thought and problem solving. Most intelligent animals are confined to relatively simple mechanisms of learning and memory.

1.2 *Specific Definitions*

Intelligence is often defined in terms of specific, often species-specific, manifestations. The reduction of response to irrelevant events (habituation) and associative learning (classical conditioning and instrumental learning) are widespread and common means of generating adaptive behavior in most intelligent species. There are also a host of more restricted forms of intelligence which include the ability of honey bees to learn about the position of food resources from other members of their hive, the ability of some bird species to acquire species-typical bird song, and tool use in chimpanzees. The human capacity for acquiring language, doing arithmetic, and forming abstract concepts are other examples. Very little is yet known about the psychological and neurological mechanisms that are the basis of any forms of intelligence. They may be similar if not identical in some cases, but are clearly different in others. For example, most psychologists agree that there has been a singular failure to identify language acquisition with associative learning. Nonetheless, all are forms of intelligence conforming to the generic definition given above.

2. *The Origins of Intelligence*

One of the few things that Darwin wrote about with uncharacteristic lack of clarity is the origin of intelligence and its relationship to other forms of behavioral adaptations. It is now believed that intelligence originates in the limitations of instincts in dealing with changes in the world that occur at certain rates. However, it must be recognized that the difficulties faced by Darwin and later writers is that evolution is an historical process. The traits that are seen in living forms now are the result of selection forces acting on variations in the past, often millions, or hundreds of

millions of years before the present. This applies to all adaptations including intelligence. Beyond the reach of empirical verification, one can never be certain about the reasons why adaptations originally evolved. The strength of any analysis rests upon the extent to which conjectured past selection forces are still present.

2.1 The Uncertain Futures Problem

Nothing in the world is constant. Change and flux are pervasive, giving rise to what C. H. Waddington called the uncertain futures problem. All problems of survival concern coping with change, and all extinctions are caused by the inability of species to cope with uncertain futures. One of Darwin's great contributions to modern thought was his recognition that explaining the dynamic, not the static, is the aim of science. Change, however, takes different forms and occurs at different rates. Intelligence is one of a family of processes that make up biological responses to dynamically changing circumstances (Plotkin 1994).

All adaptations have three characteristics. The first is they increase individual fitness (that is, the chances of survival and reproduction). This is why they are selected and become widespread in a population or species. The second is they are heritable (that is, they are part-caused by genes and hence inherited from parents by offspring). The third is they constitute a relationship between parts of organismic organization (for instance, the existence of an enzyme, or a limb of specific form) with a particular feature of the environment (for example, a nutrient, or the form of terrain). Behavioral adaptations possess all these characteristics. They increase fitness; they are genetically part-caused, finding expression in behavior through development in an appropriate environment; and they have matching relationships to specific features of the environment.

2.2 Adaptations to Different Rates of Change and the Evolution of Intelligence

Adaptations constitute matching relationships to rela tively enduring features of the world, that is, whose change is so slow that they appear to be enduring. This is why they are present in most members of a species and endure as species-specific traits over long periods. The information for the construction of the trait is present in the gene pool of that population or species. However, features of the world that change more rapidly than can be tracked by the main evolutionary program have also to be adapted to if they constitute significant features of the environment. Some behavioral adaptations show developmental flexibility (see *Genotype and Phenotype*) such that different ranges of environmental experience will result in different forms of behavior (for example, living as solitaries or swarming in locusts). Such developmental adjustments to adaptations, however, are bound to the temporal restrictions of the developmental process. Other fluctuations in the environment are so fleeting and pervasive, like non-seasonal temperature changes, that flexible development cannot provide appropriate adaptations. In such cases, short-term responses like shivering or sweating provide a means of tracking such changes.

An important and different class of change is those which occur rapidly and possibly repeatedly throughout a lifespan, but which may constitute features of the world that endure for significant periods. Neither developmental plasticity nor short-term physiological adjustments can furnish adequate adaptations to such forms of change. Examples are the position of a food resource or the identity of an ally or enemy in a social species. Such short-term stabilities concern the relationship between organisms and their environments that are central to the social and behavioral sciences. It is the need to adapt to such short-term stabilities that constitute the evolutionary origins of intelligence. Hence, the definition of intelligence as the capacity to alter CNS states because of experience now extended to the notion that such changes in short-term stabilities be tracked and matched by neural network states that generate adaptive behaviors to such short-term stabilities.

One important feature of the environment is cause-effect relationships. Some of these are so constant that adaptations in the form of innate or instinctive behaviors of approach or avoidance will evolve. Others are transient. For example, the characteristics that identify food that causes illness may and do change and such changes must be tracked. A sensitivity to such changing cause–effect relationships and the ability to conserve these in the form of changes in CNS states (that is memories) that direct appropriate behavior were likely to have been amongst the earliest forms of intelligence to have evolved.

3. The Phylogenetic Distribution of Intelligence

There are no reliable reports of intelligence in non-animal forms such as plants or fungi. Depending upon classificatory system, the animal kingdom is made up of about 25 major divisions (phyla), with reports of intelligence being present in species belonging to seven of these.

3.1 Intelligence in Vertebrates

Vertebrates comprise much the greater part of the phylum Chordata, the comparative evidence indicating that intelligence in the form of associative and other relatively simple forms of learning is widespread

throughout this subphylum, which ranges from cartilaginous and bony fish, through amphibia, reptiles, and birds, to mammals. The comparative study of intelligence, in any phylum, has never been undertaken in any systematic fashion. Most laboratory-based animal learning studies have used just two species of animal, the domesticated laboratory rat and the pigeon. The poor status of this work as comparative scholarship is attested to by few texts on animal learning carrying an index of species. That said, and noting that negative studies are seldom published, it is probably the case that positive results on simple learning have been reported in all studies undertaken in different vertebrate species.

In addition to associative learning capacity, intelligence in the form of the ability to acquire knowledge of complex spatial arrays, complicated sound patterns being incorporated into vocal signaling, problem solving, transitive reasoning, and tool use have been demonstrated in a variety of species of bird and non-human mammals. The existence of imitation learning and the ability of apes, especially chimpanzees, to attribute mental states to others and to be able to learn and use language are highly contentious issues.

3.2 Intelligence in other Phyla

The most complete review of invertebrate intelligence is to be found in three volumes edited by Corning et al. (1973a, 1973b, 1975). In general, the evidence is either weak or disputed. The unicellular Protista (for example, amoeba) may show habituation, but lacking a nervous system, the mechanisms underlying the ability to cease to respond to non-significant events will be entirely different from multicellular animals that have evolved nervous systems (see *Science and Industry*). Claims for the existence of habituation in Coelenterate species (e.g. hydra, medusa) are also disputed, though such animals do have nerve-net nervous systems. Platyhelminthes (flatworms) have received considerable study. These animals have a CNS, occupy a pivotal position in the evolution of multicellular animals, and may have been demonstrated to have simple associative learning. The evidence for associative learning in members of the phylum Annelida (earthworms, leeches) is stronger. The great majority of animals are members of the phylum Arthropoda, which includes crustaceans, spiders, and insects. The behavior of honeybees is the most extensively investigated of all invertebrates. There is no doubt that they have associative learning, which shows remarkable similarities to the learning of the laboratory rat. Gastropod molluscs (like snails) can habituate, and possibly condition. There are extensive reports of Cephalapod molluscs (octopods and squids) demonstrating associative, instrumental, and discriminative learning. Given that these are active predators with highly developed sensory and nervous systems, this might be expected. However, it should be noted that claims for learning in these species have been vigorously contested.

In general, the evidence for habituation in a number of non-chordate animals is strong. Honeybees apart, the status of work on other forms of learning is poor. There is much to be done on invertebrate intelligence in the future, an area of research that for some years now has been moribund. It should be noted that centralized nervous systems with neural networks evolved around 500 million years ago. The selective advantages of being able to detect conserved cause–effect relationships and other short-term stabilities must always have been very strong. It would be surprising if high quality research into non-chordate learning in the future did not prove positive.

3.3 Intelligence, Ecology and Life-history Strategies

The nature of the short-term stabilities that are significant for the survival and reproduction of animals will vary with the life-history strategies and ecologies of different species. This view has led to a conceptually powerful approach in recent years which seeks to place intelligence within the context of the natural history of any intelligent species, where intelligence is seen to be interwoven with instincts (Gould and Marler 1987). It is likely that much of the ambiguity of findings on animal intelligence, especially that of invertebrates, will be resolved by research carried out within this framework.

4. Human Intelligence

Intelligence has been a central area of human psychological research throughout the twentieth century. Much of it has revolved around the seeming dichotomy between general intelligence (*g*) and specific intelligences (*s*) such as verbal and numerical skills. In recent years, *s* has come to be identified with computationally specialized cognitive modules (Fodor 1983). There is, however, no reason to think of *g* and *s* as being exclusive properties of human intelligence (Plotkin 2002). It is entirely conceivable either that some property of the brain's information processing capacity is common to all computational modules, or that the outputs of individual modules are subject to further common processing which is reflected in *g*.

4.1 Current and Future Research

Since the early 1980s there has been a remarkable advance in understanding the emergence of intel-

ligence in the child. The findings are relatively uniform in supporting the existence of discrete intelligences (*s*) that appear to be largely innate cognitive systems (Hirschfeld and Gelman 1994). The notion of human intelligence as a *tabula rasa* has been shown to be wrong. Furthermore, there is a move towards relating such individual cognitive skills to more basic mechanisms, for example showing the connections between language, gesture, and manual skills (Wilson 1998). This conceptual marriage between viewing intelligence as a collective of specific, innate, cognitive devices, and yet relating it to more basic mechanisms like information processing and memory mechanisms (Deacon 1997), is likely to be the dominant form of research into human intelligence in the coming decades.

It should be noted that lineages maintain many of the features of ancestral species. For this reason, human intelligence certainly comprises evolutionarily older forms of intelligence, such as associative learning because of the continuing need to be sensitive to conserved cause–effect relations, as well as more newly evolved forms of intelligence such as numerical and linguistic competence.

4.2 Intelligence and Culture

Culture (see *Cultural Evolution: Theory and Models*) is at once an extraordinary manifestation of, and an extension of, human intelligence. The appearance of culture is one of the major evolutionary transitions in the history of life (Maynard Smith and Szathmary 1995). An important focus of research into intelligence in this century will concern the essential psychological mechanisms of culture, especially those features of intelligence that allow humans to enter into culture (Plotkin 1998), as well as the interrelationship between culture and human cognition (Donald 1991) and between culture and biological evolution (Laland et al, 1995) (see also *Genes and Culture, Coevolution of*). Intelligence will be seen as an evolved adaptation with specific manifestations in *Homo sapiens* that make it central to understanding human culture. Understanding the relationships between biological evolution, individual intelligence, and human culture delineates one of the most important developments in the social sciences in the decades to come.

See also: Aging Mind: Facets and Levels of Analysis; Brain, Evolution of; Cultural Evolution: Theory and Models; Emotions, Evolution of; Evolution, History of; Evolution of Cognition: An Adaptationist Perspective; Genes and Culture, Coevolution of; Genetic Factors in Cognition/Intelligence; Genotype and Phenotype; Human Cognition, Evolution of; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Lifespan Theories of Cognitive Development;

Bibliography

Corning W C, Dyal J A, Willows A O D (eds.) 1973a *Invertebrate Learning*, Vol. 1. Plenum, New York

Corning W C, Dyal J A, Willows A O D (eds.) 1973b *Invertebrate Learning*, Vol. 2. Plenum, New York

Corning W C, Dyal J A, Willows A O D (eds.) 1975 *Invertebrate Learning*, Vol. 3. Plenum, New York

Deacon T W 1997 *The Symbolic Species*. Allen Lane, London

Donald M 1991 *Origins of the Modern Mind*. Harvard University Press, Cambridge, MA

Fodor J A 1983 *The Modularity of Mind*. MIT Press, Cambridge, MA

Gould J L, Marler P 1987 Learning by instinct. *Scientific American* **256**: 74–85

Laland K N, Kumm J, Feldman M W 1995 Gene-culture coevolutionary theory: a test case. *Current Anthropology* **36**: 131–56

Hirschfeld L A, Gelman S A (eds.) 1994 *Mapping the Mind*. Cambridge University Press, Cambridge, UK

Maynard Smith J, Szathmary E 1995 *The Major Transitions in Evolution*. Freeman, Oxford UK

Pinker S 1994 *The Language Instinct*. HarperCollins, New York

Plotkin H 1994 *Darwin Machines and the Nature of Knowledge*. Harvard University Press, Cambridge, MA

Plotkin H 1998 *Evolution in Mind*. Harvard University Press, Cambridge, MA

Plotkin H 2002 *The Imagined World Made Real*. Allen Lane, London

Sternberg R J, Kaufman J C (eds.) 2000 *Evolution and Intelligence*. Erlbaum, Mahwah, NJ

Wilson F R 1998 *The Hand*. Pantheon Books, New York

H. Plotkin

Intelligence, Genetics of: Cognitive Abilities

Far more genetic research has been conducted on intelligence than on any other behavioral dimension or disorder. Most of this research consists of family, twin, and adoption studies (see *Behavioral Genetics: Psychological Perspectives*) that investigated the rudimentary questions of whether and to what extent genetic factors are associated with individual differences in intelligence. This research will be reviewed briefly followed by new directions for genetic research that go beyond these rudimentary questions, including molecular genetic research that aims to identify specific genes responsible for genetic influence on intelligence.

1. Genetic Perspectives on Intelligence

Nearly all genetic research on human intelligence focuses on individual differences within our species. In contrast, cognitive psychology generally considers species-typical cognitive functioning, asking questions

such as the involvement of working memory and a central executive in processing information. Imaging research asks what parts of the brain are engaged during particular tasks. Such research is more concerned about universal themes of human cognitive functioning rather than individual variations on these themes, which is the provenance of genetic research. These represent different perspectives and entail different levels of analyses. Perspectives are not right or wrong, merely more or less useful for asking particular questions. Genetic research cannot address the etiology of species-typical behavior but it is well suited to investigate the etiology of individual differences within a species.

Genetic research to date is also largely limited to a particular model of cognitive functioning called the psychometric model that considers cognitive abilities to be organized hierarchically (Carroll 1993) from specific tests to broad factors to general cognitive ability (often called *g*). There are hundreds of tests of cognitive abilities. These tests measure several broad factors (specific cognitive abilities) such as verbal ability, spatial ability, memory, and speed of processing. These broad factors are not independent—they intercorrelate modestly, about 0.40. That is, in general, people who do well on tests of verbal ability tend to do well on tests of spatial ability. General cognitive ability, that which is in common among these broad factors, was discovered by Charles Spearman nearly a century ago. It has been suggested that *g* is preferable to the word intelligence because the latter has so many different meanings in psychology and in the general language (Jensen 1998).

Most people are familiar with intelligence tests, often called IQ (intelligence quotient) tests. These tests typically assess several cognitive abilities and yield total scores that are reasonable indices of *g*. For example, the Wechsler tests of intelligence, widely used clinically, include ten subtests such as vocabulary, picture completion (indicating what is missing in a picture), analogies, and block design (using colored blocks to produce a design that matches a picture). In research contexts, *g* is usually derived by using a technique called factor analysis that weights tests differently according to how much they contribute to *g*. This weight can be thought of as the average of a test's correlations with every other test. This is not a statistical abstraction—one can simply look at a matrix of correlations among such measures and see that all the tests intercorrelate positively and that some measures (such as spatial and verbal ability) intercorrelate more highly on average than do other measures (such as nonverbal memory tests). A test's contribution to *g* is also related to the complexity of the cognitive operations it assesses. More complex cognitive processes such as abstract reasoning are better indices of *g* than less complex cognitive processes such as simple sensory discriminations (Jensen 1998).

Although *g* explains about 40 percent of the variance among such tests, most of the variance of specific tests is independent of *g*. Clearly there is more to cognition than *g*. Moreover, there may be more to *g* than can be seen from the types of psychometric tests used to assess it at the end of the twentieth century. There are other ways to study cognitive processes than psychometric tests such as information-processing approaches that typically rely on reaction time. Especially exciting are neuroscience measures that directly assess brain function such as evoked potentials, positron emission tomography (PET) scans, and functional magnetic resonance imaging (fMRI). As new tests of cognitive abilities are developed and their reliability and validity established, their relationship to *g* and their genetic and environmental origins will be investigated. However, few genetic studies have as yet used these other cognitive measures.

Just as there is more to cognition than *g*, there is clearly much more to achievement than cognition. Personality, motivation, and creativity all play a part in how well someone does in life. However, it makes little sense to stretch a word like intelligence to include all aspects of achievement such as emotional sensitivity and musical and dance ability that do not correlate with tests of cognitive ability.

Despite the massive data pointing to *g*, considerable controversy continues to surround *g* and IQ tests, especially in the media. There is a wide gap between what lay people (including scientists in other fields) believe and what experts believe. Most notably, lay people often hear in the popular press that the assessment of intelligence is circular—intelligence is what intelligence tests assess. To the contrary, *g* is one of the most reliable and valid measures in the behavioral domain, its long-term stability after childhood is greater than for any other behavioral trait, and it predicts important social outcomes such as educational and occupational levels better than any other trait (Gottfredson 1997). Although a few critics remain, *g* is widely accepted by experts. It is less clear what *g* is, whether *g* is due to a single general process such as executive function or speed of information processing, or whether it represents a concatenation of more specific cognitive processes (Jensen 1998).

2. *Historical Context*

The relative influence of nature and nurture on *g* has been studied since the beginning of psychology. Indeed in 1865, a year before the publication of Gregor Mendel's seminal paper on the laws of heredity, Francis Galton published a two-article series on high intelligence and other abilities, which he later expanded into the first book on heredity and cognitive ability, *Hereditary Genius: An Inquiry into its Laws and Consequences* (1992; originally published in 1869).

The first twin and adoption studies in the 1920s also focused on *g*. Highlights in the history of genetic research on *g* include Leahy's (1935) adoption study, in which she compared IQ resemblance for non-adoptive and adoptive families. This study confirmed an earlier adoption study that showed genetic influence, in that IQ correlations were greater in non-adoptive than in adoptive families. The first adoption study that included IQ data for biological parents of adopted-away offspring also showed significant parent–offspring correlation, suggesting genetic influence (Skodak and Skeels 1949). Begun in the early 1960s, the Louisville Twin Study was the first major longitudinal twin study of IQ that charted the developmental course of genetic and environmental influences (Wilson 1983).

In 1963, a review of genetic research on *g* was influential in showing the convergence of evidence pointing to genetic influence (Erlenmeyer-Kimling and Jarvik 1963). During the 1960s, environmentalism, which had been rampant until then in American psychology, was beginning to wane, and the stage was set for increased acceptance of genetic influence on *g*. Then, in 1969, a monograph on the genetics of intelligence almost brought the field to a halt, because the monograph suggested that ethnic differences in IQ might involve genetic differences (Jensen 1969). Twenty-five years later, this issue was resurrected in *The Bell Curve* (Herrnstein and Murray 1994) and caused a similar uproar. The causes of average differences between groups need not be related to the causes of individual differences within groups. The former question is much more difficult to investigate than the latter, which is the focus of the vast majority of genetic research on IQ. The question of the origins of ethnic differences in performance on IQ tests remains unresolved.

The storm raised by Jensen's monograph led to intense criticism of all behavioral genetic research, especially in the area of cognitive abilities. These criticisms of older studies had the positive effect of generating a dozen bigger and better behavioral genetic studies that used family, adoption, and twin designs. These new projects produced much more data on the genetics of *g* than had been obtained in the previous 50 years. The new data contributed in part to a dramatic shift that occurred in the 1980s in psychology toward acceptance of the conclusion that *g* is significantly associated with genetic differences between individuals (Neisser et al. 1996).

3. *Genetic Influence on* g

Dozens of studies including more than 8,000 parent-offspring pairs, 25,000 pairs of siblings, 10,000 twin pairs, and hundreds of adoptive families all converge on the conclusion that genetic factors contribute substantially to *g* (Bouchard and McGue 1981, Plomin et al. 2001). Correlations for first-degree relatives living together average 0.43 for more than 8,000 parent-offspring pairs and 0.47 for more than 25,000 pairs of siblings. However, *g* might run in families for reasons of nurture or of nature. In studies involving more than 10,000 pairs of twins, the average *g* correlations are .85 for identical twins and 0.60 for same-sex fraternal twins. These twin data suggest that genetic factors play an important role in the origins of familial resemblance in *g* scores. Adoption studies also indicate substantial heritability. For example, in two recent studies, identical twins reared apart are almost as similar for *g* as are identical twins reared together, with an average correlation of 0.78 for 93 such pairs (Bouchard et al. 1990). Adoption studies of other first-degree relatives also indicate substantial heritability, as illustrated by recent results from the longitudinal 25-year Colorado Adoption Project (Plomin et al. 1997).

Estimates of the effect size of genetic factors, called heritability, vary from 40 to 80 percent but estimates based on the entire body of data are about 50 percent, indicating that genes account for about half of the variance in *g*. Even a meta-analysis that attempted to ascribe as much variance as possible to prenatal effects estimated heritability as 48 percent (Devlin et al. 1997). It is not of much importance whether heritability is 30, 50, or 70 percent because few implications would follow from one or the other estimate being correct. The point is that genetic influence on *g* is not only statistically significant; it is also substantial. Although heritability could differ in different cultures, moderate heritability of *g* has been found, not only in twin studies in North American and western European countries, but also in Moscow, former East Germany, rural India, urban India, and Japan (Plomin et al. 2001).

Most of the genetic variance for *g* is additive; that is, genetic effects add up rather than interact across loci—offspring resemble their parents genetically for additive genetic effects, not for interactive effects among genes. The additivity of most genetic effects on *g* may be due to the fact that there is greater assortative mating (nonrandom mating) for *g* than for other behavioral traits. Spouses correlate about 0.10 for personality, about 0.20 for height and weight, but for *g* spouses correlate about 0.45. Because bright women are likely to mate with bright men, this double-barreled effect makes their offspring likely to be brighter on average than would be expected if mating were at random. In this way, assortative mating spreads out the distribution of additive genetic effects on *g* in the population (Plomin et al. 2001).

These same genetic data provide the best available evidence for the importance of environmental factors independent of genetics. If genetic factors account for about half of the variance then non-genetic factors account for the rest of the variance. Environment clearly is important. This is suggested as well by the

steady rise in IQ scores during the past several generations, which would seem too short a time to allow genetic explanations (Flynn 1999), and by studies in which children from abusive families show gains in IQ when adopted (Duyme et al. 1999). However, little is known about specific environmental differences as they relate to *g* independent of genetic factors.

4. *Beyond Heritability*

The answers to the rudimentary questions of whether and how much genetic factors contribute to *g* now seem sufficiently clear—the answers are, respectively, 'yes' and 'a lot.' Genetic research on *g* has moved beyond heritability, for example, to investigate developmental changes, multivariate relations among cognitive abilities, and specific genes responsible for the heritability of *g*.

4.1 *Genetic Influence on* g *Increases During Development*

When Francis Galton first studied twins in 1876, he investigated the extent to which the similarity of twins changes during development. Other early twin studies of *g* were also developmental, but this developmental perspective faded from genetic research until the late twentieth century. One of the most interesting findings about *g* is that heritability increases steadily from infancy (20 percent) to childhood (40 percent) to adulthood (60 percent). For example, a recent study of twins aged 80 years and older reported a heritability of about 60 percent (McClearn et al. 1997).

The 20-year longitudinal Colorado Adoption Project (CAP) confirms this finding using the adoption design (Plomin et al. 1997b). CAP is a 25-year study of 245 children adopted away from their biological parents at birth and adopted into adoptive homes in the first month of life. CAP includes 'genetic' (biological) parents and their adopted away children, 'environmental' (adoptive) parents and their adopted children, and 'genetic-plus-environmental' (non-adoptive or control) parents and their children who have been matched to the adoptive families. Correlations between non-adoptive parents and children increase from less than 0.20 in infancy to about 0.20 in middle childhood and to about 0.30 in adolescence. The correlations between biological mothers and their adopted-away children follow a similar pattern, indicating that parent-offspring resemblance for *g* is due to genetic factors. In contrast, parent–offspring correlations for adoptive parents and their adopted children hover around zero, which suggests that family environment shared by parents and offspring does not contribute importantly to parent–offspring resemblance for *g*. Because CAP is a parent–offspring design with adult parents and young offspring, these results can be interpreted more precisely as showing that genetic effects on adult *g* do not come on line just in adulthood, but also to a considerable extent in adolescence, to a lesser extent in childhood, and even to some slight extent in infancy.

Why does heritability of *g* increase during the life span? Perhaps completely new genes come to affect *g* as more sophisticated cognitive processes come on line. A more likely possibility is that relatively small genetic effects early in life snowball during development, creating larger and larger phenotypic effects, perhaps as individuals select or create environments that foster their genetic propensities.

4.2 *Genetic Effects are Broad (Molar) Rather than Modular*

In the widely accepted hierarchical model of cognitive abilities mentioned earlier, *g* consists of what is in common among specific cognitive abilities such as spatial, verbal, speed of processing, and memory abilities. Less is known about the genetic and environmental origins of individual differences in specific cognitive abilities, but they also appear to show substantial genetic influence, although less than *g* (Plomin and DeFries 1998).

To what extent do different sets of genes affect these phenotypically different abilities? A technique called multivariate genetic analysis examines covariance among specific cognitive abilities and yields a statistic called the genetic correlation, which is the extent to which genetic effects on one trait correlate with genetic effects on another trait independent of the heritability of the two traits. That is, although cognitive abilities are moderately heritable, the genetic correlations between them could be anywhere from 0.0, indicating complete independence, to 1.0, indicating that the same genes influence a variety of cognitive abilities. Multivariate genetic analyses have consistently found that genetic correlations among specific cognitive abilities are very high, close to 1.0 (Petrill 1997). In other words, the same genetic factors largely influence different abilities. What this finding means concretely is that if a specific gene were found that is associated with verbal ability, the gene would also be expected to be associated with spatial ability and other specific cognitive abilities.

These genetic results have major implications for current theories of cognitive neuroscience. According to one theory, the brain works in a modular fashion, that is, cognitive processes are specific and independent. Implicit in this perspective is a bottom-up reductionistic view of genetics in which modules are the targets of gene action. In contrast, the findings from multivariate genetic analyses suggest a top-down view in which genetic effects operate primarily on *g*,

rather than a bottom-up view in which genetic effects are specific to modules. Given that the brain has evolved to learn from a variety of experiences and to solve a variety of problems, perhaps it makes sense that the brain functions holistically. However, finding genetic correlations near 1.0 does not prove that genetic effects are limited to a single general cognitive process that works in a top-down way. Another alternative is that specific cognitive abilities as they are currently assessed might involve many of the same modular processes that are each affected by different sets of genes. This alternative hypothesis could be tested by multivariate genetic research on measures of more modular processes such as neuroimaging measures of brain function.

4.3 Specific Genes are Beginning to be Identified

Heritability of complex dimensions such as *g* seems likely to be due to multiple genes of varying but small effect size rather than a single gene that has a major effect. Genes in such multiple-gene systems are called quantitative trait loci (QTLs) (Plomin et al. 1994). Unlike single-gene effects like PKU that are necessary and sufficient for the development of a disorder, QTLs contribute interchangeably and additively like probabilistic risk factors. Traditional methods for identifying single-gene effects are unlikely to succeed in identifying QTLs.

It is interesting that the two best-replicated QTLs are both in the cognitive domain. Apolipoprotein-E was first reported in 1993 to be related to late-onset dementia using allelic association (Corder et al. 1993) and has since been replicated in scores of studies (Rubinsztein 1995). QTL sib-pair linkage designs were used to identify a linkage between chromosome 6p21 and reading disability (Cardon et al. 1994), a linkage that has been replicated in several subsequent studies (e.g., S. E. Fisher et al. 1999). The QTL perspective suggests that both dementia and reading disability are likely to be the quantitative extremes of continuous distributions.

If additive genetic factors are important for complex traits, then it should be possible to identify the genes responsible if we have sufficient power to detect QTLs of small effect size. The critical question is the distribution of effects sizes of these QTLs: What is the average effect size of QTLs and how are QTL effect sizes distributed? If the average effect size is 1 percent we will eventually detect many QTLs given a heritability of at least 50 percent. However, if the average QTL effect size is 0.1 percent, we will detect very few QTLs and they will be difficult to replicate.

A QTL study applying new genetic approaches to *g* yielded a replicated association in a study comparing groups of children of high *g* and children of average *g* (Chorney et al. 1998). The gene is insulin-like growth factor-2 receptor (IGF2R) on chromosome 6, which has been shown to be especially active in brain regions most involved in learning and memory. The frequency of one of the alleles was twice as high in two groups of children with high *g* as compared to two groups of children with average *g* (about 30 percent vs. 15 percent). Combining these results yielded a highly significant result ($\chi^2 = 12.41, p < 0.0004$). An ongoing systematic scan of the genome using these techniques has begun to identify several other QTLs associated with *g* (P. J. Fisher et al. 1999).

5. Implications of Identifying Genes Associated with g

Identifying replicable QTLs associated with *g* will make it possible to use measured genotypes rather than indirect inferences about heritable influence based on familial resemblance in order to address issues such as developmental change and continuity and the multivariate relationships among cognitive abilities. Replicated QTL associations with *g* will also provide discrete windows through which to view pathways between genes and *g*. Functional genomics, understanding how genes affect traits, is generally viewed in terms of bottom-up molecular biological analyses of cellular function in which the gene product is identified and its effects studied at a molecular and cellular level. However, other levels of analysis are also likely to be useful in understanding how genes affect *g* such as anatomical neuroimaging, functional neuroimaging, electrophysiology, psychophysiology, cognitive processing, and psychometrics. As an antidote to the tendency to define functional genomics at the cellular level of analysis, the phrase 'behavioral genomics' has been proposed (Plomin and Crabbe 2000). The grandest implication for science is that the functional (behavioral) genomics of *g* will serve as an integrating force across diverse disciplines with DNA as the common denominator, opening up new scientific horizons for understanding learning and memory.

As is the case with most important advances, identifying genes for cognitive abilities and disabilities will also raise new ethical issues. Finding QTLs for *g* will have important implications for society as well as science (Plomin 1999). In terms of implications for society, it should be emphasized that no policies necessarily follow from finding genes associated with *g* because policy involves values. For example, finding genes for *g* does not mean that we ought to put all of our resources into educating the brightest children. Depending on our values, we might worry more about children falling off the low end of the bell curve in an increasingly technological society and decide to devote more public resources to those who are in danger of being left behind.

Many ethical issues related to DNA are being broached at the level of single-gene disorders that are

hard-wired in the sense that a single gene is necessary and sufficient for the development of the disorder. This will benefit ethical deliberations about the genetics of *g*, which seems less pressing because genetic effects on *g* are probabilistic rather than deterministic for two reasons. First, heritability is closer to 50 percent than to 100 percent, which means that non-genetic factors make a major contribution. Second, because many genes contribute to the heritability of *g*, the system is inherently probabilistic. Potential problems related to finding genes associated with *g* have been discussed such as prenatal and postnatal screening, discrimination in education and employment, and group differences (Newson and Williamson 1999).

The fear lurks in the shadows of such discussions that finding genes for *g* will limit our freedom and our free will. In large part such fears involve misunderstandings about how genes affect complex traits like *g*. Finding genes for *g* will not automatically open a door to a genetic version of Huxley's brave new world where babies are sorted out at birth (or before birth) into alphas, betas, and gammas. Although the balance of risks and benefits to society of finding genes for *g* is not clear, basic science has much to gain from functional genomic studies of brain functions related to learning and memory. We need to be cautious and to consider carefully societal implications and ethical issues. There is also much to celebrate here in terms of the increased potential for understanding our species' nonpareil ability to think and learn.

See also: Behavioral Genetics: Psychological Perspectives; Developmental Behavioral Genetics and Education; Evolution of Cognition: An Adaptationist Perspective; Genetic Factors in Cognition/Intelligence; Human Cognition, Evolution of; Intelligence: Central Conceptions and Psychometric Models; Intelligence, Evolution of; Intelligence, Genetics of: Heritability and Causation; Intelligence: Historical and Conceptual Perspectives

Bibliography

Bouchard T J Jr, Lykken D T, McGue M, Segal N L, Tellegen A 1990 Sources of human psychological differences: The Minnesota study of twins reared apart. *Science* **250**: 223–8

Bouchard T J Jr, McGue M 1981 Familial studies of intelligence: A review. *Science* **212**: 1055–9

Cardon L R, Smith S D, Fulker D W, Kimberling W J, Pennington B F, DeFries J C 1994 Quantitative trait locus for reading disability on chromosome 6. *Science* **266**: 276–9

Carroll J B 1993 *Human Cognitive Abilities*. Cambridge University Press, New York

Chorney M J, Chorney K, Seese N, Owen M J, Daniels J, McGuffin P, Thompson L A, Detterman D K, Benbow C P, Lubinski D, Eley T C, Plomin R 1998 A quantitative trait locus (QTL) associated with cognitive ability in children. *Psychological Science* **9**: 1–8

Corder E H, Saunders A M, Strittmatter W J, Schmechel D E, Gaskell P C, Small G W, Roses A D, Haines J L, Pericak Vance M A 1993 Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**: 921–3

Devlin B, Daniels M, Roeder K 1997 The heritability of IQ. *Nature* **388**: 468–71

Duyme M, Dumaret A-C, Tomkiewicz S 1999 How can we boost IQs of 'dull children'?: A late adoption study. *Proceedings of the National Academy of Sciences USA* **96**: 8790–4

Erlenmeyer-Kimling L, Jarvik L F 1963 Genetics and intelligence: A review. *Science* **142**: 1477–9

Fisher P J, Turic D, McGuffin P, Asherson P, Ball D M, Craig I, Eley T C, Hill L, Chorney K, Chorney M J, Benbow C P, Lubinski D, Plomin R, Owen M J 1999 DNA pooling identifies QTLs for general cognitive ability in children on chromosome 4. *Human Molecular Genetics* **8**: 915–22

Fisher S E, Marlow A J, Lamb J, Maestrini E, Williams D F, Richardson A J, Weeks D E, Stein J F, Monaco A P 1999 A quantitative-trait locus on chromosome 6p influences different aspects of developmental dyslexia. *American Journal of Human Genetics* **64**: 146–56

Flynn J 1999 Searching for Justice: The discovery of IQ gains over time. *American Psychologist* **54**: 5–20

Galton F 1992 *Hereditary Genius: An Enquiry into its Laws and Consequences*. World, Cleveland, OH

Gottfredson L S 1997 Why *g* matters: The complexity of everyday life. *Intelligence* **24**: 79–132

Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York

Jensen A R 1969 How much can we boost IQ and scholastic achievement? *Harvard Educational Review* **39**: 1–123

Jensen A R 1998 *The g Factor: The Science of Mental Ability*. Praeger, Westport, CT

Leahy A M 1935 Nature–nurture and intelligence. *Genetic Psychology Monographs* **17**: 236–308

McClearn G E, Johansson B, Berg S, Pedersen N L, Ahern F, Petrill S A, Plomin R 1997 Substantial genetic influence on cognitive abilities in twins 80+ years old. *Science* **276**: 1560–3

Neisser U, Boodoo G, Bouchard T J, Boykin A W, Brody N, Ceci S J, Halpern D F, Loehlin J C, Perloff R, Sternberg R J, Urbina S 1996 Intelligence: Knowns and unknowns. *American Psychologist* **51**: 77–101

Newson A, Williamson R 1999 Should we undertake genetic research on intelligence? *Bioethics* **13**: 327–42

Petrill S A 1997 Molarity versus modularity of cognitive functioning? A behavioral genetic perspective. *Current Directions in Psychological Science* **6**: 96–9

Plomin R 1999 Genetics and general cognitive ability. *Nature* **402**: C25–9

Plomin R, Crabbe J C 2000 DNA. *Psychological Bulletin* **126**(6): 806–28

Plomin R, DeFries J C 1998 Genetics of cognitive abilities and disabilities. *Scientific American* **278**(5): 62–9

Plomin R, DeFries J C, McClearn G E, McGuffin P 2001 *Behavioral Genetics*. Worth, New York

Plomin R, Fulker D W, Corley R, DeFries J C 1997 Nature, nurture and cognitive development from 1 to 16 years: A parent-offspring adoption study. *Psychological Science* **8**: 442–7

Plomin R, Owen M J, McGuffin P 1994 The genetic basis of complex human behaviors. *Science* **264**: 1733–9

Rubinsztein D C 1995 Apolipoprotein-E—A review of its roles in lipoprotein metabolism, neuronal growth and repair and as

a risk factor for Alzheimers-disease. *Psychological Medicine* **25**: 223–9
Skodak M, Skeels H M 1949 A final follow-up on one hundred adopted children. *Journal of Genetic Psychology* **75**: 84–125
Wilson R S 1983 The Louisville Twin Study: Developmental synchronies in behavior. *Child Development* **54**: 298–316

R. Plomin

Intelligence, Genetics of: Heritability and Causation

In theory, the 'genetics of intelligence' describes the extent to which, and manner by which, an individual's mental abilities are influenced by his or her genes. In practice, there is a lot of debate and no universally accepted description of the genetics of intelligence. Intelligence is regarded as a complex trait, meaning that many genetic and environmental factors are thought to influence it. Nevertheless, we know little about the number or nature of genes that influence cognitive abilities. We know even less about how the expression of these genes depends on other genes or on environmental factors. To complicate matters even further, there are numerous facets of intelligence, each of which may be influenced in different ways by genes and the environment. This article focuses on the genetics of human intelligence and on the problems of assessing the genetic and environmental contributions to intelligence. To begin, definitions and measures of intelligence will be described. The effects of genes known to influence intelligence will then be mentioned. Finally, the degree to which variation in intelligence among humans may be genetically based will be discussed.

1. *Intelligence is a Multifaceted Trait*

Before one can study the genetics of a trait, one has to define precisely what the trait is and how it can be measured. Once a measure has been defined for a trait, then one can begin to tackle the question of how an individual's score may be influenced by genetic and environmental factors. Specifying an appropriate scale for intelligence is complicated by the fact that intelligence is a multifaceted trait. That is, there are several aspects of intelligence, involving reasoning, abstraction, problem solving, memory, linguistic aptitude, artistic ability, etc. Abilities in these different areas tend to be correlated to some extent within a population, which has led some authors to posit the existence of an underlying general intelligence factor, *g* (Spearman 1904). It is argued that the *g* factor directly influences many different aspects of intelligence, causing them to be correlated. Another possibility is that

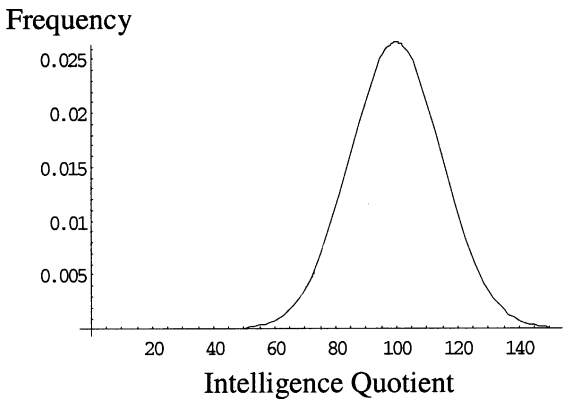

Figure 1
The frequency distribution for an idealized IQ test. Raw scores from a test are rescaled to fit the shape of a normal curve with mean at 100 and standard deviation of 15. Consequently, when the test is devised, 95 percent of the population will have IQ scores that fall between 70 and 130 by construction

different aspects of intelligence interact with one another and that these interactions generate the observed correlations (Hunt 1997). For example, strength in one area of intelligence (e.g., memory) might compensate for weakness in another area (e.g., linguistic ability), causing a person's abilities in these areas to be more similar than would be expected for uncorrelated traits. A compelling reason to believe that the various facets of intelligence are not just alternate manifestations of some underlying *g* factor is that they respond in different ways to aging and education. For example, Horn (1985) noted that measures of 'fluid intelligence' (the ability to solve new problems) generally declines with age among adults whereas 'crystallized intelligence' (the ability to apply known problem-solving methods to a problem at hand; see Cattell 1971) tends to increase with age and experience. A final possibility is that there are both general and specific cognitive factors that interact in a myriad of ways for different aspects of intelligences (e.g., Spearman 1904, Vernon 1979, Carroll 1993).

In most genetic studies, some variant of the Intelligence Quotient (IQ) is used as a single measure of intelligence, ignoring the multifaceted nature of intelligence (for exceptions see Cardon et al. 1992, Cardon and Fulker 1994, Pederson et al. 1994). The IQ measure grew out of studies by Alfred Binet on the ages at which children display skills of varying difficulty. Binet (1905) developed a series of age-scaled tests to assess the mental age of children. Wilhelm Stern (1916) used these tests to devise the IQ, defined as the mental age of a child divided by his or her chronological age multiplied by 100. Similar scales have been devised for adults, such as the Wechsler Adult Intelligence Scale and the Stanford-Binet Inte-

lligence Scale. These scales have been plagued with controversy over the degree to which the measures are biased in favor of certain classes, cultures, races, or sex (Gould 1996, Schiff and Lewontin 1986, Wahlsten 1997). It is argued that there can be no completely objective measure of intelligence, since what it means to be intelligent and who is perceived to be intelligent are social constructs subject to change.

Scores on IQ tests are adjusted so that they follow a normal distribution, or a 'bell curve,' with a mean of 100 (Fig. 1). An indication of the importance of environmental factors on IQ is that these scales must constantly be readjusted over time, a phenomenon known as the 'cohort effect.' In many countries, the average IQ score has risen over time (Flynn 1987, Wahlsten 1997). For example, the average IQ score based on Ravens Progressive Matrices of 18-year old males from The Netherlands has increased from 100 in 1952 (when the test was developed and standardized) to 106 in 1962, 112.4 in 1972, and 121.1 in 1981/2. This time-span is far too short for genetic changes in the composition of the population to account for the increases. Rather, these data are evidence that changing educational and social environments can have a dramatic effect on whatever aspects of intelligence were measured by this IQ test.

2. *Genes Known to Affect Intelligence*

Intelligence, like any human attribute, is partly a product of our genes, simply because we would not be living human beings without these genes. However, human intelligence is partly a product of our environment, because mental impairment can be caused, for example, by exposure to toxins such as lead during development. Between these two trivial statements lies an entire spectrum of possibilities, and a heated 'nature versus nurture' debate has grown over the extent to which individual differences in intelligence are genetically based (Devlin et al. 1997b, Fraser 1995, Herrnstein and Murray 1994, Jensen 1969, Lewontin et al. 1984, McInerney 1996).

The human genome contains between 30,000 and 100,000 genes. Although the number of genes that affect intelligence is unknown, there is reason to believe that the number is large. Measures of intelligence, such as IQ, are most likely quantitative genetic traits influenced by hundreds or thousands of genes. Most of these genes will not be 'intelligence' genes, *per se*, but rather genes that are critical to normal development. Severe mutations in such genes would affect several traits, not just mental ability, causing a disease syndrome with multiple symptoms. These severe mutations are the most studied, in large part because they are the easiest to detect.

Geneticists have compiled over 10,000 genes and genetic disorders affecting humans into the database, *Online Mendelian Inheritance in Man* (http://www3.ncbi.nlm.nih.gov/Omim/). A search for entries describing intellectual impairment, mental retardation, or dementia turned up almost 1,000 syndromes, although many of these might not be genetically independent. Familiar examples include Down's syndrome, Parkinson's disease, Alzheimer's disease, fragile X syndrome, Prader–Willi syndrome, Tay-Sachs disease, Wolf–Hirschhorn syndrome, and Rubinstein syndrome. Through the human genome project, more genetic details will be uncovered for these and other human traits and diseases.

Even in those cases where a mutation has been identified that 'causes' a disorder affecting intelligence, the mental development of patients carrying the mutation can be greatly improved by environmental conditions. A classic example of the interplay between genes and the environment in mental development is phenylketonuria (PKU). PKU is a recessive disease caused by a mutation at the phenylalanine hydroxylase gene (PAH) on chromosome 12. People with PKU do not produce enough of the enzyme, PAH, which oxidizes phenylalanine and converts it into tyrosine. Without this enzyme, phenylalanine levels increase to poisonous levels in the body, giving rise to a number of symptoms including mental retardation, a 'mousy' odor, light pigmentation, unusual posture, eczema, and epilepsy (Paine 1957). All of these manifestations of the disease may be prevented, however, by a straightforward change in the environment: a diet low in phenylalanine.

It is crucial to remember that understanding the genetics of severe mental disorders may reveal little about how intelligence develops normally and how this development depends on genetics. By analogy, understanding the environmental risk factors associated with brain damage from severe meningitis (inflammation of the tissues surrounding the brain) reveals little about environmental influences on normal brain development.

3. *Population-based Estimates of the Genetics of Intelligence*

Although severe mutations at many genes may affect intelligence, it does not follow that the majority of differences observed among individuals within a population are due to genetic differences. In fact, genes that, when disrupted, have large effects on mental abilities may not be responsible for any of the more minor differences seen among individuals. An entirely different approach must be taken to assess the extent to which variability in a trait is based on genetic differences within a population. In animals and plants, one would raise individuals in controlled environmental conditions and measure trait differences among these individuals. Assuming that the environmental variance has been minimized in the experiment, the remaining variability is attributed to genetic dif-

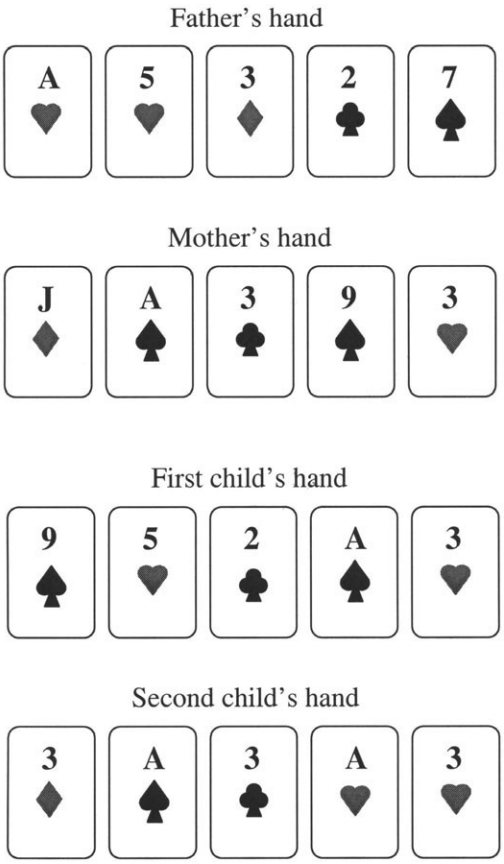

Figure 2
The attributes of an offspring may not be predictable on the basis of parental attributes with epistasis. As an illustration, consider a poker game where two parent hands are used to provide the cards for two offspring hands. In this example, the trait is 100 percent genetically determined, because all cards in the offspring hands come from either the father's hand (2/5 cards) or the mother's hand (3/5 cards, with the fifth card representing maternally inherited mitochondria). Nevertheless, an offspring's hand may be much worse or much better as a poker hand than expected based on the parents' hands. In the example shown here, the father's hand (ace high) and the mother's hand (pair of threes) are fairly poor, as is the first child's hand (ace high), but the second child has a fantastic hand (full house). Poker is a game with a large amount of epistatic interactions among cards in a hand. That is, a winning hand depends less on any particular card than on a good combination of cards. Similarly, even if a complex trait such as intelligence depends on the genes that an individual carries, it may do so in such an idiosyncratic way that only individuals with the same combination of genes are similar.

ferences among individuals. This controlled approach cannot be used for humans, however, and instead correlations based on family measures are used to infer the influence of genetic variation on the observed variability in a trait such as IQ.

3.1 Heritability

Heritability is a quantity used to measure the extent to which individual differences in a trait can be attributed to genetic differences. We must first define a few terms. The phenotype (P) of an individual constitutes the visible and/or measurable properties of that individual. In the following, for concreteness, we will take the phenotype to be an individual's score on an IQ. Therefore, our discussion will focus narrowly on the heritability of IQ rather than the heritability of intelligence. Different measures of intelligence are influenced in different ways by genes and the environment and consequently have different heritabilities (Cardon et al. 1992). Nevertheless, the difficulties described below in estimating heritability apply to any measure (or measures) of intelligence in humans.

In the simplest genetic models, phenotype is imagined to be a sum of genetic and environmental effects. Specifically, the genotype (G) of an individual is the total effect of his or her genes, and the environment (E) of that individual is the total effect of all non-genetic influences (e.g., education, diet, stress) on his or her phenotype. These effects will vary among individuals, and the differences are measured by the phenotypic (V_P), genotypic (V_G), and environmental (V_E) variance in a population for a given trait. Broad-sense heritability (H^2) measures the proportion of variance in phenotype that is attributed to genotypic variance: $H^2 = V_G/V_P$. This formulation assumes that the phenotype can be neatly partitioned into two parts and ignores genetic-by-environment interactions (GxE). All treatable genetic disorders, such as PKU, exhibit strong GxE, since symptoms depend on both predisposing mutations and insufficient treatment. Interactions between genes and environment may take on many forms and may well be important in the development of complex traits such as intelligence.

Broad-sense heritability estimates the degree to which differences among individuals are genetically based, but not all genetically based differences can be passed from parents to offspring. For example, whether a parent is homozygous (e.g., $++$) or heterozygous (e.g., $+-$) at a gene is entirely a genetic attribute (broad-sense heritability is 100 percent), but this state cannot be passed on faithfully from parents to offspring. Even if both parents are heterozygous at a gene, their offspring will only be heterozygous half of the time. In other words, there are genetic attributes of an individual that depend on the particular ways in which genes combine and interact and that cannot be inferred from the additive effect (A) of each gene

Table 1
Estimated correlations between the phenotypes of relatives[a]

Relationship	Estimated correlation in phenotype
Monozygotic twins	$V_G/V_P = H^2$
Siblings	$1/2\ V_A/V_P + 1/4\ V_D/V_P$
Half-siblings	$1/4\ V_A/V_P = 1/4\ h^2$
Parent–offspring	$1/2\ V_A/V_P = 1/2\ h^2$
Aunt/uncle	$1/4\ V_A/V_P = 1/4\ h^2$
Grandparent–offspring	$1/4\ V_A/V_P = 1/4\ h^2$

a This simple genetic model ignores epistasis, correlations among parents (assortative mating), cultural inheritance, maternal effects, and common environmental effects (due, for example, to attending the same schools). V_D equals the variance in dominance effects. See text for remaining definitions.

considered over all possible genetic backgrounds. There are two main types of genetic interactions. Dominance (D) describes interactions between the two copies of a gene (one inherited from the mother and one from the father), and epistasis describes interactions among different genes. With dominance and epistasis, knowing the phenotype of parents can sometimes provide little information about the phenotype of offspring. This point is illustrated in Fig. 2.

With genetic interactions, broad-sense heritability does not predict the extent to which offspring resemble their parents. Instead, narrow-sense heritability is used to measure how variation among individuals is influenced by genetic differences that are, on average, passed from parents to offspring. Narrow-sense heritability is defined as the fraction of phenotypic variance that can be attributed to variation in the additive effects of genes (V_A): $h^2 = V_A/V_P$. Narrow-sense heritability is always less than or equal to broad-sense heritability.

Neither heritability measure can be calculated directly since the genotypic (G), additive genetic (A), and environmental (E) scores for human individuals are unknowable. Instead heritability is inferred from formulae relating it to correlations among relatives. This creates an immediate problem, because the expected formulae depend strongly on the model one chooses. If one assumes that there is no epistasis, no social or cultural inheritance, no assortative mating (i.e., no correlation between parents in a trait), no sex differences, and no GxE interactions, then correlations among relatives can be derived in a straightforward manner (Table 1; see e.g., Lynch and Walsh 1998). With such a simplified model, heritability can be estimated from familial correlations. For complex traits such as IQ, however, we know that many of these assumptions are violated. Estimates in the literature for broad- and narrow sense heritability differ widely, depending in large part on what set of assumptions are made.

3.2 Heritability Estimates

Bouchard and McGue (1981) summarized a large number of studies measuring the correlations between relatives in IQ scores (Table 2). Heritabilities have been estimated from these data using different models and different subsets of the data (Table 3). The heritability estimates described in Table 3 have an extremely broad range, from 0.26 to 0.72 (out of a maximum range of 0 to 1). Although a large number of models appear to be included in Table 3, most of these models share several assumptions. Consequently, the range of heritability estimates is even smaller than

Table 2
Familial correlations in IQ from Bouchard and McGue (1981)

Relationship	Symbol	Number of pairs	Correlation
Spouses	M	3817	0.33
MZ twins (together)	r_{MZT}	4672	0.86
MZ twins (raised apart)	r_{MZA}	65	0.72
DZ twins (together)	r_{DZT}	5546	0.60
Sibs (together)	r_{SST}	26473	0.47
Sibs (raised apart)	r_{SSA}	203	0.24
Offspring–parent	r_{OPT}	8433	0.42
Offspring–midparent	r_{OMT}	992	0.50
Offspring–parent apart	r_{OPA}	814	0.22
Foster offspring–parent	r_{FOP}	1397	0.19
Foster offspring–midparent	r_{FOM}	758	0.24
Midoffspring–midparent	r_{MMT}	410	0.72
Adopted/biological sibs	r_{SFB}	345	0.29
Adopted/adopted sibs	r_{SFF}	369	0.34
Half-sibs	r_{HSS}	200	0.31
Cousins	r_{CZ1}	1176	0.15

Table 3
Heritability estimates based on correlations in IQ between relatives[a] (see Table 2)

Estimator	Estimate	Simple expectation	Assumptions
Estimates of broad-sense heritability			
r_{MZA}	0.72	$(V_A+V_D)/V_P$	(D, E, F[b], G; requires A, B to estimate h^2)
Estimates mixing broad-sense and narrow-sense heritability			
2 (r_{MZT}–r_{DZT})	0.52	$(V_A+3/2\ V_D)/V_P$	(B, C, F; requires A to estimate h^2)
2 r_{SSA}	0.48	$(V_A+1/2\ V_D)/V_P$	(B, C, D, E, F[b],G; requires A to estimate h^2)
2 (r_{SST}–r_{SFF})	0.26	$(V_A+1/2\ V_D)/V_P$	(B, C, D, F; requires A to estimate h^2)
Estimates of narrow-sense heritability			
2 r_{OPA}	0.44	V_A/V_P	(B, C, D, E, F[b])
2 (r_{OPT}–r_{FOP})	0.46	V_A/V_P	(B, C, D, E, F[b])
DP model	0.29	V_A/V_P	(A, B, D)
IP model	0.33	V_A/V_P	(A, B, D)
IS model	0.42	V_A/V_P	(A, B, D)
Devlin model[c]	0.34	V_A/V_P	(B, F)

a The first six estimates are described in Plomin and Loehlin (1989). The next three estimates are based on different published models of cultural and genetic inheritance with assortative mating from the analysis of Otto et al. (1994). The DP model assumes direct cultural transmission of IQ with assortative mating based on IQ (Cavalli-Sforza and Feldman 1978, Feldman and Cavalli-Sforza 1979, Martin et al. 1986). The IP model assumes indirect cultural transmission of IQ through social factors (e.g., socio-economic status) with assortative mating based on IQ (Cloninger et al. 1979a, 1979b, Feldman and Cavalli-Sforza 1979, Rice et al. 1978). The IS model assumes indirect cultural transmission of IQ through social factors and assortative mating based on social grouping (Morton and Rao 1978, Rao et al. 1976, 1979). The last estimate is from Devlin et al. (1997a), which uses a model that includes maternal effects and other pre-separation influences on siblings and twins. 'Simple expectation' is the value for the estimators under the restrictive genetic model described in Table 1. Assumptions made by the estimators include: (A) no dominance, (B) no epistasis, (C) no assortative mating, (D) no pre-separation effects on adopted offspring, (E) no correlation between biological and adoptive environments, (F) no cultural inheritance, (G) no allowance for siblings experiencing common environments (e.g., same schools).
b Cultural inheritance would affect the phenotypic variance in a way that is not accounted for in these estimates based on offspring raised apart (Otto et al. 1994).
c Devlin et al. (1997b) examined a slightly different data set, which excludes the last five rows in Table 2.

would be seen if models based on completely different sets of assumptions were included. For example, all estimates except the first one based on identical twins raised apart (r_{MZA}) ignore epistasis. Genetic interactions are difficult to include in models because the forms of these interactions are generally not known. Without this knowledge, one cannot accurately predict correlations among relatives, just as one cannot accurately predict the winnings of the poker hands in Fig. 2 without knowing the rules of poker. Furthermore, all models ignore GxE interactions. Finally, even those models that include cultural inheritance cannot capture the intricate and diverse ways in which social structures influence IQ and the many ways in which family members can influence one another's social context.

There is no consensus about the most accurate value for heritability. There are essentially two camps: those who argue that heritability lies between 0.5 and 0.8; and those who argue that heritability lies between 0.2 and 0.5. The first camp tends to focus on estimates based on twins, especially identical twins raised apart. Such estimates include dominance and epistatic interactions to some extent and therefore tend to measure broad-sense heritability. The correlation in IQ between identical twins raised apart is indeed high (0.72), but there are several nongenetic reasons to expect that identical twins raised apart may be similar. Twins raised apart tend to be placed in similar homes (Bouchard et al. 1990). They also have a certain amount of contact before and after separation (Bouchard et al. 1990). In particular, twins share the same womb and environmental influences during gestation, including any maternal effects (Devlin et al. 1997b). Furthermore, strikingly similar twins may be more easily recognized as identical twins raised apart and may be more likely to enter twin studies. The second camp, which argues for a low heritability of IQ, tends to use a wider variety of familial correlations and more explicit models. Such models have incorporated inheritance of social and cultural influences, similarities among parents (assortative mating), selective placement of adopted children, and/or maternal effects. Although narrow-sense heritability is most often reported in these expanded studies, the difference between broad-sense and narrow-sense heritability is often small when it is measured. For example, Devlin et al. (1997a) estimate a narrow-sense heritability for IQ of 34 percent and a broad-sense heritability of 48 percent. Heritability estimates are sensitive enough to the assumptions made to warrant skepticism towards any particular value.

3.3 The Limits of Heritability

Despite the fact that a lot of effort and scores of publications have been devoted to estimating the

heritability of IQ, heritability provides very limited information of relevance to human society. Heritability is a descriptor of the variability currently within a population and how this variability is partitioned into genetic and nongenetic influences. Even with an estimate of heritability, we know nothing about the reasons for the IQ of any given individual. Furthermore, heritability depends entirely on the set of environments in which it is measured. If one changes the environment, heritability may change in unpredictable ways, both because the phenotypic variance depends on the variance in environmental effects and because genotypic effects may depend on the environment (GxE). One cannot say on the basis of heritability whether or not a trait will change in response to an environmental manipulation (e.g., improved parental support and training, dietary supplements to pregnant women, different methods of schooling).

If heritability for IQ were 100 percent, one might be tempted to conclude that environmental factors are irrelevant and that there is little point in embarking upon social programs to improve the environment. This conclusion would, however, be entirely wrong. Consider the following fictitious example: imagine that there are only two types of people in equal frequencies, those with PKU and those without it. Imagine that no treatment is known, and that there are no environmental/cultural/social/maternal effects on IQ. Let the IQ of those without PKU equal 100, and the IQ of those with PKU equal some number less than 100. Under this scenario, 100 percent of the variability in IQ within the population is genetic ($H^2 = 1$). If the link between dietary phenylalanine and mental deficits in people with PKU is then discovered, and if people with PKU start controlling their diets, the average IQ would rise towards 100, despite the fact that, originally, there was only genetic and no environmental variation within the population. In short, heritability says nothing about how IQ might change in a new environmental context.

Estimating heritability will not help us know how to improve a trait, be it IQ or health or happiness. Detailed studies about how changes in social, educational, medical, and economic opportunities affect intelligence are required to determine what programs would be effective in permitting individuals to attain their highest potential.

4. *Race, Gender, and IQ*

Nowhere has the debate over the genetics of intelligence been fiercer than in the literature discussing differences in IQ between genders or among races (see, for example, the debate over *The Bell Curve* (by Allen et al. 1996, Devlin et al. 1997b, Fraser 1995, Herrnstein and Murray 1994, McInerney 1996). In those cases where a mean difference in IQ is observed between two groups, it is very difficult to determine the root causes of the IQ difference. One problem is the inherent bias in IQ tests favoring those individuals that come from the same cultural milieu and social class as the scientists that developed the tests (for examples, see Schiff and Lewontin 1986, pp. 21–3). A further problem is that the basis of differences among groups need bear no relationship to the basis of differences within groups. Even if IQ were entirely genetically heritable within groups, IQ differences among them may be entirely cultural or environmental. For example, the difference in average IQ score between a group of children exposed *in utero* to high levels of alcohol and a control group would be primarily environmental in origin (due to alcohol exposure), regardless of the heritability of IQ in either group.

Perhaps the greatest hurdle to understanding the root causes of IQ differences, however, is that it is difficult to eradicate the many socioeconomic differences faced by individuals of different ethniciticies, cultures, or genders. Nevertheless, transracial adoption studies have shown that differences in IQ among racial groups largely disappear when children are adopted into caucasian families, presumably because the children are then reared in 'the culture of the tests' (Weinberg et al. 1992). Furthermore, those programs that have enriched the educational environment of disadvantaged children have had great success in boosting their mean IQ score. For example, Wahlsten (1997) reviews two studies of children from disadvantaged homes with few financial and educational resources; a majority of these children were African American. In the studies, children were randomly assigned to a group receiving no additional educational resources and a second group receiving educational day care prior to public schooling. The average IQ score at four years old was 13.7 and 12.1 points higher in the two studies for the children receiving additional educational resources. Considering the average IQ gap between African Americans and Caucasians in the USA is about 15 points (Devlin et al. 1997b), these studies suggest that most if not all of the differences could be eradicated by providing quality educational opportunities for all children.

5. *Concluding Remarks*

Nearly the entire sequence of the human genome is now available. This genetic database will make it easier to find genes associated with human traits, including aspects of intelligence. It will be essential to remember, however, that genes do not cause intelligence. The developmental process that led to who we are today is a complex one involving the expression and repression of tens of thousands of genes in a manner that both depends on environmental conditions and alters these environmental conditions. The specific form of a gene that may enhance IQ in one family in one set of environments may be irrelevant in other families and in other environments. If future

genetic studies of intelligence are able to identify genes that enhance (or reduce) IQ under certain circumstances, then we may be able to mimic (or counter) the effects of these genes. Attempts to improve the educational and nutritional environments of children are, however, much more promising and immediate. Yet we should remember that linear, one-dimensional scales such as IQ mislead us into thinking that there are better and worse intellects. In reality, human thought is much more complex and multi-dimensional, which we must keep in mind if we are to preserve and encourage the rich intellectual diversity that characterizes human society.

See also: Genetic Factors in Cognition/Intelligence; Genetics and Development; Genotype and Phenotype; Intelligence, Evolution of; Intelligence, Genetics of: Cognitive Abilities; Memory in the Fly, Genetics of

Bibliography

Allen A, Anderson B, Andrews L, Beckwith J, Bowman J et al. 1996 The Bell Curve: statement by the NIH-DOE Joint Working Group on the ethical, legal, and social implications of human genome research. *American Journal of Human Genetics* **59**: 487–8

Binet A A S T 1905 Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux [New methods for diagnosing the intellectual level of abnormal persons]. *Année Psychologique* **11**: 191–336

Bouchard T J, Lykken D T, McGue M, Segal N L, Tellegen A 1990 Sources of human psychological differences: the Minnesota study of twins reared apart. *Science* **250**: 223–28

Bouchard T J, McGue M 1981 Familial studies of intelligence: a review. *Science* **212**: 1055–9

Cardon L R, Fulker D W 1994 A model of developmental change in hierarchical phenotypes with application to specific cognitive abilities. *Behavior Genetics* **24**: 1–16

Cardon L R, Fulker D W, DeFries J C 1992 Multivariate genetic analysis of specific cognitive abilities in the Colorado adoption project at age 7. *Intelligence* **16**: 383–400

Carroll J B 1993 *Human Cognitive Abilities*. Cambridge University Press, Cambridge, UK

Cattell R B 1971 *Abilities: Their Structure, Growth, and Action*. Houghton Mifflin, Boston, MA

Cavalli-Sforza L L, Feldman M W 1978 Dynamics and statistics of traits under the influence of cultural transmission. In: Morton N E, Chung C S (eds.) *Genetic Epidemiology*. Academic Press, New York, pp. 133–43

Cloninger C R, Rice J, Reich T 1979a Multifactorial inheritance with cultural transmission and assortative mating. II. A general model of combined polygenic and cultural inheritance. *American Journal of Human Genetics* **31**: 176–98

Cloninger C R, Rice J, Reich T 1979b Multifactorial inheritance with cultural transmission and assortative mating. III. Family structure and the analysis of separation experiments. *American Journal of Human Genetics* **31**: 366–88

Devlin B, Daniels M, Roeder K 1997a The heritability of IQ. *Nature* **388**: 468–71

Devlin B, Fienberg S E, Resnick D P, Roeder K 1997b *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve*. Springer, New York

Feldman M W, Cavalli-Sforza L L 1979 Aspects of variance and covariance analysis with cultural inheritance. *Theoretical Population Biology* **15**: 276–307

Flynn J R 1987 Massive IQ gains in 14 nations: what IQ tests really measure. *Psychological Bulletin* **101**: 171–91

Fraser S 1995 *The Bell Curve Wars: Race, Intelligence, and the Future of America*. BasicBooks, New York

Gould S J 1996 *The Mismeasure of Man*. Norton, New York

Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York

Horn J L 1985 Remodelling old models of intelligence. In: Wolman B B (ed.) *Handbook of Intelligence: Theories, Measurements, and Applications*. Wiley, New York pp. 267–300

Hunt E 1997 The concept and utility of intelligence. In: Devlin B, Fienberg S E, Resnick D P, Roeder K (eds.) *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve*. Springer, New York, pp. 157–76

Jensen A R 1969 How much can we boost IQ and scholastic achievement. *Harvard Educational Review* **39**: 1–123

Lewontin R C, Rose S P R, Kamin L J 1984 *Not in Our Genes: Biology, Ideology, and Human Nature*, 1st edn. Pantheon Books, New York

Lynch M, Walsh B 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA

Martin N G, Eaves L J, Heath A C, Jardine R, Feingold L M et al 1986 Transmission of social attitudes. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 4364–8

McInerney J D 1996 Why biological literacy matters: a review of the commentaries related to The Bell Curve: intelligence and class structure in American life. *The Quarterly Review of Biology* **71**: 81–96

Morton N E, Rao D C 1978 Quantitative inheritance in man. *Yearbook of Physical Anthropology* **21**: 12–41

Otto S P Feldman M W Christiansen F B 1994 Genetic and cultural transmission of continuous traits. In: *Morrison Institute Working Paper*. Stanford University, Stanford, CA, pp. 1–91

Paine R S 1957 The variability in manifestations of untreated patients with phenylketonuria (phenylpyruvic aciduria). *Pediatrics* **20**: 290–302

Pedersen N L, Plomin R, McClearn G E 1994 Is there G beyond *g*? (Is there genetic influence on specific cognitive abilities independent of genetic influence on general cognitive ability?) *Intelligence* **18**: 133–43

Plomin R, Loehlin J C 1989 Direct and indirect IQ heritability estimates: a puzzle. *Behavior Genetics* **19**: 331–42

Rao D C, Morton N E, Cloninger C R 1979 Path analysis under generalized assortative mating. I. Theory. *Genetical Research of Cambridge* **33**: 175–88

Rao D C, Morton N E, Yee S 1976 Resolution of cultural and biological inheritance by path analysis. *American Journal of Human Genetics* **28**: 228–42

Rice J, Cloninger C R, Reich T 1978 Multifactorial inheritance with cultural transmission and assortative mating. I. Description and basic properties of the unitary models. *American Journal of Human Genetics* **30**: 618–43

Schiff M, Lewontin R C 1986 *Education and Class: The Irrelevance of IQ Genetic Studies*. Oxford University Press, Oxford, UK

Spearman C 1904 General intelligence, objectively determined and measured. *American Journal of Psychology* **15**: 201–93

Stern W 1916 Der Intelligenz Quotient als Masz der kindlichen Intelligenz, inbesondere der Unternormal [The Intelligence Quotient as a measure of children's intelligence, especially for those below normal]. *Zeitschrift für angewandte Psychologie* **11**: 1–18

Vernon P E 1979 *Intelligence: Heredity and Environment*. Freeman, San Francisco

Wahlsten D 1997 The malleability of intelligence is not constrained by heritability. In: Devlin B, Fienberg S E, Resnick S E, Roeder K (eds.) *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve*. Springer, New York, pp. 71–87

Weinberg R A, Scarr S, Waldman I D 1992 The Minnesota transracial adoption study: a follow-up of IQ test performance at adolescence. *Intelligence* **16**: 117–35

S. P. Otto

Intelligence: Historical and Conceptual Perspectives

1. *Introduction and Brief History*

The term *intelligence*, as it is used in modern psychology, refers to individual differences in the ability to acquire information, and to use that information to solve new problems. The emphasis on individual differences distinguishes theories of intelligence from theories of cognition. Theories of cognition try to explain how the brain–mind system works in general, while theories of intelligence attempt to explain how people vary in their cognitive ability.

Comments on intelligence have a long history. Homer's *Odyssey* repeatedly refers to the hero's superior thinking. In the 1500s the Spanish philosopher Juan Huarte de San Juan wrote extensively on the difference between people who solve problems through their memory and those who rely on imagination. Huarte also speculated that the differences were due to differences in biological capabilities, but his investigations were restricted by the limited (and often erroneous) biological theories of his time. Truly scientific theories of intelligence had to wait until the nineteenth century, when Francis Galton proposed that individual differences in mental capacity were reflections of individual differences in the efficiency of the central nervous system. This lead Galton to try to develop tests of 'nervous functioning,' such as the speed with which a person could make a motor movement in response to a signal, and then to try to relate performance on these tests to performance in complex activities, such as success in college courses. At the time this effort was not seen as a success, and it was abandoned. Subsequently more sophisticated statistical analyses have shown that this conclusion may be premature. Some of our modern theories of intelligence can be traced to Galton's ideas. Shortly after Galton's proposal Alfred Binet, a French scientist, was asked to develop a test that could be used to identify children who were likely to fail in the French public school system. Binet took a pragmatic approach. He assumed that (a) mental competence grows as a child ages and (b) that children who are ahead or behind normal development at one age are likely to be similarly ahead or behind at a later age. Accordingly, Binet developed an intelligence test that was actually a 'battery' of subtests, in which children were asked to perform a variety of tasks that were 'typical of their age.' For instance, an average four-year-old child was supposed to be able to repeat back three digits: an eight-year-old should repeat back five digits.

The most widely used modern versions of intelligence tests, such as the Stanford–Binet test (derived from Binet's original work by Lewis Terman, a professor at Stanford University) and the Wechsler adult and child intelligence scales, retain Binet's emphasis on a battery of tasks, rather than relying on a single task to measure intelligence. Tests that are used to evaluate candidates for jobs and academic programs also use the battery approach. The Scholastic Assessment Test (SAT), widely used in the USA for college admission, for instance, consists of separate batteries evaluating linguistic and mathematical skills. So does the Armed Services Vocational Aptitude Battery (ASVAB), which is used as a selection device for the United States Armed Services.

There has been concern that tests such as these, which assume that all examinees have had certain experiences, are simply evaluating a person's exposure to different cultures. This is unlikely, at least for the range of countries represented by modern industrialized society, for testing seems to serve the same purpose in European, North American, and industrialized Asian nations. However there have been numerous attempts to develop tests that used the same type of problem (but at different levels of difficulty) rather than to take the test-battery approach. The most successful of these tests require the examinees to find patterns in observations, and especially in geometric patterns.

2. *One Intelligence or Many?*

As different tests were developed a surprising finding emerged. Although tests of vocabulary, paragraph comprehension, arithmetic problem-solving, or pattern recognition involve different processes on their face, in practice there are moderate to high correlations between the various types of tests and subtests. The correlation coefficients, r, usually lie in the range 0.6 to 0.9. This suggests that there is a single individual difference, 'general intelligence,' that determines a person's mental competence almost independently of the way in which that competence is expressed. While this turns out to be a reasonable approximation for

many purposes, research in the 1990s demonstrated a more complex picture. The following facts seem to be a more accurate statement of the situation:

(a) In the range from normal to high mental competence there appear to be three factors of intelligence. One is *Fluid Intelligence* (Gf), which can be thought of as the ability to see new patterns in data and develop new solutions to novel problems. The second is *Crystallized Intelligence* (Gc), which reflects the ability to apply previously learned information to the current problem. The third is *Visual–Spatial Intelligence*, which reflects the ability to manipulate visual images and see visual relationships.

(b) Measures of fluid and crystallized intelligence are correlated (r about 0.6) in most studies, suggesting but not proving that they may be partly a reflection of a general intelligence (*g*) factor. An alternative explanation for the correlation is that the Gc and Gf depend upon different combinations of elementary abilities, such as pattern recognition, short term memory, and ability to maintain attention. At the present time we do not know which of these interpretations is correct.

(c) Visual–spatial reasoning has a much lower correlation with Gc and Gf than the two intelligences have with each other. This indicates that visual—spatial reasoning is a somewhat different ability than intelligence, as indicated by Gc and Gf.

(d) The above remarks apply in the normal–above normal-gifted range. Test scores are much more highly correlated at the lower end of the scale than at the upper end. Put a slightly different way, unusually high scores in one type of test (say, Gc) may be associated with a fairly wide range of scores on a different test (say, Gf). However unusually low scores on one type of test are usually associated with low scores on another type of test. Put somewhat crudely, the case for general intelligence is weak, but the case for general stupidity is strong.

There are a few exceptions to statement (d). Turner's syndrome is a genetic disease that affects women. Turner's syndrome patients have low visual–spatial intelligence test scores, but score in the normal range on tests requiring verbal reasoning. People who have suffered damage to specific areas of the brain may lose mental competence in one area, such as memory, without necessarily losing competence in other areas. (People who suffer widespread brain damage, not surprisingly, show widespread deterioration of functioning.) These exceptions aside, statements (a) to (d) above are a reasonable summary of the distributions of test scores in normal populations.

3. *Do Test Scores Matter?*

A person's score on an intelligence or aptitude test is of little difference in itself. It becomes interesting only if test scores predict performance in some socially important arena, such as schools or the workplace. There has been considerable controversy over whether the tests (a) predict nothing at all, (b) predict performance in academic settings only, or (c) predict performance in both academic and workplace settings. The facts are fairly well known. The disagreement is over what to make of them.

In university settings the correlation between test scores and grades is in the range 0.3 to 0.4. The correlations are somewhat higher for students going through technical training, such as Armed Services technical training schools, at least partly because these students represent a wider range of abilities than university students. These correlations do not allow for the fact that performance scores can only be obtained for those people who have a high enough test score to be enrolled or employed in the first place. When statistical adjustments are made to allow for this selection effect the correlation rises to about 0.5. Moving outside of academics, similar scores are obtained in industrial settings. For instance, one study found a correlation of 0.38 between scores on an SAT-type test given to management trainees and the level of management position that the trainees achieved 20 years later. Similar correlations have been observed in many studies relating test scores to industrial performance, although the range of obtained correlations is generally higher in industrial than academic studies. This may be because it is difficult to obtain accurate measures of personal performance in many industrial settings.

Mathematically, a correlation of 0.35 indicates that about 10 percent in the variance of one score is predictable from knowledge of another. In this case, it means that 10 percent of the variance of academic or industrial performance would be predictable from knowledge of a cognitive test score. The higher 0.5 estimate raises the variance effect to 25 percent. Both of these values are some distance from perfect prediction. However, 'variance' is a statistical term whose meaning does not easily translate into the everyday concept of 'variation.'

There are three other ways of looking at the situation. One is to consider what sorts of correlations are typically found between different aspects of human variation. For instance, the correlation between adult height and weight (within sexes) is about 0.5, not far from the estimated population value for test-performance relations.

A second way to look at the issue is to ask whether there are alternative predictors that do as well as tests of cognitive competence. The evidence here is striking. The typical personality test has a correlation of about 0.20 with indices of workplace performance, just slightly more than one-half of the value for a cognitive test. However, it is important to note that the personality and cognitive tests appear to predict different aspects of workplace performance, for prediction from a combination of cognitive and personality tests is

higher than the prediction from a cognitive test alone. If just one measure of prediction of performance is to be used, a cognitive test will be a better predictor than a personality test. However, the ideal selection system would use both types of predictor.

A third way to try to evaluate the importance of intelligence is to ask how much workplace or academic performance is improved by using a test as a screening device. The answer to this question depends upon the correlation between test scores and performance and the extent to which the academic or workplace institution can select candidates. If all applicants are to be accepted, then there is no point in using any test. On the other hand, if only 10 percent of the applicants are to be accepted, then a test that has a correlation of 0.3 with performance can be used to select a workforce (or student body) that, on the average, will outperform an unselected group by more than 150 percent. This shows that the use of cognitive tests for selection is most justified when (a) the selection is rigorous and (b) there is a substantial difference between the benefits of good and poor performance.

Finally, there are interesting observations about the distributions of test scores in the population. It is well established that people who have high test scores as adolescents or young adults will, on average, tend to get better jobs and do better economically. It is even true that people with high test scores tend to live longer than those with low test scores! However, it is very difficult to establish cause and effect from such studies, because education, health, and economic success are themselves positively correlated.

4. *What Causes Intelligence?*

When we ask 'what causes intelligence?' it is important to distinguish between two different ways that this question can be asked. At the individual level, we can ask what characteristics of a person cause them to be or not to be intelligent. At the population level, we can ask what measurable demographic variables are associated with intelligent performance. We look first at the individual level and then the population level.

4.1 *The Individual Level*

At the individual level, it becomes important to distinguish between physical and social causes of cognitive competence, and between fluid and crystallized intelligence (Gc and Gf). People who perform well on tests of fluid intelligence and reasoning are, to a very large extent, people who are adept at keeping track of several things at once, and who can manage fairly large amounts of information. This ability is known as an ability to manage 'working memory' for information relevant to the problem currently being worked on. It is contrasted with 'long-term memory,' which refers to memory for information about how the world works and information about biographic events that happened some time ago. Working memory, in turn, appears to be related to the functioning of the frontal and prefrontal areas of the brain. There is now a good deal of evidence suggesting (but not yet proving) that individual differences in the ability to activate information in these areas of the brain is related to scores on tests of fluid intelligence. Just what these relations are, and how they relate to individual differences in neural functioning in other parts of the brain, is an important topic in research on intelligence. While individual differences in fluid intelligence are related to individual differences in brain functioning, it is important not to exaggerate the relationship. Statistical analyses suggest that individual differences in brain functioning can account for only a part of the wide individual differences in the ability to deal with 'new and unusual problems,' that is, fluid intelligence. By default, the remaining influences must be social (e.g., education, early training that might set a particular style for problem solving), but as of this article psychologists have been unable to determine what these environmental influences are.

The situation is quite different for tests of crystallized intelligence, that is, the ability to recognize and apply previously acquired solutions to new situations. In part this ability appears to depend upon generalized pattern-recognition abilities in the brain. However, it is also extremely responsive to education and training. This is seen most clearly by studies of *expertise*, individual differences in solving problems within a particular domain. Study after study has shown that specialized problem-solving depends upon the acquisition of schematic forms of reasoning appropriate to the domain at hand. Furthermore, these schema are very largely acquired by extensive practice. This has been shown in domains as far apart as chess, physics, and economics. Fluid intelligence appears to be one determinant of how efficiently information can be acquired during the schema-acquisition period. (This may partially account for the correlation between measures of fluid and crystallized intelligence.) However, knowledge builds on knowledge, so soon the expert has an advantage over the novice in learning how to learn within a particular field. Whether or not there is a generalized ability to learn to learn, regardless of the field being studied, is an open question. If there is, this ability is probably quite closely related to fluid intelligence.

As would be expected, any influence that causes a deterioration of brain structure is likely to have a deleterious effect on intelligence. In modern industrialized society alcoholism is undoubtedly the biggest single effect; repeated studies have shown a negative correlation between excessive alcohol use and intelligence test performance. Once again, whether or not this is causal is hard to determine, as excessive alcohol use could either produce or be produced by

low intelligence. However, the effect can be transgenerational. A major cause of mild mental retardation in children is excessive alcohol use by the mother during pregnancy. Other negative influences on intelligence include prolonged malnutrition (primarily in the developing nations) and exposure to atmospheric lead.

Aging has a paradoxical effect on intelligence. Fluid intelligence test scores drop. This is not surprising, as there is considerable evidence for an age-related drop in performance in tasks involving working memory. However there are very large individual differences in the extent of the drop. If we compare the fluid intelligence test scores of otherwise comparable individuals in their 20s and 60s we find that the top scores of the older group are only slightly below the top scores of the younger examinees, but the lowest scores in the older group are considerably below the lowest scores in the younger examinees.

Crystallized intelligence, in the sense of scores on such things as vocabulary and knowledge tests, may rise slightly during the adult working years, and declines only slowly until people reach their 70s or beyond. Because we live in a specialized society this may actually underestimate the abilities of older people. People become quite competent in those things that they practice, so during their working years people may become 'cognitive specialists' in the tasks that they encounter every day. Perhaps for this reason, general cognitive tests, such as the Department of Labor's General Aptitude Test Battery (GATB), seem to underestimate the out-of-laboratory performance of older workers. Of course, these statements apply only so long as the workplace and social environments remain the same.

4.2 Population Level Variables

Another way to answer the question 'what causes intelligence?' is to look at the correlations between intelligence test scores and various other measures of individual differences. This sort of analysis can be quite informative, although it only indirectly provides an indication of the mechanisms that produce cognitive competence at an individual level.

One of the oldest questions in the study of intelligence is whether intelligence is inherited or acquired through experience. The answer, not surprisingly, is that it is partly inherited and partly acquired. Very many studies have been conducted in an attempt to estimate the percentage of variance in a population that is due to genetics. This is called the *heritability coefficient*. Conceptually, the clearest experiment is to examine the correlation between monozygotic (MZ) 'identical' twins raised apart to the correlation between MZ twins raised together. Such situations are rare but they do occur, usually in cases of adoption immediately after birth. Correlations as high as 0.8 have been reported between MZ twins raised apart. By contrast, the test scores of dizygotic (DZ) 'nonidentical' twins raised apart is usually in the range 0.4 to 0.5. If one believed that the environments for each of the adoptees were chosen randomly this data is consistent with a heritability coefficient of 0.8. However there are a variety of reasons to believe that these correlations are inflated to some unknown degree by common environmental factors (e.g., sharing a common uterine environment prior to birth). Other studies of correlations between family members of varying degrees of relation (e.g., parent–child, siblings, grandparent–grandchild), using statistical techniques too complicated to be easily explained, have lead geneticists to conclude that the heritability coefficient is somewhere in the range 0.4 to 0.6. It is important to realize that this value applies to a modern industrialized society, where the difference between the well-to-do and the poor in terms of nutrition, disease, and schooling are not large. In a society in which there were very wide differences in living conditions and educational opportunity we would expect environmental effects on intelligence to be large, and hence the heritability coefficient would decrease.

Notwithstanding these qualifications, one fact is clear. Within the range of environments found in a modern postindustrialized or industrialized society, the single most important influence on intelligence is heritability. It accounts for slightly over one-half of the variance in intelligence test scores, on a population basis. This does not mean that you can say that one-half of an individual's mental competence is due to genetics: the variance referred to applies only to the population as a whole. Given advances in understanding the human genome it may be possible, in the future, to associate specific genes with intelligence. When this is done it will be possible to discuss genetic effects on an individual basis. As of the year 2001, such research was in its infancy.

What in the environment determines intelligence? Three effects have been noted. One is schooling: people who receive formal instruction (and especially people who are literate and read a good deal) do better on intelligence tests. Another is the physical environment: certain diseases, environmental hazards, and pathogens can cause reductions in intelligence. The most notable are alcoholism, exposure to atmospheric lead, and any disease or physical insult that harms the brain. The last influence is not inevitable. Some brain injuries lead to a drop in intelligence, others do not.

A third, and particularly puzzling, environmental influence is the 'cohort effect' phenomenon. In industrialized nations the population average on intelligence tests rose throughout the twentieth century. (We do not have comprehensive data for nonindustrialized nations.) The rise was particularly marked at the low end: on a percentage basis there are fewer people with very low scores (IQ below 80) than there used to be. The cause for this rise is unknown. However, it is

clearly an environmental effect, for the change is far too rapid to be due to genetic changes. Understanding genome patterns for intelligence (see above) and understanding the reason for the cohort effect are perhaps the most interesting questions in research on intelligence as we progress through the twenty-first century.

5. *Who Has Intelligence? Demographic Variables*

The question of whether or not different demographic groups differ in intelligence is one of the most hotly debated, socially contentious issues in the social sciences. It turns out that the facts are relatively clear. The reasons for those facts, however, are far from clear.

Male–female differences present a fairly straightforward case. While absolute equality cannot be proven, there are no marked differences between men and women in either fluid or crystallized intelligence. It is possible that women exhibit a slight advantage on some verbal tasks, but the size of the effect is not large. Men, on average, do considerably better than women on visual–spatial reasoning tasks that involve visualization or perceptual judgments of motion. This is important, because such judgments are required in tasks that involve spatial orientation, such as walking around a building and then trying to imagine what it would look like from above. The difference between population means on such tasks is about 0.5 standard deviation units. In less statistical terms this means that about 70 percent of the male population scores above the median score for the female population. Because visual–spatial reasoning has been shown to be correlated with achievement in mathematics, it has been suggested that the female difference in visual–spatial reasoning is one of the reasons that women are not proportionately represented in jobs that depend on mathematical reasoning. While this is a reasonable conjecture, with some evidence to support it, social factors are also responsible for this difference in occupational achievement.

There are racial and ethnic differences in intelligence test scores. This statement is largely based on studies in the USA, although a few studies from Europe (where there is less ethnic diversity) show generally the same pattern. In tests carried out in the late 1990s, 'African-Americans' scored about 0.8 standard deviation units below the 'White' mean (and it should be noted that this latter social category includes many disparate groups), while 'Asian-Americans' scored approximately 0.3 units above the 'White' mean. People identified as 'Hispanic-Americans' (who may be of any race) scored between the 'African-American' and 'White' mean, and there were differences between particular groups of 'Hispanic-Americans'. Furthermore, there was no evidence that the tests were unfair to particular groups, in the sense that the tests predicted performance more accurately in one group than another. The correlations between test scores and academic or job performance (and other relevant statistics) were substantially the same across all groups. This indicates that the cognitive skills that are (a) tapped by the test and (b) required in academics and the workplace are not equally distributed across racial and ethnic groups in the USA.

These statistics do not provide a reason for the differences. Environmental effects are clearly involved, for greater opportunity for social equality has been accompanied by a lessening of the difference between groups. For instance, the 'African-American' vs. 'White' difference was about 1.2 standard deviation units prior to the end of legal segregation and separate, usually inferior, educational opportunities for African-Americans. This clearly indicates a social influence. However the rate at which the gap is closing has slowed since about 1980. This and other evidence has led some individuals to claim that ethnic differences are partly due to genetic differences between groups. The claim has been strongly challenged by other researchers. The controversy is impossible to resolve at this time, largely because genetic differences between ethnic groups are confounded with social differences in economic status, living conditions, family structures, and a host of other environmental variables. To make the situation even more confusing, it is logically possible that differences between groups A and B are genetic, while the differences between A and C are environmental. Perhaps some day advances in molecular biology will make it possible to trace genetic influences on intelligence at the single-gene level. It is also possible that in the future a better understanding will be obtained of the way special environmental factors, such as childrearing practices or schooling, influence intelligence. Until this research is done any causal 'explanation' of racial and ethnic differences in cognitive competence will be highly speculative.

6. *Final Comment*

A century of modern research on intelligence has led to a great many interesting conclusions. Perhaps the most important is that individual differences in cognitive competence are the best single predictors of success in industrialized and information-age societies. This does not mean that intelligence is the *only* thing that is important, but it does mean that it is important. The changes in population levels of intelligence throughout the twentieth century are not yet understood. When they are, though, they may well lead to ways of improving human cognitive potential in the twenty-first century.

See also: Cattell, Raymond Bernard (1905–98); Genetic Factors in Cognition/Intelligence; Intelligence: Central Conceptions and Psychometric Models; Intelligence, Evolution of; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heri-

tability and Causation; Intelligence: History of the Concept; Intelligence, Prior Knowledge, and Learning; Lifespan Theories of Cognitive Development

Bibliography

Hunt E 1995 *Will We Be Smart Enough? A Cognitive Analysis of the Coming Workforce*. Russell Sage, New York
Mackintosh N J 1998 *IQ and Human Intelligence*. Oxford University Press, Oxford, UK
Sternberg R J (ed.) 1994 *Encyclopedia of Intelligence*. Macmillan, New York
Sternberg R J, Detterman D K (eds.) 1986 *What is Intelligence?* Ablex, Norwood, NJ

E. Hunt

Intelligence: History of the Concept

The direct antecedents of the word 'intelligence' lie in the Latin '*intelligentia*' or '*intellegentia*' meaning 'the action or faculty of understanding,' itself derived from the Latin '*intellegere*' meaning 'to understand.' As one of the traditional divisions of the human soul (along with the emotions and the will), intelligence was a concept that had long been of importance to Western philosophers and theologians, especially in their attempts to differentiate human beings from other species in the animal world. Nonetheless, until the eighteenth century, it was not a word that piqued much general interest on either side of the Atlantic. Intelligence began to attract attention in the West as growing curiosity about the nature of human differences meshed with the turn to scientific methods as a privileged form of explanation. This essay will explore some of that history and suggest how the development of the concept was intimately related to the particular cultural circumstances in which it was formulated and used.

1. Species, Groups, and Intelligence

Although human nature had become a topic of hot concern in the eighteenth century, of relevance not only to the emerging human sciences but also to many of the most important developments in political philosophy as well, this did not at first spur much interest in intelligence. Rather, the focus tended to be on other concepts, on 'faculties,' 'talents,' 'sensations,' and, most especially, 'reason' and 'virtue.' While it is clear that there was widespread agreement that human beings differed in their mental powers, the origin, significance, and even the sheer number of those differences were all issues that elicited either a multitude of opinions, or no particular view whatsoever. In general, single all-encompassing approaches to understanding mental capacity, such as 'intelligence,' tended to be used only in limited and very particular contexts.

Intelligence began to become an object of direct scientific investigation toward the end of the eighteenth century, when such naturalists as Linnaeus, Buffon, and Blumenbach developed taxonomic schemes that compared human beings systematically with the rest of the animal kingdom, often focusing in part on differences in overall levels of mental ability. The Dutch anatomist and artist Peter Camper, for example, proposed in the 1760s that a scale of intelligence across the animal kingdom could be erected on the basis of a particular physiological feature, the facial angle, which first allowed intelligence to be converted into a type of measurable characteristic, though one operating primarily at the level of species or groups. Over the next century and a half, a number of other physical features were also evaluated as stand-ins for direct determinations of degree of intelligence, with most physical anthropologists according particular significance to cranial capacity and brain weight. Elaborate programs of measuring hundreds of skulls were carried out throughout Europe and the United States, with the results frequently used to provide scientific sanction for common beliefs about the intellectual inferiority of certain groups, most especially Africans and women, on the basis of the greater proximity of their mean skull or brain sizes to those for the non-human primates. As more data were collected, however, extreme variations were found both within and between groups, rendering group-level rankings in the eyes of many largely meaningless. At the same time, the efforts of women in feminist movements and blacks in anti-slavery campaigns produced powerful denunciations of the entire project of creating hierarchies of intelligence based on physical characteristics.

At the level of individuals, until the latter part of the nineteenth century intelligence as a personal characteristic distinguishing one individual from another generated little systematic interest, save for medical concern with diagnosing the profoundly mentally deficient. Mental philosophy and the emerging discipline of 'scientific' psychology were preoccupied with the universal characteristics of human reason and, in the case of psychology, with the process of subjecting mental features to new laboratory techniques being imported from physiology. As the evolutionary theories of Charles Darwin and Herbert Spencer became more broadly accepted, however, interest grew in the overall mental power of an organism, because it was deemed to be a critical characteristic shaping the organism's adaptation to its environment. Greater pressure to understand the nature of individual differences was also developing as the result of a number of institutional and social changes occurring throughout the West. The spread of universal primary education, the increase in urban concentrations of workers, and

the expansion of factory methods of production, along with the rapid growth in asylums for the mentally deficient, all provided many more opportunities to engage in systematic comparisons at an individual level, and thus helped to make intellectual differences much more visible and socially meaningful.

2. Intelligence and Its Tests

During the late nineteenth century, a number of investigators began to explore the meaning of individual differences, including those associated with overall level of intelligence. In England, Francis Galton carried out an extensive program of quantifying human mental and physical characteristics (anthropometrics) toward the goal of identifying superior members of the human stock. His *Hereditary Genius* ([1869], 1972) argued that innate differences in overall superiority existed within human groups, that those differences were mental in origin, and that they were inheritable. Galton's work helped to initiate widespread enthusiasm for anthropometry, one that persisted until the end of the nineteenth century. It was also a basis for his promulgation of the science of eugenics, which conceived of human progress in biological terms and advocated improvement in the quality of the breeding stock through encouraging marriages and child bearing among the most 'fit' (i.e., most intelligent) and restricting them among the least.

On the other side of the English Channel, psychologist Alfred Binet, one of the pioneers in the development of *psychologie individuelle* in France, conducted a wide range of investigations at the turn of the twentieth century into the nature of human mental ability. As one of these projects, in collaboration with Théodore Simon, Binet developed the initial version of what would become the first successful individual intelligence assessment instrument, the Binet–Simon Intelligence Scale (1905), a group of age-related tasks whose purpose was to reveal whether an examinee's overall intelligence was developing at a 'normal' rate. The measuring aspect of the scale was enhanced following its 1908 and 1911 revisions, when results from the testing were reported in terms of mental age (MA), which facilitated conceiving of intelligence as a physical and quantifiable object.

Binet's death soon after the publication of the 1911 scale robbed the Binet–Simon of its most effective champion in France and interest in intelligence testing there soon waned. Such was not the case elsewhere in the West, however, most especially in the United States. With the help of American psychologist Henry Herbert Goddard, who introduced the 1908 version of the scale to his fellow US psychologists in 1910, the Binet–Simon was adopted with enthusiasm. Rival versions of the test were soon produced in great numbers and a vogue for testing developed that never completely vanished.

Stanford University psychologist Lewis M. Terman's 1916 revision of the scale, his Stanford–Binet (Terman 1916), did much to cement the place of intelligence testing within the American intellectual and cultural landscape. Thoroughly overhauled for an American population, the Stanford–Binet provided clearer age demarcations and popularized a new way of quantifying mental ability, the Intelligence Quotient (IQ), a ratio of mental age to chronological age (times 100) designed to remain constant over time. Terman celebrated this feature of IQ, argued that it provided a measure of an individual's innate mental ability and suggested that a range of social decisions, including the nature of the occupation an individual should take up and whether or not he/she should be institutionalized, could be made on the basis of this number.

While not every mental tester was as certain about the implications of IQ as Terman, during the 1920s use of intelligence tests in a variety of contexts spread widely, facilitated by the creation of a group mental test for the World War I war effort that allowed 1.75 million US army recruits to be assessed without needing the time-consuming one-on-one methods of the Stanford–Binet. Terman's individual scale, revised in 1937 and 1960, however, remained the gold standard for measurements of intelligence, though its preeminence began to be challenged starting in the 1940s with the creation by David Wechsler of the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Adult Intelligence Scale (WAIS), instruments that generate assessments of verbal and nonverbal intelligence as well as an IQ.

3. The Nature of Intelligence

Concomitant with the rise of intelligence testing came a series of debates over the characteristics of the object being measured. In 1904, British psychologist Charles Spearman used early test data to argue for the unitary nature of intelligence, explaining performance on mental tests in terms of general intelligence (*g*) and specific abilities (*s*). While numerous researchers—including Karl Pearson, Goddard, and Terman—accepted his analysis, others were skeptical, insisting instead that intelligence was composed of a number of primary independent abilities. Edward L. Thorndike in the United States was among the first to articulate this position, and he was soon joined during the 1920s by two statistically sophisticated psychologists, L. L. Thurstone in the US and Godfrey Thomson in the UK. In the period after World War II, psychologists continued to put forward a range of interpretations of the composition of intelligence: while Hans Eysenk remained convinced of the reality of *g*, for example, Philip E. Vernon proposed a hierarchical model of intelligence that inter-linked specific skills, general abilities, and overall intelligence, and Joy P. Guilford contended that intelligence was composed of as many as 150 independent factors. Later influential additions

to these models included Howard Gardner's theory of the existence of seven discrete types of intelligence and Robert J. Sternberg's triarchic conception of intelligence.

Overshadowing all of the arguments over intelligence, however, has undoubtedly been the nature–nurture question. Figures such as Galton, Spearman, Pearson, and Terman argued early in the twentieth century strenuously for the nature position, with Galton producing studies on identical twins that have served as a model for investigations into the relative weights of heredity and environment. During the 1900s to 1930s, when eugenics was at the apex of its popularity, arguments in favor of intelligence as an inheritable biological entity ran strong and were used to justify calls for immigration restriction and for the sterilization of the mentally 'unfit,' as well as for the creation of multi-tracked secondary schools. At the same time, however, anthropologists were beginning to put renewed emphasis on culture as the primary determinant of human behavior, claims strengthened during the middle of the twentieth century by studies carried out especially at the Iowa Child Welfare Research Station, where IQ was found to change depending on nutrition and educational environment. The post-war period continued to see the debate pressed from both sides, with increasingly sophisticated twin studies showing high IQ correlations between identical twins separated at birth, while at the same time other researchers were teasing out ever more complicated connections between intelligence development and such factors as nutrition, family child-rearing practices, socio-economic status, and quality of education received.

While few experts would deny the influence of both genes and environment, the vociferous debate in the mid-1990s over *The Bell Curve* (Herrnstein and Murray 1994), with its claims that IQ is hereditary and a prime determinant of an individual's future, indicates that broad disagreements persist about intelligence and show little likelihood of quick resolution. What they reveal as well is that the language of native intelligence has remained a powerful vehicle for discussions of a range of social issues, from the organization of the educational system to the value of affirmative action programs to the just allocation of social resources.

See also: Artificial Intelligence in Cognitive Science; Artificial Intelligence: Uncertainty; Binet, Alfred (1857–1911); Cognitive Psychology: History; Cognitive Psychology: Overview; Deductive Reasoning Systems; Eugenics, History of; Galton, Sir Francis (1822–1911); Intelligence: Central Conceptions and Psychometric Models; Intelligence, Evolution of; Intelligence: Historical and Conceptual Perspectives; Rationality in Society; Thorndike, Edward Lee (1874–1949)

Bibliography

Danziger K 1997 *Naming the Mind: How Psychology Found its Language*. Sage, London

Degler C N 1991 *In Search of Human Nature: The Decline and Revival of Darwinism in American Social Thought*. Oxford University Press, New York

Galton F [1869] 1972 *Hereditary Genius: An Inquiry into Its Laws and Consequences*. Peter Smith, Gloucester, USA

Gould S J 1981 *The Mismeasure of Man*. Norton, New York

Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York

Rose N 1985 *The Psychological Complex: Psychology, Politics and Society in England, 1869–1939*. Routledge & Kegan Paul, London

Smith R 1997 *The Norton History of the Human Sciences*. Norton, New York

Sutherland G 1984 *Ability, Merit and Measurement*. Oxford University Press, Oxford, UK

Terman L M 1916 *The Measurement of Intelligence: An Explanation of and a Complete Guide for the Use of the Stanford Revision and Extension of the Binet–Simon Intelligence Scale*. Houghton Mifflin, Boston

Zenderland L 1998 *Measuring Minds: Henry Herbert Goddard and the Origins of American Mental Testing*. Cambridge University Press, Cambridge, UK

J. Carson

Intelligence: Organizational

For Weber (1978) and his heirs the intelligent organization was a bureaucratic organization. The files served as the repository of knowledge. Intelligence resided in the institutionalization of routines (see *Organization: Overview*; *Bureaucracy and Bureaucratization*). Later it became seen as functionally specific to positions at the apex of the organization or in the technostructure. Today, it is necessary to rethink the concept in terms of virtuality and knowledge management.

1. Organizational Intelligence as a Historically Situated Term

Organizational intelligence—defined as the questions, insights, hypotheses, evidence that were relevant to policy—enters organizations through their strategic apex as clear, timely, reliable, adequate, and wide-ranging strategic information that informed embedded organizational routines. Wilensky (1967), from whom this definition is taken, regarded organizational intelligence as a variable that was directly dependent upon the degree to which there was, organizationally:

(a) Conflict or competition in relation to the external environment—typically related to the extent of involvement and dependence on government (see *Conflict: Organizational*);

(b) Internal support and unity;

(c) A belief in the rationalization of internal operations and external environment; and

(d) A contingent development of size and structure, goal and member heterogeneity, and the centralization of authority (structural contingency).

Wilensky's (1967) view was that the more organizations were developed in these terms the greater would be the need for and existence of organizational intelligence, embodied in specific types of functions, such as 'contact men,' 'internal communications specialists,' and 'facts-and-figures men.' His views were very much informed by the experiences of the military, defense, and intelligence agencies of the US in the Second World War and Cold War eras. In particular, he was concerned with the way that organization routines, embedded in discursive patterns, would often dictate policy long after it should have been evident that it was failing. He found examples in both the British and US strategic bomber commands of the Second World War as well as in the Vietnam War that raged at the time that he was writing. Were he writing more recently, he would, no doubt, have had something similar to say about the NATO offensive against Serbia that occurred in 1999. Organizational routines and discursive patterns can often mean that even where strategic intelligence is available it may be unattended to or, if it is, not understood. Thus, organizational intelligence involves more than merely information: it involves also its application, even against discursively established routines. Organizational intelligence marks the ease with which such routines are abandoned.

2. *Contemporary Views of Organizational Intelligence*

Wilensky's (1967) view of organizational intelligence was informed largely by big, US military-industrial complex organizations. The picture changed dramatically during the 1980s, as learning from Japan became the watchword. Japanese organizations were still bureaucracies—but learning bureaucracies. What they were oriented to learning was continuous improvement of quality—and it was this that became seen as the nub of their specific organizational intelligence—learning not only to do existing things better but also to be better at innovation. The key involved unlocking total organizational intelligence rather than refining strategic intelligence (see *Innovation: Organizational*). Thus, the metaphors of intelligence shifted from defense to commerce (see *Organizations, Metaphors and Paradigms in*).

Using smart machines and robots for more routine work, Japanese manufacturers tried to create smarter workers for better products (see *Technology and Organization*). Organizations sought to develop relational contracts based on similar philosophies of continuous improvement as were applied in manufacturing (see *Alliances and Joint Ventures: Organizational*). Intensified global competition and the development of new digital technologies became the drivers that saw the lessons from Japan become widely distributed in existing industry, especially in the United States and Europe, by the end of the 1980s. The knowledge based information economy had arrived in which creativity, intelligence, and ideas were the core capability for sustainable success.

New information and communication technologies are crucial to such innovation processes, helping to globalize production and speed up the diffusion of technology. Information technology and globalization are transforming organizational concepts of time and space. The convergence of computing power and telecommunications reach is providing new technological and information resources in a global, digital world. As Hamel and Prahalad (1994) insisted, against Wilensky's (1967) view of entrenched structures, size, and centralization as strategic assets, such assets become liabilities when organization competitiveness is based on radical, nonincremental innovation (see *Strategy: Organizational*).

3. *Intelligent Information*

The development of information and communication technologies not only provides the means to process and transmit vast amounts of information but also determines the shape of organizational intelligence. If information and knowledge is to be used productively and intelligently by organizations then organizational intelligence rapidly translates into knowledge management rather than contacts, internal communications, and facts-and-figures (Wilensky 1967). Too much information is too easily available so that the key issue is not gaining information but being able to manage available knowledge.

Increasingly, knowledge and information is created and used as binary digits and transported electronically, rather than in physical forms. Ultimately, digitalization creates new structures of electronically networked organizations, replacing both individual market-based relations as well as bureaucratic structures whose resource base, in terms of size, comprises their chief strategic asset. Galbraith (1967), writing in the same year as Wilensky, although he did not foresee a digital economy, was one of the first to recognize how profound the organizational implications of the knowledge economy were in his prophetic statement of the arrival of the technocracy in *The New Industrial State*:

> With the rise of the modern corporation…the guiding intelligence…(passes to)…a collective and imperfectly de-

> fined entity…(that)…embraces all who bring a specialized knowledge, talent, or experience to group decision-making. I propose to call this organization the Technostructure (1985).

The new digital technostructure allows knowledge to be developed more readily as well as to be stored, accessed, and distributed more easily, thereby simplifying organizational transactions and allowing them to be conducted remotely. Hierarchical and centralized organizations of functional specialists applying standard procedures give way to flatter, more decentralized firms with flexible arrangements by professionals who rely on real-time information, catering for specific markets or customers.

Information technology is central to each stage of this transformation, potentially enabling global coordination of enterprise networks, distributed processing, portable work, and constant access.

Enthusiasm for information technology led to a fragmentation of technologies and approaches in many organizations around local networks. Incompatibility of software, computers, and systems undermined networking capability, provoking a determined search for open systems interconnection. Integration became the driving force as organizations attempted to coordinate all operations. Constantly improving information technology promotes changing conceptions of the role of information systems. What was originally an impenetrable device of limited utility has become the main strategic resource upon which operations and strategy are founded. It is the core of contemporary organizational intelligence.

Organization intelligence involving distributed knowledge, technology, and networks focuses attention on management control. Contemporary organizations create a growing variety of linkages along the value chain by developing extensive electronic and contractual relationships with networks of suppliers, customers, and partners. The intelligence created in these systems 'is not simply about the networking of technology but about the networking of humans through technology…not just…linking computers but…internetworking human ingenuity' (Tapscott 1996).

How is 'intelligence' to be exercised in line with managerial objectives in networks? Increasingly, the rules of organization bureaucracy give way to the patterning and design of organization culture. Unlike routines produced according to formal rule, there is more scope to innovate and establish experimentation around norms. The interaction of members in a group makes it possible to exchange information on successful practices. Under these conditions groups formulate similar patterns of perception, similar interpretations of some subjects, and similar evaluation of alternatives. There is an adaptation function whereby groups arrive at successful results to satisfy their needs. By following the norms of the group, people gain acceptance from the group. Acceptance by a group signifies affiliation or cohesiveness. From the organizational viewpoint, culture aids adaptation to the environment. 'Quality first' is an example. If members accept such a slogan and internalize it, it can become a pattern for decision-making. This is not to suggest a culture dreamscape where resistance and conflict have been eliminated in favor of boundless creativity. Such a picture rarely corresponds with empirical evidence (DeCieri et al. 1991). Instead, it suggests a world in which obsessive self-surveillance becomes the norm—a neurosis at the core of organizational intelligence in a digital world.

Information technology is used extensively to make performance and quality standards visible and accountable to employees in organizations. For instance, large-scale video screens can express quality-related information in a simple code that all can understand, alerting people to their degree of success in meeting standards. Two aspects of this are important. First, everyone's intelligence is enrolled to organizational effect through their becoming a constantly self-regarding subject. It is not only others who will hold one accountable—such as bosses and colleagues—but also one must hold one's self-accountable. The fact that others are able to see, transparently, what one is accountable for, helps focus one's energies on this enormously. One becomes a specialist in self-regarding behavior that helps to ensure that the need for external surveillance and exercise of power is minimized as the functions of control become internalized in oneself as a neurotically normal organism.

Second, this modeling of organizational intelligence in terms of individually reflexive self-control breaks down both collective strength and enhances individual isolation. The union function of articulating plural and countervailing power in the workplace becomes less relevant when, formally, unions either directly (as in Japan) or indirectly through their members are involved in the quality processes that define organizational intelligence. Where everyone is assumed to have an interest in quality to oppose it can be seen only as bloody-mindedly against the interests of employees whose jobs depend on continually improving quality—including oneself. Here is another significant discontinuity with Wilensky: organization intelligence becomes embodied in everyone, not just the 'contact men,' 'internal communications specialists,' and 'facts-and-figures men.' In these new forms of organization, structure loses its historic role of managing power relations at a distance. For one thing, distance disappears electronically; for another, power relations flatten as teams proliferate, work becomes a series of projects, and the supervisory gaze is both internalized and becomes part of peer pressure.

Davidow and Malone (1992) identified *The Virtual Corporation* as a distinctive model premised on new technologies making old assumptions irrelevant. Organization design becomes virtual, enabling time and space to be collapsed and the informational

Table 1
Modern and virtual organization compared on Weber's criteria

Modern Organization	Virtual Organization
Functionality in design structure	Defunctionalized project-based design held together by network capabilities
Hierarchy governing formal communication flows and managerial imperative the major form and basis of formal communication	Instantaneous remote computer communication for primary interaction; increase in face-to-face informal interaction; decrease in imperative actions and increased governance through accountability in terms of parameters rather than instructions or rules
The files	Flexible electronic immediacy through IT
Impersonal roles	Networking of people from different organizations such that their sense of formal organizational roles blurs
Specialized technical training for specific careers	Global, cross-organizational computer-mediated projects

Source: Nohria and Berkely, 1995

controls inscribed in bureaucracy which sought to manage across them, superseded.

> [T]he file cabinets of bureaucratic ritual disappear, replaced by devices that shatter the traditional physical instantiation of information and knowledge … When employees … use electronic mail or build reports from network databases, there is no original physical reality to which this information refers (Nohria and Berkely 1994).

Some find the virtual organization an attractive, if challenging, prospect, because it will have no '(P)re-established boundaries, and it will be conspicuous by the absence of hierarchy. It will be completely horizontally structured and geographically distributed organization … (with) … small cluster groups that are distributed throughout the world in network-intensive, computer mediated, interactive environments' (Estabrooks 1995).

Others are less sure that intelligent technologies necessarily make organizational intelligence. Groth (1999) presents a less sanguine, and probably more accurate, perception of the human emotional and organizational implications of information technology in which there is a continual acceleration in the rate of obsolescence of organizations' and individuals' knowledge base and skills—their intelligence. To counter this would require provision for the continual upgrading of skills and life-long learning: whether states or organizations will deliver this or whether it is left to markets, will be a core issue of future organizational intelligence.

The virtual organization is almost the exact opposite of the modern organizations that Weber (1978) first identified (see Table 1 in which we contrast Weber's modern organization with its 'virtual' counterpoint).

Virtual organizations are invariably networked organizations (see *Network Analysis*; *Interorganizational Relationships and Networks*). Networks may be characterized in terms of the strength or intensity of linkages, as well as the symmetricality, reciprocity, and multiplexity of their flows. The strength of a network linkage depends on the extent to which it is an 'obligatory passage point' in the network: can information flow elsewhere or must it route that way? The greater the amount of information, affect, or resources flowing through the passage point, then the more powerful will be those whose knowledge decodes it (Roberts and Grabowski 1996, Clegg 1989). The relations of different knowledge around these obligatory passage points embodied in different people, may be more or less symmetrical, that is, some will be more or less dependent or independent within the flow of relations. Reciprocity refers to the degree of mutual or nonreciprocal obligation that occurs in the relationship. Multiplexity refers to the degree to which those who relate to one another do so more or less exclusively. Finally, the content of the linkages is important, in terms of its degree of 'classification' and 'framing': how strongly or loosely framed or bounded it is, and how tightly or loosely coded is its classification. These notions of classification and frame serve to replace the earlier emphasis that Wilensky (1967) placed on variables of organizational structure and resources as key variables in the understanding of organizational intelligence.

4. *The Recursive Nature of Organizational Intelligence*

One hallmark of organizational intelligence is its fundamentally recursive character: learning transforms the existing stock of organizational know-how contained in existing normative routines and competencies. Within such a framework 'learning is no longer … objectified in norms, procedures, routines

and standards. Rather it is the cognitive activity which produces images, representations, causal links, and which is sensitive not only to human passions but to the social and organizational conditioning of thought as well' (Gherardi 1997) (see *Learning: Organizational*).

Organization learning occurs as a result of individuals learning becoming institutionally embedded in organizational ways of doing things. Where it is not, then learning will remain at the individual level. For organization learning to occur there has to be a culture that is conducive as well as management devoted to intelligence gathering, as Wilensky stressed. Many traditional tasks of management are now undertaken by computer-based technologies: managers thus must possess diagnostic, interpersonal, creative, and systems-thinking skills rather than be fact-finders. Management needs to enact interpretation, particularly where radical technological innovation creatively destroy or reduce existing competencies. Increasingly, such managers will seek to practice exploratory learning to increase organizational intelligence.

Exploratory learning is associated with complex search, basic research, innovation, variation, risk-taking, and relaxed controls. The stress is on flexibility, investments in learning, and the creation of new capabilities. Exploratory learning characterizes more intelligent organization, where radical innovation, rather than refining what already exists, produces creative discontinuities. Levinthal and March (1993) propose that the survival of any organization depends upon being sufficiently exploitative of what it already knows as to insure current viability and sufficiently exploratory as to insure its future viability. Too much exploitation risks organizational survival by creating a 'competency trap,' where increasingly obsolescent capabilities continue to be elaborated. Too much exploration insufficiently linked to exploitation leads to 'too many undeveloped ideas and too little distinctive competence.'

The digital interlinking of organizations allows for learning to be distributed globally, immediately, virtually, if the intelligence acquired in one location is potentially portable to others. To do this, breakthroughs must be standardized and exploited. Standardized information is a commodity. By making tacit knowledge a commodity the dependence of exploratory insights on the individuals who produce it is eliminated. Where management can reduce individual dependency by rendering knowledge into artifacts it is possible to manipulate and combine knowledge with other factors of production in ways that are impossible if it remains a human possession. Abstract properties need to be developed for the phenomena to become a standardized commodity—so that it becomes alienable, like a property title—entailing a shift from knowledge-workers to knowledge as a pure factor of production, usually embodied in systems and software.

It is easy to transmit exploratory learning embedded in innovation throughout the world when it is embedded in a systems and software. But these may not capture the embedded tacit knowledge involved in making exploratory innovation work. A solution is at hand: work processes can be videotaped, scanned on to computer, and downloaded instantly by the globally networked organization, thus disseminating it globally, instantly, to individual desktops and workstations. Thus, we are all potential 'contact-men' now and anyone connected to the World Wide Web, intranet, or Internet, can be an 'internal communications specialist' and 'fact-finder.' Organizational intelligence no longer belongs to a specific organizational entity or to its apex: it can be virtually everywhere.

See also: Information and Knowledge: Organizational; Intelligence, Prior Knowledge, and Learning; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations, Sociology of; Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Strategic Intelligence

Bibliography

Clegg S R 1989 *Frameworks of Power*. Sage, London

Davidow W H, Malone M A 1992 *The Virtual Corporation: Structuring and Revitalising the Corporation for the 21st Century*. Harper Collins, New York

DeCieri H, Samson D, Sohal A 1991 Implementation of TQM in an Australian manufacturing company. *International Journal of Quality and Reliability Management* **8**(5): 55–65

Estabrooks M 1995 *Electronic Technology, Corporate Strategy and World Transformation*. Quorum, Westport, VI

Gherardi S 1997 Organizational learning. In: Sorge A, Warner M (eds.) *Pocket International Encyclopedia of Business and Management: The Handbook of Organizational Behaviour*. International Thomson Business Press, London pp. 542–51

Groth L 1999 Future Organizational Design. The Scope for the IT-based Enterprise, Wiley, Chichester

Hamel G, Prahalad C K 1994 *Competing for the Future*. Harvard Business School Press, Boston

Levinthal D A, March J G 1993 The myopia of learning. *Strategic Management Journal* **14**: 95–112

Nohria N, Berkely J D 1994 The virtual organization: Bureaucracy, technology, and the implosion of control. In: Heckscher C, Donnellon A (eds.) *The Post-bureaucratic Organization: New Perspectives on Organizational Change*. Sage, Thousand Oaks, CA, pp. 108–28

Roberts K H, Grabowski M 1996 Organizations, technology and structuring. In: Clegg S R, Hardy C, Nord W (eds.) *Handbook of Organization Studies*. Sage, London, pp. 409–23

Tapscott D 1996 *The Digital Economy: Promise and Peril in the Age of Networked Intelligence*. McGraw Hill, New York

Weber M 1978 *Economy and Society: An Outline of Interpretive Sociology*. University of California Press, Berkeley, CA

Wilensky H L 1967 *Organizational Intelligence: Knowledge and Policy in Government and Industry*. Basic Books, New York

S. R. Clegg and T. F. Clarke

Intelligence, Prior Knowledge, and Learning

Intelligence test scores can account for interindividual achievement differences in many content areas to a considerable extent but they lose their predictive power when measures of prior content-specific knowledge are added. When controlling for differences in prior knowledge, medium correlations between intelligence and learning outcome are found. However, it is still far from clear exactly what kinds of learning activities are affected by intelligence. Education stimulates the emergence of intelligence but does not decrease the variance in intelligence test scores.

1. Intelligence: A Valid Predictor of Achievement Differences

Individuals with similar cultural, social, and educational backgrounds differ from one another in the time they need to process certain information, in their ability to understand complex ideas, in the efficiency with which they can deal with novel, transfer-demanding tasks, and in the learning outcome that results from attending certain instructional environments. The construct of psychometric intelligence attempts to clarify what is behind such achievement variations that cannot be explained by differences in learning environments or in amount of practice.

1.1 Measuring Intelligence

About a century ago Alfred Binet constructed problems designed to determine whether children who did not meet certain school requirements suffered from mental retardation or from behavioral disturbances. Since then many psychologists have been quite successful in developing reliable verbal and nonverbal intelligence tests for children and adults. Intelligence tests contain items composed of verbal, numerical, and pictorial material, and they require various mental operations, among them inductive and deductive reasoning, pattern recognition, and memorization. So-called speed tests contain items that are comparably easy for everybody; individual differences in the numbers of correctly solved problems only occur because of time limitations. In power tests the items are ordered according to their difficulty, and limits in intelligence become apparent if people do not solve all problems despite having sufficient time.

The distribution of achievement scores in all intelligence test scales follows the bell curve (normal distribution). This reflects the fact that the majority of people resemble each other quite a lot with respect to their cognitive capabilities, and only a few people show extraordinarily low or high competencies. Normal distribution is the statistical prerequisite for measuring intelligence on interval level by indicating deviations from the mean score. To determine the intelligence quotient (IQ), test scores are converted to a scale in which by convention the mean is 100 and the standard deviation is 15. The reliability of IQ tests, revealed either by correlation coefficients based on repeated measurement or by figuring out internal consistency, is between 0.80 and 0.90, which is higher than for most other psychometric measures. Nonetheless, a reliability lower than 1.0 only allows us to interpret a range rather than a single value. For example, if a person's tested IQ is 110, and if the reliability of the test is 0.90, the IQ of this person is between 101 and 119 with a probability of 95 percent. A test reliability of 0.80 reveals a range of 97–123.

Despite broad variations in the content and the form of presentation of intelligence test items, all tests have in common that they do not presuppose knowledge that can only be acquired in particular learning environments not accessible to everybody. Moreover, differences in test scores within a group only reflect differences in intelligence between these persons if all of them have had access to similar learning environments. In other words, each member of the group must have had, in principle, the opportunity to acquire the knowledge necessary to solve the problems. Intelligence test scores can be raised considerably by practicing the respective types of items, while individual differences do not disappear but rather remain quite stable at a higher level. The Flynn effect, named for its discoverer, is the observation that mean IQ scores have increased considerably in the past 50 years. This may be at least partly due to training effects caused by widespread exposure to intelligence test items in various media. Interindividual differences in intelligence test scores between persons can only be interpreted as differences in intelligence if similar amounts of practice can be presupposed. Attempts to construct so-called culture-free or culture-fair tests, which were supposed to be unaffected by prior experience, have failed because it turns out that different cultures are not prepared in the same way even for dealing with nonverbal material and mental operations that are not part of institutional schooling. Although some studies reveal ethnic differences in mean IQ, hitherto there is no convincing evidence that these differences are genetically affected.

Within a fairly homogeneous cultural context intelligence can be considered as a personality trait mainly for two reasons. First, performance on intelligence tests that are based on different contents and mental operations show significant correlations. For

instance, correlations between tests on inductive reasoning with pictorial material and with verbal material are around 0.50, and correlations between tests that measure basic visuospatial competencies such as mental rotation and tests of verbal fluency are about 0.30. In numerous studies run all over the world, multivariate statistical methods such as factor analysis have revealed that a single factor, called factor *g*, can account for 30–40 percent of the variance in IQ-test batteries composed of various scales covering a range of content knowledge and mental operations. Second, longitudinal studies have revealed that IQ is a fairly stable measure across the life span. Long-term stability of IQ for adolescents and adults comes close to the reliability of the tests, and even in early childhood long-term correlations are around 0.50 for both verbal and nonverbal tests.

Overall, intelligence test scores predict academic performance fairly well: the correlations between IQ and grades in school and university are about 0.50. The correlations between intelligence test scores and measures of outside-school success such as income or professional status are lower but still significant. That intelligence cannot account for a larger amount of achievement variation is not at all surprising, given the importance of numerous other factors, among them social background, motivation, and effort.

Because verbal and nonverbal intelligence tests are good predictors of how well an individual will succeed in school and university, they are quite helpful for making recommendations for different educational trajectories. They allow educators to identify children who cannot be expected to gain from normal schooling and therefore need special education, or children who might benefit from an advanced learning environment. According to Guthke (1993), the validity of intelligence tests can be increased if they are presented as learning tests, which means that all test-takers practice the items in several trials by getting feedback and thinking hints so that individual differences in familiarity with tests are compensated for. Particularly for children from disadvantageous social environments, learning tests are a more reliable and valid measure of intelligence than conventional IQ tests.

Attempts to extend the construct of intelligence by including social and emotional competencies as well as striving for success are controversial because the theoretical background of these concepts is still vague, and, more importantly, because the tests designed to measure these aspects do not meet the strong diagnostic criteria that have been established in psychometric research.

1.2 Factor g, Specific Abilities, and Knowledge Acquisition as Determinants of Cognitive Development

What cognitive capabilities are behind factor *g* is still a controversial question in psychology (Neisser et al. 1996). According to low-level theories, *g* reflects individual differences in basic brain functions, for example, neuronal speed. These theories are supported by results from studies that reveal 10–25 percent common variance between intelligence test scores and simple reaction time measures. However, the search for a single brain function that determines differences in cognitive capabilities has not, as yet, met with much success.

Anyhow, the heterogeneous pattern of correlations between the range of intelligence subtests leaves no doubt as to the fact that human abilities vary beyond general basic brain functions. Though significant, many correlations between intelligence subtests are low, indicating the involvement of independent mental resources. Even between tests on inductive reasoning which are based on different forms of representation (i.e., pictures, numbers, words), only medium correlations are revealed, suggesting that cognitive processes are to a large extent guided by specific verbal, visual-spatial, or numerical abilities, among others.

Among the several attempts that have been made to integrate general and specific abilities into broader theories of intelligence is the hierarchical model developed by Cattell (1971). The author proposes two major subfactors, which are related to one another. Fluid intelligence (Gf), which describes basic information processing and reasoning is measured by content-poor nonverbal tests. Crystallized intelligence (Gc), which is particularly revealed in content-rich verbal tests, represents the accumulation of higher order knowledge over the life span of an individual. Both kinds of intelligence are closely interrelated in childhood because Gf partly guides the development of Gc. From the age of about 25 years on, however, Gf and Gc take different developmental trajectories. While Gf starts to decrease at the age of 25, first slightly and from about 50 years on more quickly, Gc is unaffected by age and even has good chances to increase until the age of around 70. It is worth noticing that because of the large individual differences, the given age information is only a rough estimation.

Developmental changes also occur with respect to the structure of intelligence. The prominence of Gf decreases during childhood while Gc as a function of repeated enlargement and rearrangement of knowledge becomes more useful. When people from their adolescence on, at the latest, start to specialize in different areas, general cognitive resources have to be completed by specific abilities. This process of differentiation is characterized by decreasing correlations between intelligence subtests from childhood over adolescence to adulthood. Only in old age is a process of dedifferentiation observed which is characterized by a growing impact of Gf (Baltes et al. 1999).

After years of controversial debate, researchers now agree that part of within-group differences in IQ is determined by genes. Twin and adoption studies conducted in North America and Europe suggest that

about 50 percent of performance variation in intelligence tests is due to genetic differences. In general, the amount of variance in intelligence test scores to be explained by genes is higher the more members in a society have access to school education, health care, and sufficient nutrition. Findings which suggest that heritability increases during the life span and is not lower for Gc than for Gf are contra-intuitive only at first glance. To understand these results one has to realize that societies which provide access to a broad variety of cognitive activities in professional as well as in private life enables adults more than children to actively select special environments which fit their genes. People who have found their niche can perfect their competencies by deliberate learning.

Genes, however, not only indirectly guide learning and knowledge acquisition via general and specific abilities, rather, there is growing evidence that during evolution the human mental architecture has been equipped with quite specific knowledge structures for instance about visual or auditory patterns, physical objects, language use, and social situations. Such knowledge structures, called input systems or modules, allow appropriate cognitive and behavioral functioning beyond learning. Moreover, the representational redescription of innate modules may provide the basis for acquiring higher order domain-specific knowledge which only emerges during cultural development (Karmiloff-Smith 1992). By integrating the differential perspective of psychometric intelligence research and the universal approach of mental modularity, Anderson (1992) presents a new theory of differential cognitive development in his very informative book. According to this approach individual differences in cognitive capabilities start to increase from birth onwards because the speed and efficiency with which modules are redescribed into higher order knowledge structures is determined by general and, to a lower extent, by specific cognitive abilities. Despite its general plausibility, however, several central claims of this theory are in conflict with some well-established and appreciated theories of cognitive development and psychometric intelligence.

2. *Prior Knowledge: The Best Predictor of Learning Outcome but not a Substitute for Intelligence*

In three lines of research domain-specific knowledge has become an important variable that has challenged the value of intelligence tests as a diagnostic instrument.

2.1 *Research on Outstanding Professional Expertise*

Nobody would challenge that most of the people who have contributed in an outstanding way to an academic domain score in the upper third on intelligence tests. However, one should note that this claim is almost all that intelligence research can contribute to explaining expertise. For outstanding expertise in various fields, among them chess, medicine, and physics, having access to an elaborate domain-specific knowledge base was proven to be necessary and sufficient. For detailed treatment of this topic and for further references see *Expertise, Acquisition of*.

2.2 *Comparisons between Intelligent Novices and Less Intelligent Experts*

When entering learning settings, learners often differ from one another with respect to prior domain-specific knowledge and IQ. Studies that have investigated the learning effects of attending mathematics lessons at school and reading texts about sports have revealed much better outcomes for persons with high prior knowledge levels (experts) and low IQ than for persons with little prior knowledge (novices) and high IQ (Weinert and Helmke 1998). Note that it was novices, rather than laypersons, who were considered. In contrast to laypersons, novices possess the necessary domain-specific knowledge in terms of rules and core concepts but differ from experts in their lack of practice.

2.3 *Domain-specific Determinants of Cognitive Development in Childhood*

Research on cognitive development suggests that variations in prior domain-specific knowledge can often better account for achievement differences between younger and older children than general cognitive capabilities. For detailed treatment and for further references see *Cognitive Development: Child Education*. Moreover, longitudinal studies suggest that within-age-level achievement differences in core elementary school subjects are to a remarkable extent determined by domain-specific prior knowledge obtained in the preschool years. Early numerical competencies could account for achievement variations in mathematics after partialing out general intelligence, and early indicators of letter identification and phonological awareness predicted later performance in reading and writing (Weinert and Schneider 1999).

The reported results are important because they show that at least in complex knowledge domains a high IQ cannot compensate for a lack of prior knowledge, and moreover, that there is no direct connection between intelligence and achievement in content domains based on rich specific knowledge. However, regression analyses based on longitudinal studies reveal that the confounded variance of prior knowledge and intelligence predicts differences in learning outcome better than each single variable. This means that prior knowledge and intelligence are to a certain extent inextricably linked with each other.

Intelligence may guide the selection of learning environments and thereby determine the acquisition of prior knowledge. A person with a low IQ will hardly follow courses on theoretical physics even if he or she is credited with extra time. Moreover, intelligence may influence in how many content areas a person is able to acquire a profound amount of prior knowledge.

3. *Intelligence and Learning*

Since Sternberg (1985) claimed in his pioneering book that research on intelligence has to focus on different aspects of information processing, as there are meta-components, performance components, and knowledge-acquisition components, several attempts have been made to find out what kinds of cognitive activities are affected by intelligence during learning and problem solving. Creative assessment paradigms and stimulating ideas will be discussed in the following section.

3.1 *Efficiency in Constructing and Accessing Knowledge: Few Knowns and Many Unknowns*

At first glance it sounds plausible that, given that individual differences in prior knowledge can be controlled, intelligence affects the speed of mastering complex systems as well as competencies in dealing with novelty. Intelligence may affect the efficiency with which a knowledge base is constructed and moreover it may determine the means of accessing it. For instance, from an early age intelligent persons may represent certain problem-solving strategies in a more abstract way so that they can be easily transferred to novel situations. In addition, efficient metacognitive strategies for accessing and combining knowledge may increase the lead of the intelligent persons. Such general claims have been somewhat supported by research on complex problem solving (Frensch and Funke 1995). In this paradigm learners are presented with complex computer scenarios that, for example, require them to maximize the profit of a company, increase the supply of water in the desert but at the same time avoid an ecological collapse, or organize a city as a mayor. Tests and interviews make sure that prior knowledge about the content of the scenarios as well as experience with the computer are the same for all learners. To reach a certain goal, learners have to manipulate various variables that show a complex pattern of interaction with each other. Correlations of 0.30–0.50 between IQ and speed of mastering a complex scenario have been obtained in various studies. Moreover, contrary to a long-held scientific opinion, the impact of intelligence on learning outcome does not decrease with practice.

Although such results are somewhat promising, they often encourage the establishment of rash and arbitrary definitions of intelligence, such as speed of information processing, working memory capacity, ability to abstract, ability to deal with novelty, or metacognitive competencies. It is, however, often difficult to distinguish between such concepts, and moreover, they often lead to inconsistent conclusions. For instance, those learners who are thought to acquire information quickly may do so at the expense of profoundness. Given the controversies in cognitive psychology about modeling general working memory capacity, this construct is hardly ready for describing individual differences. Also the plausible claim that intelligent persons represent their knowledge in an abstract way that facilitates transfer does not hold (Detterman and Sternberg 1993). On the one hand, as mentioned above, research on expertise indicated that an elaborate knowledge base is not characterized by abstraction but rather by the efficient organization of specific elements. On the other hand, one of the most consistent results of psychology is that even persons with an IQ above average only rarely show spontaneous knowledge transfer even when they are presented with problems with an isomorphic formal structure. Moreover, because novelty is a combination of hitherto unrelated knowledge, the construction of completely new demands has turned out to be unattainable. The enthusiasm about metacognition, which was appreciated as a synonym for IQ by some researchers in the 1980s, has died down considerably after it has turned out that metacognitive competencies can rarely account for achievement differences in the normal and upper performance range. In sum, despite the many attempts to identify components of information processing as sources of individual differences, hitherto more questions have arisen than have been resolved.

3.2 *Educational Implications of Differences in IQ*

Comparisons between schooled and unschooled groups reveal a strong effect of education on intelligence test scores even on nonverbal tests. Only by systematic education can individuals' intelligence emerge and approach an optimum. However, given that a basic level of education has been encountered, schooling in general and special training programs in particular increase intelligence only very modestly, if at all (Perkins and Grotzer 1997). IQ differences remain quite stable over time in groups who have been attending stimulating learning settings. Education highlights individual differences in intelligence rather than compensating for them.

Broad variance in intelligence is a challenge for designing educational environments. The question arises of whether learners of different intelligence levels gain more if they are assigned to different learning environments. A few studies have revealed so-called Aptitude-Treatment Interactions. Less intelligent students gained more from structured than from open instruction, while for more intelligent learners the reverse pattern occurred. However, most studies have

failed to reveal interactions between intelligence and educational treatment, some of them probably because they lacked the statistical power necessary for revealing interaction effects. In general, when assigning learners to different learning treatments according to their intelligence, one must remember that IQ follows the normal distribution. This means that 68 percent of the population do not differ by more than one standard deviation in either direction from the mean—they are quite similar. Therefore, assigning an unbiased group of learners to two different learning environments by median split of IQ scores is not useful at all. Moreover, research has shown that similar educational environments can equip learners with a broad range of IQs with necessary prior knowledge in various areas. Eventual deficiencies in intelligence can be compensated for by effort and deliberate practice to a considerable extent. However, given the relationship between intelligence and efficiency in learning and information processing, a higher IQ facilitates the exploitation of learning environments, leading to the acquisition of knowledge that is broad as well as deep enough to master many of the as yet unknown demands of the future.

See also: Creativity and Cognition; Educational Learning Theory; Expertise, Acquisition of; Intellectual Functioning, Assessment of; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Knowledge Representation; Learning Theories and Educational Paradigms; Science and Technology Studies: Experts and Expertise

Bibliography

Anderson M 1992 *Intelligence and Development: A Cognitive Theory*. Blackwell, Oxford, UK

Baltes P B, Staudinger U, Lindenberger U 1999 Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology* **50**: 471–507

Cattell R B 1971 *Abilities: Their Structure, Growth, and Action*. Houghton Mifflin, Boston

Dettermann D K, Sternberg R J (eds.) 1993 *Transfer on Trial: Intelligence, Cognition and Instruction*. Ablex Publishing Corporation, Norwood, NJ

Frensch P A, Funke J (eds.) 1995 *Complex Problem Solving: The European Perspective*. Erlbaum, Hillsdale, NJ

Guthke J 1993 Development in learning potential assessment. In: Hamers J H M, Sijtsma K, Ruijssenaars A J M (eds.) *Learning Potential Assessment*. Swets & Zeitlinger, Amsterdam, pp. 43–67

Karmiloff-Smith A 1992 *Beyond Modularity: A Developmental Perspective on Cognitive Science*. MIT/Bradford Books, Cambridge, MA

Neisser U et al. 1996 Intelligence: Knowns and unknowns. *American Psychologist* **51**(2): 77–101

Perkins D N, Grotzer T A 1997 Teaching intelligence. *American Psychologist* **52(10)**: 1125–33

Sternberg R J 1985 *Beyond I.Q.: A Triarchic Theory of Human Intelligence*. Cambridge University Press, New York

Weinert F E, Helmke A 1998 The neglected role of individual differences in theoretical models of cognitive development. *Learning and Instruction* **8(4)**: 309–23 (Special Issue)

Weinert F E, Schneider W (eds.) 1999 *Individual Development from 3 to 12: Findings from the Munich Longitudinal Study*. Cambridge University Press, Cambridge, UK

E. Stern

Intelligent Tutoring Systems

The main aim of intelligent computer-aided instruction (ICAI) systems or intelligent tutoring systems (ITS) is to provide sophisticated instructional advice on a one-on-one basis that is comparable to that of a good human tutor. Research on ITS also serves a second aim: to develop and test models about the cognitive processes involved in instruction. Although these two aims are closely related, they may lead to different emphases concerning the priorities dealt with in relevant research (see below).

The first ITS were built in the 1970s (for an overview of these early systems see Wenger 1987). They were a response to the rather inflexible conventional computer-aided instruction (CAI) systems which were (and are) mostly built on the principles of reinforcement learning. Whereas the reactions of CAI systems follow predetermined paths, ITS aim to provide instructional interventions that are custom tailored to the strengths, weaknesses, and the actual knowledge of a given learner. This flexibility of ITS which in theory also extends to unexpected situations and which is the justification for calling them 'intelligent' is achieved by using techniques from artificial intelligence (AI) research (see *Artificial Intelligence in Cognitive Science*). The most salient property of ITS which takes them apart from other AI applications such as expert systems (see *Expert Systems in Cognitive Science*) is the diagnosis of the current state of the student knowledge. This diagnosing facility and the other parts of ITS are described next. Then follows a description of how ITS work. After that, several problems are pointed out that have led to discrepancies between theory and practice of ITS. Finally, several new developments are mentioned.

1. System Architecture

In the classical view ITS consist of three modules that contain domain knowledge, a model of the learner's current state, and teaching knowledge. These modules have been given slightly different names by different authors. Here, using Dede's (1986) terms they will be called *knowledge base, student model*, and *pedagogical module*, respectively. The importance of the com-

munication between learner and system has been increasingly recognized and this *user interface* (see *Human–Computer Interface*) is now widely recognized as the fourth module of ITS.

1.1 Knowledge Base

The knowledge base contains the domain knowledge, both declarative (knowing what) and procedural (knowing how). For instance, in ITS that deal with geometry proofs, the declarative part may contain the fact that the three angles of a triangle add up to 180 degrees; and the procedural part may contain the 'side-angle-side' rule, that is, the rule that if two triangles coincide in two corresponding sides and the angle in between one may conclude that the triangles are congruent. To represent the domain knowledge many different kinds of representational schemes have been used such as semantic networks, frames, constraint satisfaction networks, logic, propositional representations, and variations or combinations thereof (see *Propositional Representations in Psychology*; *Schemas, Frames, and Scripts in Cognitive Psychology*). However, the most common technique used to represent domain knowledge seems to be production systems (see *Cognitive Theory: ACT*; *Production Systems in Cognitive Psychology*) which consist of a memory buffer (working memory), a set of production (if–then) rules (procedural memory), a collection of facts (declarative memory) and an interpreter that decides what to do next. Production system architectures are often used in expert systems and indeed, expert systems have served as the knowledge base in early ITS. The best known example is GUIDON which used the expert system MYCIN as its knowledge base (Clancey 1987). In more recent ITS, production systems form the basis of an ideal student, who knows how to perform the task the student is expected to perform (e.g., Anderson et al. 1995).

1.2 Student Model

The student model should represent the student's current state of knowledge. That includes a representation of the knowledge, concepts, and skills the student has acquired either fully or partially. It also includes the representation of a given student's special skills and needs. Moreover, there should be a mechanism to represent misconceptions, bugs, or erroneous information which the student might have acquired. For instance, in ITS that deal with geometry proofs, the student model might represent that the student has understood the side-angle-side rule, that the student (erroneously) thinks there is also an angle-angle-angle rule which allows one to conclude that two triangles are congruent, and it may have inferred that this given student prefers delayed over immediate feedback. All this information should be used to provide optimal teaching interventions. To come up with such a student model one needs assumptions about how people learn and how they represent in memory what they have learned. Specific contents of the student model are inferred by asking the students questions and analyzing their reactions toward prompts of the system or the way they are going to tackle a problem posed by the ITS. For the student model, basically the same AI techniques are used as for the knowledge base.

1.3 Pedagogical Module

Teaching can be considered a knowledge-based skill, guided by strategies and techniques which are selected and combined dynamically in reaction to the student's actions. The pedagogical module incorporates just this skill. It should determine the presentation method, the balance of tutor and student control, and it should give feedback when necessary. One important task of the pedagogical module is how to deal with student errors, for instance, when to interrupt and what to say. Pedagogical strategies may include presenting increasingly complex concepts or problems, Socratic tutoring, simulating phenomena, and modeling of expert problem solving via coaching. The teaching strategies and techniques used should have proved successful for the particular content matter. The pedagogical module should also be able to answer student questions.

1.4 User Interface

The user interface is essential for the success of an ITS: even if the three other parts were perfect, a weak user interface could make the ITS nearly useless. The best solution would be for the student to communicate with the ITS in natural language. This, unfortunately, is not possible to date, despite huge advances in the AI field of natural language comprehension and generation. Solutions to this problem include 'canned language,' multiple choice selection, and the almost exclusive use of graphical interfaces. The user interface must take into account attentional and memory constraints of the student but it should also use the strengths of human information processing. In many earlier ITS, the user interfaces have been constructed more or less ad hoc, but recently, theoretical considerations about transfer and domain specific issues have increasingly influenced the construction of interfaces (e.g., Anderson et al. 1995, Sedlmeier 1997).

2. How do ITS Work?

The four components of ITS described above capture the different aspects included in ITS and give a good

idea about how ITS might work in principle. Although parts of real ITS can be mostly assigned to one of these four parts there are not usually four independent modules at the level of implementation that interact with each other on equal terms. Usually one of the three classical components can be regarded as playing the central part in the teaching process.

2.1 Expert-centered Instruction

In most ITS, the knowledge base plays the central part in the teaching process. Usually, the knowledge base is conceptualized as a domain expert or an ideal student and the course of action taken is dependent on the deviations of the student model from the expert or ideal student. In an *overlay model* the knowledge of the learner is seen as a subset of the expert's knowledge (e.g., Clancey 1987) and if a deviation between the current student model and the expert knowledge is diagnosed, the system tries to remove that discrepancy by suitable instructional measures. It has been found that the student's knowledge is not just a subset of the expert's knowledge but may contain nonexpert ways to arrive at the solution of a problem. Extensive research on how students solve problems correctly and what kinds of errors they commit led to the model of an ideal student and to 'bug libraries,' that is, to possible explanations for errors (Brown and Burton 1978). This approach has been elaborated into the technique of *model tracing* (e.g., Anderson et al. 1990). In this approach, a performance model specifies how, for a particular problem, a student's current knowledge will map onto performance. Instructions address diagnosed confusions and keep the student on the right solution path. A learning model specifies how the student's current knowledge will change as a result of problem-solving experiences and this learning model can be used to trace the student's knowledge over time.

2.2 Teacher-centered Instruction

One might argue that the most natural choice of an ITS component to control the instructional process should be the pedagogical module. Unfortunately, the pedagogical module is the most neglected component of ITS. This might in part be due to still existing deficiencies in instructional theory but it might also be the result of a practical problem: Teachers or education scholars are seldom involved in the construction of ITS. One approach that puts the teacher into the center of the instructional process is to use planning techniques that make it possible to develop and flexibly adjust global teaching strategies (e.g., Peachey and McCalla 1986, Sedlmeier and Robles de Acuña-Ponseti 1992).

2.3 Student-centered Instruction

In expert and teacher-centered instruction, the learning process is fully dominated by the tutor. This contrasts with the current emphasis in instructional theory on *learning by doing* and *situated learning* (see *Situated Cognition: Contemporary Developments* and *Situated Cognition: Origins*). One way to put the student into the center of the instructional process would be to use techniques of *machine learning* (Self 1985). Research on this kind of instructional process is still in its infancy.

3. The Practice of ITS

Whereas ITS seem to be ready to be used for specialized applications on the commercial market (Norton et al. 1998), to date, by far the most ITS did not pass the status of prototypical systems and there are very few ITS employed in the classroom. One important reason for this state of affairs is probably the diversity of researchers involved in the construction of ITS. A second might be that there is still no convincing evidence of ITS's superiority over more conventional teaching procedures, and finally, there are severe problems with constructing good student models.

3.1 Diversity of Interests

Constructing ITS is a multidisciplinary endeavor, including computer scientists, mathematicians, cognitive psychologists, and educational researchers, to name the most important groups. With some notable exceptions (e.g., Anderson et al. 1995), researchers are usually most interested in those aspects of ITS that are closest to their own field of research: Educators are most interested in effective tutors but often do not know how to write computer software, and researchers in cognitive science are more interested in developing and testing theories about how humans learn than in the sometimes tedious process of implementing computer software in schools. Another problem was that many ITS were developed on workstations and Lisp machines that schools could not afford and maintain.

3.2 How Effective are ITS?

There is still a lack of thorough evaluation studies in which the effectiveness of ITS is compared to that of conventional teaching techniques. Keeping in mind that the development costs for ITS are quite high, ITS have to achieve noticeably better learning results than conventional CAI systems. One motivation to use ITS in the classroom has been the assumption that individualized instruction is much more effective than group instruction and that ITS can provide such an individualized instruction. However, a collection of

meta-analyses shows that overall, the benefits of individualized instruction are almost indistinguishable from those of group instruction (Fraser et al. 1987). This is in line with the conclusion which Legree and Gillis (1991) reached in a survey of evaluation studies: For the majority of the ITS analyzed, there was no noticeable advantage over the learning success achieved with group instruction.

3.3 Problems with the Student Model

Research on student models since the early 1970s has shown that even for relatively small and well-defined domains such as simple arithmetic it is not possible to construct a complete cognitive model (e.g., Elsom-Cook 1993). One reason is that the representation of everyday knowledge which is needed in addition to domain knowledge is an almost insurmountable task. Another reason is that the basis for diagnosis, the student's behavior, is often not clearly interpretable. A given student error encountered by the system may have many potential causes. However, the underlying misconceptions and not the observable behavior should be the basis for corrective procedures. Even if it were possible to construct an optimal student model, could it be used effectively? All available evidence suggests that it could not: Good teachers often have incomplete and even wrong 'student models' but are nonetheless very successful. They achieve their success not by a thorough diagnosis but by moving through a curriculum script, that is, a loosely ordered but well-organized and sequenced set of skills and concepts they expect students to learn, and by using adequate activities and strategies for teaching this material (Putnam 1987). Thus it seems that, at least for the purpose of effective instruction, the student model in its present state has severe deficiencies and might be discarded without much loss as far as teaching efficiency is concerned (Gugerty 1997).

4. Perspectives: Cognitive Science Research vs. Educational Practice

Research in ITS is still flourishing (e.g., Goettl et al. 1998) but the emphasis is more on the second aim stated in the beginning of this article, that is, to develop cognitive models about the components of the instructional process. This could eventually also lead to the achievement of the first aim, that is, to provide effective instructional devices for all kinds of domains. It seems that for highly specialized industrial applications, this first aim has already been reached, at least in part. The deployment of ITS in the classroom is, however, only in its starting phase. For practical purposes it might be worthwhile in the short run to de-emphasize the reliance on the problematic student model and put more effort into the construction of theory-guided flexible interfaces. Such flexible interfaces could be the front-end to existing computer tools, simulations, games, or microworlds (Cumming and Self 1990) or could take part in guiding the teaching process (Sedlmeier 2000). Technological advances and the possibility of using the Internet will greatly facilitate the deployment of ITS in schools but ultimately whether or not schools will profit from the potentially huge benefits of ITS depends on successful collaboration among educators and system builders.

See also: Computer-assisted Instruction; Digital Computer: Impact on the Social Sciences; Human–Computer Interaction; Human–Computer Interface; Instructional Technology: Cognitive Science Perspectives

Bibliography

Anderson J R, Boyle F, Corbett A T, Lewis M W 1990 Cognitive modeling and intelligent tutoring. *Artificial Intelligence* **42**: 7–49

Anderson J R, Corbett A T, Koedinger K R, Pelletier R 1995 Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences* **4**: 167–207

Brown J, Burton R R 1978 Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science* **2**: 155–92

Clancey W J 1987 *Knowledge-Based tutoring: The GUIDON Program*. MIT Press, Cambridge, MA

Cumming F, Self J 1990 Intelligent educational systems: Identifying and decoupling the conversational levels. *Instructional Science* **19**: 11–27

Dede C 1986 A review and synthesis of recent research in intelligent computer-assisted instruction. *International Journal of Man–Machine Studies* **24**: 329–53

Elsom-Cook M 1993 Student modelling in intelligent tutoring systems. *Artificial Intelligence Review* **7**: 227–40

Fraser B J, Walberg H J, Welch W W, Hattie J A 1987 Syntheses of educational productivity research. *International Journal of Educational Research* **11**: 145–252

Goettl B P, Halff H M, Redfield C L, Shute V J (eds.) 1998 *Intelligent Tutoring Systems*. Springer, Berlin

Gugerty L 1997 Non-diagnostic intelligent tutoring systems: Teaching without student models. *Instructional Science* **25**: 409–32

Legree P J, Gillis P D 1991 Product effectiveness evaluation criteria for intelligent tutoring system. *Journal of Computer-Based Instruction* **18**: 57–62

Norton J E, Jones J A, Johnson W B, Wiederholt B J 1998 Are intelligent tutoring systems ready for the commercial market? In: Bloom C P, Loftin R B (eds.) *Facilitating the Development and Use of Interactive Learning Environments*. Erlbaum, Mahwah, NJ, 259–88

Peachey D R, McCalla G I 1986 Using planning techniques in intelligent tutoring systems. *International Journal of Man–Machine Studies* **24**: 77–98

Putnam R T 1987 Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal* **24**: 13–48

Sedlmeier P 1997 Basic Bayes: A tutor system for simple Bayesian inference. *Behavior Research Methods, Instruments and Computers* **29**: 328–36

Sedlmeier P 2000 How to improve statistical thinking: Choose the task representation wisely and learn by doing. *Instructional Science* **28**: 227–62
Sedlmeier P, Robles de Acuña-Ponseti J 1992 'Intelligente' Hilfe beim Lösen von alltagsnahen Wahrscheinlichkeitsproblemen: Modellierung dynamischer Wissensinhalte für ein flexibles Tutorsystem ['Intelligent' help in solution of probability problems with everyday applications: Modeling of dynamic knowledge for a flexible tutor system]. *Kognitionswissenschaft* **3**: 24–37
Self J 1985 A perspective on intelligent computer-assisted learning. *Journal of Computer Assisted Learning* **1**: 159–66
Wenger E 1987 *Artificial Intelligence and Tutoring Systems*. Morgan Kaufmann, Los Altos, CA

P. Sedlmeier

Intensification and Specialization, Archaeology of

Many of the major transitions studied by archaeologists—the development of agriculture, urbanism, institutionalized inequality, and markets; processes of imperial incorporation, colonialism, and political collapse—either consist of, or are accompanied by, significant changes in the organization of production. As such, these changes, their causes, and consequences have been pivotal research issues. Studies of intensification and specialization build on archaeological approaches to production, and must be understood in the light of this broad construct, which differs significantly from its use in economics.

1. *The Archaeology of Production*

Archaeological concepts of production are typically situated in the anthropological perspective that productive activities are embedded in larger cultural and social systems. Production—the creating, constructing, and tending of artifacts, structures, human-modified plants and animals, even entire landscapes—is a fundamental focus of the study of the archaeological record. Production, as one dimension of human economy is both responsive to and constitutive of other domains such as processing, distribution, and consumption. This perspective is a legacy of the formalist–substantivist debate in anthropology, which centered around whether nonindustrial economies could be understood in the same terms as industrial capitalism (the formalist position), or whether the economies of nonindustrial societies typically studied by anthropologists were, in Polanyi's (1957) words, more fundamentally 'embedded and enmeshed' in social institutions. However, this picture is complicated by important strands in archaeological thinking which draw on economistic approaches or on those approaches as modified by ecology and evolution.

Research has not focused evenly on all arenas of production. Although agriculture and animal husbandry constitute important areas of research, there is a longer history of emphasis on durable goods or artifacts, objects made or modified by humans. Artifact manufacture is often termed craft production. While changes in the organization of craft production are generally conceived in terms of specialization, food production is more often viewed in terms of intensification. Part of this divergence may relate to the different kinds of information available about each arena of production. Students of craft production typically work with artifacts themselves and with evidence from production locales such as kilns, workshops, even dumps. Food production research does involve products such as seeds and animal bones, but these are often conceived as objects of consumption, with production studies focused primarily on locales of production such as fields.

2. *Intensification*

Archaeological studies of intensification have focused primarily on agriculture. In some sense, the historical trajectory of agricultural change can be viewed as a long, although variable, history of productive intensification, concomitant with the rise in global population, its clustered distribution on the planet, and the entrenchment of economic and social inequality.

2.1 *The Concept of Intensification*

The difference between intensification and simple increase is analogous to the difference between concentration and amount. Intensification of production refers to attempts to increase the concentration of production, or productive output per unit of land, labor, or some other fixed quantity. The focus on production effort in this definition highlights the salience of input, allowing failed efforts to increase productivity to be included in studies of intensification. In archaeology, the variable held constant almost always refers to land in reference to agriculture (getting more from a given area), and labor in reference to craft production (increasing efficiency of production). Situations in which both land and labor are held constant while capital inputs are increased, as in industrialized agriculture, are more rarely discussed.

2.2 *History of Approaches to Intensification*

The model of intensification set forth by economist Ester Boserup (1965) has been the most influential

formulation of the problem, and was widely adopted in archaeology (e.g., Cohen 1977). Boserup's model was parsimonious, general, and comprehensive; however, like other unilineal models of cultural evolution, it greatly generalizes and simplifies this complex process. Boserup's view of population pressure as an independent variable driving intensification inverted the earlier Malthusian formulation. While Malthus saw arable land as limiting increases in production, Boserup optimistically asserted that constantly increasing population (an assumption shared with Malthus) acted as a motor, driving technological changes in land use along an extensive–intensive continuum.

The most contested aspect of Boserup's model was the causal efficacy of population pressure, a view resting on several related assumptions. First, producers are assumed to exert the minimum effort required to meet their needs (the law of least effort). Thus, the most labor-extensive regime possible will always be employed. Second, Boserup argued that there are diminishing marginal returns to labor (declining efficiency) with increasingly intensive forms of production. The disadvantages of intensive agriculture—increased labor inputs and declining efficiency of that labor—ensure, for Boserup, that more intensive forms of production will be adopted only when strictly necessary. These assumptions have been criticized on empirical grounds, as has the prime-mover status of population pressure. There is now a vast literature considering such causal factors as population density and distribution; sedentism; market-driven, prestige-driven, or politically-driven demands for produce; risk, climate change; and many others (reviewed by Morrison 1994). Contemporary approaches to intensification tend to stress multiple and cascading causes which may be locally contextual and historically contingent (e.g., Kirch 1994).

2.3 *Consequences of Intensification*

In a proximate sense, the most common consequence of intensification is increased productivity. More significant, however, are less-proximate consequences, such as the restructuring of human labor, possibilities for differentiation of nonfood production, and the support of nonproducers. It is in this sense that intensification has been seen as underwriting social and political complexity, by which archaeologists generally mean the existence of institutionalized inequality and/or the presence of complex polities such as states or chiefdoms. Historically, intensified food production has also had dramatic ecological consequences, leading to landscape modification on a large scale, the worldwide movement of plant and animal taxa, and the existence of artifactual landscapes such as cities.

2.4 *Methodological Issues in Intensification Studies*

The study of intensification poses serious methodological challenges. The investigation of past agriculture, in particular, is difficult because of its large spatial scale and ephemeral material remains. For this reason, archaeologists consider numerous indicators of land use, including agricultural implements, historical documents, settlement distributions, agricultural facilities, and botanical remains. Identifying intensification in the archaeological record depends on the definition adopted; views which stress output require some measure of productive success, a difficult requirement for archaeology, where agricultural productivity can rarely be quantified. Output-based models are more popular in craft production studies, where such measures may be more reasonably inferred. Production effort is usually postulated on the basis of inferences about the overall form of production, using clues such as technological forms, products, or cultigens with special productive requirements, as well as knowledge of specific environmental conditions.

2.5 *The Future of Intensification Studies*

Three promising trends in the archaeology of intensification may be identified. The first is increased attention to the process of intensification itself, a process incorporating a great degree of variability. Kaiser and Voytek (1983) for example divide intensification into three components: specialization, diversification, and intensification proper. These involve changes in the amount and organization of labor and its application through technology, and require measurement of multiple variables. The second trend is increased attention to actual historical trajectories of change and the variable paths intensification may follow. Finally, in accord with moves toward a more inclusive view of production, many studies now weigh cultural meanings of food products alongside concerns of labor efficiency or energetics.

3. *Specialization*

The origins and development of specialized production have always been major concerns in archaeology. If intensified food production is often seen as undergirding social complexity, then specialization has been regarded as its index. Students of specialization draw on Marxian traditions which integrate concern for the means and social relations of production with anthropological attention to culture. This research tradition, however, draws equally from Adam Smith in its attention to the division of labor and focus on exchange, as well as its often embedded assumptions about the nature of economic behavior. In their concern for social and political implications of

specialization, archaeologists have also drawn on Durkheimian notions of solidarity and integration.

3.1 The Concept of Specialization

Specialization may be defined as the channelling of resources and/or labor into restricted ends, a definition focusing on the products and process of specialization. Other definitions stress the role of specialization in setting apart people, technologies, and production locales, both physically and socioeconomically. Most literature on specialization in archaeology focuses on craft production; such studies have been dominated by a concern for technological process, interest in the organization of production, and, more recently, an expanded interest in the social relations of production, including issues such as identity and meaning. Specialization implies exchange on some level and as such, is not intelligible outside larger political/economic contexts.

3.2 History of Approaches to Specialization

Most early studies of artifacts emphasized typology, using morphology and manufacturing technology to classify objects, creating the chronological and spatial systematics that still underlie disciplinary practice. Despite pioneering studies such those of V. Gordon Childe that stressed the social and political roles of craft specialists in early stratified societies, most observations on artifact variability were oriented toward defining the cultural/temporal units which were themselves the goals of research. With moves toward functionalism in the 1950s and 1960s however, technologies of production commanded greater attention as areas of study in their own right; this focus on technology continues into recent materials science approaches.

With the 1960s and 1970s came concerted efforts to delineate the organization of production and especially the integration of specialized producers into larger political economies. Growing from this interest, much effort has been spent investigating the scale (household production, workshops, factories) and tempo (full- vs. part-time specialists) of specialized production and the material correlates of organizational forms. Also of concern were forms of control over production, strategies of labor and resource mobilization, and the disposition of manufactured goods. Several classifications have attempted to capture some of these contextual dimensions of production, distribution, and consumption (Costin 1991). Studies of exchange have been particularly important, congruent with a focus on regional economies and with the development of new analytical methods for determining source areas of raw materials.

Specialization studies have been marked by a concern for the initial causes of specialization, although here explanations often stress socially-based demands, perhaps because many artifacts are clearly not biologically necessary for survival. More recently, archaeologists have begun to stress processes of production, especially agent-based approaches which focus on structured chains of culturally-inflected decisions made by producers (Lemmonier 1986) and on interrelations between technology and meaning. A new focus on producers, too, has followed from concern for more finely-divided aspects of producer identity, including gender and ethnicity and from recognition that multiple organizational forms of production often co-exist, even where the producers are the same.

3.3 Consequences of Specialization

From the time of Childe, archaeologists have explicitly linked specialized production with complex political forms and social inequality (Brumfiel and Earle 1987). The very existence of elites is typically inferred, in part, from sumptuary objects identified by their elaboration, restriction, and sometimes exotic provenance. Curiously, specialization in craft production (sometimes cast as intensification) is generally viewed as promoting efficiency, while intensified agriculture is depicted as producing declining marginal returns. Recent work on the meaning of production, especially its ritual context, draws more fully from an anthropological tradition de-emphasizing efficiency.

3.4 Methodological Issues in Specialization Studies

Recognizing specialized production begins with the objects and by-products of manufacturing and their distribution. The definition of specialization allows for certain ambiguities, as issues of scale and context impinge. For example, the production of artifacts in one region and exchange to another may be considered regional specialization, although artifacts were manufactured by every household. Locations such as quarries are specialized places without necessary associations with producer specialization. Although the goal of archaeological research is usually to identify producer specialization, because archaeological data are fundamentally spatial, determining the organization of production is inevitably inferential. Much of the methodological development of specialization studies grew around this dilemma. Complex technological requirements or high levels of producer skill are sometimes assumed to index specialist manufacture; another criterion advanced for identifying specialization and its scale is standardization, mass-production being thought to promote standardization (for reasons of efficiency).

3.5 The Future of Specialization Studies

The trends identified in intensification studies also hold for studies of specialization. Concerns for process are implicated in perspectives that draw specialized production into the larger field of the anthropology of technology; this concern is sometimes expressed in the language of 'crafting' and the ways in which crafting both draws from and creates social relationships (Costin and Wright 1998). Recent developments in the archaeology of both intensification and specialization tend to stress process over cause and explore the consequences of productive organization for differently situated social actors rather than classify productive forms along an evolutionary continuum.

See also: Agricultural Change Theory; Boserup, Ester (1910–99); Ceramics in Archaeology; Differentiation: Social; Indigenous Knowledge and Technology; Labor, Division of; States and Civilizations, Archaeology of; Trade and Exchange, Archaeology of

Bibliography

Boserup E 1965 *The Conditions of Agricultural Growth*. Aldine, Chicago

Brumfiel E, Earle T K (eds.) 1987 *Specialization, Exchange, and Complex Societies*. Cambridge University Press, Cambridge, UK

Cohen M N 1977 *The Food Crisis in Prehistory*. Yale, New Haven, CT

Costin C L 1991 Craft Specialization: issues in defining, documenting, and explaining the organization of production. In: Schiffer M (ed.) *Archaeological Method and Theory*, Vol. 3. University of Arizona Press, Tucson, AZ

Costin C L, Wright R P (eds.) 1998 *Craft and Social Identity. Archaeological Papers of the American Anthropological Association*, Vol. 8. American Anthropological Association, Arlington, VA

Kaiser T, Voytek B 1983 Sedentism and economic change in the Balkan Neolithic. *Journal of Anthropological Research* **2**: 323–53

Kirch P V 1994 *The Wet and the Dry: Irrigation and Agricultural Intensification in Polynesia*. University of Chicago Press, Chicago

Lemmonier P 1986 The study of material culture today: toward an anthropology of technological systems. *Journal of Anthropological Archaeology* **5**: 147–86

Malthus T R *Essay on the Principle of Population, as it Effects the Future Improvement of Society*. Johnson, London

Morrison K D 1994 The intensification of production: archaeological approaches. *Journal of Archaeological Method and Theory* **1**: 111–59

Polanyi K 1957 The economy as an instituted process. In: Polanyi K, Arensberg C M, Pearson H W (eds.) *Trade and Market in the Early Empires: Economies in History and Theory*. Free Press, Glencoe, NY

K. D. Morrison

Intentionality and Rationality: A Continental-European Perspective

1. The Text

There is an objective interrelationship between 'intentionality' and 'rationality.' As a rule, intentionality is considered to be a necessary, although not a sufficient, condition for rationality. This means that one is usually convinced that something or someone can be called rational only when the predicate 'intentional' can be applied to it or to him or her first. In addition, both concepts can be applied both in the sense of an expression of disposition and to describe specific manifest qualities. Thus, intentionality is attributed to living beings when they have the ability to produce something that can be described as intentional: actions, psychic occurrences, or linguistic utterances. Conversely, the intentional has always been understood as the actualization of the disposition 'intentionality.' Similarly, in the case of 'rationality,' that which can be seen as rational, can be traced back to the activation of rational competence which itself presupposes intentionality. Both can be applied only to people. If one considers institutions, systems, or even machines and their achievements to be rational, one uses this expression in a figurative or derived sense.

2. Intentionality

2.1 Intentionality From the Standpoint of the Theory of Action

This artificial philosophical term does not have an equivalent in colloquial language except to characterize something that someone has done as intentional, i.e., deliberate. Thus, it made sense that the theory of action use this predicate in a defining sense as well, and speak of actions only when there is sufficient reason to assume that the action or behavior took place intentionally. For the social sciences in so far as they conceive of themselves as sciences of action, but also in a court of law or in morally relevant situations, it is important to know what it is that makes occurrences in the world actions. Reflex movements or forms of instinctual behavior certainly do not belong in this category; they are objects neither of social scientific research nor of normative judgment. The specific difference of actions is usually seen in the fact that they took place deliberately, and only under this condition is responsibility taken for them. The question, then, is how this relation to intentions fits together with the causal interpretation of all world events: are intentions causes? If so, then they themselves would have to have been caused, and this is considered incompatible with our self-understanding as subjects acting freely and spontaneously.

In opposition to the neo-positivist concept of a physicalistic unified science, which also ties the social sciences to behaviorism—that is, to the description and explanation of observable behavior only—the attempt has been made since the 1950s, above all in Anglo-Saxon analytical philosophy, to place the theory of action on its own foundation. In this attempt, special meaning was attached to the concept of 'intention' (cf. Anscombe 1957). Related concepts were grouped around it, all of which were explicated more precisely with the goal of countering the physicalists' causal manner of speech with an intentional vocabulary that also enabled one to speak in scientific terms about the intentions, motives, and reasons for actions. The central differences became clear above all in the extensive debate that took place in the 1960s about causes and reasons. According to the intentionalist camp, two characteristics above all distinguish reasons, as the clearest intentions or motives for action, from causes. First, reasons are not events that are followed by other events according to general laws of nature, but rather, they are subjects' dispositions to bring about specific events under specific circumstances. Second, reasons, like intentions or motives, relate to what follows from them, and not, like cause and effect, to each other. According to Hume, one can speak of causality in the strict sense of the word only when cause and effect are two events that can be described as logically independent of one another. This is not the case here, as one cannot ascribe to someone a reason for any action without making reference to this same action. Thus, someone who did not take a specific action also cannot have had a reason for this action. (This is referred to in the literature as the 'logical connection argument': cf. Melden 1961, p. 78ff.)

The reference to the special status of intentions and their meaning for the description and explanation of events as actions also led to attempts at opposing the causal explanations of the natural and behavioral sciences with another explanation type: that of intentional or rational explanations (cf. Schwemmer 1973), which are to be determinative in the sciences of action. (When speaking of rational explanations, 'rational' refers to the reasons (*rationes*) that someone has for doing something specific.) In this context, recourse is made to Aristotle's practical syllogism, which can be considered an elementary form of practical rationality in general (cf. von Wright 1963).

The thesis 'what we call causes in the case of physical events are, in the case of actions, intentions (reasons, motives, aims)' does admittedly lead to the problem that actions are also physical events—for example, bodily movements—and in this respect, of course, also have causes. Thus, when the thing–event language and the action–reason language are used in parallel, they are simply two different descriptions of one and the same occurrence in the world, and the problem arises of how these two kinds of description can be reconciled with each other. The same holds for the compatibility of causal explanations of action with intentional explanations of action: it can be ensured only by accepting that the intentional explanation type is not a sub-type of causal explanation forms, but something else with a completely different aim. Additionally, Donald Davidson (cf. Davidson 1980, p. 7ff.) showed that it is very sensible to view not the reasons themselves, but rather the complex intentional disposition of having reasons as a causally effective factor.

2.2 Intentionality from a Psychological Standpoint

In the late scholastic period, the expression *intentio* was used as a generic term for concepts and ideas, whereby a distinction was made between *intentio prima* and *intentio secunda*. *Intentio prima* applied to things and facts, while *intentio secunda* had other *intentiones* as its object: as thoughts about thoughts. In this case, *intentio* thus has an object which itself is an *intentio* and which only exists within it. Franz Brentano characterized consciousness using this figure of the intentional inexistence of the *intentio prima* in the *intentio secunda*. According to Brentano it is always consciousness of something, and intentionality, in this sense of being directed at an inner object, was considered by him to be the specific characteristic that differentiates mental states from all others. Brentano's student Edmund Husserl adhered to his teacher's model, although with several changes, and thus also to 'intentionality' as a fundamental concept of his phenomenological philosophy. The same is true of the phenomenological schools that broke with Husserl (including Scheler, Heidegger, Sartre, and Merleau-Ponty). Furthermore, the work of Alfred Schütz is important (especially *Der sinnhafte Aufbau der sozialen Welt* 1974): he took up the ideas of Husserl and, using phenomenological means, undertook to lay the foundations of sociology as an understanding (*verstehende*) and at the same time explaining (*erklärende*) science of action in Max Weber's sense. Weber was interested in defining the difference between mere behavior and action with a 'subjective meaning,' that is, action to which agents attach their own meaning (Weber 1974, p. 542). The phenomenological concept of intentionality opened up the opportunity for Schütz to explicate more precisely what could be meant by the term 'subjective meaning.' The phenomenological tradition found its way into the discussion of the foundations of social sciences through Schütz above all, and here provided an important additional stimulus to the methodological programs of symbolic interactionism and of ethno-methodology.

2.3 Intentionality from the Standpoint of Linguistic Philosophy

In contrast to Brentano and the phenomenologists, linguistic philosophy maintains that thoughts cannot

be discussed independently of their expression in language. In this field, Rudolf Carnap in his *Logischer Aufbau der Welt* (1966) was the first to study 'intentional relationships,' that is, the logical relationships between sentences beginning with 'X thinks,' 'Y believes,' or 'Z assumes' and the clauses which, as a rule, follow the word 'that.' In designating such verbs as 'intentional,' Carnap expressly followed Brentano and Husserl, but with an important difference in interpretation: he held that an entire sentence must always follow the 'that.' The idea that intentional attitudes always relate to something that exhibits the form of an entire sentence and, only in this context, to objects has been disputed ever since. For this reason, such attitudes are also known as propositional attitudes. The grammatical forms in which we present such attitudes do admittedly have the unpleasant characteristic of not conforming to the demand of extensionality. This means that the truth value of sentences such as 'X believes that *p*' is not a function of the sentence parts contained within them, since it could be true that X believes *p*, even if *p* is false. This connection between intentionality and intensionality, which means that sentences expressing intention are extensionally opaque, has prompted many philosophers to follow the example of W. V. O. Quine in banishing the intensional context from scientific language—although admittedly at the cost of having to bring the intentional into compliance via a purely behavioral interpretation of physicalistic–extensional description and explanation.

Today, the intentional is no longer threatened by reductionism. At least since Wittgenstein's late work and Gilbert Ryle's classic, *The Concept of Mind* (1949), the philosophy of mind has striven to provide the intentional attitudes their grammatical right as well, but without falling back into the pre-lingual self-observation or introspection with which Brentano and the phenomenology of Husserl operated. Daniel Dennett (1991) in particular made important suggestions for an answer to the question of what it means to ascribe intentionality (intentional stance) to organisms or machines. His most formidable opponent for many years was John R. Searle, who does not hold intentionality to be merely something ascribed by an observer, but rather, a natural quality of specific organisms. The question of the naturalizability of consciousness (cf. Dretske 1995)—that is, the possibility to describe and explain intentionality exclusively by natural scientific means—has been debated heatedly since then.

In language theory as well, the topic of 'intentionality' has been discussed in depth, particularly in connection with the thesis of Grice (1957) which states that linguistic meaning can be understood and reconstructed exclusively from that which a speaker means and expresses in communicative actions. In simpler terms, this intention-based semantics (cf. Meggle 1994) refers back to what a speaker wants to give a hearer to understand through his or her speech act. Thus, it brings together the semantic aspect of intentionality with the aspect of intentionality represented in theory of action. In his work *Intentionality*, Searle (1983) submitted a superior version of this approach to Grice's, with the central thesis that intentionality is a basic state of mind that 'gives' the mind its linguistic expression. With this, suddenly Searle's earlier theory of the speech act appeared as a branch of the philosophy of mind. The opposing position to this intentionalism in semantics is conventionalism (e.g., Wittgenstein), according to which all that we can mean and communicate is already set out according to linguistic conventions and rules. The conventionalist asks 'How can I know what I mean before I hear what I say?' (Black 1977). The strength of the conventionalist position is that it refers to what a speaker means and can give a hearer to understand, even before the speaker has to be present in verbal or just symbolic form (cf. Schnädelbach 2000, pp. 204ff.), because 'X means that' and 'X wants Y to believe that' always needs a propositional object which is itself written in grammatical form. Not only is the relationship between intentionality and conventionality (Strawson 1971, pp. 174ff.) interesting from the point of view of language theory; it also applies to a wide variety of types of action—for example, to ritual or institutional actions—and not merely to speech acts.

3. *Rationality*

The word 'rationality' is currently used mostly in the sense of means–ends rationality (*Zweckrationalität*), which takes as its own example the economic conception of rationality. What it means is always the optimal balancing of ends and means while giving primacy to ends, which can either not be rationally justified at all or only incompletely. However, this model cannot adequately express all that is connected historically and objectively with the *rationalitas* of the *animal rationale*. The philosophical tradition, admittedly, seldom used the expression 'rationality' and always gave preference to other terms such as 'reason' or 'understanding.' Therefore, the concept of rationality has only returned to modern philosophical discourse via the detour of discussions in the social sciences and in the theory of science. It is well known from common experience with unreasonable rationalizations that a distinction can be made between the rational and the reasonable; thus, a philosophical theory of rationality should be defined broadly enough to encompass such intuitions.

3.1 *A Narrower and a Broader Concept of Rationality*

The predicate 'rational' is most often used in a normative or evaluative sense, whereas the opposite

appears as 'irrational' in the sense of being contrary to reason. Thus, we judge actions, institutional rules, and also people by means of a criterion of rationality, and discover that what is being judged is inadequate to the criterion. The fact that we are dealing here with the narrower concept of rationality becomes clear when we recall that we are also talking about the opposite of the rational even when such a judgment does not come into question at all. In this case, we use the 'irrational' to designate that which is lacking reason or is arational. This broader concept of rationality thus serves to separate from the wide variety of things and events those things which we could view as candidates for a normative or evaluative judgment according to the narrower concept of rationality. Rationality, as the opposite of unreasonableness, is probably the same thing as understandability; we use the narrower, normative concept of rationality only for behavior patterns and expressions where there is something to understand about the causal or functional explanation (cf. Schnädelbach 1992, pp. 79ff.). The relationships between this broader concept and the normative concept of rationality were the topic that was debated following the theses of Winch (1958) among philosophers and social scientists on 'rationality and relativism' (cf. Hollis and Lukes 1982, Wilson 1970). The question, then, is whether rationality as understandability simply coincides with intentionality. What is certain is that intentionality is a necessary condition for rationality qua understandability, but it is not certain whether it is a sufficient condition. The concept of the rule appears to decide this question: it is to be assumed that the understandability of behavior patterns and expressions presupposes the capacity to follow rules which the person understanding them is also in a position to follow. There is much evidence suggesting that intentionality exists below the level of following rules as well, although here the condition of understandability would not yet be fulfilled.

3.2 Types of Rationality

The fact that we always have to reckon with different types of rationality (cf. Schnädelbach 2000, pp. 256ff.) has been known to philosophy since its beginnings. Plato's differentiation between *nóesis* and *dianoia*, which was passed down into the modern age via the difference between *ratio* and *intellectus* or between reason (*Vernunft*) and understanding (*Verstand*) can be mentioned in this context, as can Aristotle's doctrine of the five dianoetic virtues, Scientific Knowledge, Art, Prudence, Intuitive Reason, and Wisdom. Even when one assumes the usual, but much too narrow, interpretation of rational competence as the capability of establishing reasons, one must make further differentiations. Because of the fact that Aristotle defined the capability of establishing reasons—which according to him constitutes the specific characteristic of scientific knowledge—in general terms as the ability to answer why-questions, we have had to clearly differentiate between causes (*causae*) and reasons (*rationes*) since the seventeenth century. Therefore, at the beginning of the twenty-first century we regard answers to questions about causes as explanations, and only answers to questions about reasons as 'the establishing of reasons' in the strict sense. We can differentiate between cognitive and normative types of explanations: while cognitive (causal, final, functional, historical, etc.) explanations tell us why something is the case, normative explanations inform us about why something should be the case. Establishing reasons, on the other hand, relates in the epistemological case to reasons that indicate why one could believe something to be true, and in the intentional or rational case, to what—within the bounds of what we can possibly know—can be said for doing or not doing something specific. The intentional or rational establishing of reasons is the locus of the social scientific model of rationality, since neither in the case of economic rationality nor in the case of Weber's means–ends rationality does it play a role whether subjective calculations correspond to reality to the best of one's knowledge or not. Thus, someone who made a mistake was not irrational but simply unlucky. This is also the locus of those intentional or rational explanations for action mentioned in Sect. 2.1 which cannot be interpreted as a special type of causal explanation because of their relation to the subjective perspective of the agents.

Often, 'establishing reasons' is used in the sense of the Latin *rationem reddere* or the Greek *lógon didónai* to mean argumentation. However, this usage often fails to recognize that argumentation, as a reflection of critique and justification, relates not directly to things, occurrences, or actions—after all, nobody criticizes or justifies the fact that the sun will rise tomorrow—but to the claims to validity that are connected with assertions about things, occurrences, or actions. The most important of these claims to validity are those of truth and rightness. The question whether or not an occurrence took place is not the theme of argumentation, but rather, the assertion that this occurrence did take place. Here, as well, the issue is not primarily whether or not an action is good, but whether the thesis that this action is good can be justified or not. In simpler terms, the rationality type 'argumentation' is to be located on the meta-level of speaking about assertions that apply to things and occurrences or actions. There is much evidence to support the idea that when Kant and Popper among others defined reason as essentially a critical faculty, they had this type of rationality in mind. However, in order for something that can be the object of argumentation—that is, of critique and justification—to come about at all requires rationality in the broad sense, that is, qua understandability. Thus, understandability is the more fundamental type of rationality.

3.3 Incompleteness of Every Theory of Rationality

The idea that it is probably impossible to explicate rationality all at once, as a whole and for all contexts is supported by the fact that every explication of rationality must itself have a claim to rationality. Therefore, it cannot discuss its own medium of explication and the implicit standards of explication contained in it as long as it makes use of them itself. Admittedly, this is possible in other contexts, but then as well something outside the theme of discussion is left behind (cf. Putnam 1981, Chap. 5). From this emerges an argument against the radical skepticism regarding rationality that has come into fashion under Nietzsche's influence; if it really takes itself seriously, it forgoes the opportunity to present itself as a serious position.

See also: Action Theory: Psychological; Action, Theories of Social; Bounded Rationality; Free Will and Action; Intentionality and Rationality: An Analytic Perspective; Motivation and Actions, Psychology of; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory: Cultural Concerns; Rational Theory of Cognition in Psychology; Rationalism

Bibliography

Anscombe G E M 1957 *Intention*. Blackwell, Oxford, UK

Apel K-O, Kettner M (eds.) 1996 *Die eine Vernunft und die vielen Rationalitäten*. Suhrkamp, Frankfurt/Main, Germany

Black M 1977 Meaning and intention: an examination of Grices' views. In: Meggle G (ed.) 1977 *Analytische Handlungstheorie Band 1, Handlungsbeschreibungen*. Suhrkamp, Frankfurt/-Main, Germany

Carnap R 1966[1928] *Der Logische aufbau der Welt*, 3rd edn. Meiner, Hamburg

Davidson D 1980 *Essays on Actions and Events*. Oxford University Press, Oxford, UK

Davidson D 1982 Rational animals. *Dialectica* **36**: 317 ff.

Dennett D 1991 *Consciousness Explained*. Penguin, London

Dretske F 1995 *Naturalizing the Mind*. MIT Press, Cambridge, MA

Elster J 1979 *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge University Press, Cambridge, UK

Elster J 1983 *Sour Grapes: Studies in the Subversion of Rationality*, Cambridge University Press, Cambridge, UK

Gosepath S 1992 *Aufgeklärtes Eigeninteresse. Eine Theorie theoretischer und praktischer Rationalität*. Suhrkamp, Frankfurt/Main, Germany

Grice H P 1957 In: Meggle G (ed.) 1979 *Handlung, Kommunikation, Bedeutung*. Suhrkamp, Frankfurt/Main, Germany

Habermas J 1981 *Theorie des kommunikativen Handelns*. Suhrkamp, Frankfurt/Main, Germany

Habermas J 1989 *Zwecktätigkeit und Verständigung. Ein pragmatischer Begriff der Rationalität*. In: Stachowiak H (ed.) *Pragmatik. Handbuch pragmatischen Denkens* Band III. Meiner, Hamburg, Germany

Hollis M, Lukes S (eds.) 1982 *Rationality and Relativism*. MIT Press, Cambridge, MA

Meggle G (ed.) 1979 *Handlung, Kommunikation, Bedeutung*. Suhrkamp, Frankfurt/Main, Germany

Meggle G 1994 *Handlungstheoretische Semantik*. De Gruyter, Berlin

Melden A I 1961 *Free Action*. Routledge and Kegan Paul, London/New York

Nozick R 1993 *The Nature of Rationality*. Princeton University Press, Princeton, NJ

Parfit D 1984 *Reasons and Persons*. Clarendon Press, Oxford, UK

Putnam H 1981 *Reason, Truth and History*. Cambridge University Press, Cambridge, UK

Rescher N 1988 *Rationality*. Oxford University Press, Oxford, UK

Ryle G 1949 *The Concept of Mind*. Hutchinson, London

Schnädelbach H 1992 *Zur Rehabilitierung des animal rationale. Vorträge und Abhandlungen 2*. Suhrkamp, Frankfurt/Main, Germany

Schnädelbach H 2000 *Philosophie in der modernen Kultur. Vorträge und Abhandlungen 3*. Suhrkamp, Frankfurt/Main, Germany

Schütz A 1974 [1932] *Der sinnhafte Aufbau der sozialen Welt*, new edn. Suhrkamp, Frankfurt/Main, Germany

Schwemmer O 1973 *Theorie der rationalen Erklärung*. Beck, Munich, Germany

Searle J R 1983 *Intentionality. An Essay in the Philosophy of Mind*. Cambridge University Press, Cambridge, UK

Strawson P F 1971 Intention and convention in speech acts. In: *Logico-Linguistic Papers*. Methuen, London 1974, pp. 149ff.

von Wright G H 1963 Practical inference. *The Philosophical Review* **72**: 159ff.

Weber M 1974 *Gesammelte Aufsätze zur Wissenschaftslehre*, 4th edn. Mohr and Siebeck, Tübingen, Germany

Welsch W 1996 *Vernunft. Die zeitgenössische Vernunftkritik und das Konzept der transversalen Vernunft*. Suhrkamp, Frankfurt/Main, Germany

Wilson B R (ed.) 1970 *Rationality*. Blackwell, Oxford, UK

Winch P 1958 *The Idea of Social Science and its Relation to Philosophy*. Routledge and Keagan Paul, London/New York

H. Schnädelbach

Intentionality and Rationality: An Analytic Perspective

Intentionality and rationality are the two primary excellences which human, and perhaps some other, minds are understood to have. Intentionality is that directedness which allows thoughts to be about other things, even about the world. Because the mind is an intentional system, it can represent how things are. Rationality is that excellence which is concerned with how that representation should work so as to be accurate, and with how one should behave given one's thoughts about the world, both those that are about how it is (beliefs) and those about how one wants it to be (desires). This article discusses the nature of

intentionality, the question of what the objects of thought actually are, what relationship if any it has with consciousness, and some modern naturalistic (broadly scientific) accounts of what makes something intentional. Under the heading of rationality, the various putative domains of rationality are discussed: belief, desire, and action. This article concludes with a brief discussion of whether we in fact are rational, and of how to justify the normative claim that we ought to be rational.

1. Intentionality

Intentionality is the term given to the mind's striking capacity to have states which are about features in the world. In addition to weighing three grams, or being comprised of ectoplasmic goo, or conforming to a certain neural networking diagram, a mental state might have the property of being *about* something: perhaps *panna cotta* flavored gelato. This property —this 'aboutness'—is an intentional one. It is sometimes thought of as a kind of directedness. The mind is directed on things; your fears are directed on their objects; your beliefs likewise.

The term itself is derived from a medieval Latin translation of an Arabic technical term. The medieval discussion was about a rather different topic than the contemporary issues that are discussed under the rubric of intentionality—perhaps slightly closer to the contemporary debate about the nature of concepts (Sorabji 1991; see also *Natural Concepts, Psychology of*).

1.1 The World or Intermediate Objects?

A straightforward and appealing account of intentionality is to claim that it consists of some kind of relation between a mental state and the object it represents. This claim is neutral between the view that the relation is inexplicable by the natural sciences, and reductive naturalistic programs that try to explain just what scientifically mandated properties this relation is constituted by.

There are difficulties: presumably it is possible to have beliefs about things that do not exist. In some cases this can be finessed away—my belief about the gelato in the fridge, should there turn out to be no gelato in the fridge, might be redescribed as a (false) belief about the fusion of all gelato, that part of it is instantiated in the fridge. But such finessing is pointless, because there are *types* of things that don't exist, as when I have beliefs about Bunyips. So if my (false) belief about the Bunyips does not gain its intentionality in virtue of a relation between a mental state and Bunyips, what does the relation hold between?

One solution is that it holds between a mental state and a special, intermediate, intentional object. The intentional object might be there even if the mental state does not succeed in directing itself onto the world. This requires that there are such intentional objects to do the representational work. Some philosophers, including Meinong and latterly Routley and Zalta (Zalta 1988) hold variations on the doctrine that there are such entities, but they do not have full fledged existence: the nonexistent (sometimes called Meinongian) objects have some other kind of being which is enough to do the work.

Some proponents of the Meinongian view can allow that intentionality is always relational, but that in the case of existing objects the relation is between the mental state and the existing object, and in the case of the nonexistent objects it is between the mental state and the nonexistent object. This requires them to say (in the case of a fear of a Yeti) that the nonexistent Yeti is the very same object as the thing you would be fearing if there really were Yeti.

Perhaps in favor of indirect objects is that it seems that the contents of our beliefs, desires and so on are more finely discriminated that the things in the world. This can be noticed in so-called referentially opaque conditions (a kind of intensional context: see *Intentionality and Rationality: A Continental-European Perspective*). If Bill desires a Rosa brand Gelato, and if it so happens that Rosa brand is the worst in Sydney, most do not think that it follows straightforwardly that Bill desires the worst gelato in Sydney, even though the gelato he desires is (unbeknown to him) the worst in Sydney. However if some indirect intentional object is taken to be the content, then there might be two distinct intentional objects, only one of which is the object of Bill's desire, but both of which are themselves directed. One of the disturbing things about this approach is that it does not seem to make the problem go away. How do the special intentional objects do the work of ensuring that our mental states, that are in relation to them, get to be about the world? If it is a relation to the object in the world, then there is a regress problem. If not and there is just a mysterious power, then the mysterious power may also be possessed by the mental state directly.

There is also difficulty for naturalists in seeing how mental states can bear relations to mysterious indirect objects. Finally a more basic complaint is that even where there are such indirect entities, it seems that our thoughts are surely at least sometimes directly about the things in the world: it is the gelato itself Bill desires, not some intermediate object.

1.2 Intentionality and Consciousness

Some have thought that intentionality is a special feature that marks off mental phenomena from nonmental phenomena: all and only mental systems display intentionality. Something like this was intended by Brentano in the slogan claiming that

intentionality is the 'mark of the mental.' This is not meant to be a nomenclatural claim which merely says that, by definition, if a system is complex enough to be intentional then it is complex enough to be mental. Instead it seems to proceed by some connection between consciousness and intentionality. Thus, John Searle (1983), for example, thinks that a physical system which could do whatever reductionists claim is enough for representation (see Sect. 2) could nonetheless lack genuine intentionality, because it wouldn't be *about* the thing it co-varied with. The reasoning is that either such systems might not have what it takes for consciousness, and thus lack intentionality, or else the reverse: that since they might lack genuine intentionality, they might lack consciousness.

1.3 Reductive Approaches to Intentionality

Most common among contemporary approaches to intentionality are reductive attempts to say what scientifically respectable property of our minds it is that accounts for the difference between intentionality and nonintentionality.

The first of these approaches is to deny that the phenomenon of intentionality exists. So-called eliminative materialists (e.g. Churchland 1981) agree that we can specify scientifically respectable properties which, if they were instantiated, would guarantee that there is intentionality. But they do not think such properties are instantiated, and thus think science will discover that there is no intentionality. Since their arguments depend on stringent principles for the existence of intentional entities, there are two ways for non-eliminativist reductionists to argue against them: by accepting the standards and making the empirical claim that there are things of this nature, or by rejecting the standards.

Among those who think there is in fact intentionality in the natural world, there are, very broadly, four basic approaches in the contemporary literature: informational semantics, teleo-biological theories, broadly functionalist theories, and instrumentalist theories.

Informational semantics is the view that intentionality occurs in the case of the mind only when there is co-variation is between mental states and things in the world. Caveats about complexity, inter-relatedness, or causal and logical dependence are often added to upgrade this condition to sufficiency. Versions of these theories can be found famously in pieces by Fred Dretske (Dretske 1983) and Jerry Fodor (Fodor 1990).

Objections to the view are legion, the most troubling being variations of the so-called misrepresentation problem. The problem is that in situations where there is obvious error, the theory denies that there is error. Suppose that we are trying to have a mental state about cats. It co-varies with cats (i.e., it appears exactly when the agent is in the presence of cats), but occasionally—say about 1 percent of the time—it misfires in the dark and the state appears in the mind when a large rat is seen in the darkness. For simplicity suppose nothing else causes this cat symbol in the head. Intuitively, it seems we should say the symbol is about cats, and occasionally we deploy that symbol to say false things of cats—as when we think that there is a cat there, but unbeknown to us it is in fact a rat which has misled us. Unfortunately, the theory notes that there is perfect co-variation: between the mental symbol on the one hand, and cats *or* rats. Thus it says the symbol is about cats or rats. And so thus we have in every case a *true* belief about cats or rats—that they are present. Various solutions to this problem can be found in the references cited above, but none uncontroversially works.

Another objection is that no purely correlational fact could account for the thought that intentional states do not just reliably indicate the state of the world, but rather such indication is what they are normatively *for*. Teleo-semantics purports to find a solution to this problem. This is the cluster of views around the work of Ruth Millikin (Millikan 1984) and Karen Neander (Neander 1991), according to which evolutionary biology plays a central role. The leading idea is that states are genuinely representational because something like co-variation is the *proper biological function* of the states (or is produced by states whose function is producing states of this sort). This in turn is cashed out evolutionarily: a state's proper biological function is whatever it did that caused it to be selected by natural selection. Thus a heart's function is to pump blood but not to make a noise, because while hearts have always made noises, these noises did not contribute to the fitness of those organisms with hearts, whereas pumping blood did. So, in the case of mental states, while there may be many mental states which correlate with external features of the world, only those that do so because it was advantageous during evolutionary history get to count as intentional states. Objections are of two kinds. The first kind is that the states discriminated in such ways may not turn out to be remotely about the sorts of things that we imagine our intentional states pick out. The second kind involves worries about the historical individuation of mental states. If Creationism is true, for example, states would not be intentional—because they would have no evolutionary history. Alternatively, if Creationism is true then intentionality is an entirely different thing from what it would be on the assumption that Creationism is false. This would raise questions about what is in common between what is called 'intentionality' on the Creationist assumption and what is called 'intentionality' on the standard assumption that would make it right to give them the same name.

Functionalist theories say that intentionality is a phenomenon which arises when systems are com-

plicated enough to interact with the environment in a complicated way. So what matters is not just how the things in the world affect the presence or absence of the state, but how the state interacts with other states to affect the world. A state gets to be a belief that icecream is around, and thus a belief, and thus intentional, not only because it is an indicator of icecream, but also because it potentially interacts with the world. Combined with desires for icecream it may cause icecream directed behavior. In most functionalist accounts, there are few constraints on what the internal nature of a mind has to be like to count as intentional, so long as that nature will cause the right interaction with the world (Armstrong 1968, Braddon-Mitchell and Jackson 1996, Lewis 1972).

The final major category is instrumentalism (see *Realisms and their Opponents: Philosophical Aspects*). An instrumentalist about intentionality sometimes thinks that there is no intentionality as such, but it is pragmatically useful to imagine there is, because it allows us to engage in intentional explanations of behavior, which are both quicker and more practical than the full neurophysical story. Strictly speaking such views are versions of eliminativism which do not seek to eliminate *talk* of the properties they hold not to exist. Other theorists who are sometimes called instrumentalists hold that there are indeed intentional states, but that all it takes for a system to have intentional states is to produce effects in the world best predicted or explained by intuitional talk (see e.g. Dennett 1991). On some construals, this kind of instrumentalism is close to functionalism, since like the functionalist the view is realist about intentional states, and like the functionalist it places few constraints on the intrinsic nature of the states insofar as that nature is not relevant to their causal or explanatory profile. The differences center around whether talk of states apt for prediction and explanation is a distinct way of typing the states from the functionalist's way of typing causally. The answers to these questions about the differences between functionalism and realist instrumentalism will depend in part on views about the relationships between causation and explanation.

2. *Rationality*

2.1 *Introduction*

If anything is common to various accounts of rationality, it is that rationality is a special excellence of the mind. Perhaps it is to have one's higher mental processes functioning well. It must be narrowed down by excluding certain kinds of functioning—functioning well just in terms of crude computational power can be excluded. More controversially moral well functioning is often excluded (see *Rationalism*).

It can be best thought of in terms of criticism: one can criticize various mental process as irrational. Many think that one can criticize as irrational someone who fails to draw obvious deductive inferences; one can criticize someone as irrational who has strange and bizarre patterns of change in desire (though this is controversial as we shall see later), and one can criticize as irrational someone who fails to act so as to bring about what she desires in the light of her beliefs. On the other hand, it is more controversial whether one can criticize as irrational behavior or thought on the grounds that it is immoral and virtually no one thinks that not being able to perform Herculean tasks of mental arithmetic is enough to count as irrationality.

This article will cover three domains: belief, desire, and practical reason, and the question of how much there is in common between what is called 'rationality' in these domains. There is also a question as to whether rationality is a feature of emotions: is it rationally required or prohibited to be in various emotional states in various circumstances?

Rationality with respect to these domains has characteristic goals. The rational believer aims at truth, the rational desirer aims at best desires (perhaps here is a connection with morality), and the rational actor aims to act in a way so as to satisfy her desire in the light of her beliefs. An unresolved issue is whether rationality is characterized as whatever processes bring about these goals, or whether rationality can be characterized by the nature of the processes themselves, and we are lucky enough to live in a world where these processes do in fact bring about the goals on average better than other processes.

2.2 *Rationality of Belief*

In the domain of beliefs, agents are held to be irrational if they form or revise beliefs in the wrong way. At least three kinds of belief forming mechanisms are at issue: the source of initial beliefs, the updating of beliefs on the basis of evidence, and the deduction of the implications of these beliefs in the absence of further evidence.

There is no clear consensus on how initial beliefs should be sourced. Bayesian probability theory holds that there is no way of assessing the rightness of the initial probabilities that agents give to various claims about the world. Instead, the method for revising beliefs in the light of evidence is something that the theory specifies. So on this tradition, an agent's beliefs are rational just if they are revised in the light of evidence the way the theory specifies (see e.g., Jeffrey 1983). There are competing systems for updating beliefs in the light of evidence.

Some philosophers, however, grant special epistemological status to some initial beliefs (or probability assignments) depending on where they come from. Initial beliefs might be held to be held rational insofar as they come from approved sources. These include the testimony of others (see *Expert Testimony*)

or claims about sense impressions (see *Logical Positivism and Logical Empiricism*).

The rational constraints on drawing inferences from existing beliefs unaided by further evidence is the domain of the various branches of deductive logic (see *Deductive Logic*). Of course there is a richer sense of rationality in this area: the capacity to draw the *important* conclusions and ignore the irrelevant ones.

Artificial intelligence researchers have long seen how irrational seeming a system will be that just sits and cranks out all the irrelevant consequences of what it believes.

2.3 Rationality of Desire

While most philosophers would allow that beliefs, or at least the process of change of belief, can be legitimately subject to criticism on the grounds of rationality, there is less consensus in the case of desires. Hume's famous dictum that it is not contrary to reason for him to prefer the destruction of the world to the scratching of his finger, is an illustration of his idea that foundational desires are not subject to constraints of reason. An agent simply has the basic desires she has; the job of reason is to modify beliefs with the goal of truth, which is turn serves the task of satisfying desires. A way of understanding the motivation for this is through considering the *point* of reason. If you think that reason has goals such as truth or the satisfaction of desire, then it is easy to see how belief formation can be irrational—it fails to track the truth—and how action can be irrational—it fails to satisfy desire in the light of belief. It is harder to see what goals basic desires have—though perhaps ease of satisfaction might accord with one of these desiderata, in which case a kind of quasi-Buddhist constraint might recommend itself, of desire minimization. It is here also that connections between rationality and ethics might be drawn, since one constraint might be that the goal of desire is to motivate oneself towards the good life, and rational desires are those which do this (see *Rationalism*; *Economics and Ethics*).

But even if there is no constraint on basic desires taken singly, there might be some constraints on combinations. In the same way that one can rationally criticize someone for inconsistent beliefs, on the grounds that not all can be true, one might criticize someone for incoherent desires, on the grounds that not all can be satisfied. But this will be a much messier business, for it is much less clear that there is anything wrong with a desire set, not all of which can be satisfied. For contingency purposes, it is never the case that all our desires are satisfied, and it is hard to see why it is worse if this is elevated to a logical truth in the case where they cannot all be satisfied because two desires are logically exclusive. Indeed one can imagine that there might be some inconsistent desire sets which lead to greater flourishing than some consistent desire sets—in which case this constraint is at odds with the ethical one above. But supposing that it is a desideratum to have consistent desires: is 'rational' the right word to describe such sets? This might be thought to be an essentially terminological issue. Certainly the situation is isomorphic with the case of consistent beliefs, but whether the differences are small enough to justify using the same term may be an unsettleable piece of semantics.

There is a further question about desire evolution. How should desires change over time? Most grant that along with changes in beliefs, there are rationally mandated changes in instrumental desires. If I desire to be cured of disease, and I learn that a certain berry contains the cure, then I have a reason to acquire a desire to eat the berries in the service of my desire to be. But what of desire change over time under the impact of experience which is not grounded instrumentally. Some think that greater general consistency across time and other coherence factors are rational constraints. On this view, someone whose desires change all the time in apparently senseless ways, wishing for world peace and bananas one day, and world war and fruitlessness the next, is irrational. Certainly such a person in practice would have inconsistent desires (most desire happiness, and fluctuating desires is a sure-fire killer of that state). In these cases the view that co-satisfiability is a requirement would be enough to claim that such a person is irrational. But what if the wildly fluctuating desires were always synchronically consistent? This would still be a very bad way to be, though the author thinks that it is again a borderline terminological matter whether there is enough in common between this malady and others to call it irrational rather than unfortunate.

2.4 Rationality of Action

Rationality of action is most commonly called practical rationality. It is the kind of rationality which concerns the relations between one's beliefs and desires on the one hand and one's actions on the other. If you desire ice-cream, if the ice-cream is in front of you, and if there are no conflicting desires or other impediments to eating it then it is a principle of practical rationality that you should pick it up and lick. The noneater is practically irrational.

Sometimes practical rationality is characterized as reasoning *about* action. Some interpreters of Aristotle and Kant see their work in this way. Similarly contemporary work on decision theory and game theory is often understood as what conclusions about what one *ought* to do should be drawn from premises about beliefs and desires. Another way of understanding practical rationality is the relationship between action itself and belief and desire. The difference is just that on the first way of understanding it one is

practically irrational if one holds wrong *opinions* about how one should act given one's beliefs and desires, and thus it is a special case of rationality about belief or theoretical reason. In the other way of understanding it, one is practically irrational if one actually acts in the wrong way given a set of beliefs and desires.

Much modern work on practical reason has focused on various calculi which either describe or prescribe how to act given one's beliefs and desires. Decision theory (see e.g. Skyrms 1986) (see *Decision Theory: Bayesian*; *Decision Theory: Classical*) is a family of theories which, on the assumption that features of the world relevant to one's decisions are themselves unaffected by those decisions, aims to give an precise account of how to choose Game theory (see *Game Theory*) is the calculus which has been developed to compare actions in situations where the situation itself depends on ones actions, as in many games, where one's own moves may determine the moves of others and thus affect one's assumptions about how the world is that are relevant to choice (see e.g., Elster 1979).

One central philosophical worry with these theories stems from an account of what desires are. Decision and game theory both aim, it seems, to prescribe how we should act. Thus we ought to be able to rationally criticize someone who has acted irrationally and flouted the favored theory. Yet the most popular account of desires (or preferences as they are usually called in the literature) is so-called revealed preference theory. According to revealed preference theory, one's desires are ascribed according to one's actions. In other words, you see what someone does, you make some assumptions about what they believe on the assumption that they are rational with respect to belief formation, and then assume that their desires are whatever it would take to produce the action. But then you can never criticize anyone's actions, for you can always attribute desires, however strange, that make sense of actions. According to revealed preference theory these are the desires that the person had.

In its most extreme version it is hard to use the theory to make recommendations about action as well. For it is hard to say what one's desires are until they are revealed in action, and in strong versions they are *constituted* by the actions (among other things) so there is no fact of the matter.

So if the theories are to have prescriptive content, then we need a substantive theory of desire internal to the workings of the mind, a theory that is not tied so closely to what one actually does.

One phenomenon, about which there has been a lot of philosophical discussion ever since Aristotle, and which is connected to these issues, is so-called weakness of the will. Sometimes it seems that we act against our own desires. You might overwhelmingly desire to loose weight, resolve to avoid icecream, but eat it nonetheless. It might seem impossible to act against your desires if desire is constituted by actions, for there is a desire posited to explain any action. Theories of weakness of will divide between those that accept this and posit sudden changes of desire or hidden desires to explain the behavior, and those that posit a nondesire-based mechanism that makes one sometimes behave against all the things considered desired one has at the time of the weak-willed act.

2.5 *Are We Rational?*

Recently there has been much focus on the question of whether, as a matter of fact, we are rational or not. Much of this has focused on theoretical reason or rationality about belief. Tests show that when examples of simple reasoning are given to even quite intelligent people, large proportions draw wrong inferences. Even simple *modus ponens* can be got wrong. Elementary statistical reason goes badly wrong, though how badly wrong depends on how the information that subjects are expected to use to draw their conclusions is presented (see e.g. Stich 1990, Chap 1).

There has sometimes seemed to be a debate about whether this view that irrationality is omnipresent conflicts with the views of some evolutionary psychologists who see the brain as a marvellously well-engineered machine with many modules for making calculations about the environments in which it evolved. But there is surely no conflict. For what is marvellously engineered for the environments we evolved in need not be so great for our current one, and modules that calculate quickly and efficiently in demanding real-time environments are unlikely to use the same algorithms as would be best for giving considered opinions when time is not an issue (see Stich and Samuels 2000).

2.6 *Ought We be Rational?*

What the question of whether we are rational or not does raise, though, is the issue of whether we ought to be. For if humans are in fact to a substantial extent not theoretically rational, then the claims that a theory of rationality has to be purely descriptive cannot be upheld. For it would be a false description. So perhaps this should be thought of as a purely normative, a theory of how we ought to think and act. And the question is why? If we lack of a foundational story about why one ought to adhere to some norms of thought, then either a scepticism about rationality or a relativism about rationality—the view that that there are many sets of norms competing in the same domains, and there is no fact of the matter about which ought generally to be adopted—become attractive. The latter view has acquired some prominence in recent years) (see *Relativism: Philosophical Aspects*).

One answer to the question of why we should adopt a set of rational norms is found in the pragmatic school: it will be somehow good for use or useful. But this is perilously subject to empirical refutation. For the evolutionary theories seem to suggest that aiming to achieve true beliefs at any cost is in fact not good for us. Perhaps quick and dirty solutions are more life preserving, even at the cost of falsehood. There is also speculation that optimism is a condition in which people systematically overestimate the likelihood of good outcomes and underestimate the likelihood of bad ones, while at the same time getting on better in life than the more accurate pessimists.

A more modest claim, but still unsatisfactory to many, is to make rationality itself a hypothetical imperative. Insofar as you care about truth, exhibit theoretical reason. Insofar as you care about coherent desires, make them coherent. Insofar as you care about satisfying your desires, act so as to make them most likely to be satisfied in the light of your beliefs. Of course even the modest claim, to have any force, requires a principle very like practical reason: insofar as you care about something, follow the rule that will give it to you.

2.7 Rationality and Reasoning

The fact that rationality is often thought of as a characteristic human excellence, and so is the power of conscious thought, has sometimes led to a confusion between rationality and conscious, deliberate reasoning. In fact it is very useful to distinguish between them. Much of what we do is not the result of conscious deliberate reasoning (sometimes called ratiocination). We make decisions, pursue our goals, and act in a way which can be praised as rational or criticized as irrational, despite the fact that it may not happen in a conscious way. If it were only conscious deliberate reason that was rational, then rationality would be even harder to find than it is. Of course, in showing that a piece of cognition is rational, it sometimes helps to provide an argument which shows how the process can be *rationalized*, i.e. described in a way that might look like a valid piece of deliberation or ratiocination. But this is not to say that this is what originally went on. Thus we might say of animals, even if we do not think they possess the power of conscious reasoning, that their mental processes are rational insofar as they process the information they receive about the world and respond in a rational way.

See also: Bounded Rationality; Cognitive Science: Overview; Decision Theory: Bayesian; Intentionality and Rationality: A Continental-European Perspective; Irrationality: Philosophical Aspects; Knowledge (Explicit and Implicit): Philosophical Aspects; Knowledge Representation; Perception: Philosophical Aspects; Practical Reasoning: Philosophical Aspects; Problem Solving and Reasoning, Psychology of; Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory: Cultural Concerns; Rational Theory of Cognition in Psychology; Rationalism; Rationality in Society; Reference and Representation: Philosophical Aspects

Bibliography

Armstrong D M 1968 *A Materialist Theory of the Mind*. Routledge & Kegan Paul, London

Braddon-Mitchell D, Jackson F 1996 *The Philosophy of Mind and Cognition*. Blackwell, Oxford, UK

Churchland P M 1981 Eliminative materialism and the propositional attitudes. *The Journal of Philosophy* **78**: 67–90

Dennett D 1991 True believers. In: Rosenthal D M (ed.) *The Nature of Mind*. Oxford University Press, Oxford, UK, pp. 339–53

Dretske F 1983 Precis of dreske knowledge and the flow of information. *Behavioural and Brain Sciences* **6**: 55–63

Elster J 1979 *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge University Press, Cambridge, UK

Fodor J 1990 *A Theory of Content and Other Essays*. Bradford Books, MIT Press, Cambridge, MA

Jeffrey R C 1983 *The Logic of Decision*. University of Chicago Press, Chicago

Lewis D 1972 Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* **50**(3): 249–58

Millikan R G 1984 *Language, Thought and other Biological Categories*. MIT Press, Cambridge, MA

Neander K 1991 Functions as selected effects: the conceptual analyst's defense. *Philosophy of Science* **58**(2): 168–84

Searle J R 1983 *Intentionality*. Cambridge University Press, Cambridge, UK

Skyrms B 1986 *Choice and Chance*, 3rd edn. Wadsworth, Belmont, CA

Sorabji R 1991 From Aristotle to Brentano: the development of the concept of intentionality. In: Blumenthal H, Robinson H (eds.) *Oxford Studies in Ancient Philosophy*

Stich S P 1990 *The Fragmentation of Reason*. Cambridge University Press, Cambridge, MA

Stich S, Samuels R 2000 Ending the rationality wars: how to make disputes about human rationality disappear. In: Elio R (ed.) *Common Sense, Reasoning and Rationality*. Oxford University Press, New York

Zalta E N 1988 *Intentional Logic and the Metaphysics of Intentionality*. MIT Press, Cambridge, MA

D. Braddon-Mitchell

Interactionism and Personality

1. Personality

Personality is a central concept in psychological theorizing and empirical research. Much has been written about personality but there still seems to be no

commonly accepted and applied definition (van Heck 1997). One consequence is fragmentation of relevant knowledge, a situation that has been described as follows: 'Contemporary academic investigation of personality makes little attempt to sustain a consistent, generally applicable view of personality, preferring an eclectic use of heterogeneous data, which are then scrutinized for patterns' (*New Encyclopedia Britannica*, 1991, **9**: 312). As emphasized by researchers from different positions, the fragmentation is detrimental and an obstacle to scientific progress in personality research. Research in this area needs the formulation of a general, integrated theoretical perspective that can serve as the framework for planning, implementing, and interpreting empirical studies on specific issues.

Although diverse, most of the definitions have an element in common; implicitly or explicitly they regard the total individual, 'the undivided personality' (Allport 1924, p. 132), as the target of observation and generalization in theorizing and empirical research. Rather than trying to contribute another semantic definition of what personality 'really is,' the presentation here starts from the view that what can be the target of scientific analysis in personality research is an active and purposeful organism, functioning and developing as a total, integrated being in interaction with different aspects of the environment. With reference to this view, the task for personality research is to contribute knowledge about how and why individuals as a total, integrated organism, think, feel, act, and react in real life.

A full discussion of the implications of this task formulation requires a distinction between three perspectives: a current, a developmental, and an evolutionary perspective. The three perspectives are complimentary. However, most empirical personality research has a current perspective and this article is focused accordingly.

Models using a current perspective analyze and explain why individuals functions, in terms of their contemporaneous mental, behavioral, and biological states, independently of the developmental processes that might have led to the present state of affairs.

2. *Dynamic Interaction*

Dynamic interaction is a basic principle in the functioning of living organisms. Two concepts are central in dynamic interaction: process and reciprocity.

(a) A fundamental starting point for theoretical discussions and empirical research on individual functioning is its character of integrated, dynamic, complex, and adaptive processes. A basic feature of these processes is time. The processes at different levels of individual functioning cannot be properly understood by looking at snapshots of states; we need films. The importance of introducing time in empirical studies has been demonstrated in several ways. For example, access to advanced techniques for brain-imaging has made it possible to investigate the current functioning of the brain while an individual is performing a mental task. This has contributed to new understanding of the interaction of mental acts, biological structures and processes, behavior, and events in the environment.

(b) Reciprocity or bidirectionality in the functioning of operating elements is a characteristic feature of dynamic processes at all levels of the integrated person–environment system, both horizontally among elements and systems functioning at the same level of the organism and vertically between systems functioning at different levels. At the cellular level of the individual, the growth, functioning, and death of a particular cell is dependent on its continuous reciprocal exchange of information with neighboring cells. The effective functioning of any subsystem of the body, for example, the coronary system or the immune system, depends on its reciprocal interaction with other bodily systems. The way an individual functions socially in a specific situation is characterized by continuous reciprocal interaction with other individuals participating in the same situation. The individual's self-identity and world views are formed and develop in a process of reciprocal interaction with the environment.

3. *Classic Interactionism*

During the twentieth century a number of prominent researchers have argued, from very different viewpoints, for what might be designated classic interactionism: Kantor (1924, 1926) the behaviorist, Murray (1938) in his need-press theory, Murphy (1947) in his biosocial theory, Tolman (1951) focusing on purposive behavior, Sullivan (1953) discussing interpersonal relations, Rotter (1954) in his social learning approach, and Sells (1963) arguing explicitly for an interactionistic position, to mention the most influential in personality psychology. The central idea expressed by these theorists was summarized in Lewin's (1936) formula $B = f(P, E)$; behavior (B) is a function of the person (P) and the environment (E). The equation reflects three common elements of classic interactionism: (a) the focus is on person–environment relations, (b) behavior is regarded as an outcome variable, and (c) behavior cannot be understood and explained without considering the environment, particularly the situation in which behavior occurs. This focus on the interplay of individuals and environments in classic interactionism is also reflected in the different contextual models which have been presented and are influential in developmental psychology.

In spite of the very skillful analyses of central issues concerning the individual–environment relations,

there is a surprising lack of explicit discussions of interactional processes in the early contributions. The term interaction seldom appears in the writings of the theorists referred to above. Depending on theoretical preference, P and E in the B = f(P, E) equation were more readily addressed than the functional relations between them.

During the 1970s the discussion was enriched by the introduction of mediating cognitive and motivational aspects of individual functioning for understanding person–environment interactions. Summarizing the state of the art of classic interactionism, Endler and Magnusson (1976) emphasized the individual as an active agent in the person–situation interaction process, the role of cognitive and emotional factors in that process, and the importance of the subjective meaning of situations and environments in person–environment interactions (Magnusson 1976).

Empirical research using an interactionistic perspective was almost nonexistent before the 1960s. Such empirical studies as were presented mainly aimed at demonstrating the role of situations in the person–situation interaction. During the 1970s and 1980s, the explicit formulations of a classic interactionist view exerted a twofold impact on personality research. First, the issue of cross-situational consistency/inconsistency in individual functioning became a central topic for theoretical debate and empirical analyses (Magnusson and Endler 1977). Second, this debate led to an interest in theoretical taxonomies and empirical analyses of situational characteristics (Magnusson 1981, Forgas and Van Heck 1992).

4. *Modern Holistic Interactionism*

A modern interactionistic view considers the way an individual and the integrated person–environment system function in a holistic perspective. Within this framework, individual functioning can be described and analyzed in terms of complex, dynamic, and adaptive processes involving mental, biological, and behavioral aspects of the individual and different aspects of the environment (Magnusson 1999, Magnusson and Törestad 1993).

The two basic concepts of interactionism are central in this perspective: process, implying the time perspective, and reciprocity among operating elements and systems. At each level, the totality derives its characteristic features and properties from the interaction of the elements involved, not from the dynamic effect of each isolated part on the totality. Each aspect of the structures and processes that are operating in the individual (perceptions, plans, values, goals, motives, biological factors, conduct, etc.), as well as each aspect of the environment, takes on psychological meaning from its role in the total functioning of the individual. A certain element or system derives its significance, not from its structure, but from its role in the system of which it forms a part.

In three inter-related ways this view is an extension of classic interactionism. First, behavior is not regarded as an outcome variable, that is, as the result of the processes. It is a certain aspect of individual functioning and an integrated part of the total dynamic, adaptive interaction process. Second, interactional processes are not restricted to the interplay between the individual and the environment. Dynamic, adaptive interaction is also a basic principle for the processes within the individual, in which mental, biological, and behavioral factors are involved. Third, the character of these individual-bound processes is important in the individual's continuous interaction with the environment. Besides being influenced by characteristics of the environment, these processes have an impact on the person–environment interaction processes.

A heated issue for debate during the 1970s and 1980s was the 'existence' and role of personality traits. Much of the debate reflected a misunderstanding of the interactionistic view as in conflict with the trait concept. There is no contradiction between the analysis and description of individual functioning at the trait level and an interactionistic view: the essential argument concerns the trait concept's explanatory contribution to understanding individual functioning in specific situations and across situations. It is a matter of level of analysis. What Allport (1937) defined as 'bona fide structures in each personality that account for the consistency of its behavior' (p. 37) form an essential basis for understanding coherence in the individual's interaction with situations of different character and for coherent continuity in the individual's interaction with the environment over time. Mischel and Shoda (1995) recently contributed to the discussion on this issue by analyzing traits in terms of processing dispositions with reference to social, cognitive, and biological models.

A holistic view of individual functioning has ancient roots. Yet it is also very modern and in line with developments in other disciplines concerned with dynamic, complex processes. In recent decades it has derived substantive content from research on both cognitive and noncognitive factors and on structures and processes of the brain and the physiological system. In fact, a holistic view is inherent in the rapid development of research in the interface of mental, biological, and behavioral aspects of individual functioning. Cognitive neuroscience is one of the most dynamic areas of relevance for understanding the way individuals think, feel, act, and react in real life.

5. *Methodological Implications*

At a general theoretical level, the holistic–interactionistic perspective in personality research is

being increasingly accepted. However, little empirical research has been performed and interpreted with explicit reference to this perspective (Endler and Parker 1992). In their review of recent personality research, Endler and Speer (1998) concluded: '… the general research methodologies employed in the personality area have not changed a great deal since the 1960's' (p. 667). This tradition has been dominated by the application of statistical models and methods for the study of relations among variables across individuals at group level; the focus is on variables.

The formulation of a holistic interactionistic theoretical framework has led to the development and application of models and methods in which the individual is the organizational unit, a theoretical approach which is designated the person approach. In a person approach, individual functioning is studied empirically in terms of patterns of operating factors at the level that is relevant for the specific study. The development of adequate methods for the application of a person approach has been fostered by the presentation of modern models for dynamic, complex processes, particularly general systems theory. In other disciplines the formulation of such models has contributed to impressive progress.

6. *Modern Interactionism as a General Theoretical Framework*

As a result of the rapid development of research on mental and biological aspects of individual functioning and the contributions from modern models for dynamic, complex processes, the modern holistic, interactionistic view now offers a stable platform for further scientific progress, enabling personality research to fall into step with recent developments in other disciplines in the life sciences.

A holistic view serves two inter-related general purposes: (a) as a common conceptual space for effective communication among researchers concerned with very different problems, and (b) as a theoretical framework for the planning, implementation, and interpretation of specific empirical studies. No single researcher can be an expert on all problems at all levels of individual functioning and development. Researchers in cognition and antisocial behavior are concerned with different problems and may apply different methods for observation and data treatment. A common theoretical framework enables them to communicate and gain from each other's empirical work more effectively. Thus, the modern holistic interactionistic perspective forms the basis for a synthesis of knowledge, gained from empirical studies. At the same time the following implications of a modern interactionistic frame of reference should be kept in mind. First, it does not imply that specific mental, behavioral, and biological aspects of individual functioning cannot or should not be the object of empirical research. Second, it does not offer a hypothesis or an explanation for all problems. Within the holistic framework general principles for scientific investigation are applicable: the Newtonian model did not answer every question about the structure and functioning of the physical world but it did serve the two purposes summarized above.

See also: Allport, Gordon W (1897–1967); Personality Assessment; Personality Psychology; Personality Psychology: Methods; Personality Structure; Personality Theories

Bibliography

Allport G W 1924 The study of the undivided personality. *Journal of Abnormal and Social Psychology* **19**: 131–41

Allport G W 1937 *Personality: A Psychological Interpretation*. Holt, New York

Endler N S, Magnusson D 1976 Toward an interactional psychology of personality. *Psychological Bulletin* **83**: 956–79

Endler N S, Parker J D A 1992 Interactionism revisited: Reflections on the continuing crisis in the personality area. *European Journal of Personality* **6**: 177–98

Endler S R, Speer R L 1998 Personality psychology: Research trends for 1993–1995. *Journal of Personality* **66**: 621–69

Forgas J P, van Heck G L 1992 The psychology of situations. In: Caprara G V, van Heck G L (eds.) *Modern Personality Psychology*. Harvester, New York, pp. 418–55

Kantor R 1924 *Principles of Psychology*. Principia Press, Bloomington, IN, Vol. 1

Kantor R 1926 *Principles of Psychology*. Principia Press, Bloomington, IN, Vol. 2

Lewin K 1936 *Principles of Topological Psychology*. McGraw-Hill, New York

Magnusson D 1976 The person and the situation in an interactional model of behavior. *Scandinavian Journal of Psychology* **17**: 253–71

Magnusson D (ed.) 1981 *Toward a Psychology of Situations*. Erlbaum, Hillsdale, NJ

Magnusson D 1999 Holistic interactionism: A perspective for research on personality development. In: Pervin L A, John O P (eds.) *Handbook of Personality. Theory and Research*, 2nd edn. Guilford Press, New York, pp. 219–47

Magnusson D, Endler N S (eds.) 1977 *Personality at the Crossroads: Current Issues in Interactional Psychology*. Erlbaum, Hillsdale, NJ

Magnusson D, Törestad B 1993 A holistic view of personality: A model revisited. *Annual Review of Psychology* **44**: 427–52

Mischel W, Shoda Y 1995 A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review* **102**: 246–68

Murphy G 1947 *Personality: A Biosocial Approach in Origins and Structure*. Harper, New York

Murray H A et al. 1938 *Explorations in Personality*. Oxford University Press, New York

Rotter J B 1954 *Social Learning and Clinical Psychology*. Prentice-Hall, New York

Sells S B 1963 An interactionist looks at the environment. *American Psychologist* **18**: 696–702
Sullivan H S 1953 *The Interpersonal Theory of Psychiatry*, 1st edn. Norton, New York
Tolman E C 1951 Psychology versus immediate experience. In: Tolman E C (ed.) *Collected Papers in Psychology*. University of California Press, Berkeley, CA (reprinted from *Philosophy of Science*, 1935)
Van Heck G L 1997 Personality and physical health: Toward an ecological approach to health related personality research. *European Journal of Psychology* **11**: 415–43

D. Magnusson

Interactionism: Symbolic

Symbolic interactionism constitutes a longstanding and resilient theoretical perspective and research tradition in sociology and particularly within sociological social psychology, wherein it has been referred to as one of the three major faces of social psychology (House 1977). In contrast to perspectives that emphasize internal (biological or psychological) or external (structural or cultural) determinants of human social behavior, symbolic interactionism argues that most nonhabituated social actions take their particular form during the course of interpretive interaction in concrete social situations. Hence, the perspective's orienting premise is that an understanding of most human social action, be it individual or collective, requires an appreciation of its situated, interactive, and interpretive character. This article provides an overview of the perspective's historical roots and development, its central principles, and its criticisms and theoretical and research variations.

1. *Historical Roots and Development*

Although the phrase 'symbolic interactionism' was coined by Herbert Blumer in 1937, the various ideas and arguments associated with the perspective are clearly rooted in the philosophy of pragmatism, and particularly in the lectures and writings of one of its chief proponents, George Herbert Mead (see Miller 1973, for a comprehensive discussion of Mead's work and a complete listing of his writings, and Joas 1985, for a contemporary assessment). Mead, along with fellow pragmatists John Dewey, William James, and Charles Pierce, the social psychologist Charles Horton Cooley, and a group of sociologists at the University of Chicago, namely Robert Park and W. I. Thomas, provided the initial seedbed of ideas that eventually germinated, mutated in different but overlapping ways, and diffused among successive generations of sociologists. The two most prominent students among the first generation were Herbert Blumer and Everett Hughes, who taught and mentored, directly or indirectly, a wave of students who matriculated at the University of Chicago, mainly in the decade following World War II. Included among these students were Howard Becker, Elliot Freidson, Erving Goffman, Joseph Gusfield, Helena Lopata, Tamotsu Shibutani, Gregory Stone, Anselm Strauss, and Ralph Turner, all of whom became prominent sociologists defining and shaping a variety of subfields, such as collective behavior, deviance, organizations, social problems, the study of face-to-face interaction in public places, and social psychology, with an emphasis on roles, self, identity, and socialization processes. It is this set of scholars that played a major role in the diffusion of the evolving perspective as they spread out across the USA at various colleges and universities and who, along with their mentors and the initial seeded scholars, comprise what has been referred to as the 'Chicago School of Sociology,' or at least a major variant thereof (Fine 1995).

The interactionist perspective blossomed in a number of other sociology departments as well, with those at the universities of Iowa, Minnesota, California at San Diego, and Indiana becoming training centers, with somewhat different emphases, for successive but smaller cohorts of students. The Iowa program, founded and guided by Manford H. Kuhn and then sustained by Carl Couch, even became designated a 'school' of interactionism, largely because it had a different emphasis than the so-called Chicago School. But there has never been any significant debate about Chicago being at the core of the development and diffusion of the perspective.

2. *Central Principles*

Neither all of the aforementioned scholars nor all of their students would identify themselves as symbolic interactionists. However, if attention is focused on the character of their scholarly research and writings, it is clear that what most of them have done and continue to do reflects a number of pragmatist themes that congeal into what can be thought of as the core principles of symbolic interactionism. Herbert Blumer, who perhaps more than anyone else was associated with the articulation of the interactionist perspective, or at least a major variant of the perspective, contended that there are three such principles: that people act towards things, including each other, on the basis of the meanings they have for them; that these meanings are derived through social interaction with others; and that these meanings are managed and transformed through an interpretive process that people use to make sense of and handle the objects that constitute their social worlds (Blumer 1969, see Denzin

1989, for an expansive discussion of this interpretive process).

Embracing these three principles and reflected in most work today that falls under the interactionist umbrella are four broader and even more basic orienting principles:

(a) The principle of interactive determination, which holds that neither individual or society, nor self or other, are ontologically prior but exist only in relation to each other and therefore can be fully understood only in terms of their interaction;

(b) The principle of symbolization, which highlights the processes through which events and conditions, artifacts and edifices, people and aggregations, and other features of the ambient environment take on particular meanings, becoming objects of orientation that elicit specifiable feelings and actions;

(c) The principle of emergence, which focuses attention on the nonhabituated side of social life and its dynamic character and thus the potential for change, not only in the organization and texture of social life but in associated meanings and feelings as well; and

(d) The principle of human agency, which emphasizes the active, goal-seeking character of human actors.

The basic insight suggested by these principles is that the objects of analysis—be they crowds, face-to-face encounters in public places, social problems, organizations, work and occupations, illness, identities, self-concept and -esteem, or emotions—cannot be fully understood apart from the interactive web or context in which they are situated and the interpretive work of the actors involved. Because of this, interactionists generally have been highly critical of theoretical perspectives, analytical schemes, and research methodologies that neglect or gloss over interpretative, interactive processes. This does not mean that the perspective dismisses the influence of biological, psychological, historical, or structural factors in the determination and explanation of behavior; rather, it suggests that such factors may constitute predispositions to or constraints on action without necessarily determining the character of the action. Thus, from the standpoint of symbolic interactionism, social actors take into account the structural and cultural factors (e.g., role constraints, social expectations, norms, values) that impinge on the situations in which they find themselves in course of developing their respective lines of action.

Given the underlying view of the human being as an active rather than merely responsive organism, and the importance of interpretive interaction to the perspective, it is not surprising that 'the self' has been center stage in symbolic interactionist theorization and research. This is due in large measure to the self's reflexive character, as manifested in the ability to define, name, and act towards oneself just as one acts towards other objects. And it is this reflexive capacity that makes the self the central mechanism through which interpretation occurs and which, in turn, allows for the formulation of lines of action.

3. *Criticisms and Variations*

Because of the emphasis placed on situated, interpretative interaction, and thus human agency, critics of the perspective have argued that it ignores organization, power, and politics and is therefore astructural (see Reynolds 1993, Chap. 9, for a review of this and related criticisms). Like most critical characterizations, this one is only partly true. As previously noted, social actors, whether individual or collective, are said to take structural factors—such as role constraints, power differences, and organization—into account when anticipating or engaging in social behavior. Since empirical analyses of this process often focus on interactions that take place at the face-to-face, microscopic level of social life, as between people in public and quasi-public places (see Lofland 1989, for a summary of relevant research), it can be argued that macrostructural and mesostructural forces and processes are overlooked. Yet, it has been argued, as well, that social order, whatever the level, is constituted and reconstituted during the course of everyday interaction (see Goffman 1983), whether it is face-to-face or mediated.

Additionally, a good deal of interactionist research has occurred in and/or focused on organizational, mesostructural contexts (e.g., hospitals, shop floors, restaurants, orchestras, and spheres of industry like the automobile industry), and concepts that reference structural entities and forces (e.g., negotiated order, networks, and constraint) have increasingly found their way into interactionist research and analyses (see Fine 1993, Maines 1977, and Reynolds 1993, for literature reviews). Issues of political power, conflict, and organization also come into play in the constructionist perspective on social problems inspired by symbolic interactionism. Proponents of this work focus on the interpretive, claims-making activities of social problem entrepreneurs. Although there is considerable debate among adherents of this perspective as to the relative importance of putatitive objective social conditions in relation to the success of claims-making activities (Miller and Holstein 1993), it is recognized that these activities do not occur in a social vacuum but in a context of competing and conflicting claims in which some claimants or crusaders are advantaged over others for various organizational and political reasons. The medicalization of various forms of deviance and social problems is a telling case in point (Conrad and Schneider 1980).

Structural factors also figure in interactionist analyses of self-concept and identity, social roles, and emotions, but with varying degrees of emphasis. The

previously noted distinction between the Chicago and Iowa schools of symbolic interactionism reflect this different emphasis, with adherents of the former considering the self as a process that can be best studied through ethnographic fieldwork, and those associated with the latter conceptualizing the self primarily as an object that lends itself to study via survey research techniques and even laboratory experimentation (see Reynolds 1993, particularly Chap. 4). More recent research on the concept of identity parallels this distinction. Linked more closely to the Iowa school are approaches to identity that accent its structural locus, arguing that the source and relative salience of our identities resides primarily in the roles we play and the social relationships in which they are embedded (McCall and Simmons 1978, Stryker 1980). In contrast is work that approaches identity from a more processual, negotiated perspective (Goffman 1963, Snow and Anderson 1987, Strauss 1959). Both approaches acknowledge the influence of structural factors (e.g., roles and social relationships) in relation to identities and their relative salience, but the latter grants the actor greater latitude in negotiating and weighting the structural forces and impositions encountered. Similar differences in interpretive and structural weighting are also found in interactionist scholarship on roles (Stryker 1980, Turner 1962, 1978) and emotions (Ellis 1991, Hochschild 1979, Smith-Lovin and Heise 1988).

Taken together, these observations indicate that there is considerable variation in the range of work that falls under the interactionist umbrella, and that blanket criticisms, such as that the perspective is plagued by an astructural bias, are therefore likely to be off the mark. One final illustration of the extent of this internal variation is the recent methodological debate between those who take a radical, postmodern approach to research, seeing all data as biased and fictionalized, as a kind of second- or third-order reality, and those who regard careful research as a means for securing reasonable approximations of social reality (see Denzin 1992 and Fine 1993, for discussion of and references to this debate).

Given these debates and variations, which can be arrayed along an interpretive or constructionist continuum ranging from unconstrained interpretation or subjectivism at one extreme to highly constrained and contextualized interpretation at the other, it is reasonable to wonder what adherents of the perspective have in common. What ties them together, however loosely? The answer is found in the perspective's view of human beings as active rather than merely responsive organisms and in its contention that an understanding of much human social action requires consideration of its situated, interactive, and interpretive character. And it is also this combination of orienting principles that makes the perspective a useful sensitizing one within sociology in general, thereby accounting in part for its resilience and persistence.

See also: Action, Theories of Social; Delinquency, Sociology of; Dramaturgical Analysis: Sociological; Ethnomethodology: General; Exchange in Anthropology; Goffman, Erving (1921–82); Identity: Social; Interactionism and Personality; Macrosociology–Microsociology; Mead, George Herbert (1863–1931); Phenomenology in Sociology; Reflexivity: Method and Evidence; Self: History of the Concept; Small-group Interaction and Gender; Social Psychology; Social Psychology: Sociological; Status and Role, Social Psychology of; Symbolic Interaction: Methodology; Traditions in Sociology

Bibliography

Blumer H 1969 *Symbolic interactionism: Perspective and Method*. Prentice-Hall, Englewood Cliffs, NJ

Conrad P, Schneider J W 1980 *Deviance and Medicalization: From Badness to Sickness*. Mosby, St. Louis, MO

Denzin N K 1989 *Interpretive Interactionism*. Sage, Newbury Park, CA

Denzin N K 1992 *Symbolic Interactionism and Cultural Studies: The Politics of Interpretation*. Blackwell, Oxford, UK

Ellis C 1991 Sociological introspection and emotional experience. *Symbolic Interaction* **14**: 23–50

Fine G A 1993 The sad demise, mysterious disappearance, and glorious triumph of symbolic interactionism. *Annual Review of Sociology* **19**: 61–87

Fine G A (ed.) 1995 *A Second Chicago School? The Development of a Postwar American Sociology*. University of Chicago Press, Chicago, IL

Goffman E 1963 *Stigma: Notes on the Management of Spoiled Identity*. Prentice-Hall, Englewood Cliffs, NJ

Goffman E 1983 The interaction order. *American Sociological Review* **48**: 1–17

Hochschild A R 1979 Emotion work, feeling rules, and social structure. *American Journal of Sociology* **85**: 551–5

House J 1977 The three faces of social psychology. *Sociometry* **40**: 161–77

Joas H 1985 *G. H. Mead: A Contemporary Re-examination of His Thought*. MIT Press, Cambridge, MA

Lofland L H 1989 Social life in the public realm: A review. *Journal of Contemporary Ethnography* **17**: 453–82

McCall G J, Simmons J S 1978 *Identities and Interaction*, rev. edn. Free Press, New York

Maines D 1977 Social organization and social structure in symbolic interactionist thought. *Annual Review of Sociology* **3**: 235–59

Miller D L 1973 *George Herbert Mead: Self, Language, and the World*. University of Texas Press, Austin, TX

Miller G, Holstein J A (eds.) 1993 *Constructionist Controversies: Issues in Social Problems Theory*. Aldine De Gruyter, New York

Reynolds L T 1993 *Interactionism: Exposition and Critique*, 3rd edn. General Hall, Dix Hills, NY

Rock P 1979 *The Making of Symbolic Interactionism*. Macmillan, London

Smith-Lovin L, Heise D (eds.) 1988 *Analyzing Social Interaction: Research Advances in Affect Control Theory*. Gordon and Breach, New York

Snow D A, Anderson L 1987 Identity work among the homeless: The verbal construction and avowal of personal identities. *American Journal of Sociology* **92**: 1336–71
Strauss A L 1959 *Mirrors and Masks*. Free Press, Glencoe, IL
Stryker S 1980 *Symbolic Interactionism: A Structural Version*. Benjamin/Cummings, Menlo Park, CA
Turner R H 1962 Role-Taking: Process versus Conformity. In: Rose A M (ed.) *Human Behavior and Social Process*. Houghton-Mifflin, Boston
Turner R H 1978 The role and the person. *American Journal of Sociology* **84**: 1–23

D. A. Snow

Interdependence: Political

As used by political scientists, 'interdependence' has two meanings. First, a group of actors is considered interdependent if the behavior of each one is contingent on the behavior of the others. Second, a group of actors is considered interdependent if it would be costly for them to forego their relationship. Defined in both ways, interdependence is a concept that has been used quite widely to analyze various aspects of international relations. Particular attention, however, has been paid to economic interdependence among countries; and in this article, the definition, measurement, causes, and consequences of international economic interdependence are reviewed.

1. Interdependence as Sensitivity and as Vulnerability

In the field of international relations, it is common to distinguish between sensitivity interdependence—which corresponds to the first meaning described above—and vulnerability interdependence—which corresponds to the second meaning (Baldwin 1980). The key difference between sensitivity and vulnerability interdependence hinges on the costs that countries would bear if relations between them were disrupted. If, for example, changes in economic conditions in one country have a speedy and marked influence on another country, then there is a high degree of sensitivity interdependence between them. If these countries also would find it quite costly to sever their relationship, then the extent of vulnerability interdependence between them is high as well. But sensitivity and vulnerability interdependence often do not move in lockstep. Frequently, there are extensive economic connections between states, yielding a high level of sensitivity interdependence. However, they would not find it especially costly to replace these connections—either by expanding economic interactions with third parties or making domestic economic adjustments—so the level of vulnerability interdependence is relatively low.

2. The Measurement of Interdependence

Interdependence has been measured in various ways, most of which are linked to the volume of interstate transactions. Among these transactions are tourism and travel, mail, and immigration. Some of the earliest work on interdependence, for example, measured its growth by assessing whether domestic transactions and communications were rising at a more rapid pace than foreign transactions and communications (Deutsch and Eckstein 1961). More typically, however, the extent of interdependence is gauged by the volume of economic exchange between states.

One particularly popular measure of economic interdependence is the ratio of international trade flows to national income. Some studies have assessed the extent of interdependence throughout the international system using the ratio of total global trade to total global income. That the value of this ratio has increased dramatically since the conclusion of World War II is viewed widely as evidence that interdependence is on the rise. Studies of international relations, however, usually focus on interdependence between countries, rather than worldwide. Such analyses often measure interdependence by dividing the annual volume of trade between a pair of countries by the annual gross domestic product (GDP) of one state or the other.

Frequently it is argued that the ratio of bilateral trade to GDP is a good gauge of both sensitivity and vulnerability interdependence. This ratio does provide a useful measure of sensitivity interdependence, since it indicates the extent to which trade partners' economies are intertwined. The argument that this ratio also furnishes an accurate indication of vulnerability interdependence rests on the claim that as commerce between a pair of countries makes up a larger portion of each country's total economic output, it is increasingly costly for either partner to replace the trade conducted with its counterpart. But this claim has been the subject of heated debate, and some observers have pointed out that states that trade extensively may not find it very difficult to locate alternative markets in the event that economic relations between them deteriorate (see Baldwin 1980, Hirschman 1945/1980, Keohane and Nye 1977). As such, considerable care should be exercised in using the ratio of bilateral trade to GDP as a measure of vulnerability interdependence. Unfortunately, however, developing more precise measures of vulnerability interdependence requires data that are not available for many countries and periods of time.

3. Determinants of Interdependence

What accounts for the degree of interdependence between states? Economic and technological factors play important roles in this regard. A group of states

vested with similar factor endowments, for example, tends to realize fewer gains from economic exchange than a group of states with very different endowments. All else being equal, both the extent of and the cost of disrupting economic exchange (and, hence, the degree of interdependence) are likely to be lower in the former than the latter group. In addition, economic growth can contribute to heightened interdependence by spurring the demand for goods and services, including those produced abroad. Furthermore, technological improvements have increased the ability to ship goods safely and quickly, and have reduced the costs of doing so, thereby promoting international economic exchange. Such improvements have also increased the ease and speed of moving financial assets between countries, again spurring the growth of interdependence (see Cooper 1968).

However, the extent of economic interdependence also depends on political factors. Particularly important are the international institutions established to regulate economic exchange. The General Agreement on Tariffs and Trade and its successor, the World Trade Organization, have helped slash trade barriers and promote the flow of commerce among participating states. So have various preferential trading arrangements (PTAs), like the European Economic Community, that were designed to foster economic integration. Many of these institutions have contributed to rising interdependence among member-states (Jacobson 1984, Keohane and Nye 1977). In a related vein, some PTAs—including many of those formed by less developed countries during the 1960s and 1970s, and the Council for Mutual Economic Assistance—were established to reduce the economic and political dependence of members on third parties (Mansfield and Milner 1999).

Political–military relations also influence the extent of economic interdependence. Economic relations tend to be more open—and both sensitivity and vulnerability interdependence tend to be higher—between allies than between adversaries. The efficiency gains from open trade stimulate the growth of national income and rising income can be used to enhance states' political–military capacity. Countries cannot ignore these political–military implications of foreign commerce without jeopardizing their well-being; and they can address these security implications by trading more freely with their political–military allies than with other states (see Gowa 1994). Open trade, in turn, promotes sensitivity interdependence and can heighten vulnerability interdependence too.

Political–military relations affect political, as well as economic, interdependence among states. Sensitivity interdependence tends to be high between both allies and adversaries, since a state's political behavior is contingent on the behavior of both its closest friends and its foes. But whereas the level of vulnerability interdependence tends to be low between adversaries, it tends to be high between allies. Since allies typically rely on each other for military security, they face a high cost if their political relations deteriorate unless each of them can quickly and easily locate substitutes for the security furnished by the others.

4. *Implications of Interdependence*

Having discussed some factors affecting the extent of interdependence, this article concludes by addressing some of its domestic and international implications. From a domestic standpoint, heightened interdependence can erode the effectiveness of a country's economic policies and, hence, undermine its economic autonomy (see Cooper 1968). If, for example, private actors face relatively few impediments to shifting assets from one country to another and governments face substantial costs if these actors actually relocate assets, then governments are constrained in setting tax and regulatory policies that affect private actors. As such, various observers maintain that one implication of growing interdependence has been greater convergence in macroeconomic policies throughout the world.

Economic interdependence also influences international relations. In a pair of countries marked by substantial vulnerability interdependence, each country is vested with some power vis-à-vis the other, since each one has the capacity to damage the other by severing their economic relationship. Rarely, however, do such countries depend on their economic relationship to the same degree. The least dependent party is vested with a source of power vis-à-vis the more dependent party because the former can more easily bear the costs of disrupting or forgoing the relationship (Baldwin 1980, Hirschman 1945/1980).

Furthermore, it is argued widely that interdependence affects the likelihood of military conflict between states. Debated for centuries, at least three competing views can be identified on this topic. The first is that economic interdependence inhibits interstate hostilities, a claim that can be traced to biblical times, but that most frequently is associated with the nineteenth century Manchester liberals. Although there are various strands of this position, its advocates have stressed a number of core arguments. One is that by increasing contact and communication among individuals and governments, economic interdependence promotes cooperative political relations. Another is that economic exchange and conquest are substitute means of acquiring the resources needed to foster political security and economic growth. As economic interdependence rises, the incentives to obtain these resources through territorial expansion and foreign aggression decline. A third argument emphasizes that since war typically disrupts economic relations among the belligerents, heightened interdependence increases the costs associated with military conflict, thus deterring its onset.

However, the view that greater trade promotes peace has not gone unchallenged. Mercantilists and economic nationalists have long maintained that, as noted earlier, open commerce can undermine the national security of states by influencing interstate power relations. Because the gains from trade tend to be distributed asymmetrically and can be used to enhance states' political–military capacity, those countries benefiting less from a commercial relationship may find their security jeopardized. Similarly, if trading partners depend on sustaining a commercial relationship to very different degrees, trade may do little to deter belligerence on the part of the less dependent state. Even if a set of states depends heavily and uniformly on trade with each other, such dependence could contribute to commercial disputes and thereby stimulate political antagonism. Furthermore, various observers have argued that the anarchic structure of the international system compels states to limit their economic dependence. As dependence rises, countries may have reason to engage in foreign expansion to manage or reduce it.

Finally, a wide variety of studies have concluded that international economic relations have no systematic bearing on political conflict. Many such studies argue that hostilities stem largely from variations in the distribution of political–military capabilities and that power relations underlie any apparent effect of economic exchange on military antagonism. Despite enduring and heated debates about the nature and strength of the relationship between economic interdependence and political conflict, systematic analyses of these issues have been relatively scarce (for more detailed treatments of the relationship between interdependence and political conflict, see Blanchard et al. 2000, Gasiorowski 1986, Stein 1993).

See also: Alliances: Political; European Integration; International Business; International Law and Treaties; International Organization; International Relations, History of; International Relations: Theories; International Trade: Economic Integration; Regional Integration

Bibliography

Baldwin D A 1980 Power and interdependence: A conceptual analysis. *International Organization* **34**: 471–506

Blanchard J F, Mansfield E D, Ripsman N M 2000 *Power and the Purse: Economic Statecraft, Interdependence, and National Security*. Cass, London

Cooper R N 1968 *The Economics of Interdependence: Economic Policy in the Atlantic Community*. McGraw-Hill, New York

Deutsch K W, Eckstein A 1961 National industrialization and the declining share of the international economic sector, 1890–1959. *World Politics* **13**: 267–99

Gasiorowski M J 1986 Economic interdependence and international conflict: Some cross-national evidence. *International Studies Quarterly* **30**: 23–38

Gowa J 1994 *Allies, Adversaries, and International Trade*. Princeton University Press, Princeton, NJ

Hirschman A O 1945/1980 *National Power and the Structure of Foreign Trade*. University of California Press, Berkeley, CA

Jacobson H K 1984 *Networks of Interdependence: International Organizations and the Global Political System*, 2nd edn. Knopf, New York

Keohane R O, Nye J S 1977 *Power and Interdependence: World Politics in Transition*. Little, Brown, Boston

Mansfield E D, Milner H V 1999 The new wave of regionalism. *International Organization* **53**: 589–627

Stein A A 1993 Governments, economic interdependence, and international cooperation. In: Tetlock P E, Husbands J L, Jervis R, Stern P C, Tilly C (eds.) *Behavior, Society, and Nuclear War*. Oxford University Press, New York, Vol. 3, pp. 241–324

E. D. Mansfield

Interest Groups

Among the requisites for democracy on a sustained basis are not only free and fair elections but also the rule of law, respect for dissent and freedom of expression, and the legitimation of the organized representation of citizen interests—often in opposition to those currently in power—through parties and interest groups. Interest groups, those myriad organizations—ranging from business and professional associations to labor unions to environmental and consumer groups to groups concerned about issues as diverse as abortion and nuclear power—that link the individual to public institutions and seek to influence public policy, are a feature of every functioning democracy. Although they may be present in nondemocratic systems as well, it is not an exaggeration to claim that democracy on a national scale is impossible without interest groups.

1. Problems of Definition

Differences of opinion arise in any discussion of how to categorize the various collectivities that mediate between the individual and the state in a democracy. Many of the institutions often described as interest groups—for example, corporations, foundations, or public bodies such as regional governments—are not groups at all, in that they do not have members in the ordinary sense. For this reason, some analysts prefer the term 'organized interests' to 'interest groups.' Interest groups have much in common with other social aggregates that represent citizen preferences to policy makers in a democracy. In fact, organized interests often have connections, formal or informal, with political parties or social movements. However, unlike political parties, they do not nominate candidates to campaign under their name for the purpose

of running the government. They contrast with social movements in their greater degree of formal organization.

2. Interest Groups and the Policy Process

Wherever they are, interest groups serve as intermediaries in a two-way process of communication linking a citizen constituency with policy makers. All interest groups perform certain functions: they provide information to public officials to assist in designing policy; they seek to persuade policy makers to pursue a preferred course of action, an activity that may be difficult to distinguish in practice from the attempt to inform; and they communicate with members—keeping them apprised as to what the government is doing, educating them about the political process, and cultivating support.

Within the context of these widely shared functions, organized interests vary substantially—in ways that reflect the degree of permeability of the political system to new interests, and the nature and relative importance of formal political institutions—with respect to the difficulty they face in gaining access, the techniques they use to influence policy, and the targets of their activities. Newly emergent groups have much more difficulty penetrating interest group politics in Israel, which has a strong, centralized state, than in the USA, with its tradition of voluntarism and a public domain that, because it is divided by federalism and separation of powers, provides many points of access. Within and across polities, there is tremendous variety in the kinds of techniques groups use to realize their objectives. France, for example, has a tradition of direct action by groups not present in, say, the UK. Within any single country, organized interests differ in the extent to which they function solely as insiders utilizing traditional, low-profile means of achieving political influence or operate as outsiders, adopting more public tactics of mobilization and protest—with unions and environmental groups more likely to engage in protest activities than, say, business associations.

In choosing where to direct their energies, organized interests inevitably gravitate to power. Hence, the primary focus of interest group activity in Japan is the bureaucracy, while in Italy it is parliament. In a system as fragmented as the US political system, interest groups often have choices, and may seek to locate a political controversy in the institutional arena in which they would have the greatest probability of success.

Political systems differ not only in terms of what interest groups do and where they channel their activities but also in terms of the structure of relations between interest groups and government. Some political scientists have suggested that there are two basic patterns of organized interest interaction with the state: neocorporatist and pluralist. According to the neocorporatist model, approximated most closely in Sweden and Austria, mechanisms providing for equal representation of individuals on a geographical basis are supplemented by mechanisms providing for functional representation of organized interests—ordinarily by a relatively limited number of peak associations. Such organizations are recognized and licensed by the state and attain a regularized role in policy making through delegations of administrative power or participation on public councils or committees. Important government decisions are made only after consultation with major economic interests, most notably workers, employers, and farmers.

In contrast, according to the pluralist model, interest groups are much more numerous—less highly aggregated and organized around multiple bases of political conflict—and much more autonomous. Pluralist interest-group politics, approximated most closely by the USA, is relatively permeable to the entry of new groups, and groups arise and get involved in politics at their own initiative, rather than at the behest of the state. In a politics of pluralism, interest groups have greater freedom in choosing which political battles to fight in which political arenas. Typically, those battles involve shifting, issue-specific coalitions taking sides on issues of narrower, more limited import.

It is essential to recognize that these alternative models, which have generated considerable scholarly discussion and some controversy, are ideal types. No polity actually conforms to either model, and many countries—for example, France where a strong, centralized state has meant comparative weakness for interest groups, or Japan where regular consultation between business and government more or less excludes labor—evidence a pattern that is neither pluralist nor corporatist. Furthermore, in most countries no single pattern obtains for the making of policy in all issue areas: the formulation of foreign policy, for example, is less likely to entail regularized bargaining with interest groups than the making of agricultural policy.

3. Interest Group Systems: Who is Represented?

Reflecting national differences in their political traditions and political culture, in the number and depth of politically relevant social cleavages, and in the number, strength, and ideological distinctiveness of the parties, there are also, across democracies, enormous differences in the nature of interest systems, the collective set of organized interests active in politics. Polities vary widely in both the number of groups and the proportion of citizens who are members of any group at all; in the Scandinavian countries, where most employees belong to some kind of union or professional association, membership is much higher

than in, for example, Italy or Greece. There are also differences in the nature and number of the axes of political cleavage encompassed in organized interest politics; the multiplicity of kinds of interests represented in the interest groups of a large and diverse country such as the USA contrasts with that in a relatively homogeneous country such as Norway.

In addition, interest group systems vary substantially in the degree to which interest representation is highly aggregated or fragmented functionally or geographically and in the extent to which interest groups have a monopoly on the organization of a particular constituency. For example, in contrast to the centralized interest group systems of Austria and Sweden are those of the Netherlands, where business associations are organized according to religious differences, or Belgium, where business associations are divided along regional lines but labor is pillarized into socialist, liberal, and Catholic national federations. Furthermore, there are marked differences in terms of the relative strength of organizations representing different kinds of interests—in particular, in the vigor of the opposition provided to traditionally powerful business organizations by organized labor and citizens' groups concerned about consumer and environmental issues.

There are, however, certain regularities in the kinds of interests represented by interest groups. It is difficult to think of a democracy with active interest group politics where producer interests are not well represented—whether by peak employer groups, sectorally based trade associations, or corporations acting individually. In contrast, with the exception of pensioners, the interests of those who benefit directly from welfare-state policies—for example, students or recipients of public assistance—rarely receive vigorous advocacy by interest groups although their causes may be championed by political parties.

4. *Interest Groups and Parties*

That the political parties in a democracy might represent interests not being championed by organized interests—and vice versa—makes clear that an overall assessment of the way in which citizen interests are represented in any particular democracy requires that its interest groups be considered in the context of its parties. Common wisdom holds that strong parties go hand in hand with weak organized interests and, conversely, weak parties with strong organized interests. In the USA, a plurality of groups broadens substantially the range of interests represented in politics by two relatively weak and ideologically moderate parties. However, strong parties do not necessarily preclude strong interest groups. Sweden has strong political parties, and interest groups are guaranteed a consultative role in corporatist policy making in the system of state commissions. And France has traditionally been characterized as having both weak parties and weak interest groups in the face of a strong state.

Democracies also differ in the extent to which interest groups have strong links to like-minded political parties. For example, at least until recently, UK trade unions have had a much closer relationship to the Labour Party than their US counterparts do to the Democrats.

5. *The Evolution of Interest Group Politics*

In spite of the substantial differences across democracies, there have been, over the past few decades, two related developments affecting interest group politics on a cross-national basis that suggest the possibility of convergence among democracies. First is the emergence of interest representation at the European Union (EU). Although group activity in Brussels often arises within individual EU nations, many of the groups involved are either individual corporations or European-level, rather than national, organizations.

The second important trend is an advocacy explosion in democracies everywhere with—in a fragmented, pluralist fashion—more concerns being represented by more groups, and a greater reliance on professional lobbyists. This fluid style of interest advocacy is especially characteristic of the EU but has appeared in the national capitals as well. With respect to the kinds of interests represented, a growing number of citizens' groups concerned about an array of issues including nuclear proliferation, the environment, immigration, and the rights of gays, women, and ethnic minorities supplement traditional economic groups representing business, farmers, and workers. How these developments will alter interest group politics is not yet clear. Nonetheless, private associational life will surely remain essential to the functioning of democracies around the world.

See also: Corporatism; Democracy; Issue Networks: Iron Triangles, Subgovernments, and Policy Communities; Lobbying; Pluralism; Political Representation

Bibliography

Baumgartner F R, Leech B L 1998 *Basic Interests: The Importance of Groups in Politics and Political Science*. Princeton University Press, Princeton, NJ

Bentley A F 1908 *The Process of Government*. University of Chicago Press, Chicago

Crouch C 1993 *Industrial Relations and European State Traditions*. Clarendon Press, Oxford, UK

Dalton R J 1994 *The Green Rainbow: Environmental Groups in Western Europe*. Yale University Press, New Haven, CT

Heinz J P, Laumann E O, Nelson R L, Salisbury R H 1993 *The Hollow Core: Private Interests in National Policy Making*. Harvard University Press, Cambridge, MA
Mazey S, Richardson J 1993 *Lobbying in the European Community*. Oxford University Press, Oxford, UK
Olson M Jr 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
Schlozman K L, Tierney J T 1986 *Organized Interests and American Democracy*. Harper & Row, New York
Schmitter P C, Lehmbruch G (eds.) 1979 *Trends toward Corporatist Intermediation*. Sage, Beverly Hills, CA
Thomas C C (ed.) 1993 *First World Interest Groups: A Comparative Perspective*. Greenwood Press, Westport, CT
Truman D B 1951 *The Governmental Process*. A Knopf, New York
Walker J L 1991 *Interest Groups in America*. University of Michigan Press, Ann Arbor, MI

K. L. Schlozman

Interest Groups, History of

Organized interests and groups, their factions or parties fighting for influence and power in a given polity or beyond, go back a long way in history. In modern political theory the notion of 'interests' was first narrowed down by enlightened philosophers from Machiavelli to Montesquieu to material and rational interests, as opposed to and counteracting the 'passions' (Hirschman 1977). James Madison and the Federalists have considered both the passions and the different material interests, particularly of 'those who hold, and those who are without property,' to be the principal sources of factions in a society which could only be contained and controlled by building adequate institutions with a sufficient number of checks and balances (Hamilton et al. 1961, 'Federalist 10'). The concepts of interest groups, although under different labels, and of interest politics in a more specific sense, began to proliferate in the Western world from the beginning of the nineteenth century when, following the American and French revolutions, the world of political parties and factions started to dissociate from the universe of interests at large; and when the unequal progress of industrialization in different parts of Europe induced individuals, firms, and groups to articulate their interests in determined policies in order to improve economic or institutional prerequisites, or cope with the social consequences of industrialization.

So the modern interest group might be defined as a voluntary association of individuals, firms, or smaller groups uniting in order to defend or fight for a common interest, with the intention of influencing and intervening in the political process, but without ambition to form a political party (exceptions notwithstanding). A number of similar terms correspond to the same definition: pressure groups, organized interests, lobbies, power groups, the German *Verbände*, the French *syndicats*, US 'chambers,' and many 'institutes' or 'associations.' (Functional equivalents in the non-Western world are here excluded.) As the groups' activities are usually directed toward the political arena asking for specific legislation, allocation of benefits, and state intervention promoting their interests, the rise of state interventionism and of the modern social and welfare state since the last quarter of the nineteenth century has been closely intertwined with a secular proliferation of interest groups of different types which, on the one hand, usually reflected the characteristics of the (more or less democratic) political systems in which they originated. On the other hand, the groups' activities have often significantly contributed to modifying and modernizing the political systems, and to establishing new rules and patterns of interaction. Their recognition and legitimation have not only been products of their performance, but have also been dependent on the general theoretical and ideological orientations of the political consensus within a given polity, which could vary from authoritarian or Rousseauist views of single or 'particular' interests as being categorically detrimental to the common good to more pluralistic Anglo-Saxon approaches of accepting different interest articulations as expressions of a vital civil society (Tocqueville) and of taking factions for granted as vehicles of political agency (cf. Fraenkel 1964). In order to overcome some of the implied shortcomings of legitimation, interest groups have usually tended to develop ideological mechanisms of their own to demonstrate that their particular interests serve the common good or coincide with the interests of the community.

1. *Types*

Interest groups have developed into a wide variety of different types. They could organize individuals, firms, associations, or public corporations (cities, states, provinces, legislatures) at a local, translocal, national, or transnational level and develop different characteristics. With regard to the contents and quality of the interests involved it could be said that, with some simplification and not accounting for the various mixes and marginal or exceptional cases, most of the groups may belong to one of the following eight types: (a) professional associations; (b) groups of business, commerce, and industry; (c) labor unions; (d) agricultural organizations; (e) single interest groups; (f) ideological interest groups; (g) public interest groups; and (h) welfare associations.

(a) The most 'classical' case seems to be the one of the professional associations which have taken over many of the functions of the premodern guilds, estates, brotherhoods, or clubs, particularly in regulating the

requirements and mechanisms of cooptation and access to a number of key professions, be it as a privilege of professional self-government as in the USA (American Bar Association, American Medical Association) or as a task delegated by the state as in most of the German *Kammern* (of lawyers, doctors, artisans). The various medical associations have become particularly differentiated with the rise of compulsory health insurance or comprehensive national health services (cf. the German *Kassenärztliche Vereinigungen*). The organization of less professionalized groups such as teachers or employees in trade and commerce has often followed the lines of professional associations, if they did not unionize in the labor ranks as in most Latin countries.

(b), (c) The two predominant camps of interest groups of the industrial era, the associations of industry and the labor unions, have typologically developed out of earlier commercial guilds or professional associations of craftspeople and workers. The labor unions in industry have usually been able to acquire more strength and influence since they began to overcome their traditional fragmentation along the lines of professions and crafts, and started to form broader industrial unions—in Germany since the 1890s, in the UK after World War II, in most other industrial countries during the 1920s and 1930s. In industry and labor it also made a great difference whether there evolved a coordinating peak association or confederation, or whether there were two, as in German industry before 1919 or in Spanish labor before 1939 and after 1975, or a more fragmented variety, as in the USA. In case labor relations became institutionalized, the industrial sector and its associations occasionally created separate employers' associations with the specific objective of coordinating negotiations with the unions. Banking, insurance, and small business interests in most countries have also organized separately.

(d) Most agricultural interest groups, which came into existence from the last third of the nineteenth century on, besides voicing the interests of 'agriculture' against industry and the industrial world, have reflected the structures and divisions of the various crop sectors or regions in their nation's agricultural economies, with the interests of larger producers usually dominating those of the smaller ones, a high influence of specialized production interests notwithstanding. In less developed countries or regions with a greater number of dependent small peasants and landless laborers, agricultural interests have also been organized in labor unions and social movements. In some countries (such as France) and in determined historical periods and constellations in others (such as Italy, Spain, the USA during the Populist Era, Germany and the USA around 1930 during the Great Depression) the mode of agricultural interest articulation could oscillate between interest-group politics and social protest.

Two other 'classical' types are constituted by groups dedicated to voicing (e) a single nonprofessional, nonproducer, and nonpublic interest—for example, the National Rifle Association (NRA) in the USA, groups advocating or fighting particular forms of exclusive or commercialized sports or entertainment, automobile associations; and (f) by more ideologically inspired groups such as those advocating imperialism, temperance, or prohibition (Anti-Saloon League), or fighting abortion or the death penalty. A number of single interest groups claim to be contributing to the public good by improving people's health—for example, most large sports associations—or by promoting human progress and empowerment of the underprivileged—for example, many organizations of the minorities and of women in the USA (such as the National Association for the Advancement of Colored People, the Leadership Conference on Civil Rights, the National Organization for Women).

(g) Some of these groups might also be considered to belong, at least partly, to the type of public interest groups, the most common issues of which in the USA—following more than a century of honorable precursors such the Anti-Slavery Society (1833) and the American Civil Liberties Union (1920)—since the 1970s have been environmental and consumer interests (Common Cause, Public Citizen, Consumer Federation of America; cf. McFarland 1984). In Germany and some other Central and West European countries antinuclear and pacifist issues would have to be added. At the transnational level, groups such as Greenpeace or Amnesty International would belong to this category.

(h) The heterogeneous groups which might be labeled as welfare associations often combine professional interests, particularly those of social workers, with those of public entities, mostly cities and counties, and the single and ideological interests of churches, religious communities, and other private agents in the welfare arena, which in parts might also be considered as serving the public interest. Many organized themselves early on into coordinating central or national associations of associations or corporations, such as the National Conference of Charities and Correction (1874/84) in the USA (professional interests prevailing) or the *Deutsche Verein für Armenpflege und Wohltätigkeit* (1881) in Germany (corporate interests prevailing), long before sectoral peak associations were formed.

The various forms of intergovernmental lobbying will not be considered here because they are about influence and allocation of benefits within the public sphere only. Hence, neither the National Governors' Association or the US Conference of Mayors, nor the German *Kultusministerkonferenz* or the *Deutsche Städtetag* will be considered as interest groups. The same applies to the often well-organized lobbies of foreign governments in Washington, DC, or to the representation of state or regional interests vis-à-vis

the European Commission in Brussels, even if they seek to increase their influence by behaving like private interests. Likewise, the numerous Political Action Committees (PACs) of American interest groups, which have basically been founded for legal reasons with the specific purpose of influencing legislation, might not be considered as being a separate type of interest group, because the interests voiced by them are the same as those of their creators (cf. Cigler and Loomis 1995, Walker 1991).

Alternative typologies might differentiate between status quo defenders and agents of change, or according to whether the groups rely more on money or on numbers, on machines and bureaucrats, or on mobilization and agitation, whether they wish to influence legislators *en petit comité* or think they need a broad public forum (which also depends on the political system), or whether or not they need a high level of identification or have to provide a variety of services in order to secure their members' loyalty (such as automobile associations). Among the most basic and influential differences are those between big and small interest groups, and between big and small constituencies. Smaller constituencies (such as steel or farming interests) are usually much better organized and represented than broader constituencies (such as consumers), which tend to fall victim to the mechanisms of the 'free rider' effect (Olson 1965, Downs 1957).

2. *Constellations and Trajectories*

The shapes of interest groups and of the evolving national systems of interest intermediation have been significantly influenced by particular economic, social, and political constellations, by institutional sets, and by the distinct paths and trajectories of modernization in different countries and societies. Of particular importance here are the peculiar relations and mixes of developments in the three most important dimensions of the modernization processes, which, simplified, might be labeled as bureaucratization, industrialization, and democratization. In the UK industrialization has triggered democratization, and bureaucratization came late, basically in order to cope with the social consequences of industrialization. In the USA, which followed a similar pattern, the elements of democratization were stronger from the beginning, although moderated by federalism, and more checks and balances. On the European continent bureaucratization has been the legacy of early state building under absolutism, but here the French Revolution has made the great difference. In France the bureaucratic factors combined with the different lines of democratization (notable representation or Bonapartism, with its peculiar 'deliberative' mechanisms), whereas in most of the German states democratization was postponed, and the characteristic patterns of modernization were dominated, down to the mid-twentieth century, by factors combined from bureaucratization and industrialization.

These different constellations have influenced the emergence and performance of interest groups with regard to their social bases and opportunity structures, their strategies and addressees, organization, modes of operation, and achievements, and their relationship with governments, parliaments, and political parties. It is important to note that of the USA and Germany—two countries which have shaped the Gestalt of modern interest groups for the twentieth century more than any others since the breakthrough of organized mass politics between 1890 and 1910—one was not democratic. In Germany the government was not accountable to parliament, and the decades of Bismarckian authoritarianism had weakened the political parties and given the (supposedly less political) interest groups better conditions to flourish than anywhere else.

3. *Different Patterns of Organization*

In Germany the interest groups developed typologically in three waves:

(a) Most of the associations of the first wave, some of which dated back to the first decades of the nineteenth century, were by 1900 closely cooperating with the state bureaucracy, and had become part of privileged semiautonomous corporations, or 'chambers,' of agriculture, commerce, and the trades, which could no longer oppose government policies.

(b) The second wave was initiated by the protectionist groups of heavy industry (*Zentralverband Deutscher Industrieller*, 1876) and, though to a lesser degree, by heavy agriculture, in response to the great deflation of the 1870s whose principal addressees were bureaucrats and legislators. The same characteristics of interest representation vis-à-vis parliament and the executive, but not much mobilization, can be found in most groups of business and industry which were emerging during the same period in the industrial countries—usually divided into protectionists and free traders. In the USA from the 1880s on, a two-tiered system was emerging, with, at its peak, national associations such as the National Association of Manufacturers (1885) and the US Chamber of Commerce, and thousands of influential trade associations to represent the various branches of production. In US agriculture, however, populist protest preceded organized lobbying (see below).

(c) The groups of the third wave in Germany went beyond the previously established patterns of interest representation in that they started massive and permanent propaganda campaigns, and began to influence elections, political parties, and individual politicians. They entered the stage at a time, from the 1890s onward, when the progress of industrialization induced by new technologies combined with

the extension of the political mass market, modernization of party organization, the proliferation of social Darwinism, and imperialistic rhetoric to create an ambience favoring a new type of vigorous, aggressive, and well-organized interest group working the public. Among these groups were the protectionist and interventionist Farmers' League (*Bund der Landwirte*, 1893), various petty bourgeois groups of the 'old Mittelstand' (artisans, shopkeepers), and a number of imperialistic propaganda associations such as the Pan German and the Navy League. These groups were at the core of the politics of the New Right against liberalism and social democracy. They contributed much to the modernization and radicalization of conservative groups, which now turned ultranationalistic and racist, and to the irreconcilable right/left polarization of the social and political forces active at the end of the monarchy that was to continue during the Weimar Republic, which it helped to undermine (Puhle 1975, Ullmann 1988).

One of the models the German interest groups of the third wave decided to follow was that of 'social democracy,' meaning the party and the labor unions, which soon after their emergence from clandestinity under Bismarck's Anti-Socialist Law in 1890 created efficient bureaucratic machines and initiated wide-ranging activities in expanding and consolidating the movement—which went far beyond the scope of what the earlier traditional crafts unions and associations of mutual assistance had achieved since the 1850s. The 'General Commission' of the social democratic unions, from the late 1890s on, came close to what might be called a peak association of labor, despite the existence of some smaller Catholic and liberal unions. The powerful industrial unions, particularly those of the metal workers and miners, could claim to be the most modern in the world. Their position was further enhanced by the fact that industry and the new employers' associations, from around 1900 on, began to negotiate with them on a regular basis, usually at the regional level. Beside bureaucratic traditions, the modern unions contributed to the early emergence of institutionalized labor relations in Germany, and vice versa.

In the USA, both the organization of industrial unions, most visible in the Congress of Industrial Organization (CIO) and institutionalized labor relations, though much more fragmented than in Germany, were not achieved before the New Deal reforms of the mid-1930s (cf. Schönhoven 1987, Bok and Dunlop 1970). Most of the unions within and outside the American Federation of Labor (AFL, 1886) were organized for much longer along the lines of the traditional crafts, although not as long as the UK unions of the Trades Union Congress (TUC), which nevertheless managed to dominate the Labour Party for the first two thirds of the twentieth century. The AFL and the CIO were reunited in 1955. In France, Southern Europe, and in other parts of the world there has often been more competition between unions of different ideological and political 'colors,' which have close ties to their respective parties.

Many European interest groups of the third wave, particularly in agriculture, also saw a model in the American Populists. It is interesting to see, however, that the sequences of different modes of organization and political performance of the agricultural interests in Germany and the USA have somehow been inverse: In Germany before 1933 the prevailing patterns went from ordinary lobbying to protest and mobilization; in the USA it has been the other way round, some limited revivals of protest in the Great Depression notwithstanding. The Populist revolts were followed by extensive and efficient lobbying techniques, as established by the American Farm Bureau Federation (1919), now also taken up by the older National Grange (1867) and National Farmers' Union (1902). In the UK and Scandinavia, traditional sectoral lobbying has prevailed throughout. In France one could always find several different modes of operation simultaneously—lobbying in the channels of notable politics by the big producers, mostly in the north, which organized themselves in Catholic and radical groups; and protest and mobilization on the part of the smaller producers, notably the winegrowers, mostly in the south. Interest representation has hence been more fragmented, and a second tier of special product interests (*associations spécialisées*) has been influential. Since the 1950s the coordinating function of the *Fédération Nationale des Syndicats d'Exploitants Agricoles* has become stronger. Small peasants and laborers in the French South were eventually organized in socialist unions; in parts of Italy and Spain also in anarchist and anarchosyndicalist unions (cf. Hervieu and Lagrave 1992).

4. *Systems of Interest Intermediation*

The various interest groups interact with one another, and with the legislative and executive branches of the governments, in national (occasionally also regional) systems of interest intermediation, the organization of which follows distinct patterns. These patterns depend, among other things, on the respective intensities and traditions of democratic and parliamentary government, and state interventionism. They usually oscillate between the two poles of more liberal and market-oriented pluralism, and more regulated and institutionalized corporatism, be it in the authoritarian 'state-corporatist' tradition (mostly of fascist extraction) or in the democratic variant of 'societal corporatism' or 'neocorporatism.'

Modern corporatism was originally defined by Schmitter as:

> a system of interest representation in which the constituent units are organized into a limited number of singular,

> compulsory, noncompetitive, hierarchically ordered and functionally differentiated categories, recognized or licensed (if not created) by the state and granted a deliberate representational monopoly within their respective categories in exchange for observing certain controls on their selection of leaders and articulation of demands and supports. (Schmitter and Lehmbruch 1979, p. 13)

Among the essential prerequisites for the emergence of corporatist mechanisms or systems are usually counted the existence of peak associations, strong reformist labor movements, a propensity for state interventionism, and a readiness for interparty cooperation, if not consociationalism, in the political arena (cf. Puhle 1984, Lijphart 1984). In reality, however, this 'pure type' of corporatism has been as rare as its opposite, the nineteenth-century model of liberal pluralism.

The systems of interest intermediation of most developed countries seem to have evolved into one of the two most frequent 'mixed' patterns. One is an organized and more or less regulated pluralism, often with strong oligarchic tendencies, which usually begins with 'regulation by the regulated,' as in the USA during the Progressive Era. The other is a syndrome displaying corporatist mechanisms in a number of important sectors, not, however, adding up to a full-fledged corporatist system except in very few countries, such as Sweden, Austria, or The Netherlands in their 'classical' postwar periods. Occasionally the two patterns even seem to converge: The 'iron triangles,' observed in certain sectors until the late 1990s, of a pluralist pattern of interest intermediation such as the American, came close to the classical 'tripartite' arrangements between industry, labor (both ideally represented by peak associations), and the state in corporatist systems.

Corporatism as well as the quasicorporatist functions of the 'iron triangles'—since the economic crisis of the 1970s—seem to have been significantly weakened by the increased tendencies toward fragmentation and segmentation, of interests and politics, under the impact of decentralization and deregulation, loss of membership, 'new lobbyism,' and the emergence of new public interest groups. This is particularly the case in the USA, where in the late 1980s about 18,000 associations (and 7,000 lobbyists) were registered before Congress, in comparison with around 1,350 in Germany (cf. Petracca 1992).

The degree of institutionalization of the mechanisms of interest intermediation could also vary on a sliding scale of at least four different intensities. These range from almost none (lobbying in the open political market) to codified advisory or deliberative functions of interest representation—either *ad hoc*, as in parliamentary hearings, or on a permanent basis, as provided by some constitutions and in the economic, social, and regional councils of the European Union—to coordinating and regulating agencies jointly organized by the interests and the state after the model of the agencies and councils established in Germany and the USA during the First World War and revitalized by the New Deal or by the (moderately neocorporatist) *Konzertierte Aktion* of the early 1970s in Germany. A variety of corporatist arrangements in other countries would also fit into this category. The highest level of institutionalization with compulsory participation, binding decisions, and some general jurisdiction has been reached in sectoral interest representation in councils or chambers with delegated public powers, and in the different regimes of institutionalized labor relations with 'tripartite' participation, in many industrial countries. In addition, in the USA and in Europe some of the newly established legal programs for the empowerment of women or minorities, although they do not recognize interest groups, have given some of the respective groups a quasipublic monitoring function.

5. *Functions*

Together with political parties, interest groups are the core agents of structured intermediacy indispensable to the functioning of a free society and a democratic polity. They contribute essentially to the aggregation, articulation, and representation of interests, serve as vehicles for identification, mobilization, and participation, and also as transmission belts for support and legitimation. On the whole, most modern interest groups during the nineteenth and twentieth centuries have contributed to the extension and preservation of liberal democracy, and to social and political change. The existence of distinct and different organized interests has been a key prerequisite of the recognition of pluralism, for setting up channels and rules for the democratic game, and for checks and balances. Active and dynamic interest groups and nongovernmental organizations (the often idealized NGOs) have contributed to consolidating the mechanisms of civil society, and their participatory and mobilizing energies have helped to revitalize political systems in cases of crisis and stagnation.

Interest groups, however, cannot be considered to be democratic *per se* in their internal structure, or to defend and promote democracy under all circumstances. In many European countries during the first half of the twentieth century, interest groups, principally of industry, agriculture, and the mobilized, crisis-ridden middle classes, substantially contributed to the breakdown and destruction of democracies, and collaborated with the dictators (e.g., in Germany, Italy, and Spain), in particular by disputing the principle of parliamentary representation and the legitimacy of political parties. As, with regard to their primary political function, many tend to be somewhat ambiguous vis-à-vis the particular form of the political order, it is important that interest groups—like political parties—be held responsive and accountable to

the general rules of the democratic game, that their performance in the public arena be subject to adequate control, and that, wherever necessary, efforts are made to integrate the regime of interest intermediation into the texture of democratic institutions and traditions (which may be more problematic in the case of groups acting transnationally). In dictatorial regimes, in contrast, interest groups, despite their instrumental function of organizing and occasionally mobilizing the different social sectors around the regime party or movement, in many cases by virtue of their participatory impact, have also been serving as vehicles of incremental liberalization, which eventually produces elite splits, transfers loyalties from the regime, and contributes to its demise. Recent research into the transitions from authoritarian rule and the consolidation of democracy has also demonstrated that the character of interest groups and the constellations within the 'partial regime' of interest intermediation have an important function for the overall process of democratic consolidation (Schmitter 1995).

Bibliography

Bentley A F 1949 *The Process of Government: A Study of Social Pressures*. Principia Press, Bloomington, IN

Bok D C, Dunlop J T 1970 *Labor and the American Community*. Simon and Schuster, New York

Cigler A J, Loomis B A (eds.) 1995 *Interest Group Politics*. Congressional Quarterly Press, Washington, DC

Downs A 1957 *An Economic Theory of Democracy*. Harper and Row, New York

Ehrmann H W (ed.) 1958 *Interest Groups on Four Continents*. University of Pittsburgh Press, Pittsburgh, PA

Fraenkel E 1964 *Deutschland und die westlichen Demokratien*. Kohlhammer, Stuttgart, Germany

Hamilton A, Madison J, Jay J 1961 *The Federalist* (ed. Cooke J E). The World Publishing Company, Cleveland, OH

Hervieu B, Lagrave R M (eds.) 1992 *Les Syndicats agricoles en Europe*. L'Harmattan, Paris

Hirschman A O 1977 *The Passions and the Interests: Political Arguments for Capitalism before its Triumph*. Princeton University Press, Princeton, NJ

Kaiser J H 1956 *Die Repräsentation organisierter Interessen*. Duncker und Humblot, Berlin

Key V O 1964 *Politics, Parties, and Pressure Groups*. Crowell, New York

LaPalombara J 1964 *Interest Groups in Italian Politics*. Princeton University Press, Princeton, NJ

Lijphart A 1984 *Democracies. Patterns of Majoritarian and Consensus Government in Twenty-one Countries*. Yale University Press, New Haven, CT

McFarland A S 1984 *Common Cause: Lobbying in the Public Interest*. Chatham House, Chatham, NJ

Meynaud J 1955 *Nouvelles études sur les groupes de pression en France*. Colin, Paris

Morlino L (ed.) 1991 *Costruire la democrazia: Gruppi e partiti in Italia*. Il Mulino, Bologna, Italy

Olson M 1965 *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge, MA

Petracca M P (ed.) 1992 *The Politics of Interests. Interest Groups Transformed*. Westview, Boulder, CO

Puhle H J 1975 *Politische Agrarbewegungen in kapitalistischen Industriegesellschaften. Deutschland, USA und Frankreich im 20. Jahrhundert*. Vandenhoeck and Ruprecht, Göttingen, Germany

Puhle H J 1984 Historische Konzepte des entwickelten Industriekapitalismus. 'Organisierter Kapitalismus' und 'Korporatismus.' *Geschichte und Gesellschaft* **10**: 165–84

Schmitter P C 1995 Organized interests and democratic consolidation in southern Europe. In: Gunther R, Diamandouros P N, Puhle H J (eds.) *The Politics of Democratic Consolidation. Southern Europe in Comparative Perspective*. The Johns Hopkins University Press, Baltimore, MD, pp. 284–314

Schmitter P C, Lehmbruch G (eds.) 1979 *Trends Toward Corporatist Intermediation*. Sage, Beverly Hills, CA

Schönhoven K 1987 *Die deutschen Gewerkschaften*. Suhrkamp, Frankfurt, Germany

Streeck W (ed.) 1994 *Staat und Verbände*. Sonderheft 25, Politische Vierteljahresschrift. Westdeutscher Verlag, Opladen, Germany

Truman D B 1962 *The Governmental Process: Political Interests and Public Opinion*. Knopf, New York

Ullmann H P 1988 *Interessenverbände in Deutschland*. Suhrkamp, Frankfurt, Germany

Von Beyme K 1980 *Interessengruppen in der Demokratie*, 5th edn. Piper, Munich

Walker J L 1991 *Interest Groups in America*. University of Michigan Press, Ann Arbor, MI

Wilson F L 1987 *Interest Group Politics in France*. Cambridge University Press, Cambridge, UK

H.-J. Puhle

Interest: History of the Concept

While modern social science is inconceivable without some notion of interest, the actual concept has a long and complicated history of shifting meanings. In this relatively brief overview, these changes will be highlighted by following the main stages of this conceptual development.

As a central notion for the understanding of human behavior, the concept of interest emerged in early modern political theory. It was part of the sceptical view of human affairs which informed the secular approach to politics and government which arose anew in the Renaissance. In the course of the seventeenth century, the concept entered a variety of other discourses as well, and gained a strategic importance by becoming linked to a theory of civic exchange. This view was subsequently elaborated by political economists and utilitarian philosophers. From the early nineteenth century onwards the conceptual changes have occurred mainly in debates accompanying the formation and the development of academic disciplines. Interest became embedded primarily in economic theories, which have been emulated as well as vividly contested in other disciplines.

1. The Politics of Interest

The word interest, derived from the Latin *inter esse*, originally referred to procedures for compensation in Roman law. The expression *id quod interest* could be applied to a variety of claims in this respect. The meaning of taking rent on loans was directly related to the legal notion of compensation. Interest in this specific sense came into use in many European languages during the fifteenth century, commonly as a euphemism for usury (Fuchs 1976, Hirschman 1992).

The more general meaning of the term emerged in the beginning of the sixteenth century. Interest then referred to the more general sense of advantage and to the human propensity for seeking benefits. It was part of a sceptical anthropology and a nonmetaphysical approach to statecraft, appearing systematically for the first time in the work of Francesco Guicciardini. Guicciardini, one of the most perceptive political theorists of Renaissance Italy, frequently observes in his *Ricordi* (1512/30) that 'self-interest' prevails in nearly all human beings. This applies especially to successful leaders and others who do well in life: they always have their 'own interest in mind and measure all their actions accordingly.' Guicciardini adds, however, that 'true interest' does not necessarily reside in pecuniary advantage: it pertains, more often, to honor and to the art of knowing how to keep a good reputation.

The general meaning which Guicciardini gives to the term allows him to use it for matters of government as well. Where particular interests are in conflict with the 'public interest,' Guicciardini advocates the primacy of the latter. The conceptions of Guicciardini and his contemporary Machiavelli were developed more fully in the reason-of-state literature, which flourished in Europe around 1600. For one of the leading theorists, Giovanni Botero, 'reason of state' meant 'reason of interest,' and state interest had to be the supreme rule of conduct for princes and statesmen. In this view religious and constitutional matters were to be treated as merely instrumental issues. For the ruler, interest of state is the only legitimate principle of action, and this interest is defined in opposition to both the interests of other states and the passions of the ruler. Control over people demands self-control: passions and other disorderly appetites need to be replaced by a rational use of state interest.

Interests thus came to be seen as the principal motive of human behavior and as the only realistic rule of political conduct. From a predominantly critical concept, directed against ecclesiastic and humanistic virtues, it had gained a more positive meaning. While preserving its amoral connotations, interests appeared to be a more stable and more reliable motive than the passions (Hirschman 1977). The French Huguenot leader Henri de Rohan expressed this view in his influential *De l'Intérêt des Princes et des États de la Chrétienté* (1638) by stating that 'princes order their people around and interest orders princes around.' The prince can be deceived, his council can be corrupt, only 'interest does not lie.'

It was through the translation of Rohan's work that the term 'interest' became established in the political vocabulary in England. And it was from the political scene that the term was then transferred to the market place and came to be applied to private behavior (Gunn 1969).

In the early modern political literature interest was not merely an analytical term. Interests were indeed perceived as the predominant motive of human behavior, but political advisors stressed that interests had to be observed in a rational and calculating way. This was not because human beings were considered to be rational decision makers, but because of the strategic advantages of rational calculation. The prudential weighing of costs and benefits was a behavioral norm for political elites rather than a form of motivational reductionism (Holmes 1995).

2. Contours and Consequences of Exchange

From the political literature, written by counselors and diplomats, the notion of interest entered other intellectual genres in the course of the seventeenth century (Heilbron 1998, Lazzeri and Reynié 1998). Among these other discourses, natural law and moral philosophy stand out; in both cases interest-driven behavior came to be related to new conceptions of politics and society.

In modern natural law interest was part of a foundational argument. The anthropology of the sceptics had given a seemingly universal role to self-interest, and more in particular to self-preservation. Natural law theorists treated this factual observation as a natural right. On the basis of the fundamental right to assure one's self-preservation, they built a system of moral and political obligation which was intended to overcome the relativism of the sceptics.

For natural law theorists such as Thomas Hobbes and Samuel Pufendorf, people are motivated primarily by self-love. But the consequences of their view differed fundamentally. For Hobbes a strong state was the only alternative for permanent and violent conflict between selfish indivuals. Pufendorf constructed more of a historical argument. Human societies arise out of individual needs. Upon this assumption, he proposed a scheme of societal development, culminating in a form of sociability proper to 'commercial' society. Commerce at the time referred to exchange or traffic and not specifically to trade. The idea of a 'commercial' society designated a relatively pacified order within which various realms of exchange had been established. Contrary to Hobbes, Pufendorf thus envisioned a form of sociability based essentially on the ties of self-interested exchange (Hont 1987).

The very idea of a social order emerging out of the self-interested acts of individuals is found more ex-

plicitly in seventeenth-century French moral philosophy. In a number of remarkable essays of the 1670s, collected in his *Essais de morale* (1671–78), the Jansenist moralist Pierre Nicole explained how human society could be well ordered and prosperous without assuming religious duties or prescribing any secular virtues.

Jansenist theology made a strict distinction between the 'city of man' and the 'city of God.' The radically separated realms were founded on two mutually exclusive drives: self-love and love of God. Nicole and some other Jansenists sought somehow to adapt their uncompromising outlook to the demands of worldly life. Especially in his moral essays, Nicole tried to define a modus vivendi for Christians who were forced to live in a corrupt world. Looking for an intermediary route between pious retreat and worldly corruption, he proposed what is probably the first consistent model of commercial society.

The Hobbesian image of a war of all against all, Nicole argued, is valid only in the very beginning of humankind. Since each human being is a threat to the other and no one wants to be a victim, people unite with other people. To affirm their union, laws are established and punishment is instituted for those who violate the laws. Fear of death is therefore the first restraint on self-love. Once open violence is excluded, human beings are forced to use artificial means to satisfy their desires. These artificial means are all forms of exchange. The result of this process is a civil society, which—given the corrupt human nature after the Fall—could not be better organized, even if true religion is banned. Human traffic alone, established and regulated by self-love, had produced this result.

From the point of view of God human society meant corruption, but from the point of view of human beings themselves, the establishment of human civility is the best possible achievement. Given the fact that very few people are prepared to withdraw from society, Nicole added that 'enlightened self-love' was the most effective policy in human affairs.

For Nicole and his fellow Jansenists, human beings could not be credited with any virtue. As in the maxims of La Rochefoucauld, they depicted what seemed virtuous (courage, love, friendship) as being a refined expression of self-interest. Since human virtue is a kind of narcissistic illusion, self-interested exchange was the most realistic alternative for the threat of civil war and a sufficient basis for an orderly society.

3. *Political Economy and Utilitarianism*

Many of the economic and the utilitarian arguments advanced in the eighteenth century were derived from these seventeenth-century writings. Mandeville's tale about private vices and public benefits, for example, originated in the work of French moralists like Nicole. It was not disinterestedness or self-denial which produced civil society, but vanity, avarice, and ambition. Adam Smith's famous phrase that it is not from the benevolence of the butcher that we expect our dinner, but from his 'regard to his own interest' was poignant but historically hardly original. The same may be said of his affirmations that 'society may subsist among different men, as among merchants, from a sense of its utility, without any mutual love or affection.'

Classical political economists like Smith transformed the argument about self-interested exchange into an economic theory of self-regulating markets. When the general argument about exchange narrowed down to a mechanism of economic markets, the political questions implied in the doctrine of interest reemerged. Utilitarianism was one of the attempts to rethink the political issues involved.

Emerging in the eighteenth century and becoming an intellectual movement with Jeremy Bentham and James Mill, the utilitarian system was built on the principle that every human being seeks pleasure and avoids pain. Human conduct is universally guided by the ideas and feelings which are associated with either one of these emotions. This behavioral principle was applied not only to individuals, but extended to the polity as a whole. Like individuals, governments should promote the amount of happiness and reduce the amount of pain. Moral arithmetic, based on the principle of the greatest amount of happiness of the greatest number, thus provided the means for assessing the utility of public institutions and initiating political reform.

The rise of political economy and utilitarian philosophy provoked several intellectual countermovements in the decades around 1800. In their opposition to the French revolution and the ensuing waves of democratic reform, conservatives appealed to the lustre of tradition to counter the cold-blooded politics of interest. Romantics similarly found ways to reaffirm the power of the passions.

4. *Disciplinary Social Science and Interdisciplinarity*

When the social sciences during the nineteenth century gradually became university disciplines, theories of interest were associated specifically with economic theories which—with the exception of Marxism—were increasingly separated from other social concerns. John Stuart Mill redefined the status of political economy by arguing that while different causes operate in society, some of the more powerful causes need to be studied separately. Political economy was legitimately restricted to the phenomena based on the desire of wealth and the law that a greater gain is preferred to a smaller. In Mill's view political economists do not deny other motives, but merely abstract

from them; self-interest is more a methodological construct than a behavioral reality.

During the 'marginalist revolution' the plea for scientific abstraction was combined with the postulate of instrumental rationality (Demeulenaere 1996, Hausman 1992). Economic behavior was modeled as rational calculation aimed at maximizing utility. This approach, which formed the heart of the neoclassical program, allowed much of the formal and technical developments that have characterized mainstream economics ever since. The notion of interest turned into a mathematical technique of maximization under constraint. With Pareto the domain of economics was redefined as the study of choice or optimization, and in Lionel Robbins's definition economists study 'human behavior as a relationship between ends and scarce means that have alternative uses.'

While this definition is, in principle, applicable to all forms of human behavior, many economists during the beginning of the twentieth century were well aware of the limits of their approach. For Vilfredo Pareto neoclassical economics had a limited validity since many aspects of human behavior do not meet the conditions of rational calculation. Many other sociologists also reacted critically to the abstract and formal nature of *homo economicus*. Emile Durkheim and François Simiand argued against the 'unrealistic abstractions' of economic theory, advocating a positive and more empirical approach which was not unlike that of institutional economics. Max Weber proposed a broader understanding of interest and rationality and integrated these in a general typology of social action. Weber reintroduced what he called 'ideal interests' besides material interests, and distinguished instrumental rationality from value rationality. In addition to affective and traditional action, instrumental rationality refers to a means–end relationship, while value rationality pertains to choice as derived from values.

While the academic division of labor between the various social sciences became more strict during many decades, it has been increasingly contested since the 1960s. Economic theories and modeling procedures expanded into many areas which were previously seen as the territory of other disciplines. In political science and sociology economic approaches have become part of the mainstream. Game theory has provided a tool for modeling problems of interactive choice; and rational choice, more generally, became the common denominator for a broad area of interdisciplinary work (Heap et al. 1992).

The progressive expansion of economics was accompanied by a movement in the opposite direction: sociologists, psychologists, organization theorists, and anthropologists have gradually invaded the domain of economics, proposing alternative accounts of core issues of economic theory. Diverse and varied as these contributions may be, they have at least demonstrated that the technical sophistication of economics depends on assumptions which can reasonably be questioned on empirical as well as on theoretical grounds.

5. *Enduring Ambiguities*

Since its early uses in Renaissance political theory, the concept of interest and its terminological derivatives (self-interest, group interest, public interest) have been part of what may be called the materialist tradition in the social sciences. But this general, realistic and hard-nosed orientation is divided over a number of fundamental issues. The development of the notion of interest uncovers at least three enduring ambiguities.

First, the term interest can be used for interpretative purposes in accounting for observable behavior; it can also be used in a normative way for defining what rational behavior is. In the former sense interest generally refers to an explanatory principle; in the latter it merely designates a logically possible and perhaps desirable course of action. The ambiguity of serving both descriptive and prescriptive functions has persisted ever since the origins of the term in political theory. Second, interest refers to material advantages as well as to political, cultural, and symbolic benefits. If the broader understanding of interest is accepted, the idea of maximization becomes problematic, since it is unclear how costs and benefits in the various domains could be calculated and compared with one another. Third, actors may pursue their interests by relying on rational calculation, but they may also do so on the basis of a process of socialization in which choice is not 'rational' in the economists' sense, but bound up with habitualization leading to the practical mastery of a social game (Bourdieu 1990). In the former case interest is part a rational choice theory, in the latter it is not.

See also: Consequentialism Including Utilitarianism; Enlightenment; Interest, Psychology of; Interests, Sociological Analysis of; Political Economy, History of; Public Interest; Rational Choice and Organization Theory; Rational Choice in Politics; Rational Choice Theory: Cultural Concerns; Rational Choice Theory in Sociology; Utilitarianism: Contemporary Applications

Bibliography

Bourdieu P 1990 *The Logic of Practice*. Polity Press, Cambridge, UK

Demeulenaere P 1996 *Homo Oeconomicus: Enquête sur la Constitution d'un Paradigme,* 1st edn. Presses Universitaires de France, Paris

Fuchs H-J 1976 Interesse. In: Ritter J, Gründer K (eds.) *Historisches Wörterbuch der Philosophie*. Wissenschaftliche Buchgesellschaft, Darmstadt, Vol. 4, pp. 480–5

Gunn J A W 1968 Interest will not lie: A seventeenth century political maxim. *Journal of the History of Ideas* **29**: 253–76

Gunn J A W 1969 *Politics and the Public Interest in the Seventeenth Century*. Routledge & Kegan Paul, London
Hausman D M 1992 *The Inexact and Separate Science of Economics*. Cambridge University Press, Cambridge, UK
Heap S H, Hollis M, Lyons B, Sugden R, Weale A 1992 *The Theory of Choice: A Critical Guide*. Blackwell, Oxford, UK
Heilbron J 1998 French moralists and the anthropology of the modern era: On the genesis of the notions of 'interest' and 'commercial society'. In: Heilbron J et al. (eds.) *The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 77–106
Hirschman A O 1977 *The Passions and the Interests: Political Arguments for Capitalism before Its Triumph*. Princeton University Press, Princeton, NJ
Hirschman A O 1992 The concept of interest: From euphemism to tautology. In: *Rival Views of Market Society*. Harvard University Press, Cambridge, MA, pp. 35–55
Holmes S 1995 *Passions and Constraint: on the Theory of Liberal Democracy*. Chicago University Press, Chicago
Hont I 1987 The language of sociability and commerce: Samuel Pufendorf and the theoretical foundations of the four-stages theory. In: Pagden A (ed.) *The Languages of Political Theory in Early-Modern Europe*. Cambridge University Press, Cambridge, UK, pp. 253–76
Lazzeri C, Reynié D (eds.) 1998 *Politique de l'Intérêt*. Annales littéraires de l'Université de Franche-Comté, Besançon

J. Heilbron

Interest, Psychology of

Interest has been conceptualized and studied as a habitual tendency, a motivational belief, a trait, a component of personality, a vocational choice, and a reader response elicited by text characteristics (Renninger et al. 1998). These various conceptualizations of interest can be grouped and defined as a psychological state and/or as an individual disposition (e.g., Krapp et al. 1992, Todt and Schreiber 1998).

Interest as a psychological state involves focused attention, increased cognitive functioning, persistence, and affective involvement. Although focusing attention and continuing cognitive engagements normally require increased effort, when interest is high these activities feel relatively effortless. Increased affect has also been associated with the experience of interest, and researchers generally consider mainly positive affect, such as enjoyment or liking, when they refer to the affective component of interest. Interest as a disposition may be central to how individuals attend, select, and persist in processing certain types of information in preference to others (Hidi 1990, 1995). It has been suggested that interest has evolutionary profit since it causes individuals to concentrate their cognitive, emotional, and motor processes on a certain part of their environment for continuous periods (Rheinberg 1998).

Common to most conceptualizations of interest is an interactive relationship between people and their environment. This feature goes back to Dewey (1913) who said that in interest 'self and world are engaged with each other in a developing situation' (p. 126). Thus, interest always refers to a person's interaction with a specific class of objects, events, or activities, and as such has been identified as having physiological, affective, and cognitive correlates. This strong focus on the specificity of the person and environment relation (as opposed to being localized in either the person, or in the task, topic, activity, etc.) distinguishes interest from other psychological constructs such as general curiosity and intrinsic motivation. Since interest develops over a history of interactions with a class of specific objects and/or activities that include both intrinsic and extrinsic aspects, it is inappropriate to describe interest as the outcome of either purely intrinsic or extrinsic factors (Hidi 2000).

1. *History*

The first theory of interest is often attributed to the German philosopher Herbart ([1806] 1965), who stated that cultivation of diversified interests was a prerequisite for learning and should be one of the primary goals of education. In the early part of the twentieth century many important thinkers were grappling with the relationship between interest and learning. In this period the many different ways in which academics looked at interest reflected the complexity of the concept. Dewey (1913) paved the way for future educational research by suggesting that interest-based learning was different from learning that lacked interest. Furthermore, Dewey argued that interest was necessary to maintain self-initiated, content-related activities that result in pleasure, satisfaction, and learning.

As the fields of education focused on concepts that could be operationalized and empirically tested, there was a noticeable decline in research devoted to the less precise concept of interest. Not until the last two decades of the twentieth century did researchers renew their focus on the concept of interest. Consequently, a renaissance of interest research occurred, with most of the work being produced in Germany and North America. In this period, researchers started to recognize that both emotional and cognitive factors can elicit interest (Kintsch 1980, Schank 1979) and to view interest as being domain and/or content-specific (e.g., Schiefele et al. 1979). The early conceptualizations were further developed into the more recent categorizations of interest.

2. *Various Categories of Interest*

Situational and individual interest are currently the two categories most commonly associated with experiencing the psychological state of interest.

2.1 Situational Interest

Situational interest is generated by particular conditions and/or objects in the environment. Such interest involves focused attention and an affective reaction that may or may not last (Hidi 1990). Although the affective reaction associated with situational interest tends to be positive, negative emotions may also be involved in experiencing interest (Hidi and Harackiewicz 2000). For example, a student who is not interested in science may watch a television show that demonstrates how sheep and cows may be cloned. The student reacts with surprise and a little concern. Her interest has been triggered and her attention has been focused. She experiences fascination mixed with some negative emotions driven by her apprehension over what might happen if humans are cloned. If the television show does not further sustain these reactions, the student will stop watching and move on to other activities. Thus, her situational interest has been linked only to that moment and proved to be short-lived. On the other hand, if the content of the show presents further images that engage her attention and she continues to watch it, her situational interest has been maintained. In this case, the student has continued to experience the psychological state of interest. She became fully engaged, and may even be riveted to the television program.

The above examples suggest that situational interest should be conceptualized as having two potential stages, one in which interest is triggered, and a subsequent stage in which interest is further maintained (Hidi and Baird 1986). Several experimental investigations have found empirical support for this two-stage model (e.g., Harackiewicz et al. 2000, Mitchell 1993).

2.2 Individual Interest

Individual interest has been described as an individual's relatively enduring predisposition to attend to certain objects, stimuli, and events, and to engage in certain activities (e.g., Krapp et al. 1992, Renninger and Wozniak 1985). Through repeated engagements over time, individuals build related knowledge structures, experience positive affects, and come to value highly the object of their individual interests. For example, a person with an individual interest in science values and seeks out opportunities to engage in scientific activities that he or she finds stimulating and enjoyable, works through difficulties that may involve negative emotions, continues to seek out opportunities to accumulate scientific knowledge, and experiences the psychological state of interest (Krapp and Fink 1992, Renninger 2000). Investigations that focused on individual interest have demonstrated that it influences learning and as such is an important component of academic motivation (e.g., Ainley 1998, Alexander et al. 1995, Prenzel 1992, Schiefele 1996).

2.3 The Relationship Between Situational and Individual Interest

Although situational interest and individual interests are distinct, they are not dichotomous phenomena, and they can be expected to influence each other's development. Situational interest, once maintained, may contribute to the development of a longer-term individual interest leading to increased knowledge, value, and positive feelings for the object of interest (Krapp 1998, Renninger 2000). For example, a student assigned a project on Freud starts reading his work only because it is required reading. As the student reads, his (situational) interest is triggered, and he becomes captivated with Freud's personality theory. He becomes excited, develops his own assumptions on how the behavior of significant others in his life could be best understood, and wants to learn more. From a cognitive point of view, a wide range of knowledge patterns have been activated. As the student continues to read about Freud's theory, makes connections, and develops new hypotheses, corresponding changes can be expected to occur across declarative, conceptual, and logical knowledge structures (Farnham-Diggory 1994). From a motivational point of view, although interest has been elicited through an externally triggered cognitive activity, it leads to continued and persistent engagement. As the activity proceeds, it is no longer externally imposed on the student, but becomes enjoyable, self-determined, and autonomous. The potential of situational interest to affect both cognitive and motivational functioning is an important consideration since environmental influences such as teachers and parents can contribute in this way to developing individual interest in students lacking such interest.

Individual interest can influence situational interest by moderating the impact of environmental factors on the psychological state of interest (Bergin 1999, Murphy and Alexander 2000, Pintrich 2000). For instance, having an individual interest in science would predispose a student to watch science television programs. If the program turns out to be a basic introduction to cloning and reiterates information the student knows well—even if the video is arresting for novices—the student will turn to something else.

2.4 Topic Interest

Another form of interest that has been investigated in the literature is topic interest. To date, the meaning ascribed to the term topic interest is ambiguous. Some researchers have considered topic interest to be a form of individual interest but others have treated it more like a form of situational interest. To demonstrate the ambiguity of the way the term has been used, consider what might happen when students are presented with a text entitled 'Cloning Sheep in the 21st Century.' For

students with a well-developed individual interest in biological science, the interest aroused by the text title (topic interest) would be closely linked to their individual interest. However, students who do not have a well-developed interest in the area might also report strong topic interest. In the latter case, this would most likely be due to situational factors such as the novelty or the uncertainty conveyed by the title. Thus, both individual and situational factors influence topic interest. Topic interest has also been investigated in the extensive literature that deals with vocational choice. However, reviewing this literature is outside the scope of this article.

3. Interest and Gender

Although parents and teachers have always been aware of gender difference in interest, only in the last decade of the twentieth century did researchers begin systematically to examine how gender influences individuals' interests. The results of a number of other investigations indicated that gender differences also play a role in the development and outcome of situational, individual, and topic interest (see Hoffmann et al. 1998, for a review of the literature) For example, boys have been found to be more interested in physics than girls, whereas interest in writing shows the opposite trend. In addition to the gender differences in interests, boys' and girls' performances are differently affected by the experience of interest. The literature suggests that it may be particularly important for boys to be interested in their school activities, since their task performance seems to be more adversely affected than that of girls by low levels of interest.

4. New Directions

4.1 Relation Between Interest and Other Motivational Concepts

Until the beginning of the twenty-first century, motivational research tended to be clustered around a few concepts such as interest, intrinsic motivation, self-efficacy, task value, and goal orientation. Whereas the considerable literature resulted in the clarification of these concepts, little understanding developed as to how they were interrelated. Many researchers argue that the time now seems ripe for investigating the interrelations among these motivational clusters (Murphy and Alexander 2000). Interest, intrinsic motivation, self-efficacy, and mastery goals seem to have similar behavioral outcomes such as focused attention, persistence, effort, and positive emotional reactions. It is incumbent upon future research to investigate and clarify how these motivational variables develop and influence each other.

4.2 Neuropsychological Aspects

It has been proposed that the psychological and physiological processes associated with interesting information have unique aspects not present in processing information without such interest (Hidi 1990). By elucidating how the brain works when engaged in interesting as opposed to uninteresting activities, current developments in neuropsychology may legitimize this claim. Thus, a multidisciplinary research agenda that coordinates the findings of the newer techniques of brain behavioral analysis with the longer tradition of educational psychological research on interest holds exciting promise for the future.

See also: Intrinsic Motivation, Psychology of; Motivation and Actions, Psychology of; Vocational Interests, Values, and Preferences, Psychology of

Bibliography

Ainley M D 1998 Interest in learning in the disposition of curiosity in secondary students: Investigating process and context. In: Hoffmann L, Krapp A, Renninger K, Baumert J (eds.) *Interest and Learning: Proceedings of the Seeon Conference on Interest and Gender*. Institut für die Pädagogik der Naturwissenschaften, Kiel, Germany, pp. 257–66

Alexander P A, Kulikowich J M, Jetton T L 1995 Interrelationship of knowledge, interest, and recall: Assessing a model of domain learning. *Journal of Educational Psychology* **87**: 559–75

Bergin D A 1999 Influences on classroom interest. *Educational Psychologist* **34**: 87–98

Dewey J 1913 *Interest and Effort in Education*. Riverside Press, Boston

Farnham-Diggory S 1994 Paradigms of knowledge and instruction. *Review of Educational Research* **64**: 463–77

Harackiewicz J M, Barron K E, Tauer J M, Carter S M, Elliot A J 2000 Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology* **92**: 316–30

Herbart J F [1806] 1964 Allgemeine Pädagogik, aus dem Zweck der Erziehung abgeleitet. In: Herbart J F (ed.) *Pädagogische Schriften*. Kupper, Dusseldorf, Germany, Vol. 2, pp. 9–155

Hidi S E 1990 Interest and its contribution as a mental resource for learning. *Review of Educational Research* **60**: 549–71

Hidi S 1995 A re-examination of the role of attention in learning from text. *Educational Psychology Review* **7**: 323–50

Hidi S 2000 An interest researcher's perspective on the effects of extrinsic and intrinsic factors on motivation. In: Sansone C, Harackiewicz J M (eds.) *Intrinsic Motivation: Controversies and New Directions*. Academic Press, San Diego, pp. 309–39

Hidi S, Baird W 1986 Interestingess—A neglected variable in discourse processing. *Cognitive Science* **10**: 179–94

Hidi S, Harackiewicz J 2000 Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*. **70**: 151–79

Hoffmann L, Krapp A, Renninger A K, Baumert J (eds.) 1998 *Interest and Learning: Proceedings of the Seeon Conference on Interest and Gender*. Institut für die Pädagogik der Naturwissenschaften, Kiel, Germany

Kintsch W 1980 Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics* **9**: 87–8
Krapp A 1999 Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education* **14**: 23–40
Krapp A, Fink B 1992 The development and function of interests during the critical transition from home to preschool. In: Renninger R A, Hidi S, Krapp A (eds.) *The Role of Interest in Learning and Development*. Erlbaum, Hillsdale, NJ, pp. 397–429
Krapp A, Hidi S, Renninger A 1992 Interest, learning and development. In: Renninger R A, Hidi S, Krapp A (eds.) *The Role of Interest in Learning and Development*. Erlbaum, Hillsdale, NJ, pp. 3–25
Mitchell M 1993 Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology* **85**: 424–36
Murphy P K, Alexander P 2000 A motivated exploration of motivation terminology. *Contemporary Educational Psychology*. **25**: 3–53
Prenzel M 1992 The selective persistence of interest. In: Renninger R A, Hidi S, Krapp A (eds.) *The Role of Interest in Learning and Development*. Erlbaum, Hillsdale, NJ, pp. 71–98
Pintrich R 2000 An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology*. **25**: 92–104
Renninger K A 2000 Individual interest and learning: Implications for practice. In: Sansone C, Harackiewicz J M (eds.) *Intrinsic Motivation: Controversies and New Directions*. Academic Press, New York, pp. 373–404
Renninger K A, Hoffmann L, Krapp A 1998 Interest and gender: Issues of development and learning. In: Hoffmann L, Krapp A, Renninger K, Baumert J (eds.) *Interest and Learning: Proceedings of the Seeon Conference on Interest and Gender*. Institut für die Pädagogik der Naturwissenschaften, Kiel, Germany, pp. 9–21
Renninger K A, Wozniak R H 1985 Effects of interests on attentional shift, recognition, and recall in young children. *Developmental Psychology* **21**(4): 624–31
Rheinberg F 1998 Theory of interest and research on motivation to learn. In: Hoffmann L, Krapp A, Renninger K, Baumert J (eds.) *Interest and Learning: Proceedings of the Seeon Conference on Interest and Gender*. Institut für die Pädagogik der Naturwissenschaften, Kiel, Germany, pp. 126–45
Schank R C 1979 Interestingness: Controlling inferences. *Artificial Intelligence* **12**: 273–97
Schiefele H, Hausser K, Schneider G 1979 'Interesse' als Ziel und Weg der Erziehung. [Interest as an aim and path in education]. *Zeitschrift für Pädagogik* **25**(1): 1–20
Schiefele U 1996 Topic interest, text representation, and quality of experience. *Contemporary Educational Psychology* **21**: 3–18
Todt E, Schreiber S 1998 Development of interests. In: Hoffmann L, Krapp A, Renninger K, Baumert J (eds.) *Interest and Learning: Proceedings of the Seeon Conference on Interest and Gender*. Institut für die Pädagogik der Naturwissenschaften, Kiel, Germany, pp. 25–40

S. Hidi

Interests, Sociological Analysis of

It is not an easy task to define the notion of interest. Throughout the tradition of sociological analysis, it is often referred to without further specification. In ordinary language, there are two general meanings of this particular idea: The first is that of an agent's special commitment toward some kind of object, or activity, like music. This meaning can be studied by sociology as far as it is concerned with tastes or cultural practices. But the second and more important meaning is the one that implies a certain selfishness. This is mainly the case with 'economic' interests, those concerned with the appropriation of scarce goods. But this idea can also be extended to noneconomical goods, like power or prestige.

1. Sociology and Economic Interests

1.1 Three Problems

Economics has been described as a science of human interests, when they are considered apart from any other motive. This definition is implicit in Adam Smith's analysis, while explicit in Pareto's general statement on the division of work inside social sciences. According to this description, what should the role of sociological analysis be regarding interests? Three major problems come to the forefront of the discussion.

The first one is that of social order. Any economic activity occurs within a social framework, made up of varied institutions, rules, and rights. There are two options here. The first one is to explain this framework in terms of interests, following the tradition of Hobbes and Hume. There is a trend, especially in contemporary sociology inspired by economics, to try to revive this kind of explanation of social norms in terms of interests. But the major part of sociological analysis has tried, on the contrary, to define a specific attitude toward social order that cannot be reduced to interests.

The second problem is that of economic attitudes themselves. In the economic tradition, these attitudes are generally considered from a nonhistorical point of view, in a constant way. For instance, in his description of economic equilibrium, Pareto makes no direct reference to any historical evolution of economic attitudes. For him there are only random distributions of different attitudes toward interest among different people (Pareto 1963). In most of the sociological tradition, this has been seen as irrelevant, since economic attitudes do vary historically. Therefore, various attempts to explain the logic of these changes can be found. Such is the aim, for instance, of Mauss's (1954) celebrated *Essai sur le don*. More generally, sociological analysis tries to relate interests to the background values that encourage or discourage them. In fact, there are two aspects to be stressed from this standpoint: The first is that the very notion of what is to be considered as an interest for an individual can vary socially. For instance, Mauss tries to show that in

archaic societies there is no real interest in personal consumption. The second level is that, any kind of interest being admitted as such, there are social values that reinforce it or restrict its expression.

Therefore, the third problem is that of the extension of the role of interests in social life. Economics mainly studies the exchange of scarce goods. But interests arise in other fields, such as political conflict, family attitudes, etc. The problem will be then to understand how far an analysis in terms of interests can go to explain social phenomena.

1.2 The Problem of Social Order

Three major sociologists have tried to describe the problem of social order as the main area, in social life, where it seems that interests alone cannot suffice to produce the norms that unify people into a society. Comte, first, in a radical attack against what appeared to him as an 'abstract' view, namely that of social contract and economic exchange, stressed the existence of a social consensus foreign to any calculation, and prior to any economic exchange. Following Comte, Durkheim built all his sociological theory on the assumption that individual interests cannot be seen as the root of social solidarity. In a Kantian fashion, he understood morality as a repressive force against individual interests. But, in opposition to Kant, he did not find the roots of such obligations in an individual rational sense of morality, but rather derived it from a specific social level that exerts coercion upon individuals. Parsons, in his attempt to build a general theory of society, continued this tradition. He stressed the existence of common shared values on which economic interests depend. The main point that these three authors made is that selfish interests cannot always lead to cooperation. These same interests can also lead sometimes to opposition among individuals. Hence, the social primacy of cooperation should not be based only on interest.

These three authors can fairly easily be grouped under the 'holist' tradition. But even in the tradition of methodological individualism, there has been a similar trend to refer interests to the values behind them that allow or restrict their development.

1.3 Economic Attitudes

Weber is famous for having, in his work on the Protestant ethic (Weber 1974), stressed the influence of religious attitudes on economic ones. He links the development of capitalism to a complex religious background. Demagicalization, universalization, and moralization of relations among people, commitment to an ascetic lifestyle, and interpretation of worldly success as a sign of eschatological salvation are the key religious elements that can contribute to explaining capitalism's rise in Western societies. These religious elements are joined to other values, namely the general importance of rationality, and its effects on other institutional contexts, like the extension of rational legal systems. The specific modern capitalist attitude cuts itself from its religious roots: It becomes an acquisitive trend in a society oriented towards constant increase of goods rather than satisfaction of stable needs (Weber 1981).

There is, however, another element important in the analysis of economic attitudes: the description of tastes which are behind interests. Parsons and Smelser (1956), for instance, in a general theorization of the relations between economics and sociology, insist upon the fact that the economic preferences that constitute interests are rooted in common social values. The analysis can apply also to the very constitution of economic markets and the value of profit making (Polanyi 1944)

1.4 The Scope of Interests

The article so far has stressed the importance of the social background that narrows or enlarges the field where interests play a role, and has also indicated that, from a sociological point of view, the very definition of interests depends on a social valuation of some kind of ends.

But to this description should be added the opposite trend, that insists on the role of typical standard interests in several aspects of social life. This move does not only refer economic exchange to selfish interests, nor does it try to explain social values or institutions in terms of interest, but it also considers that even beliefs can be referred to interests. That has been the general meaning of a Marxist analysis of ideologies. But the main point here is that a critical attitude towards social order is related to the revelation of hidden interests behind common moral or political beliefs. The main difficulty of such a conception is that it does not take into account the specifically cognitive side of any ideology that cannot be so easily reduced to noncognitive contents.

But the extension of the scope of interests can also intervene without any critical intention: It is only then an effort to show that attitudes in some areas apparently removed from any economic dimension, can be interpreted in terms of interests: not only social norms, but also attitudes towards marriage, children, voting etc. Becker's (1976) or Posner's (1981) work are representative of such ambitions.

2. A Methodological Divide

2.1 Altruism and Selfishness

There is a common assumption shared by the two traditions that try to explain social order in opposed fashions regarding interests: the selfishness of indi-

viduals. Individuals are considered to act according to their interests because they are selfish. This common assumption leads to two alternative results, as we have seen. The first is to consider that social order and social norms can be explained in terms of social interests. The second is that it is not possible to do so, but, since there is a social order as well as something like a morality, which leads individuals to act morally, in an unselfish manner, it has to be referred to something different from interests.

But why should individuals be considered selfish? There are basically two reasons for this. The first is empirical and arises from the evidence of the selfish attitudes that prevail in economic exchange, where there is a huge tendency to avoid systematically letting someone's else interest prevail over one's own.

The second element is methodological: Since commitment to social values is to be explained in a way that is not tautological, it seems convenient to try to show that these norms derive from a sense of interest, that should not, for its part, be explained.

Durkheim admits straightforwardly that individuals are selfish when they are taken alone, but he does not justify this assumption. This assumption is, however, at the core of his effort to describe a specific social coercive level that could explain why individuals act in an unselfish way on certain occasions. His argument could be worded as follows: Individuals are selfish; sometimes they devote themselves to others, but since others are just like me, I should have no reason to prefer them to myself; therefore, it is not to them that I show my respect, through my moral behavior, but to society, as a distinct entity from any individual (Durkheim 1996).

But this statement rests on the idea that such altruistic behavior is always strange. One gives another individual what one has denied oneself, despite the fact that this other being is just another version of oneself not worthy of such an attitude. But in fact, most often, this so-called altruistic behavior does not correspond to this description. Usually, it intervenes in three different scenarios. The first relates to symmetrical attitudes of reciprocity. In this case, there is an admission of the idea that different individuals should have the same rights, and therefore, be allowed to have the same opportunity to pursue their own interests within a common framework. The respect of the right of the other is not the seemingly absurd renunciation of one's self-interest for the other's benefit, with no reason to do so, since the other is just another self. It can be derived from the acknowledgement of the fact that both individuals can be granted the same rights since they have the same interests. It is not that one favors the other at the expense of oneself. It is that everyone can understand, precisely, that since someone's interest should not systematically prevail upon the other's, the solution to this problem is the adoption of norms of reciprocity. A related analysis can be find in Harsanyi (1969)

The second case for altruism occurs when a stronger or wealthier person helps a weaker or poorer one. But, in this case, it is obvious that if people renounce their own interests in favor of someone else, it is precisely because they are in a position to help someone else who does not have the same means. This indicates that it is not another equivalent self that is helped, but someone who is in a worse position.

Third, it is true that sometimes there is a 'sacrifice' of the self in favor of others. Durkheim insists strongly on this kind of situation. But it could be understood, also, by considering the specific position of the person: If the captain of a ship is legally considered to be responsible for a vessel's safety because of his or her specific competence, it is therefore consistent for the captain to remain on board until everyone has evacuated.

All this means that it does not seem absolutely necessary to assume that there is some global social force behind the individual restrictions to self-interest. It is possible to understand actors' self-limitations of interest by their ability to understand the logic of a situation and the need for social rules. Therefore, three positions can be described. The first one is an individualistic and utilitarian position, that interprets most of human actions in terms of selfish interests, as in the case of Coleman's sociology.

The second is a nonindividualistic and nonutilitarian position. As we have seen, it rests upon the idea that, as individuals are selfish, and naturally committed to their own interests, the fact that they sometimes act in an altruistic manner should lead to the recognition of a specific social (not individualistic) level that is responsible for such a change of attitude.

But since it is not obvious that individuals should definitely be considered selfish and only devoted to their own interests without the intervention of a superior force, it seems possible to introduce a third methodological position, individualistic but nonutilitarian. Actors would have then to choose to favor their interests, or instead to avoid satisfying them, because they would be able rationally to understand the appropriate course of action regarding different situations.

2.2 Rationality and Interests

It is important to introduce here the notion of rationality, which has a multisided relationship with that of interest. The classical notion of rationality that prevails in economics is an instrumental one. It designates the choice of adequate means towards ends. The important point is that ends are not integrated in this sense of rationality. They are arbitrary, given at random.

But the rational choice model, highly influential both in economics and sociology, is not entirely clear (Demeulenaere 1996). On the one hand, it is officially

maintained that ends, i.e., interests, are not rational by themselves, rationality being reserved to the choice of means. But on the other hand, the description of rational behavior seems to involve an attitude that does follow selfish interests, and denies respect to any kind of value or social norm that would go against them. There is also another ambiguity. On the one hand, interests are not considered to be substantial ones, as they depend on individuals' choice. But on the other hand, interests are described as substantial: money or wealth, power and prestige.

Thus it is not clear whether this rational choice model is strongly normative or not. It would be so if it stressed that individuals are rational only if they follow their interest, in the selfish sense. But this assumption is very costly for two reasons. The first is that it is not obvious that obeying one's own interests should be the only rational choice. The second is that individuals do not have the impression that they are not rational when submitting to a norm that cannot be referred to interests. On the contrary, they believe they have reasons for doing so.

2.3 Axiological Rationality

Therefore, it seems possible to introduce, in sociological analysis, a third point of view that maintains the principle of rationality, without restricting it to the pursuit of interests or to the limited choice of means towards ends. This principle of rationality should allow the actor to discriminate between interests that deserve to be followed and those that should be avoided.

How should we define this axiological rationality (Boudon 1997)? The main point is that it would show the reasons actors have to act in a specific way, the reasons being not necessarily their interest. Boudon, for instance, shows the importance, in the voter's attitudes, of normative aspects that should not be reduced to the actor's narrow scope of interest. Those normative aspects are not arbitrary as they can be related to a system of beliefs that seems acceptable to actors regarding their general aims. Voters know that, in a large-scale election, their individual votes have no influence on the result. But they know also that other voters are aware of the same fact. If they all abstain from voting, they will lose the benefit of the democratic system. Committed to that system, they understand the importance of voting, even if it is costly in the short run and seemingly inefficient.

Such a theoretical interpretation of the actor's behavior should thus expand the concept of rationality and depart from a narrow positivistic attitude where the ends (including interests) can never be interpreted in their own rationality.

See also: Action, Theories of Social; Altruism and Self-interest; Bounded Rationality; Consequentialism Including Utilitarianism; Durkheim, Emile (1858–1917); Interest, Psychology of; Macrosociology–Microsociology; Methodological Individualism in Sociology; Moral Sentiments in Society; Rational Choice Theory in Sociology; Rationality in Society; Solidarity: History of the Concept; Theory: Sociological; Values, Sociology of; Weber, Max (1864–1920)

Bibliography

Becker G S 1976 *The Economic Approach to Human Behavior*. University of Chicago Press, Chicago

Boudon R 1997 The moral sense. *International Sociology* **12**(1): 5–24

Coleman J 1990 *Foundations of Social Theory*. Belknap Press, Chicago

Demeulenaere P 1996 *Homo Oeconomicus. Enquête sur la constitution d'un paradigme*. Presses Universitaires de France, Paris

Durkheim E [1906] 1996 *Détermination du fait moral*. Presses Universitaires de France, Paris

Harsanyi J 1969 Rational choice models of political behavior vs. Functionalist and conformist theories. *World Politics* **21**: 4

Mauss M [1925] 1954 *The Gift*. Cohen & West, London

Pareto [1916] 1963 *The Mind and Society: A Treatise on General Sociology*. Dover Publications, New York

Parsons T, Smelser N J 1956 *Economy and Society*. Routledge, London

Polanyi K 1944 *The Great Transformation: The Political and Economic Origin of our Time*. Beacon Press, Boston

Posner R 1981 *Economics of Justice*. Harvard University Press, Cambridge, MA

Weber M [1904–5] 1974 *The Protestant Ethic and The Spirit of Capitalism*. Allen and Unwin, London

Weber M [1923] 1981 *General Economic History*. Transaction, New Brunswick, NJ

P. Demeulenaere

Interference and Inhibition, Psychology of

The concepts of interference and inhibition have a long and diverse history (see Dempster 1995, Macmillan 1996, Smith 1992), and have benefited from a new thrust of energy in cognitive psychology (Dagenbach and Carr 1994, Dempster and Brainerd 1995, Dempster and Corkill 1999, Houdé 2000). These two concepts are directly interconnected, as suggested in the title of the book edited by F. N. Dempster and C. J. Brainerd (1995), *Interference and Inhibition in Cognition*. If interference causes an alteration in performance (triggered by prior learning, a distractor, etc.), then inhibition is the cognitive process that enables subjects to resist interference, as clearly stated by Dempster in the early 1990s (Dempster 1992, 1993). Two fields of cognitive psychology where focusing on

interference and inhibition has led to the newest and greatest insights are the study of selective attention processes, and the study of cognitive development.

1. Interference, Inhibition, and Selective Attention

Traditionally, selective attention has been thought to be a function of activation. That is, after initial automatic processing, relevant information is selected (early or late, depending on the locus of the 'selective filter') by an activation or facilitation mechanism (Broadbent 1958, Deutsch and Deutsch 1963, Norman 1968, etc.). From then on, irrelevant information is no longer processed, and its effect dissipates passively over time. An alternative conception, the attention-*inhibition* view (see Bruce and Tipper 1998, Houghton et al. 1996, Neill et al. 1995), has gradually taken over from classical attention-*activation* theory. In this new approach, the essential mechanism of selective attention is inhibition (or resistance to interference), i.e., the active blocking of irrelevant information. In this case, the cognitive processing of relevant information, after selection, is not seen as being due to specific activation-facilitation, but simply to the fact that processing is no longer adversely affected by interfering irrelevant information.

For psychologists, the question becomes how can these two possible modes of cognitive selection be distinguished experimentally. The paradigm designed to do this is negative priming. Using this paradigm has become the most popular way of measuring inhibition and its interconnection with interference. Take a situation where the subject has to respond first to S1 (the relevant or target stimulus) while ignoring S2 (the irrelevant or distractor stimulus, the one that interferes). Suppose that afterwards the subject unexpectedly has to respond to S2, or in another condition to S3 (a new stimulus). The first part of this procedure is called 'the prime' and the second part is called 'the probe.' The attention-*activation* view says that during the prime, S2 dissipates passively over time since it is not selectively activated. If the effect of S2 has not yet completely disappeared from memory when the probe appears, then S2 (now the relevant stimulus) should be easier and faster to process than S3. This is the classic priming effect (positive priming or facilitatory effect), which is measured using mental chronometry (reaction time in ms). Still in the activation view, if the initial effect of S2 has completely faded when the probe arrives, then S2 processing should not differ from S3 processing. Now in the alternative view, attention-*inhibition*, the prediction for this same prime–probe sequence is exactly the opposite. In this case, S2 is inhibited on the prime, i.e., it is actively blocked in resistance to interference, so it should be more difficult (slower) to process than S3 on the probe. This is the negative priming effect.

Under the impetus of S. P. Tipper (1985), who introduced the negative priming idea into cognitive psychology, a large number of experimental studies on adults have confirmed the existence of this phenomenon in a wide variety of situations, including identification tasks (picture naming, word naming, letter identification), categorization tasks (semantic categorization, lexical decision), matching tasks (letter matching, shape matching), counting tasks, localization tasks, and so on. In Tipper's (1985) influential study using picture-naming tasks, subjects were shown superimposed line drawings. One stimulus (S1), the relevant one (target), was drawn in red, and the other (S2), the irrelevant one (distractor), was drawn in green (e.g., S1 = a kite and S2 = a trumpet). So on the prime, interference came from the green stimulus (here, the trumpet). In the negative priming condition, the prime–probe sequence was such that the prime distractor became the probe target (the same trumpet now drawn in red ink). Naming latencies were significantly slower in this condition than when the prime and probe trials were unrelated (that is, when a new stimulus, S3, was introduced as the probe target). Note that the negative priming effect was also observed with a solely semantic relation (i.e., nonidentical) between the prime distractor and the probe target (e.g., a cat and a dog, respectively).

This first work by Tipper, and the many studies that followed (for reviews, see Bruce and Tipper 1998, Neill et al. 1995), have brought the inhibitory control of information into the foreground as an important mechanism of resistance to interference, with the negative priming (observed on the probe) being indicative of the existence and effectiveness of the inhibition (generated earlier on the prime). New and more precise questions then arose, such as: What is the exact nature of interference, and hence, what is in fact inhibited (the response, the perception, the representation)? Does inhibition depend on task requirements? How is episodic memory (retrieval) involved in negative priming? Is there a single inhibition mechanism or are there many? What parameters affect negative priming, knowing that in certain cases the expected effect is not observed? All of these theoretical and experimental questions are now under study in current research.

Initially, the negative priming paradigm was defined in terms of stimuli, with S1 as the target and S2 as the distractor (see the Tipper example above), but current research as illustrated in the following section also uses a variation applied directly to cognitive strategies.

2. Interference, Inhibition, and Cognitive Development

In the study of cognitive development, new experimental research has also been aimed specifically at

analyzing interference and inhibition (Bjorklund and Harnishfeger 1990, Dempster 1992, Dempster and Brainerd 1995, Houdé 2000, Houdé et al. 2000, Pascual-Leone 1988, Tipper et al. 1989). In this framework, the underlying assumption is that cognitive development cannot be reduced to the mere coordination-activation of structural units or scheme (as in Jean Piaget's 1984 structuralist theory and in the neo-Piagetian models, see Demetriou 1988) but that developing also means learning to inhibit a competing strategy (scheme). To illustrate this approach, two examples are given below in the areas of object construction in infancy and numerical abilities in childhood.

The question of the relationship between cognitive development, interference, and inhibition becomes a relevant one as soon as the basic unit of reality, the permanent object, is in place in the infant. Research on infant oculomotor activity, which uses the violation-of-expectancy paradigm, has shown that early object permanence already exists at the age of four or five months. How, then, can one explain the well-known A-not-B error, observed by Piaget (1954) at eight months and present until the age of one year? To demonstrate this error, the experimenter puts an infant in front of two covers that are equally easy to reach (A and B), and then puts an object under cover A. The infant has no trouble finding it. After a few repetitions (A-A-A- …), the object is conspicuously moved under cover B. The infant who continues to search under A makes the A-not-B error.

According to Piaget, this error is a testimony to the lack of object permanence, in the sense that infants should know that the object continues to exist under cover B because that is where they lost sight of it. But this explanation is no longer tenable today. As indicated above, new research has shown that the object permanence scheme (acquired by the early age of four or five months) clearly precedes the A-not-B error. (Certain authors (Smith et al. 1999) even indicate that there are circumstances in which older children and adults make a similar error.) An interference-and-inhibition-based analysis is better able to resolve this paradox. Situations where oculomotor reactions to unexpected events are observed (events where object permanence is violated) can indeed be considered as optimal contexts where the 'simple' activation of object permanence suffices. The Piagetian situation, on the other hand, where the object disappears under A-A-A- … and then under B, is a misleading situation which, according to neuropsychological analyses of the connection between frontal cortex maturation and the A-not-B error, requires the inhibition of a dominant motor tendency, i.e., the interfering preprogrammed gesture towards A (Diamond 1991, 1998). The A-not-B error is thus considered to be the outcome of an executive failure to inhibit a motor response—which leads to perseveration—and not the lack of the object permanence scheme. More exactly, Diamond (1991, 1998) defends a 'memory + inhibition' interpretation. Already in infancy, then, being intelligent (no longer making the A-not-B error) means inhibiting an interfering scheme. In agreement with Diamond (1991), 'Cognitive development can be conceived of, not only as the progressive acquisition of knowledge, but also as the enhanced inhibition of reactions that get in the way [interference] of demonstrating knowledge that is already present [here, object permanence]' (p. 67).

In the area of numerical development, new research has also revealed the existence of early abilities that were unknown to Piaget. By recording oculomotor behavior, recent studies have shown that four- and five-month-olds are capable of detecting the violation or conservation of number when presented with unexpected or expected numerical events. It has also been shown that these early numerical abilities undergo a cognitive-linguistic reorganization process and then re-emerge in preschoolers after a temporary drop in performance (Houdé 1997). So why, then, do children at this age answer incorrectly on Piaget's (1952) conservation of number task? When shown two rows of objects that contain an equal number of objects but differ in length (because the objects in one of the rows have been spread apart), these children think the longer one has more objects.

We know that Piaget's interpretation was that preschool children are still fundamentally intuitive, or as he called them, 'preoperational,' and hence limited to a global and holistic perceptual way of processing information (here, based on length or, in certain cases, on density). In this view, they have not yet integrated the number strategy that enables them to perform an analytic process involving an exact calculation. The new studies mentioned above cast doubt on this interpretation, suggesting that the conservation of number task is a number/length interference task more than anything else, one that reflects the ability to resist the visuospatial length-equals-number strategy (an often relevant quantification heuristic still used by adults). As Dempster (1995) said, 'Conservation and class inclusion have more to do with the ability to resist interference than they do with the child's ability to grasp their underlying logic' (p. 15).

Here again, as in the A-not-B error in object construction, being intelligent (resisting the visuospatial interference) is essentially being capable of inhibition. This interpretation was recently confirmed using a negative-priming adaptation of Piaget's numerical task (Houdé and Guichart 2001). A chronometric paradigm (adapted from Tipper 1985; see above) was used with nine-year-olds, who succeed in Piaget's conservation of number task, to test the role of cognitive inhibition in a priming version of this classical task. The experimental design was such that the misleading strategy 'length-equals-number' to be inhibited on the prime (a Piaget-like item with number/length interference) became a congruent strategy to be

activated on the probe (a subsequent item where number and length covaried). A negative priming effect, manifested by slower reaction times, was observed for the prime–probe sequence (compared to an unrelated condition). This result thus confirmed that, first and foremost, success on Piaget-like tasks requires an inhibitory process. In this light, the solution to the enigma of cognitive development could be found by looking mainly at the interference-and-inhibition side of cognition rather than searching solely on the scheme-coordination or coactivation side, as Piaget did.

Thus, whether it be in the study of cognitive development in infancy and childhood, or in the study of selective attention in adulthood, interference and inhibition are key concepts for new approaches in cognitive psychology (and also in the field of cognitive aging, following Dempster's 1992 theoretical input). One of the most exciting research trends for addressing these questions is neurofunctional imaging, a new field where the cerebral bases of cognitive interference and inhibition are beginning to be explored (Ghatan et al. 1998), particularly in the anterior cingulate cortex (Bush et al. 2000) and in its strong reciprocal interconnections with the lateral prefrontal cortex. Another exciting perspective is the current evolutionary framework (Bjorklund 1997, Bjorklund and Harnishfeger 1995, Cosmides and Tooby 1987), where inhibition is regarded as a possible 'Darwinian algorithm.'

See also: Attention and Action; Attention: Models; Attention, Neural Basis of; Priming, Cognitive Psychology of

Bibliography

Bjorklund D F 1997 In search of a metatheory for cognitive development (or, Piaget is dead and I don't feel so good). *Child Development* **68**: 144–8

Bjorklund D F, Harnishfeger K K 1990 The resources construct in cognitive development: Diverse sources of evidence and a theory of inefficient inhibition. *Developmental Review* **7**: 93–130

Bjorklund D F, Harnishfeger K K 1995 The evolution of inhibition mechanisms and their role in human cognition and behavior. In: Dempster F N, Brainerd C J (eds.) *Interference and Inhibition in Cognition*. Academic Press, New York, pp. 141–73

Broadbent D E 1958 *Perception and Communication*. Pergamon, London

Bruce M, Tipper S P 1998 Attention and inhibition. In: Pashler H (ed.) *Attention*. Psychology Press, Hove, UK, pp. 191–21

Bush G, Luu P, Posner M I 2000 Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Science* **4**: 215–22

Cosmides L, Tooby J 1987 From evolution to behavior: Evolutionary psychology as the missing link. In: Dupre J (ed.) *The Latest on the Best Essays on Evolution and Optimality*. MIT Press, Cambridge, MA, pp. 277–306

Dagenbach D, Carr T H (eds.) 1994 *Inhibitory Processes in Attention, Memory, and Language*. Academic Press, New York

Demetriou A (ed.) 1988 *The Neo-Piagetian Theories of Cognitive Development*. North-Holland, Amsterdam

Dempster F N 1992 The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. *Developmental Review* **12**: 45–75

Dempster F N 1993 Resistance to interference: Developmental changes in a basic processing mechanism. In: Howe M L, Pasnak R (eds.) *Emerging Themes in Cognitive Development: Vol. 1. Foundations*. Springer, New York, pp. 3–27

Dempster F N 1995 Interference and inhibition in cognition: An historical perspective. In: Dempster F N, Brainerd C J (eds.) *Interference and Inhibition in Cognition*. Academic Press, New York, pp. 3–26

Dempster F N, Brainerd C J (eds.) 1995 *Interference and Inhibition in Cognition*. Academic Press, New York

Dempster F N, Corkill A J 1999 Neo-interference research and the development of intelligence. In: Anderson M (ed.) *The Development of Intelligence*. Psychology Press, Hove, UK, pp. 215–43

Deutsch J A, Deutsch D 1963 Attention: Some theoretical considerations. *Psychological Review* **70**: 80–90

Diamond A 1991 Neuropsychological insights into the meaning of object concept development. In: Carey S, Gelman R (eds.) *The Epigenesis of Mind. Essays on Biology and Cognition*. Erlbaum, Hillsdale, NJ, pp. 67–110

Diamond A 1998 Understanding the A-not-B error: Working memory vs. reinforced response, or active trace vs. latent trace. *Developmental Science* **1**: 185–9

Ghatan P H, Hsieh J C, Petersson K M, Stone-Elander S, Ingvar M 1998 Coexistence of attention-based facilitation and inhibition in the human cortex. *NeuroImage* **7**: 23–9

Harnishfeger K K 1995 The development of cognitive inhibition: Theories, definition, and research evidence. In: Dempster F N, Brainerd C J (eds.) *Interference and Inhibition in Cognition*. Academic Press, New York, pp. 176–204

Houdé O 1997 Numerical development: From the infant to the child. Wynn's (1992) paradigm in 2- and 3-year-olds. *Cognitive Development* **12**: 373–91

Houdé O 2000 Inhibition and cognitive development: Object, number, categorization, and reasoning. *Cognitive Development* **15**: 63–73

Houdé O, Guichart E 2001 Negative priming effect after inhibition of number/length interference in a Piaget-like task. *Developmental Science* **4**: 71–4

Houdé O, Zago L, Mellet E, Moutier S, Pineau A, Mazoyer B, Tzourio-Mazoyer N 2000 Shifting from the perceptual brain to the logical brain: The neural impact of cognitive inhibition training. *Journal of Cognitive Neuroscience* **12**: 721–8

Houghton G, Tipper S P, Weaver B, Shore D I 1996 Inhibition and interference in selective attention: Some tests of a neural network model. *Visual Cognition* **3**: 119–64

Macmillan C J 1996 The concept of inhibition in some nineteenth century theories of thinking. *Bruin and Cognition* **30**: 4–19

Neill W T, Valdes L A, Terry K M 1995 Selective attention and inhibitory control of cognition. In: Dempster F N, Brainerd C J (eds.) *Interference and Inhibition in Cognition*. Academic Press, New York, pp. 207–61

Norman D A 1968 Toward a theory of memory and attention. *Psychological Review* **75**: 522–36

Pascual-Leone J 1988 Organismic processes for neo-Piagetian theories. In: Demetriou A (ed.) *The Neo-Piagetian Theories of Cognitive Development*. North-Holland, Amsterdam, pp. 25–64
Piaget J 1952 *The Child's Conception of Number*. Routlege & Kegan Paul, London
Piaget J 1954 *The Construction of Reality in the Child*. Basic Books, New York
Piaget J 1984 Piaget's theory. In: Mussen P H (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. 1, pp. 103–28
Smith L B, Thelen E, Titzer R, McLin D 1999 Knowing in the context of acting: The task dynamics of the A-not-B error. *Psychological Review* **106**: 235–60
Smith R 1992 *Inhibition. History and Meaning in the Sciences of Mind and Brain*. Free Association Books, London
Tipper S P 1985 The negative priming effect: Inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology* **37A**: 571–90
Tipper S P, Bourque T A, Anderson S H, Brehaut J C 1989 Mechanisms of attention: A developmental study. *Journal of Experimental Child Psychology* **48**: 353–78

O. Houdé

Intergenerational Justice

Intergenerational justice refers to the negative and positive obligations one generation owes to another. The concept of a generation is essentially indeterminate as there is no point at which a single generation stops and another begins. Within any single population a number of generations will coexist. Intergenerational justice does not refer to relationships of right and duty between different generations within one temporally continuous population. The obligations of young to old or old to young are matters of distributive justice. Intergenerational justice refers to the relationships of obligation, right, or benevolence that ought to exist between groups of people who are not temporally continuous or to that set of issues which apply to noncontinuous populations, in the same way that international justice applies to populations that do not spatially overlap.

The fact of nontemporal continuity explains the problem of intergenerational justice. Many philosophers argue that the absence of temporal continuity casts doubt on the possibility of justifying claims of obligation or right between generations. Discussions of intergenerational justice focus on two sets of questions; first, how do we explain obligations between nontemporally continuous populations and second what obligations might we have to future generations. This article will examine the bases of these doubts about extending the idea of justice to apply to the relationship between nontemporally continuous populations and then conclude with an assessment of what we might owe to posterity.

1. *Contractualism and Asymmetry*

Contemporary discussions of distributive justice are synonymous with the resurgence of contractarian thinking under the influence of philosophers such as Rawls (1971) and Gauthier (1986). We can distinguish between 'mutual advantage' and 'impartialist' contractarianism (Barry 1989). The contract device is used for thinking about distributive questions in circumstances where the parties disagree about ultimate moral principles. If we cannot base our obligations to one another by appeal to a common standard such as utility maximization, we can use the idea of a rational bargain or reasonable agreement as the basis for distributive norms and obligations. The idea of a contract is a thought experiment that can be used to legitimize principles rather than an actual agreement process that groups of people have to collectively engage in. However, even as an hypothetical device the idea of a contract (in both versions of the theory) poses a major problem because of the asymmetrical relation between the participants in intergenerational cases.

The 'mutual advantage' theory argues that distributive principles can be derived from a rational bargain between participants in a cooperative venture, such as market society or political community. The issue of justice only arises when certain circumstances of justice obtain. If each individual could get everything they wanted without cooperation there would be no need for principles of justice. The circumstances of justice are threefold. First, (roughly) equal power among the parties. What one lacks in strength they can make up in guile so no one is guaranteed always getting their own way. Second, moderate scarcity of goods and motivations obtains, so that social cooperation is necessary or at least in the interest of all. Finally, while all benefit from the existence of social cooperation the participants are interested in the distribution of its benefits and burdens. As rational agents each has an interest to maximize their share of the benefits but minimize their share of the burdens of cooperation.

The problem for the 'mutual advantage' theorist is that the circumstances of justice do not unequivocally obtain between noncontinuous generations. If we use the idea of the contract within one generation we can make sense of the idea of rough equality of power and reciprocity as the key motivations for accepting the distribution of the benefits and burden of social cooperation (although we might still argue that this is a morally inappropriate way to think about justice). Yet when we extend this idea to noncontinuous generations the weakness of the method becomes apparent. In what way can we both speak of ourselves and future generations as part of the same 'cooperative' venture? If we just concentrate on the rough equality of power we can see how there is an asymmetrical relation between the generations. Take an issue such as energy policy. We can conceive of

examples where a present generation can unilaterally impose significant costs on future generations such as the management of toxic or nuclear wastes. It is clearly possible for a present generation to take such choices without any possible negative consequences from the future generation except perhaps disapproval and poor historical reputation. The present generation is able unilaterally to affect the interests of a future generation, whereas a future generation cannot unilaterally affect the interests of the present generation, hence the asymmetry. Without rough equality of bargaining power any contractual agreement will simply reflect the advantage of the stronger party, in this case the present generation. If this is so, the contract does nothing more than allow the present generation to pursue its own interests taking account of the future only when its own reputation is at stake and when this is of higher value than any other present interest such as standard of living. Because of the asymmetrical relationship underlying the bargain it is unclear how the present generation could be motivated to assume obligations to the future given the motivational assumptions of 'mutual advantage' contractarianism. The contract not only does not generate obligations to the future, but would appear to generate reasons for ignoring such claims.

The 'impartialist' theory departs from many of the key features of the 'mutual advantage' bargain just considered. Rawls's theory introduces the idea of a hypothetical original position in which the participants deciding which principles of justice should shape the basic structure of their society as a fair system of social co-operation, are shrouded by a 'veil of ignorance.' The point of the veil of ignorance is to filter out the possibility of exploiting 'positional' advantage or inequalities of bargaining power. By denying agents behind the veil of ignorance, information about their particular circumstances and goals, and allowing them only general knowledge about society, Rawls constructs an impartial choice situation. Rational self-interest plus ignorance delivers impartiality and fairness.

Rawls' theory is a thought experiment designed to test the fairness of candidate principles of justice. The contract with the veil of ignorance is designed to model a fair choice situation. However, once again the asymmetry problem arises. The agents behind the 'veil or ignorance' are denied information that will prevent them pursuing their own interests at the expense of each other. The agreement is effectively between members of a single generation. What is to stop the participants choosing principles in order to maximize the condition of the worst off currently in existence that will at the same time worsen the situation of those not yet born? In order to bring future generations into the agreement Rawls posits an additional motive for the participants. They are expected to reason as representatives of families who are assumed to have a concern for their descendants. Heads of families provide the intergenerational link that overcomes the asymmetrical relations between future generations. The problem with this strategy is that it merely concedes the problem of asymmetry and the inability of the contract device to generate adequate obligations to the future. At best the additional motivational assumption gives those in the present generation a reason to take account of future generations. It does not establish an obligation to future generations in the way that the contract does establish obligations within one generation. The addition of the extra motivational assumption provides a noncontractarian reason for taking account of future generations. In effect it assumes precisely what it is supposed to provide, namely a reason for taking account of the interests of future generations.

While contractarian arguments have been central to the development of current theories of distributive justice, they seem singularly ill-equipped to deal with issues of justice between generations because of the asymmetrical relationship between even hypothetical bargainers. This is a point acknowledged by Barry, one of the most significant defenders of a neo-Rawlsian contractarianism (in Dobson 1998b). The key relationship of reciprocity and the equal ability of the partners to affect the interests of each other do not apply in the case of noncontinuous generations. To overcome this the contractarian has to appeal to noncontractarian reasons.

2. *The Nonidentity Problem*

A further ground for skepticism which is related to the asymmetry objection, but which has implications beyond contractarian arguments is the nonidentity problem. This problem has been given its most forceful modern restatement by Parfit (1984). And it is as applicable to theories such as utilitarianism as it is to contractarian theories of justice.

The nonidentity problem draws on a paradoxical implication of the intuitive response to issues of intergenerational justice. Assume that policy makers are making decisions about technologies that deplete environmental resources or cause global warming. These decisions involve questions of intergenerational justice because they do not merely distribute the benefits and burdens of social cooperation within generations but also across generations. Any policy that unilaterally worsens the conditions of future generations by depleting nonrenewable resources without suitable compensations for the future might be deemed to involve an intergenerational injustice as it sacrifices the interests of future generations to serve our own temporary interests. In this case the injustice consists of harming the interests of future generations by denying them opportunities or by unilaterally imposing costs upon them. Against this a sceptic

might respond that our intuitive concern for the future does not withstand critical scrutiny and that we should discount the interests of future generations in determining between environmental policies.

The argument claims that only harms to assignable individuals constitute cases of injustice. An action becomes a case of injustice when it involves harm to the interests of an individual agent or groups of individuals. But if we turn to the issue of appraising environmental policies it is far from clear that they do distribute harms to assignable individuals.

Whichever environmental policy is chosen will have complex consequences for the future, many of which are indirect. One such consequence is on the procreation decisions of particular populations, that is, on who should come to mate with whom. The nonidentity problem depends upon the assumptions that any decision to procreate at one time rather than another will result in the birth of different people. The fact that I was born at the time I was and in the place I was in part made me the person I am now. If I was born either earlier of later my experiences would have been different and therefore I would have been a different person. If this claim is true then any policy which affects peoples behaviour will have implications for the identity of future generations. Thus, an environmental policy which forces me to use my car less and travel more on public transport may be responsible for me meeting my future partner and therefore albeit indirectly for any future persons who result from our relationship.

The important point is that whichever policy is chosen will result in different future populations and not simply different levels of welfare for the same potential population. But if different populations result from different policy choices then does it make sense to talk about obligations of intergenerational justice? We can argue that a policy of resource depletion does not involve any issue of intergenerational justice because it does not harm the interests of any assignable individual. After all whichever individuals do find themselves bearing the burdens of our choices only exist as the people they are because of choices made by us in the past. If we had chosen otherwise they would not have existed at all, but rather some other group of persons would have existed. For the contractarian the problem of showing how assignable individuals are harmed by our choices undermines the claim that we can do injustices to people in the future and this fact renders the whole idea of intergenerational justice deeply problematic. But the problem also applies to utilitarian theories that explain the idea of injustice in terms of worsening the welfare of individuals. We cannot, for example, argue that a future population would have a higher welfare if we had adopted a policy of conservation than if we had adopted one of resource depletion for either policy would result in different populations. The issue would then become one not of whether population X had a welfare level of a or b, but whether population X should exist or whether population Y should exist. The utilitarian can respond to this claim but only at the expense of opening themselves to certain repugnant conclusions.

3. *Utilitarianism and Repugnant Conclusions*

The idea behind the nonidentity problem is that intergenerational obligations must be person-regarding, that is because obligations are owed to assignable individuals. But as we have seen, as the identity of individuals in the future is itself dependent upon present actions, the idea of harming or benefiting future generations becomes problematic. Utilitarians—those philosophers who believe our most fundamental obligation is to promote, either total or average, welfare—often reject the person-regarding view of moral obligations by claiming that overall welfare is prior to questions of its distribution (see Stearns 1972). For the utilitarian we have an overriding obligation to promote utility or welfare and this avoids the nonidentity problem because we can assess overall outcomes irrespective of the identity of the recipients of utility or welfare. This strategy raises a number of important issues when applied to intergenerational questions, but its most striking consequences arise from its application to population policy.

If we have a primary obligation to promote or maximize welfare and this obligation is prior to any obligations to assignable individuals then we are inevitably faced with the question how many people should there be? This utilitarian obligation to promote welfare has direct implications for individual procreative decisions. Before considering whether to conceive a child or bring it into the world a couple will need to consider the impact of this decision on overall welfare. We can consider the impact of this procreative choice on either average or total welfare.

Let us first consider the case of average utility. Here we are required to consider the impact of our procreative choices on the average level of utility across the relevant population. If our society is a prosperous and happy one and the addition of further members who, though happy, would be slightly less happy than the average, then their impact would be negative. In such a case average utilitarianism would counsel against conceiving such a child. Such a way of conceiving procreative decisions might seem suspect but often we do apply such reasoning in thinking about population policy. Some utilitarians consider draconian population policies are justified by average utility. However, opponents of utilitarianism argue that such an approach to procreative questions is wholly inappropriate because social welfare is a morally irrelevant consideration.

Average utilitarianism does not merely have implications for whether we should allow the birth of happy children who are merely less happy than average, it also has implications for those who exist but are less happy than the average. Consider a population of 10 people two of whom are miserable while eight are happy. If we add to this consideration the fact that there is no hunger in the society so that the eight would not gain in utility if the two were removed, we could still increase the average utility by removing the two. However, we could then make similar judgments within the eight until we were left with a very small population perhaps numbering only two in which there was no difference in average utility. In existing populations there are always groups whose elimination would, other things being equal, increase average utility.

What happens when these arguments are applied across populations, as in the case of intergenerational justice? If we accept the premise of the nonidentity argument, namely that the policies we pursue will affect the identity and size of future generations then using the average utility argument we can engage in some interesting trade-offs. If we are concerned about a policy of conservation vs. one of depletion we can reason as follows. A depletion policy would have a large negative impact on a large population but a relatively low impact on a smaller future population. As long as the impact diminished the fewer people there were to be affected we would satisfy our utilitarian obligation. This could easily be achieved by coupling the depletion policy with a draconian population policy amongst our existing population, such that the average utility of the small population x is greater than the average utility of population y that would have resulted from the policy of conservation. The point here is that we can make one of the aspects of the nonidentity problem a policy objective, in order to justify our depletion of resources. As long as we can manipulate the size of future populations by regulating procreative decisions we can pursue almost any policy without harming the interests of future generations.

Average utility results in some questionable conclusions, does total utility fair any better? In the case of individual procreative decisions the issue becomes one of the duty to maximize happy people. As long as each child being born is happy to some degree they contribute to the total sum of happiness. It need not be the case that they are happier than their parents only that they are happier to exist than not to exist by however small a degree. If we grant this condition, which follows from the commitment to maximizing total utility, then we are faced with an obligation to bring as many happy people into the world as possible, however small the individual level of happiness of each person is. The most striking implication for intergenerational justice is that we act justly to the future when we make the size of future populations as big as they can be irrespective of how low the average utility of each member of those future generations might be. Whereas average utility suggests the imposition of particularly draconian population policies, total utility goes to the opposite extreme by encouraging population maximization. The full impact of this repugnant conclusion can be seen when we apply the total utility strategy to the choice between policies of depletion versus conservation. We need to be able to judge the relative utility or welfare of one population against another who would have lived if an alternative policy had been pursued. As long as the total utility of population a is greater than that of population b that might have existed as a result of a different policy choice, then we are justified in pursuing policy a over b. This can help us explain and justify some trade-offs between population size and level of utility such that we might prefer a policy which resulted in a larger population with a lower utility to a smaller population with a higher utility, for example population a with n members each having a utility of 10 as opposed to population b with $2n$ members each having a utility of eight. But this approach becomes counterintuitive and repugnant when we consider the case of a third population of $10{,}000n$ each having a utility of 0.01. In this latter case we face the repugnancy of the total utilitarian approach whereby as long as the population is large enough the utility of the individual members does not matter as long as it is above zero.

By rejecting the person-regarding aspect of justice arguments, utilitarianism can overcome the criticism that we only have obligations to assignable individuals (and therefore we cannot be said to have obligations to future generations) because of the nonidentity problem. They get around the problem only by exposing themselves to different problems resulting from their indifference to issues of distribution that are also at the heart of intergenerational justice arguments. Both forms of utilitarianism considered here address the issue of intergenerational justice by the artificial means of manipulating population size. This strategy either loads the dice in favor of existing generations as long as they are prepared to manipulate the size of future generations accordingly or to reduce our obligations to the simple duty to procreate as much as possible. Neither approach seems an adequate answer to the question of what if any obligations we have to future generations. The two main candidate theories for grounding claims of justice, contractualism, and utilitarianism have serious problems in making sense of the identity of future generations and the idea that we may have obligations to them. Does this mean that the sceptic is right and the idea of intergenerational justice is flawed?

4. *The Prospects for Intergenerational Justice*

The contemporary literature on intergenerational justice is largely focused on discussions of the paradoxical

consequences of applying existing theories of justice to intergenerational problems. Does this mean that the sceptic is correct in arguing that the idea of relationships of justice between generations is incoherent? Two contemporary political philosophers, De-Shalit (1995) and Barry (in Dobson 1998b) reject this sceptical position and suggest that there are ways we can overcome the sceptical position to make sense of intergenerational obligations. However, they do so from very different perspectives, De-Shalit is a communitarian whereas Barry is a universalist egalitarian.

4.1 Transgenerational Community

De-Shalit begins with the challenge posed by the inadequacy of contractarianism and the nonidentity problem. Faced with the problem of asymmetry in the original position, Rawls introduces an additional motivation linking the generations to justify a just savings principle as a constraint on contemporary consumption. This additional motive which requires the participants to think as heads of families is supposed to model the motive to be concerned about our posterity. De-Shalit takes this natural motive and generalises it across the idea of constitutive communities. Why should the family be the only repository of such a motive? His point is that it is not the particular institution that is the repository of this motive but the type of relationship. The contractarian and the non-identity argument posit a sharp break between generations that creates an unbridgeable moral gap between persons. De-Shalit challenges this individualization by arguing that our identities are shaped and constituted by membership of communities. Membership is itself obligation creating, in the same way that Rawls assumes that acting as head of families will provide the sufficient motive to assume obligations to the future. The motivation to be concerned about posterity is for De-Shalit a natural motivation and this is missed by the narrowly 'individualistic' approaches of contractarians and utilitarians. The task for the political philosopher is then not to construct the basis of the obligation to the future. We have this natural concern because our own identity and interests are tied up with the membership and existence of a certain type of community. The task is to expand the implications of that concern for posterity and the specific obligations that can be derived from it. Two important implications arise from De-Shalit's communitarian turn. First, he broadens the notion of those goods we have a duty to bequeath to posterity. Whereas much of the literature assumes merely that our obligations to the future concern material resources and the environment, De-Shalit's emphasis on constitutive communities entails that all those definitive aspects of such communities might be 'goods' that we have an obligation to preserve for posterity. Thus, we could have obligations to preserve languages and cultures or defining institutions such as constitutions, political rights, etc. The second implication is that De-Shalit changes the focus of intergenerational justice from a universal to a particular one. We have obligations to our posterity and not future generations as such. The British people might have obligations to preserve their constitution for future generations, or the Quebequois might have an obligation to pass on their 'distinct society' to future Quebequois, but not to future Nigerians or for the good of future Japanese. For De-Shalit it is not assignable individuals that we need worry about but only assignable groups. Obligations only extend as far into the future as it is reasonable to think a constitutive community will retain its identity.

One possible objection to De-Shalit's approach is that it makes group membership too important. If we only have obligations to our own posterity then we only have obligations to preserve the environment for our posterity. But this raises the counter intuition that it surely cannot be right that I have a duty to protect the environment for my own posterity but that if a policy of depletion falls not on my posterity but on another people then I have no obligation or responsibility. It is for this reason that other theorists are reluctant to abandon moral universalism, or the idea that from a moral point of view spatial and temporal location are not significant.

4.2 Universalism Defended

The key intuition underlying Barry's universalism is that place and time do not provide morally relevant bases on which to discriminate between the interests of those who fall under a principle. This assumption is morally basic, as such it must be defended indirectly by challenging the challenges to it and showing how it accords with our reflectively adjusted intuitions. We might, for example, show that it overcomes the narrow parochialism of De-Shalit's communitarianism. Alternatively, we might suggest that it is the intuition that underlies Rawls' additional motivation designed to supplement contractarianism. The universalist is also likely to reject the utilitarian variant because of the repugnant conclusions that follow from abandoning a person-respecting morality. A universalist might deny that we have an obligation to bring about maximal well-being if this results in the kinds of population policies that result in repugnant conclusions.

One simple way of avoiding this utilitarian paradox is to conceive of obligations of justice not in terms of promoting the good as maximal welfare but in terms of not harming the interests of the future generations. This negative obligation has implications for the redistribution of resources and benefits, it does not merely involve staying ones hand and refraining from

causing harm. In this way we can argue that we have an obligation to give weight to the interests of future generations when we make irreversible consumption decisions. We ought not merely discard future people because they are in the future. However, it is important to notice that the relationship of obligation is between those existing not to harm those in the future, it is in effect a matter of how we regard the future rather than a matter of how we and those in the future interact.

The major objection still facing this approach is the nonidentity problem. We can only have obligations to assignable individuals but the identities of these individuals are contingent on the choices we currently make such that we cannot be said to be harming them. Referring to 'future generations' does not mark out the identity of those being harmed. One way to overcome this problem is to contest the idea that only assignable individuals can be harmed. Thus, it might make sense to say that our decisions about resource depletion can harm future generations relative to the absence of that policy even though the existence of particular individuals might be dependent upon that policy. The nonidentity paradox will only work against such a view if it could be shown that the identities of all members of the relevant future generations owe their existence only to this environmental policy. We can also contest the idea implicit in the nonidentity problem that we cannot be said to harm the interests of members of a future generation as their interests will be a function of the circumstances in which they find themselves. While it is certainly true that any complex account of universal transtemporal interests will be hard to conceive of, there are certain things that it can be argued will be in any subsequent group's interest. It is clearly in no future group's interest to live in a world in which potable water is made scarce by pollution, or in which ozone damage raises levels of carbon dioxide in the atmosphere and levels of harmful radiation. Future generations are also likely to place some value on their environment and biodiversity. How much each of these things will matter is debatable, but that they will matter is uncontestable. Where this leaves the universalist is with the idea that we have obligations to take account of the 'general' interests of future generations not to be harmed in certain ways, whoever they happen to be and whatever else they happen to want. Thus, the burdens of intergenerational justice fall on contemporaries in the form of not harming those interests of future generations by over consumption, environmental depletion and unsustainable development without permanent advances to any future generation.

As subsequent generations extend far into the future our common intuitions suggest that our obligations become weaker in terms of just savings and the transfer of resources. In the case of private individuals the motive to save for offspring is strong but weakens across many generations until we reach a stage where most are indifferent to distant posterity. But the universalist approach is able to accommodate this intuition because it transforms the debate from one of just-savings and the intergenerational transfer of wealth into a present day concern for sustainable economic and political policies which do not worsen the position of those in the future. The idea of sustainable economic development has its own broad and complex literature but the idea of sustainablity gives content to how we might think about justice between generations. In this way the idea of justice between generations converges with the environmentalists' concern with respect for the environment as a common good. If we have obligations to future generations they are to respect the common resource that is the environment which we share. Where intergenerational justice differs from environmental ethics is that it regards the value of the environment as an aspect of human interests. This of course is not the only way we might regard it, but that is merely to recognise that even in the intergenerational case there is more to morality than justice.

See also: Children, Value of; Consequentialism Including Utilitarianism; Deserts, Philosophical Concept of; Family Theory: Economics of Intergenerational Relations; Fertility Theory: Caldwell's Theory of Intergenerational Wealth Flows; Generation in Anthropology; Generations, Sociology of; Justice and its Many Faces: Cultural Concerns; Justice: Philosophical Aspects; Property: Legal Aspects of Intergenerational Transmission; Rights; Social Justice; Utilitarian Social Thought, History of; Utilitarianism: Contemporary Applications

Bibliography

Barry B 1989 *Theories of Justice*. University of California Press, Berkeley, CA

De-Shalit A 1995 *Why Posterity Matters*. Routledge, London

Dobson A 1998a *Justice and the Environment*. Oxford University Press, Oxford, UK

Dobson A (ed.) 1998b *Fairness and Futurity*. Oxford University Press, Oxford, UK

Gauthier D 1986 *Morals by Agreement*. Clarendon Press, Oxford, UK

Page E 1999 Intergenerational justice and climate change. *Political Studies* **47**: 53–66

Parfit D 1984 *Reasons and Persons*. Clarendon Press, Oxford, UK

Rawls J 1971 *A Theory of Justice*. Oxford University Press, Oxford, UK

Sikora R I, Barry B (eds.) 1978 *Obligations to Future Generations*. Temple University Press, Philadelphia, PA

Stearns J B 1972 Ecology and the indefinite unborn. *Monist* **56**: 612–25

P. J. Kelly

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Intergroup Relations, Social Psychology of

For purposes of this article, *intergroup relations* will be defined in accord with Sherif's (1962) classic definition of intergroup behavior, i.e., 'individuals belonging to one group interacting, collectively or individually, with another group or its members in terms of their group identification…' (p. 5). This definition highlights the subjective nature of intergroup situations in that it depends on the participants' cognitive representation of themselves as group members or as representatives of social groups or categories. Based on this conceptual definition, the operational definition of intergroup behavior is the presence of category-based discrimination—responding differentially to others depending on whether they belong to one's own social group ('ingroup') or to another group ('outgroup'). Such discrimination can range from preferential attitudes to group-based hostility and warfare. In this article, the study of intergroup discrimination is distinguished from the study of prejudice as intraindividual cognition (see *Stereotypes, Social Psychology of* and *Social Cognition and Affect, Psychology of*) or personality (see *Personality and Social Behavior*).

Within social psychology, theory and research on intergroup relations encompasses two major research traditions. One is the theory of social categorization and social identity which focuses on the processes underlying the formation of ingroup and outgroup differentiation and discrimination. The other is research on the 'contact hypothesis,' which focuses on understanding the conditions that promote or reduce intergroup bias and discrimination. In recent years, these two traditions have been merged into a more comprehensive perspective on the origins of intergroup discrimination and reduction of prejudice.

1. Social Categorization and the Ingroup–Outgroup Schema

One perspective on intergroup relations is provided by social identity theory, as articulated by Tajfel (1978) (see *Social Identity, Psychology of* and *Social Categorization, Psychology of*). In brief, social identity theory represents the convergence of two earlier traditions in the study of intergroup attitudes and behavior—social categorization (e.g., Doise 1978) and social comparison (e.g., Pettigrew 1967). The theoretical perspective rests on two basic premises:

(a) Individuals organize their understanding of the social world on the basis of categorical distinctions that transform continuous variables into discrete classes; categorization has the effect of minimizing perceived differences *within* categories and accentuating intercategory differences.

(b) Since individual persons are themselves members of some social categories and not others, social categorization carries with it implicit *ingroup–outgroup* (we–they) distinctions; because of the self-relevance of social categories, the ingroup–outgroup classification is a superimposed category distinction with affective and emotional significance.

These two premises provide a framework for conceptualizing any social situation in which a particular ingroup–outgroup categorization is made salient. In effect, the theory posits a basic *intergroup schema* with the following characteristic features:

(a) Assimilation within category boundaries and contrast between categories such that all members of the ingroup are perceived to be more similar to the self than members of the outgroup (the *intergroup accentuation* principle).

(b) Positive affect (trust, liking) selectively generalized to fellow ingroup members but not outgroup members (the *ingroup favoritism* principle).

(c) Intergroup social comparison associated with perceived negative interdependence between ingroup and outgroup (the *social competition* principle).

The affective and behavioral consequences of this schema lead to intergroup situations characterized by preferential treatment of ingroup members, mutual distrust between ingroup and outgroup, and intergroup competition. According to this theoretical perspective, the starting point for intergroup discrimination and prejudice is a cognitive representation of the social situation in which a particular categorical distinction is highly salient.

1.1 Behavioral Consequences of Ingroup–Outgroup Differentiation

The role of category salience in intergroup bias has been well documented in experimental research using the 'minimal intergroup paradigm' (Brewer 1979, Tajfel 1978). These experiments have demonstrated that merely categorizing individuals into two arbitrary, but distinct, social groupings is sufficient to elicit discriminatory evaluations and behavior (ingroup bias). The tendency to favor members of one's ingroup over outgroup members has been demonstrated in many different forms, including evaluative biases (attributing more positive characteristics to ingroup members than outgroup members), attributional biases (holding outgroup members more responsible for negative outcomes and giving ingroup members more credit for positive outcomes), allocation biases (distributing more positive rewards and credits to ingroup members than to outgroup individuals), and interaction biases (greater cooperation toward ingroup members and competition toward outgroup members).

Outside the laboratory, the apparently universal tendency for human beings to differentiate themselves

according to group membership has been widely recognized. In 1906, William Graham Sumner coined the term *ethnocentrism* to refer to people's attachment to ingroups and their preference for everything associated with the ingroup over that of outgroups. Ethnocentric attachment underlies group loyalty, adherence to ingroup norms, and trustworthiness in dealings with fellow group members. At the same time, it sets the stage for intergroup competition, hostility, and conflict (LeVine and Campbell 1972).

1.2 Theories of Ingroup Bias

Although ingroup–outgroup discrimination is the defining characteristic of intergroup behavior and documented in many forms, a central concern of intergroup relations researchers is understanding the cognitive and motivational processes that underlie and fuel ethnocentric preference and discrimination. Some theoretical perspectives on this issue come from outside social psychology. These include sociobiological theories that postulate a genetic disposition to behave differentially toward 'insiders' (kin and extended family) vs. 'outsiders' (nonkin) as a strategy of inclusive fitness (see *Evolutionary Social Psychology*), and various psychoanalytic theoretical perspectives derived from Freud's analysis of group psychology (see also Adorno et al. 1950).

Social structural theories emphasize the role of realistic conflict between groups derived from competition over scarce resources (e.g., Sherif 1966). The basic premise of realistic group conflict theory is that intergroup attitudes and behavior reflect group interests. Where these are incompatible—where what one group gains is at the expense of another—the social psychological response is negative intergroup attitudes and hostile behavior. When group goals are compatible, the intergroup atmosphere should be more positive, characterized by tolerance and amicability. Within a given sociopolitical system, group conflicts of interests are most likely to be manifest in struggles for recognition, status, and power (e.g., Sidanius and Pratto 1999). Hence, a substantial amount of research in the intergroup relations arena has been devoted to understanding the relationship between intergroup differences in status and power and intergroup discrimination and bias.

Although realistic conflict may provide a sufficient condition for the emergence of negative intergroup relations, most social psychological theories question the assumption that realistic competition is a *necessary* condition for ingroup favoritism and intergroup discrimination. These theories hold that ingroup–outgroup differentiation and associated preferences serve basic human motives that are independent of competition over tangible resources. The original social identity theory perspective held that achieving *positive distinctiveness* of one's ingroup in the context of intergroup social comparison serves to maintain and enhance *self-esteem*. This idea has given rise to the concept of *collective self-esteem*, self-worth derived from membership in valued groups. The relationship between self-esteem and ingroup bias has proved to be a complex one. Results of cumulative studies suggest that personal self-esteem and collective self-esteem may have different relationships to intergroup discrimination. Ingroup bias may be motivated by the need to restore or enhance collective self-esteem, but particularly for those individuals who are already high in personal self-esteem.

Another motivational theory of ingroup identification and intergroup discrimination gives primacy to the need to *reduce uncertainty* in social situations. According to this perspective, ingroup–outgroup differentiation provides a basis for assigning meaning and structure to otherwise ambiguous social information and provides the individual with guidance and rules for social behavior and decision making. Intergroup discrimination serves to enhance differentiation and achieve greater certainty, particularly when uncertainty has been aroused. Finally, yet another motivational theory of ingroup identification and differentiation from outgroups is based on the idea that social groups meet human needs for inclusion and differentiation as basic social motives. Ingroup membership provides for secure inclusion at the same time that intergroup boundaries serve the need for differentiation and distinctiveness.

2. Intergroup Contact and Cooperation

Whereas theories of ingroup identification and ingroup–outgroup differentiation provide a basis for understanding and explaining intergroup discrimination, another focus of social psychological research has been on identifying conditions for controlling or reducing the prejudicial effects of ingroup–outgroup distinctions. The so-called 'contact hypothesis' is actually a general set of ideas about reducing intergroup prejudice and discrimination that developed among social scientists in the 1940s in the context of inter-racial relations in the US. Although the general idea was expressed in earlier forms, credit for the full contact hypothesis is most frequently accorded to Gordon Allport as a consequence of his presentation in his influential book, *The Nature of Prejudice* (1954). The basic idea behind the hypothesis is that hostility between groups is fed by unfamiliarity and separation and that under the right conditions, contact among members of different groups will reduce hostility and promote more positive intergroup attitudes.

Specification of the 'right conditions' for positive contact constitutes a key part of the fully elaborated contact hypothesis. According to Allport (1954), the

four most important of these qualifying conditions were (a) integration has the support of authority, fostering *social norms* that favor intergroup acceptance, (b) the situation has high 'acquaintance potential,' promoting *intimate contact* among members of both groups, (c) the contact situation promotes *equal status* interactions among members of the social groups, and (d) the situation creates conditions of *cooperative interdependence* among members of both groups. Each of these conditions was derived from results of early research on racial desegregation and intergroup contact in the US, on which the hypothesis was initially based (Miller and Brewer 1984, Pettigrew 1998).

2.1 The Robber's Cave Experiment

Of the contact conditions specified by Allport, the one that has received the most attention and research since the 1950s is the stipulation regarding cooperative interdependence between members of different social groups in the contact situation. This focus is due in large part to the influence of the now classic field experiment conducted by Muzafer Sherif and his colleagues in the summer of 1954 in a boys' camp in Robbers Cave, Oklahoma (Sherif 1966). After successfully creating two rival groups at the campsite, the researchers tested conditions for reducing the intensity of intergroup conflict and ingroup bias among the campers. Having demonstrated that contact under neutral conditions alone failed to have any positive effects, the research team then introduced a series of *superordinate goals* that brought the two groups together under conditions of mutual interdependence and shared outcomes. The descriptive record and systematic measures taken by the researchers provide rich documentation of the effectiveness of cooperative interaction in reducing conflict and promoting cross-group friendships by the close of the summer camp experience.

From Robbers Cave onward, many field studies of intergroup contact have confirmed that intergroup cooperation leads to more friendliness and less ingroup bias than situations that do not promote or require cooperative interaction. Probably the most extensive application of the contact hypothesis has been the implementation of cooperative learning programs in desegregated school classrooms. Different methods of cooperative learning strategies have been devised, but all incorporate basic principles of the contact hypothesis (Miller and Brewer 1984). There is a sizable body of evidence from these settings that demonstrates the effectiveness of cooperative learning groups for increasing attraction and interaction between members of different social categories. Meta-analyses of studies in ethnically mixed classrooms confirm the superiority of cooperative learning methods over individualistic or competitive learning in promoting cross-ethnic friendships and reduced prejudice.

2.2 Laboratory Experiments: Defining the Limits of the Contact Effect

The elements of the Robbers Cave experiment also provided a prototype for subsequent laboratory experiments on the contact hypothesis and its moderating conditions. The basic laboratory paradigm is essentially a scaled-down version of the summer camp model. Participants in a laboratory session are first divided arbitrarily into two separate groups or social categories and given an opportunity to form distinct ingroup identities. The presence of ingroup bias and outgroup derogation is assessed at the end of this stage, and then members of the two groups are brought into contact under conditions that are experimentally manipulated to test features of the contact hypothesis and its underlying assumptions.

A brief review of these laboratory experiments identifies a number of factors that either inhibit or facilitate the effectiveness of contact to reduce ingroup–outgroup biases and promote positive attitudes toward outgroup members. Among the moderating variables confirmed by experimental studies are the frequency and duration of intergroup interaction, the presence of intergroup anxiety, the structure of cooperative tasks, the outcome of cooperation, and status equalization. In general, results of laboratory experiments confirm the premises of the contact hypothesis but also indicate the complexity—and potential fragility—of effects of intergroup contact even under highly controlled conditions.

In what is probably the most comprehensive laboratory test of inter-racial contact effects, Stuart Cook (see Miller and Brewer 1984) conducted a series of experiments in which highly prejudiced white subjects worked with a black confederate in an ideal contact situation (equal status, cooperative interdependence, with high acquaintance potential and egalitarian social norms) over an extended period of time. Perceptions of the black co-worker were measured at the completion of the contact experience, and general racial attitudes were assessed before, immediately after, and up to three years following the experimental sessions. Across all variations of this experiment, white participants displayed predominantly positive behaviors toward their black co-worker and expressed highly favorable evaluations in the postexperimental questionnaires. Whether liking for this individual member of the outgroup resulted in changed attitudes toward blacks and race-related issues, however, varied across the experiments and for different attitude measures.

One major reason why generalization fails is that the newly positively valued outgroup member is regarded as an exception and not typical or rep-

resentative of the outgroup in general. In Cook's studies, significant differences in postcontact attitude change among those who participated in the contact experience compared to control subjects were obtained only in an initial experiment in which what Cook referred to as a 'cognitive booster' was introduced during the course of the experiment. This added element was a guided conversation (led by a research confederate) in which the negative effects of discriminatory policies and practices were directly connected to the now-liked black co-worker. This booster served to make salient the co-worker's category membership and to establish a link between feelings toward this individual and members of the group as a whole. This explicit linkage appears to be a necessary mechanism for the effects of contact experiences to be generalized.

3. *Combining Categorization and Contact: Models for Reducing Intergroup Discrimination*

A significant advance toward a more integrative theory of intergroup relations was achieved when contact research was combined with concepts of social categorization and social identity theory to provide a theoretical framework for understanding the cognitive mechanisms by which cooperative contact is presumed to work (see Miller and Brewer 1984, Hewstone and Brown 1986). From the social categorization perspective, the issue to be addressed is how intergroup contact and cooperation can be structured so as to alter cognitive representations in ways that would eliminate one or more of the basic features of the negative intergroup schema. Based on the premises of social identity theory, three alternative models for contact effects have been developed and tested in experimental and field settings.

3.1 *Decategorization: The Personalization Model*

According to the premises of social identity theory, a primary consequence of salient ingroup–outgroup categorization is the depersonalization of members of the outgroup. Social behavior in intergroup situations is characterized by a tendency to treat individual members of the outgroup as undifferentiated representatives of a unified social category, ignoring individual differences within the group. This perspective on the contact situation implies that intergroup interactions should be structured so as to reduce the salience of category distinctions and promote opportunities to get to know outgroup members as individual persons.

The conditional specifications of the contact hypothesis (equal status, intimate, cooperative interaction) can be interpreted as features of the situation that reduce category salience and promote more differentiated and personalized representations of the participants in the contact setting. Attending to personal characteristics of group members not only provides the opportunity to disconfirm category stereotypes, it also breaks down the monolithic perception of the outgroup as a homogeneous unit. Repeated personalized contacts with a variety of outgroup members should, over time, undermine the value and meaningfulness of the social category stereotype as a source of information about members of that group. This is the process by which contact experiences are expected to generalize—via reducing the salience and meaning of social categorization in the long run.

The personalization model is supported by the early empirical evidence for the effects of extended, intimate contact on racial attitudes. More recently, extensive data on effects of intergroup friendships have been derived from surveys in Western Europe regarding attitudes toward minority immigrant groups. Across samples in France, Great Britain, the Netherlands, and Germany, Europeans with outgroup friends scored significantly lower on measures of prejudice, particularly affective prejudice (Pettigrew 1998). This positive relationship did not hold for other types of contact (work or residential) that did not involve the formation of close personal relationships with members of the outgroup.

3.2 *Recategorization: The Common Ingroup Identity Model*

The second social categorization model of intergroup contact and prejudice reduction is also based on the premise that reducing the salience of ingroup–outgroup category distinctions is key to positive effects. In contrast to the decategorization approach, however, recategorization is not designed to reduce or eliminate categorization. Instead, the goal is to create a superordinate category that encompasses both ingroup and outgroup in a single social group representation. This approach to the reduction of intergroup discrimination is known as the Common Ingroup Identity Model (Gaertner et al. 1993). It is based on the premise that when a superordinate category has been made salient, group members are more likely to think of themselves as one unit rather than two separate groups. When this form of recategorization is successful, ingroup loyalties are transferred from the original subgroups to the common social group as a whole.

The common ingroup identity model has been tested extensively in laboratory experiments assessing the conditions under which two previously segregated work groups can be successfully merged in a superordinate team. The experimental manipulations in these studies focus on situational variables that reduce the perceptual salience of ingroup–outgroup categorizations during the cooperative contact. Conditions

that enhance the salience of the common team identity and reduce the salience of subgroup categories are found to diminish or eliminate ingroup–outgroup biases in the evaluation of fellow team members. To the extent that participants perceive the combined team as a single entity, rather than two separate groups, evaluations of former outgroup members become more positive.

3.3 The Mutual Differentiation Model and Dual Identities

Although the structural representations of the contact situation advocated by the decategorization (personalization) and recategorization (common ingroup identity) models are different, the two approaches share common assumptions about the need to reduce category differentiation and associated processes. In fact, the underlying processes may be complementary in that personalized interactions facilitate common ingroup identity and shared ingroup membership promotes positive interpersonal relationships. Because both models rely on reducing or eliminating the salience of intergroup differentiation, they involve structuring contact in a way that will challenge or threaten existing social identities. Both cognitive and motivational factors conspire to create resistance to the dissolution of category boundaries or to re-establish category distinctions across time. Although the salience of a common superordinate identity or personalized representations may be enhanced in the short run, these may be difficult to maintain across time and social situations.

Pre-existing social–structural relationships between groups may also create strong forces of resistance to changes in category boundaries. Cognitive restructuring may be close to impossible (at least as a first step) for groups already engaged in deadly hostilities. Even in the absence of overt conflict, asymmetries between social groups in size, power, or status create additional sources of resistance. When one group is substantially numerically smaller than the other in the contact situation, the minority category is especially salient and minority group members may be particularly reluctant to accept a superordinate category identity that is dominated by the other group. Another major challenge is created by pre-existing status differences between groups, where members of both high and low status groups may be threatened by contact and assimilation.

These challenges to processes of decategorization/recategorization led Hewstone and Brown (1986) to recommend an alternative approach to intergroup contact wherein cooperative interactions between groups are introduced without degrading the original ingroup–outgroup categorization. In order to promote positive intergroup experience, Hewstone and Brown recommended that the contact situation be structured so that members of the respective groups have distinct but complementary roles to contribute toward common goals. In this way, both groups can maintain positive distinctiveness within a cooperative framework.

Although ingroup–outgroup category salience is usually associated with ingroup bias and the negative side of intergroup attitudes, cooperative interdependence is assumed to override the negative intergroup schema, particularly if the two groups have differentiated, complementary roles to play. Because it capitalizes on needs for distinctive social identities, the mutual differentiation model provides a solution that is highly stable in terms of the cognitive-structural aspects of the intergroup situation. On the other hand, this model of intergroup contact also reinforces the perception that the two groups are distinctly different and thus perpetuates ingroup–outgroup differentiation, and the potential for fission and conflict along group lines remains high.

To reduce the impact of ingroup–outgroup distinctions, some social psychologists have recommended combining the mutual differentiation and common ingroup identity models—creating contact conditions that preserve category identities at both subgroup and superordinate group levels. When participants simultaneously see themselves as members of different groups but also as part of the same superordinate category, intergroup relations between subgroups are more positive than in the absence of the common identity. Moreover, dual identities may also lead to even more positive outgroup attitudes than those associated with a superordinate identity alone.

Research on the dual identity model of intergroup contact challenges the assumption, implicit in much previous work in intergroup relations, that social identities are mutually exclusive, with only one social categorization (ingroup–outgroup differentiation) salient at any one time. New research is beginning to explore the implications of holding multiple group identities, or identities at different levels of inclusiveness, simultaneously. The hope is that a better understanding of the combined effects of mutual differentiation, cross-categorization, and recategorization will provide the social psychological underpinnings for successful pluralism in multicultural societies.

See also: Cooperation and Competition, Psychology of; Discrimination; Discrimination: Racial; Marital Interaction: Effects on Child Development; Social Categorization, Psychology of; Social Identity, Psychology of; Stereotypes, Social Psychology of

Bibliography

Adorno T W, Frenkel-Brunswick E, Levinson D J, Sanford R 1950 *The Authoritarian Personality*, 1st edn. Harper, New York

Allport G W 1954 *The Nature of Prejudice*. Addison-Wesley, Cambridge, MA
Brewer M B 1979 In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin* **86**: 307–24
Brewer M B, Brown R J 1998 Intergroup relations. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, MA, Vol. 2
Brewer M B, Miller N 1996 *Intergroup Relations*. Open University Press, Buckingham, UK
Brown R J 1995 *Prejudice: Its Social Psychology*. Blackwell, Oxford, UK
Doise W 1978 *Groups and Individuals: Explanations in Social Psychology*. Cambridge University Press, Cambridge, UK
Gaertner S L, Dovidio J F, Anastasio P A, Bachman B A, Rust M C 1993 The Common Ingroup Identity model: Recategorization and the reduction of intergroup bias. In: Stroebe W, Hewstone M (eds.) *European Review of Social Psychology*. Wiley, Chichester, UK, Vol. 4, pp. 1–26
Hewstone M, Brown R (eds.) 1986 *Contact and Conflict in Intergroup Encounters*. Blackwell, Oxford, UK
LeVine R A, Campbell D T 1972 *Ethnocentrism: Theories of Conflict, Ethnic Attitudes and Group Behavior*. Wiley, New York
Miller N, Brewer M B (eds.) 1984 *Groups in Contact: The Psychology of Desegregation*. Academic Press, Orlando, FL
Pettigrew T F 1967 Social evaluation theory: Convergences and applications. In: Levine D (ed.) *Nebraska Symposium on Motivation, Vol. 15*. University of Nebraska Press, Lincoln, NE
Pettigrew T F 1998 Intergroup contact theory. *Annual Review of Psychology* **49**: 65–85
Sherif M 1962 *Intergroup Relations and Leadership*. Wiley, New York
Sherif M 1966 *In Common Predicament*. Houghton Mifflin, Boston
Sidanius J, Pratto F 1999 *Social Dominance*. Cambridge University Press, Cambridge, UK
Sumner W G 1907 *Folkways*. Ginn, Boston
Tajfel H 1978 *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations*. Academic Press, London

M. B. Brewer

Internal Migration: Developing Countries

1. Introduction

There is widespread agreement among students of population on both the importance of migration for social and demographic dynamics and the difficulty that, by its multifaceted nature, its study raises for the researcher. The intensification of migratory movements always implies societal transformations, including the accentuation of regional and social inequalities, modifications in patterns of population distribution, and changed demographic profiles in both origin and destination areas. Migration studies, however, are complex. The definition of the term itself and its measurement and interpretation are not straightforward. Depending on the type of migration and the focus of the analysis, very diverse theoretical approaches have been used.

As a component of demographic growth, migration plays a fundamental role in the rapid growth of developing countries' large cities, as well as producing significant alterations of the population's distribution over its territory. Migration is responsible, therefore, for a considerable part of the demographic concentration which characterizes many of these countries.

All of the studies concerned with measuring the role of migration show that, to a great extent, it was the volume of movements from rural to urban areas which explains the intense rhythm of urban population increase in the second half of the twentieth century. However, considering its multiple facets, internal migration is not restricted to rural–urban movements, since other types may emerge and grow in importance as a result of changes in economic and sociocultural contexts and in spatial configurations, as well as resulting from specific historical moments of differing countries. This is why, especially in Latin America, types of movements typical of developing countries, such as rural–urban or rural–rural migration, coexist in the same territories with types observed in developed countries, such as urban–rural and urban–urban migration. These movements reflect new forms of organization and structuring of space.

2. Definition and Data Sources

The difficulties of studying migration begin with its definition. Unlike mortality and fertility, this demographic phenomenon has no single definition, but varies according to historical context and the spatial and temporal referents chosen. In fact, even when migration is considered, in a general way, as the move of an individual from one geographic unit (generally defined in administrative terms) to another and as involving a change in usual residence (United Nations 1970, Bilsborrow 1998b, Oucho 1998, Courgeau 1988), several questions remain for such a definition to be used empirically. On the one hand, the type of geographic unit and the boundaries which define it must be specified. The rural/urban dichotomy is one of the most commonly used distinctions, although movements between municipalities, states, or large regions are also important. From the point of view of the spatial dimension, then, migration admits several definitions.

As if this were not enough, another element must be considered for a correct definition of migration: time (Bilsborrow 1998b). What constitutes a change of residence? How long must a person remain in the place of destination to consider the move as definitive and, therefore, a change of residence? These are not easy questions to answer, a situation reflected in another

Table 1
Proportion of population residing in urban areas, by major world areas, 1975–2015

Major area	Percent of urban population		
	1975	1995	2015
World	37.7	45.2	55.5
Africa	25.2	34.4	47.2
Asia	24.6	34.6	47.8
Latin America and the Caribbean	61.3	74.2	82.1
Europe	67.1	73.6	80.1
Northern America	73.8	76.3	81.9
Oceania	71.8	70.3	72.1

Source: Chen et al. 1998

Table 2
Percent of urban growth in developing countries due to migration or reclassification, 1960–80

Major area	Percent of urban growth due to migration or reclassification		
	1960s	1970s	1980s
Developing countries	40.3	44.1	54.3
Africa	41.2	40.6	24.9
Asia	40.4	46.7	63.6
Latin America and the Caribbean	40.1	40.5	33.9

Source: Chen et al. 1998

difficulty with migration: its measurement. Although international organizations such as the United Nations endeavor to convince countries to adopt minimum criteria which would guarantee the comparability of data, through recommendations and manuals (United Nations 1970), the measurement of migration is problematic. This is so, not only because of the considerable variation in the number and size of territorial divisions among and within countries, but also because of the various possible movements which these divisions may capture.

Classic data sources, such as the demographic censuses, due to space limitations in the questionnaires and to financial constraints, must generally decide which type of question or territorial division to use for the migration question, when and if it is possible to include this dimension in the questionnaire. There are other problems as well, such as the temporal gaps of censuses, precarious data publication in many countries (Oucho 1998, Chen et al. 1998, Bilsborrow 1998b, Lattes 1998), and financial and operational difficulties for carrying out specific and periodic surveys, sources which would be the most adequate to obtain a better understanding of the phenomenon, especially of its determinants and consequences (Bilsborrow et al. 1984).

Problems of periodicity, availability, quality, and publication of data have meant that much of the information on migration is obtained through the use of indirect demographic estimation techniques, a fact which not only limits analytic possibilities but also affects the reliability and precision of results.

3. *The Role of Rural–Urban Migration in the Urbanization Process*

The twentieth century ended with more than half of the world's population living in urban areas. Nevertheless, there remain considerable differentials among countries in this respect, especially between developed and developing countries. While European and North American countries have proportions of urban population over 70 percent, and are likely to pass 80 percent by 2015 according to United Nations estimates (Chen et al. 1998), in Africa these values are under 35 percent and will not reach 50 percent even by 2015. There are, of course, important differences among developing countries. In Latin America, the region with the highest level of urbanization—especially in countries such as Argentina, Brazil, Uruguay, Chile, and Venezuela—close to 80 percent of the population live in cities. Table 1 presents estimates of the proportion of urban population residing in major world areas.

Comparing data on rural and urban population is problematic, however, and precaution should be taken, since this definition depends on each country (Bilsborrow 1998b, Lattes 1998). There is no doubt, however, that the developing world is on the path of increasing concentration of population in cities.

In this continuous process of urbanization, internal migration of rural origin plays an important role, especially considering that the higher level of rural fertility favors greater natural growth in these areas, therefore having an inverse effect. Authors such as Chen, Valente, and Zlotnik (1998), Lattes (1998), and Lucas (1997) leave no doubt as to the preponderant weight of migratory movements from rural areas to the growth of cities in developing countries.

Table 2 shows much variation, however, in the relative contribution of migration and reclassification to urban growth in developing countries. Unfortunately, in view of the lack of adequate data, existing estimates of the effect of migration on urban growth, having been obtained by residual methods—the difference between total observed growth and estimated natural growth—do not permit separating the effect of reclassification of rural to urban areas, which also contributes to the increase of urban population. For the majority of countries for which data exist, the combined contribution of migration and reclassification is greater than 40 percent. Lucas (1997) shows that for a sample of 29 countries this proportion reached 41.4 percent in the 1960s.

The more countries urbanize and the more complex and heterogeneous are their social, economic, cultural, and political structures, the more diversified are the types of population mobility. As Bilsborrow notes, the emphasis of migration research on the rural–urban type reveals 'a severe case of myopia,' since 'this bias is inconsistent with reality' (Bilsborrow 1998b, p. 8). Indeed, besides the several variants of internal migration in terms of the temporal dimension (permanent or short-term moves—seasonal, temporary, and circular (see Bilsborrow 1998b)), the spatial perspective also leads to other considerations, such as migration of urban origin.

In fact, phenomena such as 'counter-urbanization' (Champion 1989, Frey 1988, Lattes 1998), 'rural rebound' (Johnson 1999), and 'nonmetropolitan turnaround' (Fuguitt 1985), typical of developed countries, are already felt in developing countries, especially in Latin America. In addition, there are phenomena typical of developing countries. On the one hand, the settlement of frontier regions implies migration to rural areas, while on the other hand, demographic pressure in stagnating areas, concentration of land ownership and technological and organizational changes in agriculture dislocate population from rural zones. At the same time, the peripheral growth of metropolitan areas encompasses a large variety of urban–urban and urban–rural movements, which become increasingly visible and significant in these countries, especially as urbanization reaches high levels.

Rural–rural movements should not be underestimated either, least of all in some countries of Asia and Africa, where the proportion of population living in rural areas continues to be high. Skeldon (1986) shows, for example, that for India, 57.4 percent of those who had moved in the 10 years before the 1981 census did so from one rural area to another, while other types of movement were less important: rural–urban (19.5 percent), urban–urban (15.2 percent), and urban–rural (7.9 percent). Refugee movements, driven by war and environmental disaster, continue to displace populations, often resulting in permanent dislocations; such movements are rarely examined by students of internal migration.

For the largest cities, it is not only the movements originating in the less developed regions which are important. Both intra-metropolitan moves, responsible for a major part of the growth of urban peripheries (Cunha 1998), and commuting, whose growth reflects growing distances between residence and workplace, are significant features of urban life in developing countries.

4. *Characteristics and Conditioning Factors of Migration*

In view of the multifaceted character and the many forms of the migratory phenomenon, its characteristics and conditioning factors are also considerably varied, depending, on the one hand, on whether one adopts a macro or micro level of analysis, and on the other hand, the historical moment and the economic, social, political, and cultural context in which migration occurs. Thus, conditioning factors will vary and characteristics of population movements between distant regions of a country, with different degrees of development, will certainly be distinct from those referring to moves between municipalities of a metropolitan area.

Although in terms of age and sex there are regular references in the literature concerning the tendency for migration to be concentrated among young adults, in general between the ages 15–30 years, and concerning a greater proportion of women in rural–urban streams (United Nations 1993), these generalizations are very abstract and therefore incapable of capturing the entire range of possibilities encompassed by this complex phenomenon. There are social-structural factors which condition movements, as well as specific cultural aspects and geographic realities (the distances involved in the migration, for example). Such elements as life cycle, the composition and forms of family organization and characteristics of origin and destination areas are fundamental elements for determining who moves and who stays.

In terms of migration by age, selectivity reveals the importance of economic factors, especially those associated with the forms and possibilities of the insertion of the individual in the local labor market (both in terms of employment and remuneration). The greater predominance of young adults reveals a greater incidence of movements of one person or of couples in early stages of their life cycles. From the point of view of destination areas, migration will depend also on pull factors. For large urban centers, for example, selectivity will be affected by the number of opportunities available in the formal and informal labor markets.

With the restructuring of the world economy, with the accompanying flexibilization of production and employment and the ever more intense use of technology and information (Castells 1996, Harvey 1990), the employment/migration relation will not be as strong as in the past. The levels of wealth in urban areas, creating a demand for unskilled personal services, will continue to attract migrants, able to survive on irregular odd jobs, with no prospect for real employment. On the other hand, in agricultural frontier areas, for example, family migration is more likely to predominate, which would alter the age structure of migration, making it less selective.

As for sex, the available data (United Nations 1993) show that although women predominate in rural–urban migration streams, this tendency is not so clear when data from different countries are observed, especially controlling for age. Even in African countries, where the predominance of male migration is

clear, controlling for age makes a difference: in the 10–19 and over 45 categories, women dominate, while in the highly productive ages, it is men who migrate more (Singelmann 1993). The fact that younger women represent greater proportions may be explained by the differences between the ages of men and women within marriage; for older women the data probably reflect the effect of differential mortality by sex which, as is well-known, favors women. This means that for the sex variable also, migration characteristics depend greatly on the contexts in which they occur and any generalization may be, to say the least, simplistic.

In addition to factors linked to the needs and preferences of migrants or to the structural conditions of origin and destination areas, the role of social networks must also be considered among the major conditioning factors of migration. Networks, of fundamental importance in the direction and volume of streams, supply information on destinations; reduce risk; and facilitate adaptation of the individual to the new area.

5. *Conclusions*

The late twentieth century was marked by immense population movement between regions of developing countries, especially by rural–urban migration. A slowdown in technological development in these countries at the century's end, coupled with increasing productivity in agriculture, point to a change of pace in these movements. According to Brockerhoff (1999, p. 772): 'the sluggish performance of manufacturing (compared to agriculture) remains largely responsible for the observed slower pace of urban growth in developing countries, and may have decelerated urban growth from what otherwise would have been higher rates in the 1980s and 1990s by curbing net rural-to-urban migration.'

The beginning of the twenty-first century is marked by a radically altered situation. Globalization, in both its economic and sociocultural aspects, together with the demographic transition and the environmental crisis, have greatly affected the factors which condition population mobility. The most intense phase of urbanization has been completed in many countries. The mobility and flexibility of multinational capital have weakened the relation between place of residence and place of work. New forms of mobility become more important: commuting, tourism, circular migration, and seasonal migration, in addition to the increase in short-distance, intermunicipal moves. International migration, which in the nineteenth century implied a permanent move of large populations, has become an option not radically different from internal migration, and nearly as reversible. The rapid population growth phase of the demographic transition, which contributed to internal migration, has also given way to slower growth in many countries, reducing pressures in places of origin. Finally, an 'empty world' has been substituted by a 'full world' and the resulting environmental restrictions will be ever more important conditioning factors of population mobility (Hogan 1993).

Policy responses to internal migration in developing countries have varied widely. Desires to colonize new agricultural regions, to promote industrialization or to curb what is perceived as excessive growth of 'megacities' have led governments to adopt a wide range of policies. The success of such policies has often been questioned, although Brockerhoff (1999) observes that countries which expressed an intention to effect policies for slower urban growth, such as Mexico, Egypt, and India, have had lower urban growth rates than their global subregions and especially lower than their neighbors. In the coming period of slower demographic growth, policy initiatives may be more generally successful.

The diversity which is currently observed in the world demographic situation has no historical precedent. Rates of internal migration and urbanization levels are extremely varied, but the tendencies in all major regions point to convergence and a less varied situation in the future.

Bibliography

Bilsborrow R E (ed.) 1998a *Migration, Urbanization and Development: New directions and issues*. UNFPA/Kluwer Academic Publishers, Norwell, MA

Bilsborrow R E 1998b The state of the art and overview of the chapters. In: Bilsborrow R E (ed.) *Migration, Urbanization and Development: New directions and issues*. UNFPA/Kluwer Academic Publishers, Norwell, MA

Bilsborrow R E, Oberai A S, Standing G 1984 *Migration Surveys in Low-income Countries: Guidelines for survey and questionnaire design*. Croom Helm, London

Brockerhoff M 1999 Urban growth in developing countries: A review of projections and predictions. *Population and Development Review* **25**(4): 757–78

Castells M 1996 *The Rise of the Network Society*. Blackwell, Oxford, UK

Champion A G (ed.) 1989 *Counterurbanization: The changing pace and nature of population deconcentration*. Arnold, London

Chen N, Valente P, Zlotnik H 1998 What do we know about recent trends in urbanization? In: Bilsborrow R E (ed.) *Migration, Urbanization and Development: New Directions and Issues*. UNFPA/Kluwer Academic Publishers, Norwell, MA

Courgeau D 1988 *Méthodes de mésure de la mobilité spatiale: migrations internes, mobilité temporaire, navettes*. [*Methods for measuring spatial mobility: internal migration, temporary migration, commuting*]. Editions de Institut National d'Études Démographiques, Paris

Cunha J M P 1998 New trends in urban settlement and the role of intraurban migration: The case of São Paulo/Brazil. In: Bilsborrow R E (ed.) *Migration, Urbanization and Development: New directions and issues*. UNFPA/Kluwer Academic Publishers, Norwell, MA

Frey W H 1988 Migration and metropolitan decline in developed countries: A comparative study. *Population and Development Review* **14**(4): 595–628
Fuguitt G V 1985 The nonmetropolitan population turnaround. *Annual Review of Sociology* **11**: 259–80
Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK
Hogan D J 1993 Population Growth and Distribution: their relations to development and the environment. Background Paper DDR/5. United Nations, CELADE, Santiago
Johnson K 1999 The rural rebound. *Reports on America* 1, 3. Population Reference Bureau, Washington, DC
Lattes A 1998 Population Distribution in Latin America: Is there a trend towards population deconcentration? In: United Nations, Department of Economic and Social Affairs, Population Division. *Population Distribution and Migration*. United Nations, New York
Lucas R E B 1997 Internal migration in developing countries. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam
Oucho J O 1998 Recent internal migration processes in Sub-Saharan Africa: Determinants, consequences, and data adequacy issues. In: Bilsborrow R E (ed.) *Migration, Urbanization and Development: New directions and issues*. UNFPA/Kluwer Academic Publishers, Norwell, MA
Singelmann J 1993 Level and trends of female internal migration in developing countries, 1960–1980. In: United Nations *Internal Migration of Women in Developing Countries*. United Nations, New York
Skeldon R 1986 On migration patterns in India during the 1970s. *Population and Development Review* **12**(4): 759–79
United Nations 1970 *Methods of Measuring Internal Migration. Manuals on Methods of Estimating Population, Manual VI*. United Nations, New York
United Nations 1993 *Internal Migration of Women in Developing Countries*. United Nations, New York

D. J. Hogan and J. M. Pinto da Cunha

Internal Migration (Further Types): Industrialized Countries

Residential mobility and migration are pervasive facts of life in most industrialized countries, where migration is a major mechanism through which labor resources are redistributed geographically in response to changing economic and demographic forces. The migration literature may be classified into two areas of research, one dealing with the *determinants* and one with the *consequences* of migration. Some studies treat both the determinants and consequences within the same framework. Migration research is strongly oriented toward the former.

The *determinants* of migration are the factors that affect migration, including characteristics both of places and of persons and their families. The term refers to the qualitative and the quantitative importance of each factor. Place characteristics are specific to a given area, such as employment and wage opportunities, the presence of family and friends, and location-specific amenities. Personal and family characteristics help shape individual and family responses to opportunities that may exist at different locations. The *consequences* of migration refer both to the performance of migrants in their new locations relative to a benchmark (their presumed performance in their former place) and to the impacts that migrants have on others in sending and receiving areas. Traditionally, research has addressed several questions:

(a) Who migrates?
(b) Why do these people migrate?
(c) Where are the migrants coming from and where are they going?
(d) When do they migrate?
(e) What consequences result from migration?

Many studies have aimed primarily at describing migration.

1. *What is Migration?*

The United Nations manual, *Methods of Measuring Internal Migration* (1970), has served as the basis for several definitions of migration and migrants. It proposes the following (p. 2):

> A migration is defined as a move from one migration-defining area to another (or a move of some specified minimum distance) that was made during a given migration interval and that involved a change of residence. A migrant is a person who has changed his usual place of residence from one migration-defining area to another (or who moved some specified minimum distance) at least once during the migration interval.

2. *Comparing Internal Migration Rates in Industrialized Countries*

In most countries people migrate from one place to another in pursuit of increased utility resulting from better employment opportunities, higher wages, and many other factors, but international comparisons of internal migration rates are difficult to make for several reasons.

Various countries define migrants differently in terms of whose movement is being considered, in terms of the type of border they must cross, and in terms of the interval over which the movement occurs.

The method of measuring migration differs widely. In some countries, censuses are the main source of information concerning migration, whereas in others, population registers and other types of administrative records are the major source.

The size and shape of the spatial areas between which migration is measured are not uniform either within or between countries.

Migration propensities are sensitive to national economic and demographic conditions, and conse-

quently interregional migration propensities could differ for behavioral reasons.

The most unambiguous method of making cross-national comparisons is to focus on residential moves, but residential mobility includes more than what is commonly regarded as migration. Nevertheless, such measures suggest that rates of mobility vary widely across industrialized countries. Movement rates are high for Australia, Canada, New Zealand, and the USA. A second group of industrialized countries that is far behind these four includes France, the UK, Israel, Japan, Sweden, and Switzerland. Belgium, Ireland, and The Netherlands have the lowest rates of residential mobility.

3. *The Determinants and Consequences of Migration—Overview*

3.1 *Migration Patterns*

Many factors contribute to the decision to migrate. Differential characteristics of sending and receiving regions provide a potential incentive for moving, and individual and/or family traits help condition the responses to utility differentials that may arise from these different characteristics. An individual's personal characteristics—including accumulated job skills, general labor market conditions, and employment composition—will help determine the probability of gaining employment during a period of job search. Conditions in land and housing markets also may be important, and state and local taxes and the associated availability of public goods may be critical for certain potential migrants.

A number of lifecycle considerations—such as marriage, divorce, completion of schooling, entry into the labor force, start of a career, birth, aging, leaving home of children, home ownership, and retirement—are critical in an individual's or a family's decision to migrate. Other personal circumstances, often related to the lifecycle, are also important, such as employment status, earnings, education, accumulated skills, age, job tenure, sex, and health.

One of the most universal mobility relationships is that between age and migration. Migration propensities peak during the early to mid-20s and then decline steadily, with a slight upturn at retirement age in some countries. For example, in the USA during the 1980s a person aged 25–29 with four years of college education was over twice as likely to make an interstate move as his or her counterpart aged 35–44; over three times more likely to make such a move than someone with comparable education aged 45–64; and over three times more likely to make such a move than a peer in age who had less than eight years of education.

Studies of the determinants of migration commonly have been formulated in the context of individual utility maximization, with some attention given to the family or the household as the decision-making unit. Models based upon such behavioral foundations have frequently been estimated with aggregate data relating both to migrants and to the determinants of migration. Before the general availability of micro and longitudinal data, virtually all applied migration research was of necessity based on aggregate data. Although aggregate data were and are limiting in many respects, they did not prevent a boom in migration research during the 1960s and 1970s. Even in the presence of many microdata sets, aggregate data are frequently studied today. Not only are aggregate trends and tendencies of interest in their own right, but for many countries, such aggregate data are all that is available.

During the 1960s, the main thrust of migration research began to take on a more formal tone. Most of the research was not formal in a theoretical sense, but rather intuitively generated hypotheses at first were tested formally in an econometric sense with aggregate data, typically but not always with place-to-place migration data. Such models were frequently specified in the context of modified gravity models. The models are 'gravity type' in that migration is hypothesized to be directly related to the size of relevant origin and destination populations, and inversely related to distance. The models are 'modified' in the sense that the variables of the basic gravity model are given behavioral content, and additional variables that are expected to influence migration are included in the estimated relationship. These studies showed clearly that migration diminishes rapidly with increased distance and that migration is greater between more populous regions, but migration did not always appear to be from low to high wage regions nor from those with high unemployment rates to those with low rates.

3.2 *Theoretical Perspectives on Migration*

In almost all migration research conducted by economists prior to the late 1970s, the theoretical perspective taken was that of a disequilibrium system. The perspective is called 'disequilibrium' because migration is assumed to be driven by the existence of a set of non-market clearing regional wages. Spatial variations in wages or earnings are assumed to reflect opportunities for utility gains. During more recent years, this disequilibrium perspective has been challenged by proponents of the equilibrium hypothesis, which assumes that spatial variations in wages are compensating and therefore do not reflect opportunities for utility gains.

According to the human capital approach, the potential migrant will select the locality at which the real value of the expected net benefit that accrues to him from migration is greatest. The income that the individual expects to earn at each alternative destination enters importantly into his judgment concerning the benefits associated with each location. The

relevant income measure for the individual to consider is the present discounted value of his expected future stream of net pecuniary returns.

Due in part to the fairly consistent tendency for empirical studies based on aggregate data to fail to confirm the importance of wages or income in migration decisions, the equilibrium approach has been offered as an alternative to the traditional disequilibrium perspective described above. The equilibrium theorists begin by assuming that households and firms are in proximate equilibrium at any point in time. This assumption means that the marginal household and firm, while maximizing utility and profit, respectively, are spatially arrayed so as to receive zero consumer and producer surplus from their location. Thus, any movement from the general equilibrium configuration cannot improve utility or profit.

According to the equilibrium approach, changes in lifecycle factors or generally rising real incomes continuously change the demand for consumer amenities. Because amenities are not evenly distributed spatially, migration occurs and quickly reequilibrates households. Net in-migration to amenity-rich areas tends to drive down wages and drive up the prices of locally produced goods and services and land. In amenity-poor areas, opposite patterns of change occur. Wages and local prices diverge across regions until they just compensate households for the differing amenity bundles that the various regions supply.

Family ties result in negative personal externalities that are usually internalized by the family and thus tend to discourage migration. 'Tied persons' in the family are 'those whose gains from migration are (in absolute value) dominated by gains (or losses) of the spouse' (Mincer 1978, p. 753). Presuming that their joint net returns on migrating from *i* to *j* exceed their joint net costs of migrating, a couple would presumably migrate from *i* to *j*. If, for example, the wife's expected earnings in *j* were less than in *i*, but the husband's were sufficiently greater in *j* than in *i* to offset these losses, the wife would be a 'tied mover.' On the other hand, if the husband's earnings gain in *j* were to fail to offset his wife's earnings loss in *i*, the couple would remain in *i*, and the husband would be a 'tied stayer.' Moreover, according to Mincer, such ties tend to reduce the employment and earnings of those wives who do migrate and to increase the employment and earnings of their husbands. Mincer goes on to show that increased labor force participation rates of women cause an increase in migration ties, which results in both less migration and more marital instability. Increased marital instability in turn encourages migration as well as increased women's labor force participation.

3.3 Consequences of Migration

The consequences of migration have been addressed at two levels. The first deals with the migrants themselves, where the emphasis has been on the benefits to migrating, often measured in terms of earnings gains. The second deals with migration's impact on others in the origin and the destination. Do migrants depress local wages in receiving areas and displace local residents from jobs? To the extent that migrants tend to be young and well-educated, does migration deprive source regions of critically needed human capital that ensures these regions of long periods of economic stagnation?

4. Conclusions

4.1 Pressing Questions

Who migrates? Why do they migrate? Where do the migrants come from and where do they go? When do they migrate? What consequences does migration have for the migrants themselves and others in origin and destination regions? During the last 30 years, firm answers have been provided to many of these questions, but at the same time new, narrower, and in certain ways more refined, interesting, and certainly more challenging questions have arisen.

Migration propensities appear to vary considerably across industrialized countries for reasons that are not always obvious. Moreover, estimated parameters that address the questions raised above, while generally in qualitative agreement, are quantitatively somewhat different across industrialized countries, even when the models and data are roughly comparable. With regard to internal migration, rigorous crossnational studies are virtually nonexistent, but differences in geographic size (and hence spatial distribution of economic opportunities) and culture are likely to importantly underlie observed differences.

4.2 The Human Capital Model

Until about 20 years ago, aggregate data were used almost exclusively to study various migration phenomena. Such data embody a number of shortcomings that prevented the study of many important issues bearing on migration. The relatively recent availability of micro- and longitudinal data have had a major impact on four areas of migration research.

(a) Such data have gone far toward clearing up earlier puzzles concerning the relationship between unemployment and migration.

(b) They have allowed the human capital model to be tested in the migration context by allowing the estimation of migrant earnings equations.

(c) They have permitted a clarification of the relationship between personal characteristics and the decision to migrate, and they have allowed a deeper understanding of the relationship between various lifecycle and familial factors and migration.

(d) They have permitted a detailed focus on different types of migrants, particularly primary, return, and other repeat migrants.

Researchers now know that local unemployment in the USA has a significant influence on the migration decisions of the unemployed and those who are seeking new jobs, but has little influence on individuals who are secure in their jobs. However, they do not know why this relationship fails to hold for European countries. Moreover, they do not know with great confidence why some unemployed individuals are quicker to migrate than others. To some extent, immediate migration could be discouraged if a person's spouse remains employed. What characteristics of a spouse's employment discourage migration? To what extent does it matter whether the party remaining employed is the husband or the wife? What role does the availability of unemployment insurance play in discouraging immediate migration? Institutional settings differ substantially by country, such as in the nature of unemployment insurance programs, and could cause differences in the response to personal unemployment.

The human capital model has provided a powerful analytical tool for the study of numerous important issues. It is somewhat surprising that this model has not typically provided a comparably powerful explanation of internal migration. Whether migrants are compared to otherwise comparable individuals in the origin or the destination, they appear to suffer earnings losses for at least several years after migration. This finding has been verified repeatedly for the USA and for other countries. Although several explanations have been offered for these findings, we still do not know enough about why they occur. The most likely explanation for the apparent failure of the human capital model is that the model has not failed, but rather it has just not been properly or fully implimented.

4.3 Lifecycle Impacts on Migration

The influences of lifecycle changes on migration decisions have only barely been touched by researchers. The relationships between many such changes and migration therefore have remained undiscovered. Completion of education; birth and aging of children and the anticipation of these events; changing marital status and especially divorce; death of a spouse; acquisition of a home; retirement—these are only a few potentially important lifecycle changes that are likely to importantly affect migration in industrialized countries and about which much remains to be learned.

Little is known about the interactions between migration and fertility in the contemporary USA or in other industrialized countries. A sizable literature on this topic began developing in the 1930s, but was mainly focused on rural-to-urban migration. A good deal of more recent work has also focused on rural-to-urban migrants in less developed nations, and a few studies concentrate on immigrants to the USA. However, in spite of the availability of more and better data in recent years, little has been done to study how the act of migration affects fertility in the USA or in other industrialized countries.

4.4 Migration Policies

Migration serves an equilibrating function in the economy, expediting the balancing of demand and supply forces within and across regions, and thus facilitating the operation of market economies. This facilitating role of migration has attracted the attention of policy makers. In a number of European countries, such as the UK, France, and The Netherlands, migration policy has been tied directly to regional development policy; in others, such as Sweden, migration policy has been oriented toward improving the efficiency of the labor market. In either case, a major objective has been to reduce unemployment. Subsidization of relocation expenses and employment information exchanges are the most common forms that migration policy has taken in Europe.

In the USA, policy concerns regarding migration generally are passive, although this was not the case during the nineteenth century. These concerns have been directed toward both how various national, state, and local policies have affected in- or out-migration and how migration has affected the public sector of states and localities. One policy concern that has arisen in this respect is whether the jobs created by economic development efforts go to previous residents or inward migrants.

Bibliography

Bartik T J 1993 Who benefits from local job growth: Migrants or the original residents? *Regional Studies* **27**: 297–311

Charney A H 1993 Migration and the public sector: A survey. *Regional Studies* **27**: 313–26

DaVanzo J 1978 Does unemployment affect migration?—Evidence from micro data. *Review of Economics and Statistics* **60**: 504–14

Greenwood M J 1975 Research on internal migration in the United States: A survey. *Journal of Economic Literature* **13**: 397–433

Greenwood M J 1997 Internal migration in developed countries. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier Science, Amsterdam, Volume 1B, pp. 647–720

Greenwood M J 1985 Human migration: Theory, models, and empirical studies. *Journal of Regional Science* **25**: 521–44

Hunt G L 1993 Equilibrium and disequilibrium in migration modelling. *Regional Studies* **27**: 341–9
Klaassen L H, Drewe P 1973 *Migration Policy in Europe*. Lexington Books, Lexington, MA
Long L H 1988 *Migration and Residential Mobility in the United States*. Russell Sage Foundation, New York
Long L H 1991 Residential mobility differences among developed countries. *International Regional Science Review* **14**: 133–47
Mincer J 1978 Family migration decisions. *Journal of Political Economy* **86**: 749–73
Nam C B, Serow W J, Sly D F (eds.) 1990 *International Handbook on Internal Migration*. Greenwood, New York
Plane D A 1993 Demographic influences on migration. *Regional Studies* **27**: 375–83
United Nations 1970 *Methods of Measuring Internal Migration*, Manual VI. United Nations, New York

M. J. Greenwood

Internal Migration (Rural–Urban): Industrialized Countries

1. Basic Concepts

Migration is a demographic process that moves people between places, and so contributes to population change along with fertility and mortality. Migration is the movement of individuals, households, and groups between residences or residential locations. The residences or residential locations before the move are referred to as origins and the postmove residences or residential locations are labeled destinations.

The exact definition of what constitutes a migration has proved to be more difficult to pin down than the definitions of birth and death, and alternative concepts used to measure migration are outlined below. This survey is, however, restricted to migration that takes place entirely within the boundaries of national states. This type of migration is called internal; also very important as a contribution to current demographic change is international migration, in which the migration takes place between countries. We also confine attention to internal migration within industrialized countries, broadly the rich countries housing in 2000 some 1.19 billion people out of the world total of 6.06 billions (UN 2000). Many of these countries have economies in which manufacturing industry plays a relatively small role in employment terms and in which services of an ever-increasing complexity play the biggest part. Industrialized countries include most in Europe, Northern America, Australasia, and some of those in East and South Asia.

The article focuses on the variation in levels of internal migration across countries, the changing relationships between internal migration and national settlement systems, and shows how one major process, urbanization, is being replaced in part by another, labeled counterurbanization. The article concludes by examining the way migration processes vary through the life course and by commenting briefly on the forces working to produce these new patterns.

2. Levels of Internal Migration

Internal migration is measured by national statistical offices in all industrialized countries but there is as yet no international consensus on how it should be measured and reported on a harmonized basis. No international agency compiles statistics reporting on the level, the characteristics, the distances involved, or the impact on population change. Two attempts to compare migration levels are discussed here. Long (1991, 1992) has compiled and interpreted statistics on all changes in residence within a one-year interval, drawing mainly on census statistics or national population registers which capture all migrations. Rees and Kupiszewski (1999b) have used an index proposed by Courgeau (1973a) that makes possible comparison of migration flows at different scales. In both comparisons the intrinsic differences between the types of data captured by registers (all moves within an interval) and by censuses (transitions between start and end locations in an interval) are ignored. These differences are less serious when the time interval is one year.

2.1 Rates of Change of Residence

Table 1 sets out migration intensity estimates (for all changes of residence) from censuses around 1981 derived by Long (1991, 1992), supplemented by further computations by Rees et al. (2000). The statistics derive principally from periodic population censuses, which ask questions about where respondents were living at a date in the recent past (usually one or five years ago). A comparison of residences reported at the start of the reporting period with current residence yields a table of migrant transitions. Censuses or surveys count surviving migrants and normally omit repeat and return migration events, migration events made by persons not surviving the reporting period, and migrants born within the reporting interval. Census derived migration measures are also sensitive to the length of reporting interval (Courgeau 1973b, 1980).

The countries of the English-speaking world outside Europe stand out as experiencing significantly higher internal migration activity than Europe countries or Japan, with 17 to 19 percent of their populations moving each year (Table 1). Because New Zealand, United States, and Australia have, for industrialized countries, relatively youthful populations, standardizing these rates on age reduces the gap a little: the

Table 1
One-year migration intensities based on census data for selected countries

Country	Year	Crude migration rate	Age-standardized
New Zealand	1980–1	19.4[a]	19.6[c]
United States	1981	17.5[a]	17.5[c]
Australia	1980–1	17.0[a]	
	1980–1	16.3[b]	
	1995–6	18.3[b]	15.5[d]
Great Britain	1980–1	9.6[a]	10.6[c]
	1980–1	9.0[b]	
	1990–1	8.8[b]	
Sweden	1981	9.5[a]	
Japan	1979–80	9.5[a]	9.9[c]
France	1975–82	9.4[a]	
Netherlands	1981	7.7[a]	
Belgium	1981	7.3[a]	7.8[c]
Ireland	1980–1	6.1[a]	6.4[c]

a Long (1992). These include all residents. b Rees et al. (2000). These exclude any residents who immigrated in the year to census. c Long (1992) standardized on the 1981 US Age distribution. d Rees et al. (2000) standardized on the 1991 GB Age distributions.

Australian rate, when standardized on the Great Britain population distribution, lowers from 16.3 to 15.5 percent.

2.2 A Scale Independent Measure of Internal Migration Activity

A majority of European national statistical agencies adopt the event view of migration (Rees and Kupiszewski 1999a). This requires use of a population register (centrally consolidated but locally administered), which records an individual's current usual residence and to which inhabitants are required by law to report changes of address. Population registers vary in quality from very high (e.g. Norway: Rees et al. 1999) to moderate, suffering, for example, from mismatches between reported in- and out-migration totals for countries (e.g. Italy: Rees et al. 1998a). Some countries use administrative registers as substitutes for population registers, which suffer from omission of some population subgroups (e.g., the United Kingdom: Stillwell et al. 1992).

Migration measures suffer severely from scale dependency: the larger the number of territorial units recognized the greater the number of migrations observed. It is therefore very difficult to compare migration between N territorial units in one country and M units in another. Long (1991, 1992) therefore confined his international comparisons of internal migration levels to all residential migration (moves between comparable household units). However, Courgeau (1973a) proposed a method to overcome this scale dependence which regresses the scale-specific migration rate against the logarithm of the number of spatial units. The index shown in Table 2 is the slope of this regression, when the intercept is fixed at zero. It measures the rate at which migration activity increases as the number of observation units is augmented. This method enables indexes of migration activity to be computed for countries such as Italy where intramunicipal migration is not recorded in the register statistics. Table 2 is focused on Europe but some measures for Australia are also included. All computations used data for a one-year interval although the data derived from both registers and censuses. Australia stands out in the list (as in Table 1), followed by countries in Northern and North Western Europe. Mobility falls off in Southern and Eastern Europe.

2.3 Factors Influencing the Level of Internal Migration

Why do industrialized countries differ so much in their internal migration propensity? We can only speculate, in the absence of a definitive cross-national investigation, but the following factors can be suggested as influential.

(a) The degree of flexibility and ease of transfer in the labor is of vital importance. European countries have more rigid labor markets and higher unemployment, which dampen migration.

(b) European countries have large social housing sectors which encourage locational inertia, while large private rental sectors in the United States, New Zealand, and Australia facilitate regular housing adjustment.

(c) The rules operating within social housing systems (e.g., length of residence requirements, the territory within which housing can be allocated) affect migration. For example, social housing tenants find it easy to migrate within Australian states; in Britain, migration between municipalities is difficult for social housing tenants.

(d) Traditions of frontier development and rapid adjustment to changing economic circumstances characterize the United States and Australia, whereas traditions of localism and familism, together with welfare state provision, help keep unemployed Italians in the Mezzogiorno and British ex-miners in coalfield villages.

(e) The compact size and denser settlement patterns of European countries allow an adjustment, within many metropolitan agglomerations, through commuting changes rather than residential migration. There is, for example, little relationship between unemployment and migration exchanges in the Netherlands (Rees et al. 1998b).

Table 2
Comparable measures of the intensity of internal migration in 10 European countries and Australia, using an index developed by Courgeau (various years in the early 1990s)

Country	Period	Courgeau's K index with zero intercept	Source of migration data
Australia	1980–1[a]	1.52	Census
Norway	1994	0.71	Register
The Netherlands	1994	0.61	Register
Great Britain	1980–1[a]	0.54	Census
	1990–91[a]	0.52	Census
Germany	1994	0.51	Register
Portugal	1989–91[b]	0.38	Census
Estonia	1995	0.33	Register
Czech Republic	1994	0.21	Register
Italy	1994	0.19	Register
Romania	1994	0.16	Register
Poland	1994	0.15	Register

Source: Rees and Kupiszewski (1999), p33; Bell et al. (2000), Table 7
a The year prior to the census. b A longer interval of one year and 104 days has been prorated to a year.

(f) The per capita income levels in countries influence the degree to which inhabitants in later or postworking life can make consumption-related migrations. Retirement migration is important in higher income countries but has not yet emerged in Eastern Europe.

3. *Internal Migration and Settlement Structure*

Migration within countries shifts people between settlements distinguished in two principal ways: (a) by size, density, and type of economic activity and (b) by the health of the regional economy. The first dimension is usually operationalized as an urban–rural dichotomy, but needs also to be viewed more generally as a settlement hierarchy from large metropolis through medium sized city to village, hamlet, and isolated farmstead. Settlements are recognized as urban if their populations exceed a low threshold (2,000–20,000 depending on country) and if economic activity is mainly outside farming.

3.1 *Long Run Urbanization*

Internal migration has made a vital contribution to the growth of the urban population over the past two centuries. This urbanization process is monitored and projected regularly by the United Nations. Table 3 reports summary statistics from the latest evaluation (UN 2000). In more developed regions, the half century 1950 to 2000 saw a doubling to 900 million in their urban population, though this was dwarfed by urban population increase in less developed regions. The rural population of more developed regions fell during this period. The UN's Population Division sees urbanization in industrialized countries continuing in the future but at a slower rate. The main driver of urbanization has been net internal migration from the countryside to the city.

3.2 *Is Urbanization Turning into Counterurbanization?*

In the mid-1970s, the American researcher Calvin Beale pointed out that in much of nonmetropolitan America, there was a revival of population growth after decades of decline (Beale 1975). The term counterurbanization was coined shortly afterwards by Berry (1976) to label this process, which was swiftly recognized in Australia and Canada (Bourne and Logan 1976). An avalanche of studies of this new settlement dynamic has subsequently occurred, establishing its prevalence in a large number of industrialized countries but also indicating that its strength has fluctuated since the 1970s, when it was probably most strong (see Champion 1998 and Kontuly 1998 for recent reviews).

There have been considerable difficulties in agreeing an operational definition of the counterurbanization process. Researchers often use the population of settlements and their variation as indicators, but natural increase also contributes to population change, so the indicator of choice is either net total migration or internal migration. Wherever possible, researchers use statistics on internal migration, as international migration has a very different distribution in industrialized countries with external immigrants frequently replacing the population leaving

Table 3
Urbanization prospects, 1950–2030

	Populations (billions)		
Unit	1950	2000	2030
More developed regions			
Urban	0.45	0.90	1.01
Rural	0.37	0.28	0.20
Less developed regions			
Urban	0.30	1.94	3.88
Rural	1.41	2.39	3.05
	Percent urban		
Unit	1950	2000	2030
More developed regions	55	70	84
Less developed regions	18	40	56
Unit	Urbanization rate (percent per year)		
	1950–2000	2000–30	
More developed regions	0.65	0.31	
Less developed regions	1.62	1.14	

Source: United Nations (2000), Table 1

Table 4
Net internal migration by type of region, 1976–81 and 1991–6, Australia (1,000s)

Type of region	1976–81	1991–6
Metropolitan core	−283	−135
Metropolitan rest	188	56
Near	37	35
Coast	109	150
Far	−27	−60
Remote	−26	−46

Source: Stillwell et al. (2000), from ABS migration and population data

large cities through internal migration. At first, the spatial framework of analysis used to analyze population or migration trends was a simple rural/urban dichotomy or metropolitan/nonmetropolitan county division (in the United States). It was soon realized that this was both arbitrary and inadequate for capturing the complexity of population processes under way. Researchers sought to look at migration within the urban system in two ways: by the size of town or city (a hierarchical classification) and by the degree of linkage in distance or transaction terms between places and urban centers. The definition of counterurbanization was extended to mean a process of shifting the population from larger to smaller places over time, either down the urban hierarchy or outwards within any one urban field. This latter process of migration to the less densely settled urban periphery had been underway for decades before the 1970s and so some researchers argued that counterurbanization was simply an extension of suburbanization to a much wider area than hitherto, made possible by better transportation and telecommuting links.

Table 4 shows the typical table used by researchers aiming to diagnose the presence of this new population redistribution process. The small units (Statistical Divisions) for which migration is reported in successive Australian censuses were organized firstly into time consistent units and then grouped into zones on the basis of proximity to metropolitan centers (Blake et al. 2000). The resulting 35 city region-zones are further grouped in Table 4 into six categories, two metro and four nonmetro. In the 1976–81 period there were substantial losses from metro cores and gains within the rest of Australia's five metro regions. Near metro and attractive east coast zones also gained migrants. However, the deep countryside (far regions, mainly wheat, sheep and cattle farms) and the remote outback (mainly mining settlements) lost internal migrants. Comparison with the situation in the 1991–6 period shows that these most peripheral regions exported even more people, reflecting a depression in farming and a switch in mining to manning by workers with permanent residences elsewhere. Metro losses were lower reflecting urban revival and redevelopment, and metro rest and near gains lower. However, coastal regions showed larger gains as the foci of consumption led migration.

European evidence from the 1970s is gathered

Table 5
Evidence from published studies on urbanization or counterurbanization trends in European countries during the 1970s, 1980s, and 1990s

	Urbanization			Counterurbanization		
Country	1970s	1980s	1990s	1970s	1980s	1990s
Urbanization mainly						
Austria	✓	✓				
Czechoslovakia	✓	✓	✓			
Finland	✓	✓	✓			
Germany (E)	✓	✓				
Norway	✓	✓	✓	✓		
Portugal	✓	✓	✓			
Both processes						
Belgium	✓	✓		✓	✓	✓
Iceland		✓		✓		
Ireland	✓				✓	
Italy	✓				✓	✓
Spain	✓	✓			✓	
Counterurbanization mainly						
Denmark				✓	✓	✓
France				✓	✓	✓
Germany (W)				✓	✓	✓
Greece					✓	
Netherlands		✓		✓	✓	✓
Sweden		✓		✓	✓	✓
Switzerland				✓	✓	✓

Source: Adapted from Kontuly (1998)

Table 6
The relationship between internal migration and density in selected European countries in the 1990s

Relationship	Countries
Flow from low to high density	Norway, Poland, Romania
Flow from low and high to middle density	Czech Republic, Germany, Italy, Portugal
Flow from high to low density	Estonia, Netherlands, United Kingdom

Source: Rees and Kupiszewski (1999b), Figure 12.1

together by Kontuly (1998) and presented in a generalized format in Table 5. The table makes a 'fuzzy' judgment about whether studies found urbanization or counterurbanization to be in train in different periods. The picture is a complicated one, with studies in some countries (e.g., Belgium) disagreeing about trends because of differences in exact period studied and in the spatial scale of observations. In Austria, Czechoslovakia, Finland, Norway, and Portugal urbanization is still the dominant process, though in each country there is evidence of deconcentration from capital cities. Italy, Ireland, and Spain appear to be in process of moving from an urbanization pattern (1970s) to a moderate counterurbanization in the 1980s and 1990s. Denmark, France, Germany, Sweden, and Switzerland all exhibited consistent counterurbanization patterns over the 1970s, 1980s, and 1990s.

It is, of course, difficult to compare studies by different authors, so Rees and Kupiszewski (1999b) in their work for the Council of Europe carried out a set of similar analyses for a set of 10 European countries. Table 6 summarizes their findings in terms of the degree to which countries had moved to a counterurbanization pattern by the early or mid-1990s (1994 usually). The method used was to tabulate net internal migration rates in each country by municipality (or equivalent unit) grouped into population density bands (used earlier by Fielding (1982) but for larger regions). The direction and slopes of graphing net internal migration against density resulted in the assignments shown in Table 6. The results broadly confirm the Kontuly review of previous studies and show additionally that the Netherlands and the United Kingdom, along with France, exhibit the strongest forms of counterurbanization.

We can ask whether there is a sequence in the development of national urban/rural settlement systems that leads from urbanization to counterurbanization. There is no clear answer to this question as yet. The Rees and Kupiszewski studies showed that the

Table 7
Net internal migration rate by size of settlement in Poland in 1994 (per 1,000 per year)

Size of settlement (1,000s of inhabitants)	1994
> 500	+2.0
250–500	+2.3
100–250	+1.2
50–100	+1.5
25–50	+1.3
10–25	+1.4
5–10	+2.4
Towns < 5	+1.4
Rural areas	−2.6

Source: Kupiszewski et al. (1998)

Table 8
Net internal migration rates by type of commune in Italy, 1984 and 1994 (per 1,000 per year)

Type of commune	1984	1994
Urban	−2.0	−0.3
Semiurban	+5.6	+5.3
Semirural	+4.3	+6.3
Rural	+1.9	+1.1

Source: Rees et al. (1998)

Table 9
Net internal migration rates and types of settlement in France (1982–9) (per 1,000 per year)

Settlement type	1982–90
Urban centers	−5.0
Periurban rings	+17.4
Multicenter communes	+11.8
Rural	+2.8

Source: Baccaïni et al. (2000) from 1990 Census of France

Table 10
Net internal migration rates by population size of urban region and zone within urban region, Great Britain, 1990–1 (per 1,000 per year)

Size of urban region (1,000s of inhabitants)	Cores	Zone Rings	Outer areas/ rural areas	Totals
800+	−6.7	+1.6	+6.2	−5.5
400–< 800	−3.9	+3.3	+9.4	−0.9
200–< 400	−2.5	+4.2	+6.7	+0.6
100–< 200	−0.1	+4.8	+7.4	+2.7
< 100	+0.9	+5.6	+7.0	+4.5
Totals	−3.4	+4.1	+7.2	0.0

Source: Rees et al. 1996, from the 1991 Census of Great Britain, Crown Copyright and ESRC/JISC Purchase

strength of counterurbanization in 1994 (or 1990–1) was weaker than in 1984 (or 1980–1) in Western European countries, while the strength of urbanization in transition societies was weakening. In other countries there was considerable fluctuation in the prevailing relationship and sometimes it depended on the size of the different age groups taking part. So, in Norway, urbanization at the younger ages masked some deconcentration from Oslo city in the family ages.

3.3 Relationships Between Net Internal Migration and Settlement Structure

We now look at some particular country examples from different points in the urbanization–counterurbanization spectrum. Table 7 shows an example of a country, Poland, that is still urbanizing. Rural areas experienced in 1994 net internal migration losses, but the pattern of gains to urban places is not strongly related to size. Plots of population growth in the 1984–94 period and of net internal migration in 1994 against settlement density suggests a diminution of the relationship over time. Once the poorer countryside areas have emptied of their surplus labor, then we can expect deconcentration from urban centers to become a dominant influence.

Table 8 illustrates the relationship between settlement type and net internal migration for Italy, a country with weak counterurbanization. Urban communes lost internal migrants in both 1984 and 1994 but at a lower rate in the later year. The gaining rate for semirural communes increased between these two years but this was countered by a fall in gains to rural communes. The Italian case confirms the need to focus on internal migration as the variable of interest, as large numbers of rural and semi-rural communes in central and northern Italy now experience natural decrease.

A seminal study of internal migration by Boudoul and Faur (1986) established the way in which France transited from a country of urbanization in the 1950s to one of strong counterurbanization in the 1970s. Table 9 confirms that this continues to be the case in the 1980s, with the country's main urban centers losing migrants to surrounding communes and smaller centers. However, rural gains from internal migration are relatively small.

To tease out the relative contributions of urban deconcentration within a metropolitan region from the shift of population to different and smaller settlements requires a good urban–rural classification based on commuting patterns. Such a classification is available for the United Kingdom and has been used by Champion, Atkins, Dorling, and others to demonstrate the 'counterurbanization cascade.' Table 10, taken from Rees et al. (1996), uses their classification to group small areas by urban region size and by urban region zone (core, ring, periphery). The net internal

migration rates show the simultaneous operation of both cascading down the urban hierarchy and deconcentration from urban centers. However, on balance the statistics suggest that deconcentration is more significant than cascading. Most rural areas in the United Kingdom in 1990–1 gained through internal migration, and the remote region effect noted earlier for Australia was absent.

4. *Conclusions*

The review in the previous section has looked at patterns of internal migration for the population as a whole. There are, of course, important differences between different population groups. Of particular interest has been life course variation in internal migration patterns. The relationship of migration in intensity with age is well known and has been studied exhaustively (see Rogers and Castro 1981). There is also a strong relationship between life course stage and migration direction in counterurbanizing countries (see Illeris 1996 on Denmark, for example). Table 11 shows how the direction of migration across settlement type varies by age group in the Netherlands. The all ages pattern is one of small losses from urban cores and small gains to commuter towns and the rural periphery. Young adults aged 15–29 swim against this counterurban stream, showing high internal in-migration rates to city cores and high out-migration

Table 11
Net internal migration rates by age, Netherlands, by settlement type, 1994 (per 1,000 per year)

	Ages						
Settlement type	0–14	15–29	30–44	45–59	60–75	75+	Total
Cores: main centers	−10.6	16.6	−12.3	−3.2	−4.4	−7.2	−2.2
Secondary cores[a]	−3.3	1.6	−3.9	−0.6	2.1	0.7	−1.0
Ring: commuter towns	5.2	−8.6	7.8	−0.1	2.8	9.3	2.0
Periphery	4.5	−10.6	5.7	2.8	1.6	0.5	0.7

Source: Rees et al. (1998b), computed from data supplied by Statistics Netherlands
a Very large free-standing towns.

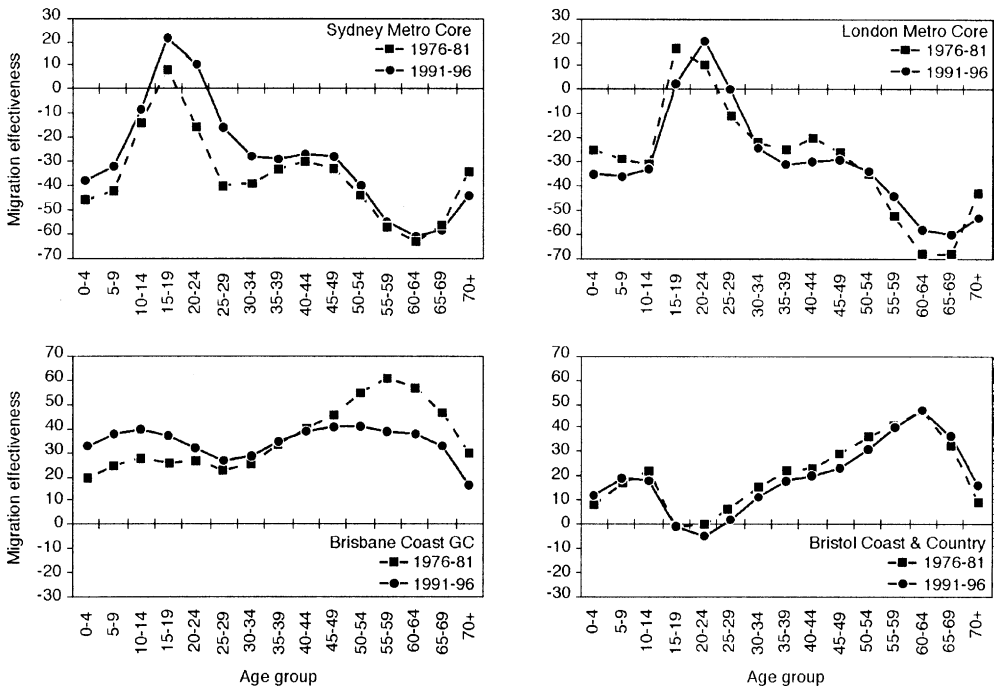

Figure 1
Net internal migration effectiveness in two capital cities and two attractive nonmetropolitan regions, Australia and the United Kingdom, 1976–81 and 1991–6

rates from commuter towns and the countryside. This pattern is repeated in most of the other strong counter-urbanizing countries (the United States, Australia, Denmark, the United Kingdom). The other ages show losses from the city cores and gains in commuter towns and the periphery, though the balance between these two varies from group to group. The strong gains of the elderly (75+) in the ring commuter towns and low gains in the periphery are linked to a desire to be close to health and support services, for example.

Figure 1 shows the way in which the redistributive effect of net migration (measured here by the effectiveness ratio, the ratio of net migration to the sum of in- and out-migration) varies by age, for two metro and two nonmetro regions in Australia and the United Kingdom. Despite vast differences in area extent, economic structure, and settlement history between these two countries, there is a remarkable similarity in internal migration patterns by age. Migration in the retirement ages is strongly effective in both countries, with people leaving the largest city at retirement and returning rarely thereafter. The metro and nonmetro curves form approximate mirror images of each other. The only major difference between countries is the greater effectiveness of younger age migration to the Brisbane Gold Coast, an attractive resort and beach region which holds on to its young people, while the Bristol Coast and Country region (Devon, Cornwall, Dorset) is less successful.

A final word is appropriate on the explanations suggested for the new pattern of counterurbanizing internal migration. A full discussion is provided by Champion (1989, 1998). One set of explanations focuses on changes in the way production is organized. New economy, high value manufacturing is no longer constrained to metropolitan centers and finding locations attractive to high skill workers is more important. Large cities suffer from many problems of growth and high density: congestion, pollution, crime, concentrations of poverty, all of which deter both enterprises and individuals from locating there. This effect, together with the restructuring of the economy, was probably most important in the 1970s and first part of the 1980s. But the emptying out of the cores of cities has created some opportunities for new enterprises and residential developments, accounting for the subdued level of counterurbanization in the 1990s. There have also been more particular 'production' factors involved in some countries: the degree to which the state has invested in rural areas and rural activities. An important boost to internal migration in the United States or in the United Kingdom has been the foundation or growth of higher education colleges or universities.

Consumption-led migration has also been important. In many countries the retirement of a generation with generous pension provision has given them the freedom to choose new residential locations. The rural–urban price gradient also provided an incentive for persons not tied to workplaces to move. However, there have had to be developments in rural and smaller towns to make them attractive to internal migrants. Rural accessibility has improved with the construction of new highways, telecommunications have spread to all parts of the national space, and in some countries improved rail services have increased accessibility. The importance of accessibility to urban services in rural areas and smaller towns is a recurring theme in many studies of internal migration patterns. Inaccessible, remote areas have been shunned by counter-urbanizing migrants. Also important, where migrants have choices, are residential preferences for living in particular environments, with the 'rural idyll' being a strong component of some national cultures while in others the preference is for the interactions, activities, and freedom afforded in urban places.

See also: Development and Urbanization; Diffusion: Anthropological Aspects; Immigration; Immigration and Migration: Cultural Concerns; Internal Migration: Developing Countries; Internal Migration (Further Types): Industrialized Countries; Migration and Health; Migration, Economics of; Migration into the Americas and between American Countries; Migration out of Europe; Migration: Sociological Aspects; Migration, Theory of; Population Pressure, Resources, and the Environment: Industrialized World; Rural Geography; Urban Geography; Urban Growth Models; Urban Sprawl

Bibliography

Baccaïni B, Pumain D, Rees P, Kupiszewski M 2000 Migrations et urbanisations dans les pays Européens. In: Mattei M F, Pumain D (eds.) *Données Urbaines* 3. Anthropos, Paris, pp. 351–61

Beale C 1975 The revival of population growth in non-metropolitan America. *Report ERS 605*, Economic Research Service, United States Department of Agriculture, Washington DC

Bell M, Blake M, Boyle P, Duke-Williams O, Hugo G, Rees P, Stillwell J 2000 Cross-national comparisons of internal migration: issues and measures. *Journal of the Royal Statistical Society*, Series A, in press

Berry B 1976 The counterurbanisation process: Urban America since 1970. In: Berry B (ed.) *Urbanisation and Counterurbanisation*. Sage, Beverley Hills, CA, pp. 17–30

Blake M, Bell M, Rees P 2000 Creating a temporally consistent spatial framework for the analysis of inter-regional migration in Australia. *International Journal of Population Geography* **6**(2): 155–74

Boudoul J, Faur J P 1986 Trente ans de migrations intérieures. *Espaces*, *Populations*, *Sociéties* **2**: 293–302

Bourne L, Logan M 1976 Changing urbanisation patterns at the margin: The examples of Australia and Canada. In: Berry B (ed.) *Urbanisation and Counterurbanisation*. Sage, Beverley Hills, CA, pp. 111–43

Cermak Z 1996 Internal migration in the Czech Republic during the period of transition. Acta Facultis, *Rerum Naturalium Universitatis Comenianae Geographica*, n. 37, Univetzita Komenskeho, Bratislava, pp. 122–30

Champion A (ed.) 1989 *Counterurbanisation: The Changing Pace and Nature of Population Deconcentration*. Edward Arnold, London

Champion A 1998 Studying counterurbanisation and the rural population turnaround. In: Boyle P, Halfacree K, Robinson V (eds.) *Migration into Rural Areas: Theories and Issues*. Wiley, Chichester, UK, Chap. 2 pp. 21–40

Courgeau D 1973a Migrations et découpages du territoire. *Population* **28**(3): 511–37

Courgeau D 1973b Migrants et migrations. *Population* **28**(1): 95–129

Courgeau D 1980 *Analyse Quantitative des Migration Humaines*. Masson, Paris

Fielding A 1982 Counterurbanisation in Western Europe. *Progress in Planning* **17**: 1–52

Illeris S 1996 Changing patterns of net migration in Denmark: An explanatory analysis. In: Rees P, Stillwell J, Convey A, Kupiszewski M (eds.) *Population Migration in the European Union*. John Wiley, Chichester, UK, Chap. 7, pp. 105–22

Kontuly T 1998 Contrasting the counterurbanisation experience in European nations. In: Boyle P, Halfacree K, Robinson V (eds.) *Migration into Rural Areas: Theories and Issues*. Wiley, Chichester, UK, Chap. 4, pp. 61–78

Kupiszewski M, Durham H, Rees P 1998 Internal migration and urban change in Poland. *European Journal of Population* **14**: 265–90

Long L 1991 Residential mobility differences among developed countries. *International Regional Science Review* **14**(2): 133–47

Long L 1992 Changing residence: Comparative perspectives on its relationship to age, sex, and marital status. *Population Studies* **46**(1): 141–58

Rees P, Bell M, Duke-Williams O, Blake M 2000 Problems and solutions in the measurement of migration intensities. *Population Studies* **54**(2): 207–22

Rees P, Durham H, Kupiszewski M 1996 Internal migration and regional population dynamics in Europe: United Kingdom case study. *Working Paper 96/20*, School of Geography, University of Leeds, Leeds, UK

Rees P, Kupiszewski M 1999a Internal migration: What data are available in Europe? *Journal of Official Statistics* **15**(4): 551–86

Rees P, Kupizsewki M 1999b *Internal Migration and Regional Population Dynamics in Europe: A Synthesis*. Council of Europe Publishing, Strasbourg

Rees P, Østby L, Durham H, Kupiszewski M 1999 *Internal Migration and Regional Population Dynamics in Europe: Norway Case Study*. Reports 99/19. Statistics Norway, Oslo

Rees P, Todisco E, Terra Abrami V, Durham H, Kupiszewski M 1998a *Internal Migration and Regional Population Dynamics in Europe: Italy Case Study*, Essays n. 3. National Institute of Statistics, Rome

Rees P, van Imhoff E, Durham H, Kupiszewski M, Smith D 1998b *Internal Migration and Regional Population Dynamics in the Netherlands*. Netherlands Interdisciplinary Demographic Institute, report No. 55, The Hague

Rogers A, Castro L 1981 *Model Migration Schedules*. Research Report RR-81-30. International Institute for Applied Systems Analysis, Laxenburg, Austria

Stillwell J, Rees P, Boden P (eds.) 1992 *Migration Processes and Patterns. Volume 2. Population Redistribution in the United Kingdom*, Belhaven, London

Stillwell J, Bell M, Blake M, Duke-Williams O, Rees P 2000 Net migration and migration effectiveness: A comparison between Australia and the United Kingdom, 1976–96. *Journal of Population Research* **17**(1): 17–38

Stillwell J, Bell M, Blake M, Duke-Williams O, Rees P 2001 Net migration and migration effectiveness: A comparison between Australia and the United Kingdom, 1976–96: Part 2, Age related migration patterns. *Journal of Population Research* **18** forthcoming

UN Population Division, Department of International Economic and Social Affairs 2000 *World Urbanization Prospects. The 1999 Revision. Data Tables and Highlights*. ESA/P/WP.161 United Nations, New York

P. Rees

Internal Validity

Internal validity refers generally to the accuracy of inferences about whether one variable causes another. In the context of an experiment, internal validity is concerned with conclusions about whether (and to what degree) the independent variable, as manipulated, makes a difference in the dependent variable, as measured. This article addresses the meaning of internal validity, the practice of achieving it, and the challenges of understanding it.

1. *Internal Validity Defined*

The concept of internal validity apparently originated with Donald T. Campbell and was popularized in his and his collaborators' work (Campbell 1957, Campbell and Stanley 1966, Cook and Campbell 1979). The term was coined as a counterpoint to 'external validity,' which deals with the generalizaSbility of a finding to persons, settings, and times other than those examined in the research (see *External Validity*; *Generalization: Conceptions in the Social Sciences*). Internal validity, in contrast, then, involves the accuracy of a causal inference pertaining to the particular persons, settings, and times examined in the research. As Cook and Campbell (1979) made clear, internal validity is also concerned only with the particular research operations used in a study, that is, with the independent variable as it was manipulated and the dependent variable as it was measured.

The term internal validity originated in the context of research methods designed to probe cause–effect relations, such as randomized experiments and quasi-experiments (see *Random Assignment: Implementation in Complex Field Settings*; *Nonequivalent Group Designs*). This remains the typical usage. However, the term is sometimes applied to procedures that do not investigate causal relations, specifically to measurement instruments such as personality scales. In this alternative usage, the apparent intention is to refer to

the structure of the scale and the interrelationship among scale items as 'internal validity,' thus differentiating these properties from the scale's relationship to other measures and behaviors. Such a different and nontraditional use of the term may invite confusion. The remainder of this article focuses on internal validity as it applies to cause-probing research.

2. *Differentiating Internal Validity from Other Forms of Validity*

Campbell (1986) noted that the term internal validity (even when restricted to cause-probing research) was often used in ways somewhat different from the original intended meaning. He proposed the infelicitous term 'local molar validity' as a substitute. This term has not been widely adopted but can help in differentiating internal validity from other forms of validity. Consider as an example a study conducted in a middle school to see whether an anger management program reduces students' aggressive behaviors such as fighting on the playground. Internal validity is 'local' in the sense that it is concerned with the immediate context of the study, that is, the specific children and school that were observed. Attempts to generalize, whether to other middle schools, to high schools, or to adults, instead involve external validity. Internal validity is 'molar' in the sense that it is concerned with whole manipulations and with measures, whatever they are, rather than with pure theoretical abstracts. Attempts to draw conclusions about theoretical concepts, say about 'affective regulation training' and 'aggression,' rather than about the program as implemented and the raters' observations of the number of fights on the playground, instead involve construct validity.

Some competing validity frameworks define internal validity in different ways. For instance, Cronbach (1982) defined internal validity as involving certain *intended* generalizations (for a summary and integration of this and other validity frameworks, see Mark 1986). Cronbach's alternative definition of internal validity includes generalizations to the categories of persons, settings, and times and to the theoretical constructs that were the original, intended targets of the research conclusions. The original concept of internal validity, as developed by Campbell and associates, continues to predominate, however.

The greatest difficulty has been in differentiating internal validity from the construct validity of the cause, that is, the proper labeling of the independent variable in abstract terms. This difficulty is seen most clearly in terms of 'threats' to validity (see Sect. 3). Cook and Campbell (1979) presented four threats to internal validity that depend upon comparative processes involving the members of the treatment and control group. For instance, the threat of 'resentful demoralization' can occur when one group, say the control group, receives a perceptibly less desirable treatment. If control group members become resentful, this resentment, rather than the intended treatment, may cause differences between the groups. Even Cook and Campbell (1976, 1979, Campbell 1986) have wavered in their judgment about whether these are internal or construct validity threats. More generally, threats to both internal validity and construct validity of the cause involve confounds with the independent variable, suggesting that it may be difficult to differentiate the two types of validity.

One possible resolution is based on the counterfactual conception of cause (Reichardt and Mark 1998; see *Causal Counterfactuals in Social Science Research*). From this perspective, a treatment effect can be defined as the difference between what happens when a (molar) treatment *has been* administered and what *would have happened* if the (molar) treatment had *not* been administered but everything else had been the same. The practical problem is that, absent time travel, *everything* cannot be the same between the treatment and control conditions except for the treatment and its effects. The same people may be compared at different times, or different people may be compared at the same time, but a researcher cannot compare the same people, with and without the treatment, at the same time. But if one allows the fiction of the ideal (but unattainable) counterfactual comparison that only time travel would make possible, the distinction between internal validity and construct validity of the cause can be clarified.

Threats to internal validity would not arise if the ideal comparison could be made. In contrast, a threat to the construct validity of the cause is a mislabeling of the cause that could arise even with the ideal comparison. For example, in studying an anger management program, it would be an internal validity problem if one compared pretest and post-test levels of aggression and if aggression changed simply because the children were older at the post-test. But this internal validity problem would disappear if the ideal counterfactual could actually be obtained, because the comparison would involve the same children at the same age, with and without exposure to the program. In contrast, it would be a construct validity problem if the program's activities modified children's affective regulation skills but the program were labeled as a 'self-efficacy intervention.' This problem would not be avoided by the ideal counterfactual. Nor would the ideal counterfactual alleviate other construct validity problems, such as various subject and experimenter artifacts (see *Experimenter and Subject Artifacts: Methodology*). In the case of resentful demoralization, this threat would not occur if the ideal counterfactual were attainable. If a researcher could travel back in time, there would be no need to construct two groups of participants who could be aware of each others'

treatment. Resentful demoralization thus is a threat to internal validity.

3. Threats to Internal Validity

The literature on internal validity consists largely of detailed lists of validity 'threats.' Internal validity threats are generic categories of causal forces that may frequently obscure causal inferences. Take as an example, once again, a researcher's efforts to determine whether an anger management program reduces aggressive behavior in a middle school. 'History' refers to the possibility that specific events, other than the intended treatment, may have occurred between the pretest and post-test observations and may obscure the true treatment effect. If the researcher observed the level of aggressive behavior on the playground before the anger management program, and again afterward, history would be a problem if a different, stricter teacher became playground monitor in the interim. 'Maturation' refers to the possibility that natural processes which occur over time within the study participants, such as growing older, hungrier, more fatigued, wiser, and the like, may create a false treatment effect or mask a real one. Less aggression may occur at the post-test simply because the children are older than at the pretest, for instance. 'Attrition' refers to the possible loss of participants in a study. For example, if children from troubled families are more likely to drop out of school or to move away in the middle of the school year, then attrition could cause a decrease in aggression from the pretest to the post-test. 'Instrumentation' arises as a validity threat when a change in a measuring instrument causes erroneous conclusions about the effects of an intervention. For instance, if observers' standards shifted over time, such that later incidents had to be more violent to be rated as aggressive, this could cause the appearance of a treatment effect when in fact there is none.

'Selection' refers to the possibility that post-test differences between a treatment group and a control group may be due to initial differences between the groups rather than to a treatment effect. Selection problems might occur if a researcher attempted to assess the effectiveness of an anger management program by comparing the level of playground aggression in two middle schools, one of which had implemented the program. In addition, more complex internal validity problems can occur, whereby some threat operates only (or more powerfully) in one group than another. For instance, 'selection by maturation' indicates that participants in the treatment condition are maturing at a different rate than those in the control condition. See Cook and Campbell (1979) for additional discussion of internal validity threats, including the threats of testing and regression to the mean.

4. Achieving Internal Validity

Most discussions about how to achieve internal validity focus on research design. Randomized experiments are generally recommended, because random assignment eliminates systematic selection bias and allows traditional statistics to estimate and account for purely random selection differences (see *Random Assignment: Implementation in Complex Field Settings*; *Selection Bias, Statistics of*). Randomized experiments also rule out most other internal validity threats, if sound research procedures are used and there is no differential attrition (but see Sect. 5.1). If random assignment is either impractical or unethical, the common recommendation is to enhance internal validity by using a strong quasi-experiment. Quasi-experiments are approximations to experiments but lack random assignment (see *Nonequivalent Group Designs*; *Time Series Quasi-experiment*; *Regression Discontinuity Design*). More generally, the process of ruling out internal validity threats can be seen as a special instance of the logic of pattern matching (see *Pattern Matching: Methodology*). In addition, especially in terms of the threats of selection and selection by maturation, the choice of proper statistical analyses can influence internal validity (see *Nonequivalent Group Designs*; *Selection Bias, Statistics of*).

5. Common Misconceptions about Internal Validity

Several misconceptions exist regarding internal validity. Three relatively common ones are discussed in this final section.

5.1 Misconception #1: Successful Random Assignment Guarantees Internal Validity

Although random assignment of participants (or other units) to treatment condition can greatly enhance the likelihood of internal validity, problems can still occur. Most widely recognized is that differential attrition may occur, with more (or different kinds of) participants dropping out of one group than another. As noted earlier, Cook and Campbell also suggested that threats such as resentful demoralization may apply in a randomized experiment. Even if these problems do not occur, internal validity threats can arise in a randomized experiment if proper research procedures are not followed. An experimenter might, for instance, have one rater observe aggression in the treatment group and another rater observe in the control group. This would create an instrumentation threat. As another example, researchers sometimes randomly assign individuals to conditions but then have all members of a group participate together. For example, in a study with mood as the independent variable, following random assignment all members of the positive mood condition may be sent to one room to

watch a funny movie, while all members of the control condition see a 'neutral' movie in another room. This can allow the independent variable to become confounded with any of a number of other factors, such as the characteristics of the experimenter conducting each session, creating a selection by history threat. To minimize internal validity threats, random assignment must be combined with careful methodology (see *Laboratory Experiment: Methodology*).

5.2 Misperception #2: Presence of a Validity Threat = Internal Invalidity

Despite clear statements by Campbell to the contrary (e.g., Campbell 1969), discussions of internal validity often make it sound as though the theoretical existence of a validity threat necessarily equates with weak internal validity. Some writings seem to suggest, for example, that, because history and several other validity threats *can* apply to a simple pretest–post-test comparison with one group, the findings from such a study are necessarily invalid. Such thinking is incorrect for at least three reasons. First, a given threat may not operate in a specific case, whether or not it is ruled out by the design. Maturation does not always cause changes in every pretest–post-test study, for instance. Second, conclusions about the causal impact of a treatment on some outcome may be accurate, even when a validity threat is operating, if the threat's effect is too small to invalidate the conclusion drawn. In a pretest–post-test design, for example, maturation may occur but be small enough not to obscure a reasonably accurate conclusion about the treatment effect. Third, even if a threat is not trivial in size, in some cases it may be possible to estimate the magnitude of the threat and adjust for it. This is the basic logic of efforts, for instance, to model selection bias (see *Selection Bias, Statistics of*; and see Reichardt 2000 for an elaboration of this logic). Still, assessing the plausibility and magnitude of validity in a nonexperimental context remains an imprecise art. The development of empirically supported theories of the conditions under which various validity threats operate, with what magnitude, would be an important future development.

5.3 Misconception #3: An Emphasis on Internal Validity = A 'Black Box' Experiment

Some critics contend that giving priority to internal validity implies a disinterest in mediating processes. In applied social research, for example, the claim has been made that those who give priority to internal validity commonly estimate the effect of an intervention without trying to peer into the 'black box' and learn about underlying processes. In fact, the methods associated with strong internal validity, such as the randomized experiment, do *enable* researchers to conduct black box experiments. However, they do not require it. Experiments (and other methods used in the service of internal validity) can be integrated with other methods used to study mediational models). Moreover, methods that maximize internal validity are also widely used in investigations designed specifically to test hypotheses about underlying causal mechanisms (see *Laboratory Experiment: Methodology*).

See also: Content Validity; External Validity

Bibliography

Campbell D T 1957 Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* **54**: 453–6

Campbell D T 1969 Prospective: Artifact and control. In: Rosenthal R, Rosnow R L (eds.) *Artifact in Behavioral Research*. Academic Press, New York

Campbell D T 1986 Relabeling internal and external validity for applied social scientists. In: Trochim W M K (ed.) *Advances in Quasi-Experimental Design and Analysis*, no. 31. Jossey-Bass, San Francisco

Campbell D T, Stanley J C 1966 *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago

Cook T D, Campbell D T 1976 The design and conduct of quasi-experiments and field experiments in field settings. In: Dunnette M D (ed.) *Handbook of Industrial and Organizational Psychology*. Rand McNally, Chicago

Cook T D, Campbell D T 1979 *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Rand McNally, Chicago

Cronbach L J 1982 *Designing Evaluations of Educational and Social Programs*. Jossey-Bass, San Francisco

Mark M M 1986 Validity typologies and the logic and practice of quasi-experimentation. In: Trochim W M K (ed.) *Advances in Quasi-Experimental Design and Analysis. New Directions for Program Evaluation*, no. 31. Jossey-Bass, San Francisco

Reichardt C S 2000 A typology of strategies for ruling out threats to validity. In: L. Bickman (ed.) *Research Design: Donald Campbell's Legacy*. Sage, Thousand Oaks, CA, Vol. 2

Reichardt C S, Mark M M 1998 Quasi-experimentation. In: Bickman L, Rog D (eds.) *Handbook of Applied Social Research*. Sage, Thousand Oaks, CA

M. M. Mark and C. S. Reichardt

Internal Warfare: Civil War, Insurgency, and Regional Conflict

Internal wars are those that take place within the boundaries of one country, making them distinct from interstate war which is generally fought between two (or more) states. Internal conflicts often—but need not—involve the state on one side of the conflict. There are different ways to categorize internal conflicts. Some think of them as breaking along the lines of the goals of the combatants (irredentism, secession, ideological), while others focus on the primary orientation of the groups in conflict (religious, ethnicity, ideology). While an increased prevalence of internal warfare seems to coincide with the end of the Cold

War, their increased frequency shows up most dramatically as the former European colonies gained their independence (Gurr 1994, Regan 2000). Because nonstate actors are involved in these conflicts and they take place within the territorial boundaries of one country, conflict management is often difficult.

1. Conceptual Issues Regarding the Causes and Frequency of Internal Conflict

Early scholarship focusing on internal conflict proposed that there are at least three categories—or levels—of conflict, and that each might be thought of as having unique precipitating conditions. The three categories are (a) political turmoil, (b) organized conspiracy, and (c) internal warfare (Gurr 1970). Political turmoil reflects conditions of spontaneous and unstructured conflict. Food riots might be one example. Such turmoil, however, holds the potential to expand into more organized violence. Organized conspiracies are those conflicts in which opposition groups have some form of structure, but the level of violence remains rather contained. A conflict generally is considered an internal war once the level of fatalities exceeds a certain threshold—most often considered 1,000 battle deaths per year (Small and Singer 1982).

Internal conflicts tend to result from disputes over current or anticipated access to resources. In many instances those controlling the distribution of the resources—usually the state—largely comprise a culturally homogeneous group that discriminates against another (or several) homogeneous groups. For example, in the Rwandan upheaval of 1994 the politically dominant Hutus engaged in the genocidal slaughter of the Tutsi. Although the two groups lived together for centuries, albeit not always peacefully, a power-sharing arrangement between them appears to have altered expectations of the future distribution of resources such that the Hutus chose conflict over accommodation. The main exception to cultural homogeneity forming the organizational lines between the groups in conflict can be found in ideological conflicts. In ideological conflicts the groups form along class lines which can breech cultural distinctions. If one thinks of conflict organized along cultural lines as incorporating two nations within one state, with the conflict waged over control of the state or the distribution of resources from it, then ideological conflict can be thought of as fighting within one nation over two ideologically opposed ideas about how the state should distribute resources.

There are several competing explanations for the causes of internal conflicts. Gurr (1993) has undertaken an extensive examination of the make-up and character of minority groups at risk, identifying 233 politically mobilized groups. The core of his explanation for internal conflict turns on issues of resource distribution, though this is not the only way to think about the etiology of internal warfare. Some propose what is called the 'contact hypothesis' which argues that increased contact between different cultural groups increases the opportunity for and frequency of violent conflict among them (Forbes 1997). Counter arguments point to multiethnic countries like the United States, Canada, and New Zealand to demonstrate that cross-cultural contact is not sufficient to cause conflict. They argue that in highly multiethnic societies the propensity for conflict is diminished. Others argue that conflict results from relative or absolute deprivation (e.g., Gurr 1970, Davies 1962). Relative deprivation reflects judgments about individual or group resources relative to group expectations, what Gurr called value aspirations vs. actual capabilities. Absolute deprivation largely reflects a condition of poverty, and would generally show up as a result of low levels of economic development and/or a skewed distribution of resources (e.g., Muller and Seligson 1987). In spite of the logical robustness of deprivation-based arguments, a cursory examination of conditions in countries around the world reveals that there are large numbers of people living under conditions of relative or absolute deprivation who do not revolt against their government. Clearly being deprived is not a sufficient condition for internal conflict. One way to think about a necessary condition is that the masses will rise up only when expected gains from violent political action outweigh the likely costs. Future expectations form a core component in thinking about the onset of internal violence from this perspective (Fearon 1998). If people hold dire expectations about their prospects for access to resources in the future they will be more likely to engage in violent political action now. Under these conditions—where people react to costs and expected benefits—the ability of the state to repress opposition movements is critical. Convergence around the conditions most associated with internal conflict demonstrates that political regime types play a considerable role in the likelihood of internal unrest (e.g., Henderson and Singer 2000). As countries become partially democratic the people begin to hold increased expectations and make increased demands on the state, yet the state apparatus lacks sufficient institutional capacity to cater to or channel these demands. Some see these internal conflicts as part of the processes contributing to state building (Cohen et al. 1981).

Empirically we know that there have been about 150 violent internal conflicts beyond simple political turmoil since the end of World War II (Regan 2000). According to the Correlates of War data between 1946 and 1992 there have been 53 of conflicts that could be categorized as an internal war (Singer and Small 1994). These conflicts range from rather small and short-lived events like Gambia (1981) to long-running and extremely violent conflicts such as Rwanda (1994), Somalia (throughout the 1990s), or El Salvador (in the

1980s). Internal conflicts are not confined to the less developed part of the world. Great Britain has been fighting an internal conflict against the IRA since the mid-1970s; the Spanish have a conflict over the Basque region, and Turkey has experienced a quite violent conflict with members of its Kurdish population. Each country is a member of the NATO alliance. Russia, furthermore, has been involved in a very aggressive conflict in its republic of Chechnia.

Recent empirical evidence on postcolonial civil wars suggests that the most important factors associated with their onset are: (a) the regime-type of the state, falling somewhere between a full democracy and a complete autocracy, (b) the level of economic development, and (c) the level of military spending by the state. Interestingly, this research does not support the idea that cultural polarization is related significantly to the onset of civil war in the post colonial period (Henderson and Singer 2000, Collier and Hoeffler 2000, Hegre et al. in press). Other recent research has attempted to develop empirical models to forecast the onset of internal rebellion (Gurr and Moore 1998, Gurr 1998). Although the forecasting capabilities of these models are not perfect, they do provide rigorous and generalizable mechanisms to evaluate the risk of internal conflict in specific settings.

The role of parties outside internal conflicts is also of central importance. Even though one could consider third party interventions as one form of conflict management, there is some evidence to suggest that outside interveners tend to increase the duration of internal conflicts (Regan 2000, in press). The internationalization of internal conflicts has become an increasingly important area of research, with emphasis placed on the role of third parties, the effectiveness of peacekeeping forces, and the conditions for negotiated settlements (see Lake and Rothchild 1998, Brown 1996, Regan 2000, Diehl 1993, Walter 2000). This has taken on central importance in international organizations such as the World Bank, which has organized a research group to study the onset and management of internal conflicts (Ibalwaldi and Sambanis 2000, Collier and Hoeffler 2000, web site can be accessed at: www.worldbank.org).

2. *Conflict Management in Internal Conflicts*

One can think of the resolution of internal wars like wars in general as being most likely when the costs of continued fighting exceeds expectations for future gains. If the probability of victory is low and/or the expected payoff from victory is insufficient, then one would anticipate that the parties would prefer a negotiated settlement. Resolving internal conflicts, then, requires generating the conditions whereby continued fighting is the least attractive course of action. Internal conflicts, however, pose different types of problems when thinking about conflict management than international wars. First, international convention inhibits outside parties from intervening on behalf of at least one side (the nongovernmental actors) in such conflicts. Although third parties do intervene on behalf of antigovernment forces, such interventions tend to violate at least the spirit of the UN charter. Second, there is little incentive for the parties involved in the conflict to demobilize if and when a settlement is reached (Walter 2000). Since difficulties with enforcement leads to a built-in incentive to violate a peace accord, such as by failing to demobilize, fully implemented peace agreements are difficult to achieve without some form of outside intervention. International management of internal conflicts requires outside actors to intervene in a manner that contributes to the conditions where conflict is too costly, the benefits from resolution are reasonably high, and all sides to the conflict hold similar expectations about implementation. The most pressing question is how outside parties can generate the conditions that lead to a stable and fully implemented peace agreement. Although the amount of empirical evidence is rather meagre there are a few studies that point toward effective mechanisms.

Using data on 139 internal conflicts between 1944–94 Regan (2000) identified a set of conditions that increase the likelihood of successful third party interventions in internal conflicts. Success was defined as stopping the fighting for a period of at least 6 months, and, therefore, reflects a short-term interest in the settlement of conflicts rather than a long-term resolution. The results of the analysis suggests that intense conflicts are more likely to be settled by outside interventions than are those with lower levels of hostilities, and that the most important predictor of successful interventions involves the interveners strategy. Walter has also shown that internal wars rarely are resolved without power sharing arrangements and security guarantees by outside parties (Walter 2000). In effect, there is little chance of the parties to an internal war resolving their conflict without the involvement of outside parties. These parties might take on the role of a supporter of one of the antagonists (in an attempt to alter the balance of capabilities), a unilateral observer of the terms of an agreement, or multilateral supervisor of an interim agreement (Walter 2000, Diehl 1993, Regan 2000). There is some empirical evidence, however, that when a third party intervenes in support of one side of an internal conflict it is likely that the conflict will have a longer expected duration than a similar conflict without an outside intervention (Regan in press). The role of unilateral interventions, therefore, may have a very narrow role as a tool of conflict management.

3. *Summary*

Since the end of World War II internal conflicts have been much more frequent than interstate wars. In fact, some estimate that the number of deaths attributable

to internal conflicts far exceed those of all interstate wars of the twentieth century. The causes of internal conflicts are rather complex, though empirical evidence and theoretical logic point toward competing access to limited resources. When resources are limited and people make demands for them, the ability effectively to channel or repress these demands determines the likelihood of violent conflict. Because of the unique aspects of territorial sovereignty and nonstate actors, the resolution of these conflicts poses problems for the international community. Peacekeeping troops under the auspices of the UN or other multinational organizations have become an increasingly visible mechanism to control internal conflict (see Diehl 1993).

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict: Anthropological Aspects; Conflict/Consensus; Conflict Sociology; Feud and Internal War: Legal Aspects; Warfare in History

Bibliography

Brown M E (ed.) 1996 *The International Dimensions of Internal Conflict*. MIT Press, Cambridge, MA

Cohen Y, Brown B R, Organski A F K 1981 The paradoxical nature of state making: The violent creation of order. *American Political Science Review*. **75**(4): 901–10

Collier P, Hoeffler A 2000 Greed and grievance in civil war. Unpublished manuscript. Post Conflict Fund of the World Bank, Washington, DC

Davies 1962 Toward a theory of revolution. *American Sociological Review* **27**(1): 5–19

Diehl P F 1993 *International Peacekeeping*. The Johns Hopkins University Press, Baltimore, MD

Elbadawi I A, Sambanis N 2000 External Interventions and the Duration of Civil Wars. Unpublished manuscript. Post Conflict Fund of the World Bank, Washington, DC

Fearon J D 1998 Commitment problems and the spread of ethnic conflict. In: Lake D A, Rothchild D (eds.) *The International Spread of Ethnic Conflict*. Princeton University Press, Princeton, NJ, pp. 107–26

Forbes H D 1997 *Ethnic Conflict: Commerce, Culture, and the Contact Hypothesis*. Yale University Press, New Haven, CT

Gurr T R 1970 *Why Men Rebel*. Princeton University Press, Princeton, NJ

Gurr T R 1993 *Minorities at Risk: A Global View of Ethnopolitical Conflict*. United States Institute of Peace Press, Washington, DC

Gurr T R 1994 Peoples against states: Ethnopolitical conflict and a changing world system. *International Studies Quarterly* **38**(September): 347–77

Gurr T R 1998 A risk assessment model of ethnopolitical rebellion. In: Davies J L, Gurr T R (eds.) *Preventive Measures: Building Risk Assessment and Crisis Early Warning Systems*, Rowan & Littlefield, Lanham, MD

Gurr T R, Moore W H 1998 Ethnopolitical rebellion: A cross-sectional analysis of the 1980s with risk assessments for the 1990s. *American Journal of Political Science* **41**(4): 1079–1103

Hegre H, Ellingsen T, Gites S, Eladitsch P in press Toward a democratic civil peace? *American Political Science Review*

Henderson E A, Singer J D 2000 Civil war in the post-colonial world, 1946–92. *Journal of Peace Research* **37**(3): 275–99

Lake D A, Rothchild R (eds.) *The International Spread of Ethnic Conflict*. Princeton University Press, Princeton, NJ

Muller E N, Seligson M A 1987 Inequality and insurgency. *American Political Science Review* **81**(2): 425–51

Regan P M 2000 *Civil Wars and Foreign Powers: Outside Intervention and Intrastate Conflict*. University of Michigan Press, Ann Arbor, MI

Regan P M in press Third party interventions and the duration of intrastate conflicts. *Journal of Conflict Resolution* (Feb 2002)

Singer J D, Small M 1994 *International and Civil War Data, 1816–1992. Correlates of War Project*. University of Michigan, Ann Arbor, MI

Walter B 2000 Negotiating Settlements to Civil Wars. Book Manuscript. University of California, San Diego, CA

P. M. Regan

International Advertising

International advertising entails dissemination of a commercial message to target audiences in more than one country. Target audiences differ from country to country in terms of how they perceive or interpret symbols or stimuli, respond to humor or emotional appeals, as well as in levels of literacy and languages spoken. How the advertising function is organized also varies. In some cases, multinational firms centralize advertising decisions and budgets and use the same or a limited number of agencies worldwide. In other cases, budgets are decentralized and placed in the hands of local subsidiaries, resulting in greater use of local advertising agencies.

International advertising can, therefore, be viewed as a communication process that takes place in multiple cultures that differ in terms of values, communication styles, and consumption patterns. International advertising is also a business activity involving advertisers and the advertising agencies that create ads and buy media in different countries. The sum total of these activities constitutes a worldwide industry that is growing in importance. International advertising is also a major force that both reflects social values and propagates certain values world wide.

1. International Advertising as a Communication Process

In international markets the process of communicating to a target audience is more complex because communication takes place across multiple contexts, which differ in terms of language, literacy, and other

cultural factors. In addition, media differ in their effectiveness in carrying different appeals. A message may therefore not get through to the audience because of people's inability to understand it (due to literacy problems), because they misinterpret the message by attaching different meanings to the words or symbols used, or because they do not respond to the message due to a lack of income to purchase the advertised product. Media limitations also play a role in the failure of a communication to reach its intended audience.

The process of communication in international markets involves a number of steps. First, the advertiser determines the appropriate message for the target audience. Next, the message is encoded so that it will be clearly understood in different cultural contexts. The message is then sent through media channels to the audience who then decode and react to the message. At each stage in the process, cultural barriers may hamper effective transmission of the message and result in miscommunication.

In encoding a verbal message, care needs to be taken in translation. Numerous examples exist of translation problems with colloquial phrases. For example, when the American Dairy Association entered Mexico with its 'Got Milk?' campaign, the Spanish translation read 'Are You Lactating?' Low levels of literacy may result in the need to use visual symbols. Here again, pitfalls can arise due to differences in color association or perception. In many tropical countries, green is associated with danger and has negative connotations. Red, on the other hand, is associated with weddings and happiness in China. Appeals to humor or sex also need to be treated with considerable care as their expression and effectiveness varies from one culture to another. The dry British sense of humor does not always translate effectively even to other English-speaking countries.

In addition to encoding the message so that it attracts the attention of the target audience and is interpreted correctly, advertisers need to select media channels that reach the intended target audience. For example, use of TV advertising may reach only a relatively select audience in certain countries. Equally, print media will not be effective where there are low levels of literacy. Certain media may also be more effective in certain cultures. For example, radio advertising has substantial appeal in South America where popular music is a key aspect of the local culture.

The cultural context also impacts the effectiveness of communication. In what has been termed 'high context' cultures, such as Japan and China, the context in which information is embedded is as important as what is said (Hall 1976). In 'low context' cultures, which include most Western societies, the information is contained in the verbal messages. In these cultures, audiences tend to demand explicit informational content (De Mooij 1998). Conversely, people in high context cultures are often more responsive to image or mood appeals, and rely on personal networks for informational content.

2. *International Advertising as a Business Practice*

When a firm develops an international advertising strategy, one important issue is whether to develop a global or regional advertising campaign or tailor communications to differences in local markets (Peebles and Ryans 1984). If the purpose of advertising is to develop a strong corporate or global image, a uniform global campaign is more likely to be used. When, on the other hand, the objective is to launch a new product or brand, or to more clearly differentiate the product or brand from other competing brands or products, local campaigns tailored to local markets are more typical.

A global campaign offers a number of advantages. In the first place, it can be an important means of building a strong and coherent global image for the firm and/or its products worldwide. Use of the same image in different countries builds familiarity and generates synergies across world markets. It allows utilization of good ideas and creative talent (both of which are scarce commodities) on a worldwide basis. In addition, use of a single campaign provides substantial cost savings in copy development and production costs. Conversely, development of multiple local campaigns can lead to duplication of effort, result in inconsistent brand images across countries, and confusion in consumers' minds with regard to the benefits offered by the brand and corporate image.

While use of uniform advertising appeals offers a number of advantages, differences in customer perceptions and response patterns across countries and cultures, as well as media availability and government regulation, are major barriers to the use of a standardized campaign. Even though technological developments allow adaptation of advertising appeals to different languages (for example, TV can have audio channels in two languages, Internet messages can be automatically translated), development of visual and verbal copy that works effectively in multiple countries poses major creative challenges.

Faced with this dilemma, firms may use a global umbrella campaign combined with local country or product-specific advertising. The global umbrella campaign develops a uniform image for the company or brand worldwide, often relying on consistent visual images and the corporate logo. Product-specific or country advertising builds on this image, modifying the appeal and providing information tailored to the local market (see Rijkens 1992). The objective of the umbrella campaign is to provide an integrating force, while local campaigns provide greater relevance to specific local customers and markets. The organiz-

ational structure of the firm often plays a key role in the choice of global vs. locally adapted campaigns. If international operations are organized on a country-by-country or geographic basis and operate as local profit centers with local advertising budgets, pressures exist for the use of local advertising campaigns. If, on the other hand, the company is organized by product divisions, with centralized advertising budgets at corporate or regional headquarters, the use of regional or global advertising campaigns is more likely (see Douglas and Craig (1995) for more information on global strategy).

3. *International Advertising as an Industry*

The world advertising industry is characterized by a large number of small and medium-sized advertising agencies that operate primarily in one country, and by a small number of very large advertising agencies with operations in many countries. These agencies have developed extensive networks of offices throughout the world in order to coordinate the advertising process in all the countries where their clients do business. These networks often include both wholly-owned subsidiaries and formal relationships with local advertising agencies to establish a presence in new markets, particularly in emerging markets.

In an effort to establish greater control over their advertising, many major advertisers are consolidating all their advertising with one agency. For some major advertisers such as IBM and Citibank, this represents annual advertising expenditures in excess of $500 million worldwide (Grein and Ducoffe 1998). As a consequence, advertising agencies that do not have a global network are at a serious disadvantage when competing for new advertising accounts or attempting to retain existing ones that are expanding globally.

The majority of these large advertising agencies are headquartered in the US. Of the 10 largest advertising agency groups, six are headquartered in the US, two in France, and one each in Japan and the UK. The largest agency group, WPP, is a British agency holding company made up of three large US-based agencies. With the exception of Dentsu, the Japanese agency, most other agency networks generate the majority of their revenues outside their home country. The second largest agency group, Omnicom, places over $45 billion of advertising for its clients around the world and derives half its revenue from outside the US. Omnicom has 891 offices in over 85 countries and employs 35,600 persons worldwide (57 percent work outside the US). US-based advertising agencies and their subsidiaries are responsible for most of the advertising throughout the world. Consolidation and concentration are major trends. The three large agency groups (WPP, Omnicom, and Interpublic) have grown through aggressive acquisitions and now account for 39 percent of the world's advertising expenditures (*Advertising Age* 2001). Of the approximately $60 billion in advertising placed by the top 25 agency networks in Europe during 1995, 89 percent of the total was placed by subsidiaries of US-based agencies. This general pattern holds in most parts of the world that do not have restrictions on foreign ownership. The major exception is Asia, where the three major Japanese agencies account for 62 percent of the advertising placed by the top 25 agency networks. Current and comprehensive information on advertising can be obtained from *Advertising Age's* website, www.adage.com.

Table 1
Advertising spending in the top 10 global ad markets

Country	1997 advertising expenditures (millions)	Percent
US	117.0	50
Japan	35.7	15
UK	20.8	9
Germany	20.3	9
France	9.7	4
Brazil	8.8	4
Italy	7.2	3
Australia	5.5	2
Canada	5.4	2
S. Korea	5.3	2
Total	235.7	100

The expenditures reflect only the 13 measured media tracked by *Advertising Age* such as TV, magazines, radio, the Internet, and Yellow Pages. The $400 billion figure cited on the next page includes unmeasured spending such as direct mail, promotion, co-op advertising, and catalogues

Worldwide over $400 billion is spent on advertising. Approximately half of that amount is spent in the US and the other half outside the US. Information on advertising spending can be obtained from *Advertising Age's* website and from McCann-Erickson's website (www.mccann.com). The bulk of expenditures outside the US take place in Europe and Japan, although Brazil, Canada, Mexico, and Australia are also important advertising markets. Outside of these markets, China is the next largest advertising market and is also growing rapidly.

Table 1 shows advertising spending in the top 10 global ad markets. The US and Japan account for 65 percent of the total advertising spending in these markets and the four European markets account for an additional 25 percent. Outside of these markets, no other market, except for Brazil, accounts for more than 2 percent of the total spending. The concentration of spending in the US in part explains the dominance of US-based advertising agencies. Not only do they work for US-based clients that continue to expand outside the US, but also they accumulate knowledge

and experience in the practice of advertising that can be applied elsewhere.

Once the advertising message has been created, a media plan must be developed and specific media vehicles purchased to deliver the message to the target audience. Media differ from country to country in their availability, effectiveness, and efficiency in delivering a message, and with relatively few exceptions, tend to be organized on a country-by-country basis. Notable exceptions include StarTV, MTV, CNN in television, *Business Week International*, the *Asia Wall Street Journal*, the *International Herald Tribune* in print, and selected industry and medical publications that are read worldwide. There is also a trend toward consolidation of media in order to achieve greater economies of scale and leverage content developed in one market to others. This consolidation facilitates the purchase of media on a regional and global basis. In addition, the Internet is emerging as a truly global medium that does not conform to country boundaries.

4. International Advertising as a Social Force

In the view of the advertiser, the primary objective of advertising is to sell products or services. In achieving this primary goal, there are often profound secondary consequences. Advertising exerts a formative influence whose character is both persuasive and pervasive. Through the selective reinforcement of certain social roles, language, and values, it acts as an important force fashioning the cognitions and attitudes that underlie behavior not only in the marketplace, but in all aspects of life. In an international setting, advertising has an important social influence in a number of ways. First, much international advertising is designed to promote and introduce new products from one society into another. Often this results in radical changes in lifestyles and behavior patterns of a society, stimulating for example the adoption of fast food, casual attire, or hygiene and beauty products. International advertising also encourages desire for products from other countries, it creates expectations about ' the good life,' and establishes new models of consumption. Advertising is thus a potent force for change, while selectively reinforcing certain values, lifestyles, and role models.

Often the symbols, ideals, and mores that international advertising portrays and promotes are those of Western society and culture. Through global advertising, brands such as Levi's, Nike, Marlboro, and McDonalds are known by and have become objects of desire for teens and young adults throughout the world. Similarly, images and scenes depicted in much international advertising are either Western in origin or reflect Western consumption behavior and values. Even where adapted to local scenarios and role models, those shown often come from sectors of society, such as the upwardly mobile urban middle class, which embrace or are receptive to Western values and mores.

Consequently, a criticism frequently leveled at international advertising is that it promulgates Western values and mores, notably from the US, in other countries. This is viewed particularly negatively in societies with strong religious or moral values which run counter to those of the West as, for example, Islamic societies in the Middle East. When Western advertising depicts sexually explicit situations or shows women in situations considered as inappropriate or immoral, it is likely to be considered a subversive force undermining established cultural mores and values. Equally, in some countries such as France, there is a strong negative reaction to the imposition of US culture, values, and use of English in advertising. Promotion of tobacco products by US and UK companies in countries where there is no legislation regulating or banning cigarette advertising has also been criticized.

At the same time, international advertising also acts as an integrating force across national boundaries. It disseminates messages using universal symbols and slogans, and establishes a common mode of communication among target audiences in different parts of the world. At the same time, multicultural values are reinforced by advertisers, who adopt images incorporating peoples of different nations and diverse cultural backgrounds, as, for example, the Colors of Benetton campaign or the British Airways 'Peoples of the World' campaign. The impact of such campaigns is further reinforced by the growth of global media such as Star TV, CNN, MTV or print media that target global audiences worldwide.

Consequently, while, on the one hand, international advertising can be viewed as a colonizing force propagating Western values and mores throughout the world, it is also an important force integrating societies and establishing common bonds, universal symbols, and models of communication among peoples in different parts of the globe.

See also: Advertising and Advertisements; Advertising: Effects; Advertising: General; Globalization and World Culture; Globalization, Anthropology of; Information Society; International Business; Mass Media and Cultural Identity; Mass Media, Political Economy of; Media Imperialism; Media, Uses of; Multinational Corporations; Telecommunications and Information Policy

Bibliography

2001 *Advertising Age* April **23**: 1

De Mooij M 1998 *Global Marketing and Advertising: Understanding Cultural Paradoxes*. Sage Publications, Thousand Oaks, CA

Douglas S P, Craig C S 1995 *Global Marketing Strategy*. McGraw-Hill, New York
Grein A, Ducoffe R 1998 Strategic response to market globalization among advertising agencies. *International Journal of Advertising* **17**: 301–19
Hall E T 1976 *Beyond Culture*. Anchor Press, Garden City, NY
Peebles D M, Ryans J K Jr. 1984 *Management of International Advertising: A Marketing Approach*. Allyn and Bacon, Boston
Rijkens R 1992 *European Advertising Strategies*. Cassell, London

S. P. Douglas and C. S. Craig

International and Transboundary Accords, Environmental

The need for international environmental accords, agreements specifying the terms of environmental policy coordination among participating governments, arises from the political fragmentation of the world into more than 190 separate states—the boundaries of which generally fail to correspond to the limits of the habitats, ecosystems and bioregions comprising the Earth's biosphere. International environmental accords range from 'soft law' declarations of goals and principles or programs of action outlining desirable priorities and guidelines for conduct, to 'hard law' agreements specifying policy measures that participating governments obligate themselves to undertake, and, in many cases, creating joint institutions for ongoing policy review and adjustment (Dupuy 1992). They can take the form of bilateral treaties between pairs of states, regional treaties among a few states, or global treaties involving many states. This article focuses on the regional and global agreements, tracing their number and subject matter, outlining their main features, and indicating the main directions of social science research into their effectiveness.

1. Subject Matter and Number of International Environmental Accords

International environmental problems can be divided into four broad types: management of international common pool resources, management of international common sinks, redress of transboundary harm and regulatory harmonization. Common pool resources—such as oil fields or fish stocks—exist in such close physical inter-relation that they cannot be easily divided into separate portions for each user, with the consequence that overhasty or overintense exploitation by any one user depletes the resource for all users. Common sinks are physical places like the atmosphere, a river, or an open tract of land which large numbers of people can reach for disposing of unwanted substances. Common sinks become environmental problems when human injection of gaseous emissions, liquid effluents or solid by-products of agriculture and industrial production are made at such a rate and/or in such aggregate amounts that the sink's assimilative capacity is exceeded. International common pool resources and international common sinks can be present either in common areas outside the jurisdiction of individual states (the high seas, the atmosphere above the high seas, outer space, Antarctica) or in areas that straddle the boundaries of two or more states.

Both domestically and internationally, common pool resources and common sinks are the exception; the majority of the world's resources and spaces can be divided into separate portions permitting each user (whether individual, corporate, or communal) to adopt effective management policies on its own. Yet overexploitation of resources that leads to environmental degradation (such as soil erosion from clear-cutting forests) and excessive emissions from industrial production or concentrated urban populations can produce negative effects for others. When these negative effects cause transboundary harm by extending into the territory of neighboring states, governments face demands that they fulfill the obligation to ensure that activity within their borders does not harm persons or places outside. Regulatory harmonization occurs when governments perceive that differences in national rules are hindering efforts to protect species or habitats, or when governments, firms, or both, believe that variation in national rules cause enough difference in the cost of economic activity to give firms in one state a significant advantage over competitors elsewhere.

Nineteenth-century international treaties addressing what we now define as environmental problems were typically bilateral or regional accords about sharing rivers that form international boundaries or traverse the territories of more than one state, reversing localized overexploitation of fish or marine mammals on the high seas, controlling the spread of disease, or protecting migratory wildlife (Ruster et al. 1983). Thus, most of them addressed common pool resources problems and a minority involved regulatory harmonization. The number of environmental treaties and the range of problems they addressed increased steadily after World War I. Of the 62 regional or global environmental agreements concluded between 1920 and 1972 (UNEP 1990), 15 addressed high-seas fishing, 12 marine pollution, 10 human health hazards, 10 wildlife or plant protection and five shared river resources. Of the 103 concluded between January 1973 and December 1998 (UNEP 1990, http://www.sedac.ciesin.org/entri/texts-home.html, http://www.tufts.edu/fletcher/multilaterals.html), 65 addressed the familiar problems: marine pollution (36), high seas fisheries (11), human health hazards (11), plant protection (3) and shared river resources (4), while the other 38 addressed new issues: general environmental

cooperation (17), atmospheric pollution (10), nonhuman diseases (6), waste transport (3), and river pollution (2). The balance among types of problems addressed also shifted. The 1920–72 treaties included 24 addressing international common pool resources (mainly high seas fishing), nine protection of international common sinks (mainly the oceans), 11 transboundary harm and 18 regulatory harmonization. The post-1973 treaties include 20 addressing common pool resources (primarily high seas fishing), 36 common sinks (oceans and the atmosphere), 14 transboundary harm and 33 regulatory harmonization.

This broadening of attention was reinforced, and the character of international environmental accords altered fundamentally, by the reconceptualization of environmental problems promoted through the 1972 United Nations Conference on the Human Environment (Stockholm Conference). Previous international environmental agreements addressed problems as isolated difficulties, usually only after the negative effects of current activity were highly visible. Later accords have been informed by clearer understandings of ecology, visions of ecological sustainability, and awareness of the interrelatedness of problems (Schachter 1991, Caldwell 1996). Some, like the 1971 Ramsar Convention on Wetlands, the 1980 Convention on the Conservation of Antarctic Marine Living Resources and the 1992 Convention on Biodiversity, explicitly adopt ecosystems management approaches; even the larger number focused on a particular problem define that problem in ecological terms and reflect growing consensus that problems should be addressed before negative effects become serious. The diffusion of new scientific knowledge informed by ecological paradigms to policy-makers and publics, and the accession of more people sharing ecologically informed beliefs to influential positions in governments, international organizations, nongovernmental organizations and business firms has helped entrench the new beliefs. Citizens, experts, business executives, and political leaders alike are now aware that natural ecosystems and human–environment interactions involve dynamic processes, and that successful environmental policy coordination requires long-term cooperation, periodic assessment of results and frequent policy revision.

2. *Features of Contemporary International Environmental Accords*

Though each global or regional multilateral environmental accord contains unique provisions (see descriptions in e.g., Haas et al. 1993, Cameron et al. 1996, Victor et al. 1998, Grubb et al. 1999, Miles 1999, Wettestad 1999, Earthscan and Fridtjof Nansen Institute 1998/99), the most prominent of them share a number of broad features. Most notably, they establish an organizational structure for ongoing consideration of the problem and potential solutions, which involves a conference of the parties (representatives of all participating governments) assisted by an array of advisory committees and a secretariat. This and other features are institutional responses to the need for frequent revision, the desire to expedite the process of bringing international agreements into effect, the character of the actual addressees of policy measures, and the need to assess whether and to what extent environmental improvement is occurring by direct observation of natural ecosystems.

The general consensus that international environmental accords must be designed for frequent revision derives from two forms of uncertainty. The first is the initial inability to identify the full contours of the environmental problem at hand and the type and extent of action needed to redress it, an inability that often makes governments hesitant to accept binding obligations. Uncertainty about the physical aspects of the problem is usually addressed before or during negotiations by convening scientific panels to secure or assemble relevant data, assess its implications, and present the results in a form that policy-makers will understand—as happened in negotiations regarding transboundary air pollution in Europe, protection of the ozone layer and climate change. Uncertainty about the remedial action needed is addressed by adjusting the type and extent of policy measures. Governments often start by agreeing only to cooperate in elucidating the problem or to adopt 'soft law' statements of desirable policy directions. Only when they are confident that they understand the basic contours of various possible solutions do they move towards the 'hard law' of specific and binding commitments. The second and more persistent form of uncertainty arises from the dynamic character of nature, human–nature interactions, technological innovation and continual revision of scientific knowledge. More than in other areas of international cooperation, participants realize that even the best current understanding is provisional and likely to be revised in light of additional experience or research.

Consensus on the need to facilitate revision of international environmental accords has encouraged extensive use of a new form of international agreement developed in a number of fields after World War II: the framework treaty plus associated protocols (Kiss 1993). The framework treaty specifies the problem to be addressed, the common objectives, and the collective procedures for organizing cooperation while the protocols (sometimes called annexes or schedules) specify the particular regulatory and programmatic measures participating governments agree to implement. This form was foreshadowed in the river-sharing and fisheries agreements of the late nineteenth and early twentieth centuries creating intergovernmental commissions whose suggestions of particular policy measures often formed the basis for additional treaties

and explicitly adopted in a number of technical fields, such as safety of navigation and use of the radio frequency spectrum. Current framework treaty-protocol arrangements link the initial and later agreements more clearly, and provide an organizational platform supporting current efforts to expedite the process of international rulemaking.

Both of the organizational forms common in contemporary international environmental accords—periodic conferences of the parties and heavy reliance on scientific and technical advice provided by committees of relevant experts—also have long antecedents. The early river and fisheries commissions met periodically and secured scientific advice either from a specially constituted scientific advisory committee or an existing international scientific body. The conferences or meetings of the parties institutionalized in contemporary accords differ from the meetings of the early international commissions mainly by involvement of a larger number of governments, participation by higher level officials, greater elaboration of procedures for securing expert advice and greater attention by a wider array of nonstate actors.

The widespread sense of urgency regarding environmental problems has encouraged extensive use of various procedures for expedited international rule-making. In the nineteenth century, multilateral agreements could be opened for acceptance only by unanimous agreement of the states participating in the conference drawing them up (a rule tempered by the deference of others to the views of the great powers), became binding only on those governments which formally ratified the result and did not come into effect even for ratifying governments until all or most of the governments represented at the negotiating conference had ratified. Even with twentieth-century relaxations of the unanimity rule, the need for ratification often means a five to ten year delay between the end of negotiations and the treaty's taking effect. Waiving the ratification requirement is one way to expedite rule-making. A few agreements concluded in the mid-twentieth century, including the 1946 International Convention on the Regulation of Whaling and agreements on aviation safety negotiated under International Civil Aviation Organization auspices, substituted an 'opt out' procedure for changing technical policy measures (the sort of detail now placed in protocols). Under this procedure, an agreed change goes into effect for all participating governments that have not filed a written objection by a particular date. What this means for implementation depends on who objects. Any objection introduces greater unevenness into implementation; objection from a government that has jurisdiction or control over the activity causing a large part of the problem effectively halts change. As the experience of international fisheries commissions shows, the prospect of objections from major governments create strong pressures for 'least common denominator' agreements. Yet an 'opt out' clause also means that objectors cannot prevent supporters from proceeding, freeing the latter from the need to persuade even the most skeptical before they try out some policy measure.

The framework treaty-protocol form facilitates taking advantage of other devices for expedited international rulemaking. Most contain two-tier amendment clauses allowing new or modified protocols to be opened for acceptance by smaller majorities of participating governments than are required for proposing changes to the framework treaty. The process of bringing new protocols or modifications into effect has been speeded up by providing for immediate 'provisional application' pending acceptance, stipulating shorter deadlines for filing objections, or even, as with 'adjustments' to deadline or emissions reductions in the Montreal Protocol, allowing a consensus or qualified majority vote in the meeting of parties to establish an immediately binding obligation.

Though international environmental accords are addressed to governments, the policy measures they specify are meant to affect the behavior of business firms, social groups, households and individuals. Participating governments therefore promise not only to instruct their own officials and agents to act in particular ways but to use their administrative apparatus to affect the actions of private persons and entities. The fact that private individuals, groups and entities are the ultimate addressees of international environmental accords affects both their initial negotiation and their subsequent implementation.

Democratic norms of domestic governance and governments' awareness that channeling behavior is much easier when policies elicit voluntary compliance, both press governments toward allowing interest groups, social movements and citizens—many of whom have strong material stakes in or strong views about potential policy measures—a voice in the formulation of regulatory measures. This complicates the negotiation of international environmental accords by linking the intergovernmental discussions to both the domestic politics of individual states (particularly the larger industrial ones) and the transnational politics of the environmental movement (for example, see *Social Movements: Environmental Movements*; *Environmental Policy: Protection and Regulation*).

Implementation of agreements is more complicated because affecting the behavior of citizens and private entities requires greater administrative capabilities than simply instructing and supervising officials. The relevant expertise, resources and administrative skills are distributed very unevenly, with the governments of many developing states hard-pressed to develop and maintain them. Those governments need help building capacity to administer the policies they promise to carry out. Awareness that noncompliance may be the result of inability rather than will has altered perceptions of the compliance problem by fostering the

realization that punishments or rewards, which rest on assumptions that the recipient has the capacity to comply and only needs the spur of changed incentives, will not work with those who lack capacity. Striking a balance between capacity-building and incentive-shifting is the subject of an ongoing debate between advocates of a 'management approach' focused on removing obstacles to compliance through capacity-building (e.g., Chayes and Chayes 1995) and advocates of an 'enforcement approach' stressing the continued relevance of rewards and punishments (e.g., Downs et al. 1996).

The general agreement that developing more systematic compliance assessment is important to improving the implementation of international environmental accords has inspired development of more coherent 'systems of implementation review' (Victor et al. 1998) drawing on precedents developed in the fields of labor standards, human rights and development aid. They combine periodic government self-reporting of activities, comments on those reports or periodic policy review, and in some cases receipt of information from private entities with the traditional devices of intergovernmental complaint and third party dispute resolution. They are managed through new organizational structures, typically small intergovernmental implementation committees reporting to the conference of the parties. Though the details of these review processes vary from treaty to treaty, implementation committees have favored a low-key, nonconfrontational approach premised on the belief that some slippage is inevitable while governments acquire needed resources and learn to cope with their new regulatory tasks.

Even intentional noncompliance is handled in new ways. Most international environmental accords do not employ one of the primary traditional remedies for noncompliance with treaties: retaliatory violation by another participating government. Counterviolation is effective when obligations are reciprocal and counterviolation exerts direct pressure on the noncomplying government, as is common in international trade. It is a less useful device in environmental affairs because few counterviolations put pressure on a violator, and even when they do they take governments further from their joint goal by producing more environmental harm. In response, governments have emphasized other forms of pressure. Implementation committee meetings and conferences of the parties offer ample opportunities for application of peer pressure and shaming. Some multilateral environmental treaties include provisions for trade sanctions. The project evaluation processes included in aid programs instituted in response to Third World governments' arguments that the industrial states, as causers of most current environmental harm, have the obligation to bear most of the cost of redressing it, provide opportunities for guiding recipient governments towards particular measures. Public concern about environmental problems and the strong transnational organization of the environmentalist movement create the conditions for supplementing—or even substituting for—inter-government pressures by nonstate actor provision of information that would otherwise go unnoticed, public shaming through media and protest campaigns, or organization of consumer boycotts.

The need to assess the results of cooperation separately from levels of compliance stems from the character of environmental problems. Most international agreements, including those on human rights, trade and arms control, establish human-designed conventions and the desired result flows directly from obeying the terms of the agreement. Environmental agreements, in contrast, deal with physical systems having autonomous dynamics. Even complete compliance with treaty commitments will fail to yield environmental improvement if those commitments are too weak or rest on a faulty analysis of the cause and effect relationships between human conduct and environmental degradation.

Determining the existence and extent of environmental improvement requires ongoing monitoring of those portions of the biosphere addressed by the treaty, and periodic scientific assessment of their current and likely future condition. Under most international environmental accords, the monitoring function is performed cooperatively, with governments or national scientific institutes providing parallel observations of conditions in different places derived from terrestrial, marine and atmospheric observation or satellite remote sensing (for example, see *Remote Sensing and Geographic Information Systems Analysis*). The integration of these data and analyses of what they convey about current and future natural conditions is performed by a scientific committee composed of relevant experts drawn from many countries that typically reports six months to a year in advance of the conference of the parties at which the assessment will be considered.

When, as is often the case, the scientific assessment indicates a need to alter policy measures, a separate process of identifying potential solutions and selecting among them comes into play, a discussion that hinges on technological feasibility and economic costs and benefits. Recent international environmental accords make provision for securing expert knowledge in these areas through advisory committees of economic and engineering experts, also drawn from various participating states, whose task is to suggest new or revised policy measures to the conference of the parties. These committees also contribute expertise that the political and diplomatic generalists at the conference usually lack and, if they include members of business firms and environmentalist organizations as well as government-employed experts, provide opportunities for expression of nonstate actors' views on potential regulatory measures.

Neither the scientific nor the technical assessment processes involves substituting 'expertise' for 'politics,' because decisions about whether and how to revise the policy measures specified in the treaty depends on governments' calculations of interest and hard political bargaining. However, the advisory committees do form an essential part of the treaty-based 'preordained learning process' (Sand 1996, p. 792) by helping policy-makers update their understanding of the biosphere, human–biosphere interactions and feasible policy measures.

3. *Factors Influencing the Success of International Environmental Accords*

Efforts to identify the conditions facilitating or inhibiting the negotiation and implementation of international environmental accords have focused on three broad sets of factors: institutions, actor beliefs and material conditions. There is as yet little consensus on how these factors, singly or in combination, contribute to success or failure. This stems partly from the scarcity, until very recently, of carefully designed comparative case studies, but even more from the use of two very different conceptions about how actors think and choose: a rational choice account, regarding actors as utility maximizers with stable interests focused primarily on material costs and benefits, and a social practices account, regarding actors as developing their understandings of themselves and their interests from membership in a wider society and socialization into its norms, discourses and collective learning processes.

Rational choice approaches treat actor interests and motivations as stable and inquire into how material conditions and institutions affect actor choice. Insights into the impact of material conditions and the possible roles of institutions in facilitating negotiation and implementation of agreements have been developed in two ways, by looking at particular features of activity and by classifying problems into broad types posing distinct cooperative challenges because of the way in which actor calculations of self-interest intersect in interaction. The first stream of work has highlighted such considerations as the greater difficulty of regulating large numbers of actors, each creating a small part of the problem, than a small number, each creating a significant part of it; the different opportunities for particular governments to regulate depending on whether those involved in an activity move from place to place (as in shipping) or remain in one place (as in mining); the impact of linkage or separation between costs and benefits of better environmental practice (for instance, the great difficulty of securing cooperation from upstream or upwind polluters); and the influence of changing cost–benefit ratios of environmentally superior action as new technology is developed or regulations alter the relative cost of continuing old rather than adopting new forms of activity.

The second stream, drawing strongly on economic theories of property rights regimes, provision of collective goods and government regulation of markets (e.g., Sandler 1997, Symes 1998), has identified the distinct regulatory challenges posed by different types of situations. In this view, the management of common pool resources and common sinks poses the toughest challenges to successful cooperation because they involve collaboration situations in which actors remain tempted to maximize short-term gain by violating their commitments. The common image of these problems as resembling a 'prisoner's dilemma' (Hardin 1968) leads analysts to conclude that effective institutions are those that reduce actor temptation through lengthening their time horizons and deterring cheating by monitoring and enforcement, tasks more difficult when dealing with large rather than small communities (see *Environment and Common Property Institutions*). Transboundary environmental problems, in this analysis, are international level expression of the familiar problem of negative externalities, complicated by the fact more than one government is involved (see *Environmental Policy: Protection and Regulation*). The range of policy solutions suggested for single countries, ranging from central government 'command and control' measures to 'market-based measures' resting on conducive property rights, taxation and regulatory regimes, can work across borders only if governments cooperate quite closely. Yet, as rational choice analysts have noticed, this cooperation is hindered because governments (and other actors) have a strong stake in continuing to use the regulatory scheme with which they are already familiar. Rational choice analysts treat regulatory harmonization as a coordination problem in which firms and governments have different preferences about regulation but little incentive to depart from an agreed regulatory scheme once it is established. Here, institutions can be used to increase the likelihood of cooperative action by augmenting the flow of information about actor preferences and behavior, providing possibilities of reducing discontent about differences in costs or benefits through concessions on related issues, or reducing transaction costs by substituting group discussions for one on one bargaining.

Rational choice analysts anticipate that the substantive provisions of international environmental accords will have a strong focus on enforcement mechanisms when addressing common pool resources or common sinks, and a strong focus on information flows, compensation mechanisms and reducing transactions costs when addressing regulatory harmonization. However, the actual development of international environmental accords has not conformed to these expectations. Through addressing these matters current international environmental accords reflect greater concern for developing

common information about the environmental problem and fostering joint deliberation regarding standards for proper conduct. This has provided openings for the social practices theorists who claim that actor choices are guided by logics of appropriateness derived from shared definitions of social roles and proper conduct by role-holders, rather than by utilitarian calculation of material interests. In this view, institutions serve primarily to define legitimate and illegitimate courses of conduct, facilitate socialization of actors into their roles, facilitate actor engagement in learning about particular situations, and promote joint deliberation about the content of the shared social beliefs and norms.

Social practices analysts expect greater cooperation on environmental problems when governments (and other actors) adopt an ecological worldview, accept protection of the biosphere as a distinct policy goal, redefine government responsibility and sound business practice to include fostering ecologically sustainable forms of human activity, and include effects on the biosphere among the criteria for judging the merits of alternate economic and social policy measures. For social practices theorists successful international environmental cooperation is not a matter of altering incentives while leaving interests unchanged, but of altering beliefs so definitions of interest will be changed. In their view, institutions are important primarily as mechanisms for facilitating redefinition of social roles and norms, socialization into and internalization of new norms, and continued actor learning.

This orientation has led social practices theorists to focus on the conditions under which learning occurs and new beliefs become disseminated from small groups of initial advocates to wider segments of society (e.g., Clark et al. 2001). These analyses highlight the importance of visible negative effects in triggering reconsideration, of scientific consensus on the extent and material causes of environmental degradation in facilitating acceptance of new information, and of activity by transnational sets of policy entrepreneurs and social movements in disseminating new worldviews, beliefs and information more widely.

Though rational choice and social practices approaches are often treated as irreconcilable theoretical rivals, they are actually complementary because each illuminates particular aspects of human behavior. Social practices analysis is strongest on the importance of causal and normative beliefs in guiding actor choices, and in beginning to identify the mechanisms of deliberation and learning through which those beliefs are maintained or changed. Rational choice analysis is strongest in indicating how material calculations factor into actor choice within the areas of discretion permitted by all but the most determinate social norms, and how different patterns of individual actor positions affect the likelihood of different compromises and the stability of various coalitions that might arise. The complementarity of the two approaches has been recognized implicitly in the stream of policy analysis inquiring into how each of the mechanisms of persuasion, alteration of information, shift in the availability of opportunities to engage in particular behaviors, reward and punishment can be used to bring about desired behavioral change. It is also acknowledged by the growing attention paid among both rational choice and social practices theorists to the ways in which international institutions are affected by and affect domestic realignments that create new constituencies for environmental policies or alter the balance between existing constituencies. Though many social scientists will continue to prefer one approach over the other, a growing number will explore integrating the two more explicitly in more careful empirical tracings of actual processes of establishing, implementing and revising international environmental accords.

See also: Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Change and State Response; Environmental Economics; Environmental Planning; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Vulnerability; International Justice: Philosophical Aspects; Nature, Concepts of: Environmental and Ecological; Precautionary Principle; Property Rights; Resource Institutions; Sustainable Development

Bibliography

Caldwell L K 1996 *International Environmental Policy*, 3rd edn. Duke University Press, Durham, NC

Cameron J, Werksman J, Roderick P (eds.) 1996 *Improving Compliance with International Environmental Law*. Earthscan, London

Chayes A, Chayes A H 1995 *The New Sovereignty*. Harvard University Press, Cambridge, MA

Clark W C, Van Eijndoven J, Jaeger J (eds.) 2001 *Learning to Manage Global Environmental Risks*. MIT Press, Cambridge, MA

Downs G W, Rocke D M, Barsoom P N 1996 Is the good news about compliance good news about cooperation? *International Organization* **50**: 379–407

Dupuy P-M 1992 Soft law and the international law of the environment. *Michigan Journal of International Law* **12**: 420–35

Earthscan and Fridtjof Nansen Institute 1998/99 Annual. *Yearbook of International Cooperation on Environment and Development*. Earthscan, London

Grubb M, Vrolijk C, Brack D 1999 *The Kyoto Act: A Guide and Assessment*. Royal Institute of International Affairs, London

Haas P M, Keohane R O, Levy M A (eds.) 1993 *Institutions for the Earth: Sources of Effective International Environmental Protection*. MIT Press, Cambridge, MA

Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8
Kiss A 1993 Les traites-cadres: une technique juridique characteristique du droit international de l'environnement. *Annuaire Francaise de Droit International* **39**: 792–7
Miles E L Jr (ed.) 1999 *Explaining Regime Effectiveness: Confronting Theory with Evidence*. MIT Press, Cambridge, MA
Ruster B, Simma B, Bock M (eds.) 1983 *International Protection of the Environment: Treaties and Related Documents*. Oceana, Dobbs Ferry, NY
Sand P H 1996 Institution-building to assist compliance with international environmental law. *Zeitschrift für Auslandisches Öffentliches Recht und Völkerrecht* **56**(3): 773–93
Sandler T 1997 *Global Challenges: An Approach to Environmental, Political, and Economic Problems*. Cambridge University Press, Cambridge, UK
Schachter O 1991 The emergence of international environmental law. *Journal of International Affairs* **44**: 457–93
Symes D (ed.) 1998 *Property Rights and Regulatory Systems in Fisheries*. Fishing News Books, Oxford, UK
UNEP 1990 *International Register of Environmental Treaties*. United Nations Environment Programme, Nairobi, Kenya
Victor D G, Raustiala K, Skolnikoff E B (eds.) 1998 *The Implementation and Effectiveness of International Environmental Commitments*. MIT Press, Cambridge, MA
Wettestad J 1999 *Designing Effective Environmental Regimes*. Edward Elgar, Cheltenham, UK
Young O R (ed.) 1999 *The Effectiveness of International Environmental Regimes: Causal Connections and Behavioral Mechanisms*. MIT Press, Cambridge, MA

M. J. Peterson

International Arbitration

International arbitration can refer to practices involving either commercial disputes or disputes between states. The practice for states, exemplified by disputes about borders, are ad hoc and infrequently invoked. Of much more practical relevance at the beginning of the twenty-first century is the system of private justice created to resolve international business disputes, which has become the standard system for international transactions. When two businesses with headquarters in different countries enter into a contractual arrangement, the contract will generally provide that any disputes relating to the contract will be resolved through private arbitration. It is the default arrangement. International commercial arbitration involves hundreds of international commercial arbitration centers competing to have the most prominent rosters of arbitrators and the largest number of high stakes international cases. However, the number of international commercial arbitrations is impossible to determine precisely since many of the arbitrations are not administered by any formal organization, and there is no clear line between domestic and international. International commercial arbitration may be institutional, under the auspices of one of the institutions competing for the business of international commercial arbitration, or it may be *ad hoc*, handled solely by the parties and their counsel. It is especially difficult to count the *ad hoc* arbitrations.

It is clear, however, that the numbers of international commercial arbitrations increased at the end of the twentieth century, particularly in China and Latin America. The most important of the institutions is the International Chamber of Commerce in Paris, France, and it generally has the largest caseload and the cases involving the highest stakes. In 1999, some 529 requests for arbitration were filed with the ICC, concerning 1,354 parties from 107 different countries. The places of arbitration assigned to the cases involved 48 different countries on five continents. Arbitrators of 57 different nationalities were appointed or confirmed under the ICC Rules, and the amount in controversy exceeded one million dollars in 58 percent of the new cases. The ICC Court, which confirms the awards, had 269 new awards submitted in that year.

International commercial arbitration—as arbitration generally—involves a process whereby a third party issues an opinion that is formally binding on the parties. Unlike domestic arbitration as it has traditionally been practiced in the USA, international commercial arbitrators typically write formal opinions. These formal opinions are typically not published nor even made public informally. They belong to the parties, who must therefore consent to any publication. The awards made by the arbitrators are binding and very difficult to challenge or appeal. The New York Convention of 1958, which has been adopted by well over 100 states, make an arbitral award in practice more easily enforced within a contracting state than would be a judgment obtained through litigation. In addition the procedures of international commercial arbitration have been converging around a set of rules that tend to detach arbitration from its site and to emphasize party autonomy. The UNCITRAL (United Nations Commission on International Trade Law) model law has been the major model for this convergence.

The most prominent feature of international arbitration is that the arbitrators are private individuals without any formal connection to the state or a state judicial system. In theory almost anyone could be selected as an arbitrator. In practice, however, there is a relatively small group that tends to dominate selection for major arbitrations. The group consists largely of European lawyers, professors, and retired judges; but there are also prominent arbitrators from other parts of the world.

The advantages of arbitration in international disputes are usually listed as follows: the ability to select elite private judges, the secrecy of the proceedings, and the ability to avoid having to submit to the judicial system of the opposing party. Not generally listed as advantages, are the cost and the duration of

the proceedings, since in fact arbitrations may cost as much or more, and take as much or more time, than litigation in a national court system.

The leading centers in addition to the International Chamber of Commerce in Paris include the London Court of International Commercial Arbitration (LCIA), the American Arbitration Association (AAA), the Cairo Regional Center for International Commercial Arbitration, the Stockholm Chamber of Commerce, and the Chinese International Economic and Trade Arbitration Commission (CIETAC). While there are many centers, most contracts that contain arbitration clauses will typically select one of the established centers. The ICC is especially useful in this regard for relatively new participants to international commercial arbitration. The ICC, while more expensive, helps to guide new parties and new arbitrators through the processes, and it also has a reputation as an institution that can ensure that its awards are respected in domestic courts.

There are a number of interesting socio-legal issues that relate to international commercial arbitration. They include: (a) how this system of private justice came to be considered the legitimate way to resolve international business disputes; (b) what this system means for issues concerned with alternative dispute resolution more generally; and (c) how people become international commercial arbitrators (Dezalay and Garth 1996).

1. *Building the International Arbitration System*

International business was long a world of complex personal relations rather than formal laws and dispute resolution processes. The number of actors was relatively small in any given field, and the actors could get to know and take account of each other. There might be contracts, but the contracts were relatively unimportant documents either in the negotiation or when disputes might arise. Histories of the petroleum industry, in particular, reveal the key actors behind the 'seven sister' oil companies and how they interacted with the countries in which they did business. Lawyers were very much part of this history, but lawyers for the oil companies—including John McCloy, the famous lawyer for the Rockefeller family who was termed the 'chairman of the establishment'—relied much more on personal influence and connections than on their expertise as lawyers.

When this relatively stable set of relationships was shaken up, especially by new entrants in the 1960s and 1970s willing to offer more favorable terms to produce the oil, the lawyers for the oil companies used contract and international law to try to hold on to their privileged positions. At the same time, a small number of elite actors in the exporting countries had taken advantage of oil company paternalism to obtain education abroad—in law and engineering, for example. They and a few maverick legal advisors began to formulate legal arguments that slowly built the legal infrastructure of the industry. The major transformation, however, came when the processes of readjustment of the relationships resulted in nationalization of the oil industries in most of the oil exporting countries. The oil concession agreements typically had arbitration clauses through an analogy to the arbitration that had long been practiced in disputes between nations. When the oil was nationalized in places such as Kuwait, Libya, and Saudi Arabia, the companies invoked the arbitration clauses and began arbitration processes.

Legal accounts of this period typically celebrate the grand arbitrations that resulted, suggesting that important legal principles supporting the stability of contracts trumped the political acts of nationalization. Other accounts, including those by journalists and historians, however, do not mention the arbitrations at all. They instead emphasize the negotiations that took place between those who had long invested in the relationships, the negotiations between governmental actors at another level, and the gunboats that were putting not so subtle pressure on behalf of the oil companies. Lawyers involved at the time suggest that their activity in the arbitrations was not highly valued or even much noticed by company executives.

The arbitration processes did go forward, however, and they produced raw material that equipped international business with people and principles to handle trans-national disputes. The raw material was in the form of written opinions that were widely circulated, academic articles about the principles in the opinions, and individuals who gained or enhanced their reputations as potential arbitrators by virtue of their participation in the arbitrations. The petroleum arbitrations thus contributed to the production and legitimization of international commercial arbitration for trans-national business disputes.

When new sets of problems arose that related to some of the themes of the petroleum arbitrations, in particular disputes involving the numerous construction projects that took place in oil-producing countries after the price of oil increased in the early 1970s, international commercial arbitration played a crucial role. The continental academics who developed these principles into a new *lex mercatoria* also developed relationships with students from the less developed world, and the students helped to reinforce the legitimacy of these principles in their own countries. Put very simply, the process produced lawyers on both sides, legal principles, and a legitimacy that made it possible to generate a demand for what this group supplied. For the International Chamber of Commerce, this increased demand meant that the while the first 3,000 requests for arbitration came between 1923 and 1976, the next 3,000 came in the following 11 years.

2. International Commercial Arbitration and Alternative Dispute Resolution

The world of international commercial arbitration was dominated in the 1960s and 1970s by continental academics who wrote about the *lex mercatoria*, but also tended to encourage the parties to find a way to settle the dispute. The *lex mercatoria* could be used by the arbitrators to find a solution that fit the business relationship and left the parties satisfied with the result. Consistent with the general orientation of the leading arbitrators and lawyers associated with the pioneers of international commercial arbitration, the process was not very adversarial, was relatively inexpensive except for the arbitrators, did not tend to produce extensive documentary evidence, and did not contemplate such US practices as cross-examination of witnesses. The model was quite consistent with Continental practices, even for others who participated in the processes.

When the arbitration business began to take off in the 1970s and 1980s, it coincided with the rise to prominence of business litigation in the USA. The US law firms in Paris, France, were long involved in international commercial arbitration, but they tended to play according to the Continental rules. As the practice of international commercial arbitration grew and expanded outside of a relatively small circle of people, US businesses became increasingly involved in arbitration, and developing countries found it to be in their interests to hire US law firms as well. The enhanced presence of US litigators had important consequences. They naturally sought to use the techniques of discovery, intensive production of documents, cross-examination of witnesses, and more generally the kind of aggressive adversarial behavior that was beginning to characterize litigation in the USA. They also sought to retain arbitrators who would permit the lawyers to use the techniques that they had developed. The result was that the processes of international commercial arbitration changed substantially. They did not become identical with US litigation, but they moved much closer to US adversarialism. International commercial arbitration became much more document intensive, much more adversarial, and much more expensive.

Litigation in the USA had also been transformed, and that transformation had produced the beginnings of the alternative dispute resolution movement in that country. New organizations such as the Center for Public Resources had begun to encourage corporations to practice more mediation as a way to avoid the high costs and adversarial nature of business litigation. The movement became quite influential in the USA and some of the individuals associated with it began to suggest that it would also be useful for international commercial arbitration. Not surprisingly, the first reaction of many within the international arbitration community was that they already practiced mediation as part of the normal process. They argued that the practice of international arbitration provided an effective alternative to litigation as practiced in the USA.

As international commercial arbitration became more like US litigation, however, the US-promoted alternative began to gain more adherents within the international arbitration community. By the end of the twentieth century, the menu of US alternatives—including mediation but a range of others as well—had become much more common. This gradual transformation illustrates the way that the formal categories of dispute resolution—litigation, arbitration, and even mediation—can disguise profound transformations in the way that the underlying practices proceed. The center of gravity of international commercial arbitration moved much closer to US-style litigation, which was itself a relatively new invention, and the shift also helped to bring the US-antidote, alternative dispute resolution.

3. Becoming an Arbitrator

The international commercial arbitration community is relatively closed, often termed a 'mafia' even by insiders. One reason for this 'club' is simply that for high stakes arbitration, the parties typically want to select arbitrators who are known quantities. Untested arbitrators are very risky from the point of view of the parties. It is also deemed important to select arbitrators who are known for their ability to 'speak the language' and be taken seriously within an arbitral tribunal. Since arbitrators often serve as counsel because of their familiarity with the processes and people, it is common for people from the relatively closed arbitral community to nominate others from the community. The key to success in the arbitration proceeding is selecting an arbitrator who will be impartial enough to persuade the chair of the tribunal but will also be able to understand and put forward the case of the party that appointed him or her. The selection of established names also serves a protective function in case the decision does not favor the selecting party. It is easier to avoid blame for a safe selection than for a novel one. For many reasons, including the fact that a relatively small supply serves the economic interests of the arbitrators, it is not easy to gain entrance into this community.

The easiest way to gain entrance for at least some small cases that may provide a basis to build a practice is to conform to the characteristics of those already in the field. The individuals who set the norms emphasized scholarly accomplishments, which put an emphasis on publications, cosmopolitanism, and linguistic ability. Those who seek to enter must probably also attend a number of the major conferences in order to display the appropriate characteristics and be judged more personally. However, the general point is

that invitations to enter the community tend to be issued most easily to those who can offer something to the existing community—most obviously persons of high status who reflect that status back on the arbitral community, persons from areas that are not yet committed to the legitimacy of international commercial arbitration (which means they offer both credibility and a potential new caseload), or sometimes disciples of those already in the arbitral community. This process helps to explain why individuals without high status may study, write, and attend conferences, but lacking a prestigious platform, they cannot gain entry. On the other hand, individuals from developing countries who serve as judges on the World Court do typically gain entry. The process also helps explain why there are so many arbitration centers despite the fact that the caseloads are concentrated in a relatively few places. Those who create new centers gain the attention of the elite of the arbitral community and also help to provide new domains where businesses will contemplate arbitration and courts will respect the awards. As a result, the new center may promote the local leader into a position closer to the core of the arbitral community. There tends to be relatively few of these peripheral arbitrators in each relevant site.

See also: Intergenerational Justice; International Business; International Justice: Philosophical Aspects; International Law and Treaties; International Relations: Theories; International Trade: Commercial Policy and Trade Negotiations; Laws, Conflict of; Lex Mercatoria; Public Goods: International; World Trade Organization

Bibliography

Craig W, Park W W, Paulsson J 2000 *International Chamber of Commerical Arbitration*. Oceana Publications, Dobbs Ferry, NY

Dezalay Y, Garth B 1996 *Dealing in Virtue International Commercial Arbitration and the Construction of a Transnational Legal Order*. University of Chicago Press, Chicago

Redfern A, Hunter M 1999 *International Commercial Arbitration*, 3rd edn. Thomson Professional Publishing, London

B. Garth

International Business

The term 'international business' refers to economic activities that take place across the boundaries of nation–states. The most important of these are trade and foreign investment. Merchandise trade has grown 12-fold since the end of World War II, and trade in services has increased even faster since the early 1980s. Foreign direct investment (as opposed to portfolio investment not seeking the managerial control of the foreign asset) has grown at a rate twice as fast as that of trade. Foreign direct investment gives rise to multinational enterprises, i.e., firms with operations in more than one country (see *Multinational Corporations*). There are some 54,000 of these enterprises in the world today which collectively control about 450,000 affiliates. They account for one third of world merchandise trade and for the vast majority of trade in technology. While theories of international trade were first put forward during the sixteenth century and largely have remained the province of economics, the study of foreign direct investment and the multinational enterprise started only after World War II and has been approached from a variety of social and behavioral perspectives. The three most important ones—economic, managerial, and sociopolitical—are covered here (see Table 1).

1. The Economic Approach

Modern economic thinking about foreign direct investment and the multinational enterprise properly starts with Hymer's (1976) thesis, posthumously published in 1976. The key insight was to see foreign direct investment not as a capital movement but as the decision of a firm to produce some good or service in a foreign location. Hymer noted that foreign direct investors did not seem to go abroad in search for higher interest rates—as the theory of investment would predict. Rather, he argued that 'control of the foreign enterprise is desired in order to remove competition between that foreign enterprise and enterprises in other countries' (Hymer 1976). Thus, Hymer was the first to observe the monopolistic behavior of the firm that engages in foreign direct investment. In other words, the multinational was seen as an anomaly that could only be explained by reference to imperfections in the markets for production factors and goods, economies of scale, and/or government intervention. The most complete rendering of the oligopolistic thesis is that of Knickerbocker (1973). He observed that firms in a loose-knit oligopoly in the home country match each other's investments in foreign locations following a pattern of action–reaction or move–countermove out of fear that not keeping up with rivals might erode their profitability and competitive position or endanger their sources of supply.

Hymer was also the first scholar to note that 'the control (of a foreign enterprise) is desired in order to appropriate fully the returns on certain skills and abilities' (1976), especially those having to do with product and process innovation. Vernon (1979) further argued that patterns of trade and foreign investment had to do with the so-called 'product cycle' and not so much with comparative advantage. Vernon saw entrepreneurs and managers as relatively myopic, i.e., most knowledgeable about, and responsive to, con-

Table 1
Three perspectives on international business and the multinational enterprise (MNE)

	Economic	Managerial	Sociopolitical
Name of the literature(s)	IB (international business)	IM (international management)	Development studies and international relations
Research community	AIB (Academy of International Business); AEA (American Economic Association).	AIB and AOM (Academy of Management)	ASA (American Sociological Association) and APSA (American Political Science Association)
Research questions	Why do firms go abroad? How do they go abroad? What are the welfare implications?	Why and how do firms go abroad? How are MNEs organized?	What role do MNEs play in economic and political development and in international relations?
Assumptions and answers	Rationality; constrained maximization; calculation of trade-offs	Capability-based strategies; organizational structure follows strategy	Power and interest drive processes and outcomes
Ways of doing research	Deductive reasoning; formal modeling; large samples	Conceptual framing; grounded theory; case studies and large samples	Mostly inductive; grounded theory; case studies; comparative–historical method
Key scholars	Hymer (1960); Caves (1971); Knickerbrocker (1973); Vernon (1966); Buckley and Casson (1976); Dunning (1979); Kogut (1983)	Perlmutter (1969); Stopford and Wells (1972); Wells (1983); Prahalad and Doz (1987)	Vernon (1971); Moran (1974); Kobrin (1982); Evans (1979); Haggard (1990)

ditions in the home market. The features of new products reflected characteristics of the home market and affected the pattern of foreign expansion of the firm. Vernon proposed that firms would only pursue foreign market opportunities after the home market became saturated. Foreign markets relatively similar to the home market in terms of purchasing power and tastes would be entered first, and exports would probably be the preferred mode of entry. Licensing or foreign production would come later in response to host country protectionism and/or rising production costs in the home country.

The economic approach to the multinational enterprise acquired a high degree of internal consistency after scholars started to draw from the insights of transaction cost economics, which has become the backbone of the widely-accepted view among economists that multinationals are an efficient governance response to certain market failures (Buckley and Casson 1976; Dunning 1979; Caves 1996). The multinational enterprise is expected to exist if two conditions are met. First, the necessary condition is that there be differences in the productivity and cost of production factors by location (in the case of vertical expansion by multinationals) or significant transportation costs, trade protectionism, adverse changes in exchange rates, and/or needs to customize products to local demand (in the case of horizontal expansion). Transaction cost theory, however, warns that a second, sufficient condition must be met in order to guarantee that foreign production will be controlled managerially by a multinational firm: market failure due to asset specificity or uncertainty (vertical expansion) or the possession of intangible assets such as patents, brands, and other proprietary skills that cannot be protected in a contractual, arm's length fashion (horizontal expansion). A final, distinctive contribution to the economic approach was made by Kogut (1983), who proposed to see the multinational enterprise as a flexible network in which resources are arbitraged across borders, a view that revisited Vernon's insight that multinationals approach foreign investment decisions as a sequential process. A significant body of empirical evidence exists corroborating the major tenets of the economic approach (Caves 1996).

2. *The Managerial Approach*

While the economic approach has focused on the questions of why and how firms go abroad, the managerial approach has contributed an understanding of how multinationals are organized and managed (see *Multinational Corporations*). This second approach relies on typological conceptual frameworks and inductive reasoning rather than formal modeling. In addition, theoretical development frequently results from case studies, although testing of the main propositions has also been undertaken with large

samples, as in the economic approach. Perhaps the first noteworthy typology of the multinational enterprise in the managerial literature was proposed by Perlmutter (1969), who distinguished among three types of multinational 'mindsets.' Ethnocentric multinationals are the ones assumed by Vernon's (1979) produce-cycle approach. They have simple but inflexible organizational structures. Polycentric multinationals, by contrast, adapt to local circumstances but have difficulty co-ordinating. Finally, geocentric multinationals are costly but very good at gathering and integrating information from various locations. Drawing on a rich data set on US multinationals assembled under the co-ordination of Vernon, Stopford, and Wells (1972) presented the first comprehensive treatment of the various factors at the firm, industry, and country levels affecting the choice of organizational structure and degree of ownership over the foreign subsidiaries of the multinational firm.

The early insights by Perlmutter and by Stopford and Wells were further developed in their respective dissertations by Prahalad and Doz as the 'integration-responsiveness' framework (Prahalad and Doz 1987). Multinationals are seen as organizations responding to environmental pressures for global integration (so as to reduce costs through economies of scale and scope) and to pressures for local responsiveness (so as to adapt to local tastes, institutions, or regulations). Following the strategy–structure paradigm in organizational theory, they identify different organizational structures and managerial processes that allow multinationals to operate under different configurations of global integration and local responsiveness. Although still in its infancy, a neoinstitutional approach to the multinational enterprise as a firm exposed to isomorphic pulls from both the home and the host country is also emphasizing the organizational problems associated with multinationality.

3. *The Sociopolitical Approach*

A third perspective on international business has built on the insights of both the economic and managerial approaches to the study of the political and social context of multinational activity. Vernon (1971) was the first to explicitly argue that the spread of multinational corporations creates 'destructive political tensions,' and that there is a 'need to re-establish balance' between political and economic institutions.

Political scientists have looked at the impact of multinationals on international relations. They argue that changes in the international location of activities due to foreign direct investment have reinforced the importance of domestic policies, as countries engaged in regionalization, sectoral protectionism, and mercantilistic competition during the 1980s so as to cope with globalizing trends. They have also analyzed the relative bargaining positions of host governments and multinationals. In a classic book, Moran (1974) analyzes Chilean efforts to seize control over its vast copper industry, finding that nominal ownership and control does not guarantee success in a worldwide oligopoly. Sociologist Evans (1979) published his landmark book *Dependent Development*. He argues that foreign investment attracted by import–substitution industrialization policies results in a 'triple alliance' among local entrepreneurs, state bureaucrats, and foreign capital that perpetuates and even enhances patterns of dependence of the host country on foreign technology and capital. Haggard (1990) summarizes the bargaining power literature writing that the host government is relatively powerful vis-à-vis multinationals making huge fixed investments (e.g., in extractive industries), relatively weak when dealing with export-oriented firms looking for cheap labor, and weakest when multinationals are attracted by import-substitution policies that create a political base of support among local suppliers, distributors, employees, and joint-venture partners. Kobrin (1982) offers a different, yet complementary, approach to the politics of multinational activity by examining how multinationals evaluate political risk and try to protect themselves against it.

Sociologists have also examined the effects of international business activity on economic growth and inequality in the world. One major controversy in the literature is whether foreign investment retards long-term economic growth and exacerbates income inequality in the host country or not. Capital dependency theorists have long argued in the affirmative using cross-national empirical evidence, while other sociologists have pointed out that the data do not support such claims (see Kentor 1998). This protracted debate—now three decades old—does not take into account the fact that the effect of foreign investment on host-country growth and income inequality is mediated by the economic development strategy of the host government. Countries allow more or less foreign investment depending on a number of ideological, political, and economic considerations, and the effects of multinational activity differ widely between import-substitution and export-oriented contexts (Haggard 1990). While it is easy to point to examples of the harmful effects of multinationals, and even of their outright wrong-doing, the social science literature has been slow to recognize that they have played a largely constructive role in many countries, e.g., Singapore, Malaysia, Puerto Rico, Ireland, Portugal, Spain, in which even organized labor and the left have welcomed them.

Since the 1980s the expansion of international business activity has acquired a new, qualitatively different outlook. Globalization in the late twentieth century is different from the internationalization of the world economy during the nineteenth century through increasing cross-border flows of trade and investment. The 'global' economy is driven by the increasing scale

of technology, the surge in cross-border collaboration between firms along the value-added chain, and the cross-border integration of information flows. While the conventional wisdom is that globalization undermines the nation–state and compels firms to converge towards 'universal best practices,' sociologists and political scientists document that governments exercise a great deal of choice in the global economy and firms continue to adopt organizational structures and patterns of behavior that are related to the institutional conditions in their home countries (Orrù et al. 1997). Thus, globalization is far from being a monolithic, uniform, and inexorable trend. Rather, it is a fragmented, contradictory, and heterogeneous process of change (Guillén 2001). More empirical and theoretical research is needed to describe the causes of globalization—including those having to do with international business activity—and to understand its effects on political, economic, and social outcomes in various parts of the world.

See also: Alliances and Joint Ventures: Organizational; Capitalism: Global; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; International Law and Treaties; International Marketing; International Organization; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; International Trade: Geographic Aspects; Market Areas; Marketing Strategies; Markets and the Law; Monetary Policy; Rational Choice and Organization Theory; Venture Capital; World Trade Organization

Bibliography

Buckley P, Casson M 1976 *The Theory of the Multinational Enterprise*. MacMillan, London

Caves R E 1996 *Multinational Enterprise and Economic Analysis*, 2nd edn. Cambridge University Press, New York

Dunning J 1979 Explaining changing patterns of international production: In defence of the eclectic theory. *Oxford Bulletin of Economics and Statistics* **41**: 269–95

Evans P 1979 *Dependent Development*. Princeton University Press, Princeton, NJ

Guillén M F 2001 Is globalization civilizing, destructive or feeble? A critique of five key debates in the social-science literature. *Annual Review of Sociology* **27**: 235–60

Haggard S 1990 *Pathways from the Periphery: The Politics of Growth in the Newly Industrializing Countries*. Cornell University Press, Ithaca, NY

Hymer S 1976 *The International Operations of National Firms: A Study of Direct Foreign Investment*. MIT Press, Cambridge, MA

Kentor J 1998 The long-term effects of foreign investment dependence on economic growth, 1940–1990. *American Journal of Sociology* **103**: 1024–46

Knickerbocker F T 1973 *Oligopolistic Reaction and Multinational Enterprise*. Harvard Business School, Boston

Kobrin S J 1982 *Managing Political Risk Assessment: Strategic Response to Environmental Change*. University of California Press, Berkeley, CA

Kogut B 1983 Foreign direct investment as a sequential process. In: Kindleberger C P, Audretsch D (eds.) *The Multinational Corporation in the 1980s*. MIT Press, Cambridge, MA

Martinelli A 1982 The political and social impact of transnational corporations. In: Maklen H, Martinelli A, Sinelser N J (eds.) *The New International Economy*. Sage, London, pp. 79–116

Moran T H 1974 *Multinational Corporations and the Politics of Dependence: Copper in Chile*. Princeton University Press, Princeton, NJ

Orrù M, Biggart N W, Hamilton G G 1997 *The Economic Organization of East Asian Capitalism*. Sage, Thousand Oaks, CA

Perlmutter H 1969 The tortuous evolution of the multinational corporation. *Columbia Journal of World Business* **Jan–Feb**: 9–18

Prahalad C K, Doz Y L 1987 *The Multinational Mission: Balancing Local Demands and Global Vision*. Free Press, New York

Stopford J M, Wells L T Jr 1972 *Managing the Multinational Enterprise: Organization of the Firm and Ownership of the Subsidiaries*. Basic Book, New York

Vernon R 1971 *Sovereignty at Bay: The Multinational Spread of US Enterprises*. Basic Books, New York

Vernon R 1979 The product cycle hypothesis in a new international environment. *Oxford Bulletin of Economics and Statistics* **41**(4) (November): 255–67

M. F. Guillén

International Communication: History

International communications includes communications between the inhabitants and between the governments of different states. Such communications have increased in speed, reliability, and volume since the beginning of civilization, until they now constitute a flood.

1. Ancient and Medieval Communications

Messengers were the first means of conveying information over long distances. Before the Battle of Marathon in 490 BC, Phidippides, 'by profession and practice a trained runner,' ran to Sparta, 120 miles away, arriving 'on the very next day.'

Inhabitants of cities and isolated communities had no choice but to rely on messengers and travelers, but kings needed more regular communications. To keep control of their far-flung domains, the Persians established a permanent postal system with relays of horses along the Royal Road that stretched from Susa in western Iran to Ephesus on the Aegean Sea. The Romans created a similar network of imperial messengers and relay posts along their major roads, called *cursus publicus*. The Chinese, Inca, Mongol, and other empires established similar systems. These early imperial networks were reserved for official messages; anyone else who had a message to send would have to

find a traveler going in the right direction, and hope for the best.

The Middle Ages saw the decline or disappearance of the governmental apparatus of classical times and the emergence of numerous unofficial channels of communication, some of them well organized. The Benedictine monks of Cluny entrusted pilgrims with messages to other monasteries from Spain to Poland. The merchants of the Hanseatic League carried documents and messages along with merchandise. Butchers, who traveled from place to place buying and slaughtering livestock, carried messages for people. The most organized of these nongovernmental networks was the one centered on the University of Paris, whose students came from all over Europe and carried and sent messages between Paris and their homelands. Although they were supposed to serve only the university community, in fact they commonly carried messages for the public as well. These networks were slow and unreliable, however, for they did not operate on a schedule, nor did they use relays of horses or messengers.

2. *Postal Communications*

The first postal system that established permanent relays and offered its services to the general public was founded in 1489 by Franz von Taxis, postmaster to the Holy Roman Emperor Maximilian I. The Thurn and Taxis family provided a regular postal service throughout most of Europe until 1867. With up to 20,000 messengers and relays of horses on all the main routes, their postal service was reliable and, by the standards of the day, fairly swift: from Brussels to Paris in 44 hours, to Innsbruck in six days, and to Toledo in 12 days. Service, however, was very costly.

The Habsburgs may have trusted the Thurn and Taxis family, but the rising national monarchies of France and England did not, nor did they trust any other private or nongovernmental message service. The first royal messenger service in France, established in 1464, was reserved for official messages. Private individuals could bribe the royal couriers to carry messages surreptitiously. The royal postal service was opened to the public in the early seventeenth century. It set the rates for private letters very high, to subsidize the official mail. Once the royal mail service was opened to the public, the government attempted to suppress private messengers, in order to protect its lucrative monopoly and permit easier surveillance of potentially subversive communications; the last nongovernmental service in France, that of the University of Paris, ended in 1719.

The evolution of the English postal system paralleled that of France. There had long been royal messengers, but the first permanent service was founded in 1512. Royal messages to and from foreign countries, however, were still hand-delivered by special messengers. Meanwhile, merchants had their own networks, such as the Merchant Adventurers' Post for English merchants engaged in foreign trade, and the Foreigners' Post for foreign merchants in England. Queen Elizabeth, suspicious of Catholic plots, ordered all mail leaving or entering the country to be carried by the royal post. This decree, though often repeated by her successors, was never fully enforced.

In 1635, the royal postal service began to serve the public; two years later it included the foreign mails as well. As in France, 'public service' was a euphemism, for postal rates were kept extremely high as a source of income for the treasury and for the king's favorites and officials had the right to open and read private letters. Surreptitious inspection of letters, especially those from abroad, was a common occurrence in the eighteenth and early nineteenth centuries. Most people who needed to correspond across borders preferred to send letters with a friend, a servant, a hired messenger, or a traveler. Surveillance abated in the mid-nineteenth century, as the increased volume of mail overloaded the postal bureaucracies.

The mid-nineteenth century saw a tremendous surge in the volume of mail, thanks to two innovations. One was the penny post with prepaid stamps, introduced in England by Rowland Hill in 1840 and soon imitated in other countries. The other was railroads and steamships, which made transportation cheap and fast. Letters from relatives or friends in foreign countries, a rarity until then, became a commonplace experience for middle-class families in the Western world.

Airplanes began transporting mail right after World War I. In parts of the world, such as Africa and South America, that had poor telegraph and rail connections, airmail represented a breakthrough comparable to the effect of railways in Europe a century earlier. In places well served by traditional postal services, airmail competed with the cheapest kind of telegram, the 'night letter.' After World War II, airmail became the standard means of sending first-class postage over long distances.

3. *The Telegraph*

Many scholars believe that modern telecommunications began with the electric telegraph. From a technical perspective, the advent of electricity was certainly revolutionary. From an information point of view, however, the revolution began half a century earlier with the optical telegraph, invented by Claude Chappe in France in 1794.

Throughout the French Revolution and after, every French government expanded the optical telegraph network and supported the family that ran it. Chappe and his brothers built lines east to Strasbourg, west to Brest, and north to Dunkirk and Brussels; under

Napoleon, they extended the network south to Lyon and Marseilles. As Napoleon's empire grew, lines radiated to Turin, Milan, and Venice, to Antwerp and Amsterdan, and to Mainz. After the defeat of Napoleon the French network was reduced to its domestic lines. Other European countries, as well as the USA, Egypt, and Australia, built individual lines linking their major cities to the nearest harbors. Some of these were government-owned lines reserved for official or military traffic, while others served the merchant communities. In the few instances where two systems met at a border, any message going from one country to the other had to be decoded, translated, hand-carried across the gap, and encoded for further transmission. International telecommunication had to await the electric telegraph.

During the 1830s, several inventors came up with methods of transmitting information over long distances by electricity. The one that eventually displaced all others was invented by Samuel Morse. Morse's contribution was not the use of electricity but a code which allowed the use of only one wire, with the earth serving as a return. The cost savings of Morse's system made up for the need for trained operators.

During the next two decades, the nations of Europe and North America began installing electric telegraph lines. At first, they used incompatible systems; messages had to be written out, translated, hand-carried across the border, then retransmitted. As traffic grew, governments strove to overcome this barrier to communication. In 1850, Austria and several German states signed the Austro-German Telegraph Union. Treaties between France and its neighbors followed. Finally in 1865, all European governments except the UK founded the International Telegraph Union, to be administered by the International Bureau of Telegraph Administrations, the first permanent international organization. It encouraged national telegraph administrations to adopt compatible technical standards and to connect their lines across border. By the 1860s, it was possible to send telegrams across Europe and between the USA and Canada.

Meanwhile, engineers had solved the technical problems of communicating across the sea. The first successful submarine cable crossed the Channel in 1851. Cables across oceans were a greater challenge, however. The first Atlantic cable failed in 1858, as did the first cable from Egypt to India two years later. After many trials and errors, cable technology finally caught up with the demand. The Atlantic cable of 1866 proved reliable. In the 1870s, British companies rushed to lay cables from the UK to India and from there to China and Australia. Cables reached South America and Africa in the 1880s. By World War I, cables criss-crossed all the oceans and connected every continent and every country.

Submarine cables were phenomenally expensive, however. Only the North Atlantic had enough traffic to justify several competing cables. In other parts of the world, the high cost of cables had to be amortized over small numbers of customers, and a few firms, most of them branches of the British conglomerate Eastern and Associated, dominated the market. The high cost of sending messages over long distances led business customers and news agencies to develop elaborate code books.

4. *Wireless Telegraphy*

Although the telegraph had 'annihilated time and space' (to use a favorite expression of nineteenth-century boosters), it left many people dissatisfied, for both economic and political reasons. Economically, international telegrams were very expensive, and transoceanic ones exorbitantly so, placing them out of reach of all but the wealthiest clients and commercial enterprises. And politically, the fact that most ocean cables belonged to British firms, and that most intercontinental telegrams passed through British hands, made all other governments very uncomfortable, especially after 1900, when the UK proved that it was not averse to intercepting, delaying, and censoring foreign messages during a crisis like the Boer War (1900–2).

In 1895, Guglielmo Marconi demonstrated the possibility of transmitting information through the ether by electromagnetic radiation. He found immediate backing from the Royal Navy and from British shipping lines, which had long wanted to communicate with their ships at sea. France, Germany, and the USA also greeted the new technology with enthusiasm, for they had fallen far behind the UK in the cable business, and saw radio as a means of building international connections independent of the UK.

Until the late 1920s, wireless used longwaves (over 1 km, or less than 300 kHz), which required power in proportion to the distance to be covered; stations powerful enough to transmit across continents or oceans needed dozens of huge towers, miles of antennas, and alternators that used as much electricity as a small town. Within countries, wireless could not compete with the cheap and efficient telegraph networks. Across oceans it was competitive only under unusual circumstances. Before World War I, the four major powers built stations linking them to their colonies. During the war, when it was impossible to lay cables, the US expeditionary force built stations in France to supplement the overloaded transatlantic cables. After the war, to meet the booming demand for transatlantic communications, companies found it faster and cheaper to build gigantic transmitters than to lay costly cables.

Longwave wireless had two flaws to which cables were immune: they were vulnerable to atmospheric static and to interception and cryptanalysis. The British, with their heavy investment in cables and their far-flung colonies and bases, were concerned with

keeping their cables viable. Thus, it came as a shock when Marconi, the man who had invented wireless, announced a new technology that threatened to make both longwave radio and submarine cables obsolete, namely shortwave radio (c. 13–120 m or 2.5–22 MHz).

Shortwave, though more vulnerable to atmospheric conditions and less reliable than longwave, had one enormous advantage: it was extraordinarily cheap; a transmitter cost a few hundred dollars instead of hundreds of thousands. Even small inexpensive transmitters could, at certain hours of the day, reach across oceans and around the globe.

The advent of shortwave coincided with the Depression. The collapse in international trade brought the British cable companies to the brink of bankruptcy, but they were rescued at the last moment by their government for strategic reasons. Although cables served a useful function in World War II for the transmission of top-secret information, they were overshadowed by the proliferation of shortwave transmitters throughout the world. Indeed, a war of movement over enormous theaters of war like the Russian plains or the western Pacific would have been inconceivable without a multitude of radios. At the same time, the reliance of armed forces on radio meant that the outcome of battles depended on communications intelligence and cryptanalysis as much as it did on ships, tanks, and planes.

5. *Broadcasting*

Broadcasting, the transmission of radio and later television programs to many receivers at once, played a tremendous role in the cultural history of every twentieth century nation. The first experiments in transmitting voice and music over the airwaves took place before and shortly after World War I. However, broadcasting had little impact on international relations before the 1960s, for two reasons. One was that most broadcasting was in the medium wave range (150–600 m, or 500–2,000 kHz), which did not carry far; only border regions could receive foreign broadcasts. The other was that programs were in the language and culture of each country (or rather, of each country's elite), with little appeal to foreigners.

International shortwave programs, such as those pioneered by the BBC in the early 1930s, were specifically designed to appeal to people in other countries. Soon, every major country had its shortwave station beaming news and propaganda to the world, often in many languages. No doubt such broadcasts appealed to the converted and reminded exiles of their homelands, but whether they persuaded more than a handful of foreigners is questionable.

Mention should be made of the interesting cultural phenomenon of ham radio operators who communicated around the world, not so much to convey information as to collect lists of other distant amateurs they had contacted, much as birdwatchers collect the names of birds they have spotted.

Television uses high-frequency waves that cannot bend with the curvature of the earth. Until the advent of satellites in the 1970s, transmitters could only reach local viewers. Satellite television, however, has opened up the world to broadcasts from foreign countries; all that is needed is a dish antenna. Countries with conservative theocratic or tyrannical regimes resent the intrusion of such foreign cultural phenomena as sex, violence, and free speech. However, prohibition is hard to enforce, and people around the world are much more interested in foreign television programs than they ever were in shortwave news broadcasts.

6. *International Telephony*

Telephony is a recent phenomenon. Though invented in the 1870s, it affected few people before the early twentieth century in North America, the 1920s in western Europe, and the post-World War II period in other parts of the world; at the beginning of the twenty-first century, there were fewer telephones in all of Africa than in Manhattan.

International telephony is an even more recent phenomenon. Due to the bandwidth requirements of the human voice, telephone messages could not use telegraph wires, but needed repeaters every few hundred miles. International calls were possible only between adjacent countries. Most telephone systems were government controlled, and calls to foreign countries had to go through an operator and were extremely costly. Calls overseas required two radio stations in each country, one to transmit and one to receive. Until the 1950s, such calls were considered newsworthy events rather than a medium of communication.

This situation began to change after the first transatlantic telephone cable was laid in 1954, and especially after the advent of satellite transmission of telephone calls in the 1970s and fibre-optic cables in the 1980s. As prices fell from dollars to pennies a minute, the number of international calls has multiplied exponentially, with no end in sight. Businesses and even many individuals now make calls to distant countries as readily as their parents once called across town.

As this is written, a new technology promises to change international telephony as dramatically as the first telephone cables: low-earth-orbit (LEO) satellites that can relay calls to and from portable telephones anywhere on earth, bypassing national networks. Although governments and national companies are likely to resent this intrusion on their monopolies, telecommunications technology is too invasive to stop. Eventually, the well-to-do will be connected around the world, regardless of governments and borders, while the rest of humankind looks on in wonder.

7. Data Communication and the Internet

The current revolution in communication is as radical as the telegraph's 'annihilation of time and space' that so amazed our ancestors a century and a half ago. By turning text, voice, music, pictures, and video images into bits, electronic digitization allows the transmission of all forms of information with equal ease and speed. Two associated technologies, microwaves and fiber-optic cables, have reduced the cost of transmission almost to zero. The results are nothing short of phenomenal. The first transatlantic telephone cable could carry a dozen or so telephone calls at once; the most recent fiber-optic cables across the Atlantic carry hundreds of thousands of simultaneous calls, or dozens of television channels, or any combination of data, voice, text, or images. Communication is not free because other costs remain, such as switching and distribution. But the price differences between local, long-distance, and international communications are shrinking fast.

The Internet began in the 1970s as a US military project to link the mainframe computers of research institutions. By the mid-1980s, when personal computers began to proliferate, engineers in Europe, North America, and Japan first created simple networks to transmit data, then linked them together into a super-network called the Internet that could transmit any information to any connected computer. The demand for this new form of communication soared beyond anyone's prediction, as the public logged onto a virtual world of chat rooms, bulletin boards, electronic mail, shopping, entertainment, advertisements, even music and video clips from around the world. And the cost is very low. For people who own a computer (even a primitive 'Web-TV' box) this amazing service costs a few dollars a month and the price of a local telephone call. All other costs (servers, cables, switches, software, and administration) are borne by governments or by advertisers. In the near future, the Internet promises to offer television and telephone service, and to compete with newspapers, libraries, and even universities. And unlike the technologies that it replaces, the Internet is oblivious to borders and political jurisdictions. We can only speculate about the effect on cultures and on the politics of nation-states.

What the public sees is only a fraction of the world of international digital data transmission. Computer networks, connected by fiber-optic cables, have revolutionized finance and commerce as well. Credit-card charges, engineering drawings, business contracts, stock market reports, late-breaking news, banking transactions, and myriad other data flit across the globe at the speed of light. The amount of money—in the trillions of dollars—that circulates at any given moment dwarfs the economies of the largest countries. The new communications media offer the potential for phenomenal economic growth and for disastrous crashes. Humans may still be bound by the laws of physics, but for information, space and time have truly been annihilated.

See also: Communication and Democracy; Communication and Transportation, History of; Communication: Electronic Networks and Publications; Communication: Geographic Aspects; Information Society; Information Technology; International Communication: Regulation; International Relations, History of; Internet: Psychological Perspectives; Postal Systems; Printing as a Medium; Radio as Medium; Telegraph; Telephone; Television: History

Bibliography

Aitken H G J 1985a *The Continuous Wave: Technology and America Radio, 1900–1932*. Princeton University Press, Princeton, NJ

Aitken H G J 1985b *Syntony and Spark: The Origins of Radio*. Princeton University Press, Princeton, NJ

Barty-King H 1979 *Girdle Round the Earth: The Story of Cable and Wireless and its Predecessors to Mark the Group's Jubilee, 1929–1979*. Heinemann, London

Douglass S J 1987 *Inventing American Broadcasting, 1899–1922*. Johns Hopkins University Press, Baltimore, MD

Flichy P 1991 *Une histoire de la communication moderne*. La Découverte, Paris

Griset P 1991 *Les révolutions de la communication, XIXe–XXe siècle*. Hachette, Paris

Headrick D R 1991 *The Invisible Weapon: Telecommunications and International Politics, 1851–1945*. Oxford University Press, New York

Headrick D R 2000 *When Information Came of Age: Technologies of Knowledge in the Age of Reason and Revolution, 1700–1850*. Oxford University Press, New York

Herodotus 1987 *The History Book VI* [trans. D. Grene]. University of Chicago Press, Chicago, pp. 105–6

Holzmann G J, Pehrson B 1995 *The Early History of Data Networks*. IEEE Computer Society Press, Los Alamitos, CA

Lubar S 1993 *InfoCulture: The Smithsonian Book of Information Age Inventions*. Houghton Mifflin, Boston

Pool I D S 1990 *Technologies Without Boundaries: On Telecommunications in a Global Age*. Harvard University Press, Cambridge, MA

Wilson G 1976 *The Old Telegraphs*. Phillimore, London

Young P 1991 *Person to Person: The International Impact of the Telephone*. Granta, Cambridge, UK

D. R. Headrick

International Communication: Regulation

The regulation of international communication was launched in Paris, in 1863, at a conference convened to lay the foundation of an international postal system

(Hamelink 1994, Mattelart 1996). For most of the next century, the regulation of international communication focused mainly on providing a few basic rules to ensure that communication resources could be used nationally, according to the goals and capacities of individual nation-states, and only minimally on relations between states. But the assumption that communication was a national affair requiring a minimum of international coordination disappeared with the emergence of a global communication environment in which communication technologies pay increasingly less attention to national borders. In the twenty-first century, any discussion of communication regulation has to be placed in an international context; thus, it is increasingly pertinent to speak of the international regulation of communication.

1. *The History of Regulation of International Communication*

In any field of endeavor, regulation implies recognition of a problem to be solved by authorities in a position to propose and eventually implement a solution. In the field of communication, regulation has typically sought to deal with issues involving the need to harmonize technical standards, allocate scarce resources, and ensure that at least a substantial portion of these resources is used in the interest of the common good.

The invention of the telegraph placed these issues on the international diplomatic agenda during the latter decades of the nineteenth century. The world's first permanent intergovernmental organization, the International Telegraph Union, was set up in 1865 to provide a framework for development of international telegraph and telegram services and the issue was problematic enough to spawn a spate of international conferences in Paris (1865), Vienna (1868), Rome (1872), and St. Petersburg (1875), resulting in a series of conventions regulating international telegraphy. Telegraph communication introduced a new set of challenges to national sovereignty, as signals paid no attention to national borders. However, the telegraph was still reliant on cable connections and these, of course, were territorially based. With hindsight, it may be said that regulating telegraphy was a relatively simple matter compared with what was yet to come.

The problem took on its modern face when wireless radiocommunication came on the scene at the turn of the twentieth century. Wireless radio made communication possible between any two points, provided that two conditions were met: each party had to be equipped with an appropriate apparatus and both had to be tuned to the same wavelength or frequency. Each of these conditions raised an important issue that soon required international regulation.

In 1901, a ship carrying Prince Henry of Prussia ran into difficulty off the coast of England. Its call for help was heard by a Marconi company receiving station on the Isle of Wight, which refused to forward it because of a company policy limiting dealings to ships equipped with Marconi transmitters. Prince Henry's ship, unfortunately, was outfitted with a rival German company's apparatus. The incident led to an international conference in Berlin in 1903, the essential purpose of which was to break the Marconi company's attempt to set up a worldwide monopoly in wireless radiotelegraphy.

It soon became clear that effective use of the 'air' would require some kind of prescription of the use of certain wavelengths for certain purposes, but it would be another 30 years before such a system was put in place. A more substantial conference—31 governments as opposed to eight—followed in 1906, and culminated in the first International Radiotelegraph Convention. The treaty opened ship-to-ship and ship-to-shore communication for the first time, obliging operators to exchange messages irrespective of their corporate affiliation. A subsequent conference in London in 1912 refined the treaty.

By the 1920s, a new and hitherto unexpected use of the new technology had appeared. Taking their lead from amateur operators, entrepreneurs in the budding US electronics industry realized that ordinary folk could be interested in acquiring radio receivers provided that some kind of attractive listening content were made available. The broadcasting industry was born, and with it a whole new set of international regulatory issues.

The main technical challenge to broadcasting was finding a way to organize the use of a radio spectrum characterized by scarcity: there were a limited number of wavelengths suitable for transmission and reception, and too many people, governments and companies who felt they had something to say (or sell). Within national jurisdictions this could be handled by licensing the use of the spectrum to specific parties. Different countries each had their own criteria for doing this. In most of Western Europe and the British Dominions, for example, licensing favored national public service broadcasting monopolies such as the BBC. In the USA, most of the broadcasting spectrum was licensed to privately owned commercial stations. However, regardless of the use to which broadcasting would be put in various countries, an overall international framework had to be put in place before national broadcast regulation could be effective at all.

With the rapid spread of radio broadcasting to multiple uses, including commercial entertainment, authorities feared the result of chaotic traffic jams unless a modicum of order were applied at the international level. In October 1927, representatives of some 70 countries (and 40 communication companies) convened in Washington at the behest of the US government to discuss the matter. US President Calvin Coolidge, opening the conference, declared that, 'An instrument of such far-reaching magnitude, fraught

with so great a power for good to humanity, naturally requires *national and international regulation and control*... In many fields our country claims the right to be the master of its own independent development. It cordially concedes the same right to all others. But in the radio field the most complete development, both at home and abroad, lies in mutual concession and cooperation. Your main endeavor will be to discover the rules which will be for the mutual advantage of all those who are connected with this great industry and who are the users of this means of communication.'

The conference's main order of business was to allocate spectrum use to different classes of activity: military, commercial shipping, broadcasting, etc. Rather than divide the spectrum among countries, it would be divided by group; all nations would be free to use the spectrum for the purposes thus decided, regulating national use accordingly. Further regulation might be necessary at the regional level, e.g., North America. But regulation would begin at the international level and everything would flow from that.

Categories of spectrum use were thus defined in a set of General Regulations annexed to the International Radiotelegraph Convention. A subsequent conference in Madrid in 1932 further refined the regulations and merged wireless and conventional telegraphy in one association, to be known as the International Telecommunication Union, or ITU. The term telecommunication has since evolved to encompass any form of distance communication—telephone, radio, telegraph, television—and more recently satellite, computer and Internet communication. Each of these technologies has added its own specific twist to the complexity of the international regulation of communication.

2. *Towards the International Regulation of Communication*

The regulation of international communication changed in character after the Second World War, with the emergence of the United Nations system of multilateral organizations (Schiller 1969). As we have seen, the oldest of these, the International Telecommunication Union, traces its lineage to 1865. However, an important change in quality was introduced in the immediate post-War period as international agreements began to deal with not only international but also domestic issues. A wide range of new issues, from trade to human rights, now came into play.

The immediate post-War period brought with it a new era in international communication. Article 19 of the Universal Declaration of Human Rights underscored the right to freedom of information. UNESCO, created in 1946, undertook a series of activities which have been crucial to developing an understanding of the links between communication and culture and their importance for human development. New regional bodies such as the Council of Europe have included communication in their sphere of concern, through periodic resolutions on specific themes and issues.

At the same time, communication issues began to crop up in the wake of the new economic multilateralism that flowed from the Bretton Woods agreements and creation of institutions such as the World Bank and the International Monetary Fund. The first General Agreement on Tarifs and Trade in 1947 was the setting for a vigorous debate that culminated in the acceptance of foreign film import quotas, notwithstanding the GATT's general thrust towards liberalization of national markets. The extent to which cultural 'products' constitute a specific type of commodity requiring its own set of international trade rules has been a constant feature of multilateral and regional trade negotiations since that time.

The question of national sovereignty over communication took on a new color in the political and technological context of the 1960s and 1970s, with the emergence of dozens of new postcolonial states and the development of satellite technologies making it possible to transmit sound and images irrespective of national borders (Smythe 1981). The unequal flow of information content from north to south and the increasing importance of technological resources led, conceptually, to the articulation of a 'cultural imperialism' thesis and, politically, to efforts to create a 'New World Information and Communication Order' (Tomlinson 1991, Galtung and Vincent 1993). These issues were highlighted by publication, in 1980, of the report of a UNESCO commission chaired by Irish jurist and human rights activist Sean MacBride (UNESCO 1980) and the subsequent withdrawal from UNESCO of the USA, UK, and Singapore. Meanwhile, with far less fanfare, an ITU report of the same period documented the unequal distribution of technical resources for communication worldwide (International Telecommunication Union 1984).

UNESCO has since adopted a more low-key communication strategy, emphasizing the training of communication professionals and development of media institutions in the 'transitional' states of central Europe, Africa and Asia. In 1995, however, a UN/UNESCO World Commission on Culture and Development published a major report which brought many of the lingering issues back on the table, with an updated analysis (United Nations/UNESCO 1995).

Meanwhile, the collapse of the Berlin Wall made it possible to extend the development of a global communication system under the hegemony of the western alliance. Cultural and communication goods and services were consistently integrated to the international trade agreements of the 1990s, despite the resistance of countries with strong traditions of national regulation in these areas, such as Canada and France. Under the auspices of the World Trade

Organization, an open global market in telecommunications is emerging—paradoxically, one in which, through a raft of mergers, there are fewer and fewer players. Foreign ownership and content regulations are increasingly under attack and threatened with extinction in blueprints for the future such as the aborted Multilateral Agreement on Investment (MAI), an initiative of the Organization for Economic Cooperation and Development that failed to crystallize in 1998. Experiences such as the MAI and the heavily contested Millenium Round of the WTO which opened in Seattle in December 1999 illustrate the highly politicized nature of international regulation in general. Communication, however, is not only an object of such regulation, it is also in many respects the catalyst for it.

Generally, the social science study of the regulation of international communication has developed along parallel lines with the area itself. For many decades it occupied, at best, a sleepy corner in the study of international relations. Still a relatively sparse field, it has grown substantially since the late 1960s, however, with the emergence of a wide range of critical studies in mass communication. A seminal work in this regard was Herbert Schiller's (1969) *Mass Communication and American Empire*. Throughout the 1970s, the work of scholars such as Armand Mattelart, Cees Hamelink, Hamid Mowlana and Kaarle Nordenstreng began to shape a more general field of international communication, which broadly focused on the previously mentioned 'new world order' debates centered within UNESCO. Issues such as inequality in east–west and north–south information flows and the absence or inadequacy of international regulatory mechanisms were often highlighted by this scholarship.

More recently, a younger generation of communication scholars has begun paying attention to international regulatory issues in the context of broader debates on globalization and global governance. This interest has generated a wide range of institutional studies of organizations such as the ITU and, especially, the WTO, from the perspective of international regulation of communication. The trend towards international media conglomerization and the spread of new global communications technologies such as the Internet have also led to more scholarly focus on international regulatory issues such as access and copyright. Much of this scholarship is policy-oriented and has a strong normative thrust.

3. *The International Regulation of Communication in a World of Trade*

Since the late 1980s and particularly in the wake of the collapse of the Berlin Wall, it is fair to say that a new global communication environment has been taking shape. From a regulatory perspective, this environment is characterized by the fact that, unlike in previous eras, there is no discernible center. Rules continue to be made at the national and regional level but these are increasingly impacted and constrained by supranational agreements.

Multilateralism is clearly one of the central trends in global governance; but different players have different stakes and, consequently, need different strategies for intervening in a political system based on multilateral relations. In this context, even large powers are recognizing the need for global coordination and regulation in order to create the state of order necessary for the flourishing of their interests. Thus, for example, at the turn of the century the European Union and the USA were each floating proposals for some kind of formal regulatory framework to govern global communication.

The need to provide some kind of stability to the context in which the Internet was developing lay at the heart of the matter. Paradoxically, over the rhetorical claims of users and entrepreneurs alike, this most open and freewheeling communication technology was crying out for the institution of an orderly governance framework—much as radio had a century earlier.

The urgency of such a project was driven home in September 1999 when CEOs of the world's most important communication companies met in Paris to discuss the global regulation of electronic commerce. The Global Business Dialogue on Electronic Commerce, or GBDe as the group is known, urged governments and international organizations to coordinate their regulatory efforts regarding the Internet, which it described as a global medium requiring a global policy approach. This is but one example of the regulatory challenges facing the emerging global communications environment.

Indeed, the powerful technology of the Internet has exacerbated many old problems related to communication policy at the national level while introducing new ones globally. Oddly, national regulators are tending towards abandoning attempts to regulate the Internet, just as the global issues it raises cry out for some kind of transnational regulatory intervention (see 'Regulating the Internet' 2000).

The example of the Internet illustrates that while most of the debate around questions of communication regulation still pit the market against the state, the issues are increasingly international. In the case of the Internet, it has been suggested (Mathiason and Kuhlman 1999) that we should be moving towards an *international framework convention*, which would articulate basic norms of how the Internet is to be governed and establish a mechanism for monitoring compliance with those norms and determining future changes. A framework convention would be different from a treaty-based regime such as the one pertaining in trade under the auspices of the WTO. Its task would be to sort out the roles and responsibilities of the various national, international and private actors

involved in the development of this 'global enabling technology'. It would need to be negotiated, these authors suggest, on the basis that the Internet is part of the global commons. This type of proposal clearly depends on a broad consensus of what communication is about. It can only hope to work to the extent that communication technology is recognized as a public good. In other words, the normative view one takes about something like the Internet is much more important in determining the limits and possibilities of regulation than the nature of the technology or the policymaking capacity of the state. Debates surrounding the normative definition of communication technologies are therefore a key prerequisite to any attempt at international regulation in this area.

4. *Models for Regulation and a Global Approach to Communication*

There are at least *four models* that can be identified with respect to the international regulation of communication:

(a) The libertarian model: no regulation. With the new digital technologies such as the Internet, this is the approach that is being taken by most national regulators (Australia's being an important exception), mainly because they do not know what to do or how to do it. It is also largely favored by grassroots activists who are benefiting from the system's present openness. However, the history of communication technologies shows that left to its own devices, this open access is not likely to last. A libertarian model of Internet governance will likely lead eventually to closed doors, restricted access, and limited communication.

(b) Self-regulation: this is the approach most often favored by industry players, with the encouragement of national regulators. It is presented as the solution to problems such as abusive content and the protection of rights, on the argument that consumers will respond if they are not satisfied. However, as we saw with the example of the GBDe, even the promoters of self-regulation recognize the need for a global structural framework for communication activity, within which industry self-regulation would take place.

(c) The closed club, or top-down institutional model: where plans are negotiated in organizations such as the OECD, G7, or WTO, as well as in the new institutions emerging as the corporate sector fills the vacuum created by the retreat of national governments from regulatory issues. One such agency is the Internet Corporation for Assigned Names and Numbers (ICANN), a private organization set up at the initiative of the US government in 1998 for the purpose of directing traffic on the Internet.

(d) The long march through the institutions: a process that is tied to the broader project of democratization of global governance, reflected in some of the initiatives around UN reform and in notions such as 'cosmopolitan democracy' (Archibugi et al. 1998). Access to global policymaking is being fostered to some extent by some important initiatives in multilateral agencies such as UNESCO and the ITU which have demonstrated some sensitivity to the concerns of civil society and the inclusion of NGO representation in their activities.

The international regulation of communication can also be seen as part of a larger sphere of intervention encompassing other aspects of policy. By thinking about *a global public policy approach to communication* it becomes possible to begin addressing a whole range of issues for ensuring the public interest in communication globally, such as the following:

(a) regulation of commercial activities in the public interest, to guarantee equitable access and basic services;

(b) funding and institutional support for the creation and sustaining of public service and alternative media;

(c) placing limits on corporate controls resulting from transnational concentration of ownership in new and conventional media and telecommunications;

(d) guarantees of access to available media channels on the basis of public interest criteria;

(e) development of universal codes and standards for curtailing the spread of abusive contents;

(f) facilitating networking capacity through communication technologies of not-for-profit organizations; and

(g) provision of public communication spaces for conflict resolution and democratic dialogue on global issues.

Before this can happen, however, credibility will need to be given to the idea that the global communication environment, from the conventional airwaves to outer space, is a public resource, to be organized, managed and regulated in the global public interest. Broadening access will require appropriate transnational regulatory mechanisms, in addition to mechanisms for a more equitable distribution of global commercial benefits. From a perspective that places high priority on the importance of the public interest obligations of communication systems, there is a need for the international appropriation of some air and space for distribution outside the country of origin of viable creative products that currently have no access to the global agora that figures so prominently in utopian discourses on information technology. The convergence of communication technologies requires a parallel convergence in programs and policies, and also the invention of new models, new concepts, and a general new way of thinking about communication.

In the twenty-first century, the international regulation of communication will be an essential part of the global governance system. This indicates a need to develop democratic mechanisms for ensuring access to the regulatory process at the global level, without which it will soon be difficult, if not impossible, to

promote a public interest through policy intervention *in any area* influenced by communication.

Communication has always been linked to democratic struggles, and is increasingly relevant to thinking about broader issues such as the role of the state and human rights. In the twenty-first century, because of the particular situation of communication in the overall environment of globalization, issues regarding the regulation of communication have an important impact on a range of related questions. The international regulatory framework for communication is therefore emerging as a key structural component of the emerging global governance framework in general.

In grappling with this, it is important to see that even if it is true that we are moving towards a global society, this does not mean that there is no longer a need for regulation. The development of economic or market globalization does not in itself negate the need for rules; indeed, the new communication environment recognizes this and is constantly searching for new regulatory mechanisms appropriate to capturing and defining what might be termed the global public interest.

Communication and information technologies, although of little interest in themselves, are strategically critical in efforts to intervene economically on the one hand, and in the cultural sphere on the other. The regulation of communication revolves around struggles over who gets to use these technologies, under what conditions, in order to promote which projects, and in who's interests. In the contemporary context of globalization, 'communication' is another way of describing the technological space at the interface of the economic and the cultural. What goes on in that space—the international regulation of communication—is thus of great import for human and social development.

See also: Broadcasting: Regulation; Communication: Electronic Networks and Publications; Global Governance; Globalization: Political Aspects; International Advertising; International Business; International Communication: History; International Law and Treaties; International Marketing; Postal Systems; Regulatory Agencies

Bibliography

Archibugi D, Held D, Kohler M 1998 *Re-imagining Political Community. Studies in Cosmopolitan Democracy*. Polity Press, Cambridge, UK

Galtung J, Vincent R C 1993 *Global Glasnost: Toward a New World Information and Communication Order?* Hampton Press, Cresskill, NJ

Hamelink C J 1994 *The Politics of World Communication: A Human Rights Perspective*. Sage, London

International Telecommunication Union, Independent Commission for Worldwide Telecommunications Development (chaired by Sir Donald Maitland) 1984 *The Missing Link*. ITU, Geneva, Switzerland

Mathiason J R, Kuhlman C C 1999 *An International Communication Policy: The Internet, International Regulation and New Policy Structures*. New York University, New York

Mattelart A 1996 *La Mondialisation de la Communication*. Presses Universitaires de France, Paris

Mowlana H 1996 *Global Communication in Transition*. Sage, Thousand Oaks, CA

'Regulating the Internet' 2000 *The Economist* 10 June, 73–9

Schiller H I 1969 *Mass Communication and American Empire*. Augustus Kelley, New York

Smythe D W 1981 *Dependency Road: Communications, Capitalism, Consciousness and Canada*. Ablex, Norwood, NJ

Tomlinson J 1991 *Cultural Imperialism: a Critical Introduction*. Johns Hopkins University Press, Baltimore, MD

UNESCO, International Commission for the Study of Communication Problems (chaired by Sean MacBride) 1980 *Many Voices, One World: Communication and Society, Today and Tomorrow*. Kogan Page, London

United Nations/UNESCO, World Commission on Culture and Development (chaired by Javier Perez de Cuellar) 1995 *Our Creative Diversity*. World Commission on Culture and Development, Paris

Vincent R C, Nordenstreng K, Traber M 1999 *Towards Equity in Global Communication*. Hampton Press, Creskill, NJ

M. Raboy

International Justice: Philosophical Aspects

International justice, like justice within borders, is largely concerned with duties of aid in response to inequality and duties of noninterference. However, the change to an international context strikes most people as profoundly altering the moral norms to be applied. They think that tax-financed aid to the needy should attend to foreigners to a far lesser extent than to compatriots. And they think that considerations of national sovereignty support stringent prohibitions against intrusion in another country's affairs, even to promote just and humane goals rightly pursued within a country's borders. Current philosophical debate over the proper norms of foreign aid and national sovereignty is a forum both for challenges to the routine assessments of the moral importance of borders and for appeals to these assessments as exposing the typical insensitivity of modern moral philosophy to the special ties of obligation binding compatriots.

1. International Inequality

Most people in countries that are relatively rich per capita think that foreign aid should be a small part of their government's total aid budget even if it is an effective means of relieving the most urgent deprivations in the world at large. Much philosophical

thinking about international justice has been prompted by arguments that this patriotic bias is incompatible with the equal valuing of everyone's life that morality demands.

1.1 Impartial Concern

This challenge to ordinary patriotism is clearest in moralities that take the moral point of view to be, fundamentally, a perspective of impartial concern for all. Utilitarianism is the most familiar and simplest example. If the right choice is always the one that creates as much happiness as any alternative, then each of us should always be willing to contribute to the needy whenever the consequent loss in happiness would be less than the consequent gain. Since the vast majority of the neediest live in countries with relatively few local means of relief, citizens of richer countries should use tax-financed aid to relieve these desperate foreign needs even if serious but lesser domestic needs for education, meaningful work, or comfortable housing go unmet.

Even within a morality of impartial concern and under the factual assumption that foreign aid effectively provides for urgent foreign needs, transfers are ruled out when they would so lower a rich country's productivity that it becomes a less efficient instrument for increasing world happiness. Moreover, if one departs from the classical utilitarian focus on individual acts and determines moral responsibilities by asking what institutional assignments of responsibility produce the most happiness, there is (as Goodin and others have noted) a powerful case for some significant duty of patriotic bias in most countries. A worldwide system of institutional responsibilities including special responsibilities toward needy compatriots relies on the spontaneous concerns of the most effective social groups, guides aid according to the best-informed deliberations, and favors the least costly transfers. Still, if responsibilities are to be determined by welfare-maximizing institutions, then the efficiency of modern international communication and transportation, the vastly unequal resources of the richest countries and poor ones, and the specially desperate neediness of the many who suffer in poor countries would seem to call for a demanding responsibility of citizens of the richest countries to help needy citizens of poor countries, a responsibility overriding the standard patriotic bias, appropriate though the latter might be in a less unequal world.

The challenge to routine attitudes toward foreign poverty is especially daunting because it does not depend on utilitarianism's distinctive view of the dictates of impartial concern. In assessments of domestic justice, many political philosophers favor, instead, the perspective of Rawls' *A Theory of Justice*, (1971) according to which the justice of a society's basic institutions is determined by standards of political choice which all participants would prefer in the so-called 'original position,' in which each seeks to promote his or her life-goals while ignorant of their specific content and of his or her special advantages or disadvantages (see *Contractarianism*). The outcome, Rawls argued, would include an egalitarian 'difference principle' requiring that the life-prospects typical of the worst-off social position be as great as possible. Although Rawls has resisted its use in assessing global economic justice, others (for example, Beitz and Pogge) have taken a cosmopolitan original position, in which the deliberants do not know where in the world they live, to be an appropriate response to worldwide interdependence and inequality. If (as some of Rawls' own arguments suggest), commitment to the original position reflects a duty to use institutions to reduce lifelong burdens of the brute bad luck of birth in a disadvantaged social situation, then this cosmopolitan original position, presumably yielding a global difference principle, seems an appropriate response to the enormous, undeserved lifelong difference it makes to be born in Burkina Faso, say, rather than in Belgium.

The case for demanding duties of foreign aid is also strengthened by a parallel, which Singer strikingly deployed, with certain well-established duties of rescue. It is wrong for a passerby to refuse to save a toddler drowning in a wading pool because rescue would ruin his expensive suit. It is hard to discern any morally relevant difference between this obligatory rescue and preventing the death from dysentery of a child in a foreign country through a similarly costly donation to a project of foreign aid.

1.2 Ethical Nationalism

Arguments that impartial concern entails stringent duties of international aid are frequently offered in criticism of routine patriotic biases. But others (for example, David Miller) take these shocking consequences to reveal the inadequacy of moral universalism, the assumption, widespread among philosophers, that all moral duties ultimately derive from some perspective of equal appreciation of everyone's worth. Perhaps this universalism underrates special obligations of loyalty to certain kinds of unchosen associates.

Alleged duties of special concern for fellow-members of one's nationality have been of special interest to critics of universalism because such claims have a powerful appeal and yet are hard to reconcile with a universalist morality. Most members of a nationality feel a special obligation to help continue the distinctive culture and history with which they identify and to look after the well-being of those who share this culture and history. According to an ethic of nationality, this sentiment is morally insightful. Those who can

be part of such a project of continuity and special concern ought to seize this opportunity. As a consequence, they will have less concern for outsiders, even though these outsiders have equally valuable lives.

Current philosophical partisans of ethical nationalism are well aware that nationalism can lead to callousness and brutality. As a moral barrier to nationalist excesses, they insist on respect for certain basic rights possessed by everyone in virtue of mere common humanity. Even if one's positive duty of aid to foreigners is quite undemanding, one must respect stringent negative duties not to steal from them or kill them. However, it is troubling that the vast distance between this humane nationalism and the outlook that motivates genocide and wars of national expansion has been created by piecemeal stipulation of a limited set of duties to humanity. If the duty of loyalty to the nationalist project of mutual concern and cultural continuity so easily overrides positive concern for outsiders' needs, why can't it justify ethnic cleansing, as a means of preventing the eventual disintegration of a nationality through intermarriage and cultural intermingling? A principled explanation of why the ethnic cleanser's nationalism is depraved would appeal to some regulative principle on the basis of which both goals of nationality and specific human rights receive their moral authority—but the principles claiming such comprehensive power are the bases for universalist moralities. Moreover, any political morality in which nationality-based concern is fundamental seems open to morally troubling discriminations in policies of domestic aid, for example, the provision of lesser health benefits to those who are not members of the dominant nationality.

1.3 International Justice and Domestic Fairness

In addition to its role in stimulating doubts about the adequacy of universalist morality, resistance to demands for global beneficence has contributed to investigations of forms of universalism that do not express a fundamental perspective of impartial concern. In general, there is no need to be impartially concerned for different people in order to regard them as having equal moral worth. Parents are usually much more concerned for their own children than for other people's children, without remotely supposing that the lives of other people's children are any less valuable.

Because of its well-established role in discussions of political obligation, the duty to treat others fairly was one of the first universalist resources used to justify patriotic bias in a framework of equal respect but not of impartial concern. Benefiting from one's engagement in a cooperative scheme that depends on most participants' willingness to subordinate immediate self-interest to the advancement of the scheme, one ought to seek a fair division of the benefits and burdens of the shared project among all participants. Free-riding shows inadequate respect. So, on general grounds of fairness, compatriots have a special mutual obligation of reciprocity for one another's contributions to their shared political community. Dagger and others drew the international moral: lacking a similar debt of reciprocity to outsiders, just participants in a political community will be less concerned for the nonparticipants' well-being.

Such duties of political community are far from excluding foreign aid. If important foreign burdens can be relieved without any serious loss to compatriots, this help might be required in order to express the equal valuing of all. Moreover, citizens of per capita rich countries participate in an international system of economic and political institutions, involving lesser, but significant, benefits and burdens, and consequent lesser duties of reciprocity.

In criticizing this appeal to mere political community, ethical nationalists have questioned its capacity to generate adequate obligations of concern for needy compatriots. It is hard to see how an obligation of reciprocity could dictate special concern for compatriots who cannot contribute to one's political community because they suffer from severe lifelong physical disabilities. (Similarly, compatriots burdened by natural, rather than social, disadvantages seem inadequately supported by a complementary argument from political fairness, the argument that the ultimate coerciveness of government makes special concern for those burdened by laws that one upholds a dictate of respect for persons.)

1.4 Universal Respect and Special Relationships

Such doubts concerning the adequacy of appeals to political fairness have contributed to growing interest in the interaction of equal respect for all with the proper valuing of special relationships. Most people would deny that someone properly values friendship if she would desert a friend as part of a strategy for introducing friendship into the lives of two other, friendless people, taking friendship to be nothing more than a good to be impartially promoted. In general (as Scheffler, Scanlon and others have emphasized), the proper valuing of a worthwhile relationship entails special concern for others to whom one happens to be so related. Since those who lack valuable and beneficent relationships, through no fault of their own, can be specially needy, the proper valuing of relationships conflicts with standards of distributive fairness requiring the minimization of burdens of brute bad luck. But it does not obviously conflict with the injunction to show equal respect for persons; indeed, the failure properly to value one's relationships, as in the case of the philanthropic desertion of the friend, seems to show disrespect. By the same token, the proper valuing

of relationships seems a dictate, rather than a competitor, of alternative fundamental precepts equivalent to the injunction to conform to rules that could express equal respect for all, precepts requiring conformity to rules that all could freely impose on themselves without lacking self-respect or that all could freely impose on themselves in light of a rationale that would be a reasonable basis for their joint agreement.

A hybrid strategy in which the demands of equal respect are specified, in part, by reflection on special relationships may be a means by which a broadly universalist morality can avoid duties of impartial global beneficence without lapsing into callousness. When compatriots' potentially demanding loyalty to their political institutions is compatible with their self-respect, perhaps, as in the case of a worthwhile friendship, proper valuing of their relationship entails special loyalty to one another, a loyalty that should not be calibrated to others' actual contributions any more than loyalty among friends. This loyalty, deriving from political obligations acquired at birth, may preclude the hard words, 'I would do nothing for you if you had been unable to contribute to our joint political project, through no fault of your own.' If so, then, on pain of arbitrariness, this special concern might have to be extended to others who cannot take part in the joint political project because of severe disability but are within the scope of an obligation to take part if one can. Assuming that special concern for associates ought to be proportionate to the importance of the shared institution and the potential demands of the institutional loyalties that sustain it, there would be a significant though lesser duty of foreign aid, reflecting engagement in international cooperation.

A related recent trend in moral philosophy has insisted on the irreducible role of one's special relationship to oneself in limiting duties to aid others. Many advocates of legitimate self-concern deny that someone is obliged to embrace rules of giving that would, foreseeably, seriously worsen his or her life, even rules solely dictating the relief of burdens much greater than the burdens of giving. Perhaps the morally crucial difference between the practices of rescuing victims close at hand that are Singer's starting point and the practices of global rescue meeting his demand for global concern is that the balance of foreseeable lifelong costs and potential benefits would violate the legitimate self-concern of the less vulnerable if the rules of rescue were global. Perhaps (as Richard Miller and others have argued) departures from patriotic bias in tax-financed aid on the part of governments of countries that are the best off per capita would impose losses on their citizens, in trusting civic relationships and other morally important goods, that would be too great a cost, even as a means of relieving more urgent needs abroad. If so, unequal concern would show no disrespect, but only a legitimate concern for one's own well-being, whose legitimacy the foreign poor could acknowledge without any lack of self-respect.

Of course, much depends on how the seriousness of costs and deprivations is assessed, on how respect, self-respect, and reasonableness are construed, and on how the line is drawn between mere nongiving and participation in coercive projects, contexts in which appeals to one's well-being or one's valuable relationships provide much less support for partiality. For example, Shue has argued that poor people in the poorest countries would lack self-respect if they accepted practices of foreign aid that neglected their basic rights, i.e., their rights to the security, subsistence, and health care needed to enjoy any further rights, even if this neglect protected cultural goods in per capita rich countries or the relatively expensive life-projects with which many people in rich countries intelligently identify as giving order and value to their lives. Pogge has emphasized the constraints on the self-advancement of people in poor countries imposed by the worldwide economic dominance of firms and institutions controlled by people in rich countries, and the tendency of international practices of investment and purchase to encourage the corrupt repression that blights many poor countries. An appropriately nuanced moral typology of coercion and complicity might reveal duties of justice toward the poor of these countries that could require large sacrifices from people in richer parts of the world. The shift from a universalism of impartial concern to a universalism of equal respect is a shift to a broad framework, whose content and coherence is the subject of heated controversy.

2. *Sovereignty*

Rights to freedom from interference play an important role in international political discourse, as they do in domestic controversies over civil liberties. However, the characteristic explicit topic of international complaints of unjust interference is the autonomy of governments (or, more precisely, national governments), rather than individuals. There is broad agreement on certain presumptions against international interference: special justifications are required to legitimate interference by one government in political life within the territory controlled by another; this burden of proof is extremely heavy when interference takes the form of armed invasion, which is never justifiable merely as a means of advancing the interests of the invading government or its citizens; defensive war is usually a just response to armed attack. Still, virtually every theorist agrees that the burdens of justifying interference can sometimes be met and that some regimes could not justly be defended against armed attack from abroad. After all, individuals, not governments, are the ultimate bearers of moral rights, and, far from uniformly protecting the rights of people in their territories, governments often violate them, sometimes in atrocious ways. Different views of the

relationship between the rights and morally urgent needs of individuals and the collective political processes in which they take part have sustained conflicting specifications of the duty not to intrude on sovereignty and the rightness of defending sovereignty.

2.1 Moral Individualism and Political Autonomy

The gap between alleged rights to noninterference of governments, on the one hand, and rights and morally significant needs of individuals, on the other, is specially wide in moralities that value political arrangements solely as instruments for promoting individualist goods, i.e., goods enjoyed by individuals which do not intrinsically depend on their participation in any specific kind of societal process. A strong presumption against invasion which does not respond to an armed attack may be justifiable on this basis. Such intrusion usually provokes resistance and deadly violence, which may disastrously expand as military alliances and rival national interests are mobilized. If the government of the invaded territory has protected people in the secure enjoyment of fundamental liberties, they lose a vital basis for well-being, and a new one may be hard to create. On the other hand, a government that violates fundamental liberties is apt to receive new popular support when nationalist sentiments are inflamed by invasion.

Still, just as a universalism of impartial concern challenged routine favoritism toward needy compatriots, impartial concern for individualist benefits challenges moral barriers to interference that most people would impose. Since the well-being of those in the invaded country has no special standing, a war of conquest could, in principle, be justified by net gains to all affected due to increased benefits to the subjects of the invading government—say, the gains to poor people in Iraq from the acquisition of oil reserves formerly controlled by rich Kuwaitis. Moreover, since no intrinsic value is ascribed to such collective goals as joint participation in government on the basis of a shared, distinctive culture, foreseeable benefits to the conquered in health, literacy, or material prosperity can yield a case for imperialist conquest that destroys popular local political institutions.

2.2 Collective Self-determination

Seeking to raise higher moral barriers to interference, many theorists have ascribed independent moral value to individuals' engagement in certain collective processes through which they determine their own political institutions within their own familiar political communities. If participation in such a collective process can enrich someone's life, apart from individualist benefits, then it is easy to explain why the conquest of another people's territory cannot be justified simply by net individualist gains. The difficulty is to avoid the overvaluation of the project of basing political outcomes on what is locally shared. In 1994, the government of Rwanda massacred a half-million Tutsis and moderate Hutus, acting solely within its borders. Foreign governments surely had no duty to permit the bloody enactment of this local drama to run its course just because it was local.

What exceptions will be made to the presumption against military intervention and how far they will extend to nonmilitary interference will depend on what forms of collective self-determination are valued. Both a strong presumption against interference and strictly limited exemptions (such as Walzer's 'rules of disregard' of the prohibition of nondefensive war) can be derived from the special value of national self-determination, the effort to base common political arrangements on the shared identities constitutive of nationality. Foreign invasion normally interferes with this process, but may promote it, in exceptional cases, by supporting the national uprisings of minorities or by countering the prior intervention of outside powers; moreover, if the value of national self-determination is contingent on avoidance of the gravest injustices in the struggle to determine national institutions, there is a room for intervention to stop massacre or enslavement.

Still, a valuing of national self-determination that supports distinctive prohibitions of military intervention must fence off some struggles in which some compatriots threaten others with injustices so grave that they are rightly resisted with deadly force. Lesser injustices, after all, are not just causes for the use of deadly force by any group, domestic or foreign. To insist that those grim struggles must be kept fratricidal strikes many as a dangerous overvaluing of the national character of political life as opposed to its justice. In addition, basing foreign military neutrality on the valuing of national self-determination often seems to combine cleareyed acceptance of nonideal national political processes with naive blindness to international economic processes. Far from relying solely on local resources, tyrannical sovereign governments have nearly always benefited from internationally recognized prerogatives to negotiate loans, sell natural resources and buy arms.

In constructing substantial but nonabsolute barriers to intervention, Rawls has relied on a different standard of valued collective self-determination, analogous to his domestic political liberalism, through which fellow-citizens of a single society respectfully respond to diversity in their comprehensive moral doctrines (see *Public Reason*). His international analogue is a 'law of peoples' that would be willingly accepted by representatives of all peoples whose dominant political culture is, at least, 'decent'—morally flawed, perhaps, through departures from liberal democracy, but still meriting respect as in-

cluding an international precept that military force is not to be an instrument for advancing the national interest and a domestic commitment to the rule of law and the pursuit of the common good.

This respect for certain illiberal political cultures has troubled many of those who share Rawls' view that such cultures are, nonetheless, seriously defective. If, for example, a political culture disrespects women, despite seeking in its own way to include their good in the common good, should there be a strict prohibition of international interference, even through trade discrimination and diplomatic pressure? A rigid rule of noninterference seems to deny that individual persons are the bearers of moral rights.

2.3 Kantian Sovereignty

In assessing international duties of noninterference, as in assessing duties of international aid, some theorists try to steer a middle course in which a perspective of universal respect for individuals ascribes moral authority both to certain local forms of collective self-rule and to norms of world community. This broadly Kantian project has an important antecedent in Kant's own arguments that world justice requires a world federation of sovereign liberal democracies.

A broadly Kantian assessment of the virtues and limits of political sovereignty takes its start in arguments for liberal–democratic forms of political association and against world government. If a morally responsible person is someone committed to living by principles expressing full and equal respect for all, then such a person will aspire to a political life through which each can freely, rationally and self-respectfully commit himself or herself to uphold the outcome of the shared process of public deliberation. Such a political order will guarantee each the civil liberties and democratic rights that no one could self-respectfully renounce and will base legislation on principled persuasion in which each can effectively participate, seeking common ground yet open to inevitable compromises. Because the diversity of cultural norms makes it impossible to base adequate legislation on principled persuasion worldwide and because the diffusion of civic activity worldwide would make it so hard to resist an oppressive central authority, a mutually respectful political life must be based on national sovereignty (which Kant contrasted with the 'soulless despotism' of world government). However, the commitment to base one's life on principles expressing equal respect for everyone everywhere also dictates support for appropriate worldwide norms of conduct among governments.

Because of the connection between worldwide respect for persons and localized republican polity, such international norms would require noninterference with foreigners who are working out the terms of their shared political life on the basis of deliberative democracy and respect for civil liberties, even if the process and its outcomes are currently defective. In addition, presumptions against interference will extend to more deeply flawed regimes. Given the dispersal of military power among national governments, a broad permission to intervene to rectify serious injustice could produce a ghastly enlargement of the perils of Locke's state of nature, in which the anarchic pursuit of justice by individuals leads to a deadly spiral of intervention and counterintervention, driven by inevitable disagreements and conflicts of interest. Moreover, an ethic based on observance of rules that all could freely and self-respectfully impose on themselves will count it as a serious loss when people who have attained Rawls's standard of decency in their political outlook are forced to conform to laws departing from their own defective understanding of the demands of justice. Finally, if peace and commerce tend to be the best background for displaying the advantages of liberal democracy and overcoming social conflicts that make it unworkable, international institutions will tend to promote an ultimate pacific confederacy of liberal democracies if they prohibit military intervention and enhance economic interaction among governments above a threshold of civil and political liberties which is well below the full demands of justice.

These considerations parallel some of Rawls' arguments for his law of peoples and are one natural extension of civic republican thinking into the international realm (see *Republicanism: Philosophical Aspects*). They would sustain an international code of conduct, expressing an aspiration to worldwide liberal democracy, which includes a strong presumption against military intervention and a corresponding right to repel invasion. The precise nature of the limited rules allowing intervention will depend on the global milieu of military capacities and political sentiments. But they would probably permit some military interventions against home-grown tyrants, invited by popular movements in the direction of liberal democracy, even when humane nationalists' threshold of atrocity has not been crossed. There would be even greater scope for nonviolent interference with foreign tyranny through diplomatic pressure, trade discrimination, or public condemnation meant to change political life within a foreign territory.

A role in affirming and preserving the distinctive culture of a nationality will not be a fundamental reason for leaving a government alone. This ultimate detachment from nationalist concerns strikes some as a recipe for cultural imperialism. In addition, ethical nationalists argue that stable and effective justice depends on pervasive sentiments of solidarity based on nationality. If so, the universalist appraisal of sovereignty leads to the alienated stance of support for the public encouragement of evaluative outlooks whose insightfulness the universalist denies.

Perhaps the most distinctive feature of Kantian and civic republican justifications of the presumption against intervention is their tendency to support systematic shifts in political authority that would reduce the scope of national sovereignty. These justifications ground the international rights of national governments on the need to support public facilities through which people can express respect for one another's autonomy. Without any intention to meddle in another government's political affairs, decisions made in the territory of one government (for example, interest-rate policies of the US Federal Reserve) can have a profound impact on the lives of people elsewhere. If we should show respect for fellow-citizens by giving them a voice in regulating processes that would otherwise subordinate their lives to the unconstrained will of others, then, on the same grounds, we ought to show respect for those whose lives are molded by transnational processes, by providing them with a voice in the collective regulation of these processes. As the needed international institutions expand their scope and their democratic legitimacy, many disputes over the morality of foreign aid and intervention may be ended, not by conclusive arguments but by obsolescence, due to new international patterns of political loyalty, reciprocity and self-respectful trust.

See also: Equality: Philosophical Aspects; Human Rights in Intercultural Discourse: Cultural Concerns; Individualism versus Collectivism: Philosophical Aspects; International and Transboundary Accords, Environmental; International Arbitration; International Law and Treaties; Justice: Philosophical Aspects; Kantian Ethics and Politics; Nationalism: General; Offenses against the Laws of Humanity: International Action; Sovereignty: Political

Bibliography

Beitz C R 1999 *Political Theory and International Relations*, 2nd edn. Princeton University Press, Princeton, NJ

Dagger R 1985 Rights, boundaries, and the bonds of community: A qualified defense of moral parochialism. *American Political Science Review* **79**: 436–47

Goodin R E 1988 What is so special about our fellow countrymen? *Ethics* **98**: 663–86

Kant I 1784/1795 Idea for a universal history with a cosmopolitan purpose; perpetual peace. In: Reiss H (ed.) 1970 *Kant's Political Writings*. Cambridge University Press, Cambridge, UK

Luban D 1980 Just war and human rights. *Philosophy & Public Affairs* **9**: 160–81

MacIntyre A 1984 *Is Patriotism a Virtue?* University of Kansas Press, Lawrence, KS

Miller D 1995 *On Nationality*. Oxford University Press, Oxford, UK

Miller R W 1998 Cosmopolitan respect and patriotic concern. *Philosophy & Public Affairs* **27**: 202–24

Nussbaum M C with respondents 1996 *For Love of Country: Debating the Limits of Patriotism*. Beacon Press, Boston

Pogge T W 1989 *Realizing Rawls*. Cornell University Press, Ithaca, NY, (especially Part III: Globalizing the Rawlsian Conception of Justice)

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Rawls J 1999 *The Law of Peoples*. Harvard University Press, Cambridge, MA

Scanlon T M 1998 *What We Owe to Each Other*. Belknap Press of Harvard University Press, Cambridge, MA

Scheffler S 1997 Relationships and responsibilities. *Philosophy & Public Affairs* **26**: 189–209

Shue H 1996 *Basic Rights: Subsistence, Affluence and U.S. Foreign Policy*, 2nd edn. Princeton University Press, Princeton, NJ

Singer P 1972 Famine, affluence, and morality. *Philosophy & Public Affairs* **1**: 229–43

Walzer M 1977 *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. Basic Books, New York

R. W. Miller

International Law and Treaties

1. *Definition*

International law and treaties is a term that in English has at least three modern meanings: (a) legal rules and legal processes generated by more than one state; (b) legal rules and legal processes that concern interstate relations; and (c) legal rules and legal processes that regulate the international transactions of both states and private parties.

2. *Nature and History*

Traces of what we now term *international law and treaties* are as old as written history. Evidences of compacts between states and rulers of states date back at least four thousand years; they chronicle agreements about, *inter alia*, war and peace, foreign merchants, and ambassadors, all still lively topics of the discipline. It is safe to say that international law and treaties figure as necessary features of human society, regardless of any juristic definition, description, or argument.

The modern intellectual tradition relating to international law and treaties owes much to Roman law. The Roman term *jus gentium*, the law of nations, described legal rules pertaining generally to foreigners when specific foreign rules were unknown or in conflict. These were rules thought to be so basic that they were shared by all nations. Hence, like natural law, the law of nations was presumed to exist without positive expression, for example, in legislation. In Roman law, *jus gentium* ordinarily was employed in

legal matters concerning private transactions, for example, marriage, wills, and commerce.

Despite its origins in Roman private law, the law of nations was adopted by European jurists in early modern times to describe certain public legal relations. For example, in the sixteenth century, Spanish writers such as Suarez and Vitoria employed principles of *jus gentium* to argue that the Spanish Crown owed legal duties, for example, of basic humanitarian conduct, to all peoples, even the indigenous peoples conquered by Spain in the Americas.

The often alleged 'father' of international law was Hugo Grotius (1583–1645), a Dutch jurist exiled to Paris following the suppression of the liberal party in the Netherlands. Already an accomplished diplomat, jurist, and theologian, Grotius devoted the remainder of his life to the cause of peace. In the midst of Europe's disastrous Thirty Years War (1618–48), Grotius published the law of nations' most famous book, *De Jure Belli Ac Pacis*, to demonstrate that legal rules and processes defined and limited the sovereign rights and aggressions of otherwise bitterly divided Catholic and Protestant states, and that such legal rules and processes even applied to relations between Christian and non-Christian states.

Grotius elaborated a legal theory that still serves as the foundation of modern international law. He argued that sovereign states were defined and bound by two kinds of legal rules drawn from the law of nations: positive law and natural law. The positive law of nations, expressed in treaties and custom, was made by the explicit and implicit agreement of sovereign states. The natural law of nations, expressed in the natural laws of reason and religion, was the result of necessary rules binding everyone, even rulers of states.

The first part, at least, of Grotius' theory of *jus gentium* was quickly married to emerging seventeenth-century political theory, like that of Hobbes, positing the idea of *sovereign* states. Grotius' consensual theory of the law of nations helps explain why treaties, such as the Peace of Westphalia that terminated the Thirty Years War in 1648, are legally binding: sovereign states have authority both to legislate rules internally for their own territory and to make rules externally for themselves through interstate agreement. Grotius' legal theory also explains why sovereign states should not interfere in the internal affairs of other sovereign states: a nation's sovereign authority must be limited so that the sovereign authority of other states is protected. Such international legal principles that weave sovereign states together with international law continue to guide international relations today.

In his age, Grotius' *jus gentium* excited many rulers, including Gustavus Adolphus of Sweden, who is said to have used Grotius's book as a pillow while campaigning on his many wars, as well as many universities. In Germany, The Netherlands, and England, professorial chairs in the new discipline of the law of nations were created, commencing an outpouring of scholarly works further describing and developing the discipline.

Not all students, of course, were pleased by their professors; one of the earliest pupils of the noted English law professor, William Blackstone, rebelled against what he was taught about the law of nations. In 1789, Jeremy Bentham (1748–1832) rejected the term the *law of nations* as professed by Blackstone; Bentham argued in *An Introduction to the Principles of Morals and Legislation* that the discipline should be renamed *international law*. Bentham's invented word, *international*, successfully passed from international law to denominate many other disciplines, for example, *international relations*, *international politics*, and *international economics*.

More or less inadvertently, Bentham changed not only a term but the very definition of a discipline. Grotius' and Blackstone's classical definition of the *law of nations* focuses on legal rules and processes generated by more than one state. Bentham's definition of *international law*, however, focuses on the subjects of the discipline; these Bentham said were states and states alone. The classic *jus gentium* regards individuals, as well as states, as its subjects, possessing legal rights and duties and accessing legal process. For the sake of theory, Bentham rejected individuals as proper subjects of *international law*.

A third definition of the discipline, along with a new term to describe it was provided by the US jurist, Philip Jessup, when in 1956 he proposed replacing both the term the *law of nations* and the term *international law* with a new term *transnational law*. Jessup defined the discipline of *transnational law* as encompassing any legal rule or process, municipal or international that has to do with international transactions of any state, entity, or individual.

It may or may not make a difference how one defines the discipline. For example, a rule about delimitation of the continental shelf not only emerges from a multistate source, a treaty, *à la* Grotius's *law of nations*, but it also relates to interstate relations *à la* Bentham's *international law*, and deals with international transactions *à la* Jessup's *transnational law*. Hence, such a rule easily fits within the discipline as described by all three definitions. However, other times the different definitions lead to conflict. For example, if one believes, along with Bentham, that only states may be subjects of *international law*, then one cannot believe that individuals have international legal rights and duties, for example, be protected by international human rights law against abusive state action. A contrary position is usually taken by those who subscribe to the ordinary definitions of the *law of nations* and *transnational law*. Nowadays in practice, although the term *international law* is very popular, some of its exclusivities have been softened both by the classical tradition of the *law of nations* and by the modern sentiment expressed in *transnational law*.

3. Domain

The domain of *international law and treaties* is vast and growing. The discipline's classical fields, many hundreds of years old, include the *laws of war* (both about going to war, *jus ad bellum*, and about the proper conduct of war, *jus in bello*), the *law of the sea*, the *law of ambassadors and diplomacy*, the *law of treaties* (how treaties are made, interpreted, and implemented), the *law of the recognition and succession of states*, and the *law of merchants* (traditionally, *lex mercatoria* was part of the law of nations). More recently, new fields in the discipline have emerged, including the *law of international organizations and institutions*, *international humanitarian law*, *international human rights law*, *international trade law*, *international financial law*, the *international law of Antarctica*, the *law of outer space*, *international environmental law*, *international refugee and immigration law*, *international criminal law*, the *law of international arbitration*, *international tax law*, and the *international law of state responsibility*. The proliferation of fields in the discipline of international law is due in large measure to the increasing numbers and kinds of international relationships among states, peoples, corporations, and institutions, as well as to the inability of national law and regulation to adequately cope with such international transactions.

4. The Legislative and Enforcement Questions of International Law

To comprehend the fundamentals of modern international law, it is important to understand that two questions have always beset the discipline. First, how is it possible to authoritatively generate international legal rules? And second how is it possible to realistically enforce international law? In short, without an effective international political system, it seems at first glance that there is neither an authoritative way to make international law nor a realistic process to enforce it. These questions are answered in practice by remarkably effective systems of international legal rules and international legal process.

5. International Legal Rules

How are international legal rules authoritatively made? International lawyers often turn to Article 38 of the Statute of the International Court of Justice (the 'ICJ') to begin to describe the formation of international legal rules. Article 38 sets forth the rules available to the ICJ, the United Nations's court, in deciding cases before it, albeit there are other kinds of international legal rules as well. The three principal sorts of international legal rules available to the ICJ are (a) rules drawn from explicit international agreements (mostly treaties), (b) rules drawn from international custom, and (c) rules drawn from general principles of law. These are sometimes described as the *formal sources* of international law in that they are the sorts of rule-making that are acceptable in form as legitimate in international law (cf. the *material sources* of international law which are the evidences that show the features of a specific rule).

Rules drawn from international agreements include international legal rules set out in treaties, compacts, conventions, protocols, or whatever may be a written agreement between or among states. Nowadays, there are tens of thousands of such instruments; their norms constitute the greatest number of international legal rules. These explicit international rules are comparable both to statutes in municipal law (in that they establish rules for general conduct) and to domestic contracts (in that they are based on party consent). Treaty rules are the international legal rules that tend to be both least ambiguous and most clearly based on state consent. All other kinds of international legal rules are more troubled by problems both of ambiguity and of authority.

The second type of international legal rule is *customary international law*. Ordinarily, customary international law is perceived as based on the implicit agreement of states and is composed of two elements. First is a consistent pattern of state practice; this defines the rule. Second is a belief, an *opinio juris*, that compliant state behavior is based on a sense of binding legal obligation. In practice, the determination of a consistent pattern of state practice is almost always controversial and involves a subjective weighing of evidences from state practice proving and disproving the rule. *Opinio juris* is even more controversial: states rarely indicate that they are behaving in international relations in a certain way because they feel compelled to do so by a sense of international legal obligation. Rather, *opinio juris* is usually an evaluation added by a judge or a commentator to a description of state practice.

The third kind of legal rule prescribed by Article 38 of the ICJ Statute is *general principles of law*. These are principles generally shared by all legal systems. This concept is rather similar to the original Roman notion of *jus gentium*: rules so basic that they are to be found everywhere. In modern practice, general principles of law are usually developed by jurists through a comparative law search. The method proceeds by examining diverse municipal legal systems. If a rule is found to be common to most or all of the municipal legal systems examined, then the rule is also presumed to exist as a general principle of law in international law.

Of the three formal sources of international law provided for in the ICJ statute, treaty and custom are, more or less, consensual in nature. Even general principles of law can be seen to have a consensual aspect since it can be argued that if states already

recognizes a legal rule internally, they may be presumed to accept that self-same rule externally. Beyond the ICJ statute, however, other formal sources of international legal rules exist without displaying such real or alleged consensual foundations.

Such nonconsensual rules may be seen as being no longer in the camp of positive law, but belonging to that of natural law. Indeed, natural law itself has long been an important formal source of international law. However, the term *natural law*, once so common, is nowadays rarely used in discussions of international law. Instead, newer equivalent terms have emerged. Key among these are *jus cogens* and *fundamental norm*, both expressing the natural law idea that some legal rules are so fundamental that they can not only be found everywhere, but that they trump any conflicting rule of positive law. *Jus cogens* or fundamental norms are often thought to include both the basic principles underpinning the international system, for example, the very notion of 'sovereignty,' as well as those basic principles protecting key human rights, for example, legal norms protecting individuals against genocide, slavery, and torture.

Another nonconsensual form of international law is *equity*. Sometimes viewed as a corrective of the strict application of a legal rule (equity *contra legem*), equity can also involve rule creation (equity *praeter legem*), and rule interpretation (equity *infra legem*). Equitable rules or principles can be based on nonlegal criteria, for example, distributive justice and basic notions of fairness.

6. *International Legal Process*

How are international legal rules really enforced? Most times when international legal rules are recognized in practice, they are applied by the nation states themselves. So, for example, all kinds of international legal rules, whether they be drawn from treaties, custom, general principles of law, *jus cogens*, or equity, are applied from time to time by municipal courts. Approaches to this process usually are thought to be either *dualistic* or *monistic*. *Dualism*, the prevalent view, holds that municipal legal systems and international legal systems are distinct from each other (dualist) and that there needs to be rules of *incorporation* providing why, when, and how a rule of international law is to be applied by a municipal court. *Monism*, the less usual view, sees both international law and municipal law as parts of a unified system of law where incorporation of international legal rules into municipal law is to be presumed.

Different municipal legal systems have different answers to questions of the incorporation of international law. Moreover, a municipal legal system's rules about incorporation may vary depending on the type of international legal rule concerned. So, for example, though treaty rules are sometimes deemed to be *self-executing* in the United States, that is, requiring no domestic legislation to put them into internal effect, in England treaty rules are never self-executing: all require an act of Parliament before they can be employed as rules of decisions by municipal courts. However, respecting customary international law, the United States and England share a similar approach. Both municipal legal systems deem some sorts of customary international law to be part of the *law of the land* and available to municipal courts without legislative act.

Besides providing rules of decision for municipal courts which then enforce the international rules alongside municipal law, international law is also enforced by procedures in the international legal system. There are, for example, an increasing number of international courts. Oldest among these is the International Court of Justice at the Hague, which dates its antecedents to the Permanent Court of Arbitration established by the Hague Peace Conference of 1899, and to the Permanent Court of International Justice, established under the auspices of the League of Nations in 1921. The International Court of Justice was established alongside the United Nations, its parent institution, in 1945. The Court is composed of 15 judges elected by the UN General Assembly. There are usually ICJ judges from China, France, Russia, the United Kingdom, and the United States.

Reference has already been made to the ICJ Statute, the treaty that constitutes the Court. The ICJ statute also provides for two kinds of jurisdiction for the Court. First is its *contentious jurisdiction*; here the Court, with the prior consent of the parties to the dispute, actually decides cases between states. However, neither international organizations nor individuals may be party to the ICJ's contentious jurisdiction. The second sort of ICJ jurisdiction is its advisory jurisdiction: at the request of some international organizations (though neither states nor individuals), the ICJ may render a legally non-binding advisory opinion on a legal matter. Since 1921, the practice of the Permanent Court of International Justice and the International Court of Justice has remained roughly steady: about three or four judgments or advisory opinions rendered each year.

Since 1950, a great many new international courts have been created. Each of these courts is limited either geographically or by its subject matter, but some have become a great deal busier than the older and more universal ICJ. Europe's two regional international courts—the European Court of Human Rights in Strasbourg (to which 41 European states belong) and the European Court of Justice in Luxembourg (numbering 15 adherent states)—nowadays decide hundreds of cases each year. Less busy, but still important are such international courts and international dispute-settlement procedures as the Law of the Sea Tribunal in Hamburg, the World Bank's

International Centre for the Settlement of Investment Disputes in Washington, the Inter-American Court of Human Rights in Costa Rica, the World Trade Organizations's dispute settlement procedures in Geneva, and the international criminal tribunals for the former Yugoslavia in the Hague and that for Rwanda in Kenya. Unlike the states-only provisions limiting the ICJ, all these newer international tribunals permit individuals in some circumstances to be parties to litigation.

In addition to the increasing number of international courts, there are a great many international arbitral procedures available to provide *ad hoc* jurisdiction to settle international legal disputes, public and private. Many of these international arbitrations are organized under the auspices of one of the international arbitral institutes. Two of the busiest are the International Chamber of Commerce in Paris and the American Arbitration Association in New York.

Finally, international law is often enforced simply by the practice of governments. Reciprocity and good faith account, outside any formal dispute settlement procedures, for the lion's share of cases in which international law is effectively employed. It would be difficult for much of the world's business, public and private, to get done unless governments followed the similar practices set out in treaties, custom, and the other kinds of international law.

7. *Prognosis*

It is likely that international law and treaties, already employing tens of thousands of international lawyers worldwide, will continue to prosper and expand. Modern trade, communications, and transportation have all been globalized much more thoroughly than modern politics that is still largely state-centered. It is unlikely that without a political or social impetus for internationalization of political decision-making, that the sovereign state will remain the principal vehicle of government. Accordingly, it will be international law and treaties that will shoulder the burden of harmonizing and facilitating the growing transactions among the nations.

See also: International Arbitration; International Business; International Justice: Philosophical Aspects; International Relations, History of; International Relations: Theories; International Trade: Commercial Policy and Trade Negotiations; Offenses against the Laws of Humanity: International Action; Public Goods: International

Bibliography

Bentham J 1970 (eds. Burns J H, Hart H L A) *An Introduction to the Principles of Morals and Legislation*. Athena Press, London

Brierly J L 1949 *The Law of Nations*, 4th edn. Oxford University Press, Oxford, UK

Brownlie I 1990 *Principles of Public International Law*, 4th edn. Oxford University Press, Oxford, UK

Franck T M 1990 *The Power of Legitimacy Among Nations*. Oxford University Press, Oxford, UK

Grotius H 1925 (ed. Scott J B) *The Law of War and Peace*. Carnegie Endowment for International Peace, Washington, DC

Higgins, Rosalyn 1994 *Problems & Process: International Law and How We Use It*. Oxford University Press, Oxford, UK

Janis M W 1999 *An Introduction to International Law*, 3rd edn. Aspen, Gathersburg/New York

Janis M W, Evans C (eds.) 1999 *Religion and International Law*. Martinus Nijhoff, The Hague/Boston/London

Janis M W, Noyes J E 2001 *International Law Cases and Commentary*, 2nd edn. West, St Paul, MI

Jennings R, Watts A 1992 *Oppenheim's International Law*, 9th edn. Longman, Harlow, UK

McNair A D 1938 *The Law of Treaties*. Columbia University Press, New York

Rousseau C 1970 *Droit International Public*. Sirey, Paris

Wheaton H 1836 *Elements of International Law*. Carey, Lea & Blanchard, Philadelphia, PA

M. W. Janis

International Marketing

International marketing has grown out of the weaknesses of international trade theory which considered no differences in local consumer tastes and preferences and did not take into account the possibility for firms to launch marketing operations directly on foreign markets through sales subsidiaries. Definitions of international marketing are first examined, followed by a review of the central debate in this field, that is, whether products and marketing strategies should be standardized for increasingly globalized world markets or whether they should be tailored to local consumer needs and marketing environments. With the emergence of global markets, culture now appears as a prominent variable for international marketing decisions. The third section explains how cross-cultural differences must be taken into account to highlight significant differences in areas such as consumer behavior, market research, and advertizing practices. An in-depth understanding of what is global and what remains local is a key input for designing successful international marketing strategies.

1. *Definitions of International Marketing*

1.1 *Traditional Definitions of International Marketing*

The early definitions of international marketing in the 1950s emphasized foreign market operations, that is,

the distribution of products abroad and the development of marketing channels and consumer franchise in foreign countries as ways and means to reach larger customer bases. As a consequence, emphasis was put on developing stable relationships with agents and dealers in foreign countries. This view of international marketing was linked largely to international business concerns such as the choice of the right mode of entry, the assessment of political risk in target countries, and the logistics of international marketing (see *International Business*).

In the 1970s, with growing foreign direct investment of multinational companies, international marketing more and more has focused on the design of marketing strategies for local markets and their co-ordination within world markets (Wind et al. 1973). The understanding of local peculiarities in consumer behavior and marketing environments increasingly was considered as a required step in the definition of marketing strategies in an international context. The 4Ps paradigm of marketing (product, price, place, promotion) has been applied extensively to foreign markets (see *Marketing Strategies*), considering how to adapt products to local tastes and requirements, to adjust price policies internationally, and manage local distribution channels and advertizing campaigns in diverse areas of the world. International marketing has, therefore, been a replication and extension of marketing knowledge 'made in the USA.' Up to the 1980s this process has been implemented pragmatically with much respect for local contexts through customized marketing strategies. However, the issue of co-ordination of marketing strategies across national markets, either at the regional or global level, has emerged as a key organizational issue in the 1980s and 1990s for multinational companies (see *Multinational Corporations*). More centralized control of exceedingly independent subsidiaries has been required because the increasing globalization of the world economy imposes large-scale operations.

1.2 International Marketing as the Diffusion Process of Marketing Knowledge

Marketing concepts and practices were developed initially and for the most part in the United States and have been popularized by Kotler's *Marketing Management* in its numerous translations and editions (Kotler 1994). Most of the books on marketing management were borrowed from the United States and then translated directly without much adaptation. As Van Raaij (1978) stated concerning consumer research: 'Consumer research is largely "made in the USA" with all the risks that Western American or middle-class biases pervade this type of research in the research questions we address, the concepts and theories we use, and the interpretations we give.'

However, the success of the word 'marketing' gave a new image to trade and sales activities in countries where previously it had been socially and intellectually devalued. Despite the success and the seemingly general acceptance of the term 'marketing' at the international level, there have been some basic misconceptions of it in many countries, marketing often being reduced to one of its dimensions such as market research, sales promotion, or advertizing. Marketing knowledge has been imported progressively, first plainly, as universal and explicit knowledge, then merged with local tacit knowledge and ways of doing. In many countries, marketing superimposed itself on local selling practices and merged with them rather than replaced long-established commercial practices. As Lazer et al. (1985) point out in the case of Japan: 'what has occurred (in Japan) is the modification and adaptation of selected American constructs, ideas, and practices to adjust them to the Japanese culture, that remains intact.'

2. The Adaptation/Standardization Debate in International Marketing

2.1 The Trend Towards the Globalization of Markets

As advocated by Levitt (1983), we see the emergence of global markets for standardized consumer products on a previously unimagined scale of magnitude. Ancient differences in national tastes would tend to disappear, while local consumer preferences and national product standards would be 'vestiges of the past.' Consumers world-wide would look for good quality/low-cost products and global competitors would seek to standardize their offer everywhere. In Levitt's words it would 'not (be) a matter of opinion but of necessity' (1983).

However, globalization is a process that occurs mostly at the competition level. The GATT treaty (General Agreement on Tariffs and Trade) and the WTO (World Trade Organization) progressively have replaced tariff and nontariff barriers by entry barriers related to scale of operations, marketing knowledge, and corporate experience. As far as consumer behavior and marketing environments are concerned, natural entry barriers related to culture and language diminish progressively and only in the long term. There are still many different marketing 'villages' rather than a 'global village.'

The globalization of consumption displays a complex pattern of mixed global and local behavior. On the one hand, there is striking convergence as concerns basic demographics, for instance, the population, comprises more and more older persons, the size of households is constantly decreasing, and the proportion of immigrants is increasing in most countries,

Table 1
Arguments in favor of international standardization or local adaptation

Product attributes	Arguments in favor of adaptation	Arguments in favor of standardization
Physical attributes	Cost-reducing adaptations Local standards and regulations, marketing environment	Experience effects Economies of scale International standards
Services attributes	Local peculiarities in service, maintenance and distribution	Significant learning effects 'global' customer base
Symbolic attributes	Unfavorable image of company, country-of-origin, or brand name Inappropriate symbolic meaning	Favorable image of company, country-of-origin, or brand name Demands for 'universals'

with higher concentration in large cities. The same convergence phenomenon is to be observed for the sociocultural environments in the form of growing equality between men and women and increasing percentages of working women, while health and environment concerns are on the rise. Convergence in consumer behavior can be observed at a broad level: services tend to replace durables in household budgets and demand is growing for healthcare, environmentally friendly, fun, and convenience products (Leeflang and Van Raaij 1995)

2.2 How to Design 'Globalized' International Marketing Strategies

The real issue is not a dichotomous choice between total adaptation or full standardization since international marketing performance has been shown to be a combination of both adaptation and standardization strategies (Shoham 1996) and to depend on a large number of factors related to the four components of the marketing mix (Baalbaki and Malhotra 1995). Around a core product offering that is standard worldwide, most global companies such as Coca Cola or McDonald's customize when needed. Product customization may also result in market differentiation thus creating a competitive advantage vis-à-vis actual competitors and raising entry barriers against potential competitors. Excellent global companies standardize as much as feasible and customize as much as needed.

Table 1 proposes a systematic description of arguments in favor of adaptation or standardization, according to the different levels of product attributes: physical, service, and symbolic attributes. Some arguments in favor of either adapting or standardizing originate from within the company which can benefit from changing its way of operating. Other arguments are related to external constraints imposed by the environment, consumer behavior, and regulations: they imply company adaptation to market demand, either by adapting or standardizing its offer.

As concerns the physical attributes, consumer behavior matters a lot, especially the frequency of consumption, the amount consumed per helping, etc. For instance, the required size of a cereal packaging will not be the same in a country where an average consumer eats 50 grams of cereal daily for breakfast and in another country where servings are larger but it is consumed less frequently. Even pizza, a product supposed to be the epitome of international standardization, largely is customized to local tastes. Brand names, although in some rare cases global, need some attention when transferred to foreign markets to avoid unintended meaning.

3. The Emergence of Culture in the Field of International Marketing

Traditional international trade doctrine has laid the foundations for a denial of culture in international marketing. Ricardo's law of comparative or relative advantage is based on the assumption that products and consumers' tastes, habits, and preferences are perfectly identical across countries. However, quantity and price do not matter solely: local consumers invest meaning beyond the utilitarian aspect of the product. Rather than being merely commoditized, generic, and indefinitely marketable, goods and services increasingly are viewed in international marketing as being singularized and invested with cultural meaning by local consumers who display both utilitarian and nonutilitarian consumption motives.

3.1 Unique Consumption Experiences

Consumption is still largely a local reality. Far from being uniquely related to culture, local consumption reflects also climate, customs, and the mere fact that much of our lives is still experienced, shared, perceived, and interpreted with near people who share the same kind of 'local knowledge' in the Geertzian sense

(Geertz 1983). Consumption experiences remain local while integrating much global influence because part of the cultural meaning vested in goods and services is now shared world-wide. As noted by Applbaum and Jordt (1996): 'Globalizing influences have bored intercultural tunnels around the world, but core meaning systems such as those wrapped up in the idea of the family, continue to differ significantly.' Consumers attribute meaning to products and services in context, especially what it means to desire, to search, to evaluate, to purchase, to consume, to share, to give, to spend money, and to dispose. Consumption experiences are full social facts based on the interaction with other participants in the marketplace such as manufacturers, distributors, salespeople, and other consumers.

For instance, consumption as disposal involves views of what is the appropriate relation to the environment, of what is clean vs. dirty, and of how cleaning efforts should be allocated. Consumption generally leads to the final destruction of goods, this being true also for consumer durables when they are obsolete or out of order. Paper-based products are a good case in point: filters for drip-coffee machines are white in France and yellow-brown in Germany (*naturbraun*), paper handkerchiefs are generally white in France and yellow-brown in Germany, and toilet paper is generally rosy or white in France and greyish in Germany. The Germans express their willingness to be environmentally friendly (*umweltfreundlich*) by purchasing paper-based products whose color exhibits their genuinely 'recycled' nature, that is, not bleached with chlorine-based chemicals used to whiten recycled paper. The same holds true for the German writing and copy paper whose greyish and irregular aspect would be considered by most of the French as 'dirty' and poor quality. The difference in consumer experience lies in the difference of continuity in the ecological concern. Germans are deeply committed to protect their natural environment because they live on a territory about three times more densely populated than France and they insist on strong coherence between pro-environment discourse and actual consumption behavior. The two peoples seem in any case to have different ways of combining and reconciling nature and culture. Understanding local consumer behavior in a cross-cultural perspective (Howes 1996; Sherry 1995) is seen more and more as a prerequisite for the design of sound international marketing policies.

3.2 International Market Research and the Issue of Cross-cultural Equivalence

A basic issue in cross-cultural marketing research is to assess whether the concepts used in questionnaires and interviews have similar meaning for respondents in different cultural contexts. For instance the issue of conceptual equivalence has to be addressed before testing the influence of certain constructs on consumer behavior. Such basic concepts as beauty, youth, friendliness, wealth, well-being, sex appeal, and so on often are used in market research questionnaires where motivation for buying many products is related to self-image, interaction with other people in a particular society, and social values. They seem universal. However, it is necessary to question the conceptual equivalence of these words when designing a cross-cultural questionnaire survey. Even the very concept of 'household' (widely used in market research) is subject to possible inequivalence: Mytton (1996) cites the case of Northern Nigeria where people often live in large extended family compounds or *gida* which are difficult to compare with the prevalent Western concept of household which reflects the living unit of a nuclear family. In general, market research measurement instruments adapted to each national culture (the Emic approach) offer more reliability and provide data with greater internal validity than tests applicable to several cultures (the Etic approach, or 'culture-free tests'). But it is at the expense of cross-national comparability and external validity: results may not be easily transposable to other cultural contexts.

3.3 Understanding How Culture Influences International Marketing Strategies

While being globalized as much as possible, international marketing strategies must be tailored to local consumer behavior, taking into account cultural influences in the design of the 4Ps. For instance, European cars overwhelmingly have manual gearboxes. Europeans do not consider an automatic gearbox as a standard feature for a car, as in the United States, for a number of reasons, certain being related to regulation (driving licenses have to be obtained on a car with a manual gear box, except for handicapped persons), others to deep-seated prejudices (automatic cars consume more petrol) or to social beliefs such as that automatic gear-boxes are only for luxury cars or for handicapped persons (Usunier 1999).

Price is a central element of relational exchange, that is, as a signal conveying meaning between buyer and seller, marketers and consumers, and between companies and their middlemen. In the area of international price policy, it is important to assess which is the dominant price-related behavior: do local consumers tend to use price as a proxy of quality? Do they tend to be price minded? Is bargaining prices a behavior that is considered socially acceptable? Answers to these questions will be key inputs in defining adequate local pricing.

The case of Japanese distribution *Keiretsus* exemplifies the cultural embedding of distribution channels and the difficulty to enter foreign channels when being a 'cultural outsider.' Relationships be-

tween channel members are rooted deeply in local patterns of human and economic relationships; distribution appears a 'cultural filter,' which must be considered carefully along with other criteria before choosing a foreign distribution channel.

Cross-national variations in sales promotion methods have to be considered when transferring promotional techniques across borders. One must check that the very purpose of the promotional technique is understood locally, a problem in conceptual equivalence. For instance, in many societies being given a free sample is difficult to interpret. Gratuity is understood either as a sign of poor quality ('they give it because they cannot sell it') or as a sign of naivety of the manufacturer ('let's take as much as possible'). In recent years, Procter and Gamble experienced major problems with free samples in Poland where some people disregarded free samples while others broke mailboxes to steal product samples.

For reasons of image consistency, many companies now try to promote their products globally through standardized advertizing campaigns which use the same advertizing strategy, themes, and execution world-wide. However, marketing communications are based on language, both verbal and nonverbal. Language shapes our world-views in as much as the words we use and the way we assemble them in speech correspond to particular experiences and assumptions about the world in which we live (see *Sapir–Whorf Hypothesis*). Advertizing, as the main tool for communicating marketing messages to customer audiences is sensitive to local cultures and languages. Thus, before transferring campaigns cross-nationally, international companies have to decide which elements should be localized and which ones can be similar world-wide in two main areas, advertizing strategy (information content, advertizing appeals, etc.) and advertizing execution (characters and roles represented, visual and textual elements, etc.).

See also: Alliances and Joint Ventures: Organizational; Capitalism: Global; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; International Law and Treaties; International Organization; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; International Trade: Geographic Aspects; Market Areas; Marketing Strategies; Markets and the Law; Monetary Policy; Rational Choice and Organization Theory; Venture Capital; World Trade Organization

Bibliography

Applbaum K, Jordt I 1996 Notes toward an application of McCracken's 'cultural categories' for cross-cultural consumer research. *Journal of Consumer Research* **23**: 204–18

Baalbaki I B, Malhotra N K 1995 Standardization versus customization in international marketing: An investigation using bridging conjoint analysis. *Journal of the Academy of Marketing Science* **23**: 182–94

Geertz C 1983 *Local Knowledge*. Basic Books, New York

Howes D 1996 *Cross-cultural Consumption*. Routledge, London

Kotler P 1994 *Marketing Management*, 8th edn.. Prentice Hall, Englewood Cliffs, NJ

Lazer W, Murata S, Kosaka H 1985 Japanese marketing: Towards a better understanding. *Journal of Marketing* **49**: 69–81

Leeflang P S H, Van Raaij W F 1995 The changing consumer in the European Union: A meta-analysis. *International Journal of Research in Marketing* **12**: 373–87

Levitt T 1983 The globalization of markets. *Harvard Business Review* **61**: 92–102

Mytton G 1996 Research in new fields. *Journal of the Market Research Society* **38**: 19–33

Sherry J F 1995 *Contemporary Marketing and Consumer Behavior*. Sage, Thousand Oaks, CA

Shoham A 1996 Marketing-mix standardization: Determinants of export performance. *Journal of Global Marketing* **10**: 53–73

Usunier J C 1999 *Marketing Across Cultures*, 3rd edn. Prentice Hall, Englewood Cliffs, NJ

Van Raaij W F 1978 Cross-cultural methodology as a case of construct validity. *Advances in Consumer Research* **5**: 693–701

Wind Y, Douglas S P, Perlmutt H V 1973 Guidelines for developing international marketing strategies. *Journal of Marketing* **37**: 14–23

J.-C. Usunier

International Migration by Ethnic Chinese

The mass international Chinese migration began in the middle of the nineteenth century and today makes up 30 million people. It has evolved towards a diasporic configuration characterized by a multipolarity of migration and an interpolarity of relations in which networks play an important role and explain the economic efficiency of this diaspora.

1. The Migration from China

For centuries, or even millenniums, there has been a migration from China to Southeast Asia. These migrations mainly comprise an economic, political, or military urban elite. The Chinese migration really reached its peak in the middle of the nineteenth century when the combination of several factors made people leave. It was the result of the two Opium Wars, the Taiping Rebellion, the abolition of slavery, and the colonial enhancement of Southeast Asia by the English, the Dutch, and the French.

The two Opium Wars (1840–42 and 1856–60) and the Taiping Rebellion (1860–65) provoked tremendous social change and terrible famines, especially in the southern provinces, which are principal centers of

departure even today. The progressive abolition of slavery during the first half of the nineteenth century brought about a demand for cheap labor in order to replace slave work in former slavery countries. But above all, the colonization of Southeast Asia and the development of new English, French, and Dutch colonies (Indochina, Indonesia, Malaysia, Singapore, and so on) also created a demand for workers in plantations (hevea, pepper plant, sugar cane, tobacco, and others) and in mines, a demand that was no longer filled by slavers or by local labor.

From then on, the geography of international Chinese migration was maintained for a long time, almost exclusively affecting the southern provinces of continental China and mainly spreading into Southeast Asia and into former slavery areas. This geography still plays a fundamental role in the general distribution of migration today, although there have also been substantial modifications. As a result of the treaties of the Opium Wars, the creation of the 'coolie trade' allowed colonial powers to hire workers in order to develop their dependent territories. The coolie trade produced flows of departure towards the Caribbean, the Indian Ocean islands, and Polynesia, but especially towards Southeast Asia. These workers were, for the most part, landless peasants. They were among the poorest and saw in these contracts, however miserable, the possibility for them and for their families to survive. The large majority of workers came from the southern provinces of China (Guangdong, Fujian, and Hainan).

Within 25 years, from 1876 to 1901, 4,850,000 emigrants, a considerable figure, are assumed to have passed in transit through ports in southern China. During that time, the number of people belonging to the Chinese diaspora was estimated at 8 million, most of whom were found in the Nan Yang areas (South China Sea) of Asia. Moreover, several dozens of thousands of people were scattered over the insular constellations of the Indian and Pacific Oceans, some hundreds of thousands went to the eastern rim of the Pacific Ocean, from Canada to Chile, and also to the Caribbean islands, in particular to Cuba and Jamaica.

1.1 Chinese International Migration Today

The ancientness of the migration, interracial marriages, the constant flow of migration and the fact that the ethnic Chinese usually take the nationality of their receiving countries make an accurate measure of the extent of the diaspora difficult because it is hard to determine what an 'ethnic Chinese' is. What differentiates ethnic identity from other social identities (professional, religious, and so on) is, as Max Weber states, the shared belief in a common origin, whether it is real or imagined. Ethnic identity is, therefore, a social construction not a natural one. It is still a valid criterion for defining a population because this belief strongly configures the relationships among individuals themselves and thus determines specific social behaviors with important effects on economic, political, and cultural patterns.

Bearing these considerations in mind, one can use the extensive study and evaluation of data made by Poston et al. (1994) from various sources: *Overseas Chinese Economic Year Book*, 1991, 1992; data from national censuses of different countries; data from *Britannica Book of the Year* and from a special issue of *Revue Européenne des Migrations Internationales*. Figures provided by Poston et al. concerning Hong Kong and Macao are not included in Table 1 because they cannot be considered places of settlement of international Chinese migrants, nor have we included figures for India (130,000) and Turkey (60,000) because these populations are mostly Tibetan refugees (India) and Khazakhs refugees (Turkey). From these numerous and diverse estimates, one can consider the number of ethnic Chinese in the world to be around 30 million (not including Hong Kong and Macao).

Ethnic Chinese are present in nearly every part of the world, ranging from nine individuals in Finland to more than 7 million in Indonesia. One can note the presence of more than 1,000 ethnic Chinese groups in almost 80 countries, exceeding 10,000 individuals in more than 40 countries, 100,000 in nearly 20 countries, and a million individuals in at least five countries (Indonesia, Thailand, Malaysia, Singapore, and the United States). The majority (85 percent) live in Asia, but their presence there is proportionally less important than it was 40 years ago (96 percent). There is now a redeployment of the diaspora, notably towards the non-Asian Pacific (North America, Australia) and towards Europe.

1.2 Main Places of Origin

One can distinguish several categories of places of origin, namely continental China, Hong Kong, and Taiwan and the other zones where a Chinese presence can be found, corresponding to the different poles of the diaspora.

In continental China, the main zones of origin remain the southern provinces: Fujian and Guangdong, to which one can add Zhejiang (north of Fujian). As usual migration relies on already existing migratory networks, which most often turns migrants towards North America, Australia, and Europe. Migration towards Southeast Asia still continues, but is no longer dominant, and migration to the countries of former Indochina had stopped for obvious reasons during the 1970s. These 'proletarian' migrations are often the result of people living in modest conditions who leave to join their compatriots (close and distant relatives, acquaintances), who have already established themselves abroad. For this reason, they are not very different from the labor migration of other popula-

Table 1
Distribution of overseas Chinese (countries of more than 10,000 ethnic Chinese)

ASIA	26,178,000
Indonesia	7,315,000
Thailand	6,000,000
Malaysia	5,471,700
Singapore	2,112,700
Vietnam	2,000,000
Burma	1,500,000
Philippines	820,000
Cambodia	300,000
Laos	160,000
Japan	150,300
Brunei	44,400
North Korea	43,600
South Korea	22,800
AMERICAS	3,226,600
United States	1,645,500
Canada	680,000
Peru	500,000
Brazil	100,000
Panama	100,000
Costa Rica	22,000
Argentina	20,000
Jamaica	20,000
Mexico	20,000
Ecuador	15,000
Venezuela	15,000
Guatemala	14,000
Chile	13,000
Surinam	10,000
EUROPE	769,500
Former USSR	274,000
France	200,000
United Kingdom	125,000
Netherlands	45,500
Germany (west)	29,500
Italy	20,700
Spain	15,000
Belgium	13,000
Sweden	12,000
OCEANIA	373,900
Australia	300,000
New Zealand	35,000
French Polynesia	18,000
AFRICA	108,000
South Africa	36,000
Mauritius	35,000
Madagascar	14,500
Réunion	14,300
TOTAL	30,656,000

tions. Given the more and more restrictive laws of immigration in the receiving countries, part of these migrations are illegal (at least from the point of view of the receiving countries), and produce a lucrative form of trafficking for those who organize them. The cost of such a move can reach the equivalent of US$20,000 to US$25,000.

Migration from Hong Kong appears to be different. It has been caused by the uncertainty concerning the future of Hong Kong and the fears about its rejoining China in July 1997. This migration is either illegal 'proletarian' migration or migration of individuals and families with British nationality who are able to migrate to the countries of the Commonwealth. Otherwise, many entrepreneurs and businessmen are able to immigrate into Canada, for example, because of the financial investments they can make in this country. (The candidates must be able to make an investment of $150,000.) Such migration can also be possible towards Southeast Asia. In fact, the inclusion of Hong Kong into China does not seem to have caused a massive flow of departures, contrary to what was expected.

Migration from Taiwan is not the result of a fear of the future. It is often due to the local *bourgeoisie* having the financial means to be accepted in the receiving countries. Such migration is notably directed towards the west coast of North America, and has produced a strong presence in Los Angeles (Monterey), in San Francisco, and to a more modest extent in Vancouver and in Seattle (Waldinger and Tseng 1992). There is also a notable migration of entrepreneurs to Southeast Asia. In recent years, a number of them have established themselves in Vietnam.

In addition to the migration from the various parts of China, there are the moves, between the various poles of the diaspora. It consists of persons already established (or born) in one of these poles who move to another country. In this context, all combinations between poles are possible. This interpolar migration also includes the flow of people during the 1970s and 1980s, provoked by the wars in Vietnam, Laos, and Cambodia. A large part of the populations with a Chinese origin in these countries were forced to flee under terrible conditions, and they settled in other countries of Southeast Asia or in Europe and North America, where they joined already established compatriots. This exodus, whose size it is difficult to measure because it was not only composed of persons of Chinese origin, may have mobilized three to four million people. It has, thus, deeply modified the geographical distribution in the Chinese diaspora.

Among all this overseas migration, one needs to add a new phenomenon for which there is little hard information, namely Chinese migration towards the territories of the former Soviet Union along the border of China, migration towards Central Asia, Siberia, and the far east of the former Soviet Union. It seems mainly to be a question of business migration, and includes people who have set up small businesses specializing in the distribution of products made in

China. Labor migration to South Korea and a flow of people to Japan should also be mentioned.

2. The Chinese Diaspora

The multipolarity of the migration and the interpolarity of relations are two morphological factors that define the international migration of ethnic Chinese as a diaspora (Ma Mung 1998). The multipolarity of the migration corresponds to the diaspora in the original sense of dispersal (from Greek, *speiro*, to sow), and the interpolarity of the relations refers to the existence of relations between the different settling poles of the Chinese diaspora.

2.1 Multipolarity of the Migration

Since the nineteenth century, Chinese migration has had a pronounced multipolar feature because of its direction towards many territories in Southeast Asia, the American continent, Polynesia, the Indian Ocean, the Caribbean, South Africa and, in a more modest way, in Europe (Thuno 1996). Throughout the twentieth century, a change in the destinations and a reinforcement of the Chinese presence was noticeable. As previously stated, the presence of more than 10,000 overseas Chinese in each of more than 40 countries is recorded, with more than one million people in at least five countries. One of the main features of this distribution is therefore a strong multipolarity of the migration. The other feature of the international Chinese migration is a strong interpolarity of relations.

2.2 Interpolarity of Relations

Interpolar migrations are often linked to major geographical trouble. The wars in Vietnam, Laos, and Cambodia, and the political take-over by totalitarian regimes that followed them, caused the departure of more than three million people, most of whom where first or later generations of Chinese origin. For the most part, they joined already established communities in Europe, North America or Asia by means of diverse itineraries, sometimes passing through various receiving countries. More recently, in Indonesia, where more than 7 million ethnic Chinese are recorded, demonstrations against the government and President Suharto have at times degenerated into anti-Chinese pogroms. During the first half of 1998, some 1,200 individuals of Chinese origin are believed to have been killed by rioters in the city of Jakarta alone, and other deadly anti-Chinese riots took place in other towns, in particular in Medan, in the north of Sumatra. From 100,000 to 125,000 ethnic Chinese are thought to have left Indonesia within 1 year, for the most part heading towards Singapore, Malaysia, Australia, or even Hong Kong (*Strait Times*, Singapore, July 18, 1999), where 38,300 Indonesians of Chinese origin were recorded in February 1999 (*AFP*, Hong Kong, May 21, 1999). More modest flows of people headed for Thailand, the United States, and other destinations where there were overseas Chinese already. Fortunately, all migration is not due to dramatic events, but is rather a search for better opportunities elsewhere. This is the case for the Chinese who settled in Italy (Campani 1994) or even in Spain (Beltran 1997), among whom a good part came from France in the 1980s and 1990s because the French government immigration laws became ever stricter, and because of the relatively greater freedom in the new receiving countries. It is worth noting that these Chinese also came to join small communities that had already been settled for several decades. This also happened between France and the Netherlands during the period between the two world wars (Pieke 1992).

The diversity of the national origins of the Chinese within the receiving countries is a good indicator of the interpolarity of the migrations. In France, and today also in Spain (Beltran 1997) and possibly in Great Britain, more than 20 different national origins are recorded among the ethnic Chinese, with even a greater number in North America: nearly 30 in Montreal alone (Chan 1991) for instance and 40 in the United States.

These interpolar migrations are the basis of relations between the different settling poles. They represent the networks of movement of persons, materials and non-material values (goods, money, information, social norms and codes, cultural models).

The classic example is the establishment of trading networks between China and the different receiving countries of the diaspora. These networks have been strengthened considerably since the opening of the Chinese economy. On a more subtle scale, such migration represents an extension of distributional networks in the various areas of settlement in a given country, as long as the migration spreads. The most interesting point is the growth of trading networks among the different receiving countries. The way a diaspora works, it induces a diversification of the geographical areas of supply and distribution. In grocery stores, to take a simple example, specialized products came from several dozen different countries in addition to continental China, namely from Taiwan, Singapore or Hong Kong, Hawaii, Thailand, Brazil, and even Spain and California. A study made out in an Asian supermarket in Paris in 1998 revealed that the products came from 37 different countries that had an ethnic Chinese population. A similar study from a Chinese supermarket in San Diego (California) in 1999 revealed more than 40 different countries of origin.

The existence of interpolar relations can allow people engaged in business to move from one country to another when working conditions worsen. This explains why Sino-Indonesian capital was particularly withdrawn from Singapore and Thailand in the 1960s and 1970s. This is also the reason why today there is a shift in capital from Southeast Asia to the American west coast in order to take advantage of economic opportunities (financial flow from Taiwan or Singapore) and/or anticipated difficulties (flow from Hong Kong in connection with its retrocession into continental China). Lin (1992) underlines the role of a Chinatown in attracting investors.

2.3 Decisive Economic Effects

Southeast Asia makes up 85 percent of the number of individuals in the Chinese diaspora. The presence of these immigrants has asserted itself in the nineteenth century already, and it has been reinforced constantly ever since. At first, this immigration included labor force for mines and plantations, but it shifted quickly to small shop-keepers, craftsmen, and sometimes to manufacturers and businessmen. Already at the beginning of the twentieth century, the rice business and commercial distribution were operated by these immigrants or their descendants even in the most remote villages. Since then, their presence in economic activities has constantly been reinforced. In order to appreciate the economic weight of the Chinese diaspora in this area of the world, one can note that the ethnic Chinese control 81 percent of the market value of listed companies; in Indonesia, they control 73 percent of such capital (KPMG International, July 1996), and 61 percent in Malaysia. As far as their banking power is concerned, more than 100 banks are entirely or partially controlled by Chinese in the diaspora; 45 of them are in Indonesia, 19 in Malaysia, 14 in Singapore, 14 in Thailand, and 13 in the Philippines (Trolliet 1994). *The Economist* (July 18, 1992) notes that 'Worldwide, the overseas Chinese (including Taiwan and Hong Kong) probably hold liquid assets worth US$1.5–2 trillion.'

Since 1979, the Chinese diaspora has increasingly had important economical relations with continental China. This has contributed in a very decisive way to the spectacular development of this country since the mid-1980s. Numerous authors agree that 70 to 80 percent of foreign investments (US$10 billion in 1995) was due to the Chinese diaspora. They also play a role in the creation of businesses, and in this way contributed to more than 100,000 joint ventures in 1994, totalling US$17 billion (Lever-Tracy et al. 1996).

3. Conclusion

The international migration of ethnic Chinese, which started in the mid-nineteenth century, has evolved into a diaspora-type configuration. Today, this migration constitutes the most important population in a diaspora (30 million people). Networks play an essential role in such a diaspora. Even if today it seems too powerful and rich, one should remember that originally it was composed of labor migration, and that this remains an element to this day. Current migrants are still mainly proletarians.

This migration, like other migrations, form new social groups—the diasporas—settled on transnational territory. A diaspora is based on an identity of ethnic features. This ethnic consideration does not exclude other elements. Studies of the Chinese diaspora constantly remind the investigator of the singularities of the individuals who have different national and/or geographical origins and speak many different languages. Zheng (1995) found that no less than eight languages were spoken among the staff of a famous Chinese restaurant in Paris, French and English included. Therefore, an 'internal' cosmopolitanism is at work even within a single diaspora, which guarantees multiple allegiances across the different countries where the diaspora has been established, as well as with China.

See also: China: Sociocultural Aspects; East Asian Studies: Economics; East Asian Studies: Geography; Migration, Economics of; Migration into the Americas and between American Countries; Migration, Theory of

Bibliography

Beltran A J 1997 Immigrés chinois en Espagne ou citoyens européens? *Revue Européenne des Migrations Internationales* **13**(2): 63–79

Campani G 1994 *L'immigrazione silensioza. Le communità cinesi in Italia*. Edizioni della Fondazione Agnelli, Milan

Chan K B 1991 *Smoke and Fire, The Chinese in Montreal*. Chinese University Press, Hong Kong

Lever-Tracy C, Ip D, Tracy N 1996 *The Chinese Diaspora and Mainland China: An Emerging Economic Synergy*. MacMillan, New York, St Martin Press, London

Lin J 1992 Flux de main d'œuvre et de capitaux chinois vers les Etats-Unis. *Revue Européenne des Migrations Internationales* **8**(3): 73–90

Ma Mung E 1998 Groundlessness and Utopia: The Chinese diaspora and territory. In: Sinn E (ed.) *The Last Half Century of Chinese Overseas*. Hong Kong University Press, Hong Kong

Pan L (ed.) 1999 *The Encyclopedia of Chinese Overseas*. Curzon Press, Richmond, UK

Pieke F 1992 Immigration et entreprenariat, les Chinois aux Pays Bas. *Revue Européenne des Migrations Internationales* **8**(3): 33–50

Poston D, Mao M X, Yu M Y 1994 The global distribution of the overseas Chinese around 1990. *Population and Development Review* **20**(3): 631–45

Thuno M 1996 Chinese emigration to Europe. Combining European and Chinese sources. *Revue Européenne des Migrations Internationales* **12**(2): 275–96
Trolliet P 1994 *La diaspora chinoise*. Presses Universitaires de France, Paris
Waldinger R, Tseng Y 1992 Divergent diasporas: The Chinese communities of New York and Los Angeles compared. *Revue Européenne des Migrations Internationales* **8**(3): 91–116
Wang G 1992a *Community and Nation*. Allen & Unwin, St Leonards, UK
Wang G 1992b *China and the Overseas Chinese*. Times Academic Press, Singapore
Zheng L H 1995 Ressources linguitiques et gestion relationnelle dans la communauté chinoise de Paris. *Revue Européenne des Migrations Internationales* **11**(1): 133–51

E. Ma Mung

International Migration by Ethnic Germans

In the late 1930s and early 1940s, members of ethnic German minorities living in Central and Eastern Europe were in part resettled in Germany at the request of the Nazi government, and in part deported from the European area of the Soviet Union to the Asian area, as a consequence of Stalinist policies. Altogether, this affected at least 2,000,000 people. Between 1945 and 1948 the post-war expulsion of 12,000,000 Germans led to the largest state-organized ethnic cleansing in the history of the twentieth century. German citizens as well as ethnic Germans from other Central and Eastern European states were deported from their traditional areas of settlement, taken mainly to the British, the Soviet and the US occupation zones within Germany. After 1950 the West German government decided to facilitate immigration for members of the remaining ethnic German minorities living in the communist countries of Europe. Officially they were called resettlers (*Aussiedler*). In total, some 4,000,000 people arrived as ethnically privileged migrants to Germany between 1950 and 1999. There are similar immigration regulations for members of co-ethnic diasporas in Israel, in Greece, and in several Central and Eastern European countries. This article analyzes the ethnic German expellees and resettlers, their regions of origin, the most important periods of the partly voluntary, but mostly involuntary, migration, the magnitude of these migration flows, as well as the legal regulations and their political context.

1. *Historical Overview*

From the fourteenth to the eighteenth century hundreds of thousands of German settlers migrated from the southern and western parts of Germany to Central and Eastern Europe. This included in part the conquest and colonization of territories, e.g., by Teutonic Knights. In other cases German settlers were recruited by local rulers. Still others left their homelands for political or religious reasons.

With the founding of nation states on an ethnic basis, the majority of ethnic Germans in Central and Eastern Europe became members of a German diaspora. In 1871, with the creation of the second German Empire, some 12,000,000 German speaking people (both ethnic Germans and Jews) remained citizens of the Habsburg and tsarist empires. After 1918, notable German-speaking minorities lived in France, Belgium, Poland, Czechoslovakia, Hungary, Romania, Yugoslavia, as well as in Italy and in the Soviet Union. Even during the inter-war period the future of the ethnic German diasporas became a key issue of German domestic politics, and later also a foreign policy issue. The German government and the ruling elites promoted two ethnopolitical solutions: the revision of borders in order to include territories with ethnic German populations, and the resettlement of ethnic Germans from their historical areas of settlement, sometimes portrayed as necessary 'evacuation.' Nazi Germany realized both: in 1940 through the annexation of areas of compact German settlement (Austria, parts of Bohemia and Moravia, Danzig, western Poland and Alsace-Lorraine); and between 1939 and 1944 through the 'transfer' of between 625,000 and 650,000 ethnic Germans living outside territories annexed to, or occupied by, Germany in 1938 and 1939. This forced resettlement affected ethnic Germans from the Baltic countries; from Polish territories that later became parts of Ukraine (Volhynia, Bucovina); from Crimea (today also part of Ukraine); from the Caucasus; from Romania (Bessarabia, Dobrudja); from Slovenia (Gottschee); and from Northern Italy (South Tyrol/Alto Adige; see Table 1).

In 1941, some 1,400,000 ethnic Germans living in the Volga region and in other parts of the USSR, alongside many other ethnic groups, became victims of Stalinist policies. They were deported mainly to Siberia, Kazakhstan and Kyrgyzstan because of alleged collaboration with Nazi Germany, or at least sympathies for the German invasion of the USSR. And, unlike other deported groups, ethnic Germans were not allowed to return to their traditional settlement areas in the European area of Russia.

2. *Post-war Expellees and Their Resettlement in the Allied Zones of Occupation*

At the end of World War II, German refugees came in large numbers from what were until then the eastern provinces of Germany and formerly annexed terri-

Table 1
Post-war German expelleees and their resettlement.

Country	Expellees 1944 to 1949 (Total)		Resettled in West Germany	Resettled in East Germany with Berlin	Resettled in Austria and other Western countries	Transferred ethnic Germans 1939 to 1944[a]
	'000s	Percent	'000s	'000s	'000s	'000s
Former German territories	6,980	55.8	4,380	2,600		
Poland (with Danzig)	980	7.8	630	335	15	150
Czechoslovakia	3,000	24.0	1,900	850	250	
Soviet Union (with the Baltic countries)[b]	270	2.2	180	55	10	1,555
Hungary	210	1.7	175	10	25	
Romania	250	2.0	145	60	45	200
Yugoslavia	300	2.4	150	35	115	35
Italy						100
Subtotal	11,990	95.8	7,560	3,945	460	2,040
Quasi-expellees[c]	525	4.2	385	125	15	
Total	12,515	100.0	7,945	4,070	475	2,040

Sources: Benz 1995, Münz 2000, Reichling 1986, Ther 1999
a Between 1939 and 1944 some 625,000 to 650,000 ethnic Germans were transferred to Germany, while another 1,400,000 were internally deported to the Asian part of the USSR. b Between 1939 and 1942 some 130,000 ethnic Germans from the Baltic countries and a relatively small number of Germans from the Crimean peninsula and from the Caucasus (25,000) were transferred to Germany. c Quasi-expellees are persons originating from territories where the expulsion took place, but who themselves in 1945 already lived in Germany (in its current borders) or had become prisoners of war.

tories, to the territory of present day Germany. Between 1945 and 1948, an even larger number of people—12,000,000 in all—were expelled from former German provinces annexed to Poland and the USSR, from other parts of Poland, as well as from Czechoslovakia, Hungary, and Yugoslavia (see Table 1). The victorious Allied powers had allowed Poland and Czechoslovakia to deport all members of their ethnic German minorities. During 1945 and 1946, Hungary and Yugoslavia carried out similar measures. To a larger public this was explained as retaliation and collective punishment for Nazi and war crimes, but the ethnic cleansing was also made to ensure that no relevant groups of people with German citizenship or of German origin would remain in East-Central Europe. The underlying rationale was that this would prevent, once and for all, any further interethnic conflicts or future territorial claims by Germany vis-à-vis its eastern neighbors.

The last all-German census of 1946 listed 5,800,000 expellees in the three western zones of occupation, and 3,600,000 expellees in the Soviet zone. Further expulsions in the years 1947 and 1948 and the migration of expellees from the Soviet zone to the Western zones led to a situation in which some 7,900,000 expellees were already living in West Germany—mostly in the British and American zones by 1950—and another 3,600,000 in East Germany (DDR) (Lemberg and Edding 1959). During the same period Austria absorbed some 430,000 ethnic German expellees (see Table 1).

By the beginning of the 1950s, some 12,000,000 German citizens and persons of German origin from former German provinces and from East-Central Europe had been resettled to East and West Germany as well as to Austria, which was also occupied by the Allies of World War II (see Table 1; see also Benz 1995, Stanek 1985). A further 500,000 people who originally

came from regions with German residents were not expelled, because they had either left the area before 1944 to 1945, were still serving in the German army, or were already prisoners of war in Allied camps in 1944 or 1945. Later these quasi-expellees, for whom return to their old homelands had become impossible, were given the same legal status as those actually expelled, at least in West Germany. This gave them, among other benefits, a legal claim to compensation payments for lost possessions and property (*Lastenausgleich*).

Of the 12,500,000 expellees and quasi-expellees of the years 1945 to 1949, more than half (56 percent) were from former Eastern provinces of Germany (now parts of Poland and Russia); 24 percent from Czechoslovakia as re-established in 1945; and 8 percent from Danzig and territories of inter-war Poland (see Table 1). In relation to population size there were more involuntary migrants in East Germany (DDR)—3,600,000 or almost 19 percent of the total population—than in West Germany (FRG)—7,900,000 or 16 percent of the total population—while the proportion of German expellees in Austria was small by comparison, at 400,000, or 7 percent of the total population.

The influx of German expellees and their integration were in no way free from conflicts, for they were unwanted in many places. In the integration of this group and its later acceptance by German post-war society the recognizably involuntary nature of the expulsion played as big a role as the fact that they were Germans by ethnic origin or by citizenship (Frantzioch 1987).

3. *German Resettlers and the Legal Basis of Their Absorption in (West) Germany*

From 1950 to 1987 the immigration of ethnic Germans from East-Central and Eastern Europe to Germany continued at a low level (see Fig. 1). Until 1987, some 1,400,000 ethnic German immigrants were registered in West Germany. Only thereafter did a clear increase begin. Between 1988 and 1999 a further 2,600,000 arrived. The (West) German Constitution (*Grundgesetz*) of 1949 defined as citizens not only former citizens of Nazi and inter-war Germany but also all ethnic Germans who had found refuge on German territory within the boundaries existing in 1937. Since the 1950s, an extensive interpretation included other ethnic Germans still living in Central and Eastern Europe as potential citizens. The first legal basis of the ethnically privileged immigration of these potential citizens to Germany was the Federal Law concerning Refugees and Expellees (*Bundesflüchtlings- und Vertriebenengesetz*). This Law of 1953 regulated the admission and absorption of former German citizens and ethnic Germans, who, after the completion of the general expulsion measures left those of Germany's (former) Eastern provinces currently under foreign jurisdiction (Danzig, Estonia, Latvia, Lithuania, the Soviet Union, Poland, Czechoslovakia, Hungary, Romania, Bulgaria, Yugoslavia, Albania, or China). The original regulation was aimed at persons whose situation was comparable with that of the expellees of 1945 to 1948. In 1957, ethnic German immigrants were officially called resettlers (*Aussiedler*) for the first time and were given the same status and access to benefits as post-war expellees. In this context the concept of ethnic origin was clarified: 'Somebody is a member of the German nation, who has professed his (her) Germanness in his (her) homeland,' the law stated. Beside this declaration, so-called objective features were brought into play in order to distinguish ethnic Germans from other persons. Ethnic Germans were defined by descent, language and cultural or ethnic orientation. All those fulfilling the criteria and living in the aforementioned countries were (and some still are) entitled to claim German resettler status and to be admitted as privileged immigrants. The publicly-declared moral and political justification for this admission to Germany were repressive measures against members of ethnic German minorities (e.g., deportation, suppression of the German language, political and economic discrimination, etc.) at least in part legitimized as retaliation measures for Nazi and German war crimes. But the fact that, until 1989 and 1990, all ethnic Germans of Central and Eastern Europe lived under communist rule was used as an additional argument in their favor.

Ethnic German immigrants (*Aussiedler*) became German citizens immediately upon arrival in Germany. They could also make use of a series of benefits designed to facilitate their integration: the claim to a council apartment; free German language courses and courses of vocational training, as well as professional retraining; and a claim to transfer payments for which most foreign immigrants would not be eligible—e.g., public pensions even for those ethnic German immigrants who did not contribute to German pension insurance (Heinelt and Lohmann 1992).

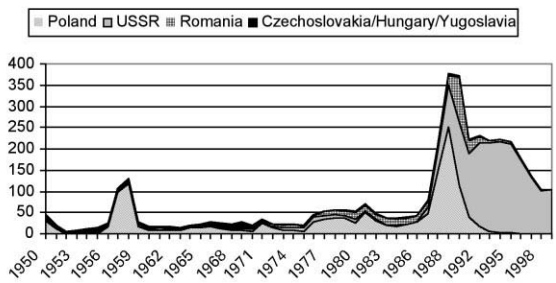

Figure 1
Immigration of ethnic Germans to Germany by country of origin, 1950 to 1999
Source: German Federal Administration Office, Ministry of the Interior

The origins of the considerable material support for ethnic German immigrants lay in the material compensation for persons who had lost or left behind all their property in Germany's former Eastern provinces, originally introduced after 1945. This can hardly justify the benefits accorded to *Aussiedler* of the late twentieth and early twenty-first centuries. They do not come from regions once belonging to pre-war Germany. And since 1990 the post-communist countries of origin no longer force them to give up real estate or other possessions. Now, the existing measures only serve the purpose of integrating the *Aussiedler* into German society as quickly as possible.

Until 1988 the chance to emigrate was offered to ethnic Germans mainly on the basis of bilateral agreements between West Germany and the governments of Poland, Romania, Czechoslovakia, and the Soviet Union. The migration was voluntary to a greater extent than the expulsions of the years 1945 to 1949, and usually only occurred when members of the ethnic German diaspora made a corresponding application. Against this background the analytical distinction between expellees (*Vertriebene*; 1945 to 1949) and resettlers (*Aussiedler*; since 1950) makes sense, although the latter term was only introduced in 1957. Nevertheless, members of the German minorities living in Poland, Romania, and the former USSR were discriminated against in many cases—not only before, but also after 1950. Large parts of the West German public interpreted the *Aussiedlers*' decisions to emigrate both as a response to political and social discrimination at home and as a clear profession of Germanness and belief in the political system of the FRG. Rarely was the emigration of *Aussiedler* primarily seen as an economically-motivated decision (Bethlehem 1982, Delfs 1993).

4. *Areas of Origin and Periods of Immigration*

From the beginning of the 1950s and until the 1980s, Central and Eastern Europe's communist regimes limited the freedom of movement of their ethnic German citizens almost to the same extent as that of citizens belonging to the ethnic majority. The borders were well guarded. Leaving the state without authorization was treated as a crime. Only privileged people received a passport valid for international travel to Western countries. After the end of the organized post-war expulsion there were, for a while, only a few cases of family unification. In 1950, 47,000 ethnic immigrants came to West Germany; in 1952, only 5,000. During the following Cold War period (1953 to 1987) on average 37,000 *Aussiedler* per year came to Germany. The annual fluctuations mirrored, on the one hand, periods of domestic political liberalization or crisis in the countries of origin—during the second half of the 1950s in Poland, 1967 to 1968 in Czechoslovakia, and after 1986 in the USSR—and, on the other hand, the communist regimes of Poland and Romania also used concessions on exit permits for *Aussiedler* as 'leverage' in order to improve their political relations with the FRG and to receive economic or financial aid. At the same time, the political elites of these two sending countries hoped that the still existing ethnic minorities would be weakened and finally disappear through emigration. This all played a role in the mid-1950s when some 250,000 persons of German and mixed ethnic background were allowed to leave Poland in the direction of West Germany. It was also true for the period before and after the conclusion of a Basic Treaty (*Grundlagenvertrag*) between Bonn and Warsaw. In 1970 the number of ethnic Germans who were allowed to leave was almost five times higher than 1968 and 1969 combined. During the 1970s, Poland received substantial financial support in return. In the case of Romania, Germany even agreed on a fee per ethnic German migrant (Münz and Ohliger 1998).

Between 1950 and 1987 almost two-thirds (62 percent) of all ethnic German immigrants came from Poland (848,000) and a further 15 percent from Romania (206,000). Although there also was a strong German minority in the Soviet Union, in that period, only 110,000 persons (8 percent of the *Aussiedler*) were able to emigrate from that country (see Fig. 1). Despite this emigration, the greater part of the still existent German minorities continued to live in their traditional areas of settlement—Upper Silesia, Transylvania, Banat—or in regions into which they were deported during World War II—in particular Siberia, Kazakhstan, Kyrgyzstan. With the fall of the Iron Curtain and the rescinding of administrative travel restrictions in East-Central and Eastern Europe at the end of the 1980s, the number of ethnic German immigrants rose considerably. In 1988, 203,000 *Aussiedler* came to Germany, almost three times as many as in the previous year. And in 1990, the influx of *Aussiedler* reached its highest mark ever, at 397,000 persons. From 1988 to 1999 a total of 2,600,000 *Aussiedler* migrated to Germany. In this phase, the post-Soviet successor states—namely Kazakhstan, Russia and Kyrgyzstan—became the most important regions of origin (1,775,000 *Aussiedler*, or 69 percent). The second largest group was the ethnic German immigrants from Poland (574,000 or 22 percent over the same period). Those from Romania ranged third (219,000 or 8 percent).

Until 1989 and 1990 members of ethnic German minorities from East-Central and Eastern Europe were not able to freely leave their home countries, but were allowed free entry to West Germany, and almost automatically became German citizens. The application for *Aussiedler* status could be made during a short-term visit to Germany, or even after illegal entry. Fast and unbureaucratic naturalization allowed ethnic Germans under such conditions an immediate stay in the country of their forefathers. But with the

fall of the Iron Curtain, Germany became a lot more restrictive. Since 1990 members of ethnic German minorities have had to apply in their country of origin for *Aussiedler* status and for a special entry permit to be issued by the relevant German consulate. The 'Germanness' of the applicants is now checked in advance. These new regulations immediately led to a much smaller influx of ethnic immigrants (221,000 people in 1991), and to a backlog of applications to be processed. In 1992 an annual quota for the admission of *Aussiedler* was established. Until 1999, the size of this quota was 220,000 persons per year, and in 2000 it was reduced to 110,000. On top of this, in 1996, the German administration introduced a language test, in which potential *Aussiedler* have to prove their proficiency in German. This led to a further reduction of the inflow: in 1994, 222,000 *Aussiedler* were given permission to enter Germany, by 1999 the number was only 105,000 (see Fig. 1). Another reason for this reduction is the pace of administrative procedures and individual decisions to actually migrate. In 1999, some 400,000 applications for *Aussiedler* status were pending. Another 180,000 ethnic Germans had already acquired *Aussiedler* status, i.e., the right to immigrate and become a German citizen, but have not exercised this right so far.

5. *Reduction and Foreseeable End to the Immigration of Ethnic Germans*

The immigration of ethnic Germans to Germany has been declining since 1990. There are several reasons for this decline. Since 1993 the unconditional right to immigrate to Germany has existed only for ethnic Germans living in the successor states of the Soviet Union. The regulation by which ethnic Germans must prove their linguistic competence also has a restrictive effect on the influx of *Aussiedler*, because people with poor or no knowledge of German are denied privileged access to Germany. In the years 1996 to 1998, some 50 percent of all applicants passed the mandatory language test. The remaining 50 percent have no right to repeat the test. However, some of those who fail the test later come to Germany as family members (i.e., spouses, parents, children, or children-in-law) of someone who successfully claimed *Aussiedler* status. Unlike those who pass the test, their spouses, their children, the other accompanying family members remain foreign nationals. They are not entitled to a number of integration measures, and they have no immediate access to the German labor market.

An end to the ethnically-privileged immigration into Germany is forseeable. In 1992 the so-called 'Law concerning late consequences of World War II' (*Kriegsfolgenbereinigungsgesetz*) limited the options of future applications for *Aussiedler* status and subsequent entry into Germany to ethnic Germans born before 1993. This regulation will become important after the year 2010, when people born in 1993 and later reach the age of 18. This younger generation of ethnic Germans will have no independent claim to *Aussiedler* status, though some of them will still be able to immigrate legally into Germany within the framework of family unification.

Between 1950 and 1999 a total of 4,000,000 *Aussiedler* came to Germany. Most of them were from the Soviet Union and its successor states (1,800,000) and from Poland (1,400,000). The potential for a further influx of *Aussiedler* is not exactly foreseeable, but it is limited. There are almost no ethnic Germans still living in the successor states of Yugoslavia, in the Czech Republic, and in Slovakia. In Romania, the great wave of emigration (1989 to 1992) reduced the German minority population to a core of older people, no longer willing to emigrate (approximately 60,000 in 1999). Estimates say that there are 500,000 to 800,000 people in Poland who see themselves as ethnic Germans, in particular those living in Silesia. Many of them have already received permission to immigrate as *Aussiedler*, but have not yet exercised this right. Others migrated but did not settle in Germany permanently and returned to Poland as dual citizens. A considerable number successfully regained German citizenship without emigrating. For this reason the number of people in Poland with both German and Polish citizenship has grown (to approximately 275,000 in 1999). In the successor states of the USSR, the remaining number of ethnic Germans and people of mixed origin with at least some German ancestry is estimated at 500,000 to 1,000,000. In the face of the crises of economic transformation as well as political and ethnic conflicts, one can expect that many of them will either have declared themselves as Germans already, or will do so in future, in order to keep open the option of later emigration to Germany or to profit from public funds earmarked by Germany for the support of ethnic Germans in the Eastern diasporas.

See also: Eastern European Studies: History; Eastern European Studies: Politics; Immigration; Immigration and Migration: Cultural Concerns; Internal Migration (Further Types): Industrialized Countries; Migration and Health; Migration, Economics of; Migration out of Europe; Migration: Sociological Aspects; Migration, Theory of; Population Pressure, Resources, and the Environment: Industrialized World

Bibliography

Bade K J (ed.) 1999 *Aussiedler. Deutsche Einwanderer aus Osteuropa*. Universitäts-Verlag Rasch, Osnabrück, Germany

Benz W (ed.) 1995 *Die Vertreibung der Deutschen aus dem Osten*. Fischer Taschenbuch Verlag, Frankfurt am Main, Germany

Bethlehem S 1982 *Heimatvertreibung, DDR-Flucht, Gastarbeiter, Zuwanderung, Wanderungsströme und Wanderungspolitik in der Bundesrepublik Deutschland*. Klett-Cotta, Stuttgart, Germany

Delfs S 1993 Heimatvertriebene, Aussiedler, Spätaussiedler. *Aus Politik und Zeitgeschichte* (Supplement to the weekly newspaper *Das Parlament*), **B48/93**, pp. 3–11
Dietz B, Hilkes P 1994 *Integriert oder isoliert? Zur Situation rußlanddeutscher Aussiedler in der Bundesrepublik Deutschland*. Olzog, Munich, Germany
Frantzioch M 1987 Die Vertriebenen—Hemmnisse, Antriebskräfte und Wege ihrer Integration in der Bundesrepublik Deutschland. *Schriften zur Kultursoziologie 9*. Dietrich Reimer Verlag, Berlin
Heinelt H, Lohmann A 1992 *Immigranten im Wohlfahrtsstaat am Beispiel der Rechtspositionen und Lebensverhältnisse von Aussiedlern*. Leske and Buderich, Opladen, Germany
Lemberg E, Edding F (eds.) 1959 *Die Vertriebenen in Deutschland*. F. Hirt, Kiel, Germany, 3 Vols
Münz R, Ohliger R 1998 Long-distance citizens: Ethnic Germans and their immigration to Germany. In: Schuck P H, Münz R (eds.) *Paths to Inclusion: The Integration of Migrants in the United States and Germany*. Berghahn, Oxford, UK, pp. 155–203
Münz R 2000 Deutschland und die Ost-West-Wanderung. In: Fassmann H, Münz R (eds.) *Ost-West-Wanderungen*. Böhlau Verlag, Köln, Germany, pp. 49–83
Reichling G 1986 *Die deutschen Vertriebenen in Zahlen. Vol. 1, Umsiedler, Verschleppte, Vertriebene, Aussiedler 1940–1985*. Kulturstiftung der Deutschen Vertriebenen, Bonn, Germany
Stanek E 1985 Verfolgt-verjagt-vergessen. *Flüchtlage in Österreich*. Europa-Verlag, Vienna, Austria
Ther P 1999 *Deutsche und polnische Vertriebene*. Vandenhoek und Ruprecht, Göttingen, Germany

R. Münz

International Migration by Ethnic Russians

Throughout most of their history, ethnic Russians have migrated from the Russian heartland to the borderlands, first, of the Russian Empire and, then, of the Soviet Union. More recently, much of their migration has involved 'repatriation' to the heartland and emigration to 'foreign' countries, i.e., to residences beyond the borders of the former Soviet Union. This article examines these migration flows.

1. The Migration of Ethnic Russians Within the Russian Empire and USSR

1.1 The Russian Empire

The expansion of the Russian Empire and development of its huge territory was accompanied by mass migration from the country's ethnic Russian heartland to its non-Russian frontiers. This process began in the seventeenth century and continued for several centuries. As the political boundaries of the state expanded, the territory populated by ethnic Russians expanded as well, beginning with the territory between the Black and the Baltic seas, the Volga region, the Urals, Northern Kazakhstan, Siberia and the Far East, and eventually including Transcaucasia, Kazakhstan, the Baltic region, and also—at the beginning of the twentieth century—Central Asia.

More or less reliable estimates of migration can be made starting from 1719, when Russia began conducting regular counts of the adult population. According to these estimates, during the eighteenth century no less than 1.7 million adults, mostly ethnic Russians, moved from the center of the country to its peripheries. In the first half of the nineteenth century, there were 1.5 million such migrants; in the second half—4 million; and between 1897 and 1916—5.3 million. A significant portion of these migrants moved outside of Russia proper into the borderlands of the enormous empire. In 1719, there were about 100,000 such migrants, by the end of the eighteenth century they numbered over 400,000 (Kabuzan 1996).

The main surge of imperial Russian expansion took place around the end of the nineteenth century. By 1897, there was a tenfold increase in the number of ethnic Russians living in the territory of the future Soviet republics—to 4.5 million. By the eve of the Bolshevik revolution, this number had reached 7.5 million, about 10 percent of all ethnic Russians. Almost half of the Russians outside Russia proper (47.3 percent) lived in the Ukraine, with 19.1 in Belarus, 16.7 in Kazakhstan, 5.8 in the Baltic countries, 3.5 in Central Asia, and 1.7 percent in Moldova (Kabuzan 1996).

The migration of ethnic Russians before the Revolution was, generally, of a colonial nature and often followed upon wars of conquest. Migration was the imperial regime's main method of settling and developing acquired land, of controlling and defending its borders, and of ensuring stability. At the same time, in many border regions, such as the Caucasus and Central Asia, the Russians became emissaries of urban culture and education.

Despite sizable migration, the share of ethnic Russians in the population of the border regions in 1917 was not very large. It was highest (21.7 percent) in the territory of what is now Kazakhstan, especially in its northern areas adjacent to Russia proper. Ethnic Russians constituted 14.5 percent of the population in Belarus, 9.6 in Latvia, 9.3 in Ukraine, 8.4 in Azerbaijan, and 2.1 percent in Uzbekistan.

1.2 The USSR

World War I and the subsequent October revolution and civil war led to a sharp reversal in the migration of Russians from center to periphery. In 1926, 25 percent fewer Russians lived in the non- Russian territories of the former Russian Empire than had lived there in

1917—only 5.8 million. Most of those concerned resettled in Russia proper, although some joined the sizable group of Russians who left the country altogether. In the late 1920s and early 1930s, however, the historic process of centrifugal migration resumed at a much accelerated pace.

Centrifugal migration in the Soviet period was spurred by the regime's commitment to rapid industrialization and by its determination to russify sensitive border regions, where native populations were sometimes deported en masse and replaced by ethnic Russians. This process began along the country's western and southern borders, and accelerated after 1939–40, when the USSR annexed the Baltic States and parts of Poland and Romania (Polyan 1996). As a result, by 1939 almost 11 million ethnic Russians lived in border regions of the USSR—one and a half times more than had lived there in 1916.

During World War II, there was a massive evacuation of population and of industrial production from central Russia to, and beyond, the Ural mountains. This gave another push to ethnic Russian migration, especially to Central Asia and Kazakhstan. After the war, the economic reconstruction in the western part of the USSR was accompanied by further mass migration by ethnic Russians, especially into the Ukraine, while the Sovietization of the reconquered Baltic states gave rise to wholesale 'ethnic cleansing' and Russification. And, somewhat later, the campaign to develop the so-called virgin lands of Kazakhstan was accompanied by another large influx of ethnic Russians. As a result, the number of Russians living in the non Russian republics of the USSR rose from 10.7 to 16.2 million between 1939 and 1959. In the same period, the ethnic Russian population of Russia proper increased by 8.8 percent.

The large-scale migration of ethnic Russians into the non-Russian republics of the USSR continued into the 1960s and was linked closely with rapid urbanization. Most migrants moved to areas where cities were growing fast but the local rural population was not yet ready to live in them. This was the case in the Transcaucasia, Central Asia, Kazakhstan, and Moldova in particular. In the Baltic states, the influx of Russians had less to do with the unpreparedness of the local population for urban and industrial life than with low population growth rates and ethnic cleansing. The migration of Russians to non-Russian territory was also facilitated by the low standard of living in Russia's villages, which is where most migrants came from.

In the later 1960s, the inflow of ethnic Russians into the non Russian republics of the USSR began to slow down and stopped altogether by the 1980s, at which time some 24 million ethnic Russians, almost 17 percent of the entire ethnic Russian population, lived outside of Russia proper. In part, this was a reflection of the virtual disappearance of 'surplus labor' in Russian rural areas as well as of the improvement of educational and employment opportunities in Russian cities. At the same time, in Central Asia and the Transcaucasia, a demographic explosion and accompanying large increases in native workforce entrants led to an intense competition for urban jobs—jobs that non-Russian natives were now psychologically and educationally ready to perform and ethnically self-conscious and self-assertive enough to demand. The magnitude of the resulting change is indicated by the increase in the percentage of non-Russians in the industrial workforce between 1967 and 1987: from 43 to 61 percent in Uzbekistan, from 26 to 41 percent in Kirghizia, from 35 to 54 percent in Tadjikistan, and from 38 to 59 percent in Turkmenistan. Similar changes took place in Azerbaijan, Georgia, Moldova, and Kazakhstan. Such changes were even larger in the case of white-collar workers and professionals.

As the 'indigenization' of the workforce in the non Russian republics of the USSR spread and accelerated, and for largely the same reasons, Russian residents of those republics began to migrate (or re-migrate) to Russia. Beginning in Georgia and Azerbaijan in the 1960s, this trend became evident in Central Asia in the 1970s, and affected most of the non- Russian republics by the 1980s. The only exceptions were Ukraine, Belarus, Moldova, and the Baltic states, where it began only in the 1990s. Claims that the collapse of the Soviet Union marked an abrupt turning point in Russia's migratory history are incorrect. This notwithstanding, the collapse did have an enormous and probably enduring impact on Russian migratory behavior.

2. *Migration to Russia From the Other Newly Independent States*

With the collapse of the USSR, the 'repatriation' of ethnic Russians ceased to be a largely voluntary process and assumed a forced character, driven by ethnic discrimination, human rights abuses, and communal violence. Many of the 'repatriates' were responding to newly imposed limitations on their civil and political rights, restrictions on use of the Russian language, displacement from administrative jobs and intellectual occupations, and curbs on their private sector activities. In many cases, they were also 'displaced persons,' fleeing from armed conflicts.

In the Ukraine and Belarus, where discrimination, coercion, and violence were absent, the overwhelming majority (well over 95 percent) of resident Russians stayed put, and those who migrated to Russia did so solely for economic reasons, as did Ukrainians and Belarussians who migrated with them. In the other newly independent states of the former Soviet Union, however, the outmigration of Russians assumed mass proportions. Thus, between 1990 and 1998 the number of Russians in Armenia and Tajikistan declined by over 50 percent. In Azerbaijan and Georgia the decline

was only slightly smaller, while the decline in the Central Asian states (Kazakhstan excluded) and the Baltic states was 25 and 14 percent, respectively. In general, the Russian Diaspora in non-Slavic countries of the former Soviet Union has fallen by 20 percent as a result of repatriation. Furthermore, although there has been some slowdown in the process since 1995, repatriation will almost certainly increase further, especially if and as Russia's economy recovers and begins to grow.

3. *Russian Emigration Beyond the Borders of the Soviet Union and the Newly Independent States*

3.1 *Pre-Soviet and Soviet Emigration*

Until the end of the nineteenth century, very few ethnic Russians settled beyond the borders of the Russian Empire. In the eighteenth and nineteenth centuries, small groups of religious sectarians resettled in Romania (Old Believers) and Canada (Dukhobors), but Poland was the only country with a sizeable Russian diaspora (about 300,000 people in 1900).

Early in the twentieth century, the flow of people migrating from Russia increased substantially. Four million people emigrated to the USA between 1901 and 1920, but the vast majority of them were Jews. According to the US census of 1910, there were 65,600 non-Jews among residents 'of recent Russian origin.'

The first large wave of ethnic Russian emigration was generated by the Bolshevik revolution and attendant civil war and Red terror. In the first years after the revolution, between 1.5 and 2 million people left the country (Polyakov 1986; Maksudov 1989). This was the so-called 'White emigration,' consisting mostly of ethnic Russians. The main stream of White emigration initially went to Poland (one million), Germany (560,000), France (175,000), and Latvia (12,000), and then on to the USA (Kulischer and Kulischer 1948). At the same time, over three million former prisoners of war returned to Russia from Germany and Austria, mostly ethnic Russians (Volkov 1930, Maksudov 1989).

World War II brought about further mass migrations of ethnic Russians. Thus, of the four million people who were sent to German labor camps from the USSR, 1.2 million were ethnic Russians. Russians also comprised the majority of the millions of Soviets in German prisoner of war camps. Of these slave laborers and prisoners, some 31,000 ethnic Russians stayed in the West after the war (Polyan 1996).

Between 1947 and 1957, they were joined in the West (chiefly Canada, the USA, and Australia) by another 128,000 emigrants from the USSR, almost all of whom were Russians or Ukrainians who were allowed to emigrate on grounds of family reunification (Mariansky 1969). Except for these people, however, almost no one was allowed to emigrate from the Soviet Union from the second half of the 1920s until the years immediately preceding the disintegration of the union.

3.2 *Post-Soviet Emigration*

Despite the complete abolition of barriers to emigration following the collapse of the Soviet Union, ethnic Russians have remained largely at (or returned) home. Between 1992 and 1998, 178,500 ethnic Russians left Russia as legal emigrants. This was 29.1 percent of the total number of documented emigrants from Russia, the majority of whom were Jews or Volga Germans. There is probably a higher percentage of ethnic Russians among undocumented emigrants from Russia, but there is no reliable information available either on the number of undocumented emigrants or their ethnic composition. What is clear is that the proportion of highly educated specialists among Russian emigrants is significantly larger than the proportion of such specialists in Russia's overall population.

Today, around 1.5 million people who consider themselves Russian live outside the borders of the former Soviet Union, including over one million in the USA, 90,000 in Canada, 80,000 in Brazil, 50,000 in Argentina, 40,000 in France, 35,000 in Romania, and 20,000 in Austria (Bruk 1986, Kabuzan 1996). Although the number of such emigrants may grow over time, it is unlikely to grow as rapidly or reach the magnitude that many in both Russia and the West once forecast and (for various reasons) feared. This is both because Russians do not seem particularly interested in emigrating and because their neighbors have created strong barriers to their doing so.

See also: Economic Transformation: From Central Planning to Market Economy; Immigration; Immigration and Migration: Cultural Concerns; Immigration: Consequences for Fiscal Developments in the Receiving Population; Internal Migration (Further Types): Industrialized Countries; Internal Migration (Rural–Urban): Industrialized Countries; International Migration by Ethnic Germans; Migration and Health; Migration, Economics of; Migration out of Europe; Migration: Sociological Aspects; Migration, Theory of; Population Pressure, Resources, and the Environment: Industrialized World; Socialism: Historical Aspects; Soviet Studies: Geography

Bibliography

Anderson B, Silver B 1989 Demographic sources of the changing ethnic composition of the Soviet Union. *Population and Development Review*, Vol. 15, issue 4, pp. 609–56

Bruk S I 1986 *Nasyelyeniye mira*. [English translation *World Population*], Nauka Publishers, Moscow, Russia, p. 825
Byerzina M Ya 1971 *Formirovaniye etnichyeskovo sostava nasyelenya kanady*. [English translation *The Formation of Canada's Ethnic Composition*] Nauka Publishers, Moscow, Russia, p. 194
Kabuzan V 1996 *Russkiye v mirye*. [English translation *Russians in the World*] Russian Baltic Information Center "Blitz", St. Petersburg, Russia, p. 352
Kulischer A, Kulischer F 1948 *Europe on the Move: War and Population Changes*. Columbia University Press, New York, p. 377
Maksudov S 1989 *Potyeri naseleniya v USSR*. [English translation *Population Losses in the USSR*] Chalidze Publisher, Vermont, USA, p. 294
Mariansky A 1969 *Sovryemyenniye migratsii nasyelyeniya*. [English translation *Contemporary Population Migration*] "Statistika", Moscow, Russia, p. 224
Polyakov Yu A 1986 *Sovyetskaya strana poslye okonchaniya grazhdanskoi voiny: territoriya i nasleniye*. [English translation *The Land of the Soviets After the Civil War: Territory and Population*] Nauka Publishers, Moscow, Russia, p. 272
Polyan P M 1996 *Zhertvy dvuk diktatur*. [English translation *Victims of Two Dictatorships*] "Vash Vybar", Moscow, Russia, p. 442
Volkov Ye Z 1930 *Dinamika narodonasyelyeniya Rossii za* 80 *lyet*. [English translation *The Dynamics of the Population of Russia over Eighty Years*] Moscow–Leningrad State Publishing, Moscow, Russia, p. 272
Wilcox W E 1929 *International Migrations*. National Bureau of Economic Research, New York, pp. 14, 18, 1086

Z. A. Zaionchkovskaya and J. R. Azrael

International Migration by Ethnic South and Southeast Asians

1. Introduction

South (including the countries of India, Pakistan, Bangladesh, and Sri Lanka) and Southeast Asia (including the countries of Indonesia, Philippines, Thailand, Vietnam, Singapore, Malaysia, Brunei, Cambodia, Laos, and Myanmar) are home to 31 percent of the world's population at the turn of the twentieth century (1,879.4 million persons) and the region has experienced massive population change since the 1980s (Leete and Alam (1993). The rapid increase in population movement between countries that has accompanied the process of globalization in this period in all of the countries of this region has been affected significantly. There has been a major increase in the levels, complexity, and impact of international migration in the region. Whereas at the end of the 1970s such movement was limited to a selected small proportion of residents in the region, it is now within the calculus of choice of a major part of the population as they weigh up their life chances. With the increase in movement has come an increased awareness of it and intervention to influence it by governments in the region. This article seeks to summarize the evolving patterns of movement and address some of the major issues relating to it.

2. Patterns

The data available to examine international migration in South and Southeast Asia are limited in the extreme. The increase in significance and intensity of movement has outpaced the development of data systems to measure it. Moreover, undocumented migration in the region is arguably as great in scale as legal migration and, by definition, the information relating to that migration is limited. While there is a great deal of complexity in the spatial patterning of the movement, the degree of permanency, the characteristics of movers, and its legality, it is possible to identify three main subsystems of movement in the region. These include a so-called south–north, more-or-less permanent flow to developed countries, a predominantly temporary labor flow to the Middle East, and a third set of interactions within the region. Each subsystem involves different balances of forced and nonforced movement, males and females, permanent and nonpermanent, documented and undocumented, labor related and other movement, etc.

South–north movement is one of the most dominant global international migration flows (UN 1997) and South and Southeast Asia is one of the dominant origin areas of such flows. The largest nation in the region, India, is estimated to have a diaspora of 15 million persons living overseas with US $150 billion in assets (*Migration News*, November 30, 1998). There are said to be 3.2 million overseas Pakistanis (*Asian Migration News*, June 15, 1999). Table 1 presents estimates of the numbers of South and Southeast Asian migrants in more developed countries. These are underestimates of the impact of immigration because the data are obviously incomplete; undocumented migrants are excluded and the children born overseas to migrants and later generations are not included. Nevertheless, the scale of the movement can be appreciated.

However, more-or-less permanent settlement in North countries is not the only type of movement out of South and Southeast Asia. International labor migration whereby workers from the region move to other countries to work, usually under contract, for a set period (usually around two years) before returning to their home country has expanded since the 1980s. One of the major streams of this type out of the region is directed toward the Middle East. There has been a long tradition of South Asians working in that region which extends back to colonial times, but the steep increase in oil prices after 1973 saw massive flows of foreign exchange into the region which fostered massive infrastructure construction which in turn saw the expansion of recruitment of foreign workers

Table 1
Reported numbers of Southeast and South Asia-born populations in 'North Countries'

Origin	USA (1998)	Canada (1996)	Australia (1999)	Japan (1997)	New Zealand (1996)	Europe[2]
Southeast Asia						
Myanmar	17,210		10,123[3]		513	
Thailand	159,555		18,936		3,348	
Laos	198,193		9,900		1,008	
Cambodia	112,597	17,965[4]	21,542		3,678	
Vietnam	918,114	139,300	175,200	11,900	3,465	192,442
Philippines	1,296,803	184,550	116,900	93,300	7,002	102,192
Malaysia	50,475	16,100[4]	94,800		11,889	
Singapore	24,010	6,285[4]	34,900		3,477	
Indonesia	64,376		65,800	11,900	2,715	22,500
Total	2,841,333	364,200	548,101	117,100	37,095	317,134
South Asia						
India	756,911	235,930	100,700	7,500	12,807	192,748
Pakistan	605,000		8,354[3]	5,600	837	932,568
Bangladesh	121,089		5,063[3]	6,100	1,212	80,600
Sri Lanka			56,400	2,900[3]	4,020	147,700
Nepal			1,483[3]			
Total	1,483,000	235,930	172,000	22,100	18,876	1,353,616

Sources : April 1998 US Census Bureau Current Population Survey Data, New Zealand 1996 Census, ABS 2000, ABS 1996 Census, Statistics Canada 1996 Census 20 percent Sample Data, Statistics Canada 1992, Garson 2000, *Asian Migration News*, June 15, 1999
1 Includes Brunei
2 Includes only figures updated to 1998 SOPEMI (OECD 1998)
3 1996 Census data
4 1991 Census data

Table 2
Estimates of migrant workers from South and Southeast Asia deployed to the Middle East

Country of Origin	Saudi Arabia	Other	Total	Year	Source
Southeast Asia					
Thailand	1,561	16,567	18,128	1998	Chalomuong 2000
Philippines	650–700,000	300–350,000	1 million	1999	*Migration News* July 1999, December 1999
Indonesia	425,000	19,630 (UAE)		1997	Indonesian Embassy Riyadh, *Migration News* December 1999
South Asia					
India	50–60,000	100,000 (Bahrain)	–	1996/7	*Migration News* December 1996, July 1997
Pakistan			1,552,330	1999	Scalabrini Migration Centre 2000
Bangladesh			800,000	1996	*Migration News* December 1996
Sri Lanka			770,000	1999	Scalabrini Migration Centre
Nepal			1,400,000	1992	Asian Migrant Centre 1999

initially from South Asia but later from Southeast Asia.

In the late 1980s this movement changed to include more women and service workers. Again, data are poor since many Middle Eastern countries do not release information on migrant workers due to the sensitivity attached to heavy reliance on foreign labor in several countries where these workers in some cases outnumber local workers. Nevertheless, Table 2 presents some information for the main labor-exporting countries. The bulk of migrant workers from Southeast Asia are unskilled while South Asia has sent both skilled and unskilled workers. Female domestic servants have become more significant in the movement over time, especially those from Indonesia, the Philippines, and Sri Lanka.

In the 1980s and 1990s there was a growing tempo of movement of international labor migrants to

Table 3
South and Southeast Asian countries: estimates of migrant workers in other countries

Origin Countries	Number	Main Destinations	Source of Information	Year
Southeast Asia				
Burma/Myanmar	1,033,111	Thailand	Asian Migrant Centre 1999	1998
Thailand	382,825[a]	Middle East, Taiwan, Myanmar, Singapore, Brunei, Malaysia	Asian Migrant Centre 1999, Scalabrini Migration Centre 1999	1998
Laos	100,000[b]	Thailand	Asian Migrant Centre 1999, Scalabrini Migration Centre 1999	1998
Cambodia	200,000	Malaysia, Thailand	Scalabrini Migration Centre 2000	1999
Vietnam	36,625	Korea, Japan, Libya, Czech Republic, Saudi Arabia, Yemen	Asian Migrant Centre 1999, *Migration News* November 1997	1997
Philippines	7,000,000	Middle East, Malaysia, Thailand, Korea, Hong Kong, Taiwan	*Migration News* June 1998, July 1999, July 2000	1999
Malaysia	250,000	Japan, Taiwan	Asian Migrant Centre 1999	1995
Singapore	5,000[a]		Asian Migrant Centre 1999	1998
Indonesia	1,600,000[a]	Malaysia, Saudi Arabia, Taiwan, Singapore, South Korea, United Arab Emirates	*Migration News* June 2000 and December 1997	2000
Total	10,607,561			
South Asia				
India	285,000	Middle East	*Migration News* March 2000, Scalabrini Migration Centre 2000	2000
Pakistan	3,180,973	Middle East, Malaysia	Scalabrini Migration Centre 2000	1999
Bangladesh	3,000,000	Saudi Arabia, Malaysia	Salabrini Migration Centre 2000	1999
Sri Lanka	1,200,000	Middle East, Malaysia	*Migration News* January 1999, Scalabrini Migration Centre 2000	1999
Nepal	1,400,000	Middle East, India	Asian Migrant Centre 1999	1992
Total	9,065,973			

a Documented
b Undocumented

Asian destinations, including other countries in the Southeast Asian region. This is evident in Table 3, which consolidates estimates of the stocks of overseas migrant workers from the various nations of South and Southeast Asia. These indicate that for many the main destination is an Asian nation. In fact, several of the East and Southeast Asian nations which began their transitions to low fertility in the 1980s and 1970s have suffered significant labor shortages in recent years. This has been a function of both rapid economic expansion and the declining numbers of new entrants into their workforces due to low fertility. This has been accompanied by a distinct segmentation of their labor forces whereby local workers have eschewed entry into low-income, low-status jobs which have created niches for overseas migrant workers in these areas. Hence, particular industries in particular Asian countries have become quite dependent on foreign labor. These include some activities in countries like Thailand and Malaysia which themselves are supplying migrant workers to Japan and Taiwan. Hence, in Thailand the rice milling and fishing industries are heavily reliant on labor from Myanmar while in Malaysia the plantation and forestry industries rely upon gaining workers from Indonesia, the Philippines, and Bangladesh.

It is apparent, then, that one can recognize three distinct subsystems in South and Southeast Asian migration. One has involved movement to the so-called northern nations; another links the region to the Middle East; while a third involves intra-Asia movement. While permanent settlement in the case of the first subsystem and labor migration in the case of the second two are dominant, there are a number of other forms of international movement involved in the subsystems. Dealing first with the south–north movement, a most important factor was the abolition of racially biased immigrant selection policies in the main immigration countries (USA, Canada, Australia, etc.) in the 1960s and 1970s. Immigration policies were developed in which selection was based upon skill/education and/or family linkages as well as acceptance of refugee/humanitarian migrants. Hitherto, migrants from South and Southeast Asia could not compete on an equal basis for migrant places with people from Europe. In addition, migration flowed along linkages established by colonial rule (e.g., Indians and

Table 4
South and Southeast Asia: refugees by country of asylum/origin 1999

	Number ('000)	
	Destination	Origin
Southeast Asia		
Brunei		
Cambodia	–	15.1
East Timor		162.5
Indonesia	162.5	–
Laos	–	0.1
Malaysia	50.5	
Burma/Myanmar	–	121.8
Philippines	0.2	45.1
Thailand	100.1	
Vietnam	15.0	293.2
South Asia		
Afghanistan	–	2,540.2
Bangladesh	22.2	
Bhutan	–	107.6
India	180.0	–
Nepal	127.9	–
Pakistan	1202.0	–
Sri Lanka	–	66.4

Source: UNHCR 2000

Pakistanis to the UK, Indonesians to The Netherlands, etc.) or by military involvement (Koreans to the USA, etc.).

An important element in the south–north movement, as well as migration within Asia, has been that of the refugees. Table 4 shows the current stock of refugees in countries in the region. While there have been several such flows since the 1950s, two have been on a particularly large scale and have differed substantially in type. The outflow of refugees from Indochina following reunification of Vietnam in 1975 was predominantly initially to other Asian (especially Southeast Asian) nations.

This involved more than two million people, the vast majority of whom have been settled in third countries (predominantly in the 'north') or repatriated back to their home countries. These movements have resulted in substantial subsequent migration since these refugee settlers functioned as anchors in destinations to bring other family members to these destinations, often using the family reunification components of migration programs.

The second major refugee migration occurred from Afghanistan, especially into the neighboring nations of Pakistan and Iran. This involved several million people, most of whom were not settled in third countries or repatriated but stayed for long periods in their countries of temporary asylum. There have been many smaller but nevertheless significant movements, mostly involving movement into neighboring countries—Myanmar into Thailand, Tibet into India, Bangladesh into India, Philippines into Malaysia, etc.

Table 5
Percentage of foreign students from Asia studying in tertiary institutions in some north countries

Country	Year	Total number of foreign students	Percentage from Asia
USA	1995/96	453,785	64.1
Australia	1999	83,111	85.8
Japan	1994/95	53,511	92.1
New Zealand	1997	6,415	70.8

Source: LINESCO 1999; DETYA 1999

Another strand in the south–north migration has been the acceleration in the migration of students to undertake secondary but especially tertiary education. Table 5, for example, provides data on the percentage of foreign students from Asia studying in tertiary institutions in some north countries. In Australia, for example, over the 1983–1999 period the number of overseas students increased from 9,098 to 83,111. There is a strong connection between such movement and eventual settlement in the destination nation. Indeed, gaining an educational qualification in the 'north' country can facilitate entry as a permanent settler.

A related element in the intra-Asian subsystem involves the movement of trainees from labor surplus nations like Indonesia, Thailand, and the Philippines to countries with shortages such as Japan and Korea. The movement ostensibly is for workers of multinational companies in the former countries to spend a period in the home country of the companies to gain training. There has been discussion of the extent to which companies are using this system to import unskilled or semiskilled workers when government policy prevents such movement.

While the net effect of the south–north movement is emphatically in favor of the north, the flow in the opposite direction is not to be ignored. This in part involves former emigrants returning to their home country but in the case of some Southeast Asian nations also involves natives of north countries to meet the demands for skilled workers in areas like accountancy, engineering, management, etc. which cannot be met locally. Most of the latter will remain only temporarily in the Asian country. The fact that there is return migration into Asian countries from the north, however, is often neglected. It often can take the form of a 'reverse brain drain' involving highly qualified people who return and take up important positions in the economy or government of their home country.

Obtaining estimates of international migration in the South and Southeast Asian regions is very difficult. Data collection systems generally are not well de-

veloped and there is a high level of undocumented migration. The highly complex nature of the movement needs to be stressed as does the fact that there are complex linkages between the different types of movement.

3. The Migrants

Migration is always a selective process in that the groups of people who move differ from those who are left behind or those which they join at the destination. Each of the three main subsystems briefly described above is distinctive not just in terms of the countries involved but also the types of people who move. The south–north migration generally can be characterized as being selective of the better-educated and skilled populations since entry criteria in the destination countries usually require this, although 'business' migration programs involving the investment of minimum levels of personal wealth at the destination are increasingly common. Under family reunification and refugee/humanitarian programs and the United States Quota Program, there are some, but limited, opportunities for unskilled people to move but the overall selectivity remains. One characteristic of the south–north movement is the increasing involvement of women such that in several north countries women immigrants from the region (especially Southeast Asia) outnumber their male counterparts.

The composition of the flows in the Asian and Middle East subsystems is quite different. While there are certainly exceptions (these include some of the flows out of South Asia and the Philippines), the outmovements are dominated by unskilled workers. These are directed at particular labor market niches at the destination which, due to low wages or low status, are unable to attract local workers. Several of these flows are selective of women so that in some sending countries (e.g., Sri Lanka, Indonesia, and the Philippines), women outnumber men among the outgoing migrant workers. These women especially are directed to a number of specific occupations which sometimes make them vulnerable to exploitation. These especially include work as domestics in private homes in destination countries but also in some cases are associated with the so-called entertainment industry. In some countries (e.g., the Philippines) it has been found that migrant workers sometimes take up jobs in the destination country which are at a level below their qualifications.

4. The Process

There can be no doubt that the level of international population movement in South and Southeast Asia has increased substantially since the 1980s. Whereas in the 1970s the possibility of international migration was only within the reach of relatively few people in the region, it is now within the calculus of choice of a much larger number of people and a wider range of socioeconomic, cultural, and gender groups. Clearly, the whole process of globalization, especially as it has impinged upon labor markets, has been an important factor. This has increased greatly the linkages between nations involving flows of goods, finance, ideas, etc., and it is inevitable that movements of people also have increased although the barriers erected by nation-states to flows of people have been reduced.

Neo-classical economics explanations have relevance for migration in the region. South and Southeast Asia contain four of the 10 largest labor surplus countries in the world. The demographic and economic gradients between richer and poorer countries both globally and within Asia have increased since the 1980s so that there are not only steep differences in income and wages between nations but also in the rate of increase of the working age populations. Labor-importing economies are undergoing economic transformation from manufacturing to service-based economies and this is creating jobs for migrant workers. Labor shortages in areas like information technology, management, and a number of skilled areas is encouraging the development of international labor markets but, also, the labor market segmentation referred to earlier is creating opportunities for migrant workers in low-skilled, low-paid, and low-status areas. The South and Southeast Asian region is able to supply workers both in the skilled areas (from South Asia, the Philippines) and the unskilled occupations.

Nevertheless the widening gaps between nations in population and economic growth and income and wage differentials are not sufficient explanation of the burgeoning of migration in the region, although they are an important component of any comprehensive accounting for it. There are a number of other elements which are also involved. Among the most important is the development of social networks between origin and destination nations. Empirical research indicates that the majority of movers from the region migrate to join family and friends who migrated earlier to the destination. A large proportion of South and Southeast Asians now have social capital in countries other than their own in the form of a family member or friend who lives and/or works in those countries. Those linkages form important conduits along which international migration, both documented and undocumented, takes place. The development of ever cheaper and better transport and communication systems enhances the operation of these networks in initiating migration.

Another element involves the so-called migration industry, which has a long history in the region but has proliferated in recent times. This involves a plethora of immigration agents, recruiters, travel agents, travel providers, immigration officials, labor brokers, and

other intermediaries who are involved in initiating and facilitating international migration. Part of this infrastructure are the people smugglers and traffickers who transport migrants outside of government regulations but it also involves many who operate within the law.

Another element is the greater involvement of government in international migration in the region. While governments in destination countries are predominantly concerned with controlling the numbers and characteristics of immigrants, those in sending countries increasingly are becoming involved in the emigration process. Realization that international migration can relieve under- and unemployment at home and enhance foreign exchange earnings has seen governments in several nations intervene to encourage emigration and to protect the rights of their nationals living and/or working in other countries. Some countries have been active in encouraging former migrants with needed skills to return while others have policies and programs to encourage their nationals in other countries to invest their earnings in their home countries.

5. *Issues*

The proliferation of international migration in South and Southeast Asia has seen the emergence of a number of significant issues both in the countries of origin and destination as well as among the migrants themselves and their families. One of the most debated of these relates to remittances—those moneys earned by migrants abroad that are sent to or brought back to the country of origin. There is debate as to the scale and impact of these financial flows but the data relating to them are largely incomplete. This is because much of the flow does not pass through official channels. Nevertheless, it is clear that the impact of remittances can be considerable. Table 6, for example, shows the significance of reported remittances in relation to foreign exchange earnings from trade in three Southeast Asian countries. While global figures such as those in Table 6 understate the scale and impact of remittances nationally, there can be no doubt that their impact within regions of South and Southeast Asian countries is much greater. This is because migrants tend to be drawn from particular *regions* rather than randomly across the nation. Hence, remittances tend to be focused on particular parts of countries such as Kerala in India or Flores in Indonesia. Moreover, these regions are often among the poorest in the origin countries and the remittances represent one of the few sources of investment in those regions. For example, a December 1999 report from the State Planning Committee of Laos indicates that remittances from abroad represented 28 percent of all household earnings in the Vientiare Valley compared with 25 percent from agriculture, 22 percent from wages, and 18 percent from business (Litner 2000, p. 48). However, remittances have also been seen to increase costs and widen income distribution in origin areas as well. There have been several interventions by governments in the region to encourage nationals overseas to put their earnings in national banks and to invest in their home countries as well as to encourage the sending of remittances to the home country.

Another important issue relates to the exploitation of migrants from the region. While it is incorrect to depict all such migrants as victims subject to exploitation, there can be no doubt that some groups are vulnerable to exploitation of a number of kinds. This is particularly the case among some groups of female migrants, especially those who work in occupations where protection is difficult, such as domestic workers in private homes and in the entertainment industry. Those who are undocumented are especially vulnerable to exploitation but lack of knowledge and local languages and customs can also increase vulnerability

Table 6
Main Southeast Asian labor exporting countries: workers' remittances relative to exports and imports in US$ million, 1980–98

			Total Merchandise		R	
Country	Year	Workers Remittances	Exports (X)	Imports (M)	X	M
Indonesia	1980	33	21,908	10,834	0.1	0.3
	1992	264	33,815	27,280	0.8	1.0
	1997	1,261	63,238	62,830	2.0	2.0
	1998	1,252	48,848	27,337	2.6	4.6
Philippines	1980	421	5,744	8,295	7.3	5.0
	1992	2,222	9,790	15,465	22.7	14.4
	1997	5,742	40,365	50,477	14.2	11.4
Thailand	1979	191	5,240	7,158	3.6	2.7
	1992	1,500	32,473	40,466	4.6	3.7
	1997	1,306	72,415	72,437	1.8	1.8

Source: Hugo 1996, p. 8, Battistella and Asis 1999, *Asian Migration News* June 30, 1999, World Bank 2000, Nazara 2000, p. 12; BPS

of immigrants. Governments of sending countries vary considerably in the efforts they have made to protect their nationals abroad. The Philippines, for example, leads the way in this area but even so significant exploitation still occurs. The issues of protection of nationals overseas is most pressing but governments find it difficult to interfere in the affairs of other countries. Accordingly, some of the most effective protection comes from the activities of nongovernment organizations and through informal networks established by the migrants. Exploitation of the migrants can occur in the process of movement when they may be forced to pay exorbitant amounts of money or suffer physical abuse. There have been several documented cases of undocumented migrants from the region losing their lives due to suffocation or drowning in the process of migration. At destinations they can be subject to exploitation and discrimination especially if they are undocumented migrants. Such migrants usually are depicted as criminals rather than as victims of the groups that exploit them in the process of migration or at the destination.

The pressing issue of protection of migrants and migrant workers from the region is related to that of controlling the intermediaries in the migration process. While all such intermediaries do not operate outside of the law, a substantial number do. Indeed, the involvement of crime syndicates in people smuggling and trafficking has increased exponentially and has become an important issue in international crime. The multinational nature of these groups makes them difficult to control and their influence is widespread, even involving officials involved in immigration. The thriving of such organizations clearly is related to the failure of documented migration systems and enhanced policing of migration is unlikely to solve the problem of undocumented migration. Improvement of official migration systems must be part of the increasing international effort to combat the people smuggling and trafficking which is especially rife in South and Southeast Asia. In countries of origin there is a need for information programs which provide potential migrants with detailed, accurate information about the costs of the migration process and the nature of the experience at the destination. Studies in several countries indicate that migrants are often poorly informed on both these counts and this can lead to them paying excessive amounts to move and to them having false expectations of the destination. The latter, for example, results in up to one-half of Indonesian women going to the Middle East to work as domestics returning to Indonesia before their contract time has expired. There needs to be a crackdown on the activities of unscrupulous intermediaries in several countries.

The issue of south–north migration representing a 'brain drain' of skill and talent needed for development in the home countries is a longstanding one in the region. In several origin countries, however, the occupational profile of migrant workers is largely an unskilled and semi-skilled one so that movement out of a country like Indonesia certainly cannot be characterized as a brain drain. On the other hand, much of the movement out of countries like India, Pakistan, Philippines, and Singapore is of skilled and highly educated groups. In those cases, however, the movement is not necessarily a net economic loss to the nation. Economic studies have shown that in some cases individuals have a net positive economic impact if they go overseas and remit money back to their home country rather than remain behind. In countries like India and the Philippines there appears to be an oversupply of some skilled groups in relation to the opportunities available locally. Nevertheless, there can be no doubt that the emigration of some skilled groups can be detrimental to local economies in some cases. With the globalization of labor markets such mobility is inevitable. Some countries in the region (e.g., Pakistan) have put in place programs to encourage former emigrants with certain skills to return to their home country. The growth of India's information technology industry, for example, has lured some former emigrants to return.

The interrelationship between international migration and development in the region is much argued. There can be no doubt that international migration can have both positive and negative effects on the well-being of communities left behind. This presents a challenge to policy makers to put in place institutional structures which maximize these benefits and minimize the negative effects. All empirical evidence points to the fact that there is little to be gained from policies and programs designed to stop population movement when there are economic imperatives behind that movement. The challenge is in creating policies and structures which enhance the benefits of the movement.

There is a range of issues which relate to the experience of migrants from South and Southeast Asia in destination countries. Several groups settling in northern countries show a tendency to concentrate in particular communities or develop ethnic enclaves, especially in major metropolitan areas. Debate on this ranges between those who see such areas as barriers to the adjustment of migrants and those who see them as having a cushioning effect in allowing migrants to adjust to life in the new country in a gradual way and with the support of community networks. Other issues relate to discrimination experienced by some groups in housing and job markets as well as nonrecognition of qualifications leading to occupational skidding and deskilling.

6. *Conclusion*

The South and Southeast Asian region represents one of the world's major reservoirs of potential international migrants accounting for more than one-

quarter of the global population and several of its largest labor surplus economies. Until recently, the impact of international migration in the region has been limited but it is now of significance in all countries in the region. Although fertility decline has been significant the labor force in the region continues to grow rapidly. Economies in several countries in the region are also growing rapidly but all of the pressures which have created greater movement since the 1980s than previously will continue to operate. Hence, the twenty-first century will see an increase in the scale and complexity of movement out of, within, and into the region.

See also: Area and International Studies: Development in Southeast Asia; Migration, Economics of; Migration, Theory of; South Asian Studies: Economics; South Asian Studies: Geography; Southeast Asian Studies: Economics

Bibliography

Australian Bureau of Statistics (ABS) 2000 *Migration Australia 1998–99*. Catalogue no. 3412.0, ABS, Canberra

Department of Education, Training and Youth Affairs 1999 *Students 1999: Selected Higher Education Statistics*. Commonwealth of Australia, Canberra

Garson J 2000 Recent trends in migration movements and policies in the OECD countries. Paper prepared for Workshop on Migration and the Labour Market in Asia, Tokyo, 26–28 January

Leete R, Alam I (eds.) 1993 *The Revolution in Asian Fertility: Dimensions, Causes, and Implications*. Clarendon Press, Oxford; Oxford University Press, New York

Litner B 2000 *Far Eastern Economic Review*

Organisation for Economic Co-operation and Development (OECD) 1998 *Trends in International Migration: Continuous Reporting System on Migration: Annual Report*. OECD, Paris

Statistics Canada 1992 1991 Census of Canada. *The Daily* December 8

United Nations 1997 *World Population Monitoring, 1997: Issues of International Migration and Development: Selected Aspects*. United Nations, New York

United Nations Educational, Scientific and Cultural Organization (UNESCO) 1999 *Statistical Yearbook 1999*. UNESCO Publishing and Bernan Press, Paris and USA

United Nations High Commissioner for Refugees (UNHCR) 2000 *Refugees and Others of Concern to UNHCR: 1999 Statistical Overview*. UNHCR, Geneva

G. Hugo

International Migration: Mediterranean Basin and the Middle East

In this article, the Middle East is defined to include primarily the Gulf oil-exporting countries and Israel, and the Mediterranean basin is defined to include the countries surrounding the Mediterranean sea. The types and volumes of international migration experienced by these regions during the 1970s, 1980s, and 1990s are quite different from one another. To the Gulf countries, temporary labor migration has been the predominant type and has been large enough to constitute a major migration system measured on a worldwide scale. Migration to the Mediterranean region has been relatively smaller and includes permanent settlers, temporary workers, and asylum seekers.

1. Data Availability and Quality

Lack of reliable and valid data are frequently a hurdle in correctly documenting the stocks and flows of international migrants. For the Middle East, several advances have been made in establishing data gathering institutions in the major sending countries, especially in South and Southeast Asia (Appleyard 1998, Arnold and Shah 1986). Information about the non-national population is also gathered by the receiving countries and is published regularly (ESCWA 1997). One weakness of the annual flow data from several sending countries is the high percentage of persons who are not reflected in official statistics. In some cases the annual flow may be twice as high as the recorded level.

Data on some countries of the Mediterranean region (France, Greece, Italy, Spain, and Turkey) is available from the annual statistics of the Organization for Economic Cooperation and Development as part of its continuous reporting system on migration (OECD 1993, 1998).

2. The Middle East

Six oil-rich Gulf countries (Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates) predominate as major receivers of migrants in the area. Migrants to the Gulf originate in many parts of the world, and some countries are host to over 100 nationalities (Serow et al. 1990).

Major features of the receiving countries that underlie the migration process are as follows. Oil is the major source of revenue for all the six Gulf Cooperation Council (GCC) countries. All of them are governed by kings or emirs who are regarded as benevolent patriarchs, and who generally share a policy of providing to their subjects free or highly subsidized social welfare services such as education, health care, housing, and employment. All but Saudi Arabia have small indigenous populations. A unique feature of their populations is the extremely high percentage of non-nationals, especially in their labor forces (Table 1). All have high birth rates, which they want to maintain, resulting in young populations and rapid rates of labor force growth.

Table 1
Total population by nationality and percentage of non-nationals in the labor force

Country	Population 1996 Total	Percent non-national	Percent non-national in the labor force (1990)
Saudi Arabia	19,814,330	27.7	60
Kuwait (1997)	2,152,775	65.4	84
Bahrain	566,726	37.4	51
Qatar	568,242	76.5	92
UAE	2,392,927	73.1	89
Oman	2,240,073	27.5	70
Total	27,735,073	51.4	74

2.1 Volume and Characteristics of Migration to the Gulf

Labor migration to the Gulf dates back to the 1930s when oil was discovered and the oil industry began to be established. A remarkable upsurge in such movement occurred after the 1973 oil embargo, which resulted in a dramatic rise in oil prices. The additional income led to ambitious development programs of the Gulf countries that had become rich in capital but were short of labor.

The majority of the migrants to the Gulf are either Asian or Arab, with a larger percentage of the former. Among Arabs, Egypt is currently a major sender. Yemen and Jordan have also been important countries of origin during 1970s, 1980s, and 1990s. Among non-Arabs, four countries in southern Asia (Bangladesh, India, Pakistan, and Sri Lanka) and two in Southeast Asia (the Philippines and Indonesia) supply most of the migrants. Less than five percent of all migrants are from European or American countries.

The Gulf War of 1991 played an important role in restructuring the nationality composition of migrants in the Gulf. Kuwait banned the return of five nationality groups whose governments had supported Iraq during the Gulf War, namely, Iraqis, Palestinians, Jordanians, Yemenis, and Sudanese. About 800,000 Yemenis were expelled from Saudi Arabia, and about 350,000 Palestinian/Jordanians left Kuwait, mainly for Jordan. The persistently high outflows from the sending Asian countries suggest that the trend towards Asianization has been maintained steadily. In Kuwait, for example, the expatriate labor force in 1988 comprised 66 percent Asians, 33 percent Arabs, and only 1 percent other nationalities.

The number of migrants can be estimated in terms of the stock at any point in time or annual flows. The stock of workers from the four major southern Asian senders was estimated to be 4–5 million in 1994–5. Data on annual flows from selected countries are shown for the latest available year in Table 2. Since these data reflect only the documented migrants, they are likely to be an underestimate of the real numbers, probably by 30–50 percent.

Table 2
Annual migration flows and percentage going to Middle East (ME), selected countries

Country of origin	Total	Percent to ME
Bangladesh (1998)	267,667	90.2
India (1997)	416,424	92.8
Pakistan (1997)	153,929	96.0
Sri Lanka (1997)	149,843	84.6
Indonesia (1994)	118,000	40.6
Philippines (1996)	660,122	62.0

Among the receiving countries, Saudi Arabia is the largest recipient and therefore holds special importance. During the late 1990s, about one-third of the Sri Lankan, one half of the Pakistani and Bangladeshi, and 63 percent of Indian migrants headed annually for Saudi Arabia.

The majority of migrant workers from India, Pakistan, and Bangladesh are male. However, 75 percent of migrants from Sri Lanka, about two-thirds from Indonesia, and half from the Philippines were female during the 1990s. One major reason for the feminization of the migrant workforce is the increasing demand for housemaids. Published data actually underestimate the number of female workers from some countries (e.g., India) who do not allow this type of migration because workers therefore move illegally and are missed by official statistics.

Among the male workers, especially those from South Asia, more than half consist of unskilled or semiskilled workers employed as laborers, servants, cleaners, and gas station attendants. Less than 10 percent are professionals.

2.2 Working and Living Conditions of Migrants

Surveys of migrant workers and returnees indicate that for a large majority, labor migration to the Gulf is a positive experience (Gunatilleke 1992). Despite the downward trend in salaries, migrants are able to earn wages that are several times higher than possible in the home country. A majority would like to stay as long as possible, and try to bring in as many of their friends and relatives as they can. For workers and their families who have become used to higher levels of income, there seem to be no better alternative than to keep their Middle East jobs by whatever means possible. The workers themselves are willing to undergo a fair amount of hardship for the sake of their families. In about 70 percent of cases the migrants are married but are not accompanied by their families since the rules often do not allow it. Absence of the migrant has been found to create marital and social problems only in a minority of families.

2.3 Reasons Underlying the Large Flows to the Gulf

In the sending countries, several political, demographic, and socioeconomic factors serve as push factors, while the large wage gap between the sending and receiving countries is a major pull factor that motivates labor migration in the region. Poverty in the sending countries is the single most important push factor. According to the UNDP, South Asia has the largest number of people in income poverty. Its human poverty index shows that 48 percent of Bangladeshis and 47 percent of Pakistanis faced various forms of deprivation in 1997 (UNDP 1997). The rapidly growing population and consequently the expanding labor force, in several countries, exacerbate the level of poverty. Other factors that have contributed to poverty are political conflicts and natural disasters.

Labor migrants often do not come from the poorest or most destitute groups. They are the relatively more enterprising and 'fit' individuals who can afford the ever-increasing financial cost of migration, and many are employed before migration. However, high unemployment rates in the sending countries are also push factor in many cases.

A major facilitator of migration is the formation of informal networks of friends and relatives, making it a self-perpetuating phenomenon. This also results in the formation of migration pockets in the sending country. Some surveys show that friends and relatives arrange work visas for 34 percent of respondents (Shah 1998).

In the receiving countries, several factors perpetuate the demand for foreign workers. The major ones are the small indigenous population and labor force, limited skills among the national workforce, the abundant and cheap supply of expatriate labor, its willingness to work in 3D (dirty, dangerous, and demanding) jobs that nationals shun, and the relatively low participation of women in the workforce in some countries. In addition, the profit that accrues to a fairly large number of intermediaries is a factor. It is well known that labor is sometimes brought in not to satisfy genuine demand but to earn transfer fees. Another important factor is the structural imbalance in the host country labor force. Nationals are employed overwhelmingly in the public sector, about 94 percent of them in Kuwait and 93 percent in Saudi Arabia. Nationals prefer to work in the public sector where work is less arduous and less competitive, and remuneration is generally as good as in the private sector.

2.4 Policies of Sending and Receiving Countries

Governments of the sending countries are highly pro-emigration (United Nations 1998). Most South and Southeast Asian countries have revised their emigration rules and procedures and developed administrative machineries to regulate the outflow, ensure worker welfare, and promote further worker emigration. However, countries differ in the degree of aggressiveness with which they promote manpower export. Bangladesh, for example, has successfully found a new market for labor export, Malaysia.

Remittances from overseas workers have been a lifesaver for the economies of several countries and continue to be a powerful force in shaping government policies (Amjad 1989). In 1994, remittances from overseas workers accounted for 34 percent of the revenues earned from all goods and services exported from Bangladesh. In the case of India, Pakistan, and Sri Lanka, this percentage ranged from 14 to 17 percent. For several Arab senders, remittances have been an equally important source of foreign exchange earnings.

In the receiving countries, the major thrust of the current policy is to reduce the percentage of expatriates. One of the major government policies is therefore to encourage employment of nationals in the private sector. A related policy is to restrict the employment of expatriates in the public sector. A third policy is to curb the visa trading that encourages the import of workers. A fourth policy consists of the crackdown on illegal migrants.

2.5 Becoming Illegal and Amnesty Return

Before the mid-1990s, illegal migration was not a major issue in the Gulf. With the worsening economic situation of the host countries, and the increasing level of unemployment among the nationals, the GCC countries have made concerted efforts to reduce the number of illegal residents. Two ways that a migrant becomes an illegal resident are as follows. The first way is if the migrant takes a job without having the

appropriate visa. This happens if a person works for an employer other than the one who sponsored the work visa, or the type of visa is different from the one required for that work. Second, the migrant becomes an illegal resident by overstaying the permitted duration. Many Gulf migrants are in an illegal status, especially of the first type.

Punishments for living in the Gulf illegally include fines ranging from US$240 in Oman to several thousand dollars in Saudi Arabia, and jail terms ranging from one to six months in different countries. In response to illegal migration, from mid-1996 to mid-1998, all the six countries declared periods of amnesty during which they allowed illegal workers to leave, or regularize their stay, without paying fines or being jailed. It is estimated that more than one million persons were repatriated during the amnesty periods.

2.6 *Protection of Workers' Rights and of the Most Vulnerable*

The Gulf migrant needs protection in the sending as well as receiving country. When abuse occurs, recruitment agents and employers are the two major violators of the migrant's rights. Recruitment agents who arrange the employment for a majority of workers sometimes charge up to ten or fifteen times more than they are allowed officially. In some cases, the workers are abandoned in the host country or a third country of transit. Rules and procedures exist for punishing fraudulent agents. Yet, many loopholes remain. What is reported to, and handled by, the protectorate offices in sending countries is only the tip of the iceberg.

In demanding the rights of their workers in the Gulf, sending country governments are usually quite tolerant of occasional mistreatment. The relationship between the sending and receiving country is not equitable; the former is in a clearly weaker position. In the receiving country, the embassy of the home country typically intervenes only in cases of dispute with the objective of achieving reconciliation rather than promoting the rights of workers. Egypt and the Philippines are known to demand the protection of their workers' rights more actively than other sending countries.

The labor laws of some host countries (e.g., Kuwait) do not cover certain categories of workers, such as housemaids, a fact which implies that this group requires special protection. Even though a large majority (about 80–85 percent) of the housemaids appear to be satisfied with their migration experience, abuse of housemaids is a reality. Among the countries that export female domestic workers, only the Philippines collects and publishes systematic data on the types of abuse suffered by housemaids reporting such violence. The usual complaints are nonpayment of wages, long working hours, and general mistreatment. However, cases of physical assault, verbal abuse, and rape are also known. Runaway maids present a special problem. The embassies of some countries have set up shelters in the host country for temporary accommodation of runaway maids.

2.7 *Migration to Israel*

Since its creation in 1948 as a homeland for the Jewish people, Israel has been a destination for Jews from many different countries, the major ones being Poland, Romania, Iraq, Iran, Morocco, and Tunisia (Appleyard 1991). In the latter years of the twentieth century, many Soviet Jews have migrated to Israel as a result of the relaxation of emigration restrictions. In 1989–96, an estimated 669,000 persons from the former Soviet Union were admitted to Israel (Zlotnik 1998).

Different origins and phases of migration have resulted in a society with distinct ethnic groups, varying in degree of social status and access to resources and power. European Jews (Ashkenazi) are at the top of the hierarchy while Arabs and some other groups, particularly recent arrivals, are at lower levels in the social and political structure.

During the 1990s Israel experienced a new type of migration, namely, one of foreign workers. Since its occupation of large amounts of Arab territory in 1967, Israel had depended on Palestinian workers to fill certain positions at the lowest end of the labor force. These workers often had work permits, but some were illegal. In 1993, the government announced a closure of the Palestinian territories and the supply of such workers became intermittent, depending on the degree of restrictions on work permits. The demand for workers in the construction and agricultural sectors necessitated the import of non-Palestinian foreign workers. In 1996, the government had issued 104,000 work permits to employ foreigners who came mainly from three countries, Romania, Thailand, and the Philippines. Together with legal foreign workers, an upsurge in illegal workers has occurred. It is estimated that there may be as many as 100,000 illegal foreign workers in Israel (Bartram 1998). In 1998, nearly 5,000 illegal workers, most of whom were overstayers, were deported.

3. *The Mediterranean Basin*

The countries surrounding the Mediterranean basin have experienced different magnitudes of migration. Three types of migrants can be considered as typical, namely, permanent settlers, temporary foreign workers, and asylum seekers.

3.1 *Volume and Types of Migration*

The percentage of foreign-born persons in the population of a country provides one indicator of migr-

ation. For seven countries in the Mediterranean basin, these percentages during the 1980s and 1990s were as follows:

France (1990)	6.3 percent
Italy (1996)	2.0 percent
Turkey (1980)	1.9 percent
Greece (1981)	1.9 percent
Spain (1996)	1.3 percent
Tunisia (1984)	0.5 percent
Morocco (1982)	0.3 percent

In March 1990, France had 3.6 million foreigners in its population. Of them, 1.5 million were foreign workers. Major countries of origin were Portugal, Algeria, and Morocco, each of whom had more than half a million nationals living in France. Other countries that had 200–50 thousand nationals in France were Italy, Spain, Tunisia, and Turkey.

Italy was host to about 1,095,000 foreigners in 1996, coming from over 170 different countries. The largest number was from Morocco (120,000). Albania and the Philippines had 64,000 and 57,000 nationals, respectively, in Italy. The above data are based on stock estimates and do not provide an indication of annual flows. Foreigners residing in Italy can obtain nationality after 10 years of residence.

Turkey has been a country of both emigration and immigration. In 1991, about 2.8 million Turkish nationals lived abroad, the largest numbers in Germany (1.8 million), and The Netherlands (0.21 million). Regarding immigration, around 1.6 million ethnic Turks have migrated to Turkey since 1923, mainly from Bulgaria, the former Yugoslavia, Greece, and Romania. Turkey has also provided asylum to refugees from the former Yugoslavia and more recently to Kosovars from Yugoslavia. It has also sent contract migrant workers to the Gulf states and several other countries, especially Saudi Arabia.

Spain had been an emigration country during the last few decades but in recent years has received increasing numbers of immigrants. Between 1989 and 1996, the foreign population increased from 250,000 to 539,000, originating largely in Europe and America. Several measures have been introduced to tighten migration inflows, especially from North Africa.

3.2 Current Issues

Preventing the entry of illegal migrants is a major issue in most receiving countries. Several Mediterranean countries have signed the Schengen Agreement, which provides guidelines on border regulation and works toward preventing the entry of undocumented and unwanted migrants. Restrictive efforts have had limited success, however, and trafficking of migrants is a booming business. Turkey and then Greece (a European Union member) are among the main routes for people from developing countries into Europe. Spain is also considered a possible corridor for immigrants to reach northern Europe (Huntoon 1998).

General anti-immigration sentiment and xenophobia in Europe have been on the rise. Some extreme-right politicians in France are trying to promote a 'national preference law' that would advocate the repatriation of Arab and African immigrants and give a preference to French citizens in housing and jobs. At the same time, however, concerted efforts are being made to regularize the stay of illegal workers and address the issues of naturalization and integration. Several new laws to this effect have been passed recently.

4. The Outlook for the Future

With budget deficits and unemployment among nationals of the immigrant-receiving countries, a slowdown in demand for foreign workers is imminent in the Gulf. Rising unemployment among nationals has raised difficult political questions about the wisdom of importing large numbers of foreign workers. Forceful actions to expel illegal workers are a vivid reflection of the changed mindset. The six GCC countries plan to 'replace 75 percent of foreign workers with their own citizens by 2020' (CISNEWS 1998). If the receiving countries are actually able to implement their restrictive policies, migration to the Gulf will be reduced markedly. The persons most affected by these policies will be the unskilled and semiskilled. However, the demand for such workers will be conditional on the profit accruing to the intermediaries and employers, and on the possibilities of replacing expatriates with nationals. It appears that the demand for housemaids and some types of service workers will continue unabated in the near future.

Policy makers in sending countries will need to recognize the changing attitudes and plans of the Gulf countries. They will also need to plan for the returnees and encourage their entrepreneurial abilities fully. In the long run, sending countries cannot regard migration as a substitute for development at home. It is only through an improvement of its economic and social sectors that the sending country will gain any leverage in the migration process.

Adequate protection of migrant workers remains a difficult issue. There is a need to devise and implement protection through policies such as a guaranteed minimum wage and adequate living and working conditions. However, instituting such policies is at odds with the labor market realities, where the employer decidedly has the upper hand and final word.

In the Mediterranean region, too, immigration policies of the receiving countries are becoming more restrictive. Illegal migration and trafficking are issues of major concern. In terms of border controls and harmonization of their policies, European Union members are faced with many difficult points, espe-

cially in dealing with family reunification and integration.

A general policy that may be recommended for both regions is the facilitation of social and economic development of the countries where migrants originate. In the meantime, mounting evidence shows that restrictions inevitably result in exploitation of migrants. The goal should therefore be to address the root causes of migration and reduce inequalities in standards of living between countries so that migration results from choice, and not compulsion.

See also: Assimilation of Immigrants; Immigration; Immigration and Migration: Cultural Concerns; Internal Migration (Further Types): Industrialized Countries; International Migration by Ethnic Germans; Jewish Diaspora; Middle East and North Africa: Sociocultural Aspects; Migration and Health; Migration, Economics of; Migration out of Europe; Migration: Sociological Aspects; Migration, Theory of

Bibliography

Amjad R (ed.) 1989 *To the Gulf and Back: Studies on the Economic Impact of Asian Labor Migration*. International Labour Organisation, Asian Employment Programme, New Delhi, India

Appleyard R (ed.) 1998 *Emigration Dynamics in Developing Countries, Volume II: South Asia*. Ashgate Publishers, Hampshire, UK

Appleyard R 1991 *International Migration: Challenge for the Nineties*. IOM International Organization for Migration, Geneva, Switzerland

Arnold F, Shah N M (eds.) 1986 *Asian Labor Migration: Pipeline to the Middle East*. Westview Special Studies in International Studies, Westview Press, Boulder, CO

Bartram D V 1998 Foreign workers in Israel: history and theory. *International Migration Review* **32**: 303–25

CISNEWS September 14, 1998, http://www.cis.org [Center for Immigration Studies]

ESCWA 1997 *Demographic and Related Socio-Economic Data Sheets*. United Nations Economic and Social Commission for Western Asia, No. 9

Gunatilleke G (ed.) 1992 *The Impact of Labor Migration on Households: A Comparative Study of Asian Countries*. United Nations University Press, Tokyo

Huntoon L 1998 Immigration to Spain: implications for a unified European Union immigration policy. *International Migration Review* **32**: 423–50

Kritz M M, Lim L L, Zlotnik H 1992 *International Migration Systems: A Global Approach*. Clarendon Press, Oxford, UK

OECD 1993, 1998 *Trends in International Migration: Continuous Reporting System on Migration*. Organisation For Economic Co-Operation and Development, Paris

Serow W J, Nam B C, SlyD F, Robert H W (eds.) 1990 *Handbook of International Migration*. Greenwood Press, New York

Shah 1998 The role of social networks among South Asian male migrants to Kuwait. In: Appleyard R (ed.) *Emigration Dynamics in Developing Countries, Volume II: South Asia*. Ashgate Publishing, Aldershot, UK, pp. 30–70

UNDP 1997 *Human Development Report, Mongolia 1997* United Nations Development Programme, Ulaanbaatar, New York

United Nations 1998 *International Migration Policies*. Department of Economic and Social Affairs, Population Division, New York, ST/ESA/SER.A/161

Zlotnik H 1998 International migration 1965–96: An overview. *Population and Development Review* **24**: 429–68

N. M. Shah

International Organization

1. Definition

The term International Organization (IO) refers to norms of conduct and processes of interaction among more than 6,000 (1999) concrete organizations and their members, bureaucracies wishing to attain objectives that are seen by them as unrealizable without cooperating across political and organizational boundaries. Of these 6,000 IOs, 254 (170 in 1962) were entities whose membership was composed of states (IGOs), such as the United Nations (UN) and the Caribbean Community; another 5,766 (1,542 in 1962) represent private associations with international objectives (NGOs), such as Doctors Without Borders and the European Confederation of Free Trade Unions; 72 percent of the IGOs and of the NGOs represent a regional, as opposed to a universal, constituency. (*Yearbook of International Organizations*, 1998–9, Vol. 3, App. 3, Table l; 1962–3, Vol. 2, pp. 15 and 19, as amended by author's computations from tables of content of Parts I, II, III, IV; count based on the *Yearbook's* categories A, B, C, D only; Murphy 1994.)

1.1 Cooperation in International Relations Theory

The terms cooperation, collaboration, and coordination are widely used but do not have identical connotations in international relations theories. Realists and neorealists doubt the possibility of institutionalizing cooperation across borders but agree with other practitioners of utilitarian rational choice approaches, notably neoliberal institutionalists, in attaching the specific technical meanings to each term used by game theory. For constructivists, peace theorists, and world systems theorists, to collaborate (or cooperate, or coordinate) means that the actors in question have achieved a consensus, often temporary, that their objectives (or interests or values) coincide sufficiently so as to make joint action more effective than individual action. Although these schools agree that in many cases consensus varies with the nature of the issues, the schools disagree in assigning reasons for the differential patterns of cooperation. But without

belief in the efficacy of and the need for cooperation there can be no IOs, either in theory or in the practice of actors.

1.2 Organization, Institution, or Regime?

The term 'international institution' is often used interchangeably with IO. IOs are instances of the attempted institutionalization of overarching norms to govern state and private behavior. But the notion of institution is much broader since it refers to any routinized and sanctioned practice in social life. Institutions may make use of organizations, but need not do so. Similarly, 'international regimes' are often confused with IOs. Regimes are highly institutionalized forms of collective behavior in specific issue-areas (e.g., trade, telecommunications, security from aggression) that may, but need not, make use of IGOs. Often several regimes are embodied in single comprehensive IOs, notably in the UN, the Organization of American States (OAS), and the European Union (EU).

2. Historical Evolution of Organizations

2.1 Main Events

After 1860 there developed a number of IGOs dedicated to coordinating national measures for dealing with communicable diseases and standardizing postal and telecommunications services. In 1919 the League of Nations was founded by the victor nations of World War I to prevent a repetition of the war they had just won. In the event of an interstate dispute, the IGO was to provide dispassionate inquiry, offer peaceful methods of settling disputes, and—failing these efforts—restrain the aggressor state by the use of collective force (sanctions). The outbreak of World War II in 1939 confirmed the failure of the League.

The victors of World War II, in 1945, founded the United Nations (UN) on the same principles of collective security that had guided the League. One reason for the League's failure had been a requirement for unanimity in its main decision-making organ; the UN sought to remedy this by stipulating only the consent of the five largest powers to make binding a decision to use force. But like the League, the UN was powerless to stop aggression by one of these five. With less faith in law and good will than the founders of the League, the creators of the UN preferred a dozen 'specialized agencies'—independent IGOs loosely linked to the UN—whose services where thought to improve the lot of humankind and thus reduce the reasons why states would wish to resort to force. Such agencies arose in the fields of food and agriculture (FAO), public health (WHO), labor standards (ILO), education (UNESCO), intellectual property (WIPO), and meteorology (WMO). In addition, hundreds of other universal membership IGOs with special missions, often not linked to the UN system, continue to be created. The tidal wave of NGOs developed in response to the growth of IGOs, to be able to press the interests of their members in IGO forums, to shape IGO decisions and to aid in their implementation (Claude 1956).

Other than the UN's measures for dealing with aggression, these IGOs have no power to coerce their members; few of them have the right to make decisions firmly committing members to specific policies. The exception is a group of universal membership IGOs set up after 1944 to free international trade and investments, stabilize currencies, and avoid large-scale international economic crises. The World Bank, the International Monetary Fund (IMF), and the World Trade Organization (WTO) do have the power to make binding decisions and to enforce them with monetary and trade sanctions.

2.2 The New Descriptive Terms for a Postidealistic IO

Many advocates of the UN had seen IOs as stepping stones gradually leading toward world government, because IGOs were expected to attain greater autonomy from their members and NGOs were thought to grow in numbers and influence, a sentiment that had never inspired many political leaders. IOs were relaunched after World War II under three different ideological banners: conservatism was used to justify the new collective security, functionalism the creation of the specialized agencies, and liberalism legitimated the new emphasis on human rights and the abolition of colonial empires.

Conflicts among these rival ideologies were never resolved and they continue to characterize the arguments employed by advocates and politicians. The creation of regional IOs—*regionalism* as the alternative to the UN's *universalism*—was urged by many governments and some scholars as a better way to assure peace and plenty than the UN's; in the 1950s and 1960s many new regional IOs were created, especially in Europe and Latin America, competing in some instances with the UN system and representing institutionalized conflicts between blocs of states.

Current writing dealing with IOs distinguishes between three core types of organization seen as expressions of their charters or constitutions. *Intergovernmentalism* refers to a system in which all major decisions are made by member states acting in conformity with the IGOs' decision-making rules; these IGOs mostly rely on their member states for the implementation of decisions. *Supranationality* describes IOs which possess some powerful organs not dominated by governments; such IOs also are able to make binding decisions relying on their own bureaucracies, such as the EU and the WTO. *World* (or

regional) *government* exists when a group of states chooses full-fledged federal institutions for itself, such as the Caribbean and East African federations which failed in the 1960s, and abortive efforts to write a world constitution (Clark and Sohn 1958).

IGOs animated by close ties among the specialized bureaucracies of their members, such as defense, agriculture, money, or public health, feature *transgovernmental* relations, to highlight the dominating role of specialized bureaucracies rather than the states' central organs. IGOs in which NGOs play a crucial role (such as the ILO) as well as ties among NGOs themselves, are called *transnational* to stress their autonomy from the nation-states who normally dominate world politics. Advocates of the growth of a global civil society stress the growing importance of transnational ties. *Epistemic communities* are networks of experts committed to knowledge-based views on the causes of phenomena they are called upon to abate; they actively shape the decisions of many IOs with science-related programs (Keohane and Nye 1977, P. Haas 1990).

3. IO and Social Science

3.1 From Advocacy to Analysis

By and large, social scientists ignored IO until the 1960s, considering the field dominated by idealist advocacy and resistant to systematic analysis by scholars other than lawyers. Idealism held out the hope that reasonable politicians, accepting as binding the international law of the day, would always avail themselves of the means of peaceful accommodation. Political scientists, sociologists, and economists in advanced industrial countries, once idealism was seen as incapable of answering complex questions, became interested in applying theories of organization, along with quantitative methods, to the systematic study of IO (E. Haas 1964, Alker and Russett 1965, Kriesberg 1968). Others sought to embed the study of IO in theories about modernization and pluralism (E. Haas 1958, Cox and Jacobson 1973), to be followed in the 1980s and 1990s with the application of rational choice and modeling constructs (Cedarman 1997). Writing on IO in formerly communist countries was almost entirely dominated by the defense of national policies, a situation which changed to some extent after 1990. Research and writing in developing countries is still dominated by advocacy of preferred positions.

3.2 Normative Evolution: from Organization to Institution

Lawyers and sociologists proposed to study change in world politics by tracing the evolution of norms of state behavior.

The term 'norm' may refer to prevalent state practices considered right and self-evident as well as to desirable standards of behavior, whether generally observed or not. In either and both senses, change can be mapped by investigating whether a more constraining set of norms is gaining adherence (Hoffmann 1981, Kratochwil 1989). The explanation of such changes became a chief interest of social scientists.

One way of conceptualizing these changes is to stipulate an evolution from organizations to institutions. Organizations are thought in all current theories to be animated, at first, by the instrumental interests of their members; a change in interest can trigger either organizational growth or decay. Institutions, however, are dominated by norms actors take for granted as being desirable. An *organization* whose members move away from the purely instrumental (and usually short-run) view of their interests becomes an *institution* when the same members come to value the IO for its own sake and then come to think in longer-run terms when defining their interests.

3.3 Aspects of Global Change

Constructivist and neoliberal scholars believe that there is no longer an agreed hierarchy among the objectives of states (Keohane 1984, Ruggie 1998). Nor is there a hierarchy among IGOs, no organizational core at the global level. Alongside ties and conflicts among central decision makers, transgovernmental and transnational ties proliferate and increasingly influence core decisions by heads of government and foreign ministries. The world of IOs is one of many overlapping and competing networks involving many types of actors. The main activity of these networks is the continuous negotiation of agreements, following agendas no longer set exclusively by a few powerful nation-states. Scholars who study forms of organizational decision making that sidestep the core executive power of member states draw increasingly on sociological network theory (Powell 1990).

In selected issue-areas, e.g., environmental rules, human rights protection, trade disputes, and loan negotiations, something approximating supranational competencies are evolving in some regions and also in the UN system, owing to commitments that member states made earlier, usually for instrumental reasons, and later find inconvenient or embarrassing to disown. In fact, although not in formal law, the vessel of national sovereignty is springing bigger and bigger leaks as the formerly sharp distinction between international and purely domestic policy issues becomes blurred: such 'local' events as famine and civil war involving genocide become matters of global concern.

Different schools of thought advance different reasons for this trend. Neoliberals stress that states

allow new, sovereignty-restricting, norms to evolve because the transaction costs of engaging in all kinds of international activity would rise too high if IOs did not provide important information and if trust and reciprocity were allowed to decay. Game theorists stress the parties' need for maintaining their reputation for honesty. Peace theorists explain the trend by suggesting that leaders are learning to accept peaceful norms. Constructivists argue that leaders, under certain conditions, learn to change their 'identities' and redefine their interests so as to stress cooperation when they experience the costs of aloofness or the limits of acting alone (P. Haas and E. Haas 1995). Because theorists disagree about the significance and roles of IO, indicators for judging change were devised.

3.4 Indicator of Organizational Change: Autonomy

The degree of autonomy of an IGO is an indicator of the distance the world (or a region) has traveled away from the initial condition under which single sovereign states, or groups of sovereign states opposing one another in alliances, were the only players in the game. Such was the global order in 1945. The IGOs founded then were entirely the creatures of the states that were their creators and paymasters. An increase in autonomy is a step toward world government, a decrease a reaffirmation of state sovereignty and hegemony.

The attainment of supranational powers by an international bureaucracy is one measure of autonomy, and so is the ability of an IGO to make binding decisions by majority vote. Autonomy is advanced when epistemic communities of experts shape decisions, or when NGOs act in this fashion, alongside instructed representatives of states. Autonomy also increases when the making of binding decisions is left to panels of bureaucrats, judges and other experts. Autonomy grows to the extent that an IO, or the network of which it is a part, is an open system whose members are able to obtain and assimilate new knowledge about the issues on its agenda, when decision making is no longer monopolized by the relatively fixed interests of a few member governments (Adler and Crawford 1991, Reinalda and Verbeek 1998). Increasing autonomy implies that international relations are mediated by forces that represent a higher degree of institutionalization than was true in 1945. Increasing IO autonomy is associated with more powerful IGOs and with growing success in solving the problems of peace and well-being IOs are expected to master. Deinstitutionalization is gauged by observing reversals in such trends.

3.5 Indicator of Organizational Change: Multilateralism

Another way to gauge the growing (or waning) influence of IOs on global affairs is to observe changes in multilateral dealings among states and NGOs. (Ruggie 1993) An increase in items on the agenda of global issues, a waxing of international negotiations and of networks of actors conducting them betoken a growth in the density of contacts, needs, and responses. Increases in multilateralism are measured by the frequency of meetings, the number of international agreements concluded, and the widening scope of these agreements, driven by the expectations of actors that unilateral action is both ineffective and improper. Like growing autonomy, increases in multilateralism imply more restraint on state behavior. Unlike autonomy, a growth in multilateralism need not imply more efficient and more effective IOs—just denser relations among IGOs, NGOS and in joint efforts by both types. Autonomy refers to institutionalized power of IGOs, intergovernmental and supranational, universal and regional; multilateralism is concerned with the density of communication networks, not with power.

4. How Much Organizational Change Has There Been?

4.1 Regional IOs and Regional Integration

Only the EU has demonstrated marked increases in both multilateralism and autonomy and has thus contributed to the rapid integration of the continent. Many IGOs in the Caribbean and Central and South America—dedicated in principle to regional integration—have lost autonomy. IOs associated with the former Soviet Union have mostly disappeared and their successors remain weak. Nor do groupings in Southeast Asia and the Pacific Rim demonstrate expanding autonomy. Only the Cono Sur in South America and the organization of West African states show some growth in autonomy and are making visible contributions to the closer integration of their regions.

The situation with respect to multilateralism is slightly different. Western hemisphere IOs are associated with a sharp rise in multilateral ties despite the poor showing regarding autonomy. The same is true of Southeast Asian and Pacific Rim IOs. On the other hand, disintegration and fewer ties are evident in most of the African and the Middle East/Islamic organizational worlds.

In the Caribbean and Central America an earlier trend toward greater autonomy and multilateralism was reversed as both regions seek greater integration with a North American region in which multilateralism is growing apace.

4.2 Universal Organizations

Since 1945, the family of UN-affiliated IGOs has clearly contributed to, and remains the core actor in, a

rapidly increasing network of multilateral ties. Multilateral practices are concentrated in these issue areas: environmental protection, use and allocation of scarce resources, articulation and protection of human rights, economic development, public health, and the stabilization of the global economic system. The phenomenal growth in the numbers and influence of NGOs active in shaping UN policy is evidence of this trend.

The same cannot be claimed of the autonomy of the UN system. The pattern is mixed. Probably, the autonomous powers of the Secretary-General and of the High Commissioners for Refugees and for Human Rights have increased slightly over the years, but not dramatically. The influence of the Economic and Social Council has waned. Such autonomous powers as had been enjoyed earlier by the ILO, FAO, and UNESCO are no longer discernible. On the other hand, the autonomy of the World Bank, IMF, and WTO has sharply increased since 1980 (Culpeper 1997), and so has the autonomy of the International Atomic Energy Agency (IAEA) as it responded to major crises and emergencies.

5. *Toward a Synthesis of Theories about IO?*

Much of the literature is still driven by contributions that advocate increases in autonomy and multilateralism. 'Global governance' is the current label under which international integration of all kinds, including world government, is being discussed. The term is also used to describe any desired reforms of IGOs to enable them better to address new issues on the global agenda (Ramphal and Carlsson 1995). Advocacy stresses the need for more popular participation in the work of IOs, for more equality in the representation of state interests, for greater transparency in decision making, and for more accountability to people as well as to governments. If acted on, these suggestions would move IO much closer to world government. As advocacy, then, idealism remains very much alive, but only peace theorists embrace this approach on analytic grounds.

Most social science theories about IO, however, are not concerned with this goal. They remain committed to including the discussion of global institutional trends in overall explanations of collective behavior, social statics, and social change. Neorealists continue to marginalize IO altogether whereas neoliberals remain wedded to utilitarian–microeconomic principles in their search for the explanations of interests animating actors in IOs. World systems theorists, like neoliberals, are unable to account for growing autonomy and multilateralism in the protection of human rights and of the environment, issue areas outside the materialist foundations of these schools.

The ontological commitments of these theories are so different as to make a synthesis very difficult. If one is possible it may be the result of constructivists' efforts to explain all actor behavior in terms of material *and* non-material 'identities,' qualities that are ontologically prior to expressed interests. Thus a commitment to norms for their own sake can be considered as genuine an interest as the devotion of wealth and security. Autonomy and multilateralism are thus given a theoretically endogenous status.

See also: Cooperation and Competition, Psychology of; Cooperation: Sociological Aspects; European Integration; Global Governance; Globalization: Political Aspects; International Justice: Philosophical Aspects; International Law and Treaties; International Science: Organizations and Associations; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; Labor Organizations: International; Multinational Corporations; Regional Government; Regional Integration; Science and Technology: Internationalization; United Nations: Political Aspects

Bibliography

Adler E, Crawford B (eds.) 1991 *Progress in Postwar International Relations*. Columbia University Press, New York

Alker H, Russett B 1965 *World Politics in the General Assembly*. Yale University Press, New Haven, CT

Cedarman L-E 1997 *Emergent Actors in World Politics*. Princeton University Press, Princeton, NJ

Clark G, Sohn L 1958 *World Peace Through World Law*. Harvard University Press, Cambridge, MA

Claude I L 1956 *Swords Into Plowshares*. Random House, New York

Cox R W, Jacobson H K (eds.) 1973 *The Anatomy of Influence*. Yale University Press, New Haven, CT

Culpeper R 1997 *Titans or Behemoths?* Lynne Rienner, Boulder, CO

Fawcett L, Hurrell A (eds.) 1995 *Regionalism in World Politics*. Oxford University Press, Oxford, UK

Haas E 1958 *The Uniting of Europe*. Stanford University Press, Stanford, CA

Haas E 1964 *Beyond the Nation-State*. Stanford University Press, Stanford, CA

Haas P 1990 *Saving the Mediterranean*. Columbia University Press, New York

Haas P, Haas E 1995 Learning to learn. *Global Governance* **1**(3): 255–84

Haas P, Keohane R, Levy M (eds.) 1993 *Institutions for the Earth*. MIT Press, Cambridge, MA

Hoffmann S 1981 *Duties Beyond Borders*. Syracuse University Press, Syracuse, NY

Keohane R 1984 *After Hegemony*. Princeton University Press, Princeton, NJ

Keohane R, Nye J 1977 *Power and Interdependence*. Little, Brown, Boston, MA

Kratochwil F 1989 *Rules, Norms and Decisions*. Cambridge University Press, New York

Kriesberg L (ed.) 1968 *Social Processes in International Relations*. Wiley, New York
Murphy C 1994 *International Organization and International Change*. Oxford University Press, New York
Powell W W 1990 Neither market nor hierarchy. *Research in Organizational Behavior* **12**: 295–336
Ramphal S, Carlsson I for the Commission on Global Governance 1995 *Our Global Neighborhood*. Oxford University Press, New York
Reinalda B, Verbeek B (eds.) 1998 *Autonomous Policy Making by International Organizations*. Routledge, New York
Ruggie J G (ed.) 1993 *Multilateralism Matters*. Columbia University Press, New York
Ruggie J G 1998 *Constructing the World Polity*. Routledge, New York
Weiss T, Forsythe D, Coate R 1997 *The United Nations and Changing World Politics*, 2nd edn. Westview Press, Boulder, CO

E. Haas

International Relations, History of

International relations comprise the ensemble of economic, demographic, cultural, and military exchanges among states as well as among peoples gathered in organized societies. Among other fields of social sciences, the theory and history of international relations have contributed to the knowledge of international relations, each one having a different focus. This article traces the latter's evolution.

1. *German Origins*

International relations was converted into a field of historical sciences study during the first half of the nineteenth century; this means that, by then, international relation's were isolated as an investigation object, whose treatment originated its own scientific procedures, capable of evolving with relative autonomy. The conception of this research area was associated with the creativity of Leopold Von Ranke considered by many to be the founder of scientific historiography. Ranke was convinced that European civilization fermented in the states system and could achieve a universal dimension. He constructed a history of international relations based on a tripod: (a) the study of European great states' internal development; (b) the study of collective movements and the forces resulting from institutional, religious, or social experience; and (c) the analysis of foreign policies, involving the relations among states, and particularly the confrontations among great powers.

From German unification, in 1870–71, to the First World War, a deviation of German historiography from international relations took place, reducing Ranke's original European and universal scope of study, and establishing the primacy of foreign policy, which served to exhibit the ascension of Prussia during the Wilhelminic Era. This type of 'germanocentric' approach acquired national equivalents in other European countries, which caused history of international relations to be reduced into a diplomatic history. This new form of history of international relations preserved, within Europe and outside it, the scientific character of exploring archival documents; however, it turned apologetic by receding its arguments to those of chancelleries belonging to the countries to which historians were connected; it became, then, poor in cognitive terms, since it merely treated the apparent movement of states' behavior, and neglected the other relevant aspects of Rankean analysis.

During the same period, the study of international relations had to face imperialism, a phenomenon related to the domination, either formal or informal, direct or indirect, of European States, as well as of the United States and Japan, over other territories, especially the southern area of the planet. This fact coincided with worldwide spread of the great powers' foreign policies. Responses to the challenge of explaining imperialism came, on one side, from imperialism theories—which were neither theory nor history of international relations—and, on the other side, from diplomatic history, this one ambitiously searching scientific explanations which were, nevertheless, attached to the interests of national states. When the Versailles Peace Treaty of 1919 blamed the German Empire for the onset of the First World War, the fact that many historians from various countries debated the causes of the conflict and the responsibility of different states reinforced the introspective character of diplomatic history. It thus distanced itself more and more from Ranke's conceptual definition and universalist proposals.

2. *Multiple Causes—the French Contribution*

It was not before the period between the World Wars that historiographical revolution took place which would rescue the study of international relations to the modern world of social sciences. This movement was strongly influenced by French historians, and it had one of its centers in the *Institut d'Histoire des Relations Internationales Contemporaines* (Institute of Contemporary International Relations History), created in 1935 at the Sorbonne, and directed by Pierre Renouvin, Jean-Baptiste Duroselle, René Girault, and Robert Frank successively. The knowledge elaborated by the Sorbonne group encompasses hundreds of Ph.D. theses, a general history of international relations, works of theoretical reflection on the theme, as well as dozens of specialized books and the periodical *Relations Internationales*, published in association with the Graduate Institute of International Relations, in Geneva, since 1974.

Despite recognizing the role of the state, particularly the great powers and empires (Duroselle), the French have penetrated the social dimension of international relations (Renouvin) and given special emphasis to the role of economics as a long-term factor conditioning the trends (Girault). Consequently, the object of study won higher precision and scientific adequacy. Reflections on how to make both a history of international relations and a critical evaluation of the knowledge produced about it have led to the method's refinement towards multiple causes. This was accompanied by the diversification of information sources in order to embrace the complexity of factors aggregated to the research.

These French historians tried to create a *corpus* of knowledge for understanding international relations by defining four complimentary investigation directions. First, historians should identify the components of international life involving statesmen—whose temperament is to be disclosed in the light of anthropology and psychology; the images people have of foreigners; the importance of borders and of groupings, such as small organized communities; nation-state and plurinational organizations. Second, the historian is expected to investigate origins and causes of international relations, comprehending a body of historical forces (nature elements, demography, economics, principles and values), whose movement and efficacy should be described and connected to foreign policy decisions. Third, the study should turn to identifying the purposes of international action, which involve the determination of agents, the role of information, as well as the means, ends, and risks of strategic planning. In order to circumscribe the object of study in all its dimensions, the historian should, fourth, follow the movement, the process of international relations through time, unveiling the scheme of symmetrical or asymmetrical relations, peaceful or conflicting negotiations, war, and diplomacy.

These four groups of factors represent diverse ways of research and illuminate the object of study in an attempt to provide it with a understanding. The multiple causes system consists precisely of this complex approach. According to Duroselle, the theory of international relations, the point to which historical investigation flows, should be based on empirical knowledge, and would have as its cognitive frontier the establishment of regularities, rules, and recipes taken from history's advice. One example is the idea that the empire—the excess of power—will always perish, either because it engenders the unbearable on its dominions, or because it provokes the coalition that destroys it.

3. *Order Over Anarchy—the British Perspective*

Since the 1950s, an intellectual effervescence comparable to the earlier one in France, motivating studies on international relations history, could be observed in England. The renovating spirit has modified its historical focus, evolving from strong empirical hues to systemic approaches. In the origins of this modernizing movement are Donald Watt, professor of International History at the London School of Economics and Political Sciences, the members of the British Committee on the Theory of International Politics, which operated for 25 years congregating researchers among whom are Herbert Butterfield, Martin Wight, Hedley Bull, and Adam Watson. *The Expansion of International Society* (1984) and *The Evolution of International Society* (1992), amid other works, have divulged new research trends and British contributions to the field of knowledge.

These British scholars on international relations kindled the faith on the classics (Hobbes, Locke, Grotius, Kant, Pufendorf, Heeren, Clausewitz, Hanan, Tocquéville, Ranke). British historians accentuated enduring aspects of human nature which reflect on the evolution of international relations, by making international life somehow predictable and emphasizing continuity over change. According to them, economic, political, and military interactions among states which depend on each other and are guided by their own interests eventually lead to establishing rules, institutions, and common relationship values which, in turn, orient the external conducts of those states. Thus, the *system of States* or *international system* is formed. In some cases, this international system intensifies its net of interactions on the basis of specific rules and values, originating the *international society*, this one having a strong cultural tone. This evolution characterizes the passage, in the eighteenth century, from the *European States system* to the *European International Society*.

Despite not admitting associates to its directory group, the *European International Society* expanded towards American countries and Japan in the nineteenth and twentieth centuries. It gave room to an international system on a worldwide scale—movable, pluralist, competition-prone—whose expansion device was the great powers' European concert. It imposed forms of internal organization and relations among States, juridical regulations, ways of trading, living and thinking, and engendered, from 1800 to 1914, the greatest global domination system ever yet known. Conflicts among its protagonists and protests against Western supremacy allowed the rise of a new scenario, marked by the emergence of the two superpowers, against the background of the Second World War, and of the Third World afterwards. A new international society was born. This worldwide international society of the twentieth century's second half, partly heir to the former, partly led by diverse rules, institutions, and values, evidences the decline of the European matrix in terms of both world order conformation and international relations. At the end of the Cold War, with the challenges of world

reorganization, British historians would do better—according to the Italian critic Brunello Vigezzi—if they insisted less on European inheritance and influence on the configuration of international society, and recognized in it the fecundity of its richness and possibilities.

The British school had, therefore, the merit of fruitfully associating the theory and history of international relations, thus creating the key concepts of *system* and *society* in order to advance in the comprehension of international life. The English connected past and present and cultivated a belief in the ability of international agents to administrate anarchy via consensus. If Duroselle was convinced that there is no possible theory of international relations without a historical basis, to Watson, Theory and History blend together. But British scholars, unlike the American political scientists, are not interested in a *prince counseling* theory. They are concerned more about the cognitive side and reflect intellectual preferences, ideals, and values, which they consider most adequate to study international life.

4. Contributions from Italy and Switzerland

In Italy, modern international relations historiography established a dialog with the French school. Federico Chabod and Pierre Renouvin conversed about methodological conceptions and approaches. This dialog broadened with the diffusion by Brunello Vigezzi of the above-mentioned British Committee works in Italy. As a result of these interactions and of its own creativity—take, for instance, the works of Mario Toscano and Ennio Di Nolfo, among others—the Italian school became fertile and diversified. Its originality lays on the insistence on *spiritual forces* inherent to the movement of international life: mentalities, sentiments, public opinion, collective passions, systems of values. The Italians' works mirrored, therefore, their peninsula's international relations and foreign affairs, that is, the divisions which have opposed conservatives to revolutionaries, monarchists to republicans, expansionists to moderates, in the last two centuries.

Even denser/tighter/closer was the collaboration between the French and the Swiss. Nevertheless, also in this case there was a singularity characterizing the studies developed especially by the Graduate Institute of International Relations at Geneva University (Antoine Fleury, Jacques Freymond, Daniel Bourgeois, Yves Collart). This Institute, organized to refine interdisciplinary studies of international relations, has distinguished itself precisely for incorporating to history contributions from other disciplines, above all political science, but also economics and law. In addition, it has not left aside the interest on neutrality, which has been one of the historical trends of Swiss foreign policy.

5. The National Imperatives in the United States

Two facts conditioned the expansion of studies on international relations history in the United States: the escalating role of the country in twentieth century international relations, and the extraordinary development of international relations theory evidenced since the Second World War. The two facts granted the field of study an introspective trait not observed in other regions, with respect to both the study method and the elaborated knowledge. This introspective trait separates American scholars from the universalist approach of the European schools. Nevertheless, it is inappropriate to conclude that the history of international relations in this country has had a linear development of its object, focus, and method. Perhaps no other historiography has presented such anguish and hesitations.

As of the 1930s, similarly to what happened in the rest of the world, a documental and scientific yet ethnocentric diplomatic history flourished in the United States, exemplified by the works of Samuel Bemis. As in other countries, the reaction came in the following two decades; it was conducted, to the greatest extent, by Thomas Bailey and Charles Beard whose studies brought back internal factors, as well as universal, economic, social, and ideological conditionings of foreign policy decisions. The movement *inwards* was substantiated in the years following the Second World War, when a strong nationalist interpretation trend was observed as a result of the Cold War and of the country's expanding international responsibilities. Realist theorists and consensualist historians such as Hans Morgenthau, George Kennan, and Walter Lippman privileged the state as an agent of international relations and appealed to the elements of national unity to confront external threats, among which were firstly nazi-fascism, and then communism. An official and ideological knowledge was produced.

The refraining of the Cold War and the painful experience of the Vietnam War strengthened critical reflections about international relations. In the 1970s, numerous historians followed researchers like William Appleman Williams and David Horowitz among others, on the path of their discoveries about the relativity of national ideals; the responsibility, shared with the Soviet Union, for the onset of the Cold War; and the prevalence of North American economic interests over international politics.

This critical refinement has gone further in the last decades of the twentieth century, particularly due to the opening of archives and the handling of classified documents in the United States, Russia, and other countries; also relevant has been the use by researchers of modern technological resources. This is how British Alan Milward rebuilt, on the base of new empirical information, the history of relations among European countries, and remade the path to European integration shifting away from the idealistic explanations

originated by Jean Monnet's conceptions. In the United States, Arthur M. Schlesinger Jr., Stanley Hoffman, Robert Gilpin, Samuel Huntington, and John Lewis Gaddis—although it is unfair to mention only these names—reintroduced discussions about important issues referring to the North American great power's relations with the world after the Cold War and the collapse of the Soviet Union. Research again revolved the decision-making process in foreign policy and began to understand the United States' role as the only hegemony on a planetary scale since the end of the Cold War.

At the end of the twentieth century, one cannot identify a hegemonic current of contemporary international relations interpretation in the United States. Nor can it be affirmed that this country's historians have released themselves from the cognitive constraints of the past. These constraints derive from both the frustrations of international theory to explain the evolution on international life, and the endless contending of interpretations. Gaddis has expressed in these circumstances North Americans' pessimism towards the scientific value of historical knowledge.

It does not seem adequate to apply to the United States the notion of an international relations history school valid for European research groups earlier mentioned, since American historians have not demonstrated continuity and coherence in the method's choice and in the definition of the study object. North Americans have accomplished a rich and diverse research experiment, whose long-term profile lies on the value of realist approaches, which restrict the observation scope of their nation-state's actions. In a certain way, they have prolonged throughout modernity and postmodernity the introspective traits of old diplomatic history elaborated before the 1930s.

6. *Brazil, Argentina—the Southern Approach*

Just as in other parts of the world, Latin America also displayed the treaty-led, factual, and apologetic diplomatic history so common before the Second World War. Pandiá Calógeras, in Brazil, and Vicente G. Quesada, in Argentina, clearly represent this approach. The History Department at the University of Brasilia and the Instituto de Investigaciones de Historia Económica y Social (Institute for Social and Economic Historical Research) at the Buenos Aires University have gathered precursory advanced study groups which allowed modernization of that field of knowledge in the region. In those institutions, studies of international relations history acquired a unique profile deriving from Latin American country's own way of viewing the world, supplied by their concerns and external designs.

Thus, the perspective of international analysis in the region diminished the role of war and security, since the subcontinent has a pacific tradition. At the same time, the Southern approach emphasized the role of the development in international studies. The central argument explored by Latin American scholars has been the fact that development is the axis of majors countries' foreign policies since the 1930s.

In addition to general histories of Brazilian and Argentinean international relations (José Paradiso Juan Archibaldo Lanús, Paulo G. F. Vizentini Clodoaldo Bueno, Amado Luiz Cervo) various studies have explored the relations between these countries and the great powers, mainly the United States (Gerson Moura, Moniz Bandeira, Mario Rapoport), between these countries and Africa (José Honório Rodrigues, José Flávio Sombra Saraiva), as well as regional relations (Paulo Roberto de Almeida, Edmundo Anibal Heredia, Francisco Doratioto). Historians have concentrated their analyses both on the economic possibilities opened by the expansion of capitalism and on mechanisms to bear or overcome historical backwardness. They had to face the local foreign policy options dilemma: to choose either an autonomous development, based on a robust and sustainable national economy, or a development associated with international economic forces, based on foreign capital and enterprises. Amidst the bipolar world and the globalization taking place in the end of the twentieth century, relevant issues were constantly discussed. Among them, development's legitimacy, inequality among nations, asymmetric relations between superpowers and developing countries, international cooperation, exploitation and dependence, the permanence of power and wealth structures, as well as other ingredients of Latin American countries' experiences of international insertion.

7. *The Specific and the General in Historians' Works*

The international relations history field of study is not merely the field of multiple causes resulting from the empirical dimension of research and from the relevance of the process over the event. It is also the field of the political world, in which peoples' necessities, interests, ideas, and values sprout when they envisage other peoples or move beyond their borders. Historians reflect this double-faced research on their work.

Cognitive improvement depended in the past and depends nowadays on the careful conservation of historical documents and on the easy access that governments and international organizations grant researchers. This improvement has evidenced the need to diversify information sources, going beyond diplomatic archives kept in chancelleries and similar institutions, which are necessary but insufficient to clear up some questions posed by international life analysts. Schools and interpretation currents spread all over the

world reveal, first, that international relations have consolidated as a study object for historians; second, that they developed coherent and efficient research methods; and third, that the amount of knowledge elaborated should be proportional to the growing importance of international relations in the lives of individuals and societies in the last two centuries.

The march towards the expansion of this field of knowledge has profited from two dialogs established in its course. Historians organized the Commission of History of International Relations within the International Committee of Historical Sciences in the 1980s, which promotes frequent colloquia where experts exchange research experiments aiming at purifying concepts and practices. Apart from this process, historians have broadened their interactions with akin sciences, such as economics, psychology, anthropology and, especially, political science and the theory of international relations.

When theorists and historians of international relations work, they meet in one *underground stream* (Brunello Vigezzi) of international life; in it, theorists realize that their explanatory formulations become more coherent provided the enlargement of the observation field's empirical base, and historians agree that theories illuminate reality and open ways to its comprehension. Notions suggested by some authors to oppose Theory to History are precisely the same ones advocated by other authors in order to unite these two areas of knowledge. The greatest masters in the study of international relations tend to unify History and Theory by using the evidence and the hypothesis, the past and the present, the internal and the external, the empiric and the theoretic, the comprehension and the decision.

Cooperation between history of international relations and social sciences has led people to believe that the evolution of this study area, taken the increasing interdependence observed in international life after the Cold War and the North–South dialog, requires new method adjustments and problem identification. Nation-state and national interest conceptual categories, which served scientific development in the past, are challenged by new concepts such as global governance and universal citizenship. In this sense, why not request that the British equip their *concepts* of international system and international society in order to follow the transformations of the study field and to step on the terrain of global responsibilities? Why not expect from Italians the application of new *ideas* and systems of *values* to fill the gap left by the collapse of ideological imperialism which impregnated the world during the Cold War period? The versatile and rapid North Americans, just as the structuralist South Americans, might be required to provide *instruments of analysis* and explanations to phenomena such as international violence and oppression, solidarity and indifference, exploitation and cooperation, the exercise of power and self-determination of peoples. Also, the French may advance towards the consolidation of multiple causes as an applicable *method* to the study of factors and agents on a perspective of interdependence. And shouldn't the Germans, in their turn, heirs of the rankean tradition, seek the combination of local impulse and universal impact of international forces over the shaping of an unequal world?

Historical knowledge bears strong consistence due to the empirical analytic method which establishes the primacy of induction over deduction. Nevertheless, it is imperative to rethink History continually and to provoke ruptures of interpretation as did German Fritz Fischer concerning First World War origins, and some North American historians regarding the role of the United States in delineation of the Cold War order.

This article has summarized the history of international relations by emphasizing the contribution of the dynamic groups within the key countries and regions. But this has not meant the exclusion of pluralism inside these countries nor common guidelines which exist among different national backgrounds. It has elected some names and some schools or interpretation currents of international relations history in intending to define its field of study, trace its evolution in the last two centuries, and underline its relevance in the ensemble of social and behavioral sciences. Beyond those whose names were mentioned, hundreds of others add the result of their works to the historians' rich and diversified contribution. Thus, come to mind the historiography from other European and non-European countries such as Russia, Australia, Japan, India, and China. For pedagogical purposes, the history of international relations should first maintain its reference to history of civilization, and second, it should purify itself and absorb the dialogs among historians and between them and other experts in social and behavioral sciences on a global scale.

See also: Balance of Power, History of; Cold War, The; Contemporary History; Diplomacy; First World War, The; Frontiers in History; Geopolitics; Globalization: Political Aspects; Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); International Communication: History; International Justice: Philosophical Aspects; International Law and Treaties; Military History; Nationalism, Historical Aspects of: The West; Nations and Nation-states in History; Political History (History of Politics); Ranke, Leopold von (1795–1886); Second World War, The; War: Causes and Patterns; Warfare in History

Bibliography

Bull H, Watson A 1984 *The Expansion of International Society*. Oxford University Press, Oxford, UK

Cervo A L, Rapoport M 1998 *História do Cone Sul*. Editora Revan, Rio de Janeiro, Brasilia
Duroselle J B 1992 *Tout empire périra; Théorie des relations internationales*. A Colin, Paris
Duroselle J B 1993 *Histoire diplomatique de 1919 à nos jours*. Dalloz, Paris
Di Nolfo E 1994 *Storia delle relazioni internazionali, 1918–1992*. Laterza, Rome
Gaddis J L 1997 *We now know. Rethinking Cold War History*. Oxford University Press, Oxford and New York
Girault R 1998 *Être historien des relations internationales*. Publications de la Sorbonne, Paris
Milward A S (ed.) 1993 *The Frontier of National Sovereignty: History and Theory, 1945–1992*. Routledge, London and New York
Renouvin P 1994 *Histoire des relations internationales*. Hachette, Paris
Vigezzi B 1991 *Politica estera e opinione pubblica in Italia dall'Unità ai giorni nostri* 1st edn. Jaca Book, Milan
Watson A 1992 *The Evolution of International Society. A Comparative Historical Analysis*. Routledge, London and New York

A. L. Cervo

International Relations: Theories

International relations analyzes the strategic interaction among countries within the framework of an international system. As such, it involves issues in the interdependence of unit and system, common to many problems in the natural and social sciences. States (or other units) interact in a situation of anarchy—the absence of a supranational government able to enforce agreements. Theorists of international relations disagree strongly on the impact of anarchy. Does it lead irresistibly to conflict as nations seek to protect themselves, or to cooperation as nations see the benefits of peace? All elements of the interaction are contested in debates, from the definition of the unit (is it the State, or elements of society, or transnational forces beyond the State) to the properties of the system (anarchic or structured, culturalist or materialist). The notion of strategy is widely shared—whatever the units are, they must decide on a course of policy where they must anticipate the actions of other units—but the influences upon strategy are in sharp dispute.

The disputes can be sorted into two major lines of cleavage: one set of disputes focuses on the nature of the system, the other on domestic influences on foreign policy.

1. *System Theories*

At present, system theory is split among three contending schools: realism, neoliberal institutionalism, and constructivism. For realists, whose leading theorist in recent decades has been Waltz (1979), anarchy compels nations to behave in ways that may lead to conflict. Without centralized authority, nations face the security dilemma. Operating in uncertainty, they cannot be certain if they will be physically attacked by neighbors, so they take action to protect themselves, such as building up their armed forces. This frightens other countries which in turn cannot be sure whether the military will not be used for offense, so they have to arm as well. The results are arms races, bristling militaries, and a propensity to fight—the condition of 1914. All states, regardless of type (democratic, peace-loving, moral, or the contrary of each) are driven by this logic to prepare for war. Only deterrence, only power balancing, can therefore keep the peace by making war too costly. Nations cooperate only if it is in their obvious self-interest to do.

In contrast, neoliberal institutionalists argue that nations can and do cooperate in conditions of anarchy. Conflict under anarchy is often the result of the problems of collective action. International institutions can overcome these problems: they may reduce transaction costs, provide incentives to trade concessions, create mechanisms for dispute resolution, facilitate information sharing, and supply processes for making decisions, and, through frequent iterations, build trust through frequently observing such mechanisms. Conflict is thus a pole on a continuum, not an inevitable outcome of anarchy. Institutions can increase cooperation even without coercive power over the units. This body of ideas, championed by Keohane, (1984) has been used to explain the substantial growth of international institutions such as WTO (World Trade Organization), the IMF, (the International Monetary Fund) the European Union, and numerous regulatory 'regimes' such as those that supervise international aviation, post services, telecommunications, and others.

A third school, constructivists and sociological theorists, complain that both Keohane and Waltz neglected the nonrationalist and nonmaterial aspects of the interaction of units: culture, ideas, values, the internalization of norms, the constitutive elements of identity, and all the tissue of human exchanges and cultures. For constructivists, the interaction of States is socially constructed: 'anarchy is what you make of it.' The units form understandings about the world and norms of conduct which guide action. Strategic interaction, for such theorists, is a cultural construction. States internalize from world society a set of norms about their goals and proper practices that shape their behavior. They learn from international society how to act: what is acceptable, what not. Discourse and communication shape behavior. Colonization and slavery were abandoned when world culture turned against them. Notice that this group of theorists stress system, not domestic politics. It is not domestic culture which constrains States here, but the culture of the 'system.'

The constructivist approach is particularly strong among European researchers, while the realist and neoliberal institutionalists approaches are strong in the US. This is to some degree a historical reversal. *Realpolitik* arose in Europe, while the Americans were strongly influenced by moralist thinkers, such as Woodrow Wilson. It may also express changing problems: the European countries in the year 2000 examine the multidimensional issues of integrating traditional national cultures into a new institutional form, the European Union; the US thinks about a global role of security and trade.

2. *Unit-centered Theories*

Unit-centered theories argue that countries have choices. The system is not wholly constraining. Most of the time, for many States, especially the large ones, the system provides some 'slack,' some possibility for alternative responses. This creates a need for theories that explain the choice among alternatives allowed by the system. Three can be noted here: individual cognition and culture; institutions and political system; and interest groups.

2.1 *Cognition and Culture*

As countries engage in strategic interaction, their leaders must evaluate complex situations. Is Japan's home Defense Force purely for protective purposes or could it threaten others? Theorists of perception, such as Jervis, examine the ways individuals process information. They see 'misperceptions,' such as intelligence failures, arising from the limitations of our cognitive capacity, and from the rational strategies people devise for managing those limitations. The need to use analytic priors, for example, to sort out 'noise' from reality, creates a vulnerability to miscalculation. Information cognition theorists stress rationality. They stress what all people have in common, as human beings, that shapes their strategic behavior.

By contrast, cultural theorists stress the differences among people. The most common version of this approach examines the culture of a nation or a civilization. Theorists of this kind argue that Germany or Japan could go to war in the future because the culture of these nations is 'militaristic.' Huntington (1996) argues that cultural differences form the basis of conflict in the coming years, as civilizations struggle to protect and defend their core values.

2.2 *Political Institutions*

Decision-making systems influence outcomes. The way in which decision making is structured shapes the content of what comes out. If preferences are held constant, varying the institution varies the outcome. To understand foreign policy we must analyze the institutions within countries.

A major distinction is between democracies and authoritarian systems. Woodrow Wilson is famous for the theory that war is caused by authoritarianism. WW II and the events that followed discredited this view in favor of realism, but it has revived in recent years with a vigorous literature arguing that democracies do not fight each other.

Another line of reasoning examines variance among democratic institutions: the relationship of executive to legislatures, voting rules and party systems, bureaucratic structures and organizational processes, the structure of the armed forces and its relationship to civilian authorities. All of the institutionalist arguments have in common that they attribute causal effect in making foreign policy to patterns of decision-making.

2.3 *Interest Groups and Social Structure*

A third line of domestic politics reasoning explains foreign policy by the demands made on governments by the social groups on which governments rely for support. Governments need backing in society. In democracies, they need to win elections and enjoy the confidence of key players in the economy; in authoritarian systems, they need the support of key members of the 'selectorate' or elite with control of major power centers. Policy will reflect the preferences of these groups.

International economic policy, concerning such issues as tariffs, trade treaties, and common markets, is examined in this approach as a function of social cleavages around the beneficiaries of policy change: competitive industries support free trade, uncompetitive ones oppose it. Similarly, defense policy has been analyzed in reference to lobbies: defense spending reflects the preferences of a military–industrial complex, a lobby that fights for contracts and a big establishment which confers influence and power.

These three approaches to domestic politics sustain research programs that call attention to quite different processes and types of evidence: on cognition, perception, psychological mechanisms, values, cultural systems, fables, myths, analogies; on the institutions of government, elections, party systems, bureaucracies, voting rules, procedures, agreements and mechanisms; and on the structure of society, the economy, religious groups, business associations, trade unions, and professional groups.

3. *Cross-cutting Approaches*

Several lines of research seek to cut across the divide between unit and system, seeking to analyze the

interaction between them. One approach examines the impact of system upon the internal properties of the unit itself. This effect has been called 'the second image reversed.' Another variant looks at the interaction of the politics of the two dimensions, or 'two-level games.'

Another brand of theory challenges the centrality of the State in the system–unit approach. The state can be seen as but one of many actors in the international arena: firms, churches, cultures, NGOs, all comprise 'transnational forces' which operate in the world, and which may constrain state behavior or supplant it altogether. Thus there are different actors operating in the world, with no need to privilege the State. Theorists in the English School posit these transnational forces as comprising a kind of 'civil society' at the international level. They are not part of formal government, either national or international, but part of the culture, organizations, and structures that define the context within which governments operate. These arguments seek to de-center the State, as well as to de-center politics, economics, and military/security variables which tend to dominate most writing on international relations.

3.1 Law, Institutions, Norms

The failure of the League of Nations to prevent WW II undermined belief in the relevance of international law to the study of international relations. In recent years, interest in law has returned. International institutions and agreements, especially on fields concerning trade and regulation, the spread of the European Community, and the revived interest in democracy and norms have all contributed to this revival. Countries appear increasingly to delegate authority to institutions which operate in a judicial framework and process. Dispute resolution is shifted away from political processes—direct negotiation by interested parties operating from positions of power—toward formalized procedures in structured institutions administering codes of law or agreements. The European Court of Justice is the most developed example of this kind, but others are evident as well. Specialists dispute whether there is more 'legalization' or rather a shift in different issue areas, more in some, less in others, and what the various forces may be at work that shape these patterns.

3.2 Civil Wars, Disintegration; New Political Forms and Integration

At the frontier of the interaction of unit and system is the transformation of the units. Some countries disintegrate as civil war or other processes erupt. Austria–Hungary, the colonies of Britain, France and the Netherlands, the USSR no longer exist, giving way to their component entities. At the same time, other processes appear to be forming new units, of a kind without clear name or label. Economic and political integration have formed the European Union; which is not quite a state nor an international government. It has some properties or functions of many different categories. Is it a harbinger for similar destinies in other parts of the world? Will there be more such entities in Latin America, Southeast Asia, Africa? Answering these questions requires the integration of international relations with issues of domestic politics, of transnational forces and local ones, of culture and rational action, of strategic action and unilateralism, of security and economics.

3.3 Globalization

World trade, the flow of capital, peoples, and culture, gives rise to debates about globalization: is the world more integrated and interdependent, if so why, is this good or bad, and who benefits from it? The processes at work involve countries, as it is they who make the policies that encourage or inhibit globalization, and transnational forces, which may influence the internal preferences of countries and the contexts in which they operate. The debates on this topic involve all the core conflicts of this field.

See also: Action, Collective; Balance of Power, History of; Balance of Power: Political; Conflict/Consensus; Conventions and Norms: Philosophical Aspects; Deterrence; Foreign Policy Analysis; Global Governance; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Institutionalism; International Arbitration; International Communication: History; International Law and Treaties; International Organization; International Trade: Economic Integration; International Trade: Geographic Aspects; Liberalism and War; Monetary Policy; National Security Studies and War Potential of Nations; Nations and Nation-states in History; Norms; Realism/Neorealism

Bibliography

Allison G T 1971 *Essence of Decision: Explaining the Cuban Missile Crisis*. Little, Brown, Boston

Goldstein J, Kahler M, Keohane R O, Slaughter A-M (eds.) 2000 *Legalization and World Politics, International Organization*, 54 (3) Summer.

Gourevitch P A 1978 The second image reversed: International sources of domestic politics. *International Organization* **32**(4): 881–911

Huntington S P 1996 *The Clash of Civilizations and the Remaking of the World Order*. Simon and Schuster, New York
Jervis R 1976 *Perception and Misperception in International Politics*. Princeton University Press, Princeton, NJ
Jervis R 1978 Cooperation under the security dilemma. *World Politics* **30**(2): 167–214
Katzenstein P J 1996 *Cultural Norms and National Security: Police and Military in Postwar Japan*. Cornell University Press, Ithaca, NY
Keohane R O 1984 *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton University Press, Princeton, NJ
Lake D A, Powell R (eds.) 1999 *Strategic Choice and International Relations*. Princeton University Press, Princeton, NJ
Putnam R D 1988 Diplomacy and domestic politics: The logic of two-level games. *International Organization* **42**(3): 427–60
Rogowski R 1989 *Commerce and Coalitions: How Trade Affects Domestic Political Alignments*. Princeton University Press, Princeton, NJ
Waltz K 1979 *Theory of International Politics*. McGraw-Hill, New York
Wendt A 1999 *Social Theory of International Politics*. Cambridge University Press, New York

P. Gourevitch

International Research: Programs and Databases

The availability and diversity of databases for international research in the social sciences increased significantly in the latter half of the twentieth century. Social scientists use international databases constructed from data obtained in many ways, ranging from multinational sample surveys to Earth observation images. They are used to compare the behavior of both individuals and institutions by examining economic, social, spatial, environmental, or political change across many countries. International databases can consist of quantitative or digital observations describing a group of nations, a multinational geographic region, or even the Earth as a whole. Comparability of the data is generally achieved through the repeated use of the same data collection instrument, such as a survey questionnaire, in a number of countries, or by combining comparable national statistics or other records into a single data set.

The creation of global or multinational databases and their use in research are related to developments both within and outside the social sciences over the past 50 years. Among these are changes in the theory and focus of social science research, the emergence of international institutions and research programs capable of assuming responsibility for obtaining and maintaining large data collections, and advances in computers and the technology of data collection, analysis, management and distribution (see, e.g., *International Science: Organizations and Associations*). Today's diversity in and widespread availability of international research data are a relatively new phenomenon. In the early twentieth century, quantitative social science analysis was frequently focused on the urban neighborhood. Often linking their research to social reform activities, social scientists obtained data by conducting intensive surveys of the housing, working, and family conditions of the poor or working-class residents of specific neighborhoods. Because the physical area of these urban neighborhoods was small, social scientists could map the data manually, thus obtaining a spatial component for their analysis (Kellogg 1914).

By mid-century, quantitative social science had broadened both its theoretical focus and its data collection methods, in part because of improvements in the techniques and the instruments used for data collection. The use of sampling, developed in the 1930s and 1940s, permitted social scientists to construct databases on national populations through the use of questionnaires distributed to a probability sample of the population. Information obtained from the members of the sample was then used to construct inferences about the population from which the sample was taken. Data obtained from probability samples of a population were not geo-referenced, although the sampling framework was often based on geographic criteria. The use of data obtained from samples of a larger population freed social scientists to examine members of a variety of social groups, regardless of their geographic location. The use of the questionnaire as a data collection instrument led to an analytic emphasis on subjective data, such as individual opinions, perceptions, and self-reported behaviors. It also reinforced the growing interest in the analysis of national or macro-social phenomena.

In the immediate postwar period, there was an increased policy and research emphasis on comparing economic development among countries. Prewar research on national income accounting by Simon Kuznets became the foundation upon which the United Nations created a framework for comparable national income accounting in 1952. These accounts, once standardized and available for many countries, constituted a multinational database for economic research as well as for policy. Shortly after this, Deutsch (1960) attempted to identify comparable non-economic as well as economic data for international research into what he called the capability and stability of nations. Significantly, the basic unit of analysis for this work, as for the national income accounts, was the nation state, and the multinational data required for analysis contained one observation per variable per state.

The new international institutions that were formed in the wake of World War II assumed responsibility for maintaining large databases comprising data on each member nation. Although the League of Nations

and the International Labor Organization had obtained national data to create international databases beginning in the 1920s, it was the creation of the United Nations and its associated institutions that provided the greatest stimulus to the development of a wide range of multinational databases. The United Nations itself stimulated member countries to conduct decennial censuses and then maintained international databases on a number of socioeconomic conditions. Specialized United Nations agencies collected data on more specific topics (e.g., the Food and Agriculture Organization, the World Health Organization, UNESCO, and others). In addition, regional economic organizations such as the Organization for Economic Co-operation and Development (OECD) published economic and social data on member states and sponsored international meetings to define data elements that were comparable from country to country. Similarly, the multilateral financial institutions, such as the World Bank and the International Monetary Fund, created and disseminated multinational economic and social databases. Nongovernmental organizations also began to produce international databases, such as the annual list of military equipment produced by the International Institute for Strategic Studies in the United Kingdom. Since that time, these international organizations have continued to update and maintain a wide range of databases for international research and analysis.

A somewhat different approach to the creation of international databases was taken by Almond and Verba (1963) in their study of political culture. (See also *Databases, Core: Political Science and Political Behavior*.) Instead of using countries as the basic unit of analysis, they were interested in comparing what was taking place within countries. They conducted sample surveys of the population in five nations (Germany, Italy, Mexico, the United Kingdom, and the United States) to examine political attitudes and democracy. The result was an international database of five noncontiguous countries that provided social scientists with far greater breadth and depth of internal or national data than could be obtained using the types of databases maintained by multilateral organizations with a single datum per variable per nation. Other international databases of this type include the Level of Living or social indicator surveys coordinated by the OECD in the 1960s and 1970s; the World Fertility Survey and the International Social Survey Project (ISSP), which used the same questionnaires in a number of countries to obtain comparable international data; and the Luxembourg Income Study (LIS), which is an international data collection comprising dissimilar national surveys that measure a single phenomenon (income distribution) over time. Like the economic databases which resulted from basic research on national income accounting in the 1930s, these international databases were extensions of data collections that were defined and initially obtained for theoretical or policy purposes. In most cases, the data originated as national databases and were subsequently combined with other data sets to create an international database. Social scientists increasingly turned to comparative or international research—and the use of international databases—to illuminate and extend their understanding of complex processes and socioeconomic relationships within nations. (See related entries *Databases, Core: Sociology*; *Microdatabases: Economic*; *Statistical Systems: Economic*; *Statistical Systems: Labor*.)

In the late 1980s, a new type of research—and new types of international databases—began to emerge in some areas of the social sciences. For example, in the physical and ecological sciences, international research programs on global environmental change resulted in a growing appreciation of the environmental significance of socioeconomic and behavioral phenomena. Increasingly, scientists recognized that anthropogenic forces were responsible for a series of interrelated changes in the Earth's environment and, in turn, that global environmental change could have a wide range of impacts on human populations. The concept of global-scale research issues, together with the international salience of environmental problems and the broad research agenda developed in international scientific programs, led social scientists to consider developing global-scale rather than multinational databases.

At the same time, new technological developments made it easier to create and use global databases. Advances in computers and software made it possible for individual social scientists to obtain and analyze very large databases. The advent of geographic information system (GIS) software provided a means of spatially integrating diverse types of data; the availability of satellite remote sensing on a global scale has provided Earth observation images that are used to create global-scale databases; and advanced scientific computing has provided a means of managing and analyzing the data. Among the earliest global databases were those developed with remote sensing data, such as the so-called 'City Lights at Night' data, which are based on images obtained through the Operational Linescan System of the Defense Meteorological Satellite Program (DMSP). Satellite instrumentation was able to observe faint sources of visible near-infrared emissions on the surface of the Earth, specifically, the light from human settlements, fires, and gas flares. The global DMSP image of night-time lights provides a dramatic image of the density and spatial distribution of human settlements on Earth. However, satellite imagery of night-time lights provides only a rough guide to the distribution of human populations and tells little about what takes place within settlements. In 1992, a report to the International Social Science Council's Standing Committee on the Human Dimensions of Global Environmental Change recommended the creation of a gridded population database for use

with remote sensing data (Clarke and Rhind 1992). The first Gridded Population of the World (GPW) database, prepared in response to that report, was released in 1995. Although the data were originally obtained from national censuses, the GPW database displays population as a function of space rather than political units. It can provide either the number of people in 2.5-minute by 2.5-minute latitude/longitude cells across the surface of the Earth or the density of settlement (persons per square kilometer), and serves as a prototype for other types of gridded or vectorized socioeconomic databases.

Not only do the new computer capabilities make it possible to obtain the data required by broadening theoretical perspectives in the social sciences, but, at times, the databases themselves stimulate changes in these perspectives. Social scientists now have at their disposal a broader range of comparable international data than in the past and far greater flexibility in analysis, visualization, and manipulation of these data. New developments in data management and analysis, such as data mining, combined with this rich diversity in international databases, constitute a valuable and rapidly expanding component in the research infrastructure of the social sciences.

See also: Data Archives: International; Databases, Core: Demography and Registers; Information Society; International Communication: History; International Science: Organizations and Associations; Science and Technology: Internationalization

Bibliography

Almond G A, Verba S 1963 *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton University Press, Princeton, NJ

Center for International Earth Science Information Network (CIESIN) 2000 *Gridded Population of the World (GPW)*, Version 2. Available from: http://sedac.ciesin.columbia.edu/plue/gpw

Clarke J I, Rhind D W 1992 *Population Data and Global Environmental Change*. International Social Science Council, Paris, France

Deutsch K W 1960 Toward an inventory of basic trends and patterns in comparative and international politics. *The American Political Science Review* **54**(1): 34–57

International Social Survey Program 2000 Available from: http://www.issp.org

Kellogg P U 1914 *The Pittsburgh Survey: Findings in Six Volumes*. Russell Sage Foundation, New York

R. B. Miller

International Science: Organizations and Associations

The formal infrastructure to support and promote international collaboration in the social and behavioral sciences consists of several associations based on disciplines or subdisciplines, the International Social Science Council (ISSC), and to a certain extent the International Council for Science which was formerly known as the International Council of Scientific Unions (ICSU). This infrastructure is largely a product of the period since World War II. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) played a major role in encouraging the establishment of the infrastructure and supporting it. Participation in the associations and their activities has broadened and become more global, and activities sponsored by the infrastructure have increased over time. Despite this, at the dawn of the twenty-first century the infrastructure designed to support the social and behavioral sciences remained weaker than that designed primarily to support the natural sciences and there were fewer international collaborative programs.

1. History

The first steps toward the creation of science organizations and associations for the social and behavioral sciences were taken in the nineteenth century (Baker 1992, Greenaway 1996). These steps included both disciplinary based and broader developments. Geographers held their first international congress in 1871. This eventually led to the formation of the International Geographical Union (IGU) in 1922. The Association for the Reform and Codification of the Law of Nations was founded in 1873. This eventually became the International Law Association. Psychologists held their first international congress in 1889. At their third congress in 1900, they created a continuing committee, the International Congress of Psychology, to organize subsequent congresses.

In 1899 nine national academies formed the International Association of Academies. In 1919 this was transformed into the International Research Council, which in 1931 was succeeded by the International Council of Scientific Unions (ICSU). ICSU was comprised of 40 national members—national academies—and eight scientific unions, one of which was the International Geographical Union (IGU). The IGU was the only social science association among the eight. Because ICSU's forerunner was an organization of national academies and because it had national academies as members, the ICSU was supported by funds from national governments from the outset.

In addition to the IGU, two other social science associations were established in the inter-war period, the International Union for the Scientific Investigation of Population Problems in 1928 and the Standing International Commission of the International Congresses of the Administrative Sciences in 1930. The former changed its name to the International Union for the Scientific Study of Population in 1961, and the

latter became the International Institute of Administrative Science in 1977.

Despite these steps, there was relatively little international collaboration in the social and behavioral sciences prior to World War II compared to the collaboration that occurred in the natural sciences, and while the social and behavioral sciences were firmly established in North America and Western Europe they were weak or nonexistent in most countries in other parts of the world.

UNESCO was created in 1945. Its constitution directed it to support science and international scientific collaboration. The social sciences had a special place in UNESCO (Laves and Thomson 1957, Sewell 1975). UNESCO's Constitution begins with the assertion '… since wars begin in the minds of men, it is in the minds of men that the defenses of peace must be constructed.' There was a strong conviction among many of those who participated in the creation of UNESCO and the development of its early programs that social science could contribute to the promotion of peace, social justice, and human dignity. They believed that the development of social science and international collaboration among social scientists was essential to realize these aims.

They were generally unwilling to recognize that scholars and many government leaders questioned whether the social sciences had, or could, or should have the same universal qualities that the natural sciences had. As the Cold War developed and tensions grew between developed and developing states, many governments and a few scholars sought to enlist UNESCO to support their causes (Dunn 1950). Although these broader conflicts have plagued UNESCO and its own programs, they have largely been avoided in the organization's efforts to promote and support international science organizations and associations.

In the early years after World War II, UNESCO promoted the formation of international disciplinary associations. With UNESCO's encouragement the World Association of Public Opinion Research was created in 1947; the International Union of Anthropological and Ethnographical Sciences, and the World Federation of Mental Health in 1948; the International Economic Association, the International Political Science Association, and the International Sociological Association in 1949; the International Committee of Comparative Law in 1950, which became the International Association of Legal Science in 1955; and the International Union of Psychological Sciences in 1951. UNESCO provided subventions to the newly established associations.

At the same time that it encouraged the formation of disciplinary associations UNESCO sought to establish a facility that would promote social sciences generally and collaboration among them. UNESCO convened two consultative meetings in 1951 to recommend a course of action. The second consultative meeting in December 1951 unanimously recommended that an International Social Science Council should be created, a recommendation that UNESCO's General Conference endorsed in a resolution. The ISSC was established the following year at an UNESCO sponsored meeting.

ISSC was created as a confederation of five disciplinary based international associations. When the constitution was revised in 1972 it became a federation. By the 1990s the number or member associations had settled at 14. Beyond the 12 associations mentioned above these included the International Peace Research Association, which was founded in 1964, and the International Federation of Social Science Organizations, which was founded in 1979.

ISSC also included 15 associate members. Among them are such bodies as: l'Association Internationale des Sociologues de Langue Française; the International Association for Applied Psychology; the International Association of Economic History; the International Committee for Social Science Information and Documentation; the International Federation of Data Organizations; the International Society of Criminology; the International Studies Association; and the World Future Studies Federation.

Unlike ICSU, ISSC did not include national or regional members until its constitution was amended in 1992. Despite this constitutional change, national academies have not rushed to join ISSC. As of 1999 there were only eight national and three regional members, and none of these was from Japan, North America, or Western Europe, the affluent areas of the world most able to support scientific collaboration. In contrast, eight of the nine academies that founded the International Association of Academies, the forerunner of ICSU, were from Western Europe and the ninth was the United States National Academy of Sciences. In 1999 ICSU had 98 national members from 96 countries.

The founders of ISSC feared that including national members might subject the council to ideological and political pressures, and when the possibility of having national members was raised again in the 1970s this fear was even more pronounced (Baker 1992). In addition, the place of social and behavioral sciences in national structures varies substantially. Which if any national body, would be an appropriate national member, was not as clear as it was with respect to the natural sciences. Not having national members has been one of the factors why ISSC has not had the same access to funds from national governments that ICSU has had. UNESCO's subvention has always been ISSC's predominant source of funds.

2. *Activities and Accomplishments*

As the twenty-first century opened, most of the formal infrastructure to support international collaboration

in the social and behavioral sciences had been in place for half a century or longer. Many things happened during this period (see IPSA 1999).

The disciplinary associations grew. They were comprised of both collective members—national associations—and individual members. They were founded by as few as a handful of national associations. By 1999 most of the 14 member associations of ISSC had from 40 to 60 collective members. This growth represented a considerable increase in social scientists in areas other than North America and Western Europe. In addition, the largest associations had from 1,000 to 2,000 individual members from about 100 states. While national associations from North America and Western Europe accounted for only about 40 percent of the collective members they accounted for from two-thirds to three-quarters of the individual members.

When ISSC was founded, UNESCO gave ISSC funds for its own operation and to disperse to the disciplinary associations. These funds were dispersed as block grants. This system continued until 1996. Since 1996 UNESCO continued to provide funds for ISSC and the disciplinary associations, but the funds were provided for projects only. Beyond providing a subvention, since ISSC's creation UNESCO has given it contracts to execute projects and office space in UNESCO House. ISSC has never had any other space.

Following the tradition set by the geographers, international legal specialists, and psychologists in the nineteenth century, the disciplinary associations have organized congresses. The frequency of the congresses varies from once a year to once every three or four years. Participation in congresses has grown to 1,500 or more. The disciplinary associations also convene smaller meetings, and sponsor roundtables, study groups, and research committees. They publish abstracts, journals, and newsletters and sponsor book series. By the end of the twentieth century, the disciplinary associations had come to manage substantial scientific exchanges.

ISSC's activities have been constrained by its limited funds. In its early years, ISSC focused on providing services for its member associations and for UNESCO. Most of ISSC's early projects and publications derived from UNESCO contracts. Both UNESCO and ISSC proposed topics for the projects. There always has been a close connection between UNESCO's social science program and ISSC's activities.

In the 1980s ISSC took steps that led to the creation of two broad international interdisciplinary research programs, the International Human Dimensions Program (IHDP) on Global Environmental Change and the Comparative Research Program on Poverty. Sponsoring large-scale research programs was a new venture for ISSC. These programs required raising additional funds granted specifically to support them.

Social scientists and social science associations also have connections with ICSU. Three social science associations are among its 25 scientific union members. The IGU was a founding member. The International Union of Psychological Science became a member in 1982, and the International Union of Anthropological and Ethnological Sciences was admitted in 1993.

ICSU was established before UNESCO was. Although UNESCO's financial assistance was crucial in creating ICSU's first professional secretariat, and ICSU has regularly received funds from UNESCO, UNESCO has not been the predominant source of ICSU's funds. The ICSU secretariat has been located in Paris since 1972, but it has had its own building.

One of ICSU's first activities was to sponsor a large-scale international collaborative research program, the Second International Polar Year in 1932–3. Starting with the International Geophysical Year in 1957, ICSU has sponsored many international collaborative research programs in the period since World War II. The largest of these in 2000 was the International Geosphere-Biosphere Program, A Study of Global Change, which began in 1986. Other programs that ICSU co-sponsors include The World Climate Research Program, the IHDP (with ISSC), and the Scientific Committee on Problems of the Environment.

ICSU could build on earlier practice. The first International Polar Year was held in 1982–3. There is great pressure for international collaboration in the natural sciences. Natural science frequently requires data from all over the world, and these data must be collected according to common protocols and measured according to identical scales. There is a deep tradition of unfettered exchange of knowledge in the natural sciences.

The three social science associations that are members of ICSU have participated actively in ICSU's international collaborative research programs. There have been discussions about additional international social science associations becoming members of ICSU and about closer collaboration between ISSC and ICSU. In a step toward greater collaboration, in 1996 ICSU became a co-sponsor with ISSC of the IHDP. A merger of ISSC and ICSU would be difficult. Not all of the 11 ISSC member associations that are not members of ICSU would fit easily into ICSU.

As a consequence of the relative weakness of the formal infrastructure to support and promote international collaboration in the social and behavioral sciences, many collaborative efforts for training, data collection and archiving, and research have been conducted under the auspices of regional and national organizations such as the European Science Foundation, the Research Directorate-General of the European Commission, the Latin American Social Sciences Council, the US National Academy of Sciences, and the Social Science Research Council or under *ad hoc* arrangements.

Interestingly most US professional associations in the social sciences have many members who are not citizens or residents of the USA. For instance, in 1999 such individuals comprised 35 percent of the membership of the American Society of International Law.

How the infrastructure for supporting international collaboration in the social sciences would evolve was an open question. The infrastructure was created and utilized to promote the development of the social sciences in what frequently was a top-down process. Development of a more robust infrastructure would probably require bottom-up pressures from both the social and natural science communities and government support.

See also: Data Archives: International; Information Society; International Communication: History; International Research: Programs and Databases; Science and Technology: Internationalization

Bibliography

Baker F W G 1992 *Forty Years History of the International Social Science Council: The Evolution of the Social Sciences*. ISSC, Paris

Dunn F S 1950 *War and the Minds of Men*. Harper, New York

Greenaway F 1996 *Science International: A History of the International Council of Scientific Unions*. Cambridge University Press, Cambridge, UK

IPSA (International Political Science Association) 1999 Fifty Years of IPSA. *Participation* **23:1**, 4–13; **23:2**, 6–13; **23:3**, 4–12

Laves W H C, Thomson C A 1957 *UNESCO: Purpose, Progress, Prospects*. Indiana University Press, Bloomington, IN

Sewell J P 1975 *UNESCO and World Politics: Engaging in International Relations*. Princeton University Press, Princeton, NJ

H. K. Jacobson

International Trade: Commercial Policy and Trade Negotiations

'Commercial policy' describes any form of government intervention towards international trade. The study of commercial policy is a branch of international trade theory, itself a subfield of microeconomics. None of this sounds likely to arouse passions, but in practice trade policy has often prompted bitterly divisive political debates and has been a central concern of domestic and foreign policy.

In the early modern period, 'Mercantilist' writers rationalized the use of restrictive trade policy by expansionist monarchs to foster exports and ensure trade surpluses. David Ricardo in 1817 provided the intellectual case against this approach. His theory of trade patterns, based on specialization according to comparative advantage (see *International Trade: Economic Integration*), predicted that no free-trading country would lose relative to autarky. This laid the foundations of the modern enthusiasm for free trade on the part of most mainstream academic economists. However, Ricardo's arguments were probably less important than the increasing ascendancy of commercial over landed interests in leading to the repeal of the UK Corn Laws in 1848, which ushered in an era of (mostly) falling barriers and expanding trade. Twentieth-century world wars and depression reversed this trend and were themselves influenced by trade policy: commercial and imperial rivalries contributed to the onset of the First World War, and the 1930 Smoot–Hawley tariff worsened the depression in the US and hastened the collapse of world trade. Postwar attempts to restore the multilateral trading system led in 1947 to the General Agreement on Tariffs and Trade (GATT), reconstituted in 1995 as the World Trade Organisation (WTO). Under their auspices, successive rounds of trade negotiations have yielded progressive reductions in trade barriers. However, the chaotic events inside and outside the abortive Seattle meeting in 1999, which it had been hoped would launch another trade round, showed that progress towards further liberalization will face opposition from critics of 'globalization.'

As this brief background makes clear, a full account of the theory and practice of trade policy would require an extensive discussion of general intellectual currents, economic history, and contemporary international relations. This article has the more limited objective of summarizing what economic theory has to say on the topic. It begins with some necessary taxonomy and then reviews the principal theoretical arguments for and against trade restrictions.

1. Varieties of Commercial Policy

The most obvious form of commercial policy, and historically often the most important, is a *tariff*, a tax on imports which raises their domestic price above the world price, and so 'protects' domestic producers at the expense of home consumers. Confusingly, an export *subsidy* has a similar effect, raising the price of an export good to domestic producers and consumers above its world price. The two measures have opposite effects on the relative price of imports to exports, which is the basis of the Lerner symmetry theorem: a uniform tariff on all imports has exactly the same effects on relative prices as a uniform tax on all exports. Both raise the relative price of imports at home and thus discourage trade. A corollary is that trade can be liberalized either by reducing tariffs or by leaving them in place and subsidizing exports: politically a more expedient route and one followed successfully by some of the newly industrializing countries of East Asia.

Tariffs have declined in importance since the Second World War relative to nontariff barriers such as import quotas, 'voluntary export restraints' (i.e., quotas imposed by exporting countries), and government procurement rules. Such policies are qualitatively similar to tariffs in their protective effects, though the conditions for exact equivalence rarely hold. Finally, most domestic policies (taxes, subsidies, health and safety regulations, etc.), even if not explicitly discriminatory, have external repercussions. Though not strictly forms of commercial policy, their effects are increasingly recognized in trade negotiations.

Constructing a true measure of trade policy is an index-number problem: how to aggregate all these different types of trade restriction into a single measure which is comparable across countries and across time. Solutions in principle to it have been devised, but implementing them in practice is extremely difficult. In applied work, levels of protection are usually measured by trade-weighted average tariffs and, even less satisfactorily, by 'coverage ratios,' the percentage of traded commodities which are subject to nontariff barriers.

This article considers only the case where commercial policy applies indiscriminately to imports from whatever source. The desirability of this is enshrined in Article I of the GATT, which requires that all trading partners be treated as favorably as the 'most-favored nation.' However, Article XXIV allows exceptions for regional trade agreements, which have grown in importance in recent years with the widening and deepening of the European Union and the signing of the North American Free Trade Agreement. For further details, see *International Trade: Economic Integration*. The GATT also tolerates tariffs imposed on exporters found guilty of 'dumping'—selling below cost or below the price charged in their home market. Such 'anti-dumping' tariffs are an important form of protection in the contemporary world economy: even the threat of imposing them can deter foreign exporters.

2. *Trade Policy in a Competitive Small Open Economy*

Devising criteria for trade policy which will hold universally is a daunting task, and it makes sense to begin with a simple benchmark case. The classical starting point is an economy which is both competitive—individual consumers and firms cannot affect domestic prices—and small—the economy as a whole cannot affect world prices. Free trade must then maximize real national income, since it removes the constraint requiring an exact match between domestic production and consumption patterns. Specialization in production increases the value of aggregate output at world prices, while consumers benefit by being able to buy from the cheapest supplier worldwide.

However, individuals are both consumers and income recipients, and aggregate gains can mask big shifts in internal income distribution. The near certainty that there will be some losers is implied by the Stolper–Samuelson theorem. This was originally formulated for a special model, where it predicts that protection will raise real wages (so trade liberalization will lower them) if imports use labor relatively intensively. More generally, the logic of the theorem implies that there are almost always some factor-owners who will lose from a reduction in trade barriers. Most obviously, this will be true of factors which are specific (even if only in the short run) to import-competing sectors.

Losers notwithstanding, the existence of national gains from trade ensures that 'aggregate welfare' must rise, meaning that it would be possible to tax some of the winners' gains, compensate the losers, and still leave no one worse off. Free trade is thus the archetype of a situation which is potentially Pareto efficient or simply 'efficient' (confusingly, the term has a more precise sense than in common parlance). The same result holds even if the government has limited taxing and spending powers, and can only redistribute income through changes in commodity (or 'indirect') taxes. Of course, all this is poor consolation for the losers if the compensation is not actually carried out. Nonetheless, trade theorists tend to emphasize the efficiency gains, and prefer to try and devise programs of adjustment assistance to help those adversely affected rather than to recommend foregoing the national gains. In this they are motivated by professional division of labor (losses to particular groups mandate changes in the tax and social welfare system, not protection), and a belief that the poor rarely gain from highly restricted trade, rather than heartlessness. For the same reasons, the remainder of this article concentrates on the effects of trade liberalization on aggregate welfare, and will not repeat these essential qualifications about its distributional consequences.

Even though the case for free trade is clear, the best way to move towards it may not be (except in the trivial case where there is only a single tariff). Abolishing all tariffs at once is unlikely to be politically feasible. Two rules of piecemeal trade liberalization are then available. The first is the uniform reduction rule: reduce all tariffs by an equiproportionate amount. Heuristically, this kind of reform leaves relative tariff rates unchanged, so it is 'as if' there is only a single tariff rate, which is steadily reduced. Hence it is not surprising that (pathological cases apart) it guarantees a welfare improvement. The second is the concertina rule: reduce the highest tariff rate. A sufficient condition for this to raise welfare is that the good in question is a substitute for all other goods subject to tariffs. Substitutability is not necessary, however. For example, if all goods subject to tariffs are complements for each other, then a reduction in any tariff (not just the highest) raises

imports of all tariff-constrained goods and a welfare gain is again assured. Finally, the concertina rule does not justify increasing the lowest tariff, unless all exports are subsidised at higher rates: only raising the lowest *distortion* guarantees a welfare gain.

This discussion illustrates the distinction between 'first-best' and 'second-best' welfare economics. Policy recommendations are more complicated in the latter case, when some pre-existing distortions cannot be abolished. Nevertheless, a general principle applies: activities which from a welfare perspective are under-supplied in the absence of intervention should be encouraged and vice versa. A related rule of thumb with many useful applications in practical policy making is the principle of targeting: intervention should be applied as closely as possible to the desired target, whether this is to offset an irremovable distortion or to attain a 'noneconomic' objective (such as restricting imports of certain types of goods or protecting industries deemed essential to cultural independence or national security). From this perspective, trade policy is rarely a first-best instrument. For example, if there is a minimum wage in the import-competing sector, protection may raise welfare because it partially offsets the minimum wage. But other forms of intervention, such as employment or production subsidies, would have the same effect at lower welfare cost.

3. *Trade Policy in a Large Open Economy*

Relaxing the assumption that the economy is 'small' admits a specifically economic argument for protection which was clearly stated by Bickerdike in 1906. Reducing home demand for imports now lowers their world price, improving the home country's terms of trade (i.e., reducing the price of imports relative to exports) and yielding a welfare gain. The optimal tariff is the tariff which just balances this gain from manipulating the world price against the loss from trading at a different price from the rest of the world. A corollary is that, if a country is a major supplier of a good, and if local producers are competitive, an export tax can improve national welfare. In effect, the home government acts as a monopolist: a role which uncoordinated private producers cannot adopt by themselves. It should be stressed that these policies are only optimal from a national point of view (which explains why some authors prefer the term 'exploitative' to 'optimal'). World welfare definitely falls, though once again transfers, this time international, would be needed to compensate losers from universal free trade.

In practice, very few if any individual countries have a major influence on world prices. Cases where a group of countries acting together would have such power are more common, especially producers of primary commodities with few close substitutes. Hence the tendency for cartels and commodity price agreements in such markets. However, such groupings are typically unstable, since each member has an incentive to 'free-ride' on its partners' output restrictions. The Organization of Petroleum Exporting Countries is the outstanding counterexample and is probably best explained by a dominant firm model, where one large producer with very low marginal production costs (Saudi Arabia) in effect sets the price, and a 'competitive fringe' of other oil-producing countries (whether OPEC members or not) adjusts its outputs accordingly.

4. *Trade Policy with Economies of Scale and Imperfectly Competitive Markets*

The arguments reviewed so far assume that firms are perfectly competitive and produce in equilibrium with constant returns to scale. A great deal of research in recent years has relaxed these assumptions and explored the implications for trade policy.

A useful starting point is the case of Marshallian 'external economies': individual firms lack market power but expansion of the industry as a whole lowers costs for all. This isolates the implications of increasing returns, while retaining the assumptions of perfect competition. In such markets, industries which enjoy significant economies of scale are likely to concentrate in large countries under free trade. However, this need not justify protection by small countries, since this would condemn them to high-cost local production and to foregoing the gains from participating in the international division of labor. It is only medium-sized countries which are likely to lose from free trade by specializing in the 'wrong' commodities. Even for them the principle of targeting continues to apply: production subsidies rather than tariffs are the optimal policy.

Relaxing the assumption that firms are perfectly competitive complicates matters considerably, since there are many varieties of imperfect competition. It is convenient to distinguish between two: monopolistic competition and oligopoly. To begin with the former, it resembles perfect competition in two important respects: individual firms are too small to influence their rivals; and they enter or leave the industry in response to profit opportunities, so that in equilibrium profits are zero. The key distinguishing features are that firms enjoy economies of scale and that each produces a distinct variety. This in turn reflects consumers' tastes, which exhibit a preference for diversity. In this kind of economy, restricting trade has an additional harmful effect: it reduces the range of choice available to consumers. Policy choice is complicated because the free market may lead to more or less varieties being produced than the social optimum. However, the principle of targeting continues to apply: provided anti-trust measures are used if necessary to

ensure an optimal number of domestic firms, free trade remains desirable.

A different set of issues arises in the case of oligopoly, where there are barriers to entry and a relatively small number of firms. Since firms perceive that they are interdependent, they behave 'strategically,' taking into account the reactions of their rivals. In such markets, there may be scope for governments to intervene in favor of home firms, an idea which has come to be known as the theory of strategic trade policy.

The key insight is that, if a home and a foreign firm make their decisions simultaneously, the home firm cannot credibly commit to a level of output which would maximize home welfare conditional only on the behavior of the foreign firm. By contrast, the home government is assumed to be able to credibly commit to policies before both firms take their decisions. Hence there is scope for the home government to make the commitment on behalf of the home firm. In the simplest example, where a single home firm competes against a single foreign firm in a third market, the implications are dramatic. The optimal policy for the home government is to provide a positive export subsidy. This allows the home firm to credibly commit to more aggressive behavior, raising its output, market share, and profits at the expense of its foreign rival. This 'profit shifting' result seems to provide a rationale for the support of 'national champions.' Similar results apply to import-competing firms: tariffs may serve to raise welfare by shifting profits from foreign to home firms.

However, the profit-shifting argument turns out to be subject to many qualifications. The underlying model assumes that firms compete in the marketplace by choosing their outputs, taking the output choices of their rivals as given (i.e., that firms engage in Cournot competition). This sort of behavior is plausible when technology requires firms to commit in advance to their capacity output levels. If instead output can be varied with little change in marginal cost, then firms are more plausibly modeled as price-setters (engaging in Bertrand competition). In that case, firms behave more aggressively in the absence of intervention, and the optimal policy is an export *tax* rather than a subsidy. The rationale for intervention is the same in the two cases: the home government uses its superior commitment power to achieve an outcome which the domestic firm cannot achieve on its own. However, the practical relevance of the theory is reduced by the sensitivity of the actual policy prescription to assumptions about how firms behave. One slight defense of intervention comes from recent research, which suggests that the ambiguity is reduced when subsidies are given to pre-production variables such as R&D or marketing expenditures rather than directly to exports.

Other criticisms of strategic trade policy are that with many home firms there is a countervailing incentive to tax them (just as in the competitive large open economy case of Sect. 3); that the gains from intervention are more than offset if foreign governments also subsidize their own firms (see Sect. 6); and that general-equilibrium interactions with the rest of the economy are ignored. For example, if applied to a number of sectors, all of whom draw on a limited supply of some factor of production (such as skilled labor), an export subsidization policy merely raises the wages of that factor with little or no effect on the pattern of output or the level of national welfare. For these and other reasons, attempts to quantify the likely gains from strategic trade policy suggest that they are very small at best. However, the gains could be much larger if subsidies made it possible for a home firm to compete in the first place, especially if the alternative was domination of the world market by a foreign monopoly. (The huge subsidies to Airbus by European governments are often justified on these grounds.)

A final issue which arises when markets are imperfectly competitive is that trade policy can itself affect the degree of competition. With price in excess of marginal cost, firms are producing below their optimal scale. Hence any policy, including protection, which raises domestic output may increase welfare. However, as noted in Sect. 2, protection is not the best form of intervention. Exposing home firms to foreign competition is likely to be more effective, reducing prices to consumers and allowing any surviving home firms to produce at a more efficient scale.

5. *Trade Policy and Growth*

To assess the effects of trade policy in a growing economy, all the issues discussed in previous sections remain relevant, and some new ones arise. Until recent years, the standard approach to modeling economic growth was the neoclassical model of exogenous growth due to Solow. In that framework, the long-run or steady-state rate of growth is determined by exogenous rates of population growth and technological progress. Hence, trade policy cannot affect the steady-state growth rate, though it may affect the rate at which the steady state is approached.

More recently, attempts have been made to provide endogenous explanations for economic growth. These stress the importance of resources devoted to research and development (R&D), both in encouraging technological innovations and in facilitating the introduction of new and higher-quality products. They also emphasize the importance of externalities, as the benefits of R&D typically cannot be fully appropriated. This has immediate implications for trade policy. For example, if a sector of the economy is disproportionately engaged in R&D, protecting that sector will raise the long-run growth rate. These arguments are related to an older, 'infant-industry' argument, which defends transitional protection to enable a new firm to benefit from learning-by-doing

and scale economies. For such firms, a tariff, by guaranteeing higher home sales, may allow a firm to compete in export markets.

Once again, the principle of targeting must be mentioned. At best, these arguments justify production or R&D subsidies: since the industries involved are likely to be oligopolistic, they provide a case for strategic industrial rather than trade policy. Many of the qualifications noted in Sect. 4 continue to apply. Success is more likely to come from general policies which foster a culture of innovation and enterprise, rather than specific interventions which seek to 'pick winners.' A final caution is that, even when income distribution and externalities are ignored, GNP per head is not the same as welfare in a growing economy. Policies which raise the growth rate in the short or long term may do so at the expense of current consumption.

Given the theoretical ambiguities surrounding the effects of trade policy on growth, it is hardly surprising that empirical studies have failed to find a conclusive link between trade and growth. Some authors have shown that various measures of 'openness' can explain the relative growth performance of different countries, but such measures are not directly related to trade policy. The case against modest restrictions on trade is not proven. Nevertheless there are strong theoretical and empirical reasons for believing that countries, especially those with small home markets, which close themselves off from international movements of goods, factors, and ideas, are likely to have lower levels of welfare and growth.

6. *Institutions, Politics, and Trade Policy*

The discussion so far has confined attention to the case of trade policy which is set unilaterally by a single country. It has also adopted an exclusively 'welfarist' perspective in evaluating trade policy: the practical relevance of this is questionable unless decisions on trade policy are devolved to public-spirited bureaucrats. A variety of approaches has been taken to relaxing these assumptions. While much remains to be done, recent research has thrown light on the political pressures for and against protection, and on the role of international institutions such as the GATT in underpinning moves towards freer trade.

Many of the potential gains from unilateral intervention identified earlier must be qualified when it is recognized that foreign governments face similar incentives. Though often described as 'retaliation,' this phenomenon is usually modeled by viewing governments as players in a simultaneous-move game, often assuming only two countries for simplicity. The effects on welfare in the cases of optimal tariffs (from Sect. 3) and optimal export subsidies to Cournot firms (from Sect. 4) are similar: at least one country must lose relative to free trade and both may lose. The latter outcome (which must ensue if the countries are relatively similar) illustrates a 'prisoner's dilemma': unless the countries explicitly cooperate on free trade, each has an incentive to adopt an interventionist policy and the result is lower welfare for both. The case of optimal export taxes on Bertrand firms is different: now, both countries may gain relative to free trade (though only at the expense of consumers). Nevertheless, it exhibits a feature that holds in all cases: cooperation between governments leads to higher welfare than noncooperative choice of trade policy.

With or without international cooperation, all the theories of trade policy discussed so far are normative, assuming that national welfare is the primary policy objective. Hence they fail to explain why governments so frequently restrict trade by more than the welfare-maximizing extent. In response, the burgeoning field of political economy has proposed a variety of positive explanations for the pattern and prevalence of trade policy. It may be determined by spending on 'rent-seeking' by factors of production for protection of the sectors which use them intensively (or, in the limiting case of immobile factors, exclusively). It may be determined by direct voting on tariff rates, in which case the rate chosen will reflect the factor ownership of the median voter. Finally, it may be the outcome of campaign contributions to politicians by lobby groups, whether to increase the probability of their preferred party being elected, or to influence the policies of an incumbent government.

All these theories rely on special assumptions about the nature of the political system (often assuming that politicians seek to maximize some weighted average of welfare and political support) and the manner in which individual preferences influence political decisions. They also assume that trade policy is used for redistributive purposes, and for the most part they allow no role for other forms of public policy. They are therefore vulnerable to the criticism that they fail to explain satisfactorily why trade policy is the preferred instrument when more efficient methods of redistributing income are available.

Finally, just as the widespread use of trade policy is a puzzle, so also is the steady trend towards greater liberalization in the period since the Second World War. It can be explained by shifts in prevailing ideology, combined with the superior economic and political performance of those countries which were first to move towards open markets. But economic explanations have also been proposed. To the extent that trade is intra- rather than interindustry (which appears to be the case for much trade between developed countries), trade liberalization imposes lower costs of adjustment and so has less distributional impact. The GATT can be rationalized as a mechanism for implementing the international transfers needed to compensate countries for foregoing nationally optimal tariffs which would lower world welfare. Free-trade agreements of all kinds may serve as a commitment device: a government has an in-

centive to join them because they provide a way to credibly distance itself from the domestic pressure groups which lobby for protection.

7. Conclusions

In the benchmark case of a competitive, small, open economy, free trade must raise aggregate national welfare, although some individual groups will lose unless compensation is actually paid. Relaxing the benchmark assumptions allows exceptions to the case for free trade: 'optimal' tariffs to manipulate world prices; 'strategic' tariffs or export subsidies when home firms engage in oligopolistic competition with foreign rivals; and infant industry protection to allow home firms benefit from learning by doing. Protection can also raise the growth rate, though it is less likely to raise welfare in a growing economy. All these possible arguments for protection are subject to many qualifications. Moreover, on closer examination, most economic arguments for protection turn out instead to be arguments against *laissez-faire*, and so must be qualified by the principle of targeting: corrective measures should be applied as close to the source of the 'distortion' as possible, suggesting that other forms of intervention (such as R&D or production subsidies) are preferable to trade protection in most cases. Overall, with due allowance for some ambiguity, both theoretical arguments and empirical evidence suggest a pragmatic case for free trade.

8. Suggestions for Further Reading

Bhagwati (1988) and Irwin (1996) provide contemporary and historical background. Bhagwati (1971), Corden (1974), and Dixit (1985) give overviews of the theory of trade and welfare, using mainly prose, diagrams, and algebra, respectively. More recent updates are given in the contributions to Grossman and Rogoff (1995): see especially the chapters by Brander on strategic trade policy, Feenstra on estimating the effects of trade policy, Grossman and Helpman on technology and trade, Rodrik on political economy, and Staiger on rules and institutions for international trade. For more details and further references on particular topics, see Anderson (1992) on dumping and anti-dumping; Anderson and Neary (1996) on index numbers of trade restrictiveness; Bagwell and Staiger (1999) on the GATT; Dixit and Norman (1980), especially Sects. 3.2 (on redistribution through commodity taxation), 9.1 (on trade and competition), and 9.3 (on product differentiation and intra-industry trade); Ethier (1982) on trade policy under increasing returns; Grossman and Helpman (1991) on trade policy and growth; Krugman (1984) on import protection as export promotion; Neary (1995) on tariffs and quotas; Neary and Leahy (2000) on strategic trade and industrial policy; and Rodriguez and Rodrik (1999) on empirical studies of trade policy and growth. Among the many important topics not covered are effective protection (see Ethier (1977)) and the interaction of trade and environmental policy (see Neary (2001)).

See also: Economics, History of; International Trade: Geographic Aspects; Ricardo, David (1772–1823)

Bibliography

Anderson J E 1992 Domino dumping I: Competitive exporters. *American Economic Review* **82**: 65–83

Anderson J E, Neary J P 1996 A new approach to evaluating trade policy. *Review of Economic Studies* **63**: 107–25

Bagwell K, Staiger R 1999 An economic theory of GATT. *American Economic Review* **89**: 215–48

Bhagwati J N 1971 The generalized theory of distortions and welfare. In: Bhagwati J N, Jones R W, Mundell R A, Vanek J (eds.) *Trade, Balance of Payments and Growth: Essays in Honor of C. P. Kindleberger*. North-Holland, Amsterdam

Bhagwati J N 1988 *Protectionism*. MIT Press, Cambridge, MA

Corden W M 1974 *Trade Policy and Economic Welfare*. Oxford University Press, Oxford, UK

Dixit A K 1985 Tax policy in open economies. In: Auerbach A J, Feldstein M (eds.) *Handbook of Public Economics*. North-Holland, Amsterdam, Vol. 1

Dixit A K, Norman V 1980 *Theory of International Trade: A Dual, General Equilibrium Approach*. Cambridge University Press, London

Ethier W J 1977 The theory of effective protection in general equilibrium: Effective-rate analogues of nominal rates. *Canadian Journal of Economics* **10**: 233–45

Ethier W J 1982 Decreasing costs in international trade and Frank Graham's argument for protection. *Econometrica* **50**: 1243–68

Grossman G, Helpman E 1991 *Innovation and Growth in the World Economy*. MIT Press, Cambridge, MA

Grossman G, Rogoff K (eds.) 1995 *Handbook of International Economics*. North-Holland, Amsterdam, Vol. III

Krugman P 1984 Import protection as export promotion: International competition in the presence of oligopoly and economies of scale. In: Kierzkowski H (ed.) *Monopolistic Competition and International Trade*. Oxford University Press, Oxford, UK, pp. 180–93

Irwin D A 1996 *Against the Tide: An Intellectual History of Free Trade*. Princeton University Press, Princeton, NJ

Neary J P 1995 Trade liberalisation and shadow prices in the presence of tariffs and quotas. *International Economic Review* **36**: 531–54

Neary J P 2001 International trade and the environment: Theoretical and policy linkages. In: Cararro C, Siniscalco D (eds.) *Advances in Environmental Economics: Theory and Policy*. Cambridge University Press, Cambridge, UK

Neary J P, Leahy D 2000 Strategic trade and industrial policy towards dynamic oligopolies. *Economic Journal* **110**: 484–508

Rodriguez F, Rodrik D 1999 Trade policy and economic growth: A skeptic's guide to the cross-national evidence. Discussion Paper No. 2143, CEPR, London

J. P. Neary

International Trade: Economic Integration

Regional integration agreements (RIAs) are groupings of countries formed with the objective of reducing barriers to trade between members. They have long historical antecedents, sometimes being a stage to nation building (most famously in the Zollverein of nineteenth-century Germany), and sometimes as part of colonial trading arrangements. In the postwar period, developments were led by the European Union (EU, originally the European Economic Community (EEC), founded in 1958), and in the 1960s and 1970s there was a number of rather inward looking (and largely unsuccessful) RIAs between developing countries. Since the mid-1980s there has been a dramatic increase in regional integration activity. Of the 194 RIAs notified to the General Agreement on Trade and Tariffs/World Trade Organization (GATT/WTO) at the beginning of 1999, 87 were notifications since 1990. Now almost all countries are members of at least one RIA, and more than one third of world trade takes place within such agreements. The new developments include the expansion and deepening of the EU; the construction of new and more open RIAs between developing countries; and the advent of RIAs in which both high-income and developing countries are equal partners, lead by the North American Free Trade Area (NAFTA) which, in 1994, extended the Canadian–USA free trade agreement to Mexico.

RIAs come in many shapes and sizes. They vary in income levels, in openness to trade, and in the share of trade that takes place within the RIA, (60 percent for the 15 members of the EU, but just 10 percent for the eight members of the West African Economic and Monetary Union). Structures vary, from the loose agreements to facilitate trade of the African Cross-Border Initiative and Asia Pacific Economic Cooperation (APEC) forum, through to the deep integration of the EU, involving the construction of shared executive, judicial, and legislative institutions. The focus of this article, in common with most of the economics literature, is on RIAs as mechanisms for trade liberalization, not addressing the integration of factor markets, monetary systems, or political institutions. Even with this focus, RIAs differ widely in structure. 'Free trade areas' remove internal tariff and non-tariff barriers, but permit members to retain independent external tariff policies. 'Customs unions' go further, in fully harmonizing external trade policy. And it is increasingly recognized that effective integration of markets requires more than reducing tariffs and quotas. Many other barriers—such as differing national product standards and a host of minor border frictions—impede trade and support market segmentation, and some RIAs are now pursuing 'deep integration' policies to eliminate these barriers.

This article is organized under four main headings. Section 1 outlines the main effects of RIAs on member countries, Sect. 2 discusses some of the wider policy choices faced in RIA formation, and Sect. 3 looks at the implications of RIAs for the world trading system as a whole. Section 4 concludes.

1. *Economic Effects on Member Countries*

1.1 *Trade Creation and Diversion*

The modern analysis of RIAs dates from Viner (1950), who drew the distinction between the trade creating and trade diverting effects of RIA formation. The classical source of gains from trade is that global free trade allows consumers and firms to purchase from the cheapest source of supply, hence ensuring that production is located according to comparative advantage. In contrast, trade barriers discriminate against foreign supply, inducing domestic import-competing producers to expand even though they have higher costs than imports. This in turn starves domestic export sectors of resources and causes them to be smaller than they otherwise would be. Since a RIA liberalizes trade, reducing at least some of the barriers, doesn't it follow that it too will generate gains from trade? Viner's contribution was to show that the answer is; not necessarily. The gains from trade argument applies if all trade barriers are reduced, but need not apply to a partial—and discriminatory—reduction in barriers, as in a RIA. This is because discrimination between sources of supply is not eliminated, it is just shifted. If partner country production displaces higher-cost domestic production then there will be gains—trade creation. But it is possible that partner country production may displace lower cost imports from the rest of the world, and this is welfare reducing trade diversion.

The analysis of trade creation and trade diversion constitutes one of the first formal analyses of the more general problem of 'second-best welfare economics.' Given that distortions remain in place in some activities in the economy, it is not necessarily the case that removing just some of the distortions (e.g., eliminating trade barriers on partner countries and leaving them in place on external countries) is welfare improving. In the literature on regional integration the response to the fundamental ambiguity created by the second best took three main forms.

First, authors established circumstances under which there is no interaction between formation of the RIA and external trade flows, so no possibility of trade

diversion. Meade (1955) pointed out that if trade barriers with nonmembers take the form of fixed quantitative restrictions, then a RIA must raise the total welfare of member countries since there is no possibility that imports from the rest of the world are displaced. Ohyama (1972) and Kemp and Wan (1976) showed how, when external trade barriers take the form of tariffs, it is possible to adjust these to hold external trade volumes constant, so preventing trade diversion from occurring.

Second, researchers identified conditions, in terms of changes in endogenous variables, for welfare gain. For example, welfare increases if the initial-tariff weighted change in trade volume is positive (Meade 1955). If internal tariffs are close to zero, then reducing them to zero raises welfare if it increases tariff revenues earned on external trade (Ethier and Horn 1984).

The third approach is to identify features of economies (in terms of their underlying exogenous characteristics) under which they are more or less likely to gain or lose from RIA membership. Lipsey (1957) argued that joining with countries that are already one's largest trading partners is unlikely to lead to diversion, since the fact that the countries were originally the largest trading partners suggests that they are the lowest cost source of supply. Similar reasoning, including transport costs in the costs of supply, leads to the 'natural trading bloc' argument (Wonnacott and Lutz 1989, Summers 1991). Venables (2000) shows that those members of an RIA with comparative advantage most different from the world average are most likely to lose from trade diversion, as their trade is diverted to partner countries with comparative costs between theirs and the world average.

Empirical work on trade creation and trade diversion has taken two main forms; econometric studies of changes in trade flows, and computer simulation studies of the full general equilibrium effects of RIA membership.

Econometric studies seek to quantify the changes in trade flows attributable to membership of a RIA, and thereby identify trade creation and diversion. A variety of different econometric models have been developed, the most common being based on the gravity model which estimates bilateral trade between countries as a function of their GDPs, populations, the distance between them, and physical factors such as sharing a land border, and being landlocked or an island. Dummy variables capture whether or not countries are in a particular RIA, their estimated effect indicating whether countries in a RIA trade more or less than would otherwise be expected. Using this technique, Bayoumi and Eichengreen (1997) find that the formation of the EEC reduced the annual growth of member trade with other industrial countries by 1.7 percentage points, with the major attenuation occurring over 1959–61, just as trade preferences were phased in. Soloaga and Winters (1999) look at a wide range of RIAs, producing a mixed picture with little evidence of widespread trade diversion. Overall, there appears to be weak evidence that external trade is smaller than it otherwise might have been in at least some of the blocs that have been researched, but the picture is sufficiently mixed that it is not possible to conclude that trade diversion has been a major problem. Furthermore, it cannot be inferred that trade diversion has been economically damaging without information on relative costs and tariff structures, variables that are not revealed in this sort of aggregate exercise.

The second empirical approach is based on computable equilibrium modeling. This involves construction of a full computer model of the economies under study and use of the model to simulate the effects of the policy changes associated with the RIA. Such a model typically contains a great deal of microeconomic detail, so can be used to predict changes in production in each sector, and changes in factor prices and real incomes. In models that assume a perfectly competitive environment, the combined effects of trade diversion and trade creation typically give very small welfare gains—just a fraction of 1 percent of GDP (see Baldwin and Venables 1997 for a survey). The strength of these models is that they have sufficient microeconomic structure for the effects of a policy change to be traced out in detail, and its real income effects to be calculated. They are also often used for prediction—to estimate the likely effects of a policy change before it is implemented. But they have the major weakness that they are not usually fitted to data as carefully, or subject to the same statistical testing, as econometric models. The cost of the microeconomic detail is a complexity that makes rigorous econometric estimation impossible.

Although the focus of the trade creation and diversion literature has been on the changes in trade flows induced by regional integration, two consequent effects are important. First, changes in trade flows may change world prices, possibly improving the terms of trade of member countries, although this gain arises at the expense of outside countries. For example, if trade diversion occurs then RIA imports from outside countries are reduced, and any reduction in import prices that this causes is a terms of trade gain. Empirical work on this issue by Winters and Chang (2000) shows that Brazil's membership in Mercosur has been accompanied by a significant decline in the relative prices of imports from nonmember countries.

Second, changes in tariffs and trade volumes will lead to loss of government tariff revenue. This can occur directly (as intra-RIA tariffs are cut) and as a consequence of trade diversion (as imports are diverted away from external, tariff inclusive, sources of supply). Its cost depends on the social cost of raising funds by alternative means, and can be severe in some developing countries. For example, in the South African Development Community, where some coun-

tries are quite heavily dependent on trade with South Africa, substantial amounts of revenue are involved, amounting to perhaps 5.6 and 9.8 percent of government revenue for Zambia and Zimbabwe respectively. Cambodia derived 56 percent of its total tax revenues from customs duties prior to its entry into the Association of South East Asian Nations (ASEAN), and Fukase and Martin (1999) argue that entry into ASEAN provided a powerful stimulus for the introduction of a value-added tax.

1.2 Scale and Competition Effects

A second mechanism through which member countries are affected by RIA membership derives from the fact that countries may be too small to support, separately, activities that are subject to large economies of scale. Regional cooperation offers a route to overcome the disadvantages of smallness, by pooling resources or combining markets. These scale benefits can arise in public projects (see World Bank 2000) and also at the level of the private firm, where they typically interact with imperfectly competitive market structures. These considerations are absent from the trade creation and trade diversion approach outlined above, which is based on the perfect competition and constant returns to scale paradigm of traditional trade theory. It was only in the 1970s and 1980s that formal analysis of the interaction between trade, economies of scale, and imperfect competition began with the 'new trade theory,' and these techniques have now been extensively applied to regional integration.

The basic argument is that there is a trade-off between the extent to which firms can achieve economies of scale, and the intensity of competition in the market. For a given size market, larger firms means fewer firms and hence more monopolistic outcomes. If regional integration combines markets, then it shifts this trade-off, potentially allowing firms to be bigger and markets to be more competitive (Smith and Venables 1988). For example, there might be an initial situation in which two economies each have two firms in a particular industry, and these firms exploit their 'duopoly' power, setting prices well above marginal cost. After formation of the RIA this becomes four firms in one combined RIA market. This increases the intensity of competition, and possibly induces merger (or bankruptcy), perhaps leaving only the three most efficient firms. The net effect is increased competition, increased firm scale, and lower costs. 'Triopoly' competition is likely to be more intense than the original duopolies, and surviving firms are larger and more efficient, so can better exploit economies of scale. A further source of gains comes from possible reductions in internal inefficiencies that firms are induced to make. If the RIA increases the intensity of competition, it may induce firms to eliminate internal inefficiencies (X-inefficiency) and raise productivity. Since competition raises the probability of bankruptcy and hence layoffs, it also generates stronger incentives for workers to improve productivity, and increases labor turnover across firms within sectors. Although these are significant potential sources of gain, they have been difficult to achieve in many RIAs. This is addressed further in Sect. 2.1.

1.3 Winners and Losers

A continuing concern is with the distribution of the costs and benefits of regional integration between member countries. Do central regions gain at the expense of peripheral ones, and do poor countries tend to catch up or get left behind? The evidence is, broadly, that RIAs composed of developed countries tend to show convergence (for example, the narrowing of per capita income differentials observed in the EU, see Ben-David 1993). However, the picture for RIAs composed of developing countries is more mixed, with some examples of divergent performance (World Bank 2000).

The analytical literature on these questions is quite sparse, but provides several clues why this might occur. First, as mentioned above, trade diversion is more likely for countries with 'extreme' comparative advantage, suggesting that in a RIA among developing countries it might be the lowest income countries that experience diversion. For example, their imports of manufactures might be diverted from nonmember countries to a partner that has a comparative advantage in manufactures within the RIA, but not relative to the world at large. Second, industries might tend to cluster in locations that have relatively good market access, or that are well supplied with business services or provision of other intermediate goods. This is more likely to occur in developing countries than in developed ones, partly because of their sparser provision of business infrastructure, and partly because the small size of their manufacturing sectors means that clustering is less likely to run into congestion and other sources of diminishing returns. The clustering may lead to wages being bid up in one member country at the expense of others.

2. Policy Making

2.1 The Depth of Integration

The simplest form of regional integration is the elimination of tariffs (or quotas) between member countries. Beyond this there is a wide range of policy options open to countries considering integration, many of which turn on the 'depth' of integration sought by member countries—ranging from modest trade liberalization, through full economic integration, to the formation of shared institutions.

The distinction between a free trade area and a

customs union, in which a common external tariff is set, has already been made. The latter involves greater sharing of sovereignty and requires establishing procedures for revenue sharing, but in return can yield much greater market integration. In a free trade area where countries set different external tariffs the free internal circulation of goods is impossible; border formalities have to be maintained to ensure that external imports do not all enter through the member with the lowest external tariff, for re-export to other member countries. Since these imports include intermediate goods that are further processed in member countries, in practice this involves enforcing complicated 'rules of origin' governing trade flows within the RIA. (The EU's agreement with Poland has 81 pages of small print in its rules of origin section, and NAFTA some 200, Krueger 1997.) These rules are not necessary in a customs union, enabling simplification—or elimination—of internal border formalities.

It is increasingly recognized that tariffs and quotas alone may be just a small part of the overall barriers to trade created by an international border. Rules of origin create frictions, and so too does contingent protection (such as antidumping rules, known to have a 'trade chilling' effect even when not used), duplicative customs procedures, differing national product standards, and simple border red tape. The cost of border formalities on intra-EU trade in the early 1990s has been estimated to be more than one percent of the gross value of internal trade, despite implementation of procedures to cut these costs; in many RIAs the costs of border formalities are much larger.

Furthermore, relatively minor border frictions may constitute large real trade barriers because it can be in the interests of firms to try and limit effective cross-border competition in order to maintain collusive market-sharing arrangements. This will prevent the 'scale and competition' effects discussed in Sect. 1.2 from being achieved.

Recognition of the importance of these barriers—and of the failure of more than 25 years tariff-free trade to truly integrate markets—was a motivation behind the 'deep integration' of the EU's 1992 Single Market Program (Flam 1992). The program involved adoption of almost 300 measures falling into the following five main types.

(a) Simplification and in some cases abolition of intra-EC border controls, involving, e.g., replacing border paperwork by an EU-wide system of administering value added tax on cross-border transactions.

(b) Adoption of the 'mutual recognition' principle for product standards, under which a product that can be legally sold in any EU country can be legally sold in all, thereby removing the need for expensive re-testing and re-certification of products.

(c) Progress towards deregulation of the transport sectors of EU countries, including measures to reduce restrictions on truckers from one country accepting loads in another.

(d) Opening of public procurement in EU countries to effective competition from suppliers in all EU countries.

(e) Deregulation of service sector activities, including opening financial services to competition and giving service providers and professionals the right of establishment in other EU countries.

Estimates of the gains from these measures range up to as much as 5 percent of EU GPD, although these are based largely on computable equilibrium studies rather than an actual survey of outcomes (see Baldwin and Venables 1997).

2.2 *Policy Reform and Commitment*

Trade policy reforms—and other sorts of reform—are often hampered by the expectation that they may be reversed. Adjusting to reform typically involves investments, but these investments will not be made unless investors are confident that the reform will persist. These problems are mitigated if a country has a 'commitment mechanism' guaranteeing that the reform will be durable, and membership of a RIA can, in some circumstances, provide such a mechanism (Fernandez and Portes 1998).

The commitment mechanism operates most obviously for trade policy—membership requires that tariffs with member countries be cut, and reneging on agreed internal liberalization is likely to bring swift retaliation by partner countries. However, it has been argued that RIAs are valuable as commitment mechanisms for a much wider range of measures. Although NAFTA was ostensibly about trade policy, an important part of its motivation was the desire on the part of both the Mexican and US governments to lock in the broad range of economic reforms that the Mexican government had undertaken in the preceding years. The EU Articles of Agreement with eastern European accession candidates are explicit in promoting 'full integration into the community of democratic nations.' And the intervention of other Mercosur countries is credited with having averted a military coup in Paraguay in 1996 (Survey on Mercosur, *The Economist*, October 12, 1996). Paradoxically, it is even suggested that the value of a RIA as a commitment mechanism is greatest in areas other than trade policy, because there is already a way committing to tariff reductions—the tariff bindings of the GATT/WTO.

3. *Regional Integration Agreements and the World Trading System*

The effect of regionalism on the world trading system as a whole has been the subject of extensive recent debate. Is the growth of regionalism part of a process towards global free trade, or is it a substitute for it,

damaging to multilateral negotiations and likely to lead to a situation of protected trading blocs? It is noteworthy that RIAs go against the fundamental principle of the GATT/WTO, the 'most favored nation' principle under which trade policy must be non-discriminatory. They are permitted only via Article XXIV of the GATT, which requires that they cover 'substantially all trade' of member countries and lead to external tariffs that 'shall not on the whole be higher or more restrictive.'

There are essentially three issues. First, does a given structure of RIAs create a force for more or less liberal external trade policy? Second, is there an inherent dynamic of RIA formation which will lead to ever larger RIAs—perhaps even a RIA of the whole world, giving global free trade? And third, is there evidence that the presence of RIAs has assisted or retarded progress in the multilateral trade system?

An insight into the effect of the number of RIAs on incentives to set external tariffs was provided by Krugman (1993) who noted that if tariffs are set to improve member countries' terms of trade, then they will be lowest—and consequently world income greatest—in two opposite circumstances. One is when there is a single world trade bloc containing all countries, i.e., global free trade. The other is when trade policy is set by many small independent jurisdictions, each so small as to have no market power and no reason to deviate from free trade. Between these extremes each trading bloc has an incentive to use external tariffs to try and improve its terms of trade, suggesting that a situation of relatively few large trading blocs might be the worst possible outcome.

This line of argument has been criticized from a number of different angles. From the theory side, it should be recognized that tariff setting is a repeated game, so the incentives to cooperate, or to deviate from cooperation, need to be taken into account (Bond and Syropoulos 1996, Winters 1998). From the institutional side, it has been pointed out that article XXIV of the GATT expressly forbids RIAs from raising external tariffs. And from the empirical side, no evidence has been found to suggest that RIAs are in fact prone to set higher external tariffs than are separate countries (Foroutan 1998).

What of the dynamics of RIA formation? Does adding a member country to a RIA increase the incentives for further countries to join, and will existing members be willing to permit unrestricted entry? A number of researchers have argued that the incentives to join increase as RIAs become larger—a phenomenon termed 'domino regionalism' by Baldwin (1995). This may be because of perception of growing benefits of membership, or because of increasing costs of being outside. These costs can arise as outsiders' terms of trade may decline (the other side of the members improving terms of trade referred to in Sect. 1.1). Perhaps more importantly, countries fear that firms may relocate, in search of the benefits of a larger market (evidence of foreign direct investment flows in Europe at the time of the Single Market Program suggests that this occurred, Baldwin et al. 1996). Another source of loss from nonmembership of RIAs is the risk of being isolated if a trade war occurs (Whalley 1998).

Although these arguments suggest a growing demand for membership of RIAs, it is less clear that this will be matched by a willingness to accommodate new members, at least within existing RIAs. The most important tradeoff here is between the benefits of 'deeper' integration, and the difficulty of achieving this with larger memberships. Thus, while the EU has continued to enlarge, worries about the difficulty of accommodating new entrants impedes progress. An important contribution to the ideas on regionalism was made when the Asia Pacific Economic Cooperation (APEC) forum announced the principle of 'open regionalism.' Although this term has been given a number of quite different meanings, the key idea is that of open access, whereby the RIA announces that any country willing to abide by its rules may join. However, at time of writing, the idea remains to be put to the test, since APEC itself has yet to develop as a RIA. Overall, then, it seems that regional integration does create its own dynamic, attracting further countries to want to join. But there is no reason to believe that all such requests should be successful, or that the end of the process is more likely to be a single world free trade area than it is to be a number of competing trade blocs.

Finally, are there any reasons to believe that the presence of RIAs has facilitated or impeded progress in multilateral trade negotiations? The argument for impedance hinges mainly on finite government capacity. Investment of time and effort in regionalism reduces the capacity of governments to invest in multilateral negotiation. On the other side, it is argued that regionalism can help, by reducing the number of separate negotiators (since members of a customs union negotiate as a single body)—an argument that begs the question of how members formulate their common position.

Also, it is suggested that regionalism has proved valuable by providing 'laboratories' for trying new approaches to issues of trade reform, particularly on more difficult issues. As for the track record of RIAs and multilateral negotiations, views are again divided, although there is a view that fear of fragmentation into regional blocs provided a spur for successful completion of the Uruguay round of trade negotiations (WTO 1995).

4. *Conclusions*

Joining a RIA is a major 'one-off' event for a country. It affects all aspects of the economy—goods prices, the structure of production, and income flows to different

economic agents. Generalizing about the implications of such occasional and large changes is inevitably difficult. Nevertheless, research has established the circumstances under which gains or losses are more or less likely, and drawn out the implications of this for the design of agreements (World Bank 2000). Future research needs to pay greater attention to strategies for developing countries, and the role of RIAs in assisting them to participate effectively in the world trading system.

See also: Commodity Chains; Economic Geography; Globalization: Geographical Aspects; Globalization: Legal Aspects; Globalization: Political Aspects; International Business; International Trade: Commercial Policy and Trade Negotiations; International Trade: Geographic Aspects

Bibliography

Baldwin R E 1995 A domino theory of regionalism. In: Baldwin R E, Haaparanta P, Kiander J (eds.) *Expanding Membership in the European Union*. Cambridge University Press, Cambridge, UK

Baldwin R, Forslid R, Haaland J I 1996 Investment creation and diversion in Europe. *The World Economy* **19**: 635–59

Baldwin R E, Venables A J 1997 International economic integration. In: Grossman G M, Rogoff K (eds.) *Handbook of International Economics*. North-Holland, Amsterdam, Vol. 3

Bayoumi T, Eichengreen B 1997 Is regionalism simply a diversion: Evidence from the evolution of the EC and EFTA. In: Ito T, Krueger A (eds.) *Regionalism versus Multilateral Trade Arrangements*. University of Chicago Press, Chicago

Ben-David D 1993 Equalizing exchange: Trade liberalization and income convergence. *The Quarterly Journal of Economics* **108**: 653–79

Bhagwati J 1993 Regionalism and multilateralism: An overview. In: De Melo J, Panagariya A (eds.) *New Dimensions in Regional Integration*. CEPR, London

Bond E W, Syropoulos C 1996 The size of trading blocs: Market power and world welfare effects. *Journal of International Economics* **40**: 411–37

Ethier W, Horn H 1984 A new look at economic integration. In: Kierszkowski H (ed.) *Monopolistic Competition and International Trade*. Oxford University Press, Oxford, UK

Fernandez R, Portes J 1998 Returns to regionalism: An analysis of nontraditional gains from regional trade agreements. *World Bank Economic Review* **12**: 197–220

Flam H 1992 Product markets and 1992: full integration, large gains? *Journal of Economic Perspectives* **6**: 7–30

Foroutan F 1998 Does membership in a regional preferential trade arrangement make a country more or less protectionist? *The World Economy* **21**: 305–35

Frankel J A 1997 *Regional Trading Blocs in the World Economic System*. Institute for International Economics, Washington, DC

Fukase E, Martin W 1999 *Evaluating the Implications of Cambodia's Accession to the ASEAN Free Trade Area: A General Equilibrium Model (CGE) Approach*. World Bank, Washington, DC

Kemp M, Wan H Y 1976 An elementary proposition concerning the formation of customs unions. *Journal of International Economics* **6**: 95–7

Krueger A O 1997 Free trade agreements versus customs unions. *Journal of Development Economics* **54**: 169–87

Krugman P 1993 Regionalism versus multilateralism: Analytical notes. In: De Melo J, Panagariya A (eds.) *New Dimensions in Regional Integration*. Cambridge University Press, Cambridge, UK

Lawrence R Z 1996 *Regionalism, Multilateralism and Deeper Integration*. Brookings Institution, Washington, DC

Lipsey R G 1957 The theory of customs unions: Trade diversion and welfare. *Economica* **24**: 40–6

Meade J E 1955 *The Theory of Customs Unions*. North-Holland, Amsterdam

Ohyama M 1972 Trade and welfare in general equilibrium. *Keio Economic Studies* **9**: 37–73

Smith A, Venables A J 1988 Completing the internal market in the European Community: Some industry simulations. *European Economic Review* **32**: 1501–25

Soloaga I Winters L A 1999 Regionalism in the 1990s; what effect on trade? Discussion paper no. 2183, Centre for Economic Policy Research, London

Summers L 1991 Regionalism and the world trading system. In: *Policy Implications of Trade and Currency Zones*. A symposium sponsored by the Federal Reserve Bank of Kansas, KS

Venables A J 2000 *Winners and Losers From Regional Integration Agreements*. CEPR discussion paper, London

Viner J 1950 *The Customs Union Issue*. Carnegie Endowment for International Peace, New York

Whalley J 1998 Why do countries seek regional trade agreements? In: Frankel J (ed.) *The Regionalization of the World Economy*. University of Chicago Press, Chicago

Winters L A 1998 Regionalism versus multilateralism. In: Baldwin R, Cohen D, Sapir A, Venables A (eds.) *Market Integration, Regionalism and the Global Economy*. Cambridge University Press, Cambridge, UK

Winters L A, Chang W 2000 Regional integration and the prices of imports: An empirical investigation. *Journal of International Economics* **51**: 363–77

Wonnacott R I, Lutz M 1989 Is there a case for Free Trade Areas. In: Schott J J (ed.) *Free Trade Areas and US trade Policy*. Institute for International Economics, Washington, DC

World Bank 2000 *Trade Blocs*. World Bank, Washington, DC

World Trade Organization 1995 *Regionalism and the World Trading System*. WTO Secretariat, Geneva, Switzerland

A. Venables

International Trade: Geographic Aspects

1. Introduction

International trade is the physical movement and electronic transfer of goods and services across national borders. It includes the movement of commodities, such as natural resources and manufactured goods, and the transfer of services, such as personal

and commercial data, banking and financial transactions, and various other kinds of professional and business-related activities. These are important because they link local, regional, and national economies with the global economy, and in so doing serve as a conduit for the international diffusion and exchange of ideas and information, culture, technology, social and political institutions, managerial know-how, business cycles, and capital investment funds.

In order to understand the nature, evolution, and significance of international trade over time, geographers analyze specific attributes of trade flows. These characteristics of trade are usually defined as the *direction* of trade, which refers to the origins and destinations of trade; the *composition* of trade, which describes the specific goods and services being transferred; the *magnitude* or level of trade, which pertains to the value and/or volume of trade; and the *means* whereby goods and services are conveyed or transmitted, which may refer to specific trade routes or to competing modes of transportation and communication. The remaining sections describe the intellectual context of the geographers' interest in trade, and portray the changing emphases geographers have given to the study of international trade over the past several decades.

2. *Geographers' Inherent Interest in International Trade*

Geographers are interested in the various characteristics of international trade because they reflect the interrelationships that exist between local, regional, and national space-economies and the world economy. Examining the dynamics of trade enables geographers to forecast the interrelationships that exist between world trade patterns and global competitiveness, the internationalization and changing spatial structure of economic activities, global–local relationships, corporate strategy and government policymaking, and economic growth and development strategies.

More specifically, the intellectual context of the geographer's interest in international trade is based upon a set of traditional themes and characteristics of spatial systems that can be described as the 'geographical aspects' or 'geographies' of international trade. These core spatial concepts are *places* (e.g., cities, regions, and countries); *locations* in space (e.g., the absolute and relative locations of producers and consumers); *distributions* (e.g., spatial patterns of human activities, such as the international locations of manufacturing and agricultural production); *spatial interactions* (i.e., the movements and interconnections among places across space and over time); *scales of resolution* (i.e., units of observation, which may refer to individuals, corporations, or national governments); *change* over time (i.e., spatial and nonspatial outcomes of human behavior); *regions* (i.e., areas at varying spatial scales, such as regional trade blocs, that have common geographic characteristics and that can be utilized to examine the relationships between spatial processes and spatial patterns over time); and *potentially constraining factors*, such as the distance between buyers and sellers, or international boundaries, which may contribute to the differentiating characteristics of places, and which may increase the transactional costs of spatial interaction.

Examples of the questions that geographers ask about the attributes of international trade, which are designed to provide an understanding of the 'geographies' of trade, include the following: (a) To what extent does the distance, as measured by transportation and other transactional costs, between consumers and producers in different countries impact the direction and magnitude of international trade? (b) In what ways has the emergence of strong regional trading blocs and international institutions within the world economy affected the direction, composition, and magnitude of international trade in goods and services among nation-states? (c) What is the relationship between the magnitude and composition of a region's international trade and its long-term economic growth and development prospects? (d) As the predominant focus of global trading activities has shifted from the Atlantic to the Pacific, what changes have taken place, or may be anticipated, in the location and spatial distribution of economic activities at the local level within the countries of the North American Free Trade Association? (e) In what specific ways have globalization and the processes of internationalization changed the competitive relationships among metropolitan areas in the world, and how have these dynamics affected the direction, composition, and magnitude of international trade generated by these urban areas? (f) What is the relationship between shifts in the location and concentration of industrial production from the industrialized economies of the world to the newly emerging countries and the changing composition and growing importance of intra-firm and intra-industry trade? Over the years, questions such as these have guided geographers in their efforts to investigate the geographical aspects of international trade. In the following section, an overview of the legacy of these efforts is set forth.

3. *Evolution and Diversity of Research by Geographers*

The history of research on international trade by geographers since the 1970s or so has progressed from a period when little or no interest has been visible to one that during the decade of the 1990s has been marked by heightened and redirected activity by an increasing number of researchers. In general terms, this legacy has emphasized two approaches. Prior to the early 1990s, most of the research has been focused

upon describing various individual attributes of the international trade system, with little emphasis upon developing an integrated, theoretical body of knowledge. In contrast, more recent research has been well grounded in theory and focused upon the geographies of international trade at different spatial scales.

3.1 Early Perspectives

Prior to the late 1980s, most of the research on the geographical aspects of international trade can be characterized as infrequent and generally descriptive in nature, with an emphasis upon commodity movements between nation-states (see McConnell 1970, 1986 for reviews). Among the investigations carried out by geographers during this early period are studies of the development of international ports, overseas trading routes, and econometric models designed to explain the direction and magnitude of country-to-country trade flows. These models attributed the spatial patterns of international trade to variations in levels of gross national product, population size, distances between countries, and so-called trading preferences, which were usually based upon the membership of individual countries, in particular trading blocs.

With few exceptions, therefore, most of the pre-1990 studies are focused upon trade as country-to-country commodity flows, with little attention given to the evolving theory of international trade that was being advanced by economists and researchers in international business schools. As noted by Johnston (1989, p. 400), '… the explanatory content [of these studies] is weak; they can tell us what the world is like, but not why it is like that.' Moreover, little recognition is given to the importance of corporate-level decision making in efforts to explain the movement of goods and services across international frontiers. Finally, most of the research during this period pays little attention to the rapidly changing global economy and the interrelationships between such dynamics and the geographies of subnational places and space.

3.2 Recent Investigations

Beginning in the early 1990s, geographers have begun to intensify their interests in international trade and have been approaching this research utilizing two potentially interrelated perspectives, which are differentiated primarily by their spatial scales. One approach examines the international trade system from a macro perspective, focusing upon the flow of goods and services among nation-states. In contrast, a second area of research views international trade from a micro perspective, and examines the relationships and interdependencies between the dynamics of the global economy and subnational economies. In both of these approaches, however, the empirical research is strongly grounded in the evolving theory of international trade. The principal characteristics and representative studies of each of these two bodies of research are presented later.

The *macro-oriented* studies examine the changing spatial structure of the global trading system by emphasizing the decisions and activities of nation-states, groups of nations allied within regional trading blocs, or countries at different stages of economic and industrial development. Most of this research is focused upon the relative importance and consequences of the regional trading blocs that have emerged since the 1940s in North America, Europe, and Asia (Michalak and Gibb 1997); the formulation of econometric models that are designed to account for statistical variations in the patterns of country-to-country trade flows (O'Loughlin and Anselin 1996); and the role of international trade, and particularly the role of government trade policies, in advancing the economic growth and development of nation-states (Gwynne 1966, Poon 1994).

An example of this research is the work by Poon and Pandit (1996), which questions the suitability of the nation-state as the relevant spatial scale for analyzing the international trade system. Following a detailed statistical analysis of inter- and intrabloc trade in 1970 and 1990, the authors identify two regional hierarchies that are emerging in the spatial structure of world trade: a Pan-Eurasian region state, which consists of Europe, the former USSR, and Africa; and a Pan-Pacific trade region comprising Asia and North and South America. Contrary to expectations, they find that trade is increasing between these two region states, which supports the premise of the new trade theory that scale economies and large, efficient markets are the major forces underlying the existing international economic system. Moreover, rather than moving toward a more closed, protectionist trade environment, which has been suggested by others, these emerging region states are outward looking spatial entities that are fostering a more open, global trading environment.

In contrast to the macrobased analyses, the *micro-oriented* investigations are typically designed to explore the interrelationships that exist between the dynamics of the global trading environment and subnational space economies. Particular attention has been focused upon states, regions, technology districts, metropolitan areas, and industry sectors within particular countries. The underlying assumption of this perspective is that the global and local space economies are intertwined, with domestic markets becoming increasingly internationalized and affected by forces beyond their national borders. At the same time, however, the individuality and uniqueness of specific sub-national places contribute in important ways to the international competitive advantages of their local business establishments and to the orien-

tation, composition, and magnitude of the international trade that is conducted. Given the increased importance of international trade to local economies, it is not surprising that politicians and business leaders are interested in ways in which this segment of the local economy can be enhanced. Hence, many of the micro-oriented studies by geographers have raised normative questions about trade policy and trade-promotion efforts under the constraints of globalization.

Several studies are representative of the micro-level emphasis. Research by Erickson (1989) and Leichenko and Coulson (1999) explores the relationship between the industrial growth of states within the US and changes in the value of the foreign export shipments and direct export employment of these states. The more recent study tests a basis tenet of the new international trade theory that a bidirectional causal relationship exists between exports and regional economic growth at the state level. In other words, the model forecasts that a strong regional economy promotes exports, and exports, in turn, promote the regional economy. The theory also postulates that the growth of certain scale-sensitive industry sectors will be more strongly tied to export growth than will others. The study confirms the complex relationship that exists between state industry growth and exporting, and finds that the relationship is indeed sector dependent; in other words, some sectors, such as electronics, chemicals, instruments, paper, and wood products are good candidates for export promotion programs, while other sectors are less likely to respond favorably. In short, this study contributes to the theoretical literature on international trade and regional growth, and, at the same time, endeavors to provide practical guidelines to local government policymakers regarding the relationship between trade and industry policies.

Another example of this more recent research is a study of the impacts of intensified international trade on regions and metropolitan areas in the US (Noponen et al. 1993). The authors examine how the forces of international trade have shaped the changing location and geographic concentration of various industries in the USA, and how this geography has impacted the economic wellbeing of people living in regions throughout the country. By investigating the recent history of the steel, auto, insurance, machine tool, shipping, and pharmaceutical industries in the USA, the authors conclude that the size and location of these industrial sectors are not the result of natural comparative advantages or domestic cost differentials, as theory would suggest. Instead, they are the result of the US government's free trade policies and similar strategies in other nation-states. The study also concludes that the economic fortunes of places within the country are closely tied to the worldwide system of trade agreements, and that the US government has played a major role in developing and sustaining industrial leadership and international competitive advantage in these key industry sectors.

Finally, two additional pieces of research are representative of research that is designed to demonstrate that despite the forces of globalization and internationalization operating in the world today, sub-national places and spaces continue to possess enormous power in shaping not only the geographies of the domestic space economy, but also the international economic system. Storper (1997) makes the case that regions within nation-states continue to play a key role in the composition of international trade. Focusing attention upon production systems that are organized to carry out continuous product innovations, such as the production of scientific instruments and aircrafts, the author argues that such systems are located in highly concentrated and distinctive sub-national regions—agglomerations that he labels as 'technology districts.' The wealth of human and resource assets of these regions creates specific advantages that enable local business establishments to compete successfully in the international marketplace.

A similar argument about the relative advantages of local space economies is advanced by Clark and O'Connor (1997). The authors examine the global operations of the financial services system and conclude that the geography of specific places shapes these global markets. The power of 'local' is attributed to the informational content of financial markets that is embedded in the local markets. In particular, well-structured international financial systems are dependent upon locally specific information networks of producers and the monitoring of firms' financial and investment performances. In other words, the effect of geography in the trade of financial services is its impact upon the production and content of information at the local level. Thus, the geographic aspects of place and location have a direct impact upon the international operations of this particular service sector.

4. *Epilogue*

The recent emphasis by geographers on research that is designed to advance theory-building and to examine further the interrelationships that exist between the dynamics of the global economy and national and sub-national space economies is encouraging. By focusing upon place and space at different spatial and temporal scales, and by recognizing the interdependence of geographic scales, geographers are in a favorable position to advance our understanding of the processes that underlie the geographic aspects of international trade and to forecast more accurately the impacts these processes are likely to have on national and regional economies at different spatial scales.

Future research by geographers on the geographic aspects of international trade may be expected to

explore the six questions put forth earlier as well as the following concerns: (a) Researchers are just beginning to investigate the dynamics of international trade in services. Given the increased importance of this sector within the global economy, more research is needed to understand the influence trade in services is having upon places and space economies. (b) In addition, the interdependence of spatial scales is a two-way process: the geographies of places are outcomes of the diversity of processes, and yet these places are part of the processes themselves. For example, economic and political processes operate across spatial scales and cannot be fully understood when looking at a particular scale in isolation. Therefore, more research is needed to understand how these processes are related to international trade and how they interact across geographic scales. (c) Attention should also be given to the steps that can be taken by local places to strengthen and exploit their local assets and relative economic and political power in the global system. This seems especially important in the coming years for metropolitan areas, which are increasingly the primary actors in the global trading environment.

See also: Commodity Chains; Economic Geography; Globalization: Political Aspects; International Business; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; Multinational Corporations; World Trade Organization

Bibliography

Clark G L, O'Connor K 1997 The informational content of financial products and the spatial structure of the global finance industry. In: Cox K R (ed.) *Spaces of Globalization: Reasserting the Power of the Local*. Guilford, New York

Erickson R A 1989 Export performance and state industrial growth. *Economic Geography* **65**: 280–92

Gwynne N 1996 Trade and developing countries. In: Daniels P W, Lever W F (eds.) *The Global Economy in Transition*. Addison, UK, pp. 239–62

Johnston R J 1989 Extending the research agenda. *Economic Geography* **65**: 338–47

Leichenko R M, Coulson N E 1999 Foreign industry exports and state manufacturing performance. *Growth and Change* **30**: 479–506

McConnell J E 1970 A note on the geography of commodity trade. *The Professional Geographer* **22**: 181–4

McConnell J E 1986 Geography of international trade. *Progress in Human Geography* **10**: 471–83

Michalak W, Gibb R 1997 Trading blocs and multilateralism in the world economy. *Annals of the Association of American Geographers* **87**: 264–79

Noponen H, Graham J, Markusen A R 1993 *Trading Industries, Trading Regions: International Trade, American Industry, and Regional Economic Development*. Guilford, New York

O'Loughlin J, Anselin L 1996 Geo-economic competition and trade bloc formation: United States, German, and Japanese exports, 1968 to 1992. *Economic Geography* **72**: 131–60

Poon J 1994 Export growth, economic growth, and development levels: An empirical analysis. *Geographical Analysis* **26**: 37–53

Poon J, Pandit K 1996 The geographic structure of cross-national trade flows and region states. *Regional Studies* **30**: 273–85

Storper M 1997 *The Regional World: Territorial Development in a Global Economy*. Guilford, New York

J. E. McConnell

Internet: Psychological Perspectives

A remarkable transition in the penetration of computing and communications concepts into the routines of both professionals and the public has been seen since 1990. It is arguable that no other innovative technology (such as radio, television, video recorders, or fax machines) has entered the societal consciousness as rapidly and pervasively as has the Internet and its most famous application, the World Wide Web. As recently as 1990, the Web was unknown and the Internet remained largely the domain of academics, scientists, and government agencies. Today it is rare to see a television commercial that does not provide a Web address for the company; electronic commerce constitutes a multibillion-dollar economy on the Internet; the Web has penetrated the educational system at all levels (including K-12 classrooms in the public schools); and it is common for individuals to turn first to the Internet for information that they once would have sought from friends or from libraries—or failed to find at all.

The Internet and the Web have also profoundly affected the conduct of science. First, these technologies now support new kinds of scientific cooperation and distributed collaboration, with major implications for the conduct of research in cognitive psychology and cognitive science. Second, the societal changes that are resulting from the technology are rich areas for study and analysis by cognitive scientists. This article offers brief introductions to both these topics.

1. A Brief Internet Tutorial

Like the international telephone networks and the banking infrastructure for electronic funds transfer, the Internet is technically complex. Yet, like the telephone and ATM machines used to access those other networks, the basic user interface for the Internet (and especially the browser concept used for the Web) is simple and requires no detailed understanding of the network's technical underpinnings. A few basic principles can support a robust conceptual grasp of the Internet and the Web; once a user understands these notions, conquering the Internet and the Web can be straightforward.

The Internet is made possible by collegial agreements on standards for networking connectivity and information exchange; individual networks can connect to the Internet (and access machines on all other networks that are connected to the Internet) if they agree to abide by the standards that assure compatibility. The individual networks in the Internet must use a standard protocol for exchanging information, and this is known as the Internet protocol (IP). Coupled with another standard known as the transmission control protocol (TCP), these two standards are now the dominant communications protocols for all networking. References to TCP/IP indicate that a network has adopted the Internet standard for communications.

The World Wide Web is only one application on the Internet. Others include electronic mail, file transfers, news groups, and remote login to computers. The basic protocol that supports the Web is known as the hypertext transfer protocol (or HTTP). Addresses for sites on the World Wide Web are known as uniform resource locators, abbreviated to URLs. If you know the URL for a site of interest, it is trivial to 'visit' that site by typing the address into a Web browser on a machine that has been connected to the Internet. Access to Web pages can also be simplified by allowing a user to click on an active area or *link* on one page, which automatically requests the retrieval of the associated page, even if it is on a different computer somewhere else on the Internet.

2. *Internet Support for Collaboration*

With the recent emergence of distributed cognition as a framework for considering group decision making (see *Cognition*, *Distributed*), there has been increasing research regarding the role of technology in supporting cognitive activities distributed among a number of agents, potentially consisting of both humans and machines. This perspective emphasizes understanding the role of computers and technology, including the Internet, in the context of workplaces and for tasks that often involve collaboration.

Two new disciplines have emerged as a result of this trend. One is computer supported cooperative work (CSCW), concerned with developing new technologies to support work among members of groups. The other is computer mediated communication (CMC), concerned with the psychological, sociological, and cultural aspects of computer based collaboration. Combined with the disciplines of computer science and psychology, CSCW and CMC form the basis for a multidisciplinary perspective for studying human interaction with technology. Use of the Internet to support communication and collaboration is a natural extension of such work.

In many problem-solving domains, the immensity of the research challenges, coupled with the opportunities for redundancy when research is not coordinated, have led to recognition of the need for multi-investigator and multi-institutional collaboration (Cerf and The Committee on a National Collaboratory 1993). Internet-based collaboration can allow groups of researchers at diverse locations to share resources, to pool expertise, and to offer standardized programs and tools for use at their local sites (Kouzes et al. 1996). Formal experimental work has been undertaken accordingly to understand better the impact of the Internet when scientific collaborators are geographically distributed (Patel et al. 1999, Shortliffe et al. 1998). Results from such studies suggest changes in skills required with this new medium of communication, and hence the implied challenges for education and training.

The collaboratory notion has emerged as a cohesive aggregation of individuals and institutions that work towards a common set of objectives across disparate geographic locations (Cerf and The Committee on a National Collaboratory 1993). One of the central goals of a collaboratory is to create a synergistic and interdependent research entity whose whole is greater than the sum of its parts. At present, collaboratories can best be construed as grand-scale social experiments in the development of collective intelligence, 'as if the conceptual object were produced by a single good mind' (Smith 1994). Evidence suggests that, although computer mediated collaboration is a difficult enterprise, fraught with numerous cognitive, cultural, social, and technical challenges, effective network supported collaboration at a distance is attainable (Patel et al. 1999).

3. *The Internet as a Subject of Study*

As with all computing technology, effective use of the Internet requires attention to cognitive issues in the optimal use of computers, but the distributed nature of the environment has created certain unique elements that distinguish the Internet and its use from other computing and communications environments.

3.1 *Human–Computer Interaction*

An important challenge in the development of technology based environments, including those delivered via the Internet, is the design of effective user interfaces. The World Wide Web provides an unprecedented opportunity for the widespread access to information in professional domains. However, as the number of sites increases, both professionals and lay people are finding themselves overloaded with a wealth of information that is in need of automated filtering.

A number of additional issues center around human factors and cognitive issues related to meeting the

information needs of extremely varied audiences and end-users. For example, in healthcare these users range from medical specialists to patients, each having varied background knowledge (Patel et al. 2000). The development of methods for evaluating Internet applications regarding their effectiveness and utility, even though these systems are accessed from locations ranging from the workplace to the home, is accordingly an immense challenge. Work in this area typically addresses the application of usability–engineering methods, including advances in the remote recording and analysis of end-user interactions with web-based systems.

3.2 The Social Consequences of the Internet

The Internet has permeated almost all facets of life. The societal transformations are likely to be immense and somewhat unpredictable. Much has been written about the potential consequences and promises in the area of healthcare (Shortliffe 1998). Education is another arena in which the Internet is likely to produce profound changes. The new networking technologies not only afford great opportunities for distance learning, but are also beginning to have a substantial impact in the classroom. Perhaps this impact is most strongly felt in science education, where new tools allow instant access to a wealth of information resources plus collaboration with peers and experts (Linn et al. 1998). New computer mediated collaborative learning paradigms are beginning to emerge (Koschmann 1996). However, the benefits of such an approach over traditional classroom practices has yet to be clearly established.

How does Internet use affect social and psychological well being? Kraut and colleagues (Kraut et al. 1998) studied 93 families (256 individuals) who were provided with free Internet access for two years in Pittsburgh. The results indicate that as people used the Internet more, they reported communicating with fewer friends and spent less time with their families. These subjects also experienced more stress, reported feeling more lonely and depressed even though interpersonal communication was their most important reason for using the Internet. The authors concluded that despite the remarkable advantages afforded by this new medium, it could cause too much disengagement from real life. Individuals need to moderate their use of this technology and endeavor to strike an appropriate balance.

4. Requirements for New Human Abilities

If information becomes more readily accessible in the near future, will the need to engage in effortful learning be diminished? If answers to questions can be obtained more easily through the Internet, will it be as important to have such knowledge in one's head? Many scientists believe that diminishing one's knowledge base or skills is inappropriate despite these new technologies for knowledge access. Individuals will have to be capable of discerning relevant from irrelevant information (filtering) and interpreting information within a particular context. A clear lesson from a number of years of cognitive science and artificial intelligence research is that prior knowledge critically shapes one's perception of new information. In general, the better one's understanding of the domain, the more useful and powerful information-finding resources are likely to be.

Yet new skills for technology use are clearly needed. In the world of rapid technological change, no workplace has all the skills today that are going to be needed in 20 years from now. As a result, people have begun preparing for these new skills now, such that they are continually being updated with the changing society. Literacy will need to evolve continuously, since emerging technologies are increasingly part of the routine fabric of life.

See also: Cognition, Distributed; Communication: Electronic Networks and Publications; Human–Computer Interaction; Media and History: Cultural Concerns

Bibliography

Cerf V and The Committee on a National Collaboratory 1993 *National Collaboratories: Applying Information Technology for Scientific Research*. National Academic Press, Washington, DC

Koschmann T 1996 Paradigm shifts and technology: An introduction. In: Koschmann T (ed.) *CSCL (Computer-supported Collaborative Learning): Theory and Practice of an Emerging Paradigm*. Erlbaum, Mahwah, NJ, pp. 1–23

Kouzes R, Myers J, Wulf W 1996 Collaboratories: Doing science on the Internet. *IEEE Computer* **August**: 40–6

Kraut R, Patterson M, Lundmark V, Kiesler S, Mukophadhyay T, Scherlis W 1998 Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist* **53**(9): 1017–31

Linn M, Bell P, Hsi S 1998 Using the Internet to enhance student understanding of science: The knowledge integration environment. *Interactive Learning Environments* **6**(1-2): 4–38

Patel V, Cytryn K, Shortliffe E, Safran C 2000 The collaborative health care team: The role of individual and group expertise. *Teaching and Learning in Medicin* **12**(3): 117–32

Patel V, Kaufman D, Allen V, Shortliffe E, Cimino J, Greenes R 1999 Toward a framework for computer-medicated collaborative design in medical informatics. *Methods of Information in Medicine* **38**: 158–76

Shortliffe E 1998 Health care and the next generation Internet (editorial). *Annals of Internal Medicine* **129**(July 15): 138–40

Shortliffe E, Patel V, Cimino J, Barnett G, Greenes R 1998 A study of collaboration among medical informatics research laboratories. *Artificial Intelligence in Medicine* **12**: 97–123

Smith J B 1994 *Collective Intelligence in Computer-Based Collaboration*. Erlbaum, Hillsdale, NJ

E. H. Shortliffe and V. L. Patel

Interorganizational Relationships and Networks

1. Interorganizational Relationships and Networks

The study of interorganizational relationships and networks focuses on the antecedents, patterns, content, forms, and outcomes of longer-term ties among organizations, be they suppliers, customers, competitors, or other entities. Research on interorganizational relations with regard to the level of analysis mainly concentrates on individual organizations and their bilateral linkages, both vertical and horizontal. Objects of study include, for example, alliances, joint ventures, and long-term supplier or customer relations. Research on interorganizational networks focuses predominantly on the overall patterns of relationships among a larger group of organizations and their members. For example, it examines social networks among members of different organizations, interlocking directorates, or regional firm and industry clusters.

Studies of interorganizational relations and networks have their roots in three distinct research traditions. A tradition of business research, mainly in the fields of strategic management and organization science, predominantly in the form of studies of interorganizational relationships. Emerging since the 1980s, this research tends to focus on the reasons for collaboration between organizations, on the particular forms of collaboration, and the impact on participating organizations.

In research on interorganizational networks, there exist two broad perspectives. The tradition of social network analysis is mainly founded in social-psychology and sociology. In particular, it stems from sociostructural analysis, namely sociometry, developed from the 1930s. This research concentrates on how the structural patterns formed by the relations that exist among multiple actors (be they organizations or individual organization members), and an actor's position within a network, influence the actor's behavior and performance. Implications of networking for individual actors thus constitute the focal point of these studies. The research tradition highlighting clusters of organizations, mainly based in economics, economic geography, and sociology. This literature focuses on the emergence, functioning, and effects of local agglomerations of co-specialized firms, including supporting institutions such as educational or financial organizations.

Viewing recent developments in the field of research on interorganizational relations and networks, it seems that the disciplinary boundaries separating these three research traditions have become more permeable. Scholars from all three research traditions have begun mutually to recognize their research and to adopt concepts and methods from one another. Specifically, organization and strategy researchers have begun to acknowledge that the way in which organizations are embedded in wider networks of relationships and in specific institutional contexts enables and constrains organizational behavior. Researchers exploring local clusters of firms have likewise investigated how social networks impact on the functioning of such clusters. Finally, social network analysis has begun to complement its traditional focus on network analytical methods by theoretical insights often derived from strategy and organizational research. Nevertheless, the following sections will discuss the three noted research traditions separately, as sufficient differences remain between them.

2. Interorganizational Relationships

The phenomenon of interorganizational collaboration is not new. Since the 1980s, however, one could witness in many industries a fast proliferation of joint ventures, alliances, consortia, and other forms of collaboration that has triggered a burgeoning literature (see *Alliances and Joint Ventures: Organizational*). Among the prominent research questions addressed by this research are the following: Why do firms engage in interorganizational cooperation, and when? With whom are firms likely to ally, and why? How do firms organize and manage their cooperative relationships? What are the outcomes and implications of particular forms of interorganizational relationships for the participating firms and for third parties?

2.1 Motives

Actors' motives for engaging in interorganizational relationships can be quite varied (see Glaister and Buckley 1996). Some firms cooperate in order to collude against common rivals or reduce competition by binding competitors as allies. Others use cooperation in an attempt to access needed resources and/or capabilities in order to enhance their competitiveness, e.g., in terms of improved products, better or wider market access, or faster market entry. Moreover, the formation of interorganizational relationships is seen as a means of fast, effective, and efficient learning, thus short-circuiting the process of acquiring and appropriating knowledge. In addition, cost savings in various ways can motivate the formation of interorganizational relationships. Through joint research, procurement, production or marketing, firms strive

to realize economies of scale and/or scope. Furthermore, firms can also cooperate in order to achieve specialization benefits, when they decide not to engage in the full spectrum of a particular product's value chain but to specialize in particular stages of production, and then to cooperate in the provision of the final product. A better information flow and improved coordination of resource flows among cooperating firms can also lead to cost and time savings. Finally, reduction of risk and uncertainty represent motives for forming interorganizational relationships when organizations seek to spread financial or other risks, for example when shouldering (mostly large) innovations or other projects (Nohria and Eccles 1992).

2.2 Forms of Interorganizational Relationships

There exists considerable variation in the formal and informal structure of interorganizational relationships. Research has suggested numerous and diverging conceptualizations and measurements in order to capture this variety. Perhaps the most widely used conceptualization refers to the contractual and organizational arrangements by which collaborating firms govern their relationship. Specifically, different forms of collaboration are distinguished on the basis of the distribution of ownership- and decision-rights between the parties, and the coordination mechanisms employed in the relationship. In this vein, research distinguishes more hierarchical, hybrid and more market-like forms of governance (Ebers 1997, Powell 1990; see *Hierarchies and Markets*). Other conceptualizations focus on the collaborators' motives, the types of resource interdependencies that exist between the collaborating organizations, or on the norms, rules and processes that characterize the collaboration.

2.3 Consequences of Interorganizational Relationships

Compared to the numerous studies on the antecedents and forms of interorganizational relationships, the consequences of interorganizational relationships for the collaborating firms, let alone for third parties such as consumers, have received considerably less scholarly attention. Depending on the specific rationale and design of a relationship, interorganizational collaboration may enhance firms' innovativeness through mutual learning, augment their resource base and capabilities through privileged access to their partners' knowledge and resources, enhance their reputation through association with reputed partners, or improve their cost position through better coordination of interdependencies and production cost savings. However, as a consequence of collaborating, firms may also experience an erosion of their competitive position by imitating or faster-learning partners. Moreover, lock-in effects of collaboration may impede necessary adjustments due to partner-specific investments and limited access to new information. In addition, a number of studies have addressed indirectly some consequences of interorganizational relationships by studying the stability and termination of interorganizational relationships (Doz and Hamel 1998).

2.4 Theoretical Approaches and Methods

As in the field of organization studies as a whole, research examining when, why, how, and to what effects organizations engage in interorganizational relationships has employed a number of different approaches and theories. Most prominently, these include resource dependence theory, organizational economics, industrial marketing and purchasing, strategic management, and institutional theory (see Oliver and Ebers 1998). Resource dependence theory views interorganizational relations as a means for securing needed resources and of enhancing control over resource supply. Power relations and the effective management of resource dependencies are at the center of researchers' interest. Organizational economics approaches, mostly informed by agency or transaction cost theory (see *Transaction Costs and Property Rights*), offer comparative assessments of various forms of interorganizational relationships. These stress the economizing effects of particular arrangements that result from an alignment of particular transaction and actor attributes with contractual and organizational features of the relationship. Industrial marketing and purchasing approaches, often with regard to technological collaboration, study how the resource, informational and social interdependencies that exist between the collaborating organizations impact on their behavior (Håkansson and Snehota 1995). Scholars from the field of strategic management focus on the exchange and allocation of resources and information within interorganizational relationships, and on how interorganizational relationships can be utilized for improving a firm's conduct and profitability (Jarillo 1993). Neo-institutional theory, by contrast, emphasizes ideological and cultural factors. This school of thought views interorganizational relationships in the context of organizations seeking support and legitimacy. It studies how legitimacy pressures, social institutions and institutionalized behaviors shape the forms and functioning of interorganizational relationships. With regard to methods employed, case study and survey research dominate the study of interorganizational relationships. Longitudinal studies are still rather rare, as are experiments and simulations.

3. Interorganizational Networks: Social Network Analysis

Social network analysis views interorganizational networks as a set of linkages (e.g., resource, friendship, informational ties) among a set of actors (individuals,

groups, or organizations). Social network analysis offers various algorithms designed to describe and analyze positions of nodes within, and structural properties of, interorganizational networks. The basic aim of this research tradition is to understand how the overall structure of an interorganizational network, and an actor's position within that network, provides opportunities and constraints for actors' behavior (see Scott 1991; see also *Network Analysis*). Compared to research on interorganizational relationships, this strand of research thus pays more attention to the overall embeddedness of organizational and individual actors in networks of relationships than to individual attributes of the actors and of their individual relations.

Social network analysis suggests two broad ways in which overall structural features of social networks influence actors' behavior. First, particular network structures and positions have implications for the information that is available to actors and thus shape their decision-making and behavior. Second, particular network structures and positions in various ways invest actors with, or deprive them of, power to exert control over other actors. Such power may stem, for example, from utilizing informational advantages, privileged access to other actors, invoking obligations associated with particular relations, or from being able to mobilize sanctions by third parties (see Burt 1992).

3.1 Principal Methods of Network Analysis

Simulation and experiment represent important methods within social network analysis. However, more and more social network analysts have begun to complement these traditional methods with survey research and sometimes ethnographic analysis. Important concepts that have been explored include network and actor centrality (i.e., measures of the connectedness of actors within a network); measures concerning structural properties of networks such as the density of a network (indicating the number of established ties in relation to the possible ties among the actors in a network); and the concept of clique (denoting a subset of actors within a network all, or most, of whom are directly tied to one another); or similarity measures: for instance, structural equivalence (two actors occupy structurally equivalent positions if they entertain identical linkages with the same third parties in the network) (see Wasserman and Faust 1994).

3.2 Major Research Subjects

Prior to the 1980s, the bulk of research in social network analysis concentrated on developing and analyzing mathematical models of social structures and ensuing behaviors. Since then, however, there has been a proliferation of studies that apply these models to empirical phenomena (see Mizruchi and Galaskiewicz 1993). Two issues in particular have received considerable attention: interlocking directorates and social capital.

3.2.1 Interlocking directorates. This research studies how the composition of firms' boards of directors influences firm behavior. If firms' directors sit on the boards of other firms, the directors of which in turn sit on the boards of other firms, this creates interlocking directorates. There exist several competing views as to the role that such interlocks play. However, many scholars suggest that interlocks represent an indicator of social relations among organizations, and of the extent to which firms are integrated into particular networks. Research has shown that interlocking directorates influence, for example, the use of 'greenmail' and of 'golden parachutes' policies for top management, the prevalence of structural defenses against takeover attempts, participation in mergers and acquisitions, and other elements of firm strategy such as the extent of research and development and the form of firm financing (see Mizruchi 1996).

3.2.2 Social capital. The notion of social capital originates from community studies that highlighted the role of strong, intense personal relationships among the members of a local community that invested them with mutual trust and the power for collective action. Social capital endows actors with a credential that entitles them to credit, and with a resource that can be utilized in action. At least two competing conceptualizations of social capital can be distinguished. A *structural view* sees social capital as a function of the information and control advantages that an actor enjoys who is in a broker position between otherwise unconnected actors; this position exploits a 'structural hole' in a network (Burt 1997). A *relational view* sees social capital as a function of the intensity and extensiveness of personal bonds among actors. In this perspective, strong ties among actors, intense relationships characterized by mutual trust, a sense of obligation, common norms and expectations, create social capital on which the respective actors can draw (Coleman 1990).

In examining the consequences of social capital, research has identified two main effects. First, social capital can be utilized for increasing the efficiency of action through increasing the efficacy of information diffusion. Strong ties and intense relationships among

actors that share many characteristics make information exchange easier and less ambiguous; likewise, exploiting the position of a structural hole enables a more efficient diffusion of information, because redundant contacts are minimized. Second, social capital can be utilized as a productive resource that enhances actors' goal achievement. This can be realized, for instance, by exploiting the position of a structural hole for learning about, and exercising control over, more rewarding opportunities; for instance, with regard to career advancement or firm profit. Likewise, in densely connected networks characterized by strong ties among actors, information about actors' behavior is disseminated quickly and evaluated on the basis of similar norms; thus, behavior can readily be socially sanctioned, cooperation is enabled and opportunism constrained. Interorganizational networks whose members command significant relational social capital were thus found to thrive and prosper.

4. *Interorganizational Networks: Clusters of Organizations*

It has long been a well-known phenomenon that firms are not evenly distributed in geographic space but to a greater or lesser extent cluster locally. From the streets and quarters occupied by specialized trades in medieval towns, to the diamond-cutting industry of Antwerp and Tel Aviv; the coal and steel industry in the Ruhr, Lothringia and Silesia regions; the cutlery industry in Sheffield and Solingen; the financial services industry in London and New York City; to the software industry in California's Silicon Valley, firms during different periods of time have formed regional or local clusters. While regional interorganizational networks thus by no means are a novel phenomenon, research interest in regional interorganizational networks since its inception in the late 1800s (Marshall 1920) has experienced a revival beginning in the 1980s.

4.1 *The Formation of Clusters*

The literature identifies two broad reasons why firms, together with some supporting organizations, cluster regionally or locally. First, through clustering firms can realize agglomeration economies. These result from both urbanization advantages and localization advantages. Urbanization advantages include economies of scale with regard to a shared infrastructure (e.g., for communication or transport), economies of scope (e.g., technological spillovers and informational externalities), specialization benefits (leading to technological or skill advantages), and transaction cost savings (e.g., reduced search costs for buyers and suppliers) (Krugman 1995). Localization benefits comprise most importantly easy access to locally concentrated resource inputs (e.g., raw materials, specialized skills, specific attitudes of a local workforce, or co-located suppliers). A related line of reasoning stresses that clustering among firms is enabled and supported by regionally embedded institutions such as chambers of commerce, employers' unions, banks, science parks, universities, and training centers. These institutions often provide important resources, e.g., capital or access to distribution channels and potential customers. Moreover, they often act as informational brokers that support the exchange of information among firms, encourage and facilitate mutual learning, and can thus foster the responsiveness, adaptability and innovativeness of networked firms.

Second, relational advantages emanating from the business and social relationships among geographically concentrated firms and individuals can also foster clustering. Vigorous competition from nearby competitors and face-to-face contact with demanding local customers can put pressure on the competitiveness and innovativeness of firms. Frequent contact with employees from rival firms and their products can further mutual learning. Family and friendship ties among local business people, common membership in local trade associations, sports clubs and political institutions can be a source of knowledge and learning and provide better access to resources, e.g., capital and political influence, that are required for successful business development. Moreover, intense social ties, behavioral norms and routines may smooth cooperation, enable the exchange of tacit knowledge, encourage a ready and more trustworthy exchange of information among network members, and may thus support mutual learning and innovation (Porter 1990, Lane and Bachmann 1998).

4.2 *Outcomes of Local Clusters*

In line with these driving forces, research has identified a number of consequences associated with regional clusters of firms. Perhaps the most widely debated outcome concerns the increased competitiveness of regional cluster firms. Two major driving forces can be identified. On the one hand, clusters enhance competition through intense competitive pressure from nearby rivals and customers; this puts high pressure on firms and their employees to compete. On the other hand, localization and relational benefits provide competitive advantage over nonclustered rivals.

A second theme widely discussed in the literature concerns learning and innovation advantages. Local proximity and shared norms and traditions are said to allow for better cooperation even among rival firms, and to promote technology and informational spillovers. Together with competitive pressure, this can

contribute to enhanced incentives as well as opportunities for learning and innovation. In addition, it is noted that clusters of small co-specialized firms that can flexibly recombine their resources and capabilities can be lighter on their feet than larger, more integrated competitors. Collectively, these firms are therefore said to be able to adapt more quickly to changes in demand, both quantitatively and qualitatively (Piore and Sabel 1984, Nooteboom 2000).

Both postulated effects make regional clusters a favored theme in discourses on regional economic development. Endogenous growth and development of regions as well as sustained regional competitiveness are among the promises that have been associated with regional clusters (Staber et al. 1996).

However, there is evidence from research that the above-noted effects of regional clusters can, but need not necessarily, arise. There exist numerous examples of clusters that thrived at one point in time, only to fail later. Many coal and steel regions in the USA and Western Europe have shared this fate, as has the once admired mini-computer cluster along Route 128 in Massachusetts, USA. Several factors may contribute to the decline of regional clusters. Some of these are inherent to the functioning of regional clusters, some associated with external developments. They include: failing demand for a cluster's product, obsolescence of particular modes of production, competition from similar clusters in other regions, loss of the ability to coordinate co-specialized activities, and inertia due to a loss of developmental abilities. As this debate about the consequences of regional clusters for the participating firms and their respective regions shows, performance evaluations of business networks and regional districts are extremely difficult and problematic.

4.3 Methods and Approaches

Most research to date builds on case study evidence. A large number of conceptual studies engage in transferring more general insights from particular disciplines, for instance, economics or sociology, to the field of regional clusters.

At the end of the nineteenth century, economics was among the first disciplines to analyze the phenomenon of regional clusters. Paradoxically, the increasing globalization in the 1980s revived interest in internationally competitive, regional clusters, both in economics and economic geography (Porter 1990, Krugman 1995). Evolutionary economics developed a particular interest for the learning and innovation characteristics of regional clusters (Nooteboom 2000) that is also shared by the innovative milieu approach (Maillat 1991). Pioneered by the work of Piore and Sabel (1984), sociology too took note of regional clusters, particularly of so-called industrial districts. The new industrial spaces approach to interorganizational networks (Storper 1997) builds on this tradition. Piore and Sabel's work was further complemented by institutional theory sociologists, who emphasized the social and institutional structures underlying regional clusters (Herrigel 1996). One important branch of this work concentrates on national and regional systems of innovation (Lundvall 1992). Despite a network approach within cluster research that emphasizes the importance of social bonding and social networks (Fisher et al. 1999), however, compared to the research on interorganizational relationships and social network analysis, in research on regional clusters cross-disciplinary dialogue seems to be far more uncommon.

See also: Alliances and Joint Ventures: Organizational; Boundaries and New Organization Forms; Competitive Strategies: Organizational; Conflict: Organizational; Economic Geography; Network Analysis; Networks and Linkages: Cultural Aspects; Policy Networks; Social Capital; Strategy: Organizational

Bibliography

Burt R S 1992 *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA

Burt R S 1997 The contingent value of social capital. *Administrative Science Quarterly* **42**: 339–65

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Doz Y L, Hamel G R 1998 *Alliance Advantage: The Art of Creating Value Through Partnering*. Harvard Business School Press, Boston, MA

Ebers M 1997 Explaining interorganizational network formation. In: Ebers M (ed.) *The Formation of Interorganizational Networks*. Oxford University Press, New York, pp. 3–40

Fisher M M, Suarez-Villa L, Steiner M (eds.) 1999 *Innovation, Networks, and Localities*. Springer-Verlag, Berlin

Glaister K W, Buckley P J 1996 Strategic motives for international alliance formation. *Journal of Management Studies* **33**: 301–32

Håkansson H, Snehota I (eds.) 1995 *Developing Relationships in Business Networks*. Routledge, London

Herrigel G B 1996 *Industrial Constructions: The Sources of German Industrial Power*. Cambridge University Press, New York

Jarillo J C 1993 *Strategic Networks: Creating the Borderless Organization*. Butterworth-Heinemann, Oxford, UK

Krugman P R 1995 *Development, Geography, and Economic Theory*. MIT Press, Cambridge, MA

Lane C, Bachmann R (eds.) 1998 *Trust Within and Between Organizations: Conceptual Issues and Empirical Applications*. Oxford University Press, New York

Lundvall B-A (ed.) 1992 *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning*. Pinter Publishers, London

Maillat D 1991 The innovation process and the role of milieu. In: Bergman E, Maier G, Tödtling F (eds.) *Regions Reconsidered:*

Economic Networks, Innovation and Local Development in Industrialized Countries. Cassel, London, pp. 103–17
Marshall A E 1920 *Principles of Economics, An Introductory Volume*, 8th edn. Macmillan, London
Mizruchi M S 1996 What do interlocks do? An analysis, critique and assessment of research on interlocking directorates. *Annual Review of Sociology* **22**: 271–98
Mizruchi M S, Galaskiewicz J 1993 Networks of interorganizational relations. *Sociological Methods and Research* **22**: 46–70
Nohria N, Eccles R G (eds.) 1992 *Networks and Organizations: Structure, Form, and Action*. Harvard Business School Press, Boston, MA
Nooteboom B 2000 *Learning and Innovation in Organizations and Economies*. Oxford University Press, Oxford, UK
Oliver A, Ebers M 1998 Networking network studies: An analysis of conceptual configurations in the study of interorganizational relationships. *Organization Studies* **19**: 549–83
Piore M J, Sabel C F 1984 *The Second Industrial Divide: Possibilities for Prosperity*. Basic Books, New York
Porter M E 1990 *The Competitive Advantage of Nations*. Free Press, New York
Powell W W 1990 Neither market nor hierarchy: Network forms of organization. *Research in Organizational Behavior* **12**: 295–336
Scott J 1991 *Social Network Analysis: A Handbook*. Sage, Newbury Park, CA
Staber U H, Schaefer N V, Sharma B (eds.) 1996 *Business Networks: Prospects for Regional Development*. Walter de Gruyter, Berlin
Storper M 1997 *The Regional World: Territorial Development in a Global Economy*. Guilford Press, New York
Wasserman S, Faust K 1994 *Social Network Analysis: Methods and Applications*. Cambridge University Press, New York

M. Ebers

Interpersonal Attraction, Psychology of

In this article we consider the nature and history of the study of interpersonal attraction in social psychology; theoretical perspectives on attraction; predictors of attraction, including perceiver, situational, and dyadic factors, and characteristics of the object of attraction; and future research directions.

1. *What Is Interpersonal Attraction?*

Interpersonal attraction is traditionally defined in social psychology as a positive attitude or evaluation regarding a particular person, including the three components conventionally ascribed to attitudes: behavioral (tendency to approach the person), cognitive (positive beliefs about the person), and affective (positive feelings for the person). Another approach treats attraction as the desire to form a friendly or romantic relationship with a particular person. Attraction is often treated as equivalent to liking. Loving, particularly being 'in love,' with someone, is sometimes seen as a very strong or special kind of attraction—'romantic attraction'—including exclusivity and sexual interest. Attraction in the above senses is distinguished from attractiveness—characteristics of people such as good looks or desirable personality that make others be attracted to them.

2. *History of the Study of Interpersonal Attraction*

Social psychologists have focused mainly on friendly and romantic attraction between previously unacquainted peers, with most research being conducted in North America. The first major research programs, conducted in the 1960s, consisted mainly of experiments on the effects of perceived attitude similarity and, to some extent, appearance. Interest in attraction between strangers largely waned by the mid-1970s as researchers became more interested in ongoing relationships. However, a new wave of research, beginning mainly in the mid-1980s, focused on a possible evolutionary basis of mate selection, and on the specific facial and bodily features that are normatively seen as attractive. Among the most recent systematic reviews of the attraction literature are a textbook chapter by Simpson and Harris (1994) and a section of a chapter by Berscheid and Reis (1998) in the most recent *Handbook of Social Psychology*. Shorter reviews of this literature are included in most social psychology and personal relationship textbooks.

3. *Theoretical Perspectives on Attraction*

The most influential theory in the early period of studying attraction, put forward by Byrne (1971) in relation to his large and successful research program on perceived similarity effects, is based on classical learning theory approaches: we are attracted to those who provide valuable rewards. Thus Byrne argues that one is attracted to someone perceived to share one's attitudes because it is rewarding to be agreed with. Other theorists have used reward approaches to explain why people are attracted to those with good looks (e.g., aesthetic values and that we gain status from being seen with them; Hatfield and Sprecher 1986). Recently, Aron and Aron (1997) posited that people are attracted to being in relationships with people who provide, specifically, the reward of 'self-expansion'—increasing one's potential efficacy by being able to include in one's self the others' desirable resources, perspectives, and identities.

A competing, influential theoretical approach from the early period of studying attraction is based on Gestalt-influenced cognitive models. For example, balance theory, and especially Newcomb's (1961) A-B-X model, posits that people are attracted to similars because people spontaneously organize cognitive elements in a balanced fashion. If person A has a positive relation to attitude object X, and A knows that person

B also has a positive relation to X, then A's cognitive elements will be balanced only if A has a positive relation to B. (Balance theories also predict that A will like B if both A and B have a negative relation to X.) A related model, cognitive dissonance theory (Festinger 1957), predicts that A will come to like B in the above situation in order to reduce or avoid the dissonance (distressing psychological contradiction) if A were not to like B.

Another set of models focus on the role of stereotypes and culture. The idea is that in a particular cultural context certain features (such as good looks and intelligence) are normatively linked—so that when a person has one of these characteristics, they are assumed to have the others—in order that attraction is strongly facilitated (e.g., Dion et al. 1972).

Finally, much recent work employs models from evolutionary psychology (see Buss and Kenrick 1998, for a review). The idea is that modern individuals are likely to be those whose ancestors had psychological tendencies that made them most successful at producing offspring. According to this view, women have evolved to prefer mates who show signs of being able to protect and support the upbringing of the child (e.g., status) while men have evolved to prefer mates who show signs of vitality (e.g., physical beauty), suggesting that they are more likely to bear healthy children.

4. *Perceiver Characteristics*

Individuals who are extroverted, dispositionally lonely, or who have low self-esteem, a preoccupied attachment style, or high need for affiliation are especially likely to become attracted to other people (see Simpson and Harris 1994). The physically attractive are also more likely to be sociable, date, and engage in sexual relationships earlier (see Berscheid and Reis 1998).

5. *Situational Factors*

Attraction requires at least some exposure to the object of attraction and attractions usually occur among people who live in the same environment, and thus who also share similar backgrounds and experiences (see Berscheid and Reis 1998). Attractions are also more likely the more interaction one has with the person—indeed familiarity effects show up even comparing exposure times varying only by seconds (Zajonc 1968). Further, attractions are likely to occur among people who are seen as appropriate for the particular kind of relationship, so that attractions are most common among age peers, among people of the same social class, and among people who are approved of by one's friends (see Aron et al. 1989). Romantic attractions are also more likely when one has just ended another romantic relationship (Aron et al. 1989), but are less likely when one is deeply committed to a current relationship (Johnson and Rusbult 1989). Finally, romantic attractions are more likely under conditions in which one is physiologically aroused (the arousal is misattributed to the person) or under unusual or challenging circumstances (which become associated with the person) (Aron et al. 1989).

6. *Dyadic Factors*

As noted earlier, a long-standing, thoroughly replicated finding is that attraction is greater toward people perceived to share one's attitudes than to those perceived not to share one's attitudes (Byrne 1971). Much of the effect is due to a dislike of those who do not share one's attitudes, though shared attitudes also play some positive role (see Berscheid and Reis 1998). Actual shared attitudes are less important than perceived shared attitudes, and similarity of personality and backgrounds has generally shown little direct effect on initial attraction (see Berscheid and Reis 1998).

A number of studies have attempted to test the idea that opposites attract, or at least that there is an attraction between those with complementary needs (such as liking to take care of people and liking to be taken care of). For the most part, this research has not yielded impressive results (see Berscheid and Reis 1998). One suggestion, with some support, is that under conditions in which a relationship is unsure, people seek those who are similar, but when people are confident that a relationship is likely, similarity plays less of a role (Aron and Aron 1997).

Another kind of similarity that has been studied is matching on social value—such as similar levels of good looks (e.g., Walster et al. 1977). Research has shown that pairs of friends and pairs of lovers do indeed tend to be matched on attractiveness and that people tend to assume such matches on desirability overall (so that if one is shown a couple where one is much more good-looking than the other, people presume the less good-looking partner must be very successful or intelligent or have some other offsetting characteristic). However, studies of initial attraction suggest that when given the opportunity, everyone would prefer the most desirable partner. Only if the possibility of rejection is highly salient does one's own level of desirability play much of a role in whom one is attracted to (or at least whom one will approach).

A highly influential influence on attraction is the discovery (or imagined discovery) that the other person likes oneself. Being liked or admired by a reasonably appropriate and attractive person plays a role in the vast majority of initial strong friendship and romantic attractions (Aron et al. 1989). Indeed, in accounts of friendship formation and falling in love, the moment of discovery of being liked is often emphasized as the moment of the onset of strong feelings of attraction.

7. Characteristics of the Person to Whom One is Attracted

Buss (1989) asked participants (mainly college students) in 37 different cultures to rate the importance of various characteristics in selecting a marital partner. The characteristics most preferred by both women and men were being a good companion, considerate, honest, affectionate, dependable, and intelligent. However, women, more than men, preferred mates described as college graduates and as having a good earning capacity; and men, more than women, preferred mates described as physically attractive. These findings, including the gender differences, have held up well in subsequent studies employing diverse research approaches. Other research has focused on more general characteristics. For example, Sadalla et al. (1987) found women and men both judged men displaying dominant behavior to be more sexually attractive, but that dominance played little or no role in either gender's judgments of a woman's sexual attractiveness.

Regarding physical appearance, a major area of research interest has been facial attractiveness. In one series of studies conducted with US samples, Cunningham (e.g., 1986) demonstrated that having large eyes, prominent cheekbones, a big smile, a large chin, and high-status clothing were all associated with a man's attractiveness as judged by women. Women's facial attractiveness as judged by men includes having large eyes, prominent cheekbones, and a big smile. However, unlike women's preferences with regard to men, men prefer women with a small chin, a small nose, narrow cheeks, high eyebrows, and large pupils. Researchers studying facial attractiveness often interpret their results in terms of general traits of attractive faces being ones that are average and symmetrical, and that such traits serve as indications of evolutionary fitness.

Another recently researched aspect of physical appearance is body shape, particularly the waist-to-hip ratio. This work, conducted mainly by Singh (e.g., 1993), posits that people are attracted to other-sex individuals with specific mathematical ratios of body fat distribution. Key findings are that women judge waste-to-hip ratios around 0.85–0.90 as most attractive in men, and that men judge ratios of about 0.7 as most attractive in women (in both cases, these ratios hold regardless of overall weight).

8. Future Directions

There is likely to be continued periodic interest in traditional issues such as similarities and differences, and there seems to be a growing interest in evolutionary approaches and the defining features of physical attractiveness. Two possibly emerging areas are understanding the role of attractiveness in long-term relationships and identifying neural systems involved in attractiveness perception.

See also: Family Theory: Economics of Marriage and Divorce; Love and Intimacy, Psychology of; Marriage; Marriage: Psychological and Experimental Analyses; Partner Selection across Culture, Psychology of; Personality and Marriage; Social Relationships in Adulthood

Bibliography

Aron A, Aron E N 1997 Self-expansion motivation and including other in the self. In: Ickes W, Duck S (eds.) *Handbook of Personal Relationships*. Wiley, Chichester, UK, pp. 251–70

Aron A, Dutton D G, Aron E N, Iverson A 1989 Experiences of falling in love. *Journal of Social and Personal Relationships* **6**: 234–57

Berscheid E, Reis H T 1998 Attraction and close relationships. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, pp. 193–281

Buss D M 1989 Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences* **12**: 1–14

Buss D M, Kenrick D 1998 Evolutionary social psychology. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, pp. 982–1026

Byrne D 1971 *The Attraction Paradigm*. Academic Press, New York

Cunningham M R 1986 Measuring the physical in physical attractiveness: Quasi-experiments on the sociobiology of female facial beauty. *Journal of Personality and Social Psychology* **50**: 925–35

Dion K, Berscheid E, Walster E 1972 What is beautiful is good. *Journal of Personality and Social Psychology* **24**: 285–90

Festinger L 1957 *A Theory of Cognitive Dissonance*. Row, Peterson, Evanston, IL

Hatfield E, Sprecher S 1986 *Mirror, Mirror: The Importance of Looks in Everyday Life*. State University of New York Press, Albany, NY

Johnson D J, Rusbult C E 1989 Resisting temptation: Devaluation of alternative partners as a means of maintaining commitment in close relationships. *Journal of Personality and Social Psychology* **57**: 967–80

Newcomb T M 1961 *The Acquaintance Process*. Holt, Rinehart and Winston, New York

Sadalla E K, Kenrick D T, Vershure B 1987 Dominance and heterosexual attraction. *Journal of Personality and Social Psychology* **52**: 730–8

Simpson J A, Harris B A 1994 Interpersonal attraction. In: Weber A L, Harvey J H (eds.) *Perspectives on Close Relationships*. Allyn and Bacon, Boston, pp. 45–66

Singh D 1993 Adaptive significance of female physical attractiveness: Role of waist-to-hip ratio. *Journal of Personality and Social Psychology* **65**: 293–307

Walster [Hatfield] E, Walster G, Berscheid E 1977 *Equity: Theory and Research*. Allyn and Bacon, Boston

Zajonc R B 1968 Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology* **9**: 269–74

A. Aron and G. Lewandowski

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Interpersonal Psychotherapy

Interpersonal psychotherapy (IPT) is a brief, time-limited psychotherapy (i.e., 12–16 sessions) that was developed in the 1970s for the treatment of nonbipolar, nonpsychotic depressed outpatients by Klerman and colleagues (Klerman et al. 1984). IPT was initially formulated not as a novel therapy, but as an attempt to represent the current practice of psychotherapy for depression (Klerman and Weissman 1993). IPT makes no assumptions about the causes of depression; however, it does assume that the development of clinical depression occurs in a social and interpersonal context and that the onset, response to treatment, and outcomes are influenced by the interpersonal relations between the depressed patient and significant others (see *Depression*). IPT moves through three defined phases, each of which is associated with specific strategies and tasks for the therapist and patient. It is similar to many other therapies in terms of techniques and stance but is distinct in terms of strategies. Its well-defined treatment strategies are aimed at resolving problems within four social domains: grief, interpersonal role disputes, role transitions, and interpersonal deficits. Although the initial goal of IPT is to reduce symptoms of depression, the overarching goal is to improve the quality of the patient's current interpersonal relations and social functioning. IPT has been found efficacious for major depression and has also been successfully adapted to treat other types of mood and nonmood disorders. The extensive empirical background and theoretical foundation, including the strategies and techniques of IPT are fully described in a comprehensive book (Weissman et al. 2000).

1. Development

Adolf Meyer's (1957) theory that psychopathology is a result of maladaptive adjustment to the social environment lay the groundwork for IPT. Harry Stack Sullivan (1953) stands as the theorist who most clearly articulated the interpersonal paradigm and popularized the term 'interpersonal' as a balance to the then-dominant intrapsychic approach. Sullivan believed that psychiatry involves the scientific study of people and the processes that go on among them, rather than the exclusive study of the mind or of society. According to Sullivan, the unit of clinical study is the patient's interpersonal relationships and people cannot be understood in isolation from them. In his theory, Sullivan (1953, p. 13) posited that people have 'relatively enduring patterns of recurrent interpersonal situations' which can either foster self-esteem or result in hopelessness, anxiety, and psychopathology.

Sullivan developed a comprehensive theory of the connections between psychiatric disorders and interpersonal relations for the developing child in the family and for the adult in the multiple transactions of life. The roles of major interest to interpersonal psychotherapy occur within the nuclear family (as parent, child, sibling, partner); the extended family; the friendship group; the work situation (as supervisor, supervisee, or peer); and the neighborhood or community. The interpersonal approach views the relationship between social roles and psychopathology as occurring in two ways: disturbances in social roles can serve as antecedents for psychopathology; and mental illness can produce impairments in the individual's capacity to perform social roles. IPT is also associated with the work of John Bowlby (1982), originator of attachment theory, who acknowledged the importance of early attachment to subsequent interpersonal relationships and psychopathology. In sum, IPT is derived from theories in which interpersonal function is recognized as a critical component of psychological adjustment and well-being.

IPT is based not only on theory but on empirical research linking change in the social environment to the onset and maintenance of depression. Considerable research exists that documents support for the key interpersonal problem areas. For instance, scientists have demonstrated that people become depressed in the circumstances of complicated bereavement, marital disputes, and the life changes associated with interpersonal role transitions, particularly in the absence of social ties. Moreover, several longitudinal studies have focused on the interpersonal consequences of depression (e.g., separation and divorce, withdrawal from social activities), showing that not only can life events trigger depression, but indeed depression can trigger adverse life events (for a review see Frank and Spanier 1995, Weissman et al. 2000). In combination, these data provide strong support for use of an interpersonal approach for the understanding and treatment of depression.

Depression is conceptualized as having three component processes: symptom function, social and interpersonal relations, and personality and character problems. Symptom function entails the development of depressive affect and the neurovegetative signs and symptoms (low energy, sleep and appetite disturbance, etc.). These are believed to have both biological and psychological precipitants. Social and interpersonal relations entail interactions in social roles with other persons derived from learning based on childhood experiences, concurrent social reinforcement, and personal mastery and competence. Personality and character problems refer to enduring traits such as inhibited expression of anger or guilt, poor psychological communication with significant others, and difficulty with self-esteem. These traits determine a person's reactions to interpersonal experience. IPT intervenes in symptom formation and social dysfunction associated with depression, rather than enduring aspects of personality, because of its relatively short duration and low level of psychotherapeutic intensity.

There is little expectation that IPT will have a pronounced effect upon personality structure; however, many IPT patients do acquire new social abilities that may help offset personality difficulties. An important component of IPT for depression is the deliberate avoidance, during the treatment of the acute symptomatic episode, of issues related to personality functioning (Weismann et al. 2000).

The methods and techniques of many schools of psychotherapy share common ground (Frank 1973), including attempts to help patients gain a sense of mastery and reduce social isolation. A major difference among the therapies lies in their conceptualization of the causes of the patient's problems, and in the typical length and focus of treatment. For instance, IPT differs from traditional psychodynamic approaches in that it is time-limited, focused, examines current rather than past relationships, and recognizes but does not focus on intrapsychic defense mechanisms and internal conflicts. IPT also differs from cognitive and behavioral approaches in that maladaptive thoughts and behaviors are addressed only as they apply to problematic interpersonal relationships. The goal is to change the relationship pattern rather than associated depressive cognitions, which are acknowledged in IPT as depressive symptoms. Although the therapeutic style is less directive in IPT than in cognitive and behavioral approaches (i.e., less didactic/prescriptive), the therapies are similar in terms of their active current focus on specified target areas. In contradistinction to other psychotherapies, IPT recognizes but does not directly focus on the patient's personality characteristics.

2. *Treatment Phases, Goals, and Strategies*

The initial phase, ordinarily the first one to three sessions, includes diagnostic assessment and psychiatric history and establishes the context for the treatment. After a standardized diagnostic symptom review is conducted, the patient is diagnosed as depressed and assigned the 'sick role.' Patients are given the 'sick role' to exempt them from additional social pressures, increase their awareness of the need for help, and elicit their cooperation in the process of recovery. A detailed examination of the patient's interpersonal history (the interpersonal inventory) is conducted, which includes a review of the patient's current social functioning and current close relationships, their patterns and mutual expectations. During this review, changes in relationships are illuminated that were proximal to the onset of symptoms (e.g., the death of a significant other, changing to a new job, increasing marital discord, or disconnection from a friend). The interpersonal inventory provides a structure for elucidating the social and interpersonal context of the onset and maintenance of depressive symptoms and delineates the focus of treatment.

An evaluation of need for medication is based on the severity of symptoms, previous experience with medications, and patient preference. The therapist and patient then discusses the diagnosis and what the patient might expect from treatment. Symptom relief starts with helping the patient understand that the vague and uncomfortable symptoms are a part of a known syndrome, which responds to several treatments and has good prognosis. The therapist then links the depressive syndrome to one of four interpersonal problem areas: grief, role transitions, interpersonal role disputes, or interpersonal deficits. After the major interpersonal problem area associated with the onset of the depression is identified, the therapist makes a specific treatment plan with the patient to work on this problem area.

During the intermediate phase of treatment, typically the middle 7 to 11 sessions, the therapist implements treatment strategies that are specific to the identified problem area. Grief is identified as the problem area when the onset of the patient's symptoms are associated with the death of a loved one, either recent or past. The goals for treating complicated bereavement include facilitating mourning and helping the patient to find new activities and relationships to substitute for the loss. Role transition includes any difficulties resulting from a change in life status (e.g., divorce, retirement or change in one's work role, moving, leaving home, diagnosis of medical illness). The patient is helped to deal with the change by recognizing positive and negative aspects of the new role they are assuming, and pros and cons of the old role this replaces. Interpersonal role disputes are conflicts with a significant other (e.g., a partner, other family member, co-worker, or close friend) which emerge from differences in expectations about the relationship. The therapist assists the patient to identify the nature of the dispute and generate options to resolve it. If resolution is impossible, the therapist assists the patient in dissolving the relationship and in mourning its loss. Interpersonal deficits include patients who are socially isolated or who are in chronically unfulfilling relationships. The goal is to reduce the patient's social isolation by helping to enhance the quality of existing relationships and encouraging the formation of new relationships. The problem area may change during the course of treatment. The patient may have several related problem areas and may work on more than one or may select the most prominent or changeable.

In the termination phase of treatment, usually the last few of the 12 to 16 sessions, the therapist assists the patient in evaluating and consolidating gains, acknowledging the feelings associated with termination, detailing plans for maintaining improvements in the identified interpersonal problem area(s), and outlining remaining work. Patients are also encouraged to identify early warning signs (e.g., low energy) and to identify plans of action.

Throughout IPT, focus stays on the interpersonal context of a patient's life. For example if a patient is having depressive symptoms, the therapist will focus on relationship difficulties that exacerbate the symptoms rather than review cognitions or inner conflict associated with the depression. Interpersonal goals are derived from specified problem areas and are formulated within the first three sessions. Meetings should not pass without reference to these goals, as they are unique and require specific and directive interventions by the therapist. Indeed, research on IPT maintenance treatment for recurrent depression has demonstrated that the therapist's ability to maintain focus on interpersonal themes is associated with better outcomes (Frank et al. 1991). In session, unfocused discussions are redirected to the key interpersonal issues, and abstract and general discussions are minimized in order to preserve focus. Therapists refrain from making inquiries that evoke vague or passive responses, such as general questions about the patient's week. Rather, sessions begin with questions such as 'What would you like to work on today?' and 'How have things been since we last met?' These questions provide more direction for patients and focus them on recent interpersonal events and recent mood, which the therapist attempts to connect. Patients who describe problem areas will be asked about recent mood and other depressive symptoms; alternatively, if the patient focuses on symptoms, the therapist asks about recent life events and interactions (Weissman et al. 2000).

3. *Therapeutic Stance and Techniques*

Each model of psychotherapy can be characterized in part by the positioning of the therapist in terms of the nature of the relationship with the patient and the level of therapeutic activity. The IPT therapeutic stance is one of warmth, support, and empathy. The IPT therapist is active and advocates for the patient rather than remaining neutral. Moreover, the therapeutic relationship is not viewed as a manifestation of transference, nor is it to be treated as a friendship (see *Therapist–Patient Relationship*). Whereas IPT is distinct at the level of strategies, it is similar to many other therapies at the level of techniques. Some of the most frequently used techniques include exploratory techniques, encouragement of affect, clarification, and communication analysis; other techniques include use of therapeutic relationship, behavior change techniques, and adjunctive techniques.

4. *Outcome Research*

IPT has demonstrated efficacy for the treatment of several mood disorders including nonpsychotic major depression and recurrent depression (Weissman et al. 2000), as well as bipolar mood disorder (Frank et al. 1999). IPT has also shown efficacy for bulimia nervosa (Agras et al. 2000), and binge eating disorder (Wilfley et al. 1993) (see *Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder*). In addition, a number of new research applications are currently under investigation for various disorders including dysthymia, post-traumatic stress disorder, social phobia, body dysmorphic disorder, chronic somatization, borderline personality disorder, and anorexia nervosa. For a comprehensive review of empirical support and studies in progress see Weissman et al. (2000).

IPT has been adapted for various populations including the elderly, adolescents, and patients with comorbid medical conditions (e.g., human immunodeficiency virus) (Weissman et al. 2000). IPT also has been modified for groups (Wilfley et al. 2000), long-term treatment, couples, telephone intervention, primary care, and adjunctive self-help, and has been translated into several languages (Weissman et al. 2000). Together, these findings indicate that IPT is not only an effective and promising treatment for a range of psychiatric disorders, but it also has the potential for wide dissemination.

5. *Future Directions*

Four key areas are in need of further investigation. First, although a number of outcome studies in IPT clearly document its efficacy, little is known about the mechanisms by which IPT exerts its effects (Frank and Spanier 1995). A greater understanding of the mechanisms by which change occurs in IPT would assist in further refinements of the treatment and yield insights about the nature of the psychiatric syndrome under investigation. Second, increased efforts to improve the effectiveness of IPT are warranted. Frank and Spanier (1995) suggest that specific efforts to improve outcome in IPT could entail altering the structure of the treatment itself (e.g., timing, duration, frequency of the sessions) and identifying specific therapists' behaviors and patient responses and features that are associated with better outcome. For instance, initial data suggest that high treatment specificity (i.e., consistent focus on the interpersonal problems and consistent use of IPT techniques) is associated with better outcome (Frank et al. 1991). However, it still remains unclear how treatment specificity translates into specific actions of the therapist and patient. Moreover, data from a comparative treatment trial in patients with bulimia nervosa (BN) suggests that IPT is slower acting than cognitive behavior therapy (Agras et al. 2000). Further research is needed to determine whether refinements can be made to increase the rapidity of BN patients' response in IPT which may lead to improvements in ultimate outcome as well. Third, continued research is

warranted to examine whether IPT is effective for other populations and disorders. For some conditions and formats the original form of IPT will be sufficient whereas for others unique adaptations may need to be employed to increase the overall effectiveness and applicability of IPT. It will be critical to determine the optimal length and dosing for particular disorders as well. Fourth, IPT was developed as a research intervention and still remains to be well disseminated among clinicians (Weissman et al. 2000). Thus, there exists a need to translate IPT efficacy data to effectiveness studies and routine clinical practice.

See also: Cognitive and Interpersonal Therapy: Psychiatric Aspects; Depression; Depression, Clinical Psychology of; Group Psychotherapy, Clinical Psychology of; Psychotherapy, Brief Psychodynamic; Psychotherapy Integration; Therapist–Patient Relationship

Bibliography

Agras W S, Walsh B T, Fairburn C G, Wilson G T, Kraemer H C 2000 A multicenter comparison of cognitive-behavioral therapy and interpersonal psychotherapy for bulimia nervosa. *Archives of General Psychiatry* **57**: 459–66

Bowlby J 1982 *Attachment and Loss: Vol 1. Attachment*. Basic Books, New York

Frank E, Kupfer D J, Gibbons R, Hedeker D, Houck P 1999 Interpersonal and social rhythm therapy prevents depressive symptomatology in bipolar1 patients. Paper presented at the 3rd International Conference on Bipolar Disorder, Pittsburgh, PA

Frank E, Kupfer D J, Wagner E F, McEachran A B, Cornes C 1991 Efficacy of interpersonal psychotherapy as a maintenance treatment of recurrent depression. *Archives of General Psychiatry* **48**: 1053–9

Frank E, Spanier C 1995 Interpersonal psychotherapy for depression: Overview, clinical efficacy, and future directions. *Clinical Psychology: Science and Practice* **2**: 349–69

Frank J D 1973 *Persuasion and Healing: A Comparative Study of Psychotherapy*, rev. edn. Johns Hopkins University Press, Baltimore, MD

Klerman G L, Weissman M M 1993 Interpersonal psychotherapy for depression: Background and concepts. In: Klerman G L, Weissman M M (eds.) *New Applications of Interpersonal Psychotherapy*, 1st edn. American Psychiatric Press, Washington, DC, pp. 3–26

Klerman G L, Weissman M M, Rounsaville B J, Chevron E S 1984 *Interpersonal Psychotherapy of Depression*. Basic Books, New York

Meyer A 1957 *Psychobiology: A Science of Man*. Thomas, Springfield, IL

Sullivan H S 1953 *The Interpersonal Theory of Psychiatry*, 1st edn. Norton, New York

Weissman M M, Markowitz J C, Klerman G L 2000 *Comprehensive Guide to Interpersonal Psychotherapy*. Basic Books, New York

Wilfley D E, Agras W S, Telch C F, Rossiter E M, Schneider J A, Cole A G, Sifford L, Raeburn S D 1993 Group cognitive-behavioral therapy and group interpersonal psychotherapy for the nonpurging bulimic individual: A controlled comparison. *Journal of Consulting and Clinical Psychology* **61**: 296–305

Wilfley D E, MacKenzie K R, Welch R R, Ayres V, Weissman M M 2000 *Interpersonal Psychotherapy for Group*. Basic Books, New York

D. E. Wilfley

Interpersonal Trust across the Lifespan

An examination of the literature in psychology indicates that interpersonal trust (referred to as trust) bears on virtually every facet of social functioning and has profound affects on mental and physical health. Trust plays a critical role in the formation and maintenance of friendships (Rotenberg 1991), romantic relationships (Lamm 1998), and relationships in business organizations (Kramer and Tyler 1996). Individuals' willingness to trust others is linked to longevity, physical health, psychological well-being, and academic performance (Bernath and Feshbach 1995, Barefoot et al. 1998). Trust is negatively associated with lying, stealing, engaging in delinquent acts, loneliness, and depression (see Rotenberg 1994, Bernath and Feshbach 1995). Establishing trust with others plays a critical role in the successful treatment of various psychological problems, such as attempted suicide and abuse (see Bernath and Feshbach 1995). Finally, scientific progress involves a fundamental trust in the integrity of researchers and authors in the scientific community. The purpose of this article is to examine trust as it changes across the lifespan.

1. A Theoretical Framework

The present examination of trust is guided by Rotenberg's (1994) 3 (bases) $\times$ 2 (domains) $\times$ 2 (target dimensions) framework. According to Rotenberg (1994), there are three bases of trust: (a) reliability, which refers to the fulfillment of word or promise, (b) emotional trust, which refers to the reliance on others to refrain from causing emotional harm, such as being receptive to disclosures, maintaining confidentiality of them, refraining from criticism, and avoiding acts that elicit embarrassment, and (c) honesty, which refers to telling the truth and engaging in behaviors that are guided by benign rather than malicious intent and by genuine rather than manipulative strategies.

The three bases of trust are further differentiated with respect to two domains, cognitive/affective and behavioral. The cognitive/affective domain pertains to individuals' beliefs/attributions of the three bases of trust or of trust *per se*, and the emotional experiences accompanying those beliefs or attributions. The behavioral domain pertains to individuals' behavioral

tendencies to rely on others to act reliably, in an emotional trustworthy fashion and honestly. In addition, both the bases and domains are further differentiated by dimensions of the target of trust that are specific qualities of trusted–distrusted persons. The dimensions of the target of trust are (a) specificity, which ranges from generalized to a specific person, and (b) familiarity, which ranges from somewhat unfamiliar to very familiar. Included in Rotenberg's (1994) framework was the principle that interpersonal trust has a strong reciprocal quality.

2. *Trust Across the Lifespan*

The following is a description of trust across the lifespan derived from an integration of Rotenberg's (1994) trust framework and the relevant research.

2.1 *Infancy* (*0–2 Years*)

During this period, individuals establish (a) basic trust versus mistrust, according to Erikson (1963), (b) different qualities of attachment with caregivers, (c) working mental models of relationships (on the basis of the latter) that affect the development of subsequent social relationships (see Bernath and Feshbach 1995), and (d) shyness arising from temperamental differences. Accordingly, infants primarily base their trust on behavioral honesty in which they attempt to determine whether persons have benign rather than malicious intentions (behaviorally identified). Caregivers and others in the immediate family are the targets of infants' trust but, as a result of internal working models and temperament, trust is generalized in a rudimentary fashion toward unfamiliar others.

2.2 *Early Childhood* (*2–6 Years*)

During early childhood, children develop (a) 'theory of mind' in which they understanding that persons can engage in acts of deception designed to evoke false beliefs in others, and (b) the ability to comprehend the link between promises and behavior by the conclusion of the period. During early childhood, children establish within a cognitive domain an honesty basis of trust, predicated on the intention to deceive, and a reliability basis of trust, predicated on the fulfillment of promises (see Rotenberg 1991, Bernath and Feshbach 1995). Because children's social network grows, they extend interpersonal trust to a larger range of adults and peers.

2.3 *Middle Childhood* (*7–12 Years*)

During this period, parents continue to affect children's social cognition and behavior and children become increasingly involved in school, peer friendships, and peer groups (primarily same-sex). Across this period, children demonstrate increasing sophistication in their reasoning about violations of social expectations and moral issues (see Kahn and Turiel 1988). As a consequence, children's honesty, reliability, and emotional trust are (a) affected by parents' trust beliefs and behavior, (b) an integral part of same-sex peer friendship, (c) part of same-sex peer group interactions, (d) linked to their reasoning about the violation of social expectations and moral principles (see Selman 1980, Kahn and Turiel 1988), and (e) associated with academic achievement (Rotenberg 1991, Bernath and Feshbach 1995). Because of the increasing exchange of intimacies among peer friends and peer group, an extensive emotional basis of trust is developed during middle childhood.

2.4 *Adolescence* (*13–19 Years*)

Adolescents' sexual development, moral development, increased ability to comprehend abstract concepts, and their involvement in broader social networks and social events result in further changes in trust. As a consequence, adolescents' trust extends to sexual matters such as birth control and the contraction of sexually transmitted diseases including AIDS (see Lear 1997). Also, adolescents begin to include sociopolitical events and the moral issues pertaining to wider groups such as society (see Colby and Kolhberg 1987) in their trust domain (generalized unfamiliar others). As a consequence, adolescents' trust is particularly susceptible to sociopolitical changes. For example, Hochreich and Rotter (1970) found that the trust beliefs of college students decreased from 1964 to 1969, spanning the social revolution of the 1960s. Adolescents' trust in the establishment decreased (national and international politics, judiciary, the mass media) and in society in general decreased (the hypocrisy, self-seeking, and competitiveness of people in general). Adolescents' trust may be a 'barometer' of social change.

2.5 *Early and Middle Adulthood* (*20–29 Years and 30–55 Years, Respectively*)

During this period, individuals most commonly become involved in extended romantic relationships and enter the workplace. Accordingly, trust during this period is highlighted by trust within the context of romantic relationships (Rempel et al. 1985). Individuals' trust in marriage partners has been found to be associated with their love of their partners and the belief that partners are concerned about their (the individuals') welfare (Rempel et al. 1985). This period is also marked by trust issues concerning superiors and co-workers in the workplace (Kramer and Tyler 1996). All bases, domains, targets, and dimensions of trust in Rotenberg's (1994) framework are viable throughout the course of adulthood.

2.6 *Later Adulthood* (*55–75 Years*)

This period is marked by individuals' concerns over quality of life and physical safety. Older adults' trust centers on the extent to which their offspring (adult children), social networks, and social agencies will provide care and help to establish or sustain a satisfactory quality of life (Bierhoff 1992). During this period, individuals are concerned about their vulnerability to being victims of crime and thus their trust centers on their physical safety (Rotenberg 1990).

2.7 *Very Old Age* (*75 + Years*)

During very old age, individuals confront issues of distribution of inheritance and, as a consequence, re-evaluate their trust in family members. They deal with the issue of their immortality and specifically whether the trust they place in family members and perhaps in society or mankind to construct a better world will be confirmed after death.

3. *Directions for Future Research*

The major problem faced by researchers is the lack of a coherent theory to guide the examination of trust, notably a theory that outlines the development of trust across the lifespan. This article marks one attempt at developing such a theory (see also Bernath and Feshbach's (1995) theory of trust across childhood and adolescence). Nevertheless, a more structural model of trust development is necessary, one that *predicts* the relations between trust and other aspects of interpersonal functioning at each age period and *predicts* developmental changes in trust. Also, it is important to note that there is a scarcity of research on trust during the first (infancy) and the final periods (later adulthood and very old age) of development. Despite Erikson's (1963) emphasis on trust during infancy, researchers have not attempted to examine that phenomenon directly. Rather, researchers have assessed trust indirectly, primarily as an implicit facet of the quality of attachment. Also, gerontologists are encouraged to examine a critical facet of interpersonal functioning during old age, that of trust.

See also: Trust: Cultural Concerns; Trust: Philosophical Aspects; Trust, Sociology of

Bibliography

Barefoot J C, Maynard K E, Beckham J C, Brammett B H, Hooker K, Siegler I C 1998 Trust, health and longevity. *Journal of Behavioral Medicine* **21**: 517–26

Bernath M S, Feshbach N D 1995 Children's trust: Theory, assessment, development, and research directions. *Applied and Preventative Psychology* **4**: 1–19

Bierhoff H W 1992 Trust and trustworthiness. In: Montada L, Filipp S H, Lerner M J (eds.) *Life Crisis and Experience of Loss in Adulthood*. Erlbaum, Hillsdale, NJ, pp. 411–33

Colby A, Kolhberg L 1987 *The Measurement of Moral Judgment*. Cambridge University Press, Cambridge, UK, Vol. 1

Erikson E H 1963 *Childhood and Society*, 2nd edn. Norton, New York

Hochreich D J, Rotter J B 1970 Have college students become less trusting? *Journal of Personality and Social Psychology* **15**: 211–4

Kahn P H Jr, Turiel E 1988 Children's conceptions of trust in the context of social expectations. *Merrill-Palmer Quarterly* **34**: 403–19

Kramer R M, Tyler T R (eds.) 1996 *Trust in Organizations: Frontiers of Theory and Research*. Sage, Thousand Oaks, CA

Lamm H, Wiesmann U, Keller K 1998 Subjective determinants of attraction: Self-perceived causes of the rise and decline of liking, love, and being in love. *Personal Relationships* **5**: 91–104

Lear D 1997 *Sex and Sexuality: Risk and Relationships in the Age of AIDS*. Sage, Thousand Oaks, CA

Rempel J K, Holmes J G, Zanna M P 1985 Trust in close relationships. *Journal of Personality and Social Psychology* **49**: 95–112

Rotenberg K J 1990 A measure of the trust beliefs of elderly individuals. *International Journal of Aging and Human Development* **30**: 141–52

Rotenberg K J (ed.) 1991 *Children's Interpersonal Trust: Sensitivity to Lying, Deception, and Promise Violations*. Springer, New York

Rotenberg K J 1994 Loneliness and interpersonal trust. *Journal of Social and Clinical Psychology* **13**: 152–73

Selman R L 1980 *The Growth of Interpersonal Understanding*. Academic Press, New York

K. J. Rotenberg

Interpretation and Translation: Philosophical Aspects

If we are to make sense of people's behavior, we need to understand their utterances and identify their beliefs and desires, as well as grasp the meanings of other cultural artifacts and practices. Philosophers have described this as finding an 'interpretation' of the agents' behavior and mental states; 'translation' is the more specific task of finding sentences in a person's own language that express the same meanings as sentences in some another language. Both issues have been central to philosophical investigations since the 1950s.

The topic raises important issues about how knowledge is possible in the social sciences, and it also engages with some central philosophical issues about how humans can have knowledge of the minds of other people, and how linguistic communication is possible. It presents a pair of closely related problems:

(a) The first is an issue in the philosophy of language. How are we able to discover the meanings of words and sentences? This problem arises in different forms: how are people able to be sure that fellow English

speakers understand words as they do? Also, how (if at all) can they understand the speech of those whose languages are different from their own?

(b) The second is a problem in the philosophy of mind. How is it possible to understand the behavior of other people by identifying the beliefs, desires, and other mental states that lead them to act as they do? Once again, this can take several forms. How is it possible to understand the people we live with, people who share a common culture' and how are we able to understand those whose ways of life are very different from our own?

These two sets of issues are closely related, and seem to be two aspects of the same problem. Unless people's beliefs and desires can be identified, it will be difficult to make sense of their speech. Moreover, unless account is taken of what people say, it will often be difficult to make sense of their mental states.

These issues became fundamental to the philosophies of language and mind when Willard van Orman Quine challenged the underlying dogmas of logical positivism in books such as *From a Logical Point of View* (Quine 1953) and *Word and Object* (Quine 1960). He did so by arguing that the positivists made assumptions about language and mind that could not be combined coherently with the sort of account of interpretation and translation that was required by their empiricist epistemology. In addition, Donald Davidson's distinctive anti-reductionist account of mind, developed in a series of papers in the 1960s (reprinted in Davidson 1980, 1984), grew out of related reflections about how we are constrained in interpreting people's behavior, verbal and otherwise. For these philosophers, and those influenced by them, examining problems about translation and interpretation was our best clue to the nature of linguistic meaning and of mental states such as beliefs and desires. These discussions have focused on how translation can capture the cognitive aspects of meaning, leaving to one side issues arising when we consider the translation of, for example, literary works.

As these remarks suggest, investigation of these topics has had two different kinds of focus or motivation. In Quine's work, it was originally a device for criticizing the views of others: examining the nature of translation shows that meanings are much messier and more indeterminate than most analytical philosophers have supposed. Davidson's efforts, largely endorsed by Quine in his later writings, have a more constructive intent: studying the standards that guide our practice of interpretation is supposed to show that our thought about mind is guided by acceptance of normative standards of rationality rather than by a search for natural laws that govern mental phenomena. The remainder of this introduction explores these motivations further, with a view to seeing why this focus on meaning and interpretation has seemed so important.

Logical positivists typically endorsed two views. First, rationality depends upon the use of principles and standards that are analytic. They are 'true by virtue of meaning,' reflecting the conventional rules that fix the meanings of words and sentences. For example, my grasp of the word 'bachelor' reveals that that bachelors must be adult unmarried males, and this fixes the criteria for the truth of claims about some person being a bachelor. Once I grasp the rules determining the meaning of a proposition, I recognize some propositions as necessarily true and I have standards for evaluating arguments and propositions. For most positivists, rationality depends upon a set of rules and principles that are analytic. Second, understanding a proposition involves knowing what evidence would show that it was true. What implications does this have for an account of how we understand other speakers? In order to obtain knowledge of their beliefs and desires, and of the meanings of their utterances, we must discover which propositions they treat as analytic, and what verification conditions they associate with their assertions. Also, since we understand people on the basis of their verbal and nonverbal behavior, we must rely solely upon that. The position is coherent only if this is indeed possible. Quine argued that it is not possible: behavior does not provide a sufficient basis for identifying propositions held to be analytic, and identifying which verification conditions are associated with particular sentences. He argued for this by examining the epistemology of translation: that there are alternative translations of one language into another which are intuitively nonequivalent, but which observations of verbal behavior are powerless to choose between. The consequence of this is the indeterminacy of translation: the idea that our utterances have determinate *meanings*, like the view that our beliefs have determinate contents, is 'metaphysical' and nonrespectable (Quine 1960, Chap. 2).

Even if we do not share Quine's empiricism, it is plausible that we understand other people's speech on the basis of their behavior. Although I have so far treated issues about the identification of people's mental states and issues about the meanings of their words as on a par, Quine's more recent writings suggest that this is potentially misleading (Quine 1990). There may be limits to how much we can know with certainty about people's thoughts and other mental states. This need not undermine the thought that studying interpretation is the best way of explaining just what our grasp of concepts such as belief and thought consists of. We have already noted that Davidson adopts this strategy, and a similar view is found in the work of Daniel Dennett. But the case of meanings and language is different. Language is a *public* means of communication, and meanings are 'open to view.' It makes little sense to suppose that all speakers of some language may associate different conventional meanings with its expressions. Hence we need an account of the epistemology of language use that in some way makes room for the publicity of these

linguistic facts. Quine has claimed that although behaviorism is implausible as an account of the mind, it is compulsory in linguistics (Quine 1990 pp. 37–8).

In *Word and Object*, Quine emphasized that he was concerned with 'radical' translation. Also, Davidson's writings on related topics were focused on 'radical' interpretation (Davidson 1984, Chap. 9). Hence we must begin by explaining what this means, and why it is important. Ordinary domestic issues of translation arise when we wonder whether one of our fellows understands properly some word or idiom: we can discover that some person uses the word 'refute' with the same meaning as 'deny,' and we can discover that this does not conform to normal usage. Or we work out the meaning of some word or construction that we do not currently understand. In doing this we may be guided by the explicit explanations of other speakers, and we exploit our understanding of some of the other things that the people say. Our problem then arises against a background of shared knowledge of the meanings of at least some words and sentences. Something similar often arises when translating a language that is unfamiliar to us. If we already understand languages that are related to the one we wish to translate, this can provide valuable clues that enable us to get started. We may even rely upon the expertise of bilinguals who speak both the language we are studying and some other language with which we are already familiar. In such cases, once again, we take for granted extensive knowledge of meanings, mental states, and cultural practices, and we use this as a springboard for making sense of particular words and utterances in familiar or unfamiliar languages. Radical translation occurs when none of these aids is available: We approach a wholly unfamiliar language and understand it without relying upon any dictionaries or bilinguals, or prior understanding of languages we know to be related to the object of our study. We have nothing to go on, it seems, but verbal behavior. Quine believed that if our aim is to advance our philosophical understanding of meaning and mind, this is the fundamental case. Similarly, Davidson approached questions of interpretation by asking how we arrive at an account of the beliefs, desires, and meanings of people without any prior knowledge of their mental states or their language. This makes our epistemological position wholly transparent when we attempt translation or interpretation.

This emphasis on radical translation is controversial. Some philosophers influenced by Wittgenstein—for example Peter Winch (1958, and see Wilson 1974)—argue that understanding of a culture or community is only possible 'from the inside,' that we must enter a community and learn to cooperate in its practice before understanding is possible. Otherwise, interpretation will be a distortion that imposes our values and standards of rationality upon the community we wish to understand.

So far, I have presented these issues as if their interest was wholly epistemological. But there is a further motivation for focusing on radical translation, one that is metaphysical (see, especially, Lewis 1973). Many philosophers believe that human beings are fundamentally physical objects; indeed, that the universe is a physical system. There are important questions about the place of mental states, such as beliefs and desires, and facts about the semantics of languages in this physical world. For thinkers such as Dennett (1979, Chap. 1) and Davidson, studying our practice of interpretation helps us to see how phenomena such as minds and meanings are related to physical phenomena. Radical interpretation relies on no evidence about the contents of the subjects' thoughts and the meanings of their words. Hence, it encourages us to focus on the relations between evidence that is characterized in nonsemantic, nonpsychological terms, and propositions that have a semantic or psychological subject matter. This leads to another reason for taking radical interpretation seriously, and another way of characterizing what makes such interpretation *radical*. The behavioral evidence for interpretation is characterized initially in ways that make no use of problematic psychological or semantic vocabulary. Thus, it provides one way of studying the relations between the physical facts and these distinctive ways of thinking about them. Davidson, for example, uses his study of interpretation to argue that mental phenomena are real, and are grounded in physical facts without being reducible to them (Davidson 1980).

1. *Quine's Claims and Arguments*

As the remarks above may suggest, Quine remained true to the empiricism of the positivists. He held that the main evidence for arriving at translations of an unfamiliar language is information about the experience that speakers take to be relevant to the acceptability of claims expressed in their language (Quine 1960, Chap. 2). We can describe the 'sensory stimulations' that would lead someone to assent to a sentence, as well as the stimulations that would lead them to reject it. Putting this into terminology inherited from behaviorism, Quine claimed that such descriptions would identify the positive and negative 'stimulus meaning' of the sentence in question. 'Observation sentences,' such as 'that is green' or 'there is a rabbit over there' will have a fairly determinate stimulus meaning, varying little among speakers and across time; but relatively few sentences will meet this condition. Some sentences will be agreed upon by all speakers in pretty much any stimulus conditions; for example, 'green is a color': these are 'stimulus analytic.' Some pairs of sentences will turn out to be accepted in the same circumstances—they are 'stimulus synonymous.' Translations should then be tested by seeing whether they preserve 'stimulus meanings,' translation observation sentences by sentences which

are accepted in the same circumstances, stimulus analytic sentences by stimulus analytic sentences, and so on. But this only gets us so far. A huge variety of sentences will be stimulus analytic: consider 'three plus three is six,' 'birds can fly,' and 'the Earth has existed for many years.' Hence there is much freedom in how they are translated. Also, many sentences will neither be stimulus analytic nor have a definite stimulus meaning. Some sentences (both occasion and standing) will be controversial, speakers disagreeing about whether they should be accepted in particular circumstances. Others will be problematic, because most speakers simply will be agnostic about whether they can be accepted, and will have no disposition to accept or reject them in any circumstances. If we adhere to the requirements of an empiricist account of meaning, the possible evidence for the correctness of a translation seems very limited.

In *Word and Object*, Quine seems to claim that these facts about stimulus meanings, stimulus synonymy, and stimulus analyticity provide the only evidence that can be relevant to translating the cognitive meanings of the utterances of some alien language. When we recall that one target of his account is to attack some positive ideas about meaning, this may not be implausible. For empiricists, who defend a verificationist theory of meaning, these links between the acceptance of sentences and experience may indeed be all that could be relevant to identifying the meanings of words and sentences. A suitably empiricist account of translation would then require us to map sentences of the alien language on to sentences of the home language in a way that preserves empirical content: we map observational sentences on to observational sentences with the same stimulus meaning; we map stimulus analytic sentences on to stimulus analytic sentences; and so on.

Quine's famous thesis of the indeterminacy of translation holds that, given this empiricist account of what makes a translation correct, there will be many intuitively nonequivalent manuals of translation between any two languages that fit all possible evidence equally well (Quine 1960, Chap. 2). There is no objective basis for choosing between them, and hence no fact of the matter which is correct. This challenges the assumption which grounded positivist epistemology: that an adequate scientific language embodies a determinate structure of analytic principles and a body of sentences associated with determinate criteria of verification. There is not enough space to examine in detail Quine's reasons for supposing that this is true, but all exploit the idea that the relations between language and experience have a holistic character. Our evidence for translation will always involve associations between experience and particular sentences or sets of sentences. But whether a stimulation will lead a speaker to assent to some sentence can depend on background knowledge. Consider Quine's famous example 'Gavagai,' supposedly accepted whenever rabbits are present. However, in practice it can make a difference whether the informant believes that there are other animals superficially like rabbits, or believes that some insect is a sure sign of the presence of rabbits. Once we consider very theoretical sentences, the links with experience are so remote and dependent on the support of other sentences taken to be true, it would be surprising if there were not ways of adjusting translations systematically that were equally compatible with the evidence. Even when we could associate a fixed stimulus meaning with some observation sentence, the translation of many non-observation sentences will depend on how we read syntactic and semantic structure into that sentence, and how we assign meanings to its subsentential parts. Again, it would be surprising if there were not ways of doing that which led to radically distinct translations for sentences which are more remotely linked to experience (see Hookway 1988, Chaps. 8–9, Wright 1997).

As arguments against a fairly austere empiricist view about distinct meanings, these arguments—albeit schematic—may carry conviction. But what this shows is unclear. Quine concludes that meanings and propositions, synonymies and analyticities are not scientifically respectable: if translation is indeterminate, then so is meaning and intentionality. Others may interpret it as a *reductio ad absurdum* of the empiricist views about meaning that Quine employed in his argument. As a constructive account of how translation proceeds, Quine's story seems to fail to take account of all the resources that are normally available to a translator or interpreter. He appears to conclude that we do employ criteria that help us to select one among the 'correct' translation manuals, but that these guide a practical choice among correct manuals rather than a further criterion of correctness. However, it is not wholly clear that this is his view, and, in his most recent writings, his commitment to the indeterminacy of translation is qualified. Influenced by the views of Davidson that will be discussed below, Quine now acknowledges that there are more constraints on the correctness of translation than were listed above, and even grants that the indeterminacy of translation is an empirical thesis that could turn out to be false (Quine 1990).

The discussion of *Word and Object* makes passing use of another 'maxim of translation,' one that subsequently has acquired much greater prominence. The 'principle of charity' rests on the idea that 'one's interlocutor's silliness … is less likely than bad translation' (Quine 1960, p. 59), and this urges us not to trust translation manuals according to which our subjects make assertions that are 'startlingly false on the face of them.' Occurring in a discussion of the anthropological idea of a 'pre-logical mentality,' Quine argues that it is rational to expect speakers of other languages to adopt the same logical standards as we do, and to reject translations that defeat this expectation. So a good translation should preserve

stimulus meanings, etc., and it should avoid attributing to our informants beliefs that are strikingly and inexplicably false.

2. Davidson and Radical Interpretation

Although Donald Davidson has always emphasized the continuities between his work and Quine's, many readers find in his writings some innovative ideas that transform the position found in *Word and Object* radically. Most strikingly, Davidson abandoned Quine's empiricism and gave center stage to a development of Quine's claims about the principle of charity (Davidson 1984). The result was a very distinctive epistemological and metaphysical position that provided underpinning for a general philosophical vision which appeared in a series of papers expressing views that have been enormously influential. Davidson's contributions can be divided into two kinds: he offered a new way of describing the enterprise of translation and interpretation; and he transformed our understanding of the evidence that was available for testing interpretations. We shall examine these in turn.

First, Davidson was emphatic that he wanted a theory of interpretation, not a theory of translation. As we have seen, this means that he seeks a systematic specification of the beliefs, desires, and other mental states of an agent, along with an account of what they mean by the words and sentences that are uttered. These tasks are inseparable: unless we have opinions about a speaker's beliefs and desires (etc.), we shall be unable to make a reasonable guess about what they are saying; and since speech is often a crucial clue to beliefs and desires, we cannot decipher someone's beliefs and desires *before* beginning to make sense of their linguistic behavior. The two tasks must be attempted as a unified whole. Hence we defend a particular assignment of meanings to someone's sentences as part of the best overall account of their mental states and linguistic behavior.

From early papers such as 'Truth and Meaning' (Davidson 1984), Davidson emphasized that a semantic theory for a natural language should use the notion of truth: understanding a sentence requires knowledge of its (possibly context relative) truth conditions. So an adequate theory should assign truth conditions to all the sentences of the language, drawing on axioms about the semantic properties of parts of sentences, and taking into account the significance of the ways in which the parts are combined in meaningful sentences. His proposal for doing this, drawing on the work of Alfred Tarski in formal semantics, need not detain us here. The correct interpretation of a sentence in an unfamiliar language will be the statement of its truth conditions that forms part of the best overall interpretation of the beliefs, desires, and behavior of its speakers. Like Quine, Davidson asks what evidence we can rely upon in establishing the correctness of interpretations (Davidson 1984, Chaps. 10–11).

Davidson's early account of how we test such theories is apparently simple. He assumed that we could identify when our subjects accept sentences as true, when they *take* them to be true. Ignoring the worry that this presupposes there is no difficulty in identifying expressions of acceptance and dissent, its value as evidence appears limited: even if we can identify the circumstances in which the informants believe that these sentences are true, it does not follow that we can identify when they *are true*—that we can pin down their truth conditions. The speakers may have systems of beliefs that contain extensive falsehood. Davidson simply denies this possibility: our practice of interpretation is guided by a principle of charity which requires us to assume, other things being equal, that when agents assent to sentences, those sentences are in fact true. We can indeed make sense of people having beliefs that are false, but we can only do so by relying on the background assumption that their other beliefs are predominantly true. Error is possible about some abstract or theoretical matters, and about everyday matters in unusual circumstances. But they are normally reliable about uncontroversial everyday matters: whether something is a blade of grass; whether it is green when seen in good lighting conditions; whether two plus two is four, and so on. Without this bridgehead of unproblematic true belief, we could not interpret our subjects' behavior by reference to beliefs and desires, and making utterances at all. Hence the principle of charity: 'We take the fact that speakers of a language hold a sentence to be true (under observed circumstances) as *prima facie* evidence that the sentence is true under those circumstances' (Davidson 1984, p. 152). Interpretation is possible because 'We can dismiss a priori the chance of massive error' (Davidson 1984, p. 169). Charity extends further than that: Davidson held that to treat someone as a person *is* to make a normative assessment—to hold that we can expect their beliefs to be (for the most part) true, their desires to be for things we can see the value of, their inferences to be sound (by our standards), and so on. Interpretation involves expecting people (*ceteris paribus*) to conform to familiar, normative standards of belief and reasoning, allowing for error and foolishness only in unusual cases. It makes no sense to think of someone as a person whose beliefs are overwhelmingly false, and whose actions seem pointless (see Hookway 1988, Chap. 10, Heal 1997).

However, things are more complex: that hunter-gatherers possess truths about fundamental particles would be just as silly as that they should fail to acknowledge things that are plainly open to view. There are many cases where error is more likely than truth: we expect people to be deceived by perceptual illusions; and psychological research teaches that we

often go wrong in thinking about probabilities, and so on. So even if the principle of charity embodies an important truth, the formulations we have employed are too simple to capture it.

This can be accommodated in two different ways. First, Richard Grandy urged that we replace the principle of charity by a principle of humanity (Grandy 1973): a good translation or interpretation should enable us to view our subjects as believing what we would believe, were we to be in their shoes, to share their experience and background beliefs. This explains why we should not expect hunter-gatherers to hold beliefs about nuclear physics, and why we expect others to share many of the tendencies to error that come naturally to us.

Second, an adequate account of interpretation and translation must incorporate an epistemology. Quine's theory did that explicitly: the role played by stimuli and observation sentences revealed his commitment to empiricism. Davidson rejected empiricism and, in his earlier writings, provided no alternative epistemological position. The principle of charity can recognize that our beliefs result from our causal interactions with our surroundings. Beliefs tend to be true when they are caused by the states of affairs that make them true, or when they are caused by reliable signs of those states of affairs. This provides guidance concerning when we should expect our subjects to possess truths, and when we should be surprised that they do so. A properly qualified principle of charity can make allowance for this, exploiting information about when we should expect agents to possess truths, and when we should expect them to be in error.

Grandy's principle is a partial acknowledgement of this, one that can operate when we are ignorant of the causal processes involved in forming our knowledge. The result that it is *a priori* that we are reliable in forming beliefs, other things being equal, can also be secured. If utterances containing 'London' typically are caused by exposure to London, then this is evidence that 'London' refers to that city. In that case, our practice of interpretation will ensure that we find the causal interactions between thought and the world that are required to ensure that our beliefs are generally true. But we should also expect that over time we would build up a philosophical anthropology, an account of people's capacities and tendencies, which will further constrain our judgments about what is being said and about what agents believe.

Exploring these various possibilities reveals that we rely on many sources of information about the anchorage of our beliefs and utterances in formulating and assessing translations and interpretations. In his writings from the 1970s, Davidson continued to claim that interpretation was indeterminate, that there would be many distinct interpretations between which no objectively grounded choice could be made. However, as Quine has insisted recently, the more we see the range of considerations that constrain translation and interpretation grow, the less obvious is this claim: it may be an empirical question about how much indeterminacy there is, if any.

3. Conclusion

Thinkers other than Quine and Davidson have also developed the idea that it is in some way constitutive of interpretation that we rationalize people's behavior, interpreting people as logical and generally truthful. It is fundamental to the 'Intentional Systems Theory' of Daniel Dennett (1979), an account of the relations between a functionalist account of mind and our ordinary vocabulary of psychological interpretation and explanation. In a series of papers, Martin Hollis (Wilson 1974, Chaps. 10–11) has applied a similar idea to problems about explanation in anthropology, challenging a variety of relativist views about rationality, and about the interpretation of ritual and magical practices.

As more and more information is taken to be relevant to reducing the range of possible interpretations and translations, the distinction between radical and nonradical interpretation may be less important than was once supposed. That a translation enables us to cooperate with those we wish to understand, to participate in their practices, should clearly be a point in its favor. However, this does not undermine the underlying insight that interpretations which follow Davidson's strategy are not thereby false. Our grasp of the mental is, indeed, shaped by a sense of the norms that govern belief, desire, and action, and interpretation is an attempt to understand behavior as reasonable in context in the light of those norms.

See also: Empiricism, History of; Language and Society: Cultural Concerns; Language and Thought: The Modern Whorfian Hypothesis; Language, Continental Philosophy of; Logical Positivism and Logical Empiricism; Meaning and Rule-following: Philosophical Aspects; Theory of Mind; Word Meaning: Psychological Aspects

Bibliography

Arrington R, Glock H-J (eds.) 1996 *Wittgenstein and Quine*. Routledge, London

Davidson D 1980 *Essays on Action and Events*. Oxford University Press, Oxford, UK

Davidson D 1984 *Inquiries into Truth and Interpretation*. Oxford University Press, Oxford, UK

Dennett D 1979 *Brainstorms*. Harvester, Hassocks, UK

Grandy R 1973 Reference, meaning, and belief. *Journal of Philosophy* **70**: 439–52

Heal B J 1997 Radical interpretation. In: Hale R A, Wright C J G (eds.) *Companion to the Philosophy of Language*. Blackwell, Oxford, UK

Hookway C 1988 *Quine: Language, Experience, and Reality*. Polity Press, Cambridge, UK
Hookway C, Pettit P (eds.) 1978 *Action and Interpretation*. Cambridge University Press, Cambridge, UK
Kirk R 1986 *Translation Determined*. Oxford University Press, Oxford, UK
Lepore E (ed.) 1986 *Truth and Interpretation*. Blackwell, Oxford, UK
Lewis D 1973 Radical interpretation. *Synthese* **23**: 331–44
Miller A 1998 *Philosophy of Language*. UCL Press, London
Quine W V O 1953 *From a Logical Point of View*. Harvard University Press, Cambridge, MA
Quine W V O 1960 *Word and Object*. MIT Press, Cambridge, MA
Quine W V O 1990 *Pursuit of Truth*. Harvard University Press, Cambridge, MA
Wilson B 1974 *Rationality*. Blackwell, Oxford, UK
Winch P G 1958 *The Idea of a Social Science and its Relation to Philosophy*. Routledge & Kegan Paul, London
Wright C J G 1997 The indeterminacy of translation: In: Hale R A, Wright C J G (eds.) *Companion to the Philosophy of Language*. Blackwell, Oxford, UK

C. Hookway

Interpretation in Anthropology

Since the beginnings of the discipline, interpretation has occupied and affected everyone who practices anthropology. Being all-pervasive, a concept or practice can become invisible such that in most standard works on the theory and history of anthropology 'interpretation' does not rate an entry in the index, and that goes, curiously, also for the book that more than others put interpretation on the map of debates in recent anthropology (Geertz 1973). Half a decade later, anthropology figured prominently among the social sciences that were said to have turned to an 'interpretive or hermeneutic approach' (Rabinow and Sullivan 1979, but see Bauman 1978 for the long and deep history of the challenge of hermeneutics). Significantly, this claim was made in a Reader, a collection of authoritative essays. As a genre of scholarly books, we expect Readers to appear when a theoretical concept and its attendant practices of research and writing are either programmatically announced or have already congealed as a distinctive (sub)discipline. The latter, as far as we can tell, has not happened. Since a clearly defined hermeneutic anthropology does not exist, a statement about 'interpretation in anthropology' is a peculiar task, and one which calls for narrative rather than systematic presentation.

1. *The Burden of History: Hermeneutics*

A way to get a grip on elusive 'interpretation' in anthropology is to trace the concept to the discipline's modern beginnings. Most historians locate them in the Age of Enlightenment, and this proves helpful provided that the scope is not narrowed unduly (and anachronistically) to a search for eighteenth-century predecessors of anthropology as a natural or social science. Thinkers whom we recognize as proto-anthropologists formulated their ideas in a context of theological, philosophical, and (often overlooked) philological debates. These debates were carried out in a field of tension that developed between the Enlightenment and, largely contemporary, Romantic thought (a complex story still best covered in Gusdorf 1973). From this broader view of our genealogy one can derive a heuristic model whereby modern conceptions of interpretation are seen to go back to moves between two poles. One of them is a long tradition of the exegesis of authoritative texts such as the Scriptures, the Classics but also corpora of Law. Typically, interpreters of such texts partook in the authority of their sources and had an interest in promoting and defending a transhistorical status for their sources. At the other pole, we find a tradition of secular critique. It emerged from moves to historicize authoritative texts by submitting them, often building on methods and techniques developed by theological and philological exegetes, to literary, social, and political analysis and interpretation (a conceptual pair that shows up also in most modern writings). Biblical criticism and what was called the Homeric question had demonstrated that these texts of quasi-transcendent religious and aesthetic authority had a mundane history of creation, collection, and editing under multiple individual and collective authorship. Because they had a past they were relevant to a critical understanding of the present. In a line leading from Hegel to Marx and Weber, to reveal the historicity of human affairs became a foremost task of social science.

Anthropology owes its openness to ideals of rational inquiry into, and appreciation of, the creations of other traditions ('cultures') at least as much to this secular tradition of critical interpretation as to the principles and habits of natural science. Given this double inheritance, distinctions between interpretive hermeneutic and explanatory scientific positions that keep the two facing each other in binary opposition become stultifying. The same applies when exegesis, interpretation working close to the material surface of texts, is contrasted with hermeneutics concerned with deep meanings. Least of all do such divisions correspond to that between a subjective quest for authority and power and the objective pursuit of scientific truth. Those who interpret and those who explain have about the same potential for turning into authoritative pundits. In sum, to acknowledge the double genealogy of anthropology in a hermeneutic and a scientific lineage maps a conceptual space necessary to discuss interpretation in anthropology; distinctions between them do not validate or invalidate, or even classify approaches.

While unproductive oppositions should be rejected, it is important to keep in mind that, in its modern

usage, interpretation has been a critical and often militant concept. To cite one of Bauman's conclusions: '*The method of sociological hermeneutics*, like that of empirical science, *can serve the practice of communication only in its negative capacity, as the method of criticism*' (1978, p. 241, emphasis in the original).

2. *Interpretation and the Object of Anthropology*

Anthropology may have owed much to a tradition of critical interpretation, and J. G. Herder, a thinker steeped in that tradition, can be credited with having bequeathed to anthropology essential elements that went into its guiding idea, culture. Yet interpretation or hermeneutics did not assume a prominent position in debates about the production of anthropological knowledge until the late 1960s. Of the two paradigms under which anthropology gained admission as an academic discipline toward the end of the nineteenth century, evolutionism (as represented by E. B. Tylor) emerged from the Scottish Enlightenment and embraced principles and taxonomic methods of natural history. It derived its methods above all from disciplines such as geology, biology, and comparative anatomy. Diffusionism can be traced to the Romantic project of a universal cultural history but it practiced a hermeneutic that had been narrowed to a positivist methodology of interpreting historical sources, as exemplified in F. Graebner's (1911) adaptation of the work of E. Bernheim (1970, orig. 1889), a medieval historian. In both paradigms, notwithstanding the literary ambitions of some proponents, such as James Frazer, and the influence of literary critic Matthew Arnold on E. B. Tylor, anthropology seemed to lack an object for critical interpretation comparable to the great texts, the Scriptures, or the Classics. This did not change with the succeeding paradigms of British functionalism, French structuralism, or even (and at first) American culturalism. Primitive, nonwestern societies were not to be interpreted; they called for functional explanation, structural decoding, or structural-functional systemic analysis, until the study of cultural 'meaning' took the center stage.

This last development was tied to the concept of symbol that, unbeknownst to many who wielded it (leaning on Freud, Cassirer, Kenneth Burke, or Susanne Langer), always threatened to take anthropology back to precritical hermetic traditions of interpretation. Most daringly stated by Roy Wagner (1975), a 'symbolic anthropology' staked its claims—also with a Reader (Dolgin et al. 1977)—at about the time when calls for an 'interpretive and hermeneutic' approach were first heard (possibly prepared by Palmer's excellent general introduction to hermeneutics in English published in 1969). Clifford Geertz was invoked as an authority by the symbolic as well as the interpretive movement. He remained somewhat aloof of both and continued to call his approach 'semiotic' (that is, sign-centered) although he went further in the direction of hermeneutics than any of his predecessors when he suggested that anthropology's object of study, culture, might be approached as (or like) a text (an 'ensemble of texts' in his words). Geertz referred to an essay by Paul Ricoeur (1971, reprinted in Rabinow and Sullivan 1979, pp. 73–101). Ricoeur had advanced the 'hypothesis' that the human sciences, by adopting a hermeneutic approach, could proceed from 'Explanation to Understanding' (for a broader statement of the relevance of hermeneutics and interpretation to the social sciences, see Ricoeur's influential study of Freud, 1970). Earlier influences from Dilthey and Weber had prepared the ground for giving to 'understanding' this new and sharp focus.

3. *Interpretation and Method*

Thus, by the mid-1970s, one of its most influential proponents seemed to have steered anthropology back to the Romantic tradition of interpretation. However, important as this move may have been at the level of theory (that is, in redefining the object of anthropology and in clarifying the status of the discipline as an interpretive rather than explanatory science), its practical consequences were limited. The model of the text did not generate a new distinctive method; thick description and a knack for discerning 'webs of signification' continued to work as the ethnographer's 'magic.' For those who embraced the hermeneutic stance, text remained essentially a metaphor and few among these writers were compelled by this figure to direct their ethnographic research and writing to actual texts and their interpretation. The latter happened in a movement that was roughly contemporary with Geertzian hermeneutics and had intellectual roots that, acknowledged or not, go back to W. von Humboldt. It took its departure from a critique of formalist structural linguistics, of purely classificatory approaches in folklore, and of positivist views of ethnography. It was in this field of innovative endeavors where critical thought in language-centered anthropology (Hymes 1974) and the philosophy of social science (Habermas 1988, orig. 1967) converged that we began to see how interpretation might develop from a theoretical figure to a practical guideline, a method organizing the production of anthropological knowledge from field research to writing.

Whether or not certain choices, steps, procedures, and strategies devised for interpretation can count as an anthropological method is an epistemological question: how must we conceive the nature of anthropological inquiry, how the nature of its object? From a hermeneutic position the answer would roughly be this: anthropology, though aware of, and in some of its branches concerned with, neural, etiological, and certain statistically demonstrable

determinations of human behavior, starts with the assumption that human experience as well as the study of that experience are always mediated. Mediated means that experiences (and our understandings) are made, shared, and transmitted by means (lit. things in the middle, media) that include language (in the broadest sense of the term), practices of communication and representation, and material objects. This is an epistemological, not an ontological view: When we conceive of an ensemble of mediations as culture this is not to posit culture as an entity. Culture is a discourse on mediations and practices. In this view, mediations are objectifications of human action somewhat but (as we shall argue) not quite in the sense envisioned by Ricoeur when he stated that 'action itself, action as meaningful, may become an object of science, without losing its character of meaningfulness, through a kind of objectification similar to the fixation which occurs in writing. By this objectification action … constitutes a delineated pattern which has to be interpreted according to its inner connections' (1971, p. 538). The analogy with fixation in writing (the reason to introduce the model of the text in the first place) 'prepares the detachment of the *meaning* of the action from the *event* of the action' (1971, p. 538). In the passages from which these quotations are drawn Ricoeur deploys the vocabulary of Weberian–Parsonian culturalism (pattern, meaning, meaningful action, interaction) and thereby advocates interpretation as an operation on elements of culture that, though objectified, are disembodied—a point often made in the critique of American culturalism and echoed in Bourdieu's (1977, p. 96) qualifying the hermeneutic stance as 'objectivism.'

In contrast, approaches to text and interpretation in anthropology that, explicitly or not, focus on mediation and on culture as praxis are concerned with maintaining or restoring the event-character of action. Interpretation then becomes a method of translating the texts and documents, that is, those 'fixations in writing' that are the 'data' produced by interactive, communicative ethnography (notes, recordings, collected texts, including some that made their way into historical or administrative archives), into events and eventually into processes. Not the model of 'the' text but models of text making are required for such interpretive work, much of which has been inspired by Dell Hymes's (1974, orig. 1964) seminal observations on components of speech events. His ideas, sometimes called 'sociolinguistic,' were further developed by the contributors to Bauman and Sherzer (1974). Linking the notion of speech event to that of genre, Dan Ben Amos has shown methodological consequences for folklore studies (1976), and a classic in this vein is Dennis Tedlock's collection of essays (1983). Fabian has stated the philosophical (anti-positivist) point of departure, offered a model of text production, and a warning against tendencies to grammaticalize Hymes's pragmatic approach to texts (1991, Chaps. 1, 3, 5).

4. *Interpretation and Performance*

The move away from semiological to pragmatic concerns reflected and promoted yet another understanding of interpretation. 'Interpretation' used to be a discipline and skill taught in speech and drama departments; today, many of them have become part of a multifaceted field called Performance Studies. Performance, as a theoretical concept usually paired with 'competence,' was adopted by Hymes from generative linguistics and applied to his ethnography of speaking but it undoubtedly gained its wider currency in anthropology in the more specific, theatrical understanding promoted, above all, by Victor Turner (stated most comprehensively in 1986). Actually, as his work shows, attention to performance in the interpretation of culture has a complex intellectual history. Turner began by giving focus and life to drama, a pervasive but mostly dead theatrical metaphor in social theory (for instance in Parsonian functionalism–structuralism) that was at the origin of such crucial sociological terms as role, actor, and enactment. He saw drama above all in ritual, and came close (as did Max Gluckman) to postulating a distinctive type of premodern society whose social dynamic was not determined by linear accumulation of change and innovation but by cyclical enactment of a dramatic script. Undoubtedly, this presented a gain over static views of primitive or traditional societies. However, not unlike the model of culture as a text, the drama metaphor had its limits when it came to understand its practical consequences for the production of ethnographic knowledge. What was needed was an epistemological conception of performance as both a way in which a culture works and a method by which it may be studied and represented. Taking an increasingly literal approach to performance as theatrical performance (including experimental collaboration with Richard Schechner, a producer and theorist in the field of theater), Turner ended in an impasse that eventually led him to advocate a biological (neurological) position not unlike that earlier taken by Lévi-Strauss (see Turner 1986, pp. 156–77). Nevertheless, his work helped to make anthropologists aware that large parts, perhaps most expressions, of cultural knowledge are performative in the sense that what they say is also constituted by how contents are enacted. An emerging 'ethnography of reading' (Boyarin 1993) applies this to texts fixated in writing and opens to anthropology a wide field of interpretation no longer predicated on categorical distinctions between orality and literacy; for methodological consequences and procedures, see Finnegan (1992) and Tonkin (1992). What changed, and was gained, by including pragmatic (sociolinguistic) and performance aspects in methods of interpreting oral sources is best illustrated by comparing Vansina's classic work on oral tradition (1961) with its radical revision (1985).

Finally, it should be mentioned that a pragmatic

and critical stance toward the interpretation of texts also leads to a re-evaluation of 'context,' a concept that has had a long history in anthropology. It was applied mainly in the validation of ethnography and often, much like text, in a metaphorical manner in admonitions to place specific facts or data in their historical, political, economic, religious, etc. 'context.' In-as-much as this suggested that, though relevant to an interpretation of findings, context was itself not subject to interpretation (at least not at the moment when it was invoked), this has become an untenable proposition. Context poses problems of interpretation, it does not solve them.

5. *The Politics of Interpretation*

A review of theoretical and practical views of interpretation in anthropology would not be complete without developments that emerged in the 1980s. They consisted of moves giving anthropology a 'literary turn,' an apt designation if it is understood as promoting attention to writing in the production of ethnography and anthropological knowledge. Not metaphorical invocations of text as a model, but calls for critical interpretation of ethnographies as texts were formulated by a group resolved to lift the veil under which modern, scientifically minded anthropology had kept ethnographic authorship (Marcus and Cushman 1982, Clifford and Marcus 1986, Clifford 1988). Postmodernism became a convenient but doubtful label for this movement—doubtful because it obscures its roots in decidedly modern advancements in the historiography of our discipline, in the theoretical confrontation with its colonial and postcolonial political role that had occupied critical anthropology since the 1960s, and, though less obviously, in feminist critique of anthropological discourse.

As regards interpretation, the subtitle of the Clifford and Marcus volume, 'The Poetics and Politics of Ethnography,' pointed to the directions further work could, and did, take. Along the lines of a hermeneutical understanding of the term, interpretation now applied to the canon of anthropological writing the tools developed by literary theory and criticism (exemplified by Geertz 1988). Interpretation in the performative sense encouraged, beyond literary deconstruction, 'experimentation with genre' which resulted, if not quite in its demise, then in relativizing the status of the monograph as the authorized form of ethnography: narrative, dialogue, poetry, collage, and autobiographic inclusion of the ethnographer in his or her account have begun to shape much of anthropological writing. Perhaps it is not always easy to draw a line between hard-won freedom from scientism and gratuitous literary posing; postmodern ethnography is at its best when it is able to justify its choice of representational forms on epistemological grounds. How an author translates practices of research into writing has become a foremost task of critical interpretation in anthropology. To be able to discern narrative plots and strategies, identify tropes and other rhetorical devices, and to recognize intertextual connections is now required when we discuss the validity of ethnographies. Conversely, though it would be foolish to cast away all our prior habits and techniques, field research can no longer rely largely on preconceived methods that once were thought to yield data without being themselves subject to constant interpretation (which would have made methods useless in a classical positivist research setup). Attempts to formulate these insights have reached the level of textbooks and handbooks that steer a difficult course between opening a back door to flat-footed methodology and scaring anthropological novices with hermeneutic hypersensitivity.

Distinguishable but never separable from attention to poetics is the concern with the politics of ethnography, that is, with relations of power that determine production of anthropological knowledge in all its phases from research to writing. Therefore the question of interpretation in anthropology cannot be raised without (taking a lesson from Michel Foucault) questioning the devices of control and domination deployed in the discourse anthropology pronounces on others. Interpretation may be 'just interpretation,' but neither as a hermeneutic alternative to scientism nor as a pragmatic method should it make anthropology complacent about its achievements.

See also: Ethnography; Hermeneutics, History of; Hermeneutics, Including Critical Theory; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Verstehen und Erklären, Philosophy of

Bibliography

Bauman R, Sherzer J (eds.) 1974 *Explorations in the Ethnography of Speaking*. Cambridge University Press, Cambridge, UK

Bauman Z 1978 *Hermeneutics and Social Science: Approaches to Understanding*. Hutchinson, London

Bernheim E 1970 *Lehrbuch der historischen Methode und der Geschichtsphilosophie*. Burt Franklin, New York (Reprint, original edition 1889)

Ben Amos D (ed.) 1976 *Folklore Genres*. University of Texas Press, Austin, TX

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK

Boyarin J (ed.) 1993 *The Ethnography of Reading*. University of California Press, Berkeley, CA

Clifford J 1988 *The Predicament of Culture: Twentieth Century Ethnography, Literature, and Art*. Harvard University Press, Cambridge, MA

Clifford J, Marcus G E (eds.) 1986 *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA

Dolgin J L, Kemnitzer D S, Schneider D M (eds.) 1977 *Symbolic Anthropology: A Reader in the Study of Symbols and Meanings*. Columbia University Press, New York
Dundes A 1980 *Interpreting Folklore*. Indiana University Press, Bloomington, IN
Fabian J 1991 *Time and the Work of Anthropology: Critical Essays 1971–1991*. Harwood Academic Publishers, Chur, Switzerland
Finnegan R 1992 *Oral Traditions and the Verbal Arts: A Guide to Research Practices*. Routledge, London
Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York
Geertz C 1988 *Works and Lives: The Anthropologist as Author*. Stanford University Press, Stanford, CA
Gusdorf G 1973 *L'Avènement des sciences humaines au siècle des lumières*. Payot, Paris
Habermas J 1988 *On the Logic of the Social Sciences* (trans. Nicholsen S W, Stark J A). Polity, UK
Hymes D 1974 *Foundations in Sociolinguistics: An Ethnographic Approach*. University of Pennsylvania Press, Philadelphia, PA
Marcus G E, Cushman D 1982 Ethnographies as texts. *Annual Review of Anthropology* **11**: 25–69
Palmer R E 1969 *Hermeneutics: Interpretation Theory in Schleiermacher, Dilthey, Heidegger, and Gadamer*. Northwestern University Press, Evanston
Ricoeur P 1970 *Freud and Philosophy: An Essay in Interpretation* (trans. Savage S). Yale University Press, New Haven, CT
Ricoeur P 1984 The model of the text: Meaningful action considered as a text. *Social Research* **51**(1–2): 185–218
Tedlock D 1983 *The Spoken Word and the Work of Interpretation*. University of Pennsylvania Press, Philadelphia, PA
Tonkin E 1992 *Narrating Our Pasts: The Social Construction of Oral History*. Cambridge University Press, Cambridge, UK
Turner V W 1986 *The Anthropology of Performance*. Performing Arts Journal Publications, New York
Vansina J 1961 *De la tradition orale. Essai de méthode historique*. Musée Royal de l'Afrique Centrale, Tervuren
Vansina J 1985 *Oral Tradition as History*. University of Wisconsin Press, Madison

J. Fabian

Interpretive Methods: Macromethods

The goal of all interpretive methods, both macro and micro, is to make sense of some slice of reality—to render it understandable or intelligible. Since the work of Max Weber (1949, 1958, 1978) has been central to articulating macrointerpretive methods, this essay uses Weber's work to elucidate key features of these methods and to help outline their key tensions.

1. *Webers* Protestant Ethic *as an Illustrative Case*

Weber's (1958) *The Protestant Ethic* exemplifies the meaning-oriented (verstehende) and explanatory goals and strategies of macrointerpretive work. Noting a historical association between the Protestant reformation and the rise of capitalism in Western Europe, Weber outlines a puzzle. How could a religion leading to more *conformist* behavior than did Catholicism become associated with radically *innovative* economic behavior? Unraveling this historical and theoretical puzzle requires understanding the cultural beliefs occasioned by the Reformation and the dilemmas they created for the Protestant believer. Detailing exactly how the new beliefs provided for social control allows Weber to make sense of a culturally significant and historically specific pattern of institutional change.

To do this, Weber constructs the ideal types 'capitalist ethos' and 'Protestant ethic.' *Ideal types* are analytic tools for interpretive, explanatory work. The social scientist creates ideal types by abstracting elements from concrete historical situations, then combining the elements into a meaningful conceptual whole. The ideal type does not exist in empirical reality, but it is constructed inductively from that reality to illuminate something of cultural significance (Weber 1958, p. 47, 1949, pp. 91–111, 1978, pp. 18–22).

According to Weber (1958, pp. 51–68), the spirit of capitalism combines two elements: an absolute ethical duty to accumulate capital and to do so in a rigorously calculating rational way. The ideal-typical capitalist must increase his or her capital as an end in itself, methodically identifying and following courses of action that lead to profit-making, while strictly avoiding any enjoyment of life. Weber argues that any account of the rise of modern capitalism must explain how this ethos, which then helped transform economic institutions, evolved.

The ideal-type Protestant ethic analytically combines three elements (Weber 1958, pp. 79–154). First is Luther's idea of the calling—that people have a moral duty to their work. Second is ascetic mastery of the world, the idea that at each minute of every day, the individual must bring God's glory to fruition in the world, rather than by withdrawing to a monastery or nunnery. Third is the Calvinist doctrine of predestination, arguing that by the decree of God for the manifestation of God's glory, some people are predestined for grace and others for damnation. Only God knows which people are which, and nothing people do in their lives can change their preordained fate.

Once Weber has theoretically constructed his object of inquiry—the capitalist ethos—and his explanatory concept—the Protestant ethic—he creates the interpretive links tying the two together. He asks and answers the question, what does predestination *mean* and how is it likely to orient behavior for the people who believe in it, given that they *also* believe in the calling and ascetic mastery of the world? He then argues as follows.

> In its extreme inhumanity, this doctrine must above all have had one consequence for the life of a generation [that] surrendered to its magnificent consistency. This was a feeling of unprecedented inner loneliness of the single individual. In what was for man

> of the age of the Reformation the most important thing in life, his eternal salvation, he was forced to follow his path alone to meet a destiny which had been decreed for him from eternity … the complete elimination of salvation through the Church and the sacraments was what formed the absolutely decisive difference from Catholicism (Weber 1958, p. 104).

Attaining salvation is so central that believers will question constantly whether they are saved. Since there is no way of knowing, most people will feel intolerable psychological strain. Strain is amplified because at the same time that participation in Protestant churches and rituals cannot guarantee salvation, believers are told by religious leaders that it is 'an absolute duty to consider oneself chosen … lack of self confidence is the result of insufficient faith, hence of imperfect grace' (Weber 1958, p. 111). It becomes imperative to resolve this 'Catch-22' situation and justify certainty in one's salvation.

What will the believer do? Faced with a preordained outcome, some people might become fatalistic and others might commit suicide. To explain why believers do not take these paths, Weber situates predestination in its broader belief context. Because of the conjunction of belief in predestination with beliefs in the calling and ascetic mastery of the world, the Protestant believer is likely to try and relieve his or her incredible psychological insecurity by bringing God's glory to fruition through efficacious action in a calling in the world. 'Good works' cannot bring salvation, but they *do* provide an effective way to get rid of the fear of damnation, assuming that the believer lives an entire 'life of good works combined into a unified system' (Weber 1958, pp. 115–17).

In short, predestination and the need for proof of salvation provide a psychological basis for constant self-regulation and the rational ordering of moral life. The other elements of the Protestant ethic explain how and why the solution to the need for proof will be found in the systematic rationalization of moral life in the world rather than being found elsewhere. As a meaningfully interpreted whole, the Protestant ethic is powerful brew. The interpretive link to the capitalist spirit now is clear: purposive-rational orientation to capital accumulation as an end in itself is a strategy to provide the believer with a clear, continuous and unambiguous sign of his or her own salvation. As a collective attribution of meaning and action orientation, the strategy—and the macro religious structure that created it—unintentionally reorders economic behavior and institutions.

2. *Key Features and Tensions in Macrointerpretive Work*

As illustrated by Weber, imputation of chains of meaning, motivation, and consequent likely courses of action are a hallmark of macrointerpretive explanation. Because actors are front and center, with social institutions providing the cultural scripts that orient actors' attributions of meaning and their behavior, macrointerpretive work is ideally suited to reconciling agency and structure (Sewell 1996). Similarly, macrointerpretive work is profoundly historical, although it is not necessarily historicist in the narrowest meaning of that term.

Because most macrointerpretive work is concerned with imputing actors' meanings, culture and ideology loom large as objects of inquiry or as explanatory factors (Wuthnow 1989, Mahoney 1999). However, macrointerpretive methods also signal the interrelated construction of historical narratives and social science concepts to interpret historical wholes, processes or particularities, so that they are rendered meaningful to us as social scientists (Ragin 1987). These two analytically distinct aspects of macro interpretive work tend to be found together, as in Weber's *Protestant Ethic*, but they sometimes appear separately (Mahoney 1999).

Macrointerpretive scholarship is enriched by multiple analytic techniques including diverse formal and nonformal narrative and comparative methods (see, e.g., Ragin 1987, 2000, Griffin 1993, Stryker 1996, Mahoney 1999). Some scholars view strategies such as narratives stressing sequence, path dependence, and 'eventful temporality' and comparisons emphasizing the divergent structural and cultural contexts of events as alternative approaches to doing macrointerpretive work (Griffin 1992, Sewell 1996). However, others argue these strategies are complementary (Quadagno and Knapp 1992, Stryker 1996, Mahoney 1999).

Like Weber's own work, macrointerpretive scholarship embodies a tension between the concrete and particular and the abstract and general (compare Bendix (1978) with Moore (1966)). Often, macro interpretivists resolve the tension between specificity and generality, as did Weber, by using the concepts they create to resolve puzzles and explain historically specific events and event sequences, while also seeking limited historical, or conditional, generalizations (Quadagno and Knapp 1992, Stryker 1996, Mahoney 1999, Paige 1999). The underlying assumption about the relationship between history and social science theory guiding macrointerpretive work is that history is theorized at the same time as theory is historicized. In other words, the two are 'mutually constructed' and 'mutually adjusted' so that they are 'logically interdependent processes' rather than 'analytically independent things' (Stryker 1996, p. 312).

Macrointerpretive work often is portrayed as wholly inductive. But as Weber's *Protestant Ethic* demonstrates, it cannot be characterized fairly in this way. Neither can its practitioners be characterized as formalistically applying Mill's method of agreement, indirect method of difference or some combination of the two. Notwithstanding much literature describing

comparative variants of macrointerpretive methods in terms of Mill's canons, Goldstone (1997, p. 108) shows why 'Mill's methods cannot be used and are not used by comparative case-study analysis.' Instead, Goldstone (1997) likens case-oriented comparative methods to sophisticated detective work. Like all macrointerpretive analysis, these methods involve a constant interplay of ideas and evidence to solve anomalies, they carefully combine inductive with deductive reasoning, and they involve attending to the complexities of multiple and conjunctural causation (Stryker 1996, Goldstone 1997, Ragin 1987). A final hallmark of macrointerpretive scholarship is selection of research topics according to their 'value-relevance' and 'cultural significance' for the scholar and intended audience (Weber 1949, pp. 76–7, Sewell 1996, Mahoney 1999).

This run-down of multiple currents and tensions within macrointerpretive work is not exhaustive, but it should suffice to show that macrointerpretivists debate (a) whether or not interpretation must involve causal explanation as well as imputing meaning and (b) the prominence that should be accorded to narrative as opposed to comparative techniques. Much, but not all, macrointerpretive work is explicitly or implicitly causal-analytic (Skocpol and Somers 1980, Ragin 1987, Goldstone 1997). Similarly, much macro interpretive work is explicitly or implicitly comparative, even as it relies heavily on narrative strategies (Skocpol and Somers 1980, Mahoney 1999). For a good introduction to diverse causal-analytic strategies and to the multiplicity of comparative techniques employed by macrointerpretivists, see Skocpol and Somers (1980), Skocpol (1984), Ragin (1987), Rueschemeyer et al. (1992), Goldstone (1997), and Mahoney (1999). For a good introduction to the diversity of narrative techniques, including very different concepts of narrativity and its role in macro interpretive social science, see Stone (1979), Griffin (1993), Sewell (1996), Stryker (1996), and Somers (1997).

A final tension involves questions of subjectivity and objectivity in social science. Some contemporary scholars—notably feminist standpoint theorists such as Smith (1990) and Collins (1990)—suggest that it is neither possible nor desirable to transcend the lived experience of the actors one studies. Meanings should be imputed through dialogic techniques practiced with real subjects or, if this is not possible, with historical texts, so that the actors studied would recognize and verify these meanings as their own.

Deconstructionist critiques such as these have been useful to elucidate how class, race, gender, and other categories of domination and subordination are implicated in the production of social science knowledge. However, most macrointerpretive scholars follow Weber (1949) in assuming that *interpretive objectivity* is both possible and desirable. Although there is substantial disagreement over which concepts and procedures are best, most macrointerpretive scholars presume that if the analyst is clear about his or her concepts and how they are operationalized and if he or she employs an explicit set of procedures to establish the interpretations he or she makes, he or she can arrive at interpretive objectivity. Here Weber's (1949, 1978) discussion of the use of ideal types, counterfactuals, and principles of objective possibility to arrive at explanations that are causally adequate and adequate on the level of meaning is seminal. Contemporary promotion of techniques of 'strategic narrative,' (Stryker 1996), of formal qualitative comparative analysis (QCA) and its extension into fuzzy set logic (Ragin 1987, 2000), and of event structure analysis (Griffin 1993), can be seen as updates on Weber's (1949) rules for achieving objectivity in macrointerpretive social science. In these extensions, issues of validity, reliability, and replicability are emphasized.

3. *Conclusion*

In sum, there is no one macrointerpretive method. Macrointerpretive scholarship is built around a diversity of techniques. Both old methodological stalwarts and new developments are playing their role in generating the culturally significant findings of macrointerpretive research. In the twenty-first century, there is an enormous body of writing illuminating the rich variety found within macrointerpretive methods. What all varieties share with Weber's *Protestant Ethic* is an emphasis on imputing meaning to render large-scale social actions and processes intelligible. As comparisons of such diverse studies as Wallerstein's (1974) on the world system and Somers (1997) on English working-class formation suggest, macrointerpretive scholarship of all sorts is well suited to forming and empirically examining concepts, hypotheses, and theories that help us understand how social structures work, how they reproduce themselves and how they change. Although macrointerpretive scholarship relies on analytic logic completely different from that of frequentist statistical methods, it proceeds based on coherent methodological principles and procedures.

See also: Analytic Induction; Archival Methods; Case Study: Methods and Analysis; Case-oriented Research; Causal Counterfactuals in Social Science Research; Causation (Theories and Models): Conceptions in the Social Sciences; Classification: Conceptions in the Social Sciences; Comparative Studies: Method and Design; Content Analysis; Explanation: Conceptions in the Social Sciences; Generalization: Conceptions in the Social Sciences; Grounded Theory: Methodology and Theory Construction; Ideal Type: Conceptions in the Social Sciences; Idealization, Abstraction, and

Ideal Types; Interpretive Methods: Micromethods; Narrative, Sociology of; Weber, Max (1864–1920)

Bibliography

Bendix R 1978 *Kings or People? Power and the Mandate to Rule*. University of California Press, Berkeley, CA

Collins P H 1990 *Black Feminist Thought: Knowledge, Consciousness and Empowerment*. Unwin-Hyman, Boston

Goldstone J 1997 Methodological issues in comparative macrosociology. *Comparative Social Research* **16**: 107–20

Griffin L J 1992 Temporality, events and explanation in historical sociology: An introduction. *Sociological Methods and Research* **20**: 404–27

Griffin L J 1993 Narrative, event-structure and causal interpretation in historical sociology. *American Journal of Sociology* **98**: 1094–133

Mahoney J 1999 Nominal, ordinal, and narrative appraisal in macrocausal analysis. *American Journal of Sociology* **104**: 1154–96

Moore B 1966 *The Social Origins of Dictatorship and Democracy*. Beacon Press, Boston

Paige J M 1999 Conjuncture, comparison, and conditional theory in macrosocial inquiry. *American Journal of Sociology* **105**: 781–800

Quadagno J, Knapp S 1992 Have historical sociologists forsaken theory? Thoughts on the history/theory relationship. *Sociological Methods and Research* **20**: 481–507

Ragin C 1987 *The Comparative Method*. University of California Press, Berkeley, CA

Ragin C 2000 *Fuzzy-set Social Science*. University of Chicago Press, Chicago

Rueschemeyer D, Huber Stephens E, Stephens J D 1992 *Capitalist Development and Democracy*. University of Chicago Press, Chicago

Sewell W Jr. 1996 Three temporalities: Toward an eventful sociology. In: McDonald T J (ed.) *The Historic Turn in the Human Sciences*. University of Michigan Press, Ann Arbor, MI, pp. 245–80

Skocpol T (ed.) 1984 *Vision and Method in Historical Sociology*. Cambridge University Press, New York

Skocpol T, Somers M 1980 The uses of comparative history in macrosocial inquiry. *Comparative Study of Society and History* **22**: 174–97

Somers M 1997 Deconstructing and reconstructing class formation theory: Narrativity, relational analysis, and social theory. In: Hall J R (ed.) *Reworking Class*. Cornell University Press, Ithaca, NY, pp. 73–105

Smith D E 1990 *The Conceptual Practices of Power: A Feminist Sociology of Knowledge*. Northeastern University Press, Boston

Stone L 1979 The revival of narrative: Reflections on a new old history. *Past and Present* **95**: 3–25

Stryker R 1996 Beyond history versus theory: Strategic narrative and sociological explanation. *Sociological Methods and Research* **24**: 304–52

Wallerstein I 1974 *The Modern World System*. Academic Press, New York

Weber M 1949 [1903–17] *The Methodology of the Social Sciences*. Free Press, New York

Weber M 1958 [1904] *The Protestant Ethic and the Spirit of Capitalism*. Scribner, New York

Weber M 1978 [1921], Roth G (ed., trans.) *Economy and Society*. University of California Press, Berkeley, CA, Vol. 1

Wuthnow R 1989 *Communities of Discourse: Ideology and Social Structure in the Reformation, the Enlightenment and European Socialism*. Harvard University Press, Cambridge, MA

R. Stryker

Interpretive Methods: Micromethods

In this article, recent developments in the uses and representations of interpretive, research methods are analyzed. Micromethods are focused on those interpretive approaches which yield close readings of texts and social situations. This will require serious attention to forms of narrative analysis, and the uses to which narrative material can be put by a society and its members, including social scientists. Accordingly, my intentions are fourfold. First, to analyze the problems and issues confronting the use of micro-qualitative methods and research evidence. Second, to define and locate this sprawling, interdisciplinary field of qualitative inquiry within the history of interpretive methods in the United States in the twentieth century. Third, to discuss the multiple uses and forms of qualitative inquiry, including its uses in participatory, applied action research settings. Speculations on where interpretive narrative inquiry will go in the twenty-first century form the conclusion.

1. *The Field of Interpretive Research and its History*

Interpretive research methods produce descriptions and accounts about the ways of life of the writer and those written about. There are multiple forms, or strategies of qualitative research (see below), including ethnography, participant observation, and case study analysis. The use of these strategies is connected to specific methodological practices, ranging from interviewing, to participant observation, to the use of visual, narrative, and personal experience methods. These methods generate specific types of empirical materials. These materials come in multiple forms, ranging from transcribed interviews and field observations, to life stories, cultural texts (stories, music, cinema, soap operas, TV news, televised sporting events), the artifacts and symbols of material culture (building and burial sites, pottery, tools, clothing), photographs, archival records, biographical, and historical documents. Traditionally, these materials are seen as producing the evidence that qualitative researchers use when they formulate interpretations of social and cultural phenomenon. Three interconnected activities define the interpretive research process. They go by a variety of labels or names, including theory,

method, and analysis. Behind these terms stands the researcher who has a gendered biography, and who speaks from a particular class, race, cultural, and ethnic perspective. This researcher confronts the world with a set of ideas, or basic beliefs. These beliefs can be called a paradigm (Guba 1990). A paradigm consists of an epistemology, an ontology, and a set of research strategies. Specific types of methodological practice are specified by a paradigm's epistemology. Ontology raises basic questions about reality and the nature of human nature. Methodology addresses the issue of how we gain knowledge about the world. Today in qualitative inquiry multiple paradigms, each with its particular ontology, epistemology, and methodology, compete for attention. These paradigms include logical postivism, postpositivism, constructionism and feminism, Marxism and critical theory, cultural studies and ethnic models. (Logical positivists believe in an external reality that can be apprehended by the methods of objective science. In contrast, postpositivists assert that there may be an external reality, but it cannot be known perfectly.)

The positivist paradigm, within interpretive methodology, presumes an objective world that can be studied with objective, micro-research methods. The constructivist paradigm presumes multiple realities, and contends that the researcher and those studied create understandings about the world and its meanings. Terms like trustworthiness, credibility, transferability, dependability, and confirmability replace the traditional positivist criteria of internal and external reliability and validity. Feminist, ethnic, critical theory, and cultural studies models privilege a real world that makes a material difference in the lives of men, women, and children. Interpretive methods are used, and researchers experiment with new ways of writing and presenting empirical evidence.

These paradigms, which structure qualitative research, operate in a complex historical field which cross-cuts five historical moments. These five moments simultaneously operate in the present. They are defined as the *traditional* (1900–50), the *modernist*, or golden age (1950–70), *blurred genres* (1970–86), the *crisis of representation* (1986–90) and *postmodern*, or present moments (1990–present). The present moment is defined by the refusal to privilege any method or theory, or methods claim to validity.

Successive waves of critical, epistemological discourse move across these five moments. The *traditional period* is associated with the logical positivist, quantitative research paradigm. In this period qualitative researchers attempted to justify their research methods in terms of traditional criteria of reliability and validity. Researchers sought to maintain a detached, objective stance toward the world and their objects of study. Research methods were the means by which evidence about the world was collected. The other who was studied was alien, foreign, and strange.

The field worker, during this period, was lionized, made into a larger-than-life figure who went into and then returned from the field with stories about strange people. These accounts were structured by the norms of classical ethnography. This sacred bundle of terms (Rosaldo 1989, p. 31) organized ethnographic texts in terms of four beliefs and commitments: a commitment to objectivism, a complicity with imperialism, a belief in monumentalism (the ethnography would create a museum-like picture of the culture studied), and a belief in timelessness (what was studied never changed). This model of the researcher, who could also write complex, dense theories about what was studied, holds to the present day.

The *modernist* and *blurred genres* moments are connected to the appearance of postpositivist arguments. The logic of multiple methods (and triangulation) was adopted, embodied in the development of empirically grounded theory. During this period, researchers started to move away from the detached observer stance of the traditional moment. At the same time a variety of new interpretive, qualitative perspectives made their presence felt, including hermeneutics, structuralism, semiotics, phenomenology, cultural studies, and feminism.

The modernist phase or second moment builds on the canonical works from the traditional period. Social realism, naturalism and slice-of-life ethnographies are still valued. This was the golden age of rigorous qualitative analysis.

By the beginning of the third stage (1970–86), *'Blurred Genres,'* qualitative researchers had a full complement of paradigms, methods, and strategies to employ in their research. Geertz's two books, *Interpretation of Cultures* (1973) and *Local Knowledge* (1983), defined the beginning and end of this moment. In these two works he argued that the old functional, positivist, behavioral, totalizing approaches to the human disciplines was giving way to a more pluralistic, interpretive, open-ended perspective. This new perspective took cultural representations and their meanings as its point of departure. Calling for 'thick descriptions' of particular events, rituals, and customs, Geertz suggested that all qualitative writings were interpretations of interpretations. The observer had no privileged voice in the interpretations that were written. The central task of theory was to make sense out of a local situation.

In the blurred genre phase the humanities became central resources for critical, interpretive theory, and the interpretive research project broadly conceived. The blurred genres phase produced the next stage, the crisis of representation. Here researchers struggled with how to locate themselves, their empirical materials, and their subjects in reflexive texts. It became increasingly clear that research methods were not neutral tools that produced value-free observations. Rather all observations were understood to be theory-laden; facts were not independent of the observer's values; and observers could not sustain an objective,

detached view of themselves, the world, and the research process.

The *fifth moment* is the present. It is characterized by a sensibility which doubts all previous paradigms and historical moments. Theories are now read in narrative terms, as tales from the field (Geertz 1988). Preoccupations with the representation of the 'Other' remain. New epistemologies from previously silenced groups emerge to offer solutions to this problem. The concept of the aloof researcher has been abandoned. Action, activist-oriented research is on the horizon, as are more social criticism and social critique. The search for grand narratives is being replaced by more local, small-scale theories fitted to specific problems and specific situations.

1.1 Reading History

The following conclusions can be drawn from this brief history. Each of the earlier historical moments continues to operate in the present, either as legacy, or as a set of practices that researchers follow or argue against. The multiple and fractured histories of qualitative research make it possible for any given researcher to attach a project to a canonical text from any of the previous historical moments. Multiple criteria of evaluation now compete for attention. There have never been so many paradigms, strategies of inquiry, or methods of analysis to draw upon and utilize. Researchers are in a moment of discovery and rediscovery, as new ways of looking, interpreting, arguing, and writing are debated and discussed. The qualitative research act can no longer be viewed from within a neutral or objective perspective. Class, race, gender, and ethnicity shape the process of inquiry, making research a multicultural process.

1.2 Interpretive Inquiry as a Site of Multiple Methodological Research Practices

Any definition of micro-interpretive research and the evidence generated by qualitative methods must work within this complex historical field. Qualitative research and qualitative evidence mean different things in each of these moments. Nonetheless, an initial, generic definition can be offered.

Interpretive inquiry is multimethod in focus, involving a naturalistic approach to its subject matter. This means interpretive researchers study things in their natural settings, attempting to make sense of these things in terms of the meanings people bring to them. *Qualitative research* involves the studied use and collection of case study, personal experience, introspective, life story, interview, observational, historical, interactional, and visual texts which describe routine and problematic moments and meanings in individual's lives.

Qualitative research, as a set of interpretive practices, privileges no single methodology over another. It has no theory, or paradigm that is distinctly its own. Multiple theoretical paradigms, as argued above, claim use of qualitative research methods and strategies, from constructivism, to cultural studies, feminism, Marxism, and ethnic models of study. Qualitative research does not belong to a single discipline.

These separate and multiple uses and meanings of the methods of qualitative research make it difficult to agree on any essential definition of the field, for it is never just one thing. Still a definition must be made. I borrow from, and paraphrase Nelson, Treichler, and Grossberg's attempt to define cultural studies (1992, p. 4).

> Qualitative research is an interdisciplinary, counter-disciplinary field. It cross-cuts the humanities, the social and the physical sciences. Qualitative research is many things at the same time. It is multi-paradigmatic in focus. Its practitioners are sensitive to the value of the multi-method approach. They are committed to the naturalistic perspective; and to the interpretive understanding of human experience. At the same time the field is inherently political and shaped by multiple ethical and political positions. Qualitative research embraces two tensions at the same time. On the one hand it is drawn to a broad, interpretive, critical sensibility. On the other hand it is drawn to more narrowly defined positivist, postpositivist, humanistic and naturalistic conceptions of human experience and its analysis.

This rather awkward statement means that qualitative research, as a set of interpretive practices, embraces within its own history, constant tensions, and contradictions over the project itself, including its methods, and the forms its findings and interpretations take.

2. The Crisis of Interpretation

Today qualitative researchers confront a double crisis of representation and legitimation. Embedded in the discourses of poststructuralism and postmodernism (Denzin 1997, p. 6, Lather 1993), these two crises are, as Lather (1993) notes, coded in multiple terms, variously called and associated with the narrative, 'interpretive, linguistic, and rhetorical turns' in the social sciences. This linguistic turn makes problematic two key assumptions of social theory and interpretive research. The first assumption, as suggested above, presumes that theorists and researchers can no longer directly capture lived experience; such experience, it is argued, is created in the social text written by the researcher. This is the representational crisis. It confronts the inescapable problem of representation, but does so within a framework that makes the direct link between experience and text problematic.

The second assumption makes the traditional criteria for evaluating and interpretive theory and re-

search problematic. This is the legitimation crisis. It involves a serious rethinking of such terms as validity, generalizability, and reliability, terms already re-theorized in postpositivist, constructionist-naturalistic, feminist, and interpretive discourses. This crisis, asks, 'How are interpretive, ethnographic studies to be evaluated in the poststructural moment?' Clearly these two crises blur together.

2.1 The Representational Crisis

A single but complex issue defines the representational crisis. It involves the assumption that much if not all social science and ethnographic writing is a narrative production, structured by a logic that separates writer, text, and subject matter. Any social text can, accordingly, be analyzed in terms of its treatment of four paired terms: (a) the 'real' and its representation in the text, (b) the text and the author, (c) lived experience and its textual representations, and (d) the subject and his or her intentional meanings. The text presumes a world out there (the real), that can be captured by a 'knowing' author through the careful transcription (and analysis) of field materials (interviews, notes, etc.). The author becomes the mirror to the world under analysis. This reflected world then re-presents the subject's experiences through a complex textual apparatus which typically mingles and mixes multiple versions of the subject. The subject is always a textual construction, for the 'real' flesh-blood person is always translated into either an analytic subject as a social type, or a textual subject who speaks from the author's pages.

Qualitative researchers have historically assumed that their methods probe and reveal lived experience. They have also assumed that the subject's word is always final, and that talk directly reflects subjective-lived experience. The literal translation of talk thus equals lived experience and its representation.

Poststructuralism challenges these assumptions. Language and speech do not mirror experience, they create it and in the process of creation constantly transform and defer that which is being described. The meanings of a subject's statements are, therefore, always in motion. There can never be a final, accurate representation of what was meant or said, only different textual representations of different experiences. As Lather (1993, p. 3) observes, these arguments do not put an end to representation, they signal instead the end of pure presence (Lather 1993, p. 3). The task at hand is to understand what textually constructed presence means, since there is only ever the text. This leads to the question of a text's authority.

2.2 The Legitimation Crisis

Many contemporary qualitative researchers challenge postpositivist arguments concerning the text and its validity. Lather (1993) and Scheurich (1997, p. 84) argue that validity is a text's call to authority and truth. Lather calls this version of validity epistemological. That is, a text's authority is established through recourse to a set of rules concerning knowledge, its production, and representation. These rules, as Scheurich (1997) notes, if properly followed, establish validity. Without validity there is no truth, and without truth there can be no trust in a text's claims to validity. With validity comes power, and validity becomes a boundary line 'which divides good research from bad, separates acceptable (to a particular research community) research from unacceptable research … it is the name for inclusion and exclusion' (Scheurich 1997, p. 84).

Poststructuralism reads the discussions of logical, construct, internal, ethnographic and external validity, text-based data, triangulation, trustworthiness, credibility, grounding, naturalistic indicators, fit, coherence, comprehensiveness, plausibility, truth, relevance, as attempts to reauthorize a text's authority in the postpositivist moment. Such moves still hold (all constructionist disclaimers aside) to the conception of a 'world-out-there' that is truthfully and accurately captured by the researcher's methods and written text.

These words, and the methodological strategies that lie behind them, represent attempts to thicken and contextualize a work's grounding in the external empirical world. They represent efforts to develop a set of transcendent rules and procedures that lie outside any specific research project. These rules, if successfully followed, allow a text to bear witness to its own validity. Hence a text is valid if it is sufficiently grounded, triangulated, based on naturalistic indicators, carefully fitted to a theory (and its concepts), comprehensive in scope, credible in terms of member checks, logical, truthful in terms of its reflection of the phenomenon in question. The text's author then announces these validity claims to the reader. Such claims now become the text's warrant to its own authoritative representation of the experience and social world under inspection.

2.3 Resistances to Qualitative Studies

The academic and disciplinary resistances to interpretive, micro-research further illustrate the politics embedded in this field of discourse. The challenges to qualitative research are many. Qualitative researchers are called journalists or soft scientists. Their work is termed unscientific, or only exploratory, or entirely personal and full of bias. It is called criticism not theory, or it is interpreted politically, as a disguised version of Marxism, or Humanism.

These resistances to qualitative inquiry reflect an uneasy awareness that its traditions commit one to a critique of the positivist project. But the positivist

resistance to qualitative research goes beyond the 'ever-present desire to maintain a distinction between hard science and soft scholarship' (Carey 1989, p. 99). The positive sciences (physics, economics, and psychology) are often seen as the crowning achievements of Western civilization, and in their practices it is assumed that truth can transcend opinion and personal bias (Carey 1989, p. 99). Qualitative inquiry is seen as an assault on this tradition, leading some positivists to retreat into a 'value-free objectivist science' (Carey 1989, p. 104) model to defend their position. But these critics seldom attempt to make explicit, and critique the 'moral and political commitments in their own contingent work' (Carey 1989, p. 104).

The opposition to positive science by the postpositivists and the poststructuralists is seen as an attack on reason and truth. At the same time, the positive science attack on qualitative research is regarded as an attempt to legislate one version of truth over another.

2.4 Action Research

In the contemporary period greater use is being made of qualitative evidence in participatory and applied action programs. Life stories, case histories, and personal narratives are often used as key documentary evidence to establish the need for social change. Such materials are also used as proof that an applied action program works. Indeed qualitative (and narrative) evidence lends itself especially well to those applied action programs which stress the subjective, reflexive dimensions of social experience. They further imply that such materials are critical to the implementation of participatory action research as a practical, collaborative, participatory, emancipatory, and reflexive project. First-person narrative texts allow Third World and indigenous persons to share in the ownership of the research endeavor. Such texts anchor research in the context of ongoing community life. They help community members develop a shared orientation toward collective action aimed at addressing social injustice.

Participatory ethno-dramas, forum, and political theater can be used to mobilize community consciousness around an instance of perceived injustice or repression. For example, Brecht's Epic or Dialectical Theatre is deliberately disruptive and political, and thus anticipates the more radical forms of contemporary postmodern, political theater that has connections back to Artaud's concepts of a pure theatre (Birringer 1993, p. 217). Third World popular theater has been used by 'oppressed Third World people to achieve justice and development for themselves' (Etherton 1988, p. 991). This form of dramatic representation extends what theater and narrative can do politically.

3. In Conclusion

The legacies of the present move in several directions at the same time. They all turn on social texts and narratives, their uses and how they are read. The narrative turn in the social sciences presumes that social texts, including recorded or visual texts, interviews, fieldwork notes, transcribed conversations, speeches, film, music, advertisements, personal, electronic, and archival documents, can be rigorously and scientifically analyzed through the methods of narrative analysis.

But Trinh (1989, p. 141) warns us that traditional, empiricist narrative methods represent an approach to storytelling that must be avoided. They turn the story told into a story analyzed. In so doing they sacrifice meaning for analytic rigor. They only hear and read the story from within a set of predetermined structural categories. They do not hear the story as it was told. The goal is to recover these lost stories.

See also: Interpretation and Translation: Philosophical Aspects; Interpretation in Anthropology; Interpretive Methods: Macromethods; Thick Description: Methodology

Bibliography

Birringer J 1993 *Theatre, Theory, Postmodernism*. Indiana University Press, Bloomington, IN

Carey J W 1989 *Culture as Communication*. Unwin Hyman, Boston

Denzin N K, Lincoln Y S 1994 Introduction. In: Denzin N K, Lincoln Y S (eds.) *Handbook of Qualitative Research*. Sage, Thousand Oaks, CA, pp. 1–17

Denzin N K 1997 *Interpretive Ethnography: Ethnographic Practices for the 21st Century*. Sage, Thousand Oaks, CA

Etherton M 1988 Third World popular theatre. In: Banham M (ed.) *The Cambridge Guide to Theatre*. Cambridge University Press, Cambridge, UK, pp. 991–2

Geertz C 1973 *Interpretation of Cultures: Selected Essays*. Basic Books, New York

Geertz C 1983 *Local Knowledge*. Basic Books, New York

Geertz C 1988 *Works and Lives*. Stanford University Press, Stanford, CA

Guba E G 1990 The alternative paradigm dialog. In: Guba E C (ed.) *The Paradigm Dialog*. Sage, Newbury Park, CA, pp. 17–30

Lather P 1993 Fertile obsession: validity after poststructuralism. *Sociological Quarterly* **34**: 673–93

Nelson C, Treichler P A, Grossberg L 1992 Cultural studies. In: Grossberg L, Nelson C, Treichler P A (eds.) *Cultural Studies*. Routledge, New York, pp. 1–16

Rosaldo R 1989 *Culture and Truth*. Beacon, Boston

Scheurich J J 1997 *Research Method in the Postmodern*. Falmer Press, Washington, DC

Trinh T Minh-ha 1989 *Woman, Native, Other: Writing Postcoloniality and Feminism*. Indiana University Press, Bloomington, IN

N. K. Denzin

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Intrinsic Motivation, Psychology of

Motivation concerns the processes through which behaviors are initiated and directed. Intrinsic motivation is a type of motivation that does not require reinforcers, pertains to activities that individuals find interesting, and requires the nutriments of psychological need satisfaction to function optimally. Herein, we review the emergence of the concept, discuss the field of inquiry as it has evolved, and point to its real-world relevance.

1. Historical Roots of Intrinsic Motivation

During the middle decades of the twentieth century, the initiation and direction of behavior was addressed primarily by learning theory (Hull 1943) and operant theory (Skinner 1953)—the two strands of behaviorism that dominated empirical psychology at that time. The two theoretical approaches shared several fundamental beliefs and theoretical propositions, the most important being that both claimed external reinforcers were the basis for behavioral regulation. The most important difference was that Hullian theory gave content to motivational processes by defining reinforcements in terms of the reduction of universal physiological needs, whereas Skinnerian theory defined reinforcements functionally as consequences of behavior that change its likelihood of recurring.

During that same period, studies began to uncover anomalies related to exploration and manipulation that could not be explained by either behavioral approach (Harlow 1953). Thus, White (1959) stated that a new type of motivational concept was necessary. Calling it effectance motivation and giving it a critical role in development, White suggested that people are inherently motivated to be effective in dealing with their surroundings. Subsequently, de Charms (1968) proposed an additional motivational propensity that does not operate by reinforcement principles, calling it personal causation.

Deci (1975) used the term intrinsic motivation to encompass both ideas about people's inherent tendency towards effective, volitional engagement with the environment. Suggesting that people are intrinsically motivated to take interest in novelty, assimilate experience, and apply their acquired skills flexibly, he defined the concept both operationally and theoretically. The operational definition, which was formulated in contrast to the Skinnerian assertion that all 'voluntary' behaviors are controlled by reinforcements in the environment, stated that intrinsically motivated behaviors are self-rewarding and thus do not require reinforcers that are operationally separable from the behaviors. The theoretical definition of intrinsic motivation, which was formulated in contrast to the Hullian assertion that all behaviors are derived from the four universal physiological needs, defined intrinsically motivated behaviors as those that people do out of interest when they experience satisfaction of the universal psychological needs for competence and autonomy.

At about the time reinforcement processes were found inadequate for explaining a variety of behaviors, intention emerged as the central concept for explaining behavioral regulation (Heider 1958). Accordingly, people undertake intrinsically motivated activities expecting to find them interesting and enjoyable, assuming they feel competent and self-determined while doing them.

2. The Organismic Metatheory

The concept of intrinsic motivation was a dramatic departure from motivation as it had been viewed in experimental psychology, for it eschewed the ideas of passivity and conditioning in favor of natural proactivity and growth. As such, it necessitated a significant shift from a mechanistic toward an organismic metatheory within the psychology of motivation. It became essential to view people as self-initiating beings who learn and develop as they proactively engage the environment. Intrinsic motivation represents the prototype of self-initiated behavior.

The explication and early investigation of intrinsic motivation involved contrasting it with extrinsic motivation, which refers to engaging in an activity for consequence other than the spontaneous feelings that accompany it. The classic example of extrinsic motivation is doing an activity to attain a reward or avoid a punishment. Employing attribution theory, de Charms (1968) stated that intrinsically motivated behaviors have an internal perceived locus of causality, while extrinsically motivated behaviors have an external perceived locus of causality, and Lepper et al. (1973) defined intrinsic motivation as behaviors for which people infer internal causes, and extrinsic motivation as those for which they infer external causes. Kruglanski (1975) proposed that when intrinsically motivated, rewards are endogenous to the activity, whereas when extrinsically motivated, rewards are exogenous.

3. Research on Intrinsic Motivation

Initial research on intrinsic motivation examined the effects of extrinsic rewards on intrinsic motivation, finding that extrinsic rewards did undermine intrinsic motivation within limiting conditions (e.g., Deci 1975). Specifically, the undermining occurred when the activity was initially interesting and the rewards were both contingent on doing the activity and expected while doing it.

This general finding, which has now been replicated numerous times was controversial when it first appeared and has remained controversial ever since. In fact, the controversy was very apparent recently when Eisenberger and Cameron (1996) published a high visibility article in which they reported having done a meta-analysis of experiments examining reward effects on intrinsic motivation and having found little evidence for the undermining effect. However, Deci et al. (1999) subsequently pointed out many errors and inappropriate procedures in that Eisenberger and Cameron meta-analysis and did a new meta-analysis that corrected the previous shortcomings. The newer meta-analysis, involving 128 studies, confirmed, in fact, that tangible extrinsic rewards do reliably undermine intrinsic motivation for the rewarded activity, as had been concluded previously in narrative reviews of the literature (e.g., Deci and Ryan 1985).

Part of the significance of the undermining effect is that it indicates that motivation is not a unitary concept, varying only in amount, because these two types of motivation—intrinsic and extrinsic—are not additive. Indeed, it has turned out that the orientation of motivation—whether a behavior is intrinsically or extrinsically motivated—is more important than the overall amount of motivation for predicting behavior and its consequences (Ryan and Deci 2000).

In fact, numerous studies have shown that intrinsic, relative to extrinsic, motivation was associated with better conceptual learning, greater creativity, more cognitive flexibility, greater behavioral persistence, and enhanced well-being (Deci and Ryan 1985, Utman 1997). Given these findings, a thorough investigation of the social–contextual conditions that facilitate versus impede intrinsic motivation seemed essential.

Because the needs for competence and autonomy were theorized to be crucial nutriments for intrinsic motivation, contextual factors that allowed satisfaction of the needs for competence and autonomy were predicted to enhance intrinsic motivation, whereas those that thwarted satisfaction of these basic needs were predicted to undermine intrinsic motivation.

Studies concerned with competence showed that positive performance feedback enhanced intrinsic motivation, that negative feedback diminished it, and that perceived competence mediated these effects. Other research found that increases in perceived competence must be accompanied by feelings of autonomy in order for the perceived competence to enhance intrinsic motivation. In fact, far more studies have examined autonomy than competence as a mediator of contextual effects on intrinsic motivation. This research has shown that threats, deadlines, directives, and pressured competition, like extrinsic rewards, diminished intrinsic motivation, presumably because people tend to experience them as limiting their autonomy (i.e., as controlling their behavior). On the other hand, choice, the opportunity for self-direction, and acknowledging people's feelings have been found to enhance intrinsic motivation because they provide a greater sense of autonomy (Deci and Ryan 1985).

Other research conducted in real-world settings such as homes, classrooms, and work groups revealed a general interpersonal style referred to as autonomy supportive (in contrast to controlling) that, when used by parents, teachers, and managers, catalyzed greater intrinsic motivation and self-determination in their children, students, and employees. Thus, both specific external events, such as the offer of rewards, and general interpersonal ambiences can affect intrinsic motivation, enhancing it by supporting people's feelings of autonomy and competence and undermining it by diminishing those feelings (Deci and Ryan 1985).

4. *Extrinsic Motivation*

The abundance of research showing that extrinsic motivators such as rewards, deadlines, threats, and evaluations tend to undermine intrinsic motivation would seem to imply that extrinsic motivation is invariantly nonautonomous. However, research has shown that extrinsic motivation can become autonomous or self-determined through the processes of internalizing and integrating behavioral regulations and their underlying values (Ryan and Deci 2000). Internalization is the process of taking in a value or regulation, and integration is the process of transforming that value or regulation into one's own.

Self-determination theory (Deci and Ryan 1985) suggests that when extrinsic motivation has merely been taken in but not transformed or integrated it will continue to control and diminish feelings of autonomy. An example is ego-involvement (i.e., internal pressure to do a behavior in order to feel worthy), and studies have shown that this type of motivation is antagonistic to autonomy (Deci and Ryan 1985).

However, extrinsic motivation that has been more fully integrated allows the feeling of autonomy and has been found to predict more positive outcomes, such as better school performance, greater persistence at healthy behaviors, and enhanced psychological well-being not only in the USA but in various countries including Japan (Ryan and Deci 2000).

Research has also considered the effects of social contexts on internalization and integration, focusing primarily on whether contextual factors tend to be controlling versus supportive of self-determination. In both laboratory and field settings, these studies demonstrated that providing a meaningful rationale for an uninteresting behavior, along with supports for autonomy and relatedness, promoted internalization and integration. Controlling contexts (relative to contexts that allowed need satisfaction) yielded less overall internalization, as evidenced by less subsequent behavior, and the internalization that did occur in controlling contexts was less well integrated (Deci et

al. 1994). Field studies using both cross-sectional and longitudinal designs also found fuller internalization of values and regulations when parents (Grolnick and Ryan 1989) and instructors (Williams and Deci 1996) were more supportive of autonomy, competence, and relatedness, thus suggesting that supports for these three basic needs facilitate the internalization and integration of extrinsic motivation.

5. Conclusions

The central element in the psychology of intrinsic motivation is the critical distinction between behaviors that are volitional and accompanied by the experience of freedom and autonomy—those that emanate from one's sense of self—and behaviors that are accompanied by the experience of pressure and control and are not manifestations of self (Ryan and Deci 2000). Intrinsically motivated behaviors, which are performed out of interest and are characterized by perceived competence and autonomy, are the prototype of self-determined behaviors.

Research examining internalization and integration of extrinsic motivation and the conditions that facilitate it is important for the psychology of intrinsic motivation for two main reasons. It shows first, that although the naturally occurring state of intrinsic motivation is the prototype of volitional, self-determined activity, extrinsic motivation can approximate that state if the organismic processes of internalization and integration have operated effectively, and second, that the inherent processes of intrinsic motivation and integration of extrinsic motivation are similarly facilitated in social contextual conditions that allow satisfaction of people's innate psychological needs for competence, autonomy, and relatedness.

Thus, it seems that interpersonal contexts created by teachers, parents, managers, healthcare providers, and other individuals in positions of responsibility can have a substantial impact on the degree to which others, such as students, children, employees, and patients maintain their intrinsic motivation and become more volitional and self-determined with respect to extrinsic motivation.

See also: Academic Achievement Motivation, Development of; Curiosity, Psychology of; Motivation and Actions, Psychology of; Motivation: History of the Concept; Motivation, Learning, and Instruction; School Achievement: Cognitive and Motivational Determinants; Schooling: Impact on Cognitive and Motivational Development

Bibliography

de Charms R 1968 *Personal Causation: The Internal Affective Determinants of Behavior*. Academic Press, New York

Deci E L 1975 *Intrinsic Motivation*. Plenum, New York

Deci E L, Eghrari H, Patrick B C, Leone D R 1994 Facilitating internalization: the self-determination theory perspective. *Journal of Personality* **62**: 119–42

Deci E L, Koestner R, Ryan R M 1999 A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* **125**: 627–68

Deci E L, Ryan R M 1985 *Intrinsic Motivation and Self-determination in Human Behavior*. Plenum, New York

Eisenberger R, Cameron J 1996 Detrimental effects of reward: reality of myth? *American Psychologist* **51**: 1153–66

Grolnick W S, Ryan R M 1989 Parent styles associated with children's self-regulation and competence in school. *Journal of Educational Psychology* **81**: 143–54

Harlow H F 1953 Mice, monkeys, men, and motives. *Psychological Review* **60**: 23–32

Heider F 1958 *The Psychology of Interpersonal Relations*. Wiley, New York

Hull C L 1943 *Principles of Behavior: An Introduction to Behavior Theory*. Appleton-Century-Crofts, New York

Kruglanski A W 1975 The endogenous–exogenous partition in attribution theory. *Psychological Review* **82**: 387–406

Lepper M R, Greene D, Nisbett R E 1973 Undermining children's intrinsic interest with extrinsic reward: a test of the 'overjustification' hypothesis. *Journal of Personality and Social Psychology* **28**: 129–37

Ryan R M, Deci E L 2000 Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* **55**: 58–68

Skinner B F 1953 *Science and Human Behavior*. Macmillan, New York

Utman C H 1997 Performance effects of motivational state: a meta-analysis. *Personality and Social Psychology Review* **1**: 170–82

White R W 1959 Motivation reconsidered: the concept of competence. *Psychological Review* **66**: 297–333

Williams G C, Deci E L 1996 Internalization of biopsychosocial values by medical students: a test of self-determination theory. *Journal of Personality and Social Psychology* **70**: 767–79

E. L. Deci and R. M. Ryan

Introspection: History of the Concept

The term *introspection* refers to the self-observation of mental events. Historically, it is a relatively recent term that achieved prominence in the course of debates about the status and the subject matter of psychology as a science. Critics stressed the incompatibility of self-observation and objective, scientific, observation; defenders proposed versions of introspection that could lead to reliable results.

1. The Philosophical Background

1.1 Origins

There is no documented occurrence of the term *introspection* before the late seventeenth century.

Serious discussion of the phenomenon described by this term began a century later and only gathered momentum during the course of the nineteenth century. Positivists like Auguste Comte now began to contrast the reliable, objective methods of science with the speculations of philosophers supposedly based on self-observation. Introspection would then be as old as philosophy, a claim that is hardly supported by an analysis of the relevant early texts.

The origins of the modern concept of introspection are closely tied to the rise of an empiricist philosophy at the end of the seventeenth century. For John Locke (1690) all knowledge was based on experience, but there were two sources of experience: the external senses and 'the perception of our own mind within us' which he likened to an 'internal sense.' The modern concept of introspection emerged out of the empiricist philosophers' belief that self-knowledge was based on acts of internal observation that were closely analogous to the acts of observation by which we gain knowledge of the external world.

1.2 The Classical Controversy

Objections to the empiricist concept of introspection were launched from two entirely different directions, though there was some convergence on specific criticisms. One set of objections came from German Idealist philosophers, the other from scientific positivists like Comte. Among the first group a beginning was made by Immanuel Kant, who gave several reasons why an empirical study of the psyche could never amount to a natural science. Among these reasons was the consideration that one could not observe one's mental states without thereby altering them. This position was based on a rejection of the empiricist analogy between knowledge of the natural world and self-knowledge. At best, self-observation would yield information about the self as it appears to itself, not as it really is. Idealism led to the development of methods of investigation in the human sciences that were neither introspective nor natural scientific in character.

Scientific *positivism*, on the other hand, rejected introspection along with any knowledge gained by methods other than those of the natural sciences. These methods demanded a strict separation of the observer and the objects observed, a separation that, according to Comte, no one could achieve in regard to their own mind. Introspection did not merely modify what was observed, it was actually impossible.

Safeguarding the foundations of philosophical empiricism, John Stuart Mill replied to Comte by offering what came to be the classical defense of introspection: mental events do not disappear in a flash, they are remembered. What is actually observed during introspection are the memories of recently past mental events, not the events as they occur. In other words, introspection is really retrospection, the examination of memory images. This defense was taken over by William James (1890) and other empiricists, though by the closing years of the century their advocacy of introspection was qualified by warnings about possible errors of memory. Some conceptual clarification was introduced by Franz Brentano (1973/1874), who distinguished between the immediate awareness of ongoing mental events, which he called 'internal perception,' and 'self-observation,' the retrospective attending to past mental events. It can be confusing to refer to both as 'introspection.'

2. Introspection and Experiment

The turn to experimental methods of psychological investigation was closely linked to the controversial nature of introspection. Wilhelm Wundt, who played a major role in the launching of experimental psychology, believed that the psychological laboratory provided the only conditions under which reliable introspective data could be gathered (Danziger 1990). He accepted the distinction between inner perception and self-observation but, unlike his predecessors, totally rejected the latter. Instead, he advocated a control of the conditions of internal perception so as to approximate the conditions of external perception. This could be accomplished in the laboratory by such measures as the repetitive presentation of relatively simple stimuli and the requirement that reactions be immediate and automatic. However, adherence to these conditions would necessarily limit the scope of experimental psychology (and of valid introspection) to the investigation of 'elementary' psychological events, mainly in the areas of sensation and reaction times.

At the beginning of the twentieth century these restrictions were rejected by a new generation of experimental psychologists who greatly extended the scope of experimental introspection so as to enable them to investigate more complex psychological processes like thinking and problem solving. The reliability of results was now thought to depend on the adoption of certain attitudes while introspecting. But there was no unanimity about the nature of these attitudes. E. B. Titchener (see *Structuralism*, *History of Structuralism*), a major exponent of experimental introspectionism, demanded that introspective descriptions should be in terms of simple, irreducible, units and should abstract from any meaning the stimulus might have. In Europe, members of the Wurzburg School (see Humphrey 1951) emphasized the importance of clearing one's mind of preconceptions while introspecting. (Titchener's requirements were prime examples of such preconceptions). Later, the *Gestalt* psychologists extended this *phenomenal* approach by insisting that perceptions should be described exactly as they appear in daily life, i.e., in terms of interrelated, meaningful, patterns.

The investigations of the Wurzburg School raised a new kind of question: what is the relationship between the conscious experience of the introspector and the verbal report he or she makes of this experience? In the empiricist tradition it had always been assumed that the relationship is one of description, not different in principle from a description of objects experienced as located outside oneself. However, it now appeared that some of the verbal reports in introspection experiments could not be regarded as descriptions of inner states, they were much more like *expressions* of these states, e.g., when someone reports feeling: 'Oh no, not that again!' The German term for this is *Kundgabe*, which has connotations of broadcasting and proclaiming. This distinction pointed towards recognition of the communicative features of introspective reports. It also suggested that the result of treating introspective reports purely as descriptions of mental events was likely to be a misleadingly intellectualistic account of such events.

3. *Behaviorism and its Aftermath*

Divergent attitudes to the practice of experimental introspection led to divergent results among different laboratories. Titchener claimed that sensory images could always be discovered in any thought activity but this was disputed by the Wurzburg School, by Binet in France, and by others. This became known as the 'imageless thought controversy.' Situations in which the results from one laboratory are at variance with those from other laboratories are far from unknown in the history of science. However, the imageless thought controversy acquired a special significance because its existence was used as evidence for the uselessness of introspection by a new movement in American psychology, namely, *behaviorism*.

Early behaviorists like J. B. Watson considered introspection to be entirely bogus, while others, like K. S. Lashley, conceded that it might provide 'cues to physiological problems.' Certainly, no scientific hypothesis could ever be verified by means of introspection. But in retrospect it seems that the behaviorists drew unwarranted conclusions from the practical difficulties of introspection (Mackenzie 1977, Howe 1991). Both pragmatic and metaphysical reasons were operative here. Pragmatically, it seemed advisable to dispense with a practice that would never yield the kind of knowledge that the science of psychology increasingly aspired to, i.e., knowledge that would be useful in dealing with practical psychological problems on a large scale. Metaphysically, the behaviorists were committed to the elimination of mind from the category of things that could have effects in the real world. Therefore, the study of mental events was pointless.

Behaviorism was more successful at denigrating introspection than at offering an alternative explanation of what actually happened when someone claimed to be introspecting. Its attempts to do so generally involved assumptions about the existence of subtle physical processes like subliminal movements of the vocal cords and other as yet undetectable examples of 'covert' behavior. It was these, rather than mental states, that supposedly gave rise to introspective reports. Behaviorism effectively accepted the original philosophical analogy between internal and external observation. It differed from those it called introspectionists only with regard to the kind of stuff that was the target of internal observation.

There was a strong rhetorical element in the behaviorists' critique of introspection which led to the construction of historical legends. One such legend was that of a category of 'introspective psychology' that comprised prebehaviorist and antibehaviorist psychologies in general. This blurred the significant distinctions that existed among those not prepared to outlaw all forms of introspection. None of them had characterized themselves as 'introspectionists' and all had often been critical of many forms of introspection.

Debates about the nature of introspection reached their height during the first quarter of the twentieth century and were followed by a period during which the topic was usually regarded as a closed chapter in the history of psychology. Subjects in psychological experiments could still be questioned about the reasons for their responses, but their replies were treated as 'verbal reports,' a form of overt behavior, not a description of internal mental events. Theoretical discussions about the nature of introspection were largely restricted to professional philosophers like Russell, Wittgenstein, and Ryle and had little or no influence on scientific psychology. One exception was the latter day behaviorist B. F. Skinner (1953) who continued to regard introspective reports as an effect of various forms of covert behavior, but also added the possibility that at times they might be no more than a claim by a person to have accomplished a particular discriminative task. Disappointing evidence in the area of 'covert behavior' as well as the unscientific multiplication of highly speculative hypotheses eventually robbed the behaviorist position of its early plausibility.

More recently, the question of introspection has resurfaced in the context of computer models of mind and consciousness (Dennett 1978; see also *Digital Computer: Impact on the Social Sciences*). Here, introspective reports are seen as the output of some kind of language processing system that receives input from various sources, notably, one or more memory systems, an attentional perceptual system, and perhaps a central 'control' system. The nature of the output will depend on the 'information' that is in these systems and on their functional inter-relationships. In some versions of the model the final subsystem only has access to the results of previous information processing, not the processing itself. This accounts for

certain inadequacies of introspective evidence; problems of memory account for others (see Ericsson and Simon 1993).

Other recent contributions have begun to be concerned with the obvious fact that all experimental introspection and much everyday introspection occurs in a social context and that introspective reports are always limited by the linguistic resources available to the reporter (Lyons 1986). The content of introspective reports is therefore circumscribed by social and cultural factors.

See also: Behaviorism; Empiricism, History of; Experimentation in Psychology, History of; Philosophy, Psychiatry, and Psychology; Positivism, History of; Psychoanalysis: Overview; Reflexivity in Anthropology; Self: History of the Concept; Self-knowledge: Philosophical Aspects; Unconscious: History of the Concept; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Brentano F C 1973/1874 *Psychology from an Empirical Standpoint*. Routledge, London

Danziger K 1990 *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge University Press, Cambridge, UK

Dennett D C 1978 *Brainstorms*, 1st edn. Bradford Books, Montgomery, VT

Ericsson K A, Simon H A 1993 *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA

Howe R B K 1991 Introspection: A reassessment. *New Ideas in Psychology* **9**: 25–44

Humphrey G 1951 *Thinking*. Methuen, London

James W 1890 *Principles of Psychology*. Holt, New York

Lyons W 1986 *The Disappearance of Introspection*. MIT Press, Cambridge, MA

Mackenzie B D 1977 *Behaviourism and the Limits of Scientific Method*. Routledge, London

Skinner B F 1953 *Science and Human Behavior*. Macmillan, New York

K. Danziger

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Inv

Invariance in Statistics

1. *Invariance in Statistics*

The use of the word invariance in a statistical context most often refers to situations where a statistical model transforms in a suitable way when potential observations are transformed. Describing this context requires the specification of two ingredients. The first is the collection of probability distributions which specify the statistical model. The second is the group of transformations which describes what type of invariance is under consideration. Using the invariance of a statistical model to suggest inferential schemes, such as estimators or tests of hypothesis, is one of the most important applications of invariance reasoning. Such uses of invariance, via concrete examples, form the connective tissue in this article. These examples provide applications of the *principle of invariance* which asserts that invariant problems should have invariant solutions.

The ideas of invariance provide not only the motivation for many statistical procedures but supply the insight and techniques for establishing both distributional and optimality properties of modes of inference. This is most pronounced in the area of linear statistical models—univariate as well as multivariate. For example, in standard regression and analysis of variance models, the affine invariance is reflected in maximum likelihood estimation, in classical F-testing procedures as given by ANOVA tables, and in the structural equation approach to linear models (Fraser 1968) (see *Fiducial and Structural Statistical Inference*). In addition, nonparametric procedures based on ranks and inference based on permutation distributions receive their basic statistical justification from invariance considerations. A full explication of the role of invariance in statistics requires no mean amount of mathematical and statistical theory. However, much of the motivation and pioneering work can be appreciated without all the formal details. Readers wishing just to get the main ideas are advised to skip Sect. 4 and focus on the concrete examples given in Sects. 2, 5, 7, and 8 of what follows.

2. *An Example Due to E. J. G. Pitman*

One of the first full fledged invariance arguments to appear in the statistical literature of estimation is due to E. J. G. Pitman 1939). Here is a special case of his argument which provided motivation for later work on invariant estimation.

Example 1. Suppose f is probability density function on the real line R^1. For each real number $\theta \in R^1$,

$$f(u-\theta), \quad u \in R^1 \tag{1}$$

is a density with a *translation parameter* θ. The density in (1) serves as a *translation model* in some very simple experimental situations. Now, suppose that $X_1, \ldots, X_n$ are independent and identically distributed according to (1). To describe the statistical model for the data $X_1, \ldots, X_n$, first form the $n \times 1$ column vector X with coordinates $X_1, \ldots, X_n$. On Euclidean n-space R^n, X has a density

$$p(x \mid \theta) = \prod_{i=1}^{n} f(x_i - \theta) \tag{2}$$

where x is an element of R^n with coordinates $x_1, \ldots, x_n$. Equation (2) gives the likelihood of the data $x = (x_1, \ldots x_n)'$ where the prime denotes vector transpose. The density p in (2) together with $\theta \in R^1$ specifies the *statistical model* for the vector X.

Now, let c be a real number and consider the transformed random variables and parameter

$$\tilde{x}_i = X_i + c, \quad \tilde{\theta} = \theta + c \quad \text{for } i = 1, \ldots, n \tag{3}$$

If X has a distribution corresponding to θ, then $\bar{X} = (\tilde{x}_1, \ldots, \tilde{x}_n)'$ has a distribution corresponding to $\tilde{\theta}$. In other words, if X_i are all translated by c, then the model parameter is translated by c. For this reason the model for X is said to be *invariant under translation*.

Next suppose the statistical problem is to estimate θ. Given an estimator

$$T = t(X_1, \ldots, X_n) \in R^1 \tag{4}$$

let

$$r_T(\theta) = \mathscr{E}_\theta (T-\theta)^2 \tag{5}$$

be the mean squared error of T. Here, $\mathscr{E}$ denotes expectation under the distribution of X.

In essence Pitman now argued as follows:

> Consider an estimate $t(x_1, \ldots, x_n)$ of θ. From the invariance of the model, $t(x_1+c, \ldots, x_n+c)$ is an estimate of $\theta+c$. Thus,

$t(x_1+c, \ldots, x_n+c)-c$ is also an estimate of θ. Equating these two estimates yields

$$t(x_1+c, \ldots, x_n+c) = t(x_1, \ldots, x_n)+c \tag{6}$$

for all $x_1, \ldots, x_n$ and all c.

An estimate satisfying (6) is called an *invariant estimate*. Pitman asserted that the invariance of the model strongly suggested that only invariant estimates of θ should be considered. He then showed that, within the class of invariant estimators, there is one which is *best* in the sense that it has the smallest mean squared error (see (5)) for every θ. This best invariant estimate is given by the formula

$$t_0(x) = \int_{-\infty}^{\infty} \theta p(x \mid \theta)\,\mathrm{d}\theta \Big/ \int_{-\infty}^{\infty} p(x \mid \theta)\,\mathrm{d}\theta, \quad x \in R^n \tag{7}$$

To say that the *Pitman estimator* $T_0 = t_0(X)$ is *best invariant* means that if T is any invariant estimator, then $r_{T_0}(\theta) \leqslant r_T(\theta)$ for all $\theta \in R^1$.

The argument leading to the best invariant estimator T_0 is based on two observations. First, the model given by (2) for the random vector X transforms in a very simple way when the data are translated (see (3)). Second, the problem of estimating θ also transforms in a simple way and rather naturally leads to estimates satisfying (6)—namely, the invariant estimates. Generalizing Pitman's argument to cover other inferential problems is much of what invariance is about, although the earlier work of Hotelling (1936) had already introduced invariance into hypothesis testing. The next sections describe in mathematical terms what is meant by an invariant statistical model. It is helpful to keep the Pitman example in mind.

3. Statistical Models

Consider an experimental situation which will give rise to a vector $X = (X_1, \ldots, X_n)'$ of observations. In essence, a statistical model for this situation is a collection of probability distributions for X, one of which is supposed to be an adequate description of the potential variability that might be expected in X. Because X is an n-dimensional column vector, the distributions for X are defined on R^n. The collection of distributions is often written $\{P(\cdot \mid \theta) \mid \theta \in \Theta\}$ and the index θ for the collection is a *parameter*. The set Θ is called the *parameter space*. The collection $\{P(\cdot \mid \theta) \mid \theta \in \Theta\}$ is the *statistical model* for X. Statistical models are typically based on a detailed knowledge of the science underlying the generation of the vector X, the type of data similar experiments have yielded, plus a variety of other considerations including analytic tractability.

In Example 1, the sample space is R^n, the parameter space is $\Theta = R^1$, and $P(\cdot \mid \theta)$ is given by

$$P(A \mid \theta) = \int_A p(x \mid \theta)\,\mathrm{d}x \tag{8}$$

where A is a subset of R^n. The density $p(\cdot \mid \theta)$ is given in (2). Therefore $\{P(\cdot \mid \theta) \mid \theta \in \Theta\}$ is the model for the Pitman example when $P(\cdot \mid \theta)$ is given by (8).

4. Groups of Transformations

A simple but very useful type of algebraic system is a group. In essence, a *group* is a set G together with a binary operation $\circ$ so that the following hold:

(a) if g_1 and g_2 are in G, then $g_1 \circ g_2$ is in G. Thus G is *closed* under the binary operation;

(b) there is an element e in G called the *identity* so that $e \circ g = g \circ e = g$ for all $g \in G$;

(c) each element g has an *inverse*, g^{-1}, which satisfies $g \circ g^{-1} = g^{-1} \circ g = e$; and

(d) for g_1, g_2 and g_3 in G, $g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$. Thus the binary operation is *associative*.

It is common notational practice to delete the binary operation $\circ$ and write $g_1 g_2$ for $g_1 \circ g_2$. This convention is followed here. A standard example of a group is the set R^n where the binary operation is addition. A more interesting example is the set Gl_n of all $n \times n$ nonsingular matrices where the binary operation is matrix multiplication. For further examples see MacLane and Birkhoff (1967).

The notion of a group of transformations arises when each element g of a group G can also be thought of as defining a bijective (one-to-one and onto) transformation on a set $\mathscr{X}$. Here are two familiar examples.

Example 2. Take G to be the set of real numbers and the binary operation to be addition. Let $\mathscr{X}$ be Euclidean n-space R^n and let x_0 be a fixed vector in R^n. Each $g \in G$ defines a bijection of R^n to R^n given by $x \to x + gx_0$. Here gx_0 is the real number g times the vector x_0. The above notation means that the function defined by g maps x onto $x + gx_0$. For example, if x_0 consists of the vector of all ones, then $x + gx_0$ is just the vector x with g added to each coordinate. This case arose in Example 1—see (6).

Example 3. Take G to be all $n \times n$ nonsingular matrices with the binary operation of matrix multiplication. With $\mathscr{X} = R^n$, each g defines a bijective transformation $x \to gx$ where gx is the matrix g times the vector $x \in R^n$.

Although not completely obvious in the above two examples, the group operation 'fits together' with function composition when the group element is thought of as a bijective function. In Example 3, what this means is that

$$A_1(A_2 x) = (A_1 A_2)x \tag{9}$$

The left hand side of (9) is the 'function' A_1 evaluated at $A_2 x$. Hence $A_1(A_2 x)$ is the composition of the two functions A_1 and A_2 in that order. The right hand side of (9) is the 'function' $A_1 A_2$ evaluated at x where $A_1 A_2$ is the group element formed by composing the two

group elements A_1 and A_2. This 'fitting together' is what defines a *group of transformations*.

Let $\mathscr{X}$ be a set, G be a group, and assume each $g \in G$ defines a bijective transformation from $\mathscr{X}$ to $\mathscr{X}$. The value of the transformation g at x is written gx. For notational convenience, the action of the function g on $\mathscr{X}$ is written

$$x \to gx \tag{10}$$

That is, x is mapped onto gx by the function g.

The group G is said to *act on* $\mathscr{X}$ if

(a) $g_1(g_2x) = (g_1g_2)x$ for $x \in \mathscr{X}$ and $g_1, g_2 \in G$.

(b) $ex = x$ for $x \in \mathscr{X}$ where e is the identity in G.

Conditions (a) and (b) imply that the group inverse of g is the function inverse of g when g is thought of as a bijective transformation. When G acts on $\mathscr{X}$, G is a *group of transformations*.

In statistical applications, it is common for a group G to act on more than one set in the same problem. This happens in Example 1 where G acts on the sample space R^n as noted in Example 2 and G acts on the parameter space $\Theta = R^1$ (3). More complicated examples are given in the next section.

When the group G acts on $\mathscr{X}$, the set $\{gx \mid g \in G\}$ is called the *orbit* of x. The orbits decompose $\mathscr{X}$ into disjoint sets. Further, G defines an equivalence relation on $\mathscr{X}$, say $\cong$, where $x_1 \cong x_2$ if and only if $x_1 = gx_2$ for some $g \in G$. The partition of $\mathscr{X}$ induced by $\cong$ consists exactly of the orbits (MacLane and Birkhoff 1967). A function h defined on $\mathscr{X}$ is called *invariant* if $h(x) = h(gx)$ for all $x \in \mathscr{X}$ and $g \in \mathscr{X}$. In other words, h is invariant if and only if h is constant on orbits. Functions h which are invariant and which take different values on different orbits are called *maximal invariants* because such functions identify the orbits in $\mathscr{X}$. Maximal invariants are important objects in the study of invariant hypothesis testing problems, which are discussed later in this article. Here is a simple example of a maximal invariant.

Example 4. Take $\mathscr{X} = R^n$ and $G = O_n$, the group of $n \times n$ orthogonal matrices. Then G acts on R^n by matrix multiplication of vectors

$$x \to gx, \quad x \in R^n \quad \text{and} \quad g \in O_n$$

The orbits in R^n are $S_r = \{x \mid x'x = r^2\}$ where $r \geqslant 0$. The orbit S_r is just the rim of the sphere of radius r in R^n. Then $h(x) = x'x$ is a maximal invariant.

5. *Invariant Models*

Coupling the idea of a statistical model with that of a group acting on a set leads naturally to the notion of an invariant statistical model. Consider a random vector X with values in a sample space $\mathscr{X}$ and a statistical model $\{P(\cdot \mid \theta) \mid \theta \in \Theta\}$ for the distribution of X. The parameter θ may also be a vector. Assume a group G acts on both the sample space and the parameter space, although the action of G on the two spaces may be quite different.

Definition. The statistical model for X is *invariant under* G if for all $g \in G$, when X has distribution $P \cdot \mid \theta)$, then gX has distribution $P(\cdot \mid g\theta)$, for all $\theta \in \Theta$.

Example 5. Suppose the real valued random variable X is normal with mean μ and variance σ^2. With $\theta = (\mu, \sigma^2)$, X is $N(\mu, \sigma^2)$ when the parameter value is θ. In this example, the sample space $\mathscr{X}$ is R^1, while Θ is

$$\Theta = \{(\mu, \sigma^2) \mid \mu \in R^1, \sigma > 0\}$$

The group G is

$$G = \{(a, b) \mid a > 0, b \in R^1\}$$

and the binary group operation is

$$(a_1, b_1)(a_2, b_2) = (a_1a_2, a_1b_2 + b_1)$$

The action of G on $\mathscr{X}$ when $g = (a, b)$ is

$$x \to gx = ax + b$$

while the action of G on Θ is

$$\theta = (\mu, \sigma^2) \to g\theta = (a\mu + b, a^2\sigma^2) \tag{11}$$

Of course, when X is $N(\mu, \sigma^2)$, then $aX + b$ is $N(a\mu + b, a^2\sigma^2)$. This motivates the group action in (11) and shows that the normal model with arbitrary mean and variance is invariant under the action of G.

The above example is far too simple to be of much interest in inferential problems. A natural extension of Example 5 follows and is of direct relevance in both estimation and hypothesis testing—topics to be considered in sections to follow.

Example 6. The notation established in Example 5 is used here. Suppose $X_1, \ldots, X_n$ are independent and identically distributed as $N(\mu, \sigma^2)$. The random column vector $X = (X_1, \ldots, X_n)'$ takes values in R^n and the distribution of X is the n-dimensional normal distribution $N_n(\mu\mathbf{1}, \sigma^2 I_n)$ where I_n is the $n \times n$ identity matrix and $\mathbf{1}$ is the vector of ones in R^n. The parameter space for this model is Θ given in Example 5. Take G to be the group in Example 5, so G acts on Θ as in (11).

For $g = (a, b) \in G$, the transformation on R^n defined by this g is

$$x = (x_1, \ldots, x_n)' \to ax + b\mathbf{1}$$

The model for X is $\{P(\cdot\,|\,\theta)\,|\,\theta = (\mu, \sigma^2) \in \Theta, P(\cdot\,|\,\theta) = N_n(\mu\mathbf{1}, \sigma^2 I_n)\}$. This is an invariant model for X when G acts on R^n and Θ in the manner indicated.

6. *The Principle of Invariance*

When a statistical model $\{P(\cdot\,|\,\theta)\,|\,\theta \in \Theta\}$ on a sample space $\mathscr{X}$ is invariant under a group G, it is common statistical practice to search for good inferential procedures, such as estimation schemes or hypothesis tests, which are also invariant. The normative form of this practice leads to the principle of invariance: If a statistical model is invariant under a group, then invariant inference problems should have invariant solutions.

The argument in Example 1 which led to invariant estimators (Eqn. (6)) is a direct application of this principle applied to an estimation problem. Although a precise definition of what an invariant inference problem is would take us far afield, the estimation and testing examples of the following sections will well illustrate the principle. For a thorough and rigorous treatment of invariant inference problems, see Ferguson (1967), Eaton (1983, 1989), Berger (1985), Lehmann (1986), Lehmann and Casella (1998).

7. *Invariant Estimation*

Given an observable random vector X whose distribution belongs to an invariant statistical model $\{P(\cdot\,|\,\theta)\,|\,\theta \in \Theta\}$, the search for a good invariant estimation of the parameter θ is a common inferential problem. Formally an estimator is a function t defined on the sample space $\mathscr{X}$ and taking value in the parameter space Θ. The abstract version of the argument used by Pitman is the following:

> The model for X is invariant under the group G so if X has distribution $P(\cdot\,|\,\theta)$, then gX has distribution $P(\cdot\,|\,g\theta)$. If $X = x$ is observed then $t(x)$ is the estimated value of θ. But if $gX = gx$ had been seen rather than x, then $t(gx)$ should be the estimated value of $g\theta$ since gX has distribution $P(\cdot\,|\,g\theta)$. Therefore $g^{-1}t(gx)$ should estimate θ. Equating these two point estimates of θ leads to $t(x) = g^{-1}t(gx)$, or equivalently
>
> $$t(gx) = gt(x) \quad \text{for} \quad x \in \mathscr{X}, g \in G \tag{12}$$

Estimators which satisfy (12) are called *invariant* by some authors and *equivariant* by others. It should be emphasized that in (12), the expression gx means the transformation determined by g evaluated at $x \in \mathscr{X}$ where as $gt(x)$ means g evaluated at the point $t(x) \in \Theta$. Equation (12) involves the action of G on both $\mathscr{X}$ and Θ.

Example 7. The notation and assumptions of Example 6 hold here. The maximum likelihood estimators (MLE) for μ and σ^2 are

$$\hat{\mu}(x) = \frac{1}{n}\sum_1^n x_i = \bar{x}$$

$$\hat{\sigma}^2(x) = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$$

Then, written in vector form, the MLE for $\theta = (\mu, \sigma^2)$ is

$$t_0(x) = (\hat{\mu}(x), \hat{\sigma}^2(x)), \quad x \in R^n$$

Thus t_0 is an estimator defined on R^n taking values in Θ. This estimator satisfies equation (12) so that the MLE is invariant.

Example 7 is a special case of a general phenomenon—namely, when statistical models are invariant under a group, then the MLE is an invariant estimator. This is not to say that the MLE is necessarily a good (or the best) invariant estimator, but just that finding MLEs respects the principle of invariance. For example, the Pitman estimator (7) is usually not the MLE.

8. *Invariant Hypothesis Testing*

In multiparameter hypothesis testing problems, it is rather unusual to find a testing procedure with optimality properties when compared with all other tests. However, certain tests in multivariate statistical analyses (Muirhead (1982), Eaton (1983), Anderson (1984)) do have optimality properties within the class of invariant tests when the testing problem is invariant.

Again suppose an invariant model $\{P(\cdot\,|\,\theta)\,|\,\theta \in \Theta\}$, for $X \in \mathscr{X}$ is given where the group G acts on both $\mathscr{X}$ and Θ. Consider a hypothesis testing problem

$$H_0\colon \theta \in \Theta_0 \quad \text{vs.} \quad H_1\colon \theta \in \Theta - \Theta_0$$

such that H_0 is the null hypothesis and H_1 is the alternative. This testing problem is called *invariant* if the model under H_0, $\{P(\cdot\,|\,\theta)\,|\,\theta \in \Theta_0\}$ is also invariant under G. A testing procedure is invariant if it can be expressed in terms of a maximal invariant (see Sect. 4).

Example 8. Let $X_1, \ldots, X_n$ be a random sample from a $N_1(\mu, \sigma^2)$ population and consider the classical Student-t problem of testing $H_0\colon \mu = 0$ vs. $H_1\colon \mu \neq 0$. With $\bar{x} = n^{-1}\sum x_i$ and $s^2 = \sum(x_i - \bar{x})^2$, the Student-$t$ test rejects H_0 for large values of $(\bar{x}/s)^2$. This testing problem is invariant under the group $G = \{a\,|\,a \in R^1, a \neq 0\}$ where the group operation is multiplication. The action of G on the sample space is

$$x_i \to ax_i, \quad i = 1, \ldots, n$$

and the action of G on the parameter space is

$$\mu \to a\mu, \quad \sigma^2 \to a^2\sigma^2$$

The Student-t test is not uniformly most powerful (UMP) within the class of all tests but it is UMP within the class of invariant tests (Lehmann 1986). It is also true that the Student-t test is UMP within the class of unbiased tests—a fact which does not extend to the multivariate case. However, the invariance argument does extend as the following example shows.

Example 9. Let $X_1, \ldots, X_n$ be a random sample from a p-variate normal distribution $N_p(\mu, \Sigma)$. The mean vector μ is in R^p and the $p \times p$ covariance matrix Σ is assumed to be nonsingular. The multivariate analog of the Student-t problem is the hypothesis testing problem H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$. Set $\bar{X} = n^{-1}\Sigma X_i$ and $S = \Sigma(X_i - \bar{X})(X_i - \bar{X})'$. The likelihood ratio test (LRT) of H_0 vs. H_1 rejects H_0 for large values of Hotelling's T^2 statistic (Anderson 1984) given by up to constants)

$$T^2 = \bar{X}' S^{-1} \bar{X}$$

This is a natural multivariate extension of the Student-t statistic. The above hypothesis testing problem is invariant under the group $G = Glp$ of all $p \times p$ nonsingular matrices with matrix multiplication as the group composition. The action of G on the sample space is for $A \in G$

$$X_i \to AX_i, \quad i = 1, \ldots, n$$

and on the parameter space is

$$\mu \to A\mu, \quad \Sigma \to A\Sigma A'$$

Hotelling's T^2 test is invariant and within the class of invariant tests, it is UMP.

Example 9 contains elements that are common to a variety of invariant testing problems. First, the LRT is always an invariant test for any testing problem which is invariant. Therefore restricting attention to invariant tests never excludes the LRT. Second, within the class of invariant tests, there is one test that is optimal. This happenstance occurs in some, but certainly not all, invariant testing problems in multivariate analysis (see Eaton 1983).

9. *Concluding Remarks*

The examples given above reflect only the simplest invariance arguments. In fact invariance has found many applications outside the hypothesis testing and estimation contexts described here. For example, Lehmann (1986) illustrates such arguments in nonparametric problems while other applications can be found in Eaton (1983, 1989), Berger (1985), and Wijsman (1990).

A review of the statistical literature shows that many of the best known statistical procedures have attractive invariance properties. These properties not only appeal to our aesthetic sense of symmetry, but also provide powerful theoretical tools for the study and development of new inferential methods. Although the theory of invariance was well developed by the 1970s, it continues to be an active area of research in mathematical statistics. For example, some recent work in statistical prediction theory has been heavily influenced by invariance considerations (see Eaton and Sudderth 1999).

See also: Likelihood in Statistics; Multivariate Analysis: Overview; Probability: Formal; Probability: Interpretations; Statistics: The Field

Bibliography

Anderson T W 1984 *An Introduction to Multivariate Statistical Analysis*, 2nd edn. Wiley, New York

Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Springer-Verlag, New York

Eaton M L 1983 *Multivariate Statistics: A Vector Space Approach*. Wiley, New York

Eaton M L 1989 *Group Invariance Applications in Statistics*. The Institute of Mathematical Statistics and The American Statistical Association, Hayward, CA

Eaton M L, Sudderth W D 1999 Consistency and strong inconsistency of group-invariant predictive inferences. *Bernoulli* **5**: 833–54

Ferguson T S 1967 *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York

Fraser D A S 1968 *The Structure of Inference*. Wiley, New York

Hotelling H 1936 Relations between two sets of variates. *Biometrika* **28**: 321–77

Lehmann E L 1986 *Testing Statistical Hypotheses*, 2nd edn. Springer, New York

Lehmann E L, Casella G 1998 *Theory of Point Estimation*, 2nd edn. Springer, New York

MacLane S, Birkhoff G 1967 *Algebra*. Macmillan, New York

Muirhead R J 1982 *Aspects of Multivariate Statistical Theory*. Wiley, New York

Pitman E J G 1939 The estimation of location and scale parameters of a continuous population of any given form. *Biometrika* **30**: 391–421

Wijsman R A 1990 *Invariant Measures on Groups and their use in Statistics* (*Lecture Notes—Monograph Series, Vol. 14*). Institute of Mathematical Statistics, Hayward, CA

M. L. Eaton

Ion Channels and Molecular Events in Neuronal Activity

The most fundamental mechanism for imparting information in the nervous system is the opening of ion channels. This responsibility begins early in life, where opening of ion channels in progenitor cells

induces neuronal differentiation. Indeed, ion channels mold the growth and behavior of neurons throughout each phase of the life span. While there are many different classes of ion channels involved in regulating nervous system activity, Ca^{2+} entry serves as the internal ionic message of neural experience. Developmentally, calcium triggers differentiation, guides the growth of neuronal processes, and contributes to the selective elimination of neurons. Ca^{2+} mediates neurotransmitter release, the simplest form of interneuron communication, and its entry is the primary intracellular event for encoding the frequency and number action potentials generated pre- and postsynaptically. Activity- or use-dependent increases in Ca^{2+} serve as an intracellular message for initiating neuronal plasticity which may be involved in circuit modifications required for learning a memory. Finally, completing the cycle of neuronal ontogeny, Ca^{2+} in excess triggers a cascade of mechanisms leading to neuronal degeneration during senescence and age-related disease.

1. *Activity-dependent Increases in Neuronal Ca^{2+}*

Internal Ca^{2+} is maintained at extremely low concentrations creating an enormous driving force for its movement into the cytoplasm from internal stores or the extracellular space. Ca^{2+} levels are estimated to be at about 100 nM in neurons at their resting hyperpolarized state, while extracellular Ca^{2+} is in the 2–3 mM range and 10 μM to 1 mM inside the endoplasmic reticulum (Ghosh and Greenberg 1996, Mattson et al. 2000). The gating mechanisms for Ca^{2+} entry into the cell's cytoplasm includes plasma membrane voltage-dependent Ca^{2+} channels (VDCC), plasma membrane neurotransmitter or ligand-activated channels, and Ca^{2+} release from internal stores. All of these mechanisms are recruited by neuronal activity, they experience Ca^{2+}-mediated feedback regulation, and they are modulated by neurotransmitter activation of G-proteins (Lee et al. 1999, Zhang et al. 1998, Zamponi and Snutch 1998, Zuhlke et al. 1999). Each mechanism of Ca^{2+} entry is compartmentalized to produce localized increases in intracellular Ca^{2+} and site-specific activation of Ca^{2+}-sensitive pathways (Magee et al. 1998, Neher 1998, Yuste et al. 2000).

Adaptive thresholds for Ca^{2+} entry ensure its action as a second messenger will reflect high levels of neuronal activity. Voltage-gated and neurotransmitter gated channels interact during periods of neuronal activity to determine the degree of Ca^{2+} entry and thus the direction of Ca^{2+}-mediated adaptation. The classic sequence for initiating Ca^{2+} entry begins with presynaptic release of neurotransmitter acting postsynaptically to depolarize the cell beyond the threshold for triggering an action potential (Colquhoun and Sakmann 1998). The pioneering work of Hodgkin, Huxley, and Katz demonstrated action potentials were initiated by activating voltage-dependent Na^+ channels and rapidly terminated by the combined action of Na^+ channel inactivation and the opposing activation of voltage-dependent K channels (Hodgkin et al. 1952). Activation of VDCCs is facilitated by the rapid depolarization initiated by Na^+ channel activation, but it is not an absolute requirement. High- and low-threshold Ca^{2+}-mediated action potentials can be triggered by depolarization independent of Na^+ channel activation (Magee et al. 1998, Huguenard 1996). The duration and peak depolarization of the action potential, and consequently the extent of Ca^{2+} entry, is limited by activation of voltage-dependent K^+ channels and intrinsic kinetics of Ca^{2+} channel inactivation.

Unique patterns of Ca^{2+} can be achieved by the selective expression of ion channels in different populations of nerve cells. For example, the expression of inwardly rectifying K^+ (K_{ir}) channels raises the threshold for action potential generation by hyperpolarizing the resting membrane potential of striatal neurons, while the expression and interaction of low-threshold Ca^{2+} channels (T-type), Ca^{2+}-activated K^+ channels (slo), and hyperpolarization activated K^+ channels (I_H) contribute to regular, intrinsic firing in neurons of the substantia nigra and thalamus (Huguenard 1996, Santoro et al. 2000, Wilson and Callaway 2000). Regional differences in ion channel distribution can also create microenvironments of Ca^{2+} control within the cell. Clustering of Na^+, K^+, and Ca^{2+} channels in dendrites create local sites for integrating synaptic input (Magee et al. 1998, Safronov 1999), and direct association of P/Q and N-type Ca^{2+} channels with vesicle fusion proteins mediates activity-dependent release of neurotransmitter in synaptic terminals (Goda and Sudhof 1997, Neher 1998).

The molecular diversity of ion channels present in neurons creates opportunity for exquisite control of Ca^{2+} influx through differences in gating mechanisms and kinetics of activation and inactivation (Catterall 1995, Ghosh and Greenberg 1995). Calcium channels are heteroligomeric complexes containing α_1, $\alpha_2\delta$ and β subunits. Ca^{2+} channels contain a single α_1 subunit which serves as a pore and voltage sensor. Nine α_1 subunit genes have been identified, corresponding to each type of VDCC channel previously categorized by pharmacology. α_{1A} encodes for ω-Aga IVA sensitive P-type, α_{1B} encodes for ω-Ctx GVIA sensitive N-type, α_{1C}, α_{1D} and α_{1S} all encode for dihydropyridine sensitive L-type and α_{1G} and $_{\alpha 1H}$ encode for Ni^{2+}-sensitive T-type calcium channels (Zamponi and Snutch 1998). K^+ channels represent a much larger and more diverse collection of ion channels. At present, nine families of voltage-activated K^+ channels are described (Kaczorowski and Garcia 1999). Many are derived from homologues of *Drosophila* channel genes *Shaker*, *Shah*, *Shaw*, and *Shal*, referred to as K_V1, K_V2, K_V3, and K_V4 in mammals. They usually form homotetrameric channels, but association with β-subunits can alter channel kinetics (Kaczorowski and Garcia 1999). A separate class of K^+ channel funda-

mental in regulating neuronal excitability is Ca^{2+}-activated K^+ channels. They are classified by single channel conductances of 2–25 pS (small or SK channel), 25–100 pS (intermediate or IK channel), and 100–300 pS (large or BK channel) (Kaczorowski and Garcia 1999). Ca^{2+}-activated K^+ channels serve as an activity-dependent feedback mechanism for reducing neuron firing frequency and returning the membrane potential back to its resting state.

Excitatory synaptic input increases intracellular Ca^{2+} directly, by activating ionotropic and metabotropic glutamate receptors, and secondarily by driving the membrane potential beyond the voltage-dependent threshold for activating VDCCs. The cooperation of these mechanisms ensures robust Ca^{2+} signals in response to strong synaptic drive. Ionotropic glutamate receptors (AMPA, kainate, and NMDA receptors) are the postsynaptic mediators of fast excitatory synaptic transmission. Functional ionotropic glutamate receptors contain four or five structurally related subunits. The combination of subunits GluR1–4 make functional AMPA receptors, GluR5–7 and KA1–2 make kainate receptors, and NR1 and NR2A-D make NMDA receptors (Bigge 1999). Flip or flop splice variants of AMPA receptor subunits influence gating kinetics. AMPA receptors generally have low Ca^{2+} permeability, due to the Q/R site of GluR2 subunits. GluR2 is the most common subunit found in the brain, but examples of receptors without GluR2 exist and are implicated in Ca^{2+}-dependent changes in neuronal physiology (Conti and Weinberg 1999, Bigge 1999). NMDA receptors have greater glutamate affinity than AMPA receptors, slower activation and deactivation kinetics, and they are highly permeable to Ca^{2+}. The Ca^{2+}conductance/Na^+ conductance is > 5 for NMDA receptors, 2 for AMPA receptors lacking GluR2, and 0.1 for AMPA receptors containing GluR2 (Conti and Weinberg 1999). The combined action of greater Ca^{2+} permeability and slower channel inactivation results in dramatically more Ca^{2+} influx for NMDA receptors. NMDA receptors are made of NR1 and NR2 heterodimer complexes and generally composed of 4–5 subunits. NR1 is essential for channel function and NR2 confers the voltage-dependent Mg^{2+} block and glycine modulation of NMDA receptors (Conti and Weinberg 1999). The voltage-dependent Mg^{2+}-block acts as an activity sensor, allowing large increases in intracellular Ca^{2+} only when the depolarization threshold for releasing the Mg^{2+} block is exceeded. Kainate receptors, like AMPA receptors, lack voltage dependency. Compared to AMPA receptors, kainate receptors contribute little to the peak postsynaptic response, but their slower decay kinetics results in comparable charge transfer (Ferking and Nicoll 2000). Kainate receptors also mediate presynaptic inhibition of excitatory and inhibitory synapses in the hippocampus.

Slow excitatory synaptic events mediated by activation of metabotropic glutamate receptors (mGluRs) is another important mechanism for elevating intracellular Ca^{2+}. Group-I mGluRs stimulate phospholipase C via GTP-binding proteins. Phospholipase C cleaves phosphoinositol-(4,5)-bisphosphate (PIP_2) into diacylglycerol (DAG) and inositol 1,4,5-triphosphate (IP_3). DAG activates protein kinase C (PKC) and $I(1,4,5)P_3$ activates IP_3 receptors located on the surface of the smooth endoplasmic reticulum (Fagni et al. 2000, Mattson et al. 2000). IP_3 receptors trigger the release of Ca^{2+} stored in the endoplasmic reticulum at concentrations of 10–100 μM (Mattson et al. 2000). Other neurotransmitters use the same mechanism for releasing Ca^{2+} from intracellular stores, but glutamate is by far the most ubiquitous.

2. *Activity-dependent Control of Neuronal Development*

Ca^{2+} influx stimulates differentiation and directs the course of development in excitable cells (Spitzer et al. 1994). VDCCs, ionotropic glutamate receptors, and mGluRs all participate in regulating intracellular Ca^{2+} at key steps in neuronal development. Glutamate activation of NMDA and AMPA receptors induces proliferation of progenitor cells during neurogenesis and their migration to appropriate cell layers (Haydar et al. 2000). Evidence also exists for neurogenesis mediated by Ca^{2+} entry through voltage-dependent L-type Ca^{2+} channels and Ca^{2+} waves mediated by release of Ca^{2+} from intracellular stores. In postmitotic cells, Ca^{2+} continues to direct neuronal differentiation by influencing ion channel expression, neurotransmitter phenotype, growth factor expression, and neurite extension (Spitzer et al. 1994). Ca^{2+} influx through VDCCs (L- and N-type) stimulate cell adhesion molecules (CAMs) (NCAM, cadherin, and L1) to promote neurite outgrowth via CAM-mediated cell to cell interactions and the binding of CAM to actin-based cytoskeleton (Doherty and Walsh 1994). Ca^{2+} may initiate neurite extension by binding to calmodulin (CaM) and phosphorylating microtubule associate proteins (MAP2) through the activation of Ca^{2+}/CaM-dependent protein kinase II and IV (Kuhn et al. 1998). A parallel mechanism implicated in MAP phosphorylation is the mitogen activated protein kinase (MAPKs) pathway stimulated by mGluR activation (Quinlin and Halpain 1996). It is clear that Ca^{2+}-mediated remodeling is based upon a Ca^{2+} set point, since differences in responses exist depending on the Ca^{2+} level achieved and the neuronal population examined. Some neurons experience growth cone collapse with increases in Ca^{2+}, due to severing of microfilaments by Ca^{2+}-activated gelsolin (Neely and Nichols 1995). Support for this mechanism comes from the observation of downregulated VDCCs coincident with neurite extension and increases in neuronal activity producing decreases in CAM expression.

Protein synthesis is also critical to activity-dependent plasticity and growth, a process Ca^{2+} mediates through interacting steps of protein phosphorylation (Ghosh and Greenberg 1995). Activity-dependent increases in Ca^{2+} stimulate the expression of immediate early genes like c-fos and c-jun, as well as delayed 'effector' genes for neurotrophins and, in a mechanism of positive feedback, genes for VDCCs, NMDA receptors, and AMPA receptors (Ghosh and Greenberg 1995). The pattern of gene expression is dependent on the pattern of neuronal activity and Ca^{2+} accumulation. Intracellular pathways appear to act in parallel to alter gene expression, but they may respond differentially to changes in Ca^{2+} concentration. A final point of convergence for Ca^{2+}-stimulated pathways is phosphorylation of cAMP-responsive element binding protein (CREB), which activates RNA polymerase by forming a complex with CREB-binding protein (CRP). The independent pathways proposed to trigger this mechanism include Ca^{2+} activation of Ras/mitogen activated protein kinase (MAPK) pathway, Ca^{2+} activation of adenlyate cyclase and subsequently protein kinase A (PKA) and Ca^{2+} binding to CaM, and activation of Ca^{2+}/CaM kinase II and IV (Ghosh and Greenberg 1995).

A clear division does not exist between the actions of Ca^{2+} and neurotrophins in driving activity-dependent change in neurons. Activity not only increases intracellular Ca^{2+}, but it also causes the release of neurotrophins, and neurotrophins and Ca^{2+} produce similar responses in nerve cells (McAllister et al. 1999). Ca^{2+} and neurotrophins both trigger the Ras-MAPK pathway to change gene expression, neurotrophins by binding to membrane-bound tyrosine kinase (*trk*) receptors, and Ca^{2+} by activating Ras directly (Heumann 1994). Neurotrophins have both direct-rapid and slow modulatory effects on ion channels involved in excitation (Benedikt and Poo 1996). Neurotrophins cause rapid depolarizations by activating Na^{+} channels linked to *trk* receptors and they modulate glutamate-activated channels and intrinsic voltage-activated channels to increase intracellular Ca^{2+} (Jarvis et al. 1997, Kaflitz et al. 1999). Neurotrophins also increase neuronal excitability and thus Ca^{2+} influx through acute inhibition of $K_{ir}3.1$, $K_{ir}3.4$, and $K_{ir}2.1$ (Wischmeyer et al. 1998). Neurotrophins also act presynaptically to increase neurotransmitter release (McAllister et al. 1999). Testament to the intimate relationship between neurotrophins and Ca^{2+}, neurotrophins produce long-term changes in Ca^{2+} entry by increasing the expression of VDCCs, NMDA receptors, and AMPA receptors, and activity-dependent increases in Ca^{2+} produce a reciprocal increase in neurotrophin expression (McAllister et al. 1999).

The gearing of neurons toward excitation during development is clearly illustrated by the unique behavior displayed by neurotransmitter activated channels early in development. Embryonic synaptic AMPA receptors are highly Ca^{2+} permeable and mEPSC (minature excitatory postsynaptic current) amplitude can be more than twofold larger than at mature synapses (Rohrbough and Spitzer 1999). Postnatal neurons preferentially express flip variants of AMPA receptors, which increase the depolarization-mediated Ca^{2+} signal by slowing desensitization (Conti and Wienberg 1999). It should be noted, however, that many synapses show only functional NMDA receptors and only with activity to 'silent' AMPA receptors appear (Petralia et al. 1999). Early postnatal development also shows a shift in the molecular structure and physiological properties of NMDA receptors. NR2B subunits are preferentially expressed early in development and they show less voltage-dependent block by Mg^{2+} and slower kinetics, properties which lower the threshold for Ca^{2+} entry during synaptic excitation. With maturity, there is a shift to NR2A and NR2C subunits which increase the Mg^{2+} block and speed up the NMDA activated response. This postnatal disposition toward synaptic excitation is augmented by a developmental shift in the property of $GABA_A$ receptor mediated inhibition (Ben-Ari et al. 1997). Rather than inhibiting action potential generation and VDCC activation, as is seen in adults, immature $GABA_A$ synapses cause slow depolarizations triggering bursts of action potentials and Ca^{2+} influx through VDCCs (Ben-Ari et al. 1997). Later in development, the Cl^{-} gradient changes resulting in hyperpolarizing $GABA_A$ responses and an inhibition of action potential generation and Ca^{2+} channel activation (Ben-Ari et al. 1997).

3. *Activity and Synaptic Function*

Ca^{2+} entering presynaptic terminals during action potentials stimulates the fusion of synaptic vesicles with the presynaptic membrane and the release of neurotransmitter into the synaptic cleft. P/Q and N-type Ca^{2+} channels are the primary VDCCs subserving this function in the central nervous system (Goda and Sudhof 1997). Entrance of Ca^{2+} through individual channels is proposed to form microdomains, where Ca^{2+} levels may reach as high as 100 μM (Neher 1998). It is proposed that 'core complex' proteins of the synaptic fusion machinery maximize their proximity to these domains by binding syntaxin to Ca^{2+} channels (Goda and Sudhof 1997). Prior to and independent of Ca^{2+} influx, synaptic vesicles dock to the presynaptic membrane in active zones and undergo priming to make them ready for releasing neurotransmitter with Ca^{2+} influx (Goda and Sudhof 1997). The docking and priming steps are mediated by an interaction of a 'core complex' of NSF attachment protein receptors (SNAREs) located on the vesicle (v-SNARE) and attachment SNAREs located on the target membrane (t-SNARE). The core complex or SNARE is made up of syntaxin and synaptosome-associated membrane

protein of 25 kDa (SNAP-25), which are localized in the synaptic plasma membrane, and the vesicular protein synaptobrevin (also known as vesicle associated membrane protein; VAMP) (Goda and Sudhof 1997). The process of vesicle fusion and neurotransmitter release is believed to be mediated by the interaction of the synaptotagmin Ca^{2+} sensor with the 'core complex.' Synaptotagmin I and II are the leading candidates as Ca^{2+} sensors, although other presynaptic proteins with Ca^{2+} C_2-binding domains are proposed to serve the same function (Goda and Sudhof 1997). The mechanism for fusion is proposed to be a rapid electrostatic or conformational change in the fusion complex mediated by Ca^{2+} binding to synaptotagmin (Goda and Sudhof 1997). The vesicle collapses into the membrane and is retrieved away from the active zone by a clathrin-dependent mechanism involving the GTPase dynamin I (Cremona and De Camilli 1997).

Invasion of the presynaptic terminal by multiple action potentials results in a short-term change in the probability of vesicle fusion and the release of neurotransmitter. Depending on the number and timing of the action potentials, and thus the degree and duration of Ca^{2+} increase, the processes of facilitation (time constant less than one second), augmentation (time constant of several seconds), and potentiation (time constant of a few minutes) are initiated (Zucker 1999). One mechanism proposed for mediating short-term synaptic enhancement is the binding of Ca^{2+} to synapsin I. Synapsin I is a Ca^{2+}-regulated ATPase which binds to actin and may facilitate vesicle transport via the actin matrix (Zucker 1999). Other mechanisms proposed to participate in short-term synaptic plasticity include facilitating actions of synaptotagmin and Ca^{2+}-dependent facilitation of presynaptic Ca^{2+} currents (Zucker 1999). An exciting new explanation for the longest form of short-term plasticity, PTP, is based on Ca^{2+} dumping by mitochondria following mitochondrial sequestration of Ca^{2+} during spike trains (Zucker 1999). New studies indicate another key role for Ca^{2+} in mediating short-term synaptic plasticity lies in a clathrin-independent form of rapid endocytosis of synaptic vesicles. The proposed mechanism suggests that Ca^{2+} binds to synaptophysin during repeated excitation and Ca^{2+}-stimulated synaptophysin binds to dynamin I to produce rapid endocytosis (Daley et al. 2000).

Large and sustained increases in intracellular Ca^{2+} on the postsynaptic side of the synapse is the primary event for inducing long-term forms of synaptic plasticity or long-term depression (LTD) and long-term potentiation (LTP). Critical to determining whether LTP or LTD occurs are the extent and kinetics of the Ca^{2+} signal and the molecular mechanism underlying the Ca^{2+} increase (Zucker 1999). These factors and other mechanisms vary considerably between brain areas. Based on cortical and cerebellar studies, a few general themes emerge. Ca^{2+} entry through NMDA receptors is a critical event in the induction of LTP and the Mg^{2+} block of these receptors sets a stimulus and depolarization threshold. Weaker stimuli, which activate VDCCs, often lead to the induction of LTD (Zucker 1999). LTD also requires the activation of group I mGluRs at many synapses and the subsequent activation of nitric oxide synthetase (Daniel et al. 1998). VDCCs can augment the NMDA Ca^{2+} signal if back-propagating action potentials precede activation of NMDA receptors (Zucker 1999). This receptor based scheme depends on the parameters of the Ca^{2+} signal, and thus examples of NMDA receptor-dependent LTD and mGluR-dependent LTP also exist (Nicoll et al. 1998).

The mechanisms used to induce synaptic plasticity in adult animals often follow the same pathways utilized for development and growth in the nervous system. Phosphorylation plays a critical role in LTP and it is triggered by Ca^{2+} influx (Soderling and Derkach 2000). Ca^{2+} activation of Ca^{2+}/CaM kinase and Ca^{2+} activation of the Ras/MAPK pathways play distinct roles in the maintenance phase of LTP. Autophosphorylation of Ca^{2+}/CaM kinase makes it Ca^{2+} independent and creates a mechanism for maintaining long-term phosphorylation. Active Ca^{2+}/CaM kinase increases AMPA receptor responses and may be involved in the unmasking of silent receptors and the trafficking of new receptors to the extracellular membrane (Malinow et al. 2000, Soderling and Derkach 2000). Keeping Ca^{2+}/CaM kinase phosphorylated and active is opposed by the activity of protein phosphatase I. Herein lies another critical step in LTP: inhibiting protein phosphates. Phosphatase inhibition is mediated by Ca^{2+} activation of adenlyate cyclase, activation of PKA, and phosphorylation of inhibitor 1 (I1). I1 shuts off protein phosphatase I and Ca^{2+}/CaM kinase thus stays phosphorylated and active (Soderling and Derkach 2000). Another critical component of LTP expression is protein production. Ca^{2+}/CaM kinase is also implicated in gene expression as well as Ca^{2+}-activation of Ras/MAPK and, just like observed developmentally, phosphorylation of CREB (Platenik et al. 2000). Another similarity to development is the clear involvement of neurotrophins in LTP expression. Neurotrophins directly modulate Ca^{2+} physiology and they activate the Ras/MAPK pathway (McAllister et al. 1999, Klintsova and Greenough 1999).

4. *Ca^{2+} in Aging and Degeneration*

The role of Ca^{2+} in development and plasticity requires flawless control over its intracellular concentration. Aging in the brain, and in many other tissues, occurs in part through a disruption in the ability to maintain Ca^{2+} homeostasis (Thibault et al. 1998). A system level theory for explaining aging poses that chronic exposure to glucocorticoid stress hormones creates a disruption of Ca^{2+} homeostasis (Thibault et

al. 1998). In the hippocampus, stress hormones and/or aging increases the expression of L-type Ca^{2+} channels (Thibault et al. 1998). The increased Ca^{2+} signal suppresses hippocampal neuron activity by stimulating Ca^{2+}-activated K^+ channels and, consequently, the hippocampus becomes less effective in its role of terminating the stress response. Circulating levels of stress hormones therefore progressively increase over an animal's lifetime and consequently the deleterious effects of Ca^{2+} overload also become progressively worse. Elevated Ca^{2+} in aged neurons is proposed to create dysfunctional plasticity and to reduce the threshold for Ca^{2+}-mediated cell death.

Disruption of Ca^{2+} homeostasis is also proposed to be a critical event in apoptosis and necrosis produced in many age-related neurodegenerative disorders (Mattson 1998, Mattson et al. 2000, Thibault et al. 1998). Periods of excessive excitation, as might occur during stroke or in response to β-amyloid stimulation, cause abnormally high levels of Ca^{2+} ions to enter the cell. The candidate mechanisms for the pathological increase in Ca^{2+} influx include VDCCs, NMDA receptor gated channels, and mGluR activation and downstream release of Ca^{2+} from internal stores. The consequent Ca^{2+} overload stimulates Ca^{2+}-sensitive proteases and it plays a significant role in the generation of reactive oxygen species (ROS) (Mattson et al. 2000). High levels of cytoplasmic Ca^{2+} cause mitochondrial Ca^{2+} loading which stimulates mitochondrial ROS production and release of apoptotic factors (Mattson 1998).

The flood of Ca^{2+} triggers abnormal activation of calpain and free radicals activate caspases, both of which are viewed to create excessive proteolysis. In addition, Ca^{2+} sequestration by mitochondria during excessive excitation reduces the production of ATP with a critical cost of shutting down ATPases responsible for maintaining the cell's membrane potential (Albin and Greenamyre 1992). Plasma membrane Na^+/K^+-ATPases require huge amounts of energy and their reduced activity leads to cellular depolarization, activation of VDCCs, and the removal of Mg^{2+} voltage-dependent block from NMDA gated channels.

Ca^{2+} also triggers cytoplasmic production of ROS through the activation of Ca^{2+}-dependent enzymes. Ca^{2+}-activated proteases convert xanthine dehydrogenase into xanthine oxidase which produces superoxide anions and H_2O_2. High levels of cytoplasmic Ca^{2+} triggered by NMDA receptor activation also activate nitric oxide synthetase (NOS). NO produced by the activation of NOS is a free radical, as well as a ubiquitous mediator of biological processes (Dawson and Dawson 1998). NO turns into a much more potent oxidant, however, after it reacts with the superoxide anion under conditions of oxidative stress to produce highly reactive peroxynitrite. The ROS generated by these mechanisms cause lipid, protein, and nucleic acid oxidation and ultimately cell death.

See also: Long-term Potentiation (Hippocampus); Neural Plasticity; Neurons and Dendrites: Integration of Information; Neurotransmitters; Psychophysiology; Synapse Ultrastructure; Synaptic Efficacy, Regulation of; Synaptic Transmission

Bibliography

Albin R L, Greenamyre J T 1992 Alternative excitotoxic hypotheses. *Neurology* **42**: 733–8

Ben-Ari Y, Khazipov R, Leinekugel X, Caillard O, Gaiarsa J L 1997 $GABA_A$, NMDA and AMPA receptors: A developmentally regulated 'menage a trois'. *Trends in Neurosciences* **20**(11): 523–9

Benedikt B, Poo M-m 1996 Fast actions of neurotrophic factors. *Current Opinion in Neurobiology* **6**: 324–30

Bigge C F 1999 Ionotropic glutamate receptors. *Current Opinion in Chemical Biology* **3**: 441–7

Catterall W A 1995 Structure and function of voltage-gated ion channels. *Annual Reviews in Biochemistry* **64**: 493–531

Colquhoun D, Sakmann B 1998 From muscle endplate to brain synapses: A short history of synapses and agonist-activated ion channels. *Neuron* **20**: 381–7

Conti F, Weinberg R J 1999 Shaping excitation at glutamatergic synapses. *Trends in Neurosciences* **22**: 451–8

Cremona O, De Camilli P D 1997 Synaptic vesicle endocytosis. *Current Opinion in Neurobiology* **7**: 323–30

Daniel H, Levenes C, Crepel F 1998 Cellular mechanisms of cerebellar LTD. *Trends in Neurosciences* **21**(9): 401–7

Daley C, Sugimori M, Moreira J E, Ziff E B, Llinas R 2000 Synaptophysin regulates clathrin-independent endocytosis of synaptic vesicles. *PNAS* **97**: 6120–5

Dawson V L, Dawson T M 1998 Nitric oxide in neurodegeneration. *Progress in Brain Research* **118**: 215–29

Doherty P, Walsh F S 1994 Signal transduction events underlying neurite outgrowth stimulated by cell adhesion molecules. *Current Opinion in Neurobiology* **4**: 49–55

Fagni L, Chavis P, Ango F, Bockaert J 2000 Complex interactions between mGluRs, intracellular Ca^{2+} stores and ion channels in neurons. *TINS* **23**: 80–8

Ferking M, Nicoll R A 2000 Synaptic kainate receptors. *Current Opinion in Neurobiology* **10**: 342–51

Ghosh A, Greenberg M E 1995 Calcium signaling in neurons: Molecular mechanisms and cellular consequences. *Science* **268**: 239–46

Goda Y, Sudhof T C 1997 Calcium regulation of neurotransmitter release: Reliably unreliable? *Current Opinion in Cell Biology* **9**: 513–18

Haydar T F, Wang R, Schwartz M L, Rakic P 2000 Differential modulation of proliferation in the neocortical ventricular and subventricular zones. *Journal of Neuroscience* **20**: 5764–74

Heumann R 1994 Neurotrophin signalling. *Current Opinion in Neurobiology* **4**: 668–79

Hodgkin A L, Huxley A F, Katz B 1952 Measurements of current–voltage relations in the membrane of the giant axon of Loligo. *Journal of Physiology (London)* **116**: 424–48

Huguenard J R 1996 Low-threshold calcium currents in central nervous system neurons. *Annual Reviews in Physiology* **58**: 329–48

Jarvis C R, Xiong Z G, Plant J R, Churchill D, Lu W Y, MacVicar B A, MacDonald J F 1997 Neurotrophin modulation of NMDA receptors in cultured murine and isolated rat neurons. *Journal of Neurophysiology* **78**(5): 2363–71

Kaczorowski G J, Garcia M L 1999 Pharmacology of voltage-gated and calcium-activated potassium channels. *Current Opinion in Chemical Biology* **3**: 448–58
Kaflitz K W, Rose C R, Thoenen H, Konnerth A 1999 Neurotrophin-evoked rapid excitation through TrkB receptors. *Nature* **401**: 918–21
Klintsova A Y, Greenough W T 1999 Synpatic plasticity in cortical systems. *Current Opinion in Neurobiology* **9**: 203–8
Kuhn T B, Williams C V, Dou P, Kater S B 1998 Laminin directs growth cone navigation via two temporally and functionally distinct calcium signals. *Journal of Neuroscience* **18**(1): 184–94
Lee A, Wong S T, Gallagher D, Li B, Storm D R, Scheuer T, Catterall W A 1999 Ca^{2+}/calmodulin binds to and modulates P/Q calcium channels. *Nature* **399**: 155–9
Magee J, Hoffman D, Colbert C, Johnston D 1998 Electrical and calcium signaling in dendrites of hippocampal pyramidal neurons. *Annual Review of Physiology* **60**: 327–46
Malinow R, Mainen Z F, Hayashi Y 2000 LTP mechanisms: From silence to four-lane traffic. *Current Opinion in Neurobiology* **10**: 352–7
Mattson M P 1998 Free radicals, calcium, and the synaptic plasticity-cell death continuum: Emerging roles of the transcription factor NF- B. *International Review of Neurobiology* **42**: 103–68
Mattson M P, LaFerla F M, Chan S L, Leissring M A, Shepel P N, Geiger J D 2000 Calcium signaling in the ER: Its role in neuronal plasticity and neurodegenerative disorders. *Trends in Neurosciences* **23**: 222–9
McAllister A K, Katz L C, Lo D C 1999 Neurotrophins and synaptic plasticity. *Annual Review of Neuroscience* **22**: 295–318
Neely M D, Nichols J G 1995 Electrical activity, growth cone motility and the cytoskeleton. *Journal of Experimental Biology* **198**: 1433–46
Neher E 1998 Vesicle pools and Ca^{2+} microdomains: New tools for understanding their roles in neurotransmitter release. *Neuron* **20**: 389–99
Nicoll R A, Oliet S H, Malenka R C 1998 NMDA receptor-dependent and metabotropic glutamate receptor-dependent forms of long-term depression coexist in CA1 hippocampal pyramidal cells. *Neurobiology of Learning & Memory* **70**(1–2): 62–72
Quinlan E M, Halpain S 1996 Emergence of activity-dependent, bidirectional control of microtubule-associated protein MAP2 phosphorylation during postnatal development. *Journal of Neuroscience* **16**(23): 7627–37
Petralia R S, Esteban J A, Wang Y X, Partridge J G, Zhao H M, Wenthold R J, Malinow R 1999 Selective acquisition of AMPA receptors over postnatal development suggests a molecular basis for silent synapses. *Nature Neuroscience* **2**(1): 31–6
Platenik J, Kuramoto N, Yoneda Y 2000 Molecular mechanisms associated with long-term consolidation of the NMDA signals. *Life Sciences* **67**(4): 335–64
Rohrbough J, Spitzer N C 1999 Ca(2+)-permeable AMPA receptors and spontaneous presynaptic transmitter release at developing excitatory spinal synapses. *Journal of Neuroscience* **19**(19): 8528–41
Safronov B V 1999 Spatial distribution of Na^+ and K^+ channels in spinal dorsal horn neurones: Role of the soma, axon and dendrites in spike generation. *Progress in Neurobiology* **59**(3): 217–41
Santoro B, Chen S, Luthi A, Pavlidis P, Shumyatsky G P, Tibbs G R, Siegelbaum S A 2000 Molecular and functional heterogeneity of hyperpolarization-activated pacemaker channels in the mouse CNS. *Journal of Neuroscience* **20**(14): 5264–75
Soderling T R, Derkach V A 2000 Postsynaptic protein phosphorylation and LTP. *Trends in Neurosciences* **23**: 75–80
Spitzer N C, Gu X, Olson E 1994 Action potentials, calcium transients and the control of differentiation of excitable cells. *Current Opinion in Neurobiology* **4**: 70–7
Thibault O, Porter N M, Chen K C, Blalock E M, Kaminker P G, Clodfelter G V, Brewer L D, Landfield P W 1998 Calcium dysregulation in neuronal aging and Alzheimer's disease: History and new directions. *Cell Calcium* **24**(5–6): 417–33
Wilson C J, Callaway J C 2000 Coupled oscillator model of the dopaminergic neuron of the substantia nigra. *Journal of Neurophysiology* **83**(5): 3084–100
Wischmeyer E, Doring F, Karschin A 1998 Acute suppression of inwardly rectifying Kir2.1 channels by direct tyrosine kinase phosphorylation. *Journal of Biological Chemistry* **273**(51): 34063–8
Yuste R, Majewska A, Holthoff K 2000 From form to function: calcium compartmentalization in dendritic spines. *Nature Neuroscience* **3**(7): 653–9
Zamponi G W, Snutch T P 1998 Modulation of voltage-dependent calcium channels by G proteins. *Current Opinion in Neurobiology* **8**: 351–6
Zhang S, Ehlers M D, Bernhardt J P, Su C-T, Huganir R L 1998 Calmodulin mediates calcium-dependent inactivation of *N*-methyl-D-aspartate receptors. *Neuron* **21**: 443–53
Zucker R S 1999 Calcium- and activity-dependent synaptic plasticity. *Current Opinion in Neurobiology* **9**: 305–13
Zuhlke R D, Pitt G S, Deisseroth K, Tsien R W, Reuter H 1999 Calmodulin supports both inactivation and facilitation of L-type calcium channels. *Nature* **399**: 159–62

J. P. Walsh

IPAT (Impact, Population, Affluence, and Technology)

For a given population group, the impact of human activity on the natural environment can be decomposed into the per capita impact × population size. Per capita impact, in turn, can be decomposed into the per capita level of consumption or 'affluence' × the environmental impact per unit of consumption under the existing technology. These statements can be combined as the identity $I = PAT$ (Impact = Population × Affluence × Technology). In this form the equation can be attributed to Ehrlich and Holdren (1971, 1972), and is sometimes called the Ehrlich–Holdren equation. IPAT refers to the equation and to the broader view of population–environment relations that it connotes.

'Impact' here indicates a persisting rather than transitory environmental effect. In Commoner's (1971) definition, it is an external intrusion into an ecosystem which tends to degrade the system's natural capacity for self-adjustment. For population–

environment interactions the impacts of most concern are those that reduce the system's capacity to provide environmental services. An example of an environmental impact is a country's carbon dioxide emissions, which degrade the environmental service provided by the atmosphere in regulating heat radiation from the earth's surface. The decomposition in that case would be population × per capita GDP × the 'carbon intensity' of the economy. At a given level of affluence and carbon intensity, emissions rise in proportion to population.

1. *IPAT as a Causal Relationship*

The IPAT equation is more widely employed not as an identity but as a causal relationship, intended to point to the environmental damage that results from one or more of population size, a high-consumption lifestyle and environmentally destructive technology, each amplifying the others. Implicitly, it asserts that these factors can together be seen as the main human causes of degradation. That assertion is debatable: in particular, social organizational and behavioral factors would often warrant separate scrutiny as causes of degradation rather than being subsumed within A and T.

This causal use is mostly confined to elementary or rhetorical discourse about the environment. A familiar instance is in contrasting the environmental impacts of population growth in rich and poor regions. If per capita energy use is a proxy for $A \times T$, the environmental impact of an additional person in America or Europe consuming at the average level for those societies would be an order of magnitude greater than that of an average person in Asia or Africa—a disparity cited in many popular writings that seek to assign responsibility for degradation or resource depletion (see *Consumption, Environmentally Significant*). The relationship is also used simply to organize a discussion, sometimes expressed in a more general functional form: $I = f(PAT)$.

If P, A, and T were independent of each other, the multiplicative relationship would be equivalent to an additive relationship among growth rates. In the case of carbon dioxide, the growth rate of emissions would equal the sum of the growth rates of the three components. However, P, A and T are not independent of each other. For any defined population and environment, they are variables in a complex economic, demographic and sociocultural system. Each also has major distributional dimensions and is a function of time. Consumption—or any other measure of human welfare—is an output of this system; environmental effects, both intended and unintended, are outputs as well. And even at the global level the system is not autonomous: it is influenced by 'natural' changes in the environment and by environmental feedbacks from human activity.

One kind of dependency among P, A and T follows from the common situation of diminishing returns: successive increments to population may do more than proportional environmental damage if they require recourse to ever more powerful technologies to maintain consumption levels. The IPAT equation might then be seen as illustrating a best-case scenario. But other kinds of dependency would tend to lessen the impact. For example, as affluence increases, so typically does a society's demand for a clean environment and wilderness preservation—that is, there is a high income elasticity of demand for environmental quality. This would support a policy response emphasizing conservation or restoration, and the development and adoption of 'clean' technologies (see *Greening of Technology and Ecotechnology*). Similarly, population growth rates generally fall, even to negative values, with rising incomes. If such interactions are acknowledged, the PAT decomposition retains heuristic and pedagogical value but does not bear much empirical scrutiny. In particular, it cannot resolve disputes on the relative contributions of factors responsible for environmental degradation.

For this latter task, Preston (1994) has proposed looking at the variances of the growth rates of I, P, A and T over different regions or countries. Writing these as $\sigma^2{}_I$, etc., the additive relationship among growth rates implies the following relationship among variances and covariances:

$$\sigma^2{}_I = \sigma^2{}_P + \sigma^2{}_A + \sigma^2{}_T + 2\ \mathrm{cov}_{PA} + 2\ \mathrm{cov}_{PT} + 2\ \mathrm{cov}_{AT} \tag{1}$$

The covariance terms are the interaction effects. If each is relatively small in a given case, there is a simple decomposition of the impact variance into the variance imputed to each factor. Otherwise, the one or more significant interaction terms can be explicitly noted. In Preston's analysis of carbon emission data for major world regions between 1980 and 1990, population growth makes a minor contribution to the total variance; the major contributors are the growth of A and T, with a substantial offsetting effect from the interaction of A and T.

2. *Environmental Spillovers*

With a strong management regime or effective community norms and sanctions, population growth in a region need not adversely affect the local environment. Access to a resource can be rationed, behavior strictly governed; or the growth can be diverted elsewhere, usually to urban centers. Ultimately, limits on the environmental services that can be drawn on would be reached and thereafter degradation would probably ensue. Well short of that limit, however, conditions of rapid economic or political change can undermine a management regime or erode norms and sanctions,

with adverse environmental results. Excessive deforestation can often be traced to such institutional breakdowns (or to ill-considered efforts at institutional reform) rather than to population growth. In other cases, a resource may have been so abundant that no management or sanctions were needed: this is a setting where the familiar 'tragedy of the commons' may unfold as the number of claimants to the resource or their exploitative abilities increase (see Hardin 1968).

Environmental effects of human activities may extend well beyond the region of origin, as with the case of acid rain or watershed destruction. In addition to such literal spillovers, trade and other economic relationships among societies may allow adverse effects to be 'exported' to another region: a poorer society may be more willing to incur environmental damage in return for economic gain, or be less able to prevent it. The concept of a community's 'ecological footprint' offers one means of accounting for such displaced effects in terms of the total area required to sustain the community's population and level of consumption (see Wackernagel and Rees 1996; see also *Human–Environment Relationship: Carrying Capacity*). Where the environmental effects involve degradation of a global commons, as with carbon emissions into the atmosphere, an effective response requires devising an international regulatory regime that takes account of economic level, demographic scale and technological capacity.

At the global level, a calculation of human environmental impact, in the spirit of IPAT, is that of the extent of human domination of various components of the global ecosystem. Examples would include the proportion of the land surface transformed by human activity, the anthropogenic share of atmospheric carbon dioxide, and the proportion of marine fisheries that are fully exploited or depleted (Vitousek et al. 1997). A summary impact measure proposed by Vitousek et al. (1986) is the proportion of total organic material produced through photosynthesis that is used directly by humans or used in human-dominated ecosystems, which they calculated at the time to be around 30 percent of the terrestrial total and 2 percent of the marine total.

3. *Nonlinearities in Environmental Systems*

An evident difficulty for impact assessment is the nonlinearity of environmental systems. Ecosystems may have multiple equilibria; they may exhibit cyclical or chaotic behavior. Small human populations and primitive technology do not ensure small environmental impact: in prehistoric times minuscule populations of hunter–gatherers may have been responsible for some mammalian extinctions and possibly for large-scale ecological change following the introduction of new fire regimes (see Martin and Klein 1984; see also *Human–Environment Relationship: Prehistoric*). Ecosystems may be resilient under the pressure of human activity until a point is reached at which there is sharp discontinuous change (see Holling 1986). Kasperson et al. (1995) identify a series of thresholds in nature–society trajectories as human activity in a region intensifies beyond sustainability: first a threshold of impoverishment, then endangerment and finally criticality—the stage at which human wealth and well-being enter an irreparable decline.

The thresholds may not be evident in advance. Under some greenhouse scenarios, for example, gradual ocean warming leads to an abrupt shift in the course of ocean currents—possibly over a period as short as a few decades—with major effects on weather patterns and the climates of coastal states. As systems models of the environment, such as general circulation models of global climate, expand in size and sophistication, their value in prediction is expected to increase. However, they may always be defeated in this task by the inherent uncertainties created by nonlinearities in nature itself. Application of the precautionary principle, aimed at keeping human environmental impact below the threshold of sustainability, may lower the likelihood of a surprise outcome but cannot eliminate it. (See also *Environmental Surprise*; *Precautionary Principle*.)

4. *Subjectivity of Impact Assessments*

The transformation of natural environments by human activity is typically identified with damage or degradation. In the case of agricultural practices leading to erosion, salination, laterization, etc, that assessment is clearly justified. The spread of crop and forest monocultures is seen as ecologically damaging through the consequent loss of biodiversity and increase in vulnerability. Some human interventions, however, can be seen as positive in ecological terms—and not only in cases of restoration of earlier damage. It is also important to note that scientific judgment in characterizing environmental impact does not necessarily accord with public sentiment. Radically transformed environments—say, English hedged fields or Balinese rice terraces—may be regarded as aesthetically pleasing, provided their ecological stability (at a new level) is maintained. Indeed, Dubos (1980) notes that some of today's most admired landscapes are products or by-products of human activity, representing 'areas deforested, swamps drained, hillsides gouged of their stones and sand.' Prior to the Romantic age, 'wilderness' tended to evoke feelings of horror more than of beauty.

Thus, it is necessary to distinguish between an environmental impact of human activity as construed by ecologists (summarized in stylized fashion by the I of the IPAT equation), and the subjective assessment of that effect. In common usage the environment is

nature viewed through human eyes, valued by human preferences. Preferences differ among people and over a person's life; they may be socially constructed; they tend to adapt to or track realities. An environmental impact that is seen by ecologists as an effectively irreversible change in a natural system—for example, extinction of a species or loss of a tropical rainforest, or the much earlier loss of temperate forests—may offer apparently compensating gains at the time and in the locality; the full costs and accompanying regret may only be felt much later or by groups far removed geographically. Moreover, regret wanes as people's expectations adjust to the new environmental circumstances. An active and timely public response in limiting adverse impacts calls for heightened ecological awareness at the various different levels of society and fuller realization that ecological stability and environmental services cannot be taken for granted.

See also: Cassandra/Cornucopian Debate; Environment and Common Property Institutions; Environment Regulation: Legal Aspects; Environmental Change and State Response; Environmental Policy; Environmental Security; Environmental Vulnerability; Human–Environment Relationships; Population Ecology; Population, Economic Development, and Poverty; Population Pressure, Resources, and the Environment: Industrialized World; Sustainability Transition: Human–Environment Relationship; Sustainable Development

Bibliography

Commoner B 1971 *The Closing Circle*. Knopf, New York

Dubos R J 1980 *The Wooing of Earth*. Athlone, London

Ehrlich P, Holdren J 1971 The impact of population growth. *Science* **171**: 1212–17

Ehrlich P, Holdren J 1972 One-dimensional ecology. *Bulletin of the Atomic Scientists* **28**: 16–27

Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8

Holling C S 1986 The resilience of ecosystems: Local surprise and global change. In: Clark W C, Munn R E (eds.) *Sustainable Development of the Biosphere*. Cambridge University Press, Cambridge, UK

Kasperson J X, Kasperson R E, Turner II B L (eds.) 1995 *Regions at Risk: Comparisons of Threatened Environments*. United Nations University Press, Tokyo

Martin P S, Klein R G (eds.) 1984 *Quaternary Extinctions: A Prehistoric Revolution*. University of Arizona Press, Tucson, AZ

Preston S H 1994 *Population and the Environment*. International Union for the Scientific Study of Population, Liège, Belgium

Vitousek P M, Ehrlich P R, Ehrlich A H, Matson P A 1986 Human appropriation of the products of photosynthesis. *BioScience* **36**: 368–73

Vitousek P M, Mooney H A, Lubchenco J, Melillo J M 1997 Human domination of Earth's ecosystems. *Science* **277**: 494–9

Wackernagel M, Rees W 1996 *Our Ecological Footprint: Reducing Human Impact on the Earth*. New Society Publishers, Philadelphia

G. McNicoll

Irrationality: Philosophical Aspects

1. The Aristotelian Tradition vs. the Heuristics and Biases Program

There is a philosophical tradition, reaching at least as far back as Aristotle, which maintains that man is a rational animal. According to this Aristotelian traditional, human beings are intrinsically rational, though of course even the most ardent defender of this view would admit that people's decisions and thought processes are sometimes very irrational indeed. When people are tired, or drunk, or in the grip of rage, they can think and act in ways that no account of rationality would condone. To accommodate this obvious fact, contemporary defenders of human rationality often appeal to a distinction between competence and performance similar to the one invoked by contemporary linguists (Chomsky 1965, Samuels et al. 1999). The claim that human beings are rational is interpreted to mean that normal humans posses a rational reasoning competence—a tacit knowledge of normatively appropriate principles for reasoning and decision-making. Under ordinary circumstances these principles are reflected in people's thinking and their actions, though sometimes, for various reasons, our performance can fail to reflect this underlying rational competence.

In the early 1970s, however, Amos Tversky, Daniel Kahneman, and a number of other psychologists began reporting findings suggesting that the Aristotelian view of human rationality is seriously mistaken. What these studies demonstrated is that even under quite ordinary circumstances where fatigue, drugs, and strong emotions are not factors, people reason and make decisions in ways that appear systematically to violate familiar canons of rationality on a broad array of problems. Those first surprising studies sparked the growth of an enormously influential research program whose impact has been felt in a wide range of disciplines including psychology, economics, political theory, and medicine. The expression 'heuristics and biases' is often used as a label for this tradition, since it holds that much of our reasoning and decision-making relies on heuristics and biases rather than rational principles.

In the sections that follow we will offer some examples of the sort of experimental findings that many have interpreted as showing that the reasoning and decision-making competence of normal humans is

far from rational. (For more comprehensive reviews of the literature see Dawes 1988 and Baron 1994.) But first we need to say what standards of rationality are being applied by those who think the experimental findings demonstrate widespread irrationality. Though writers in the Heuristics and Biases tradition rarely offer an explicit and general theory of rationality, the most plausible reading of their work is that they are assuming some version of what Edward Stein (1996) calls the 'standard picture'. On this view, to be rational is to reason and make decisions in accordance with principles that are based on the rules of logic, probability theory, and decision theory. Sometimes, of course, it is far from clear how these theories are to be applied—a problem that we will return to in Sect. 4. For the moment, however, we will follow the lead of those who work in the Heuristics and Biases tradition and assume that the irrationality manifested in the experiments we describe is intuitively obvious.

2. *Irrationality in Reasoning*

2.1 *The Conjunction Fallacy*

In what has become perhaps the most famous experiment in the Heuristics and Biases tradition, Tversky and Kahneman (1982) presented people with the following task.

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.
>
> Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable.
>
> (a) Linda is a teacher in elementary school.
> (b) Linda works in a bookstore and takes Yoga classes.
> (c) Linda is active in the feminist movement.
> (d) Linda is a psychiatric social worker.
> (e) Linda is a member of the League of Women Voters.
> (f) Linda is a bank teller.
> (g) Linda is an insurance sales person.
> (h) Linda is a bank teller and is active in the feminist movement.

In a group of naive subjects with no background in probability and statistics, 89 percent judged that statement (h) was more probable than statement (f) despite the obvious fact that one cannot be a feminist bank teller unless one is a bank teller. When the same question was presented to statistically sophisticated subjects—graduate students in the decision science program of the Stanford Business School—85 percent made the same judgment! Results of this sort, in which subjects judge that a compound event or state of affairs is more probable than one of the components of the compound, have been found repeatedly since Tversky and Kahneman's pioneering studies, and they are remarkably robust. This pattern of reasoning has been labeled 'the conjunction fallacy.'

2.2 *Base Rate Neglect*

On the familiar Bayesian account, the probability of a hypothesis on a given body of evidence depends, in part, on the prior probability of the hypothesis. However, in a series of experiments, Kahneman and Tversky (1973) showed that subjects often seriously undervalue the importance of prior probabilities. One of these experiments presented half of the subjects with the following 'cover story.'

> A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

The other half of the subjects were presented with the same text, except the 'base-rates' were reversed. They were told that the personality tests had been administered to 70 engineers and 30 lawyers. Some of the descriptions that were provided were designed to be compatible with the subjects' stereotypes of engineers, though not with their stereotypes of lawyers. Others were designed to fit the lawyer stereotype, but not the engineer stereotype. And one was intended to be quite neutral, giving subjects no information at all that would be of use in making their decision. Here are two examples, the first intended to sound like an engineer, the second intended to sound neutral:

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.
>
> Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

As expected, subjects in both groups thought that the probability that Jack is an engineer is quite high. Moreover, in what seems to be a clear violation of Bayesian principles, the difference in cover stories between the two groups of subjects had almost no effect at all. The neglect of base-rate information was even more striking in the case of Dick. That description

was constructed to be totally uninformative with regard to Dick's profession. Thus the *only* useful information that subjects had was the base-rate information provided in the cover story. But that information was entirely ignored. The median probability estimate in both groups of subjects was 50 percent.

Before leaving the topic of base-rate neglect, we want to offer one further example illustrating the way in which the phenomenon might well have serious practical consequences. Here is a problem that Casscells et al. (1978) presented to a group of faculty, staff, and fourth-year students at Harvard Medical School.

> If a test to detect a disease whose prevalence is 1/1,000 has a false positive rate of 5 percent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? — percent.

Under the most plausible interpretation of the problem, the correct Bayesian answer is 2 percent. But only 18 percent of the Harvard audience gave an answer close to 2 percent. Forty-five percent of this distinguished group completely ignored the base-rate information and said that the answer was 95 percent.

3. *Irrationality in Decision-making*

3.1 *Framing*

In a study that is widely believed to illustrate a deeply irrational feature of human decision-making, Tversky and Kahneman (1981) presented a group of subjects with the following problem:

> Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.

A second group of subjects was given an identical problem, except that the programs were described as follows:

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die.

On the first version of the problem most subjects chose Program A. But on the second version most chose Program D, despite the fact that the outcome described in A is identical to the one described in C.

3.2 *The Ultimatum Game*

Consider a game that works as follows: There are two players, an 'Allocator' and a 'Recipient' whose identities are unknown to each other. At the beginning of the game the Allocator is given a sum of money (say $10) and asked to divide it between himself and the other player. The only restriction is that the Allocator must offer *something*. So, if the smallest unit of money available is $1, then the Allocator must offer at least this much to the Recipient. Once the Allocator has divided the money, the Recipient gets to decide whether to accept or reject the amount offered. If the Recipient accepts, each player gets to keep the sum specified by the Allocator. If the Recipient rejects the offer, then both players receive *nothing*. The game is not repeated, and the players know from the outset that they will play only once.

On one widely accepted analysis of the problem, it is never rational for a Recipient to reject an offer, since it is better to get *some* money rather than none. If this is right, then it is always rational for the Allocator to offer the Recipient the smallest amount of money possible, since rational Recipients will always accept the offer, and this will maximize the amount of money the Allocator receives. However, people who have played the game rarely make these choices. Indeed, in many studies the modal offer is to divide the money equally. Moreover, Recipients frequently reject offers of less than 25 percent of the total, preferring instead to go home empty-handed. (For an overview of the literature see Thaler 1992.)

4. *'Bleak Implications', or 'Ecological Rationality'?*

The results we've sketched are just a tiny sample of an enormous and growing literature on human reasoning and decision-making. But what conclusions about human rationality—and irrationality—should we draw from these findings? On this question opinions are sharply divided. In one often quoted comment, two leading investigators in the Heuristics and Biases tradition claimed that the experimental results have 'bleak implications' for human rationality (Nisbett and Borgida 1975). Other researchers have concluded that 'individuals are generally affected by systematic deviations from rationality' (Bazerman and Neale 1986). Still others have suggested that the fault may be in our genes, or at least in our evolutionary history—'that people lack the correct programs for many important judgmental tasks … [because we] … have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty' (Slovic et al.

1976). But many other theorists have argued that conclusions like these are unwarranted. In what follows we discuss two of the more important kinds of argument proposed by those who reject the bleak implications view. (For an extended review of the debate see Samuels et al. 2001.)

4.1 Problems with Specific Experiments

There is a substantial literature arguing that one or another experiment in the Heuristics and Biases tradition is flawed and thus does not really demonstrate that the subjects' performance is irrational. Some of the putative flaws are methodological. One charge that obviously has some merit is that it is often less than clear that the subjects understand the problem in the way that the experimenters want them to. For example, principles of conversational implicature might lead subjects in the 'feminist bank teller' experiment to assume that choice (f) (Linda is a bank teller) means that she is a bank teller who is *not* active in the feminist movement (Schwarz 1996). Other putative flaws turn on the way that the experimenters propose to apply the normative principles to particular problems. For instance, Gigerenzer (2000) argues that there are typically several different ways in which the principles of statistics and probability can be applied to a given problem and that these different analyses of the problem lead to different answers—or in some cases to no answer at all. If this is correct, then obviously we cannot conclude that subjects are being irrational simply because they do not give the answer that the experimenters prefer. Similarly, it has been urged that it is far from clear that a rational Recipient in the Ultimatum Game ought to accept all offers.

4.2 Darwinian Problems

A second response to the claim that the human mind is prone to 'systematic deviations from rationality' comes from recent work by evolutionary psychologists. These theorists argue that the ability to use probabilistic information effectively would have been highly advantageous to our hominid forebears, and therefore, they maintain, it would be surprising if we had not evolved 'an intellect capable of dealing conceptually with uncertainty.' But, the evolutionary psychologists continue, we should expect this capacity to manifest itself most clearly when probabilistic information is presented in a format that would have been common in ancestral environments. In these environments, probabilistic information would almost invariably have been presented in the form of information about frequencies as opposed to single events. Thus, evolutionary psychologists argue, we should expect people to do much better on probabilistic reasoning tasks if the problems are presented in a frequency format. It appears that they do. In the 'feminist bank teller' problem, for example, if the description of Linda is followed by a question like the one below, the number of subjects who commit the conjunction fallacy drops from over 90 percent to only about 10 percent.

> There are 100 people who fit the description above. How many of them are:
> ...
> (f) bank tellers?
> ...
> (h) bank tellers and active in the feminist movement?
> ...

Further evidence comes from Cosmides and Tooby's (1996) systematic exploration of the 'Harvard Medical School problem' discussed in Sect. 2.2. They showed that subjects find it much easier to use base rate information rationally on 'frequentist' versions of the problem, like the one that follows.

> 1 out of every 1,000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1,000 people who are perfectly healthy, 50 of them test positive for the disease.
> Imagine that we have assembled a random sample of 1,000 Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people.
> Given the information above:
> on average,
> How many people who test positive for the disease will *actually* have the disease? __ out of __.

On this version of the problem, 76 percent of subjects—as opposed to the mere 18 percent in Casscells et al.'s original study—gave the correct Bayesian answer.

In light of findings like these, evolutionary psychologists have introduced a notion of 'ecological rationality' (Gigerenzer et al. 1999). An inferential or judgmental strategy is ecologically rational if it is both accurate and efficient on the sorts of tasks that were important in the environments in which we evolved. If we interpret Aristotle's thesis as the claim that man is an ecologically rational animal, evolutionary psychologists maintain that there is a growing body of evidence indicating that he was right.

See also: Decision Theory: Bayesian; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Relativism: Cognitive; Relativism: Philosophical Aspects; Sociobiology: Overview; Sociobiology: Philosophical Aspects

Bibliography

Baron J 1994 *Thinking and Deciding*, 2nd edn. Cambridge University Press, Cambridge, UK

Bazerman M, Neale M 1986 Heuristics in negotiation. In: Arkes H R, Hammond K R (eds.) *Judgment and Decision Making: An Interdisciplinary Reader*. Cambridge University Press, Cambridge, UK

Casscells W, Schoenberger A, Grayboys T B 1978 Interpretation by physicians of clinical laboratory results. *NEJM* **299**: 999–1001

Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Cosmides L, Tooby J 1996 Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* **58**: 1–73

Dawes R M 1988 *Rational Choice in an Uncertain World*. Harcourt Brace Jovanovich, San Diego, CA

Gigerenzer G 2000 *Adaptive Thinking: Rationality in the Real World*. Oxford University Press, New York

Gigerenzer G, Todd P M, ABC Research Group 1999 *Simple Heuristics That Make Us Smart*. Oxford University Press, New York

Kahneman D, Tversky A 1973 On the psychology of prediction. *Psychological Review* **80**: 237–51

Nisbett R, Borgida E 1975 Attribution and the social psychology of prediction. *Journal of Personality and Social Psychology* **32**: 932–43

Samuels R, Stich S, Tremoulet P 1999 Rethinking rationality: From bleak implications to Darwinian modules. In: LePore E, Pylyshyn Z (eds.) *What is Cognitive Science*. Blackwell, Malden, MA

Samuels R, Stich S, Faucher L 2001 Reason and rationality. In: Nijniluoto I, Sintonen M, Wolenski J (eds.) *Handbook of Epistemology*, Kluwer, Dordrecht, The Netherlands

Schwarz N 1996 *Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation*. Erlbaum Associates, Mahwah, NJ

Slovic P, Fischhoff B, Lichtenstein S 1976 Cognitive processes and societal risk taking. In: Carol J S, Payne J W (eds.) *Cognition and Social Behavior*. Erlbaum Associates, Hillsdale, NJ

Stein E 1996 *Without Good Reason*. Clarendon Press, Oxford, UK

Thaler R H 1992 *The Winner's Curse: Paradoxes and Anomalies of Economic Life*. Free Press, New York

Tversky A, Kahneman D 1981 The framing of decisions and the psychology of choice. *Science* **211**: 453–8

Tversky A, Kahneman D 1982 Judgments of and by representativeness. In: Kahneman D, Slovic P, Tversky A (eds.) *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK

R. Samuels and S. Stich

Irrigation Societies

The notion of 'irrigation societies' derives from one of the most durable ideas in social science, that the construction and operation of large-scale irrigation systems leads to a concentration of power in the state. An early formulation of this proposition is found in Marx's writings on 'Asiatic societies': 'This prime necessity of an economical and common use of water—necessitated in the Orient—the centralizing power of Government' (Marx 1853, p. 7). In the 1950s, Karl A. Wittfogel (1957) elaborated Marx's thesis, correlating irrigation scale and political hierarchy. Beginning in the 1960s, archaeologists incorporated these ideas into theories of state formation in the Old and New Worlds. The control of water was linked to the appearance of civilizations in Mesopotamia, Peru, the Valley of Mexico, the Indus Valley, China; and to chiefdoms in the prehistoric American Southwest, Polynesia, and elsewhere (Wright 1986). In the 1970s, sociocultural anthropologists began to take issue with the Wittfogel thesis, noting that in some cases small-scale irrigation systems actually appear to promote egalitarian social systems, not despotic states. Contemporary research on irrigation societies consists of case studies; comparative cross-cultural investigations of the relationship between irrigation and power; analysis of the 'design principles' (Ostrom 1992) of successful irrigation systems; and studies of the ecological impact of irrigation.

1. Overview of Irrigation Systems

It is estimated that humanity now uses 26 percent of total terrestrial evapotranspiration and 54 percent of runoff that is geographically and temporally accessible. Most of this water is used for irrigation. The average water application rate in 1990 was estimated at 12,000 cubic meters per hectare for 240 million hectares of irrigated lands worldwide (Postel et al. 1996). In the past three decades, there has been an enormous expansion of new irrigation systems. According to the World Bank, an average of 885 large dams ($>$ 15 m high) were constructed per year between 1950 and the 1980s. During the same interval, the total area of irrigated agriculture in the world increased by about threefold (Cernea 1985, p. 23). Recently, the rate of construction of large dams has begun to decline, averaging no more than 500 per year, and this number is expected to decline further because of rising economic, social, and environmental costs (Postel et al. 1996).

Meanwhile, small-scale, hand-built, gravity-flow canal systems managed by small groups of farmers still account for about 85 percent of the world's irrigated cropland, according to World Bank estimates. Small-scale systems operated by local irrigators are still used to grow most of the crops in Asia. For example, more than 15,000 systems of less than 50 hectares are managed by farmer's associations in Korea (Oh 1978). In southern India (Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu), more than 125,000 community tanks provide water for 30 percent of the irrigated croplands (Sakthivadivel and Shanmugham

1987). More than six million small tanks and reservoirs are managed at the village level in China (Yu and Buckwell 1991). Similarly, in Morocco 830,000 hectares are irrigated by locally managed systems utilizing perennial flows and seasonal floods, which range in size from 50 to 3000 hectares (Abdellaoui 1989). About a million hectares are cultivated in Nigeria using traditional irrigation or flood-recession systems, while an estimated quarter million hectares of pre-Hispanic Peruvian terraces are still irrigated in the Andean highlands (Mabry 1996).

Small-scale irrigation systems are also found in the developed world. Fifty thousand local irrigation systems exist in Japan, some in continuous operation since the early seventeenth century, and together constitute about a quarter of Japan's total irrigated area (Takeuchi 1980). Centuries-old irrigation systems also thrive on the Mediterranean coast of Spain (Glick 1970), in the Swiss Alps (Netting 1981), and in the Southwestern USA (Mabry 1996).

2. *Environmental Constraints on Irrigation*

The physical environment constrains the kinds of irrigation systems that can be constructed. Wittfogel's thesis was based on the demands imposed by large-scale, pre-modern canal irrigation systems, fed by large rivers. Such systems, typical of arid and semi-arid regions such as those adjoining the Tigris, Euphrates, and Indus rivers, required the coordinated labor of thousands to be constructed and maintained. Tropical, semitropical, island, and desert environments impose quite different constraints, leading to different technical and managerial solutions. Scarborough (1998) has proposed an 'accretional' model for the modification of tropical and semitropical environments, where there is great species diversity but little abundance of one specific species in any single niche or patch. Archaic civilizations evolving in these settings adapted to this dispersed resource base in a manner markedly different from that of early states located in semiarid settings with access to large rivers. According to Scarborough, the political economies of tropical archaic states adapted to the interconnectivity of dispersed resources, discovering and refining the ecological pathways that made useable resources accessible to humans. For example, the civilizations of the archaic Lowland Maya, and in many regions of South and Southeast Asia developed irrigation systems based on tanks designed to capture wet-season flows and release them in the dry season. Such tanks are still in use in southern India and Sri Lanka.

Oasis civilizations of Central Asia, China, and Arabia developed tunnels called *qanats* to capture the snow melt from the Himalayas and other mountain chains, transporting the flow for great distances under the deserts. In parts of North Africa, including the Nile Valley and the American Southwest, flood recession systems were invented to capture and control scarce seasonal rainfall. Mandelbrot and Wallis (1968) showed that data on the annual flood of the Nile River exhibits chaotic behavior. Using this historic data from the Nile as well as modern data from the Senegal river basin, Park (1992) has argued that farmers adapted to the risks of too much or too little water by creating institutions to reallocate land among cultivators annually, fostering control by the elite and class stratification long before population pressure reached significant levels.

3. *Management of Irrigation*

The relationship between the technical and social aspects of irrigation is an ongoing topic of research. There are two major schools of thought. One tradition, emanating from Wittfogel and pursued primarily by archaeologists, is concerned with the power-centralizing effects of large-scale irrigation systems. A second tradition approaches small-scale irrigation systems from the perspective of common property management (or 'common-pool resources'). A key issue for both research traditions is the relationship between the scale of irrigation systems and the structure and complexity of management institutions. Hunt (1989) distinguishes between 'irrigation communities' and 'water-user associations': the former are autonomous, self-contained organizations that are supported by consensual, mutually beneficial institutional arrangements while the latter are units within larger bureaucratic structures that require communication and coordination across organizational levels.

It appears that many small-scale systems of irrigation management are self-governing. Hunt (1988) compared fifteen irrigation systems in industrialized and developing countries, ranging from 700 to 30,000 hectares in extent, and found no strong relationship between system size and the structure of authority. But he predicted that any system larger than about 100 hectares has a 'very high probability' of being managed by a 'unified administrative authority structure' (Hunt 1988, p. 347). Netting has argued that water scarcity can be a stimulus for the centralization of decision making, when it threatens widespread conflict that will seriously reduce the efficiency of the system. Kappel identified 5,000 as the upper limit of the number of people who can be supported by local self-governing systems, based on a study of 17 historic and contemporary cases. However, Uphoff found that irrigation systems with command areas of up to 40 hectares tend to be managed by the entire group of irrigators; those between 40 and 400 hectares are managed by a central official, either elected by the irrigators or appointed by the state, and systems between 400 and 4,000 hectares often have three layers of management (cited in Mabry 1996).

Turning to the dynamics of management in small-scale irrigation systems, Ostrom (1992, p. vii) notes that 'a self-governing irrigation system is a prime example of a public enterprise in which a segment of society governs itself for itself.' A frequent starting-point for the analysis of irrigation communities is Hardin's (1998) 'tragedy of the commons' thesis, which holds that rational self-interest will motivate individuals to take as much of a common resource as possible. This is linked to the 'tail-ender' problem in irrigation systems: irrigators closer to the source can control the flow to their downstream neighbors, so the farmer at the tail end is always vulnerable. How then are sustainable cooperative systems of water sharing possible?

One answer is suggested by game theory, specifically the iterated Prisoner's Dilemma, which shows how it can be in one's interest to cooperate with other users if there is an expectation that such cooperation will result in continuing access to a valuable, shared resource. Irrigation systems do not only provide benefits, they often entail high labor costs, and this can also be a motive for cooperation. Ostrom (1992) found that in negotiations over the rules of water use between the members of a local water-user association, the bargaining power of tail-enders is greater if their labor is needed to maintain the system.

Ecological factors may also influence patterns of cooperation. Rice farmers on the Indonesian Island of Bali must cope with a suite of rice pests, including rats, insects, and bacterial and viral diseases. The traditional method for controlling these pests is to plant large blocks of rice terraces at the same time. After harvest the fields are flooded, depriving the pests of their habitat, and in some cases killing them outright. Thus upstream farmers, who control the flow of water, may benefit by cooperating with their downstream neighbors. In this way, the scheduling of irrigation can be used to manage pests (Lansing 1991).

4. *Irrigation and Agro-Ecology*

Salinization resulting from over-irrigation and poor drainage has significantly reduced yields on more than 60 million hectares, about a quarter of the total worldwide, and 25 million hectares have been abandoned because of salt accumulation (Mabry 1996, p. 229). Annually, for every additional hectare brought into production by new irrigation projects, another goes out of production because of salinization (Umali 1993).

Flowing water can pick up and transport mineral nutrients needed by plants, such as nitrogen, phosphate, and potassium. Traditional Asian rice paddies use irrigation systems to bring these nutrients, as well as water, to the plants. For example, volcanic ash from Indonesian volcanoes contains about 1 percent potassium, and relatively high levels of phosphate. Until the advent of the 'Green Revolution' in rice in the 1970s (the spread of high-yielding crops designed to rely on chemical fertilizers), most of the nutrients needed in Asian rice paddies were supplied by this process of mineral leaching. In Western Europe and the USA, the spread of 'Green Revolution' farming techniques has created an over-abundance of nutrients: too much of a good thing. Worldwide, between 1960 and 1980 the application of nitrogen fertilizers increased more than five-fold, and in the decade that followed, more synthetic fertilizer was spread on land than had been applied in the entire previous history of agriculture. Rivers like the Mississippi and the Colorado acquire heavy loads of nutrients like phosphate, and generate enormous pollution problems downstream. Every summer, approximately 18,000 square kilometers of the deep waters of the Gulf of Mexico become a 'dead zone.' European countries like The Netherlands have begun to implement strict regulations restricting fertilizer use in order to protect freshwater ecosystems, but meanwhile pollution problems have scaled upwards from lakes and streams to the Baltic and North Sea. Worldwide, virtually all coastal zones are experiencing rapidly increasing pollution from agricultural runoff and domestic waste (Nixon 1998). Archaeologists generally agree that environmental problems associated with the expansion of irrigation systems led to the decline of many ancient states. Today, the spread of irrigated agriculture on which the continuing expansion of the human population depends poses a 'tragedy of the commons' problem on a global scale.

See also: Civilizations; Land Tenure

Bibliography

Abdellaoui R M 1989 Small-scale irrigation systems in Morocco. In: *Public Intervention in Farmer-Managed Irrigation Systems*. International Irrigation Management Institute, Digana, Sri Lanka, pp. 165–73

Cernea M M 1985 *Putting People First: Sociological Variables in Rural Development*. Oxford University Press, New York

Glick T 1970 *Irrigation and Society in Medieval Valencia*. Harvard University Press, Cambridge, MA

Hardin G 1998 Extensions of the tragedy of the commons. *Science* **280**: 682–3

Hunt R C 1988 Size and structure of authority in canal irrigation. *Journal of Anthropological Research* **44**(4): 335–55

Hunt R C 1989 Appropriate social organization? Water user's associations in bureaucratic canal irrigation systems. *Human Organization* **48**: 79–90

Lansing J S 1991 *Priests and Programmers: Technologies of Power in the Engineered Landscape of Bali*. Princeton University Press, Princeton, NJ

Mabry J B (ed.) 1996 *Canals and Communities: Small-Scale Irrigation Systems*. University of Arizona Press, Tucson, AZ

Mandelbrot B B, Wallis J R 1968 Noah, Joseph and operational hydrology. *Water Resources Research* **4**(5): 909–18

Marx K 1968 [1853] The British Rule in India. In: Avineri S (ed.) *Karl Marx on Colonialism and Modernization*. Doubleday, Garden City, New York
Netting R M 1981 *Balancing on an Alp*. Cambridge University Press, Cambridge, UK
Nixon S 1998 Enriching the sea to death. *Scientific American Quarterly* **9**(3): 48–63
Oh H-S 1978 Customary rules of water management for small irrigation reservoirs in Korea. *Journal of Rural Development* **1**: 96–110
Ostrom E 1992 *Crafting Institutions for Self-Governing Irrigation Systems*. ICS Press, San Francisco, CA
Park T K 1992 Early trends toward class stratification: Chaos, common property and flood recession agriculture. *American Anthropologist* **94**(1): 90–117
Postel S L, Daily G C, Ehrlich P R 1996 Human appropriation of renewable fresh water. *Science* **271**(February): 785–7
Sakthivadivel R, Shanmugham C R 1989 Issues relating to interventions in farmer-managed irrigation: Rehabilitation of a tank irrigation system. In: *Public Intervention in Farmer-Managed Irrigation Systems*. International Irrigation Management Institute, Digana, Sri Lanka, pp. 113–30
Scarborough V L 1998 Ecology and ritual: Water management and the Maya. *Latin American Antiquity* **9**(2): 135–59
Takeuchi S 1980 Japan. In: Takeuchi H (ed.) *Farm-level Water Management in Selected Asian Countries*. Asian Productivity Organization, Tokyo, pp. 76–90
Umali D L 1993 *Irrigation-induced Salinity*. Technical Paper No. 215. World Bank, Washington, DC
Wittfogel K A 1957 *Oriental Despotism: A Comparative Study of Total Power*. Yale University Press, New Haven, CT
Wright H T 1986 The evolution of civilizations. In: Meltzer D, Fowler D, Sabloff J A (eds.) *American Archaeology, Past and Future*. Smithsonian Institution Press, Washington, DC, pp. 323–68
Yu C L, Buckwell A 1991 *Chinese Grain Economy and Policy*. CAB International, Wallingford, UK

J. S. Lansing

Islam and Gender

In Islam, religious texts define the parameters of the relationship between Islam and gender; modern scholarship then interprets and gives meaning to this relationship. This essay on Islam and gender comprises two parts; part one discusses the pertinent Islamic religious texts, and part two reviews their interpretation in modern scholarship.

The gender experience in Islam is defined mainly through the *Koran*, the Book of God according to the Muslims, which is regarded as perfect, eternal, and unchangeable. Containing 114 chapters (*sura*s) and 6000 verses (*ayat*s), the *Koran* often addresses all believers, and, in specific instances, women alone. These latter instances are located specifically in chapters (*sura*s) 4 (*Al-Nisa'*), 5 (*Al-Ma'ida*), 24 (*Al-Nur*), and 60 (*Al-Mumtahana*). Chapter 4 contains the most number of verses on women and Chapter 5 spells out the conditions of marriage, dowry giving, divorce, and inheritance. Among these, three instances that affirm the male dominance over the female have been challenged, particularly by feminist scholars. One instance on polygam states that 'you may marry other women who seem good to you: two or three or four of them. But if you fear that you cannot maintain equality among them, marry only one'…(4:3). Another instance on inheritance claims 'a male shall inherit twice as much as a female' (4:11). The last instance is on punishing women that 'men have authority over women because Allah has made the one superior to the other, and because they spend their wealth to maintain them. Good women are obedient. They guard their unseen parts because Allah has guarded them. As for those from whom you fear disobedience, admonish them and send them to beds apart and beat them. Then if they obey you, take no further action against them (4:34).' There is also a reference to homosexuality in this chapter stating 'if two men among you commit indecency punish them both. If they repent and mend their ways, let them be (4:16).'

Chapter 24 refers to conditions of adultery and modesty, whereby those who defame honorable women and cannot produce four witnesses will be punished (24:4). The portions on modesty have also become contentious in that they provide the only possible reference to justify veiling and seclusion. Specifically, 'enjoin believing women to turn their eyes away from temptation and to preserve their chastity; to cover their adornments; to draw their veils over their bosoms and not to reveal their finery except to their husbands, fathers … (24:31).' However, it should be noted that restrictions are also brought upon men in the preceding part which states 'Enjoin believing men to turn their eyes away from temptation and to restrain their carnal desires (24:30).' Chapter 60 in turn asks believers to test believing women who seek refuge from them to attest if they are true believers and to protect them if they are. It is interesting to note that even though men and women are treated unequally in society, they are equal in the eye of God as believers. It is important to note that male dominance also existed in the early Jewish and Christian scriptures as well (Genesis 3:16, 1 Timothy 2:11), reflecting the status of women in the times in which they were written (Cook 2000, p. 38).

The religious texts also help portray the ideals of gender behavior in a Muslim society. According to the pre-Islamic tribes of Arabia, the 'ideals and norms of their ways were of life had been contained and preserved in their social practices (*sunna*, trodden path), the customs or oral laws handed down from the previous generations by word and example (Esposito 1988, p. 14).' The lives of the prophet Muhammad, his wives, companions, and other pious Muslims therefore provided models for all believers to emulate. Since the recorded practices of the prophet (termed *hadith*) in particular revealed God's will, they became the norm

for community life alongside the Koran. Yet these *hadith*, unlike the *Koran*, could contradict one another since they were narrated by humans rather than given by God; most of the political and social debates in the Muslim world do indeed revolve around different *hadith* versions. Nevertheless, the *Koran* and the *hadith* together form the main sources of Islamic law, followed by legal reasoning (*ijtihad*), consensus of the community (*ijma*), as well as other influences such as custom, public interest, and jurist preference or equity.

Since Islam literally means submission to the will of God, Muslims particularly emphasize correct practice; gender practice is therefore based on the lives of exemplary males and females (Ahmed 1986, 1992, Esposito 1988, Zuhur 1992). According to these authors, marriage emerges as mandatory on every Muslim man and woman unless they are financially or physically unable to marry. In marriage, husband and wife then have complementary roles, as the men mainly participate in the public sphere to secure the livelihood, and the women principally manage the family and household. Marriage recognizes the needs and wants of both partners, paralleling the story of the 'fall' in the *Koran* that identifies Adam, not Eve, as the one personally tempted by the devil. Hence neither is the woman portrayed as the cause of the 'fall' nor is there a notion of an inherited original sin for which all humanity suffers. The wives of the prophet provided, through their lives, different qualities of the ideal Muslim woman. Khadija, the first, and at the time only wife of the prophet, was a wealthy widow who employed the prophet to oversee her caravan; her prominent economic and social position as a wealthy merchant provided a strong female model. 'Aisha, whom the prophet married at age ten after Khadija's death, and who became his favorite wife, was also a forceful woman who participated in the political affairs of the community and who also narrated many of the *hadith*. Fatma, the Prophet's only recorded surviving child, provides the primal mother figure for all virtuous women to follow; one who is chaste, immaculate, and sinless. The generational transformation from Khadija to 'Aisha to Fatma captures the changing position of women. During and after the prophet's lifetime, women were gradually removed from leadership positions within the community and the prophet's wives were secluded from the public sphere to preserve the wives' honor. The public prominence of women became limited to exceptional individuals who could stand out in spite of the exclusion and seclusion. In the contemporary period, Islamist reformists such as Muhammad Abduh and Ahmad Khan did indeed identify the plight of women as the cause of the decay of the Islamic family and society, and attempted to reason that the *Koran* contained social practices of its time that needed to be reformulated in the modern period to overcome the biases against women. Western colonial practice used the same plight of women to justify their political intervention in the Middle East. The reformulation of women's position in response to both has been contested by the secular Westerners as inadequate and by the fundamentalists as inappropriate. Feminist scholars instead proposed another rereading of the sources, suggesting that in the modern period Khadija and 'Aisha rather than the obedient Fatma, need to be identified as the ideal for Muslim women (Ahmed 1986, 1992, Smith 1985).

The modern scholarship on Islam and gender centers around issues concerning the status of women in Middle Eastern societies in general, and the public participation of veiled women in particular. This selective focus, scholars argue, has largely resulted from the historical experience of the region with European colonialism when differences between the East and West were highlighted at the expense of similarities, and the image of the veiled women was specifically employed to justify the 'civilizing mission.' Historically, the publication of Edward Said's book (1978) entitled *Orientalism* alerted scholars to the European portrayal of Islam as a different, traditional, immutable, monolithic force resisting change, and the veiled oppressed image of Muslim women as its visual proof, one that also justified Western intervention. Criticism of gender analysis in other non-Western contexts (Mohanty 1984, Ong 1988, Spivak 1990) also proved helpful in further deconstructing the Western gaze in Islam and gender. These criticisms specifically engaged the growing literature of Western feminist analyses of non-Western women, pointing out how the Western gaze was uncritically replicated in these.

The literature criticizing the Western gaze over the field of Islam and gender thus identified overcoming the untexturedness of the non-Western subject and conveying her complexities as the first step in reversing the Western domination over the field of Islam and gender. The two most influential works that set the stage for the studies on gender in the Middle East (Fernea and Bezirgan 1977, Beck and Keddie 1978) were published as edited, collaborative volumes, letting the Middle Eastern women speak for themselves. More in-depth, specialized studies on particular dimensions of women's existence followed in problematizing Middle Eastern woman's alleged immutability or silence, as they studied archival court records (Tucker 1986) or women's poetry and songs (Abu-Lughod 1986). Following the queries generated by feminist theory, scholars also reflected on their own subjectivities in doing research in the field and commented on the dual structuring of knowledge by gender (Altorki and El-Solh 1988). Attempts to situate Middle East feminism in the larger context of feminist theory followed (Badran and Cooke 1990). After defining their own social space, Middle Eastern women were now located in the larger society, especially in relation to the state (Kandiyoti 1991). Redrawing the temporal and spatial boundaries of gender analysis followed as scholars reanalyzed his-

tory (Keddie and Baron 1991). More recently, the relation of religion and gender (Ahmed 1992, pp. 1–2) was studied, where it was argued that in the Middle East, women's realities are no longer captives of religion but equal partners, and that women now hold their own in the religious processes that had epistemologically obliterated their existence in the mid-twentieth century.

The most recent work in the field of Islam and gender juxtaposes theoretical abstractions with empirical realities, employing a multiplicity of academic, literary, and personal narratives to highlight the spectrum of gender experience in the Middle East both spatially and temporally (Göçek and Balaghi 1995). These range from the emergence of Islamic Eve and the reconstruction of motherhood (Spellberg 1996) to the construction of the relation between Islamic law and gender in Ottoman Syria and Palestine (Tucker 1998) to the origins of modernity and feminism in the Middle East (Abu-Lughod 1998). New efforts compare experiences within the field as they analyze the relationship between gender and support for Islamist movements in Egypt, Kuwait, and Palestine (Tessler and Jolene 1996), gender and identity construction across Turkey, Central Asia, and the Caucasus (Acar and Güneş-Ayata 2000), gender and politics in the Middle East (Berkovitch and Moghadam 1999), or women's human rights in the Muslim world (Afkhami and Friedl 1997). Other contemporary issues that engage scholars' interests range from the impact on gender relations of migration (Erman 1997), the Gulf War (Mojab 1997) or citizenship and parliamentary elections (Göçek 1999).

Works on Iran need to be mentioned as a separate category since the relation of gender and Islam has been particularly problematic after the Iranian revolution. Even though women had to re-veil in Iran, the changed nature of their participation in the public sphere became hotly debated. In spite of the religious rhetoric of returning women to their homes, the number of women in the public sphere has actually increased since the revolution whereby women have evolved a new gender politics that is empowering within the limits set by religion. The various dimensions of this new politics is currently being analyzed in relation to national memory (Najmabadi 1998), to secular versus religious gender politics (Sedghi 1996), and to the religion debate in contemporary Iran (Mir-Hosseini 1999).

A significant new effort differentiates, for the first time, the social history of women and gender (Meriwether and Tucker 1999) while another (Joseph 1999) focuses on the interaction among gender, identity, and the self. While these works provide new insights, there are still two important shortcomings in the field of Islam and gender. In defining Islam, scholars rarely take into account the vast variation in theory and praxis throughout the world. Islam is often spatially limited to the place where Islam emerged, namely the Middle East, often shortchanging its presence in the West, the emerging Muslim communities in North America and Europe, in North Africa and Africa proper, the East, Central Asian republics, the Muslim communities in China and India, and, of course, in Asia proper. Social issues such as the Islamic veil in France (Troper 1999) or women's participation in Islamist parties in Bangladesh (Shehabuddin 1999) need to be brought more to the center stage. In defining gender, even though the category theoretically includes the social interpretation of the biological divide into female and male, in practice scholars focus overwhelmingly on women often to the exclusion of men. This emphasis is again a reaction to the skewed narration in human history, which naturalized men's actions as the norm and silenced the women's contributions. More work needs to be undertaken in the field of Islam and gender emphasizing men's experiences and sexuality.

See also: Accidents, Normal; Buddhism and Gender; Catholicism and Gender; Christianity Origins: Primitive and 'Western' History; Feminist Theology; Fertility Transition: Middle East and North Africa; Islam: Asia; Islam: Middle East; Islam: Sub-Saharan Africa; Islamic Fundamentalism; Judaism; Judaism and Gender; Near Middle East/North African Studies: Gender; Near Middle East/North African Studies: Religion; Protestantism and Gender; Religion and Gender; Religion: Family and Kinship; Women's Religiosity

Bibliography

Abu-Lughod L 1986 *Veiled Sentiments: Honor and Poetry in a Bedouin Society*. University of California Press, Berkeley, CA

Abu-Lughod L (ed.) 1998 *Remaking Women: Feminism and Modernity in the Middle East*. Princeton University Press, Princeton, NJ

Acar F, Güneş-Ayata A (eds.) 2000 *Gender and Identity Construction: Women of Central Asia, the Caucasus and Turkey*. Brill, Leiden, Germany

Afkhami M, Friedl E (eds.) 1997a *Faith and Freedom: Women's Human Rights in the Muslim World*. Syracuse University Press, Syracuse, New York

Afkhami M, Friedl E (eds.) 1997b *Muslim Women and the Politics of Participation: Implementing the Beijing Platform*. Syracuse University Press, Syracuse, New York

Agmon I 1998 Women, class and gender: Muslim Jaffa and Haifa at the turn of the twentieth century. *International Journal of Middle East Studies* **30**: 477–500

Ahmed L 1986 Women and the Advent of Islam. *Signs* **11**(4): 665–91

Ahmed L 1992 *Women and Gender in Islam: Historical Roots of a Modern Debate*. Yale University Press, New Haven, CT

Altorki S, Fawzi El-Solh C (eds.) 1988 *Arab Women in the Field: Studying Your Own Society*. Syracuse University Press, Syracuse, New York

Badran M, Cooke M (eds.) 1990 *Opening the Gates: A Century of Arab Feminist Writing*. Indiana University Press, Bloomington, IN

Beck L, Keddie N (eds.) 1978 *Women in the Muslim World*. Harvard University Press, Cambridge, MA
Berkovitch N, Moghadam V 1999 Middle East Politics and Women's Collective Action: challenging the status quo. *Social Politics* **6**(1–3): 273–91
Cook M 2000 *The Koran: a Very Short Introduction*. Oxford University Press, New York
Erman T 1997 The meaning of city living for rural migrant women and their role in migration: The case of Turkey. *Women's Studies International Forum* **20** : 267–73
Esposito J 1988 *Islam: the Straight Path*. Oxford University Press, New York
Fernea E (ed.) 1985 *Women and the Family in the Middle East: New Voices of Change*. University of Texas Press, Austin, TX
Fernea E W, Bezirgan B Q (eds.) 1977 *Middle Eastern Women Speak*. University of Texas Press, Austin, TX
Göçek F M 1999 To Veil or Not to Veil: the contested location of gender in contemporary Turkey. *Interventions* **1**(4): 521–35
Göçek F M, Balaghi S (eds.) 1995 *Reconstructing Gender in the Middle East: Tradition, Identity, Power*. Columbia University Press, New York
Helman S 1999 From soldiering and motherhood to citizenship: a study of four Israeli peace protest movements. *Social Politics* **6**(3): 292–313
Hirsch S 1998 *Pronouncing and Persevering: Gender and the Discourses of Disputing in an African Islamic Court*. University of Chicago Press, Chicago
Joseph S (ed.) 1999 *Intimate Selving in Arab Families: Gender, Self and Identity*. Syracuse University Press, Syracuse, NY
Kandiyoti D 1991 *Women, Islam and the State*. Temple University Press, Philadelphia
Keddie N, Baron B (eds.) 1991 *Women in Middle Eastern History: Shifting Boundaries in Sex and Gender*. Yale University Press, New Haven, CT
The Koran 1974 [trans Dawood N J]. Penguin Classics, Middlesex, UK
Meriwether M, Tucker J (eds.) 1999 *Social History of Women and Gender in the Modern Middle East*. Westview Press, Boulder, CO
Mir-Hosseini Z 1999 *Islam and Gender: the Religion Debate in Contemporary Iran*. Princeton University Press, Princeton, NJ
Mohanty C T 1984 Under western eyes: feminist scholarship and colonial discourses. *Boundary* **12/13**: 333–58
Mojab S 1997 Women and the Gulf War: a critique of feminist responses. In: Denean Sharpley-Whiting T, White R T (eds.) *Spoils of War: Women of Color, Cultures, and Revolutions*. Rowman and Littlefield Publishers, Lanham, MD, pp. 61–82
Najmabadi A 1998 *The Story of the Daughters of Quchan: Gender and National Memory in Iranian History*. Syracuse University Press, Syracuse, New York
Ong A 1988 Colonialism and modernity: feminist re-presentations of women in on-western societies. *Inscriptions* **3/4**: 79–93
Said E 1978 *Orientalism*. Random House, New York
Sedghi H 1996 Women, the state, and development: appraising secular and religious gender politics in Iran. In: Turpin J, Lorentzen L A (eds.) *The Gendered New World Order: Militarism, Development and the Environment*. Routledge, New York, pp. 113–26
Shehabuddin E 1995 Beware the bed of fire: gender, democracy and the Jama'at-i Islami in Bangladesh. *Journal of Women's History* **10**(4): 148–71
Smith J 1985 Women, religion and social change in early Islam. In: Haddad Y, Findly E (eds.) *Women, Religion and Social Change*. SUNY Press, Albany, pp. 19–35
Spellberg D A 1996 Writing the unwritten life of the Islamic Eve: menstruation and the demonization of motherhood. *International Journal of Middle East Studies* **28**: 305–24
Spivak G C 1990 *The Post-Colonial Critic: Interview, Strategies, Dialogues*. Routledge, New York
Tessler M, Galen J 1996 Gender and support for Islamist movements: evidence from Egypt, Kuwait and Palestine. *Muslim World* **86**: 200–28
Troper M 1999 The problem of the Islamic veil and the Principle of school neutrality in France. In: Sajo A, Avineri S (eds.) *The Law of Religious Identity: Models for Post-Communism*. Kluwer Law, London, pp. 89–102
Tucker J 1986 *Women in Nineteenth Century Egypt*. American University in Cairo Press, Cairo, Egypt
Tucker J 1998 *In the House of the Law: Gender and Islamic Law in Ottoman Syria and Palestine*. University of California Press, Berkeley, CA
Zuhur S 1992 *Revealing Reveiling: Islamist Gender Ideology in Contemporary Egypt*. SUNY Press, Albany, New York

F. M. Göçek

Islam: Asia

Islam is often understood as a Middle Eastern and, even more narrowly, an Arab religion. While the origins and core scriptures of Islam are indeed Arabic, the majority of the world's Muslim population is located in South, Southeast, and East Asia. This article concerns the Islamicate civilizations and societies of these regions. It examines theoretical and methodological approaches to the study of Islam in the region, the historical processes through which Islam came to be an Asian religion, relationships between ethnicity and religion in Asian societies, popular religion and vernacular literatures in Islamic Asia, the diverse roles played by Islam and Muslims in public life in contemporary Asian States, and the significance of Islamist or fundamentalist movements.

1. 'Islam' and 'Muslim' Definitional and Theoretical Concerns

In both western and Muslim writing the terms 'Islam' and 'Muslim' are used in various and often contradictory ways. Conflicts concerning their use and definition are apparent in, and have done much to shape, scholarly discourse about religion and culture in Muslim Asia. In a theological sense the term Islam means 'submission to god.' A Muslim is a person or *jinn* (spirit) who practices submission. Debates concerning the meaning of submission are a central feature of Islamic discourse. Muslim jurists define Islam as an all encompassing behavioral code and do not distinguish between religious and other modes of social behavior. Mystics (*Sufis*) define Islam as an emotional or spiritual state and are less concerned with personal

and social behavior. Proponents of both positions often describe their opponents as *kafir* (non-Muslim).

The line between social scientific and confessional scholarship in and about the Muslim world is a fine one. Many social scientific studies share the Muslim juristic assumption that Islam is first and foremost a legal tradition and that there is a universal and normative core against which the 'Islamicness' of particular cultures can be evaluated. The result has been that societies in which the influence of Muslim mysticism predominates have often been described as nominally and even trivially Muslim (see, for example, Snouck Hurgronje 1906 and Geertz 1960). Since the early twentieth century many of the basic reference works on Islam in European languages has been written by Muslim apologists whose definitions of 'Islam' are theological motivated (for a contemporary example see Rahman 1980). For the most part their works assume a legalistic understanding of Islam. To the extent that the social science literature has been shaped by these works, it employs Muslim theological categories. Muslim scholars have, in turn, begun to use social scientific works by Western scholars with whom they are inclined to agree, as authoritative references in overtly apologetic treatises.

Recent scholarship has challenged this approach and the related view that Arabs are the guardians of Muslim orthodoxy. Increasingly, social scientists and textualists have been concerned with the dynamic tension between textual and lived Islam in diverse cultural and historical settings. The textual tradition is now understood as the product of discursive practice. Recent scholars have examined the ways in which this tradition describes, and argues for, a variety of possible 'Islams.' This perspective attempts to capture both the variety of Islamic texts and cultures and historical and doctrine linkages among them. The term 'Islamicate' is now used in reference to societies and civilizations whose development has been shaped by the interaction of Islam, as a religion, and a variety of local cultures. The use of this term circumvents concerns with orthodoxy inherent in early scholarship.

The chief difficulty with this approach is that by privileging the local character of Muslim discourse, it runs the risk of accepting the local as normative. Progress in understanding the local Islams of Asia, and elsewhere, will require additional attention to the ways in which elements of the pan-Islamic textual tradition are interpreted, understood, and put into practice, and alternatively the ways in which specific histories and cultures constrain and limit the implementation of the social ideals described in universally recognized Muslim texts. There is also a need for comparative studies of the various Asian Islams.

2. *Historical Background*

Political, economic, and military factors all contributed to the establishment of Islam as an Asian religion and the foundation of Asian Islamicate civilizations. In the seventh and eighth centuries Arab Muslim armies carried Islam into central Asia and western India. Muslim military traditions resulted in the conquest of most of South Asia. Here, Muslim expansion was motivated as much by the religious conviction that Muslim rulers have an obligation to expand the Muslim world (*dar al Islam*) as by economic considerations. Commerce spread Islam further. Arab control of the western terminus of overland and maritime trade routes linking the Mediterranean basin and China led to the expansion and consolidation of Islam in Central, East, and Southeast Asia.

2.1 *South Asia*

There has been a prominent Muslim presence in what are now Afghanistan and Sind (western Pakistan) since the early eighth century. It was not, however, until approximately 1000 CE that the conquest of the plains of what is now India began in earnest. A long series of raids and invasions based in Afghanistan led to the establishment of Islamicate empires centered in the Gangetic plain. The expansion of Muslim polities continued until the establishment of British power early in the eighteenth century. The Central Asian origins of South Asian Islam account for Turkish and Persian influences on South Asian Muslim literature and popular religion and for the presence of a significant *Shia* minority not found elsewhere in Asia. British conquest of Burma in the nineteenth century (1825–85) led many South Asian Muslims to emigrate to this Southeast Asian nation. Arab Muslim traders reached the southwest coast of India and Sri Lanka in the eighth century. Here, trade, rather than conquest, was the primary factor facilitating the expansion of Islam. The Islamicate cultures of this region resemble those of Southeast Asia and southern China more closely than those of northern portions of the Indian subcontinent.

2.2 *Southeast Asia*

Islam came to Southeast Asia in the thirteenth century. It was brought to the Malay world by Arab and South Asian Muslim traders who dominated the Indian Ocean trade during that period. Muslim states were established on both sides of the straits of Malacca and later spread southward to Java and as far east as the southern Philippines. Largely owing to the power of coastal Sultanates it rapidly became the dominant religion in the Malay-speaking world. Muslim communities in Vietnam, Cambodia, southern Thailand, and Burma originated in similar ways, but never displaced the political dominance of Indianized and, in the case of Vietnam, Sinicized states. Islam came to Burma from two additional sources in the precolonial period, from Bengal in the west and from Southern

China. During the period of British rule large numbers of south Asian Muslims migrated to Rangoon, Mandalay, and other Burmese urban centers. There are also scattered Chinese Muslim populations in northern regions of Thailand, Laos, and Vietnam.

2.3 China

Arab Muslim armies reached China's western border in AD 714. Muslim control of the silk road stretching across central Asia to western China was the avenue through which Islam came to be the religion of the various Turkic-speaking communities of central Asia, and substantial portions of northwestern China. The Mongol conquest of China and the Islamic lands of central Asia promoted closer contacts between China and the Middle East. The Mongol conquest of Persia and the sack of Baghdad in 1258 CE was a devastating blow to Islamicate cultures. However, the subsequent conversion of the Mongol elite to Islam enhanced the position of Chinese Muslim communities and resulted in the Islamicization of most of central Asia.

Legends preserved by Muslims in Southeastern China speak of Muslim emissaries and traders arriving during the lifetime of the Prophet Muhammad. While the veracity of these accounts is difficult to determine, Chinese records and archaeological evidence point to a significant Muslim presence in Southern China from the seventh century onwards. Muslims played a major role in Chinese international and internal commerce and politics from the eighth century until the establishment of the Ching dynasty in 1644. Muslims were a sizable and 'convenient' minority. Islam enabled them to establish and maintain relationships with the outside world that was both feared and belittled by the Chinese elite. At the same time Islam precluded its adherents from competition for political power within China. Javanese and other Southeast Asia chronicles mention China, as well as Arabia and India, as being among the sources of Islam. This speaks of the importance of the premodern Asiatic commercial system in the expansion of Islam in Asia.

3. Islam and Ethnicity in Asian Societies

Sociologically, religion and ethnicity are closely intertwined in the Asian Muslim world, though in complex and varied ways. In some instances Islam transcends ethnic distinctions, in others Muslim communities are divided along ethnic lines, despite the normative doctrine and political ideology of 'Muslim brotherhood.'

In nation states with Muslim minorities 'Muslim' has become an ethnic, and often political category, even where Muslims themselves would not recognize it as such. In China, the Philippines, and the Buddhist societies of mainland Southeast Asia, majority populations and governments define Islam as an ethnic category, despite linguistic and religious differences within the Muslim community. Burma provides a clear example. As the result of the inclusion of Burma in the British Indian empire, the Muslim community is extremely diverse. In addition to communities with historical ties to many regions of India, there are Chinese and ethnic Burmese Muslim communities. Typically each of these communities has its own mosque and religious officials. Burmese Buddhists and the national government ignore these distinctions and categorize all of these diverse communities as simply 'Muslim.' In nation-states with Muslim majorities Islam is generally understood as a unifying force that transcends ethnic differences.

In colonial and postcolonial societies Islam has often been used as a unifying ideology. The use of Islam as an organizing principle contributed to the rise of nationalist movements in British India and Malaysia and the Dutch Indies (Indonesia). The effectiveness of Islam as a political principle has proved to be uneven. It was largely responsible for the division of British India into the states of India and Pakistan in 1949. It provided the basis for Malay nationalism in Malaysia, but also contributed to Singapore's succession and for ongoing tension between Malay Muslim, South Asian Hindu, and Chinese communities.

3.1 South Asia

Religion and ethnicity have long been salient features of social and political discourse in South Asia. Pakistan was founded on the basis of a common religious identity, obscuring ethnic and cultural relationships between Hindus and Muslims in both the Northwest and Eastern regions as well as ethnic and linguist differences among Muslims. Pakistan subsequently split into Pakistan and Bangladesh, a division that reflects, geographic, ethnic, and linguistic differences. Bangladesh is an ethnic as well as a religious state. Pakistan is a multiethnic state that uses Islamic ideology to transcend ethnic divisions. The rise of militant Hindu nationalism as a potent political force in India has led to the identification of Islam and Muslims as an ethnically distinct and 'foreign' element of Indian society. Contemporary Indian Muslims are inclined to disregard sectarian and ethnic differences in the face of a mounting tide of Hindu religious nationalism.

3.2 Southeast Asia

Southeast Asia is characterized by extreme ethnic and religious diversity. Islam is the dominant religion in Indonesia, Brunei, and Malaysia. There are important Muslim minorities in the Buddhist countries (Burma, Thailand, Laos, and Cambodia), Sinicized Vietnam, and the Christian Philippines.

The role of Islam in the political systems of the majority Muslim nations of the region, Indonesia, Malaysia and Brunei differs fundamentally. Malaysia is officially an Islamic state, although little more than 50 percent of its citizens are Muslims. Political parties compete to appear 'more Islamic than each other.' Despite the fact that it is the world's most populous Islamic nation, Indonesia defines itself as being neither a religious or a secular state, but rather as one with the state ideology *Panca Sila* (the Five Principles) which comprise: belief in the one true God, a just and civilized humanity, national unity, democracy based on consensus of the people, and wise leadership and social justice for all of the Indonesian people. This ideology seeks to transcend religious differences in a Muslim state in which there are small, but important, Christian and Hindu minorities. Islamic political parties, several of which advocated the creation of an Islamic state, play significant roles in Indonesian politics. These groups understand Islam as an ideology that transcends ethnic distinctions. Other, more liberal, Muslim organizations have attempted to establish an Islamic theology of tolerance promoting ethnic and religious pluralism. Brunei is by far the smallest of the Muslim states of Southeast Asia. It is officially defined as a Malay Islamic Monarchy. Power is located with the Sultan and members of his immediate family. The current Sultan is known as a pragmatist and has sought to limit the influence of Muslim scholars and fundamentalists.

Other Southeast Asian nations have significant Muslim minorities. There is, however, considerable variation in the social and economic status of Muslim minority populations. There is a basic distinction between immigrant and native Muslim communities, that is, between those who define themselves in terms of ethnic categories indigenous to the region and those of Arab, Persian, and South Asian origin. The largest communities of native Muslims are found in the Southern Philippines and Southern Thailand. While Muslims make up less than 10 percent of the Philippine population, they are concentrated in the Sulu archipelago and the southern Island of Mindanao. Phillippino Muslims are culturally and linguistically related to those of eastern Indonesia and the Indonesian/Malaysian island of Kalimantan (Borneo). They fiercely resisted both Spanish and US colonialism and have never been fully integrated into the modern Philippine state. The conquest of these territories was not completed until 1914. The modern state has continued the anti-Islamic policies of Spanish and US colonialists, and encouraged Christian immigration to Muslim majority areas. The result has been continued armed struggle between the army and the Moro National Liberation front. Muslims make up approximately 10 percent of the population of Thailand and are of diverse ethnic origin. Most of the Muslim population of the Southern provinces are Malay, although there are also significant Muslims of the Thai origin. Much of this area was not part of the Thai state until the mid-nineteenth century. There are still periodic outbreaks of resistance to Thai rule. In the vicinity of Bangkok there are diverse Muslim communities of Malay, Thai, Persian, and Cham (Cambodian) origin. In the northern region there is a substantial community of Chinese Muslims originating in Yunan. In general the Thai Muslim community is not well organized as it faces continuous pressure to assimilate Thai cultural norms and Theravada Buddhism. Prior to the Khmer Rouge persecutions there were approximately 500,000 Cambodian Muslims Most are the descendants of Muslim refugees from the kingdom of Champa which was destroyed by the Vietnamese in the fifteenth century. They retain distinct ethnic and linguistic identity. Very few survive.

3.3 China

Chinese Muslims (*Hui*) are recognized by the People's Republic as an official nationality, despite considerable cultural and religious differences within the community. Some southern Chinese communities that trace their descent to Arab Muslim traders have sought to be officially recognized as *Hui* to obtain economic benefits granted to minority nationalities. Descent and refusal to eat pork are the primary factors distinguishing them from the larger *Han* Chinese community. In northwest China Turkic Muslim communities employ a combination of ethnic and religious criteria to distinguish themselves from the *Han* majority. The collapse of the Soviet Union led to an Islamic revival among the Turkic people of central Asia. This has led to an increased concern with ethnic and Islamic identity among the Muslim peoples of northwest China. The Chinese Nationalist government of Taiwan views relationships between religion and ethnicity differently than the People's Republic. It does not officially recognize the distinctiveness of Muslim Chinese. Nor does it limit or regulate expressions of Muslim religiosity.

4. Popular Religion and Vernacular Literatures

Arabic religious texts are known throughout Islamicate Asia. The *Qur'an Hadith* (traditions attributed to the Prophet Muhammad), legal, mystical, and theological works are studied in theological academies throughout the region. There are also significant vernacular literatures in Urdu, Bengali, Tamil, Javanese, and Malay. Vernacular literatures are often based on a combination of pan-Islamic and pre-Islamic components. Many recount tales of the spread of Islam and the exploits of local Muslim saints, kings, and warriors. Throughout the twentieth century there have been concerted efforts to translate

Arabic theological texts into Asian languages. This has contributed to an increased awareness and appreciation of Islamic theology and law in almost every Asian Muslim society.

The ritual complex mandated by Islamic law—the five daily prayers, alms, fasting during the month of *Ramadan*, and the pilgrimage to Mecca—are known and practiced throughout the region. Other kinds of popular religion reflect the significant role played by Muslim mystics (*Sufis*) in the conversion process. These include pilgrimage to the tombs of saints, ritual recitation of the *Qur'an*, healing rites, and the celebration of the *maulid* or birthday of the Prophet Muhammad. Sufi orders continue to play significant parts in social, religious, and political life in most Asian Islamicate societies. For many Asian Muslims Sufi devotionalism is more important than that required by Muslim law.

5. *Islamist Movements*

Islamist movements advocate the 'purification' of Muslim religious life. In general they seek to restrict devotional acts to those prescribed by Islamic law. They strongly oppose *Sufism* and rituals rooted in or associated with pre-Islamic traditions. They also seek to establish the Islamic law as the basis for social and political life. They are significant voices in intellectual, social, and political discourse throughout Islamicate Asia. Most have intellectual roots in the theology of the nineteenth century Egyptian modernist Muhammad Abduh. They emphasize a combination of modern education and religious zealotry. Many believe that the realization of the goals can be facilitated by the establishment of Islamic states. Support for these movements is found primarily among urban, middle-class movements. They are strongly opposed by traditional religious elites and in most instances ignored by rural populations.

See also: India, Religions of; International Migration by Ethnic South and Southeast Asians; International Migration: Mediterranean Basin and the Middle East; Islam and Gender; Islam: Middle East; Islam: Sub-Saharan Africa; Islamic Fundamentalism; Religion and Economic Life; Religion: Evolution and Development; Religion: Nationalism and Identity; Religious Fundamentalism: Cultural Concerns; South Asian Studies: Politics

Bibliography

Bowen J R 1993 *Muslims through Discourse: Religion and Ritual in Gayo Society*. Princeton University Press, Princeton, NJ

Dhofier Z 1999 *The Pesantren Tradition*. Program for Southeast Asian Studies Monograph Series. Arizona State University Press, Tempe, AZ

Esposito J 1987 *Islam in Asia. Religion, Politics and Society*. Oxford University Press, New York

Geertz C 1960 *The Religion of Java*. Free Press, Glencoe, IL

Gladney D C 1991 *Muslim Chinese. Ethnic Nationalism in the People's Republic*. Harvard University Press, Cambridge, MA

Israeli R, Johns A (eds.) 1984 *Islam in Asia. Volume II South and East Asia*. Jerusalem University Press, Jerusalem

Lapidus I M 1988 *A History of Islamic Societies*. Cambridge University Press, Cambridge, UK

McNeill W S, Waldman M (eds.) 1973 *The Islamic World*. The University of Chicago Press, Chicago

Rahman F 1980 *Major Themes of the Qur'an*. Bibliotheca Islamica, Minneapolis, MN

Hurgronje C S 1906 *The Acehnese, Vol. 2*. (O'Sullivan AWS trans.) Brill, Leyden, The Netherlands

Woodward M R 1989 *Islam in Java: Normative Piety and Mysticism in the Sultanate of Yogyakarta*. University of Arizona Press, Tuscon, AZ

M. R. Woodward

Islam: Middle East

This article outlines some of the major principles and practices of Islam, connecting them to the Middle Eastern historical and cultural context. It also briefly discusses the alternative forms offered by Shi'ism and Sufism.

1. *Islam in Historical Context*

Islam claims to be the direct word of the almighty and omniscient God (Allah), as directly revealed to His Messenger, Muhammad. Written down in the Quran, Allah's commands provide Muslims (those who submit) with the final and authoritative pathway to salvation, as well as a code of proper conduct covering all aspects of ordinary life. Islam is arguably the fastest growing of all religions, daily continuing its rapid spread throughout every corner of the world. Yet, despite its cosmopolitan character, Islam remains inextricably linked in language and in character to its past in the desolate heart of Saudi Arabia, where it arose in the beginning of the seventh century AD (for standard histories, see Hodgson 1974, Lapidus 1988).

This arid place was an extraordinarily unlikely spot for a new world religion to bloom. In fact, Arabia had avoided conquest by the stronger surrounding powers primarily because of its geographical isolation and physical harshness. The marginalization of Arabia meant that none of the major religions of the surrounding world had gained ascendance there, although the Arabs traced their own genealogical charter back to Ishmael, the son of Abraham and

Hagar, and therefore had a historical connection with Judaism and Christianity. But in practice they worshipped a plethora of local gods (who remain powerful today under the guise of *jinns*). These divinities offered their devotees protection from the unpredictable environment and acted as guardians during holy months when clan feuds were set aside and markets were organized on sanctified ground. One of these markets, according to many scholars, was to be found in the central Arabian crossroads of Mecca, site of the kaaba, the holy black stone that is now the goal of the *hajj*, the annual Muslim pilgrimage (for a contrary view, see Crone 1987).

There is great controversy among Western scholars over the accuracy of Islamic records concerning the early years of the religion (c.f. Crone and Cook 1977, Humphreys 1991, Wainsbrough 1977, Watt 1973, 1988), but most accept that Muhammad, the Prophet of Islam, was born in Mecca in approximately the year 570 to a member of the Quraysh clan, the kin group that controlled the market and its gods. As a posthumous child he had to rely on the guardianship of his paternal uncle, Abu Talib, and in his later revelations Muhammad was especially concerned with the protection of widows and orphans. With the help of his first wife, the successful and much older tradeswoman Khadija, he eventually became a well-to-do merchant, much respected for his integrity, good judgment, and fairness by his fellows, who called him al-Amin, the trustworthy.

At the age of 40, during a meditative retirement on Mount Hira, Muhammad was gripped by the first of the uncanny revelations that would eventually be collected into the Quran. In his vision, the angel Gabriel spoke to him, commanding him to 'Proclaim! In the name of thy Lord and Cherisher, who created—created man out of a leech-like clot.' Later revelations occurred at intervals throughout his life. One hundred and fourteen of these revelations, presented in rhymed verse of surpassing beauty and novelty, were definitively collected and collated in sections or *suras* of descending order of length (not chronology) shortly after the Prophet's death. These became the Quran, the central holy text of Islam. For the Arabs, whose major art form is poetry, the extraordinary aesthetic power of the Quranic verses was a convincing proof of their divine source. Oral recitation from the Quran remains today as the most important way for the faithful to experience the sacred (Nelson 1986).

According to Muslim hagiography, Muhammad's revelations originally had little influence in Mecca. His first converts were Khadija and some of his closest relatives, including Abu Talib's son Ali, who was later to marry Khadija's daughter Fatima. But most of the early believers were those who were poor, disenfranchised, and humble. They were drawn to the Quran's condemnation of excessive riches, to its advocacy of generous donations to care for the disadvantaged, and to its repudiation of the arrogance and selfishness of the wealthy. Yet the Quran also had a place for the interests of the upper strata. According to Muhammad's revelation, wealth and power were not bad in themselves—only greed, arrogance and an absence of community responsibility.

Nonetheless, most of the Meccan populace were unwilling to follow the Islamic demand that they pay *zakat* (a tax to be redistributed to the community of believers), cease to hoard, participate in community charity, and refuse interest on loans. According to legend, many were especially angered at Islam's absolute denial of the might of their clan gods in favor of the complete domination of Allah—a threat to the basic structure of their patriarchal society. Also galling for them was the Quranic demand that the great show submission by prostrating themselves in prayer to the Almighty God who would judge them on the Last Day.

After the death of his defender and uncle, Muhammad and his followers suffered increasing persecution in Mecca and looked for refuge elsewhere. Eventually, in 622 the Aws and Khazraj tribes of the agricultural oasis of Yathrib (later called Medina), anxious for a leader who could mediate their interminable feuds, begged Muhammad to migrate there. Fearing for his life in Mecca and hopeful of increasing his following, Muhammad accepted this invitation, and undertook the *hijra* to Medina with his followers—an event marked as the beginning of the Muslim calendar. It was there that he built the foundations of his holy empire, attracting converts from the surrounding tribes as he sought to conquer the Meccan confederation that had repudiated him. Following his inspiration, he rallied his following to struggle against the odds, defeating a much larger Meccan army at the battle of Badr in 624, successfully raiding Meccan caravans, and threatening Meccan commercial dominance. Finally in 630 the Muslim army occupied Mecca itself, as Muhammad's Qurayshian cousins were obliged to convert to his cause. When Muhammad died in 632, his followers had conquered the whole Arab peninsula, and had begun their great and ultimately successful war against the surrounding Sasanid and Byzantine empires. All of these events remain in the forefront of Muslim imagination to this day.

The Muslim explosion which burst from the hinterlands of Arabia was historically unparalleled; it united the Middle East into one open trade area, paving the way for the rapid creation of new prosperity. At the same time, the radical Muslim message of the equality of all believers struck a cord with the common people of the empires, who, theoretically at least, were liberated from their inferior status by the simple act of conversion. The rise of Islam was thus both an economic and social revolution, offering new wealth and freedom to the dominions it assimilated under the banner of a universal brotherhood guided by the

message of the Prophet of Allah. Of course, in practice the situation was more complex and ambiguous, as Arabs—proud of their early support of Muhammad—claimed superiority over the more recently converted Persians and Berbers, who in response demanded equal rights for themselves under the *sharia* or Muslim law. There was also rivalry between the old Qurayashian elite and the upstarts who had gone into exile with Muhammad, and now claimed special precedence for themselves. Nonetheless, Islam expanded rapidly, and soon predominated throughout the Middle East, from Morrocco to Afghanistan (for accounts, see Bulliet 1994, Gibb 1962, Shaban 1971, Wellhausen 1927).

In the Muslim imagination this was a period when the 'City of God' was actually realized in historical reality, under the authority of the Prophet himself and the four pious Caliphs after him: Abu Bakr (d. 634), Umar (d. 644), Uthman (d. 656), and his patrilateral parallel cousin Ali (d. 661)—who led Islam in its holy war against the unbelievers. To compare the Muslim experience with Christianity, it is as if Jesus had commanded an army that subdued Jerusalem, and Peter had become Emperor of Rome. Muslim thought is saturated with longing for a return of this idealized era when ordinary men and women worked together selflessly to realize the will of Allah under the leadership of just and divinely guided leaders, and were rewarded for their efforts by vast power and wealth.

Recalling their millennial past, the Muslim devout, unlike their Christian cousins, have never inwardly consented to the disjuncture between the religious experience of the community of believers (equal before God, led by the Prophet and his deputies) and the later reality of power-seeking secular rulers given to political intrigue and physical coercion. The remembered unity of politics and religion during the first years of Islam has always remained available as an inspiration to Muslim reformers and as a rebuke to politicians. That unity is evoked today by Muslim reformers, who inevitably decry the corruption of the times and call for a return to the pure faith and practice of the past (for an example, see Mitchell 1969).

To a great degree, the extraordinary victories achieved by early Islam were a result of the manner in which Muhammad's message transformed and expanded the worldview of the tribesmen who became fighters for the faith. As in a tribe, within the Muslim community—the *umma*—co-responsibility was required, and all believers were strictly prohibited from exploiting, injuring, or enslaving one another; they were obliged as well to share their resources through charitable contributions redistributed to the community at large. They were also expected to unite together in holy war (*jihad*) against the enemies of the faith. Joining the *umma* was a voluntary matter, but once the faith was accepted then a Muslim was forever committed, just as a tribesman was forever linked to his kin.

Also ratified and transformed by Islam were the mercantile values Muhammad had grown up with in Mecca. The Quran does not repudiate commerce, nor did the growth of Islam destroy trade and markets. On the contrary, Muhammad's message favored entrepreneurship, and integrated the ethics of the merchant into Islamic virtue. Profit, if fairly made and appropriately spent in the service of God and to help the deprived, is praised in the Quran, while the absolute right of individuals to own and dispose of personal property is assumed, as is the right to make contracts. In fact, the Quran is permeated with the imagery of the marketplace, which must have appealed directly to the individualistic mercantile values of the people of the time, both urban and rural, who were all so dependent on trade for their livelihoods (Rodinson 1973, Torry 1892).

Most importantly, the Quran emphasized the spiritual equivalence of all persons, who are held to be responsible for their own destinies and who will be judged as individuals by Allah on the Last Day. In principle, Islam is the least hierarchical of all religions: there are no priests, no church, no monks, no difference between clergy and laymen. Everyone is alike in the eyes of the transcendent and absolute God to whom all must submit. This aspect of Islam applied even to the Prophet, who was at pains to deny any sacred or superhuman status to himself. When doubters asked Muhammad to verify his revelation by miracles, he replied that the miracle of the Quran was sufficient, and that he was but an ordinary man, who had been singled out by God as the messenger of His word. In its egalitarianism and individualism, Islam accurately reproduced the dominant mores of its tribal and mercantile Arabian environment (Lindholm 1996).

Yet Muhammad's community, though it reflected the surrounding culture, was also something completely different than had existed previously. In their own imagination of themselves, the early Muslims were not yet another tribal grouping gathered around a sacred marketplace; they belonged to the 'party of God'—opponents were infidels. Nor was Muhammad one judge among many; he was the ultimate and final authority over a millenaristic nativist movement that aimed at nothing less than transformation of the world. Despite his protestations, for his followers Muhammad was not like everyone else; he was the perfect man who had actually seen the face of God (an event celebrated throughout the Muslim world).

In its prophecy, Islam followed in the tradition of the Middle Eastern exemplary religions already proclaimed by Moses, Jesus, and Zoroaster, who had all called for believers to submit to the moral codes imposed by an all-powerful yet personalized God: for them, faith and deeds, not heredity and power, were the true measure of spiritual enlightenment; the ethical action of individuals in this world was a prerequisite for passage into paradise (see Weber 1978 on the

nature of exemplary prophecy). Far from wishing to repudiate his predecessors, Muhammad believed his mission as 'the seal of the Prophets' was to correct and complete these earlier revelations, which had been misinterpreted and garbled. Because Muslims believe that the Quran preserves God's direct words in a pristine state, most are extremely reluctant to apply modern Western methods of textual criticism to the holy book.

1.1 The Rise of Shi'ism

While Muhammad lived, the *umma* was united under his charismatic leadership, but after his death, it inevitably began to lose its moral cohesion. Rebellions and internal schism rent the community, and after the death of Ali in 661 the realm of Islam was plunged into a cycle of tyrannical rulers and foreign invasions, punctuated by intermittent religiously inspired insurrections. Ali's own son, Husain, led one such revolt, but was pitilessly slaughtered in the battle of Karbela in 680—an event re-enacted and mourned today in the great self-lacerating celebrations of Muharram held among Shi'ites—the 'partisans' of Ali—who are the largest sectarian group in the Middle East (Chelkowski 1979). Today Shi'ites of different types compose nearly all the population of Iran, over 50 percent of the population of Iraq, 70 percent of Bahrain, and one-third of the populace of Lebanon. Altogether, the Shi'ites are estimated to make up 11 percent of all Muslims.

Ever since Husain's martyrdom, uprisings against the corruption of the faith have periodically been undertaken in the name of Ali and his descendants, who were believed by their partisans to be the legitimate inheritors of Muhammad's sacred authority. By restoring Ali's line to the throne, the schismatics hoped to re-establish the realm of God on earth. Usually, these movements fell victim to repression by powerful rulers, and they also suffered from internal disorganization, as there was disagreement over which of the possible claimants deserved the title of the Imam—the divinely appointed ruler. Even when successful, Ali's partisans suffered from problems of maintaining a sacred polity over several generations. For example, the great Fatimid dynasty of Egypt (969–1171) disintegrated over issues of succession, as one Imam denied his own divinity, while a later one had no sons (for a short history of Shi'ism, see Richard 1995).

The most successful Shi'ite strategy was followed by the so-called twelvers, who presently dominate in Iran. After their eleventh Imam died without issue, it was argued that the next Imam was actually living in occlusion, and would not return until the final judgement, though he could sometimes be contacted in dreams and visions. This sect then developed a complex and rigorous educational system that aimed to produce religious scholars capable of both interpreting sacred texts and of attaining mystical communion with the hidden Imam. Such men, acclaimed by their fellows for their special spiritual gifts, became exemplary figures in themselves, and were recognized as *ayatollahs*, 'miraculous signs of God.' Every *ayatollah* had his own loyal disciples, as well as access to endowments and to donations from followers; most importantly, each retained the right to independent judgment. Only with the political ascendance of Ayatollah Khomeini in the 1979 revolution was this pluralistic system overthrown, and a new centralized spiritual authority inaugurated by the state, which proclaimed itself to be the final arbiter of truth. The future of this experiment remains to be seen, but it must be noted that it flies in the face of twelver Shi'ite tradition (Arjomand 1988).

1.2 Sufism: the Mystical Path

A second reaction to the disintegration of the Caliphate was Sufism: so called, it is said, because practitioners wore rough woolen robes to indicate their unworldliness. Like the Shi'ites, Sufis also hoped for a renewal of the faith, but for them this could only be accomplished by a spiritual return to what they believed were the original practices of Islam. The more ardent of them withdrew into the wilderness where they practiced severe austerities in pursuit of an immediate experience of the divine. Many devoted themselves to repetition of devotional phrases (*dhikr*) which were combined with special breathing patterns and rhythmic movement to induce spiritual illumination (Arberry 1950, Schimmel 1975).

By the twelfth century, the Sufi pursuit of enlightenment had become institutionalized, as saintly figures (often called *shaikhs*) founded various lodges (*tariqas*) where they taught followers their own mystical practices. Students (*murids*) were immersed in a community of brothers detached from the ordinary world, obedient to the will of their spiritual guide, who alone could provide the blue patched 'frock of blessing' and dervish cap to pupils judged to have achieved a measure of divine grace. These newly enlightened individuals could then act as spiritual guides to their own pupils, adding to the *silsila*, or chain of knowledge, reaching back to the original teacher. While the Shi'ites argued for a hidden Imam, the Sufis claimed there was a *qutb*, a pivot of the universe, a perfect man veiled from ordinary men, to be discerned only by the spiritual elite, who themselves were organized in a secret hierarchical order of the 'friends of God,' manifested on the local level in the *tariqas* (Trimingham 1971).

Sufism had great influence throughout the premodern Middle East, and throughout the Muslim world, as wandering mystics furnished a cohesive vision of the cosmos in the face of a splintering polity while simultaneously providing new converts with an

experiential religion suitable for an unsophisticated audience. Establishing their lodges in the hinterlands around the tombs of saints, Sufis regularly served as the mediators in local tribal disputes and as guardians of the peace, gaining disciples and considerable power —so much so that they could occasionally become leaders in revolts, uniting tribes against the state (Gellner 1981). The greatest historical example were the Safavids, who originated as followers of a Sufi saint, and who ruled Persia from 1501 to 1722. Sufism also found a receptive audience in the cities, where the ranked organizations of worker's guilds and local clubs paralleled the hierarchical order of the *tariqas*. Military groups too found the Sufi message to their taste, while intellectuals and scholars were attracted to the controlled 'sober' forms of Sufism, so much so that it was quite common for an orthodox Muslim cleric to be a Sufi adept as well—the best known example being al-Ghazali (d.1111) (Watt 1953). In contrast, radical 'drunken' forms of Sufi worship appealed to the impoverished and sensation-seeking urban masses, who could lose themselves in ecstatic performances of *dhikr* (see Massignon 1982).

Sufism spread with Islam to the rest of the Muslim world, where it still is a powerful spiritual force. However, in the modern Middle East, Sufism has more or less lost its significance. It only remains powerful in Turkey, where the complete assimilation of orthodox Islam into the Ottoman state meant a repudiation of orthodoxy when that state lost legitimacy, leaving Sufism as the major alternative to secularism. Elsewhere, the oppositional role traditionally played by Sufism has been taken over by various fundamentalist groups which loudly accuse Sufis of being anti-Islamic saint worshippers. Yet, like Sufi *tariqas*, these groups too are founded by zealous individuals who believe they have direct access to a purer faith, and who demand the absolute and self-abnegating loyalty of a cadre of followers. The Sufic social structure thus lives on, even though its mystical content has been denied (Roy 1994).

2. *Islam in Theory and Practice*

While there has always been a powerful impulse in Middle Eastern Islam toward charismatic leadership, for the vast majority of Muslims the Quran remains the primal source of spiritual knowledge. Ideally there should be no need at all for human leadership when God has sent down rules to regulate and evaluate all human actions; rules which can be discovered by anyone who can read the holy text. But the Quran, being finite, does not and cannot explicitly cover all circumstances. There are many areas of behavior left out, some questions are not answered, other precepts are ambiguous, and sometimes there are contradictions or apparent impossibilities in the text.

In the face of these difficulties, some early Islamic thinkers (Mutazilites), influenced by Greek thought, attempted to force the Quran to conform to reason, but this effort was defeated in the ninth century by the 'traditionalists' led by the great theologian Ibn Hanbal (d. 855), who argued that the omnipotence of God meant that any constraints on Him and on His word—even the constraints of reason—were a form of idolatry. This did not mean that reason was banished —in fact, the use of analytic argument flourished in Muslim scholarship—but reason was firmly limited to explication of the legal implications of the sacred text; it could not be extended to put limits on God.

In place of an explanatory theology that attempted to unite logic and religion, Islam concerned itself with the obligations and prohibitions commanded by Allah: individual devotional practice and acceptance of the Word are what is crucial, not reflection or rationalization. In its emphasis on application, Islam has been appropriately described as a religion not so much of orthodoxy as of orthopraxy, reliant on ritual actions to reveal and inculcate faith (Smith 1957). The very word 'Islam' signifies an act—the initiatory act of surrender and submission to God. The essential practices required of all Muslims are five: the public profession of the faith (God is one, and Muhammad is his Prophet—recitation of this is sufficient for conversion), five daily prayers, the pilgrimage to Mecca for those who are able, donation of *zakat*, and the month-long fast of Ramadan. Some Muslims now say holy war should also be seen as compulsory, but this is a matter of controversy, and in any case can be interpreted as a struggle against one's own base instincts.

In all instances, these shared practices are simple, direct, and spiritually compelling. The aim of each is to infuse the faithful with the necessity of a willed subordination of the self to the dictates of the overarching Muslim community while offering in return both the promise of reward in the afterlife and immediate compensation in the here and now. The human rhythm of Muslim practice is clearest in the fast of Ramadan, which is the ninth month of the Islamic lunar calendar. During this period it is strictly forbidden to eat, drink, smoke, or have sexual contact. However, fasting is not regarded by Muslims as a way of atoning for sins through suffering, as it is in the Judeo-Christian tradition. Rather, all restrictions are lifted as soon as the sun sets, so that a greater quantity of food is eaten during the evenings of Ramadan than during the rest of the year, and the meals are of higher quality. Ramadan feasts also involve celebratory gatherings of friends and family, and a public redistribution of alms and food to the needy.

The rhythm of constraint and release, solitude and community, culminates in the great Muslim festival of *Eid al-fatr*—literally the lesser festival, but in actuality the major celebration of the year (*Eid al-attar*, the 'greater festival,' is a re-enactment of Abrahamic sacrifice in which a sheep is slaughtered and shared out

with family, dependents and the poor). *Eid al-fatr* is very much an affirmation of participation and renewal, as Muslims everywhere put on new clothes, visit, embrace, share food and drink, and gather for prayer. Forgiveness of all insults and enmity is mandated; debts are canceled and old enemies embrace to show their renewed brotherhood. The music and celebrations that marked the whole month are intensified, and the tombs of the ancestors are visited, reaffirming again the continuity of the community with its sacred past. High and low are leveled through the giving of alms; sharing in hunger gives the wealthy a sense of the contingency of their plenty and brotherhood with the impoverished, who, like all other Muslims, take part in both the fast and the feast. The egalitarian and individualist message of Islam is reaffirmed through action, as salvation is achieved by the deeds of individuals in this world, who unite with their fellows in the performance of socially and spiritually rewarding acts of self-sacrifice.

Within this theology of practice there is no original sin. Mankind has been placed on the pathway and given directions, he only needs to heed and obey. The Quran is a light to guide travelers as they work their way through the darkness of earthly life toward the climactic moment of the final judgment. Redemption, atonement, an irredeemable inner sense of evil—these aspects of Christian dogma are left aside in Islam where the punishment for turning away from the path is simple and terrible—God withdraws Himself, and leaves the individual alone to wander without a guide in the moral wilderness, destined never to reach paradise.

Yet despite the absence of original sin and the emphasis on practice, Islam is not a religion of positive thinking. Election to the ranks of the saved cannot be assumed to follow from proper practice or good works; salvation is solely a consequence of God's ineffable and unpredictable grace. Fear of the utter power and unknowability of God is mitigated only by reliance on His positive qualities. In the Quran Allah has continually described Himself as merciful and compassionate, and Muslims hope for His forgiveness. Notwithstanding God's compassion, the impossibility of pleasing an utterly transcendent, unpredictable, and indeterminable Deity necessarily leaves Muslims in a state of trepidation and perplexity as to the state of their souls.

2.1 The Primacy of Memory

In their effort to please God, and fearing their own all-too-human fallibility and hypocrisy, Muslims have naturally sought other means to secure salvation beyond the study of the Quran and following its precepts. In particular, they have tried to emulate the actual practices of Muhammad and his followers (*Sunna*—the 'trodden path'), reasoning that these must serve as the most reliable template for behavior approved by Allah. As a result, Islam, more than any other religion, has always been preoccupied with discovering, preserving, and imitating its own past. Such historical memories serve, with the Quran, as the major source for Muslim moral consciousness.

So it is that where other religions interpret their holy texts through the decisions of authoritative legislative bodies, or through deduction from general principles, Muslims instead discovered principles and authority through the collection of historical records of the words and deeds of the founder and his companions. This was the task of the preservers of reported history or, as they are conventionally known, the *ahl al-hadith*, people of 'tradition,' who at first were simply individuals who had lived in the original *umma*, and could report on its practices.

Over the generations, reciters became professionals who concentrated on developing their memories and eliminating personal biases so they could repeat, verbatim, what they had heard from those before them, without interpretation, comment, elaboration, or condensation. They also had to be moral exemplars, since *hadith* were judged reliable not only on the basis of historical criterion, but also on the basis of the integrity of the reciter. Those approved could then enter into the *isnad*, or 'chain' of reliable transmitters that was recited whenever the tradition was spoken. As Islam spread, the traditions were codified, validated, and published in six so-called 'canonical' collections, which served, with the Quran, as the basis for legal decisions.

2.2 Legalism and the Problem of Interpretation

The jurists (*faqih*) who interpreted the sacred texts at first overlapped almost completely with the *ahl al-hadith*. Neither had any formal education, and both were devoted to the study of the Quran, the *sunna*, and the traditions. The difference was that the lawyers developed regulations for matters left unclear by scripture, such as the exact nature of a property transfer. They also had to clarify the weight of any particular rule: was a practice obligatory, merely recommended, or neutral; if obligatory, should failure be punished by men, or left to the discretion of Allah?

However, from the point of view of the traditionalists, the jurists' use of reason and analogy to fill in lacunae or resolve apparent contradictions in the holy texts was frivolity at best, idolatry at worst. The potentially disastrous dispute between literalists and interpretivists was classically resolved by the theologian al-Shafi (d. 820), who argued that *ijma*, the agreement of knowledgeable scholars, ratified by the community itself, decides the manner in which ambiguities in sacred texts are to be understood. In other

words, Islamic theological disputes ought to be decided democratically, by consensus—albeit a consensus arrived at over generations.

This consensus is never final, since an authoritative dissent can, in principle, cause a reopening of the debate if that dissent manages to gain scholarly sanction and public endorsement. Shafi's compromise retained the egalitarian and populist character of Islam while simultaneously recognizing the importance of expert legal knowledge as the arbiter of tradition. Continuity through collective consensus still permitted the theoretical possibility for new interpretation—a possibility that continues to be the source of intense debate among contemporary Muslims.

Two centuries after Shafi's reforms the legal schools or *madhhab* (literally a 'chosen way') were whittled down to four, each named after its supposed founder: al-Shafi, Ibn-Hanbal, Malik b. Anas (d. 795), and Abu-Hanifah (d. 767), and each teaching its version of the law in specialized colleges (*madrasas*), and producing scholar-judges (*ulema*—people of knowledge). After some violent wrangling, it was eventually accepted that the schools should be considered equally accurate in their interpretations of divine law, since 'only God knows' the truth (Makdisi 1981). Though ideally one could choose any school to follow, most Muslims accepted the *madhhab* prevalent in their area. The central Middle East remains Hanafite or Shafite, the Malikites prevail in North Africa, while the Hanbalite school remains fairly small, but is intellectually influential. Shi'ites too have their own legal schools, slightly variant from those of the Sunnis—in fact, the first college in the world, al-Azhar in Cairo, was founded in 970 by the Shi'ite Fatimids.

Even today, despite public efforts at rationalization, Islamic law has remained remarkably personalized and open to interpretation, since there is still no central ecclesiastical organization enforcing rules, nor any standard procedure for appeal to a higher ecclesiastical court (indeed, there is no hierarchy of courts whatsoever). For these reasons, different juristconsults can issue *fatwas* (legal opinions) at odds with one another—much to the confusion of Western observers (Coulson 1964, Goldziher 1981).

2.3 Muslim Resistance to the State

Efforts by Middle Eastern states to incorporate, rationalize, and secularize Islamic law have always met with ambiguous results, as the majority of Muslims remain suspicious of the motives of secular authorities. For a recent example, when al-Azhar was completely integrated into the Egyptian state in 1961, its graduates were increasingly seen as minions of the government. In response, many self-taught Muslims have opened their own local schools for the training of religious scholars (Gaffney 1994). Similar patterns have occurred elsewhere in the Middle East, where popular mistrust of the official *ulema* has periodically led laymen and reformers to try to find a pathway to a purer and more egalitarian form of Islam.

This is so much the case today that, except in Iran, those offering an 'Islamist' critique of the state are usually autodidacts self-consciously distancing themselves from the class of the traditionally religiously educated, who are tainted by their graduation from state schools. But although ordinary Muslims are suspicious of the state, they are equally distrustful of those zealots who say they have privileged access to the truth. The best course, they believe, is to avoid conflict with the secular authorities, submit to the precepts of the Quran, *sunna*, and *hadith*, and rely on Allah's infinite mercy.

See also: Historiography and Historical Thought: Islamic Tradition; Islam and Gender; Islamic Fundamentalism; Middle East and North Africa: Sociocultural Aspects; Near Middle East/North African Studies: Religion; Prophetism; Religion: Definition and Explanation; Religion: Evolution and Development; Religion: Family and Kinship; Religion: Morality and Social Control; Religion: Nationalism and Identity; Religion: Peace, War, and Violence; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State: Cultural Concerns

Bibliography

Arberry A J 1950 *Sufism: The Religious Attitude and Life in Islam*. University of Chicago Press, Chicago

Arjomand S A 1988 *The Turban and the Crown: The Islamic Revolution in Iran*. Oxford University Press, New York

Bulliet R 1994 *Islam: The View from the Edge*. Columbia University Press, New York

Chelkowski P (ed.) 1979 *Ta'ziyeh: Ritual and Drama in Iran*. New York University Press, New York

Coulson N 1964 *A History of Islamic Law*. Edinburgh University Press, Edinburgh, UK

Crone P 1987 *Meccan Trade and the Rise of Islam*. Princeton University Press, Princeton, NJ

Crone P, Cook M 1977 *Hagarism: The Making of the Islamic World*. Cambridge University Press, Cambridge, UK

Gaffney P 1994 *The Prophet's Pulpit: Islamic Preaching in Contemporary Egypt*. University of California Press, Berkeley, CA

Gellner E 1981 Flux and reflux in the faith of men. In: Gellner E (ed.) *Muslim Society*. Cambridge University Press, Cambridge, UK

Gibb H A R 1962 *Studies on the Civilization of Islam*. Princeton University Press, Princeton, NJ

Goldziher I 1981 *Introduction to Islamic Theology and Law*. Princeton University Press, Princeton, NJ

Hodgson M 1974 *The Venture of Islam: Conscience and History in a World Civilization*. University of Chicago Press, Chicago, 3 Vols.

Humphreys R S 1991 *Islamic History: A Framework for Inquiry*. Princeton University Press, Princeton, NJ
Lapidus I 1988 *A History of Islamic Societies*. Cambridge University Press, New York
Lindholm C 1996 *The Islamic Middle East: An Historical Anthropology*. Blackwell Publishers, Oxford, UK
Makdisi G 1981 *The Rise of Colleges: Institutions of Learning in Islam and the West*. Edinburgh University Press, Edinburgh, UK
Massignon L 1982 *The Passion of al-Hallaj: Mystic and Martyr of Islam*. Princeton University Press, Princeton, NJ, 4 Vols.
Mitchell R P 1969 *The Society of the Muslim Brothers*. Oxford University Press, Oxford, UK
Nelson K 1986 *The Art of Reciting the Quran*. University of Texas Press, Austin, TX
Richard Y 1995 *Shi'ite Islam: Polity, Ideology and Creed*. Blackwell Publishers, Oxford, UK
Rodinson M 1973 *Islam and Capitalism*. Pantheon, New York
Roy O 1994 *The Failure of Political Islam*. Harvard University Press, Cambridge, MA
Schimmel A 1975 *Mystical Dimensions of Islam*. University of North Carolina Press, Chapel Hill, NC
Shaban M A 1971 *Islamic History A.D. 600–750 A.H. 132. A New Interpretation*. Cambridge University Press, Cambridge, UK
Smith W C 1957 *Islam in Modern History*. Princeton University Press, Princeton, NJ
Torry C 1892 *Commercial-Theological Terms in the Koran*. E. J. Brill, Leiden, The Netherlands
Trimingham J 1971 *The Sufi Orders of Islam*. Oxford University Press, Oxford, UK
Wainsbrough J 1977 *Quranic Studies: Sources and Methods of Scriptural Interpretation*. Oxford University Press, Oxford, UK
Watt W M 1953 *The Faith and Practice of al-Ghazali*. Allen and Unwin, London
Watt W M 1973 *The Formative Period of Islamic Thought*. Edinburgh University Press, Edinburgh, UK
Watt W M 1988 *Muhammad's Mecca: History in the Qur'an*. Edinburgh University Press, Edinburgh, UK
Weber M 1978 *Economy and Society*. University of California Press, Berkeley, CA
Wellhausen J 1927 *The Arab Kingdom and its Fall*. Curzon Press, London

C. Lindholm

Islam: Sub-Saharan Africa

An estimated 140 million Muslims live in tropical Africa, also referred to as sub-Saharan and Black Africa. The majority are concentrated in the Sahelian and Savannah regions that lie to the south of the Sahara proper, between the Atlantic to the west and the Red Sea to the east, in the Horn of Africa and along the East African coast from Somalia to the northern region of Mozambique. Nigeria's Muslim population, generally estimated to be over 50 million, is the largest in this part of Africa, and the focus in this article will be on the growth and impact of Islam there, and in East Africa, especially in Tanzania where, as in Nigeria, there is also a heavy concentration of Christians, a situation that affects the character, policies, and objectives of both religions in both countries. Elsewhere either Muslims or Christians are in a clear majority.

The advance of Islam and Christianity throughout tropical Africa has not undermined all traditional beliefs and rituals. On the contrary, as modernization has brought new challenges to community and personal life, people have not rejected the traditional belief systems for the 'newer' religions but have tended to juxtapose—the term syncretize is misleading—them, and to have recourse to either one or the other depending on the concerns that require a spiritual solution.

Broadly speaking, there have been three stages in the development of Islam in tropical Africa: quarantine, mixing, and reform. In the first of these stages Muslims, mainly merchants from North Africa, tended to live in closely-knit communities and apart from the indigenous population. The second stage was characterized by the mixing of Muslim and traditional beliefs and rituals, a process which gave rise to the third stage, that of reform.

1. West Africa Until the Nineteenth Century

It was through commerce that, over 1,000 years ago, Islam began to spread to West Africa and indeed to most of tropical Africa. During the eighth century, Muslim merchants of both Arab and Berber origins began to work their way across the trans-Saharan trade routes to the Ancient kingdoms of Ghana and Kanem Borno situated within the frontiers of what is today western Africa.

After gradually establishing itself as the religion of court and commerce, Islam came to be seen as a progressive, modernizing force in Tropical Africa. The traveler Ibn Battuta (1304–68/77) left an account of the healthy condition of Islam in Mali in the middle years of the fourteenth century. He was impressed by the meticulous observance of the Friday prayers, the importance attached to the learning of the *Qur'an* by heart and the fidelity with which *zakat* or almsgiving was practiced. Ibn Battuta also pointed to what has been a feature of Islam in West Africa since its beginnings there: mixing. He noted the retention in Muslim circles of traditional religious customs, including divination, prostration before a ruler, and the matrilineal system of inheritance, considered to be contrary to Islam, and given by reformers as the principal reason for *jihad*.

Early on Islam encountered Christianity in tropical Africa. By the middle of the seventh century, Arab Muslim traders had come into contact with the

Christian kingdom of Nubia which, when unified, stretched from the second to the sixth cataract of the Nile, although there was little by way of Islamization until the arrival of the Mamluk dynasty in Egypt in the second half of the thirteenth century. A combination of military pressure from Egypt and royal marriages between the Egyptian and Nubian dynasties resulted in the establishment of Muslim rule in the capital of the Nubian state of Muqurra in the first half of the fourteenth century.

2. *The Scholar-mystic as Reformer and Nationalist*

While Islam in tropical Africa has been spread largely by peaceful means and has used the same means to reform itself, the launching of the Almoravid movement in the southern sector of the Sahara in 1054 by Abdullah b. Yasin initiated a tradition of militant reform that was to influence the course of West African Islam down the centuries, inspiring future *jihads* or holy wars.

In the fifteenth century, this tradition was given intellectual and doctrinal weight by the Muslim scholar Al-Maghili (1425–1502) from Tlemcen in southern Algeria, who spent many years teaching Islam and counseling rulers in Hausaland (present day northern Nigeria) and the central Sudanic kingdom of Songhay (present day Niger). His radical teachings on the Shari'a (Islamic law), Muslim government and education, gender and taxation were communicated in reform manifestos to scholars and rulers across the western Sudan.

The seventeenth century saw the beginnings of a series of *jihads* in western Africa culminating in the Sokoto *jihad* which was to change the cultural and religious landscape of that region during the first half of the nineteenth century. It was launched by the scholar and mystic from the hinterland of Hausaland in northern Nigerian, Shaykh Uthman dan Fodio (1754–1817) early on in the thirteenth century of the Muslim *hijra* (1785–1882)—the century which, it was widely believed, would witness the advent of the Mahdi (God-guided one) who would ensure the triumph of Islam everywhere. The principal target of this jihad, one of the most thoroughgoing in the history of Islam, were the nominally Muslim urban based government administrators and the jurists who were accused of every kind of corruption and oppression. Oppression of fellow Muslims, Shaykh Uthman preached, was tantamount to infidelity on the ground that a believer was duty bound to befriend a fellow believer.

The Madhi, many were convinced, had arrived in the eastern Sudan in 1881 in the person of Muhammad Ahmad. A member of the puritannical *Sammaniyya* Sufi order, Muhammad Ahmad, in a brief period, overthrew Egyptian rule in the Sudan and established an Islamic state there which lasted until 1898. Like that of Shaykh Uthman dan Fodio, the Sudanese Mahdi's religious nationalist movement was both rural based and directed against the corrupt urban based elites, mostly Turco-Egyptian officials. The careers of both men highlight the crucial role of the mystic as reformer in African Islam.

After the Madhi the Sudan was to become, in theory, a condominium under joint British and Egyptian rule until 1952. The Sudan has suffered from a civil war between the largely Muslim North and the Christian, traditionalist South since 1955. Since that time several administrations have been engaged in extending Islamic law and education across the South.

3. *Islam on the East African Coast Until the Nineteenth Century*

As in West Africa, Islam's early development on the East Coast was slow. It spread more easily and evenly among the nomadic and seminomadic populations of Somalia and the Horn generally, while further south it remained confined to many of the island populations including those of Lamu, Pate, Pemba, Comoros, Mombasa, and Zanzibar, and to a few settlements on the coast itself. The best documented of the early coastal settlements is Kilwa, founded by Shirazi immigrants from the Arabian peninsula in either the tenth century or in the twelfth century. Though early Islam was largely confined to the coast Arab immigrants had established a Muslim state far inland in Shoa in Ethiopia by the ninth century which was later absorbed by Ifat, one of seven Muslim kingdoms all tributaries of Christian Ethiopia.

By the first half of the fourteenth century a well-developed Muslim society had been established on parts of the coast of East Africa. Once again Ibn Battuta, the chronicler and traveler, who visited Mogadishu in Somalia in 1331, provides a picture of the state of Islam at the royal palace at this time. He describes a deeply Islamized court around the Arabic speaking ruler and/or shaykh who was assisted by a Muslim judge or *qadi* from Egypt. Furthermore, East African students had already begun to travel to study in the Islamic university of Al-Azhar in Egypt, a tradition begun in West Africa much earlier.

Though historically less militant than western African or Sudanese Islam, the fusion of Islam with local coastal customs and traditions created the context for reformers to engage in *jihad* or holy war. The reformer Ahmad b. Ibrahim (1506–43), better known as Ahmad Gran, waged several *jihads* from the Muslim state of Adal, situated in the interior, into the Christian kingdom of Ethiopia which involved not only the purification of Islam but also the conversion of Christians. Paradoxically, the indirect consequence

of those *jihads* was the disintegration of the largely Muslim region of Adal, the main area of relatively settled Islamic life in the interior, although Harar, to the south, remained for Somalis a center of Muslim orthodoxy. In the second half of the sixteenth century the Galla, originally from south central Ethiopia, began to expand and by the nineteenth century a number of ethnic groups further north including the Wallo were largely Muslim.

Though Islam made little further progress in the East African interior until the nineteenth century, Arab Muslim culture had begun to interact with the indigenous Bantu cultures through marriage and concubinage to produce the new and distinctive culture known as Swahili, an Arabized-Bantu lingua franca spoken in Kenya, Tanzania, and in the eastern sector of the Democratic Republic of the Congo. The Muslim presence on the coast, on the other hand, was strengthened by the arrival of the Ottomans who occupied Masawa in 1557, and came into conflict with the Portuguese who had burnt down Zayla' in 1517.

4. *East African Islam Turns Inland*

The Yao people, situated midway between Lake Nyasa and the Indian Ocean and part of a trading network that included slave trading stretching from the Indian Ocean west to Katanga, were the only major ethnic group south of Somalia to turn Muslim prior to the nineteenth century. During the first half of that century increasing economic expansion inland began to effect a profound reorientation in outlook, attitude, and priorities among the Muslims of the coastal settlements. Muslim traders and agriculturalists began to move into the interior to Tabora, Ujiji, Nyangwe, Buganda, Lake Nyasa, Katanga, and elsewhere, and while not actively proselytizing they were, nonetheless, spreading Islamic culture by imitation.

The first Muslims arrived in Buganda, part of present day Uganda, from Zanzibar in the mid-nineteenth century during the reign of Kabaka Suna. Arab influence increased in the 1860s under Suna's successor Mutesa, not a full Muslim himself (he was not, for example, circumcised) and this was manifest in Buganda's adoption of the Islamic calendar, the observance of Ramadan, and the wearing of Arab dress which had become fashionable.

Tension created by the uneasy relationship between the demands of Islam and local custom and culture was never far below the surface and erupted in violence as protests by strict, orthodox Muslim visitors to Buganda against Mutesa's leading of the Friday prayers led to a backlash in which 100 Muslims were massacred. The situation was further complicated by the arrival of Christian missionaries in Buganda in 1877. Though in 1888 Christians and Muslims united to expel Mutesa's successor Mwanga, too much a pagan for either, unity soon turned to hostility as the Muslims expelled the Christians and installed Kalema as Kabaka. A civil war followed and by 1890 the Christians had gained the upper hand and were to receive the backing of the colonial regime. By the treatises of 1890 and 1892, Captain, later Lord, Lugard, acting in the name of the Imperial British East African Company, allocated Bugandan landed chieftainships along religious lines in favor of Christians and this was confirmed by the Ugandan Agreement of 1900. Christianity's dominance in Uganda had been institutionalized.

5. *Colonialism and Islam*

Muslim resistance to colonialism has been mentioned in the case of the Sudan and Muslim opposition, both passive and militant, emerged in many other parts of tropical Africa including the Senegambia, Somalia, and Northern Nigeria.

While profoundly disturbing culturally, psychologically, and spiritually everywhere in tropical Africa, the indirect consequences of colonial rule on the actual spread and institutional development of Islam were uneven both in time and place and depending on colonial aims, interests, rivalries, and manpower requirements. At first the German use of Muslims for administrative purposes in East Africa contributed to Islam's progress in the interior, while later this same practice worked against its advance by dividing government personnel from the indigenous peoples opposed to foreign domination. The strongest opposition to German colonial rule came in the form of the Maji-Maji (*maji* in Swahili means water) rebellion (1905–7), organized by Kinjikitile, once again a Sufi or Muslim mystic.

British colonialism which replaced the German system after World War I tended to slow the progress of Islam in Tanzania and elsewhere in East Africa. The system of Indirect Rule at the level of local government tended to protect the power of the local chiefs from being undermined by wealthy Muslim merchants from the coast. Moreover, Christian schools became the main supply of local administrators, teachers, police, and army personnel. The perception was formed of Christianity as the religion of progress, while Islam, as it was gradually sidelined, came to be seen as an obstacle to advancement in the modern sector of the economy and politics.

By contrast with East Africa, British colonial rule in Northern Nigeria relied heavily on Muslim personnel to implement the system of Indirect Rule and even extended Muslim authority to cover the so called 'pagan areas' such as the Jos Plateau. Muslim courts and schools were improved and, as in the Sudan, Colleges of Islamic Law were opened. What mostly assisted the spread of Islam during the colonial era was

the improvement in communications, while what mostly disadvantaged Islam and assisted Christianity was the growing importance of western education.

6. Postindependence Islam in the Context of the Nation State

The independence struggle and the regaining of independence itself provided Muslims, previously under represented in the modern sector of political life and in the civil service in many parts of Tropical Africa, with the opportunity to insert themselves in a more positive and sometimes aggressive manner into the political affairs of the new nation state.

Since independence in 1963, Nigerian Muslims have created a nationwide network of Islamic institutions and have supported the territorial integrity of the federation when threatened by seccessionist movements which continue to exist and the most serious of which, in recent times, came from the eastern region and resulted in the Biafran Civil War (1967–70). On the other hand, in the late 1970s as the country prepared to return to civilian rule after more than a decade of military rule, Muslim members of the Constituent Assembly appointed to approve the Constitution of the Second Republic became embroiled in a serious controversy over the establishment of a Supreme Shari'a Court at the Federal level that threatened the unity of the nation. Many Nigerians not only resented and felt threatened by the Muslim demand for a Federal Shari'a Court but also by the Muslim call for the removal from the preamble to the 1978 draft of the Constitution of the Second Republic the term 'secular state' on the grounds that it promoted secularism, atheism, and immorality. Those skeptical of northern Muslim intentions and more favorable, if not to seccession, then to greater regional autonomy, claim that Muslim enthusiasm for a strong central government reflects a concern not with harmony and unity but over the resources of the country, principally oil, which are located in the South.

Evident in tropical Africa since the 1970s is not only the greater role of Islam in the affairs of the nation state but also the steady increase in Muslims in what were traditionally non-Muslim areas such as Meru country in Central Kenya. The precise number of Muslims in Kenya is unknown but the best estimate would seem to be somewhere between 10 and 20 percent of the population. Islam is also expanding among the black population in South Africa where it was first introduced by slaves and political prisoners from the Dutch East Indies, Madagascar, and West Africa in the late seventeenth to early eighteenth century. Numbers were boosted by the influx of Muslim traders, many of them from Bombay, who distinguished themselves from the indentured laborers from India, a majority of whom were Hindu, who began arriving from 1860.

In the 1980s Muslims joined with Christians of all denominations, Jews, and Hindus in South Africa against the enforced pluralism of apartheid. Today most of South Africa's major cities have a sizeable Muslim presence, and increasing attention is given to education, community integration, and links with the wider Muslim world in Africa and beyond from which it was largely cut off during the apartheid era.

7. Conclusions

It is usual to stress the fit between Muslim culture and the African condition, and while there was, and is, undoubtedly much common ground between the two, Islam, like Christianity, has been both a destructive and creative force. It has produced many distinguished scholars, mystics, and philanthropists, among them a Muslim William Wilberforce in the person of Shaykh Uthman dan Fodio (1754–1817) of Hausaland in present day northern Nigeria whose writings and program of reform continue to influence and motivate contemporary Muslims not only in Africa but also in the United States and Europe.

Islam has been a major agent of globalization in tropical Africa. Wherever it has settled it has brought numeracy and literacy in Arabic and turned what were regional and/or national trading and commercial networks into international ones. It has also introduced new forms of religious community and spirituality in the form of the *turuq* or Sufi brotherhoods that link believers in tropical Africa to those in the wider Muslim world.

Whether Islam, or for that matter Christianity, has been an innovator or simply a catalyst, where the introduction of such core beliefs as belief in the Supreme Being as totally unique and independent is concerned, is both a fascinating and much debated point among social scientists and historians of religion, but one which space prevents us from discussing here.

See also: Historiography and Historical Thought: Islamic Tradition; Islam and Gender; Islam: Asia; Islam: Middle East; Islamic Fundamentalism; Middle East and North Africa: Sociocultural Aspects; Near Middle East/North African Studies: Religion; Religion, Sociology of

Bibliography

Clarke P B 1990 Islam in tropical Africa in the twentieth century. In: Clarke P B (ed.) *Islam*. Routledge, London, pp. 180–92

Hunwick J O 1990 Islam in tropical Africa to c. 1900. In: Clarke P B (ed.) *Islam*. Routledge, London, pp. 164–79

Lohdi A Y, Westerlund D 1999 Tanzania. In: Westerlund D, Svanberg I (eds.) *Islam Outside the Arab World*. Curzon Press, Richmond, UK, pp. 97–110

Nimtz A H 1980 *Islam and Politics in East Africa: The Sufi Order in Tanzania*. University of Minnesota Press, Minneapolis, MN

Pouwels R L 1987 *Horn and Crescent: Cultural Change and Traditional Islam on the East African Coast*. Cambridge University Press, Cambridge, UK
Rasmussen L 1993 *Christian–Muslim Relations in Africa: The Cases of Northern Nigeria and Tanzania Compared*. Academic Press, London
Trimingham J S 1964 *Islam in East Africa*. Books for Libraries, New York
Westerlund D 1980 *Ujamaa na Dini: A Study of Some Aspects of Society and Religion in Tanzania*. Almqvist and Wiksell International, Stockholm

P. B. Clarke

Islamic Fundamentalism

Islamic fundamentalism grew among sections of Muslim peoples worldwide in the twentieth century. It is not a unitary and coherent movement but represents a constellation of political, social, ideological, and intellectual trends with varying orientations and national settings.

1. General Features

Islamic fundamentalism is too much a product of modern history to admit to being regarded, as is often the case, as *sui generis*. It is much too severed from Muslim learned and popular traditions over centuries to admit to being considered, as is habitually presumed, an authentic, albeit condensed, resultant of these traditions. In so far as it represents the political expression of a particular attitude towards scriptural material, it is a phenomenon generically belonging to the history of religions, with distinctive affinities to the other monotheistic fundamentalisms, Judaism and Christianity, as well as to other fundamentalisms that are recasting themselves along monotheistic patterns, namely Buddhism, especially in Sri Lanka, and to a lesser extent Hinduism. In so far as this attitude towards scripturalist religious origins is made to converge with the modern political ideologies that power modern political movements, Muslim fundamentalism subtends conceptions of history, society, state, and of political action that quite naturally freely draw on the universal modules of populism, vitalism, and the entire repertoire of standard anti-Enlightenment motifs in the history of European and non-European peoples.

Finally, in so far as Muslim fundamentalism, in its wide variety of national settings, ideological and sectarian inflections, and political orientations, was an ideology of increasingly wide appeal in the closing decades of the twentieth century, it shares the common socio-economic characteristics of contemporary subalternist populist movements everywhere, both as diffuse sentiment and as centrally-organized, party-political movements of contestation and of revolution, mobilizing crowds as well as building up organizational structures and, in some extreme cases in Egypt, Algeria, Afghanistan, and some Western countries, re-socializing their members completely and constructing counter-societies.

2. Intellectual Origins

2.1 Muslim Reformism

The ultimate intellectual, as distinct from social, origins of Muslim fundamentalism go back to the latter part of the nineteenth century. Much like right-wing European nationalist and royalist ideologies on the ascendant at the time, and in line with the long tradition of sentiment and polemic against the Enlightenment and the French Revolution associated with, among others, Herder, Burke, and de Maistre, these were premised upon a vitalist conception of society and history. At the heart of this conception lay the contention that societies and histories constituted organic wholes sundered and rendered into a condition of morbid distemper by the historical processes of modernity. While these processes in Europe were identified as French Republicanism, socialism, and anticlericalism, in Muslim domains, and particularly in Ottoman lands, which were at the forefront of rapid change among Muslim peoples, they were identified with the state-led reforms in the domains of law reform, educational and cultural advancement, and modernist social engineering—all of which were galvanized by a positivist notion of historical evolution correlative with the assumption by the Ottoman state of direct control on the Bonapartist pattern, which saw cultural and legal hegemony as a primary task and prerogative.

Just as Herder and other conservative Germans derided the court of Frederick the Great as effeminate and 'inauthentic,' so the subalten rungs of the emergent Ottoman bureaucracy and modern, non-clerical intelligentsia saw the reformist Ottoman state and the upper rungs of the state administration as being captive to European domination. These groups likewise saw this domination personified in westernized life-styles and progressive ideas. Expressed initially by the Young Ottomans, these sentiments were vigorously expressed and most consequentially by the Persian Jamal al-Din al-Asadabadi, alias al-Afghani (1839–1897) and his then disciple, the Egyptian Muhammad Abduh (1849–1905), through the medium of the short-lived, but highly suggestive and influential review, *al-Urwa al-Wuthqa*, *The Sure Bond* (a Koranic expression), published in Paris in 1882–1883.

It was suggested in this review that nations are corporate entities which cohere around moral drives, a notion which bears close comparison with Herder's

notion of *Kräfte*. Once adulturated by foreign influences, these drives tend to atrophy, and to weaken the body-national which, in this social-Darwinist world of nations, is perforce made subject to stronger parties. Thus history is a cyclical process, an alternance of purity and adulteration, corresponding to phases of power and powerlessness: in the history of Islam, *al-Urwa al-Wuthqa* traced the causes of adulturation of the early Muslim empire to non-Arab newcomers to the faith, and to heterodox sects, and the adulteration of the Ottoman polity to the infiltration of modernist ideas and institutes of European provenance. Afghani and Abduh discounted race and ethnicity from their definition of a nation, and viewed religion as 'the sure bond' of Muslim nations. It was the Muslim religion, very broadly defined in culturalist and utilitarian rather than in devotional or dogmatic terms, which caused the organic unity of the nation to cohere, to acquire direction, and to gather force. Political action was thus conceived as an educational enterprise for the broad mass of people captive to superstition, and a conspiratorial political task directed at winning over prominent statesmen who might provide the authoritarian armature of this historical project.

Over subsequent decades, and until the late 1920s, the two prongs of this enterprise bifurcated, with Afghani, a charismatic and highly complex arch-conspirator but a superficial thinker, failing to build a Muslim internationalist coalition. Abduh, who became Grand Mufti of Egypt, independently charted the intellectual and cultural highlights of a lasting religious reform, explicitly inspired by the example of Luther, whose main thrust was to combat what he regarded as superstitious and irrational accretions that overcame what he took for the original Koranic foundations of the Muslim religion: folk religiosity, traditionalism, clericalism, and virtually all interpetations of religion that were contrary to modern science (including Darwinism) and modernist forms of political, legal, and social organization. At Abduh's hands, Islam became highly latitudinarian and episodically regarded almost as a natural religion corresponding to the social evolution of humanity.

2.2 From Reformism to Fundamentalism

It was with Abduh's star pupil, the Syrian Muhammad Rashid Rida (d. 1937), initially as liberal as his teacher, that what might be termed political Afghanism became ascendant. In his monthly review *al-Manar* (1898–1934), published from Cairo and prodigiously influential all the way fom Morocco to Java, he laid the foundations of an alliance between social conservatism and a cultural-political activism in the name of a religious reform shorn of its previous latitudinarianism, retained as a slogan voided of practical consequence, and with a pronounced anti-modernist polemical tone that sought to expand the definition of religion to encompass the entire surface of social life, rendering politics, in the words of a prominent follower, 'social devotions.' The emphasis on the strict observance of ritual, the emblematic use of politico-cultural spectacle such as extreme forms of punishment, the vituperative strictures against equally emblematic manifestations of modernity, particularly secularism and the social liberation of women, vigorous opposition to nearly all but literalist interpretations of the Koran and to miraculous and salvational views of Islamic history, and the hostility to modern legal reforms in the name of fidelity to Koranic and classical norms, propounding the 'restoration of the *Shari' a*' as a fundamental political slogan, and the revival of the notion of apostasy. These became leitmotifs of all subsequent Muslim fundamentalism in the Arab World and beyond.

This expansive definition of religion was influenced by the example of the literalist medievalism of the Wahhabi movement which led to the foundation of the Kingdom of Saudi Arabia in 1925. This movement, born in a deprived and isolated territory during the era of the French Revolution, represented a ritualistic puritanism which referred directly to the Muslim scriptures and certain fundamentalist figures in Muslim traditions, at the same time impoverishing these considerably, and reducing them to formulism.

3. Culture into Politics

To this expansive definition of the religious sphere was added an organized political moment in the 1920s and 1930s with the emergence of the Society of Muslim Brothers, first in Egypt, later spawning similar movements elsewhere.

3.1 Incipient Organization

Like subsequent moments in the history of Muslim fundamentalism, this political emergence followed the rhythms and patterns of international history which, in the interwar period, was characterized by the anti-Enlightenment and anti-Communist vitalist motifs that provided ideological sustenance to racism, hyper-nationalism, xenophobia, and cognate trends, both organized and socially diffuse. Thus, the Society of Muslim Brothers and its Muslim and Hindu homologs in India, the Jamaat-i Islami and the RSS, for instance, conceived themselves as sporting and paramilitary youth organizations, ostensibly beyond party and class politics in the name of the unity and homogeneity of the nation, conceiving of the state as a natural and organic emergence from a homogenous society, fully corresponding to its essence, momentarily in diremption and in the course of restoration to an arcadian past prior to its violation and corruption by inauthentic modernity.

This rhetoric of history underlying a political project of hegemonism and of the seizure of state power by means both gradual (emphasizing re-education and re-socialisation) and insurrectional, was given added ideological and systematic force and relevance in the period of the Cold War, with large-scale borrowing from state-led nationalism in Muslim countries and from populist conceptions of popular will, interpreted as fundamentally and abidingly religious, whereby Muslim peoples were thought to be, by natural instinct, Islamist in political orientation. This explains to a large degree the extraordinary degree of social, symbolic, and physical violence needed for nurturing these movements in their quest for power. Correlatively, these movements derived concrete logistical, institutional and organizational sustenance from the US and from Saudi Arabia, and may be said to have constituted the cultural plank of the Truman Doctrine, to which military and political aid was added, most dramatically in the Malay Peninsula in the 1950s, in Indonesia in 1965, and in Afghanistan in the 1980s.

3.2 Radical Islamism and Society

The main ideological innovation in this third phase of Islamic fundamentalism resides in the notion of *takfir*, the notion that Muslim societies are not only aberrant and unfaithful to their natural history, but that they are territories of degeneration and godlessness in need of a total recasting along models derived from the prophetic period in Islamic history, now transformed into a Utopia. First propounded by Abul-A'la Mawdudi (1903–79), the founder of the Jamaat-i Islami in India and later its leading light in Pakistan, it was taken over by the Egyptian Sayyid Qutb (1903–66), crossed with the notion of degeneration current in right-wing European political thought during the *fin-de-siècle* and the inter-war periods (prominent here was the influence of Alexis Carrel), and provided the cornerstone of the extremist fundamentalism of the 1990s: among certain groups in Egypt, Algeria, and Afghanistan, and more episodically elsewhere. Most specifically, the civil war in Afghanistan spawned a class of itinerant professional Islamist revolutionaries of various nationalities who took up arms in Algeria, Egypt, the US, Bosnia, Kashmir, and elsewhere.

At the close of the twentieth century, further elements were added to the ideological and activist repetoire of these movements, some of which had taken over state power (as in Iran, Afghanistan, and Sudan) or became potent players on the national scene (as in Pakistan), where they have been directing operations of social engineering aimed at recasting laws and public mores. The most important of these elements, again reflecting international flavors of the time, is the hyper-nationalist use of the vitalist notions of 'identity,' of cultural difference, and their cognates in a situation marked by the waning of projects of development and their cultural and social correlates, the weakening of states in the developing world under conditions of global deregulation, including cultural and social deregulation. These derive sustenance from armies of the slum-dwellers, the structurally unemployed, small businessmen belonging to a previous age, and subaltern intellectuals including members of the clerisy, most potently and successfully in Iran, where the break-up of the modenist state was most spectacularly displayed in the revolution of 1979 which has highly specific ideological and organizational characteristics.

An expansion of religious thought, in the name of culturally specific protocols of knowledge, led to an auxiliary and most important element that sustains fundamentalist bids for intellectual hegemony, namely the institutional and conceptual development of the 'Islamization of knowledge': Islamic social science, Islamic economics, Islamic democracy, the renewed hostility to Darwinism, and Islamic feminism. These currents of thought exercised collateral influence on sectors of political thought in many Muslim countries, some of them irreligious and hostile to political Islamism, and certain of these vocabularies seeped into the mainstream of political discourse, quite apart from state sponsored religion which had hitherto been marked by a latitudinarian reformism, now feeling the need to protect its flanks and pushed into a more conservative direction.

See also: Islam and Gender; Islam: Middle East; Middle East and North Africa: Sociocultural Aspects; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Politics; Near Middle East/North African Studies: Religion; Religion, Sociology of; Religious Fundamentalism: Cultural Concerns

Bibliography

Al-Azmeh A 1996 *Islams and Modernities*, 2nd edn. Verso, London

Appleby R S, Marty M E (eds.) 1991 *The Fundamentalism Project*, 5 Vols. Chicago University Press, Chicago

Ayubi N 1991 *Political Islam*. Routledge, London and New York

Choueiri Y 1990 *Islamic Fundamentalism*. Frances Pinter, London

Dabashi H 1993 *Theology of Discontent. The Ideological Foundations of the Islamic Revolution in Iran*. New York University Press, New York

Enayat H 1982 *Modern Islamic Political Thought*. Macmillan, London

Kerr M 1966 *Islamic Reform*. University of California Press, Berkeley and Los Angeles

Roy O 1994 *The Failure of Political Islam*. Harvard University Press, Cambridge, MA

A. Al-Azmeh

Issue Constraint in Political Science

The simplification of political alternatives—to left vs. right, conservative vs. liberal, authoritarian vs. democratic—often structures public discourse and policy making on important questions, such as the size of government, the authority of the central government, the extent of redistribution or market regulation, or the balance between individual liberties and social institutions. For the social scientist, this notion—ideology—provides a simple empirical metric for evaluating political alternatives, such as candidates and parties. It also offers a mathematically tractable basis for deriving clear, intuitively appealing theoretical predictions, such as are found in spatial models of elections and of legislative politics (Downs 1957, Krehbiel 1998). Most social scientists and historians not wedded to these mathematical rigors also use the shorthand of 'ideology' to interpret the complexity of politics (Beer 1978, Lowi 1979).

The adequacy of this description of politics has been the subject of intense empirical scrutiny. Perhaps the most significant concern is whether ideology adequately characterizes the way people think about political choices. Does a person's belief or opinion about one issue predict or constrain his or her beliefs and opinions on other issues? Is there, as Philip Converse termed it, 'issue constraint'?

Issue constraint consists of three separate conjectures: (a) rationality of individuals, (b) existence of general belief systems for individuals, and (c) existence of aggregate or widely held belief systems that simplify political debate.

1. *Individual Rationality*

Issue constraint presumes a high degree of rationality. Specifically, at a minimum it is posited that individuals have consistent preferences among directly comparable alternatives, such as spending priorities on the budget and candidates running for a given office. A strong definition of rationality holds that if person prefers alternative A to alternative B and alternative B to alternative C, then he or she, in turn, prefers alternative A to alternative C. (For a more complete discussion of concepts of rationality, see Brady and Ansolabehere (1989).)

If individuals cannot order alternatives that are immediately comparable, they likely do not possess more general belief systems.

Do individuals reason consistently? Some observers deny this basic conjecture of rationality, citing the low correlations across issue items in national surveys and low correlations in responses to specific items over time (Converse 1964). Response instability, however, might reflect problems with question wording, rather than actual inconsistencies in people's opinions. (The literature on measurement problems and response instability is lengthy. Zaller (1992, Chap. 4) provides an excellent discussion of such problems).

A more appropriate approach to measuring and testing rationality uses a strong definition of consistency. A number of researchers have designed survey instruments that offer people pairs of choices, among multiple choices. For example, for the alternatives A, B, and C, the survey would ask people their preferences between A and B, between B and C, and between A and C. The answers to all possible combinations of pairs are, then, combined to examine the ability of people to sort the alternatives in a way consistent with some underlying structure of well-ordered preferences. In psychometrics this question-wording technique is known as paired comparisons. Studies of primary election candidates (Brady and Ansolabehere 1989), taxing and spending priorities (Hansen 1998), and national defense strategies (DeNardo 1995) show high degrees of individual rationality. Use of paired comparisons to study attitudes about public policies is relatively new, and much wider use of the approach is in order for measuring the consistency and structure of preferences.

2. *Individual Belief Systems*

The idea of issue constraint posits that individuals use general rules in their thinking about politics. They apply their own general conceptions of justice and the public good to each new issue they encounter. As a result, an individual's preferences on one topic constrain preferences on other topics, problems or policies. Do individuals use principles or rules, or logic that they apply across issues, such as standards of fairness, or do individuals reason about each issue differently?

There is considerable debate about whether individuals have belief systems, or instead have preferences that are specific to given social problems. Research on learning and on partisanship is broadly consistent with the notion that individuals have belief systems: individual people do interpret political communications through partisan or ideological lenses (Berleson et al. 1954, Zaller 1992). However, it is difficult to say whether such behavior reflects the use of principles in reasoning about politics or other causes, such as habitual behavior.

The theoretical difficulties lie in establishing the bases for general political beliefs and distinguishing them from factors that are specific to certain domains. Specifically, do people reason in a consistent way regardless of which interests are at stake?

The search for a few basic principles that people use has proved difficult. Various schools of thought posit different foundations for people's beliefs. Beliefs, for example, may be structured entirely around social institutions and attitudes of tolerance (Herbert McClosky (1964) goes so far as to tie these beliefs to personality types).

Beliefs may also be structured around economic interests, such as income or class (a classic expression of the foundation of belief systems in economic interests is Downs (1958)).

One particularly important conjecture derives from social contract theory. Every society embraces a set of principles on which consent is based. Political scientists in the middle of the twentieth century, for example, posited that Americans adhere to three principles—commitment to majority rule, to equality of opportunity, and to individual liberty (minimal government)—and applied these principles to specific policies. The give and take in politics is over the interpretation of the terms and how conflicts among these principles are resolved. (McCloskey and Zaller (1984) provide an excellent discussion of the basis for the American creed. For dissent about the basis of consent about fundamentals, see Prothro and Griggs (1960) and Dahl (1961 Chap. 28).)

Racial politics has served as the locus for the most intense debate over the notion that individuals have stable, general belief systems. Research by Donald Kinder and others suggests that individuals expressed preferences about public policies, such as jobs programs and school funding, depend on which racial groups those programs are perceived to benefit (Kinder and Sanders 1996). Paul Sniderman and his collaborators argue, contrary to Kinder and others, that people use general principles about fairness and the authority of government in thinking about politics that affect specific groups. Among other things, Sniderman and his colleagues demonstrate that racially intolerant 'liberals' can be prompted using political statements to support more tolerant positions, whereas racially intolerant 'conservatives' cannot (Sniderman and Carmines 1997). How much racial differences undercut the power of more general ideological belief systems remains a hotly debated question, with fundamental implications about whether people reason about public issues using general principles or rely on specific interests and stereotypes.

How survey researchers elicit beliefs when sensitive matters such as fairness and race are involved raises a troublesome set of methodological concerns. Sensitive matters may lead to biased answers. Individuals may wish to give the 'correct' response to a survey question because of social stigmas associated with certain attitudes. Looking across questions, people may exhibit more (or less) constraint than they truly have if some of the survey questions are corrupted by response biases. The challenge is to develop unobtrusive measures of preferences that will allow researchers to separate specific interests and stereotypes from general principles or beliefs. This remains a substantial methodological challenge.

3. *Aggregate Constraint*

Ideology, or more generally the notion of issue constraint, is useful to the extent that it allows people to sort out the political beliefs of those involved in politics in some simple way, say, from liberal to conservative. Shorthand assessments of politics are extremely common. The Americans for Democratic Action, the American Conservative Union, and many other interest groups publish annual ratings of legislators, which rank politicians from 0 to 100 according to legislators' consistency with the groups' ideology. Survey research organizations and interest groups publish similar classifications of voter 'types' based on clusters of answers to questions, ranging from abortion to zero-based budgeting. In fact, several public interest groups wishing to improve voter decisions offer simple on-line questionnaires that allow people to evaluate their own 'ideology' and compare it with the candidates' (e.g., GoVote.Com (www.govote.com)).

A central question for political scientists who analyze ideology is whether politics boils down to one dimension or multiple dimensions. The greatest degree of aggregate issue constraint is provided by positing one dimension; that is, everyone sorts out along a continuum from one extreme (say, liberal) to another (say, conservative). The least degree of aggregate constraint is when every issue differs so the number of dimensions over which people take differing political stances equals the number of issues on the public agenda. For social choice theory, larger numbers of dimensions create analytical difficulty and lead social scientists to the conclusion that institutions that constrain the political agenda and thus open political discourse are needed.

What does the dimensionality of politics look like? The most intensive study of the subject has been of roll-call votes in the US Congress. The US Congress typically takes over 1,000 roll-call votes each year covering hundreds of different policy concerns and issues. Using statistical scaling techniques, political scientists analyze the correlations across all roll call votes to determine how many different factors describe the behavior of legislators. (Two different techniques have been developed expressly for this purpose and are widely used by congressional scholars. Poole and Rosenthal (1997) use a nonlinear maximum likelihood technique, which assumes probabilistic voting and normal probability distributions for each vote. Heckman and Snyder (1997) assume probabilistic voting in a spatial model with uniform probabilities, which implies a linear factor analysis model with boundary constraints.) The number is very small,

possibly just one. One factor dominates congressional roll call voting, and it is most strongly distinguished by legislators' votes on important economic issues, such as taxation, Social Security, education, welfare, and the minimum wage.

This factor explains seven times more of the variation in roll-call votes than any other factor. Other, much less important dimensions, tap conflicts over civil liberties and international relations. In the 1950s 1960s, and 1970s, race served as a much stronger second dimension of conflict in American politics; it has since been subsumed by the first dimension. (For details on the dimensionality of Congress, see Poole and Rosenthal (1997) and Heckman and Snyder (1997).)

Studies of aggregate constraint among the public have been limited by the nature of the survey enterprise. The costs of interviewing large samples within legislative districts or asking detailed questions about dozens of policies are prohibitive. However, within the present limitations, only a handful of dimensions typically appear, with the main dimension tapping attitudes about the economic authority or size of government. A fundamental conjecture of representative politics is that elected officials are responsive to the citizenship. This general idea has been applied to questions of issue constraint. Do elected officials show the same sort of issue constraint on their beliefs that the public exhibits? This question has proved incredibly hard to study, owing to the enormous cost of doing extensive elite and mass surveys. Many studies of this question suggest that the degree of constraint differs between the elite and mass levels and that the degree of agreement between elites and masses varies across issues. In his classic article 'On the nature of mass belief systems,' Philip Converse (1964) documented significantly lower correlations across issues among the masses than among elites, and suggested that people with higher education levels exhibited higher degrees of constraint. It is unclear whether there is much association, then, between the belief systems of mass publics and the belief systems of elites.

Evidence of responsiveness or aggregate constraint comes, again, from the US Congress. A number of important studies have attempted to measure the responsiveness of elected officials to voters by associating roll-call voting behavior with mass attitudes. Political scientists have found strong association between district ideology, proxied by presidential vote, and roll-call voting behavior (e.g., Ansolabehere et al. 1999). Perhaps the most striking evidence of linkages between mass beliefs and elite beliefs and behavior comes by taking a broader historical perspective. Over time, the policies embraced by the courts, Congress, the President, and the electorate move strongly together. Across a wide range of concerns on the public agenda, as the populace moved to the left from the 1950s to the 1960s, so too did American institutions and policy, and as the populace moved to the right in the 1980s, so too did institutions and policy (Mayhew 1991). Without settling who—masses or elites—moved first, these broad historical patterns provide strong evidence of important issue constraint in American politics.

Despite empirical concerns about the degree of issue constraint in public thinking, ideology remains a central organizing principle for political science. The reason is parsimony. For observers of politics, ideology is a useful shorthand; for voters, who have to sort out complicated choices, it may be essential.

See also: Public Opinion: Political Aspects

Bibliography

Ansolabehere S, Snyder J M Jr, Stewart C III 1999 *Candidate Positioning in Congressional Elections. Working Paper on American Politics*. Department of Political Science, MIT, Cambridge, MA

Beer S 1978 The search for a new public philosophy. In: King A (ed.) *The New American Political System*. AEI, Washington, DC, pp. 5–44

Berleson B, Lazarsfeld P, McPhee W 1954 *Voting*. University of Chicago Press, Chicago

Brady H E, Ansolabehere S 1989 The nature of utility functions in mass publics. *American Political Science Review* **83**: 143–64

Converse P 1964 The nature of mass belief systems. In: Apter D (ed.) *Ideology and Discontent*. Free Press of Glencoe, London, pp. 206–61

Dahl R 1961 *Who Governs?* Yale University Press, New Haven, CT

DeNardo J 1995 *The Amateur Strategist*. Cambridge University Press, New York

Downs A 1957 *An Economic Theory of Democracy*. Harper, New York

Hansen J M 1998 Individuals, institutions, and public preferences over public finances. *American Political Science Review* **92**: 513–25

Heckman J, Snyder J M Jr 1997 Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *Rand Journal of Economics* **28**: S142–89

Kinder D, Sanders L 1996 *Divided by Color*. University of Chicago Press, Chicago

Krehbiel K 1998 *Pivotal Politics*. University of Chicago Press, Chicago

Lowi T 1979 *The End of Liberalism*, 2nd edn. Norton, New York

Mayhew D 1991 *Divided We Govern*. Yale University Press, New Haven, CT

McClosky H 1964 Consensus and ideology in American politics. *American Political Science Review* **58**: 361–82

McClosky H, Zaller J 1984 *The American Ethos*. Harvard University Press, Cambridge, MA

Poole K, Rosenthal H 1997 *Congress*. Oxford University Press, New York

Prothro J, Griggs C 1960 Fundamental principles of democracy: Bases of agreement and disagreement. *Journal of Politics* **22**: 276–94

Sniderman P, Carmines E 1997 *Reaching Beyond Race*. Harvard University Press, Cambridge, MA

Zaller J 1992 *The Nature and Origins of Mass Opinion.* Cambridge University Press, New York

S. Ansolabehere

Issue Evolution in Political Science

Issue evolution refers to a model of partisan change in which issue competition among political elites gradually leads to a transformation in mass party alignments along the issue dimension. There are two characteristics that distinguish this model of partisan change. The first is the prominent role played by elites. Unlike most models of elite–mass linkage political elites are not seen as merely responding to the demands of the mass electorate but as shaping and establishing the reputation of political parties on a political issue. The masses are seen as eventually reacting to these elite cues (MacDonald and Rabinowitz 1987, Zaller 1992). However, since it takes time for parties to establish a reputation on a political issue and for citizens to perceive and act on this difference between the parties, mass partisan change typically occurs gradually over a lengthy span of time. Ultimately, the transformation of mass party alignments occurs when the issue moves current voters to change their partisan identifications and biases the recruitment of new partisans.

1. *The Standard Model of Partisan Change: The Critical Election Perspective*

The standard model of partisan change was provided by V. O. Key in the 1950s (1955). His model of critical elections postulates that dramatic events can quickly transform the fortunes of political parties, leading to a realignment during the course of a single election. His account seems to fit almost perfectly the New Deal realignment. The Republicans, ascendent in national politics since the 1860s, were held responsible for the Great Depression and, as a result, millions of voters deserted the 'Grand Old Party' and became Roosevelt Democrats during the 1932 presidential election. The New Deal realignment thus seems to be a case of partisan change that was both dramatic and quickly realized, resulting in the rapid replacement of an old majority party by a new one.

1.1 *Criticisms of the Standard Model*

Key's model of critical election realignments, especially as amplified in Burnham's *Critical Elections and the Mainsprings of American* Politics (1970), became the standard framework for understanding changes in the US party system. Despite its popularity and influence, however, several major criticisms have been directed at the critical election realignment perspective. In the first place, it has been impossible to specify exactly what constitutes a critical election. This vagueness has led to claims of realignment for virtually every recent presidential election (Bass 1991). The generalizability of the theory has also been questioned since it is difficult to build a general theory based on a single historical case (Carmines and Stimson 1989, Silbey 1991). Finally, it has been argued that the critical election perspective overemphasizes short-term partisan change, while neglecting more gradual changes that extend over a long period of time (Carmines and Stimson 1989).

2. *The Issue Evolution Model*

Carmines and Stimson (1981, 1986, 1989) have developed an alternative model of partisan dynamics which they refer to as issue evolution. Unlike the classic realignment model, it presumes that most partisan change occurs gradually and incrementally over a lengthy span of time, unfolding over several elections and perhaps decades instead of a single election. At no single point in time is partisan change necessarily striking but its culmination through time can leave an indelible imprint on the party system.

Most issues, according to Carmines and Stimson, lack the capacity to generate significant partisan change. This is because most issues either reinforce the existing party alignment, never capture public attention, or have effects that are entirely short term. It is only those relatively rare issues that transcend these limitations that have the capacity to lead to fundamental and permanent change in the party system. In other words, an issue evolution only occurs when a new issue emerges that cuts across the existing party–issue alignment, captures the public's attention for an extended period of time, and cannot be resolved in the normal policy-making process.

Carmines and Stimson (1980) put particular emphasis on the amount of cognitive processing required to deal meaningfully with a political issue as a critical dividing line between those issues that can and cannot lead to issue evolutions. Issues that are complex, technical, and obscure are not good candidates for issue evolution. Instead, if issues are to transform mass party alignments, they must be able to be understood at a 'gut' level, engaging the raw emotions of the citizenry.

Two basic aspects of an issue evolution are its structure and sequence.

2.1 *The Structure of Issue Evolution*

There are two main actors in issue evolution: partisan elites and the mass electorate. Partisan elites include presidents, members of Congress, candidates for major political offices, as well as delegates to the nominating conventions. A party's position and reputation on a political issue is established through the public acts and pronouncements of their elites. When

a president sets out his position on an issue in the State of the Union Address or when senators express their opinions on an issue through their roll call votes, they are helping to build the position of their party on a given issue. When this process is repeated over time and in different political arenas, the party establishes a reputation on the issue.

A party's reputation on the issue, in turn, serves as a partisan cue to the mass public.

But why do elites of different parties adopt contrasting positions on at least some political issues and how does this process begin? Riker (1982) suggests a plausible answer. A political party, finding itself temporarily in the minority, always has an incentive to defeat the majority winning party and, according to Riker, political issues are a natural vehicle for this purpose. New issues—if they can split the coalition of the majority party and are sufficiently attractive to the electorate—can turn old losers into new winners (and old winners into new losers) (Carmines 1991). Thus, for Riker, losing minority parties always have an electoral incentive to adopt distinctive positions on new issues in their never-ending quest to upset the current status quo. Riker's logic can also be applied to minority factions within majority party coalitions. For example, the new left operating within the majority Democratic coalition appealed to anti-war forces and those on the left on the new moral and cultural issues of abortion and gay and women rights to help it take control of the party in 1972 (Layman 2001). The point is that it makes sense for losing and disadvantaged political elites to develop and advertise new issue proposals in an effort to improve their public standing and gain political leverage.

The other major group in the issue evolution process is the mass electorate. The political masses are not only less active in the political process than elites they also have no particular incentive or capacity to create new issue alternatives themselves. Their key role in issue evolution is to respond—or, more frequently, fail to respond—to the issue cues of party elites. Moreover, even when the citizenry does respond to elite issue cues, it does not do so instantaneously. Instead, the response is gradual, delayed, inert.

2.2 *The Sequence of Issue Evolution*

According to Carmines and Stimson's model of issue evolution, changes in elite partisan behavior do not lead directly to mass partisan response. Rather, two intervening steps are necessary to link elite policy shift to mass issue realignment. First, the mass electorate must perceive a difference in the issue positions of the parties. But awareness of party differences on an issue, by itself, is not likely to lead to mass issue realignment. Citizens must care about the issue. The changed perceptions of the parties on the issue must carry with them a strong dose of affection and disaffection for the parties if they are to overcome the stubborn inertia of existing partisan identifications (Campbell et al. 1960). Only when these two intervening conditions are met—clarified mass cognitive images of the parties and polarized affection toward them—will issue reorientation among partisan elites lead to changes in mass issue alignment.

Time ordering is critical. That is, the issue evolution process should follow a specific temporal sequence with changes in the cognitive and affective images of the parties occurring after elite reorientation on the issue but before mass partisan response.

3. *Illustrations*

What partisan changes have been illuminated by the issue evolution model? Three cases of issue/partisan change in contemporary US politics have been examined through the lens of issue evolution.

3.1 *Racial Issues*

Carmines and Stimson (1989) offer a detailed account of racial politics in mid-twentieth century America as an example of issue evolution. Historically, Republicans—the party of Lincoln—had a more progressive racial tradition than did the Democrats with their support first of slavery and later of segregation. But during the 1950s and 1960s their positions on race became reversed. The Democrats became the party of racial liberalism while the Republicans became the party of racial conservatism.

Carmines and Stimson show that this partisan transformation on race closely followed the model of issue evolution. In the first place, the change was gradual and incremental, unfolding over several decades instead of within a single critical election. Through an analysis of presidential speeches, party platforms, and congressional roll-call votes, they also demonstrated that partisan elites altered their racial positions years before mass partisans reversed their positions. There was strong evidence, in other words, that elites initiated the issue evolution on race. Finally, the racial issue case follows the sequence of changes postulated in the issue evolution model. The process began with reorientation among partisan elites and ended with mass issue realignment but bracketed between this beginning and ending was the electorate's increasingly accurate perceptions of the parties' racial positions and polarized affections based on them. All in all, the partisan transformation of racial issues fits the issue evolution model extremely well.

3.2 *Abortion*

Changes in the abortion issue during the 1980s and 1990s provide another example of an issue that has undergone a partisan evolution, according to a study

by Adams (1997). Examining roll-call votes in the House and Senate, Adams shows that Democrats and Republicans were only moderately divided over abortion throughout the 1970s. But, beginning in the early 1980s, the difference between these partisan elites grew steadily during the next 15 years. By the mid-1990s, congressional Democrats clearly had adopted the pro-choice position just as Republicans had become overwhelmingly pro-life. This change unfolded gradually and incrementally over a two-decade period, a pattern predicted by the issue evolution model.

Mass partisans have also become more polarized on abortion during this period, as Adams demonstrates through an analysis of public opinion polls. Throughout the 1970s, and up until the mid-1980s, Democrats, in the aggregate, were actually more pro-life than Republicans. But during the 1990s a partisan shift occurred, resulting in Democrats becoming significantly more pro-choice than Republicans.

Adams shows, finally, that there is likely a causal connection between the elite and mass changes, with elite level changes producing mass level responses. Not only did the shifts in abortion attitudes occur earlier among partisan elites than among the partisan masses but the degree of issue polarization is sharper and more pronounced in the former than the latter.

In sum, in every major regard, abortion is consistent with the issue evolution model. Issue polarization gradually develops among both partisan elites and masses and as the issue cues of partisan elites become clearer and more distinct over time the partisan masses respond more readily to them. The process culminates in a mass issue realignment.

3.3 The New Religious Divide and Moral and Cultural Issues

In *The Great Divide: Religious and Cultural Conflict in American Party Politics* (2001) Layman provides a comprehensive account of how the new traditionalist–modernist cleavage in American religion has come to shape the contemporary American political party system, the extent to which the religious and cultural divide between the Democratic and Republican parties has grown over time, and the process through which this growth has occurred. Similar to Carmines and Stimson's discussion of the types of issues capable of creating issue evolutions, Layman contends that the traditionalist–modernist religious cleavage came to affect party politics because it is associated with political issues that are easily understood, arouse strong emotions in a large number of people, reflect a deep-seated social conflict, cut across the lines of party conflict that existed in the 1970s, and appealed to the strategic incentives of certain groups of politicians within both parties.

He draws on the issue evolution framework to develop a model of the process through which polarized stands by party leaders on cultural and moral issues such as abortion, gay rights, and the role of women have been translated into changes in the religious composition of the parties' coalitions. Party activists have played a significant role in making this connection by sustaining and extending elite polarization and by providing cues to an inattentive public about the nature of issue change at the elite levels of party politics. The mass electorate gradually responded to elite- and activist-level change by first adjusting its perceptions of and affect toward the parties, and ultimately by reshaping its partisan ties along traditionalist–modernist religious lines.

Layman's research shows that the process of partisan religious change generally has followed the patterns predicted by the issue evolution model. It began among party elites and activists and activists also provided the link between elite-level issue change and the response by a relatively inattentive public. Over time, the mass public has come increasingly to see differences in the positions of the parties on cultural and moral issues and to care about these differences. These changes in mass perceptions of the parties and affect toward them have led ultimately to a substantial shift in the religious and cultural composition of the parties' coalitions.

See also: Change: Methods of Studying; Executive Branch, Government; Issue Constraint in Political Science; Law: Change and Evolution; Political Communication; Political Elites: Recruitment and Careers; Public Opinion: Political Aspects; Social Change: Types

Bibliography

Adams G D 1997 Abortion: Evidence of an issue evolution. *American Journal of Political Science* **41**: 718–37

Bass H F Jr 1991 A reader's guide and bibliography. In: Shaffer B E (ed.) *The End of Realignment?* University of Wisconsin Press, Madison, WI

Burnham W D 1970 *Critical Elections and the Mainsprings of American Politics*, 1st edn. Norton, New York

Campbell A, Converse P, Miller W, Stokes D 1960 *The American Voter*. Wiley, New York

Carmines E G 1991 The logic of party alignments. *Journal of Theoretical Politics* **3**: 65–80

Carmines E G, Stimson J A 1980 The two faces of issue voting. *American Political Science Review* **74**: 78–91

Carmines E G, Stimson J A 1981 Issue evolution, population replacement, and normal partisan change. *American Political Science Review* **75**: 107–18

Carmines E G, Stimson J A 1986 On the structure and sequence of issue evolution. *American Political Science Review* **80**: 901–20

Carmines E G, Stimson J A 1989 *Issue Evolution: Race and the Transformation of American Politics*. Princeton University Press, Princeton, NJ

Key V O Jr 1955 A theory of critical elections. *Journal of Politics* **17**: 3–18

Layman G 2001 *The Great Divide: Religions and Cultural Conflict in American Party Politics*. Columbia University Press, NY
MacDonald S E, Rabinowitz G 1987 The dynamics of structural realignment. *American Political Science Review* **81**: 775–96
Riker W H 1982 *Liberalism Against Populism*. W H Freeman, San Francisco
Silbey J H 1991 Beyond realignment and realignment theory: American political eras, 1789–1989. In: Shafer B E (ed.) *The End of Realignment?* University of Wisconsin Press, Madison, WI
Zaller J R 1992 *The Nature and Origins of Mass Opinion*. Cambridge University Press, Cambridge, UK

E. G. Carmines

Issue Networks: Iron Triangles, Subgovernments, and Policy Communities

Studies of issue networks and other forms of networks examine the relationships between interest groups and the state. As political scientists analyzed policy making, they found that simple models of representative government in which only elected officials and political parties took decisions and had a clear, coherent overview of policy across government were inadequate. Instead, interest groups were active participants in decision making and implementation, building up close relationships with elected and unelected officials. Policy making was fragmented, with differing participants and relationships in each segmented policy field.

1. *Early Development*

Initially, in the 1960s and 1970s, there were two distinct views of the relationships between interest groups and officials. 'Iron triangles' and 'policy communities' suggested that very closed 'subgovernments' of stable participants in long-term relationships effectively made policy in well-delineated policy fields such as agriculture or defense. In the USA, 'iron triangles' were composed of the interest groups active in a particular policy field, members of Congress (especially those sitting in the relevant committees) and officials (particularly unelected) in the relevant parts of the bureaucracy with powers in the field. These actors knew each other, were in constant interaction, and exchanged concessions, advantages, and information. They were a closed group, able to exclude outsiders. The concept of 'policy community' was used and developed in studies of Britain from the 1970s onwards (Jordan 1990). Policy formation was argued to be segmented into subsystems composed of policy communities of civil servants from government departments (or units within them) and interest groups. They were built on exchange relationships and a sense of 'community,' and possessed a set of dominant values (Richardson and Jordan 1979). Both iron triangles and policy communities paid close attention to interest groups with powerful and continuing material interests, the basis of exchange relationships.

The concept of 'issue networks' was developed by Hugh Heclo in opposition to the 'iron triangles' literature in the USA, which he argued monopolized attention on closure and subgovernments, but ignored forces for fragmentation and the development of new forms of interest groups (Heclo 1978). An issue network consists of a large number of issue-skilled 'policy activists' drawn from conventional interest groups and sections of the government, together with academia and certain professions but also comprising expert individuals regardless of formal training. Participants are constantly changing, and their degree of mutual commitment and interdependence varies, although any direct material interest is often secondary to emotional or intellectual commitment. No-one is in control of the policies and issues covered by a network and it is very difficult to identify clearly the dominant actors.

Neither policy communities nor issue networks offered a free-standing, developed model of policy making (Thatcher 1998). Rather, they represented enriching metaphors and useful concepts that were empirically applicable, and provided counterbalances to previous well-established views of policy making. They encouraged disaggregated analysis and drew attention to the effects of the fragmentation of government into subsystems, the interdependence of government and interest groups and increased specialization in policy making. However, issue network, iron triangle, and policy community literatures suffered from several limitations: they covered only specific categories of relations between groups and officials; they omitted key aspects of policy making, such as ideas, the distribution of power among actors, and change; and the factors explaining the emergence of different types of issue network or policy community remained general.

2. *Network Typologies*

In the face of these limitations, more general network approaches were developed. One path involved creating typologies of networks. The concept of 'policy network' or sometimes 'policy community' became generic, covering different forms of relationship between interest groups and officials arising from variations in network characteristics. Rhodes and Marsh (1992) offer a six-category typology covering a continuum from issue networks to policy communities and based on differences in the degree of integration, membership, and distribution of resources among members. Jordan and Schubert (1992) claim that different terms for state/group relations in fact refer to

variants of networks and hence they place terms such as iron triangles, clientelism, subgovernment, policy community, and various types of corporatism and pluralism in relation to each other using three 'dimensions,' namely number of members, whether the network is sectoral or transectoral, and the level of its stability.

In extending the scope of networks, the typologists pursue more ambitious theoretical aims. They seek to relate existing different terms and categories to each other within one overarching typology in order to avoid a jumble of overlapping and confusing terms. They claim to assist in ordering and categorizing information, permitting policy subsystems to be described, classified, and compared. Explanatory aims have ranged from specifying the key characteristics of policy networks in order to allow the systematic ordering of material, linking the 'structural' features of state and society and policy making (Atkinson and Coleman 1989), and offering a 'model' of the role of interest group intermediation based on different forms of power dependence embodied in the types of network (Rhodes and Marsh 1992).

Yet network typologies have suffered from considerable difficulties (Dowding 1995). Terminological disputes over whether the generic term should be policy network or policy community have created confusion. Placing differing forms of state-interest group relations under the heading of 'policy network' has not led to a common terminological currency. As the concept of policy network/community became generic, definitions have become loose and frequently contradictory. The empirical applicability of network typologies has encountered important obstacles. Several of the 'dimensions' that form the basis of categories are difficult to operationalize, or are unlikely to generate agreement between different researchers; examples include the 'degree of integration,' 'pre-eminence,' and 'serving the interests of producers' (Rhodes and Marsh 1992). Categories are poorly defined and identifying a type of network in practice is difficult, since categories are not mutually exclusive and different types of policy network or aspects of several network types can exist in the same policy area.

3. *Interorganizational Analyses*

An alternative generalizing strategy has been to develop interorganizational analyses of networks. Drawing on social network analysis and work on interorganizational relations, they view policy making as an example of decision making by interdependent actors who engage in exchange relations (Knoke et al. 1996). Policy is seen as the product of complex interactions among actors drawn from both 'public' and 'private' sectors. Such actors are partly interdependent but also partly autonomous. They are linked horizontally without being part of a single organizational hierarchy; their relations are based on exchange, thereby producing policy networks, but nevertheless combine elements of conflict with those of cooperation (Marin and Mayntz 1991). Analysis focuses on the interaction of actors—the unit of analysis consists of the relations among network participants.

Interorganizational network analyses 'map' network structures, using qualitative and quantitative methods to reveal linkages between actors. Frequently, mathematical techniques derived from formal network analysis are used to establish the network position of actors and relationships amongst them, applying indicators/measures such as intensity of communication, reputation among network participants, or resources. Mapping networks to provide a systematic picture of linkages serves several purposes. It draws attention to the interrelations and interdependence between public and private actors. Quantification enables complex patterns to be revealed that would be difficult to find because the events studied are too numerous, vast, and/or scattered and allows comparison of networks across sectors, countries, and time periods. The most important ambition of interorganizational approaches is to offer a 'structural analysis' by looking at patterns of network relations. It is insufficient merely to analyze the attributes of individual actors, as the influence of such attributes on behavior depends on the network of relations of each actor (Emirbayer and Goodwin 1994). Instead, relations amongst actors must be studied as a whole. A structural analysis of private and public actor configurations in policy formation identifies the position of actors and the ways in which networks influence actors' capacity to use their resources. It therefore permits explicit consideration of the ways in which different structures create different mechanisms or logics in policy making.

The scope of interorganizational approaches is very broad. Thus, for example, policy network refers to 'structures' located 'beyond or between policy markets and policy hierarchies' that are composed of 'relatively autonomous action units' (Marin and Mayntz 1991). From an 'organizational state' perspective, Knoke et al. (1996) have conceived policy networks as relations linking organized interests with public officials, transforming heterogeneous communities of policy actors into webs of common benefit-seeking actors. To operationalize and apply in empirical studies, important choices must be made to narrow the scope of interorganizational networks. Similarly, applying quantitative techniques requires the choices over matters such as questions posed, categories created, the relationships to be analyzed, and variables to be controlled for. Social network analysis may embody a particular 'theoretical orientation' towards the structure of the social world, notably the need to study behavior as a set of relationships, but it provides a particular set of concepts and methods and a frame-

work for testing arguments/models rather than a substantive set of hypotheses or a specific body of substantive theory.

In response to criticisms of important omissions in the policy network literature, current research has seen diversification (Thatcher 1998). This has taken two forms. First, additional explanatory factors are incorporated into network frameworks, often with attempts to analyze changes in networks. Thus, for instance, modifications of the institutional framework within which network actors operate, the technological and economic features of a policy problem, and the ideas, values, and knowledge of network actors have been examined. Second, in search of greater explanatory power, network concepts have been linked to other models of policy making (*Journal of Theoretical Politics* 1998). Thus, for instance, using rational choice theories, networks have been analyzed as institutions, notably in their capacity to establish rules and long-lasting patterns of social relations for the interactions of actors in the policy process. They are argued to reduce uncertainty and transactions costs by aiding the creation of relations of trust that allow exchange. Ideas-based models, notably the 'advocacy coalition framework', have provided a different partner for network analyses. The effects of networks on coordination of actors under conditions of different forms of beliefs have been considered. Hence networks offer an intervening variable for the effects of ideas.

4. Conclusion: Constant Expansion

Analyses of networks have become increasingly broad: from modest origins as concepts of issue networks and iron triangles, they have expanded to overarching frameworks of relations between public officials and private interests.

See also: Interest Groups; Issue Constraint in Political Science; Issue Evolution in Political Science; Networks and Linkages: Cultural Aspects; Networks: Social; Policy Knowledge: Advocacy Organizations; Policy Networks

Bibliography

Atkinson M M, Coleman W D 1989 Strong states and weak states. *British Journal of Political Science* **19**: 47–67

Dowding K 1995 Model or metaphor? A critical review of the policy network approach. *Political Studies* **43**: 136–58

Emirbayer M, Goodwin J 1994 Network analysis, culture, and the problem of agency. *American Journal of Sociology* **99**: 1411–54

Heclo H 1978 Issue networks and the executive establishment. In: Beer S H, King A (eds.) *The New American Political System*. American Enterprise Institute, Washington, DC, pp. 87–124

Jordan G 1990 Sub-governments, policy communities and networks: Refilling the old bottles? *Journal of Theoretical Politics* **2**: 319–38

Jordan G, Schubert K 1992 A preliminary ordering of policy network labels. *European Journal of Political Research* **21**: 7–27

Journal of Theoretical Politics 1998 Special issue on policy networks **10**

Knoke D, Pappi F U, Broadbent J, Tsujinaka Y 1996 *Comparing Policy Networks: Labour Politics in the U.S., Germany, and Japan*. Cambridge University Press, Cambridge, UK

Marin B, Mayntz R 1991 *Policy Networks: Empirical Evidence and Theoretical Considerations*. Campus Verlag, Frankfurt am Main, Germany

Richardson J J, Jordan A G 1979 *Governing Under Pressure: The Policy Process in a Post-Parliamentary Democracy*. Robertson, Oxford, UK

Rhodes R A W, Marsh D 1992 Policy networks in British politics: A critique of existing approaches. In: Marsh D, Rhodes R A W (eds.) *Policy Networks in British Government*. Clarendon Press, Oxford, UK, pp. 1–26

Thatcher M 1998 The development of policy network analyses. *Journal of Theoretical Politics* **10**: 389–416

M. Thatcher

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

J

Jackson, John Hughlings (1835–1911)

John Hughlings Jackson was born in Providence Green, Yorkshire, England on April 4 in 1835, of English antecedants on his father's side and Welsh on his mother's (Critchley M and Critchley E A 1998). His father was a gentleman farmer and a brewer who, as a surviving letter shows, was a wise and kindly man. Jackson's mother died aged 29 when he was one year old. He was educated at an undistinguished boarding school which he left at the age of 15, and did not go to university. In later years 'he held that by exemption from over-teaching his mind had retained more of freedom and energy than might otherwise have been the case' (Hutchinson 1925).

When Jackson left school he was at once apprenticed to a general practitioner in York. Two years later, in 1852, he enrolled at the York Medical School and in 1855 he proceeded to St. Bartholomew's Hospital in London. He qualified in 1856 and returned to the York Dispensary for three years as a resident medical officer. In 1860 he submitted a thesis (the title of which is unrecorded) to the University of St. Andrews and was awarded the MD.

By this time Jackson had returned to London. He had become increasingly interested in philosophy and seriously considered abandoning medicine for this field. He was dissuaded by his friend Jonathan Hutchinson (who later became one of the leading surgeons and ophthalmologists of his day). About this time (Hutchinson records) Jackson came under the influence of C E Brown-Sequard who 'told him strongly, that it was foolish to waste his efforts in wide observation of disease in general, and that if he wished to attain anything he must keep to the nervous system' (Hutchinson 1925). As a result, Jackson was appointed in 1862 to the staff of Brown-Sequard's hospital, the National Hospital for the Paralysed and Epileptic in Queen Square. In the same year he became a clinical assistant at the Royal London Ophthalmic Hospital (now known as the Morefields Eye Hospital) in London and in the following year he was appointed to the London Hospital. These three institutions provided him with the experience on the basis of which he made his fundamental contributions to our understanding of the organization of nervous function.

Jackson was a shy, modest, self-effacing individual, much loved by his students and colleagues. His wife died after only eleven years of marriage. A place at the lunch table was always set for him at the home of his colleague Dr. Thomas Buzzard. Jackson came to lunch irregularly and without notice. As Buzzard's son recalled, '… he would pull a chair up to the table, exchange a few remarks, perhaps tell a humorous story with only a twinkle of the eye disturbing the solemnity of his features, and after five, perhaps ten minutes he would mumble something about remembering an engagement and leave the room' (Buzzard 1934). In his last years Jackson lived an even more solitary, though serene life. He retained an interest in the lives of his pupils and their families to the end. He died on the October 7, 1911. His contribution was widely acknowledged by the academic world during his lifetime, but to the surprise of all he received no national honor.

Jackson's contemporaries and pupils recorded that he was a meticulous observer of clinical phenomena, and from his earliest days in medicine he was committed to the necessity for pathological verification of clinical inferences. His understanding of the nervous system based on this approach was molded by two main influences. The first derived from his teacher at medical school, Dr. Thomas Laycock (later Professor of Medicine in Edinburgh) who was interested in psychology and in nervous function and dysfunction. The other influence was the writings of the philosopher Herbert Spencer with whom Jackson corresponded up to the beginning of the twentieth century. It was from Spencer that Jackson derived the notion of the evolution of nervous function, which led to his idea of a hierarchical arrangement of functions, and that disease (neurological and psychiatric) led to a successive regression from the 'least automatic' and 'least deeply organised' to the 'most automatic' and 'most deeply organised' levels. As Walshe (1961) has pointed out, these ideas derived from Jackson's life-long study of five kinds of clinical phenomena: focal seizures, chorea, hemiplegia, speech disorders, and mental disorders.

A remarkable consequence of Jackson's work was his prediction, on the basis of unaided clinical observation, of a number of physiological principles which were later confirmed experimentally. One of them is the concept of localized cerebral discharges generating focal seizures, and that such focal seizures are caused by local brain disease. The comparison between the clinical symptomatology and subsequent neuropathology showed that the distribution of focal motor and sensory signs depended on the site of the

lesion. The topographical structure–function relationships laid the ground for the topological functional organization of the cortex long before cytoarchitectonic anatomical maps were established. The general idea of the functional specialization of the cortex had already been proposed by Gall (1825). His phrenology was derived from the attribution of certain personality features to the individual 'microstructure' of the scalp as inferred from palpation. Gall's postulate that cerebral function must be localized was correct in principle but based on inadequate evidence. The first real evidence came from Jackson, who published his first account of epilepsy in 1861, almost 10 years before his clinically based concept was confirmed by using electrical stimulation of the cortex (Fritsch and Hitzig 1870, Ferrier 1876).

The first pioneers of cortical neurophysiology indeed recognized the pivotal role of Jackson in the development of the concept of epileptogenic discharges in distinct areas generating focal seizures as the following two quotations show: '… I was precisely the first one, now some thirty or more years ago, to confirm by *physiological experiment* and to define more closely that which Hughlings Jackson had concluded from clinical facts' (Hitzig 1900); '… Dr. Hughlings Jackson who from a clinical and pathological standpoint anticipated many of the more important results of recent experimental investigation into the functions of the cerebral hemispheres… (Ferrier 1886).

Another concept put forward by Jackson related to the effect of lesions on function. In order to understand the pathophysiology of the effects of lesions he introduced the term 'negative and positive elements.' The negative elements were derived from the more global effects of epileptic seizures as apparent from dreamy states and the loss of consciousness. This was regarded as a dissolution 'of so many nervous elements of the highest centres' by the epileptic discharge, so that the 'most special and complex sensorimotor arrangements' would affect the adjustment of the organism as a whole to the environment. Convulsions were recognized as due to excessive neuronal discharge. From his clinical observations Jackson deduced that 'there are two ways in which nerve tissue suffers. It may be destroyed, and then there is loss of function. It may be unstable, and then there is disorder of function—'discharge' (Jackson 1931, Vol. 1, p. 29).

The 'positive element' was exemplified by speech disturbances, where the inability to express oneself demonstrates the negative effect of a brain lesion, whereas the paraphasia—the insertion of wrong words or inadequate sentence formations—would show the effect of the damaged part on the function of the remaining neural circuitry. These views later inspired Sigmund Freud's contributions to aphasia, in particular his early insight into the emotional determinants of speech disruptions. The positive and negative elements also had an impact on subsequent hypotheses on psychotic disorders.

Jackson's explanation of the clinical deficits associated with lesions was based on his view of the three levels in the evolution of the nervous system. Each element of each level was seen to contain a complete representation of the next lower level. Consequently, each element of the middle and highest levels would contain a representation of the entire body. If the nervous system is damaged the clinical impairment is not due just to the damaged part, but represents the compensatory action of the lower level circuitry that would then partially compensate for the function of the destroyed tissue.

The concept of evolution of the nervous system was therefore instrumental to his view that dissolution of the nervous system due to the effects of lesions results from the function of the remaining, nondamaged parts being liberated from control by the damaged part. This approach to understanding the pathophysiology and nature of the neurological signs and symptoms was later replaced by 'localizationists' who considered these to be deficits in a mechanical framework, i.e., saw them as loss of function of the destroyed tissue. It soon became clear that phenomena like spasticity, tremor etc., could not be explained in this way. With the advent of new, experimentally-based conceptions of the brain as a complex, densely interconnected and re-entrant network, the Jacksonian views were basically confirmed.

Jackson's contribution was not limited to the conceptual advances we have described. His skill in precise clinical observation, together with his analytical and deductive powers, led to advances of practical value in clinical neurology. For example, he was one of those who pioneered the introduction of the Helmholtz ophthalmoscope to medicine, his first paper on the topic appearing in 1863. His papers of 1871 and 1873 on 'optic neuritis' from intracranial disease, and the diagnosis of brain tumours, represented a major advance. He made contributions to neuro-ophthalmology, particularly in relation to eye movement, and to the effects of syphilis on the nervous system.

The writings of Jackson have a reputation for being difficult to read, being filled as they are with qualifications in the text and in extensive footnotes, as he strove to get as close as possible to the truth about inherently complex problems; they are still illuminating. Many appeared in *Brain* (of which he was a cofounder) but many others appeared in journals which are now difficult to obtain. Fortunately a comprehensive selection was published in 1931 and reprinted in 1958.

See also: Brain, Evolution of; Cerebral Cortex: Organization and Function; Cognitive Functions (Normal) and Neuropsychological Deficits, Models of; Epilepsy; Long-term Potentiation and Depression (Cortex);

Neural Circuits: Models of Emergent Functions; Neural Systems: Models of Behavioral Functions

Bibliography

Buzzard E F 1934 Hughlings Jackson and his influence on neurology. *Lancet* **2**: 909–13

Critchley M and Critchley E A 1998 *John Hughlings Jackson. Father of English Neurology*. Oxford University Press, New York, p. 228

Ferrier D 1876 *The Function of the Brain*. Smith, Elder, London

Ferrier D 1886 *The Function of the Brain*, 2nd edn. Putnam, New York, p. 498

Fritsch G T, Hitzig E 1870 Über die elektrische Erregbarkeit des Großhirns. *Archiv für Anatomie und Physiologie* 300–32 (Engl. Trans. Wilkins H 1963 Neurosurgical classics. *Journal Neurosurgery* **20**: 904–16)

Gall F J 1825 *Fonction du Cerveau*. Ballière, Paris

Hitzig E 1900 Hughlings Jackson and the cortical motor centres in the light of physiological research. *Brain* **23**: 545–81

Hutchinson J 1925 Recollections of a lifelong friendship. In: Taylor J (ed.) *Neurological Fragments*. Oxford University Press, Oxford, UK

Jackson J H 1925 *Neurological Fragments*. In: Taylor J (ed.) Oxford University Press, Oxford, UK

Jackson J H 1931 *Selected Writings of John Hughlings Jackson*. Hodder and Stoughton, London

Spurzheim G 1825 *Phrenology in Connection with the Study of Physiogonomy. Part I. Characters*. Treuttel, Wurtz and Richter, London

Walshe F M R 1961 Contributions of John Hughlings Jackson to neurology. *Archives of Neurology* **5**: 119–31

W. I. McDonald and H.-J. Freund

Jakobson, Roman (1896–1982)

Roman Jakobson was one of the major linguists, literary theorists, and semioticians of the twentieth century. He contributed to linguistics such concepts as feature, binary opposition, markedness, redundancy, and universals. He also showed the importance for linguistics of child language acquisition, aphasia, poetry, the act of communication (language usage), the meaning of grammatical categories, and the systematicity of language change. And he pointed out the relationship of linguistics to semiotics, cultural anthropology, literary studies, neurology, and so forth. (Most of Jakobson's work is reprinted in his *Selected Writings*. For a one-volume introduction to his work in linguistics, see his *On Language*. For more detail and further bibliography for this article, see Waugh and Monville-Burston 1990.)

1. Brief Overview of Jakobson's Life

Jakobson was born in Moscow on October 11, 1896 and died in Cambridge, Massachusetts on July 18, 1982. While still a student he helped to found the two groups known now as Russian Formalism. The Formalists called for an immanent analysis of literary works, with a focus on the properties that distinguish literary material from any other kind. In 1920, Jakobson went to Prague and in 1926 helped to found the Prague Linguistic Circle, the cradle of (functional) structuralism. For the Prague School, language serves for communication and thus its internal structure has to be studied from the standpoint of the tasks it performs. Jakobson first coined the term structuralism in 1929 as a cover term for this viewpoint and was a leading proponent of this highly influential twentieth century trend in linguistics and in many allied humanistic, social, and behavioral fields.

In 1939, Jakobson fled the Nazis and in 1941 went to the USA. He taught at the *Ecole Libre des Hautes Etudes*, New York, from 1942–46 and in 1943 he began teaching at Columbia University. In 1949, he obtained a professorial Chair at Harvard University. From 1957 on, he was also Institute Professor at MIT. He was President of the Linguistic Society of America in 1956 and received the International Prize for Philology and Linguistics in 1980 and the Hegel Prize in 1982.

2. Structural Phonology, Child Language, and Aphasia

During his Prague period, Jakobson and fellow Russian linguist Nikolai Trubetzkoy argued that phonological systems are structural wholes, which have to be studied in terms of their relational properties, and that language change has to be seen in terms of systems (see *Phonology*). They investigated types of phonological systems (see *Linguistic Typology*) and established the study of linguistic affinities between languages (*Sprachbunde*)—how languages that are geographically and culturally related may acquire traits in common (see Jakobson 1971a).

During this time Jakobson also turned his attention to the acquisition of phonology by children and its dissolution in aphasics. His monograph—*Child Language, Aphasia and Phonological Universals* (Jakobson 1941)—demonstrated that the order of acquisition follows the laws of typology: from universal to nonuniversal, unmarked to marked, simple (optimal) structures to more complex ones. He also showed that aphasic loss follows a reverse order.

3. Theory of Distinctive Features and Functional View of Sound

In his American period and with the aid of advances in acoustic phonetics, Jakobson developed his famous theory that phonemes such as /t/ and /d/ in *ten* and *den* and /b/ and /m/ in *bat* and *mat* are not the ultimate (smallest) constituents of language. Rather, the distinctive features are: the smaller properties of

sound by which phonemes are differentiated from each other. For example, /t/ is differentiated from /d/ as unvoiced (voiceless) vs. voiced in many languages, and /b/ from /m/ as oral (nonnasal) vs. nasal. Jakobson also defined a small set of such (acoustic-perceptual) features that underlie the phonological systems of all the languages of the world (Jakobson et al. 1952, Jakobson and Halle 1956, Jakobson and Waugh 1979). He also showed the relevance of these notions for grammatical meaning, which he decomposed into semantic features (see Jakobson 1984).

Influenced by communication (information) theory and cybernetics, Jakobson explored more deeply the function of language for communication. He insisted, in particular, that all facets of the speech sound have a function (Jakobson and Waugh 1979): while distinctive features differentiate between words which are different in meaning, redundant features serve to support and enhance the distinctive features (see *Phonology*). Other features may show the boundaries of words (configurative), inform about the attitude of the speaker (emotive), or indicate the age, sex, and social class of the speaker (physiognomic).

Since distinctive and redundant features have no positive content, but all other types of signs in language do (including words and sentences), there are two types of signs in language; and language and the genetic code are the only systems to have this kind of structure. This differentiation led to a new dichotomy for language (see Jakobson and Waugh 1979, Jakobson 1980): mediacy (elements with an indirect relation to meaning, such as distinctive features) vs. immediacy (direct relation between sound and meaning: phenomena such as sound symbolism, synesthesia, glossolalia, and mythic, poetic, and magical uses of sound. In many cases of mediacy, there is an iconic (nonarbitrary) relation between sound and meaning (see also Jakobson 1966). This dichotomy is also related to the hemispheres of the brain (left vs. right, respectively) in Jakobson (1980).

All of the feature types together represent the entire makeup of a sound and there is nothing that is pure sound (Jakobson and Waugh 1979). Consequently, the oppositions traditionally made between linguistic and nonlinguistic, abstract and concrete, functional and nonfunctional, (phon)emic and (phon)etic are inappropriate.

4. *Invariance and Relational Structure*

Invariance was, for Jakobson, the dominant topic and methodological device underlying his research: any sign is defined by invariant properties, which differentiate it from the other elements in the system of which it is a part. Across languages too, he also insisted on the importance of universal invariants (language universals), establishing the equivalence of diverse signs. In addition, he correlated invariance with contextual variation: any sign evidences variation as it enters specific contexts. Contextual variants are not equal, but hierarchized: some are more basic, others more marginal. Invariance and variation apply to the *signans* (e.g., variation across phonemes) and the *signatum* (e.g., the semantic properties of the case and verbal systems of Russian (see Jakobson 1984). In semantics, by incorporating invariance (context-independent meaning, general meaning) and variation (context-specific meaning, polysemy, reference), he gave the basis for a rigorous semantics.

Invariance for Jakobson was always invariance in relationship (equivalence); and the primary type of relation in language is that of binary opposition, phonological (e.g., nasal vs. oral), grammatical (e.g., singular vs. plural), or lexical (e.g., near vs. far), etc. (see Jakobson 1971a, 1971b, 1987, Jakobson and Waugh 1979). Such binarism is based on an asymmetry in the relationship between a marked (focused or weighted) and an unmarked (neutral) term, for example, marked nasal, plural, near vs. unmarked oral, singular, far). For Jakobson, markedness is a fundamental way by which we organize not just language, but cultural systems in general (see *Linguistics: Markedness*).

5. *Functions of Language*

For Jakobson, the *raison d'être* for language is communication and language is suited to various communicative goals, which in turn are correlated with the act of communication in which language is used. In his famous article, *Linguistics and Poetics*, he defined the six primary factors of any speech event: (a) speaker, (b) addressee, (c) context (including the thing referred to), (d) message (*parole*), (e) code (*langue*), and (f) contact (medium) of communication. In conjunction with these, he then defined the functions of language in terms of a focus in the message on one of the factors: (a) emotive function (focus on the speaker), for example, intonation showing anger; (b) conative function (focus on the addressee), for example, imperatives and vocatives; (c) referential function (focus on the context), for example, talk about the real world; (d) poetic function (focus on the message), for example, poetry; (e) metalingual function (focus on the code), for example, definitions of words; and (f) phatic function (focus on the contact), for example, 'hello, do you hear me?' These functions may be dominant in a message or may be subsidiary, for example, a referential message may carry expressive information about the opinion of the speaker.

6. *Metaphor and Metonymy: Similarity/Contiguity*

In 1956, Jakobson published his famous *Two Aspects of Language* where he analyzed the relation between communicative processes and properties of linguistic

structure. First, he distinguished the two fundamental operations: selection (substitution) for encoding (production) and combination for decoding (comprehension). Then, he contrasted two types of relations: similarity (all types of equivalence) and contiguity (temporal and spatial neighborhood), and said that they underlie the two operations. Put simply, the elements in a selection set are normally associated by similarity, and those in combination by contiguity. Since in aphasia, the normal patterning of speech is disturbed, Jakobson proposed that these factors be used as a way of categorizing types of aphasic disturbances (see *Aphasia*).

For Jakobson similarity/contiguity is a fundamental polarity of language, texts, culture, and human thought (cognition) in general. He used it for much of his work in semiotics of the cinema and music. He also used it to refine his definition of the poetic function (1960): in poetry, where focus on the message is dominant, equivalence (similarity) relations are used to build the combination rather than only to underpin the elements of the selection set. Thus, rhyme, alliteration, parallelisms, metaphor, etc. help to structure the poetic text, whereas in prose contiguity is the essential constructional principle.

7. Grammar

Jakobson also focused on the function of grammatical categories. His famous article on *Shifters* (1957) focuses on those elements whose general meaning in the code can only be specified by taking into account their use in messages, because the codal meaning includes information about particular elements of the speech event. For example, deictic pronouns designate speaker (I) and addressee (you). Thus, language encodes pragmatic factors of the context of utterance and linguistics necessarily includes syntax, semantics, and pragmatics (see *Deixis*; *Pragmatics: Linguistic*).

Grammatical categories (both morphological and syntactic), for Jakobson, are defined as those which are obligatorily present in the construction of acceptable messages (1959a, 1959b), whereas particular lexical categories (e.g., words referring to space) are optional. Through this view of grammar, Jakobson provided a semantic and operational approach to the relation between language and cognition: grammatical categorizations provide the necessary patterns of thought.

Jakobson was also inspired by Charles Sanders Peirce's notion that the essence of a sign is its interpretation, that is, its translation by some further sign. Henceforth, he defined the *signatum* as that which is interpretable or, better, translatable (Jakobson 1959b) by a potentially unlimited series of signs (see *Interpretation and Translation: Philosophical Aspects*). He characterized the Peircian approach as the only sound basis for a strictly linguistic semantics.

8. Jakobson's Legacy in Linguistics and Allied Disciplines

Jakobson's influence on twentieth century thought was considerable and many of his contributions have become a permanent part of American and European views of language. He has had such a towering role in linguistics that his work has defined the field itself (see *Linguistics: Overview*). Many of Jakobson's concepts and discoveries are now so deeply ingrained in the theoretical basis of modern linguistics that they are thought to be commonplace or self-evident. This discussion will consider, in particular, his influence on linguistics in the last quarter-century and later offer some brief notes on other disciplines. (For more details of Jakobson's influence, see Waugh and Monville-Burston 1990.)

8.1 Influence on Linguistics

Jakobson's view of language as a relational whole and a communicative tool with many functions opened the door to functionalist approaches more generally and abetted the development of linguistic studies that look at discourses and texts and consider their wider linguistic and 'real-world' environment (see *Functionalism in Anthropology*). His work on *Shifters* showed that grammar, semantics, and pragmatics are but different sides of the same phenomenon and aided present-day interest in pragmatics (see *Pragmatics: Linguistic*). Work in pragmatics has led linguists to explore the boundaries of their discipline with the neighboring fields of anthropology, sociology, mythology, philology, and philosophy in a truly Jakobsonian interdisciplinary spirit. His writings on grammatical meaning have inspired others to investigate the interdependence of invariance and variation, the semantics of case systems, and the interrelation of tense and aspect with syntax, discourse and the lexicon (see *Semantics*).

Jakobson's insistence that language is a complex and dynamic whole with a variety of subcodes (speech styles) helped to pave the way for the advent of sociolinguistics (see *Sociolinguistics*). The ethnography of communication has also taken its inspiration from Jakobson's multifunctional perspective on communication (see *Ethnography*). His work on the functions of language and their relation to the speech (communication) act remains central in the investigation of speech in interaction. Anthropological linguists—who adopt a functional and pragmatic perspective with an interest in folklore, poetics (oral traditions), and discourse analysis—draw on and develop the ideas of the Prague School and of Jakobson in particular (see *Anthropology*).

Jakobson stressed the need to search for general laws governing all linguistic systems and thus helped to foster the current interest in generalizations. He saw a definite equivalence between the search for ex-

planatory adequacy in generative grammar and the elicitation and interpretation of the entire network aimed at by Praguian structuralism (see *Generative Grammar*). His monograph on child language and aphasia, generally regarded as one of his most original and influential works, has been signaled as the point of departure of a new era in linguistics and has been lauded by many because it showed the relation between child language and phonology, launched psycho- and neurolinguistics, and ushered in a new focus on universal and law-governed, general (explanatory) properties of language (see *Psycholinguistics: Overview*).

The widespread movement that presently investigates universal properties of language, both universal grammar and language universals, is deeply indebted to Jakobson. When the linguist Noam Chomsky came in contact with Jakobson and worked with Jakobson's student and coauthor Morris Halle, he fully realized the explanatory power of universals and their relation to language acquisition: thereafter, he set out to determine universal constraints on the form of grammars. The other stream of research in universals, practiced by Joseph Greenberg and his associates, examines concrete materials from a great variety of languages in order to establish typological classifications and discover the general laws that underlie the structure of human language.

Many scholars agree with Jakobson that typology is important for the study of language change, areal linguistics, and historical reconstruction (see *Linguistic Typology*). His work in historical and geographical linguistics continues to be reprinted in readers in these areas. The seminal idea that changes must always be treated in view of the system that undergoes them is the essential premise for textbooks in historical linguistics. And Jakobson's view of the teleological basis for language change has been instrumental in inspiring further work (see *Historical Linguistics: Overview*).

By revealing the close relationship between developmental studies and vital questions of linguistic theory, Jakobson's *Child Language* also launched psycholinguistics. Many psycholinguists now recognize that the broad outlines of this bold synthesis continue to be confirmed. Following Jakobson's example, scholars interested in language pathology have studied aphasia from a linguistic point of view and, for many, Jakobson's distinction between similarity disorders and contiguity disorders is regarded as fundamental and firmly supported by contemporary ideas of the functional organization of the human brain (see *Psycholinguistics: Overview*; *Aphasia*). Jakobson was the first modern linguist to investigate in depth the concurrent constituents of linguistic elements. Many linguists have claimed that the distinctive feature was Jakobson's greatest insight and, after the phoneme, the most significant step forward in modern phonology. Generative phonology recognizes Jakobson as one of its founding fathers, in particular through Halle. Markedness and the correlated notion of binary opposition are now widely used in linguistics and other disciplines. Componential analysis—analyzing an item into smaller properties and representing them as a combination—was generalized as a methodological model for semantics and lexical representation and has been adopted in other approaches as a means of describing the structure of a variety of semantic fields (see *Phonology*).

8.2 *Influence on Other Fields*

Jakobson's influence on contemporary thought has also spread to disciplines such as literature, poetics, semiotics, social anthropology, mythology, and folklore. His theoretical papers *Linguistics and Poetics* and *Poetry of Grammar and Grammar of Poetry* and his analytical studies of poems by Baudelaire and Shakespeare established him as an important theoretician of literature. He is the source of not only structuralist but other contemporary trends in literary theory. He also helped to develop semiotics as a discipline in its own right and has been called the major catalyst in the contemporary semiotic reaction (see Jakobson 1973). And he is credited with introducing the American semiotician Charles Sanders Peirce to Americans and with showing them the importance of Peirce's insights. His work in semiotics has also had repercussions in fields as diverse as economics, mathematics, especially catastrophe theory, sociology, and philosophy (see *Semiotics*).

Jakobson's influence on anthropology is considerable (see *Anthropology*). In American anthropology, for example, three convergent Jakobsonian ideas have proven to be particularly fruitful: the structure of the poetic text, the closeness of ties between poetry and mythology, and the relationship between discourse and cultural context. His greatest impact, however, has been on structural anthropology, especially through Claude Lévi-Strauss, who considers himself a disciple of Jakobson and has repeatedly acknowledged his indebtedness to him. He used many Jakobsonian ideas in his work on kinship systems, cooking, folk tales, myths, and primitive thought.

With the breadth and depth of his knowledge, his brilliant imagination, his prodigious output, the originality of his accomplishments, and his international influence on a variety of fields, Roman Jakobson has been recognized worldwide as one of the major creative minds of the twentieth century.

Bibliography

Jakobson R 1941 *Kindersprache, Aphasie und allgemeine Lautgesetze*. Almqvist and Wiksells, Uppsalsa, Sweden [1968 *Child Language, Aphasia, and Phonological Universals*. Mouton, The Hague and Paris]

Jakobson R 1956 Two aspects of language and two types of aphasic disturbances. In: Jakobson R, Halle H (eds.) *Fundamentals of Language*. Mouton, The Hague
Jakobson R 1957 Shifters, verbal categories, and the Russian verb. In: Jakobson R (ed.) *Selected Writings II: Word and Language*. Mouton, The Hague, pp. 130–47
Jakobson R 1959a On linguistic aspects of translation. In: Brower R (ed.) *On Translation*. Harvard University Press, Cambridge, MA (1966 2nd edn. Oxford University Press, New York)
Jakobson R 1959b Boas' view of grammatical meaning. In: Goldschmidt W (ed.) *The Anthropology of Franz Boas. Essays on the Centennial of His Birth*.
Jakobson R 1960 Linguistics and poetics. In: Sebeok T (ed.) *Style in Language*. MIT, Cambridge, MA
Jakobson R 1966 Quest for the essence of language. *Diogenes* **51**: 21–37
Jakobson R 1971a *Selected Writings I: Phonological Studies*, 2nd edn. Mouton, The Hague and Paris
Jakobson R 1971b *Selected Writings II: Word and Language*. Mouton, The Hague and Paris
Jakobson R 1973 *Main Trends in the Science of Language*. George Allen and Unwin, London
Jakobson R 1980 *Brain and Language: Cerebral Hemispheres and Linguistic Structure in Mutual Light*. Slavica, Columbus, OH
Jakobson R 1981 *Selected Writings III: Poetry of Grammar and Grammar of Poetry*. Mouton, The Hague, Paris, and New York
Jakobson R 1984 *Russian and Slavic Grammar: Studies, 1931–1981*. Mouton, Berlin and New York
Jakobson R 1987 *Selected Writings VIII: Major Works, 1976–1980*. Mouton, Berlin
Jakobson R 1990 *On Language*. Harvard University Press, Cambridge, MA
Jakobson R, Fant C G M, Halle M 1952 *Preliminaries to Speech Analysis*. MIT, Cambridge, MA
Jakobson R, Halle M 1956 *Fundamentals of Language*. Mouton, The Hague
Jakobson R, Jones L G 1970 *Shakespeare's Verbal Art in 'Th'Expence of Spirit*. Mouton, The Hague and Paris
Jakobson R, Lévi-Strauss C 1962 Les chats de Charles Baudelaire. *L'Homme* **2**: 5–21
Jakobson R, Waugh L R 1979 *The Sound Shape of Language*. Indiana University Press, Bloomington, IN
Waugh L R, Monville-Burston M 1990 Introduction: The life, work and influence of Roman Jakobson. In: Jakobson R (ed.) *On Language*. Harvard University Press, Cambridge, MA

L. R. Waugh

James, William (1842–1910)

William James, American philosopher and psychologist, remains a vivid presence in psychology because of his two classics, *Principles of Psychology* (1890/1981) and *Varieties of Religious Experience* (1902/1985). His work at once embodies the culmination of nineteenth century psychology, in his critical review of the older philosophical psychology and of the new experimental psychology, and the launching of twentieth century psychology, especially the functionalist tradition. He is remembered in philosophy as one of the last philosophers in the grand style who also sought to communicate his insights to a broad audience, an audience attracted by the informal felicity of his writing, public lecturing, and teaching. Distinctively American in his promotion of a pragmatism that viewed abstract concepts critically in terms of their 'cash value,' he participated actively in the European intellectual scene and like his eminent novelist brother Henry James advanced the linkage of the American and European intellectual and cultural worlds.

His enduring significance at the end of the twentieth Century is reflected in the publication of definitive scholarly editions of his complete writings in 17 volumes (1975b–88) and of his correspondence (1992) of which 12 volumes were planned. As members of an unusually creative family, he and his brother Henry and sister Alice have been the subject of extensive biographical attention (e.g., Lewis 1991, Myers 1986, Perry R. B. 1935).

1. Family Background and Early Life

The James family was economically privileged. William's father, Henry James Sr. (1811–82), who had inherited a comfortable living, was a restless, questing intellectual who found inspiration in the writings of the Swedish mystic Emanuel Swedenborg. He attracted visits from the New England transcendentalists Emerson and Thoreau while William was an infant in New York City, where he was born January 11, 1842. William's education was irregular, including periods of schooling or tutoring in England, France, Germany, and Switzerland. The intense moral and intellectual climate of the family seemingly entailed severe costs in physical and mental health for William and his gifted siblings. Teasing out the sources of creativity and of pathology in the James family has challenged its biographers.

When William was 18, he embarked on the study of painting in the Newport studio of William Hunt, a prominent American artist. But he gave up painting after about 6 months, turning next to science. He pursued studies in chemistry and comparative anatomy at Harvard, and, in a smaller shift in direction, entered Harvard Medical School when he was 22. These were years of increasing doubt, depression, and psychosomatic discomfort. James interrupted his medical training for a year to take part in an expedition to the Amazon with the Harvard biologist Louis Agassiz, and interrupted it once more to spend a year and a half in Germany taking therapeutic baths and reading widely in German literature, philosophy, and science. Returning to complete his M.D. degree in

1869, he spent several years of mental distress and near collapse living with his parents.

It was during this period that, by his own account, he drew on philosophy in a turn toward recovery, as he recorded in a subsequently famous passage in his diary of April 30, 1870:

> I think that yesterday was a crisis in my life. I finished the first part of (Charles) Renouvier's second *Essais* and see no reason why his definition of free will—'the sustaining of a thought because I chose to when I might have other thoughts'—need be the definition of an illusion. At any rate, I will assume, for the present—until next year—that it is no illusion. My first act of free will shall be to believe in free will (Perry, 1935, vol.1, p. 323).

This passage heralds the theme of voluntarism that pervades his later writings, as well as suggesting roots of his pragmatic philosophy.

James's productive life began in 1873 when he was appointed instructor in anatomy and physiology at Harvard. By 1875 he was teaching his first course in psychology, in connection with which he established a teaching laboratory that, as American historians of psychology like to note, preceded Wilhelm Wundt's laboratory at Leipzig in 1879, a date often enshrined as the founding of scientific psychology. By 1878, when he was 36, he had contracted to write the textbook on psychology that was not to see publication until 1890. It was also in 1878 that he married Alice Gibbens, a teacher who shared his intellectual and spiritual interests and had spent 5 years in Heidelberg with her widowed mother. This event reflected the stabilization of his personal life and also contributed greatly to it. In the following year he offered his first course in philosophy. He was effectively launched in his dual career as psychologist and philosopher at Harvard where he remained for the rest of his life. He died on August 26, 1910, at his family retreat at Chocorua, New Hampshire.

2. *Major Themes*

Three themes pervade James's contributions to psychology and philosophy. One is his commitment to voluntarism, the primacy of human agency, which led him to reject mechanistic doctrines. A second is his continual struggle to find a place for religious spirituality as exemplified by his father in the world revealed by Darwinian science. The third, an essential ingredient in his handling of the first two, is his commitment to pluralism and openness in theorizing, as shown in his distaste for abstract systemization and in his opposition to what he saw as the absolutism of both scientific materialism and Hegelian idealism. These themes characterize James's thought throughout his career as psychologist and philosopher, but in the spirit of the third theme, paradoxical inconsistencies recur in his psychology and philosophy.

3. *James's Psychology*

In his great book, *The Principles of Psychology* (1890/1981), James presented an account of mental life in the positivistic spirit of natural science. That meant for him that speculative entities beyond experience like the soul or psychological faculties were to be avoided. He took a commonsense dualism of mind and body for granted, though he was explicit that fundamental metaphysical issues would have to be faced eventually. The principal psychological method for him was introspection, by which James mostly meant the informal and avowedly fallible self-observation that he did so well, not the disciplined and quantified introspection of the new German experimental laboratories, which he respected while noting wryly that the method 'could hardly have arisen in a country whose natives could be *bored*.' (p. 192).

3.1 *James's Functionalism*

To apply a term that emerged with his immediate American successors, James was presenting a functional psychology, in the senses subsequently intended by the Chicago functionalists. His focus was on mental processes (functioning), not on the analysis of mental contents into constituent sensory elements. Further, he emphasized the adaptive functions of psychological processes, in line with evolutionary theory. Consciousness, he proposed, emerged to organize and direct neural processes in the brain that had become too complex to be guided by a mechanistic concatenation of reflexes. His vision of scientific psychology was committed to a mind–body interactionism troublesome to both scientific materialists and philosophers.

James viewed the nervous system and the brain as organs of mind, and devoted early chapters of the *Principles* to the functions of the brain. There follows a chapter on habit, foreshadowing the later preoccupations of J. B. Watson's behaviorism. James's account of habit was primarily physiological. Currents in the cerebral cortex originating in the sense organs 'leave their traces in the paths which they take. The only thing they can do … is to deepen old paths or to make new ones. … (A) simple habit … is, mechanically, nothing but a reflex discharge; and its anatomical substratum must be a path in the system' (p. 112). From this physiological base James proceeded with recommendations for the cultivation of good habits, putting neural mechanism to voluntary use for moral purposes, a characteristically Jamesian manoeuver.

3.2 *The Stream of Thought*

If James's treatment of the brain and of habit pointed toward later developments in neuropsychology and behaviorism, his classic treatment of the stream of

thought provided a sensitive and persuasive alternative to the prevalent approaches of introspective psychology that sought to decompose ideas and images into sensory elements. For James, the proper starting point in psychology is not sensations or feelings, as in prevalent elementaristic psychologies, but the sheer fact that thought goes on. 'Every thought tends to be part of a personal consciousness,' he wrote (p. 220). Thought seems continuous but is always changing. It deals actively and selectively with objects independent of itself. The idea that thought is actively selective underlies James's later discussion of attention, which in turn provides the rationale for his treatment of will, in terms of holding a thought of action in focus long enough for the action to occur.

No two states of consciousness are ever the same. 'A permanently existing "idea"… which makes its appearance before the footlights of consciousness at periodical intervals is as mythological an entity as the Jack of Spades' (p. 230). Relationships and transitional states are given in the primary data of introspection just as directly as the more substantive states on which consciousness may momentarily 'perch.' James's views here anticipated those of the Würzburg school regarding imageless thought, and foreshadowed Gestalt psychology in their rejection of sensory elements. James's evocative term, 'the stream of consciousness,' found a place in the context of world literature.

3.3 The Consciousness of Self

Since the stream of thought is always a personal consciousness in which the objects of awareness are divided into 'me' and 'not me,' James followed his treatment of the stream of thought with a chapter on 'The Consciousness of Self' that was very influential on later scholarship. If 'I' is taken to label the person as agent and introspective observer, 'me' is the self as object of introspective observation. For the purposes of his psychology, the 'I' is simply the passing thought in the personal stream, which appropriates the person's memories of the past and anticipations of the future. The 'me' comprises a hierachy of three selves. First is the material self—one's body, clothes, and the material possessions that one regards as 'mine.' The social self, or more properly, selves, consists of the views of oneself held by others that one cares about, as one understands them. The spiritual self is 'a man's inner or subjective being, his psychic faculties and dispositions, taken concretely… the most enduring and intimate part of the self' (p. 283). These three aspects of 'me' are the focus of motives of self-seeking and self-preservation, and of self-feelings of self-complacency and self-dissatisfaction. James provided a famous 'formula' for self-esteem as the ratio of Success to Pretensions, anticipating much later research and theory about Level of Aspiration.

The distinctively human attribute of reflective self-awareness highlighted in James's treatment of selfhood was subsequently developed mainly by social psychologists in the sociological tradition, especially C. H. Cooley and G. H. Mead, but returned as a central concern of psychologist students of personality and social behavior in the latter half of the twentieth century.

3.4 Emotion

Of all the special psychological topics that James discussed, his theory of emotion evoked more notice and controversy than any other. It was known as the James–Lange theory because in the *Principles* James linked his ideas as previously published in the 1880s with similar views of the Danish physiologist Carl Georg Lange published at about the same time. James's version held that the stronger or 'coarser' emotions of grief, fear, rage and love arise from feeling bodily changes that are aroused directly by an 'exciting fact.' In contrast with folk belief, we do not weep because we feel sad; we feel sad because we are weeping. We feel fear because we are set to run away. We feel because we act.

During the first half of the twentieth century, the James–Lange theory did not fare well. In one important critique, the physiologist Walter Cannon showed in animal research that centrally mediated emergency emotional responses do not depend on feedback from the viscera, a finding that bore more critically on Lange's version of the theory than on James's. More importantly, in a period dominated by behaviorism, James's central interest in emotion as conscious feeling seemed scientifically irrelevant. With the subsequent decline of behaviorism, each of several competing treatments of emotion looked back to James as anticipating their conclusions. Paul Ekman and Carroll Izard focused on the role of sensory feedback from the facial musculature in emotional expression as crucial to the distinctiveness of felt emotions, whereas Stanley Schachter launched a cognitively focused line of research emphasizing the interpretive context in which arousing stimulation is encountered. As cognitive perspectives on human psychology developed one-sidedly, James's emphasis on the centrality of feeling became more attractive.

3.5 Other Aspects

As embodied in the *Principles*. James's psychology dealt with many other topics, some familiar in later textbooks, others less so. At first encounter, James's account of the will seems foreign to present students of psychology, who are accustomed instead to non-consensual chapters on motivation. For James, voluntary movements are a secondary development from

the primary organic functions involved in reflex action, instincts, and emotional movements (p. 1099). The link between thought and voluntary action is provided by ideo-motor action, the concept that thoughts of a movement lead directly to making the movement, unless contrary inhibiting thoughts intervene. With the resurgence of cognitive and neuropsychology, James's focus on the selectivity of attention and on the relation between thought and action returned as central to the discipline. His century-old formulations are no longer useful, but his identification of significant problems for psychology earns renewed respect.

Other topics broached in the *Principles*, particularly hypnosis and divided consciousness, anticipated concerns beyond the boundaries of positivistic science that continued to preoccupy James during his mainly philosophic period that followed its publication (see Taylor 1983). James was much interested in the work of the French psychopathologists Charcot, Janet, and Liébeault, who preceded psychoanalysis in giving attention to nonrational processes outside the scope of normal consciousness.

4. *James as a Philosopher*

In spite of James's expressed intent in the *Principles* to avoid metaphysics, the book was infused with his philosophic thinking, particularly in his critique of earlier philosophic psychology, and contains the germ of most of the philosophic positions on which he elaborated subsequently. His specific impact on philosophy is identified with pragmatism and radical empiricism. Pragmatism, launched by James, became a vigorous philosophic movement which enjoyed a renaissance late in the twentieth century. Radical empiricism was James's attempt at the fundamental metaphysical construction that he postponed in the *Principles*. The work was incomplete at the time of his death, and the book presenting it (James 1912/1976) was assembled posthumously from journal articles.

4.1 *Pragmatism*

When James published *Pragmatism* (1907/1975a), a popular book that was much reprinted and translated, he drew on his earlier writing and thinking over three decades. Here, he offered the pragmatic method as a way of settling metaphysical disputes that might otherwise be interminable, such as those between tenderminded religious and toughminded materialistic world views: 'The pragmatic method…is to try to interpret each notion by tracing its respective practical consequences. What difference would it make practically to anyone if this notion rather than that notion were true?' (1907/1975a, p. 28) The pragmatic method entails a perspective on meaning. As he writes, 'to develop a thought's meaning, we need only determine what conduct it is fitted to produce: that conduct is for us its sole significance' (p. 29). And with pragmatic meaning comes a pragmatic conception of truth, as characterizing beliefs that receive confirmation when acted upon. 'The truth of an idea is not a stagnant property inherent in it. Truth *happens* to an idea. It *becomes* true, is *made* true by events. Its verity *is* in fact an event, a process…' (p. 97). With characteristic generosity, James attributed his initial ideas about pragmatism to his friend the semiotic philosopher C. S. Peirce, who did not agree with James's developed approach.

The extensive philosophical controversy about pragmatism in the first decade of the twentieth century appears to reflect considerable mutual misunderstanding on the part of both James and his critics, as well as some problems with his mostly informal exposition (see Myers 1986). He can be interpreted as treating claims of pragmatic truth as similar to the provisional truth claims of empirical science, in the spirit of Peirce and, later, John Dewey, or alternatively, as rationalizing wishful thinking (religious faith has good personal and social results, so it may be accepted legitimately as true by its believers). The idea that philosophical abstractions and assertions should be judged in terms of their practical consequences came to seem typically American, to both his supporters and his critics.

Late in the twentieth century, interest in pragmatism revived, with similar diversity of interpretation. One strand of thought prominently represented by the philosopher Rorty (1979) became part of the 'postmodernist' attack on the scientific world view to which some disaffected psychologists subscribed. Knowledge, in this version, is a human construction for human purposes and cannot be taken as a 'mirror of nature,' a view with roots in James. Some other humanistically inclined applied psychologists called for a pragmatic psychology focused on useful contextually limited truths rather than general laws. This version of pragmatism retains a commitment to evidential empiricism like that of the early James.

4.2 *Radical Empiricism*

As James entered his final decade increasingly troubled by his failing heart, he turned to the fundamental metaphysical issues that he had tabled when he wrote the *Principles*. In pragmatism, he had crystalized a theory of knowledge developed over his full career. Radical empiricism, in contrast, proposed an ontology that continued his attack on Hegelian absolute idealism but broke sharply with the commonsense mind–body dualism of his scientific psychology. To transcend this Cartesian dualism, and the cognitive dualism of thought and its object or consciousness and its contents, he now proposed pure experience as the fundamental reality. Pure experience prior to analysis

or conceptualization does not distinguish between thing and thought. The commonsense categories, he believed, could be recovered from the contexts in which ingredients of pure experience are embedded and their various relationships. His apparent move toward philosophical idealism was qualified by his continuing commitment to pluralism: '… there is no *general* stuff of which experience at large is made. If you ask what any one bit of pure experience is made of, the answer is always the same: "It is made of *that*, of just what appears, of space, of intensity, of flatness, brownness, heaviness, or what not'" (1912/1976, p. 14–15).

5. *Religion and Psychic Research*

James continually struggled to find a place for the moral and hopeful view of humanity's place in the universe that religions can provide. Pragmatism was one route through which he sought to reconcile the world views of religion and science. Another was psychic research. For two years he was president of the Society for Psychical Research, to which he belonged from 1884 until his death. He was ever hopeful of finding supportive evidence for spiritual realities, and never satisfied that the evidence actually obtained was adequate (cf. James 1986).

James was already esteemed as a leading philosopher and psychologist when he delivered invited lectures on 'natural religion' at the University of Edinburgh in 1901–2. Their publication as *Varieties of Religious Experience* (1902/1985) became his second enduring classic, one that, a century later, is not only cited and respected but also read. Displaying the meaning of religion in the religious experience of individuals, it draws richly on accounts by saints, mystics, and others who provide evidence of their relation to what they regard as the Divine. The conceptual distinctions around which he organized his description and analysis have entered our common culture: the religion of healthy-mindedness vs. that of the sick soul; the 'divided self' and 'once-born' and 'twice-born' characters.

Although his sympathetic feeling for the many forms of religious belief is clear throughout, his presentation of religious experience is rightly esteemed by believers and unbelievers alike. His way of deploying 'case materials' to display the nature of a human phenomenon is a model for contemporary qualitative research to emulate.

James regarded the subconscious mental states revealed by the French psychopathologists as important in religious conversion. Speculatively, he declared his conviction that subliminal consciousness overlaps other realms of consciousness in mystical experience. By 'being faithful in my poor measure to this over-belief, I seem to myself to keep more sane and true' (1902/1985 p. 408). Here, James made radical and pragmatic use of his developing 'radical empiricism.' In the late twentieth century, this mystical side of James appealed to disaffected psychologists identified with a 'transpersonal' psychology with religious undertones.

6. *On Public Issues: Imperialism and War*

As a public intellectual, James was involved in the social issues of his day. A political as well as a philosophic liberal, he was opposed vehemently to emerging American imperialism in the wake of the Spanish American War. It was as an advisor to the peace movement that he wrote *The Moral Equivalent of War* (1910/1982), one of his last essays published in his lifetime, which remains a classic document in the psychology of war and peace.

After an evolutionary account of human warlike propensities, he noted that the militarists of his day held warfare to be the mainstay of admirable, martial virtues, and that as a 'pleasure economy' supplants the earlier 'pain and fear economy,' these virtues are likely to degenerate. He agreed that war fosters 'manliness' and surrender of private interest, but, given the monstrosity of modern war, he proposed planning for alternative sources of energizing challenge. Specifically, he suggested 'instead of military conscription a conscription of the whole youthful population to form for a certain number of years a part of an army enlisted against *nature*…' (p. 171). That he focused on 'manliness' and cast nature in the role of enemy shows James as a creature of his time, but his idea of the moral equivalent of war remains attractive. The equivalence is not in regard to the release of aggression, however, but to self-testing for the common good against difficult challenge.

7. *William James Today*

James's vivid and unconventional writings assure his continued influence, as they won him acclaim but evoked controversy in his lifetime. In North American psychology, James has become such a revered ancestor that his name gets called upon in many divergent contexts. His pragmatism is still very much alive. At one point, James considered humanism as an alternative term to pragmatism. Certainly, the influence of James in psychology and philosophy has favored humanistic responsiveness to the richness and variety of human experience, in contrast with the abstractness of absolute idealism in his own day and the dehumanized barrenness of behaviorism, logical positivism, and analytic philosophy that followed. His sensitive phenomenological treatment of thought and self remains unsurpassed, though it has been built on and superseded in various ways. His openmindedness and pluralism set a valuable counter-ideal to the dogmatism that is always a potential liability of mainstream doctrine. This counter-ideal inspired mid-

century psychologists of personality such as Gordon Allport, Gardner Murphy, and Henry Murray. At the turn of the millenium, it further inspires transpersonal psychologists who find in James respectable company for their own tendencies toward mysticism. The controversy surrounding James continues.

See also: Darwin, Charles Robert (1809–82); Dewey, John (1859–1952); Freud, Sigmund (1856–1939); Functional Explanation: Philosophical Aspects; Functionalism in Anthropology; Intellectual History; Pragmatism: Philosophical Aspects; Pragmatist Social Thought, History of; Psychology: Historical and Cultural Perspectives; Religion and Science; Religion: Definition and Explanation; Religion: Psychiatric Aspects; Religion, Psychology of; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

James W 1907/1975a *Pragmatism: A New Name for Some Old Ways of Thinking*. Harvard University Press, Cambridge, MA

James W 1975b–1988 *The Works of William James* (17 volumes). Harvard University Press, Cambridge, MA and London

James W 1912/1976 *Essays in Radical Empiricism*. Harvard University Press, Cambridge, MA

James W 1890/1981 *The Principles of Psychology*. Harvard University Press, Cambridge, MA

James W 1910/1982 The moral equivalent of war. In: James W (ed.) *Essays in Religion and Morality*. Harvard University Press, Cambridge, MA

James W 1902/1985 *The Varieties of Religious Experience: A Study in Human Nature*. Harvard University Press, Cambridge, MA

James W 1986 *Essays in Psychical Research*. Harvard University Press, Cambridge, MA

James W 1992 *The Correspondence of William James* (12 volumes). University Press of Virginia, Charlottesville, VA and London

Lewis R W B 1991 *The Jameses: A Family Narrative*. Farrar, Straus and Giroux, New York

Myers G E 1986 *William James: His Life and Thought*. Yale University Press, New Haven, CT

Perry R B 1935 *The Thought and Character of William James*. Little, Brown, Boston

Rorty R 1979 *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ

Taylor E 1983 *William James on Exceptional Mental States: The 1896 Lowell Lectures*. Scribners, New York

M. B. Smith

Janet, Pierre (1859–1947)

Pierre Janet, pioneering French psychologist and psychiatrist, was born May 30, 1859 in Paris. He studied philosophy at the elite *Ecole Normale Supérieure* and began his career teaching philosophy in aprovincial *lycée*. While teaching at Le Havre from 1883–9, he worked with local physicians to conduct experiments on hypnosis with hysterical patients. These experiments led to several publications, including the thesis for his doctorate in philosophy, *L'automatisme psychologique* (1889). In the same year, Janet obtained a position teaching in a Paris *lycée* and began formal training in medicine. He worked at the Salpêtrière hospital under the direction of the foremost authority of the day on hysteria, the neurologist Jean-Martin Charcot. Janet received his medical degree in 1893 and continued to work at the Salpêtrière until 1910. The cases he studied there formed the foundation for some of his most important works. In 1902, Janet was named to the chair of experimental psychology at the Collège de France, where he taught until 1935. He also co-founded and edited the *Journal de psychologie normale et pathologique*. He died on February 24, 1947 in Paris.

Janet's thought can be separated into an early and a late period, with 1910 as a rough dividing line. In the early period, which was by far the more influential, Janet focused on extending the domain of mental life and on elaborating a psychology based on a distinction between automatic and synthetic mental activity. As a result of his investigations into hypnotism and post-hypnotic suggestion, Janet rejected the then-common assumption in scientific and medical circles that unconscious mental functioning was physiological in nature. He insisted that it depended on the intervention of ideas, just as conscious mental functioning does. These ideas simply are not connected with the consciousness of the self. Because the word unconscious implied 'neurological' for so many at the time, Janet coined the term 'subconscious' to characterize ideas separated from consciousness.

The distinction between automatic and synthetic mental activity separated Janet from British associationist psychology. Janet held that association could only recall connections previously established by the mind; in other words, association explained automatic functions such as habit and conditioned response. Associations initially are established by mental synthesis, which combines the disparate elements of perception into related wholes. Synthesis is the basis for the integration and interpretation of experience. Beginning at the simplest level of combining sensory data into perceived objects, synthesis works on its own products to build ever higher levels of object, culminating in the concepts of the self and the world. By conceiving the self as the product rather than the origin of synthetic activity, Janet rejected the traditional Cartesian notion of the self as simple and primary; rather, the self is one of the most complex mental phenomena, and it appears only at an advanced stage of mental development. In addition, Janet dismissed the philosophical tendency to characterize abstract reason as the pinnacle of human mental

functioning. Influenced by evolutionary theory, Janet insisted that adaptive behavior in the real world required the greatest mental effort and that, in the real world, the social world posed more challenges than the natural world.

Synthesis requires energy, which Janet characterized through a concept of psychological force. An individual's psychological force determines the quantity of psychological stimuli they can hold together in synthetic unity. To describe the quality or level as well as the quantity of synthetic activity, Janet developed the idea of psychological tension. When individuals lose psychological force, they can no longer synthesize all of the stimuli impinging upon them. The inability to integrate perception leads to various kinds of mental illness, which Janet grouped into two major classes. One was hysteria, in which sufferers respond by abandoning or dissociating whole classes of perception. They may lose consciousness and control of parts of their body (producing hysterical paralysis), memory (amnesia), or behavior patterns (tics and fugues). In extreme cases, synthesis continues to operate on the dissociated portions of mental life beyond the awareness or control of the self, producing what was then known as double consciousness and what today would be called multiple personality or dissociative identity disorder.

To describe the other major class of mental illnesses, Janet coined the term 'psychasthenia,' which he used instead of the existing term 'neurasthenia,' to emphasize the psychological rather than biological nature of the condition. Psychasthenics retain consciousness of their uncontrollable thoughts and actions even as they are powerless to stop them. Such individuals suffer from obsessive-compulsive disorders and from feelings of inadequacy related to their inability to control themselves.

Janet attributed the lack of psychological force to two main causes—constitutional deficiency and traumatic emotions. Following in a long hereditarian tradition, Janet held that some people are born with less psychological force than others. Such individuals are more prone to mental illness, because they are more likely to be overwhelmed by the external world. In addition, traumatic incidents may produce emotions that sap energy and disrupt the normal process of synthesis. Individuals lose the ability and the desire to integrate into consciousness painful memories, which continue to play a disruptive role as dissociated subconscious ideas.

Janet's treatment of patients began with what he called 'psychological analysis,' during which the therapist attempts to separate symptom complexes into their elements and to uncover the origins and development of the ideas behind these elements. Janet used a variety of techniques in psychological analysis. He began his career relying heavily on hypnotism to reveal subconscious memories and dissociated personalities. He supplemented hypnotism with automatic writing and other techniques designed to elicit responses from the patient's subconscious. Such methods had limitations, however. Janet became sensitive to issues of suggestion in the production of hysterical symptoms. Indeed, he was one of the first historians of hypnotism, tracing the origin of the classic hysterical symptoms described by Charcot to the effects of earlier magnetizers on the patients Charcot examined. Janet also found that psychasthenics tended to be less susceptible to hypnotism. He came to rely more on repeated observations of and conversations with the patient over long periods. Through this interaction, Janet was able both to discover the origins of neurotic symptoms and to establish a relationship with the patient that would aid in the therapeutic process.

Once the origins and elements of the symptoms were discovered, Janet attempted to eliminate their pathogenic effects by a variety of methods. By enabling patients to reintegrate dissociated memories into consciousness, Janet was able to reduce their disruptive power. He also tried to dissolve such memories through suggestion and substitution. One case he reported was that of Justine, who had hysterical crises provoked by an intense fear of cholera. Through analysis, Janet discovered that as a child Justine had developed a fear of death by helping her mother, who was a nurse, care for dying patients. This fear was further focused by an incident in which Justine saw the putrefying corpses of two cholera victims. Janet recovered these memories, and then, through an elaborate and gradual process of suggestion, substituted more acceptable images—a Chinese general with the name Cho-le-ra for a diseased corpse. Eventually, the word cholera lost its power, and the symptoms disappeared.

Janet also tried various measures to adapt a patient's psychological force to the demands of her or his environment in order to prevent a recurrence of symptoms and to avoid the development of new illnesses. He used mental exercises to increase the patient's psychological force and recommended reductions in the stresses imposed by the patient's occupation and social life. Though powerless in most cases to change the sufferer's situation, Janet recognized the significance of environment in contributing to the onset of mental illness.

After about 1910, Janet began to redescribe his psychology in terms of conduct rather than ideas. He did this because he came to view the proper goal of psychology as the explanation of the adaptation of the organism to its environment. Ideas are instruments that lead to adaptive action. Janet also expanded and systematized his conception of the hierarchy of psychological levels. In addition, he became increasingly interested in the psychology of religious belief. These new directions came together in his most important late work, *De l'angoisse à l'extase* (1926), based largely on a case study of a highly religious patient with hysterical symptoms.

Janet's influence was considerable at the end of the nineteenth and beginning of the twentieth centuries. His concepts of dissociation and of subconscious ideas found adherents from Carl Jung to William James. Alfred Adler acknowledged his indebtedness to Janet's analysis of feelings of inadequacy. Jean Piaget cited Janet's work on the levels of psychological activity as inspiration for his own developmental theory. In 1906, Janet gave a series of lectures at Harvard Medical School, which made him better known in the US.

However, Janet's influence diminished rather than grew with time. Janet failed to train many followers. He also suffered in competition with Sigmund Freud. Not only was Freud more adept at forming a devoted following, he was more successful at defending the issues over which he differed with Janet, such as the sexual etiology of neuroses and the origin of repressed memories in fantasy. The rise of behaviorism, which despite its focus on action owed little directly to Janet, also eclipsed Janet's own psychology of conduct. In psychiatry, neurological and pharmacological approaches became dominant where Freudian psychoanalysis did not prevail, and in psychology, experimental methods replaced Janet's reliance on psychopathology.

In the last two decades of the twentieth century, a number of psychologists and psychiatrists began to study Janet anew. As Freud's abandonment of the seduction theory was challenged, some found a corrective in Janet's emphasis on actual traumatic events. Theorists who questioned Freud's insistence on the sexual origins of neuroses turned to Janet for a more inclusive view of psychological trauma. The resurgence of dissociative identity disorder as a diagnostic category also led researchers to seek guidance in one of the early students of that condition. Renewed interest in these and other areas to which Janet made substantial contributions assure him an important place in the history of psychology and psychiatry.

See also: Freud, Sigmund (1856–1939); James, William (1842–1910); Jung, Carl Gustav (1875–1961)

Bibliography

Brooks J I III 1998 *The Eclectic Legacy: Academic Philosophy and the Human Sciences in Nineteenth-century France*. University of Delaware Press, Newark, DE

Ellenberger H F 1970 *The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry*. Penguin, London

Janet P 1889 *L' automatisme psychologique: Essai de psychologie expérimentale sur les formes inférieures de l'activité humaine*. Alcan, Paris

Janet P 1893 *L'état mental des hysteriques*, 2 Vols. Rueff, Paris, (1977 *The Mental State of Hystericals*, 2 Vols. University Publications of America, Washington, DC)

Janet P 1898 *Névroses et idées fixes*, 2 Vols. Alcan, Paris

Janet P 1903 *Les obsessions et la psychasthénie*, 2 Vols. Alcan, Paris

Janet P 1907 *The Major Symptoms of Hysteria: Fifteen Lectures Given in the Medical School of Harvard University*. Macmillan, New York

Janet P 1919 *Les médications psychologiques*, 3 Vols. Alcan, Paris, (1925 *Psychological Healing*, 2 Vols. Macmillan, New York)

Janet P 1926 *De l'angoisse á l'extase: Etudes sur les croyances et les sentiments*, 2 Vols. Alcan, Paris

Prevost C-M 1973 *La psycho-philosophie de Pierre Janet: Economies mentales et progrès humain*. Payot, Paris

Sjövall B 1967 *Psychology of Tension. An Analysis of Pierre Janet's Concept of 'tension psychologique.' Together with an Historical Aspect*. Svenska Bokförlaget, Stockholm

J. I. Brooks III

Japan: Sociocultural Aspects

A sociocultural overview of Japan appears straightforward. Yet, shifting political boundaries of the nation state called Japan, its multi-ethnic composition of the population, and global movements of the 'Japanese,' together with recent epistemological and theoretical questions in various fields have reminded us that such an undertaking is a highly complex enterprise (Ivy 1995). The anthropology of Japan, practiced by an increasing number of scholars all over the world, is an enormously active field now that anthropologists no longer specialize only in nonliterate 'primitives.' It is a testing ground for new approaches to anthropology, including various attempts to historicize anthropology (Kelly 1985, Ohnuki-Tierney 1987, 1993). Presented below is a brief summary of the 'anthropology of Japan,' identifying important contributors but focusing primarily on scholarship based in Japan, the United States, and the United Kingdom for reasons of space (see also Hendry 1992, Eisenstadt 1996).

1. Anthropology of Japan: In the Beginning

Anthropology *in* Japan may be represented by two major approaches, both called *minzokugaku*, each having a different character for *zoku*, one referring to folklore or folk traditions, and the other, 'people.' 'Folklore' is equivalent to 'ethnography,' rather than 'folklore' as practiced in the United States. The basic approach was founded by Yanagita Kunio (1875–1962), who emphasized the fieldwork aspect of the discipline. Its emphasis on rural and other marginalized populations was later extended to urban populations under Miyata Noboru (1936–2000). Orikuchi Shinobu (1887–1953), another pillar of this school,

made extensive use of classic literature as ethnographic texts, thus introducing history to ethnography.

The other major approach is sociocultural anthropology, which began with the influence of the Viennese school, then turned to British social anthropology, and most recently to American cultural anthropology. Most Japanese scholars in this tradition specialize in non-Japanese cultures.

American anthropology in Japan is largely a product of World War II, when the American government commissioned scholars to decipher its enemy, with a most celebrated, and often criticized, case being *The Chrysanthemum and the Sword* (1946) by Ruth Benedict. Early works focused on 'village Japan,' where they sought a more 'traditional' Japan than modernized city life.

2. *Making the Majority and Marginalizing the Internal 'Others'*

Wet-rice agriculture, introduced from the Asian continent to Japan around 400 BC, gradually supplanted the previous hunting–gathering subsistence economy and provided the economic and symbolic foundation for the Yamato state and the imperial system, which lost its political power after the twelfth century, until it was nominally 'restored' through the Meiji Restoration in 1868.

The rice-based agrarian political economy and cosmology provided a lasting impact on the social structure and cosmological tradition, which became *the* Japanese cosmology *sui generis* for all the Japanese, including non-agrarian peoples. This development contributed to the representation of Japan as 'Agrarian Japan,' writing off many social groups from Japan's history and society. We now see a burgeoning interest in 'minorities' and counter-discourses within Japanese society, especially among foreign scholars. Since the cultural, linguistic, and biological identities of the Japanese, Koreans, Ainu, and Gilyaks remain uncertain, and anthropology no longer recognizes 'race' as such, the question of Ainu identities remains controversial. The Ainu engaged in a hunting–gathering economy without a writing system, with, however, a highly developed oral tradition (Ohnuki-Tierney 1981). Subscribing to the scheme of social evolution, the Japanese as well as Westerners were eager to identify the Ainu as 'primitive hunter-gatherers' or, even 'archaic Caucasoids.' Their way of life and physical characteristics were exoticized. As they were colonized (Takakura 1960), they were forced to abandon their way of life and became a minority, the Other within the Japanese nation state. Previously occupying the Kuriles, southern Sakhalin, Hokkaido, and northeastern Honshu, the Ainu are now confined mostly to Hokkaido. Their population dropped from a peak of 80,000 in the eighteenth century to anywhere between 25,000 and 50,000 at the end of the twentieth century. Since the 1960s, various Ainu political groups have made great strides in asserting their indigenous rights.

The so-called *hisabetsu-burakumin* is another by-product of the development of 'Agrarian Japan' (Ohnuki-Tierney 1987). Until the late medieval period, they constituted loosely defined social groups in non-agrarian occupations dealing with culturally defined 'impurity,' including animal and human deaths, and the visual and performing arts. In 1582 they were codified as 'outcastes,' becoming increasingly marginalized in later periods. Despite legal emancipation in 1870, social discrimination has continued into the twenty-first century. The 1987 government census figures are: 4,603 districts with a total population of 1,166,733, while the Buraku Liberation League suggests at least 6,000 districts with a population of 3,000,000. The Okinawans also constitute a minority whose 'distinctiveness,' if there is any, is based primarily on regional differences. Okinawa's past isolation from the main islands and its location as the gateway of influence from outside created some cultural differences, and their dialect was once regarded as non-Japanese.

There are various 'foreign' residents, of which 85 percent (700,000) are Koreans. Others include Chinese, Westerners, and more recent immigrants from Southeast Asia and the Middle East. In other words, Japan, although stereotyped as demographically homogeneous, has been a multi-ethnic nation throughout its history.

Although scholarly work on women had previously been available, studies undertaken at the end of the twentieth century have included feminist and other new perspectives, foregrounding their voices, their life experiences, and women as social agents who actively carve out or change courses of their lives (Hara Chieko, Iwao Sumiko, Karen Kelsky, Takie Lebra, Ueno Chizuko). The issue of gender preferences, institutionalized or culturally sanctioned as well as individual, has also received considerable attention (Robertson 1998).

3. *'Culture'—Non-localized*

Another burgeoning genre of scholarly work derives from an emphasis on how 'culture,' or a particular people such as the Japanese, is not localized. Our awareness of the nonisomorphic nature of culture, locality, and people and an unprecedented flow of population across borders of nation states has led to a number of studies of 'transnationalism' and transnationals in the past and in the contemporary context of globalization (Mannari and Befu 1983). These have included studies of overseas Japanese communities, young Japanese who have returned to Japan after being raised overseas due to their parental occupation (Goodman 1990), Japanese businessmen overseas (E. Ben-Ari), and emigrants who have returned to Japan

(Takeyuki Tsuda), as well as transnational economic activities (T. Bestor; H. Miyazaki). Studies of Japanese colonization have become another powerful lens to understand Japanese culture and society, along with Japanese cultural and political nationalisms (K. Yoshino).

4. 'Culture/History' at the Confluence of the Local and the Global

If the political economy and cosmology of rice agriculture were born as a result of the interaction between the global flow and the local development, so were subsequent major historical watersheds of Japanese history. Japan's 'modern' era was ushered in because of the threats of Western colonialism all over Asia. In 1868 Meiji oligarchs used the 'return of the imperial system,' which had been virtually nonexistent since the end of the twelfth century, as a means of overthrowing the shogunal government, which had proved incapable of coping with Western threats. The Constitution of Imperial Japan, drafted by German (Prussian) and Austrian scholars, transformed the emperor from an apolitical deity in the Japanese pantheon to a sovereign, Almighty God, and Commander of the military. With Japan's defeat in World War II, another constitution was written by the American occupying forces, creating a 'symbolic' emperor and introducing equal rights for women, coeducation in public schools, and land reform which drastically leveled social inequality.

Changes in the educational system too are the result of the local/global interactions. With a high literacy rate, less than 4 percent leave school at the age of fifteen, while almost all the rest complete senior high school education. Some 500 four-year universities and 580 two-year colleges in the 1990s contrast with the prewar system, which consisted of the National Higher Schools and the Imperial Universities, whose graduates, all male, dominated the key posts in the government, business, and intellectual fields. Today, in addition to the University of Tokyo, University of Kyoto, other major national universities, and the top private universities, such as Keio and Waseda, there are alternative routes to success, offered by the International Christian University or Jōchi University, run by Jesuits, Chūō University for law, and others. Holders of doctorates from US and UK universities receive increasing recognition. Despite democratization of the educational system in some sense, the recent affluence has produced more economically based inequality, with wealthy parents sending their offspring to costly 'cram schools' to increase their chances of passing the entrance examination to prestigious universities (J. Hendry; T. Rohlen).

Late twentieth-century debates over the enforcement of the national anthem and the national flag, textbook revisions, and curriculum changes form an ongoing debate about Japan's military past. After World War II, the American occupation forces enforced the well-known 'blacking out' of passages about Japan's aggression from school textbooks. Half a century later, the Japanese face the problem of how to reintroduce the Pacific war experiences, including atrocities committed by the Japanese military, such as the Nanking Massacre, 'comfort women,' and chemical experiments. Many Japanese are enormously sensitive to any signs of nationalism, and even patriotism, yet also remain sensitive to the 'foreign pressure.' A major factor in potential curriculum changes in pre-university education and also in teacher training is the call for 'globalization' or 'internationalization.' Another factor is pressure to locate the Japanese more solidly among other Asians, rather than upholding Westerners as their mirror, as has been the case in the past. Culture as a product of local–global interactions is well articulated in Japanese religions. Nominally, the Japanese religions consist of native Shintoism and two religions of foreign origin—Buddhism and Confucianism. In addition, the mountain ascetic practice (*shugendo*) continues to be practiced, while numerous 'new religions' came into being often during periods of unrest. At the end of the twentieth century several new religions enjoy large memberships, including Soka Gakkai, which also forms a successful political party.

Mahayana Buddhism came from India via China in the late sixth century and was first embraced by the elite but later became influential among common people. With the rise of political and military nationalism in the early Meiji period, it was targeted as a 'foreign' religion, and many Buddhist temples and works of art were destroyed. In contrast, Shintoism was 'elevated' to be the exclusive national religion. These changes at the institutional levels, however, had little to do with the common folk who subscribed to Shintoism and Buddhism, both of which had become a part of their daily lives. Most rites of passage celebrating life are Shinto rites, while most mortuary rituals are Buddhist. In Japanese folk religiosity, humans and superhumans constitute a continuum—exceptional humans are assigned power over humans who nonetheless manipulate and cajole the superhumans to their advantage. Japanese religions are decisively 'this-worldly,' rather than focusing on salvation after life. The major 'functions' of deities and buddhas are to provide business prosperity, traffic safety, recovery from illness, good health, success in one's undertaking, including high school and university entrance examinations, and other worldly gains (J. Kitagawa).

Christianity was introduced to Japan when Jesuit missionaries arrived in 1549. The converts were persecuted during the sixteenth and seventeenth centuries, when the rulers saw Christianity as potentially subversive of their authority, and again later during the nationalistic period. However, many Meiji Christian

leaders were highly respected by a broad spectrum of people as intellectuals and pillars of moral conscience under the increasingly totalitarian government.

Numerous practices that fall under the rubric of 'magic' and 'superstitions' are found in profusion in urban centers and are practiced by a large number of Japanese, including the young and the educated, who carry amulets in their cars, and hang them on children's rucksacks and hospital beds. Children and high school students hang votive plaques on which they write a prayer, for example, for success in examinations and dating. Young and old alike practice divination of all kinds when choosing a marriage partner, a date for a wedding, a date for opening of a store, etc.

Although worldly concerns occupy a central place in Japanese religiosity, at its core is a belief in the soul, especially that of the deceased. Various practices associated with souls of the deceased constitute the most important rituals at the end of the twentieth century. These include funerals and later rituals connected with them, as well as the daily offering of water, tea, freshly cooked rice, and favorite foods of the deceased on their ancestral alcove. Municipalities own large cemeteries where the family tombs of the residents are located. During the Bon festival, when the ancestors are supposed to return to this world, and at other prescribed times, families in cars form caravans heading toward these cemeteries. This practice continues in intensity despite the almost complete absence of a vision of the afterworld (Smith 1974).

With a population of about 123 million, Japan's demographic pattern is characterized by a large proportion of people over 65 (about 20 percent at the end of the twentieth century), the world's lowest infant mortality rate, and a very low birth rate. The household, *ie*, consisting of an extended family with its emphasis on generational continuation and the primogeniture system, has been considered the basic social unit, although it was codified only in the Meiji period and outlawed in 1947. This image is misleading, since the 'continuation' of *ie* does not mean a continuation through blood line—a high incidence of adoption characterizes Japanese kinship, and the system was most important only for families with substantial wealth and/or important occupation. For most, the nuclear family has been far more important. Since the mid-nineteenth century, industrialization accelerated the migration of nuclear families to urban centers where well over three quarters of the population is concentrated today, leaving the rural areas primarily in the hands of the elderly.

Abortion, legalized in 1948 although practiced with high frequency even earlier, and infanticide were formerly the primary means for birth control. At the end of the twentieth century the condom (80 percent of contraceptive users) and the rhythm method are used, with the oral contraceptive legalized only in 1999. In spite of the lowest infant mortality rate in the world, the Japanese population growth is alarmingly low. The physical and psychological involvement of fathers in child rearing has increased in recent years, so that they spend more time with their families, especially on weekends.

While studies of workplaces by anthropologists are by no means neglected (Kondo 1990, C. Turner), popular culture now receives much attention (Moeran 1989; A. Allison; J. Callas; I. Condry; M. Creighton; K. Kelsky; P. Noguchi). 'Globalization' is most visible in 'youth culture,' a product especially of American influence. Whether or not they are in the mainstream of society, today's youth flourish in their youth culture, listening to the latest pop music, following the latest youth fashion, and eating out with their friends at popular restaurants. Many color their hair, sometimes blond. They have adopted body piercing after the fashion of American youth, in contrast to the centuries-old Japanese refusal to imitate Chinese ear piercing, which was seen as a form of impurity. Before the last decades of the twentieth century, even married couples did not show their affection in public. At the beginning of the twenty-first century, public places are crowded with young dating couples who openly display their feelings.

Young and old, twenty-first century Japanese spend a great deal of money and time on leisure activities, including 'traditional' nature viewing—viewing of cherry blossoms, chrysanthemum, and colored leaves. Television and games constitute the major sources of leisure activity. Sports have become primarily spectator sports watched on TV, with baseball and sumo the most popular. Both professional and high school baseball are enormously popular (W. Kelly). Sumo, the oldest professional sport in Japan, received a great deal of attention when it broke with tradition by giving the highest ranking of *yokozuna* to foreign-born wrestlers. There are now sumo organizations throughout the world (R. K. Tierney). One of the most popular participant sports is golf. Domestic and foreign tourism has become a major leisure activity for Japanese of all ages (N. Graburn). The young sometimes combine tourism with their passion for consumerism, going to France, Italy, and the United States to purchase 'brand name' merchandise. A 'unique' leisure activity is the reading of comic books, with some comic books having a weekly circulation of over 4 million—the only arena of recreation competing against the electronic media.

The healthcare system, too, has developed as a result of Japan's interaction with various others. After the introduction of Chinese medicine, Dutch medicine was introduced during the early eighteenth century, followed by the German system. Today most Japanese go to a biomedical doctor for major acute diseases and surgery, while using Chinese medicine for day-to-day health maintenance and chronic illnesses, in addition to going to temples and shrines to pray for speedy recovery (J. Breslau; M. Lock; E. Ohnuki-Tierney). In

1959–60, all Japanese became legally eligible for pension coverage and health insurance, provided either by one's employer or the government. During the affluent mid-1970s, a national system of free medical care for the aged was introduced, and there was a massive increase in the pension benefits. When the 'bubble' burst, the national scheme of free care for the elderly was discontinued in 1983, and various other contractions were enforced. At the beginning of the twenty-first century, state of the art biomedical technology is available in Japan, although brain death was legally recognized only in 1999, and the Japanese have shown considerable resistance toward organ transplantations in general.

Japan enjoys one of the lowest crime rates in the industrialized world. However, organized crime is a serious problem without any realistic solution in sight. The targets of organized crime are usually highly specific, without threatening the general public. However, periodically, Japan is shaken by outbursts of crime such as the Aum Shinrikyo cult, which in 1995 spread poisonous gasses in a Tokyo underground station, causing 5,000 deaths and injuries.

5. *Conclusion*

The critiques of 'totalizing' and 'essentializing' representations of culture/society/people have led the anthropology of Japan in several productive directions, foregrounding diversity and multiple voices, whereas the awareness of the nonlocalized nature of culture/society/people has stimulated anthropologists to come to terms with various new ways of looking at what indeed is the 'anthropology of Japan.' These new perspectives place Japanese culture and society in the global historical flow and geopolitics, rather than treating them as a hermetically sealed entity.

See also: Meiji Revolution, The; Southeast Asian Studies: Culture; Southeast Asian Studies: Economics; Southeast Asian Studies: Politics; Southeast Asian Studies: Society

Bibliography

Benedict R 1946 *The Chrysanthemum and the Sword*. Houghton Mifflin, Boston

Eisenstadt S N 1996 *Japanese Civilization*. University of Chicago Press, Chicago

Goodman R 1990 *Japan's 'International Youth.'* Clarendon Press, Oxford, UK

Hendry J 1992 *Understanding Japanese Society*. Routledge, London

Ivy M 1995 *Discourses of the Vanishing*. University of Chicago Press, Chicago

Kelly W 1985 *Deference and Defiance in Nineteenth Century Japan*. Princeton University Press, Princeton, NJ

Kondo D K 1990 *Crafting Selves*. University of Chicago Press, Chicago

Mannari H, Befu H 1983 *The Challenge of Japan's Internationalization*. Kwansei Gakuin University Press, Hyōgo, Japan

Moeran B 1989 *Language and Popular Culture in Japan*. Manchester University Press, Manchester, UK

Ohnuki-Tierney E 1981 *Illness and Healing Among the Sakhalin Ainu*. Cambridge University Press, Cambridge, UK

Ohnuki-Tierney E 1987 *The Monkey as Mirror*. Princeton University Press, Princeton, NJ

Ohnuki-Tierney E 1993 *Rice as Self*. Princeton University Press, Princeton, NJ

Robertson J E 1998 *Takarazuka*. University of California Press, Berkeley, CA

Smith R 1974 *Ancestor Worship in Contemporary Japan*. Stanford University Press, Stanford, CA

Takakura S 1960 *The Ainu of Northern Japan*. American Philosophical Society, Philadelphia, PA

E. Ohnuki-Tierney

Jeffreys, Sir Harold (1891–1989)

Sir Harold Jeffreys, F. R. S. was born on April 22, 1891 in Fatfield, County Durham, England and died on March 18, 1989 in Cambridge, England. He was educated at Armstrong College, Newcastle-upon-Tyne and at St. John's College, Cambridge University. After brilliant scholastic performance, he became a Fellow of St. John's College in 1914. From 1946 to 1958, he was Plumian Professor of Astronomy and Experimental Philosophy at Cambridge University. In 1925, he was elected a Fellow of the Royal Society, later became its Bakerian Lecturer and Royal Medalist and received the Copley Medal in 1960. He was knighted in 1953, President of the Royal Astronomical Society, 1955–57, and President of the International Association of Seismology and Physics of the Earth's Interior, 1957–60. In 1962, the Harold Jeffreys Lectureship was created by the Royal Astronomical Society and Jeffreys presented the inaugural lecture in 1963.

The obituary in the Times of London, March 23, 1989, noted that Jeffreys's 'famous treatise, *The Earth: Its Origin, History and Physical Constitution*, which with re-editing and improvements, went through seven editions between 1924, when it first appeared, and 1976, did perhaps more than any other book to make the geophysics of the solid earth a coherent study. His researches in seismology, in particular, established his reputation as one of the leading authorities on that subject.'

His early research was centered on astronomical and meteorological dynamics. An accomplished mathematician, he wrote two monographs, entitled *Cartesian Tensors*and *Operational Methods in Mathematical Physics*. He and his wife, Dr. Bertha Jeffreys, a very able mathematician, wrote *Methods of Math-*

ematical Physics, which was published in several editions by Cambridge University Press and became a highly respected and widely cited classic. In a review of the work in *Nature*, it is stated that 'The authors distinguished for their knowledge of mathematical physics and cosmogony, and by their researches in these fields, possess also a remarkably comprehensive and deep understanding of pure mathematics. Their purpose has been to synthesize the mathematical knowledge for the service of physics … The work is a benefaction to the cause of progress in natural philosophy.'

The Collected Papers of Sir Harold Jeffreys on Geophysics and Other Sciences was published by Gordon and Breach, London, in six volumes. Volume 6 is entitled, *Mathematics, Probability and Miscellaneous Other Sciences*. Work reported in this volume and in his books, *Scientific Inference* (1931 and later editions) and *Theory of Probability* (1939 and later editions) reflected Jeffreys's concerns with the foundations of science and, in particular, scientific inductive procedures. His book, *Theory of Probability* presents an axiom system for inductive inference applicable in all the sciences and operational procedures for description, measurement, modeling, data analysis, estimation, testing, and prediction. Many applications of these procedures to important statistical problems encountered in many sciences are provided in his books and papers.

Since he began research on probability theory after World War I in a series of papers written jointly with Dr. Dorothy Wrinch, his contributions have been of fundamental importance to the philosophy of science, scientific method, and theoretical and applied statistics. The general nature of this research has been stated in the preface of the second edition of *Scientific Inference* as follows: 'The general standpoint that scientific method can be understood if and only if a theory of epistemological probability is provided, remains unaltered. Consequently, I maintain that much that passes for theory of scientific method is either obscure, useless or actually misleading.'

In his book, *Theory of Probability*, Jeffreys presents a theory of epistemological probability, discusses its relevance for work in all areas of science, and applies it to solve many practical, important estimation, testing, prediction, and other statistical problems including those considered by R. A. Fisher, Joey Neyman, Karl Pearson, and others. Irving J. Good, an eminent statistician who reviewed Jeffreys's 1961 edition of *Theory of Probability* wrote that Jeffreys's book '…is of greater importance for the philosophy of science, and obviously of greater immediate practical importance, than nearly all the books on probability written by professional philosophers lumped together.'

As regards to Jeffreys's work on scientific method, he along with Pearson (1938) stressed the 'unity of science principle,' namely, that any area of study, for example, economics, physics, psychology, etc., can be scientific if scientific methods are employed in analyzing data and reaching conclusions. Or as Pearson (1938, p. 16), cited by Jeffreys states, 'The unity of all science consists alone in its method, not in its material.' With respect to scientific method, Jeffreys, in his book, *Theory of Probability*, provides an axiom system for probability theory which is useful to scientists in their efforts to learn from their data, explain past experience, and make predictions regarding as yet unobserved data, fundamental objectives of science, that are applicable in the natural, biological, and social sciences. He states that scientific induction involves observation and measurement and the production of generalizations or models to explain past data and predict new data, a view of induction that is fundamentally different from those of Mach and Popper. Further in induction, deduction plays a role but alone is inadequate for scientific work, mainly because deduction provides just three attitudes towards a proposition, namely, proof, disproof, or complete ignorance. As Hume and others have stressed, there is always the possibility of an exception to a scientific proposition. Thus, Jeffreys saw a need for statements less extreme than those of deductive logic, namely proof, disproof, or complete ignorance. His inductive logic provides such statements that involve reasonable degrees of belief in scientific propositions.

Jeffreys considers a probability to be a measure of the reasonable degree of confidence that an individual has in a proposition given the individual's available background information and data. This individualistic, subjective definition is compared to other definitions, including the axiomatic, long-run frequency, and hypothetical infinite population definitions in Chap. 7 of *The Theory of Probability*. He points out that the first definition is widely employed, for example, in the work of J. Neyman, the second is the Venn limit and strongly advocated by R. Mises, and the third is usually associated with R. A. Fisher although it had appeared earlier in statistical mechanics in the work of Gibbs and others. His criticisms of these concepts of probability are ingenious and devastating and have not been refuted in the literature, perhaps because they are irrefutable. In addition to specific detailed examples illustrating the inadequacies of these three definitions, Jeffreys comments, 'No probability has ever been assessed in practice, or ever will be, by counting an infinite number of trials or finding the limit of a ratio in an infinite series … A definite value is got on them only by making a hypothesis about what the result would be. The proof even of the existence is impossible. On the limit definition, … there might be no limit at all … the necessary existence of the limit denies the possibility of complete randomness, which would permit the ratio in an infinite series to tend to no limit.'

After further detailed criticisms of the above three definitions of probability, he writes, 'The most serious

drawback of these definitions, however, is the deliberate omission to give any meaning to the probability of a hypothesis.' (1998, p. 377). Also, he remarks, '…in practice no statistician ever uses a frequency definition, but that all use the notion of degree of reasonable belief, usually without even noticing that they are using it and that by using it they are contradicting the principles they have laid down at the outset' (1998, p. 369). See also Jeffreys's (1998 pp. 30–3) discussion of the 'personalistic' or 'moral expectation' views of probability.

Having provided a definition of probability, Jeffreys's axiom system provides procedures for using such probabilities in a coherent way. That is, he provides basic analysis that yields the addition and product rules and other results, including Bayes's Theorem or the principle of inverse probability that Jeffreys states (1998, p. 28) was '…first given by Bayes (1763). It is the chief rule involved in the process of learning from experience.' It is invaluable in using data to revise our initial beliefs in propositions regarding the values of parameters, alternative hypotheses and models, future values of variables, etc. With this analysis, Jeffreys presents a logically consistent framework for learning from data to explain the past and to make predictions about future outcomes, primary objectives of science, as mentioned above. He demonstrates, through analyses of many applied problems using data, that his estimation, prediction, and testing procedures work well in practice and makes many comparisons with results provided by the maximum likelihood approach of Fisher and others, Neyman-Pearson testing procedures, Fisher's p-value approach to testing, etc.

Central in Jeffreys's work on statistical estimation, prediction, and testing problems is Bayes's Theorem. Whatever the estimation problem, use of Bayes's Theorem enables an investigator to combine his prior information regarding possible values of the parameters, represented by a 'prior density' with the likelihood function that incorporates the information in the data to obtain a 'posterior density' for the parameters that incorporates all the prior and sample information. That is, using Bayes's theorem, the posterior density is proportional to the prior density times the likelihood function with the factor of proportionality a numerical constant such that the integral of the posterior density over its range is equal to one. Thus for whatever problem, given the inputs, a prior density and a likelihood function, a posterior density for the parameters is available. This posterior density can be employed to compute the probability that a parameter's value lies between any two given values, say between 0.4 and 0.6, a solution to the famous inverse problem posed by Bayes. Optimal posterior point estimates for parameters are obtained easily from a given posterior density, say a posterior mean, median or modal value that are optimal in terms of minimizing posterior expected loss *vis à vis* squared error, absolute error and zero-one loss functions, respectively. Further, in large samples, Jeffreys (1998, p. 193) shows that under general conditions, the posterior density assumes a normal shape centered at the maximum likelihood estimate with a covariance matrix equal to the inverse of the estimated Fisher information matrix (1998, p. 193) and remarks that R. A. Fisher's maximum likelihood methods yields '…results which are indistinguishable from those given by the principle of inverse probability (in large samples), which supplies a justification for it' (1998, p. 194).

Since a suitable prior density is needed to utilize Bayes's Theorem and has been a bone of contention for many years, Jeffreys devoted considerable attention to the problem of formulating appropriate prior densities. In particular, he put forward a reproducible way of formulating invariant prior densities appropriate for use in situations in which an investigator is ignorant, or wishes to proceed as if he were ignorant, about the possible values of parameters. His famous recommendation in such circumstances is to take the prior density for the parameter(s) proportional to the square root of the determinant of the Fisher information matrix (see p. 179ff for material on invariant priors). Use of this 'ignorance' or 'diffuse' or 'non-informative' prior density provides invariant posterior probability statements for all continuous one-to-one transformations of the parameters, a remarkable and highly original solution to the choice of prior problem. Jeffreys derived many invariant priors and used them in analyses of applied problems. He is usually given credit for being a pioneer in considering and solving the problem of formulating invariant priors.

With respect to prediction, in general a probability density function for future, as yet unobserved, observations is formulated conditional on parameters with unknown values. If we have, for example, a proper prior or posterior density for the parameters, the product of it and the conditional predictive density can be formed and the parameters integrated out to yield an operational predictive density. It can be used to make predictive probability statements, for example, to compute the probability that a future, as yet unobserved, observation will lie between, for example, 3.4 and 4.7. Also, optimal point predictions can be derived for given predictive loss functions, for example, a mean for quadratic loss functions, etc. For more on uses of predictive densities for forecasting and control, see the general references cited above.

As regards to the evaluation of alternative models or hypotheses put forward to explain empirical phenomena, Jeffreys recommended workers to start with simple models and complicate if necessary. He and Dorothy Wrinch put forward the Simplicity Postulate, namely a statement that simpler models have higher probabilities of performing well. Jeffreys considered all possible models ordered with respect to degree of complexity and assigned probabilities to each, with

simpler models receiving higher probabilities, in an infinite sequence that is assumed to converge.

Within this broad modeling framework, on considering two alternative hypotheses, Jeffreys used Bayes's Theorem to derive an expression for the posterior odds on these two hypotheses that is equal to the prior odds on the two hypotheses times the Bayes's factor, the ratio of the predictive densities for the observations under the two alternative hypotheses. Jeffreys used posterior odds in his pathbreaking work on significance tests and compared his results for fundamental testing problems with those provided by use of Fisher's p-values and Neyman-Pearson testing procedures. Since Jeffreys's pioneering work on significance testing, many have employed and extended his methods to deal with a wide range of testing and model selection techniques, see, for example, the review paper by Kass and Raftery (1995). Given Jeffreys's brilliant formulation, interpretations and uses of significance tests based on posterior odds, much improved methods for evaluating and/or combining hypotheses and models evolved and are now in widespread use.

Since Jeffreys's initial work in the 1920s, many have come to appreciate his outstanding contributions to the philosophy and practice of science. It is noteworthy that he analyzed much nonexperimental astronomical and geophysical data. Yet his analysis of randomization in experimental design in his book *Theory of Probability* is extremely penetrating and valuable. In addition, he has made important contributions in the areas of distribution theory, asymptotic theory and expansions, robust estimation, invariance theory via differential geometry, information measures for distributions, mixture distributions, contingency tables, a key nonparametric problem, and in modifying the Pierre Simon Laplace Rule of Succession. And last, that Jeffreys's methods work not only in solving physical science problems but also in solving problems in biology, the social sciences, business, law and other areas is noteworthy. Finally, while many orthodox statisticians were critical of Jeffreys's work years ago, it is the case that his approach and results are currently widely employed and valued.

See also: Bayesian Statistics; Geographic Thought, History of; Probability: Interpretations; Statistics, History of; Statistics: The Field

Bibliography

Bayes T 1763 An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* **53**: 370–418

Berger J O 1985 *Statistical Decision Theory*, 2nd edn. Springer, New York

Bernardo J M, Smith A F M 1994 *Bayesian Theory*. Wiley, New York

Berry D A, Chaloner K M, Geweke J K 1996 *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*. Wiley, New York

Box G E P, Tiao G C 1993 *Bayesian Inference in Statistical Analysis*. Wiley Classics Library, New York

Cook A H 1990 Sir Harold Jeffreys biographical memoirs. *Fellows of the Royal Society of London* **36**: 303–33

Jaynes E T 1983 *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. Reidel, Dordrecht, The Netherlands

Jeffreys H 1931 *Scientific Inference*. Cambridge University Press, Cambridge, UK (Reissued with addenda, 1937, 2nd edn. 1957, 3rd edn. 1973)

Jeffreys H 1939 *Theory of Probability*. Oxford University Press, Oxford, UK (2nd edn. 1948, 3rd edn. 1961, 3rd rev. edn. 1967, Reprinted 1988 and 1998)

Jeffreys H 1971 *Collected Papers of Sir Harold Jeffreys on Geophysical and other Sciences*, Vols. 1–6. Gordon and Breach, London

Jeffreys H 1998 *Theory of Probability*, 3rd rev. edn. Oxford University Press, Oxford, UK

Kass R E (ed.) 1991 Tribute to Sir Harold Jeffreys. *Chance* **4**(2)

Kass R E, Raftery A E 1995 Bayes factors. *Journal of the American Statistical Association* **90**(430): 773–95

Kass R E, Wasserman L 1996 The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**: 1343–70

Pearson K 1938 *The Grammar of Science*. Everyman, London

Zellner A (ed.) 1980 *Bayesian Analysis in Econometrics and Statistics: Essays In Honor of Harold Jeffreys*. North-Holland, Amsterdam

Zellner A 1996 *An Introduction to Bayesian Inference in Econometrics*. Wiley Classics Library, New York

Zellner A 1997 *Past and Recent Results on Maximal Data Information Prior Distributions in Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*. Edward Elgar, Cheltenham, UK

A. Zellner

Jewish Diaspora

The study of international migration focuses on individuals or groups of people moving between countries. National boundaries constitute one key criterion in assessing which geographical mobility constitutes international migration, as well as the sociodemographic profile of the different relevant populations: the whole residents of a given country of emigration, the particular subset of actual movers, and the residents of immigration-receiving countries. Within this framework, the study of a diaspora is atypical. Rather than a clearly defined territorial entity, a diaspora is a globally dispersed set of subpopulations mutually related by common values, symbols, and historical experiences. Besides describing the experiences of each local subpopulation, diaspora studies aim to uncover the existence, determinants, and consequences of common patterns, if any.

The Jewish diaspora is a particular case of such a territorially dispersed sociodemographic collective de-

fined by religious–ethnic–cultural criteria. Variable definitions of 'who is a Jew' may affect evaluation of sociodemographic change (DellaPergola et al. 2000). While during earlier history group boundaries were rather rigid, namely concerning the status of the Jews, in the modern era of political emancipation, modernization, secularization, intensive communication and social interaction, religious, ethnic, national, or cultural identities tended to blur. While a declining number of countries have provided relevant information on Jews (or Hebrews) as a religious or ethnic category in their censuses and migration statistics, more recent sources such as population surveys increasingly rely on the respondents' self-definitions of group identity. The identificational boundaries and consequent quantitative assessment of Jewish populations become increasingly operative as people may declare different identifications under variable circumstances. However, since 1950 the state of Israel has used a broad normative definitional framework through the Law of Return (Hok hashvut) granting immediate Israeli citizenship to Jewish immigrants, their children, grandchildren, and respective spouses, even if they are not Jewish. The related concepts of a core Jewish population, and an enlarged Jewish population also including non-Jews of Jewish origin and other non-Jewish household members, should be kept in mind in the study of the Jewish diaspora's sociodemographic change, including international migration.

1. *Geographical Change in Historical Perspective*

Like other diasporas but to a greater extent than for other groups, Jewish international migration has spread over a very long time span. The unique amount of geographical mobility experienced by Jews over history stands behind the negative stereotype of the wandering people—a rootless ideal type presumably associated with other negative characteristics. While repeated large-scale migrations did occur along Jewish history, explanations stem from a complex array of political, economic, social, and cultural determinants. In a historical perspective, minority status often exposed Jewish populations to legal, socioeconomic, and cultural constraints. These influences deeply cut across Jewish communities worldwide, though in each locale significantly different sociopolitical and cultural environments generated unique incentives or constraints to stay or move. Repeatedly in the Jewish experience this took the form of expulsions or deportations.

The main focus of this article is since 1850 with an emphasis on the period after World War II. It suffices, therefore, to note that Jewish populations had high levels of geographical mobility from antiquity throughout the middle ages and the early modern period. The initial stages of Jewish peoplehood formation and consolidation occurred through wanderings across the area between Mesopotamia and Egypt. The land of Canaan—later known as Eretz Israel, the Holy Land, and Palestine—eventually became the center of ancient Jewish settlement and political hegemony. A Jewish diaspora was permanently established by since the deportation to Babylon (eighth to sixth centuries BC). The Jewish presence expanded to North Africa and Southern Europe during the last centuries BC and the first centuries CE after the fall of the second Jewish Temple under the Roman Empire, and especially following the westward expansion of Islam since the seventh century. By the end of the first millennium northbound migration had led to the consolidation of the culturally influential Jewish communities of *Ashkenaz* (Germany in Hebrew). Jewish population growth in the Iberian peninsula created the other main cultural center of *Sepharad* (Spain in Hebrew). It is only after the thirteenth century that significant numbers of Jews moved eastward to Poland and Lithuania, eventually expanding to the whole of Eastern Europe.

Reflecting these migrations, the main centers of Jewish settlement repeatedly shifted. Evidence from the late twelfth century still points to a predominantly Middle Eastern geographical distribution. The subsequent rise of Western European Jewry was curtailed by numerous expulsions from European cities, provinces and countries, culminating in the major exodus of Jews from Spain and Portugal in the fifteenth and sixteenth centuries and their dispersal throughout Europe, the Mediterranean, and the Caribbean. Eastern Europe consequently became the main regional center of Jewish presence.

Following at least ten centuries of stagnation at an estimated level of a few hundred thousands to slightly above one million, since the early seventeenth century world Jewish population rapidly increased, mostly in Eastern Europe. The estimated global total grew from 1 million in 1700 to 2.5 million in 1800, 10.6 million in 1900, and 16.5 million in 1939. Demographic change reflected a comparatively early transition to lower levels of mortality in the presence of widespread marriage and moderate to high fertility. Rapid population growth and a consequent impoverishment along with the social–environmental pressure of antisemitism accompanied by outbursts of physical violence stimulated large-scale Jewish transoceanic migration during the last two decades of the nineteenth century and the first two decades of the twentieth century. Major Jewish communities thus rose in the US, Canada, Latin America, Southern Africa, and Australia.

After World War II and the destruction of Jewish communities in Central and Eastern Europe (6 million estimated victims), North America became the largest regional concentration of world Jewry. Since independence in 1948, large-scale immigration was the leading factor in Israel's rapid rise to have the second largest Jewish population worldwide, drawing from all diaspora communities but especially from Eastern

Table 1
World Jewish population estimates, by major regions, 1825–2000

Region	1825	1900	1939	1948	2000
Total (thousands)	3,281	10,600	16,500	11,185	13,192
Total (percent)	100.0	100.0	100.0	100.0	100.0
East Europe	69.2	71.6	49.4	23.8	4.1
West Europe	14.0	11.1	8.2	9.3	7.9
Palestine/Israel	0.2	0.5	2.7	5.8	37.0
Other Asia	8.9	3.6	3.4	4.2	0.4
North Africa	7.3	2.9	3.0	5.3	0.1
North America	0.2	9.7	30.0	45.6	45.9
Latin America	0.1	0.2	2.6	4.7	3.2
South Africa, Oceania	0.1	0.4	0.7	1.3	1.4

Europe, North Africa and the Middle East. Table 1 summarizes the main changes in world Jewish population distribution over the nineteenth and twentieth centuries.

2. *Volume of Jewish Migration, Countries of Origin and Countries of Destination*

The changing structure and interrelated character of modern Jewish international migration fits a systemic perspective (Kritz et al. 1991). The number, direction and characteristics of Jewish migrants at any time were significantly determined by the existing worldwide distribution of Jews—in turn, to a large extent, the product of previous migration (DellaPergola 1998; Lestschinsky 1960; Schmelz 1971).

About a quarter of a million Jews migrated mostly from Central Europe between 1840 and 1880. Estimates of the changing volume of Jewish international migration between 1880 and 2000 are presented in Fig. 1. Five-year moving averages were computed

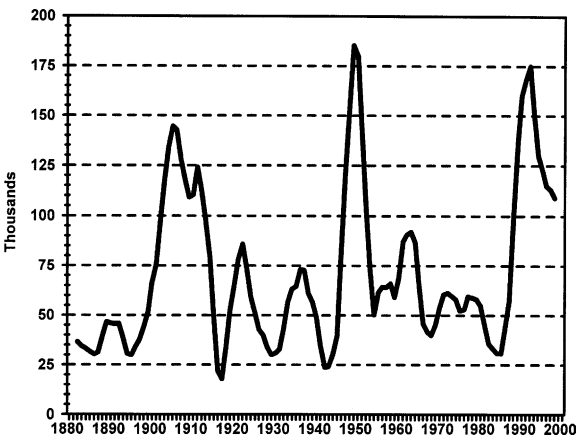

Figure 1
International migration of Jews, 1880–2000—five-year moving averages

from yearly migration data and estimates for each of the main countries of destination. Overall, over eight and a half million Jews migrated between continents, including all Jewish migration to Israel and to the US. Of these, about 2.4 million moved between 1880 and 1918, 1.6 million between 1919 and 1948, 1.9 million between 1948 and 1968, and about 2.6 million between 1969 and 1999. These figures do not include several hundred thousands more who migrated between neighboring countries within the same continent, nor does it provide a full account of return migration.

The continuous wave-like pattern in Jewish international migration demonstrates recurrent crises affecting the position of Jewish communities in different parts of the world and the consequent need for prompt relocation, which in turn suggests dependency of the Jewish diaspora on changing political, socioeconomic, and cultural circumstances in the respective countries. Some of the stimulating factors of large-scale emigration operated in recent years in essentially the same way as they did at the turn of the twentieth century in spite of the deep intervening sociopolitical changes across world society and Jewish communities alike.

Three major waves were: (a) the pre-World War I mass migration from Eastern Europe to North America, peaking in fiscal years 1905–1906 (154,000 Jews admitted), 1906–1907 (149,000) and 1913/14 (138,000); (b) the massive transfer to Israel of displaced and refugee Holocaust survivors and of sizeable Middle Eastern and North African Jewish communities between 1948 (239,000 in 1949) and 1951; and (c) the more recent exodus from the former Soviet Union (FSU) (228,000 in 1990 and 195,000 in 1991). These major moments of crisis or post-crisis adaptation occurred at comparable intervals of about 40–45 years, much in response to major events in the global polity (pre-World War I and post-World War II adjustments, the end of the Soviet Union). While no mechanical or deterministic assumption is suggested about the alignment of global geo-political forces, economists and demographers have discussed the possible cyclical unfolding of economic change and its

Table 2
Estimated Jewish international migration, by major areas of origin and destination—absolute numbers, percent distribution, and rates per 1000 Jewish population in countries of origin, 1881–1999

Areas of origin and destination	1881–1918	1919–1948[a]	1948[a]–1968	1969–1999	Total
Total (thousands)	2,400	1,615	1,880	2,626	8,521
Yearly average (thousands)	63	55	91	85	72
Total (percent)	100	100	100	100	100
From Eastern Europe[b]	97	86	33	55	68
To Western countries	95	65	6	22	47
To Palestine/Israel	2	21	27	33	21
From Asia–Africa[b]	3	5	52	16	18
To Western countries	2[c]	2	15	3	5
To Palestine/Israel	1	3	37	13	13
From Palestine/Israel[d]	[c]	3	10	16	8
From Western countries[e]	0	6	5	13	6
Regional subtotals:					
To Western countries	97	70	31	41	60
To Palestine/Israel	3	30	69	59	40
Rate per 1000 Jews, total	6	4	8	6	6
From Eastern Europe[b]	12	8	12	48	18
To Western countries	12	6	3	18	9
To Palestine/Israel	0	2	10	30	9
From Asia-Africa[b]	3	3	83	94	38
To Western countries	2[c]	1	28	28	14
To Palestine/Israel	1	2	55	66	25
From Palestine/Israel[d]	[c]	5	5	4	4
From Western countries[e]	0	1	1	1	1

a May 15 (Israel's independence day). b Since 1990, Asian regions of FSU included in Asia-Africa. c Emigration from Asia-Africa to West countries includes emigration from Palestine. d All emigration attributed to Western countries. e To Palestine/Israel

impact on national political systems (Kuznets 1958). Consequences for specific social or ethnoreligious groups tended to stimulate mass migration.

The volume of the main Jewish migration flows is summarized in Table 2 according to a fourfold typology: Eastern Europe; Muslim countries in Asia and Africa; Western countries including America and Western Europe; Ottoman and British Palestine turning since 1948 into the state of Israel. Until 1918, Eastern Europe was the main sending region and the US the main recipient. During the inter-war period, Eastern Europe continued to be the main area of origin but migration was sparser and Palestine emerged as a main country of destination. Since 1948, Israel became the dominant country of destination, especially between 1948 and 1968 when the majority of Jewish migrants came from Muslim countries in Asia and Africa. Since 1969 Eastern Europe again became the main supplier. Over the whole period 1881–1999, 60 percent of all Jewish migrants went to Western countries, primarily the US, and 40 percent went to Palestine/Israel.

Against the predominant Eastern European origin of Jewish migration, its intensity as reflected by yearly rates of emigration per 1000 Jews in the countries of origin was highest from North Africa and the Middle East, fairly low and declining from Israel, and lowest from migration-receiving western countries. At the beginning of the twenty-first century, the overwhelming majority of world Jewish population was located in countries with low emigration propensities. In prospective, and short of significant changes in the current global and regional circumstances, Jewish international migration was bound to decline.

Jewish migration to Israel, usually called by the value-oriented term of *aliyah* (ascent in Hebrew) consistently followed a wave-like pattern. Historiography has numbered such migration waves serially, starting with the 'first' (1881–1903) through the 'fifth aliyah' (1932–1938). The same cyclical pattern actually prevailed throughout the nineteenth century, though with relatively small numbers of migrants, and it clearly continued after Israel's independence. In absolute numbers the greatest single wave occurred in 1948–1951, followed by others of declining size until the major surge of the early 1990s. However, relative to a rapidly expanding receiving population, the highest and quite unique impact of immigration on the receiving society was recorded during the 'fourth aliyah' (283 immigrants per 1000 Jewish residents in

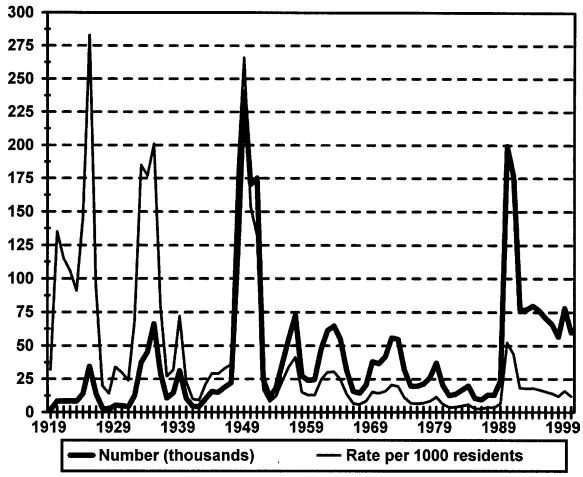

Figure 2
Immigration of Jews to Palestine/Israel, 1919–2000—numbers and rates per 1,000 resident population

1925), followed by Israel's 'great aliyah' (236 per 1000 in 1948 and 266 per 1000 in 1949) and the 'fifth aliyah' (201 per 1000 in 1935) (see Fig. 2).

3. *Main Determinants of Jewish Migration*

International migration has usually represented a major mechanism of adjustment for Jewish communities endangered by disruption of the orderly relationship with the surrounding society and severe worsening of civic rights and economic opportunities. Besides the variable intensity of pro-migration forces reflecting the socioeconomic and political situation within countries of origin (DellaPergola 1998), migration policies were an important factor (Bachi 1977, Friedlander and Goldscheider 1979). Regarding the actual possibility to leave, during the inter-war period and after World War II restrictive Jewish emigration policies were implemented by the Soviet Union and its allied countries in Eastern Europe (in the framework of more general limitations) and other mostly Middle Eastern countries. The effect of border opening is clearly demonstrated by the migration wave from the FSU in the 1990s.

Availability of alternative countries of destination crucially affected the volume and direction of Jewish migration. The landmark was the introduction of restrictive immigration policies by the US in 1921 and 1924. The latter established annual quotas for each country of origin amounting to no more that 2 percent of the respective population already present in the US by 1890. The number of Jewish immigrants to the US was forced down drastically to 10,000 per year. Restrictions followed in Canada, Argentina, Brazil, South Africa, Australia, Palestine (based on 'economic absorptive capacity'), and in Western Europe. Opportunities radically changed after Israel's independence in 1948 and its opening to virtually unrestricted Jewish migration. Israel's ideological call for mass Jewish immigration came together with instant citizenship rights and a modicum of socioeconomic help aimed at easing the initial integration of new immigrant households. Since World War II repeated changes in the Jewish quota again affected the number of Jewish migrants to the US, namely from the FSU. Around 2000, Germany was the second main recipient of Jewish migration after Israel.

Assistance provided by Jewish international and Israeli agencies generally enhanced migration. Various bodies promoted, directed, supported, and subsidized Jewish migrants, such as the Jewish Colonization Association (JCA), the World Zionist Organization, the Jewish Agency for Israel, the American Jewish Joint Distribution Committee (JDC), the Hebrew Immigrant Aid Society (HIAS), and Nativ, the liaison bureau of Israel's government in the FSU (Avni 1991, DellaPergola 1998, Wischnitzer 1948).

The recent migrants' feedback reflected their satisfaction with the situation in the new countries. Information sent back to family and other social networks in the countries of origin significantly affected the latter's decisions about further emigration. The final impact of migration was determined by the amount of return migration. The ratio of net to total immigration was generally higher among Jews than among other comparable groups. At the peak of migration to the US in 1908–1924, the amount of return migration was 5 percent among Jews versus 33 percent among all immigrants (Hersch 1931, Sicron 1957). Later, the rate of immigration retention was higher in Israel (78 percent in 1948–1998, including emigration of Israeli-born) than in the US, Argentina, Brazil, Australia, New Zealand, and countries in West Europe at the peak of the respective immigrations. Short-term economic determinants were the main explanatory factor of emigration from Israel.

4. *Characteristics of Migrants*

The demographic, socioeconomic, and sociocultural characteristics of the Jewish population significantly affected differential propensities to emigrate, and the choice of country of destination.

4.1 *Demographic Characteristics*

Because of its political–cultural character and its more irreversible, rather than merely economic and temporary nature, Jewish migration was comparatively less demographically selective than other migration streams. Entire families moved, including men and women of the whole age range rather than just young single individuals. This meant comparatively high

dependency ratios. Wide demographic differences separated Jewish diaspora communities that had undergone different degrees of modernization, such as the contrast between Jewish populations in Europe and populations in North Africa and the Middle East. In comparison, Jewish migration to Israel included more child and elderly dependents than migration from similar countries of origin to western countries such as the US or France.

4.2 Socioeconomic Characteristics

Jewish migrants included a broad cross-section in terms of socioeconomic stratification. Jews usually had higher-than-average education in the respective environments, with special concentrations in trade, selected branches of industry, and the liberal professions. Jewish migration propensities were higher at the lower and upper ends of the social spectrum. Here again structural differences that had emerged in the diaspora affected the characteristics of migrants to different countries. Israel absorbed a comparatively higher proportion of migrants previously employed in trade and blue-collar occupations. The difference was especially notable in the case of migration from North Africa to France and to Israel during the 1950s and 1960s (Bensimon and DellaPergola 1984, Sicron 1957). Immigration from Western countries and the more recent wave from the FSU contributed unusually high proportions of university graduates that enhanced the quality of Israel's labor force.

4.3 Ethno-Religious Identification

While the general definition of Jewish migration applies to the whole of migration flows reported here, not all of the migrants who moved together as families or as communities were actually Jewish. Intensive assimilation experienced by Jews in the diaspora in recent decades translated into high percentages of mixed marriages. The enlarged Jewish population inclusive of non-Jewish household members in the FSU at the close of the twentieth century amounted to twice or more the size of the core Jewish population. The number and proportion of people entitled to emigrate under special provisions for Jewish migrants established in Israel under the Law of Return, in the US quota system, and in Germany, therefore tended to grow significantly.

5. Consequences for Jewish Communities and for Countries of Destination

Mass relocation played a crucial role in shifting the main centers of Jewish presence globally. Even more significantly, the passage from less to more modern environments produced significant changes in demographic behavior, socioeconomic opportunities, and cultural outlook among Jewish communities. In the US and other western countries, successful immigrant absorption over a three-generation span stimulated a fast rise of Jews from one of the lowest to one of the highest-ranked ethnic and religious groups in the social ladder.

In Israel, absorption of mass migration created the backbone of a new society and significantly stimulated national economic growth and complexity (Metzer 1998). Between 1948 and 2000, net migration contributed about one half of Israel's total population growth of 4.3 million, while much of the natural increase was produced by immigrants and their descendants. Diverse levels of modernization and socioeconomic development attained by Jewish diaspora communities in the respective countries translated into internal social gaps within the new framework of Israeli society. The ingathering of the exiles and fusion of the diasporas became central tenets of Israeli social policies both in the immediate and in the longer term (Bachi 1977, Schmelz et al. 1991).

See also: Diaspora; Historiography and Historical Thought: Jewish Tradition; International Migration: Mediterranean Basin and the Middle East; Migrations, Colonizations, and Diasporas in Archaeology; Zionism

Bibliography

Avni H 1991 *Argentina and the Jews: A History of Jewish Immigration*. The University of Alabama Press, Tuscaloosa, AL

Bachi R 1977 *The Population of Israel*. The Hebrew University, Jerusalem

Bensimon D, DellaPergola S 1984 *La Population Juive de France: Sociodémographie et Identité*. The Hebrew University, Jerusalem, and Centre National de la Recherche Scientifique, Paris

DellaPergola S 1998 The global context of migration to Israel. In: Leshem E, Shuval J T (eds.) *Immigration to Israel: Sociological Perspectives*. Transaction, NB, Canada, pp. 51–92

DellaPergola S, Rebhun U, Tolts M 2000 Prospecting the Jewish future: population projections 2000–2080. *American Jewish Year Book* **100**: 103–46

Friedlander D, Goldscheider C 1979 *The Population of Israel*. Columbia University Press, New York

Hersch L 1931 International migration of the Jews. In: Wilcox W (ed.) *International Migration. National Bureau of Economic Research, New York*, Vol. 2, pp. 471–520

Kritz M M, Lim L L, Zlotnik H 1991 *International Migration Systems: A Global Approach*. Oxford University Press, Oxford, UK

Kuznets S 1958 Long swings in the growth of population and in related economic variables. *Proceedings of the American Philosophical Society* **102**: 25–52

Lestschinsky J 1960 Jewish migrations, 1840–1956. In: Finkelstein L (ed.) *The Jews: Their History, Culture and Religion*. Harper, New York, Vol. 2, pp. 1536–96

Metzer J 1998 *The Divided Economy of Mandatory Palestine*. Cambridge University Press, Cambridge, UK
Schmelz U O 1971 Migrations. *Encyclopaedia Judaica*. Keter, Jerusalem, Vol. 16, pp. 1518–29
Schmelz U O, DellaPergola S, Avner U 1991 *Ethnic Differences Among Israeli Jews: A New Look*. The Hebrew University and American Jewish Committee, Jerusalem
Sicron M 1957 *Immigration to Israel, 1948–1953*. Falk project for economic research in Israel and Israel central bureau of statistics, Jerusalem
Wischnitzer M 1948 *To Dwell in Safety: The Story of Jewish Migration since 1800*. Jewish Publication Society, Philadelphia

S. DellaPergola

Jewish Law

The term 'Jewish law,' the modern English epithet for halakha, covers legal arrangements in different geographic areas and historical eras, from biblical times to the present. Jewish law, the foundational texts of which are the Torah (the Pentateuch) and the Talmud, is a normative system regulating all aspects of daily life, private and public, including ritual, property, family, commercial and penal law, and the law of obligations. In the public domain, it encompasses the judiciary, taxation law, communal regulations and constitutional law. Although some segments of the halakha are valid only in the land of Israel, the jurisdiction of Jewish law is personal, not territorial, and applies to Jews everywhere. Jewish law also makes a claim on humanity in general through its code known as the seven Noahide laws: the injunction to establish courts, and the prohibition of blasphemy, idolatry, sexual transgressions, bloodshed, robbery, and eating flesh cut from a living animal (bSanhedrin 56a). (Tractate names prefaced by b, j, and t refer to the Babylonian Talmud, the Jerusalem Talmud, and the Tosefta respectively). Today, not all Jews accept the authority of Jewish law, though in previous millennia it was accepted as binding by Jews everywhere.

This article will provide an overview of Jewish law from the perspective of legal theory; descriptions of institutions and trends are therefore general and do not mention exceptions and unique cases.

1. *Preliminary Remarks*

The role of the halakha in postbiblical Jewish civilization cannot be overestimated. It is not only the force that held Jews together, both in the land of Israel and throughout the Diaspora, but its study and interpretation are also the primary focus of Jewish intellectual endeavor, by far overshadowing philosophy, homiletics, and Kabbala. Study of the law is a paramount religious duty, to be carried out daily by laymen as well as scholars. Social and personal success are measured by achievement in studying the law. Furthermore, students are encouraged to speak their mind, even if their opinions diverge from the received view, empowering individuals to participate in the evolution of the law. Though limited to those actually engaged in Torah study, this participation has a ripple effect, generating a sense in the community that the law, being an ongoing enterprise rather than dictated from above, is democratic, so to speak. Most halakhic literature is authored by individuals who hold no official position, and whose authority is accepted by the community due to their erudition and integrity. Another element that reflects the participatory dimension of Jewish law is its cardinal, all-embracing decision-making principle that the law is in accordance with the majority view.

1.1 *Historical Survey*

The following periods can be distinguished in the history of Jewish law: the biblical period, the talmudic period (first century BCE to fifth century CE), the Geonic period (sixth century to twelfth century), the period of the Early Authorities (Rishonim) (thirteenth century to sixteenth century); and the period of the Later Authorities (Aharonim) (seventeenth century to the present). Four main geographic centers of legal activity are similarly distinguished: the land of Israel and Babylonia, in the talmudic and Geonic periods, and later, Sfarad the Iberian peninsula (north Africa, and Ashkenaz Franco-Germany). Peripheral centers were largely under the influence of these four centers. As to the biblical era, on the internal self-understanding of Jewish law, the Torah is an expression of a unified divine revelation, a claim unacceptable to contemporary biblical research. Hence a convention has emerged among most scholars of Jewish law to the effect that biblical research is considered a field distinct from research in Jewish law.

1.2 *Jewish Law as a Legal System*

From the jurisprudential point of view, it is difficult to see the different periods and centers as constituting a single system of law, since the hierarchical institutional structure that is the hallmark of any legal system is absent. Jewish law developed despite the fact that it is not the positive law of any state, and its central institution, the Great Sanhedrin, the high court—the institution responsible for maintaining the system's uniformity and coherence—has not existed for about 2,000 years. Certain parts of Jewish law, especially those corresponding to modern criminal law, have not been in force since. A tension between this anomalous

reality and the ideal of a sovereign Jewish regime in its own territory with its own institutions, reflected in the frequently encountered term 'in our day,' which connotes the inferiority of our times, pervades Jewish law.

In the absence of a constitutive institution, Jewish law is generated by the following elements, between which there is yet more tension: (a) the rabbinical establishment, the teachers and students of the law at the rabbinical academies (*yeshivot*), where they study the canonical texts of Jewish law, engaging in debate and dialogue as questions arise; (b) the Jewish communal establishment, responsible for conducting the civic affairs of Jewish communities; (c) the members of the community, who, by adopting certain customs and comportments on their own, force the establishment to acknowledge these behaviors, sometimes condoning them, sometimes attempting to repress them; and (d) the Gentile societies within which Jews live. The extent of the autonomy of Jewish communities varied from place to place and era to era, but there was always a certain trickledown effect by which Gentile mores penetrated the Jewish communities; it may explain, to some extent, the diversity of the different Jewish communities.

A more direct source of this diversity of customs and behaviors, which is tantamount to a state of *de facto* legal pluralism, is the fact that communities are guided by different halakhic authorities. Special status is accorded the local authority, the *mara deatra*. The halakhic establishment validates the community's acceptance of the halakhic authority of its choice, recognizing his decisions as binding on residents of that locale, even if at variance with the prevailing views on the issues at hand.

At the same time, no rigid schools of thought were created in the Jewish world, and the works and rulings of every halakhic authority, regardless of the place and time of their composition, are considered relevant to the halakhic dialogue. Even if it is difficult, from a conceptual point of view, to justify the claim that the various centers, throughout the different periods, make up a single system, from the sociological point of view there is no doubt that the communities see themselves as constituting a single system.

1.3 Legislative and Judicial Institutions

The two main organs of any legal system, the judicial and the legislative, are also found in Jewish law. Whereas rulings of the Sanhedrin were universally valid, 'in our day' all legislation (*takanot*) is of local validity only. Nevertheless, it is abundant and of wide scope, covering all areas of civil and ritual law. The judiciary, too, is highly developed. The following issues pertaining to courts (*batei din*) are extensively discussed in the halakhic literature: alternative methods of dispute resolution, such as arbitration and compromise; appointment of judges; qualifications for judges; payment for judging; composition of the court; self-representation and representation by an agent; due process and equal treatment of all sides; the prohibition against personal or financial interest in the case on the part of the judge; fitness of witnesses and their examination; issuing rulings in writing; and the publication, finality, and implementation of rulings. The extent to which these theoretical discussions can be applied in practice, however, is a function of the degree of autonomy granted the community by the Gentile authorities. Adjudicating disputes within the Jewish court system, where possible, is zealously insisted upon, resort to Gentile courts being considered tantamount to renunciation of the halakha's authority.

2. Talmudic Law

2.1 The Talmudic Period as the Formative Era of Jewish Law

The formative era of Jewish law is the talmudic period, during which foundational texts of Jewish law—the Mishna and the Talmud—were composed. The Mishna is the pre-eminent compilation of the teachings of the Tannaites, sages who lived in Palestine in the first two centuries CE. Arranged thematically, it covers all aspects of Jewish law. The Mishna itself became a focus of interpretative analysis by later sages, the Amoraites, whose work was not limited to commentary on the law, but also addressed the general culture of the time. Palestinian Amoraic academies produced the Jerusalem Talmud, compiled around 400 CE; Babylonian Amoraic academies produced the Babylonian Talmud about a century later. Redaction is far more evident in the latter than in the former.

2.2 Divine Origin and Human Elaboration

Talmudic law is premised on the divine revelation of the Torah to Moses (first half of the thirteenth century BCE) at Mount Sinai (the written law), and the accompanying oral revelation (the oral law), later committed to writing. Together, they record the divine precepts (*mitzvot*) according to which the Jewish people must conduct themselves. Despite the Torah's divine origin, it was subjected to human scrutiny and elaboration. A distinction emerged between *deoraita* and *derabanan*, the former term denoting that deemed of divine origin, the latter, that deemed of human origin. Rabbinic interpretation of the law is not arbitrary, but applies what are known as 'the 13 exegetical principles by which the Torah is interpreted.' Some of these principles of logical inference determine the meaning of the written law, others

enable new rules to be extrapolated from it. There is, however, some debate as to whether these rules were actually employed by the sages of the Talmud to generate new laws, or were but a literary device used to link new laws to the biblical text (*asmakhta*).

Talmudic law distinguishes between precepts pertaining to interpersonal relations (*mitzvot bein adam lehaveiro*) and precepts pertaining to the relationship between individuals and the Almighty (*mitzvot bein adam lamakom*). These realms are not altogether distinct, being interwoven, both conceptually and in practice. The precepts can also be divided into those falling under monetary law (*mamonot*) and those falling under ritual law (*isur veheter*). Monetary law, deemed *jus dispositivum*, is generally approached with greater flexibility, allowing adjustments necessitated by changing circumstances to be readily instituted. Despite this flexibility, in monetary disputes extreme caution is exercised before ruling that someone's assets are to be taken away.

2.3 Jewish Law as an Elaboration of the Talmud

The evolution and philosophy of Jewish law cannot be understood without a firm grasp of talmudic law. The unique position of the Babylonian Talmud vis-à-vis later works is grounded in the fact that it is the latest text universally accepted as binding by the halakhic authorities. Responsible for establishing its authority were the Geonim, heads of Babylonian talmudic academies after the Muslim conquest of Babylonia. From then on, every legal discussion takes as its starting point the Babylonian Talmud's position on the question at hand, though it is not necessarily followed. Accordingly, the remainder of this article will focus on the features of Jewish law that emerge from talmudic law.

Due to its interpretative nature, halakhic creativity emerges in successive layers, in tree-ring fashion. As earlier contributions are always retained, the evolution of the law can be traced relatively easily. Most of the halakhic literature is essentially elaboration of and commentary on the Talmud—the pre-eminent commentator being Rashi, R. Solomon Yitzhaki (France, 1040–1105)—and its application to changing circumstances. The Tosafists, students of Rashi, were pre-eminent in this enterprise. At various times, so-called 'legal codes,' expositions of the law up to that point, were complied, particularly in the Sfardic tradition. Notable examples are the code of Maimonides (Spain-Egypt, 1135–1204), the *Mishne Torah*, the sole completely comprehensive treatment of Jewish law; the *Arbaa Turim*, by R. Jacob b. Asher (Franco-Germany, Spain, c. 1269–1343); and the *Shulhan Arukh*, by R. Joseph Caro (Israel, 1488–1575). These codes themselves became the focus of interpretive activity, generating a new genre of commentaries that came to be known as 'armor bearers,' the codes being the 'knights.'

3. Characteristic Features of Jewish Law that Emerge from Talmudic Law

3.1 Exposition and Implementation

A potentially misleading duality between the exposition of the law and its implementation in practice has always existed in Jewish law. The use of exact measurements and precise formulations (*midot veshiurim*), the meticulous phrasing of laws, the avoidance of such abstract formulations as 'reasonable behavior' and 'bona fide conduct,' could create the impression that talmudic law is a dry legalistic system within which rights and obligations are decided pedantically. But this impression is fundamentally incorrect. The nature of talmudic law as implemented in practice (*halakha lemaase*) cannot be inferred from its theoretical exposition (*halakha*). This distinction is not imposed upon the sources, but found in the sources themselves. Thus, for example, the Talmud explains a misunderstanding as follows: 'He was talking in terms of theoretical law, and they understood him as if he was talking about the implementation of the law' (jBeitza 2a); and states: 'One who inquires about practical law must state that he is inquiring for practical purposes…[given] a practical inquiry and a theoretical inquiry, the practical takes precedence' (tSanhedrin 7:7).

Any attempt to describe Jewish law must therefore carefully distinguish between law intended for implementation and theoretical law. An account of the former for the talmudic period must be based on episodes, precedents, and rulings related in the Talmud; for the post-talmudic period, the responsa literature is a far more reliable source for ascertaining the law intended for implementation than the other genres of halakhic writing. The responsa literature, which originated in the Geonic period, and has flourished ever since, consists mainly of legal opinions issued in response to queries submitted by courts regarding cases before them. Over the years, a corpus of some 250000 responsa has accumulated; part has been electronically indexed by the Bar llan Responsa Project. Another source of information about actual application of the law is extant records of the legal activity, judicial and legislative, of various communities (*pinkasei hakehilot*) from centuries past, as well as extant collections of legal instruments.

3.2 Judicial Deviation from the Law

An accepted truism in western legal thinking is that a judge's ruling must accord with the law. It must be presented as grounded in a well-defined body of norms supplied by the system, known in advance, and applied to the case at hand. The jurisprudential approach that emerges from the Babylonian Talmud does not fit this model. Rather, the judge has the power to deviate from the law and base his ruling on extralegal

considerations. A tradition reported in the name of R. Eliezer b. Jacob encapsulates this aspect of the judicial process: 'I heard that the *beit din* punishes… not according to the Torah' (bSanhedrin 46a).

Obviously, in most cases there will be no discrepancy between the theoretical law and concrete rulings. What is significant is not the frequency of deviation, but the fact that it is an option. The judge does not see himself as constrained by the legal rules, and if, in his eyes, the circumstances of the case justify deviating from them, he need not hesitate to act accordingly. Of course, this does not sanction arbitrary, unreasoned deviation. When a judge deviates from the law, he does so in order to impose the spirit of the Torah, which is traditionally perceived as not given to reduction to a set of rules. Over time, this became the norm in the application of the law throughout the halakhic world, and while judges did not always feel comfortable deviating from the law, their authority to do so was never questioned.

Thus, three levels of law, the first two of which, will, more often than not, overlap, can be discerned: theoretical law (*halakha*), law intended for implementation (*halakha lemaase*), and concrete rulings, which will occasionally diverge from the law intended for implementation.

3.3 Governed by Men, Not by Rules

There is, therefore, a salient difference between the approach of Jewish law and that of western law with respect to the judicial process. The dictum, 'governed by rules, not by men,' which reflects the western ideal, does not reflect the halakhic outlook (see Ben-Menahem 1991). The western approach is based on the idea that a litigant is one who claims his due, not one who seeks succor. The litigant's claim is premised on the court's submission to a well-defined, publicly known body of norms, and on a legal rule that, in the eyes of the litigant, entitles him to a favorable ruling. In contrast, a litigant appearing before a rabbinical court claims no entitlement to a favorable ruling, even if there is a legal rule that, in his eyes, entitles him to such a ruling. Because the judge is permitted to deviate from the rule and decide the case as he sees fit, the litigant's status is that of one who seeks the court's aid. The converse adage, 'governed by men, not by rules,' thus more accurately expresses the worldview of Jewish law.

This analysis may explain the absence of three almost universal legal institutions from Jewish law: equity courts, courts of appeal, and the doctrine of binding precedent. Given the potentially *ad hoc* nature of a halakhic legal decision, as outlined above, the doctrine of binding precedent is beside the point. Courts of appeal are intended to ensure the uniformity of the law and the supremacy of legal rules, objectives that are not a priority for the halakha. Equity courts are a corrective device to redress possible inflexibility on the part of the law. Given the freedom accorded the judge in the Jewish legal system, no need for them was perceived.

From the judge's perspective, the freedom to deviate from the law if necessary is anchored in the notion that judging is a religious precept, the fulfillment of a divine calling. While the weightiness of this charge may induce a paralyzing fear that the law will be distorted and justice miscarried should he misrule (bSanhedrin 6b), it may liberate the judge from the sense that he is bound by legal rules, which, formulated in human language, are subject to human limitations. The judge's mission transcends the simple application of a rule to a given case. He becomes, as it were, God's partner in administering the law: 'Judges should know whom it is that they are judging, before whom they are judging, and who will call them to account [if they pervert justice], as it is written. 'Consider what ye do, for ye judge not for men, but for the Lord' (II Chron. 19:6)' (bSanhedrin 6b). The religious—transcendental dimension of the halakha is most apparent here.

A judicial system of this type can thrive only in a community that accepts the judge as a spiritual leader and role model, and not just an expert technocrat hired to administer the law. At the same time, this concept of the judge is to be distinguished from that of the cadi under the palm tree, as the outlook of the judge in a halakhic proceeding is thoroughly informed by the halakhic discourse.

4. Jewish Law in the State of Israel

With the renewal of the idea of the return to Zion, the idea of constituting Jewish life in the land of Israel on the basis of Jewish law was put forward out of ideological motives by nonreligious thinkers. At the beginning of the twentieth century, the now-popular phrase *mishpat ivri*, Hebrew law, was coined, to denote that part of Jewish law corresponding to civil and criminal law, excluding the halakha's ritual aspects. The foundational work in Hebrew law, which sought to provide a comprehensive picture of Jewish law from a legal perspective, was Rav-Tzair's (Tchernowitz's) *History of Jewish Law* (1945–53, Hebrew); Menachem Elon's *Jewish Law: History, Sources, Principles* (1994) is essentially an updated and much-expanded version of this work. Overall, however, this agenda never gained acceptance. When the founding of the state of Israel began to appear imminent, the idea was reintroduced, this time by national religious circles. Again, it failed to catch on. Today the positive law in Israel recognizes Jewish law as the law of the state only in the realm of family law, which is adjudicated in state-run rabbinical courts, and in a few isolated laws pertaining to religious observance. Otherwise, the influence of Jewish law on Israeli law is minuscule. However, rabbinical courts unconnected to the state thrive in Israel, as they have throughout history in the Diaspora.

Nevertheless, Jewish law is the focus of much scholarly interest at Israeli universities and colleges, as well as, of course, in Israel's numerous *yeshivot*. Contemporary work in Jewish law treats, among other subjects, bioethics and halakha, feminism and halakha, and economic analysis of the halakha.

5. *Non-orthodox Understanding of Jewish Law*

Beginning in the nineteenth century, the hegemony of the halakha as understood by the rabbinical establishment was challenged. The most radical critique came from the Reform movement, which altogether rejected the halakha. The Conservative movement sought to accelerate the pace of halakhic change, in view of the rapid technological and social changes of the twentieth century, and to articulate more explicitly the role of conscience in the halakha. Given the preceding analysis of the halakha as pluralistic, it appears to be unnecessary to dwell on the Conservative understanding of Jewish law. While the divide between Conservative and Orthodox Jews may be highly significant from a sociological perspective, there is no divide with respect to acceptance of the halakha's authority. The divergences between the Orthodox and Conservative communities are essentially no different than those between other communities on the halakhic landscape. As for the Reform movement, to the extent that it has rejected the halakha, it is outside the purview of this article.

See also: Folk, Indigenous, and Customary Law; Historiography and Historical Thought: Jewish Tradition; Jewish Diaspora; Judaism; Judaism and Gender; Law: Anthropological Aspects; Legal Systems, Classification of

Bibliography

Ben-Menahem H 1991 *Judicial Deviation in Jewish Law*. Harwood, Chur, Switzerland

Ben-Menahem H, Hecht N S (eds.) 1988 *Authority, Process and Method*. Harwood, Amsterdam

Cohen B 1966 *Jewish and Roman Law*. Jewish Theological Seminary of America, New York

Elon M 1994 *Jewish Law: History, Sources, Principles*. Jewish Publication Society, Philadelphia

Elon M (ed.) 1995 *The Principles of Jewish Law*. Keter, Jerusalem

Hecht N S, Jackson B S et al. (eds.) 1996 *An Introduction to the History and Sources of Jewish Law*. Clarendon, Oxford, UK

Herzog I 1980 *The Main Institutions of Jewish Law*. Soncino, London

Jewish Law Annual, Vols. 1–13

Rakover N 1990 *The Multi-language Bibliography of Jewish Law*. Jerusalem

Rav-Tzair (Tchernowitz) C 1945–53 *Toledoth Ha-Halakah* [sic], History of Hebrew Law (Hebrew). Jubilee Committee, New York

Silberg M 1973 *Talmudic Law and the Modern State*. Burning Bush Press, New York

Weisbard P H, Schonberg D 1989 *Jewish Law: Bibliography of Sources and Scholarship in English*. Rothman, Littleton, CO

H. Ben-Menahem

Job Analysis and Work Roles, Psychology of

1. *Introduction*

Work analysis is essential if we are to understand and change the way people work and their work environment. If we wish to solve work-related problems, work analysis is not normally a matter of choice: we are either obliged to perform it explicitly or at least base our decisions on implicit work-analysis information.

Job analysis is part of work analysis. It is necessary (see also Ghorpade 1988) because

(a) jobs and their constituent tasks are the basic (design) elements of an organization, forming the basis for the constitution of work groups, departments and other larger divisions;

(b) jobs are consciously designed and changed; by people (e.g., industrial engineers) who do not normally perform them; and

(c) modern organizations are constantly changing (job analysis allows a continuous adjustment of job-related requirements to changing organizational goals).

There are a large number of jobs (e.g., in the production, office/administration and service sectors) and elements of the work situation (e.g., organizational structures, agreements on working hours, work objects, and means of work), and a wide variety of ways in which job-related information can be used, that practically preclude the possibility of adopting a universal approach or analysis method here. Accordingly, there is no one job-analysis method, but rather a number of systems, techniques, and methods that can be used to analyze jobs (see Dunckel 1999, Gael 1988).

Nor are there any universally valid definitions; terms such as job, position, task, and element being used differently in different countries and approaches. This fact should be taken into account by the reader, for this paper, too, uses terms in a way that would certainly not be endorsed by all job analysts.

Job analysis is the systematic, empirically oriented, problem-based collection of information on the interaction between humans and the work tasks and conditions that constitute a job or are connected with it.

Depending on the respective problem or goal (e.g., organizational planning and design, human-resource management), different approaches and different tech-

niques and methods are normally required (Dunckel 1999, Gael 1988, Ghorpade 1988). A distinction should be drawn between job analysis and job evaluation: job evaluation in the narrower sense aims to determine compensation levels for a job (Landy 1993); in a broader sense, job evaluation is also the evaluation of jobs according to given criteria, e.g., the extent to which jobs are conducive to health or personality development.

Jobs are composed of tasks, i.e., the same jobs are characterized by largely identical tasks. Task analysis is therefore a key element in job analysis (see Landy 1993). Task is generally taken to mean the purpose of a work system or the goal to be achieved under given conditions. Tasks specify goals to be attained by the worker, taking into account different overall conditions. Elements of jobs are therefore defined as work tasks if they relate in each case to different goals (see more detailed treatment in Oesterreich and Volpert 1987, Ghorpade 1988). Tasks are the points in a job at which the organization meets real individuals and makes concrete demands on the activities of these individuals. The task is thus the 'interface' between organization and individual and an essential element in psychological job analysis.

Psychological job analysis systematically captures and evaluates information on a working individual's activity. Here, it is not only concerned with the externally visible activity and the conditions surrounding it, but also with the psychological processes (e.g., thinking, motivation) and structures (e.g., memory) determining and regulating this activity. Psychological job analysis thus focuses on the concrete work activity as a psychologically regulated activity—i.e., the activity of a working individual.

Unlike 'classical' work studies or technical (job) analyses, psychological job analysis is inconceivable without involvement of the real working individual concerned. Supplying the workers affected with comprehensive information on the aims and objectives of an analysis, taking into account the fears and anxieties caused by it and providing feedback to the workers on the results are (or should be) essential elements of any psychological job analysis.

2. *Classification Criteria*

Job-analysis methods and techniques can be distinguished on the basis of various criteria (see also McCormick 1976): the unit of analysis, intended applications, theoretical foundation, method users and methods of data collection, methodological standards and results.

2.1 *Unit of Analysis*

The job-analysis unit is the job itself and the tasks that comprise it. For this reason, many job-analysis systems relate directly to the tasks and the immediate work conditions. More comprehensive job-analysis systems (see Dunckel 1999, Gael 1988) take into account the fact that essential features (e.g., the potential of the individual for defining goals self-reliantly) and results of the tasks become comprehensible only if the overall organizational conditions and the knowledge, skills and attitudes of the workers are taken into consideration. They therefore extend the analysis to include both the organizational conditions (e.g., degree of responsibility, codetermination potential, values and norms, leadership climate) and the working individuals with their respective knowledge, skills and attitudes.

Methods can be classified, based on widely accepted criteria, according to whether they are more job/task-oriented or worker/person-oriented. Job-oriented analysis is concerned with analyzing tasks, job conditions or job features, without regard for the concrete individuals involved and their different knowledge, skills, abilities and attitudes. Person-oriented analysis, on the other hand, is centred on the person and specifically concerned with differences between individual workers in the perception, interpretation and performance of the job.

Typical job-oriented approaches are: Functional Job Analysis, Task Inventories, Health Services Mobility Approach; typical person-oriented approaches are: Position Analysis Questionnaire, Critical Incident Technique, Ability Requirement Scales (see Gael 1988, Ghorpade 1988).

The distinction between job-oriented and person-oriented analysis is of conceptual and practical significance. In stress research, for example, it is conceptually important to begin by determining stress factors independently of the person involved (and their individual perceptions and coping behavior) in a job-oriented manner, then going on to examine how objectively identical stress factors are perceived and dealt with differently by different individuals and how they affect different people. However, the (person-oriented) analysis of these inter-individual differences means first determining objective stress factors, because we are concerned here with inter-individual differences in relation to objectively identical stress factors. This distinction is of practical significance when planning future workplaces for which there are not yet any workers. Here, the job-oriented approach is the only way of obtaining job analysis information.

2.2 *Intended Applications*

Techniques and methods have been and are being developed for different goals and potential applications. Job-analysis literature contains numerous different lists of goals and proposed uses for job-analysis information (see Ash 1988, Lees and Cordery

2000). Basically speaking, two goals complexes can be distinguished: work/job and organization design, and personnel development. More concrete aims of job analyses are, for instance, comparing work activities, changing and planning the work situation and organization, determining skill requirements and factors defining aptitude requirements, technology assessment, maintenance of industrial health and safety standards, and job evaluation.

Job-analysis methods and techniques play an important part in human-resources management and personnel selection (see Algera and Greuter 1998). There are used to determine more precisely the requirements a person must meet to perform their work tasks. These requirements may be specified as tasks to be accomplished, as behavioral requirements (e.g., required behavior or behavioral repertoire), as eligibility requirements (e.g., knowledge and skills) or as trait requirements (e.g., abilities and interests). For each of these requirement types, there are a number of techniques available (see Gael 1988).

If the results of such analyses are combined, job analysis can also be used to describe (work) roles as defined by Katz and Kahn (1978). According to these authors, roles are standardized behavior patterns demanded of all persons involved in a given functional relationship. Here, job analysis can also help identify when and under what conditions role expectations (of the different organization members with respect to individual workplace occupants) can lead to conflicts and job stress.

The intended applications of job analysis are not necessarily independent of one another. If, for example, the goal is to change the work organization, it will often be necessary to analyze the technological implications and the consequences for the workers involved as well.

Besides the application purpose, the application area also plays an important role. Job-analysis techniques and methods consider different levels and units of an organization. A distinction must be made between:

(a) sectors (e.g., industrial, administrative, service);

(b) levels of an organization (e.g., enterprise as a whole, business division, department, workplace group, workplace, work task);

(c) professional groups (e.g., executives, specialized professional groups); and

(d) activity classes (e.g., assembly, control and monitoring activities, administrative activities, service activities).

Depending on the specific concern, the emphasis will be on different information. Job analysis, too, faces the problem of breadth vs. depth. The more detailed the information, the more limited the application purpose. It may therefore be a good idea to proceed in several steps, starting with rough analyses to determine the key analysis areas, which are then analyzed in greater detail.

2.3 Theoretical Foundation

The theoretical foundation largely determines which information is captured at which level or with which analysis unit.

Work studies as defined by Taylor or Gilbreth (see Ghorpade 1988) are based on an additive *movement* (or motion) model. Techniques working on this basis thus attempt to define elementary movement units (e.g., grasping with the hand), combining these additively in order to then determine, say, the standard time required to perform the work.

Approaches rooted in behaviorism also attempt to identify elementary units (of behavior) (e.g., processing materials, recognizing optical differences). Their units are bigger, though; they are therefore not movement- but *behavior*-oriented.

These approaches, however, fail to take into account the fact that movements merge to form 'wholes,' are integrated in complex webs of activity and the regulating mental processes and representations and are codetermined by these. This is why many recent developments, especially in Germany, are based on the '*action regulation* theory' (see Frese and Sabini 1985, Oesterreich and Volpert 1987), thus giving priority to questions relating to the psychological regulation of action, the level of psychological regulation, the completeness of actions, the degrees of freedom (Hacker 1998) or the scope for action or decision. The guiding idea here is that of humane work, i.e., work geared to human strengths and enabling individual workers to perform their job under permanently tolerable conditions, without impairment of their well-being and in a manner conducive to their personal development.

In the emphasis they place on characteristics such as scope for action, variability, identity and importance of the task, action-theory approaches are in keeping with the traditions of industrial sociology, e.g., the work of Turner and Lawrence (1965) and the work of Hackman and Oldham (1975) that builds on this. Since the latter approaches are specifically concerned with questions relating to the 'motivation potential of work,' they can also be classified as oriented to *motivation theory*.

Besides drawing on approaches based on behavior, action, and motivation theory, job-analysis systems also have recourse to concepts of stress theory, ergonomics and human engineering; in addition, worker-oriented approaches draw on concepts of personality theory.

2.4 Method Users

The main users of job-analysis methods are the workers themselves, first-level supervisors, higher-level supervisors, job analysts, technical experts and other company experts, but also works and staff councils.

Whether and to what extent a method can be used depends, among other things, on the application requirements that must be met by the users. Important factors here, besides formal qualifications, are the amount of experience the users need in analysis techniques, whether they can teach themselves how to use such techniques or whether special training is required.

2.5 Methods of Data Collection

There are also fundamental differences between techniques in terms of the data-collection methods used, e.g.:

(a) interview methods (e.g., individual and group interviews, technical conferences with experts, more or less structured questionnaires and check lists);

(b) observational methods (e.g., direct and indirect observation, continuous observation, work sampling);

(c) analyses of company data (e.g., working hours lost, accident statistics, workplace descriptions);

(d) analyses of documents (e.g., file analyses, form analyses);

(e) work activities performed by job analysts.

Each of the methods has its advantages and disadvantages. It is therefore a good idea to combine several methods (see Gael 1988). These advantages and disadvantages can be highlighted by comparing interview and observation methods.

2.5.1 Interview methods. It is a good idea to interview workers, for they are the ones that know their own work activities best. Furthermore, such interviews are indispensable when the workers' subjective assessment of the work is needed or psychological processes are to be evaluated because these can only be accessed directly by introspection. In addition, interview methods, especially questionnaires, are frequently the method of choice because they are relatively easy to develop and use. Interview methods are the most frequently used job-analysis technique.

This should not, however, blind us to the fact that interview methods have a number of weaknesses. Some typical problems are:

(a) comprehension problems of workers who are not so accustomed to dealing with the written language (e.g., in the case of questionnaires);

(b) the ambiguity of everyday language;

(c) difficulties in translating scientific terms into everyday language; and

(d) the problem of putting into words many aspects of psychological regulation processes (Hacker 1998).

2.5.2 Observation methods. Observation methods are generally used in cases where it is important to avoid the sort of errors that can occur in interview methods or 'bias' as a result of evaluation and interpretation processes on the part of the workers, or when, in future workplace design, no workers are yet available for the planned jobs.

Observation methods are often seen as a way of getting round the problems inherent in interview methods and obtaining 'more objective' data. In reality, they are subject to the same sort of problems as interview methods, in some cases giving rise to additional problems:

(a) the quality of job observations deteriorates for complex work activities.

(b) certain temporally dynamic aspects of the work activity (e.g., pressure of time) are harder to observe.

(c) infrequent events, which are nevertheless of significance for the job, (e.g., starting and stopping machines, annual accounts) are often not included.

(d) observers, too, are subject to evaluation, interpretation and 'biasing' processes. For instance, observers tend to rate workplaces as uniformly good or bad.

These typical advantages and disadvantages mean that proper job analysis involves considering precisely which methods are suitable. This also means that users must be aware of the problems inherent in these methods, carrying out, where necessary, appropriate training measures to reduce them.

It is also a good idea—whenever this is feasible—to combine different methods, e.g., questionnaires, interview, and observation methods. For this reason, many techniques also include the *observational interview* as a proven data-collection method, based on structured observation of the work processes and related interviews with the workers involved at their workplace.

2.6 Methodological Standards

Job-analysis methods differ, among other things, in their degree of standardization—ranging from non-standardized 'free' descriptions to semi-standardized interviews to observations and interviews following exactly prescribed rules. They differ in the amount of time they take (from 30 minutes to several hours), their psychometric quality, the number of dimensions captured, etc. Which method to choose is not generally decidable; this depends largely on the intended application and on available knowledge and theory building with respect to the problems under investigation.

Scientifically based methods should be reliable and valid:

(a) The *reliability* of a method shows whether and to what a extent a method can be used to obtain 'stable,' 'reliable,' or 'replicable' results. Ideally, repeated measurements of the same object should show as little deviation as possible (see Oesterreich and Bortz 1994).

(b) The *validity* of a method indicates the extent to which it actually measures what it is supposed to measure.

Examining these quality criteria is not only of scientific interest, but also of utmost practical importance. If the results of the analysis are to have practical consequences, e.g., for job design, decision-makers must be able to depend on the fact that the results are reliable and valid.

2.7 Results

Job-analysis methods provide quantitative and qualitative results. Qualitative results are mostly verbal, narrative descriptions of the job or the tasks; quantitative results are presented in numerical—and in some cases graphical— form. Complex methods generally present results in both qualitative and quantitative form.

3. Outlook

In the future too, job analysis will continue to constitute a major basis for organization and personnel development. Given the numerous potential applications, there will be no one universal method covering all application purposes, even though the latest techniques allow more effective forms of information analysis and evaluation.

Job-analysis methods were developed principally for the industrial and administrative sectors. They exhibit shortcomings, for example, when analyzing activities in the service sector, domestic and family work and the self-employed sector. Furthermore, job analysis, too, will be obliged to respond to changes in the working world. New forms of work, e.g., jobs in virtual and networked organizations, teamwork within virtual structures, teleworking, and electronic business, also present new challenges for job-analysis techniques and methods.

In addition, in the area of human-resources management, job analysis is still needed (see Ash 1988) to help reduce gender discrimination with respect to pay by showing that the still existing differences in income between men and women on the same jobs cannot normally be justified by differing job requirements and conditions.

See also: Human Resource Management, Psychology of; Industrial–Organizational Psychology: Science and Practice; Job Design and Evaluation: Organizational Aspects; Job Design, Psychology of; Personnel Selection, Psychology of; Well-being and Burnout in the Workplace, Psychology of; Work, History of; Workplace Environmental Psychology

Bibliography

Algera J A, Greuter A M 1998 Job analysis. In: Drenth P J D, Thierry H, Wolff C J de (eds.) *Handbook of Work and Organizational Psychology*, 2nd edn. Psychology Press, Hove, UK, Vol. 3

Ash R A 1988 Job analysis in the world of work. In: Gael S (ed.) *The Job Analysis Handbook for Business, Industry and Government*. Wiley, New York, Vols. 1 and 2

Dunckel H (ed.) 1999 *Handbuch psychologischer Arbeitsanalyseverfahren (Handbook of Psychological Job-analysis Techniques)*. Vdf, Zürich, Switzerland

Frese M, Sabini J (eds.) 1985 *Goal Directed Behavior: The Concept of Action in Psychology*. Lawrence Erlbaum, Hillsdale, NJ

Gael S (ed.) 1988 *The Job Analysis Handbook for Business, Industry and Government*. Wiley, New York, Vols. 1 and 2

Ghorpade J 1988 *Job Analysis. A Handbook for the Human Resource Director*. Prentice Hall, Englewood Cliffs, NJ

Hacker W 1998 *Allgemeine Arbeitspsychologie* (General Industrial Psychology). Huber, Bern, Switzerland

Hackman J R 1970 Tasks and task performance in research on stress. In: McGrath J E (ed.) *Social and Psychological Factors in Stress*. Holt, Rinehart & Winston, New York

Hackman J R, Oldham G R 1975 Development of the Job Diagnostic Survey. *Journal of Applied Psychology* **60**: 159–170

Katz D, Kahn R L 1978 *The Social Psychology of Organizations* (2nd rev. edn.). Wiley, New York

Landy F 1993 Job analysis and job evaluation: The respondent's perspective. In: Schuler H, Farr J L, Smith M (eds.) *Personnel Selection and Assessment*. Lawrence Erlbaum, Hillsdale, NJ

Lees C D, Cordery L 2000 Job analysis and design. In: Chmiel N (ed.) *Introduction to Work and Organizational Psychology. A European Perspective*. Blackwell, Oxford, UK

McCormick E J 1976 Job and task analysis. In: Dunnette M D (ed.) *Handbook of Industrial and Organizational Psychology*. Rand McNally, Chicago

Oesterreich R, Bortz J 1994 Zur Ermittlung der testtheoretischen Güte von Arbeitsanalyseverfahren (On determining the test-theoretical quality of job-analysis techniques). *Arbeit Betrieb Organization Aktuell* **3**: 2–8

Oesterreich R, Volpert W 1987 Handlungstheoretisch orientierte Arbeitsanalyse (Action-theory-oriented job analysis). In: Kleinbeck U, Rutenfranz J (eds.) *Arbeitspsychologie* (Enzyklopädie der Psychologie, Themenbereich D, Series 111, Band 1). Hogrefe, Göttingen, Germany

Turner A N, Lawrence P R 1965 *Industrial Jobs and the Worker*. Harvard University Press, Cambridge, MA

H. Dunckel

Job Design and Evaluation: Organizational Aspects

Job evaluation is a method used to produce a hierarchy of jobs in an organization as the basis for determining relative pay levels. It seeks to measure the relative value of jobs, not that of the jobholders. Ideally, the performance of the individual should not enter into job evaluation although, in practice, it may be difficult

to dissociate individuals from their jobs. Job evaluation is based on the assumption that components of work can be isolated and that these can be decomposed into elements considered worthy of remuneration.

The main aim of job evaluation is to provide an acceptable rationale for determining the pay of existing hierarchies of jobs and for slotting in new ones. It may be implemented unilaterally or with varying degrees of participation by the workforce. Acceptability, consensus, and the maintenance of traditional hierarchical structures are normally the goals of such schemes.

Job evaluation schemes do not directly determine rates of pay. The rate for the job or the salary market for a job grade is influenced by a number of factors outside the scope of most schemes. Often, the pay determinants and indeed hierarchies are linked to external market pay rates, the relative bargaining strengths of the negotiating bodies, and traditional patterns of pay differentials between jobs. Job evaluation is concerned with relationships, not absolutes.

1. *The Rise of Job Evaluation*

Job evaluation grew up in the first half of the twentieth century, in an era of large, bureaucratic organizations and increasing unionization. There was a shift toward more 'rational' and 'scientific' methods of determining pay differentials. A major concern was to justify pay levels since many industrial disputes were concerned with wages. Collective bargaining, unilateral wage determination, or individual negotiations were seen as unsystematic compared to job evaluation. The rise of mass production and sometimes-militant industrial unionism was also important. As a result of new production methods introduced at the end of the 1930s and during World War Two, the number of varied yet unskilled occupations increased sharply. The clear line between mass production and traditional craft jobs had become blurred. Pay on the basis of job content, not personal attributes, complemented this system of industrial relations.

2. *Main Emphases in Job Evaluation Research*

For many years the main interest in job evaluation has come from scholars in the fields of industrial relations, labor economics, and industrial psychology.

During World War II some academic economists had been members of, or attached to, the US National War Labor Board, a body which encouraged the use of job evaluation. Subsequently, some of these scholars, notably Clark Kerr, wrote about the industrial relations aspects of the subject (Kerr and Fisher 1977). One of his conclusions was that the design of a scheme is of less practical importance than the context into which it is introduced and the way in which it is administered. Harold Northrup, another scholar with links to National War Labor Board, concluded that job evaluation was used by management 'partly to deter or prevent unionization, partly to rationalize its wage scales prior to unionization … and partly to stabilize the wage structure and eliminate continuous bargaining over particular rates after unionization.' (quoted in Hutner 1986, p. 25)

In the 1940s a very different strain of research was started by industrial psychologists interested in the design of rational, equitable pay systems and psychometric issues of reliability and validity. This early research generally concluded that trained individuals, familiar with the organization's jobs and given detailed instructions on how to analyze them, could reach similar conclusions on overall job worth (see, for example, Ash 1947). Interest in sex discrimination in pay systems stimulated further research on the validity of different job evaluation methods in the 1980s. This produced results which were much less clear (Arvey 1986). Madigan (1985) concluded that, while the reliability of various methods tested was in the range from .70 to .90, the accuracy with which the predicted pay grade matched the actual grade was much lower.

The validity of job evaluation has been investigated by looking at the results of different methods, using predictive validity procedures, factor analysis techniques, and testing for potential sources of bias. Results have been contradictory, since some researchers have concluded that the results will be similar whatever the method used (for example, Robinson et al. 1974), whereas others have found significant differences (see Madigan and Hoover 1986). There is some consensus that the skills and knowledge dimension, present in nearly all methods, is an overriding predictor of job evaluation results (Rogers 1946, Madigan 1985). It has been argued that this means that job evaluation mainly measures abilities that a person brings to the job, rather than the job content (Rogers 1946). Another research finding is that where the evaluators know the job title this significantly influences the results (Smith et al. 1989). A review of the psychological literature has proposed, unsurprisingly, that 'research models need to be expanded beyond focusing on the psychometric properties of job evaluation' (Gerhart and Milkovich 1992).

Recent research has drawn upon organization theory to study how individual and group power may influence the outcomes of job evaluation. The coalition model of organizations suggests that negotiation, bargaining power, and politics are key determinants of decision behavior, particularly resource allocation decisions (Pfeffer and Salancik 1974). Using this approach Welbourne and Trevor (2000) examined whether power in a university affected job evaluation outcomes. They concluded that both departmental and individual positional power helped predict the results. Departmental power predicted the number of job grade allocations and reallocations achieved and

positional power predicted the probability of favorable reallocation decisions. Powerful departments were more likely to approach the Human Resources Department with marginally acceptable requests and indulge in more extensive negotiations about the appropriateness of job descriptions and the proposed grade. The school's dean was likely to intervene prior to an official submission, particularly for higher graded positions. Welbourne and Trevor conclude that a rational model of behavior to explain outcomes is inadequate and that job evaluation captures the dynamics of the job's environment as well as its worth. Writing over sixty years earlier Viteles was less measured, claiming that 'beneath the superficial orderliness of job evaluation techniques and findings, there is much that smacks of chaos' (Viteles 1941, p. 165).

3. *Equal Pay and Job Evaluation*

In the 1960s equal pay legislation was introduced in many western economies. The concept of equal pay for equal work has been widened in many jurisdictions to include equal pay for work of equal value. In the European Union, where a dispute arises as to whether work is of equal value, a country is required to provide a process by which an assessment of value can be made. Such a process involves some form of comparison based on job evaluation or classification. Where job evaluation is used it must be shown not to discriminate against women. In some countries, particularly the US, equal pay for work of equal value, also known as comparable worth, has proved to be a highly contentious issue. Economists, labor lawyers, historians, and sociologists have joined the debates. The role of job evaluation in bringing about equal pay for work of equal value has been scrutinized and vigorously disputed.

Among economists there are many that dismiss the use of job evaluation as a remedy for pay discrimination against women. They believe that to attempt to compare the value of different jobs using job evaluation is like 'comparing apples with oranges.'

The power dimension in job evaluation has been made more transparent by historical analyses of the way in which job evaluation has been used to maintain traditional male differentials, reflecting the prevailing values or relative power in society. Job evaluation, despite its rhetoric of fairness, does not necessarily lead to pay equity for women. During World War Two a US National War Labor Board case by the United Electrical Workers against General Electric (GE) and Westinghouse sought equal pay for work of equal value (Milkman 1987). The job evaluation system used for all GE jobs gave different pay rates for men and women. The company responded that these gender-based rate differentials were longstanding, accepted by the union during bargaining rounds, and in line with industry practices. They argued that to pay women more would put the firms at a competitive disadvantage and that the board should consider 'sociological factors,' including women's different social and economic roles. Although the board decided in the union's favor, the companies ignored the decision (Milkman 1987, pp. 80–1). Although explicit references to male and female rates disappeared gendered concepts, such as the distinction between 'light' and 'heavy' work, became the basis for unequal pay.

Until recently most job evaluation schemes continued to incorporate the gender bias of the early twentieth century (Treiman and Hartmann 1981). Common practices were: the separation of men and women into different job evaluation schemes; greater weight being given to the attributes of male-dominated occupations; and using different scales for men and women when converting job evaluated points to pay rates (Baker and True 1947).

4. *The Law and Job Evaluation*

Equal pay legislation has exposed job evaluation to sometimes rigorous analysis in the courtroom. Out of this are emerging principles that seek to ensure that job evaluation schemes are constructed and implemented transparently and without obvious bias. A distinction accepted in the European Court is between analytical and non-analytical methods. In analytical schemes jobs are broken down into their component elements for the purposes of comparison. In non-analytical schemes the relative worth of the jobs may be based on a whole job comparison. More formal methods, particularly analytical schemes, have been judged to be more transparent and less susceptible to bias. However, it has been accepted that no scheme can ever be fully objective since the whole process is based on a series of judgments made about the facts presented to the evaluators. These judgments reflect each evaluator's background, experience, and attitudes.

5. *Future Directions*

By the middle of the 1980s job evaluation was being used by about 75 percent of US firms (Beer et al. 1985). However, for the past two decades increased individualization of the employment relationship has been developing and unions in most western economies have been on the retreat. Employers have been exhorted to be more flexible and willing to consider fundamental changes in employment practices. Increasingly job evaluation has been portrayed as bureaucratic, inflexible, reinforcing hierarchy, and not in tune with the needs of the times. Lawler has argued that in a rapidly changing, highly competitive environment, particularly where companies depend on

advanced technology and the knowledge of their employees, pay systems should emphasize growth, development, and performance (Lawler 1986, 1994). He claims that this is inhibited by job evaluation, which depersonalizes individuals by equating them with a set of duties, rather than concentrating on who they are or what they can do.

These views have been influential and there has developed a greater emphasis on paying people for their skills, knowledge, and performance. However, job evaluation, after becoming less fashionable in the 1980s and early 1990s, is in many countries more popular than ever. One reason for this is that the case against job evaluation has been overstated. Many of the problems associated with it can be overcome by redesigning schemes to fit the needs and values of modern organizations. A second reason is that there exists a flourishing job evaluation consultancy industry which has been quick to respond to some of the criticisms by modifying existing schemes and introducing new ones which incorporate more emphasis on skills, knowledge, competency, and performance. Broadbanding has introduced fewer, wider pay bands, thereby reducing hierarchical levels and the need for constant reevaluations.

A further influence has been that in many countries organizations have introduced job evaluation in order to be seen to be responding to legislative demands for pay equity between men and women. Job evaluation, carried out analytically, transparently, and with the participation of those concerned, normally provides a defence against potentially very expensive legal action.

Job evaluation will continue to change, but is unlikely to disappear. Employees expect equitable treatment. Even in an age of flexibility and individualism, it remains the most reliable means of achieving a pay structure that is perceived as both rational and fair by employees and their managers.

Bibliography

Arvey R D 1986 Sex bias in job evaluation procedures. *Personnel Psychology* **39**: 315–35

Ash P 1947 The reliability of job evaluation procedures. *Journal of Applied Psychology* **32**: 313–20

Baker H, True J M 1947 *The Operation of Job Evaluation Plans: A Survey of Experience*. Princeton University, Industrial Relations Section, Princeton, NJ

Beer M, Spector B, Lawrence P R, Mills D Q 1985 *Human Resource Management*. Collier Macmillan, London

Gerhart B, Milkovich G 1992 Employee compensation: research and practice. In: Hough L M, Dunnette M G (eds.) *Handbook of Industrial and Organizational Psychology*. Consulting Psychologists Press, Palo Alto, CA, Vol. 1, pp. 481–569

Hutner F C 1986 *Equal Pay for Comparable Worth*. Praeger, New York

Kerr C, Fisher L H 1977 Effects of environment and administration on job evaluation. In: Kerr C (ed.) *Labor Markets and Wage Determination*. California University Press, Berkeley, CA, pp. 53–86

Lawler E E 1986 What's wrong with point-factor job evaluation. *Compensation and Benefits Review* **18**: 20–8

Lawler E E 1994 From job-based to competence-based organizations. *Journal of Organizational Behavior* **15**: 3–15

Madigan R M, Hoover D J 1986 Effects of alternative job evaluation methods on decisions involving pay equity. *Academy of Management Journal* **29**: 84–100

Milkman R 1987 *Gender at Work: The Dynamics of Job Segregation by Sex during World War II*. University of Illinois Press, Urbana, IL

Pfeffer J, Salancik G R 1974 Organization decision making as a political process: The case of a university budget. *Administrative Science Quarterly* **19**: 135–51

Robinson D D, Wahlstrom O W, Mecham R 1974 Comparison of job evaluation methods. *Journal of Applied Psychology* **59**: 633–7

Rogers R C 1946 Analysis of two point rating job evaluation plans. *Journal of Applied Psychology* **30**: 579–85

Smith B N, Hornsby J S, Benson P G, Wesolowski M 1989 What is in a name: The impact of job titles on job evaluation results. *Journal of Business and Psychology* **3**: 341–51

Treiman D J, Hartmann H I (eds.) 1981 *Women, Work, and Wages: Equal Pay for Jobs of Equal Value*. National Academy Press, Washington, DC

Viteles M S 1941 A psychologist looks at job evaluation. *Personnel* **17**: 165–76

Welbourne T M, Trevor C O 2000 The roles of department and position power in job evaluation. *Academy of Management Journal* **43**: 761–71

A. J. Arthurs

Job Design, Psychology of

Jobs are created by people for people. Whether deliberately or by default, choices are made about which tasks to group together to form a job, the extent to which job holders have to follow prescribed procedures in completing those tasks, how closely the job incumbent will be supervised, and numerous other aspects of the work. Such choices are the essence of job design, which may thus be defined as 'the specification of the content and methods of jobs' for individuals or teams' (Wall 1995). Other terms often used as synonyms for job design include 'job' and 'work restructuring,' 'work design,' and 'work organization.' It is difficult to distinguish among these terms, although those who employ the word 'work' in preference to 'job' tend to take a broader perspective linking the design of jobs more explicitly to the wider organizational context. Psychological interest in this topic is concerned with the effects of job design on employee attitudes and behavior, and with the technological and organizational factors that influence the choice of job design. In this article, the psychology of job design is described with regard to its main research emphases, historical context, proposals for job redesign, key theoretical approaches, research findings, and current issues and future directions.

1. Main Emphases in Job Design Research

In principle, the concept of job design covers all types of work and job properties. Within applied psychology, however, a more particular perspective has developed. First, with respect to types of work, attention has centered mainly on lower-level jobs in larger enterprises, e.g., those involving office and especially shopfloor manufacturing work. Second, with regard to job properties, the focus has been on key generic job characteristics, such as the variety of tasks in jobs and the amount of discretion or autonomy job incumbents have in completing those tasks. Those emphases reflect the history of job design practice, as described next.

2. Historical Context

Since the turn of the twentieth century, the trend in job design within enterprises has been one of 'job simplification' or 'deskilling.' The move from small craft-based industries to larger factories, the emergence of mass production, and the ideas of Adam Smith (1776) and Charles Babbage (1835) on the 'division of labor,' together with Frederick Winslow Taylor's (1911) exposition of the 'principles of scientific management,' are among the influences which led to the design of narrow jobs with closely prescribed tasks. (The above seminal writings are now difficult to find, but key extracts, along with other historically important contributions, are reproduced in Davis and Taylor 1972). The reasoning was that simplifying work in this way would reduce costs by minimizing the risk of errors, allowing less skilled labor to be used, and reducing training requirements. The epitome of this process is the assembly line, where jobs can be so simplified that they involve the continuous repetition of a single operation with a short cycle time, no discretion over how to carry out the task, and no control over the pace of work.

Concern about the human costs of job simplification inspired some of the earliest research in applied psychology. In the UK this was the focus of work conducted during the 1920s under the auspices of the government-funded Industrial Fatigue Research Board. Employees in highly repetitive jobs, such as cigarette making, tobacco weighing, and bicycle-chain assembly, were shown to be bored and dissatisfied with their work (Wall and Martin 1994). As this area of inquiry developed in the UK, the USA, and elsewhere over the next few decades, evidence also began to accumulate of more serious consequences in terms of a link between repetitive work and employee stress or mental health (e.g., Kornhauser 1965). Studies in the 1950s and 1960s extended the agenda by considering also how the restriction of job discretion or autonomy inherent in job simplification affected job holders, and showed similar and often stronger effects.

3. Job Redesign

Evidence of the negative consequences of job simplification fostered proposals for 'job redesign.' This normative term reflects the historical background outlined above. Job simplification is cast as the traditional form of job design, and job redesign is used to denote deliberate attempts to reverse its deleterious effects by building into jobs more task variety, autonomy, or associated properties. Understandably, suggestions for job redesign paralleled the history of job design research. Thus the earliest proposals focused on reducing repetitiveness by increasing the number of different tasks experienced by employees. One such proposal was for 'job rotation,' which involves moving employees at regular intervals between different (simplified) tasks. Another was for 'horizontal job enlargement,' which increases task variety by including a wider range of tasks in jobs.

Later job redesign proposals in the 1960s and 1970s were for 'job enrichment' or, as it is less commonly called, 'vertical job enlargement.' This approach reflected the concern about the low levels of discretion in simplified jobs, and focused on increasing employee autonomy over the planning and execution of their own work (e.g., by giving job holders responsibility for decisions that otherwise would be undertaken by support and supervisory staff). The term job enrichment, originally coined by Herzberg (1966) to denote the approach derived from his 'two-factor theory,' is now used more generally. That theory distinguished between intrinsic or 'motivator' job factors, such as achievement and responsibility; and extrinsic or 'hygiene' factors such as supervision and work conditions. The core proposition was that intrinsic factors were the key determinants of satisfaction and performance, hence the emphasis on redesigning jobs to promote these properties. A final proposal, aimed at enhancing the discretionary component of work at the team level, stems from the sociotechnical systems approach to work organization, and is for the implementation of 'autonomous work groups' or 'self-managing teams' (see below). It is effectively job enrichment for teams.

4. Major Theoretical Approaches

Two theoretical approaches have dominated research on job design, and have yet to be superseded. One of these, concerned with individual job design, is the 'Job Characteristics Model' (Hackman and Oldham 1976). This identifies five 'core job characteristics,' namely skill variety, task identity, task significance, autonomy, and feedback from the job itself, as determinants of work motivation, job satisfaction, work performance, labor turnover, and absence. The strength of the effects of the job characteristics on the outcomes is predicted to be affected by individual differences, in particular being stronger for employees with greater

'growth need strength' and (in later formulations) also for those with higher contextual satisfaction and more knowledge, skill, and ability.

The other dominant approach derives from the more general sociotechnical systems approach to organizational design. The emphasis in this case is on the design of work for teams, with the key proposal being for the implementation of 'autonomous work groups.' Six key criteria specified for such groups are that the work should be reasonably demanding and provide variety, afford the opportunity to learn and continue learning, include an area of decision making that employees can call their own, offer social support and recognition, be of wider social relevance, and lead to a desirable future (Cherns 1987). There is no particular acknowledgment of individual differences in the sociotechnical approach, but otherwise the specified properties of well-designed jobs are very similar to those of the Job Characteristics Model.

5. *Research Strategies*

Research on job design has been of two main types: cross-sectional field studies examining the relationship between job characteristics and outcomes, and change or intervention field studies investigating the impact of job redesign on outcomes. The latter are exemplified by two field experiments described by Wall, Clegg and colleagues (Wall and Martin 1994). In the first, previously simplified individual jobs in a confectionery department were replaced by autonomous group working. Team members were given greatly expanded responsibilities, such as allocating tasks among themselves and resolving every-day operational problems, within the constraints of agreed production targets and health and safety requirements. Effects were investigated over a period of 18 months and showed substantial increases in output and job satisfaction, and a reduction in strain. The second study compared the introduction of autonomous work groups in a greenfield site with a sister site within the same parent company which operated with a more traditional job design. Once again job attitudes (e.g., job satisfaction) were found to improve, but in this case output did not (though productivity was higher because of fewer support staff?).

The difference between these studies highlights a major feature of research findings more generally. Whereas the effects of job redesign on job attitudes appear highly reliable, the consequences with regard to performance are less consistent (Parker and Wall 1998).

6. *Current Issues and Future Directions*

Interest in job design, which peaked during the 1970s, making it then one of the most prominent topics of inquiry within industrial and organizational psychology, waned during the 1980s. However, at the start of the twenty-first century it is resurfacing in response to new strategies, practices and technologies emerging in manufacturing and elsewhere. Organizations see improved quality, flexibility, and responsiveness to customer demand as the route to greater competitiveness, as well as cutting costs, and are supporting this through the use of computer-based technology, just-in-time inventory control, total quality management, business process re-engineering, delayering, and other initiatives. It is widely appreciated that the success of these depends on their being supported by appropriate job designs.

This renewed emphasis on job design has been further strengthened by other developments, and in particular by more general proposals for 'high performance work systems,' 'high involvement organizations,' and 'empowerment' (e.g., Lawler 1992, Hardy and Leiba-O'Sullivan 1998). These proposals incorporate the traditional core principles of job design, inasmuch as they recommend that employees take on a wider range of tasks and be given much more autonomy and discretion in their work. At the same time they adopt a wider perspective than that usually encompassed within job design theory and research. In so doing they help to make explicit the limitations of existing knowledge and clarify directions for future development.

Several areas of development are high on the agenda (Parker and Wall 1998). The first concerns the relationship between job design and performance. Traditional job design theory, and its manifestation within the broader notions of high-involvement organizations and empowerment, is implicitly universalistic in this regard. It assumes positive effects of increased autonomy on performance irrespective of external circumstances. Yet, as noted above, evidence shows the impact of job redesign on performance is variable. Such inconsistency points to the existence of unspecified contingencies that need to be identified. One such contingency, suggested by organizational theory and recent research findings, is that of 'operational uncertainty.' It is likely that providing autonomy for employees benefits performance most under changing and unpredictable work conditions, but yields less benefit as requirements become more standard and predictable (Wall and Jackson 1995).

A second line of development concerns the need to investigate the effect of job design on a wider range of outcomes. Most studies focus on outcomes such as job satisfaction, organizational commitment, and mental health. Considerably less attention has been given to the consequences of job designs for health and safety (e.g., accidents, overexertion injuries, incidents of violence), employee learning and development (e.g., the development of knowledge, personal initiative), and outside-work outcomes (e.g., leisure activities, family relationships).

A third area for development is to broaden the range of work characteristics encompassed. At present the job characteristics considered are those stemming from traditional forms of work, in particular simplified jobs within manufacturing settings. The growth of service and knowledge work, the emergence of new computer-based technologies, and the opportunities being opened up by telecommunications suggest that a range of other job properties likely to affect employee well-being and performance will come into play. For example: many computer-based systems can substantially reduce the physical demands of work, but enhance learning and problem-solving demands; teleworking can accentuate social isolation; modern manufacturing initiatives such as lean production can increase employee work load and time pressure; and new forms of 'contingent working' (e.g., short-term, project-based contracts) raise questions of security and employment continuity. Job design theory will need to adapt to such changes.

Job design theory will also need to adapt to changes in the work force itself. Increasingly, employees are 'less young, less male, and less white' (Howard 1995, p. 33). Existing work design theory takes little account of age, gender, and race, yet these demographic factors could have an important influence on people's requirements for, and responses to, job design. For instance, autonomy over working hours might be especially key for working women with children because it allows them more easily to juggle work demands with domestic and child-care tasks (for which women still typically have the major responsibility). For older employees working in physically demanding jobs, autonomy over work breaks might be especially important because it allows for rests to alleviate any physical strain.

Perhaps the least-well understood aspect of job design is that concerned with the mechanisms, or processes, through which job properties affect outcomes. Historically, job design theory and practice have been founded on motivational assumptions, that gains from job redesign result from enhancing the effort people put into their work. More recent research suggests that motivation does not provide a complete explanation, and that consideration needs also to be given to the effect of job design for peoples' understanding of their work, their proactivity, their work orientations, and their feeling of self-efficacy (Parker et al. 1997, Parker 1998).

A final issue concerns the role of job design principles in the design of new systems. The emphasis within the job design literature has been on how jobs can be better adapted within their technological context. However, the technology may itself preclude job design alternatives that might be of value. It is equally important to ask how knowledge of job design might be exploited to design better systems in the first place, and to encourage social scientists and users to work alongside development engineers in the design and implementation of new systems (Clegg et al. 1996).

See also: Engineering Psychology; Job Analysis and Work Roles, Psychology of; Job Design and Evaluation: Organizational Aspects; Psychological Climate in the Work Setting; Stress at Work; Worksite Health Promotion and Wellness Programs

Bibliography

Babbage C 1835 *On the Economy of Machinery and Manufacturers*. Charles Knight, London

Clegg C W, Coleman P, Hornby P, Maclaren R, Robson J, Carey N, Symon G 1996 Tools to incorporate some psychological and organisational issues during the implementation of computer-based systems. *Ergonomics* **39**: 482–511

Cherns A 1987 The principles of socio-technical systems design revisited. *Human Relations* **40**: 153–62

Davis L E, Taylor J C 1972 *Design of Jobs*. Penguin, Baltimore

Hackman J R, Oldham G R 1976 Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance* **15**: 250–79

Hardy C, Leiba-O'Sullivan S 1998 The power behind empowerment: Implications for research and practice. *Human Relations* **51**: 451–83

Herzberg F 1966 *Work and the Nature of Man*. World Publishing, Cleveland, OH

Howard A (ed.) 1995 *The Changing Nature of Work*. Jossey-Bass, San Francisco

Karasek R, Theorell T 1990 *Healthy Work: Stress, Productivity and the Reconstruction of Working Life*. Basic Books, New York

Kornhauser A 1965 *Mental Health and the Industrial Worker*. Wiley, New York

Lawler E E 1992 *The Ultimate Advantage: Creating the High Involvement Organization*. Jossey-Bass, San Francisco

Parker S K 1998 Enhancing role breadth self-efficacy: The roles of job enrichment and other organizational interventions. *Journal of Applied Psychology* **83**: 835–52

Parker S K, Wall T D 1998 *Job and Work Design: Organizing Work to Promote Well-being and Effectiveness*. Sage, Thousand Oaks, CA

Parker S K, Wall T D, Jackson P R 1997 'That's not my job': Developing flexible employee work orientations. *Academy of Management Journal* **40**: 899–929

Smith A 1974 [1776] *The Wealth of Nations*. Penguin, Harm-ondsworth

Taylor F W 1911 *The Principles of Scientific Management*. Harper, New York

Wall T D 1995 Job design. In: Nicholson N (ed.) *Encyclopedic Dictionary of Organizational Behavior*. Blackwell, Cambridge, MA

Wall T D, Jackson P R 1995 Changes in manufacturing and shopfloor job design. In: Howard A (ed.) *The Changing Nature of Work*. Jossey-Bass, San Francisco

Wall T D, Martin R 1994 Job and work design. In: Cooper C L, Robertson I T (eds.) *Key Reviews in Organizational Behavior: Concepts, Theory and Practice*. Wiley, Chichester, UK

T. D. Wall and S. Parker

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Job Stress, Coping with

Coping with job stress is a key concept in understanding people's adaptation to their work roles. In their jobs employees are confronted with various kinds of demands, which may become 'stressors' when they tax or exceed the employee's adaptive capabilities. Examples of common job stressors include work overload, role problems, poor job control, lack of support from supervisors and co-workers, and interpersonal conflicts. These stressors may lead to negative psychological (e.g., depression, irritability, burnout), physical (e.g., headaches, heart palpitations, hyperventilation) and behavioral (e.g., absenteeism, turnover, violence) symptoms or 'strains.' The psychological process that is involved in dealing with these job stressors in order to reduce strain is called 'coping.' More specifically, coping with job stress refers to the things employees do to handle the stressors they encounter in their work roles (see *Stress and Coping Theories*). These individual coping efforts may or may not be supported by the employing organization.

1. *The Concept of Coping*

Rather than a homogeneous concept coping represents a diffuse umbrella term. Nevertheless, a widely used definition of coping exists that refers to the constantly changing cognitive and behavioral efforts a person makes to manage demands that tax or exceed their personal resources (Lazarus and Folkman 1984). Coping refers to an interplay between person and environment that occurs when an individual appraises a situation or an event as stressful. As far as coping with job stress is concerned, this interplay involves the employee and his or her job environment or work setting that is appraised as stressful. It is important to note that coping refers merely to the effort made and not to the quality of the outcome. That is, coping may be either effective or successful in preventing, avoiding, or controlling individual distress or it may be ineffective or unsuccessful. For instance, compensatory or comfort behaviors such as excessive eating, smoking, drinking, or working commonly are cited as examples of maladaptive coping strategies. Furthermore, coping may refer to overt actions or observable behaviors as well as to covert, unobservable thoughts and events.

2. *The Historical Development of Coping*

As far as the conceptualization of coping is concerned, three approaches can be distinguished. Initially, coping was studied from a psychodynamic point of view in the general context of human adaptation. In the first half of the twentieth century, leading psychoanalysts like Freud, Adler, and Jung described adjustment and defense mechanisms by which individuals deal with intrapsychic conflict. Building upon this earlier work, in the 1960s and 1970s ego-psychologists distinguished between coping as an advanced or mature ego process leading to successful adaptation and defense as an unsuccessful neurotic mode of adaptation.

The second approach emerged from the 1960s and conceived coping in terms of personality traits. Relatively stable dispositions or coping styles (i.e., denial, repression, vigilance) were expected to determine the individual's coping behavior across various situations. Although empirical research has shown that people possess stable and consistent coping preferences it is likewise true that individuals are able to vary their responses to stressful situations in accordance with the demands posed by these situations. Personality traits have made a comeback in the 1970s and 1980s as so-called personal coping resources like 'hardiness,' 'dispositional optimism,' 'internal locus of control,' 'self-efficacy,' 'sense of coherence,' and 'self-esteem.' Such coping resources were expected to foster successful coping since the individual can draw upon them when under stress.

The currently dominant transactional approach originates from the work that Lazarus and his colleagues have performed since the 1950s. In contrast to the psychoanalytic approach, coping is defined in this approach by an effort that is independent from its success and includes pro-active as well as defensive strategies. Furthermore, coping is considered to be a dynamic and transactional process involving person as well as environment rather than the result from a relative personality disposition.

Coping has been applied to various aspects of human life. Initially, it was used predominantly in clinical and health psychology to understand how patients deal with chronic disease, substance use, loss, chronic pain, and trauma. More recently, minor coping efforts involving so-called daily hassles have been investigated. Coping with job stress (including unemployment) has been studied since the end of the seventies and involves long-term chronic job stressors (e.g., future job insecurity), short-term daily hassles at the job (e.g., interpersonal conflict), as well as acute severe stressors (e.g., armed robbery).

3. *General Coping Theory*

The leading cognitive-behavioral theory of stress and coping of Richard Lazarus and his co-workers (see *Stress and Coping Theories*) is usually applied to work situations as well. It assumes that a particular situation or event is appraised negatively as stressful in terms of threat, loss, or harm. Alternatively, the situation or event may also be appraised neutral or positive as a

challenge. In this so-called 'primary appraisal,' the person asks 'What do I have at stake in this encounter?' For instance, in a conflict with a colleague the employee's self-esteem may be threatened, whereas unemployment is likely to be perceived as a loss, rather than a challenge. In 'secondary appraisal' the concern of the person is 'What can I do? What are my options for coping?' The answer depends to a large extent on personal characteristics, such as motivation, beliefs about oneself and the world, and recognition of personal and situational resources for coping such as financial means, social and problem-solving skills, particular personality traits, and a favorable work situation that includes for instance social support from one's colleagues and supervisor. It is important to note that both appraisal processes—primary (demand) appraisal and secondary (resource) appraisal—are two components of one cognitive process that take place simultaneously, instead of one process (secondary appraisal) taking place after the other (primary appraisal).

Generally speaking a main distinction is made between two targets: (a) problem-focused coping, which involves attempts to deal with stress by directly altering the situation and (b) emotion-focused coping, which involves attempts to regulate the emotional responses to a stressful situation. For instance, employees might consult their supervisors after a conflict with a colleague (problem-focused) or they may try to relax or distract themselves (emotion focused). In addition, various methods of coping are distinguished such as active–cognitive (i.e., intrapsychic attempts to manage the appraisal of the stressor through, for instance, constructive dialogue or positive comparison), active–behavioral (i.e., overt attempts to deal directly with the stressor), and avoidance (i.e., avoiding or denying the stressor). Although targets and methods of coping can be distinguished theoretically this distinction is usually quite difficult to make at the practical level.

After the coping effort has been completed, a reappraisal of the initial stressful event takes place along the same lines as the original primary appraisal: 'Is there still something at stake in this encounter?' The conflict with one's colleague may be solved or one may have found a new job. In that case the reappraisal is positive and there is no need for further coping efforts. However, in case initial coping efforts failed and the event is reappraised as stressful coping attempts continue.

4. *The Individual and Organizational Perspective*

In Fig. 1 a taxonomy of individual ways of coping with job stress is displayed, including some examples of coping behaviors. Typically, in occupational settings workplace strain and coping are individual *as well as* organizational issues. This means that interventions to reduce workplace strain may be targeted at the individual or the organizational level (Quick et al. 1997). Interventions that focus on the individual include job stress education activities, relaxation programs, cognitive-behavioral programs, and employee skills training in such areas as time management and assertiveness (see *Stress Management Programs*). Interventions that focus on the organization seek to reduce strain by changing and improving macrolevel factors operating within organizations (e.g., job-redesign, selection and training policies, organizational development). Although organization-based interventions offer considerable potential for reducing or eliminating job stressors, most workplace initiatives focus on the individual employee. That is, their adaptability to the existing work environment is improved by changing their behavior and improving their life-style or stress management skills and augmenting their coping resources. In the last decades different programs have been developed to increase individual stress resilience and promote health and wellbeing at the workplace, including on-site fitness facilities, health screening, dietary control, smoking cessation, cardiovascular fitness programs, relaxation classes, stress-management training, antiburnout workshops, stress and health education, and psychological counseling.

Generally speaking, individual-based interventions, particularly relaxation programs and cognitive-behavioral programs, seem to be effective in reducing employee strain (i.e., burnout, distress, depressed mood, anxiety, and psychosomatic symptoms), whereas the effectiveness of organization-based strategies is still to be demonstrated unambiguously.

5. *Research*

Earlier studies on coping with job stress often found that employees use fewer problem-focused than emotion-focused strategies directed at changing the problems encountered in the context of work. This was interpreted by noting that the very nature of problems at work are perhaps not ameliorated easily by attempts to change them. It was reasoned that employees are perhaps better off by developing a renewed cognitive perspective and by regulating their negative emotions. More recent studies, however, point to the opposite: currently employees seem to favor problem-focused coping instead of emotion-focused coping. More detailed investigations have shown that the level of job control plays a crucial role. When a high level of job control exists, problem-focused strategies prevail, but when job control is poor employees seem to favor emotion-focused strategies. Unfortunately, the latter are generally less effective compared to the former. For instance, studies on organizational changes, such as corporate restructuring and downsizing, suggest that employees cope

METHOD	Problem/Task	Emotions/Reactions
Cognitive	Control Planning, organizing, and prioritizing assignments	Control Try to think of oneself as a winner, as someone who always comes through
	Escape Try to pay attention only to your duties in order to overlook difficulties	Escape To tell oneself that difficulties are unimportant
Behavioral	Social Sit down and talk things out	Social To express ones irritation to other colleagues to let off steam
	Solitary To do what has to be done, one step at a time	Solitary Spend time on a hobby
	Control Delay or leave some of ones normal job responsibilities	Control Show no emotional reaction and do not communicate distress to anyone
	Escape Get busy with other things to keep ones mind off the problem	Escape Change to nonwork activity; take pills, smoke more

Figure 1
Ways of coping with job stress (adapted from Latack and Havlovic 1992)

rather poorly with these events which are largely beyond their control. They use predominantly emotion-focused and avoidant coping strategies and suffer long-term adverse mental health.

As a general rule, the more active, problem-focused, and control-oriented the coping strategy the less (mental) health symptoms and the higher employee wellbeing. In contrast, emotion-focused strategies as well as escape-oriented or avoidant strategies generally are associated with poor (mental) health and unwellbeing. However, two factors seem to be important for successful coping: time frame and availability of coping resources. It seems that emotion-focused or escape/avoidant-oriented strategies are effective when dealing with short-term stressors, whereas problem-focused and control-oriented strategies are more effective when dealing with chronic or recurrent job stressors. Furthermore, employees who have more coping resources at their disposal appear to use active coping efforts more and avoidant coping less. Consequently, employees who can draw upon coping resources are more likely to deal successfully with job stress.

As far as the relationship between employee coping and sociobiographical factors (e.g., age, gender, education) and job factors (e.g., level and type of job) is concerned, no conclusive empirical evidence is available (Cartwright and Cooper 1996). For instance, among employees no consistent gender-specific or job-specific coping patterns have been observed. Most likely this lack of conclusive evidence has to be attributed to the complex interaction patterns that exist among the factors involved.

6. *Methodological Issues and Future Directions for Research*

The measurement of coping is fraught with problems that are partly imputed to the fact that almost exclusively self-report questionnaires are used. More specifically, inductively or empirically developed coping scales yield different dimensions or coping behaviors compared to deductively or theory-based coping scales. Furthermore, the number of coping dimensions included in various questionnaires ranges

from two to 28 and there is neither agreement on the hierarchy of these dimensions nor on the fact whether coping should be assessed dispositionally (as a trait) or episodically (as a state). Most importantly, to date, no generally accepted measure exists to assess coping in an occupational context.

At the theoretical level, the crucial distinctions between coping and coping resources, between coping styles and coping behaviors, and between appraisal and (cognitive) coping needs further clarification.

See also: Occupational Health; Stress at Work; Workplace Safety and Health

Bibliography

Cartwright S, Cooper C L 1996 Coping in occupational settings. In: Zeidner M, Endler N S (eds.) *Handbook of Coping: Theory, Research, Applications*. Wiley, New York, pp. 202–20

Latack J C, Havlovic S J 1992 Coping with job stress: A conceptual evaluation framework for coping measures. *Journal of Organizational Behavior* **13**: 497–508

Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York

Quick J C, Quick J D, Nelson D L, Hurrell J J (eds.) 1997 *Preventive Stress Management in Organizations*. American Psychological Association, Washington, DC

W. B. Schaufeli

Joint Action

For the purposes of this article a 'joint action' is the action of two or more people. Thus, two or more people may hold a conversation, go for a walk (together), and so on. The participants will naturally refer to what 'we' did, as in 'We had a conversation,' and 'We went for a walk.' Joint actions are ubiquitous and it is hard to imagine a human life without them. At the same time it is not easy to say precisely what is required for the performance of a joint action. A variety of accounts have been envisaged by philosophers. This article explains some of the relevant considerations and issues, and reviews some of the central answers that have been given.

1. Some Initial Considerations on Joint Action

1.1 Joint Intentions are Integral to Joint Actions

Contemporary philosophy of action has tended to focus not on joint action but on the action of an individual person. A notion central to the philosophy of action is that of intention. It is commonly argued that what distinguishes an action (such as a man raising his arm) from a mere bodily movement (such as his arm rising) is an intention. When a man raises his arm, his arm rises in part by virtue of his intention to raise it. The sociologist Max Weber made a similar point when he characterized action as meaningful behavior.

Joint actions, similarly, appear to presuppose intentions, in this case joint intentions. Thus, for example, people cannot be conversing unless there is a joint intention to converse. The parties to a conversation act in light of a joint intention to converse, so as to implement that intention. A joint intention may be referred to by such statements as 'We intend to go shopping,'where our intending to go shopping is not, or is not simply, a matter of each of the people in question personally intending to go shopping. Precisely what is at issue is not obvious. Much of the growing body of philosophical work relevant to joint action is concerned with the nature of joint intention, or, in less formal terms, what it is for us to intend.

1.2 Joint Action Must be Cooperative Only in a Broad Sense

Joint action is not necessarily a matter of amicable cooperation. We may be quarrelling with one another or engaged in a fight. Quarrelling and fighting are species of joint action. Though it need not be amicable, there is clearly a broad sense in which joint action is always cooperative or collaborative. I cannot quarrel with you unless you quarrel with me; I cannot fight with you unless you fight with me. If you refuse to take me seriously we cannot quarrel. If you lay down your arms, we are no longer fighting.

1.3 Some Consequences of Involvement in Joint Action

Informal observation of the way people behave when doing things together makes it clear that the parties often rebuke one another for behavior contrary to the relevant joint intention. Such rebukes are commonly understood to be in order. For example, Jill and John are walking together, John begins to lag behind, and Jill rebukes John for doing so. John understands that, in the context of their walking together, Jill has the standing to rebuke him.

Less punitive responses are also commonplace, and routinely understood to be in order. For example, Jill asks John why he is lagging behind. John may not like being questioned like this, but he understands that Jill has a legitimate interest in the matter.

Underlying these responses is the understanding that, all else being equal, participation in a joint intention entitles each of the parties to relevant actions

from the others. If one party does not produce the relevant actions, therefore, the others have the standing to question and to rebuke that party, as they see fit.

A further point is this. Suppose John and Jill jointly intend to go for a five mile walk. At some point along the way John decides that this will be too tiring for him. He may then announce that he is turning back, and he may indeed turn back. If he does these things, then, failing some special background circumstances, he will be acting contrary to understandings that have been set up between him and Jill. It would be different if he said or did something that amounted to asking for and gaining Jill's permission for his turning back.

In general terms, no one party to a shared intention is in a position to cancel or change that intention unilaterally. Though I am in a position unilaterally to change my own mind, I am not in a position unilaterally to change our mind, that is, to change our joint intention.

1.4 Inferences from Statements about What 'We Intend'

A final observation on joint action is this. From 'We intend to walk for five miles' and 'In order for us to walk for five miles, I must walk for five miles,' it seems that John can infer that he has reason to walk for five miles. In other words, our intention can give me a reason for acting. It is at least not obvious that there need be any interposition in this reasoning of a premise about what John himself intends.

2. Two Central Questions

A satisfying account of joint action will incorporate an explanation of the observations just given. Two relevant questions deserve particular emphasis.

2.1 What is the Relation of Joint Intentions to Personal Intentions?

Are the joint intentions that underlie joint actions basically concatenations of personal intentions of the participants, perhaps personal intentions in favor of the joint action? This may be the conjecture that most immediately springs to mind. As we shall see, however, serious problems attend this idea.

2.2 Of What Kind are the Entitlements Associated with Joint Intention?

The parties to a joint intention understand their participation to entitle each of them to actions in conformity with the relevant joint intention. How is this understanding to be explicated? In particular, what kind of entitlement is at issue? One might assume that it is moral entitlement founded on moral principle. On the other hand, it may seem that joint intention is too basic and primitive a phenomenon to require moral understanding in all participants. It is certainly not obvious that it does require such understanding.

3. Some Problems with Accounts of Joint Intention in Terms of Meshing Personal Intentions

A carefully constructed account of joint intention that puts personal intentions at center stage is that developed by Bratman (Bratman 1999 contains a number of relevant articles). This may serve as an example of this type of account for the present purposes.

A somewhat simplified version of Bratman's account runs as follows. We jointly intend to go for a walk if and only if (a) (i) I intend that we go for a walk, (ii) you intend that we go for a walk. (b) I intend that we go for a walk in accordance with and because of (a) (i) and (b) (ii); you intend likewise. (c) (a) and (b) are common knowledge between us.

Bratman does not define the 'common knowledge' that he requires in condition (c). What is intended is, roughly, that (in the example) the fact that conditions (a) and (b) are fulfilled is completely out in the open as far as the two people involved are concerned. (For the first published discussion of 'common knowledge' see Lewis 1969.)

There are several problems with this account of joint intention. Two will be detailed here. Both are common to accounts that put personal intentions of the participants at the foundation of joint intention and hence joint action.

One problem has to do with the entitlements associated with joint intention. It does not appear that anyone is entitled to relevant actions from anyone else just by virtue of the satisfaction of the conditions laid down in the account. My intending that we go for a walk, and intending that this occur by virtue of my so intending and your so intending, does not seem to entitle you to any particular action from me. Nor does the fact that my various intentions are entirely out in the open between us seem sufficient to bring any such entitlement into existence.

If you know that I intend that we go for a walk, and have reason to think that I won't change my mind, then there is a sense in which you are 'entitled to expect' that I am going to make the appropriate contribution to our walking together. But this just means that you have a good basis for predicting that I am going to act thus. It does not mean that you are entitled to my acting thus in such a way that questioning or rebuking me is in order.

In related discussion, Bratman has invoked a complex moral principle to argue for the presence of the relevant entitlements in many situations in which his conditions on joint intention hold. The principle in

question is the 'principle of fidelity' proposed in Scanlon (1990). This principle runs as follows (Scanlon 1990, p. 208): If (a) A voluntarily and intentionally leads B to expect that A will do x (unless B consents to A not doing x); (b) A knows that B wants to be assured of this; (c) A acts with the aim of providing this assurance, and has good reason to believe that he or she has done so; (d) B knows that A has the beliefs and intentions just described; (e) A intends for B to know this, and knows that B does know it; and B knows that A has this knowledge and intent; then, in the absence of some special justification, A must do x unless B consents to x not being done.

Scanlon's principle of fidelity is not applicable in all situations where there are joint intentions on Bratman's account of them. So appeal to this principle will not rescue Bratman's account of joint intention insofar as joint intention by its nature entitles its participants to conforming actions.

Another problematic aspect of Bratman's type of account is that it does not explain the need for the concurrence of all parties in order to dissolve the joint intention. One with a personal intention is in a position to change it unilaterally. It follows that an individual who is party to a joint intention in Bratman's sense is in a position unilaterally to do away with one of the foundational personal intentions at any time. There is no need for the permission of the other party or parties. Thus a joint intention in this sense appears to be something relatively flimsy, flimsier than joint intentions as these are ordinarily understood.

4. *The Irreducibility of 'We Intend to Do Such-and-such'*

John Searle has claimed that at the core of joint intention and action lie irreducible thoughts of the form 'We intend to do such-and-such' in the mind of each participant (see, for instance, Searle 1995; see also Sellars 1963). These thoughts are irreducible in the sense that they cannot be reduced or broken down into thoughts about what 'I intend' and what 'you intend.' Searle's central concern is to describe the inner or subjective side of joint intention with respect to each participant. He evidently believes that there are requirements for the correctness of thoughts of the form 'We intend to do such-and-such,' though he does not focus on this issue.

What he writes at one point suggests that, in his view, one person's thought of the form 'We intend to do such-and-such' is correct as long as the other person or persons in question also think that thought with respect to the same people. This proposal does not seem to be right, if 'We intend to do such-and-such' is understood in accordance with everyday understandings about joint intention.

Suppose that Ben is currently thinking, with respect to himself and Elaine, 'We intend to get married.' Elaine is in a similar position. According to the proposal under discussion, it follows that each one is correct in thinking 'We intend to get married.' Yet according to everyday understandings it would not be plausible to claim that, solely by virtue of the facts described, Ben and Elaine jointly intend to get married. If one learned that they had never discussed the matter one would judge that their thoughts were not properly grounded.

5. *A Role for Joint Commitment*

5.1 *Joint Commitment*

When is one right in thinking an irreducible thought of the form 'We intend to do such-and-such,' according to everyday understandings? Margaret Gilbert has proposed that, in a sense she explains, the parties need to have established a joint commitment to intend as a body to do such-and-such (see Gilbert 1999).

The relevant notion of commitment may be introduced by reference to a personal decision. If Jane decides to go shopping this afternoon, she is then subject to a personal commitment to go shopping. She is not subject to such a commitment if she merely inclines to go shopping or hopes to do so.

According to Gilbert, people can be subject both to personal and to joint commitments. A joint commitment in Gilbert's sense is not an aggregate or conjunction of corresponding personal commitments of each party. Rather, it is a commitment of all the parties, or of the whole. By its nature a joint commitment requires the participation of each party for its creation and its dissolution.

Gilbert has argued that the idea of a joint commitment is a fundamental concept that underpins not only joint intentions and goals but also joint beliefs, and social rules and principles (see Gilbert 1989, 1996, 1999). Those engaged in joint action act in light of the joint commitment underlying the relevant joint intention or goal.

There are both basic and derived forms of joint commitment. In order to create a joint commitment of the basic type, each party must express to the rest his or her readiness to be committed with them through a joint commitment whose content is specified. When this has been done in conditions of common knowledge, the relevant joint commitment is in place. In general, when a joint commitment is formed the parties to it are committed simultaneously. They are committed through the joint commitment. Similarly, no one can be rid of their commitment through the joint commitment unless the others are. Once it is rescinded for one it is rescinded for all.

Without prior authorization, no one can unilaterally rescind or dissolve a joint commitment: all must rescind it together. Otherwise it stands as a constraint

on the parties. A joint commitment can of course be violated by one party acting alone, but this is precisely a matter of violation as opposed to dissolution. Unlike those who rescind or dissolve it, the violator acts against a standing joint commitment.

Gilbert construes the idea of being jointly committed to intend as a body somewhat as follows. The parties are jointly committed to constitute together, as far as is possible, a single body that intends to do such-and-such. Though they are plural, or many, they are jointly committed to make themselves as much like a singular subject of intention as they can. A joint commitment to intend as a body is sufficient to create, in Gilbert's terminology, a 'plural subject' of intention. In other words, in her terms, we intend to do such-and-such if and only if we constitute a plural subject of intention.

How do people express to one another their readiness to be jointly committed to intend as a body to do such-and-such? There are many such ways. Among these are such humdrum exchanges as 'Shall we go for a walk this afternoon?' with a following 'Sure, let's.'

Joint commitments are not necessarily set up through verbal exchanges. Two men who jump into a rowing boat in hot pursuit of a fugitive may signal by their attempts to coordinate the movements of the oars that, in effect, they stand ready to be jointly committed to intend as a body to row the boat. As each moves the oars in light of their ensuing commitment they engage in the joint action of rowing the boat.

In principle it may take some time for a particular joint commitment to be established. Each party's readiness to be party to the joint commitment may be made clear only gradually and over the course of many interactions with the other people in question.

5.2 Derived Joint Commitments

The foregoing examples of the formation of a joint commitment involve joint commitments of the basic type. A joint intention in Gilbert's sense may also involve a derived commitment, as in the following example.

Don asks his wife Sally 'Where do we plan to go on vacation this summer?' He presumes that though he is unaware of any such plan, there may already be such a plan, and Sally will know what it is. This presumption is based on the existence of a background commitment of the basic type. Don and Sally are jointly committed to accept as a body that Sally may make vacation plans for the two of them.

With this background commitment in place, Sally is in a position unilaterally to create joint vacation plans. In other words, by virtue of the background joint commitment and Sally's planning, Don and Sally may take on a derived joint commitment to plan or intend as a body to go to Spain this summer, for instance.

5.3 Large Populations

Members of large populations, such as the many inhabitants of a large island, are unlikely to know one another personally, or even to know of one another as individuals. If there are to be population-wide joint commitments in such populations, this will be through expressions of readiness to be jointly committed with all others with a certain common feature, such as 'those living on this island.'

By virtue of background joint commitments of the kind discussed in Sect. 5.2, the members of a large population can have a particular joint intention in Gilbert's sense though many of the members do not know of the intention in question. This accords with everyday understandings according to which it may be said that, for instance, a particular country intends a particular action, without it being assumed that everyone in that country knows of the intention, or, indeed, of the ensuing action.

Those who form intentions for the whole population may be authorized to ensure that these intentions are implemented. Their implementation need not involve all members of the population in order that the action in question be considered an action of the population as a whole.

5.4 Obligations and Entitlements of Joint Commitment

Gilbert has argued at some length that the concept of a joint commitment underlies a central concept of obligation, a concept of obligation for which there is a corresponding concept of entitlement or right (see Gilbert 1996, 1999). A brief indication of her position might note that if one violates a commitment that is joint, one is evidently answerable to the other parties to the commitment. They can observe that one has violated 'our' commitment as opposed to a commitment of one's own. They have a clear basis for questioning the violator and for administering rebukes on their own behalf. A natural way to characterize the situation in familiar terms is to say that the parties to a joint commitment are all entitled to conformity from the other parties. These others have corresponding obligations to conform: they are obligated to all parties to the commitment, their obligation being to conform to the commitment.

This and related arguments are arguments from the nature and structure of joint commitment. It may thus be best not to think of the relevant entitlements and obligations as moral entitlements and obligations. All that one needs to understand, in order to understand that they are there, is the nature of a joint commitment. One need not appeal to the value of anything such as, in the case of Scanlon's principle of fidelity, the value of what Scanlon refers to as 'assurance.'

Going back now to Gilbert's account of joint intention and action, it offers an explanation of how

those who are party to a joint intention have associated entitlements to conformity from one another. These are entitlements of joint commitment. There are derivative entitlements to ask questions, rebuke one another, and so on, should deviation from the joint intention be mooted. Gilbert can allow that in some situations where there are joint intentions it is morally permissible not to act in accordance with them. Such situations may include those where one party coerced another into participating in the relevant joint commitment, and those where the joint commitment is to intend something immoral. All things considered, then, it may be morally permissible not to fulfil the relevant obligations of joint commitment. This does not entail that those obligations no longer exist. They exist as long as the sustaining joint commitment exists (see Gilbert 1996).

5.5 *Joint Intentions as the Intentions of Plural Subjects*

Gilbert's plural subject account of joint intention and action does not appeal to personal intentions in favor of some joint action. She holds, in effect, that one who says 'We intend to do such-and-such,' referring to a joint intention, does not say something reducible to statements about what 'I' intend and what 'you' intend. To that extent she and Searle are in agreement.

The plural subject account of joint intention provides an explanation of how from an irreducible 'We intend to row to the other side' and 'My rowing more slowly than I am doing will best facilitate our rowing to the other side,' I can infer 'I should row more slowly.' My participation in the relevant joint commitment gives me reason to do what I can to implement the joint intention. It is not necessary, given this account, that I first form a personal intention that we row to the other side.

6. *Applications of the Plural Subject Theory of Joint Intention and Related Phenomena*

Concurring with the judgement of sociologists, Gilbert has argued that those who participate in joint intentions and actions constitute a social group, though the group may be a small and transient one (see Gilbert 1989). She argues, more generally, for the claim that plural subjects in her sense constitute an important, central kind of social group. Thus it can be argued that obligations of a special kind are part and parcel of group membership.

This line of argument makes it possible to bring Gilbert's plural subject theory of joint action and related phenomena to bear on a classic question of political philosophy, the question of political obligation. Is there an obligation to obey the laws of one's own country in particular, and, if so, what is the source of this obligation? Evidently, one possible source is participation in a joint commitment. Thus one might participate, for instance, in a joint commitment to accept as a body the rules prescribed by a certain person as the rules of the jointly committed population as a whole (see Gilbert 1996, 1999).

Gilbert's plural subject theory is also relevant to the topic of collective guilt in the sense of collective moral responsibility. According to Gilbert, an act is constituted as a joint act by virtue of an underlying joint commitment. That joint commitment unifies the people in question into a group. It may then sometimes be reasonable to ascribe moral responsibility for a joint act to a group as such, quite apart from the question of the moral responsibility of particular group members. This perspective on joint action also allows for the intelligibility of a common phenomenon: the experience of pride or guilt over the actions of one's group, as opposed to one's own actions (see Gilbert 1996).

7. *The Nature and Role of Agreements*

Another philosopher who has written at length on joint intention and action is Raimo Tuomela. In Tuomela (1984), he put what he called 'we-intentions' at the center of an account of joint action. Each party to a joint action must have a relevant we-intention. Tuomela's complex definition of a we-intention runs roughly as follows: a member Ai of a collective G we-intends to do X, where X is some act of G, if and only if (a) Ai intends to do his own part of X, given that every member of G will do X his own part of X; (b) Ai believes that every member of G will do his own part of X; (c) there is a mutual belief in G to the effect that (a) and (b).

Tuomela (1995) introduces the 'underlying assumption that we-intentions are formed in a process involving the making of an explicit or implicit agreement (or at least believed agreement)' (p. 127). Agreements determine what joint action is afoot; we-intentions ensue. If all goes as the parties we-intend, joint actions follow.

Clearly, agreements are a paradigmatic context for the production of joint intentions and actions. However, as the example of the rowers suggests, agreements as these are ordinarily understood and are not essential. If one says that there is an implicit agreement in such cases, the question arises as to what precisely this claim amounts to, since an implicit agreement does not seem to be an agreement proper. One might do well to start with an account of paradigmatic agreements. There has been less concentrated philosophical focus on such agreements than might have been expected, given the undoubted importance of agreements within the scheme of everyday social concepts. A standard claim, perhaps echoing statements in contract law, has been that an agreement

is an exchange of promises. Gilbert has argued at length against this view, on one interpretation of it, and in favor of the proposal that typical everyday agreements are, in effect, joint decisions, where a joint decision is constituted by a joint commitment to endorse as a body the decision in question (see in particular Gilbert 1996, Chap. 13).

Suppose that Phyllis and Joe jointly commit to endorse as a body the decision to get married next year. Presumably it is reasonable then to regard them as jointly committed to intend as a body to get married next year. Given Gilbert's account of agreements and joint intentions, then, agreements give rise to joint intentions, providing a basis for joint action. Agreements are not necessary for joint action, however, insofar as a joint commitment to intend as a body can be brought into being in the absence of a joint commitment to endorse as a body a certain decision. This accords with the way the relationship of agreements to joint actions is commonly understood.

8. *Individualism, Holism, and Joint Action*

Annette Baier has objected to the 'individualistic bias' she finds in recent philosophical work on joint intention and action (see Baier 1997). She objects in particular to the assumption that multiperson actions are constructed out of single persons, their abilities, and intentions. Baier proposes that doing things with others is more basic than doing things alone. To say this, however, is not yet to explain what doing things with others amounts to.

Clearly, the range of existing philosophical accounts of joint intention and action is wide. They include accounts in terms of personal intentions, with appeal to moral principles such as Scanlon's principle of fidelity to account for the associated entitlements and obligations. (This appeal is to be found in both Bratman's and Tuomela's work.) They also include Gilbert's account, which derives the associated entitlements and obligations from a joint commitment. Gilbert's account is a nonindividualistic, or holistic, account, at least insofar as a joint commitment in her sense is not simply a conjunction of personal commitments.

Accounts that make personal intentions central accord with the approach to social phenomena of Max Weber who put at the center of his sociological concepts the concept of an action that is oriented to the actions of others (called by Weber a 'social action'). They harmonize well, also, with the somewhat more recent idea of characterizing social processes in terms of the mathematical theory of games. Gilbert's plural subject account of social phenomena accords rather with the approaches of Emile Durkheim and Georg Simmel, among sociologists, and to Jean-Jacques Rousseau among political philosophers. (Gilbert 1989 proposes an interpretation of Durkheim that suggests a commonality between Gilbert's approach and that of Durkheim.) These thinkers all emphasized in one way or another that people engaged in joint action were unified in a significant sense.

See also: Cooperation: Sociological Aspects; Group Processes, Social Psychology of; Individualism versus Collectivism: Philosophical Aspects; Interactionism and Personality; Interactionism: Symbolic; Problem Solving (Joint), Psychology of

Bibliography

Baier A C 1997 *The Commons of the Mind*. Open Court, Chicago

Bratman M E 1999 *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press, Cambridge, UK

Durkheim E 1982 *The Rules of Sociological Method*, 1st American edn. Free Press, New York

Gilbert M 1989 *On Social Facts*. Routledge, London

Gilbert M 1996 *Living Together: Rationality, Sociality, and Obligation*. Rowman and Littlefield, Lanham, MD

Gilbert M 1999 *Sociality and Responsibility: New Essays in Plural Subject Theory*. Rowman and Littlefield, Lanham, MD

Lewis D K 1969 *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA

Rousseau J-J 1983 *On The Social Contract*. Hackett, Indianapolis, IN

Scanlon T 1990 Promises and practices. *Philosophy and Public Affairs* **19**: 199–226

Searle J R 1995 *The Construction of Social Reality*. Free Press, New York

Sellars W 1963 Imperatives, intentions, and the logic of 'ought.' In: Castaneda H-N, Nakhnikian G (eds.) *Morality and the Language of Conduct*. Wayne State University Press, Detroit, MI

Simmel G 1971 *Georg Simmel: On Individuality and Social Forms*. University of Chicago Press, Chicago

Tuomela R 1984 *A Theory of Social Action*. Dordrecht, Boston

Tuomela R 1995 *The Importance of Us*. Stanford University Press, Stanford, CA

Weber M 1922 *Economy and Society*. University of California Press, Berkeley, CA, Vol. 1

M. Gilbert

Jokes and Joking

The terms joking, humor, laughter, and the comic are often used interchangeably. They refer to expressive culture. A joke is something done or said in a playful manner to excite mirth. It may take the form of a short narrative or action in which boundaries are transgressed. The capacity to joke and perceive something as funny is a universal. Humor is relative: what is held to be funny may differ from time to time, place to place, person to person. Joking is a complex behav-

ioral and cultural process in which the what is as important as the who, how, and when. This article reviews some forms, domains, and functions of joking.

1. The Nature of Humor

Humor is omnipresent yet elusive, nonsensical yet serious, friendly yet hostile, a universal yet specific. A joke can make sense across two millennia, as this 'stupid peasant' joke from ancient Greece testifies: 'A peasant, having heard that parrots live for a hundred years, bought one to see if it was true.' A joke may also change or lose its point as it moves between cultures. The 'dirty jokes' that provoke laughter in a Tamil popular theater hardly draw a smile from Dutch observers. Worldwide topics of joking relate to the boundaries between nature and culture (sex, food, health, death) and to persons on social boundaries (strangers and lunatics).

Understanding a joke requires intellectual effort, yet scholarly attention partly spoils the punchline. The comic is part of everyday life yet stands apart. Humor creates a momentary world of make-believe. It is closely related to laughter, yet one can appreciate a joke without laughing, and not all events and actions which trigger laughter are funny.

This paradoxical nature of humor is further complicated by the fact that it is both perception and performance, cognition and action, intellectual and emotional, institutionalized and spontaneous. Cognition is involved in grasping a *double entendre* or perceiving an incongruence. On the other hand, much humor touches the emotional stratum of cultural expectations and involves mimic, gestures, pantomime, and horseplay. An apt one-sentence summary of humor is 'a play upon form' (Douglas 1968) or 'a play upon meanings' (Zijderveld 1983).

1.1 The Study of Joking

The paradoxical nature of humor helps to explain why philosophers are interested in it and social scientists have neglected it. Philosophers including Plato, who detested humor, Erasmus and Rabelais, who celebrated it, and Bergson and Plessner, have devoted learned treatises to the mechanisms of joking. In their perspective, there is a bias towards intellectual and linguistic mechanisms, especially on the awkward incongruity, perceived in surprise, which most thinkers define as the essence of humor.

Freud promoted the psychological study of humor and considered joking a regression to infantile modes of thinking and a substitute for aggression and libido. Freud concentrated on *der Witz*, which is the quintessence of Yiddish humor: a short, pointed, witty story, or pun.

During the 1980s and 1990s humor research has expanded both in terms of disciplines involved and number of publications. In 1988 *Humor: International Journal of Humor Research* was founded. Psychologists and linguists dominate the inquiry into joking.

1.2 Anthropology

It is said that Bronislaw Kaspar Malinowski, a founding father of social anthropology, defined his discipline as the science of the sense of humor. Indeed, the anthropologist shares with the humorist a strategy of making the familiar strange and the strange familiar. Common sense is disrupted, the unexpected evoked in order to enhance the awareness that reality is socially and culturally constituted. This relativist perspective may be essential for the cross-cultural study of humor (Driessen 1997).

Yet, despite Malinowski's claim, until recently anthropologists have neglected joking as a topic. So have sociologists (Zijderveld 1983). Some early anthropologists studied humor from a folklorist and antiquarian perspective, but rarely as a resource for cultural analysis. In some fieldwork guides joking and laughter are not even mentioned. Radcliffe-Brown's article (1940) was a breakthrough, although his interest was in the structural implications and functions of joking relationships rather than their symbolic content. Mary Douglas found the essence of the joke in its attack on control. In the 1970s and early 1980s anthropologists used semiotic and narrative approaches to study joking *per se*, not just its functions (Handelman and Kapferer 1972). The first comprehensive, comparative anthropological study of humor and laughter appeared in the mid-1980s (Apte 1985), and since then there have been numerous articles and books. Yet, joking remains to the most part a spin-off of other research topics, despite the knowledge that it provides clues to what really matters in groups and societies.

2. The Faces of Jokes

2.1 Genres of Joking

In most cultures there is a distinction between the diffuse joking of common people in everyday life and institutionalized joking by experts at specific times and places, such as the clown in wedding ceremonies in Polynesia, the television comedian on New Year's Eve, or the trickster in a Javanese *wajang* play.

Humor ranges from mild jokes to biting satire. Benign joking is simple, informal, and harmless, intended for diversion and repose. It does not unsettle the course of daily life. In metropolitan Egypt, for instance, joking is considered the lubricant of social life. Persons who lack a sense of humor 'have thick

blood' and are scorned. A second type of joking is wit, banter, persiflage, or punning. This is a game of intellect and language, using irony to tackle discrepancies between appearance and reality. A third type is black or gallows humor which challenges and mocks tragedy and misery, for instance the 'Darwin Awards' on the Internet. A fourth form is satire, militant or aggressive joking, meant to criticize authority, common sense, and conventions, as in Carnival. The trickster of the Winnebagoo, Pueblo, and other Native Americans is a satirist *pur sang*, as were the Eurasian court jokers. The political cartoonist is a modern example of a satirist.

2.2 Domains of Joking

It is often hard to translate a joke from one language to another, especially in the case of puns and witticisms. Language is both a vehicle and topic of jokes. Linguistic play and rhetorical devices, such as incongruity, paradox and irony, are central to puns, tall tales, and riddles. Anthropologist Chagnon was the victim of verbal humor struggling to learn Yanomamö. Informants made a sport of inventing obscene and ridiculous names for people such as 'dirty rectum' or 'feces of the harpy eagle' which Chagnon wrote down and used without knowing their meaning (Driessen 1997). Apaches stage joking imitations of the 'whiteman' with exaggerated, loud, and fast speaking (Basso 1979). In many societies, for instance the Inuit, verbal virtuosity is expressed in duels in which humorous and obscene insults are recited to ridicule opponents. In Egypt these mock fights are ritualized in male coffee houses, for instance: 'Your father puts on his underpants with a shoehorn'; 'when your father sees a garbage truck coming he lies down like a heap.'

There is also a strong interdependence between gender and joking. Sexuality is a core subject of jokes everywhere and humor is an important element in attraction among the sexes. In most societies there is gender inequality in joking. While men tend to use more aggressive humor and laugh louder in public spaces, women remain more passive and restrained. Men play more prominent roles in ritual public joking. There are exceptions. Older women may publicly engage in humor and even compete with men. In Pacific island communities female clowns play important ceremonial roles. In many Mediterranean societies women perform for women mock imitations of men, their sexuality included. In Western society female comedians are gaining prominence.

Age is important in joking. Psychologists have noted the crucial role of humor in personality development and language learning. It serves as a general educational device. Children and adolescents have a special sense of humor with which they mock adults. Adults often find amusement in encouraging children to imitate adult etiquette. There are striking gender differences in the nature, devices, and approval of children's jocular behavior.

A widespread kin-related form of institutionalized humor is the joking relationship. Some categories of kin are licensed, sometimes even obliged, to tease, ridicule, and insult persons and groups belonging to another category, who must accept it without taking offence. In many societies kin of the opposite sex who are potential sexual partners are involved in joking relationships. In Western society mild joking is allowed with grandparents.

Where governments fear freedom of expression and rigidly demand conformity, political humor is repressed yet flourishes. 'Joking up' is a weapon of the subordinated whereas 'joking down' may act as a means of the powerful to express their superiority.

One of the main domains of joking in the globalizing world is ethnicity. Ethnic jokes make fun of supposed and farcically exaggerated defects in certain ethnic groups or nations. The deficiences appear in pairs, e.g., the stupid and canny, the stupid and cowardly, militarism and cowardice. Such dualities relate to commonly experienced ambiguous and contradictory situations (Davies 1990). The telling of ethnic jokes may be, but is not necessarily, an expression of racism and discrimination. Ethnic and many other jokes have a complex relationship with everyday experience.

Religion is replete with paradoxes and ironies. Consequently, joking, laughing Gods, and holy fools are to be found in the beliefs and rituals of many religions, global and local. There are counter-virtuosi who disturb and suspend the rigid ritual order by burlesque, inversion of roles, sexual, and scatalogical humor. Berger (1997) even argues for a 'theology of the comic' in which the experience of the comic is a promise of redemption.

Joking is not only a feature of face-to-face communication but also a major element in newspapers and magazines, movies, radio, television, and the Internet. Comic books have become a popular art form. There is musical humor and humor in novels and poetry. Humor is an effective device in advertising. In sum, joking permeates all domains of public and private life.

3. Meanings and Functions

Given the Protean nature of humor, it is not surprising that there is no universal structure of meaning to joking. The content of many jokes relates to commonly experienced ambiguities, dilemmas, and paradoxes that are part of the human condition. Joking is one of the ways to deal with such experiences, and this is a universal meaning of humor. Much joking involves entertainment. It produces pleasure and smoothes social interaction. One of the most widespread psychological and functionalist theories has it that joking

plays an important role in conflict resolution by releasing tension. Related theories hold that joking is a tool of cultural critique and social control by highlighting taboos and key values. Another social function of jokes is that they mark group boundaries and help to promote a sense of belonging, sometimes by testing and challenging social cohesion. Jokes can also effect greater commitment, regulate embarrassing behavior, and sustain serious interaction. And jokes are relativizers, creating in a playful manner distance from turmoil and toil, misery and misfortune.

Trends in humor research show an increased interest in the healing power of joking and laughter in cross-cultural communication, psychotherapy, neurology, and physiology. Ethnic humor is a topic of growing relevance as are differences in joking styles and tastes.

See also: Cultural Relativism, Anthropology of; Entertainment; Folklore; Freud, Sigmund (1856–1939); Functionalism in Anthropology; Liminality; Malinowski, Bronislaw (1884–1942); Play, Anthropology of

Bibliography

Apte M L 1985 *Humor and Laughter: An Anthropological Approach*. Cornell University Press, Ithaca, NY

Basso K H 1979 *Portraits of 'The Whiteman.' Linguistic Play and Cultural Symbols Among the Western Apache*. Cambridge University Press, Cambridge, UK

Berger P L 1997 *Redeeming Laughter. The Comic Dimension of Human Experience*. Walter de Gruyter, New York

Davies C 1990 *Ethnic Humor Around the World. A Comparative Analysis*. Indiana University Press, Bloomington, IN

Douglas M 1968 The social control of cognition. Some factors in joke perception. *Man* **3**: 361–76

Driessen H 1997 Humor, laughter and the field: reflections from anthropology. In: Bremmer J, Roodenburg H (eds.) *A Cultural History of Humor*. Polity Press, Oxford, UK

Handelman D, Kapferer B 1972 Forms of joking activity: A comparative approach. *The American Anthropologist* **74**: 484–518

Nilsen D L F 1993 *Humor Scholarship. A Research Bibliography*. Greenwood Press, Westport, CT

Radcliffe-Brown A R 1940 On joking relationships. *Africa* **13**: 195–210

Zijderveld A C 1983 The sociology of humor and laughter. *Current Sociology* **31**: 1–103

H. Driessen

Journalism

Journalism is generally defined as the activity of gathering and presenting information and commentary about current events for public distribution. From a social science point of view, it can be conceived in a number of ways: as a distinct *occupation* with its own patterns of recruitment and socialization and occupational subculture; as a distinct form of social *activity* involving the gathering and processing of information; as a social *institution* with a set of relations with other institutions like the state; or as a *cultural form* or form of discourse.

Disputes about the boundaries of journalism have been common among both scholars and practitioners. Chalaby (1998), for example, has argued that journalism properly so-called emerged in Britain and the US with the development of commercial mass media; the production of news as a commodity, in his view, is radically distinct from the earlier practice of using the press to publicize political ideas. Journalists' organizations, particularly in countries with state-owned broadcasting systems, have sometimes excluded those who worked in broadcast news, on the grounds that they did not practice journalism independently, but merely transmitted the views of others. In many countries today there is also sharp debate about whether those who produce new forms of 'infotainment'—talk show hosts or producers of 'tabloid' news magazines, for example—can legitimately be called journalists. There is little to be gained conceptually by trying to adjudicate these disputes and establish the 'true boundaries' of journalism. These disputes are enlightening precisely because multiple forms of journalism have always coexisted, and because journalism has never been rigidly differentiated from other social activities, but has overlapped with and existed in relations of mutual interdependence with a variety of political, economic, social and cultural institutions, and practices.

1. Historical Evolution

The term 'journalist' emerged in the seventeenth century; originally, 'journalism' referred simply to the activity of producing a printed periodical. One can trace many of the social functions involved in journalism to earlier eras, but it seems best to say that journalism emerged with the development of the newspaper and other forms of print culture. The emergence of journalism was connected not only with the development of print technology, but also with an inter-related set of social and cultural changes. First, it was connected with the development of capitalism and the market. This was true in two senses. The market, first of all, increased the demand for specialized sources of information. Those involved in market relationships had an intensified need to know about occurrences beyond their own world of face-to-face relationships. Merchants were among the first producers and consumers of newspapers, and as increasing parts of the population were drawn into the market economy, newspaper readerships grew. Second, news-

papers were themselves distributed through the market; this was true even in their early days, when the primary motivation for their production was often political or cultural rather than economic.

The emergence of journalism was connected with the rise of the nation-state and political citizenship. It is connected of course with the emergence of democratic politics; early newspapers were often primarily vehicles for public political participation. Expansions of newspaper readership were often closely connected with expansions of the electoral franchise. In the US, for example, the rise of the 'penny papers,' which greatly expanded the audience for news, came on the heels of the Jacksonian revolution and the elimination of property qualifications for voting. In Uruguay, one of the few Latin American countries in which a mass circulation press has developed, newspaper circulations increased from one newspaper for every 11 inhabitants in 1870 to one for every four in 1916, a period which also saw the rise of mass-based political parties.

The notion of citizenship is broader that of democracy, however, and the connection between journalism and the rise of the nation-state goes beyond the development of democratic political institutions. Another way in which the modern nation-state draws the mass public into political life is through mobilization for war; newspaper circulations have typically expanded in times of war as well as in times of expanded political participation. As Anderson (1991) has pointed out, the culture of nationalism involves a sense of belonging to an imagined community of compatriots; journalism is one of the key cultural forms in the creation of this sense of national collectivity, and it can play this role with or without political democracy.

Finally, journalism is connected with the rise of realism as a cultural form; it emerges simultaneously with a number of related realist genres, including the novel and scientific writing, and indeed has in certain historical periods overlapped with these, as novels and scientific information were published in newspapers and novelists often wrote for the press. All three of these historical changes, of course, are connected with the rise of the public.

Max Weber (1947) once characterized the journalist as a 'type of professional politician.' Indeed in its early days journalism was substantially—though not exclusively—a part of the world of politics. Journalists were often advocates for political causes; Benjamin Franklin, for example, who like many others in his age was simultaneously a journalist and a politician, described journalism as essentially a form of oratory and thus part of a tradition that could be traced back to the political assemblies of Greece and Rome. In democratic countries, newspapers typically became associated with political parties, as these began to emerge, or factions of parties, and often were supported financially by parties or individual politicians. In authoritarian countries newspapers were often either political mouthpieces of the state or of insurgent movements. In Africa, for example, many leaders of the independence movements of the twentieth century were involved in journalism.

The identification between journalism and politics was substantially changed, however, by the rise of the commercial newspaper industry. This began in the United States in the 1830s, with the development of the 'penny papers,' which expanded newspaper circulations by an order of magnitude and became profitable cultural industries by selling the attention of a mass audience to advertisers (Schudson 1978). In Britain commercialization of the press was delayed by restrictive taxes, but got under way in the 1850s. In Japan, similarly, commercial papers began to compete with the earlier, politically oriented papers at the end of the 1870s. Commercialization did not push politics aside altogether. Most newspapers in the United States, for example, continued to have strong partisan loyalties throughout the nineteenth century at least, and their owners and editors to be active in politics. They were, however, financially independent of parties, and their owners independent political powers in a way their predecessors often had not been. In this sense it could be said that commercialization increased the power of journalism as an institution. Commercialization also changed the agenda and style of journalism, diminishing political advocacy in favor of the competition to provide news in a timely and sometimes sensational fashion.

A sharp debate has emerged in media studies over the implications of commercialization of the news media for democratic politics. One view holds that commercialization gives the press the independence it needs to serve as a 'watchdog' of the state and to provide the public as a whole with information 'without fear or favor.' The contrary view holds that commercialization tends to drive political content out of the press, replacing it with sensational human interest material, and also to consolidate control of the press in the hands of the upper class, leading to unequal political competition between competing social interests (Curran and Seaton 1997). This argument has been developed especially from the case of Britain, where a highly politicized working class press flourished in the period when taxes inhibited the development of commercial media. Weber (1947) also observed that the commercial papers in Germany 'have been regularly and typically the breeders of political indifference.'

Commercial newspapers were the first to employ hired reporters, editors, and eventually artists, photographers, and other specialized journalistic personnel. Throughout the nineteenth century journalism developed as a distinct occupation with specialized practices and a sense of identity. Many of the conventions for gathering and presenting information which we now think of as distinctively journalistic

emerged during these years, including for example the summary lead, which synthesizes the 'most important' information in a news story for the reader, and the practice of interviewing public figures, which was important in part because it gave journalists a more active role in the creation of news, and in some cases increased the public profile of individual journalists.

In the twentieth century the notion of journalism as a 'profession' began to emerge, most distinctively, perhaps, in the United States, where the idea of neutral expertise became a particularly strong element of political culture. Professionalization meant, above all, the development of the idea that the journalist serves the public as a whole, rather than particular political tendencies, owners, or other interests. It is associated with the shift, noted by Siebert et al. (1956), from the older 'libertarian' conception of press freedom to the 'social responsibility' model, which sees the press as the holder of a public trust. Professionalization was also associated with a shift toward balance and 'objectivity' as primary journalistic values, with corresponding changes in writing styles, and with increased autonomy of journalists, as against owners or managers within news organizations. Journalistic autonomy has nevertheless always been limited: despite aspirations in some countries for direct journalistic control of news organizations (perhaps most notably France after World War II; the Journalists at *Le Monde* still elect the paper's Director), ultimate authority in virtually all news organizations remains outside the hands of journalists.

Several decades of scholarship have painted a complex picture of the implications of the professionalization of journalism. Professional norms can be seen, for example, both as a mechanism of social control, constraining journalists and often serving to exclude nonconforming content from the news, and as a basis for journalists' autonomy, legitimizing their right to a space of action not constrained by their owner's political views, the pressures of advertisers, etc. (Soloski 1989). The concept of professional routines has become central to the analysis of the social role of journalism. Routines are the standard practices which enable news organizations to operate efficiently and which legitimate the many choices that must be made in the production of news, allowing them to be treated as matters of shared professional judgment rather than political debate (Tuchman 1978). Where professionalism is strong, these routines are powerful enough that particular individuals, including both journalists and owners, have limited ability most of the time to shape news content. Those routines, however, often have social biases built into them, and most scholars consider the biases built into journalistic routines to be the most important explanation for the politics of news (Gitlin 1980). Among the most important routines is the use of certain kinds of sources of information which are considered 'authoritative'—principally government officials and other elite members of society (Gans 1979, Hall et al. 1978).

Where professional autonomy—or 'internal press freedom,' as it is often called in Europe—is weaker, professional routines become less important for explaining news content, and more instrumental explanations, focusing on control by private owners or the state, are often more appropriate.

There are many views about what kinds of social biases prevail in contemporary journalism. Some, for example, have argued that journalism generally undermines the authority of established social institutions (this view began to grow in the 1970s, which probably saw the high point of journalistic autonomy in most Western countries, and also saw a decline in the legitimacy of many social institutions). Some have argued that they serve more or less consistently to support established social interests, and some that they are essentially neutral, and represent the full range of contending interests. The most common view among scholars is probably the view that in liberal societies media routines tend to reproduce the existing structure of power, but not in an entirely consistent way; under certain circumstances, for example, the dominant routines can open the media to penetration by new social movements or lead to periods of contestation over the boundaries of political debate.

2. *Differentiation*

Structural functionalism holds that human societies tend to evolve toward increased differentiation, in which institutions become increasingly specialized in the functions they perform. Some scholars—most notably Alexander (1981)—have argued that the development of journalism should be seen in this light. This does make considerable sense in societies where commercialization and professionalization are relatively advanced. Journalism in these societies has increasingly become separated from party politics, and has developed norms and routines of its own, a 'media logic' distinct from the logic of politics. In other ways, however, the notion that journalism tends to evolve toward differentiation seems doubtful. Most importantly, such a view neglects the tendency for journalism to become integrated into market mechanisms. Professionalization, to be sure, limits this to some degree, but the 'media logic' that has emerged in contemporary societies is very much a joint product of professional norms and the pressures of the market, which many argue are growing increasingly strong. Other scholars thus give a very different account of the history of journalism than the structural functionalists. Habermas, for example, decries the 'colonization' of the public sphere by the cultural industries, and Bourdieu argues that the journalistic field once heavily influenced by the political field is now heavily influenced by that of the market.

3. Journalism and the State

In some societies the state controls the news media directly. This is most common with broadcast news, but is also common with newspapers, particularly in developing countries where there are few sources of capital outside the state. In societies with clientelist political relations, journalism will often be tied to the state in various ways by systems of political patronage, as in Mexico, where until recently both journalists and newspapers were heavily dependent on state funding (Orme 1997), or in Greece, where industrialists often use their papers to bargain for favors from the state (Papathanasopoulos 1999). A similar pattern has emerged in Russia, and to varying degrees in other parts of the former Soviet Union and Eastern Europe.

In liberal societies journalism is generally thought of as standing apart from the state, serving to check its power as a watchdog or fourth estate. This is accurate to a significant degree, as reflected in incidents like the publication by *The New York Times* and other papers of the 'Pentagon Papers' (a secret government study of US decision-making on the Vietnam war 'leaked' to the press by a dissident former official). But close ties between the media and the state also exist in liberal societies. State subsidies and restraints do exist in such societies, as do clientelist relationships between politicians and media owners. But more important is the close relationship between journalists and government officials as sources and subjects of news, a relationship which has proved mutually beneficial to officials seeking publicity and control over the news agenda, and journalists seeking news (Bennett 1990). This relationship is highly developed, with the news media organizing much of the newsgathering process around the institutions of the state, and the state devoting considerable resources to accommodating them. In many cases the relationship is formally organized, as in the lobby system in the British parliament or the press clubs of Japan (Pharr and Krauss 1996). The closeness of the relationship between contemporary journalists and state officials means that journalists have access to considerably more information about the inner workings of government than they would as outsiders. It also means that they are often in an important sense part of the community of political elites, and reflect the prevailing viewpoints within that community. Here, again, doubts can be raised about the differentiation theory of journalism history: it could be argued that journalism is more entangled with the state than it was a century or two ago, when journalists had no privileged access to the corridors of power.

4. Varieties of Journalism

Journalism has always been plural; indeed one of the distinctive characteristics of the modern newspaper is the fact that it mixes widely disparate types of content, not only in terms of subject matter but in terms of the functions it fulfills for readers. Journalism provides both information and entertainment. It also provides moral and ideological orientation. Even in countries where objectivity became a dominant professional value, there continue to be many occasions when journalists communicate value judgments to their readers, particularly when the society's consensus or hegemonic values are considered to be at stake, in covering international conflict, for example, or in investigative reporting. Some scholars have also argued that journalists perform a ritual function, promoting a sense of community solidarity, particularly in the coverage of ceremonial events and crises

In most societies there is also some differentiation of journalism into distinct styles, often serving different parts of the community. The most striking example is probably the differentiation of the British press between the 'quality' or broadsheet press, serving middle and upper class readers and emphasizing informational journalism, and the mass or tabloid press, serving primarily working class readers with a much more entertainment-oriented form of journalism. In France a distinction has often been drawn among 'rationalist,' 'sensationalist,' and 'ideological' newspapers, the latter including Catholic and Communist organs. In many countries local and regional papers are more commercially oriented and lack distinct political orientations, while national (or capital city) papers reflect distinct political tendencies. In the US, ethnic minority media have played an important role, and often practice a form of journalism that was intended less to provide neutral information than to defend the interests of a particular part of the community.

Diverse social and political systems have also given rise to diverse forms of journalism. In Europe, for example, although partisan journalism in the traditional sense has declined dramatically since the Second World War (the longest important survivor was the Italian Communist paper *L'Unità*), newspapers nevertheless have more clearly differentiated political orientations than those in the US. European journalists are also more likely to see commentary and to some degree advocacy as integral to the journalist's role than their US counterparts (Mancini 1993, Donsbach and Klett 1993). An important distinction can also be drawn between societies where newspapers have mass circulations (mainly in Northern Europe, North America, and parts of East Asia) and those (in Southern Europe, for example, and most of Latin America) where their circulation is restricted to a relatively small, usually educated and politically-active elite. Print journalism tends to be more a part of the political world and less a part of the market in the latter, and it emphasizes 'horizontal' communication among political elites more than 'vertical' communication between elites and a mass public.

Journalism in Communist societies had its roots in the Western European tradition of party journalism, oriented toward political mobilization. But with the fusing of party and state, journalists became official figures and distinctive functions evolved. Journalists have often served a function similar to that of the ombudsman, for example—as someone to whom ordinary people would complain, hoping to be heard by those in power. Journalism was an important institution in Communist societies, well supported by the state, trade unions, and other official institutions to which it was attached. In countries where the communist party remains in power, particularly Vietnam and China, a complex hybrid of party control and market orientation is emerging (Zhao 1998).

In postcolonial societies, the most distinctive form of journalism to emerge was 'development journalism,' particularly prominent during the 1970s, which rejected the notion of the journalist as disinterested observer, and saw it as the function of the journalist to play an active role in promoting economic and political development. Development journalism was widely criticized for subservience to state power, though to a large extent the political caution which development journalists exercised was probably dictated by the same conditions of political life that affected all journalists in postcolonial societies, whatever model of journalism they espoused. In many ways, development journalism is similar to the tradition of 'boosterism' in US journalism, which similarly saw the newspaper as a force for the economic development of the community, and often similarly emphasized consensus over conflict.

5. *Current Developments*

A number of powerful forces are currently transforming journalism and its social role, described in the following sections.

5.1 *The Rise of Television*

In most of the world, television news has increased in importance at the expense of print journalism. The mass public increasingly relies on television as a source of news, and political communication, which was once channeled mainly through newspapers and party organizations, flows increasingly through television. The direction of influence within journalism has also shifted. Television once clearly followed the lead of print journalism. Today, there is a complex reciprocal influence, and newspapers still increasingly emulate the agenda and style of television. In some countries, the US, for example, this trend is decades old. In others it is more recent, dating from the 1980s or early 1990s when the public service monopoly on broadcasting ended.

5.2 *Commercialization*

In many countries the news media have become increasingly commercialized and journalism more strongly influenced by market factors. In the West the news media have been substantially commercialized since the nineteenth century. But for much of the twentieth century the effects of market forces on the news media were partly counterbalanced by professionalization, by public service broadcasting, and by press subsidies, which many European countries expanded in the 1970s to keep politically oriented papers alive and prevent market forces from eliminating competing voices (Gustafsson 1980). All three have become relatively weaker in recent years. The rise of television and, in Europe and some other parts of the world, the displacement of public service by commercial broadcasting is an important reason. Commercial television is primarily an entertainment medium, and journalism is much less central than in the newspaper industry.

The culture of journalistic professionalism may also have been undercut by a general decline in the prestige of political life. A key element of that culture was the idea that journalists should make independent judgments about what the citizens needed to know about public affairs. Today it is increasingly assumed that it is the journalist's job to provide whatever content interests the media consumer. The news agenda has shifted away from public affairs in the traditional sense, and toward human interest and 'service' journalism useful to ordinary people in everyday life (investment advice, health reporting) Journalistic autonomy has been undercut in important ways, as the business and editorial sides of news organizations are increasingly integrated.

Some argue that these changes have made journalism less elitist, more responsive to the concerns of the majority of the population, others that they undermine the integrity of the system of public communication. Journalism clearly has important functions for the society as a whole that most private industries do not have. It is central to political life, and to the process of opinion formation more generally; some have argued that it is increasingly central, as other institutions, political parties, and trade unions, for example, have weakened. It remains one of the key questions of media studies whether market mechanisms can provide the fulfillment of these functions, and if not, what other social mechanisms can be devised.

5.3 *Democratization*

In developing and newly industrialized countries, there has been a general move toward liberal democracy in recent years, with many authoritarian regimes giving way to multiparty systems and greater political openness. An increased emergence of independent news

media has been associated with these changes, probably as both cause and effect. Opposition parties have increasingly gained access to the media in many countries, political scandals have proliferated (Waisbord 1994), and elites have lost much of the tight control over public discourse they once held. Often commercialization of the media has accompanied democratization, and this has led to a dramatic reorientation of journalism away from serving government officials and toward the mass audience, sometimes resulting in rapid shifts from 'protocol news' to sensational crime reporting. Many questions remain about whether and how democratic media systems will be consolidated in these societies. To what extent will advertising-based media in societies where the middle class remains relatively small serve the majority of the population? To what extent will media owners—often small in numbers as the economic base of the media is limited—attempt to use the media to support their own political interests? Under what conditions can standards of professionalism evolve that will limit demagoguery or manipulation of political information?

5.4 New Technologies

Most discussion of the impact of new, particularly digital technologies remains speculative. It is clear that the boundaries between technology-based genres of journalism—print, radio, television—will be broken down. Some have argued that proliferation of channels will make journalism less central as a mediator of information—people will watch events live or exchange information directly on the Internet. Multiplication of channels has indeed created competition for the role of mediator as, for example, MTV or Internet gossip sheets compete with network television and mainstream newspapers. It is also quite possible, however, that the proliferation of information and its centrality in economic and social life will make those who collect and synthesize it more rather than less important. It is also not clear whether new technologies and the new information industries arising around them will result in fragmentation of journalism into distinct genres targeting different demographic groups, or in new forms of competition for a mass audience.

Bibliography

Alexander J 1981 The mass news media in systemic, comparative and historical perspective. In: Katz E, Szecskö T (eds.). *Mass Media and Social Change*. Sage, Beverly Hills, CA

Bennett W L 1990 Toward a theory of press-state relations in the United States. *Journal of Communication* **40**(2): 103–25

Chalaby J 1998 *The Invention of Journalism*. Macmillan, New York

Curran J, Seaton J 1997 *Power Without Responsibility: The Press and Broadcasting in Britain*. Routledge, London

Donsbach W, Klett B 1993 Subjective objectivity: How journalists in four countries define a key term of their profession. *Gazette* **51**: 53–83

Gans H J 1979 *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek and Time*. Pantheon, New York

Gitlin T 1980 *The Whole World is Watching: Mass Media in the Making and Unmaking of the New Left*. University of California Press, Berkeley, CA

Gustafsson K-E 1980 The press subsidies of Sweden: a decade of experiment. In: Smith A (ed.) *Newspapers and Democracy*. MIT Press, Cambridge, MA

Hall S, Crichter S, Jefferson T, Clarke J, Roberts B 1978 *Policing the Crisis*. Macmillan, London

Mancini P 1993 Between trust and suspicion: How political journalists solve the dilemma. *European Journal of Communication* **8**: 33–51

Orme W A (ed.) 1997 *A Culture of Collusion: An Inside Look at the Mexican Press*. Lynne Reiner Publishers, Boulder, CO

Papathanasopoulos S 1999 Effects of media communication and politics in Greece. *The Communication Review* **3**(4):

Pharr S J, Krauss E S (eds.) 1996 *Media and Politics in Japan*. University of Hawaii Press, Honolulu, HI

Schudson M 1978 *Discovering the News: A Social History of American Newspapers*. Basic Books, New York

Siebert F S, Patterson T, Schramm W 1956 *Four Theories of the Press*. University of Illinois Press, Urbana, IL

Soloski J 1989 News reporting and professionalism: Some constraints on reporting the news. *Media, Culture & Society* **11**: 207–28

Tuchman G 1978 *Making News: A Study in the Construction of Reality*. Free Press, New York

Waisbord S 1994 Knocking on newsroom doors: The press and political scandals in Argentina. *Political Communication* **11**(1): 19–33

Weber M 1947 Politics as a vocation. In: Gerth H H, Wright Mills C (eds.) *From Max Weber: Essays in Sociology*. Oxford University Press, New York

Zhao Y 1998 *Media, Market and Democracy in China: Between the Party Line and the Bottom Line*. University of Illinois Press, Urbana, IL

D. C. Hallin

Journalism and Journalists

The journalist is the key employee of news organizations who, in the definition adopted by the International Labor Organization, 'collects, reports and comments on news and current affairs for publication in newspapers and periodicals or broadcasting by radio and television.' Due to national peculiarities in labor statistics a global estimate of the number of journalists—500,000 in 1989—remains quite unreliable.

Journalistic work is primarily based on the *symbolic manipulation* of the content (sense making) in contrast to those, like typographers, responsible for the *mechanical* production of the newspaper or other news medium. In large and complex electronic media

systems, journalism is contextualized and co-determined by other media output and occupations. Several groups of employees, who are occupied with symbolic manipulation or involved in administrative and organizational tasks, are conventionally not considered journalists. The development of electronic production and distribution of media contents, and the growth of new forms of journalism (e.g., tabloids and talk shows, the Internet) brought about numerous occupational groups (technicians and other employees in broadcasting, documentary production, audio-visual workers, computer programmers, designers) whose work cannot be classified exclusively into either side, i.e., they participate both in symbolic and technical (re)production of contents. The boundaries between journalism and nonjournalism, and between full-time and part-time journalists become even more problematic in the cyberspace dominated by a constant flux of information produced, reproduced, and distributed by millions of Internet 'visitors.'

There is no uniform 'journalistic culture' or concerted form of professional operation of journalists that would lead to a single concept of journalistic work and norms. In contrast to the rather technical meaning of journalism in English and most other languages of Western civilization referring to the workforce in institutionalized forms of (mass) communication, the Greek concept of journalism, *δημοσιογραφία* (*demosiografía*), means 'writing about/for the people' and is focused on the social substance of journalism by emphasizing people rather than a particular class, or nature. Since *demos* means more than population but *citizenship*, journalism implies political relevance, which indeed was its most significant common trait until very recently. A critical labor perspective in the 1960s led to the concept of newsworkers, suggesting a distinction between news, and entertainment or fiction without solving the problem of a clear sociological referent of the concept. In the Muslim world, journalists are required, similarly to those in the West, not to 'mix truth with falsehood' and to verify news, but reporting is not allowed to be conceived on suspicion, to ruin people's reputation or to mock them, and most importantly, news must have a 'use value' in that it must provide benefit (*Nafa*') to the people. The latter idea was central to 'development journalism' emerging in the 1960s in developing countries, emphasizing that reporting should contribute to national development of the country concerned. Similarly, 'public' or 'civic journalism' movements in the West in the 1990s suggest that journalists should not merely report, but discover a community's 'relevance' by meeting the people and experiencing their lives.

People have always passed on information from one to the next, but journalism as a distinctive occupational task did not begin until the seventeenth century, when it became a supplementary, part-time occupation of otherwise professionally committed individuals in the social communication process—e.g., printers, postmen, or tradesmen. Modern journalism, as a result of industrial and bourgeois revolutions, became fully institutionalized with the mass media. During that period, the press was largely liberated from censorship and party subservience. R. Park noted that the rise of the reporter had been one of the most important events in the 'age of news.' While in earlier centuries journalists largely belonged to self-sustained free professions of the middle class, with industrialization they joined the class of workers employed by the press publishers. The separation between the occupations of publisher and journalist was followed by the establishment of national and international professional organizations and trade unions of journalists.

During the nineteenth century, newspapers were generally still closely linked to political parties and journalists subject to the political credo of newspapers. The idea of objectivity represented an alternative to partisan journalism and a way to legitimate the professional status of journalists. The quest for objectivity was related to the need to establish a sort of professional, scientifically validated method for journalists to ensure their professional autonomy based on a 'standard' university education, also typical of traditional professions. The idea of journalistic objectivity as a professional norm was not widely recognized until the end of the First World War. In the early 1900s, the demand for objective reporting was embedded in the broader cultural movement of 'scientific naturalism,' which stressed the importance of methods in giving evidence of facts.

The conflict between subjective storytelling (the blending of facts and opinions, real events and trivial fictional material, news, and entertainment into a tightly packaged story) and 'objective news' (the classic function of journalism to sort out factual and reliable accounts of daily, particularly politically relevant events) was never solved. The growing number of journalistic *codes of ethics* since the beginning of the twentieth century to mark professional standards in journalism has never resulted in a single code that would universally define the standard of 'objectivity,' including the professional standards adopted by International Federation of Journalists (IFJ). There are two main reasons: (a) the two tendencies of storytelling and objective reporting are not mutually exclusive but intermingling extreme strategies in reporting events; (b) the competition for consumers causes competition for the latest news and makes journalists vulnerable to politics. Political authorities enter journalism either directly through interviews and press conferences or indirectly through 'information subsidies' (making information available to journalists on a *quid-pro-quo* basis), influence the journalistic agenda setting, if not attitudes and actions, and eventually make journalism subsidiary to *public relations*. For example, in Japan, news sections of newspapers are produced in close cooperation with

powerful news sources in government, industry, and the civil service, organized through the 'press club' system.

Journalism lacks the objective criteria—stressing the importance of universal norms, service to clients or altruism, intellectual activity and scientifically based expertise, autonomy and responsibility, and ethics—that would place it in the same social position as 'true' professions such as medicine or law. Although the social influence of journalism is rising, journalists in many countries have no specific professional standards to meet. Nevertheless, national codes of ethics share some common aims such as prohibiting editorial advertising and misleading audiences, and protecting individuals against invasion of privacy, defamation, and damaging publicity in general. Specific forms of (self-)regulation such as courts of honor, press or media councils, or press or media ombudsmen exist in a great number of countries to supervise the implementation of ethical standards by journalists and the media. They may be composed of journalists alone (as in the case of courts of honor within journalist associations) or representatives of the general public and media organizations.

There are invariably controversies between scholars, media workers, and media industries about the structure of journalism studies and the type of education needed for journalists. Early courses in journalism were mostly practical or vocational and connected with on-the-job training. In the 1920s and later, journalism curricula expanded into social sciences and included courses such as communication theories, public opinion, history of journalism, and ethics. The number of educational institutions in journalism has grown throughout the twentieth century, starting with the University of Missouri in 1908 in the United States (which at present has 350 schools of journalism) and soon followed by other regions in the world, where educational programs and materials were largely based on those from the USA and Europe. One of the first journalism schools linked to universities was set up in 1918 at the National Peking University in China. In Latin America the number of journalism faculties and schools increased from 13 in 1950 to almost 300 by the end of the century. There is no 'typical' journalism curriculum, but it usually combines journalistic and nonjournalistic courses, with the prevalence of the latter (e.g., in the USA in a ratio 25:75). Many schools offer specialization in print, radio and television journalism, public relations, and advertising; fewer have more specialized programs, e.g. in photojournalism, magazine journalism, or television documentary. Generally, two seemingly contradictory trends may be found in the development of journalism programs: more and more specialized and even fragmented curricula in different fields of journalism on the one hand, and on the other a more 'generic' approach to journalism, stressing the need for instruction in communication skills cutting across specific job demands and combined with nonjournalistic specialization such as arts, agriculture, economics, law, or politics. Both trends are restricted to larger and developed countries where many universities and schools can compete to 'supply' journalists to widely differentiated media.

In the developed world, a large proportion (often exceeding 50 percent) of newcomers to the profession graduate in journalism at a university or its equivalent. An international comparative study indicates that on average students choose journalism as senior teenagers, primarily because they believe that the work of journalists is interesting and useful for society, and that they personally are talented for this job. However, many of those who graduate in journalism never enter a journalistic career but rather other communication-related or even noncommunication occupations. Neither is specific education a condition for entering a journalistic occupation. The lack of entry conditions such as agreed upon professional qualifications and licensing may be related to the liberal idea that freedom of expression and press freedom should not be restricted, in order not to violate fundamental citizens' freedoms. Almost all countries outside the United States have a body of press or media law, but journalists have no special legal status other than as defined by the constitutional rights and freedoms of citizens and, in some countries, some minor state legislation, e.g., the protection of information sources. However, this 'ideology of nonprofessionalism' is under threat from reality, in which media owners and editors may define specific qualities and conditions that candidates for journalism must fulfil, and journalists may use mechanisms such as codes of ethics and forms of self-regulation to protect not only group autonomy but also monopoly in the labor market.

The lack of formal education in terms of theoretically grounded knowledge within a particular field contributes to the controversy about professionalism in journalism. Acquisition of such knowledge is central to the selection processes (search for the 'right' person for a job), and instrumental (skills) and expressive socialization (norms, ethics) by which entry into a profession may be controlled. The latter dimension of socialization gives primacy to the intellectual component of professionalism and to the production of expert knowledge resulting from research. In 1922, W. Lippmann complained that no research had been conducted in the field of journalism despite its immense impact on democratic life. Although the German *Zeitungswissenschaft* studied systematically the press beginning in the nineteenth century (e.g., journalistic genres, economics of newspapers), M. Weber was the first to suggest in 1910 to pay attention to journalism as a profession. He explained that a sociology of newspapers ought to include the study of journalists, part-time journalists, and other newspapers employees, 'collectivism and individualism in the creation of the content of the newspaper,' 'the

anonymity of newspaper' and its effects upon journalists, cultural specificities influencing journalism and the social functions of the press, and 'the fate and the situation of journalism as a profession' (particularly the questions of education and professional demands upon the journalist). Weber's ambitious research program was not systematically carried out until the 1950s, when during the rise of 'mass communication as a scientific discipline' in the United States journalists became the subject of empirical research. Similarly to university education, however, studies in media occupations and professionals were not always well received by working journalists.

In most countries journalists are closer to lawyers and politicians than teachers or public servants in terms of prestige and salaries. They are part of local and national élites although they are on average much younger if compared to other professions. Particularly journalists in developing countries begin with rather high social status. Generally, however, the journalists' opportunity for internal promotion within the profession is limited. Studies indicate a tendency to work in the field for a relatively short time and then move on to other activities because of comparatively low pay, inadequate work conditions, routine work, and limited opportunities for vertical mobility. Journalists often move to other media fields such as public relations or advertising, and less frequently they leave the field of mass media, e.g., by moving to politics. A strong professional mobility between the worlds of journalism and politics, which was particularly characteristic of the former socialist countries, and a too-close relationship of journalists with advertising, entertainment, and public relations may severely endanger the autonomy of journalists and mute their critical voices.

However, when journalists are critical to political and commercial events and actors, as in 'investigative journalism,' they are often exposed to intimidation and harassment by the elites, and even assassinations are used to silence unwanted journalistic voices not only in wartime or against war correspondents. During the 1990s, 662 journalists and media staff were killed or murdered in the course of their work, and many other disappearances were listed as under investigation by the IFJ and the International Press Institute. The killings are only the tip of an iceberg of physical assaults and jailings which affect journalists every year. However, the most widespread and epidemic peril of independent journalism in many countries is an increasing number of lawsuits against journalists, based on the legal provisions that the media should not harm the rights to privacy and reputation of individuals (including public officials); libel actions may be a disguised way of preventing journalists from performing a 'watchdog function' in society. Civil libel compensations became a sword of Damocles particularly in Eastern Europe and Russia in the 1990s, but prosecutions of journalists and high compensations threaten the media with bankruptcy and generally dampen investigative journalism in many other countries with a longer democratic tradition.

In terms of social and demographic characteristics journalists do not represent the population of citizens who rely on their reporting and interpretations, but are expected—like other professions—to carry out their work in the interests of the audiences rather than their own interests. Such expectations may be disputed by the facts that professional journalists disproportionately originate from the families in which professional and white-collar occupations prevail, live in elite neighborhoods, are disproportionately male (although this disparity is lessening during the last few decades), and their views of the world commonly depart from those of an 'average citizen,' usually leaning toward more liberal and democratic political views, in contrast to technical staff in media organizations who are usually more conservative. For example, in a 2000 *Los Angeles Times* survey, 82 percent of US journalists were found to back abortion compared to only 49 percent of the general population. This is not to argue that journalists are a politically coherent population. In some countries, such as Italy, journalists are clearly 'divided' among different political affiliations, a tendency advocated by some contemporary regulatory measures supporting (in the case of press also subsidizing, e.g., in Norway and Sweden) diversity and political plurality of the media.

The composition of the journalistic workforce is generally becoming more diversified in terms of gender, but men still represent the majority of the journalistic workforce. The increase in the number of female journalists indicates that journalism is no more a male preserve, but it also reflects a tendency of the intimization that permeates journalism. Yet the proportion of female journalists is still lower than the proportion of female students of journalism; female journalists are concentrated in the junior posts and in the non-news sector (e.g., magazines), and they tend to be paid less than men despite the fact that they are, on average, better educated than their male counterparts. The differences may be partly a consequence of the fact that female journalists relatively recently entered the profession, but they also may have to do with discrimination against women in journalism.

The least ambiguous mark of 'professionalism' in journalism is the establishment of professional organizations. Journalists are commonly associated nationally either into trade unions or professional associations; frequently, their organizations combine both these functions. Yet a long-term tension exists between professional associations, which stress the professional status of journalists (and usually also include employers and managers), and trade unions, concerned primarily with collective negotiations on behalf of working journalists (newsworkers) with employers. Journalists often tend to define themselves as experts claiming to be above politics, but in some

countries, journalistic associations also pursue political goals. The criteria for membership of journalists' trade unions and professional associations are fluid, particularly in relation to part-time journalists and 'boundary occupations' such as public relations. Trade unions of journalists negotiate for salaries, working conditions, and fringe benefits, and in some cases they also protect journalists' independence, e.g., by financial support to journalists penalized for refusing to do work incompatible with the professional standards (NUJ, National Union of Journalists in the UK) or, as in France, by effectuating the same benefits to journalists who would leave jobs because of the substantial political changes in the editorial policy as to those who lose their jobs involuntarily. In many situations unions also serve as 'tools of the employer' because an imbalance in power exists between unions and employers, due to the lack of unity among unions and guilds, their weak economic position, and goals narrowed to individuals' rewards.

The largest international associations of journalists and media staff are the IFJ and the Media and Entertainment International (MEI). First established in 1926 and relaunched in 1946 and 1952, IFJ is the largest international journalistic organization, with 460,000 members and 140 member-unions in 103 countries on all continents (July 2000). The aims of IFJ include the promotion of international actions to defend press freedom through autonomous national trade unions or (as in the USA) guilds of journalists and other 'authors,' representation of journalists in the international trade union movement, and assistance to journalists and their unions in fighting for their professional and industrial rights. Journalists from socialist and some other countries established in 1946 the International Organization of Journalists (IOJ) as the counterpart to the IFJ. The IOJ, which also claimed the IFJ of 1926 as its predecessor, was dissolved in 1991 after the collapse of communism. Members of 'technical' occupations in the arts, audiovisual, cultural, entertainment, and mass media sector represent the bulk of the membership of Media and Entertainment International which includes 130 trade unions around the world.

See also: Audience Measurement; Information Society; Media Ethics; News: General; News Interview; Radio as Medium; Television: History

Bibliography

Bourdieu P 1998 *On Television—The Power of Journalism*. The New Press, New York

Dahlgren P, Sparks C (eds.) 1991 *Communication and Citizenship: Journalism and the Public Sphere in the New Media Age*. Routledge, London

Hardt H, Brennen B (eds.) 1995 *Newsworkers: Toward a History of the Rank and File*. University of Minnesota Press, Minneapolis, MN

Kunczik M 1988 *Concepts of Journalism North and South*. Friedrich-Ebert-Stiftung, Bonn, Germany

Javnost–The Public 1996 Special issue. Journalism at the crossroads (3, 3)

Journal of Communication 1992 Special issue. Journalism in crisis and change (42, 3)

Lippmann W [1922] 1960 *Public Opinion*. Macmillan, New York

Merrill J C 1983 *Global Journalism*. Longman, New York

Park R E 1967 *On Social Control and Collective Behavior*. University of Chicago Press, Chicago

Schudson M 1978 *Discovering the News: A Social History of American Newspapers*. Basic Books, New York

Siebert F, Peterson T, Schramm W 1956 *Four Theories of the Press*. University of Illinois Press, Urbana, IL

Splichal S, Sparks C 1994 *Journalists for the 21st Century*. Ablex, Norwood, NJ

Weber M [1910], 1976 Toward a sociology of the press. *Journal of Communication* **26**(3): 96–101

S. Splichal

Judaism

Judaism, a religion of ethical monotheism in the class of Christianity and Islam, encompasses diverse, but kindred religious systems ('Judaisms') that in common exhibit these indicative traits: (a) they maintain that God is unique and made manifest in self-revelation, set forth to Moses at Sinai in the Torah; (b) they privilege the Pentateuch (Five Books of Moses, aka the Torah) among the Israelite Scriptures; and (c) they regard the Jews, or some Jews, at later times and in other places as the continuation of Scripture's 'Israel' in the Land of Israel. All Judaic religious systems—cogent theories of the social order, grounded in the Pentateuch, comprised by a way of life, a world view, and an account of who and what is 'Israel'—exhibit those three indicative traits. In antiquity a number of Judaic religious systems flourished, one represented by the Dead Sea Scrolls, another by the Elephantine Papyri, a third by the writings of Philo of Alexandria. The most influential and enduring is called Rabbinic Judaism (after the title of its principal authority, the rabbi), or Talmudic Judaism (for its principal document beyond Scripture), or classical or normative Judaism (for its paramount status). Rabbinic Judaism attained normative status in ancient times and retained a paramount position until the nineteenth century, when competing Judaisms took shape.

1. Rabbinic Judaism

Rabbinic Judaism in the fourth and fifth centuries AD responded to the crisis of the Christianization of the Roman Empire. With the triumph of Christianity through the conversion of Constantine in 312, Christianity's explicit claims, now validated in world-

shaking events of the age, demanded a reply. The sages of the Talmud of the Land of Israel provided it. At those very specific points at which the Christian challenge met head-on old Israel's world-view, sages' doctrines responded. The response took three forms. The Torah ('Judaism') was defined in the doctrine, first, of the status, as oral and memorized revelation, of the Mishnah, and, by implication, of other rabbinical writings. The Torah, moreover, was presented as the encompassing symbol of Israel's salvation. The Torah, finally, was embodied in the person of the Messiah who would be a rabbi. The Torah in all three modes confronted the cross, with its doctrine of the triumphant Christ, Messiah and king, ruler now of earth as of heaven. Judaism in the rabbis' statement did endure in the Christian West, imparting to Israel the secure conviction of constituting that Israel after the flesh to which the Torah continued to speak. When Islam gained its victory, Christianity throughout the Middle East and North Africa gave way, Judaism in those same vast territories retained the loyalty and conviction of the people of the Torah.

1.1 From Israel, the Nation Dwelling Alone to, the Jews, an Ethnic Group of Individual Citizens of the Nation State

A drastic shift in the political circumstance of a Judaism in the West, represented by the American Constitution of 1787 and the French Revolution of 1789, affected Jews' thought about perennial questions. From the end of the eighteenth century the secularization of political life and institutions defined a question not easily answered by Rabbinic Judaism. Earlier modes of organizing matters had recognized as political entities groups and guilds and classes, and, among them, the Jews found a place. In the hierarchical scheme, with church and monarchy and aristocracy in their proper alignment, other political entities could likewise find their location. With church disestablished, monarchy rejected, and aristocracy no longer dominant in politics, the political unit became the (theoretically) undifferentiated individual, making up the nation state. Within that theory was no room for a collectivity such as 'Israel,' the holy community, viewed as a political unit. But, political theory held, there might be room for the Jewish individual, in rightful place alongside other undifferentiated individuals. The individuation of politics produced a considerable crisis for the Rabbinic Judaism.

1.2 Modern and Contemporary Judaisms in the Context of Rabbinic Judaism

Rabbinic Judaism encompassed the whole of the existence of the Jews, who found its truth self-evident, its definition of life ineluctable. In the way of life of that system a Jew was not always a Jew, he or she was only a Jew. The other Judaisms acknowledged the former. Jews never stopped being Jews. The world would not let them, even if they wanted to. But Rabbinic Judaism made slight provision for Jews to be anything but Jews: the holy people had no other vocation, no alternative, to its holiness. Its history as a people different in kind from other peoples, its destiny at the end of time; these matched its distinctive holy way of life in the here and now. So Israel was always Israel and only Israel. But in modern times Israel became one of several things that Jews would be: also Americans, also workers, also Israelis, among the twentieth-century Judaisms, but never only Israel, God's people. And that theory of Israel matches in social terms the conception of the individual person as well. For in the received Torah, Israel—the Jew—lived out life in the rhythm of sanctification of the here and now, realizing in concrete deeds the Torah's words, once more, not only always but also Israel. It was not a romance, it was a marriage.

2. Emancipation

A process called 'emancipation,' part of a larger movement of emancipation of serfs, women, slaves, Catholics (in Protestant countries, for instance, England and Ireland), also encompassed the Jews. Benzion Dinur defines this process of emancipation as follows:

> Jewish emancipation denotes the abolition of disabilities and inequities applied specially to Jews, the recognition of Jews as equal to other citizens, and the formal granting of the rights and duties of citizenship. Essentially the legal act of emancipation should have been simply the expression of the diminution of social hostility and psychological aversion toward Jews in the host nation … but the antipathy was not obliterated and constantly hampered the realization of equality even after it had been proclaimed by the state and included in the law (Dinur, in: 'Emancipation,' *Encyclopaedia Judaica* **6**: 696–718).

The political changes that fall into the process of the Jews' emancipation began in the eighteenth century, and, in a half-century affected the long-term stability that had characterized the Jews' social and political life from Constantine onward. These political changes raised questions not previously found urgent, and, it follows, also precipitated reflection on problems formerly neglected. The answers to the questions flowed logically and necessarily from the character of the questions themselves.

Dinur traces three periods in the history of the Jews' emancipation, from 1740 to 1789, ending with the French revolution; then from 1789 to 1878, from the French revolution to the Congress of Berlin; and from

1878 to 1933, from the Congress of Berlin to the rise of the Nazis to power in Germany. The adoption of the American Constitution in 1787 confirmed the US position on the matter. Jewish males enjoyed the rights of citizens, along with all other whites. The first period marks the point at which the emancipation of the Jews first came under discussion, the second marked the period in which Western and Central European states accorded to the Jews the rights of citizens, and the third brought to the fore a period of new racism that in the end annihilated the Jews of Europe.

In the first period advocates of the Jews' emancipation maintained that religious intolerance accounted for the low caste status assigned to the Jews. Liberating the Jews would mark another stage in overcoming religious intolerance. During this first period the original ideas of Reform Judaism came to expression, although the important changes in religious doctrine and practice were realized only in the earlier part of the nineteenth century.

In the second period, the French revolution brought Jews political rights in France, Belgium, the Netherlands, Italy, Germany, and the Austro-Hungarian Empire. As Germany and Italy attained unification and Hungary independence, the Jews were accorded the rights and duties of citizenship. Dinur explains: 'It was stressed that keeping the Jews in a politically limited and socially inferior status was incompatible with the principle of civic equality … it is the objective of every political organization to protect the natural rights of man, [hence] all citizens have the right to all the liberties and advantages of citizens, without exception.' Jews at that time entered the political and cultural life of the Western nations, including their overseas empires (hence Algerian Jews received French citizenship). During this second period Reform Judaism reached its first stage of development, beginning in Germany. It made it possible for Jews to hold together the two things they deemed inseparable: their desire to remain Jewish, and their wish also to be one with their 'fellow citizens.' By the middle of the nineteenth century, Reform had reached full expression and had won the support of a sizable part of German Jewry. In reaction against Reform ('the excesses of …'), Orthodoxy came into existence. Orthodoxy no less than Reform asked how 'Judaism' could coexist with 'Germanness,' meaning citizenship in an undifferentiated republic of citizens. A centrist position, mediating between Reform and Orthodoxy, was worked out by theologians in what was then called the Historical School, and what, in twentieth-century America, took the name of Conservative Judaism. The period from the French Revolution to the Congress of Berlin therefore saw the full efflorescence of all of the Judaisms of political modernization. All of these Judaisms characterized the Jews of Western Europe, and, later on, America. But in America Reform, Orthodoxy, and the Historical School or Conservative Judaism radically changed in character, responding to the urgent issues of a different circumstance, producing selfevidently valid answers of a character not compatible with the nineteenth-century statements of those same systems.

In the third period, antiSemitism as a political and social movement attained power. Jews began to realize that, in Dinur's words, 'the state's legal recognition of Jewish civic and political equality does not automatically bring social recognition of this equality.' The Jews continued to form a separate group; they were racially 'inferior.' The impact of the new racism would be felt in the twentieth century. Judaisms of the twentieth century raised the questions of political repression and economic dislocation, as these faced the Jews of Eastern Europe and America.

3. *The Nineteenth Century Judaisms and the Political Question*

Reform, Orthodoxy, and Conservative carried forward Rabbinic Judaism, each in its own claim forming the natural and legitimate next step in the ongoing life of that Judaism. All make explicit their claim to stand in a linear, unitary, harmonious relationship with the dual Torah of Sinai, to form the necessary development of the Torah. Each of these three Judaisms negotiated issues of secularity, on the one side, and social and cultural and political change, on the other. None affirmed an essentially secular view, nor did any one of them formulate its systemic statement outside the framework of the dual Torah. The three main Judaisms born in the aftermath of the advent of modern politics have now to be specified. All of them continued the Torah as set forth in Rabbinic Judaism and adopted the Torah as their generative symbol and myth, its law as their norms, its theology as their touchstone. Between 1800 and 1850 all had taken shape.

Chronologically first is Reform Judaism, coming to expression in the early part of the nineteenth century and making changes in liturgy, then in doctrine and in way of life of the received Judaism of the dual Torah. Reform Judaism recognized the legitimacy of making changes and regarded change as reform, yielding Reform. Second was the reaction to Reform Judaism, called Orthodox Judaism, which in many ways was continuous with the Judaism of the dual Torah, but in some ways as selective of elements of that Judaism as was Reform Judaism. Orthodox Judaism reached its first systematic expression in the middle of the nineteenth century. Orthodox Judaism addressed the same issue, that of change, and held that Judaism lies beyond history; it is the work of God; it constitutes a set of facts of the same order as the facts of nature. Hence change is not reform, and Reform Judaism is not Judaism—so Orthodoxy. Third in line, and somewhat after Orthodox Judaism, was positive Historical

Judaism, known in America as Conservative Judaism, which occupied the center between the two other Judaisms of continuation of the dual Torah. This Judaism maintained that change could become reform, but only in accord with the principles by which legitimate change may be separated from illegitimate change. Conservative Judaism would discover those principles through historical study. In an age in which historical facts were taken to represent theological truths, the historicism of Conservative Judaism bore compelling weight.

Each of these Judaic systems exhibited three characteristic traits: (a) each asked how one could be both Jewish and something else, that is, also a citizen, a member of a nation; (b) it defined 'Judaism' (that is, its system) as a religion, so leaving ample space for that something else, namely, nationality; (c) it appealed to history to prove the continuity between its system and the received Judaism of the dual Torah. The resort to historical fact, the claim that the system at hand formed the linear development of the past, the natural increment of the entire 'history' of Israel, the Jewish people, from the beginning to the new day—that essentially factual claim masked a profound conviction concerning selfevidence. The urgent question at hand—the political one—produced a selfevidently correct answer out of the history of politics constituted by historical narrative.

That appeal to history, particularly historical fact, characterizes all three Judaisms. The Reformers stated explicitly that theirs would be a Judaism built on fact. The facts of history, in particular, would guide Jews to the definition of what was essential and what could be dropped. History then formed the court of appeal—but also the necessary link, the critical point of continuity. The Historical School took the same position, but reached different conclusions. History would show how change could be effected, and the principles of historical change would then govern. Orthodoxy met the issue in a different way, maintaining that 'Judaism' was above history, not a historical fact at all. But the Orthodox position would also appeal most forcefully to the past in its claim that Orthodoxy constituted the natural and complete continuation of 'Judaism' in its true form. The importance of history in the theological thought of the nineteenth-century Judaisms derives from the intellectual heritage of the age, with its stress on the nation state as the definitive unit of society, and on history as the mode of defining the culture and character of the nation state. History as an instrument of reform, further, had served the Protestant Reformation, with its appeal to Scripture as against (mere) tradition, its claim that it would restore Christianity to its (historical) purity. Finally and most important, the supernaturalism of the inherited Judaism of the dual Torah, its emphasis upon God's active intervention in history, on miracles, on a perpetual concern for the natural implications of the supernatural will and covenant—that supernaturalism contradicted the rationalism of the age. The one thing the Jewish thinkers wished to accomplish was to show the rationalism, the reason—the normality—of the Judaisms they constructed. Appealing to (mere) facts of history, as against the unbelievable claims of a Scripture placed upon a positive and this-worldly foundation that religious view of the world that, in the received system of the dual Torah, rested upon a completely supernatural view of reality.

Just as the questions before all three important Judaisms of the nineteenth century were the same, the answers of the three systems remarkably congruent to one another. The three Judaisms of continuity exhibit striking traits in common. All looked backward, at the received system of the dual Torah. All sought justification in precedent out of a holy and paradigmatic past. All viewed the documents of that system as canonical, differing, of course, on the relative merit of the several components. They concurred that texts to prove propositions deemed true should derive from those canonical writings (or from some of them). All took for granted the enduring, God-given authority of those writings. None doubted that God had revealed the (written) Torah at Sinai. All looked for validating precedent in the received canon. Differing on issues important to both world view and way of life, all three Judaisms concurred on the importance of literacy in the received writings, on the lasting relevance of the symbolic system at hand, on the pertinence of the way of life (in some, if not in every detail), on the power of the received Judaism of the dual Torah to stand in judgment on whatever, later, would serve to continue that Judaism.

True, the differences among the three Judaisms impressed their framers and with good reason. The Reformers rejected important components of Rabbinic Judaism and said so. Written Torah, yes, Oral Torah, no. The Orthodox explicitly denied the validity of changing anything, insisting on the facticity, the givenness, of the whole. The Conservatives, in appealing to historical precedent, shifted the premise of justification entirely. Written Torah, yes, Oral Torah, maybe. They sought what the Orthodox thought pointless and the Reform inconsequential, namely, justification for making some few changes in the present in continuation of the processes that had effected development in the past. None of these points of important difference proved trivial. But all of them, all together, should not obscure the powerful points of similarity that mark all three Judaisms as continuators of the Judaism of the dual Torah. The points at which each Judaism took its leave from the received system do not match. In the case of Reform, the break proved explicit: change carried out by articulate, conscious decision, thus change as a matter of policy, enjoys full legitimacy. And as for the positive Historical School and of its continuators in Conservative Judaism, the gulf between faith and fact took the measure of the

difference between the received system of the dual Torah and the statement of mere historical facts that, for the Historical School, served to document the faith. In saying that 'things have changed in the past, and we can change them too,' Reform established its primary position. It pointed to precedent, and took as implicitly conceded the power of the received system to stand in judgment. All the more so did the Orthodox and Conservative theologians affirm that same power and place themselves under the judgment of the Judaism of the dual Torah. All three established a firm position within the continuation of that Judaism. Each claimed to take priority as the next step in the linear and incremental history of Judaism.

4. *The Twentieth Century and its Mythic Ideologies*

Three Judaisms were born in the twentieth century, two in 1897, one in 1967. The first was Jewish Socialism and Yiddishism; the second, Zionism; and the third, three generations later, the American Judaic system of Holocaust and Redemption. Jewish Socialism took shape in the Bund, a Jewish union organized in Poland in 1897. Zionism was founded in the World Zionist Organization, created in Basel in 1897. American Judaism—'the Judaism of Holocaust and Redemption,' explaining in ethnic and political terms why and where one should be Jewish in the formulation under discussion—came to powerful expression in the aftermath of the 1967 war in the Middle East. All three Judaic systems answered profoundly political questions. Their agenda attended to the status of the Jews as a group (Zionism, American Judaism), the definition of the Jews in the context of larger political and social change (Jewish Socialism, Zionism). It follows that the urgent questions addressed by the twentieth-century Judaisms differed in kind from those found acute in the nineteenth century. In the twentieth century, powerful forces for social and economic change took political form, in movements meant to shape government to the interests of particular classes or groups, the working classes or racial or ethnic entities, for instance. The Judaic systems of the century responded in kind.

Jewish Socialism presented a Jewish-ethnic system congruent to the political task of economic reform through state action. The Jews would form unions and engage in mass activity of an economic, and ultimately, therefore, of a political character. In that same century the definition of citizenship, encompassing ethnic and genealogical traits, presented the Jews with the problem of how they were to find a place in a nation state that understood itself in an exclusionary and exclusive, racist way—whether Nazi Germany, nationalist Poland, Hungary, Romania or France. Zionism declared the Jews 'a people, one people,' and proposed as its purpose the creation of the Jewish State. Later on, shifting currents in American politics, a renewed ethnicism and emphasis on intrinsic traits of birth, rather than extrinsic ones of ability, called into question the Jews' identification with the democratic system of America as that system defined permissible difference. A Jewish ethnicism, counterpart to the search for roots among diverse ethnic groups, responded with a tale of Jewish 'uniqueness'—unique suffering—and unique Jewish ethnic salvation, redemption in the Jewish State. So three powerful and attractive movements, Jewish Socialism, Zionism, and American Judaism, presented answers to critical issues confronting groups of Jews. All of these movements addressed political questions and responded with essentially political programs. Zionism wanted to create a Jewish state, American Judaism wanted the Jews to form an active political community on their own, and Jewish Socialism in its day framed the Jews into political, as much as economic, organizations, seeing the two as one, a single and inseparable mode of defining economic activity and public policy.

5. *Comparing the Nineteenth- and the Twentieth-century Judaisms*

On the surface the three Judaic systems of the twentieth century took up political, social, economic, but not theological questions. While the nineteenth-century Judaisms addressed issues particular to Jews, the matters of public policy of the twentieth-century Judaic systems concerned everyone, not only Jews. All of the new Judaisms intersected with comparable systems—like in character, unlike in content—among other Europeans and Americans. Socialism then is the genus, Jewish Socialism the species; American ethnic assertion the genus, American Judaism the species. The issues addressed by the Judaisms of the twentieth century, the crises that made those issues urgent, did not affect Jews alone or mainly. The crises in common derived from economic dislocation, which generated socialism, and also Jewish Socialism; the reorganization of political entities, which formed the foundation of nationalism, and also Zionism; and the reconsideration of the theory of American society, which produced, alongside the total homogenization of American life, renewed interest in ethnic origins, and also American Judaism.

The nineteenth-century Judaisms made constant reference to the received system of the dual Torah, its writings, its values, its requirements, its viewpoints, its way of life. The twentieth-century Judaisms did not. True, each Judaism born in the nineteenth century faced the task of validating change that all of the nascent Judaisms in one way or another affirmed. But all of the new Judaisms articulated a principle of change guiding relationships with the received system,

which continued to define the agenda of law and theology alike, and to which, in diverse ways, all the Judaisms recognized themselves as answerable. We cannot point to a similar relationship between the new Judaisms of the twentieth century and the received Judaism of the dual Torah. For none of them made much use of the intellectual resources of that system, found important issues deemed urgent within that system, or even regarded itself as answerable to the Judaism of the dual Torah.

For the twentieth-century systems birth came about within another matrix altogether: the larger world of socialism and linguistic nationalism, for Jewish Socialism and Yiddishism; the realm of the nationalisms of the smaller peoples of Europe, rejecting the government of the international empires of Central and Eastern Europe, for Zionism; the reframing, in American culture, of the policy governing social and ethnic difference, for American Judaism. None of these Judaic systems of believing and behaving drew extensively on the received Judaic system of the dual Torah, and all of them for a time vastly overshadowed in acceptance among the Jewish group the Judaisms that did. So the passage of time, from the eighteenth to the twentieth century, produced a radical attenuation of the bonds that joined the Jews to the Judaism of the dual Torah. Our interest concerns not explaining the social change at hand, nor even in invoking the political explanation that seems to me paramount, but only in reviewing the simple program with which we began.

The Judaisms of the nineteenth century attained a high measure of self-consciousness because they had before their eyes the image of the innocent faith of their predecessors, and many in their own time as well. The Judaisms of the twentieth century, abandoning all pretense at a connection to the received Judaism, lost also the awareness that change took place, people made choices, and, in all, a dimension of decision-making took the measure of their Judaisms. So they entered a new phase of selfevidence, appealing now for vindication not to received texts but the obvious facts of the everyday world. So history now proved propositions for the new Judaisms, and the text of those Judaisms was the world out there. Proof texts derived from headlines in newspapers. The continuators of Rabbinic Judaism developed systems of belief and behavior that invariably fell into the category of religions, in our setting, Judaisms. Whether or not the twentieth-century successor systems constitute religions, Judaisms, presents a question bearing only slight consequence. Clearly, Jewish Socialism and Zionism provided the deep meaning for the lives of millions of Jews, so, defining ways of life and world views and the Israel subject to realize both, they functioned entirely as did the religions, the Judaisms, of the nineteenth century and of the twentieth as well. But Socialism–Yiddishism and Zionism differ from the continuator-Judaisms, because neither invoked a supernatural God, revelation of God's will in the Torah, belief in Providence, or any other indicator of the presence of the family of closely-related religions, the Judaisms; and, moreover, their framers and founders did not claim otherwise. Had the framers alleged that theirs was a continuator-Judaism, we should have to introduce that fact into our analysis and interpretation, but none did. The Socialists took a position actively hostile to religion in all forms, and the Zionists compromised with the religious Judaisms of the day but in no way conceded that theirs was a competing Judaism. As to American Judaism, in its contemporary form, in its appeal the salvific myth of Holocaust and Redemption, it crosses the border between a genuinely religious and an entirely secular system addressed to an ethnic group, the Jews, not a religious community, 'Israel.'

5.1 Influence of Twentieth-century Systems

No Judaic religious systems of broad influence took shape in the twentieth century. Three factors explain why.

5.1.1 The Holocaust. First let us consider the demographic factor. The most productive sector of world Jewry perished, and the social and political and cultural conditions that put forth the great systemic creations vanished with the six million who died. Judaic systems in all their variety emerged in Europe, not in America or in what was then Palestine and is now the State of Israel, and, within Europe, they came from Central and Eastern European Jewry. We may account for the systemopoia of Central and Eastern European Jews in two ways: (a) the Jews in the East, in particular, formed a vast and coherent population, with enormous learning and diverse interests; (b) the systems of the nineteenth and twentieth centuries arose out of a vast population living in self-aware circumstances, not scattered and individual but composed and bonded. The Jews who perished formed enormous and self-conscious communities of vast intellectual riches.

5.1.2 The demise of intellect. The second explanation for the end of systemopoia is the as yet unappreciated factor of sheer ignorance, the collapse of the system of cultural transmission of Judaism. That second factor is constituted by the utter loss of access to that permanent treasury of the human experience of Jewry preserved and handed on in the canonical Torah. The Judaisms that survive—political and ethnic in the main—do provide ready access to emotional or political encounters, readily available to all by definition. But they offer none of that confrontation of taste and judgment, intellect and reflection,

that takes place in traditional cultures and with tradition. The twentieth-century systems resort mainly to the immediately accessible experiences of emotions and of politics. The repertoire of human experience in Rabbinic Judaism presents as human options the experience of the ages. By contrast Israeli nationalism and American Judaism—the two most influential systems that move Jews to action in the world today—scarcely concern themselves with that Judaism. They work with the raw materials made available by contemporary experience: emotions on the one side, politics on the other. Access to realms beyond requires learning in literature, the only resource for human experience beyond the immediate. But the Judaic systems of the twentieth century do not resort to the reading of books as a principal act of their way of life, in the way in which Rabbinic Judaism and its continuators did and do. The consequence is a strikingly abbreviated agenda of issues, a remarkably one-dimensional program of urgent questions.

5.1.3 The triumph of large-scale organization. Third and distinct from the other two is the bureaucratization of Jewry in consequence of the political tasks it rightly has identified as urgent. Jews have had to work through very large organizations and institutions. The contemporary class structure of Jewry therefore places in positions of influence those Jews who place slight value on matters of intellect and learning, and that same system accords no sustained hearing to Jews who strive to reflect. The exemplary experiences of those who exercise influence derives from politics, through law, from economic activity, through business and industry, from institutional careers, through government, politics and law. As the gifts of establishing routine take precedence over the endowments of charisma of an intellectual order, the experiences people know and understand—politics, emotions of ready access—serve, also, for the raw materials of Judaic system-building. The great war conducted against the Jews, beginning not in 1933 but with the organization of political antiSemitism joined to economic exclusion, from the 1880s onward, effected the change. In a profound sense the type of structure now characteristic of Jewry represents one of the uncounted costs of the Holocaust.

See also: Anti-Semitism; Buddhism and Gender; Diaspora; Genocide: Anthropological Aspects; Historiography and Historical Thought: Jewish Tradition; Islam and Gender; Jewish Law; Judaism; Judaism and Gender; Religion: Evolution and Development; Religion, Sociology of; Zionism; Zionism, History of

Bibliography

Adler H G 1969 *The Jews in Germany from the Enlightenment to National Socialism*. University of Notre Dame Press, IN

Albert P C 1977 *The Modernization of French Jewry: Consistory and Community in the Nineteenth Century*. Brandeis University Press, Hanover, NH

Baron S W 1942/1972 *The Jewish Community: Its History and Structure to the American Revolution*. Greenwood Press, Westport, CT, 3 Vols.

Baron S W 1973 *A Social and Religious History of the Jews*, 2nd edn. Columbia University Press, New York, 15 Vols.

Chazan R (ed.) 1980 *Church, State and Jew in the Middle Ages*. Behrman House, New York

Dawidowicz L S 1984 *The Golden Tradition: Jewish Life and Thought in Eastern Europe*. Schocken, New York

Dawidowicz L S 1976 *The War against the Jews, 1933–1945*. Rinehart and Winston, New York

Eisenstadt S N 1992 *Jewish Civilization: The Jewish Historical Experience in a Comparative Perspective*. State University of New York Press, Albany, NY

Encyclopedia Judaica 1972 Macmillan, New York, 16 Vols.

Feldman L H 1993 *Jew and Gentile in the Ancient World: Attitudes and Interactions from Alexander to Justinian*. Princeton University Press, Princeton, NJ

Finkelstein L (ed.) 1979 *The Jews: Their History, Culture, and Religion*. Greenwood Press, Westport, CT, 2 Vols.

Goldscheider C, Zuckerman A S 1986 *The Transformation of the Jews*. University of Chicago Press, Chicago

Grabbe L 1992 *Judaism from Cyrus to Hadrian. I. The Persian and Greek Periods. II. The Roman Period*. Fortress Press, Minneapolis, MN

Grayzel S 1960 *History of the Contemporary Jews from 1900 to the Present*. Meridian Books, New York

Guttmann J 1964 *Philosophies of Judaism: The History of Jewish Philosophy from Biblical Times to Franz Rosenzweig*, 1st edn. Holt, Rinehart and Winston, New York

Kedourie E (ed.) 1979 *The Jewish World: History and Culture of the Jewish People*. Abrams, New York

Lewis B 1984 *The Jews of Islam*. Princeton University Press, Princeton, NJ

Mahler R 1971 *A History of Modern Jewry*. Schocken Books, New York

Malamat A, Ben-Sasson H (eds.) 1976 *A History of the Jewish People*. Harvard University Press, Cambridge, MA

Mendes-Flohr P R, Reinharz J (eds.) 1980 *The Jew in the Modern World: A Documentary History*. Oxford University Press, New York

Meyer M A 1972 *The Origins of the Modern Jew: Jewish Identity and European Culture in Germany, 1749–1824*. Wayne State University Press, Detroit, MI

Neusner J 1990 *Self-fulfilling Prophecy: Exile and Return in the History of Judaism*. Scholars Press, Atlanta, GA

Schwarz L W (ed.) 1956 *Great Ages and Ideas of the Jewish People*. Random House, New York

Seltzer R M 1982 *Jewish People, Jewish Thought: The Jewish Experience in History*. Macmillan, New York

Sharot S 1976 *Judaism: A Sociology*. Holmes & Meier, New York

Stillman N A 1979 *The Jews of Arab Lands: A History and Source Book*, 1st edn. Jewish Publications Society of America, Philadelphia

Talmage F (ed.) 1975 *Disputation and Dialogue: Readings in the Jewish-Christian Encounter*. Ktav Publishing House, New York

J. Neusner

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Judaism and Gender

Contemporary Jewish feminist scholarship (Hyman, Hyman and Moore, Peskowitz and Levitt, Plaskow, Alpert, Heschel, Prell, Boyarin, Baskin, Adler, Kaplan) and activism (Pogrebin, Klepfisz and Kaye/Kantrowitz, Brettschneider, Beck, Balka and Rose, Barack Fishman, Weidman Schneider) are transforming notions of Jews, Jewishness, and Judaism. Through these efforts Jewish culture, religion and history are being transformed. The simple logic of inclusion is giving way to more radical reconfigurations of the Jewish past, present, and future.

Jewish feminist activists and scholars were a part of the broader feminist movement not only in the so-called second wave of US feminist activism but even earlier efforts (Kuzmack, Hyman, Kaplan). By the 1970s they were demanding that Jewish women be included in all aspects of Jewish communal life as equals (Pogrebin).

Building on Virginia Woolf's suggestion that we add a supplement to history as a way of remedying the inequities between men and women, feminist historian Joan Scott calls our attention to the ambiguities within what she has termed 'the logic of the supplement.' As Scott explains, 'women are both added to history and they occasion its rewriting; they provide something extra and they are necessary for completion, they are superfluous and indispensable' (Scott 1993, p. 241). This contradictory logic is at the heart of what it means to think about gender, Jews, Jewishness, and Judaism in the beginning of the twenty-first century. A consideration of women much less the larger category of gender in the context of Jewish communal, religious and/or scholarly life, has resulted in a radical rethinking of the basic assumptions that have defined all dimensions of contemporary Jewish life. On the one hand, the inclusion of women within Jewish life constitutes a simple expansion of existing narratives. On the other hand, it has made clear that any overview of Jewish life without a consideration of gender would be both inaccurate and incomplete.

1. *Some Historical Background*

As feminist scholars have argued (Hyman 1995, Kaplan 1991, Levitt 1997), the call for inclusion has its origins in the legacy of Jewish emancipation in the West where individual middle-class Jewish men were granted rights as citizens of liberal states. They were not emancipated as Jews. By giving up various forms of Jewish communal authority, they came to see themselves reflected in the universal democratic vision of the French and American revolutions.

Although these revolutions excluded many, their universal vision offered the promise of emancipation to many who remained disenfranchised within Western culture including middle-class women. Jewish women whose husbands, brothers, uncles, and fathers had become citizens, began to ask questions about their own status. Using the logic of these liberal revolutions, by the late nineteenth century Jewish feminists became active in larger women's movements. With other women they fought for the right to own property and the right to vote (Kaplan 1979, Kuzmack 1990, Hyman 1995) but these struggles also inspired them to challenge various inequities within Jewish communal life as well. These included the dramatic increase in the number of Jewish women attending synagogues for public worship and the opening up of once gender segregated seating (Peskowitz and Levitt 1997). By the 1970s Jewish women used notions of liberal justice and equality to demand their inclusion in all aspects of communal worship. By arguing that specific Jewish communal practices either excluded women or treated them as inferior to men, they demanded their rightful place on the *bema*, reading torah and leading prayer services. In these ways, they were able to critique the masculinist bias of traditional worship. This liberal critique has expanded to include Jewish social and communal practices from religious observance to the various fields of Jewish studies scholarship.

The increasing presence of Jewish women and attention to issues of gender have occasioned the rewriting of Jewish history and the rereading of classical Jewish texts in the various fields of Jewish Studies scholarship (religion, literature, history, sociology, political science, and language study) as well as a resurgence in Jewish cultural production including art, poetry, fiction, film, music, and dance.

2. *Jewish Religious and Communal Life*

In Jewish religious worship these changes have been palpable. They include a transformation in the language of prayer. Not only is there now gender inclusive language but whole new liturgies that take account of the experiences of Jewish women (Falk 1996, Adler 1998). Since the 1970s, in increasing numbers of Jewish communities, women are engaging in once restricted religious practices. They are chanting sacred texts, participating in worship services, having honors once granted only to Jewish men and they are leading communal worship as rabbis and cantors (Nadel 1998). They are also becoming presidents of congregations and Jewish religious organizations. Although many of these efforts remain contested especially among various Orthodox Jewish communities, among liberal Jews, reform, conservative, reconstructionist and renewal Jewish communities, the inclusion of women in all of these realms including their role as rabbis and cantors are now very much established practices especially in the United States. Even among certain modern orthodox communities these roles and practices are changing. The growth of orthodox Jewish

feminist prayer groups, study centers, conferences, and the creation of women rabbinic interns throughout the 1990s all attest to the growth of an Orthodox Jewish feminist movement (Greenberg 1981, Grossman and Haut 1992). Even within ultra-orthodox communities, women are reclaiming the rituals of family purity, for example, as attractive Jewish practices for women (El Or 1994, Davidman 1991, Kaufman 1991, Ochs 1990).

Jewish women are also very much a part of passing on Jewish religious traditions through the work of Jewish education. Since the nineteenth century, Jewish women have playing an ever increasing role in the education of Jewish children in what had been an exclusively male domain. Jewish women are increasingly running Jewish day schools sponsored by the various religious movements and training day school teachers and administrators in graduate programs in Jewish education. Among liberal Jews, women are playing an increasing role in the education of future rabbis and cantors. The number of women on the faculties of the conservative, reform, and reconstructionist rabbinical colleges are growing. Jewish women are also taking on leadership roles as administrators at these institutions.

In all of these realms there is a growing commitment to include women in the vital life of Jewish religious practices and Jewish communal life more broadly understood (Barack Fishman 1993, Prell 1999). Increasingly Jewish women are taking on a variety of leadership roles both in North America and in Israel, running Jewish agencies, social service organizations, and federations.

Through these efforts other changes are also being fostered. Feminist efforts have also opened up Jewish communal life to gay and lesbian Jews. Here lesbian Jewish feminists have made substantial contributions as rabbis, lay leaders, activists, and scholars (Alpert 1997, Plaskow 1990, Beck 1982, Balka and Rose 1989). They are powerfully challenging notions of family, community, and religiousness by their strong presence within and outside of organized Jewish communities. They are challenging the criteria for membership in Jewish congregations and communal organizations; constructing new rituals of recognition such as commitment ceremonies and events to celebrate 'coming out' in a Jewish context. They have also demanded that the Jewish community address homophobia and the impact of the AIDS epidemic on Jews.

3. *Jewish Studies Scholarship*

The various fields of Jewish Studies scholarship have also seen the addition of women as a precursor to more critical transformations of scholarly practice particularly in the fields of textual study, religion, history, literature, and the various social sciences. Although Jewish women scholars were already active in the academic study of Judaism beginning in the late nineteenth century, the most powerful and dramatic changes came with the second wave of US feminist activism. Feminist scholars have not only added women to the study of ancient Jewish text, Jewish philosophy, theology, history, literature, and social study, they have raised questions about both the content of Jewish studies courses and the ways they are being taught in colleges and universities (Peskowitz and Levitt 1997, Davidman and Tenenbaum 1994, Rudavsky 1995).

Work on classical Jewish texts, Hebrew bible (Pardes 1992) and especially rabbinic literature (Peskowitz 1997, Boyarin 1993, Hauptman 1998, Wegner 1988, Adler 1998, Biale 1984) have changed how these sources are read. Feminist scholars have not only looked at how Jewish women figure in these texts, they are increasingly raising questions about how gender (both masculinity and femininity) is produced in these texts and how these visions of gender have become normative. Here reading practices are crucial. As Miriam Peskowitz notes, such normative texts can and should be read as legal fantasies following the lead of recent scholars in critical legal theory. As Peskowitz argues, these texts tell us about the desires of those who produced them.

Jewish feminist scholars have also taken on the traditions of Jewish theology, ethics, and philosophy to offer new approaches to these various fields of Jewish religious studies scholarship (Adler 1998, Plaskow 1990, Alpert 1997, Levitt 1997, Falk 1996). These scholars not only question the gender of God but the subject of ethical discourse. Who is the person making moral choices and from what standpoint are these issues being addressed. They also challenge the mind–body dualisms that mark medieval as well as modern Jewish thought and their implications for Jewish women.

In Jewish history, feminist scholars have used the addition of women to their readings of the Jewish past to rethink some of the grand narratives that have shaped Jewish historical study. By demonstrating the role women played in emancipation and enlightenments, for example, modern Jewish historians have shown us how vital Jewish women were to these transformations in Jewish life (Hyman 1995, Kaplan 1979, 1991, Kuzmack 1990). Jewish feminist historians have also called our attention to the role of Jewish women in assimilation into America culture (Hyman and Moore 1997, Stahl Weinberg 1988, Glenn 1990, Prell 1999).

Broader works on twentieth-century Jewish history have also been transformed by accounts of gender. Included in this scholarship are powerful accounts of zionism as a masculinist national movement (Berkowitz 1993, Boyarin 1997, D. Biale 1992) as well as a growing body of literature about the role of women in the devastation of the holocaust (Ofer and Weitzman 1998, Rittner and Roth 1993). As Daniel Boyarin

points out, Zionism was a way of remasculinizing the Jewish people. On the other hand, feminist scholars have begun to examine how the experience of genocide was gendered. They point to how and in what ways pregnancy, menarche, menstruation, and motherhood marked Jewish women in the holocaust. They also address the particular devastation of sexual assault in this context. Finally, women's historians have called attention to the various roles of Jewish women in the ancient past. Not only were they wives, daughter, and mothers, they were also workers. They were dyers and weavers, spinners and crafts-people alongside men (Peskowitz 1997, Meyer 1988) and, in some instances, they were also communal leaders (Kraemer 1992, Brooten 1982). The work of feminist scholars has offered us new evidence of the lives of Jewish women in the past. In the fields of the social sciences, feminist scholars have called our attention to the role of Jewish women in Jewish communal and political life including their role in the process of Jewish assimilation. How did they dress themselves and their families or decorate their homes? What did they serve their families for dinner? These are all crucial questions for feminist ethnographers who are re-examining American Jewish assimilation (Prell 1999, Barack Fishman 1993, Davidman and Tenenbaum 1994).

The work in Jewish literature is massive and complex. Feminist scholars have not only discovered heretofore forgotten or unknown texts by Jewish women in various languages including Hebrew, Ladino, Yiddish, and English (Alcalay 1996), they have also offered new accounts of more well-known texts (Fuchs 1987, Seidman 1997). Scholars like Naomi Seidman have also asked us to reconsider the relationship between Jewish languages and how gender is used to naturalize the privileging of Hebrew over Yiddish. As she explains, like the relationship between husband and wife, Yiddish, the language of daily life was subservient to Hebrew, the masculine language of prayer and traditional Jewish study.

4. *Jewish Cultural Production*

Recent works of feminist art and literature have not only built on the work of Jewish scholars in a variety of fields, they have also used popular and material culture to offer critical perspectives on Jewishness as a cultural and ethnic as well as a religious identity. Some works highlight the absence of Jewish women from certain forms of religious or cultural production like the reading and writing of sacred texts in Helene Aylon's work (Kleeblatt 1996). Others have used these artifacts including texts to create new theatrical or cinematic productions on or about Jewish women.

Since the 1970s there has been a growing body of Jewish feminist literature especially by lesbian Jewish feminists (Alpert 1997, Beck 1982, Klepfisz and Kaye/Kantrowitz 1986). These writers often focus on Jewishness as a cultural and political and not a religious practice. More recent efforts have included the reclaiming of Klezmer music and the resources of Yiddish film for new cultural production (Fox 1999).

In the arts the inclusion of women has expanded the realm of Jewish cultural production in the United States as well as Israel. Increasingly Jewish women are playing an active role in these efforts by creating new works, sponsoring exhibits, and participating in various forums to present feminist artistic works to a broader audience (Kleeblatt 1996, Pellegrini 1997, Antler 1998, Prell 1999). Here Jewish museums have become a crucial site for new forms of Jewish feminist cultural production.

5. *Future Prospects*

Although Jewish women are, now more than ever, a part of Jewish communal life, there is much more work to be done. Efforts to make explicit the inequities at the heart of many Jewish communal practices much less the exclusion of women from sacred texts and the problematic roles they do have when they are included in these texts all demand ongoing critical examination. Among Jewish feminists the idea of giving up these traditions is too simple. Although they are critical of these texts, they continue to claim them and try to find ways of forging new and more powerful relationships with these legacies. Jewish women are doing this labor as university scholars, day school teachers and administrators, as cantors and rabbis, and as poets and artists.

See also: Accidents, Normal; Anti-Semitism; Buddhism and Gender; Catholicism and Gender; Feminist Theology; Feminist Theory; Historiography and Historical Thought: Jewish Tradition; Holocaust, The; Islam and Gender; Jewish Law; Judaism; Nazi Law; Protestantism and Gender; Religion and Gender; Religion, Sociology of; Women's Religiosity; Zionism, History of

Bibliography

Adler R 1998 *Engendering Judaism: An Inclusive Theology and Ethics*. Jewish Publication Society, Philadelphia, PA

Alcalay A (ed.) 1996 *Keys to the Garden: New Israeli Writing*. City Lights, San Francisco

Alpert R 1997 *Like Bread on the Seder Plate: Jewish Lesbians and the Transformation of Tradition*. Columbia University Press, New York

Antler J (ed.) 1998 *Talking Back: Images of Jewish Women in American Popular Culture*. Brandies University Press, Hanover, NH

Balka C, Rose A (eds.) 1989 *Twice Blessed: On Being Lesbian, Gay and Jewish*. Beacon Press, Boston

Barack Fishman S 1993 *A Breath of Life: Feminism in the American Jewish Community*. Free Press, New York

Baskin J (ed.) 1991 *Jewish Women in Historical Perspective*. Wayne State University Press, Detroit, MI

Beck E T 1982/1984/1989 *Nice Jewish Girls: A Lesbian Anthology*. Beacon Press, Boston
Berkowitz M 1993 *Zionist Culture and West European Jewry Before the First World War*. Cambridge University Press, Cambridge, UK
Biale D 1992 *Eros and the Jews: From Biblical Israel to Contemporary America*. Basic Books, New York
Biale R 1984 *Women and Jewish Law: An Exploration of Women's Issues in Halakhic Source*. Schocken, New York
Boyarin D 1993 *Carnal Israel: Reading Sex in Talmudic Culture*. University of California Press, Berkeley, CA
Boyarin D 1997 *Unheroic Conduct: The Rise of Heterosexuality and the Invention of the Jewish Man*. University of California Press, Berkeley, CA
Brettschneider M (ed.) 1996 *The Narrow Bridge: Jewish Views on Multiculturalism*. Rutgers University Press, New Brunswick
Brooten B J 1982 *Women Leaders in the Ancient Synagogue: Inscriptional Evidence and Background Issues*. Scholars Press, Chico, CA
Davidman L 1991 *Tradition in a Rootless World: Women Turn to Orthodox Judaism*. University of California Press, Berkeley, CA
Davidman L, Tenenbaum S (eds.) 1994 *Feminist Perspectives on Jewish Studies*. Yale University Press, New Haven, CT
El Or T 1994 Watzman H (trans.) *Educated and Ignorant: Ultraorthodox Jewish Women and their World*. Reinner, Boulder, CO
Falk M 1996 *The Book of Blessings: New Jewish Prayers for Daily Life, the Sabbath, and the New Moon Festival*. Harper Collins, San Francisco
Fox H N 1999 *Eleanor Antin*. Los Angeles County Museum of Art and Fellows of Contemporary Art, Los Angeles
Fuchs E 1987 *Israeli Mythogynies: Women in Contemporary Hebrew Fiction*. State University of New York Press, Albany, NY
Glenn S A 1990 *Daughters of the Shtetl: Life and Labor in the Immigrant Generation*. Cornell University Press, Ithaca, NY
Greenberg B 1981 *On Women and Judaism: A View from Tradition*. Jewish Publication Society, Philadelphia, PA
Grossman S, Haut R (eds.) 1992 *Daughters of the King: Women and the Synagogue: A Survey of History, Halakhah, and Contemporary Realities*. Jewish Publication Society, Philadelphia, PA
Hauptman J 1998 *Rereading the Rabbis: A Woman's Voice*. Westview Press, Boulder, CO
Heschel S (ed.) 1983/1995 *On Being a Jewish Feminist: A Reader*. Schocken, New York
Hyman P E 1995 *Gender and Assimilation in Modern Jewish History: The Roles and Representations of Women*. University of Washington Press, Seattle, WA
Hyman P E, Moore D D (eds.) 1997 *Jewish Women in America: An Historical Encyclopedia*. Routledge, New York, 2 Vols.
Kaplan M A 1979 *The Jewish Feminist Movement in Germany: The Campaign of the Jüdisher Frauenbund, 1904–1938*. Greenwood Press, Westport, CT
Kaplan M A 1991 *The Making of the Jewish Middle Class: Women, Family, and Identity in Imperial Germany*. Oxford University Press, New York
Kaufman D R 1991 *Rachel's Daughters: Newly Orthodox Jewish Women*. Rutgers University Press, New Brunswick
Kleeblatt N L 1996 *Too Jewish? Challenging Traditional Identities*. Rutgers University Press, New Brunswick
Klepfisz I, Kaye/Kantrowitz M (eds.) 1986/1989 *The Tribe of Dina: A Jewish Women's Anthology*. Beacon Press, Boston
Kraemer R S 1992 *Her Share of the Blessings: Women's Religions Among Pagans, Jews, and Christians in the Greco-Roman World*. Oxford University Press, New York
Kuzmack L G 1990 *Woman's Cause: The Jewish Woman's Movement in England and the United States 1881–1933*. Ohio State University Press, Columbus, OH
Levitt L 1997 *Jews and Feminism: The Ambivalent Search for Home*. Routledge, New York
Meyer C L 1988 *Discovering Eve: Ancient Israelite Women in Context*. Oxford University Press, New York
Nadel P S 1998 *Women who would be Rabbis: A History of Women's Ordination, 1889–1985*. Beacon, Boston
Ochs V L 1990 *Words on Fire: One Woman's Journey into the Sacred*. Harcourt Brace Jovanovich, New York
Ofer D, Weitzman L J (eds.) 1998 *Women in the Holocaust*. Yale University Press, New Haven, CT
Pardes I 1992 *Countertraditions in the Bible: A Feminist Approach*. Harvard University Press, Cambridge, MA
Pellegrini A 1997 *Performance Anxieties: Staging Psychoanalysis, Staging Race*. Routledge, New York
Peskowitz M 1997 *Spinning Fantasies: Rabbis, Gender, and History*. University of California Press, Berkeley, CA
Peskowitz M, Levitt L (eds.) 1997 *Judaism Since Gender*. Routledge, New York
Plaskow J 1990 *Standing Again at Sinai: Judaism from a Feminist Perspective*. Harper Collins, San Francisco
Pogrebin L C 1991 *Deborah, Golda and Me: Being Female and Jewish in America*. Crown Publishers, New York
Prell R-E 1999 *Fighting to Become Americans: Jews, Gender and the Anxiety of Assimilation*. Beacon, Boston
Rittner C, Roth J K (eds.) 1993 *Different Voices: Women and the Holocaust*. Paragon House, New York
Rudavsky T M 1995 *Gender and Judaism: The Transformation of Tradition*. NYU Press, New York
Scott J 1993 Women's history. In: Kauffman L S (ed.) *American Feminist Thought at Century's End: A Reader*. Blackwell, Cambridge, UK pp. 234–57
Seidman N 1997 *A Marriage Made in Heaven: The Sexual Politics of Hebrew and Yiddish*. University of California Press, Berkeley, CA
Stahl Weinberg S 1988 *The World of Our Mothers: The Lives of Immigrant Women*. Schocken, New York
Umansky E M, Ashton D (eds.) 1992 *Four Centuries of Jewish Women's Spirituality: A Sourcebook*. Beacon, Boston
Wegner J R 1988 *Chattel or Person? The Status of Women in the Mishnah*. Oxford University Press, New York
Weidman Schneider S (ed.) 1984/1985 *Jewish and Female: A Guide and Sourcebook for Today's Jewish Woman*. Touchstone, New York
Weissler C 1998 *Voices of the Matriarchs: Listening to the Prayers of Early Modern Jewish Women*. Beacon, Boston

L. Levitt

Judges

A judge is a public officer appointed to administer the law in a Court of Justice impartially and independently. Judicial independence is instrumental to judicial impartiality, since only an independent judge can be expected to be fully impartial: hence 'his and it

usually is his' configuration as an officer within some governmental structure, a measure originally meant to protect him against pressures from the confronting parties (Shapiro 1981). However, while this official status of the judge may provide protection against the parties it may also facilitate pressures and influences from those to whom he owes his office. In some highly infrequent cases (such as early Roman Law) the judge is freely elected by the parties who commit themselves to abide by his decisions. However in most historical instances the judge's authority and binding capacity is independent from the will of the involved parties and derives from the backing granted to his actions by the political system. This creates a direct link between the legitimacy of judicial institutions and the type of legitimacy of the prevailing form of political government. Judicial impartiality can thus be expected to be reasonably guaranteed only under strict democratic rule, one of the basic requisites of a democratic polity being the existence of an effectively independent judiciary.

Officially empowered to determine in a binding way with which of the contending parties lies the legal right, the judge constitutes a key, central figure in all legal systems, past or present. However, the judge is but one among other possible conflict-solving institutions. Rather than as a discrete, fully differentiated category, he can be better understood as one of the extremes of a continuum that would run: *go-between*, *mediator*, *arbitrator*, *judge* (Shapiro 1981, p. 3). All these roles require an impartial implementation but differ in the degree of formalization and depersonalization presiding the relation with the contending parties. The *go-between*, who operates on a pure consent basis, represents the more informal, face-to-face, and equity-oriented end of the continuum; the *judge*, professionally, trained, vested with official authority and required to lack all personal acquaintance with the parties, occupies the opposite end.

As societies become more complex and differentiated, and as the state-controlled legal system increases its centrality in social life, the role of the judge gains predominance. However, the other conflict-solving roles do not disappear; instead, they subsist as alternative—and usually very active—instances. The continuum can be thus perceived as a cumulative process, allowing for the emergence of progressively more complex and formalized agencies of conflict resolution that co-exist with the preceding ones (see *Dispute Resolution in Economics*).

1. Types of Judges

A basic typology of judges can be attempted from the combination of three factors intervening decisively in the shaping of judicial roles. In the first place, the *type of legal system* within which the judge operates: a basic distinction can be made between Western legal systems (comprising both Common Law and Civil Law systems), and all other types of systems, such as *Far Eastern* (Chinese and Japanese), *Religious* (Islamic or Hindu), or *Socialist* (see Zweigert and Kötz 1998). In the second place, the type of political legitimacy behind the body of laws to be applied, and consequently behind the political system which upholds those laws (Weber 1947). Third, the basic function that is attributed to the judge: a basic distinction can be made between authoritative adjudication and mediating arbitration. The following classificatory grid (Table 1) results:

Three basic configurations of the judicial role can thus be established:

(a) *Democratic judge*. A type of judge who adjudicates cases applying rules of general validity, enacted by duly elected representatives of the citizenry. This type of judicial system corresponds to representative democracies (legal-rational type of domination in Weberian terms). There are two possible formats for the *democratic* type of judge: Common or Civil Law.

(b) *Traditional or predemocratic judge*. A predominantly arbitrating, conciliatory type of Judge who applies norms embodying a traditional legacy, contained in religious, moral, or ethical codes, or in local customs, *mores*, or folkways. This type of judge would correspond to the Weberian category of 'traditional' political domination. Several classical types of judicial roles could be included within this category: the Islamic type of judge (who attempts conciliation—if and when the compromise does not permit what is forbidden or does not forbid what is permitted by the religious law—or who adjudicates on a case-by-case basis according to equity); the traditional Far Eastern Judge (Chinese or Japanese), more concerned with mending and preserving social harmony and equilibrium than with legal arguments or questions of individual rights; and the African or Native-American type of customary judge.

(c) *Nondemocratic judge*. A type of judge corresponding to autocratic political situations where the norms to be applied may be claimed to express the will of the people, *but as interpreted and elaborated by an autocratic ruler* (or ruling body), self-appointed as exclusive, natural spokesman for the collective will, and as such above any need for elective confirmation. This third situation would be the equivalent to Weber's concept of *charismatic* rule.

2. Increasing Convergence

This threefold typology provides basic global models or *ideal-types* but is not meant as a description of real contemporary judicial systems. The judicial systems of the more economically developed and more democratically stable countries would generally fit well into the first ('*democratic judge*') category. Most other countries, striving in many cases to consolidate modernized and democratic polities, and as such immersed in

Table 1
Judges: a basic typology

			Prescribed Basic Judicial Function		
			Adjudication (Judge imposes a solution on the parties)		Arbitration/ Conciliation (Judge assists the parties in finding an arrangement)
			Applying rules with universal validity	On a case-by-case basis	
Type of Legitimacy of the Body of Laws to be applied (The laws are considered legitimate because they embody…)	(a) The Will of the People, as expressed by duly elected representatives		Civil Law Judge; Common Law Judge		
	(b) Tradition, i.e., a traditional legacy contained in religious or moral codes or in local customs	Religious norms		Islamic Judge, *Cadi Justice*	
		Ethical norms			Far Eastern Judge (Chinese, Japanese)
		Customary norms			Customary Judge (African)
	(c) Mandates of a Ruler or Ruling Body, of an autocratic ('charismatic') nature who claims to express the true will of the people		'Socialist' and 'Totalitarian' type of Judge		

transitional situations, represent mixed, intermediate stages. However, the generalization can be made that as processes of economic, social, and cultural globalization take place at a world scale, and as political democratization and the rule of law tend to become increasingly undisputed, universal goals, the tendency can be detected in most developing countries towards the adoption of judicial systems parallel to the two models of Justice prevailing in stable democracies (predominantly, the Civil Law type of judicial system). In quite a few cases this has implied some degree of *denaturalization* of the imported model, since the modern Civil Law (as well as the Common Law) type of Judge is designed to function under democratic conditions. Certainly this was not always the case: both systems (Civil and Common law) had emerged and consolidated well before the modern idea of the rule of law had even begun to develop. But their current configuration results from the substantial redefinitions they both experienced in the wake of the democratic revolutions of the eighteenth century.

3. *Judicial Systems in Developing Countries*

Purely customary judicial systems can be now said to have become a rarity of merely anthropological or historical interest. By 1990, all Francophone African countries, with the only exception of Cameroon, had suppressed all types of customary courts, establishing a State system of Courts of the Civil Law type (Mangin 1990). In Anglophone Africa, customary courts are still allowed to exist in most countries, but as marginal institutions, their decisions being in any case appealable to State Courts, in most cases modeled along Common Law lines (Coldham 1990). One of the most conspicuous instances of Far Eastern traditional type of Judge, the Japanese, is now extinct and has given way to a Civil Law type system of Justice, with some elements of Common Law (Oda 1992, quoted in Zweigert and Kötz 1998).

Apart from China, one of the few remnants of the 'socialist' category of judicial system, most traditional systems are experiencing assimilation to the Western models. In India, where a process of *hybridization* gave birth to 'Anglo-Muhammadan' and 'Anglo-Hindu' law (Zweigert and Kötz 1998, pp. 310, 318), these bodies of law are administered by State judges in Common Law style.

In the case of the Islamic world a distinction has to be made between: (a) countries (such as Iran or Afghanistan) which in the wake of Islamic fundamentalist revolutions reorganized as Islamic Republics

and which would still fit into the traditional/religious category of judicial systems; and (b) countries (such as Egypt, Tunisia, or Jordan) which though experiencing in different degrees a trend towards the re-Islamization of social, legal, and political institutions still maintain a secular State-courts system created in most cases in the 1950s and 1960s, along the Civil Law model (Botiveau 1994).

4. *Civil and Common Law Judges*

The two types of judicial systems increasingly predominant in the contemporary world (Civil and Common Law) share a basic common, though distant, root: Roman Law, Justinianean and Classic, respectively (van Caenegem 1988). Their historical development led to differing legal and judicial traditions: Common Law is 'judge-made law, that is to say (consisting of) rules to be collected from the judgments of the Courts' (Dicey 1905/1963), whereas codification stands out as the basic identifying feature of Civil Law (Merryman 1969). But they both embody what seems to be the final stage of the centuries-long process of evolution of the Western legal tradition. As pointed by Berman (1983) the emergence of Canon Law (which gave rise to the first full legal system), the Protestant Revolution (which secularized the legal world) and the English Revolution (which established the supremacy of Parliament) paved the way for the two versions of democratic judicial systems predominant today and brought about by the American (1776) and French (1789) Revolutions. As democratically reborn judicial systems, both the Civil Law (as reshaped by the French Revolution) and Common Law (as amended in American soil) share a similar contextual prerequisite for their adequate operation: the existence of the rule of law within a pluralistic, representative democracy.

However, both systems diverge in the way in which they attempt to keep the organization of Justice in line with the commonly shared dogma of popular sovereignty. The variant of the Civil Law originated in the French Revolution was basically intended to prevent judges from interfering with the popular will as expressed in the statutes enacted by duly elected representative bodies—a reaction against the preeminence of *Parlements* under the *Ancien Régime*. In the new, democratic situation the judge had to be in Montesquieu's words just 'the mouth of the law.' Only from an unswerving application of democratic laws would the judge acquire a democratic character. In contrast, the main purpose of the judicial system established by the 1787 American Constitution was to make of the judge an essential element of the system of checks and balances that the new democratic polity was intended to be. To this effect, the judge had to apply the new statute law which had been democratically enacted (and which should be given priority in cases of collision with judge-made law); at the same time he was empowered to insure that statutes passed by ephemeral, temporary majorities would not go against the basic covenant contained in the 'supreme law of the land' (i.e., the Constitution). (See *Judicial Review in Law*.)

As a consequence, two clearly differentiated models of 'democratic Justice' crystallized. On the one hand, the Common Law judge, as watchman of the Law, that is, of the 'patchwork quilt of legislative provisions held together by thread woven out of general principles and sewn by the judges' (Bell et al. 1998, p. 5), performs his functions with a basic policy-implementing orientation (Damaska 1986) predominantly within an adversarial system. On the other hand, the Civil Law judge, as a specialized anonymous civil servant decides the cases brought to him by applying a body of written legal norms, which is assumed to provide a complete and coherent repertoire of solutions; his functional orientation is inquisitorial and conflict solving. Intended to be a system of Justice rigidly respectful with the popular will, the Civil Law system turned out in practice to be a machinery of Justice Administration insulated from politics. Although this sounds favorable, what it means is that the Civil law model of Justice, devised for a democratic setting, proved equally functional when the overall political system ceased to be democratic. The system had been designed to provide a dutiful application of the laws in force, whatever their nature, not to scrutinize their legitimacy or interpret their contents. A change from democracy to autocracy meant no difference for the internal logic of the system, which could continue operating practically unchanged. The cases of Italy, Germany, and Spain in the twentieth century, to quote but some particularly conspicuous cases where breakdown of democracy, autocratic rule, and return to democracy did not entail any proportionally correlative reorganization of the judicial system (and even just marginal changes in the composition of the judiciary) provide a telling example (Toharia 1975). This unintended political floating ability and adaptability of the Civil Law model explains the successive attempts to amend its original late-eighteenth-century design in order to reinforce its original democratic structure. The increasing importance given to *jurisprudence* (or binding judicial interpretation of norms), a word which Robespierre wanted to see removed from the French vocabulary; the gradual introduction of varying forms of judicial review (in most cases in the form of Constitutional Courts) largely under Hans Kelsen's influence in the 1920s and 1930s; the increasing questioning of the inquisitorial principle (which in the early twenty-first century is still in force, even though it is likely to disappear soon in three European Civil Law systems—France, Spain, and Portugal); and new forms of judicial activism (Guarnieri and Pederzoli 1996) are clear examples of the reorganization that the original structure has been experiencing. In a sense,

these changes reflect an increasing pattern of convergence between Civil and Common Law models (Cappelletti 1989), a convergence which goes both ways: the impact that the legal system of the European Union is having on the legal and judicial practices of the United Kingdom has also meant substantial and gradual redefinitions of some basic Common Law features (White 1999).

5. *The Selection of Judges*

A key element in the configuration of any judicial system is the way in which judges are selected. An ideal system of judicial recruitment should allow for: (a) the selection of competent and qualified persons; (b) thorough mechanisms unlikely to impair their impartiality and independence while in office; (c) making room for mechanisms of discipline, evaluation, and accountability of the judicial personnel; and (d) aiming to create a judiciary as representative as possible of the social body (in terms of ideology, social class, gender, race, religion, or other relevant social attributes). These four requirements are obviously difficult to meet and most existing systems of judicial recruitment attempt to achieve them through a wide variety of arrangements, which probably explains why 'no other governmental official is selected by so many different methods' as the judge (Stumpf 1998, p. 135).

A basic, preliminary distinction can be made between *professional* and *bureaucratic* systems of selection (Guarnieri and Pederzoli 1996), corresponding respectively to Common Law and Civil Law judicial systems. The *professionalized* selection looks for judges among already experienced, practicing lawyers which produces an older body of judges who have relatively short tenure as judges. (On mechanisms of selection see *Judicial Selection*.) By contrast, the *bureaucratic* system of selection prevailing in Civil Law systems allow for much longer judicial careers: from an average entrance age barely around or well below 30, to compulsory retirement around age 70 or even 75. The system is not meant to select mature, experienced lawyers and make them judges but rather to select young law graduates, without any previous legal practice, and prepare them for a lifelong career in the judiciary. The sense of achievement within the legal profession and the social aura and moral *auctoritas* that characteristically go with the image of the Common Law judge, is lacking in his Civil Law counterpart.

6. *Judges' Social Representativeness*

As professionals called to perform a highly specialized task implying extensive legal training, judges have tended historically to be recruited among members of educated, well-to-do sectors of society. This has originated an almost universal double structural bias within all judiciaries: a *class bias* (middle- to lower-class people, as well as people from minority groups, were likely to be strongly under-represented) and a *gender bias* (women were particularly likely to be strongly under-represented, as a result of the late recognition of their rights to equal education and equal access to professional life). A third, additional bias might even be said to have traditionally characterized the Judiciary: a *conservative bias* (as suggested originally by Max Weber, who pointed to the inherent conservative implications of legal reasoning and thence its more likely appeal for conservative-minded persons). As a result women, members of minorities, persons of lower-middle- or lower-class origin and people with ideologically progressive orientations were only rarely found within judiciaries.

To a large extent, all these biases seem to be gradually fading, as educational systems have become more inclusive and affirmative action programs have promoted higher levels of social equality. As for the social origin of their members, Civil Law judiciaries have been traditionally less elitist than their Common Law counterparts: in countries such as France, Italy, or Spain most judges tended to come from rural and middle- to low-middle-class families whereas in the US or in the United Kingdom urban upper- or upper-middle-class origin usually prevailed. But in both cases the presence of judges with lower- or working-class origin was practically nil. This situation is changing substantially in both systems as judiciaries gradually become more representative of the prevailing social structure (Toharia 1987). The data in Table 2 reflect the gradual increase of social pluralism within most judiciaries since the 1970s. The change is particularly perceptible with respect to the increased presence of women, especially in civil law systems. The universalistic, competitive nature of the *bureaucratic* process of selection (generally based purely on personal performance in public entrance examinations, which leaves less room for blatant discrimination), and the fact that it takes place at a much earlier point in life seem to account for this seemingly more advanced degree of *feminization* of most Civil Law judiciaries. The presence of women within the judiciary may be said in any case to have passed well beyond the 'tokenization' phase, and to be becoming a fully institutionalized pattern. Even in Arab countries such as Tunisia, in spite of the controversy reopened by the religious fundamentalists' questioning of the eligibility of women for judicial office, in 1994 one out of every four judges was a woman (Botiveau 1994). Racial and religious minority groups appear to have also increased their presence within the judiciary, though to a somewhat lower degree. As for the *ideological bias* the issue seems to have lost much of its former intensity. In Common Law countries, appointment or direct election of judges on the basis of their ideological

Table 2
How socially representative is the Judiciary?

Women and other social groups within the Judiciary. Selected dates and countries		
Date	Country	
1835	USA	First Roman Catholic appointed to the Supreme Court (Roger B. Taney)
1916	USA	First Jewish-American appointed to the Supreme Court (Louis Brandeis)
1961	USA	African-American judges represent 0.8 percent of all State and Federal Judges
1962	UK	First woman appointed as County Judge (Elizabeth Lane)
1965	UK	First woman appointed to the High Court Bench (Elizabeth Lane)
1967	USA	First African-American appointed to the Supreme Court (Thurgood Marshall)
1967	Tunisia	First woman admitted into the Judiciary
1977–80	USA	President Carter's positive action program results in the appointment of 100 women, African-Americans, and Hispanic-Americans,to the Federal Bench
1979	UK	England's first Sikh judge appointed
1979	California, USA	Governor Brown appoints America's first openly homosexual judge
1980–1	USA	3.1 percent of all US State Supreme Court Justices are women
1981	USA	First woman appointed to the Supreme Court (Sandra Day O'Connor)
1981	California, USA	Governor Brown appoints first openly lesbian judge
1986	USA	First Italian-American appointed to the Supreme Court (Antonin Scalia)
1989	Germany	18 percent of all judges are women (compared with 3 percent in 1960)
1990	The Netherlands	28 percent of all judges are women (compared with 11 percent in 1973)
1992	Georgia, USA	First African-American woman appointed to the Supreme Court of a former Confederate State
1992	Costa Rica	42 percent of all judges are women
1992	Buenos Aires, Argentina	26 percent of all judges are women (compared with 11 percent in 1985)
1993	USA	10 percent of all federal judges are African-Americans
1993	USA	3.3 percent of judges in all levels of the US Judiciary are African-Americans
1994	Tunisia	23 percent of all judges are women
1994	Venezuela	First woman ever appointed as Chief Justice of a nation's Supreme Court (Ms Cecilia Sosa)
1994	UK	Percentage of women in the different levels of British Judiciary: Heads of Division: none;Lords Justices of Appeal: one out of 32; High Court Judges: 6 percent; Circuit Judges:6 percent; District Judges: 10 percent
1995	Italy	41 percent of all judges are women
1996	France	47 percent of all judges are women
1996	USA	Of all Judges within the Federal Judiciary, 9 percent are African-American, 4 percent Hispanic-American, 0.6 percent Asian-American and 0.2 percent Native-American. These minority groups add up to 13.8 percent, compared with 4.9 percent in 1976
1996	USA	Within the Federal Judiciary, 17 percent are women (compared with 1 percent in 1976 and 9 percent in 1989). During President Clinton's first administration (1992–6) 31 percent of all appointments to the Federal bench corresponded to women (as compared with 0.4 percent under Nixon, 16 percent under Carter, 8 percent under Reagan and 19 percent under Bush), and 33 percent to members of minority groups
1999	Spain	34 percent of all judges are women, compared with 12 percent in 1987. Among the last class of new judges selected, 58 percent are women

orientation is perceived as a legitimate way of creating a bench that 'looks like the country.' In Civil Law countries, where judicial systems are supposed to be person proof, the scant data available indicate nonetheless that the ideological composition of the judiciary tends to reflect quite closely the ideological pluralism prevailing in the global society (Toharia 1987). The pattern of affiliation to the ideologically

oriented unions of judges existing in countries such as Italy, France, or Spain provides an additional evidence of such pluralism.

See also: Constitutional Courts; Courtroom Consultation: Psychology; Courts and Adjudication; Judicial Selection; Juries; Lawyers; Legal Education; Legal Professionalism; Supreme Courts

Bibliography

Bell J, Boyron S, Whittaker S 1998 *Principles of French Law*. Oxford University Press, Oxford, UK

Berman H J 1983 *Law and Revolution. The Formation of the Western Legal Tradition*. Harvard University Press, Cambridge, MA

Botiveau B 1994 La loi islamique et jugement moderne. *Droit et Cultures* **28**: 25–45

van Caenegem R C 1988 *The Birth of the English Common Law*, 2nd edn. Cambridge University Press, Cambridge, UK

Cappelletti M 1989 *The Judicial Process in Comparative Perspective*. Clarendon Press, Oxford, UK

Coldham S 1990 Les systèmes judiciaires en Afrique anglophone. *Afrique Contemporaine* **156**: 27–37

Damaska M R 1986 *The Faces of Justice and State Authority*. Yale University Press, New Haven, CT

Dicey A V 1905/1963 *Law and Public Opinion in England*. Macmillan, London

Guarnieri C, Pederzoli P 1996 *La Puissance de Juger*. Editions Michalon, Paris

Mangin G 1990 Quelques points de repère dans l'histoire de la Justice en Afrique francophone. *Afrique Contemporaine* **156**: 21–6

Merryman J H 1969 *The Civil Law Tradition*. Stanford University Press, Stanford, CA

Pannick D 1987 *Judges*. Oxford University Press, Oxford, UK

Shapiro M 1981 *Courts. A Comparative and Political Analysis*. The University of Chicago Press, Chicago

Stumpf H P 1998 *American Judicial Politics*, 2nd edn. Prentice-Hall, Upper Saddle River, NJ

Toharia J J 1975 *El Juez Español. Un Análisis Sociológico*. Tecnos, Madrid, Spain

Toharia J J 1987 *Pleitos Tengas. Introducción a la Cultura Legal Española*, 1st edn. CIS/Siglo Veintiuno de Espana Editores, Madrid, Spain

Weber M 1947(1925) *The Theory of Social and Economic Organizations*. Oxford University Press, New York

White R C A 1999 *The English Legal System in Action. The Administration of Justice*. Oxford University Press

Zweigert K, Kötz H 1998 *An Introduction to Comparative Law*. Clarendon Press, Oxford, UK

J. J. Toharia

Judicial Review in Law

Judicial review may be defined as the institutional capacity of courts of law to assess the validity of actions taken by both inferior and coordinate branches of government. In the United States, judicial review has been associated historically with Supreme Court decisions regarding the allocation of powers among the branches of the national government, and between the national and state governments. In the last half of the twentieth century, the federal judiciary also used the power of judicial review to play a prominent role in protecting individual rights. Most modern nation-states now incorporate some variant of this power into their laws and constitutions.

1. Historical Origins

Judicial review is an outgrowth of efforts to limit the scope of power held by those authorized to interpret both sacred and secular texts. Societies, particularly (but not exclusively) those with complex governing structures, have created systems of law that constrain or influence official and social behavior through expressed codes or instructions. These, in turn, are overseen by special, trained officials. The fear that written laws and those empowered to enforce them will subvert the basic structures and values of a society form a prominent chapter in the history of self-government. Enabling government while protecting against its abuses is both an age-old and a universal concern.

In the Judeo-Christian tradition, for example, one finds such concerned in the books of Exodus, Leviticus, and Kings I and II. There, following the example of the primary law-giver, Moses, the prophets often contested the interpretations of both sacred and civil law provided by the kings and other secular rulers. Again, intellectuals in ancient China engaged in an extensive debate as to whether *li* (ritual) or *fa* (law) should govern human behavior. Mo Tsu, one such intellectual, argued that government was necessary to establish and govern meanings, including those pertaining to human conduct. Other intellectuals countered that the attempt to codify essential rituals or forms of conduct would lead to sophistry in political life. The Prince's advisors, it was reasoned, would confuse the Prince with various interpretations of essential political terms (Pocock 1971).

2. Tension with Democracy

In modern democratic polities, the dilemma presented by judicial review is often referred to as the 'counter-majoritarian difficulty.' This is a recognition of the inherent tension between democracy and the rule of law. Courts and the law would seem to be ideal guardians of a political system's fundamental principles and values. Yet, it is also recognized that, structurally, courts are not majoritarian democratic institutions, and their decisions and interpretations of laws and constitutions may be at variance with popular values. There is always the fear that, in the guise of

interpretation, a judge will substitute his or her personal values for those of a democratic majority.

The emergence and practice of judicial review in the United States is a textbook example of this continuing tension. Although the American Constitution does not mention judicial review, it was not unknown to the framers. In *Marbury v. Madison* (1803), the Supreme Court held that 'it emphatically is the province and duty of the judicial department to say what the law is,' and that, under a written constitution of enumerated powers, the Court could not enforce a law repugnant to it. An act of Congress that contravened the Constitution, the Court held, was null and void.

Only a few laws were held invalid in the nineteenth century, but the scope and incidence of judicial review expanded rapidly in the twentieth century. It is today a centerpiece of a 'judicialized' American political system. There is general agreement at the theoretical level about the utility of judicial review, but also constant disagreement about its application in particular cases and the alleged perils of an 'imperial judiciary.'

3. *Globalization of Judicial Review*

As Americans continue to debate the use and consequences of judicial review, it has taken root elsewhere. Most postwar constitutions in Europe, as well as India and Japan, incorporated some variant of it. And it now has been adopted, at least formally, by many of the nations of the former Soviet empire. Many of these nations had, in fact, once deliberately rejected it. Post-Napoleonic Europe, for example, firmly placed the judiciary under the control of the positivistic codes and enactments of the parliaments. Asian countries struggled throughout the nineteenth and early twentieth centuries to develop independent judiciaries that would be 'acceptable' to the United States and the other Western powers (Grovogui 1996, Millard 1931).

The rise of authoritarian and totalitarian regimes with little regard for basic human rights in the twentieth century led to an 'Americanization' of the governing regimes in Turkey, the Philippines, Japan, France, and Germany after the Second World War. Each of these nations has joined many others in adopting some measure of constitutional and liberal reform in their political arrangements, reforms that include an expressed power of judicial review. In Africa, judicial review and 'constitutional courts' have taken root in places like Benin, South Africa, Ghana, and many other countries where, previously, the judiciary was almost completely subservient to the executive or legislature (Shapiro and Stone 1994, Prempeh 1999).

But even in Great Britain, which has no written constitution, no bill of rights, a strong tradition of parliamentary supremacy, and no formal mechanism for judicial review, in the same period that the US was undergoing a transition toward a more 'activist' judiciary with respect to civil rights, British judges increasingly policed the administrative realm for violations of governing statutes. Some scholars have looked to internal factors, such as intergovernmental support and rights-advocacy organizations, as well as the influence of the increasingly rights-oriented and litigious culture of the United States, to explain the development and persistence of what might be called clandestine review (Shapiro and Stone 1994, Epp 1998).

The British experience is but one indicator that, while the United States has consciously sought to export its 'unique' institution, the results have not been uniform. Courts in Europe, Africa, and Asia continue to retain much of their traditional character even if consciously aware of the American example. This is particularly true of countries with civil code rather than common law systems. European constitutional courts, for example, are usually special courts that hear cases brought by the government, not private litigants. In France, the Constitutional Council may only review legislation in the immediate time period after passage and prior to promulgation. After that time, the statutes in question cannot be reviewed by that body. Moreover, the fact that members of both the Assembly and Senate can, with great ease, bring a constitutional challenge to legislation has further, blurred the edges separating legislating from adjudicating. The legislator proposes and debates a law 'as if' a judge will hear the same; and judges write decisions 'as if' the audience will be a legislator. Whether the parliament is the creature of the judge or the judge is the creature of the parliament is, in all respects, an important question (Shapiro and Stone 1994, pp. 400–2).

Characteristics such as those retained by the French point to important differences from American cultural and jurisprudential conventions. The US Supreme Court has never viewed itself as a court of 'instruction.' It is not empowered to issue merely advisory opinions; judicial review of constitutionality can only be undertaken in the context of an adversary 'case or controversy' proceeding. Organizing principles of government such as federalism and the separation of powers also vary widely in their acceptance, meaning, and application, insuring many variants of judicial review.

4. *Judicialization of Politics?*

The fact that judicial review has become a global phenomenon has led some social scientists to speak of a 'judicialization' or 'juridicalization' of politics. At its root, the thesis suggests that political institutions such as legislatures and administrative agencies have come to expect and even welcome courts as arbiters or overseers of their actions. They tailor their actions accordingly to accord with an expected response by the

judiciary (Stone Sweet 1999, p. 164, Shapiro 1994). The electorate itself has come to have similar expectations. In the American political context, for example, judicialization has had the paradoxical effect of legitimizing, *de facto*, the role of judicial review in a democratic process. While most Americans lack a deep knowledge of American history, politics, or law, they are well able to speak the language of rights and can, with some measure of accuracy, defend or dispute the validity of the legal holdings involved in such famous cases as *Brown v. Board of Education* (1954), *Miranda v. Arizona* (1966), and *Roe v. Wade* (1973) (Friedman 1990. Grossman and Epp 1991). At the same time, the idea that the US Supreme Court is both the final arbiter of legal disputes, and should be so, is not seen by a significant portion of the electorate as contrary to basic democratic values. Judges, from this view, are the lesser of many evils in a world where state and national institutions have come to be mistrusted. What citizens oppose most is decisions with which they disagree!

Outside of the United States, the tension between judicial review and democracy has been less of an issue. This is due, without doubt, to the association of judicial review with postauthoritarian reforms and constitutions that establish the primacy of fundamental human rights principles, and entrust their protection to constitutional courts. The rise of supranational courts such as the European Court of Justice has not only created a second level of review that transcends the courts and law of individual countries, but has also added considerable force to the argument that the enforceability of basic rights is critical to the maintenance of democracy and a civilized world order.

When judicial review transcends national boundaries, the problems of legitimacy and acceptance may be even greater. As the legal philosopher H. L. A. Hart pointed out, judicial review (or some variant of it) is likely to be a problem in any political community or arrangement in which those who are authorized to make or interpret the laws are institutionally (and even more so, culturally) distinct from the members of that community (Hart 1961). Governing structures and particular social and political conflicts are continually readjusted to meet the shifting strengths and needs of the community. This is particularly true where, as with the European Court or the new constitutional courts of eastern Europe, the role of courts is still evolving and not yet fully understood or accepted. But it is also true, if to a lesser extent, in nations or communities where judicial review and an activist judicial role is endemic and expected.

5. *Contemporary Debate*

The wide-ranging debates that pervaded the American academy from roughly 1940 to 1990 tapered off as the US Supreme Court drifted rightward, and as the *de facto* presence of judicial review in American political life came to be more or less taken for granted. While the American literature on judicial review might seem to be of immediate and obvious interest to legal reformers in Europe, Asia, and Africa, it remains to be seen whether the American experience (or the 'counter-majoritarian difficulty' for that matter) will be of deep interest in those places. Rather, as judicial review has become less a species of *American exceptionalism* and a broader, global issue, current scholarship has turned increasingly to a hybrid of the American debate. At issue is the empirical question of *why* judicial review develops in some times and places and whether, normatively speaking, judicial review may not be better employed in some systems than in others.

Some scholars have argued that for judicial review to take root there must be some level of commitment to the rule of law, and to government that is both democratic and limited. In polities where pure majoritarianism prevails there is, of course, no necessary limit to what the legislature can do in the electorate's name. Where the government is limited by some overarching set of principles, the ability of courts of law to intervene and veto legislation in apparent conflict with such principles is more pronounced (Shapiro 1998). In a twist on the American experience, then, the inversion of the relationship of rights-to-the-state in regimes throughout Europe after the Second World War has enshrined their constitutional courts as defenders of pre-existing and constitutive rights held by the citizenry. Many of these rights find their sources not merely in written constitutions (such as the German Basic Law's protections of 'Human Dignity,' which many nations lack) but in sources both internal or external to the political system (e.g. treaties and conventions) (Shapiro and Stone 1994). Conversely, in some African countries like Ghana, the long-standing tradition of judicial deference to executive authority has led the judges to use their new powers to adopt a jurisprudence seemingly at odds with the more recent, liberal written constitutions. In these cases, commentators have suggested, new 'constitutional courts' with judges learned in foreign jurisprudence and committed to constitutional reform will have to be substituted for a judiciary that is, in large part, a holdover from the *ancien régime* (Prempeh 1999).

Other oft-cited factors include the existence of commerce and internal geographic and cultural divisions, such as federalism. Commercial actors conventionally use legal devices to secure both formal agreements and sanctions should the terms of those agreements be violated. Written law, and the search for a neutral third party who is an expert in such law, has long been an important part of trade. Social scientists have argued that commerce was of great importance in accounting for the historical emergence of judge-made law. The origins of the professional lawyer, Max Weber argued, are to be found in the *cavere*: the Roman practice of drafting contracts, wills,

and other legal devices. The lawyers, seeking to make commercial relations more predictable for their clients, would change the meanings of extant legal terms or invent new legal devices and then seek out judges who would institutionalize the changes as authoritative (Weber 1954, p. 72).

A recent study of the World Trade Organization (WTO) lends support to this observation. The WTO's predecessor, the General Agreement on Tariffs and Trade, consisted originally of a Panel System of diplomats who mediated issues between states. With time, the treaty came to be viewed as authoritative law and the Panel's decisions regarding that law were seen as both binding and precedential. Accordingly, the argument was made that legal specialists were needed who might then study the treaty and relevant Panel decisions to further enforce its terms. Beginning in the early 1970s, the Nixon administration pushed to place trade disputes into the hands of lawyers familiar with trade laws. And in fact it was the American bar that, in the 1970s and 1980s, spearheaded a shift toward the more legalistic model that is at the heart of the WTO today (Stone Sweet 1999, pp. 164–71).

The existence of federative or federal government is also cited as a factor. Divisions of power among disparate electoral, legislative, and executive bodies, and vertical divisions of power, invariably occasion conflict that is itself caused by, and therefore cannot be corrected easily by, 'normal' political procedures. As the judge is often regarded as a sort of theorist of the state and its basic arrangements, her appeal is in clarifying or mediating the existing division of power among such entities (Dworkin 1988).

Supranational courts, such as the European Court of Justice, have added yet another layer of review. That Court, for example, has claimed a sweeping power of judicial review for itself by converting the various treaties it oversees into a sort of overarching constitution binding on both member nations and the courts and citizens of those nations. Such a practice undermines the close connection between judicial review and national sovereignty; and it may also contravene basic principles of international law, at least in theory. Yet the European Court has been successful in part because it provides stability and some manner of coherence, and balances national sovereignty interests with the goal of integration.

Scholars also cite more temporally specific factors. In the United States, commercial elites dominated constitutional litigation for most of US history. Yet in the post-Second World War era, nontraditional litigants were able to pool time, monetary resources, and available legal talent to press for important constitutional guarantees (Epp 1998). And the Supreme Court, over time, 'adjusted' the Constitution to facilitate the development and legitimacy of the American welfare state. In Europe, somewhat differently, the devolution of legislative power in parliamentary systems and the concomitant rise in cabinet/party-enforced government, has occasioned a role for courts, particularly in those countries (e.g. Germany) where even a governing majority cannot prevent legislation from being referred to a constitutional court (Shapiro and Stone 1994, p. 405).

The French social historian and philosopher, Michel Foucault, argued that, since the early modern period, juridical institutions and discourses have been displaced gradually by administrative institutions and their discourses. The world of sovereigns and their landed jurisdictions had given way to a world of populations and institutions to 'normalize' these populations (Foucault 1976). If we take both the American experience and the recent proliferation of legal and constitutional norms in Europe and elsewhere as our guide, the thesis would have to revised to account for the appearance of administrators who come garbed as judges. The influence of legal and constitutional norms, vocabularies, and institutions on the socioeconomic and political structure of much of the planet has been considerable; the effects of this process, are, however, poorly understood and in need of further study. In this respect, 'postmodern' arguments have to be seriously qualified. Whether justified in fact and theory, the expectation that judges and legal norms can ground individual rights and provide a constitution for transnational and intercultural contact is shared increasingly today. The very American *Marbury v. Madison*, as it were, has found a place for itself in the global environment.

See also: Balance of Power, History of; Balance of Power: Political; Constitutional Courts; Constitutionalism; Constitutionalism, Comparative; Courts and Adjudication; Democracy; Democratic Theory; Fundamental Rights and Constitutional Guarantees; Power: Political; Supreme Courts

Bibliography

Dworkin R 1988 *Law's Empire*. Belknap Press of Harvard University Press, Cambridge, MA

Epp C R 1998 *The Rights Revolution: Lawyers, Activists, and Supreme Courts in Comparative Perspective*. University of Chicago Press, Chicago

Foucault M 1978 *The History of Sexuality, Volume 1: An Introduction*. Pantheon Books, New York

Friedman L 1990 *The Republic of Choice*. Harvard University Press, Cambridge, MA

Grossman J B, Epp C 1991 The reality of rights in an 'atolerant' society. *This Constitution* 20–8

Grovogui S N'Z 1996 *Sovereigns, Quasi-Sovereigns, and Africans: Race and Self-Determination in International Law*. University of Minnesota Press, Minneapolis, MN

Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK

Millard T F 1931 *The End of Extraterritoriality in China*. The A.B.C. Press, Shanghai, China

Pocock J G A 1971 *Politics, Language, and Time*. Atheneum, New York
Prempeh H K 1999 New jurisprudence for Africa. *Journal of Democracy* **10**(3): 135–49
Shapiro M 1994 Juridicalization of politics in the United States. *International Political Science Review* **15**(2): 101–12
Shapiro M 1998 The European Court of Justice: Of institutions and democracy. *Israel Law Review* **32**(1): 3–50
Shapiro M, Stone A 1994 The new constitutional politics of Europe. *Comparative Political Studies* **26**: 397–420
Stone Sweet A 1999 Judicialization and the construction of governance. *Comparative Political Studies* **32**: 147–84
Weber M 1954 *On Law in Economy and Society*. In: Rheinstein M (ed.) Harvard University Press, Cambridge, MA

J. B. Grossman

Judicial Selection

Judicial selection is a topic and research area that is intimately associated with the assessment of the quality of justice a legal system provides through its courts of law. What is the link between judicial selection and judicial independence? What should be the goals of a selection system? What are the principal methods for choosing judges? Do different methods produce different kinds of judges and justice? These questions need to be addressed in order to understand how justice is dispensed; they are explored here primarily with reference to American courts but with some attention to the courts of other countries.

1. *Judicial Selection and Judicial Independence*

Judicial selection was very much on the minds of the American colonists in the eighteenth century when they included among the list of 'repeated injuries and usurpations' in the Declaration of Independence that the king 'has made judges dependent upon his will alone, for the tenure of their offices, and the amount and payment of their salaries.' The framers of the American Constitution, mindful of this and also influenced by Montesquieu's work, incorporated the principle of the separation of powers among the branches of government, with the judiciary envisioned as a separate and distinct governmental entity. Unlike the English form of government, as they understood it, the judiciary was not to be linked to the legislative and executive branches. Neither the executive nor the legislative branch was to be solely responsible for the selection of judges; they were to share that power. The President was charged with nominating and the US Senate given the task of advising and consenting by either confirming or not confirming nominees by majority vote. To ensure the independence of the federal judiciary, judges were given lifetime tenure removable only by the cumbersome process of impeachment by the House of Representatives and conviction by two-thirds vote of the Senate after a trial. Thus, for the American framers, judicial selection was inextricably associated with judicial independence.

Can there be justice from a judiciary whose members owe their jobs to an executive who can remove or discipline them at will? Can a judge whose selection and retention is by election be as neutral and objective as a judge with life tenure? Are civil service methods of selection superior to more political selection methods in producing fair and objective justice?

There are a variety of judicial selection methods used to staff courts in countries around the world. Legislatures, the chief executive or heads of executive branch departments or agencies, and even members of the judiciary can be found participating in selection processes. Are there consequences of differing types of selection methods for the quality of justice?

2. *Goals of a Selection System*

To begin to answer the questions raised above, it is necessary first to formulate what ought to be the goals of a selection system. It is reasonable to argue that at bottom, a selection system ought to produce 'good' judges, that is, professionally outstanding individuals who have demonstrated that they have judicial temperament, including a reputation of being even-tempered and fair-minded, and, if already on the bench, a track record as competent and impartial. A 'good' judge will exercise his or her discretion so as not to be obviously biased at either end of the ideological spectrum. Furthermore, a good selection system will *not* discriminate against qualified women and ethnic/racial minorities, something that it can be argued is essential for the legitimacy of a judiciary.

3. *Methods for Selecting Judges in the States of the USA*

Five methods for selecting judges are used by the states and one method used by the federal government. Close to half the states elect some or all of their judges at general elections where individuals run for judicial office as members of a political party (partisan elections) or are required by law to run without a party affiliation (nonpartisan elections). About one-third of the states use the merit plan whereby a nominating commission often of lawyers and judges submits to the governor a list of names to fill a particular judicial position and the governor by law must make the

selection from that list. Eight states and the federal government select their judges by appointment by the chief executive with the approval of another body, typically the executive council or senate. Three states use the legislature to select those for judicial office with the governor having no constitutionally mandated role in the process.

In practice, scholars have found that all these methods involve politics of one sort or another. There is the most obvious form of party organizational politics that at the local level is evident with the electoral methods of selecting judges. Even where the nonpartisan ballot is used, in some states local party leaders endorse a candidate or otherwise make known their support for the individual running for judicial office. At the state level, party organizational politics are often prominent in the legislative and gubernatorial appointment systems. Governors not only are central to the gubernatorial selection method but they can influence the merit selection process by naming friends and associates to the nominating commissions. Interest groups not only can influence the outcome of judicial elections, they can exert influence on the governor and legislature. And still another form of politics is bar association politics which cuts across all selection methods—for example, the bar may rate or even endorse candidates for judgeships. And, of course, the organized bar has great influence in the merit selection process.

Electoral methods have been studied and scholars have found that judicial elections tend to be ignored by the electorate, with low voter interest and low turnout on election day. Furthermore, the majority of judges in electoral systems are initially selected for interim appointments by the governor to fill vacancies on the bench that occur between elections due to resignation, retirement, or death. This means that at the next election such judges run as incumbents and are often uncontested (which is true for most judicial elections across the board). Research has found that the party label is the most significant factor in the voter's choice.

Merit selection plans have also been the subject of scholarly investigation. The commissions themselves have been found to be characterized by intense negotiation and in some instances commissions have been found to stack the list of names so that the favorite of the commissioners would be the likely person chosen by the governor. The partisan makeup of the commissions has occasionally resulted in partisan considerations being taken into account in preparing the lists.

4. *Selection of Federal Judges in the USA*

Judicial selection of federal judges in the USA contains some similarities with selection in the states. Party organizational politics, the politics of policy, and bar association politics all come into play. But because federal judges are nominated by the President, the center of selection activity lies in the executive branch. Participants in the negotiations, particularly for lower-court judgeships include not only Justice Department and White House officials, but also US Senators and other major leaders of the President's party from the state in which a federal judgeship is to be filled or the state that by tradition expects to be represented on a federal court of appeals. Furthermore, since 1953 the American Bar Association (ABA) Standing Committee on Federal Judiciary has investigated and provided ratings that are transmitted to the executive branch of leading candidates for judgeships. An ABA rating of 'not qualified' frequently will result in the nomination being abandoned. But note that, in March 2001, President George W. Bush removed the ABA from the pre-nomination process.

Senators of the President's party expect to play a decisive role in the nomination of federal district judges from their states. Judgeships in the federal territories (including the District of Columbia), as well as those for the specialized courts and, of course, the US Supreme Court are considered more within the prerogative of the President. But the Senate institutionally has a key role to play in the process as it must confirm a nominee before he or she becomes a federal judge. When the President's party is in control of the US Senate, confirmation typically is a smooth process. But when there is divided government, the confirmation process has been bogged down in partisan and ideological battles. This is certainly a drawback of federal judicial selection.

In the nomination process, there are several considerations that ought to be mentioned. First is the professional credentials of the candidates. The legal qualifications of candidates are subject to close scrutiny by the Justice Department and the ABA making it less likely that unqualified individuals are appointed. Second is the political reality that typically 90 percent of judicial appointments go to members of the President's party. Third, senators from the state of the prospective judicial nominee must approve the nominee. Although rarely invoked, a senator can claim senatorial courtesy if the nominee is objectionable to that senator and that means that the senator's colleagues would refuse to confirm. Fourth is the policy or ideological outlook of the candidate. If an administration sees the judiciary as affecting its policy agenda, as was notable with the Franklin Roosevelt and Reagan–Bush administrations, it will tend to screen judicial nominees to eliminate those antithetical to its policies.

Because the Supreme Court is widely acknowledged to be the judicial policy maker for the nation, appointments to the nation's highest court have been the subject of keen interest and occasional controversy. A frequent presidential campaign issue has been what sorts of people will be named by the person elected president. Presidents have always been per-

sonally involved in the selection of justices and have made their choices based on a variety of considerations including the personal, the political (such as party affiliation and affect on constituent groups in the President's coalition), and the ideological. During the nineteenth century close to one out of three nominations failed in the Senate. In the twentieth century the proportion was one out of twelve (with four of the five Senate rejections occurring in the last third of the century). Thus confirmability must also be added as a consideration, especially when the Senate is in the hands of the opposition party.

5. *Evaluation of Selection Methods*

Does any one selection method produce more 'good' judges than any other method? The evidence is scanty and inconclusive that one method produces superior judges or judges with vastly different backgrounds or qualifications, with the exception that appointive methods appear to result in more women and minorities being chosen for the bench. However, not one selection method has been shown to be a superior method for the appointment of judges with the qualities that make for 'good' judges or judges whose on-the-bench performance is considered to meet the highest standards. Likewise, the evidence linking the method of judicial selection with judicial decisional trends is inconclusive.

6. *Judicial Selection and its Impact in Countries Outside the USA*

For countries other than the USA, relatively little research has been conducted on judicial selection and its impact. The principal judicial selection mode in civil law countries is that which considers the judiciary a form of civil service and part of the government bureaucracy. The judiciary is not considered a co-equal branch of government. Judges work their way up the judicial hierarchy—all outside the political appointment process. Judges are typically viewed as conservative supporters of the regime insofar as the regime is responsible for the law that judges must apply. The concept of judicial review is for the most part alien to civil law systems.

The bureaucratic route to a career judiciary may be by a competitive entrance examination that enables those who pass to attend a school for judges (the selection method, for example, in France, Spain, and Portugal) or directly by a competitive entrance examination (the method found, for example, in Belgium, Italy, Peru, and Brazil). Grades on examinations will typically determine first judicial appointments, and subsequent promotions are usually based on seniority and the assessments of senior judges or state ministries of justice. An alternative bureaucratic selection method utilizes a type of apprenticeship without examinations and schooling (as for example in Germany, Argentina, Ecuador, and Chile). Many civil law systems have recently allowed lawyers with a minimum level of professional experience to be candidates for a set proportion of judicial vacancies. Those selected are thus able to substitute their years of practice for the examination, schooling, or apprenticeship requirements.

In England, a common law country, there are some similarities to the USA in that judgeships are not considered a form of civil service and that there is a tradition of political experience and party patronage playing a role in the Lord Chancellor's selection of judges for the appellate courts (e.g., between 1832 and 1906 Members of Parliament accounted for 58 percent of appointments). In more recent decades the proportion has been dramatically lower. Higher court judges are drawn from the legal elite and tend to be drawn from the upper class. In general, partisan considerations are much more muted in England than in the USA although both countries tend to place an emphasis on professional qualifications. Women and minorities are slowly achieving representation in the ranks of the English judiciary but at a far slower pace than found in the federal judiciary in the USA. For example, in 1990 the Law Lords were all male, while there was one woman on the Courts of Appeals and on the High Court. By contrast, in the USA there was one woman on the Supreme Court, 17 on the US courts of appeals, and 50 on the district courts (roughly 10 percent of judgeships as compared to under 2 percent of the higher English judiciary).

The Appellate Committee of the House of Lords is Britain's highest court, the functional equivalent of a supreme court although its membership is somewhat flexible. The Law Lords, chosen by the Lord Chancellor although officially appointed by the Queen after consultation with the Prime Minister, are drawn from a narrow pool of lower appellate judges who themselves have risen through the judicial hierarchy. Entry into the judicial profession is from the elite group of barristers known as Queen's Counsel. The selection process generally favours the appointment of judicial conservatives.

In India, lower court judges are civil servants chosen by competitive examination, but the higher judiciary is appointed from eminent members of the bar. The president has sole authority to appoint Supreme Court justices but is constitutionally required to consult certain government officials, including the Chief Justice of India, before making the appointment. Previous judicial experience is generally required and in practice only those who have served on the High Court of one of India's states are chosen. There is also a mandatory retirement age of 65. Turnover tends to be higher and here, too, there has been a tendency for judicial conservatism to prevail.

A number of questions initially asked earlier in this article cannot be answered definitively on the basis of research conducted thus far. Much more is needed by way of systematic empirical analysis to be able to sketch in the details of judicial selection and its impact around the world. Yet it is clear that any assessment of how justice is administered needs to take into account precisely those considerations of judicial selection.

See also: Constitutional Courts; Courts and Adjudication; Judges; Legal Education; Legal Professionalism; Supreme Courts

Bibliography

Abraham H J 1998 *The Judicial Process*, 7th edn. Oxford University Press, New York

Epp C R 1998 *The Rights Revolution*. University of Chicago Press, Chicago

Gerhardt M J 2000 *The Federal Appointment Process*. Duke University Press, Durham, NC

Goldman S 1997 *Picking Federal Judges: Lower Court Selection from Roosevelt through Reagan*. Yale University Press, New Haven, CT

Griffith J A G 1991 *The Politics of the Judiciary*, 4th edn. Fontana, London

Jacob H, Blankenburg E, Kritzer H, Provine D M, Sanders J 1996 *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT

Sheldon C H, Maule L S 1997 *Choosing Justice: The Recruitment of State and Federal Judges*. Washington State University Press, Pullman, WA

S. Goldman

Jung, Carl Gustav (1875–1961)

Carl Gustav Jung was born in Switzerland at Kesswil, near Lake Constance, on July 26, 1875, a parson's son. He died at the age of 85 on June 6, 1961 as one of the great pioneers of depth psychology, in his house in Küsnacht on Lake Zurich, where he had lived and worked for more than 50 years. His external life was relatively uneventful, in contrast to the riches of his inner life, as he stated at the end of his life. It was the inner life—psychic reality—as he called it, which captured his tireless and careful attention, and formed his life and work.

In 1895, Jung began his medical studies at the University of Basle. The sudden death of his father in 1986 caused some financial troubles, and left him to take care of his mother and sister Gertrud, who was nine years younger than him. His poverty contrasted with the continued reputation of both of his late well-known grandfathers, both fathers of 13 children and both professors of Basle University. Carl Gustav Jung senior (1794–1864), a political refugee from Mannheim, Germany, had been a professor of medicine and president of Basle University. Samuel Preiswerk (1799–1871), C. G. Jung's grandfather on his mother's side, professor of theology and ancient Hebrew at the university, had served as the president of the reformed ministers in the Basle Canton. After his exams in December 1900 Jung decided to leave Basle and to follow a professional career as a psychiatrist at the psychiatric hospital at Burghölzli in Zurich.

The Burghölzli hospital, a representative building on a hill in the south of the city of Zurich, was affiliated with the newly founded University of Zurich. Since 1898, the hospital had been directed by Eugen Bleuler (1857–1936), who followed his teacher Auguste Forel. Forel was the doyen of Swiss psychiatry and had established professional standards not only for his institution but for Swiss psychiatry as a whole. Confronting strong mistrust towards the new institution from the Zurich population, he established control of the medical direction of the administration of the hospital, set legal conditions for the admission, treatment, and release of patients, and last but not least, emphasized, with regard to the discipline of psychiatry, an unbiased empirical approach to mental phenomena, focusing on the neurological, psychological, and social aspects of mental disorders. Most importantly, Forel (a Swiss francophone) introduced hypnosis (as practiced by the French School of Nancy) as a first psychotherapeutic treatment for inpatients. In hypnosis he found the proof for the existence of subconscious processes which were equivalent to externally observable neurological phenomena (scientific monism).

Jung experienced his first time at Burghölzli hospital, where he lived as an intern together with about 400 mentally ill patients, about 80 attendants, and only three other medical colleagues (including Bleuler; see *Bleuler, Eugen* (*1857–1939*)), as quite a shock: 'My inability to understand gave me such a sense of inferiority, that I couldn't bear to go outside the hospital' (Jung 1989). He read through the available literature and dedicated himself to the study of his cases. Soon he began his dissertation, '*On the psychology and pathology of so-called occult phenomena*' (1902).

In this first scientific publication Jung studied his spiritualistic experiments with his cousin Hélène Preiswerk carried out at the end of his medical studies, and treated her case in the context of other (hysterical) patients of the Burghölzli hospital. In a quite minor context he referred to Freud and Breuer's *Studies on Hysteria* and to Freud's *Interpretation of Dreams*, but mainly he relied on the new studies of Pierre Janet (see *Janet, Pierre (1859–1947)*), and Théodore Flournoy. Flournoy, professor of psychology in Geneva, had just published a scientific best-

seller *Les Indes à la planète Mars* (1900, see Flournoy 1994), a long-term study of a spiritualistic medium. Jung got in touch with Flournoy and showed special appreciation for his 'objective-distanced' approach to mental phenomena (which was influenced by his personal friend William James; see *James, William (1842–1910)*) and his having arranged the new findings in a bigger context. Jung's untroubled friendship with Flournoy outlasted the much more famous, but never as untroubled, friendship with Freud, which began several years later.

In October 1902, Jung quit his strenuous job at Burghölzli for the first time. He went to Paris to study at the Collège de France with Pierre Janet and also learnt English at the Berlitz School. He then married the 21-year-old Emma Rauschenbach (1882–1955), the daughter of a factory owner, and settled in a town residence in Zurich. At the end of the 1903, he returned to Burghölzli hospital, initially as an external representative for absent colleagues. In December 1904, he finally took over the regular post of resident doctor, and moved into the resident doctor's family apartment at the hospital. Three of his five children were born there and grew up together with Bleuler's family. Bleuler had also recently married the philologist and writer Hedwig Waser, one of the first academic women in Switzerland. Both women shared the interests of their husbands and took part in their scientific discussions.

Bleuler's psychiatric approach was formed by his previous experience as director of the remote Rheinau asylum, where he had lived as a bachelor in very close contact with more than 500 'incurable' chronically mentally ill patients. Concerning therapy, he emphasized the therapeutic value of the 'active community' (*tätige Gemeinschaft*), in which everyone worked according to their abilities in meaningful activities. For his special scientific interest—the dynamics of psychotic behavior—he favored a hypothesis of neurological conditions which caused a discontinuity in the forms of association.

It was with some enthusiasm that, in 1904, Jung and Bleuler set up a psychological laboratory for association experiments. These played a prominent part in psychological research at the time, especially in Wundt's (see *Wundt, Wilhelm Maximilian (1832–1920)*) school. Their findings soon led to passionate debates in the debating club which Bleuler had established with his friend and colleague, the Swiss Russian pioneer neurologist C. V. Monakow. In giving account not only to the regularities of the test person's responses to the stimulus words, as was usual in the Wundtian approach, but also to the 'intruders in the mind,' the irregularities (delays of reaction times, repetitions and other disturbances), Jung and his coworkers (initially Franz Riklin, Sr., Jung's wife, and several volunteers) identified the 'emotional-toned complexes' (*gefühlsbetonte Komplexe*) of the test persons as autonomous expressions of their repressed or dissociated unconscious conflicts. These results were particularly decisive for Jung. Behind completely 'mad' behavior he discovered quite normal hopes and desires. Paranoiac ideas and hallucinations revealed their hidden sense in the context of the patient's life stories. 'At the bottom we discover nothing new and unknown in the mentally ill; rather, we encounter the substratum of our own nature,' he stated. 'That was at that time a very moving experience' (Jung and Jaffé 1962).

When Jung was preparing a paper for his application as lecturer at Zurich University in 1905, he realized that his findings threw a new light on Freud's theory of repression. He decided to stand up for him, despite Freud's bad reputation in academic psychiatry. In 1906, Jung published a first volume of Association Studies, which he sent to Freud in Vienna. In 1907, he published *On the Psychology of Dementia Praecox*, which included a detailed case study with association experiments of an old female patient of the clinic, who since Forel's time had served as a demonstration case for students. These two publications inaugurated a remarkably passionate personal relationship with Freud which is documented in hundreds of letters (McGuire 1974). Bleuler shared a special enthusiasm for Freud, although, like Jung, he was always somewhat critical of Freud's recently elaborated sexual theory. In October 1907, Jung—against some initial opposition from Bleuler—founded an informal 'Freud Club' at the Burghölzli, which excluded Freud critics. In April 1908, Jung organized a first Psychoanalytical Congress. On this occasion a first joint publication was founded—the *Jahrbuch*, edited by Bleuler and Freud, and compiled by C. G. Jung. Thus Jung, with the decisive support of Bleuler, inaugurated the reception of Freudian psychoanalysis in academic psychiatry, and the beginning of the so-called 'psychoanalytic movement.'

The culmination of—as well as the turning point in—the relationship of Freud and Jung was their joint trip in September 1909 to the 20th Anniversary Conference of Clark University, Worcester, USA. Both had to lecture and were honored with their first honorary degrees. Jung had been invited independently, because of traditional connections between the Burghölzli hospital and Clark University through Forel's relations. In the six weeks together with Freud during the trip to USA, some strange personal behavior and some critical points in mutual dream interpretations inevitably revealed to Jung the limits of a friendly and cooperative relationship with Freud. Nevertheless, he continued to act for Freud as organizer and founder of the International Psychoanalytical Association, a role which he later remarked on as *collaboration à toux prix*. In 1910, he was elected president of the International Psychoanalytical Association. Shortly after Freud's provoked break with Alfred Adler in 1912, Freud challenged Jung's leadership with the organization of a 'secret committee,'

consisting only of very 'confidential' followers—excluding Jung. Jung subsequently resigned as the editor of the *Jahrbuch* in October 1913, and as president of the International Psychoanalytic Association in April 1914.

In March 1909, some months before his trip to the USA with Freud, Jung quit his job at the Burghölzli clinic and opened a private practice as psychiatrist in Küsnacht, Lake Zurich. In the 1920s and 1930s, his popularity began to grow. In 1933, he accepted the presidency of the German General Medical Society for Psychotherapy and the editorship of the affiliated *Zentralblatt für Psychotherapie*. He defended this compromising collaboration with German psychotherapists with his aim to reorganize the society as an international institution in order to help some colleagues maintain their functions, and to prevent a general abolition of psychotherapy as a medical discipline in Germany. Indeed, Jung resigned in 1940 when he saw that his endeavors had been in vain.

In 1935, he was appointed professor for psychology at the Federal Institute of Technology, Zurich (ETH); from 1944–5 he held a professorship in medical psychology at the University of Basle. In 1948, the Jung Institute in Zurich (later Küsnacht) was established by an international committee of some of his associates as a research and training center for Jungian analysts.

1. *Analytical Psychology: Contribution to Science*

Immediately after his return from the USA, Jung immersed himself in archeological and mythological literature in order to differentiate his own view of the unconscious from Freud's. The result was his major work *Wandlungen und Symbole der Libido* (1912), later published in English as *The Psychology of the Unconscious* (1916). In contrast with Freud's personalistic view of the psyche—i.e., his attempt to reduce the unconscious to a hiding place of the personal, sexual conflicts of the dreamer—for Jung, sexuality was not the sole determining factor; rather, the main point was that the unconscious had an autonomous life, which reflected not only the personal, but also the common, collective history of the mind. The manifestations of the unconscious stood not only for a causal, reductive purpose, but also had a final i.e., forward-looking meaning. Relying on some products of automatic writing (two hymns and the dramatic myth of an Indo-American hero) produced by a young American, Miss Frank Miller (first published in 1905 by Flournoy), Jung presented his own conception of the libido as a general 'psychic energy' with a dual nature, split into positive and negative poles with a steady flux between construction and de(con)struction. He also outlined an 'anatomy and embryology of mental development,' which followed common patterns of changing imaginary processes in the history of mankind as well as in the personal history of each individual person. Wavering between an analysis of Miss Miller's fantasies, some hidden self-analysis, a detailed analysis of religious hero myths of humankind from Mithraism to Christianity, and an analysis of classical and romantic philosophy and literature, he developed his theory of the structure and energetic functioning of the psyche without yet coining a decisive terminology of his own.

In the years between 1912 and 1918, Jung experienced a time of personal insecurity and scientific isolation. He resigned from his post as lecturer at the university; instead, he dedicated himself to his international practice as a psychotherapist, and concentrated on his private studies and on his confrontation with his own unconscious.

In 1921, Jung published his major book, *Psychological Types*, in which he linked his historical researches with the findings of his self-analytical process. He outlined the key concepts which later became characteristic of analytical psychology such as 'types,' 'collective unconscious,' and 'individuation process.'

He had become interested in the question of psychological types because of the differences he perceived between Adler and Freud. With critical reference to historical material, Jung elaborated a model of typology, with eight psychological types—two main attitude types of extroverts and introverts, further subdivided by the predominance of one of four psychological functions: thinking, feeling, sensation, and intuition. While sensation conveys concrete reality through the senses, intuition conveys the hidden possibilities in the background. And while thinking leads to a judgment about the meaning of concrete reality, feeling leads to a judgment of its worth. Normally one attitude or function prevails (superior function) and the opposite (inferior) attitude or function remains undeveloped, more unconscious and undifferentiated. The value of this model lies less in its contribution to differential psychology than in the theory of knowledge and the theory of the therapeutic process as a quest for wholeness.

From this experience of the relativity of every standpoint with respect to psychic reality, Jung highlighted the significance of the 'personal equation' as a prior condition of any psychological statements—i.e., the position that what a psychologist claimed to be true of human psychology as a whole was first and foremost a subjective confession. Additionally, he developed a conception of the therapeutic process, which involved not only becoming conscious of repressed contents of the unconscious but also integrating the psyche and developing a much more complete approach to inner and outer realities. This process of individuation involved withdrawing projections which are structured according to certain autonomous images of the unconscious (archetypes); these reflect the common history of mankind (archetypes of the collective unconscious) as well as the repressed images of one's own personal history (personal com-

plexes). Common and ubiquitous archetypes are the shadow, the unconscious negative side of the personality; the hero, the idealized picture of the person; the anima, the animus; and the repressed female/male soul picture in men and women. This process provokes an ethical conflict in each patient: they have to take responsibility for their mental attitudes, and develop a fuller approach to reality not only in ideas, but also in feelings, sensations, and intuition. As a result, the center of a person moves from the ego (the center of consciousness), towards the self (the center of the whole, individual, conscious and unconscious psyche). As therapeutic methods for this process, Jung developed active imagination (as a method of assimilating unconscious contents by several forms of self-expression) and amplification (circumambulation) of the symbols (instead of free association in Freudian analysis).

The foundation of the dynamic concept which became known as analytical psychology and the institutionalization of Jungian analysis were connected with the founding of the Psychological Club in 1916. This was made possible by the generous financial gift of Jung's US patient, Edith Rockefeller McCormick. After Jung's separation from the Burghölzli clinic and the Freudian circle, the 'Club' reestablished an important debating forum in which he could communicate his ideas to interested patients, colleagues, and scientists. Here many of his former patients—among them his life-long special assistant and friend Toni Wolff (1888–1953), his secretary and the compiler of his biography, the psychologist Aniela Jaffé (1903–91), and the philologist in ancient languages Marie Louise von Franz (1915–98)—were encouraged to take up public lecturing. In the 'Club' he also met outstanding academics in his main subjects of interest, such as mythology, religious studies, philosophy, psychology, and archeology. 'Club' lecturers included Karl Kerényi (1897–1973), pioneer in Greek mythology, the young Mircea Eliade (1907–86; see *Eliade, Mircea (1907–86)*), pioneer in the science of religions, and other representatives of different Christian, Indian, Jewish, and Eastern religious traditions. Richard Wilhelm (1873–1930), missionary to China and sinologist, had special impact on Jung. Wilhelm shared with him the old mystical Taoist text *The Secret of the Golden Flower*. It was in this text that Jung first discovered the correspondence of his individuation concept with Eastern philosophy and worked out his mandala symbolism for psychic wholeness or the self.

From 1933 on, the 'Club' lectures and Jung's public lectures at ETH were supplemented also by the Eranos Conferences, which took place every summer for a fortnight in Ascona, Lake Maggiore, at the house of Olga Fröbe, an enthusiastic admirer of Jung, Richard Wilhelm, and the Jewish philosopher Martin Buber (1878–1965). In the atmosphere of these meetings with contemporary authorities and pupils, Jung first presented his view of traditional Christianity and medieval alchemy, and the archetypes of the collective unconscious.

An encounter with death in 1944 marked the beginning of the period of Jung's final elaboration of ideas. His main work on traditional Christianity, *Answer to Job* (1952) and his major work on alchemy *Mysterium Conjunctionis* (Mystery of Conjunction) (1955) appeared at this time. The common principles of the reconciliation of the opposites of the dual nature of the psyche, which he had worked out in this major book, led to the final formulation of his main pioneering concept, *Synchronicity as a Principle of Acausal Relations* (1952), which he published in collaboration with Wolfgang Pauli (1900–58), professor of physics at ETH. The idea of an ultimate psychoid archetype represents the connecting principle of psyche and matter, which are two aspects of the same thing and can realize themselves in both dimensions 'synchronistically.' The relative autonomy and unpredictability of the synchronistic phenomena correspond with the new findings of the subatomic and cosmic physics, and hint toward a metaphysical order to which both the psyche of the perceiving person as well as nature itself submit.

2. *Contemporary Influence*

In contrast with Freud, who gave the impression that psychoanalysis 'may be likened to Athena's springing forth fully armed from the head of Zeus' (Eissler 1965), Jung always tried to trace back his ideas to older, universal traditions. His pioneering work was the rediscovery of an uninterrupted chain of psychic manifestations and spiritual traditions from gnosticism through alchemy to the present. His early claim for a constructive, forward-looking meaning in unconscious manifestations led to some changes in early psychoanalytic treatment which are now common in psychotherapy: for instance, the symmetrical setting (instead of the patient lying down), focus on present conditions as responsible for the outbreak of neurosis, and focus on transference and counter-transference as the crucial in changing mental attitude. In addition, Jung's conception of the spiritual nature of psychic life influenced the later transpersonal and humanistic schools of psychotherapy. His concern with spiritual mystical traditions often led to controversial claims that he himself was a mystic, selling false dreams of spiritual redemption. The alleged evidence of a recent recriminatory voice (Noll 1997) has been traced back to an obviously misidentified and misinterpreted document and shown to be fallacious (Shamdasani 1998). Jung emphasized that he had no 'doctrine' to offer, but rather some methods of looking at facts, phenomena, and natural laws. His approach to these facts is empirical in the sense of a critical reflection on

the nature of all experience which relates to self-experience and the nature of the perceiving psyche.

This view, which valuates the psyche as 'the mother of all science and arts,' reveals its special importance in the approach to religious phenomena. The unconscious psyche with its autonomous operators (collective archetypes) reflects the numinous, awful, frightening, and timeless spirit in changing images which correspond to changing religious images from ancient times to the present. In differentiating between religious images and religion itself, Jung pointed to the inner experiences open to everyone independently of every external belief system.

See also: Bleuler, Eugen (1857–1939); Erikson, Erik Homburger (1902–94); Freud, Sigmund (1856–1939); Janet, Pierre (1859–1947); Lifespan Development, Theory of; Myth in Religion; Personality Psychology; Psychiatry, History of; Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis, History of; Psychoanalysis: Overview; Psychology: Historical and Cultural Perspectives; Psychotherapy, History of: Psychiatric Aspects; Unconscious: History of the Concept

A. Graf-Nold

Bibliography

Bleuler E 1923 *Textbook of Psychiatry*. Allen and Unwin, London

Eissler V R 1965 *Medical Orthodoxy and the Future of Psychoanalysis*. International Universities Press, New York

Ellenberger H F 1970 *The Discovery of the Unconscious. The History and Evolution of Dynamic Psychiatry*. Penguin Books, London

Flournoy T 1900 *Les Indes à la Planète Mars. Atar*, Genève 1994 *From India to the Planet Mars. A Case of Multiple Personality with Imaginary Languages*. (ed. S. Shamdasani, transl. Vermilye D)

Flournoy T 1994 *From India to the Planet Mars: A case of multiple Personality with Imaginary Languages* [ed. Shamdasani, trans. Vermilye D]. Princeton University Press, Princeton, NJ

Fordham F 1953 *An Introduction to Jung's Psychology*. Penguin Books, Melbourne, Vic

Forel A 1907 *Hypnotism; or Suggestion and Psychotherapy*. Rebmann Company, New York

Jones E 1953 *The Life and Work of Sigmund Freud*. Basic Books, New York

Jung C G 1902 *Zur Psychologie und Pathologie sogenannter occulter Phänomene*. O. Mutze, Leipzig [*On the Psychology and Pathology of So-called Occult Phenomena*]

Jung C G 1916 *Psychology of the Unconscious. A Study of the Transformation and Symbolism of the Libido. A Contribution of the History of the Evolution of Thought*. Moffat, New York

Jung C G 1921/1923 *Psychologische Typen*. Rascher Verlag, Zürich 1921 [*Psychological Types, or: The Psychology of Individuation*. Routledge and Kegan Paul, London 1923, CW 6]

Jung C G 1944 *Collected Works* [ed. Read H, Fordham M, Adler G, McGuire W]. Princeton University Press, Princeton, NJ

Jung C G 1952 *Antwort auf Hiob*. Rascher Verlag, Zurich [*Answer to Job*. 1954, Routledge and Kegan Paul, London]

Jung C G 1955 *Mysterium Coniunctionis. Untersuchung über die Trennung und Zusammensetzung der seelischen Gegensätze in der Alchemie*. Rascher Verlag, Zürich [*Mysterium Coniunctionis. An Inquiry into the Separation and Synthesis of Psychic Opposites in Alchemy*]

Jung C G 1989 *Analytical Psychology, Notes on a Seminar, Given 1925* [ed. McGuire W]. Princeton University Press, Princeton, NJ

Jung C G, Jaffé A 1962 *Memories, Dreams, Reflections*. Pantheon, New York

Jung C G, Pauli W 1952 *Naturerkläung und Psyche*. Racher Verlag, Zürich [1955 *The Interpretation of Nature and the Psyche*. Pantheon, New York

McGuire W (ed.) 1974 The Freud/Jung letters [trans. Mannheim R, Hull R F C]. Princeton University Press, Princeton, NJ

Noll R 1997 *The Aryan Christ: The Secret Life of Carl Jung*. Random House, New York

Shamdasani S 1998 *Cult Fictions. C. G. Jung and the Founding of Analytical Pychology*. Routledge and Kegan Paul, London

Smith C D 1990 *Jung's Quest for Wholeness. A Religious and Historical Perspective*. State University of New York Press, New York

Juries

The jury is a group of laypersons recruited by the legal system to resolve a dispute. In contrast to professional judges, jury members are amateur and nonrecurring decision makers. They are selected from the community at large rather than for their legal expertise, and they adjudicate disputes only on occasion rather than on a routine basis. The jury is both a cultural icon and a popular scapegoat, praised as a champion of democracy, and denounced as incompetent or biased when it reaches an unpopular verdict. Empirical research offers a more measured view of the jury, revealing a motivated set of decision makers, often admirable in the way they handle the complex and contentious disputes they are asked to resolve, but always human in their struggles to deal with limited resources and imperfect information and to balance competing values in reconstructing events and assessing mental states.

The ideal jury too conjures up multiple images: a thorough and impartial judge of the evidence presented at trial, a faithful representative of the community from which it is drawn, a compliant follower of its legal instructions, a protector of citizens against the power of the state, and a repository of commonsense justice that tempers the harsh hand of the law. Critics charge that juries fail to meet one or more of these standards. Behavioral scientists who study the jury

find some variation from this set of ideal attributes, in part because the attributes can be inconsistent with one another, in part because the jury's cognitive and motivational characteristics influence its behavior.

1. Jury Variations

The jury has gone through numerous incarnations over time and exists internationally in a variety of forms. The early English jury was limited to male property owners who were summoned to decide a case because they were familiar with the facts. They acted as witnesses, investigators, and tribunal, determining both the facts and the law. The modern American jury is drawn from the citizenry at large, with the parties entitled to eliminate some prospective jurors if they have preconceived notions about what the verdict should be before the trial begins. Jurors generally are expected to sit quietly through the trial, apply the law as it is described in the judge's instructions, deliberate until they reach a verdict or an impasse, and disperse without personal consequences, except for the occasional inquiries of the American press in high-profile trials. The right of citizens to have both criminal and civil controversies resolved by groups of ordinary fellow citizens is guaranteed by the US Constitution and by every state constitution.

Although the majority of criminal trials decided by a jury occur in the USA, juries are also part of the criminal justice system in the countries of the British Commonwealth. Moreover, a resurgence in juries is taking place (e.g., in Spain and Russia), reflecting the perception of citizen adjudication in the justice system as a mark of democratic reform. In civil cases, jury trials occur almost exclusively in the USA and Canada. England eliminated juries in most civil cases by the 1930s, retaining the civil jury primarily in defamation cases.Most modern juries are small, compared to the ancient Greek jury of several hundred that tried Socrates. For example, the Brazilian jury has seven members, and some American jurisdictions have juries of six. Under the ancient Greek jury system and the modern Brazilian jury system, deliberations are omitted and verdicts are determined by the position taken by the majority in a secret ballot vote. Like the American and English juries, the Spanish jury includes deliberation, but it adds the requirement that the jury provide a rationale for its verdict.

Although the jury is prominent in discussions about the American legal system, legal disputes are infrequently resolved by the verdict of a jury, even in the USA. Most criminal cases in the USA end in dismissals or guilty pleas; most civil cases are dismissed or settled. An estimated 150,000 jury trials occur annually in state courts and an additional 10,000 in federal courts. The influence of the jury extends far beyond the trials it actually decides, however, as litigants and attorneys anticipate what a jury would do if asked to resolve the dispute. Inaccurate expectations are common, based on impressions gleaned from high-profile but idiosyncratic trials and from trials resulting in unusually high awards that attract media attention. The absence of direct and reliable feedback on what would have happened had a jury actually decided a case that was resolved without trial provides little opportunity to correct misimpressions.

2. Political Roles of the Jury

In addition to adjudicating disputes, the jury is a potential source of legitimacy for the legal system. To the extent that the jury is viewed as representing a fair cross-section of the community, its verdict is likely to be seen as the product of fair consideration and can carry a legitimacy that the decision of the judge, as an employee of the state may lack. Even when its verdict is unpopular and even if observers believe that the jury does not fairly represent the community, the jury acts as a lightning rod, insulating the judge and other parts of the state legal system from criticism.

The jury also can act as a conduit for community standards. For example, in evaluating a claim of self-defense, the jury must determine what a reasonable person would be expected to believe, as well as what the particular defendant did believe. Although the jury is charged with applying to the facts the law it receives from the judge, the self-defense example illustrates the fuzziness of the division between law and facts. The jury must often inject its understanding of appropriate standards into its fact-finding even while scrupulously following the instructions that the judge provides.

A final political role for the jury is its educative function, identified by De Toqueville as the jury's great strength. Surveys suggest that more than half of American adults have had some personal involvement with the courts, and of this subgroup, half have served on a jury. Citizens also receive information and misinformation on the courts from other sources, including the media. The extent to which jury service provides additional or corrective information is unclear, but jury experience tends to make jurors feel more positive about the jury system

3. Methodological Challenges in Studying the Jury

Study of the jury is a particularly thorny project because the jury speaks publicly only to deliver its verdict: deliberations occur outside the presence of any nonjuror. Although not all jurisdictions have legislation that bars access to jury deliberations, there have been only a few instances in which deliberations have been observed or recorded. Courts have been

reluctant to permit observation even when no law forbids it. Researchers interested in the jury have generally been left to study the jury more indirectly. They have used a variety of approaches (each with strengths and limitations): archival studies of jury case characteristics and verdict patterns (records provide information on large samples of cases, but limited variables are recorded by courts, and local commercial jury verdict reporters systematically under-report cases); post-trial interviews with jurors (the jurors are reporting on the actual process of reaching a verdict, but what they report is limited to what jurors noticed, accurately recorded, remember, and are willing and able to report); surveys of other trial participants, such as judges and attorneys (these respondents are informed court observers, but they have only indirect information on jury behavior and their reports may be influenced by the verdict); field experiments (these can offer strong evidence on the impact of a legal reform like juror note-taking, but provide little information on process and are hard to implement correctly); and simulation experiments (simulations are strong on process information and provide unambiguous causal inference, but whether the findings can be generalized depends on correspondence with dynamics of real jury behavior).

4. *How Jurors Evaluate Evidence*

The traditional legal model of the jury trial portrays jurors as passive recipients of the evidence and legal instructions. Empirical studies of jury behavior find jurors to be active processors of incoming information. Early models of jury decision making included, first, averaging models in which jurors assess and weigh pieces of evidence, combining the results to reach a verdict and, second, Bayesian models in which jurors consider and evaluate each new piece of information, revising their position on the appropriate verdict in light of their prior position and the additional evidence. These formal models have enjoyed limited success as descriptions of how jurors actually decide cases. Explanation-based models of jury decision making, such as Pennington and Hastie's (1986) story model, provide an account of jury behavior that comports better with empirical evidence. Consistent with the story model, jurors do not simply record and store the evidence for later use as they receive it. Rather, they actively select and organize the trial evidence to construct a story about what happened. The story they construct is based on the evidence, but jurors also use it to fill in gaps in the evidence with inferences based on their understandings about how the world works. Jurors arrange evidence in the form of a sequence of motivated human actions that include important events, the circumstances of the case, inferences about character, and the parties' motivations and states of mind. By influencing jurors' understanding of what took place, the order in which facts are presented (i.e., in story order rather than witness order) can affect verdicts.

5. *Jury Composition*

The modern American jury is far more heterogeneous and representative of the citizenry at large than was the early English jury or even the jury in the early twentieth century. Nonetheless, the jury is not a random sample of citizens. It is the product of a multistage selection process that typically begins with a list of potentially eligible jurors drawn from voter registration lists and often supplemented by individuals holding drivers' licenses in the general geographic area where the court sits. Prospective jurors may be excused from jury service on the basis of hardship, but losses also arise from geographic mobility, a failure to update the lists, and non-response by prospective jurors to a court summons. The loss of prospective jurors in the qualification and summons process results in a systematic under-representation of minorities, younger individuals, and those at lower income levels.

The final stage in jury selection occurs when prospective jurors are brought into the courtroom and questioned to determine whether they will serve in the particular case. Those who clearly express preconceived notions about what the verdict in the case should be or those with clear conflicts of interest are excused by the judge (the challenge for cause). In addition, the parties can excuse a limited number of prospective jurors without giving a reason (the peremptory challenge). The US Supreme Court has ruled that peremptory challenges based on race or gender are constitutionally prohibited, but it is likely that the prohibition fails to eliminate racially and gender motivated challenges. Courts generally do not require much justification from the party initiating the peremptory challenge when the opposing party charges that the challenge was improperly motivated. Also, the small number of challenged jurors in the typical trial means that an attorney can generally identify a unique and non-discriminatory reason for each challenge.

The result of this variety of shaping and sometimes cross-cutting forces is that juries tend to be somewhat more educated, wealthier and older, and less likely to include a representative number of minorities than the distribution of these groups in the adult population. The selection process that occurs before the jurors enter the courtroom is primarily responsible for the difference between the composition of the American jury and the adult population as a whole. Although the difference is likely to persist, the American jury today is more representative than it has ever been, and is more heterogeneous than the juries of other coun-

tries with jury system. Moreover, jury participation is extensive. Surveys indicate that 25 percent of American citizens are likely to serve on a jury trial at some point in their lives.

6. *Jury Size and Decision Rule*

Traditional English and American juries consisted of twelve members who were required to reach a unanimous verdict. Some American jurisdictions now permit juries with as few as six members and non-unanimous verdicts of 9:3 or 10:2; England permits a 10:2 verdict if the jury has been unable to reach unanimity within a specified period of time. Reducing jury size increases the likelihood of an aberrant jury verdict. A majority decision rule tends to reduce the rate of hung juries and to shorten deliberations because jurors in the majority do not need to gain the support of all jury members. It may also result in less thorough deliberations if the deliberators can arrive at a verdict without considering the reasons why there is disagreement among the members of the jury.

7. *Individual Differences*

Attempts to predict juror verdict preferences from background characteristics have generally been able to explain only a modest amount of variance in outcomes. Demographic characteristics like gender, race, and age generally account for very little of the variation in response. Attitudinal characteristics can be more powerful, albeit also modest, predictors.

Expectations, beliefs, and values affect the way that jurors react to evidence. In that respect, jurors are no different from any other decision makers, because people ordinarily scrutinize more carefully and are more likely to reject information that is inconsistent with their beliefs and expectations, because it is generally easier for people to remember theory-consistent information than theory-inconsistent information, and because ambiguous information tends to be interpreted as theory-consistent.

Some types of cases are more likely than others to implicate strongly held beliefs or values. A primary example involves the death penalty: even among jurors willing to impose the death penalty in some cases, the strength of their support for the death penalty can strongly influence the likelihood that they will vote for a death sentence.

8. *Pretrial Publicity*

The American constitutional right to a free press occasionally provides the public with information or misinformation about the case that a jury will be asked to decide. Under such circumstances, the constitutional rights to a free press and to a fair trial are potentially at odds, a conflict not faced in countries like Canada and Great Britain where the press is prohibited from writing about impending trials. Pretrial publicity about issues like a defendant's alleged confession can affect the expectations and beliefs of prospective jurors in the USA. Although some limited research indicates that the effects of some types of pretrial publicity are generally small, it also suggests that jury selection and the passage of time cannot be depended upon to eliminate them. In some high-profile cases, the legal system faces not only the cost of moving a trial, but the increasing difficulty of identifying a location that media coverage has not saturated with pretrial coverage.

9. *Competence, Including Reactions to Experts*

Both criminal and civil trials increasingly involve the testimony of experts who present technical and scientific evidence (e.g., DNA analysis, econometric modeling, epidemiological and toxicological research). Such complex evidence presents a challenge for the experts who must communicate with a lay audience, for the judge who acts either as a gatekeeper in ruling on which experts can testify and what they can say or as the trier of fact, and for the lay jury.

Although jurors typically exert considerable effort in trying to understand the content of expert testimony, questions arise about the ability of jurors (and judges) to evaluate technical and scientific evidence. Jurors (and judges) can be confused by expert testimony, but that confusion can often be reduced or avoided if the expert is also a good teacher. Moreover, jurors frequently express skepticism about the trustworthiness of experts presented by parties within the adversary system and they adopt reasonable strategies to evaluate expert credibility. In addition, juries can and do draw on the expertise of their most competent members to assess the strength of the evidence, providing a resource that is unavailable to the trial court judge.

10. *Reactions to the Law, Including Nullification*

Jurors are instructed to apply the law as the judge describes it to the facts as the jury finds them. Juries may not always follow the instructions, but the primary cause of that failure is the inadequacy of the instructions rather than a willful disregard by the jury. Jurors may fail to follow the judge's instructions on the law if they are either unmotivated or unable to apply the instructions, but they typically see themselves as obligated to apply the law and spend a significant portion of their time during deliberations discussing the law. Nonetheless, legal instructions often fail to provide jurors with helpful legal guidance. The major goal in writing jury instructions on the law is to produce language that will accurately reflect the relevant law so that appellate courts will find the instruction acceptable. Little attention is generally

given to ensuring that the meaning of the instruction is clearly communicated. Although comprehension of instructions is often low, it can be significantly improved by using plain language, simplifying sentence construction, providing definitions for technical terms, and specifically correcting jurors' inaccurate expectations and misconceptions.

Although the most common source of jury deviations from legal standards is a failure of the legal instructions to convey clearly what the appropriate legal standard is, jurors also may deviate from the path outlined in the instructions due to cognitive or motivational biases. For example, jurors admonished to disregard particular information may find it difficult to do so. Other legal instructions may ask the jurors to engage in demanding mental gymnastics—for example, to use a defendant's criminal record only to assist in evaluating his credibility, but not as evidence of bad character; to forget that they learned about damaging evidence that the judge ruled inadmissible—and jurors may be unwilling or find it impossible to perform the required cognitive adjustments. It is unclear whether these failures are significantly less likely when the trier of fact is a judge, or whether they represent heuristic patterns of using information that neither can be expected entirely to overcome.

Finally, jurors may depart from the judge's legal instructions when the application of the legal standard to the particular case so substantially violates the jurors' sense of justice that they are persuaded to temper the letter of the law in that case. Although such deviations are rare, explicit jury nullification plays a central role in conceptions of the appropriate role of the jury and has been a source of extensive debate. Although courts have long recognized the power (as opposed to the right) of the jury to nullify, courts and commentators have argued about whether juries should be told about that power. When jurors are explicitly told that they have the right not to apply the law as the judge describes it, they are more willing to reach verdicts that avoid the literal application of the law.

11. Deliberations

Deliberations resulting in a group verdict distinguish the jury from its chief alternative, the trial court judge, in two ways. By requiring agreement from multiple jurors, the jury verdict in principle reduces the likelihood that the decision will represent an idiosyncratic view of a single deviant decision maker. Moreover, in theory, deliberations give the jury an opportunity to profit from the resources of its multiple members, to pool its knowledge and sensibilities, and to resolve differences, ultimately producing a verdict that reflects more than what could be achieved either by a single decision maker or by mechanically combining or averaging the preferences of the individual members.

The extent to which deliberations actually do affect jury verdicts, however, is in dispute. Some scholars (e.g., Kalven and Zeisel 1966) have suggested that jury verdicts simply reflect the position of the majority before deliberations begin. The picture they present is consistent with the verdict-driven jury that takes an immediate vote to see where each juror stands and then focuses its attention on persuading the minority to join the position initially held by a majority of the jurors. When a vote is immediate, it is likely to reflect predeliberation preferences. When a discussion of the evidence precedes a vote (the so-called evidence-driven jury), that vote will be affected by any changes that have occurred as a result of the discussion. Although jurors often call for an immediate vote, discussions can interrupt before a vote is completed, so that first votes often are not immediate and they imperfectly reflect the individual predeliberation preferences of the jurors. Nonetheless, most juries probably do end up with a verdict that reflects the majority that is visible by the time most of the jurors have expressed a verdict preference in deliberations. The majority, using both normative and informational pressure, persuades the minority to accept its position. The outcome of deliberations is less predictable when a majority position is less likely to emerge because there are multiple verdict choices or in civil cases where damages are at issue.

12. Civil Liability and Damages

Civil juries are sometimes viewed as pro-plaintiff when it comes to liability, although there is little evidence to support the claim that they are more likely than judges to find the defendant liable. Although jurors do hold corporations to a high standard of behavior, the standard primarily arises from an expectation that corporations are better able to control their behavior than an individual might be. On the issue of damages, the pattern of jury behavior is less clear. Compensatory damage awards generally vary with the severity of the injury (i.e., vertical equity), but there is some evidence that plaintiffs with small losses tend to be overcompensated and those with large losses tend to be undercompensated. Compared to special damages (e.g., lost wages, medical expenses), the difficult-to-measure general damages (e.g., pain and suffering), show greater variation in the amounts awarded for injuries at the same level of severity (i.e., horizontal inequity). In the absence of a standard for the 'correct' amount that will compensate for pain and suffering, some scholars have suggested that juries be supplied with information on the pattern of pain and suffering awards in previous jury trials involving similar injuries. Although the problem of defining 'similar injuries' remains, this additional information may reduce some unwarranted disparity across cases.

Civil juries are sometimes instructed that they can award damages for punishment and deterrence (i.e.,

punitive or exemplary damages) as well as damages for compensation. Punitive awards are rare, however. They most frequently occur in cases involving intentional harms and financial rather than physical harms.

13. The Jury vs. Alternative Triers of Fact

Discourse on the performance of the jury frequently ignores a basic reality of legal systems: in evaluating the jury, the relevant comparison is not some hypothetical ideal decision maker, whatever qualities such a model decision maker would have. Rather, the appropriate comparison is the human alternative or set of alternatives that might be used in the jury's stead. Alternatives to the jury include professional, legally trained judges and mixed tribunals composed of laypersons and legally qualified professionals. Studies that have compared the reactions of laypersons and legal professionals to the same criminal trials indicate, first, high levels of agreement on verdict (74–90%), and, second, a greater inclination to acquit by laypersons when they disagree with the professionals. In addition, agreement rates do not vary with the complexity of the trial, suggesting that disagreement is not a product of an inability of the laypersons to understand the evidence.

The assumption of the legal system that judges can 'compartmentalize' and discount legally inadmissible information when they are exposed to it appears to be an inflated expectation. However, the few studies that compare the process of decision making by judges and jurors show a surprising correspondence even in response to case characteristics that might be expected to influence laypersons more than professionals. Both laypersons and professional judges fail to disregard inadmissible evidence, are influenced by outcome information in their judgments about the probability of that outcome (the hindsight bias), and are reluctant to find for the plaintiff when presented with statistically based evidence. If judges are influenced by inadmissible evidence, the jury trial may provide an effective way to avoid burdening the trier of fact with much of the potentially biasing information that judges regularly confront. In addition, the trial judge in a jury trial retains the power to override a conviction by a jury if the judge determines that the conviction is unwarranted by the evidence, providing a rarely used but available extra protection for the defendant in a criminal jury trial.

14. Reactions to Jury Trial and the Reform Movement

As a lightning rod for popular debate, the jury is regularly the subject of calls for reform. Some proposed reforms, such as those advocating that jurors be permitted to take notes and to submit questions for witnesses during the trial, do not assume that juries are failing to perform adequately but rather are designed to assist them in reaching well-considered judgments, to improve the comfort of the conscripted citizens who serve as jurors, and generally to optimize jury performance and juror satisfaction. Other proposed reforms aim at reducing the right to jury trial or limiting the power of the jury, such as calls for a complexity exception to the right to jury trial in civil cases and for legislative caps that place a ceiling on jury awards. High-profile trials resulting in unpopular verdicts have stimulated some of the proposed changes in the jury system as have the concerted lobbying efforts of particular interest groups (e.g., the insurance industry). It is likely that calls for reform of all kinds will continue to follow the jury. The real question is whether the increasing scholarly literature on jury behavior will inform popular and political discussions about the jury.

See also: Courtroom Consultation: Psychology; Courts and Adjudication; Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Judicial Selection; Procedure: Legal Aspects; Rules in the Legal Process; Social Insurance: Legal Aspects.

Bibliography

Abramson J 1994 *We, The Jury: The Jury System and the Ideal of Democracy*. Harper Collins, New York

Daniels S, Martin J 1995 *Civil Juries and the Politics of Reform*. Northwestern University Press, Evanston, IL

Diamond S S, Casper J D 1992 Blindfolding the jury to verdict consequences: damages, experts, and the civil jury. *Law & Society Review* **26**: 513–63

Ellsworth P C 1989 Are twelve heads better than one? *Law and Contemporary Problems* **52**: 205–24

Galanter M 1993 The regulatory function of the civil jury. In: Litan R (ed.) *Verdict: Assessing the Civil Jury System*. Brookings, Washington, DC

Greene E 1989 On juries and damage awards: the process of decision-making. *Law and Contemporary Problems* **52**: 225–46

Hans V P, Lofquist W S 1992 Jurors' judgments of business liability in tort cases: implications for the litigation explosion debate. *Law & Society Review* **26**: 85–115

Hastie R, Penrod S D, Pennington N 1983 *Inside the Jury*. Harvard University Press, Cambridge, MA

Heuer L, Penrod S D 1996 Increasing juror participation in trials through note taking and question asking. *Judicature* **79**: 256–62

Horowitz, I 1985 The effect of jury nullification on verdicts and jury functioning in criminal trials. *Law and Human Behavior* **9**: 25–36

Kalven H Jr, Zeisel H 1966 *The American Jury*. University of Chicago Press, Chicago

Lempert R O 1993 Civil juries and complex cases: taking stock after twelve years. In: Litan R (ed.) *Verdict: Assessing the Civil Jury System*. Brookings Institute, Washington, DC

Lloyd-Bostock S, Thomas C 1999 Decline of the 'Little Parliament': juries and jury reform in England and Wales. *Law and Contemporary Problems* **62**: 7–40

MacCoun R J 1989 Experimental research on jury decision making. *Science* **244**: 1046–50

Munsterman G T, Munsterman J T 1986 The search for jury representativeness. *Justice System Journal* **11**: 59–78

Pennington N, Hastie R 1986 Evidence evaluation in complex decision making. *Journal of Personality & Social Psychology* **51**: 242–58

Saks M J, Marti M W 1997 A meta-analysis of the effect of jury size. *Law and Human Behavior* **21**: 451–67

Sandys M, Dillehay R 1995 First-ballot votes, predeliberation dispositions, and final verdicts in jury trials. *Law and Human Behavior* **19**: 175–95

Smith V L, Studebaker C A 1996 What do you expect?: The influence of people's prior knowledge on crime categories on fact-finding. *Law and Human Behavior* **20**: 517–32

Thaman S 1999 Europe's new jury systems: the cases of Spain and Russia. *Law and Contemporary Problems* **62**: 233–59

Vidmar N 1995 *Medical Malpractice and the American Jury: Confronting the Myths about Jury Incompetence, Deep Pockets, and Outrageous Damage Awards*. University of Michigan Press, Ann Arbor, MI

S. S. Diamond

Justice and its Many Faces: Cultural Concerns

Claims for justice and protests against injustice are ubiquitous in social life. Political movements, revolutions, and wars are initiated under the banner of justice. Justice is a prominent issue in all fields of politics. The courts are swamped with law suits, and the demand for new forms of conflict settlement is increasing. The injustice suffered by individuals is the core of everyday conflict in private life, and the victims of misfortune have to cope with the injustice of their fate. The concern for justice seems to be an anthropological universal.

However, there does not appear to be a universal consensus on what is considered to be just or unjust. We speak of justice in the singular as if there were only one just solution for any problem. Yet there may be diverging views on what would be a just solution for any given injustice, and on which principles of justice should be applied. Many conflicts concerning justice become intense because people are not aware of the basic dilemma structure of such conflicts. This will be outlined in the first section of the present article. The second section will deal with procedural justice and the settlement of justice conflicts. In the third section, the justice motive will be depicted as a primary motive which may assume many different faces.

1. Principles of Justice and Disputes about Their Validity

Justice is typically regarded as the *normative standard for decision making and valuation*. In actual fact, there is no single standard of justice, but instead many different ones. This is true for all three domains of social life in which justice is critical: distribution, social exchange, retribution and the acknowledgment of deeds. Moreover, the application of different standards of justice results in diverging and conflicting outcomes.

1.1 Distributive Justice

In *the domain of distribution*—i.e., the allocation or distribution of material resources or symbolic goods, rights, duties, positions, power, opportunities, taxes, etc., within or between groups or populations—*equality* certainly constitutes the basic idea of justice. However, equality can be specified in many different ways. It may mean *equal shares for all human beings* or, if that is impossible or dysfunctional, equal opportunities. Or else, according to Aristotle, it may mean *equal shares or equal opportunities for all equal human beings*. This second view implies that human beings differ and that certain differences justify unequal allocations and inequalities in the overall distribution of resources, rights, duties, etc., (for instance, according to their membership to a group, organization, or population; according to their social status, specific merits, professional qualifications and productivity, age, gender, neediness, etc. Another rule of distributive justice is the conservation of the status quo and of previously acquired rights). In this sense, equality means equal shares for all those in the same position, for all those with the same kinds of merits or needs, and so forth. The equal opportunity rule can be specified, for instance, as the equal chance rule, in which a lottery procedure is used when it is impossible for goods to be split up (e.g., in the allocation of organs for transplantation), or the rotation schedule, which may be appropriate when it would be dysfunctional for a position to be split.

Walzer (1983) somewhat neutralized the problems of justice with the inequalities in allocation and distribution by his concept of '*complex equality*,' which postulates that distributions in different 'spheres of justice' (material wealth, social recognition in various contexts, political power, education, kinship and love, recreation time, etc.,) are not perfectly correlated. Thus, a lower rank in one sphere may be compensated by a higher ranking in another. Moreover, the subjective importance of the different spheres of justice varies, so that perceived overall inequalities may be reduced further.

However, the question of which differences between people justify which inequalities in allocations and

distributions remains open. In the ongoing discussion, arguments are inspired by cultural traditions and social philosophies such as egalitarianism, liberalism, social welfarism, utilitarianism, and the human and civil rights movement. We are far from having reached a general consensus in this question: the preferred standards of distributive justice vary between and within cultures.

This is true in all fields of distributive justice. For instance, the populations of postcommunist states are less tolerant of inequalities, and ascribe more responsibility for individual welfare to the government than do the populations of states with a liberal tradition (Kluegel et al., 1995). The rules of justice for the allocation and withdrawal of scarce goods such as scholarships, state-subsidized housing, jobs, etc., vary greatly within and across states (Elster 1992). In the dismissal of employees, for instance, the criteria applied include the seniority of the employees, their acquired skills, current productivity, neediness, responsibility for a family, age, and gender. Though it may be possible to justify the application of each of these criteria, each will result in different decisions, and not in a single just solution.

1.2 Justice within Social Exchange Relationships

Modified forms of the equality criterion for appraising justice can also be found in *other domains of social and societal life*. In *exchange relationships* the equality criterion is found in *the principles of reciprocity and equity* as well as in *the concept of fair contract*. Social exchanges are regarded as just if *reciprocity* is established. This is true for positive exchanges, where equal mutual advantage is the normative standard, as well as for negative exchanges, as portrayed in the biblical rule of 'an eye for an eye and a tooth for a tooth.' Equal mutual advantages have also been postulated in the social contract conception of the state (Hobbes 1651, 1968) and in the sociological theory of role-bound exchanges (teacher-student, physician-patient, employer-employee, leader-follower; in Parsons, [1951]). However, the general truth of this postulate of equal mutual advantage has been criticized with good reason (Gouldner 1960): Some social roles or positions are certainly more attractive, powerful, and profitable than others. *Equity* has also been proposed as the normative standard for informal social exchanges, implying equal ratios of investments, costs and outcomes, and benefits for all parties involved (Homans 1961). The equity principle appears to be more clear-cut than it actually is, however, since the evaluation of justice or injustice is guided by personal balances of subjectively focused exchanges (of goods, services, love, respect, trust, loyalty, harm, negligence, hostilities, etc.).

Moreover, indirect exchanges (e.g., services to third parties or to the community which are of indirect benefit to the partners in the exchange) may or may not be included in these subjective evaluations—the same is true for same chained exchanges, e.g., in the generation contract (Emerson 1976). (The generation contract involves members of one generation supporting the older generation in the expectation that, in their own old age, they will be supported by the next generation.)

The *contract* is a prototypical form of social exchange. Contracts are regarded as just when the partners are equally informed and equally free to consent to them (Nozick 1974). Justice is threatened if relevant information is withheld, if pressure is exerted, or if one party is not free to refuse to enter into the contract on account of a certain predicament. But even contracts made by informed and freely consenting parties may be unjust with respect to third parties (e.g., cartel contracts made at the expense of others).

1.3 Retributive Justice and the Justice of Awards

Within the *domain of retributive justice* equality is specified firstly as *impartiality* in blaming and punishing people for their misdeeds or in praising and rewarding people for their achievements and heroic or moral deeds. Impartiality means that the same standards are applied to similar people in similar circumstances. This allows different standards to be applied to different people. For example, different standards of achievement may be applied in schools so that culturally or socially disadvantaged or less gifted pupils have any chance of being praised and rewarded. Secondly, equality in the domain of retributive justice is specified as *proportionality*, e.g., of punishment and guilt, reward and excellence. The major problem with proportionality is to determine how blame, guilt, praise, and achievements can be reliably gauged. Another crucial question with respect to retributive justice remains unanswered, however, namely which norms should be applied in the evaluation of actions. For instance, the atrocities legalized by the laws of a totalitarian regime cannot be punished if these laws are observed, but taking human and civil rights as normative standards here would violate the juridical prohibition against enacting laws retrospectively.

1.4 The Justice of the Social System and Politics

Since *the constitution*, *the legal system*, and *the institutions* of a society impact on distribution, social exchange, and retribution, these components of the societal system are also the objects of justice appraisals, e.g., the economic system, the labor laws, the health and welfare system, the educational system, environmental protection laws, immigration rules, the

generation contract—all of these maybe valued as basically just or may be criticized as unjust as well as *the politics* responsible for the implementation and the adaptation of the system. Constitutions may be criticized for failing to guarantee human rights, the protection of the environment, or animal welfare; the economic and tax systems may be criticized for allowing the development of huge inequalities in wealth or for demotivating individual productivity; the educational system may be blamed for failing to provide equal opportunities for the socially disadvantaged or failing to meet the needs of gifted students. Arguments concerning justice can be found to support or censure any given policy. And all parties concerned usually seem convinced that their view is the only valid one. This is how normative standards are conceived—as generally valid and as binding for all concerned.

1.5 What is Unjust, What Would be Just?

It is easier to reveal injustices than to establish justice within complex social and societal systems. In the following, a few examples will be given to illustrate this thesis.

An action such as the distribution of grades may be evaluated as unjust with respect to a specific standard or principle of justice, e.g., the correspondence of grades and performance. Any failure to establish this correspondence would be unjust and subject to correction. This is an easy case. It becomes much more complicated when the performances are not assessed objectively, or when the principle of justice is questioned, e.g., when those giving the grades are asked to consider the individual, social, and contextual conditions of achievement which may have been favorable or unfavorable, fairly or unfairly allocated, justly or unjustly distributed.

In many cases, efforts to correct one injustice result in the creation of another. A prominent example of this is the affirmative action taken to correct the indisputable historical disadvantages of women on the labor market which still persists to this day. For instance, woman are still underrepresented in top-level positions. Interpreting this fact as an unjust discrimination against women, various types of affirmative action have been taken to give preference to female applicants over their male competitors. The problem with this policy as far as justice is concerned is that it aims to correct a historical problem by reversing gender privileges in today's generation of students and young professionals. Today's young women have certainly suffered fewer injustices in education and on the labor market than earlier generations of women, and today's young men are less privileged than their predecessors. Therefore, it is problematic for these historical injustices to be corrected by affirmative action affecting only the present generation of young women and men, and having no impact on preceding generations. The costs of this measure should not be borne exclusively by young men. Moreover, it would be an illusion to believe that the whole group of women will vicariously participate in and profit from the success of some young women. Women are not a social group, but a social category. Individually, most are members of gender-mixed social groups such as families, where they are bound not only to daughters, sisters, and mothers, but also to sons, brothers, and fathers. They may be proud of the careers of the male as well as the female members of their family.

Another example is the environmental protection policy. There is little doubt that current and future environmental pollution not only entails health risks, but the risk of climate changes with incalculable consequences. There is no doubt that the pollution of air, soil, and water entails gross injustices. As pollution is a side-effect of productive processes in agriculture, industry, and business, and of traffic, air conditioning, etc., the concern of justice is about the allocation of the profits, costs, and risks resulting from these processes. It goes without saying that some people and populations benefit more than others from these processes, and those who profit most are unlikely to be those who have most to lose from the risks and disadvantages of pollution. The costs of pollution are still typically externalized, meaning that the costs and risks are not borne by those responsible for their causation, or by those who benefit from the processes in question. One of the policies for reducing pollution is the internalization of the costs by means of pollution taxes. However, the problems for justice created by pollution are not solved by raising taxes, and this measure may result in new problems of justice. Firstly, there is no guarantee that the taxes raised will be used to compensate the victims of pollution. In fact, not all of the victims of pollution can be identified, as its harmful effects are long-lasting, may interact with many other factors, and may be cumulative. Furthermore, air and water pollution extend across rgional and state boundaries and may involve delayed risks for the future, when coming generations may be affected by the greenhouse effect. Secondly, pollution taxes may intensify social inequalities. They may be ruinous for poorer firms, but not for richer ones, and prohibitive for poorer car owners, but not for more affluent drivers. The result may be the bankruptcy of some firms, the consequent dismissal of employees, and the aggravation of economic inequalities, potentially causing new problems of social justice.

A final example: It is a rule of retributive justice that crimes are to be punished. Nevertheless, what exactly constitutes a just punishment for a concrete crime is constantly under discussion. The assessment of personal guilt is difficult, as is the assessment of what constitutes an efficient punishment with respect to various purposes—general and individual deterrence, resocialization, protection of the general public, or the victims' claims for retaliation. Equal punishment for

the same category of crime and similar degrees of guilt may seem just, but may have an entirely different impact depending on the personality, life perspective, and social context of the offender. For some it may represent nothing but a small hiccup in the life plan, for others it will destroy the whole life perspective. Moreover, the impact of legal punishments is frequently not restricted to the offender; those who are close to the offender are also affected, and may suffer social discrimination and economic hardship which they have not deserved.

Claims for justice result from perceived injustices. Injustices are typically perceived as the violation of a single principle of justice, and from a restricted view of reality. The intensity of resentment about perceived injustices does not corroborate the validity of the personal or collective views. Efforts to correct perceived injustices frequently result in new injustices if further principles of justice are not taken into account, or if the complexities of social and physical reality are not acknowledged. Awareness of existing justice dilemmas and of the complexity of social systems is usually necessary for sustainable and broadly acceptable solutions to be found.

2. *Procedural Justice and the Settlement of Conflicts*

Given this broad spectrum of options for appraising and constructing justice in the various fields of social life, conflicts about justice are preprogrammed. How can these conflicts be settled? How can peace be preserved? One answer is to ensure fair procedures of decision making, procedures which are broadly accepted as just within the population (Luhmann 1983). This holds for parliamentary procedures, elections, institutional decisions, decisions of courts, arbitrators and other authorities. The separation of powers is a crucial element in the appraisal of justice in a given society—this includes the right to appeal against institutional and parliamentary decisions in administrative and constitutional courts. Extensive empirical research has confirmed the impact of fair procedures on conflict resolution and on the acceptance of authoritative decision making. It is not only the impartiality of the authorities, the consistent use of norms and arguments, and the objectivity in the use of relevant information which people consider to be criteria of fairness—the granting of a 'voice' to all concerned, meaning that all concerned are given the opportunity to present their views and claims and to influence the decision-making process, is also seen as an essential criterion (Thibaut and Walker 1975, Leventhal 1980). There is ample general evidence of what has been dubbed 'the fair process effect,' describing the phenomenon that perceived procedural fairness helps the parties involved to accept decisions or outcomes which are less favorable than they had expected or hoped for. Decent and polite treatment by the authorities may add to this effect (Lind and Tyler 1988). There is also evidence that perceived procedural fairness bolsters trust in the justice of institutions.

The settlement of conflicts—in the political arena as well as in all other fields of social life—requires the parties to acknowledge that conceptions of justice different from their own are not completely erroneous, and that the realm of justice is not without contradictions. It might be helpful to demonstrate the divergent conceptions of justice in different fields of life, social contexts, historical periods, and cultural traditions. It is necessary for the parties involved to understand that *dilemmas of justice* are unavoidable since conflicting standards of justice all have some validity. Conflict settlement therefore requires all conflicting principles and views of justice to be considered. Moreover, solutions which harmonize them have to be sought. Diverging constitutional principles such as freedom, equality, and solidarity have to be harmonized in the same way, since one-sided applications of each individual principle in societal practice would violate the others.

It is the *wisdom of institutions* to pay heed to several contradicting justice standards in their policies. The 'social market economy,' for instance, is conceived as a way of harmonizing liberal capitalist principles with social welfare maxims. In his theory of justice, Rawls (1971) has tried to integrate the right of all citizens to free economic activities with the right of all citizens to participate in the general welfare. His 'maximin' principle justifies both freedom of economic activity for all and the resulting inequalities in economic wealth, but restricts freedom and social inequality to the point that the maximum benefit will result for the least privileged category of citizens. When asserting every citizen's right to participate in the general economic wealth, Rawls also justifies redistribution and regulation by the state.

Similarly, most of the procedures for the allocation or withdrawal of scarce goods described by Elster (1992) at the level of institutions and corporations combine several principles of distributive justice. Together with elements of procedural fairness, this might explain why these procedures are broadly accepted without considerable public protest in all of the countries studied.

3. *The Justice Motive and Its Many Faces*

It is not generally acknowledged that the *concern for justice* is *an anthropological universal*. When asked why people care about justice, one of two answers are generally given: (a) humans have a primary or primordial motive for justice (Lerner 1980), (b) they believe that it is in their own interest to apply and to respect the standards of justice. These two answers are radically different. The first implies that people

want or need justice as an end in itself. The second answer implies that the concern for justice is merely a function of another motive, namely self-interest (Walster et al. 1978).

This second answer mirrors the *rational choice hypothesis* of social contract theories in political philosophy. From Hobbes to Rawls, the metaphor used has been that of a fictitious original pre-state situation, in which a crowd of individuals consensually contracts to establish a society with a binding normative system and powerful institutions restricting individual freedom because they believe that this would be mutually advantageous to all and that it would best serve their own interests in the long run.

By applying the rational choice model, concerns for justice are reduced to self-interest. For instance, it may be argued that equity rules are observed in business exchanges because the actors expect this approach to pay off in the long run of continued exchanges. Similarly, politicians may commit themselves to the cause of social justice because they hope to win the votes of those whose claims they are voicing. An employer may pay fair wages in the hope that this will enhance employees' motivation and performance. In such cases, justice is not the primary concern of the actors—not an end in itself—but a useful 'rational' choice for individuals aiming to maximize their own outcome.

There is no doubt that standards of justice may be observed for such reasons. But *once internalized, justice is no longer an option which is rationally deliberated*. Instead, *it imposes itself as an ought*, valid for oneself and for others, in the same way as Shweder and Haidt (1993) have postulated for moral norms. As long as justice is only an optional element in making rational choices, it is not an unconditionally valid ought, and can be disregarded without any moral disquiet if other useful options arise. As an ought, however, every unjustified violation will evoke moral reactions—guilt when the subject him- or herself has failed to meet a norm of justice by either action or omission, and resentment when others have violated such a norm (Montada 1998).

Justice concerns and self-concerns may be *confounded*, e.g., in cases of relative personal deprivation when compared with similar others, relative deprivation of one's own group when compared with other social groups (Crosby 1976), tangible cuts in salaries, increases of taxes or the prices of basic foodstuffs (Moore 1978), or when one's own interests are frustrated (Pastore 1952). In such cases, individual diagnostics will be required to decide whether self-interest or the justice motive is under primary threat. Empirical evidence indicates that violent responses only become probable in any of these cases when the frustration of self-interest is perceived as unjustified and illegitimate (Major 1994).

However, *self-interest* and *justice concerns* may also *conflict* with one another. Many of those who live in privileged conditions or who have been advantaged by fortune have *'existential' guilt feelings* when thinking about the fates and hardships of less fortunate others. Evidence of existential guilt feelings has been gathered from Holocaust survivors, Hiroshima survivors, released prisoners of war, and 'survivors' of layoffs, as well as from heterogeneous samples of well-situated people confronted with the hardships of poor people in the Third World, unemployed individuals in their own country, and other socially and physically disadvantaged groups (see Montada 1998 for a review).

Compelling evidence for the *prescriptive nature of justice* is provided by observations that subjects without any vested interest of their own resent injustices inflicted upon others, and engage in the support of these victims. It is not unusual for *social movements* to be initiated and *supported by people without any vested interests of their own*. Keniston's (1968) study of young radicals involved in the anti-Vietnam war movement and the study by Haan et al. (1968) of activists involved in the 1960s civil rights movement in USA serve as examples here.

However, other manifestations of the justice motive suggest that views of justice and injustice are personal or social constructions within specific contexts and realities. This can be observed in processes of emotional coping with experiences of loss, e.g., by denying obvious injustices or emphasizing questionable cases of injustice (Montada 1994). It can also be observed in the social emergence of perceived entitlements and political protest (Major 1994), as well as in mechanisms by which personal and social deprivation can be ignored (Wegener 1987).

The three psychologically fascinating phenomena described in the following can be interpreted as examples of *trade-offs between concerns of justice and self-interest*.

Blaming innocent victims was a phenomenon observed in many experiments conducted by Lerner and his students (Lerner 1980), as well as in studies carried out elsewhere (Montada and Lerner 1998). Blaming victims is plausibly interpreted as subjects doing an injustice in an attempt to preserve their Belief in a Just World which is, or would be, threatened by the victimization of innocent people. Belief in a Just World can be assumed to be a psychological resource which may be defended by attributing responsibility and blame to the victims themselves, thus reframing the injustice of their victimization.

Another phenomenon is the *exchange fiction* (Lerner 1980). It has been observed that many people tend to prefer buying an overpriced article, when they know that the profits will go to needy people, rather than directly donating an amount of money corresponding to the price difference. The explanation for this is that helping a needy person establishes a commitment and personal responsibility for this and other needy persons. Thus, any act of helping is problematic in several respects: It implies an acknowledgment of undeserved

neediness, thus threatening the Belief in a Just World, it creates a continuing and generalized responsibility for needy people which may interfere with personal concern and, finally, further injustices would emerge with respect to all other similarly needy people who have not been helped. Nevertheless, provided that they are not to blame for having inflicted their hardships upon themselves, needy people deserve to be helped. This conflict is best solved by an exchange fiction which allows the donation to be masked as a purchase.

The third phenomenon is called the *free riding dilemma*. Everyone will agree that free riding—profiting without investing—serves self-interest. Those who do invest resent free riders for their selfishness. A relevant observation here is that opinion polls conducted in Europe during the 1990s revealed that two-thirds of employees were, in principle, willing to reduce their working hours and income (on average by 10 percent to 20 percent), if this would result in the creation of new jobs for the unemployed. However, this willingness was very rarely translated into action. At least two hypotheses explain this discrepancy: (a) self-interest (preserving one's level of income) finally outweighs the justice motive, (b) moving the targets of social comparison changes the objects and contents of the justice motive: When considering mass unemployment and its undeserved sequel, it would be a just decision for the relatively privileged to share their working time and income. When comparing themselves with free riders (other full-time employees who are not willing to share their privileges), however, subjects who do choose to share would feel relatively deprived. Thus, not sharing can be justified as long as free riding is not prevented at the societal level. What at first glance appears to be selfish behavior may well be motivated by justice concerns.

In sum, the justice motive is an anthropological universal, but one which may assume many different faces.

See also: Criminal Justice, Ethics of: Aspects of Human Dignity; Equality and Inequality: Legal Aspects; Injustice: Legal Aspects; International Justice: Philosophical Aspects; Justice and Law; Justice: Philosophical Aspects; Justice: Political; Justice: Social Psychological Perspectives; Social Justice

Bibliography

Ackermann B 1980 *Justice in the Liberal State*. Yale University Press, New Haven, CT

Crosby F A 1976 A model of egoistical relative deprivation. *Psychological Review* **83**: 85–131

Elster J 1992 *Local Justice*. Russell Sage Foundation, New York

Emerson R M 1976 Social exchange theory. *Annual Review of Sociology* **1**: 335–62

Gouldner A 1960 The norm of reciprocity. *American Sociological Review* **25**: 1011–78

Haan N, Smith M B, Block J 1968 Moral reasoning of young adults: Political-social behavior, family background, and personality correlates. *Journal of Personality and Social Psychology* **10**: 183–201

Hobbes T 1648/1968 *Leviathan*. Penguin, Harmondsworth

Homans G C 1961 *Social Behavior: Its Elementary Forms*. Harcourt, Brace & World, New York

Keniston K 1968 *Young Radicals: Notes on Committed Youth*. Harcourt Brace, New York

Kluegel J R, Mason D S, Wegener B (eds.) 1995 *Social Justice and Political Change: Public Opinion in Capitalist and Post-Communist States*. Aldine de Gruyter, New York

Lerner M J 1980 *The Belief in a Just World: A Fundamental Delusion*. Plenum Press, New York

Leventhal G S 1980 What should be done with equity theory? New approaches to the study of fairness in social relationships. In: Gergen K J, Greenberg M S, Willis R H (eds.) *Social Exchange: Advances in Theory and Research*. Plenum Press, New York

Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum Press, New York

Luhmann N 1983 *Legitimation durch Verfahren*. Suhrkamp, Frankfurt/Main

Major B 1994 From social inequality to personal entitlement: the role of social comparisons, legitimacy appraisals, and group membership. In: Zanna M (ed.) *Advances in Experimental Social Psychology* **26**: 283–355. Academic Press, New York

Montada L 1994 Injustice in harm and loss. *Social Justice Research* **7**: 5–28

Montada L 1998 Justice: Just a rational choice? *Social Justice Research* **11**: 81–101

Montada L, Lerner M J 1998 (eds.) *Responses to Victimization and Belief in a Just World*. Plenum Press, New York

Moore B 1978 *Injustice: The Social Bases of Obedience and Revolt*. Macmillan, London

Nozick R 1974 *Anarchy, State and Utopia*. Basic Books, New York

Parsons T 1951 *The Social System*. Free Press, Glencoe, IL

Pastore N 1952 The role of arbitrariness in the frustration-aggression hypothesis. *Journal of Abnormal and Social Psychology* **47**: 728–31

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Shweder R A, Haidt J 1993 The future of moral psychology: Truth, intuition, and the pluralist way. *Psychological Science* **4**: 360–65

Thibaut J W, Walker L 1975 *Procedural Justice: A Psychological Analysis*. Erlbaum, Hillsdale, NJ

Walster E, Berscheid E, Walster E W 1978 *Equity: Theory and Research*. Allyn & Bacon, Boston

Walzer M 1983 *Spheres of Justice: A Defense of Pluralism and Equality*. Basic Books, New York

Wegener B 1987 The illusion of distributive justice. *European Sociological Review* **3**: 1–13

L. Montada

Justice and Law

Justice has been called the fundamental moral virtue of social institutions (Rawls 1971). In similar fashion, 'legal justice' is the name commonly given to the fundamental moral virtue of legal institutions.

Theorists through the centuries have differed substantially over what this virtue is, and what relation legal justice bears to other forms of justice. In its commonsense meaning, according to which legal justice means treating people as the law provides, legal justice helps ensure that people are treated consistently and predictably. At the same time, however, it seems unlikely that treating people as the law provides invariably means treating them justly, because the law is at best an imperfect instrument of justice in its broader sense. Some theorists have responded to this concern by offering alternative accounts of legal justice. Others, of more skeptical bent, have doubted that legal justice has much to do with real justice, either because they deny that legal institutions have moral aims, or more commonly because they believe that other virtues, such as economic efficiency, pragmatic problem-solving, or compassion, are more fundamental aims of legal institutions.

1. *The Commonsense Conception of Legal Justice*

In its primary everyday sense, legal justice means treating people as the law provides. This view may be referred to as the *commonsense conception* of legal justice. Under the commonsense conception, law, rightly interpreted and properly applied to facts determined rationally, yields legal justice—indeed, 'legal justice' turns out to be nothing more than the name for the state of affairs in which law is rightly interpreted and properly applied to facts determined rationally. The phrase has no meaning independent of this. Under the commonsense conception, legal justice is roughly synonymous with legality itself. It is what most people mean when they speak of the 'rule of law,' an abbreviation for 'the rule of law, not of men.' Where the rule of men subjects people to the will of rulers, which may well be partial or capricious, the rule of law impartially subjects people only to the law itself (see *Rule of Law*).

In Western philosophy, it was Plato (428–347 BC) who in his late dialogue *The Laws* first introduced the idea that human rulers should be 'slaves of the law.' His student Aristotle (384–322 BC) also endorsed the rule-of-law idea in his work *Politics* (Bk. 3, ch. 16); it is also Aristotle who, in *Nichomachean Ethics* (Bk. 5) provided the first systematic analysis of the forms of justice. It is noteworthy, however, that Plato did not endorse the rule of law wholeheartedly. Drawing on the analogy of physicians who change their course of treatment when they find it is killing their patients (Statesman 295c–297c), he argued that officials should depart from legal rules when the good of the city requires it. The fear that rigid legality may lead to injustice haunts the commonsense conception of legal justice.

The commonsense conception raises large jurisprudential questions. How is law interpreted *rightly*? How is it applied *properly*? How are facts determined *rationally*? Competing philosophical schools offer various theories of legal interpretation; proceduralists disagree about how to adjudicate properly; and evidence scholars balance epistemological accuracy, investigative cost, and the protection of personal rights in rational fact-determination. As a result, the commonsense view that legal justice means treating people as the law provides is not as straightforward as it seems. It is fraught with ambiguity—an essentially contested concept if ever there was one. The commonsense conception of legal justice turns out to be an entire family of conceptions.

Perhaps the most fundamental divide among theories of legal interpretation and law application exists between views that try to bind decision-makers through objective general criteria and views that encourage decision-makers to exercise moral judgment. An example of the former is so-called *intentionalism*, the idea that laws should be interpreted according to the intention of those who drafted them, ascertained through objective historical evidence. An example of the latter is the theory of the distinguished philosopher Ronald Dworkin (b. 1931; see for example Dworkin 1977, 1980), who believes that judges are required to interpret law in the morally best light it can sustain. (For further discussion on this point see the articles *Legal Reasoning and Argumentation*; *Procedure: Legal Aspects*; *Scientific Evidence: Legal Aspects*)

2. *Justice and Legal Justice*

Regardless of how its ambiguities are resolved, the commonsense conception implies that legal justice will never be any better than the law itself. If the law itself is unjust, then treating people as the law provides will produce injustice. US Supreme Court Justice Oliver Wendell Holmes wrote in a letter, 'I have said to my brethren many times that I hate justice, which means that I know if a man begins to talk about that, for one reason or another he is shirking thinking in legal terms' (Holmes 1929). Legal justice, understood as legality, differs essentially from what might be called extralegal justice, or *justice simpliciter*. If Holmes is right that legal institutions do not deliver *justice simpliciter*, then two rather disquieting conclusions appear to follow. First, the term 'legal justice' is fundamentally misleading: the word 'justice' in it really doesn't mean justice. Second, there may be no reason to think of legal justice as a moral virtue at all. A good murderer is not good, and legal justice may be unjust.

Moreover, even when a law is good, inflexible application may yield outcomes in particular cases that are inequitable—unjust, in another everyday sense of the term—because of the exceptional facts of the case. Laws are general in scope, but, as Aristotle observes, even a good general rule may yield an unjust outcome when the case doesn't fit the rule. Thus,

alongside the problem of unjust laws there is a problem of unjust outcomes. Both problems generate a tension between justice and legal justice.

2.1 The Natural Law Alternative to the Commonsense Conception of Legal Justice

The most ambitious approach to the problem of unjust law simply denies the possibility that law can be unjust. At first blush, this sounds absurd, but the proposition that unjust law is not really law at all (*lex injusta non est lex*) has had powerful defenders in the so-called *natural law* tradition, among them St. Augustine and St. Thomas Aquinas (see *Natural Law*). The natural law argument may be put as follows. Inherent in the idea of law is the claim that the law ought to be obeyed, that the law creates a moral obligation over and above whatever sanction attaches to it. That is why a tax law is simply not the same as a gunman's command to hand over your wallet, and lawful authority is not merely the 'gunman writ large.' The difference is that one does nothing blameworthy by evading the gunman, but it is commonly believed that tax fraud is wrong—more generally, it is commonly believed that people have a moral obligation to obey the law. If the law is unjust, however, the moral obligation vanishes, and with it the status of law (Soper 1984).

Roman Catholic natural law theorists put the argument somewhat differently. Legal systems are instituted for a purpose, namely to promote the common good. The common good requires that there be law-makers and law-enforcers, who solve the various coordination problems that arise in organized society. Their role in fulfilling the common good is the sole reason that law-makers and law-enforcers have authority over others, and why others ought to obey them. The law-making authority therefore has an implicit limit: if legislators are not aiming at the common good, they have exceeded their authority, and the laws they enact are laws in form alone (Finnis 1980).

The intuitive basis to the natural law argument is that every recognizable legal system is built around a core set of rules, such as the proscriptions of violence, theft, and fraud. These are moral requirements recognizably directed toward the common good; they are 'natural law' in the sense that reason alone seems to dictate that they belong in every legal system. To natural lawyers, the idea that any arbitrary set of rules could constitute a legal system is unrealistic, because it denies that reason itself reveals the core of natural law at the heart of legal systems.

The natural law position can best be understood as an alternative to the commonsense conception of legal justice. Where the latter understands legal justice to mean treating people as the law provides, the natural law conception understands legal justice to require treating people in a manner consistent with natural law. The commonsense conception imbues legal justice with the virtues of impartiality and predictability; the natural law alternative imbues legal justice with substantive concern for the common good.

To the natural-law argument there are at least two responses. The most prominent is the argument of legal positivists (see *Legal Positivism*). Positivists hold that law consists at bottom of nothing other than posits of the sovereign authority. Any valid enactment by the sovereign authority is law, whether or not it is just. From the positivist point of view, natural lawyers are merely playing definitional games when they insist that a duly-enacted statute should not be honored with the title 'law' unless it is just. Taking away the word 'law' does not take away the problem of unjust law, because officials remain duty-bound to treat people as the statute provides. The commonsense conception of legal justice cannot be merely defined away.

Furthermore, even if the natural law argument is right that not just any capricious set of rules counts as law, and that only a system devoted to the common good will be recognizable as a legal system, it does not follow that every single law in a legal system is just: a system devoted to the common good can contain flaws and blind spots. The natural law argument applies most plausibly to entire legal systems, not individual laws. Perhaps a largely unjust system is not really law; but an unjust enactment of a largely just system is presumptively valid. In that case, the problem of unjust law reappears within the natural-law conception of legal justice.

A distinctive version of the natural law argument was devised by the American philosopher Lon Fuller. According to Fuller, a legal system must satisfy eight procedural requirements. To count as laws, enactments must be general in scope, public, prospective rather than retroactive, clear, consistent, relatively constant, capable of being obeyed, and enforced as written. Fuller believed that law serves as a baseline around which citizens organize their activities, and citizens must be able to know, understand, and comply with the baseline. The eight procedural requirements ensure that law satisfies these conditions. For Fuller, then, legal justice has more specific content than the commonsense conception—it includes the eight procedural requirements—but less content than the traditional natural law view.

Fuller's eight requirements have been widely accepted as an analysis of the rule of law. Fuller also believed that any legal system satisfying the eight procedural requirements would be substantively just, but his various efforts to demonstrate this ambitious conclusion are unpersuasive. American slave law satisfied the eight procedural requirements; so did many legal regimes that treated women as mere property of their fathers or husbands. Such legal regimes were procedurally just but substantively unjust, and most theorists agree that Fuller's theory of

'procedural natural law' fares no better than traditional natural law theories in eliminating the problem of unjust law.

2.2 Building Justice-enhancing Devices into the Law

An alternative approach to the problem of unjust law concedes that unjust law is possible, but tries to narrow the gap between justice and legal justice by building into legal systems devices for bringing law into line with justice. For example, legal language can be interpreted flexibly rather than rigidly, or interpreted in its morally best light, as Dworkin proposes. To take another example, many legal systems permit statutes to be invalidated when they conflict with constitutional principles, which are often framed in broadly moralistic language. In addition, common-law criminal juries have the power to nullify unjust law by acquitting a defendant in defiance of the law. Officials, too, may decline to enforce unjust laws, nullifying them through a kind of benign neglect. Arguably, nullification defeats legal justice in the name of *justice simpliciter*; but some theorists have pointed out that the discretion to nullify law is itself sanctioned by law, and may be regarded as an escape clause built into the concept of legal justice (e.g., Kadish and Kadish 1973). Law must be administered, and because laws are stated in general form, officials will always have some discretion to determine when laws apply; thus, the power to nullify will always exist.

Devices like these may well lessen the tension between justice and legal justice, but few theorists believe that they eliminate it. No method of interpretation is flexible enough to yield any desirable result, nor is every unjust law automatically unconstitutional. Furthermore, it seems unlikely that juries and officials will nullify unjust laws in every case—nullification is a device better suited to the exceptional case than to the rule itself.

2.3 Law and Equity

What about situations in which a law is just, but applying it strictly in a particular case yields an unjust outcome? Aristotle defines the word *equity* as 'a correction of law where it is defective owing to its universality,' and argues that justice requires that law should yield to equity. This, arguably, is what nullifiers are up to. In Aristotle's memorable image, rules of law should be like the soft lead rulers that builders use to measure irregularly shaped molding: 'the rule adapts itself to the shape of the stone, and so too the decree is adapted to the facts' (*Nichomachean Ethics*, bk. 5, ch. 10). (For further discussion, see *Equity*.)

But while the utility of soft rulers is clear, soft rules are more problematic. In contemporary terms, Aristotle was arguing that law should consist of context-dependent standards rather than rigid rules. A law that says 'drive no faster than 100 kilometers per hour' is a rule; a law that says 'drive at a reasonably safe speed' is a standard. Standards, unlike rules, are adaptable to facts in the case-by-case way that Aristotle recommends.

There are two problems with replacing rules by standards, however. The first is that standards are useful only when there is widespread social agreement on how they apply across a broad range of cases; where no such agreement exists, law must resort to rules if it is not to become fatally vague. For that reason, mature legal systems always contain both rules and standards.

The second problem is that the shift from rules to standards will not eliminate the need to supplement law with equity. Standards, like rules, must be general, and injustice can therefore result from applying them in particular cases. When a driver is rushing a stroke victim to the hospital, for example, he or she may be driving faster than 100 kilometers per hour *and* driving at an unsafe speed. If it seems unjust to punish the driver, that will not be because the speed limit takes the form of a rule rather than a standard—rather, it is unjust to punish the driver because his or her unusual situation does not fit any universal speed limit, either a rule or a standard.

Perhaps, then, the universality of laws should be abandoned. The problem with this approach is that the universality of legal rules and standards is taken by most theorists to be the hallmark of laws, distinguishing them from orders and directives. Furthermore, the virtues of legal justice lie precisely in the universality of law. Because law is general, legal justice treats like cases alike, and in that way it is fair and impartial; in addition, generality makes the law predictable, which allows people to plan their affairs around the law. Thus, treating people as the law provides honors the values of fairness and predictability, whereas allowing decrees to adapt to the facts undermines both these values.

2.4 Alternative Dispute Resolution

One way to reduce the tension between law and equity is to divert a dispute from the formal legal system to mediation, arbitration, out-of-court settlement, or other forms of alternative dispute resolution (ADR). Proponents argue not merely that ADR can resolve disputes more quickly, cheaply, and cooperatively than formal adjudication, but that also by particularizing outcomes to the unique facts of each case ADR creates higher-quality justice—it creates something that looks more like equity than law.

Here as elsewhere the argument is very ancient. In Greek authors of the Homeric period, the word

'justice' (*diké*) was synonymous with a mutually-agreeable settlement. In *The Laws*, Plato distinguished between the best judge, who destroys the wicked, the second-best judge, who subordinates the wicked to the good, and the third-best judge, who reconciles the contending parties with each other. Plato argued that in a real, non-utopian city the third-best judge, the one who settles cases through compromise, is better than the other two (627e–628a).

However, the virtue of ADR and the settlement process—that ADR makes an end run around legal formalities—raises questions about how to evaluate the justice of settlements. Negotiations, which often turn on nothing but naked bargaining power, can yield outcomes fully as inequitable as trials. Furthermore, critics of ADR suspect that informal processes may reinforce unequal power relationships, and that the lack of formal protections will leave vulnerable groups even more vulnerable. Should the justice of a mediated settlement be evaluated against the standard of an ideally fair trial (the 'shadow verdict'), or an ideally fair negotiation (the 'shadow bargain'), or something entirely *sui generis*? How are any of these standards to be determined? Does a compromise promote extra-legal justice, or does it merely compromise legal justice? Such questions yield no easy answers. (For further discussion of this point, see *Mediation, Arbitration, and Alternative Dispute Resolution (ADR)*; *Negotiation and Bargaining: Role of Lawyers.*)

2.5 Origins of the Conflict between Justice and Legal Justice

How grave the tension between *justice simpliciter* and legal justice is depends on what *justice simpliciter* consists of. Philosophical theories of justice differ dramatically from each other (see *Justice: Philosophical Aspects*), and none commands anything resembling a consensus. When Holmes wrote 'I hate justice,' his skepticism can be read as a complaint about the vagueness and debatability—hence, the contentlessness—of the concept of extralegal justice.

Yet despite the philosophical controversies, there is a recognizable Western concept of justice of which the philosophical theories may be said to offer alternative interpretations or conceptions. This concept draws on the two chief sources of Western moral thought, Greek philosophy and the Bible.

For Plato and Aristotle, and likewise for the biblical authors and redactors, justice is primarily a virtue of persons rather than of institutions. Thus, Plato's *Republic* investigates justice in the city only because Plato believes that the city is a plausible analogue to the individual soul, his real topic of interest; so too, Aristotle places his account of justice in his treatise on ethics rather than his treatise on politics. In the Bible, a just community is simply a community filled with just people. In this respect, the classical and biblical sources differ profoundly from contemporary theorists, who treat justice as a property of institutions rather than of persons.

It was Aristotle who first observed that the words 'unjust' and 'just' are ambiguous—the former refers both to 'the lawless man and the grasping or unfair man,' and similarly the latter refers both to the law-abiding and the fair man (*Nicomachean Ethics* Bk. 5, Ch. 1). One form that the virtue of justice assumes, then, is obedience to the law. Related to this form are special virtues of officials—to punish lawbreaking, and to judge impartially (which is a requirement found both in the Hebrew Bible and in Aristotle).

What about *justice simpliciter*, that is, justice apart from law-observance? Both the Jewish and Christian traditions locate the core ideal of justice in the Golden Rule and the injunction 'Love your neighbor as yourself' (Leviticus 19:18; Matthew 22:39), which commentators often regard as alternative formulations of the same duty. In addition, both traditions lay special emphasis on the injustice of neglecting or oppressing the poor. For Jews, loving your neighbor as yourself is so closely identified with the duty to care for the needy that the Hebrew word *tzedakah* carries the double meaning of 'justice' and 'alms-giving.' *Tzedakah* is a duty of the just individual, and a community will be unjust to the degree that it neglects the needy. In contemporary discussions, people often use the term 'social justice' in precisely this way. Similar themes appear in non-Western religious traditions as well (including a version of the Golden Rule in the Analects of Confucius).

This sense of *justice simpliciter* as solidarity with the least fortunate will inevitably generate conflicts with legal justice. Even within Leviticus 19, the duty to care for the needy lies in tension with the duty of judges not to favor the poor. Furthermore, as functionalists argue, laws aim to maintain and preserve a social order; law is an innately conservative institution, whereas the biblical conception of social justice is innately utopian and radical. One cannot really expect that laws in any society will be systematically formulated to enforce wealth-transfers from the prosperous and powerful to the poor and powerless.

There is another reason as well to anticipate divergence between legal and extralegal justice. Since Aristotle, it has been traditional to distinguish distributive justice from corrective justice. The former concerns fundamental questions of distribution in society; the latter deals with the rectification of injuries, both civil and criminal. Legal institutions such as courts deal principally with issues of corrective justice, and have very little to do with distributive justice (which is not to deny that legal decisions often have distributive effects). As writers from Aristotle onwards have observed, corrective justice may run contrary both to moral virtue and distributive justice—if a good man accidentally injures a wicked one, or a deserving poor man injures an unjustly wealthy one, the injurer

will owe compensation, even though in the former case the good man may be ruined for the sake of the bad, and in the latter case the distribution of wealth becomes even more lopsided than it was before. Legal institutions take established distributive principles and patterns as a given, regardless of whether they are just. In this sense, then, even when courts achieve corrective justice, they may well reinforce distributively unjust practices.

3. *Alternatives to Legal Justice*

So far this article has surveyed four conceptions of legal justice: the commonsense conception, according to which legal justice secures the goods of impartial consistency and predictability in the law; the traditional natural law conception, according to which legal justice secures the common good; Fuller's 'procedural natural law,' according to which legal justice provides an intelligible baseline against which people may order their own affairs; and the Aristotelian idea that genuine legal justice combines universal legal standards with equity. However, none of these is able to relieve the tension between legal and extralegal justice. The question remains whether other alternatives do better. Three alternatives have become prominent in contemporary debates.

3.1 *Law and Economics*

One response is simply to deny that achieving extralegal justice is an important aim of legal systems. A particularly influential contemporary version has been offered by law-and-economics theorists. According to Richard Posner (b. 1939), a founder of law and economics, the aim of legal systems is to maximize aggregate social wealth. If legal rules systematically fail to maximize social wealth, disputing parties could realize shared gains by bargaining around the law; inevitably, then, the law gravitates toward wealth-maximizing rules. Furthermore, wealth maximization is a normatively desirable aim, better than utilitarianism because it cultivates productive virtues rather than pleasure-seeking (Posner 1983). Critics, however, have found both the descriptive and normative claims of law and economics exaggerated. Inefficient legal rules sometimes linger for centuries, and the argument that a law prohibiting murder or rape can only be justified if it can be shown to maximize social wealth seems unsupportable (Dworkin 1980, Kronman 1980). (For further discussion of this point, see *Law and Economics: Empirical Dimensions*.)

3.2 *Pragmatism*

Another prominent alternative to the commonsense conception of legal justice is *pragmatism*. As one proponent defines pragmatic adjudication, 'a pragmatist judge always tries to do the best he can for the present and the future, unchecked by any felt duty to secure consistency in principle with what other officials have done in the past' (Posner 1998, p. 237). To the objection that this is an invitation to lawlessness, pragmatists respond that doing the best for the present and the future will in some cases require the judge to apply law and precedent in a rigidly formalistic manner; in other cases, however, pragmatism requires judges to look past formalities.

Pragmatists collapse legal justice into extralegal justice, where the latter is understood as achieving the best consequences overall. In one respect, pragmatism descends from the legal realists (see *Legal Realism*), who argued that legal formality is merely a disguise for implicit policy-choices, which should be made explicit and be guided by concern for consequences. In another sense, pragmatism represents a radical version of Aristotle's plea for context-specific equitable standards. The attractions of pragmatism are plain, but the general objection to broad standards applies in an acute form to pragmatism: there will seldom be agreement on what 'doing what is best for the present and the future' requires in any controversial case; thus, the virtues of predictability, transparency, and treating like cases alike will erode from the law.

3.3 *Feminism*

Feminists influenced by the work of psychologist Carol Gilligan have argued that the ethic of impartial justice is a quintessentially male construct—one that is in no way superior to the feminine 'ethic of care.' To illustrate the ethic of care, feminist legal scholars have fastened on the image of Shakespeare's Portia, who pleads that the 'quality of mercy' should replace rigid legal justice. The ethic of care focuses on repairing relationships and is directed toward individual needs rather than rules, policies, or the best overall consequences; unsurprisingly, this leads some proponents of the ethic of care to favor ADR over formal adjudication. On the other hand, other feminists caution that the ethic of care is part of legal justice rather than an alternative to it; conversely, care itself becomes corrupted without the virtues of consistency, integrity, and even-handedness—precisely the virtues promoted by legal justice in its commonsense conception (West 1997). (For further discussion on this point, see *Feminist Legal Theory*).

4. *Conclusion*

It is often unclear whether law and economics, pragmatism, and feminism are intended as alternatives *to* legal justice or as alternative analyses *of* legal justice. In a sense, all the theories surveyed here are responses to two fundamental tensions. The first is the tension between the rule of law, which subordinates particular cases to general rules, and the moral

demand to treat particular cases according to their individualized merits. The second is the tension between the values of consistency, predictability, and stability that legal justice serves and the more utopian values implicit in the broader notion of justice *simpliciter*. It may well be that no intellectual approach can eliminate this tension, or reconcile the competing senses of the word 'justice'

See also: Aristotle (384–322 BC); Common Sense, Anthropology of; Justice and its Many Faces: Cultural Concerns; Justice: Philosophical Aspects; Justice: Political; Justice: Social Psychological Perspectives; Pragmatism: Philosophical Aspects

Bibliography

Aristotle *Nicomachean Ethics*. (Many editions; 1998 Dove Publications, Mineola, NY)

Aristotle *Politics*. (Many editions)

Dworkin R M 1977 *Taking Rights Seriously*, rev. edn. Harvard University Press, Cambridge, MA

Dworkin R M 1980 Is wealth a value? *Journal of Legal Studies* **9**: 191–226

Finnis J 1980 *Natural Law and Natural Rights*. Clarendon Press, Oxford

Fuller L L 1969 *The Morality of Law*. Rev. edn. Yale University Press, New Haven, CT

Gagarin M 1986 *Early Greek Law*. University of California Press, Berkeley, CA

Gilligan C 1982 *In A Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA

Holmes O W 1929 Letter to John C H Wu, 1 July 1929. In: *Justice Holmes to Dr. Wu: An Intimate Correspondence 1921–32* [1947]. Central Books, New York

Kadish M R, Kadish S H 1973 *Discretion to Disobey: A Study of Lawful Departures from Legal Rules*. Stanford University Press, Stanford, CA

Kronman A T 1980 Wealth maximization as a normative principle. *Journal of Legal Studies* **9**: 227–42

Menkel-Meadow C 1985 Portia in a different voice: speculations on a women's lawyering process. *Berkeley Women's Law Journal* **1**: 39–63

Plato *The Laws*. (Many editions)

Plato *Statesman*. (Many editions)

Posner R A 1983 *The Economics of Justice*. Harvard University Press, Cambridge, MA

Posner R A 1998 Pragmatic adjudication. In: Dickstein M (ed.) *The Revival of Pragmatism: New Essays on Social Thought, Law, and Culture*. Duke University Press, Durham, NC

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Soper P 1984 *A Theory of Law*. Harvard University Press, Cambridge, MA

West R 1997 *Caring for Justice*. New York University Press, New York

D. J. Luban

Justice, Access to: Legal Representation of the Poor

History and political theory both underscore the fundamental importance of providing equal access to justice for all citizens no matter what their economic and social status. It was promised to the subjects of emperors and kings long before those subjects began to even dream of the possibility of governing themselves. Then democratic theory, especially the social contract theory so influential with the founders of democracies in Europe and North America, emphasized that no man would surrender his natural right to resolve disputes through force unless government provided a forum to resolve those disputes where all citizens, rich and poor, had an equal chance. Thus, it is no surprise that constitutions of nearly every democracy guarantee their citizens 'due process' or 'fair hearings' and 'equal protection of the laws' or 'equality before the law' (Johnson 2000).

Despite these solemn guarantees, however, for most democratic nations it was well into the twentieth century before governments became serious about offering their lower income citizens equal access to justice. Progress has been uneven, with some countries forging far ahead of others. As of the beginning of the twenty-first century, this has become a field of study for political scientists, sociologists, historians, and legal scholars—seeking to analyze the problems, to devise solutions, to evaluate experimental approaches, to comprehend the political opposition to this goal, and to measure the progress.

1. An Overview Of Access to Justice Strategies

The study of access to justice issues in modern societies begins with an appreciation of the essential role lawyers play in the regular civil courts in every industrial society. In these societies, courts place most of the burden on the parties to investigate the facts, find the applicable law and present the case to the neutral, passive decision-makers—the judges or juries—supplied by the government. This is a heavy, often expensive burden the private parties must discharge and one typically requiring the expertise of lawyers. As a result of this division of responsibility between the government-funded courts and the private litigants, the latter bear most of the cost of the dispute resolution function as carried out in the courts of most modern nations.

This article analyzes how modern nations seek to supply counsel to those unable to afford their own, and also how some nations have sought to restructure the dispute resolution function to reduce the need for counsel. Three basic strategies are covered: (a) pro-

viding government-funded lawyers to disputants who are unable to afford their own legal counsel (II); (b) providing incentives to encourage private lawyers to represent people unable to afford to pay legal fees but without using government funds (III); and (c) providing dispute resolution forums where disputants can participate fairly without the assistance of a lawyer (IV).

2. *Increasing Access to Justice by Providing Government-paid Lawyers to Disputants Unable to Afford Their Own Legal Counsel*

Empirical studies conducted in the USA support the hypothesis that litigants who lack lawyers are severely disadvantaged in the courts, at least as those courts operate at the beginning of the twenty-first century (Galanter 1974). Debtors sued in collection cases who were represented by counsel had a six times better chance of winning than those who defended themselves. A study of unrepresented litigants who brought suit in the Washington, DC courts showed they suffered dismissal at the early stages of the proceedings and not one succeeded in even getting to trial. Also, low-income tenants representing themselves before New York's housing courts suffered eviction in almost every case while 90 percent of those represented by lawyers were saved from that fate.

Most governments did not require empirical research to reach the conclusion that their lower income citizens required free counsel if they were to enjoy effective access to justice in the regular courts. But the level of commitment to equal justice for all citizens and the willingness to fund the lawyers necessary to that goal vary greatly from nation to nation. Nonetheless, there was a trend toward greater access backed by larger government investments during the second half of the twentieth century in nearly all industrial democracies.

However, the legal status of the right to counsel in the regular courts is far different in the USA than elsewhere in Western democracies. While nearly all European and Commonwealth nations have enacted a broad statutory right to counsel in civil cases, as well as criminal cases, neither the federal nor any state government has done so in the USA. England's first such statute was enacted in 1495, France in 1852, Germany in 1877, the rest of Northern Europe in the early twentieth century, Italy in 1923, and most Canadian provinces and Commonwealth countries in the 1960s and early 1970s. In these countries, statutes guarantee low-income people that the government will supply them with a lawyer in any case where they qualify financially for such assistance, although they typically require the litigant's position to have some merit as well. In the past, most countries did not compensate the lawyers appointed to represent needy litigants but expected them to representation on a pro bono basis (Cappelletti et al. 1975). But over the course of the twentieth century more and more of the industrial democracies began to pay the lawyers serving poor people with government funds until by the end of the century a majority of European nations and all of the Commonwealth countries have chosen to do so (Johnson 2000).

Europe and the USA also have gone along different routes in recognizing a 'constitutional' right to counsel for indigent litigants. The courts on both continents have declared a constitutional right to counsel in criminal cases. But despite interpreting nearly identical constitutional language, the US Supreme Court and the European Court on Human Rights have reached opposite conclusions on the right to counsel in civil cases. In 1979, the European Court ruled the 'fair hearing' guarantee found in the European Convention on Human Rights and Fundamental Freedoms required member governments to supply free counsel to indigent civil litigants (Airey vs. Ireland 1979). Two years later, however, in a five-four decision the US Supreme Court held the constitutional guarantee of 'due process' did not require the provision of free counsel to civil litigants in the USA (Lassiter vs. North Carolina 1981). Switzerland, although not part of the European Community, anticipated the decision of the European Court on Human Rights over 40 years earlier when in 1937 the Swiss Supreme Court held that nation's constitutional guarantee that 'All Swiss are equal before the law' required government to provide free lawyers to those Swiss citizens who were too poor to hire their own (Judgment of Oct. 18, 1937, O'Brien 1967).

Lacking either a constitutional or statutory guarantee, poor people in the USA and many other countries around the world have legal assistance with their civil problems only if they can get help from a limited number of lawyers supplied through private or legislative charity (Zemans 1979). Indeed private charity was the only source of legal counsel for the poor during most of US history. It took the 'War on Poverty' declared in 1965 to create even a minimal level of government funding of legal services for the poor in civil cases (Johnson 1978).

As might be expected, comparative research documents that a statutory or constitutional right to counsel usually leads to a much higher financial commitment to government-paid lawyers for lower income citizens. Table 1 lists several industrial democracies in Europe, North America, and Oceania. The table first compares their total and per capita investments in government-funded civil legal services for those unable to afford lawyers. But the more salient comparison takes account of the relative prosperity of these different countries and reflects how much of a nation's gross national product its government commits to insuring access to justice for its lower income citizens (as measured by legal services investment per $10,000 of per capita GNP). Finally, the last two

Table 1

Comparative civil legal services investments [in US$] [Nations ranked by relative share of GNP invested in publicly-funded civil legal services—lowest to highest]

Nation (or political subdivision of nation, e.g., province, state)	This nation's total government investment in civil legal services (in US dollars) [In US includes Federal, State, local governments, and IOLTA expenditures]	Government's per capita civil legal services investment (in US dollars)	Government's civil legal services investment per $10,000 of GNP (in US dollars)	Total US governmental investment in civil legal services if US invested as much of its GNP as this nation does on civil legal services	Total English governmental investment in civil legal services if England only spent as much of its GNP as this nation does on these services
United States (FY 1998)	$600 million [population = 270 million]	$2.25	$0.70 [= 70 cents]	$0.6 billion [e.g., $600 million]	$80 million
Germany (1996)	$390 million [population = 80 million]	$4.86	$1.90	$1.6 billion	$200 million
France (1994)	$270 million [population = 59 million]	$4.50	$1.90	$1.6 billion	$200 million
Australia (FY 1998–99)			[Each state has its own program]		
–New South Wales	$31 million [population = 6 million]	$5.12	$2.75	$2.3 billion	$320 million
Canada (FY 1998–99)			[Each province has its own program]		
–Quebec	$52 million [population = 7.3 million]	$7.07	$3.50	$3.0 billion	$400 million
–Ontario	$82 million [population = 11.5 million]	$7.06	$3.60	$3.0 billion	$400 million
–British Columbia	$32 million [population = 4 million]	$7.80	$4.00	$3.34 billion	$465 million
Netherlands (1998)	$150 million [population = 15.5 million]	$9.70	$4.20	$3.5 billion	$480 million
New Zealand (FY 1998–99)	$27 million [population = 3.8 million]	$7.10	$5.10	$4.25 billion	$560 million
England (1999)	$2 billion $1.35 billion [population = 53 million]	$39.00 $26.00	Gross = $17.00 Net = $12.00	$14.2 billion $10.1 billion	$1.35 billion

1.1 *Sources of Legal Aid Expenditure Data: Australia – Legal Aid Commission of New South Wales*, Annual Report '99, Financial Overview, Program Expenditure Chart. *Canada – Legal Aid in Canada: Resource and Caseload Data Tables, 1998–99, Table 5-Legal Aid Expenditures by Object (Statistics Canada-Catalogue No. 85F0028, Ottawa, 2000) Source*: Legal Aid Survey, Canadian Centre for Justice Statistics. *England – Lord Chancellor's Department, Judicial Statistics—England and Wales for the Year 1999 (CM4786, July 2000). Table 10.7*-Legal Aid Expenditures: Receipts and Payments, 1999. *France – Projected figures for 1994 reported in National Resource Center for Consumers of Legal Services*, France Beefs Up On Legal Aid, *Legal Plan Letter, Jan. 15, 1993; Germany – Gerhard Danneman*, Access to Justice: an Anglo-German *Comparison, 2 European Public Law 271 (1996) Table 5. Netherlands – Peter van den Biggelaar*, Legal Aid in the Netherlands, in "Legal Aid in the New Millennium" *(papers presented at the International Legal Aid Conference, University of British Columbia 16–19 June 1999) p. 74. (Statistics for 1998). New Zealand – Legal Aid Board, Annual Report for Year Ending June 30, 1999, p. 33. Appendix 1*-Overview of legal aid costs and recoveries. *United States – Legal Services Corporation*, Annual Report 1999. *Sources of GNP and per capita GNP data: World Bank, 2000 World Development Indicators and World Development Indicators database. Exchange Rates: Chart uses exchange rates in effect in year of legal aid expenditure, e.g., 1996 exchange rate for Germany and 1999 exchange rate for England, etc.*

1.2 *Note: An online version of this chart [updated periodically] can be found at http://www.equaljusticeupdate.org*

columns of the table illustrate what the total budget would be for two countries—England and the USA—were their investments the same per GNP as the other nations portrayed in this study. (These two nations are chosen because they represent the most generously funded (England) and least generously funded (USA) civil legal services programs included in these comparisons.)

Nations also differ in how they deliver their government-subsidized legal services. There are three common models: compensated private counsel, salaried staff attorney, and mixed systems.

Under the compensated private counsel model private lawyers represent lower-income people just as they would paying clients. But when serving those low-income clients the lawyers submit their bills to the government. The low-income clients, in turn, can usually choose any lawyer who has expressed a willingness to participate in the government program. This is the predominate delivery system in most countries, including England, the largest nations of continental Europe, and the Commonwealth countries.

Under the salaried staff model the government establishes separate legal aid offices, often in 'clinics' or 'neighborhood law offices' in close proximity to concentrations of low-income people. The offices are staffed with salaried lawyers, usually employed full-time. When poor people have legal problems they bring them to one of these offices and are represented by these salaried lawyers. The USA is the largest country that primarily relies on this delivery model.

A number of empirical studies have evaluated the costs and benefits of compensated private counsel vs. salaried staff delivery systems. Most comparisons suggest salaried staff programs deliver services at lower cost per case (Currie 2000). Both models score high on client satisfaction, but a substantial percentage also place a high value on the ability to choose their own lawyer—a feature of the compensated private counsel model. At the same time, salaried offices are better able to address community issues faced by the residents of low-income areas (Cappelletti et al. 1975).

Partly in response to the results of these studies and also in order to reduce program costs, several nations have moved to a mixed system that attempts to combine the virtues of the other two. Some of these mixed systems, such as those in Quebec, Canada and Sweden, adopt a 'client option' approach, allowing clients to choose between salaried staff lawyers and compensated private attorneys to represent them in a given case. Others, such as those in Ontario and British Columbia, divide legal problems between the two systems on a subject matter basis—typically with clients required to take all 'poverty law' cases to salaried attorney offices and the more traditional legal problems such as divorce to compensated private counsel.

Some countries using the compensated private counsel model to implement a right to counsel have experienced ever-expanding legal aid budgets. These budgetary considerations have led some governments, most notably in England, to begin exploring modifications calculated to reduce costs. These measures include 'franchising' where the government signs contracts with certain private lawyers and law firms to provide services at reduced fees. They also include variations of a 'capitation' approach where lawyers agree to handle all the legal problems generated by the entire low-income population in a defined geographic area for a fixed per capita sum (Moorhead 1998). British Columbia in Canada uses a form of 'managed care' common in some health delivery systems in which staff members closely monitor each litigation investment decision a compensated private lawyers might make during the course of a representing a poor person. In 1999 and 2000, England introduced several cost-control measures as part of a major overhaul of legal aid and the entire civil litigation process in that country (Regan et al. 1999).

3. Arrangements Which Encourage or Require Private Attorneys to Represent Lower Income People Without a Government Subsidy

In some nations, government-subsidized lawyers coexist with fee arrangements that encourage private attorneys to represent lower income clients without government funding in certain categories of cases. The most common of these are: (a) contingent fees, (b) conditional fees, and (c) court-awarded fees. In addition to these economic incentives, in many countries—especially those like the USA where government-funded legal services are in short supply—public policies encourage or require private lawyers to devote some of their time to the representation of poor people on a pro bono (uncompensated) basis (see Sect. 3.4).

3.1 Contingent Fees

The USA is almost unique in allowing lawyers to represent clients on a 'contingent fee' basis. Under fee contracts allowing them a percentage (typically 33 to 40 percent, but sometimes lower for early settlements and higher for trials) of the client's recovery, lawyers agree to assume the risk of failure, not charging any fee if they lose. Thousands of the poorest citizens are represented each year in cases where they are suing for a substantial sum of money and have a good chance of winning. At the same time, the contingent fee system offers no help to low-income people who are defendants, or who are plaintiffs in the many categories of legal disputes that do not involve monetary stakes or where the potential financial recovery is too modest to cover the risk of loss (Regan et al. 1999).

The contingent fee system is not without its critics. Some charge the fee percentages are too generous to lawyers, leaving their clients with less than full compensation for their injuries. Others, however, respond that if the percentages are set too low, lawyers will not find it economically feasible to take cases offering only modest potential recoveries or posing a significant risk of loss. Without the bonus of substantial fees when they win, contingent fee lawyers cannot afford to take riskier or smaller cases and thus can contribute less to increasing access for the low-income population (Regan et al. 1999).

Other studies have applied economic analysis to suggest the typical contingent fee structure places lawyers and their clients in a conflict situation. The lawyers maximize their profits when they get a decent recovery with a minimal investment of time while the client is best served by the largest possible recovery no matter how much time and effort the lawyer must expend. As a result, contingent fee lawyers may be inclined to settle too early and for too little, from the client's perspective. A recent empirical study, however, suggests contingent fee lawyers may come closer to maximizing their clients' returns than economic analysis predicted (Kritzer 1998a).

3.2 Conditional Fees

Most countries, other than the USA, use a 'loser pays' rule in allocating legal fees. That is, losing litigants pay not only their own lawyers but are required to pay at least a 'reasonable fee' to the winners' lawyers. A few nations, Ireland and at the beginning of the twenty-first century England, modify this rule in a way that may allow lawyers to serve low-income clients without a government subsidy in some cases.

In 1995, England introduced a modified form of the 'conditional fee' model, one that as of 2000 applies to nearly all civil litigation involving monetary stakes no matter the economic means of the parties. Under this model lawyers do not charge their clients a fee if they lose, but can charge an 'uplift' of up to 100 percent, in effect doubling their fee, if they win, with that 'uplift' payable out of the clients' monetary recovery. However, the total amount of the fee including the 'uplift' cannot exceed 25 percent of that recovery. Should they lose, clients are not relieved of their obligation to pay the other side's fees. But, if they can afford it, they can purchase insurance to cover this risk.

The UK government expects the 'conditional fee' system to supplant legal aid for the claimant's side of most cases offering the prospect of a monetary recovery. Critics point out the prospects of success must be unusually favorable before lawyers will be willing to provide representation under a 'conditional fee' system. They will not get a fee from either side if they lose and the 'reasonable fee' they do receive will ordinarily be less generous than the 33 to 50 percent of the judgment that contingent fee lawyers typically earn when they win at trial (Moorhead 2000).

3.3 Court-awarded Fees

Many jurisdictions in the USA depart from the general 'American' rule and shift fees to the losing party in certain categories of cases: (a) Statutory fee-shifting—in lawsuits under statutes specifically authorizing fee-shifting in favor of private parties whose cases enforce the public policies embodied in those statutes; (2) Public interest fee-shifting—in lawsuits where the court, rather than the legislature, determines that the public interest will be promoted by awarding winning litigants the bonus of requiring the loser to pay their legal fees.

Both statutory fee-shifting and public interest fee-shifting offer an alternative source of funding for those unable to afford their own counsel. They are especially important in major litigation not offering the possibility of a substantial monetary recovery, such as those seeking injunctive relief ordering the government or some other institutional defendant to change its policies in the future (Cappelletti et al. 1978–79). Where a case satisfies one of these standards, these fee awards can be more generous than either contingent or conditional fees, and frequently are. The court is typically not limited by the size of the winning party's economic recovery, which often is nothing, but can award winning lawyers their regular full hourly fees for the time they spent on the case. Beyond that, they can apply a 'multiplier,' doubling or tripling the lawyers' fee in order to recognize the riskiness, complexity, or significance of a particular piece of litigation (Regan et al. 1999).

3.4 Voluntary and Mandatory Pro Bono Legal Services

As mentioned earlier, most European countries originally implemented the statutory right to counsel by appointing lawyers to provide free legal representation to poor people without compensation—a form of mandatory pro bono services. During the twentieth century most of these nations found this approach unsatisfactory and shifted to a system that compensated the lawyers representing the poor (Huls 1994, Johnson 1994), although a few, like Italy, still depend on drafting lawyers to provide legal aid (Cappelletti et al. 1975).

Voluntary pro bono services depend entirely on the charitable impulses of the private legal profession. Faced with a shortage of government-funded legal aid the USA in the last decades of the twentieth century has attempted to encourage more private lawyers to volunteer time to represent the poor. Many states and cities have organized pro bono programs supervised

by staff. Others require private lawyers to report annually on the hours devoted to pro bono services. A national program issued a 'pro bono challenge' to the 500 largest law firms, asking them to commit 3 to 5 percent of the firm's billable hours to pro bono representation of the poor—and had mixed results. Some other nations—Sweden and Australia, among them—also have begun to look to pro bono services to fill some of the gaps created by reductions in government legal aid budgets (Regan et al. 1999).

There is serious question, however, whether societies can expect pro bono legal services to play more than a limited, supplementary role in the delivery of legal representation to low-income people. Is there any more reason to expect lawyers to supply enough services on a pro bono basis to meet the low-income population's total need for legal services than to expect medical doctors and hospitals to supply enough of their services pro bono to replace the present government-funded health care systems for the poor? —e.g., Medicare and Medicaid in the USA, the National Health Service in the UK etc.

4. *Increasing Access to Justice Without Supplying Free Lawyers, Government Paid or Otherwise*

Professional lawyers appear essential to fair proceedings in the regular courts of most modern countries, as the European Court on Human Rights held in Airey vs. Ireland (discussed in Sect. 2). The legal rules governing the decisions are complex, the procedures intricate, and investigatory techniques difficult to master. For all these reasons, unrepresented parties seldom have a chance for a 'fair hearing' in the regular courts.

Yet the European Court on Human Rights also recognized it may be possible for governments to create forums—for some cases at least—where the assistance of lawyers is not crucial and thus effective access to justice can be achieved without providing free counsel to lower income disputants (Airey vs. Ireland 1979). It is possible to identify four approaches that various countries have tried, alone or in combination, in an attempt to de-professionalize dispute resolution: (a) simplified forums; (b) law simplification; (c) technological substitutes for legal expertise; and (d) the use of less expensive, non-lawyer advocates.

4.1 *Simplified Forums*

For simpler disputes it may be feasible to construct simpler dispute resolution processes. Early in the twentieth century, several countries started 'small claims courts' to hear disputes involving minor stakes. Using simple procedures and expecting the judges to take a more active role in developing the facts, these courts were designed to function well without lawyers. Some criticize small claims courts for pitting poor people against lawyers hired by institutions such as credit companies, landlords, and businesses. A few jurisdictions responded by banning lawyers entirely from these courts. Studies of small claims courts have arrived at mixed findings about their fairness to individual litigants, especially in cases involving institutional opponents (Yngvesson et al. 1979).

Some jurisdictions also have created specialized courts roughly based on the small claims court model to handle certain narrow categories of cases (housing courts, labor courts, social welfare courts, and the like) with the expectation that their simplified processes would allow many disputants to represent themselves. Once again, the results are mixed (Cappelletti et al. 1978–79).

In addition to simplified courts within the court system itself, some jurisdictions have created administrative bodies empowered to resolve certain categories of disputes. For example, in England, 'tribunals' composed of part-time volunteers decide most disputes arising out of public housing and other government benefit programs. In British Columbia, Canada, an agency called the 'Rentalsman' determines most landlord–tenant disputes. In many jurisdictions administrative hearings decide whether a worker is entitled to compensation for job-related injuries and how much. Once again, the procedures are simple compared to the regular courts and typically disputants are encouraged to appear without lawyers—and sometimes lawyers are barred entirely (Cappelletti et al. 1978–79).

In recent years, private mediation and arbitration (often called Alternative Dispute Resolution or simply 'ADR') have emerged as major alternatives to the regular civil courts, especially in the USA. The primary motive behind the growth of these alternative forums, however, arises not from a desire to increase access to justice for lower income individuals, but to reduce delay and expense for more affluent litigants (Abel 1982).

While private mediation and arbitration have not proved a significant strategy for increasing access to justice when the disputes pit individuals against institutions, they have offered a useful alternative to litigation in disputes between evenly matched individuals (Menkel-Meadow 1999). Some communities in the USA have started neighborhood (or community) dispute resolution centers to help people settle disputes—especially those between family members or neighbors—without resorting to the courts. The services are usually free or require only nominal fees (Singer 1979).

4.2 *Simplification of Substantive Legal Rules Governing Disputes*

One of the major reasons people must employ lawyers to represent them is the complexity of the legal rules

the judge or jury will apply to decide who wins. Some suggest we can reduce the need for lawyers and thus improve access for lower income disputants by simplifying those rules. In some jurisdictions, for example, legislatures enacted no fault divorce laws and authorized couples to obtain 'summary divorces' if they had no children, limited property, and short-term marriages. Presumably, these statutory changes allow most people who qualify for such divorces to do so without a lawyer (Cappelletti et al. 1978–79). In an ever more complex society, however, it is questionable whether simplification of the legal rules will emerge as an important strategy for increasing access to justice in any but very limited circumstances.

4.3 Technological Substitutes for Legal Expertise

Starting in the late 1980s software developers produced a host of computer programs designed to offer average citizens the ability to generate legal documents such as wills and contracts. In the 1990s court systems and others desiring to increase access to justice for lower income individuals began developing programs which use similar technology to generate litigation documents that can be filed in common cases such as landlord–tenant, child custody, and spousal abuse cases. As of 2001, however, there has been no assessment of the issue whether these computer-generated litigation documents help litigants if they lack legal representation when they appear in court.

However, technology, computer technology in particular, may have only begun to transform the dispute resolution process in ways which could improve access for lower income litigants. Conversely, technology may reduce access to justice by increasing still further the cost of the litigation process and compounding the advantages higher income parties already enjoy. In fact, this phenomenon so evident in the health care field, can also be observed in the law. Only the party with considerable resources can afford many of the expensive scientific tests and fancy computer-generated graphics which enhance the chances of victory in the modern court room.

4.4 Use of Less Expensive, Non-lawyer Representatives

Some specialized lower-cost forums permit non-lawyers to represent parties appearing before them. Some even rely exclusively on non-lawyer advocates for this role. Presumably the use of these less educated and less expensive personnel enhances access by lowering the cost of representation to the parties or to the government employing the advocates.

A recent study tested the difference between representation by lawyers and representation by lay advocates in these administrative forums. The study suggests only marginal differences in the quality of representation or the results achieved. It says little if anything, however, about the comparative performance in the regular courts or other forums that decide a broader range of disputes requiring advocates possessing a greater body of legal knowledge and skills (Kritzer 1998b).

5. Future Issues in the Movement Toward Equal Access to Justice

The major issues for government-funded legal aid and equal access to justice in the twenty-first century include: (a) whether a 'globalization of constitutional values' will lead courts in the USA and elsewhere to join the majority of industrial democracies in recognizing a constitutional right to counsel in civil cases; (b) whether those nations which already have such a right as a matter of statute or constitutional interpretation will develop delivery mechanisms and alternative dispute resolution policies allowing them to guarantee effective access for lower income citizens at politically sustainable levels of public investment; and (c) whether the vital importance of equal access to justice for democratic government and political stability will cause developing countries and former communist nations to institute significant legal aid programs. As it has in the past, social science will have much to contribute not only to the design of these initiatives, but more importantly to their evaluation. The insistence on empirical proof will offer the most effective guarantee that what emerges indeed grants every citizen, irrespective of means, the right to equal justice that democratic societies promise to all.

See also: Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Inequality; Inequality: Comparative Aspects; Justice and its Many Faces: Cultural Concerns; Justice and Law; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Social Justice

Bibliography

Abel R (ed.) 1982 The politics of informal justice: the American experience, Vol. 1. *Comparative Studies* (Vol. 2). Academic Press, New York

Airey v. Ireland. Eur. Court H.R. (Ser. A) Judgment of Oct. 1979

Cappelletti M, Gordley J, Johnson E Jr 1975, 1981 *Toward Equal Justice: A Comparative Study of Legal Aid in Modern Societies*. Oceana, New York

Cappelletti M, Bryant G (eds.) 1978–79 *Access to Justice*. 3 Vols. Giuffre, Milan

Currie A 2000 Legal aid delivery models in Canada: Past experience and future developments. *University of British Columbia Law Review* **33**: 285

Galanter M 1974 Why the 'haves' come out ahead. *Law and Society Review* **9**: 95

Huls N 1994 From pro deo practice to a subsidized welfare state provision: Twenty-five years of providing legal services to the poor in the Netherlands. *Maryland Journal of Contemporary Legal Issues* **4**: 326
Johnson E Jr 1978 *Justice and Reform: The Formative Years of the American Legal Services Program*. Transaction Press, New Brunswick, NJ
Johnson E Jr 1994 Nordic legal aid. *Maryland Journal of Contemporary Legal Issues* **4**: 301
Johnson E Jr 2000 Equal access to justice: Comparing the United States and other industrial democracies. *Fordham International Law Journal* **24**: 83
Kritzer H 1998a Contingent-fee lawyers and their clients: Settlement expectations and settlement realities. *Law and Social Inquiry* **23**: 795
Kritzer H 1998b *Legal Advocacy: Lawyers and Nonlawyers at Work*. University of Michigan, Ann Arbor, MI
Lassiter v. North Carolina, 452 US 18 (1981)
Lawrence S 1990 *The Poor in Court: The Legal Services Program and Supreme Court Decision-making*. Princeton, Lawrenceville, NJ
Menkel-Meadow C 1999 Do the 'haves' come out ahead in alternative judicial systems? 'Repeat players' in ADR. *Ohio State Journal on Dispute Resolution* **15**: 19
Moorhead R 1998 Legal aid in the eye of the storm: Rationing, contracting and a new institutionalism. *Journal of Law and Society* **25**: 365
Moorhead R 2000 Conditional fee agreements, legal aid, and access to justice. *University of British Columbia Law Review* **23**: 471
O'Brien W 1967 Why not appoint counsel in civil cases: The Swiss approach. *Ohio State Law Review* **28**: 1
Regan F, Paterson A, Goriely T, Fleming D (eds.) 1999 *The Transformation of Legal Aid: Comparative and Historical Studies*. Oxford University Press, Oxford, UK
Schefergegen Appenzell A.R.H. Regierungsrat, *Judgment of Oct. 8, 1937*, BGE 63 Ia 209 (1937)
Singer L 1979 Nonjudicial dispute resolution mechanisms: The effects on justice for the poor. *Clearinghouse Review* **13**: 569
Yngvesson B, Hennessey P 1975 Small claims, complex disputes: A review of the small claims literature. *Law and Society Review* **9**: 219
Zemans F 1979 *Perspectives on Legal Aid*. Frances Pinter, London

E. Johnson

Justice: Philosophical Aspects

Justice is a central moral notion, associated with fair and impartial decision procedures (procedural justice) as well as with persons and groups being treated even-handedly (formal justice, treating like cases alike) and in a morally fitting way (material or substantive justice). The demands of justice are frequently presented as uncompromising (strictly required) and contrasted with the more flexible—and sometimes conflicting—demands of mercy, charity, humanity, beneficence, and equity, which may be merely supererogatory.

Justice is predicated mainly of three kinds of iudicanda, namely subjects (persons and collectivities), the conduct of subjects (actions and omissions: decisions, accusations, policies, wars, etc.), and social rules (e.g., social institutions, laws, conventions, practices, economic arrangements, educational and health-care systems). In recent decades, discussion has focused on the justice of social rules (i.e., on social justice).

Justice is often differentiated by theme. Thus distributive justice governs the assignment of benefits and burdens (e.g., rights, duties, property, income, taxes), commutative justice governs economic exchanges, while corrective and retributive justice govern, respectively, compensating the victims and punishing the perpetrators of prior wrongdoing. International justice is typically viewed as a separate area, in which the four mentioned themes are overshadowed by reflections on the just use of force. Just war theories are generally organized around two themes: just cause in going to war (*ius ad bellum*) and just ways of fighting a war (*ius in bello*). Kant adds a third theme: the just way of concluding a war so as to lay the foundation for a stable peace (*ius post bellum*; Kant 1996, § 58).

The manifest complexity of the notion of justice makes it doubtful that any unified account of justice can be adequate across iudicanda and themes. In response to such doubts, it is tempting, as numerous historical examples attest, to simplify matters through reductionist moves. One such move, prominent among deontological conceptions of ethics, is to reduce judgments of character to judgments of conduct, e.g., by holding that a just subject is one stably disposed toward just conduct, independently defined. (One may object that the justice of subjects is in part a function also of their inner states: of their feelings and attitudes in the case of individuals and of their internal mode of organization in the case of organized collectivities. Thus, the city-state of Athens may be called unjust not merely on account of its unjust conduct toward Socrates and the Melians, but also on account of the injustice of its social rules, which severely disadvantaged women, metics, and slaves.) Following Aristotle, some conceptions of ethics have proposed the reverse: to reduce judgments about conduct to judgments about the character, or virtues, of subjects. Just conduct is then conduct a just agent might engage in. (One difficulty here is that, to avoid circularity, the virtue of justice must not be characterized in conduct terms.) Locke and Nozick have proposed a third kind of reduction: that social rules are just if and only if they were, or at least could have been, established through just conduct. Conversely, Rawls has suggested that persons and conduct are just when they support and follow just social rules that apply to them and help establish just social rules insofar as such are still lacking (Rawls 1971, pp. 115, 334).

Other reductionist attempts seek to reduce judgments about various kinds of iudicanda to judgments about states of affairs. One simple consequentialist reduction of this sort holds that subjects, conduct, and social rules are just in proportion to their tendency to promote the overall goodness of the world (or of that part of the world to which they have a special responsibility). A more complex such reduction, of judgments about social institutions to judgments about distributions (Rawls), is discussed below.

While it is quite possible to embed some such reductions within a theory of justice, their plausibility still depends on the overall plausibility of this theory, which in turn depends on whether the evaluative and normative judgments derivable from it are morally sound. So far no detailed theory of justice, reductionist or otherwise, has managed to cover several themes and kinds of iudicanda in a plausible way. And contemporary writers have therefore tried, more modestly, to develop limited conceptions of justice. The paradigmatic effort in this vein (Rawls 1971), is focused narrowly on how distributive justice might be instantiated in the most important and pervasive rules ('basic structure') of a self-contained society.

1. *The Theory of John Rawls*

A criterion of justice should entail plausible absolute and comparative judgments about feasible basic structures, actual or proposed. Rawls seeks a criterion that, in addition, facilitates agreement among citizens accepting it on how to maintain or reform their shared basic structure. This second desideratum speaks against any criterion that, like the principle of utility, is too difficult to apply and hence too easily abused. Rawls seeks the morally best *public* criterion of social justice.

His central idea is to answer this moral question through a prudential one: morally best is that public criterion of justice which prospective citizens would choose on prudential grounds. The criterion and the basic structures based on it should be as good as possible for the persons living under them; nothing else need be considered. This central thesis is, of course, itself a moral claim—to be developed and examined in what follows.

Citizen's interests and preferences vary and are shaped by the actual basic structure of their society. For his prudential question to have a determinate answer, Rawls needs therefore a vantage point that attains some independence from the status quo and also ensures agreement. This vantage point is the original position—an imagined deliberation among 'parties' who, acting in the interest of prospective citizens, are to agree on a public criterion of justice without knowing (due to a 'veil of ignorance') the distinguishing features (creeds, values, tastes, desires, and endowments) of their clients or even the natural and historical context of their clients' society.

To have grounds for deliberation, the parties need some information about the interests of their clients. Rawls makes them assume that all clients have three interests, which he portrays as closely connected to their role as citizens in a democratic society (and hence as not biased toward any particular religious, philosophical, or ethical world-view or way of life). Rawls calls these the three higher-order interests, suggesting both that they are interests in the content and fulfillment of other interests (like second-order desires are desires about desires) and also that they are deep, stable, and normally decisive. These interests are (a) to develop and exercise a sense of justice, which involves a sincere allegiance to some conception of justice (whose content remains unspecified so as not to prejudge the parties' decision), (b) to develop and exercise the capacity to form, to pursue, and rationally to revise a conception of the good, and (c) to protect and advance one's particular conception of the good (Rawls 1993, pp. 19, 74). The main question in the original position is then: which social order—consisting of a public criterion of justice and of the basic structures this criterion would actually bring forth in various alternative contexts—affords the best prospects for higher-order interest fulfillment?

Rawls makes the dramatic claim that the parties would, because of their severely limited information, make their choice pursuant to the maximin rule, seeking to optimize the worst social position that the public acceptance of their criterion might generate. They would examine candidate criteria of justice, asking of each how badly off, in terms of the three higher-order interests, persons might come to be in societies in which it is the accepted public criterion of justice. And they would prefer the criterion that promises the higher floor (*max*imum *min*imorum), even if it would also lead to a much lower ceiling or lower average.

Rawls argues that the parties would then choose a criterion of justice formulated in terms of certain social primary goods, such as rights and liberties, income and wealth. Access to these all-purpose means is not only a good proxy for, but also much easier to measure than, higher-order interest fulfillment, which enhances the criterion's capacity to foster agreement on institutional design. Rawls claims then that the parties would choose these two principles as the public criterion (Rawls 1993, pp. 5f):

> *First principle*. 'Each person has an equal claim to a fully adequate scheme of equal basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value.'
> *Second principle*. 'Social and economic inequalities are to satisfy two conditions: first, they are to be attached to positions and offices open to all under conditions of fair

equality of opportunity [the opportunity principle]; and second, they are to be to the greatest benefit of the least advantaged members of society [the difference principle].'

The first principle is to have lexical priority over the second and, within the second, the opportunity principle is to have lexical priority over the difference principle (Rawls 1971, pp. 302f). 'Lexical' is short for 'lexicographical' and thus invokes how words are arranged alphabetically, giving priority to the first letter over the second, to the second over the third, and so on. Rawls conceives priorities among principles accordingly: in the comparison of alternative feasible basic structures, a superiority in terms of a lexically prior principle overrides any inferiority in terms of lexically subordinate ones.

The proposed public criterion of justice requires, then, that a society's basic structure be arranged so as to secure, above all, the basic liberties. These include the political liberties, freedom of thought and liberty of conscience, freedom of association, freedom and integrity of the person (including freedom of movement and the freedom to hold personal property), and the rights and freedoms included in the rule of law. Except in the case of the political liberties, the first principle does not require that citizens have sufficient means to take full advantage of their basic liberties; the distribution of such means (e.g., income and wealth) is governed by the second principle. Yet it does require more than legal recognition: citizens must have secure access to their freedoms; attempts to disturb their enjoyment or exercise must be effectively prevented and deterred.

The second principle addresses the social and economic aspects of basic structures. The opportunity principle requires that persons have roughly equal chances of success irrespective of their initial (parental) socioeconomic position. The difference principle requires that rules affecting socioeconomic inequalities (in particular, the tax system) be designed so as to optimize the worst socioeconomic position. Existing poverty is justified if, and only if, no institutional reform decreasing or increasing inequality would raise the lowest socioeconomic position.

2. *The Bias Toward Recipient-oriented Theorizing*

While Rawls's theory has received much criticism, it has also in the last 30 years shaped the debate about justice and, most importantly, the prevailing understanding of the concept of justice. There is little awareness now of the distinctive features of, or alternatives to, the specific concept of justice employed by Rawls and his many critics and defenders.

While the word 'just' can, and often does, function as a one-place predicate, the concept of justice (like the concept of mother) essentially involves a second place: to be unjust is to be unjust to someone; A is just or unjust to B, to certain recipients.

The idea of recipients, of those who have justice claims on a particular iudicandum, invokes the questions what their claims on the iudicandum are, what these recipients have and do not have, and what they ought to have or not to have. The concept of justice essentially involves then also a third place for relevant goods and ills, broadly conceived: A is just or unjust to B by affecting whether and to what extent B has C. (As the category of conduct includes omissions, so the category of goods and ills includes the null-case of receiving nothing, being denied a good or being spared an ill.) For B to suffer an injustice, there must be goods or ills that B either has-but-should-not-have or lacks-but-should-have.

The concept of justice as explicated thus far—fundamentally a three-place concept, with frequent but derivative one-place and two-place employments—frames current discussions of justice. According to it, justice is a feature of certain iudicanda, which depends on their effects on the distribution of certain goods and ills to or among their recipients. One might call this the passive concept of justice, because it directs moral attention to those by whom justice and injustice are suffered. Employing this passive concept, one will focus on a iudicandum's effects on its recipients—specifically on its impact on what these recipients are able to have or to be (relative to what they ought to be able to have or to be). One will consider a iudicandum just in proportion to its impact on the moral quality of the distribution of relevant goods and ills among its recipients: To be just is to promote good distributions. The passive concept thus leads easily to a conception of justice as distributional; and much of the current justice literature is indeed about how to code and rank distributions.

This framing of the debate may well be a remnant of long-dominant consequentialist, especially utilitarian, theorizing (Bentham, Mill, Sidgwick), which Rawls has superficially done so much to displace. The central consequentialist idea is that subjects, conduct, or social rules should be assessed solely by their effects on how well the world goes. Most current theorizing about justice involves the closely related thought that the justice of subjects, conduct, or social rules should be assessed by their effects on the moral quality of the overall distribution of goods and ills among the iudicandum's recipients. In both cases, the interest in outcomes is indirect. How well the world goes is of little practical interest in itself, but can be of great practical interest if one judges the moral quality of subjects, conduct, or social rules by their impact on how well the world goes. Similarly, the moral quality of overall distributions is of little practical interest in itself, but can be of great practical interest if one judges the justice of subjects, conduct, or social rules by their impact on the moral quality of distributions.

Does it make sense to take distributions as the most basic iudicanda and then to judge the justice of subjects, conduct, or social rules by their distributional effects? With respect to subjects and their conduct, this mode of assessment is widely rejected. Among the timeworn staples of ethics classes are gruesome organ redistribution stories designed to elicit the response that it is not true that subjects and conduct that aim at or produce the best overall distributions are always just. Abstractly considered, a situation in which everyone has at least one eye and one kidney is surely morally better than (an otherwise similar) one in which some, through no fault of their own, have no functioning eye or kidney while many others have two. But redistributive conduct and subjects aiming at such an abstractly better distribution are nevertheless considered gravely unjust.

With respect to subjects and their conduct, cases of this sort may be used to draw the conclusion that we ought to distinguish between treating recipients justly and promoting a good distribution among recipients and, more generally, between how subjects treat their recipients and what impact they have on them. The moral significance of this distinction can be illustrated with matched doing/allowing scenarios such as the following: in Scenario One, Art has two functioning kidneys and Sue has none; we must decide whether to give one of Art's kidneys to Sue. In Scenario Two, Art has one functioning kidney and Sue has none; we must decide whether to give a donated kidney to Art or to Sue. (Other things are presumed equal across scenarios, with both Art and Sue urgently wanting the kidney in question.) In terms of their distributional effects, the two decisions are on a par: in each scenario, we decide whether the final distribution of kidneys will be 2:0 or 1:1; and a decision in Sue's favor is then in both scenarios equally promoting of a good distribution. Intuitively, however, the scenarios differ dramatically: a decision for Sue treats Art unjustly in One but not in Two and a decision for Art treats Sue unjustly in Two but not in One.

With respect to social rules, a similar distinction seems called for, and for similar reasons. Just social rules for allocating donated kidneys favor those who, through no fault of their own, have no functioning kidney over those who have one, and thereby promote a better distribution of kidneys. But just social rules do not mandate the forced redistribution of kidneys from those who have two to those who have none, though doing so would likewise promote a better distribution of kidneys. Once again, treating recipients justly does not boil down to promoting the best distribution—what matters is how social rules treat their recipients, not merely what their effects on them are.

This distinction has been remarkably neglected in current work on social justice. It is not surprising that it plays no role in consequentialist theorizing, which is committed to the view that social rules should be judged by their effects on the overall outcome, irrespective of how they produce these effects. Remarkable is that Rawls' theory, which he bills as a deontological alternative, likewise makes the justice of social rules depend exclusively on the overall distribution they tend to produce. Rawls assigns no moral relevance to *how* basic structures affect the overall distribution of goods and ills. And he is followed in this by most of his critics.

At least this decision is clearly built into the centerpiece of Rawls's theory: 'the persons in the original position try to acknowledge principles which advance their systems of ends as far as possible. They do this by attempting to win for themselves the highest index of primary social goods, since this enables them to promote their conception of the good most effectively whatever it turns out to be. The parties … strive for as high an absolute score as possible' (Rawls 1971, p. 144). By judging social orders in this purely recipient-oriented way, Rawls ensures from the start that they are judged exclusively by their 'output' in terms of what overall distributions they generate among their participants—without regard to how they may bring it about that particular social goods or ills end up with particular persons. While Rawls differs from utilitarians in regard to how distributions should be coded and assessed—fulfillment of higher-order interests rather than happiness, maximin rather than sum-ranking or averaging—he shares with them the purely recipient-oriented mode of theorizing about justice.

Consistently followed, this approach should lead Rawls to embrace strong compensation requirements with regard to natural handicaps. It is evident that the extent to which a person's higher-order interests (especially the third) are fulfilled depends not merely on her share of social goods but also on her natural endowments. Someone genetically predisposed toward good health, cheerfulness, intelligence, and good looks is better able to advance her conception of the good than someone genetically predisposed toward sickliness, melancholy, low intelligence, and ugliness. How, if at all, would the parties take account of this fact? Even leaving aside genetic modifications, the parties can still raise the floor of higher-order interest fulfillment by formulating a criterion of justice that takes account of the natural distribution—by requiring basic structures to favor the naturally disfavored in the distribution of social goods and ills insofar as doing so raises the floor (by ensuring that no one is very badly off in terms of both natural and social goods).

These considerations notwithstanding, Rawls claims that his parties would adopt a criterion of justice that assesses each basic structure solely on the basis of the distribution of social primary goods it produces: 'a hypothetical initial arrangement in which all the social primary goods are equally distributed … provides a benchmark for judging improvements' (Rawls 1971, p. 62). The two principles do not take into

account natural primary goods, such as 'health and vigor, intelligence and imagination' (Rawls 1971, p. 62): 'The natural distribution is neither just nor unjust' (Rawls 1971, p. 102).

How might one react to Rawls' apparent inconsistency? First, one might reject and revise his exclusive emphasis on the distribution of social goods and ills. In this vein Amartya Sen has argued that, since natural differences among persons evidently affect their quality of life, we should focus on the distribution not of social primary goods but of capabilities. How badly off someone is depends on her access to being well-nourished, adequately clothed and sheltered, etc., which, in turn, depends not merely on her income, but also on her physiology as well as on the climate and culture she confronts. To this critique, Rawls has responded that he had, for purposes of a first approximation, chosen to leave severe natural handicaps and disabilities aside and that the remaining natural disadvantages—Sen had mentioned differential caloric and textile needs—are so minor that their accommodation within a criterion of justice is not worth the cost in terms of increased complexity and consequent divisiveness. But this response is unconvincing: even ordinary natural handicaps—such as ugliness or a tendency toward melancholy and depression—can be quite devastating. And even if severe natural handicaps of mind and body can be ignored in a first approximation, they cannot be left aside in the final version of a theory that is to serve as a public standard by reference to which citizens shape and reform their society's basic structure.

Second, one might defend Rawls by arguing that natural inequalities are of little importance to a person's true quality of life: her ability to fulfill her higher-order interests. This response is also unpromising. Innate mental disabilities, resulting in low intelligence or an inability to concentrate, can surely undermine a person's capacity to understand and apply a conception of justice as well as her capacity to form, to revise, and rationally to pursue a conception of the good. Such disabilities, and genetically based ugliness and melancholy as well, surely also affect profoundly a person's success in advancing her particular conception of the good.

Following Rawls' proposed criterion of justice at the expense of his purely recipient-oriented machinery for deriving this criterion, one might thirdly argue that the distribution of natural goods and ills, however important to persons, is irrelevant to the assessment of basic structures, which should be based exclusively on the distribution of social goods and ills each generates: the task of social rules is to do justice to, or to treat justly, all those whose lives are regulated by this order. This requires a just allocation of the benefits and burdens of social cooperation, not promotion of the best attainable distribution of all (social and natural) goods and ills that may be relevant to higher-order interest fulfillment or quality of life. Natural inequalities precede social rules, so to speak, as persons already have their genetic endowments and handicaps when they enter society. A just basic structure should not then make the naturally favored subsidize the naturally disfavored any more than a just parent or administrator or civil judge should seek to allocate benefits and burdens under her control so as to even out natural inequalities. This view faces the difficulty of drawing precisely enough the distinction between social and natural inequalities, to which it gives so much moral significance. It is not clear how one can, without considerable arbitrariness, measure the extent to which a given disease or inequality in earning power is due to genetic and to social factors. Taken strictly, this view would moreover lead to the controversial conclusion that there is nothing unjust about a society that has a just economic system but then leaves the allocation of medical care (as well as of the means of mitigating melancholy, homeliness, low native intelligence, etc.) entirely to the market.

3. *Toward a More Adequate Theory*

A fourth option is to develop a conception of social justice that neither (like the two principles) ignores natural goods and ills nor (as the original position would suggest) takes account of them on a par with social goods and ills. The key idea is that a criterion of social justice must be sensitive not merely to the intrinsic features of goods and ills (such as their kind, magnitude, and frequency), which make them relevant to the recipients' quality of life, but also to an extrinsic feature, namely to the way in which their distribution is affected by social rules. Such a criterion must take account of—and its application thus requires information about—the particular relation between social rules and human quality of life, which may affect whether some institutionally avoidable shortfall is an injustice at all and, if so, how great an injustice it is.

The intuitive moral significance of this relational feature can be shown by distinguishing, in a preliminary way, six ways in which a social order may affect the goods and ills of its recipients. The following illustration uses six different scenarios, arranged in order of their intuitive moral significance, in which, due to social rules, certain innocent persons are avoidably deprived of some vital nutrients V (the vitamins contained in fresh fruit, say). First-class shortfalls are officially mandated, paradigmatically by the law (legal restrictions prevent certain persons from buying foodstuffs containing V). Second-class shortfalls arise from legally authorized conduct of private subjects (sellers of foodstuffs containing V lawfully refuse to sell to certain persons). Third-class shortfalls are foreseeably engendered through the uncoordinated conduct of subjects under rules that do not specifically mention them (certain persons, suffering severe poverty within an ill-conceived economic order, cannot afford to buy foodstuffs containing V). Fourth-class

shortfalls arise from private conduct that is legally prohibited but generally tolerated (sellers of foodstuffs containing V illegally refuse to sell to certain persons, but enforcement is lax and penalties are mild). Fifth-class shortfalls arise from natural factors whose effects social rules avoidably leave unmitigated (certain persons are unable to metabolize V due to a treatable genetic defect but are not receiving the treatment that would correct their handicap). Sixth-class shortfalls, finally, arise from self-caused factors whose effects social rules avoidably leave unmitigated (certain persons are unable to metabolize V due to a treatable self-caused disease—brought on, perhaps, by their maintaining a long-term smoking habit in full knowledge of the medical dangers associated therewith—and are not receiving the treatment that would correct their ailment).

Even without all necessary qualifications and refinements, this list makes clear that it is highly relevant to the justice of social rules *how* they affect the distribution of goods and ills, that it is no more true of social rules than of subjects and conduct that they are just if and insofar as they promote a good overall distribution. How can the moral significance of this relational feature be explained? Social rules are not merely imposed on people (the 'recipients'), but also imposed by people. Purely recipient-oriented theorizing ignores these imposers of social rules; and the passive concept of justice has no place for them. Once the assessment of social rules takes account of both recipients and imposers, the sought explanation is at hand: social rules that cause harms by mandating them are more unjust than social rules that engender (equally frequent and severe) harms through overly mild penalties and lax enforcement, because they reflect a different attitude on the part of their imposers. The first injustice is close to the negative-duty end of the spectrum (bringing harm about), while the second injustice is closer to the positive-duty end (letting harm happen).

This hypothesis also explains why justice does not require social rules that mandate beatings whenever the total number of beatings is thereby reduced—and why it may (*pace* Rawls 1971, Chap. 4) oppose official restrictions of basic liberties which, by making other such liberties more secure, enhance citizens' freedom overall. Rawls must count official and private threats to basic liberties on a par: from the prudential perspective of prospective citizens, it is no worse for one's basic liberties to be threatened by laws or government agents than for them to be equally threatened by zealots, criminals, or drunk drivers (cp. Rawls 1971, p. 242). Here Rawls collides with the widely held belief that official violations are morally more significant, so that, for example, a society should not allow its government to restrict the basic liberties of (some) citizens whenever doing so will prevent greater threats to basic liberties by private actors. 'Reforms' that weaken constraints on policing or introduce lower standards of evidence or disproportionate punishments (e.g., the death penalty for recidivist drunk drivers) are not thought to accord with justice merely because risks to citizens are reduced on balance.

This diagnosis suggests that a genuine alternative to purely recipient-oriented theorizing can be facilitated by developing a more active concept of justice, one that includes a fourth essential place for (what one might call) the agents of justice, for those who have or share moral responsibility for the justice or injustice of the iudicandum. Pursuant to this alternative concept, the injustice of any iudicandum presupposes some agent(s), D, responsible for its injustice or for making it (more) just. Calling this the active concept of justice alludes to its diverting some attention from those who suffer justice and injustice to those who produce them.

While the passive concept easily leads to a conception of justice as distributional, the active concept suggests a conception of justice as relational, as involving a relation between particular agents and recipients mediated through a iudicandum. By drawing attention to what we owe one another (see Scanlon 1999), to each agent's specific responsibilities for particular iudicanda and, through them, to their particular recipients, the active concept leads away from the idea that a concern for the justice of a iudicandum is simply a concern for its promoting a good distribution among its recipients. Rather than tie justice to an outcome and thus view it as something that obtains to a greater or lesser extent (what would justice be like, what ought recipients to be able to have or to be?), the active concept presents justice as something that is done to a greater or lesser extent: by particular agents, through a iudicandum, to particular recipients.

To conclude, unjust subjects, conduct, and social rules involve victims unjustly harmed by them. But it does not follow that the magnitude of the injustice is a function of the magnitude of these harms. It may also depend on how these harms are produced. The active concept of justice invites sensitivity to such relational factors. In regard to social rules, such sensitivity can take the form of a scalar version of the familiar distinction between negative and positive duties. Such a scale is needed to weight the harms rules produce by the way in which they produce them (mandate, authorize, engender, …). A similar such scale is needed to gauge the degree to which different agents contribute to imposing unjust rules. Moral responsibility for the justice of some iudicandum may vary from agent to agent. So there may be competition among iudicanda—one may have responsibilities with regard to achieving and maintaining the justice of several iudicanda and one may then have to decide how much of an effort one ought to make with regard to each. These issues concerning responsibilities and their prioritization are crucial for giving justice a determinate role in the real world. And they tend to be

overlooked from the start, or grossly oversimplified through a generic duty to promote just iudicanda, when the topic is approached in terms of the concept of passive justice.

Bibliography

Ackerman B 1980 *Social Justice and the Liberal State*. Yale University Press, New Haven, CT

Aristotle 1985 [ca. 330 BCE?] *Nicomachean Ethics* [trans. Irwin T]. Hackett, Indianapolis, IN

Barry B 1989 *Theories of Justice*. University of California Press, Berkeley, CA

Bentham J 1982 [1789] In: Burns J H, Hart H L A (eds.) *An Introduction to the Principles of Morals and Legislation*. Methuen, London

Cohen G A 1997 Where the action is: On the site of distributive justice. *Philosophy and Public Affairs* **26**: 3–30

Dworkin R 1981 What is equality? *Philosophy and Public Affairs* **10**: 185–246, 283–345

Gauthier D 1987 *Morals by Agreement*. Cambridge University Press, Cambridge, UK

Hume D 1983 [1751] In: Schneewind J B (ed.) *An Enquiry Concerning the Principles of Morals*. Hackett, Indianapolis, IN

Kant I 1996 [1797] *The Metaphysics of Morals* [trans. Gregor M]. Cambridge University Press, Cambridge, UK

Locke J 1960 [1690] In: Laslett P (ed.) *Two Treatises of Government*. Cambridge University Press, Cambridge, UK

Nozick R 1974 *Anarchy, State, and Utopia*. Basic Books, New York

Plato 1974 [ca. 380 BCE] *The Republic* [trans. Grube G M A]. Hackett, Indianapolis, IN

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Rawls J 1993 *Political Liberalism*. Columbia University Press, New York

Scanlon T M 1999 *What We Owe To Each Other*. Harvard University Press, Cambridge, MA

Sen A 1982 Equality of what? In: Sen A (ed.) *Choice, Welfare and Measurement*. Cambridge University Press, Cambridge, UK

Sen A 1992 *Inequality Reexamined*. Harvard University Press, Cambridge, MA

Sidgwick H 1981 [1874] *The Methods of Ethics*. Hackett, Indianapolis, IN

Smith A 1979 [1759] In: Raphael D D, Macfie A L (eds.) *The Theory of Moral Sentiments*. Oxford University Press, Oxford, UK

Walzer M 1984 *Spheres of Justice*. Basic Books, New York

T. Pogge

Justice: Political

Political justice is a term whose meaning varies according to the different aspects of justice that it is used to highlight. This article discusses four different ways in which the term is used: (a) to characterize the principles that govern a distinctly political sphere of distributive justice; (b) to point to a distinctly political way of arriving at principles of distributive justice; (c) to describe the departures from the ordinary forms of criminal justice that are required by political trials and war crime tribunals; and (d) to characterize the distinctive forms of justice that emerge within political as opposed to other forms of community. After looking at each of these different ways of talking about political justice, the article concludes by exploring some of the controversies that emerge from the tensions among them.

1. Political Justice

1.1 Political Justice as a Sphere of Distributive Justice

The simplest and most familiar way of talking about political justice is as a particular sphere of distributive justice (Walzer 1983, Chap. 13). Justice is most frequently described as a set of principles that guide the distribution of goods within a community. As a sphere of distributive justice, political justice refers to the principles by which distinctly political goods, such as citizenship, voting rights, and access to political office, are distributed. Debates about these principles, with their clash of democratic, meritocratic, and aristocratic arguments, are a familiar feature of everyday political life as well as the history of political thought, so they do not need much elaboration here.

1.2 Political Justice as a Source of Distributive Principles

This way of talking about political justice is much less familiar, at least beyond the walls of the academy. John Rawls, the author of the influential book, *A Theory of Justice* (1971), has popularized it within the academic community. A political conception of justice, according to Rawls, is one in which shared principles of distributive justice are derived from the agreement of rational people who find themselves in a common political situation or culture, rather than from what Rawls calls 'comprehensive' claims about human nature, rationality or well-being. Political justice, in Rawls' sense, is contrasted with metaphysical justice or philosophic justice rather than with principles of justice that govern the distribution of goods in other spheres of life, such as the economy, the family, or civil society. Rawls argues, in particular, that rational and truly impartial participants in a democratic public culture would respond to the challenges of a pluralistic society by choosing the two principles of justice that he defended *A Theory of Justice*. Rawls developed this conception of political justice in order to defend his theory of justice against complaints that it rested upon untenable assumptions about human nature and rationality (Rawls 1985, 1993). It has become popular among some scholars

who welcome it as a way of philosophizing about justice without metaphysics. There is, however, a great deal of controversy about whether Rawls can reach his substantive conclusions successfully without invoking the kind of assumptions about human nature and rationality that he declares off-limits to a 'political' theory of justice.

1.3 Political Justice as Extralegal Procedures of Criminal Justice

A third way of talking about political justice focuses on the exercise of justice in criminal trials. It refers to those difficult cases in which extraordinary circumstances make it impossible to follow the ordinary norms of legal procedure that are designed to shield criminal justice from partisan political beliefs and passions. Such cases include, most prominently, the trials of political leaders for acts that they committed in previous regimes and of members of defeated nations for acts that its members committed in war (Kirchheimer 1961, Shklar 1986). In political trials or war crime tribunals many well-known and generally respected norms of criminal procedure, such as judgment by neutral third parties and the prohibition against retroactive legislation, are hard, if not impossible, to follow. As a result, these exercises of justice look questionable from a purely legal point of view, the imposition of the political preferences of the powerful rather than an application of the rule of law.

The need for this particular conception of political justice has been questioned from two opposing perspectives. On the one hand, many scholars believe that we can extend the familiar norms of the rule of law even into the controversial arena of political trials and war crime tribunals. On the other hand, other scholars believe that the distinction between political and ordinary cases, or between politics and law, is a myth. They argue that all criminal cases are, in effect, political because rule of law norms invariably subordinate legal judgments to partisan political goals and opinions.

1.4 Political Justice as Communal Practice

A fourth and final way of talking about political justice treats it as the distinctive way in which judgments about right and wrong are made and enforced within political communities. Political justice, in this sense, is contrasted with domestic justice or imperial justice, i.e., with the way in which justice is practiced in communities that lack the conceptions of citizenship and sharing of power that are usually associated with the idea of political community. The classic model for this version of political justice can be found in Book 5 of Aristotle's *Nicomachean Ethics*. Aristotle argues there that we need to distinguish the form of justice that develops among those who take turns in ruling and being ruled, i.e., among members of a political community, from the forms that develop among members of families or subjects of large empires. He goes on to argue that in political communities justice takes two different forms: conventional right, which concerns judgments about matters—such as which side of the road we should drive on—where we are looking for agreement rather than an intrinsically correct standard of justice; and natural right, which concerns judgments about matters—such as how to punish murder—where we debate about intrinsically correct standards (Yack 1993, Chap. 5). Understood in this way, political justice represents a way of exercising power rather than one aspect of distributive justice.

2. Comparisons and Controversies

These four ways of talking about political justice are logically compatible with each other. In practice, however, there are some tensions among them, tensions that inspire interesting controversies in contemporary debates about justice.

The most interesting of these tensions is the difference between the first two conceptions of political justice described above, each of which treats political justice as a component of distributive justice and the last two conceptions, each of which treats it as something more than distributive justice. In most theories the 'normal model' of justice (Shklar 1990), is a distributive model, a set of rules and principles governing the distribution of goods within a particular community. The third and fourth conceptions of political justice discussed above, however, challenge this normal, distributive model of justice in a way that the first two conceptions do not. For if you believe that there is a realm of criminal justice in which political judgments about the good of communities should take precedence over pre-existing rules and procedures, then you are treating justice as something more than the distribution of goods according to the impartial application of generally recognized rules and principles.

And if you believe that political justice concerns a distinctly political way of exercising power to further the good of a community, then distributive models of justice look like attempts to de-politicize the exercise of justice, attempts to hide the assertion of power that accompanies every attempt to impose standards of justice upon a community.

Viewed from the perspective of the third and fourth conceptions of political justice outlined above, the Rawlsian conception is strikingly *apolitical*. Rawls may start out from our shared political condition, what he calls 'the fact of pluralism.' But in order to deal with this situation he asks us to *imagine* how rational and truly impartial participants in a shared democratic public culture would choose to order the

basic institutions of their society. As a result, there are no politics, at least as politics are ordinarily understood, in Rawls' conception of political justice: no competition for power, no efforts at persuasion or co-optation, no tests of our ability to use power wisely. Politics, as ordinarily understood, are subordinated to a set of principles of justice in Rawls' conception of political justice.

But viewed from the perspective of Rawlsian and other 'normal' models of justice, there is strikingly little *justice* in conceptions of political justice that treat justice as more than the application of a body of pre-existing rules and principles. For making justice depend on the partial and contestable judgments that inform the exercise of political power takes away the determinacy that most theorists associate with the concept of justice. Without stable standards equally available to all, how, they ask, can we use justice to settle our conflicts and establish social order? There may well be, from their point of view, a sphere of activity in which the exercise of power cannot be guided by pre-existing rules and principles. But in that sphere we go beyond the reach of justice and have to call on other virtues, such as prudence, benevolence, or courage, to guide us.

Where does this controversy leave us with regard to the concept of political justice? The first conception of political justice, as a particular sphere of distributive justice, is relatively uncontroversial. As long as political rights and privileges are distributed within a community, there will always be debate about the best way to do so. Such debates continue today, most often among various versions of democratic and meritocratic principles of distributive justice.

The other three conceptions of political justice are, in contrast, quite controversial, as we have seen. All three conceptions grow out of the loss of the sense that we can derive the content and legitimacy of our standards of justice from a set of principles sanctioned by God or nature. But Rawls responds to that loss by trying to show how human beings confronted by a common political situation can construct such a set of principles for themselves without invoking higher authority. And those who employ the last two conceptions of political justice respond by insisting that we must acknowledge and accept the partial and uncertain use of power that inevitably accompanies the exercise of justice in political life. The theoretical difficulties that years of criticism have identified in Rawls' attempts to justify his two principles of justice raise serious doubts about his conception of political justice. But the lack of determinacy and stability in the opposing conceptions of political justice raise very legitimate concerns about our ability to use them to maintain a decent political order. Until these concerns are met by a more systematic and convincing theory, these conceptions of political justice seem unlikely to replace distributive models as the mainstream approach to theorizing about justice.

See also: Aristotle (384–322 BC); Criminal Justice, Sociology of; Equity; Justice and its Many Faces: Cultural Concerns; Justice and Law; Social Justice

Bibliography

Aristotle 1956 *Nicomachean Ethics*. Loeb Classical Library, Cambridge, MA

Kirchheimer O 1961 *Political Justice: The Use of Legal Procedure for Political Ends*. Princeton University Press, Princeton, NJ

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Rawls J 1985 Justice as fairness: Political, not metaphysical. *Philosophy and Public Affairs* **14**: 223–51

Rawls J 1993 *Political Liberalism*. Columbia University Press, New York

Shklar J N 1986 *Legalism: Law, Morals, and Political Trials*. Belknap Press of Harvard University Press, Cambridge, MA

Shklar J N 1990 *The Faces of Injustice*. Yale University Press, New Haven, CT

Walzer M 1983 *Spheres of Justice*. Basic Books, New York

Yack B 1993 *The Problems of a Political Animal: Community, Justice and Conflict in Aristotelian Political Thought*. University of California Press, Berkeley, CA

B. Yack

Justice: Social Psychological Perspectives

1. Definition and Subject-matter of Social Psychological Discussion of Justice

Justice means that people get what they are entitled to, or deserve, on the basis of who they are and what they have done. Although clear in abstract terms, this definition leaves undecided what exactly particular people are entitled to get. Similarly, the basic rule of distributive justice, that people have the right to be equally treated as others who are alike them, leaves unsettled the criteria according to which the likeness of people should be assessed and the kind of treatment that the people should get. Justice can be unambiguously defined only on an abstract level. The abstract definition is open to multiple translations into concrete terms. For that reason, it is likely that different people or groups differ in their justice judgments of given conditions or circumstances. Even if views of justice are socially shared, this does not change the basically subjective nature of judgments of justice and injustice.

In contrast to other scientific disciplines studying justice, social psychology does not take a normative approach. It deals with justice in a descriptive rather than a prescriptive way. The aim is not to define what is just and unjust, and how justice can be achieved. The

focus on the contrary is on the subjective sense of justice and injustice and its impact on human action and judgment. Social psychologists study what people regard as just and unjust under given circumstances, how people deal with the concept of justice, how they react to situations that they regard as unjust, and under which circumstances, and why, people care about justice.

2. *Distributive Justice: Early Beginnings and Development of the Field*

The social psychological discussion of justice originated from relative deprivation research (Stouffer et al. 1949, Runciman 1966) and Homans' (1961) propositions about distributive justice. Although in slightly different terms, all these writings proposed that people evaluate outcomes they receive relative to some kind of comparison standard rather than in absolute terms. The comparison standards define the quality of outcome people expect to get based on the outcomes that other people receive. The evaluation of one and the same outcome differs depending upon its position relative to the comparison standard used. If the outcome is equal to the comparison standard it will be experienced as satisfying and appropriate, if it is lower it will be experienced as dissatisfying and inappropriate.

An important step in the development of the field was the research and theorizing by Adams (1965) on affective, cognitive and behavioral consequences of perceived inequity. Similar to Homans, the author suggested that people consider what they contribute to an interaction or relationship (i.e., their inputs) and what they get in return (i.e., their outcomes). They compare the ratio of their own outcomes and inputs with the outcome–input ratio of others and expect the ratios to be equal. If the outcome–input ratios of a person and the comparison other do not match, people experience inequity. Inequity is an aversive state of distress that motivates people to restore equity either 'actually' by behavioral means or 'psychologically' by cognitive reconstruction. Interestingly, equity theory claims that this holds equally for people who are worse off and those who are better off than the comparison other. Adams used the equity model to explain workers' reactions to under- and overpayment at work. Walster et al. (1978) developed equity theory as a general theory of social justice and applied it to various different kinds of social interactions and relationships.

Another important impetus came from the just world hypothesis (see Lerner 1980). According to this view, people have a basic desire to believe in a just world in which people get what they deserve. This basic desire leads people to act, or engage in cognitive distortion, in order to maintain or re-establish the belief that the world around them is just.

Early social psychological conceptualizations of justice focused on one particular form, namely *distributive justice*, i.e., the justice of the distribution of outcomes or resources. In addition, the early conceptualizations concentrated on one particular principle of distributive justice, the equity or contribution principle. This principle prescribes that people's outcomes should be proportional to their inputs, and the outcome-input ratios of different people should be equal. The narrow focus of early theory and research was later extended in several respects. Within the domain of distributive justice, alternative conceptualizations proposed the existence of multiple justice norms, the most prominent being the equality principle (to each the same), the contribution or equity principle (to each according to his contributions), and the need principle (to each according to his needs). The multi-principle approaches to distributive justice assume that different distribution rules are regarded as just, and serve as principles of action and evaluation, under different conditions (see Deutsch 1975, Mikula 1980). One of the decisive conditions is the basic orientation of the respective social system or group. According to this view, the equality principle predominates in groups with solidarity orientation, the equity or contribution principle within production oriented groups, and the need principle in groups with caring orientation. Empirical studies in this tradition of research were designed to identify situational and personal conditions that determine which distribution principles are used and regarded as just (see Törnblom 1992).

3. *Procedural Justice*

Further developments in the field extended the focus and considered additional forms of justice. Justice judgments do not only refer to the way in which outcomes are distributed, but also to the decision-making procedures by which the distributions are arrived at, i.e., *procedural justice*. The social psychological study of procedural justice was stimulated by the work of Thibaut and Walker (1975). These authors studied people's responses to different third-party procedures of conflict resolution. They found that people regard procedures that provide them with control of the decision-making process, and the decision itself, to be fairer. In addition, they found that the perceived fairness of the procedures shaped people's satisfaction with their outcomes. Subsequent research studied particular features of procedures that contribute to the perception of procedural justice and injustice, and the consequences of these perceptions (see Lind and Tyler 1988, Tyler et al. 1997).

Justice researchers identified a variety of procedural elements that contribute to perceived fairness of procedures. For instance, Leventhal (1980) proposed that procedures will be regarded just if they ensure:

consistent treatment across persons and over time, utilization of accurate information, suppression of personal biases of decision makers, existence of appeal mechanisms by which wrong decisions can be corrected, representation of the affected parties in the decision-making process, and compatibility with fundamental moral and ethical values. Empirical evidence suggests that three procedural features are particularly relevant to the perception of procedural justice: the opportunity for participation of the parties concerned, decision makers' accounting for the decisions made, and treatment of concerned parties with dignity and respect. People consider procedures that allow them to participate in the decision-making process and express their opinions and views ('voice') to be fairer. Interestingly, this seems to occur even when the expressed opinion does not influence the final decision, as long as the respective parties feel that their views have been seriously considered. In addition, people regard decisions more fair if the decision-making authorities explain and account for their decisions. This gives the impression of neutrality and increases people's trust in the authorities. Finally, people regard procedures more fair when the authorities treat them with dignity and respect. Taken together, people seem to be concerned both about structural aspects of the decision procedures and social aspects, i.e., the enactment of the procedures.

Perceived procedural justice has a variety of positive consequences. Procedural justice promotes people's acceptance of decisions, their long-term adherence to agreements, and their willingness to cooperate with groups they belong to. Perceived procedural justice also enhances the perceived legitimacy of civil authorities, institutions and rules, and people's willingness to defer to authorities and their decisions. Finally, perceived procedural justice improves people's satisfaction with the outcomes they receive. Strikingly, these positive consequences result even if the decision procedures lead to unfavorable outcomes for the persons concerned.

4. *Reactions to Injustice*

Social psychological theory and research distinguished three major types of responding to injustice: actual restoration of justice by behavioral means, psychological or cognitive restoration of justice by changing the interpretation of the situation, and nonacting or resignation (see Tyler et al. 1997). People may prefer cognitive over behavioral reactions to injustice for various reasons. People who caused injustice, and people who are advantaged unfairly, may be inclined to justify the given situation to avoid the need for compensation or restitution and to protect a positive self-image. People who are disadvantaged unfairly may be motivated to deny being victims of injustice in order to protect their self-esteem. In addition, they may come to the conclusion that they themselves are responsible for their own fate. They may also view their negative situation as a result of a legitimate process and thus not regard it as unjust. Finally, they may perceive their situation as normal and appropriate because they compare themselves predominantly with similar others who share the disadvantage. People also seem to be inclined to regard 'what is' as 'what ought to be.' Another important reason why people may not behaviorally react to injustice is that they feel powerless and have resigned to their fate.

Behavioral reactions to injustice include attempts to restore justice either in the form of seeking compensation or retaliating against the responsible agent, or leaving the unjust relationship or group. The likelihood of behavioral reactions of the former type increases when people see a chance of improving the situation, and when they assume that the situation would not change without any action. In addition, perceived personal benefits and costs will influence whether people react behaviorally or not. However, the intensity of feelings of injustice seems to be the most relevant factor. People with strong feelings of injustice are more likely to react behaviorally to injustice independently of the costs involved and the likelihood of success. Behavioral reactions to injustice can occur as individual-level responses and collective or group-level responses. Collective reactions are more likely when the injustice is perceived as a collective rather than an individual issue.

5. *Subjectivity, Divergent Views of Justice, and Social Conflict*

The subjective nature of justice judgments makes it plausible that different people or groups differ in their assessments of given situations. Divergent evaluations and interpretations are particularly likely among people who hold different perspectives or roles in relation to a perceived injustice (Mikula et al. 1998, Tyler et al. 1997, Walster et al. 1978). Three particularly relevant perspectives are the perspective of those who suffer from the injustice (i.e., the 'victim'), the perspective of those who caused the situation in question (i.e., the 'actor' or 'perpetrator'), and the perspective of people who perceive the critical situation without being personally affected (i.e., the 'observer'). Actors typically tend to regard their own behavior as appropriate and just, and to attribute any negative behavior on their part to external circumstances. Victims and observers are more likely to regard the same behavior as unjust, and attribute responsibility and blame to the actor. Similarly, majorities and minorities sometimes disagree about what constitutes a fair decision-making procedure in political settings.

Divergent views about what is just and unjust can evoke social conflicts. The notion of justice can be

helpful for, but also detrimental to, the limitation and resolution of social conflicts (Mikula and Wenzel 2000). To be helpful, the conflicting parties have to reach a shared view of the situation and their entitlements or to recognize the existence of differing but equally legitimate views of justice. Rigid ideas about what is just, and the firm conviction that the own view is the only true and right one, contribute to the escalation of social conflicts. Conflicting parties may even exclude each other from the moral community within which moral values and rules of justice apply. In this case, justice loses any conflict resolving function.

6. *Why Do People Care About Justice?*

The question of why people care about justice received different answers in different social psychological theories. One group of theorists assume that people are concerned about justice because it serves their self-interest of maximizing their outcomes in the long run. For instance, equity theorists (Walster et al. 1978) proposed that social systems develop rules of justice, and induce members to follow these rules, because this can maximize collective outcomes. People expect others to follow the rules and follow the same rules themselves as long as it is in their interest to do so. Early theorizing on procedural justice (Thibaut and Walker 1975) argued in a similar way. People care about fair procedures, and regard procedures that give them control over outcomes to be fairer, because they help them to maximize their outcomes in the long run. The relational or group-value model of procedural justice (Lind and Tyler 1988, Tyler and Lind 1992) provides an alternative perspective. According to this theory, people are concerned about their position in groups. They use experiences with their treatment by authorities as a source of information about their position. The evidence that they are treated justly indicates that they are regarded as valuable and worthy members of the group. This has positive implications for their self-esteem and self-worth. While providing different explanations of people's caring about justice, due to instrumental concerns about outcomes versus relational concerns about group status, both groups of theorists regard justice as a means rather than an end in itself. Other authors criticize that these views do not capture the essential feature of justice as a moral ought. Without denying that justice may be used as an instrumental means, they argue that justice often is an end in itself. People are motivated intrinsically to behave justly, and sometimes act in ways which are not in line with self-interested motives and maximizing own outcomes. Future justice research will have to consider, and try to integrate, the different theoretical positions and analyze the conditions which stimulate the two manifestations of justice concerns in people's conduct and thinking.

7. *Areas of Social Justice Research*

Social justice research expanded into many different areas of social life: work organizations, legal institutions, educational settings, civil institutions, social policy and economics, gender issues, environmental issues, international conflicts, family issues, close personal relationships, and coping with life crisis and victimization. People's perceptions of justice and their reactions to injustice in work organizations have been one of the main areas of social justice research from the very beginnings of equity theory. The rapid growth of this line of research has led some authors to use the generic term 'organizational justice' (see Cropanzano and Greenberg 1997). Empirical evidence shows that the perception of distributive justice shapes people's satisfaction with the outcomes they receive, work performance, absenteeism, and turnover. The perception of procedural justice influences workers' organizational citizenship behaviors and reactions to a variety of organizational measures such as personnel selection procedures, performance evaluations, layoff decisions, strategic organizational planning and change, and workplace drug screening.

8. *Concluding Remarks*

Social psychological research has substantiated impressively that people care about justice and proved that justice judgments influence human conduct and thinking. Justice matters on the interpersonal level, the intergroup level, and the societal level, even though differences exist with respect to the justice principles that are used in evaluations of justice and injustice, the perception of personal and collective injustice, and the conditions that lead to personal level and group level reactions to perceived injustice (Tyler and Smith 1998). Some authors argue that moral values and rules of justice apply only within a limited range of social relationships. People are less concerned about justice with respect to people who are outside their own social group or category. People can be excluded from the moral community within which moral values and rules of justice apply because they are dissimilar or different or because they are perceived as competitors or enemies.

See also: Conflict and Conflict Resolution, Social Psychology of; Justice and its Many Faces: Cultural Concerns; Justice, Access to: Legal Representation of the Poor; Justice: Philosophical Aspects; Social Justice

Bibliography

Adams S J 1965 Inequity in social exchange. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*, Academic Press, San Diego, pp. 267–99

Cropanzano R, Greenberg J 1997 Progress in organizational justice: Tunneling through the maze. *International Review of Industrial and Organizational Psychology* **12**: 317–72
Deutsch M 1975 Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues* **31**: 137–49
Homans G C 1961 *Social Behavior: Its Elementary Forms*. Harcourt, Brace and World, New York
Lerner M J 1980 *The Belief in a Just World: A Fundamental Delusion*. Plenum, New York
Leventhal G S 1980 What should be done with equity theory? New approaches to the study of fairness in social relationships. In: Gergen K J, Greenberg M S, Willis R H (eds.) *Social Exchange. Advances in Theory and Research*. Plenum, New York, pp. 27–55
Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum, New York
Mikula G 1980 On the role of justice in allocation decisions. In: Mikula G (ed.) *Justice and Social Interaction: Experimental and Theoretical Contributions from Psychological Research*. Springer-Verlag, New York, pp. 127–66
Mikula G, Athenstaedt U, Heschgl S, Heimgartner A 1998 Does it only depend on the point of view? Perspective-related differences in justice evaluations of negative incidents in personal relationships. *European Journal of Social Psychology* **28**: 931–62
Mikula G, Wenzel M 2000 Justice and social conflict. *International Journal of Social Psychology* **35**: 126–35
Runciman W G 1966 *Relative Deprivation and Social Justice: A Study of Attitudes to Social Inequality in Twentieth-Century England*. Routledge & Kegan Paul, London
Stouffer S A, Suchman E A, De Vinney L C, Star S A, Williams R A, Jr. 1949 *The American Soldier: Adjustments during Army Life*. Princeton University Press, Princeton, NJ
Thibaut J, Walker L 1975 *Procedural Justice: A Psychological Analysis*. Erlbaum, Hillsdale, NJ
Törnblom K Y 1992 The social psychology of distributive justice. In: Scherer K R (ed.) *Justice: Interdisciplinary Perspectives*. Cambridge University Press, New York, pp. 177–236
Tyler T R, Lind E A 1992 A relational model of authority in groups. In: Zanna M (ed.) *Advances in Experimental Social Psychology*. Plenum, New York, pp. 115–91
Tyler T R, Smith H J 1998 Social justice and social movement. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*. McGraw-Hill, Boston, pp. 585–629
Tyler T R, Boeckmann R J, Smith H J, Huo Y J 1997 *Social Justice in a Diverse Society*. Westview Press, Boulder, CO
Walster E, Walster G W, Berscheid E 1978 *Equity. Theory and Research*. Allyn and Bacon, Boston

G. Mikula

Juvenile Justice: International Law Perspectives

Since the 1980s there has been a quiet revolution spreading through the world of child criminal justice. It has spread globally as far as Australia (Young Offenders Act (1993)) and New Zealand (Children, Young Persons and their Family Act 1989) in the South Pacific, Canada (Young Offenders Act 1985, as amended), and Costa Rica in the Americas, Ghana, South Africa (Constitution of South Africa 1996) and Uganda in Africa (Child Statute 1996).

The catalyst for this reform is the United Nations Convention on the Rights of the Child 1989—the first global treaty protecting children's rights (see Van Bueren 1998a). The Convention on the Rights of the Child enshrines the full range of civil, political, economic, social, and cultural rights. It stresses a holistic approach to children's rights—all rights are indivisible and related. This has important implications for child criminal justice, as it implies that all rights must be considered for children in the criminal justice system, from the right to freedom of expression to the right to the highest attainable standard of health.

Freedom of expression, protected in Articles 12 and 13 of the Convention, envisages that children must be consulted in all decisions which affect them. This includes the planning of child criminal justice policy from prevention through detention policies and conditions. Indeed, Sri Lanka was specifically criticized by the Committee for not taking children's views into account.

The Convention on the rights of the child focuses states' attention on children as victims of crime, an aspect of criminal justice that, given the demonization of children in some countries, is often overlooked. Article 39 obliges all States Parties to take appropriate measures to promote 'physical and psychological recovery and social reintegration' for child victims of abuse, neglect, torture, cruel, inhuman and degrading treatment, and punishment. Such recovery and reintegration services ought to occur in an environment which fosters the health, self-respect, and dignity of the child.

The principle of nondiscrimination under Article 2 is also applicable to the child criminal justice system. In particular, the committee has noted discrimination against socially and economically disadvantaged children. In relation to Bolivia, the UN Committee on the Rights of the Child recommended that alternatives to institutional care for economically and socially disadvantaged children should be created.

As a human rights treaty, the Convention removes the debate about child criminal justice from the bitter sphere of party politics and places the discussions and campaigns firmly in the arena of legal commitments, which a country is required to implement. In the 191 countries bound by the Convention, the provisions of child criminal justice are removed from party politics but are used by campaigners to help shape the political debate, particularly when drafting bills of rights and children's statutes.

In addition, there are three sets of rules adopted by the global community which provide greater detail on the daily operation of child criminal justice. These

rules are the United Nations Standard Minimum Rules for the Administration of Juvenile Justice 1985 (Beijing Rules), United National Rules for the Protection of Juveniles Deprived of their Liberty 1990, and United National Guidelines for the Prevention of Juvenile Delinquency 1990 (Riyadh Guidelines). These, together with the United Nations Guidelines for Action on Children in the Criminal Justice System 1997, and the United Nations Declaration of Basic Principles of Justice for victims of Crime and Abuse of Power 1985, provide a comprehensive framework of child criminal justice. The texts are reproduced in van Bueren (1988b).

The provisions in the Convention on the Rights of the Child, together with the provisions in the international instruments, have resulted from experience-gathering and reflection over several decades on the best ways to protect children's rights.

States from different regions are also bound by regional treaties, which include principles on child criminal justice. These embrace the European Convention on Human Rights 1950, the American Convention on Human Rights 1969, and the African Charter on the Rights and Welfare of the Child 1990.

1. *Children, not Juveniles*

We have begun to shift away from using the term 'juvenile.' The United Nations Programme of Action does not concern 'juvenile justice' but rather 'children in the criminal justice system.' If, as Rorty and others argue, experience is shaped by language (Truth and Progress 1988), then the adoption and use of the term 'juvenile justice' by an number of English-speaking jurisdictions also has to be queried.

Because of international legal developments, the time has come to question the rationale behind describing a young person in conflict with the law as a juvenile offender, while a person under 18 in need of state protection is described in terms of child welfare. The only acceptable justification for retaining the distinction between juveniles and children is that juvenility sometimes includes those over the age of 18, and some states wish to extend additional levels of protection to this age group. However, this can be done without retaining two separate terminologies for those under 18.

The use of an inappropriate judgmental layer makes it harder to argue against the more punitive suggestions being made in the American Counter Reformation. Distinguishing between adult and under-18-year-old offenders resulted from differences in approach between the civil and the criminal legal systems in relation to children. However, as the international laws on children make clear, all of the special protections apply to children under the age of 18 in the justice system, regardless of whether the field is civil or criminal. In international law, the special protections attach because of childhood, and childhood is defined solely with regard to age.

2. *The Benefits of Using a Human Rights Approach*

The international laws on child criminal justice also act as powerful but peaceful catalysts for change. The Convention on the Rights of the Child takes the argument beyond questioning who should have priority—children or society. The international law is clear—the best interests of the child dictate that children have priority; by prioritizing children, society benefits. According to Riyadh Guideline 1.4, child criminal justice 'should be conceived as an integral part of the national development process of each country.' It is no longer permissible under international law to place child criminal justice at 'the bottom of the pile' of resource allocation. States are under an immediate legal duty to ensure that children's civil, economic, social, and cultural rights in child criminal justice are fully protected. This can only be done if sufficient resources are allocated for all stages of the justice system, not just in courts but also in prevention and detention. A well-resourced child criminal justice system operating in accordance with international law should not be regarded as a drain on national resources. Indeed, it should reduce the risk of reoffending.

The Convention calls for the establishment of a child-oriented system which recognizes the child as the holder of rights and freedoms. The Convention stresses the need for all actions in child criminal justice to be guided by Article 5, which provides that the best interests of the child should be the primary consideration. Indeed, the United Nations Committee on the Rights of the Child, the body entrusted with implementing the Convention, criticized Mongolia because Mongolian legislation ranked the recovery and reintegration of children as only the third goal behind the protection of society and the punishment of the child.

One of the attractions of the international law on child criminal justice is that it requires a psychological shift for society. International law now merges what were once the punishment and rehabilitation functions of juvenile justice with the service provision and protections of child welfare. The trend implies that much of child criminal justice should be governed by the same principles as child protection. The unified child rights approach is the same for the civil and the criminal justice fields. Some states, such as Uganda, have one children's statute governing both areas of justice. It is also reflected in the return, albeit with modification, of indigenous values and approaches, for example, Family Group Conferencing, developed

from Maori traditions in New Zealand, places child criminal justice within the family unit. This is a very different approach from punitive penalties for parenting.

The international laws on child criminal justice have also been used by states when drafting their constitutions. The 1996 South African Constitution has adopted verbatim provisions from the Convention on the Rights of the Child in relation to detention. Hence, the international laws on child criminal justice have helped to inspire legislative reform and innovative and creative thinking. Those campaigning for legislative reform have not only been adults. In June 1985, the National Street Children's Movement was established and held its first Street Children's Congress in Brasilia, with over 500 street and working children taking part (see Rizzini et al. 1993). It attracted much media attention and the children were articulate in denouncing the State Foundation for the Welfare of Minors (FUNABEM), and institutional and police violence against them in the criminal justice system. They demanded urgent reforms.

The Movement began to give the children a more positive public image. Two constitutional amendments, inspired by the Convention on the Rights of the Child, were proposed, and supported by petitions signed by 1,300,000 children and 200,000 voters. These proposals became the Brazilian Constitution's Chapter on the Rights of Children and Adolescents 1988. Children were also consulted on the provisions in what became the Children and Adolescent Act, Hence the mobilization of different sections of society is seen as indispensable to raising public awareness and participation about child criminal justice issues.

Change has not only occurred through popular participation. In some countries, the courts have reacted to campaigns for improving child criminal justice. In India, a case was brought to the Supreme Court by a social worker and freelance journalist, to remedy the appalling conditions in which children were held in jail (*Sheela Barse v. Union of India*). The Supreme Court of India directed that every district and session judge should visit the district jail at least once every two months and, in the course of the visit, consider the conditions of children in prison and draw the attention of their High Court and the administration to penal conditions. The Court also directed the State Legal Aid Boards to make legal advice available to children detained in prison awaiting trial on criminal charges (Rizzini et al. 1993, p. 599).

The Court referred to an order made on 12 July 1986 relating to children with disabilities, and destitute and abandoned children. It directed the Director General of All India Radio to give publicity seeking the cooperation of nongovernmental organizations and social services in the reintegration and recovery of these children. The Court remonstrated that these children should not be in jail, but should be given adequate medical treatment and receive training.

The Court also suggested that, rather than each state having its own Children's Act, it would be desirable to have national parliamentary legislation. Just as the Inter-American Commission on Human Rights was concerned about the different standards prevailing in child criminal justice in American states, so the Indian Supreme Court was concerned about the different standards found in Indian states.

The court also ruled that the petitioner could visit jails, borstals and all child criminal justice accommodation institutions in all parts of India. The Supreme Court proclaimed, '[t]he petitioner need not be looked upon as an adversary. She had in fact volunteered to do what the State should have done' (Rizzini et al 1993 p. 600). The Court saw the petitioner as helping it fulfil its goal in protecting children's rights.

Similarly, the South African Constitutional Court in Hugo upheld President Mandela's decree that all mothers in prison on 10 May 1994 with children under the age of 12 years should benefit from a remission of sentences. In other words, rather than have children in prison with their mothers, or children left abandoned outside prison without their mothers, the best interests of the child dictated that mothers should be released. The Constitutional Court, in upholding Mandela's decree, took into account that the release of mothers would help prevent abandonment and poverty of children whose mothers were incarcerated.

The President also decreed the release of two other groups: prisoners with disabilities, and young people.

There is still a very long way to go, but the Convention on the Rights of the Child has operated as a very effective catalyst in protecting the rights of children in the justice system.

See also: Adolescent Vulnerability and Psychological Interventions; Antisocial Behavior in Childhood and Adolescence; Children and the Law; Children, Rights of: Cultural Concerns; Criminal Justice, Ethics of: Aspects of Human Dignity; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Social Competence: Childhood and Adolescence

Bibliography

Rizzini, Rizzini, Munhoz, Galeano 1993 *Childhood and Urban Poverty in Brazil: Street and Working Children and Their Families*. UNICEF

1998 Truth and progress. *Philosophical Papers* **3**

Van Bueren G 1998a *The International Law on the Rights of the Child*. Kluwer

Van Bueren G 1998b *International Documents on Children*, 2nd edn. Kluwer, Amsterdam

G. Van Bueren

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

K

Kant, Immanuel (1724–1804)

'The character of a true philosopher is such that the philosopher does nothing other than to exercise his natural powers and capacities—this namely through the searching investigation of critique.' This claim cannot be put more humbly—it is the very claim which guided Kant on his way to the *Critique of Pure Reason* and effected a 'revolution in style of thought' which continues until this day. According to Kant, this exercise and development of the best human powers deserves the title of philosophy only then, however, if such activity emanates from independent thinking: 'my philosophy must be grounded in my own self, and not in the understanding of others.' Philosophy cannot merely serve the satisfaction of individual curiosity, but rather should be related to the 'interests of humanity.' At the outset of his critical career, Kant wrote that he would deem himself 'more useless than a common laborer' if he couldn't believe that philosophy would contribute to 'the establishment of the rights of humanity.'

Kant was born on April 22, 1724 in Königsberg into a family of manual laborers of very little means. His schooling took place in the strict spirit of Pietism. At the age of 16, Kant began his studies at the university of his native city, financing himself in part by giving private lessons and playing at the billiard table. The breadth of his subjects of study—mathematics, natural science, theology, philosophy, and Latin—was later to be reflected in the wide spectrum of Kant's fields of inquiry. Initially, his interest in modern physics was particularly strong; he viewed Newton's *Philosophiae naturalis Principia mathematica* as the paradigm of rigorous science.

Kant's first academic publication, *Thoughts on the True Estimation of Living Forces* (1747), owes its importance above all to its central claim, which is very much indicative of Kant's later philosophy. Here the 24-year-old attempts to settle a dispute between Cartesians and Leibnizians by proposing a mediating position relative to the point of view of each party. Kant exclaimed: 'I have already drawn up the path which I would like to follow. I shall set upon it, and … nothing should prevent me from pursuing it further.' Having lost the basis of his financial support through the death of his father, Kant left the university in 1746 in order to earn a living as private tutor in various families. During this time, Kant had occasion to pursue several extensive studies in natural science, the results of which he published anonymously in 1755 under the title *Universal Natural History and Theory of the Heavens*. Here, both the birth of the cosmos and solar systems, as well as the development of the earth up until the appearance of life and the rational animal, man, are deduced solely from physical laws. In positing the creation of heavenly bodies out of rotating nebula, Kant anticipated Laplace's theory; and the insight that spiral nebula are actually very distant galaxies secured Kant a sure place in the history of astronomy. In the year 1755, Kant returned to the university. Having attained a doctoral degree with a meditation 'On Fire' (*De igne*), he earned his *venia legendi* with a postdoctoral treatise on 'First Principles of Metaphysical Cognition.' In this work Kant marked out that field of problems with which he was to occupy himself for his entire life. Even the so-called 'transcendental philosophy' of the critical period remains connected to the 'first principles,' in that it poses the question after the 'conditions of the possibility' of our knowledge. Although these 'conditions of possibility' are no longer located in a divine understanding—later, Kant came to understand them as the functional conditions of human understanding and in this respect as *a priori*—they nevertheless continue to be the object of metaphysics.

For Kant, all scientific knowledge is metaphysical insofar as it requires us to go, by means of concepts, beyond purely empirical experience. This is always the case in those passages where Kant speaks of knowledge in general, of 'world' and 'reality,' of morality, beauty, or history. Understood in this sense, transcendental philosophy is a 'metaphysics of metaphysics.' Moses Mendelssohn's famous statement that Kant was a 'Crush-All' of metaphysics should therefore not be so construed, as if all metaphysics had been discredited by critique. Actually, Kant disabled only the claims of a theorational and empirist metaphysics, while setting 'critical' standards which all scientific metaphysics must be able to satisfy.

In the over 25-year period between Kant's postdoctoral thesis of 1755 and the publication of *Critique of Pure Reason* in 1781, Kant wrote several works concerning the foundation of theoretical and practical metaphysics. He also experimented with empirically designed procedures, and began to show a certain affinity for skeptical methods—an affinity which was to prove lasting. Two works from this period give us the impression that Kant might have considered abandoning metaphysics altogether: the elegantly parleying *Observations on the Feeling of the*

Beautiful and the Sublime (1764) and the satirical *Dreams of a Spirit-seer* (1766). Kant's appearance of having an antimetaphysical stance must have been strengthened for his contemporaries when, having completed yet another obligatory dissertation 'On the Form and Principles of the Sensible and Intelligible World' (*De mundi sensibilis atque intelligibilis forma et principiis*) in 1770, Kant published nothing worthy of mention for 11 years. In retrospect, we can recognize the approach of critical thought in this second dissertation: space and time are introduced as human forms of intuition which do not belong to things as understanding apprehends them. At this point, however, Kant still thought human understanding capable of a kind of cognition which could apprehend the nature of things as a divine understanding might.

The continuing 'critical' insights of the following years forced Kant to renounce the claim that we are able to know the nature of things. From a human perspective, nothing can be said about how 'things-in-themselves' are constituted. At this point in Kant's philosophical development, such 'things-in-themselves' as a divine understanding might be able to know can no longer serve us as the idea of a substance which could provide a basis for the things we can know. It does not even make sense to pose the question concerning the reality of 'things-in-themselves,' for they are mere concepts—we do not even know if anything corresponds to them. As such, the 'thing-in-itself' is simply an expression of our own limitation. The 'thing-in-itself' may serve as a limiting concept, however, in order to prevent us from deeming human knowledge absolutely certain. Even the pure concepts of understanding and reason do not have access to absolute truth.

Kant by no means contests the possibility of knowledge; he is not an agnostic on this count. Rather, he emphasizes the relativity of human knowledge: every insight, even in the objective natural sciences, is functionally relative to human cognitive capacities. We perceive reality in the manner in which it 'appears' to us. Every objective perception is dependent on the sensible intuition in space and time, and remains in this respect limited to experience. But renouncing experience leads either to analytical judgements, which only express what is already contained in concepts, or it misleads us to speculation, in which all certainty is lost. Yet the speculative thinking of pure reason is not sterile. Under conditions of disciplined self-critique, pure reason may contribute to the guidance and ordering of empirical knowledge (its 'regulative function'), it can ward off dogmatic claims (its 'critical function'), and it can set practical goals in those areas where certain knowledge is not attainable anyway. In this last case, pure reason becomes 'practical reason.'

The *Critique of Pure Reason* (1781), in which Kant systematically developed these insights, initially received little attention. The second edition (1787), elucidated by the Prolegomena (1783) and expanded by the *Grounding for the Metaphysics of Morals* and the *Metaphysical Foundations of Natural Science*, gave rise to a sea change, the likes of which philosophy had seldom seen. Kant's experimental-philosophical proposal for the foundation of the results of human knowledge proceeded no longer from the character of the object, but rather from the conditions of the subject—an innovation which came to be understood as a revolution in science. The 'cardinal propositions' of human knowledge cannot be extracted from 'things-in-themselves,' but rather have their fundaments in our very acts of cognition, and only become the fundament of nature *qua* 'appearance' through human activity. This went very well with the heightened self-confidence of the Enlightenment. The critique of reason declares man originator of his world—without contradicting the results of natural science or questioning the traditional claims of the belief in divine origins. At the same time, the delimitation of knowledge by means of critique serves to assign belief a genuine place. God and world, freedom and the person come to be interpreted as concepts of pure reason; they do not denote any object, but are imperative for the conceptual self-determination of the individual.

This attempt to ground human procedural capacities *a priori* had consequences for practical action; Kant showed these in the *Critique of Practical Reason* (1788). In ethics, too, the cardinal propositions are not grounded in the world, but are the product of our own capacities. It is not possible to speak of any pre-existing 'good' or 'evil' in the world: one must seek the origin of these terms solely in the human will. Insofar as this will conceives itself as a rational one, it determines by itself what is to count as morally good. The principle of the will is therefore that of self-determination (autonomy) according to the principle of practical reason in the will itself. So long as we acknowledge a commitment to this principle—and we have already done so when we ask ourselves what we 'should' do—we understand ourselves as free and autonomous. In this way, we have the possibility to orient the principles of our action (maxims) on rational insight and submit them to the general self-jurisdiction of reason, which consists in the categorical imperative. Like the categorical elements of the perception of nature, the categorical imperative does not emanate from external authority but from man's own conception of himself as a rational being.

No new virtues are made compulsory in Kant's critical refoundation of ethics. Instead, principles from Platonic, Aristotelian, and Stoicist ethics are made subordinate to specific criteria of reason. In the last analysis, the emancipated individual alone must determine these criteria. Justification of behavior through one's own insight (and therefore through one's own reason), which first came into philosophical consciousness through Socrates, ascends to its full conceptual conclusion here. The autonomy of ethical

self-determination aspired to since antiquity, while being defined over and against empirical explanation, thereby gains a pointedly individual expression. The 'categorical imperative' has meaning only in relation to the 'maxims' or 'subjective principles' of the individual. It is to be noted that justification always refers only to the 'form' or conceptual part of a moral judgement. The concept alone is meaningless without content or 'material,' however, and this is also the case in ethics.

There are therefore always certain 'preferences' involved in that 'material' which must be adduced to every moral judgement and hence to every moral act. Feelings and passions belong to the moral world. Thus, there must be some feeling which gives impetus to moral action—Kant calls it the 'feeling of reverence.' He is careful not to conflate the necessary motivation of acting with its justification; as in Plato, justification may only proceed from rational insight. Solely by means of this kind of insight can man experience himself as free.

Kant's third critique, the *Critique of Judgement* (1790), contains in Part 1 a grounding of aesthetics, once again departing from human standards alone. Aesthetic judgement is the expression of a harmony of cognitive powers which stimulate the individual pleasurably: whenever sensibility and reason come together and mutually intensify one another, an aesthetic experience is possible. Aesthetic experience also expresses itself in a 'judgement' in which the individual, existing in unrestricted subjectivity, communicates himself to others. Kant frees aesthetic judgement from all claims of scientific truth, moral valuation, and purely sensible satisfaction. In aesthetic experience, everything depends on one's own self-autonomous ('heautonomous') judgement. This faculty of judgement binds all sensible and mental powers together in itself and is capable of producing its own pleasure. What is more, it demands expression among those who are like-minded. Hence the faculty of judgement does not at all isolate the individual, but is the very thing which brings him into a conscious relation with those who are congenial to him.

In Part 2 of the *Critique of Judgement* Kant draws up a theory of living nature. The individual constantly judges nature in analogy to his own experience as an active being. An organism is perceived in accordance with its functional purpose, 'as if' a rational and end-oriented will were operating within it—a will which the individual is admittedly only acquainted with through himself. Kant's theory of the organism attempts to grasp the dynamic complex unity of living beings. In so doing, he draws upon the principle of causation whenever this is possible. In this vein, Kant establishes the definition that the parts of the organism 'join one another in the unity of the Whole by virtue of being mutually the cause and effect of one another's form.' Indeed, a living product of nature stands in itself 'reciprocally as cause and effect.' Kant thereby made the way for a theory of 'self-organization' which found much attention in the twentieth century, in biology as well as in physics and sociology. The third critique contains the most important impulses for the following generation of German philosophers, particularly for Schelling and Hegel. These philosophers sought to overcome the subjective limitation of aesthetic and teleological judgements emphasized by Kant within the framework of an extensive objectivity of nature or of mind.

Once he had completed his critical works, the elderly Kant went about the business of drawing systematic consequences for particular areas of research. He only managed this, however, in the domain of practical action. Against the resistance of the Prussian Censor, Kant showed these consequences for religious belief (*Religion Within the Limits of Reason Alone*, 1793) and published in 1796/7 the *Metaphysics of Morals*, containing not only a 'doctrine of virtue' for self-responsible action but also doctrines of law and state. Together with the brief essay *Perpetual Peace*, this work exercised direct influence on the political opinions of Kant's contemporaries.

Kant put the productivity of his principle of critical justification to the test in numerous smaller writings on political philosophy and the philosophy of history, also concerning the concept of mankind or the relation among the sciences. Though he added no new model to the pre-existing political constructions of his time, he did develop a new principle of legitimization for political rule under the auspices of a conception of law and the idea of international treaty. He showed how, under general conditions of law, morality, and politics are connected through the medium of the principle of publicity without being derived from one another. He probably was the first to see that the action of states is not only to be justified with regard to the obligations to its own citizens, but is rather to be measured by 'cosmopolitan' standards which require international protection. Where once was the primarily religiously motivated idea of a Christian and European peace order, Kant placed the concept of an international peace order founded upon legal institutions—a concept upon which was drawn in the foundation of the League of Nations in the twentieth century.

Kant's theory of world peace was an innovation in political theory. It was based upon a historically and economically well-founded diagnosis of the state of mankind on the verge of the complete colonialization of the earth. There is no longer a single political entity on the globe which is independent from the political events in any other given part of the world. Because the surface of the earth offers only limited space, people can no longer make way for one another. War, the historical importance of which Kant recognized, can therefore no longer be a means for political ends. The French Revolution had shown that a people's claim to freedom is capable by itself of causing overdue changes in law. Hence Kant arrived at a compre-

hensive conception of the right, of morality and politics on the basis of 'human rights.' These rights, which guarantee freedom and equality for all, are to be made binding within states by means of a republican and representative constitution, and are to be secured among different states through a system of treaties. A stop is put to colonial politics by 'world citizens' rights'; at the same time, each individual citizen has the basic right to humane treatment, even in states other than his own. The individual states remain intact as indispensable units of social self-determination, but they are compelled by the dynamics of self-binding law to make internal reforms and to rely upon a federal alliance in their external relations. In his brief treatise on *Perpetual Peace*, Kant sketched a theory of politics appropriate to modern conditions of action through the combination of juridical, moral, rhetorical, and power-political elements.

Kant was prevented from completing his system of natural philosophy by numerous obligations which resulted from the broad public debate over his works. Inspired by the French Revolution, Germany experienced a euphoric decade of critique before the turn of the eighteenth century. Kant was thus exposed not only to a growing following, but also to a rapidly formed and primarily religiously motivated polemic. He gave proof of his literary ability in his equally polemical responses to this polemic. It is an ability which, in Kant's main works of critique, only reveals itself to those who have penetrated the difficult coherence of his thought.

In 1796 Kant ceased to teach and in 1801, at the age of 77, retired. His physical strength, already weak, deteriorated rapidly thereafter. Kant could hardly influence the edition of his lectures on anthropology, pedagogy, logic, and geography which were prepared by his students. Nevertheless, he continued working on the yet incomplete philosophy of nature. In doing so, Kant returned once again to the premises of his epistemological critique and conceived a unity of man and nature which bears a remarkable resemblance to the systems of Schelling and Hegel. While writing, Kant's strength suddenly failed him. The speculative notes break off with increasing regularity, and comments on everyday incidents are interspersed among Kant's notes. These writings, which today are collected and published under the title *Opus postumum*, document the way in which Kant's body gradually failed his mind.

Kant's final years were marked by a deterioration of physical and mental powers which was nevertheless born with dignity. On February 12, 1804 Kant died of old age in his city of birth. Despite numerous offers from other universities, Kant never left Königsberg. Yet he always took an active interest in world events, and kept company with businessmen, administration officials, and officers. He knew the travel literature of his time so well that one might think he had been to Java, America, or London himself. He was an enthusiastic supporter of the ideas of the French Revolution, both in the daily society gathered at his table and in public. In Kant's main *œuvre* of critique as well as in his wide-ranging biological, social, and political interests, one driving incentive of Kant's philosophy is manifest: the exploration of the possibilities and boundaries of man.

In his lectures on logic Kant often summarized the 'cosmopolitan importance' of philosophy in four central questions: (a) 'What can I know?' (b) 'What should I do?' (c) 'What may I hope for?' (d) 'What is man?' To the first question, Kant added by way of explanation, 'answer *metaphysics*; to the second, *morality* answers; to the third, *religion*; and to the fourth, *anthropology*.' The true meaning of this catalogue of questions is only recognized then when Kant adds that the last question encompasses all the others. In conclusion Kant writes that one could basically count metaphysics, morality, and religion to anthropology, 'since the first three questions are related to the last.' Thus Critical Philosophy appears to be, above all, a comprehensive inventory of the best human faculties—and these are the faculties of reason.

See also: Cognitive Psychology: History; Cognitive Science: History; Romanticism: Impact on Social Thought

Bibliography

Allison H E 1990 *Kant's Theory of Freedom*. Cambridge University Press, Cambridge UK

Allison H E 1983 Kant's *Transcendental Idealism. An Interpretation and Defense*. Yale University Press, New Haven, CT

Beck L W 1974 *Kants 'Kritik der praktischen Vernunft'. Ein Kommentar*. Fink Verlag, Munich, Germany

Cassirer E 1918 *Kants Leben und Lehre*. B. Cassirer, Berlin

Friedmann M 1992 *Kant and the Exact Sciences*. Harvard University Press, Cambridge, MA

Gerhardt V 1995 *Immanuel Kants Entwurf 'Zum ewigen Frieden' Eine Theorie der Politik*. Wissenschaffliche Buchgesellschaft, Darmstadt, Germany

Guyer P 1987 *Kant and the Claims of Knowledge*. Cambridge University Press, Cambridge, UK

Guyer P 1989 The unity of reason. Pure reason as practical reason in Kant's early concept of the transcendental dialectic. *The Monist* **72**: 139–67

Guyer P 1992 *The Cambridge Companion to Kant*. Cambridge University Press, Cambridge, UK

Höffe O 1983 *Immanuel Kant*. C H Beck. Munich, Germany

Kaulbach F 1969 *Immanuel Kant*. De Gruyter, Berlin

Kersting W 1984 *Wohlgeordnete Freiheit. Immanuel Kants Rechts- und Staatsphilosophie*. De Gruyter, Berlin/New York

Langer C 1986 *Reform nach Prinzipien. Untersuchungen zur politischen Theorie Immanuel Kants*. Klett-Cotte, Stuttgart, Germany

O'Neill O 1989 *The Constructions of Reason: An Exploration of Kant's Practical Philosophy*. Cambridge University Press, Cambridge

Prauss G 1973 *Kant. Zur Deutung seiner Theorie von Erkennen und Handeln*. Kiepenneus und Witsch, Cologne, Germany
Shell S M 1996 *The Embodiment of Reason: Kant on Spirit, Generation, and Community*. University of Chicago Press, Chicago
Sommer M 1988 *Identität im Übergang: Kant*. Suhrkamp, Frankfurt am Main, Germany
Wood A W 1978 *Kant's Rational Theology*. Cornell University Press, Ithaca, NY

V. Gerhardt

Kantian Ethics and Politics

Kantian ethics judges an action as morally good if it is done autonomously out of duty and its maxim could be the foundation of a general law. Neither the purpose nor the outcome of an action have any impact on its moral value. This is why Kantian ethics is normally seen as contrary to any type of consequentialist ethics. Politics, according to Kant, has always to be strictly under the rule of law as it works to achieve the happiness of the people as an end of its own. Implementing the rule of law also in international politics, Kant hopes that international relations will move slowly toward eternal peace. (Because it is not in all cases clear whether Kant means 'right' or 'law,' and because he sometimes even seems to make use of the double meaning of the German word *Recht*, it is more appropriate to use the word 'right' instead of 'law' from now on, as Mary Gregor does in her translation; Kant 1996.)

Kant draws a clear distinction between the fields of morality and legality. Hence, it is possible to discuss these topics separately, without denying the important connections that are due to the common origin of their key principles in (practical) reason (cf. e.g., Kersting 1993, 175ff).

1. *Ethics of Duty—Ethics of Autonomy*

Although, as with any other label, it is an oversimplification to reduce Kantian ethics either to an ethics of autonomy or to a mere ethics of duty, it might help in finding an initial approach to say that somehow it is both. The close connection between these two central concepts will be clarified later. However, they both signify that Kant does not develop an eudaimomian or any type of consequentialist ethics. Kant does not think that happiness is unimportant, as we will see later. But he deliberately rejects ancient eudaimonian ethics as a possible foundation for morality, and centers his postcritical works on moral philosophy around the concepts of duty, autonomy, and moral law (Forschner 1993, Wike 1994). Simplifying again, it seems helpful to characterize Kantian ethics within three steps, more or less identifying each with one of the central Kantian works on moral philosophy.

1.1 *Extracting and Vindicating the Categorical Imperative: The Groundwork of the Metaphysics of Morals* (*GMM*)

In his first larger postcritical work on ethics, starting from the statement that nothing can be called good without restrictions except a good will, Kant shows that a will is good if and only if it acts autonomously out of duty and according to the famous categorical imperative (CI). Everything other than a good will may be used with bad intentions. Even some of the virtues can be held by immoral people as the 'coolness of a scoundrel' proves to us (Kant 1996, p. 50). By saying that lack of success does not diminish the moral value of good will 'as something that has its full worth in itself' *as long as good will acts 'with its greatest efforts'* (Kant 1996, p. 50), Kant immediately answers some of the most tenacious objections to his ethics.

The main topic of the GMM is an attempt to clarify the relations between duty, autonomy, and the law of morals. First, he shows that the concept of duty is the best way to explain what is meant by a good will; particularly, how the moral worth of an action does not lie in its expected effect but in the principle of volition. An action is morally good if it is not only done in *conformity with duty* but *from duty* (Kant 1996, p. 53). Duty is defined as 'the necessity of an action from respect for law' (Kant 1996, p. 55) (according to the principle that I ought to act in such a way that my maxim could become a universal law) and respect is defined as a feeling that is 'not received by means of influence' but 'selfwrought by means of a rational concept and therefore specifically different from all feelings … that can be reduced to inclination or fear' (Kant 1996, p. 56). In Kant's view, the question of whether our maxims could become a universal law provides us, 'inexperienced in the course of the world' as we are, with an easy test of whether the maxim of our volition is morally good. This is what 'common human reason' has 'always before its eyes and uses as the norm for its appraisals,' although it 'admittedly does not think so abstractly in a universal form' (Kant 1996, p. 58). It is important to notice Kant's regard for common human reason and the morality of the common people, which is, according to his own testimony, due to the influence of Rousseau (Schneewind 1998, 487ff., Kant 1902, ff. XX, 44). Kant does not see the CI 'act only in accordance with that maxim through which you can at the same time will that it become a universal law' (Kant 1996, p. 73), as an invention of philosophers, but as a precise and short formula of how ordinary people think about morality. However, what he claims is that, until now, no philosopher has found a clear and precise description of what common reason thinks about ethics.

Nevertheless, common human reason might indeed find some difficulties in the abstractness of this account, just as philosophers until now. This can be seen from the comment of Friedrich Schiller who felt immoral because he loved his friends from inclination and not from duty. Hence, it should be mentioned that actions in conformity with duty that are motivated by inclination are not immoral. We just have no means to judge whether they are moral. Kant does, however, distrust inclination which he characterizes in the *Critique of Practical Reason* (*CPrR*) as 'blind and servile, whether it is kindly or not' (Kant 1996, p. 235).

On the other hand, as Kant explains in the outset of the second part of his *Groundwork*, we never really know whether we, or anybody else, acts from duty. If we look closely enough at all empirical actions, even at those 'from love of humankind' that are generally in conformity with duty, we 'come upon the dear self, which is always turning up' (Kant 1996, p. 62). Therefore, we require an 'isolated metaphysics of morals' that is made for all rational beings. Thus, there are the two results: (a) that morality is something for rational beings and (b) that a will is a good will not from love of mankind or any other inclination but from its accordance with the moral law. With good will reason itself becomes practical.

An ongoing subject that is discussed several times in this and in the later works is the relation of happiness to morality. Kant constantly argues that morality cannot be based on any principle about happiness. He has different reasons for saying this. First, because instinct is a much better guide to happiness than reason, it would have been an unnecessary and useless effort of nature to provide us with reason if it were for the sake of happiness. While this argument has been criticized for its obscure 'naturalism,' one may in fact agree with the view that happiness is a natural aim for rational beings, so it is useless to set the pursuit of happiness as a norm. According to another argument 'it is a misfortune that the concept of fortune is such an indeterminate concept that, although every human being wishes to attain this, he can still never say determinately and consistently with himself what he really wishes and wills' (Kant 1996, p. 70). Furthermore, because it is very difficult to predict the result of an action, it would be very difficult to decide whether it is right or wrong, whereas, in Kant's view, it is very easy to know how the categorical imperative handles this. However, it is important to notice that morality and happiness are by no means necessarily in conflict. They are just logically independent, insofar as happiness is the most general term for an object of the will, i.e., for the matter of the will, while the will's moral goodness is due to its form (*CPrR*, Kant 1996, p. 160). It is even a moral duty to preserve one's own happiness just to avoid the temptation to act immorally out of poverty or envy.

The rejection of happiness as the basic moral end in itself leads Kant to the characterization of the CI as the one and only unconditionally valid imperative. In contrast to the imperatives of skill and prudence (prudence teaches us how to achieve happiness), there is no external end that justifies the categorical imperative. Another result is the special status of *rational* beings as ends in themselves. Kant's hypothetical construction of a 'kingdom of ends' (*Reich der Zwecke*), which is somehow included in the concept of every rational being, may be the best tool to show how closely duty, autonomy, and the moral law are interrelated in his moral philosophy. In the kingdom of ends, every rational being, including God, i.e., every person, is seen as a lawgiver of the moral law. This is the reason why they are autonomous. Every finite rational being has the duty to follow this law. Every rational being is an end in itself and thus benefits from the moral law. Because of its deeply egalitarian character, obviously influenced by the Rousseauist view of the *volonté générale*, it might be better to follow Mary Gregor's suggestion for an alternative translation of *Reich der Zwecke* by 'commonwealth of ends' (Kant 1996, p. 83, on the 'moral force of reciprocity,' cf. Korsgaard 1996, 194f.). In the discussion of the few last decades on animal and medical ethics, the question has been raised whether the circle of those who benefit from the moral law should be expanded to beings that are not rational or only potentially rational.

Due to the inordinate reliance on the *Groundwork* in the attempt to clarify Kant's moral philosophy, the charges of formality and emptiness have been laid against Kant. An integral reading of his works shows, however, that this charge is misplaced. It even becomes clear that it would be misleading to label Kant a 'rule ethicist.' True, maxims can be tested as to their universality, but on the ground of maxims alone it is not always possible to decide what to do in a particular situation. Kant is perfectly aware of this, as his stressing of the importance of the faculty of practical judgment and his casuistical questions in the *Doctrine of Virtues* (*DV*) in the *Metaphysics of Morals* make clear. Before we deal with these issues, we must look at why the CI is valid or why we should be moral, in Kant's words, *how* the categorical imperative is possible (cf. Guyer 1998). Although Kant's main thrust on this issue is to be found in the *CPrR*, it is worth seeing how he deals with the question in the *Groundwork*.

1.2 The Moral Law, Free Will, and the Highest Good: The Critique of Practical Reason

In the third part of the *Groundwork* that points to a critique of pure practical reason, Kant himself poses the question: 'But why, then, ought I to subject myself to this principle and do so simply as a rational being, thus also subjecting to it all other beings endowed with reason?' (Kant 1996, p. 96). Here he answers this question by removing a 'kind of circle' (Kant 1996,

p. 97, Schönecker 1998). It seems that, on the one hand, we must be free from causal determination to be subject to the moral law and on the other hand, we must declare that we are subject to the moral law because we are free. However, freedom, i.e. autonomy, consists of the self-constraint of the will through the moral law.

Kant finds his solution by making the distinction between two 'standpoints.' In one, we are in the world of senses, having only cognition of appearances, never cognition of things in themselves. In the other, we are, especially in a practical context, members of the world of understanding (Kant 1996, p. 98). However, 'as a rational being, and thus as a being belonging to the intelligible world, the human being can never think of the causality of his own will otherwise than under the idea of freedom' (Kant 1996, p. 99) and this means submitting to the moral law. However, Kant admits that we can never explain *how* pure reason becomes practical (Kant 1996, p. 104). This seems to be the problem of his moral philosophy in general: either the 'good will' through which pure reason becomes practical is in the intelligible world of understanding. Then it is difficult to understand how it can be *my* will. Or it is *my* will, and then it is difficult to see how it can be completely separated from all my emotional attitudes. To talk of two standpoints does not help much while there can be no explanation of their inter-relationship.

The close connection between the freedom of will and the rule of the moral law—the first is what makes the law hold, the second what makes freedom known to us (*CPrR*, Kant 1996, p. 140)—is one of the main topics of the *CPrR* where Kant maintains several times that, with moral law, we have in practical contexts as a reality what could in theoretical philosophy only be proved to be not impossible in the solution of his 'third antinomy' in the *Critique of Pure Reason* (Kant 1902, ff. III, 366ff., cf. Wood 1999, 172ff., Ertl 1998).

In a comment in Sect. 7 of the chapter on the principles of pure practical reason, he explains the fundamental law of pure practical reason, i.e., the CI, as a 'fact of pure reason' (Kant 1996, 164f.) that is not empirically found. Although it is not quite clear what is meant by this cryptic formula, which is repeated several times within the *CPrR* and other Kantian works, sometimes with slight variations, it appears not to mean that something is revealed to pure reason or to us by pure reason but rather that the use of pure reason in practical contexts implies that we are under obligation by moral law.

Because happiness is the natural object of desire of human beings, i.e., of all rational and finite beings, whereas the moral law rules the formal structure of the will, happiness, like any other material content, cannot be the incentive for moral law. Kant explicitly rejects the moral relevance of any distinction between higher and lower pleasures. Nevertheless, Kant tries to reconcile the human wish for happiness with his concept of morality in his theory of the 'highest good' where he finds in the doctrine of Christianity (in contrast to the teachings of the Epicureans and the Stoics (Kant 1996, 241ff.)) a chance to give to those who are worthy of happiness the 'hope of some day participating in happiness to the degree that we have been intent upon not being unworthy of it' (Kant 1996, p. 244). To support this hope, he proposes in his doctrine of postulates not only the freedom of will but also the immortality of our soul that continues its way towards perfection even after our death and the existence of an almighty god who can and will give deserved rewards to those who lived in conformity with moral law. Perhaps one cannot be sure that a *postulated* god can give *real* hope to everybody and an incentive to follow the moral law. Perhaps, instead, some may find this incentive in the hope for glory or in the hope to contribute to the moral improvement of the world or in something similar.

1.3 Living the Categorical Imperative: The Doctrine of Virtues

In the *Doctrine of Virtues* (*DV*) of his late masterpiece *The Metaphysics of Morals*, Kant deals with the different types of duties that cannot be ruled by outer laws—duties of man concerning himself and his perfection and duties concerning the happiness of others. In the *DV* it becomes clear that for Kant there are obligatory ends for human beings: this indicates that his moral philosophy, which is surely not a variation of consequentialism, is not a pure *Gesinnungsethik* either, i.e., the outcome of an action does matter, although it is not always in our hands what this outcome might be. To be sure, that does not mean that the moral quality of an action is determined by anything else other than the quality of the motivation and the universality of the maxim in question. This includes the idea, however, that the agent must do everything he or she can to make morality (in the form of moral actions) a reality in the world. Because we have a duty to follow the CI, we have a duty to pursue our own perfection and the happiness of others. We also have duties toward ourselves. These are due, again, to our 'double existence' in the world of appearances and the intelligible world. We have duties toward ourselves as rational beings, i.e., toward the humanity in us (§ 3, Kant 1996, p. 544).

A second important achievement of the *DV* compared to earlier works is the casuistic element that demonstrates how Kant's ethics might be applied, at least in some cases. There are also considerations on the use of judgement (*Urteilskraft*) in the *CPrR* (Kant 1996, 194ff.). But here, 'the rule of judgment under laws or pure practical reason is this: ask yourself whether, if the action you propose were to take part, you could indeed regard it as possible through your will' (*CPrR*, Kant 1996, p. 196, on moral judgment see Herman

1993). In his later work, however, Kant uses the humanity formula, i.e., our duty to respect the moral law and our duty to respect every man as an end in himself to deal with particular subjects. Among these questions and subjects are suicide—that is to be seen as 'debasing humanity in one's person' and therefore morally forbidden (§ 6, Kant 1996, p. 546), as well as 'defiling oneself by lust' (Kant 1996, 548ff.), avarice (Kant 1996, 555ff.) and servility, 'self-esteem is a duty of the human being to himself' (§ 11, Kant 1996, 558), and also duties like beneficence and gratitude (Kant 1996, 571ff.) and vices like envy and malice (Kant 1996, 576ff.). He also deals with friendship (Kant 1996, 584ff.). Particularly in the 'casuistical questions' that are added in different places, one can find that Kant's views on these subjects are somehow strict. However, they are much more diversified than might be expected after reading the *Groundwork* and even the *CPrR*. Nevertheless, a remaining problem appears to be that Kant supplies no guidelines on how to interpret the moral law in 'hard cases.' This is partly due to his view that a law court, which he often uses as a metaphor for the internal judgment of the conscience (*DV* § 13; Kant 1996, p. 560) must *apply* the law via imputation, not *interpret* it, i.e., the court has to decide whether this individual case falls under the law, not to explain what the law means. This seems to have been a common view of eighteenth century continental legal theory. On the other hand, Kant may be the first philosopher who recognized that to apply a rule we need a faculty of its own, which he called the faculty of judgment, a discovery that is sometimes attributed to Wittgenstein's reflections in rule-following.

2. *Right and Politics*

According to Kant, any ethically acceptable politics has to take place under the rule of right. It is the duty of 'moral politicians' who are opposed to hypocritical 'political moralists' to do their best to bring constitutions of states and political relations between states 'into conformity with natural right, which stands before us as a model in the idea of reason' Toward Perpetual Peace, TEP, Kant 1996, p. 340). Nevertheless, there is an aim of politics on its own, i.e., the happiness of the citizen (TEP, Kant 1996, p. 351)—something that might be surprising to some of Kant's critics. Pivotal to Kant's legal and political philosophy is, therefore, first his treatment of right and how to determine its realm and, second, his hope on historical development towards peace and republican government. The key engine for bringing about legal progress is the public use of reason (Toward Perpetual Peace, TEP, Kant 1996, p. 351).

2.1 *Right as Compatibility of Freedom*

Defining right in the *Introduction to the Doctrine of Right* (*DR*) of his *Metaphysics of Morals* as the condition for the compatibility of one person's outer freedom with that of the other (Kant 1996, p. 387), Kant gives an account of every human being's innate right of freedom as 'independence from being constrained by any other's choice' (Kant 1996, p. 393). In fact, according to Kant, 'right is connected with an authorization to use coercion' (Kant 1996, p. 388) as long as this coercion is opposed to a wrongful 'hindrance to freedom in accordance with universal laws' and so is understood as 'a hindering of a hindrance to freedom' (Kant 1996, p. 388). 'Strict right' that is 'not mingled with anything ethical' can even 'be represented as the possibility of a fully reciprocal use of coercion that is consistent with everyone's freedom in accordance with universal laws' (Kant 1996, 388f.). The strict reciprocity of rightful coercion shows how crucial the egalitarian element is in Kant's legal philosophy, together with the emphasis on freedom from any other's arbitrary coercion.

Kant's doctrine of right begins with private right 'concerning what is externally mine or yours in general.' There have been vehement discussions within the few last decades about whether the fact that private right is treated before public right together with Kant's view that there is possession already in the state of nature (§ 9, Kant 1996, 409f.) justify looking at Kant as a possessive individualist (Saage 1994, Zotta 2000) or whether one should stress his statement that this possession is only provisional and 'conclusive possession' is possible 'only on a law of a common will' (§ 9, Kant 1996, p. 410), such that Rousseauist and therefore egalitarian parts would gain more weight. Anyway, Kant rejects Locke's founding of property on work, because work is something 'accidental' to the land that cannot decide whose property it is (§§ 15, 17), and bases provisional possession on first occupation, conclusive possession on the common will (cf. Williams 1983, p. 94, cf. Kersting 1993, 337ff.). Perhaps one might say that although Kant's primary concern is with individual freedom—according to his definition of right—he nevertheless has a sense for public welfare and distributive justice (Rosen 1993).

Because in the state of nature property and any right are insecure 'however well disposed … human beings may be' (§ 44, Kant 1996, p. 456), we have the moral and legal duty to enter a state of public right (§ 42). What Kant describes in his *DR* is 'the state in idea' that 'serves as a norm for every actual union into a commonwealth' (§ 45, Kant 1996, p. 457). Kant clearly prefers a republic with division of powers that is the only alternative to despotism and 'any true republic is and can only be a system representing the people' (DR § 52, Kant 1996, p. 481) which is obviously incompatible with the absolutist and feudalist systems of his times. Therefore, there is another discussion on whether he really rejects every resistance against unjust regimes so bluntly as it looks like in the published works (DR § 49 A, Kant 1996, 463f., Of the Common Saying: That May Be Correct in Theory, but is of no Use in Practice,

OCS, Kant 1996, p. 303) or whether there might be some more or less hidden hints that he sees it in a more nuanced manner (Beck 1993, Westphal 1992). Anyway, it seems to be clear that Kant prefers reform of the constitution (Goyard-Fabre 1996, 217f.) and for this he claims again and again the transcendental principle of publicity, of freedom of the pen (OCS, Kant 1996, p. 302).

According to an interesting and attractive interpretation, Kant accepts the political systems of his time—and perhaps he would do the same with those of our time—mainly because they provide us with the possibility of a change toward more just societies that are closer to the state in the idea (Brandt 1982, Kersting 1993). But this change has to be done by the sovereign via reform and not by the people via revolution (DR § 49 A, Kant 1996, p. 465). As we have to accept the state of affairs in the distribution of property at least provisionally for being able to enter a civil state with distributive justice (DR § 41), we have to accept the actually existing legality for being able to transform it by rightful means until it comes close to a true republic. On an international level, we have to accept large parts of the actual distribution of power for being able to approach eternal peace via a commonwealth of free republics. The common tool for this treatment of nonperfect social situations is the permissive law that allows us to live in circumstances that are contrary to the principles of right as long as a quick change by means of right is impossible and as long as there is a serious effort to improve the state of affairs.

2.2 Toward Eternal Peace

One of the most influential and effective elements in Kant's political philosophy might be his proposal of an international legal order that could be able to guarantee eternal peace that is not only 'a suspension of hostilities' (TEP, Kant 1996, p. 317). There had been considerations how to realize a peaceful world through a council of the sovereigns like the *Projet pour rendre la paix perpétuelle en Europe* of the Abbé St. Pierre. What Kant does—again under the influence of Rousseau—is to connect the international order with the inner structure of the state. His first 'definitive article for perpetual peace' therefore says that 'the civil constitution of every state shall be republican' (TEP, Kant 1996, p. 322) and the second that 'the right of nations shall be based on a federalism of free states' (Kant 1996, p. 325). This 'surrogate' for a world-republic—which seems to be unachievable under current conditions—provides a way to maintain the sovereignty of the participating states with the rule of right, the only alternative to the unacceptable rule of war, 'right cannot be decided by war and its favorable outcome' (Kant 1996, 327f.).

On the other hand, Kant is not a categorical pacifist. He accepts that wars had played an important role in the development of mankind (*Idee zu einer allgemeinen Geschichte in weltbürgerlicher Absicht*, Kant 1902, ff., VIII, 24ff.), and he would prefer the warlike state of nature to a repressive world monarchy (TEP, Kant 1996, p. 336). He is not an unrealistic dreamer who expects world peace within a short time. Anyway, as Doyle (Doyle 1986) and others have shown, it was not an unrealistic assumption that democratically governed states tend much less to bellicose behavior than others. Kant rather thinks, or hopes, that with the help of the 'great artist nature' (*die große Künstlerin Natur*, TEP, Kant 1996, p. 331), human societies will move slowly but steadily towards the federation of free republics guaranteeing eternal peace. Kant's idea seems to be that even immoral actions ultimately lead to an improvement of the legal order without discharging the culprit of his responsibility, however. Somehow behind the back of the agents who are looking for their personal advantage, the artist nature works in a way like Smith's 'invisible hand' towards this federalism of republics that might even turn into a world republic. Anyway, this historical development prepares conditions under which persons may act morally much more easily than it is for them now in a situation that is closer to the state of nature.

Kant does not consider this 'mechanism of nature' (ib.) as a natural law. Rather it is a projection of practical reason, a hope we must have because we have a duty to move toward eternal peace. Elsewhere, Kant says, if we can't prove that a development will take place but can't prove the contrary either, a need of reason (*Vernunftbedürfnis*) gives us the right to expect what seems more rational (Kant 1902, ff. VIII, Kleingeld 1995, p. 94).

See also: Ethics and Values; Justice: Philosophical Aspects; Kant, Immanuel (1724–1804); Law and Morality: A Continental-European Perspective; Law and Morality: An Analytic Perspective; Moral Reasoning in Psychology; Political Science: Overview; Political Thought, History of; Research Ethics: Research; Virtue Ethics

Bibliography

Beck L W 1993 Kant's two conceptions of the will in their political context. In: Beiner R, Booth W J (eds.) *Kant & Political Philosophy. The Contemporary Legacy*. Yale University Press, New Haven, CT

Brandt R 1982 Das Erlaubnisgesetz, oder: Vernunft und Geschichte in Kants Rechtslehre In: Brandt R (ed.) *Rechtsphilosophie der Aufklärung*. de Gruyter, Berlin

Doyle M W 1986 Liberalism and world politics. *American Political Science Review* **80**: 1151–69

Ertl W 1998 *Kants Auflösung der 'dritten Antinomie.'* Alber, Freiburg

Forschner M 1993 Moralität und Glückseligkeit in Kants Reflexionen. In: Forschner M (ed.) *Über das Glück des*

Menschen. Wissenschaftliche Buchgesellschaft, Darmstadt, pp. 107–50
Goyard-Fabre S 1996 *La philosophie du droit de Kant*. Librairie philosophique J. Vrin, Paris
Guyer P 1998 The possibility of the categorical imperative. In: Guyer P (ed.) *Kant's Groundwork of the Metaphysiscs of Morals*. Rowman & Littlefield, Lanham, MD
Herman B 1993 *The Practice of Moral Judgment*. Harvard University Press, Cambridge, MA
Höffe O (ed.) 2000 *Grundlegung zur Metaphysik der Sitten. Ein kooperativer Kommentar*. Vittorio Klostermann, Frankfurt/Main, Germany
Kant I 1902 *Kants gesammelte Schriften*. Königlich Preußischen Akademie der Wissenschaften (ed.). Walter de Gruyter, Berlin
Kant I 1996 *Practical Philosophy* [trans. Gregor M]. Cambridge University Press, Cambridge, UK
Kersting W 1993 *Wohlgeordnete Freiheit*. Suhrkamp, Frankfurt/Main, Germany
Kleingeld P 1995 *Fortschritt und Vernunft: Zur Geschichtsphilosophie Kants*. Königshausen & Neumann, Würzburg, Germany
Köhl H 1990 *Kants Gesinnungsethik*. de Gruyter, Berlin, Germany
Korsgaard Ch M 1996 *Creating the Kingdom of Ends*. Cambridge University Press, Cambridge, UK
Rosen A D 1993 *Kant's Theory of Justice*. Cornell University Press, Ithaca, NY
Saage R 1994 *Eigentum, Staat und Gesellschaft bei Immanuel Kant*. Nomos Verlagsgesellschaft, Baden-Baden, Germany
Schneewind J B 1998 *The Invention of Autonomy. A History of Modern Moral Philosophy*. Cambridge University Press, Cambridge, UK
Schönecker D 1998 *Kant: Grundlegung III. Die Deduktion des kategorischen Imperativs*. Alber, Freiburg
Westphal K 1992 Kant on the state, law, and obedience in the alleged 'Antirevolutionary Writings.' *Journal of Philosophical Research* **XVII**: 383–427
Wike V S 1994 *Kant on Happiness in Ethics*. State University of New York Press, Albany, NY
Williams H 1983 *Kant's Political Philosophy*. St. Martin's Press, New York
Wood A W 1999 *Kant's Ethical Thought*. Cambridge University Press, Cambridge, UK
Zotta F 2000 *Immanuel Kant. Legitimität und Recht*. Alber, Freiburg

M. Kaufmann

Key, Valdimir Orlando (1908–63)

1. Biographical Summary

Valdimir Orlando Key, Jr. was born in Austin, Texas on 13 March 1908. He spent his childhood in the very small West Texas town of Lamesa, later obtaining a BA from the University of Texas at Austin in 1929, and an MA degree at the same institution in 1930. V. O. Key, Jr. then moved to the University of Chicago—then perhaps the premier institution for studying American politics in the United States—and received his doctorate there in 1934. His early professional years were devoted to teaching and, later, to work in Washington for the Social Science Research Council and, subsequently, as part of the technical staff of the National Resources Planning Board during the heyday of the New Deal. Further Washington experience came his way during World War II with a three-year stint at the US Bureau of the Budget (now the Office of Management and Budget). His academic career took him first to Johns Hopkins University and thence to Yale and Harvard Universities, the latter from 1951 until his untimely death in Cambridge, Massachusetts, on 4 October 1963. The long-standing illness to which he eventually succumbed did not prevent him from moving to the very front rank of American political scientists of his generation. The reasons for this fame—a reputational survey of 1965 by A. Somit and J. Tanenhaus placed him then in a clear first position—will be concisely discussed below.

2. Scientific Contribution

As a Southerner, and at a time when the South was far more a region apart from the rest of the American polity than it is now, Key proved the ideal scholar to provide a comprehensive description and analysis of the South's *apartheid* political order, and its massive influence on national politics, very shortly before its demise. This study, *Southern Politics in State and Nation* (1949), provided a detailed portrait of politics in each on the eleven former Confederate States at the time of writing, as well as analyses of the region's influence on congressional politics; and—an important contribution in its own right—an extended summary statement on the nature and consequences of one-party systems for the representation of public opinion through elections. The craftsmanship is meticulous and innovative, with a mixture of 'thick description' of persons and forces at work based on extensive interviewing, along with the skilful deployment of charts and maps based upon aggregate returns from referenda and primary and general-election returns. More than half a century later, *Southern Politics* still repays the reader's attention.

Quite a lot of Key's basic approaches to the study of American politics, and his personal value commitments, can be found in these pages. First, and obviously, lies the exceptional historic and contemporary importance of 'sectionalism' (major regional differences based on economy, political subculture and experience) in the study of American politics. This point is worth stressing, since with the advent of national scientific surveys of American voters and their behavior in and after the 1960s, this dimension strikingly faded from view in Americanist political science for a protracted period. Secondly, *Southern Politics* underscored the importance of history and

historical experience in shaping the confines of today's electoral politics. It is thus perfectly appropriate to view Tennessee politics in the 1940s (and later) in light of the polar oppositions that were displayed in a June 1861 referendum on the state's secession from the Union. This concern to cite present reality in its developmental context was also, however, to run counter to mainstream preoccupations within political science, as the quest for 'scientific regularities' variously led to embracing concepts and models derived from social psychology or economics coming to the fore.

Finally, Key's democratic value commitments, which informed his entire published output across his professional lifetime, are presented in clear view in this magisterial work. Several points about how these views were expressed deserve mention here. Key made no pretension to 'grand theory,' and imported no elaborate conceptual schemes from great Europeans such as Marx or Weber (see *Marx, Karl (1818–89)*; *Weber, Max (1864–1920)*). Like many in his and a somewhat earlier generation of American political scientists, his intellectual focus was American to the core, and concentrated on the gap between the democratic promise embedded in the national political culture on one hand, and real-world antidemocratic practices and realities on the other. These latter obviously abounded in the Southern politics he studied, and this gap represented the chief analytic problematic of his lifetime. As he demonstrated in great detail, in the kind of disorganized and exclusionist politics of his South the have-mores profited and the have-lesses lost out as a matter of empirical fact.

At the same time, Key's was a mind which was strikingly subtle and nuanced. Not for him were the sledgehammer blows of Beard (1913), or other notable writers in the American Progressive tradition. Rather, he very carefully spelled out the structure of a given state of political affairs, equally carefully pointed out the vastness of the gap between ideals and empirical reality, and invited his readers to reflect on this gap and how it might be overcome. He was an optimist about the extent this could be accomplished in time—as scholars shaped by the political crises of the 1960s and 1970s often were not. Thus, his hopes for his native region rested to a significant degree on the longer-term effects of industrialization and urbanization on its society and politics. In retrospect, these hopes were to be realized, though only in the context—he was no prophet, after all—of a massive federal intervention in the 1960s, the 'Second Reconstruction,' fuelled by the civil-rights revolution; and, not least, by the USA's foreign-policy requirements in a postcolonial world.

His later major studies displayed his awareness that there were other impediments to the realization of his ideal, the expansion of the electorate's effective role in politics through the instrument of party competition and political choice. In *American State Politics* (1956), Key devoted his attention to several of these impediments: the impact of direct primaries and other regulations of party activities on party organization, including the multiplication of uncontested seats in legislative bodies; what he calls the 'lottery of the long ballot,' i.e., the separate election of a wide range of state office-holders; and the shifting of major state elections away from coincidence with presidential elections and 'national tides' to off years, where, incidentally, voter turnout is nearly always substantially smaller than in presidential years. Key was clearly concerned about the implications of such developments for the integrity of parties as vehicles for individual and collective choice at elections. As Table 1 below shows on just one of these dimensions, his forebodings have been more than amply realized in the long generation since his death.

Table 1
Insulation effects: states having at least some gubernatorial elections coinciding with presidential elections, 1936–2000

Year	No. of states	Total no. of US reps. in these states	Total US reps. in these states (%)
1936	35	313	72.0
1956*	29	225	51.7
1984	13	58	13.3
2000	11	52	12.0

* Year of publication of *American State Politics*

In his last years, V. O. Key increasingly devoted his attention to the challenges he saw in the development of survey research, both academic and applied. This was to issue in two works, *Public Opinion and American Democracy* (1961) and—posthumously—*The Responsible Electorate* (1966), and a number of articles where criticism was perhaps more pointed. He was vigorously opposed to political-science approaches that appeared to eliminate politics from their analytic schemes. He was alarmed by the potential for manipulating the public that was already (if by today's standards dimly) visible in applied 'market research' by candidates and their media and polling staffs. For, in his view, public opinion ultimately was an 'echo chamber'; and the quality of its output was heavily dependent upon the input provided by political elites and their consultants. Finally, he was convinced that the survey-research portrait of voters as in the main know-nothings (or know-very-littles) drastically underestimated voters' practical political intelligence and capacity for purposive, rational pursuit of political choice in accordance with their interests. On one major point raised in *The Responsible Electorate*—the distinction between interelection 'switchers' and self-identified partisan independents—his view was later to receive substantial confirmation in a 1981 study of panel surveys by Morris Fiorina (1981). In any

event, Key's last work made its own real contribution to the subsequent explosion of 'issue-voting' studies and, eventually, to a much more sophisticated and politically-centered survey-research enterprise.

3. *Assessment of Key's Professional Impact and Current Importance*

As the consummate professional that he was, V. O. Key devoted himself to the well-being and improvement of the political-science profession in countless ways. One example, at a time when statistical knowledge within the profession was still pretty much at a primitive level, was his little 1954 volume, *A Primer of Statistics for Political Scientists*. Another was his textbook, *Politics, Parties and Pressure Groups* (Crowell, NY, 1942–1964), which was far the best in its field and extended his teaching to generations of students. He sought to expand the frontiers of knowledge, not only through these and other works, but through myriad contacts with his peers and service to such organizations as the Social Science Research Council and the American Political Science Association.

It may be said that, for two decades or so after his death, Key remained more revered than followed by the political-science mainstream. His was always a sterling example of outstanding craftsmanship in specifying and executing important research projects that provided new insights into the American electoral process; of an enormous mastery of the data of American politics past and present; and a most exceptional capacity to break new ground by using it in novel ways. At the same time, regionalism and history largely fell out of mainstream concerns; his older South rapidly became, or seemed to become, a museum piece; and for a long time, analysis of aggregate election data seemed to be under W. S. Robinson's 1950 ban on the uses of ecological correlation. Since then, as problematics and professional interests have shifted somewhat, regionalism and historical depth in the explanation of current states of affairs have seen a notable comeback, and a whole field, American Political Development, has emerged into view. At the same time, ecological regression has emerged as a valuable estimation technique for analyzing voter transitions, e.g., during the critical realignment of the 1850s.

In a real sense, Key's work is more significant today than it seemed to be for some years after his death. Nowhere is this more true than in one major area of research by political scientists and historians alike: the study of critical realignment as a major change process across the span of American electoral, and more broadly political, history. As of 1991 (the last formal count), more than 500 books and articles had been devoted to this subject. While some historians had recognized qualitatively the exceptional turning-point significance of such elections as 1800, 1828, 1860, 1896, and 1932. Key was the first to provide an examination of this as an empirically specifiable problem in his truly seminal 1955 article, 'A theory of critical elections' (1955). It is noteworthy that Key never saw fit to explore the subject further, and one can doubt that he would find the elaboration of models in this field always to his taste. But his was the pathbreaking innovation that truly founded a vast research enterprise that is still underway, and has also attracted the attention of non-American researchers as well (Pierre Martin, 2000).

Bibliography

Bensel R F 1980 *Sectionalism and American Political Development*. Madison, WI

Cummings M C (eds.) 1988 *V. O. Key, Jr. and the Study of American Politics*. Washington

Fiorina M P 1981 *Retrospective Voting in American National Elections*. New Haven, CT

Gienapp W E 1987 *The Origins of the Republican Party, 1852–1856*. New York

Key V O 1949 *Southern Politics in State and Nation*. Knopf, New York

Key V O 1955 A theory of critical elections. *Journal of Politics*. **17**: 3–18

Key V O 1956 *American State Politics*. New York

Key V O 1959 Secular realignment and the party system. *Journal of Politics*. **21**: 198–210

Key V O, Munger F 1959 Social determinism and electoral decision: The case of Indiana. In: Burdick E, Brodbeck A (eds.) *American Voting Behavior*. Free Press, Glencoe, IL, pp. 281–99

Key V O 1961 *Public Opinion and American Democracy*. New York

Key V O 1961 Public opinion and the decay of democracy. *Virginia Quarterly. Rev*. **37**: 481–94

Key V O 1966 Cummings M C (ed.) *The Responsible Electorate*. American Political Science Association, Cambridge, MA

Kleppner P 1987 In Appendix: Realignment theory after Key. *Change and Continuity in Electoral Politics, 1893–1928*. Westport, CT, pp. 239–50

Martin P 2000 *Comprendre les évolutions électorales: La théorie des realignements revisitée*. Paris

Natchez P 1985 Images of Voting/Visions of Democracy. New York

W. D. Burnham

Keynes, John Maynard (1883–1946)

John Maynard Keynes was the most influential economist of the twentieth century. He invented, almost single-handedly, a new branch of economics, macroeconomics, or the theory of what determines output as a whole, as contrasted with microeconomics which studies the forces determining the composition

of a *given* output. He did so in order to destroy the implicit assumption of 'classical' economics that a competitive market economy would always ensure full employment; and thus to justify government intervention to maintain full employment. The task seemed the more urgent in the wake of the Great Depression of 1929–1933. The result was *The General Theory of Employment, Interest and Money* (hereafter GT), whose publication in 1936 is conventionally taken to be the start of the Keynesian Revolution. Its main policy fruit was the commitment, till the mid-1970s, by most Western governments, to maintain high and stable levels of employment in their countries. The lack of secure microfoundations for Keynesian macrotheory, as well as the somewhat contentious implications of that theory for the role of governments, has continued to make Keynesian economics controversial; while economic historians have disputed the contribution of Keynesian policy to the 'golden age' of global prosperity running from the 1950s to the early 1970s.

1. *His Life*

John Maynard Keynes was born at 6 Harvey Road, Cambridge, England on June 5, 1883 into an academic family. He was the eldest son of John Neville Keynes, a logician and economist, and Florence Ada Brown, the daughter of a Congregational divine. He had an outstandingly successful school career at Eton College which was followed by an equally glittering undergraduate one at King's College, Cambridge, where he gained a first-class honours degree in Mathematics, wrote papers on the medieval theologian Peter Abelard and the Conservative political philosopher Edmund Burke, and became President of the Cambridge Union and the University Liberal Club. Of crucial importance to his intellectual and moral formation was his election, in 1902, to the Cambridge Apostles, an exclusive 'conversation' society, where he fell under the influence of the philosopher G. E. Moore, and which brought him the friendship of Lytton Strachey. Moore's *Principia Ethica* remained his 'religion under the surface' for the rest of his life. It taught that the highest forms of civilized life were friendship, aesthetic enjoyments, and the pursuit of truth.

In 1906 Keynes was placed second in the Civil Service Examination, his worst marks being in economics, which he had studied briefly under Alfred Marshall (see *Marshall, Alfred (1842–1924)*). After two years in the India Office, in which he wrote a thesis on probability in his spare time, he started lecturing on monetary economics at Cambridge; in 1909, his thesis won him a fellowship at King's College, which remained his academic home for the rest of his life. His membership of the Bloomsbury Group, a commune of Cambridge-connected writers and painters who lived in the Bloomsbury district of London, dates from the start of his friendship with the painter Duncan Grant, Lyttton Strachey's cousin, in 1908. In 1913 he published his first book, *Indian Currency and Finance*, and served on the Royal Commission on Indian Finance and Currency.

Keynes helped to avert the collapse of the gold standard in the banking crisis of August 1914 which accompanied the outbreak of the first world war. From January 1915 to June 1919 he was a temporary civil servant at the Treasury, showing a notable ability to apply economic theory to the practical problems of war finance. He was against military conscription, and would have been a conscientious objector had his Treasury work not exempted him from military service. When Lloyd George succeeded Asquith as Prime Minister, Keynes (in January 1917) became head of the Treasury's new 'A' Division, set up to manage external finance. He helped to build up the system of Allied purchases in external markets, while chafing at Britain's growing dependence on American loans and the failure to arrange a compromise peace. Keynes was chief Treasury representative at the Paris Peace Conference in 1919, where he tried unavailingly to limit Germany's bill for reparations, and to promote an American loan for the reconstruction of Europe. His resignation from the Treasury on June 5, 1919 was followed by the publication, in December, of *The Economic Consequences of the Peace*, the book which first brought him international fame. A bitter polemic, informed by both moral passion and economic argument, against the Allied policy of trying to extort from Germany an indemnity which it could not pay, it reflected his revulsion against Lloyd George's leadership in both war and peace, and his fears for the future of European civilisation. Unless the Versailles Treaty were drastically revised, 'vengeance, I dare predict, will not limp' (Keynes 1919, p. 170)

Between the wars, Keynes's life was divided between Cambridge, London, and East Sussex. He was a spectacularly successful investment Bursar of King's College, and despite some major reverses, made about $400,000 for himself over his lifetime—worth $12m. today—out of which he financed a fine collection of pictures and rare books and the building of the Cambridge Arts Theatre in 1935. In London, where he rented a house at 46 Gordon Square, he was, at various times, on the Boards of five investment and insurance companies, the chief one being the National Mutual Life Assurance Company, which he chaired from 1921 to 1937. Between 1923 and 1931 he was chief proprietor and chairman of the Board of the weekly journal, the *Nation and Athenaeum*, contributing regular articles on financial and economic topics. (He remained chairman of the Board of the *New Statesman and Athenaeum* when the two journals merged in 1931.) Between 1911 and 1945, he edited the *Economic Journal*. In the 1920s his ideas on economic policy permeated Whitehall through monthly meet-

ings of the Tuesday Club, a dining club started by his friend and stockbroker, Oswald Falk. In the 1930s, he sought to influence policy through his membership of the Prime Minister's Economic Advisory Council. In 1925, he took the lease of Tilton, a farmhouse in East Sussex, next door to Charleston, where Duncan Grant lived with the painter Vanessa Bell. This move coincided with his marriage to the Russian ballerina, Lydia Lopokova, who gave his life the emotional stability it had previously lacked, and which provided the necessary background for sustained intellectual effort.

In the 1920s, the postwar European inflations, succeeded in Britain by heavy unemployment, formed the background to his two theoretical books *A Tract on Monetary Reform* (1923) and *A Treatise on Money* (1930) which dealt with the causes and consequences of monetary instability, and their remedies. These theoretical exercises were punctuated by two notable polemical pamphlets *'The Economic Consequences of Mr. Churchill'* (1925) and *'Can Lloyd George Do It?'* (1929), written with Hubert Henderson. The first attacked Churchill's decision, as Chancellor of the Exchequer, to put the pound back on the gold standard at an overvalued exchange rate against the dollar; the second was a plea for a large public-works program. Reconciled to Lloyd George in 1926, Keynes attempted to provide the Liberal Party with a social philosophy of the 'middle way' between individualism and state socialism, suitable for an inflexible industrial structure. Regulation of demand, he would later write, was the only way to maintain capitalism in conditions of freedom (Keynes 1936, p. 381, 1940, p. 123).

The Great Depression of 1929–1932, together with technical flaws in *A Treatise on Money*, took Keynes back to the theoretical drawing board. What now seemed to be needed was not an explanation of Britain's 'special problem' of persisting unemployment, but an explanation of how aggregate output could collapse, labor remain unused, and global depression persist in a world in which resources remained scarce. From the autumn of 1931 to the summer of 1935, Keynes worked on a new book of theory, initially entitled '*The Monetary Theory of Production*.' He was helped not just by older economists like Ralph Hawtrey and Dennis Robertson, but by Roy Harrod and a 'Cambridge Circus' of young disciples led by Richard Kahn. There was just one major pamphlet in these years, '*The Means to Prosperity*,' written in June 1933. On a trip to the United States in 1934 to study the New Deal first-hand, he wrote: 'Here, not in Moscow, is the economic laboratory of the world.'

The GT, published in February 1936, tried to demonstrate both that 'underemployment equilibrium' was logically possible, and how monetary manipulation by the central bank combined with an extensive 'socialisation' of investment could maintain full employment. The first proposition, in particular, divided the economics profession, since it rejected the 'classical' thesis of a self-equilibrating economy. The publication of the GT confirmed the breach between Keynes and Dennis Robertson and Hubert Henderson, his chief collaborators in the 1920s; at the same time it marked the birth of a 'Keynesian school' of economics led by Richard Kahn and Joan Robinson (see *Robinson, Joan (1903–83)*) at Cambridge, Roy Harrod and James Meade at Oxford, and Nicholas Kaldor at the London School of Economics. In the United States, the GT supplied the younger generation of (mainly Harvard-trained) economists with a theoretical rationale for the New Deal. Keynes himself joined in the fierce controversies which his book generated, even though he was severely incapacitated, from May 1937 to March 1939, with heart disease. As another European war, and with it, the return to deficit-financed full employment became increasingly likely, Keynes sought to win acceptance for his Revolution by showing how the managment of aggregate demand to avert depression could just as easily be used to contain inflation in a war economy.

The upshot was his pamphlet '*How to Pay for the War*' (1940), which won the approval of his arch critic Friedrich Hayek (see *Hayek, Friedrich A von (1899–1996)*), and whose logic influenced Kingsley Wood's war budget of 1941. Restored to a semblance of health by his doctor, Janos Plesch, Keynes himself returned to the Treasury in August 1940 as unpaid adviser to the Chancellor of the Exchequer, and remained its dominating force for the rest of his life. Elevated to the House of Lords as Baron Keynes of Tilton, his influence was felt in the Beveridge Report on Social Security in 1942, and in the Employment White Paper of 1944 which pledged the government to maintain a 'high and stable level of employment' after the war. In 1942, Keynes became chairman of the Council for the Encouragement of Music and the Arts (CEMA), a wartime innovation which, transformed into the Arts Council of Great Britain shortly before he died, inaugurated permanent state patronage of the arts.

The American demand that, in return for Lend-Lease, Britain scrap its imperial preference system after the war, inspired Keynes to his last great constructive effort, his plan for an International Clearing Union (1942). This was designed to shift balance-of-payments adjustment from debtor to creditor countries, so as to avoid the externally-generated deflationary shocks which had spread depression under the gold standard. The Bretton Woods Agreement of 1944, which set up a system of fixed, but adjustable exchange-rates and two new institutions, the International Monetary Fund and the International Bank for Reconstruction and Development, fell short of Keynes's hopes. The abrupt US cancellation of Lend-Lease in August 1945 led him to undertake the fifth of his six Treasury missions to the United States to secure an American grant or interest-

free loan of $5bn. Forced to accept a semi-commercial loan for an amount far less than he had requested, Keynes gave a brilliant defence of his policy in his last speech to the House of Lords on December 7, 1945.

On April 21, 1946, worn out by his labors, he suffered a fatal heart attack at Tilton, a little short of his 63rd birthday. In an imposing memorial service at Westminster Abbey, Clement Attlee, the prime minster, headed a list of mourners drawn from all walks of life.

2. *Major Contributions*

Keynes's main contributions to economics can be grouped under three heads: the theory of deep economic fluctuations; the theory of effective demand; and the technique of stabilization policy based on the second.

2.1 *The Theory of Economic Fluctuations*

This cannot be wholly understood apart from Keynes's distinctive theory of probability. In his *A Treatise on Probability* (1921), his main contention was that probability is not statistical, but is a logical relationship between the premise and conclusion of an argument which fell short of demonstrative proof. Like Frank Knight (who however used a statistical theory of probability), Keynes distinguished between risk, or situations when agents can form probabilities and thus insure against loss, and uncertainty, when 'there is no scientific basis on which to form any calculable probability.' He wrote in 1937: 'Thus the fact that our knowledge of the future is fluctuating, vague and uncertain, renders wealth a peculiarly unsuitable subject for the methods of the classical economic theory' (Keynes 1937, p. 113). This may be contrasted with the modern mainstream view that agents maximise with known probability distributions, so that market-clearing models are the most appropriate for analyzing economic behavior. Keynes himself identified as a tacit axiom of the classical theory of the self-regulating economy that 'at any given time facts and expectations were … given in a definite and calculable form' (Keynes 1937, p. 112).

Keynes's theoretical work in the 1920s is devoted to unpackaging the Quantity Theory of Money (QTM) which ruled out, by assumption, the possibility that monetary disturbances can have 'real' effects on the economy. This might be true in the long-run, Keynes remarked acidly, 'but in the long-run we are all dead' (Keynes 1923, p. 65). Keynes used the quantity of money *equation* to show that changes in the quantity of money can unsettle business expectations and the distribution of income, thereby causing short-run fluctuations in the level of business activity. In doing so, he emphasised the use of money as a 'store of value' or hedge against uncertainty, and the distinction between fixed and flexible prices. Thus in the *Tract on Monetary Reform* he wrote: 'The *fact* of falling prices injures entrepreneurs; consequently the *fear* of falling prices causes them to protect themselves by curtailing their operations' (i.e., adding to their cash balances) (Keynes 1923, p. 34). In the *Treatise on Money*, he argued that the disturbing effects of money can arise not just from inflationary or deflationary policies, but from *autonomous* changes in business sentiment. Much of the *Treatise* is devoted to analyzing the consequences of short-run changes in the 'propensity to hoard' or 'velocity of circulation,' which Keynes identifies, by means of his 'Fundamental Equations,' with the emergence of windfall profits and losses, and a disequilibrium between saving and investment. Keynes shows how disequilibrium prices produce an oscillation of boom and slump round a (notional) full-employment equilibrium. But such equilibrium is achieved accidentally, in the absence of appropriate monetary policy.

The chief theoretical point which emerges from these two works is the instability in what Keynes would later call the liquidity-preference schedule. His claim was that, in a variety of situations readily encountered in economic life, hoarding money or savings can yield superior utility to investing them, and that this, combined with 'sticky' wage and price contracts, can explain quite prolonged lapses from full employment. His contribution up to1930 was largely a summation of the monetary theories of Irving Fisher (see *Fisher, Irving (1867–1947)*) and Knut Wicksell, and is properly regarded as a contribution to the theory of economic fluctuations, with increased emphasis on volatile expectations. Like Fisher, Keynes advocated a monetary policy of keeping domestic prices stable; this might be incompatible with adherence to the international gold standard, an institution which Keynes dismissed as a 'barbarous relic' (Keynes 1923, p. 138).

2.2 *The Theory of Effective Demand*

Keynes's revolutionary contribution to economics lies not in his theory of business fluctuations but in his theory of how aggregate output is determined. This is the subject matter of his GT. The QTM was a theory of the price level, when the levels of output and employment were explained in real, nonmonetary terms. Following a suggestion from Ralph Hawtrey, Keynes agreed in 1930 that 'it will probably be difficult in the future to prevent monetary theory and the theory of short-period supply from running together' (Keynes 1930, p. 146). This was the start of macroeconomics.

The core of the GT is the theory of effective demand. It tells how, when something happens to reduce aggregate demand, aggregate supply and aggregate

demand are reconciled, not through falls in the price-level of output, but by a reduction in its quantity. It is a short-period theory: prices adjust only after, and as a result of, the fall in output and employment, and even then not completely. In the short-run, the volume of output in an economy is determined by the money demand for that output. The GT is the theory of what determines the level of money demand for output.

In Keynes's notation $Y = C + I$. This tautology simply means that National Income (Y?) is equal to the sum of the expenditure on National Output, divided into consumption spending (C) and investment spending (I). Consumption expenditure is governed by Keynes's 'consumption function,' based on the 'psychological law' that consumption (and therefore saving, which is Y-C) is a stable fraction of total income, and that the marginal propensity to consume is less than unity. In a progressive economy the gap beween consumpion and production must be filled by investment if full employment is to be maintained. The amount of investment is determined by the relationship between what the investment is expected to yield over its lifetime ('the marginal efficiency of capital') and the cost of borrowing the money needed to finance it (the rate of interest). Both are subject to uncertain expectations. Consumption is the stable element in demand; investment the volatile element. Thus, what the community wishes to save may at a particular moment exceed what businessmen want to invest. The revolutionary claim of the GT is that when this happens, the divergent plans are brought into equality by a reduction in the level of income (output). Moreover, if the consumption function is known, it is possible to show, by a mathematical formula (Kahn's multiplier), by how much the community's income or alternatively employment must change to reconcile divergent saving and investment plans. The novelty of Keynes's treatment, as compared not just with the older theory, but his own *Treatise on Money*, was to demonstrate that an economy might be in equilibrium, with saving equal to investment, and aggregate demand equal to aggregate supply, at less than full employment.

Critics fastened on the most contentious feature of the new doctrine, the theory of 'under-employment equilibrium.' They denied that this was a genuine equilibrium; it was a disequilibrium state, frozen by the assumption that certain prices were fixed. Flexible interest rates and wage rates would always ensure the desired level of employment, whatever was happening to aggregate demand.

Keynes could have defended his position by saying that his fixed-price assumptions were realistic but he set out to expose the logical flaws underlying the implicit classical price-adjustment story. He denied, first of all, that the interest rate was, or could be, determined in the market for saving and investment, since the quantity of saving depends on the level of current income. This was the basis of his view that it is changes in income that equilibrate saving and investment plans. In place of the classical theory of interest, he offered his liquidity-preference theory of interest, which makes interest the price for giving up cash. The greater the liquidity-preference of wealth-holders, the higher the yield they will demand for switching from cash into bonds or other securities. Insofar as the rate of interest *does* move in response to a fall in investment demand, it will be in the wrong direction, since the 'same circumstances which lead to pessmistic views about future yields (of capital assets) are apt to increase the propensity to hoard' (Keynes 1937, p. 118).

As far as wage flexibility is concerned, Keynes agreed with the orthodox view that the *real* wage is inversely related to the quantity of employment, but denied that real wages are determined in the labor market by money-wage bargains, since an all-round reduction in money-wages might, by reducing the general price level, leave the real wage unaffected. The effect of a 'flexible wages policy' thus depended entirely on its impact on the components of aggregate demand-consumption and investment—which might well be unfavorable.

Keynes conceded that, in time, the decline in employment (or money-wage rates) would, by decreasing the amount of money required to satisfy the business demand, lower the rate of interest required to satisfy the 'propensity to hoard' (Keynes 1937, p. 118); he also thought that, after a slump, a shortage of capital goods would develop which would revive profit expectations'. (Keynes 1936, pp. 317–18). In short, a slump might eventually bring about the price adjustments which according to the classical economists were supposed to prevent the slump from occurring. He doubted, though, whether even longer-run forces would ensure a gravitational pull towards full employment. 'We oscillate,' he wrote, 'round an intermediate position, appreciably below full employment and appreciably above the minimum employment a decline below which would endanger life' (Keynes 1936, p. 254).

Despite nonacceptance by the leading economists of key elements in Keynes's macroeconomic model (Dennis Robertson, in particular, made a damaging attack on his theory of interest), critics were driven to argue for their positions within the analytical framework Keynes had set up. Macroeconomic analysis became part of the economists' tool kit, even if Keynes's own conclusions were rejected.

2.3 Stabilization Technique

Keynes's three main theoretical books were all aimed at providing justifications for conscious attempts to stabilize output and employment at a high level. The quest for output stabilization led directly to the development of National Income and Expenditure

Accounts, to enable governments to estimate the size of the 'output' gap—the gap between what the economy was producing and what it could produce at full employment (or, in later refinement, what it could produce when it was growing to trend)—which needed to be plugged by extra spending. Keynes also insisted that the aggregate demand and supply framework of the GT could be used to work out how much spending needed to be withdrawn from an economy to prevent inflation if aggregate demand exceeded aggregate supply.

In principle, stabilization policy can be either monetary or fiscal, or some combination of both. In practice, Keynes rejected the monetary route to managing demand. He did not deny that investment was interest-elastic. But he thought that interest-rate changes, produced by varying the quantity of bank credit, were too slow-moving in their effects on investment; and he was worried that if interest rates were raised to check a boom it would be difficult to bring them down to check a slump, since the expectation of falling rates would increase the 'propensity to hoard.' So he advocated a policy of permanently cheap money, to be buttressed by capital controls and 'under-funding' of the National Debt. This left fiscal policy as the main instrument for preventing both slumps and inflationary booms. Keynes's fiscal philosophy must be distinguished from Keynesian fiscal policy as understood (and sometimes practiced) after his death. It was based on a sharp distinction between the government's 'ordinary' or current-account budget and what he called the 'capital' budget, made up not just of government spending on capital projects but the economy's total investment. The government's ordinary budget should be balanced at all times, with a surplus earmarked for debt repayment in 'normal' times. It was the task of the capital budget to balance the 'national accounts.' Keynes believed that a great deal of investment was already 'socialized,' that is, subject to public policy even if it was technically private. As he put it in 1943: 'If two-thirds or three quarters of total investment is carried out or can be influenced by public or semi-public bodies, a long-term programme of a stable character should be capable of reducing the range of fluctuation to much narrower limits than formerly…If this is successful it should not be too difficult to offset small fluctuations by expediting or retarding some items in this long-term programme' (Keynes 1943, p. 322).

Finally, Keynes attached only minor importance to exchange-rate adjustments. His Clearing Union Plan (1942) provided for fixed, but adjustable rates, and this was a feature of the Bretton Woods Agreement (1944). Keynes's attitude—like that of most economists of his day—was governed by price-elasticity pessimism. He preferred countries to use, if necessary, capital controls to protect their balance of payments in the framework of a monetary regime of fixed exchange rates, low tariffs, and automatic creditor lending through international institutions.

Keynes was not an ivory-tower theorist. His theorizing was controlled by real world events. His own investment experience enabled him to identify the 'speculative' motive for holding money; his sense of political fragility led him to concentrate on the economics of stabilization; his civil service experience enabled him to turn theories into workable plans. His economic theorizing was less directly, but still importantly, controlled by his philosophical beliefs, particularly by his conception of the good life and of the conditions of just exchange.

3. *The Legacy*

Economics was transformed by its encounter with Keynes. Macroeconomics became for many years a dominant part of the subject; macroeconomic forecasting the main tool of government policy. However, Keynes's doctrines never won universal acceptance, and key aspects of his theoretical and policy legacy have been challenged. The debated issues can be grouped under four heads.

3.1 *Acceptance of Keynes's Theory*

What Keynes bequeathed was not the same as what was accepted. The first theoretical breach came with demonstrations by Pigou (1942) and by Modigliani (1944) that the Keynesian slogan 'quantities adjust, not prices' was true only if money wages were rigid. This became the basis of the 'neoclassical synthesis,' which grafted Keynesian macroeconomics on to classical theory, but left the rigidities unexplained. Milton Friedman's counter-attack was both methodological and theoretical. On the one hand, he argued that it was not acceptable to posit *ad hoc* supply and demand functions. Second, his own application of neoclassical standards of method to Keynes's aggregate equations—seen in his permanent income hypothesis (1957), the stable demand for money function (1956, 1963), and the theory of the 'natural rate' of unemployment (1968)—undermined the case for Keynesian stabilization policy. Economies were more cylically stable than Keynes had supposed; multipliers were small, or nonexistent; government manipulation of aggregate demand had no permanent 'real' effects, but only raised the inflation rate. These 'policy ineffectiveness' propositions were to be hardened still further by the 'rational expectations' school of Robert Lucas and Thomas Sargent. The tendency of Friedman's critique (popularly called 'monetarism') was to reinsert an updated version of the Quantity Theory of Money into the heart of macroeconomics. It revived the pre-Keynesian notion (adumbrated by Keynes himself in the *Tract on*

Monetary Reform) that the most important macroeconomic function of governments was to keep stable the purchasing power of money.

3.2 Acceptance of Keynesian Policy

Contrary to widespread mythology, this was patchy. Both the British government and the US Administration committed themselves to targetting 'high' levels of employment, but it is often asserted that US policy only became Keynesian in the early, and German policy in the late, 1960s, and that both Keynesian episodes were fairly brief. Much of this discussion begs the question of what one means by 'Keynesian' policy. Running a budget deficit is no more a sign of Keynesian virtue than is running a budget surplus in boom conditions anti-Keynesian. An alternative argument is that Keynes's influence was exerted not so much through national policies as through the US willingness to provide the rest of the world with reserves and liquidity. However, this likewise begs the question of how 'Keynesian' this willingness was.

3.3 The Impact of Policy on Events

For a long time, the canonical view was that Keynesian demand-management policies and Keynesian-inspired institutions (the Bretton Woods system) were mainly responsible for the uniquely successful employment and growth performance of most countries in the 1950s and 1960s. Today most of the credit for the 'golden age' is given to opportunities for 'catching up' with American technology, recession-proof military spending by the US, and high levels of 'social' spending.

In the 1950s and 1960s it was common to argue that Keynesian policy helped to save capitalism by removing the scourge of mass unemployment. (In Marxist terms, it legitimized the capitalist order.) By the 1970s it was being argued that it endangered the long-run survival of capitalism by producing rising inflation, an expanding public sector, and increasingly draconian wage and price controls.

3.4 The Current Debate Over Keynesian Economics

According to the monetarist-cum-rational expectations schools, Keynesian economics failed the predictive test: it led to inflation, still worse 'stagflation.' In the 1970s, Keynesian policies were attacked for ignoring the existence of a 'natural' rate of unemployment, and (by the Virginia or Public Choice school) for assuming that politicians wanted to maximize the collective social welfare, rather than their own individual utilities. Taken together, these two attacks offered a forceful argument against the use of discretionary fiscal and monetary policy to balance economies.

The use of models of economies with nominal rigidities is still general, but whether these rigidities are to be taken as given is much more questioned than in 1950s and 1960s. The dominant 'supply side revolution' of the 1980s chiefly was concerned to dissolve rigidities seen as institutional by deregulating labor markets. Against this, the 'new Keynesians' explained how sticky prices are rational because of transactions and information costs, and how shocks to demand can destroy both physical and human capital. These explanations seemed both to strengthen and weaken the case for Keynesian macroeconomic policy. On the one hand, they gave renewed intellectual respectability to stabilization policy. On the other hand, by explicitly introducing inflation into their analyses, they conceded the existence of a rate of unemployment (the so-called NAIRU or Non-Inflation Accelerating Rate of Unemployment) below which unemployment could not be pushed by manipulating demand. Finally, against the mainstream profession's use of Bayesian statistics and decision theory to model agents' behavior, a minority school of 'Post-Keynesians' continues to assert the fundamental nature of Keynes's attack on the rationality axiom.

An interim judgment on the Keynesian Revolution would be that the main body of classical economics was too well-entrenched to be overthrown by the frontal assault he mounted. The notion that uncertainty was at the heart, rather than at the financial margins, of economic processes, proved too subversive of the science economics claims to be, to be acceptable. Most economists would say today of Keynes what Marshall said of Jevons: 'His success was aided even by his faults … he led many to think he was correcting great errors; whereas he was really only adding important explanations.' Whether this will be the final verdict is still questionable.

See also: Economics, History of; Economics: Overview; Elicitation of Probabilities and Probability Distributions; Gender Ideology: Cross-cultural Aspects; Keynesianism and State Interventionism; Macroeconomic Data; Monetary Policy; Policy History: State and Economy; Political Economy, History of; Post-Keynesian Thought; Probability and Chance: Philosophical Aspects; Probability Density Estimation; Probability: Formal; Probability: Interpretations; Unemployment: Structural

Bibliography

Davidson P 1992 Would Keynes be a new Keynesian? *Eastern Economic Journal* **18**(4):

Harrod R F 1951 *The Life of John Maynard Keynes*. Macmillan, London

Hicks J R 1974 *The Crisis in Keynesian Economics*. Blackwell, Oxford, UK
Keynes J M 1919 *The Economic Consequences of the Peace, Vol. 2. Collected Writings of John Maynard Keynes*. Macmillan and Cambridge University Press, London
Keynes J M 1923 *A Tract on Monetary Reform, Vol. 4. Collected Writings of John Maynard Keynes*. Macmillan and Cambridge University Press, London
Keynes J M 1930 *The General Theory and After: Part I Preparation, Vol. 8. Collected Writings of John Maynard Keynes*. Macmillan and Cambridge University Press, London
Keynes J M 1936 *The General Theory of Employment, Interest and Money, Vol. 7. Collected Writings of John Maynard Keynes*. Macmillan and Cambridge University Press, London
Keynes J M 1937 *The General Theory and After: Part II Defence and Development, Vol. 14. Collected Writings of John Maynard Keynes*. Macmillan and Cambridge University Press, London
Keynes J M 1940 *Activities 1939–1945: Internal War Finance, Vol. 22. Collected Writings of John Maynard Keynes*. Macmillan and Cambridge University Press, London
Keynes J M 1943 *Activities 1940–1946 : Shaping the Post-War World, Vol. 27. Collected Writings of John Maynard Keynes*. Macmillan and Cambridge University Press, London
Leijonhufvud A 1969 *Keynes and the Classics: two Lectures on Keynes' Contribution*. IEA, London
Moggridge D E 1992 *Maynard Keynes: An Economist's Biography*. Routledge, London
O'Donnell R M (ed.) 1991 *Keynes as Philosopher–Economist*. Macmillan, London
Patinkin D 1976 *Keynes's Monetary Thought: A Study of Its Development*. Duke University Press,
Schumpeter J A 1952 *Ten Great Economists: from Marx to Keynes*. Allen and Unwin, London
Skidelsky R 1983 *John Maynard Keynes: Hopes Betrayed 1883–1920*. Macmillan, London
Skidelsky R 1992 *John Maynard Keynes: The Economist as Saviour 1920–1937*. Macmillan, London
Skidelsky R 1995 Keynes and the quantity theory of money. In: (ed.). *The Quantity Theory of Money from Locke to Keynes*. Edward Elgar, Cheltenham, UK
Skidelsky R 2000 *John Maynard Keynes: Fighting for Britain 1937–1946*. Macmillan, London
Walters A 1987 *Friedman, Milton. The New Palgrave Dictionary of Economics*. Macmillan, London

R. Skidelsky

Keynesianism and State Interventionism

Keynesianism, whatever its variants, assigns to the State the responsibility to regularly intervene in the operation of the economy so as to maintain its performance at a high level of employment and growth (Hall 1989). Keynesianism believes in the effectiveness of this intervention. From Keynes' analysis of the preference for liquidity and the economic crisis of the 1930s it has retained a definite skepticism concerning the effectiveness of monetary policy. Investment and consumption are held to be little affected by interest rates; it therefore suffices to have a monetary policy that maintains interest rates at a sufficiently low level. On the other hand, Keynesianism stresses countercyclical management of overall demand by means of fiscal and budget policies: financing expenditures by public borrowing in case of recession; seeking a budget surplus to hold down inflation in the rising phase of the cycle. Fiscal constraint can be relaxed to finance welfare programs. This is a decisive change with respect to classical economics that deemed a balanced budget to be the highest priority. The Keynesian stimulus works via a sort of self-fulfilling prophecy concerning investment and demand. Prewar policy held to the contrary that the modest means available to the government was a guarantee for democracy, allowing it to stand up to populist excesses and to unlimited demands from special-interest groups.

Keynes would have been hostile to Keynesianism. In all likelihood he would have espoused, although in different terms, the criticism expressed by Friedman (1968) and Lucas (1976) towards the Keynesians' conception of State intervention in the economy. But this criticism, albeit radical (and essentially well taken), is aimed less at the principle than at the modalities of an effective State intervention. Keynesianism, for all its '*naïveté*,' has managed to make this intervention accepted. This is the paradox that ensures a long life for intellectual and political debate about Keynes' writings. This contribution is divided into three sections: a history of the reception encountered by Keynesian ideas: criticism of these ideas, and the relevance of Keynes today.

1. *The Reception of Keynesian Ideas*

Keynesianism was not born with Keynes. Hicks (1937) notably formalized the IS-LM curves that jointly determine the global income and the rate of interest, via the relations between investment, savings, demand and supply of money. Samuelson (1948) provided the analytical frameworks of macroeconomic policy. Put briefly, the synthesis thus established, neoclassical in spirit, resituates Keynes' work in the conventional framework of market equilibrium. It eliminates the role of agents' expectations, which was decisive for Keynes (as well as eliminating the role of the 'state of confidence'), and replaces it with a mechanistic vision of the economy that is reduced to a few global parameters. Phillips (1958) then added the curve linking wage increases, rising prices and the unemployment rate. The development of national accounting and statistical apparatus provided the basis for econometric modeling. Little by little economists, following in this respect the theory of economic policy elaborated by Tinbergen (and drawing politicians and political discourse in their wake), came to espouse the use of macroeconomic models for the quantitative evaluation of State policy. Keynesian policy is funda-

mentally a policy designed to fine-tune short-term fluctuations in the economy. If short-term forecasts given by a model are accurate, then the long-term evaluation of the effects of a policy is presumed to be reliable. The long-term outcome is engendered by a carefully constructed chain of short-term periods. Keynes thought, on the contrary, that the agents' view of the long term had a direct effect on their current behavior.

From the 1930s to the 1950s the reception of Keynesian ideas spread through three channels: the influence of professional economists' expertise on political decision making (seeking new explanations for the economic crisis); new demands formulated by political parties (themselves conditioned by previous experience); and the capacity of Keynesian ideas to spawn broad political coalitions. The configuration of history fostered the convergence of these three channels: the heritage of the war economy (when the 'young' Keynesians had demonstrated the effectiveness of State organization and intervention); the needs of reconstruction (which called for a high level of public investment), and the expansion of social policies aimed at achieving full employment. In the light of the prewar risk of confrontation between the labor agenda (leading to nationalization of the means of production) and that of capitalism (*laissez-faire* liberalism), Keynesianism offered a compromise. Full employment and sustained growth could be achieved in a capitalist economy through the judicious use of macroeconomic management (without interfering with the management prerogatives of private capital). Labor could see full employment guaranteed, and capital could continue to control investment.

But in order for Keynesian policy to be effective two structural conditions were necessary: a system of appropriate institutions; and a common language and beliefs shared among the actors in the economy. These conditions were fulfilled to various degrees depending on the country—more highly developed in Great Britain and France though differently, than in the United States or Germany. The State must be linked to the actors by a set of institutions that structure the flow of information and resources. These institutions ensure a common assessment of the economic situation and provide security for anticipations (notably via property rights and contract law). A relationship of confidence between the central bank, the State, and financial markets is crucial for the financing of public debt and establishing regulation of the banking system. At the same time, shared views of the expected and legitimate role of the State must spread throughout the economy and society. This 'conventional' foundation creates the state of confidence enabling each actor to constitute favorable anticipations (of demand and investment). The combination of these anticipations is self-fulfilling of the forecast projected by the macroeconomic model; it allows State intervention to attain its goals. This double system of institutions and shared conventions paradoxically ensures the effectiveness of Keynesian policy. Without this system, Keynesian policy is nothing. This is what the Keynesians forgot. They consequently come to believe in the State as a *deus ex machina*. In neglecting the question of expectations they left the way open for the school of rational anticipations. The return in force of market economics and neo-classical rationality was the outcome.

2. *Criticism of Keynesianism*

A series of attacks, some old (Hayek, Buchanan) and others recent or renewed (Friedman, Lucas, Sargent), converged in the late 1960s and early 1970s, leading to a collapse of Keynesianism.

Hayek and Buchanan had paved the way, one by rejecting any sort of validity for macroeconomics in itself, the other by demolishing the myth of the State as an enlightened despot. Keynesianism opens the floodgates of inflation by facilitating uncontrolled creation of credit. Then come wage and price control policies that eliminate inflation at the cost of much higher unemployment than that which would have resulted under a stable currency regime (Hayek). Economics can be seen as a science that gives people greater control over their environment, but it has no command over politics. Governments are more interested in collecting the maximum number of votes than in pursuing the public good (Buchanan).

The Phillips relationship had given Keynesian policy its missing link: that of setting wages. Fine tuning efforts could focus on the inflation/unemployment dilemma: how much unemployment is needed to ensure price stability? The stagflation of the 1970s (high unemployment and high inflation occurring simultaneously) was its downfall. Friedman reintroduced anticipations. He showed that the trade-off between inflation and unemployment is only temporary: it stems from inflation that is not anticipated by agents. Eventually unemployment returns to its natural rate, which is determined by structural socio-economic characteristics, but at the price of higher inflation. Discretionary monetary policy cannot durably stabilize either interest rates or unemployment at low levels. Keynesians situate themselves in a world where everyone anticipates that prices will remain stable, and where this anticipation persists no matter what happens to prices and wages. This world has vanished. Friedman updated his 1948 program: replacing a discretionary monetary policy with public adoption by the monetary authorities of a set rule. In the present instance this means: setting a growth target for the money supply (using a specified monetary aggregate); avoiding sudden changes, not overreacting to events. In 1948 this 'monetarist' framework had received no attention. Economists had adopted the highly flattering position of 'advisor' to the managers

of the economy (and thereby justifying the development of their profession). The contrast between the massive unemployment of the Great Depression and the success of the post-war economy had persuaded the economy's managers of the efficacy of systematic State intervention. 'Friedmanian' rules are now widely followed by monetary authorities (like the Federal Reserve Bank or the European Central Bank).

Lucas's criticism was aimed at the use of macroeconometric models. The theory of economic policy (Tinbergen 1952) underlying this use is invalidated by the fact that actors adapt their rules of decision to modifications in economic policy. The variables controlling economic policy are thus endogenous to the system; the structure of the models varies with economic policy. No reliable projection of future behavior and trends can be obtained by extrapolation of series and correlation with the past. In order for a change in the rules to have a chance of being effective (i.e., leading to foreseeable trends) it must be publicly debated and understood by the actors in the economy. Once again, despite a few pockets of resistance, this criticism ultimately dealt a fatal blow to substantive macroeconomics 'defined by itself to be the body of expertise … devoted to the development and refinement of forecasting and policy evaluation methods' (Lucas 1980, p. 250). This discipline failed to foresee the return of economic growth at the dawn of the twenty-first century. But this has not prevented the teaching of macroeconomics to continue in academia.

Now that the din of the dogmatic battles has ceased, a more balanced assessment is possible. The structuring role of the anticipations of agents is now recognized. They introduce an irreducible element of unpredictability. What counts in the dynamics of the economy is action and coordination. Economic policy must abandon the aim of influencing the course of events by substantial and discretionary means. Its field of action relates to the rules that serve as the framework of reference for private forecasts and actions. At bottom, for these rules to be effective they must be debated and made public. They leave room for interpretation in action and for innovation. They provide a framework that can be publicly revised and adjusted in keeping with the results obtained. This does not imply, however, an absence of concern for joblessness and full employment, but rather a clearer consciousness of the properties of various policy instruments and how they complement each other. The unemployment rate can no longer be a direct target of economic policy. Measuring an unemployment rate under full employment is meaningless. Firms and employers have many options other than unemployment at their disposal for adjusting to fluctuation in the demand for products and to the outlook for innovation. Satisfactory allocation of social risks is the task of welfare economics. Providing a framework of stability and security that safeguards the anticipations and terms of agreements concluded–this should be the essence of economic policy (Lucas 1976). But can this modest objective satisfy the body of professional economists, who are much more numerous than in the past and permanently accustomed to the comforts and remuneration of policy expertise?

3. *The Relevance of Keynes Today*

Keynesianism is dead, but what of the writings of Keynes himself? In his ferocious criticism of Tinbergen in 1939 Keynes already expressed the arguments that were later expanded by Lucas (who did so without making any reference to the writings of Keynes). This criticism originates in Keynes's *Treatise on Probability* (1921). As early as this work he understood that the problem is a double one, that of measurement and that of 'passing from statistical description to inductive generalisation' (Keynes 1939, p. 566). If all causal factors must be measurable for economic policy to be subjected to statistical tests and forecasting, 'it withdraws from the operation of the method (of Tinbergen) all those economic problems where political, social and psychological factors, including such things as government policy, the progress of invention and the state of expectation, may be significant' (1939, p. 561). Keynes was likewise prudent concerning the arithmetic of the multiplier. Lastly, for Keynes the prime target of monetary policy was the value that actors accord to the conventions that they follow on financial markets. It is the state of confidence that is important, and not directly interest rates and unemployment.

'Neo-Keynesians' have counter-attacked. They integrated into the Keynesian scheme rational anticipations and the possibility of exogenous shocks and of constraints on individual supplies and demands. But in so doing they remain open to these criticisms of Keynes. And by insisting on the rigidity of costs, their recommendation of wage flexibility paradoxically converges with that of liberals attached to the idea of reducing the natural unemployment rate. The 'post-Keynesians' (Davidson 1991) insist on the radical and non-probabilistic uncertainty that surrounds the decision-making of economic agents. They thus move closer to the institutional (and not solely political) content of Keynes' message. They attempt to express his relevance to the current context. This uncertainty stems from an instability in the overall trajectory of the economy. Institutions must be created to construct a 'corridor' that restricts the range of directions that actions can take. Following Keynes' intuition on inference and probability, it should probably be added that institutions allow people to make judgement and inference on situations of action which are well-founded and secure.

It is time to leave behind the paralysis of opposing schools and dogmatic assumptions (in particular regarding States vs. the market, or the nature of the

market, or rationality). Too often a worthwhile idea is rejected, on the pretext that it comes from the opposing school. Little by little a macroeconomics of rules is expanding in which government of the economy by the State is permissible (a posthumous victory for Keynesianism). But, as thought Keynes, this government must be limited in scope and in ambition (Skidelsky 1999). The dynamics of the economy are above all the result of the actions, anticipation and coordination of the actors. These actors are diverse in terms of their identities, their means and ends (and in the relationship they establish between means and ends). The knowledge they have acquired of their terrain renders their action irreplaceable in a context of uncertainty. A framework of democratically debated, stable and foreseeable rules seems necessary for the implementation of collectively effective actions. But the debate is shifting to the status of these rules (are they prescriptive rules, or guidelines open to interpretation in action?) and to the status of macroeconomic variables (objectives, constraints, or simple by-products of market activity?).

Bibliography

Davidson P 1991 Is probability theory relevant for uncertainty? A post-Keynesian perspective. *Journal of Economic Perspectives* **5**: 129–43

Friedman M 1968 The role of monetary policy. *American Economic Review* **58**: 1–17

Hall P A (ed.) 1989 *The Political Power of Economic Ideas. Keynesianism Across Nations*. Princeton University Press, Princeton, NJ

Hicks J R 1937 Mr Keynes and the 'Classics': A suggested interpretation. *Econometrica* **April**: 147–68

Keynes J M 1921 *A Treatise on Probability*. Macmillan, London (1973 reprinted in *The Collected Writings of J. M. Keynes, Vol. VIII*. Macmillan, London)

Keynes J M 1936 *The General Theory of Employment, Interest and Money*. Macmillan, London (1973 reprinted in *The Collected Writings of J. M. Keynes*, Vol. VII. Macmillan, London)

Keynes J M 1939 A review of *A Method and its Application to Investment Activity* by J. Tinbergen (1939 Statistical Testing of Business-cycle Theories, tome I, League of Nations, Geneva). *The Economic Journal* **49**: 558–68 (1973 reprinted in *Collected Writings*. Macmillan, London, Vol. XIV)

Leijonhufvud A 1968 *On Keynesian Economics and the Economics of Keynes—A Study on Monetary Theory*. Oxford University Press, Oxford, UK

Lucas R E 1976 Econometric policy evaluation: A critique. In: Lucas R E (ed.) 1981 *Studies in Business-cycle Theory*. MIT, Cambridge, MA

Lucas R E 1978 Unemployment policy. In: Lucas R E (ed.) 1981 *Studies in Business-cycle Theory*. MIT, Cambridge, MA

Lucas R E 1980 Rules, discretion and the role of the economic advisor. In: Lucas R E (ed.) 1981 *Studies in Business-cycle Theory*. MIT, Cambridge, MA

Phillips A W 1958 The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* **25**: 283–99

Samuelson P A 1948 *Economics*. MIT, Cambridge, MA

Skidelsky R 1999 The conditions for the reinstatement of Keynesian policy. In: Pasinetti L L, Schefold B (eds.) *The Impact of Keynes on Economics in the 20th Century*. Edward Elgar, Cheltenham, UK

Tinbergen J 1952 *On the Theory of Economic Policy*. North Holland, Amsterdam

R. Salais

Kimura, Motoo (1924–94)

Motoo Kimura was Japan's foremost population and evolutionary geneticist. He is best known for his neutral theory of molecular evolution, which asserts that the great majority of evolutionary changes at the DNA level are driven by mutation and random survival of mutant genes, rather than by natural selection. He was also a major contributor to the mathematical theory of evolution and population genetics.

1. Life

Motoo Kimura was born on November 13, 1924, in Okazaki, Japan. Early in his life he was fascinated by plants and had a deep interest in their structure. He also displayed an early talent for mathematics. He was admitted to the National High School in Nagoya in 1942. Here, thanks to a knowledgeable and sympathetic teacher, he studied the chromosomes of lilies. This institution also provided his first contact with biometry, and he realized for the first time that he could use his mathematical skills to study biological problems. In 1944 he was admitted to the Kyoto Imperial University.

Normally Kimura would have enrolled in the Faculty of Agriculture to study with the internationally respected geneticist, Hitoshi Kihara. But agriculture students were being called into military service, while botany students were not. So, on Kihara's advice, Kimura enrolled in the Faculty of Science. By the time of his graduation the war was over, and he joined Kihara who encouraged his study of mathematical problems. During this time, Kimura had no specific duties and took the opportunity on his own to read systematically in mathematical genetics. His mathematics, formidable as it sometimes seems, was mainly self-taught. Life in wartime Japan was miserable, and remained so for some time afterward. Kimura was able to supplement a sparse physical and intellectual diet by visiting his cousin, a quantum physicist who lived in the suburbs and could supply both food and scientific talk. Kimura hoped for a theory of population genetics comparable to that of physics.

In 1949, he joined the staff of the National Institute of Genetics in Mishima, and retained this position for the rest of his life. Thanks to help from some American geneticists studying the effects of the atomic bomb in Hiroshima, he was able to obtain support for graduate study in the USA. After a year at Iowa State College Kimura transferred to the University of Wisconsin in 1954 as the author's student. This also provided an opportunity for daily association with America's most eminent population geneticist, Sewall Wright, who had recently moved to the University of Wisconsin.

Except for his graduate years and several short periods abroad, Kimura spent the remainder of his life in the National Institute of Genetics in Mishima, Japan. Late in his life, he developed amyotrophic lateral sclerosis and died on his 70th birthday, November 13, 1994.

2. *Major Contributions*

Kimura's scientific life can be divided into two periods. The first began while he was still a student and continued until 1968. During this time he made major contributions to the mathematical theory of population genetics. In 1968, he introduced his neutral theory of molecular evolution. Although he never stopped making theoretical contributions, his major efforts from this time on were devoted to further developments and defense of the neutral theory.

The mathematical theory of evolution had its heyday in the period roughly from 1920 to 1950. It was almost completely dominated by three men, R. A. Fisher and J. B. S. Haldane in England, UK and Sewall Wright in the US. Kimura became the logical successor. Much of Kimura's work utilized the stochastic equations of the Russian mathematician, A. Kolmogrov. Kimura showed a remarkable inventiveness in solving these difficult equations and applying them creatively to significant evolutionary problems.

Among the problems that Kimura solved are: the probability that a new mutant gene will ultimately spread through the population; the number of generations required for this process; if the mutation is lost, the number of generations it persists before loss; the number of individuals that carry a mutation during the time until it is fixed or lost; and the average age of a mutant gene segregating in a population.

At the same time he contributed to many other areas of population genetics theory. These included: a comparison of inbreeding systems, showing that systems minimizing the immediate increase in homozygosity are not the best in the long run; a further development of Fisher's 'fundamental theorem of natural selection' with more explicit treatment of gene interactions; a new and influential model of population structure, the 'stepping stone' model, in which migrants are restricted to neighboring colonies; the efficiency of rank-order selection in changing gene frequencies and removing deleterious mutations from the population.

By 1968 the subject of molecular evolution was in a period of rapid development, which still continues. This was the year of Kimura's bombshell. He argued that the great majority of amino acid and nucleic acid changes are selectively neutral. Instead of natural selection as the main directive force, these changes occur by mutation and whether they persist or are lost is a matter of chance. Remarkably, Kimura's early work on stochastic processes turned out to be preadapted for the study of molecular evolution. He had a set of ready-made tools available, and he proceeded to exploit them with great vigor.

The neutral theory made a number of qualitative and quantitative predictions. Among these was that molecular changes occur at a rate independent of structural and physiological changes; Kimura noted that sharks, whose external form has hardly changed in an enormous time period, have as much molecular variation as rapidly evolving animal species, such as mammals. There were a number of similar predictions. One of the simplest and most remarkable is that the rate of neutral gene substitution in the population is equal to the rate of mutation of this gene per individual. This has permitted predictions of evolutionary rates; departures from these predictions provide evidence for natural selection. Rates that are slower than the neutral expectation can be attributed to 'selective constraint;' that is, the amino acids in this region are so specifically fitted to their function that any replacement doesn't work as well. Thus, essentially all mutations are harmful and are removed by natural selection (sometimes called purifying selection). Several proteins, for example some histones, are in this category. Any change in the molecular structure is harmful. In contrast, regions that evolve faster than the neutral rate are attributed to positive selection. An appreciable number of the mutations are beneficial. One of the best understood examples is the DNA region governing tissue transplants.

The neutral theory has provided a rational basis for the idea of a 'molecular clock' guiding the rate of evolution. The rough constancy of the mutation rates for comparable genes in different organisms—in particular that they are much more stable than the forces of selection—has led to the clock assumption becoming the basis for construction of phylogenetic trees, a field of great activity as molecular data are increasingly used.

At first, Kimura's theory was rejected out of hand by most evolutionists. But gradually, over the years, it has come to be widely accepted. It is now realized that a great deal of the DNA of higher organisms has no known function and that the actual genes constitute a small part of the DNA. At present it is accepted that most nucleic acid substitutions follow the neutral paradigm. As for amino acid changes, the jury is still

out. Clearly some changes follow neutral kinetics, others clearly are selected, and the proportions are yet to be sorted out.

3. An Assessment

Let it be said at the outset that Kimura's work has had its greatest impact on evolution theory rather than on the social and behavioral sciences. His work is very general and is applicable to the whole living world, of which that of greatest interest to social scientists, human evolution, and human behavioural evolution in particular, are only a part. Molecular evolution has yet to make the impact on human behavioral evolution that it is certain to make one day.

Kimura's greatest contribution has been to impose a stochastic viewpoint on the study of evolution. Beneath surface regularities (and irregularities) of form and function are a multitude of molecular changes, governed by the laws of chance. At present, research in behavioral genetics and evolution is based mainly on deterministic theory. Problems such as separating genetic from environmental effects are usually studied by such means as observing adopted children or identical twins reared apart, and the techniques are usually those derived from analysis of variance. But as evolutionary and population genetics theory becomes further developed and richer, and as the molecular basis of behavioral traits (including those in humans) are better understood, stochastic aspects will play an increasingly important role.

The study of natural selection and adaptation has long had difficulty in formulating clearly testable and falsifiable hypotheses. This has been particularly evident in the abundance of 'just so' stories; because we can think of a plausible reason for a structure or function doesn't necessarily mean that it arose for this reason (although it often has). One attribute of the neutral theory is that it can provide a null hypothesis for testing causes of evolutionary changes. For example, as mentioned above, the neutral theory makes a simple evolutionary rate prediction, against which departures can be measured. Thus far, this kind of reasoning has been used mainly for genes with physiological functions, but as the molecular basis of behavioral changes becomes better understood this kind of analysis will be increasingly applicable.

During his lifetime, Motoo Kimura received virtually every honor for which an evolutionary geneticist is eligible. In addition to numerous honors both in Japan and elsewhere, he was a foreign member of the United States National Academy of Sciences, the Royal Society of London, and the Chevalier de L'Ordre National du Mérite in France. He received honorary doctorates from the Universities of Chicago and Wisconsin.

Kimura's lasting impact is twofold. The first, best appreciated by population geneticists, is his enrichment of population genetic and evolutionary theory, especially as these are influenced by stochastic processes. The second, appreciated by a much larger audience, is his discovery of the enormous role that random processes play in molecular evolution. Kimura never denied the importance of Darwinian natural selection as the main evolutionary force shaping form and function, but beneath this, he emphasized, is an enormous play of random changes at the molecular level. In addition to Darwin's 'survival of the fittest,' he added the slogan, 'survival of the luckiest.'

See also: Adaptation, Fitness, and Evolution; Darwin, Charles Robert (1809–82); Evolution, History of; Evolutionary Selection, Levels of: Group versus Individual; Fisher, Ronald A (1890–1962); Galton, Sir Francis (1822–1911); Genetics and Development; Human Evolutionary Genetics; Lorenz, Konrad (1903–89); Natural Selection

Bibliography

Kimura M 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK

Kimura M 1994 *Population Genetics, Molecular Evolution, and the Neutral Theory*. University of Chicago Press, Chicago

J. F. Crow

Kindling, Neural Basis of

The epilepsies are a heterogeneous group of chronic brain disorders characterized by an enhanced susceptibility to spontaneous, recurrent seizures due to the excessive activation of neuronal networks. In recent years a number of widely diverse causes of epilepsy have been identified, including specific gene mutations, developmental defects, and acquired metabolic and structural abnormalities. For example, a mutation in a subunit of the nicotinic acetylcholine receptor is the basis of autosomal dominant nocturnal epilepsy, while other types of epilepsy have been linked to calcium channel defects (Clark and Wilson 1999). Abnormal brain development resulting in dysplasias and other cortical malformations may also lead to epilepsy (Guerrini et al. 1999). At almost any age, injury to the brain due to stroke, trauma, infection, tumors, as well as the adverse effects of alcohol and other neurotoxic substances may be epileptogenic. Despite this diversity, the multiple causes of epilepsy share at least one key feature i.e., they all produce long-lasting changes in neuronal physiology that result in recurrent spontaneous electrical discharges. Identifying the mechanism(s) of epileptogenesis will there-

fore lead to a more thorough understanding of brain function and contribute to the development of more effective pharmacologic treatments for epilepsy. While the underlying cellular processes responsible for the epileptic state remain unknown, a great deal of progress has been made in delineating the critical events associated with seizure induction and propagation, as well as epileptogenesis, primarily through the use of animal models. In this regard, the kindling model of epilepsy has played a dominant role.

1. *Definition of Kindling*

In 1967 Goddard demonstrated that daily administration of a subconvulsant electrical stimulus in rats caused a long-lasting change in brain function characterized by the subsequent development of generalized motor seizures. Since this initial report, electrically induced kindling in the mammalian brain has been one of the most widely employed methods to produce recurrent seizures both *in vivo* and *in vitro*. During the initial stages of kindling the inciting electrical stimulus usually produces no overt changes in behavior or electrical activity as measured by an electroencephalogram (EEG) (McNamara et al. 1992). However, with successive stimuli there is associated EEG evidence of hyperexcitability in the form of an after-discharge. Initially, the electrical seizure activity is unaccompanied by behavioral seizures. However, with continued periodic stimulation the duration of the afterdischarge increases along with the development of behavioral seizures. The behavioral stages of kindled seizures in rats can be classified according to the following classes: class 1, facial clonus; class 2, head nodding; class 3, contralateral forelimb clonus; class 4, rearing; class 5, rearing and falling. Once a class 5 seizure is reached, the subject is considered to be fully kindled. In this state, seizures can be elicited by low-level electrical stimulation even after a prolonged stimulus-free interval, indicating that kindling induces permanent alterations in neuronal physiology. Importantly, elicitation of an afterdischarge is required for kindling to occur. In addition to electrical stimulation, after-discharges can be produced by carbachol or other convulsant agents, as well as alcohol, which are used in models of chemical kindling.

2. *Kindling Properties*

Experimental evidence has demonstrated that different brain regions have varying capacities to develop kindled seizures (Sato et al. 1990). The regions most susceptible to kindling include limbic structures such as the hippocampus, amygdala, and piriform cortex, as well as certain areas of neocortex (Sato et al. 1990, McNamara et al. 1992). Interestingly, there is substantial overlap between epilepsy-prone regions in the rodent brain and those in humans, lending further support to the validity of using kindling models to study epileptogenesis in humans. However, it is important to point out that there are variations between kindling models. In this regard, there are clear differences between the electrophysiological and behavioral features following neocortical stimulation and those following stimulation of limbic structures, such as the hippocampus or amygdala (McNamara et al. 1992). For example, in contrast to amygdala kindling, the afterdischarge produced by stimulation of anterior frontal neocortex may be accompanied by behavioral changes at the outset. The behavioral features, however, are less pronounced than with limbic stimulation. Additionally, seizures evoked by stimulating frontal neocortex are relatively brief when compared to amygdala- or hippocampal-evoked seizures. Although not well understood, these differences may eventually be found to have clinical relevance in terms of modulating the responses to pharmacologic treatment.

The epileptogenic potential of a specific brain region appears to be closely correlated with the rate of kindling i.e., the amount of stimulation required to elicit a class 5 seizure (Sato et al. 1990). Thus, both the amygdala and piriform cortex require the least number of stimulations to reach the kindled state, while hippocampal kindling occurs relatively slowly. Evidence has suggested that increased kindling efficacy, particularly in the amygdala/piriform cortex region, may be related to strong connections with motor and other forebrain systems, as well as the presence of highly reactive neurons with pacemaker-like properties (Sato et al. 1990).

3. *Kindling Induces Changes in Gene Expression and Cellular Morphology*

Investigations employing the compound, 2-deoxyglucose, which labels metabolically active brain regions, have been performed in kindled animals. The results of these studies indicate that the regional pattern of activation following electrical stimulation appears to be stage-dependent i.e., during stages characterized by partial seizures the amygdala and piriform cortex are metabolically labeled, while other areas are subsequently recruited following further stimulation (Ackermann et al. 1986). Activation of brainstem structures such as the substantia nigra has also been reported in kindled animals. Regional increases in metabolic activity in the kindled brain can be correlated with the induction of molecular markers such as the immediate early gene, c-fos (Dragunow and Robertson 1987).

Immediate early genes encode transcription factors that are thought to play critical roles in regulating long-term changes in neuronal function and structure (Sheng and Greenberg 1990). These alterations are likely important in normal brain functions such as

learning and memory, as well as pathological conditions such as epilepsy (Sheng and Greenberg 1990). Transient activation of the c-fos gene has been demonstrated in both the hippocampus and piriform cortex of kindled animals (Dragunow and Robertson 1987, Dragunow et al. 1988, Simonato et al. 1991). Moreover, the spatial and temporal patterns of expression of c-fos and other immediate early genes, such as c-jun and NGFI-A, may undergo complex changes following a single afterdischarge (Simonato et al. 1991). These findings suggest that the transcriptional targets of Fos protein and other immediate early gene products may be critical effectors of the long-term modifications in neuronal function that are associated with kindling.

In addition to immediate early genes, kindling-induced alterations in the expression of growth factors, neuropeptides, and other proteins have been documented (Perlin et al. 1993, Kopp et al. 1999). Increased activity of calcium binding proteins has also been demonstrated in the dentate gyrus of kindled animals, suggesting that intracellular calcium may be an important modulator of epileptogenesis. Increased intracellular calcium associated with oxidative stress may lead to free radical formation and neuronal injury in some kindling paradigms (Rauca et al. 1999).

Kindling-induced effects have also been connected with morphological changes in neurons and astrocytes. Although the contribution of astrocytes and other glial elements needs to be clarified, both astrocyte proliferation and activation have been observed in the hippocampus and amygdala/piriform cortex of kindled rats (Sato et al. 1990, Adams et al. 1998). Studies in knockout mice suggest that the astroglial glutamate transporter may play a role in defining the duration of kindled seizures (Watanabe et al. 1999). Timm staining, a histochemical technique that is widely employed to evaluate synaptic reorganization following neuronal injury in the central nervous system (CNS), has been used in kindling studies (Sutula et al. 1988). An increase in the density of Timm-positive granules in the dentate gyrus of kindled rats supports the notion that mossy fiber collaterals terminate back on their cells of origin and thereby enhance the reactivity of the hippocampus (Sutula et al. 1988). In addition, the compelling results of ultrastructural studies have revealed kindling-induced changes in both the number and type of synaptic terminals, providing a possible anatomical basis for the long-lasting alterations in synaptic function (Geinisman et al. 1992).

4. *Excitatory Neurotransmission and Kindling*

A discussion of the neural basis of kindling would not be complete without addressing the role of exctitatory and inhibitory synaptic transmission. It is widely held that epileptogenic mechanisms are associated with alterations in the balance between excitatory and inhibitory synaptic functions in the CNS. Excitatory neurotransmission in the mammalian CNS is primarily due to the activation of glutamate receptors (Michaelis 1998). Although the molecular biology of glutamate receptors is beyond the scope of this entry, it is important to recognize that glutamate receptors are commonly classified on the basis of pharmacological profiles into ionotropic i.e., *N*-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-4-isoazole propionate (AMPA), and kainic acid (KA) subtypes, as well as ligand-gated metabotropic receptors (Michaelis 1998). Each type of glutamate receptor is formed by an assembly of several subunits (i.e., NMDA receptors have five subunits), all of which have been cloned, sequenced, and functionally characterized. As expected, specific combinations of subunits can assemble into functional glutamate receptors with distinctive physiological and pharmacological properties (Michaelis 1998). Several features distinguish NMDA receptors from the other subtypes of glutamate receptors. These include a voltage-dependent Mg^{2+} block of the ion channel that is relieved by depolarization, sensitivity to glycine, and, most importantly, the ability to allow entry of Ca^{2+}, which is itself a key modulator of neuronal excitability (McBain and Mayer 1994).

Among the classes of glutamate receptors, the activation of NMDA receptors appears to play the most prominent role in modulating long-term alterations in synaptic function in the nervous system, a process that is referred to as synaptic plasticity (Mody 1999). NMDA receptor activation has been documented in the kindled state, and changes in specific channel properties have been proposed to form the underlying basis of enhanced NMDA function after kindling (Mody 1999). Among the NMDA receptor-related alterations that have been reported with kindling are prolonged channel opening, increased channel phosphorylation, enhanced agonist potency, and alterations in the Mg^{2+} block (Mody 1999). Evidence has also indicated that such alterations may last for up to one month or longer. In contrast, other studies have demonstrated either no changes in the expression of NMDA subunits, or minimal changes in the binding of L-[^{3}H]glutamate to NMDA receptors after electrical or chemical kindling (Ekonomou and Angelatou 1999, Mody 1999). Additional evidence supporting a prominent role for NMDA receptor activation in the development of kindling is provided by studies employing NMDA antagonists. NMDA receptor antagonists such as aminophosphonovaleric acid (AP5), and channel blockers such as MK-801, are highly effective at inhibiting kindling development (McNamara et al. 1988, McNamara et al. 1992). For reasons that are at present poorly understood, these same antagonists are relatively weak suppressors of seizures in previously kindled animals.

5. *Inhibitory Neurotransmission and Kindling*

As ligand-gated $GABA_A$ receptors are responsible for most of the inhibitory neurotransmission in the mammalian brain, it is reasonable to surmise that repression of GABAergic synaptic inhibition may promote seizure development and kindling. Analogous to glutamate receptors, $GABA_A$ receptors are composed of different subunits that assemble into functional channels that regulate Cl^- ion conductance, ligand affinity, and sensitivity to modulators (Luddens et al. 1995). Although a decrease in synaptic inhibition may lead to increased neuronal excitation, the role of altered GABAergic synaptic transmission in the development of epilepsy remains controversial. Changes in the expression of $GABA_A$ receptor subunit proteins, as well as increased sensitivity of channels to modulation by Zn^{2+} have been demonstrated (Mody 1999). However, both enhanced GABA binding as well as an increase in the number of $GABA_A$ receptors have been reported in the kindled hippocampus (Mody 1999), consistent with increased synaptic inhibition.

Taken together, the results of these and other studies suggest that kindling induces alterations in glutamatergic or GABAergic synaptic functions characterized by effects on either intrinsic channel properties or synapse topology, respectively. Thus, altered synaptic excitation is primarily due to changes in the physiological properties of NMDA receptor channels, while changes in synaptic inhibition are due to alterations in $GABA_A$ receptor numbers, subunit composition, and ligand affinity. Whether epileptogenesis is associated with an imbalance between glutamatergic and GABAergic functions, however, remains to be determined.

6. *Relevance of Kindling to Human Epilepsy*

Because of the strong correlation with long-term alterations in brain function the kindling model is highly relevant to the study of chronic epilepsy in humans. In this regard, it is important to recognize that the permanent changes produced by periodic subconvulsant stimuli distinguish kindling from other seizure models i.e., electroshock or the administration of compounds such as pentylenetetrazol, in which single seizures are produced in an otherwise normal brain. This has important ramifications for pharmacotherapy, as agents that may be effective anticonvulsants in single seizure models may not prevent kindling-induced seizures. One example is the NMDA receptor channel blocker, MK-801, which inhibits electroshock-induced but not kindled seizures (McNamara et al. 1988). In addition, a distinction must also be made between drugs that are anticonvulsant and those that are antiepileptogenic. Accordingly, compounds such as MK-801 may be more effective at inhibiting the development of kindling (i.e., antiepileptogenic) rather than kindled seizures (anticonvulsant), while other drugs such as carbamazepine may block kindled seizures but not kindling development (Sato et al. 1990). Further, drugs that have both antiepileptogenic and anticonvulsant properties (i.e., valproic acid) may provide greater therapeutic efficacy against epilepsy. It is therefore important to include the kindling model in drug-screening programs aimed at identifying antiepileptogenic compounds for use in humans.

Additional similarities to human epilepsies provide further rationale supporting the use of the kindling model to study human epileptogenesis. Many of the behavioral and electroencephalographic features of kindled seizures in rodents bear a striking resemblance to those of temporal lobe epilepsy in humans. Surprisingly, attempts to produce complete kindling (i.e., class 5 seizures) in nonhuman primates have met with limited success. Although class 4 and class 5 seizures can be elicited in baboons, experimental evidence has suggested that there may be strain-dependent differences in the susceptibility of nonhuman primates to kindling (Sato et al. 1990). Nonetheless, the ability to induce kindling in nonhuman primates supports the idea that a kindling-like mechanism plays an important role in the development of epilepsy in humans. In addition, several groups have reported the development of recurrent seizures in patients who have received repeated periodic electrical stimulation to the brain for the treatment of either schizophrenia or chronic pain (Sato et al. 1990). These observations lend strong support to the idea that kindling does occur in humans.

If the kindling pathway is a critical component in human epileptogenesis, then the identification of pathophysiological conditions that may promote kindling becomes increasingly important. There is accumulating evidence suggesting that a focal brain injury may lead to increased neuronal excitability and repetitive discharges that, under certain circumstances, may kindle surrounding structures. In addition, it has been recently proposed that hippocampal damage due to febrile seizures in early childhood may lead to focal neuronal hyperexcitability resulting in the development of temporal lobe epilepsy (Chen et al. 1999). This idea is supported by investigations using glutamate analogues such as kainic acid (KA). Selective neuronal degeneration following KA treatment results in sprouting of hippocampal granule cell axons and formation of recurrent excitatory synapses that may result in abnormal repetitive discharges and a kindling-like process (Routbort et al. 1999).

7. *Summary*

Kindling is a phenomenon that is characterized by a progressive and long-lasting increase in seizure susceptibility, or epileptogenesis, resulting from the administration of periodic subconvulsant electrical or

chemical stimuli. The electrophysiological and behavioral features of kindling have been well defined and may vary with brain region. Limbic and specific neocortical regions have the greatest potential for kindling. Kindling-induced brain activation is associated with altered gene and protein expression in neurons and glia, as well as morphological changes that contribute to long-lasting modifications in neuronal structure and function. Additionally, changes in both excitatory glutamatergic and inhibitory GABAergic synaptic neurotransmission have been demonstrated in the kindled brain. Similarities between the kindling model and human epilepsies provide further justification for using kindling to study epileptogenesis in humans. Recent evidence suggests that a kindling-like process may lead to the development of epilepsy following a focal injury to the brain. A more thorough delineation of the cellular and molecular events underlying kindling will therefore provide a greater understanding of epileptogenesis in the human brain, and could lead to the development of more effective antiepileptic treatments.

See also: Electrical Stimulation of the Brain; Epilepsy; Synaptic Transmission; Electroencephalography: Basic Principles and Applications

Bibliography

Ackermann R F, Chugani H T, Handforth A, Moshe S, Caldecott-Hazard S, Engel J Jr 1986 Autoradiographic studies of cerebral metabolism and blood flow in rat amygdala kindling. In: Wada J A (ed.) *Kindling 3*. Raven Press, New York, pp. 73–90

Adams B, Von Ling E, Vaccarella L, Ivy G O, Fahnestock M, Racine R J 1998 Time course for kindling-induced changes in the hilar area of the dentate gyrus: Reactive gliosis as a potential mechanism. *Brain Research* **804**: 331–6

Chen K, Baram T Z, Soltesz I 1999 Febrile seizures in the developing brain result in persistent modification of neuronal excitability in limbic circuits. *Nature Medicine* **5**: 888–94

Clark S, Wilson W A 1999 Mechanisms of epileptogenesis. *Advances in Neurology* **79**: 607–30

Dragunow M, Robertson H A 1987 Kindling stimulation induces c-fos protein(s) in granule cells of the dentate gyrus. *Nature* **329**: 441–2

Dragunow M, Robertson H A, Robertson G S 1988 Amygdala kindling and c-fos protein(s). *Neuroscience Letters* **102**: 261–3

Ekonomou A, Angelatou F 1999 Upregulation of NMDA receptors in hippocampus and cortex in the pentylenetetrazol-induced 'kindling' model of epilepsy. *Neurochemical Research* **24**: 1515–22

Geinisman Y, Morrell F, De Toledo-Morrell L 1992 Increase in the number of axospinous synapses with segmented postsynaptic densities following hippocampal kindling. *Brain Research* **569**: 341–7

Goddard G V 1967 Development of epileptic seizures through brain stimulation at low intensity. *Nature* **214**: 1020–1

Guerrini R, Andermann E, Avoli M, Dobyns W B 1999 Cortical dysplasias, genetics and epileptogenesis. *Advances in Neurology* **79**: 95–121

Kopp J, Nanobashvili A, Kokaia Z, Lindvall O, Hokfelt T 1999 Differential regulation of mRNAs for neuropeptide Y and its receptor subtypes in widespread areas of the rat limbic system during kindling epileptogenesis. *Molecular Brain Research* **72**: 17–29

Luddens H, Korpi E R, Seeburg P H 1995 $GABA_A$/benzodiazepine receptor heterogeneity: Neurophysiological implications. *Neuropharmacology* **34**: 245–54

McBain C J, Mayer M L 1994 *N*-methyl-D-aspartic acid receptor structure and function. *Physiological Reviews* **74**: 723–60

McNamara J O, Morrisett R, Nadler J V 1992 Recent advances in understanding mechanisms of the kindling model. *Advances in Neurology* **57**: 555–60

McNamara J O, Russell R D, Rigsbee L, Bonhaus D W 1988 Anticonvulsant and antiepileptogenic actions of MK-801 in the kindling and electroshock models. *Neuropharmacology* **27**: 563–8

Michaelis E K 1998 Molecular biology of glutamate receptors in the central nervous system and their role in excitotoxicity, oxidative stress and aging. *Progress in Neurobiology* **54**: 369–415

Mody I 1999 Synaptic plasticity in kindling. *Advances in Neurology* **79**: 631–43

Perlin J B, Gerwin C M, Panchision D M, Vick R S, Jakoi E R, DeLorenzo R J 1993 Kindling produces long-lasting and selective changes in gene expression of hippocampal neurons. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 1741–5

Rauca C, Zerbe R, Jantze H 1999 Formation of free hydroxyl radicals after pentylenetetrazol-induced seizure and kindling. *Brain Research* **847**: 347–51

Routbort M J, Bausch S B, McNamara J O 1999 Seizures, cell death, and mossy fiber sprouting in kainic acid-treated organotypic hippocampal cultures. *Neuroscience* **94**: 755–65

Sato M, Racine R J, McIntyre D C 1990 Kindling: Basic mechanisms and clinical validity. *Electroencephalography and Clinical Neurophysiology* **76**: 459–72

Sheng M, Greenberg M E 1990 The regulation and function of c-fos and other immediate early genes in the nervous system. *Neuron* **4**: 477–85

Simonato M, Hosford D A, Labiner D M, Shin C, Mansbach H H, McNamara J O 1991 Differential expression of immediate early genes in the hippocampus in the kindling model of epilepsy. *Molecular Brain Research* **11**: 115–24

Sutula T, Xiao-Xian H, Cavazos J, Scott G 1988 Synaptic reorganization in the hippocampus induced by abnormal functional activity. *Science* **239**: 1147–50

Watanabe T, Morimoto K, Hirao T, Suwaki H, Watase K, Tanaka K 1999 Amygdala-kindled and pentylenetetrazole-induced seizures in glutamate transporter GLAST-deficient mice. *Brain Research* **845**: 92–6

S. Schreiber

Kinship in Anthropology

Studies in kinship, and the related institutions of marriage and the family, have until recently been central to anthropological investigation and debate.

The problem of handling the topic of kinship cross-culturally is the key to understanding the historical development of a large majority of anthropology's central analytic concepts, theories, and methods. Deliberations over the puzzles of kinship and marriage gave rise to the discipline's most sophisticated technical and theoretical elaborations, and its most virulent, ever-present controversies. It was also assumed to be the area of technical competence, the most demanding of rigor in thought, through which anthropology could best defend its scientific respectability. As a result, kinship was that aspect of social life that became the linchpin for the unfolding of all the grand paradigms of thought within anthropology, whether it be Morgan's narrative of evolutionism (1871), Malinowski's functionalism (1930), Radcliffe-Brown's structural-functionalism (1952), or Lévi-Strauss' structuralism (1969). As Fox (1967) could still comment, 'kinship is to anthropology what logic is to philosophy or the nude is to art; it is the basic discipline of the subject.' The situation has, however, changed.

Nowadays at least two major questions can be asked. First, why did anthropologists predominantly prioritize kinship over politics, economics, and religion? Second, why lately have they shown a decided *lack of interest* in kinship, or rather in 'kinship' as it was conceived traditionally within anthropology? The answers will take us to problems in the major presuppositions underlying anthropology's highly valued analytic constructs of kinship and society. Although it has been the case that kinship from the start has always been the object of strident debate, it is only since the 1960s that the legitimacy of the topic itself began to be queried. In more recent years the process of doubting has accelerated due to very basic epistemic shifts within anthropology in the wake of feminism and other modes of disciplinary self-inspection about its claims to knowledge. One prevalent conclusion forthcoming from such transformations is that much of received kinship theory is no longer seen as justifiable.

1. *The Narrative of Kinship*

1.1 *Law and Order by Another Name*

The primary puzzle for the anthropologist throughout the first half of the twentieth century was how to explain the maintenance of order within the 'simple societies' of far-flung regions where anthropologists conducted their research. Such societies lacked the basic law and order organizing institutions of Western society. They had no government to speak of, no marketplace as we know it, and no law courts, police, or armies. It was clear that they did not compartmentalize their social life into the distinct and separate institutions that we recognize as kinship, economics, politics, and religion. Anthropologists found instead that these peoples used the *idiom of kinship* to frame most of their activities, including those with political, economic, and religious intent. Analytically, the step from this insight was to view kinship to be the major institution of 'tribal' societies, and the kinship tie to be the one that compelled all others in social relations. Kinship, as the strongest of social bonds, became seen as the basis through which 'primitive' societies maintained order; it was through kinship ties that people created relations of social solidarity. Thus, 'social structure,' that is, those rules regulating the kinship, marriage, and residential institutions of a people that endow social role and identity, and which, therefore, perpetuate societal relationships, became anthropology's proper object of inquiry (e.g., see Radcliffe-Brown 1965). Everything else, a society's morals, law, etiquette, religion, politics, and education, was to be studied as but an aspect of social structure (Radcliffe-Brown 1965), or in other words its kinship system.

The emphasis anthropologists placed upon the problem of 'societal order' cannot be stressed too much. As Firth comments, the perception of order was fundamental to their investigation (Firth 1961). An underlying concern was over what could replace the authority of government in 'simple' societies, and the answer was to view kinship as having this coercive power. In other words, it was the kinship system that came to be understood as playing the same official function among 'primitive' peoples as government does in western societies. The reasoning leading to such a conclusion is circular, for kinship as a system became defined as the primary source for the rules and regulations that provided for order and continuity within 'native' society. It was presumed that the status, role, rights, duties, and obligations of a person in a 'simple society' were forthcoming from and ascribed by the person's place within the kinship system. The primary societal organization of these societies was then understood to be 'kinship-based,' and it was the 'kinship polity' (see Fortes 1969) forthcoming from either patrilineality or matrilineality that was the key concern. The slippage was simple: we have government, while they have the politics of kinship.

The evolutionist agenda underpinning such kinship theory is obvious, despite the Functionalists' claim to the contrary. It was a bias that assumed that the history of humankind's social development discloses a progression that moved from a reliance upon the natural facts of kinship, and the cultural elaboration of them, to a Western style of development that increasingly compartmentalizes the societal institutional ordering of kinship, economics, politics, and religion in such a way that kinship eventually becomes deprioritized. Anthropology's main object for study has been modern Europe's alien other, all those 'primitive' peoples attached to worlds marked as an uncivilized part of nature to be transcended and dominated by modern civilization. In large part

anthropology's technical vocabulary has denoted primitivism, and 'kinship' is no exception. First of all, the institution of kinship, more than any of the other primary domains of society, was understood to be the one most closely linked to the natural in human activities: while kinship can modify nature, it cannot transcend it. As Schneider (1984) argues, an axiom critical to kinship theory has been that the social and cultural attributes of kinship are derivative of the biological relations of reproduction. Thus, if all those alien societies studied by anthropologists were kinship-based, and if for their people the idiom of kinship took priority over economics, politics, and religion, their primitive status was further confirmed.

1.2 The Plot Thickens: The Distinction between the Domestic and the Jural

The 'law and order' thrust of traditional kinship studies, which equated the kinship system (of 'primitives') to 'society' itself, also came to include the critical distinction between the 'domestic' and 'jural' domains of kinship. The realm of the domestic was composed of relations of filiation between parents and children; it was the domain of the hearth, the family, the husband, his wife, and his children. In Fortes' terms (1969), it was the domain where the 'axiom of amity' reigned. In contrast, within the 'jural' domain the everyday relations of amity and filiation were for the most part irrelevant, for the principles of descent and lineage ruled its membership and provided the backbone for its jural structures of dominance and subordination. It was the jural domain that comprised the polity and, thus, provided society with its order and continuity. The prescriptions and regulations of the kinship polity (composed of its 'corporate descent groups') were what ruled and constrained 'primitive' people. A man's jural status, rights, and obligations within society were in essence provided by his place within the lineage of his birth. Genealogy determined one's political status, and one's rights to land and other entitlements.

Given the above narrative of the place of kinship in 'native' societies, we can understand that the received wisdom of kinship theory until as late as the 1970s was that unilineal descent systems were necessary as a sticking plaster of 'primitive' societal order—despite increasing evidence to the contrary. As Radcliffe-Brown (1965) asserts, 'unilineal institutions in some form, are almost, if not entirely, a necessity in any ordered social system.' Even Lévi-Strauss, in his major critique of descent theory in kinship studies, finds the existence of unilineal descent essential to the logic of his model of elementary structures of marriage exchange. He says that this is because the social cohesion of elementary systems of kinship that are premised on the notion of groups of men exchanging wives requires a rule of descent, for the groups themselves must be defined by such a 'stable' rule of descent (Lévi-Strauss 1969). One should be aware, however, that Lévi-Strauss began his formidable attack upon the prevalent view of 'descent as societal order' in 1949, and that his stress upon alliance over descent as the salient ordering principle of 'primitive society' did not become part of mainstream debate until the 1960s (e.g., Leach 1961, Needham 1964), but also see the earlier debate between Radcliffe-Brown (1953) and Dumont (1953). The concern of Lévi-Strauss, it is to be noted, was with creating a minimal model of society by showing the ways in which kinship groups were integrated through rules of marriage exchange, that is, he wished to demonstrate how classifications of kinship and marriage logically provided a broader level of societal integration than that achieved through rules of descent alone.

The rather chauvinistic reductionism of the structural functionalist and structuralist grand paradigms of society and societal ordering, where society itself was to be equated with male structures of order and control, took another couple of generations to unveil and unravel.

2. The Question of Definitional Rigor

2.1 Kinship, Descent, and Marriage

In itself, a fuller ethnographic record began steadily to undermine many of the major analytic constructs of kinship and societal order, particularly the idea that 'primitive society' was universally based upon exogamous, corporate, land-holding, unilineal descent group structures. By the 1950s and 1960s reports of field research especially from the Pacific on kinship systems that were not premised on a unitary rule of unilineal descent but upon cognatic, bilineal, ambilineal, or double descent principles became legion (cf. Bohannan and Middleton 1968). The debate over the unitary view of unilineal descent basically was closed by the influential article by Scheffler (1966), who was able to demonstrate through the ethnography by then at hand that notions of descent were used among different peoples, and by the same people, toward highly varied ends, and not necessarily toward that of corporate group structure.

By the 1970s, for instance, through the ethnography from Amazonia that began to enter mainstream debates on kinship and marriage, it became clear that the notion of descent itself could hardly be declared a universal principle of 'primitive' social ordering, for there were peoples who did not recognize a principle of descent as relevant for any social or political purpose (e.g., see Rivière 1969, Overing 1975). Even kinship and marriage as analytical constructs *per se* came also under attack, especially the notion of achieving any type of unitary view of either (but also, see Leach's

(1961) earlier hacheting of these sacred constructs of societal ordering). We find Needham (1971) announcing flamboyantly that 'there is no such thing as kinship, and it follows that there can be no such thing as kinship theory!' He was referring to 'minimal' definitions of kinship when framed in the context of genealogically reckoned jural rights, such as their allocation and transmission from one generation to the next. Cross-culturally, the ethnographic evidence could not uphold a totalizing view that assumed a predictable relationship between the cultural constructs of kinship and its classification, social roles, rights, and obligations, and the allocation of individuals to particular types of social groups. It was this unitary package of kinship as part and parcel of particular politico-jural orders that earlier anthropology had indeed upheld.

Even more courageous for the times are Rivière's (1971) queries into the analytic concept of marriage. He argues against any jural definition of marriage, and suggests instead that the institution of marriage first be viewed structurally, as one of many relationships conceived possible between men and women. It was his reading of the ethnographic literature that anthropologists had been defining institutions cross culturally as 'marriage' when the said institutions in fact had 'no feature in common other than that they are concerned with the conceptual roles of male and female' (Rivière 1971). In other words, to understand what marriage is for any given people the question of the cultural construction of gender relationships must be understood, rather than the jural relations between groups of men that entail their exchange of women.

It is significant that Rivière's fieldwork experience had been with indigenous peoples of Amazonia, for whom anything approaching a 'jural' relationship usually would be stretching the point, as too would lineages and descent-group ordering as normally discussed in the literature. It was as difficult to find corporate land holding groups among Amazonian people, as the elders who might rule them. There were no *groups* of men forming ties of alliance through the exchange of their women. Instead, ties of marriage, which moreover were highly salient to Amazonian constructs of sociality, were linked to a principle of cognation and not descent. As a consequence the contributions of Amazonian specialists, in line with Dumont's (1957) reading of marriage alliance in India, played a major part in the later reinterpretation and unraveling of both alliance theory as first formulated by Lévi-Strauss and descent theory as formulated by Fortes and Radcliffe-Brown.

2.2 The Structural Analysis of Kinship Terminologies: The Power of the Paradigm

It is by now obvious that we cannot achieve an analytic definition of the construct of kinship and related institutions that would be both universally adequate and at the same time respectful of indigenous understandings and knowledge. In short, anthropology cannot, even if it wished, arrive at a universal definition of kinship. Part of the dilemma is of course linguistic, insofar as most of the important analytic terms of anthropology have not only been highly abstract but also continued to carry the complex historical baggage of western thought and practice. Terms such as society, community, family, kinship, descent, lineage, or structure, function, and system, are to be used at the peril of totally eluding another people's understanding of what they are doing socially—which, one would think, is the very *raison d'etre* of the anthropological task. As anthropology slowly came to realize, any 'definitional rigor' to be achieved through the use of any of these terms is well-nigh impossible. Such constructs tend to sit within particular and forceful paradigms of social order, and therefore carry all the litter of such grand narratives. Structural functionalism was followed by structuralism, and the structural analysis of kinship terminologies was a particularly obvious case of a paradigm so powerful that its highly reductive results killed for the time being the possibility of interesting further advances being made in kinship theory, the very area where anthropology was once so creative and rich in debate.

The overwhelming allure of structural analyses of kinship can be ascribed to the power of their methods which wed anthropology to advances in modern linguistics, a field considered to have become the most scientific of the human sciences. The methods and models of formal analysis were so enlightening of previously misunderstood structures that they gave promise to a mathematical rigor and definitional clarity that would transform anthropology into a 'true science.' Their initial success was so stupendous (e.g., see Lounsbury 1968) that they made seeming child's play of previous attempts to provide order to the complicated structures of many 'native' kin terminologies. Anthropologists were taught to be more rigorous in discerning the logical differences between systems. Such sophistication in method was greatly needed in anthropology, and for its example we can be only grateful.

However, as Overing (1987) has noted, the method became confused with world view. The logic of the method through sleight of hand became equated to the logic of terminological use, and thus also with indigenous thought. As a result, we arrived once more at the 'universal' to what Schneider so aptly derides as the anthropological 'Doctrine of the Genealogical Unity of Mankind' (Schneider 1984) or the genealogical meaning of kinship terms. Kinship is, everywhere, first and foremost about genealogical relatedness (see especially, Scheffler 1978), a resoundingly uninteresting conclusion to come in the wake of such dazzling structuralist performances. It

was also a judgment that was highly suspect. It appears that the method itself allowed for no other interpretation because the meanings of terms were made to fit not only the domain of logic, but also western commonsense notions about what kinship is. The demand itself of formal analysis for logical rigor reduces the number of its elements that account for meaning to very few: affinity as well as consanguinity can be allowed, but not the further complications of what such notions possibly mean from the indigenous point of view. It was the niggling doubt over this issue (where are all the people?) that drove many anthropologists away from the technical chore of analyzing kinship logics. As Campbell (1985) remarks, the very abstract level at which structural analysis operates is about the tenth remove from anything going on in daily practice and thought.

The more recent formal analyses of kinship terminologies in the hands of scholars such as Héritier (1981) move in the more promising direction of not assuming indigenous meaning. The aim instead is to specify the structural possibilities of a given classificatory system that in turn frame or limit the choices made in reality. In short, the exercise is no longer directed toward the task of uncovering hidden layers of meaning, but to explore the interplay between structure and the semantics of actual use.

3. *Is There Hope for Kinship?*

Schneider (1984) advises anthropologists not only to stop looking for 'kinship' but also all the other metacultural categories of economics, politics, and religion which, together with kinship, form a quartet of societal domains embedded in European culture and have been willy-nilly incorporated into the European social scientist's analytic schemes. While we can heartily agree with Schneider's full blown rampage against the anthropological treatment of the topic of kinship, such concordance does not entail the dismissal of the study of those social relationships, and their classifications, that were once, more or less, subsumed under the label of 'kinship.' People do bear children, and there is a social framework through which they do so, and through which these children are raised to become adult members of human social groups. The members of these social groups follow particular practices in the course of which relationships that are highly significant for them are developed, as too are very interesting ways of thinking about them. With all this we can agree. The overriding question still remains—how do we understand and translate such practices, relationships, and ways of thinking?

Happily, since the 1980s, anthropologists have developed a myriad of ways to approach subjects that formerly would have been classified under the general rubric of 'kinship.' While nowadays the topic of something called 'kinship' does not loom large in the literature, such issues as 'self,' 'agency,' 'gender,' and the life of values and affect, do. The topics of personhood, emotions, and aesthetics are much more likely to take their place in the titles of doctoral theses than those of kinship, affinity, and jural rules. Thus, we see that the 'technical' language of anthropology has been transformed in the wake of the shifts of attention away from something we once called the 'jural-political domain,' with its contrast to something we labeled 'the domestic group,' and equally removed from the notions of social structure and prescriptive behavior, and those of rights and obligations. In their place, idioms of equality and inequality are now being explored, and such values as nurture, sharing, generosity, and those of peace and violence. The stress is upon ambiguity, flux, the personal everyday, and a multiplicity of voices, rather than upon grand structures of mind and society and societal rules and regulations. The emphasis tends to be upon context and the performative, and not underlying and hidden rules of practice and thought. Such shifts in direction have often been undertaken in a spirit of rebellion, by feminists but also many others, against the ethnocentrisms underpinning the grand narratives of anthropology. The gain has been that anthropologists are presenting very different pictures, certainly in their richness, from those of yesterday of the ways other people view, act, and experience the world of the social. These depictions in themselves are further enlightening of the previous 'sins' of reduction, prejudice, and ethnocentrism.

3.1 *The Impact of Gender Studies*

It has made all the difference to the ethnographic endeavor to include the subject of women and especially women's voices. In the process we have time and again discovered a rich symbolic world that demands an understanding of the interplay of men and women, and their respective knowledge, that makes travesty of the simplistic 'male only' models of yesteryear. It is only by understanding both male and female (often complementary or conflicting) perspectives that gender relations among another people can begin to be comprehended. As gender studies progressed, it became absolutely clear that the authoritative voice of the ethnographer was insufficient, as too that of the male 'informant' (not a designation that we can use today!). Nor can we speak from the perspective of a generic female, for there is no such thing. There is ever a plurality of voices, a seemingly helter-skelter chiming that provides, significantly, fertile layers of evaluative contextualization, the recognition of which has in the end changed our anthropological visions of culture and society.

All those kinship structures through which men established important relationships with each other through the exchange of their muted women, which

became the model of 'society'; all those 'political' roles and statuses through which men who controlled the political domain came to be the knowledge-holders of their culture—these were the topics that once were recognized as primary anthropological concerns. From the 1970s onwards, once women were inserted firmly within the context of 'establishment' discourse in anthropology, the consequence was a true 'Kuhnian' shift of paradigm. The epistemological and methodological foundations of the discipline were indeed shaken when females, fifty percent of the world's population, finally entered the picture. Thus, the 'question of women' did play a large part in upsetting established ways of thinking and practice in anthropology, and not least in regard to the topic of kinship.

One major contribution of the early writings on gender was to make anthropology clearly aware of its exclusion of women as a topic for study, and how such oversight was, moreover, linked to the male bias of dominant theoretical and methodological assumptions of the discipline, and beyond that to major strands of western thought in which anthropology played its part. In *Women, Culture & Society*, its authors show how women, like men, transculturally 'are social actors whose goals and strategies are intrinsic to the process of social life' (Rosaldo and Lamphere 1974), and that 'even in situations of overt sex role asymmetry women have a good deal more power than conventional theorists have assumed (Rosaldo and Lamphere 1974). It became quickly obvious that major presuppositions and procedures of the discipline needed to be reconsidered radically in order to include the agency of women. In other words, studies of social life and its organization needed to include women as actors and speakers and not as mere passive, muted objects of men. It slowly came to be seen that women as well as men played a major role in the creation of social ties of community life.

There is by now a large amount of ethnography telling of indigenous peoples among whom the women, both ideologically and in practice, control their own labor, and the products of it (see much of the contemporary literature on Amazonia). Likewise, values—and the structures of equality and inequality linked to them—may not be so straightforward, but ambiguous, as among the Hagen of New Guinea where there is constant play between egalitarian and hierarchical principles (e.g., Strathern 1988). There are also now many ethnographic examples showing that the quality of the relations between the genders may well be subject to perspective, where females have a very different view from males of the strength of their respective roles and participation within the social life of the community (e.g., Keesing 1987). This stress in studies of gender upon the critical importance of presenting perspective and thus allowing for the 'authorship of many' has led to energetic debate over the epistemological foundations of anthropology, which in turn has transformed the question, the topics, and the methodology of the discipline—and in the end its own self image as having the right of authorial privilege.

3.2 Kinship by Another Name? Networks of Relationships and Personal Kind Terms

We find that Schneider and Needham, as pioneers in the deconstructing of key concepts of kinship theory, were merely tapping the surface of a modernist creation for which the very notion of kinship was but one aspect of a multifracteled edifice filled with assumptions about society and the social order, that in turn were tied to networks of ideas about the relation of the family to other societal institutions, and the relationship between the sexes, the private and the public, the dominant and the subordinate, all of which were implicated further by presuppositions about the nature of human existence, and its progress, which in addition were premised upon ontological notions of the priority of reason over the emotions, and so it continued through an enormous number of other dualisms and bundles of relations pertinent to the western imagery of society and the world, and the elements of which they were comprised. Our notion of kinship carried with it the interarticulations of this entire structure.

The interesting lesson that has been learned more recently through changing the types of questions anthropologists' ask is that other people's views of the social relationships of everyday life are as enmeshed as our own within wider networks of meaning. They also include ontological presuppositions about the nature of human existence and capabilities for sociality that are linked to bundles of relationships and arrays of ideas about the world that are as complex as our own. How do we then understand these other sets of linkages, so unlike our own? That has become a primary question. The quest for understanding their interconnections has in part required the unpeeling of our presuppositions about reproductory and biological processes, parenting, the nature of the material world, and the interconnections of all of these things to what we call Society (cf. Strathern 1992).

Overing (1985) has suggested that to understand better the complexity of indigenous social thought, we should change the label of what we have been calling 'kinship terms' to 'personal kind terms.' This involves a radical switch in perspective that concomitantly raises the conceptual status of these terms to one more closely aligned to the indigenous view and practice. As with many of our own constructs of 'natural kinds,' 'personal kind terms' are also highly abstract, philosophically important concepts that defy unitary definition. They share the openendedness and elusiveness that is typical of all abstract terms that comprise complex relational properties. The difference is that

personal kind terms do not refer to the world of nature, which is the western domain of competence, but to qualities of personal relationships, the area that indigenous people have opted for theoretical elaboration (cf. Horton 1979). We have reduced their personal kind terms to our own very weak language of kinship, one that speaks of 'consanguinity,' 'affinity,' 'social category,' and 'amity,' which is often a bad mistake. These terms, as used in everyday life, can have metaphysical weight that goes well beyond our western notions of a 'biological' relationship or one through marriage. Moreover, the quality of the relationship they demonstrate, as, for instance, one of nurturing, teaching, treachery, competition, or predation, may override a more 'physical' kind of relating. Will this man or woman work tranquilly with me, or have predatory designs that will make me ill?

Certainly in Amazonian ethnography, the emphasis today in investigations of peoples' use of relationship terminologies is often upon the metaphysical and/or moral loading of the classifications (e.g., Teixero-Pinto 1997, Viveiros de Castro 1992, Overing 1985). For instance, Kidd writes about the Enxet of Paraguay that:

> Their understanding of why they act as they do centres on their concept of the *waxok*, an aspect of the self that is both intensely private and inherently social. They insist that their social behaviour—both appropriate and inappropriate—can be explained by the physical—or metaphysical—state of the *waxok*. Furthermore, because the *waxok* is also the centre of cognition, people can also consciously transform it so as to enable themselves to act in either a self-centred or other-regarding manner. It is an explanation that, I believe, we should take seriously if we want to understand indigenous social life and it is one that finds its root in the practice of child-raising, in the creation of 'good/beautiful' people who have been taught not only how to think but how to feel. It is this *waxok*-centred combination of thinking and feeling that enables the Enxet to act appropriately and which, ultimately, guides them as they strive to generate sociality and engender tranquillity (Kidd 1999).

The peoples of Amazonia have been difficult for anthropologists to believe, and to an interesting extent still are (cf. Overing 1999). For many of these people the personal ties of parenting, nurturing, sharing, and pooling are not premised primarily on a notion of 'biological linkage' or a 'linkage through blood,' or membership within a jural group, but which instead are generated over time through consistent and processual action. As such, the personal quality of a relationship as made manifest through everyday practice is paramount to its classification. In other words, the ongoing quality of a relationship (as in the 'growing' of a child by a parent and child, the mutual care of brothers, or sisters, or of any other personal relation) may have generative value in a material sense that goes far beyond the minimal possibilities that are endowed through the act of sex (e.g., Gow 1991). For sure, the personal kind terms that are here being applied to their nearest and dearest, and to those further afield, are about reproductory possibilities, but it is not 'reproduction' in the sense that the anthropologist imagined it in traditional kinship theory.

4. *Kinship is Alive and Well!*

In fact, we can say that kinship studies are as alive and well as always in anthropology, in that the personal relationships and activities of parenting, nurturing, and the processes of generating and gendering of bodies into social adulthood hold center attention. Such studies are, however, unrecognizable as pertaining to the kinship theory of yesteryear. A similar reorientation of concentration is as clear in Melanesian studies as in those of Amazonia (e.g., Strathern 1988) where personal relationships are discussed through categories very different from the former ones of prescriptive rule, roles and statuses, and social structure, and where the talk instead, as among Amazonianists, is upon the indigenous understandings of such matters as gender distinctions, the contents of the self and the mastery of them, and the construction of human bodies; it is about indigenous ambiguities over the nature of personhood and the various possibilities of agency in this world and others, and the elaborate relation of these issues to indigenous practice, the life of affect, and their metaphysics. In other words, a dialogue is being created between us, the anthropologists, and them, the peoples of New Guinea or Amazonia, over what it means to be human in this world, or female humans, male humans, or betwixt and between? What does it mean to be social beings in this world? How do we go about attaining this state? If for other people the western grand distinctions between society and nature do not hold, what other possibilities of an interesting kind are there? These are very different questions from those asked in mainstream anthropology in the 1970s and 1980s.

See also: Ancestors, Anthropology of; Clan; Family, Anthropology of; Genealogy in Anthropology; Marriage

Bibliography

Bohannan P, Middleton J (eds.) 1968 *Kinship and Social Organization*. Natural History Press, Garden City, NY

Dumont L 1953 The Dravidian kinship terminology as an expression of marriage. *Man* **53**(224): 143

Dumont L 1957 Hierarchy and marriage alliance in South Indian kinship. Occasional Papers of the Royal Anthropological Institute, no.12

Firth R 1961 *Elements of Social Organization*. Beacon, Boston [1951 Watts & Co., London]

Fortes M 1969 *Kinship and the Social Order: The Legacy of Lewis Henry Morgan*. Routledge & Kegan Paul, London
Fox R 1967 *Kinship and Marriage: An Anthropological Perspective*. Penguin, Harmondsworth, UK
Gow P 1991 *Of Mixed Blood*. Clarendon Press, Oxford, UK
Héritier F 1981 *L'Exercice de la Parenté*. Gallimard: Seuil, Paris
Holy L 1996 *Anthropological Perspectives on Kinship*. Pluto Press, London
Horton R 1979 Material-object language and theoretical language: Towards a Strawsonian sociology of thought. In: Brown S C (ed.) *Philosophical Disputes in the Social Sciences*. Harvester Press, Brighton, UK
Keesing R 1987 Ta'ageni: Women's perspective on Kwaio society. In: Strathern M (ed.) *Dealing with Inequality: Analysing Gender Relations in Melanesia and Beyond*. Cambridge University Press, Cambridge, UK
Kidd S Love and hate among the people without things: The social and economic relations of the Enxet people of Paraguay. 1999 PhD. thesis, University of St Andrews, UK
Leach E R 1961 *Rethinking Anthropology*. Athlone Press, London
Lévi-Strauss C 1969 *The Elementary Structures of Kinship*. Beacon, Boston [1949 *Les Structures Élémentaires de la Parenté*. Presse Universitaires de France, Paris]
Lounsbury F 1968 The structural analysis of kinship semantics. In: Bohannan P, Middleton J (eds.) *Kinship and Social Organization*, 1st edn. Natural History Press, Garden City, NY
Malinowski B 1930 Kinship. *Man* **30**: 9–29
Morgan L H 1871 *Systems of Consanguinity and Affinity of the Human Family*. Smithsonian Institution, Washington, DC
Needham R 1964 Descent, category, and alliance in Siriono society. *Southwestern Journal of Anthropology* **20**: 229–40
Needham R 1971 Remarks on the analysis of kinship and marriage. In: Needham R (ed.) *Rethinking Kinship and Marriage*. Tavistock Publications, London
Overing J 1985 Today I shall call him 'Mummy': Multiple worlds and classificatory confusion. In: Overing J (ed.) *Reason and Morality*. Tavistock Publications, London
Overing J 1987 Translation as a creative process: The power of the name. In: Holy L (ed.) *Comparative Anthropology*. Blackwell, Oxford, UK
Overing J 1999 Elogio do cotidiano: A confiança e a arte da vida social em uma comunidade amazônica (In praise of the everyday: Trust and the art of sociality in an Amazonian community). *Mana* **5**(1): 81–108
Overing K J 1975 *The Piaroa, a People of the Orinoco Basin*. Clarendon Press, Oxford, UK
Radcliffe-Brown A R 1953 Dravidian kinship terminology. *Man* **53**(169): 112
Radcliffe-Brown A R 1965 *Structure and Function in Primitive Society*. Free Press, New York [1952 Cohen & West, London]
Rivière P 1969 *Marriage among the Trio: A Principle of Social Organisation*. Clarendon Press, Oxford, UK
Rivière P G 1971 Marriage: A reassessment. In: Needham R (ed.) *Rethinking Kinship and Marriage*. Tavistock Publications, London
Rosaldo M Z, Lamphere L 1974 Introduction. In: Rosaldo M Z, Lamphere L (eds.) *Woman, Culture and Society*. Stanford University, Stanford, CA
Scheffler H W 1966 Ancestor worship in anthropology or, observations on descent and descent groups. *Current Anthropology* **7**: 541–51
Scheffler H W 1978 *Australian Kin Classification*. Cambridge University Press, Cambridge, UK
Schneider D M 1984 *A Critique of the Study of Kinship*. University of Michigan Press, Ann Arbor, MI
Strathern M 1988 *The Gender of the Gift: Problems with Women and Problems with Society in Melanesia*. University of California Press, Berkeley, CA
Strathern M 1992 *After Nature: English Kinship in the Late Twentieth Century*. Cambridge University Press, Cambridge, UK
Teixeira-Pinto M 1997 *Ieipari*. ANPOCS/HUCITEC, São Paulo, Brazil
Viveiros de Castro E 1992 *From the Enemy's Point of View: Humanity and Divinity in an Amazonian Society*. University of Chicago, Chicago

J. Overing

Kinship Terminology

Kinship terminology forms the most complex, orderly, and cross-culturally variable, part of a language's vocabulary. These characteristics make it central to semantics (Sect. 1). At the same time the use of kinship terms as address terms brings in considerations of footing, which can detach the logic of address usage from that of referential use, necessitating more contextual accounts of kin term use (Sect. 2).

Kinship terminologies often reflect the culturally patterned ordering of social categories of descent, marriage, and alliance, and this has ensured their enduring interest to anthropologists (see *Kinship in Anthropology*). It is now clear that the diversity of kinship systems around the world boils down to a restricted set of variants on a few basic types. These typologies of kinship systems are discussed in Sect. 3.

1. Semantics of Kin Terms

The challenges posed by kin terms can be illustrated by comparing English and Marri-Ngarr, an Australian language (Scheffler 1978). Fig. 1 illustrates the Marri-Ngarr system schematically, while Table 1 lists the full set of denotata for selected terms. In analyzing kinship, we distinguish *kin terms*, which are emic (language-internal) categories, from etic (universally discriminable) *kin types*, the set of denotata a particular term includes. Thus the English kin term 'uncle' includes (at least) the kin types father's brother (FB) and mother's brother (MB). Other abbreviations we will use are C (child), S (son), D (daughter), H (husband), W (wife), Z (sister), m (man speaker), w (woman speaker), e (elder), y (younger).

To avoid clutter, many kin types known by the same kin term are omitted from Fig. 1. For example, in Marri-Ngarr same-sex siblings always receive the same

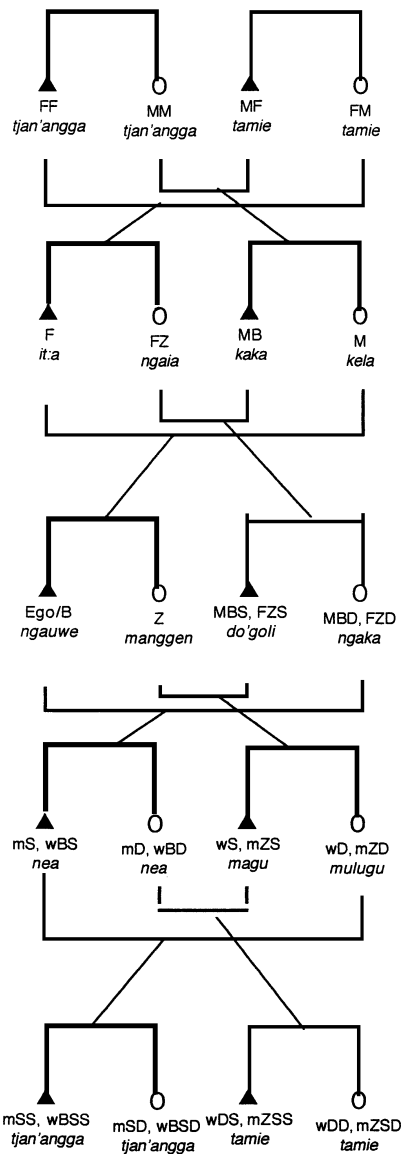

Figure 1
Marri-Ngarr kinship system (source: Falkenberg 1962, p. 39)

kin term: F and FB are both known by *it:a*; but only the first is given in Fig. 1, though both are listed in Table 1. Further, since a FB is equivalent to a F, his sons are equivalent to father's sons, in other words to brothers, meaning that this type of cousin (FBS) also falls into the 'brother' category, *ngauwe*.

The first dimension on which these systems differ, then, is that Marri-Ngarr but not English is 'classificatory'—by using certain principles of extension, it extends indefinitely outwards to encompass the entire social universe. A corollary is that the distinction between consanguineal (or 'blood') and affinal relatives (related by marriage) is hard to maintain, since any potential in-law is already related as a blood relative: the rule favoring marriage to certain cousins (FZC) means that a man's FZH who, as in English, is a type of uncle (*kaka*), is at the same time a potential father-in-law. In English, on the other hand, kin terms take in only a fraction of the social universe, and in general consanguineal and affinal terms are distinct.

The second striking difference is in how the denotational space is carved up. We focus on just one of the many differences that exist. Both languages have four terms to cover parents and their siblings, but in English father and mother, who are 'lineal' relatives lying directly on the descent line, are distinguished from the 'collateral' relatives uncle and aunt, which take in any siblings of parents, whether same sex or opposite sex. However, in Marri-Ngarr, parents' same-sex (*parallel*) siblings—FB and MZ—are known by the same term as the parents themselves, in contrast to parents' opposite-sex (or cross-) siblings, who receive distinct terms (Fig. 2).

The cross vs. parallel distinction, although irrelevant to English, pervades the Marri-Ngarr system. The distinction between 'cross-aunts' and 'cross-uncles' (*ngaia* and *kaka*), related to one's parents by an opposite-sex sibling link, and 'parallel' aunts and uncles (*kela* and *it:a*) is the key to understanding how 'cousins' get divided into two categories: cross-cousins (*manggen* if female, *do'goli* if male)—children of cross-aunts and cross-uncles and therefore potential spouses—vs. parallel cousins, the children of one's parallel aunts and uncles, who are treated as siblings and are not possible marriage partners. The cross vs. parallel distinction is reflected in kinship terminologies around the world and there are good social reasons for its importance. If membership in social groupings is passed down a patrilineage or matrilineage, then you are in the same group as your parallel relatives but in a different group from your cross-relatives.

The complexity and cross-linguistic variability of kin terms makes them a fertile testing ground for theories of semantics. During the structuralist semantics period of the 1950s and 1960s a common approach was to represent meanings as clusters of abstract features, a technique known as *componential analysis* (see D'Andrade 1995 for a recent evaluation). The appeal of these analyses lay in their economical use of small numbers of components which combine to generate large numbers of terms, and in the way in which particular terms with wide reference could be defined through the use of abstract components which range over a large class of kin types.

Burling (1970) illustrates the application of componential analysis to Njamal, whose kinship system closely resembles the Marri-Ngarr system above. The Njamal term *mama*, for instance, which takes in a comparable range of kin types to the Marri-Ngarr *it:a*,

Table 1
Full set of denotata for selected Marri-Ngarr kin terms

	Term	Gloss	Primary denotata	Other denotata	English translation equivalents
1	*tjan'angga*	Parallel grandkin	FF, MM	FFB, FFZ, MMB, MMZ, FMBW, MFZH, mSC, wDC, ZDC, BSC, HZCC, WBDC	Grandfather, grandmother, great uncle, great aunt, grandson, granddaughter, great nephew, great niece
2	*it:a*	Father	F	FB, FFBS, FMZS, MMBS, MFZS, WMB	Uncle, second cousin once removed, mother-in-law's brother
2a	*it:a il:ingeri*	Wife's mother's brother	WMB		Mother-in-law's brother
3	*ngaia*	father's sister	FZ	FFBD, FMZD, MMBD, MFZD, MBW, WM	Aunt, second cousin once removed, mother-in-law
3a	*ngaia il:ingeri*	Wife's mother	WM		Mother-in-law
4	*kela*	mother	M	MZ, MMZD, FMBD, FFZD, FBW, WFZ	Mother, second cousin once removed, father-in-law's sister
5	*kaka*	Mother's brother	MB	MFBS, MMZS, FMBS, FFZS, FZH, WF	Uncle, second cousin once removed, father-in-law
5a	*kaka boi*	Wife's father	WF		Father-in-law
6	*ngauwe*	Brother	B	S of (3), (5)	Brother, cousin, second cousin
7	*ngaka*	Sister	Z	D of (3), (5)	Sister, cousin, second cousin
8	*manggen*	Female cross-cousin	MBD, FZD	D of (4), (6), WZ, BW, HZ	Cousin, sister-in-law
9	*do'goli*	Male cross-cousin	MBS, FZS	S of (4), (6), HB, WB, ZH	Cousin, brother-in-law

Source: Adapted from Scheffler 1978, pp. 120–1.

aunt		*mother*	*father*	*uncle*	
FZ	MZ	M	F	FB	MB
ngaia	*kela*		*it:a*		*kaka*

Figure 2
Parent and uncle/aunt terms in English and Marri-Ngarr

is defined as G^{+1} $M^{e}S^{m}$, i.e., as a relative in the first ascending generation who is male and belongs to the same patrimoiety as ego. (Patrimoieties divide Njamal society into two halves, with membership in each patrimoiety inherited from one's father; marriage must be with a member of the opposite patrimoiety, which means that, for example, your mother and her brother will be in the opposite patrimoiety to you.) Abstract features such as 'same/opposite patrimoiety' enable the analyst to account for the wide denotational range of such terms as Njamal *mama* or Marri-Ngarr *it:a*.

However, componential analysis of kin terms has been attacked from a number of quarters. Kay (1975) points out that the use of semantically uninterpreted features without a predicate-argument structure removes kin terms from standard logical relations such as entailment, and obscures the fact that kin terms are two-place predicates (and in many languages kin terms are actually verbs—see Evans 2000). He proposes an alternative logic-based system of representation.

From another angle, Wierzbicka (1986) attacks the hidden circularity of using terms such as 'ascending generation' to define a word such as *mama* without mentioning 'father,' since 'ascending generation' can only be defined with respect to parenthood. She proposes a definition of 'father' which refers outside the kin domain, drawing on his role as begetter (at least for English and Njamal). By allowing the analyst to spell out the relevant cultural components in the definition, this approach has the advantage of showing exactly how a concept such as 'father' may vary across cultures—for example, is the father the begetter, or merely the man married to the woman at time of giving birth to the child? Non-kin primitives are also useful in predicting metaphorical extensions of kin terms, such as the use of 'father' to mean 'priest' (Kronenfeld et al. 1985).

An attack from a third direction came from extensionist approaches (Lounsbury 1964, Scheffler 1978), which heralded a paradigm shift away from 'categorical' theories of meaning, which hold that concepts can be defined satisfactorily in terms of necessary and sufficient conditions applying equally to all members, toward prototype theories which allow for varying degrees of membership from focal to peripheral. For example not all 'uncles' in English are equally good exemplars—all speakers will agree that MB and FB should be included, while for FZH and MZH speakers may disagree, and some will prefer hedging expressions such as 'sort of an uncle' or 'uncle by marriage.' MB and FB are therefore *focal* members of the uncle category, while FZH and MZH are *peripheral* members.

Scheffler still uses componential analysis for focal category members: the primary denotatum of Marri-Ngarr *it:a*, F, is defined as K.L.G1.+.♂ i.e., parallel lineal male kin of first ascending generation. However, to get the remaining members of the category, he employs various *extension rules*. For example, a *same-sex sibling merging rule*, which places same-sex siblings in the same category, extends *it:a* from F to FB. Half a dozen extension rules, chained together if necessary, can generate all the extensions in column 5 from the primary denotata given in column 4 of Fig. 1.

2. *Kin Terms in Interpersonal Context*

So far we have been maintaining a traditional idealization, stemming from the use of the 'genealogical method,' which has long been used to elicit kin terms while obtaining genealogies: one asks how the informant addresses and is addressed by each person in the genealogy (Barnard and Good 1984). In fieldwork this method is a necessary step in obtaining an initial set of terms, but it has the drawback of examining kin terms in isolation from their actual use, thus bracketing out the contributions of politeness, conversational footing, and discourse goals to usage (Zeitlyn 1993). In Ungarinyin, for example, which normally distinguishes even- and odd-numbered generations, some terms are extended to take in all generations in a lineage when discussing patrilineal clan estates (Rumsey 1981). Within the functionalist program in linguistics, which sees language structure as resulting from the pressures of actual use, context holds the key to understanding changes in the system itself.

Once kinship terms get used in address, further factors come into play. Often languages have distinct kin terms for address ('Dad!)' and reference ('My father was wise'), although the line may be hard to draw. For reasons of politeness, kin terms often have a wider range when used as address terms, extending to nonkin, as with Indonesian (*i*)*bu* 'mother,' which in address may extend to older females more generally (then dropping the first syllable).

With its long documented heritage, Chinese offers an unparalleled opportunity to trace how address and reference terms change over time. Kryukov (1998) shows how certain innovations start among address terms, later spreading to reference terms. In the Book of Song (first millennium BC), both father and his

siblings were referred to by the general term *fu*, but in address the distinct term *shufu* could be used for FyB. By mid-millennium, however, *shufu* also appears as a reference term. This is typical of how, because they respond to interpersonal context, address terms are usually the more dynamic segment of the kinship system, giving them a special importance for understanding how kinship terminologies change.

Address, however, is not the only way interpersonal context impacts on kinship terms. As mentioned, kin terms are fundamentally two-place predicates, and in many languages there is obligatory marking of the reference point or *anchor* by a possessor suffix: for example, Dalabon *bulu-ngan* 'father-my.' In languages such as English, by contrast, the anchor is typically implicit and must be recovered by pragmatic inference. Speakers often have the choice of whether to take themselves, the hearer, or some third person as anchor. According to context, the English word 'Mum' can be interpreted with an anchor that is first-person (e.g., said by a child to an adult), second-person (e.g., said by an adult to a child), or third-person (e.g., said by a child-care worker taking the child's perspective).

In all these examples the kinship predicate is calculated with respect to just one anchor. However, kinship terms may also make two anchors simultaneously explicit in special 'trirelational' systems, such as the Mayali terms known as Gun-dembui: alongside basic terms like *garrard*, 'mother,' are terms such *al-garrng*, 'the one who is your mother and my daughter,' and its converse *al-doingu*, 'the one who is your daughter and my mother.' See Merlan (1989) for discussion of a similar system in Jawoyn.

The ability to take another's perspective is a hallmark of mature 'social intelligence,' and studies of how speakers manipulate the choice of anchor have important implications for the development of social cognition. Concomitantly, the need constantly to shift away from an egocentric viewpoint in reckoning kinship makes it a difficult semantic field for children to master. Piagetian-style studies of how children acquire kin terms have examined the development of relational abilities (Danziger 1957, Luong 1986).

3. *Typologies of Kinship*

The systematic comparison of kinship terminologies around the world, begun with Morgan's *Systems of Consanguinity and Affinity of the Human Family* (1871), has made kinship vocabulary the most widely sampled and analyzed semantic field, leading to the finding that 'the thousands of systems of kinship terminology in human societies around the world are but variants of a limited number of different types' (Godelier et al. 1998, p. 5).

Consider sibling terminologies. Nerlove and Romney (1967) examined 245 languages for the patterning of terms across the eight logical kin types given by the product of *sex of ego* × *sex of referent* × *relative age*. Common patterns are the use of a single term ('sibling') for all eight kin types, and the division into two terms based on sex of referent (as in English 'brother' vs. 'sister'). Strikingly, they found that of the 4,140 logically possible types only 10 were attested in more than one language of the sample.

The dominant typologies of kinship terminology take the patterning of parent/aunt/uncle terms to be the best predictive feature of the rest of the system. The relevant patterns are illustrated in Table 2, which gives the three commonest type-terms used in the literature.

While the first four types are each widely attested, the fifth, although logically possible, is unattested, presumably because it would require a disjunctive definition (F *or* MB), whereas the extended categories in the other types allow one to find a common factor (e.g., 'male relative of one's patriline' for the Iroquois type). Here the systematic absence is probably attributable to cognitive rather than social causes, since disjunctive categories are widely dispreferred (Greenberg 1980, Hage 1997).

The central role of parent/uncle/aunt terms in typologies of kinship follows from their part in regulating marriage possibilities. The commonest type of marriage partner among relatives is some type of cousin, and the figuring of cousin relationships depends in turn on the classification of parents' siblings.

The Dravidian system is a widely cited example of the relation between kin terms and marriage rules, since it is widely associated with exchange of spouses between descent groups, in the form of reciprocal cross-cousin marriage (i.e., between FZC and MBC), as in southern India (Trautmann 1981). If a man marries his cross-cousin (e.g., FZC), then his cross-uncle is simultaneously his father-in-law, and in fact the terminological equation of these two kin types is widespread in Dravidian terminologies.

This example shows how choices made in one part of the system—here parent/uncle/aunt terms—predict what choices will be made elsewhere. The development of implicational statements—cross-linguistically robust logical relations between one linguistic feature and another—is a central concern of linguistic typology, and one can revisit typologies of kinship in these terms, testing out formulations such as 'if a language distinguishes parallel and cross-cousins, it will distinguish parallel and cross-uncles.'

Characterizing possible kinship systems by suites of implicational statements rather than simply by idealized types has the advantage that, although in 'type cases' large numbers of features co-occur, this is not always the case. First, not all kinship systems display full type consistency. For example, there is a natural correlation between a bifurcate merging pattern in the first ascending generation and a parallel versus cross-cousin distinction in ego's generation. But not all languages display this: Pitjantjatjara has a bifurcate merging pattern (F = FB ≠ MB; M = MZ ≠ FZ) but

Table 2
Major typologies of kinship system, based on father/uncle term

			Morgan	Lowie/Kirchhoff	Murdock
F	FB	MB	Classificatory	Generational	Hawaiian
F	FB	MB	Classificatory	Bifurcate merging	Iroquois (including Dravidian, Kariera)
F	FB	MB	Descriptive	Bifurcate collateral	Sudanese
F	FB	MB	Descriptive	Lineal	Eskimo
F	FB	MB	(Not named)	(Not named)	Not attested

Source: adapted from Trautmann 1981, p. 84.

does not distinguish cross-cousins from parallel cousins or siblings (Scheffler 1978). In terms of the typology in Table 2, we would have to say that Pitjantjatjara is 'Iroquois' in the first ascending generation, but 'Hawaiian' in ego's generation. A comprehensive typology thus needs to allow for many parameters to be varied independently, rather than assuming they always co-occur; the sorts of extension rule outlined in Sect. 1, being logically independent, allow this.

Second, the types given in Table 2 are too crude. Murdock (1949), extending his focus into other generations, recognized a further two types, Crow and Omaha, based on cross-generational mergings (e.g., Crow: FZ = FZD = FZDD). The basic types also neglect the affects of relative age of the father's sibling, yet| Gui, for example, treats FyB as a type of F while lumping FeB and MB with FF (Ono 1997). The most detailed explorations of differences between systems supposedly of the same type have concerned bifurcate merging systems. In Kariera systems, for example, spouses in the ascending generation always differ in crossness, whereas in most Dravidian systems they do not (Viveiros de Castro 1998).

All kinship systems employ a limited series of kin terms for a potentially infinite set of kin types. But although the number of attested systems way exceeds the classic fourfold or sixfold typologies of Lowie and Murdock, it is nonetheless striking that most logically possible systems are not found. While extensionist approaches manage to parametrize this patterned diversity by not requiring groups of characteristics to go together, they do not make explicit the implicational relationships between one type of extension and another—here we need to bring in tools developed by linguistic typologists. Only by combining these approaches will we solve the puzzle of how human beings, despite being evolutionarily equipped for diversity in their constructions of social categories, repeatedly come up with similar solutions, due to a tangling of cognitive and social factors that we have yet to unpick.

See also: Color Terms, Linguistics of; Kinship in Anthropology; Numeral Systems; Politeness and Language; Word Meaning: Psychological Aspects

Bibliography

Barnard A, Good A 1984 *Research Practices in the Study of Kinship*. Academic Press, London

Burling R 1970 *Man's Many Voices. Language in its Cultural Context*. Holt, Rinehart and Winston, New York

D'Andrade R 1995 *The Development of Cognitive Anthropology*. Cambridge University Press, Cambridge, UK

Danziger K 1957 The child's understanding of kinship terms: A study in the development of relational concepts. *Journal of Genetic Psychology* **91**: 213–32

Evans N 2000 Kinship verbs. In: Vogel P, Comrie B (eds.) *Explorations in Word Class Typology*. Mouton de Gruyter, Berlin, pp. 103–72

Godelier M, Trautmann T R, Tjon Sie Fat F E (eds.) 1998 *Transformations of Kinship*. Smithsonian Institution Press, Washington, DC

Greenberg J 1980 Universals of kinship terminology: Their nature and the problem of their explanation. In: Maquet J (ed.) *On Linguistic Anthropology: Essays in Honor of Harry Hoijer*. Undena Publications, Malibu, FL, pp. 9–32

Hage P 1997 Unthinkable categories and the fundamental laws of kinship. *American Ethnologist* **24**: 652–67

Kay P 1975 The generative analysis of kinship semantics: A reanalysis of the Seneca data. *Foundations of Language* **13**: 201–14

Kirchhoff P 1932 Verwandtschaftsbezeichnungen und Verwandtenheirat. *Zeitschrift für Ethnologie* **64**: 47–72

Kronenfeld D B, Armstrong J D, Wilmoth S 1985 Exploring the internal structure of linguistic categories: An extensionist semantic view. In: Dougherty J D W (ed.) *Directions in Cognitive Anthropology*. University of Illinois Press, Urbana, IL, pp. 91–110

Kryukov M V 1998 The synchro-diachronic method and the multidirectionality of kinship transformations. In: Godelier M, Trautmann T R, Tjon Sie Fat F E (eds.) *Transformations of Kinship*. Smithsonian Institution Press, Washington, DC, pp. 294–313

Lounsbury F G 1964 The structural analysis of Kinship semantics. In: Hunt H G (ed.) *Proceedings of the Ninth International Congress of Linguistics*. The Hague, Mouton, pp. 1073–93

Lowie R H 1928 A note on relationship terminologies. *American Anthropologist* **30**: 263–7

Luong H V 1986 Language, cognition and ontogenetic development: A reexamination of Piaget's premises. *Ethos* **14**: 7–46

Merlan F 1989 Jawoyn relationship terms: Interactional dimensions of Australian kin classification. *Anthropological Linguistics* **31**: 227–63

Morgan L H 1871 *Systems of Consanguinity and Affinity of the Human Family*. Smithsonian Institution, Washington, DC

Murdock G P 1949 *Social Structure*. McMillan, New York
Nerlove S, Romney A K 1967 Sibling terminology and cross-sex behavior. *American Anthropologist* **74**: 1249–53
Ono H 1997 How do the |Gui categorize people? *Gengo Kenkyu* (*Journal of the Linguistic Society of Japan*) **111**: 42–58
Rumsey A 1981 Kinship and context among the Ngarinyin. *Oceania* **51**: 181–92
Scheffler H 1978 *Australian Kin Classification*. Cambridge University Press, Cambridge, UK
Trautmann T H 1981 *Dravidian Kinship*. Cambridge University Press, Cambridge, UK
Viveiros de Castro E 1998 Dravidian and related kinship systems. In: Godelier M, Trautmann T R, Tjon Sie Fat F E (eds.) *Transformations of Kinship*. Smithsonian Institution Press, Washington, DC, pp. 332–85
Wierzbicka A 1986 Semantics and interpretation of cultures: The meaning of alternating generation devices in Australian languages. *Man* **21**: 34–49
Zeitlyn D 1993 Reconstructing kinship or the pragmatics of kin talk. *Man* **28**: 199–224

N. Evans

Klein, Melanie (1882–1960)

Within psychoanalysis worldwide, Melanie Klein has had more impact than any other psychoanalyst besides Freud. Her wide-ranging contributions not only reshaped psychoanalytic understandings of mental life; they also have fundamental import for social scientific understandings of human action and the creation of meaning. That Klein is indisputably the most important psychoanalyst after Freud also points to the unusual prominence and participation of women in psychoanalysis from almost the beginning (see Chodorow 2000).

Born in Vienna on March 30, 1882, Melanie Reizes Klein moved to Budapest in 1910, where she began to read Freud. She entered analysis and analytic training under Freud's colleague Sandor Ferenczi (on Klein's biography, see Grosskurth 1986, and Segal 1979). She moved to Berlin in 1921, joining the Berlin Psychoanalytic Society and entering analysis with Karl Abraham. Her first formal work as a child analyst began in Berlin. At the invitation of Ernest Jones, Klein visited London in 1925 to lecture on child analysis, and she subsequently moved to England in 1926. Until her death in London on September 22, 1960, she was an active clinician, a leading and innovative child analyst, and a prolific writer who had many followers and colleagues.

Understanding Melanie Klein's life requires an understanding of the history of psychoanalysis. Klein had no direct contact with Freud, but was trained and supported by Ferenczi, Abraham, and Jones, the leaders of the main psychoanalytic centers outside of Vienna and Freud's most important colleagues. These several centers allowed more possibility for independence of thought and action than Vienna itself. While Klein was pioneering the practice and theory of child analysis in Berlin and London in the 1920s, Anna Freud was engaged in a parallel project, based on somewhat different understandings and methods, in Vienna. After Freud developed cancer in 1923, he came to depend almost entirely on his daughter as both professional heiress and personal companion and caretaker. Klein's theory, technique, and stature as a child analyst directly challenged Anna Freud's primacy in this same arena, and Freud was drawn to support his daughter. Thus, even as Klein considered herself the true theoretical heir to Freud, her work from the outset was strongly contested by Anna Freud and her father.

As a result of the Anschluss of March, 1938, in which Hitler, with the approval of many Austrians, seized and incorporated Austria into Nazi Germany, the Freuds moved to London, thereby setting the stage for a confrontation and theoretical and organizational sharpening of differences. Freud died in 1939, but the followers of Anna Freud and those of Melanie Klein in London spent the latter war years in 'Controversial Discussions' (King and Steiner 1991), with an 'Independent' Group holding a middle ground and developing its own theories. Debate posed Klein's espousal of Freud's late instinct theory, with its focus on aggression, her theories of early infant development, her advocacy of the analysis of very young children, and her insistence that the analysis of children (and adults) should begin immediately from interpretation of unconscious fantasy, against Anna Freud's focus on ego defenses and ego functions, her following of the traditional structural and libidinal phase theories, and her advocacy of analyzing from the most surface layer of defense rather than from the deepest layer of unconscious fantasy. The result of the 'Controversial Discussions' was a settling in of a Kleinian Group, led by Klein until her death and still an active force in British psychoanalysis (Spillius 1988 includes writings on theory and technique by Klein's students and later colleagues).

Since the 1940s, no psychoanalytic center has developed without grappling with Klein. Kleinians continue to be central to British psychoanalysis, and Kleinian thinking has shaped psychoanalysis throughout Latin America. Kleinian ideas were initially anathema in the USA, where emigre analysts were closely allied with Anna Freud, but since the 1980s, Kleinian thinking has become widely influential. In Europe too, psychoanalytic theory grapples with, accepts, or challenges Klein.

Kleinian theory is a complex, intertwined whole resting on several foundations (see Klein 1975). This account focuses on Kleinian theory rather than on Klein's founding contributions to child analysis and psychoanalytic clinical technique. As a psychoanalyst, Klein like Freud begins from the discovery and

investigation of the dynamic unconscious, the postulate that conscious thinking only expresses a small fraction of what goes on in the human mind. The beginning datum of Kleinian theory is the psychic reality, or internal reality, of what Klein called 'unconscious fantasy.' The dynamic unconscious consists of repressed or split off fantasies that have a powerful effect in shaping conscious life, feeling, and behavior. Unconscious fantasy operates from the very beginning of life and consists of affect-laden stories or proto-stories about the self, the body, and one's relation to others. These stories are not linguistic or narrative, but are forms of implicit knowing and being that portray in direct, unmediated, affectively charged ways, a mental picture of an inner world of self and others. Klein described how this internal reality is created and its constituent effects on all bodily experience, interpersonal relations, and social and cultural experience (for an extended reading of Klein along these lines, see Chodorow 1999b). It is an 'object-relations theory,' picturing the self, or ego, in relation to others, or objects. Klein's method of play analysis and work with very young children led her to a 'developmental theory' based on changes in prevalent unconscious fantasies. The management of anxiety about 'aggression' in self and others is a central goal of most unconscious fantasy.

Although Freud discussed the ego's relations to its objects and the effects of the object upon the ego—for instance, in his claim that the ego contains a history of object choices—Klein is considered the original object-relations theorist. For Klein, the ego is object-related from birth, all drives are directed towards objects rather than toward gratification and release, and all psychological life revolves around internal pictures of the ego in relation to its objects. Objects are both external—actual people like mother, father, or analyst with whom the self interacts—and internal, parts of the outside world that are unconsciously experienced as within the self. Objects in this 'internal world' may be 'part-objects'—an aspect of another, a body part of another, like the mother's breast or a fantasy of her inside organs—or 'whole objects'—others in all their different aspects. Psychic life especially consists in the building of an internal world through what the ego does psychologically with objects, using such mechanisms as 'splitting,' 'projection,' and 'introjection.' Through these processes, internal and external realities are continuously recreated in relation to one another.

Whereas for Freud, repression of unacceptable thoughts and feelings is the key defense, for Klein, splitting, introjection, and projection are the central defenses. Klein's developmental account revolves around the uses of these defensive processes. In the 'paranoid-schizoid position,' the self and the earliest internal objects—initially the mother's breast, and throughout life, objects that represent the breast—are psychically experienced as either all good or all bad. The ego's goals are to keep bad and good parts of the self and object separate. The ego attempts to preserve its own goodness by keeping good, or loved, objects within, and ejecting and rejecting bad, or hated, objects. 'Projective identification' then leads the ego to feel that the object, now containing bad parts of the self, is persecutory. To control the now persecutory objects, they may be reintrojected, in turn bringing hatred back into the self, which projects badness back into the object. And so on. Split-off good feelings and parts of the self may also be defensively projected into the object, so that the self can feel protected and loved. 'Persecutory anxiety'—the fear of being attacked by bad objects or of annihilation of the self—is the dominant anxiety of the paranoid-schizoid position and leads to the spiral of splitting, projection, and introjection that characterize it. The paranoid-schizoid position is paranoid because of the fear of attack it contains and schizoid because it is based on separating different aspects of the self and the other.

When the child reaches the 'depressive position,' he or she recognizes people as whole objects and painfully realizes that an attack on the bad object is also an attack on the good one, since they are one and the same. Guilt arises, along with the desire to make 'reparation.' Reparation is one root of creativity throughout life as well as of the desire to parent and take care of others. 'Depressive anxiety'—the fear that one will damage the good object—is the characteristic anxiety of the depressive position. Unlike Freud's libidinal stage theory, Klein's theory of positions is not a stage theory in which the achievement of the depressive position involves the giving up of the paranoid-schizoid. All people who have reached the depressive position continue throughout life to move between the two positions.

A complex inner world of self and objects constitutes unconscious fantasy and is the basis of 'transference.' For Klein, transference is not just the putting of early feelings about a parent onto the analyst but a total way of being in the world. Although it is mainly investigated in analysis, transference gives meaning to and animates all experience. It functions as an internal prism through which the world is perceived, becomes subjectively meaningful, and is an object of action. Klein, and especially later Kleinians, also show how forms of thinking themselves transferentially express psychic structure and psychic fantasy, for example, whether a patient can make connections among thoughts, can keep anxiety-laden ideas in mind or must split them off and evacuate them, can keep ideas whole or must destroy or fragment them, can take in ideas from the analyst, can symbolize generatively, and other similar processes.

For Klein, the most anxiety-laden and intolerable unconscious fantasies concern aggression. Klein holds to the later Freudian dual-drive theory, rejected by many analysts, of the life and death instincts. The death instinct constitutes a threat of annihilation or fragmentation from within, creating basic existential

anxiety from the beginning of life. The basic wish of the psyche is to get rid of this threat, leading to the projection that creates the first bad object. Persecutory anxiety thus arises originally as a direct response to the internal threat of destruction from within and thereafter also in response to the aggression that has been through projective identification put into objects. The paranoid-schizoid and depressive positions are both organized largely in response to anxieties generated by these fears of annihilation, aggression, and hatred that emerge from self and object.

A major manifestation of the death instinct is found in envy, which Klein describes as the most toxic emotion. Klein discovered the importance of envy in patients' attempts to destroy and dismantle the help offered by the analyst. Envy originates in the perception that the other possesses and enjoys something good and desirable—originally imaged as the creativity, fullness, and bounty of the maternal breast. The envious impulse is to spoil or destroy that goodness. Envy is felt especially virulently against those who appear not to be envious, those who can enjoy creativity and happiness in others. The envious cannot feel gratitude or serenity, and in fact envy is perhaps the least tolerable of emotions, because its goal is to destroy or spoil not that which is bad, but that which is good. Envy is manifested in attempts to destroy the offerings of parents, of teachers, of analysts, of any whom seem contented and happy. In literature, Iago represents envy in its purest form.

Although she did not explore these implications herself, Klein's work has profound implications for social science (see e.g., Chodorow 1999b). Klein's account of unconscious fantasy and transference challenges the sociological assumption that the material or structural reality of politics, society, and culture come first and create or determine human life. By documenting how people create the subjective meaning of the world from within, through the dynamics of unconscious fantasy, and how perception, action, and thought result from a transferential shaping that has real effects, Kleinian theory demonstrates that action in the external world is not only taken through the conscious, rational assessment of options, nor is it determined through structural or material constraints. Action is as much a result of an internal, psychologically created picture as of society or culture, and meanings are as much created psychodynamically through affectively-charged fantasy as they are culturally presented.

Splitting, projection, and introjection are psychological mechanisms that characterize many social processes. Many of the major political events, crises, and controversies of our time can only be fully understood if the operation of rage, splitting, projection, and introjection is taken into account (see Chodorow 1998). For example, political and cultural demonization of the enemy is psychologically based on projection and introjection, in which all good resides in the nation, the ethnos, the group, or the political party, and all bad resides in those who are not part of this group. Splitting bad from good, and containing all good within the self while putting all bad into the other, underpins all virulent racism, nationalism, ethnic conflict, and attempts at ethnic cleansing or genocide. Splitting may also be used to ward off elements in the self that are threatening because they are too attractive. Thus, misogyny and homophobia express splitting, projection, and denial, in this case of misogynist and homophobic men's feminine identifications or desires for other men—attractions and desires that are too anxiety-laden to be contained within the self. Finally, Klein describes the great envy of women's creative insides and the psychic threat of images of the mother who contains everything good and powerful—babies, milk, the father's penis, and the ability to give and take life. Such images of maternal power and the mother's body help to explain the psychological intensity of abortion politics and perhaps the extreme violence of a few anti-abortion activists (for a brief overview of Kleinian theory in relation to gender, see Chodorow 1999a).

Kleinian theory is therefore a basis both of a theory of psychoanalytic treatment and a description of how the human mind develops and operates. It describes how people create themselves and give the world subjective meaning. It explains motivations for action and perception. It is thus an essential component of all investigations of society, politics, and culture.

See also: Feminist Theory: Psychoanalytic

Bibliography

Chodorow N J 1998 The enemy outside: Thoughts on the psychodynamics of extreme violence with special attention to men and masculinity. *Journal for the Psychoanalysis of Culture and Society* **3**(1): 25–38

Chodorow N J 1999a From subjectivity in general to subjective gender in particular. Introductory essay to Melanie Klein's mourning and its relation to manic-depressive states. In: Bassin D (ed.) *Female Sexuality: Contemporary Engagements*. Jason Aronson, Northvale, NJ

Chodorow N J 1999b *The Power of Feelings*. Yale University Press, New Haven, CT

Chodorow N J 2000 Psychoanalysis and women psychoanalysts. In: de Mijolla-Mellor S (ed.) *Women in the History of Psychoanalysis*. Karnac Books, London

Grosskurth P 1986 *Melanie Klein*: *Her World and Her Work*. Knopf, New York

King P, Steiner R (eds.) 1991 *The Freud–Klein Controversies, 1941–45*. Routledge, London and New York

Klein M 1975 *The Writings of Melanie Klein*. The Hogarth Press, London, Vols. 1–4

Segal H 1979 *Klein*. William Collins Sons and Co., Glasgow, UK

Spillius E B (ed.) 1988 *Melanie Klein Today: Developments in Theory and Practice*. Routledge, London and New York, Vols. 1 and 2

N. J. Chodorow

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Knowledge Activation in Text Comprehension and Problem Solving, Psychology of

In order to successfully comprehend written discourse, readers must have the necessary resources to manage the vast amounts of information that continually become available during reading. Ultimately, this means that at some point in the comprehension process this information must be reduced and organized into a coherent representation, one that reflects the intended meaning of the writer. Two major components of the comprehension process are those involved in the activation of the knowledge necessary for comprehension and those involved in the integration of that information. Although they are not necessarily discrete components, many current theories do treat them as such (e.g., Ericsson and Kintsch 1995, Kintsch 1988a, 1998b, Myers and O'Brien 1998, O'Brien and Myers 1999). The focus of this article will be on the current state of knowledge concerning the activation component.

One of the earliest and still most influential models of the comprehension process was developed by Kintsch and van Dijk (1978). According to this model, text is processed over a series of cycles. On each cycle, a small portion of the text is transformed into a set of propositions that are connected in working memory on the basis of argument overlap. Once connected, a subset of these propositions is held over in working memory in order to provide possible connections for the set of propositions input on the next cycle. As long as each set of input propositions can be connected to those currently in working memory, the text is considered locally coherent and comprehension proceeds smoothly. If, however, there are no readily available connections in working memory, the reader searches the long-term memory portion of the representation in order to find a proposition that provides a connection. Once found, that proposition is returned to active memory and the connection is made.

Many of these assumptions concerning when and if readers search memory, as well as what information is readily available to the reader, set the stage for our current theories and understanding of the knowledge activation component of the comprehension process. For example, many studies (e.g., Dell et al. 1983) have demonstrated that when a reader experiences a local coherence break, a search of long-term memory does indeed take place. Inactive portions of the representation are reactivated, providing connections that facilitate the re-establishment of local coherence. However, we are still grappling with the conditions under which this knowledge activation process occurs, and whether the activation process is active or passive.

For a brief period of time, these issues were best captured by contrasting the minimalist (e.g., McKoon and Ratcliff 1992) and constructionist (e.g., Graesser et al. 1994) views of how a reader gains access to earlier portions of a discourse, as well as information from general world knowledge. In the original paper presenting the minimalist view, McKoon and Ratcliff argued that the amount of information available to the reader was quite limited—primarily the immediately preceding one or two sentences. As long as this 'readily available' information allowed the reader to maintain local coherence, no further search or activation processes were necessary, nor would they occur. In contrast, Graesser et al. argued that the reader was motivated by the 'search after meaning' principle. A critical assumption of this principle was that readers were not satisfied with maintaining local coherence; they also sought to maintain global coherence. Maintaining global coherence involves the activation and integration of relevant information from earlier portions of the episodic memory representation of the text, as well as knowledge-based relations derived from general world knowledge.

Consistent with the constructionist view, there have been several demonstrations that information presented earlier in a text may be available even when a text is locally coherent. For example, Albrecht and O'Brien (1993) presented participants with passages containing a description of a particular characteristic of a protagonist (e.g., 'Mary was a strict vegetarian'). Following several sentences that served to background the characteristic, participants were presented with a target sentence in which the protagonist engaged in an action that was locally coherent but contradicted the earlier-described characteristic (e.g., 'Mary ordered a cheeseburger and fries'). Reading times on this target sentence were inflated when contrasted against appropriate controls. This indicated that, although not necessary for local coherence, the backgrounded characteristic was reactivated and disrupted the comprehension process.

In response to these kinds of findings, McKoon and colleagues proposed the memory-based text processing view (e.g., McKoon and Ratcliff 1995). According to this view, concepts currently in working memory make contact with other concepts processed earlier, as well as related concepts from general world knowledge through a fast-acting, passive resonance process (e.g., Ratcliff 1978). The primary difference between this view and the earlier minimalist view was refined as a result of the definition of 'readily available' to now include easy access to vast amounts of related information without the need to invoke any sort of strategic search processes on the part of the reader. With respect to the constructionist view, this passive activation component is an uncontroversial bottom-up component of the full comprehension process. Inclusion and acknowledgment of this component have muted much of the debate between the two contrasting views and shifted the focus toward a combined attempt to assess the kinds of information that are likely to be readily available to a reader. This

bottom-up view of the activation component is shared by several prominent theories of memory activation during reading (e.g., Ericsson and Kintsch 1995, Goldman and Varma 1995, Kintsch 1988a, 1998b, Myers and O'Brien 1998).

It is easiest to explain the strongest assumptions of the memory-based text processing view within the context of the resonance model proposed by Myers and O'Brien (1998). It is important to note that the model is limited to processes and factors governing only the (re)activation of concepts and propositions during reading. Other models do extend this process to how that information is evaluated and ultimately integrated into the evolving text representation, most notably Kintsch's (1998b) Construction-Integration (C-I) model and Goldman and Varma's (1995) CAPS-CI model. The resonance model is based on the assumption that concepts and propositions derived from the sentence currently being processed, or information residing in working memory as a result of reading earlier portions of text, serve as signals to all of memory. The intensity of those signals may depend on the degree of attention given to the text elements they reflect but the process initiated by the signals is not under the control of the reader; it is both passive and dumb. As in many models of memory (e.g., Ratcliff 1978), Myers and O'Brien assume that this process is one in which concepts and propositions in the discourse representation and in the reader's knowledge base resonate as a function of the degree of match to the input. This match depends on the featural overlap among concepts, and on the argument overlap of propositions. Memory elements that are contacted by the initial signal in turn signal to other memory elements. During this resonance process, activation builds and when the process stabilizes, the most active elements enter working memory.

There are two critical features of the resonance process. First, it is continual; a signal is constantly being sent to all of memory, not just when triggered by a coherence break. What is resonating in long-term memory changes constantly with each change in the content of working memory. More important, the resonance process is dumb. Any information that resonates sufficiently is returned to active memory independent of whether that information will ultimately facilitate or hinder comprehension. Note that Kintsch's (1998b) C-I model has the same assumption; because the construction (or reactivation) phase is both passive and dumb, it is quite possible that the output of the construction phase will be incoherent, and possibly contradictory.

This bottom-up view of knowledge activation eliminates the distinction between local and global coherence. Because most information related to any current input into working memory automatically resonates and is reactivated in response to that input, the reader simply needs to monitor the coherence of the information that is in working memory. By definition this will involve not only recently read information, but also related information from earlier portions of the text, as well as related information from general world knowledge.

There is mounting evidence that this knowledge activation is indeed passive and dumb. Several studies have shown that inactive portions of memory are activated and become part of the integration process, even when that information contradicts and/or disrupts comprehension (e.g., O'Brien et al. 1998). Even stronger evidence exists in studies in which information that is related—in terms of featural overlap—but irrelevant becomes active and disrupts comprehension. For example, Cook et al. (1998) found that if a narrative described a secondary character as a vegetarian (e.g., Joan), a later reference to the primary character (e.g., Mary) eating a hamburger resulted in the reactivation of the vegetarian characteristics assigned to the secondary character. It is difficult to explain these types of findings in terms of a reader actively searching memory for relevant information. But these types of findings become obvious predictions for comprehension models that include a passive activation component.

In many earlier models of knowledge activation there was an order in which the episodic memory trace and general world knowledge were activated. Kintsch and van Dijk (1978) assumed that the episodic memory trace was activated first, whereas van Dijk and Kintsch (1983) assumed that general world knowledge was activated first. Current models (e.g., Kintsch's 1998b C-I model) now routinely assume that both sources of knowledge are activated in parallel—an assumption that must be made if knowledge activation is passive. In earlier work examining knowledge activation, researchers often examined activation of the episodic memory trace while holding general world knowledge constant (e.g., O'Brien 1987) or vice versa, holding the episodic memory trace constant while varying general world knowledge (e.g., Spilich et al. 1979). Because most models now assume that the two sources are activated in parallel, pressure is mounting to examine the interaction of these two knowledge sources. The most promising work in this area is that of Kintsch and his colleagues' use of Latent Semantic Analysis (LSA, see the special issues of *Discourse Processes*, 1998, **25** devoted to LSA). LSA is a method for assessing the impact of general world knowledge on the comprehension process. It involves the building of large, but simple, associative networks of word meanings. This is done by measuring co-occurrences of words in large text corpora. The mechanism for activation is consistent with Kintsch's C-I model (i.e., passive). LSA has proven quite successful in accounting for several cognitive phenomena including word recognition, sentence-word priming, and text comprehension. It is very exciting and promising that a simple associative network can go so far in accounting for the impact of general knowledge activation on the

comprehension process. However, although computationally it provides a good fit to data, its value as a representational theory may be limited (see Perfetti 1998). Nevertheless, it provides an excellent starting point for considering the interaction of the two sources of knowledge activation.

The activation and instantiation of inferences, either necessary or elaborative, plays a crucial role in the comprehension process. Clearly, before any sort of inferential information can be incorporated into an evolving representation of a text, it must be activated (see Kintsch 1998b, for an excellent discussion of the distinction between activation and instantiation, p. 189). It must also be validated against general world knowledge (Singer and Halldorson 1996). By definition this means knowledge-based relations not explicitly stated in the text but already present in general world knowledge must be activated. Two critical questions are. What information is activated? And how does activation occur? Within the memory-based text processing view, any concept that is semantically related to what is currently being read should be passively and automatically activated. This makes a vast amount of information quickly and readily available (e.g., Ericsson and Kintsch 1995), some of it useful to the reader, some of it not. Currently this account of the initial knowledge activation phase can account for the majority of results showing activation of information that ultimately plays a role in inference generation (e.g., Murray et al. 1993). However, there is also evidence that some inferences require a slower-acting conscious process (e.g., Calvo et al. 1999). The memory-based text processing view cannot account for these sorts of active generation processes. The question of active vs. passive inferences remains an open and important issue for future research.

The issues and principles raised in knowledge activation in text comprehension also apply to issues in problem solving. Within Ericsson and Kintsch's (1995) model of long-term working memory, activation of the knowledge necessary for problem solving is initially guided by the textual input. Automatic activation processes enrich explicitly stated information. This activation links information in working memory to large, pre-existing structures in long-term memory. The result is a large expansion of the information that is readily available to be used in problem solving. This is an important first step in understanding problem solving processes. However, beyond the initial activation stage, the issues in problem solving become enormously complex and messy. Problem solving is typically assumed to involve complex, conscious, resource demanding operations. Other than in situations involving well-practiced experts, those operations that can, or will, be utilized are difficult to constrain, and as a consequence are difficult to predict (see Kintsch 1998b).

The primary goal of models consistent with the memory-based text processing view is to explain as much of the overall comprehension process as possible without invoking strategic processes on the part of the reader. These models are incomplete because they typically lack a full specification of the mechanisms for evaluating and integrating the large amount of information they assume to be readily available to the reader. Some of those mechanisms are well specified in the top-down components of current constructionist models. Any complete model of the comprehension process will have to include many of the components identified in the memory-based text processing models as well as many of the top-down components identified in constructionist models. The current state of the field of text comprehension has clearly identified these two ends of the comprehension continuum. Future work that begins to identify the gray area between these two end points is likely to prove the most informative to a complete theory of the comprehension process.

See also: Knowledge (Explicit and Implicit): Philosophical Aspects; Knowledge Representation; Knowledge Spaces; Learning from Text; Memory for Text; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning; Text Comprehension: Models in Psychology

Bibliography

Albrecht J E, O'Brien E J 1993 Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**: 1061–70

Calvo M G, Castillo M D, Estevez A 1999 On-line predictive inferences in reading: Processing time during versus after the priming context. *Memory & Cognition* **27**: 834–43

Cook A E, Halleran J G, O'Brien E J 1998 What is readily available? A memory-based view of text processing. *Discourse Processes* **26**: 109–29

Dell G S, McKoon G, Ratcliff R 1983 The activation of antecedent information during the processing of anaphoric reference in reading. *Journal of Verbal Learning and Verbal Behavior* **22**: 121–32

Ericsson K A, Kintsch W 1995 Long-term working memory. *Psychological Review* **102**: 211–45

Goldman S R, Varma S 1995 CAPping the construction-integration model of discourse comprehension. In: Weaver III C A, Mannes S, Fletcher C R (eds.) *Discourse Comprehension: Essays in Honor of Walter Kintsch*. L Erlbaum, Hillsdale, NJ

Graesser A C, Singer M, Trabasso T 1994 Constructing inferences during narrative text comprehension. *Psychological Review* **101**: 371–95

Kintsch W 1988a The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* **95**: 163–82

Kintsch W 1998b *Comprehension: A Paradigm for Cognition*. Cambridge University Press, New York

Kintsch W, van Dijk T A 1978 Toward a model of text comprehension and production. *Psychological Review* **85**: 363–94

McKoon G, Ratcliff R 1992 Inference during reading. *Psychological Review* **99**: 440–66
McKoon G, Ratcliff R 1995 The minimalist hypothesis: Directions for research. In: Weaver III C A, Mannes S, Fletcher C R (eds.) *Discourse Comprehension: Essays in Honor of Walter Kintsch*. L Erlbaum, Hillsdale, NJ
Murray J D, Klin C M, Myers J L 1993 Forward inferences in narrative text. *Journal of Memory and Language* **32**: 464–73
Myers J L, O'Brien E J 1998 Accessing the discourse representation during reading. *Discourse Processes* **26**: 131–57
O'Brien E J 1987 Antecedent search processes and the structure of text. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **13**: 278–90
O'Brien E J, Myers J L 1999 Text comprehension: A view from the bottom up. In: Goldman S R, Graesser A C, van den Broek P (eds.) *Narrative Comprehension, Causality, and Coherence: Essays in Honor of Tom Trabasso*. L Erlbaum, Mahwah, NJ
O'Brien E J, Rizzella M L, Albrecht J E, Halleran J G 1998 Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**: 1200–10
Perfetti C A 1998 The limits of co-occurrence: Tools and theories in language research. *Discourse Processes* **25**: 363–77
Ratcliff R 1978 A theory of memory retrieval. *Psychological Review* **85**: 59–108
Singer M, Halldorson M 1996 Constructing and validating motive bridging inferences. *Cognitive Psychology* **30**: 1–38
Spilich G J, Vesonder G T, Chiesi H L, Voss J F 1979 Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior* **18**: 275–90
van Dijk T A, Kintsch W 1983 *Strategies of discourse comprehension*. Academic Press, New York

E. J. O'Brien, A. E. Cook, and K. A. Derepentigny

Knowledge, Anthropology of

1. *First Approaches: The Late Nineteenth and Early Twentieth Centuries*

Most nineteenth century formulations of the problem of knowledge focused on what entire social groups shared. These formulations also implicitly or explicitly assumed an evolutionary schema of social development from 'elementary' to more complex societies, a notion as common to Charles Darwin as it was to Karl Marx. Thus William Robertson Smith (1972), a Biblical scholar whose work profoundly influenced later sociologists and anthropologists, including Émile Durkheim and E. E. Evans-Pritchard, saw a close relationship between the stage of development of a social group and the nature of its intellectual, religious, and moral life. For example, Smith argued that to understand the message of the prophet Ezekiel, the social organization of his time had to be delineated, together with the cosmology and modes of thought contemporary with it. If Ezekiel's prophetic message was found to be different from that of Samuel, Smith argued, it is not because of any contradiction in the Bible but because Samuel addressed himself to groups and situations at different stages of human development.

This dual sense of evolution in history and in the complexity of social forms was basic to Émile Durkheim's *Elementary Forms of the Religious Life* (1915) and his *Primitive Classification* (1963), co-authored with his nephew, Marcel Mauss. Durkheim's goal was to study the most 'elementary' societal forms in order to discern the basic social elements out of which ideas of time, space, causality, category, and person are constructed. Using the example of Australian aborigines, which he considered the best documented 'elementary' society, Durkheim argued that religious thought, like all forms of thought, derived from collective social action.

Durkheim sought to synthesize the two major prior approaches to categories of thought, the 'apriorist' and the 'empiricist'. The apriorist approach, exemplified by Immanuel Kant, viewed categories of thought as immanent in the human mind. The empiricist approach, exemplified by John Locke, saw these categories of thought as derived from experience, with individuals independently constructing similar perspectives. Mediating between these two approaches, Durkheim argued that categories of understanding were neither innate nor derived from individual experiences. Instead, they were the product of a 'collective consciousness' created and sustained through shared ritual action. This social basis of knowledge and categorization applied as much to the classification of objects and the natural world as it did to religious thought, at least in 'elementary' societies.

For Durkheim, the notion of the 'collective consciousness' ('conscience collective') joined assumptions of the fundamental faculties of the human mind with the creation of collective representations. Durkheim's argument was elegant but in synthesizing the main currents of European social philosophy, he was unable to demonstrate the primacy of the social in 'causing' forms of classification. At the least, his argument suggested an analogy between the social order and forms of knowledge, and his psychological assumptions concerning collective representations were quickly superceded. Nonetheless, his work profoundly influenced successive generations of anthropologists interested in the ethnographic representation of cosmologies and religious beliefs and their relation to the social order.

2. *The Ethnography of Knowledge*

An example of how the work of Robertson Smith and Durkheim influenced ethnographic writing in the first

half of this century is E. E. Evans-Pritchard's study of the Nuer of the southern Sudan (1940). Evans-Pritchard cites neither of these predecessors in his text, although their influence is evident in the structure of his argument. He depicted the Nuer as a cattle-herding people with a simple material life and minimal social and political differentiation, highly dependent on their local environment and seasonal migrations. Evans-Pritchard described these in detail, indicating how ecological conditions shaped Nuer social life and cosmology, or ideas of religion, causality, and the order of things. He showed how the structure of Nuer lineages, age-grades, and political units corresponded to their ideas of time, space, and political relationships.

Like Durkheim, Evans-Pritchard allowed readers to infer that Nuer categories and classifications derived from the structure of their social and political life. As with Durkheim, however, all Evans-Pritchard demonstrated was common schemas or homologies of form among a 'traditional' social structure, ideas of time and space, and cosmology.

The anthropological tradition represented by Evans-Pritchard remained dominant through the mid-twentieth century, inspiring other work on cosmology and social structure in other African societies, including John Middleton on the Lugbara and Godfrey Lienhardt on the Dinka.

The roots of Durkheim's influence continued well beyond mid-century and could be seen in the work of anthropologists such as Edmund Leach and Mary Douglas (1966). Such studies suggest the extent to which people go to maintain or distinguish conceptual kinds and boundaries among animals, plants, genders, foods, peoples, and types of activity. Claude Lévi-Strauss (1966) also focused on the forms of classification implicit in both the kinship structures and in the myths of 'elementary' societies, exploring similarities across cultural traditions that suggested common 'structures' in myths, for example, akin to those studied by structural linguists. Lévi-Strauss's work encouraged a focus on how analogies are made between different cultural domains.

Some anthropologists, such as Maurice Godelier (1986), sought to join Durkheimian concerns with categories and classification to a Marxist-inspired linking of ideas of knowledge to modes of production. As with Durkheim's work, however, the emphasis in these studies remained primarily in the classificatory schema implicit in systems of knowledge and in analogies between symbolic representations and material/economic life rather than in how people created, imposed, and altered, ways of knowing.

3. *Psychology and Anthropology*

Until the 1920s, the lines between psychology and anthropology significantly overlapped. It was often psychologists more than anthropologists who engaged in comparative work across societies and explored the extent to which humans thought alike. Some anthropologists argued that human intellectual and cognitive processes were essentially similar and varied only on the level of evolution of their culture. Others, such as the French philosopher, Lucien Lévy-Bruhl, argued that there was a great divide between the pre-rational and 'primitive' forms of thought and the 'rational' thought of modern societies. Lévy-Bruhl's argument had a strong resonance in French colonial lobbies, and was even offered to explain the presumed inability of the Muslim societies of North Africa to abandon social forms and beliefs supposedly inadaptive to the colonial, modern world.

After the 1920s, however, there was an increasing separation between psychological and anthropological approaches to cognition and knowledge. Psychologists more than anthropologists explored issues of cognition, the capacity and processes by which individuals learn, classify, and make associations and analogies. Psychologists thus focused on how the brain is 'hard-wired' for certain thought processes, including the ability to learn languages. Anthropologists dealt more with culturally learned forms of knowledge and classification. Only a few anthropologists, such as Gregory Bateson (1972), continued to envision a common ground between the two disciplines.

As Dan Sperber (1985) noted, anthropological work always had two dimensions. Some anthropologists have been more interested in the interpretation and analysis of particular cultural and social groups. Others have been more concerned with the cultural abilities and dispositions of humankind in general. Studies in cognitive anthropology, as the latter focus became known, underscored this distinction. These were the anthropologists most inclined to explore the link with psychology. The 1960s again brought scholars from the two disciplines into close collaboration. In one key study, for example, color perceptions, a domain that anthropologists previously assumed to be culturally transmitted, became recognized as having a more limited range of variation than was initially supposed, suggesting underlying cognitive principles (Berlin and Kay 1969).

The focus of the Durkheimian tradition of the anthropology of knowledge was always in the 'outer' frame of knowledge—how classificatory schemes for understanding society and the world related to the social order. The work of the cognitive anthropologists studied the order internal to particular schemes of knowledge—such as botany, color, disease types, firewood, and kinship and relationship systems. Michael Cole (1996) called this renewed collaboration between anthropologists and psychologists the 'second psychology,' in which the two disciplines explored the frontier between culture and 'mind' in understanding cognitive development, the acquisition of literacy and

numeracy, and language acquisition. In this vein, Holland and Quinn (1987), for example, use the term 'schema' to refer to standard learned patterns of activity, such as 'going to the post office,' by which members of a society learn cultural patterns and habits to accomplish routine tasks, freeing individuals to focus instead on less predictable situations.

4. *Intellectuals*

Another component of the anthropology of knowledge is the role of intellectuals in society and the social production of knowledge. Until recent decades, anthropologists were primarily concerned with those bodies of cultural knowledge shared by all members of a society. Anthropologists such as Radin (1957) in American anthropology and Griaule (1965) in the French tradition portrayed intellectuals as possessing an especially intense awareness of the sacred center of their society and an ability to reflect and explain its categories of knowledge. A classic essay by Victor Turner (1967) on the knowledge and character of one of his principal informants on Ndembu symbolism, comparing him to an 'Oxford don' also broke from an earlier tradition in which individual producers of knowledge were subsumed in larger collectivities.

An influential approach to defining intellectuals in relation to theories of modernization and modernity was Edward Shils (1972). For him, intellectuals were those imputed with an 'innate need' to understand and articulate ideas and concepts that are more general and profound than those faced in practical, everyday concerns. For Shils, all intellectual work originally arose from religious occupations, but religious orientations in modern times have attracted a diminishing share of the creative capacities of an elite oriented toward more 'modern' concerns. Shils's approach assumed that not all knowledge was equally shared, and it also encouraged anthropologists and other scholars to think of varied types of intellectuals and their relation to the different political and social contexts in which knowledge is produced and valued. Whether they were called intellectuals or entrepreneurs, anthropologists shifted their focus away from discerning classificatory categories and cosmologies and toward the economic and political contexts in which knowledge is produced and shaped.

In this shift to understanding knowledge and its social distribution as process and as historically rooted, the strategies used for comprehending 'elementary' societies—to invoke Émile Durkheim's term—were adapted to depict ideas of knowledge in advanced, 'complex' ones. Thus Pierre Bourdieu compared the cognitive style implicitly learned at the Sorbonne to that transmitted by Bororo elders in the Brazilian Amazon. For Bourdieu, the verbal maneuvers learned by students preparing for the *leçon* at the Sorbonne furnished a model for the 'right' mode of intellectual activity in the French context as much as the dualistic spatial layout of Bororo villages embodied Bororo cosmology as interpreted by dominant elders. The dualistic method of the *leçon*, which establishes a traditional 'two views' on any subject, is subsequently applied to discussion of a wide range of intellectual problems, to the exclusion of alternative, less culturally valued approaches.

5. *Knowledge and Power*

Bourdieu's work on education, social class, and knowledge in the 1960s linked together several long-standing themes in the anthropology of knowledge, but also indicated a shift from earlier concerns with categorization to the practice of creating and distributing knowledge. Bourdieu, like Bloch (1985), Fardon (1985), Godelier (1986), and others, related knowledge and social classifications to the distribution and control of resources. In the specific case of education in France, Bourdieu depicted how the style and forms of knowledge of French higher education converted social classifications into school ones, relating mastery of the 'symbolic classifications' and habits of thought of formal schooling to both class and social capital.

The theme of the historical problem of change in dominant ideologies of higher education and how these ideologies were related to the conflicting interests of class and social groups in complex societies was raised at the beginning of the century. Durkheim's (1977) neglected lectures on the evolution of higher education in France is a case in point. It was only by the 1970s, however, that such ideas rose to the forefront of anthropological thought. Studies such as Bourdieu's on higher education in France, and others on themes such as the internal logic and changing social context of Islamic ideas of knowledge (Eickelman 1985) and the process by which dominant ideas of religion and tradition shifted in Indonesia (Hefner 1985) joined the long-standing anthropological concern with tacit knowledge in religious experience with how social actors derive meaning from public symbolism and continually shape and renew bodies of knowledge in differing social and historical circumstances.

Fredrik Barth's (1987) account of local variations in the cultural traditions of the Ok of highland New Guinea illustrates both the shift to understanding how people produce and use knowledge, and the erosion of the earlier distinction elementary and complex societies. The Ok use their ritual occasions to create and alter systems of knowledge as much as seminars and innovative lectures accomplish the same objective in contemporary European societies. Departing from earlier accounts such as Evans-Pritchard on the Nuer, Barth emphasized the processes of thought and innovation at work within a system of knowledge, producing multiple variations and developments.

Approaches such as Barth's also had profound implications for the idea of culture. Rather than viewing culture implicitly as internally homogeneous, shared, and sharply bounded, culture instead is conceived as a perpetually interactive social organization of communication in which individuals reproduce and alter representations and symbols. Knowledge is thus produced in the enactment of the 'regulated improvisations' (Bourdieu 1977) by which people express thoughts, organize, and interpret social action. Social practice in specific historical contexts thus produces 'objective' structures (economic and power relations) that in turn produce cognitive and affective dispositions of the social world common to members of given social groups or classes.

Current trends in the anthropology of knowledge thus encompass two related concerns. One builds on new collaborations between anthropology and psychology, exploring new ways of relating public and institutionalized cultural expression to cognition and mental representations (Shore 1996). The other focuses on the social production of knowledge and its relation to power, value, communications, and control. The earlier concern with the values, cosmology, and ways of knowing that are, common to all people in a given society, has given way to how ideas and innovation are communicated in more bounded and explicit fields of knowledge. Such studies include the diffusion of knowledge and ideas in Silicon Valley, the genome project and in biogenetic engineering, and the implications of new communications media on ideas of religious and political authority. The processes by which change is embodied in implicit shared ideas of gender and sexuality, and authority and responsibility, also forms a component of current concerns in the anthropology of knowledge.

See also: Anthropology; Anthropology and History; Collective Beliefs: Sociological Explanation; Collective Memory, Anthropology of; Cultural Critique: Anthropological; Cultural Diversity, Human Development, and Education; Cultural Relativism, Anthropology of; Culture: Contemporary Views; Ethnography; Ethnology; Historiography and Historical Thought: Current Trends; Interpretation in Anthropology; Knowledge (Explicit and Implicit): Philosophical Aspects; Knowledge, Sociology of; Linguistic Anthropology; Modernity: Anthropological Aspects; Phenomenology: Philosophical Aspects; Psychological Anthropology; Self: Philosophical Aspects; Tradition, Anthropology of

Bibliography

Barth F 1987 *Cosmologies in the Making: A Generative Approach to Cultural Variation in Inner New Guinea*. Cambridge University Press, Cambridge, UK

Bateson G 1972 *Steps to an Ecology of Mind*. Ballantine Books, New York

Berlin B, Kay P 1969 *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA

Bloch M 1985 From cognition to ideology. In: Fardon R (ed.) *Power and Knowledge: Anthropological and Sociological Approaches*. Scottish Academic Press, Edinburgh, UK

Bourdieu P 1967 Systems of education and systems of thought. *International Social Science Journal* **19**: 338–58

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, New York

Cole M 1996 *Cultural Psychology: A Once and Future Discipline*. Harvard University Press, Cambridge, UK

Douglas M 1966 *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. Frederick A. Praeger, New York

Durkheim É 1915 *The Elementary Forms of the Religious Life*. George Allen and Unwin Ltd., London

Durkheim É 1977 *The Evolution of Educational Thought*. Routledge and Kegan Paul, London

Durkheim É, Mauss M 1963 [orig. 1903] *Primitive Classification*. University of Chicago Press, Chicago

Eickelman D F 1985 *Knowledge and Power in Morocco: The Education of a Twentieth Century Notable*. Princeton University Press, Princeton, NJ

Evans-Pritchard E E 1940 *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People*. Clarendon Press, Oxford, UK

Fardon R (ed.) 1985 *Power and Knowledge: Anthropological and Sociological Approaches*. Scottish Academic Press, Edinburgh, UK

Godelier M 1986 *The Mental and the Material*. Verso, London

Griaule M 1965 [orig. 1948] *Conversations with Ogotêmmeli*. Oxford University Press for the International African Institute, London

Hefner R 1985 *Hindu Javanese: Tengger Tradition and Islam*. Princeton University Press, Princeton, NJ

Holland D, Quinn N 1987 *Cultural Models in Language and Thought*. Cambridge University Press, Cambridge, UK

Lévi-Strauss C 1966 *The Savage Mind*. University of Chicago Press, Chicago

Radin P 1957 [orig. 1927] *Primitive Man as a Philosopher*. Dover Books, New York

Robertson Smith W 1972 [orig. 1889] *Lectures on the Religion of the Semites*. Schocken Books, New York

Shils E 1972 *The Intellectuals and the Powers*. University of Chicago Press, Chicago

Shore B 1996 *Culture in Mind: Cognition, Culture, and the Problem of Meaning*. Oxford University Press, New York

Sperber D 1985 *On Anthropological Knowledge*. Cambridge University Press, Cambridge, UK

Turner V 1967 [orig. 1957] Muchona the Hornet, Interpreter of Religion. In: Turner V. *The Forest of Symbols: Aspects of Ndembu Ritual*. Cornell University Press, Ithaca, NY

D. F. Eickelman

Knowledge, Education, and Skills, Geography of

Knowledge, professional skills, inventiveness, competence, creativity, technological capabilities, and educational achievement have never been evenly distributed in populated space. Apart from everyday

routine information and common sense, most forms of knowledge and skills display large spatial disparities. Since knowledge is regarded as a key factor for competitiveness, development and modernization, a geographical perspective is required to analyze and to explain the spatial structures, processes, and disparities involved in the creation, adoption, distribution, and application of knowledge and skills.

1. Spatial Concentration of Knowledge and Skills

1.1 Diffusion of Knowledge and Information

New knowledge (whether in the form of technologies or social values) is created in particular places and contexts and through interaction within space. The speed at which new knowledge and information diffuses over space depends on the type of knowledge, the institution within which the new knowledge is produced, the interest of the producer (inventor) to share his or her knowledge, the previous knowledge necessary to understand the contents of new information, the availability of technology necessary for the production and application of knowledge, and the inclination to accept the knowledge.

The first type of knowledge and information which is most quickly distributed in space is the kind of 'public news' which is easily understood by almost everyone, which does not require previous knowledge or expensive technology, and whose free distribution is in the provider's best interests. This type of information and knowledge is highly mobile, can be spread around the globe in seconds and is easily available by telecommunication.

The second type of knowledge and information is more complex and cannot readily be used even by those who have access to it if these constituencies lack the previous knowledge, experience or training necessary to recognize, understand and evaluate the contents of the information. This type of knowledge can only be transferred from one person to another if the receiver's training, skills or capability are broadly equivalent to that of the sender. Previous knowledge necessary to understand an information may consist of knowledge of a foreign language or of a specialized scientific discipline. The third type of knowledge and information is kept secret by its producer in order to gain an economic, political, scientific or military advantage.

These three forms of knowledge and information vary in the degree of their spatial concentration and in the speed of their diffusion. Only the first type of information, which is the least important for economic competitiveness and innovation, is (theoretically) ubiquitously available, assuming all places have the technical equipment and all people an equivalent inclination to receive the message. The second type displays a much higher spatial concentration in a few centers or areas, and often circulates only between the upper levels of the urban hierarchy (Hägerstrand 1966, Törnqvist 1968, 1970, Pred 1973). The third type has the highest degree of spatial concentration. The term 'spatial concentration' does not imply a tight, everlasting linkage to a certain place but rather a conglomeration of functions that may change its location over time.

1.2 Center and Periphery as Spatial Representation of Power and Knowledge

Terms such as power, authority, dominance, and control are only applicable in the framework of asymmetric social relations. Asymmetric social relations manifest themselves spatially, both in the horizontal and in the vertical dimension. Due to the importance of spatiality in the construction and representation of difference, social structures and social relations are often described spatially by terms such as center, core, periphery, marginality, outsider, top, bottom, segregation or distance. In spatial representation knowledge and power regularly take an elevated or central position. Dependent positions with low status are at the margin.

A center is a point of reference, orientation, and authority. It is the nodal point (field) of interaction and communication whence the elements of a social or spatial system are ruled, coordinated, and controlled (Strassoldo 1980). It collects and distributes resources, sets the rules, norms, and standards for the members of the system; it controls, legitimates, and accumulates knowledge (Latour 1987); it provides a conspicuous public platform from which ideas emanate; it constructs realities and creates memories and identities. The center is the place where the highest levels of authority, power, decision making and knowledge are located. Most of the work places located at the periphery are lowly skilled and are the site of lowly paid routine activities coordinated and controlled by external decision makers. The antagonism between core and periphery is a spatial representation of an uneven distribution of power and knowledge.

Why do power and knowledge so frequently build coalitions and why do they seek proximity to each other? Both instrumental power (the capacity to make others do our will and the ability to control, regulate and dominate), and association power, i.e. the power to do things by acting in concert or using institutional mediation (Agnew 1999), have to rely on experts, scientists, academics and 'think tanks' to achieve and sustain their position. They also need representatives of spiritual knowledge (priests, intellectuals, artists, ideologues) who share their norms and values and legitimate their power. It is in the interest of power to bind knowledge into networks of assent (Agnew 1999). On the other hand, both scientific knowledge (ex-

pertise) and spiritual knowledge seek proximity to (political) power for various reasons. First, centrality is the spatial manifestation of authority, importance, and prestige. Authority fascinates most experts and intellectuals, they want to be concerned with the essential in existence, and strive for influence and recognition. Second, network-centrality determines access to valued resources and information unavailable to those at the margin. Third, experts, scientists, and artists in many cases have to prove their usefulness to power in order to secure their privileges and the resources they need for their work. Fourth, proximity to power may also increase the opportunities to influence important decision-makers. Close spatial proximity between power and knowledge seems to be indispensable. Even intellectuals (dissidents) opposing those in power have to operate in centers in order to get access to communication networks and to obtain response in the media. Decision makers, professionals, experts, artists, and intellectuals striving for prestige, influence, and success, sooner or later find their place of work in one of the centers or will create for themselves a new center in their chosen field.

Work places of important decision-makers, highly skilled specialists, and high-technology industries show a strong tendency toward spatial concentration, whereas lowly skilled routine activities in production and administration show a trend towards dispersion and decentralization. The main characteristic of peripheries is that they constantly loose the largest part of their highly educated to the centers.

Knowledge is rooted in people, but it is also embodied in technologies, products, services, management performance and institutions of schooling, education and training. The broad range of research topics is paralleled by the large variety of theoretical approaches that the geography of knowledge, education, and skills has to apply. The following chapter presents one example, namely theoretical approaches explaining the spatial concentration of work places for top-level decision-makers, intellectuals, and highly skilled specialists. Work places present important settings and contexts for a great variety of actions.

2. *Theoretical Approaches to Explain the Spatial Concentration of Knowledge*

2.1 *Organization Theory*

For an organization the production and acquisition of new knowledge and the ability to learn, adapt and reorganize itself are the best means to cope with uncertainty. The acquisition of new knowledge and new information is no guarantee for successful decisions and actions but it increases, at least for a certain time, transparency, predictability, and efficiency while sustaining flexibility and competitiveness and enhancing survival chances in an uncertain environment. According to this approach, it is not the 'fittest' that survive but those who constantly acquire new knowledge, skills and information and are therefore able to adapt quickly to new situations and to cope with new challenges. 'It is evolution through learning that clearly dominates and accounts for the major variations in human social behavior' (McClintock 1988, p. 60). There is no doubt that incompetence, ignorance, and lack of experience are the main factors leading to the collapse of social systems and the decline of centers. The question of how much incompetence, ignorance or lack of skills a given social system can afford depends primarily on the intensity of competition, on the uncertainty of the environment, and on the available resources of the system.

As centers and peripheries are socially constructed and as space is a product of relations and interactions, both the status and the locations of centers and peripheries change over time. In spatially *mobile* social systems, such as nomadic tribes or armies in war, centers constantly change their location in space. Most social systems, however, are spatially rooted and change the location of their centers and peripheries only gradually. Large and complex social systems show a hierarchy of centers. The definition of a place as center or periphery and its ranking in the hierarchy vary with the spatial scale; a center at the micro-scale may be defined as periphery on a macro-scale.

Since an organization can only compensate for a certain amount of incompetence for any extended period of time, it acts in its own interest when it fills the key positions of information processing, decision making, planning, coordination, and control with highly skilled, educated, and experienced persons. In particular, those positions and subsystems that are constantly confronted with uncertainty and whose decisions have long-lasting consequences for the entire system require highly developed skills, knowledge, and experience. In social systems, knowledge, skills, and experience are a value-adding source and have the same function as redundancy in technical systems. They reduce uncertainty and enhance survival chances.

In a dynamic and competitive society, the acquisition of knowledge and skills is an ongoing process that is never finished. Much of our knowledge is transitory and looses its value after a certain time. In a competitive society, it is not knowledge itself that counts, but the possession of prior, specialized, unique, better or rare knowledge that justifies the role of experts and advisers, that makes knowledge attractive to those in power. Advisers, counselors, experts, priests, and intellectuals derive their power from their claim to know *better* or *earlier* than the majority of people, from their ability to reduce the complexity of knowledge and to make knowledge useful. In a dynamic and competitive society, the skills and knowledge needed for the key functions of a system striving

for success (survival) will always be scarce and expensive commodities.

The first crucial question is where to locate scarce knowledge and skills within a social system's structure of decision making. In bureaucratic systems and in systems confronted with low degrees of uncertainty, decision-making, problem-solving, research, development, and planning will shift to the upper levels of the system's hierarchy and consequently the lower levels will be characterized by routine activities and work places for those with lower skills. In systems confronted with a dynamic and complex environment and dealing with constantly changing, unpredictable, one-time transactions, decentralization of competence and authority within the system is more effective. For a more detailed discussion of various types of organizations, environments and means of coordination see Mintzberg (1979).

The second question is where to locate scarce knowledge and skills spatially. Routine activities for which there are plans, rules, and regulations call for lower skills and less face-to-face contacts with other organizations. They are predominantly regulated and coordinated by indirect contacts such as letters or telecommunication. Therefore, they can be situated in a great number of smaller cities or even rural areas, as long as the traditional location factors such as transportation and wage costs make it feasible. The harder it is for decisions to be governed by guidelines, plans, regulations or so-called hard information and the greater the uncertainty about the future, the more necessary it is to have frequent and spontaneous face-to-face contacts with knowledgeable and well-informed decision-makers and highly skilled specialists of other organizations. The needs of the highest management levels for face-to-face contacts and valuable networks can only be fulfilled in a few large centers offering a large diversity and density of knowledgeable specialists and prominent decision-makers.

Trustful relations and frequent face-to-face contacts to leading decision makers in government, research, finance, consulting, media or industry offer access to early knowledge about important tendencies and innovations and provide the top positions of an organization with crucial information, thus facilitating their quick response and adaptation to new situations and developments.

2.2 Symbol Orientated Approach

The spatial distribution of work places for the highly and lowly skilled, for the powerful and powerless cannot be fully explained by functional arguments. Distributions are also influenced by the symbolic meaning of places. A place can be a symbol for prestige, reliability, credit-worthiness, and power. Another address only short distance away may suggest danger, poverty, failure or crime. The terms center (core) and periphery have a strong symbolic meaning. A center is associated with social attributes such as power, authority, dominance, prestige, control, and influence. The term periphery symbolizes dependency, marginality, backwardness, lower skills, lower prestige, weakness and subordination.

It is not only the functional and spatial division of labor that creates centers and peripheries. Anthropology, social psychology and the history of religions also offer a large number of ritualistic constructions of centers. In many religions, the center is a sacred place where the gods revealed themselves or where the connection with the gods was initiated. In early civilizations, the magicians, sages, and priests were assembled at the symbolic center of power or represented themselves as the center of the social system. Their claim to have a connection with gods or with the ancestors or to represent god's will on earth gave them a central position in their social system. Such ritualistic constructions of centers are also to be found in modern societies. In many cases centers claim pre-eminence with regard to knowledge and competence. It is not only the functional significance, but also the symbolic and psychological meaning of the terms center, core, heart, and top that attracts intellectuals and other successful knowledge producers.

2.3 Conflict Theory

Knowledge and information are not uncontested. Places of knowledge production compete with each other, newly fabricated knowledge may disrupt or destroy existing values, memories, and belief systems, and is not appreciated everywhere.

Knowledge can be used as a means to discipline, control, regulate, exclude, and manipulate. Production and distribution of knowledge and information are often associated with censorship, forgery, and propaganda. Many national memories, historical 'realities', social statistics, and 'truths' are socially constructed according to the interests of dominant groups and contested by other groups. Those in power have always tried to exercise tight control over institutions of knowledge production and information distribution, fortunately with limited success. The shamans, magicians, and priests of earlier centuries tried to ensure their monopoly of holy knowledge. A number of European universities were established not to improve knowledge and culture, but to strengthen the political power of the rulers, to generate an élite cadre or to erect an 'intellectual fortress' against unwanted influences.

The representations of the culture, history, and geography of colonies by imperial powers (Said 1978, Gregory 1998) are perfect examples of the assertion that culture (knowledge) underwrites power and power elaborates culture (knowledge). After the in-

troduction of compulsory schooling by the nation state schools, universities, and academies were constantly used by the central authority to enforce the 'universal' knowledge of those in power against the 'particular' knowledge and narratives of minorities. Since the school was considered an important instrument in the socialization process of the younger generation and of immigrants, its mission included consolidating the nation state by assimilating ethnic minorities, by supporting ideological and political propaganda, by transmitting historical narratives, and by creating national identities. The construction of belief systems, the legitimation of new knowledge, the creation of a regional or national identity, as well as the production, distribution, control, and suppression of information, is guided and controlled by the 'centers of calculus' (Latour 1987). The use of knowledge as a means to discipline and control, the fact that the spatial diffusion of knowledge may be accompanied with disjunction, contestation, opposition and debate, and the contest between various forms of knowledge, have contributed to the spatial concentration and separation of knowledge.

3. *Early Roots, Research Topics, and Possible Future Directions for the Geography of Knowledge, Education, and Skills*

3.1 *Early Roots and History*

The difficulties in measuring various types of knowledge across space may have contributed to the fact that empirical studies for a long time focused rather on educational achievement and school systems than on knowledge itself. The study of the relation between knowledge and action, of creative milieus or of the role of indigenous and religious knowledge started much later.

The historical roots of a 'geography of knowledge, education, and skills' can be traced back to the first decades of nineteenth century. Some of the negative consequences of the industrial revolution such as poverty, child labor, crime, alcoholism, and a lack of decent housing aroused an interest in moral statistics, first in France and Great Britain, and later in other countries. Many nineteenth century social reformers believed that poverty, crime, and alcoholism were caused by ignorance and a lack of moral education. They therefore studied social and spatial disparities of illiteracy, the provision and quality of schools, the skills and salaries of teachers, the availability of books in households, and the educational attainment of children.

France was the first country comprehensively to study regional differences in illiteracy. The main sources were the examinations of military recruits and marriage registers (Furet and Ozouf 1977, Heffernan 1989). In 1826, Charles Dupin gave a lecture about the interrelation between the population's educational achievement and economic well being. In 1827, he published the *Carte figurative de l'instruction populaire de la France*. This map demonstrated large regional disparities in educational attainment between northern and southern France (reprint in Meusburger 1998, p. 193). The tables added to this map compared the educational attainment, the number of patents for inventions, and the membership in the *Académie Française* with various economic indicators suggesting a correlation between educational achievement and economic performance.

In Britain, a number of *Statistical Societies* and *Friendly Societies* emerged in the 1830s and 1840s to study the living conditions and social situation of the poor and to foster social reform. Literacy was again a main concern of these studies. In 1849, J. Fletcher published a map showing 'Ignorance in England and Wales' in the *Journal of the Statistical Society of London*.

Apart from occasional studies of drop out rates and truancy, of busing problems, and of the locations of schools and universities, human geography has neglected spatial disparities in knowledge, education, and research until the mid-1960s. This changed when Robert Geipel (Munich) and other geographers initiated and institutionalized a 'geography of education'. For detailed treatment of its history and development in various countries see Meusburger (1998).

3.2 *Research Topics and Possible Future Directions*

The newly emerging 'geography of knowledge, skills and education' in its early period of the 1960s and 1970s first focused on spatial diversities of school systems (size, equipment, location criteria, and catchment areas of schools and universities), on spatial patterns of expenditure on education, on the evolution of patterns of school provision, on cultural and economic effects of schools and universities, on spatial effects of educational policies and school planning concepts, on the effects of school closures, and on spatial disparities of educational achievement.

On the spatial macro-scale, indicators of educational achievement are used to describe the spatial dimension of social stratification and the spatial distribution of human resources. The educational achievement of the adult population is measured by indicators such as the proportion of illiterates, the proportion of the adult population who completed a certain level of education (e.g. proportion of university graduates) or by median school years completed. Literacy is often regarded as a precondition for the modernization process. Spatial disparities in the educational attainment and behavior of pupils (students) can be measured by indicators such as the proportion of an age group attending (a certain type of?) school, by transition rates of an age group to various types of

secondary schools and universities, by school completion rates, by the scale of dropout and truancy, the suspension rates and many other indicators.

From the 1980s onward geography studied the personnel employed at schools, universities and research institutions. This included spatial disparities of the age-structure, skills, income, gender, social, ethnic, and regional origin, careers, and regional mobility of teachers, scientists, and other knowledge workers.

Influenced by research in economy and sociology, geography started to study the spatial variations in the transition from the school system to the occupational system in the 1980s and 1990s. The main research interests dealt with the correlation between educational achievement, first positioning in the labor market and professional career, and the spatial discrepancies and balancing mechanisms between supply and demand of skilled and unskilled labor. Migration motive patterns and distances of highly and lowly educated, of skilled and unskilled persons, and the interrelationship between career and regional mobility are central issues in this research field. The 'brain drain' and the transfer of talent from peripheries to centers strongly contribute to the long-term persistence of spatial disparities at all scales.

The more recent research topics and possible future directions include issues such as: (a) spatial disparities of research input (financial, technical and human resources, infrastructures), and research output (e.g. patent density, innovation density); (b) spatial differences in the quality of schools (measured by achievement tests); (c) the necessity of adjusting curricula to local economic needs or cultural traditions; (d) the role of schools (teachers) for local cultures and local empowerment; (e) the definition and spatial delimitation of creative (innovative) milieus, regional, and international differences of high-tech competitiveness; (f) the effects of telecommunications on the spatial distribution of work places for knowledge workers; (g) the role of indigenous and religious knowledge; (h) the evolvement and long term spatial effects of scientific networks; (i) the relation between human resources, endogenous economic development and economic conduct; (?j) the role of knowledge producers in the construction of historical narratives, national memories, and regional identities; and (k) the use of knowledge or enforced ignorance as a means to discipline and control. This list of research topics is never closed and strongly depends on the region, culture and time period studied. As knowledge is 'a capacity for social action' (Stehr 1994), and as it influences the perception, evaluation, decision making and actions of individuals, and the conduct of constituencies, a knowledge-based approach can be applied to a broad range of topics in human geography.

See also: Authority, Social Theories of Collective Memory, Psychology of; Communication: Geographic Aspects; Diffusion, Sociology of; Education: Skill Training; Elites: Sociological Aspects; Equality of Opportunity; Ethnic Conflicts; Identity: Social; Illiteracy, Sociology of; Information Society, Geography of; Innovation, Theory of; Intellectuals, Sociology of; Knowledge, Sociology of; Minorities; Modernization, Sociological Theories of; Power in Society; Social Change: Types; Social Evolution, Sociology of; Social Geography; Space and Social Theory in Geography

Bibliography

Agnew J 1999 Mapping political power beyond state boundaries: territory, identity, and movement in world politics. *Millennium Journal of International Studies* **28**: 499–521

Feldman M P, Florida R 1994 The geographic sources of innovation: technological infrastructure and product innovation in the United States. *Annals of the Association of American Geographers* **84**: 210–29

Furet F, Ozouf J 1977 *Lire et Ecrire. L'Alphabétisation des Français de Calvin à Jules Ferry*. [*Reading and writing. The rise of literacy in France from Calvin to Jules Ferry*]. Les Editions de Minuit, Paris, 2 Vols.

Geipel R 1966 *Angewandte Geographie auf dem Feld der Bildungsplanung* [*Applied geography on the field of educational planning*]. *Tagungsbericht des Deutschen Geographentages* 1965 **39**: 448–57

Gould W T S 1993 *People and Education in the Third World*. Longman, Scientific & Technical, Harlow, New York

Gregory D 1998 Power, knowledge and geography. *Geographische Zeitschrift* **86**: 70–93

Hägerstrand T 1966 Aspects of the spatial structure of social communication and the diffusion of information. *Papers and Proceedings of the Regional Science Association* **16**

Heffernan M 1989 Literacy and geographical mobility in nineteenth century provincial France: Some evidence from the Département of Ille-et-Vilaine. In: *Local Population Studies* **42**: 32–42

Latour B 1987 *Science in Action*. Open University Press, Milton Keynes, UK

Livingstone D N 1995 The spaces of knowledge: contributions towards a historical geography of science. *Environment Planning D* **13**: 5–34

Maillat D 1995 Territorial dynamic, innovative milieus and regional policy. *Entrepreneurship and Regional Development* **7**: 157–65

Malecki E J 1997 *Technology and Economic Development*, 2nd edn. Addison, Wesley, Longman, Essex

Marsden W E 1987 *Unequal Educational Provision in England and Wales: The Nineteenth-Century Roots*. Woburn Press, London

McClintock Ch G 1988 Evolution, systems of interdependence, and social values. *Behavioral Science* **33**: 59–76

Meusburger P 1998 *Bildungsgeographie. Wissen und Ausbildung in der räumlichen Dimension* [*Geography of Education. Knowledge and Education in the Spatial Dimension*]. Spektrum, Heidelberg, Germany

Mintzberg H 1979 *The Structuring of Organizations. A Synthesis of the Research*. Prentice-Hall, Englewood Cliffs, NJ

Pred A R 1973 *Urban Growth and the Circulation of Information*, 1st edn. Harvard University Press, Cambridge, MA

Said E 1978 *Orientalism*. Penguin, London

Stehr N 1994 *Knowledge Societies*. Sage, London

Strassoldo R 1980 Centre-periphery and system boundary: Cultural perspectives. In: Gottmann J (ed.) *Centre and Periphery. Spatial Variation in Politics*. Sage, Beverly Hills, CA,, pp. 27–61

Thrift N 1985 Flies and germs: A geography of knowledge. In: Gregory D, Urry J (eds.) *Social Relations and Spatial Structures*. Macmillan, Basingstoke, Hampshire, pp. 366–403

Törnqvist G 1968 *Flows of Information and the Location of Economic Activities*. Gieerup, Lund

World Bank 1999 *World Development Report 1998/99. Knowledge for Development*. Oxford University Press, Oxford, UK

P. Meusburger

Knowledge (Explicit and Implicit): Philosophical Aspects

Explicit knowledge is knowledge that the knower can make explicit by means of a verbal statement: 'Someone has explicit knowledge of something if a statement of it can be elicited from him by suitable enquiry or prompting' (Dummett 1991). Implicit knowledge can then be defined simply as knowledge that is not explicit. On this construal, implicit knowledge corresponds roughly to what Polanyi calls 'tacit knowing': 'we can know more than we can tell' (1967).

The distinction between explicit and implicit knowledge might seem to be connected closely to the familiar distinction (Ryle 1949) between knowledge of a proposition ('knowing that') and possession of a skill ('knowing how'). In everyday cases of 'knowing that' (such as knowing that the atomic number of lead is 82) the knower can say what it is that they know. In contrast, it is manifestly possible to have a skill (such as the ability to tie one's shoelaces) while being quite unable to give any verbal account of how the performance is achieved. So 'knowing how' may provide examples of Polanyi's 'tacit knowing.'

But implicit knowledge, as it is considered here, is closer to 'knowing that' than to 'knowing how.' In cases of implicit knowledge, a proposition or rule is known, but this is not available to the knower for verbal report. In the absence of verbal report, an attribution of implicit knowledge must be supported by other kinds of empirical evidence. But more fundamentally, we need some account of what it is for a subject to possess knowledge but to be quite unable to make it explicit (cf. Dienes and Perner 1999).

1. *Implicit Memory and Perception*

In cognitive psychology, a distinction is drawn between explicit and implicit tests of memory. Explicit memory tests 'make explicit reference to and demand conscious recollection of a specific previous experience.' But in an implicit memory test 'memory for a recent experience is inferred from facilitations of performance … that need not and frequently do not involve any conscious recollection of the prior experience' (Schacter 1989). For example, a subject might be shown a list of words and then tested in various ways. In an explicit test, the subject might be asked to say which words were on the list (recall) or might be presented with words and asked whether they appeared on the list (recognition). In an implicit test, the subject might be given a task that involves responding to words (e.g., by reading them aloud). Memory for the words on the list would be inferred from the fact that the subject reacts faster to those words than to words that were not on the list. The subject's reaction to a word that was on the list is said to be 'primed' by the earlier exposure to that word.

Intuitively, explicit tests of recall or recognition probe whether the subject still knows, in the everyday sense, which words were on the list. In contrast, implicit tests seem to provide evidence that information about which words were on the list is in some way present in the subject, perhaps in a form that can influence the subject's performance only on particular tasks. This intuitive idea that the two kinds of tests probe rather different kinds of memory, or information storage, is supported by findings that there are factors that affect performance on explicit tests and implicit tests differently (Schacter 1989). For example, extending the interval between the initial exposure to the list and the subsequent test may impair performance on an explicit test but not on an implicit test. Conversely, changing the sensory modality of presentation between initial exposure to the list and subsequent test (e.g., from auditory to visual presentation) significantly reduces priming effects in an implicit test but has little or no effect on recall and recognition in an explicit test. In addition, under some conditions performance on an implicit test has been shown to be uncorrelated with success or failure on an explicit test.

Priming effects may be observed even in the total absence of recall or recognition. When a word is initially presented with a very short exposure, and is followed by another stimulus to produce 'backward masking', the subject is quite unable to report anything about the word. But that initial presentation may still facilitate responses to the same word in a subsequent task. The results of masked priming experiments demonstrate implicit memory without explicit memory.

In cases where a subject is unable, even at the time of the first presentation, to report which word appears on the screen, we may say that perception itself is implicit rather than explicit: it is 'perception without awareness.' The influence of perceptually presented information of which the subject is unaware is also apparent in an experiment involving very brief presentations of photographs of faces that show expressions of emotions (Dimberg et al. 2000). Subjects are shown a happy, neutral, or angry face, rapidly followed by a

neutral face (to produce backward masking). They report seeing nothing before the neutral face. Yet their facial reactions, detected by measuring activity in facial muscles, are significantly different, depending on the emotional expression of the briefly presented face.

2. Knowledge and Belief

It is natural to describe these cases of implicit memory and implicit perception by saying that the tests reveal that the subject knows more than she can tell. But it is also natural to raise a query about the use of the term 'know' here. In epistemology, knowledge is usually taken to be true belief that meets some further condition involving justification of the belief or reliability of the method by which the belief is produced. The mechanisms of information gathering in perception and information storage in memory are plausibly reliable. But the notion of justification does not seem to be applicable in cases where the subject is unaware of the presence or influence of the information. So, if knowledge always involves justification, then so-called implicit knowledge does not seem to meet the requirements for knowledge. And even if knowledge is construed in a more inclusive way, there is room for doubt as to whether implicit knowledge amounts to a belief on the part of the subject.

At least three factors contribute to this doubt. One is the subject's lack of awareness. But this is not, by itself, decisive since it is possible to make sense of the notion of an unconscious belief. A second factor is that implicit knowledge seems to be limited in its influence. A belief may figure in many different ways in the subject's subsequent reasoning and planning, depending on their other beliefs and desires. But implicit knowledge that a happy face has been presented has only been shown to affect facial muscles. States of implicit knowledge do not appear to be inferentially integrated with the subject's beliefs (Stich 1978); they are, rather, inferentially isolated. However, this again is not quite decisive, since we can make sense of the idea of compartmentalized beliefs.

A third factor contributing to the doubt about the applicability of the folk psychological notion of belief is that believing requires the possession of concepts. No one can believe that penguins waddle without having some concept of penguins and of waddling. But, in at least some of the cases of implicit knowledge that we have mentioned, the subject does not need to possess the concepts that a theorist would use to characterize the information that has been gathered and stored. Thus, for example, the involuntary facial response to a happy or angry face does not seem to depend on the subject having any conceptualized grasp of the emotions of happiness or anger. A subject who lacks the concepts of happiness or anger cannot have any beliefs about happiness or anger. Yet such a subject may, it seems, be described as possessing implicit knowledge that a happy face or an angry face has been presented.

These considerations do not show that cases of implicit knowledge should never be grouped together with ordinary knowledge and belief. But examples of implicit memory and perception illustrate how, in cognitive psychology, the notion of implicit or tacit knowledge encompasses unconscious storage and processing of information whether or not this falls within the domain of the folk psychological notion of belief.

3. Preserved Implicit Knowledge in Neuropsychological Patients

Patients suffering from brain damage as a result of a stroke or head injury often show preserved implicit knowledge in the absence of explicit knowledge (Schacter et al. 1988).

Amnesic patients cannot remember recent experiences. If, for example, an amnesic patient is given a list of words to study she will perform poorly on a subsequent test of recall or recognition. But studying the list, even though it goes unremembered, may affect subsequent performance. In some cases, amnesic patients who are asked to complete three-letter sequences with the first words that come to mind, show the influence of the studied list in much the same way as happens with normal subjects (Warrington and Weiskrantz 1968). This is implicit memory in the absence of explicit memory.

Some patients are blind in a region of their visual field as a result of damage to their visual cortex. They are unable to report whether a light stimulus is present or absent in that region, to say whether a presented stimulus is, for example, an X or an O, or to discern the emotional expression on a face presented in the blind field. The patients say that they do not see the stimulus at all. Yet some of these patients, when asked to guess whether a stimulus is present or absent, or whether the stimulus is an X or an O, or whether a video clip presented in the blind field shows a happy or a fearful face, perform significantly above chance levels. This is the phenomenon of 'blindsight' (Weiskrantz 1986, de Gelder et al. 1999). It is a case of implicit perception in the absence of conscious or explicit awareness of the salient properties of the presented stimulus.

Prosopagnosic patients are unable to recognize and identify familiar faces. If, for example, a patient is asked to classify photographs as being of familiar or unfamiliar faces or to match names to faces, they perform at chance levels. Yet their autonomic affective responses (in particular, skin conductance responses) to familiar faces are significantly different from those to unfamiliar faces. A prosopagnosic patient may be unable to classify the faces of famous people according to their occupation (e.g., politician or television personality) although they can, of course, correctly assign occupations to the names of these people. But,

in some cases, presenting the face of a television personality alongside the name of a politician or vice versa interferes with the patient's performance when asked to assign an occupation to the name, just as it does in normal subjects (Young 1998). Information about the occupation of the person whose face is presented affects performance even though it is not available for verbal report.

Blindsight and other neuropsychological phenomena present challenges for both philosophical and empirical theories about consciousness. They also raise questions about the extent to which normal subjects are authoritative about the workings of their own minds. A normal subject would probably find it compelling to suppose that, when he sees the face of a television personality, his conscious recognition of the face interferes with his classification of a simultaneously presented name as that of a politician. Although this is likely to be correct, the fact that the same pattern of interference is found in patients who are unable to recognize faces raises an alternative possibility. It is at least conceivable that, contrary to what it is so compelling to suppose, the interference is produced in normal subjects in the same way as in prosopagnosic patients, by a process of which the subject is quite unaware.

4. *Implicit Learning*

Subjects may be trained to classify items as belonging or not belonging to some category by being provided with a principle or rule of classification. That is, the teacher may make the principle or rule explicit. Alternatively, subjects may be provided with a set of exemplars of the category; or they may simply be provided with feedback as to whether their classifications are correct or incorrect. On the basis of studying a set of exemplars, or after a period of training, a subject may achieve performance on the classification task that is significantly better than guessing at random. Indeed, the subject's classifications may be reliably correct. A subject who has attained a reliable level of performance may then be able to state the principle or rule of classification. In this case, even without being explicitly told the rule, the subject has gained explicit knowledge of the rule. But it may happen that the subject is quite unable to state the rule, even though his classifications reliably conform to the rule. In this case, the subject may be said to have learned the rule implicitly. The subject has achieved conformity to the rule without being able to state the rule.

Research in the area of implicit learning involves many complex issues (Berry and Dienes 1993, Shanks and St. John 1994). An experimenter may assign items to a category on the basis of a particular rule of classification, R, and may seek to discover whether a subject who has learned to make the same category assignment has arrived at explicit knowledge of the same rule. But asking the subject whether he is using rule R to make his classifications is problematic since considering that question may itself bring about a change in the subject's explicit knowledge.

Instead, the subject may simply be asked what rule they are using. But, if they do not produce a statement of rule R in response to this question, it by no means follows that the subject's classifications depend on implicit rather than explicit knowledge. One reason is that a rule that is explicitly known might not be volunteered because, for example, the subject is not entirely confident that it is correct. A second reason applies even if the subject does not have explicit knowledge of rule R. It may be possible to achieve classification performance that is significantly better than guessing at random, or even reliably correct over a large range of cases, by using explicit knowledge of some other rule.

To at least some extent, problems with the notion of implicit learning can be overcome by identifying independently well-motivated characteristics of implicit and of explicit knowledge. If a subject does not provide a statement of rule R but nevertheless makes classifications in conformity with rule R then their classification performance can be assessed in the light of these characteristics.

There is, however, a more general problem with the notion of implicit learning of a rule. So far, this has been characterized as learning to make classifications that conform to the rule, yet without explicit knowledge of the rule. We have observed some difficulties in deciding empirically whether explicit knowledge is really absent. But the more general problem is that any battery of performance that conforms to a rule inevitably conforms to more than one rule. If a subject's classification performance conforms to rule R_1 and to rule R_2 and the subject has explicit knowledge of neither rule, then the subject may be said to have implicitly learned something. But that does not settle whether the attained state of the subject should be described as implicit knowledge of R_1 or of R_2. In a case where two rules require just the same classifications, there is a question of principle whether any sense can be given to the notion of implicit knowledge of one rule rather than the other. A considerable body of research on implicit learning concerns the task of learning an artificial grammar (Reber 1989). But it is in the context of discussions of knowledge of natural language grammars that this question of principle has its original home.

5. *Tacit Knowledge of Rules*

In a famous passage, Chomsky (1965) says: 'Obviously, every speaker of a language has mastered and internalized a generative grammar that expresses his knowledge of his language. This is not to say that he is aware of the rules of the grammar or even that he can become aware of them.' Knowledge of the rules or

principles of a generative grammar is usually described as tacit knowledge and it counts as implicit knowledge according to our definition. The claim that ordinary speakers of natural languages tacitly know a grammar for their language is a core element of Chomskyan linguistics (e.g., Chomsky 1986, 1995).

Some philosophers have argued that the claim involves a conceptual confusion on the grounds that the notion of a rule of language belongs with the idea of a normative practice in which people advert to rules to justify, criticize or excuse their actions (Baker and Hacker 1984). But the notion of a tacitly known rule should be distinguished from the notion of a rule that figures in a normative practice. Once the distinction is made, there remains an important question about the relationship between the rules that are tacitly known by individual speakers and the normative practices in which those speakers participate. But that question does not indicate any conceptual incoherence in the very idea of tacit knowledge.

Other critics have urged that the ideas of tacit knowledge of rules and of unconscious information processing in general are problematic because of a deep connection between intentionality and consciousness (Searle 1990). But in response to this worry it can be said that, while there may be a connection between consciousness and the intentionality of beliefs, the states of tacit knowledge that Chomsky talks about are importantly different from belief states (Stich 1978, Davies 1989).

5.1 Quine's Challenge

Quine challenges Chomsky's introduction of the notion of tacit knowledge by making use of the distinction between behavior that conforms to a rule and behavior that is guided by a rule. A subject can behave in a way that conforms to a rule without using the rule to guide his behavior for, as Quine (1972) uses the notion of guidance: '(T)he behavior is not *guided* by the rule unless the behaver knows the rule and can state it.' Guidance requires explicit knowledge.

Chomsky's tacit knowledge is supposed to require less than explicit knowledge; but it cannot be equated with mere conformity. In fact, conformity to rules is neither necessary nor sufficient for tacit knowledge of those rules. It is not necessary, since the presence of tacit knowledge of rules does not guarantee perfect deployment of that knowledge in actual performance. It is not sufficient, since a tacit knowledge claim is not offered as a summary description of behavior but as a putative explanation of behavior. There will always be alternative sets of rules that require just the same behavior for conformity; but it is part of the idea of tacit knowledge that a speaker's actual behavior might be correctly explained in terms of tacit knowledge of one set of rules rather than the alternatives.

It is at this point that Quine (1972) poses his challenge. He insists that, if an attribution of tacit knowledge is an empirical claim that goes beyond a summary of conforming behavior, then it should be possible to indicate what kinds of evidence would count in favor of or against that empirical claim. He also insists that this evidence should involve the subject's behavior. To this latter point, it is reasonable to reply that there can be no a priori limit on the kinds of evidence that might be relevant to an empirical claim. So it is not legitimate to restrict evidence to the behavior of the very subject to whom the attribution of tacit knowledge is being made. But the more general point about evidence is a fair one.

5.2 Sketch of an Account of Tacit Knowledge

More fundamental, however, than the question of what evidence would support an attribution of tacit knowledge is the question what the correctness of such an attribution would consist in. We can sketch one answer to this fundamental question by using an example that involves not rules of syntax but very simple letter-sound rules of the kind that could be employed in reading words aloud. Suppose that one of these rules states that if a word begins with the letter 'b' then its pronunciation begins with the sound /B/. If a subject's pronunciation behavior conforms to this rule then it displays a pattern. Whenever a presented word begins with 'b,' the subject's pronunciation begins with /B/. But the transitions that concern us are not these transitions from presentation to pronunciation. Rather, we focus on transitions amongst beliefs or states of information. Whenever the subject starts out with the information that the presented word begins with 'b,' the subject ends up with the information that the word's pronunciation begins with /B/.

If these states were beliefs, then the subject's pattern of transitions from belief state to belief state would be accounted for if the subject possessed explicit knowledge of the 'b'-/B/ rule. This piece of explicit knowledge would figure as a common factor in the causal explanations of the subject's transitions from belief to belief. In contrast, there would be no such common factor if the subject had merely memorized the pronunciation of each of a large number of words beginning with 'b.' The difference between having explicit knowledge of the rule and having an independent piece of knowledge for each of the instances that fall under the rule corresponds to a difference in causal-explanatory structure.

In fact, we do not assume that the subject has beliefs about words and their pronunciations; the transitions may involve states of the kind that figure in information-processing psychology. Still less do we assume that the subject either has explicit knowledge of pronunciation rules or else has explicitly memorized the pronunciations of words. But we do make use of the idea of a common factor in the causal explanations of transitions. An attribution of tacit knowledge of the

'b'-/B/ rule can be construed as the claim that there is a single state of the subject that figures in the causal explanations of the various particular transitions that the subject makes (Evans 1981).

In general, a state of tacit knowledge is a state that figures as a common factor in causal explanations of certain transitions amongst states of information (or beliefs). Where different sets of rules require the same transitions for conformity, the attribution to a subject of tacit knowledge of a particular set of rules is made correct by the presence of a particular structure in the causal explanations of the subject's transitions (see also Davies 1987, Peacocke 1989).

Once an account of tacit knowledge in terms of causal-explanatory structure has been given, it is a relatively straightforward matter to give examples of empirical evidence that would confirm the attribution to a subject of tacit knowledge of a particular set of rules such as a grammar. Indeed, some of this evidence meets Quine's additional requirement of concerning the behavior of the subject to whom the attribution is being made. Relevant behavioral evidence could come from experimental studies of language acquisition, language perception, and language dysfunction following brain damage, while further evidence would be available from neural imaging.

6. *Explicit and Implicit Representation*

According to the basic notion of explicit knowledge introduced at the beginning of this article, explicitness is a matter of the subject being able to present information in linguistic form. It is not a matter of how the information is stored in the subject's mind or brain when it is not being called upon.

Suppose that someone knows, in the everyday sense of that term, the axioms of some theory. Provided that this knowledge can be verbally articulated it counts as explicit knowledge. Now consider some theorem that is derivable from those axioms. The person who knows the axioms may well, with some suitable enquiry and prompting, be able to see that the theorem follows from the axioms and to state it verbally. For example, someone who explicitly knows some elementary arithmetical facts may be able to work out, and to state, that 68 plus 57 equals 125. On Dummett's account, this latter piece of knowledge, even though it is computed when needed rather than being stored in memory, counts as no less explicit than the stored elementary facts from which it is derived. Indeed, it counts as explicit knowledge even before the knower works it out; for explicitness is defined in terms of the possibility of eliciting a verbal statement by enquiry or prompting. Once some propositions are classified as explicitly known, the category of explicit knowledge also includes at least some of the as-yet-undrawn consequences of those propositions.

But the terms 'explicit' and 'implicit' are also used to mark a number of distinctions that are specifically related to the storage and processing of information. Thus, Dennett introduces a quite different distinction from Dummett's when he says (1983): 'Let us have it that for information to be represented *implicitly*, we shall mean that it is *implied* logically by something that is stored explicitly.' On this different usage, the as-yet-undrawn consequences of propositions that count as explicitly stored would be classified as implicitly represented.

In fact, relative to any given notion of explicit representation or storage of information, it is possible to define a whole family of notions of implicit representation or knowledge. The members of the family differ over the inferential resources that can be used in drawing out consequences from the information that is explicitly stored. We might, for example, consider only deductive inferences that can be carried out by the person in question. Then, if a person explicitly stores elementary arithmetical facts in memory and is able to infer that $68 + 57 = 125$, this would count as implicit representation of that latter fact. Alternatively, we might allow any valid deductive inference, whether or not the person can actually carry it out. In that case, implicit representation or knowledge would outrun the possibility of eliciting from the person a verbal statement. We might go beyond deductive logic and allow methods of induction or rule extraction, with or without a restriction to methods that are available to the person in question. In that case, a person who stores in memory only facts about particulars might count as having implicit knowledge of generalizations or rules that subsume those facts.

The notion of implicit representation also depends on a prior conception of explicit storage. But there is no uniquely correct account of what explicit storage amounts to. One idea that is, however, fairly naturally associated with explicitness in the context of theories of information storage and processing is that explicitly stored information has to be accessed before it can be used. An access problem may result in a failure to use stored information that is relevant to the task in hand. A second idea that sometimes goes along with the notion of explicitness is that explicit storage involves the use of a language-like code. Putting the two ideas together, we have a picture in which explicit storage of information is a matter of linguistic inscriptions on pages that are not always easy to locate. Drawing out the consequences of explicitly stored information would then be naturally conceived as manipulating inscriptions in ways that are analogous to the formal procedures involved in the derivation of theorems from axioms in logical systems (Fodor 1987).

The notions of explicit storage and implicit representation under discussion in this section belong with mechanistic theories about the storage and processing of information. They could be applied to any information-processing systems and have no

special connection with human subjects. Even in the case of human subjects, information might be explicitly stored yet not accessible to consciousness and not available to the subject for verbal report.

A state of tacit knowledge (as sketched in Section 5) could be constituted by the explicit storage of a rule that was not accessible to consciousness. Indeed, it sometimes seems to be assumed that, within the picture of the storage and processing that we have just sketched, tacit knowledge of a rule is always embodied in a mechanism by having the rule explicitly stored. There is a clear distinction between explicitly storing information and using that information; use of stored information involves processes for accessing that information. But such access processes are not typical of the operation of connectionist networks (McClelland et al. 1986): 'Using knowledge in processing is no longer a matter of finding the relevant information in memory and bringing it to bear; it is part and parcel of the processing itself.' So, against the background of the assumption that tacit knowledge involves explicit storage, it is often said that what is distinctive of connectionist networks is that they perform cognitive tasks without having tacit knowledge of rules.

The account of tacit knowledge in terms of causal-explanatory structure certainly allows for the case of tacit knowledge that is a matter of explicit storage. But the key notion in the account is that of a common factor in causal explanations of certain transitions amongst states of information. So the account also allows for the possibility of tacit knowledge that is 'part and parcel of the processing,' and it is left open that tacit knowledge may sometimes be embodied in connectionist networks (Davies 1995).

See also: Cognitive Science: Overview; Implicit Learning and Memory: Psychological and Neural Aspects; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Meaning and Rule-following: Philosophical Aspects; Perception: Philosophical Aspects; Reference and Representation: Philosophical Aspects

Bibliography

Baker G P, Hacker P M S 1984 *Language, Sense and Nonsense: A Critical Investigation into Modern Theories of Language*. Blackwell, Oxford, UK

Berry D C, Dienes Z 1993 *Implicit Learning: Theoretical and Empirical Issues*. Lawrence Erlbaum Associates, Hove, East Sussex, UK

Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Chomsky N 1986 *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York

Chomsky N 1995 *The Minimalist Program*. MIT Press, Cambridge, MA

Davies M 1987 Tacit knowledge and semantic theory: Can a five per cent difference matter? *Mind* **96**: 441–62

Davies M 1989 Tacit knowledge and subdoxastic states. In: George A (ed.) *Reflections on Chomsky*. Blackwell, Oxford, UK

Davies M 1995 Two notions of implicit rules. In: Tomberlin J E (ed.) *Philosophical Perspectives, 9: AI, Connectionism, and Philosophical Psychology*. Ridgeview, Atascadero, CA

de Gelder B, Vroomen J, Pourtois G, Weiskrantz L 1999 Non-conscious recognition of affect in the absence of striate cortex. *NeuroReport* **10**: 3759–63

Dennett D C 1983 Styles of mental representation. *Proceedings of the Aristotelian Society* **83**: 213–26. Reprinted in: *The Intentional Stance*. MIT Press, Cambridge, MA, 1987.

Dienes Z, Perner J 1999 A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences* **22**: 735–808

Dimberg U, Thunberg M, Elmehed K 2000 Unconscious facial reactions to emotional facial expressions. *Psychological Science* **11**: 86–9

Dummett M 1991 *The Logical Basis of Metaphysics*. Harvard University Press, Cambridge, MA

Evans G 1981 Semantic theory and tacit knowledge. In: Holtzmann S H, Leich C M (eds.) *Wittgenstein: To Follow a Rule*. Routledge and Kegan Paul, London. [Reprinted In: *Collected Papers*. Oxford University Press, Oxford, 1985]

Fodor J A 1987 Why there still has to be a language of thought. In: *Psychosemantics*. MIT Press, Cambridge, MA

McClelland J L, Rumelhart D E, Hinton G E 1986 The appeal of parallel distributed processing. In: Rumelhart D E, McClelland J L, the PDP Research Group (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 1: Foundations*. MIT Press Cambridge, MA

Peacocke C 1989 When is a grammar psychologically real? In: George A (ed.) *Reflections on Chomsky*. Blackwell, Oxford, UK

Polanyi M 1967 *The Tacit Dimension*. Routledge and Kegan Paul, London

Quine W V O 1972 Methodological reflections on current linguistic theory. In: Davidson D, Harman G (eds.) *Semantics of Natural Language*. Reidel, Dordrecht

Reber A S 1989 Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General* **118**: 219–35

Ryle G 1949 *The Concept of Mind*. Hutchinson's University Library, London

Schacter D L 1989 Memory. In: Posner M I (ed.) *Foundations of Cognitive Science*. MIT Press, Cambridge, MA

Schacter D L, McAndrews M P, Moscovitch M 1988 Access to consciousness: Dissociations between implicit and explicit knowledge in neuropsychological syndromes. In: Weiskrantz L (ed.) *Thought Without Language*. Oxford University Press, Oxford, UK

Searle J 1990 Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences* **13**: 585–95

Shanks D R, St. John M F 1994 Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences* **17**: 367–95

Stich S 1978 Beliefs and subdoxastic states. *Philosophy of Science* **45**: 499–518

Warrington E K, Weiskrantz L 1968 New method of testing long term retention with special reference to amnesic patients. *Nature* **217**: 972–4

Weiskrantz L 1986 *Blindsight: A Case Study and Implications*. Oxford University Press, Oxford, UK
Young A W 1998 *Face and Mind*. Oxford University Press, Oxford, UK

M. Davies

Knowledge Representation

1. Knowledge Representation: Ideal and Reality

The fundamental conjecture of knowledge-based artificial intelligence (AI), known as the physical symbol system hypothesis (Newell and Simon 1976), posits that the cognitive processes of any broadly intelligent agent, natural or artificial, can be largely characterized in terms of a knowledge base: a transparent, symbolic representation of the agent's beliefs, intentions, and value judgments, together with reasoning processes that modify the knowledge base in rational ways, and perceptual and motor processes that connect the knowledge base to the outside world. The field of *knowledge representation* (KR) is the study of how such a knowledge base can be constructed.

The physical symbol system hypothesis requires that a knowledge base be symbolic, transparent, and modular. That is, the atomic elements of a knowledge base are symbols, each of which has an associated meaning. It is possible to identify small substructures of the knowledge base that correspond to specific beliefs, judgments, and intentions. For example, the belief that Holbein painted a portrait of Thomas More in oil might be represented in a predicate calculus formula like

$$\exists_X \text{ oil_painting}(X)$$
$$\wedge \text{ painter(holbein, } X) \wedge \text{ depicts } (X, \text{thomas_more})$$

or in a fragment of a semantic net, such as that shown in Fig. 1. Here the symbols are '$\exists$' (meaning 'there exists'), 'X' (a variable), '$\wedge$' (meaning 'and'),

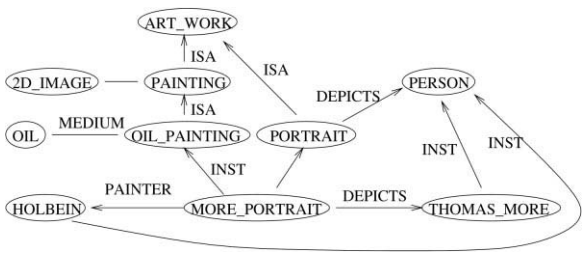

Figure 1
Fragment of a semantic net

'oil_painting,' 'painter,' 'holbein,' and 'thomas_more'; the parentheses and the comma indicate the syntactic structure. (The fact that some of these symbols are English words is purely for the convenience of the human reader; the formula could just as well be notated

$$\exists_X \text{ g001}(X) \wedge \text{ g002(g003, } X), \wedge \text{ g004}(X, \text{g005})$$

if it is accepted that 'g001' denotes the property of being an oil painting, that 'g002' denotes the relation of a person painting a painting, and so on.) It should be noted that not every belief of the agent can be explicitly expressed in the knowledge base. For example, a person who knows that Holbein painted a portrait of Thomas More will also know that Holbein knew how to paint; that Thomas More was painted; that Thomas More was born before Holbein's death; and so on. Clearly it is not plausible that each of these ancillary beliefs is separately represented in the knowledge base. Rather, they are represented implicitly, in the sense that they can be derived as needed.

Of course, the construction of the broadly intelligent artificial agent described above is today just a pipe dream for the distant future. Any work aimed at achieving a working system any time soon must be much more modest in its ambitions. Within the context of such limited objectives, the problem of knowledge representation is one of finding data structures that express the knowledge needed for the particular application.

Thus broadly construed, the problem of knowledge representation would appear to be equivalent to computer programming generally. Any program, after all, uses some data structures, and incorporates some kind of knowledge. What distinguishes a data structure as interesting from the point of view of knowledge representation is that its design reflects some of the features of the broadly intelligent knowledge-based agent described above. As artificial intelligence is the study of tasks that are easy for people (or animals) to carry out but difficult to implement in a computer, so knowledge representation is the design of data structures that express knowledge that people deal with easily, but are hard to encode in a data structure.

The design of a knowledge representation can be very roughly divided into four parts

(a) *Architectural* issues involve defining general-purpose schemas for knowledge representation and procedures for reasoning with these schemas. Most of the research in knowledge representation has been addressed to architectural issues, as they are largely independent of the particular domain and application (Reichgelt 1991).

(b) *Content* issues address such questions as what is the knowledge to be represented, how should the domain be conceptualized, what are the key concepts in the domain, and what kinds of partial knowledge must be dealt with. Since content issues are always

domain-dependent, and often task-dependent, they are much less susceptible to general treatment. However, there are some domains—such as time, space, epistemic states, plans, and so on—that are so ubiquitous in intelligent reasoning, and so difficult, that it is worthwhile studying them extensively in the abstract, without reference to a particular application (Davis 1990).

(c) The *implementation* of the knowledge representation as a data structure and of the reasoning mechanisms as procedures is usually carried out in a uniform way for a given architecture; alternatively, it may be highly specialized to a particular class of problems.

(d) Finally, any knowledge representation requires an *interface*, either to user queries or to application programs such as robot control. As interface issues are entirely task-dependent, there is very little that can be said about them in general.

This article will begin (Sect. 2) by discussing some features that are desirable in a knowledge representation. It will then (Sect. 3) survey some of the important architectures in knowledge representation. Section 4 will discuss the issue of content, illustrating it with a brief discussion of representations of time. Finally, Sect. 5 will discuss two alternative approaches to knowledge representation: neural networks and digital libraries.

2. *Desiderata*

The design of a knowledge representation, whether for a broadly intelligent agent or for some limited application, will generally aim toward the following objectives:

(a) *Scope*. A comprehensive knowledge representation scheme should ideally be extensible to cover any possible object of human thought, and any humanly possible belief system.

Table 1
Table of values

	Works		
Work	Medium	Genre	Artist
Last Supper	Fresco	Religious	da Vinci
Melancholia	Engraving	Allegory	Durer
Thomas More	Oil	Portrait	Holbein
Mona Lisa	Oil	Portrait	da Vinci
	Artists		
Artist	Country	Birth	Death
Durer	Germany	1471	1528
Holbein	Germany	1497	1543
da Vinci	Italy	1452	1519
O'Keefe	USA	1887	1986

Of course, the designer of a KR theory need not explicitly state how the theory can express knowledge about the Parthians, the blue whale, and the Crab Nebula. But a convincing case should be made that beliefs about such things can be expressed in the representation, if necessary. A complete KR theory should explain in detail how to represent fundamental and universal subjects of human thought, such as time, space, simple physical categories, the human mind, and human relations (Davis 1990, Hayes 1979, McCarthy 1968).

A knowledge representation for a limited application should cover all relevant aspects of relevant domains.

(b) *Expressivity*. The KR scheme should be able to express the various forms of partial and incomplete knowledge that naturally arise for the desired application.

For example, a standard form of representation is a collection of tables of values, known in computer science as a 'relational database.' For example, Table 1 displays a table of paintings and a table of artists.

However, even within its narrow scope, such a representation may not achieve reasonable expressivity, because there is no way to express plausible partial states of knowledge such as:

Turner lived in the nineteenth century.

Rembrandt painted many oil paintings.

Vermeer did not make any sculptures.

The Sistine Chapel and the sculpture of David were created by the same artist.

If such partial states of knowledge are important in the application, then a more expressive language must be found.

Another example: it is tempting to represent spatial information using a representation that corresponds to a single picture of the scene (see *Computational Imagery*). This is often called *diagrammatic representation*, (Glasgow et al. 1995), though this term is used in many different conflicting ways. Such representations are easily understood and support efficient inference. However, they cannot be used to represent partial information. For instance, an agent might know that A is a mile from B and that A is a mile from C without knowing the distance from A to C. However, it is not possible for a single picture to display the distances from A to B and from A to C without committing itself to some particular distance between B and C. Therefore, unless all the program's geographic knowledge can always be expressed in a precise map, this representation will not be sufficiently expressive.

(c) *Precision*. The meaning of a knowledge representation should be unambiguous, and, as far as the nature of the subject matter allows, well defined.

(d) *Support for reasoning*. It must be possible to model the processes of reasoning involved in cognition in terms of well-defined operations over the knowledge representation. For example, as mentioned above, an agent who knows that Holbein painted a portrait of

More should be able to conclude that Holbein knew how to paint, that More was born before Holbein died, and so on.

(e) *Efficiency of reasoning*. The implementation of reasoning operations over the knowledge representation must also be efficient. Conclusions must be derived in the time that they are needed, often a matter of seconds or fractions of a second.

(f) *Learnability*. A broadly intelligent agent must be able to learn over time from experience and from instruction. Therefore, modules must exist that can learn from experience and can generalize individual cases to general rules. The knowledge representation should support such learning modules (Mitchell 1997).

(g) *Support for uncertainty*. If an agent can only use knowledge that it knows with 100 percent certainty, it will be paralyzed. Rather, an agent must be able to compare relative likelihoods, and to take action on the basis of the best information it can get. A knowledge representation should therefore support the expression of relative degrees of certainty, and reasoning modules should be able to take these degrees of certainty into account (Pearl 1988).

(h) *Robustness*. The information in a knowledge base will sometimes be simply wrong or corrupted. The effects of a single error in the knowledge base should be limited in scope.

(i) *Ease of construction*. In the best case, it may be possible to build a large knowledge base automatically or semiautomatically from existing data sets, such as conventional databases, online libraries, or collections of digitized images. If the knowledge base must be constructed by hand, then a notation that is simple and natural enough to be used by programmers with little technical expertise is clearly preferable to one that can be used only by an expert in knowledge representation. Other things being equal, the less human labor, especially expert labor, that is required to build a knowledge base, the better.

(j) *Interface*. A knowledge representation should support a suitable interface to the user or the application.

Expressivity and precision are features, both of the architecture and of the content of a knowledge base. Scope is chiefly a feature of content, though expressivity and therefore architecture do bear on scope; some domains inherently require a very expressive language. Robustness, ease of construction, and support for reasoning, for learning, and for uncertainty are primarily features of the architecture. Efficiency of reasoning depends on the implementation, which, in turn, reflects the architecture, and on the class of problems being solved.

3. Architectures

A number of KR architectures are briefly surveyed below. The particular architectures included here have been included, in part because of their importance in KR research and applications, and in part to illustrate the range of possible approaches to KR.

3.1 Logic-Based Systems

A logic-based system represents knowledge as a set of sentences in a logical language. Reasoning is implemented as inference procedures that carry out deductive proofs authorized within the logical theory (Geneserth and Nilsson 1987) (see entry *Deductive Reasoning Systems*).

A logical theory characteristically consists of three parts

(a) *A syntax*, which specifies what kinds of symbols may be used, and how they can be combined to form well-formed sentences.

(b) *A semantics*, which specifies what kinds of meanings may be attached to symbols, and how these meanings combine to form the meaning of sentences.

(c) *A proof theory*, which specifies what manipulations over sentences constitute deductive proof.

For example, first-order logic (also known as the predicate calculus) is the most widely used logic in knowledge representation. In first-order logic, there are seven kinds of symbols.

(a) *Constant* symbols represent individual entities. For example 'holbein' may represent Hans Holbein. The symbol 'ad1517' may represent the year 1517.

(b) *Predicate* symbols represent properties of relation. For example, 'painter' may represent the two-place relation of a person painting a painting. The predicate 'precedes' may represent the two-place relation of one event preceding another.

(c) *Function* symbols represent mappings. For example, 'birth_of' may represent the one-place mapping of a person to the event of his/her birth. The function 'creation_of' may represent the one-place mapping of an artifact to the event of its creation.

(d) The *Boolean operators* are '$\wedge$' meaning 'and'; '$\vee$' meaning 'or'; '$\neg$' meaning 'not'; and '$\Rightarrow$' meaning 'implies.'

(e) *Variables*, such as X, Y, Z, represent indeterminate objects.

(f) The *quantifiers* are $\forall$ meaning 'for all,' and '$\exists$' meaning 'there exists.'

(g) Finally, the grouping symbols, which serve to specify the structure of a sentence, are the comma and the open and close parentheses.

Constant, predicate, and function symbols are known as *non-logical* symbols; they may be chosen, and their meanings may be assigned, by the designer of the knowledge base. Boolean operators, quantifiers, variables, and grouping symbols are known as *logical* symbols; their meanings are fixed by the logic. The distinction is roughly analogous to the distinction between open and closed categories in linguistics, or to the distinction between identifiers and reserved words in programming languages.

The syntactic rules for first-order logic establish that these symbols can be put together to form sentences such as

(1) $\forall_{X,Y}$ painter $(X, Y) \Rightarrow$ precedes(birth_of(X), creation-of(Y))

The semantic rules establish that, given the above assignments of meanings to the symbols 'painter,' 'precedes,' 'birth_of,' and 'creation_of,' the meaning of this sentence is that the birth of a painter precedes the creation of any of his works. The proof theory establishes that, given formula (1) above and the fact, 'painter(holbein, more_portrait),' it is possible to infer

(2) precedes (birth_of(holbein), creation_of(more_portrait)),

i.e., that Holbein was born before the portrait of More was painted. Details of these rules may be found in any logic textbook (e.g., Mates 1972)

First-order logic is probably the most common foundation for a logic-based knowledge representation, but many other logics have also been applied. Some of these extend first-order logic to achieve greater expressivity. *Modal logics* define a variety of operators on sentences. For instance, given a base sentence such as 'Bill Clinton is President,' a *temporal* operator might generate the sentence 'Bill Clinton has been President,' an *epistemic* operator might generate the sentence, 'Al Gore knows that Bill Clinton is President'; a *deontic* operator might generate the sentence 'Bill Clinton ought to be President.' These operators can be combined, as in the sentence 'Al Gore should have known that Bill Clinton would be President' (Turner 1984). *Fuzzy logic* allows the use of vague terms that are hold to greater and lesser degree; for example, 'George Washington was a tall person' is reasonably true, 'Abe Lincoln was a tall person' is quite true, 'Wilt Chamberlain was a tall person' is absolutely true (Zadeh 1987). *Probabilistic logics* allow sentences to be tagged with a measure of certainty (Bacchus 1990).

Other knowledge representations have used logics less expressive than the predicate calculus in order to achieve greater efficiency in a general purpose inference engine. For instance, in the *propositional calculus*, the only logical symbols are Boolean operators, and a separate symbol must be used for each atomic proposition. Thus there is one symbol for the proposition, 'Leonardo painted the Last Supper,' another for the (false) proposition, 'Rodin painted the Last Supper,' and so on. General rules like formula (1) above can then only be stated every instance for every pair X and Y. Despite this proliferation of symbols and formulas, however, as long as the underlying domain is not too large, it is often more effective to use the propositional calculus than the predicate calculus because the best methods of inference known are much faster for the propositional calculus than the predicate calculus (Selman et al. 1992).

The chief virtues of logic-based representation are precision and modularity. The meaning of a formula in a logic-based representation is completely and exactly fixed by the logic and the meaning of the symbols included. It is independent of any other formula in the knowledge base, and any other aspect of the knowledge base. Another important virtue is that procedures for deductive inference over logic-based knowledge bases are known. These procedures characteristically generate only valid inferences and all valid inferences (in the technical jargon, they are *sound* and *complete*) and, in practice, they can often be implemented so as to find useful answers quickly.

Their chief failing is difficulty of use. Building a knowledge base in a logic-based representation inevitably requires a large investment of expert labor. Moreover, even experts in knowledge representation find the encoding of complex domains in logical languages surprisingly difficult and error-prone. Logic-based languages also tend to be fragile. In principle, any inconsistency in any part of a logical theory invalidates the entire theory; and in practice small and subtle inconsistencies in foundational parts of a logic-based knowledge base can lead to bizarre and unpredictable results.

3.2 Semantic Networks

A *semantic network* consists of labelled nodes connected by labelled arcs (Fig. 1). A node may represent an individual such as Hans Holbein or a category, such as oil paintings. An arc represents a two-place relation between nodes, such as the relation 'painter' between a work of art and a person.

The most important arcs in any semantic network are those labelled 'INST' and 'ISA.' An 'INST' link goes from an individual I to a category C and asserts that I is an instance of C. An ISA link goes from category C to category D and asserts that C is a subset of D. Implicitly, therefore, if there is a chain of ISA links from C to D, then it can be inferred that C is a subset of D. If there is an INST link from I to C and then a chain of ISA links from C to D, then it can be inferred that I is an instance of D. For instance, in Fig. 1, it can be inferred that oil paintings are works of art, and that the portrait of Thomas More is a work of art.

The central reasoning mechanism in a semantic network is *inheritance*: a node inherits all the properties of the nodes that contain it. For instance, in Fig. 1, the node 'MORE_PORTRAIT' inherits the 'MEDIUM' arc to 'OIL' from 'OIL_PAINTING'; it inherits the 'FORM' arc to '2D-IMAGE' from 'PAINTING'; it inherits the 'DEPICTS' arc to 'PERSON' from 'PORTRAIT.' In other words, the system can infer that the portrait of More is a two-dimensional image from the facts that it is a painting, and that all paintings are two-dimensional images.

The semantic network architecture, as described above, has a number of advantages. It is easy and natural to use; its meaning can be defined precisely;

Table 2
Two frames

```
frame more_portrait
inst:   oil_painting;
genre:   portrait;
depicts:   thomas-more;
shape:   rectangle (29.5 * inch, 23.7 * inch);
surface:   wood;
date:   1527;
painter:   holbein;
owner:   frick;
image:   more.gif {file with displayable image};
endframe

frame oil_painting
isa:   painting;
medium:   oil;
surface:   ⟨OR⟩ [canvas, wood];
date:   ⟨CONSTRAINT⟩ [@.painter.birth <@.date <
        = @.painter.death];
painter:   ⟨OR⟩ [⟨INSTANCE⟩ person,
           ⟨INSTANCE⟩ workshop];
owner:   ⟨OR⟩ [lost, destroyed, ⟨INSTANCE⟩ person,
         ⟨INSTANCE⟩ institution];
endframe
```

and the inheritance mechanism is easy to implement and efficient. The chief disadvantage is that it is not very expressive. Also, the apparent simplicity of the representation can end up working against precision; unless the user adheres rigorously to the specifications of use, it is very easy to construct a semantic network that looks plausible, but that uses the same notation with two completely different meanings (Woods 1975).

It is possible, and indeed common, to increase the expressivity of the very minimal architecture described above by adding additional features, such as 'meta-arcs' that connect two arcs and describe the logical relations between them (Sowa 2000). However, the more such features are added, the less simple and natural the representation becomes.

3.3 Frame-based systems

A frame represents either an individual or a category, and associates with its subject a collection of pairs of a *slot* with a corresponding *filler*. A slot is just the name of a feature of an individual. A filler of a slot for an individual is just the value of that feature; for example, the 'color' slot for 'statue of liberty' has filler 'green.' The filler of a slot for a category contains information of some kind about the fillers of the same slot for individual instances of the category. This can take many different forms; it can be a particular value, a range of values, a constraint on values, or a function that computes the value. A tag, or structure of tags, indicates which type of information is involved. For instance, in frame 'oil-painting' in Table 2, the value of the 'date' slot is a constraint asserting that the date of a painting must take place within the painter's lifetime. The '@' sign used here is a special-purpose variable meaning the particular instance of the frame. The value of the 'painter' slot must either be a particular human or a particular workshop. The value of the 'owner' slot is either a particular person or institution, or the values 'lost' or 'destroyed.' (Note: the specific syntax used here is invented, but the features are typical of frame-based systems.) As in semantic nets, frames inherit fillers through 'inst' links, which contact an individual to a category, and through 'isa' links, which connect a subcategory to a super-category. Other forms of inference in frame-based systems include computing or verifying the value of a slot using inherited procedures or constraints and categorizing a new individual as an instance of a category based on its features (Minsky 1975, Borgida et al. 1989).

Obviously, frame-based systems are in many ways similar to object-oriented programming languages; indeed, the two theories interacted strongly in their development.

The chief advantages of frame-based architectures are expressivity, flexibility, and ease of use. The chief disadvantages is lack of precision, and lack of a well-defined model of inference. The architecture provides a wealth of features and options for both representation and inference, but only a weak underlying model. Hence, in a complex case, it is difficult to predict how these features will interact, or to explain unexpected interactions, which makes debugging and updating difficult.

4. Content

The most difficult problems facing the designer of a knowledge base are those of content: What knowledge is needed for a given application, what concepts are needed, and how can the knowledge be effectively encoded in a given architecture?

For specialized knowledge-based systems, known as expert systems (see *Expert Systems in Cognitive Science* and *Expert Systems in Medicine*), the accepted procedure of encoding knowledge involves an extended collaboration between a human expert and a 'knowledge engineer.' The knowledge engineer observes the expert solving a problem, and interviews him in as great depth as possible as to how he found the solution, how his reasoning was justified, what additional factors could have led him to other conclusions, and so forth. The knowledge engineer then abstracts a system of concepts and rules, and implements a prototype reasoning program. The knowledge engineer tests the program on a collection of examples, the expert provides a critique of the behavior of the program, and the knowledge engineer uses this critique to refine and improve the program (Stefik 1995).

In encoding commonsense knowledge, a popular methodology is to begin by defining a microworld, a small, coherent domain of study. Aspects of the real world that lie outside the microworld will either be idealized away or dealt with in a coarse, *ad hoc* manner. A collection of commonsensically obvious inferences in the microworld is assembled. The researcher constructs a formal model that suffices to justify these inferences; defines symbols to express the concepts used in the inferences; and expresses the knowledge involved in the inferences in terms of these symbols in the chosen knowledge-base architecture. Over the course of time, the microworlds considered become broader in scope and richer in detail (Davis 1990, 1998).

4.1 Time

The issues involved in content may be illustrated in a discussion of representations of time, certainly the most ubiquitous domain in automated reasoning.

Ontologically, time may be taken to be discrete or continuous, to be linear or branching, to consist of points or of intervals, to be Newtonian or relativistic. Which choice is made depends on the applications being considered. For instance, in a chess-playing program, it is natural to take time to be discrete (as nothing significant happens between successive states of the board), and to be branching (to express the options available to each player at each stage). In reasoning about most physical processes, it is usually desirable to take time to be continuous, in order to describe continuous change, and linear, since no choices are being made. (Uncertainty about the future, as opposed to choices about the future, is easily accommodated in a linear theory of time.) The relativistic model of time need only be invoked if the application involves multiple observers whose relative speed is a large fraction of the speed of light.

Many different languages have been devised for time. Some of the most popular include:

(a) *Time stamps*. If a linear model of time is used, and clock times are known precisely and are important, then each event can be marked by date and clock time.

(b) *Numerical constraints on time*. If a linear model of time is used, and clock times are important but are not fully known, then a reasonable representation can use a collection of symbolic constraints between the times of events, such as 'The date of the painting of More's portrait comes after Holbein's birth and before his death.'

(c) *The situation calculus*. This theory uses a branching, discrete model of time. There are three primary types of entities: situations, which are essentially instant of time; fluents, which are proposition whose value changes over time, such as 'Clinton is President'; and actions which mediate between one situation and a successor situation. The two principal symbols are 'results(S, A),' which maps a situation S and an action A to the situation that results if A is performed in S; and 'holds(S, F)' the assertion that fluent F holds in situation S (McCarthy and Hayes 1969).

(d) *The event/interval calculus*. Another representation for time popular in KR research is based around intervals and events. There are 13 possible temporal relations between intervals; interval I may entirely precede interval J; or the end of I may coincide with the beginning of J; and so on. The interval language contains primitives to describe these 13 relations, plus the reaction, 'occurs(I, E),' meaning event E occurs in interval I (Allen 1983).

(e) *Tense logic*. This was developed in the philosophical literature to formalize the tenses of natural language. It is a modal logic, containing operators like, 'ϕ will always be true,' 'ϕ has been true at some time in the past,' and so on (van Benthem 1983).

(f) *Dynamic logic*. This representation is popular in formal theories of programming languages. It uses a discrete, branching model of time. Statements have the form 'after P has been executed, ϕ will be true' where P is a plan or a computer program.

4.2 Very Large Knowledge Bases

Several recent projects have attempted to construct knowledge bases that encode some substantial fraction of commonsense knowledge, including the wellknown *CYC* project (Lenat 1995) and the HPKB and Ontolingua projects. The architectures of these systems combine logic-based and frame-based features, and supply powerful and well-designed inference engines and user-interfaces. The associated knowledge bases are extremely large, and have been successfully applied in a number of different tasks, including natural language interpretation, and front ends to commercial databases. No systematic evaluation of the contents of these systems has been published, and the degree of success that has been attained in addressing the fundamental challenges of knowledge representation and commonsense reasoning is therefore difficult to determine.

5. Alternative Approaches

The KR architectures we have considered above, together with many other proposals of a more or less similar flavor, such as production systems, constitute what may be called the 'classical,' or (with some question begging) the 'knowledge-based' approach to AI. Knowledge representation, in this view, involves large, complex structures of symbols, defined and assembled by hand. This approach to AI essentially derives from a line of philosophical thought running from Descartes through Leibnitz, Frege, and Russell.

In the late 1980s and 1990s, however, as a result of the inherent difficulty of this line of research, and of the limited progress that has been made, this approach to AI has been challenged by two alternative methodologies: neural networks, and statistical analysis of large corpora (see *Neural Networks and Related Statistical Latent Variable Models* and *Statistical Analysis of Large Corpora*).

Neural network research (Haykin 1999) attempts to construct computational models that resemble, with greater or lesser fidelity, the structure of the brain as currently understood. These models vary greatly in detail, but, generally speaking, they consist of a fixed structure of very many, interconnected, primitive computational parts. Each such part models a single neuron. The behavior of a part is controlled by a small set of numerical parameters. In *task execution* mode, the network begins with an input state, representing the task to be accomplished, and evolves toward an output state, representing the desired behavior. In *learning* mode, the network tries to adjust the values of the control parameters so as to achieve optimal behavior during task execution.

The permanent knowledge of such a network is thus encoded in the value of the parameters that govern the neurons. This encoding is typically opaque and distributed; that is, it is not possible to say how the parameter values encode the knowledge or even to localize the knowledge to a particular part of the network. Rather, the knowledge is an emergent property of the network as a whole.

The strength of a neural network lies in the strong theory of learning that applies. Neural networks are also generally very robust; they can often survive well if some of the training data is incorrect, or if some of the components are unreliable. Scope and expressivity are characteristically much more limited than in conventional knowledge-based systems: a neural network generally carries out a small set of narrowly focused tasks.

The difficulties in constructing a neural network are very difficult from those of constructing a symbolic representation. It is not necessary, or indeed possible, to handcraft the representation to fit one's understanding of the domain. However, it is often necessary to tune the structure of the network and the learning algorithms rather carefully to fit the application. This tuning is much more an art than a science.

The statistical approach to AI involves taking very large corpora of data, and analyzing them in great depth using statistical techniques. These statistics can then be used to guide new tasks. The resulting data, as compared to the knowledge-based approach, are extremely shallow in terms of their semantic content, since the categories extracted must be easily derived from the data, but they can be immensely detailed and precise in terms of statistical relations. Moreover, techniques—such as maximum entropy analysis—exist that allow a collection of statistical indicators, each individually quite weak, to be combined effectively into strong collective evidence.

From the point of view of knowledge representation, the most interesting data corpora are online libraries of text. Libraries of pure text exist online containing billions of words; libraries of extensively annotated texts exist containing hundreds of thousands to millions of words, depending on the type of annotation. Now, in 2001, statistical methods of natural language analysis are, in general, comparable in quality to carefully hand-crafted natural language analyzers; however, they can be created for a new language or a new domain at a small fraction of the cost in human labor (Charniak 1993, Armstrong 1994).

Viewed as a knowledge representation, natural language text is, of course, unparalleled in terms of scope, expressivity, and ease of construction. (A few domains of human knowledge, interestingly, lie largely outside the scope of natural language; natural language is notoriously poor at expressing spatial knowledge or mechanical knowledge, such as how to tie a shoelace.) It is exceptionally weak in terms of precision and support for inference. Natural language constructs are constantly, massively ambiguous, and the disambiguation that human understanders carry out automatically and effortless depends, in many cases, on precisely the kinds of world knowledge whose representation has been found to be so difficult.

See also: Decision Support Systems; Expert Systems in Medicine; Intelligence: Historical and Conceptual Perspectives; Intelligence: History of the Concept; Knowledge (Explicit and Implicit): Philosophical Aspects; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Logics for Knowledge Representation; Semantic Knowledge: Neural Basis of

Bibliography

Allen J 1983 Maintaining knowledge about temporal intervals. *Communications of the ACM* **26**(11): 832–43

Armstrong S (ed.) 1994 *Using Large Corpora*. MIT Press, Cambridge, MA

Bacchus F 1990 *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach to Probabilities*. MIT Press, Cambridge, MA

Borgida A, Brachman R, McGuiness D, Resnick L 1989 CLASSIC: A Structural data model for objects. In: *Proceedings of SIGMOD '89*. Portland, OR, pp. 59–67

Charniak E 1993 *Statistical Language Learning*. MIT Press, Cambridge, MA

Davis E 1990 *Representations of Commonsense Knowledge*. Morgan Kaufmann, San Mateo, CA

Davis E 1998 The naive physics perplex. *Artificial Intelligence Magazine* **19**: 51–79

Genesereth M, Nilsson N 1987 *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann, San Mateo, CA

Glasgow J, Narayanan N, Chandrasekaran B 1995 *Diagrammatic Reasoning*. MIT Press, Cambridge, MA
Hayes P 1979 The naive physics manifesto. In: Michie D (ed.) *Expert Systems in the Microelectronic Age*. Edinburgh University Press, Edinburgh, UK, pp.
Haykin S 1999 *Neural Networks: A Comprehensive Foundation*. Prentice Hall, New York
Lenat D 1995 CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, **38**(11): 33–8
Mates B 1972 *Elementary Logic*. Oxford University Press, Oxford
McCarthy J 1968 Programs with common sense. In: Minsky M (ed.) *Semantic Information Processing*. MIT Press, Cambridge, MA, pp. 403–18
McCarthy J, Hayes P 1969 Some philosophical problems from the standpoint of artificial intelligence. In: Metzler B, Michie D (eds.) *Machine Learning 4*. Edinburgh University Press, Edinburgh, UK, pp.
Minsky M 1975 A framework for representing knowledge. In: Winston P (ed.) *The Psychology of Computer Vision*. McGraw-Hill, New York, pp.
Mitchell T 1997 *Machine Learning*. McGraw-Hill, New York
Newell A, Simon H 1976 Computer science as empirical inquiry: Symbols and search. *Communications of the ACM* **19**: 113–26
Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA
Reichgelt H 1991 *Knowledge Representation: An AI Perspective*. Ablex Publishing, Norwood, NJ
Selman B, Levesque H, Mitchell D 1992 A new method for solving hard satisfiability problems. *Proceedings of AAAI-92*, pp. 440–46
Sowa J 2000 *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole, Pacific Grove, CA
Stefik M 1995 *Introduction to Knowledge Systems*. Morgan Kaufmann, San Mateo, CA
Turner R 1984 *Logics for Artificial Intelligence*. Wiley, New York
van Benthem J 1983 *The Logic of Time*. Reidel, Dordrecht, Netherlands
Woods W 1975 What's in a link: foundations for semantic networks. In: Bobrow D, Collins A (eds.) *Representation and Understanding: Studies in Cognitive Science*. Academic Press, New York, pp. 35–82
Zadeh L 1987 Commonsense and fuzzy logic. In: Cercone N, McCalla G (eds.) *The Knowledge Frontier: Essays in the Representation of Knowledge*. Springer-Verlag, New York, pp. 103–36

E. Davis

Knowledge Societies

The analysis of contemporary societal change toward information society or knowledge society can emphasize two different sets of variables: technological innovation and institutional transformation. If it is theoretically controversial as to how both sets may be causally interrelated—ranging from technological determinism to the social shaping of science and technology—empirical oberservation certainly benefits from focusing on both: the impact of technological change on the organizational and cultural institutions of society as well as on the enormous monetary and cultural investments of corporate and individual agencies in developing and using new knowledge. The terms 'information society' and 'knowledge society,' though overlapping in many respects, may help to distinguish these perspectives. The first relates to the pervasive influence of computer-based networks and thereby the globalization of information, communication, and control technologies. It assumes that collecting, copying, storing, transmitting, incorporating, manipulating, simulating, and managing information about every aspect of collective and individual life penetrate and shape all these aspects. The concept of 'knowledge society' emphasizes in a complementary way the problems and strategies of making sense of information. The knowledge work of researchers, experts, analysts, and users depends much on and contributes to information—but what makes this work important is its use of theories, models, scenarios, evaluation criteria, decision strategies, experimental designs, and implicit experience in order to establish bits and pieces of orientation and certainty. Orientation and certainty are needed by individual and corporate actors who face the risks of decision-making in an environment that has become complex, opaque, and rapidly changing precisely because of the infinite amount of information it contains. The instantaneous availability of ever increasing resources of information and information about information vastly increases the need to know something about its value and relevance (Earl 1996).

When the term knowledge society was coined, the path of societal development was conceived much more clearly. Fritz Machlup (1962) was the first to estimate the contribution of knowledge work to the economy; Peter Drucker (1969) developed guidelines for mastering the discontinuity brought about by information technology and knowledge work; Daniel Bell (1971, 1973) predicted the stage of postindustrial society, in which theoretical knowledge would be the new 'axial principle' of society and strategic resource of progress. A parallel line of reasoning was developed by reform Marxists in the so-called Richta report (1969) and the Japanese 'Plan for an Information Society' was presented to its government in 1972 (Masuda 1981). These followed a period of improving the empirical data base (Porat 1977), pondering the validity of concepts (Lyon 1988), and comparing different countries (Edelstein 1978).

Scholarly and public attention became steadily intensified and extended the general themes of the societal centrality of knowledge to a broad variety of fields of investigation: the reconstruction of class structure in knowledge society (Schiller 1981, Schiller and Alexandre 1992), its relation to postmodernism (Lyotard 1984, Poster 1990), globalization and world

society (Castells 1997), the spread of expert culture (Stehr and Ericson 1992, Jasanoff 1990), the relevance of intellectual capital and its impact on the reorganization of business (Beninger 1986, Stewart 1997), and the role of science in knowledge society (Gibbons 1994). Usually the term is used generically, denoting a comprehensive societal change, but with respect to the uneven distribution of conditions it may be analytically preferable to split the generic term and speak of various patterns of knowledge societies (Stehr 1994).

1. Defining the Concept of Knowledge Societies

The reality of knowledge societies might have pleased Plato as putting into practice his ideal of philosophy governing society. It could also be taken as fulfilling the predictions of Condorcet (1743–1794) and Comte (1789–1857) according to which knowledge about the (laws of the) development of society would be put in control of shaping its structure. Or else, the machinery of information and communication technology could be taken to have realized the ideas of integrative and long-range planning developed under the auspices of the OECD in the postwar period (Jantsch 1968). None of these visions has even approximately come true. Neither wisdom, nor generally valid law-like insights, nor integrative and comprehensive scenarios are relevant, but knowledge is expected to be helpful in singling out the risks of decision and action. The permanent modernization of societies leaves all actors in successive states of uncertainty, insecurity, and ambiguity. The more information resources are available for interpreting past experiences, signs of change, and options to adapt, the less likely it is that information itself condenses into reliable knowledge and patterns of orientation. Rather, actors are left alone with the choice between doing anything or nothing—except where there is an agency promising to be specialized in knowing how to read and use information.

Whether, of course, such a promise is reliable, may again be a matter of information which can perhaps be obtained by another agency specialized in watching closely the operations of other knowledge interpreters. The scope of work performed by knowledge agencies covers a broad variety of activities. In many cases knowledge operations become standardized and implemented into machinery and computer-aided systems. To this extent they operate as reliably as other technological tools. Still, there seems to exist a permanent tension between modeling and routinizing processes and structures, and their appropriate use in organizational and individual action. Changing conditions and situational specifities ensure that knowledge tools 'are not simply applied but processes to be developed' (Castells 1997, p. 32). The tension between knowledge as application of certified truth and knowledge as risky investigation of what has become and is going to become uncertain can be discussed on three different levels: industrial routinization of knowledge (Sect. 2.1); professional expertise (Sect. 2.2); and learning by designing tests and experiments (Sect 2.3).

2. The Production, Distribution, and Consumption of Knowledge

2.1 Routinization and Management of Knowledge Work

Knowledge work will always include certain features of individual creativitiy and imagination. But its integration into the institutional framework of society presupposes its transformation from craftsmanship to industry. Though the term 'industrialization of knowledge' may sound anachronistic—at least insofar as knowledge society is conceived to be postindustrial (Bell 1973)—some characteristic changes of knowledge work are similar to the transition from craftsmanship to industrial production. One of the essential nineteenth century breakthroughs of the industrial revolution was the development of a branch specialized in producing machine tools (Rosenberg 1972). The general tendency was to dis-integrate the manufacturing of all parts of a product and to specialize in the production of standardized parts of multiple use and high precision. Similarly, the industrialization of knowledge work proceeds by dis-integrating the craftsman style of manufacturing knowledge—a style that is best represented in the tradition of scientific research groups which start by collecting data from events, arranging them in information packages, and providing interpretation by theoretical models. Contemporary knowledge work transgresses the limits of this tradition. This does not imply a fundamental difference between science and knowledge work but assumes science and its institutions to become part of a new production and distribution system of knowledge. The most important step to be taken in this respect is to understand knowledge as something that can be designed, assembled, shaped, combined, controlled, and adjusted by means of knowledge instruments and that can be certified, licensed, and sold (Fuller 1992). The paradigm of such kinds of knowledge is the algorithm, the software core of all intelligent products and procedures. Algorithms are objectified know-how structures that instruct men, machines, and themselves what to do under certain conditions. In this sense, the history of computer and software industry resembles many traits of the machine tool and other hardware industries. Several authors have rightly stressed the differences between 'Fordist' mass production and 'post-Fordism,' but they overlook the structural similarity and material continuity between industrial and knowledge economies. Viewing know-

ledge as a produceable thing allows one to attach to it the juridical label of being a property. Where the boundary should be drawn between knowledge as either a collective and public or a corporate and proprietary good is highly controversial (Weil and Snapper 1989). Traditional philosophical understandings of knowledge, as well as the received view of professional work as applying disciplinary academic knowledge to single problems, are no longer applicable to most variants of knowledge work. The concept now refers to activities that can be characterized by importing, improving, combining, assessing, and testing knowledge by means of informatized machines, instruments, toolsets, networks, and media in order to produce expertise (Reich 1991). Having stressed the industrial aspects of knowledge work as they become manifest in material and immaterial knowledge machinery, it is only fair to point out the differences. The dynamics of knowledge work tend to create individualistic cognitive styles with respect to methods, problem selection and definition, and communication with partners and clients. It is hence by no means certain that different approaches will produce identical results. This, in turn, provokes a fundamental conflict for the management of knowledge work. On the one hand management must try to increase complexity by bringing together different approaches to defining problems, framing options, and designing solutions. On the other hand it is important to quickly reduce complexity by generating trust in prototypes, which can be tested and improved with selected users (Leonard-Barton 1995). If the main task of knowledge management is the organization of innovation, it does not leave untouched organizational design (Myers 1996). Bureaucracy needs to be replaced by knowledge links, flexible networks, and strategic alliances. Organizations not only have to support the idiosyncratic aspects of knowledge work, but also to turn into learning institutions, which work with visions, experiments, and assessments in order to permanently define the environment to which they want to adapt (Argyris and Schön 1996).

2.2 Facing the Unknown

There is a complementary aspect to the economic and organizational characteristics of knowledge society. Certainly, knowledge would not have become a predominant resource if it could not become transformed in marketable goods and services. It is not less true, however, that every state of knowledge opens up even more notions of what is not known. The more society relies on knowledge, the more it navigates into a dynamics of change. To understand, describe, predict, construct, and control this change becomes illusory—simply because of lack of knowledge.

This is exactly the reason why knowledge experts (Jasanoff 1990) and symbolic analysts (Reich 1991) are needed and expected to create islands of certainty to allow orientation and planning. Demand for knowledge expertise may differ in several respects: by its purposes, ranging from understanding to predicting and constructing; its form of communication, which may be more propositional or more practically embedded; and its scope of application. It may be a rather general theoretical belief system, a generalized rule, a simulation model, a special case experience, or just the solution of an urgent problem. Wherever such expertise is invested, the expectation is to obtain some certainty. But certainty does not exist only in degrees of quality ranging from absolutely valid, to reliable under certain conditions, to better than nothing. It is also exposed to a second-order knowledge about knowledge, or expertise on expertise. Every form of expertise depends on choices with respect to heuristics, methods, or data bases and could have been formulated differently. Where there are choices there are risks. But if, in turn, risks can only be reduced by additional expertise, then the net result may be a steady loss of certainty in society. Even if infinite amounts of information were accessible and could be processed by advanced knowledge technologies, the well-known problem of acting under bounded rationality would not be solved but enhanced. The constant examination and reformulation of social practices in the light of incoming information about those very practices is also conceived in Giddens' model of 'reflexive modernization' (Giddens 1990).

2.3 Knowledge and Science

It is important to relate the characteristics of knowledge society to those of science in society. While on the one hand large-scale expertise can be interpreted as spreading the methodical and theoretical potential of science to highly specialized problems and as a consequence of a job-seeking academically trained labor force (Stehr 1992), the differences between expertise and research are obvious. Research is performed in an institutionally (laboratories) and cognitively (theories) separated world, in which failures and errors are highly welcome. Everything can hypothetically and experimentally be pushed to the extreme and only in a few cases are legal and social norms violated. Society is only affected indirectly by research, namely by transforming research results into technologies or ideologies. Even if this view meets a convenient self-description of the scientific community rather than historical facts, expertise could rarely be sold as a piece of hypothetical reasoning to which those who apply it provide the testing field. Still, it is instructive to underpin the intersections of expertise and research.

While many branches of science tend to derive expert knowledge directly by extrapolating from isolated theoretical worlds to the real world (of health, education, international affairs, social structure, de-

velopment, etc.), expertise tends to not only apply certified knowledge but to do research. Consequently, knowledge society can be interpreted as being exposed to hypothetical assumptions and experiments designed by experts and analysts. While expertise effectively reduces the complexity of the real world for an observer, it inevitably contributes to increase the risks of those who decide to rely on it.

3. Critical Aspects of Knowledge Society

This ambivalent status of expert knowledge between disinterested scientific neutrality and action-oriented advice has given rise to criticism of expert culture and its 'politics of knowledge' (Cozzens and Woodhouse 1994). Another problem to be addressed in the future is the deficient quality management of expert knowledge and the inadaequacy of licensing and accreditation mechanisms. A profound cognitive aspect of criticism relates to the loss of personal experience and judgement and its substitution by mediated images, simulations, screens, and spectacles (Morley and Robins 1995) along with a new potential for ideological manipulation (Schiller 1981). The intra- as well as international boundaries between those who have or do not have technical and intellectual access to information (information rich vs. information poor) have given rise to neo-Marxian class analysis of knowledge societies (Schiller 1992). These and other controversial aspects (see Webster 1995) indicate the deep structure and far-reaching consequences of the ongoing transformation toward knowledge society.

See also: Expert Systems in Cognitive Science; Finalization in Science; Humboldt, Wilhelm von (1767–1835); Information Society; Information Society, Geography of; Information Technology; Knowledge, Sociology of; Risk, Sociology and Politics of; Science and Industry; Science and Technology Studies: Experts and Expertise; Scientific Knowledge, Sociology of; Sociology, Epistemology of;

Bibliography

Argyris C, Schön D 1996 *Organizational Learning II. Theory, Method, Practice*. Addison-Wesley, Reading, MA

Bell D 1971 The post-industrial society: The evolution of an idea. *Survey* **17**(2): 102–68

Bell D 1973 *The Coming of Post-industrial Society*. Basic Books, New York

Beninger J R 1986 *The Control Revolution. Technological and Economic Origins of the Information Society*. Harvard University Press, Cambridge, MA

Castells M 1997 *The Rise of the Network Society*. Blackwell, Oxford, UK

Cozzens S, Woodhouse E 1994 Science, government, and the politics of knowledge. In: Jasanoff S, Markle G, Petersen J, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA, pp. 533–53

Drucker P 1969 *The Age of Discontinuity. Guidelines for our Changing Society*. Harper and Row, New York

Earl M 1996 *Information Management. The Organizational Dimension*. Oxford University Press, Oxford, UK

Edelstein A S 1978 *Information Societies: Comparing the Japanese and American Experiences*. International Communication Center, Seattle, WA

Fuller S 1992 Knowledge as product and property. In: Stehr N, Ericson R V (eds.) *The Culture and Power of Knowledge. Inquiries into Contemporary Societies*. De Gruyter, Berlin, Germany, pp. 157–90

Gibbons M 1994 *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. Sage, London

Giddens A 1990 *The Consequences of Modernity*. Stanford University Press, Stanford, CA

Jantsch E 1968 Perspectives of planning. *Proc. OECD Working Symp. Long-Range Forecasting and Planning*. Bellagio, Italy, 27th Oct.–2nd Nov. 1968. Organisation for Economic Co-operation and Development, Paris

Jasanoff S 1990 *The Fifth Branch: Science Advisers as Policy-makers*. Harvard University Press, Cambridge, MA

Leonard-Barton D 1995 *Wellsprings of Knowledge*. Harvard Business School Press, Boston

Lyon D 1988 *The Information Society: Issues and Illusions*. Polity Press, Cambridge, UK

Lyotard F 1984 *The Postmodern Condition*. University of Minnesota Press, Minneapolis, MN

Machlup F 1962 *The Production and Distribution of Knowledge*. Princeton University Press, Princeton, NJ

Masuda Y 1981 *The Information Society as Post-Industrial Society*. World Future Society, Bethesda, MD

Morley D, Robins K 1995 *Spaces of Identity. Global Media, Electronic Landscapes and Cultural Boundaries*. Routledge, London

Myers P S 1996 *Knowledge Management and Organizational Design*. Butterworth-Heinemann, Newton, MA

Porat M U 1977 *The Information Economy*. Office of Telecommunications, Washington, DC

Poster M 1990 *The Mode of Information: Poststructuralism and Social Context*. Polity Press, Cambridge, UK

Reich R 1991 *The Work of Nations*. Knopf, New York

Richta R 1969 *Civilization at the Crossroads: Social and Human Implications of the Scientific and Technological Revolution*. International Arts and Sciences Press, White Plains, NJ

Rosenberg N 1972 *Technology and American Economic Growth*. Harper and Row, New York

Schiller H 1981 *Who Knows: Information in the Age of the Fortune 500*. Ablex, Norwood, NJ

Schiller H, Alexandre L 1992 *The Ideology of International Communications*. Institute for Media Analysis, New York

Stehr N 1992 Experts, counsellors and advisers. In: Stehr N, Ericson R V (eds.) 1992 *The Culture and Power of Knowledge. Inquiries into Contemporary Societies*. de Gruyter, Berlin

Stehr N 1994 *Arbeit, Eigentum und Wissen: zur Theorie von Wissensgesellschaften*. Suhrkamp, Frankfurt/Main, Germany

Stehr N, Ericson R V 1992 *The Culture and Power of Knowledge. Inquiries into Contemporary Societies*. De Gruyter, Berlin, Germany

Stewart T A 1997 *Intellectual Capital. The New Wealth of Organizations*. Nicholas Brealey, London

Webster F 1995 *Theories of the Information Society*. Routledge, London

Weil V, Snapper J (eds.) 1989 *Owning Scientific and Technical Information: Value and Ethical Issues*. Rutgers University Press, New Brunswick, NJ

W. Krohn

Knowledge, Sociology of

The sociology of knowledge is a summative term for an assortment of theoretical statements and empirical studies dealing, in diverse ways, with the relationship between various processes and products of human cognition and other sociocultural factors. On the organizational map of the discipline of sociology in Europe and in the USA, the sociology of knowledge has generally occupied a marginal position, its warrant and scope sufficiently contested 'that the history of the subdiscipline … has been the history of its various definitions' (Berger and Luckmann 1966, p. 4). Focusing on manifestations of knowledge that include ideas, ideologies, scientific theories, religious and political doctrines, moral, ethical and philosophical beliefs, world views, mental categories, cultural and organizational discourses, languages, social norms, and the forms and practices of everyday knowing, sociologists of knowledge (whether self-identified or retrospectively claimed for the area by other contributors) have concerned themselves with the sociocultural consequences of this range of intellectual products and processes and, particularly, with their origins and transformations—developments that have been linked to sociohistorical contexts, material and cultural conditions, modes of production, power relationships in various forms, institutional arrangements, internal organizational processes, and a vast array of different social positions and interests. Sociologically relevant reflection on this set of issues may be periodized in four broad stages.

1. *Antecedents of the Sociology of Knowledge*

While observations about the social foundations and consequences of various cognitive products and processes have a complex history that extends back to the Greeks and forward through Renaissance figures such as Machiavelli and Francis Bacon, the modern roots of the sociology of knowledge are generally dated from the French Enlightenment of the eighteenth century, when thinkers like Helvetius, Holbach, and Voltaire turned attention to the social bases and effects of religious—as contrasted with scientific—conceptions, while Montesquieu, Rousseau, and Condorcet adumbrated the broader thesis that human consciousness is socially determined. Working in this tradition, Destutt de Tracy advocated a 'science of ideas' that would trace human thoughts to their sensory and material conditions, ultimately making possible the construction of a new morality and a perfected political order. For this science, in 1796 he coined the term 'ideology'—a neologism that his critics swiftly converted into a pejorative word for doctrinaire dreams and illusions. (For fuller treatment of this background, see Eagleton 1991, Kennedy 1978, Larrain 1979.)

In the early nineteenth century, Auguste Comte declared the evolution of modes of human thought the main topic of the new discipline of sociology, erecting the field on the Law of Three Stages, viz., the assertion 'that each of our leading conceptions—each branch of our knowledge—passes successively through three different theoretical conditions: the theological or fictitious; the metaphysical, or abstract; and the scientific, or positive' ([1830] 1975, p. 71). Substituting French positivism for German idealism, G. W. F. Hegel ([1807] 1967) likewise presented the history of society as a sequence of intellectual stages—as successively developing forms of human consciousness, dialectically progressing to a condition of absolute (genuinely scientific) knowledge.

From the mid nineteenth to the early twentieth century, these early modern anticipations of the sociology of the knowledge were elaborated in many directions, particularly in connection with the development of 'classical' sociological theory. Rejecting Hegelian idealism for historical materialism, Karl Marx conceptualized the connections between human thought and social conditions in several ways. In his most encompassing formulation, he asserted that dialectically changing 'modes of production [formed] the basis of all history' and that the 'origins and growth [of] all the different theoretical products and forms of consciousness, religion, philosophy, ethics, etc., etc.'—the idealistic superstructure—were explainable 'from that basis' (Marx and Engels [1846] 1970, pp. 57–8). Taking the social relations of production as 'the realm of real history,' Marx's early work cast this superstructure as the derivative 'realm of ideology,' using the latter term in the pejorative sense of 'illusions' and distorted 'inversions' of reality that 'veil the class struggle' and thus promote existing patterns of class domination (Marx and Engels [1846] 1970, pp. 47, 119; Marx [1852] 1963, pp. 46–7). Relatedly, Marx linked particular ideas to specific 'material interests' and held that 'the class which has the means of material production at its disposal has control … over the means of mental production,' so that 'the ideas of the ruling class are in every epoch the ruling ideas,' as these are developed by 'conceptive ideologists' perfecting the illusions that benefit their class (Marx and Engels [1846] 1970, pp. 64, 65, 99). Marx's later work, in contrast, presented ideology less as a superstructural inversion of reality than as a reflection of inversions already embedded in particular

modes of production. Characterizing capitalism itself, for example, as an 'enchanted and perverted world,' Marx held that, whatever their class interests, conceptualizers under capitalism generally 'remain more or less in the grip of the world of illusion,' a world that 'corresponds' with 'the forms of illusion in which they move about and find their daily occupations' ([1894] 1967, pp. 827–31; for discussion, see Eagleton 1991, Mepham 1979). Subsequently, many of Marx's followers overlaid this negative view of the social functions of ideology with a broader 'sense of ideology as the mental forms within which men and women fight out their social conflicts' (Eagleton 1991, p. 90). This is a tendency evident in the work of Bernstein, Lenin, and Lukacs, and in Hoare and Smith's conception of ideologies as forces that, under sociohistorical conditions that he sought to specify, can 'organize human masses, and create the terrain on which men move, acquire consciousness of their positions, struggle, etc.' ([1948–51] 1971, p. 377).

Coming from the French tradition, Emile Durkheim approached knowledge from a different angle. Extending his concern with the social bases (and functions) of such 'collective representations' as religion and morals to the 'extra-logical origin' of 'logical thought,' Durkheim argued that the 'mental operations' by which members of preliterate societies 'classify … things, events, and facts about the world into kinds and species [are] modeled on [their] form of social organization,' i.e., that 'the first logical categories were social categories; the first classes of things were classes of men, into which these things were integrated' (Durkheim and Mauss [1903] 1963, pp. 3, 4, 8, 82). Generalizing the point, Durkheim subsequently proposed that the form and the content of all the 'categories of human thought'—of the most basic 'conception[s] of time, space, cause, number, and so on,' even in their modern scientific incarnations—'depend on the way in which the collectivity is organized, upon its morphology, its religious, moral, and economic institutions, and so on' ([1912] 1995, pp.14–16). Controversial to this day, the claim proved widely influential, inspiring work by Halbwachs (1925), Granet (1934), and others.

2. *Emergence of the Sociology of Knowledge as a Specialty Area (1920 to c. 1965)*

As a specialized theory-and-research area with its own name, the sociology of knowledge originated in Weimar Germany when Max Scheler and Karl Mannheim, reacting (*inter alia*) to Marxian theories of ideology and to the cultural crisis of their own time, advocated the creation of *Wissenssoziologie*. Scheler did so as an extension of his interests in philosophy. Seeking to recognize 'the social–historical development of [even] the highest types of human knowledge' but to 'escape [from the] philosophical relativism' that had led contemporaries to deny the truth-claims of ideas 'conditioned by a historical and sociological standpoint,' he challenged Marx's thesis that one 'can "explain" th[e] ideal world from the world of real history' by asserting that 'the history of real factors'—understood as a sequence in which kinship, political, and finally economic factors successively hold centrality—'does not at all determine the positive meaning-content of works of mind, though it does hinder, release, delay, or hasten [their] development,' opening and closing 'the sluices of the spiritual stream' ([1924] 1980, pp. 41–2, 53–4; see also Frisby 1983).

Mannheim took another tack. Excepting mathematics and portions of the 'exact sciences' from his analysis, he sought to problematize the 'social conditioning' of the forms and contents of 'every product of thought'—'perceptions and interpretations,' moral concepts of 'duty, transgression, and sin,' political 'opinions, statements, [and] propositions,' and the total 'modes of thought' of different epochs, cultures, and social groups. According to Mannheim, such thought-products stand in 'multiple interconnections' with 'the structure of historical reality,' with 'the social position of given groups'—a phrase primarily denoting class position, but also including 'generations, status groups, sects, occupational groups, schools, etc.'—and with 'the historical–social situation of the intellectually active and responsible members of the group' ([1931] 1936a, pp. 56–7, 81–2, 86, [1926] 1936c, pp. 264–5, 276). In putting forth this view, Mannheim wished to call forth a wide-ranging academic field that would subject 'the relationship between thought and existence' to systematic 'empirical research' ([1931] 1936a, p. 69, [1929] 1936c, p. 264; see also 1936b), and occasionally he engaged in such research of himself, most notably in his study of German conservative thought ([1925] 1986). But Mannheim set for the sociology of knowledge an even larger agenda, borne from his belief that examination of 'the relationship between the intellectual positions held and the social position occupied' would necessarily reveal that 'every point of view'—Marxism included—'is particular to a social situation.' From this it followed, for Mannheim, that 'the thought of all parties in all epochs is of an ideological character,' where he took 'ideological' to mean 'narrow' and 'one-sided'—properties with important epistemological and political implications. By exposing the one-sidedness of all perspectives, the sociology of knowledge would lay the basis for 'a new type of objectivity,' which was the dynamic synthesis of different viewpoints, and would bring political rivals into dialogue, checking their pernicious tendency to unmask each another's ideas as mere illusions (1936a, pp. 77–85, [1929] 1936b, p. 5; for discussion, see Kettler and Meja 1995).

Mannheim's statements drew broad attention to the sociology of knowledge in Europe and the USA and set the new field on an unsteady course. In Europe, his epistemological and political views became the focus

of extensive critical discussion and made the field, for decades, largely a site of meta-theoretical controversy (Frisby 1983, pp. 174–224). In the USA, where engagement with sociology-of-knowledge type issues had long been underway (Fuhrman 1980), Mannheim's call for empirical studies had greater resonance and fueled efforts to constitute the sociology of knowledge as an specialized subdiscipline dedicated to 'fruitful empirical research,' rather than to debate over 'philosophical' questions of relativism, objectivity, and the like (quoting Wirth 1936, pp. xxi, xxvii; see also Kettler and Meja 1995, pp. 193–246).

Particularly significant here were Robert Merton's writings in the early 1940s, which sought to move the sociology of knowledge from 'speculative insight' to 'rigorous test,' by purging Mannheim's ideas of 'their epistemological impedimenta' and associating the field with a defined 'list of problems [open to] empirical investigation' ([1945] 1968a, p. 562, [1941] 1968b, p. 542). To this end, Merton elaborated a 'paradigm for the sociology of knowledge,' in which he replaced Mannheim's 'amorphous category of knowledge' with a systematic inventory of different types and aspects of 'mental productions' and then differentiated the many possible sociocultural or 'existential bases' for human thought-products ([1941] 1968a, p. 552, 1968b, pp. 514–5). Most distinctively, Merton gave prominence to the question (which he faulted his predecessors for neglecting) '*how* are mental production related to the existential basis?'—distinguishing, in reply, a wide range of possible modes of relationship, analysis of which 'promise[d] to take research in the sociology of knowledge from the plane of general imputation to that of testable empirical inquiry' ([1945] 1968b, pp. 515, 537).

This promise proved difficult to realize, however. In the course of the next twenty years, students of Merton occasionally took up the challenge—Gouldner (1965) with research on Plato's social theory as a response to the social tensions present in ancient Greece, Coser (1965) with work on the general societal and institutional conditions that favored the development of Western intellectuals. These studies joined writings by Mills ([1943] 1964) on the social roots of US pragmatism (see also Mills 1963), by Znaniecki (1940) and Shils ([1955–68] 1972) on intellectuals, by members of the Frankfurt school on the ideology of capitalist societies (see especially Marcuse 1964), and by a scattering of other scholars on a miscellany of topics (see the collection by Curtis and Petras 1970) to steer the sociology of knowledge in more empirical directions. But these remained sporadic efforts, intermixed with works offering programmatic definitions of the sociology of knowledge and further discussion of Mannheim's more philosophical concerns (see, e.g., DeGre [1939–43] 1985, Stark 1958, Wolff [1943–65] 1983). Such works continued to distance empirically oriented sociologists from the field, even as various sociology-of-knowledge type questions resurfaced in the sociology of religion, the sociology of education, the sociology of occupations and professions, the sociology of public opinion and mass communication, and the sociology of science (Coser 1968, p. 432), as well as in research by historians of science and historians of ideas (see especially Ringer 1969).

The era that began with the writings of Scheler and Mannheim and extended to the mid-1960s remains, nonetheless, the most active period in the sociology of knowledge as a specialty area, and it stamped the area with a set of basic assumptions that were common to most work in the field at the time. These included, first, the assumption of a division between sociocultural factors, on the one side, and the substantive contents of thought-products and processes, on the other, where the former factors were 'external' to the realm of knowledge, while the latter were 'internal'; second, the supposition that the 'internal' factors were easily interpretable by the sociological investigator, their meanings more-or-less transparent in the statements where they were contained; and, third, the belief that sociological explanations of knowledge chiefly involved appeals to 'external' factors at the macro-level, i.e., to societal-wide economic, political, and cultural developments, or combinations thereof (Camic and Gross 2000; see also Kuklick 1983). Traceable to the earliest beginnings of the sociology of knowledge, these assumptions as of yet met little challenge.

3. *The Specialty in Retreat (c. 1965 to c. 1985)*

For a variety of reasons, largely exogenous to the area itself, work in the sociology of knowledge as such waned significantly from the mid 1960s to the early 1980s. This was due partly to a continuation of the tendency, already evident in the pervious period, for topics that fell to the area to undergo partial incorporation into other sociological subfields. Expanding interest in the sociology of science proved especially consequential here, for in the process of focusing on scientists' career patterns and on the organizational and normative conditions that sustain scientific institutions, sociologists of science effectively denied—despite contrary suggestions in the influential work of Kuhn (1962)—that there was even much possibility for a 'sociology of the conceptual and theoretical contents of science' and related domains of knowledge (Ben-David 1971, p. 14). In the face of this challenge, Berger and Luckmann attempted to redirect the sociology of knowledge to the study of 'what is taken for granted as "knowledge" [by] the man in the street' (1966, p. 3); but it was not long before this topic was itself transplanted, in part into ethnomethodology, another subfield then on the rise (Maynard and Clayman 1991).

Compounding these developments within specialty areas were broader trends in the social sciences during this period. Most far-reaching among these was a

growing commitment to quantitative research, based increasingly on types of survey data that seemed remote from the traditional concerns of sociologists of knowledge. This trend, to be sure, coexisted with a heightening of interest in theory and in historical scholarship, but the Marxist renaissance that was then underway generally pulled this interest towards questions about the dynamics of modes of production and state apparatuses and toward debates about the role and concept of social class (and, subsequently, about gender and race). While new theories of ideology did emerge (see, in particular Althusser 1969, 1971, Therborn 1980), they did so, almost without exception, as appendages to work on class and state structures; as such, their effect was more to divert attention from, than to turn it to, examination of the sources and consequences of cognitive products and processes in their own right. In this context, little new work in the sociology of knowledge came forth, aside from reinterpretations of figures from earlier periods (see Simonds 1978) and various editions of their original writings.

4. The Contemporary Revival (c. 1985 to the Present)

Just as exogenous developments during the two preceding decades imperiled the sociology of knowledge, intellectual trends since the mid 1980s have begun to resuscitate the field. Most generally these have included: burgeoning interest in the interdisciplinary field of cultural studies, stimulated by a declining focus on Marxist structuralism and the growing influence of works on discourse and linguistic theory, post-structuralism, and post-modernism (for discussion, see McCarthy 1996); the effort within Marxism itself to transform general claims about ideology into empirical studies of discursive formations and practices (Laclau and Mouffe 1985, Thompson 1990); the move within feminist scholarship to problematize knowledge and the standpoint and social position of knowledge producers (Harding 1991, Smith 1987); the increased production of sociologically-informed studies in the history of science and intellectual history (see Shapin 1982, Toews 1987); and the redirection of the sociology of science toward the examination of the contents of scientific knowledge and the social processes shaping them (see Collins 1983). Reinforcing these broad trends has been an expansion of interest in European social theorists whose works give considerable attention to knowledge-production processes, particularly Michel Foucault and his archeological and genealogical studies of the reciprocal constitution of various knowledge systems and webs of power relations (see Prado 1995), and Pierre Bourdieu and his writings on the dynamics of intellectual and scientific fields (see Ringer 1990).

From this confluence of exogenous developments has come a renewed concern with the sociology of knowledge, a concern that is currently pulling the field in the direction of two different approaches. The first of these is the broad-constructionist approach, exemplified in programmatic statements by McCarthy (1996), Stehr and Meja (1984), and Swidler and Arditi (1994), and in studies by Griswold (1986), Lamont (1987) and Wuthnow (1989). Pointing the sociology of knowledge outward toward the contemporary sociology of culture rather than toward the study of knowledge specialists, this perspective treats 'political and religious ideologies as well as science and everyday life, cultural and organizational discourses along with formal and informal types of knowledge' (Swidler and Arditi 1994, p. 306); and while amenable to a middle-level or institutional analysis of the social sources and consequences of these various forms of knowledge, it often retains the underlying suppositions of the 1920–65 period, viz. assumptions about the division between external (sociocultural) and internal (thought-content) factors, about the transparency of internal meanings to the sociological investigator, and about the explanatory importance of macro-level external factors. Moreover, while strongly committed to empirical research, this approach has also favored reopening some of the meta-theoretical issues that occupied earlier generations (see the contributions in Stehr and Meja 1984).

In these ways, the broad-constructionist approach diverges from the recent narrow-constructionist alternative. Found in the work of Camic (1983, 1992, Camic and Xie 1994), Collins (1998), Kusch (1995) and Shapin (1994), this second approach focuses the sociology of knowledge mainly on men and women who specialize in the production of ideas and on the particular social processes by which their ideas emerge and develop—a move that, in effect, transforms the field into a sociology of ideas. This perspective has tended to reject the core assumptions of the older sociology of knowledge, building instead on scholarship that argues that sociocultural processes are as much internal to the content of ideas as they are external (Bloor 1976, Shapin 1992), that the meanings of ideas are only understandable to an investigator after careful contextual reconstruction (Skinner 1969), and that local, micro-level settings are often the main sites for the development of ideas (Geertz 1983, Whitley 1984). Like the broad-constructionist approach, this narrow-constructionist perspective presently provides a foundation for several lines of empirical research (see Camic and Gross 2000). No forecast can yet be made, however, as to which approach, if either, will rescue the sociology of knowledge from its traditionally marginal position in the discipline of sociology.

See also: Comte, Auguste (1798–1857); Constructivism/Constructionism: Methodology; Culture, Soci-

ology of; Durkheim, Emile (1858–1917); Foucault, Michel (1926–84); History of Science; History of Science: Constructivist Perspectives; Ideology, Sociology of; Intellectuals: History of the Concept; Intellectuals, Sociology of; Knowledge Representation; Mannheim, Karl (1893–1947); Marx, Karl (1818–89); Positivism: Sociological; Science and Religion; Science, Sociology of; Sociology: Overview; Theory: Sociological

Bibliography

Althusser L 1969 *For Marx*. Penguin Press, London

Althusser L 1971 *Lenin and Philosophy and other Essays*. New Left Books, London

Ben-David J 1971 *The Scientist's Role in Society: A Comparative Study*. Prentice-Hall, Englewood Cliffs, NJ

Berger P L, Luckmann T 1966 *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Doubleday, Garden City, NY

Bloor D 1976 *Knowledge and Social Imagery*. Routledge and Kegan Paul, London

Camic C 1983 *Experience and Enlightenment: Socialization for Cultural Change in Eighteenth Century-England*. University of Chicago Press, Chicago

Camic C 1992 Reputation and predecessor selection. *American Sociological Review*. **57**: 421–45

Camic C, Gross N 2000 The new sociology of ideas. In: Blau J (ed.) *The Blackwell Companion to Sociology*. Blackwell, Oxford, UK

Camic C, Xie Y 1994 The statistical turn in American social science – Columbia University, 1890 to 1915. *American Sociological Review* **59**: 773–805

Collins H M 1983 The sociology of scientific knowledge – studies of contemporary science. *Annual Review of Sociology* **9**: 265–85

Collins R 1998 *The Sociology of Philosophies: A Global Theory of Intellectual Change*. Belknap Press of Harvard University Press, Cambridge, MA

Comte A [1830] 1975 *Auguste Comte and Positivism*: *The Essential Writings*. Harper & Row, New York

Coser L A 1965 *Men of Ideas: A Sociologist's View*. Free Press, New York

Coser L A 1968 Knowledge, sociology of. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 8, pp. 428–35

Curtis J E, Petras J W (eds.) 1970 *The Sociology of Knowledge: A Reader*. Duckworth, London

DeGre G [1939–43] 1985 *The Social Compulsions of Ideas: Toward a Sociological Analysis of Knowledge*. Transaction Books, New Brunswick, NJ

Durkheim E [1912] 1995 *The Elementary Forms of Religious Life*. Free Press, New York

Durkheim E, Mauss M [1903] 1963 *Primitive Classification*. University of Chicago Press, Chicago

Eagleton T 1991 *Ideology: An Introduction*. Verso, London

Frisby D 1983 *The Alienated Mind: The Sociology of Knowledge in Germany, 1918–1933*. Heinemann Educational Books, London

Fuhrman E R 1980 *The Sociology of Knowledge in America, 1883–1915*. University of Virginia Press, Charlottesville, VA

Geertz C 1983 *Local Knowledge*: *Further Essays in Interpretive Anthropology*. Basic Books, New York

Gouldner A W 1965 *Enter Plato*. Basic Books, New York

Gouldner A W 1970 *The Coming Crisis of Western Sociology*. Basic Books, New York

Granet M 1934 *La Pensée Chinoise*. La Renaissance du livre, Paris

Griswold W 1986 *Renaissance Revivals City, Comedy and Revenge Tragedy in the London Theatre, 1576–1980*. University of Chicago Press, Chicago

Halbwachs M 1925 *Les Cadres Sociaux de la Mémorie*. F. Alcan, Paris

Harding S G 1991 *Whose Science? Whose Knowledge? Thinking from Women's Lives*. Cornell University Press, Ithaca, NY

Hegel G W F [1807] 1967 *The Phenomenology of Mind*. Harper & Row, New York

Hoare Q, Smith G N (eds.) 1971 *Selections from the Prison Notebooks of Antonio Gramsci*. Lawrence & Wishart, London

Kennedy E 1978 *A Philosopher in the Age of Revolution: Destutt de Tracy and the Origins of Ideology*. University of Pennsylvania Press, Philadelphia

Kettler D, Meja V 1995 *Karl Mannheim and the Crisis of Liberalism*: *The Secret of These New Times*. Transaction Publishers, New Brunswick, NJ

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Kuklick H 1983 The sociology of knowledge: Retrospect and prospect. *Annual Review of Sociology*. **9**: 287–310

Kusch M 1995 *Psychologism: A Case Study in the Sociology of Philosophical Knowledge*. Routledge, London

Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy: Towards a Radical Democratic Politics*. Verso, London

Lamont M 1987 How to become a dominant French philosopher – the case of Derrita Jacques. *American Journal of Sociology* **93**: 584–622

Larrain J 1979 *The Concept of Ideology*. University of Georgia Press, Athens, GA

Mannheim K [1925] 1936a Ideology and utopia. In: Mannheim K (ed.) *Ideology and Utopia: An Introduction to the Sociology of Knowledge*. Harcourt Brace Javanovich, New York

Mannheim K [1926] 1936b Preliminary approach to the problem. In: Mannheim K (ed.) *Ideology and Utopia: An Introduction to the Sociology of Knowledge*. Harcourt Brace Jovanovich, New York

Mannheim K [1929] 1936c The sociology of knowledge. In: Mannheim K (ed.) *Ideology and Utopia: An Introduction to the Sociology of Knowledge*. K. Paul, Trench, Trubner, London

Mannheim K [1931] 1986 *Conservatism: A Contribution to the Sociology of Knowledge*. Routledge & Kegan Paul, London

Marcuse H 1964 *One-Dimensional Man*: *Studies in the Ideology of Advanced Industrial Society*. Beacon, Boston

Marx K [1852] 1963 *The 18th Brumaire of Louis Bonaparte*. International Publishers, New York

Marx K [1894] 1967 *Capital*. International Publishers, New York, Vol. 3

Marx K, Engels F [1846] 1970 *The German Ideology*. International Publishers, New York

Maynard D W, Clayman S E 1991 The diversity of ethnomethodology. *Annual Review of Sociology* **17**: 385–418

McCarthy E D 1996 *Knowledge as Culture: The New Sociology of Knowledge*. Routledge, London

Mepham J 1979 The theory of ideology in capital. In: Mepham J, Ruben D-H (eds.) *Issues in Marxist Philosophy, Vol 3, Epistemology, Science, Ideology*. Harvester, London

Merton R K [1941] 1968a Karl Mannheim and the sociology of knowledge. In: Merton R K (eds.) *Social Theory and Social Structure*. Free Press, New York
Merton R K [1945] 1968b The sociology of knowledge. In: Merton R K (eds.) *Social Theory and Social Structure*. Free Press, New York
Mills C W [1943] 1963 *Power, Politics, and People: The Collected Essays of C. Wright Mills*. Oxford University Press, New York
Mills C W 1964 *Sociology and Pragmatism*. Oxford University Press, New York
Prado C G 1995 *Starting with Foucault: An Introduction to Genealogy*. Westview Press, Boulder, CO
Ringer F K 1969 *The Decline of the German Mandarins: The German Academic Community, 1890–1933*. Harvard University Press, Cambridge, MA
Ringer F K 1990 The intellectual field, intellectual history, and the sociology of knowledge. *Theory and Society* **19**: 269–94
Scheler M [1924] 1980 *Problems of a Sociology of Knowledge*. Routledge and Kegan Paul, London
Shapin S 1982 History of science, and its sociological reconstructions. *History of Science* **20**: 157–211
Shapin S 1992 Discipline and bounding: The history and sociology of science as seen through the externalism–internalism debate. *History of Science* **30**: 333–69
Shapin S 1994 *A Social History of Truth: Civility and Science in Seventeenth-Century England*. University of Chicago Press, Chicago
Shils E [1955–68] 1972 *The Intellectuals and the Powers and Other Essays*. University of Chicago Press, Chicago
Simonds A P 1978 *Karl Mannheim's Sociology of Knowledge*. Clarendon, Oxford
Skinner Q 1969 Meaning and understanding in the history of ideas. *History and Theory*. **8**: 3–53
Smith D 1987 *The Everyday World as Problematic: A Feminist Sociology*. University of Toronto Press, Toronto
Stark W 1958 *The Sociology of Knowledge: An Essay in Aid of a Deeper Understanding of the History of Ideas*. Free Press, Glencoe, IL
Stehr N, Meja V (eds.) 1984 *Society and Knowledge: Contemporary Perspectives in the Sociology of Knowledge*. Transaction Books, New Brunswick, NJ
Swidler A, Arditi J 1994 The new sociology of knowledge. *Annual Review of Sociology* **20**: 305–29
Therborn G 1980 *The Ideology of Power and The Power of Ideology*. NLB, London
Thompson J B 1990 *Ideology and Modern Culture: Critical Social Theory in the Era of Mass Communication*. Polity Press, Cambridge, UK
Toews J E 1987 Intellectual history after the linguistic turn. *American Historical Review* **92**: 879–907
Whitley R 1984 *The Intellectual and Social Organization of the Sciences*. Oxford University Press, Oxford, UK
Wirth L 1936 Preface. In: Mannheim K (ed.) *Ideology and Utopia: An Introduction to the Sociology of Knowledge*. Harcourt Brace Jovanovich, New York
Wolff K H [1942–65] 1983 *Beyond the Sociology of Knowledge*. University Press of America, Lanham, NY
Wuthnow R 1989 *Communities of Discourse*. Harvard University Press, Cambridge, MA
Znaniecki F 1940 *The Social Role of the Man of Knowledge*. Columbia University Press, New York

C. Camic

Knowledge Spaces

Knowledge spaces are combinatorial models for encoding the conceptual organization of a given body of knowledge. Their theory leads to the design of efficient algorithms for assessing the concepts acquired by an individual in the given domain. Computer software implementing such algorithms is now in use in schools and colleges (e.g., Aleks, briefly presented later). Their output is by far richer than just a crude number (e.g., 8 out of 10) or mark (e.g., B). Indeed, it describes the present state of knowledge of the individual together with advice for further progress. Thus, by its combinatorial emphasis, knowledge space theory appears as a multidimensional generalization of classical test theory. It can also be applied to other settings (for instance, system failure analysis). Probabilistic extensions of the theory are covered in the article *Markov Processes for Knowledge Spaces*, while here the basic, deterministic concepts are introduced and some applications are sketched.

1. *Organizing a Body of Knowledge*

To start with an example, suppose that a teacher wants to assess the acquired capabilities of the pupils in a classroom as regards arithmetical multiplication. A list of items will be set up, starting from a simple question such as

Item *a* $\quad 9 \times 8 = ?$

and ending with a more complicated problem such as

Item *b*
$$\begin{array}{r} 278.9 \\ \times\, 3.98 \\ \hline = \quad ? \end{array}$$

In classical practice, all items will be proposed to all pupils. The outcome of such a test boils down to a mark assigned to each pupil, which provides no other information than the number of correct answers. The performance of such a test is very poor. First, the pupils have to work on all problems (be these problems fairly easy or on the contrary out of reach for their actual competencies). Second, the marks convey a rather inaccurate description of the concepts and techniques mastered by each individual pupil.

These drawbacks are circumvented in oral testing. Here, the teacher can improve the examination procedure by taking into account answers previously collected in order to carefully select the next question to be asked. Moreover, this flexibility helps in building a more accurate description of the pupils' strengths and weaknesses. To illustrate these two points with our example, a teacher who checks that a pupil masters Item *b* will probably avoid asking Item *a* to that pupil.

On the other hand, if the pupil does not master Item b, other items will be proposed in order to diagnose which mistakes or misconceptions hamper the pupils' acquisitions. By carefully selecting each time the new item, the teacher will gradually refine a description of the concepts and techniques mastered by the pupil. Such a description can be complemented with advice for remediation and a plan for further study.

Which are the factors that potentially make oral examination superior to written testing? Notice first that our analysis assumes a decomposition of the body of knowledge under testing into elementary pieces. To make things simple, we may suppose that each such piece corresponds to an item. Then, we assume some structure on the set of all these items. Back to our example, we declare Item a to be a prerequisite for Item b. As a consequence, we take for granted the following assertion:

from a correct answer to Item b, we may surmise a correct answer to Item a (if the latter is asked).

In short: mastery of Item a is surmised from mastery of Item b. By contraposition, this rule is reformulated as

a false answer to Item a implies a false answer to Item b.

Mathematically speaking, we have a (binary) relation P on the set Q of all items. The above assertions can now be written as aPb. We call P the surmise relation on Q and assume that P is a quasi order, that is: P is reflexive (qPq holds for all q in Q) and transitive (for all $p, q, r \in Q$, pPq and qPr imply pPr). Such a relation puts severe restrictions on the possible states of knowledge.

2. Knowledge States

Given the set Q of items, we identify the state of knowledge of an individual with the subset $\mathscr{K}$ of items that this person masters; thus $\mathscr{K} \subseteq Q$. However, in view of the surmise relation P, not every subset of Q can be a state of knowledge. Indeed, if a knowledge state contains q and pPq holds, then $\mathscr{K}$ must also contain p. We illustrate this with an example where $Q = \{a, b, c, d\}$ and P is the relation depicted in Fig. 1.

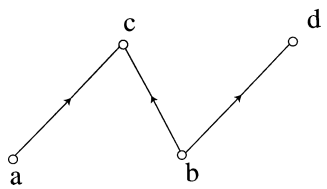

Figure 1
Diagram for a surmise relation.

Subset $\{a, c\}$ is not a knowledge state since bPc holds, while $\{a, b, c\}$ is a knowledge state as can be checked. The full family $\mathscr{K}$ of knowledge states in this example is as follows:

$$\mathscr{K} = \{\varnothing, \{a\}, \{b\}, \{a, b\}, \{b, d\}, \{a, b, c\}, \{a, b, d\}, Q\}$$

where $\varnothing$ denotes the empty set. Notice $\{a, b\} \in \mathscr{K}$ and $\{b, d\} \in \mathscr{K}$, and also $\{a, b\} \cap \{b, d\} = \{b\} \in \mathscr{K}$.

This example can be used to illustrate some general properties of the family $\mathscr{K}$ of knowledge states of any surmise relation. The empty set $\varnothing$ and the whole set Q belong to $\mathscr{K}$; moreover, if $\mathscr{K}, L \in \mathscr{K}$, then $\mathscr{K} \cap L \in \mathscr{K}$ and $\mathscr{K} \cup L \in \mathscr{K}$, that is: $\mathscr{K}$ is closed under intersection and under union. It turns out that these properties characterize families of subsets deduced from surmise relations. Indeed, this is a reformulation of Birkhoff's Theorem (for infinite ground sets Q, closure under general intersection and union, not restricted to two subsets, need to be considered). We now argue that a more general setting is required.

3. Knowledge Spaces

Suppose that a student produces correct solutions to systems of linear equations in several unknowns. The examiner cannot infer from this information whether the student has mastered computation of determinants, Gauss elimination procedure, or the substitution scheme. At least one of these three methods is mastered, but which one is not clear from the student's answers. A generalization of the concept of a surmise relation is thus needed. Again, let Q be a set of items. To each item q, several subsets $C_1, C_2, \ldots, C_k$ of Q will be attached. These clauses for q represent all possible ways of acquiring the mastery of q; their number k may of course vary with q. We denote by σ the mapping associating to q the collection $\{C_1, C_2, \ldots, C_k\}$. For an example, assume $Q = \{a, b, c, d\}$ and

$$\sigma(a) = \{\{a, b, c\}, \{a, d\}\} \tag{1}$$

$$\sigma(b) = \{\{b, c\}\} \tag{2}$$

$$\sigma(c) = \{\{c\}\} \tag{3}$$

$$\sigma(d) = \{\{d\}\}. \tag{4}$$

The expression for $\sigma(a)$ thus encodes the following fact: mastery of item a implies mastery of a itself plus

(b and c) or d.

In fact, a mapping σ is another formalization of an AND/OR graph from artificial intelligence. We will now make the concept more precise by taking into consideration four axioms (satisfied by our last example).

(a) for each item q, there is at least one clause for q;

(b) for each item q and each clause C for q, we have $q \in C$;

(c) for each pair of items q, r and each clause C for q containing r, there is a clause D for r such that $D \subseteq C$;

(d) any two clauses C, C' for the same item are incomparable, i.e. neither $C \subset C'$ nor $C' \subset C$ hold.

Axiom (b) clearly generalizes the reflexivity axiom of a surmise relation, while Axiom (c) extends transitivity. Axiom (d) translates our assumption that a clause for item q encodes a minimal way of acquiring the mastery of q.

Given a surmise system σ on Q, we define a knowledge state as any subset $\mathscr{K}$ of Q satisfying:

for each item q in $\mathscr{K}$, there is a clause C for q that is contained in $\mathscr{K}$.

In the four-item example, the collection $\mathscr{K}$ of all knowledge states is as follows:

$$\mathscr{K} = \{\varnothing, \{c\}, \{d\}, \{a, d\}, \{b, c\}, \{c, d\}, \{a, b, c\}, \{a, c, d\}, \{b, c, d\}, Q\}.$$

Any surmise relation P on Q becomes a surmise system if we attach to any item q the single clause $C = \{r \in Q | rPq\}$. Notice that the knowledge states defined from P coincide with those defined from the surmise system. We have moreover a neat variant of Birkhoff's Theorem. Indeed, a family $\mathscr{K}$ of subsets of a finite set Q constitutes the collection of all knowledge states of a surmise system if and only if $\mathscr{K}$ is closed under union, and contains $\varnothing$ and Q (an infinite version of this result is presented in Doignon and Falmagne 1999). A structure $(Q, \mathscr{K})$ satisfying these two conditions is called a knowledge space.

Closure under union is not required in most practical applications. A knowledge structure $(Q, \mathscr{K})$ is a finite set Q of items together with a family $\mathscr{K}$ of subsets of Q, with $\varnothing$ and Q belonging to $\mathscr{K}$. The elements of $\mathscr{K}$ are the knowledge states.

4. *Wellgradedness*

Knowledge assessment in our context assumes that a knowledge structure $(Q, \mathscr{K})$ has been built; the goal then is to design an algorithm that determines which state in $\mathscr{K}$ represents the knowledge of a student. This algorithm will efficiently select questions to be asked, thus reducing at each step the remaining possible states. Clearly, if all answers collected truely reflect the student's knowledge, the process will end after a few steps with a precise element of $\mathscr{K}$.

In real life, things are not so simple. Answers may be biased by lucky guesses (even if the items are not of the multiple-choice type this can happen) or by careless errors (a slight flaw in computation may alter the correctness). To take these undeterministic aspects into account, a probabilistic extension of the theory was conceived (see the article *Markov Processes for Knowledge Spaces*). It relies on a combinatorial notion that we now sketch.

After some time, an assessment procedure will have proposed one state $\mathscr{K}$ from $\mathscr{K}$ as the accurate representation in the structure $(Q, \mathscr{K})$ of the student's knowledge. A further answer of the student, say to item q, may call for a modification of $\mathscr{K}$. If $q \notin \mathscr{K}$ and q is correctly answered, then $\mathscr{K} \cup \{q\}$ should become the state selected in $\mathscr{K}$—provided $\mathscr{K} \cup \{q\}$ belongs to $\mathscr{K}$. Similarly, if $q \in \mathscr{K}$ and q is incorrectly answered, then $\mathscr{K} \setminus \{q\}$ should be selected—provided it belongs to $\mathscr{K}$. We may not assume that always $\mathscr{K} \cup \{q\}$ and $\mathscr{K} \setminus \{q\}$ belong to $\mathscr{K}$, because otherwise $\mathscr{K}$ would be the collection of all subsets (which is a much too large collection for our purposes). On the other hand, the assessment procedure should be able to explore the whole collection $\mathscr{K}$ by elementary steps consisting in the addition or deletion of a single item. To state a precise definition, we use for any two subsets $\mathscr{K}$, L of Q their (symmetric-difference) distance

$$d(\mathscr{K}, L) = |\{q \in Q | q \in \mathscr{K} \setminus L \text{ or } q \in L \setminus \mathscr{K}\}|$$

which counts the number of items q in which $\mathscr{K}$ and L differ. By definition, a knowledge structure $(Q, \mathscr{K})$ is wellgraded when for any two knowledge states $\mathscr{K}$, L in $\mathscr{K}$ with $d(\mathscr{K}, L) = h$, there exists a sequence $\mathscr{K} = \mathscr{K}_0, \mathscr{K}_1, \ldots, \mathscr{K}_h = L$ of states in $\mathscr{K}$ such that $d(\mathscr{K}_j, \mathscr{K}_{j+1}) = 1$ for $j = 0, 1, \ldots, h-1$.

For $\mathscr{K}$ a state in a knowledge structure $(Q, \mathscr{K})$, we define the inner fringe and outer fringe of $\mathscr{K}$ as

$$\mathscr{K}^I = \{q \in \mathscr{K} | \mathscr{K} \setminus \{q\} \in \mathscr{K}\}, \tag{5}$$

$$\mathscr{K}^O = \{q \in Q \setminus \mathscr{K} | \mathscr{K} \cup \{q\} \in \mathscr{K}\}. \tag{6}$$

For example, if

$$\mathscr{K} = \{\varnothing, \{c\}, \{d\}, \{a, d\}, \{b, c\}, \{c, d\}, \{a, b, c\}, \{a, c, d\}, \{b, c, d\}, Q\}$$

then

$$\{a, d\}^I = \{a\} \tag{7}$$

$$\{a, d\}^O = \{c\} \tag{8}$$

In real-life applications, the outer fringe $\mathscr{K}^O$ collects those items that a student in knowledge state $\mathscr{K}$ could directly learn without any further prerequisite. This notion plays a crucial role in the design of the Aleks software.

It can be proved that a knowledge structure $(Q, \mathscr{K})$ is wellgraded if, and only if, any state in $\mathscr{K}$ is

characterized by its inner and outer fringes. The latter condition can be stated as follows: if $\mathscr{K}$, $L \in \mathscr{K}$ and $\mathscr{K}^{\mathrm{I}} \subset L$, $\mathscr{K}^{\mathrm{O}} \subset Q \backslash L$, then $\mathscr{K} = L$.

5. *Aleks*

Aleks is a software under development at Irvine, California by a team headed by Jean-Claude Falmagne (other projects with similar goals are in an earlier stage). At its completion, it will be able to assess the mathematical knowledge of pupils and students from Kindergarten to Grade 12. Presently in use in California schools and American colleges, it covers arithmetic and algebra by a batch of more than 200 items. For a user, a session typically presents about 10 items, carefully selected on the basis of previously acquired information on that particular user, and also on collected answers. At the end of the session, the student receives information on her progress and advice for faster study. Of course, the teacher is also informed of the monitoring process. Aleks is more than an assessment system, since it provides on demand the user with explanation and access to a dictionary. From experiments, it appears that students using the Aleks system progress much further in the material than students in a conventional class.

6. *Bibliographic Orientation*

The first publication on knowledge spaces was a 1985 paper by Doignon and Falmagne. The same authors published a survey book in 1999. Various applications of the theory are covered in two books edited by Albert (1994), and Albert and Lukas (1999). An evolutive and searchable bibliography is maintained by Cord Hockemeyer at the following web address:

http://wundt.kfunigraz.ac.at/hockemeyer/bibliography.html

while Aleks website is at:

http://www.aleks.com

See also: Instructional Technology: Cognitive Science Perspectives; Intelligent Tutoring Systems; Knowledge (Explicit and Implicit): Philosophical Aspects; Knowledge Representation; Logics for Knowledge Representation; Markov Processes for Knowledge Spaces; Tacit Knowledge, Psychology of

Bibliography

Albert D (ed.) 1994 *Knowledge Structures*. Springer-Verlag, Berlin

Albert D, Lukas J (eds.) 1999 *Knowledge Spaces: Theories, Empirical Research, and Applications*. Erlbaum, Mahwah, NJ

Doignon J-P, Falmagne J-C 1999 *Knowledge Spaces*. Springer-Verlag, Berlin

J. P. Doignon

Kohlberg, Lawrence (1927–87)

Kohlberg was born on October 25, 1927, in Yonkers, New York, the son of a Protestant mother—a well-known chemist—and a Jewish father—a successful self-made businessman with strong civic commitment and of conservative leaning. Kohlberg grew up in a liberal Jewish community. In 1945 he joined the US Merchant Marines; he then helped to smuggle East European Jewish refugees to Palestine, was seized by the British army, and was interned in Cyprus. He fled to Palestine and spent some time in a kibbutz—pondering issues of justice and violence. In 1947 he went to the University of Chicago, reading moral philosophy from Plato to Dewey, studying Locke, Mill, and Jefferson in search of a basic moral principle that might bridge disagreements—and discovered Kant's imperative: 'treat every human being as an end in himself.' He did graduate work in psychology, studying with Bettelheim, Rogers, Gewirtz, Morris, Newgarten, Havighurst, and Strauss—'the first American scientist who attempted to integrate Piaget's structural theory with Mead's symbolic interactionism' (Kohlberg 1984, p.vii).

After two years of clinical work he started research for his doctoral dissertation on the development of moral judgment that he finished in1958. From 1959 to 1961 he was assistant professor at Yale University; 1961–2 fellow at the Institute for Advanced Study in the Behavioral Sciences in Palo Alto; 1962–8 professor at the University of Chicago; 1968–87 professor at Harvard University. Having an incurable disease, on January 17, 1987 Kohlberg ended his life.

Kohlberg changed US psychology. His influence is due to intellectual brilliance, but also to an inspiring personality attracting scientists from all over the world. Kuhmerker et al.(1991) elicited personal letters from friends and colleagues describing what Kohlberg had meant for their lives. Some core characteristics emerge: Kohlberg was 'loveable'—and had 'a remarkable ability to insure this characteristic in those who worked with him…all of whom had not only a peculiarly fanatical interest in the study of moral development but also a deep personal affection for Lawrence Kohlberg, the man.' (Grimley) Kohlberg was persistent; what struck Kegan was 'that an idea can be compelling enough and an individual devoted enough for the two to spend 25 years together in a relationship as fresh and combustible today as when it began.' Kohlberg was stern—demanding extraordinary efforts of himself and of others; he 'would call another scoring workshop whenever he felt his

assistants were working less than 80 hours a week.' His commitment was contagious: 'At the Center of Moral Education there was a sense of being at the right place at the right time—that terribly new and important breakthroughs were just around the corner…that questions of social justice were on everyone's mind and we thought we had the answers—or nearly had the answers.' He was 'a visionary' who kept 'the major questions in focus.' Had it not been for him 'many of us would not have become interested in psychology at all' (Rest).

For Kohlberg the unity of moral judgment and action was not just a theoretical stipulation—it was his very personality. A colleague recalls an occasion when he had initiated a resolution against the invasion to Cambodia. When the question arose who would accompany him in handing it over in Washington:

> …the meeting became silent. Motionless. Eerily quiet. These giants of education who daily proclaimed the virtues of equality, of human rights, of democratic citizenship and had themselves just voted overwhelmingly in favor of the resolution now sat in perfect pose, looking at their shoes. Numbed and dazzled by this educational experience. I finally noticed one lone figure somewhat unkempt in attire standing next to me. Although I had met him a number of prior occasions, that was the time I really met Larry. (Sprinthall)

Kohlberg addressed important issues authentically and in an innovative way. This carried others along. Thus, Kohlberg's theory springs also from collective endeavors—initiated, inspired, and directed by Kohlberg, but advanced by his students, friends, and colleagues as well.

1. *Kohlberg's Approach*

Kohlberg's approach ran counter to the basic assumptions about the nature of human beings, of norms, and of science dominant at the time. Humans were seen as basically passive (shaped by reinforcement schedules in behaviorism, by parental authorities in psychoanalytic theories, or by the gradual unfolding of genetic endowments in maturationism) and primarily motivated by self-serving concerns (avoiding external or superego sanctions). In contrast, Kohlberg—following Piaget—viewed people as active interpreters of their world and also motivated by intrinsic concerns (for mastery, truth, or other values).

Norms were seen as externally given contingent facts. Kohlberg, instead, viewed norms as propositions raising the validity claim of rightness, that is, in need of and capable of being justified. Psychology followed a natural science epistemology: copying or explaining reality is aimed for by standardizing measurement procedures or experimentally isolating and manipulating variables. In contrast, Piagetian analyses seek to trace universal developmental changes in subjects' interactive constructions of reality; hermeneutic interviews aim at objectivity by guaranteeing not the physical identity of the stimulus, but the identity of its meaning. Kohlberg was one of the pioneers in bringing this Piagetian cognitive frame to bear on academic psychology in the USA.

Also, his interest in moral reasoning was new. Up until then, only normconforming behavior (not moral action) or moral emotions (not moral judgment) had been studied—both from an observer's point of view. Kohlberg, instead, looked at morality from a first-person perspective. And he did so with a specific—not uncontested—understanding of morality by focusing on judgment he excluded questions of moral character, of habitualized moral dispositions, of virtues. By defining morality on deontic terms he excluded questions of the good life that he left to ethics and/or to the personal domain. By identifying morality with a justice procedure he largely neglected content. By describing this procedure as 'moral musical chairs' he sketched a hypothetical role-taking procedure that remains entirely monological.

2. *Moral Development*

2.1 *The Stage Model*

Piaget had described young children's moral development as progress from a heteronomous stage (characterized by egocentrism; unilateral respect; a focus on consequences rather than on intention; an understanding of rules as objective, absolute and unalterable; a belief in immanent justice) to an autonomous morality produced by peer interaction and cognitive decentering. From his study of adolescent development Kohlberg derived a refined description of the stage model and of developmental mechanisms.

Development consists in changes in the reasoning structure justifying judgments of 'right' or 'wrong' and in the concerns motivating norm conformity. Its structural core is a widening of the sociomoral perspective. Universally, there are three levels with two 'hard' stages each. The model is based on the following assumptions:

– each stage is a structured whole that is qualitatively different from the others;

– the sequence of stages is invariant, irreversibly progressive without stage skipping;

– the stages form a hierarchy: higher stages are better differentiated and integrated structures that allow morally better solutions to dilemmas.

The stages are described as follows:

At the preconventional level (typical for most children under nine) situations are judged from a concrete individual perspective. Rules are seen to be external to the self. 'Right' is defined as obeying rules and authority, avoiding punishment, not doing physical harm (Stage 1) or as meeting one's own interests and letting others do the same (Stage 2); the reasons for doing right are to avoid punishment and to bow to

the superior power of authorities (Stage 1) or to serve one's own needs or interests (Stage 2).

At the conventional level (typical of most adults) morality is judged from a 'member of society' perspective and is identified with socially shared internalized norms. 'Right' means living up to the expectations of one's group (Stage 3) or doing one's duty in society (Stage 4); the reasons for doing right are the need to be 'good' in one's own or others' eyes and caring for others (Stage 3) or the need to maintain a good conscience by meeting defined obligations (Stage 4).

At the postconventional level (reached by few adults only) moral values are defined from a 'prior to society' perspective in terms of self-chosen universal principles. 'Right' is upholding legal contracts (Stage 5) or is defined in terms of universal ethical principles such as equality and respect for the dignity of human beings as individuals (Stage 6). The reasons for doing right are obligations to the social contract (Stage 5) or a commitment to principles that have been rationally understood and accepted as valid (Stage 6).

2.2 Tests of Moral Development

In the Moral Judgment Interview (MJI, Colby and Kohlberg 1987) hypothetical moral dilemmas are given which allow of no 'right' answer (e.g., Should a man steal a drug to save his dying wife?) Subjects are first to decide what the protagonist ought to do (uphold the law or preserve life?) and then to justify and defend this decision against moral counterarguments advanced by the interviewer who aims at eliciting the subject's highest competence level. The responses are matched to structurally similar criterion judgments listed for each stage in the scoring manual.

In the Social Reflection Measure (SRM) Gibbs et al. (1992) developed a brief interview and coding manual that can be used for younger children. Objective tests have also been developed in which subjects rank their preference for a set of standardized responses that express stage-specific considerations (Rest's Defining Issue Test, DIT, Lind's Moral Judgement Test, MUT). Preference measures produce scores that are consistently (about one stage) higher than spontaneously produced answers, yet do correlate in medium range with scores in MJI.

2.3 Mechanisms of Development

In contrast to social learning and psychoanalytic theories which assume identification or modeling as the primary developmental mechanisms, cognitive theories see development as a result of internal restructuring in response to cognitive conflict. There is experimental evidence in support of this mechanism (cf. Walker 1986). The need to reconcile conflicting views may be the reason why taking leadership roles advances moral development.

This understanding of developmental mechanisms has inspired pedagogic interventions. Classroom discussions of moral dilemma tend to produce (albeit slight) moral growth. The institutionalization of just communities, that is, structures allowing for democratic self-government, have been shown to improve the moral atmosphere of schools or prisons considerably and generate feelings of reciprocal responsibility (Higgins et al. in Kurtines and Gewirtz 1984).

2.4 Empirical Support and Criticisms of the Stage Model

2.4.1 Sequence. Age and stage have been shown to correlate in cross-sectional studies. Regression or stage skipping have rarely been found in longitudinal studies and not been induced in experimental studies designed to produce moral development (Snarey 1985, Walker 1986). Thus, the claim of a developmental sequence and of an invariant order stands.

2.4.2 Hierarchy. There is conceptual and empirical support for the claim of hierarchy. Conceptually, the stage sequence comprises an increasing universalization of perspectives (isolated individual; dyad; group; a given society; a democratically organized society; humanity at large) and an expansion of relevant concerns (negative and positive consequences to the actor; intentions; role obligations; societal institutions and legality; universal rights). Postconventional thinking makes it possible to integrate and balance these different concerns in view of universal principles of equality and respect for the individual. Empirically, subjects prefer higher and reject lower stage reasoning.

2.4.3 Universality. Kohlberg has been accused of ethnocentrism and gender bias in claiming universal validity for a stage model that was initially developed from interviews with males in the USA only. Two issues are at stake: first, is there a universal core to morality, and second, is Kohlberg's stage sequence universal? The first issue is dealt with theoretically in moral philosophy and practically in endeavors to win worldwide acceptance for human rights. The second issue sparked extensive empirical research (see *Ethnocentrism*; *Gender and Feminist Studies in Anthropology*; *Gender Differences in Personality and Social Behavior*).

The universal presence of stages 1–4 is reported in a review of 45 studies from 27 widely diverse cultures involving over 5,000 subjects. Postconventional thinking was found in urban samples only, not in traditional tribal or village societies either in Western or non-Western cultures (Snarey 1985). Principled thinking thus seems a 'metalevel' feature necessary for

adjudicating conflicts between subculture-specific norms and not required in normatively integrated and isolated cultures.

Gilligan (1982) has claimed the existence of two moralities: a rigid justice orientation more typical for males (corresponding to stage 4) and a flexible morality of care and responsibility more typical for females (corresponding to stage 3). Empirically, reviews of research involving 19,000 subjects have shown that either there are no stage differences between the sexes (Lind et al. 1987) or else they tend not to disadvantage women, or to disappear when education and employment are controlled for (Walker 1984). Conceptually, Gilligan starts from a more encompassing understanding of morality than Kohlberg by including questions about the good life. Also, there are some theoretical confusions: 'care' often is experienced as a 'duty' by women, and flexibility—if not a mere reflection of powerlessness—seems a correlate of a modern secularized moral understanding ('ethics of responsibility', Max Weber) rather than of sex membership.

2.5 Another View of Moral Development?

One core issue concerns the relationship between judgment and action. This problem has been dealt with in several ways. First, the claim was raised and empirically supported that consistency will increase with moral development (Kohlberg and Candee 1984, Blasi 1980). This makes sense given that subjects referring to self-serving concerns will be assigned to lower stages. Later, responsibility judgments (expressing the personal relevance of moral judgments, Higgins et al. 1984, Blasi 1984), and the distinction between types A (considering pragmatic concerns) and B (exclusively focusing on moral values) have been introduced (Kohlberg 1984, Colby and Kohlberg 1987). These additions are remedies to deal with the fact that individuals differ in intensity of moral commitment—as is especially clearly evidenced by Colby's and Damon's (1994) 'moral exemplars'.

A more far reaching theoretical revision, however, may be called for in view of recent studies on children's moral understanding. Turiel (1983) and Nucci and Turiel (1993) have shown that young children distinguish between moral rules (enjoying universal, unalterable, intrinsic validity) and conventional or religious rules (that depend on authorities or God and are valid only for in-group members). Nevertheless, most young children expect a wrongdoer to feel good and only gradually begin to ascribe negative emotions (Nunner-Winkler and Sodian 1988). Until about six or seven years (i.e., before self-reflexive role-taking is developed) these predict moral behavior. By about 10 years almost all children (even those low in moral motivation) say they expect a wrongdoer to feel bad. In justifying these negative emotions most children tend to refer neither to impending sanctions nor to empathic identification with the victim but to the fact that a wrong had been committed. These departures from Kohlberg's description of preconventional thinking result from differences in problems addressed and measurement procedures used: Turiel and Nucci and Lee questioned children's understanding of norm validity. Nunner-Winkler focused on the intensity and type of moral motivation. Kohlberg asked for action recommendations in moral dilemmas. Until they begin to care about morality children will guide these along prudential rather than moral considerations.

A new reading of moral development suggests itself. Universally, children acquire an early understanding of the categorical ought: they know that moral norms are intrinsically valid and motives should not be self-serving. Thus, from early on, all have a good cognitive grasp of the constitutive feature of the moral domain—its deontic character. Moral competence, however, requires more. Sociocognitive development (which constitutes the structural core of Kohlberg's stage descriptions) and the acquisition of knowledge are needed for passing judgments in more complex situations. Moral motivation is needed to translate knowledge into action. These requisites are acquired in later and differential learning processes.

3. Kohlberg's Legacy

In discussing Kohlberg's contribution, Rest (in Kuhmerker et al. 1991) notes that the stage concept may dwindle away and additional variables (e.g., content, motivation, or character) may prove of import. What will remain is Kohlberg's part in the paradigm shift from behaviorism to cognitive psychology. More specifically, what will remain is his emphasis on the cognitive side of morality; his striving for reconstructing it from a participant perspective and linking it with philosophical quests for justifiability; his insistence that intrinsic motivation exists (still an urgent claim given rational choice reductionism); his emphasis that there is more to morality than 'being friendly with your friends' (Rest). In theory and practice Kohlberg was committed to an understanding of morality that demands building just institutions and treating all human beings with equal respect.

See also: Dewey, John (1859–1952); Erikson, Erik Homburger (1902–94); Gender Differences in Personality and Social Behavior; Hall, Granville Stanley (1844–1924); Hegel and the Social Sciences; James, William (1842–1910); Kant, Immanuel (1724–1804); Moral Development: Cross-cultural Perspectives; Moral Education; Moral Reasoning in Psychology; Moral Sentiments in Society; Morality (East and West): Cultural Concerns; Piaget, Jean (1896–1980); Piaget's Theory of Human Development and Education

Bibliography

Clark Power F, Higgins A, Kohlberg L 1989 *Lawrence Kohlberg's Approach to Moral Education*. Columbia University Press, New York

Colby A, Kohlberg L1987 *The Measurement of Moral Judgment. Vol. 1: Theoretical Foundations and Research Validation. Vol. 2: Standard Issue Scoring Manual*. Cambridge University Press, New York

Gibbs J C, Basinger K S, Fuller D 1992 *Moral Maturity. Measuring The Development of Sociomoral Reflection*. Erlbaum, Hillsdale, NJ

Kohlberg L 1969 Stage and sequence. The cognitive-developmental approach to socialization. In: Goslin D A (ed.) *Handbook of Socialization Theory and Research*. Rand McNally, Chicago pp. 347–480

Kohlberg L 1981 *Essays on Moral Development: Vol. 1. The Philosophy of Moral Development. Moral Stages and The Idea Of Justice*. Harper & Row, San Francisco

Kohlberg L 1984 *Essays on Moral Development: Vol. II: The Psychology of Moral Development. The Nature and Validity of Moral Stages*. Harper & Row, San Francisco

Kohlberg L 1987 *Child Psychology and Childhood Education*. Longman, New York.

Kohlberg L, Levine C, Hewer A 1983 *Moral Stages: A Current Formulation and a Response to Critics*, Basel

Kuhmerker L, Gielen U, Hayes R L (eds.) 1991 *The Kohlberg Legacy for the Helping Professions*. REP Books, Birmingham, AL

Kurtines W M, Gewirtz J L 1984 *Morality, Moral Behavior, and Moral Development. Basic Issues in Theory and Research*. Wiley, New York.

Kurtines W M, Gewirtz J L (eds.) 1991 *Handbook of Moral Behavior and Development. Vol. 1: Theory. Vol. 2: Research, Vol. 3: Application*. Erlbaum, Hillsdale, NJ

Modgil S, Modgil C (eds.) 1986 *Lawrence Kohlberg. Consensus and Controversy*. Falmer Press, Philadelphia.

New Directions for Child Development 1990 The legacy of Lawrence Kohlberg **47:** 1–106

Nunner-Winkler G 1998 Development of moral understanding and moral motivation. In: Weinert F E, Schneider W (eds.) *Individual development from 3 to 12. Findings from the Munich Longitudinal Study*. Cambridge University Press, New York pp. 253–90

Rest J R 1986 *Moral development. Advances in Research and Theory*. Praeger, New York

G. Nunner-Winkler

Köhler, Wolfgang (1887–1967)

Wolfgang Felix U. Köhler, a leading German–US psychologist in the twentieth century, was born to Franz E. Köhler, a German gymnasium (high-school) teacher and his wife Wilhelmine on January 21, 1887 in Reval (today, Tallinn), Estonia, at that time a part of the Russian Empire. Köhler's experiments with apes made him internationally renowned early during his long scientific career. As a cofounder of a school called Gestalt psychology, together with Max Wertheimer (1880–1943) and Kurt Koffka (1886–1941), he became an influential scholar many years before he left Europe to join the faculty of Swarthmore College, Pennsylvania in the USA.

1. Major Events in Wolfgang Köhler's Life and Work

The development and highlights of psychology at the University of Berlin during the first decades of the twentieth century are strongly associated with the name of Wolfgang Köhler (see Table 1). Major dates and events in his life include his time in Frankfurt (the birth-place of Gestalt psychology), his research years on the island of Tenerife during World War I, his professorship in Berlin, and his emigré years in the USA (1935–67).

1.1 Köhler's Assistantship at Frankfurt's Institute of Psychology

After his graduation from high school in 1905, Köhler undertook university studies in philosophy, history, and natural science in Tübingen, Bonn, and then Berlin, where he concentrated on psychology under Carl Stumpf's (1848–1936) supervision. Besides his studies in experimental psychology, especially in psychoacoustics with Stumpf, he continued to pursue his interests in natural science and physics under Max Planck (1858–1947) and Walther Nernst (1864–1941). In his doctoral dissertation, which he finished in 1909, he combined physical acoustics with psychoacoustics. This research, published in the same year in the *Zeitschrift für Psychologie*, demonstrated his great talent as a scientist. His findings forced his teacher Carl Stumpf to revise and do further research on his own psychoacoustical theories.

As an assistant at Frankfurt's institute of psychology (1909–1913) under its director Friedrich Schumann, Köhler worked with Max Wertheimer and Kurt Koffka, who also had been trained in natural science and experimental psychology. This cooperation lead to a life-long exchange among these three men, each of them involved in the advancement of psychological theory and research. Köhler worked on the then widely-accepted '*constancy hypothesis*,' continued his psychoacoustical investigations in Frankfurt, and in his paper on '*unnoticed sensations and errors of judgment*' (1913) merged his individual line of research with that of Max Wertheimer's pioneering Gestalt work.

1.2 Köhler's Research on Tenerife

From 1913 to 1920, Köhler conducted research on the mentality of apes, chickens, and children at an

Table 1
Chronology: major events and dates in Wolfgang Köhler's life and work

1887	Born on January 21, 1887 in Reval (Tallinn), Estonia
1905–9	University studies in Tübingen, Bonn, and Berlin
1909	Doctoral dissertation thesis with Carl Stumpf at Berlin's Institute of Psychology
1909–13	Assistant at Frankfurt's Institute of Psychology; collaboration with Max Wertheimer (1880–1943) and Kurt Koffka (1886–1941)
1911	*Privatdozent* (Lecturer) at Frankfurt
1913–20	Director of the Anthropoid Station on Tenerife; first epoch-making publications on 'transposition' and 'insight' in the chimpanzee
1920	Book: Treatise on Gestalt phenomena in natural sciences (*Die physischen Gestalten in Ruhe und im stationären Zustand: Eine naturphilosophische Untersuchung*)
1921	Professor at Göttigen Founding of the journal *Psychologische Forschung* (*Psychological Research*), together with Max Wertheimer, Kurt Koffka, and others; major journal of Gestalt psychology (until 1938)
1922–35	Full professor of psychology and director of Berlin's Institute of Psychology
1925	Book: *The Mentality of Apes* (trans. from the 1921 German book)
1929	Book: *Gestalt Psychology* (later edn. 1947)
1935	Final clash with the Nazi authorities in Berlin and emigration to the USA; psychology professor at Swarthmore College until 1958
1938	Book: *The Place of Value in a World of Facts*
1944	Influential research publication on 'figural after-effects' (together with Hans Wallach)
1954–6	Perceptual research at the *Institute for the Advanced Studies* in Princeton, NJ
1958–9	President of the *American Psychological Association* (USA)
1959–67	Perceptual and psychophysiological research at Dartmouth College, New Hampshire (USA)
1967	Deceased on June 11, 1967 in Enfield, New Hampshire (USA)

anthropoid research station of the Prussian Academy of Sciences on Tenerife, Canary Islands. He began his investigations with perceptual and problem-solving tasks, demonstrating the ape's and the chicken's size constancy and their transposition of perceived stimulus relationships. He also demonstrated the chimpanzee's fabrication and use of simple 'tools.' Most famous became the success of one of his apes named Sultan in fitting together two sticks to make them long enough to reach the goal object, a banana. Köhler (1917, 1921) also did extensive work on the chimpanzees' ability to solve a problem by stacking boxes on top of each other, thus demonstrating, once again, 'insightful' animal behavior. Köhler (1920; see Table 1) expanded the Gestalt-theoretical perspective when he wrote his treatise on '*Die physischen Gestalten in Ruhe und im stationären Zustand*' ('The physical configurations at rest and in stationary states'), a book in which he aimed to integrate basic principles in psychology and physics in a new psychoneural solution to the mind-body problem (Sect. 2.1 below). He became a professor at Göttingen's psychology institute for a short time (1921) before he followed Carl Stumpf in the chair at Berlin.

1.3 Köhler's Professorship at the University of Berlin

In 1922 Köhler became full professor of philosophy and director of the Psychological Institute at the University of Berlin, where he worked together with other eminent scientists like Max Wertheimer and Kurt Lewin (1890–1947). Here the international fame of the Berlin School of Gestalt psychology grew until its peak during the early 1930s. It was during this time that the Gestalt principles of the organization of perception were published and extended to new research issues (Wertheimer 1923) like memory, cognition, motivation, and aesthetics.

During these years Köhler also lectured abroad several times, in the USA at Clark University (1925–6) and at Harvard University (1934–5), in South America in Uruguay (1930), as well as Brazil and Argentina (1932). When Köhler encountered the harassment of university scholars by the Nazi government in 1933, he fought racial discrimination and struggled against the political authorities of those days; he may have been the only psychology professor in Germany who protested in public against the totalitarian system (Henle 1978). When in 1935 his assistants were fired, he left Germany and followed his colleagues into exile in the USA.

1.4 Köhler's Emigré Years in the USA (1935–67)

In 1935 Köhler became a psychology professor at a small but prestigious undergraduate college, Swarthmore, near Philadelphia. There he continued his research, but on a smaller scale since his efforts were hampered by the absence of advanced facilities

and the lack of a doctoral program at this institution (Ash 1995, Sarris 1995).

Nevertheless, Köhler, together with his American colleagues and students, published important new findings on figural after-effects and on related perceptual phenomena; he also studied the cortical activity associated with pattern vision (Sect. 2.2 below), and wrote three further books on Gestalt psychology, all in English (Köhler 1938).

After World War II, Köhler received worldwide recognition for his life work, including several honorary doctorates from American and European universities. He was a recipient of the APA's (American Psychological Association) Distinguished Scientific Contribution Award and was elected a Member of the National Academy of Sciences. He was also named an 'honorary member' (*Ehrenbürger*) of the Free University of Berlin, where he lectured several times, first as a visiting professor. He was elected president of the APA for the 1958–9 term, an unusual honor for an emigré scientist in the USA.

2. *Köhler's Accomplishments and his Impact on Psychology*

Gestalt theory had begun after the turn of the twentieth century as a scientific revolt, first against Wilhelm Wundt's (1832–1920) ruling elementaristic ('atomistic') approach, and later opposed to John Broadus Watson's (1878–1958)—then dominant—behaviorism. It took several decades before Gestalt theory became a fully acknowledged perspective in psychology.

2.1 *Köhler's Work on Cognition, Perception, and Memory*

Cognition and 'Insight' in the Chimpanzee: Köhler's ideas about thinking were opposed strongly to those of the other schools of psychology. Some of his research was specifically directed against E. L. Thorndike's (1874–1949) influential behavioristic interpretation of learning and problem solving in animals. According to the latter view, a correct behavioral response occurs in the course of blind trial-and-error activity; on repeated trials the correct response is reinforced and thus strengthened simply by associations in an 'automatic' way (Thorndike's 'law of effect'). Köhler, in sharp contrast to that approach, argued that Thorndike's puzzle-box task did not allow any perceptual-cognitive insight into the whole procedural arrangement and thus represents no possibility of real problem solving; 'insight' requires a situation which is completely 'natural' and 'visible' to the animal (Köhler 1921).

According to Köhler (1951), problem solving is based on the organization of meaningful relations. It begins with a situation and a goal that cannot be directly reached (the so-called 'detour' problem). The process of problem solving requires a re-examination of the situationally given materials which must be 'reorganized,' quite analogously to the reversal of perceptual patterns. Köhler was criticized by his opponents for not considering the influence of past experience on problem solving. He did not deny the role of past experience in principle, but considered it of minor significance in insightful learning. This issue has still not been fully resolved today.

2.1.1 *Human perception and psychophysical theory.*

Basic to Köhler's perceptual research was the study of stimulus relationships, starting with his comparative work on transposition in apes, chickens, and children (Köhler 1917); for example, a chicken is presented with two gray samples, a darker one, *A*, and a lighter one, *B* (the positive stimulus): after successful discrimination training the chicken is tested with a new pair of samples, *B* and *C* (*C* being lighter than *B*). The majority of Köhler's hens chose *C* over *B*, thus showing relative choice behavior ('transposition'). Köhler obtained similar experimental results with his chimpanzees as subjects, using size as well as brightness as stimulus dimensions, thereby demonstrating the priority of 'structural' properties in perception and learning. From the perspective of Gestalt-theory, the effective stimulus for a given perception or perceptual choice is always relational in character; what is perceived and reacted to is fundamentally a given 'pattern,' which is conceived of as the interaction (relation) between a particular stimulus and its surround. This idea, although challenged by the behaviorist school, has been proved to be accurate, by and large; it fits with the Gestaltists' emphases on the importance of *frame of reference* in perception (cf. Sarris 1994).

Köhler's interest in the physical processes in the brain, which dates back to his years on Tenerife (or even before), generated an influential book on a special psychoneural theory of perception. This monograph (1920) dealt with the neurophysiological processing during a given experience, in an attempt to relate the laws of physics to the brain's physiochemical characteristics. It was to this end that he later directed his research on figural after-effects and so-called satiation (Köhler and Wallach 1944). His explanation offered for these much—investigated phenomena was based on the notion that direct electric currents are generated in the cortex, resulting in different active 'fields' for various perceptual phenomena (Köhler and Held 1949). However, this theory was not accepted widely by the scientific community; on the contrary, it was later refuted on both experimental and theoretical grounds in criticisms of Köhler's speculative psychophysical isomorphism.

2.1.2 *Memory and learning from a Gestalt perspective.*

Köhler expanded the Gestalt principle of perceptual and cognitive organization to the province of

memory and learning by assuming that recognition and recall are functionally another case of grouping. When a form or pattern is perceived it tends to be subsequently recognized and recalled in a well-structured, organized way; memory traces in the brain are conceived of as neurophysiological correlates of recognition and recall.

Köhler (1947) developed several criticisms of classic associationist psychology. For instance, he attacked the 'law' of association by contiguity as being merely a mechanistic principle:

> Two processes A and B happen to occur together and, whatever the nature of A and B may be, a bond is formed between them! I do not know a single law in physics or chemistry which could in this respect be compared with the law of contiguity (Köhler 1947).

Köhler argued that meaningful pairs of words are always associated more readily than are pairs of nonsense elements since the former can more easily ('naturally') be organized into a unified pattern ('Gestalt'). According to this account, recognition and recall depend on an interaction between the assumed memory traces that are relationally determined (Köhler 1951; see the reference details in Murray 1995).

2.2 *The Impact of Köhler's Work on Psychology*

Köhler's and his followers' work gradually influenced the direction and content of modern psychology. Gestalt psychology turned out to offer an important alternative to behaviorism and its related perspectives, especially in the fields of perception and cognition, since the 1960s in the twentieth century.

2.2.1 Köhler's lasting research contributions. Köhler was a research-oriented scientist aiming at an integrative and multidisciplinary development of psychology as a science. Today his research is acknowledged widely by mainstream psychology as a major contribution to the advancement of psychology as a science. His work is seen as a theoretical and experimental breakthrough in animal problem solving and in human perception and cognition. Nowadays modern psychophysical and neurobiological theory are being used to re-examine the classic Gestalt issues.

Some of the central tasks of the brain—such as figure-ground coding, perceptual organization, object identification, and the formation of object memory—concern the integration of separately coded stimulus features into coherent and meaningful object representations. These fundamental problems in explaining the basis of perception and cognition, namely the understanding of the computational or informational significance of neurobiological activity, have not been solved. Nowadays these issues are approached by the concerted efforts of different disciplines (notably, psychophysics, neurobiology, and neurocomputing).

2.2.2 Köhler on ethics and human values. Köhler also envisioned the science of psychology as an expanding area to be developed according to an ethical understanding (Köhler 1969). His analysis of values and human ethics led Köhler to reject the commonly-held assumption that values are merely subjective; on the contrary, his position offers an alternative to the relativism that seems to dominate discussions of values in science and life.

3. *Final Notes*

Köhler died on June 11, 1967, at the age of 80 at his home in Enfield, NH, USA. There has been a Wolfgang Köhler Society on Tenerife, Spain, since 1992; and the Wolfgang Köhler Memorial Lectures are organized biannually by the *Deutsche Gesellschaft für Psychologie*, Germany's national psychology association. The Wolfgang Köhler Archives are housed in the Archives of the Philosophical Society of America in Philadelphia. The complete bibliography of Köhler is published in Henle (1971).

See also: Animal Cognition; Cognitive Psychology: History; Heider, Fritz (1896–1988); Lewin, Kurt (1890–1947); Primate Behavior: Significance for Understanding Humans; Psychology: Historical and Cultural Perspectives; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Ash M G 1995 *Gestalt Psychology in German Culture 1890–1967*. Cambridge University Press, Cambridge, UK

Henle M (ed.) 1971 *The Selected Papers of Wolfgang Köhler*. Liveright, New York

Henle M 1978 One man against the Nazis—Wolfgang Köhler. *American Psychologist* **33**: 939–44

Köhler W 1917 Die Farbe der Sehdinge beim Schimpansen und beim haushuhn (The color of visual objects in the chimpanzee and in the chicken). *Zeitschrift für Psychologie* **77**: 248–55

Köhler W 1921 *Intelligenzprüfungen an Menschenaffen (Intelligence testing of apes)*. Springer, Berlin (trans. into English, 1925 *The Mentality of Apes*. Harcourt, Brace, New York)

Köhler W 1947 *Gestalt Psychology*. H Liveright, New York

Köhler W 1951 Relational determination in perception. In: Jeffress L A (ed.) *Cerebral Mechanisms in Behavior*. Wiley, New York

Köhler W 1958 Perceptual organization and learning. *American Journal of Psychology* **71**: 311–15

Köhler W 1969 *The Place of Value in a World of Facts*. Liveright, New York
Köhler W, Held R 1949 The cortical correlate of pattern vision. *Science* **110**: 414–19
Köhler W, Wallach H 1944 Figural after-effects: An investigation of visual processes. *Proceedings of the American Philosophical Society* **88**: 269–357
Murray D J 1995 *Gestalt Psychology and the Cognitive Revolution*. Harvester Wheatsheaf, New York
Sarris V 1994 Contextual effects in animal psychophysics: Comparative perception. *Behavioral and Brain Sciences* **17**: 763–64
Sarris V 1995 *Max Wertheimer in Frankfurt: Beginn und Aufbaukrise der Gestalt psychologie* (Max Wertheimer in Frankfurt: Beginnings and foundational dilemmas of Gestalt psychology). Pabst, Lengerich
Wertheimer M 1923 Untersuchungen zur Lehre von der Gestalt (Investigations concerning the theory of Gestalt' patterns). *Psychologische Forschung* **4**: 301–50

V. Sarris

Korea: Sociocultural Aspects

1. *Political and Institutional Contexts of the Anthropology of Korea*

The development of the English-language anthropology of Korea, a post-liberation (from Japanese colonialism in 1945) venture, must be understood in its political and institutional contexts.

In the cold war context of the ideological division of the Koreas, the representation of 'Korean' culture has been necessarily political. The anthropology of 'Korea' has been overwhelmingly the anthropology of South Korea. While the South Korean developmentalist state has depicted Koreanness to its own ends, decades of oppositional movements have contested official representations.

The recent and simultaneous academic institutionalization of the anthropology of Korea in and beyond South Korea constitutes another critical context. Importantly, sustained Western anthropological discourse on Korea does not predate the emergence of South Korean anthropological scholarship on Korea. Hence the English language anthropological portrayal of Korea reflects considerable scholarly exchange between South Korean, Korean émigré, and non-Korean anthropologists. Also, the vast majority of US anthropological dissertations on South Korea are the works of South Korean or recent émigré (South) Korean American scholars. This history aside, however, much of South Korea's anthropological history squarely reflects the Western-centered development of social cultural anthropology. Indeed, the vast majority of anthropologists in South Korea have received their Ph.D. abroad, primarily in the United States.

2. *Village Ethnography: Korean Culture/Korean Development*

In keeping with anthropological history, the early ethnography of South Korea turned to villages as exemplary of Korean culture. Many early works sought to distinguish 'Korean' culture from Chinese and Japanese culture. Indeed, anthropology and folklore in Korea often wrote against colonial discourses which denied the historical legitimacy and cultural autonomy of 'Koreanness' (Janelli 1986). In a different vein, an interest in Korean distinctiveness outside of Korea in part reflects the marginalization of Korean studies in the Western study of East Asia.

The early portrayal of Korean culture centered on several related dichotomies: elite (*yangban*, referring loosely to membership in a genealogy with office-holding ancestors) vs. popular/non-elite culture; Confucianism vs. shamanism/popular religion; and male vs. female. These dichotomies turn on understandings of the (Chinese) Confucianization of Korea (both ideologically and institutionally) over the last four centuries. The cornerstones of Confucian practice and ideology include patrilineal descent and primogeniture, structural separation and subordination of women, and text-based ritual observance, including particularly the veneration of patrilineal ancestry. It is critical to note that the early anthropology of Korea was most comfortable locating 'Korean culture' in villages organized as *yangban* patrilineages.

In the consideration of the dichotomies above there are points of considerable disagreement reflecting theoretical developments in anthropology. While some anthropologists have been comfortable in contrasting elite hegemonic Confucian culture with popular indigenous culture (e.g., Brandt 1990 (1971), Osgood 1951), others have insisted that lived reality obscures or even obviates these scholarly constructs (e.g., Chun 1984, Kendall 1985, Janelli and Janelli 1982). Nonetheless, anthropology has continued to lend considerable attention to the persistent presence of these very categories in Korean public and private discourse (see, for example, the Korean-language writings of Kwang-Kyu Lee). These discussions echo longstanding debates in anthropology: on culture (norm/prescription/ideology vs. practice/reality), social structure (organization vs. cognitive models), gender (isolated vs. integrated worlds), and history (cultural continuity vs. the invention of tradition).

A number of early post-liberation anthropological and sociological works on Korea consider village variation, foremost the distinction between *yangban*-dominated vs. non-dominated villages (or consanguineous vs. non-consanguineous villages) (see, for example, the writings of prominent Korean ethnographer Taik-Kyoo Kim). For many investigators these distinctions took on interest in relation to the apparent collapse of the old status system under the influences of land reform (1950), the Korean War,

modern education, technological transformation, changes in economic organization, and burgeoning rural exodus. This literature, however, asks very different questions of villages, looking less for cultural essences, and more to village 'development' or 'modernization' in a rapidly changing South Korea.

2.1 Korean Cultural Patterns and Tensions

Cornelius Osgood's (1951; fieldwork, 1947) brief fieldwork in occupied South Korea becomes the first post-liberation village ethnography. In an attempt to distinguish the uniqueness of Korea, Osgood examines the culture and personality of the Korean 'man'. He writes of an 'innately emotional' Korean personality strained under the strictures of Chinese Confucian formalism. Vincent Brandt's (1990 (1971); fieldwork, 1966) village ethnography seeks to reconcile village communalism and egalitarian social relations with Korea's dominant ideological system with its rigid lineage, age, and occupational hierarchies. Like Osgood, Brandt is thus interested in cultural tensions in Korea. Prescient of later works, Brandt struggles over how to portray these tensions: an 'other' ideological system that competes alongside the Chinese, Confucian elite system; practices and values that comprise a 'different level of consciousness'; the ideational vs. the behavioral; a lower-class tradition vs. the dominant one; or in terms of ideological and cultural village variance (e.g., in accordance with *yangban* concentration, and village mode of subsistence).

2.2 Questioning Dichotomies

Roger and Dawnhee Yim Janelli's (1982; fieldwork, 1973–7) ethnography of ancestor worship rejects the facile equation of Korean ancestor worship, social organization (including gender organization), and filial piety with Confucianism. Rather, the Janellis analyze ritual ideas and practice in relation to Korean social organization—reflective of, but never reducible to, Confucianism. The Janellis are particularly interested in how the dependent, passive ancestor of prestigious formal ancestor worship can be reconciled with the active and potentially hostile ancestor of domestic ancestral rites and shamanism. They correlate women's much greater interest in ancestral malevolence with their structural vulnerability in Korean patrilocal, patrilineal, and patriarchal social organization. Shamanism and ancestor worship emerge as separate and internally coherent systems, neither wholly Confucian nor anti-Confucian.

Kyung-soo Chun (1984; fieldwork, 1975–6) was similarly interested in social practice, considering reciprocity as an organizing principle in a Korean island village. He scrutinized 'community' boundaries by considering social relationships beyond kinship, thus examining 'practical' rather than ideal patterns.

Laurel Kendall's (1985; fieldwork, 1977–8) ethnography of village shamanism and women's domestic ritual employs the insights of feminist anthropology to challenge similarly the dichotomies in question here. Writing against classical anthropological portrayals of women's ritual as revenge or resistance, Kendall argues that women's ritual domains contrast with, but by no means subvert, men's practices and objectives. Kendall nonetheless analyzes women's ritual roles and practices as a window through which Korean society as a whole can be reconsidered. She highlights social features that challenge the limits of formal patrilineal organization and ideology: the house and household (*chip*) as women's space and ritual domain; women's power as matrons of autonomous households; and women's roles as the link between bilateral kinship groups.

2.3 Village Variation and 'Development'

Village variation took on two meanings in the early anthropology of Korea. First, variation allowed some anthropologists to explore gaps between 'ideal' structures and social realities; important here are writings on villages that do not feature the dominant cultural pattern or ideology of patrilineal kinship organization. Second, village variation allowed for controlled comparison of developmental outcomes. Sang-Bok Han (1977; fieldwork, 1968–9), for example, examines ecological adaptation in three fishing communities distinguished for not having permanent patrilineal kinship organization (see also the writings of rural sociologists Man-Gap Lee and Ki-hyuk Pak). Building on this legacy, Clark W. Sorensen (1988; fieldwork, 1976–7, 1983) examines peasant households' adaptation to rapid industrialization, considering transformations of farm labor, village economy, land tenure, and out-migration.

3. The Discursive Turn: The Ethnography of Social-cultural Contests and Class/Gender

As anthropology has followed the vast majority of South Koreans out of the village to urban and modern organizational venues, 'traditional' or 'Korean' culture has been revisited. In keeping with their anthropological times, this next generation of Korean ethnographies is less interested in characterizing Korean culture or considering Korean development: it is committed instead to portraying the various discursive contexts and contests in which South Koreans live and make sense of their lives. Important to this endeavor is attention to the state—particularly its nationalist and development rhetorics—as consti-

tutive of personal subjectivities and social life. It is perhaps because of South Korea's decades of political authoritarianism that Korean anthropology features the state to an extent that many other anthropologies do not. These works challenge the facile portrayal of cultural continuity (from the village to the city or modern organization), considering instead the constructed character of cultural discourses. Persistent parameters of cultural and value difference in the contemporary anthropology of Korea have been class and gender.

3.1 *The Politics of Culture*

The new generation of works pays scrupulous attention to the political-economic context of social practice and cultural life. Exemplary is Chungmoo Choi's (1995) examination of competing appropriations of shamanism and other cultural performances by both the state and oppositional groups. In a similar vein, Roger Janelli (with Dawnhee Yim) (1993; fieldwork, 1987) portrays daily company life in a major corporation in the context of corporate and state representations of 'Korean culture'. Highlighted are the ways in which company discourses distort the reality of Korean family, lineage, and village life.

A number of works situate contemporary practice in the *longue durée* of anthropologically informed cultural history. Kendall (1996), for example, sets her ethnography of marriage in South Korea in the context of the twentieth-century sweep of Korean cultural discourses of tradition and modernity, especially apt because wedding styles are often so easily labeled one or the other.

In the 1980s, a number of anthropologists turned their attention to the cultural and historical coordinates of social movements. Abelmann (1996; fieldwork, 1987–8), for example, examines the range of urban and rural actors in a farmers' movement in order to consider struggles over history and memory (see also H. M. Kim 1997 and S. K. Kim 1997 on the women's labor movement). Also noteworthy are new ways of reading cultural practices: exemplary is Seong Nae Kim's (1995) reading of the history of ideological strife in Cheju Island through the dreams of a shaman.

3.2 *Class and Gender*

Drawing on the theory and methods of anthropology's discursive turn, anthropologists of Korea have theorized the contingent, complex character of class, and its articulation with gender.

A central debate has considered how to conceptualize the hegemony of middle-/upper-class norms and values, and the counter-hegemony of working-class life and culture. Many works portray the lived reality of class in such a way that it is impossible either to romanticize unbridled class struggle on the one hand, or to consider all Koreans as dupes of hegemonic culture, on the other hand. Exemplary is Seung Kyung Kim's (1997; fieldwork, 1986–8) examination of the lives of women factory workers in an export industrial zone. This work details women's contribution to the labor movement and recognizes prevailing views that obscure these women's powerful desires for social mobility.

Understanding that class and gender are inextricable from one another, a number of anthropological studies examine the workings of class in women's lives, and women's roles in the making and social reproduction of class, asserting particularly the importance of women's work beyond the formal economy (Myung Hye Kim 1992). Anthropologists have also wrestled with how to understand the persistence of hegemonic patriarchal ideologies and social organization. In her examination of women's reproductive ideologies across class, for example, Eun Shil Kim (2001) considers the complex and subtle play of patriarchy, class dominance, and resistance (see also H. M. Kim 1997). Also important have been analyses of the discursive production of femininity and motherhood over time and across class (see also the Korean-language writings of Oakla Cho and Taek-Lim Yoon). Particularly influential have been the prolific Korean-language writings of anthropologist Haejoang Cho (2001), who has argued for the multivocality of 'traditional' cultural legacies for women (see also Soh 1991).

4. *'Korean' Ethnography into the Future*

As anthropology has come increasingly to question its objects of study, many have begun to look beyond South Korea to reconceptualize 'Korea'. Richard Grinker (1998), for example, considers South Korean cultural and political identity through South Korean state and popular representations of North Korea (see also the prescient work of Choong Soon Kim (1988) on divided families). Also noteworthy are burgeoning transnationally focused studies of Korean diasporic communities that demand reconceptualization of South Korea and the Koreas, including studies of Koreans in China and their 'return' to Korea as migrant laborers; the Korean American transnation; North Koreans in Japan; South Korean multinational corporations abroad; and global social activism concerning 'comfort women'. Given the escalating interest in the transnational production of Korean modernity, still surprisingly unexplored is Japanese colonial ethnography of Korea, as a window on both the Japanese colonial project and post-liberation anthropological knowledge. Similarly, the missionary encounter in Korea remains little studied in and beyond anthropology.

The above horizons aside, it is important to note that for the anthropology of Korea there remain many domains almost entirely unexplored ethnographically

(in English), including all levels of schooling; childhood and adolescence; Christianity and Buddhism; the military (with its universal conscription for men); sexuality; leisure (including tourism); the media; and medical systems.

Finally, it is important to note that the anthropology of Korea has become increasingly bold in asserting its contributions to general anthropological debate. The field is ever more keenly aware of the multiplicity of global processes that converge in the Koreas, and of the important anthropological issues presented by the Koreas' postcoloniality and national division.

See also: Area and International Studies: Development in Southeast Asia; East Asia, Religions of; East Asian Studies: Culture; East Asian Studies: Economics; East Asian Studies: Geography; East Asian Studies: Politics; East Asian Studies: Society

Bibliography

Abelmann N 1996 *Echoes of the Past, Epics of Dissent: A South Korean Social Movement*. University of California Press, Berkeley, CA

Brandt V 1990 (1971) *A Korean Village: Between Farm and Sea*. Waveland Press, Prospect Heights, IL

Cho H 2001 In: Kendall L (ed.) *Under Construction: The Gendering of Modernity, Class, and Consumption in the Republic of Korea*. University of Hawaii Press, Honolulu, HI

Choi C 1995 The Minjung Culture Movement and the construction of popular culture in Korea. In: Wells K (ed.) *South Korea's Minjung Movement: The Culture and Politics of Dissidence*. University of Hawaii Press, Honolulu, HI

Chun K 1984 *Reciprocity and Korean Society: An Ethnography of Hasami*. Seoul National University Press, Seoul, South Korea

Grinker R R 1998 *Korea and its Futures: Unification and the Unfinished War*. St. Martin's Press, New York

Han S 1977 *Korean Fisherman: Ecological Adaptation in Three Communities*. Seoul National University Press, Seoul, South Korea

Janelli R L, Janelli D Y 1982 *Ancestor Worship and Korean Society*. Stanford University Press, Stanford, CA

Janelli R L 1986 The origins of Korean folklore scholarship. *Journal of American Folklore* **99**(391): 24–49

Janelli R L, Yim D 1993 *Making Capitalism: The Social and Cultural Construction of a South Korean Conglomerate*. Stanford University Press, Stanford, CA

Kendall L 1985 *Shamans, Housewives and Other Restless Spirits: Women in Korean Ritual Life*. University of Hawaii Press, Honolulu, HI

Kendall L 1996 *Getting Married in Korea: Of Gender, Morality and Modernity*. University of California Press, Berkeley, CA

Kendall L (ed.) 2001 *Under Construction: The Gendering of Modernity, Class, and Consumption in the Republic of Korea*. University of Hawaii Press, Honolulu, HI

Kim C S 1988 *Faithful Endurance: An Ethnography of Korean Family Dispersal*. University of Arizona Press, Tucson, AZ

Kim E S 2001 In: Kendall L (ed.) *Under Construction: The Gendering of Modernity, Class, and Consumption in the Republic of Korea*. University of Hawaii Press, Honolulu, II

Kim H M 1997 Gender/sexuality system as a labor control mechanism: gender identity of Korean female workers in a US multinational corporation. *Korea Journal* **37**(2): 57–70

Kim M H 1992 Late industrialization and women's work in urban South Korea: an ethnographic study of upper-middle-class families. *City and Society* **6**(2): 156–73

Kim S K 1997 *Class Struggle or Family Struggle?: The Lives of Women Factory Workers in South Korea*. Cambridge University Press, New York

Kim S N 1995 In: Wells K (ed.) *South Korea's Minjung Movement: The Culture and Politics of Dissidence*. University of Hawaii Press, Honolulu, HI

Osgood C 1951 *The Koreans and their Culture*. Ronald Press, New York

Soh C S 1991 *Women in Korean Politics*. Westview Press, Boulder, CO

Sorensen C W 1988 *Over the Mountains are the Mountains: Korean Peasant Households and Their Adaptations to Rapid Industrialization*. University of Washington Press, Seattle, WA

Wells K (ed.) 1995 *South Korea's Minjung Movement: The Culture and Politics of Dissidence*. University of Hawaii Press, Honolulu, HI

N. Abelmann

Korsakoff's Syndrome

Organic amnesia is a syndrome, caused by specific kinds of brain damage, in which an impairment in the ability to recall and recognize recently experienced facts and personally experienced events (anterograde amnesia) is accompanied by a similar memory deficit for premorbidly encountered facts and events (retrograde amnesia). These memory deficits occur in the face of relatively preserved intelligence and ability to recall information encountered in the previous few seconds (short-term or working memory) so that the memory deficits are disproportionate to other cognitive impairments. Korsakoff's syndrome, in which memory impairment results primarily from depletion of thiamine, is an important subvariety of organic amnesia.

1. Origin of the Name 'Korsakoff's Syndrome'

Although the syndrome is named after the Russian doctor Korsakoff, earlier authors had alluded to the disorder and it was Lawson in 1878 who first gave a formal description of a disorder in which there was a severe loss of recent memory that was frequently associated with chronic alcoholism. Lawson also distinguished between cases with relatively selective

memory loss and those in whom alcohol consumption caused a progressive dementing condition, which is nowadays referred to as 'alcoholic dementia.' Between 1887 and 1891, Korsakoff then gave the classic description of a much larger series of amnesia cases, two-thirds of whom were chronic alcoholics. The other third of the cases comprised individuals with several etiologies including neoplasm, carbon-monoxide poisoning, and diabetes, as well as persistent vomiting. Korsakoff identified how the memory disorder varied in severity, and that it had a retrograde as well as an anterograde component. He also noted that patients often falsely recollected information, a disorder known as confabulation. Korsakoff further noted that the memory disorder was commonly preceded by a state of confusion, accompanied by eye movement disorders such as nystagmus, as well as gait disturbance (ataxia).

This triad of symptoms had first been described in 1881 by Wernicke, who also reported that the syndrome, which now bears his name, was associated with alcoholism. Although the syndrome is often incomplete, it is not diagnosed unless at least one of the complexes of motor symptoms is present. There has been controversy about the frequency with which the appearance of the usually chronic memory disorder is preceded by Wernicke's encephalopathy. Although the issue is not fully resolved, it is clear that in many cases of alcoholic origin, perhaps over half, Wernicke's encephalopathy is not detected prior to the onset of amnesia. Consistent with this, it is clear that many of these cases have a relatively insidious onset often lasting over a year, and that these cases are weakly associated with Wernicke's encephalopathy. In contrast, more acute onset cases frequently show the Wernicke ocular signs. Emergence of the memory deficit may be preceded by acute coma, Wernicke's encephalopathy (both of which are life threatening), or it may emerge gradually without either of these prodroma being detected. There even appear to be cases where the memory disorder is not identified in life, but is suspected retrospectively at postmortem on the basis of neuropathology and a history of alcoholism. No evidence indicates that the presence or absence of detectable Wernicke's encephalopathy affects the memory disorder or its underlying neuropathology differently. It is, however, still possible to argue that the persistent memory disorder, caused by thiamine deficiency, is always preceded by one or more episodes of Wernicke's encephalopathy, but that these episodes may often go undetected. If correct, the brain damage underlying Wernicke's encephalopathy should be found in Korsakoff's syndrome regardless of whether or not it is preceded by a detectable Wernicke's encephalopathy. This encephalopathy does not always lead to a persisting amnesia.

Although the term 'Korsakoff's syndrome' is sometimes loosely used as a synonym for amnesia, it is more correctly reserved for those cases of relative persistent amnesia which are triggered by thiamine depletion, the most prominent cause of which is chronic alcoholism. Alcoholism is not the only cause of thiamine depletion, however, and the evidence about this cause as well as other causes of thiamine depletion has only emerged gradually over the last century. Diagnosis depends on identifying clinical and neuropsychological evidence of a disorder largely confined to memory and learning in which there is a history of alcoholism or nutritional problems.

2. *Evidence for the Role of Thiamine Depletion in Organic Amnesia*

Both Wernicke's encephalopathy and amnesia can result from a range of nutritional and other disorders related to malabsorption, which include self-starvation, intravenous feeding, particularly in the presence of glucose, persistent vomiting, surgical removal of parts of the gastrointestinal tract, and various carcinomas of the alimentary canal. De Wardener and Lennox (1947) notably observed Wernicke's encephalopathy in 52 malnourished prisoners in Southeast Asia during World War II. Over 60 percent of the prisoners also developed memory loss within six to fourteen weeks of the start of captivity, often preceded by diarrhea and vomiting. This corresponds to the time it takes beriberi, which is caused by thiamine deficiency, to develop. Symptoms of other vitamin deficiencies take longer to develop. Also, as is now well known, the symptoms responded very well to thiamine treatment when this was available and could be instituted early before brain damage was irreversible. In recent years, cases of amnesia caused by malnutrition have been reported relatively rarely, although there may be under-reporting given the prevalence of anorexia. Nevertheless, Korsakoff's syndrome with a typical neuropathology has been reported to result from dieting, particularly when this is followed by high consumption of carbohydrates, which presumably overloads the already marginal thiamine stores.

A common feature of the alcoholic and nonalcoholic patients who develop the memory deficits and neuropathology typical of Korsakoff's syndrome is very probably thiamine deficiency. Not only is alcoholism often associated with poor nutrition, which causes thiamine deficiency, but alcohol directly impairs thiamine metabolism in many ways (e.g., by distruption of the gastrointestinal transport of the vitamin as well as its storage and use). Although thiamine treatment may be less effective with alcoholic than nonalcoholic cases, this could be related to incidental factors such as the patients' ages and when the treatment was introduced. Nevertheless, it is often argued that some aspects of alcoholic Korsakoff's syndrome may be related to the direct neurotoxic effects of alcohol, and that nicotinic acid deficiency

may also contribute to development of the disorder in some cases.

Korsakoff's syndrome is a rare complication of alcoholic abuse, although its frequency seems currently to show marked regional variations, as it is apparently much more frequent in Australia than in the USA and the UK. This variation could relate to differences in factors such as binge drinking or the amount of thiamine intake. However, the syndrome's rarity has led to the suggestion that it is associated with a hereditary disorder affecting the enzyme transketolase, which needs thiamine pyrophosphate (TPP), the active form of thiamine, as a cofactor. The evidence about this is conflicting, although it is likely that some kind of transketolase abnormality is related to the development of Korsakoff's syndrome. Exactly how thiamine deficiency causes the brain damage that leads to amnesia is not known. However, TPP is involved in DNA synthesis, three enzymatic reactions vital for glucose metabolism, and the production of several neurotransmitters including acetylcholine, glutamate, gamma-aminobutyric acid (GABA), noradrenaline, and serotonin. All three of the thiamine-dependent enzymes have been found to be depressed in patients with Wernicke-Korsakoff syndrome, but not in other alcoholics. Some evidence suggests that thiamine deficiency depresses oxidative glucose metabolism more severely in brain regions implicated in the syndrome, although this does not exclude the possibility that the deficiency may also disrupt specific neurotransmitters as well.

2.1 *Evidence of Neurotransmitter Abnormalities in Korsakoff's Syndrome*

No work has related glutamate and GABA abnormalities to Korsakoff's syndrome. There is, however, evidence that reduced acetylcholine turnover occurs when thiamine activity is low, and that this may also cause reduced acetylcholine production. It has also been argued that Korsakoff patients suffer a nearly 50 percent reduction of neurons in their nucleus basalis, a basal forebrain structure which contains ascending projections that use acetylcholine as a neurotransmitter. However, other Korsakoff patients have shown minimal changes in basal forebrain structures at postmortem. This relates to an important feature of the syndrome, which is that the location of brain damage is very variable across patients so presumably only some will show marked abnormalities in cholinergic neurons. Furthermore, although studies using anticholinergic drugs have shown that the anterograde amnesia of patients who have cholinergic deficiencies may be partially caused by the transmitter deficit, their retrograde amnesia cannot be. This is because anticholinergic drugs do not disrupt premorbid memory.

Five patients with alcoholic Korsakoff's syndrome have been reported to show improved memory when given a serotonin reuptake inhibitor, and a similar effect has been reported in alcohol-treated volunteers. However, the patient study needs replicating and whether a serotonergic abnormality is linked to the syndrome remains unproved. The same can be said about the possible involvement in the syndrome of a noradrenergic disorder, as the relevant evidence is conflicting. Some work has found that levels of the noradrenergic metabolite, 3 methoxy 4-hydroxy phenylglycol, were reduced in the cerebrospinal fluid of Korsakoff patients, that these levels correlated with memory test performance, and that treatment with adrenergic agonist drugs produced a small improvement in memory. However, other work has not replicated these findings and animal studies suggest that noradrenergic neurons are more concerned with the mediation of attentional and arousal processes. If the animal studies are correct, then any effects on memory must be secondary to attentional disturbances. As with cholinergic abnormalities, noradrenergic deficiencies are also a variable feature of Korsakoff's syndrome because the level of neuronal loss in the locus coeruleus is highly variable. This small brainstem nucleus is the major source of noradrenergic projections to the forebrain regions implicated in memory.

3. *The Neuropathology of the Syndrome*

Postmortem examination of patients who have died during the acute stage of Wernicke's encephalopathy, or following a relatively prolonged Korsakoff's syndrome, show a similar, although variable, pattern of lesions that is distributed in a bilaterally symmetrical fashion within the midline diencephalon, brain stem, and cerebellum. Pathology is most common in the mammillary bodies, midline thalamic nuclei such as the dorsomedial and anterior nuclei and medial pulvinar, the periventricular and periaqueductal grey, the walls of the third ventricle, the floor of the fourth ventricle, and the cerebellum. Cortical atrophy with associated neuronal loss and increased hydration, which may be particularly prominent in the frontal lobes, has also been reported in a minority of cases with the full syndrome. This atrophy is similar in alcoholic patients regardless of whether or not they have Korsakoff's syndrome. In the acute stages of the disease, gross lesions are usually small and circumscribed with nonhemorrhagic softening or, less commonly, hemorrhages, whereas in the chronic disease atrophy is prominent, but hemorrhages may not be present. Microscopic examination may be necessary to confirm the presence of Korsakoffian pathology because gross lesions are not identifiable in all patients.

The location of the pathology that underlies the triad of symptoms which Wernicke's encephalopathy comprises is not fully resolved. The ocular symptoms perhaps involve damage to the tectal area as well as to several brain stem nuclei including the oculomotor, abducens, and vestibular nuclei, whereas the ataxia

involves cerebellar damage. Identifying the location of the lesions underlying the memory deficits of Korsakoff's syndrome has proved particularly difficult because of the range and variability of the structures that may be compromised in this disorder. Also, many non-Korsakoff alcoholics have milder memory deficits. Victor and his colleagues (1989) have reported that all of their alcoholic Korsakoff cases in whom the dorsomedial thalamic nucleus was affected showed amnesia, whereas a few patients in whom this nucleus was intact, but in whom the mammillary bodies were affected, did not show clinically detectable amnesia. There is, however, good evidence from both animal and human lesion studies that damage to either the dorsomedial thalamic nucleus or the mammillary bodies is sufficient to cause some memory deficits, so, when present, damage to either of these structures is likely to be contributing to the amnesic deficit in Korsakoff's syndrome.

Korsakoff patients with an etiology of chronic alcoholism have nevertheless shown severe amnesia even when postmortem examination only found paratenial nucleus damage anterior and dorsal to the intact dorsomedial thalamic nucleus. These patients also had mammillary body damage as well as other cortical and subcortical lesions. Korsakoff patients show variable damage to structures of Papez circuit that includes demyelination of the fornix as well as atrophy of the mammillary bodies and anterior thalamus. However, damage in alcoholic Korsakoff patients compared to alcoholic Wernicke's cases with milder memory problems has only been convincingly shown to be worse in one region: the anterior thalamus. Damage to Papez circuit structures is sufficient to cause a mild amnesia. It remains unknown how severe the amnesia produced by a combination of this damage with damage to other diencephalic nuclei, such as the dorsomedial nucleus and mammillary bodies, actually is. Most probably, extradiencephalic damage also contributes to the severe amnesia found in some patients with Korsakoff's syndrome.

3.1 Neuroimaging the Brains of Korsakoff Patients in Life

Magnetic resonance imaging (MRI) has shown that, relative to healthy people, alcoholic Korsakoff patients show widening of the third ventricle, the Sylvian fissure, and frontal sulci, as well as reductions in thalamic density. The evidence is conflicting about whether these changes are greater in Korsakoff's syndrome than in non-Korsakoff alcoholic patients, although the postmortem data would suggest that it should not be. Similarly, there is conflict about whether measures of frontal and thalamic atrophy correlate with memory measures in the Korsakoff patients. This conflict may be resolved by improvement of MRI measurement techniques in the future. Resolving this issue is important because it has often been argued that frontal atrophy may contribute to the Korsakoffian memory impairment.

Functional imaging has tended to show a picture of reduced cortical activation in Korsakoff's syndrome, which, although prominent in frontal regions, extends into posterior neocortical regions, but minimally into the medial temporal regions where lesions are known to cause amnesia. One possibility is that the reduced neocortical activation does not disrupt basic processing ability, but impairs the ability of neocortical structures to store fact and event information. Further investigation needs to determine whether the extent of reduced neocortical activation correlates with memory.

4. Neuropsychology of Korsakoff's Syndrome

The degree of impairment of intelligence and other, noncognitive functions varies widely across Korsakoff patients. Patients often show impairments of visuoperceptive and visuospatial functions, which are also found in nonamnesic alcoholics and may be caused by neocortical atrophy. Korsakoff patients are also usually impaired on several tests of executive function, particularly when the ability to shift strategy is tapped. Such executive functions are believed to depend primarily on the integrity of the frontal lobes.

Immediate or working memory as tested by the ability to repeat a string of items is usually held to be intact in Korsakoff's syndrome. However, subtle span deficits have been claimed and performance on short-term forgetting tasks has been highly variable across studies. Evidence suggests that this variability may relate to the degree of frontal lobe dysfunction. Unrelatedly, the apathy and lack of insight often reported in patients may also relate to frontal lobe dysfunction.

The anterograde amnesia of Korsakoff's syndrome comprises impaired recall and recognition of recently studied facts and events. Although recognition is lost at a normal rate even over short delays, there is evidence that free recall of complex associative information is lost pathologically fast over such short delays. There is also evidence that recall of semantically related words is more impaired than equivalent recognition after short delays, and that memory for temporal and other aspects of context is more impaired than item recognition. It remains controversial whether these disproportionate deficits, which are also found in amnesics with medial temporal lobe lesions, relate to frontal dysfunction, the severity of amnesia, or both.

Like other amnesics, Korsakoff patients also often show intact implicit or unaware memory. Thus, they show normal classical conditioning when the conditioned and unconditioned stimuli overlap in time, normal acquisition and retention of perceptual, motor, and cognitive skills, and normal priming. Priming is shown when subjects respond differently to previously

shown verbal and visual stimuli, even when these are not consciously recognized. However, as with other amnesics, Korsakoff patients may show impaired priming for certain complex stimuli that were novel prior to study.

Although, like other amnesics, Korsakoff patients show normal semantic memory when this is defined as memory for factual information that was massively overlearnt premorbidly, they are impaired at learning new word definitions and public information even when given reasonable exposure over a long postmorbid time. Their retrograde amnesia is characterized by an impaired ability to recall and recognize personal and public incidents, and facts acquired as memories up to several decades prior to the clear onset of their memory disorder. Although there is typically a temporal gradient in which earlier acquired memories are less affected, amnesia usually extends back at least 25 to 30 years.

Confabulation is rare in Korsakoff's syndrome except at the initial confusional stage when spontaneous and florid confabulation may be found. Patients may make intrusional errors when retrieving autobiographical memories, but the rare, more florid confabulation that persists beyond the confusional stage probably reflects a disorganized and disinhibited retrieval of memories and their associations. This is believed to be caused by the frontal lobe dysfunction that may be severe in some Korsakoff patients. More typically, patients show a nonfluent pattern of retrieval that is characteristic of patients with frontal lobe pathology.

Retrograde and anterograde memory deficits share little variance in patients, and it has been argued that the functional deficits are relatively unrelated and that retrograde amnesia, at least for more remote premorbid memories, is mainly caused by frontal lobe dysfunction, which produces disorganization of retrieval processes. Although Korsakoff's syndrome may be associated with a more severe temporal order memory deficit than is typically found in amnesics with medial temporal lobe lesions, it is unclear whether this too may not result from Korsakoff patients' frontal lobe dysfunction. It seems most likely that a failure to consolidate certain kinds of associations lies at the heart of the anterograde amnesia found in the syndrome, and that this is mainly caused by midline diencephalic lesions.

See also: Alcohol-related Disorders; Alcoholism: Genetic Aspects; Amnesia; Amnesia: Transient and Psychogenic; Learning and Memory, Neural Basis of

Bibliography

Brody B A 1996 The Wernicke-Korsakoff syndrome: Neuropathology and pathogenic basis. *International Journal of Neuroradiology* **2**: 216–30

Butters N, Cermak L S 1980 *Alcoholic Korsakoff's Syndrome: An information Processing Approach to Amnesia*. Academic Press, New York

De Wardener H E, Lennox B 1947 Cerebral beriberi (Wernicke's encephalopathy): review of 52 cases in a Singapore PoW hospital. *Lancet* **1**: 11–17

Harding A, Halliday G, Caine D, Kril J 2000 Degeneration of anterior thalamic nuclei differentiates alcoholics with amnesia. *Brain* **123**: 141–54

Harper C 1998 The Neuropathology of alcohol-specific brain damage, or does alcohol damage the brain? *Journal of Neuropathology and Experimental Neurology* **57**: 101–10

Harper C G, Sheedy D L, Lara A I, Garrick T M, Hilton J M, Raisanen J 1998 Prevalence of Wernicke-Korsakoff syndrome in Australia: Has thiamine fortification made a difference? *Medical Journal of Australia* **168**: 542–5

Hunt W A, Nixon S J (eds.) 1993 *Alcohol-induced Brain Damage*. Research Monograph 22. NIH Publication no. 93-3549. The Institute, Rockville, MD

Kopelman M D 1995 The Korsakoff syndrome. *British Journal of Psychiatry* **166**: 154–73

Mayes A R 1988 *Human Organic Memory Disorders*. Cambridge University Press, Cambridge, UK

Mayes A R, Meudell P R, Mann D, Pickering A D 1988 Location of lesions in Korsakoff's syndrome: Neuropsychological and neuropathological data on two patients. *Cortex* **24**: 367–88

Parkin A J, Leng N R C 1993 *The Neuropsychology of the Amnesic Syndrome*. Lawrence Erlbaum Associates, Hillsdale, NJ

Talland G A 1965 *Deranged Memory*. Academic Press, New York

Victor M, Adams R D, Collins G H 1989 *The Wernicke-Korsakoff Syndrome and Related Neurological Disorders due to Alcoholism and Malnutrition*. 2nd edn. F A Davis Company, Philadelphia

Witt E D 1985 Neuroanatomical consequences of thiamine deficiency: A comparative analysis. *Alcohol and Alcoholism* **20**: 201–21

A. Mayes

Kraepelin, Emil (1856–1926)

1. Biographical Note

Emil Kraepelin was born on February 15, 1856 in Neustrelitz (Mecklenburg, West Pomerania, Germany). He studied medicine in Leipzig and Würzburg from 1874 until 1878. During his medical studies in Würzburg, Kraepelin worked as a guest student at the psychiatric hospital under the directorship of Franz Rinecker (1811–83). He began his professional life in 1878 working with Bernhard von Gudden (1824–86) at the District Mental Hospital in Munich, where he stayed until 1882. Kraepelin then moved to Leipzig, where he worked with Paul Flechsig (1847–1929) and Wilhelm Erb (1840–1921). He was promoted to university lecturer there in 1883.

His lifelong personal and scientific relationship with Wilhelm Wundt (1832–1920) began in Leipzig. Encouraged by Wundt, Kraepelin wrote his *Compendium of Psychiatry* in 1883, the precursor of his influential textbook *Psychiatry*, which was to be published in nine editions between 1883 and 1927. Kraepelin stayed in contact with Wundt (see *Wundt, Wilhelm Maximilian* (1832–1920)) by correspondence and paid him several visits before Wundt's death in 1920; in many of his publications he emphasized and acknowledged the importance of this relationship.

In 1884 Kraepelin married Ina Schwabe. After a short period of employment in Leubus (in Silesia) and Dresden, Kraepelin was appointed professor of psychiatry at the University of Dorpat in 1886. In 1891 he took over the chair of psychiatry at the University of Heidelberg. From 1903 until 1922 Kraepelin was ordinary professor of psychiatry in Munich where, in 1904, he opened the new building of the psychiatric hospital of the Ludwig Maximilian University, the main part of which is still in use. Despite the difficult conditions caused by World War I, Kraepelin founded the *Deutsche Forschungsanstalt für Psychiatrie* (the German Research Institute for Psychiatry) in Munich in 1917 to encourage and improve psychiatric research (Weber 1991). During his time in Munich, Kraepelin's colleagues at the university hospital and the *Forschungsanstalt* included Alois Alzheimer (1864–1915), Franz Nissl (1860–1919), Korbinian Brodmann (1868–1918), Walter Spielmeyer (1879–1935), August Paul von Wassermann (1866–1925) and Felix Plaut (1877–1940)—to mention but a few. In 1924 the *Deutsche Forschungsanstalt* was integrated into the Kaiser Wilhelm Society and became in 1945 the Max Planck Institute for Psychiatry as part of the Max Planck Society. Emil Kraepelin died in Munich on October 7, 1926.

2. *The Concept of Psychiatric Research*

Wilhelm Griesinger (1817–68) had marked the turning point in nineteenth-century psychiatry by calling for thorough clinical and pathophysiological research based on the premise that 'mental illness is a somatic illness of the brain.' But Griesinger's theory was by no means as simple as this one statement, which is so often quoted. He held clearly differentiated views on the problem of somato- and psychogenesis, although favouring the first in the case of what was later to be called 'endogenous psychoses.'

Karl Ludwig Kahlbaum (1828–99) continued the traditions of French psychopathology as represented by J. Falret and A.L.J. Bayle and developed a clinically orientated research method in the second half of the nineteenth century in Germany, taking the course of illness into special consideration. This, like Griesinger's theory, was believed to be a victory over the speculative concepts of the so-called 'romantic medicine' and even more so an argument for a critical distance to the position of the harsh and unreflected 'somaticists,' who were no less dogmatic than some of the 'psychicists.' Kahlbaum had clearly recognized the methodological differences between pathological-anatomical and clinical-psychopathological work. With 'progressive paralysis of the insane' as an example he explained the route from the 'syndrome-course unit' (*Syndrom-Verlaufs-Einheit*) to the—postulated—etiologically-based 'disease unit' (*Krankheitseinheit*).

The influence on Kraepelin of the founder of experimental psychology, Wilhelm Wundt, can hardly be overestimated. Wundt's aim was to establish psychology as a kind of natural science which relied on experimental data. He criticized the highly speculative approach of 'philosophy of nature' in the sense of Schelling and Schleiermacher, but did not agree with materialism or association psychology in the sense of Herbart either. At least in his earlier writings, Wundt favored a parallelistic point of view in the mind–body problem. Experimental research—and Wundt's ideas were obviously the most fascinating for the young Kraepelin— may successfully be used in natural sciences as well as in psychology, without ignoring the epistemological differences between the mental and the physical. Kraepelin modified Wundt's ideas by extracting what he regarded as useful for empirical research in psychiatry. That is why Wundt's psychology, viewed through the 'filter' of Kraepelin's texts, seems much more unified and straightforward than it really was. Kraepelin simplified and, so to speak, 'smoothed out' Wundt's concept, but he did not falsify it.

Emil Kraepelin was not very interested in the philosophical basis and implications of psychiatric theory and practice. His view of what (natural) science was, and what impact it had or should have on social and political developments, was, as Engstrom (1991, 1995) has shown convincingly, typical for the way natural scientists saw themselves in Germany at the turn of the last century. Despite his skeptical attitude toward theoretical considerations in psychiatry, the following major theoretical frameworks can be identified as underlying—seldom explicitly, most often very implicitly—Kraepelin's concept of psychiatry: realism, parallelism, experimental approach, naturalism.

2.1 *Realism*

In sharp contrast to the philosophical tradition of what is usually called 'German idealism,' Kraepelin, like most of his contemporaries in the field of natural sciences, believed in an independently existing 'real world' that includes other people and their healthy or disturbed mental processes. Kraepelin repeatedly pointed out that the psychiatric researcher has to

describe objectively what really exists and what 'nature presents to him'—the formulations differ, but the essence is a strictly realistic philosophy. The consequences for psychiatric nosology are evident: such realism will lead to the concept of 'natural disease entities' which exist completely independently of the researcher. The scientist describes what he finds, or—in stronger terms—describes 'given things.' His own activity in constructing scientific hypotheses or diagnostic entities is underestimated.

2.2 Parallelism

Kraepelin advocated psychophysical parallelism. Like Wilhelm Griesinger, whom he admired for his highly critical attitude toward speculative psychiatric theories, he disapproved of reductionistic materialism which identifies mental events with neurophysiological processes. Kraepelin spoke about two kinds of phenomena, somatic and psychological, which are decidedly different, but closely connected. Kraepelin defended the existence of mental phenomena against all kinds of what he, like Karl Jaspers, called 'brain mythologies.' Contrary to Wundt, however, Kraepelin, although calling himself a parallelist, did not enter the philosophical controversy about this concept. In particular, he did not critically differentiate between parallelism and interactionism and did not realize that any strictly defined parallelism makes it more than doubtful that mental life can still be regarded as an independent sphere and not just as having a one-to-one relationship with the somatic level; and this, of course, means (causal) determinism.

As a consequence of his somewhat ambivalent position in the mind–body debate, there is indeed an implicit tendency toward monism in Kraepelin's writings, particularly when considering his ideas about psychology as a natural science. It should, therefore, be emphasized that this 'monistic tendency' was definitely not a metaphysical one, but a weak version of what may be called 'methodological monism,' insofar as he decidedly favored the quantitative methods brought forward by the natural sciences (Verwey 1985).

2.3 Experimental Approach

Kraepelin argued that the psychological experiment should become a major scientific tool, not only for the understanding of disturbed mental processes, but also for the understanding of healthy mental life. Both Wundt and Kraepelin realized the difference between a physical and a psychological experiment, but the experimental design did not differ significantly in the two areas. Kraepelin seems to have considered the experimental approach a kind of guarantee for the scientific status of psychiatric research. From this it may be inferred that he rated the experimental approach higher than the mere description of clinical phenomena, although the latter method was considered to be indispensable, especially if combined with follow-up examinations. Kraepelin developed and maintained a skeptical attitude toward subjective—especially biographically-determined—aspects of mental disorders, which could not be studied experimentally.

2.4 Naturalism

In his early writings—mainly in those on forensic topics—Kraepelin clearly expressed his opinion that *a priori* ideas, freedom of the will, and unchangeable moral values do not exist. Everything is more or less dependent on the time and the specific social and cultural context in which it is used. For Kraepelin, man is nothing but a part of nature, and anything man can do is a product of this natural existence: thus Kraepelin should be seen as an exponent of the evolutionism theory (Hoff 1998). Later in his life, he became somewhat more cautious concerning these matters, but there is no reason to believe that he substantially changed his mind. This naturalistic, 'anti-metaphysical' point of view of course made Kraepelin feel sympathetic with Darwinistic and biologistic theories, although he always rejected oversimplifications such as those in the monistic theories of Ernst Haeckel (1834–1919), Jakob Moleschott (1822–93), and Ludwig Büchner (1824–99).

3. Nosology

3.1 Basic Aspects

On the clinical level, Kraepelin changed the details of his diagnostic system over and over again. On the basic level, however, his nosology showed remarkable stability over time: between the second and the ninth editions of his textbook (i.e. from 1887 to 1927) Kraepelin did not change the central postulate. He stated that the essential features of all psychotic disorders will eventually be classified in a 'natural' system, no matter which scientific method is applied: anatomy, etiology and symptomatology, if developed sufficiently, will necessarily converge in the same 'natural disease entities' (see Fig. 1).

This very strong hypothesis is limited to a certain extent in three important theoretical papers, written between 1918 and 1920: 'Ends and means of psychiatric research' ('*Ziele und Wege der psychiatrischen Forschung*'; 1918)), 'Research in the Manifestations of mental illness' ('*Die Erforschung psychischer Krankheitsformen*'; 1919) and 'Clinical Manifestations of Mental Illness' ('*Die Erscheinungsformen des Irreseins*'; 1920). In these papers Kraepelin took into account

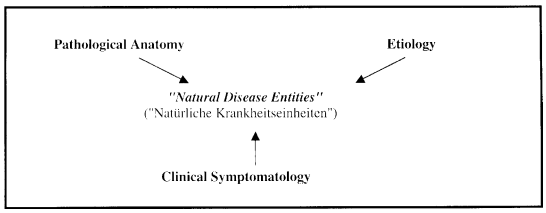

Figure 1

contemporary arguments such as K. Birnbaum's differentiation between pathogenetic and pathoplastic factors in Mental Illness or R. Gaupp's hypothesis of the possibility of psychogenic delusion; he now acknowledged the value of defining certain syndromes as a middle course between nosologically unspecific symptoms and specific diseases. But—and this is the essential point—at no time did he abandon his postulate of underlying distinct and natural disease entities (Hoff 1994, 1995).

3.2 Clinical Aspects

Kraepelin's clinical nosology is best separated into three periods.

3.2.1 The early period, 1880–91. This period can be characterized as the search for a reliable and valid psychiatric system between clear-cut naturalistic beliefs and the methodological framework of experimental psychology in the sense of Wundt. As regards nosology, Kraepelin slowly moved away from earlier nineteenth-century concepts which he criticized as being unreliable and ill-defined from a clinical and, especially, prognostic point of view. In these years the term *'dementia praecox'* was not yet used. A group of clinically heterogeneous paranoid and hallucinatory psychoses tending to become chronic was labelled *'Wahnsinn'* (insanity) and resembled the evolutions now known as schizophrenic psychoses leading to a residual state. In addition, Kraepelin introduced *'Verrücktheit'* (madness) as a chronic psychosis with a better prognosis, that explicitly did not lead to residual states. The affective psychoses were split into the three groups: melancholia, mania, and periodical or 'circular' psychosis.

3.2.2 The middle period, 1891–1915. During these years Kraepelin's thinking reached its most systematic and influential level as far as its clinical and scientific implications are concerned: Kraepelin significantly broadened his clinical experience and self-consciously created a complete nosological system. He finalized his concept of natural disease entities as discussed above. The main clinical result of this period—first clearly proposed in the sixth edition of 1899—was the well-known dichotomy of endogenous psychoses: that is, the separation of *dementia praecox* (schizophrenia) with—as he saw it—a poor prognosis, from manic-depressive illness (bipolar disorder) with a good, or at least better, prognosis. With respect to *dementia praecox*, he supposed an organic defect as the basis of the illness, a kind of 'autointoxication,' leading to the destruction of cortical neurons. The patient's personality may promote the development of the psychotic illness, but it is not a central factor; contrary to most other nosological areas, 'degeneration' was believed to be of much less etiological importance in *dementia praecox*. 'Paraphrenia' was conceptualized as a psychosis with acute and heterogeneous clinical symptomatology, including the development of lasting deficits, but its separation from typical cases of *dementia praecox* was justified by the absence of massive disturbances of volition and by a much lesser degree of affective flattening.

In manic-depressive illness the etiology was said to be much less clear than that of *dementia praecox*. Kraepelin proposed a genetically-determined irritability of 'normal' affects, so that the psychosis itself emerged from certain predisposing 'basic states' (*Grundzustände*). The concept of degeneration was an integrative and central part of this hypothesis. In this period, Kraepelin integrated different types of circular or recurrent affective illness into the overarching concept of manic-depressive insanity (*Manisch-depressives Irresein*) (6th edition 1899).

Kraepelin's concept of paranoia was modified several times in this period. After the very broad concept of *Verrücktheit* of the early editions of *Psychiatry*, which was of restricted clinical use, he significantly narrowed it, especially in the 5th edition of 1896. Here, paranoia was defined as a severe and chronic delusional illness without constant alteration of personality and volition; the existence of abortive or benign cases was denied up to the 7th edition (1903/04). In the 8th edition (1915) this very rigid concept was broadened again, but not as much as in the early versions. Kraepelin now accepted cases with low severity and a comparably good prognosis, but he maintained the strict separation of *dementia praecox* and paranoia.

In Kraepelin's view—highly typical of the way of thinking within the theoretical framework of degeneration theory—disorders of personality resulted from a circumscript retardation of psychological development. He argued that not all patients with personality disorders reach a 'normal' or mature level of affective and cognitive functioning. Therefore, in this field, it made no sense to delineate clear-cut disease processes as in *dementia praecox* and manic-depressive psychosis.

3.2.3 The later period, 1916–26. During this final period, Kraepelin reacted to several commentaries and criticisms of his nosology, for example, A. E. Hoche's syndromatic theory and E. Kretschmer's suggestion to supplement the Kraepelinian system with a multidimensional approach. He moved towards an internal broadening of his system by reformulating his disease concept as discussed above; he combined a more differentiated view of pathogenesis and the role of individual psychological factors with a strong and unchanged position on the existence of 'natural disease entities.' Compared to this differentiation of the theoretical basis during these years, changes in clinical nosology were small.

4. *Kraepelin and Present-day Psychiatry*

Kraepelin's psychiatry became so influential during his lifetime, because it offered a pragmatical, clinical and prognosis-oriented nosology, developed by a self-confident author who focused on straightforward quantitative and naturalistic research methods. He claimed to abandon speculative aspects of psychiatry as far as was possible, although he himself did—unintentionally—'import' a number of implicit theoretical and, in part, highly speculative aspects into psychiatry.

In the decades after World War II, Kraepelinian ideas—and 'biological' concepts in general—lost much of their influence and were discussed mainly from the historical point of view. However, Kraepelin's concepts were 'rediscovered' in the 1960s and 1970s when 'biological psychiatry' became the most influential (and most self-confident) field of psychiatric research. In particular, researchers and clinicians from the English-speaking countries viewed themselves as 'Neo-Kraepelinians' (Blashfield 1984). But 'Neo-Kraepelinianism' is not a scientific theory in the strict sense. At present, it is an important and heterogeneous set of concepts, trying to create a basis for biological research in clinical psychiatry. Of course it does not just repeat Kraepelinian ideas. Central to this set of concepts is the intention to identify the biological basis—the 'natural' basis, in Kraepelin's words—of mental disorders. But these findings may well lead to a nosological system quite different from Kraepelin's, a topic which has attracted much interest at the end of the twentieth century. Many authors have postulated a 'denosologization' of psychiatric research, especially in biological studies (e.g., van Praag et al. 1987). The idea is to prevent nosological prejudices from hampering research by restricting the interpretation of data to the limits of 'classical' concepts, especially the Kraepelinian dichotomy of endogenous psychoses. There might (and will) be, these authors argue, biological findings that reveal an entirely different distribution in psychiatric patients than that expected within the framework of classical nosology. Such an approach, though modifying Kraepelinian nosology, does not stand in real contrast to his basic ideas. On the contrary, Kraepelin himself often stressed that all diagnostic criteria and categories were due to change according to the actual state of the art in psychiatric research—leaving the postulate of the existence of natural disease entities untouched.

'Neo-Kraepelinian' authors are—as was Kraepelin himself—at risk of overestimating the explanatory power of biological findings; for example, by accepting only biological criteria for nosology, and dismissing psychopathology as 'unscientific.' Nosological prejudices or dogmas do, of course, exist in psychopathological approaches as well as in biological ones, and they always severely obstruct research.

In the context of 'Neo-Kraepelinianism,' the modern operationalized diagnostic systems *ICD*-10 and *DSM-IV* should be considered. Their main impetus is to improve the reliability of psychiatric diagnoses by establishing clear diagnostic criteria. They focus on the behavioural level by means of clinical description and try to avoid interpretation of symptoms. Whereas Kraepelin's nosological system—at least implicitly—included a number of etiological hypotheses and epistemological presuppositions, operationalized diagnostic systems aim to avoid etiological considerations (a concept which is often, incorrectly, called an 'atheoretical approach'). But in spite of these theoretical differences, both Kraepelinian nosology and modern diagnostic systems run the risk of uncritically reifying their constructs called 'diagnostic entities.' This is why the authors of *DSM-IV*, for example, warn their users not to regard diagnostic categories as definite or 'natural' entities, but as conventions which need further verification. Especially in forensic psychiatry, diagnostic systems are sometimes misinterpreted as collections of natural constants, quasi-automatically leading to diminished responsibility (e.g., pathological gambling or kleptomania).

In summary, Kraepelin's psychiatry was not only highly influential in his lifetime, but gained wide acceptance toward the end of the twentieth century, as the focus of psychiatric research again shifted to biological topics. It is of crucial importance that present-day psychiatry—be it 'neo-Kraepelinian' or not—adopts a comprehensive view of Kraepelin's scientific work, embedded in a thorough knowledge of its complex conceptual history (Berrios and Hauser 1988). Otherwise, as well as benefiting from Kraepelin's ideas, psychiatry will also be restricted by their conceptual limitations.

See also: Bipolar Disorder (Including Hypomania and Mania); Bleuler, Eugen (1857–1939); Dementia: Overview; Dementia: Psychiatric Aspects; Depression; Freud, Sigmund (1856–1939); Janet, Pierre (1859–1947); Mental and Behavioral Disorders, Diagnosis and Classification of; Mental Health and Normality; Nosology in Psychiatry; Philosophy, Psychiatry, and

Psychology; Psychiatry, History of; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Schizophrenia; Syndromal Diagnosis versus Dimensional Assessment, Clinical Psychology of; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Berrios G E, Hauser R 1988 The early development of Kraepelin's ideas on classification: a conceptual history. *Psychological Medicine* **18**: 813–21

Blashfield R K 1984 *The Classification of Psychopathology—Neo-Kraepelinian and Quantitative Approaches*. Plenum Press, New York

Engstrom E J 1991 Emil Kraepelin: psychiatry and public affairs in Wilhelmine Germany. *History of Psychiatry* **2**: 111–32

Engstrom E J 1995 *Kraepelin—Social Section*. In: Berrios G E, Porter R (eds.) *A History of Clinical Psychiatry. The Origin and History of Psychiatric Disorders*. Athlone, London, pp. 292–301

Hoff P 1994 *Emil Kraepelin und die Psychiatrie als klinische Wissenschaft. Ein Beitrag zum Selbstverständnis psychiatrischer Forschung*. Springer, Berlin

Hoff P 1995 *Kraepelin—Clinical Section*. In: Berrios G E, Porter R (eds.) *A History of Clinical Psychiatry. The Origin and History of Psychiatric Disorders*. Athlone, London, pp. 261–79

Hoff P 1998 Emil Kraepelin and forensic psychiatry. *International Journal of Law and Psychiatry* **21**: 343–53

Kraepelin E 1883–1927 *Psychiatrie*. Nine editions: 1883, Abel, Leipzig; 1887, Abel, Leipzig; 1889, Abel, Leipzig; 1893, Abel, Leipzig; 1896, Barth, Leipzig; 1899 (2 vols.), Barth, Leipzig; 1903/04 (2 vols.), Barth, Leipzig; 1909–15 (4 vols.), Barth, Leipzig; 1927 (2 vols.), Barth, Leipzig

Kraepelin E 1918 Ziele und Wege der psychiatrischen Forschung. *Zeitschrift für die gesamte Neurologie und Psychiatrie* **42**: 169–205

Kraepelin E 1919 Die Erforschung psychischer Krankheitsformen. *Zeitschrift für die gesamte Neurologie und Psychiatrie* **51**: 224–46

Kraepelin E 1920 Die Erscheinungsformen des Irreseins. *Zeitschrift für die gesamte Neurologie und Psychiatrie* **62**: 1–29

Kraepelin E 1983 *Lebenserinnerungen*. Hippius H, Peters G, Ploog D (eds.), with the collaboration of P Hoff and A Kreuter. Springer, Heidelberg Berlin New York. English translation (1987): *Memoirs*. Springer, Berlin

Van Praag H M, Kahn R S, Asnis G M, Wetzler S, Brown S L, Bleich A, Korn M L 1987 Denosologization of biological psychiatry or the specificity of 5-HT disturbances in psychiatric disorders. *Journal of Affective Disorders* **4**: 173–93

Verwey G 1985 *Psychiatry in an Anthropological and Biomedical Context—Philosophical Presuppositions and Implications of German Psychiatry 1820–1870*. Reidel, Dordrecht

Weber M M 1991 Ein Forschungsinstitut für Psychiatrie. Die Entwicklung der Deutschen Forschungsanstalt für Psychiatrie München 1917–1945. *Sudhoffs Archiv* **75**: 74–89

P. Hoff

Kuhn, Thomas S. (1922–96)

1. *Life and Works*

Thomas S. Kuhn was born July 18, 1922 in Cincinnati, Ohio. He earned a Bachelor's degree (1943) in physics at Harvard University. During his graduate work in theoretical solid state physics with John H. van Vleck who later won a Nobel prize, he was introduced to the history of science by James Conant, then President of Harvard. In 1947, while preparing a course on the development of mechanics before Galileo and puzzling over passages of Aristotle's physics, he discovered something that set the agenda of his further work. When he read Aristotle with the set of contemporary physical concepts in mind, much of Aristotle's physics seemed either awfully wrong or even outright meaningless. But by changing the ascribed meaning of some of the Aristotelian key concepts (in some cases only slightly and in others more deeply), a whole new world of physics became accessible, namely that of Aristotle and his contemporaries. After completing his Ph.D. in theoretical physics in 1949, Kuhn switched entirely to the history of science where he began by applying the insights he gained through his experience of Aristotelian physics. From 1948 to 1951, he was a Junior Fellow of the Harvard Society of Fellows, and from 1951 to 1956 Assistant Professor for General Education and History of Science. From 1956 to 1964, he was a member of the faculty of the University of California at Berkeley, mainly in the History of Science Department. In 1957, he published *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought*, which became a very successful textbook. Kuhn spent the year 1958–9 as a Fellow at the Center for Advanced Study of the Behavioral Sciences in Stanford where he became aware of the different patterns typical of the interactions among natural scientists on the one hand, and among social scientists on the other. The difference typically concerns a broad consensus of natural scientists with respect to the foundations of their discipline, in contrast to social scientists who typically dispute them.

This observation became very important when Kuhn wrote *The Structure of Scientific Revolutions* (*SSR*) which was published in 1962. This is the most widely read book ever published in the history and philosophy of science, and it also had a tremendous impact in the social and behavioral sciences, most of all in the sociology of science. Its key term, paradigm, is now a household word. Some of its other central concepts like normal science, paradigm shift, or incommensurability are now commonly used in a large number of diverse disciplines. However, many of the interpretations the book (or its parts) has been given were vigorously rejected by Kuhn as deep misunderstanding of what he had intended to say. By now, more than one million copies of the book have been

sold in English, and it has been translated into at least 26 languages. In 1964, Kuhn joined Princeton University as a Professor of History of Science; he was also a member of the Institute of Advanced Study from 1972 to 1979. Kuhn's historical work after *SSR* mainly concerned the genesis of quantum mechanics. *Sources for the History of Quantum Physics* (Kuhn et al. 1967) contains an inventory of relevant material including interviews with many of the contributors to quantum mechanics. *Black Body Theory and the Quantum Discontinuity, 1894–1912,* (Kuhn 1978) is a controversial narrative of the introduction of the quantum hypothesis. Some of Kuhn's papers both on the history and the philosophy of science are contained in his *The Essential Tension*, published in 1977. In 1979, he became Professor for Philosophy and History of Science at the Massachusetts Institute of Technology until he retired in 1991. From the 1980s on, Kuhn worked on a book whose final working title was *The Plurality of Worlds: An Evolutionary Theory of Scientific Development*. In this book, he tried to elucidate the foundations of scientific change by a new theory of meaning of empirical concepts, especially of taxonomic kind terms. Furthermore, he elaborated on the evolutionary metaphor for scientific development with which *SSR*'s last chapter ended. Kuhn could not finish the book due to his death on June 17, 1996, in Cambridge, MA.

2. *Kuhn's Theory of Scientific Development*

Kuhn's main subject in the philosophy of science was a developmental scheme, or a schematic description, of scientific development which primarily concerned the basic (i.e., curiosity-driven) natural sciences. The intended description is schematic in the sense that it presents scientific development as a succession of different phases whose main features are independent of the specific subject matter of the respective science. For each one of these phases a specific mode of scientific practice is characteristic. This theory was first presented in *SSR* but was later refined in various respects.

The developmental scheme that Kuhn suggests should hold with only minor variations for all basic scientific disciplines. It consists of the following. Before reaching maturity, nascent scientific fields typically are characterized by controversies between competing schools. There is no consensus among the practitioners of the emerging field. Each one of these schools has a particular view of the respective research domain, and typically these views derive from extra-scientific sources. Kuhn has called this phase of scientific development 'preparadigmatic.' The competition among the schools eventually may end when one group produces an exemplary solution to a pre-eminent research problem with two characteristics: it is sufficiently unprecedented to attract the members of the other schools, and it is sufficiently open-ended to leave enough interesting problems for further scientific work. These model solutions are called paradigms. They serve to guide research implicitly in the succeeding period called normal science. When a science has reached its first normal science phase it is said to have matured.

Normal science is characterized by a broad consensus of the practitioners of the field about fundamental questions, and consequently, by a particular mode of research. This mode of research can be described by a five-dimensional analogy to puzzle-solving where exemplars of puzzles include chess problems and crossword puzzles. The analogy concerns:

(a) The existence of regulations constraining acceptable approaches to and solutions of problems. In the case of science, these regulations are given only implicitly, namely as part and parcel of the paradigmatic solutions.

(b) The expectation of the solubility of appropriately chosen problems. In the case of science, this expectation is fed by the conviction that the appropriateness of the paradigmatic solutions extends beyond the particular problems they are solutions of.

(c) Missing intentions of fundamental innovation of the guiding regulations. In the case of science, normal science intends to exploit the cognitive potential intrinsic to the paradigmatic solutions, and does not intend to overturn them.

(d) Absence of test or confirmation of the guiding regulations. In the case of science, this is due to the fact that in normal science, failure to solve a research problem is seen as a failure of the researcher and not of those regulations that guide research. Either the research problem was not chosen well, or the researcher was unable to solve it despite its solubility.

(e) Individual motivations: to prove oneself an expert puzzle-solver. In the case of science, this is evidenced by the extremely elaborate system by which scientific achievements are rewarded (e.g. scientific prizes including the Nobel prize, honorary degrees, invited lectures, nominations, etc.).

Normal science is always confronted with anomalies, i.e., with phenomena or problems that behave contrary to the expectations supplied by the paradigm. Anomalies do not usually call the validity of the guiding regulations of normal research into question. But under special circumstances they may, and then they become 'significant anomalies.' In such cases, the practice of science changes into 'extraordinary science' or 'science in crisis.' It aims at amending or even overthrowing the yet binding regulations. Its research focuses on the significant anomalies and their context. Extraordinary science resembles prenormal science in that it tends to develop competing schools. However, it is more focused than prenormal science as all these different schools have to deal with the same set of

significant anomalies and they all want to retain as much as possible from the earlier period of normal science. If this research leads to a new theory that is accepted by the scientific community because it can lead to a new phase of normal science, a scientific revolution has occurred. Scientific revolutions in Kuhn's sense are thus 'the tradition-shattering complements to the tradition-bound activity of normal science' (Kuhn 1962). The rejection of the older theory is accompanied by a change of the problem-field and its related standards of solution, and by a corresponding change in basic scientific concepts. Some old concepts are discarded, some new ones are introduced, and some change meaning, sometimes in a subtle way. Kuhn even describes revolutions as transformations of the world in which scientific work is done, although it is not easy to make explicit and plausible what is exactly meant by this locution. Kuhn compresses these features of revolutions into the concept of 'incommensurability': a relation that holds between successive traditions of normal science. In *SSR*, the concept of incommensurability had not become entirely clear. It was, therefore, the subject of much criticism as well as misunderstanding. Most of Kuhn's philosophical work after *SSR* aims at a clarification and further explication of the concept of incommensurability.

Because of incommensurability, we must, according to Kuhn, rethink the concept of scientific progress in the natural sciences. First, scientific progress is not cumulative, due to conceptual changes during revolutions. Cumulativity implies that something that is a part of science at some point in time will, neglecting small corrections, remain a part of science forever. But the conceptual changes that occur during revolutions are much more than small corrections of the existing body of knowledge. They amount to a thorough conceptual reorganization, and typically to a changed ontological perspective. This holds in spite of the fact that very often, much of the older knowledge is retained, especially for practical applications, and that, in the case of quantitative theories, limiting relations exist between the older and the newer theory. Furthermore, Kuhn denies that scientific progress is an approach to truth. Instead of conceiving of scientific progress as a teleological process, i.e., one that is goal-directed, we should think of scientific progress in an analogous way as Darwinian evolutionary theory conceives of evolution. Darwinian evolutionary theory states that there is no goal of evolution towards which it is directed. In a similar way, in scientific development there is no 'set goal' which would be 'a permanent fixed scientific truth' that science approaches (Kuhn 1962). However, Kuhn stresses that this does not imply that there is no progress in the sciences. There is progress in the sciences, though not in the form of an increase of verisimilitude. Rather, there is progress in the form of 'an increase in articulation and specialization' of scientific knowledge (Kuhn 1962). Therefore, the widespread characterization of Kuhn's theory as entirely relativist is simply false.

There are several breaks between Kuhn's theory and the previous tradition in the philosophy of science, some of which explain the extraordinary resonance his theory found in the social and behavioral sciences. Six of them will be discussed briefly. The first is Kuhn's altered view of scientific progress that has just been mentioned. Second, this view implies the untenability of those forms of realism which assert that science at least approximately describes what is really 'out there,' independently of any observer. Instead, theories describe the world in terms of concepts that are historically contingent and which may change in the future. Third, because of this change of basic scientific concepts through revolutions, the classical conception of reductionism is also hardly tenable. In this classical conception, the redefinition of the concepts of the reduced theory by means of the concepts of the reducing theory played a key role. But if incommensurability prevails, some of the required redefinitions are impossible because of meaning shifts. Fourth, many of *SSR*'s assertions stood in marked opposition to Popper's critical rationalism. For instance, Kuhn's normal science, as seen from Popper's perspective, is bad science because it is not continuously directed towards critically testing the guiding regulations. Instead, some anomalies are not even pursued, but are ignored or even pushed aside. Alternatives to the paradigmatic framework are also typically ignored. Yet in Kuhn's view, critical evaluation of fundamental theories is restricted to the periods of extraordinary science, and even then, scientific practice is not simply an attempt to falsify theories by confronting them with basic statements about nature.

Rather, theory evaluation is a comparative procedure in which at least two theories are assessed with respect to their cognitive abilities. Thus according to Kuhn, theory falsification as described by Popper is a stereotype that does not occur in the actual history of science. A fifth consequence of Kuhn's theory is the abolishment of the idea that science is guided by the scientific method, as construed as a set of rigorous rules. This idea has dominated the understanding of modern science from its very beginning. But according to Kuhn, exemplary problem solutions guide scientific research in its normal phase. Their cognitive potential for research is not exploited by explicit (or fully explicable) rules, but rather by implicit analogies. New problems are identified in the light of solved ones, and new solutions are judged as legitimate in a like manner.

Finally, and most importantly in our context, for Kuhn the principal agents of science are communities, not individuals. In the philosophy of science tradition before Kuhn, no one questioned the identity of the principle agent of science. But armed with the distinction between individuals and communities, it is obvious that in the older tradition, the individual

scientist was taken to be the principal actor. The possibility of two scientists rationally disagreeing was not permitted, nor were there discussions about the gradual formation of scientific consensus out of disagreement, or of the development of disagreement from a previously established consensus. In other words, scientific communities were not present in the discussion at all. Science was seen as a one-person game, but for Kuhn it is, at its heart, a social enterprise. It is communities who are the ultimate evaluators of knowledge claims. It is communities who ultimately decide between competing theories and their associated modes of research, and it is communities who are seen by themselves and by others as responsible for some domain of scientific knowledge. Kuhn, therefore, repeatedly refers to the 'sociological base of my position' (for references, see Hoyningen-Huene 1993), but at the same time, he insists that this fact does not invalidate the essential epistemological component of his theory, i.e., its normative element. How is this to be understood?

Kuhn characterizes scientific communities by the values they hold. Typically, these values comprise accuracy, scope, consistency, fruitfulness, and explanatory power, among others. These values are instrumental in the evaluation of the application of theories and, partly derivatively, in the evaluation of the theories themselves. However, these communal values are not defined so sharply that each member of the community reaches exactly the same evaluation in a concrete case of application. Individuals may differ in the relative weight they attribute to these values as well as in the specific articulation of any one of them. Yet in normal science, these differences typically do not become apparent because there is a consensus about the paradigms. However, with the advent of significant anomalies, this latent difference between scientists belonging to the same community may become manifest due to different evaluations of the anomalies, resulting in dissent. The dissent may continue with respect to the different candidates proposed to resolve the crisis. A new consensus about some theory and its ability to solve outstanding problems can only emerge if, and when, (almost) all members of the community prefer it to its competitors, although their individual differences with respect to the values have not disappeared. Thus, during crisis resolution, the individual differences of the members of the community become irrelevant. It is the communal core of the values that prompts the choice of the new paradigmatic theory. Thus, the choice of the new theory is essentially an event that has to be described in sociological terms, as opposed to psychological ones. On the other hand, the normative dimension of theory choice has not disappeared. The reasons for the choice may still be normatively evaluated. Scientific values are an essential part of the reasons for choosing some theory. They are good reasons if they promote the aims of science. This is typically the case if the values in question are those values mentioned above (i.e., accuracy, fruitfulness, consistency, and the like).

3. *Kuhn's Impact on the Behavioral and Social Sciences*

Kuhn's work has been extremely widely received in the social and behavioral sciences, more than any other philosopher's work in the twentieth century. Before Kuhn, mainstream social science had adopted a picture of the natural sciences that was based mainly on logical empiricism, and many social scientists attempted to assimilate their disciplines to that picture. However, success was limited. Neither the predictive, nor the explanatory power of a natural science like physics could be achieved. In addition, methodological controversy did not cease to exist even if in some field, a particular school became dominant for some period. In this situation, the alternative view of the natural sciences that Kuhn offered also seemed to suggest, as a corollary, a different view of the social sciences. More specifically, as Kuhn had incorporated elements in his description of the natural sciences which had been barred by the earlier tradition in the name of the objectivity of science, such as group processes and the like, the difference between the natural and the social sciences seemed to have decreased. In this way, the already comparatively high interest that social scientists have in the philosophy of science, given the epistemologically insecure status of their disciplines, became even more intense in the case of Kuhn's philosophy. At least to some authors, Kuhn seemed to suggest a kind of fresh start for the social sciences.

Three general facts about the reception of Kuhn's work in the social and behavioral sciences are striking. First, the reception was by and large uncritical, and in many cases very superficial. For example, there are literally hundreds of articles in which the reference to Kuhn appears to be a purely rhetorical device, making no substantial point. Second, Kuhn's developmental model was intended to apply only to the basic disciplines in the natural sciences. Kuhn says very little about the social sciences, basically only that they are typically in the pre-paradigmatic state (Kuhn 1962, 1977). This did not prevent many social and behavioral scientists from applying parts of Kuhn's theory to different aspects of their disciplines. Third, Kuhn's ideas have been used for the most diverse purposes. On the basis of his work, even outright contradictory assertions have been made. This is partly due to the fact that Kuhn's writings, especially *SSR*, have admitted of very different interpretations, which in turn derives from the inner tensions, ambiguities, and metaphors that they contain. In particular, Kuhn's evaluation of the rationality of science appeared as highly ambivalent, and his use of the paradigm concept

seemed at times to cover just about everything with which scientists deal. Thus, Kuhn's texts appear to have a high degree of plasticity, a fact that has been subsequently deplored by their author. Only a very close and patient reading of the texts can reduce this plasticity. But on top of the plasticity of Kuhn's texts, social scientists, due to their own agendas, have been seduced to read many things into Kuhn's work. Thus, many of the claims in the literature about what Kuhn does or does not support are not really contained in Kuhn's texts.

The predominant use of Kuhn's work in the social and behavioral sciences has concerned the current status of some discipline or its history. Is the discipline still in its preparadigmatic state or does it already have a paradigm? Or put another way, is it a 'mature' science which has reached the state of normal science? Has it had paradigms in the past? Are the changes that the discipline experienced in the past Kuhnian revolutions? Is the discipline currently in a Kuhnian crisis? Or are the social sciences fundamentally different from the natural sciences in being 'multiple paradigm sciences'? As mentioned above, these questions received the most diverse answers by different authors. Elements from Kuhn's theory were used by some to defend mainstream social science, and by others to develop a critical attitude towards it, or to defend or attack alternatives. Some authors tried to show that on Kuhnian terms, a given social science is very nearly like natural science, whereas others tried to demonstrate the opposite, and so on. No consensus has been reached in any of the social or behavioral sciences about these and related questions.

Finally, that part of the reception of Kuhn's work to which it was immediately pertinent, namely the history and the sociology of science, should be discussed. Perhaps surprisingly, in the history of science proper, the reception of Kuhn's metahistorical work does not show up heavily in the literature. This may be partly due to the same reasons that Kuhn did not apply his characteristic concepts (like paradigm, normal science, etc.) in his own historical work. Historians very often try to avoid heavily theory-laden concepts when doing historical work because of the imminent distortions of what may be gleaned from the sources themselves. In addition, historians in the second half of the twentieth century have typically been quite critical about any generalizations about the course of history (such as a developmental scheme).

But in the sociology of science, Kuhn's impact has been tremendous. Pre-Kuhnian sociology of science either discussed the social structure of science or the impact of science upon society. But in contrast to other areas of sociology, the scientific content seemed to lie entirely outside the scope of sociology because of the purported objectivity of science, which precluded any direct social contingency. The most relevant sociological question asked with respect to scientific knowledge was 'what are the social conditions of a scientific community that are conducive to the production of such knowledge?' Answers to this question typically consisted in the description of a purportedly universal value system holding in scientific communities that virtually guaranteed the desired outcome of the activities of the community's members. However, with the advent of Kuhn's work, this view was challenged in various respects. The production of scientific knowledge now seemed not at all determined by time-independent, universal methodological rules. Instead of binding rules, research is guided by values which only influence, but do not determine scientific behavior. Such values are influenced by group-specific and even idiosyncratic factors. In other words, scientific knowledge now indeed seemed socially contingent, which licenses and even appears to demand, sociological analysis. Various schools developed analytical tools for this purpose, and a great many case studies of contemporary and past science were performed. Typically, the content of science was now seen as the result of negotiations among scientists in which human interests of a personal or political nature, or power relations, played an important, and sometimes even decisive, role. Due to an intense debate within sociology and interventions mainly from philosophers, these positions showed rapid development that continues today. It should be noted, however, that despite all of the changes in these positions, Kuhn himself remained extremely skeptical towards them. In his view, the epistemological dimension of scientific knowledge, i.e., the possibility of its normative evaluation, must be an essential part of an integrated image of science, but it had been entirely dismissed by the sociology of science (Kuhn 1991, 1992).

See also: History of Science; History of Science: Constructivist Perspectives; Mathematical Models in Philosophy of Science; Physical Sciences: History and Sociology; Scientific Controversies; Scientific Culture; Scientific Disciplines, History of; Scientific Instrumentation, History and Sociology of

Bibliography

Barnes B 1982 *T. S. Kuhn and Social Science*. Macmillan, London

Bernstein R J 1976 *The Restructuring of Social and Political Theory*. Blackwell, Oxford, UK

Gutting G (ed.) 1980 *Paradigms and Revolutions. Applications and Appraisals of Thomas Kuhn's Philosophy of Science*. University of Notre Dame Press, Notre Dame, IN

Horwich P (ed.) 1993 *World Changes. Thomas Kuhn and the Nature of Science*. MIT Press, Cambridge, MA

Hoyningen-Huene P 1993 *Reconstructing Scientific Revolutions. Thomas S. Kuhn's Philosophy of Science* (trans. Levine A T). University of Chicago Press, Chicago

Kuhn T S 1957 *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought*. Harvard University Press, Cambridge, MA
Kuhn T S 1962/1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago
Kuhn T S 1977 *The Essential Tension. Selected Studies in Scientific Tradition and Change*. University of Chicago Press, Chicago
Kuhn T S 1978 *Black Body Theory and the Quantum Discontinuity, 1894–1912*. Clarendon Press, Oxford, UK
Kuhn T S 1991 The road since structure. In: Fine A, Forbes M, Wessels L (eds.) *PSA 1990. Proceedings of the 1990 Biennial Meeting of the Philosophy of Science Association*, Vol. 2. Philosophy of Science Association, East Lansing, MI, pp. 3–13
Kuhn T S 1992 *The Trouble with the Historical Philosophy of Science. Robert and Maurine Rothschild Distinguished Lecture, November 19, 1991, An Occasional Publication of the Department of the History of Science*. Harvard University Press, Cambridge, MA
Kuhn T S 1993 Afterwords. In: Horwich P (ed.) *World Changes. Thomas Kuhn and the Nature of Science*. MIT Press, Cambridge, MA, pp. 311–41
Kuhn T S 2000 *The Road Since Structure. Philosphical Essays, 1970–1993, with an Autobiographical Interview*. Conant J, Hangeland J (eds.) University of Chicago Press, Chicago
Kuhn T S, Heilbron J L, Forman P, Allen L 1967 *Sources for the History of Quantum Physics. An Inventory and Report*. American Philosophical Society, Philadelphia, PA
Lakatos I, Musgrave A (eds.) 1970 *Criticism and the Growth of Knowledge*. Cambridge University Press, London

P. Hoyningen-Huene

Kula Ring, Anthropology of

First set out in detail in Malinowski's classic *Argonauts of the Western Pacific* (1922), the 'kula ring'—a translocal, long distance exchange network connecting islands off the southeast coast of Papua New Guinea (Fig. 1)—has become a classic ethnographic reference point for the study of non-Western exchange practices. Yet apart from more limited work by Fortune (1932), Roheim (1950) and Belshaw (1955), field research yielding significant publications on kula was not carried out again after Malinowski until the 1970s (see Leach and Leach 1983). This article explains key features of kula, updating Malinowski's account where necessary on the basis of the more recent work, and modifying some common over-simplifications; in conclusion, it comments briefly on anthropological approaches to kula.

1. Some Defining Features of Present-day Kula

Modern kula centers in the exchange of white (*Conus*) armshells and red (*Spondylus* or *Chama*) shell necklaces in opposite directions around the islands (Fig. 1). Only middle and upper ranked shells have personal names, but all shells are qualitatively graded. There are other kinds of kula media (see Sect. 2), but only these shells currently circle the region establishing its basic socio-geography. Kula transactors, however, should exchange directly in a limited sector with partners or potential partners in participant communities on either side of them (although illicit bypassing occurs). Working a strategic politics of influence in the inter-island world, they attempt to make names for themselves that also impact on their standing in their own community.

Transactors obtain shells of one category on visits to partners in one direction and later transact them to their own visitors from the opposite kula direction. Typically these exchange practices engage long term debt involving a year or much more between a gift and its equivalent repayment. A major transaction with multiple, additional shell exchanges pinned to its passage may go on for years. Nevertheless, less prestigous exchanges may occasionally be completed with a visitor who brings the appropriate type of shell for immediate exchange, thus avoiding the risks of default inherent in time and distance.

Although kula participants are primarily men, recent research points to the contemporary importance of women traders in the Bwanabwana region (Macintyre 1983a) and, less prominently, in some other areas such as northern Mwadau and Dobu (J. Leach in Leach and Leach 1983, p. 17); nevertheless, women participants do not usually sail on kula. This research has also revealed a crucial type of shell ownership not noted by Malinowski (but probably operative in his time) called 'kitomu' (or cognates), which is fundamental to the *modus operandi* of kula. A shell's first owner is its maker. Owners may do as they will with their kitomu shells, keeping or else alienating them to obtain high value items like pigs, magic and cash or to pay off a kula debt; the shell then becomes another person's kitomu. Alternatively one may circulate one's kitomu in kula; an equivalent return becomes, on receipt, one's kitomu, the prior shell becoming the kitomu of the equivalent's donor. Thus particular kitomu are alienated, but, apart from default, one still owns a shell as kitomu (Damon in Leach and Leach 1983, Munn 1977, 1986, Weiner 1988, 1992 cf. also Godelier 1996, p. 90–1). Kitomu shells, empty of prior obligations, provide an element of freedom in transaction that variously aids owners in the incrementation of their kula.

2. Adjustment of Some Common Oversimplifications of Kula

Following Malinowski's emphasis, a common assumption is that modern kula entails a ritual

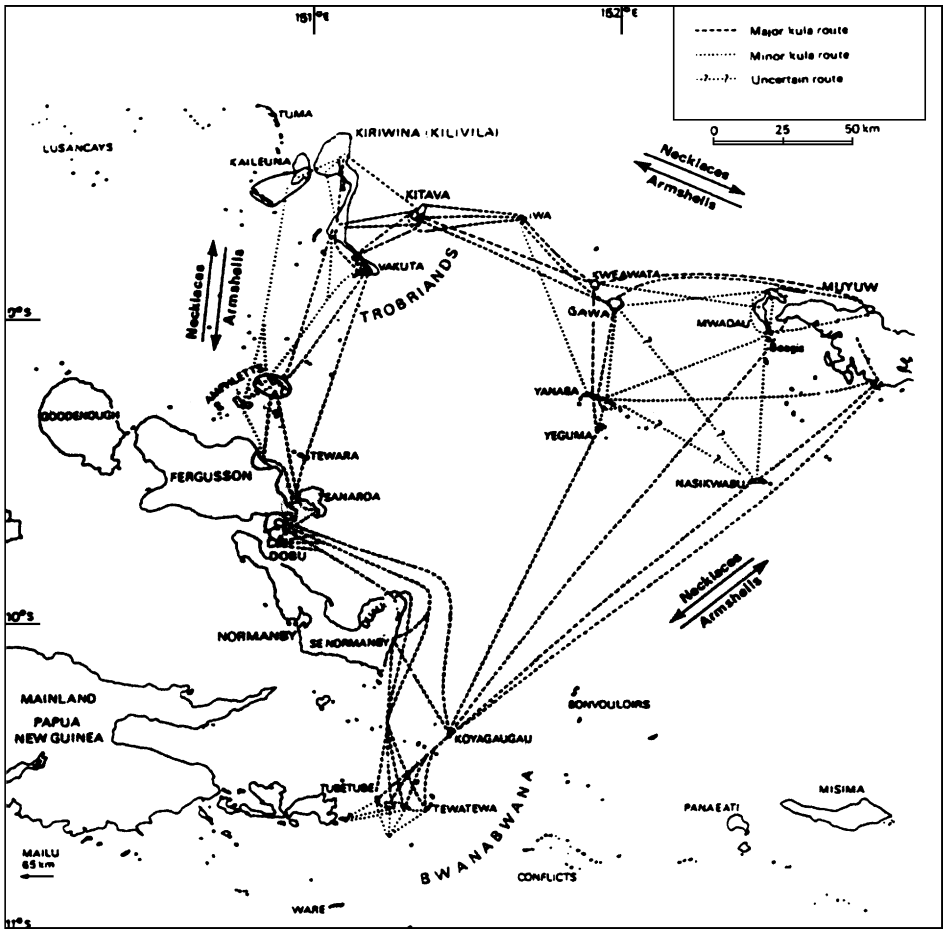

Figure 1
The kula ring in the 1970s (from Munn 1986, p. 4)

exchange of just the two types of shell set apart as a system from trade with other non-ritual goods (see Sect. 3). But a variety of evidence indicates that such a sharp line cannot be drawn. For instance, in Bwanabwana (Fig. 1) pigs are intrinsic to kula (there called 'kune') although they obviously do not circulate around the whole ring (Macintyre 1983a, pp. 105–6). Overseas directional transactions of Gawan-built canoes can also be described as 'kula-ing' one's canoe. Armshells from southern canoe recipients become the canoe owners' kitomu; with the exchange finished, canoes are alienated from their makers (Munn 1986).

If kula media are not as exclusive as sometimes supposed, so also the circulation of the shells themselves is not exclusive to Massim islands participating in the ring: other islands make use of these shells, leaking them into and draining them out of the kula region, thus affecting kula circulation. Indeed, as Malinowski reported, a source of the necklaces (Rossel Island) is itself far south of the ring. Kula is in this sense an open not a closed system.

Neither is kula static. In its present form, it is a modern development radically affected by British colonization. One key change has been the widening of participation in kula. Formerly confined to men of influence, kula effloresced, becoming increasingly 'democratized' as endemic warfare gave way to British 'pacification' (Macintyre 1983b; Young 1983). Of course change is ongoing; for instance, in the 1990s the increased use of western boats by Bwanabwanans was reducing the southward extension of canoe transactions, thus negatively affecting Gawan acquisitions of kitomu from canoes (Damon, personal communication 1999).

At another level, change is inherent in the way individual kula paths and routing are constituted. A path of partners is initiated and maintained only through the passage of shells: without shells a path disappears. *Pace* Malinowski (1922, p. 83), partner-

ships are not necessarily 'permanent.' In a politics that entails the constant testing of trust, shifting personal relations ongoingly alter paths across the ring. Nor are routes between communities fixed. Over time kula's competitive practices can yield not just individual path alterations, but also alterations in routes. Thus by the 1970s Duau (north-east Normanby) villages no longer did much kula directly with Bwanabwana: new southern coastal villages of people from the interior had managed to insert themselves into the route (Thune in Leach and Leach 1983, p. 352).

This dynamic view of kula also illuminates non-circular aspects of its socio-geography. Although the overall routing is circular, at any given moment many paths of partners do not circle the region. A path may never attain circularity because closure can be produced only by making reliable connections via the passage of shells.

Just as kula is ongoingly subject to change, so also it is marked by variability throughout the ring. Notable are different local articulations with marital and other domestic exchanges. In Bwanabwana, kula is not sharply separable from marriage exchanges, but in the Trobriands the two are kept apart (Macintyre 1983a, Weiner in Leach and Leach 1983). Elsewhere, articulations vary. Gawan men take shells from their kula paths to give to affines for the latters' paths; shell returns ideally go back to donors and the original paths (Munn 1986). But on Vakuta, shells taken off paths for affinal recipients in return for gifts of pigs become these affines' kitomu, even though they may not be the donors' kitomu (Campbell in Leach and Leach 1983, pp. 206–7); presumably donors must find path replacements elsewhere. These sorts of local differences, among many others, have ripple affects on trans-local shell movements (cf. also Wagner 1989, p. 270, Persson 1999).

3. *Anthropological Approaches to Kula*

Serious focus upon kula first emerged within two formative studies of anthropological thinking on exchange Malinowski's (1922) first major work developing his use-oriented functionalism, and Mauss' (1950) classic essay on the gift. Basic problems posed by typical, pre-1970s kula commentaries drew on Malinowski's questionable dichotomy between kula as a ceremonial practice involving the transaction of 'useless' things and the economic trade of 'useful' things (for a brief critique see Young 1983, p. 8). The distinction between ceremonial (kula) and non-ceremonial trade was also the prototype for the later development of more complex models of economies with multiple separate 'exchange spheres' (cf. Firth 1957 p. 218).

The same Western economic dualism framed the search for some single explanatory function, or 'purpose' of kula such as 'prestige competition' (J. W. Leach in Leach and Leach 1983, pp. 5–6). In particular, it fuelled the concern with finding an economic function for kula by putting it in the service of the 'useful' commerce—for instance, as establishing 'the peace of the market' enabling utilities to circulate (Singh Uberoi 1962, p. 148; see Macintyre's 1983b critique of the association of kula with peacemaking). The period's most serious reexamination combined some of these ideas in a more broadly based structural-functional study (Singh Uberoi 1962) which viewed Trobriand kula as a sociopolitical practice periodically releasing men from internal descent group relations to establish inter-island bonds.

It is, however, Mauss's concerns that pervade contemporary kula analyses. Mauss situated kula in a theory of moral–juridical contract empowered by bonds between persons and things. His attention to the cultural symbolism and ontology of 'person-thing' relations in gift exchange (rather than simply to exchange relations between things, as in the 'exchange spheres' model) has been influential in recent exchange and kula studies. Mauss' contrast between 'archaic' gift exchange and the modern Western market-based system is also one root of current arguments involving another dualism, that of 'gift' and 'commodity' exchange. Kula figures prominently in these arguments—for instance, in Gregory's (1983) widely cited, if problematic, model of gift exchange as the transaction of 'inalienable objects' in contrast to the 'alienable objects' of commodity exchange. Yet the dynamics of kitomu (Sect. 1, 2) alone suggest the difficulties of applying this dichotomous framework to kula.

No single orientation dominates contemporary kula studies. Rather, kula has been subject to varied forms of rethinking: for example, in terms of a Marxian concern with grounding circulation in production and labor value (Damon 1980), a symbolic formulation of action, spacetime and value (Munn in Leach and Leach 1983, Munn 1986), or gendered models of power and social reproduction (Weiner 1992). Most recently Persson's (1999) detailed comparison of selected Massim societies approaches kula through the study of regional variation proposing 'a holistic perspective going beyond societies as closed entities' (Persson 1999, p. 224). Since its introduction into anthropology at the beginning of the twentieth century, kula has been a sounding board for anthropological issues, and continued to be so at the century's end.

See also: Exchange in Anthropology; Melanesia: Sociocultural Aspects

Bibliography

Belshaw C 1955 In search of wealth: A study of the emergence of commercial operations in the Melanesian society of south-east—Papua. *American Anthropological Association Memoir* **80**, Vol. 57, no. 1, pt. 2

Damon F H 1980 The Kula and generalised exchange: Considering some unconsidered aspects of the elementary structures of kinship. *Man* **15**: 267–92

Damon F H 1990 *From Muyuw to the Trobriands: Transformations Along the Northern Side of the Kula Ring*. University of Arizona Press, Tuscon, AZ

Firth R 1957 The place of Malinowski in the history of economic anthropology. In: Firth R (ed.) *Man and Culture*. Routledge and Paul, London

Fortune R K 1932 *Sorcerers of Dobu*. Routledge and Sons, London

Godelier M F 1996 *L'enigme du don*. Librairie Artheme Fayard [1999 *The Enigma of the Gift*. University of Chicago, Chicago]

Gregory C H 1992 *Gifts and Commodities*. Academic Press, London

Leach E, Leach J W (eds.) 1983 *The Kula: New Perspectives on Massim Exchange*. Cambridge University Press, Cambridge, UK

Macintyre M 1983a Changing paths: An historical ethnography of the traders of Tubetube Ph.D. thesis (unpubl.), Australian National University, Canberra, Australia

Macintyre M 1983b Warfare and the changing context of *kune* on Tubetube and Papua–New Guinea. *Journal of Pacific History* **18**: 11–34

Malinowski B 1922 *Argonauts of the Western Pacific*. Routledge and Sons, London

Mauss M 1950 Essai sur le don. Formes et raison de l'echange dans les sociétés archaiques. In: *Sociologie et anthropologie*, 1st edn. Presses Universitaires de France [1990 *The Gift: The Form and Reason for Exchange in Archaic Societies*. Routledge and Kegan Paul, London]

Munn N 1977 The spatiotemporal transformation of Gawa canoes. *Journal de la société des Océanistes* **33**: 39–53

Munn N D 1986 *The Fame of Gawa: A Symbolic Study of Value Transformation in a Massim (Papua New Guinea) Society*. Cambridge University Press, Cambridge, UK

Persson J 1999 *Sagali and the Kula: A Regional Systems Analysis of the Massim*. Department of Sociology, Lund University, Lund, Sweden

Roheim G 1950 Psychoanalysis of a Matrilineal Society. Psychoanalysis and Anthropology. International Universities Press, New York

Singh Uberoi J P 1962 *Politics of the Kula Ring: an Analysis of the Findings of Bronislaw Malinowski*. Manchester University Press, Manchester, UK

Wagner R 1989 The exchange context of the Kula. In: Damon F H, Wagner R (eds.) *Death Rituals and Life in the Societies of the Kula Ring*. Northern Illinois University Press, De Kalb, IL

Weiner A B 1976 *Women of Value, Men of Renown*. University of Texas, Austin, TX

Weiner A B 1988 *The Trobrianders of Papua New Guinea*. Holt, Rhinehart and Winston, New York

Weiner A 1992 *Inalienable Possessions*: *The Paradox of Keeping While Giving*. University of California Press, Berkeley, CA

Young M 1983 The Massim: An introduction. *The Journal of Pacific History* **18**: 4–10

N. D. Munn

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

L

Labor History

Labor history has a double meaning. In a narrow sense it denotes the history of social movements of wage earners, including their leaders, their organizations (e.g., mutual aid societies, trade unions, consumer, and producer cooperatives, and working-class parties), and their collective actions (petitions, strikes, boycotts, electoral campaigns, etc.). In a broad sense, labor history comprizes the history of the working classes at large, including the history of family life, demography, everyday culture, leisure activities, housing, religion, migration, and so on (see *Working Classes, History of*). In the early phases of its development as an academic discipline since the 1870s practitioners of labor history predominantly used the narrow concept.

1. Beginnings

Labor history as an academic practice came into being after the classical Wage Fund theory had become discredited. According to this theory (of which many variations have existed) there is at any point in time a fixed amount of money out of which wages can be paid; it was therefore 'plainly impossible, in any given state of capital and population, to bring about any genuine and permanent rise of wages, otherwise than in the slow course of generations' (Webb and Webb 1897, II, p. 603). Trade-union action could thus have two results. Either it would lead to a redistribution of wages from one segment of the working class to another without the total available amount increasing: 'the increased wages secured for merely absorb a larger share of the capital, leaving less to be distributed among others.' Or, trade-union action was more or less useless: 'If wages are below equilibrium, a union can raise them, thus doing what ought to be done and in time would be done by the market itself' (McCulloch 1851, pp. 50–2). In the 1870s the decline of the Wage Fund theory started with criticisms by John Stuart Mill and others.

With the decline of the Wage Fund theory trade unions and strikes came to be seen in a different light. While most scholars previously had considered trade-union activities as false in theory and pernicious or futile in practice, it now became a much more serious topic deserving careful analysis, both contemporary and historical. Pioneer efforts were, amongst others, Lujo Brentano's *On the History and Development of Gilds, and the Origins of Tradeunions* (London 1870), Richard T. Ely's *The Labor Movement in America* (New York 1886) and the *History of Trade Unionism* (London 1894) by Sidney Webb and Beatrice Webb (born Potter).

For a long time two patterns of interpretations were dominant. On the one hand there was the liberal tradition, represented by, for instance, John R. Commons and Selig Perlman in the United States, the Hammonds in Britain, or scholars like Lujo Brentano, Gerhard Schulze-Gävernitz, and others in Germany. They reconstructed the development of labor movements as the history of the civil emancipation—and consequently the integration—of the working class within capitalism.

The other 'classical' tradition was, of course, the Socialist approach in its many forms, varying from moderates to radicals, and including people like the Webbs in Britain, Edouard Dolléans in France, Eduard Bernstein and Carl Grünberg in Germany, or Edvard Bull in Norway. What these socialists had in common was their interpretation of labor history as a history of attempts to transcend capitalism—in whatever way.

Both 'classical' approaches emphasized the labor movements' institutional aspects. Classical labor history consisted of the application of the approaches, methods, formats, and styles of the traditional historiography of ideas and politics to the field of labor history. Insofar as it had addressed 'social issues,' it had been an offshoot of economics and economic history; and insofar as it had addressed the movement itself and its organizations, leaders, and ideas, it had been an offshoot of political or intellectual history. Both versions of the classical approach 'tended to produce both a model and an accepted version of history, both national and international, which ranged from an informal but not very flexible to a formal and highly inflexible orthodoxy' (Hobsbawm 1984, p. 4) I should add, however, that some authors were more or less dissident. There have always been historians who paid attention to atypical aspects, like Werner Sombart in his essays on the Italian proletariat (1895); John Le Breton Hammond and his wife Barbara Bradby in their trilogy *The Village Labourer, 1760–1832* (1911), *The Town Labourer, 1760–1832* (1917) and *The Skilled Labourer, 1760–1832* (1919), or Ivy Pinchbeck in her book on *Women Workers and the Industrial Revolution, 1750–1850* (1930).

2. Broad Interpretation

The classical approaches, whatever their precise substance, served as syntheses for large groups of labor historians. These old syntheses began to be undermined in the 1950s and the 1960s—a process that continued with even greater force in the following decades. In Britain, Asa Briggs, Eric Hobsbawm, Edward Thompson, and others tried to contextualize workers' struggles. As Hobsbawm wrote in 1964, these historians accentuated 'the working classes as such (… and) the economic and technical conditions that allowed labor movements to be effective, or which prevented them from being effective' (Hobsbawm 1964, p. vii). Somewhat later, similar trends became apparent in other countries. In the United States, Ira Berlin, David Brody, Herbert Gutmann, and David Montgomery, among others, were pioneers.

In France, where the writers of 'evenemential' (*événementiel*) labor movement history had long been opposed by the followers of the *Annales*, whose structural and serial historiography had largely 'left the people out' (to borrow Raphael Samuel's expression), some rapprochement was, especially after 1968, apparent; that was, however, such a slow process that Perrot felt the need to remark in 1979: '(In French historiography) the study of the workers' movement has polarized historians for a long time and eclipsed other problems, such as the development of the working class and its culture. However, this is changing rapidly' (Perrot 1979, p. 150). In Germany, the *Strukturgeschichte*, which was founded in the 1950s by Werner Conze and others (and which incorporated some elements from the National Socialist *Volkssoziologie*, including its focus on the 'structured totality' of the social 'order'), was modified and transformed into a modern approach combining Weberian and Marxian elements by Hans-Ulrich Wehler, Jürgen Kocka and others since the late 1960s.

Gradually, and especially since the 1960s, a broader approach to labor history has thus become widely accepted. The discipline started to incorporate studies of, for example, social mobility, migration, demography, labor processes, schooling, households, urbanization, social insurance, or sexuality. Race, ethnicity, and gender became generally accepted as essential elements of any analysis, but a full integration into mainstream historiography has not yet been accomplished.

Parallel to labor history's thematic expansion its geographical scope has widened. Labor history originated in Western Europe and North America, though there have been important contributions from other parts of the world early on. From the 1920s it became established in the Soviet Union and later in other Communist countries, and gradually struck roots in Latin America (Argentina, Brazil, and Mexico especially), Africa (especially South Africa), the Middle East, India, and Australasia. Labor history's present situation is uneven. On the one hand it has become an acknowledged and integral part of historiography in general and of social history in particular. But on the other hand, its academic conjuncture is strongly linked to political changes, like the collapse of 'socialism' in the Soviet Union and Eastern Europe, and the rise of new workers' movements in Asia.

Both thematic and geographical expansions have made labor history an extremely rich and diversified field. By now, labor history overlaps into women's history, urban history, agrarian history, cultural anthropology, folklore, social economics, history of technology, historical government theories, industrial relations, history of law, and the like. While this disciplinary expansion clearly embodies a wealth of intellectual opportunities, it also leads to a lack of orientation. Already in 1973, Thompson observed that 'The central concern of history, as a relevant humane study—to generalize and integrate and to attain a comprehension of the full social and cultural process—becomes lost' (Thompson 1973).

3. Conceptual Plurality

Labor history in the broad sense reached its peak during the 1970s when former student militants began to write their theses and books. The productivity of this generation was immense and impressive. Perhaps there will never again be written so much about labor history in so few years as at that time. Numerous substantive controversies enlivened the discipline: How should one conceptualize social classes? Are Karl Marx and Max Weber compatible? Is "labor aristocracy" a useful concept? Have 'organized' labor movements suppressed the interests of unskilled workers? How can one integrate women's history and labor history? What is the societal, cultural, and political importance of the working classes? A wide range of issues were hotly debated, in professional journals and in the general media as well.

This peak period did not last long. Most importantly, the organizing categories of the 1960s and 1970s were increasingly called into doubt, partly because they contained continuities with the founding period. Frequently the teleology of the first phase had simply been reserved. While the 'classic' labor historians had been inspired by an optimistic perspective, celebrating the emergence of mass organization and stressing the dynamics of unity and organization, the generation of the 1960s and 1970s tended to ask what had gone wrong with the movement. From a methodological point of view an 'epistemology of absence' became predominant: 'Rather than seeking to explain the presence of radically varying dispositions and practices, [labor historians] have concentrated disproportionately on explaining the absence of an expected

outcome, namely the emergence of a revolutionary class consciousness among the Western working class' (Somers 1989, p. 325).

Ever since, many labor historians regard the state of their discipline as a protracted crisis. First, the emerging paradigms of women's and ethnic history showed that there had been giant blank spots on labor history's map and that the filling in of these spots made a complete rewriting of the old narratives unavoidable. And second, the unilinear conception of class consciousness that had been dominant for a long time was questioned. 'Once it was established that class was a construction and not a predetermined consequence of structural forces, then the notion that it could be seen as the product of an autonomous culture reacting to certain political and economic circumstances ultimately had to be rethought. Class and class action, it became apparent, could work in many different and contradictory ways' (Price 1991, p. 253).

As a result of the growing uncertainty about its organizing categories, labor history is beginning to lose its character as a 'discipline.' On the one hand, the distinction with contiguous disciplines, such as women's studies, ethnic studies, anthropology, institutional economics, sociology, or social psychology, is beginning to dissolve. On the other hand, conceptual difficulties and political disappointments both have stimulated postmodern approaches, especially in the United States.

In North America and Western Europe, including Britain, the debate between labor historians is therefore dominated by a paradigmatic crisis—a crisis that, as a matter of fact, has apparently not always resulted in a strongly reduced output of the 'older generation' of labor historians, but has often caused a diminished interest among students. In Eastern Europe and the former Soviet Union, the problems are even more serious. The collapse of so-called 'socialism' has led to the almost complete demise of classical labor historiography, since the latter is with justification perceived as a constituent of the old dictatorship; generally Communist regimes used a distorted 'official' history of the national workers' movement as part of their ideology of legitimation.

At the same time, exciting new developments take place. In Latin America labor history is booming (French 2000), and in Asia it is rapidly advancing as well. After a group of Indian intellectuals had already published a series of very innovative collections of essays on the social history of the 'subaltern' classes, which focused especially on the peasants, now a generation of sophisticated labor historians has emerged in that country. Interest in labor history is increasing in several other Asian countries as well, parallel to the emergence of large groups of new wage laborers. Partly, these new labor historians grapple with the same conceptual questions as their colleagues in the rich countries, but that does not prevent their intense research activity.

4. *Core Controversies*

There seems to be four interconnected central issues dominating debates within the discipline.

4.1 *Central Categories*

The geographical and thematical broadening of labor history's scope, plus a wider time frame (which now starts to include earlier periods, roughly from the thirteenth century on) makes it necessary to rethink elementary categories, including the discipline's most central notion: labor. In the Anglo-Saxon world, this term has a double meaning. It refers simultaneously to work and to workers' movements.

Labor historians have traditionally conceived 'work' as wage labor, performed by legally free individuals in factories, mines, ports, etc., thus reflecting occupational and gendered views of classical labor movements. Implicitly not included were (a) other forms of free waged labor like domestic service, office work or voluntary military and police service; (b) unfree waged labor of indentured or otherwise bonded workers; (c) hidden wage labor, for example, through self-employment; and (d) nonwaged labor for subsistance or market purposes. Labor history's broadened perspective suggests a reconceptualization of 'work' as any human effort adding use value to goods and services; use value is understood here as 'a good or service that could sustain any activity carried on by a person other than its producer—whether or not we approve of the activity' (Tilly 1998, p. 22–3). This excludes 'purely destructive, expressive, or consumptive acts,' which might be considered as 'antiwork' (Tilly 1998, p. 23) Such a catholic approach makes it possible to overcome old fixations of an institutional or Eurocentric nature.

Traditional labor history has defined a workers' movement as 'all of the organized activity of wage earners to better their own conditions either immediately or in the more or less distant future. In all countries it has run along three lines—political, economic, and cooperative' (Commons 1932, p. 682). A broadening of the concept of 'work' and 'worker' would imply that the notion of a workers' movement has to be broadened as well, and that it should include collective survival activities of all previously excluded categories of workers, whether these are covered by political parties, trade unions, and cooperatives, or not.

4.2 *Class/Gender/Ethnicity*

The relationship between class, gender, and ethnicity has been discussed for several decades. While most labor historians originally considered 'class' as central, and 'gender' and 'ethnicity,' and 'race' as categories of a secondary nature, gradually a new consensus has grown which stresses that gender identity, race ident-

ity, and class identity have 'shaped each other' (Roediger 1993, p. 131) But despite this new understanding, the *simultaneous* historical analysis of the three identities has proved to be extremely difficult. Following Curthoys one is tempted to see a fundamental indeterminacy problem here: 'Trying to keep just two of these concepts in play has proved extremely difficult. […] But if keeping two such concepts in play is hard enough, look what happens when the third concept, be it ethnicity or class or sex, is brought seriously into play. The system, the analysis, becomes too complex to handle' (Curthoys 1991, p. 15).

4.3 Structure vs. Actor

A second fundamental indeterminacy problem has a strong influence on labor history debates as well: the relationship between the actor-oriented and the structure-oriented perspective. Classical labor history focused primarily on the agents of social changes: workers, their organizations, and leaders. More recent approaches have paid much attention to broader developments, like, for example, demographic shifts, socioeconomic structures, or governmental bureaucracies. Although few would contest that both modes of interpretation are essential for labor history, their integration is highly problematic.

Even authors who try to combine the two levels (e.g., by paying equal attention to the *longue durée* and short-term events), have to accept that they are in fact telling two separate but combined stories instead of constructing one integrated narrative. The more a scholar focuses on one aspect, the more he or she tends to lose sight of the other aspect. Elton once observed: 'Despite attempts to deny this, it [quantitative history] can effectively operate only by suppressing the individual by reducing its subject matter to a collectivity of human data in which the facts of humanity have real difficulty in surviving' (Elton 1983, pp. 118–9). The reverse could be said as well.

4.4 Great vs. Small Stories

Classical labor history interpreted the history of working classes and labor movements against the background of major long-term material and cultural changes, for example, the general process of modernization or the rise of capitalism. Since the 1970s this perspective has been challenged from at least two sides.

First, there has grown a strong tendency to critize such broad interpretative frameworks as 'centrist.' It has been argued that such a way of seeing 'positions historical phenomena either to the center or the margins of the historical process, according to their supposed role in the 'great transformation.' Against this unilinear approach, the critics juxtaposed the emphasis on experiences and processes that do not fit in general schemes. Such a 'de-centered' research strategy is expected to lead to 'a greater awareness of the uniqueness, difference, and otherness of historical phenomena, something which is rather obscured by the application of universalizing, passepartout categories (such as role, economic growth, bureaucracy)' (Medick 1987, p. 82).

Second, classical labor history has been disputed by the so-called 'linguistic turn.' This point of view claims that the backgrounds and interests of the members of a workers' movement cannot explain its development. Instead the language used by the movement's members should be considered as the driving force, as 'a complex rhetoric binding together, in a systematic way, shared premisses, analytical routines, strategic options, and programmatic demands' (Jones 1983, p. 107, cf. Scott 1988, pp. 53–67). Discourse analysis therefore replaces the study of larger material and cultural developments.

Both challenges have directed scholarly attention to previously neglected aspects of labor history. But they have also contributed to the discipline's fragmentation. If discourse is disconnected from its context and considered as 'non-referential' (Jones 1983, p. 22), or if the concept of historical master processes is contested, then labor history becomes the study of disjointed processes and texts, confirming the lack of a 'central concern' deplored by Thompson (1973). An alternative way of developing labor history further might be through maintaining the idea of master processes and of the contextual situatedness of discursive activities, but by simultaneously taking 'non-central' and linguistic aspects seriously (Kocka 1997).

5. Globalization

While the development of labor history is so problematic and uneven, in addition and in connection to this, another trend is apparent: the strong increase in the need to go beyond the national boundaries. The more we know about the history of the working classes in an increasing number of countries, the more tempting it is to place the various national developments in a broader context. Two basic reasons for that can be put forward.

In the first place, we can only discover what is specific and what is generic in our own history by looking beyond national borders. A second important reason for the crossing of national borders is that the object of the investigation is hardly concerned with these borders. Working-class formation is not neatly restricted to certain national borders, but is a process in which voluntary and forced immigration and emigration have a great influence. Dramatic developments in one country may cause turbulences in other countries; strike waves often have a transnational character; new forms of campaigning are imitated elsewhere; national labor movements communicate

with each other, learn from each other, and create international organizations.

The growing need for studies that look beyond national borders has resulted in a tremendous growth of the number of contributions to transnational labor history, which started in the 1970s. Yet the work of most labor historians is still characterized by a distinct 'methodological nationalism' which prestructures their writings, with an often rather strict separation of studies of the United States and of other countries.

See also: Gender History; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Labor Movements, History of; Labor Unions; Social History; Social Inequality in History (Stratification and Classes); Social Movements, History of: General; Social Protests, History of; Socialism: Historical Aspects; Working Classes, History of

Bibliography

Commons J R 1932 Labor movement. In: *Encyclopaedia of the Social Sciences*. Macmillan, New York, Vol. 8; pp. 682–96

Curthoys A 1991 The three body problem: Feminism and chaos theory. *Hecate* **17**(1): 14–21

Elton G R 1983 Two kinds of history. In: Fogel R W, Elton G R (eds.) *Which Road to the Past? Two Views of History*. Yale University Press, New Haven, CT, pp. 71–121

French J D 2000 The Latin American labor studies boom. *International Review of Social History* **45**: 279–308

Hobsbawm E J 1964 *Labouring Men: Studies in the History of Labour*. Weidenfeld and Nicholson, London

Hobsbawm E J 1984 *Worlds of Labour: Further Studies in the History of Labour*. Weidenfeld and Nicholson, London

Jones G S 1983 *Languages of Class: Studies in English Working Class History, 1832–1982*. Cambridge University Press, Cambridge, UK

Kocka J 1997 New trends in labour movement historiography: A German perspective. *International Review of Social History* **42**: 67–78

McCulloch J R 1851 *A Treatise on the Circumstances which Determine the Rate of Wages and the Condition of the Labouring Classes*. London

Medick H 1987 Missionaries in the row boat'? Ethnological ways of knowing as a challenge to social history. *Comparative Studies in Society and History* **29**: 76–98

Palmer B D 1990 *Descent into Discourse: The Reification of Language and the Writing of Social History*. Temple University Press, Philadelphia, PA

Perrot M 1979 The three ages of industrial discipline in nineteenth-century France. In: Merriman J M (ed.) *Consciousness and Class Experience in Nineteenth-Century Europe*. Holmes & Meier, New York and London, pp. 149–68

Price R 1991 The future of British labour history. *International Review of Social History* **36**: 249–60

Roediger D 1993 Race and the working-class past in the United States: Multiple identities and the future of labor history. In: Van Der Linden M (ed.) *The End of Labour History?* Cambridge University Press, Cambridge, UK, pp. 127–43

Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York

Somers M R 1989 Working class formation—19th century patterns in Western Europe and the United States. *Contemporary Sociology* **18**: 325–9

Subaltem Studies 1982–96 Oxford University Press, New Delhi, 9 Vols

Thompson E P 1973 Responses to Reality. *New Society* **4 October**

Tilly C, Tilly C 1998 *Work under capitalism*. Westview Press, Boulder, CO/and Oxford, UK

Webb S, Webb B 1897 *Industrial Democracy*, 2. Vols. Longmans, Green, New York and Bombay, India

M. van der Linden

Labor Markets, Labor Movements, and Gender in Developing Nations

The women's movement has made the world aware of the importance of gender in the world's labor markets. However, definitions and methods of enumeration still leave out a large section of the female work force. Even where women are included as part of the labor force they face highly segregated labor markets. Since about 1980, globalization and structural adjustment programs have negatively impacted labor movements in developing nations. However, in the process of dealing with change, international trade unions have realized the importance of organizing women workers especially in the areas of agriculture and the informal sector. These are areas of growth of the labor movement for women in the future.

1. The World's Labor Force

The world's labor force has been increasing at a fast rate in the recent decades and almost doubled between 1965 and 1995, from 1.3 billion to 2.5 billion (World Bank 1995). As Table 1 below shows, the largest labor force is in the poorest countries.

In most parts of the world, most employment is still in the nonwage, mainly agricultural sector. As Table 2 reveals, whereas employment in industrialized countries tends to be in industry or services and workers are earning a wage (USA nearly 90 percent, France 74 percent), the employment in the developing world tends to be in the nonwage sector (China 84 percent, Ghana 85 percent). So, it must be remembered that when we discuss the labor force in the developing world we are talking about a labor force that is either

Table 1
The world's labor force by region

Region	No. of Workers (in millions)	Percent of total labor force
High income OECD	373	15
Non-OECD Europe and Central Asia[a]	239	10
Middle East and North Africa	80	3
Sub-Saharan Africa	214	9
South Asia	440	18
East Asia and Pacific	964	39
Latin America and Caribbean	166	6

Source: World Bank 1995
a The European countries in this category are the non-OECD countries, generally lower income

Table 2
Distribution of labor force by sector in selected countries

Country	Agriculture wage	Agriculture nonwage	Industry wage	Industry nonwage	Services wage	Services nonwage
China	1	60	4	14	11	10
Indonesia	6.5	43.9	10	5.8	14.3	19.5
India	1.5	61.7	3.7	10.5	9.4	13.2
Egypt	6.3	36.2	15.6	5.1	27.2	9.6
Ghana	3.2	57.8	3.2	9.6	9.7	16.5
Brazil	9.4	14.8	19.4	4	37.3	15.1
United States	1.6	1.3	24.2	1.6	65.5	5.8
France	2.3	5.9	27.7	2.5	56.2	5.5

Source: World Bank 1995. Numbers of workers in millions. Data pertain to years between 1988 and 1993

predominantly self employed, in market or subsistence activities, or works in family nonpaid employment.

2. *Women and Work*

Women are part of the labor force in every country, but many aspects of the work they do are very different from men. These differences include type of work, place of work, hours of work, and even attitude towards work. Because of these differences, the work done by women is often not recognized as work, and sometimes women are undercounted in official labor force statistics. In the 1990s there was a concerted attempt from women's organizations and academics to move towards a wider definition of work which would include much of the work that women do.

Women work in fields, in factories, in offices, in mines, on waterways, much as men do, but unlike men, the household is also a major center for economic activity for women. Home-based workers, such as garment workers or leather makers, work within their homes, as do subsistence farmers and animal tenders. Women often undertake unpaid production of goods for the market as part of the family business. In addition, women usually do unpaid household work like cleaning and cooking. Government enumerators and even the women themselves often do not understand that other than unpaid housework, all other types of work should be counted as part of the market economy according to the official definitions of the 'labor force' stipulated by governments and international agencies. Because of this wide variety of work, and because of the differing perceptions of what constitutes work, women's active participation in the economy tends to remain undercounted.

Table 3
Women's percentage share of the total labor force in 1990, by region

Former USSR	59
North America	41
Eastern Asia	40
Australia, Europe, Japan and New Zealand	39
Sub-Saharan Africa	38
Latin America and Caribbean	29
Oceania	29
South Asia	20
North Africa and West Asia	18

Source: United Nations 1995

The total number of economically active women in 1990 was estimated to be 828 million. However, women's share of the labor force, as counted in the population censuses, varies widely from region to region, as Table 3 shows.

Women's share of the labor force is higher in those countries where the majority of the labor force is in the

formal wage sector. In agricultural societies and where the informal sector is a high proportion of the labor force, women's work tends to be around the household and the family farm and hence undercounted.

> It is not easy in a population or household survey to decide whether a person is working or not, and what people are working at, particularly women in developing regions. Terms such as employment, job, work and main activity mean different things to different people. Interviewers also contribute to bias when they assume that women or children are not economically active. Persons who work for a wage or a salary are usually enumerated as economically active regardless of how the questions are phrased. The complications arise as one moves from wage and salary employment to self employment—especially in the informal sector—and to unpaid labor in a family farm or business: that is from more male-dominated to more female-dominated forms of economic activity. For example, a methodological survey in India found that 13 percent of adult women were wage or salary earners. When market oriented production was included—for example, family business, self-employment, crafts and agricultural activities—the figure rose to 32 percent. When the new ILO (International Labour Organization) standard was fully applied, 88 percent of women were found to be economically active (United Nations 1995).

Even women in the formal wage economy generally are segregated in different occupations from men. A recent cross-national study (Anker 1998) found that occupational segregation by sex is extensive in every country. Approximately one half of the workers in the world are in an occupation where one sex dominates to such an extent that the occupation could be considered a 'male' or 'female' one (an extent of more than 80 percent). In every country, female occupations generally are seen as less 'valuable' with lower pay, lower status, and fewer advancement opportunities. For example, 88 percent of managers, legislative officials, and government administrators are men. Occupations open to women are consistent with stereotypes of women. However, levels of occupational segregation vary across regions with the Asia/Pacific region having the lowest level and the Middle East/North Africa region, the highest.

3. *Globalization*

Since the 1970s, there has been dramatic growth of global trade. Markets across countries and regions are becoming increasingly linked, so those workers in even remote rural areas of developing countries are being affected increasingly by the trends in the global product and finance markets. Financial and capital flows also increased as a proportion of GDP. There has been a large increase of capital outflow from industrialized countries since 1980. Among the causes of this growth of trade are reductions in trade barriers through regional or multilateral trade agreements; rapid expansion of foreign direct investment by multinational companies; reduced barriers to international capital transfers, resulting in large flows of equity funds and other speculative investments across national borders; and diffusion of communication facilities that have dramatically increased the speed of transfers.

Since about 1980, globalization, along with structural adjustment programs, brought many changes in the structure of employment and the opportunities available. In many of the developed countries, the unskilled jobs have been transferred to the developing world, concomitant with a rise in opportunities for the skilled and professional workers at home. In Asia there was a rise in employment opportunities, wage levels, and skills, especially for younger women. However, after the 1990s economic crisis in Asia, there has been a major setback for employment. Poverty levels, which had fallen below 20 percent, have risen dramatically, to as much as 60 percent in Indonesia. Latin American countries too, had major upsets with very high inflation rates, high unemployment, and falling well-being. Countries in sub-Saharan Africa were hit the hardest with falling income per capita, high unemployment, and social unrest.

4. *Informal Sector*

The period of globalization has also seen the growth of the informal sector in all regions of the world. Since the term 'informal sector' was coined in the 1970s, it has been defined in various ways and gone in and out of favor in development circles. One dominant school of thought predicted its demise, arguing that as GDP per worker increased, the share of the work force in the formal sector would rise. However, due to slow growth, economic reforms, and/or changing patterns of production, the informal sector has grown steadily in the developing world (and in some countries of the developed world). In the 1980s and 1990s, the share of the work force in the formal sector slowed down or absolutely declined in most developing countries. It was only the rapidly growing economies of East and Southeast Asia that experienced substantial growth of formal sector employment (although the Asian economic crisis of the 1990s led to lay-offs in the formal sector and dramatic growth in the informal sector).

As an umbrella concept for economic activities that fall outside the modern and (in many countries) the agricultural sector, the informal sector has been defined in various ways. Alternative definitions emphasize the fact that informal sector activities rarely comply with official and administrative requirements; are not adequately reflected in official statistics; and are not covered by protective legislation or trade union organization. In an attempt to encompass the heterogeneity and complexity of the informal sector, the 15th International Conference of Labor Force Statisticians in 1993 adopted an international definition of the informal sector to be used in the collection

Table 4
Percentage of enumerated labor force that is in informal sector

Informal sector	La/C	Africa	Asia
Share of:			
Total employment	27	89	85
Urban employment	25	61	40–60
Poor employment	50	91	78
New jobs	54	93	NA

Source: WIEGO 1999. La/C refers to Latin America and the Caribbean

and presentation of labor force data. According to this 1993 international definition, the informal sector comprises (a) informal self-owned enterprises (with only family workers or temporary employees) and (b) informal enterprises with more permanent employees. Under this definition, whether 'informal' is restricted to nonagricultural activities, nonregistered activities, and/or enterprises of a given size is left up to individual countries. Whichever defining characteristics are adopted by individual countries, several sets of activities are particularly difficult to capture in national statistical surveys: itinerant or seasonal or temporary jobs (e.g., on construction sites or road works); secondary activities; and street vending. In addition, no matter what defining characteristics are adopted by individual countries, one set of workers is left out of the 1993 definition: subcontract workers or outworkers who work for formal firms. These limitations of official data notwithstanding, the estimated size of the informal sector is quite large, as shown in Table 4.

In India, where the official statistics are relatively good, it is estimated that the informal sector accounts for 90 percent of the total workforce and 63 percent of GDP and the household sector (of which the informal sector is the major share) accounts for 76 percent of savings (WIEGO 1999).

If home-based workers and street vendors (particularly women in these sectors) were more adequately reflected in official statistics, the size and contribution of the informal sector would be larger still. In Latin America and Asia, microstudies suggest that home-based workers (both own account and subcontract workers) account for 40–50 percent of the workforce in key export industries—textiles, garments, and footwear. In West Africa, national surveys suggest that street vendors account for more than 33 percent of the informal urban labor force and more than 30 percent of the total urban labor force.

5. *Genesis of Labor Movements*

The Labor Movement as we know it today began to emerge in Europe and the United States in the middle of the nineteenth century as a response to the massive industrialization which gave rise to a new class of the 'proletariat.' The first trade unions represented skilled male workers, and reflected their origins in earlier craft guilds. These trade unions developed support and welfare provisions (e.g., in times of unemployment) for their members and helped them find employment. Trade unions that developed later in the nineteenth century were more likely to organize all workers in the industry, particularly unskilled and general workers. This later trend was very much associated with general ideas of socialism that were strong at the time and linked to ideas of cooperation as well as trade unionism.

As industrialization came to the developing world there were spontaneous struggles by workers to organize for better work conditions. However, it was only after 1920 that the labor movement developed in the third world, most of whose countries were still under colonial rule. 1920 was momentous for the labor movement for two reasons. First, the International Labor Organisation was set up as a tripartite body with membership of National Trade Unions, Governments, and National Association of Employers, to regulate the conditions of work and provide a forum for dialogue. At the same time the newly formed USSR had followers in communist parties in most countries, who began organizing within the labor movement. The next 50 years saw the consolidation of labor movements in most developing countries along with freedom movements, which led to newly independent countries in most of Asia, Africa, and Latin America. Many of the newly independent countries included the Right of Association as a fundamental right in their Constitutions, creating a legal basis for the newly formed trade unions. Labor laws were also passed in most countries to protect the working conditions of the workers. At the same time most countries, influenced by the economic successes of the Soviet Union, instituted central economic planning and large public sector organizations. Trade unions flourished in these. At the national level in many countries, individual trade unions attempted to come together to form national federations. These federations, with their large memberships, became politically powerful in many countries. The mid-twentieth century was the heyday of the labor movement worldwide.

6. *International Labor Movements*

The labor movement has always recognized the importance of international links and trade unions in particular trades have attempted to organize across countries. For example, the IUF is one of 14 international trade secretariats that unite workers on the basis of their industry, craft, or occupation. Founded in 1920 through the merger of international federations of bakery, brewery, and meat workers, it was first known as the International Union of Food and Drink Workers. Its headquarters were located in

Zurich and membership was almost entirely European until World War II. Today the IUF represents 329 organizations in 118 countries, in all regions of the world. These organizations represent approximately 10 million workers. However, it was during the period when the trade union movement was at its height that that the major international trade union federations were formed. The International Confederation of Free Trade Unions (ICFTU) was established in 1949, with National Federations of noncommunist trade unions as its members. The World Federation of Trade Unions (WFTU), based in Moscow, was the international Federation of trade unions with a communist philosophy. In addition there were smaller international federations such as the Christian Labor Federation. Today the ICFTU represents 213 organizations in 143 countries on all five continents, with a membership of 124 million, of whom 43 million are women.

7. *Decline of the Trade Unions*

Globalization of the international economy has coincided with the decline of the power of trade unions worldwide. In the industrialized countries, the profile of production began to change. Multinational corporations moved their production into the developing countries. In these affluent nations, self-employment began to increase, employment in manufacturing and blue collar jobs declined, and service sector, professional, and white-collar employment grew. As a result, the proportion of the labor force in trade unions declined.

In developing countries, new trends began to be seen. East and South East Asia began developing at a fast pace with growing exports. The governments of these countries more or less suspended democratic functioning and with it the freedom for workers to create and join trade unions of their own choosing. In South Asia, the trade unions had always been confined to the formal sector—the private corporate and public sector enterprises—covering barely 10 percent of the workforce. As globalization and liberalization policies spread, these sectors shrank instead of expanding as had been expected, so this shrank the proportion of the national workforce unionized. In most of Africa (other than South Africa, which has had a unique history of unions being in the forefront of anti-apartheid struggles), the small trade unions, which in any case were confined to a small section of industry, were suppressed as antidemocratic forces gained the upper hand. In Latin America, countries with large industrial sectors or with large plantations had sizable unions with militant histories. However, as liberalization and globalization policies spread many of these unions were attacked. In some cases, individual unionists were killed. In the USSR and Eastern Bloc, trade unions were controlled by the State, but in many countries these unions became the vehicles for anti-government activity leading to a collapse of the USSR and change of regimes in the former socialist countries. Table 5 shows that only a small minority of workers are covered by unions in a number of regions.

Table 5
Trade union membership as a percentage of labor force in selected countries

Country	Percent
Australia	31
India	8
Kenya	9
Morocco	12
Mexico	32

Source: World Bank 1995

The reasons for decline of the labor movements have been both external and internal. One factor has been the globalization of production which led to decline of production in the countries of strong unionization, as well as the casualization of production which led to the closing of unionized factories and work going into small and contract production. A second factor has been the antilabor policies followed by governments in most developing countries. This has been most pronounced in South East Asia, Latin America, and some of the African countries.

The labor movement has been subject to a number of major internal weaknesses. Perhaps the major weakness has been the confining of the movement to large public and private enterprises. In the earlier years of development, it was optimistically believed that the economic structures of developing countries would change to become more like those of industrialized countries, i.e., with most of the workforce in manufacturing or services in the formal sector with a formal employer-employee relationship. However, instead of shrinking the informal sector has in fact expanded, and the difference in earnings, access to social security, and security of work, between the formal and informal, unionized, and nonunionized workers has grown. The labor movement was unable to reach out and organize the workers of the informal sector and as a result their own influence has continued to shrink.

8. *Women and Labor Movements*

In the early days of industrialization, when working conditions were bleak and hours were long, women were an important, though generally unskilled part of the industrial work force. However, the first trade unions were meant for skilled workers and based on a principle of exclusivity—protecting the employment of the skilled workers and keeping out unskilled. In practice, this often meant excluding women. This tradition went on until the early twentieth century

when, for example, the engineering union would not admit women to their membership, even when they made up a majority of the workforce in a factory. Male union workers also adopted the ideology of the 'family wage,' arguing that employers should pay male workers enough that they could support wives and children. Part of this ideology was that the woman belonged in the home and the man was the breadwinner.

The second wave of trade unionism started with the textile industries in England where a majority of workers were women. At the end of the nineteenth century there was a great movement to organize all workers and this is what led to the development of the general unions which organized the 'unskilled' and 'semiskilled,' including some women. However, even in this period, many women workers were excluded. Most industries and professions excluded married women and the biggest source of employment for women was still domestic service, completely unorganized. As a result of the exclusion of women from some trade unions and the fact that women were in many different forms of informal employment, in addition to trade unions there were strong women's organizations—the Women's Trade Union League and Women Workers' Federations. Some advocated women joining general unions, others favored women having their own unions.

After World War I, unions gradually dropped their ban on women members and the separate women's organizations declined. However, married women were still not usually allowed to work formally. Moreover, as the trade unions were able to get better wages and working conditions for their members the numbers of women in most unionized industries declined due to a number of reasons. First, the societal belief was that the males in the family should earn while the women were homemakers. Unions demanded a 'family wage' so that women could stay home. As wages and working conditions improved, men replaced women in the workforce. Second, protective legislation, such as limiting number of hours worked, night shifts, prohibition of entry into certain jobs, began to limit the scope of work available to women. Finally, new technologies demanded new skills for which men were able to get the training denied to women. Women's employment in modern and formal industries declined, and with it the participation of women in the trade unions.

The 1970s, however, saw the growth of a powerful world wide women's movement. This feminist movement, as it spread to developing countries, brought more skilled and educated women into the labor force. Women's employment in white-collar work began increasing, and with it women's consciousness of their own rights. In the 1980s women in the trade unions too began to voice the need for greater representation. Women began to demand more power within the trade unions and in the 1990s the international trade unions as well as trade unions at the national level began to focus on drawing more women as active members of trade unions. The ICFTU, Programme of Action for Integration of Women into Trade Union Organisations was first adopted in 1978, and updated in 1985 and 1988. The 15th World Congress in 1996 adopted resolutions on 'Equality' and 'Women and Development' calling on ICFTU and its affiliates to integrate gender perspectives into all their work, and to adopt a positive action program for women. The ICFTU reports that women's share of its affiliates is 34 percent and that women's membership has been growing in all regions in the 1990s.

However, the majority of women in developing countries are in agriculture and the informal sector. These are areas where the labor movement is still weak, and so they have not yet been integrated into labor movements.

9. *The Future of Women in Labor Movements*

Unions in developing countries have begun to recognize the importance of organizing the informal sector. However, these small unions have little or no voice at the national or international level. A number of international movements are now developing that are trying to represent women who work in the informal sector in developing nations. Among these are Home-Net, the international alliance of home-based workers, the International Alliance of Street Vendors, and the World Fisherpeople's Federation. If these alliances are examples of the future, the new labor movement will be one in which workers of all types, not just 'employees,' will be represented, and where the whole concept of work and worker will undergo a change to bring the large majority of informal sector women into the mainstream of the economy.

See also: Economic Development and Women; Economic Globalization and Gender; Feminist Movements; Labor History; Labor Movements and Gender; Labor Movements, History of; Labor Unions; Lone Mothers in Nations of the South; Nation-states, Nationalism, and Gender; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender; Social Class and Gender; Social Movements and Gender; Trade Unions, Economic Behavior of; Trade Unions: Empirical Analyses; Work, Sociology of

Bibliography

Ackerman F, Goodwin N R, Dougherty L, Gallagher K 1998 *The Changing Nature of Work*. Island Press, Washington, DC

Anker R 1998 *Gender and Jobs: Sex Segregation of Occupations in the World*. International Labor Office, Geneva, Switerland
Asian Development Bank 1998 *Annual Report*. Oxford University Press for the African Development Bank, Oxford, UK
Baksh-Soodeen R 1994 Caribbean feminism in international perspective. *Mumbai Economic and Political Weekly*, October 29
Brtiskin L, McDermott P 1993 *Women Challenging Unions*. University of Toronto Press, Toronto, ON
Coo A H, Lorwin V R, Daniels A K 1992 *The Most Difficult Revolution: Women and Trade Unions*. Cornell University Press, Ithaca, NY
Cunnison S, Stageman J 1995 *Feminizing the Unions*. Avebury, Aldershot, UK
Figueirendo J, Shaheed Z 1995 *New Approaches to Poverty Analysis and Policy—II*. International Labour Office Publications, Geneva, Switzerland
FNV/CNV 1997 *Organizing Change: Strategies for Trade Unions to Organise Women Workers in Economic Sectors with Precarious Labour Conditions*. Stichting FNV Press, Amsterdam
Hosmer Martens M, Mitter S 1993 *Women in Trade Unions, Organizing the Unorganized*. ILO, Geneva, Switzerland
International Confederation of Free Trade Unions 1970 *ICFTU Twenty Years (1949–1969)*. International Confederation of Free Trade Unions, Brussels, Belgium
International Confederation of Free Trade Unions 1986 *Working Women: ICFTU Policies and Programmes*. International Confederation of Free Trade Unions, Brussels, Belgium
International Labour Office 1995 *World Employment*. International Labour Office Publications, Geneva, Switzerland
International Labour Office 1997 *World Employment*. International Labour Office Publications, Geneva, Switzerland
International Labour Office 1999 *World Employment Report 1998–99*. International Labour Office Publications, Geneva, Switzerland
International Labour Office (FEMMES) 1999 *Gender Issues in Labour Market Policies*. Unpublished www document.
Mariam H Y 1994 Ethiopian women in the period of social transformation. Mumbai Economic and Political Weekly, October 29
Mathur V S 1976 *Changing Role of Trade Unions in Developing Economies*. International Confederation of Free Trade Unions Asian Regional Organisation, Green Park
Ramaswamy E A 1995 *Six Essays for Trade Unionists*. India Office South, New Delhi, India
Rodgers G 1995 *New Approaches to Poverty Analysis and Policy—I*. International Labour Office Publications, Geneva, Switzerland
Rodgers G, van der Hoeven R 1995 *New Approaches to Poverty Analysis and Policy–III*. International Labour Office Publications, Geneva, Switzerland
Shah H, Gothoskar S, Gandhi N, Chhachhi A 1994 Structural adjustment, feminisation of labour force and organisational strategies. *Mumbai Economic and Political Weekly* October 29
Thomas H (ed.) 1995 *Globalization and Third World Trade Unions: The Challenge of Rapid Economic Change*. Zed Books, London
United Nations 1995 *The World's Women*. United Nations, New York
WIEGO 1999 *Report of the First Annual WIEGO Meeting*. Ottawa, Canada, April 1999
Worku Z 1994 Organising women within a national liberation struggle. *Mumbai Economic and Political Weekly*, October 29
World Bank 1995 *World Development Report*: *Workers in an Integrating World*, World Bank, Washington, DC

R. Jhabvala

Labor Movements and Gender

Labor movements include trade unions and other formal organizations that seek to improve the wages and conditions of workers in industrial societies, as well as a variety of less durable social movements that share this goal. The relationship of labor movements to gender is complex and highly variable. In some times and places, labor movements have actively advanced the struggle for gender equality; in other instances, they have contributed instead to the reproduction and maintenance of gender subordination. The scholarship on the subject reflects this ambiguous legacy.

Like much of the early interdisciplinary feminist research that appeared in the 1970s and 1980s, the first wave of scholarship on the topic of gender and labor movements was largely compensatory. It sought to demonstrate that from the earliest period, all over the world, women had been actively engaged in strikes, unionizing efforts, and other collective actions intended to advance the interests of working people—even though these female contributions were nearly invisible in the literature on labor history. This work emerged alongside and was nurtured by the (then) 'new' social historians' critique of institutional labor history, as well as by the revisionist labor history and neo-Marxist sociology of the 1970s and 1980s that constructed trade unions as essentially conservative institutions (see Buhle 1990).

Challenging the conventional wisdom that women were less easily mobilized into labor movements and less active in trade unions than their male counterparts, this early scholarship recuperated the previously hidden history of women's labor struggles. It also documented the ways in which labor movements themselves had been historically—and were in the present—controlled by men, who seldom attempted to organize women, and in many cases actively excluded them from union membership. But when opportunities to mobilize and to join trade unions were available, women workers were responsive and by some accounts even more militant than men, this first wave of research showed (Kessler-Harris 1975). This perspective turned on its head the earlier view that women were a problem for labor movements, arguing instead that labor movements were problematic for women, due to institutionalized male domination. That, in turn,

raised the issue of why labor movements were male dominated. Two distinct (although not mutually exclusive) explanations emerged from the initial phase of inquiry. Both linked male domination of unions and other labor movement organizations to gender inequality in the workforce and in the wider society, but one set of commentators emphasized structural and the other cultural factors. In the structural perspective, best exemplified by Hartmann's work (1976), women's exclusion from trade unions and their subordination within labor movements was understood as a linchpin of the larger patriarchal order. In this view, male-dominated unions play an active role in maintaining and perpetuating gender inequality throughout the society, most importantly confining women to marginal, poorly remunerated forms of paid employment and at the same time reinforcing the asymmetrical gender division of unpaid domestic labor. Male workers, Hartmann and others argued, have a material interest in perpetuating their power 'as men,' and labor organizations like trade unions are a key instrument through which they pursue this interest.

The parallel, cultural explanation for the relative marginality of women to labor movements focused less on the material interests of male workers than on their cultural domination of trade unions and other labor movement institutions. It took as its starting point the notion that male and female workers, due to their contrasting social positions and their sex-segregated experience in the labor force, rely on different, gender-specific understandings of their relationships to work and to collective action focused on labor issues. In this view, labor movements have typically been linked to a specifically masculine iconography and discourse that celebrate aggressiveness and violence, and to an informal work culture rooted in exclusively masculine cultural repertoires. Women workers, to be sure, also have a rich work culture that is often sharply opposed to management, but this is organized around such concerns as family, children, and romance which are quite remote from the agendas of unions and other labor organizations (Westwood 1984). Indeed, this cultural perspective suggested, labor movements are often perceived by women workers as exclusively masculine, not only because of their militant rhetoric and overtly adversarial posture toward employers, but also because they typically operate on male cultural terrain, for example, holding union meetings at night, and in bars, so that women must compromise their respectability if they wish to attend.

These two approaches, separately or in combination, dominated the first wave of literature on gender and labor movements. They were appealing in their apparent comprehensiveness. They seemed to explain not only women's under-representation in the ranks of labor activists and among trade union members, but also their general exclusion from leadership positions in labor organizations and the scant attention paid to 'women's issues' by labor movements. Yet despite the insights these perspectives offered, they could not explain the wide range of variations in the relationship of labor movements to gender documented in the growing body of historical and sociological feminist scholarship that burgeoned in the 1970s and 1980s. As this limitation became increasingly obvious, new approaches to the subject emerged.

One of the new perspectives re-interrogated the cultural terrain that was previously viewed as separating women from male-dominated labor movements. Drawing on the concept of 'women's culture' in feminist historiography as well as on ethnographic and historical research, such scholars as Kaplan (1982) argued that 'female consciousness,' rooted in traditional domestic concerns, was not antithetical to labor movement participation but in fact could actively propel women into broad, community-based labor struggles. Other scholarship has linked women's work culture to a distinctively female style of leadership in labor organizing and to the mobilization of women workers into otherwise conventional modes of collective action, suggesting that women's culture and labor movements may not be incompatible—contrary to the earlier claims that stressed the decisively masculine cultural character of labor movements. From this apparent paradox a key question emerged, namely, under what conditions have labor movements been effective vehicles of collective action for women workers?

Not only was the presumption that the culture of labor movements and 'women's culture' are incompatible falsified, but also, the notion that the structure of labor movements was inherently male-dominated and a critical force contributing to the reproduction of patriarchy was increasingly questioned in the 1980s. Scholars began to uncover instances where male workers and labor movements dominated by them struggled on behalf of women workers and for gender equality, rather than the reverse. Indeed, sometimes male workers' gender interest in maintaining male domination has taken precedence, but in other instances men's class interest in gender equality has prevailed instead (Milkman 1987). Once again the question became: under what conditions is each of these possible outcomes likely to occur? Thus the deterministic grip of the structural and cultural explanations for male domination of labor movements and female marginality within them was increasingly loosened in favor of a more contingent approach.

Some commentators, accordingly, began to focus on variations over time in the relationship of labor movements to gender. One common line of argument constructed the shift from a craft-union dominated labor movement (which was comprised mainly of skilled workers and often excluded women and ethnic

minorities) to a more inclusive industrial union movement as the basis for a significantly more woman-friendly labor movement in many countries, even if some aspects of male domination persisted (Cook et al. 1984). Indeed, an evolutionary logic underlies much of the literature, explicitly or implicitly, according to which there is progress over time toward more gender-egalitarian and inclusive forms of labor mobilization and unionism. As the workforce has become increasingly feminized, and with the increasing challenge to gender inequality in the wider society, the claim here is that labor movements too have become increasingly open to women's participation and that their commitment to gender equality has grown (Milkman 1990).

But recent research suggests a more complex and uneven pattern of historical development. For example, Cobble (1991) has shown how 'occupational unionism' among US food service workers in the craft union era offered important opportunities for female mobilization that were later obliterated as industrial unionism became the hegemonic mode. While the unions she describes were sex-segregated, all-female groupings, such an arrangement proved to have important advantages for women workers in this particular context. For early twentieth-century waitresses, Cobble argues, 'an organizational structure based on the logic of craft, rather than being incompatible with female mobilization, provided instrumental in its creation and maintenance.' Other scholars have highlighted the effects of labor movement bureaucratization, arguing that the more diffuse, community-based organizing that is typical of the pre-institutional phase of labor movement development was more accessible to women than the later, more bureaucratic phases. In the US, for example, Frank's work (1993) shows how the broad-based labor movement that existed in Seattle in the 1920s, which included labor cooperatives, consumer boycotts, and 'union label' campaigns, offered a vehicle for women's involvement rooted in domestic and community-oriented concerns, all of which became peripheral with the later triumph of a highly bureaucratic type of 'business unionism.' More generally, recent literature documents substantial variation in the gender dynamics of labor movements and in the gender composition of labor organization memberships, intra- and internationally as well as over time. While there have been some efforts to systematically explain such variation there appears to be a strong element of historical contingency as well.

A theoretical approach that emphasizes the embeddedness of gender in social institutions, practices and culture has become increasingly influential in recent years (Scott 1988; Baron 1991). Deeply influenced by poststructuralism and postmodernist theory, feminist scholars working from this perspective excavate the ways in which gendered subjectivities are constituted both inside and outside the workplace and how these influence labor struggles. They see gender itself as an object of contestation and at the same time as a primary signifier of power relations more generally. This work has been valuable in underscoring the variable content of gender categories over time and space and the ways in which labor struggles both reflect and transform those categories.

Another focus of increased scholarly attention in recent years is the way in which gender interacts with race, ethnicity, and immigration status in the context of labor movements. The picture that emerges from the accumulated case studies is complex. Like their male counterparts, women workers often have constructed or defended exclusionary ethnic, racial, and citizenship-based boundaries within labor movements. Yet in the US, African–American women are far more likely to be union members than white women, in part because they are more often employed in unionized occupations and industries (especially in the public sector). And perhaps because today job segregation by gender tends to be even more extensive than job segregation by race, ethnicity, or citizenship, there are many examples of labor solidarity among women of diverse racial, ethnic, and national backgrounds.

In the final decades of the twentieth century, labor movements came under vigorous attack all over the world, a development that largely coincides with the global trend toward increased (although cross-nationally uneven) female labor force participation. The ironic result is that in many countries, labor movement membership has become increasingly feminized, and unions have become more concerned with gender issues and with promoting gender equality, in precisely the period in which the power and influence of labor movements have been dramatically eroded. In the US, for example, in 1999 only 14 percent of all wage and salary workers were union members (less than half the figure at the postwar peak), but the gender gap in unionization rates is lower than at any other time since records have been kept. In 1999, 16.1 percent of male and 11.4 percent of female wage and salary workers in the US were union members, compared to 24.7 and 14.6 percent respectively in 1983 (US Bureau of Labor Statistics 2000). In part as a result of its increasingly female membership, in recent years the US labor movement has embraced issues like pay equity and parental leave, and more generally has become a leading advocate of gender equality. Yet this development has had limited impact, because the overall effectiveness of the labor movement has declined so precipitously. An historically informed approach suggests the likelihood of labor movement resurgence at some future date, however, in which case it seems likely that such movements will help advance the struggle for gender equality to a greater degree than in the past.

See also: Feminist Economics; Feminist Movements; Gender and Feminist Studies in Economics; Gender

and Feminist Studies in Sociology; Gender, Economics of; Gender Ideology: Cross-cultural Aspects; Labor History; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements, History of; Labor Unions; Rural Industrialization in Developing Nations and Gender; Social Class and Gender; Social Movements and Gender; Trade Unions, Economic Behavior of; Trade Unions: Empirical Analyses; Work, Sociology of

Bibliography

Baron A (ed.) 1991 *Work Engendered: Toward a New History of American Labor*. Cornell University Press, Ithaca, NY

Buhle M J 1990 Gender and labor history. In: Moody J C, Kessler-Harris A (eds.) *Perspectives on American Labor History: The Problems of Synthesis*. Northern Illinois University Press, DeKalb, IL

Cobble D S 1991 *Dishing it Out: Waitresses and Their Unions in the Twentieth Century*. University of Illinois Press, Urbana, IL

Cook A H, Lorwin V R, Daniels A K (eds.) 1984 *Women and Trade Unions in Eleven Industrialized Countries*. Temple University Press, Philadelphia, PA

Frank D 1993 *Purchasing Power: Consumer Organizing, Gender, and the Seattle Labor Movement, 1919–1929*. Cambridge University Press, New York

Hartmann H 1976 Capitalism, patriarchy and job segregation by sex. In: Blaxall M, Reagan B (eds.) *Women and the Workplace: The Implications of Occupational Segregation*. University of Chicago Press, Chicago

Kaplan T 1982 Female consciousness and collective action: The case of Barcelona, 1910–18. *Signs: Journal of Women in Culture and Society* **5**: 545–66

Kessler-Harris A 1975 Where are the organized women workers? *Feminist Studies* **3**: 92–110

Milkman R 1987 *Gender at Work: The Dynamics of Job Segregation by Sex during World War II*. University of Illinois Press, Urbana, IL

Milkman R 1990 Gender and trade unionism in historical perspective. In: Tilly L A, Gurin P (eds.) *Women, Politics, and Change*. Russell Sage Foundation, New York

Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York

US Bureau of Labor Statistics 2000 *Employment and Earnings* **47**(1)

Westwood S 1984 *All Day, Every Day: Factory and Family in the Making of Women's Lives*. University of Illinois Press, Urbana, IL

R. Milkman

Labor Movements, History of

'Labor movement' is a comprehensive concept of workers' organised efforts, in evolutionary or revolutionary ways, to change capitalistic societies in order to improve the economic, social, political, and cultural situation of working class people. Although based on the working class, labor movements include all kinds of socialist and some non-socialist organizations; mutual aid societies, trade unions, political parties, cultural organizations, consumers' cooperatives, and certain women's and youth organizations are the basic organizations of the labor movement.

Being a world-wide movement today, the labor movement has its roots in the first industrial revolution in Great Britain in the eighteenth century, when workers protested against developing industrial capitalism by destroying capitalists' property (Luddism). In 1812, the death penalty was introduced in Great Britain for machine breaking, and some Luddites were executed. Despite this, Luddites were active in Great Britain in the 1830s. On the European continent, machine breaking continued until the early twentieth century, even among agricultural workers. Its value as a weapon for workers in the class struggle was limited, but it was an important part of the formation of workers as a class in opposition to capitalists and the bourgeoisie.

Skilled industrial workers in Great Britain, the foremost being spinners and weavers, developed a more successful strategy in the late 1700s by starting trade unions. Unions seemed to the establishment to be a great menace, especially when they were considered to be 'contaminated' by the ideas of the French revolution. In 1815, British workers were denied the right to organize trade unions, and the class struggles of this period culminated with the Peterloo massacre in 1819. But radical political ideas were defeated all over the world, and trade unions were once again tolerated in 1824. On the European continent French and Silesian weavers also went on strike for better conditions. All over Europe and the US, workers went on strike, sometimes spontaneously, sometimes well organized; sometimes workers could achieve a wage rise or avoid a wage decrease; sometimes they could achieve or defend their right to unionize, sometimes they failed. Sometimes, striking workers not only had to fight capitalists and the police but also strike breakers, so-called 'black legs.' On some occasions skilled male workers tried to prevent capitalists from hiring female workers.

The political dimension of workers' early struggles was most obviously indicated in the French revolution in 1789, in which working-class women took part as well, but also in the European revolutions of 1830 and 1848. At the same time, socialist utopians like Henri de Saint-Simon, Charles Fourier, Robert Owen and Joseph Proudhon tried to formulate ideas of alternative societies. Even if the ideas of the Utopians were not transformed into political programs, their importance for future socialist—and feminist—thoughts should not be underestimated. The Working Men's Association in England (1835) established a comprehensive political program as a reaction to the

disappointments with the reform movement of 1830–1832. In the People's Charter (1838) secret elections, equal suffrage for men, and social reforms were demanded. A new Factory Act was introduced in 1850, which limited the industrial work time to ten hours a day and to six days a week, and the Corn Laws (grain duty) were repealed (1846). As a political program the People's Charter was succeeded in 1848 by the Communist Manifesto, which had much more far-reaching demands and had a more international perspective than the People's Charter. It argued for a take-over of the means of production by workers, and the slogan 'Proletarians of all countries, unite!' was to become a symbol for socialists for many years to come. Also, the consumers' co-operative movement, founded in Rochdale in 1844, was of more lasting importance for the working class than the chartist movement.

Being disarmed and defeated in the European revolutions of 1848, the position of the working class and the labour movements was weakened. Political activists on the continent, of which many were intellectuals, had to flee to Great Britain and the US. London became a center for European radicalism. Not much was achieved until 1863, when workers and intellectuals from some Western European countries met in London to organize a trade union and a political co-operation between workers of 'civilized' countries and to prevent the recruitment of strike breakers from the continent to England. The International Working Men's Association—the First International (FI)—was established in 1864; its most prominent members were Karl Marx, Friedrich Engels, and Michael Bakunin. At its peak the FI had 25,000 members, but differences in opinions between trade unionists and politicians about the task of the organization weakened it, as did disputes between socialists and anarchists about ideology and strategies. After the crushing of the Paris Commune in 1871 the FI faded and was finally dissolved in 1876.

Before its dissolution, the FI encouraged workers of industrialized countries to intensify trade union activities and to start national socialist parties. Skilled male workers thus organized trade unions in most European countries and in the US in the 1870s and 1880s, keeping unskilled workers as well as women out of the unions. In Great Britain these unions co-operated within the Trades Unions Congress (TUC) (1868) in the US within the American Federation of Labor, AFL (1886) and in Germany in the 'General-Kommission' (1830) (after several forerunners). Semi-skilled and unskilled male and female workers were organized or organized themselves later and usually to a lesser degree, and in some countries female workers organized women's trade unions; in Denmark the Women's Trade Union remains in 2000 as one of the biggest trade unions in the country. In several cases traveling funds (which supported individual workers in finding work elsewhere) were the origins of trade unions for skilled workers, and many workers unionized in connection with a strike. In the beginning strikes were often started spontaneously but gradually were more carefully planned.

In Great Britain the 'new unionism' challenged the older unions of skilled workers and developed a more socialistic ideology, based on class identity rather than craft, that changed the British labour movement in the 1890s. In the US unskilled workers, in 1905, organized 'One Big Union', Industrial Workers of the World. Like the Confédération Générale du Travail in France (1902), it was based on syndicalistic ideas. In Germany and the Scandinavian countries, national trade unions began to co-operate within national federations in the 1890s. At the turn of the century, 20–22 percent of all workers in Denmark and the Netherlands were organized in local and national trade unions, 10–12 percent in Great Britain, Germany, and Sweden, 6–7 percent in France, Norway, and the US, 2 percent in Australia and Italy. Not all trade unions were inspired by socialist ideas of class struggle and anti-capitalism. Several were inspired by liberalism, especially in Great Britain where trade unions had an affiliation with the Whigs. Other unions had a religious, usually catholic, affiliation, especially in Germany, the Netherlands and the Latin countries of Europe. In the US there were ethnic unions among Europeans in the early days, and among Black and Mexicans in more recent times. All unions demanded the workers' right to unionize, collective bargaining, minimum wages, limitations of the workday, and prohibition of child labour. They differed in their attitudes to strategies and political affiliations.

As was realized by the British trade unionists of the FI, trade unions should act internationally to be successful. In the late 1860s, the Amalgamated Society of Engineers had sections in the US, France and some of the British colonies. Glovers in Austria, Germany, and the Scandinavian countries had an agreement of solidarity in the early 1870s. Between 1889 and 1911, 28 International Trade Secretariats were started for international co-operation on the basis of crafts (typographers, printers, hatters, cigar makers, tobacco workers, shoe makers, glass workers, and so on). Most of the secretariats were based in Germany, a new center of industrial capitalism with many foreign labor migrants. They transferred a lot of money from the national unions to striking and/or locked-out colleagues, and tried to stop the use of blacklegs. They also tried to handle the questions of international labor migration and of Franco-German antagonism.

The international Federation of Transport Workers (1896) is the oldest international trade union remaining in existence in 2000. It started on the basis of international co-operation before national unions existed. Beginning in Copenhagen in 1901, trade union leaders from different European countries met at conferences to discuss labor issues like work time, piece work, women's work, and international labor

migration. In 1902, the International Secretariat of National Trade Union Centers was founded, having its headquarter in Berlin. From 1909, the AFL also took part in that international co-operation, and in 1912 the organization was reorganized as the International Federation of Trade Unions (IFTU).

Workers and socialist intellectuals also organized for a political struggle. While the FI had its headquarters in London, socialist parties were started in Germany in the 1860s by Ferdinand Lassalle, August Bebel, and Wilhelm Liebknecht. At a conference in Gotha in 1875, the two parties joined and formed the Sozialistische Arbeiterpartei Deutschlands (SAPD). Thereby the first nation-wide socialist political party of the world was founded, accepting a class-oriented program. The SAPD challenged the establishment in elections for Parliament, getting 9 percent of the votes in 1877. The establishment was frightened, and it tried to stop socialism by anti-socialist laws. Meanwhile, between 1874 and 1903, socialist parties were founded in all European countries; socialist parties were also founded in the US, Australia, and New Zealand. In Germany, after the abolition of the anti-socialist laws and the re-establishment of the party as Sozialdemokratische Partei Deutschlands (SPD) (1890), the socialist labor movement became a political force of increasing importance. Every large German town had important trade unions, a socialist newspaper, consumers' co-operatives, proletarian cultural associations, but Hamburg was 'No City of Socialism.' In 1914, the SPD was the largest political group in the German Parliament and the establishment was depending on it for its war policy. Trade unionists and socialists in other countries were inspired by the German development, and other socialist parties grew in importance, so much so that the parliamentarian way to socialism seemed realistic to many socialists in Western and Central Europe. This choice of strategy has been one of the most debated issues of labor history in Europe.

Britain was somewhat of an exception, but after the breakthrough of the new unionism the trade unions' connection to the Whigs was questioned. The Independent Labour Party was founded in 1893, but the Committee for National Trade Unions (1895) and the Labour Representation Committee (1900) paved the way for the Labour Party (1906). In the US, the IWW supported the Socialist Party (1901), but both organizations were pressed back by the sharp anti-socialist campaigns during World War I.

Ideologically, there were differing views regarding organizational forms among fractions and individuals within the socialist labor movements before World War I. The trade unionist strategy said that trade unions should organize political parties and control them; the British Labour Party was subsequently built based on that strategy. Many social democrats (socialists) held the opinion that trade unions should act to defend workers' rights on workplaces, while socialist parties should take part in political elections, and control the state with the help of a majority of parliamentarians. Trade unions should have great influence on the party, but not control it. Inspired by Karl Kautsky, several social democrats held a deterministic view of the development of socialism. The German and the Scandinavian labor movements were most outspoken on that idea. Leninist strategy said that the party was the avant-grade of the working class, that should bring workers' consciousness beyond trade union consciousness into class-consciousness. The idea that the party was superior to the unions in the class struggle was formulated among the Bolsheviks within the Russian social democratic party. It became the lodestar among Communists, while anarcho-syndicalists denied the value of parliamentarian activities. They argued for spontaneous and direct action, like the occupation of mills and land, and it had its stronghold in France, Spain, and Italy.

Since the dissolution of the FI in 1876, socialists from several countries had met for conferences on international labor issues. In 1889, on the anniversary of the storming of the Bastille, they met in Paris and decided to reorganize an international socialist labor movement. Even if Latin American and Asian delegates took part in the Paris conference, the Second International (SI) became a white, male, Western European dominated organization. Its main demands were the introduction of equal and universal suffrage, workers' right to unionize, and an eight hour working day, but its ultimate objective was the overthrow of capitalism. At the Paris conference, it was decided to organize May Day rallies all over the world in the memory of the Haymarket victims (Chicago 1886) and in support of the demand for an eight hour working day. Every year since 1890, the international labor movement, where and when possible, has made political demands on employers and on the state throughout the world. The hymn 'The International' has been sung by millions of demonstrating workers at the May Day rallies and at other kinds of meeting. Before World War I the manifestations were impressive, but since then they have been of fluctuating importance in the Western European democracies, while they turned into state managed events in Communist countries. In Nazi Germany, the May Day manifestations were not prohibited, but taken over by the Nazis, and in Japan the company-based trade unions have held rallies in support of 'their' company.

Saying that workers had nothing to gain from patriotism, the SI argued for proletarian internationalism and anti-militarism, and criticized imperialism. Those ideas were more or less dashed in 1914, when most German socialists supported the emperor's plea for war credits. Colleagues in the other belligerent countries took the same position, supporting the war policy. The SI also was split by different opinions on international labor migration, which complicates the idea of proletarian internationalism. The Second

International was, indeed, mostly a debating rather than a decision-making organization. World War I put an end even to that. With the outbreak of the war, also the Golden Age of Reformism ended. In the belligerent countries trade unions partly acquired a new role by mobilizing for war efforts. In the international organizations the TUC and AFL took over the dominating role that German trade unions had before the war.

World War I had vital importance for the socialist labor movement. The left wing opposed ministerial socialism, i.e., socialist members of a government dominated by liberals, as well as the party truce that most socialists accepted during the war. In 1915, representatives for the socialist opposition, mostly young people, met in Zimmerwald to protest against the war. In 1917, facing tremendous class cleavages, strikes and hunger riots, the socialist labor movement split in most countries into two parties. The Bolshevik Revolution in Russia and the failed attempts to raise revolutions in Germany, Finland, and Hungary confirmed the split, and it was made permanent in 1919, when the left-wing parties joined under Russian Bolshevik rule in the Communist International (Comintern), or the Third International. Several groups left the Comintern in 1921, after the introduction of Lenin's 21 theses and the transformation of the left-wing socialist parties into sections of the Comintern. In 1919, the social democrats (socialists) tried to reorganize the Second International, while some noncommunist left-wing parties (led by the Austrian and the French socialist parties) joined in the International Working Union of Socialist Parties (the Vienna International). In 1923 the Second International and the Vienna International merged into the Labour and Socialist International, including most of the socialist and social democratic parties in the world.

After World War I, the international trade union movement also revitalized; by 1920 about 50 percent of all workers were unionized in Austria, Belgium, Denmark, France, Germany, Great Britain, and Italy. Women workers and white-collar workers also organized gradually. The IFTU was reorganized in 1919. In 1924, it started a sub-committee on women's issues, the International Committee of Trade Union Women, but in reality women workers' rights received little attention. White-collar workers unionized, some of them by themselves in national and international trade unions, while others unionized with blue-collar workers of the same industry.

After the war, the trade union movement radicalized, for instance by claiming industrial democracy and socialization of the means of production. This radicalization made the AFL leave the IFTU. On the other hand, communist-dominated trade unions assembled into the Red International of Trade Unions in 1921, an organization that inspired strikes in noncommunist countries, and prevented strikes in the Soviet Union, where they instead became a part of the state. Syndicalistic workers in 1922 joined in the International Working Men's Association. Religious trade unions organized internationally after the war in the International Federation of Christian Trade Unions, later the World Confederation of Labor. Thus, the international labor movement was politically very split, a situation that was even more distinct after the new policy in the Soviet Union in 1928, when the communists proclaimed social democrats being 'social fascists' and their main enemies. This split was of great importance for the coming to power of fascists in Germany in 1933 and in Austria, where the Austromarxists' social experiment 'The Red Vienna' was put to an end in 1934. The whole spectrum of a split labor movement was demonstrated in the Spanish Civil War. In the US, on the other hand, semi-skilled and unskilled workers started to unionize in the Congress of Industrial Organisation (1935). US trade unions also raised some successful strikes, and in the Upper Midwest the Farmer Labor Party had some, temporary, successes in the 1930s.

The coming to power of fascism in central Europe caused two different strategies within the European labor movement. For a short time, Popular Front government in France and Spain effected important social reforms for workers, but were defeated by conservatives and fascists. In Scandinavia, the social democratic parties chose class co-operation as a new strategy, politically with farmers' parties, and with respect to industrial relations with employers (the Scandinavian model). Important social reforms were achieved in the late 1930s and after World War II. The Scandinavian countries also developed into the most unionized countries of the world, where women and white-collar workers also are highly unionized (80–85 percent at the turn of the millenium).

On the European continent, with the rise of fascism and the outbreak of World War II, the labor movements collapsed. The relations between socialists and communists sharpened even more after the Molotov–Ribbentrop pact, but towards the end of the war socialist–communist relations improved and it looked as though a united labor movement could come out of the anti fascist struggle. The postwar program of the British Labour Party attracted socialists in most industrialized, noncommunist countries. These programs were supported in elections in several countries. Important reform programs were introduced and some socialization of the means of production was done. A unified international trade union movement was reorganized in 1945, the World Federation of Trade Unions (WFTU); of important trade unions only the AFL did not join it. Capitalist and bourgeois societies seemed to be threatened in Europe, and extensive strikes challenged the social order of the US, but the Cold War put an end to that. The Red Army forced social democrats and communists in East Germany to merge into the Sozialistische Einheitspartei Deutschlands (SED). In northern, western,

and southern Europe there was once again a total split between social democrats/socialists and communists. The Soviet Union and the communist parties in the Soviet Union controlled parts of Europe, and in France and Italy founded the Cominform in 1947. In the US the Taft–Hartley law (1947) weakened the trade unions. The socialist trade unions left the WFTU in 1949 and started the anti-communist International Confederation of Free Trade Unions. The ICFTU also included the AFL, which was weakened by the Taft–Hartley law. In 1951, socialist/social-democratic parties founded the Socialist International, based on the principles of 'democratic socialism.' In the 1970s the SI added to its Euro-centrist policy more worldwide activities to counteract communist influence in the Third World, where the labor movements were mostly based amongst landless peasants.

In the Third World, the labor movement has faced enormous difficulties. Probably, activists were treated more brutally there in the twentieth century than their colleagues in Western Europe in the nineteenth century. Also, the seasonal and migratory nature of labor, and company-controlled housing counteracted unionization and political mobilization. In British colonies, however, trade unions were of some importance, even in colonial times; trade unions were organized in mining regions, in ports and at railways, and in major urban concentrations, and they took part in the liberation movement even before the 1920s. In India it was of great importance for the liberation in 1947, but since then the trade union movement has been split along both political and religious lines. After the fall of the Soviet Union, trade unions in India and in some Arab counties dominate the WFTU.

After the Soviet invasion of Hungary in 1956, some communists in Western Europe criticized Soviet communism, a criticism that accelerated after the Soviet invasion of Czechoslovakia in 1968. Inspired by the Czech communists, Euro-communism developed in the 1970s as an ideology for an independent, democratic road to socialism. In Western Europe strike activities increased in the 1960s, accelerating for a few years after 1968. In several countries, social democratic governments were the targets for this radicalization of workers and students. The social democratic labor movement was influenced and radicalized, and the social democratic and the reformed communist wings of the labor movement converged. Once again industrial democracy was demanded, but the demands gradually faded away. In some Western European countries there was political co-operation in the final 20 years of the twentieth century between social democrats (socialists) and reformed communist parties, which have cooperated in the defence of the welfare state.

The radicalization in the 1970s also meant the breakthrough of women's questions in the labor movements, partly based on a feminization of the workforce. Many women workers were employed in the public sphere with, in some respects, contradicting interests with industrial workers. Demands for equal pay for equal work were raised, women workers opposed sexual discrimination, and they politically mobilized. Most socialist parties have a minimum quota of women in party organizations and/or female candidates for parliamentary elections. Labor historians have started to focus on gender in the study of labor movements. Another consequence of postwar capitalism has been an ongoing increase of white-collar trade unions, and blue- and white-collar workers have often cooperated in trade union matters. Politically, many white-collar workers have voted for socialist/social democratic parties in Western Europe.

The labor movement, however, still has its most significant influence in Western Europe, where it focuses on questions related to the European Union. European labour movements have, however, given economic and political support to trade unions and political parties in the Third World. These relations have, however, been a little hesitating, as the labor movements in many Third-World countries have been part of the national liberation struggle. Socialist ideas, including Marxist–Leninist ideas, have been guidelines for many liberation movements, but normally they have survived only the first decade of independence. After liberation, many new rulers have tried, as was done in the Soviet Union and Communist-ruled Eastern European states, to subordinate trade unions to state policies. In Africa, however, the Congress of South African Trade Unions (COSATU) was of great and lasting importance for the overthrow of political dictatorship, as was, for instance, Solidarnosc in Poland. Today, the COSATU organize about 40 percent of all workers in South Africa, far beyond what, for instance, the AFL-CIO is capable of (15 percent).

After the fall of the Soviet Union, the communist parties of the former Soviet-controlled regions have been reformed, and have gained quite significant support in democratic elections. Trade unions have also been reformed, or new unions have started. Western colleagues have supported them a little, most of them belonging to the ICFTU.

See also: Bourgeoisie/Middle Classes, History of; Capitalism; Class Consciousness; Class: Social; Communism, History of; First World War, The; Labor History; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Labor Unions; Political Parties, History of; Russian Revolution, The; Social History; Social Movements, History of: General; Social Movements, Sociology of; Socialism: Historical Aspects; Trade Unions, Economic Behavior of; Trade Unions: Empirical Analyses; Welfare State, History of; Working Classes, History of

Bibliography

Carew A et al (eds.) 2000 *The International Confederation of Free Trade Unions*. Lang, Bern, Switzerland
Forman M 1998 *Nationalism and the International Labor Movement: The Idea of The Nation in Socialist and Anarchist Theory*. Pennsylvania State University Press, University Park, PA
Hobsbawm E J 1964 *Labouring Men; Studies in the History of Labour*. Weidenfeld and Nicholson, London
Hyman R, Price R (eds.) 1983 *The New Working Class? White Collar Workers and their Organisations*. Macmillan, London
Padgett S, Paterson W E 1991 *A History of Social Democracy in Postwar Europe*. Longman, London
Sassoon D 1996 *One Hundred Years of Socialism*. Tauris, London
Thompson E P 1963 *The Making of the English Working Class*. Pantheon Books, New York
Upham M 1996 *Trade Unions of the World*, 4th edn. Cartermill, London

L. Olsson

Labor Organizations: International

> You shall not abuse a needy and destitute laborer, whether a fellow countryman or a stranger in one of the communities of your land. You must pay him his wages on the same day, before the sun sets, for he is needy and urgently depends on it; else he will cry out to the Lord against you, and you will incur guilt. *Deuteronomy*, ch. 24, v. 14, 15.

Rapid globalization succeeded the Cold War as the axis around which international labor organizations worked at the beginning of the twenty-first century. The International Labor Organization (ILO), under vigorous new leadership, sought a place at the table where the intellectual and policy architecture of the global economy is determined. The International Confederation of Trade Unions (ICFTU) and the international trade secretariats (ITSs) strengthened ties with national labor organizations to counter the power of multinational corporations that reshape the international division of labor, to influence the international institutions that provide credit to national governments, and to share in making the rules that govern trade and international investment. These bodies historically have been subordinate to the national institutions they represent; globalization may change this balance.

1. *Historical Background*

The nineteenth-century industrial revolution evoked a strong response as exploitation and social injustice grew, first in the UK and then in Europe and North America. The communitarian aspirations of Robert Owen and the reformist fervor of the Chartists was followed by more specifically class-based organizing in the 1840s and 1850s, culminating in the First International, the International Workingmen's Association (IWMA), founded in London in 1864, in which Karl Marx played a leading role. As strikes grew in number and fervor, efforts to prevent employers from importing 'blackleg' workers to break strikes strengthened international connections in Europe and the UK.

Local unions began to grow in numbers and geographic scope, reflecting the expansion of markets. Closely linked at first to socialist organizations, they developed more pragmatic activities in response to problems in the workplace. In most of Europe, the unions were created by, and were close to, socialist or social democratic movements; in the UK, by contrast, the Trades Union Congress preceded and created its political arm, the Parliamentary Labour Party.

The Chicago Haymarket affair and the prosecution of anarchists accused of the bombing strengthened efforts to build international solidarity. Major influence in the Second International, largely stimulated by Haymarket, went to socialist and social democratic parties in Germany, Austria, and Sweden, all of which sponsored newspapers, housing, theater, music, and other activities for the working class and the public. It fell victim to World War I (Windmuller 1979, pp. 3–4).

2. *The International Trade Secretariats (ITSs)*

These institutions link unions representing workers in cognate skills or sectors. The first was the Federation of Boot and Shoe Operatives, founded in1889 in Paris at the time of the first congress of the new Socialist International. By 1914, there were almost 30; they included the International Transport Federation, founded in London in 1896 by the London dockworkers, the first ITS to include semi-skilled and unskilled workers. In later years, mergers and closures reduced this number to 14 at the beginning of the twenty-first century. Except for the US mineworkers union, these were almost completely European bodies before World War I (Windmuller 1979, pp. 3–4).

Efforts to build broader international solidarity among national labor bodies separate from the socialist international began with a 1901 conference in Copenhagen organized by Scandinavian, Belgian, German, and British union leaders; the next year in Stuttgart they founded what later became the International Federation of Trade Unions (IFTU) (Van Hooten and Van Der Linden 1988, p. 11). It too fell victim to World War I, which largely ended visions of international solidarity; as Eric Hobsbawm wrote (in Van Hooten and Van Der Linden 1988, p. 11):

> For socialist leaders the outbreak of war and the collapse of their international in August 1914 were utterly traumatic … The same Welsh miners who both followed rev-

olutionary syndicalist leaders and also poured into the army as volunteers in 1914 brought their coalfield out in a solid strike in 1915, deaf to the accusation that they were being unpatriotic in doing so.

2.1 International Labor Legislation

The International Association for Labor Legislation was proposed in 1897 and formed in 1901, a 'direct forerunner of the I.L.O.' (Morse 1969, p. 7). Like its counterparts, the International Association on Unemployment, and the International Association for Social Progress, it developed model legislation designed to help governments create effective programs. Here too, World War I ended this work, though some labor unions continued to meet during the war years.

3. Between the Wars

By 1917 the American Federation of Labor (AFL), under Samuel Gompers' leadership, had established itself as the dominant national labor organization. It consisted almost completely of skill-based craft unions in a loose confederation. Efforts to organize the unskilled and semi-skilled by the Industrial Workers of the World (IWW) had been crushed by employers, unchallenged or actively supported by government at all levels. The AFL had initially resisted involvement in the war, but by 1917 it was actively involved. This brought Gompers closer to the national centers of power. In 1917 he arranged for President Woodrow Wilson to address the AFL annual convention, the first US president to do so (Larsen 1975, pp. 77–106).

In September 1918, Gompers attended the Interallied Labor Conference in London; he played a major role and was successful in sponsoring a motion in favor of full surrender by the Central Powers, against the views of some European unionists. Gompers was elected chair of the international labor commission organized to recommend measures to deal with labor questions in the peace treaty.

3.1 The AFL and Versailles

A socialist in his early years Gompers later argued that trade unions should focus on the needs of their members and eschew any broader ideological views; he once said, 'I have no quarrel with Socialists, but I have no use for their proposals.' He insisted that the commission he chaired meet in Paris, to be close to the treaty discussions, and as he said, it 'should not be held any place except Paris if we wanted to protect our deliberations from a Bolshevik stampede (Gompers 1957, p. 302).

The conference produced the basic design and purpose of the ILO, and a clause on labor rights that was included in the Versailles treaty. Efforts to revive the IFTU were less successful; the new Soviet unions wanted their own structure, as did union bodies with Catholic or Protestant denominational identities. Gompers withdrew the AFL because he saw no way to coexist with the Soviets, and because he insisted on complete autonomy for each national union body. He never wavered from the view that unions' only legitimate concern was the well-being of workers through collective bargaining.

4. The ILO Between the Wars

British delegates, impressed by the achievements of a tripartite structure of cooperation among business, labor and government in the war effort, took the initiative in incorporating tripartism into the governance of the ILO. Gompers chaired the 1919 founding conference in Washington. The Senate's refusal to ratify the Versailles treaty made it impossible, or impolitic, to propose US membership; Republican presidents in the 1920s shunned both the League of Nations and its agency, the ILO.

4.1 Early ILO History

Even without US membership, the ILO was able to move quickly and effectively. It adopted 16 conventions and 18 recommendations in its early years, dealing with the 8-hour day and the 48-hour week, protection of women and young people at work, maternity protection, and protection of maritime and agricultural workers. (Conventions are debated in two successive annual conferences before they can be adopted. When ratified by states they have the force of international law. Recommendations carry less weight.) By 1939 67 conventions and 66 recommendations had been adopted, and the tripartite governance system was firmly in place. The International Court of Justice affirmed the ILO's jurisdiction in cases brought by national governments. The ILO's program of research, utilizing experts and investigating in the field, had built a reputation for quality. As governments began to experience the potential effects of conventions, the pace of adoption slowed, testimony to the ILO's effectiveness, Membership expanded beyond the core of advanced economies to include nations in Asia, Africa, and Latin America (Morse 1969, pp. 4–23, Ghebali 1989, pp. 9–14).

4.2 US Affiliation

Reversing earlier policies, the AFL supported US membership at its conventions in 1932 and 1933.

Secretary of Labor Frances Perkins, took the lead in persuading President Roosevelt that the USA should join, and in lobbying both houses of Congress, arguing that ILO expertise could help in the design of social welfare and labor legislation, as indeed it did (Ostrower 1975, Martin 1976, pp. 427–30). The Soviet Union also affiliated in 1934, but ceased to send delegates to the annual conference after 1937, while the USA continued to participate; John G. Winant served as Assistant Director from 1935 to 1938, and as the first American Director General between 1939 and 1941.

5. *World War II, The Philadelphia Declaration and the Cold War*

Exiled to Canada during the war, the ILO convened in Philadelphia in 1994, its twenty-fifth anniversary. President Roosevelt sent a warm message, and held a reception after the conference ended. The 'Philadelphia Declaration' charted an ambitious course for the ILO in the postwar era. It revised some of the constitutional language, reaffirmed its basic principles, and defined the priorities of social justice and 'a just share of the fruits of economic progress' as central to the avoidance of the return of depression and war. These included full employment, the right of workers to collective bargaining, social security protection, access to medical care, protection of women and children, adequate housing and nutrition, and safe and healthy working conditions. Its human rights emphasis anticipated the language of the UN Declaration of Human Rights in 1948 (ILO 1944). It emphasized, as ILO leaders continue to do, that the ILO is the only international institution that provides a franchise and governance role to workers and employers.

5.1 *Post-War Growth*

Between 1949 and 1968, the number of member nations increased from 55 to 118, and the budget grew fivefold; the ILO moved much closer to universal membership. It developed a relationship with the UN Development Program for cooperative work in developing economies. Its World Employment Program adopted in 1968 focused greater attention on the problems of unemployment and underemployment. Its attack on apartheid led to the withdrawal of South Africa in 1966. Its technical and professional reputation grew; it built a reputation as 'the world's centre of excellence for research and understanding of labour-related issues' (Langille 1999). David Morse, who served as Director General from 1948 to 1970, raised the visibility and influence of the ILO in the USA.

6. *The Cold War and the Crisis of International Labor*

The USA withdrew from the ILO in 1977, but rejoined in 1980. A crisis of confidence had developed around two issues: the increasing bitter US–Soviet Cold War rivalry, and the international effort to isolate and punish Israel; the ILO was weakened seriously by both.

6.1 *AFL and CIO International Competition*

Large-scale organization by Congress of Industrial Organizations (CIO) unions of masses of unskilled and semi-skilled workers posed a major challenge to the AFL's historic dominance in labor matters. CIO officials played major roles in the economic war effort, and built close relationships with the Roosevelt administration. The CIO opposed the AFL's views in international labor affairs and developed its own relationships, building alliances with French and Italian union bodies that were anathema to the AFL's virulently anti-communist positions.

The AFL's views were close to those of the intelligence and foreign policy agencies in the postwar government, and the two forged close financial and working relationships. The CIO played an active role in the formation of the World Federation of Trade Unions (WFTU), which included Soviet unions and unions close to the Soviets, especially in France and Italy. CIO leaders were not friends of communism; their actions were pragmatic efforts to achieve relative equality in international union matters with the AFL.

The 1947 Marshall Plan proposal completed the rift between the Soviet bloc and most unions of the free world. The CIO supported the Marshall Plan and President Truman's 1948 election campaign. It withdrew from the WFTU; both the AFL and the CIO participated in the founding of the International Confederation of Free Trade Unions (ICFTU) in 1949, and were elected to two of the seven vice presidencies (Radosh 1969, pp. 304–47, Reuther 1976, pp. 329–59, 411–27). The AFL–CIO's Cold War obsessions strained relationships with many ICFTU affiliates and weakened its capacity to respond to changing economic and political challenges; this ended only after new AFL–CIO leaders were elected in 1995.

The effort by the WFTU to control the ITSs raised strong resistance. ITS leaders recognized that the achievement in wartime USA of full employment and rising standards of living for workers provided a viable alternative to the Soviet model that had boasted of its progress during the Great Depression. Their lifelong loyalty to social democratic or socialist principles had earned them the hostility of the communist movement, which they reciprocated, in the 1920s and 1930s. Soviet opposition to the Marshall Plan reinforced

their resistance, and their support of a new international body that respected the independence of these long-established institutions. Even in France, where support of the Soviet unions was strong, old divisions reemerged; 'Force Ouvrière,' one observer remarks, 'is a child of France, not of the United States' (MacShane 1992, p. 277). Konrad Ilg, the longtime secretary of the International Metalworkers Federation (IMF), pointed in 1945 to an 'iron curtain … behind which developments are taking place about which we know hardly anything (MacShane 1992, p. 15). (This iron curtain image preceded by several months the more famous words of Winston Churcll.) MacShane describes the opposition of the ITSs to the WFTU as the 'submerged rock that ripped out one of the sides of the WFTU,' while the Marshall Plan was 'the more visible reef' (MacShane 1992, p. 26).

7. *The ILO in the Post-Philadelphia Era and the Cold War*

Early on, the ILO defined two conventions adopted in 1948 and 1949 as basic: No. 87 (freedom of association and the right to organize), and No. 98 (right to collective bargaining). In later years two others were added: No. 29 (elimination of forced or compulsory labor), and No. 111 (elimination of discrimination in employment and choice of occupation). They have acquired a quasi-constitutional standing; even states that have not ratified them are expected as a condition of good standing to honor them, and are subject to scrutiny if they are charged with failure to do so.

Employment issues were highlighted when in 1969—the year the ILO was awarded the Nobel Peace Prize—it launched a multi-year World Employment Program, one of whose products is the annual World Employment Report.

7.1 *US Withdrawal and Reaffiliation, 1977–80*

Cold War tensions reached a peak with the appointment in 1970 of a Soviet official as Assistant Secretary General, drawing strong US objections. US payments were suspended for a year in protest, then resumed. Admission of the Palestine Liberation Organization (PLO) to observer status, over US objections, led to regular attacks on Israel by Arab and other members; the USA argued that these were political in nature, not germane to the ILO's work, but they continued. US employers objected to the admission of Soviet citizens to the employer group, on the grounds that they were no more than agents of the Soviet state. In 1977, the USA endorsed an ILO study critical of forced labor in the Soviet Union, but it failed to pass—'the US had asked for a vote of confidence, and none was forthcoming' (Galenson 1981, p. 65).

The USA filed a letter in 1975 giving 2-year notice of its intention to withdraw. Withdrawal was damaging both financially and politically; the level of interest and awareness of the ILO, even among labor and legal scholars, remains low. The Carter administration concluded that withdrawal had served its purpose, and that US interests would be better served by rejoining, which occurred in 1980. George Meany's retirement from the AFL–CIO presidency helped.

8. *The ILO, the ICFTU, the Trade Secretariats and the Changing Global Economy*

The world's economy and politics were transformed by the end of the Cold War and the large-scale globalization of economic life. Discussion in the ILO resulted in the adoption in 1998 of the 'ILO Declaration on Fundamental Principles and Rights at Work and its Follow-Up.' This followed the 1995 Copenhagen World Summit for Social Development, whose principal organizer, Juan Somavia of Chile, took office in March 1999 as the first ILO Director General coming not from an advanced economy but from 'the South,' symbolic of increasing divisions in levels of economic development and well-being.

Under the terms of the Declaration, the ILO was committed to producing annual reports, beginning in 2000, under the rubric of 'Your Voice At Work,' focusing on each of the basic rights that form the ILO's bedrock principles:

2000—ensuring the right of all workers to form and join unions, with similar rights to employers;
2001—elimination of all forms of forced or compulsory labor;
2002—elimination of child labor; and
2003—elimination of discrimination in employment and choice of occupation.

8.1 *New ILO Leadership*

In his first year in office the new Director General began to address three major challenges that the ILO faced:

(a) Organizational reform and renewal to ensure that the International Labor Office has identified accurately the major objectives to be addressed, and has the staff and structure needed to address them effectively.

(b) Strengthening relations with international and national labor bodies and other institutions to build more effective cooperation in the shared effort to achieving the right balance between growth and equity, market power and social protection, efficiency and justice for working people.

(c) Participation in the processes by which the global economy defines and effectuates the structure of rules and governance for nations and independent

entities, private and public, that wield power and influence, making the ILO, in the words of the Director General, 'a leading organization of the twenty-first century.'

A comprehensive reorganization of the International Labor Office, around a set of specified goals and priorities, was well under way in 2000. The new structure differs significantly from that which preceded it. (An earlier organizational structure is shown in Ghebali 1989, pp. 164–5.)

An ambitious schedule of conferences, speeches and consultations took the Director General to many parts of the world in his first year. Audiences included the ICFTU world congress in Durban, South Africa; the annual convention of the AFL–CIO; the congress of the European Trade Union Congress (ETUC); the TUC annual conference; the May Day workers' jubilee in Rome; the UN General Assembly; the World Trade Organization (WTO) ministerial conference; the Organization of African Unity; and the follow-up conference to the Copenhagen World Summit. The schedule continued at the same pace in 2000 and beyond.

The ILO was able to broker discussions involving the Secretary General of the UN, international labor leaders, and international business organizations on the concept of a 'global compact' which addresses the central issues of equity and representation that the global economy continues to raise as it restructures economic activity and redefines the international division of labor.

Participation in governance is the most difficult of these objectives. The ILO has consultative status with the International Bank for Reconstruction and Development (IBRD) and the International Monetary Fund (IMF) but relations with the WTO remained at the discussion level in 2000.

9. *International Unions and the International Economy*

Many of the difficulties and challenges that face the ICFTU and the ITSs have long histories. Because economic development occurred at different rates in different countries, inevitably unions were engaged in domestic struggles for recognition, power, and influence. Some national union bodies were the offspring of political parties; others, notably in the USA, developed independently of the parties. Some are focused on national political struggles to protect and advance workers' rights and interests, while others see the workplace and the employer as the focus of the struggle, with the state serving primarily as rule-maker, not decision-maker. In some countries, union density has consistently been high, while in others it fluctuates. Some have a single national body and a centralized structure, while others confront multiple claims and competition for workers' support.

In Europe, the ETUC has been successful in developing cross-national forms of cooperative activity and effective collaboration; a shared history, close geographical links, and decades of involvement in the process of economic integration, deregulation of labor markets, and increased cross-national mobility play important enabling roles. Some of the ITSs have developed new leadership and greater ability to organize cooperative activities of mutual support; International Transport Workers was helpful in the US strike by the International Brotherhood of Teamsters against United Parcel Service in 1997, an important milestone in the effort by US unions to restore balance after the erosion in strength that greatly reduced union density and effectiveness after 1980.

Because even the most globally organized multinational corporations operate in countries with governments, unions, political parties, and social movements they are in principle subject to laws that govern labor affairs, social protection, health and safety, environmental integrity, and other subjects of national legislation and attention by the media and the public. Advocates and consumer organizations have been able to expose patterns of exploitation in both underdeveloped and domestic 'sweatshops.'

10. *Conclusions*

The power to affect or control the direction of economic and social policies will remain the central issue, and will be determined, as the past shows, by the relative balance of forces with different goals and priorities, and the degree of unity or division among them. For international labor, the objective remains an economic order that protects and promotes the range of human rights embodied in the ILO basic conventions, and a structure of governance and rule-making that ensures a voice which represents and advocates these rights.

10.1 *The ILO*

The ILO has built over the decades a rich web of international and national legislation, investigations, and interpretations of its conventions and recommendations. Its role in developing social welfare policies, and standards for the regulation of labor conditions in a wide range of occupations, has helped to build credibility in countries in every area and at different levels of economic, social, and political development. It remains the only international institution where the workers and employers have a voice in decision-making and policy determination. It has survived serious crises that threatened its continuity and viability. (For information on the ILO in the year 2000, see Table 1.)

Table 1
Basic information on the in the year 2000

Member nations: 175		
Total budget: US$481.8 million		
Total staff:	Geneva headquarters	Field
Professional:	492	184
General service:	600	438
Number of conventions adopted*: 183		
Number of recommendations adopted*: 191		
Conventions ratified 1995–2000		
Forced labor (No. 29, 1930): 15		
Freedom of association and protection of the right to organize (No. 87, 1948): 14		
Right to organize and collective bargaining (No. 98, 1948): 20		
Equal remuneration (No.100, 1951): 20		
Abolition of forced labor (No. 105, 1957): 29		
Discrimination in employment (No. 111, 1958): 22		
Minimum age (No. 138,1973): 38		
Worst forms of child labor (No. 182, 1999): 8		

* Conventions are debated in two successive years prior to adoption; when ratified they acquire the status of international law. Recommendations carry less legal weight and are not binding

For its first three decades, the ILO was the sole international voice advocating and defining human rights. In the postwar era others have taken up this role; UN human rights covenants set standards for governments. Codes of conduct are negotiated with individual firms that operate in more than one country. The ILO's decision-making structure is based on representation by nations in a period when strong economic and demographic forces challenge the ability of national governments to deal with them. The ILO's efforts to gain a seat at the rule-making table had achieved only limited results at the beginning of the twenty-first century.

10.2 International Trade Union Bodies

The ICFTU and the ITSs were able to move forward after Cold War tensions abated in the last decade of the twentieth century, but they lacked the financial support and organizational linkages needed to engage effectively the corporate and financial interests that came to dominate the international economy in the late twentieth century. International capital deployed a wide range of strategies that continued to build more concentrated corporate power; labor's inventory is more limited. Use of strikes, boycotts, sharing information, building popular support, use of media, and domestic political work all play a role, but unions remained on the defensive in most areas of the world. Union densities declined and labor-linked political parties shifted toward the right or center. The center of gravity of intellectual work in international economics endorsed the values of markets and competition; progressive views about social welfare, labor markets and industrial relations had lost ground and confidence in the closing decades of the twentieth century.

This had begun to change as the costs and consequences of globalization were examined and doubts raised about the validity of these global paradigms. Large-scale and long-lasting unemployment, increased inequality and poverty, illiteracy, and malnourishment (UNDP 1999) all raise the level of skepticism and the degree of openness to new or dissenting views. Popular movements showed signs of new life, and the academic analysis of global issues found more room for critical and dissenting views. In the future, as in the past, the balance of political forces will determine the outcomes of these struggles.

Bibliography

Boyd R E, Cohen R, Gutkind P C W (eds.) 1987 *International Labour and the Third World: The Making of a New Working Class*. Avebury, Aldershot, UK

Bulletin of Comparative Labour Relations 2000 Multinational enterprises and the social challenge of the XXIst century. The ILO Declaration on Fundamental Principles at Work, Public and Private Codes of Conduct. *Kluwer Law Journal* **37**, The Hague

Compa L, Friedman S F 1996 *Human Rights, Labor Rights and International Trade*. University of Pennsylvania Philadelphia

Galenson W 1981 *The International Labour Organization: An American View*. University of Wisconsin Press, Madison, WI

Ghebali V-Y 1989 *The International Labour Organisation: A Case Study on the Evolution of U.N. Specialized Agencies*. Martinus Nijhoff, Dordrecht, The Netherlands

Gompers S 1957 *Seventy Years of Life and Labor: An Autobiography*, Taft P, Sessions J A (eds.). E. P. Dutton, New York

Gross J A 1999 A human rights perspective on United States labor relations law: A violation of the right of freedom of association. *Employee Rights and Employment Policy Journal* **65**

ILO 1944 Declaration concerning the aims and purpose of the International Labour Organization, Geneva

Langille B 1999 The ILO and the new economy. *International Journal of Comparative Labour Law and Industrial Relations* **5**(3) (Autumn): 254

Larsen S 1975 *Labor and Foreign Policy: Gompers, the AFL, and the First World War, 1914–1918*. Associated University Presses, Canterbury, NJ

Lubin C R, Winslow A 1990 *Social Justice for Women: The International Labor Organization and Women*. Duke University Press, Durham, NC

MacShane D 1992 *International Labour and the Origins of the Cold War*. Clarendon Press, Oxford, UK

Martin G 1976 *Madam Secretary: Frances Perkins*. Houghton–Mifflin, Boston, pp. 427–30

Morse D 1969 *The Origin and Evolution of the ILO and its Role in the World Community*. Cornell University Press, Ithaca, NY

Ostrower G B 1975 The American decision to join the International Labour Organization. *Labour History* **16**(4) (Fall): 495–504

Radosh R 1969 *American Labor and United States Foreign Policy*. Vintage Books, New York

Reuther V 1976 *The Brothers Reuther*. Houghton-Mifflin, Boston, MA, pp. 329–59, 411–27
Taft P 1957 *The AF of L in the Time of Gompers*. Harper & Brothers, New York
Taft P 1959 *The AF of L from the Death of Gompers to the Merger*. Harper & Brothers, New York
Taylor R 1999 *Trade Unions and Transnational Industrial Relations*. International Labor Organization, Geneva, Switzerland
UNDP (United Nations Development Program) 1999 *Human Development Report*. Oxford University Press, New York
Van Hooten F, Van Der Linden M (eds.) 1988 *Internationalism in the Labour Movement 1830–1940*, E. J. Brill, Leiden
Windmuller J P 1995 *International Trade Secretariats: The Industrial Trade Union Internationals*. US Department of Labor, Bureau of International Labour Affairs, Washington, DC

S. M. Rosen

Labor Supply

In economics, the study of 'labor supply' has come to mean the analysis of work behavior where the questions addressed include: How many hours do people work for pay and what factors affect work hours? and What proportion of the population works for pay and what factors affect this proportion? The first question, about the number of hours or weeks that individuals work in the labor market, concerns the intensive margin of work. The second question, whether people work or not, concerns the extensive margin of work. The extensive margin may be measured by the employment–population ratio or by the labor force participation rate. Questions such as: How hard do people work per hour? cannot be examined systematically given the paucity of data on work effort.

The questions above stress work 'for pay.' This excludes voluntary work and unpaid household work. In fact, data are available on both voluntary work (see Freeman 1997) and household work (see Juster and Stafford 1991), and the very same conceptual framework summarized below that is used to study paid work is also employed to understand these other types of work. However, this article concentrates on work for pay and, subsequently, whenever the word 'work' is used, it should be understood that 'work for pay' is meant. Most of the research by economists during recent decades has addressed the supply of labor to the labor market as a whole, not the supply of labor to particular occupations or to particular firms or industries. Research on labor supply to particular occupations or to particular firms uses models similar to the one sketched below to understand the supply of labor to the entire labor market, although there will necessarily be some differences in the analysis. For a review of some research on occupational labor supply, see Freeman (1986). For a study of the labor supply to a single firm, see Nelson (1973). In this article, the research reviewed concerns the supply of labor to the labor market as a whole.

1. *Key Empirical Regularities*

Any satisfactory theory of work behavior should explain the following empirical regularities:

(a) Men work more than women.

(b) In North America and Western Europe, white men work more than black men.

(c) In recent years and in relatively rich economies, those who have been at school longer work more.

(d) Unmarried men work less than married men while, until recently at least, married women work less than unmarried women.

(e) Work hours and employment–population ratios fall in recessions and rise in booms though this procyclical movement is stronger for low-skill than high-skill workers.

(f) Over the past century, in North America and Western Europe, for most types of workers, work hours have fallen and vacations have lengthened. A particularly instructive illustration of this for the USA is provided in Fig. 1, which graphs the usual length of the workday by wage decile. The unbroken line in this graph draws upon information from various surveys in the 1890s while the broken line provides information from a special Current Population Survey in 1991. The length of the workday has fallen for all workers except those at the highest wage decile. Indeed, whereas low-wage workers once worked the longest hours, high-wage workers now tend to do so. Figure 1 describes US women, but the data for US men are similar (see Costa 1999a).

(g) Among men, work rises with age until middle age and then falls.

(h) Although comparisons across countries are frustrated by differences in data collection methods and definitions, it appears that, in the 1990s, average hours worked in the USA exceeded those in Europe.

(i) Average hours worked by self-employed people tend to exceed those by employees (see Aronson 1991). No doubt, instances will be found where these patterns are contradicted, but these empirical findings constitute the norm. A rich literature exists in economics on these issues.

One empirical regularity has received much less attention, the fact that individuals synchronize their work schedules: almost 80 percent of full-time workers in the USA start work between 7 a.m. and 9 a.m. and about the same proportion finish between 3 p.m. and 6 p.m. (Mellor 1986); about 60 percent of all those employed do not work weekends (Smith 1986). Presumably there are strong complementarities in work and leisure activities (i.e. the benefits to one individual from engaging in a given activity are greater if other

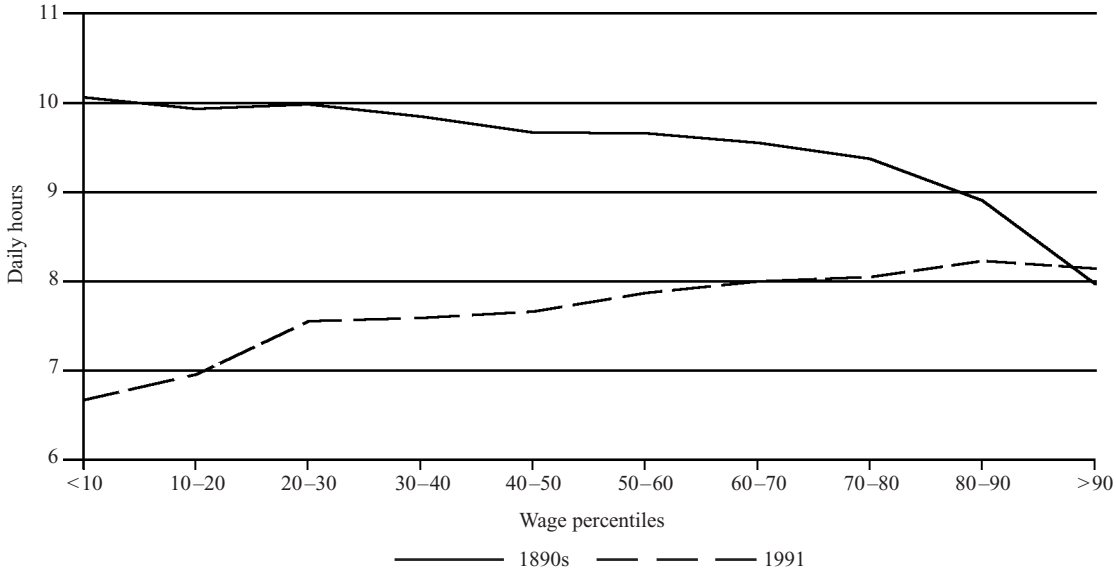

Figure 1
Usual length of work day by hourly wage percentile: women aged 18–64, 1890s and 1991

individuals also engage in that activity at the same time) although little research effort has been directed towards identifying the key complementarities. Weiss (1996) has been one of the few to grapple with this issue.

2. *An Informal Statement*

2.1 *Classifying the Variables*

In essence, economists think in terms of an individual's time spent at work as depending upon four classes of variables: (a) pay per unit time at work; (b) other income; (c) certain demographic characteristics that are observed by researchers; and (d) characteristics of the individual that are unobserved by researchers.

Consider these variables in turn. First, the dependent variable: time spent at work. There are many different dimensions of work time, including annual hours of work, annual weeks worked, weekly hours, daily hours, and whether to work or not.

Now the independent variables. 'Pay per unit time at work' is designed to measure the incentive to work or, equivalently, the price of not working. It is usually measured by an individual's average hourly earnings. Economists usually presume that after-tax earnings matter, not pre-tax earnings. Where average hourly earnings do not vary with time at work, then the use of hourly earnings is appropriate to characterize the incentive to work for pay. However, there are a number of instances where the wage per hour is not fixed but varies (often increases) with work hours so average hourly earnings may provide only a local indicator of work incentives. For instance, in many countries, by statute or by union-negotiated collective bargaining agreements, or by the operation of market forces, wage premiums are paid for long hours worked per week. Or job offers come in tied wage-hours packages: one employer may offer a job specifying a wage of US$10 per hour for 40 hours per week while another may specify a wage of US$12 per hour for 42 hours. In these instances, the incentive to work is not provided by a single wage rate that describes all work opportunities.

The variable 'other income' is sometimes represented by the individual's rent, dividends, and interest, so-called nonlabor income. This type of income for an individual is often the consequence of labor income earned at an earlier age and, therefore, will tend to be higher for those who habitually work long hours. In this way, high non-labor income will tend to be a result of, rather than a cause of, long work hours. Hence, some researchers prefer to use some measure of wealth at a young age (if the data are available to measure it). Or 'other income' is sometimes represented by the labor income of other family members.

It is routine to examine the relationship between time at work and wages holding constant certain demographic characteristics, such as gender and race, characteristics that are typically included in surveys and that are available to researchers. Even allowing for differences in wages and other income, work patterns vary by race and gender, suggesting either systematic differences in 'tastes for work' or different work and nonwork opportunities for these groups.

Finally, it is recognized that much of the variation in work behavior across individuals remains unaccounted for after holding constant all the variables that are usually provided in large data sets. Expressed differently, variables unobserved by researchers 'explain' most of the variation in work across individuals. These unobservables are sometimes called 'tastes for work,' but it should, be understood that this is simply giving a name to our ignorance.

2.2 The Nature of the Available Data

It is important in understanding any empirical labor supply research to recognize the essential source of variation in the data that is being exploited to measure the effects of various variables on work. One source of variation is across people: the correlation between work and other variables (especially wages and other income) is computed across people observed at approximately the same time. This correlation is often calculated separately for men and women and for different racial groups. A different source of variation is over time: here, in essence, the average values of each individual's work and other variables are first computed, then variations over time around each individual's average in work and other variables are calculated to form the relevant correlations.

Although it is straightforward to compute these correlations, they are sometimes difficult to interpret. Consider the effect of wages on work behavior, a subject that has been the focus of most of economists' research on labor supply. An ideal experiment from which to infer the effect of wages on work would be to select people in the population at random, assign wages to them prior to their work choices, and then observe whether and how much they work. In this experiment, the parts played by the random selection of the population and the prior assignment of wages to people are very important. Unfortunately, nature does not provide us with these ideal experimental conditions.

First, the wages that people face at any moment are determined in part by individuals' past decisions about how long to stay at school and what skills to acquire. These past schooling and training decisions may have been made with the expectation of capitalizing on them subsequently by working longer hours in the labor market. People who are less averse to market work are more likely to stay at school longer and to acquire the skills that result in relatively high wages later in life. In this event, the correlation between work hours and wage rates incorporates a causation that runs from preferences about work hours to wages rather than the other way around. Wages are not assigned to people in the manner described by our ideal experiment; in part, wage differences across people reflect people's prior decisions and their past and current preferences.

Second, when computing the correlation between work hours and wage rates, the wage rates facing people who choose not to work for pay are usually unobserved. Expressed differently, information on the wage offers that nonworkers are rejecting is not available to researchers. This problem is typically addressed by computing the hours–wage correlation among those working positive hours only. However, this generates the problem that the sample of people who choose to work does not represent a random sample of the population but a highly nonrandom and self-selected sample. The requirement of our ideal experiment that the correlation between work and wages be computed for people who constitute a random sample of the population is not met. This is one manifestation of what has become known as the problem of 'sample selection' and it has been emphasized in Heckman's influential research (e.g. Heckman 1979). The sample selection problem arises whenever factors unobserved by the researcher affect both (a) the selection of the sample over which the hours–wage correlations are to be computed and (b) the hours of work decisions of the individuals. In this case, the ideal experimental condition—the random selection of the sample—is not being satisfied, and the meaning of the hours–wage correlation is subject to more than one interpretation (as shown succinctly by Heckman 1993).

Therefore, the data available to researchers to compute the effect of wages and other variables on work behavior probably do not satisfy certain ideal experimental conditions and this frustrates measuring the behavioral parameters governing the impact of these variables on work choices.

3. Static Models of Labor Supply

3.1 Conceptual Framework

To model an individual's work decision, let his or her preferences for work hours (h) and consumption (c) be embodied in a utility function $U = U(c, h; X, e)$ where X denotes fixed characteristics of the individual such as gender and race that are observed by researchers. Characteristics affecting preferences that are not observed by researchers are given by e. The individual's income–expenditure constraint is expressed as $p \cdot c = w \cdot h + y$, where p is the price of consumer goods, y is nonlabor income, and w is average hourly earnings.

In the simplest and most tractable of cases, p, w, and y are fixed to the individual and independent of choices of c and h. In this instance, provided U has some convenient technical properties, selecting c and h to maximize U subject to the income–expenditure constraint yields an expression that relates hours worked, h, to the budget constraint variables that are given to the individual, namely, p, w, and y, and to the fixed characteristics conditioning his preferences, X and

e: $h = f(p, w, y; X, e)$. This relationship is the labor supply function and it provides a formal justification for the characterization given in Sect. 3, whereby time spent at work is related to four classes of variables: w, y, X, and e. (Because the budget constraint may be written $c = (w/p) \cdot h + (y/p)$, it is evident that there are not three independent price and income variables, namely, p, w, and y, but only two such variables, namely, w/p and y/p.) As usual in economics, for the key empirical propositions, it is not essential that any individual behaves in this easily caricatured, maximizing manner, but that it facilitates the derivation of relationships and provides a convenient benchmark if we assume that the individual behaves as if he were deliberately optimizing. Less terse explanations of this model can be found in many good textbooks such as Ehrenberg and Smith (1997) or Filer et al. (1996).

This framework applies to the individual who chooses to work a positive number of hours ($h > 0$) and to the individual who chooses not to work at all ($h = 0$). In the case of the decision of whether to work, the individual who is currently not working compares the wage per hour he would receive if he worked, the 'offered market wage,' with his reservation wage which is defined as the implicit value of his nonwork time. When his reservation wage exceeds his offered wage, he does not work for pay ($h = 0$); when his reservation wage falls short of his offered wage, he works some positive hours in the market ($h > 0$). Therefore, the higher is the offered wage, the greater the likelihood of an individual working for pay.

The previous statement concerns an individual's work decisions. Many people live in households where joint work decisions have to be made. In these cases, the standard treatment postulates a household utility function that depends upon the household's consumption and the work time of each household member. If incomes and expenditures are pooled, then each household member's work behavior will depend not only on his wage rate but also on the wage rates of other household members.

This household formulation has been the subject of criticism in recent years for failing to specify how the household utility function is derived and for being silent on how resources are allocated within the household (see, for instance, Chiappori 1988). Recent models start with the utility functions of each household member and postulate that the household reaches efficient decisions with respect to its consumption and work behavior. 'Efficiency' here means that, given prices and wages, the work and consumption decisions of the household members have exploited all opportunities to enhance their welfare without cost so that one person's welfare can be improved only by lowering another member's welfare.

One version of these models characterizes household decision making as a bargaining process. Each individual has a level of welfare that can be enjoyed in the absence of cooperation with other members of the household. This is the individual's 'threat point' and is sometimes equated with the welfare the individual would enjoy if he or she left the household. Provided the household remains intact, anything that increases an individual's threat point will increase the individual's share of the benefits from remaining in the household. For instance, if the market wages paid to unmarried women rise, this increases the wife's threat point in bargaining with her husband and, provided the wife remains married, she should enjoy a greater share of resources within the household. This model has been used not only to understand husband–wife decision making and the growing incidence of divorce, but also intergenerational behavior in developing countries when adult children live with their parents. A formal treatment of these issues is found in McElroy (1990) and a more accessible discussion is in Lundberg and Pollak (1996). These models are likely to be a popular topic for future research especially if they are able to add materially to our understanding of household formation and dissolution.

3.2 Empirical Research

A large amount of empirical research has been directed towards the relationship between h and w. This is because changes in income taxes and other sorts of taxes and changes in welfare benefits are tantamount to changes in w (and sometimes in y, too) and the cost and benefits of these tax and welfare policies depend critically on whether and by how much hours worked and employment–population ratios increase or decrease as w is altered.

In labor economics research in the 1970s and 1980s, the most common source of information to examine the relationship between hours worked and average hourly earnings consisted of large cross-sectional data sets, that is, data on individual people observed at approximately the same calendar time. In essence, after controlling for differences among people in many characteristics, the residual correlation was computed between time spent at work and average hourly earnings. In other words, the impact of wages on work was inferred from differences among people in their wages and work after controlling for other variables.

The most common finding in this research was that, among adult men, hours worked and wage rates were mildly negatively correlated while, among adult women, hours worked and wage rates were positively correlated. A summary and evaluation of these studies appear in Killingsworth (1983), Pencavel (1986), and Killingsworth and Heckman (1986). However, the particular results were disturbingly fickle: relatively small changes in the composition of the sample or in estimating methods sometimes had large effects on the magnitude of the work–wage correlations. Mroz (1987) provides an excellent demonstration of this. Recent research has tended to yield somewhat more

consistent and robust results. This is because the more recent research has followed people over time so, in addition to the variation in the data across people, there is also variation in the data for each person over time. When understanding such intertemporal variation, it becomes important to consider life-cycle or dynamic models of labor supply.

4. *Life-cycle Models of Labor Supply*

In life cycle models of the effect of wages on work behavior, it is crucial to ask: What is the source of the variation in observed wages to which individuals are adjusting their labor supply? One source of variation is the life-cycle variation in wages for a given individual, that is, the evolutionary changes in wages as a person ages. Typically, as an individual ages, their wages rise, sometimes at a decreasing rate, until later in life when wages begin to fall. A second source of variation in wages is the difference in wages across people at the same stage in their life cycle. This second source of variation in wages is called the parametric difference in wages across individuals. At any moment, differences in work behavior among people will be a mixture of responses of labor supply to, first, the path of wages as a given individual ages and, second, differences across individuals in wages observed at the same age.

The first of these responses may be observed if the work behavior and wages of a cohort of people born at about the same time are followed as they age. The proportional change in hours worked that follows a proportional change in wages as these individuals age is called the intertemporal substitution elasticity. It is the life-cycle effect of wages on work. Instead of following people born at about the same time, sort these individuals by age group and compute the correlation between work and wages across people within each age group. These people are about the same age, but they are born in different years (as they belong to different cohorts) so that the observations on their work and wages must be drawn from different calendar years. This correlation incorporates the response of labor supply to parametric differences in wages across people.

The first study to lay out these distinctions with clarity and to compute the various labor supply responses was MaCurdy (1981). His estimates of the intertemporal substitution elasticity for adult US men were in the order of 0.23, meaning that a 10 percent growth in wages as an individual ages would induce a 2.3 percent increase in annual hours worked. His estimates were imprecise, but more recent estimates for both men and women derive similar values for the intertemporal substitution elasticity (see Pencavel 1998, 1999).

When individuals are sorted by age and the correlation between work and wages computed across people holding age approximately constant, the subsequent estimate of the impact of wages on work is often (though not invariably) positive: for both US men and women, a representative estimated elasticity of annual hours with respect to wages is of the order of 0.10. (The elasticity of hours with respect to wages is defined as the proportional change in hours induced by a proportional increase in wages.)

Whereas in the research in the 1970s and 1980s it was common to compute negative estimates of labor supply elasticities, in the 1990s the estimates were more frequently positive and appeared less sensitive to modifications in specification and methods. Why should this be? First, there is a better understanding about the need to distinguish conceptually among different possible wage effects on work behavior. In particular, researchers differentiate the incentives for an individual to work at ages when his wages are relatively high (this is the life-cycle effect) from the incentives to work that come from wage differences arising for reasons that are independent of age (this is the effect of parametric differences in wages). Second, the 1990s research had an important time-series dimension, which may be a critical feature given that, over the past 20 years or so, the structure of wages has changed considerably. This generates a variation in wages that was denied to researchers in the 1970s and 1980s who relied exclusively on wage differences among people at a given moment.

Though there remains some disagreement among economists regarding the value of the elasticity of hours worked with respect to wages, it is undeniable that positive estimates of the effect of wages on work would help to explain many features of changes in work behavior over the past 25 years in the USA. Work levels have increased more for women than men, they have increased more for well-educated people than poorly educated people, and they have increased more for married people than for single people. In each of these three comparisons, the wage structure has changed favoring those who work more: the wages of women have risen relative to those of men, the wages of well-educated workers have risen relative to those with less education, and the wages of married people have risen relative to the wages of unmarried people. One simple story compatible with these facts is that changes in technology and other factors have caused changes in the wage structure and—if work–wage relationships are positive—these wage changes have induced relatively more work from those whose wages have risen relative to those whose wages have fallen.

Positive work–wage estimates may also explain some of the empirical regularities listed in Sect. 2. Men work more than women and, other things equal, men are paid more per hour than women. White men work more than black men and, other things equal, white men are paid more than black men. Those with more schooling work more and those with more

schooling have higher average hourly earnings. Unmarried men work less than married men, and unmarried men tend to have lower wages than married men. Unmarried women work more than married women, and unmarried women tend to have higher hourly pay than married women. The work–age patterns for men broadly correspond to their wage–age patterns. No doubt, there are other competing explanations for these regularities, but the naïve positive association between work and wages across these groups is striking.

5. Other Empirical Findings

5.1 Labor Supply and Taxes

One aspect of labor supply research that has occupied a large amount of attention is the impact of income taxation on work behavior. If t is the tax rate, w represents gross (that is, pre-tax) hourly earnings, and y is gross (pre-tax) nonlabor income, then net or post-tax wages are defined as $w^n = (1-t) \cdot w$ and net or post-tax nonlabor income is $y^n = (1-t) \cdot y$. If each individual responds to net wage rates and net nonlabor income, then the relevant budget constraint involves the post-tax variables: $p \cdot c = w^n \cdot h + y^n$.

The problem posed for researchers measuring the effect of income taxes on labor supply is that, in most income tax systems, tax rates change with income (which, in turn, changes with work hours) so that the net wage rate varies as an individual alters his work hours. This means that the incentive to work is not represented by a value for hourly earnings that is fixed to the individual and independent of their work choice. Instead, there is no single value that represents the incentive to work because the net hourly wage at, say, 20 hours of work per week will diverge from the net hourly wage at 40 hours per week. In this instance, the budget constraint, $p \cdot c = w^n \cdot h + y^n$, is no longer linear.

It is not only the explicit income tax system that produces nonlinearities in the budget constraint. Most income-maintenance programs designed to provide support for low-income people also result in nonlinear budget constraints for individuals eligible for such programs. This is because most income-maintenance programs do not transfer a fixed sum that is independent of the recipient's income. On the contrary, most transfer programs are income conditioned: less money is transferred to the recipient as the recipient's income rises, until a point is reached, the so-called break-even point, where transfers are reduced to zero and the individual moves off the welfare program and is faced with just the (positive) income tax system. The rate by which transfer income is reduced as the welfare recipient earns one more dollar is the welfare system's implicit tax rate. Virtually all implicit tax rates are very high, often as high as 100 percent when transfer income is reduced by one dollar for every dollar that the individual earns, so that budget constraints of low-income people are highly nonlinear.

The research on the labor supply effects of income taxes and transfers tries to reproduce the essential features of individuals' after-tax and after-transfer budget constraints. When the tax system can be characterized as generating an after-tax budget constraint for an individual in which after-tax wage rates are continually falling as more hours are worked (that is, a budget constraint that is strictly concave from below as implied by a smoothly progressive tax system), the statistical procedures for estimating the effect of wages and taxes on work behavior are relatively straightforward and tractable. When, however, the after-tax budget constraint has kinks and net wages may rise over some regions as hours increase, the statistical methods become more recondite and the results more sensitive to alternative estimation methods. The variation in the estimates of different studies reported in Blundell and MaCurdy's (1999) survey is striking. The essential source of the difficulty for researchers is that most data sets provide little information about the income tax and transfer programs that households face. Consequently, with the information available on gross wages and income, in effect, researchers try to complete the income tax and transfer algorithms for each individual, but the inaccuracies in this process and the errors introduced must be extensive.

A similar problem exists in trying to determine the impact of welfare or income maintenance programs on labor supply. A popular hypothesis is that some people reduce their work levels and, therefore, their market earnings to become eligible for welfare. This is revealed, it is argued, by the fact that welfare participation (the fraction of households receiving welfare payments) is higher in those states in the USA where the welfare program parameters imply relatively generous benefits. The problem with this argument is that, even in the absence of any work incentive effects of the welfare system, an association between welfare participation and the parameters of the welfare system are to be expected. Where welfare program parameters imply relatively generous welfare benefits, households will accept opportunities to raise their disposable income and participate in the welfare program, even if they do not change their work behavior. The association between welfare participation and welfare program parameters (especially the welfare system's implicit tax rate) has to be specified carefully to identify work incentive effects (see Ashenfelter 1983). Unfortunately, because very few differences across states exist in the statutory implicit tax rate on welfare income, there is little persuasive evidence from welfare programs of work disincentive effects.

Therefore, at present, probably the most reliable estimates of the effect of income taxes and transfers on

Table 1
Joint and unconditional probabilities of employment of white husbands and wives in the USA, 1990

		Probability of wife…		
		…not working	…working	Unconditional probability
Probability of husband…	…not working	0.075	0.059	0.134 (probability of husband not working)
	…working	0.309	0.557	0.866 (probability of husband working)
Unconditional probability		0.384 (probability of wife not working)	0.616 (probability of wife working)	

work behavior are those inferred from studies that compute the impact of changes in gross wages on work. The assumption here is that the effect of a proportional increase in gross wages on work hours equals the negative of the effect of a proportional increase in tax rates on work hours. This implies, given the estimates reported above of the impact of wages on work, that a 10 percent increase in income tax rates will induce individuals belonging to one cohort to work between one and three percent fewer hours at each age than they would otherwise choose to do.

5.2 Family Labor Supply Decisions

The employment probabilities of husbands and wives do not appear to be independent of one another. To show this, consider caucasian households in the USA where both the husband and the wife are aged between 25 and 64 years. (The data are drawn from the 1960 Census of Population.) Table 1 presents the joint and unconditional probabilities of market work of these husbands and wives. Thus, the probability that both the wife and the husband work is 0.557 and the probability that the wife works but the husband does not is 0.059. Hence the sum of these two joint probabilities is the unconditional probability of the wife working which is 0.616 (given in the bottom line of Table 1). Similarly the probability that the wife does not work but the husband does is 0.309 and (as already noted) the probability that both the wife and the husband work is 0.557. Hence, the sum of these two joint probabilities yields the unconditional probability of the husband working which is 0.866 (given in the last column of Table 1).

From these joint and unconditional probabilities can be derived the conditional probabilities and these are listed in Table 2. For instance, the probability that the wife works conditional on the husband working is 0.643, given in the first line of Table 2. Or, the probability that the husband works conditional on the wife also working is 0.904 given in the third line of Table 2. If the husband's and wife's employment probabilities were independent, then the conditional employment probabilities in Table 2 would equal the unconditional probabilities in the final column and the final line of Table 1. In fact, the conditional probabilities of the wife working are 0.643 and 0.441 (given in the first two lines of Table 2) whereas the unconditional probability of the wife working is 0.616 (from Table 1). The conditional probabilities of the husband working are 0.904 and 0.805 (given in the last two lines of Table 2) while the unconditional probability of the husband working is 0.866 (from Table 1). These data suggest that employment independence is closer to being satisfied for husbands than it is for wives. However, for both husbands and wives, the probability that one is employed is greater if his or her spouse is also employed. Consistent with this, working husbands and wives tend to retire at or close to the same time (see Hurd 1990 and Costa 1999b).

There are a number of reasons why employment probabilities of husbands and wives are positively correlated. First, this may reflect something about marital choices: women less averse to market work tend to marry men with similar work preferences. Second, the correlation may be that each party's responses to their shared economic environment. For instance, if both husbands and wives work more in households possessing little nonlabor income, then the positive correlation in work behavior may reflect the common influence of this source of wealth. Third, the correlation may be the outcome of both marriage patterns and economic environments. For example, it has already been reported that those men facing higher wages are more likely to work in the market than lower- wage men and those women facing higher wages are more likely to work in the market than lower-wage women (a). Furthermore, high-wage men tend to marry high-wage women (b). Together, (a) and (b) will generate a positive correlation between the employment probabilities of husbands and wives.

The positive correlation in the market work of husbands and wives may also reflect cross-wage effects. That is, hold the wages of the husband constant: then, do those husbands married to relatively high-wage wives work more than those husbands married to relatively low-wage wives? Conversely, hold the wages

Table 2
Conditional probabilities of employment of white husbands and wives in the USA, 1990

Probability of wife working conditional on husband working	= probability of both husband and wife working divided by the unconditional probability of husband working	= 0.557 ÷ 0.866	= 0.643
Probability of wife working conditional on husband not working	= probability of wife working and husband not working divided by the unconditional probability of husband not working	= 0.059 ÷ 0.134	= 0.441
Probability of husband working conditional on wife working	= probability of husband and wife working conditional on wife working divided by the unconditional probability of wife working	= 0.557 ÷ 0.616	= 0.904
Probability of husband working conditional on wife not working	= probability of husband working and wife not working divided by the unconditional probability of wife working	= 0.309 ÷ 0.384	= 0.805

of the wife constant: do those wives married to relatively high-wage husbands work more than those wives married to relatively low-wage wives? This issue has arisen in the last few decades as a possible reason for the growth in the employment–population ratio of married women. A popular explanation has focused on the slow growth in the wages of married men since the early 1970s: are married women working more because the wages of their husbands have increased so little or, in some instances, have actually decreased? The broad trends in the key variables support such an interpretation because there has been a steady increase in the market work of married women while, for the past 25 years or so, median real wages of married men have increased hardly at all.

Indeed, the work decisions of husbands and wives do seem to be related to the wages of their spouses. In particular, when one spouse's real pay falls, the other spouse tends to work more (see Juhn and Murphy 1997, Pencavel (1998, 1999)). However, these cross-wage effects—the relation between one individual's work behavior and his or her spouse's real earnings–are much less powerful than the own-wage effects (that is, the relation between each individual's work behavior and his or her own real wages). According to these findings, the employment of married women has increased principally because their real offered wages have been increasing and the movements of their husband's earnings have been a factor, but a factor of less importance in accounting for the movements in their work levels.

If women were working more to augment the stagnant pay of their husbands, then the increases in the work of married women should be greatest for those men whose wages have fallen, not for those men whose wages have risen. In fact, the largest increases in work among women have occurred for those women married to high-wage men and the real earnings of high-wage men have tended to increase in real terms since the 1970s. By contrast, the market work of wives of low-wage men has increased much less and yet the real wages of low-wage men have tended to fall. These facts are reported in Juhn and Murphy (1997).

Hence this evidence suggests that husbands and wives tend to substitute their work time for one another: when the wage rate of one party increases (decreases), that person works more (less) and his wife or her husband works less (more). The positive correlation in the employment probabilities of husbands and wives is driven more by marriage patterns and by own-wage effects than by cross-wage effects.

5.3 Work Behavior of Older People

One of the most striking of long-term developments in work behavior is the decline in employment of older men. Of men 65 years and older, 68 percent were in the

labor force in 1890, 42 percent in 1940, 31 percent in 1960, and fewer than 20 percent in the mid-1980s. Interestingly, from the mid-1980s to the mid-1990s, this decline appears to have flattened out. Similar declines in hours of work among those men employed are also observed. By contrast, the labor force participation rate of women aged 65 years and over has changed very little: it was 8 percent in 1890 and a little more than 8 percent 100 years later.

The drop in the amount of work allocated by older men to the market is one definition of retirement. Another definition, often accompanying the first, turns on the receipt of pension or Social Security income. It is tempting to convert this association into a causal relationship: retirement in the sense of the reduction in labor supply among older workers is induced by the opportunity to draw on annuities from accumulated pension funds.

Because the dominant retirement income for the median elderly household is Social Security income, a much researched issue has been the impact of Social Security on retirement. For most of the past 50 years or so, Social Security wealth has been increasing over time while male labor force participation rates have been falling so the aggregate time-series movements in these variables are compatible with a causal role for public pensions. However, monthly income from Social Security was first paid in 1940 to a relatively small fraction of the elderly population at which time labor force participation rates of men 65 years and over had already fallen by more than 20 percent since the beginning of the century. As Costa has noted (1998, pp. 20–21), in the USA and other countries (Britain, Germany, and France), retirement trends do not display a discontinuity around years when, as a consequence of new legislation, public pension benefits start being paid.

However, a role for Social Security affecting retirement behavior is suggested from an examination of the retirement patterns of the members of a given cohort as they age. In particular, since the mid-1960s when early Social Security benefits were first introduced, there has been a spike in male labor force participation rates at age 62 years. It is difficult to understand why there should be such a spike at 62 years other than the discontinuous retirement incentives embodied in the Social Security benefit rules.

In addition, certain private pension programs create incentives to retire before the age of 65 years. Thus, defined-benefit pension plans specify the pension annuity as a fixed proportion of a worker's highest annual earnings and often the actuarial value of such defined-benefit plans is higher for people younger than 65. The fraction of workers covered by defined-benefit pension programs has been falling and that of defined-contribution pension plans increasing. Because pension wealth grows until the funds are withdrawn, defined-contribution plans embody no incentives for early retirement.

The growth in Social Security and in private pensions represent specific manifestations of a more general phenomenon, namely, the increase in wealth that many cohorts have enjoyed this century. This increase in wealth discourages market work at some time during an individual's life and it may be efficient for the typical individual to take this increased leisure not in the form of fewer work hours each year, but in the form of fewer years at work and of earlier retirement.

5.4 *Differences in Work–Wage Elasticities*

In undergraduate textbooks, for many years it has been common to draw graphs of the relationship between hours worked and wages that imply a positive association at relatively low wages and a negative association at relatively high wages. This nonlinear relationship is called a 'backward bending labor supply curve.' Although this curve is a staple of every economics student's education, the empirical support for this relationship is meager. In some instances, researchers have specified the relationships in such a way that a nonlinear and nonmonotonic relationship between wages and work hours was not investigated. In other instances, however, researchers have explored the possibility of this backward-bending curve and found little evidence for it.

With the richer, more recent, data that combine cross-section and time-series variation, the issue has come up for re-examination. It surfaced because the impact of wages on work appeared to be larger for men with relatively low levels of schooling than for well-educated men, and it was a short step to move from this observation to a specification in which the impact of wages on labor supply fell as wages rose. Indeed, according to one set of estimates, the elasticity of annual hours worked with respect to wages is calculated to be of the order of 0.33 at the average wage of men with less than a high school education, and it falls to 0.14 at the average wage of men with more than 16 years of schooling (see Pencavel 1999). Note, however, that although the slope of the hours–wage relationship becomes less positive at higher wages, it never becomes negative, except for a very few people with extremely high wages. A similar result is obtained by Juhn and Murphy (1997).

6. *Conclusion*

Over the past 40 years or so, there has been a very large volume of research on labor supply by economists. This is because the work decisions of individuals represent one of the essential components of labor market analysis (in the classical metaphor, these labor supply decisions represent 'one blade of the scissors,' the other blade being the hiring and employment

decisions of firms) and because questions involving labor supply are closely tied to important issues of the public policy (such as the effect of income taxes on how much individuals work). This research has been increasingly sophisticated in economic and statistical modeling. In addition, much more is known today about work patterns than was known in 1960. Some components of the economist's standard model of labor supply (such as the intertemporal substitution elasticity) have been estimated with confidence. However, there remains some disagreement over the values of other key parameters such as the effect across different individuals of increases in wages on hours worked. Most economists probably believe that, for the typical worker, increases in wages generally induce relatively small increases in work hours although researchers would refrain from confident statements about precise magnitudes.

See also: Economic Geography; Employment and Labor, Regulation of; Labor, Division of; Sex Differences in Pay; Spatial Labor Markets; Unemployment: Structural; Work: Anthropological Aspects; Work, Sociology of; Workplace Environmental Psychology

Bibliography

Aronson R L 1991 *Self-employment: A Labor Market Perspective*. ILR Press, Ithaca, NY

Ashenfelter O 1983 Determining participation in income-tested social programs. *Journal of the American Statistical Association* **78**: 517–25

Blundell R, MaCurdy T 1999 Labor supply: a review of alternative approaches. In: Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*. North-Holland, Amsterdam, Vol. 3a, pp. 1559–1696

Chiappori P-A 1988 Rational household labor supply. *Econometrica* **56**: 63–90

Costa D 1998 *The Evolution of Retirement: An American Economic History, 1880–1990*. University of Chicago Press, Chicago

Costa D 1999a *The Wage and the Length of the Work Day: From the 1890s to 1991*. Unpublished paper, Department of Economics M.I.T. and National Bureau of Economic Research

Costa D 1999b *The Rise of Joint Retirement, 1940–1998*. Unpublished paper

Ehrenberg R G, Smith R S 1997 *Modern Labor Economics: Theory and Public Policy*, 6th edn. Addison-Wesley, Reading, MA

Filer R K, Hamermesh D S, Rees A E 1996 *The Economics of Work and Pay*, 6th edn. HarperCollins, New York

Freeman R B 1986 Demand for education. In: Ashenfelter O, Layard R (eds.) *Handbook of Labor Economics*. North-Holland, Amsterdam, pp. 357–86

Freeman R B 1997 Working for nothing: the supply of volunteer labor. *Journal of Labor Economics* **15**: S140–S166

Heckman J J 1979 Sample selection bias as a specification error. *Econometrica* **47**: 153–62

Heckman J J 1993 What has been learned about labor supply in the past twenty years? *American Economic Review, Papers and Proceedings* **83**: 116–21

Hurd M 1990 The joint retirement decision of husbands and wives. In: Wise D A (ed.) *Issues in the Economics of Aging*. University of Chicago Press, Chicago, pp. 231–54

Juhn C, Murphy K M 1997 Wage inequality and family labor supply. *Journal of Labor Economics* **15**: 72–97

Juster F T, Stafford F 1991 The allocation of time: empirical findings, behavioral models, and problems of measurement. *Journal of Economic Literature* **39**: 471–522

Killingsworth M R 1983 *Labor Supply*. Cambridge University Press, Cambridge, UK

Killingsworth M R, Heckman J J 1986 Female labor supply: a survey. In: Ashenfelter O, Layard R (eds.) *Handbook of Labor Economics*. North-Holland, Amsterdam, pp. 103–204

Lundberg S, Pollak R A 1996 Bargaining and distribution in marriage. *Journal of Economic Perspectives* **10**: 139–58

MaCurdy T E 1981 An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* **89**: 1059–85

McElroy M B 1990 The empirical content of Nash-bargained household behavior. *Journal of Human Resources* **25**: 559–83

Mellor E F 1986 Shift work and flextime: how prevalent are they? *Monthly Labor Review* **109**: 14–21

Mroz T A 1987 The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* **55**: 765–99

Nelson P 1973 The elasticity of labor supply to the individual firm. *Econometrica* **41**: 853–66

Pencavel J 1986 Labor supply of men: a review. In: Ashenfelter O, Layard R (eds.) *Handbook of Labor Economics*. North-Holland, Amsterdam, pp. 3–102

Pencavel J 1998 The market work behavior and wages of women. *Journal of Human Resources* **33**: 771–804

Pencavel J 1999 A cohort analysis of the association between work and wages among men. Unpublished paper, Department of Economics Stanford University

Smith S J 1986 The growing diversity of work schedules. *Monthly Labor Review* **109**: 7–13

Weiss Y 1996 Synchronization of work schedules. *International Economic Review* **37**: 157–79

J. Pencavel

Labor Unions

Labor unions are interest associations of workers in waged employment. They are formed to improve the market situation and the life chances of their members, by representing them in the labor market, at the workplace, and in the polity, and in particular by collectively regulating their members' terms of employment. Unions emerged in the transition to industrial society in the nineteenth century, together with the de-feudalization of work, the rise of free labor markets, and the commodification of labor. While employing modern means of formal organization, they represent an element of traditional collectivism in a market economy and society. Unions have taken a wide variety of forms and adopted different strategies in different historical periods, countries, and sectors. They are therefore favorite subjects of comparative social science.

1. Unions in the Social Sciences

The literature on unions is vast and extends into several disciplines. In the Marxist and socialist tradition, authors in the late nineteenth and early twentieth centuries debated whether and how unionism might contribute to the overthrow of capitalism or, to the contrary, its stabilization. Democratic reformers like Beatrice and Sidney Webb explored the potential contribution of unionism to the governance of a democratic industrial society. Institutional economists, such as Commons and Perlman in the United States or Brentano in Germany, tried to understand the economic causes and consequences of unionization. Neoclassical economists like Böhm-Bawerk explored the capacity of unions to override market prices and achieve a lasting redistribution of resources, while their successors in the 1960s applied theories of monopolistic competition to analyze the impact of collective bargaining on wages (Chamberlain). Early econometric analyses studied the impact of economic fluctuations on union organization and union activities, including strikes (Ashenfelter). The discipline of industrial relations, which grew after World War II in the US (Dunlop) and Britain (Flanders, Clegg), focused on unions as participants in collective bargaining under the 'New Deal' and in the postwar 'mixed economy.' Sociologists took an early interest in the transformation of unions into centralized and bureaucratic mass organizations with a staff of full-time officials (Cassau). Postwar American research on union democracy was in particular interested in the sources of communist influence in unions, while macro-sociological studies have investigated the origins of unionism in the context of social protest movements and their integration in the modern nation-state (Tilly). Political sociology and political science display a lasting interest in the role of unions in the polities of industrialized democracies. Economists study the economic effects of unions, in particular with respect to relative factor prices, inflation, productivity, and employment (Freeman and Medoff, Calmfors and Drifill, Soskice). Similar concerns figured prominently in historical-institutionalist research and theory on neo-corporatism in the political economy of advanced European welfare states (Schmitter, Crouch). Labor law has long been concerned with the rights of unions *vis-à-vis* employers, and with reconciling collective bargaining and collective industrial agreements with civil law and the freedom of trade and contract. Constitutional lawyers, in particular on the European continent, have tried to clarify the status of unions and collective bargaining in relation to the state and public legislation.

2. The Rise of Unionism in Industrial Countries

Unions emerged in conflict with both economic liberalism and political authoritarianism, striving simultaneously for economic regulation and political freedom. As a result, they were originally treated as conspiracies against free trade, the state, or both. But as unions established themselves as effective labor market cartels, they also became suppliers of labor to those employers willing to deal with them, and in this capacity slowly turned into 'managers of industrial discontent' (Flanders). Similarly, while unions represented the interests of a class opposed to the advance of capitalism, their toleration in the course of democratization contributed to the transformation of liberal into organized capitalism, and was central to the institutionalization of compromise between capital and labor. The integration of unionism and working class collectivism in the institutional framework of modern industrial societies turned unions into legitimate intermediary organizations between their members on the one hand and employers and the state on the other, and transformed class conflict into joint regulation of work and employment. It represents the foremost example of the stabilization of the capitalist economy through democratic politics and institution building.

Early unions saw themselves as democratic organizations of self-help and self-government of workers independent from the feudal or bourgeois pre-democratic state. Often they belonged to broader labor movements that included political parties, consumer cooperatives, mutual assistance funds, educational associations, sports clubs, etc. While unions generally resented interference of the state and the law in their internal organization and activities, they differed widely in structure and ideology. Thus, syndicalist and anarcho-syndicalist unions, which in a number of countries remained significant well into the twentieth century, regarded themselves as constituent units of a direct democracy of producers set to replace both capitalist employers and the bureaucratic apparatus of the modern state. These traditions, which equally opposed capitalism and parliamentary democracy and favored direct action over both collective bargaining and political-electoral lobbying, culminated in the militant council movements of World War I and the revolutions following it, especially in the Soviet Union.

Integration of unions in democratic capitalism, and union recognition by governments and employers, was greatly advanced by the two world wars. Economic mobilization and the governance of the war economy required the collaboration of union leaders, who in many countries came to be co-opted into positions of quasi-public authority. Also, enlisted soldiers had to be promised a better life in a fairer society upon their return from the battlefields, and in defeated countries traditional elites were replaced in the aftermath of war by liberal or socialist governments. Generally, the end of World War I brought political democratization and, precipitated by the threat of socialist revolution, widespread acceptance of collective bargaining. But the first postwar settlement proved fragile in countries

like Germany, Japan, Italy, and Spain, where unionism was soon suppressed by authoritarian regimes and replaced with state-controlled mass organizations of workers that were put in charge of administering state social policy. Similarly, in the Soviet Union workers councils were incorporated in a repressive state machinery and unions turned into 'transmission belts' from the state to the working class. In the United States, by comparison, the New Deal extended union organizing rights while the Swedish Social-Democratic government of the 1930s and the British war cabinet of the 1940s began to develop the contours of the labor-inclusive Keynesian welfare state of the second postwar settlement after 1945.

The 'golden age' of capitalism in the second half of the twentieth century saw the worldwide ascendance of a 'mature' type of unionism, centralized at the level of the national state and pursuing economic and social policy goals through collective bargaining and political-electoral lobbying within the confines of capitalism and parliamentary democracy. This development was part of the consolidation of both democratic capitalism and the nation-state in the countries under American hegemony, where legal recognition of unions and free collective bargaining, extensive social welfare provision, a sizeable public sector, and politically guaranteed full employment made possible the coexistence of liberal democracy and the market economy. The normalization of unionism under postwar 'embedded liberalism' (Ruggie) coincided with national regulation and standardization of the employment relationship and of the status of wage earners as distinguished from employers and the self-employed, which in turn corresponded to the advance of industrial mass production. Far-reaching legal and political regulation of the labor market, introduced to insulate employment and employment conditions as much as possible from economic fluctuations, stabilized union power. The worldwide outburst of labor militancy in the late 1960s, which took national union leaders by surprise, further added to union power, as governments, still feeling politically constrained to provide for full employment, more than ever depended on union cooperation for restoring economic and political stability.

Economic, political, and social change in the 1980s and 1990s caused a roll-back of unionism in most countries. In the 1980s, the British government proved that elections could be won even with high levels of unemployment. Under the leadership of the post-New Deal United States, governments also found less politically and economically costly ways to combat inflation than 'social contracts' with union leaders who had often turned out unable to deliver on their promises of wage moderation. Low inflation and high unemployment coincided with deep industrial restructuring, in response to economic internationalization and a new wave of technological change. Gradually this undermined the Fordist organization of work and the normalized employment relationship unions had helped shape and with which they had learned to live comfortably. Also, the breakdown of Communism in the late 1980s eliminated the systemic opposition to capitalism that had made Western governments willing to make concessions to unions after the two world wars. While private sector unionism has almost disappeared in the world's leading economy, the United States, it is, however, still firmly entrenched in many European countries, due to strong institutional supports left over from the social compact of the postwar era.

3. *Union Organization and Union Growth*

The first unions were local protest movements of the 'working poor.' In the mid-nineteenth century, they were superseded by formally organized craft unions of skilled workers which thrived particularly in the liberal political environment of early industrializing Britain and America. Commanding considerable market power, craft unions represented their members mainly through economic action. In addition to maintaining sizeable mutual assistance funds, many of them were cartels that unilaterally fixed the price of their members' services. From the state craft unions did not expect much apart from freedom of organization. While politically liberal, they were socially exclusive. Not only was membership in them reserved to the skilled. To control the labor supply, draft unions also limited apprenticeship training, prevented employers from hiring non-union members for craft jobs ('pre-entry closed shop' for specific 'job territories'), and defended an organization of work that ensured demand for their members' skills (job control unionism).

Where craft unions became firmly established, they survived beyond the second wave of unionization at the end of the century, when the growing masses of unskilled factory workers began to get organized. Given the weak market position of their members, the 'general' unions that undertook to organize the large factories of the beginning age of mass production had to place their hope on the strength of large numbers. They therefore tried to recruit as many members as possible regardless of skill, although where craft unionism was strong, they mostly remained unions of the unskilled. While craft unions preferred a 'voluntarist' mode of operation keeping the state at arm's length, the new general unions emphasized political action and pressed for social legislation and state intervention in the economy.

In Anglo-America where industrialization and democracy had come early, craft and general unions coexisted well into the twentieth century, with the former deeply influencing the practices and ideologies of the latter. By comparison, in late industrializing countries with authoritarian governments and nar-

rowly limited political and economic opportunities for a market pursuit of collective interests, craft unions were mostly absorbed in inclusive 'industrial unions' organized by economic sectors. As these had originated in often-bitter political struggles especially for universal suffrage, their unionism was embedded in universalistic concepts of social and political reform that differed profoundly from craft particularism or American 'business unionism.' Frequently the new unions were allied to political parties, especially socialist and Catholic, and while craft unions had been fragmented by skill, industrial unionism was often fragmented by political orientation. Indeed as authoritarian governments had initially denied them the opportunity to represent their members economically, some of the new unions remained for long subordinate to political parties.

Craft unions, while narrowly based and normally small, were highly centralized where their members faced national labor markets. With the rise of the nation-state industrial unions, representing their members through political action for organizational rights, favorable labor law, and social welfare legislation, also developed nationally centralized structures. In principle, these were usable for centralized collective bargaining as well, in solidaristic egalitarian pursuit of nationally standardized wages and employment conditions. However, getting members to support common demands in spite of often widely different situations at their workplaces required strong organizations. While political ideology helped, the capacity of national unions to aggregate the interests of a diverse membership depended in large part on whether employers were supportive of negotiation at national level. In countries with an established craft legacy bargaining remained mostly decentralized and fragmented as industrial unions had to allow their members independent pursuit of sectional interests, on the model of and in competition with skilled unions. Sometimes, especially in the United States, unskilled unions managed to establish rights to 'job property' for their members as well and enforced on employers their own version of a closed shop, under which new hires had to join the union within a short period or lose their job (post-entry closed shop).

Union organizational development in the nineteenth and twentieth centuries included the rise of national federations of local and sectoral unions, which occurred in close correspondence with the formation of the modern nation-state and the establishment of liberal democracy. It also involved growing independence of unions, not just from governments but also from friendly political parties, enabling even the industrial unions of the European continent to deal with democratically elected governments of conservative political complexion. Related to this was increasing reliance of unions on collective bargaining with employers as their main mode of operation, with either individual firms or employers associations. This presupposed the elimination of 'yellow' company unions controlled by employers as well as recognition of collective bargaining in labor law, for both of which government support was crucial. Collective bargaining, in turn, required that unions established organizational authority over the strike—the collective withdrawal of the labor power of workers from employers—just as political recognition of unions demanded that strikes were called only for economic purposes, abandoning the syndicalist tradition of the 'general' political strike. Union organizational development further involved integrating the growing number of white-collar workers, either in separate unions or inside encompassing industrial unions, which followed different trajectories in different countries and generally remained much less successful than the organization of blue-collar industrial workforces.

The literature offers a variety of explanations for the growth and decline of union membership over time and the differences in union density rates between sectors and countries. Unions tend to be strong in the core sectors of export-oriented manufacturing, like metal engineering, in large establishments, and in the public sector, while they are weak in private services and in small firms. But cross-country differences are considerable, and indeed on hardly any parameter of social structure do developed industrial societies differ as much. While in some Scandinavian countries around 80 percent of wage earners are union members, in France and the US density had fallen to about 10 percent by the end of the twentieth century. And while union membership relative to the workforce declined during the 1980s in the United Kingdom and the Netherlands, it increased in Belgium and Finland. Fluctuations in membership over time can also be enormous, as they were in Britain where in the three years between 1920 and 1923, union density declined from 45 to 30 percent.

Econometric studies have found that in the United States and Britain in particular, union density tended to increase with nominal wages and inflation, while it was negatively affected by unemployment and an already high level of unionization (Bain and Elsheikh). However, studies of this sort cannot explain the variation in unionization between countries, or ruptures in historical continuity. Institutionalist approaches take into account that union growth includes the acquisition by unions of recognized status, based on agreements with employers or the state helping unions recruit members ('union security'). Often such status is gained in exchange for political and economic moderation. Differences in union density then reflect different opportunities for unions to organize and keep members ('workers don't organize unions; unions organize workers'). In addition to the closed shop in its many official and unofficial versions—which makes membership compulsory by making access to employment dependent on it—employers may assist unions in collecting their membership subscriptions.

Similarly, favorable legislation on union recognition or workplace representation through works councils may improve unions' ability to recruit. In some countries, like Switzerland, non-members are legally obliged to pay union dues, and in Belgium and Denmark unions administer parts of the public social insurance system, enabling them to discriminate against non-members and thereby create additional incentives to join.

More generally, the organizing problems of unions may be described in terms of generic problems of collective action in the rational pursuit of group interests. Union achievements, like higher wages or favorable social policy legislation, are often collective goods in that they cannot be withheld from workers who have not contributed to their production. Apart from sentiments of fraternal solidarity, there is then little incentive for workers to join a union, even if they find the outcomes of union activities desirable. Indeed, given the risk that others will 'free-ride,' members may contribute in vain, as the union may not find enough support to be successful. Like other collective organizations, unions therefore typically offer potential members 'outside inducements' (Olson), i.e., benefits other than the collective goods for which they were originally founded, which unlike these can be limited to members (for example, healthcare benefits). They also strive to make membership less voluntary and obligatory, often with the help of the state or the employers.

4. *Unions as Collective Actors*

As intermediary organizations, modern unions represent their members in the labor market, the workplace, and the polity. Depending on their membership, their organizational structures and traditions, and the constraints and opportunities offered by employers and the state, different unions attach different significance to the three arenas and coordinate their activities in them differently.

4.1 *Unions in the Labor Market*

Unionism, and indeed much of modern labor law, assumes that individual workers are at a disadvantage vis-à-vis employers, and that redress of this imbalance requires that workers combine to face their employer as a collectivity. By regularizing wages and conditions of employment, sometimes also through state legislation, unions try to take wages and conditions out of competition, to make the market more predictable and protect workers' living conditions from its fluctuations. In this sense, Polanyi regards unionism as an element of the social counter-movement against the 'satanic mill' of the 'self-regulating market' of early liberalism.

Economists have long debated whether unions can raise wages above market level. Clearly, they can get nominal wage increases that exceed the increase in productivity, which depending on the government's monetary policy may result in either inflation or unemployment. With respect to the wage structure, in fragmented union systems skilled unions may use their independence to defend wage differentials, causing a wide wage spread between skill groups and industries. Fragmented ('competitive') bargaining may also enable well-organized groups of strategically placed unskilled workers to make high relative wage gains, which in turn may result in a 'leapfrogging' pattern of general, and potentially inflationary, wage increases. Wage differentials are also high in economies in which unions have little or no influence.

By comparison, industrial or general unions typically entertain an egalitarian ideology and strive to compress wage differentials, raising the wages of their less well-paid members more than those of the others. The result tends to be increased capital intensity and unemployment at the lower end of the labor market. Industrial unions therefore favor extensive training and retraining, funded by employers or the state, to increase the productivity of the less skilled to a level where they can earn the high wages negotiated for them. In favorable circumstances, this may lead to a general upgrading of a country or sector's production system, in the direction of more skill intensive and less price competitive production.

Since the 1970s, there has been a debate on the relationship between centralization of collective bargaining and various dimensions of macro-economic performance, such as inflation, growth and employment. In the 1970s centralized bargaining by national or industrial unions was believed by many to be more economically beneficial than fragmented bargaining, as it made it more difficult for bargainers to disregard the costs of redistributive bargaining or productivity-restraining work rules. Later it was suggested that bargaining agents in a decentralized wage setting system were more likely to be disciplined by market pressures, the problem being intermediate levels of centralization and unions both large enough to override the market, and small enough to be able to externalize the costs of their behavior to society. Statistical analyses have remained inconclusive and suggest that the economic effects of bargaining structures are highly conditional on economic and institutional circumstances.

4.2 *Unions in the Workplace*

Comparative research suggests that employers are willing to agree to centralized bargaining if this helps them keep unions out of the workplace. Workplace unionism, especially in a craft tradition, detracts from managerial prerogative and may interfere with management attempts to increase productivity. On the

other hand, even industrial unions need to be present in the workplace, if only to recruit members, monitor the implementation of collective agreements, and generally insert themselves in the governance of the employment relationship and the 'wage-effort bargain' inherent in the labor process. Balancing central and workplace representation, or external and internal union organization, was and is one of the main problems of unionism and industrial relations.

Least of all employers seem to like multi-unionism in the workplace. This has often led to agreement with industrial unions on unified arrangements for plant level workforce representation, like in Italy. In some countries, like the US, single-union representation is established by a legal procedure. In others workplace unionism is merged into institutions of legally based participation, giving works councils elected by the workforce as a whole rights to information, consultation, and co-decision making while ensuring that management has a single counterpart with which to negotiate. While works councils in most countries cannot call strikes, they tend to have close relations with internal and external unions, not least since most of their members are usually also union members.

In a number of Asian countries, unions have evolved into 'enterprise unions' that are only loosely federated above enterprise level. This corresponds to an environment of closed internal labor markets and lifetime employment of core workforces. Enterprise unions tend to be highly cooperative with managerial efforts to increase productivity and competitiveness as they represent the close community of fate between their members and the enterprise that employs them. Often their officials are managers whose career is advanced by serving for a time in a union position. Works councils in European countries also tend to cooperate with management. Unlike enterprise unions, however, they exist next to an external union that can call strikes without regard to the competitive position of individual workplaces.

The economic effects of different forms of workplace unionism are still being debated. For some time, decentralized unionism and workplace bargaining were associated with high strike rates whereas centralized and politically influential unions, like those in Scandinavia, were found to be highly effective without much recourse to strikes. Especially where distributive wage bargaining is moved to the national or industrial level, or where distributive conflict is suspended by jointly faced competitive market pressures, workplace representation was shown to enhance efficiency, among other things by enabling workers to voice complaints as an alternative to quitting. The same effect has been attributed to works councils. By comparison, in countries with adversarial industrial relations traditions, external labor markets, decentralized wage bargaining, freedom of workplace unions to call strikes, and a history of craft unionism and restrictive practices, workplace unionism came to be regarded in the 1960s and 1970s as a source of inefficiency. Here employers began to devise their own structures of workplace cooperation and 'worker involvement,' in the context of the developing practice of 'human resource management,' which often but not necessarily presupposes or aims at a 'union-free environment.'

4.3 Unions in the Polity

In most European countries, the rise of unionism was closely intertwined with nation-building and state formation. T. H. Marshal regards unionism as a step in the evolution from civil, to political, to social rights of citizenship, with unions using political rights to collective organization in the pursuit of social rights, albeit in the civil sphere of the marketplace and through freely negotiated—collective—contracts ('industrial citizenship'). As early repression of unions gave way to inclusion in shared public spaces (Crouch), unions became firmly integrated in the social and political order of the Western European nation-state, especially under the second postwar settlement after 1945.

The high point of union inclusion in advanced industrial countries was the 'neo-corporatist' period of the 1970s and early 1980s. Especially in countries with a tradition of political unionism and centralized collective bargaining, unions were able to get major concessions from governments and employers in 'political exchange' for wage restraint, which in turn was to enable governments to protect full employment with Keynesian means without having to accept excessive rates of inflation ('incomes policies'). Among the concessions gained by unions in this period were legislation on union recognition and access to the workplace; various other forms of 'union security'; legal extension of collective agreements to all firms in an industry; higher social security and public pension benefits; and industrial democracy legislation. Neo-corporatism was more likely to work where governments were of a social democratic complexion; but most conservative governments were also eager not to antagonize the unions. The more politically powerful unions became under corporatist arrangements, the lower as a rule was the strike rate as strong unions were able to get in the political arena what others had to struggle for, with uncertain event, in the industrial arena.

The historically close involvement of union movements with their respective nation-states explains to a large extent why unions and collective bargaining have so far resisted economic pressures for institutional convergence. In fact, the structure of unionism and its status inside its institutional context are an important dimension distinguishing between the different 'models of capitalism' that have attracted attention in the 1990s. The lasting symbiosis between union move-

ments and national states also accounts in large parts for the weakness of international unionism. Such weakness is found even in the European Union, where in spite of deep economic integration national unions have remained highly distinctive in their organizational structures and *modus operandi*, and are jealously guarding their independence.

5. *The Prospects of Unionism Today*

Much of the literature on unions in the early 1960s observed and predicted a decline of unionism in leading industrial societies. In most Western countries, with the notable exception of the United States where union decline continued unabated, this was contradicted by the 'hot autumns' of 1968 and 1969, which were followed by a sharp reversal of membership losses and by unprecedented influence of unions in politics. In the 1990s, however, union membership and power were again declining worldwide. With the American private economy on the way to effective de-unionization, the British government in the 1980s attacked unions' legal and institutional supports, with considerable success. Outside the Anglo-American world, union influence was weakened by the end of politically guaranteed full employment combined with intensified international competition, while the growing importance of the service sector and the 'knowledge economy' and the rise of a more diverse and better educated workforce made it more difficult for unions to recruit new members. Even where institutional conditions remained unchanged, unionism tended to be increasingly confined to the traditional sectors of the economy, which have long been shrinking in size and importance.

In their history unions have become deeply involved in the functional and social integration of the modern nation-state: in the national standardization of employment practices, the organization of redistributive solidarity at national level, and the political governance of the national economy. The decay of the second postwar settlement coincides with strong market pressures for more flexible modes of governance, less political interference, and internationalization of economic transactions. It is not clear whether and how unions, and indeed the nation-state itself, will find a response to the de-nationalization of the global economy. With their membership aging, unions may become defenders of an industrial employment and social security regime that no longer fits the needs of the rising post-industrial sector of the workforce. Being unattractive to the winners of structural change in the highly skilled and female segments of the labor force, unions also seem less than successful in organizing the new underclass of low-paid service workers, many of them immigrants, that has been growing for some time in many industrial societies.

See also: Class Consciousness; Class: Social; Conflict: Organizational; Conflict Sociology; Industrial Policy; Industrial Relations and Collective Bargaining; Industrial Relations, History of; Industrial Sociology; Labor History; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Labor Movements, History of; Labor Supply; Work, History of; Work, Sociology of

Bibliography

Bain G, Price R 1980 *Profiles of Union Growth: A Comparative Statistical Portrait of Eight Countries*. Blackwell, Oxford, UK

Crouch C 1993 *Industrial Relations and European State Traditions*. Clarendon Press, Oxford, UK

Crouch C, Pizzorno A 1978 *The Resurgence of Class Conflict in Europe*. Macmillan, London

Dunlop J T *Industrial Relations Systems*, rev. edn. Harvard Business School Press, Boston, MA

Ebbinghaus B, Visser J 2000 *Trade Unions in Western Europe since 1945*. Macmillan, London

Lipset S M (ed.) 1986 *Unions in Transition: Entering the Second Century*. Institute for Contemporary Studies, San Francisco

Marks G 1989 *Unions in Politics: Britain, Germany and the United States in the Nineteenth and Early Twentieth Centuries*. Princeton University Press, Princeton, NJ

Rogers J, Streeck W (eds.) 1995 *Works Councils: Consultation, Representation, and Cooperation in Industrial relations*. The University of Chicago Press, Chicago

W. Streeck

Labor, Division of

Division of labor is a feature of social structuring found in all human societies. Its form varies widely and changes over time. In a narrow sense the concept is used in an economic context to describe the breakdown of a complex task into a number of detailed and specialized tasks. In a social context the meaning is much wider, because labor and hence the consequences of its division impacts decisively societal structures. An example familiar to everyone is the household division of labor. This refers to the differential allocation of tasks to women and men in the family and in childrearing. Following a definition of *labor*, this article will consider the complexity of this concept by focusing on different levels and dimensions of the *division of labor*. Then its role in social development will be dealt with. The division of labor in modern capitalist-industrial societies will be granted the main attention. Finally, the development of social thought on the concept will be briefly examined.

1. *Labor*

Labor or work can be defined as purposeful activity necessary in any human society to secure its existence. Only humans do work in the sense of using creativity, conceptual and analytical thinking together with manual aptitudes to transform nature for their use (see *Work and Labor: History of the Concept*). A distinction between work and labor is often made. The word 'labor' is derived from the Latin *labor* implying toil and distress, and hence labor alludes to arduous work, done under some duress and control by others. But the use of words is in fact not that strict. With regard to the specialization of activities, it is common to speak of 'division of labor' (and not of 'division of work'). Similarly, it is common for example to speak of the 'labor process,' the 'labor market,' the 'labor movement' (including the 'labor unions' in the economic, the 'labor party' in the political sphere), or the 'International Labor Organization' (ILO) and the expression 'work' is not used here.

2. *Divisions of Labor*

2.1 *Concepts: Economic Division of Labor vs. Social Division of Labor*

All empirical evidence shows that labor or work always entails some specialization. *Division of labor* refers to separation of activities and the specialized allocation to different individuals. It is a universal trait of human existence. This does not, however, imply that it is caused by natural differences (biological differences between women and men, for example). Division of labor is always human-made, its forms are socially shaped.

Any definition of division of labor basically must start with the recognition of two different connotations. In its narrow and simple sense, the concept is used in an *economic context*. It describes the splitting up of a complex productive task into a number of specialized, simpler tasks. The most renowned example is that of Adam Smith (1776) for pin needle production. The *increase in productivity* is exactly the ultimate reason for the separation and specialization of tasks in manufacturing.

This form is known as *detailed* or *technical division of labor*. It made its appearance on the stage of human history with all-pervasive force only three hundred years ago in Europe with the establishment of conditions not in use previously: that is manufacturing and the 'invention' of capitalist principles of production.

In a broad sense, division of labor is a precondition for conceptualizing society, as used in a *social* or *sociological context*. Reference to the *social division of labor* implies divisions at different levels of society which comprise its complex structure. Here the attention is on social differentiation such as class, gender, or ethnicity; on the role of power; on forces of social cohesion and disintegration; and on the importance of solidarity and morale. All the major institutions of a modern complex society play a part in the *social division of labor*: in the *economic system* with its elements like the market, competition, capital, contract law, labor market, even differences between (paid) employment and non-paid labor; in the *political system* with its various specialized institutions of the legislative, the executive, and the judiciary; in the *cultural system* with its various socializing institutions for the creation of skills, value orientations, and spiritual meaning.

In a schematic form the *concepts* may be listed this way:

economic division of labor	*social division of labor*
detailed division of labor (e.g., technical division of labor)	*division of labor* (e.g., by gender, occupations)

2.2 *Levels*

It is clear from the above that division of labor is a complex concept and can refer to different levels of human activity. It extends from the household or family on the *micro level*, through work organizations like enterprises on the *meso* (*intermediate*) *level*, divisions in society at large on the *macro level*, to the entire world on the *global level*. Examples of divisions of labor on the various levels are the domestic division of labor, the organizational division of labor, the occupational division of labor, or the international division of labor.

Here is an overview on the levels in schematic form:

Micro-level: e.g., domestic or familial division of labor
Meso- (*intermediate*) *level*: e.g., organizational division of labor
Macro-level: e.g., occupational division of labor
Global level: e.g., international division of labor

2.3 *Dimensions*

It is necessary to trace division of labor in various different dimensions. The most obvious dimension is a broad division of labor between women and men, which all known societies exhibit in some manner or other. This *sexual* (or *gendered*) *division of labor* is obviously important in the area of work, but it also reaches beyond that to social, political, cultural, and religious functions.

Any analysis of the social structure of a society, or a comparative study of different societies, must certainly consider the distinction between these varying dimensions. In addition, some dimensions are relevant at more than one level. The sexual division of labor,

for example, is of tremendous importance at the domestic or family level, but it also plays an important role in the economic realm. It is inseparably connected with opportunities and status in employment on the societal level, and is in evidence even on the international level. (When, for example, the work of British or German male textile workers is nowadays performed by possibly lower paid women in developing countries).

An overview of the more important *dimensions* of the division of labor in schematic form:

division of labor by sex (or *gender*)
division of labor by age
division of labor by occupations
division of labor by skill
division of labor by hierarchy
division of labor by space
division of labor by time.

3. *Division of Labor and Development*

3.1 *The Basis for Complexity*

Taking the broad sweep of human history, the contrast between very simple and vastly extended complex social structures springs to mind. Division of labor and concomitant social differentiation are the processes that underlie shifts towards more complex societies.

Until around ten thousand years ago hunting and gathering bands were small and there was little role specialization beyond a division of labor by sex and age. Only since settling down to agrarian provisioning for survival has human social life slowly developed towards the complexity known from recorded history.

3.2 *Traditional and Modern Societies*

Agrarian (pre-industrial) societies are characterized by the fact that the overwhelming portion of productive tasks are performed in agriculture and self-provisioning of the household. The (extended) *family* is also the productive unit. The social division of labor is low, insofar as only a small proportion of the population is engaged in other occupations, such as specialized crafts, trading, military functions, and religion. *Tradition* is of eminent importance. The distribution of functions in the societal labor process is based on *ascription* (being born into ones role and status), not on *achievement*. The pre-condition for a considerable increase in the division of labor is production of a larger surplus beyond the self-provisioning of the individual families (see *Societies, Types of*).

There was, however, no unilinear progression in history towards ever-greater complexity, no continuous evolution from primitive to modern societies. The development of social forms shows long periods of remission and reversals, as well as change proceeding along more than one line.

Traditional agrarian societies typically are *feudal societies*, and the division of labor is often based on coercion and power. Serfdom (especially among the agrarian population) is widespread, and slavery (for labor that is painful and exacting) is not uncommon.

For thousands of years in human history static forces were dominant in the division of labor. This came to an abrupt end when, in Europe (beginning in England), *capitalism* and *industrialization* began to transform the world. The last two to three hundred years have seen by far more changes in the division of labor than all the previous ages. Compared with development up to then, capitalist transformation marks a quantum leap for the division of labor.

4. *The Division of Labor in Modern Societies*

4.1 *The Institutional Frame*

The capitalist-industrial mode of production was established on the basis of a number of institutions which did not previously exist. There is a very close interconnection between the division of labor and these institutions. The following can be listed as the major specific conditions (e.g., Weber 1922/1968; Marx 1867/1976) (see *Capitalism*).

4.1.1 Separation of household and workplace. With the introduction of manufacturing in special premises, gainful employment and the family become separated. A previously unknown division is introduced between the *public sphere* of work and the *private sphere* of home. The function of the *family* (or the household) changes dramatically as it is no longer a combined unit of production and reproduction. The division implies the separation between *unpaid labor* in the household and *paid labor* in the workplace. This is responsible for a thorough change in the *sexual division of labor*, as it provides a new basis for power relations in the family and for the *gender typing* in employment.

4.1.2 Concentration of labor in the manufacture and the factory. Assembling of workers in great numbers in one place under one direction was the prerequisite to splitting up complex craft work into a multitude of detailed tasks. With this *detailed division of labor*, enormous gains in productivity were possible. This usage of labor carries a number of (economic) advantages: unskilled workers who receive relatively low wages and are easily replaceable can be engaged; and, control over speed and output is easy. Fragmentation into simple repetitive functions is also the basis

for the subsequent mechanization and automation which again pushes productive output. However, the price paid for rising productivity is *alienation* (see *Alienation, Sociology of*).

4.1.3 Private property. In the process of industrialization, property acquires a new social function and is important for the division of labor in two respects. First, it provides for those who plan and control the *legitimacy* to enforce the manufacturing (detailed) division of labor. Second, it provides the basis for the accumulation of wealth. *Profit* becomes the main motive for organizing a labor process. This implies that the social function and the importance of *money* changes completely: from a generalized exchange medium on local markets to a carrier of value and power that makes the world go round.

4.1.4 Free labor. A *labor market*, where labor can be bought and sold as a commodity, was not common in human history. Earlier times depended on domination and power for the use of labor that was often unfree labor (patriarchy, serfdom, slavery). Those were the typical characteristics of a *forced division of labor*. The modern development relies on free contracting, i.e., on *wage labor*. For labor, this implies that social relations of a double character become relevant: on the labor market, the relation is between equals, whereas, inside the workshop, the relation is one of subjugation under domination and power (Rueschemeyer 1986).

4.1.5 Liberalization of markets and competition. The importance of liberalism in greatly increasing the wealth-creating capacity of a nation was strongly emphasized in the classical 'political economy.' Unrestrained by administrative rules of government, a free market encourages producers to specialize in activities where they have a relative or absolute productive advantage. Through specialization, they benefit from greater dexterity, more efficient use of materials, time, expertise, and technology. At the same time, the invisible hand of competition penalizes ineffective producers by removing them from the market and encourages the rational (cost-effective) exchange of goods and services.

4.1.6 Application of science, technological change, mechanization, and automation. In the pursuit of profit and under the constant pressure of competition, science is applied systematically for the creation of ever more sophisticated machines and procedures to render production processes more efficient. Technological change thus becomes a permanent feature of working life. Human labor that remains in the process is forced to fill the gaps left by machinery and to adapt constantly to changing requirements.

4.2 Division of Labor and the Dynamics of Change

4.2.1 Structural change of society. Change has remained a fundamental feature of the division of labor ever since industrialization altered occupational structures in an unprecedented manner. This in turn affects division of labor on all levels, and in all dimensions. In the last decades the highly developed societies have changed to *post-industrial societies*. Leaving industrial production behind as the major field for employment has ushered in the *service society* or *knowledge society*. The occupational division of labor expands into new fields. Provision of knowledge-driven services, evaluations, information, guidance, analysis, etc. outnumbers now by far industrial production work. The huge diversification of the labor structure through industrialism and the rise of service and knowledge work is reflected in the fact that up to 30,000 distinct jobs are listed now in some countries, as compared to no more than 50 crafts and trades in pre-industrial societies.

Knowledge and software have become the main themes in the 'new economy' (Leadbeater 1999). Increases in the professional level of many jobs and the professions as such gain importance as knowledge-occupations (Lane et al. 2000). Production work and hardware decrease, relatively speaking, in importance. This is reflected in the continuous decline in the share of blue-collar work and the growth of white-collar work. Often, manual labor is either relocated in other countries or is eliminated altogether through automation. Together with the often-practised outsourcing of work functions, this brings about considerable shifts for the spatial (or geographical) division of labor. The new information and communication technologies have furthered these trends of global expansion of the division of labor, under which the whole world potentially becomes one workshop.

4.2.2 Changes in work organization. Assembling people in large numbers and assigning them detailed tasks requires coordination and planning to reach the common goal. As a result, organizations have become an indispensable feature of modern life in many areas. At the same time they are of foremost importance for the conditions of work.

Organization theory and management have developed into vast fields of their own. This has developed on three major influences that shaped the division of labor thoroughly.

Emphasis on 'rationality' was the main concern of Max Weber (1922/1968) when he developed the principles of *bureaucracy* as a design for modern administration. For production work, *The Principles of Scientific Management* were developed by Frederick Taylor (1911) which aimed at improved industrial efficiency. Among Taylor's principles, the separation of planning work from execution imposed a division of labor that relied on tight control and hierarchy, destruction of skills and discretion of the workers. Taylor's whole system based the division of labor on a low level of *trust*. The industrialist Henry Ford added the mechanical disciplining power of the assembly line. Furthermore, he combined *mass production* with the cultivation of *mass markets*. Since then, bureaucratic organisation, *Taylorism* and *Fordism*, have been so successful that they shaped the form of the division of labor and working conditions in industry all over the world. This includes the communist or state-socialist countries which thus did not seize their unique opportunity to realize a different form for the division of labor (e.g., Sayer and Walker 1992). Widespread application of Taylorist and Fordist principles of work rationalization led to the 'degradation of work' (Braverman 1974) in mass industry in the twentieth century. The impact on working conditions has been discussed deeply in the ensuing 'labor process debate.'

The grounds for a *New Division of Labour* (Littek and Charles 1995) arose from a number of far reaching changes in the markets, a new quality of competition, and growing pressure for flexibility and customer services which together marked the 'end of mass production.' The abolition of detailed fragmentation of work, tight control, and discretionary power of the hierarchy in order to raise creativity, quality standards, and service orientation were important for the new division of labor. Elements that have gained in importance are decentralization, networking, flexibility, upgraded skills, motivation, self-organization, responsibility, and autonomy. These are conditions which are based on *trust* as a means of coordination, quite the opposite to the systems of Taylor and Ford. The new model is a 'learning organization' with self-regulation of autonomous individuals.

The spread of these new forms was pushed by the technological development which made simple and repetitive tasks redundant though automation. The trend to these new forms, flexible specialization and self-regulated creativity, was further encouraged and enhanced with the development of the 'new economy' (see Sect. 4.2.1).

4.3 Division of Labor, Social Differentiation, and Cohesion

Social differentiation is so closely connected with the division of labor that they could almost be treated as synonymous. The wide development of the division of labor under modern industrialism (mostly in capitalist, but also in socialist or communist form) caused a similarly wide expansion of *differentiation* in society. Along with this expansion, social and economic *interdependence* has increased enormously: each individual in society is dependent on so many others for its existence.

Social differentiation itself has to be examined from various angles, the most important of which are discussed below.

Class and *social stratification* are unthinkable without a connection to the division of labor. They are relevant as a pre-condition, as well as a result of, the division of labor. With regard to *class*, most social scientists would agree that the number of classes needed to describe societal structure in complex societies has to be extended beyond the two basic ones (bourgeoisie and wage-labor) Marx had in mind. Often the concept is replaced altogether by *stratification*. Nevertheless, ownership of productive property is important as a pre-condition for the right to organize labor or otherwise being organized. The gains from economic activity shape wealth distribution. As a result of the differentiated division of labor, the *inequality* of the wealth distribution has grown in most countries over time. At the global level, a similar trend of growing disparity can be observed in the distribution of wealth between rich industrialized countries and poorer developing countries as a result of *international division of labor* in practice.

As a major element in differentiation, *gender* remains a cause for inequality. Even though patriarchy has lost ground and equal opportunity rights are enforced in most modern societies, the effects of a *gendered division of labor* still are to be found in practically all societies. These often include lower income, confinement to lower skilled jobs, reduced access to leading functions, broader employment in precarious jobs and in part time, and so on. This can be explained by the fact that the *household* continues to be the basic social unit around which people conduct their lives (Pahl 1984), and that an unequal distribution of tasks in household work between women and men obviously continues to be practiced more or less throughout the world (e.g., Hakim 1996).

Finally, *social integration and cohesion* become much more important with the growth in inequality that follows the extension of the division of labor. *Cooperation* (as a passive medium) and *coordination* (as an active medium) are necessary to counterbalance the diverging forces and to secure social integration at least at a minimum level. Social integration on the societal level can be based above all on two institutions: trust, motivation, values on the one hand and power on the other (which is based on legitimacy in modern states). The strategy in most highly developed countries is to secure *both* by providing a minimum of social security and administering some

form of social regulation to the market economy. A socially responsible division of labor thus may be seen as the guiding vision for modern countries.

5. *Division of Labor in Social Thought*

Because the division of labor is a fundamental structuring principle of complex societies, it has played an important role in theories about society and social relations. Interest in division of labor was particularly strong in the late eighteenth and the nineteenth century in connection with the unprecedented social changes brought about by capitalist industrialism. This was also the time of the formation of *sociology* and *political economy*. Later, the division of labor as a category in social theory was somewhat neglected (Sayer and Walker 1992), but interest has risen again in the last 25 years. This is mainly due to:

— a broad expansion of studies into the sexual or gendered division of labor and the connections between household work and employment (see *Gender, Economics of*);

— studies on the so-called globalization, which extends the reach of economic activities to a scale of global or international division of labor based on Western business values (see *Globality*);

— studies on the growing pressure for labor flexibility in the new economy, which indicates a loss of dependable employment conditions and of elements that formerly structured work like defined time, space, or occupation (see *Industrial Sociology*).

In the history of theory on the division of labor, a number of key authors should be mentioned *Ancient Greek philosophy*, and particularly Plato and Aristotle, recognized the division of labor or specialization in occupations as a necessity for social well-being. This was linked to exchange which requires a market and a currency (as a medium of exchange). The socializing effect stood in the foreground for Aristotle when he noted that no society made up of two physicians could emerge, but a society containing a physician, a farmer, and other occupations would exist.

At the beginning of modern times, *Adam Smith* (1723–1790) is of the utmost importance for the development of a discipline which became known as political economy. No one has ever placed more emphasis on the division of labor. For Smith, it is practically the only factor driving economic progress. In his book *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776), Smith explored the possibilities for raising productivity along two avenues: the detailed division of labor within the manufacturing process, and the liberalization of markets in the economy as a whole. He recognized the egoistic pursuit of individual interests, but not any feelings of altruism or care for the needs of others, as the driving force of economic activity.

For his contemporary Adam Ferguson (1723–1816), social differentiation through the division of labor stood in the foreground, mainly with regard to occupational differentiation. He considered it as a specific structuring type for human solidarity, as did Durkheim.

Both the economic and the social aspects of the division of labor were analyzed by Karl Marx (1809–83). In a similar way to the classical political economy theorists. Marx elaborated the enormous potential of labor for productivity in manufacturing and big industry. But, drawing on very similar insights, he came to a different conclusion to that of Smith: that there is a *fundamental difference* between division of labor among autonomous parties, 'division of labor in society,' and a division of labor that takes place under the authority and in the interest of one dominant party, 'division of labor in manufacture.' The contrast between 'anarchy' of the market (reflected in the social division of labor) and 'despotism' in the workshop (reflected in the manufacturing division of labor) was highlighted by Marx as mutual conditions and an essential feature of *capitalism* (1867). The separation between owners and non-owners of the means of production found recognition as the basis for class formation in society and for the asymmetry of power in the labor process. He was also strongly concerned with *deskilling* (cf. Braverman 1974) and with *alienation* as a result of the specific capitalist fragmentation of work.

Emile Durkheim (1858–1917), like Ferguson, directed attention mainly to the social division of labor (1893), for which he above all gained reputation. In contrast to more pessimistic views, he recognized that division of labor does not have to create disintegration, but may have a socializing effect. Durkheim thus discerns *mechanic solidarity* in 'primitive' societies from *organic solidarity* in 'complex' societies, while observing the danger of *anomie* and of a 'forced division of labor' in modern development (see *Solidarity: History of the Concept*).

In his analysis of the modern (Western) world Max Weber reached findings very similar to those of Marx. He also held (1922) that capitalist division of labor is something quite specific, based on the role of private property, on profit as the motive for production, and on the application of rational organizing principles. However, he put the weight differently. Weber was interested mainly in fundamental prerequisites of 'modernity,' among which legitimate power and 'rationalization' of all spheres of social life (including dependable rational administration in the form of 'bureaucracy') are of importance.

The *division* of labor by gender or sex received only scant attention from the so-called founding fathers of these theories. Among the few exceptions were Friedrich Engels and Thorstein Veblen. However, attention to this theme has changed considerably over the last decades with the emergence of an

extensive literature of feminist analysis or gender studies (see e.g., Hakim 1996; Pahl 1984).

See also: Capitalism: Global; Differentiation: Social; Durkheim, Emile (1858–1917); Feudalism; Gender, Economics of; Industrial Sociology; Industrialization; Labor Supply; Marx, Karl (1818–89); Modernity; Sex Segregation at Work; Slavery as Social Institution; Smith, Adam (1723–90); Social Stratification; Spencer, Herbert (1820–1903); Weber, Max (1864–1920); Work, Sociology of

Bibliography

Braverman H 1974 *Labor and Monopoly Capital. The Degradation of Work in the 20th Century*. Monthly Review Press, New York

Castells M 1996 *The Rise of the Network Society*. Blackwell, Malden, MA

Durkheim E 1893 *De la division du travail social*. Alcan, Paris (1984 *The Division of Labour in Society*. Macmillan, London)

Hakim C 1996 The sexual division of labour and women's heterogeneity. *British Journal of Sociology* **47**: 178–88

Lane C, Potton M, Littek W 2001 *The Professions Between State and Market*. CBR Working Paper 189, University of Cambridge, Cambridge, UK

Leadbeater C 1999 *Living on Thin Air. The New Economy*. Penguin, Harmondsworth, UK

Littek W, Charles T (eds.) 1995 *The New Division of Labour. Emerging Forms of Work Organization in International Perspective*. de Gruyter, Berlin, Germany

Marx K 1867 *Das Kapital* I (1976 *Capital*). Penguin, Harmondsworth, UK

Pahl R E 1984 *Divisions of Labour*. Blackwell, Oxford, UK

Rueschemeyer D 1986 *Power and the Division of Labour*. Blackwell, Oxford, UK

Sayer A, Walker R 1992 *The New Social Economy: Reworking the Division of Labor*. Blackwell, Cambridge, MA

Smith A 1776 *An Inquiry into the Nature and Causes of the Wealth of Nations*. Strahan and Cadell, London

Standing G 1999 *Global Labour Flexibility. Seeking Distributive Justice*. Macmillan, London

Taylor F W 1911 *The Principles of Scientific Management*. Harper, New York

Weber M 1922 *Wirtschaft und Gesellschaft*. (1968 *Economy and Society*). Bedminster Press, New York

W. Littek

Laboratory Experiment: Methodology

Experiments are special empirical investigations. They differ from nonexperimental empirical research methods. This article examines this difference (see Sect. 1).

Many researchers feel that only the experimental method offers the advantage of specifying causes and effects. This advantage is based on the unique possibility of controlling extraneous variables that can invalidate the empirical results (see Sect. 2).

By means of this control, an attempt is made to meet the various quality criteria of an experiment. The criteria of internal and external validity were developed within the context of an inductivistic theory, which has been criticized. Sect. 3 presents an alternative deductivistic theory of experiments.

Books dealing with the experimental method refer to the ideal control experiment. Results of empirical studies show the experiment to be a social interaction between the experimenter and the subject and thus to differ greatly from the ideal experiment. Sect. 4 focuses on these differences and their consequences for experimental research in the area of social and behavioral sciences. The experiment has been criticized for various other reasons (see Sect. 5).

1. Experimental vs. Nonexperimental Research

Let us use an example to illustrate the difference between experimental and nonexperimental research. The example concerns whether a specific film documenting the life of Jews in the Warsaw Ghetto affects the attitude of viewers towards Jews. A first study investigates the attitude of subjects who already had seen the film, and compares it to the attitude of subjects who had not. Imagine that this comparison shows that those who had seen the film have a more positive attitude towards Jews. A second study comes to the same conclusion, but in this study the investigator randomly assigns the subjects to viewing the 'film' (experimental condition) or to a 'no film' control condition and then measures their attitudes.

Both studies include two variables. The first variable is dichotomous ('film viewed' vs. 'film not viewed'), whereas the second (social attitude) has more than two values. How can the result of the first study be interpreted? Does the film positively influence attitude, or did those who watched the film have a more positive attitude towards Jews prior to the study? This question cannot be answered. It is therefore impossible to classify the variables as independent and dependent. In the second study, the investigator manipulates the first variable and randomly assigns the subjects to the 'film' and 'no film' conditions using what is in essence, a coin toss. In this way he creates a temporal sequence between the manipulated conditions (independent variable) and the dependent variable 'social attitude.' The subjects are randomly assigned to the conditions in order to prevent pre-experimental attitudes towards Jews differing between the two groups of participants. The result of this study can be interpreted in such a way that the viewing of the film had a positive causal effect on the attitude.

The second study is an example of experimental research. An experiment can be defined by the following criteria. The experimenter creates the con-

ditions, systematically varies them, and applies the principle of randomization. Experiments exhibit at least one independent and at least one dependent variable. The same also applies to quasiexperiments, though here the subjects are by definition, not assigned to conditions according to the principle of randomization (Cook and Campbell 1979). By contrast, the first study is an example of a nonexperimental, correlational study. The variable, film/no film, was not manipulated (varied) but selected by the investigator. The characteristic of 'randomization' is also missing. 'Selected' means that the investigator selected subjects for the study who already had assigned themselves to the conditions.

The terminology just outlined is widely accepted. However, sometimes the term 'experiment' is used with an adjunct. In these cases, the study being referred to is not always an experiment as defined above. This applies, for example, to the term 'ex *post facto* experiment' (Chapin 1947), where the conditions are not manipulated but selected by the investigator. The ex *post facto* experiment differs from a correlational study in that the former includes subsequent controls in order to permit the inference of a cause-and-effect relationship. A criticism of this procedure can be found in Campbell and Stanley (1963). On the other hand, the terms 'laboratory experiment' and 'field experiment' refer to true experiments. The adjuncts refer to different settings in which the experiments take place.

The present article deals with true (randomized) laboratory experiments and excludes other research methods. It describes the principles of experimentation in ways that are independent of any given discipline within the social and behavioral sciences.

2. *Experimental Error and Experimental Validity*

Not all research questions can be experimentally examined. Certain variables such as sex and age cannot be experimentally manipulated. Thus, if the relationship between age and intelligence is of interest, there is no opportunity for experimentation. However, this opportunity did exist for studying whether a specific film influences attitudes towards Jews. In the opinion of many researchers, such an opportunity should be taken because the experiment is superior to other methods for defending causal interpretations (e.g., Carlsmith et al. 1976). Random assignment permits us to conclude that viewing the film is the cause of the more positive attitude (see *Causation (Theories and Models): Conceptions in the Social Sciences*).

Experiments are subject to errors that jeopardize interpreting the relationship between an independent and dependent variable as unequivocally causal. A differentiation must be made between systematic and random errors (Cox 1961). An internal systematic error is present if the subjects who are observed under various experimental conditions also differ with respect to other variables. In this case the independent variable is confounded with extraneous variables. The criterion of internal validity (Campbell and Stanley 1963) has not then been met (see *Internal Validity*).

An external systematic error is present if the conditions of the experiment deviate 'from the conditions under which it is proposed to apply the conclusions from the experiment' (Cox 1961, p. 44). Such an error would exist if the experiment in the example given in Sect. 1 was performed using college students, but the conclusions regarding the effect of the film were not limited to persons with a higher education. In cases of external errors, the criterion of external validity (Campbell and Stanley 1963) has not been met (see *External Validity*).

Although random errors are not correlated with the independent variable, their control is important. Random errors affect the variation of the dependent variable within the experimental and control conditions. The attitude values of the subjects within the experimental condition 'film' deviate more or less from a mean, and the same applies to the attitudes under the control condition 'no film.' These dispersions are the result of so-called random errors. The smaller these dispersions are, the greater the precision of the experiment will be (Bredenkamp 1969). Since the relationship between the independent and dependent variable is statistically analyzed, the reduction of the dispersion resulting from random errors is very important. The concept of a random error differs from that of an error of measurement in classical test theory (see *Classical (Psychometric) Test Theory*). The dispersion of the random errors within the experimental and control conditions includes not merely the variance of the measurement errors, but also that of the true scores as defined in classical test theory.

2.1 *Internal Validity*

Controls to ensure internal validity promote an unequivocal causal interpretation of the relationship between the independent and dependent variables. In order to avoid the confounding of a known extraneous variable with the independent variable, the control techniques of 'elimination' and 'constancy of conditions' are employed. An internal error exists, for example, if a female experimenter measures the attitude of the subjects under the experimental condition, while a male experimenter measures the attitude under the control condition. The sex of the experimenter may have an effect on the dependent variable. To control this error, a single experimenter could collect all the data (constancy of conditions). An extraneous variable is eliminated, for example, if background noise that might reduce the audibility of speech is removed.

Unknown extraneous variables can be controlled by randomization. Randomization ensures that the expected values of the extraneous variables are identical under different conditions. Specific instructions exist concerning the random assignment of the subjects to the experimental conditions (e.gq., Keppel 1973 see *Random Assignment: Implementation in Complex Field Settings*).

Despite these controls, there remains the possibility that a factor is present that jeopardizes the internal validity of the experiment. Thus, for example, simply watching a film—regardless of its content—may have an effect on the social attitude. An experimenter who compares the attitudes under the conditions of film/no film, will overlook this possible error, even if the method of randomization was employed to control internal errors. There should therefore be at least one further condition included under which subjects view a film that is neutral with regard to its attitude toward Jews.

2.2 External Validity

In their highly influential article on the evaluation of experimental and quasi-experimental designs, Campbell and Stanley (1963) also introduced the criterion of external validity. External validity asks the question of generalizability: To what populations, variables, and settings can the experimental effect be generalized? Holding an extraneous variable constant in order to control internal errors always brings along the risk of reducing the external validity of an experiment. If a single experimenter performs the experiment on social attitudes towards Jews, the observed effect might then exist for this experimenter only. In order to examine this question, a second independent variable must be added to the experiment as a control factor so that its effects can be separated from the effects of the experimental independent variable. The control factor is introduced in such a way that it is not confounded with the experimental independent variable (see *Control Variable in Research*).

Of critical importance to external validity is the finding that the more positive attitude resulting from the 'Ghetto film' condition only applies to certain experimenters, not others. This is an example of a so-called 'disordinal interaction' between two independent variables (Bracht and Glass 1968). An interaction is called ordinal if the more positive attitude under the 'Ghetto film' condition occurs with all experimenters, but to varying degrees. In this case, who the experimenter is influences the dependent variable differentially by correlation, but the causal sign of the difference stays the same across all experimenters. The analysis of variance is frequently used for the statistical analysis of multifactorial experiments (e.g., Keppel 1973), and it can determine whether an interaction between the independent variables is present. However, it cannot identify the type of interaction involved.

Only disordinal interactions are problematic for the external validity of an experiment because they identify situations where the causal sign between the conditions of an independent variable varies, being positive with some experimenters in the example chosen but negative with others. Thus, identifying the type of interaction is important, and statistical procedures for doing this are presented in Bredenkamp (1982).

Questions about the population to which an experimental effect can be generalized often concern disordinal interactions (Bracht and Glass 1968). If only persons with college education participate in an experiment, then its external validity is endangered. It would be better to include variation in the level of education in the experiment as a control factor.

The question concerning the variables to which an experimental effect can be generalized presumes that these variables are fallible indicators of a theoretical construct. 'Social attitude' is a construct that can be operationalized in diverse ways. If different dependent variables, all assumed to be indicators of the same construct, are included in the investigation, it is called a multivariate experiment. This type of experiment allows us to examine whether the effects observed for a specific dependent variable also apply to other indicators of the same construct. Indeed, one could also compare several experiments using different indicators of the same construct (Bredenkamp 1980). In order to avoid a mono-operation bias, wherever it is feasible, a given construct should be operationalized in different ways (Cook and Campbell 1979).

Whether an experimental effect can be generalized to experimenters other than those used in a study touches on only one aspect of the setting in which a cause–effect relationship is examined. With regard to generalizability, field experiments are more externally valid than laboratory experiments whose validity for 'the real world' is questionable. Although laboratory experiments offer the advantage of controlling internal errors, they have the disadvantage of limited external validity. However, such considerations presuppose an inductivistic theory of experimental validity, and this has been criticized (see Sect. 3).

2.3 Precision

Increasing the precision of an experiment is important since statistics are used to make the decision about whether there is any relationship between the independent and dependent variable in a given experiment. Two errors can arise in this statistical decision. If there is no relation between the independent and dependent variable (i.e., the statistical null hypothesis is valid), but the statistical analysis

suggests such a relationship, then a type I error has been made whose probability is called α. The second error occurs if a relationship between the independent and dependent variable exists (i.e., the statistical alternative hypothesis is valid), but the statistical test fails to detect this relationship (type II error). The probability of this error is called β. The greater the experimental precision, the lower is the probability β. In other words, the greater the precision, the lower is the risk of retaining a false null hypothesis.

Controls reducing the dispersion of random errors serve to increase the precision of an experiment. One method consists of adding a control factor to the experiment. As mentioned, this measure is also employed in order to check external validity. If, for example, different experimenters collect the experimental data, taking them into account as a control factor can contribute to reducing the variance of random errors (error variance). The effects of the experimenters and their interaction with the treatment (see Sect. 2.2) can then be separated from the effect of the experimental independent variable. The larger the effect of the control factor and the interaction, the higher is the degree to which the error variance is reduced.

In some cases, this statistical procedure is also employeo increase precision when pretests are available. It is conceivable that one's attitude towards Jews could be influenced by a test carried out prior to the application of the treatments. To deal with this, subjects are first ranked according to their pretest values. Subjects with the first k values are then randomly assigned to the k experimental conditions. The same applies to subjects with ranks $k+1$ to $2k$, etc. This method is called matching. The degree to which precision is increased depends on the magnitude of the correlation between pretest and dependent variable (Feldt 1958). Under the aspect of external validity, the performance of a pretest has the disadvantage that the result may be generalizable only to pretested subjects (Campbell and Stanley 1963).

Still other procedures for increasing the precision of an experiment exist. Among these is the possibility of applying all treatments to each subject in a random sequence. In this case, the inter-individual variation of random errors is controlled. The effect of the experimental independent variable is then tested against the intraindividual error variance that is generally smaller than the interindividual error variance. A possible disadvantage of this procedure is the risk of decreasing internal validity due to repeated testing (e.g., carryover effects, fatigue).

3. *A Deductiviristic Theory of the Experiment*

Campbell and Stanley's theory of internal and external validity (1963) is based on inductive inference. This is particularly apparent with regard to external validity. Since no conclusive induction principle exists but statements about internal and external validity presuppose just such a principle, the theory has been criticized by Gadenne (1976).

An alternative, deductiviristic theory of the experiment has been proposed by various authors (Bredenkamp 1980, Gadenne 1984, Hager and Westermann 1983, Erdfelder and Bredenkamp 1994). According to Erdfelder and Bredenkamp (1994) various methodological rules of experimentation can be derived from the general principle of maximizing the fairness and strength of an experiment. Fairness is defined as the probability that a scientific hypothesis will be confirmed in an experiment, if it is actually true. Strength is defined as the probability that a false scientific hypothesis will not be confirmed. If scientific hypotheses are tested statistically, and the acceptance of the statistical null hypothesis is the result that is in accordance with the scientific hypothesis, the following applies (see Erdfelder and Bredenkamp 1994):

$$\text{Fairness} = (1-\alpha)-(1-\alpha-\beta)\,h_0 \qquad (1)$$

$$\text{Strength} = (1-\beta)-(1-\alpha-\beta)\,g_0 \qquad (2)$$

If acceptance of the statistical alternative hypothesis is the result that supports the scientific hypothesis, then Eqns. (3) and (4) hold:

$$\text{Fairness} = (1-\beta)-(1-\alpha-\beta)\,h_1 \qquad (3)$$

$$\text{Strength} = (1-\alpha)-(1-\alpha-\beta)\,g_1 \qquad (4)$$

α and β are the probabilities of type I and type II statistical errors (see Sect. 2.3), g_0 is the probability that the statistical null hypothesis is valid if the scientific hypothesis is not true, and h_0 is the probability that the statistical null hypothesis is not valid if the scientific hypothesis is true. Accordingly, g_1 is the probability that the statistical alternative hypothesis is valid if the scientific hypothesis is not true, and h_1 is the probability of the statistical alternative hypothesis being untrue if the scientific hypothesis holds.

Further discussions will be restricted to Eqns. (3) and (4). They can be similarly applied to Eqns. (1) and (2). If the scientific hypothesis and the statistical alternative hypothesis are logically equivalent, then $g_1 = 0$ and $h_1 = 0$. In this case, fairness and strength correspond to the probabilities of correct statistical decisions, $1-\beta$ and $1-\alpha$. This equivalence seldom applies in psychology. Psychological hypotheses are often related to processes occurring within individuals, while statistical hypotheses refer to populations. Nevertheless, the statistical analysis does make sense provided that the scientific hypothesis (S) implies a statistical hypothesis, for example, the alternative hypothesis (H_1). Randomization is indispensable for establishing such an implication. If the implication $S \rightarrow H_1$ holds, then $h_1 = 0$. In this case, fairness is completely determined by the probability of a correct

decision in favor of the statistical alternative hypothesis. The strength remains unknown because the value for g_1 is not known. However, the falsification theory put forth by Popper (1982) contains methodological rules that can be applied to reduce g_1. The smaller g_1 is, the greater is the strength for a given pair of values of α and β. If $g_1 < 1$ and $S \rightarrow H_1$ holds, then a reduction of the statistical error probability, α, at a fixed value of experimental fairness will increase the strength.

Up to this point, we have ignored the fact that scientific hypotheses usually refer to theoretical constructs that must be operationalized. This operationalization leads to the empirical hypothesis that is actually tested. Obviously, the result of a statistical test cannot be used to assess a scientific hypothesis if its logical relation to the empirical hypothesis is unknown. This problem can be largely solved by stochastic modeling techniques. The stochastic model comprises a structural and a measurement model. The structural model contains the central statements of the scientific hypothesis, while the measurement model formalizes the relationships between the latent constructs and empirical variables (Erdfelder and Bredenkamp 1994).

Within the context of the deductiviristic theory, some of the problems originally formulated by Campbell and Stanley (1963) need to be reformulated (Bredenkamp 1980). In psychology, scientific hypotheses are interpreted as deterministic or statistical universal hypotheses that refer to an unlimited number of individuals. Each subject belonging to the population the hypothesis refers to is fully representative for the purpose of the experiment. If an experiment is performed with more than one subject, this is done merely to control internal errors by randomization and not for generalization since random samples cannot be taken from an unlimited population. In this view, random assignment serves to justify the use of statistical tests and has nothing to do with generalization (Edgington 1995). Indeed, no generalization of the experimental results to extra-experimental settings is attempted. Experiments serve the single purpose of testing scientific hypotheses.

Practical problems are solved by means of technological statements derived from a scientific hypothesis that has been confirmed experimentally. Technological hypotheses refer to treatments that must be applied in order to have specific, desired effects (Gadenne 1976). The technological statement can be the subject of empirical studies in field settings. On the other hand, controlled tests must be performed in order to test scientific hypotheses. The laboratory offers the best conditions for controlling extraneous variables.

Such a deductiviristic theory of the experiment combines aspects of Popper's falsification theory (1982), Lakatos' methodology of scientific research programs (1970), as well as Neyman and Pearson's theory of statistical hypothesis testing (Neyman 1950). The latter theory permits the (compulsory) control of the statistical error probability, β. The book by Cohen (1988) is very useful for social and behavioral scientists in applying this theory.

Recently, Westmeyer (1998) has criticized the deductiviristic theory of experiments. He doubts the compatibility of the theories of Popper, Lakatos, and Neyman and Pearson. In addition, he demands single-subject experiments to test hypotheses that refer to individuals. This latter demand does not contradict principles of the deductiviristic theory. However, for vast areas of experimental psychology, internally valid single-subject experiments (see *Single-subject Designs: Methodology*) that require repeated measurements cannot be carried out (see Sect. 2.3). In these cases, the logical link between scientific and statistical hypotheses and the resolution of the resulting problems are of great significance.

4. *The Ideal and the Real Experiment*

Textbooks contain rules for conducting the ideal experiment. In the end, it is the aim of all rules to make the experiment distinguishable from other procedures by controls that make it possible to trace variation in the dependent variable back to the independent variable while excluding all extraneous factors from the interpretative system.

Empirical studies seem to violate these requirements. According to Friedman (1967) the ideal experiment is a non-social situation. The view that all extraneous variables are controllable assumes that both the experimenter and the subjects are conceived as nonpersons whose behavior towards one another remains constant. The suspicion that this is not the case can be inferred from the fact that textbooks concerning the experimental method set forth no rules for the experimenter regarding his or her behavior during the pre-experimental phase (greeting the participants, questions regarding the individual). An experimenter can induce experimental errors in many different ways. Friedman (1967) discovered variations in behavior during the pre-experimental phase (how often the experimenter smiles at the subject, looks at the subject, etc.) both between and within experimenters. These differing behaviors can be effective extraneous variables (Friedman 1967).

If the experiment is viewed as a social situation in which the experimenter and subject interact, various extraneous factors on the part of the experimenter and the subject could well come into play. Detailed information concerning this can be found in the book by Rosenthal and Rosnow (1969). On the part of the experimenter, it is the expectations that can influence the results if he or she communicates his or her hypothesis to the subjects in a nonverbal manner (experimenter bias). In addition, the experimenter's

personality traits can influence the results (experimenter effect).

On the other side of the dyadic relation stands the participant, who has taken his or her role as subject either voluntarily or involuntarily, who is searching for the hypotheses of the experiment, who aims at producing results that will confirm or disconfirm the hypothesis, and who is more or less apprehensive about evaluation. All these factors can become effective extraneous variables that endanger the clear interpretation of an experimental result. For this reason, various controls have been suggested to minimize these errors (see *Experimenter and Subject Artifacts: Methodology*).

Does the fact that the experiment is actually a social situation affect internal and external validity in the inductive theory or strength and fairness in the deductiviristic approach? It is obvious that not all-potential disruptiviristic factors in an investigation can be controlled. The brightness of the room in which a social psychology experiment is performed usually remains uncontrolled because the experimenter assumes no influence of brightness on the results. However, the stakes are quite different in experiments on visual perception. Potential sources of experimental errors that cannot be controlled by randomization have to be controlled by use of other techniques if there is a well-founded assumption they could produce artificial results. This also applies to the control of possible artifacts produced by the experimenter or subject. Only the confirmation of general artifact hypotheses (e.g., 'in most studies, the results can be traced back to the experimenter's expectations'; 'in many studies, the results can be traced back to the subject's attempts to produce findings that confirm the hypothesis,' etc.) would make general control of the artifacts necessary. To date, general artifact hypotheses have not been confirmed (e.g., Bredenkamp 1980). Of course, this does not preclude the possibility of artificial results in single investigations.

5. *Objections to the Experiment*

Among all empirical research methods, laboratory experiments have most often been criticized. The controls carried out in the experiment serve the single purpose of deciding about a scientific hypothesis with as little bias as possible. The objection to this is that results obtained under conditions of strict control are artificial and cannot be generalized to extra-experimental settings where many factors that were experimentally controlled now come into play. Section 3 has already discussed this problem. It should be reiterated that laboratory experiments test scientific hypotheses, and that they must employ control techniques in order to avoid false conclusions. They serve the scientific goal of explanation (see *Explanation: Conceptions in the Social Sciences*). The solution of practical problems poses other questions that must be answered within the context of technological research.

A second objection refers to the fact that the real experiment is a social situation that differs significantly from the ideal experiment. It has been argued that for this reason all advantages of the experiment refer to a method that does not exist in reality. However, there is no basis for arguing that social processes generally bias experimental results (see Sect. 4).

A final objection posits that experimentation with humans is simply morally unacceptable. It cannot be reconciled with human dignity because, in the interest of reaching scientific goals, the human is reduced to an 'object.' However, a similar objection can also be raised to nonexperimental investigations. If a ban on empirical research using human subjects is derived from this argument, behavioral and social sciences become impossible. Indeed, in many experimental studies, human dignity does not suffer to any greater extent than in everyday situations in which, for example, children are expected to cope with the effects of educational reforms or changes in social relationships resulting from a move to a different place. When carrying out empirical studies, principles No. 9 (research with human participants) and No. 10 (care and use of animals) of the American Psychological Association (1990) must in all cases be observed. Of particular importance is that 'the investigator protects the participant from physical and mental discomfort, harm, and danger that may arise from the research procedures' (American Psychological Association 1990, p. 395).

The problem cannot be resolved by simply relinquishing experimentation. For example, the application of certain therapeutic procedures whose effectiveness has not been demonstrated by controlled experiments or quasi-experiments appears questionable. Thus, while experiments with scientific goals can pose ethical problems, so too can the failure to perform such experiments. Deciding not to experiment may pose ethical problems of equal weight to deciding to do them and must be justified every bit as clearly and cogently.

See also: Experimental Design: Overview; Laboratory Studies: Historical Perspectives; Medical Experiments: Ethical Aspects

Bibliography

American Psychological Association 1990 Ethical principles of psychologists. *American Psychologist* **45**: 390–5

Bracht G H, Glass G V 1968 The external validity of experiments. *American Educational Research Journal* **5**: 437–74

Bredenkamp J 1969 Experiment und Feldexperiment [Experiment and field experiment]. In: Graumann C F (ed.) *Sozialpsychologie* (= Handbuch der Psychologie, Band 7.1) {Social

psychology (= Handbook of psychology, Vol. 7.1)}. C. J. Hogrefe, Göttingen, Germany
Bredenkamp J 1980 *Theorie und Planung psychologischer Experimente* [*Theory of Psychological Experiments and their Planning*]. Steinkopff, Darmstadt, Germany
Bredenkamp J 1982 Verfahren zur Ermittlung des Typ der statistischen Wechselwirkung [Procedures for the identification of detecting the type of statistical interaction]. *Psychologische Beiträge* **24**: 56–75, 309
Campbell D T, Stanley J C 1963 Experimental and quasi-experimental designs for research on teaching. In: Gage N L (ed.) *Handbook of Research on Teaching*. Rand McNally & Company, Chicago, IL
Carlsmith J M, Ellsworth P C, Aronson E 1976 *Methods of Research in Social Psychology*. Addison-Wesley Publishing Company, Reading, MA
Chapin F S 1947 *Experimental Designs in Sociological Research*. Harper, New York and London
Cohen J 1988 *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. L. Erlbaum Associates, Hillsdale, NJ
Cook T D, Campbell D T 1979 *Quasi-Experimentation—Design & Analysis for Field Settings*. Rand McNally College Pub. Co., Chicago
Cox D R 1961 Design of experiments: The control of error. *Journal of the Royal Statistical Society Series A* **124**: 44–8
Edgington E S 1995 *Randomization Tests*, 3rd rev. and expanded edn. M Dekker, New York
Erdfelder E, Bredenkamp J 1994 Hypothesenpruefung [Hypothesis testing]. In: Herrmann T, Tack W H (eds.) *Methodologische Grundlagen der Psychologie* [*Methodological Foundations of Psychology*] (= Enzyklopaedie der Psychologie, Band B I 1 [Encyclopedia of psychology, Vol. B I 1]. Hogrefe, Goettingen, Germany
Feldt L S 1958 A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika* **23**: 335–53
Friedman N 1967 *The Social Nature of Psychological Research: The Psychological Experiment as a Social Interaction*. Basic Books, New York
Gadenne V 1976 *Die Gueltigkeit psychologischer Untersuchungen* (*The Validity of Psychological Investigations*). Kohlhammer, Stuttgart, Germany
Gadenne V 1984 *Theorie und Erfahrung in der psychologischen Forschung* (*Theory and Experience in Psychological Research*). JCB Mohr, Tuebingen, Germany
Hager W, Westermann R 1983 Planung und Auswertung von Experimenten (Planning and analyzing experiments). In: Bredenkamp J, Feger H (eds.) *Hypothesenpruefung* (*Hypothesis Testing*) = Enzyklopaedie der Psychologie, Band BI 5 (Encyclopedia of Psychology, Vol. B I 5). Hogrefe, Goettingen, Germany
Keppel G 1973 *Design and Analysis: A Researcher's Handbook*. Prentice-Hall, Englewood Cliffs, NJ
Lakatos I 1970 Falsification and the methodology of scientific research programs. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK
Neyman J 1950 *First Course in Probability and Statistics*. Holt, New York
Popper K R 1982 *Logik der Forschung. Mohr, Tuebingen* (1959 *The Logic of Scientific Discovery*. Hutchinson & Co., London)
Rosenthal R, Rosnow R L (eds.) 1969 *Artifact in Behavioral Research*. Academic Press, New York
Westmeyer H 1998 Psychologische Methoden und soziale Konventionen (Psychological methods and social conventions). In: Klauer K C, Westmeyer H (eds.) *Psychologische Methoden und soziale Prozesse* (*Psychological Methods and Social processes*). Pabst, Lengerich, Germany

J. Bredenkamp

Laboratory Studies: Historical Perspectives

1. *The Origin and Meaning of 'Laboratory Studies'*

The term 'laboratory studies' dates back to the late 1970s, when it began to be applied to studies of the production of scientific knowledge at its 'natural' site, the scientific laboratory (Latour and Woolgar 1979, Knorr Cetina 1981, Lynch 1985, Traweek 1988). Not all natural scientific knowledge is produced in laboratories, but the laboratory nonetheless epitomizes modern science and knowledge. It is what the factory was to industrialization, the place where particular forces of production are brought together, organized, and unleashed. Laboratories do not produce industrial commodities, but, according to laboratory studies, they do create products, the objects and facts of science. A first characteristic of the original laboratory studies was that they refocused questions of knowledge on the context of knowledge production, moving them away from questions of the validity and representational correctness of scientific findings. Philosophers of science, until then the main authorities on matters of scientific procedure and content, had largely neglected the context of knowledge production, which they called the context of discovery. Historians often defined issues of scientific content as questions pertaining to the history of ideas, detached from local settings. To be sure, both groups also enriched their accounts with studies of crucial experiments, but experiments, as we shall see, are not laboratories. The whole process of knowledge production in the contemporary fact factory of the natural sciences, and the role of the fact factory itself, were, up to the late 1970s, untrodden territory in the social and behavioral sciences. One reason for the interest generated by laboratory studies lies in the fact that they 'conquered' this new territory, that they unlocked a hidden and central chamber in our understanding of knowledge and science. It needs to be recalled that the development of scientific method and the ensuing advances in knowledge are often associated with the transition from traditional to modern societies, carrying much of the burden of explaining the dynamics and the progress of industrialization. Knowledge and

technology also lie at the core of the current trend to global systems and transformation to a 'knowledge society.' Any perspective that contributes to the understanding of these forces will be of interest in the larger context of these developments. A second characteristic of laboratory studies is that they grounded their analyses empirically, using a mix of 'molecular' level methods, most importantly observation and ethnography, ethnomethodological techniques, and discourse analysis. Since much of the knowledge that experts have of their own ways of working is implicit, interview methodologies are not conducive to clarifying these. This empirical approach to knowledge environments and their embedded processes stands in contrast to postmodernist and transformation theory perspectives and their theoretical assessments and critiques of science. It also contrasts with much of the knowledge society literature (Bell 1973), which tends to render knowledge and technology the independent variable—for example, it treats knowledge as the engine of economic change, responsible for such widespread changes as shifts in the division of labor, the development of specialized occupations, the emergence of new enterprises, and sustained growth. In these accounts, 'knowledge' is sometimes conceptualized to fit in with long-standing beliefs about science (an example is Bell's attempt to explicate knowledge in terms of theory, 1973, p. 44), but it is in effect the last thing to be explained and has no reality outside an analyst's models (Stehr 1994).

If the interest in opening the black box of scientific practice is what distinguishes laboratory studies, then the interest in not restricting what is to be looked at inside the black box distinguishes them from previous studies of scientific experiments. The third characteristic of laboratory studies is their inclusiveness with respect to all practices and events implicated in knowledge processes. Experiments have been largely defined methodologically in the history and philosophy of science, where they were investigated; notions like the testing of theories, experimental design, blind and double-blind procedure, control group, factor isolation, and replication are linked to experiments. Students of laboratories focus on a space rather than on particular experiments. This brings to the fore the full spectrum of activities going on in this space and involved in the production of knowledge. Studies of laboratories have shown that scientific objects are not only 'technically' created in laboratories but also symbolically and politically construed. For example, they are construed through the literary techniques of persuasion that one finds embodied in scientific papers, or through the political stratagems of scientists in forming alliances and mobilizing resources. An implication of this has been the awareness that, in reaching its goals, research 'intervenes' not only in the natural world but also—and deeply—in the social world. Another implication is that the products of science have come to be seen as cultural entities rather than as natural givens 'discovered' by science. If the practices observed in laboratories are cultural in the sense that they cannot be reduced to the application of methodological rules, the 'facts' that result from these practices must also be seen as having been shaped by culture.

2. *The Laboratory as a Theoretical Notion*

The significance of the notion of a laboratory lies not only in the fact that it has opened up this field of investigation and offered a cultural framework for 'plowing' this field. It lies also in the fact that the 'laboratory' itself has become a theoretical notion in our understanding of knowledge. Accordingly, the laboratory is not only a 'house' for the conduct of experiments or the physical setting in which knowledge processes take place, but the locus of mechanisms and processes which can be regarded as necessary for the 'success' of science and knowledge.

2.1 *Laboratories as Reconfigured Environments*

Characteristically, these mechanisms and processes are nonmethodological and mundane. One hallmark of the mechanisms and processes with which the notion of a laboratory has been associated is that they imply, to use Merleau-Ponty's terminology, a reconfiguration of the system of 'self-others-things' ('Moi-Autrui-les choses,' 1945, p. 69), and a reconfiguration of the *phenomenal field* in which experience is acquired in science. The system of 'self-others-things' is not the objective world independent of human actors or the inner world of subjective impressions, but the world-experienced-by or the world-related-to agents. Laboratory studies suggest that the laboratory is a means of changing the world-related-to-agents in ways which allow scientists and other knowledge workers to capitalize on human constraints and sociocultural restrictions. The improvement derives from the *malleability* of knowledge objects. For example, there are at least three features of natural objects which a laboratory science does not need to accommodate: first, it does not need to put up with the object *as it is*; it can substitute all of its less literal or partial versions. Second, it does not need to accommodate the natural object *where it is*, anchored in a natural environment; laboratory sciences bring objects '*home*' and manipulate them 'on their own terms,' in the laboratory. Third, a laboratory science does not need to accommodate an event *when it happens*; it does not need to put up with natural cycles of occurrence, but can make them happen frequently enough for continuous study.

Laboratories, then, allow for some 'homing in' on natural processes; the processes are 'brought home' and made subject to the conditions of the local social order. The power of the laboratory (but, of course,

also its restriction) resides precisely in its 'enculturation' of natural objects. The laboratory subjects natural conditions to a 'social overhaul' and derives epistemic effects from the new situation. Social forms, to be sure, are also subject to 'overhauls' in laboratory processes. Questions raised in this context are, who 'is' the epistemic subject, and how is this subject configured in a particular setting? The epistemic subject is the bearer of knowledge, the unit that is productive of scientific findings. This unit cannot be automatically assumed to be the individual scientist, nor even a division of labor among laboratory leader, postdoc, grad student, and technician. In some areas, for example in experimental high-energy physics, this 'unit' now encompasses collectives of considerable size (up to 2,000 physicists), a large and complex machine (a detector), and a level of distributed cognition in which machine analysis and human analysis merge into a single, continuing, circulating discourse. The productive role of the detector in this joint discourse is recognized by participants, who attribute to it the sole epistemic authority to observe and describe the particle processes instigated in its internal environment. But the productive role of discourse as a platform for integrating machine analysis and human analysis and for creating a collective consciousness that is at the same time material and human also needs to be recognized; it is a circulating consciousness that appears to replace the individual as the bearer of knowledge. More generally speaking, purely human orders appear to be transformed in laboratory processes into object-relation regimes in which the objects are machine complexes, organisms, and the like. Laboratory studies have approached these through network notions, where it is assumed that the creation of links between the nodes in a network of human and nonhuman entities is also what stabilizes and universalizes scientific outcomes (e.g., Callon 1986, Latour 1987). They have also looked at these regimes in terms of their contribution to understanding a more general shift toward object relations in contemporary society—a shift fueled and sustained by an enormous expansion of object worlds within the social world and by an erosion and 'emptying out' of primordial social relations. Objects, in this approach, are thought to be the winners of the relationship risks that are manifest in the problems of married and family life and that many analysts find inherent in contemporary human relations (Coleman 1993, Knorr Cetina 2000).

Laboratories can be distinguished from organizations along similar lines. Organizational sociology is relevant to laboratory studies since most laboratories are or are part of complex organizations. But the concept of organization in the social sciences is essentially a concept of the coordination of human groups. Laboratory studies, on the other hand, have tended to see governance structure as inseparable from patterns of coordination that are obtained between persons and objects of investigation. The sociology of organizations pays attention to organizational 'subdivisions,' the sort of formal units and subgroups into which organizations can be decomposed. The subdivisions of interest to the laboratory approach are alternate knowledge environments that are not hierarchically but 'ontologically' related to the larger framework of a lab: the difference between alternate knowledge environments derives from the line of reality developed and 'farmed.' One example of such an alternate reality and of a laboratory within the laboratory is a simulation environment. In experimental high-energy physics, simulation constitutes a parallel world in which nearly all aspects of the physical detector, the particle clashes, and the physics analysis are recreated and worked out. Such 'detour' environments need not be of a different nature, for example, symbolic rather than physical, as in simulation. They may also emerge from particular natural forms that are articulated further in the laboratory. For example, in molecular biology the fruit fly and the slime mold have become worlds of investigation in their own right and have evolved as model realities separated off from the rest of the biological universe. The emergence of alternate realities has brought with it the emergence of special knowledges and new specialists who are the ones dealing with the environment. In the case of the fruit fly, these include a 'Drosophila elite' of several generations bound together by formative psychosocial relationships; they are distinguishable from other scientists who pass through fly rooms during their training through their careers, their gender, and the problems they work on (Kohler 1994, 92ff.).

2.2 *The Constructionism of Laboratory Studies*

A second feature associated with laboratories pertains more to knowledge processes within them, which are seen to be constructive rather than merely descriptive. Constructionism was one of the first and most important results of laboratory studies, and the analytic answer these studies gave to the microprocesses they observed in real-time episodes of scientific work (see *Technology*, *Social Construction of*). Constructionism has been taken as a principle of investigation in laboratory studies: one asks how much of the success of a knowledge claim can be explained in terms of the relevant productive processes before this success is attributed to the representational content of the claim. Given that most knowledge claims that are considered true will eventually become false, it is the representational hypothesis which is thought to be implausible by students of laboratories. The constructionist thesis has also been taken as a substantive assertion about natural reality being humanly constituted in a wider sense. Science, in this view, secretes an unending stream of entities that make up 'the world'—it holds, one might say, a monopoly on specifying and delimiting the material world in which

we live. What exists before scientifically delimited objects are culturally delimited objects, those which humans pick out, shape, and deal with in everyday life. In both cases, the 'material world' flows from human designations. Laboratory constructionists have also granted the existence of an unspecified material world beyond cultural and scientific designation; but this presupposed world has to be distinguished from concrete objects and forces which can be said to exist only after they have somehow been designated and distinguished, built into our accounts, and inserted into daily practice. In that sense constructionists subscribe to an ontology that anchors existence in social and cultural life, not outside it. While the existence of the material world as a physical entity independent of us is granted on principle, the existence of specific objects identified in terms of their character cannot so be granted.

Constructionist thinking in general existed before laboratory studies, but it had not been extended to science. First, as already implied, constructionist metaphors take on a much stronger flavor when they are applied to natural reality, to knowledge and science. To say that kinship systems or gender relations are social conventions and cannot be attributed to 'human nature' is widely accepted—what undergirds this understanding is, among other things, the cultural variation of these systems. But to make the same claim for scientific facts is a different matter, raising the specter of relativism and skepticism. Constructionism, though, should not be assimilated with these doctrines. It is best understood as the attempt to carve out a middle ground between relativism and realism, or as an attempt to point at space beyond the two poles. This space is the one where the complex chains of laboratory transformations of natural and social objects take place, each link involving alternative options, perspectives, interpretations, negotiations, strategies, conflicts, resources, obstacles, and surprises (e.g., Latour and Woolgar 1979, Knorr Cetina 1981, Lynch 1985, Latour 1987)—as well as frequently social, economic, and political groups outside science (e.g., Fujimura 1996, Rabinow 1996). Only if this space remains invisible and the beginning of an experimental series is short-circuited with its end, do realist interpretations become plausible, as do relativist interpretations of knowledge—which interpretation results depends on whether one cuts human work or material recalcitrance and resistance out of the picture. Constructionism, on the other hand, is the view that the space in which a particular intermesh between human work and material recalcitrance is forged holds the key to understanding knowledge.

3. *Knowledge Cultures*

More recent laboratory studies have supplemented the interactional and writing-centered articulation of the constructive character of knowledge processes with a measure of aggregate-level analyses, and they have also unfolded further the material side of laboratories. At the same time, a more narrowly defined interactional model of knowledge production has been expanded in several directions, for example, in the direction of taking into account contextual cultures. The notion of culture had been used in the early laboratory studies to describe the intent of the work; it was also indicative of the methodology used in these studies (ethnography) and of the phenomenon that laboratories appeared to involve whole cultural worlds, that is a large variety of factors not reducible to sociological variables. Nonetheless, there is a sense in which the early laboratory studies lacked a cultural perspective. They limited themselves to one knowledge area and, as a consequence, ignored the cultural diversity of knowledge. They were concerned, one might say, with knowledge rather than with knowledges. With one exception (Traweek 1988), they also investigated laboratories in only one country, the USA; the studies of Latour and Woolgar (1979), Knorr Cetina (1981), and Lynch (1985) were all done in California. Thus these studies also ignored the possible influence of national cultures. Traweek (1988) compared the Stanford Linear Accelerator laboratory with its equivalent in Japan (the KEK near Tokyo), contrasting the two settings with respect to laboratory organization, detector design and building, leadership style, and models for good working conditions in the science she was studying—high-energy physics.

This study also addressed the aspect of the making of a male physics culture, in addition to analyzing specificities having to do with the way institutions are set up in different countries; but it did not extend the analysis to questions of whether and how knowledge cultures differ between sciences. The assumption long made of the unity of science may in fact not be warranted, and may have been sustainable only in the face of a lack of comparative investigations of knowledge. A recent comparative study of high-energy physics and molecular genetics locates differences between these fields in the understanding of measurement and experiment, in the general epistemological approach of a field, in the ontology of objects and machines, and in the conception of epistemic subjects (Knorr Cetina 1999). For example, the study finds a 'liminal' approach in which the regions of positive, phenomenal knowledge are studied and narrowed down through investigating the errors and limits of knowing in place in one science but not in the other. Liminal epistemologies forge a coalition with the evil that bars knowledge by turning these barriers into a principle of knowing. 'Referent' epistemologies, on the other hand, enhance the role of natural objects in selecting experimental results, and are committed to the use of experiential registers as strategies of knowing. Such 'epistemic cultures' may be contextualized by investigations of 'epistementalities' that prevail in

the larger society—beliefs and attitudes about the correct distribution of knowledge, the rights of access to it, the requirement and meaning of transparency, etc. The importance of further research on these topics can be linked back to the idea of a knowledge society as one where expert systems penetrate all areas of contemporary life. In such societies, knowledge cultures may in fact become what territorial cultures had been to industrial society: defining environments which enable and shape social and economic development.

4. Types of Laboratories

A second expansion of laboratory studies beyond the inclusions of knowledge cultures concerns the fact that diverse types of laboratories are now recognized to exist and are included in the analysis. The spaces of knowledge vary historically, ranging from the 'houses of experiment' in seventeenth-century England investigated by historians (Shapin 1988) to the virtual collaboratories stretching across continents in astronomy today. The locally situated benchwork laboratories of molecular biology have to be distinguished from laboratory 'centers' that tie a whole field to a place, and those from interlaboratory fields of interconnected but functionally differentiated stations of work.

4.1 Benchwork Laboratories

Benchwork laboratories correspond to the most common idea of what a laboratory is: a repository of processing materials and instruments in rooms in which research tasks are performed continuously and simultaneously, typically by more than one researcher and by technicians. Such laboratories are specific to particular sciences such as molecular biology (e.g., Lynch 1985, Fujimura 1987, Jordan and Lynch 1992). They can be found at any university but they are also the core units of large research institutions. One characteristic of such laboratories is a two-tier organization differentiating between the laboratory level and the experiment level of activities. On the first level, such laboratories become focused on and identified with their leaders, who spend much time representing and promoting 'their' laboratory, and recruiting resources and personnel for it. On the level of scientists, the laboratory fragments into projects associated with individual researchers around which other components are assembled. Together, these create small lifeworlds comprising— besides a scientist—materials, instruments, bench space, and help from technicians and students. A second characteristic of such laboratories is the embodied character of the work and the intervening technology of experimentation. Intervening in a lab setting can be contrasted with the technology of experimentation in the social sciences, which is based on a theory of nonintervention: in blind and double-blind designs, researchers attempt to eradicate the very possibility that the outcome will be influenced by the scientists involved.

In benchwork laboratories, on the other hand, this tendency is reversed: one of their most notable features is that the objects of experimental work are subject to almost any imaginable procedural intrusion and physical manipulation. A third characteristic of benchwork laboratories is that they encapsulate within themselves a continued traffic of substances, materials, equipment, and observations that spills beyond laboratory boundaries and ties single laboratories into interlaboratory fields.

4.2 Center Laboratories

With the second type of laboratory, such fields may become more strongly structured in terms of centers and peripheries and semiperipheral units. One identifying detail of center laboratories appears to be that resources important to a field become centered in a particular location and space (see also Hilgartner 1995). One reason for this may be the sheer size and expense of the equipment needed. High-energy physics provides an example; in this area, there have long been regional and continental centers explicitly motivated in terms of the resources needed to build the accelerators and detectors that can provide and deal with the energy levels of particular forefront experiments. Given the further rise of energy levels, there is now a global center; perhaps for the first time in the history of science there is one laboratory worldwide (at the turn of the millennium the CERN in Geneva, Switzerland) in which resources and expertise from Europe, the USA, and Asia have been concentrated. This laboratory provides access to groups from all countries and continents equipped to participate in the respective experiments. Such centers are line items in national budgets; they are maintained through regular contributions dedicated to the lab, which is itself the size of a village (for a history of big science, see Galison and Hevly 1992, Galison 1997). Global center laboratories of this sort very nearly absorb within them a whole field, with as yet largely unstudied consequences for epistemic questions such as how consensus is formed about scientific outcomes when fields are interiorized in laboratories.

If their structural position in a field and the interiorization of scientific communities are characteristic features of some center laboratories, the role of experiments in them is another. In the benchwork laboratories considered, experiments have little ontological identity. They simply dissolve into ongoing streams of experimental work that is also continuous with the service work provided for the laboratory. In high-energy physics, on the other hand, experiments are highly visible units which dispose over their own financial resources, which are separate from those of

the larger laboratory. As a consequence, laboratories and experiments are technically, organizationally, and socially divorced from each other. Technically, laboratories build, maintain, and run accelerators and colliders which provide particle clashes that experiments investigate. Experiments, on the other hand, build, maintain, and run the detectors in which the clashes leave their mark. Organizationally, experiments are run by large 'collaborations,' comprising, at the turn of the millennium, up to 2,000 physicists from up to 200 physics departments and laboratories located worldwide, who join up for the purpose of conducting the experiment. Laboratories, on the other hand, dispose over a permanent staff of accelerator physicists, computer specialists, technicians, and others, who conceive of and tend to the scientific and technical infrastructure of the laboratory. Researchers on both sides of the divide depend on one another but often know little of one another. Experiments become relatively closed, total units that may also decide, as independent collaborations and collective actors, to migrate between laboratories (to submit an experimental proposal to another laboratory).

4.3 Network Laboratories and Distributed Laboratories

Though the specific configuration of such center laboratories may be historically distinctive, the idea of a central big science laboratory is not distinctive. An early example is Tycho Brahe's observatory, begun in 1576, where a dozen assistants, along with mechanics and others, collected great quantities of data from which a 'computing division' derived parameters of planetary motion (Heilbron 1992). The Danish king financed this empire, and cheap labor from the surrounding inhabitants took care of its construction, tended Tycho Brahe's garden, and raised his crops. A second example may have been the Society of Jesus, which maintained seminars and collections and which possessed a network of scientific collaborators with whom they conducted extensive correspondence. Perhaps surprisingly, similar 'network' and 'platform' laboratories have reappeared today in a variety of settings. They can be seen as a third type of laboratory, which may be distinguished from bench laboratories and center laboratories by two features: the laboratory is no longer situated in a particular place but extends to and involves a number of parties and locations, and what happens between these locations is central to understanding the outcomes of the lab.

One example of network laboratories is the electronically mediated 'virtual laboratory' exemplified by work in mathematics, theoretical physics, and astrophysics, where scientists use electronic linkages and e-mail communication to join their research efforts on symbolic objects. The notion of a laboratory here captures the phenomenon that the electronic space becomes a *workbench* for participants; a space in which transitory research objects may be stored, particular resources can be found, and dispersed materials can become interactively related and integrated. Center laboratories, when they involve ter-ritorially dispersed collaborating groups, also intermittently expand into such e-mail co-laboratories, but the groups also frequently meet at the center, where they spend prolonged periods of time tending to machines and where a local workforce is maintained. Special cases of such virtual laboratories are beginning to appear in the humanities and social sciences, for example in history, where digitization projects erect interactive platforms for historical research. These platforms make available archival materials that are categorized, cross-linked, and presented with navigation systems for research uses. Such platforms are not simply electronic versions of alphabetically stored books and archival materials. Through carefully thought out and theorized category systems and through the interlinkages they allow for, new questions open up, comparisons become possible, and contextual information can be called forth and integrated with other pieces of information. In addition, the platforms offer possibilities of virtual manipulation, for example, the virtual reconstruction of historical sites or experiments that cannot easily be reconstructed in naturalistic ways. Interestingly, the history of laboratories may be elucidated further through some of these platforms, which assemble information on sites of knowledge production in previous centuries.

5. The Laboratory Perspective

A third extension of laboratory studies takes its lead from family resemblances between laboratories and other physical and virtual spaces. In this case the notion of a laboratory, and the concepts emerging from laboratory studies, are transferred to areas that are not literally laboratories but can be seen as spaces of knowledge. When the above mentioned platforms make available material on historical research sites they include among the sites clinics theaters, lecture rooms, field stations, and even farms—places into which knowledge crossed from laboratories or from whence practices moved into them. Such sites can be studied using a laboratory perspective. This perspective also simply brings into view the phenomenon that matters with which students of laboratories have concerned themselves, for example, the construction of facts, are actually ongoing occurrences in a variety of settings. This is the sense in which the industrial factory itself begins to resemble a laboratory as a site for invention and intervention in which new realities are created (Miller and O'Leary 1994).

We can illustrate this use of the laboratory approach by means of two examples, one from transsexual

research and one pertaining to the factory. In the first case the new sex of a person who desires to have a different sex from the one he or she is born with is seen as a 'fact' that is being constructed in a laboratory made up of the different stations of the treatment program that transsexuals undergo (Hirschauer 1991). The laboratory approach sheds light on the multiplicity of stations that make up the lab, on the constructive and transformative work contributed by the stations, and on the heterogeneity of the frameworks of knowledge involved. In the second case, that of the factory, Japan is brought to Illinois, as it were, in the attempt to reconfigure the American factory in the image of global factory modernization, by redesigning shop floors, recalculating new spatial orderings of production, and molding the worker according to the ideal of a 'new economic citizenship' for plant personnel. Here the factory itself becomes a laboratory for acting upon its own assemblage of locales and internal relations and for refashioning the person who participates in the assemblage (Miller and O'Leary 1994).

The notion that we now live in a knowledge society is relevant here once again. If this view is correct, the range of applications of laboratory studies as a particular perspective along the lines just indicated is large. In a knowledge society, the sites where knowledge is produced and applied multiply, creating an open series. These sites are not limited to science, but rather move outside science. As the above examples also indicate, the laboratory approach is not concerned with knowledge processes as defined in terms of an end result ('correct knowledge') but with what could be called the epistemic attributes embedded in, and, in a knowledge society, conferred upon, many processes—exemplified by institutional reality construction, by object relations, and by the reconfiguration of human subjects. The three extensions of the laboratory approach described here indicate the transformation of what was originally an approach to particular natural sciences into a perspective applicable to such broader questions.

See also: Experiment, in Science and Technology Studies; Experimentation in Psychology, History of; Laboratory Experiment: Methodology; Scientific Instrumentation, History and Sociology of

Bibliography

Bell D 1973 *The Coming of Post industrial Society: A Venture in Social Forecasting*. Basic Books, New York

Callon M 1986 Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. In: Law J (ed.) *Power, Action and Belief: A New Sociology of Knowledge?* Routledge and Kegan Paul, London

Coleman J 1993 The rational reconstruction of society: 1992 Presidential Address. *American Sociological Review* **58**: 1–15

Fujimura J H 1987 Constructing 'do-able' problems in cancer research: Articulating alignment. *Social Studies of Science* **17**: 257–93

Fujimura J H 1996 *Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer*. Harvard University Press, Cambridge, MA

Galison P 1997 *Image and Logic*. The University of Chicago Press, Chicago

Galison P, Hevly B (eds.) 1992 *Big Science: The Growth of Large-Scale Research*. Stanford University Press, Stanford, CA

Heilbron J L 1992 Creativity and big science. *Physics Today* **45**(11): 42–7

Hilgartner S 1995 The Human Genome Project. In: Jasanoff S, Markle G E, Petersen J C, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA

Hirschauer S 1991 The manufacture of bodies in surgery. *Social Studies of Science* **21**(2): 279–319

Jordan K, Lynch M 1992 The sociology of a genetic engineering technique: Ritual and rationality in the performance of the plasmid prep. In: Clarke A, Fujimura J (eds.) *The Right Tools for the Job: At Work in Twentieth-century Life Sciences*. Princeton University Press, Princeton, NJ

Knorr Cetina K 1981 *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Pergamon, Oxford, UK

Knorr Cetina K 1999 *Epistemic Cultures. How the Sciences Make Knowledge*. Harvard University Press, Cambridge, MA

Knorr Cetina K 2000 Postsocial theory. In: Ritzer G, Smart B (eds.) *Handbook of Social Theory*. Sage, London

Kohler R 1994 *Lords of the Fly*. University of Chicago Press, Chicago

Latour B 1987 *Science in Action*. Harvard University Press, Cambridge, MA

Latour B, Woolgar S 1979 *Laboratory Life: The Social Construction of Scientific Facts*. Sage, Beverly Hills, CA

Lynch M 1985 *Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory*. Routledge and Kegan Paul, London

Merleau-Ponty M 1945 *Phénoménologie de la perception*. Gallimard, Paris

Miller P, O'Leary T 1994 The factory as laboratory. In: Power M (ed.) *Accounting and Science*. Cambridge University Press, Cambridge, UK

Rabinow P 1996 *Making PCR: A Story of Biotechnology*. University of Chicago Press, Chicago

Shapin S 1988 The house of experiment in seventeenth-century England. *Isis* **79**: 373–404

Stehr N 1994 *Knowledge Societies*. Sage, London

Traweek S 1988 *Beamtimes and Lifetimes: The World of High Energy Physics*. Harvard University Press, Cambridge, MA

K. Knorr Cetina

Laicization, History of

During the second half of the twentieth century, sociologists often used the notion of *secularization*. This notion was basically considered as the decrease of religion's influence on modern society. Nowadays, not

only can one notice that the decrease is not complete but one can also criticize that the notion of *secularization* is too broad. For a better understanding, one might want to distinguish a process of secularization from a process of laicization.

When the process of secularization is predominant, the relative decline of religion's influence takes place in parallel to other evolutions, with no major tensions between religious and political or other social forces. Certainly, religious changes, as well as economic, political, etc. changes, may produce internal tensions. But, when all is said and done, the triumphant forces participate in the same social dynamic. Therefore, there is no important clash between the changes within the religious sphere and other social changes. For that matter, Scandinavian countries are a typical example of secularization. The switch to Lutheran Protestantism, particularly linked to the Bible's translation, has favored the development of a national culture. Theological arguments ('theology of the two reigns,' absence of essential differences between clerics and laymen, rejection of the papacy, etc.) have allowed the autonomy of the nation-state towards religion not to provoke important conflicts. Moreover, a joint and progressive democratization of State and Church has taken place. In the end, for example, a national Danish church emerged (*Folkekirken*)—this continues to be the symbol of the identity of nations that have been largely secularized.

The process of laicization differs: the tensions between various social forces take on the aspect of an open conflict where the religious becomes a politico-cultural stake. The symbolization of the national culture, among others, is controversial. Either the State imposes religion to society (*clericalism*) or refuses that religion to continue to be the symbol of national identity (*laicization*). The French case is the main example of such a process.

1. Revolution in France and After: the First Threshold of Laicization

In France, the roots of this conflict are one of the results of the Wars of Religion (sixteenth century) as well as the persecution of Protestants and dissenting Catholics (Jansenists) (end of the seventeenth century and eighteenth century). As a reaction to this, a certain anti-religious dimension appears in the French philosophy of Enlightenment that had permeated a part of the nobility and bourgeoisie. This spirit differed from the English Enlightenment (or the German *Aufklärung*). The latter appeared as a confrontation within a diversified Protestantism, wanting an 'enlightened religion,' the former directly attacked the Roman Catholic Church, '*l'infâme*' as wrote Voltaire, who was a deist.

In 1789, religion pervaded French society. The monolith Roman Catholic faith was made compulsory for the French people. Communion at Easter and confession to a priest were required from everyone. The clergy was the first of the three 'Orders' of the kingdom. The Catholic Church was an inclusive institution exercising power over other institutions, which often only existed in embryonic states. Registration of births, deaths, and marriages, medical and social welfare and education were in the hands of or controlled by the clergy. The monarchy had a religious justification with the theory of the divine right of kings. The king possessed a politico-religious power and the defense of Roman Catholicism was one of the monarchy's basic duties. But gradually the monarchy also established a large autonomous political and administrative structure. A Gallican movement developed the idea of a French Catholic Church, which would be autonomous from the Pope and protected but also partly controlled by the State.

The first collapse came with Article X of the *Declaration of Human Rights*. It proclaimed the principle of religious freedom (August 1789), which was progressively applied to minorities. The Constitution divested the monarchy from its religious character, which was considered as a contract and entered into the nation. The fact that the King did not respect this contract led to the conflicting establishment of the Republic (August–September 1792). The state no longer had religious foundations.

As early as 1789, the Revolution celebrated the new order. First, Roman Catholicism lent a religious foundation to these celebrations. But progressively, a breach between the Catholic Church and the Revolution appeared. There were three reasons for this: the proclamation of religious freedom, the bill of the Civil Constitution of the Clergy (1790)—which accentuated gallicanism and decided that priests and bishops should be elected as political deputies—and the establishment of civil registers for births, marriages, and deaths. The 'religion of the fathers' would not be more capable of celebrating the new society, and the antirevolutionary propaganda became more and more associated with Roman Catholicism. By the year II (autumn of 1793), the institution of revolutionary cults went with a violent struggle (period of the Terror) directed against revealed religions.

Napoleon Bonaparte inherited a chaotic situation. He thought it would be possible to restore a Roman Catholic Church and maintain some of the Revolution's great achievements. After delicate negotiations with the Pope, a Concordat was signed (July 1801). The Roman Catholic Church was proclaimed 'religion of the great majority of the French' but did not have the status of a state religion; the 'organic Articles' recognized two Protestant Churches (Lutheran and Reformed) (1802) and the Jewish religion (1807). In 1804, the Civil Code made no reference to religion.

It is possible to evaluate this situation, using the abstract notion of the first threshold of laicization

(constructed on the basis of Weber's ideal-type), which was the permanent result of revolutionary upsets and is marked by three characteristics.

(a) *Institutional fragmentation*: Roman Catholicism was no longer an inclusive institution. Religious activities were clearly distinguished from 'profane' activities. Educational and health needs assumed gradual autonomy, related to religious needs, and were provided for by specific institutions that were going to develop progressively.

(b) *Recognition of legitimacy*. Revolution did not destroy religious needs that continued to exist objectively. The service provided by religion was a public service and the state paid ministers of recognized religions. Religions were socially recognized as a foundation of social morality.

(c) *Religious pluralism*. The state recognized four religions, it protected and controlled them for they could satisfy the religious needs of their followers and develop moral values. Other religions were more or less tolerated. Relative freedom was even given to those who decided to do without 'the help of religion.'

The French Revolution did not succeed in establishing a durable system of separation of Church and State, whereas the American Revolution did.

In the United States, the Declaration of Independence stipulates that all human beings *are created equal*, *that they are endowed by their Creator with certain unalienable Rights*, which is not the case in the French Declaration of Rights. Why? Because in the United States, religious pluralism allows the dissociation of God from every church.

In France, religious monopoly would let the Roman Catholic Church become the owner of human rights, being the mediator between God and humans. This creates a structural difference in the long run. The American situation may have specific characteristics linked to the master function of religion: social space for cultural pluralism. Stephen Warner (1993) indicates the main differences:

(a) Religion is constitutive for some American subcultures.

(b) Religion in America has historically promoted the formation of associations among mobile people.

(c) Religion in America serves as a refuge of free association and autonomous identity, a 'free social space.'

(d) The second generation of immigrants often transmutes ethnicity into religion because it allows immigrants to assimilate and conserve their identity.

2. *From the First Threshold to the Second Threshold of Laicization*

In France, the compromise of the first threshold of laicization delineates a logic that dominated for almost a century. But the situation that it describes was not static. At first, this was due to the fact that at the beginning of the nineteenth century, a number of aspects of the first threshold were not entirely realized: other institutions were not sufficiently developed to become totally autonomous. Then, especially during the second half of the nineteenth century, there was a growing conflict between the partisans of revolutionary values and those who thought of France mainly as a Roman Catholic nation.

The first conception won. During the 1880s, the Republicans introduced a religiously neutral educational system, which included *a morale laique*. At the end of the century, a new crisis arose from the Dreyfus case. The impression prevailed that the Republic was threatened by the *clerical peril*. The important law of July 1901 on voluntary societies excluded religious congregations from the right to free associations. A new law voted in July 1904 prohibited members of religious orders from teaching. The separation of Church and State appeared to be inevitable and took place in a climate of violent confrontation.

But two models of separation were possible. One model was antireligious and combative. Such was the bill of Emile Combes (Prime Minister). However, it was refused, especially by Jean Jaurès, the socialist leader, who hoped to achieve a law that would bring peace and eventually make it possible to fight social problems. A liberal model prevailed and ensured the freedom of conscience, guaranteed the free exercise of religion, and respected the self-organization of each church (art. 1 and 4) even though it neither 'recognized' nor paid subsidies to any religion (art. 2). This marks the end of the system of 'recognized cults.'

Even in liberal dominance, this model of separation completed a religiously neutral educational system and found a new logic. This can be seen as the second threshold of laicization. As for the first threshold, there are three characteristics:

(a) *Institutional dissociation*: religion was entirely optional because it was no longer considered as the institutional structure of society but rather as a mode of free association.

(b) *Social neutral about religious legitimacy*: religious needs having become a private matter and morality officially *laique*. The question as to the usefulness of religion for society was no longer publicly relevant. (In a logic of the conflict, the opponents to Roman Catholicism asserted that religion and religious morality were dangerous for the Republic and its ideals. This point was a compromise.)

(c) *Freedom of religion*: various religious societies belonged to the sphere of public liberties. The state guaranteed each citizen freedom of conscience and allowed each one to meet with religious societies or associations.

The separation of 1905 was marked by a twofold aspect: the victory of the *laique* camp in the hard and long conflict of the *two Frances* but also the result of a conflict within the Republican party. An implicit

covenant with opponents was searched for in order to reach peace in the long run so that complete victory in the short term would predominate. The conflict did not die immediately: it vanished progressively because religion ceased to be a political problem.

The *Résistance* had united unbelievers and believers in a common struggle. By 1945, the bishops held that there could be a positive meaning to the term *laicity* and, in the next year, *laicity* was officially proclaimed by the Constitution. It was a tripartite government—communist, socialist, and MRP (a political party that is close to Christian Democracy)—that included *laicity* in the Constitution.

3. *Laicization in Various Countries*

Other countries also show cases of the process of laicization which comprise analogies with the French case as well as specificities. Here are three examples. First, Belgium became an independent state in 1831, in a 'unionist' atmosphere, which gathered Catholics and Liberals rather than anticlerics. The outcome was a regime of separation–recognition, accompanied by 'recognized cults' (but no Concordat) which tended to resemble the first threshold of laicization in France—this was possible because the situation was more democratic than that of Napoleon Bonaparte. In 1879, the Liberals attempted to create a Laique school but the Catholic Party, victorious at the 1884 elections, progressively abolished these measures. Therefore, Belgium was never situated in a similar logic to the French second threshold but slowly integrated laicity and *laique* morals in its system of 'recognized cults': 'nonconfessional' morals, school-classes, *laique* advisers in hospitals or prisons, *laique* youth fetes that were similar to the first communion, etc.

For a very long time, Spain lived through a refusal of laicization of any sort. The 1912 Constitution declared that 'the Nation's religion is (…) the apostolic and Roman Catholic religion, the only real one.' But in the 1931 Constitution, the Spanish republic imposed both elements of the first French threshold (abolition of the religious Catholic monopoly, institution of the civil marriage) and the second French threshold (separation, *laique* schools, suppression of religious symbols on public buildings), and even a few measures that go beyond the second threshold in France (for example, chaplaincies were suppressed). During the Civil War, this laicization resembled the situation of the French Revolution rather than that of the Third Republic. The issue lay in Franco's national Catholicism, which re-affirmed Roman Catholicism as 'the only real' religion. Nevertheless, since 1978 there has been no State religion but 'relations based on cooperation with the Catholic Church and other confessions.'

The Italian situation of the mid-twentieth century was more ambivalent than the Spanish one but was not *laique* either. The 1948 national reconciliation respected the agreements of Latran (1929) which was a compromise between the Fascist regime and the Catholic Church. On the other hand, during the same time, it affirmed the mutual independence of the State and Catholicism (State religion). This lasted until 1971, when the constitutional court affirmed the priorities of constitutional norms on the agreements with the Catholic Church. After that, Catholicism ceased to be the State religion, divorce was authorized, and in 1984, the Concordat was re-negotiated and an official pluralism began to exist.

Today's Spanish and Italian systems tend to resemble a laicity that corresponds to the first French threshold of laicization. On the other hand, England possesses an established Church that coexists with a maintained Catholicism as well as nonconformist Churches. This constitutes an original variation of the process of secularization. By the end of the seventeenth century, politico-religious conflicts no longer took place. In Germany, after the Thirty Years' War, the main process was that of secularization but it also integrated elements of laicization (*Kulturkampf*, in particular).

We then could think that the French situation is the most *laique* but this is not the case. The separation between State and Church in Mexico (1917) was very polemical towards religion. The Constitution shows, for example, that State education 'should remain absolutely away from any religious doctrine' and should fight 'fanaticism and prejudices.' It forbids monastic orders, makes sure that religious ceremonies only take place in Churches, which 'will always be under the authority's supervision,' refuses the right to vote to the ministers of cults, etc. Some measures were recently made less strict, but Mexican laicity was historically harsher than French laicity.

The Turkish laicity also stands in a polemical logic when it comes to religion, which explains why it is backed up by the army. Some liberal personalities have tried to render it more liberal and make it come closer to the French laicity but today's situation does not favor such an evolution.

4. *New Problems and European Construction*

Nowadays, whatever the degree of laicization, various societies must face new problems. For example, culturally speaking, the dynamic element in laicity had been the development of a morality independent of any religious authority. But this morality was frequently going to be based on science. Now science and morality tend to be more or less dissociated in many respects. Science was increasingly understood as a *techno-science*, the functional efficiency of which was undeniable but which, far from helping to resolve moral questions, created new ones, more difficult to resolve because of their intrinsic power. The devel-

opment of genetics is significant, and in bio-ethical committees, representatives with a *religious outlook* are often asked to give their point of view. In today's society, there are various debates on morals: questions concerning sexuality, marriage, the representation of violence on television, etc. are brought up. Religious opinions participate in these debates but their solutions are not necessarily imposed on the whole of society.

In several countries, debates concern laicity and Islam directly: for some immigrants' children, this opposes the culture that is taught at school and the culture that is transmitted within the family, which rather finds its roots in tradition. In France the *headscarves affair* is a typical example. But this affair finds significance by showing how difficult it can be for the immigrant populations of Muslim belief or origin—who often settle in the outskirts of a city; areas that are prone to break up and directly subject to increased unemployment and the crisis of the Welfare State—to integrate. On the other hand, Buddhism has integrated without difficulty.

Moreover, the relations between the State and religion in most of the countries of the European Union tend to become a specific case of a larger European model. Italian Professor Silvio Ferrari defines it with the following co-ordinates (Ferrari 1998):

(a) The State is impartial towards the various individual religious subjects.

(b) A religious subsector is singled out within the public sector. This may be understood as a playing or protected area field. (…). (This is the area in which lay most of the differences between France and the other European countries because in France this area is not institutional.)

(c) The State has the right to intervene in this area only to see that the players respect the rules of the game and the boundaries of the playing field.

This model is now deeply rooted in the political and legal culture of contemporary Western Europe. The model is ingrained to such an extent that it has become the basis for the very formulation of the laws in force concerning religious liberties and church–state relationships. This may be seen both in internal conventions and in the constitutional laws of the main European nations. These conventions and laws are formulated to ensure the impartiality of the public powers, which are legally bound to respect the freedom to profess any religious faith (besides art. 9 of the European Convention of Human Rights) and to avoid any discrimination based on religion (art. 14 ECHR). In this entirety, secularization can be put side by side with some elements of laicity by diminishing the differences linked to the existence of an official religion.

To conclude, the ideal type of *laicity* can first be characterized as a twofold refusal: the refusal of an official religion, the standards and prescriptions of which would be imposed on all by means of civil law with the state's affirmation of religious neutrality, and the refusal of an official atheism which would limit religious freedom. More positively, *laicity* can be seen as a close relation between the freedom of conscience and the freedom of thought. The freedom of conscience implies that every individual has the right to believe or not to believe, to retain or to change his beliefs. Therefore, the freedom of conscience leads to the pluralism of beliefs and philosophies within society. The freedom of thought implies that none of these beliefs or philosophies are privileged or imposed by the State to the detriment of other convictions. The State is to respect them all and is to make sure that they can all be expressed freely.

See also: American Revolution, The; Church and State: Political Science Aspects; French Revolution, The; Religion: Definition and Explanation; Religion: Evolution and Development; Religion, History of; Religion, Sociology of; Secular Religions; Secularization

Bibliography

Baubérot J (ed.) 1994 *Religions et Laïcité dans l'Europe des Douze*. Syros, Paris

Baubérot J 1997 *La Morale Laïque Contre l'Ordre Moral*. Le Seuil, Paris

Baubérot J 2000 *Histoire de la Laïcité Française*. Presses Universitaires de France, Paris

Cavana P 1998 *Interpretazioni della Laicita*. Esperienza francese ed Esperienza italiana a confronto, Roma, editrice a.v.e.

Davie G 1994 *Religion in Britain Since 1945, Believing without Belonging*. Blackwell, Oxford, UK

Ferrari S 1998 Tolerance, religion and the law in contemporary Europe. In: Ferrari V, Heller Th, de Tullio E (eds.) *Citizenship and Immigration*. Giuffre editore, Rome, pp. 159–77

Hera A de la, Martinez de Codes R M 1998 *Spanish Legislation on Religious Affairs*. Ministerio de Justici, Madrid

Larkin M 1974 *Church and State after the Dreyfus Affair, The Separation Issue in France*. Macmillan Press, London

Poulat E 1987 *Liberté Laïcité*. Le Cerf/Cujas, Paris

Rémond R 1998 *Religion et Société en Europe*. Le Seuil, Paris

Warner R S 1993 Work in progress toward a new paradigm for the sociological study of religion in the United States. *American Journal of Sociology* **98**(5): 1044–93

Weill G 1929 *Histoire de l'Idée Laïque en France au XIXème Siècle*. Alcan, Paris

J. Baubérot

Land Degradation

'Land degradation' is a composite term. It encompasses an array of biophysical processes, each of which may result in a reduction in the quality of the

land. It also includes a range of socioeconomic outcomes, any one of which implies a worsening of the human condition. Specifically, it is now usually defined as the temporary or permanent decline in the productive *capacity* of the land, and the diminution of the productive *potential*, including its major land uses (rainfed arable, irrigated, rangeland, forestry), its farming systems (e.g., smallholder subsistence) and its value as an economic resource. 'Capacity' and 'potential' are, however, elusive to measure: they involve many other issues of identification of environmental change and long-term effects on society (see *Food Security*; *Human–Environment Relationship: Indicators*; *Sustainable Development*). Consequently, land degradation is more usually perceived in terms of its symptoms and biophysical process, from the scenically dramatic (gullies and badlands) to the often invisible (sheet erosion by water, loss in soil physical structure). Other definitions of land degradation abound, and many authors find difficulty distinguishing between it and *desertification*. The confusion accounts for the undermining of the credibility of the debate over the worldwide seriousness of land degradation (Pimentel 1993). Here, land degradation will be viewed in terms of the environmental changes that reduce the productivity and sustainability of landscapes and of the primary renewable natural resource components, especially the soil and vegetation.

1. *Land Degradation: An Old Problem with New Urgency*

Land degradation has been advanced as 'the single most pressing current global problem' (O'Riordan 2000). Yet, it is an old problem (e.g., Jacks and Whyte 1939), that has undergone a series of often-emotive revivals every decade since the Dust Bowl era in the mid-West USA (e.g., Osborn 1948, Carson 1962, Commoner 1972; Blaikie and Brookfield 1986). The historical stereotype of land degradation is that it is a process ruining the planet, and a destruction caused by ignorant peasants, who will in short time reap the folly of their degrading activities. Hailey's (1938) *African Survey* called it the 'scourge of Africa,' providing the reader with images of sediment-choked rivers and barren hillsides. Concern about degradation, particularly the processes of soil erosion, has fuelled many campaigns to combat it, and spawned numerous institutions to address it. In colonial Africa, for example, the Federation of the Rhodesias and Nyasaland funded the largest ever soil conservation research program on the continent between 1953 and 1965 to investigate rates of erosion and runoff, as well as the effectiveness of conservation and types of land use. In Southern Rhodesia, research evidence of the potential seriousness of erosion directly led to the formation of the Department of Conservation and Extension (CONEX), which carried out major conservation planning and design works throughout the country. A somewhat different institutional base was set in Malawi in response to perceived land degradation. The Land Husbandry Branch was formed in 1960 to undertake land use planning. This led to the more holistic concept of 'land husbandry,' including all farm-level production activities, argued as being a more effective tool for delivering conservation than technical, often structural, measures such as earth bunds (contour embankments to intercept runoff and sediment).

In the US, worries about land degradation supported by pictures of (usually) wind erosion engulfing farmhouses and fields, triggered massive expenditure on research and new institutions. The Soil Conservation Service (now the Natural Resources Conservation Service) was the largest single section of the US Department of Agriculture. Farmers were obliged by law to comply with conservation planning regulations and procedures. Compliance involved using research outputs such as the 'Universal Soil Loss Equation' which calculated rates of soil loss for various planned land uses, comparing these with a benchmark known as the 'tolerable soil loss,' a rate at which it was said that future production would not be jeopardized. Some of the science underpinning these procedures is now known to be flawed (Stocking 1996a). Yet, there is no doubt that the sustained research and institutional effort in the US, spurred originally by the Dust Bowl era, has not only raised awareness of land degradation but also done much to conserve that nation's soils today.

At international level, there is now a new urgency in addressing land degradation. While the World Commission on Environment and Development (Brundtland 1987) highlighted sustainable development, land degradation was only then seen as a shadowy adversary in achieving long-term productive agriculture. It was an insidious enemy, often masquerading under other names (see *Desertification*). The series of Sahel droughts in the mid-1980s illustrates the image problem of land degradation. Headlined 'drought,' its main impact came through widespread crop and grazing land failure due to the reduced capability of the degraded soils to support plant growth. The water-holding capacity of a degraded soil is only a fraction of a virgin soil. Land degradation therefore increases the vulnerability of marginal environments and societies.

Nevertheless, in the search for sustainable global futures, land degradation has once again come to the fore. The 1992 United Nations Conference on Environment and Development in Rio de Janeiro revived interest. However, it was not until 1998 that enough signatories drew their pens to ratify the *Convention to Combat Desertification*, the major components of which are national and international proposals to combat land degradation. Today, the fastest-growing

Table 1
Degradation status of global land in crops, permanent pasture, forest, and woodlands

Degradation category	Amount of land affected (million ha)	Lost production (percent)
Total usable land	8,735	
Not degraded	6,770	0
Degraded	1,965	
– lightly	– 650	5
– moderately	– 904	18
– strongly	– 411	50

Sources: Crosson 1997; O'Riordan 2000

section of the global funding mechanism (the Global Environment Facility, GEF), which supports the incremental cost of developing countries' efforts to comply with the international environment-development conventions, is for land degradation projects. The driving force for the renewed interest comes from new data from international projects such as the Dutch-based *Global Assessment of Soil Degradation* (see Table 1). The International Food Policy Research Institute uses such data to show that, while food production may largely be maintained by ever-increasing inputs, there are 'hotspots' of serious land degradation in vulnerable places such as north-east Thailand, northern China, and many African drylands (Scherr and Yadav 1996).

2. *Land Degradation: Differing Perspectives and Contexts*

The strident rhetoric of land degradation continues today. Bolstered by the power of numbers, land degradation is pictured today as getting ever worse: 'each year, 75 billion metric tons of soil are removed from the land' (Pimentel, quoted in Reij et al. (1996) who provide a good critique of the selective use of 'scientific' findings). This 'crisis' perspective is, however, becoming tempered with a new realism (Glanz 1999). One recent study questions estimates such as Pimentel's: over 20 years in an intensively cultivated drainage basin in Wisconsin, USA, sediment accretion was only about 6 percent of the rate that occurred in the 1930s, despite large stream and valley changes, while sediment yield was relatively constant (Trimble 1999). If predictions of the past had been true, whole continents would now be laid waste. Manifestly, they are not. So what is happening?

As agriculture intensifies, soils are certainly changing for the worse. Their intrinsic biological functions to support plant growth (e.g., in the natural processes of making nitrogen available to plants) have diminished. Research shows conclusively that land degradation produces poorer soils and reduced crop yields (Tengberg et al. 1998). Mediterranean environments and the tropics in general are especially vulnerable (Conacher and Sala 1998; Eden and Parry 1996). Typically, the first few tonnes of soil loss show the greatest decline in yields. Depending on actual soil type, up to 75 percent of the starting yield can be lost with 1 cm loss of soil (about 130 tonnes per hectare). Production is only maintained by compensating for the diminished state of the soil resources—through providing nutrients artificially, increasing use of irrigation to supply water that would otherwise have been supplied by a rich soil, new crop breeds that can tolerate degradation-induced toxicities, and so on. So successful are the new technologies that for most farmers they work out cheaper to implement than the equivalent conservation measures to maintain the intrinsic soil quality. Perversely, the economics of commercial farming mean that it is more financially viable to increase production while allowing land degradation than to farm conservatively. Nowhere is this better illustrated than in some intensive farming in Europe, where sandy soils have seen a decline in organic matter to as little as a quarter of 1 percent. By standard guidelines from the FAO, these soils are now severely degraded. Yet, never have they produced greater yields (Pretty 1998). With sophisticated rotary boom irrigators, multiple applications of fertilizer, and relatively cheap soil amendments such as lime, 6–10 tonnes of wheat per hectare is typical. The consequences elsewhere in the environment of substituting technology for combating land degradation can be dire (e.g., see *Water Pollution: Terrestrial and Groundwater*). Nevertheless, society appears to prefer to clear up the consequences than tackle the root causes and processes of land degradation.

Land degradation is best perceived today as the net result of a web of causes and consequences, some of which are easily and cheaply manipulated, while others are fixed. With soil productivity providing the core variables of concern that are affected by land degradation, Fig. 1 demonstrates the complex interactions on biophysical, socioeconomic, political, institutional, and cultural factors. Farming systems, the implementation of soil and water conservation practices, and assistance from external agencies (lower two right hand boxes in Fig. 1) could be manipulated relatively easily for the benefit of soil productivity and the reduction in land degradation. Whether it would be economic so to do, or viable with current institutions, is another matter. On the other hand, geology, climate, the international economy and cultural inheritance are all relatively fixed components, providing a backdrop of largely unchangeable influences that may cause degradation when the land is used. Appreciating the scale of the complex influencing factors and their amenability to external intervention is now seen as essential to developing approaches to environmental problems acceptable both to scientists and society (see *Ecology, Political*).

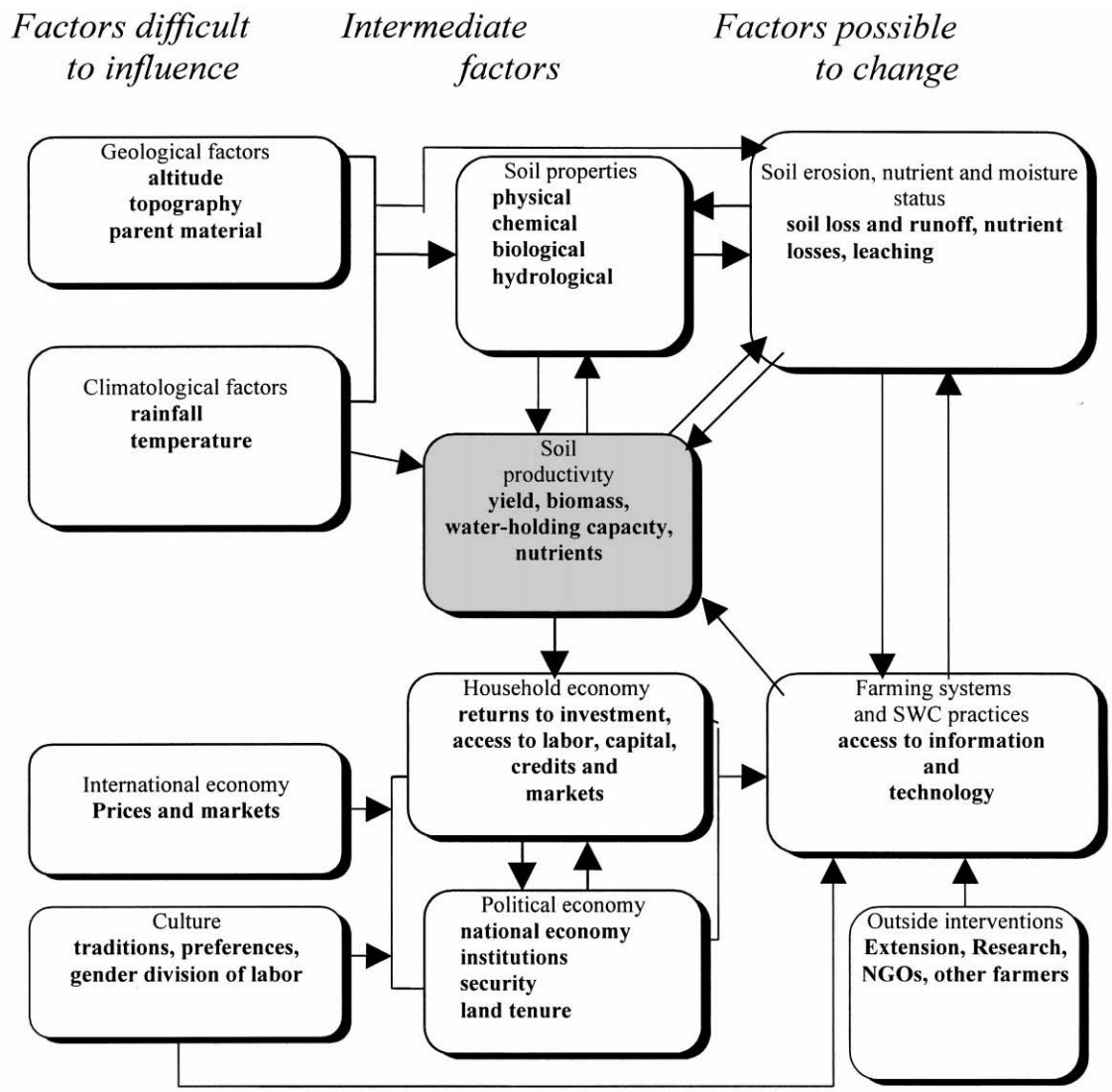

Figure 1
Causes and consequences of land degradation and its impact on soil productivity

3. *Land Degradation: Biophysical Processes*

Land degradation is usually described by the natural resource that is being depleted (e.g., soil/vegetation/environmental degradation) or the biophysical process by which it operates (e.g., soil erosion by wind or water, sodication, salinization, deforestation). In all these processes of environmental change, the soil is normally seen as the focal resource that diminishes in quality with land degradation. As vegetation degrades in quantity and species composition, for example, the soil changes chemically, physically, and biologically. A vicious cycle is evident in most degradation processes: as the soil degrades, so is its ability to support plant growth or other life-support functions. In soil degradation six processes usually are recognized:

(a) *Water erosion*. This includes the splashing of soil particles by raindrop impact; sheet erosion whereby a layer of topsoil is removed by flowing water; and gully erosion where a channel is formed. Gullies are often perceived as the most serious form of water erosion because they are obvious features in the landscape. However, sheet erosion by water removes far greater quantities of soil. Worldwide it is by far the most

common land degradation process in both amount of soil lost and impacts on production.

(b) *Wind erosion*. This is the removal and deposition of soil by wind. It is commonest in, but by no means restricted to, arid and semi-arid areas. A principal cause is the removal of vegetation by, for example, overgrazing or preparation for cultivation.

(c) *Excess of salts*. Salts may accumulate in the water held in the soil (salinization) or sodium cations may increase in proportion to nutrient cations such as calcium attached electrochemically to solid soil particles (sodication or alkalinization). The first is typical of semi-arid areas and of poorly managed irrigation schemes, while the second is largely a natural process accelerated by water runoff and associated concentrations of sodium induced by human land use.

(d) *Chemical degradation*. This includes a variety of processes. Some are related to the loss of plant nutrients by percolating water (e.g., calcium) or by chemical fixing into a form unavailable to plants (e.g., phosphorus); another is the build-up of toxic levels of chemicals. The pH-related change of aluminum into a free form that can be accessed by plants causes the severest impact on production in certain vulnerable tropical soils.

(e) *Physical degradation*. Adverse changes in properties such as porosity, permeability, bulk density and structural stability may arise through farming practices. The commonest physical degradation is the formation of a surface crust by raindrop impact, causing decreased water infiltration and greater runoff.

(f) *Biological degradation*. Organic matter is a transient component of soils. Where it is lost through natural processes such as mineralization at a rate significantly faster than it is resupplied by vegetation, the soil deteriorates in its overall biological functions. This happens on most intensively farmed agricultural soils.

These processes of degradation in soils all threaten the sustainability of agriculture. Usually, several processes occur simultaneously. Water erosion, for example, results in loss of soil structure, surface crusting, waterlogging, reduction in organic matter, and breakdown of stable aggregates. In the face of such an onslaught, soil resources very quickly deteriorate. Farming becomes more difficult and more costly.

4. *Land Degradation: Economic and Social Impact*

Explanations of land degradation have been made at a number of spatial and temporal scales. They express the influence of, and the impact on: (a) the resource endowments of land users; (b) the nature of agrarian society and of the state; and (c) even the imperfections in the international world economy (see Blaikie 1989 for his original presentation of a political economic analysis of land degradation; also O'Riordan 2000 for an updated version). Such an analysis gives a useful entry to socioeconomic understanding of the impacts of land degradation, helping to identify where interventions are possible and by whom. They help to counterbalance the excessively technical and place-based focus of most attempts at combating land degradation. They also explain why many soil and land conservation projects have failed because of inadequate attention to the root social and economic causes of degradation.

A partial listing of social impacts would have to include examples of major disruptions that accompany land degradation. War and conflict, as in the Horn of Africa, and around Liberia and Sierra Leone, have caused massive migrations and concentrations of people in adjacent areas. Pastoralists, in particular, live in sensitive semi-arid environments, where overgrazing of vegetation and trampling by animals is a constant danger. In normal times, society has its local institutions that manage access to the pool of common property resources of rangeland; in times of conflict, these institutions are often first to disappear. Political decisions can also have unintended side effects of promoting land degradation. *Ujamaa* (or villagization), forcibly promoted by Julius Nyerere's Tanzanian Government in 1972–73 on the grounds of better provision of education, health, and communication services, had the effect of concentrating formerly dispersed settlements into villages up to several thousand people. In vulnerable but fertile areas such as Mwanza and Shinyanga, this had disastrous consequences for the soil and land resources close to villages.

The economic impact of land degradation has received attention only recently. Valuation of the quality of land is fraught with uncertainty, and depends upon the perspective of the analyst. There are two principal choices for valuation: the production or the resource value approaches. The first assesses the impact of land degradation in the value of production lost, usually in market prices of the staple crop. This is a useful device for commercial farming situations, where production foregone is a real loss to the land user. The greater cost of farming the land or the increased investments in fertilizers or irrigation to compensate for land degradation have also been added to the value for lost production. The resource value approach, which has been the commonest because the data have been collected from experimental plots in the past, values land degradation in terms of the quantity of nutrients lost in the sediment. The means for placing a monetary value is through fertilizer prices. This second approach tends to give much larger and often unrealistic costs to land degradation than the first. Environmentalists would, however, argue that the resource value approach calculates a cost for society and future generations, and not just the immediate impact on land users.

5. *Combating Land Degradation: Future Directions*

Land degradation is, therefore, a complex biophysical, social, and economic issue, which affects many lands and peoples, particularly in the tropical developing world. The processes involved have immediate on-site impacts, rendering land diminished in quality (see *Human–Environment Relationship: Carrying Capacity*; *Sustainability Transition: Human–Environment Relationship*) and ability to support plant growth (see *Food Security*). Off-site impacts are also significant in, for example, reservoir sedimentation, eutrophication, and other downstream damages. Combating land degradation needs a commensurately broad armory of approaches and techniques, targeted not just at root biophysical causes, but also at social and economic controls, which may persuade farmers that it is in their interests to farm more conservatively.

Seeking the right intervention for the appropriate target is the major challenge for land managers. There are no shortages of techniques to combat land degradation. However, the heterogeneity of local society, the varying wealth status and resource endowments of local people, and property rights that affect security of tenure, all determine whether a technique is rational and viable for any one household. Future research will have to concentrate on matching solutions to specific environmental situations, societies, and economic demands. A better appreciation of indigenous technologies has already led to rehabilitation projects that combine elements of local knowledge with formal science. Low cost technologies which use primarily biological means will be the main focus of interest for the most seriously degraded lands.

The term *land husbandry* has been coined recently to emphasize that control of land degradation must be seen as part of the agricultural production enterprise and environmental management practice. For too long, conservation has been an additional activity. It may have attracted subsidies and incentives, but these in the longer term have simply reinforced the burdensome and uneconomic nature of many of the techniques being promoted. Land husbandry fosters the notion that integrated, grassroots approaches to the land involve the total production cycle, land users' constraints and opportunities, access to land, labor, and capital, and the technical appropriateness of solutions.

Bibliography

Blaikie P 1989 Explanation and policy in land degradation and rehabilitation for developing countries. *Land Degradation and Rehabilitation* **1**: 23–37

Blaikie P, Brookfield H 1986 *Land Degradation and Society*. Methuen, London

Brundtland G H 1987 *Our Common Future*. Oxford University Press, Oxford, UK

Carson R 1962 *Silent Spring*. Houghton Mifflin, Boston

Commoner B 1972 *The Closing Circle: Confronting the Environmental Crisis*. Cape, London

Conacher A J, Sala M (eds.) 1998 *Land Degradation in Mediterranean Environments of the World: Nature and Extent, Causes and Solutions*. Wiley, Chichester, UK

Crosson P 1997 Will erosion threaten agricultural productivity? *Environment* **39**(8): 4–9, 29–31

Eden M J, Parry J T 1996 *Land Degradation in the Tropics: Environmental and Policy Issues*. Pinter, London and New York

Glanz J 1999 Sharp drop seen in soil erosion rates. *Science* **285**: 1187–9

Hailey, Lord 1938 *An African Survey*. Oxford University Press, Oxford, UK

Jacks G V, Whyte R O 1939 *The Rape of the Earth: A World Survey of Soil Erosion*. Faber and Faber, London

O'Riordan T 2000 *Environmental Science for Environmental Management*, 2nd edn. Prentice-Hall & Pearson Education, Harlow, UK

Osborn F 1948 *Our Plundered Planet*. Faber and Faber, London

Pimentel D (ed.) 1993 *World Soil Erosion and Conservation*. Cambridge University Press, Cambridge, UK

Pretty J N 1998 *The Living Land: Agriculture, Food and Community Regeneration in Rural Europe*. Earthscan, London

Reij C, Scoones I, Toulmin C 1996 *Sustaining the Soil: Indigenous Soil and Water Conservation in Africa*. Earthscan, London

Scherr S, Yadav S 1996 Land degradation in the developing world: Implications for food, agriculture, and the environment to 2020. Food, Agriculture and the Environment Discussion Paper 14, International Food Policy Research Institute, Washington, DC

Stocking M A 1996a Challenging conventional wisdoms about soil erosion in Africa. In: Leach M, Mearns R (eds.) *The Lie of the Land: Challenging Received Wisdom in African Environmental Change and Policy*. Curry, London

Stocking M A 1996b Soil erosion. In: Adams W M, Goudie A S, Orme A R (eds.) *The Physical Geography of Africa*. Oxford University Press, Oxford, UK

Tengberg A, da Veiga M, Dechen S C F, Stocking M 1998 Modelling the impact of erosion on soil productivity: A comparative evaluation of approaches on data from southern Brazil. *Experimental Agriculture* **34**: 55–71

Trimble S 1999 Decreased rates of alluvial sediment storage in the Coon Creek Basin, Wisconsin, 1975–1993. *Science* **285**: 1244–6

M. A. Stocking

Land Reclamation

The term land reclamation can be used in two main ways. On the one hand it can be used to describe actions undertaken by humans to reclaim land from its natural state so that it can be put to some particular use. On the other it is employed to describe ways of restoring ecosystems that have been degraded by human actions.

1. Reclamation From a Natural State

The first of these uses has been employed by historical geographers and others who have sought to investigate the ways in which landscapes have been transformed by human actions. Of particular note here was the work of the American conservationist Marsh (1864). In this magnificent and ground-breaking work he not only described the ways in which humans had modified nature by such actions as deforestation, thereby causing land degradation, but also looked at the ways in which more deliberate actions had been undertaken to transform the face of the Earth. So, for example, in his chapter on '*The Waters*' he described 'works of harbour and coast improvement,' exclusion of the sea by the construction of dikes, as in The Netherlands, the draining of lakes and marshes, the lowering of the levels of lakes, surface and under-draining of soils, irrigation of the world's drylands, and the embanking of rivers to control torrents and inundations. In his chapter on '*The Sands*' he described the ways in which dunes, such as those of the Landes of Gascony in France, could be controlled, stabilized, and made productive by judicious planting.

Since he wrote, there have been many studies that have followed up this theme of the deliberate transformation of landscapes, including notable studies of the draining of the marshes of the Fenland in East Anglia, England, by Darby (1940), or the draining of the Somerset Levels in western England by Williams (1970). Humans have exerted great energy to 'reclaim' certain landscapes, be they the industrious Dutch who have built large tracts of their country by protecting it from the sea, the land-hungry inhabitants of Hong Kong who have built out their high rise urban structures on 'reclaimed land,' the engineers who have lined the Mississippi River with enormous embankments, or the hydraulic engineers who constructed the massive canal systems that irrigate the plains of the Indian subcontinent and elsewhere (Table 1).

A particularly important type of land reclamation is the transformation of landscapes from their forested state into other types of land cover and land use (e.g., agriculture and pastoralism). This is one of the oldest forms of reclamation, which dates back to the Mesolothic and Neolithic. More recent examples include the medieval clearances of forest from central Europe, the removal of native vegetation from extensive tracts of North America since European settlement, and the current spasm of clearance of the rain forest. On a global basis over a third of the earth's land surface has been converted to 'domestic' use.

Table 1
Estimates of the increasing area of irrigated land on a global basis

Year	Irrigated area (10^6 ha)
1900	44–48
1930	80
1950	94
1955	120
1960	168
1980	211
1990 (estimate)	250

2. Reclamation of Degraded Land

The second meaning of the term is one that has come into prominence as it has become apparent that there is an increasing need to restore or reclaim land that has been degraded by human activities. Human actions frequently have rendered land unproductive or have reduced its ecological value. Examples of such actions include the destruction of the land surface by mining activities, the pollution of the ground and of water resources by various types of chemical or biological substances, the salinization of soil and water as a result of irrigation or land clearance, soil erosion resulting from over-cultivation or over-grazing, the loss of woodlands, the spread of desert conditions (desertification), and habitat destruction and associated reduction in biodiversity. The significance and extent of many of these problems is now well known (see, e.g., Turner 1990, Goudie 2000, Thomas 1956, Mannion 1998) and had led to an increasing concern with means of environmental management (Owen and Unwin 1997), of which land reclamation is a part. Connected with land reclamation is the concept of ecological restoration—meaning the desire to return an area to its original state prior to degradation so that it becomes a fully-functioning, self-sustaining ecosystem. Not all land reclamation, however, goes as far a being full ecological restoration. To give an example; an old mining waste tip may be reclaimed for agricultural use rather than being restored as a fully natural and pristine forest. Likewise, a badly salinated and waterlogged tract of irrigated land may be reclaimed for agriculture by the installation of an effective network of drains, but is not restored to its original state as a tract of desert. Concepts and means of land restoration and reclamation have been reviewed by Harris et al. (1996).

3. Examples of Land Reclamation and Restoration

In order to give an indication of the nature and scope of land reclamation and restoration techniques, it is appropriate to provide a small selection of examples. The first of these is to look at means of rehabilitating land that has been damaged by soil erosion. Morgan (1995) provides a review of some of the methods that have been employed, and draws upon the experience

Table 2
Means of reclaiming areas eroded by water action

Revegetation:
(a) deliberate planting;
(b) suppression of fire, grazing, etc., to allow regeneration

Measures to stop stream bank erosion (e.g., stone banks and rip-rap)

Measures to stop gully enlargement:
(a) planting of trailing plants, etc.;
(b) weirs, dams, gabions, etc.

Crop management:
(a) maintaining cover at critical times of year;
(b) rotation of crops;
(c) growing cover crops;
(d) agroforestry

Slope runoff control:
(a) terracing;
(b) deep tillage and application of humus;
(c) digging transverse hillside ditches to interrupt runoff;
(d) contour ploughing;
(e) preservation of vegetation strips (to limit field width)

Prevention of erosion from point sources such as roads and feedlots:
(a) intelligent geomorphic location of roads, feedlots, etc.;
(b) channeling of drainage water to non-susceptible areas;
(c) covering of banks, cuttings, etc., with vegetation

Table 3
Control techniques for reclaiming areas from drifting sand and mobile dunes

Problem	Control methods
Drifting sand	Enhancing of deposition of sand through creating large ditches, vegetation belts and barriers, and fences
	Enhancement of sand transport by aerodynamic streamlining of surface, or changing surface materials
	Reduction of sand supply by surface treatment, improved vegetation cover or erection of fences
	Deflection of moving sand by fences, barriers, or vegetation belts
Moving dunes	Removal by mechanical excavation
	Destruction by reshaping, trenching through dune axis or surface stabilization of barchan arms
	Immobilization by trimming, surface treatment and fences

that was developed not only by pioneers like Bennett (1939), who established and directed the US Soil Conservation Service from the 1930s to deal with some of the problems encountered in the Dust Bowel years, but also upon experience gained in the degraded lands of Africa and Asia. The prime means of controlling erosion (Table 2) is to undertake revegetation, either by deliberate planting or by the suppression of fire and grazing to allow regeneration. Associated with this is the need for crop management, whereby an adequate level of plant cover is maintained at critical times of year, rotation is practised, and intercropping is employed. It is also necessary to prevent erosion from point sources like roads and tracks by the careful use of drains, to protect stream banks with structures like gabions, to stop gully encroachment by means of weirs and dams, and to control slope runoff by terracing, deep tillage, and the application of humus, the digging of transverse ditches to interrupt runoff, contour ploughing, and the preservation of grass strips to limit field width. Wind erosion requires its own techniques (Table 3), including the planting of shelterbelts or windbreaks and the preservation of soil moisture levels (Watson 1990, Middleton 1990). Because of the widespread and detrimental effects of soil erosion, attempts at its control have a very long history, and in Africa, for example, there are many indigenous strategies that have proved remarkably effective (Reij

Table 4
Global estimate of secondary salinization in the world's irrigated lands

Country	Cropped area (Mha)	Irrigated area (Mha)	Share of irrigated to cropped area (percent)	Salt-affected land in irrigated area (percent)	Share of salt-affected to irrigated land (percent)
China	96.97	44.83	46.2	6.70	15.0
India	168.99	42.10	24.9	7.00	16.6
CIS	232.57	20.48	8.8	3.70	18.1
USA	189.91	18.10	9.5	4.16	23.0
Pakistan	20.76	16.08	77.5	4.22	26.2
Iran	14.83	5.74	38.7	1.72	30.0
Thailand	20.05	4.00	19.9	0.40	10.0
Egypt	2.69	2.69	100.0	0.88	33.0
Australia	47.11	1.83	3.9	0.16	8.7
Argentina	35.75	1.72	4.8	0.58	33.7
South Africa	13.17	1.13	8.6	0.10	8.9
Subtotal	842.80	158.70	18.8	29.62	20.0
World	1473.70	227.11	15.4	45.4	20.0

Source: reproduced by permission of CAB International and the University of New South Wales Press from *Salinisation of Land and Water Resources* 1995.

Table 5
Methods available for reclamation of salinized land

Improvement of sodic soils by the replacement of excessive adsorbed sodium by calcium or magnesium
Provision of adequate drainage to prevent waterlogging
Removal of excess salts by leaching
Water conservation (e.g., use of sprinklers, drip irrigation, cultivation of plants with low water needs, etc.)
Efficient delivery systems that minimise seepage loss from canals and deliver an appropriate amount of water at the night time
Careful leveling of land to enable more uniform application of water and, hence, better salinity control
Use of interception basins to trap saline water

1991) even in the face of quickly rising human population levels (Tiffen et al. 1994). There are also many examples of soil conservation measures that have been successful in reversing degradation, as is made evident, for instance, by the studies of Trimble in the US (Trimble and Lund 1982).

Another example is the reclamation of land that has been afflicted by salinization and waterlogging as a result of the introduction of irrigation. This is a problem that affects large tracts of land in the Indus valley of Pakistan, the Tigris–Euphrates system of Iraq, and many other potentially productive arid areas (Ghassemi et al. 1995) (Tables 4 and 5).

The third example one can give is the reclamation of contaminated ground. The reclamation methods available broadly divide into two types: traditional methods of isolation, covering or removal of contaminated materials; and decontamination techniques. Among the means of achieving isolation are capping the deposits to reduce rainfall infiltration, the installation of vertical barriers to minimize migration of contaminated waters, the digging of diversion trenches to intercept water and/or pollutants from contaminated ground, horizontal barriers beneath the deposits to prevent downward leaching, and break layers of coarse stones to prevent upward movement of contaminants by capillary action. Decontamination techniques can be divided broadly into thermal, separation, solidification/stabilization, chemical, and biological methods. Once the ground material has been treated in these various ways it will be necessary to revegetate the site, though this is not always a simple process, because the materials may be poor or lacking in nutrients, there may be some residual level of toxicity, the materials may have adverse moisture conditions (i.e., too wet or too dry), they may be compacted, or they may be too unstable to permit vegetation to be established. For all these reasons reclamation of contaminated land is often a difficult and lengthy business that may involve the use of fertilisers and essential minerals, the importing of soil, the installation of drains and mulches, stabilisation with geotextiles or by landscaping (e.g., terracing), and an understanding of the best pioneer plants to employ. Leguminous plants, for example, may be especially useful.

4. A Multidisciplinary Endeavor

What these three examples indicate is that land reclamation is a multidisciplinary endeavor that requires the involvement of a wide range of disciplines. Anthropologists may be required to understand and explain social mores, environmental historians may be needed to ascertain what conditions existed before degradation occurred, plant scientists will be required so that appropriate planting strategies are followed, engineers will be required to undertake some of the technical tasks that are necessary, geomorphologists may be useful for providing insights into appropriate terrain shapes, hydrologists will be needed to deal with water flows, and soil scientists will use their expertise to make sure that soils are used in an effective manner. Connected with this is the observation that land reclamation is not merely a technical matter that can be achieved solely by engineering solutions. Much land reclamation involves human decisions (e.g., deciding not to use too much irrigation water, deciding to introduce crop rotations or to plant and maintain shelterbelts, deciding to reduce grazing pressures, deciding to what uses contaminated land should be put, etc.). It is, therefore, an area where social and physical scientists need to come together.

See also: Desertification; Ecology, Political; Environmental Adaptation and Adjustments; Environmental and Resource Management; Environmental Planning; Environmentalism, Politics of; Environmentalism: Preservation and Conservation; Human–Environment Relationship: Carrying Capacity; Human–Environment Relationship: Indicators; Land Degradation; Sustainability Transition: Human–Environment Relationship; Sustainable Development

Bibliography

Bennett H H 1939 *Soil Conservation*. McGraw Hill, Chicago

Darby H C 1940 *The Draining of the Fens*. Cambridge University Press, Cambridge, UK

Ghassemi F, Jakeman A, Nix H A 1995 *Salinisation of Land and Water Resources*. CAB International, Wallingford, UK

Goudie A S 2000 *The Human Impact on the Environment*, 5th edn. Blackwell, Oxford, UK

Harris J A, Birch P, Palmer J 1996 *Land Restoration and Reclamation*. Longman, Harlow, UK

Mannion A M 1998 *Global Environmental Change*, 2nd edn. Longman, Harlow, UK

Marsh G P 1984 *Man and Nature*. Scribner, New York

Middleton N J 1990 Wind erosion and dust storm control. In: Goudie A S (ed.) *Techniques for Desert Reclamation*. Wiley, Chichester, UK, pp. 87–108

Morgan R P C 1995 *Soil Erosion and Conservation*, 2nd edn. Longman, Harlow, UK

Owen L, Unwin T 1997 *Environmental Management: Readings and Case Studies*. Blackwell, Cambridge, MA

Reij C (ed.) 1991 *Sustaining the Soil. Indigenous Soil and Water Conservation in Africa*. Sustainable Agriculture Programme of the International Institute for Environment and Development, London

Rhoades J D 1990 Soil salinity—causes and controls. In: Goudie A S (ed.) *Techniques for Desert Reclamation*. Wiley, Chichester, UK, pp. 109–34

Thomas W L (ed.) 1956 *Man's Role in Changing the Face of the Earth*. University of Chicago Press, Chicago

Tiffen M, Mortimore M, Gichuki F N 1994 *More People, Less Erosion: Environmental Recovery in Kenya*. Wiley, Chichester, UK

Trimble S, Lund S W 1982 Soil conservation and the reduction of erosion and sedimentation in the Coon Creek Basin, Wisconsin. US Geological Survey professional paper 1234

Turner B L (ed.) 1990 *The Earth as Transformed by Human Action*. Cambridge University Press, Cambridge, UK

Watson A 1990 The control of blowing sand and mobile desert dunes. In: Goudie A S (ed.) *Techniques for Desert Reclamation*. Wiley, Chichester, UK, pp. 35–85

Williams M 1970 *The Draining of the Somerset Levels*. Cambridge University Press, Cambridge, UK

A. S. Goudie

Land Rights and Gender

In largely agrarian economies arable land is the most valued form of property and productive resource. It is a wealth-creating and livelihood-sustaining asset, and for a significant majority of rural households it is the single most important source of security against poverty. Traditionally, it has been the basis of political power and social status. For many, it provides a sense of identity and rootedness within the village, and has a permanence which few other assets possess. In some communities ancestral land also symbolically stands for continuity of kinship and citizenship.

Despite its critical importance, however, the issue of *women's* rights in land (and more generally in property) has until recently largely been neglected in social science scholarship and public policy. In fact, large-scale surveys and agricultural censuses in almost all developing countries collect property-related information by households not disaggregated by gender. At present, we thus have to draw upon small-scale surveys and the numerous village studies: these reveal that typically few women own arable land and even fewer effectively control some. Several reasons (outlined further below) have been put forward on why it is important for women to have effective and independent land rights. But first consider some definitions.

1. Defining Land Rights

Rights (in any form of property) can be defined as claims that are legally and socially recognized and enforceable by an external legitimized authority, be it

a village-level institution or some higher-level body of the State. Land rights can stem from inheritance, transfers from the State, tenancy arrangements, purchase, and so on. They can be in the form of ownership or usufruct (rights of use), and can encompass differing degrees of freedom to lease out, mortgage, bequeath, or sell.

Three additional distinctions are relevant here. First, there is a difference between the legal recognition of a claim and its social recognition, and between recognition and enforcement. A woman may have a legal right to inherit property, but this may remain merely on paper if the law is not enforced, or if the claim is not recognized socially as legitimate. Second, is the distinction between ownership and effective control. It is sometimes assumed incorrectly that legal ownership carries with it the right of control in all its senses. In fact, legal ownership may be accompanied by restrictions on disposal, as among the Jaffna Tamils of Sri Lanka and several communities in Latin America, where a married woman needs her husband's consent to alienate the land she legally owns. Third, we need to distinguish between rights vested in individuals and those vested in groups.

'Effective rights' mean rights not just in law but in practice. Some scholars also stress women's 'independent rights', meaning rights independent of male ownership and control.

2. *Why it is Important for Women to have Effective Land Rights*

Effective and independent land rights for women have been emphasized in recent years on four broad counts: welfare, efficiency, equality, and empowerment (as elaborated in Agarwal 1994).

2.1 *Welfare Implications*

It is well accepted that land access can significantly reduce a household's risk of poverty, but for several reasons land solely in men's hands need not guarantee female welfare.

First, evidence from many countries reveals persistent gender inequalities and a bias favoring males in the distribution of resources within households, including allocations for basic necessities such as healthcare and education and, in some regions, even food. Biases in food and healthcare are revealed especially in anthropometric measures, morbidity rates, and most starkly in female adverse sex ratios (Dreze and Sen 1995).

Second, women without independent resources are highly vulnerable to poverty and destitution in case of desertion, divorce, or widowhood. In parts of South Asia, not uncommonly, even women from rich parental and marital families, deprived of their property shares on widowhood, can be found working as agricultural laborers on the farms of their well-off brothers-in-law. Indeed, widows and the elderly often recognize that entitlement to family care and support can depend critically on whether or not they have property to bequeath (Basu 1999).

In contrast, direct land transfers to women are likely to benefit not just women but also children in greater degree. Evidence from many regions shows that women, especially in poor households, spend most of their earnings on basic household needs, while men spend a significant percentage of theirs on personal goods, such as alcohol, tobacco, etc (Dwyer and Bruce 1988). Studies reviewed by Strauss and Beegle (1996) indicate that in some African countries, in given contexts, the prevalence of child malnutrition is lower among female-headed households compared with male-headed households, even though the latter have higher incomes. In urban Brazil, the effect on child survival probabilities is many times greater from asset income that accrues to the mother than that which accrues to the father (Thomas 1990). Children in rural India are found more likely to attend school and receive medical attention if the mother has more assets (Strauss and Beegle 1996). In other words, women's and children's risk of poverty would be reduced and their welfare enhanced if women had direct access to land, and not just access mediated through male family members.

Land can provide women both direct and indirect benefits. Direct advantages stem from the possibilities of growing not just crops, but trees, a vegetable garden, or grass for cattle. Indirect advantages arise from land serving as collateral for credit or as a mortgagable or saleable asset during a crisis; increasing the probability of women finding wage employment; providing an asset base for rural nonfarm enterprises, and so on. Those with land tend to have substantially greater rural nonfarm earnings from self-employment than the totally landless. In other words, owning even a small plot can be a critical element in a diversified livelihood system, and can improve women's and the family's welfare significantly, even if the plot is not large enough to provide full family subsistence.

These factors are likely to become increasingly important as marital and kin support erodes, and female-headed households multiply.

2.2 *Efficiency Implications*

Secure and more gender-equal land rights would also have implications for productive efficiency. First there is the incentive effect. Although it is widely recognized that security of tenure can be critical for motivating farmers to make productivity enhancing investments in their fields, the need for incentives within the family has received little attention. Some studies suggest that incentives could be as important within families. In

Kenya, for example, the introduction of weeding technology in maize production raised yields on women's plots by 56 percent where women controlled the output, and only by 15 percent on their husbands' plots where too women weeded but men got the proceeds (Elson 1995). Secure land rights for women farmers and their control over the produce could thus prove to be an important element in enhancing production. Second, titles (which serve as collateral in many regions) could help increase output by improving women's access to production credit. This can prove especially critical where women are the principal farmers, as where male outmigration is high, or where widows (or wives) are cultivating separate plots still formally owned by kin. Third, research from Africa suggests that women might use land more efficiently than men in certain contexts. In Burkina Faso, for instance, due to their choice of cropping patterns women achieved much higher values of output per hectare on their own plots than did their husbands on theirs (Udry et al. 1995). Although women's yields for given crops were lower than men's, this was due to their lesser access to inputs such as fertilizers which were concentrated on the men's plots. The study estimated that output could be increased by as much as 10–20 percent if such inputs were reallocated from plots controlled by men to those controlled by women in the same household. In other words, allocative efficiency could improve with more gender equal distribution of inputs. A literature review of the effect of gender on agricultural productivity in several countries of Africa and Asia also concludes that output could be increased notably if women farmers had the same access to inputs and education as male farmers (Quisumbing 1996). Fourth, women are often better informed about traditional seed varieties and the attributes of trees and grasses. This knowledge could be put to better use if they had greater control over farming. Fifth, titles can empower women to assert themselves better with agencies that provide inputs and extension services.

Some oppose women's land claims on the grounds that it will reduce output by reducing farm size or increasing fragmentation. However, there is no consistent evidence of an adverse size effect. In fact, in many regions smaller farms have higher productivity (Lipton and Longhurst 1989, Berry and Cline 1979); and fragmentation can arise equally with male inheritance. Also, where necessary, farmers have dealt with fragmentation in various ways: consolidation through purchase and sale; land leasing arrangements to bring together cultivation units even where ownership units are scattered; joint investment and cultivation by small groups, etc.

While welfare arguments for women's land rights have received some policy attention, there is yet little recognition of the efficiency implications. These are likely to be especially important in regions of high female headedness, or where feminization of agriculture is moving apace as more men than women enter nonfarm occupations. Over time, we can thus expect an increasing proportion of farmers to be female, and women would need special targeting in policies for increasing production and improving food security.

2.3 *Equality and Empowerment Implications*

In the equality argument for land rights, two aspects have been especially emphasized. Gender equality as a measure of a just society, and equality in land rights as a measure of women's economic empowerment. The word 'empowerment' is now used widely in the literature, usually without being defined. One way to define it is would be 'as a process that enhances the ability of disadvantaged ("powerless") individuals or groups to challenge and change (in their favor) existing power relationships that place them in subordinate economic, social and political positions' (Agarwal 1994).

Endowing women with land would empower them economically as well as strengthen their ability to challenge social and political gender inequities. An illustrative example is women's experience in the state of Bihar (eastern India). Here, in the 1970s, after a peasant struggle in which both men and women participated, when women in two villages received land, they said: 'We had tongues but could not speak, we had feet but could not walk. Now that we have the land, we have the strength to speak and walk' (Agarwal 1994). This sense of empowerment accompanying improved land rights also enhances women's ability to assert themselves within the home, in the community, and with the State (the latter for demanding their due in various government schemes, infrastructure, and services). The effect of land access on women's 'bargaining power' within and outside the household has also been theorized (Agarwal 1997).

3. *Sources of Arable Land for Women*

Given its importance, to what extent do women have effective land rights? Consider the three major ways by which women can gain land: inheritance, State transfers, and the market.

3.1 *Land via Inheritance*

In most countries land is largely privatized: in India, for instance, 86 percent is privatized. Inheritance is thus the most important source of land. Typically inheritance is post-mortem, with some rare examples (e.g., from Sri Lanka) of premortem transfers via dowry. Traditionally, three types of inheritance systems prevailed: patrilineal (where ancestral property passed through the male line), matrilineal (where

ancestral property passed through the female line), and bilateral (where ancestral property passed to and through both sons and daughters). Patrilineal inheritance was the most common.

During the twentieth century, in some regions, inheritance laws underwent significant transformation and moved toward gender equality. In other regions age-old and highly gender-unequal inheritance customs continue, or the laws, despite improvement, have yet to provide full gender equality. In sub-Saharan Africa, for instance, most women's land access is mediated by marriage. Women usually receive separate plots for cultivation through their husbands, but tend to lose those usufruct rights on divorce. Only rarely do women have claims through natal kin such as fathers, or through matrilineal inheritance systems (Davidson 1988). In most Latin American countries, by contrast, inheritance laws are relatively gender equal (Deere and Leon 1999). In between come countries such as those of South Asia (where inheritance laws vary by both region and religion): here women can legally inherit from both fathers and husbands, but in most communities their rights are not equal to men's (Agarwal 1994). In addition, there are the transition economies, many of which are still formulating private property laws.

Even where women enjoy legal equality, however, there is a large gap between law and practice. The exact extent of this gap is not known due to the noted dearth of gender-disaggregated data on property, but small-scale surveys and village studies indicate that most women do not inherit what is their legal due. In rural India, for instance, a recent sample study found that only 13 percent of women whose fathers owned land inherited any as daughters, and only 51 percent of those whose husbands owned land inherited as widows (Chen 2000). In Latin America, despite legal equality, daughters are less likely to inherit land, and when they do, they inherit less and poorer quality land (Deere and Leon 1999).

Several social and administrative factors underlie the gap between law and practice. A significant constraint is postmarital residence. Where women are married to strangers in distant villages parents tend to disinherit daughters from fear of losing control over the land. Many women, especially in South Asia, also give up their shares in favor of brothers whom they see as potential sources of economic and social security in case of marital breakup, in the absence of a state social security system. Where women do not voluntarily give up their claims, male relatives may file court cases, forge wills, use threats, and even resort to violence. These constraints are compounded by the approach of many government functionaries who often share the prevalent social biases and obstruct the implementation of laws in women's favor.

The gap between legal rights and actual ownership is often matched by that between ownership and effective control. Marriages in distant villages can make direct cultivation by women difficult. Constraints also stem from illiteracy, high fertility, and social restrictions on women's mobility (as in regions of high female seclusion). These difficulties are compounded by women's limited control over cash and credit for purchasing inputs, and male bias in agricultural extension services.

However, many of the social constraints vary by region. Distant marriages to strangers are more a feature of some groups (e.g., Hindu communities in northern South Asia), than others. Where in-village marriages with close kin are allowed, there is less opposition to women inheriting land. Again female seclusion is not universal and is more a feature of Muslim societies and small pockets of north Indian Hindu communities. More subtle and widespread is the cultural construction of women's behavior and the gendering of social space which restricts women's public participation indirectly.

3.2 Government Land Transfers

A second potential source of land is through State transfers. Typically, however, government land reform and resettlement programs tend to allot land almost exclusively to males. Underlying this is usually the assumption that men are the primary cultivators and breadwinners and women the helpers and dependents. For instance, in the Latin American agrarian reforms carried out between the 1960s and late 1980s, a study of eight countries found that women beneficiaries ranged from negligible in Chile and Peru to 15 percent in Mexico, other countries coming in between (Deere and Leon 1999). Although the allocation criteria appeared gender-neutral, they embodied gendered norms and perceptions that effectively excluded most women. For instance, most schemes required beneficiaries to be household heads (presumed to be male, excepting widows or single mothers); or be permanent wage agricultural workers (who were usually men, women typically being seasonal workers); or to be well educated (where again women were disadvantaged).

In South Asia, similarly, land transfers by governments, whether under land reform or irrigation resettlement schemes, have been almost entirely to men, even in communities that traditionally practiced bilateral or matrilineal inheritance.

Recent policy shifts in most countries have typically taken the form of granting joint titles to both spouses. While this is a step forward in reducing existing gender inequalities, joint titles can also restrict women in gaining control over the produce, or claiming their shares in case of marital conflict or divorce, or exercising freedom in bequeathing the land as they wish.

Gendered assumptions have also guided land transfers following peasant movements. Women have usually been significant participants in such move-

ments but have seldom been recognized as independent claimants. Exceptions are few and far between, such as the earlier-mentioned 1970s Bodhgaya struggle in eastern India, in which women demanded and received independent land shares in two villages (Manimala 1983).

3.3 Lease or Purchase

The third source of land for women is through lease or purchase. This is a limited option, in that individual women usually have much less access than men to monetary resources. Also, rural land markets are often constrained and land is not always available for sale. This option therefore cannot compensate for the gender inequalities in inheritance or government transfers. It can, however, supplement those means, especially if done through collective endeavor. For instance, in parts of South Asia, groups of landless women have been using state-provided subsidized credit for leasing in or purchasing land in groups, and cultivating it jointly. A collective approach to cultivation and management has several advantages over an individual approach. It helps women mobilize funds for capital investment (which they are less able to do individually); provides economies of scale in input use, labor sharing, and marketing; and increases women's self-confidence in demanding technical information and inputs from public agencies (Agarwal 1998).

4. Conclusion

Research on the potential positive impact of effective land rights on rural women and children's well being, women's social and political empowerment, gender equality, and productive efficiency, suggests that concerted efforts by policy makers and grassroots activists are warranted to promote women's land rights. Among the aspects that would need attention are the following: promoting gender-equal inheritance in law and practice; formulating public policies which recognise women as significant farmers, with claims both to land and to infrastructural services such as inputs, credit, and information; initiating gender-disaggregated data collection to monitor shifts in command over landed property; and so on.

Many of these aspects would be relevant for women's property rights in general, even where the importance of arable land as property is declining (or has declined) with industrialization and urbanization.

Equally, efforts will be needed to ensure that existing biases are not replicated in new land reform initiatives and property rights formulations. For instance, agrarian reform is a major policy issue in postapartheid South Africa; and new private property rights in land and other assets are now being constituted in Eastern Europe and the former USSR. Here there are dangers of creating new gender inequalities, as is already happening to some degree (Verdery 1996). But there is also a window of opportunity for turning the tide and taking steps to ensure gender equality in both land and other property.

See also: Divorce and Gender; Domestic Life and Material Culture; Economic Development and Women; Economic Globalization and Gender; Gender, Economics of; Gender Ideology: Cross-cultural Aspects; Household Production; Labor Markets, Labor Movements, and Gender in Developing Nations; Lone Mothers in Nations of the South; Matrifocality; Nation-states, Nationalism, and Gender; Poverty and Gender in Developing Nations

Bibliography

Agarwal B 1994 *A Field of One's Own: Gender and Land Rights in South Asia*. Cambridge University Press, Cambridge, UK

Agarwal B 1997 Bargaining and gender relations: within and beyond the household. *Feminine Economics* **3**(1): 1–51

Agarwal B 1998 Disinherited peasants, disadvantaged workers: a gender perspective on land and livelihood. *Economics Political Weekly* **33**(13): A2–A14

Basu S 1999 *She Comes to Take Her Rights: Indian Women, Property and Propriety*. State University of New York Press, Albany, New York

Berry R A, Cline W R 1979 *Agrarian Structure and Productivity in Developing Countries*. Johns Hopkins University Press, Baltimore, MD

Chen M 2000 *Perpetual Mourning: Widowhood in Rural India*. University of Pennsylvania Press, Philadelphia, PA

Davidson J (ed.) 1988 *Agriculture, Women and Land: The African Experience*. Westview Press, Boulder

Deere C, de Leon M 1999 *Gender, Property and Empowerment: Land, State and Market in Latin America*. Department of Economics, University of Massachusetts, Amherst, MA

Dreze J, Sen A 1995 *India: Economic Development and Social Opportunity*. Oxford University Press, Delhi, India

Dwyer D, Bruce J (eds.) 1988 *A Home Divided: Women and Income Control in the Third World*, Stanford University Press,Stanford, CA

Elson D 1995 Gender awareness in modeling structural adjustment. *World Development* **23**(11): 1851–68

Lipton M, Longhurst R 1989 *New Seeds and Poor People*. Johns Hopkins University Press, Baltimore, MD

Manimala 1983 Zameen Kenkar? Jote Onkar! Women's participation in the Bodhgaya land struggle. *Manushi* **14**: 2–16

Quisumbing A R 1996 Male–Female differences in agricultural productivity: methodological issues and empirical evidence. *World Development* **24**(10): 1579–95

Strauss J, Beegle K 1996 Intrahousehold allocations: a review of theories, empirical evidence and policy issues. *MSU International Development Working Paper No. 62*. Department of Agricultural Economics, Michigan State University, East Lansing, MI

Thomas D 1990 Intrahousehold resource allocation: an inferential approach. *Journal of Human Resourses* **24**(4): 635–63

Udry C, Hoddinott J, Alderman H, Haddad L 1995 Gender differentials in farm productivity: Implications for household efficiency and agricultural policy. *Food Policy* **20**(5): 407–23
Verdery K 1996 *What Was Socialism, and What Comes Next?* Princeton University Press, Princeton, NJ

B. Agarwal

Land Tenure

1. Introduction

1.1 Scope and Emphasis

Land tenure, connecting to much-debated issues of human territoriality and sovereignty, can be approached at any level of scale, from domestic to cosmic. As presented in this article, land tenure moves beyond the confines of conventional English legal terms like 'rights and duties' and political-economic ones like public or private 'property' or 'ownership.' A wider array of concepts, taking real access, attachment, and mobility into account, is needed. Visas and border blockages matter no less than plot markers or title deeds. Terms and translations—for people, things, and links between—are critical throughout.

Land can refer to space or place, often denoting material substance deemed immovable, and sometimes connoting belonging. In some tongues it takes several distinct words to translate, and corresponding terms may or may not include water, plants, animals, people, and built or dugout structures. 'Tenure' (from L. *tenere*, to hold) or landholding need not refer just to a particular form of legal, political, or economic regime (such as freehold, fiefhold, or collective village ownership) but can indicate human-land attachment in a broader sense, taking account of unofficial, illegal, or occasional use, and of persons living, dead, or unborn.

1.2 Sovereignty and Strife

'Nearly all social struggles in Antiquity,' claimed Max Weber, 'were essentially for the ownership and use of land' (1988, p. 343). He might say the same for recent times. Particular regions have hallmark forms of land tension and strife, simplified and sharpened in mass media if not also in scholarly representation. Such are ethnoreligious struggles over autochthony in the Levant and Northern Ireland; forced transmigration in Indonesia; latifundio, minifundio, squatting, and private militias ('goon' squads) in Hispanic America; small-tribe survival struggles in Amazonia; landlord-tenant exploitation in South Asia; plantation oppression in the Caribbean and island southeast Asia; collectivization and decollectivization in Russia and China and states in their hegemonies; farmer-herder conflicts in the American west or the Sahel; individual titling that threatens women, children, and the rural poor in the African tropics. But none of these regions monopolizes the tenure issue for which it is famed. Tasks for social scientists and humanists include finding particularities and exceptions, bringing to light indigenous practical solutions, and assessing the performance of currently sovereign authorities.

2. Historical Background

Records of land claims and transfers have been dug up among some of the earliest known writings in Mesopotamia from four to five thousand years ago, along with royal edicts canceling subjects' debts and reversing land forfeitures to creditors. Plato's and his student Aristotle's debate about the merits of public versus private landholding in the Aegean *polis* echoes two millennia later in European and American traditions. John Locke, Adam Smith, and Herbert Spencer are still widely cited by private property's upholders, while the young Jean-Jacques Rousseau, Karl Marx, Pierre Proudhon, and (in America) Henry George are champions of its critics.

2.1 Evolutionist Political Philosophy

The spread of European imperial power and the westward expansion of Euro-American settlement in the late nineteenth and early twentieth centuries, accompanied as they were by wholesale expropriation or disenfranchisement of foragers, mobile herders, and swidden (slash-and-burn, long fallow) farmers, led sociologists and anthropologists into the study of land tenure as their disciplines coalesced. Jurist Henry Maine's vision of Roman property law's and human history's evolution from ascribed 'status' to achieved 'contract' was qualified by lawyer-anthropologist Lewis Henry Morgan's conclusion in *Ancient Society* (1877), that individual property was an inadequate basis for civilization's further progress. Societies moved, said Morgan, from (L.) *societas*, i.e. ties based on kinship and other particular ties, to *civitas*, ties based on territorially defined polity, but this did not always mean upward progress either. Karl Marx and Friedrich Engels, inspired by Rousseau and Morgan among others, envisaged history as a one-way flow through stages of prefeudal tribalism, feudalism, capitalism, socialism, and finally a future communism with abolition of private property and state government.

Evolutionist theories based on a largely imaginary prehistory and assumptions of unilinear change continue to resound in legal, political, and economic

circles. But twentieth-century anthropology has made it more apparent how tenure change can zigzag and backtrack, and how many different forms of tenure can be 'modern' at once, sometimes in close co-existence.

2.2 Persons, Things, Systems

Early ethnography in Australia and Native America, and Bronislaw Malinowski's work on Trobriand Island gardening in Melanesia in the 1910s, revealed indigenous concepts in which human groups and spaces (however defined) were deemed aspects of each other, often inextricable from spiritual beings or forces. Research in colonial Africa in the mid-twentieth century depicted how land tenure tied tightly into cognition, language, and translation; kinship and friendship; political control and hegemony; religion and ritual; aesthetics and myth (Biebuyck 1963). Monarchs' and chiefs' control over land (as in Buganda or Ashantiland) appeared to be part and parcel with control over people, but their rights excluded alienation by sale and other rights often identified with ownership. Landholding was justified by clearing, residing, or working on the land, by conquest, by membership in a kin or local group, by chiefly assignment and tribute, or a mix of these. Multiple rights of different groups to use land for different purposes, for example farming, grazing, and thatch, firewood, or clay collection, overlapped in complex ways not captured by terms like 'property' or 'ownership.' In 'acephalous' (politically uncentralized) societies like those of Tiv in Nigeria or Chagga in Tanzania, land territories were often found closely linked to genealogy of clans or lineages, shifting and stretching as these moved, grew, or shrank. Here and elsewhere in Africa, land was spoken of as an active interest not only of the living, but also the dead and unborn. To Audrey Richards, Lucy Mair, and other British-trained social scientists, land tenure appeared as part of complex, functioning cultural systems. Max Gluckman's southern Africa-based work on 'hierarchies of estates' over land, for instance of use, transfer, and administration, influenced a generation of tenure studies. Gluckman (1965) observed that property is not just about rights of people over things, but also about duties between people with respect to things.

2.3 Structure and Process of Inequity: Critical Rethinkings

In the 1960s and early 1970s, Latin American-inspired scholarship on dependency and 'modes of production,' in a Marxist vein, criticized the exploitation of rural and tropical peripheries by metropolitan capitals through agribusiness or plantation owners or the co-optation of comprador (buyer) elites. South America, the continent with the longest history of European colonialism, had, some observed, the most uneven distribution of landholdings. Parisian Marxist structuralists in the late 1970s revised the view that capitalism eroded all other forms of tenure and exchange (quasi-feudalism, family farmholding, etc.), contending instead that it encapsulated and preserved these for its own profit.

In the 1980s and 1990s social scientists with political-economic interest turned their attention to attempts by national regimes to monopolize power over land relations, particularly in Africa. State attempts to nationalize, privatize, or socialize landholding now looked like raw or subtle power grabs (Downs and Reyna 1988, Le Bris et al. 1991). Large-scale resettlement efforts, as in Ethiopia or Indonesia, seemed to some analysts like attempts to remove or disorient potential dissidents. Attempts to 'villagize' scattered farmers or sedentarize mobile herders appeared as attempts more easily to tax, conscript, and regulate them (Scott 1998). Postmodernist and post-structuralist scholars, meanwhile, eschewed the notion of timeless tenure 'systems,' preferring to see complexity, fragmentation, indeterminacy, negotiability, contestability, and historical contingency, in human-land as in human-human interaction. In town and country, they criticized tenure regimes based on race, class, and gender.

Toward the turn of the millennium, a plurality of newly empowered indigenous voices in southern countries helped force scholars and jurists concerned with land tenure everywhere to reassess their most basic terminology and philosophical premises. Assumptions of human dominion over land, of individualistic rights and duties, and of the sanctity of arbitrarily drawn borders and boundaries that blocked human movement, all came under criticism.

3. Metaphor and Rhetoric

Human communication about land draws heavily on metaphoric imagery from the human body (headwaters, heartlands, fertility, etc.), kinship (mother-, fatherland), domicile (gateway to Asia), and other human-centered sources. Such metaphorizing imbues poetry and politics, while imagery of 'grounding' or 'terra firma' imbues legal and scientific discourse. English terms like 'property,' 'rights,' or 'ownership' sometimes carry culture-bound overtones of civilization, justice, or morality. To speak of land as a commodity denies psychological attachments to space and place. To speak of land as a 'resource' implies it exists to serve human use. Between people who assume that humans can own land and master it, and those who assume on the contrary that humans belong to

the land and must respect it and conform to its requirements, communication easily goes awry.

People who derive their livelihood from land in different ways perceive and communicate about it in different geometric and geological idioms. Whether in the Andes, the Sahel, or Lapland, transhumant herders who must focus on linear access routes tend to clash with farmers who focus on fields along the way, when animals eat or trample crops. Seasonal and tidal wetlands and mudflats (as in the Niger delta, in the Sudanese Sudd, or along the Senegal River) are frequently contested by farmer-fishers, ecologists, and developers using different terminology.

4. *Land and Tenure Reform*

While land reform, in English, conventionally refers to redistribution of holdings, land *tenure* reform refers to changes in the more basic rules and understandings by which land is occupied, used, or transferred. Among many kinds of tenure reform are proclamation of sovereignty or eminent domain, implementation of titling (i.e. registration), and change in rules about transfer (loan, rental, sale, inheritance, etc.). Dozens of twentieth-century national tenure reforms sought to individualize or collectivize landholding.

Their histories are full of ironies and surprises. Frequently, reforms aimed at improving distribution of holdings have ended up worsening it, whether because of unequal information, implementer bias, deceit, or reneged deals. Moreover, reforms aimed at boosting production have often reduced it instead, for instance because new owners or occupants lacked farming experience or acquired the land only for speculation. Land tenure is so tightly bound up with kinship, religion, and other aspects of culture that a newly imposed tenure regime seldom expunges what was there before. More often, it just crowds new procedures together with old, allowing the wealthy, powerful, astute, or well connected to gain at others' expense by manipulating confused or conflicting principles.

There are exceptional cases of successful land and tenure reform. One celebrated example is Taiwan's land redistribution in 1953, swiftly and efficiently implemented after a period of Japanese occupation had weakened local resistance. Attempts to emulate it elsewhere have seldom succeeded.

4.1 *Bureaucratic Simplification*

Bureaucratic administration, tending to push toward abstraction and universalism in its publicly professed goals, simplifies complex realities of human culture and ecology, sometimes beyond recognition, to make people or spaces manipulable for taxation, military conscription, etc. Empty spaces on maps, or designations like 'vacant' or 'underproductive,' help justify land grabs, disfavoring swidden farmers, transhumant herders, foragers, and other occasional users. Titles issued to individuals ignore claims of dependents. But subjected people may thwart imposed tenure regimes, for instance by continuing to transfer fields without informing state authorities, as happened in Kenyan and other African cases of titling reforms over the second half of the twentieth century. Real and paper tenure diverge further.

Nowhere do real people live their lives consistently by capitalist or socialist principles, whatever they may think they do. In agropastoral Africa, for instance, land farmed by families or individuals may be expected to be opened for grazing by the animals of some broader public after harvest. Property shifts between public and private, or remains subject to multiple tenure 'systems' simultaneously.

4.2 *Collectivization and Villagization*

Collective ownership and production have been part of utopian visions since Plato's *Republic* or before, and many different attempts have been made to bring these into reality in monasteries or other single- or mixed-sex religious or secular communes, sometimes modeled upon families. Mexico's *ejidos* after the 1910 revolution, and Israel's kibbutzim more recently, have been closely watched by admirers and critics alike. Marxist-inspired attempts to collectivize landholding under state aegis in the Soviet Union between 1917 and 1931, and in China after 1949, have been emulated in various parts of the world, notably eastern Europe, Vietnam and Cambodia (Kampuchea), Cuba, Tanzania, Ethiopia, and Lusophone African countries including Angola and Mozambique.

While late twentieth-century Chinese history has shown that some forms of state-organized collective farming can raise productivity, improve distribution of food and other vital resources, or both, if starting from a feudal or quasi-feudal baseline, people in few parts of the world have been able to emulate socialist tenure reforms with lasting success. Commonly, work-incentive problems get in the way, and local populations chafe under centralized bureaucratic control of their lives.

Notable failures have been attempts to concentrate rural people into villages where the land could not sustain them, as in Tanzania's compulsory villagization from 1973 to 1976 (an experiment duplicated in Ethiopia a few years later). Similarly, savannah borehole well projects, for instance in Botswana or across the Sahel, have led to overconcentration of animals and overgrazing. Two policy conclusions on herders and farmers are clear enough: (a) people who live spread out, or mobile, usually have good reasons

for wanting to stay that way; (b) good intentions, in land tenure reforms, are about as likely to produce harmful as helpful results.

4.3 Titling and Individuation

Another kind of tenure reform is the titling or registration of individual claims for the first time. 'Private' property is often used to mean individual property, but the two need not be the same. State and local governments and landholding companies have several ways of registering land titles and transactions. Among the most commonly applied in former British colonies is the Torrens system, devised in Australia and New Zealand in the 1860s, where it allowed grid allotments for new settlers.

Where people already live, titling land in the names of individuals, under state guarantee, does not necessarily increase security of tenure as often assumed. It may decrease it. Instructive was the General Allotment Act (or Dawes Act) of 1887, by which the US government allowed for the awarding of individual land titles to Native Americans (or First Peoples). Between 1887 and 1932, about two-thirds of Native American lands were lost to immigrant settlers and speculators, who were not slow to exclude its former inhabitants and users.

British, French, and Belgian colonial administrations appear to have paid this story little heed in launching land titling programs in their colonies. Early twentieth-century titling initiatives were expanded to nationwide scale in the Kenyan model established under the Swynnerton Plan in 1954 and passed into law in 1957. That model, which still continues, was being emulated elsewhere in Africa at the turn of the millennium.

At least as often as not, land titling in new states and perhaps old seems to speed the concentration of holdings into fewer hands. Individuals newly freed of social constraints on selling, mortgaging, or pledging erstwhile family or clan land do not always make wise decisions. Speculators and accumulators use urban capital and rural off-farm incomes to buy out disadvantaged rural people in periods of drought, plague, or unfavorable prices.

The mortgage, a deadlined pledge linking land or other property to credit, plays a key part in this process. Borrowers who fail to repay by a specified date forfeit their land to the lender or lender's agent. The mortgage appears regularly in histories of rural dispossession, not least in England, where law about it has changed many times, and in North America. While seemingly vital to urban and industrial economic life, in rural areas mortgage loans are more treacherous to farming people, particularly where rainfall, infrastructure, and crop markets are unreliable. Although rich like poor may lose lands, mortgage foreclosure contributes heavily to class stratification. Inland tropical settings often lack sufficient industrial or commercial opportunity to absorb newly landless or land-poor people. Even where land titling is accepted, mortgaging is morally debated in many cultural and legal traditions. *Shari'a*, the sacred law of Islam, forbids mortgaging as sin.

When, and whom, is titling likely to help? Threatened indigenous minorities frequently seek titles to defend their home territories—and by extension, their livelihood and way of life—from more numerous and powerful outsiders who threaten to dispossess or displace them. These may include fortune seekers like ranchers and rubber tappers, as in Amazonia or Malaysia; homesteaders and land speculators, as in the American west; planters, as in the Caribbean; or politicians and their favorites, as in equatorial Africa.

Land titling is never a politically neutral process of recording information. Usually someone gains and someone loses. Gainers are often richer, better-connected, and better-schooled people; erstwhile land borrowers or land clients, who may become owners; and people who emigrate or seek to speculate in land without having to occupy or farm it. Losers include ancillary rightholders and dependents, and members of fast-reproducing families. Just as rural people have circumvented or resisted state directives to villagize or collectivize their landholding under socialist policies, as in Tanzania or Ethiopia, they have reasserted the rights of collectivities—clans, lineages, age sets, or voluntary associations—under individualistic capitalist ones, as in late twentieth-century Kenya or Côte d'Ivoire.

4.4 Fragmentation and Consolidation

While economists and agronomists have often favored large consolidated blocks of land for mechanization and economies of scale, anthropologists and ecologists, observing their subjects' resistance, have found that fragmented holdings may make sound ecological and economic sense in some settings. Unreliable rainfall, acute pest problems, and dicey markets encourage a cultivator to keep a hand in diverse cultivars and microclimates to avoid a total loss. Maintaining multiple holdings also helps conceal wealth from neighbors or tax collectors.

4.5 Irrigated Lands

Nowhere is control over land more hotly political than in irrigated areas, where central authorities may seek to dictate settlement, land and water allocation, inheritance, etc. The centripetal draw of high farm yields in irrigation schemes is sometimes counteracted by the centrifugal push of struggles over control over community and family life, as in the Jahally Pachar

rice scheme in the Gambia. Where flood control is a main concern, as in much of irrigated China (or Bangladesh), bureaucratic polities tend to be highly centralized and authoritarian. Where it is not, as in much of Africa, regimes attempting to emulate such models have enjoyed little popular support, often proving fragile and short-lived—particularly where ethnic mixture compounds the adjustment problems of living in close, compulsory synchrony.

All irrigation but the smallest scaled involves inequalities in access to water. The privilege of upstream location in the trench network, implying superior control, often coincides, as cause or effect, with personal or familial wealth or power.

4.6 *Displacement and Resettlement*

Development programs like irrigation schemes and hydroelectric dams; military travel and war; and environmental hardships like droughts and floods all entail resettlement of large numbers of people. If the land of destination is unoccupied, usually there is something undesirable about it. If it is occupied, then someone has to move over and friction is likely. Late twentieth-century anthropological studies of resettlement, notably by Elizabeth Colson, Thayer Scudder, and collaborators among Gwembe Tonga in Zambia (formerly Northern Rhodesia), showed how resettled people adjust to extreme stress and trauma in patterned and predictable phases. Among their problems are the shaken confidence in whatever leaders are seen to have allowed the misfortune to happen. Helping or allowing communities or neighborhoods to move intact, in so far as their members wish, seems to mitigate the ill effects by keeping at least their social 'landscape' familiar. But no kind of human 'development' intervention is harder to orchestrate, or more prone to produce catastrophe, than forcing resettlement.

4.7 *Agrarian Encroachments*

Many contemporary 'foraging' peoples reside in difficult environments not of their choosing, for instance in southern and central Africa and northern Canada, where 'race' discrimination has been a factor. Attempts to use these peoples as windows into human prehistory—to see how hard people worked, how peacefully they lived, or how well they cooperated—have often led to spurious conclusions.

Foraging, swidden farming, and mobile herding peoples worldwide suffered a varyingly steady diminution of their lands, and sometimes severe dislocation, from before the eighteenth until the late twentieth centuries, often in episodes where racism played a part. The history of Native American foraging and farming peoples after Columbus—a story of military conquest, deceit, misrepresentation of consent, and treaty reneging by Euro-Americans—hit one of its low points in President Andrew Jackson's Indian Removal Act of 1830, which forced Cherokee, Seminole, and other settled agricultural Native Americans from their southeastern homes to Oklahoma, at the loss of thousands of lives. In another kind of extreme case, Australian law did not recognize the rights of indigenous (or Aboriginal) peoples to land until 1976. In the twentieth century, swidden farmers in tropical and subtropical zones worldwide faced serious threats to their livelihoods under encroachment from sedentary farmers, agribusinesses and other concerns, and under their own population growth. Loss of land for these peoples, as for foragers and pastoralists, has often meant disruption of culture and society in myriad, sometimes indirect ways. The new century and millennium present the challenge, if not of reversing this history, of allowing dignified livelihood to persons constricted or displaced.

4.8 *Reserves and Exclusion*

Official attempts to fence off wilderness reserves have achieved checkered results, particularly in poorer regions and countries in Africa south of the Sahara and in mainland Southeast Asia. One clear lesson from these attempts is that foragers, herders, or swidden farmers excluded from accustomed lands will tend to subvert the new classification unless somehow included in rewards accruing from tourism, mining, or other commercial exploitation of new parks or reserves.

5. *Advocacy and Practice*

The activist international defense of land rights of threatened indigenous minorities, particularly in sparsely settled regions pressed for new resource exploitation by outsiders (Amazonia, Alaska, central Africa, northern Australia, Irian Jaya, and elsewhere), crystallized in the Copenhagen-based International Work Group for Indigenous Affairs in 1968, London-based Survival International, in 1969, and Cambridge, Massachusetts-based Cultural Survival, Inc., in 1972, as well as other human rights defense groups. A worldwide movement of indigenous human rights groups, mutually emboldening, has raised public consciousness and achieved changes in national legislation like the Alaska Native Claims Settlement Act in 1971, Australia's Aboriginal Land Rights (Northern Territories) Act in 1976, and from Canada, the demarcation of Nunavut at the end of the century as a largely autonomous territory for First Peoples. Still, the International Labor Office's Convention 107,

claiming that 'the right of ownership, collective or individual, of the members of the populations concerned over the lands which these populations traditionally occupy shall be recognized,' and the World Bank's tribal peoples policy issued in 1982, have mainly been honored in the breach (Bodley 1999). Participating in the active design of constructive tenure processes along lines more faithful to indigenous systems and aspirations, and showing respect to indigenous terminology, remains a challenge.

6. *New Frontiers*

As social scientists and humanists look beyond simplistic dichotomies like 'public vs. private' (or collective vs. individual), or 'rights and duties,' they perceive many yet unstudied forms of tenure. These may offer important solutions to work-incentive problems typical of socialism, on the one hand, and maldistribution problems typical of capitalism on the other. Other terms, concepts, and policy frameworks are needed that respect access, stewardship, and sentiments of attachment, and that allow for human population flows between territories.

Shifting attention between micro and macro scale, geographers, archaeologists, and others have sought to connect their work on the ground with findings newly accessible from 'remote sensing,' of landforms and land occupation and use, from air and space. Rural land tenure remains better studied than urban and peri-urban, and ground tenure better than water tenure. Border and boundary issues are land tenure issues on a large scale. Problems of land quality and distribution can be addressed not only by altering tenure regimes within nations, but also in allowing people to move between nations or continents more freely. Land tenure is not just about getting in, but also about getting out.

A challenge for social and biological scientists alike is to connect the understandings about territoriality observable from human action on the land, with understandings becoming obtainable through ethology and the electrochemical study of the human brain. Human biology has yet produced few if any conclusive findings about territorial instincts in humans, but enough species exhibit recognizable social and cultural patterning in their territorial relations to make it seem unlikely that human spatial perceptions and attachments result from uniquely human social or cultural processes alone.

Meanwhile, as more languages have become involved in the academic study of land and tenure, the reexamination of the most familiar, but deeply cultural and potentially exotic, concepts and categories—like 'land,' 'property,' and 'rights'—has continued.

See also: Agricultural Change Theory; Agriculture, Economics of; Irrigation Societies; Property: Legal Aspects; Property Rights; Rights: Legal Aspects; Rural Sociology.

Bibliography

Agarwal B 1994 *A Field of One's Own: Gender and Land Rights in South Asia*. Cambridge University Press, Cambridge, UK

Biebuyck D (ed.) 1963 *African Agrarian Systems: Studies Presented and Discussed*. Oxford University Press, Oxford, for the International African Institute, London

Bodley, J H 1999 *Victims of Progress*, 4th edn. Mayfield Publishing, Mountain View, CA

Downs R E, Reyna S P (eds.) 1988 *Land and Society in Contemporary Africa*. University Press of New England, Hanover, NH

Gluckman M 1965 *The Ideas in Barotse Jurisprudence*. Yale University Press, New Haven, CT

Hann C M 1998 *Property Relations: Renewing the Anthropological Tradition*. Cambridge University Press, Cambridge, UK

Ingold T, Riches D, Woodburn J (eds.) 1988 *Hunters and Gatherers: Property, Power and Ideology*. Berg, Oxford, UK

Lawrence D L, Low S M 1990 The built environment and spatial form. *Annual Review of Anthropology* **19**: 453–505

LeBris E, Le Roy E, Mathieu P (eds.) 1991 *L'Appropriation de la Terre en Afrique Noire: Manual d'Analyse de Décision et de Qestion Foncières*. Editions Karthala, Paris

Malinowski B 1935 *Coral Gardens and their Magic: A Study of the Methods of Tilling the Soil and of Agricultural Sites in the Trobriand Islands*. George Allen and Unwin, London

Malmberg T 1980 *Human Territoriality: Survey of Behavioural Territories in Man with Preliminary Analysis and Discussion of Meaning*. Mouton, The Hague, The Netherlands

McCay B M, Acheson J M (eds.) 1987 *The Question of the Commons: The Culture and Ecology of Communal Resources*. University of Arizona Press, Tucson, AZ

Meek C K 1949 *Land Law and Custom in the Colonies*, 2nd edn. Oxford University Press, London

Morgan L H 1877 *Ancient Society*. Henry Holt, New York

Netting R McC 1993 *Smallholders, Householders: Farm Families and the Ecology of Intensive, Sustainable Agriculture*. Stanford University Press, Stanford, CA

Orlove B S, Custred G (eds.) 1980 *Land and Power in Latin America: Agrarian Economies and Social Processes in the Andes*. Holmes and Meier, New York

Rousseau J-J 1965 *Discours sur l'Origine et les Fondements de l'Inégalité parmi les Hommes* (*Essay on the Origin and Foundations of Human Inequality*). Gallimard, Paris

Scott J C 1998 *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT

Shipton, P 1994 Land and culture in Tropical Africa: soils, symbols, and the metaphysics of the mundane. *Annual Review of Anthropology* **23**: 347–77

Tuma, E H 1965 *Twenty-Six Centuries of Agrarian Reform: A Comparative Analysis*. University of California Press, Berkeley, CA

Weber M 1988 *The Agrarian Sociology of Ancient Civilizations*, trans. Frank R I. Verso, London

P. Shipton

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Land Use and Cover Change

The rather ungainly term 'land use/cover change' is fitting, for it conveys the conjoint nature of this branch of human–environment research. This article considers the merging of scientific perspectives that has been necessary to understand the dynamics of human land-use practices and the social and natural environments that shape, and are shaped by, those practices. This endeavor is aimed at estimating past human impacts on the earth, with a view to projecting possible future impacts.

1. *Land Use and Land Cover*

While closely related, the terms land use and land cover refer to quite different dimensions of the earth's surface that have largely been investigated in distinct scientific traditions. Land cover refers to the biophysical attributes of the earth's surface and immediate subsurface. This includes vegetative layers (both natural and artificial), exposed soil or rock, and ice, as well as human constructions such as buildings and pavement. While some land covers are considered pristine or at least minimally touched by the human hand (e.g., remote tropical rain forests), others are recognized as transformed or radically altered (e.g., urban areas).

Land use, meanwhile, refers to the human intent to which the land surface is put. While use registers as an imprint on a given location in terms of the change in land cover that would otherwise exist, it also includes the motivations, constraints, and enabling conditions that give rise to the use. Land cover and land use each imply both state and process variables, and have multiple and overlapping properties whose appearance depends to a great degree on the particular ways in which they are observed.

Because of the compound nature of the questions relating human land use practices to land cover states and changes, addressing them involves an amalgam of analytical perspectives spanning, for example, the social and physical sciences, quantitative and qualitative approaches, and academic and policy concerns. Each of the research traditions involved brings its own theoretical perspective and modes of inquiry, and their coupling is not always easy. In particular, the need to merge locally specific explanations of how and why people use land in particular ways, with broader patterns observed at regional or global scales brings into sharp relief a long-standing conundrum of scientific inquiry—scale dependence: the dynamics of land use/cover changes vary depending on the spatial and temporal scale at which they are observed (see *Local–Global Linkages: Environmental Aspects*). Nevertheless, the enterprise is bound together by the desire to understand how biophysical conditions, particularly 'initial' land cover, shapes human activities (opportunities and constraints), and how these activities, or land uses, alter these pre-existing conditions, thereby shaping future livelihood options.

2. *Human–Environment Relations in Land Use/Cover Change*

The relationship between humans and the natural world appears to be a universal concern of Western thought, occupying a central place in Greco-Roman, Arab, and European mythology and philosophy (see *Human–Environment Relationships*). Modern attempts to characterize the impact of human activities on the earth include George Perkins Marsh's seminal work *Man and Nature* (1864), which has been updated by major multidisciplinary efforts, including *Man's Role in Changing the Face of the Earth* (Thomas 1955) and *The Earth as Transformed by Human Action* (Turner et al. 1990). During the 1990s, the public discussion over the threat of human-induced alterations to the earth's bio-geochemical cycles (e.g., 'global warming') has shifted from speculation to skepticism and finally to widespread acceptance and debate over the details of where, when, and with what consequences such changes may occur.

The bulk of global change science remains dedicated to improving our understanding of the functioning of the earth's climate system, for example through the modeling of carbon dynamics (see *Climate Impacts*). However, as the biophysical systems are better understood, greater attention has been focused on the human dimensions, including the major anthropogenic sources of global change, the consequences of that change in terms of the sustainability of livelihood systems, of the vulnerability of particular groups of people and places, and of the societal perceptions of and responses to change (Kasperson et al. 1995; see *Human Dimensions of Global Change*). These concerns require understanding the human systems—economic, political, cultural, and sociotechnical—associated with human activities at particular places and times (Moran 2000).

Many human activities are known to cause significant disruption of global bio-geochemical cycles, particularly the combustion of fossil fuels. The modification of land cover is also a significant anthropogenic source of bio-geochemical disruption, especially the replacement of forest with grasslands and croplands, and the application of chemical fertilizers and pesticides to croplands. Other important questions driving global change research concern the reduction of biological diversity, both directly through species extinction resulting from human activities, and indirectly through the destruction of habitat (see *Mass Killings and Genocide, Psychology of*). Here again a solid understanding of land use/cover dynamics is required.

These concerns are of more than academic interest; policy is actively being crafted to limit human impacts on the environment at both regional and global scales. The effectiveness of environmental laws and agreements depends on sound scientific understanding of the impact of land use practices, in order to target critical areas and to monitor compliance. Accordingly, research programs have been developed at national and international levels, including the Land Use Cover Change (LUCC) Project, under the aegis of the International Geosphere-biosphere Programme and the International Human Dimensions Programme (Lambin et al. 1999). Land use and cover change research constitutes a bridge between the understanding of the biophysical aspects of global change and the human dimensions of sustainability (see *Sustainability Transition: Human–Environment Relationship*).

Two fundamental, linked, sets of questions underlie land use/cover change research: the first concerns the nature, location, and rates of anthropogenic modification of land cover; while the second seeks to understand the reasons behind land-use practices leading to such modifications. The first set of questions is generally examined at relatively broad scales, often using satellite imagery to determine spatial and temporal patterns of change, while the explanation of land use practices requires local-scale research, and often involves interviews with land managers and direct observation of land management practices. As answers to these two sets of questions take shape, it becomes possible to model the trajectories of change that have been observed, and to project future land-cover outcomes.

3. *Land Cover Change*

Despite having been studied for more than a century, land cover change is still poorly understood, but recent technological and methodological advances show great promise. Two advances in particular bear noting: the ability to observe the current state of the earth's surface from space, and the ability to reconstruct past land-cover patterns through the techniques of the paleo- and historical sciences.

Remote sensing of the earth refers to the use of instruments carried on aircraft and earth-orbiting satellites that detect energy reflected or emitted from the earth's surface in different portions of the electromagnetic spectrum. The land cover of most inhabited portions of the earth has been periodically recorded through aerial photography since the 1950s, facilitating the production of land-cover maps. Since the 1970s, satellite remote sensing data have provided more frequent views of large areas (see *Remote Sensing and Geographic Information Systems Analysis*).

Land-cover research has utilized two sets of satellite imagery, though the distinction between them is disappearing. One set of satellites (e.g., the Landsat, SPOT, and ERS platforms) provides relatively high spatial resolution, sampling the earth's surface on the order of tens of meters, along swathes of 100–200 km. While such imagery is capable of distinguishing several types of land cover, the satellites take weeks to observe the entire surface of the earth, and views of land cover are often inhibited by cloud cover or other atmospheric conditions, especially in the tropics.

Nevertheless, several efforts have achieved reliable views of land cover from high-resolution imagery at continental, or subcontinental, extents by creating mosaics of images (e.g., North America, Europe, Amazonia, and East Africa). Most of the focus to date has been on those areas where transformation of the earth's surface has been most dramatic in recent times—the tropics. The clearance of forestlands has garnered special attention, both because of the potential impact on climate change and biodiversity, and because of the relative ease with which this particular form of land-cover conversion can be detected through remote sensing. Estimates of rates of deforestation are becoming more reliable, and researchers have identified so-called hotspots, where habitat destruction—usually deforestation—is most rapid and biodiversity therefore threatened (see *Deforestation–Forestation*).

Another set of remote sensing systems, primarily intended for the observation of weather, has been acquiring daily views of the earth for over a decade. These images can be compiled to create cloud-free images for ten-day or monthly intervals. These data sources typically suffer from poor spectral and spatial resolution, however advances in processing this rich time series are yielding unique information on seasonal and longer-term climatic patterns, such as El Niño. Separating these natural cycles and rhythms is crucial to the reliable detection of anthropogenic land-cover change. Over the last few years of the twentieth century, significant achievements have been made in the development of global land-cover databases. Global land-cover maps have recently improved in spatial resolution by about an order of magnitude, with reliable databases now produced from data from a single year of observation at a spatial resolution of about 1 km. New satellite systems promise to improve this spatial resolution by a factor of four in the near future, with spectral definition better tuned to land cover change detection. These land-cover data are complemented by similar, related global databases at similar spatial resolutions concerning, for example, human population density and the incidence of fire. (Reviews can be found in special issues of the *International Journal of Remote Sensing* **21** (2000, no. 6/7) and *Photogrammetic Engineering and Remote Sensing* **65** (1999, no. 9).)

The detail and reliability of the information provided by both sets of sensors are constantly being enhanced through improvements both in the instruments themselves and in the technology and skill with which the data are processed. As the data from the

various satellite systems are better integrated, the gap will close between the global coverage and local case studies using aerial photography and high-resolution imagery. Attempts at such integration have highlighted the disparities between analytical schemes for characterizing land cover, as well as the difficulties of merging information about biophysical and social dimensions of land-use/cover dynamics (National Research Council, NRC 1998).

In addition to the very detailed record of land use and related land-cover impacts provided by remote sensing over the second half of the twentieth century, our knowledge of prior land-use impacts is also growing through the analysis of pollen in sediment cores and other paleo-ecological and archaeological techniques. Such techniques have enabled the estimation of the earth's land cover 6000 years ago (GAIM 2000), and detailed reconstructions are being made for the past three centuries by projecting current land cover back in time using historical agricultural and demographic information. The convergence of remotely-sensed and other historical sources with prehistoric reconstructions will provide a much fuller picture of the earth's land cover throughout the time of significant human impact, yielding valuable insights on long-term trajectories of change, and of possible path-dependency in the evolution of regional land-cover patterns.

4. *Land Use Change*

Elements of land-use practices have been a staple topic of investigation for numerous disciplines in the social sciences, including anthropology, archaeology, demography, economics, geography, history, political science, and sociology. Through different lenses, scholars in these fields have sought to explain how and why people in different parts of the world have developed technology and social systems in response to local environmental conditions, how these human actions have in turn shaped the local environment, and how people then respond to these changing conditions (Redman 1999). Thousands of case studies have been undertaken, usually concerning communities or small regions, and some efforts have been made to synthesize these cases in order to extract general principles and to develop a unified theory. While the history of land use in Europe is well documented (e.g., Darby 1956), land-cover transformation has been perceived to be most dramatic in the tropics in recent years, and a great deal of attention has been accorded these regions, with particular attention to agricultural land-use practices (e.g., Ruthenberg 1976).

A broad range of factors has been proffered to explain local land-use practices, with perhaps the greatest effort spent examining agricultural change and the role of population growth (see *Agricultural Change Theory*). The now-famous work of British economist Thomas Malthus, who predicted famine as an inevitable result of population growth in Europe, has fostered huge debate within the social sciences. Another major landmark was the work of Ester Boserup (1965), who turned the Malthusian thesis on its head, attributing agricultural development in Asia to land pressure driven by population growth. Subsequent refinements have described a range of land use change trajectories that may emerge from land pressure and suggested how the driving forces of change are mediated by social and environmental factors at many levels (Turner and Ali 1996).

In addressing the relationship between population growth and land degradation, political-economic analyses hold external forces responsible when land managers engage in environmentally destructive land use practices (see *Land Degradation*). Likewise, cases have been articulated of population growth leading to the rehabilitation of natural resources, including afforestation (see *Land Reclamation*). A great deal of work has been done to examine the role of demographic factors in land use change, with particular emphasis on government-subsidized settlement programs in Latin America, and rural–urban migration in Africa and Asia. Results suggest complex relationships between land-cover dynamics and the provision of infrastructure (especially roads), family development cycles, urban wage employment, and other factors that vary strongly by region (e.g., Lambin et al. in press; see *Property Rights*).

In the case of forest cover, global-scale empirical evidence reveals a complex relationship between forest cover and population density, suggesting that as population density rises in newly settled regions, forest cover has often been removed for the purposes of agricultural expansion. Another period ensues during which wood fuel is replaced by other sources (e.g., fossil fuels) and agricultural intensification allows for the production of sufficient food on less land. Eventually affluence rises to the point where attention to environmental rehabilitation becomes feasible and afforestation becomes the dominant trend.

5. *Land-use/Cover Dynamics*

Perhaps the main challenge of land use/cover change science is to forge robust linkages between the broad views of land cover change provided by remote sensing and other techniques, and the fine-grained understandings of land-use dynamics. In weaving together these diverse strands of inquiry, land use/cover change research examines the observed paths, or trajectories, and projects them into the future. While this may not lead to prediction, it will suggest what might happen should present trends continue and, perhaps more importantly, suggest other possible futures. There are significant obstacles to achieving this integration of land-use and land-cover perspectives, but advances in geographic information technologies and new analytical techniques show great promise.

Most early models of land cover change were either empirical (mimicking the observed relationships between variables, such as forest cover and population density), or mechanistic (describing a limited range of landscape processes according to simple equations) (Kaimowitz and Angelsen 1998). While such models have the advantage of requiring minimal, empirical evidence (e.g., maps of forest cover and population density), their utility is limited by the assumption that such relationships will persist and are the main determinants of landscape change (Lambin 1994). In order to incorporate the information gleaned through research on land-use dynamics more realistically, a new generation of dynamic simulation models are being developed that model the behavior and even the learning processes of individual actors in the landscape. Other developments include the generation of land use/cover change scenarios and the analysis of syndromes of vulnerability to environmental change.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Climate Impacts; Deforestation–Forestation; Desertification; Development: Sustainable Agriculture; Ecology, Cultural; Ecology, Political; Extinction of Species; Global Environmental Change: Human Dimensions; Greening of Technology and Ecotechnology; Human–Environment Relationships; Land Degradation; Land Reclamation; Local–Global Linkages: Environmental Aspects; Planning Issues and Sustainable Development; Remote Sensing and Geographic Information Systems Analysis; Sustainability Transition: Human–Environment Relationship

Bibliography

Boserup E 1965 [1993] *The Conditions of Agricultural Growth: The Economics of Agrarian Change under Population Pressure*. Earthscan Publications, London

Darby H C 1956 The clearing of the woodland in Europe. In: Thomas W (ed.) *Man's Role in Changing the Face of the Earth*. University of Chicago Press, Chicago

GAIM 2000 *Summary report for the GAIM 6000 yr BP experiment and BIOME 6000. Global analysis, integration and modeling*. http://gaim.unh.edu/Structure/Key-Issues/Biome6000/index.html

Kaimowitz D, Angelsen A 1998 *Economic Models of Tropical Deforestation: A Review*. Center for International Agroforestry Research, Bogor, Indonesia

Kasperson J X, Kasperson R, Turner B L II (eds.) 1995 *Regions at Risk: Comparisons of Threatened Environments*. United Nations University Press, Tokyo

Lambin E F 1994 *Modelling Deforestation Processes: A Review*. The European Commission, Luxembourg, TREES Series B: Research Report No. 1

Lambin E F, Baulies X, Bockstael N, Fischer G, Krug T, Leemans R, Moran E F, Rindfuss R R, Sato Y, Skole D, Turner B L II, Vogel C 1999 *Land Use and Cover Change Implementation Strategy*. International Geosphere-biosphere Programme Secretariat, Stockholm, IGBP Report No. 48, IHDP Report No. 10

Lambin E F, Turner B L II, Geist H, Agbola S, Angelsen A, Bruce J W, Coomes O, Dirzo R, Fischer G, Folke C, George P S, Homewood K, Imbernon J, Leemans R, Li X, Moran E F, Mortimore M, Ramakrishnan P S, Richards J F, Skånes H, Steffen W, Stone G D, Svedin U, Veldkamp T, Vogel C, Xu J in press Our emerging understanding of the causes of land-use and -cover change. *Global Environmental Change* **4**

Marsh G P 1864 [1965] *Man and Nature: Or, Physical Geography as Modified by Human Action*. Belknap Press, Cambridge, MA

Moran E 2000 *Human Adaptability: An Introduction to Ecological Anthropology*. Westview Press, Boulder, CO

National Research Council 1998 *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC

Redman C L 1999 *Human Impact on Ancient Environments*. University of Arizona Press, Tucson, AZ

Ruthenberg H 1976 *Farming Systems in the Tropics*. 2nd edn. Oxford University Press, Oxford, UK

Thomas W L Jr (ed.) 1956 *Man's Role in Changing the Face of the Earth*. University of Chicago Press, Chicago

Turner B L II, Clark W C, Kates R W, Richards J F, Matthews J T, Meyer W B (eds.) 1990 *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere Over the Past 300 Years*. Cambridge University Press, Cambridge, UK

Turner B L II, Ali A M S 1996 Induced intensification: agricultural change in Bangladesh with implications for Malthus and Boserup. *Proceedings of the National Academy of Sciences USA* **93**: 14984–91

W. J. McConnell

Landsat Imagery in Geography

The Landsat series of satellites have produced one of the most valuable remote sensing records of change in the earth's land surface. The particular value of their data lies in the length of their record, which goes back almost continuously to 1972. Their value also arises from their relatively detailed scale or resolution compared with many space-borne sensors, since this means that the impacts of human activities such as deforestation or urbanization can be readily detected and monitored. Much useful information can be obtained from Landsat images through visual inspection of photographic hard copy. But the derivation of reliable quantitative information usually relies on careful processing of digital data from the satellites' sensors. Landsat data have been used for numerous different applications, many relating to land cover. This term refers to the biophysical state of the earth's surface cover and provides a valuable input for the mapping and monitoring of land use. Other uses of the data include water and air pollution monitoring, and mineral discovery (Jensen 2000, Lillesand and

Kiefer 1999). Landsat data and the products derived from them are often best used when placed within the framework of a Geographic Information System (GIS).

1. The Landsat Series

The first Landsat satellite was placed in orbit on September 23, 1972 (Table 1). Several subsequent satellites have been launched successfully, with the exception of Landsat 6, which failed to reach orbit. All the other satellites have exceeded their nominal lifetimes. The first three Landsat satellites (given the original name of ERTS or Earth Technology Resources Satellite) had a Multispectral Scanner System (MSS) as the prime sensor, yielding data with spatial resolution of 79 m. Subsequently Landsats 4 and 5 not only had the MSS sensor but also included the sensor called the Thematic Mapper (TM), which has a resolution of 30 m. On Landsat 7 the Enhanced Thematic Mapper (ETM) was the sole sensor on board; its properties are similar to those of the TM, but it has some capability for providing more detailed images (see Table 1). All instruments collect data by focusing radiation from the earth on detectors; the resultant recorded signals represent the amount of incoming radiation in digital form. Radiation from several different wavebands is measured simultaneously. These signals can then be used to create hard-copy pictures such as those illustrated here, or can be made available as digital files that can be subsequently processed by the user. The spatial resolution of the sensors indicates the smallest size of object detectable, but the latter is also dependent on the contrast between the object and its surroundings. Also, digital processing of the data can generate products with different pixel sizes. Hence, although the basic resolution of the MSS is 79 m and the TM is 30 m, many of the standard products purchased by users will have pixel sizes of 57 m and 28.5 m respectively. The frequency of images acquired from Landsat varies considerably. Landsats pass over every part of the earth's surface approximately once every 16–18 days but in practice images are collected far less frequently, due to factors such as cloud cover and lack of recording capacity of satellites. In the early 1970s a systematic attempt was made to acquire global coverage. In contrast, during much of the 1980s and 1990s, when the system was operated by commercial entities, data were often only acquired when a firm customer had been identified for a particular scene. Only with the launch of Landsat 7 was an acquisition strategy implemented to ensure total global acquisition of cloud-free images on a seasonal basis.

All Landsats have a sun-synchronous orbit, which means they can take images of every location at the same time each day. This greatly assists in change detection. One of the important progressive improvements in the Landsat sensors during their history has been their better calibration, resulting in an increasingly accurate record of the earth's surface.

2. Distinctive Role of Landsat Data

The high resolution and long record of Landsat data make them especially useful for the detection and monitoring of human impacts on the earth's surface. There are other satellite systems with similar sensing capabilities to those of Landsat. These include the French SPOT (Satellite Probatoire pour l'Observation de la Terre) series of satellites with a spatial resolution of 20 m (and 10 m for a panchromatic band) and the IRS (Indian Remote Sensing Satellite) series launched by the Indian Space Agency. Images from these are capable of being used in the same way as data from Landsat. However, neither has the longevity or comprehensiveness of the spatial coverage provided by Landsat's sensors.

Table 1
Landsat satellites, their sensors and characteristics

Landsat Satellite Number	Launch date	Finish of data collection	Sensors on board satellite (see below for properties)
1	September 23, 1972	January 6, 1978	MSS
2	January 22, 1975	February 5, 1982	MSS
3	March 5, 1978	March 31, 1983	MSS
4	July 16, 1982	July 1987	MSS and TM
5	March 1, 1984	Operational	MSS and TM
7[a]	April 15, 1999	Operational	ETM

a Landsat 6 never achieved orbit.
Characteristics of sensors
MSS bands: visible (0.4–0.5 μm, 0.6–0.7 μm); visible/near-IR boundary (0.7–0.8 μm); near IR (0.8–1.1 μm) and thermal infrared (10.41–12.6 μm)
Spatial resolution: 79 m, except 237 m for thermal band
TM bands: visible (0.45–0.53 μm, 0.52–0.60 μm, 0.63–0.69 μm); near-IR (0.76–0.9 μm); mid-IR (1.55–1.75 μm, 2.08-2.35 μm); thermal (10.40–12.5 μm)
Spatial resolution: 30 m, except 120 m for thermal band
ETM bands: visible (0.45–0.515 μm, 0.525–0.605 μm, 0.63–0.69); near IR (0.75–0.90 μm), mid IR (1.55–1.75 μm, 2.09–2.35 μm); thermal IR (10.4–12.5 μm); panchromatic (0.52–0.9 μm). Spatial resolution: 30 m, except 60 m for thermal band and 15 m for panchromatic band

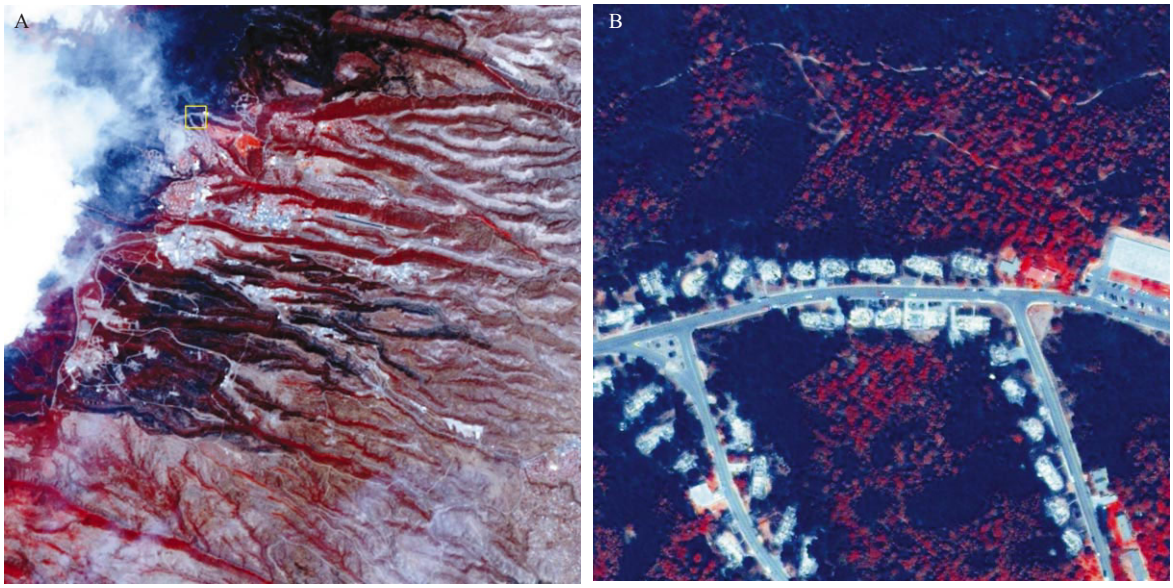

Figure 1
Comparison of images from the Landsat Enhanced Thematic Mapper and high-resolution data from IKONOS of Los Alamos, New Mexico, shortly after the Cerro Grande fire, May 2000 (includes material © Space Imaging L.P.). (A) A Landsat ETM scene. The dark areas show areas burnt during the fires of 2000. The area is approximately 15 km across. (B) An image of a small part of the Thematic Mapper image (see small box in Fig. 1a). This image was taken by the IKONOS satellite and is shown here with a resolution of 1 m

Complementing Landsat data are those at coarser resolutions, notably from the NOAA series of satellites. Advanced Very High Resolution Radiometer is of particular value in monitoring seasonal variations in crops and other types of vegetation (D'Souza et al. 1996), but only at the coarse resolution of around 1 km. Much finer resolution data from commercially launched satellites (or from intelligence data released to the general public) provide much more detailed images than Landsat in which human activities can more readily be interpreted (Fig. 1); their availability is much more restricted than Landsat data.

3. *What do Images from Landsat Show?*

Successful use of Landsat data requires an understanding of the basics of how radiation interacts with the earth's surface and its atmosphere (Lillesand and Kiefer 1999). The higher the amount of incoming radiation in a given waveband (Table 1) the higher the digital number recorded by the sensor. All the bands of the sensors except for the thermal infrared band sense the amount of reflected radiation, including the near and mid infrared. TM bands 1 through 3 sense radiation in the part of the spectrum sensed by the human eye, TM band 4 senses the near infrared which is sensitive to the reflectance by the physical structure of healthy vegetation, and TM bands 5 and 7 are sensitive to the amount of water in leaves as well as the physical and chemical composition of soil materials. For the thermal infrared bands, higher recorded values relate to the temperature of the surface.

Pictures derived from Landsat images are usually false-color images in that they represent the sensed radiation by use of colors readily recognizable by the human eye, but do not show the actual appearance of the surface to the unaided human eye (Parkinson 1997). Figure 2 shows TM data. The atmosphere can substantially distort the signal received by the Landsat sensors.

4. *Extracting Information from Landsat Imagery*

Extracting information from Landsat data basically involves finding out what amount of recorded radiation corresponds to samples of known ground conditions radiation and using these relationships to extrapolate throughout the whole image. In practice, if reliable results are to be obtained a quite complex set of technical procedures have to be performed (Lillesand and Kiefer 1999). First, even obtaining images of the right place and date may require interrogation of several different archives, because as yet there is no single comprehensive archive; the

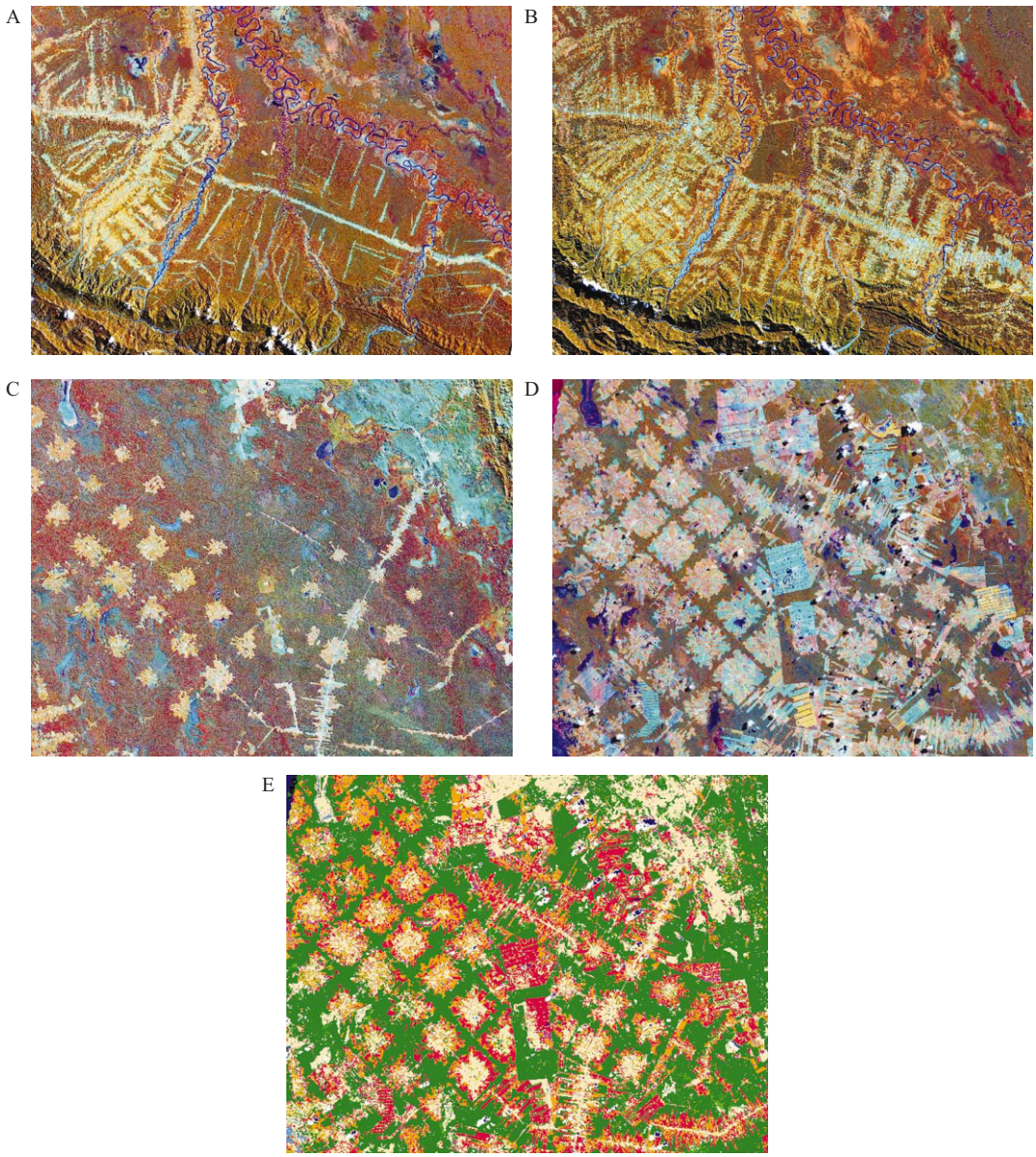

Figure 2
Examples of deforestation depicted on Landsat images. (A) Deforestation in Cochabamba State Bolivia, image was obtained by the Thematic Mapper on October 4, 1986; the impact of the road pattern on deforestation patterns can be clearly seen. Similar patterns of land use change are found in many other tropical areas. The area is approximately 40 km across. (B) The same area as for Fig. 2B, obtained on July 12, 1999. (C) Deforestation near Santa Cruz in central Bolivia, resulting from government resettlement of people from the highlands in settlements known as the San Julian colonies. The image was obtained on July 2, 1986 and is approximately 53 by 42 km in extent. (D) The same area as for Fig. 2C on March 26, 2000, showing the agricultural expansion between the two dates. (E) Classification of the two Landsat images depicted in Figs. 2C and 2D.

United States Geological Survey has the most comprehensive set of images. Ground reference data will often have to be collected, though higher resolution remotely sensed data can sometimes act as a useful surrogate for ground data (Fig. 1). Pre-processing of the Landsat data is commonly required. Correction of

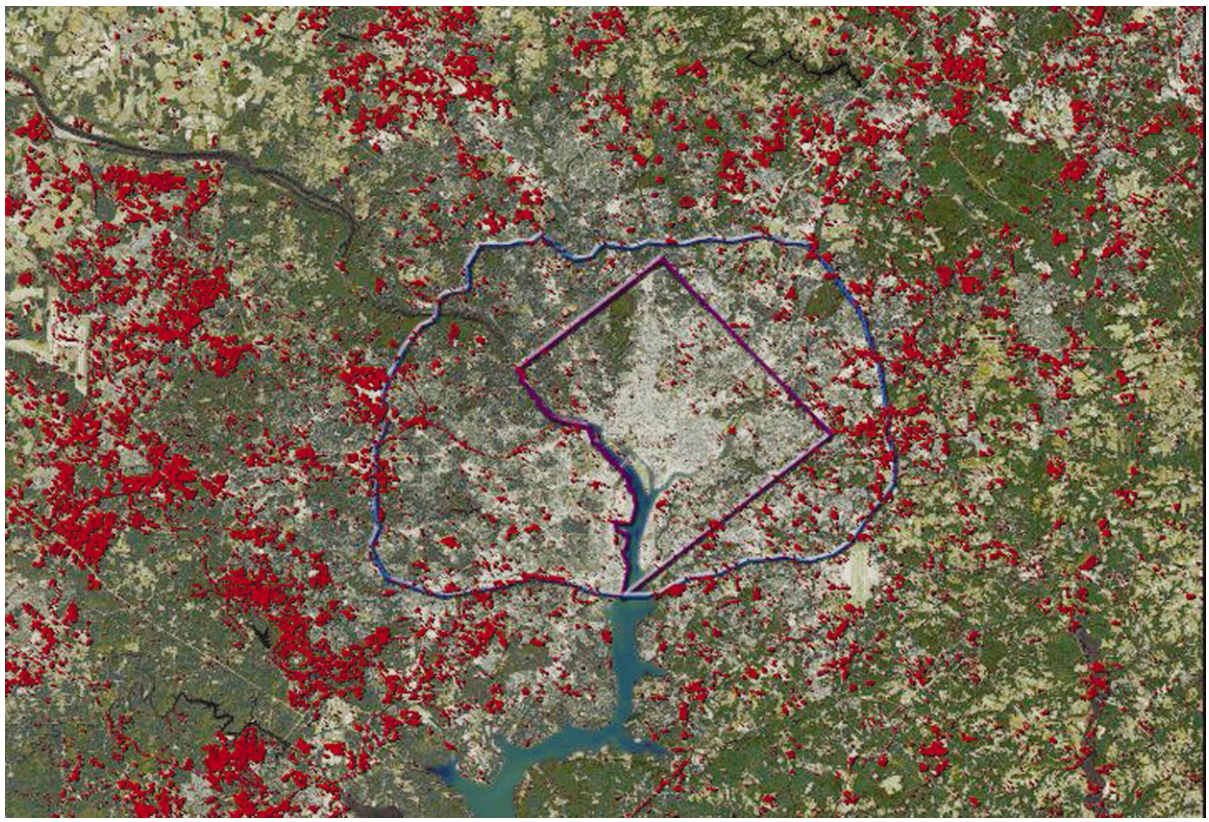

Figure 3
Urban growth from Landsat imagery for Washington, DC and the surrounding areas of Maryland and Virginia. (See Masek et al. 2000 for further details)

the data may be required to avoid problems such as variability in sensor performance over time and reduction of atmospheric distortions, when change detection is being carried out. Reprojection of the data will often be necessary to ensure the data are co-registered with conventional map bases and are co-registered with each other if more than one set of Landsat images are used. Because of the importance of a common spatial framework for Landsat data, and other spatial data sources as well as field data, GIS is commonly used.

A choice will need to be made between different information extraction procedures. Reliance solely on traditional photo-interpretation methods is relatively uncommon nowadays. Even where photo-interpretation methods are adopted, users will employ digital enhancement methods to aid the interpretation. Computer-based methods, normally using some type of statistical decision rule, are most commonly applied, since they allow the most internally consistent and detailed results. Once the information has been extracted using such methods, editing by skilled interpreters can enhance the quality of the result. Finally an assessment of the accuracy of the result should be carried out using field or other reference data not used in earlier stages, so that the reliability of the end products is understood.

5. *Applications of Landsat Imagery*

Landsat images have been used for a wide range of applications, including many concerned with the transformation of the biophysical environment by humans and the impact of the physical environment on human society. One of the more important applications has been in the monitoring of forests, especially in tropical regions. Deforestation in the tropics and sub-tropics is a longstanding phenomenon, but one which has accelerated greatly in recent years. Its significance in the impact on the global carbon cycle and hence on global warming, on massive reductions of biodiversity (because of the species richness of these areas), and on its impact on the sustainability of these regions.

Until relatively recently there was considerable controversy about the actual rates of tropical deforestation. The analysis of Landsat data has done much to reduce uncertainties over both the rate and location of deforestation. Some of the higher estimates of deforestation for the Brazilian Amazon were shown to be incorrect (Skole and Tucker 1993). The need for spatially comprehensive Landsat (or similar resolution) data to obtain accurate estimates has been confirmed (Tucker and Townshend 2000). Different types of land management can be readily discerned from the resultant spatial patterns depicted on Landsat data (Fig. 2). The overall pattern of deforestation is clearly linked to the location of transportation corridors and settlement. Deforestation within countries is often a direct result of government-sponsored resettlement, as in Brazil and Indonesia, as well as international investment in agriculture, as in southern Bolivia, where World Bank loans have encouraged extensive industrial soybean production (Steininger et al. 2000).

Urbanization is another process of land transformation that can be monitored using Landsat data (Masek et al. 2000). Figure 3 shows the expansion of built-up areas near Washington DC between the 1980s and 1990s. Although reliable aggregated data are available in many developing countries, the benefit of products derived from Landsat lie in the flexibility of aggregation and disaggregation of spatial units. In developing countries, in particular, products derived from Landsat data will often be much more up to date than locally derived official data.

The use of Landsat and other forms of remote sensing in the social sciences can be found in the collection of papers edited by Liverman (1998).

6. *Conclusions*

Data from the Landsat series of satellites provide an invaluable source of information on the earth's environment, especially properties of the land surface, and how it has been affected by human activities. Not only do the resultant images provide timely information on current conditions, but as time passes they provide an increasingly useful historical record.

Bibliography

D'Souza G, Belward A S, Malingreau J-P 1996 *Advances in the Use of NOAA AVHRR Data for Land Applications*. Kluwer Academic Press, Dordrecht, The Netherlands

Jensen J R 2000 *Remote Sensing of the Environment: An Earth Resource Perspective*. Prentice-Hall, New York

Lillesand T M, Kiefer R W 1999 *Remote Sensing and Image Interpretation*, 4th edn. Wiley, New York

Liverman D (ed.) 1998 *People and Pixels: Linking Remote Sensing and Social Science*. National Research Council, Washington, DC

Masek G, Lindsay F, Goward S N 2000 Dynamics of urban growth in the Washington DC metropolitan area, 1973–1996, from Landsat observations. *International Journal of Remote Sensing* **21**: 3473–86

Parkinson C L 1997 *Earth from Above: Using Color-Coded Satellite Images to Examine the Global Environment*. University Science Books, London

Skole D L, Tucker C J 1993 Tropical deforestation and habitat fragmentation in the Amazon: Satellite data from 1978 to 1988. *Science* **260**: 1905–9

Steininger M K, Tucker C J, Townshend J R, Killeen T R, Bell V, Ernst P 2000 Tropical deforestation in the Bolivian Amazon *Environmental Conservation*(in press)

Tucker C J, Townshend J R G 2000 Strategies for monitoring tropical deforestation using satellite data. *International Journal of Remote Sensing* **21**: 1461–71

J. Townshend

Landscape Architecture

A landscape is the synthesis of all the natural and cultural features—fields, buildings, hills, forests, farms, deserts, and water—that distinguish one part of the surface of the earth from another part. Landscape architecture is the art and science of arranging land so as to adapt it most conveniently, economically, functionally, and aesthetically to any of the varied wants of people.

Frederick Law Olmsted, Sr (1822–1903) was largely responsible for creating the discipline. He adapted European landscape gardening ideas to American places. The driving force for Olmsted was how to make increasingly industrial and crowded cities more livable. Olmsted's work as well as his goals for the discipline defined the field. Beginning in the mid-nineteenth century, he and his colleagues designed urban parks; new suburban communities; parkways; campuses; state, national, and international parks; city and regional plans; cemeteries; and estates. In the process, a truly American art form was born, which has had international influence.

Landscape architecture was developed in four periods. First, varied European styles emerged in landscape gardening, including the Moorish in Spain, the Italian Renaissance, the French Baroque, and the highly original English picturesque, as well as distinct, yet synthetic, styles in the Netherlands, Germany, and Poland. Second, the contributions of Olmsted and his collaborators established a broader social agenda. Third, a modern movement emerged just before the World War II, which reached fruition in the decades that followed the war. Finally, landscape architecture has played an influential role in the environmental movement from the 1960s to the present. In each of these four periods an Asian influence has been present.

For example, especially the English landscape style exhibited an awareness of Chinese and Japanese gardens.

1. European Precedents

In Islam, the Mediterranean city, building, and garden design tradition of the Greeks and Romans was maintained as the Dark Ages descended on Europe. The Islamic tradition emphasized the organization of space based on an application of principles from geometry and other branches of mathematics. Interior courtyards were favored for housing, with the judicious use of water. Water played an important role, not only for drinking and irrigation, but also for cooling. The arrangement of space, the manipulation of water flows, and the careful application of building materials, such as masonry and ceramic tiles, contribute to a pleasant living environment.

The Islamic garden is the 'Paradise Garden' of the *Koran*. These places were to be a manifestation on earth of a heavenly condition. Its rigorous ordering according to geometrical principles is a logical corollary and makes it equivalent to a mandala. The Islamic approach had a significant impact in the Iberian peninsula. As Spain was conquered by the Christians, Islamic ideas about design and planning filtered into European thinking. Among the structures built by the Moors in Spain, their dramatic, fortress palace at Granada, called the Alhambra and built between the thirteenth and fourteenth centuries, would make an especially important impression. In particular, its inner courtyards (the Court of Myrles and the Court of the Lions) provide models for how water and shade can be manipulated for human comfort and delight.

The Islamic world was one of the influences on the Italian renaissance. The early phase of this learning revival was especially strong in Tuscany, where the influence of Moorish design on the hillside villas around Florence is evident. For example, the powerful Medici family built hillside gardens at Fiesole (1458) overlooking Florence. They were also responsible for the extensive Boboli gardens behind the Pitti Palace in Florence. The outdoor spaces of these villas were remarkable both in Tuscany and later around Rome as the Medici introduced the style there and eventually elsewhere in Italy. While there were adaptations of Islamic ideas and uses of materials, the Italian designers also drew on ideas from ancient Rome, especially those of Marcus Vitruvius Pollio, and forged a new style. In works such as the gardens of Villa Lante (1518) and Villa D'Este (begun c. 1550), architects such as Donato Bramante (1444–1514), Pirro Ligorio (1491–1580), and Jacopo Barozzi da Vignola (1507–73), sculpted earth, water, and plants in new ways. Geometric patterns were carved into hillsides, with water flows directed to erupt strategically at dramatic fountains, framed by rows of trees and shrubs that guided the eye to various focal points. Water is an essential feature of Italian gardens, of almost all designed landscapes really, just as are hard stone and living plants, because the polarities of the senses can be exposed.

The great French garden designer André Le Nôtre (1613–1700) took these Italian forms and enlarged them with Baroque proportions. The natural expanses of the French countryside were not as constrained as the hillsides of Italy and the power of the Sun King eclipsed that of even the Medici. Begun in 1661, Louis XIV's palace at Versailles recreated a vast, five-square-mile portion of the French countryside into an elaborate system of grand vistas. A main east–west axis and several crossing axes of various sizes divide the space into a series of parterres with connecting *allées* bordered with massive, carefully manicured trees. As with the Italian gardens, Le Nôtre created fountains and other focal points at strategic points along the sight lines, but whereas trickles of waters collect into gushing fountains in Italy, a grand canal an elaborate, mechanical pumping system was created at Versailles. From the apex at the château, the landscape underscored who controlled and ruled this place.

A contrasting landscape style emerged in England, as it became a sea power connected to the Americas and the Orient and influenced by growing ideas about democracy. While John Locke (1632–1704) wrote about the inalienable rights of people to pursue life, liberties, and property, poets like Alexander Pope (1688–1744) waxed on about learning from the genius of the place. As the Englishman Horace Walpole (1717–97) observed of William Kent (c. 1685–1748), one of the innovators of his nation's landscape design tradition, 'He leaped the Fence and saw that all Nature was a garden.'

Kent, Lancelot 'Capability' Brown (1716–83), and Humphrey Repton (1752–1818) sought to improve the English landscape. They arranged estates with asymmetrically placed clumps of trees, broad undulating greenswards, serpentine lakes, and surrounding belts of woodlands in the grounds of Stowe (1736), Rousham (1738–41), Kensington Gardens (c. 1744), Bowood (1763–71), Blenheim Palace (c. 1765), and Ashburnham Palace (1767). Their romantic designs used small classical and Asian influenced structures as focal points. The English landscape garden became known as an imitation of nature. Where straight lines dominated the Moorish, Italian, and French designs, curves preponderated in the British Isles from the eighteenth century on.

Watkin (1982) argues persuasively that the English landscape tradition drew its inspiration for its non-geometric 'picturesque' qualities from the Chinese garden. The English and other Europeans were exposed to the Chinese approach especially by returning French Jesuit priests. The Chinese garden contains complex symbolic meanings for its cultural and phys-

ical context. These gardens, as well as subsequent adaptations and transformations in Korea and Japan, are richly symbolic as microcosms.

Ancient Roman gardens, as well as rising English nationalism, also influenced the picturesque approach. Watkin notes that the third Earl of Shaftesbury, 'called for the creation of a national style and a national taste based on the spirit of national freedom, which he believed was enshrined in the Whig oligarchy …' and to achieve this style the model was 'the free commonwealth of republican Rome as well as the culture of China' (1982, p. 2).

This growing philosophy, affected largely by the teachings of Locke, coincided with the English colonization of the New World and would have a lasting imprint in North America. Meanwhile, the English landscape style and the contrasting geometric forms from the continent impacted approaches in other European nations, which also developed their own styles, notably in the Netherlands, Germany, and Poland. In the Netherlands especially, the term landscape (*landschap*, in Dutch) has deep cultural significance, because much of the land in the nation has literally been made. The Germans borrowed from the English approach, but contributed a more scientific understanding of nature, laying the groundwork for contemporary concepts of ecology and sustainability. The Poles combined formal and informal spaces in conscious melding of nature and culture, at places such as the grounds at Wilanów Palace and Lazienki Park.

2. *Olmsted and his Collaborators*

American landscape architecture came into being after several decades of marination in the English style. Andrew Jackson Downing (1815–52) was certainly an adherent to the English school of landscape gardening. Frederick Law Olmsted, Sr. had been influenced profoundly by the English style, especially Birkenhead Park, which he visited during an 1850 trip to the British Isles. Birkenhead is a suburb of Liverpool. Its park was designed specifically for the public by Joseph Paxton (1803–65) in 1844. Previous parks in England and elsewhere in Europe were the estate gardens of the aristocracy donated to a city or loaned to the public.

A second direct impression from the British Isles on Olmsted's work was his partnership with the English architect, Calvert Vaux (1824–95). In spite of the strong English influences, Olmsted was an original. He altered the name of the field from 'landscape gardening' to 'landscape architecture' and established a new discipline that transcended making gardens for the very rich and powerful.

Olmsted's incredible start in landscape architecture was the design and construction supervision of Central Park in New York City with Vaux beginning in 1857. Through Central Park, Olmsted sought to bring the benefits of open space to the urban masses. The urban park would provide a healthy antidote to the crowding and pollution brought on by the Industrial Revolution. Central Park was an immediate success. This led to Olmsted to establish a landscape architecture practice, initially with Vaux and later with his sons. The practice was responsible for urban parks in Brooklyn, Boston, Chicago, Montreal, and, eventually throughout North America.

The discipline that Olmsted founded was defined by the scope of his own work as well as that of his son (Frederick Law, Jr., 1870–1957) and nephew–adopted son (John C. Law, 1852–1920). In addition to urban parks, Olmsted designed Riverside, Illinois—a prototype suburban new community; contributed to the preservation of Yosemite and Niagara Falls—setting the state for national and international parks; and participated in the planning of the World's Columbian Exposition of 1893 which created the inspiration for the 'City Beautiful Movement.' Toward the end of his prolific life, Olmsted, Sr. also designed an estate for the very wealthy George W. Vanderbilt, but Biltmore was also an experimental forest and farm. At Vanderbilt's Biltmore, the French-trained American forester and conservationist Gifford Pinchot (1865–1946) first applied his 'multiple use, sustained yield' concept, a precursor to contemporary sustainability.

While the Olmsteds broadened the scope of landscape architecture, the style they employed was strongly English-influenced. The 'pastoral aesthetic,' as well as the associated Whig political philosophy, resonated with Americans. The aesthetic was appropriate for large portions (but not all) of the North American continent. When Olmsted worked in California on the design of the Stanford University campus plan, the style, more Spanish than English, was more suitable for the climate and history of the state.

Stanford was but one of the campuses that Olmsted, his sons, and their collaborators designed. In addition to designing campuses, Olmsted was involved in other aspects of higher education. He was a strong supporter of the land-grant colleges, established by Justin Morrill (1810–98) in 1862, to offer education in the agricultural and mechanical arts and sciences. Morrill, Olmsted, and others believed that the interests of agriculture and industry had been ignored by the established universities in the United States which followed more elitist goals. Many of these land-grant schools began requiring landscape gardening courses in the late nineteenth century and one (Iowa State) offered Olmsted its presidency.

Landscape architecture impacted the elite schools too. Perhaps, Olmsted's brightest young collaborator was Charles Eliot, son of the president of Harvard. The younger Eliot organized probably the first comprehensive, multidisciplinary ecological inventories of Mount Desert Island in Maine during the 1880s. With the Olmsted firm, Eliot developed a plan for the Boston metropolitan region in 1893. He proposed an 'emerald necklace' for the region that linked park and

open space systems while making environmental improvements for flood control and water quality. Eliot died of spinal meningitis prematurely at age 38 in 1897. In his memory, his father established the landscape architecture program at Harvard that would lead the discipline through much of the twentieth century.

3. *The Modern Movement in Landscape Architecture*

In the early twentieth century, landscape architecture established itself as a profession in North America and northern Europe and Japan. Professional associations were founded, including the American Society of Landscape Architects (1899), and leading universities, such as Harvard and the growing land-grant schools, initiated departments or programs. However, much of the idealism of the Olmsteds dwindled, as the progressive branch of the field embraced the new profession of city planning. The leading landscape architecture historian Norman Newton noted that 'single-track eclecticism' took over especially in the Harvard department as practitioners flocked to build estates and exclusive neighborhoods for the wealthy.

However, it was at Harvard where a counter movement against this eclectic garden-design dominance began. In the late-1930s, students Garrett Eckbo (1910–2000), Dan Kiley (b. 1912), and James Rose (1910–91) rebelled against the Harvard curriculum. Their quest for a more modern, a more creative, approach to landscape design coincided with the arrival of Walter Gropius (1883–1969) at Harvard in 1937. Gropius helped institutionalize the Bauhaus philosophy of design within this leading American academic institution.

Meanwhile, in California, Thomas Church (1902–1978) had developed a practice revolving around gardens for the middle and upper-middle class. Celebrated in *Sunset* magazine, the Church garden provided an ideal for the new American outdoor, suburban lifestyle. Between the rebellious Harvard trio and Church, a modern movement in landscape architecture developed. This modernism is characterized by both its democratic spirit, that is, making the benefits of landscape design available to a wide spectrum of people, as well as by its spare cubist- and surrealist-inspired forms. For example, Eckbo designed Depression-era multifamily housing in California for the Farm Security Administration. These schemes are functional, spatial arrangements, designed to provide shelter and a delightful environment for working people in challenging economic situations. Eckbo was one of many landscape architects who contributed to the economic and social recovery of the nation through various public works projects from the greenbelt new towns to the national parks and forests.

American modernists found inspiration in the works of the Brazilian Roberto Burle Marx (1909–94) and the Mexican Luis Barragán (1902–88). However, arguably modernism in landscape architecture coalesced through the work of Hideo Sasaki (1919–2000) and Lawrence Halpern (b. 1916). As both student and teacher gravitating between the University of Illinois and Harvard, Sasaki defined landscape architecture during the 1950s. In his San Francisco home, Halpern's work reflected a free spirit, in harmony with the 1960s. Their work returned in scope and scale to that established in the previous century by the Olmsteds—urban parks, new communities, campuses, and, increasingly, corporate headquarters.

4. *The Ecological Era*

During the 1960s, the way people viewed their relationship to the Earth changed fundamentally. Since the World War II, the specter of nuclear destruction clouded the future. In the quest to prevail in the arms race, the Soviet Union and the United States raced to the moon. Something unexpected happened. Images of the Earth from space were transmitted into living room televisions and breakfast newspapers. On July 20, 1969, Neil Armstrong stepped on the moon and altered our view of Earth. From that barren orb, it became clear that the Earth is alive. Boundaries are human impositions. The world we inhabit is both fragile and beautiful.

Even before the *Apollo* images were broadcast, voices for protecting the Earth had been raised. Aldo Leopold (1886–1948) urged people to view land as community, and Rachel Carson (1907–1964) demonstrated the dangers of poisoning our nest. To this chorus, Ian McHarg (1920–2001) urged that we 'design with nature.' His book by the same name provided a new theory for landscape architecture, and altered other fields as well, notably planning, architecture, and the environmental sciences.

From his University of Pennsylvania (Penn) base, McHarg married a traditional landscape architecture technique with contemporary concepts from ecology for his new theory. Map overlays had been used by landscape architects since the Olmsteds and Eliot. McHarg suggested that information be collected in a systematic, chronological order to reveal interactions, relationships, and patterns. Natural phenomena, such as climate and geology, influence processes like water flow and soil development, which in turn affect the location of plants and animals. McHarg suggested that by conducting such an ecological inventory of a place opportunities and constraints for various possible land uses can be identified. The values of people living in the place can then be used to determine how these opportunities and constraints indicate a range of suitabilities for potential land use.

This approach influenced how environmental impact assessments are undertaken and provides the

underlying theory for geographic information systems (GISs). GISs are used for displaying, analyzing, and storing spatially-related data in computers. The initial technological and theoretical advancement of GISs largely came from landscape architecture (and, sometimes, planning faculty) such as Carl Steinitz (Harvard), Bruce MacDougall (Penn then the University of Massachusetts), Julius Fabos (University of Massachusetts), Lewis Hopkins (University of Illinois), and Thomas Dickert (University of California–Berkeley).

McHarg recognized that in order for human communities to take full advantage of ecological knowledge, a fundamental value shift was necessary especially in western nations. He argued that the Judaic–Christian philosophy of 'multiplying and subduing the earth' has deleterious consequences for the environment. McHarg found hope in Native American and Asian religious beliefs, worldviews that stressed harmony with nature, rather than dominance.

5. *Prospects: Following Nature's Lead*

The ecological revolution promulgated by McHarg at Penn went beyond influencing techniques and technology. The ecological prospective advocated a fundamental shift in the way we design and plan. Ecology is the study of how all living creatures, including people, interact with each other as well as with their physical and biological environments. Such understanding is essential for sustainable development and regenerative design. Ecological design involves understanding the regional context in which humans live, a recognition of natural limits and capacities, and an ability to build at an appropriate scale.

Landscape architecture practices, such as Andropogon in Philadelphia, Jones and Jones in Seattle, and Design Workshop in Colorado have continued to refine and advance McHarg's concepts about ecological design and planning. Iconoclastic practitioners such as Rich Haag in Seattle, A. E. Bye in Connecticut, and Laurie Olin in Philadelphia have adopted a less empirical, more intuitive, yet still ecological, approach to their work. Other practices, such as Hargreaves Associates in San Francisco and Civitas in Denver, initially rejected an ecological based approach in favor of a more fine arts orientation, but eventually gravitated toward addressing ecology, because of the complexity of the urban and suburban sites where they work demanded it.

Meanwhile, theorists such as Anne Whiston Spirn, Elizabeth Meyer, and James Corner suggest new insights into how to read landscapes, with ecology providing the vocabulary for such readings. Landscape architects are unique among artists and architects in that their mediums for creation are mostly living entities. They work with natural processes and materials, which is a complex undertaking. As the Roman architect Vitruvius observed in his second book on architecture '… all things are generated as the Nature of Things has determined, not for the pleasure of man, but disparate as though by chance.'

The necessity to understand the nature of places is basic to the art of landscape architecture. Ecology provides the language for reading landscapes. Such understanding leads to an appreciation of critical regionalism. Understanding and acknowledging the diverse cultures of places is fundamental to such regionalism. American landscape architecture appears to be adaptable to many places internationally. Such internationalism is neither an international 'style' nor a form of cultural imperialism, rather it is a flexible approach, sensitive to cultural pluralism.

Landscape design as a distinct art can first be recognized in the Moorish gardens of Spain and in the hillside Renaissance villas of Italy as well as the contemplative places of China, Korea, and Japan. All the early examples represent cultural adaptations to climate and topography. The Italian style was expanded to grotesque proportions in the French Baroque, and was then muted to the practical scale of the Dutch polders. The English landscape school leaped the garden wall, and the Americans introduced landscape aesthetics to the city, the suburb, and the wilderness. Landscape architecture is now crossing continental divides and finding Nature in the whole world.

See also: Art and Culture, Economics of; Art, Sociology of; Development and Urbanization; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Strategic Planning; Urban Growth Models; Urban Life and Health; Urban Places, Planning for the Use of: Design Guide; Urban Studies: Overview

Bibliography

Birnbaum C A (ed.) 1999 *Preserving Modern Landscape Architecture* (papers from the Wave Hill–National Park Service Conference). Spacemaker Press, Cambridge, MA

McHarg I L, Steiner F (eds.) 1998 *To Heal the Earth*. Island Press, Washington, DC

McHarg I L 1969 *Design with Nature*. Natural History Press, Garden City, New York (2nd edn.1992 Wiley, New York)

Newton N T 1971 *Design on the Land: The Development of Landscape Architecture*. The Belknap Press of Harvard University, Cambridge, MA

Olin L 2000 *Across the Open Field: Essays Drawn from English Landscape*. University of Pennsylvania Press, Philadelphia, PA

Rybczynski W 1999 *A Clearing in the Distance: Frederick Law Olmsted and America in the Nineteenth Century*. Scribner, New York

Steenbergen C, Reh W 1996 *Architecture and Landscape: The Design Experiment of Great European Gardens and Landscapes*. THOTH, Bussum, The Netherlands

Walker P, Simo M 1994 *Invisible Gardens: The Search for Modernism in the American Landscape*. MIT Press, Cambridge, MA
Watkin D 1982 *The English Vision: The Picturesque in Architecture, Landscape and Garden Design*. Harper and Row, New York

F. Steiner

Language Acquisition

The ability to learn language is present in almost every human child (Lenneberg 1967). Children who are born blind have few problems learning to speak, although they may occasionally be confused about words for colors or geographic locations. Children who are born deaf readily acquire a rich system of signs, as long as they are exposed to native sign language speakers. Even a child like Helen Keller, who has lost both hearing and sight, can still acquire language through symbols expressed in touch and motion. Children with neurological disorders, such as brain lesions or hydrocephalus, often acquire complete control over spoken language, despite a few months of early delay. Children with the most extreme forms of mental retardation are still able to acquire the basic units of human communication. Given the pervasiveness and inevitability of first language acquisition, we often tend to take the process of language learning for granted. But language is the most complex skill that a human being can master. The fact that nearly all of us succeed in this task indicates how remarkably well the structure of language adapts to our underlying abilities. Language is immensely complex, but its complexity is accessible to all of us.

Linguists tend to think of language as having a universal core from which individual languages select out a particular configuration of features, parameters, and settings (Chomsky 1982). From this perspective, child language is an interesting slice of the universal pie. The shape of this slice is presumably limited both by formal universal constraints and by the child's mental abilities or developmental status.

Psychologists (Fletcher and MacWhinney 1995) view language learning from a very different perspective. To the psychologist, language acquisition is a window on the operation of the human mind. This window allows us to view the structure and functioning of neural circuits in the brain. It also allows us to understand how these circuits support processes such as reinforcement, generalization, imagination, and thinking. To understand these processes better, psychologists conduct controlled experiments in which children learn new words, sounds, and rules. They may measure these processes using neural imaging techniques, or they may simply study the changes in the language of the child across time.

Important and engaging though these scholarly perspectives may be, the best way to appreciate the dynamics of language development is to assume the perspective of the child. By taking this perspective, we can understand the challenges the child faces and the ways in which each are overcome.

1. *Early Audition*

William James (1890) described the world of the newborn as a 'blooming, buzzing confusion.' However, on the auditory level, the newborn's world is remarkably well structured. The cochlea and auditory nerve provide extensive preprocessing of signals for pitch and intensity. In the 1970s, researchers discovered that human infants were specifically adapted at birth to perceive contrasts such as that between /p/ and /b/, as in *pit* and *bit*. However, subsequent research showed that even chinchillas are capable of making this distinction (Werker 1995). Thus, it appears that much of the basic structure of the auditory world can be attributed to fundamental processes in the mammalian ear.

Beyond this basic level of auditory processing, it appears that infants have a remarkable capacity to record and store sequences of auditory events. For example, if the six-month-old hears a sound such as /badigudibagadigudigagidu/ repeated many times, the parts that are repeated will stand out and affect later listening. In this example, the repeated string is /digudi/. If infants are trained on such strings, they will grow tired of this sound and will come to prefer to listen to new sound strings rather than those that have the boring old /digudi/ string (Saffran et al. 1996). These habituation effects are strongest for stressed syllables and syllables immediately following stressed syllables. Recent studies of these effects in auditory memory suggest that we are born with an ability to store and recall the sounds of human language. It is as if the infant has a tape recorder in the auditory cortex that records input sounds, replays them, and accustoms the ear to their patterns.

Babies also demonstrate preferences for the language that resembles the speech of their mothers. Thus, a French infant will prefer to listen to French, whereas a Polish infant will prefer to listen to Polish (Jusczyk 1997). In addition, babies demonstrate a preference for their own mother's voice, as opposed to that of other women. Together, these abilities and preferences suggest that, during the first eight months, the child is remarkably attentive to language. Although children are not yet learning words, they are acquiring the basic auditory and intonational patterns of their native language. As they sharpen their ability to hear the contrasts of their native language, they begin to lose the ability to hear contrasts not represented in the native language. If the child is growing

up in a bilingual world, full perceptual flexibility is maintained. However, if the child is growing up monolingual, flexibility in processing is gradually traded off for quickness and automaticity.

2. Early Articulation

During the first three months, a baby's vocalizations are nothing more than cries and vegetative adaptations. However, around three months, at the time of the first social smile, babies begin to make the delightful little sounds that we call 'cooing.' These sounds have no particular linguistic structure, but their well-integrated intonation makes them sure parent pleasers. By six months, the baby is producing structured vocalizations, including a larger diversity of nasals, vowel types, and syllables with the canonical consonant–vowel (CV) structure. The basic framework of early babbling seems to be constructed on top of patterns of noisy lip-smacking that are present in many primates (MacNeilage 1998). These vocal gestures include some form of vocal closure followed by a release with vocalic resonance. Essentially, this is the CV syllable in which a consonant is followed by a vowel.

Until the sixth month, deaf infants continue to babble normally. However, by the age of nine months, deaf infants lose their interest in babbling. This suggests that earlier babbling is sustained largely through proprioceptive and somaesthetic feedback, as babies explore the various ways in which they can play with their mouth. After six months, babbling relies increasingly on auditory feedback. During this period, the infant tries to produce specific sounds to match up with specific auditory impressions. It is at this point that deaf children no longer find babbling entertaining, since they cannot obtain auditory feedback. These facts suggest that, from the infant's point of view, babbling is essentially a process of self-entertainment.

In the heyday of behaviorism, researchers viewed the development of babbling in terms of reinforcement theory. They thought that the reinforcing qualities of language would lead a Chinese baby to babble the sounds of Chinese, whereas a Quechua baby would babble the sounds of Quechua. This was the theory of 'babbling drift.' However, closer observation of the babbling of eight-month-olds indicates that virtually no such drift occurs. By 12 months, there is some slight drift in the direction of the native language, as the infant begins to acquire the first words. Proponents of universal phonology have sometimes suggested that all children engage in babbling all the sounds of all the world's language. Here, again, the claim seems to be overstated. Although it is certainly true that some English-learning infants will produce Bantu clicks and Quechua implosives, not all children produce all of these sounds.

3. The First Words

The child's ability to produce the first word is based on three earlier developments. The first is the infant's growing ability to record the sounds of words. The second is the development of an ability to control vocal productions that occurs in the late stages of babbling. The third is the general growth of the symbolic function, as represented in play, imitation, and object manipulation. Piaget (1954) characterized the infant's cognitive development in terms of the growth of representation or the 'object concept.' In the first six months of life, the child is unable to think about objects that are not physically present. However, as infants learn more about objects, they become able to associate their properties with their own actions and other features of the context. In this way, subtle cues can be used to dredge up fuller representations from memory. For example, a child may see a dog's tail sticking out from behind a chair and realize that the rest of the dog is hiding behind the chair. This understanding of how parts relate to wholes supports the child's first major use of the symbolic function. When playing with toys, the 12-month-old will begin to produce sounds such as *vroom* or *bam-bam* that represent properties of these toys and actions. Often these phonologically consistent forms (Peters 1983) appear before the first real words. Because they have no clear conventional status, parents may tend to ignore these first symbolic attempts as nothing more than spurious productions or babbling.

Even before children have produced their first conventional word, they have already acquired an ability to comprehend as many as 10 conventional forms. They learn these forms through frequent associations between actions, objects, and words. Parents often realize that the prelinguistic infant is beginning to understand what they say. However, they are hard-pressed to demonstrate this ability convincingly. Researchers deal with this problem by bringing infants into the laboratory, placing them in comfortable highchairs, and asking them to look at pictures, using the technique of visually reinforced preferential looking (Woodward et al. 1994). A name such as *dog* is produced across loudspeakers. Pictures of two objects are then displayed. In this case, a dog may be on the screen to the right of the baby and a car may be on the screen to the left. If the child looks at the picture that matches the word, a toy bunny pops up and does an amusing drum roll. This convinces babies that they have chosen correctly and they then do the best they can to look at the correct picture on each trial. Some children get fussy after only a few trials, but others last for 10 trials or more at one sitting and provide reliable evidence that they have begun to understand a few basic words. Many children show this level of understanding by the tenth month—often 2–3 months before the child has produced a recognizable 'first word.'

Producing the first word is a bit like stepping out on stage. When babbling for their own entertainment, the only constraints infants faced were ones arising from their own playfulness and interest. If they wanted to allow a sound to vary within certain limits, that was fine. Now, when faced with the task of producing word forms, the articulation has to be extremely accurate and within conventional limits. The forms of early words often deviate radically from the adult standard. The many simplifications that the one-year-old introduces to adult phonology are well known to students of phonological development (Menn and Stoel-Gammon 1995). Children tend to drop unstressed syllables, producing *hippopotamus* as *poma*. They repeat consonants, producing *water* as *wawa*. And they simplify and reduce consonant clusters, producing *tree* as *pee*. These phonological processes echo similar processes found in the historical development and dialectal variation of adult language. What is different in child language is the fact that so many simplifications occur at once, making so many words difficult to recognize.

4. Early Semantics

As the child's stock of words grows, it becomes harder to keep words apart from each other. To solve this problem, children have to choose between two opposing strategies. On the one hand, children may try to be conservative in their first uses of words. For example, a child may use the word *dog* to refer only to the family dog and not to any other dog. Or a child may use the word *car* to refer only to cars parked outside a certain balcony in the house and not cars in any other context. This tendency toward undergeneralization can only be detected if one takes careful note of the contexts in which a child avoids using a word. The flip side of this coin is the strategy of overgeneralization. It is extremely easy to detect overgeneralizations. If the child calls a tiger a *kitty*, this is clear evidence for overgeneralization. Of course, it is always possible that the children really meant to say something like, 'That animal over there reminds me a lot of my kitty.' However, if the child intended this, they would be operating with a rather strange set of ideas about how words are used.

At first, both undergeneralization and overgeneralization are applied in a relatively uncontrolled fashion. Children's first applications of undergeneralization are unreasonably rigid and they soon learn that words are meant to be generalized. At the same time, the child's first attempts at generalization are also often wildly overproductive. For example, a child may use the word 'duck' first to refer to the duck, then to the picture of an eagle on the back of a coin, then to a lake where there were ducks, and finally to other bodies of water. These 'pleonastic' extensions of forms across situations are fairly rare, but they provide interesting commentary regarding the thinking of the toddler when they do occur.

Scholars from Plato to Quine have considered the task of figuring out word meaning to be a core intellectual challenge. Quine (1960) illustrated the problem by imagining a scenario in which a hunter is out on safari with a native guide. Suddenly, the guide shouts 'Gavagai!' and the hunter, who does not know the native language, has to quickly infer the meaning of the word. Does it mean 'Shoot now' or 'There's a rhino' or perhaps even 'It got away'? Without some additional cues regarding the likely meaning of the word, how can the hunter figure this out?

The problem facing the toddler is similar to that facing the hunter. Fortunately, the toddler has some good cues to rely on. Foremost among these cues is the parent's use of joint attention and shared eye gaze to establish common reference for objects and actions. If the father says *hippo* while holding a hippopotamus in his hand, the child can use the manual, visual, verbal, and proxemic cues to infer that the word *hippo* refers to the hippopotamus. A similar strategy works for the learning of the names of easily produced actions such as falling, running, or eating. It also works for social activities such as *bath* or *bye-bye*. The normal child probably understands the role of shared eye gaze even before learning the first words. At three months, children maintain constant shared eye gaze with their parents. In normal children, this conta maintains and deepens over time. For autistic children, contact is less stable and automatic. As a result, autistic children may be delayed in word learning and the general growth of communication.

Shared reference is not the only cue toddlers use to delineate pick out the reference of words. They also uses the form of utterances to derive the meanings of new words. For example, if toddlers hear 'Here is a zav', they know that *zav* is a common noun. However, if they hear 'Here is Zav', then they know that *Zav* is either a proper noun or perhaps the name of a quantity. If they hear 'I want some zav', they know that *zav* is a quantity and not a proper or common noun. Cues of this type can give a child a rough idea of the meaning of a new word (Smith 1999). Other sentential frames can give an even more precise meaning. If the child hears, 'This is not green, it is chartreuse', then it is clear that *chartreuse* is a color. If children hear, 'Please don't cover it, just sprinkle it lightly', then they know that *sprinkle* is a verb of the same general class of *cover*. The use of cues of this type leads to a fast, but shallow, mapping of new words to new meanings.

5. Word Combinations

Throughout the second year, the child struggles with perfecting the sounds and meanings of the first words. For several months, children produces isolated single words. However, eventually they need to go beyond

this one-word level. They need to associate predicates such as *want*, *more*, or *go* with arguments such as *cookie* or *Mommy*. The association of predicates to arguments is the first step in syntactic development. As in the other areas of language development, these first steps are taken in a very gradual fashion. Before producing a smooth combination of two words such as 'my horsie', children will often string together a series of single-word utterances that appear to be searching out some syntactic form. For example, a child might say 'my, that, that, horsie' with pauses between each word. Later, the pauses will be gone and the child will say 'that horsie, my horsie.' This tentative combination of words involves groping on both intonational and semantic levels. On the one hand, children must figure out how to join words together smoothly in production. On the other hand, they also must figure out which words can meaningfully be combined with which others.

As was the case in the learning of single words, this learning is guided by earlier developments in comprehension. We have to assess children's early syntactic comprehension by controlled experiments in the laboratory. Here, again, researchers have used the preferential looking paradigm (Golinkoff et al. 1999). To the right of the child, there is a TV monitor with a movie of Big Bird tickling Cookie Monster. To the child's left, there is a TV monitor with a movie of Cookie Monster tickling Big Bird. The experimenter produces the sentence 'Big Bird is tickling Cookie Monster.' If children look at the matching TV monitor, they are reinforced and a correct look is scored. Using this technique, researchers have found that 17-month-olds already have a good idea about the correct word order for English sentences. This is about 5–6 months before they begin to use word order systematically in production.

The grammar of the first word combinations is extremely basic. The child learns that each predicate should appear in a constant position vis-à-vis the arguments it requires. For example, in English, the word *more* appears before the noun it modifies and the verb *run* appears after the subject with which it combines. Slot-filler relations can control this basic type of grammatical combination. Each predicate specifies a slot for the argument. For example, *more* has a slot for a following noun. When a noun, such as *milk*, is selected to appear with *more*, that noun becomes a filler for the slot opened up by the word *more*. The result is the combination 'more milk.' Later, the child can treat this whole unit as an argument to the verb *want* and the result is 'want more milk.' Finally, the child can express the second argument of the verb *want* and the result is 'I want more milk.' Thus, bit by bit, children build up longer sentences and a more complex grammar. This level of simple combinatorial grammar is based on individual words as the controlling structures. This type of word-based learning is present even in adults. In languages with strong morphological marking systems, word-based patterns specify the attachment of affixes, rather than just the linear position of words. In fact, most languages of the world make far more use of morphological marking than does English (Slobin 1985). In this regard, English is a rather exotic language.

5.1 Missing Glue

The child's first sentences are almost all incomplete and ungrammatical. Instead of saying, 'This is Mommy's chair', the child produces only 'Mommy chair' with the possessive suffix, the demonstrative, and the copula verb all deleted. Just as the first words are full of phonological deletions and simplifications, the first sentences include only the most important words, without any of the glue. In some cases, children simply have not yet learned the missing words and devices. In other cases, they may know the 'glue words' but find it difficult to coordinate the production of so many words in the correct order.

These early omissions provide evidence for two major processes in language development. First, children try to make sure that the most important and substantive parts of the communication are not omitted. Unfortunately, children make this evaluation from their own, egocentric perspective. In an utterance like 'Mommy chair' it is not clear whether the child means 'This is Mommy's chair' or 'Mommy is sitting in the chair', although the choice between these interpretations may be clear in context (Bloom 1973). The second factor that shapes early omissions is phrasal frequency. Children tend to preserve frequent word combinations, such as 'like it' or 'want some.' These combinations are often treated as units, producing errors such as 'I like it the ball' or 'I want some a banana.'

5.2 Productivity

Productivity can be demonstrated in the laboratory by teaching children names for new objects. For example, we can show a child a picture of a funny-looking creature and call it a *wug*. As we noted before, the positioning of the word *wug* after the article *a* induces the child to treat the word as a common noun. Children can then move from this fact to infer that the noun *wug* can pluralize as *wugs*, even if they have never heard the word *wugs* (MacWhinney 1978).

Three-year-olds also demonstrate some limited productive use of syntactic patterns for new verbs. However, children tend to be conservative and unsure about how to use verbs productively until about age five years. From the child's perspective, these laboratory experiments with strange new toys and new words tend to encourage a conservative approach. As they get older and braver, children start to show productive use of constructions such as the double object, the passive, or the causative (Brooks et al.

1999). For example, an experimenter can introduce a new verb like *griff* in the frame 'Tim griffed the ball to Frank' and the five-year-old will productively generalize to 'Tim griffed Frank the ball.'

The control of productivity is based on two complementary sets of cues: semantics and co-occurrence. When children hear *a wug*, they correctly infer that *wug* is a count noun. In fact, because they also see a picture of a cute little animal, they infer that *wug* is a common, count, name for an animate creature. These semantic features allow them to generalize their knowledge by producing the form *wugs*. However, we could also view this extension as based on co-occurrence learning. The child learns that words that take the indefinite article also form plurals. On the other hand, words that take the quantifier *some* do not form plurals. In this way, the child can use both semantic and co-occurrence information to build up knowledge about the parts of speech. This knowledge can then be fed into existing syntactic generalizations to produce new combinations and new forms of newly learned words. The bulk of grammatical acquisition relies on this process.

5.3 The Logical Problem of Language Acquisition

The problem with productivity is that it produces overgeneralization. For example, an English-speaking child will soon learn to form the past tense of a new verb by adding one of the variant forms of *-ed*. This knowledge helps the child produce forms such as *jumped* or *wanted*. Unfortunately, it may also lead the child to produce an error such as *goed*. When this occurs, we can say that the child has formulated an overly general grammar. One way of convincing the child to reject the overly general grammar in which *goed* occurs is to provide the child with negative feedback. This requires the parent to tell the child, 'No, you can't say *goed*.' The problem here is that children seem to ignore parental feedback regarding the form of language. If the child calls a hamburger a *hot dog*, the parent can say, 'No, it is a hamburger.' The child will accept this type of semantic correction. But children are notoriously resistant to being corrected for formal grammatical features.

The fact that children tend to ignore formal correction has important consequences for language acquisition theory. In the 1960s, work in formal analysis convinced some linguists that the task of learning the grammar of a language was impossible unless negative feedback was provided. Since negative feedback appeared to be unavailable or unused, this meant that language could not be learned without some additional innate constraints. This argument has led to many hundreds of research articles exploring the ways in which children's learning places constraints on the form of grammar. Referring back to Plato's *Republic* and the Allegory of the Cave, Chomsky and others have characterized the task of language learning as a logical problem.

In fact, children have more resources available to them than Chomsky seems to suggest. Using these resources, the child can recover from overgeneralization without negative feedback. In the case of *goed*, everyone agrees that recovery is easy. All the child has to do is to realize that there is only one way of producing the past tense of *go* and that is *went*. In other words, the irregular form *went* comes to block production of the overregularized form *goed*. Here, recovery from overgeneralization is based on the competition between the regular pattern and the irregular form.

Consider another example. Suppose that a child decides that the verb *recommend* patterns like the verb *give*. After all, both verbs involve a beneficiary and an object being transferred. However, only *give* allows a double object construction, as in 'John gave the library the book.' Most people find 'John recommended the library the book' ungrammatical. One solution to this error is to avoid making it in the first place. However, a more general solution is to record the strength of the competing syntactic patterns. The correct way of saying 'John recommended the library the book' is to say 'John recommended the book to the library.' This correct formulation should be strengthened whenever it is heard. As the strength of the frame for the verb *recommend* grows in comparison to the ungrammatical frame, the use of the competing frame is blocked. This solution assumes that the child realizes that the two frames are in competition. It may be that reaching this realization requires some attention to syntactic form. However, this solution does not require children to pay attention to corrective feedback. Instead, they only need to attend to correct sentences and to make sure that they understand that these are competing ways of saying roughly the same thing.

6. Emergentist Accounts

In the 1980s, a sharp division existed between accounts of language acquisition that emphasize learning and those that emphasize innate knowledge of universal grammar. At the start of the twenty-first century, there is no longer a clear separation between alternative approaches. Instead, the challenge facing researchers is to formulate a sets of mechanisms that are rich enough to account for the details of these emergent processes and which are grounded on the facts of human neurophysiology. To achieve this, many psychologists rely on the formalisms of neural network theory, which is also known as connectionism (McClelland and Rumelhart 1986). This framework uses large numbers of units and the connections between these units to capture the patterns of language. This web-like architecture of nodes and con-

nections is intended explicitly to resemble the structure of the human brain with neurons, synapses, and weights on synaptic connections. Neural network models have provided good accounts for the acquisition of morphology (MacWhinney et al. 1989), phonology (Plaut and Kello 1999), and syntax (Allen and Seidenberg 1999).

A major limitation of many current models is their failure to pay sufficient attention to the child's perspective on learning. In the late 1990s, researchers began to correct this problem. Models that rely on concepts derived from dynamic systems theory (Smith 1999) and the theory of attention (Merriman 1999) have proven useful in accounting for word learning by focusing on the ways in which concrete situations provide learners with cues. Newer formal models also relate word learning to children's physical control of their own body (Bailey et al. 1997) and their own perspective as a causal agent (MacWhinney 1999). These new developments may also allow us to understand how language works to embed children into the fabric of social life (Vygotsky 1962).

7. Conclusion

Language is a unique marker of humanity. It distinguishes the human species from the rest of the creation, and it allows us to share our thoughts and feelings. Language is the most complex skill that any of us will ever master. Despite this complexity, nearly every human child succeeds in learning language. This suggests that language is optimally shaped to mesh with the abilities of the human mind and body. On the one hand, the universals of human language match up with our neurological, cognitive, and physical abilities. At the same time, parents provide rich contextual and emotional support to guide children through the process of language discovery. By studying language learning, we learn more about universals of human language, the shape of social interaction, and the structure of the human mind.

See also: First Language Acquisition: Cross-linguistic; Language Learning Impairment

Bibliography

Allen J, Seidenberg M 1999 The emergence of grammaticality in connectionist networks. In: MacWhinney B J (ed.) *The Emergence of Language*. Erlbaum, Mahwah, NJ, pp. 115–52

Bailey D, Feldman J, Narayanan S, Lakoff G 1997 Modeling embodied lexical development. In: Gernsbacher M (ed.) *Proceedings of the 19th Meeting of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ, pp. 18–22

Bloom L 1973 *One Word at a Time: The Use of Single Word Utterances*. Mouton, The Hague

Brooks P J, Tomasello M, Dodson K, Lewis L B 1999 Young children's overgeneralizations with fixed transitivity verbs. *Child Development* **70**: 1325–37

Chomsky N 1982 *Some Concepts and Consequences of the Theory of Government and Binding*. MIT Press, Cambridge, MA

Fletcher P, MacWhinney B J (eds.) 1995 *The Handbook of Child Language*. Blackwell, Oxford, UK

Golinkoff R M, Hirsh-Pasek K, Hollich G 1999 Emergent cues for early word learning. In: MacWhinney B J (ed.) *The Emergence of Language*. Erlbaum, Mahwah, NJ, pp. 305–30

James W 1890 *The Principles of Psychology*. Holt, New York

Jusczyk P W 1997 *The Discovery of Spoken Language*. MIT Press, Cambridge, MA

Lenneberg E H 1967 *Biological Foundations of Language*. Wiley, New York

MacNeilage P 1998 The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences* **21**: 499–546

MacWhinney B J 1978 The acquisition of morphophonology. *Monographs of the Society for Research in Child Development* **43** (whole issue No. 1): 1–123

MacWhinney B J 1999 The emergence of language from embodiment. In: MacWhinney B J (ed.) *The Emergence of Language*. Erlbaum, Mahwah, NJ, 213–56

MacWhinney B J, Leinbach J, Taraban R, McDonald J L 1989 Language learning: cues or rules? *Journal of Memory and Language* **28**: 255–77

Menn L, Stoel-Gammon C 1995 Phonological development. In: Fletcher P, MacWhinney B J (eds.) *The Handbook of Child Language*. Blackwell, Oxford, UK, pp. 335–60

Merriman W 1999 Competition, attention, and young children's lexical processing. In: MacWhinney B J (ed.) *The Emergence of Language*. Erlbaum, Mahwah, NJ, pp. 331–58

Peters A M 1983 *The Units of Language Acquisition*. Cambridge University Press, New York

Piaget J 1954 *The Construction of Reality in the Child*. Basic Books, New York

Plaut D C, Kello C T 1999 The emergence of phonology from the interplay of speech comprehension and production: a distributed connectionist approach. In: MacWhinney B J (ed.) *The Emergence of Language*. Erlbaum, Mahwah, NJ, pp. 381–416

Quine W V O 1960 *Word and Object*. MIT Press, Cambridge, MA

Rumelhart D E, McClelland J L 1986 *Parallel Distributed Processing*. MIT Press, Cambridge, MA

Saffran J R, Aslin R N, Newport E L 1996 Statistical learning by 8-month-old infants. *Science* **274**: 1926–8

Slobin D I 1985 Crosslinguistic evidence for the language-making capacity. In: Slobin D (ed.) *The Crosslinguistic Study of Language Acquisition. Vol. 2: Theoretical Issues*. Erlbaum, Hillsdale, NJ, pp. 1157–256

Smith L B 1999 Children's noun learning: how general processes make specialized learning mechanisms. In: MacWhinney B J (ed.) *The Emergence of Language*. Erlbaum, Mahwah, NJ, pp. 277–304

Vygotsky L S 1962 *Thought and Language*. MIT Press, Cambridge, MA

Werker J. F 1995 Exploring developmental changes in cross-language speech perception. In: Gleitman L, Liberman M (eds.) *An Invitation to Cognitive Science. Language*. MIT Press, Cambridge, MA, Vol. 1, pp. 87–106

Woodward A L, Markman E M, Fitzsimmons C M 1994 Rapid word learning in 13- and 18-month-olds. *Developmental Psychology* **30**: 553–66

B. J. MacWhinney

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Language and Animal Competencies

If language were only a unitary dimension, we might, indeed, insist that it is language that sets humans apart from animals—immutably and forever. But language is not a digital phenomenon. Just as 'Love is a many splendored thing…,' as a song proclaims, so is *language*! It is a highly complex system that seemingly culminates with the eloquent use of voice in oration, poetry, and song as well as in debate, lectures, argument, and conversation. As a consequence, language has frequently been equated with speech and studied accordingly. We are convinced, however, that speech is only one of a set of media that allows for the expression of language, hence not a requisite to language. If the highly effective and coordinated mechanisms of speech and hearing had not evolved, our species would likely have refined other basic media for language—graphics, whistling, signing, drumming, and so on. Thus, we hold that although learning to talk/speak is a developmental imperative unique to humans, the basic requisites of language are not. Indeed, research documents that several animal species have the potential to acquire processes and skills that are essential to language as we know it in our own species.

At least one notable scholar of human language has asserted, in a recent lecture, that the question, 'Do animals have language?' is not scientific (Pinker 1999). Why? Because language cannot be defined to everyone's satisfaction. By that token, of course, then no one (and certainly not the scholar of reference) should assert that animals do *not* have language. The fact is that definitions of constructs (e.g., personality, learning, thinking, intelligence, language, etc.) can never be honed to everyone's satisfaction. Notwithstanding, definitions of terms are necessary to science.

Here, our working definition is that

> language is a biobehavioral system that provides for the learning of meanings of symbols and for their use to exchange information within a social context. Symbols may be based in any modality—sound (e.g. speech, hearing, drumming, whistling, etc.), vision (e.g. writing, geometric figures, movements with hands and body, etc.), touch (e.g. Braille), and so on.

It is meaningfulness that differentiates symbols/words from other stimuli. The meanings of words (e.g., semantics) are the result of social usage, hence the meanings of words are subject to radical change across time. The language system is open in that new words may be added and old words deleted. The language system also provides for the generation of novel messages through the interactive use of words (e.g., rules of grammar and syntax provide for 'I'm buying a house dog' to differ radically in meaning from 'I'm buying a dog house,' though the words used remain constant). The ability to comprehend language is an absolute requisite to competence in productive use/expressions of language. Hence, the first essential process of language acquisition by the human child is comprehension, that is, understanding language used by others. Later, the child typically begins to use/produce that language through speech and gesture.

The following review of animal language research, conducted over the course of the past half-century, documents that language is a continuum, a highly complex biobehavioral system that is clearly traceable to animals, and notably to the great apes (Gardner et al. 1989, Premack 1970, Rumbaugh 1977, Fouts and Fouts 1989, Miles 1990, Savage-Rumbaugh and Lewin 1994), pinnipeds (Herman 1987, Schusterman et al. 1993), and parrots (Pepperberg 1990), all having demonstrated the ability to learn and use symbols with semantic competence (e.g., for symbols to acquire meaning and to represent things/events not necessarily present). The primary foundation of language is in the comprehension of speech or of the units of any other language system, and not in the production of speech. Because of its openness and generativity (e.g., to the addition of new word-symbols and to the modulation of their meanings through social commerce), language should not be viewed as or called an instinct (e.g., unlearned, highly predictable expression of genetically defined behaviors, synchronized and expressed through the course of biological development and maturation). The acquisition of language, whether by the human child or ape, is highly contingent on specific aspects of early rearing; and apes, if reared from birth in a language-structured environment (e.g., one in which language is used throughout all dimensions of their rearing experiences to the end of encouraging their comprehension and anticipation of all events), can understand hundreds of novel sentences of request that they hear even for the first occasion in tightly controlled research procedures. Formal training procedures, based on stimulus–response–reinforcement, used to cultivate language competence in animals are less effective than appropriate social/language rearing to the cultivation of language and notably to the ability to comprehend human speech. Apes can use their language skills to coordinate problem solving, either between themselves or with caregivers, that otherwise would be impossible (Menzel 1999). Apes' brains are lateralized both in structure and function, though the circuits/systems used for language might not, in detail, be identical to ours in language acquisition and use (Rilling et al. 1999).

1. *Language Acquisition in Animals*

Research by Keith and Cathy Hayes (Hayes and Nissen 1971) clearly demonstrated that chimpanzees could not speak with any significant facility and that if they had any potential for language, another medium of expression would have to be developed in order to study it. Allen and Beatrix Gardner (1969) used a modification of the American Sign Language for the Deaf, a manual and gestural language system of a kind

that Yerkes and Learned (1925) suggested might be fruitful for the study of apes' language potential. Manual signing systems were also used by Miles (1990) and Terrace (1979). Premack (1971) used a set of plastic tokens as units of a vocabulary and made significant contributions to defining the essence of words.

Rumbaugh's (1977) group succeeded in developing a computer-monitored keyboard and system with which the chimpanzee, Lana, became very proficient in learning 'stock sentences of request' that were effective in obtaining food (i.e., Please machine give piece of banana), viewings of movies, slides, and the outdoors (i.e., Please machine make window open), and special human companionship and other privileges (i.e., Please Tim tickle Lana). All keys of her keyboard were embossed with distinctive lexigrams that functioned as words—hence they are termed 'word-lexigrams.' Lana's many accomplishments included the learning of names of six colors and six objects (with each object being painted each of the six colors, for a total of 36 items), with which she could give, upon request, either the color or the name of a designated object from a set of several objects. For instance, with a random array of the objects displayed to her, she could accurately answer questions, such as 'What name of this that's (red, green, etc.)?' vs. 'What color of this shoe?'

Somewhat later, chimpanzees Sherman and Austin (Savage-Rumbaugh 1986) learned functionally how to use lexigrams in order to make requests for specific items. Eventually it was clear that for them their lexigrams actually represented the items that they had earlier simply requested. This conclusion was based on their near errorless ability to classify via two lexigrams, one standing for 'food' and the other for 'tool,' each of 17 lexigrams on a trial-1 basis in a controlled test. With these basic skills of using lexigrams to request items or to name them, they then were able to move on to a number of remarkable interactions that entailed their coordinated communications and behaviors to solve problems that netted prized incentives for them and, also, to announce what it was that they were about to do—to announce which item it would be that they would retrieve from a random set of foods and drinks (Savage-Rumbaugh 1986).

An even more significant finding was obtained in work with bonobos, a rare form of chimpanzee (*Pan paniscus*). Sue Savage-Rumbaugh and her colleagues (Savage-Rumbaugh et al. 1993, Savage-Rumbaugh and Lewin 1994) worked with Matata, a bonobo, with training methods that had worked so well with Sherman and Austin (common chimpanzees, *P. troglodytes*). With Matata, however, they had almost no success. Serendipitously, however, it was discovered that Matata's adopted son, Kanzi, who as an infant only observed Matata's training situation, had spontaneously learned (e.g., without formal training) it all! It was only when Matata was separated from him, when he was $2\frac{1}{2}$ years old, that his learning became manifest. Then, at his own initiative, he formulated novel requests, announced intended actions, and proposed games that were not even part of Matata's sessions. By age four years, he understood several dozen spoken words. His learning as an infant was noted for its 'silence' in that his behaviors had given no indication to the researchers that he was mastering such complex relationships and language skills. By age four years, Kanzi was competently understanding and using lexigrams with his caregivers and, quite unexpectedly, gave evidence of understanding much of his caregivers' spoken English. Formal tests revealed that he knew the referents/meanings of at least 150 spoken words.

Kanzi's accomplishments go well beyond the highly significant fact that by observation alone he learned his initial lexicon (Savage-Rumbaugh and Lewin 1994, Savage-Rumbaugh et al. 1998) and also eventually came to understand spoken English words. Although there have been long-standing assertions that an ape, here and there, 'understands' a few English words, none began to equal Kanzi's understanding of hundreds of words—not just individual words, but novel sentences of request as well. Kanzi, age seven years, was tested with Alia, a $2\frac{1}{2}$-year-old child, with 660 novel sentences (e.g., neither familiar, trained, or modeled by others). Some requested that they act in reference to places, persons, and materials, while others called for them to act with reference to objects and locations (Savage-Rumbaugh et al. 1993)—'Take the gorilla (doll) to the bedroom'; 'Give Karen a carrot,' 'Take the vacuum cleaner outdoors,' 'Hammer the snake,' or 'Get the lettuce that's in the microwave oven,' and so on.

In addition, there were a variety of requests that entailed reversals such as 'Make the doggie bite the snake,' vs. 'Make the snake bite the doggie.' Alia and Kanzi promptly fulfilled about 70 percent of the novel requests upon their first presentation. On reversal requests, Kanzi was 81 percent correct, whereas Alia was 64 percent correct.

Kanzi's remarkable achievements, traceable to his early rearing, are in sharp contrast to the devastating effect that impoverished rearing during early infancy has upon the cognitive and social competence of common chimpanzees. If their rearing environments (e.g., the first two years or so) are sorely limited and socially impoverished, they never fully recover. They remain both cognitively and socially compromised (e.g., poor at breeding, parenting, and social communication commerce), even when tested at age 14 years as young adults (Davenport et al. 1973). Primates' general social and cognitive competence are very dependent on their early environment. The same is true for the human child.

Significant research has also been reported with a parrot. Pepperberg (1990) trained Alex, an African gray parrot, using what she called the 'model-rival'

method. Here, Alex observed linguistic interactions between two people who served as 'models' and 'rivals.' The model/rivals talked about and exchanged objects of interest to Alex. Alex entered the conversation at will and eventually was highly skilled in asking for, choosing, or describing the color, shape, and materials of the objects with which he was familiar. For both Kanzi and Alex, the language acquisition contexts were designed to encourage spontaneous responding and designed such that uses of words were efficacious. Typical classical or operant conditioning experiments have emphasized the mastery of a single response in a specific context. By contrast, the training contexts for Kanzi and for Alex drew upon a great variety of different lexical responses.

The data of these projects and successful replications thereof (Brakke and Savage-Rumbaugh 1995, 1996, Savage-Rumbaugh et al. 1998, Pepperberg 1999) have had major affects upon how we now should view animals' abilities.

2. *Tool Making by Observational Learning*

Kanzi also learned by observation how to make and use stone tools to gain access to baited food sites. Kanzi was never trained to create the flakes in any specific manner. His model was an anthropologist, Nick Toth, who created stone flakes through striking two rocks together (Savage-Rumbaugh and Lewin 1994, Toth et al. 1993). Over time, Kanzi acquired skill both in creating stone flakes and evaluating their probable effectiveness for cutting ropes of varying thickness. As a tool lost its cutting edge, Kanzi would make a new, more effective one. Kanzi's causal understanding of tool fabrication and use was demonstrated in three very impressive ways. Early in his efforts, he innovatively started to throw one rock on to a hard surface or on to another rock on the ground. Thus, Kanzi first appeared to understand that it was the fundamental force of one rock striking another that created the flakes. Second, he rapidly learned how to assess whether a given flake would be sufficient to the task at hand—be it to cut a rope or a piece of leather. Inadequate chips were discarded and replaced with new ones as needed. Third, eventually Kanzi came to hold the cobble so as to hit it on an edge of less than 90 degrees, thereby producing large and very sharp chips of flint.

3. *Language and Inferences of Causality*

There is evidence that language serves to make animals more competent in general—as though they understand more readily cause–effect relationships and how things work (Michotte 1963, Piaget 1930, 1974). For instance, Kanzi quickly learned, by brief observation, how to use a joystick to control the movement of a cursor-icon on a monitor and to play video games, including PacMan. Earlier, both Sherman and Austin had quickly learned the use of a joystick, again as a result of a brief demonstration (< 15 min) (Savage-Rumbaugh 1986, Savage-Rumbaugh and Lewin 1994). We had anticipated that they would have great trouble in learning to use the joystick (Rumbaugh et al. 1989) and were surprised that they learned so quickly and by observation alone. They seemingly saw/understood 'how it worked' very readily and what was to be gained by using it with precision. By striking contrast, apes with little or no language skills from our laboratory have generally required either protracted operant training or months of observing a competent cohort.

4. *Cooperation Enabled by Symbolic Communication*

Savage-Rumbaugh's (1986) project with Sherman and Austin was designed to cultivate symbolic communication so as to coordinate problem-solving efforts. After considerable training on a variety of tasks, the apes became quite accomplished in their cooperative efforts. For example, one of the chimpanzees could observe food being placed in a container that required a specific tool for opening it—a requisite to getting the prized incentive within. The other chimpanzee was not allowed to watch the placement of the food into a secured container, and, thus, did not know what tool was needed. Tools needed varied with containers and included a lever, key, a sponge, a cord, and so on. The chimpanzee who saw the food secured thus could infer what tool would be needed. A request by that chimpanzee was then made of the other one to supply it with a tool, specified by selection of the appropriate word-lexigrams at the keyboard that contained more than 100 symbols. If the symbol-encoded request was fulfilled, then the chimpanzee receiving it used the tool to open the secured container. The incentive, thus obtained, was then shared with the chimpanzee who complied with the tool-users' request. Both chimpanzees were competent in both tasks—asking for and vending the tool as specified across trials. The knowledge base with which they coordinated their efforts was documented on the several occasions on which Austin by mistake gave Sherman an incorrect tool (e.g., not the one requested). On those occasions, Sherman would throw the tool down (e.g., reject it) and then bring Austin's attention once again, via the lexigrams keyboard, to the name of the tool that had been requested and needed for the incentive to be obtained.

Clearly, this task could be solved only by social communication through the use of learned symbols. Savage-Rumbaugh (1986; also Savage-Rumbaugh et al. 1998) reports many other strong instances of goal-

directed communication. For instance, when Sherman and Austin spontaneously used labels from food cans and boxes to communicate when their keyboard was denied to them (e.g., turned off) and also when Kanzi and others helped novice research assistants find word-lexigrams that they were having difficulty finding for their own use on the keyboard.

For these and other reasons, we believe it constructive to add the category of emergents to the traditional dichotomy of learning embraced by the traditional dichotomy of respondent and operant conditioning. Emergents are exemplified in the appearance of highly complex and even unanticipated competencies on the basis of protracted generalized, rather than specific, experiences and through observation (Rumbaugh et al. 1996).

5. *Summary and Conclusions*

At the dawn of a new millennium and consonant with Darwin's postulate of psychological as well as biological continuity between humans and other animals, there is strong evidence that language should be viewed as a continuum (Deacon 1998), with its fundamental roots being firmly planted most clearly in apes—our nearest living relatives (Andrews and Martin 1987). This evidence, rooted in 50 years of dedicated effort measured in decades and careers of scientists and staff, is solid. It has survived countless peer reviews for funding and publication. Apes, pinnipeds, and parrots have potentials for language skills that can be cultivated through early rearing and social exchanges. Yet there are a few influential scholars of note who, quite matter of factly, reveal at best only a limited understanding of the research or even total disinterest. They fail to recognize the evolutionary and neurobiological significance of the animal research and seemingly are satisfied to accept rhetoric in lieu of data and scientific reason. Yet, time surely will vindicate those who argue, from scientific data, that the evidence of animal species' potential for language is of inestimable significance from the perspectives of science, philosophy, language, and public policy. For science, the data vindicates the Darwinian perspective of psychological continuity from animal to human, a continuity afforded by biological continuity (Andrews and Martin 1987). For philosophy, the data declare that human quality of life, thought, and culture need to be understood from a comparative perspective. For students of language, the data declare that the roots, primitive though they might be, for language and cognition are traceable to nonhuman genes (e.g., to animals). For public policy, the data declare that animals are not the mindless, senseless beast machines that Descartes declared them to be. Unique we humans are, to be certain; but then, so are all other species of life. More powerful than they we are, though whether we are more worthy merits reconsideration. At least some animals, notably the great apes, are not properly viewed as chattel, for they are sentient thoughtful beings, in measure. Their behaviors tell us of the evolutionary history of eons past of animals' ground that they share in common with us.

See also: Animal Cognition; Language Acquisition; Linguistics: Overview; Sign Language: Psychological and Neural Aspects

Bibliography

Andrews P, Martin L 1987 Cladistic relationships of extant and fossil hominoids. *Journal of Human Evolution* **16**: 101–18

Brakke K E, Savage-Rumbaugh E S 1995 The development of language skills in bonobo and chimpanzee—I. Comprehension. *Language and Communication* **15**: 121–48

Brakke K E, Savage-Rumbaugh E S 1996 The development of language skills in Pan—II. Production. *Language and Communication* **16**: 361–80

Davenport R K, Rogers C M, Rumbaugh D M 1973 Long-term cognitive deficits in chimpanzees associated with early impoverished rearing. *Developmental Psychology* **9**: 343–7

Deacon T 1998 *The Symbolic Species*. Horton, New York

Fouts R S, Fouts D H 1989 Loulis in conversation with the cross-fostered chimpanzees. In: Gardner R A, Gardner B T, Van Cantfort T E (eds.) *Teaching Sign Language to Chimpanzees*. State University of New York Press, New York pp. 293–307

Gardner R A, Gardner B T 1969 Teaching sign language to a chimpanzee. *Science* **165**: 664–72

Gardner R A, Gardner B T, Van Cantfort T E (eds.) 1989 *Teaching Sign Language to Chimpanzees*. State University of New York Press, New York

Hayes K J, Nissen C 1971 Higher mental functions of a home-raised chimpanzee. In: Schrier A M, Stollnitz F (eds.) *Behavior of Nonhuman Primates: Modern Research Trends*. Academic Press, New York, Vol. 4, pp. 59–115

Herman L M 1987 Receptive competencies of language-trained animals. *Advances in the Study of Behavior* **17**: 1–60

Menzel C 1999 Unprompted recall and reporting of hidden objects by a chimpanzee (Pan troglodytes) after extended delays. *Journal of Comparative Psychology* **113**: 426–34

Michotte A 1963 *The Perception of Causality* (trans. Miles T R). Basic Books, New York

Miles H L W 1990 The cognitive foundations for reference in a signing orangutan. In: Parker S T, Gibson K R (eds.) *'Language' and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*. Cambridge University Press, New York pp. 511–39

Pepperberg I M 1990 Conceptual abilities of some nonprimate species, with an emphasis on the African Grey parrot. In: Parker S T, Gibson K R (eds.) *'Language' and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*. Cambridge University Press, New York, pp. 469–507

Pepperberg I M 1999 *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots*. Harvard University Press, Cambridge, MA

Piaget J 1930 *The Child's Conception of Physical Causality*. Routledge and Kegan Paul, London

Piaget J 1974 *Understanding Causality*. Norton, New York

Pinker S 1994 *The Language Instinct*. Morrow, New York

Pinker S 1999 *The Origins of Language*, Lecture given at the Living Links Origins Symposium, Emory University, January 15–16, 1999
Premack D 1970 A functional analysis of language. *Journal of the Experimental Analysis of Behavior* **14**: 107–25
Rilling J, Kilts C, Williams S, Kelley J, Beran M, Giroux M, Hoffman J M, Savage-Rumbaugh S, Rumbaugh D 1999 Function neuroimaging of linguistic processing in chimpanzees. *Society of Neuroscience Abstracts* **25**(2): 2170
Rumbaugh D M (ed.) 1977 *Language Learning by a Chimpanzee: The Lana Project*. Academic Press, New York
Rumbaugh D M 1997 Competence, cortex and animal models: a comparative perspective. In: Krasnegor N, Lyon R, Goldman-Rakic P (eds.) *Development of the Prefrontal Cortex: Evolution, Neurobiology and Behavior*. Brookes, Baltimore, MD, pp. 117–139
Rumbaugh D M, Richardson W K, Washburn D A, Savage-Rumbaugh E S, Hopkins W D 1989 Rhesus monkeys (*Macaca mulatta*), video tasks, and implications for stimulus response spatial contiguity. *Journal of Comparative Psychology* **103**: 32–8
Rumbaugh D M, Washburn D A, Hillix W A 1996 Respondents, operants, and emergents: toward an integrated perspective on behavior. In: Pribram K, King J (eds.) *Learning as a Self-Organizing Process*. Erlbaum, Hillsdale, NJ, pp. 57–73
Savage-Rumbaugh E S 1986 *Ape Language: From Conditioned Response to Symbol*. Columbia University Press, New York
Savage-Rumbaugh E S, Lewin R 1994 *Kanzi*. Wiley, New York
Savage-Rumbaugh E S, Murphy J, Sevcik R A, Brakke K E, Williams S L, Rumbaugh D M 1993 Language comprehension in ape and child. *Monographs of the Society for Research in Child Development* **58**(3–4): R5–000
Savage-Rumbaugh E S, Shanker S, Taylor T J 1998 *Apes, Language, and the Human Mind*. Oxford University Press, New York
Schusterman R L, Gisiner R, Grimm B K, Hanggi E B 1993 Behavior control by exclusion and attempts at establishing semanticity in marine mammals using match-to-sample paradigms. In: Roitblat H L, Herman L M, Nachtigall P E (eds.) *Language and Communication: Comparative Perspectives*. Erlbaum, Hillsdale, NJ, pp. 249–74
Terrace H S 1979 *Nim*. Knopf, New York
Toth N, Schick K D, Savage-Rumbaugh E S, Sevcik R A, Rumbaugh D M 1993 Pan the tool-maker: investigations into the stone tool-making and tool-using capabilities of a bonobo (*Pan paniscus*). *Journal of Archaeological Science* **20**: 81–91
Yerkes R M, Learned B W 1925 *Chimpanzee Intelligence and its Vocal Expressions*. Williams and Wilkins, Baltimore, MD

D. M. Rumbaugh and E. S. Savage-Rumbaugh

Language and Ethnicity

As 'language' is defined elsewhere in this volume, only 'ethnicity' needs to be defined here before the relationship between the two is discussed.

Ethnicity is a form of collective identity, and thus belongs to the same class of phenomena as religious affiliation, lineage, clan, or class membership. It is the awareness of belonging to an ethnic group, and the belief that others belong to other groups of this kind. Ethnic groups differ from other groups, like voluntary associations or age-sets, by the following criteria: they comprise people of both genders and all ages, or, as Elwert (1989) puts it, entire families. Even where exogamy extends far beyond the nuclear family, ethnic groups comprise enough such exogamous units to guarantee self-sufficiency in biological reproduction: marriage partners can be provided within the ethnic unit. This means that ethnic groups have the potential to recruit members by birth, and, indeed, that is how they recruit most of their members. It does not mean that this is the only form of recruitment and that there is no ethnic conversion.

The criterion of self-sufficiency in biological reproduction sets the lower limit of the size of an ethnic group. The upper limit is set by the criterion of alterity. As ethnic groups define themselves and are defined in contradistinction to other such groups, an ethnic group can never comprise all of humankind. Ethnic groups tend to be medium-to large-scale human aggregates, typically ranging from thousands to millions. Many ethnic groups, including those which do not have a 'nation-state' to their name, exceed many independent states in the size of their populations. 'ethnic group' and 'nation' form a pair of concepts, the first stressing cultural aspect, the second political ones. 'nation' includes the additional element of statehood in one or the other form. One can say that an ethnic group plus a state in which it predominates or a claim to such separate statehood is a nation. The fact that political rights like self-determination are attributed to nations in the political and juridicial discourse does not help to define which ethnic group qualifies as a nation and which one does not. Any ethnic group can claim to be a nation and become one through such a claim. The only difference between a mere ethnic group and a nation without state is that the latter claims political sovereignty and the former does not necessarily do so. This distinction may not be very helpful in solving practical political issues, but social science does not provide a more helpful one.

Ethnicity is a cultural construct and its construction materials are cultural as well. There are other cultural constructs which are partly based on biological givens, like gender on biological sex, but in all cases in which ethnicity claims to be based on biological facts like discontinuities in the distributions of genes (race), such claims can be shown to be without a biological basis. Genes may cluster in different ways, but they never mark ethnic boundaries. Ethnic groups are never based on race, but very frequently on beliefs about race. Another element ethnic groups tend to attribute to themselves, often counterfactually, is a venerable age, and therefore ethnicity can in some ways be said to be a pseudobiological category: ethnic groups view themselves as natural and stable in time like biological

species. Other parts of culture, in addition to beliefs about race, which play a part in the construction of ethnic groups, include the reference to real or imagined historical events, peculiarities of customs, and the like. No two ethnic groups have a set of criteria which is alike, but typically ethnic discourses make reference to a plurality of such criteria. If only one such criterion is stressed, like religious affiliation, one would not speak of an ethnic group, but, rather, of a religious one.

For our present topic it is of special importance that language tends to be (but by no means always is) high up in the catalog of cultural criteria used to mark the boundaries between ethnic groups. This makes the question of the relationship between language and ethnicity appear tautological in some cases. Where language is used as the primary marker of an ethnic boundary (Barth 1969), ethnic and linguistic units tend to be congruent. In post World War I Europe, the language census has in numerous cases simply replaced the declaration of political will of local communities in effecting boundary corrections between nation-states (Dench 1986). On the other hand, through the school system and the exclusive use of the majority language for administration, national languages or languages which enjoy a special status in certain areas, have been imposed on linguistic minorities (Gellner 1981). Congruence between linguistic communities and ethnic units/nations has thus been promoted on the interstate level by adjusting political boundaries to language distribution areas and on the intrastate level by extending a language policy to the boundaries. In spite of the fact that these processes have been at work for some time, still only a small fraction of the world's nations are linguistically homogeneous (Ra'anan 1989).

While in some cases ethnic and linguistic units are coterminous, at the other extreme we find ethnic units, well defined by other criteria, which have nothing to do with language. The Garre of north eastern Africa conceive of themselves as one ethnic group because of their common genealogy and clan organization, their shared pastoral nomadic culture, and their desire, not always fulfilled, to keep peace with each other. The Garre are divided in four linguistic clusters which cross-cut other criteria of differentiation like clanship. Some of them speak an Oromo dialect close to the one of the Boran, while some speak Af Rahanweyn, yet others Af Garreh Kofar. The latter two are closely related Somali-like languages but are kept clearly apart by their speakers. There are also Garre who speak Somali proper. Oromo is a different language well beyond comprehension. It belongs to the same lowland branch of the east Cushitic languages as the Somali-type languages, but internal differentiation within this branch is high. The fact that the Garre are also divided between three nation-states (Kenya, Ethiopia, and Somalia) has nothing to do with this linguistic differentiation, since speakers of all four languages are found among the Garre of all three states. The only language which is spoken exclusively by Garre appears to be Af Garreh Kofar, but to the outside observer it is difficult to distinguish that language from Af Rahanweyn which is spoken by hundreds of thousands of non-Garre, namely Somali of the Rahanweyn clan-cluster. It does happen that Garre who do not share one of these Cushitic languages are obliged to converse with each other in languages from totally different language families, like Swahili (Bantu) or English (Germanic) which they have acquired at school, an institution frequented by only a minority of them for mostly short periods.

If there are enough other elements (political, economic, cultural) on which a feeling of ethnic belonging and commonality is based, such an ethnic unit can tolerate a high degree of linguistic diversity. If these other foundations are absent, no amount of linguistic homogeneity can prevent disruption. Even within a language formerly perceived as one, now three different languages can be distinguished if the political will so dictates. This is exemplified by the split of the Serbo-Croatian language into Serbian, Croatian, and Bosnian.

The close association between a nation and its national language is a fairly recent development of political ideology even in Europe, and more so in other parts of the world to which this idea has been exported. Originally the term mother tongue (*Muttersprache*) referred to the language learned from the mother or wet nurse as distinct from the language of the male sphere, of officialdom and scholarship, i.e., Latin (Ahlzweig 1994). Through romanticism and the evolving linguistic nationalism this term later acquired political overtones. The modern national states of Europe have, in their formative period, taken little account of the map of language distribution. Personalities claimed in retrospect as forerunners of German, Russian, or other nationalisms, in their real lives often preferred to speak French. The ethnicization of the state, its culturalization and its evolving association with a particular language was a gradual process in the development of modern Europe. Linguistic obstacles were sometimes artificially raised between nations. As their national standard the Norwegians have chosen from among their many dialects the one most different from Danish.

Not only languages but also language families have been politicized by identifying them with groups of people who were often ascribed racial characteristics, as a result of which they were either accorded privileges, or denied such privileges and even such basic rights as freedom and the right to live. Nowadays nobody outside the lunatic fringes would use the term Semite, to the extent that it is still used at all, in any other sense than speaker of a Semitic language. But there are other such terms of Biblical derivation like Cushites and Hamites which still continue in political careers which were unforeseen by the late eighteenth-, nineteenth-, and twentieth-century European his-

torical linguists who applied them to clusters of languages they found to derive from shared ancestral forms. One can hear Oromo nationalists in Ethiopia arguing for the inclusion of minority groups and their territories into Oromia on the grounds that, although they are not Oromo, they are also Cushites. Ideologies about the repulsion of Hamitic intruders are at the root of the genocide against the Tutsi in Rwanda, although the Tutsi, in fact, speak the same Bantu language as their Hutu murderers, and historical theories that they ever spoke a different language are questionable to say the least. Ironically, too, the term Hamitic is no longer used by linguists. This may serve as a recent example that linguistic classifications, once they cross the boundaries of the academic sphere, can develop a dynamic of their own and that such a dynamic can be quite a murderous one.

Linguistic nationalism often appeals to romantic notions about the relationship between a particular language and other aspects of the culture of its speakers. The language we speak is believed to influence our thinking in a subtle way, giving language a much deeper cultural implication than that of a mere boundary marker between different cultural/ethnic units. Such ideas, which go back to Johann Gottfried Herder (1744–1803), cannot be explained away by pointing to the ways in which they are instrumentalized by nationalists. In Herder's view, a language is the 'collective treasure' of a nation and, since each language shapes the perspective on the world in a different way, the intellectual wealth of mankind consists in the diversity of its languages, i.e., in the number of such treasures (Schlesinger 1991, p. 12f). Or, in the words of a modern Herderian: 'It is ethnic and linguistic diversity that makes life worth living. It is creativity and beauty based upon ethnic and linguistic diversity which makes man human' (Fishman 1989, p. 15). Very differentiated assumptions about the interrelationship between language and thought have also been put forward by Humboldt (1988), while a more simplistic view, basically a linguistic determinism, was proposed much later by Whorf (1962). Whorfians have also claimed the earlier Sapir as one of their persuasion and speak of a Sapir–Whorf hypothesis. A fuller analysis of the interrelationship of language and thought or language and cognition would go beyond the scope of an article on language and ethnicity (for a summary see Schlesinger 1991).

The above-mentioned possibilities, that an ethnic group is completely coterminous and identified with the community of speakers of a certain language, and the opposite case, of a total lack of such a congruence and the absence of any association between an ethic group and the language(s) spoken by it, are two extremes of a range of variation. Between these extremes we find language and ethnicity interacting in a great number of ways. To examine how language relates to ethnicity, and which languages/ethnic identifications persist in which functions in a plurilingual and culturally heterogeneous setting, i.e., to study the ecology of languages, the following variables need to be taken into account: demographic variables, like the sizes of language communities and their growth rates, and their compact or dispersed form of settlement including the forms of mixture with speakers of other languages; political and juridical variables, like the status of a language and the institutions in which it is used, and the presence or absence of ethnic stratification of the labor market with which language status often correlates; ethno-cultural variables, like the association of a language with a group's history or religion, and the potential of a language in terms of range of vocabulary and tradition of writing; interactional variables like communicational mobility and degree of multilingualism of speakers, specialization of the functions of speech varieties in diglossic and polyglossic settings, and routines of interactions between ethnic groups/speech communities; variations on topic, to explore the influence of different topics on the choice of a language; predominant/exclusive use of a language in the public/private sphere, and the linguistic distance between the languages in contact (Haarmann 1986, pp. 11ff). This list is not meant to be exhaustive.

Some of these variables are illustrated. In contact situations, languages may undergo functional differentiation. The original language in bilingual communities may become restricted to sacral functions. Ethnicity and language choice should not be studied in isolation from other dimensions of social identity like gender and age. Minority people may revert in their old age to the language of their childhood, while their professional lives have been dominated by languages of wider currency (Dow 1991). The past interaction between languages, their differences in potential and prestige in different semantic domains, and the relationships between the speakers of those languages and the cultural values attached to their ethnic identities, are reflected by loanwords. Contrary to the widespread assumption that loanwords are typical for specialized concepts, these often extend far into the basic vocabulary and may designate parts of the body or kinship terms (Haarmann 1986, pp. 186ff, Schlee 1994). In intensive language contact situations, borrowing is not restricted to the lexicon and there may be light or heavy structural borrowing (Thomason and Kaufman 1988, pp. 74f).

Written language and writing systems show a stronger tendency to be maintained as sacred, and/or symbols of ethnic identity, than does spoken language. Aramaic-speaking Christians became speakers of Arabic, which they also used for writing and in which they produced a considerable literature, but for a long time they continued to use their earlier Syriac alphabet because they identified the Arabic alphabet with the Qur'aan and Islam. The mirror case is provided by Muslims in the post-*reconquista* Spain who wrote

Spanish in Arabic characters. Jews throughout Europe for centuries wrote the respective majority languages of the countries in which they lived, with a sprinkling of Hebrew words, in Hebrew characters. Where the new alphabet, like the Arabic alphabet in non-Arabic countries converted to Islam, is associated with holy texts, the opposite may occur. Many different languages have been written in the Arabic script, and other writing systems, some of them ancient, have fallen into disuse (Lewis 1998).

Also linguistic variation within the performance of one speaker in the same speech act can highlight aspects of language and ethnicity. A narrator may switch in a story back and forth between languages and/or more or less standardized/creolized varieties of these languages to depict characters to whom different ethnic and other social identities are attributed. Also, in presenting themselves, polyglot people, or people who master more than one register of a language, can play with options of identification rather than being tied to one form of speech and the ethnicity, or more generally the identity, it connotes (LePage and Tabouret-Keller 1985).

See also: Cultural Evolution: Ethnogenesis; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Ethnology; Gender, Class, Race, and Ethnicity, Social Construction of; Language and Social Class

Bibliography

Ahlzweig G 1994 *Muttersprache–Vaterland: die deutsche Nation und ihre Sprache*. Westdeutscher Verlag Opladen, Germany

Barth F (ed.) 1969 *Ethnic Groups and Boundaries*. Allen and Unwin, London

Cooper R L, Spolsky B (eds.) *The Influence of Language on Culture and Thought*. Mouton de Gruyter, Berlin

Dench G 1986 *Minorities in the Open Society: Prisoners of Ambivalence*. Routledge and Kegan Paul, London

Dow J R (ed.) 1991 *Language and Ethnicity*. Benjamins, Amsterdam

Elwert G 1989 Nationalismus and ethnizität. Über die Bildung von Wir-Gruppen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* **3**: 440–64

Fardon R, Furniss G (eds.) 1994 *African Languages, Development, and the State*. Routledge, London

Fishman J A 1989 *Language and Ethnicity in Minority Sociolinguistic Perspective*. Multilingual, Clevedon, PA

Gellner E 1981 *Muslim Society*. Cambridge University Press, Cambridge, UK

Haarmann H 1986 *Language in Ethnicity. A View of Basic Ecological Relations*. Mouton de Gruyter, Berlin

Humboldt W von 1988 *On Language: The Diversity of Human Language Structure and its Influence on the Mental Development of Mankind*. Cambridge University Press, Cambridge, UK

LePage R B, Tabouret-Keller A 1985 *Acts of Identity: Creole Based Approaches to Language and Ethnicity*. Cambridge University Press, Cambridge, UK

Lewis B 1998 *The Multiple Identities of the Middle East*. Weidenfeld and Nicholson, London

Montville J V (ed.) 1989 *Conflict and Peacemaking in Multiethnic Societies*. Lexington Books, Lexington, MA

Ra'anan U 1989 The nation-state fallacy. In: Montville J V (ed.) *Conflict and Peacemaking in Multiethnic Societies*. Lexington Books, Lexington, MA, pp. 5–20

Schlee G 1994 Loanwords in Oromo and Rendille as a mirror of past interethnic relations. In: Fardon R, Furniss G (eds.) *African Languages, Development, and the State*. Routledge, London, pp. 191–212

Schlesinger I M 1991 The wax and wane of Whorfian views. In: Cooper R L, Spolsky B (eds.) *The Influence of Language on Culture and Thought*. Mouton de Gruyter, Berlin

Thomason S G, Kaufman T 1988 *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley, CA

Whorf B L 1962 *Language, Thought, and Reality: Selected Writings*. Wiley, New York

G. Schlee

Language and Gender

In all known societies, gender is a key construct, involving binary distinctions such as man–woman, male–female, masculine–feminine, and in all known societies language plays a crucial role in constructing and maintaining these distinctions. This is achieved in part through the language system itself (note, for example, the inherent bias of the so-called English generics *man* and *he*), but this article will focus on the ways in which language is used by speakers to construct gender and to maintain gender differences.

1. Gender-exclusive vs. Gender-preferential Differences

As long ago as the seventeenth century, differences in the language used by women and men were remarked upon in anthropological writings. Missionaries and explorers came across societies in which they claimed to find distinct languages for male and female speakers. In fact, their accounts exaggerated the reality; what can be found in some languages are phonological, morphological, syntactic, or lexical contrasts where the speaker's gender determines which form is chosen. An example of this is Yanyuwa, an Australian aboriginal language; in this language, the choice of particular case-marking suffixes depends on the gender of the speaker.

Gender-exclusive differences of this kind mean that speakers who use a form inappropriate to their gender will be strongly reprimanded. Bradley (1998) quotes a

Yanyuwa male who reminisces 'I spoke like a woman and [my mother] yelled at me, "Hey! you are a man, you have no foreskin, why do you talk like a woman?" I was shamed.' Such linguistic differences clearly function to keep gender roles distinct. However, most of the linguistic variation associated with gender found today involves gender-preferential rather than gender-exclusive differences. This means that, rather than there being linguistic forms associated exclusively with one gender, there is instead a tendency for women or for men to use a certain form more frequently. The difference between gender-exclusive and gender-preferential usage seems to correlate with differences between nonindustrialized societies, such as the Yanyuwa, and industrialized societies, such as Britain and the USA. Nonindustrialized societies tend to have clearly demarcated gender roles, whereas in modern industrialized societies, gender roles are much less rigidly structured.

2. *Quantitative Sociolinguistic Research*

Researchers investigating gender-preferential differences in grammar and pronunciation use a quantitative approach; results showing phonological or grammatical differences are presented in terms of either raw numbers or percentages, and statistical tests are used to demonstrate the significance of these findings. This approach is common throughout the sciences and social sciences, and it has proved a valuable tool for sociolinguists exploring gender-preferential differences (as it can demonstrate gender difference despite the fact that women and men are using the same linguistic forms). This approach was developed by Labov in his pioneering book, *Sociolinguistic Patterns* (1972), and it has provided sociolinguists with a model for work on linguistic variation; the model is often referred to as the Variationist Paradigm.

In terms of gender, the most robust finding to emerge from research in a wide range of speech communities is that female speakers tend to use linguistic forms closer to the prestige norm more frequently than male speakers of a similar social background (see, for example, Trudgill 1998, Eisikovits 1998, Milroy 1980). It has even been claimed that this should be viewed as a sociolinguistic universal (Holmes 1998). Many explanations have been put forward for this finding, none of them wholly satisfactory; four of these will be examined briefly here.

2.1 *Explanations for Gender-differentiated Language Use*

First, the notion of covert prestige was developed to describe the positive value assigned to vernacular (nonstandard) forms by working class and male speakers. It is argued that male speakers tend to orient to the covert prestige of vernacular forms while female speakers tend to orient to the overt prestige of standard forms.

A second explanation derives from social network theory which demonstrates how individual speakers are members of social networks which have the power to enforce group norms (Milroy 1980). In the case of language, social network theory predicts that group pressures will enforce linguistic norms. Since women and men tend to belong to different (gender-based) social networks, then different norms are maintained.

A third explanation emphasizes the role of economic factors, that is, the pressures exerted by the job market which bring some speakers, but not others, into contact with more standard speech forms (see, for example, Nichols 1998).

Finally, a more recent explanatory framework is proposed by Eckert (1998, 1999) in her research into the speech of adolescent speakers in Detroit, USA. Eckert found that girls' linguistic usage was more extreme than that of boys: girls with more middle-class values used forms closest to the prestige standard, while girls with more working-class values used the most advanced vernacular forms. Drawing on Bourdieu's notion of a symbolic market, Eckert argues that female speakers have to use more standard (or more vernacular) speech because they are marginalized in the linguistic marketplace. (For further discussion of language variation and change, see *Sociolinguistics*.)

3. *Gendered Communicative Competence*

Early research on language and gender concentrated on what were seen as core features of language: phonetics, phonology, syntax, and morphology. Later, researchers began to turn their attention to broader aspects of talk, the conversational strategies characteristic of male and female speakers which constitute their communicative competence (see *Communicative Competence: Linguistic Aspects*). Communicative competence is a phrase coined by Dell Hymes. He argued that it was essential to incorporate social and cultural factors into linguistic description, since when children learn to talk they learn not just grammar but also a sense of appropriateness. It is not sufficient for a child to be linguistically competent; he/she must also known when to speak, when to remain silent, how to tell a joke, how to be polite, etc.

As children become more competent as speakers, so they learn to talk in ways appropriate to their gender. Becoming a girl or a boy means talking like a girl or a boy. Women and men characteristically draw on different strategies in conversational interaction, strategies such as paying compliments, hedging, apologizing, swearing. Women's and men's behavior in conversation suggests that they have a different understanding of how a compliment or an apology is done. (Examples of such differences can be found in

Holmes' paper on compliments in New Zealand English, Brown's paper on politeness in a Mayan community in Mexico, and Goodwin's paper on directives in Black peer groups in Philadelphia, USA, all in Sect. 2 of Coates 1998.) Such differences have led some researchers to talk of different female and male 'styles' in conversation (see Maltz and Borker 1998).

At a broader level, it even seems that male and female speakers prefer different modes of conversational organization (see Coates 1996, 1997a). Male speakers tend to choose a one-at-a-time model of turn-taking, where speaker turns are carefully demarcated, and where speakers occupy the conversational floor one at a time, often for long solo turns. By contrast, female speakers in relaxed informal conversation with friends characteristically opt for a more collaborative model of turn-taking, which allows speakers to share the conversational floor and construct talk jointly. This latter mode of conversational interaction can be called the 'jam session' model.

3.1 Conversational Dominance

One cluster of conversational strategies has been observed to function in conversation as a way of enabling certain speakers to dominate their partners in talk. Asymmetrical patterns of talk are always a sign of power imbalance, and asymmetry has been found in research which focuses on mixed talk in a variety of social contexts, with men's greater usage of certain strategies being associated with male dominance in conversation (see the papers in Coates 1998, Sect. 3).

The following linguistic strategies can be involved in conversational dominance: interrupting another speaker; failing to respond to another speaker's utterance, or delaying a response; talking too much (that is, occupying the conversational floor for long stretches of time and thus preventing other speakers from participating); or talking too little (silence can be a very effective device to control the direction in which conversation goes, or whether conversation can take place at all).

The research evidence is that male speakers regularly use these strategies to dominate talk. Research on language in the classroom reveals that boys get to talk on average twice as much as girls, and this pattern persists in public arenas with adult speakers. In the private sphere, male speakers in a range of studies have been found to dominate their female friends or partners through using these strategies.

4. Difference or Dominance?

Two approaches to gender variation in language have dominated sociolinguistic research in the area of language and gender until very recently; these are the difference approach and the dominance approach. To put it very simply, research which takes a dominance perspective interprets the differences between women's and men's linguistic usage as reflexes of the dominant–subordinate relationship holding between men and women. Research which takes a difference perspective, by contrast, sees the differences between women's linguistic usage and men's linguistic usage as arising from the different subcultures in which women and men are socialized (this approach is sometimes called the subcultural or two-cultures approach).

Early work on language and gender took a dominance approach. The reaction against this, and the subsequent flourishing of the difference approach, arose not because researchers denied the existence of dominance and oppression in male–female relationships, but because researchers, particularly feminist researchers, became unhappy at the negative portrayal of women in work using the dominance approach. Women's language was described as weak, unassertive, tentative, and women were presented as losers, as victims.

Researchers who began using the difference model in the 1980s argued that the dominance model had become a deficit model, that is, a way of interpreting the linguistic facts which represented men's language as the norm and women's language as deviant. The advantage of the difference model is that it allows researchers to show the strengths of linguistic strategies characteristic of women and to celebrate women's ways of talking. For those carrying out research involving mixed talk, the dominance approach provides a useful explanatory framework, but for researchers investigating same-sex talk it seems less appropriate, since dominance and oppression are not obviously helpful categories for the description of all-female talk (or of all-male talk, for that matter, but it is only very recently that sociolinguists have turned their attention to the informal talk of men—see Johnson and Meinhof 1997).

Probably the main problem with the difference approach is its assumption that all interactional difficulties between women and men can be called 'miscommunication.' However, the idea that miscommunication is at the heart of male–female problems has immense popular appeal, and has become widely known throughout the English-speaking world since publication of the book *You Just Don't Understand* (Tannen 1990) and other similar works.

The popularization of the discussion about male–female differences and miscommunication means that the difference approach is now seen as problematic, because it is associated with a political stance which ignores male dominance. However, interesting work on same-sex talk continues to be carried out which implicitly draws on a difference or subcultural approach. But in many areas researchers have moved on, assimilating ideas from European social theory. In particular, current research into linguistic variation takes the view that gender is not a given but is

accomplished through talk, and that speakers have available to them a whole range of (often conflicting) discourses (see Coates 1997b, Fairclough 1992, Weedon 1987).

5. *Language and Gender Research: New Directions*

In the early years (the 1960s and 1970s), research into the interaction of language and gender relied on a predominantly essentialist paradigm which categorized speakers primarily according to biological sex, and used mainly quantitative methods. Next, in the 1970s and 1980s, came a period which recognized the cultural construction of categories such as gender. During this period, more qualitative, ethnographic approaches predominated. In recent research, a more dynamic social constructionist approach has emerged which makes possible the combination of quantitative and qualitative research.

Researchers are now urged to 'think practically and look locally' (Eckert and McConnell-Ginet 1998). In particular, researchers are urged to work in terms of communities of practice, a concept which they adopt to assert the dynamism of living communities, and to underscore the fact that personal identity is not constructed in a vacuum, but in the many communities of practice individuals belong to. Gender, one aspect of personal identity, is not a given, but is constantly created and recreated in social interaction with others in these communities of practice.

Moreover, gender is complex: the binary distinctions men–women, male–female, masculine–feminine tend to distort—and oversimplify—our thinking. It is unhelpful to think of men and women as being totally different: male and female speakers differ in many ways, but there are also many areas of overlap. It is also important to realize that using the terms 'masculinity' and 'femininity' in the singular is a stylistic convenience: there is no singular masculinity or femininity. Gender is fluid and variable and at any point in time there will be a range of masculinities and femininities extant in a culture, and these will vary in terms of class, sexual orientation, ethnicity, age, etc., as well as intersecting in complex ways. Finally, masculinity and femininity cannot be understood on their own: the concepts are essentially relational. In other words, femininity (for example) is only meaningful when it is understood in relation to masculinity and to the totality of gender relations (see Connell 1995).

At the end of the twentieth century, sociolinguistic research has moved on from the simple correlation of linguistic form with social category. Researchers now analyze spoken and written data with the aim of understanding how gender is constructed in everyday life and of assessing the role of language in the creation and maintenance of contemporary masculinity and femininity. There is increasing emphasis on the need to be aware of similarities as well as differences between male and female speakers, as well as to assert the plurality of masculinities and femininities. Finally, the spread of language and gender research to non-English speaking communities is widening understanding of the key issues, since it seems that whatever the language of a community, gender politics manifest themselves in similar ways all over the world.

See also: Education (Higher) and Gender; Education (Primary and Secondary Schools) and Gender; Gender and Feminist Studies in Anthropology; Gender and Language: Cultural Concerns; Gender, Class, Race, and Ethnicity, Social Construction of; Gender Differences in Personality and Social Behavior; Gender-related Development; Language and Social Class; Politeness and Language; Small-group Interaction and Gender

Bibliography

Bradley J 1998 Yanyuwa: Men speak one way, women another. In: Coates J (ed.) *Language and Gender: A Reader*. Blackwell, Oxford, UK, pp. 13–20

Coates J 1996 *Women Talk. Conversation Between Women Friends*. Blackwell, Oxford, UK

Coates J 1997a One-at-a-time: the organisation of men's talk. In: Johnson S, Meinhof U (eds.) *Language and Masculinity*. Blackwell, Oxford, UK, pp. 107–29

Coates J 1997b Competing discourses of femininity. In: Kotthoff H, Wodak R (eds.) *Communicating Gender*. Benjamins, Amsterdam, pp. 285–314

Coates J (ed.) 1998 *Language and Gender: A Reader*. Blackwell, Oxford, UK

Connell R W 1995 *Masculinities*. Polity Press, Cambridge, UK

Eckert P 1998 Gender and sociolinguistic variation. In: Coates J (ed.) *Language and Gender: A Reader*. Blackwell, Oxford, UK, pp. 64–76

Eckert P 1999 *Variation and Social Practice: The Linguistic Construction of Social Meaning in Belten High*. Blackwell, Oxford, UK

Eckert P, McConnell-Ginet S 1998 Communities of practice: Where gender, language and power all live. In: Coates J (ed.) *Language and Gender: A Reader*. Blackwell, Oxford, UK, pp. 484–94

Eisikovits E 1998 Girl-talk/boy-talk: Sex differences in adolescent speech. In: Coates J (ed.) *Language and Gender: A Reader*. Blackwell, Oxford, UK, pp. 42–54

Fairclough N 1992 *Discourse and Social Change*. Polity Press, Cambridge, UK

Holmes J 1998 Women's talk: The question of sociolinguistic universals. In: Coates J (ed.) *Language and Gender: A Reader*. Blackwell, Oxford, UK, pp. 461–83

Johnson S, Meinhof U (eds.) 1997 *Language and Masculinity*. Blackwell, Oxford, UK

Labov W 1972 *Sociolinguistic Patterns*. Blackwell, Oxford, UK

Maltz D, Borker R 1998 A cultural approach to male–female miscommunication. In: Coates J (ed.) *Language and Gender: A Reader*. Blackwell, Oxford, UK, pp. 417–34

Milroy L 1980 *Language and Social Networks*. Blackwell, Oxford, UK
Nichols P 1998 Black women in the rural south: Conservative and innovative. In: Coates J (ed.) *Language and Gender: A Reader*. Blackwell, Oxford, UK, pp. 55–63
Tannen D 1990 *You Just Don't Understand: Women and Men in Conversation*. William Morrow, New York
Trudgill P 1998 Sex and covert prestige. In: Coates J (ed.) *Language and Gender: A Reader*. Blackwell, Oxford, UK, pp. 21–8
Weedon C 1987 *Feminist Practice and Poststructuralist Theory*. Blackwell, Oxford, UK
West C, Zimmerman D 1998 Women's place in everyday talk. In: Coates J (ed.) *Language and Gender: A Reader*. Blackwell, Oxford, UK, pp. 165–75

J. Coates

Language and Literature

Literary linguistics is a type of linguistics which studies language in literature, and asks whether the form and use of language in literature are different from the form and use of language in general. This can lead to a further question about function: whether special kinds of form and use of language serve the particular functions of literary texts—to be culturally valued artefacts, to stimulate emotion, and to communicate important meanings. Alongside these questions there is another question about literary form (such as the division of a poem into lines), and its relation to linguistic form (such as the division of a text into sentences): whether these kinds of division into constituent parts are fundamentally similar or different in kind.

1. *Literary Linguistics*

Linguistics looks for generalizations and so treats any text, any instance of language, just as an example of general processes at work and hence of interest only to the extent that it demonstrates those general processes. In contrast, literary criticism, at least traditionally, is interested in the text in itself, its characteristics considered unique and special, carrying traces of the historical moment of the text's production and of the individual genius of the author who produced it. This is a fundamental distinction between linguistics and literary studies, abstraction as opposed to concreteness, generalization as opposed to specificity, and means that any combination of linguistics and literary studies must compromise in one direction or the other.

Linguistics might put itself at the service of literary studies by offering to analyze the underlying linguistic system which is instantiated in a specific text; this is an approach often associated with more functionalist orientations in linguistics. Functionalist literary linguistics characteristically uses linguistics to illuminate the workings of language in a particular text, and in particular to establish an interpretation of the text. This 'critical linguistics' spreads beyond literary studies, into the analysis of media texts (in 'critical discourse analysis'). It has also been influential in language teaching, with linguist-educationalists arguing that the student better understands a language by analyzing its use in literature (see Widdowson 1975).

Alternatively, literature might serve linguistics by offering the literary text as evidence for generalizations about a particular kind of language: the language of literature. This is particularly associated with formalist literary linguistics, such as Generative Metrics, closely allied to Generative Phonology. Formalist literary linguistics looks for generalizations specific to language in literature, and ways in which literary practice develops from the abstract principles underlying all language. Formalist literary linguistics has its roots in Russian Formalism and related movements, and in its early stages was particularly associated with the work of Jakobson (1960). Through Jakobson, it influenced the structuralist anthropology of Claude Lévi-Strauss, and then the structuralist literary analysis of Roland Barthes, in which literary studies briefly adopted the generalizing and universalizing tendencies of linguistics. The history of formalist literary linguistics can be traced in a series of anthologies and conference proceedings: Sebeok (1960), Freeman (1970, 1981), Fabb et al. (1987), Kiparsky and Youmans (1989); see also Fabb (1997).

2. *What Makes 'Literature' Different?*

The language of literary texts may sometimes differ from the language of nonliterary texts, but this is not a defining characteristic of literature. What makes a text literary is not form, but function. Jakobson (1960) proposed that verbal behavior might perform any (and any combination) of six distinct functions, one of which was the poetic function. In literature, the poetic function dominates. For Jakobson, a text functions poetically when it draws attention to its own form—when our attention is drawn to the text itself and not just to what it means. Bauman (1975) suggests that the poetic function itself has a further purpose, which is to enable a literary text to be performed; for Bauman it is performance, understood as a type of social behavior requiring evaluation by an audience, which defines a text as literary. The poetic function draws attention to the form of the text in order to permit evaluation by the audience according to the text's adherence to rules of form. For both Jakobson and Bauman a literary text can differ from a nonliterary text only in terms of function; in principle the same text could be made

literary or nonliterary depending on the extent to which it is made to draw attention to its own form, by being presented or performed in different ways.

Though it is function rather than form which makes a text literary, the poetic function can nevertheless give rise to distinctive kinds of literary form. A literary text tends to differ from a nonliterary text in two fundamental ways relating to form: in having specialized kinds of *constituent structure*, and in having 'difficult form,' form which is not straightforward to ascertain. Specialized kinds of constituent structure include lines of verse, metrical constituents such as the foot, and narrative constituents such as 'orientation,' 'coda,' or narrative episodes. These specialized kinds of form are regulated by rules which the text can overtly adhere to, thus drawing attention to form. Difficulties for interpretation arise from formal ambiguities, as well as from exploitations of the ways in which verbal communication works to produce effects such as metaphor and irony. By making the text difficult, attention is drawn to the workings of the text and thus again the poetic function is performed. These formal complexities and difficulties may also be responsible for some of the aesthetic effects of literary texts, if the experience of complexity constitutes one of the kinds of aesthetic experience.

'Literature' can exist either in the medium of writing or speech. (Because the term implies writing, 'literature' is sometimes substituted in literary linguistics by the medium-neutral term 'verbal art.') Language in written literature and language in oral literature differ, just as writing and speech differ. Any medium both constrains and enables certain kinds of aesthetic possibility in its own medium-specific way. Thus we might ask whether spoken literature makes special use of the characteristics of speech in order to carry out literature-specific functions; the same question arises for written literature. Some kinds of literary–linguistic form are restricted to speech. Metrical verse is the most obvious example, because it organizes the components of sound: loudness, pitch, and duration. There have been proposals that certain kinds of linguistic form are restricted to writing; thus Banfield (1982) proposes that certain kinds of sentence, expressing a character's inner experience but at the same time describing that experience from another's point of view and so paradoxically combining two centers of perspective, are 'unspeakable,' found only in writing (Fludernik 1993 and other authors have suggested that complexities of this kind can also be understood as exploitation of the 'looseness' of communication, see Sect. 7).

3. *Metrical and Parametrical Form*

Verse is a kind of text defined by its having a particular constituent structure: it is divided into lines. The line is marked out as a constituent by different means in different types of verse: in free verse it is marked out as a constituent by sentence structure, layout on the page, and performance; in canonically parallelistic verse the line is characterised by parallelism with other lines (see Sect. 4); in metrical verse the line is regulated by metrical rules which fix the length of the line, measured in phonological units such as the syllable or the mora (a 'weight unit' contained with the syllable). The metrical rules can also determine a rhythm within the line, in terms of the sequencing of relatively stressed and unstressed syllables, or relatively heavy and light syllables.

One of the key questions about language in literature is whether literary language ever has its own distinctive kind of linguistic form, not just a way of using general linguistic form, but a separate rule-system which governs literary language. Metrical verse perhaps offers the best opportunity of finding specifically literary-linguistic form of this kind, because in metrical verse the rules which govern the meter appear to be similar to linguistic rules, while also differing from linguistic rules. The similarities involve phonological rules, the rules which govern how sounds are combined into words. First, metrical rules build 'tree structures' (hierarchical constituent structures) which resemble the tree structures of phonology. Second, the metrical rules form a set of interacting rules so that apparent exceptions can be understood in terms of rule interaction and even appear to operate on abstract structures, like phonological rules. Third, the metrical rules differentiate strong and weak phonological constituents, just like the phonological rules which organize the rhythmic structure of speech in general. Formalist literary linguistics has always seen metrical verse as the most important kind of text for understanding the relation between linguistic rules in general and specifically literary linguistic rules, and there has often been a close relation between work in phonology (the sound-structure of the language) and metrics; as new theories of phonology have emerged, so new insights become available into metrical structure. (Representative samples of linguistic work on meter include Halle and Keyser 1971, Attridge 1982, Hayes 1989, Hanson and Kiparsky 1996. See *Language and Poetic Structure*.)

Metrical rules control small phonological units. These include the morae which are counted in Japanese verse; the syllables counted in Welsh verse, organized into stress patterns in English verse, and weight patterns in Latin verse; and the phonological feet (combinations of syllables grouped around a strong syllable) counted in older Germanic verse like *Beowulf*. But it is also fairly common to find another kind of rule operating simultaneously with the metrical rules, and governing larger phonological units such as the word. These word-boundary rules—caesura and bridge rules—govern where words can begin or end within the line, and may contradict the constituent structures built by the meter. Thus Homer's line is

divided into six constituents ('feet'), and thus can be divided into two halves with three feet in each half. But word-boundary rules operate to ensure that the middle of the line does not coincide with the end of a word; instead a word must end just before or just after the middle boundary as defined by the foot constituent structure. The fact that word-boundary rules sometimes contradict the metrical rules with which they co-occur suggests that contradiction and complexity are design features of literary linguistic rules, perhaps because contradiction and complexity are a source of aesthetic experience. Poetic word-boundary rules are of interest to linguists partly because of their interaction with metrical rules, but also because they are sensitive to different levels of linguistic structure; Devine and Stephens (1984) have shown for Greek that in tragedies, word-boundary rules govern the lexical word, while in comedies the rules govern the 'phonological word' (which contains a lexical word plus neighboring grammatical words). This shows that literature-specific rules are able to differentiate abstract linguistic structure.

4. *Parallelism*

Most verse traditions are governed by one of two basic organizatory systems: meter or canonic parallelism (free verse, a third organizatory system, has become increasingly common in the twentieth century). *Canonic parallelism* (the term is Jakobson's) is parallelism as a structuring principle in verse holding throughout a text. Parallelism draws on the linguistic structure of the text: parallel pieces of text (usually lines, or half-lines) differ in certain respects but have some fairly salient linguistic component in common. In phonological parallelism, a sequence of sounds is repeated: the Welsh system of *cynghanedd* (harmony) requires the consonant sequence in the first half of a line to be repeated in the second half of the line; in some Gaelic songs the sequence of stressed vowels is the same in every line; and in Efik tonal riddles the sequence of tone changes is the same in both halves of the riddle. In syntactic parallelism, phrase structure is repeated; the words are different but they belong to the same parts of speech organized in the same manner. Syntactic parallelism is often tied to the third type of parallelism: lexical parallelism, in which words with related meanings are paired in adjacent sections of the text (see Fox 1988).

Parallelism is another instance of a literary practice where linguistic form is controlled in a literary text. Canonic parallelism has some similarities to meter, and functions in a similar way; both, for example, draw attention to form (hence define the text as verbal art) through ostensive repetition of form. However, parallelism and meter clearly operate by different principles; in parallelism the only general rule is 'imitate the preceding line/half-line,' while in meter the line has its structure because it is governed by a complex set of metrical rules. Thus, parallelism does not appear to involve a distinctive kind of linguistic form, but instead is a way of using linguistic form to build literary form. It is worth noting that parallelism has emerged apparently independently in widely dispersed literatures, for example, Mayan, Russian, Tswana, Egyptian, and the isolated cultures of New Guinea. In contrast, almost all known metrical traditions can be historically connected, and traced back to Indo-European or Arabic origins (and these may themselves be connected). It is relatively uncommon for metrical traditions to emerge apparently spontaneously elsewhere (rare exceptions include Somali and the East Australian Dyirbal). This correlates with a hypothesized difference between parallelism as a development of possibilities inherent in language, and meter as a rule system separate from language in general and not implied by language in general.

5. *Narrative Structure*

Narrative is a type of discourse, including as a subtype 'literary narrative.' Questions about language in literary narratives are thus to some extent the same as questions about language in narrative as a general discourse genre. One of the most influential linguistic works on narrative told to interviewers, Labov and Waletzky (1967), describes true stories in which the narrator was in danger. Though these are not literary narratives, they nevertheless have characteristics which literary narratives share. Labov and Waletzky drew particular attention to three characteristics. First, sentences in the narrative can be separated into ('storyline') sentences whose sequence mimics the sequence of events and ('non-storyline') sentences whose role is to provide contextual information. Subsequent research has shown that storyline and nonstoryline sentences typically differ in their linguistic form (storyline clauses tend to be higher in transitivity; may have the verb in a more prominent position, etc.). Second, the narrative as a discourse has a clearly structured trajectory; it often begins with an 'orientation' constituent, and ends with a 'coda' constituent, and there is a movement from a complication to a resolution. Third, the point at which the complication is resolved is often made salient by particular linguistic devices; this practice is called the 'evaluation' by Labov and Waletzky, and may involve an explicit judgment on the narrative by the narrator; more generally it draws the hearer's attention to the point of the narrative. This strand of work on narrative offers a way of thinking about literary narratives as developments of narrative as a general discourse genre, but does not involve a specifically literary kind of form. Nevertheless, it is possible to see how literary narratives exploit these basic possibilities to generate

effects such as suspense, to create complex and ambiguous structures, and to draw attention to narrative form (thus performing the poetic function).

Focus on the language of specifically literary narrative comes from another tradition, particularly associated with Dell Hymes's work on American Indian narratives (Hymes 1981). In this tradition, the narrative is seen to have a constituent structure analogous to verse constituent structure (typically nonmetrical, but sometimes with canonic parallelism), in which the text's constituent structure is signaled by the salient use of grammatical material—connective words ('and,' 'so,' 'then'), hearsay elements ('it is said'), shifts in tense, and so on. As Tedlock (1983) argues, components of speech, such as intonational contours or the placement of pauses, can also have an important role in marking out constituency. This research tradition also looks at the organization of content into episodes within the narrative, and ways in which episode boundaries are signalled by grammatical means. This research, sometimes involving the reanalysis and reconstruction of texts gathered in the early twentieth century from now lost languages, has as one of its goals to assert the literariness of the texts, by revealing their complexity. This clarifies one of the ideological goals which is characteristic of literary linguistics in its more ethnographic and anthropological modes: to promote the artistry of traditional verbal arts, demonstrating that traditional literatures from marginalized cultures can be as complex as the canonical literatures of the 'first world.'

6. *Access to Abstract Linguistic Form*

With Transformational Generative Grammar, Noam Chomsky invented a way of thinking about a sentence or a word as the 'surface' manifestation of a set of interconnected representations (see *Generative Grammar*). Each instance of spoken or written language is in this sense like the tip of an iceberg, with multiple representations underlying each sentence and each word. Users of the language have no direct access to those underlying representations, in which a sentence might have its phrases ordered differently from its surface representation, or a word might have sounds which are different from those which appear on the surface; this is an 'encapsulated' type of knowledge which is hidden away from introspection.

In this tradition, one of the things which has interested linguists about literary texts is the possibility that underlying sentence or word structures might be manipulated directly in the construction of some literary texts. This would be an extraordinary access to linguistic material which is normally hidden. Along these lines, Austin (1984) argues that instances of distorted phrasing in poetry by Shelley might involve the violation of specific constraints on transformational rules. Thus the line 'And the swift boat the little waves which bore' sounds odd because there is leftward movement of 'the swift boat' which is the object of the verb 'bore,' out of a relative clause in a noun phrase; this is a violation of the 'complex noun phrase constraint,' which was proposed in 1960s transformational grammar to regulate abstract syntactic structure. Austin shows that Shelley violates constraints as though he has access to the underlying workings of the language, which ought to be hidden to introspection or intervention. Similar kinds of argument have been made about sound structure. Thus Malone (1988) argues that rhymes in Turkish poetry seem to be somewhat dissonant on the surface but can be seen to be underlying perfect rhymes, if the linguistic process of 'vowel harmony' is factored out. Vowel harmony alters the abstract vowels which make up a word to ensure that surface vowels in the same word 'harmonize' to have similar features. Thus in these poems the rhyme shows more regularity at an abstract level than on the surface of the text.

The possibility that abstract and supposedly 'inaccessible' form is manipulated in literary texts might interestingly suggest that authors (and audiences) can have privileged though presumably unconscious access to kinds of linguistic form which are normally not available to introspection. This suggestion that literature is special in its relation to language would find favor with some traditional (specifically Romantic) views of literature. But in some cases, it might also be suggested that authors use a slightly different 'language' from the ordinary language, which operates according to slightly different rules. In some cases, these different literary languages might correspond to archaic versions of the contemporary language.

7. *The Communication of Literary Meaning*

The communication of meaning is a primary function of most ways of using language. Communication has two components. The encoding of meaning into form (performed by the vocabulary and syntax of the language, such that a sentence encodes a proposition) is one component which provides evidence for the informative intention of the communicator. But other kinds of evidence can be added, and may override the evidence offered by the encoded meaning; these kinds of evidence can come from the way in which the form is used, and rely on the context in which the communicative act occurs (including the cognitive context of the interlocutors—what they know). This second component involves inferencing, and falls under the scope of pragmatics. It is here that most of the interesting questions about literary meaning have been asked.

Metaphor is a characteristic 'meaning effect' in literary texts. In a metaphor, the encoded meaning (what is actually said) is not the communicated meaning (what is intended and can be inferred). Two lines of enquiry into metaphor have proved particularly fruitful for linguistics; both take the position

that metaphor is not specific to literature but is characteristic of the way language works in general. In the approach particularly associated with the work of George Lakoff and co-workers (Lakoff and Johnson 1980), the vocabulary of a language is claimed to be structured by metaphor, collected into sets of relations between words and meanings which develop basic metaphors, the 'metaphors we live by.' Thus, for example, the use of the term 'journey' to stand for 'life' is a metaphorical substitution which underpins many other substitution (we get a start in life, reach the end of our life, choose a path, etc). Metaphors in literature draw upon the 'metaphorical' structure of the vocabulary in general, using this as raw material and potentially innovating on it, as when Robert Frost writes a poem on *The Road Not Taken*; thus language in literature makes special use of the potential of language in general. A second approach is taken by Sperber and Wilson (1995), who see the possibility of metaphor as a development of the way in which communication works. In verbal communication, a literal meaning is encoded into the linguistic form of a sentence, but that literal meaning need not itself be the intended meaning; instead the literal meaning can constitute indirect evidence for an intended meaning which is not actually stated. This is true of all verbal communication; it is also the basis of metaphor, where the literal meaning is not the intended meaning, but instead offers evidence by which the intended meaning can be inferred.

A second characteristic 'meaning effect' in literary texts involves the attribution of a source for any thought which is communicated, and the attribution of an attitude or a set of attitudes towards that thought. Again Sperber and Wilson (1995) have proposed an influential way of seeing the literary uses of attribution and attitude—in effects such as 'free indirect speech' or irony—as developments of the possibilities of communication in general. The simplest case in verbal communication is for the thoughts expressed by the communicative act to be attributed to the communicator, along with the assumption that the communicator believes those thoughts. These attributions are part of the communicative process; they are worked out by the hearer or reader alongside the working out of the referential content of the communicative act. The attribution of source and attitude is affected by the same 'looseness' which gives rise to metaphor. In literature, the possibility of looseness in the attribution of speaker is exploited in techniques such as the use of hearsay particles (which displace the source of a thought onto persons unknown, with various literary consequences), or free indirect speech (which blur the source of a thought between narrator and character). Literature also exploits the possibility of 'looseness of attitude' in modes such as irony; in irony, the communicator does not have the 'unmarked' attitude of belief towards the thoughts which she expresses. Instead, she typically communicates her own disbelief and at the same time re-assigns the source of the thoughts to another person or set of persons who are assumed to believe those thoughts.

8. *The Communication of Literary Form*

A fundamental element of linguistic form is linguistic constituent structure (see *Constituent Structure: Syntactic*; *Phonology*). The interpretation of speech involves the assignment of a phonological constituent structure to the sequence of sounds, and the interpretation of a sentence involves the assignment of a syntactic constituent structure to the sequence of words. Literary texts have their own distinctive kinds of constituent structure, examples of literary form (rather than linguistic form). Examples include lines and stanzas in verse, or the orientation and coda and division into episodes in narrative. Each of these kinds of constituent structure is specific to literary texts, each adding noticeably to the formal complexity of the text and thus drawing attention to form and performing the poetic function. An interesting question about these kinds of constituent is whether they are similar in kind to types of syntactic or phonological constituent.

As a way of thinking about this question, it is helpful to consider two distinct ways in which a text can have a specific constituent structure. A text has phonological and syntactic constituent structure because it is processed by phonological and syntactic rules; these structures enable words to exist, to be spoken and to relate to one another in an interpretable manner. These kinds of constituent structure are in principle determinate; they can be discovered by linguistic investigation though they may not be obvious on the surface. In contrast, constituent structure might also exist as a description of the text only to the extent that it is interpreted as being present in the text, by analogy with the interpretation of implied meanings. The constituent structure in this sense is a self-description which is evidenced by the text, basically a kind of meaning, rather than being like linguistic form. Evidenced constituent structure of this kind might sometimes be vague or indeterminate, ambiguous or contradictory, and there is some reason to think that many kinds of literary constituent structure are evidenced in this manner. This includes the division of verse into lines, or the division of a narrative into functional constituents like 'evaluation' or structural constituents like 'episode.' In all cases, the text contains evidence that these constituents exist—usually some element at the beginning or end of the constituent (such as a connective like 'so' at the beginning of a narrative, or a rhyme at the end of the line). Of the various kinds of literature-specific constituent structure, only metrical constituent structure seems to be 'inherent' and determinately present in the manner of phonological structure. If these speculations are right, then most kinds of literary constituent are best

understood under a theory of pragmatics, which looks at interpretation—in this case, the interpretation of form.

Sperber and Wilson (1995) suggest that the experience of weak implicature—a large number of weakly evidenced interpretations stimulated by a text—is a 'poetic effect'; perhaps this is the cognitive component of a certain kind of aesthetic experience. This might be linked with a position held by some within literary studies and particularly postmodern literary studies. This is that textual complexity is a source of aesthetic experience, with complexity manifested in ambiguity, uncertainty, and vagueness, contradictoriness, looseness, and so on. 'Difficulties' in meaning and in form, both of which are literary exploitations of the pragmatic aspects of language, might thus be seen as ways of generating aesthetic experience. When we consider also that even the most regulated kind of literary linguistic form—metrical form—seems to be designed to give rise to complexity and contradiction, we might conclude that whereas linguistics seeks the underlying order in language, a literary linguistics might look at how the ordering principles of language give rise to the kinds of productive disorder which we experience as aesthetic.

See also: Emotion and Expression; Language and Poetic Structure; Literary Texts: Comprehension and Memory; Models, Metaphors, Narrative, and Rhetoric: Philosophical Aspects; Oral and Literate Culture

Bibliography

Attridge D 1982 *The Rhythms of English Poetry*. Longman, Harlow, UK

Austin T R 1984 *Language Crafted. A Linguistic Theory of Poetic Syntax*. Indiana University Press, Bloomington, IN

Banfield A 1982 *Unspeakable Sentences. Narration and Representation in the Language of Fiction*. Routledge & Kegan Paul, Boston

Bauman R 1975 Verbal art as performance. *American Anthropologist* **77**: 290–311

Devine A M, Stephens L D 1984 *Language and Metre. Resolution, Porson's Bridge and their Prosodic Basis*. Scholar's Press, Chico, CA

Fabb N 1997 *Linguistics and Literature: Language in the Verbal Arts of the World*. Blackwell, Oxford, UK

Fabb N, Attridge D, Durant A, MacCabe C (eds.) 1987 *The Linguistics of Writing: Arguments Between Language and Literature*. Manchester University Press, Manchester, UK

Fludernik M 1993 *The Fictions of Language and the Languages of Fiction*. Routledge, London

Fox J J (ed.) 1988 *To Speak in Pairs. Essays on the Ritual Languages of Eastern Indonesia*. Cambridge University Press, Cambridge, UK

Freeman D C (ed.) 1970 *Linguistics and Literary Style*. Holt, Rinehart and Winston, New York

Freeman D C (ed.) 1981 *Essays in Modern Stylistics*. Methuen, London

Halle M, Keyser S J 1971 *English Stress: Its Form, its Growth, and its Role in Verse*. Harper and Row, New York

Hanson K, Kiparsky P 1996 A parametric theory of poetic meter. *Language* **72**: 287–335

Hayes B 1989 The prosodic hierarchy in meter. In: Kiparsky P, Youmans G (eds.) *Phonetics and Phonology 1. Rhythm and Meter*. Academic Press, San Diego, CA

Hymes D 1981 *'In Vain I Tried to Tell You.' Essays in Native American Ethnopoetics*. University of Pennsylvania Press, Philadelphia, PA

Jakobson R 1960 Linguistics and poetics. In: Sebeok T (ed.) *Style in Language*. MIT Press, Cambridge, MA

Kiparsky P, Youmans G (eds.) 1989 *Phonetics and Phonology 1: Rhythm and Meter*. Academic Press, San Diego, CA

Labov W, Waletzky J 1967 Narrative analysis oral versions of personal experience. In: Helm J (ed.) *Essays on the Verbal and Visual Arts. Proceedings of the 1966 Annual Spring Meeting of the American Ethnological Society*. University of Washington Press, Seattle, WA

Lakoff G, Johnson M 1980 *Metaphors We Live By*. Chicago University Press, Chicago

Malone J L 1988 Underspecification theory and Turkish rhyme. *Phonology* **5**: 293–7

Sebeok T (ed.) 1960 *Style in Language*. MIT Press, Cambridge, MA

Sperber D, Wilson D 1995 *Relevance: Communication and Cognition*. Blackwell, Oxford, UK

Tedlock D 1983 *The Spoken Word and the Work of Interpretation*. University of Pennsylvania Press, Philadelphia, PA

Widdowson H G 1975 *Stylistics and the Teaching of Literature*. Longman, London

N. Fabb

Language and Philosophy

Linguistics arose partly from the need to teach Greek to Egyptian children during the Hellenistic period in Alexandria, and partly from the philosophical analysis of the notion of truth by, mainly, Plato and Aristotle. The philosophical origin has been with the study of language ever since. In the twenty-first century, philosophy has renewed its interest in language in various ways (the 'linguistic turn'). Language has been seen as providing a way to break through the Kantian dilemma of the unknowability of bare reality, and as an instrument for the perceptible expression of thoughts, and thus as a key to the inner workings of the mind and the position of humans in the world. The philosophical concern with language is not always distinguishable from the study of meaning: semantics (with its logical leanings) and philosophy are still intimately connected.

1. The Paradox of Knowledge

The fact that whatever there is supposed to be 'out there' in the 'real' world cannot be known directly, without the intervention of human cognitive processing, and indeed that it cannot be proven that there is a world 'out there,' has been a main concern in philosophy since Descartes. Kant gave what seems to

be the final formulation: the thing in itself (*das Ding an sich*) cannot be proven to exist and cannot be known. Yet humans generally assume that there is a world 'out there,' and that they have reliable knowledge of it. Many philosophers have since tried to find a way out of this paradox.

Following Leibniz, Bertrand Russell developed the program of *logical atomism*, meant to reduce all thoughts and all sentences to logically atomic propositions and their truth-functional compositions—in fact, to expressions in his language of Predicate Calculus now used universally in mathematics and the sciences. By hypothesis, this systematic logical analysis would then be the closest humans can get to bare reality (the 'furniture of the world'). By 1930 Russell had given up this idea, considering it unrealistic. His student Wittgenstein pursued the same idea in his *Tractatus Logico-philosophicus* (1922), though in a somewhat different and not altogether coherent fashion, and also concluded around 1930 that it was unfounded. Even so, this program has been very influential in subsequent philosophy. For example, Russell's Theory of Descriptions (1905), which was part of his logical atomism, dominated the theory of reference throughout the twentieth century.

After 1930, Wittgenstein attempted a different approach to the paradox of knowledge. He took to reflecting about language in a nonformalist way, against the background of questions of knowledge and being. He taught that language had no rules, that linguistic interaction was a 'game,' and that word meanings were based on 'family resemblance.' These views (Wittgenstein 1953) have since been rejected as false or imprecise. Yet his influence has been considerable, especially among students of literature and art, and also in the Oxford-based Ordinary Language Philosophy (OLP).

OLP was another influential confrontation with the paradox of knowledge, dominating much of Anglo-Saxon philosophy between 1945 and 1970. Its program, partly inspired by Wittgenstein's teachings about language after 1930, consisted in exploring to what extent a meticulous analysis of the ways words are used in ordinary discourse will reveal the reality of things, and also whether philosophical questions about them are or are not vacuous. Both aims remained unfulfilled. The age-old philosophical questions proved resistant to this kind of linguistic analysis, and the reality of things did not come much closer.

By 1970 there was a consensus among Oxford philosophers that they should reorient themselves and widen their scope. The reasons for the failure were mainly the exclusive focus on English, neglecting the other thousands of languages in the world, the lack of formal theory, and the refractory character of the questions dealt with. Even so, OLP led to two developments that have proven to be of great importance.

The first was *presupposition theory*, initiated by Strawson (1950). For the first time in the history of philosophy, presuppositions were recognized as a special kind of entailment of sentences. Since 1950 presupposition theory has been an important part of either semantic or pragmatic theory (the matter is controversial; see, for example, Carston 1998, Seuren 2000).

The second was *speech act theory*, initiated by Austin (1962). Austin raised the question of the truth value of 'performative' sentences such as 'I hereby promise to marry you.' If they are true, they are so because they have been pronounced seriously: they make themselves true by being pronounced. It appears that they cannot be false. This led to many studies, mainly in pragmatics, of the socially committing character of seriously uttered sentences. An integration with semantics, however, has not so far taken place.

A specific and much publicized angle on the paradox of knowledge was taken by Quine in Chap. 2 of *Word and Object* (1960). Here he discussed the difficulty of explaining how a language learner can learn the precise meaning of a new word, for example *gavagai* said by an exotic native on the passing of a rabbit. No amount of repetition in different circumstances will ensure that the learner gets exactly the meaning intended by the speaker. Yet the precise meaning will be learnt after a while. Quine called this the problem of the 'indeterminacy of radical translation.' It appears that the *gavagai* problem is but an instance of the general and yet unsolved problem of the acquisition of adequate knowledge of what is in 'other minds,' and indeed of adequate knowledge of anything.

2. *Intensional Contexts*

For Aristotle, truth consists in a correspondence of what is said (or thought) and what is the case. This implies that the choice of a word makes no difference for the truth value of a sentence, as long as it refers to the entity intended. This is the principle of Substitution Salva Veritate (SSV), formulated by Leibniz as (roughly) 'Things referred to by words that can stand in for each other without change in truth value are identical.'

Frege (1892) found that SSV does not work for referring expressions in *intensional contexts*, i.e., in clauses embedded under a verb reporting a third party's thoughts, such as 'John believes that the Morning Star is inhabited.' Since the Morning Star is identical with the Evening Star, both being the planet Venus, if the sentence given is true, the sentence 'John believes that the Evening Star is inhabited' should automatically also be true. This, however, is not so, since John may think that the two are not identical, and he may think that the one is inhabited but the other is not.

Frege solved this by positing that clauses embedded under verbs of thought refer not to their truth or falsity in the real world but to the thought expressed, and by observing that 'the Morning Star is inhabited' is a different thought from 'the Evening Star is inhabited,' which blocks SSV. Subsequent philosophers of language, being more of a positivist mind, rejected Frege's appeal to thoughts and sought alternative ways of solving the problem of intensional contexts, which thus became the starting point for the development of formal semantics in the twentieth century.

3. *Formal Semantics*

Since Aristotle, philosophers and linguists have striven to define the logical structure of sentences, i.e., the translation of sentences into some logically analytical language. In modern times this has been the Language of Predicate Calculus (LPC) devised by Russell, with its quantifiers and variables. These efforts are motivated by the wish to specify as exactly as possible under which conditions the propositions expressed by sentences are true or false. Logic is taken to provide such a specification; language is not; hence the desire to relate the latter to the former by translation.

The most explicit way of achieving this is formal or model-theoretic semantics, which defines how linguistic elements are to be translated into LPC, and how, for any given 'world,' LPC expressions are to be valued true or false. By generalizing over all possible 'worlds,' the truth conditions of sentences should, in principle, be statable, including those that involve intensional contexts. Clauses under verbs of thought are now taken to refer not to thoughts but to the sets of those possible worlds in which they are true. Sophisticated formal machineries should then account for the phenomena.

This program was initiated by Montague (1973). Although it seemed very promising during its first years, it has met with serious difficulties. It seems in principle unable to grasp certain cognitively determined aspects of lexical meaning (e.g., why can a speed bump be called *Schwelle* in German, but not 'threshold' in English?). More seriously, it fails to solve the problem of intensional contexts. Since a clause *C* containing a contradiction is false in all possible worlds, it should be synonymous with any other contradictory clause *D*. Yet 'John believes that not all circles are round' is not synonymous with 'John believes that not all bachelors are unmarried,' even though both embedded clauses are necessarily false and hence refer to the empty set of possible worlds; likewise for necessarily true clauses, which refer to the set of all possible worlds. This problem has so far proved unsolvable in the context of model-theoretic semantics (Dowty et al. 1981, pp. 170–5), for which reason many semanticists are now looking for a more cognitively oriented semantic theory.

4. *Thoughts and Propositions*

Talk about 'thoughts' is hazardous. There is not only the difficulty of grasping deep cognitive processes but also the fact that what we call thoughts occur at different levels of explicitness and consciousness. Here, a thought is taken to be (the result of) a mental process whereby the subject assigns a property to one or more entities, creating a representation of a possible situation. A thought that is ready for conversion into a sentence is a *proposition*. The property is expressed by the *predicate* of the proposition, the entities by *terms*. Propositions are always in the scope of a speech act operator SAO defining the speaker's social commitment with regard to it. [SAO + proposition] is converted into a sentence by the lexicon and grammar of the language in question.

5. *The Semantics of Predicates*

Predicates signify a property of one or more elements belonging to a certain class. For each predicate *P* there is a set of conditions to be fulfilled by any entity *e* for it to deserve *P*. These are the *satisfaction conditions* of *P*. In forming a proposition, the speaker considers one or more entities *e* as having the property signified by the predicate, thus forming a mental representation of a possible situation. Propositions are normally written as 'P(a),' where *P* stands for (refers to) the property and *a* refers to the element(s) *e*. If *e* actually has the property signified by *P*, the proposition is true; if not, it is false.

Some predicates are applicable only to sets of individuals. Predicates such as 'congregate,' 'disperse,' 'be numerous,' etc. require subject terms referring to groups, not individuals. Likewise, a predicate such as 'collect' requires an object term referring to a group, as one cannot collect an individual stamp, but one can collect stamps. Such predicates are higher-order predicates with regard to one or more of their terms.

It is the lexicographer's daily business to define satisfaction conditions of predicates, and the lexicographer's daily experience to find that this is extremely difficult. Often an appeal is required to what is *prototypical* in a community. Thus, the predicate 'bald' requires the (relative) absence of hair in prototypical places of the body (the top of the head for humans). Often conditions are approximative, as for the predicate 'flat,' which expresses an approximation to two-dimensionality. More often, conditions are vague, as with predicates such as 'good,' 'bad,' 'please,' 'love.' Subjective evaluations often play a central role, as with 'pleasant' or 'beautiful.' Whether this is also the case with predicates such as 'good' or 'true,' or whether these predicates are satisfied exclusively by the ontological properties of the subject term's referent is a much debated philosophical issue.

It seems useful to divide satisfaction conditions into two classes: first, the preconditions, which define the

right category of entities; then the update conditions, which define the entities that fall under the predicate. Thus, the predicate 'married' requires preconditionally that *e* be a person of marriageable age; then requires as an update condition that *e* be bound by legal marriage to another person, with all that this entails. 'Divorced' requires preconditionally that *e* has been married before and states that this is no longer so.

When *e* is semantically not of the right kind, as in 'The river Thames is divorced,' we speak of a *category mistake*. Lexical opposites, such as 'polite' and 'impolite,' generally stay within the same category ('rational being').

The preconditions play an important role in the *interpretability* of the resulting proposition. Whether the preconditions are actually satisfied in the world is irrelevant for mere interpretation: actual truth values are immaterial for comprehension. What counts is that the preconditions are fulfilled in the cognitive context preceding the utterance in question. The conditions to be fulfilled by (preceding) discourse for an uttered sentence *S* to be interpretable are the presuppositions of *S* (Seuren 1994). Presuppositions result from the preconditions of predicates.

Some predicates, in particular those of the possessive complex, such as 'have,' 'with,' or 'of,' need an appeal to common knowledge for their satisfaction. Compare the following sentences with the predicate 'have':

(1a) Each room in our hotel has a shower.
(1b) Each student in our faculty has a supervisor.

(1a) requires a one-to-one correspondence of rooms and showers for truth, whereas for (1b), a one-to many relation between students and supervisors suffices. It does not seem feasible to record such differences in the lexicon, since the decisive criterion is found not in the linguistic description of 'have,' but in general world knowledge: one has to know how things work in hotels and faculties to decide on the truth or falsity of sentences such as (1a,b).

Lexical meanings are, moreover, inherently dynamic: they are naturally extended, restricted, shifted, and even misapplied in ways that are nevertheless perfectly understood and often even become conventionalized to the point of becoming part of the lexicon. 'Sad,' for example, can be used intransitively, as in 'a sad mood,' but also causatively, as in 'a sad story.' Yet this process is not freely productive: 'nervous' can only be used intransitively, as in 'a nervous person,' but not causatively, as in 'a nervous story.'

Other forms of dynamic use are known as figures of speech or 'tropes,' mainly metonymy and metaphor. Metonymy is the use of a noun for something which is closely related to that for which the noun is normally used: 'crown' for 'monarch,' or 'grave' for 'death.' Metaphor is a much more complex trope, perhaps best defined as an implicit comparison. Deliberate category mistakes often lead to metaphor, as in 'The train was waltzing at sunset round the walls of Verona' (E. M. Forster 1975). Although 'waltz' requires a subject referent that is capable of dancing, a condition not fulfilled by trains, the choice of this predicate forces the reader momentarily to consider the train referred to as such a being, which provokes the artistic effect intended. Other metaphors arise as a result of applying an incorrect predicate to an entity referred to, as when someone asks you where you left your 'horse,' meaning your car. Tropes often become conventionalized ('dead metaphors'), and may enter the lexicon of the language. Conventionalization of tropes is a principal source of lexical enrichment.

The most important philosophical point about lexical meaning is the necessity to appeal to cognition. Satisfaction conditions cannot be defined purely in terms of 'objective' properties of the entities concerned. The cognitive aspect is particularly important with regard to the discourse dependency of lexical meanings, a lexical reflex of the general fact that thoughts do not occur in isolation but always in a wider context. This is bad news for model-theoretic formal semantics, which assumes that meanings are fully determined by conditions of truth and satisfaction in models containing no minds, but only a language, worlds, and an interpretation function from elements in the language to elements in the world(s).

6. *Reference*

Reference is the relation between definite argument terms of a predicate and things in the world. The reference relation is *intended* by the speaker and *comprehended* by the listener. Reference is a necessary condition for the proposition to have a truth value. A host of problems has arisen in connection with the reference relation. We shall make global mention of the main problem areas. Detailed analyses are to be found in the specialized literature.

6.1 *The Identification Problem*

The first problem is how speakers and listeners fix upon the intended entities by the use of definite terms. A term such as 'the dog' is naturally rendered as 'the *x* such that *x* is a dog' (Bach 1968). The problem is how the determiner 'the *x*' can pick out, from the set of dogs, the one particular dog intended by the speaker. The answer is trivial when the set contains just one element, but human speakers refer successfully even when the sets in question contain many elements.

This problem has occupied countless philosophers, most of whom recognized that contextual and cognitive factors must play a role. Russell (1905) proposed to eliminate definite terms by refinterpreting them as instances of existential quantification. This Theory of Descriptions should not only solve the identification

problem, but also the problems of reference to nonexisting entities and of reification (see below). Russell's famous example is (2a), rendered by him in logico-semantic terms as (2b):

(2a) The present king of France is bald.

(2b) $\exists x\ [\mathrm{Now}[\mathrm{KoF}(x)] \wedge \mathrm{Bald}(x) \wedge \forall y[\mathrm{KoF}(y)\ x = y]]$

(there is an x such that x is now king of France and x is bald and such that for all y, if y is king of France, x is identical with y)

The definite term 'the present king of France' has been replaced by an analysis in terms of the existential quantifier and a 'uniqueness clause' saying that for all y, if y is king of France, x is identical with y, which means that there is exactly one king of France.

Russell's analysis, though widely accepted for a long time, is inadequate for semantic analysis. This appears from embedding in emotive-intensional contexts. A sentence such as (3a) does not mean what (3b) means:

(3a) Nora was disappointed to find that the king was bald.

(3b) Nora was disappointed to find that there was a person who was both king and bald and that there was no other king.

The recognition that the Russellian reduction program does not work for language, and that definite terms must be reinstated is gaining ground. The identification problem has thereby returned in full force.

Further problems to do with the identification of reference objects are Quine's distinction between opaque and transparent reference, and Donnellan's between referential and attributive use of a definite term. The former, first presented in Quine (1953, pp. 139–59), has to do with the choice of the predicate P in a definite description. When P is selected by the speaker, so as to facilitate identification by the listener, there is *transparent reference*. It is, however, also possible for P to be selected by a person whose thoughts are spoken about, in which case the reference is *opaque*. For example,

(4) Sarah is convinced that the stupidest boy will get the scholarship

is ambiguous. On the transparent reading, the boy referred to is considered stupid by the speaker, but not by Sarah, while on the opaque reading Sarah takes a dim view of the award committee.

In line with Russell's Theory of Descriptions the difference has been analyzed as one of scope. The two readings of (4) would thus be (leaving out the uniqueness clause; 'SB': 'stupidest boy'):

(5a) $\exists x\ [\mathrm{SB}(x) \wedge \mathrm{Convinced}(\mathrm{Sarah}, [\mathrm{Get}(x, \text{the scholarship})])]$

(5b) $\mathrm{Convinced}(\mathrm{Sarah}, [\exists x\ [\mathrm{SB}(x) \wedge \mathrm{Get}(x, \text{the scholarship})]])$

The problem with this solution, however, is that the negation of (4):

(6) Sarah is not convinced that the stupidest boy will get the scholarship

still has the transparent-opaque ambiguity, whereas the negations of (5a,b), respectively, lead to a different result: (7a) means that there is no stupidest boy of whom Sarah is convinced that he will win the scholarship, which is not what (6) means; only the presumed opaque (7b) seems viable:

(7a) $\neg\exists x\ [\mathrm{SB}(x) \wedge \mathrm{Convinced}(\mathrm{Sarah}, [\mathrm{Get}(x, \text{the scholarship})])]$

(7b) $\neg\mathrm{Convinced}(\mathrm{Sarah}, [\exists x\ [\mathrm{SB}(x) \wedge \mathrm{Get}(x, \text{the scholarship})]])$

Be that as it may, after the 1970s the issue more or less disappeared from the literature.

The second distinction (Donnellan 1966) is that between attributive and referential uses of definite terms. In what Donnellan calls *referential use*, the predicate P of the definite term merely serves to identify a given entity. When, at a party, one uses the expression 'the man holding a martini' and there is only one person who can possibly be intended, the reference is successful, even if it turns out that what the man was holding was in fact a glass of water. In *attributive use*, on the other hand, P is used to refer to whatever unknown entity will satisfy P, as when I say 'Smith's murderer is unknown.' 'Smith's murderer' is used here as a parameter expression (like 'temperature,' 'name,' 'phone number'), whose value is still to be identified. Now P cannot turn out to be false or mistaken while the reference is still successful.

Donnellan recognized that Russell's analysis of definite terms does not account for attributive uses, but he failed to provide an alternative analysis that does. The subsequent philosophical literature, where Donnellan's distinction is often mentioned but never accounted for, likewise fails to provide a solution.

6.2 Reference and Ontology

A second cluster of reference problems is to do with the ontological status of the entities referred to, in particular reifications and nonexisting reference objects.

There is reification when speakers know that a definite description does not refer to an actually existing entity but to whatever is captured by a mental construction, as in 'the average London cabdriver,' 'the military-industrial complex,' or 'the difference between A and B.' In such cases, the (usually complex) predicate P of the definite description does not denote a set of entities considered to be ontologically real, but requires a (vague or precise) formula of interpretation that will reduce the satisfaction conditions of P to what may be considered ontologically bare entities in certain relations to each other.

Russell (1905) presents his Theory of Descriptions as a solution to the reifications problem as well, but he fails to show how it works. The subsequent literature has, accordingly, neglected this aspect of Russell, as it has neglected reifications in general.

Quine (1960) speaks of 'ontological commitment' to the entities referred to in the use of a natural or formal language. However, whatever 'ontological commitment' speakers of natural languages have with regard to the things they talk about is entirely a matter of language-independent belief. Speakers using reifications know that the world contains no 'entities' of the kind suggested by the reifying terms. Yet they use them as if they really were referring terms.

Reification resembles what in computer science is called 'object-oriented' programming, i.e., a method whereby complex descriptions are bundled together under a formal definition and treated as if they were terms referring to objects in the model, whereas their definition says that they are not. Object-oriented programming has turned out very efficient, because it enables the user to handle complex relations as if they were single objects. It looks as if Nature made the same discovery when it was working at the development of human language. Exactly *how* Nature implemented its version of object-orientedness is a matter for future research.

A second central problem is that of nonexisting reference objects, Russell's main concern in his Theory of Descriptions. In (2a), the definite description 'the present king of France' does not refer to an existing entity susceptible to baldness. Does it, then, refer to a nonexisting entity? Russell, like Quine, rejected the notion of nonexisting entities on the grounds that they might be said to both have and not have any given property, which would make them unfit for rational analysis. In Russell's view, which is considered standard, the being asserted by the existential quantifier is actual, real existence, and not some form of virtual being.

While Russell's analysis might seem to solve the problem of nonexisting entities for extensional sentences such as (2a), it fails for cases such as:

(8) Jones always talks about the present king of France

which may well be true even if there does not now exist a king of France. Formal semantics stipulates that a predicate such as 'talk about' may take an 'intensional' object term referring to possible entities in possible worlds, but here the problem of intensional contexts repeats itself: talk about a square circle is not the same as talk about a living corpse, yet they are both uninstantiated in any possible world. The problem of nonexisting reference objects has not so far been solved.

6.3 Donkey Sentences

Since Geach (1962) the problem of so-called *donkey* sentences, such as (9a,b), has dogged the theory of reference:

(9a) Every farmer who owns a donkey feeds *it*.
(9b) If a farmer owns a donkey *he* feeds *it*.

The name 'donkey sentences' derives from the relevant examples in Geach (1962, p. 116ff.), their modern source, which all involve donkeys. Geach, however, copied Walter Burleigh (1988, p. 92), who introduced donkey sentences around 1328 in the context of his theory of reference (*suppositio*): *Omnis homo habens asinum videt illum* (every man who has a donkey sees it).

The problem is that donkey sentences cannot be translated into standard LPC, which has only two kinds of terms: constants (definite referring terms) and variables. 'It' in (9a,b) and 'he' in (9b) cannot be constants, since they need not refer for the sentences to be true. So they must be variables. But variables must be bound by quantifiers, and this is impossible, due to scope problems. The LPC translations of (9a,b) should be (10a,b), respectively:

(10a) $\forall x[\text{Farmer}(x) \wedge \exists y[\text{Donkey}(y) \wedge \text{Own}(x,y)] \text{Feed}(x,\mathbf{y})]$
(10b) $\exists x[\text{Farmer}(x) \wedge \exists y[\text{Donkey}(y) \wedge \text{Own}(x,y)]] \text{Feed}(\mathbf{x},\mathbf{y})$

The bold-faced variables are unbound, which makes (10a,b) uninterpretable. Attempts at binding the variables under wide-scope universal quantification ('for all farmers *x* and for all donkeys *y*, if *x* owns *y*, then *x* feeds *y*' for (9a,b) alike) founder not only because they violate the principle of 'parity of form' (Russell 1905: 483), but also because they fail to work when intensional operators are inserted, as in:

(11a) Every farmer who is glad he owns a donkey is sad he has to feed *it*.
(11b) If a farmer is glad he owns a donkey, *he* is sad he has to feed *it*.

Due to scope problems, there is no way in which (11a,b) can be adequately analyzed, either analogously to (10a,b) or under wide-scope universal quantification.

Any solution will require an extension of standard LPC with referring pronouns that are not bound variables. In fact, a simple test shows that the problematic pronouns in (9a,b) are not bound variables at all, but referring terms. In general, referring pronouns can be replaced with an *epithet*, i.e., an unaccented descriptive noun phrase expressing some evaluative judgment. Bound variable pronouns or reflexive (logophoric) pronouns do not allow for such a replacement: in (12) the epithet cannot function as a bound variable:

(12) No man expects his friend to betray *him* (!*the poor devil*).

Yet epithet replacement functions perfectly well for (9a,b):

(13a) Every farmer who owns a donkey feeds *the wretched animal*.
(13b) If a farmer owns a donkey, *the man* feeds *the wretched animal*.

The problem for LPC and standard semantics is, however, that if the pronouns in question are referring terms, interpretation of (9a,b) is blocked when there

are no farmers or no donkeys. Yet these sentences are considered true in such cases. This constitutes a serious dilemma.

6.4 The Cognitive Turn

The main affliction of twentieth-century philosophy of language and semantics is the inability to come to terms with the cognitive factor in language and language use. Sentences and terms do not physically sit on situations and reference objects, respectively. The relation between the former and the latter is mediated by cognizing humans. Yet philosophers and semanticists have pretended that the mind was not there. We are faced today with the effects of this aberration. Now that the view of the mind-brain as a massive computation plant has opened new perspectives, solutions come in sight for many of the problems enumerated above. It would be premature, however, to try and specify in what ways a cognitive account of meaning and language use can be of help in solving the extant problems.

See also: Communication: Philosophical Aspects; Language, Continental Philosophy of; Linguistics: Overview; Meaning and Rule-following: Philosophical Aspects; Pragmatics: Linguistic; Reference and Representation: Philosophical Aspects; Semantics

Bibliography

Austin J L 1962 *How to Do Things with Words* [ed. Urmson J O]. Clarendon Press, Oxford, UK

Bach E 1968 Nouns and noun phrases. In: Bach E, Harms R T (eds.) *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York, pp. 90–122

Burleigh W 1988 *Von der Reinheit der Kunst der Logik. Erster Traktat. Von den Eigenschaften der Termini* [trans. and ed. Kunze P]. Felix Meiner, Hamburg, Germany

Carston R 1998 Negation, 'presupposition,' and the semantics/pragmatics distinction. *Journal of Linguistics* **34**: 309–50

Davidson D 1980 *Essays on Actions and Events*. Clarendon Press, Oxford, UK

Donnellan K 1966 Reference and definite descriptions. *Philosophical Review* **75**: 281–304

Dowty D, Wall R E, Peters S R 1981 *Introduction to Montague Semantics*. Reidel, Dordrecht, The Netherlands

Forster E M 1975 *Where Angels Fear to Tread*. Penguin Books, Harmondsworth, UK

Frege G 1892 Ueber Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* **100**: 25–50

Geach P T 1962 *Reference and Generality. An Examination of Some Medieval and Modern Theories*. Cornell University Press, Ihaca, NY

Linsky L (ed.) 1971 *Reference and Modality*. Oxford University Press, Oxford, UK

Montague R 1973 The proper treatment of quantification in ordinary English. In: Hintikka K J J, Moravcsik J M E, Suppes P (eds.) *Approaches to Natural Language. Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*. Reidel, Dordrecht, The Netherlands, pp. 221–42

Platts M (ed.) 1980 *Reference, Truth and Reality*. Routledge and Kegan Paul, London

Quine W V O 1953 *From a Logical Point of View. Logico-philosophical Essays*. Harvard University Press, Cambridge, MA

Quine W V O 1960 *Word and Object*. MIT Press, Cambridge, MA

Quine W V O 1973 *The Roots of Reference. The Paul Carus Lectures*. Open Court, La Salle, IL

Rosenberg J F, Travis C (eds.) 1971 *Readings in the Philosophy of Language*. Prentice Hall, Englewood Cliffs, NJ

Russell B 1905 On denoting. *Mind* **14**: 479–93

Schiffer S R 1972 *Meaning*. Clarendon Press, Oxford, UK

Seuren P A M 1994 Presupposition. In: Asher R E, Simpson J M Y (eds.) *The Encyclopedia of Language and Linguistics*. Pergamon Press, Oxford, pp. 3311–20

Seuren P A M 2000 Presupposition, negation and trivalence. *Journal of Linguistics* **36**: 261–97

Steinberg D D, Jakobovits L A (eds.) 1971 *Semantics. An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*. Cambridge University Press, Cambridge, UK

Strawson P F 1950 On referring. *Mind* **59**: 320–44

Wittgenstein L 1922 *Tractatus Logico-philosophicus* [trans. Ogden C K]. Routledge and Kegan Paul, London

Wittgenstein L 1953 *Philosophical Investigations* [trans. Anscombe G E M]. Blackwell, Oxford, UK

P. A. M. Seuren

Language and Poetic Structure

Language is the material out of which poetry is made. Language itself is, of course, a complex phenomenon, involving both a characteristic structure, consisting of a grammar and a lexicon, and characteristic uses, such as communication and expression. All of these elements play a role in poetic structure. This entry, however, will focus mainly on grammar, and its relationship to such characteristic aspects of poetic structure as meter, rhyme and alliteration, and syntactic parallelism. These define the core forms of versification found across the world's languages, and create poetic effects even when they are not used regularly enough to constitute versification. In recent decades, they have been the subject of striking new directions of research, as a result of new directions of research in grammar.

1. Grammar and Verse

Versification by definition imposes formal constraints on language beyond those to which it conforms intrinsically. To describe those constraints for any given verse form is a surprisingly challenging intellectual problem, because individual lines within a form are all different from one another, often artfully so for expressive purposes. It is thus a problem, as Jakobson (1981a) influentially put it, of abstracting the invariant 'verse design' away from the variation

embodied in all the 'verse instances.' Even for those verse forms, especially in written traditions, for which some constraints have been articulated, the problem remains a challenging one, in that poets' actual practice may not necessarily match the articulated constraints: some lines which appear to violate them may nonetheless be included in compositions in the form, at the same time that some which appear to conform to them may nonetheless never be used. Such mismatches suggest that the articulated constraints may not constitute a complete and true description of what a poet instinctively knows about a form.

Construing the problem of describing a verse form as making explicit the unconscious as well as conscious knowledge a poet has of what defines the form makes it, as Halle and Keyser (1972) point out, analogous to the problem of describing grammar itself. In Chomsky's (1965) influential formulation, grammar is the knowledge a fluent speaker of a language has, partly consciously and partly unconsciously, of what would or would not be well-formed instances of the language. In phonology, for example, the subsystem of grammar dealing with sound structure (see *Phonology*), an English speaker knows not only that the combination of sounds in 'brick' is an actual word and that in 'blick' is not, but also that 'blick' is a possible word, whereas '*bnick' is not only not an actual word but also not a possible word; the '*' indicates its ill-formedness. Similarly in syntax (see *Syntax*), an English speaker knows that a sentence like 'The cat stretched sleepily' is a possible sentence of English while '*Sleepily cat stretched the' is not. To describe the grammar of a language is to make the constraints which constitute such knowledge fully explicit. The research program within linguistics which adopts that goal is 'generative linguistics' (see *Generative Grammar*), and its analog for verse is sometimes termed 'generative poetics.'

Describing a verse form involves grammar in substantive aspects as well. Although meter, rhyme and alliteration, and syntactic parallelism are often described metaphorically as contributing music to poetry, the entities which define them—particular configurations of syllables and stress, vowels and consonants, noun phrases and verb phrases—are, of course, quite specific to grammar. This means that as research into grammar has sought to discover the fundamental entities and organizational principles of grammars, new formal structures have become available to serve in descriptions of verse forms, sometimes leading to striking new insights into familiar forms. By the same token, research into verse forms has sometimes suggested or supported hypothesized entities or principles of grammar.

In this regard, two ideas which have shaped grammatical research have been particularly significant for research into versification. One is the assumption, going back to de Saussure (1983) and carried on in generative grammar, that grammar is at some level purely formal: phonology is not reducible to sound, and syntax is not reducible to meaning. The entities of grammar defining verse forms may thus sometimes prove to be surprisingly abstract.

The other is the conceptualization of the grammar of any particular language as one instance of a common phenomenon of human language. Although this idea also has a long history, including notably for poetics the interest in it taken by Jakobson and his circle, it was given a particularly galvanizing formulation in Chomsky's (1965) suggestion that the most basic formal properties of grammars might be essentially similar across all languages because they are innately determined. This hypothesis of 'universal grammar' motivates the relevance of wide-ranging cross-linguistic phenomena to the formalization of any single language's grammar. For versification, this means that evidence for particular ways of analyzing the structures figuring in a verse form may come from languages unrelated to that of the verse in question.

The observation that the goals and vocabulary of the description of verse are similar to those of the description of grammar invites the deeper question of whether there can be a theory of versification, as there are theories of grammar, and, if so, what specific relationship of versification to grammar such a theory might posit. After all, there are differences as well as similarities: we can usually tell in an instant whether language is verse or not. Again it was Jakobson (1962–81a) who influentially suggested that it is specifically the sequencing of entities which are in some way equivalent in the language which constitutes poetic effects. Interpreting this in light of the formal and universalist understanding of grammar characteristic of generative theories, Kiparsky (1981b, 1987) ventured one of the first testable theories of versification, hypothesizing that verse forms might be constituted by repetition of exactly those phonological and syntactic structures to which the constraints which constitute grammars universally make reference.

One of the simplest observations in support of this idea is that each of the major forms of versification corresponds to a major subsystem of grammar. Let us turn then to some specific recent examples of descriptive and theoretical claims following this line of inquiry, and then consider some of its consequences and limitations.

2. *Meter*

The phonology of every language involves two distinct though interrelated subsystems, one dealing with rhythm, such as the quantitative grouping of sounds into stressed and unstressed syllables, and the other dealing with what is technically termed 'melody,' referring not only to tones, but also to the qualitative characteristics of segments, such as what makes a 'p' different from a 'b.' Verse forms which impose

constraints on rhythmic structure beyond those given by the phonology are meters.

The dominant meter of the modern English poetic tradition, the iambic pentameter, affords a well-developed example of how changes in the goals and substance of grammatical description have wrought changes in the conceptualization of a familiar verse form. Traditionally, the iambic pentameter has been described as having a basic form of five feet, each consisting of an unstressed syllable (^{x}) followed by a stressed one ($^{/}$):

(a) $^{x/|x/|x/|x/|x/}$

While many lines, such as (bi), have exactly that rhythmic structure, many others, such as (b) (ii)–(iii), do not:

(b) (i) So long as men can breathe and eyes can see (Shakespeare, *Sonnets* 18, 13)

(ii) Pluck the keen teeth from the fierce tiger's jaws, (19, 3)

(iii) For how do I hold thee but by thy granting, (87, 5)

Such variations have traditionally been attributed to the possibility of substituting different kinds of feet with different patterns of syllables and stress, such as trochees ($^{/x}$), spondees ($^{//}$), phyrrics (^{xx}), or anapests ($^{xx/}$), so long as any sense of the basic form doesn't get lost. But as Halle and Keyser (1972), building on Jespersen (1933), point out, unless it makes explicit under what conditions the basic form would indeed get lost, such a description fails to capture what the poet knows that leads to acceptance of some variations as metrical, but avoidance of other imaginable ones as unmetrical. The construct in (c), for example, sounds like prose, not verse, and in fact has a rhythmic structure not found in any lines in Shakespeare's sonnets:

(c) *Imaginable lines are avoided.

Yet it can be analyzed into basic and substituted feet just as the lines in (b) (ii)–(iii) can; nothing in that description of the iambic pentameter captures its fundamental difference of unmetricality.

Much research has gone into the question of how the iambic pentameter can be modeled so as to in fact represent this knowledge. Most fundamentally, Halle and Keyser (1972) suggest that the meter needs to be distinguished from the actual rhythmic structure of the line, and formalized in two components: an abstract rhythmic template, composed of weak (W) and strong (S) metrical positions, and simple constraints on how certain aspects of the rhythmic structure of the line can be mapped into the template. Observing that lines such as those in (d), which are rhythmically similar to those of (b) (ii)–(iii) but have polysyllabic words where (b) (ii)–(iii) have monosyllables, do not appear in Shakespeare's sonnets, Kiparsky (1977) proposes specifically the template and constraints in (e), informally restated:

(d) (i) *Pluck immense teeth from enraged tigers' jaws

(ii) *For how do I prosper but by the granting

(e) (i) ws ws ws wsws

(ii) Each metrical position corresponds to a syllable.

No weak metrical position corresponds to a syllable which is strong relative to another syllable in the same word.

This proposal is shaped in substantive ways by a model of linguistic rhythm itself as essentially relative, with elements at successive levels grouped into constituents in which one element is strong and any others weak (Liberman and Prince 1977), such that a polysyllabic word involves a relationship of relative strength between syllables that a monosyllabic word does not. Thus, lines like (b) (ii) can be mapped into the template in (e) (i) in accordance with the constraints in (e) (ii), while (d) (i), let alone (c), cannot:

s w

(f) (i) Pluck the keen teeth from the fierce tiger's jaws,

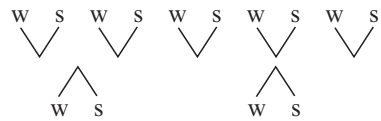

(ii) *Pluck immense teeth from enraged tiger's jaws

The cross-linguistic study of rhythmic systems within Liberman and Prince's model has led to the discovery of a rich range of abstract structures, with surprising consequences for the description of many other aspects of metrical structure as well. For example, the familiar possibility of 'feminine endings,' the inclusion of an extra unstressed syllable at the end of a line as in (b) (iii), proves to have an analog in the grammar of rhythm, in that in many languages final syllables are regularly 'extrametrical,' or outside the basic groupings which define stress patterns (Hayes 1995). And the possibility in some iambic meter of 'trisyllabic substitution,' or inclusion of extra unstressed syllables within a line as in (g), has been attributed to the role of a particular linguistic grouping of syllables called a 'phonological foot,' which in many languages takes the form of either a single stressed syllable, like 'sea' or '-ract,' or two syllables of which the first is stressed and light, like 'cata-' (Hayes 1995, Hanson and Kiparsky 1996).

(g) In cataract after cataract to the sea (Tennyson, 'OEnone,' 9)

This possibility has in turn illuminated the structure of other English meters which were previously ob-

scure, such as Hopkins' 'Sprung Rhythm' (Kiparsky 1989), as well as meters of entirely unrelated traditions, such as Finnish. In fact, the universal theory of grammatical rhythm is so well-developed at this point that it has been possible to propose a universal theory of meter based on it (Hanson and Kiparsky 1996), though its cross-linguistic explanatory value remains largely still to be explored.

A template and constraints as in (e) are, of course, only the most basic elements of a metrical description. Even for English iambic pentameter, there are systematic exceptions to the constraints of (eii), as well as additional strictures that they do not capture, and the details are different for every poet and genre; as Kiparsky (1977) observes, describing an individual poet's practice within a metrical tradition is analogous to describing dialects within a language. While formalizing these complexities has likewise made progress by drawing on insights into grammatical structure, many open questions remain. For example, knowledge of a meter clearly includes recognition not only of the limits of well-formedness, but also of specific effects of metrical complexity and simplicity that can be created within those limits. How is that knowledge expressed in models like that in (e)? Is even knowledge of well-formedness best modeled through identification of a limit, or of gradient distance from a prototype (Youmans 1989)? Interestingly, while such questions pose challenges for metrical description, they do not necessarily challenge the theoretical assumption of the close relationship of meter to grammatical rhythm; rather, parallel questions have been posed about modelling grammar itself.

3. *Alliteration and rhyme*

The family of poetic effects which includes alliteration and rhyme is often referred to as 'sound patterning.' The term is somewhat unsatisfactory, however, since it is not clear why these forms have any better claim to be described as 'sound' than meter does, at the same time that sound is not really what any such forms involve, so much as phonological structure. Since the structure they involve is specifically melodic, a more appropriate collective term for them might be 'poetic melody,' parallel to 'poetic meter.'

Again it was Jakobson (1963) who, in an influential analysis of Germanic alliteration, showed how explicit characterization of such forms as repetitions can lead to discoveries about their relationships to grammar. In Old English verse, for example, syllables in certain positions must alliterate, meaning they begin either with the same consonants, as in (h)|(i); or with any vowels, as in (h)|(ii):

(h) i. Da <u>m</u>iddangeard <u>m</u>oncynnes Weard
'then middle earth mankind's Guardian'
ii. <u>e</u>ce Drihten <u>æ</u>fter teode
'eternal Lord afterwards made'

Jakobson observes that no characterization of alliteration in terms of identities of sounds can generalize across this differential treatment of consonants and vowels, and instead relates it to syllable structure. Segments which precede the most sonorous element in each syllable, typically though not always the vowel, constitute an 'onset': in 'rate,' for example, it would be the [r], or in 'trait' the [tr]. Modifying Jakobson's own analysis somewhat, Kiparsky (1970, 1981) points out that if , in 'ate' there is simply no onset, or an empty one, then alliteration can be characterized simply as requiring structural identity of corresponding syllables up through the first segment in the onset. Moreover, onset structure explains another peculiarity of Germanic alliteration, in that exceptionally for the clusters [sp, st, sk], the second segments must also be the same. As Jakobson noted, there is evidence that the [s] in these configurations is outside the core onset structure, or 'extrasyllabic' (compare the fact that in modern English, onsets normally contain no more than two segments, but initial [s] in an onset like that of 'straight' is excepted); thus, the second segment in the cluster really is the first in the onset, and its required identity follows. This explicit account of the locus of Germanic alliteration stands as a model for the analysis of other melodic forms: while rhyme, for example, in all its various forms like full rhyme, slant rhyme or assonance are all assumed to be formalizable, and in ways that similarily draw on grammatical structure, the details of poets' actual practice have seldom been fully worked out in conjunction with explict theories of syllable structure.

Another way in which poetic melody draws on grammatical structure involves the explicit characterization of ways in which sounds can be similar.

In language, segments are often pronounced systematically differently depending on their phonological context. Modelling such knowledge has typically relied on two key assumptions. First, segments are not monolithic, but composed of distinctive features (see *Phonology*). Second, individual lexical items have an 'underlying structure' which may be distinct from their 'surface structure' which determines pronounciation, because phonological processes may alter some features in some contexts (Chomsky and Halle 1968). Poetic melody exploits this kind of knowledge in a variety of ways. First, as Zwicky (1976) suggests for the rhymes of English rock lyrics, a form may simply require corresponding sounds to differ by no more than one feature: in *made* and *babe*, for example, [d] and [b] differ only in their place of articulation feature, and in *made* and *late*, [d] and [t] differ only in their feature of voicing, or vibration of the vocal cords. Second, Kiparsky (1970) suggests that poets' practice is sometimes based on underlying rather than surface forms. For example, Malone (1987) notes that Irish has a grammatical process which nasalizes voiced stops ([b, d, g]) in certain contexts, which explains why

poets allow a word like *bo* 'cow' to alliterate with *mban* 'white (gen.pl.)': the initial sound of the latter is underlyingly [b], though the nasalization process makes it different on the surface Third, poetic forms sometimes require identity of some features but not others, a phenomenon also found in Irish. Irish rhymes require that segments in corresponding positions belong to the same class of consonants; for example, one such class is the voiced stops [b,d,g], such that *éc* 'death' (in which the 'c' spells [g]) can rhyme with *bét* 'misdeed' (in which the 't' spells [d]). That this class is targeted by the nasalization process which gives *mban* makes it a particularly clear instance of what Kiparsky hypothesized to be the general case in verse forms, of the equivalences that define them being the same equivalences which figure in grammar.

Collectively, the foregoing ideas and the cross-linguistic studies which support them suggest that melodic forms support a good deal of variety across verse instances in the same way that metrical forms do, but formalizing their verse designs has generally received less attention. In English, for example, Zwicky's (1976) study has inspired many interesting observations of variation in the high art rhyme tradition as well, but few studies have worked out explicit models of the possible complexities and their limits in the practice of individual poets, let alone characterized them as variations on an overall tradition. A particularly interesting theoretical impetus to such studies has recently come, however, from cross-linguistic studies of reduplication, systematic processes in which all or part of the beginnings or ends of words are copied to derive new words or forms of words, as illustrated in (i):

(i) i. Gothic past tense: haitan 'call' hai-hait 'called'
slepan 'sleep' sai-slep 'slept'
ii. Chamorro intensives: dankolo 'big' dankolo-lo 'really big'
bunita 'pretty' bunita-ta 'very pretty'

As Kiparsky (1973) points out, melodic verse forms often involve the same kinds of grammatical analyses that reduplication does; thus, as the intricate variety of reduplication systems has begun to be studied, universal theories of poetic melody analogous to those advanced for meter have begun to be possible for the first time (Holtman 1994, Yip 1999).

4. *Syntactic Parallelism*

Finally, in many languages it is not phonology, or at least not phonology alone, which defines verse forms, but morphology and syntax, through requirements of parallel structures from line to line. One of the most influential traditions of this kind is that of ancient Hebrew poetry, exemplified by the passage in (j), together with a translation of it by Jakobson (1981b):

(j) ʔittī milləbānōn kallāh
with me from Lebanon bride
ʔittī milləbānōn tābōʔī
with me from Lebanon come!
tāšūrī mērōʔš ʔămānāh
depart from the peak of Amanah,
mērōʔš śənīr wəḥermōn
from the peak of Senir and Hermon,
mimmə ōnōt ʔărāyōt
from the lairs of lions
mēharərēy nəmērīm
from the mountains of leopards!
('Song of Solomon' 4:8)

Here again, making explicit what the evident repetition consists of raises interesting grammatical questions. The fifth and sixth lines transparently have a common morphological and syntactic structure: the individual words differ, but the grammatical classes of their component parts and resulting wholes are identical: prepositions, plural nouns, etc. But the structures of the third and fourth lines appear to differ in that there is a verb, tāŝūrī 'depart' in the third which is absent from the fourth, and a conjoined noun in the fourth, *w′hermon* 'and Hermon,' which is absent from the third. From the point of view of syntax, however, a noun phrase consisting of a single noun can be identical to one consisting of conjoined nouns: in English, for example, a preposition which requires a noun phrase object will be indifferent to the internal structure of the noun phrase, so that [[from]$_P$ [Mary]$_{NP}$]] and [[from]$_P$ [[Mary]$_{NP}$ and [John]$_{NP}$]$_{NP}$] are structurally identical at the level of their [P NP] structure. Similarly, in certain constructions involving parallelism in language itself, a verb which is absent in the surface structure may be understood to be present underlyingly, under identity with a verb in a conjoined clause: in [[Mary [ate]$_V$ fish]$_S$ and [John [ ]$_V$ meat]$_S$], the syntactic well-formedness of the second clause requires the assumption that the unpronounced verb is present at some level of analysis. Even in the first two lines, as Jakobson (1966) points out, the vocative *kallah* 'bride' and the imperative *tabo/i* 'come' share a function of direct address, which may entail further syntactic similarities. Thus, even parallelism which appears to be principally semantic may involve identical syntax at an abstract level. The forms of syntactic parallelism across languages—Finnish, Chinese, Galician, for example—are as diverse as the languages themselves, but all merit further study from a syntactic point of view.

5. *Conclusion*

Versification represents, of course, only one of many ways in which language contributes to poetic structure. For every subsystem of grammar of which some elements define a verse form, other elements remain free to create poetic effects of other kinds: phrasal

rhythms (Cureton 1992); syntactic suspensions (Austin 1984). And even the most traditional elements of verse are not necessarily used pervasively as verse forms, but sometimes only fragmentarily, or in traditional terminology 'ornamentally' rather than 'structurally.' In Shakespeare's sonnets, for example, alliteration and syntactic parallelism come and go within a form versified by meter and rhyme, in a manner which is not fundamentally different from the way meter and rhyme come and go in poems of Eliot which are not versified at all. And not just the grammar but also the lexicon can be the source of significant patterns of repetition. But even as only a limiting case, discoveries of the nature of the relationship of versification to grammar provide remarkable evidence of how forms which seem the height of artifice may in fact be rooted in ways of perceiving which, if the hypothesis of universal grammar is correct, operate as naturally as vision or hearing (Fodor 1983). Such discoveries have consequences for poetic effects of all kinds, whether traditional or innovative, pervasive or fragmentary, natural or genuinely artificial.

Perhaps more strikingly, the focus on versification here has not touched on such fundamental functions of language as communication or expression, which are particularly complex and interesting in the case of poetry. To some extent, this reflects the observation that communication is not unique to language: it has been argued by Sperber and Wilson (1986), for example, that in ordinary language, the phonological and syntactic parsing of a sentence makes only a minimal contribution to what it actually communicates, which in fact depends on an open-ended, context-dependent process of general reasoning which unlike the structural analysis of grammar is common to how information of all kinds is interpreted. At the same time, however there is at least one way in which versification invites its own kind of reasoning: as Wimsatt (1954) notes, the repetitions which consititute versification consistently invite the question of how whatever is designated by similar forms might themselves be alike. In this way, as Jakobson (1956) suggests, there is a metaphorical way of thinking inherent in poetic structure, which is not only a matter of its meaning, but also of its form.

Bibliography

Austin T 1984 *Language Crafted*. Indiana University Press, Bloomington, IN

Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Cureton R 1992 *Rhythmic Phrasing in English Verse*. Longman, London

Fabb N 1997 *Linguistics and Literature*. Blackwell, Oxford, UK

Fodor J 1983 *Modularity of Mind*. MIT Press, Cambridge, MA

Halle M, Keyser S J 1972 The iambic pentameter. In: Wimsatt W K (ed.) *Versification: major language types*. Modern Language Association, New York

Hanson K, Kiparsky P 1996 A parametric theory of poetic meter. *Language* **72**: 287–335

Hanson K, Kiparsky P 1997 The nature of verse and its consequences for the mixed form. In: Reichl K, Harris J (eds.) *Prosimetrum: cross-cultural perspectives on narrative in verse and prose*. D. S. Brewer, Cambridge, UK

Hayes B 1995 *Metrical Stress Theory*. The University of Chicago Press, Chicago

Holtman A 1994 A constraint-based approach to rhyme. *Linguistics in the Netherlands* **1994**: 49–60

Jakobson R 1971 Two aspects of language and two types of aphasic disturbances. In: Rudy S (ed.) *Roman Jakobson: Selected Writings II*. Mouton, The Hague, The Netherlands 239–59

Jakobson R 1979 On the so-called vowel alliteration in Germanic verse. In: Rudy S (ed.) *Roman Jakobson: Selected Writings V*. Mouton, The Hague, The Netherlands 189–97

Jakobson R 1981a Linguistics and poetics. In: Rudy S (ed.) *Roman Jakobson: Selected Writings III*. Mouton, The Hague, The Netherlands 19–51

Jakobson R 1981b Grammatical parallelism and its Russian facet. In: Rudy S (ed.) *Robin Jackobson: Selected Writings III*. Mouton, The Hague, The Netherlands 98–135

Jespersen O 1993 In: Jespersen, O *Linguistica*. Levin and Munksgaard, Copenhagen

Kiparsky P 1970 Metrics and morphophonemics in the *Kalevala*. In: Freeman D C (ed.) *Linguistics and Literary Style*. Holt, Rinehart and Winston, New York 165–81

Kiparsky P 1977 The rhythmic structure of English verse. *Linguistic Inquiry* **8**: 189–247

Kiparsky P 1981 The role of linguistics in a theory of poetry. In: Freeman D C (ed.) *Essays in Modern Stylistics*. Methuen, London 9–23

Kiparsky P 1987 On theory and interpretation. In: Fabb N, Attridge D, Durant A, MacCabe C (eds.) *The Linguistics of Writing*. Methuen, New York 185–98

Kiparsky P 1989 Sprung rhythm. In: Kiparsky P, Youmans G (eds.) *Phonetics and Phonology, Vol. 1: Rhythm and Meter*. Academic Press, San Diego, CA

Liberman M, Prince A 1977 On stress and linguistic rhythm. *Linguistic Inquiry* **8**: 249–336

Malone J 1987 Muted euphony and consonant matching in Irish verse. *General LinguisticsI* **27**: 133–43

Saussure F 1983 *Course in General Linguistics*. Duckworth, London

Sperber D, Wilson D 1986 *Relevance*. Blackwell, Oxford, UK

Wimsatt W K 1954 One relation of rhyme to reason. In: Wimsatt W K (ed.) *The Verbal Icon*. Kentucky University Press, Lexington, KY

Youmans G Milton's meter. In: Kiparsky P, Youmans G (eds.) *Phonetics and Phonology, Vol. 1: Rhythm and Meter*. Academic Press, San Diego, CA

Zwicky A 1976 Well, this rock and roll has got to stop. Junior's head is hard as a rock. *Chicago Linguistic Society* **12**: 676–97

K. Hanson

Language and Social Class

Linguists have known for some time that differences in language are tied to social class. Ross (1954) suggested that certain lexical and phonological differences in

English could be classified as U (upper class) or non-U (lower class), e.g., *serviette* (non-U) vs. *table-napkin* (U), one of the best known of all linguistic class-indicators of England at the time.

Similarly, in the United States, some surveys of regional dialect recognized the importance of social status in geographical variation, and distinguished three categories of subjects based on the fieldworker's classification: Type I—little formal education, little reading and restricted social contacts; Type II—better formal education (usually high school) and/or wider reading and social contacts; and Type III—superior education (usually college), cultured background, wide reading and/or extensive social contacts. These types correspond roughly to social status.

Until the 1960s, however, most studies of variability were concerned primarily with regional variation or dialectology, following a tradition established in the nineteenth century. These studies concentrated their efforts on documenting the rural dialects which it was believed would soon disappear. Only during the latter half of the twentieth century would the concern for status-based differences in language become a primary rather than a secondary focus, when sociolinguists turned their attention to the language of cities, where an increasing proportion of the world's population lives in modern times. The rise of urbanization is connected with an increase in social stratification reflected in linguistic variation.

1. *Beginnings of Sociolinguistics*

Research focusing on social dialects is sometimes referred to as social dialectology, and occupies a central place in quantitative sociolinguistic research on urban speech varieties, beginning with William Labov's (1966) work in New York City. He was the first to introduce a systematic methodology for investigating social dialects and the first large-scale sociolinguistic survey of an urban community. Unlike previous dialectological studies, which generally chose one person (usually an older male) as representative of a particular area, this survey was based on tape-recorded interviews with 103 people who had been chosen by random sample as being representative of the various social classes, ages, ethnic groups, etc., to be found in New York City. This approach solved the problem of how any one person's speech could be thought of as representing a large urban area.

Previous investigations had concluded that the speech of New Yorkers appeared to vary in a random and unpredictable manner. Sometimes they pronounced the names *Ian* and *Ann* alike and sometimes they pronounced post-vocalic /r/ (i.e., *r* following a vowel) in words such as *car*, while at other times they did not. This fluctuation was termed 'free variation' because there did not seem to be any explanation for it. Labov's study and subsequent ones modeled after it, however, showed that when such free variation in the speech of and between individuals was viewed against the background of the community as a whole, it was not free, but rather conditioned by social factors such as social class, age, sex, and style in predictable ways. Thus, while idiolects (or the speech of individuals) considered in isolation might seem random, the speech community as a whole behaved regularly. Using these methods, one could predict that a person of a particular social class, age, sex, etc., would pronounce post-vocalic /r/ a certain percent of the time in certain situations.

Through the introduction of these new methods for investigating social dialects by correlating sociolinguistic variables with social factors, sociolinguists have been able to build up a comprehensive picture of social dialect differentiation in the United States and Britain in particular, and other places, where these studies have since been replicated.

2. *Methods*

In order to demonstrate a regular relationship between social and linguistic factors, we have to be able to measure them in a reliable way. Of the principal social dimensions sociolinguists have been concerned with (e.g., social class, age, sex, style, and network), social class has probably been the most researched. Patterns of social class differentiation are often assumed to be fundamental and other so-called sociolinguistic patterns of variation, e.g., stylistic and gender variation, are regarded as derivative of them (Labov 1972).

Many sociolinguistic studies have followed the generally accepted sociological practice of relying on indicators of social status, such as education, occupation, income, etc., grouped individuals into social classes on the basis of these factors, and then looked to see how certain linguistic features were used by each group. The method used in New York City to study the linguistic features was to select items which could be easily quantified; in particular, phonological variables such as post-vocalic /r/, which was either present or absent. This was one of the first features to be studied in detail by sociolinguists.

3. *Results*

Varieties of English can be divided into two groups with respect to their treatment of this variable: those that are *r*-pronouncing (rhotic) and those that are not *r*-pronouncing (non-rhotic). Today in Britain, accents that have lost postvocalic /r/ as a result of linguistic change generally have more prestige than those, like Scottish English, that preserve it. In many parts of the United States the reverse is true, although this has not always been the case. Table 1 compares the pro-

Table 1
Percent of postvocalic /r/s pronounced

New York City	Reading	Social class
32	0	Upper-middle class
20	28	Lower-middle class
12	44	Upper-working class
0	49	Lower-working class

Table 2
Percent of nonstandard forms in Norwich

Social-class	(ing)	(t)	(h)
Middle-middle class	31	41	6
Lower-middle class	42	62	14
Upper-working class	87	89	40
Middle-working class	95	92	59
Lower-working class	100	94	61

Table 3
Incidence of glottal stops in all environments compared to medial position

Class	I	IIa	IIb	III	Total
All environments	48.4	72.9	84.3	91.7	74.3
Medial position	0	7.2	42.5	68.8	29.6

Table 4
Verbs without -s in Detroit and Norwich

Class	Detroit	Norwich
Upper-middle class	1	0
Lower-middle class	10	2
Upper-working class	57	70
Middle-working class	–	87
Lower-working class	71	97

nounciation of postvocalic /r/ in New York City with that of Reading, England. The results show that in New York City the lower one's social status, as measured in terms of factors such as occupation, education, and income, the fewer postvocalic /r/s one uses, while in Reading the reverse is true. Like many features investigated by sociolinguists, the pronounciation of postvocalic /r/ shows a geographically as well as socially significant distribution.

Just as the diffusion of linguistic features may be halted by natural geographical barriers, it may also be impeded by social class stratification. Similarly, the boundaries between social dialects tend for the most part not to be absolute. The pattern of variation for postvocalic /r/ shows 'fine stratification' or continuous variation along a linguistic dimension (in this case a phonetic one) as well as an extralinguistic one (in this case social class). The indices go up or down in relation to social class, and there are no sharp breaks between groups. A major finding of urban sociolinguistic work is that differences among social dialects are quantitative and not qualitative.

There are many other variables in English which show similar sociolinguistically significant distributions, such as those studied by Trudgill (1974) in Norwich in an urban dialect study modeled after the New York research. Table 2 shows the results for (ing), (t) and (h). The numbers show the percentage of nonstandard forms used by different class groups. The variable (ing) refers to alternation between alveolar /n/ and a velar nasal /ɔ/ in words with *-ing* endings such as *reading, singing*, etc. The lower a person's social status, the more likely he/she is to use a higher percentage of alveolar rather than velar nasal endings. This is often referred to popularly as 'dropping one's g's,' and is a well-known marker of social status over most of the English-speaking world.

The variable (h) refers to alternation between /h/ and lack of /h/ in words beginning with /h/ such as *heart, hand*, etc. Unlike received pronunciation (RP), most urban accents in England do not have initial /h/ or are variable in their usage of it. For these speakers who 'drop their h's,' *art* and *heart* are pronounced the same. Again, the lower a person's social status, the more likely he/she is to drop h's. Speakers in the north of England, Scotland and Ireland retain /h/, as do speakers of American English. The variable (t) refers to the use of glottal stops instead of /t/, as in words such as *bottle*, which are sometimes stereotypically spelled as *bot'le* to represent the glottalized pronounciation of the medial /t/. Most speakers of English glottalize final /t/ in words such as *pat*, and no social significance is attached to it. In many urban dialects of British English, however, glottal stops are more widely used, particularly by younger working class speakers in London, Glasgow, etc.

By comparing the results for the use of glottal stops in Norwich with those for (ing) and (h), some interesting conclusions can be drawn about the way language and social class are related in this English city. Looking first at frequency, even the middle class in Norwich uses glottal stops very frequently, i.e., almost 50 percent of the time, but this isn't true of (h). There is of course no reason to assume that every instance of variation in language will correlate with social structure in the same way or to the same extent. Most sociolinguistic variables have a complicated history. Some variables will serve to stratify the population more finely than others; and some cases of variation do not seem to correlate with any external variables, e.g. the variation between /i/ and /ε/ in the first vowel of *economic* is probably one such instance. Phonological variables tend to show fine stratification and there is more socially significant variation in the pronunciation of English vowels than in consonants.

Table 5
Social class, style and sex differentiation in (ing) in Norwich

		Style			
Class	Sex	Word list	Reading	Formal speech	Casual speech
Middle-middle	m	0	0	4	31
	f	0	0	0	0
Lower-middle	m	0	20	27	17
	f	0	0	3	67
Upper-working	m	0	18	81	95
	f	11	13	68	77
Middle-working	m	24	43	91	97
	f	20	46	81	88
Lower-working	m	66	100	100	100
	f	17	54	97	100

In the case of glottal stop usage, what is socially significant is how frequently a person uses glottal stops in particular linguistic and social contexts. The use of glottal stops is particularly socially stigmatized in medial position, e.g., *bottle, butter*. A hierarchy of linguistic environments can be set up which seems to apply to all speakers. Glottal stops are more likely to occur in the following environments.

Most frequent
- Word final + consonant — e.g., *that cat*
- Before syllabic nasal — e.g., *button*
- Word final + vowel — e.g., *that apple*
- Before syllabic /l/ — e.g., *bottle*

Least frequent
- Word medially — e.g., *butter*

Although all speakers are affected by the same internal constraints in the same way, they apply at different frequency levels, depending on social class membership and other external factors. Table 3 shows the incidence of glottal stops in relation to social class in Glasgow for all environments compared with that occurring only in medial position (Macaulay 1977). Class I is the highest and contains professional people, while Class III is the lowest and contains unskilled workers. The results show that glottal stops are the norm for this community (74.3 percent), if we look at all the environments. Even the highest social class uses glottal stops nearly half the time, and the lowest class, almost all the time. However, if we look at medial position, the highest social class uses no glottal stops in this environment, while the lowest class uses 68.8 percent.

Although grammatical variables have been less frequently studied than phonological ones, they have tended to show sharp stratification; that is, they display a large social barrier between the middle class and the working class. Table 4 shows the results of a study of a grammatical variable in Detroit (Wolfram 1974) and Norwich (Trudgill 1974). The variable concerns the use of nonstandard third person singular present tense verb forms without *-s*, e.g., *he go*. Only working class speakers use these forms with any great frequency and this is more so in Norwich than in Detroit. The gap between the middle and working class norms is also greater in Norwich than in Detroit, reflecting the greater social mobility of the American social system. There are also other varieties of British English, e.g., in parts of the north, south-west and south Wales, where the present tense paradigm is regularized in the opposite direction and all persons of the verb takes *-s*, i.e., *I goes, you goes, he goes*, etc.

4. *Social Class Differentiation in Relation to Other Sociolinguistic Patterns*

The intersection of social and stylistic continua is one of the most important findings of quantitative sociolinguistics: namely, if a feature occurs more frequently in working class speech, then it will occur more frequently in the informal speech of all speakers. There are also strong correlations between patterns of social stratification and gender, with a number of now classic findings emerging repeatedly. One of these sociolinguistic patterns is that women, regardless of other social characteristics such as class, age, etc., tend to use more standard forms than men.

Table 5 shows the results of Trudgill's (1974) study in Norwich of the variable (ing), i.e., alternation between alveolar /n/ and a velar nasal /ɔ/ in words with *-ing* endings such as *reading, singing*, etc. in relation to the variables of social class, style, and sex. The scores represent the percent of nonstandard forms used by men and women in each social group in four contextual styles, i.e., when reading a word list, reading a short text, formal speech, and casual speech.

Generally speaking, the use of nonstandard forms increases the less formal the style and the lower one's social status, with men's scores higher than women's. Although each class has different average scores in each style, all groups style-shift in the same direction

in their formal speech style; that is, in the direction of the standard language. This similar behavior can be taken as an indication of membership in a speech community sharing norms for social evaluation of the relative prestige of variables. All groups recognize the overt greater prestige of standard speech and shift towards it in more formal styles.

Summing up these sociolinguistic patterns involving social class, gender, and style, sociolinguists would reply to the question of who is likely to speak most nonstandardly in a community: working-class men speaking in casual conversation. Conversely, middle-class women speaking in more formal conversation are closest to the standard. In Table 5, for instance, we can see that middle-middle-class women in word-list style never use the nonstandard form, while lower working-class men use it all of the time. Note, however, that the differences between men and women are not equal throughout the social hierarchy. For this variable they are greatest in the lower middle and upper working class. Such patterns reveal basic linguistic fault lines in a community, and are indicative of the uneven spread of the standard and its associated prescriptive ideology in a speech community. Similar results have been found in other places, such as Sweden and the Netherlands.

5. *Social Class Differentiation and Language Change*

Because variability is a prerequisite for change, synchronic variation may represent a stage in long-term change. By examining the way in which variation is embedded into the social structure of a community, we can chart the spread of innovations just as dialect geographers mapped variation and change through geographical space.

Sociolinguists have distinguished between 'change from above' and 'change from below' to refer to the differing points of departure for the diffusion of linguistic innovations through the social hierarchy. Change from above is conscious change originating in more formal styles and in the upper end of the social hierarchy, and change from below is below the level of conscious awareness originating in the lower end of the social hierarchy. Gender is critical here too. Women, particularly in the lower middle class, lead in the introduction of new standard forms of many of the phonological variables studied in the United States, the UK and other industrialized societies such as Sweden, while men tend to lead in instances of change from below.

Where hypercorrection occurs, as with post-vocalic /r/ in New York City, the lower middle class shows the most radical style shifting, exceeding even the highest group in their use of the standard forms in the most formal style. The behavior of the lower middle class is governed by their recognition of an exterior standard of correctness and their insecurity about their own speech. They see the use of postvocalic /r/ as a prestige marker of the highest social group. In their attempt to adopt the norm of this group, they manifest their aspirations of upward social mobility, but they overshoot the mark. The clearest cases of hypercorrection occur when a feature is undergoing change in response to social pressure from above, i.e., a prestige norm used by the upper class. In New York City the new /r/ pronouncing norm is being imported into previously non-rhotic areas of the eastern United States. Hypercorrection by the lower middle class accelerates the introduction of this new norm. The variable (ing), on the other hand, has been a stable marker of social and stylistic variation for a very long time, and does not appear to be involved in change and hence does not display hypercorrection.

See also: Class: Social; Economic Status, Inheritance of: Education, Class, and Genetics; Gender, Class, Race, and Ethnicity, Social Construction of; Language and Ethnicity; Language and Social Inferences; Language and Society: Cultural Concerns; Language Policy: Linguistic Perspectives; Language Policy: Public Policy Perspectives; Sociolinguistics

Bibliography

Chambers J K 1995 *Sociolinguistic Theory. Linguistic Variation and its Social Significance*. Blackwell, Oxford, UK

Labov W 1966 *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, DC

Labov W 1972 *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, PA

Macaulay R K S 1977 *Language, Social Class, and Education. A Glasgow Study*. University of Edinburgh Press, Edinburgh, UK

Ross A S C 1954 Linguistic class indicators in present-day English. *Neuphilologische Mitteilungen* **16**: 171–85

Trudgill P 1974 *The Social Differentiation of English in Norwich*. Cambridge University Press, Cambridge, UK

Wolfram W 1974 *A Sociolinguistic Description of Detroit Negro Speech*. Center for Applied Linguistics, Washington, DC

S. Romaine

Language and Social Inferences

A social event can be represented linguistically in a great variety of different ways. For instance, we can talk about a specific act of economic transaction by using one of a number of interpersonal verbs (to buy, to sell, to pay, to purchase, etc.). The very same act can

be represented in terms of the motives that gave rise to the act (e.g., to like, to want, to desire). Referring to the qualities of the actors (untrustworthy, keen, greedy, stingy, generous, etc.) is yet another set of available linguistic options to represent the transaction. The choice of verbs and/or adjectives that is made about how a social event is represented influences the types of thematic features of the situation that are attended to.

Thematic features of an event refer to a variety of social inferences that are tacitly indicated to a reader or listener of such a communication. The choice of interpersonal terms (verbs and adjectives) in representing an event has been shown to lead to a variety of inferences about a social event. These are about who the implied causal agent for the event is, how long the event is implied to have lasted, inductive and deductive inferences about the qualities of the persons in the event, *inter alia*.

Interest in this field emerged in the 1960s and 1970s when social psychologists (e.g., Abelson and Kanouse 1966) and linguists (e.g., Garvey and Caramazza 1974) discovered that interpersonal verbs influence the inferences people make in an orderly fashion. It was in the 1980s that this field was more systematically charted. This started with the discovery that interpersonal verbs referring to visible actions—and consequently classified as verbs of action (e.g., to help, to buy, to cheat)—systematically imply that the subject in a sentence is the agent of the event or the person who caused the event (as in 'John cheats Edward'). In contrast, verbs of state (e.g., to love, to respect, to abhor)—which refer to unobservable interpersonal states—systematically convey the impression that the causal origin for the event is to be found in the sentence subject (e.g., 'John respects Edward'). Thus, the verb used in such sentence constructions can systematically direct the thematic focus of an event to either the sentence subject or object as the causal origin of the event (e.g., Brown and Fish 1983). Another illustration of this can be seen in the use of action and state verbs in question formulation. These verb types tacitly issue instructions to a respondent by conveying information about how the thematic agency of an answer should be structured. Thus given the very same event, an interviewer's choice of 'Why did John *trust Edward?*' (state verb) instead of 'Why did *John confide in* Edward?' (action verb) imposes a specific focus on the theme of the interviewee's answer. The question requests an answer that is primarily about the qualities that make Edward trustworthy, rather than John's motives and reasons for confiding in Edward (Semin 2000).

Implied causation and thematic focus are not the only systematic inferences that are mediated by interpersonal language. We communicate about a social event in order to pass on what we saw, heard, or believed. In talking about an event we make implicit choices about how to represent it in our communication. By means of these choices a communicator structures the linguistic representation of an event implicitly, albeit strategically, in order to convey a diversity of inferences to the recipient of the message. Generally, we do not consciously reflect upon our particular choice of verbs and adjectives when we talk about a particular event. Thus, we might refer to the same event by saying 'John pushed David,' 'John challenged David,' 'John hates David,' or 'John is aggressive,' to name just a few options. These choices vary from a term that is very concrete, and refers to a directly perceivable and contextually specific act (pushing), to a term that is most abstract, that maintains no references to any specific context, relationship, or situation and expresses a nonvisible quality in the form of an adjective (aggressive). 'Challenge' and 'hate' occupy intermediate positions. The former verb still refers to a concrete act without preserving the direct reference to the act. The latter refers to an unobservable psychological state that characterizes a specific relationship between two people. The concrete–abstract dimension on which interpersonal language can be systematically classified has been shown to carry a number of tacit, systematic, and general inferences (Semin and Fiedler 2001). It is important to note that while any of the four speech acts about the 'hostile' nature of an event is intended to impart information about what has occurred, the particular choices of interpersonal terms in formulating the speech act conveys additional tacit information that is not consciously accessed. This information can be about who caused an event, who was responsible for the event, whether or not the behavior was under the voluntary control of the agent, whether it was determined by the personality make-up of the person, or situationally caused, and so on. Thus, different properties of interpersonal language such as verb type or abstraction–concreteness lead to a variety of tacit social inferences. So far we have discussed the types of social inferences mediated by interpersonal language. Two issues arise. The first is about the context within which language and social inferences are relevant. The second issue concerns the psychological processes that lead to specific patterns of language use in a message and those processes that are induced by the tacit properties of a message.

The context issue becomes a comprehensible one when one inserts 'language and social inferences' into a communication framework. The simplest exemplification of a communication framework requires (a) *a transmitter* who produces a message about an event (e.g., a hostile act), (b) a medium upon which the *message* is mapped—language, and (c) *a recipient*. The transmitter of a message has plans about an intended action in the form of cognitions that are mapped upon language. The behavioral execution of cognition as intended action constitutes a speech act. The particular message that is structured by an intended action can thematically focus an event in such a way that A (A hit

B—action verb) rather than B (A hates B—state verb) is the causal locus for the event. Such a message is designed to structure the representation of the event such that the recipient of the message makes the appropriate social inferences (e.g., A engaged in behavior under voluntary control, A intended the act, A caused B to be harmed, etc.) Let us consider another example, to illustrate the point in a different context. A police investigator interviews a rape victim who is expected to be trustworthy or untrustworthy. In the former case, the interviewer is likely to pose questions that imply the agency of the perpetrator to the events that led to the incident (e.g., 'Did he dance with you?—implying 'Did he initiate the dancing?'). In the latter case, where the investigator expects the victim to be untrustworthy, the question may actually be like 'Did you dance with him,' implying that the victim was the initiator for the ensuing event.

Thus, one function of language is to give overt expression to cognitions as intended actions, by providing the *transmitter* with the means to formulate a verbal act. The second related function is to structure the specific way in which the *recipient* of the message represents the event. This representation involves also the social inferences that the transmitter intended when structuring the message. Thus, a communication framework helps locate the functions that interpersonal language and associated social inferences have in a transmitter and recipient context.

So far we have covered the following:

(a) the inferential properties of language and

(b) a communication framework that specifies

(i) how a transmitter uses language to give public shape to intended actions, and

(ii) how language in the form of a message shapes the representation (and inferences) of a recipient.

The second issue involves the nature of the psychological processes. This issue has two facets. First, what are the psychological processes by means of which the transmitter produces messages with a specific set of inference inducing patterns rather than others? Second, what types of psychological processes are induced in the recipient by the specific linguistic properties of a message?

We begin by illustrating the types of psychological processes involved in message production and how a closer examination of the linguistic properties of a message facilitate highlighting cognitive and motivational processes underlying message production. One specific example is to be found in the research on the transmission and maintenance of stereotypes. This is best represented by the research on the 'linguistic intergroup bias' (Maass 1999). This bias refers to the tendency to describe positive behaviors by members of the ingroup and negative outgroup behaviors by the use of abstract language. In contrast, negative ingroup behaviors and positive outgroup behaviors are communicated by the use of concrete language. By describing positive ingroup and negative outgroup behaviors in abstract terms (e.g., adjectives/traits such as 'friendly,' 'hostile'), the communicator implies that the behaviors in question stem from enduring qualities that are unlikely to change and are likely to be repeated in the future. In contrast, using concrete language (e.g., verbs of action such as 'to help,' 'to hinder') suggests situationally specific behaviors that are not likely to recur at a future point in time. The bias has been shown to be driven by motivational and cognitive factors. Thus, the differential abstract versus concrete language use is motivated in order to preserve a positive ingroup image and derogate the outgroup and has been shown to be much stronger when the ingroup identity is threatened. Moreover, it has also been shown that stereotype consistent behaviors are communicated at a more abstract level than stereotype inconsistent behaviors. This expectancy driven bias constitutes the more general case in that it suggests that all other things being equal (i.e., motivational factors) the abstractness–concreteness of a message is driven by expectancies. Basically, the 'linguistic intergroup bias' is responsible for the transmission and maintenance of stereotypes and prejudice (Maass 1999). The 'linguistic intergroup bias' is one illustration of how transmitter-specific cognitive and motivational processes shape the linguistic properties of a message. Of course, by the reverse logic it is also possible to infer what specific psychological transmitter processes are at hand by analyzing the linguistic properties of a message.

We now turn to recipient-specific processes. Here the focus is on controlling the linguistic properties of a message and examining its psychological implications. First, messages controlling for verb type or abstractness concreteness have systematically been shown to induce the social inferences described earlier. Thus, recipients of messages that arise through the linguistic intergroup bias do interpret the message as intended—confirming that the linguistic properties of the message are responsible for the transmission of stereotypes. Further research suggests that if an event is represented by the use of concrete language (e.g., verbs of action) rather than abstract terms (e.g., verbs of state or adjectives), then it is more vividly represented, processed more systematically, and recalled better and with fewer intrusions. In contrast, representing an event abstractly leads to more heuristic processing, poorer recall, and more intrusions in memory. Moreover, if people are asked to recall an event by using a concrete term ('When was the last occasion that you *helped* somebody?') as a retrieval cue, then they recall events that are significantly more recent than when they are asked to recall an event with an abstract retrieval cue ('When was the last occasion when you were *helpful?*'). Thus, properties of interpersonal language also have recipient-specific cognitive processing implications in terms of recall, or representation, aside from the specific social inferences that they mediate.

See also: Communication and Social Psychology; Inferences in Discourse, Psychology of; Inferences, Psychology of; Language and Social Class; Language and Society: Cultural Concerns; Stereotypes, Social Psychology of

Bibliography

Abelson R P, Kanouse E E 1966 Subjective acceptance of verbal generalizations. In: Feldman S (ed.) *Cognitive Consistency: Motivational Antecedents and Behavioral Consequences*. Academic Press, New York, pp. 171–97

Brown R, Fish D 1983 The psychological causality implicit in language. *Cognition* **14**: 237–73

Garvey C, Caramazza A 1974 Implicit causality in verbs. *Linguistic Inquiry* **5**: 459–64

Maass A 1999 Linguistic intergroup bias: Stereotype perpetuation through language. *Advances in Experimental Social Psychology* **31**: 79–121

Semin G R 2000 Language as a cognitive and behavioral structuring resource: Question–answer exchanges. In: Stroebe W, Hewstone M (eds.) *European Review of Social Psychology*. Wiley, Chichester, UK, pp. 75–104

Semin G R, Fiedler K 2001 *The Linguistic Category Model*. Psychology Press, Philadelphia, PA

G. R. Semin

Language and Society: Cultural Concerns

1. Introduction

Every language provides its speakers with a unique, intricate and constantly changing medium of expression. The structure of any language embodies a myriad of prepackaged meanings, a large proportion of which are language-and-culture specific, in two respects: first, in not having exact counterparts in other languages of the world, and second, in reflecting, embodying and helping to perpetuate a particular social, cultural, and historical experience. The purpose of this article is to demonstrate this in relation to three linguistic domains—words and phraseology, grammar, and language in use—and to delineate some of the conceptual and methodological issues involved. A somewhat similar theme is pursued under the aegis of the Sapir–Whorf hypothesis (see *Sapir–Whorf Hypothesis*), which concerns the influence of language on world-view. This article, however, is concerned with more subtle and fine-grained effects.

These issues have been of pressing concern to language study from at least the time of Johann Herder and Wilhelm von Humboldt. They were carried into American linguistics and anthropology in the early twentieth century by Franz Boas and then Edward Sapir. Similar concerns continued to occupy anthropology, especially cognitive and symbolic anthropology, in the second half of the twentieth century, but in linguistics these concerns became marginalized with the ascendency of Chomskyan generative grammar; though to some extent the flame was kept alight by the ethnographic wing of sociolinguistics in the work of Dell Hymes and associates (cf. Bauman and Sherzer 1974; Gumperz and Hymes 1986). In the 1980s and 1990s, however, the dominance of generative grammar began to wane, leading to what one commentator dubbed 'the greening of linguistics,' with renewed interest in pragmatics, functionalism, anthropological linguistics, and cross-cultural semantics (cf. Foley 1997; Goddard 1998). In a parallel development, mainstream psychology became more interested in ethnopsychology and in the new field of cultural psychology (Shweder 1990, Schweder and Sullivan 1993). These developments set the stage for a revitalization of studies into the interconnections between language, culture, and society.

A complex of methodological problems connected with cross-cultural semantic analysis confronts this emergent interdisciplinary field. How can one ensure that any set of semantic descriptors is not biased by the language from which it is drawn? How to achieve the maximum resolution of complex meanings? How can the process of semantic analysis be freed from subjectivity and vagueness? According to the most well-developed model of cross-cultural semantic analysis, the 'natural semantic metalanguage' (NSM) approach, the most promising solution to these problems depends on identifying a viable set of universal semantic primes (or primitives), i.e., simple concepts which have linguistic exponents (as words or word-like elements) in all languages (cf. Wierzbicka 1996; Goddard and Wierzbicka 1994, in press). Cross-linguistic research indicates the existence of at least 60 semantic primes, including (using English exponents): I, YOU, SOMEONE, SOMETHING, PEOPLE, DO, HAPPEN, SAY, THINK, KNOW, WANT, GOOD, BAD, BIG, SMALL, THIS, ONE, TWO, OTHER, THE SAME, WHEN/TIME, WHERE/PLACE, BECAUSE, IF, CAN, NOT, LIKE. This restricted and universal vocabulary of simple meanings can be used as a kind of 'semantic bridge' in terms of which to articulate complex language-specific and culture-specific meanings in maximum detail. Because it is grounded in simple words of ordinary language, this approach makes hypotheses about subtleties of meaning accessible to the intuitions and judgments of native speakers, thus improving the verifiability of the analyses. The NSM approach is not universally accepted, however, and much valuable work in cross-cultural semantics is still conducted in essentially informal terms.

2. Words and Phraseology

Even between genetically related languages spoken by people of relatively similar cultures (such as the languages of Europe), there are sizable semantic differences in all aspects of the lexicon. Sapir (1949) was thus right when he said that 'vocabulary is a very

sensitive index of the culture of a people.' Most people have no difficulty in accepting the culture-specific nature of words for items of material culture and for social institutions, but many find it hard to accept the extent of cross-linguistic differences in terms for value concepts, emotions, speech acts, and social categories, to mention only a few examples.

Often it is possible to nominate certain highly salient and deeply culture-laden words in a language as 'key words' of that culture, i.e., words whose meanings can provide a key to understanding whole domains of cultural attitudes, values, and behaviors (Williams 1976; Wierzbicka 1997). For example, it can be argued that among terms for value concepts, English *freedom*, Japanese *wa* 'harmony,' and Malay *kesetiaan* 'loyalty,' 'faithfulness' have key word status. Among ethnopsychological concepts the same can be argued for English *mind*, Malay *hati* 'heart,' and Russian *duša* 'soul.' Single-word glosses are never adequate, however, and a reasonable understanding of the key words of any culture requires dedicated and disciplined semantic investigation. Cultures vary across time, as well as across space, so a close analysis of meaning shifts can be revealing of sociohistorical change. For example, it can be shown that the English word *friend* has weakened considerably in meaning since Shakespeare's time, when it designated an exceptional, intimate, and life-long bond of trust, affection, and mutual assistance. This shift can be seen as linked with a shift towards more flexible and broadly-based (if 'shallower') mode of social relationships.

Languages differ not only in their inventories of culture-specific words, but also in the number of words they have for speaking about a particular domain of meaning. A relatively high number of words in a single domain (lexical elaboration) is often reflective of sociocultural facts. For example, compared with most indigenous cultures, European languages have a very large stock of expressions to do with measuring and reckoning time (words such as *clock*, *calendar*, *date*, *second*, *minute*, *hour*, *week*, *Monday*, *Tuesday*, etc. *January*, *February*, etc.). At the same time, there are linguistic universals of time, such as WHEN, AFTER, and BEFORE (cf. Keesing 1994).

Two further ways in which vocabulary and phraseology can reflect sociocultural facts concern the innovation of new words and the borrowing of words from other languages. In general, new words and phrases are indexes of social change, for example, the expressions *housemate* and *online* reflect social and technological developments of mainstream English-speaking culture of the late-twentieth century. Linguistic borrowings index the extent and mode of cultural influence, as witness, for example, the massive borrowing of Chinese words into Japanese, of Latin and Greek words into English, or Sanskrit and Arabic loans into Malay. Some scholars argue that permeability to loans is itself an index of the social openness of a community.

3. *Grammar*

Proponents of linguistic relativity argue that inasmuch as speaking a language obliges individuals to continually attend to particular semantic and conceptual distinctions, this must to some extent 'shape' the habitual thinking of its speakers. Approaching the issue this way means focusing on obligatory grammatical rules and processes. On the other hand, it can be argued that optional choices in grammar, i.e., aspects over which speakers may exercise some choice and control, are more likely to reflect subtle sociocultural meanings. A familiar example is the contrast between so-called 'intimate' and 'formal' second-person pronouns in European languages, such as French *tu* vs. *vous*. Much more extensive is the honorific system of Japanese, whereby verbs carry suffixes which indicate degrees of respect felt by the speaker in relation to the individuals being spoken about. In Slavic languages, systems of expressive derivation reflect the central place of warmth and affection in Slavic cultures; in Polish, for example, a personal name like *Maria* can have as many as 10 different derivatives (*Marysia*, *Marysieńka*, *Maryśka*, *Marysiuchna*, and so on), each reflecting a slightly different emotional attitude and mood. It will be useful to look in some detail at two more examples, from Russian and from English.

3.1 '*Fatalism*' *in Russian grammar*

The Russian language has a large family of impersonal dative-infinitive constructions which refer to inexplicable things that happen to people against their will or irrespective of their will (Wierzbicka 1992). For example, the sentence *Ne budet tebe pasporta* 'There'll be no passport for you' involves a dative pronoun, a negated existential verb, and a noun in the genitive case. The sentence conveys a subtle message, not fully evident in the English translation. Essentially, a construction of this kind rules out the possibility of someone being able to obtain something beneficial and desired, implying at the same time that this is due to the fact that someone in a superior position does not want it to happen.

Another construction combines negation, an infinitive verb, and a human noun in the dative case, for example, *Ne vidat' tebe ėtix podarkov* 'You'll never see those presents.' This construction is not so common in everyday speech, but is highly characteristic of Russian folk literature and folk speech: *byliny* (folk epic), *plač* (lamentation), in folksong, in proverbs and sayings. It conveys the impression that a desired outcome can never eventuate due to some elusive and mysterious force or power that is, so to speak, 'above' humanity.

A third construction combines a human noun in dative case with a mental verb in the 3sg reflexive form, for example, *Segodnja mne vspomnilas' Praga-sady* 'Today I was reminded of Prague, of its gardens.' This

depicts a mental event simply happening inside a person inexplicably, and, in a sense, irresistibly. The most important Russian expression of this sort is the ubiquitous *xočetsja* lit. 'it wants itself to me,' which suggests a spontaneous and involuntary desire. In some ways this parallels English constructions like *It occurred to me* … or *It seemed to me* …, but whereas the pattern has a limited scope in English, in Russian it is both fully productive and extremely common.

In these and other constructions, Russian furnishes an excellent example of the grammatical realisation of a semantic theme; in this case, the theme of *sud'ba* 'fate,' of not being in control, of living in a world which is unknowable and which cannot be rationally controlled.

3.2 Personal Autonomy and the English wh-imperative

Compared with many languages and cultures, English makes only scant use of its imperative construction, e.g., a simple *Do this*!, whose core meaning is 'I want you to do this.' Much more frequent is the use of one of a brace of more elaborate locutions (so-called wh-imperatives), such as *Will you do this? Would you do this? Could you do this? Would you mind doing this?* and so on. These constructions can be seen as hybrids: in form they are questions, but since they convey the speaker's wish that the addressee do something, they have a similar function to imperatives.

It can be argued that the English wh-imperatives represent a linguistic reflection of the importance, in Anglo culture, of a cultural principle that can be labeled 'personal autonomy.' Roughly speaking, this calls upon people to respect the right of other people to make their own decisions about what they will and will not do. To use a bare on-the-record imperative could be seen to infringe this principle, hence the potentially confronting message that 'I want you to do something' is embedded into a more complex configuration which acknowledges the addressee's autonomy by inviting them to say whether or not they will comply. In form at least the addressee is presented with the option of refusing.

Similar examples of grammatical elaboration being driven or motivated by culture-specific pragmatics or cultural values could be given from many languages and cultures (cf. Enfield in press). Often such phenomena arise from the routinization of particular culture-specific patterns of usage.

4. Language in Use

4.1 Discourse Particles and Interjections

These elements exist in the border zone between grammar and discourse. While they are usually not obligatory from a strictly grammatical point of view, they are often an essential part of native-like fluency. Interjections stand as independent utterances, whereas discourse particles are integrated into the syntax of a larger utterance; but both express a speaker's here-and-now feelings, thoughts, wants, and reactions to what is happening or to what is being said. When their semantics are examined in close detail, both types of element tend to be highly indicative of sociocultural priorities, values, and attitudes.

The very common Japanese particle *-ne*, for example, conveys the message 'I think you would say the same,' consistent with the Japanese cultural value placed on harmony and consensus-seeking. English interjections and interjectional phrases such as *Well* and *on the other hand* reflect the Anglo emphasis on balance and tolerance of opposing views. A multiplicity of Polish discourse particles (such as *ależ*, *skądże* and *przecież*) and related interjectional phrases express disagreement, exasperation, and impatience, consistent with the Polish preference for spontaneous expression of emotion and lack of any inhibition on expressions of conflict or disagreement.

4.2 Linguistic Routines

These are fixed, formulaic utterances or sequences of utterances used in standardized communicative situations, for example, greetings, partings, thanks, excuses, condolences, compliments, jokes, curses, small-talk, and so on. They can range in size from a single word to lengthy interchanges, and usually occur very frequently. The overall meaning of a routine cannot be 'read off' from the literal meaning of the individual words involved, e.g., *How do you do?* is not a question about health.

In general, routines are highly culture-specific both in form and in the way they relate to the sociocultural context. To illustrate, one can compare some fixed expressions in English and Ewe (Ghana and Togo, West Africa). In many societies, when one realises that something good has happened to another person, it is usual to say something expressing one's own good feelings at the news. For example, in English one might say *Congratulations!* or *Well done!*; both expressions imply that the addressee is responsible to some extent for the happy event. In Ewe, on the other hand, appropriate things to say refer to *Máwú* 'God,' *tɔ́gbéwó* 'ancestors,' or *ŋú-wò núwó* 'beings,' expressing messages such as 'God is strong,' 'Ancestors are strong,' or 'Beings around you have worked.' Although one function is to register and display one's happiness at the good event, the Ewe formulae are not focused on the individuals concerned. This reflects the religious belief system which holds that everything that happens is ultimately the work of *Máwú* who may work in diverse ways through the ancestors or other spirits and divinities.

4.3 Speech Genres

Bakhtin (1986) defined speech genres as 'relatively stable and normative forms of the utterance' into which we 'pour' our everyday speech. The concept is a very broad one, subsuming examples which differ in size and complexity as widely as (to illustrate from English) *joke*, *toast*, *speech*, *letter*, *eulogy*, *submission*, and *lecture*. Bakhtin stressed that the repertoire of genres available to a speech community changes according to historical and sociocultural conditions.

This conveniently can be illustrated with the (now passé) Polish 'micro genre' of the *kawał* (plural *kawały*) which, roughly speaking, was a kind of conspiratorial joke. Most *kawały* were political, expressing national solidarity vis-à-vis foreign powers which have been so much a part of the Polish historical experience: the foreign partitioning powers in the nineteenth century, the Nazi occupation during World War II, the Soviet-imposed communist regime in postwar Poland. *Kawały* circulated widely, the anonymous creations of an oral culture. Like English *jokes*, they were intended to promote pleasant togetherness, i.e., to make the speaker and the addressee feel good together, but one did not value a *kawał* for its ingenuity, novelty, or sophistication (as one did, and does, *dowcipy* 'witty jokes'). More important was the feeling it gave of belonging to an in-group, the implication being 'I can say this you, but there are some people I couldn't say this to.' For example, in 1981 after martial law had been imposed in an effort to suppress the Solidarity movement, every new demonstration, strike, or protest was ascribed to 'Solidarity extremists.' A *kawał* from that time gave examples from a 'TV Dictionary' as follows: 'two Poles—an illegal gathering; three Poles—an illegal demonstration: 10 million Poles—a handful of extremists.' Clearly, though the Polish genre *kawał* may overlap in some respects with the English genre of a *joke*, it is not identical with its nearest English counterpart. Equally clearly, the precise nature of the Polish genre was a response to (and an index of) social conditions of Poland.

4.4 Speech Styles

Speech styles represent specialized 'ways of speaking,' that are recognized (and, usually, named) by the speech community as appropriate for use with certain categories of interlocutors or in certain situations. When examined in detail, both the stylistic conventions and social parameters of the speech styles can show much about sociocultural conditions. One common pattern can be illustrated from Yankunytjatjara (Central Australia), which has a special indirect speech style *tjalpawangkanyi* (*wangkanyi* 'talk') for use with people whose kin relationships call for constrained behavior, and a boisterous joking or teasing style (*inkatjingani*) for use with people whose kin relationships imply mutual acceptance and a lack of any power relationships (Goddard and Wierzbicka 1997). *Tjalpawangkanyi* has a distinctive vocal delivery (softer, slower, higher pitch) and an exaggerated rising intonation, as if to give the impression that the speaker is merely musing aloud. Direct references to the addressee, including imperatives and vocatives, are carefully avoided. Overt expressions of denial, refusal, or disagreement are also scrupulously avoided. There is liberal use of discourse particles and interjections that express uncertainty, hesitation, and minimization. Another striking feature, reported also of respectful speech styles in many other places, is generality of reference. Speakers avoid using any expression which unambiguously indicates a person, place, or thing, preferring vague locutions like 'that one (person)' and 'around here.'

The social meaning of *tjalpawangkanyi* refers to the Yankunytjatjara social emotion *kunta*. Usually glossed as 'shame,' 'embarrassment,' or 'respect,' this involves a sense of social difference and discomfort in the other person's presence, and the desire to avoid doing anything which might cause the other person to think anything bad about one. The strongest *kunta* is evoked by the respectful *umari* ('avoidance') relationship between a man and his parents-in-law, which absolutely prohibits personal contact of any kind. Less severe *kunta* is felt in the presence of the siblings or cousins of *umari*, and in other relationships where propriety is important, the very relationships for which *tjalpawangkanyi* is appropriate. In this culture, then, the indirect speech style can be seen both as a kind of partial avoidance and a way of giving voice to *kunta*. In other societies, indirect speech styles can serve quite different purposes.

Yankunytjatjara joking styles are at the other end of the spectrum to *tjalpawangkanyi*. They bend, flaunt, or even parody the normal conventions of interaction. All the linguistic forms excluded from *tjalpawangkanyi*, including imperatives, vocatives, contradiction, exclamations, and sensitive vocabulary items, flourish in this style. Participants delight in banter, mock insult, and make the most of any chance to defy, challenge, or demean each other. Sometimes there will be sexual innuendo and risqué comments that consciously play on kin-role expectations. In extreme versions such banter becomes blatantly obscene. Joking, uninhibited speech styles are reported from many societies around the world. Although at a general level one can say that they express 'solidarity,' 'intimacy,' or the like, the rich social meanings involved cannot really be summed up in a few words, and detailed study reveals subtle but important differences.

4.5 Discourse Styles

At an even broader level, entire speech communities can be characterized as having certain tacitly-shared

understandings about how it is appropriate to speak in particular, culturally-construed situations. In some societies it is normal for conversations to be loud, animated, and bristling with disagreement, while in others people prefer to avoid contention, to speak in well-considered phrases, and guard against exposure of their inner selves. In some societies, silence is felt to be awkward, while in others it is welcomed. In some societies it is considered very bad to speak when another person is talking, while in others this is an expected part of a co-conversationalist's work (Tannen 1981). Researchers in the ethnography of communication and in cross-cultural pragmatics are concerned to spell out the differing norms, values and attitudes (sometimes called cultural scripts) which underlie and explain these behavioral differences (Saville-Troike 1989).

Cultural norms about expressing emotions (what one feels) vary widely. For example, in mainstream Anglo societies the expression of negative feelings is relatively constrained (cf. 'cheerful' English routines such as *How are you? I am fine*). In Russian and Polish societies, by contrast, a greater value is placed on the immediate and spontaneous expression of emotions, 'bad feelings' as well as good ones. In yet other societies (for example, in Malay society and in Japanese society) the cultural ideal is to maintain a calm equanimity of expression, moderating the expression of good feelings as well as bad ones. These differences are reflected also in facial expressions and other aspects of nonverbal communication. Needless to say, these characterisations are only approximate. Many modulations and variations are found across cultures.

Even within a single language group there can be significant differences. For example, it can be argued that contemporary Anglo–American culture has developed some distinctive routines and expressions for expressing enthusiasm, optimism, and a 'positive attitude': *Wow! Great! How nice! That's fantastic! I had a terrific time! It was wonderful! Have a nice day!* Sociological studies of cultural attitudes toward emotions have also shown that Americans place an exceptional emphasis on enthusiasm.

Along with differences in emotional expression, other broad parameters of social variation in discourse style concern the expression of conflict or opposition between interlocutors' views (what one thinks) or wishes (what one wants). Japanese society, for example, is a society in which overt expression of differences and disagreements is discouraged, on account of the prevailing cultural value of *wa* (roughly, 'harmony'). For example, a well-known article by a Japanese author (Ueda 1974) is titled *Sixteen Ways to Avoid Saying 'No' in Japanese*. In Jewish culture (Schiffrin 1984, Blum-Kulka 1982), on the other hand, an overt and intense clash of views and evaluations is not only considered nonthreatening but is often actively welcomed.

Differences like these not only contribute powerfully to the texture of everyday social interaction and to the character of interpersonal relationships, but, arguably, exert an influence, through socialization, on the formation of personality types.

See also: Cognitive Anthropology; Communicative Competence: Linguistic Aspects; Cultural Evolution: Overview; Language and Social Class; Language and Social Inferences; Language Policy: Linguistic Perspectives; Language Policy: Public Policy Perspectives; Lexical Semantics; Pidgin and Creole Languages; Politeness and Language; Semantics; Sociolinguistics

Bibliography

Bakhtin M 1986 [trans. McGee V W] Emerson C, Holquist M (eds.) *Speech Genres and other Late Essays*. University of Texas Press, Austin, TX

Bauman R, Sherzer J (eds.) 1974 *Explorations in the Ethnography of Speaking*. Cambridge University Press, London

Blum-Kulka S 1982 Learning to say what you mean in a second language: A study of the speech act performanace of learners of Hebrew as a second language. *Applied Linguistics* **3**: 29–58

Enfield N J (ed.) in press *Ethnosyntax*. Oxford University Press, Oxford, UK

Foley W A 1997 *Anthropological Linguistics. An Introduction*. Blackwell, Oxford, UK

Goddard C 1998 *Semantic Analysis: A Practical Introduction*. Oxford University Press, Oxford, UK

Goddard C, Wierzbicka A (eds.) 1994 *Semantic and Lexical Universals. Theory and Empirical Findings*. Benjamins, Amsterdam

Goddard C, Wierzbicka A 1997 Discourse and culture. In: van Dijk T A (ed.) *Discourse as Social Interaction (Discourse Studies: A Multidisciplinary Introduction, Vol. 2)*. Sage, London

Goddard C, Wierzbicka A (eds.) in press *Meaning and Universal Grammar*. Benjamins, Amsterdam

Gumperz J J, Hymes D (eds.) 1986/1972 *Directions in Sociolinguistics. The Ethnography of Communication*. Blackwell, Oxford, UK

Keesing R M 1994 Radical cultural difference: Anthropology's myth? In: Pütz M (ed.) *Language Contact and Language Conflict*. Benjamins, Amsterdam

Sapir E 1949 *Selected Writings of Edward Sapir in Language, Culture and Personality*. Mandelbaum D (ed.) University of California Press, Berkeley, CA

Saville-Troike M 1989 *The Ethnography of Communication*, 2nd edn. Blackwell, Oxford, UK

Schiffrin D 1984 Jewish argument as sociability. *Language in Society* **13**: 311–35

Shweder R A 1990 Cultural psychology: What is it? In: Stigler J W, Shweder R A, Herdt G (eds.) *Cultural Psychology: Essays on Comparative Human Development*. Cambridge University Press, Cambridge, UK

Shweder R A, Sullivan M A 1993 Cultural psychology: Who needs it? *Annual Review of Psychology* **4**: 487–523

Tannen D 1981 New York Jewish conversational style. *International Journal of the Sociology of Language* **30**: 133–49

Ueda K 1974 Sixteen ways to avoid saying 'No' in Japanese. In: Condon J C, Saito M (eds.) *Intercultural Encounters With*

Japan: Communication, Contact and Conflict, 1st edn. Simul Press, Tokyo
Wierzbicka A 1992 *Semantics, Culture and Cognition*. Oxford University Press, Oxford, UK
Wierzbicka A 1996 *Semantics: Primes and Universals*. Oxford University Press, Oxford, UK
Wierzbicka A 1997 *Understanding Cultures through their Key Words: English, Russian, Polish, German, Japanese*. Oxford University Press, New York
Williams R 1976 *Keywords*. Oxford University Press, New York

C. Goddard and A. Wierzbicka

Language and Thought: The Modern Whorfian Hypothesis

Let us begin with something that means nothing. The opening lines of Lewis Carroll's nonsense poem *Jabberwocky*, are, in the original English:

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe.
All mimsy were the borogoves,
And the mome raths outgrabe.

And in French:

Il brilgue: les tôves lubricilleux
Se gyrent en vrilland dans le guave.
Enmîmes sont les gougesbosqueux
Et le mômerade horsgrave.

Finally in German:

Es brillig war. Die schlichten Toven
Wirrten und wimmelten in Waben.
Und aller-mümsige Burggoven
Die mohmen Räth' ausgraben.

The poem means nothing, other than the images it evokes. In a Pulitzer Prize winning series of essays on thought, Douglas Hofstader (1979, p. 372) maintained that each translation will evoke different thoughts in the user, because terms like the (according to Hofstader) earthy 'slithy' (slime, slither, etc.) evoke different associates than the associates evoked by the intellectual-sounding French *lubricilleux*. To the English ear, *Die schlichten Toven* sounds ominous. Is language controlling our thought? More particularly, do the forms of the approximately one thousand natural languages spoken today exert any influence upon the thoughts of their speakers?

Leonard Bloomfield (1933), a prominent linguist during the 1920s and 1930s, said that thinking was 'talking to oneself.' One way to interpret this statement is to say that thinking is an internal communication that takes place in the same language that external communication does. If so, the denotative properties of external language should influence internal thought: if you can't say it, you can't think about it. The anthropologist Edward Sapir (1921) took a stronger position. He maintained that people could only perceive those categories of the external world that are coded in their language, i.e., that if you cannot say it you cannot see it! Sapir's student, Benjamin Lee Whorf, adopted a rather more flexible view. For brevity, I shall refer to the position that language exerts a controlling influence on thought as the Whorfian hypothesis. (Some authors maintain that the term Sapir–Whorf hypothesis would be more accurate, but I believe that Whorf's amplification was great enough to justify the single reference.)

The Whorfian hypothesis is closely related to Bloomfield's notion of thinking as an internal conversation, so it is appropriate that we first look at this idea. Bloomfield's picture of the relation between language and thought was consistent with behaviorist theories that dominated psychology in the 1920–50 period. When the behaviorists spoke about thinking at all they spoke about the manipulation of response-produced internal stimuli, i.e., subvocal words. The strong position that all thought is internal language cannot be maintained, for there is ample evidence that people can think by manipulating internal visual images, and perhaps other sensory images as well. It has also been denied that the forms of the external language have any pervasive influence on thought at all.

One of the major theses of the 'cognitive revolution' that swept over psychology and linguistics in the 1950s was the contention that language and thought are independent. This argument derives from the seminal work of Noam Chomsky (1957, 1963) on the role of generative grammar in language. The clearest exposition of the language–thought distinction may be in the writings of Jerry Fodor (1975). Following Chomsky, Fodor argued that internal thought takes place in a special language, 'mentalese,' that is roughly equivalent to the predicate calculus. Expressions in mentalese are then translated into expressions in external languages according to the rules of the particular external language involved. An important assumption of the Fodor–Chomsky approach is that any expression in mentalese can be expressed by at least one (and usually many) expressions in any natural language. Therefore, any expression in language is translatable into an expression in any other language. Because the thoughts behind these translations occur in mentalese, the form of the external expression plays little role in thinking. This will be called the intertranslatability hypothesis. It will play a crucial role in the argument to be presented here.

The intertranslatability hypothesis is a statement about the competence of two languages to express each other's ideas. It says nothing about the difficulty of language-specific information-processing actions

required to go from the mentalese expression to the external statement in the two languages. The distinction is important.

To begin the current discussion, let us consider how language is used. One way is as a communicative act: a speaker (or writer) tries to create a thought in another person. Many studies have shown that it is possible to overwhelm the information-processing capacities of a receiver by presenting linguistic messages that are too complex to understand. Therefore, language contains devices to call attention to different aspects of the message to be communicated. For instance, attention can be focused on a topic by the device of first mention. The sentences

(a) John loved Mary.

(b) Mary was loved by John.

say the same thing, in one sense, but in (a) John is the focus of attention, while (b) is about Mary. Similarly, the provision of undisputed facts can shape meaning in subtle ways. Consider:

(c) President Eisenhower, the former five-star general, spoke to his cabinet.

This suggests that President Eisenhower did not conduct cabinet meetings in the folksy way that Abraham Lincoln is said to have done.

Whorf himself cited examples like (c) to show that the categories available in a language influence thoughts about the world. For instance, early in his professional life, Whorf worked as a safety engineer. He observed people smoking around gasoline drums that were labeled empty, disregarding the fact that a drum that has been emptied of liquid gasoline may contain highly volatile fumes. Whorf argued that the semantic association of the English term 'empty' led workers to the incorrect interpretation that the drums were inert. This is very close to Hofstader's observation that the meaningless terms 'slithy' and 'lubricilleux' will evoke different associates, and thus a different interpretation of Lewis Carroll's nonsense poem.

All languages use devices like this, and there is no doubt that thinking can be framed by subtleties of language. What is more questionable, and to the psychologist more interesting, is how the language available to a perceiver can be used to influence the way that perceivers can describe a situation to themselves.

At this time we return to Sapir's extreme position that a person literally cannot perceive a physical difference unless the person has a word category that depends on it. This proposition was tested, and apparently falsified by Eleanor Heider (Rosch) (Heider and Oliver 1972). Rosch studied color perception in the Dani, an aboriginal people on the island of New Guinea, in what is now the Indonesian province West Irian. The Dani language has only two terms that are roughly equivalent to 'dark' and 'light.' However, Rosch found that color discrimination in the Dani was the same as the color discrimination of English speakers. Rosch's work was widely cited in textbooks as conclusive evidence against Sapir's position that language can influence perception. Then, almost 30 years later, Roberson et al. (2000) conducted a follow-up study that raises questions about the earlier work. Roberson et al. used somewhat improved experimental techniques in a study of color perception in the Berinmo, a Papua New Guinea group whose language contains only five basic color terms. The Berinmo were most sensitive to color changes at the borders of the color categories in their language, whereas English speakers showed maximal color discrimination at the borders in their language. For instance, English speakers were maximally sensitive to colors that did or did not cross the blue–green boundary line, while Berinmo speakers were maximally sensitive to crossing the *nol–wor* boundary, which lies within the range of the English term green.

Another aspect of color naming has often been overlooked. Naming can affect memory even though it does not affect perception. Suppose that an English speaker is induced to call a particular object whose color is in the blue–green range either 'blue' or 'green.' Subsequently, the speaker's memory for the object's color will be distorted toward the prototypical hue of the name used to describe the object. This phenomenon can be thought of as a specific case of a much more general phenomenon. Examples of a category are remembered as more prototypical in appearance than they actually are. The general phenomenon has been demonstrated for categories as disparate as the position of a point in a clearly marked region of a page and the personality characteristics of people assigned to social roles, such as 'professor' or 'neo-Nazi.' Because our thought is heavily dependent on memory, the categories that language provides are important.

The decisions required to construct a linguistic statement can focus attention on selective aspects of a situation. This argument has been developed by Daniel Slobin (1996), who has coined the term 'thinking in speaking.' In order to describe a situation to ourselves or to another we have to decide what linguistic constructions (morphemes, syntactical forms, etc.) to use. The decisions required to choose between linguistic constructions focus attention on certain aspects of a situation. If it can be shown that different natural languages force a person's attention toward different aspects of a situation then we would have evidence for a Whorfian effect.

Whorf himself thought that he had observed such an effect in his studies of the way that the Hopi, a Native American people (and one of the larger tribal groups still extant) thought about time. According to Whorf, Hopi verbs stress cyclic stages (e.g., growth, decay). This sort of linguistic frame is used to discuss time. By contrast, English verb forms specify discrete states; the past, present, and future. Whorf believed that as a result, the Hopi had a different view of time than did speakers of 'standard European languages.'

Whorf's fieldwork with the Hopi has been criticized on the grounds that he did not 'really' understand Hopi. This claim is almost impossible to evaluate. Whorf did his fieldwork in the 1920s. The modern Hopi are much more integrated into the standard US culture than they were in 1920, and few speak the language as a first language. Therefore, we cannot interview the current Hopi and, with confidence, either object to or confirm Whorf's fieldwork. However, a similar claim has been made for the contrast between two of the most widely spoken languages in the modern world; Spanish and English. In fact, because of a unique feature, the claim is *for* Spanish speakers against at least all romance and Germanic languages.

Spanish has two forms of the verb 'to be,' *ser* and *estar*. The *ser* form is generally used to indicate permanent states and properties, while *estar* is used for continuing actions or temporary states. Two examples are:

(d) *Soy un hombre* = I am a man.

(e) *Estoy en mi oficina* = I am in my office.

Sera and her colleagues (Sera et al. 1997) have found that the use of *ser* or *estar* assists Spanish-speaking children in distinguishing between permanent and temporary properties of objects, in a manner that is unavailable to English-speaking children.

The use of classifier systems provides a second example of how 'thinking in speaking' can produce a Whorfian effect. Some of the languages distributed around the Pacific Rim require the use of classifiers before certain types of nouns. For example, Japanese requires the use of the term *bo* before things that are long and narrow, as in:

(f) I have three bo pencils.

In this example, shape classification, the property upon which classification is based, is perceivable. This is not always the case. Chinese uses a classifier for 'things that can be grasped by the hand' (Zhang and Schmitt 1998). Dyribll, an Australian aboriginal language, has a classifier for things derived from certain myths. It is applied, among other things, to 'women, fire, and dangerous things' (Lakoff 1987).

Languages that use classifier systems force the speaker to attend to certain features of objects. As a result, speakers of languages that use classifiers give higher similarity ratings to pairs of objects that share a classifier than do people whose language does not require classifiers. For example, Chinese speakers in Hong Kong, whose language classifies door keys, rulers, and pliers as 'things that can be grasped by the hand,' rate these objects as being more similar to each other than do English speakers, who do not use classifiers (Zhang and Schmitt 1998).

'Thinking in speaking' can provide effects if the development of appropriate linguistic forms calls attention to a nonlinguistic relation that is important in the world. Miller and his colleagues (Miller et al. 1995) report a more subtle effect involving number names. In every language, digit names up to the base term are arbitrary. That is, there is no reason other than history that 1 = one, 2 = two and so on up until 10. By contrast, the formation of English number names from twenty (2×10) onwards incorporate a mathematical relation; twenty-one = $20 + 1$, and even three hundred and forty-five = $300 + 40 + 5$. The number names for 11–19 are irregular in two ways. 'Eleven' and 'twelve' are arbitrary names, and a special regularity, thirteen = $3 + 10$, etc. is used instead of stating the most significant digits first, as is done for integers greater than twenty. Spanish is irregular also, although in a slightly different way. The integers from 11 to 15 have special names (*once*, *doce*, *trece*, *catorce*, *quince*), followed by the regular form with the most significant digits first (*dieciséis*, *diecisiete* = $10 + 6$, $10 + 7$). By contrast, in Mandarin Chinese, the spoken number names for all digits follow a regular convention. The Chinese equivalents of the digits eleven and twelve are $10 + 1$, $10 + 2$, as if English speakers said 'teenone, teentwo' or Spaniards said 'dieciuno, deicidos.'

Miller et al. observed that when children in English- and Spanish-speaking countries learn elementary arithmetic they have an unusual amount of trouble with operations involving numbers whose names have special terms, 11–12 in English and 11–15 in Spanish. Chinese-speaking first and second graders have no such trouble. Miller attributed this to the fact that irregular number names failed to reinforce numerical relations in the way that regular number names do.

Miller's research can be looked upon as a case in which different languages vary in the facility with which they express facts about the external world, but the facts can still be expressed. Research on terms for spatial relations provides an example of a situation in which a language may not be able to express certain facts. Most languages have words for expressing relationships in both an egocentric frame of references (right, left in English) or in an abstract geographic frame of reference (north, south, etc.). However, there are some languages that only have terms for one of these frames. Speakers of languages that are restricted to geographic frames interpret 'looks the same' to mean that two views of objects maintain their geographic orientation, while speakers of languages that have egocentric terms interpret 'looks the same' to mean that egocentric orientations are retained in two views (Levinson 1996). On the other hand, if a language only has terms for egocentrically defined relations it can be difficult to describe spatial layouts to people who have not observed them (Hunt and Agnoli 1991).

Hunt and Agnoli (1991) offered another argument for the Whorfian hypothesis. They pointed out that working (short-term) memory capacity is a major bottleneck in speech comprehension. Furthermore, the bottleneck seems to depend on the number of word forms that are to be held in short-term memory rather than the complexity of the concepts to which the

words refer. Therefore, specialized vocabularies evolve to provide a convenient way to let people talk about concepts of interest. Borrowing an argument form Whorf, Hunt and Agnoli suggested that having such specialized vocabularies facilitates thinking about theme. To a computer specialist the statement:

(g) The hacker put a cookie on the client's machine.

is certainly more compact than the equivalent in more standard English:

(h) The person attempting to break the security of a computer system for nefarious purposes put a program that surreptitiously records and reports the activity of a computing machine on the machine of a person who is using a desktop computer to gain entry to the Internet via another machine, called a server.

There has been a running argument between supporters and opponents of the Whorfian hypothesis over whether or not the existence of specialized vocabularies such as these constitutes a facilitation of thought by language. Whorf himself believed it did, arguing that Eskimos (Inuit) were better able to think about snow than English speakers. (Whorf overestimated the richness of the Inuit vocabulary for snow, but they do make distinctions that English does not, so his point is still valid.) Hunt and Agnoli do the same, arguing that words are used to package ideas, as illustrated by examples (g) and (h). As these examples show, it is possible to translate computer jargon into normal English, but the result is a pragmatically incomprehensible sentence.

Granted that examples like this can be made up, do they occur? The answer is that they can, and in situations in which the topic may be quite serious. In the late 1999–early 2000 period, a serious political dispute broke out between the government of Taiwan, which calls itself the Republic of China (ROC) and the mainland government, which refers to itself as the People's Republic of China (PRC). The two had been separate and hostile for over 50 years, but in the 1990s fairly extensive commercial relations developed between Taiwan and the mainland. Taiwanese (ROC) politicians said that, therefore, it was time to begin formal relations on a state-to-state basis. PRC officials reacted in fury, and literally threatened war if Taiwan attempted to establish two Chinese nations.

According to *New York Times* reports of the incident, part of the problem was that the Mandarin Chinese language uses the same word for 'state = formal machinery of government' and 'nation = unitary cultural and social entity.' On a de facto basis, PRC officials had been dealing with the ROC as a separate state for years. Unacknowledged contacts already existed between the formal machineries of the two governments. On the other hand, PRC leaders have an official and deeply held claim to be the only legitimate Chinese nation. To them the ROC is the government of a temporarily renegade province. Therefore, any claim to separate nationhood was, in the PRC ideology, a cause for war. ROC officials then claimed that they had only proposed government-to-government, rather than nation-to-nation, discussions. In order to expand on this claim, some officials resorted to English, even when speaking to other Chinese speakers.

The anti-Whorfian psychologist has two rejoinders to examples like these. One is to claim that these examples refer to the communication of ideas rather than the formation of them. This is consistent with the notion that 'basic thinking' takes place in an alinguistic, unobservable mentalese. Bouchard (1997) has remarked, in quite another context, that you can always explain an observation by appealing to unmeasured variables. The remark applies here. The notion of mentalese has a great deal of appeal because it fits with certain notions about generative models of language. However, it is not clear whether or not mentalese could be replaced with a reduced form of natural language terms, permitting thought directly in a speaker's language. And it is even less clear how this question could be investigated objectively.

The other anti-Whorfian rejoinder is that all these examples show is that language adjusts to cultural demands. Words and their related concepts are easily added to the lexicon. Modern English defines 'hacker' in a different way than it was originally used in computer science. More dramatically, a gay bachelor of the 1920s would not be gay in the sense the word is used in 2000. The anti-Whorfian argument is that cultural demands force language to adjust to cultures, rather than letting language restrict cultural change. Therefore, any restriction imposed on thought by lack of lexical entries for concepts like hacker or state versus nation is only temporary and probably not very important (Pinker 1994).

This objection is a value judgment rather than a hypothesis about facts, so it cannot be supported or refuted by scientific argument. It also brings us to another important point. We have been discussing specific points at which the form of a natural language may influence the thinking of its speakers. There is a 'romantic' version of the Whorfian hypothesis that claims that language may reinforce and/or restrict certain very broad cultural ways of thinking. This assertion is based on the assumption that there are definable cross-cultural ways of thinking, itself an assumption that requires analysis.

No one doubts that different cultures vary in their beliefs and value systems. This is part of the definition of culture. At the same time, modern psychologists and neuroscientists are confident that the content-free aspects of information processing, such as the distinction between short-term and long-term memory, work in the same manner for all normal human beings. If the differences between languages are regarded as a matter of setting parameters to govern the operation of a species-general universal grammar, then it is unlikely that the parameter setting would determine behavior outside of language. This is the approach

taken by linguists and psycholinguists working in the generative grammar tradition. On the other hand, if both linguistic abilities and cultural knowledge are due to general learning experiences, mutual interactions are more likely.

Some popular beliefs appear to acknowledge a belief in the romantic Whorfian hypothesis. French is said to be the language of lovers, and German the language of soldiers and engineers. In the extreme, this is silly, for differences between natural languages certainly do not preclude broad categories of activity or belief. Frenchmen built the Suez Canal, and the German romantic song 'Lili Marlene' was hummed by homesick soldiers on two sides in two wars.

What we need is more precise evidence of a linguistic effect. Let us take a fairly frequent accusation against US society. European commentators have remarked that contemporary US society mixes superficial personal familiarity with an almost cold treatment of close acquaintances. It is also true that English, unlike many other languages, uses the same form of address, 'you,' for formal meetings and for conversations between close friends. We have no objective measure showing that French people or Spaniards, whose language does include different forms of personal address, are in fact more formal in business and more intimate with their friends than are English speakers.

The closest that we can get to scientific evaluation of romantic Whorfianism is to look at correlations between definable properties of a language and global impressions of a culture. For instance, A. Wierzbicka, a Polish expatriate to Australia, has written extensively on the contrast between the Polish and English languages and what she sees as the relative emphasis on group solidarity and intimacy in Polish society and the personal speaking autonomy in the English-speaking societies. One of her favorite examples is the English convention of using indirect requests, as in

(i) Could you close the window?

This would normally be taken as a request for action rather than a request for information. Wierzbicka (1991) claims that Polish speakers prefer the more direct

(j) Close the window.

The problem is that we do not have any objective definition of 'emphasis on personal autonomy.' The problem is compounded by the fact that different subgroups in a large society may use their language in different ways. They may also vary in general cultural attitudes. Several of my Polish-speaking acquaintances dispute Wierzbicka's assertion about Polish, saying they would normally use a form like (i) rather than (j). Similarly, I, an English speaker, regard both (i) and (j) as appropriate uses of English, depending on the circumstances. I also suspect that the relative frequencies of direct and indirect requests vary a great deal across social groups and situations within English-speaking cultures. Besides, even if you could establish a correlation between language uses and situations, the causal relations would remain unknown. Does the culture adjust to the language or the language to the culture?

My point is not to dispute Wierzbicka's particular assertion. I am more concerned with the general point. Scientific hypotheses about the relationship between language and thought have to point to specific features of both language and cognition. Without adequate operational definitions we cannot investigate general assertions about relationships between language and thought.

Where, then, are we with respect to the Whorfian hypotheses? Linguistic effects on thought have been shown in specific domains, such as spatial reasoning, color perception, and memory, and the control of attention. The latter effect is particularly important, because the deployment of attention determines what goes into memory, and what goes into memory very largely determines how we think about things. The effects are generally rather small, but they occur every day a speaker uses language. Whether or not language exerts an important control on thought depends upon how one defines 'important.'

Grand romantic statements about language controlling, or at least heavily influencing, thought are harder to prove or disprove. This is at least as much because of vagueness in specifying what objective acts of thought are to be controlled as it is because of vagueness in specifying what aspects of language do the controlling.

I do not think that this state of affairs would have surprised Whorf. In his own work he was very much a student of the particular. The broader notions of what I have called romantic Whorfism come largely from his interpreters. The Whorfian hypothesis is not a proposition to be verified or disproven. It is a scorecard, telling us the extent to which language and thought are entwined.

See also: Bloomfield, Léonard (1887–1949); Color Terms, Linguistics of; Psycholinguistics: Overview; Sapir, Edward (1884–1939); Sapir–Whorf Hypothesis; Space: Linguistic Expression

Bibliography

Bloomfield L 1933 *Language*. Holt, Rinehart, & Winston, New York

Bouchard T 1997 IQ Similarity in twins reared apart. In: Sternberg R J, Grigorenko E L (eds.) *Intelligence, Heredity and Environment*. Cambridge University Press, Cambridge, UK, pp. 126–60

Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands

Chomsky N 1963 Formal properties of grammars. In: Luce R D, Bush R R, Galanter E (eds.) *Handbook of Mathematical Psychology*. Wiley, New York, Vol. 2, pp. 323–418

Fodor J 1975 *The Language of Thought*. Thomas Crowell, Cambridge, MA

Heider E R, Oliver D C 1972 The structure of the color space in naming and memory for two languages. *Cognitive Psychology* **3**: 337–54
Hofstader D R 1979 *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books, New York
Hunt E, Agnoli F 1991 The Whorfian hypothesis: A cognitive psychology perspective. *Psychological Review* **98**(3): 377–89
Lakoff G 1987 *Women, Fire, and Dangerous Things*. University of Chicago Press, Chicago
Levinson S C 1996 Language and space. *Annual Review of Anthropology* **25**: 353–82
Miller K F, Smith C M, Zhu J, Zhang H 1995 Preschool origins of cross-national differences in mathematical competence: The role of number-naming systems. *Psychological Science* **6**(1): 56–60
Pinker S 1994 *The Language Instinct*. William Morrow, New York
Roberson D, Davies I, Davidoff J 2000 Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General* **128**(3): 369–98
Sapir E 1921 *Language, an Introduction to the Study of Speech*. Harcourt, Brace, New York
Sera M D, Bales D W, del Castillo-Pintado J 1997 Ser helps Spanish speakers identify 'real' properties. *Child Development* **68**(5): 820–31
Slobin D I 1996 From 'thought and language' to 'thinking for speaking.' In: Gumpertz J J, Levinson S C (eds.) *Rethinking Linguistic Relativity*. Cambridge University Press, Cambridge, UK, pp. 70–96
Whorf B L 1956 *Language, Thought, and Reality*. MIT Press, Cambridge, MA
Wierzbicka A 1991 *Cross-cultural Pragmatics: The Semantics of Human Interaction*. Mouton de Gruyter, Berlin
Zhang S, Schmitt B 1998 Language-dependent classification: The mental representation of classifiers in cognition, memory, and ad evaluations. *Journal of Experimental Psychology: Applied* **4**(4): 375–85

E. Hunt

Language Contact

Language contact is any situation in which speakers of two or more languages communicate with each other—usually in person, but sometimes in writing (including email). This article gives a brief overview of the general topic in both its social (Sect. 1) and its linguistic aspects (Sect. 2).

1. *Language Contacts in Time and Space*

Languages have been in contact ever since human populations began spreading out into new territories and splitting into independent subgroups. No community in today's world is so isolated that its language remains untouched by outside influence over a long period of time; probably no language has been so isolated at any time in the last several thousand years. Every so often news reports trumpet the discovery of a 'stone-age tribe' that has been isolated from all other communities (and their languages) for thousands of years. A prominent late-twentieth-century case is that of the Tasadays, a small group that was discovered in the rainforest of the island of Mindanao in the Philippines. But the Tasadays, although they are not a hoax (as has been claimed), are certainly not a stone-age tribe (as was originally claimed for them): they are estimated to have been isolated from their nearest neighbors only for five to ten generations (Reid 1997; see also Headland 1991).

Most of the world's people are bilingual or multilingual. Attitudes toward the use of two or more languages within a community vary widely, however. Sometimes the highly visible symbolic status of language makes it a convenient pretext for interethnic hostilities: conflicts over language policies have erupted over the past few decades in many countries, among them the former Yugoslavia, where Croatian nationalists demanded linguistic autonomy years before the country's break-up; South Africa, where lethal riots broke out in the 1970s over the apartheid regime's insistence that the use of Afrikaans be increased in the townships' school systems; and Sri Lanka, where violent Tamil separatist activities were triggered by the 'Sinhalese-only' policy that the government established in the mid-1950s.

But a great many language contacts all over the world display no tendency toward conflict. Long-term bilingualism and multilingualism often flourish as a result of such social institutions as exogamy and regular trading relations with neighbors. Nor is cooperative multilingualism confined to small hunter–gatherer cultures. In some modern nation-states, for instance Switzerland, languages coexist peacefully, even productively.

For the vast majority of contact situations around the world we have no specific information about their onsets. Nevertheless, we can identify a variety of ways in which languages come into contact. At an earlier time in human history, some language contacts must have been established when groups met after moving into previously unoccupied territory; now that the world has no unoccupied habitable territory, this mode of contact onset is no longer available.

In more recent human history, probably the most common way in which languages come into contact is the movement of groups or individuals into other people's territory. Such movements can be peaceful immigrations, by people who wish to become integrated with the host population and people who want to work in agriculture or other industries. They can also be forced immigrations, as in the notorious Atlantic slave trade and in smaller-scale cases of slavery and indentured servitude. Or the movements can be more or less hostile encroachments by invading armies. The USA, Australia, and northern Europe, for

instance, have all been prime immigration destinations over the past century, and the resulting linguistic contacts have been the subject of a rich scholarly literature (to mention just three of innumerable examples, see Haugen 1953, Clyne 1997, and Backus 1996). Language contacts resulting from forced immigration have also received a great deal of scholarly attention, especially in the study of pidgin and creole languages (see e.g., Arends et al. 1994). Not surprisingly, given the long history of European and other colonial expansions, the literature on contacts initiated by invasions is the largest of all—covering contacts ranging from ancient Egyptian expansions to the European Age of Exploration, to Arab conquerors and indigenes in the Maghreb, to Russians and the peoples of the Russian Far East, and many others.

Some language contacts are ephemeral, others are extremely stable, and still others are in between. The stability of a language contact situation depends on many different social factors, among them the relative status and size of immigrant and host groups or of invaders and invaded; the degree to which one or more of the groups in contact desire complete amalgamation with the other group(s); the balance between 'other-directed' and 'self-directed' world views, and attitudes toward multilingualism itself, among multilingual groups in contact (Foley 1986, p. 27ff); and, when there is a dominant group, the amount of pressure it exerts on subordinate groups to shift to its own language.

2. *Linguistic Results of Language Contact*

The major linguistic aspects of language contact can be divided into several categories, including at least contact-induced language change, mechanisms of interference, the emergence of contact languages, and language death. But a preliminary comment is in order: one possible linguistic result of language contact, even under the most intense contact conditions, is no change at all. It is true that the great majority of long-term intense contacts all over the world involve significant contact-induced changes; but there are cases where lexical and even structural borrowing are rigorously avoided, so it is not safe to predict that contact-induced change is inevitable.

2.1 *Contact-induced Language Changes*

We can point to a number of linguistic and especially social predictors in trying to understand when and why contact-induced change occurs. Only the most trivial predictor is absolute: there can be no contact-induced change without contact. But other predictors also permit generalizations that are robustly supported by the linguistic evidence. One of the most important social predictors is the presence vs. the absence of imperfect learning of a target language by a group of people (Thomason and Kaufman 1988, Thomason 2001); another is the admittedly vague concept of intensity of contact. Strictly linguistic predictors, such as universal markedness and typological distance between source and receiving languages, are also important, but they are easily overridden by social factors.

If there is imperfect group learning—typically but not always in a case of group language shift—and a sufficiently intense contact situation, the first interference features to appear in the receiving language are phonological and syntactic features; morphology lags behind, and so, though for other reasons, does lexicon (except, sometimes, in the case of a shift by a superstrate group to a substrate group's language). So, for example, Latvian, a member of the Baltic subgroup of the Balto-Slavic branch of Indo-European, has undergone significant shift-induced interference in phonology, syntax, and to a lesser extent morphology as a result of shift by speakers of Uralic languages, especially Livonian; among the interference features in the most-affected dialects are its fixed initial stress rule, nominal object construction, and loss of gender distinctions (see Comrie 1981, pp. 147, 149, 152, 154). Shift-induced interference comprises both features carried over from the native language (L1) and failure to learn some target language (TL) features. Whether the shifting group's version of the TL will ultimately cause any changes in the TL as a whole depends on the extent to which the learners are integrated into the larger TL community. Note that imperfect group learning cannot be predicted with any confidence, though such factors as the relative sizes of the shifting and TL communities permit tentative predictions.

If imperfect learning plays no role in an interference process, the prediction about transferred features differs sharply: in this case the first interference features to appear in the receiving languages are nonbasic lexical items; structural features and basic vocabulary may appear later, if contact is intense enough, but in all or almost all known cases, borrowed nonbasic vocabulary predominates. In these interference processes, which have been labeled 'borrowing' in a narrower sense than the traditional usage (Thomason and Kaufman 1988), the people responsible for the interference are fully fluent in relevant aspects of both the source language and the receiving language.

Unlike the prediction of shift-induced interference in cases of imperfect group learning, even the most intense contact does not guarantee that significant borrowing will occur: communities in various parts of the world refuse to borrow vocabulary (or, in some cases, anything at all) from the language of their neighbors, even when (for instance) all members of a subordinate community are fully bilingual in a dominant group's language. So, for instance, the Montana Salish language in northwestern USA has been in intimate contact with English for 150 years, and the 60

or so remaining fluent speakers of the language are fully bilingual in English; and yet Montana Salish has only a tiny handful of loanwords from English, and no detectable structural interference of any kind.

Still, examples of extensive lexical and even structural borrowing are very easy to find in contact situations all over the world. A typical case is borrowing by Ossetic (Iranian, Indo-Iranian subgroup of Indo-European) from languages of the Caucasus, e.g., the Kartvelian language Georgian: much lexicon, glottalized ejective phonemes even in native vocabulary, several added cases in noun declension, agglutination in noun declension, and sentence structure that is more typically SOV (subject–object–verb) than in other Iranian languages of the former Soviet Union and that has a 'highly developed system of postpositions' (Comrie 1981, pp. 167, 171, 179).

Two-language contact situations, complex as they can be both socially and linguistically, are nevertheless easier to understand, in a historical sense, than multilingual contact situations, in particular regions called linguistic areas or *Sprachbünde*. Although the processes by which a Sprachbund arises are the same as those by which interference occurs in a two-language contact situation, establishing the sources of many areal features in a Sprachbund is difficult and often impossible, at least with our present knowledge of such diverse linguistic areas as the Balkans, South Asia, and the Pacific Northwest of North America.

2.2 Mechanisms of Interference

The only two mechanisms of interference that have received much attention in the literature are code-switching (especially as a mechanism of borrowing) and second-language acquisition strategies (as a mechanism of shift-induced interference). Code-switching becomes a route to interference when a code-switched element, such as a noun inserted into a sentence that is otherwise in a different language, turns up more and more frequently in the other language's sentences until it is fixed as a part of that language, with or without phonological and morphosyntactic adaptation to the receiving language's structure. One strategy used by second-language learners is the employment of L1 lexicon and structure to fill gaps in their knowledge of the L2, as when L1 English speakers use an English pronunciation of /r/ in speaking L2 German, or when they use English SVO word order in a German subordinate clause instead of the native German SOV order. A related strategy is guessing what patterns the L2 has on the basis of a perception that the L2 differs from the L1, as when Hungarian learners of a Serbo-Croatian dialect ended up with fixed penultimate stress—different both from fixed initial Hungarian stress and from free (other) Serbo-Croatian stress. These strategies lead to shift-induced interference if and when the learners' errors become a permanent part of a community's language. (The type of guessing that is done in such cases may be subsumable under a separate mechanism, called 'negotiation'; see Thomason 2001).

But although these are by far the best known mechanisms, there are others as well. One is code alternation, in which a bilingual speaker uses one language with some people and the other with different people—but never both languages with the same person, so that no code-switching is involved. In contrast to code-switching, this mechanism seems likely to be responsible for the importation of more structure than lexicon into one of a bilingual's languages; but as it has not yet been studied in any systematic way, it's hard to be sure of its typical effects. An example would be the borrowing of phonological features (e.g., aspiration of initial voiceless stops) and lexical semantic changes in the Italian of an Italian/English bilingual (M. Shigley-Giusti, personal communication, 1993). This example doesn't necessarily point to permanent interference even in this one speaker's idiolect; but it isn't hard to see how such local effects could eventually snowball into permanent community-wide interference.

A fourth mechanism of interference is passive familiarity, in which the language of a person or group is altered through contact with a language or dialect that the relevant speakers understand but do not ever speak. An obvious example is the borrowing of numerous words into whites' American English from African-American vernacular English, e.g., *Drop a dime* (on someone) 'to inform (on)' (as in *That dude'll drop a dime on ya in a heartbeat*), *boogie* 'go, leave,' and *chill* 'relax, calm down.' But structural features are also transferred by this mechanism. In fact, it is presumably part of the complex process through which shift-induced interference features make their way into the language of an integrated community consisting of L2 learners and original L1 speakers of the TL: the eventual community's language will contain a subset, but not all, of the shifting group's version of the TL—and those features were in effect borrowed by the original TL speakers, even though they are very unlikely to have spoken the shifting group's TL version.

A fifth mechanism of interference is bilingual first-language acquisition. Again, however, L1 acquisition as a mechanism of interference has not been studied systematically, so it is hard to know just how widespread its effects are. The clearest examples in the literature to date concern either syntactic features like word order or nonsalient phonological features such as intonation patterns, as in Queen's (1996) discovery that young Turkish/German bilinguals kept two phrase-final intonation patterns that are isofunctional in the two adult languages, and introduced a functional distinction between the two patterns in each of their L1's.

A sixth mechanism of interference, finally, is speakers' deliberate decision to alter their languages. This mechanism is best attested for changes in a whole language if the community is quite small, but examples like that of the eighteenth-century grammarians who introduced the Latin-inspired ban against split infinitives into Standard English at least show that the phenomenon isn't confined to hunter–gatherer societies.

2.3 Contact Languages and Language Death

The field of language contact is vast, and in a limited space only a few of its aspects can be touched on. To round out the picture a bit, this survey will conclude with very brief comments on two other aspects of great importance, contact languages and language death. Contact languages can (in my opinion) only be adequately defined by reference to their histories; there are no promising approaches that would permit them to be identified by their linguistic features alone. The one thing that all contact languages have in common is the fact that their structures and lexicons cannot all be traced back primarily to a single source language. Given the historical linguist's standard criterion for membership in a language family—descent with modification from a single parent language—this means that a contact language does not belong to any language family.

There are two basic kinds of contact languages: pidgins and creoles (see, for example, Arends et al. 1994), on the one hand, and bilingual mixed languages on the other (see e.g., Bakker and Mous 1994). The genesis of pidgins and creoles involves imperfect group learning and also a negotiation mechanism; but often there is no true target language in the sense of a language that the creators of the contact language try to learn. A creole, more or less by definition (although, as with most issues in this subfield, the claim is controversial), serves as the main language of some speech community. Prototypical pidgins serve a variety of functions: they permit communication among linguistically diverse groups that share no common language but need to communicate with each other about a limited range of topics; but many pidgins also serve to keep outsiders at a distance. Hiri Motu in New Guinea, for instance, probably developed as a means of fulfilling both purposes. It was a means of intergroup communication, but it also ensured that outsiders would not learn the true Motu language (Dutton 1997).

In sharp contrast to pidgins and creoles, bilingual mixed languages show no effects of imperfect learning: they are clearly created by bilinguals. They also serve a very different function from pidgins and creoles, because they are in-group languages, not intergroup languages.

Some contact languages arise abruptly, probably crystallizing into a language that has to be learned as such within a few years or at most one or two generations; others appear to be gradual developments. Abruptly created contact languages almost certainly include many pidgins and the creole Pitcairnese (which developed after 1790 among the Bounty mutineers, their Tahitian-speaking partners, and the offspring of those unions); they also include such bilingual mixed languages as Michif, a language created by mixed-blood Cree/French bilinguals (French noun phrases, Cree verb phrases and sentential syntax), and Mednyj Aleut, the language of mixed-blood Aleut/Russian bilinguals (Russian finite verbal inflection, otherwise mostly Aleut). In all such abrupt cases of language genesis there is a sharp break in transmission; the resulting mixed language is not the product of normal language transmission, which involves the learning of a whole language from speakers of that language (see Thomason and Kaufman 1988). It is not the product of ordinary bilingual L1 acquisition, either, because bilingual L1 learners keep their languages separated from a very early age.

Other contact languages, however, seem to be best understood as the product of gradual development: at every stage in the process, a whole language is passed down from one generation to the next, but so many interference features are accumulated over a long period of time that, ultimately, too little is left of the original language for it to be classifiable as belonging to its original family. Contact languages of this type comprise at least some creole languages born of the slave trade, where successive waves of imported slaves learned versions of a European TL that was increasingly distant from the original TL as spoken by Frenchmen (see Arends 1993, Chaudenson 1989), and also bilingual mixed languages which developed through long-term intensive borrowing that ultimately replaced most of the inherited structure and much of the inherited lexicon. A prime example is Ma'a (northeastern Tanzania), which started as a non-Bantu language (perhaps originally an ordinary Cushitic language, though the evidence is unclear—Mous 1993) but, over a period of three hundred years or so, incorporated so much lexicon and structure from neighboring Bantu languages that by the 1980s its lexicon was about half Bantu and its structure almost entirely Bantu, including a full set of Bantu inflectional affixes (Thomason 1997, but see Mous 1994 for a different historical scenario).

Finally, the saddest outcome of language contact, often for the speakers and always for linguists' hope of understanding human language in all its immense diversity, is language death. This outcome is becoming increasingly common in our modern world, and in spite of intensive language preservation efforts in many places, the process is probably irreversible for a shockingly high percentage of the languages currently alive (Hale et al. 1992). The linguistic study of language death has focused on attrition, though other aspects of

the topic have also received a considerable amount of attention; for a representative sampling, see Dressler and Wodak-Leodolter (1977) and Dorian (1989).

See also: Bilingualism and Multilingualism; Code Switching: Linguistic; First Language Acquisition: Cross-linguistic; Foreign Language Teaching and Learning; Historical Linguistics: Overview; Language Acquisition; Language and Ethnicity; Language Endangerment; Language Obsolescence; Pidgin and Creole Languages; Second Language Acquisition; Sociolinguistics

Bibliography

Arends J 1993 Towards a gradualist model of creolization. In: Byrne F, Holm J (eds.) *Atlantic Meets Pacific: A Global View of Pidginization and Creolization*. Benjamins, Amsterdam, pp. 371–80

Arends J, Muysken P, Smith N (eds.) 1995 *Pidgins and Creoles: An Introduction*. Benjamins, Amsterdam

Backus A M 1996 *Two in one. Bilingual Speech of Turkish Immigrants in the Netherlands*. Tilburg University Press, Tilburg

Bakker P, Mous M (eds.) 1994 *Mixed Languages: 15 Case Studies in Language Intertwining*. Uitgave IFOTT (Institute for Functional Research into Language and Language Use), Amsterdam

Chaudenson R 1989 *Créoles et enseignement du français*. L'Harmattan, Paris

Clyne M 1997 Multilingualism in Australia. *Annual Review of Applied Linguistics*. **17**: 191–203

Comrie B 1981 *The Languages of the Soviet Union*. Cambridge University Press, Cambridge, UK

Dorian N C (ed.) 1989 *Investigating Obsolescence: Studies in Language Contraction and Death*. Cambridge University Press, Cambridge, UK

Dressler W, Wodak-Leodolter R (eds.) 1977 Language death. *International Journal of the Sociology of Language* **12** (Special issue)

Dutton T 1997 Hiri Motu. In: Thomason S G (ed.) *Contact Languages: A Wider Perspective*. Benjamins, Amsterdam, pp. 9–41

Foley W A 1986 *The Papuan Languages of New Guinea*. Cambridge University Press, Cambridge, UK

Hale K L, Krauss M E, Craig C, England N, Jeanne La V, Watahomigie L, Yamamoto A 1992 Endangered languages. *Language* **68**: 1–42

Haugen E 1953 *The Norwegian Language in America*. University of Pennsylvania Press, Philadelphia, PA, 2 Vols.

Headland T N (ed.) 1991 *The Tasaday Controversy*. American Anthropological Association, Arlington, VA

Mous M 1993 The making of Ma'a. *Colloquium on Synchronic and Diachronic Sociolinguistic Methods and Interpretations, Universität Bayreuth*

Mous M 1994 Ma'a or Mbugu. In: Bakker P, Mous M (eds.) *Mixed Languages: 15 Case Studies in Language Intertwining*. Uitgave IFOTT (Institute for Functional Research into Language and Language Use), Amsterdam, pp. 175–200

Queen R M 1996 *Intonation in Contact: A Study of Turkish-German Bilingual Intonation Patterns*. Dissertation, University of Texas, Austin, TX

Reid L A 1997 Essay on the Tasaday controversy. http://www2.hawaii.edu/ling/fac_pages/reid.html

Thomason S G 1997 Ma's (Mbugu). In: Thomason S G (ed.) *Contact Languages: A Wider Perspective*. Benjamins, Amsterdam, pp. 469–487

Thomason S G 2001 *Language Contact*. Edinburgh University Press, Edinburgh, UK

Thomason S G, Kaufman T 1988 *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley, CA

S. G. Thomason

Language, Continental Philosophy of

Although philosophical interest in language is as old as philosophy itself, during the twentieth century this interest flourished to such an extent that the analysis of language has now become one of the most productive philosophical fields. Besides the general interest in language present in virtually all different philosophical traditions of the past century, it is possible to distinguish at least two very different ways of approaching linguistic analysis, namely, the so-called 'Anglo-American' and 'continental' philosophies of language. The first (also called 'analytic' philosophy) usually refers to the kind of analysis of language initiated at the end of the nineteenth and the beginning of the twentieth century by authors such as G. Frege, B. Russell, L. Wittgenstein, the logic positivists, etc. The second, continental philosophy of language, refers to the conception of language that finds its roots in the eighteenth century, in the works of authors such as J. G. Hamann and J. G. Herder, and was further developed in the nineteenth century by W. v. Humboldt. Through the work of M. Heidegger, it has extended its influence to contemporary hermeneutic philosophy (H.-G. Gadamer, K.-O. Apel, J. Habermas, etc.). In this article the specific features of the continental conception of language are examined.

In general terms, one of the crucial differences between the Anglo-American and the continental philosophies of language is the explicit attempt found in all the authors of the continental tradition to break with what Humboldt termed the 'primacy of logic over grammar,' that is, the predominance given from the very beginnings of Greek philosophy to the cognitive function of language at the expense of its communicative function. As a consequence of conceiving language less as a vehicle of knowledge and more as a means of understanding, this tradition analyzes the role that language plays not only in our relationship with the objective world (by allowing us to have propositional knowledge about it) but also in our relationship with the social world (which is constituted through communicative interaction) and with our

own subjective worlds (which are expressible through linguistic articulation). In this way, language is considered in its multidimensional world-disclosing function. In keeping with this focus, continental philosophy of language has always concentrated on the analysis of natural languages, developed through contingent historical processes, and has considered these as *constitutive* of our experience of the world.

In order to understand the roots of the distinctive features of the continental conception of language, it is important to keep in mind that this conception was developed against the background of an explicit confrontation with Kant's philosophy. This is important because, in spite of all criticisms of Kant, continental philosophy of language remains to this day under the decisive influence of transcendental philosophy. This influence explains two of its most characteristic features: its idealist tendencies (which in some cases lead to a clear hypostatization of language, most remarkably in the work of the late Heidegger) as well as one of the crucial problems that have concerned the authors of this tradition from the very beginning, namely the problem of linguistic relativism. The connection between these features can be understood best through an analysis of the different historical phases of development within this tradition.

1. *The Hamann–Herder–Humboldt Tradition*

As already suggested above, most of the reflections on language found in the works of the authors of this tradition take place in more or less explicit confrontation with Kant's philosophy. In the cases of Hamann and Herder this confrontation takes the specific form of a metacritique of Kant's critique of pure reason (see Hamann 1784/1988, Herder 1799/1960). One important target of both metacritiques is Kant's attempt to purify reason from its internal connection with language. The constant reference that Hamann and Herder make to the common origin of both expressions 'language' and 'reason' in the Greek word *logos* is by no means a merely etymological remark. With the explicit identification of language and reason (i.e., the claim that 'reason is language, *logos*' or 'without the word, neither reason nor world'), these authors initiate a way of thinking about both language and reason that moves constantly between the extremes of universalism and relativism, rationalism and irrationalism. On the one hand, the identification of language and reason supports the attempt to transform Kant's idea of a critique of reason into the romantic idea of the essential limitations of reason (due to the limits of the natural languages that contingently developed in history). This is a powerful motive behind the critique of the Enlightment in which many authors of this tradition participate (from Hamann to Heidegger and Gadamer). On the other hand, however, the identification of language and reason also supports the attempt to preserve the universalism of Kant's conception of reason in an equally universalist conception of communicative rationality (from Humboldt to Apel and Habermas). But beyond the remarkable differences in the philosophical programs among the authors of this tradition, the common presupposition behind their identification of language and reason is the view of language as constitutive of our experience in the strict sense given by transcendental philosophy. Accordingly, language is conceived as having the transcendental power to constitute experience that Kant erroneously ascribed to an abstract, transcendental ego.

Hamann's critique of Kant can be seen in retrospect as the initiator of such a view. Hamann identifies language as the common root of sensibility and understanding, which was sought by Kant without success. In this way, he considers language as both empirical and transcendental at the same time. By conceding to language such a quasi-transcendental status, Hamann certainly disagrees with Kant's view of language. However, he does not disagree with the most basic assumptions of transcendental philosophy. Hamann is actually taking for granted Kant's transcendental idealism, namely, the idea that the conditions of possibility of experience are at the same time the conditions of possibility of the objects of experience. As a consequence of Hamann's critique, this idea is now understood to mean that the conditions of possibility of our linguistically mediated experience of the world are, at the same time, the conditions of possibility of that which is experienced in the world. Now, the claim that language does not merely mirror but actually shapes our experience of the world faces a problem completely foreign to transcendental philosophy. For the various historical languages do not seem to offer an equivalent for the Kantian transcendental ego. First of all, these languages are plural. Second, they do not allow for a sharp distinction between the empirical and the transcendental, between that which is supposed to count as valid in them *a priori* (knowledge of the language) and that which is valid *a posteriori* (knowledge of the world). Thus if languages shape our experience of the world, they do so in merely contingent and plural ways. As Hamann points out, the meaning and use of words of different languages is *a priori* arbitrary and indifferent, but *a posteriori* necessary and indispensable. Seen in this light, Hamann's critique of Kant seems to force the relativist conclusion that Humboldt will express later in one of his most famous theses, namely, that 'every language contains a peculiar world-view.'

Hamann never addressed the problem of linguistic relativism entailed in his transformation of transcendental philosophy. Moreover, given the decisively antisystematic impulse behind his critique of Kant's philosophical system, he never attempted to work out this new view of language into a systematic conception. His reflections on language are limited to more or less

isolated remarks spread throughout his writings. A much more elaborate account of this new conception of language can be found in Humboldt's work. Although Humboldt's general reflections about language do not amount to a philosophical theory of language—they are mostly confined to the introductions to empirical works in comparative linguistics—one of their specific aims is to bring about a radical break with the traditional conception of language. Humboldt is convinced that without such a break a science of language is not possible at all. Thus his writings contain a very articulated account of the essential features that distinguish the new conception of language from the received view.

In the context of criticizing the traditional conception of language in his work *Ueber den Nationalcharakter der Sprachen*, Humboldt enumerates the basic features of his own conception of language when he remarks 'that the diversity of languages is more than a mere diversity of signs, that words and their syntax simultaneously shape and determine our concepts and that, given their systematicity and their influence on knowledge and sensibility, different languages are in fact different world-views' (Humboldt 1903–36, IV, p. 420). In this brief summary, we can see the three fundamental claims that separate Humboldt's position from the received view: (a) the rejection of the view of language as a mere system of signs, as a tool for the transmission of prelinguistic thoughts or the designation of objects that exist independently of language, (b) the thesis of the identity of thought and language that justifies the critique of the instrumental conception of language, and as a consequence (c) the claim that different languages are in fact different world-views. The problem of linguistic relativism implicit in this consequence can be traced back to the very core of Humboldt's critique of the instrumental conception of language.

The traditional view of language as a mere instrument for the designation of entities independent of language is based on an assimilation of two relations that are actually irreducible to one another: the relation of *designation* between a name and an object and the relation of *satisfaction* between a predicate and the objects to which the predicate applies. From this assimilation there arises the problem of determining what the predicates actually 'designate,' a problem encountered already in the Middle Ages; the ensuing dispute over the universals offers clear testimony as to its depth. Nonetheless, this highly problematic conception of language can be considered as the ruling view from Aristotle to Kant. Humboldt seeks a decisive break with this reduction of the powers of language to their designating function, advocating a more stratified analysis of language. In contemporary terms, such an analysis would have to deal not only with the name/object distinction, but also with the distinction between concept (or class) and object. By the same token, concerning the interpretation of signs, this analysis would have to differentiate between their meaning and reference. Humboldt states this last distinction as follows: 'The word acts in such a way that the object given in it is represented by the soul. This representation must be distinguished from the object; … along with the objective part *that relates to the object*, it has a subjective part that *lies in its manner of conceiving*. … In turn, we scarcely need to mention that this division rests upon [an] abstraction; [further,] that the word can have no place outside of thought, and neither can its object, if it is non-physical.… This remains so even in the case of material objects, *since it is never the objects directly, but always only the representation of them provided by the word, that is made present in the soul*' (Humboldt 1903–36, V, p. 418).

It is important to note that by introducing the useful distinction between the referent of the word and its meaning, Humboldt also excludes *every* possibility of a pure designating relation between a name and its object, which was the paradigm for the traditional explanation of the functioning of language. He generalizes the meaning/reference distinction as applicable to all linguistic signs. Accordingly, Humboldt extends this distinction even into the realm of names, including proper names. He continues: 'The word conceives of *every concept as general*, always designating, strictly speaking, classes of reality, *even if it is a proper name*; for thus it comprehends … all the various states, with respect to time and space, of that which it designates' (Humboldt 1903–36, V, p. 419).

Thus, Humboldt counteracts the problematic reduction of predicates to mere names in the traditional view with the inverse reduction, namely, the assimilation of names to predicates. This move can be considered the systematic point from which the problem of linguistic relativism arises. If this is not immediately evident, it soon becomes clear by analyzing its implications. According to this assumption, names are actually general concepts. Instead of designating an object, they too, by strict analogy with predicates, can relate only in a mediated fashion to their referent, as if to an object falling under a concept. Humboldt explains: 'the picture of language as merely designating *objects*, already perceived in themselves, is also not sustained by examination of what language engenders as its product.… Just as no concept is possible *without language*, so also *there can be no object for the mind*, since it is only through the concept, of course, that anything external acquires full being for consciousness. But the whole *mode of percieving* things *subjectively* necessarily passes over into cultivation and the use of language. For the word arises from this very perceiving; it is a copy, not of the object in itself, but of the image thereof produced in consciousness' (Humboldt 1903–36, VII, p. 60). Seen in this light, the idealistic conclusion that Humboldt draws from this assumption seems entirely consistent: namely, that

'man [lives] with the objects … *exclusively* in the way that they are conveyed to him by language' (Humboldt 1903–36, VII, p. 60). This in turn explains why different languages are in fact different world-views. As Humboldt remarks in a similar context: 'In this way, in words of identical meaning that are found in a number of languages, there arise diverse representations of the same object, and this characteristic of the word contributes chiefly to the fact that *every language yields a particular world-view*' (Humboldt 1903–36, V, p. 420).

However, once the diversity of conceptual systems or world-views contained in the different languages is recognized, it is no longer clear how speakers of different linguistic communities can talk about the same things and thus communicate successfully. Humboldt has to admit that 'every language places *definite boundaries* upon the spirits of those who speak it, and insofar as it provides a determinate orientation, excludes others' (Humboldt 1903–36, VII, p. 621). The commensurability of different linguistic world-views does not become a problem for Humboldt because he has skeptical doubts about the possibility of translation between languages. Actually Humboldt remarks repeatedly that 'experience with translation of the most diverse languages … shows that *it is possible to express every concatenation of ideas in every language*' (Humboldt 1903–36, IV, p. 17). Instead, the problem of linguistic relativism originates in the fact that through his criticism of the traditional conception of language Humboldt has destroyed the plausibility of the traditional answer to this problem. His conception of language as world-disclosing precludes the possibility of explaining interlinguistic understanding by appealing to a world-in-itself accessible to all human beings in the same way and to which all speakers refer through the different conventions of their languages.

Humboldt is aware of this challenge and indeed tries to articulate an alternative answer to the question of interlinguistic understanding. His answer is also based on a criticism of the traditional view of language, but this time from a different angle. Humboldt opposes the reduction of language to a mere system of signs entailed in the traditional view. In the introduction to his work on the Kawi language, Humboldt expresses this opposition with the help of one of his most famous claims, namely that language is not a product (*ergon*) but an activity (*energeia*). To take this claim seriously implies transforming the traditional conception in two important respects.

First, in the sense already mentioned, Humboldt's view of language as an *activity* underscores the fact that language does not passively mirror the world but it is actually world disclosing. Language is what makes the world appear as an ordered whole. For it is through the web of linguistic concepts that what would otherwise be a disconnected flux of perceptions is transformed into an order of objects properly so called. In this sense Humboldt claims that 'the generation of language is, from the beginning, a *synthetic process* in the most genuine sense of the term, *where the synthesis creates something that did not exist before in any constituent part*' (Humboldt 1903–36, VII, p. 95). Humboldt terms this genuine synthesis 'articulation' and regards it as 'the authentic essence of language, the lever *through which language and thought are produced*, the culmination of their intimate mutual connection' (Humboldt 1903–36, VI, p. 153).

However, there is a second, more radical sense in which Humboldt sees language as an activity. According to Humboldt, the synthetic process through which both language and thought are produced is not something given once and for all, i.e., fixed through the grammar and lexicon of a language. It is rather the result of a praxis, namely the process of communication between speakers. As he remarks repeatedly: 'Language only exists in ongoing speech; grammar and lexicon are hardly comparable to their dead skeleton' (Humboldt 1903–36, VI, p. 148). Humboldt opposes the reduction of language to a mere vehicle of thought that underlies the analysis of language in logic. In this context language is considered in abstraction from the use of language in speech for the purposes of analyzing the general structure of thoughts. But Humboldt's opposition to this abstraction is not based on the obvious fact that language, in addition to being a vehicle of thought, is also a means of communication. His view is rather that language can only be a vehicle of (intersubjective) thought to the extent that it makes communication and understanding between speakers possible. Thus a proper analysis of the cognitive function of language is impossible if language is considered in isolation from its communicative function. Behind Humboldt's claim is the idea that the different communicative perspectives, generated through the system of personal pronouns present in all languages, are the condition of possibility of any intersubjectively valid thought. For these different perspectives make it possible that speakers both express their views in their irreducible singularity and achieve a common understanding through conversation as well. Thus the commesurability of linguistic world-views is not guaranteed in advance by a unique syntax or lexicon; it is actually the result of a different feature of all languages, namely its dialogical structure. It is in virtue of the pragmatic structure of speech that a free synthesis of perspectives among speakers can come about in conversation.

Humboldt is fascinated by the double nature of language as the *source* of both *generality* and *individuality*. Due to its general character language makes it possible that different speakers share identical concepts (and in this way fulfills a socializing function). But due to its expressive potential it also allows speakers to maintain their genuine individuality (i.e., it makes individualization possible). Humboldt explains: 'in language the *individualization* within a

general conformity is so wonderful, that we may say with equal correctness that the whole of mankind has but one language, and that every man has one of his own' (Humboldt 1903–36, VII, p. 51). However, there is an obvious tension in this view. On the one hand, Humboldt claims that in virtue of its intrinsic generality (the fact that we can share it), language entails the condition of possibility of understanding among speakers. It guarantees, in Humboldt's words, that: 'nobody may speak differently to another from the way in which the latter, under similar circumstances, would have spoken to him' (Humboldt 1903–36, VII, p. 47). But on the other hand, the possibility of understanding among speakers is not a mere function of a shared grammar or lexicon. Given that the only reality of language is speech, genuine understanding can only be the lucky outcome of the process of communication itself, in which different individuals express their individuality or particular world-view. For 'only in the individual does language get its last determination.... A nation as a whole has indeed the same language, but not all the individuals in it have ... exactly the same one, and if one looks more carefully, every person actually has his own' (Humboldt 1903–36, VI, p. 183). But given that speakers' linguistic differences are actually differences in world-view, it is not clear within Humboldt's account how speakers can talk about the same things despite their genuinely different linguistic world-views. Without such an explanation the possibility of understanding in dialogue remains puzzling. Humboldt has to recognize that 'nobody means by a word precisely the same as his neighbor does, and the difference, be it ever so small, vibrates, like a ripple in water, throughout the entire language. Thus all understanding is always at the same time a not-understanding, all agreement in thought and feeling at the same time a divergence' (Humboldt 1903–36, VI, p. 183). The tension present in Humboldt's writings between the roles ascribed to language in its world-disclosing and in its communicative dimensions remains a characteristic feature of the works of most of the authors of this tradition.

2. *Further Developments in the Twentieth Century*

M. Heidegger is undoubtedly the most influential defender of Humboldt's conception of language as world-disclosing in twentieth century philosophy. However, despite the deep similarities between the conceptions of language developed by both authors, Heidegger was rather critical of Humboldt's approach to language at different times during his philosophical development and for different reasons (see Heidegger 1927, 1986). The view of language as responsible for world disclosure is not articulated explicitly in Heidegger's early writings, although some features of what would be his later conception of language already appear in them, including in *Being and Time* (see Lafont 2000). Most of Heidegger's reflections on language are contained in a collection of articles that appeared in English under the title *On the Way to Language*.

As is characteristic of all the authors in this tradition, Heidegger articulates his view of language as world-disclosing in direct opposition to the traditional conception of language as a mere tool for the designation of objects that exist independently of language. He argues that the linguistic activity of naming something with a word cannot be understood as a pure relation of designation between a name and an entity. It does not consist of the fact 'that something already familiar to us is provided with a name, rather ... through this naming, the entity is first nominated as that which it is. It is in this way that it is known as an entity' (Heidegger 1944). In *The Origin of the Work of Art*, Heidegger explains: 'When language names the entity for the first time, such naming alone brings the entities to word and to appearance. This naming nominates the entity to its being from out of this being. Such saying is a projecting of the clearing, in which it is announced as what the entity comes into the open' (Heidegger 1986, pp. 59–60). In *On the Way to Language* he expresses the same idea in a much more provocative way using a line of Stefan George's poem, namely, that 'there is no thing where the word is lacking.' Heidegger explains: 'The thing is a thing only where the word is found for the thing. Only in this way *is* it.... The word alone supplies being to the thing, [for] ... something only *is*, where the appropriate word names something as existing and in this way *institutes* the particular entity *as such*.... The being of that which is resides in the word. For this reason, the following phrase holds good: language is the house of being' (Heidegger 1986, pp. 164–6).

As Humboldt had already argued before, entities can only be identified as such once they are linguistically categorized. It is in this sense that language is responsible for world-disclosure: language determines our possible experience with entities and thus determines in advance what kind of world is disclosed to us. With regard to the world-disclosing power of language, Heidegger basically agrees with Humboldt. However, on this very same basis he criticizes Humboldt's view of language as an activity. As mentioned above, Humboldt tried to conceive the production of a linguistic world-view as a result of the human activity of speech. This would explain how speakers can synthesize their different world-views and revise them through dialogue. However, as Heidegger argues, in order for a dialogue to be possible at all, it is first necessary that one and the same world be disclosed to all speakers. As he had already claimed in *Being and Time*, 'only he who already understands can listen.' Heidegger interprets Humboldt's emphasis on the communicative dimension of language as just a remi-

niscence of the instrumental conception of language. Heidegger directs one of his most provocative claims against any attempt to transfer language's power of world disclosure to the human activity of speech. He writes: 'Man acts as though he were the shaper and master of language, while in fact language remains the mistress of man … For strictly, it is language that speaks. Man first speaks when, and only when, he responds to language by listening to its appeal' (Heidegger 1954, p. 184).

As a result of this extremely idealistic conception of language, Heidegger sees world disclosure, which is anchored in the language of a specific people or culture, as the most significant event of their historical development (see Heidegger 1980). Based on this idea, Heidegger interprets the history of Western culture as the fateful development of the understanding of being as presence, which is disclosed in Western languages and may bring this culture towards its own destruction. From a Humboldtian point of view, the reification of language implicit in Heidegger's view seems just a consequence of ignoring that language is not a product but an activity, i.e., that language is nothing other than the communicative activity between human beings. However, it is hard to question Heidegger's criticism of this Humboldtian claim, given that for both authors language has the power to constitute our experience in the strict sense set by transcendental philosophy. So long as this presupposition is not questioned, achieving the balance sought by Humboldt between the world disclosing and the communicative dimensions of language remains an unresolved task in this tradition.

H.-G. Gadamer, Heidegger's most influential student in contemporary German philosophy, tries to achieve this balance in his main work *Truth and Method*. In this work Gadamer articulates the essential features of the approach to philosophy that he calls 'philosophical hermeneutics.' At the core of this approach is an analysis of hermeneutic experience, i.e., of the interpretative relationship that human beings entertain with each other and with their symbolic products (e.g., texts, works of art, and the like). To support his account of hermeneutic experience he appeals to the conception of language developed in the Hamann–Herder–Humboldt tradition, to which he wishes to contribute with his own reflections on language. He uses the Humboldtian model of dialogue between speakers who try to reach agreement about something in the world and generalizes it as the appropriate model for describing any hermeneutic experience whatsoever (e.g., the understanding of canonical texts, works of art, and any other symbolic products by an interpreter, most significantly in the context of the human sciences). In this respect, Gadamer explicitly focuses on the communicative dimension of language. However, he looks at the Humboldtian model of dialogue with Heideggerian eyes.

Gadamer emphasizes that a genuine dialogue by its very nature is not just the product of the speakers' activity. First of all, as Gadamer remarks, 'no one knows in advance what will come out of a conversation. Understanding or its failure is like an event that happens to us' (Gadamer 1960, p. 383). But more importantly, a conversation is only possible if the participants share a linguistically disclosed world. Thus, a shared world disclosure, i.e., a common tradition, is the precondition for any understanding or agreement that speakers try to bring about in conversation. Although Gadamer opposes, as much as did Humboldt, the thesis of the incommensurability of different linguistic world disclosures, he has never seriously confronted the challenge of explaining how intercultural understanding is possible, that is, how a dialogue between participants who do not belong to the same tradition and thus do not share the same linguistic world disclosure is possible at all.

More recently, K.-O. Apel and J. Habermas have tried to work out Humboldt's and Gadamer's model of dialogue into a systematic theory of communication (see Apel 1994, Habermas 1981). Both authors identify in the communicative function of language a specific kind of rationality (communicative rationality) that allows participants in communication to reach a rational agreement about something in the world. Although they share the view of language as world disclosing that characterizes the continental tradition, their approaches introduce a strong dichotomy between form and content that aims to conceive rational understanding as a universal possibility (and thus to avoid the problem of incommensurability between different linguistic world disclosures). In clear opposition to Heidegger and Gadamer's claim that understanding is only possible on the basis of a factual agreement among speakers with a shared linguistic world disclosure, Apel and Habermas claim that understanding depends only on a 'counterfactual agreement' that all speakers share just in virtue of their communicative competence. This agreement is based on *formal* presuppositions and thus it does not depend on a shared content or shared world disclosure among the participants in conversation. A rational reconstruction of these formal presuppositions of communication is the core of the analysis of language that these authors offer in their respective approaches. Apel understands this analysis as a transformation of transcendental philosophy within the linguistic paradigm and thus calls it 'transcendental semiotics.' Habermas disagrees with Apel's claim that the formal presuppositions of communication have a transcendental status in the strict Kantian sense; thus he prefers the designation of 'formal pragmatics' for his approach. But beyond these methodological differences their reconstructions of the formal presuppositions of communication are very similar.

According to both approaches, speakers who want to reach an agreement about something in the world

have to presuppose the truth of what they are saying, the normative rightness of the interaction they are establishing with the hearer through their speech acts, and the sincerity or truthfulness of their speech acts. Complementary to these three validity claims (truth, normative rightness, and truthfulness), speakers share the notion of a single objective world that is identical for all possible observers. This is just a *formal* notion that is freed of all specific content. It is important to note that this presupposition does not correspond to the naïve belief in a world-in-itself accessible to all speakers that characterizes the traditional conception of language. As Habermas argues, if speakers had such a naïve belief they would identify the *contents* of their respective linguistic world-views with the assumed world-order itself. But with such a dogmatic attitude they would not admit the possibility that in case of disagreement the claims raised by other speakers may turn out to be right about the way the world is. In order for a rational discussion to be possible at all, speakers have to be able to adopt a hypothetical attitude vis-à-vis their own beliefs about the world. This is only possible if they distinguish, however counterfactually, between the assumed world-order itself and their respective claims about how the world is. Seen in this light, the formal presupposition of a single objective world is just a consequence of the universal claim to validity built into the speakers' speech acts. It is just an expression of the communicative constraint that makes rational criticism and mutual learning possible, namely, that from two opposed claims only one can be right. Thus the formal notion of world and the three universal validity claims build a system of coordinates that guides the interpretative efforts of the participants in communication towards a common understanding, despite their differences in beliefs or world-views.

It is an open question whether this proposal is actually successful in explaining the possibility of interlinguistic understanding, as much as it is open whether any current theory within contemporary philosophy of language is so successful. However, Apel's and Habermas' approaches are intentionally designed to meet this challenge in a way unprecedented in prior continental philosophy of language. Equally unprecedented is the scope and sophistication of the philosophical analysis of language and communication that these authors offer. This is in part due to the reception and creative integration in their own work of most of the approaches developed within contemporary philosophy of language across the borders of the different philosophical traditions.

See also: Communication and Social Psychology; Communication: Philosophical Aspects; Language and Philosophy; Language and Thought: The Modern Whorfian Hypothesis; Linguistic Turn; Linguistic Turn and Discourse Analysis in History; Linguistics: Overview

Bibliography

Apel K O 1963 *Die Idee der Sprache in der Tradition des Humanismus von Dante bis Vico*. Bouvier Verlag, Bonn

Apel K O 1973 *Transformation der Philosophie*. Suhrkamp Verlag, Frankfurt, 2 Vols. (English trans. *Towards a Transformation of Philosophy*. Routledge & Kegan Paul, London, 1980)

Apel K O 1994 *Selected Essays. Volume One: Towards a Transcendental Semiotics*. Humanities Press International, New Jersey

Apel K O 1998 *From a Transcendental-Semiotic Point of View*. Manchester University Press, Manchester, UK

Gadamer H G 1960 *Wahrheit und Methode*. J C B Mohr, Tübingen (English trans. *Truth and Method*. Continuum, New York, 1994)

Gadamer H G 1986 *Hermeneutik II*, *Gesammelte Werke*, *Vol. 2*. Möhr, Tübingen

Habermas J 1981 *Theorie des kommunikativen Handelns*. Suhrkamp Verlag, Frankfurt (English trans. *Theory of Communicative Action*, 2 Vols., Beacon Press, Boston, 1984/87)

Habermas J 1984 *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns*. Suhrkamp Verlag, Frankfurt

Habermas J 1998 *On the Pragmatics of Communication*. MIT Press, Cambridge, MA

Habermas J 1999 Hermeneutic and analytic philosophy: Two complementary versions of the linguistic turn? In: O'Hear A (ed.) *German Philosophy since Kant*. Cambridge University Press, Cambridge, UK

Hamann J G 1949–57 In: Nadler J (ed.) *Sämmtliche Werke*. Vienna, 6 Vols.

Hamann J G 1784/1988 Metakritik über den Purismus der Vernunft. In: Majetschak J (ed.) *Vom Magus im Norden und der Verwegenheit des Geistes. Ein Hamann Brevier*. Deutscher Taschenbuch Verlag, Munich, pp. 205–12

Hamann J G 1988 In: Majetschak J (ed.) *Vom Magus im Norden und der Verwegenheit des Geistes. Ein Hamann Brevier*. Deutscher Taschenbuch Verlag, Munich

Heidegger M 1927 *Sein und Zeit*. Max Niemeyer Verlag. Tübingen (English trans. *Being and Time*. Harper & Row, New York, 1962)

Heidegger M 1944 Hölderlin und das Wesen der Dichtung. *Erläuterungen zu Hölderlins Dichtung*. Klostermann, Frankfurt

Heidegger M 1954 *Vorträge und Aufsätze*. G. Neske, Pfullingen

Heidegger M 1980 Der Ursprung des Kunstwerkes. *Holzwege*. Klostermann, Frankfurt

Heidegger M 1986 *Unterwegs zur Sprache*. G. Neske, Pfullingen (English trans. *On the Way to Language*. Harper & Row, New York, 1971)

Herder J G 1799/1960 Eine Metakritik zur Kritik der reinen Vernunft. In: Heintel E (ed.) *Sprachphilosophische Schriften*. Felix Meiner Verlag, Hamburg, pp. 183–227

Herder J G 1960 In: Heintel E (ed.) *Sprachphilosophische Schriften*. Felix Meiner Verlag, Hamburg

Humboldt W v 1903–36 In: Königlich Preussische Akademie der Wissenschaften (ed.) *Gesammelte Schriften*. B. Behr's Verlag, Berlin, 17 Vols.

Humboldt W v 1963 Schriften zur Sprachphilosophie. In: Flitner A, Giel K (eds.) *Werke in fünf Bänden*. Klett-Cotta Verlag, Darmstadt, Vol. 3

Humboldt W v 1988 *On Language. The Diversity of Human Language—Structure and Its Influence on the Mental Development of Mankind*. Cambridge University Press, Cambridge, UK

Lafont C 1999 *The Linguistic Turn in Hermeneutic Philosophy*. MIT Press, Cambridge, MA
Lafont C 2000 *Heidegger, Language, and World-Disclosure*. Cambridge University Press, Cambridge, UK
Taylor C 1985 Theories of meaning. In: *Human Agency and Language: Philosophical Papers*. Cambridge University Press, Cambridge, UK, 248–92
Trabant J 1990 *Traditionen Humboldts*. Suhrkamp Verlag, Frankfurt

C. Lafont

Language: Dependency

1. Notion of Linguistic Dependency

A sentence is organized by syntagmatic relations holding between its words; these relations are involved in arranging words into strings, supplying them with appropriate prosody and giving them a necessary form. (Syntagmatic relations can be contrasted with paradigmatic relations holding between words in the language and involved in choosing semantically/syntactically appropriate words.)

The main type of syntagmatic relations is dependency: *B depends on A* if *A* controls *B*, i.e., determines *B*'s presence, position or form; we also say that *A governs B*. ('Depend on' and 'govern' are converse relations, as 'be a parent of' ~ 'be a child of.') In sentence (1)

(1) *The boy cares for his mother*.

boy depends on *care*: as the Subject, *boy* must precede the Main Verb [MV], i.e., *cares*; *for* also depends on *cares*: (if MV were *thinks* or *likes*, the preposition should be *of* or *nothing*), and *for* must follow *care*; *mother* depends on *for*, since *mother* follows *for*; etc. Dependency is symbolized with an arrow pointing from the governor to the dependent: *boy* ← *cares*, *care* → *for*, *for* → *mother*. It is an anti-reflexive and anti-symmetrical binary relation $\mathscr{D}$ ['depend on']; $\mathscr{D}$ holds between *A* and *B* such that:

no *A* can depend on itself: $\neg A\mathscr{D}A$, i.e., $A\mathscr{D}A$ is always false [anti-reflexivity];

if *B* depends on *A*, then *A* cannot depend on *B*: $B\mathscr{D}A$ entails $\neg A\mathscr{D}B$ [anti-symmetry].

Some syntagmatic relations are not dependencies, for instance, coreferentiality relation: in sentence (1), *boy* and *his* have the same referent; this relation is reflexive and symmetrical.

In what follows, only dependencies are considered.

2. Three Major Types of Linguistic Dependency

Speaking about dependency, people often mean just syntactic dependency. However, this is not the only dependency between wordforms of a sentence. Three types of linguistic dependency are distinguished: semantic, syntactic, and morphological (Mel'čuk 1981, 1988, p. 105).

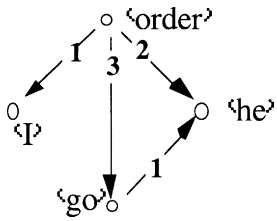

Figure 1
The SemS of the sentence
I ordered him to go

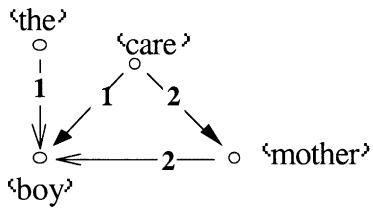

Figure 2
The SemS of the sentence
The boy cares for his mother

2.1 Semantic Dependency [Sem-$\mathscr{D}$]

2.1.1 Definition of semantic dependency The wordform $\mathbf{w}_2$ semantically depends on the wordform $\mathbf{w}_1$ in the given sentence ($\mathbf{w}_1$–**sem** → $\mathbf{w}_2$), if the meaning of $\mathbf{w}_1$ is a predicate and the meaning of $\mathbf{w}_2$ is an argument of it: '$\mathbf{w}_1$' ('$\mathbf{w}_2$').

A Sem-dependent of a wordform is called its Sem-Actant.

2.1.2 Properties of semantic dependency: As all $\mathscr{D}$s, Sem-$\mathscr{D}$ is anti-reflexive and anti-symmetrical. As for transitivity, Sem-$\mathscr{D}$ is neither transitive nor anti-transitive: in most cases, $\mathbf{w}_1$–**sem** → $\mathbf{w}_2$ and $\mathbf{w}_2$–**sem** → $\mathbf{w}_3$ entail neither $\mathbf{w}_1$–**sem** → $\mathbf{w}_3$, nor ⌜($\mathbf{w}_1$–**sem** → $\mathbf{w}_3$). Thus, from *I saw* [$\mathbf{w}_1$] *Alan's* [$\mathbf{w}_3$] *wife* [$\mathbf{w}_2$] it does not follow that I saw Alan, but it does not follow, either, that I did not. However, for some predicates, Sem-$\mathscr{D}$ is transitive: $\mathbf{w}_1$–**sem** → $\mathbf{w}_2$ and $\mathbf{w}_2$–**sem** → $\mathbf{w}_3$ entail $\mathbf{w}_1$–**sem** → $\mathbf{w}_3$, e.g., '[to] order' in *I order* [$\mathbf{w}_1$] *him* [$\mathbf{w}_3$] *to go* [$\mathbf{w}_2$] (see the Sem(antic) S(tructure) in Fig. 1). In other cases, Sem-$\mathscr{D}$ is anti-transitive: in *I wrote* [$\mathbf{w}_1$] *down Alan's* [$\mathbf{w}_3$] *address* [$\mathbf{w}_2$], $\mathbf{w}_1$–**sem** → $\mathbf{w}_2$ and $\mathbf{w}_2$–**sem** → $\mathbf{w}_3$ entail ⌜($\mathbf{w}_1$–**sem** → $\mathbf{w}_3$).

Sem-$\mathscr{D}$s are *labeled*: a Sem-$\mathscr{D}$ arc carries a symbol identifying the corresponding Sem-Actant.

Sem-$\mathscr{D}$ does not presuppose the uniqueness of the governor: a wordform can semantically depend simultaneously on several other wordforms, as shown by Figs. 1 and 2.

Sem-$\mathscr{D}$ is universal: it is present in all languages, appears in all sentences of a language, and embraces all full wordforms of a sentence. In a sentence, Sem-$\mathscr{D}$s form a connected structure (there is a Sem-'path' between any wordform and any other wordform): a network (see Figs. 1 and 2).

2.2 *Syntactic Dependency* [Synt-$\mathscr{D}$]

Sem-$\mathscr{D}$ is mentally perceived by the speakers, and Morph-$\mathscr{D}$ is perceived physically; but Synt-$\mathscr{D}$, which mediates between them, is not directly perceivable. It is an abstract construct that must be specified by some formal criteria.

2.2.1 Criteria for Synt-$\mathscr{D}$: To establish a Synt-$\mathscr{D}$ relation between wordforms $\mathbf{w}_1$ and $\mathbf{w}_2$ in a sentence three groups of criteria are needed: (a) criteria for Synt-connectedness (= for the presence of a Synt-$\mathscr{D}$); (b) criteria for Synt-dominance (= for the orientation of the Synt-$\mathscr{D}$); and (c) criteria for the specific type of the Synt-$\mathscr{D}$.

(a) *Criteria A: Synt-Connectedness*

For $\mathbf{w}_1$ and $\mathbf{w}_2$ to be syntactically directly linked ($\mathbf{w}_1$–**synt**–$\mathbf{w}_2$), both Criteria A1 and A2 must be satisfied.

Criterion A1: Linear arrangement of wordforms

Wordforms $\mathbf{w}_1$ and $\mathbf{w}_2$ have a direct Synt-$\mathscr{D}$ link only if the linear position of one must be specified with respect to the other.

'The position of $\mathbf{w}_1$ is specified with respect to $\mathbf{w}_2$' means that $\mathbf{w}_1$ precedes $\mathbf{w}_2$, follows it, or else can precede or follow it (optionally or under some conditions). If the manner adverbial can precede or follow the verb it modifies, the possible positions of the adverbial are still specified with respect to the verb, rather than the other way around.

Criterion A2: Potential prosodic unity

Wordforms $\mathbf{w}_1$ and $\mathbf{w}_2$ can have a direct Synt-$\mathscr{D}$ link only if:

(1) Either $\mathbf{w}_1$ and $\mathbf{w}_2$ can form a phrase in $\mathscr{L}$ [general case]:

e.g., N + V, N + ADJ, V + N, PREP + N, ADV + ADJ, NUM + N, etc.

(2) Or the $\mathbf{w}_1$ and $\mathbf{w}_2$ cannot form a phrase, but $\mathbf{w}_1$, $\mathbf{w}_2$ and a set of wordforms $\{\mathbf{w}_i\}$ can, such that in this phrase, $\mathbf{w}_1$ is the Synt-head and $\mathbf{w}_2$ and $\{\mathbf{w}_i\}$ also form a phrase in which $\mathbf{w}_2$ is the Synt-head [special case]:

e.g., *escape* [$\mathbf{w}_1$] *from* [$\mathbf{w}_2$] *many problems* [the set $\{\mathbf{w}_i\}$], where *escape from many problems* is a phrase, with *escape* as the Synt-head; *from many problems* is also a phrase in which *from* is the Synt-head; therefore, *escape*–**synt** → *from*.

(b) *Criteria B: Synt-Dominance*

In a phrase $\mathbf{w}_1$–**synt**–$\mathbf{w}_2$, one wordform is the Synt-governor of the other (and the Synt-head of the phrase). The Synt-governor is the wordform that determines to a greater extent than the other wordform (its Synt-dependent) properties of the phrase: syntactic, morphological, and semantic.

Criterion B1: The Synt-valency of the phrase

The passive (Synt-)valency of a phrase is the set of syntactic constructions of $\mathscr{L}$ in which this phrase can appear.

In the phrase $\mathbf{w}_1$–**synt**–$\mathbf{w}_2$, the wordform $\mathbf{w}_1$ is the Synt-governor of $\mathbf{w}_2$ if the passive valency of the whole phrase is determined to a greater extent by the valency of $\mathbf{w}_1$ than by that of $\mathbf{w}_2$.

The Synt-head of a phrase determines the possible external syntactic links of the phrase more than any other of its wordforms.

(i) The passive valency of the phrase *for decades* is fully determined by the preposition; therefore, *for*–**synt** → *decades*.

(ii) A phrase like *has escaped* or *does not escape* plays the same Synt-role as *has/does* rather than that of *escaped* or *escape*; therefore, *has*–**synt** → *escaped*, does–**synt** → *escape*.

Criterion B2: Morphological links between the wordforms of the phrase and its external context

In the phrase $\mathbf{w}_1$–**synt**–$\mathbf{w}_2$ in which the passive valency does not allow us to establish the Synt-governor (because the valencies of $\mathbf{w}_1$ and $\mathbf{w}_2$ are identical) $\mathbf{w}_1$ is the Synt-governor of $\mathbf{w}_2$ if $\mathbf{w}_1$ controls the inflection of wordforms external to the phrase or if $\mathbf{w}_1$'s inflection is controlled by such wordforms.

The Russian phrase *jubka-štany* (literally, 'skirt-pants') does not allow for the application of Criterion B1 (both wordforms have the same passive valency); but Criterion B2 singles out *jubka* as the Synt-governor: *èt**a*** [FEM.SG] *jubka-štany byl**a*** [FEM.SG] … 'this skirt-pants was …', where the external agreement is with *jubka* [FEM.SG], and not with *štany* [PL] ⟨**èt**i*** *jubka-štany byl**i*** …⟩; therefore, *jubka*–**synt** → *štany*.

Criterion B3: The semantic content of the phrase

In the phrase $\mathbf{w}_1$–**synt**–$\mathbf{w}_2$ in which neither the Synt-valency nor the morphology allows us to establish the Synt-governor $\mathbf{w}_1$ is the Synt-governor of $\mathbf{w}_2$ if $\mathbf{w}_1$–**synt**–$\mathbf{w}_2$ means 'a kind/instance of $\mathbf{w}_1$' rather than 'a kind/instance of $\mathbf{w}_2$.'

(i) In *jam sandwich*, the Synt-governor is *sandwich*, because 'jam sandwich refers to a kind of sandwich, rather than to a kind of jam' (Hudson 1990).

(ii) In Chinese *shí bàng ròu* 'ten pounds [of] meat', *ròu* 'meat' is the Synt-governor: *shí bàng ròu* refers to meat, not to pounds; therefore, *shí* ← **synt**–*bàng* ← **synt**–*ròu* [both wordforms *bàng* and *ròu* are nouns, and Chinese has no inflection].

The Synt-governor is thus the syntactic category determinant, or (if there is no syntactic determinant) the morphological behavior determinant, or (if both syntactic and morphological determinants are absent) the semantic content determinant (Zwicky 1993, p. 4).

Most approaches dealing with Synt-$\mathscr{D}$s require concord between these three properties, i.e., between Criteria B1–B3. In sharp contrast, we do not expect such a concord. Only Criterion B1 is genuinely syntactic; B2 is morphological, and B3 semantic. And it is known that the orientations of Sem-$\mathscr{D}$, Synt-$\mathscr{D}$ and Morph-$\mathscr{D}$ between two wordforms can differ (as shown in Sect. 3); therefore, these criteria must be expected often to be in conflict. Criteria B1–B3 form a hierarchy: B1 > B2 > B3. If B1 is applicable, its indication is sufficient. Only if it is not applicable, B2 applies, but only in $\mathscr{L}$ having inflection and only for $\mathbf{w}_1$ and $\mathbf{w}_2$ with different morphological properties. Otherwise, B3 applies. Therefore, Criteria B are never applied simultaneously and thus cannot contradict each other. For Criteria B to be satisfied, at least one of Criteria B1–B3 must be satisfied.

The criteria for the orientation of Synt-$\mathscr{D}$ ('Head-vs.-Dependent' problem) are discussed in Zwicky 1985, 1993, Hudson 1987, 1990, pp. 106–7, Mel'čuk 1988, and in Corbett et al. 1993.

(c) *Criteria C: Labeled Synt-Dependencies*

For each pair of wordforms $\mathbf{w}_1$–**synt** → $\mathbf{w}_2$ which are syntactically linked in a particular direction, the type of the Synt-$\mathscr{D}$ relation that links them must be established. In order to represent the SyntS of utterances, different types of Synt-$\mathscr{D}$ are needed. Thus, I ← **synt**–LOVE–**synt** → YOU does not distinguish between 'I love you' and 'You love me'; [JOHN] I ← **synt**–SEND–**synt** → YOU can be implemented as 'John sends me to you' or 'John sends you to me'; etc. In all such cases, different types of Synt-$\mathscr{D}$s, or different Syntactic Relations [SyntRels], are needed:

I ← $\mathbf{r}_1$–LOVE–$\mathbf{r}_2$ → YOU, [JOHN] I ← $\mathbf{r}_2$–SEND–$\mathbf{r}_3$ → YOU, …

SyntRels are *labeled*, and the label **r** of a SyntRel refers to a family of syntactic constructions that implement the SyntRel **r**. Thus, consider the label '**subj(ectival**' of the SyntRel that appears in Russian phrases V_{fin} [$\mathbf{w}_1$]–**subj** → N [$\mathbf{w}_2$] (*Mal'čik prixodit/ prišël* vs. *Devočka prixodit/prišla* 'The boy/girl comes/ came'). This label identifies a set of Synt-rules that make the verb $\mathbf{w}_1$ agree with the noun $\mathbf{w}_2$ in person and number (in the present/future) and in number and gender (in the past/subjunctive); these rules also position $\mathbf{w}_2$ with respect to $\mathbf{w}_1$.

In phrases $\mathbf{w}_1$–**r[?]** → $\mathbf{w}_2$, the Synt-$\mathscr{D}$ that links the two wordforms can be labeled **r** only if it satisfies the following three criteria C; if one of them is not satisfied, the presumed SyntRel **r[?]** must be split in different SyntRels.

Criterion C1: Absence of semantic contrast [**w**(L) is a wordform of lexeme L]

A SyntRel must not describe two phrases

$\mathbf{w}_i(L_1)$–**synt** → $\mathbf{w}_j(L_2)$ and $\mathbf{w}_m(L_1)$–**synt** → $\mathbf{w}_n(L_2)$ which contrast semantically ['$\mathbf{w}_i(L_1) \rightarrow \mathbf{w}_j(L_2)$' ≠ '$\mathbf{w}_m(L_1) \rightarrow \mathbf{w}_n(L_2)$'] and differ formally by purely syntactic means of expression (i.e., by word order, syntactic prosody, or syntactic grammemes).

The construction Rus. DESJAT′ ← **r[?]**–DOLLAR has two different implementations with different meanings:

desjat′ dollarov '10 dollars' vs. *dollarov desjat′* 'approximately 10 dollars';

the formal difference here is purely syntactic: word order; therefore, the presumed SyntRel **r** must be split in two different SyntRels:

DESJAT′ ← **quantitative**–DOLLAR ⇔ *desjat′ dollarov*

DESJAT′ ← **approx-quantit**– DOLLAR ⇔ *dollarov desjat′*

Criterion C1 corresponds to the '*minimal pair test*,' widely used in linguistics.

Criterion C2: Syntactic substitutability

Substitutability of syntactic subtrees for establishing SyntRels was proposed by Kunze (1975, p. 235) and is known as the *Kunze property*. Here, a weaker version of it is used: the *quasi-Kunze property*.

The quasi-Kunze property [$\Delta_{(\Xi)}$ is a subtree with the top node of the class Ξ]

A SyntRel **r** has the quasi-Kunze property if, and only if, there exists in $\mathscr{L}$ a part of speech X such that for any Synt-configuration L–**r** → $\Delta_{(Y)}$, replacing $\Delta_{(Y)}$ by $\Delta_{(x)}$ (but not necessarily vice versa!) in any Synt-structure does not affect its syntactic well-formedness.

Any SyntRel **r** must possess the quasi-Kunze property.

In phrases *have*–**r[?]** → *been* and *be*–**r[?]** → *going* the presumed SyntRel **r** does not possess the quasi-Kunze property: **have*–**r[?]** → *going* and **be*–**r[?]** → *been*; therefore, there are two different SyntRels:

HAVE–**perfect-auxiliary** → BEEN and BE–**progressive-auxiliary** → GOING

Criterion C2 is a particular case of what is known as the '*substitution test*:' it deals with the substitution of subtrees that hang on the same SyntRel **r**.

Criterion C3: Repeatability with the same Synt-governor

Any **r** of language $\mathscr{L}$ can be either non-repeatable or unlimitedly repeatable.

Non-repeatable SyntRels

A SyntRel **r** is non-repeatable if, and only if, no more than one branch labeled **r** can start from any governor.

Actantial SyntRels, which describe the Subjects, the Objects and the Complements, are non-repeatable. (In some languages, an actant can be duplicated by a resumptive pronoun, but such duplication is not considered as repeatability: it has a completely grammaticalized character.)

(Unlimitedly) Repeatable SyntRels

A SyntRel **r** is (unlimitedly) repeatable if, and only if, several branches labeled **r** can start from a governor.

The **modificative** and **circumstantial** SyntRels in English are unlimitedly repeatable; so is the **obl(ique)-**

obj(ectival) SyntRel. For a repeatable SyntRel **r** the number of branches labeled **r** starting from a governor is theoretically unlimited, although practically it is limited by pragmatic considerations or the properties of governors (e.g., by their Government Pattern, with the **obl-obj** SyntRel).

Criterion C3

Any SyntRel **r** must be unlimitedly repeatable or non-repeatable.

In Persian, expressions of the following type are extremely widespread:

Rāmin + rā ← **r[?]**–*kard*–**r[?]** → *bedār*,

Ramin DirO made awakening [Noun]

'[He] awoke Ramin.'

They include verbal collocations like *bedār kard* 'awakening make' = 'awake' or *dars dad*, (literally, 'lesson give' = 'teach') and seem to contain a DirO (*bedār* or *dars*), but behave as transitive verbs and take a 'genuine' DirO (with the suffix **-rā**). The presumed SyntRel **r[?]** in such expressions would be limitedly repeatable—just twice. Therefore, there are two different SyntRels:

RĀMIN ← **dir-obj**–KARD–**quasi-dir-obj** → BEDĀR.

Criterion C3 corresponds to the '*co-occurrence test*:' in morphology, an element of a morphological category is either non-repeatable (tense or number in English or French) or unlimitedly repeatable (the causative in Turkish).

The SyntRels of a language form a systematic inventory, just like phonemes or grammemes; Criteria C1–C3 are part of a methodology for establishing SyntRels' inventories.

Now the definition of Synt-$\mathscr{D}$ can be formulated:

The wordform $\mathbf{w}_2$ *syntactically depends* on the wordform $\mathbf{w}_1$ via SyntRel **r** ($\mathbf{w}_1$–**r** → $\mathbf{w}_2$) if Criteria A, B and C are satisfied for this pair of wordforms and **r**.

2.2.2 Properties of syntactic dependency: Like Sem-$\mathscr{D}$, Synt-$\mathscr{D}$ is anti-reflexive and anti-symmetrical: ⌜($\mathbf{w}_1$–**synt** → $\mathbf{w}_1$); $\mathbf{w}_1$–**synt** → $\mathbf{w}_2$ entails ⌜ ($\mathbf{w}_1$ ← **synt**–$\mathbf{w}_2$). Anti-symmetry means that $\mathbf{w}_1$ cannot be the Synt-governor of $\mathbf{w}_2$ and simultaneously have $\mathbf{w}_2$ as its own Synt-governor. In addition, Synt-$\mathscr{D}$ is anti-transitive: $\mathbf{w}_1$–**synt** → $\mathbf{w}_2$ and $\mathbf{w}_2$–**synt** → $\mathbf{w}_3$ (in one sentence) entail ⌜ ($\mathbf{w}_1$–**synt** → $\mathbf{w}_3$). This fact underlies the principle of unique Synt-governor, see below.

Synt-$\mathscr{D}$ is labeled: to properly represent *Mary loves John*, in the phrases *Mary* ← $\mathbf{r}_1$–*loves* and *John* ← $\mathbf{r}_2$–*loves* the SyntRels $\mathbf{r}_1$ and $\mathbf{r}_2$ must be different; otherwise, the semantic contrast will not be preserved in the Syntactic Structure [SyntS].

Synt-$\mathscr{D}$ presupposes the uniqueness of the governor: a wordform can syntactically depend only on one other wordform (or it can be independent, as is the top node of a SyntS).

Synt-$\mathscr{D}$ is universal: it is present in all languages, appears in all sentences of a language, and embraces all wordforms of a sentence. In a sentence, Synt-$\mathscr{D}$s form a connected structure: a (dependency) tree.

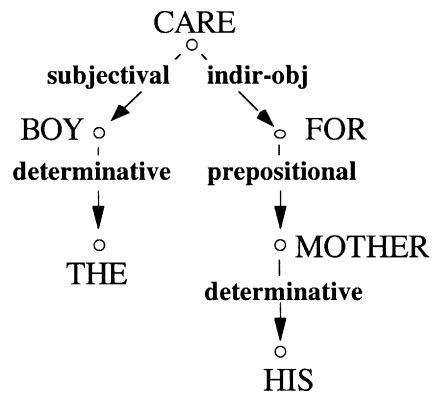

Figure 3
SyntS of sentence (1)

A dependency tree is a connected graph in which each node but one depends on only one node (= the uniqueness of the Synt-governor), while one and only one node does not depend on anything—the top node, or the root, of the SyntS. The linear order of the nodes in a SyntS is not defined; the $\mathscr{D}$-description consistently separates the Synt-links between wordforms and their linear order. (During sentence production, word order is computed by syntactic rules of $\mathscr{L}$ on the basis of Synt-$\mathscr{D}$s.) See the SyntS of sentence (1) in Fig. 3.

2.3. *Morphological Dependency* [*Morph–$\mathscr{D}$*]

2.3.1 A definition of morphological dependency: The wordform $\mathbf{w}_2$ morphologically depends on the wordform $\mathbf{w}_1$ according to the inflectional category C ($\mathbf{w}_1$–**morph/C** → $\mathbf{w}_2$) if at least one grammeme **g** of $\mathbf{w}_2$, such that $\mathbf{g} \in$ C, is determined by $\mathbf{w}_1$.

2.3.2 Properties of morphological dependency: Like Sem-$\mathscr{D}$ and Synt-$\mathscr{D}$, Morph-$\mathscr{D}$ is anti-reflexive: the inflection of a wordform cannot be determined by the wordform itself. But in contrast to Sem-$\mathscr{D}$ and Synt-$\mathscr{D}$, Morph-$\mathscr{D}$ is neither symmetrical, nor anti-symmetrical. (This does not contradict the notion of dependency because the anti-symmetry of Morph-$\mathscr{D}$ is violated only for different categories $C_1 \neq C_2$.) For agreement of ADJ with an N or for government of the case of N by V/PREP, Morph-$\mathscr{D}$ is anti-symmetrical: $\mathbf{w}_1$–**morph** → $\mathbf{w}_2$ entails ⌜ ($\mathbf{w}_1$ ← **morph**-$\mathbf{w}_2$). However, in some cases Morph-$\mathscr{D}$ is symmetrical (i.e., reciprocal): $\mathbf{w}_2$ is inflected, for the category C_1, as a function of $\mathbf{w}_1$, and, at the same time, $\mathbf{w}_1$ is inflected, for a different category C_2, as a function of $\mathbf{w}_2$, so that $\mathbf{w}_1$–**morph** → $\mathbf{w}_2$ entails $\mathbf{w}_1$ ← **morph**–$\mathbf{w}_2$; as a result $\mathbf{w}_1$ ← **morph** → $\mathbf{w}_2$ obtains, with $C_1 \neq C_2$.

In Russian:
dv + ***a*** *stol* + ***a*** 'two tables'
two MASC.NOM table(MASC) SG.GEN
The noun *stola* morphologically depends for its singular and genitive case on *dva*, while *dva* depends on *stola* for its masculine gender; for example, *pjat′ stol* + ***ov*** [PL.GEN] 'five tables' or *dvadcat′ odin stol* + ∅ [SG.NOM] (literally, 'twenty-one table'), but *dv* + ***e*** [FEM.NOM] *krovati* 'two beds'.

Like Sem-$\mathscr{D}$ and Synt-$\mathscr{D}$, Morph-$\mathscr{D}$ must be *labeled*: to fully specify a Morph-$\mathscr{D}$, we have to indicate the inflectional category C of $\mathbf{w}_2$ whose grammeme is imposed by $\mathbf{w}_1$.

Morph-$\mathscr{D}$ does not presuppose the uniqueness of the governor: a wordform can morphologically depend simultaneously on several wordforms—for different inflectional categories.

Morph-$\mathscr{D}$ is not universal: it is completely absent from many languages (e.g., Vietnamese); in a language where it exists it is not present in all sentences; and in a sentence where it is present it does not necessarily embrace all wordforms. Morph-$\mathscr{D}$s do not form, generally speaking, a connected structure in a sentence.

3. *Possible Combinations of the Three Types of Linguistic Dependency*

The semantic, syntactic, and morphological dependencies are logically autonomous and can co-occur in fourteen logically possible combinations (Mel'čuk 1981, 1988, pp. 118–28, Nichols 1986).

CASE 1: $\mathbf{w}_1$ $\mathbf{w}_2$, i.e., no syntagmatic dependency between two wordforms. Example: *the* and *mother* in sentence (1).

CASE 2: $\mathbf{w}_1$**–sem→** $\mathbf{w}_2$, i.e., Sem-$\mathscr{D}$, unsupported by any Synt-$\mathscr{D}$ or Morph-$\mathscr{D}$.

Example: In *Alan had a drink* the wordforms *Alan* and *drink* are semantically directly related, ['drink'–**sem→** 'Alan'], yet there is no Synt-$\mathscr{D}$ or Morph-$\mathscr{D}$ between them.

CASE 3: $\mathbf{w}_1$**–synt→** $\mathbf{w}_2$, i.e., Synt-$\mathscr{D}$ without any Sem-$\mathscr{D}$ or Morph-$\mathscr{D}$.

(a) In Japanese, a quantifier, while bearing semantically on the Subject as in '**Five** people were injured', depends syntactically on the verb, with which it has neither semantic nor morphological links:
Sono ziko + *de* *keganin* + *ga*
this accident LOC injured.people SUBJ(ective)
GO + NIN *deta*
five CLASS(ifier) emerged
'In this accident, five people were injured.'

(b) In English, a measure noun used as a DirO depends syntactically on the verb, but has no semantic or morphological link with it (semantically, the verb dominates the noun quantified):
John bought–**synt** → [*five*] ***kilos*** [*of potatoes*].

CASE 4: $\mathbf{w}_1$**–morph→** $\mathbf{w}_2$, i.e., Morph-$\mathscr{D}$, without any Sem-$\mathscr{D}$ or Synt-$\mathscr{D}$.

In Alutor, a verb of perception that governs a DirO clause can show object agreement either with the Subject or with the DirO of this clause (depending on their communicative roles):
(a) *Qəmavə* + *nak* *na* + *laʕu* + *tkəni* + *γət*
Qamav SG.LOC 3 SG.SUB see PRES 2SG.OBJ
γən + *annə*
you.SG INSTR
∅ + *kəlγatətkə* + *na* + *wwi* *qura* + WWI
2SG.SUB harness 3.OBJ PL reindeer PL
'Qamav sees that YÓU are harnessing the reindeer.'
(b) *Qəmavə* + *nak* *π* + *laʕu* + *tkəni*
Qamav SG.LOC 3 SG.SUB see PRES
+ NINA + WWI
3.OBJPL
γən + *annə* ∅ + *kəlγatətkə* + *na*
you.SG INSTR 2SG.SUB harness 3.OBJ
+ *wwi* *qura* + WWI
PL reindeerPL
'Qamav sees that YÓU are harnessing the REINDEER'.

CASE 5: $\mathbf{W_1} \xrightarrow{\textbf{Sem}} \mathbf{W_2}$, $\mathbf{W_1} \xrightarrow{\textbf{Synt}} \mathbf{W_2}$ i.e., Sem-$\mathscr{D}$ and Synt-$\mathscr{D}$ oriented the same way, with no Morph-$\mathscr{D}$.

Example: objects/complements in caseless languages, cf. *for* and *mother* in (1).

CASE 6: $\mathbf{W_1} \xrightarrow{\textbf{Sem}} \mathbf{W_2}$, $\mathbf{W_1} \xleftarrow{\textbf{Synt}} \mathbf{W_2}$ i.e., Sem-$\mathscr{D}$ and Synt-$\mathscr{D}$ oriented the opposite ways, again with no Morph-$\mathscr{D}$.

Example: an adjective and the modified noun in ADJ + N phrase in a language with invariable adjectives, as in *new* and *areas* in *new areas*.

CASE 7: $\mathbf{W_1} \xrightarrow{\textbf{Sem}} \mathbf{W_2}$, $\mathbf{W_1} \xrightarrow{\textbf{Morph}} \mathbf{W_2}$ i.e., Sem-$\mathscr{D}$ and Morph-$\mathscr{D}$, oriented the same way, with no Synt-$\mathscr{D}$.

Example: a pronominal clitic which semantically and morphologically depends on an infinitive, while syntactically it depends on a higher verb, which 'hosts' it cf.:

Sp. ***Le*** ← **synt**–*quisiera poder enviar este libro*, [literally, 'To-him [I] would-like to-be-able to-send this book.'

CASE 8: $\mathbf{W_1} \xrightarrow{\textbf{Sem}} \mathbf{W_2}$, $\mathbf{W_1} \xleftarrow{\textbf{Morph}} \mathbf{W_2}$ i.e., Sem-$\mathscr{D}$ and Morph-$\mathscr{D}$, oriented the opposite ways, with no direct Synt-$\mathscr{D}$.

Example: an attributive adjective as in French, cf.:
Elle semblait ***fatiguée*** 'She seemed tired';

here, semantically *elle* depends on *fatiguée* ['fatigué'–**sem** → 'elle'], but morphologically *fatiguée* depends on *elle* for its singular and feminine; syntactically, the two are not directly linked.

CASE 9: 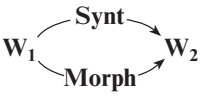

i.e., Synt-$\mathscr{D}$ and Morph-$\mathscr{D}$ oriented the same way, with no Sem-$\mathscr{D}$.

(13) A measure noun as a DirO in a language with cases depends syntactically and morphologically on the verb, but does not have a semantic link with it.

Example: in Russian: *Ivan kupil* ***tonnu*** *kirpičej* 'Ivan bought a ton of bricks'.

CASE 10: 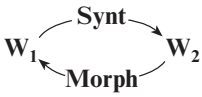

i.e., Synt-$\mathscr{D}$ and Morph-$\mathscr{D}$ oriented the opposite ways, again with no Sem-$\mathscr{D}$.

Example: a phasal verb syntactically dominates its Subject, but morphologically depends on it (= agrees with it in person and number), while there is no Sem-$\mathscr{D}$ between them, because the Subject semantically depends on the lexical verb, as in *Alan* ***begins*** *to read.*

CASE 11: 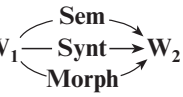

i.e., all three types of dependency oriented the same way.

Example: a verb and its nominal object in a language with cases, as in the Russian *ljubit Mašu* '[He] loves Mary'.

CASE 12: 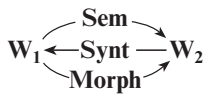

i.e. Sem-$\mathscr{D}$ and Morph-$\mathscr{D}$ oriented the same way, while Synt-$\mathscr{D}$ goes in the opposite direction.

Example: a negative particle and the verb in Arabic, cf.:

	ya+ktub+u	'writes'

vs.

la	*ya+ktub+u*	'does not write'
	INDICATIVE	
	katab+a	'wrote'

vs.

lam	*ya+ktub+∅*	'did not write'
	JUSSIVE	
	sa+ya+ktub+u	'will write'
lan	*ya+ktub+**a***	'will not write'
	SUBJUNCTIVE	

Semantically, the negative particle bears on the verb and morphologically controls its form (*lam* requires the jussive and *lan*, the subjunctive; *la* is neutral); syntactically, it depends on the verb.

CASE 13: 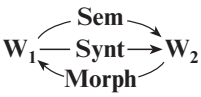

i.e., Sem-$\mathscr{D}$ and Synt-$\mathscr{D}$ oriented the same way, while Morph-$\mathscr{D}$ is opposite.

Example: a verb and its nominal actants in a caseless language with polypersonal agreement of the MV; cf. e.g., Abkhaz:

Sara Nadš'a ***i+l+əs+teiť*** *aš^w q'^w ə*
I Nadsha it her I gave book
'I gave Nadsha [a] book.'

CASE 14: 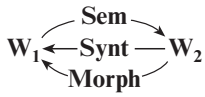

i.e., Synt-$\mathscr{D}$ and Morph-$\mathscr{D}$ are oriented the same way, while Sem-$\mathscr{D}$ is opposite.

Example: an adjective and the modified noun in a language with adjectival agreement, where the adjective bears semantically on the noun, but syntactically and morphologically depends on it.

4. *Syntactic Dependency and Syntactic Constituency*

A systematic comparison of the $\mathscr{D}$-approach with the constituency, or $\mathscr{C}$-, approach is impossible here (see Vennemann 1977, Hudson 1980a, 1980b, Dahl 1980, Matthews 1981, pp. 71–95, Mel'čuk 1988, pp. 13–17, Sgall and Panevova 1988–89, Rambow and Joshi 1997). Only two points will be touched upon.

4.1 *Syntactic Dependency and Word Order*

There are several linguistic phenomena for whose description Synt-$\mathscr{D}$ is crucial: valency, voice, restricted lexical cooccurrence, ellipses, etc. However, Synt-$\mathscr{D}$ is especially convenient for the description of word order: it forces strict separation of the hierarchical (= genuinely syntactic) 'order' from the linear order, the latter being a surface means for the expression of the former. The main task of syntax—linearizing a two-dimensional Synt-structure (Tesnière 1959, pp. 19–20)—can be solved more easily with Synt-$\mathscr{D}$s than in any other way:

(a) Word order rules can be naturally formulated in terms of positioning a Synt-dependent with respect to its Synt-governor (before or after it).

(b) Synt-$\mathscr{D}$ has allowed for the discovery (Lecerf 1960, Iordanskaja 1963, Hays 1964, Robinson 1970) of an important property of word order: projectivity. If a sentence is supplied with its $\mathscr{D}$-SyntS and a perpendicular is drawn from each wordform to the

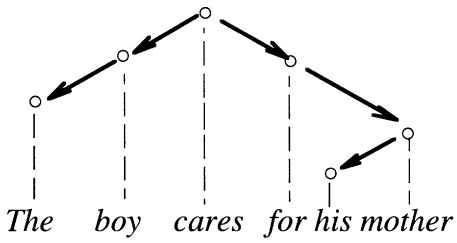

Figure 4
Sentence (1) supplied with its Synt-Structure

node that represents it, then no branch of the SyntS intersects with another branch or with a projection perpendicular. Sentence (1) is projective:

A sentence appears as a 'projection' of its SyntS; hence the name 'projectivity.'

The majority of sentences in any language are projective. This fact ensures a more elegant formulation of word-order rules and facilitates computer processing of texts: except for particular cases, only projective sentences must be produced from a given SyntS, and only SyntSs that guarantee projectivity must be associated with a given sentence.

However, projectivity can be systematically violated in many special cases, for instance:

All non-projective constructions (optional or obligatory) of ℓ must be specified in the syntactic description. Within the $\mathscr{D}$-approach this poses no problem whatsoever (while the $\mathscr{C}$-approach is unable to represent non-projective constructions without some additional machinery—e.g., transformations).

The $\mathscr{D}$-approach has the inherent ability to accommodate long-range links. The perturbations introduced into the word order of a sentence by its communicative structure—frontings, extractions, postponings, etc., in such languages as German or Russian—can wreak havoc on a $\mathscr{C}$-structure, since even the closest-knit phrases can be torn apart and permuted. But a $\mathscr{D}$-structure, completely free of linearity and contiguity, is insensitive to such permutations. As a result, the $\mathscr{D}$-approach does not face problems in representing discontinuities, which, in the literal sense of the word, simply do not exist in a $\mathscr{D}$-structure.

Dealing with the above phenomena is related to the trend in modern linguistics that Hudson (1990) aptly dubbed *lexicalism*: putting at the center of the description facts about lexical units rather than facts about syntactic constructions, so that the lexicon is given a place of honor. Stressing the importance of the lexicon goes quite well with the $\mathscr{D}$-approach in syntax, because in this approach all the links are established between wordforms and based, in the final analysis, on their lexicographic properties.

4.2 Crossing Syntactic Dependency with Syntactic Constituency

For years, linguists have been looking for a hybrid between $\mathscr{D}$- and $\mathscr{C}$-representations (e.g., Vennemann 1977). The incentive for integration comes primarily from such phenomena as coordination (see Sect. 5), extractions (*I know* ***which girl*** *you told my wife Alan*

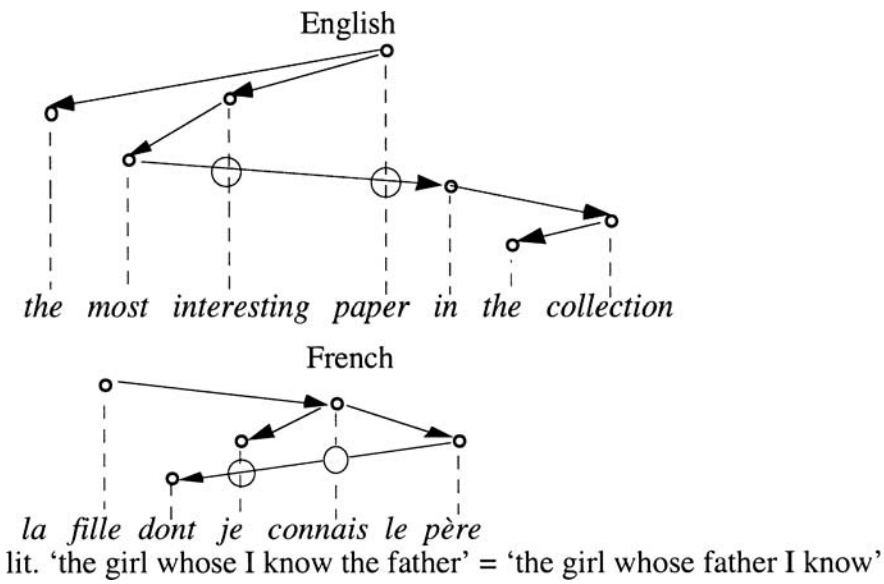

Figure 5
Non-Projective Synt-Structures

was going out with); analytical forms (*has been detected*; *the book*); collocations (*make headway, pay a visit, launch an attack*); and the like. The main idea is to introduce for any of these constructions a special type of subtree that is allowed to occupy—as a whole—one node of a $\mathscr{D}$-tree.

The first formal proposal for a 'mixed' $\mathscr{D}$-/$\mathscr{C}$-representation—by means of syntactic groups—was advanced by Gladkij (1966). The most recent one was made by Kahane (1997), in which he introduces the concept of bubble: a subset of nodes of a $\mathscr{D}$-tree which is treated as a node, while having inside a completely specified $\mathscr{D}$-structure of its own, including other bubbles.

5. *Insufficiency of Syntactic Dependency: Coordination*

If we agree to use two levels of representation for syntactic structures (Deep-SyntS and Surface-SyntS) plus communicative structures on each level, then a pure $\mathscr{D}$-representation in syntax seems to be sufficient for all syntactic phenomena, except for some types of coordination (see Hudson 1990, p. 97). A wordform **w** may 'relate' either to a whole conjoined phrase or to its Synt-head alone, such that the two constructions are formally distinct and have different meanings; however, within the strict $\mathscr{D}$-approach, both types of structure can be represented only by the direct Synt-$\mathscr{D}$ of **w** on the Synt-head of the conjoined phrase (the pure $\mathscr{D}$-formalism does not allow for the dependency on a phrase as a whole). Consequently, one SyntS corresponds in such cases to two different meanings:

The SyntS *old* ← *men* → *and* → *women* has:
—either a joint reading [the adjective bears on both nouns]:
'old {men + women}' = '{old men} and {old women}';
—or a disjoint reading [the adjective bears on one noun only]:
'{old men} and women [who are not necessarily old]'.

Different implementations that formally distinguish intended meanings are in principle available, depending on the language and particular lexical means used; for instance ('‖' stands for a pause): *old men and women* [without a pause] vs. *old men* ‖ *and women*. The semantic contrast accompanied by a formal contrast requires that the semantic distinction be maintained in the SyntS (Criterion C1). The only way to do so is to complement the $\mathscr{D}$-approach with groupings—specification, within the SyntS, of $\mathscr{D}$-subtrees relevant in such cases (Mel'čuk 1974, pp. 214–16, 1988, pp. 28–33): *old* ← ***modif***–*men* → *and* → *women* (without grouping) stands for the disjoint reading ('{old men} and women'), but old ← ***modif***–[–men → and → women] (with a grouping indicated by square brackets) for the joint reading ('old men and old women'). A grouping is not a constituent: there is no higher node to represent it as a whole (it is not a projection of a complete subtree), and it does not participate in $\mathscr{D}$-links as such (in a consistent $\mathscr{D}$-approach, only single nodes do). The use of groupings makes the $\mathscr{D}$-formalism sufficient for all syntactic purposes.

For a general view of the main topics related to Synt-$\mathscr{D}$, see Heringer 1993, Hudson 1993a, 1993b; a description of German syntax in terms of Synt-$\mathscr{D}$ is offered in Engel 1977 and Heringer 1996, and of English syntax in Mel'čuk and Pertsov 1987.

See also: Morphology in Linguistics; Syntax

Bibliography

Corbett G, Fraser N, McGlashan S (eds.) 1993 *Heads in Grammatical Theory*. Cambridge University Press, Cambridge

Dahl Ö 1980 Some arguments for higher nodes in syntax: A reply to Hudson's 'Constituency and Dependency.' *Linguistics* **18(5/6)**: 485–8

Engel U 1977 *Syntax der deutschen Gegenwartssprache*. 3rd edn. Schmidt, Berlin

Gladkij A 1966 *Lekcii po matematičeskoj lingvistike*. Novosibiisk University, Novosibirsk, Russia

Hays D 1964 Dependency theory: A formalism and some observations. *Linguistics* **40(4)**: 511–25

Heringer H J 1993 Dependency syntax: Basic ideas and the classic model. In: Jacobs J, von Stechow A, Sternefeld W, Vennemann T (eds.) *Syntax. An International Handbook of Contemporary Research*. Vol. 1. De Gruyter, Berlin and New York, pp. 298–328

Heringer H J 1996 *Deutsche Syntax Dependentiell*. Stafenburg, Tübingen, Germany

Hudson R 1980a Constituency and dependency. *Linguistics* **18(3/4)**: 179–98

Hudson R 1980b A second attack on constituency: A reply to Dahl. *Linguistics* **18(5/6)**: 489–504

Hudson R 1987 Zwicky on heads. *Journal of Linguistics* **23**: 109–32

Hudson R 1990 *English Word Grammar*. Blackwell, Oxford, UK

Hudson R 1993a Do we have heads in our minds? In: Corbett G, Fraser N, McGlashan S (eds.) *Heads in Grammatical Theory*. Cambridge University Press, Cambridge, pp. 266–91

Hudson R 1993b Recent developments in dependency theory. In: Jacobs J, von Stechow A, Sternefeld W, Vennemann T (eds.) *Syntax. An International Handbook of Contemporary Research*. De Gruyter, Berlin and New York, Vol. 1, pp. 329–38

Iordanskaja L 1963 O nekotoryx svojstvax pravil′noj sintaksiãeskoj struktury. *Voprosy jazykoznanija* no. 4, pp. 102–12

Jacobs J, von Stechow A, Sternefeld W, Vennemann T (eds.) 1993 *Syntax. An International Handbook of Contemporary Research*. De Gruyter, Berlin and New York, Vol. 1

Kahane S 1997 Bubble trees and syntactic representations. In: Becker T, Krieger H-U (eds.) *Proceedings of the 5th Meeting on Mathematics of Language*. Saarbrücken, pp. 70–6

Kunze J 1975 *Abhängigkeitsgrammatik*. Akademie-Verlag, Berlin

Lecerf Y 1960 Programme des conflits, modèle des conflits. *TA* **1(4)**: 11–18; **1(5)**: 17–36

Matthews P 1981 *Syntax*. Cambridge University Press, Cambridge

Maxwell D, Schubert K (eds.) 1989 *Metataxis in Practice. Dependency Syntax for Multilingual Machine Translation*. Foris, Dordrecht, The Netherlands
Mel'čuk I 1974 *Opyt teorii: lingvistiãeskix modelej 'Smysl–Tekst'*. Nauka, Moscow
Mel'čuk I 1981 Types de dépendance syntagmatique entre les mots-formes d'une phrase. *BSLP* **76(1)**: 1–59
Mel'čuk I 1988 *Dependency Syntax*. SUNY Press, Albany, NY
Mel'čuk I, Pertsov N 1987 *Surface Syntax of English*. Benjamins, Amsterdam
Nichols J 1986 Head-marking and dependent-marking grammar. *Lg* **62(1)**: 56–119
Rambow O, Joshi A 1997 A formal look at dependency grammars and phrase-structure grammars. In: Wanner L (ed.) *Recent Trends in Meaning-Text Theory*. Benjamins, Amsterdam, pp. 167–90
Robinson J 1970 Dependency structures and transformational rules. *Language* **46**: 259–85
Sgall P, Panevová J 1988–89 Dependency syntax—A challenge. *ThL* **15(1)**: 73–86
Tesnière L 1959 *Éléments de syntaxe structurale*. Klincksieck, Paris
Vennemann T 1977 Konstituenz und Dependenz in einigen neueren Grammatiktheorien. *Sprachwissenschaft* **2(3)**: 259–301
Zwicky A 1985 Heads. *Journal of Linguistics* **21(1)**: 1–30
Zwicky A 1993 Heads, bases and functors. In: Corbett G, Fraser N, McGlashan S (eds.) 1993 *Heads in Grammatical Theory*. Cambridge University Press, Cambridge, pp. 292–315

I. Mel'čuk

Language Development, Neural Basis of

1. *The Neurobiology of Language Processing in Adults*

The classic theory of the neural basis of language function was based on observations of adults with acquired brain injuries. Approximately 95 percent of adults who develop aphasia after stroke have damage to the left hemisphere (LH). The LH was, therefore, assumed to play a crucial role in mediating expressive and receptive language. Within the LH, Broca's area (the inferior frontal operculum) was thought to be the neural substrate for speech production and Wernicke's area (the posterior superior temporal area and supramarginal gyrus of the parietal lobe) the neural substrate for speech comprehension (Caplan 1987).

New experimental methods that assess the neural substrate of language in awake and processing individuals have revolutionized the classic theory. Language areas are not well circumscribed and are not homogeneous. Researchers have proposed that Wernicke's area is involved in phonological processing (Petersen and Fiez 1993) and may include the neurocognitive network for the written and spoken word lexicon (Howard et al. 1992). Posterior inferior temporal cortex has also been implicated in lexical and semantic aspects of language comprehension and production (Alexander 1997). Broca's area has been implicated in syntactic processing (Zurif and Swinney 1994), in verbal short-term memory (Braver et al. 1997), and in phonological processing (Demonet et al. 1992, Zatorre et al. 1992). Other areas of the brain seem to participate in language functioning. The middle-superior temporal region has been implicated in semantic processing and anterior-superior temporal area in syntactic processing (Mazoyer et al. 1993, Price et al. 1996). Moreover, the dorso-lateral prefrontal cortex (DLPFC) is active during language and reading tasks. The DLPFC may be responsible for modulation of processing in posterior semantic regions, such as Wernicke's area (Frith et al. 1991, Raichle et al. 1994) or for maintaining verbal information in memory during comprehension of complex sentences (Jonides et al. 1997).

Three lines of evidence suggest that language processing in the normal state may involve bilateral representation, at least under some circumstances. The superior temporal areas in both the LH and right hemisphere (RH) are specialized for processing auditory information (Heffner and Heffner 1984). Adult stroke patients undergo a period of recovery that leads to increased activation bilaterally during auditory and language tasks (Weiller et al. 1995, Engelien et al. 1995). Finally, fMRI studies with normal adults show that increasing task difficulty increases the magnitude of bilateral activation. Homologous areas of the RH become active as the intensity and volume of the neural response within the LH increases. This pattern has been documented for language and working memory (Braver et al. 1997, Just et al. 1996). If bilateral activation emerges during difficult cognitive tasks, these findings suggest that the relevant neural wiring for language processing may be in place in both hemispheres and utilized when a single hemisphere is unable to complete the task or complete it in a timely manner.

Though research over the last several years has documented the neural systems involved in language processing, very little is known about the processing at more molecular levels of analysis. Obviously, animal models cannot be used to consider these questions. New methods such as Magnetic Resonance Spectroscopy may provide information into energy metabolism in language processing and expand our current knowledge.

2. *Challenges in Studying the Neural Basis of Language Development*

The main question that researchers have focused on regarding the neural basis of language development has been to what extent are neural systems organized in advance of language learning? Are the involved

areas of brain, even if they are widespread and nonhomogeneous, nonetheless committed prior to language development? Or is the system organized by the very act of learning language? A secondary issue of interest has been whether the dramatic changes in language during the first few years of life are stimulated by changes in brain development, such as myelination, or whether the behavioral changes can be accounted for as the outcome of learning. At the present time this second question is addressed only by correlational studies of brain and behavioral developments and is thus highly speculative in terms of causality. This article will not address that issue.

Currently, our knowledge about the neural basis of language development is extremely limited. A major impediment to the development of the field is the paucity of methods appropriate for use in human infants and young children. In adults, neural mechanisms are investigated through a variety of different noninvasive functional imaging techniques, such as event-related potentials (ERP), positron emissions tomography (PET), and functional magnetic resonance imaging (fMRI). Each of these methods has advantages and limitations; considering them together assists in piecing together a full account of the neural basis of language function. In children, PET scanning has been avoided—except in children with clinical indications for its use—because the tests require exposure to ionizing radiation. Functional MRI involves no exposure to ionizing radiation but requires an awake and responding subject to hold still. Even minimal movement by a young child renders it impossible to detect the minimal differences in activation between baseline and experimental conditions. Children less than 7 years of age cannot usually comply with the stringent requirements of the scanning. Therefore, for infants and young children, the only direct method of functional imaging that has been used successfully is ERP.

An indirect method for addressing the neural underpinnings of language development has been to study children with early lesions impinging on classical language areas of the cortex. If, the argument goes, children with damage to the usual language areas have serious delays and disabilities in aspects of language functioning, then the implication is that the neural mechanisms for those aspects of language development were probably present prior to the injury. If the damage occurred around the time of birth and language performance is impaired, then the inference is that specialization is congenital, and possibly determined or initiated by genetic mechanisms. If children with early brain injuries successfully master language, then the implication is that, at least under extreme circumstances, alternative organizations within the human brain for language are possible.

This experimental strategy also has serious limitations. Early focal brain injury in childhood is rare and, unlike the situation in adults, the causes are diverse. Studies in the 1970s and 1980s enrolled children with hemidecortication for intractable seizures in Sturge–Weber syndrome, a vascular malformation of the brain and overlying skin (Dennis and Kohn 1975, Dennis and Whitaker 1976), and children with embolic strokes sustained during cardiac procedures to treat congenital cardiac disorders (Aram et al. 1986, Aram and Ekelman 1987). In these populations, the age of onset and extent of injury both vary. More recently, studies have enrolled children with strokes that occurred *in utero* or in the perinatal period. Though the age of onset of this population is narrow and the injury preceded any language development, the locations of injury still vary substantially. Furthermore, the study of children with early focal lesions is challenging because the language capabilities of the young infant had no injury occurred are uncertain, whereas the capacities of adults can be inferred from their previous functioning and occupation. Other variables and coexisting conditions such as seizure disorder and use of anticonvulsant medications appear to mediate outcome (Dall'Oglio et al. 1994).

3. *Evidence from Behavioral Data*

Behavioral data suggest that language development has a strong biological basis. Language development begins when a child approaches one year of age and reaches almost adult levels of competence by the time that the child enters school. The rapidity of that development, at a time when other skills—basic adding and substracting, playing chess or riding bikes—are rudimentary suggests a strong biological predisposition or preparedness. Prelinguistic infants demonstrate processing biases before they have had much exposure to language. For example, they categorically discriminate phonetic contrasts, even those that are not found in the native language and therefore not part of the linguistic environment (Eimas et al. 1971, Best et al. 1988). They become tuned to the phonetic contrasts in their native language in the second half of the first year, as their babbling approaches the sound patterns of the adult language (Best et al. 1988). These processing biases may provide the mechanisms that infants use to begin to segment and organize the sound stream to which they are exposed. Though it is tempting to assume that the mechanisms are innate, neonates also show preferences for their mother's voice, though they do not recognize their mother's face (Burnham 1993). Thus, prenatal exposure to the mother's voice *in utero*, rather than some innate bias, is likely responsible for the early preference. Interestingly, when infants start learning to map words to meaning, they confuse similar sounding words, suggesting that they are not able to capitalize immediately on their speech discrimination skills when they begin to acquire verbal language. The neural basis of language development

may involve forming connections between the neural substrates involved in speech perception and language production.

Behavioral data also suggests a strong environmental component to language development. In the area of vocabulary development, the rate of vocabulary growth is a function of the amount of maternal input the child hears (Huttenlocher et al. 1991). In the realm of syntax, children first learn features that are highly regular and prosodically prominent. Differences in the features of the adult language seem to explain some differences in the order of acquisition across languages.

4. *Evidence from ERP and fMRI*

Using ERP techniques on normal children learning language, a picture emerges that initial development of language requires a wide and bilateral neural network and that progressive neural specialization and lateralization occurs in association with language experience rather than strictly with age. Specific ERP components in toddlers 13 to 20 months of age demonstrate differences between words that the child understands and words that are incomprehensible. Word comprehension shows both age and experience components. At 13 months of age, the comprehension effects can be found over both the right and left hemisphere and over anterior and posterior regions. By 20 months of age, the effects are detectable only over the posterior left hemisphere. However, younger toddlers who are more advanced in language development show this laterality effect at a younger age than do less advanced children (Mills et al. 1997).

Data from children with early focal brain injury corroborate the view of initial bilateral involvement in language development. Children with focal injury to either hemisphere experience developmental delays in prelinguistic skills, including the onset of babbling and communicative gestures (Marchman et al. 1991) and in the onset of vocabulary development and use of word combinations in parent–child conversations (Feldman et al. 1992, Thal et al. 1991). Once the children with focal injury begin to acquire functional vocabulary and syntactic skills, their rate of developmental progress as measured by vocabulary use and mean length of utterance is comparable to each other and to children developing typically (Feldman et al. 1992). By age 4, children with focal LH damage can even master the complex morphosyntactic structures of a language like Hebrew (Levy et al. 1994). However, at school age, children with damage to either the LH or RH show delays compared with children developing typically in the creation of narrative discourse (Reilly et al. 1998). These findings are suggestive that both hemispheres participate in advanced language skills such as narrative discourse in school aged children.

Research also indicates that individual differences in developmental rate relate to the location of injury. Children ages 10 to 17 months of age with RH damage have been shown to have greater initial delays in word comprehension and production than do children with LH damage. However, at the point of word and syntax development children with LH temporal lobe injury demonstrate slower development than do those with damage to other areas of the RH or LH (Thal et al. 1991). These data are compatible with a view of increasing specialization during language development. However, it is important to note that individual differences also relate to other variables besides the size and site of the lesion, such as the presence of seizures and the use of anticonvulsant medications (Dall'Oglio et al. 1994).

Studies using fMRI in children developing typically demonstrate that predominant LH activation during language tasks is present by about age 7 to 9 years, and as with adults, bilateral activation occurs under some circumstances. The activation of neural networks on fMRI was studied during sentence processing in adults, normal school aged children and also children with perinatal focal brain injury (Booth et al. 1999, 2000). In a relatively natural linguistic task, children listened to sentences and then responded true or false to a statement about each sentence. Adults and children show lower accuracy levels to difficult object-relative sentences, such as 'The pig that the dog jumped ate the trash in the street,' intermediate accuracy to moderate subject-relative sentences, such as 'The principal that tripped the janitor used the phone to call home,' and highest accuracy levels to relatively easy conjoined-verb-phrase sentences, such as 'The cat chased the rabbit and enjoyed the hunt in the yard.' However, children are less accurate than are adults. Children, as well as adults, are more accurate when true–false statements test for the subject of the first verb, for example, 'The dog jumped the pig,' than for the subject of the second verb, for example, 'The principal used the phone.'

In adults, the fMRI showed that the sentence comprehension task produced more activation in the LH than in the RH. The greatest activation was found in the superior temporal, middle temporal, and inferior frontal areas, the traditional language areas but also in prefrontal areas, the brain region associated with working memory. Healthy children activated similar neurocognitive networks as did the adults with LH predominance. However, the distribution of activity across these networks was related to response accuracy, strategy use, and task difficulty. Greater activation was associated with higher response accuracy and with consistent use of a comprehension strategy, whether correct or incorrect. Improved performance was also associated with more activation in the anterior middle temporal area suggesting that this area was important in the syntactic processing demanded by the task. Activation of occipital regions

was greater in children who were less accurate in the task. These findings suggest that some children may have been using a different and immature strategy in complex sentence processing, possibly trying to visualize the sentence as an aid to comprehension. In both normal adults and children, the number of areas activated correlated with the difficulty of the sentences. The RH was recruited for difficult sentences (Booth et al. 2000).

Children with LH damage showed patterns of activation consistent with organization of cognitive processing into homologous areas of the contralateral hemisphere. The size of their lesion was associated with the degree of their cognitive deficit. These results confirmed earlier findings of RH language in many individuals with left hemisphere damage undergoing a carotid amytal infusion (Rasmussen and Milner 1977).

Studies of children with LH damage indicate that though the RH can participate in language processing, the LH has a more central role in language tasks at school age. Dennis and co-workers found that despite comparable intelligence, children with LH hemidecortication were inferior to children with RH hemidecortication on the ease and speed of syntactic discrimination, interpretation of passive negative sentences, and repetition of syntactically complex sentences (Dennis and Kohn 1975, Dennis and Whitaker 1976, Dennis 1987). Aram and co-workers similarly found expressive and receptive syntactic difficulties in children with LH focal damage from embolic events (Aram et al. 1986, Aram and Ekelman 1987). However, it is important to note that children with LH damage acquired in early infancy or childhood have only subtle deficits, and are not frankly aphasic as are adults with comparable lesions.

It is difficult to assess if syntax is particularly vulnerable in children with these lesions acquired later in life. An alternative hypothesis is that subjects with LH damage had greatest difficulty with the most developmentally advanced items on the assessment (Bishop 1983). Children with LH injuries and LH perinatal damage also show difficulty in tests of comprehension of difficult verbal material, formulating sentences, and lexical retrieval (MacWhinney et al. 2000).

To determine the specificity of underlying information-processing deficits after early injury, MacWhinney et al. (2000) used on-line reaction time methodology. School-aged children underwent visual and auditory tasks of signal detection, recognition, and choice as well as rapid naming to visual and auditory stimuli. The results showed that children with LH, RH, and with mixed lesions generally were as accurate as age-matched controls but consistently had slower reaction times. The reaction times of normal learners lawfully decreased as the children went from 5 to 10 years of age. Similarly, in the subjects with early lesions, the reaction times became shorter as the children got older and the slopes suggested that by adulthood the reaction times might converge. The two tasks that best distinguished children with LH injuries from children with other lesions were verbally repeating numbers presented in the auditory mode and naming numbers presented in the visual mode, though there was considerable overlap across groups. Interestingly, both these tasks required phonological access and verbal output.

Taken together, these findings on school-aged children suggest that despite adequate functional communication skills, children with LH injuries may have more difficulty than do children with other lesions performing precise, constrained, high-level language tasks, particularly in tasks that require speed. The areas of greatest difficulty seem to be developmentally challenging syntactic skills, comprehension of difficult oral directions, and formulating sentences with a constrained vocabulary. Phonological access, as evidenced in naming and repeating verbal items, also differentiates subjects with damage to the LH from other children without injuries.

5. *Summary*

Study of the neural basis of language function is advancing rapidly in the era of neural imaging and related methods. Study of the neural basis of language development is lagging behind because many of the techniques are not appropriate for children. The early emerging picture is that sensory processing mechanisms that are called on for initial segmentation and organization of the linguistic stream are present at or even before birth. However, these processing biases are not immediately available to true verbal development. Language development seems to use a broad and bilateral neural network at each stage. That network becomes more constrained and specialized to the LH as a function of language experience. However, bilateral activation remains possible in normal functioning, for example, when tasks are difficult and in circumstances of brain injury. Adult patterns of neural activation are present by school age and show LH predominance. However, children with LH damage may use homologous regions of the RH for sentence processing. Hopefully, in the near future new experimental techniques will allow greater detail and understanding of these processes and will begin to explain individual differences in children developing typically and children with a variety of developmental disorders.

See also: Aphasia; Event-related/Evoked Potentials; First Language Acquisition: Cross-linguistic; Functional Brain Imaging; Infant Development: Physical and Social Cognition; Language Acquisition; Language Learning Impairment; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Sentence Comprehension, Psychology of; Speech Per-

ception; Speech Production, Neural Basis of; Speech Production, Psychology of; Syntactic Aspects of Language, Neural Basis of

Bibliography

Alexander M P 1997 Aphasia: Clinical and anatomic aspects. In: Feinberg T E, Farah M J (eds.) *Behavioral Neurology and Neuropsychology*. McGraw-Hill, New York, pp. 133–50

Aram D M, Ekelman B 1987 Unilateral brain lesions in childhood: Performance on the Revised Token Test. *Brain and Language* **32**: 137–58

Aram D, Ekelman B, Whitaker H 1986 Spoken syntax in children with acquired unilateral hemisphere lesions. *Brain and Language* **27**: 75–100

Best C T, McRoberts G W, Sithole N M 1988 Examination of perceptual reorganization or nonnative speech contrasts: Zulu click discrimination by English-speaking adults and children. *Journal of Experimental Pyschology: Human Perception and Performance* **14**: 345–60

Bishop D V M 1983 Linguistic impairment after left hemidecortication for infantile hemiplegia? A reappraisal. *Quarterly Journal of Experimental Psychology* **35**: 199–207

Booth J R, MacWhinney B, Thulborn K R, Sacco K, Voyvodic J, Feldman H M 1999 Functional organization of activation patterns in children: Whole brain fMRI imaging during three different cognitive tasks. *Progress in Neuropsychopharmacology and Biological Psychiatry* **23**: 669–82

Booth J R, MacWhinney B, Thulborn K R, Sacco K, Voyvodic J, Feldman H M 2000 Functional activation patterns in adults, children, and pediatric patients with brain lesions: Whole brain fMRI imaging during three different cognitive tasks. *Developmental Neuropsychology* **18**: 139–69

Braver T S, Cohen J S, Nystrom L E, Jonides J, Smith E E, Noll D 1997 A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage* **5**: 49–62

Burnham D 1993 Visual recognition of mother by young infants: Facilitation by speech. *Perception* **22**: 1133–53

Caplan D 1987 *Neurolinguistics and Linguistic Aphasiology: An Introduction*. Cambridge University Press, Cambridge, UK

Dall'Oglio A M, Bates E, Volterra V, Di Capua M, Pezzini G 1994 Early cognition, communication and language in children with focal brain injury. *Developmental Medicine and Child Neurology* **36**: 1076–98

Demonet J F, Chellet R, Ramsay S, Cardebat D, Nespoulous J L, Wise R, Rascol A, Frackowiak R 1992 The anatomy of phonological and semantic processing in normal subjects. *Brain* **115**: 1753–68

Dennis M 1987 Using language to parse the young damaged brain. *Journal of Clinical and Experimental Neuropsychology* **9**: 723–53

Dennis M, Kohn B 1975 Comprehension of syntax in infantile hemiplegics after cerebral hemidecortication: Left-hemisphere superiority. *Brain and Language* **2**: 472–82

Dennis M, Whitaker H 1976 Language acquisition following hemidecortication—linguistic superiority of the left over the right hemisphere. *Brain and Language* **3**: 404–33

Eimas P D, Siqueland E R, Jusczyk P, Vigorito J 1971 Speech perception in infants. *Science* **171**: 303–6

Engelien A, Silbersweig D, Stern E, Huber W, Doring W, Frith C, Frackowiak R S J 1995 The functional anatomy of recovery from auditory agnosia— a PET study of sound categorization in a neurological patient and normal controls. *Brain* **118**: 1395–1409

Feldman H M, Holland A L, Kemp S S, Janosky J E 1992 Language development after unilateral brain injury. *Brain and Language* **42**: 89–102

Frith C D, Friston K J, Liddle P F, Frackowiak R S 1991 A PET study of word finding. *Neuropsychologia* **29**(12): 1137–48

Heffner H E, Heffner R S 1984 Temporal lobe lesions and perception of species-specific vocalizations in macaques. *Science* **226**: 75–6

Howard D, Patterson K, Wise R, Brown W D, Friston K, Weiller C, Frackowiak R 1992 The cortical localization of the lexicons: Positron emission tomography evidence. *Brain* **115**: 1769–82

Huttenlocher J, Haight W, Bryk A, Seltzer M 1991 Early vocabulary growth: Relation to language input and gender. *Developmental-Psychology* **27**: 236–48

Jonides J, Schumacher E H, Smith E E, Lauber E, Awh E, Minoshima S, Koeppe R A 1997 Verbal working memory load affects regional brain activation as measured by PET. *Journal of Cognitive Neuroscience* **9**: 462–75

Just M A, Carpenter P A, Keller T A, Eddy W F, Thulborn K R 1996 Brain activation modulated by sentence comprehension. *Science* **274**: 114–6

Levy Y, Amir N, Shalev R 1994 Morphology in a child with a congenital, left-hemisphere brain lesion: Implications for normal acquisition. In: Tager-Flusberg H (ed.) *Constraints on Language Acquisition*: *Studies of Atypical Children*. Erlbaum, Hillsdale, NJ

MacWhinney B, Feldman H, Sacco K, Valdes-Perez R 2000 Online basic language measures in children with early focal brain lesions. *Brain and Language* **71**: 400–31

Marchman V A, Miller R, Bates E E 1991 Babble and first words in children with focal brain injury. *Applied Psycholinguistics* **12**: 1–22

Mazoyer B M, Tzourio N, Frak V, Syrota A, Murayama N, Levrier O, Salamon G, Dehaene S, Cohen L, Mehler J 1993 The cortical representation of speech. *Journal of Cognitive Neuroscience* **5**: 467–79

Mills D L, Coffey C S, Neville H 1997 Language comprehension and cerebral specialization from 13 to 20 months. *Developmental Neuropsychology* **13**: 397–445

Petersen S E, Fiez J A 1993 The processing of single words studied with positron emission tomography. *Annual Review of Neuroscience* **16**: 509–30

Price C J, Wise R J S, Frackowiak R S J 1996 Demonstrating the implicit processing of visually presented words and pseudowords. *Cerebral Cortex* **6**: 62–78

Raichle M E, Fiez J A, Videen T O, MacLeod A K, Pardo J V, Fox P T, Petersen S E 1994 Practice-related changes in human brain functional anatomy during non-motor learning. *Cerebral Cortex* **4**: 8–26

Rasmussen T, Milner B 1977 The role of early left-brain injury in determining lateralization of cerebral speech functions. *Annals of the New York Academy of Sciences* **299**: 355–69

Reilly J S, Bates E A, Marchman V A 1998 Narrative discourse in children with early focal brain injury. *Brain and Language* **61**: 335–75

Thal D J, Marchman V A, Stiles J, Aram D, Trauner D, Nass R, Bates E 1991 Early lexical development in children with focal brain injury. *Brain and Language* **40**: 491–527

Weiller C, Isenee C, Rijntjes M, Huber W, Muller S, Bier D, Dutschka K, Woods R P, Diener H C 1995 Recovery from Wernicke's aphasia: A positron emission tomographic study. *Annals of Neurology* **37**: 723–32

Zatorre R J, Evan A C, Meyer E, Gjedde A 1992 Lateralization of phonetic and pitch discrimination in speech processing. *Science* **256**: 846–9

Zurif E, Swinney D 1994 The neuropsychology of language. In: Gernsbacher M A (ed.) *Handbook of Psycholinguisitics*. Academic Press, New York, pp. 1055–74

H. M. Feldman

Language Endangerment

About 6000 languages are spoken in the world. Major languages, such as English, French, Spanish, Portuguese, German, Russian, Arabic, Hindi, and Chinese, which have millions of speakers, constitute a tiny fraction of them (less than 1 percent). More than 95 percent of the world's languages are spoken by minority peoples (Krauss 1992, pp. 5, 7), and many of these minority languages have just a few remaining speakers. In every part of the world, minority languages are disappearing at an alarming speed, being replaced by major languages such as those listed above. This article examines this language crisis.

1. *Degree of Language Endangerment*

In terms of the degree of their viability, languages of the world constitute a continuum, with fully thriving ones at one end and extinct ones at the other. Various terms—both adjectival and nominal—have been proposed to characterize this continuum. Some of the terms may be arranged as in Fig. 1.

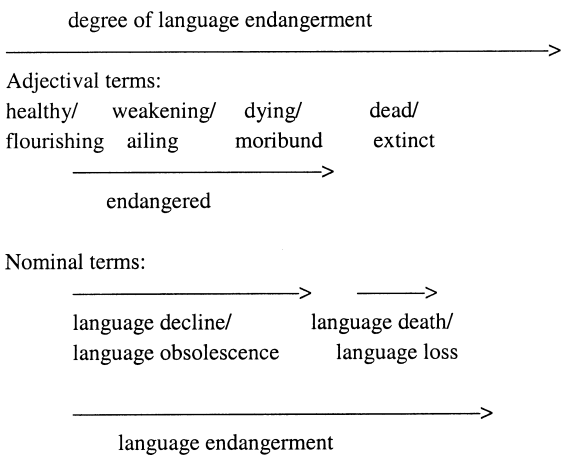

Figure 1
Degree of language endangerment, intensifying left to right

Consider Schmidt's (1990, p. 54) terms and their characterizations.

(a) Healthy languages: all generations actively use the language in a wide range of activities.

(b) Weakening languages: usually spoken by old people, but not fully transmitted to the younger generation.

(c) Dying languages: only a few speakers remain.

(d) Extinct languages: no speakers remain.

The use of the terms as proposed in Fig. 1 is as follows. Endangered languages comprise weakening/ailing and dying/moribund languages. Language decline and language obsolescence concern both groups of endangered languages, while language death and language loss refer to dead/extinct languages. The term 'language endangerment' is for the general subject.

2. *Current State of Language Endangerment*

The gravity of the current state of language endangerment can be seen from the following reports of selected areas.

In Australia, of the 250 languages spoken at the time of the first British colonization in 1788, about 100 are alive, mainly spoken by old people, and only about 20 are being passed on to children (Dixon 1991, pp. 235–36). In the USA and Canada, there must have been many hundreds of languages at the beginning of the sixteenth century, but only 187 have survived, and 149 of 187 are moribund (Zepeda and Hill 1991, p. 135, Kinkade 1991, p. 158, Krauss 1992, p. 5). In Russia, 45 of 65 indigenous languages are moribund (Krauss 1992, p. 5). In Africa, nearly 200 languages are endangered (Sasse 1992, p. 7).

To sum up, Krauss (1998, p. 103) estimates that, of 6000 languages of the world, only 5–10 percent are 'safe,' and 40 percent up to 75 percent are weakening, and at least 20 percent to perhaps 50 percent are moribund. Krauss (1998, p. 105) then forecasts that, if current trends are not reversed, during the next 150 years the number of languages is likely to be reduced to as few as 300.

3. *Two Approaches to Language Endangerment*

One approach is to document endangered languages. This is the most important thing to be done about them, and it is the most urgent task of linguists. Ideally, such documentation should include grammars, vocabularies, and texts/stories, employing audio and video recording. The other approach examines the process of language endangerment, in terms of (a) its causes, (b) its sociolinguistic aspects, and (c) structural changes (Sasse 1992). These three

Table 1
Functional differentiation

Minority language	Dominant language
Traditional life	Modern life
Within the community	With the outside world
Inside families	Outside families (inside the community)
Private, domestic	Public
Informal	Formal
Intimate	Nonintimate
Solidarity	Power
Secrecy	Nonsecrecy

aspects of language endangerment will be considered in the following three sections.

4. *Causes of Language Endangerment*

Language endangerment is caused by multiple factors, such as those listed below. They are all closely interrelated, and are often difficult to separate from one another.

(a) Dispossession of the land by invaders for settlement, farming, etc.

(b) Decline or loss of the population, caused by (i) natural catastrophes, e.g., volcanic eruption, earthquakes, (ii) diseases imported by invaders, e.g., venereal disease, smallpox, measles, influenza, and (iii) violent acts, e.g., massacres, genocide.

(c) Relocation of the people away from their traditional land.

(d) Socioeconomic oppression, by economic deprivation, discrimination, etc.

(e) Low status of the language, accompanied by ridicule, stigma, shame, etc.

(f) Assimilation policy, language policy by the government, e.g., (i) ban on the use of the indigenous language at schools, and (ii) placing children in boarding schools, which separates the children from their home environment and interrupts the transmission of the language.

(g) Switch to modern life, e.g. (i) schooling where the dominant language is used to the exclusion of the minority language, (ii) introduction of mass media, such as TV, and (iii) improvement of transport.

(h) Religion. For example, with the spread of Islam, Arabic replaced many languages of the Middle East and north Africa.

(i) Mixture of speakers of different languages, e.g., in boarding schools and reservations, where people were often forced to speak the dominant language, such as English, as a lingua franca.

(j) Language attitude. For example, if old speakers consider their language as an invaluable treasure, they may prefer not to transmit it to the younger generation for fear that the latter, who have often been westernized and 'modernized,' may not treat it with due care and respect (Schmidt 1990, p. 20).

5. *Sociolinguistic Aspects of Language Endangerment*

If a minority language and a dominant language coexist in one community, often they are used in different functional domains. Such functional differentiation is shown in Table 1. Needless to say, these domains are all closely inter-related; most of them are self-obvious. Consider the following examples.

(a) In Paraguay, Paraguayan Guarani and Spanish show a functional split, roughly as follows (Rubin 1972). If the location is rural, then Guarani will be used. If the location is nonrural, then Spanish will be used in a formal setting. In a nonformal setting, Spanish will be used if the speakers are not intimate. Between intimates, Guarani will be used if the discourse is not serious. If the discourse is serious, then the speaker may use, for instance, his/her first language or the language the hearer is estimated to be the more proficient in. Paraguayan Guarani has secured its position in this stable bilingualism, and consequently it is not endangered. It is the only native language of South America that is not endangered (Adelaar 1991, p. 73).

(b) Nahuatl of Mexico was used as an 'inside' language, and Spanish as an 'outside' language. This functional split has then developed into the 'solidarity–power' split: Nahuatl functioning as a marker of ethnicity and as a language of solidarity, as against Spanish, which is now a language of power, used to express ambition and mobility (Hill 1983, pp. 265–66).

(c) A minority language is often used as a secret language, for talking about something secret or private in the presence of outsiders.

Now, if the bilingualism in a given community is stable, with the two languages exhibiting a functional differentiation, then the minority language has a chance of survival, as is the case with Paraguayan Guarani. However, in most cases the functional range of the minority language diminishes, while that of the dominant language increases, with the result that the people shift from the minority language to the dominant language. (This phenomenon is called language shift.) When the language shift is completed, the minority language is extinct.

6. Structural Changes in Language Endangerment

6.1 Dyirbal of Australia

As a language is endangered and nears its demise, it undergoes structural changes, in phonology, morphology, syntax, semantics, lexicon, and speech styles, among other features. See *Language obsolescence*. Most such changes appear to be in the nature of simplification and reduction. Opinions differ, however, as to whether it is possible to generalize about structural changes in language endangerment. In the following, just one case study will be presented, namely, Schmidt (1985) on Dyirbal of northeast Australia. Due to the space limitation, only a very small portion of Schmidt's brilliant work can be cited. (TD refers to the traditional Dyirbal, and YD (young Dyirbal) refers to varieties of Dyirbal that deviate from the traditional norm.)

6.2 Phonology

One of the interesting changes concerns *r*-sounds. TD has an opposition of retroflex *r* and trill *rr*; consider the minimal pair of *yara* 'man' and *yarra* 'fishing line.' English has no such opposition. Now, in YD, this opposition is maintained, but it is maintained only in minimal pairs; elsewhere this opposition is not adhered to consistently, and speakers are not certain whether to use *r* or *rr*.

6.3 Morphology

One of the issues Schmidt examines is the ergative case, and it will be useful to provide a brief introduction to the phenomenon of ergativity.

Ergativity is a phenomenon in which the intransitive subject (S) and the transitive object (O), are treated alike, in distinction from the transitive subject (A). This pattern is called an ergative–absolutive pattern. It contrasts with, for example, a nominative–accusative pattern, in which the S and the A are grouped together, separately from the O. Ergativity may be manifested at various levels in grammar, e.g., morphology and syntax.

In TD, in terms of case-marking morphology, nouns, adjectives, etc. have an ergative–absolutive pattern, with the S and the O in the absolutive case and the A in the ergative case, e.g., (a), (b), while first- and second-person pronouns have a nominative–accusative pattern, see Table 2. (The ergative case does not occur in 'familiar' languages, such as English.)

(a) *nguma-Ø* *banaga-nyu.*
father-ABS(S) return-NONFUT
'Father (S) returned.'

(b) *nguma-Ø* *yabu-nggu* *bura-n.*
father-ABS (O) mother-ERG (A) see-NONFUT
'Mother (A) saw Father (O).'

In TD, the ergative suffix has six allomorphs, including *-nggu* for disyllabic vowel-final stems, e.g., *yabu-nggu* 'mother-ERG,' cf. (b); *-gu* for trisyllabic or longer vowel-final stems, e.g., *girimu-gu* 'snake-ERG'; *-ju* for *y*-final and *ny*-final stems, e.g., *walguy-ju* 'taipan (snake species)-ERG' and *binyjirriny-ju* 'lizard-ERG'; and, *-du* for *n*-final stems, e.g., *midin-du* 'possum-ERG.'

Schmidt shows that, among YD speakers, this morphological complexity was gradually reduced as a given speaker's proficiency in TD diminished, to the point where one single suffix, *-gu*, was used for all stems, irrespective of their stem-final phoneme. There were even less proficient speakers, who did not use an ergative suffix at all, and instead relied on word order, more specifically, on the English-type word order in a nominative–accusative pattern, in which the S and the A occur before the verb, and the O after the verb. (In TD, word order is free; syntactic function is indicated by case affixes of NPs.)

(c) *gugar* *buga-bi-n.*
goanna dead-INTR.VERBALIZER NONFUT
S V
'The goanna is dead.'

(d) *gugar* *baja-n* *ban* *jugumbil.*
goanna bite-NONFUT her woman
A V O
'The goanna bit the woman.'

6.4 Syntax

TD exhibits a rare phenomenon called syntactic ergativity. Due to the space limitation, a highly simplified account will be given below (for details, see Dixon 1994, pp. 160–72). Consider:

Table 2

	A (transitive subject)	S (intransitive subject)	O (transitive object)
Nominative–accusative pattern	nominative case	nominative case	accusative case
Ergative–absolutive pattern	ergative case	absolutive case	absolutive case

(e) *nguma-Ø* *yabu-nggu*
father-ABS (O) mother-ERG (A)
bura-n *banaga-nyu.*
see-NONFUT return-NONFUT
'Mother (A) saw Father (O) and [Father (S), not Mother] returned.'

It is important to stress that (e) does NOT have the meaning of the English sentence (f):

(f) Mother (A) saw Father (O) and [Mother (S), not Father] returned.

That is, in English, the understood S refers to the A and not the O; the S and the A are treated alike, in distinction from the O. In contrast, in Dyirbal, the understood S refers to the O and not the A. That is, in Dyirbal, the S and the O are grouped together, separately from the A, in complex sentences as well. This phenomenon—the grouping of the S and the O, separately from the A, in complex sentences—is called syntactic ergativity. It is rare among the world's languages.

Now, Schmidt shows that the use of syntactic ergativity, too, was on the decline among YD speakers.

6.5 Lexicon

Schmidt's study indicates that certain areas of the vocabulary were prone to loss. They include (a) types of animals and trees, e.g., 'ironbark,' 'bottlebrush,' (b) culture-specific items, e.g., 'yamstick,' 'fighting ground,' and (c) kinship terms for which there is no corresponding term in English. However, there were also areas of the vocabulary that were resistant to borrowing, such as (d) human body, e.g., 'hair,' 'tooth,' 'eye,' (e) human classification, e.g., 'man,' 'woman,' 'child,' and (f) well-known animates, e.g., 'carpet snake,' 'black bream.'

7. Language Maintenance and Revival

There are many attempts to maintain or even revive languages (see Fishman 1991). The types of these programs include the following.

(a) Language classes. For example, more than a dozen Ainu language classes for Ainu people are conducted in Japan.

(b) Bilingual education. In Australia, for instance, a number of schools are conducted using the local Aboriginal language and English.

(c) Language immersion program. An environment, e.g., preschool, is provided in which children will only hear and speak the language. The most successful and best-known is 'the language nest' program for the Maori language of New Zealand.

(d) 'Neighborhood method.' A most drastic example is a group of eleven families in Belfast, who—in order to create a cohesive community—bought houses in the same neighborhood in the middle of an English-speaking community in Belfast, and raised their children as bilinguals (in Irish and English) (Maguire 1991).

Factors that contribute to the success of a language maintenance program include the following. (i) Determination and often even sacrifice on the part of the community (cf. (d)). (ii) Government support, e.g., funding. (iii) Good documentation of the language. (iv) Availability of human resources, e.g., teachers, and advice from experts. (v) Land/area where the people concerned can live together (cf. (d)), preferably away from speakers of the dominant language. (vi) Existence of just one language for maintenance, as is the case with Maori of New Zealand. In contrast, language maintenance is much more difficult in Australian Aboriginal communities, where often several languages are spoken. (vii) Existence of a sufficiently large number of speakers. (viii) Inclusion of cultural activities, e.g., music, dance, handcraft. (ix) Tolerance of language change, as against linguistic purism. Attempts to preserve a 'pure' language may hinder the language's maintenance by discouraging non fluent speakers from speaking it.

8. Value of Linguistic Heritage

As seen above, a large number of people are making efforts, with dedication and determination, to revitalize languages in danger. Why is it important to do so? What appear to be three representative and interrelated views are cited below.

(a) 'Community view': ethnic identity. In this view, language is one of the most central components of a given group's culture. Their language is the symbol of their ethnic identity, expressing their culture in terms of worldview, ceremonies, songs, kinship terms, etc.

(b) 'Global view': linguistic diversity. This view maintains that linguistic diversity is important to human intellectual life. Minority languages often exhibit phenomena quite unimaginable if our attention were limited to, say, English. An example is syntactic ergativity of Dyirbal. With the death of minority languages, these aspects of rich cultural heritage of mankind die.

(c) Language rights. In June 1996, the draft of the Universal Declaration of Linguistic Rights was proclaimed at the World Conference on Linguistic Rights in Barcelona. The declaration is based on the principle of the equality of all peoples and all languages, and considers the following, among others, to be an inalienable human right:

(i) the right to use one's own language, and

(ii) the right for one's own language and culture to be taught.

The concept of language rights is no doubt one factor that motivates the community view, cited above, although the members of the community may not always be conscious of it.

9. Organizations

There is now an increasing number of organizations devoted to endangered languages, e.g., Endangered Language Fund (elf@haskins.yale.edu, http://www.ling.yale.edu/~elf); Foundation for Endangered Languages (nostler@chibcha.demon.co.uk, http://www.bris.ac.uk/Depts/Philosophy/CTLL/FEL/); Society for Endangered Languages (GBS@uni-koeln.de, http://www.unikoeln.de/gbs); and Terralingua (gws@mail.portup.com, http://cougar.ucdavis.edu/-nas/terralin/home.html). The following mailing list is available: endangered-languages-l@cleo.murdoch.edu.au, http://cleo.murdoch.edu.au.lists/endangered-languages-l. The following website is useful: http://www.tooyoo.ac.jp/ichel.html.

See also: Bilingualism and Multilingualism; Code Switching: Linguistic; Diglossia; Language and Ethnicity; Language Contact; Language Obsolescence; Language Policy: Linguistic Perspectives; Language Policy: Public Policy Perspectives; Pidgin and Creole Languages

Bibliography

Adelaar W F H 1991 The endangered languages problem: South America. In: Robins R H, Uhlenbeck E M (eds.) *Endangered Languages*, Berg, Oxford

Brenzinger M (ed.) 1992 *Language Death*. Mouton de Gruyter, Berlin

Dixon R M W 1991 The endangered languages of Australia, Indonesia, and Oceania. In: Robins R H, Uhlenbeck E M (eds.) *Endangered Languages*, Berg, Oxford

Dixon R M W 1994 *Ergativity*. Cambridge University Press, Cambridge

Dorian N C 1981 *Language Death*. University of Pennsylvania Press, Philadelphia, PA

Dorian N C (ed.) 1989 *Investigating Obsolescence*. Cambridge University Press, Cambridge, UK

Fishman J A 1991 *Reversing Language Shift*. Multilingual Matters, Clevedon, UK

Grenoble L A, Whaley L J (eds.) 1998 *Endangered Languages*. Cambridge University Press, Cambridge, UK

Hill J H 1983 Language death in Uto-Aztecan. *International Journal of American Linguistics* **49**: 258–76

Kinkade M D 1991 The decline of native languages in Canada. In: Robins R H, Uhlenbeck E M (eds.) *Endangered Languages*, Berg, Oxford

Krauss M 1992 The world's languages in crisis. *Language* **68**: 4–10

Krauss M 1998 The scope of languages endangerment crisis and recent responses to it. In: Matsumura K (ed.) *Studies in Endangered Languages*, Hituji Syobo, Tokyo, Japan

Maguire G 1991 *Our Own Language*. Multilingual Matters, Clevedon, UK

Matsumura K (ed.) 1998 *Studies in Endangered Languages*. Hituji Syobo, Tokyo, Japan

Robins R H, Uhlenbeck E M (eds.) 1991 *Endangered Languages*. Berg, Oxford

Rubin J 1972 Bilingual usage in Paraguay. In: Fishman J A (ed.) *Readings in the Sociology of Language*, Mouton, The Hague, The Netherlands

Sasse H-J 1992 Theory of language death. In: Brenzinger M (ed.) *Language Death*. Mouton de Gruyter, Berlin

Schmidt A 1985 *Young People's Dyirbal*. Cambridge University Press, Cambridge, UK

Schmidt A 1990 *The Loss of Australia's Aboriginal Language Heritage*. Aboriginal Studies Press, Canberra, Australia

Shoji H, Janhunen J (eds.) 1997 *Northern Minority Languages*. National Museum of Ethnology, Osaka, Japan

Wurm S A (ed.) 1996 *Atlas of the World's Languages in Danger of Disappearing*. UNESCO Publishing, Paris

Zepeda O, Hill J H 1991 The conditions of native American languages in the United States. In: Robins R H, Uhlenbeck E M (eds.) *Endangered Languages*. Berg, Oxford

T. Tsunoda

Language Learning Impairment

Early reports of children who failed to learn language at or near the expected age, but with no history or evidence of acquired brain lesions, began to emerge only in the mid 1960s. Such children were first diagnosed as 'developmentally aphasic.' The diagnosis of 'developmental aphasia' emphasized the congenital or developmental, rather than acquired, nature of the language learning impairment. Similarly, it was emphasized that the specificity of the deficit must lie in the linguistic domain. Language learning problems secondary to other developmental disabilities such as hearing loss or deafness, mental retardation, evidence of frank neurological problems (such as seizure disorder or hemiplegia), or structural or motor deficits effecting the oral musculature (such as cleft palate or cerebral palsy), were traditionally excluded from the classification of developmental aphasia. One exception to this diagnosis by exclusion was notable deficits, not only in speech and language, but also in higher order central auditory processing.

Over the years the term 'developmental aphasia' has been replaced by other terms such as 'congenital aphasia,' 'specific language impairment' and, most recently 'language learning impairment.' The most recent classification, 'language learning impairment,' acknowledges the frequent continuum between oral and written language learning impairments in children. More recently, research has moved away from a focus on exclusionary definitions of specific language impairments to a broader perspective that focuses on better elucidating the linguistic, cognitive, and neurobiological concomitants of developmental language learning impairments.

Developmental language learning impairments (LLI) can occur in children for a wide variety of reasons, and are one of the most prevalent of all

developmental disabilities. It has been estimated that approximately 20 percent of all children have some form of language learning impairment. The most common known causes of developmental language learning impairments are hearing loss (including intermittent hearing loss resulting from chronic otitis media), general mental retardation, and motor or structural defects affecting the oral musculature. Approximately 7.5 percent of children have developmental language impairments of unknown origin.

1. Diagnosis

The differential diagnosis of developmental language learning impairments is based on behavioral evaluations that include audiological, neurological, psychological, educational, speech, and language assessments. Developmental oral language learning impairments are divided into two basic categories; Expressive Language Disorder, affecting language production, and Mixed Receptive–Expressive Language Disorder, a disorder encompassing both language comprehension and production deficits. Comprehension and/or production problems may occur within one or more of the components of language; including phonology, morphology, semantics and/or syntax. Problems with pragmatics (interactive conversational skills) also occur frequently. A high proportion of children with developmental language learning impairments also have difficulty with speech articulation; that is, difficulty clearly and correctly producing one or more of the speech sounds of their language. However, speech articulation defects and developmental language learning impairments can occur independently of each other.

Many developmental cognitive disorders, such as mental retardation, pervasive developmental disability (including autism), attention deficit disorder, central auditory processing disorder, or specific syndromes such as Down's Syndrome, Fragile X Syndrome, Klinefelter's Syndrome, etc., may include delay in language development. In addition to these known concomitants of developmental language impairment, a recent epidemiological study of monolingual English-speaking kindergarten children in the US found that approximately 8 percent of boys and 6 percent of girls have a specific developmental language impairment of unknown origin. This epidemiological study also showed that the clinical identification of developmental language impairments remains low. Over 70 percent of the parents or teachers of children identified in this epidemiological study as language impaired had previously been unaware that their child had a speech or language problem (Tomblin et al. 1997). These data emphasize the importance of early referral for differential diagnosis for children failing to meet language milestones.

Children with developmental language learning impairments also have been shown to be at risk for childhood behavioral disorders. For example, clinical research studies showed that children referred to child guidance clinics for a variety of behavioral and emotional problems were found to have a higher than expected incidence of developmental language learning disorders. Conversely, children diagnosed with developmental language disorders also have been found, upon examination, to have a preponderance of behavioral and emotional disorders. Over 80 percent of juvenile delinquents were found, upon testing, to be language learning impaired (Beitchman et al. 1986, Hall and Moats 1999).

Longitudinal research studies that have followed children with early developmental language impairments, from the preschool though elementary school years, consistently have demonstrated a striking link between early developmental oral language impairments and subsequent academic achievement problems, especially dyslexia, which is characterized by reading, writing, and spelling deficits. Research that has compared children classified as 'language impaired' with those classified as 'reading impaired' has shown that both groups are characterized by a variety of oral language deficits, specifically phonological and memory problems, as well as language comprehension deficits (Snow et al. 1998). Whether these common deficits derive from speech-specific mechanisms, or from more basic neural processing deficits, has been the focus of considerable research and theoretical debate (see Tallal et al. 1993, Farmer and Klein 1995, Leonard 1998 for review).

2. Etiology

Considerable research attention has been focused on trying to understand the causes of developmental language learning impairments. Given the lack of homogeneity amongst children showing delayed language development, and the diversity of symptomatologies that can accompany developmental language learning impairments, it is unlikely that a single etiology is responsible. Etiological research to date has focused on linguistic, cognitive, and neurobiological studies.

2.1 Linguistic Studies

There have been considerable advances made in understanding the specific linguistic deficits of children with language impairments. Research has demonstrated particular patterns of difficulty within the linguistic components of phonology, syntax, and morphology. Specifically, deficits in both receptive and expressive grammatical morphology have been dem-

onstrated (Leonard 1998, Bishop 1997). However, results of longitudinal studies have shown that children with developmental language impairments develop the linguistic structures of language along a similar linguistic trajectory to that observed in normally developing younger children, albeit considerably more slowly. To date, there is little evidence that these children make linguistic errors that are deviant or substantially different from younger children at the same stage of language development. Rather, children with language learning impairments take much longer to progress through the stages of normal language development. A similar pattern of delay, rather than deviance, occurs across most populations of children with developmental language impairment, regardless of clinical diagnosis.

Looking cross-linguistically at children with developmental language impairments learning different languages in different countries, it has been shown that whichever linguistic structures are brief, unstressed, or of weak phonetic substance, and occur less frequently in a particular language, are the most difficult for children to learn, and also the most delayed in children with language learning impairments (Leonard 1998). That the same order of development occurs across populations and language environments gives strong support that there is a potent metric of sheer difficulty (whether representational or perceptual) that is imposed on the learning of linguistic structures and contents.

2.2 Cognitive Studies

The most extensive etiological studies to date have utilized behavioral methodologies and have resulted in the formulation of two prominent hypotheses: the 'linguistic specific' hypothesis and the 'rate processing constraint' hypothesis. The rate processing constraint hypothesis assumes a direct impact on the learning of the linguistic system, specifically on learning phonology and grammatical morphology. This hypothesis claims that central auditory processing constraints, particularly in the areas of auditory attention and serial memory, may result from a more basic impairment in the rate of neural information processing (Tallal et al. 1993). Tallal and colleagues have shown that many children with phonologically based speech, language, and reading disorders need significantly more neural processing time (hundreds of milliseconds instead of tens of milliseconds) to process dynamic (rapidly changing) sensory stimuli. They have hypothesized that this dynamic sensory processing rate constraint has a particularly detrimental effect on phonological processing, as many acoustic changes occurring in rapid succession within syllables and words, necessary to distinguish the individual phonological elements (speech sounds), occur within the tens of milliseconds time window. Research has demonstrated that perception of those speech syllables that incorporate brief, rapidly successive acoustic cues is most problematic for these children.

There is growing evidence from family and twin studies that developmental language learning impairments may aggregate in families and may be genetically transmitted. Twin studies have shown a particularly high concordance rate (heritability) for measures of phonological analysis (Bishop 1997). Family studies show that, as a group, infants born into families with a positive history of developmental language learning impairment demonstrate longer processing times for both nonverbal and verbal acoustic stimuli than matched infants born into families with a negative family history. When followed prospectively, nonverbal acoustic processing rates established at six months of age have been shown to predict the rate of language development, with infants showing the slowest processing rates subsequently being most likely to be delayed in language learning (see Tallal et al. 1993 for review).

In older children and adults, highly significant positive correlations between nonverbal auditory and visual dynamic sensory sensitivity and reading decoding (phonological and orthographic, respectively) also have been found across a number of studies (see Farmer & Klein 1995, Stein & Talcott 1999 for reviews). Importantly, highly significant correlations between dynamic auditory and visual sensitivity and literacy skills has been reported in normally developing children as well (Talcott et al. 2000). These recent findings provide strong evidence that individual differences in dynamic sensory sensitivity play an important role in the development of both oral and written language, across the entire range of development. As such, children with LLI may fall at the disadvantageous end of a normal distribution, rather than representing a distinct impairment.

The proponents of the 'linguistic-specific' hypothesis strongly dispute these data. They exclude any role for nonlinguistic processing constraints as a possible etiology for language learning impairments. Rather, they support a modular view of language processing and propose that language learning impairments are caused by failure of innate, 'language specific,' modules to function normally (see Leonard 1998 for review).

2.3 Neurobiological Studies

With the recent advent of *in vivo* brain imaging technologies, these powerful new methods have been applied to the study of the neurobiological basis of developmental language impairments. Electrophysiological studies support the results of behavioral studies showing specific deficits in phonological processing as well as acoustic analysis (Leppanen & Lyytinen 1997). Results from structural magnetic

resonance imaging (MRI) show that pars triangularis (Broca's area) is significantly smaller in the left hemisphere of children with language learning impairments and that these children are also more likely to have rightward asymmetry of language structures. The opposite pattern is seen in children developing language normally (Gauger et al. 1997).

Studies using functional neuroimaging such as magnetoenchephalography (MEG), have the advantage of providing both the spatial and temporal resolution most compatible with studying *in vivo* the real time neural processing involved in language. Recently, MEG has been used in an attempt to help resolve the hotly debated 'linguistic-specific' vs. 'rate processing constraint' hypotheses. These MEG studies have uniformly demonstrated significant physiological processing rate differences in individuals with a history of language learning disability, compatible with the rate processing constraint hypothesis (see Tallal et al. 1998 for review). However, the precise neural mechanisms that link these physiological rate processing constraints and the phonological and other linguistic deficits that characterize individuals with language learning disabilities, remain to be determined.

New evidence from neuropathological studies as well as functional magnetic resonance imaging (fMRI) studies has suggested that reduced cell size and density in magnocellular regions of the thalamus, leading to aberrant cortico-thalamic transmission, may underlie both the rate processing constraints and phonological processing problems of individuals with language learning impairments (see Stein and Talcott 1999 for review). Furthermore, differential cortical axonal myelination thickness has recently been postulated to underlie left hemisphere specialization for language function in general (Anderson et al. 1999), as well as for language learning impairments in particular (Klingberg et al. 2000).

3. *Remediation*

Remediation of language learning impairments is generally provided by speech language pathologists, psychologists, or special education teachers. Traditionally, speech and language therapy has been guided by a detailed linguistic assessment of the pattern of receptive as well as expressive errors made by an individual child using both standardized testing instruments and language sampling methodologies. Remediation has focused on helping the child to use age-appropriate linguistic structures in a variety of contexts. Reading remediation has focused on explicitly training children to recognize and manipulate phonemes within words (called phonological awareness) in an attempt to improve reading decoding skills, as well as on methods aimed at improving fluency and reading comprehension. Longitudinal outcomes research has shown that children with developmental language learning impairments generally require approximately two years of traditional intervention to acquire the equivalent of one year of gain in oral and written language development (see Leonard 1998, Lyon & Moats 1997 for reviews).

Electrophysiological studies of sensory fields in the cortex of primates have shown that the underlying neural circuitry can be altered after specific, temporally cohesive training regimens, leading to 'remapping.' These mapping studies demonstrate the dynamic plasticity of the brain (Jenkins et al. 1990). These studies recently have led to the development of new approaches to treating language learning impairments that incorporate temporally cohesive computerized training programs. Remediation studies, incorporating for the first time explicit dynamic auditory training, have been coupled with language training using speech that has been computer-modified to enhance and extend dynamic changes within the ongoing acoustic waveform. These studies demonstrated that explicit training, aimed at enhancing dynamic auditory sensitivity, results in highly significant improvement in temporal processing thresholds, speech discrimination, and listening comprehension (Merzenich et al. 1996, Tallal et al. 1996). These new remediation studies, together with those that have focused attention on phonological awareness training (see Lyon and Moats 1997 for review), have stimulated remediation research that aims to develop more scientifically validated approaches for the treatment of oral and written language learning impairments.

4. *Goals for Future Research*

Future research will need to continue to focus on etiology, specifically understanding the neurobiological, genetic, and environmental influences that contribute interactively to developmental language learning disabilities. It will be particularly important to better understand how nonlinguistic neural processing mechanisms (attention, dynamic processing, memory, and other executive functions) interact developmentally with linguistic systems in the brain during both typical and atypical language development. Integrating new data from advanced imaging technologies, genetics, and cognitive methodologies should lead to a better understanding of the causes and determinants of developmental language learning impairments. Determining whether language learning impairments represent a distinct impairment or rather lie at the disadvantageous end of individual differences along a normal distribution will also remain an important focus of research. The ultimate goal will be to understanding better the multiple factors affecting the broad array of individual differences that characterize the language learning process, which should in turn lead to the development of more effective methods for accelerating oral as well as written language

learning in all individuals, specifically those with developmental language learning impairments.

See also: Language Acquisition; Language Development, Neural Basis of; Speech Perception; Syntactic Aspects of Language, Neural Basis of

Bibliography

Anderson B, Southern B S, Powers R E 1999 Anatomical asymmetries of the posterior superior temporal lobes: A postmortem study. *Neuropsychiatry Neuropsychology and Behavioral Neurology* **12**: 247–54

Beitchman J H, Nair R, Ferguson M, Patel P G 1986 Prevalence of psychiatric disorders in children with speech and language disorders. *Journal of American Academy of Child Psychiatry* **24**: 528–35

Bishop D V M 1997 *Uncommon Understanding: Development and Disorders of Language Comprehension in Children*. Psychology Press, Cambridge, UK

Farmer M E, Klein R M 1995 The evidence for a temporal processing deficit linked to dyslexia: A review. *Psychonomic Bulletin & Review* **24**: 460–93

Gauger L M, Lombardino L J, Leonard C M 1997 Brain morphology in children with specific language impairment. *Journal of Speech Language Hearing Research* **40**: 1272–84

Hall S L, Moats L 1999 *Straight Talk About Reading. How Parents Can Make a Difference During the Early Years*. Contemporary Books, IL

Jenkins W M, Merzenich M M, Ochs M T, Allard T, Guie R E 1990 Functional reorganization of primary somatosensory cortex in adult owl monkeys after behaviorally controlled tactile stimulation. *Journal of Neurophysiology* **63**: 82–104

Klingberg T, Hedehus M, Temple E, Salz T, Gabrieli J D E, Moseley M E, Poldrack R A (2000). Microstructure of temporo-parietal white matter as a basis for reading ability: Evidence from diffusion tensor magnetic resonance imaging. *Neuron*

Leonard L B 1998 *Children with Specific Language Impairment*. MIT Press, Cambridge, MA

Leppanen P H T, Lyytinen H 1997 Auditory event-related potentials in the study of developmental language-related disorders. *Audiology and Neuro-Otology* **2**: 308–40

Lyon R G, Moats L C 1997 Critical conceptual and methodological considerations in reading intervention research. *Journal of Learning Disabilities* **306**: 578–88

Merzenich M, Jenkins W, Johnston P S, Schreiner C, Miller S L, Tallal P 1996 Temporal processing deficits of language-learning impaired children ameliorated by training. *Science* **271**: 77–80

Snow C E, Burns S M, Griffin P 1998 *Preventing Reading Difficulties in Young Children*. National Academy Press, Washington, DC

Stein J, Talcott J 1999 Impaired neuronal timing in developmental dyslexia—The magnocellular hypothesis. *Dyslexia* **5**: 59–77

Talcott J, Witton C, McClean M, Hansen P, Rees A, Green G, Stein J 2000 Dynamic sensory sensitivity and children's word decoding skills. *Proceedings of the National Academy of Science*. **97**: 2952–7

Tallal P, Galaburda A M, Linas R R, von Euler C (eds.) 1993 *Temporal Information Processing in the Nervous System: Special Reference to Dyslexia and Dysphasia*. Academy of Sciences Press, New York, p. 682

Tallal P, Merzenich M, Miller S, Jenkins W 1998 Language learning impairments: Integrating basic science technology and remediation. *Experimental Brain Research* **123**: 210–19

Tallal P, Miller S L, Bedi G, Byma G, Wang X, Nagarajan S S, Schreiner C, Jenkins W M, Merzenich M M 1996 Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science* **271**: 81–4

Tomblin J B, Records N L, Buckwalter P, Zhang X, Smith E, O'Brien M 1997 The prevalence of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research* **40**: 1245–60

P. Tallal

Language Obsolescence

Languages passing out of vernacular use are said to be obsolescent. As one of the possible outcomes of language contact, obsolescence may emerge both in conditions of universal multilingualism, as in the case of Tariana in the Brazilian Amazon (Aikhenvald 1996), and where a substantial part of an ethnic population has previously been monolingual, as in the case of the Gaelic-speaking fisherfolk of East Sutherland, Scotland (Dorian 1981).

The obsolescent phase of a language accompanies speech community members' shift to the use of some other language. Such a shift often comes about as a response to changes in political conditions (e.g., introduction of state-controlled schooling, transfer of territory to the control of a different state) or in economic conditions (e.g., heightened contact with outsiders through decline of traditional livelihoods or through new exploitation of local resources), or both. Language shift may also be prompted primarily by ideological factors, as in the case of Taiap, a language spoken only in Gapun village, Papua New Guinea. Despite physical isolation, limited school exposure, and continued pursuit of largely traditional subsistence activities, shift to the geographically expanding creole Tok Pisin (see *Pidgin and Creole Languages*) proceeds apace among Taiap-speaking villagers, chiefly because ideology links language, gender, and affect in ways that have steadily devalued the local vernacular as indexical of women (Kulick 1992).

1. The Emergence of a Proficiency Continuum

Except where a physical calamity (volcanic eruption, infectious disease, genocidal attack, etc.) brings about the death of most speakers within a short period, obsolescence advances gradually, accompanied by the development of a proficiency continuum. The oldest members of a speech community typically speak the ancestral language with high proficiency and some other language(s) equally or less well, while younger

members speak the ancestral language with diminishing proficiency and some other language(s) with higher proficiency.

1.1 Segmenting the Proficiency Continuum

While such continua are prominent in accounts of the structural effects of gradual obsolescence, no general consensus prevails on the levels of proficiency that should be recognized across the continuum. In one model, speakers of obsolescent languages are characterized as strong, with full or nearly full competence; imperfect, with less than full competence; weak, with restricted speaking competence; and at the low end as rememberers, capable of producing only words and isolated phrases (Campbell and Muntzel 1989). In another model, fully fluent speakers are divided into older and younger fluent speakers, according to age-related degrees of retention of conservative grammatical features; those who make grammatical mistakes recognized as such by the fully fluent are considered semispeakers and are again subdivided according to level of control of conservative grammatical features (Dorian 1981). The existence of age-atypical individuals can complicate the segmentation of a proficiency continuum. While there is normally an overall correlation with age, individuals within any age cohort may by reason of personal history be atypical; childhood residence with grandparents sometimes produces age-exceptional proficiency in the receding language, for example, whereas periods of residence away from the linguistic heartland may have the opposite effect.

1.2 Assessing Structural Changes by Sampling along the Proficiency Continuum

By sampling the usage of speakers at various points along the proficiency continuum, ongoing changes in phonological and grammatical structure can be identified in an obsolescent language. Caution is necessary both in identifying and in interpreting evidence of change, however. In assessing change, the norms against which any deviation is measured must be established strictly in terms of the most conservative local usage. Local norms are essential, since any standardized or official form of the language may be historically irrelevant to the local form, based mainly on quite different dialects, for example, or relatively recently devised. Furthermore, structural changes found to be in progress are not necessarily the unequivocal result of obsolescence as such since intensive contact with another language often promotes structural change, whether language shift is underway or not. Reports of structural convergence appear, for example, both where a language long in contact with others has become obsolescent (as with Tariana, cited above) and where languages long in contact with one another all continue to be maintained (as in the Urdu, Marathi, and Kannada case discussed by Gumperz and Wilson 1971). Generally speaking, the broader the age-range encompassed by the proficiency continuum and the larger the number of speakers sampled at various points along it, the greater the likelihood that changes to which obsolescence contributes substantially can be identified. Where circumstances permit, longitudinal sampling can offer important additional evidence, by determining the degree to which innovative structural features once absent or rare among the oldest and most conservative speakers establish themselves subsequently among those who were previously the younger speakers.

2. The Rise of Interest in Language Obsolescence

Interest in language obsolescence arises in large part from the very structural deviation that generally accompanies reduction in use and incomplete transmission. This is a striking reversal of earlier attitudes and preoccupations. Until the last few decades of the twentieth century researchers were intent on describing a language in the most intact form available and avoided working with speakers who controlled an ancestral language imperfectly. But insights from the study of language acquisition (see *First Language Acquisition: Cross-linguistic*), pidginization, and speech disorders suggested that language obsolescence, as another arena in which structural shortfall appeared, might make a similarly useful contribution to a growing body of knowledge about the nature of human language capacities. Interest grew, therefore, in exploring such questions as whether grammatical features acquired late would also be lost early, whether morphological complexity would be any more sustainable in obsolescence than it was in pidginization, whether obsolescent-language features without parallel in the language to which speakers were shifting would be maintained, and so forth.

2.1 Focus on Reduction and Loss

Because rising interest in obsolescence stemmed from questions such as these, early research focused strongly on reduction and loss in obsolescent forms of a given language. Dressler (1972) looked at phonological rule loss in Breton, Hill (1973) at decreasing subordinate clause density in Luiseño and Cupeño, Dorian (1973) at loss and merger in nominal case and passive voice in Scottish Gaelic, and Dorian (1978) at decreasing morphological complexity in plural and gerund formation in Scottish Gaelic. The first two larger-scale studies of obsolescent languages were similar in focus. Dorian (1981) reported weakening and loss of grammatical gender in pronoun replacements, weakening of case in the noun phrase, progressive loss in verbal inflection, reduction of allomorphy in the formation of

noun plural and gerund, and substitution of analytical constructions for synthetic ones among imperfect speakers of an obsolescent variety of Scottish Gaelic. Schmidt (1985), investigating obsolescent Dyirbal in Queensland, Australia, noted similarity of reductive and replacive phenomena on the morphological and syntactic levels: replacement of ergative structure by nominative/accusative structure; allomorphic reduction in the nominal case system; loss of a distinction between alienable and inalienable possession; generalization of a single case affix over various case functions; English prepositions used in place of certain suffixes, or marking simply lost; substitution of free-standing words for tense-marking verbal affixes; regularization of an irregular verb inflection and of certain complex noun-marker paradigms; breakdown in agreement rules in the noun phrase and verb complex; rigidification of word order; and simplification of the noun classification system.

In these and subsequent accounts of the structural phenomena that accompany increased language contact and ongoing language shift (e.g., Silva Corvalán 1994, Holloway 1997), certain processes are conspicuous by their recurrence, e.g., outright loss of certain morphological markers, allomorphic reduction, paradigmatic and syntactic regularization, substitution of analytic structures for synthetic structures, and convergence with the phonological and grammatical structure of an expanding language in the contact setting. These are changes long familiar from the study of language history, including cases in which obsolescence was not at issue. In consequence, obsolescence researchers have generally favored the position that the changes characteristic of obsolescence are well recognized historical processes appearing with higher incidence in conditions of obsolescence and advancing more rapidly (Dorian 1981, Schmidt 1985, Silva Corvalán 1994).

2.2 *Focus on Retention of Complexity*

Early research on language obsolescence brought to light anticipated reduction and loss but produced at the same time considerable evidence of retained complexity. Despite increased favoring of a single suffix in imperfect Scottish Gaelic speakers' plural and gerund formations, for example, suffixation did not become the sole formation process and allomorphy remained diverse. Despite reductions in inflectional morphology among young Dyirbal speakers, the same speakers preserved notable complexity in derivational morphology. Particularly contrary to expectations was the finding that grammatical structures without parallel in the expanding language could survive in the usage of speakers toward the lower end of the proficiency continuum. With no reinforcement of grammatical gender available from English, for example, and despite the disappearance of gender-differentiated pronoun replacement in their Gaelic, gender differentiation in high-frequency diminutive formation remained relatively strong among imperfect English-dominant Gaelic speakers in East Sutherland. Young Dyirbal speakers maintained two verb conjugations surprisingly well in some structures, even though English has only one productive verb conjugation.

These findings have been important in establishing the structural phenomena of language obsolescence as distinct from those of pidginization or creolization and from those of first-language acquisition. Imperfect adult speakers of obsolescent languages are by definition bilingual at least. They have the cognitive equipment of adulthood as well as extensive experience with at least one other language to draw on at communicative need. Borrowing and code-switching are available to them, of course, and they have the inventive capacity of adults to extend in potentially novel ways the structural or semantic reach of whatever ancestral-language elements they have available. A weak semispeaker proved for example to have developed her own system for keeping grammatical genders distinct to some extent, in the nominal case system of East Sutherland Gaelic, despite the fact that she lacked one of the central morphophonemic devices used for that purpose by fully fluent speakers (Dorian 1986).

3. *Evidence of Linguistic Creativity in Obsolescence*

Apart from such individual cases of creativity as this last, more generalized instances have been noted. Imperfect younger speakers from a Hungarian population in a predominantly German-speaking region of Austria demonstrated the ability to use a structural property of Hungarian, expression of one grammatical or derivational category per morpheme in concatenative word-formation, to create grammatically regular constructions that substituted intelligibly for irregular formations which they lacked (Gal 1989). More strikingly, recent research has demonstrated that obsolescence is not a barrier to the development of new grammatical structures, even for speakers whose proficiency in an expanding language exceeds their proficiency in the recessive language. Ongoing potential for grammatical elaboration therefore joins retention of grammatical complexity in establishing that the phenomena associated with language obsolescence are not solely those of progressive reduction and loss.

3.1. *Development of New Grammatical Structures in Obsolescent Languages*

When certain grammatical distinctions cease to be made by speakers of an obsolescent language, grammatical elements (particles, nominal or verbal affixes,

functionally distinct forms of adverbs, etc.) may be emptied of their original grammatical content, after which they become subject to loss or to reinterpretation. While loss is certainly commonplace and reinterpretation sometimes amounts to no more than the development of variant forms, reinterpretation can also give rise to novel grammatical structures. Morphemes which formerly served as nominal case markers have emerged as focus-markers in Jingulu, for example; prominent now in this obsolescent Australian language, the focus-markers can be shown to have developed during the past four decades (Pensalfini 1999). New verbal morphology has similarly appeared in obsolescent Tariana, via grammaticalization of elements derived from compounded verbal roots and serial verb constructions (Aikhenvald 2000). Speakers of an obsolescent Walser German dialect in northern Italy have developed a complete set of verbal inflections for person via the familiar process of reanalysis of postverbal clitic pronouns as personal endings (Dal Negro 1998). As these instances demonstrate, the linguistic processes involved are once again not unique to obsolescent languages.

4. *Methodological Problems in the Study of Obsolescence*

4.1 *Inherent Difficulties in Assessing Imperfect Speakers' Linguistic Competence*

Some of the difficulty in establishing parameters of obsolescence stems from methodological obstacles to assessing low-proficiency speakers' competence. Even where the least proficient speakers willingly undertake assessment exercises such as translating set sentences or supplying a story-line to accompany a series of pictures, their performances may be affected by psychological or cultural factors: degree of acquaintance and compatibility with the researcher, anxiety about the adequacy of ancestral-language knowledge, uncongeniality of question-and-answer interactions, and so forth. Since failure to produce particular material on request can by no means be equated with inability to produce it, this remains a serious methodological shortcoming. Certain types of social encounter can evoke previously unavailable evidence of imperfect-speaker skills, as a chance encounter with a lonely Gaelic-dominant hospital patient did in the case of an East Sutherland Gaelic semispeaker; but by their very nature such encounters can seldom be staged or recorded. In Gapun village the attractiveness of Kulick's recording machinery allowed him to coax from a young child who routinely used Tok Pisin for extended speech an unexpectedly intact short narrative in Taiap (Kulick 1992). Receptive bilingualism in near-passive bilinguals uncomfortable with testing can sometimes be evaluated via macaronic interviews, with the researcher speaking only the receding language and the interviewee responding in the expanding language (Dorian 1999).

4.2 *Cultural Differences as Obstacles to Assessing Imperfect Speakers' Linguistic Competence*

In some settings the cultural norms for transmission of knowledge disfavor the question-and-answer approaches researchers have often used to isolate the individual elements of complex grammatical systems. Speakers may resist answering strings of questions, preferring to offer instruction by companionable activity or by story-telling. Recording narratives from speakers willing to go over them in detail with the researcher subsequently is a more successful and productive technique for phonological and grammatical analysis in some settings (Dixon 1984), although it may preclude direct comparison across the proficiency continuum unless a number of different speakers can offer versions of the same narrative.

There appear also to be differences in the degree to which imperfect speakers in late-stage obsolescence settings are willing to make spontaneous use of what active skills they have. In some Arvanitika-speaking districts of Greece, extremely low-proficiency individuals willingly produce this Albanian speech variety at a structural level that one researcher characterizes as 'pathologically' malformed (Sasse 1992). Spontaneous speech at this low a proficiency level, independently confirmed for Arvanitika (Tsitsipis 1989), would be highly unusual in some settings and therefore unavailable for study; Arvanitika currently stands as an extreme case for spontaneous low-proficiency speech production in an obsolescent language.

5. *Lacunae in Obsolescence Studies to Date*

In an important attempt to distinguish the effects of language obsolescence from those of language contact more generally, Sasse (1992) has pointed to acquisition by exposure alone (hearing the language spoken in one's vicinity) rather than by active transmission (being addressed in the language and expected to reply in it) as the crucial factor in giving rise to the low-proficiency producers of 'pathological' Arvanitika. One well-studied individual who had an exposure-only acquisition history for East Sutherland Gaelic produces far less malformed spontaneous speech than do the weakest Arvanitika semispeakers (Dorian 1999), but the effect of acquisition history merits closer study, especially in incipient obsolescence settings where families differ in the extent to which they use the ancestral language with their children. Studies of whole communities which share an ancestral language, but differ in degree of contact with a particular expanding language, and hence in the degree of shift,

are needed for what they may reveal of progressive stages in convergence, reduction, and loss of specific structural features. Again to illuminate the effects of contact, studies of a single language in contact with a variety of expanding languages would be useful.

See also: Communication and Transportation, History of; First Language Acquisition: Cross-linguistic; Historical Linguistics: Overview Language and Ethnicity; Language and Society: Cultural Concerns; Language Contact; Language Endangerment; Language Policy: Linguistic Perspectives; Language Policy: Public Policy Perspectives; Second Language Acquisition

Bibliography

Aikhenvald A Y 1996 Areal diffusion in Northwest Amazonia: The case of Tariana. *Anthropological Linguistics* **38**: 73–116

Aikhenvald A Y 2000 Areal typology and grammaticalization: the emergence of new verbal morphology in an obsolescent language. In: Gildea S (ed.) *Reconstructive Grammar: Comparative Linguistics and Grammaticalization*. Benjamins, Amsterdam

Campbell L, Muntzel M C 1989 The structural consequences of language death. In: Dorian N C (ed.) *Investigating Obsolescence*. Cambridge University Press, Cambridge, UK

Dal Negro S 1998 Spracherhaltung in der Beiz—Das Überleben der Walsersprache im Pomatt/Formazza. *Wir Walser* **36**: 13–16

Dixon R M W 1984 *Searching for Aboriginal Languages: Memoirs of a Fieldworker*. University of Queensland Press, St Lucia, Australia

Dorian N C 1973 Grammatical change in a dying dialect. *Language* **49**: 413–38

Dorian N C 1978 The fate of morphological complexity in Scottish Gaelic language death: evidence from East Sutherland Gaelic. *Language* **54**: 590–609

Dorian N C 1981 *Language Death*. University of Pennsylvania Press, Philadelphia, PA

Dorian N C 1986 Making do with less: some surprises along the language death proficiency continuum. *Applied Psycholinguistics* **7**: 257–75

Dorian N C (ed.) 1989 *Investigating Obsolescence*. Cambridge University Press, Cambridge, UK

Dorian N C 1999 The study of language obsolescence: stages, surprises, and challenges. *Langues et Linguistique* **3**: 99–22

Dressler W U 1972 On the phonology of language death. In: *Papers from the Eighth Regional Meeting of the Chicago Linguistic Society*. Chicago Linguistic Society, Chicago

Gal S 1989 Lexical innovation and loss: the use and value of restricted Hungarian. In: Dorian N C (ed.) *Investigating Obsolescence*. Cambridge, Cambridge University Press, UK

Gumperz J J, Wilson R 1971 Convergence and creolization: a case from the Indo-Dravidian border. In: Hymes D (ed.) *Pidginization and Creolization of Languages*. Cambridge University Press, Cambridge, UK

Hill J H 1973 Subordinate clause density and language function. In: Corum C, Smith-Stark T C, Weiser A (eds.) *You Take the High Node and I'll Take the Low Node: Papers from the Comparative Syntax Festival*. Chicago Linguistic Society, Chicago

Holloway C 1997 *Dialect Death: The Case of Brule Spanish*. Benjamins, Amsterdam

Kulick D 1992 *Language Shift and Cultural Reproduction: Socialization, Self, and Syncretism in a Papua New Guinean Village*. Cambridge University Press, Cambridge, UK

Pensalfini R 1999 Case suffixes as discourse markers in Jingulu—a case of innovation in an obsolescent language. *Australian Journal of Linguistics* **19**: 225–40

Sasse H-J 1992 Language decay and contact-induced change: similarities and differences. In: Brenzinger M (ed.) *Language Death*. Mouton de Gruyter, Berlin

Schmidt A 1985 *Young People's Dyirbal*. Cambridge University Press, Cambridge, UK

Silva Corvalán C 1994 *Language Contact and Change: Spanish in Los Angeles*. Oxford University Press, Oxford

Tsitsipis L D 1989 Skewed performance and full performance in language obsolescence: the case of an Albanian variety. In: Dorian N C (ed.) *Investigating Obsolescence*. Cambridge University Press, Cambridge, UK

N. C. Dorian

Language Policy: Linguistic Perspectives

The planned allocation of resources to language is but one of many instances of organized and purposive intervention into areas of social usage and social change. Although the discussion here will be restricted to governmental policy in connection with language, given that such governmental interventions affect the largest numbers, language policies are also adopted and implemented by nongovernmental organizations and interest groups. Indeed, even individuals frequently implement their own language policies of the various kind described below, and/or advocate such policies for more general adoption. Language policies are of two major types: (a) language status policies, pertaining to the desired or permitted public societal functions of particular languages, and (b) language corpus policies, pertaining to the desired or permitted features of languages themselves, whether in connection with their writing systems, orthographies, pronunciations, lexicons, and/or grammars. Language policy is both an applied and a theoretical field of endeavor and the influence of the one on the other has been both substantive and productive (Kaplan and Baldauf 1997).

1. Language Status Policy

Language status policies are either prohibitory, permissive, or facilitative. If policies are prohibitory they oppose and penalize the public use of particular languages for particular societal functions. Thus, the 'English Only' movement in the United States opposes any requirement that American governments (federal,

state, or local) use any language other than English in the discharge of their governmental functions (including direct services to the public, whether or not its members have a command of English). During the years of the fascist Franco regime in Spain, the regional Iberian languages (primarily Catalan, Basque, and Gallego) were prohibited in public, whether spoken or written. The Communist regime of the Soviet Union prohibited publication in, or the teaching of, Hebrew (thereby effectively rendering inoperative religious Jewish services or traditional schooling) due to that language's perceived theological and bourgeois (i.e., antirevolutionary) associations. One of the earliest examples of modern prohibitive language policy is the 'abolition' of the regional languages (pejoratively referred to as 'patois' or 'dialects') by the French Revolution toward the very end of the eighteenth century (Schiffman 1996, Tabouret-Keller 1999).

Permissive language policies are presumably 'neutral.' While they permit nongovernmental efforts on behalf of, or in opposition to, particular languages (or in conjunction with all languages other than the national or official languages which are, naturally, supportively treated), they neither obligate the government positively nor negatively with respect to such languages. During most of its history, the USA has had a permissive language policy, making it possible for citizens of various non-English mother tongues to advocate on behalf of these languages publicly, seeking for them recognized functions above and beyond any religious functions which some of them may have and which are separately protected by the Constitution. The search for additional functions may aim at and utilize both print and nonprint media, and may aspire to functions such as becoming the medium of instruction in the public education of their young, or advocacy of governmental use of these languages in discharging various official services to speakers of these languages (e.g., in the courts, in the fire, police, and health services, as well as in social services more generally).

However, the absence of supportive policies has meant that when the mother-tongue speakers of non-English languages finally learn English (as all of them have done within a generation, or at most two), the Government has permitted these languages to fade away in public as well as in private life, and has in no way been committed to their functional continuation as a potential resource for American diplomatic, commercial, or intellectual/cultural growth. Because of such passively permissive policies, even major world languages such as German, French, and Spanish have thus far failed to be intergenerationally continuous in the US, even though all of them predate the establishment of the United States on American shores. The unwritten nature of permissive policies has also meant that little if any prior supportive legislation has had to be annulled on the few occasions when prohibitive policies have been implemented (e.g., the post-World War I, anti-German or anti-foreign language legislation in Nebraska and in several other mid-Western states—later overturned as unconstitutional by the Supreme Court decision *Meyer vs. Nebraska*, 262 US 390, 1923).

Since to post-World War II sensibilities prohibitive policies smack of anti-minority discrimination, such policies have often been replaced by purportedly permissive ones. The effective differences are often minor or totally absent, since the lack of outright supportive policies usually leaves demographically, functionally, and attitudinally threatened languages bereft of tangible opportunity to reverse their deteriorating fortunes.

Supportive policies are either across the board (multifunctional), or functionally specific in their focus. The Swiss government's support of Romansh has been multifunctional, spanning support for nursery schools, elementary and secondary schools, courses for adults, the preparation of visual and auditory media, periodical publications, and book publications. The Canadian government's Ministry of Multicultural Affairs supports cultural programs, publications, creative artists and curricular materials preparation for the immigrant languages (termed 'heritage languages' to differentiate them from the founding languages—English and French—or the indigenous languages of Canada), although the latter alone are still responsible for providing most of the basic funding for the schools, theaters, and media that utilize their languages. During most of the 1980s and 1990s, Australia's language policy with respect to immigrant languages has gone beyond that of Canada by also providing funds for teacher training and certification, as well as for student competence certification, in connection with ethnic community supplementary schools—for LOTEs (languages other than English).

In addition, many of these very same community languages have been added to the foreign language programs of elementary and secondary schools servicing ethnic neighborhoods in large urban centers. Australia has also had an unusually effective program of supporting radio and television in many of the LOTEs. However, toward the end of the 1990s many of the above supportive policies have been curtailed or phased out entirely, as the Liberal Party has assumed the leadership previously in Labor Party hands. In the US, Title VII Bilingual Education remains the most obvious example of supportive policy toward other languages, even though this policy has had transitional purposes (subtractive bilingualism leading to English alone) as part of its avowed purpose. Since all three policies (those of the US, Canada, and Australia) obtain in English-mainstream and immigrant-based societies, the differences between their policies is indicative that the omnipresence of English (or any other language of wider communication) need not

necessarily lead only to prohibitory or permissive policies vis-à-vis sidestream or less powerful languages.

2. Reversing Language Shift

One of the most demanding (and, often, most disappointing) areas of language status policy-efforts deals with the attitudinal, demographic, and functional recovery of languages that have become seriously weakened by virtue of competition with their neighbors. Prolonged and debilitating reverses often lead to language death, on the one hand; and to re-ethnification, on the other hand. Indeed, both these outcomes may themselves be the conscious or unconscious consequences of prior, or even still ongoing, language policies, at the very same time that permissive or even supportive efforts are underway for one or more threatened or weakened languages within the very same polity. When the lost or weakened functions are narrowly specific (e.g., press, media, curricular recognition as subjects of instruction), similarly specific and delimited efforts can be undertaken to reverse or, at least, more nearly balance the social forces that are at play. This is not the case when intergenerational mother-tongue continuity (rather than mere language reproduction) is in question. Mother tongues (taken as the most frequent language of early home, family, neighborhood and community life) are, by definition, acquired early, quickly and intimately (i.e., affectively). As a result, it is difficult, for formal and authoritative policies to rationally determine either their acquisition, or their retention through the period from one generation's mother-tongue acquisition and the time some 20 years off when the next intergenerational transition occurs (Fishman 1991).

The difficulties encountered in the long-term planning of intimate and informal interactions (or even in the reinforcement of such interactions via formal institutions such as schools, governments, economies, and media) has resulted in very few successes along the lines of reversing language shift in twentieth-century historical perspective. The more recent partial successes along such lines by efforts on behalf of French in Quebec, Catalan in Spain, and the clearly unusual case of Hebrew in Palestine/Israel (where prestigeful, formal [written/read], and theological functions predated the language's revernacularization) have encouraged renewed and better informed efforts on behalf of Basque, Frisian, Maori, and even Irish. (The last named suffering from approximately 80 years of unsuccessful efforts to foster widespread intergenerational mother-tongue transmission, notwithstanding an array of supportive policies.) In order to achieve the desired results, the realm of intimacy-and-infancy must be successfully ideologized as to its consequences for language transmission, but even the institutions that are not normally effective within, nor focused upon, the infancy/intimacy realm (school, media, worksphere, and government) must be focused upon language interaction with home, family, and neighborhood processes; and the threatened language must be explicitly related to the rewards to which these institutions can lead. This difficult and long-term intergenerational feedback process—(a) within, and (b) upon the informal nexus of language transmission—has been conceptualized in terms of a Graded Intergenerational Dislocation Scale (Fishman 1991). This scale highlights the priority of re-establishing informal and intergenerational interactions, the interruption or attenuation of which is viewed as particularly mother-tongue dislocative (just as the higher numbers on the Richter earthquake scale indicate greater damage and threat to life), in contrast to dislocations of formal functions normally distant and compartmentalized from the mother-tongue transmission nexus, and not normally directly related thereto.

All in all, whether in connection with reversing language shift, or in connection with other policies pertaining to the authoritative recognition of particular languages in particular societal functions, the policy culture in connection to language issues closely corresponds to the policy culture that local authorities are accustomed to, or are obliged to, utilize more generally. Democratic authorities generally begin with open policy debate (including the exchange of promises and commitments, 'bargaining,' related to forthcoming other-than-language decisions and resource allocations). This is followed by policy adoption ('decision making'), codification (the formal wording of the law), and elaboration (the law as interpreted in practical terms of 'regulations') of the decisions reached; implementation of the foregoing; ultimate evaluation by other authorities or specialists than those entrusted with formulation and implementation of the adopted language policies (wherever such separation of formulation, implementation, and evaluation are built into the political culture); and, finally, iteration of language policy debate in subsequent areas highlighted for attention.

The culture of policy formulation and implementation in the language field is clearly an instance of the more general policy culture that is locally and contemporaneously in place. Where political stability is lacking, newly elected or empowered language policy authorities may counteract those that have been previously enforced and may undertake policy replanning along different and even quite contrary lines. Such replanning commonly has both status consequences and corpus consequences for the languages to which they pertain (Clyne 1997).

3. Language Corpus Policy

Languages that are not customarily used to discharge certain societal functions commonly lack the corpus features that are required for such functions, all the

more so if they have never (or not for prolonged periods) been associated with the implementation of these functions. Accordingly, language policies may be called upon to devise writing systems, establish spelling norms, promulgate standardized grammatical conventions (particularly for nonfiction prose), foster different styles for different kinds of writing (e.g., popular newspaper style vs. serious adult reading style for languages that have exercised neither function before), create or recover terminologies (e.g., for the natural sciences), or repair lexical gaps (e.g., in revising the pronoun system in order to reflect and foster democratic and egalitarian interpersonal address and interaction). Such corpus policies may be enacted by the very same governmental or quasi-governmental authorities that are involved in status policy formulation and implementation; but, because of their more technical nature, more commonly, they are entrusted or delegated by the latter to special boards, committees, or 'academies' of linguistic specialists. On occasion, various illustrious leaders and exemplary writers and intellectuals have been instrumental in propelling corpus planning in particular directions closely allied to their own preferred usage (e.g., Luther for High German, Pushkin for Russian, Bialik for Modern Hebrew, Perets for Yiddish, Cervantes for Castilian). This frequent prominence of individual 'influentials,' particularly in early corpus planning stages, is the reason why such planning efforts often precede those undertaken by governments, or even precede the establishment of governments and governmental corpus planning authorities (Rubin et al. 1977).

Although linguistically technical, the corpus planning of academies is commonly exposed to two different and possibly contradictory sources of influence. The first is the acceptability of corpus recommendations insofar as the intended target group is involved. In this connection, considerations of euphony, transparency, simplicity, absence or presence of cognates, or other esthetic and 'least effort' notions may prevail (Fishman 1997). Another powerful (and even a more powerful) direction of influence upon the efforts of corpus planning and corpus policy is the ideological interpretation given to such efforts by the political authorities who are engaged simultaneously in culture planning (i.e., in national identity planning) more broadly. Accordingly, corpus policy is regulated in accord with bipolar considerations such as uniqueness, authenticity, purity, and classicization, on the one hand, and modernization, regionalization, vernacularization, and internationalization, on the other hand. These emphases all differ to some degree, but they may all be subsumed under the general bipolar rubric of independence vs. interdependence. Since the facilitation of economic and political interaction within the modern world is generally a prime underlying motive in all modern corpus planning, the need to maximize linguistically both the interactive advantages and the independent challenges of modern life provide language policy, particularly with respect to its corpus concerns, with a built-in tension which can often be very anguished in those language-policy settings where the rejection of internal or external colonialism and the liberation from other foreign influences come to be of primary concern.

The 'authenticity' pressure upon late-modernizing languages that are faced either by confrontation with 'genetically related' Big Brothers (i.e., with long-established and well-modernized languages of similar lexical and structural characteristics to those of the late-modernizing language), or even faced by other, genetically similar, recent arrivals at the gates of modernity have long been studied. Languages in interaction and ethnicities in interaction are both dependent on considerable constructivist effort—namely policy interventions *per se*—in order to render them recognizably and consensually 'separate,' at least within their own authoritative frameworks. The policy and planning effort on behalf of authenticity-motivated distancing has been termed 'Ausbau' i.e., the building away of one language from another, by emphasizing the real, imagined or preferred differences between them. Endlessly exploitable sources of such differences may be found in older (archaic) periods of the language, in the varieties of the rural countryside that have suffered from fewer foreign/urban influences, or in a favored classical language of particularly acceptable historical, religious, or cultural significance. Ausbau efforts have been engaged in as part of the corpus policy of Hindi and Urdu relative to each other, of Macedonian vis-à-vis Bulgarian, of Yiddish vis-à-vis German, of Ukrainian vis-à-vis Russian, of Belorussian vis-à-vis both Russian and Ukrainian, of Moldavian (under Soviet influence) vis-à-vis Romanian, of Slovak vis-à-vis Czech, of Rusyn vis-à-vis Ukrainian, etc. More general 'purification' movements exist, e.g., for French (vis-à-vis English), for North Korean (vis-à-vis Mandarin borrowings), for Catalan vis-à-vis Castilian, for Czech vis-à-vis German, for the newly independent Baltic languages vis-à-vis Russian, for Navaho vis-à-vis English, for Hebrew vis-à-vis Yiddish and English, for Yiddish vis-à-vis Russian, etc.

Of course, all languages whose communities are engaged in interactive modernization with the Western World are exposed to linguistic influences as a result thereof. Since the material and intellectual gains attributed to the general interactive stance often seem incontrovertible, the corpus policies that attempt to limit or counteract the lexical and structural borrowings that are rooted in, and that derive from, interaction with and exposure to the West are faced by a dilemma from which there is no simple or easy escape. Indeed, one school of linguistics (the so-called 'Prague School') explicitly recognizes this tension but regards the 'national language' to be the supradialectal and generally written variety of the local linguistic rep-

ertoire which is most consistently monitored by nationally oriented linguists with the goal of keeping it as authentically pure as possible. In North India this process takes the approach of Sanscritization, in South India: of Classical Tamilization, i.e., of genuinely local classicals which can both intellectualize and de-Americanize or de-Westernize the corpuses of their modernizing local national languages. Language corpus authenticity (usually experienced as primordial) and language status Westernization (usually as a result of constructivist experiences) are often pursued simultaneously (or seriatim), eliciting tensions in either case.

4. *Linguistic Culture*

Obviously, status planning and corpus planning policies are also pragmatically (rather than only ideologically) closely interdependent. What would be the good of achieving the status goals of any given language policy if the corpus remained as inadequate to discharge or implement those statuses as it had, perforce, been before the status goals had been pursued? Conversely, what would be the good of endless corpus refinement if there were no societally recognized and protected functions in which the new lexicons, grammars, spellers, and writing systems could be put to use? Indeed, when one policy emphasis (whether status or corpus) is too far in advance of the other, then this lack of synchronization is itself a major policy problem. Accordingly, all language policy efforts, even in advance of any governmental recognition, require both linguistic expertise and political expertise in order to move ahead productively in connection with their language goals.

The diversity of talents required to bring language policies into operational rather than merely theoretical existence is well illustrated by the 'first congresses' that have been organized all over the world on behalf of different late-modernizing languages (Fishman 1993). Those organized primarily under academic auspices have least commonly attained their stated status goals (e.g., re the Mayan languages or Yiddish). Those convened by jubiliant nationalist movements have often found themselves without technical linguistic expertise or consensus, even when they were triumphantly swept into power (e.g., re Macedonian). Those organized by pre-existing cultural or political organizations have most frequently combined both the political constituencies and the technical scholarly expertise needed to attain their goals (e.g., re Catalan, Hindi, or Hebrew).

In any case, the goals and means employed in conjunction with linguistic policy reflect the historically deeply ingrained linguistic culture of its sponsoring power base. Language policies are expressions of a long-lasting 'linguistic culture—i.e., of the cultural myths and cultural values pertaining to the types of linguistic accommodations, programs, and efforts that are in accord with the brunt of local historical experience and aspiration (Schiffman 1996). Rather than being expressions of conscious or unconscious linguistic determinism, language policies are generally expressions of widespread convictions that the furtherance of crucial national policies more generally also require language status policies and language corpus policies that can contribute to national policies as a whole.

See also: Cultural Policy; Educational Policy, Mechanisms for the Development of; Language Endangerment

Bibliography

Clyne M 1997 *Undoing and Redoing Language Planning*. Mouton de Gruyter, Berlin

Fishman J A 1991 *Reversing Language Shift: Theory and Practice of Assistance to Threatened Languages*. Multilingual Matters, Clevedon, UK

Fishman J A 1993 *The Earliest Stage of Language Planning*. Mouton, Berlin

Fishman J A 1997 *In Praise of the Beloved Language: A Comparative View of Positive Ethnolinguistic Consciousness*. Mouton de Gruyter, Berlin

Fishman J A (ed.) 1999 *Handbook of Language and Ethnic Identity*. Oxford University Press, New York

Kaplan R B, Baldauf R B Jr. 1997 *Language Planning from Practice to Theory*. Multilingual Matters, Clevedon, UK

Rubin J, Jernudd B H, Das Gupta J, Fishman J A, Ferguson C A (eds.) 1977 *Language Planning Processes*. Mouton, The Hague, Netherlands

Schiffman H 1996 *Linguistic Culture and Language Policy*. Routledge, London

Tabouret-Keller A 1999 Western Europe. In: Fishman J A (ed.) *Handbook of Language and Ethnic Identity*. Cambridge University Press, New York, pp. 334–49

J. A. Fishman

Language Policy: Public Policy Perspectives

Language policy involves the governance of diversity arising from historical configurations, recent immigration, or both. Whereas it is possible for a state to adopt a neutral stance with regard to a variety of religions, or to make itself blind to ancestry, it cannot choose to become mute or deaf. Because of the widespread belief that language differences are a source of political instability and societal conflict, as well as that they constitute an economic liability, policies are commonly designed to reduce the actual diversity encountered among the population and to minimize its potential growth. This reductive tendency is reinforced by the costs of multilingual services and

institutions. Consequently, over three-quarters of the world's countries have central governments that are officially or effectively monolingual; but half of them contain linguistic minorities of 10 percent or more; and in about one-third of the cases, a majority of the inhabitants do not use the official language as their everyday speech. Even where central governments are officially multilingual, the languages recognized rarely encompass the full range of everyday speeches used by the population (Laponce 1987, pp. 95–102).

The development of languages, especially in written form, has been closely interrelated with state formation. For example, Chinese ideographic writing, elaborated over 4,000 years ago into a system for ruling a vast empire that encompassed a wide variety of spoken tongues, contributed to the unique resiliency of China as a polity despite many upheavals and regime transformations and still serves as an instrument of communication between groups whose spoken languages are not mutually intelligible. By contrast, the European pattern, which was set when the Athenian and Roman empires elaborated the vernaculars of their respective territorial centers into onomatopoeic written languages, rendered these polities more vulnerable to disaggregation into territorial entities constituted around other vernaculars (Diamond 1999 pp. 215–38, 322–33). Vastly facilitated by the advent of printing in the late fifteenth century, linguistic unification within territories moved to the fore as Europe's rival states sought to transform their populations into 'nations' sharing a distinctive identity (Anderson 1983, Gellner 1983). In both France and the United Kingdom, state expansion by way of the incorporation of peripheries or of adjoining foreign territories, the advent of mass armies, democratization, and industrialization, rendered linguistic unification ever more urgent (Nordman 1998, de Certeau et al. 1975, Weber 1976, Dillard 1985, Colley 1992). Both states imposed strikingly similar repressive language policies on their respective Celtic peripheries, and both also adopted analogous language policies in their colonies with regard to central administration and the education of indigenous elites.

By virtue of the hegemonic position the states of Western Europe achieved in the modern world system, linguistic unity came to be equated with modernity and emerged as a global norm that shaped policy among the states arising from European settlement in the Americas and Oceania, notably the United States, where the founders expressed concern over the persistence of German-speaking communities, as well as among the modernizing empires of the European periphery and, later on, the postcolonial states of Asia and Africa. But the fact that most of the latecomers had more heterogeneous linguistic configurations than Western Europe during the equivalent period of state formation rendered the pursuit of uniformity illusory. Paradoxically, Europe's equation of state with nation also fostered a contradictory norm, whereby distinct linguistic groups were entitled to political autonomy. The tensions between these two norms, as well as the discrepancies between them and socio-linguistic realities, are at the root of the wide variety of language policies observable in the world at large today and of the contentions they evoke.

Although language has long been of concern to states because of its instrumental and symbolic functions, issues of language policy have recently achieved unprecedented prominence as a concomitant of the expansion of governmental activity and of citizen participation, as well as of the advent of postindustrial conditions that enhance the value of linguistic skills (Bell 1973). Among older self-acknowledged multilingual countries such as Belgium, Switzerland, and Canada, historical arrangements reflecting earlier power relationships between the groups have been called into question. This is true also of situations involving surviving indigenous minorities such as the Saami in Norway and Native Americans in Canada, the United States, or the Amazon basin. Concurrently, the decolonization of imperial possessions in Asia and Africa has resulted in the emergence of a large number of new states with heterogeneous and complex configurations involving both indigenous and imperial languages. Previously established arrangements have been called into question also among the new states of the post-Communist world. The quantitative growth and expanding scope of international migrations has produced additional diversity in both old and new states, while changing attitudes toward the incorporation of immigrants have broadened the range of policy responses, notably with regard to the education of immigrant children. Beyond the state level, entirely new language questions are being raised by the formation of international organizations engaged in governance, notably the European Union. Finally, policy makers throughout the world must deal with the implications of the ascent of English as the first global language.

A basic policy choice is whether or not to declare one or more languages 'official.' Among older states in which one language was clearly dominant, as in the Netherlands, the United Kingdom, Italy, Thailand, or the United States at the time of the founding, the default stance was silence, reflecting the fact that the language's hegemonic status was taken for granted. In relation to this baseline, the specification of an official language expresses determination to maintain or reinstate the status quo in the face of changing practices or perceived challenges. In this vein, in the last two decades of the twentieth century, approximately half the states of the United States adopted laws or constitutional amendments declaring English their official language, and there were efforts to enact a constitutional amendment to that effect at the national level as well. In 1992 the French government also initiated a constitutional amendment establishing French as the national language. Among multilingual

states, 'officialization,' which usually involves an invidious distinction between privileged languages and others, signifies the establishment of a language regime providing general rules for policy making.

Whether or not a regime has been formally established, language policy involves at the minimum a determination of the language(s) in which public services are provided, ranging from the posting of street signs to the educational system and law enforcement; of the linguistic qualifications of the appropriate officials and government employees; of the language(s) in which citizens can exercise their rights, notably voting, contacting representatives, or officials, or defending themselves against charges; and of the language(s) in which citizens are expected to exercise their obligations, such as military service, paying taxes, or maintaining business records subject to inspection. Language policy usually extends also to the specification of criteria for membership in the society, as expressed in linguistic requirements for naturalization and occasionally for immigration itself. Beyond this, the advent of the welfare state has drawn government deeply into the private sector, notably to monitor the appropriateness of language requirements for various types of employment and to insure the availability of consumer information, such as labels, to the several language communities. State intervention may extend to the regulation of language use in printed and electronic media, and even in private speech. The persecution of human beings merely for speaking their mother tongue is an extreme but unfortunately not uncommon manifestation of language policy.

In all types of states and at all levels of governance, language policy is often acutely controversial because almost any institutional arrangement entails an uneven distribution of costs to the individuals affected, considered as members of distinct language groups (Pool 1991, De Swaan 1991). The costs are both economic and psychological because language not only carries considerable weight as capital in the determination of socioeconomic status (Bourdieu 1979), but also is intimately linked to the formation of personal and collective identities. Overall, language institutions reflect prevailing class and majority-minority relations. The development of policy in this as in other spheres tends to be path-dependent. Established conceptualizations, power relationships, and institutional arrangements, set a course from which it is difficult to depart; consequently, there is likely to be a considerable discrepancy between political and social conditions at the time these institutions were established and the present. Understandably, challenges to the status quo tend to arise from efforts by the disadvantaged to lessen the costs they bear. Characteristically, the resulting confrontations take on the features of a constant-sum game, even if the proposed policies do not impose very great economic costs on the larger community, because those who hold the upper hand perceive the change as a decline of hegemony. The relatively modest costs imposed on the majority tend to be inflated, and the new arrangements magnified into a threat to the community's integrity.

The agenda of language policy arises in the first instance from a population's actual socio-linguistic configuration, itself shaped by past policies. Relevant aspects include not only the number of languages and the size of the groups, but also their spatial distribution (territorially concentrated vs. dispersed); their sociopolitical power (languages that have achieved written form and are used in government, business, and science, vs. others); and whether the relevant groups are native or immigrant. Other significant elements of the sociolinguistic configuration include the language groups' location in relation to the poles of political and economic development (center vs. periphery), as well as the relationship of language differentiation to the overall stratification system and other cultural signifiers, notably religion and physical appearance or ancestry (Laponce 1987, pp. 120–35).

Another weighty factor is linguistic culture, the set of behaviors and beliefs a speech community has about its own language and the others with which it comes into contact, whose explicitly political aspects may be termed linguistic ideology (Schiffman 1996, p. 5, Stevens 1999). The culture or ideology of the dominant group(s) naturally carries extensive weight in the determination of policy. Of special import in the contemporary period are the cognitive elements of linguistic culture, notably scientific discourse. For example, in the early twentieth century, British psychologists reported, on the basis of studies of Welsh children, that bilingualism interfered with the development of intelligence. Their findings supported the maintenance of an educational system that brutally suppressed Celtic speech on behalf of total immersion in English. The harmfulness of bilingualism was subsequently confirmed by American psychologists investigating Japanese children in California and immigrant European children in New York. This theory prevailed until the 1960s, when Canadian research, undertaken in the context of a reformist national administration more responsive to French-speakers' grievances, not only exposed the studies on which it had been based as deeply flawed, but suggested instead that bilingualism produced beneficial effects (Hakuta 1986, pp. 14–27). The new theory was then quickly invoked by Hispanic advocates of bilingual education in the emerging US debate on the subject. Of special relevance in the Third World is the ambiguous cultural loading of the colonial language in postcolonial countries: on the one hand it is resented as a language of subjection, but on the other admired as a language of prestige, which is also instrumental for accessing the world at large. The resulting ambivalence is often manifested in the educational sphere. For example, Algeria has undertaken to replace French with Arabic in its elementary schools—without concessions to the country's large Berber-speaking

minority—but is maintaining French as the language of instruction at higher levels.

Language policy varies also as a function of the general political regime. The historical expansion of political participation to encompass the more modest strata of the population imparted to language minorities greater weight in the political arena, fostering demands to lower what they perceived as an unduly large share of language costs by providing public services in additional languages as well as reducing or eliminating linguistic obstacles to upward mobility by way of education. Similar challenges tend to arise among hitherto authoritarian countries in the course of democratization. Another important regime variable is its degree of centralization. By virtue of the prevailing normative equation 'language = nation,' where linguistic minorities are territorially concentrated, demands for greater linguistic equity tend to engender, sooner or later, aspirations to greater political autonomy, which are much easier to fulfill within a federalist framework (Switzerland, Canada) than a unitary one (e.g., pre-1970 Belgium). Indeed, democratization tends to propel hitherto centralized multilingual societies with spatially concentrated groups toward federalism (Belgium, post-Franco Spain). Devolution of authority to the regional level with concomitant control over linguistic policy generally satisfies demands for autonomy (Fishman 1986), but separatist aspirations may persist (e.g., Quebec), and occasionally result in the breakup of existing states (Czechoslovakia, Yugoslavia).

The combined effect of these factors in determining policy variation can be highlighted by way of cases illustrating characteristic situations encountered in the world today.

1. One-language Dominant 'First World' Democracies

1.1 United States

Despite the absence of a formal national policy, English reigned unchallenged at the federal level from the founding onward and was imposed as a matter of course on non-English-speaking dependencies such as Indian reservations, Hawaii, and Puerto Rico (Fishman 1986). However, in the course of the nineteenth century states and local bodies often diverged from this norm in response to local conditions: Spanish was recognized in New Mexico's constitution; numerous school boards provided a modicum of bilingual public education to their immigrant communities; the use of foreign languages in the press and private education was unquestioned; language requirements for naturalization were minimal, if they existed at all; and foreign language ballots occasionally appeared as well. German was particularly widespread among public institutions throughout the Midwest. However, hostility to the 'new immigration' from eastern and southern Europe of the late nineteenth century triggered demands to restrict admissions and to Americanize newcomers forcefully. The movement achieved considerable success in the wake of World War I, when German was in effect eradicated from the public sphere, including education. However, its triumph was limited by the *Meyer v. Nebraska* decision (1924), in which the US Supreme Court ruled on due process grounds that states did not have the authority to prevent the teaching of foreign languages (Wexler 1996, p. 346). Nevertheless, as noted earlier, the educational community overwhelmingly supported monolingualism on pedagogic grounds. Notwithstanding acceptance of cultural pluralism in other spheres, an English-only regime prevailed throughout the country for the next half century, except for the recognition of Spanish as the language of Puerto Rico in the New Deal era.

Policy changes, amounting to the elaboration of a mildly multilingual regime, were brought about by distinct developments in the post-World War II era: the civil rights movement; the massive importation of Mexican workers under the 'Bracero' program (1942–65); the enactment of a less restrictive immigration law in 1965; and the adoption of a generous policy toward refugees from the Communist world. The turning point was the Bilingual Education Act of 1968, introduced on behalf of Spanish-speaking pupils in the Southwest, which added a Title VII to the first large-scale federal program in support of local schools enacted three years earlier. This provided a small demonstration program that was vastly expanded following the *Lau v. Nichols* decision (1974), in which the US Supreme Court held the Civil Rights Act to mean that the failure to provide instruction in a language students can understand constitutes unlawful discrimination (Foster 1982, p. 293). Advocates of bilingual education seized upon the case to press for the expansion of their programs, and a task force subsequently convened by the US Office of Education to prescribe 'Lau Remedies' placed bilingual education in a preferred position (Leibowitz 1971, 1982, Schmidt 2000, p. 13). In the wake of these actions, as well as the continued flow of federal funding, bilingual education programs proliferated in a variety of forms, mostly designed to facilitate transition to English instruction, but with a minority dedicated to the maintenance of non-English mother tongues. Additional federal programs were launched in the 1990s to overcome the deficiencies of earlier ones.

Concurrently, a lobbying campaign initiated by the Mexican-American Legal Defense and Education Fund (MALDEF) to obtain coverage for language minorities under the Voting Rights Act of 1965 led to the addition in 1975 of Titles II and III, requiring that registration forms, ballots, and other election materials be provided in a language other than English if more than five percent of the voters in a given district

spoke the same non-English language, and if the district's English illiteracy rate surpassed the national average. Applicable to American Indians, Asian Americans, Alaskan Natives, and citizens of Spanish heritage, the legislation covered 384 counties. Titles II and III were reauthorized in 1992 in strengthened form. Parallel developments occurred with regard to the language rights of defendants, as well as of employees in both the public and private sectors. In a dramatic reversal of past policy, in 1990 the Congress adopted the Native American Languages Act, which committed the United States to preserve and enhance its indigenous languages. The development of foreign language media, especially Spanish, in response to market opportunities has also contributed to the institutionalization of a more multilingual configuration.

Contrary to contentions that the new policies reduce incentives to learn English, research indicates that recent newcomers rapidly learn English and that, as with past waves, use of the country of origin's language declines sharply in the second generation and almost disappears in the third (Stevens 1994, Espenshade & Fu 1997, Portes and Hao 1998). The steady shift to English is somewhat hidden from view because the number of American residents who used a language other than English at home rose by virtue of continuing immigration from 23 million in 1980 to nearly 32 million in 1990, with Spanish accounting for slightly over half of the total.

Not surprisingly, the dawning regime change provoked widespread opposition, often invoking the specter of conflict ridden multilingual countries (Petrovic 1997), and triggered moves to counteract the trend by officializing English at the state and national levels, as well as eliminating bilingual education. Hostility was especially high in California, which enacted several referenda to that effect, including Proposition 227 (1998) eliminating nearly all bilingual education; Arizona followed suit two years later. As against this, the rapid transformation of immigrants into electoral clienteles fostered support for the new regime. On balance, multilingualism of services is likely to persist, so long as substantial replenishment of foreign-language speakers by way of new immigration continues; however, the future of bilingual education is less certain. As of 1997, the United States had approximately 3.5 million students with limited English proficiency (LEP), a 57 percent increase since the beginning of the decade; but they had higher rates of repetition and dropped out of school four times more frequently than their English-fluent peers (United States General Accounting Office 1999). Research findings regarding the effectiveness of various approaches to the education of LEP children, including the effects of the elimination of bilingual education in California in favor of 'immersion,' were indecisive; but run of the mill transitional programs, many of which lacked qualified staff, were widely criticized for their ineffectiveness and possible perverse effects, notably fostering within-school segregation of newcomers.

1.2 France

Markedly different from American concerns, the issues driving French language policy focus on the changing role of regional languages such as Breton, Alsatian, Occitan, Basque, and Corsican; the protection of the national hegemony of French against the intrusion of English; and the promotion of the national language abroad as an instrument for enhancing France's international standing. Nevertheless, overall movement has also been toward a softening of the traditional strict monolingual regime.

Notwithstanding the Jacobin centralist tradition, successive republics have also cultivated the image of a nation consisting of 'small fatherlands' and embraced a domesticated version of regionalism in the form of folklore, often expressed in the regional languages. The Loi Deixonne (1951) provided for the limited teaching of these languages as optional subjects in public secondary schools. The identity quests of the 1970s fostered demands for the expansion of such programs. Endorsed by a new leadership generation within the Socialist Party, they were translated into policy when it gained power in the 1980s and again in the 1990s. As of 1997, about three percent of pre-university students in both public and private education were enrolled in regional language classes (Le Monde 1999, p. 12). However, in reaction to this, in 1992 a Gaullist government enacted a constitutional amendment declaring French the national language and invoked the amendment four years later as grounds for refusing to sign the European Charter of Minority Languages, on the grounds that this would undermine the hegemony of French (Wexler 1996). Language policy was made into an electoral issue the following year, when Socialist leader Lionel Jospin expressed support for linguistic and cultural pluralism. After his victory, France signed the Charter, albeit with some reservations. The Socialist government also granted an unprecedented degree of autonomy to Corsica and announced a plan to offer bilingual education to the island's population.

Initiated on a personal basis by Charles de Gaulle after he returned to power in 1958, the promotion of special ties with French speakers abroad constitutes a counterthrust to the ascent of English as a world language, itself denounced early on by de Gaulle and prominent literary intellectuals. The francophone policy, which encompasses France's former colonies, Belgium, Switzerland, and overseas communities of French descent, notably in Canada and the United States (Louisiana, New England), was endowed with a permanent administrative apparatus in 1966, entrusted to a distinct Secretariat of State in 1986, and

elevated to cabinet level in 1993 as one of the responsibilities of the Minister of Culture. Protection of French within France was institutionalized in 1975 by way of a law requiring its use throughout the public and private sectors, including contracts, marketing, and advertising. However, the measure proved difficult to enforce, and the courts insisted on a narrow interpretation of its authority. The Loi Toubon, enacted by a Gaullist government in 1994, further imposed the use of French by private citizens in all aspects of public discourse (Wexler 1996, p. 311). But although these actions have evoked considerable attention in France and abroad, English has continued to gain ground in many spheres of society, notably within the scientific and financial communities and at the level of popular culture.

The fact that France is as much of an immigration country as the United States, with a similar proportion of foreign-born residents, has had little or no impact with regard to language in the public sphere. Albeit attributable in part to the fact that a majority of the immigrants originate in the Maghreb, where French is widespread both as a language of instruction and as a lingua franca, the absence of change largely reflects the lasting power of France's assimilationist stance. No provisions have been made for public services in languages other than French, except to a limited extent with regard to court translators for the benefit of the accused. Bilingual education, even of a transitional sort, does not exist; the only concession made to immigrant communities at the elementary level is to allow the teaching of 'languages and cultures of origin' as an elective subject outside of regular school hours. In 1989–90, attendance ranged from 13 percent for children of Algerian origin to 35 percent for Turks (Tribalat et al. 1996, pp. 188–213).

2. *Multilingual 'First World' Democracies with Large Spatially Concentrated Groups*

2.1 *Belgium*

Issued from the vagaries of Europe's dynastic tribulations and united by the Roman Catholic religion, the kingdom of Belgium straddles the long-established boundary zone between Germanic and Romance languages. Around the time of the state's founding in 1830, approximately 60 percent of Belgians, concentrated in the north, spoke varieties of Flemish, closely related to Dutch; and another 30 percent, concentrated in the south, used mostly varieties of Walloon, a group of Romance speeches with more Germanic root words than standard French. However, following the introduction of French as a language of state by their Austrian Habsburg rulers in the eighteenth century and the subsequent annexation of the Belgian provinces by revolutionary France, the traditional elites of both regions, amounting to perhaps 10 percent of the population altogether, had by then mostly adopted French as their usual language. It is therefore hardly surprising that they unhesitatingly imposed French as the country's sole official language. French exclusively was used in government as well as secondary and higher education, limiting Flemish to local administration and terminal elementary education in the north, and relegating Walloon to the margin. The capital city, situated within the Flemish region, functioned as a French-speaking transmission center. The resulting linguistic inequality was compounded by uneven economic development, centering on the Walloon coalmining region. Hence until well into the twentieth century, Flemings striving for upward mobility had no choice but to learn French, and in time many of them abandoned their ancestral language altogether (Zolberg 1974).

Demands for change arose in the second half of the nineteenth century from the ranks of the Flemish region's expanding middle class, severely disadvantaged by the burden of having to function in French rather than in their mother tongue. Initially demanding recognition of Flanders itself as a bilingual region, they subsequently shifted to the establishment of official bilingualism at the national level. This was formally achieved in 1897. Nevertheless, Flemish retained in effect second-class status; most notably, university education continued to be conducted exclusively in French, and French remained the language of command in the army. In the early decades of the twentieth century, the more radical wing of the Flemish movement took on a nationalist cast, demanding the establishment of a federal system; a few even advocated secession, possibly followed by reunion with the Netherlands. However, Flemish nationalism was tarnished by its association with German occupation in the two world wars. In the mid-1930s, a language settlement was achieved whereby the country was divided into two monolingual regions, with a bilingual capital district and national administration; as part of this, the state university at Ghent was converted into a Flemish-language institution. However, language conflict returned to the fore in the post-World War II period, and Flemish power increased as the economic standing of the regions was reversed. After a lengthy period of perennial low-level conflict, in the 1970s Belgium finally transformed itself into a complex federal system, with a language regime founded on 'dual monolingualism' rather than bilingualism. The country was divided into two regions and Brussels with regard to economic decision making, as well as into linguistic/cultural communities (Flemish, French, and a small German one) with regard to educational and cultural policy. Parity prevailed at the center, with a mandatory equal number of cabinet posts allocated to each of the two major communities and parallel administrative services in the two languages (Zolberg 1976, Witte and Craeybeckx 1987, pp. 423–518).

2.2 Canada

Paralleling the status of Flemish in Belgium, Quebec French was long associated with backwardness in the eyes of Canadian officials and English-speaking businessmen, as well as their US neighbors, despite the near identity of its written form with standard French. Paradoxically, however, the conquered French Canadians fared better under Protestant British imperial authorities than the Flemings at the hands of their own countrymen. As part of the post-conquest settlement, in an attempt to secure French Canadian loyalties in the impending conflict with the unruly American colonies, the British granted the Roman Catholic Church and the Province of Quebec a considerable degree of autonomy. This enabled French Canadians to create from the very outset institutions of secondary and higher education that fostered the formation of a French-speaking professional middle class. Although many educated Quebecois learned English, outright language shift occurred only among those who left the province. After a period of repressive direct rule imposed in the wake of proto-nationalist agitation, the creation of the Canadian Confederation in 1867 accorded Quebec along with the other provinces a degree of self-government long unthinkable in centralized Belgium (Latouche and Poliquin–Bourassa 1977). Consequently, around the turn of the twentieth century French Canada possessed many more of the elements of a distinct society than did Belgian Flanders. Nevertheless, in the 1960s, lingering insecurity concerning the survival of Quebec's distinctive culture within the Anglophonic North American environment stimulated a movement for outright independence, which experienced electoral ups and downs in the last three decades of the twentieth century. This prompted in turn reforms to implement full linguistic and cultural parity at the federal level, as well as perennial negotiations to provide Quebec with yet greater autonomy within the Canadian framework (Burnaby and Ricento 1998).

3. Postcolonial Multilingual Countries

The array of former colonial countries in Asia and Africa that achieved independence in the second half of the twentieth century display a policy pattern that has been formulated in game-theoretical terms as a '3 + / − 1 language outcome' (Laitin 1992, 1993). The basic model is drawn from India. Britain's adoption of Urdu as the official language of colonial administration for the South Asian subcontinent in the early nineteenth century fostered a movement among Hindi-speaking provinces on behalf of parity for their own language. The resulting alignment prefigured the partition of India and Pakistan in 1947. However, within the Hindi-speaking zone, other languages such as Tamil emerged concurrently as rallying points for distinct regional nationalities. From the 1920s onward, the Indian National Congress adopted Hindi as the language of national business, but allowed its provincial committees to use regional languages at their own level. The arrangement anticipated India's post-independence policy; but because of widespread English bilingualism within the bureaucracy, English was maintained also as a 'temporary' accommodation. However, the growth of regionalism after 1947 challenged the status of Hindi as India's national language and concomitantly fostered support on behalf of English as a 'neutral' alternative. Hence, although the 1950 constitution provided for switching to Hindi as the exclusive national language after 15 years, opposition rendered this politically impossible and the status quo was maintained indefinitely (Dua 1993). The outcome was thus Hindi-English bilingualism at the national level, plus the appropriate Indian vernacular at the regional level (i.e., a total of three languages); but where Hindi was the vernacular, the outcome was 3 − 1, whereas for minorities whose mother tongue differs from the official regional language, the repertoire extends to 3 + 1 (Laitin 1992, p. 45).

Similar policies prevail among the postcolonial states of sub-Saharan Africa as well, notably in Nigeria, which has a federal structure similar to India's, as well as in Congo, to the extent that any policy may be said to exist, Senegal, and Ghana. Ivory Coast might have followed the '3 + / − 1' pattern as well, thanks to the widespread use of Dyula as a lingua franca; however, in the first four decades of independence the country's leadership insisted on maintaining French exclusively as part of a general policy of close association with the former colonial power, and their choice was reinforced by the association of Dyula with a widely resented trading minority, constructed as 'foreign.' Where there is no obvious vernacular, lingua franca, or pidgin that could serve to symbolize national values, remain politically neutral, and be widely understood, the European colonial language tends to remain the interregional link language, while vernaculars tend to prevail regionally, thus resulting in a two-language outcome. Examples include Zimbabwe, Mozambique, Namibia, and possibly South Africa, with Afrikaans figuring as a 'regional' or 'group' language. A major exception is Tanzania, which thanks to the prevalence of Swahili as a lingua franca for most and as a vernacular for many has been able to pursue a traditional European-type monolingual policy; as against this, neighboring Kenya is leaning toward Swahili-English bilingualism (Laitin 1992, pp. 119, 132, 141, Mazrui and Mazrui 1993, Sure 1999).

4. Language Policy at the International Level

Despite the absence of comprehensive authority at the international or global level, the imperatives of communication have fostered voluntary arrangements

regarding language and within international organizations such arrangements have achieved more formal status. French emerged as the international diplomatic language at the time of that country's European hegemony in the late seventeenth century; but English gained ground after the Napoleonic Wars in keeping with Britain's leading role within the Concert of Europe, and its standing was reinforced by the emergence of the United States as a major international actor in the twentieth century. Accordingly, the League of Nations used French and English as its official languages; however, reflecting the waning of European hegemony and aspirations to globalize the international community, the United Nations subsequently added Spanish, Russian, and Chinese (Wexler 1996, pp. 305–6).

Arrangements involving a greater exercise of authority amounting to 'international language policy' have been evolving since the launching of the European Economic Community in 1957. The original Treaty of Rome limited itself to declaring that language matters regarding the organs of the Community shall be resolved unanimously by the Council, which declared in 1958 that the official languages of the six original members would serve as its 'official' as well as 'working' languages. This served as a guiding principle for the organization's subsequent development. With the accession of Finland and Sweden, the number of official languages reached 11, but the combinations to be covered by the translation and interpreting services reached 110, and the likelihood of a substantial number of additional East European members affords the prospect of another quantum jump in complexity and cost. However, for purposes of internal communications, de facto pragmatic arrangements prevail: Irish and Luxemburgian have official status but do not serve as working languages, whereas Catalan—with more speakers than Danish—is recognized symbolically, but not officially. In practice, Eurocrats operate largely in French, German, and English, with the latter steadily moving to the fore (De Swaan 1993b, Kraus 2000, Wright 1999).

The emergence of English as a lingua franca is a worldwide and unprecedented development, fostered by globalization. However, the latter process is also promoting certain regional languages to international status, notably Spanish, so that the hegemony of English may be limited. The wave of the future may well be widespread multilingualism, where each language is assigned its own distinctive societal function; and this will probably be facilitated by the concomitant development of electronic translation programs. A likely pattern in practice and in policy will be '2 + / − 1,' with an increasing incidence of individual-level bilingualism involving a national language and one of a small number of world languages (English and Spanish, as well as possibly Arabic and Chinese). Whereas speakers of regional languages will use 2 + 1, those whose national language ranks also as a world language will have little incentive to learn another, and hence are likely to remain monolingual (2 − 1). Although these trends have stimulated the development of nongovernmental organizations urging the enactment of national and international policies dedicated to the preservation of 'threatened' minority languages, it is unlikely that they will succeed in arresting their decline (Fishman 1998–9, Laitin 1993, De Swaan 1998).

See also: Assimilation of Immigrants; Bilingual Education: International Perspectives; Bilingualism and Multilingualism; Culture, Sociology of; Hegemony: Anthropological Aspects; Hegemony: Cultural; Language and Ethnicity; Language and Society: Cultural Concerns; Language Endangerment; Language Policy: Linguistic Perspectives

Bibliography

Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. New Left Books, London

Bell D 1973 *The Coming of Post-industrial Society*. Basic Books, New York

Bourdieu P 1979 *La distinction: Critique sociale du jugement*. Les Éditions de Minuit, Paris

Burnaby B, Ricento T (eds.) 1998 *Language and Politics in the United States and Canada: Myths and Realities*. Lawrence Erlbaum Associates, Mahwah, NJ

Colley L 1992 *Britons: Forging the Nation 1707–1837*. Yale University Press, New Haven, CT

de Certeau M, Julia D, Revel J 1975 *Une politique de la langue: La révolution française et les patois*. Gallimard, Paris

De Swaan A 1991 Language politics in India and Europe: A comparison based on a model of conflict of language interests. *Journal of Education and Social Change* **V**(3): 13–32

De Swaan A 1993a The emergent world language system: An introduction. *International Political Science Review* **14**(3): 219–26

De Swaan A 1993b The evolving European language system: A theory of communication potential and language competition. *International Political Science Review* **14**(3): 241–55

De Swaan A 1998 A political sociology of the world language system (1): The dynamics of language spread. *Language Problems & Language Planning* **22**(1): 63–75

Diamond J 1999 *Guns, Germs, and Steel: The Fates of Human Societies*. W. W. Norton, New York

Diamond S 1982 *Historical Aspects of Bilingualism in the United States*. Unpublished paper, Immigration Research Program. Washington, DC

Dillard J L 1985 *Toward a Social History of American English*. Mouton, New York

Dua H R 1993 The national language and the ex-colonial language as rivals: The case of India. *International Political Science Review* **14**(3): 293–308

Espenshade T, Fu H 1997 An analysis of English-language proficiency among US immigrants. *American Sociological Review* **62**: 288–305

Fishman J A 1986 Bilingualism and separatism. *Annals of the American Academy of Political and Social Science* **487**: 169–80

Fishman J A 1998–9 The new linguistic order. *Foreign Policy* **113**: 26–40

Gellner E 1983 *Nations and Nationalism*. Cornell University Press, Ithaca, New York
Hakuta K 1986 *Mirror of Language: The Debate on Bilingualism*. Basic Books, New York
Kraus P 2000 Political unity and linguistic diversity in Europe. *Archives européennes de sociologie* **XLI**: 138–63
Laitin D 1992 *Language Repertoires and State Construction in Africa*. Cambridge University Press, Cambridge, UK
Laitin D 1993 The game theory of language regimes. *International Political Science Review* **14**(3): 227–39
Laponce J A 1987 *Languages and Their Territories*. University of Toronto Press, Toronto, Canada
Latouche D, Poliquin-Bourassa D 1977 *Le manuel de la parole*. Editions du Boréal Express, Montréal, PQ
Le Monde 1999 July 21
Leibowitz A H 1971 *Educational Policy and Political Acceptance: The Imposition of English as the Language of Instruction in American Schools*. Center for Applied Linguistics, Washington, DC
Leibowitz A H 1982 *Federal Recognition of the Rights of Minority Language Groups*. National Clearing House for Bilingual Education, Rosslyn, VA
Mazrui A M, Mazrui A A 1993 Dominant languages in a plural society: English and Kiswahili in post-colonial East Africa. *International Political Science Review* **14**(3): 275–92
Nordman D 1998 *Frontières de France: De l'espace au territoire*. Gallimard, Paris
Petrovic J E 1997 Balkanization, bilingualism, and comparisons of language situations at home and abroad. *Bilingual Research Journal on line* (http://brj.asu.edu) **21**(2–3): 103–24
Pool J 1991 The official language problem. *American Political Science Review* **85**(2): 495–514
Portes A, Hao L 1998 E Pluribus unum: Bilingualism and loss of language in the second generation. *Sociology of Education* **71**: 269–94
Schiffman H F 1996 *Linguistic Culture and Language Policy*. Routledge, London
Schmidt R 2000 *Language Policy and Identity Politics in the United States*. Temple University Press, Philadelphia
Stevens G 1994 Immigration, emigration, language acquisition, and the English language proficiency of immigrants in the United States. In: Edmonston B, Passel J (eds.) *Immigration and Ethnicity: The Integration of America's Newest Arrivals*. The Urban Institute, Washington, DC
Stevens G 1999 *The Ideology of Language and the Incorporation of Immigrants in the United States*. International Center for Migration, Ethnicity, and Citizenship, New York
Sure K 1999 Bilingual education on an uneven playing field: The Kenyan case. *Journal of Third World Studies* **16**(1): 147–63
Tribalat M, Simon P, Riandey B 1996 *De l'immigration à l'assimilation: Enquête sur les populations d'origine étrangère en France*. La Découverte, Paris
United States General Accounting Office 1999 *Report to Congressional Requesters. Public Education. Title I Services Provided to Students with Limited English Proficiency*. GAO/HEHS-00-25, GAO, Washington, DC
Weber E J 1976 *Peasants into Frenchmen: The Modernization of Rural France, 1870–1914*. Stanford University Press, Stanford, CA
Wexler L S 1996 Official English, nationalism, and linguistic terror: A French lesson. *Washington Law Review* **71**(2): 285–377
Witte E, Craeybeckx J 1987 *La Belgique politique de 1830 à nos jours: Les tensions d'une démocratie bourgeoise*. Labor, Brussels, Belgium
Wright S 1999 A community that can communicate? The linguistic factor in European integration. *Sociological Review* **42**: 144–75
Zolberg A R 1974 The making of Flemings and Walloons, 1830–1910. *Journal of Interdisciplinary History* **5**(2): 32–74
Zolberg A R 1976 Splitting the difference: Federalization without federalism in Belgium. In: Esman M (ed.) *Ethnic Conflicts in the Western World*. Cornell University Press, Ithaca, New York

A. R. Zolberg

Languages: Genetic Classification

The genetic classification of languages is a method by which language families can be identified, that is, a group of languages can be shown to have evolved from a single earlier language. A simple example is the Romance family, which includes Romanian, Italian, French, Catalan, Spanish, and Portuguese, all of which are later changed versions of the Latin language spoken in Rome two millennia ago.

1. Methodological Preliminaries

Genetic classification is based on the fundamental property of language: the arbitrary relationship between sound and meaning. Each word (or grammatical ending) may be thought of as a coin with meaning on one side and the sounds that represent that meaning on the other. With a few exceptions there is no intrinsic connection between the two sides. Any series of sounds can represent any meaning just as well as any other; and there are hundreds, if not thousands, of possible sequences for each meaning. It is therefore unlikely that two languages will independently pick the same (or similar) sequence of sounds to represent the same meaning.

If, then, one finds similar words (that is, words that are similar in both sound and meaning) in different languages there are four possible explanations: (a) common origin, (b) borrowing, (c) chance, (d) onomatopoeia. In classifying languages one seeks to use only those similarities due to common origin, weeding out those resemblances due to the other explanations. Borrowings may be identified by a number of linguistic techniques, the most basic of which is that fundamental vocabulary—pronouns, body parts, and natural phenomena—is rarely borrowed. Borrowings may also be detected by outgroup comparison; the Romance words in English always resemble French, never Spanish or Rumanian. Accidental resemblances will of course crop up once in a while, but numerous similar words, shared by the same set of languages, cannot be accidental. Finally, onomatopoeia is quite

marginal in language precisely because it violates the arbitrary sound–meaning relationship.

In classifying languages in this manner, using only resemblances due to common origin, one arrives at families (or taxa) each of which is defined by a unique set of exclusively shared innovations (synapomorphies in biological terminology). It is important to emphasize that this requirement of exclusively shared innovations applies to all families at all levels; the Romance, Indo-European, and Eurasiatic families are all defined by a unique set of such innovations.

2. History

The first evolutionary explanation for linguistic diversity is usually attributed to Sir William Jones, an English judge stationed in India, who observed in 1786 that similarities shared by Sanskrit, Greek, and Latin could only reasonably be explained by the assumption that all three 'have sprung from some common source, which, perhaps, no longer exists.' The kinds of resemblances that led Jones to this conclusion were the similar roots and grammatical endings shared by these languages. Jones also noted that Gothic, Celtic, and Persian probably belonged to the same group, or family, which is today known as Indo-European. It includes almost all the languages of Europe (excluding Basque, Hungarian, Finnish, and Estonian) as well as languages in Iran, Afghanistan, Pakistan, and India. Extinct languages such as Hittite (Turkey) and Tocharian (western China) also belong to this family.

There were in fact many precursors to Jones' genetic hypothesis; even before he identified the Indo-European family in 1786, other families—Romance, Celtic, Finno-Ugric, Austronesian, Bantu, Iroquoian, and Algonquian—had been recognized by others. During the nineteenth century additional families were discovered in various parts of the world, and historical work on those families already identified, especially Indo-European, was carried out. Of prime concern to scholars during the latter part of the century was the reconstruction of the original Proto-Indo-European language and the discovery of the regular sound correspondences that connected the various branches of Indo-European.

Other families discovered by the nineteenth century included Uralic (Samoyed, Hungarian, Finnish), Altaic (Turkic, Mongolian, Tungus), and Dravidian (Telugu, Tamil, Malayalam). In fact, even a genetic connection between Uralic and Altaic was widely accepted, and as early as 1817 Rask had pointed out similarities shared by Eskimo-Aleut and Uralic.

During the twentieth century taxonomic work at the higher levels of classification largely came to a standstill as historical linguists, who were almost always Indo-Europeanists, decided that Indo-European represented the temporal limit of the comparative method. They believed that classification could not proceed beyond Indo-European, because language changes so rapidly that after around 6,000 years (the putative age of Indo-European) there is nothing left to compare. Also required, according to this view, is a complete reconstruction of the proto-language, as well as reconstructions of all the intermediate languages between the original proto-language and the modern languages. It is this view that has dominated historical linguistics in the twentieth century, and still prevails at the start of the twenty-first century.

Yet there have always been, from the beginning of the century to the present day, a number of scholars who argued against the supposed isolation of Indo-European, pointing out that Indo-European did indeed share a significant number of roots with certain other families spread across Eurasia. At the start of the century Pedersen called this larger family Nostratic and this name is still used today, though the precise membership of Nostratic has changed during the century. Recent works on Nostratic include Dolgopolsky (1964, 1998), Illich-Svitych (1971–84), and Bomhard and Kerns (1994). In addition, Greenberg (2000) presents evidence for a family that he calls Eurasiatic which is similar to, but not identical with, Nostratic. Greenberg's Eurasiatic includes Indo-European, Uralic, Altaic, Korean, Japanese, Ainu, Gilyak, Chukchi-Kamchatkan, and Eskimo-Aleut. One of the most salient traits of the Indo-European family is a pronoun system in which the first-person pronoun is based on M and the second-person pronoun on T (e.g. English 'me/thee,' French *moi/toi*, Russian *menya/tebya*). As early as 1905 Trombetti pointed out that this same pronominal pattern was shared by other families spread across Eurasia and he maintained that the explanation could only be genetic. Also characteristic of Eurasiatic is a system of plurality in which K represents the dual (i.e., exactly two) and T the plural (i.e., more than two); this was the similarity that Rask noted in 1817. The Nostraticists have discovered hundreds of lexical roots characterizing this family and Greenberg (2000) presents 72 Eurasiatic grammatical elements. There are thus numerous grammatical and lexical items that are found in Eurasiatic languages, but not elsewhere; these are the exclusively shared innovations that define the family.

Without doubt the most important contribution to linguistic taxonomy in the twentieth century has been the work of Greenberg. Beginning with his classification of African languages (1963), followed by a classification of New Guinea languages (1971), and then Native American languages (1987), Greenberg provided an overall classification where none previously existed. In his African classification Greenberg proposed four families: Khoisan, Niger-Kordofanian, Nilo-Saharan, and Afro-Asiatic. While this classification was initially controversial (Greenberg provided neither the reconstructions nor sound correspondences demanded by the traditionalists), subsequent research showed it to be correct in all but a few details.

It is today the general framework for all African historical research.

Greenberg's classification of New World languages into just three families—Eskimo-Aleut, Na-Dene, and Amerind—has been the subject of intense debate during the past decade, with traditional historical linguists defending their belief that there are not three, but over 200 independent families in the Americas, with no hint of genetic connections among any of them. While the Eskimo-Aleut and Na-Dene families have long been accepted as valid taxa by virtually everyone the Amerind family provoked considerable controversy. Yet Greenberg (1987) provided over 300 lexical and grammatical comparisons among various Amerind languages. In contrast to the Eurasiatic pronoun pattern, M/T 'I/you,' the most common pattern in Amerind languages is N/M 'I/you.' Elsewhere in the world an N/M pattern is virtually nonexistent, yet in the Americas it runs through Amerind languages of both North and South America. Indeed, as early as 1905, Trombetti (1905) devoted an entire appendix of his book to the evidence for this pattern in the Americas; a decade later Sapir also noted the same pattern.

While this distinctive pronominal pattern is, by itself, a powerful indicator of the validity of Amerind, there is a lexical item that provides even more compelling evidence in support of Amerind (Greenberg 1987, Ruhlen 1994). Throughout the (Amerind) languages of the Americas one finds traces of a root T–N whose meaning is some kind of child: son, daughter, child, brother, sister, sibling, etc. The vowel connecting the two consonants seemed at first to vary haphazardly; subsequent research revealed that the particular vowel was correlated—originally at least—with the gender of the child: *t'ina* 'son, brother,' *t'ana* 'child, sibling,' *t'una* 'daughter, sister.' (The apostrophe after the *t-* indicates that this consonant was originally glottalized.) Such an idiosyncratic lexical item—involving the arbitrary association of a particular vowel with a particular gender—can hardly be explained by anything other than common origin. This single innovation, by itself, defines the Amerind family.

While the controversies concerning Amerind and Nostratic/Eurasiatic were swirling around over the past 15 years, yet another large and ancient family was proposed uniting languages and families that had hitherto been considered isolates like Indo-European. As always, there were precursors, including Trombetti, to the discovery of this family, but its modern conception began with the work of Starostin (1984), who proposed that the Caucasian, Sino-Tibetan, and Yeniseian families constituted a higher-level family that he named Sino-Caucasian. Around the same time Starostin's colleague, Nikoleav, argued that Caucasian was also related to Na-Dene, thus making Na-Dene a member of Sino-Caucasian, a name that was replaced by Dene-Caucasian. In the late 1980s Bengtson provided evidence that two language isolates—Basque (Spain) and Burushaski (Pakistan)—should also be included in Dene-Caucasian. The most recent articles on this new family may be found in Shevoroshkin (1991). Like Eurasiatic and Amerind, Dene-Caucasian remains controversial and the Basque affiliation has been attacked by several scholars.

Yet another controversy that has surfaced in the past decade has been the very question of monogenesis, that is, whether all extant (and historically attested) languages share a common origin. For traditional historical linguists, who believe that even a family as shallow as Indo-European has no recognizable relatives, the question of monogenesis cannot even be approached. But for taxonomists such as Greenberg the question can be investigated simply by asking whether such large and ancient families—Eurasiatic, Australian, Dene-Caucasian, Amerind—themselves have certain words in common. In other words, have some of the putative exclusively shared innovations that define these families really been inherited from an even earlier language? If so, they are not really innovations, but derived traits. Bengtson and Ruhlen (1994), again with Trombetti as the main precursor, discuss 27 roots that are attested around the world, in language families from Africa to the Americas. Two of the most striking examples are the roots TIK 'finger, one' and PAL 'two.'

A final controversy that has involved linguistic taxonomy in the last decade has been the discovery by Cavalli-Sforza et al. (1994) that the taxonomic structure of the human populaton, based on gene frequencies, correlates to a remarkable degree with the linguistic taxonomy outlined by Greenberg and others, including such heavily criticized groups as Amerind. In fact, Cavalli-Sforza arrived at precisely the same three families that Greenberg found using language, and similar correlations have been noted for the African families. These recent findings recall Darwin's (1871) observation that 'the formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel.'

The foregoing is but a brief summary of the history of genetic classification; for a complete history of linguistic taxonomy, as well as a complete classification of the world's languages, see Ruhlen (1991).

3. *Current Status*

Though traditional historical linguists estimate there are over 250 unrelated language families in the world (Campbell 1999), Greenberg posits only a dozen. Figure 1 presents a map of these 12 families. The following sections will address, in the main, outstanding unresolved issues. It should be noted, from the outset, that what is currently most lacking is the higher-level subgrouping of these 12 families. Conse-

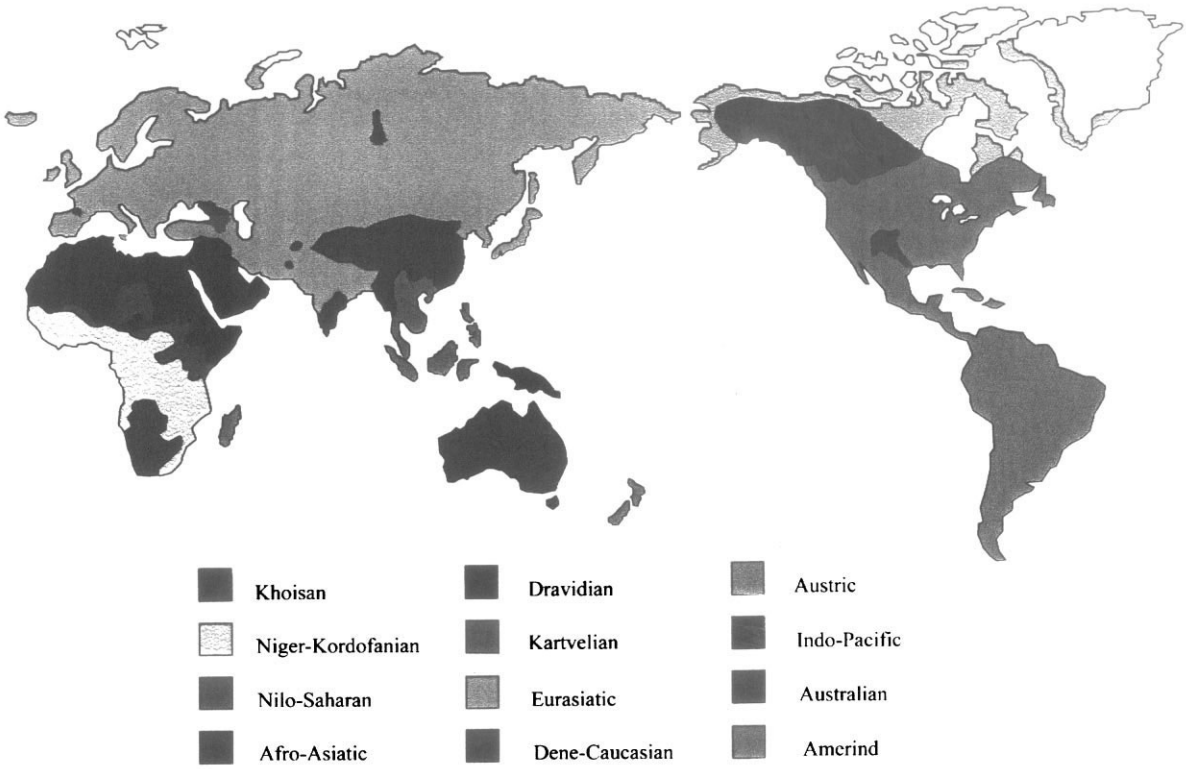

Figure 1
Map of the world's language families

quently, there is, as yet, no strong linguistic support for the 'Out of Africa' theory of human origins; the linguistic evidence, however, is certainly compatible with such a theory. On the other hand, global roots such as TIK 'finger' and PAL 'two' would appear to be incompatible with the multiregional theory of human evolution.

3.1 Khoisan

The presence of clicks in this family, but no others (except for borrowings such as in Zulu), complicates the comparison of this family with others. Internally there is a large southern group and two outliers in Tanzania (Hadza and Sandawe) that are quite different both from each other and from the southern group.

3.2 Niger-Kordofanian

It has been suggested that this family and Nilo-Saharan form a higher-level family that has been called Congo-Saharan. This possibility, which would unite most Black Africans in a single family, merits further study. Internally the highest levels of the family remain in some dispute. Greenberg posited a primary division into Kordofanian and Niger-Congo, with the latter divided into Mande vs. all the rest. Other scholars believe, however, that Kordofanian may well be closer to non-Mande Niger-Congo than is Mande; yet others consider Kordofanian, Mande, and Niger-Congo to be coordinate branches.

3.3 Nilo-Saharan

The proposed connection with Niger-Kordofanian was mentioned above. Internally Nilo-Saharan consists of around ten subgroups whose relationship to one another remains unclear.

3.4 Afro-Asiatic

Originally the Nostraticists included Afro-Asiatic in Nostratic, but, with the exception of Dolgopolsky, today most Nostraticists consider it a sister of Nostratic, rather than a daughter, which is essentially Greenberg's position. The internal subgrouping of Afro-Asiatic remains a focus of investigation, but the

divergent position of the Omotic branch seems favored. It is also possible that Semitic, Berber, and Chadic form a higher-level subgroup.

3.5 Eurasiatic

Ruhlen (1994) offers evidence connecting Eurasiatic with both Afro-Asiatic and Amerind; Starostin, however, has presented evidence linking Nostratic and Dene-Caucasian (Starostin 1989). The internal structure of Eurasiatic is unclear, but Greenberg posits a subgroup consisting of Korean, Japanese, and Ainu. The Nostraticists exclude Ainu from their family and consider Korean and Japanese as part of Altaic, rather than coordinate to it.

3.6 Kartvelian

The Nostraticists include Kartvelian in Nostratic; Greenberg considers it related to Eurasiatic but not part of it, pointing especially to similarities with Afro-Asiatic.

3.7 Dravidian

Most Nostraticists include Dravidian in Nostratic, whereas Greenberg considers it, like Kartvelian, to be related to Eurasiatic but not a member. Starostin considers Dravidian to be the most divergent member of Nostratic. The precise external relationships of both Kartvelian and Dravidian require further study.

3.8 Dene-Caucasian

Starostin's proposal for an external connection with Nostratic was noted above. Internally, Bengtson has proposed a subgroup composed of Basque, Caucasian, and Burushaski, and I have suggested that Yeniseian and Na-Dene also form a subgroup. The precise position of Sino-Tibetan in the family is unclear.

3.9 Austric

There are few proposed external connections. Internally, Daic and Austronesian seem to form a subgroup coordinate with the other two branches, Austroasiatic and Miao-Yao.

3.10 Indo-Pacific

Greenberg (1971) is the only comprehensive attempt to classify these languages. All other historical work on this very diverse family has been at a low taxonomic level. Essentially everything remains to be done. External connections with Australian have been suspected, but to date little evidence supporting this idea has appeared.

3.11 Australian

Possible external connections with Indo-Pacific were mentioned above. Internally Australian consists of some 15 subgroups, one of which—Pama-Nyungan—covers most of Australia; the 14 other subgroups are all found together in the extreme north. The reason for this peculiar distribution is not known. Internal subgrouping is unclear.

3.12 Amerind

Suggested links with Euroasiatic and Afro-Asiatic were mentioned above. Greenberg (1987) proposed 11 subgroups and suggested that three of them (Almosan-Keresiouan, Penutian, Hokan) formed a higher-level subgroup he called Northern Amerind. In South America he suggested that three branches (Macro-Carib, Macro-Panoan, Macro-Ge) form a Ge-Pano-Carib subgroup. The internal structure of the family as a whole requires further study.

See also: Evolution and Language: Overview; Evolution: Diffusion of Innovations; Evolution, History of; Genes and Languages; Historical Linguistics: Overview; Linguistic Typology; Linguistics: Overview

Bibliography

Bengtson J B, Ruhlen M 1994 Global etymologies. In: Ruhlen M (ed.) *On the Origin of Languages: Studies in Linguistic Taxonomy*. Stanford University Press, Stanford, CA, pp. 277–336

Bomhard A R, Kerns J C 1994 *The Nostratic Macrofamily: A Study in Distant Linguistic Relationship*. Mouton de Gruyter, Berlin

Campbell L 1999 *Historical Linguistics: An Introduction*. MIT, Cambridge, MA

Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

Darwin C 1871 *The Descent of Man*. Murray, London

Dolgopolsky A 1964 Gipoteza drevnejšego rodstva jazykovyx semei severnoj Eurasii s verojatnostnoj točki zrenija. *Voprosy Jazykoznanija* **2**: 53–63

Dolgopolsky A 1998 *The Nostratic Macrofamily and Linguistic Palaeontology*. McDonald Institute for Archaeological Research, Cambridge, UK

Greenberg J H 1963 *The Languages of Africa*. Indiana University, Bloomington, IN

Greenberg J H 1971 The Indo-Pacific hypothesis. *Current Trends in Linguistics* **8**: 807–71

Greenberg J H 1987 *Language in the Americas*. Stanford University Press, Stanford, CA

Greenberg J H 2000 *Indo-European and Its Closest Relatives: The Eurasiatic Language Family*. Stanford University Press, Stanford, Vol. 1, Grammar

Illich-Svitych V M 1971–84 *Opyt sravneniia nostraticheski iazykov*. Nauka, Moscow, 3 vols.

Ruhlen M 1991 *A Guide to the World's Languages*. Stanford University Press, Stanford, CA, Vol. 1, Classification

Ruhlen M 1994 *On the Origin of Languages: Studies in Linguistic Taxonomy*. Stanford University Press, Stanford, CA
Shevoroshkin V (ed.) 1991 *Dene-Sino-Caucasian Languages*. Brockmeyer, Bochum, Germany
Starostin S A 1984 Gipoteza o genetičeskix svjazjax sinotibetskix jazykov s enisejskimi i severnokavkazskimi jazykami. *Linguističeskaja rekonstrucktsija i drevnejšaja istorija vostoka* **4**: 19–38
Starostin S A 1989 Nostratic and Sino-Caucasian. In: Shevoroshkin V (ed.) *Explorations in Language Macrofamilies*. Brockmeyer, Bochum, Germany, pp. 42–66
Trombetti A 1905 *L'unità d'origine del linguaggio*. Beltrami, Bologna, Italy

M. Ruhlen

Laplace, Pierre-Simon, Marquis de (1749–1827)

Pierre-Simon de Laplace was born on 23 March 1749 in Beaumont-en-Auge into a well-to-do family of the Norman landed gentry which occasionally married into the nobility. Destined for the priesthood, he studied first at the College de Beaumont under the direction of the Benedictines of Saint-Maur. Endowed with an excellent memory and a keen intellect, Laplace seems to have excelled at the College, where he received the classical education of the time—the great Latin and French authors, in particular Virgil, Cicero, and Racine, whom he liked to cite and who are perhaps the origin of his obvious liking for well-turned phrases and his passion for physical astronomy. At the age of sixteen, he entered the College of Arts at the University of Caen for the two years of philosophy prior to the Degree in Theology. Here he discovered mathematical analysis in his dealings with natural philosophy, Euler, Clairaut, d'Alembert, and Daniel Bernoulli, and at the same time discovered his astonishing aptitude for handling formulae. On the advice of his tutors, he abandoned theology and moved to Paris in 1769, where he was very quickly noticed by d'Alembert. From then on, he was never to leave either Paris or analysis.

1. *Programme of Studies*

As Professor of Mathematics at the Ecole Royale Militaire, Laplace was literally to inundate the Académie des Sciences with mathematical papers, some fifteen in four years, certain ones of which ran to well over one hundred pages in quarto. On 24 April 1773, he was elected *adjoint mécanicien* at the Académie des Sciences in Paris. The subjects which he tackled were those which he would continue to explore for the whole of his scientific life, a life which was to prove exceptionally long, especially as he would continue publishing original and fundamental papers right up to his death in Paris on 5 March 1827.

Becoming very quickly aware of his intellectual superiority, Laplace set about tackling all the problems of his day in the three principal academies where exact sciences were fostered—Paris, Berlin, and St Petersburg. Indeed, one need only read the titles of his first papers to get a good idea of the state of mathematical sciences in the last quarter of the eighteenth century. First and foremost came differential and integral calculus, and in particular the study of differential equations and partial differential equations which governed the mechanics of solids as well as fluids and the stars. In a paper presented in 1773 Laplace 1912, Vol. 9, pp. 5–68), it was Laplace who was the first to demonstrate that in very general cases one could determine the whole integral of linear partial differential equations, and that in exceptional cases it was still possible to obtain from them particular solutions in the form of *defined integrals* that is, *integral transforms* which fitted with the initial data. This method would allow him, for example, in 1809, to solve the equation of heat in more general cases than those envisaged by Fourier.

The second subject tackled by Laplace equally concerns the solutions of differential systems, but in those cases where whole integrability is beyond reach, such as in the theory of the Moon or the large planets. These theories had not only eluded Newton, who had not used the infinitesimal algorithm in its full extension, but had also eluded Euler, Clairaut, d'Alembert, and Lagrange who were ingenious rivals in their use of it. Laplace, right from the outset, was to demonstrate his extraordinary faculty—often spoken of in terms of intuition—for reducing by approximation the irreducible difficulties present in equations dealing with the problems of multiple bodies. In March 1773, Laplace presented a paper (Laplace 1912, Vol. 8, pp. 201–75) on the 'principle of universal gravitation' in which he showed that Euler and Lagrange had been wrong in their planetary theory and that he, Laplace, at the age of twenty-four, through Newton's single principle that 'the planets gravitate against each other,' could show the absence of secular inequalities in planetary mean motions, so that the planets could not accelerate their motion around the sun irreversibly by themselves. Furthermore, they only deviated from their perfectly elliptical trajectories through periodic and temporary inequalities, thus assuring a stability in the 'system of the world' which Laplace would encounter again in all areas of his work. The mechanics of the planets is quasi-stable, and the God of Newton played no useful role in maintaining them within rational limits. The foundations for Laplace's astronomical work were established, a work which he never once doubted and which he would pursue from then on, right up to his death.

The third subject, the 'theory of chance,' is more surprising coming from a protégé of d'Alembert, but it

was of great topical interest in 1770. It was in fact in this year that the great academies suddenly resumed this subject, which had to some extent been abandoned for fifty years, and which was now brought back into fashion by works in political arithmetic—notably those of Süssmilch—and at the same time by the requirements of astronomy. Lambert, Euler, Daniel Bernoulli, Lagrange, and Condorcet simultaneously published important papers on *probability calculus*. It was therefore inevitable that the young Laplace would turn his efforts to it, and his first contributions were magnificent, notably the great paper 'on the probability of causes' published in 1774 (Laplace 1912 Vol. 8 pp. 27–65), which marked the beginning of Laplacian mathematical statistics. In this paper, one of the classic works of the theory, Laplace developed the analytical results of the proportionality principle from the probability of causes to that of effects, and, for the first time, proposed, as a criterion for estimation, a quantity known only from observations subject to unavoidable errors, this criterion serving to minimize the sum of potential errors through their respective probability. This method, which Laplace called the *method of situation* and which appeared to him to be the result of the nature of things, should, he believed, be applied to research in astronomical fields. In fact, it led to formidable mathematical difficulties which Laplace would, in the next fifty years, work on breaking down into a series of fundamental works which would culminate in 1812 in the publication of the *Analytical Theory of Probability* (Laplace 1820).

Analysis, *celestial mechanics*, and the *theory of probability* were three overlapping subjects which Laplace did not invent and which he already found in the academic papers of his time, but which he grappled with and never abandoned, to such an extent that he was to change the view scholars had formed of the world which surrounded them, a physical world, of course, but also a moral world from the moment when it could be converted into numbers.

2. *Celestial Mechanics*

The years which were to follow were a long succession of scientific triumphs. One after another, Laplace solved the most difficult problems in celestial mechanics posed since Newton, problems which had foiled the attempts of the most eminent scholars of the eighteenth century to such an extent that it was sometimes doubted whether the single principle of attraction could really account for the inequalities of motion—such as the secular inequality of the Moon, the inequalities of Jupiter and Saturn, the theory of the satellites of Jupiter, and so on—which were being brought to light by the increasingly precise observations of modern astronomy. Laplace's works as a whole—which were spread over twenty years—were brought together in a vast treatise, *Celestial Mechanics* (Laplace 1825), the publication of which began in 1799, and which Laplace would continue to expand and amend up to the end of his life. Concurrently, in 1796 he published *Exposition of the System of the World* (Laplace 1984), which was aimed at a wider readership and presented, in layman's terms, the principal results which demonstrated with the utmost clarity that Newton's law of universal gravity 'represents all celestial phenomena, right down to their smallest details' and that 'through this, empiricism has been entirely removed from Astronomy, which is now a major problem of mechanics.'

However, Laplace was well aware of the fact that whole areas of the system of the world existed where empiricism seemed difficult to exclude entirely, such as the theory of oceanic or atmospheric tides; in other words, the areas where 'irregular causes' occurred in an anarchic fashion which could, if one did not bear them in mind, conceal the true causes of phenomena and their fixed order. Laplace encountered this problem from the beginning of the 1770s to the moment when the theory of errors began to take shape. Observations of all parts of nature, even the most precise, were affected by 'unavoidable errors' which could prove themselves predominant if left to accumulate by chance, and remove all scientific value from the apparent objectivity. According to Laplace, it was therefore necessary to multiply the observations, adjust the value of the uncertain elements progressively, and reveal gradually the phenomena which were 'until this point shrouded in the errors of observations.' This progressive revealing could not be done without a method, and this is precisely the purpose of Laplace's probability calculus.

3. *The Analytical Theory of Probability*

For Laplace, probability theory was first conceived as a means of critically analysing data from observation, a critical analysis which gradually became more refined and revealed the true system of the world. Two periods in Laplace's development are usually distinguished. The first, which runs from 1774 to 1785, saw the development of Laplace's first method. In the spirit of Bayes, but essentially asymptotic in that it was independent of the special nature of *a priori* laws, it did not allow Laplace to deal with the major problem of the theory of errors, that is, to determine in the best possible way the elements known from a large number of indirect observations. It did, however, allow him to solve the population theory problems posed by Daniel Bernoulli and Condorcet by giving political arithmetic the scientific rigor it lacked, left open as it was to the most trivial empirical digressions; for example, in deciding whether in truth more boys were born in London than in Paris for the same number of births, or even whether the population of France was increasing or decreasing.

The second Laplacian asymptotic theory, this time non-Bayesian, began in 1810 after its author had solved one of the major problems of probability theory, the *central limit problem* as it is called today, where under very general conditions the total sum of errors is asymptotically normal, which allows the calculation of probability even if the particular law of errors is unknown. This remarkable theorem and its applications in the *method of least squares*, published the following year, makes up Chapter IV of Book II of the *Analytical Theory of Probability*, published by Laplace in 1812 (Laplace 1820), and bringing together his work on probability as a whole. This second great treatise from Laplace would be expanded with various different chapters until 1825. From 1816, Laplace gradually extended the applications of probability calculus into increasingly wider fields, anticipating a large number of the applications which the twentieth century would develop. The theory of probability, therefore, became for Laplace 'the most felicitous addition to the ignorance and weakness of the human spirit' as he asserts in conclusion to his *Philosophical Essay on Probability* (Laplace 1986), the first edition of which dates back to 1814 and the last to 1825.

Laplace did not limit the possibilities for the application of probability calculus to traditional fields such as games, the theory of errors, and annuities. From the moment when a shred of calculation seemed to him possible, whatever the field and however tenuous it may be, he got to grips with it and generally exceeded previous results by a great deal. For example, inspired by the works of Daniel Bernoulli, Laplace moved from population theory to a general reflection on the chance series of draws from a ballot box which, at first, only display disorder and chaos and gradually adjust and give way to the simplest and most admirable order, the irregular causes cancelling each other out to display the eternal action of constant causes. These reflections on order in chaos anticipate the work on statistical physics at the beginning of the twentieth century. What is more, Laplace, inspired by Condorcet, proposed a theory of decisions in majority voting where the basic concepts of modern test theory can be seen. Further still, inspired once again by Daniel Bernoulli, Laplace established a general result for the concave utility functions, and applied it to questions of maritime insurance and so on. One can also see many more things in the *Analytical Theory* (Laplace 1820), which make this work a monument of science for all time. (See Bibliography for complete studies on this subject.)

4. *Career*

After the flamboyant beginnings we have seen, Laplace was very quickly recognised by his peers, and if his arrogance had caused him some trouble at the beginning, he had an exceptional career. Elected in 1783 to the position of Associate Academician, he became a member of the Académie des Sciences in Paris in 1785. During the French Revolution, Laplace retired to a property near Paris to draft out *Celestial Mechanics* and kept himself at a distance from events. But with the fall of Robespierre, he returned to Paris and took an active part in the creation of three great scholarly institutions emerging out of the Revolution—the Institut de France, the Bureau de Longitudes, and the Ecole Polytechnique. Bonaparte, who was concerned about his image and the scientific glory of France, was to make Laplace, albeit short-lived, his First Minister of the Interior and soon one of the highest dignitaries of the Empire, the Chancellor of the Senate and Count of the Empire.

In 1814 Laplace voted with the majority of the Senate for the forfeiture of the Emperor. He was immediately made a peer of France, then marquis, by King Louis XVIII. In the numerous positions he held, Laplace made himself the herald and the defender of exact sciences. He involved himself little in politics, and when he did, only with questions he could reduce to numbers.

In his lifetime, Laplace enjoyed an immense scientific prestige. As founder of the Société d'Arcueil in 1807, he addressed all the scientific debates of his time, and created a brilliant school which would make his mark on part of his century, up to the point at which Laplacian physics of corpuscular fluids was superseded by positive mathematical physics and experimental physics (on this subject, see Gillispie 1997).

In 1788, Laplace married Charlotte de Courty (1769–1862) with whom he had two children, Charles (1789–1874), general and peer of France who died without descendants, and Sophie (1792–1813) who died in childbirth. Angelique (1813–1889), the only daughter of Sophie, was brought up by her grandparents and married the marquis de Colbert Chabanais and had three children, through whom the family line was established.

5. *Influence*

Laplace was one of the most influential men of science of all time, as much through his scientific work as through his two expository works which met with considerable success. There was no public or private library which did not possess them, and Laplace remains among the most cited scholars in present-day literature. His contribution to the formal assertion of mechanical determinism is sometimes summarized thus: 'All events, even those which by their insignificance do not seem to be due to the great laws of nature, are a consequence of them which is as necessary as the revolutions of the sun.' This introductory sentence from the *Philosophical Essay* (Laplace 1986) of 1814 which can already be found in one of Laplace's first philosophical texts written in 1773, is without doubt the most frequently occurring scholarly quotations in the whole of world literature.

In fact, as is well known, the assertion of determinism is a recurrent theme in the history of human thought, and countless authors in the eighteenth century in particular could be credited with it. Laplace's originality lies in producing a brilliant proof; that the tiniest disturbances in the system of the world—for example the slow acceleration of the Moon—are a necessary consequence of the single best established law of nature, that of universal gravity. It was because such a deduction was possible, as Laplace demonstrated, that determinism could be asserted and, even more, that it was then a question of one of the highest scholarly ambitions of humanity. Science in the nineteenth century became a race for precision made possible by the local determinism of phenomena in the whole of nature.

However, Laplace's influence extends well beyond that, particularly in the social and moral sciences of the nineteenth century, most notably in the work of Quetelet. Indeed, 'Laplacian determinism' was not limited to determinist phenomena. In the system of the world, there are numerous phenomena which are subject to irregular accidental causes which constantly alter the course of these phenomena. Necessity, order, and determinism are nevertheless not absent from them. For example—and it is an example dear to Pastor Süssmilch which Daniel Bernoulli took up again—the sex at birth does not seem to be subject to any determinism, yet one observes that always or almost always slightly more boys are born than girls as soon as one considers quite a large number of births. Laplace concludes that the invariability involved here is 'a result of regular causes which override the anomalies due to the chance' of irregular causes Laplace 1912, Vol. 7, p. xlviii. Chance is thus furnished with a special meaning; it is the combination of irregular causes which take effect in all directions, without favouring any one in particular. Regular causes, on the other hand, compel nature along enforced paths which become apparent in the long term. This is Laplace's *theory of constant causes* which one can clearly see in the works of other authors, for example Buffon, Moivre, and the Bernoullis, but also in the works of the ancient Stoics. It is Laplace, however, who, just as in the case of determinism in minor astronomical perturbations, provides this theory with a mathematical provision which justifies its implementation. This is the full import of his theorem of 1810; the variations due to *irregular causes* are enclosed in narrow limits determined by the Laplace formula, which gradually move closer to genuine elements in such a way that the constant causes determine them. The analysis of causes which Adolphe Quetelet undertook of the 'phenomena which organised beings present us with' is no different, except in that it integrates into the Laplacian constant causes–accidental causes schema a third type of causes, *variable causes* which, for example, produce seasonal effects, causes which Poisson, Cournot, and Bienaymé had envisaged in order to account for irregularly periodic statistical data. Quetelet's 'average man,' and the 'tendencies to crime,' can only be understood in the perspective of the constant causes in Laplace's *Essay*. In a still more obvious way, Poisson's 'law of large numbers' is a triumphant metaphor for Laplacian theory, or at least for certain aspects of it—showing as it does the order of constant causes within the apparent chaos of accidental causes.

The decision-making aspect of Laplace's work—the theory of probability seen as 'common sense reduced to calculation'—experienced a less glorious fate. After the Laplace school—Poisson, Cournot, and Bienaymé—had attempted to give it a more acceptable form in the first half of the nineteenth century, it was for a long time rejected, and with this rejection went an increasingly significant part of Laplace's work on probability. Causes were multiple and, moreover, 'variable' since this aspect a century later was reinstated at the highest level in the way that it is best known today—the theory of games, the theory of decisions, the theory of tests, and judgments based on samples. For Laplace, chance was also the name for our ignorance, an ignorance which was often incurable owing to those multiple accidental causes which distort the natural order of things and of our reason, so that the eminent role of the theory of chance, of probability calculus, is to 'criticise' the statistical data from all parts of nature; and even in the absence of data, by analogy, image, model, and calculation when possible, the theory still allowed one to find one's way more confidently through the maze of social and economic life. In France one would have to wait for Borel and his theory of games before one considered afresh along with Laplace that the theory of probability 'leaves nothing arbitrary in the choice of opinions and in the choice of decisions to make, every time one can determine, in one's way, the most advantageous choice.' Laplace's legacy—to perfect everything through the powerful means that mathematical analysis places at the scholar's disposal—is in no danger of being forgotten.

See also: Bayesian Statistics; Bayesian Theory: History of Applications; Decision Theory: Bayesian; Elicitation of Probabilities and Probability Distributions; Errors in Statistical Assumptions, Effects of; Measurement Theory: Probabilistic; Probability and Chance: Philosophical Aspects; Probability Density Estimation; Probability: Formal; Probability: Interpretations; Socialism: Historical Aspects; Statistical Methods, History of: Pre-1900; Statistics, History of; Unemployment: Structural

Bibliography

Gillispie C C 1997 *Pierre-Simon de Laplace 1749–1827. A Life in Exact Science*. Princeton University Press, Princeton, NJ

Hald A 1998 *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York
Laplace P-S de 1820 *Théorie Analytique des Prob (Analytical Theory of Probability)*. Ve Courcier, Paris
Laplace P-S de 1825 *Traité de Mécanique Céleste (Celestial Mechanics)*. Gauthier-Villars, Paris
Laplace P-S de 1912 *Oeuvres Complètes de Laplace (Complete Works)*, 14 vol, (1878–1912). Ve Courcier, Paris
Laplace P-S de 1984 *Exposition du Système du monde (The System of the World)*. Arthème Fayard, Paris
Laplace P-S de 1986 *Essai philosophique sur les probabilités* (*Philosophical Essay on Probability*). Bourgois, Paris
Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, Cambridge, MA

B. Bru

Large Data Sets in Statistical Computing

Statistics can be viewed as the scientific process of building a stochastic model which 'explains' a dataset. Data analysis is an inherently iterative (and preferably interactive) process with the similar goal of understanding a dataset; this process is not restricted to purely statistical (inferential) methods. Data mining is the process of searching a large dataset to find nuggets (interesting subsets or patterns). Because data mining is such a new field its precise aims are still changing (see *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)*). Unlike statistics or data analysis, data mining seems to require a large dataset. And, there is an emphasis on computer science issues such as algorithm design and database organization. As more statisticians become involved in this field it is changing to look more and more like data (or statistical) analysis for large datasets. Because our interest here is in large datasets, the distance among statistics, data analysis, and data mining is small. When datasets could be written by hand on a single piece of paper, one could reasonably contemplate 'all' possible models or at least a very large family of models for a dataset. As datasets have grown we can no long consider such a possibility.

Table 1
Units of measurement (SI prefixes) for large and small sizes with equivalent powers of 2 and 10

Decreasing	Increasing
10^{-3}–2^{-10} = milli- (m)	10^{3}–2^{10} = kilo- (K)
10^{-6}–2^{-20} = micro- (μ)	10^{6}–2^{20} = mega- (M)
10^{-9}–2^{-30} = nano- (n)	10^{9}–2^{30} = giga- (G)
10^{-12}–2^{-40} = pico- (p)	10^{12}–2^{40} = tera- (T)
10^{-15}–2^{-50} = femto- (f)	10^{15}–2^{50} = peta- (P)
10^{-18}–2^{-60} = atto- (a)	10^{18}–2^{60} = exa- (E)
10^{-21}–2^{-70} = zepto- (z)	10^{21}–2^{70} = zetta- (Z)
10^{-24}–2^{-80} = yocto- (y)	10^{24}–2^{80} = yotta- (Y)

In 1980, perhaps the largest dataset in the social sciences was the US Census; in 1990 it was still one of the biggest. The official 1990 Census dataset (the Public Law 94-171 dataset known as the 'PL94 data') consists of ten CD-ROMs, each containing roughly 600 megabytes of data for a total of 6 gigabytes; these are the block level counts for every Census block, plus higher level aggregations. The Summary Tape File 3A (STF3A) containing the long form data is larger; for example, the STF3A for the Commonwealth of Pennsylvania totals over 1 gigabyte itself. (Table 1; discussed in Sect. 2, gives the various units of size.) The 2000 US Census dataset is dwarfed by any number of other datasets and a few are mentioned here. An important factor influencing the recent rapid growth in the size of datasets is the automated collection of data. This process was begun in the physical sciences (think of high energy physicists collecting data on every nuclear event detected in one of their large atomic accelerators), spilled over into the commercial world (think of AT&T, the phone company, collecting data on every single phone call passing through their system), and is now pervasive, even in the social sciences.

For example, a market researcher collects point-of-scale (POS) data in a large retail chain (think of Wal-Mart or your favorite grocery store); it can easily produce a hierarchical dataset with tens or hundreds of millions of transactions per week, each containing tens of individual items. Allowing only 100 bytes per item, yields a dataset which grows (!) at the rate of 100 gigabytes per week.

As another example, a student of anthropometry obtains three-dimensional mesh images of human bodies (3D People) using an automated imaging system. An actual study in which each such image is more than 10 megabytes is described at http://www.sae.org/technicalcommittees/caesar.htm and indicates there will be a total of 24,000 images totaling over 300 gigabytes.

As a third example, a cognitive scientist studies human brain function using magnetic resonance imaging (MRI). The raw data collected on one subject in an hour or two exceeds four gigabytes. An active research center might collect several terabytes of raw data in one year.

As a final example, the World Wide Web (WWW) is, without question, the single largest social science dataset. Various estimates of its size exist; the variation in the size of those estimates is very large. As a working value let's assume it contains many tens of terabytes.

The methods by which one can study and learn about datasets of such sizes are simply not the same ones that were learned in Statistics 101. Many of the issues relevant to this point are discussed in this entry.

The next section gives definitions of the various terms of size and discusses them briefly. Sect. 3

describes why the speed of computation and the speed of data access (probably the more important for large datasets) matter. In Sect. 4 data types and methods of analysis also matter. Visualization as a primary tool for analysis is considered in Sect. 5.

1. Size Does Matter

First, a few words on words. In order to talk about the size of large datasets we need a vocabulary of sizes. An important fact in the computer world of today is that ten raised to the power of three, which equals one thousand ($10^3 = 1000$) is approximately equal to two raised to the power of ten which equals one thousand and twenty-four ($2^{10} = 1024$). The metric system uses prefixes to denote each three powers of ten (or roughly equivalently, each ten powers of two) as given in Table 1. The table also includes the standard abbreviation for each unit. As a specific example of a metric weight, one yottagram equals 10^{24} grams; it can be written as 1 Yg and is very heavy. In the other direction, one hydrogen atom, for example, has a mass of almost two yoctograms. By way of history, yotta and yocto (along with zetta and zepto) were adopted by the *Conférence Générale des Poids et Mesures* in 1991 as permissible prefixes used to modify SI units. (SI is an abbreviation for *Système International d'Unités*.) In the metric system, the seven base SI units are: the ampere, candela, kelvin, kilogram, meter, mole, and second. Other SI units are derived from these base units.

Here we are going to talk about bits and bytes. A bit (a term created by the late statistician John Tukey) is a binary integer, either a zero or a one. Eight bits in a group is a byte, the standard unit of measurement in the computer world. A half of a byte is a nibble (although this term is not used much anymore).

Table 2 gives some common descriptive terms to describe sizes of datasets together with some example storage devices with that size. This choice of names for sizes above the horizontal line was first given by Huber (1992). Four names have been added below the horizontal line for ease of reference; the reader will note that common language does not ordinarily distinguish among these four additional names so their order is somewhat arbitrary. Throughout the remainder of this article the word 'large' is used in its generic sense and not in the sense implied by this table.

Thinking back to the examples mentioned in the introduction, we see that each falls below the horizontal line in Table 2. The Census itself is roughly 'enormous,' although one can always quibble about double counting and data compression. For the POS data we can usefully refer to one year of such data as being 'enormous' and note that there are interesting things to be learned from multiple years of such data. For the 3D People data we see that the experiment cited there is 'enormous' and the potential obviously exists for larger such datasets. The fMRI data is also 'enormous,' again with the potential for growth as the engineers and physicists increase the capability of MRI technology. The WWW data seems to clearly fall between 'gigantic' and 'immense' and it continues to grow daily.

Table 2
Descriptive names and typical storage devices for datasets of various sizes (in bytes). (The terms ZIP and JAZ are trademarks of Iomega Corporation)

Descriptive name	Size	Storage device
	10^1	Fingers
Tiny	10^2	Blackboard
	10^3	Sheet of paper
Small	10^4	
	10^5	
Medium	10^6	Floppy disk
	10^7	Printed book
Large	10^8	ZIP drive
	10^9	JAZ drive
Huge	10^{10}	Hard disk
	10^{11}	
Enormous	10^{12}	Jukebox
	10^{13}	
Gigantic	10^{14}	100 terabytes
	10^{15}	
Immense	10^{16}	World Wide Web
	10^{17}	
Colossal	10^{18}	One exabyte

Note that physical scientists currently discuss even larger datasets; a number of groups are planning the collection and analysis of datasets that would be characterized as 'immense' or larger. There is no end in sight to this growth; technology provides continual downward pressure on the cost of data collection and storage, and upward pressure on the size of possible datasets that can be collected.

Current commercial statistical software cannot handle enormous data sets, and there is no particular reason that it should be able to. First, much of this software is designed to be interactive; this means that an entire data set must be held in the random access memory of the computer and immediately limits the size of possible data set to merely large. Second, all of this software has a history which typically began with one individual writing a small program to perform some specific analysis on a particular data set; size was of no concern. When the software became commercial, it was 'ruggedized' and a language was invented but no special concern was given to the size of data sets. The net result is that little attention is paid to the speed of algorithms.

2. Speed Also Matters But Not Too Much

Building on tools developed to process smaller datasets, Huber (1994) has identified five possible appro-

Table 3
Computing time as a function of problem size and algorithm speed (assuming 1 problem is solved in a microsecond). Time units are second (s), day (d), and year (y). The last column gives the multiplicative factor or additive term showing the speedup that would be achieved by a thousand-fold increase in computer speed. The entries marked by *** are extremely large and were not computed exactly

Rate	10	100	1,000	10,000	100,000	1,000,000	Speed up
N	10^{-5}s	10^{-4}s	10^{-3}s	10^{-2}s	10^{-1}s	1s	*1,000
N^2	10^{-4}s	10^{-2}s	1s	10^2s	10^4s	10^6s	*33
N^4	10^{-2}s	10^2s	12d	300y	***	***	*6
N^6	1s	12d	300y	***	***	***	*3
2^N	10^{-3}s	10^{22}y	***	***	***	***	+10
3^N	10^{-2}s	10^{40}y	***	***	***	***	+6

aches to the computational problems engendered by the statistical analysis of large datasets.

Use existing algorithms
Select smaller subsets
Process summaries
Derive smaller datasets
Develop new methods

Each of these approaches is discussed further.

Table 3 gives the running time required by various algorithm speeds on datasets of various sizes assuming that one million problem instances can be solved each second. The lesson to be learned is that lots of things are essentially impossible. Despite the sophisticated algorithms developed by computer scientists, and despite the tremendous speeds achieved by current computer hardware there are many problem-size/algorithm-speed combinations that simply cannot be realized in practice. Using the equivalence of one thousand and 2^{10}, and Moore's law, which predicts a doubling of computer speed every 18 months, the last column of Table 3 shows what one can expect from 15 years of continued technological improvement. Obviously and sadly, technology is not the answer.

Turning back to Huber's five approaches we can ask, 'What is possible?' Thinking about the first approach and restricting our attention to data which are contained in a rectangular array (a data matrix), the short answer is that we can essentially perform any algorithm that is linear in the larger dimension (usually the rows or 'cases.' This means that simple summary statistics can be computed, tables of counts, regressions (conditional expectations). Beyond that essentially nothing is possible on datasets of the size we are contemplating.

To think about this more carefully, suppose that we have a matrix with r rows and c columns with a total of $n = r \times c$ entries. Today a fast workstation can achieve 1 billion operations per second. We can ask for column sums and expect an answer in a few seconds if $n = 10^9$, no matter what the value of c. On the other hand, a regression calculation takes $r \times c^2$ calculations. If $r = 10^6$ and $c = 10^3$ then this requires 10^{12} calculations and will take several minutes. If $r = 10^5$ and $c = 10^4$ then it might take several hours. Repeating this arithmetic with $n = 10^{12}$ suggests that for $r = 10^9$ and $c = 10^3$ the regression calculation will take a week. The reader can consider other combinations of r and c. Of course, CPU speed is not the only constraint. With datasets of this size one needs to worry about the speed of data transfers from disk and the details of operating system performance such as memory paging and swapping.

This latter point is very important. Currently, the fastest data transfer speed from a hard disk drive into the main memory of a computer is no larger than 10 MB/s. More than one day will be required in order to read in an 'enormous' dataset. A 'gigantic' dataset could take a year without special tricks (simple tricks such as multiple disks, or multiple disks and multiple CPUs).

The second approach is 'select smaller subsets.' The beauty of this approach is that it can handle any size dataset no matter how large; simply take a random subsample and proceed. The drawback is that for large datasets this subset will contain a vanishingly small fraction of the data. If the subset happens to not contain some interesting part of the data, the analysis will obviously miss it too. On the other hand, if the subset happens to contain some really unusual feature the analysis can be easily misled. And, for data miners, any fairly rare combination is very likely to be overlooked because it won't appear in the subset; they won't find the nuggets.

Working with selected summaries of the data, Huber's third approach, has the great advantage that the analyst has a lot of control over the procedures used to produce the summaries, and hence the number of summaries produced. It also has the advantage that the lack of homogeneity in large datasets can be dealt with by using the subset identifier as a covariate (see Section 5). For the POS data set, it is natural to study shopping baskets. Consider simply the number of items in a basket. I am currently working with a POS dataset in which there will be roughly 10^7 summaries

generated from a dataset containing over 200 GB of data; analysis of these summaries will be a serious task in itself.

One example of Huber's fourth approach occurs when studying migration from census data. In the USA the census divides the country into roughly 5 million 'blocks.' These are amalgamated into 'block groups' and larger collections. The migration study needs to construct a table counting the number of individuals who have moved from location i to location j. Constructing this table is a daunting task; it requires processing records on the hundreds of millions of people appearing in each of two censuses. The size of the resulting table is determined by the analyst; for state-to-state migration studies it is only 50 by 50; for block-to-block studies it would be 7,000,000 by 7,000,000. Constructing the table could take massive quantities of computer time; analyzing it will take much less.

Devising new procedures, Huber's fifth suggestion, is a job for the statistical and computational scientists and will not be discussed further here.

3. Types of Data

The single most important type of data is the data matrix. We normally think of it as being organized with the columns being 'variables' and the rows being 'cases.' The discussion in Sect. 3 focused on the $r \times c$ data matrix as the basic unit. It is very often the case that large datasets have an intrinsically more complex structure. Mutli-media datasets are the most common example. Here we discuss a few of the more common types.

The generalized data matrix is a hierarchical dataset where some variables are only reported (observed) if other variables take on certain values. Often, for purposes of analysis hierarchical datasets are 'unrolled' into a data matrix. This has the unfortunate property of making them substantially larger to go along with the motivating property that processing a matrix is often a straightforward task that can be performed with pre-existing software.

With the variety of digital data collection devices currently available, ranging from television cameras, to digital audio recorders, to POS terminals, the kinds of datasets collected has an equally wide range. Television and other digital cameras can produce two-dimensional color images with a high spatial, temporal, and color resolution. Complex systems of cameras can produce three-dimensional surface mesh images (this was most recently demonstrated during the football game, Superbowl XXXV, when multiple cameras allowed the game analysts to replay full three-dimensional video and interactively change the point of view) and medical imaging devices such as computed tomography (CT) or magnetic resonance imagers (MRI) can produce fully three-dimensional volume images. The combination of sound with pictures can produce multimedia data.

4. Methods of Analysis

Small datasets are often homogenous. Large datasets are never homogenous. The major task in the analysis of a large dataset is finding homogenous subsets.

Raw data is raw. A fundamental and very time-consuming part of any analysis of a large dataset is getting it cleaned up and organized so that it can be analyzed.

At its heart, statistical data analysis is not simply a procedure which is applied to data but rather it is an iterative process involving repeated attacks on a dataset to tease out the message. For large datasets this creates a compound problem. The dataset is large but each step of the process produces additional datasets which are typically the same size as the original or larger.

Large datasets almost always, as a corollary of their inhomogeneity, are best explained by complicated models. Complicated models are hard to understand. The human visual system is ideally adapted to make sense of complex inputs and this leads naturally to the consideration of visualization methods for the analysis of large datasets. There are many points of view, many different technologies, and a wide variety of methods available for use. A quick review of traditional graphical methods is given by Jacoby (1998). An interesting collection of papers which sketches the history of dynamic methods is given by Cleveland and McGill (1988). A wonderful coffee-table book on graphical methods is Tufte (1983). (See also: *Graphical Methods: Presentation*; *Graphical Methods: Analysis*.)

5. Discussion

The problems of statistical analysis of large data sets are myriad. Here we have tried first to give a few examples of the very largest data sets. We have considered the limitations of computers and computer programs in providing analyses for these data sets. In particular we have looked at five methods for dealing with such large data sets and tried to clarify the costs and benefits of each.

There is little statistical literature to which one can turn, beyond that already cited, to learn more about statistical analysis of very large data sets. The proceedings of the *Annual Interface Symposium*, now in its thirty-third year, is one place to begin; the proceedings of the *Knowledge Discovery and Data Mining* conference, now in its seventh year, is another. Also, the annual Conference on Very Large Databases has

much that is relevant. Sadly, those readers who want to do something as conceptually simple as compute least squares linear regression estimates for a data set with 500,000 regressors, are on their own at present. Perhaps the situation will change before too long.

See also: Sample Surveys: Methods; Sample Surveys: Nonprobability Sampling

Bibliography

Cleveland W S, McGill M E (eds.) 1988 *Dynamic Graphics for Statistics*. Wadsworth and Brooks/Cole, Pacific Grove, CA

Fisher R A 1925 *Statistical Methods for Research Workers*. XYZ Publisher, London

Huber P 1992 Issues in computational data analysis. In: Dodge Y, Whittaker J (eds.) *COMPSTAT: Proceedings in Computational Statistics*. Physica-Verlag, Vienna

Huber P J 1994 Huge data sets. In: Dutter R, Grossmann W (eds.) *COMPSTAT: Proceedings in Computational Statistics*. Physica-Verlag, Vienna

Jacoby W G 1998 *Statistical Graphics for Visualizing Multivariate Data*. Sage University Press, Thousand Oaks, CA

Nicholson W L, Carr D B, Hall D L 1980 The analysis of large data sets. *Proceedings of the Statistical Computing Section of the American Statistical Association*, 59–65, Washington, DC

Tufte E R 1983 *The Visual Display of Quantitative Information*. Graphics Press, Chesire, CT

Tukey J W 1962 The future of data analysis (Corr: **33**, 812). *Annals of Mathematical Statistics* **33**: 1–67

W. F. Eddy

Large-scale Social Survey, Analysis of

Large-scale social surveys provide an inexpensive and easily accessible source of information on a wide range of topics including education, health, economics, demography, politics, and criminal justice. These surveys may serve as a primary data source, a preliminary data source to be analyzed prior to conducting a more tailored survey, or in combination with smaller, more specific data sources. After summarizing the advantages of using large-scale social surveys, we discuss how to confront the challenges they present.

There are many advantages to analyzing data from large-scale social surveys. The data are usually extensive in content and coverage of the target population. Sample sizes are often very large and the data are often easily obtained from the sponsoring agency on CD-ROM or from the internet, in many cases at no cost to the researcher. Moreover, these surveys are typically conducted by large government agencies or private survey organizations that have experienced interviewers, well-designed and tested survey instruments, and the expertise to design efficient sampling plans. As a result, the response rates are typically high (e.g., 75 percent–90 percent) and sophisticated nonresponse adjustments are often incorporated into the data to try to reduce nonresponse bias that results from systematic differences between respondents and nonrespondents. The sample data are also often reweighted, using poststratification, to agree with known population distributions of characteristics such as age, sex, and race. Finally, many surveys contain panel (longitudinal) components in which the same sample units are interviewed repeatedly over time, allowing researchers to follow individual changes over time.

The analysis of large-scale social survey data also presents several difficulties. One of the greatest technical challenges arises because sampling in these surveys is usually carried out according to complex multistage designs that combine stratification, cluster sampling, and unequal probability sampling. Many surveys that are repeated on a regular basis also rotate respondents in and out of the sample. Often point estimates of simple estimands such as population means and proportions can be calculated using a minimal amount of information about the sampling design, but standard error calculations must incorporate the key sampling design elements explicitly. Survey weights, provided along with the survey data, contain important sampling design information. Correct interpretation and use of these weights is key to the analysis of large-scale social survey data, and particularly key to standard error estimation. Analysts must also be aware of the impact of nonresponse adjustments—such as weighting class adjustments for unit nonresponse and imputation for item nonresponse—on standard error estimates. Complex multistage sampling must also be taken into consideration when calculating confidence intervals for population parameters. In particular, the degrees of freedom used when calculating confidence intervals using a t-distribution depends on the sampling design, not solely on the sample size. Methods of incorporating survey weights into estimates of simple population parameters such as means, totals, and proportions are well established; however the appropriate approach is less obvious when fitting more complex statistical models such as linear regression models and loglinear models. We address each of these issues in the following sections.

1. Complex Sample Designs

Although simple random samples are the easiest to design and analyze, large-scale social surveys often take advantage of complex multistage sampling designs to increase the precision of the resulting estimates. These designs typically involve multiple stages of sampling. For example, the first stage of sampling may select a sample of counties; the second stage may select a sample of subregions within each

sampled county; the third stage may select a certain number of sample units (e.g., individuals, households, or hospitals) within each sampled subregion. Units selected at the first stage of sampling are called primary sampling units (PSUs). Units selected at the final stage are called ultimate sampling units. One reason for multistage sampling is to minimize the travel costs of interviewers in face-to-face surveys. Another reason is that a sampling frame (a list of all the ultimate sampling units in the population) may not exist for the target population, but may be constructed sequentially as needed. For example, a list of all school-age children in a state may not be readily available, but a list of all school districts in a state may be available; then for a sample of school districts, a list of all the schools in those districts may be obtained; then for a sample of schools, a list of children may be obtained.

Multistage designs often involve cluster sampling in which groups of sampling units are selected, rather than individual units. Examples of cluster sampling include selecting several students in a class, several households in a city block, or several patients in a hospital. Cluster sampling tends to increase the variability of survey estimates compared with a simple random sample of the same size. This occurs because units in the same cluster tend to be correlated and, as a result, each additional unit sampled within the same cluster provides less additional information about the overall population parameters than an element selected at random from the population. However, due to travel costs, frame construction costs, and other overhead costs associated with each sampled cluster, a simple random sample may be much more expensive than a cluster sample of the same size. Therefore, given a fixed budget, a cluster sample may provide the most efficient estimates.

Another common design element used in large-scale multistage surveys is stratification, which involves dividing the population into nonoverlapping segments, called strata, and sampling independently within each stratum. Typical stratifying factors include region of the country and urbanicity (urban, suburban, rural). If the variable of interest tends to take on different values, on average, in each stratum, then stratified sampling can lead to more efficient estimates. In particular, the sample can be optimally allocated so that a larger sample is obtained in strata where there is larger variability in the variable of interest. Another reason for stratification is to ensure adequate representation of all segments of the population, avoiding the possibility of a simple random sample that, by chance, omits part of the population. Often certain strata, such as those containing a large proportion of racial minorities or elderly people, are sampled at a disproportionately greater rate (oversampled) in order to ensure an adequate sample size for research hypotheses that involve these groups. For example the British Crime Survey oversamples inner city areas by a factor of two (Mirrlees-Black et al. 1998). Sometimes a census is conducted in some strata, called self-representing or take-all strata. This type of stratification is particularly useful for very skewed populations that contain a few large units and many small units (e.g., Hidiroglou 1986). Stratification can also be combined with cluster sampling. For example, neighborhoods may be stratified by geographic region and then a sample of households within neighborhoods may be selected; or schools may be stratified by size and then a sample of classes may be selected from each school.

Unequal probability sampling is often used when sampling clusters. In particular, clusters are usually selected with probabilities proportional to their size (PPS sampling), with larger clusters having greater probabilities of selection. The size measure used will depend on the survey, but, as examples, a school can be measured in terms of the number of students enrolled, a company can be measured in terms of the numbers of employees, a farm can be measured in terms of acreage, and a city block can be measured in terms of the number of households. PPS sampling will lead to more efficient estimates whenever cluster totals are proportional to the size of the cluster, which is often the case. For example, larger schools tend to have more students with learning disabilities, larger companies tend to have larger revenue, larger farms tend to have a larger crop yield, larger city blocks tend to have more single parents. Unequal probability sampling is also sometimes used for convenience. For example, random-digit-dial telephone surveys actually select households with probabilities proportional to the number of voice telephone lines in the household. Selecting a simple random sample of households would require a list of the individual households, rather than a list of telephone-numbers, and this may be much more difficult to obtain or construct than a list of telephone numbers,

Finally, many large surveys that are repeated on an ongoing basis (e.g., monthly) use rotation groups. Rather than select new nonoverlapping samples each month, rotation groups remain in the sample for several months in a row (Fienberg and Tanur 1983, Iachan and Jones 1987). Every month a new rotation group enters the survey and an old one drops out so that each month only a fraction of the sample units are new. This reduces costs by not having to draw a new sample each month (there are administrative costs with drawing a new sample, locating the new sample members, explaining the survey to them, etc.) and also provides some longitudinal information. Rotation groups present no complication to cross-sectional analysis of a single wave of data but can complicate analyses that combine several months of data since the same sample units may appear multiple times in these combined datasets. Analysis should then account for correlations between repeated measurements on some sample units. In contrast, some surveys, called rolling surveys (Kish 1998) are specifically designed with

nonoverlapping samples each month so that the resulting data can be analyzed for a single month or combined over several months to form larger datasets.

2. Survey Weights

The simplest analysis of survey data assumes that all elements in the population are equally likely to be included in the sample, Unfortunately, this is not generally true for two reasons. First, complex sampling designs often lead to unequal selection probabilities, as discussed in the previous section. Second, even if units are equally likely to be selected, they generally have different probabilities of response; for example, in household surveys, women are usually easier to reach and more likely to cooperate than men.

Survey weights are used for correcting for known or expected differences between sample and population: each unit i is given a weight, w_i, and the population mean $\bar{Y}$ of any variable y is estimated by the weighted sample average,

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

As the formula indicates, only the relative values of the weights are relevant. The weights are usually calculated by the survey producers and provided along with the survey data. An informal interpretation of these weights is that w_i is proportional to the number of units in the population represented by respondent i. Estimates of population proportions are calculated using the same formula, where y_i is a variable taking on values of 0 and 1.

Survey weights in large-scale surveys are often constructed so that sample estimates agree with current census estimates of key variables such as age, sex, and race, through a process called poststratification and raking (Oh and Scheuren 1983, Gelman and Carlin 2001).Voss, Gelman, and King (1995) detail the various design and weightings strategies used by various national polling organizations for pre-election surveys. Survey weights are also often adjusted to account for nonresponse, through a process called weighting class adjustment (Oh and Scheuren 1983, Kalton and Kasprzyk 1986, Bethlehem 2001).

In most complex multistage samples, sampled units have different survey weights. However, it is possible to design these samples so that all sampled units have the same weight. These designs are called self-weighting (Cochran 1977, p. 91) because the usual unweighted estimators, the same ones used for simple random samples, will produce unbiased estimates of population parameters such as means and proportions. However, the usual unweighted standard error estimation formulas are not necessarily appropriate in these designs. Therefore, although self-weighting designs may simplify point estimation calculations, they do not simplify standard error estimation calculations.

Finally, many surveys produce separate sets of weights for different types of analyses. For example, the US General Social Survey (GSS) (Davis and Smith 1992) provides separate weights for household-level estimates and for individual-level estimates. The GSS is designed to give each household an equal probability of inclusion in the sample so that the sample is self-weighting for household-level variables. However, since only one individual in each household is interviewed, people living in large households are less likely to be interviewed than people in smaller households. Therefore, even though each household has an equal chance of being selected, individuals have different chances of being selected depending on the size of their households, with these inclusion probabilities varying by the type of survey (Gelman and Little 1998). Longitudinal studies may also have separate cross-sectional and longitudinal weights. For example, the 1993 US Panel Study on Income Dynamics (PSID) provides both 'cross-sectional analysis weights' and 'longitudinal analysis weights' (Hill 1992). Cross-sectional analysis weights can be used to obtain unbiased estimates of population statistics for a given year. Longitudinal analysis weights, on the other hand, can be used to estimate models of changes over time. In particular, the 1993 PSID longitudinal analysis weights assign non-zero weights only to individuals who were respondents during each of the preceding five years. The Canadian Survey of Income and Labour Dynamics also provides both longitudinal and cross-sectional weights (Statistics Canada 1997). When using surveys with multiple sets of weights, consult survey documentation for advice on appropriate choice of survey weight.

3. Standard Errors

Although estimates of simple population parameters such as means and proportions can easily be constructed as weighted averages using survey weights, the corresponding standard error estimation procedures are not included in most standard statistical packages. When the data have been collected from a survey with a complex sample design, standard error estimates of survey statistics derived under simple random sampling assumptions generally underestimate the true standard error, resulting in confidence intervals that are too narrow and hypothesis tests that are too likely to reject the null hypothesis. In particular, most standard statistical packages can handle 'analytic' weights that are used in regression estimation when the data values themselves each have different (known) variances, and 'frequency' weights that are used to store data concisely by recording the number of times an observation is duplicated in the dataset rather than recording each observation individually, but these weights should not be confused

with survey weights. Incorrectly treating survey weights as analytic weights or frequency weights will generally result in unbiased point estimates but underestimated standard errors. Moreover, ignoring the weights altogether and using a standard statistical package to perform an unweighted analysis of complex survey data will generally yield biased estimates of population parameters and underestimated standard errors. Furthermore, standard statistical packages do not account for stratification, cluster sampling, or adjustments such as post-stratification and raking. Brogan (1998) and Wang et al. (1997) illustrate some of the potential consequences of incorrectly analyzing survey data using standard statistical packages.

As a measure of the effect of the sampling design on the variance of survey estimates some surveys report design effect estimates. The design effect (deff) (Kish 1965, p. 162, Lohr 1999, p. 239) measures the increase (or decrease) in variance obtained using the given sample design relative to the variance that would have been obtained if a simple random sample of the same size had been used. For example, Davis and Smith (1992, p. 41) report that the average design effect for 60 variables in the GSS is 1.76, with slightly larger design effects on average for demographic variables which tend to have more clustering and smaller design effects, on average, for attitude variables. Similarly, Hill (1992, p. 67) reports design effects for estimates from the PSID ranging from 0.95 to 2.53, with most in the range 1.1 to 2.2. A design effect of 1.76 indicates that the variance of an estimate from the current design is 1.76 times greater than the variance that would have been obtained using a simple random sample of the same size. The design effect can be used to determine the 'effective sample size' by dividing the actual sample size by the design effect. The effective sample size is the number of observations that would yield the same level of precision from a simple random sample. For example, a design effect of 1.5 implies that the precision obtained by sampling 1500 units under the current complex sampling design could be obtained by sampling only 1000 in a simple random sample design. The design effect does not, however, account for differential costs of the two sampling designs.

Several software packages are available to analyze data from complex multistage surveys. Summaries and reviews of these packages are provided by Lepkowski and Bowles (1996), Cohen (1997), Carlson (1998), and Zaslavsky (2000). Stata (StataCorp 1999) and the latest release of SAS (An and Watts 1998) are multipurpose statistical packages that include procedures for analyzing complex survey data. Because variance estimation calculations for sampling designs that use without-replacement unequal probability sampling are complex, most survey software packages approximate these variances using formulae appropriate for unequal with-replacement sampling. This approximation is usually acceptable since, for these survey designs, the resulting variance estimates will slightly overestimate the true variance, resulting in slightly conservative confidence intervals and hypothesis tests. Finally, survey software packages generally use variance formulae that are technically only appropriate when the survey weights represent only the sampling design and do not incorporate nonresponse or poststratification adjustments. A few packages (WesVarPC and VPLX, see Lepkowski and Bowles 1996) use a variance estimation method call replication or resampling (Wolter 1985, Chap. 2, Korn and Graubard 1999, p. 29) that can incorporate the impact of nonresponse weighting adjustments and post-stratification adjustments. To take advantage of these capabilities, data users must be provided with 'replicate weights' (e.g., Korn and Graubard 1999, p. 34) by the data producers. For example, the US Census Bureau releases replicate weights for the Survey of Income and Program Participation.

Finally, many large-scale surveys provide generalized variance functions (Korn and Graubard 1999, p. 218, Lohr 1999, pp. 308–10) in summary reports of the survey data. To reduce publication costs and calculation effort, these surveys do not publish standard errors for every estimate in the summary report. Instead, generalized variance functions are provided to enable data users to calculate standard errors for their estimates of interest. These functions estimate the standard error of an estimate as a function of the estimate. These standard error estimates are very approximate and their results should be interpreted with caution. Direct calculation of standard error estimates, using the raw data, survey weights, and other sampling design information is always preferable to using generalized variance functions. However, generalized variance functions permit standard error estimation when either the raw data or the sampling design information cannot be released for confidentiality reasons.

4. *Confidence Intervals*

When constructing confidence intervals for population parameters using survey data, it is usually assumed that survey estimates of means and proportions are approximately normally distributed, so that the usual normal confidence intervals or, for small sample sizes n, t-distribution intervals with $n-1$ degrees of freedom are appropriate (Cochran 1977, p. 27). For complex multistage sampling, the degrees of freedom for t-intervals is typically equal to the number of sampled primary sampling units minus a correction, c, where $c = 1$ for unstratified cluster sampling and c equals the number of strata for stratified cluster sampling (Korn and Graubard 1999, p. 62). The normality approximation may break down for estimates of proportions that are very close to 0 or 1. In these situations, Korn and Graubard (1999, p. 65)

provide approximations to exact binomial confidence intervals that can be used with data from complex multistage sampling schemes.

5. Statistical Modeling with Survey Data

Much has been written on the use of the sample design and weights when fitting statistical models such as linear regression, logistic regression, loglinear models, and survival models to survey data (Fienberg 1980, Groves 1989, Sect. 6.7, Little 1991, Pfeffermann 1993, Lohr and Liu 1994, Winship and Radbill 1994, Korn and Graubard 1995, Lemeshow et al. 1998). In general, it is recommended to include, as explanatory variables in any regression-type analysis, all variables that are used in the design or the weighting of the survey. These variables include stratum indicators, cluster indicators, and demographics (such as age, sex, ethnicity, and education) that are used to construct the weights. It is thus important in such an analysis to know how the weights were constructed, not just their numerical values. If all these variables are included, one can fit the model without survey weights (see DuMouchel and Duncan 1983, Korn and Graubard 1999, p. 177). It may then be necessary to put the weighting information back into the model—after having estimated the parameters—to estimate population quantities of interest. For example, Gelman and Little (1997) present an example in which estimates are obtained for each of the states in the US, summing over demographic categories, which requires the use of census demographic breakdowns for each state. Difficulties can arise when the number of clusters or strata is large relative to the sample size or when stratum or cluster indicators are not released in public-use data for confidentiality reasons. It can also make sense to perform both weighted and unweighted analyses (DuMouchel and Duncan 1983, Groves 1989, p. 290, Lohr 1999, p. 365; see also *Sample Surveys: Model-based Approaches*). Large differences between the two analyses may indicate model misspecification.

Modeling is also used for 'small-area estimation' in which inferences are made about subsets of the population (e.g., small geographic regions or demographic segments) whose sample sizes are so small that simple weighted mean estimates are too variable to be useful. Hierarchical Bayes and related methods have been useful here (Fay and Herriot 1979, Dempster and Raghunathan 1987), and one area of ongoing research is combining these methods with classical weighting adjustments (Gelman and Carlin 2001).

6. Missing Data

In a sense, all sampling is about missing data, since the goal is to learn about the entire population from a subset. Usually, however, missing data in sample surveys refers to unit nonresponse, which occurs when some sampled units do not participate at all in the survey, either because they refuse or cannot be reached for interview, and item nonresponse, which occurs when a sampled unit responds to some but not all questions in the survey. Survey organizations usually adjust for unit nonresponse by weighting, as discussed earlier in this article.

Item nonresponse is typically handled in one of three ways (Little and Rubin 1987). In 'complete-case' analysis, one works only with the data from respondents who provided responses to all questions of interest. 'Available-case' analysis is the other extreme, where, for each question, one uses all the responses in the data. Both these approaches have the serious problem that item nonresponses do not generally occur completely at random, so it can be misleading to simply ignore them (Rubin 1976). A more flexible approach to item nonresponse is 'imputation': filling in values for missing responses.

Imputation can be done in a number of ways, and for large-scale surveys, it is often performed by the survey organization itself. Missing values may be imputed using single or multiple imputation. In single imputation (Kalton and Kasprzyk 1986 Little and Rubin 1987, Sect. 4.5), a single value is imputed for each missing value. In multiple imputation (Rubin 1987), m plausible values are imputed for each missing value to reflect the uncertainty about the missing data (usually m is between 5 and 10). These plausible values create m completed datasets that can each be analyzed seperately as if they had complete response. Simple combining rules (Rubin 1987, p. 76) can be used to combine results from each of these m separate analyses into one overall estimate along with its standard error. If data have been imputed using single imputation, special variance formulae (e.g., Rao 1996) are needed to account for the extra variability introduced into survey estimates due to the uncertainty in the missing values. Most large-scale surveys are currently singly imputed, but multiple imputation is becoming more popular because of the relative ease of analyzing multiply imputed datasets compared to singly imputed data. For example, the US National Health and Nutritional Examination Survey plans to release a multiply-imputed dataset (Schafer et al. 1993, 1996).

Related to imputation are various data-adjustment and correction procedures. For example, survey questions such as income that have occasional very high, unreliable responses can be 'topcoded'—that is, truncated at threshold values. Difficulties with analyzing topcoded data can sometimes be avoided by working with medians or quantiles (e.g., upper fifth, second fifth, etc.), as these are less sensitive than averages to extreme values.

7. Summary

Large-scale social surveys are an important source of information for a wide range of topics. In analyzing such surveys, it is important to be aware of the

complexity of the sampling design and the data adjustments that are used by survey organizations, including weighting to adjust for differences between sample and population and imputation to fill in missing responses. For estimating population means and proportions, the analyst should calculate weighted averages, using survey weights. In regression analysis, it is best to include, among the predictors, the variables that are used in the design and the weighting. Hierarchical models can be used to obtain inferences for small subpopulations.

In general, simple standard error estimates, which assume simple random sampling, are too small because of stratification, clustering, and weighting. These calculations can often be adjusted reasonably well by using design effects reported in the survey documentation or in previous analyses. For more involved analysis, it is advisable to use statistical packages that are specifically designed to correctly handle complex survey data.

See also: Longitudinal Data; Longitudinal Data: Event–History Analysis in Discrete Time; Sample Surveys: Survey Design Issues and Strategies; Sample Surveys: The Field; Survey Research: National Centers

Bibliography

An A B, Watts D L 1998 New SAS procedures for analysis of sample survey data. In (*SUGI*) *SAS User Group International Proceedings*. SAS Institute, Cary, NC

Bethlehem J G 2001 Weighting adjustments for ignorable nonresponse. In: Groves R, Eltinge J, Little R J A (eds.) *Survey Nonresponse*. Wiley, New York

Brogan D J 1998 Pitfalls of using standard statistical software packages for sample survey data. In: Armitage P, Cotton T (eds.) *Encyclopedia of Biostatistics*. Wiley, New York, Vol. 5, pp. 4167–74

Carlson B L 1998 Software for statistical analysis of sample survey data. In: Armitage P, Colton T (eds.) *Encyclopedia of Biostatistics*. Wiley, New York

Cochran W G 1977 *Sampling Techniques*, 3rd edn. Wiley, New York

Cohen S B 1997 An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *The American Statistican* **51**: 285–92

Davis J A, Smith T W 1992 *The NORC General Social Survey: A User's Guide*. Sage, Newbury Park, CA

Dempster A P, Raghunathan T E 1987 Using a covariate for small area estimation: A common sense Bayesian approach. In: Platek R, Rao J N K, Sarndal C E, Singh M P (eds.) *Small Area Statistics: An International Symposium*. Wiley, New York, pp. 77–90

DuMouchel W H, Duncan G J 1983 Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association* **78**: 535–43

Fay R E, Herriot R A 1979 Estimates of income for small places: An application of James–Stein procedures to census data. *Journal of the American Statistical Association* **74**: 269–77

Fienberg S E 1980 The measurement of crime victimization: Prospects for panel analysis of a panel survey. *The Statistician* **29**: 313–50

Fienberg S E, Tanur J M 1983 Large-scale social surveys: Perspectives, problems, and prospects. *Behavioral Science* **28**: 135–53

Gelman A, Carlin J B 2001 Poststratification and weighting adjustments. In: Groves R, Eltinge J, Little R J A (eds.) *Survey Nonresponse*. Wiley, New York

Gelman A, Little T C 1997 Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* **23**: 127–35

Gelman A, Little T C 1998 Improving on probability weighting for household size. *Public Opinion Quarterly* **62**: 398–404

Groves R M 1989 *Survey Errors and Survey Costs*. Wiley, New York

Hidiroglou M A 1986 The construction of a self-representing stratum of large units in a survey design. *The American Statistician* **4**: 27–31

Hill M S 1992 *The Panel Study of Income Dynamics: A User's Guide*. Sage, Newbury Park, CA

Iachan R, Jones B I 1987 Classification and properties of rotation sampling designs. *Journal of Official Statistics* **3**: 169–81

Kalton G, Kasprzyk D 1986 The treatment of missing survey data. *Survey Methodology* **12**: 1–16

Kish L 1965 *Survey Sampling*. Wiley, New York

Kish L 1998 Space/time variations and rolling samples. *Journal of Official Statistics* **14**: 31–46

Korn E L, Graubard B I 1995 Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician* **49**: 291–95

Korn E L, Graubard B I 1999 *Analysis of Health Surveys*. Wiley, New York

Lemeshow S, Letenneur L, Dartigues J-F, Lafont S, Orgogozo J-M, Commenges D 1998 Illustration of analysis taking into account complex survey considerations: The association between wine consumption and dementia in the PAQUID Study. *American Journal of Epidemiology* **148**: 298–306

Lepkowski J, Bowles J 1996 Sampling error software for personal computers. *The Survey Statistician* **35**: 10–17

Little R J A 1991 Inference with survey weights. *Journal of Official Statistics*: 405–24

Little R J A, Rubin D B 1987 *Statistical Analysis with Missing Data*. Wiley, New York

Lohr S L 1999 *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA

Lohr S L, Liu J 1994 A comparison of weighted and unweighted analyses in the National Crime Victimization Survey. *Journal of Quantitative Criminology* **10**: 343–60

Mirrlees-Black C, Budd T, Patridge S, Mayhew P 1998 *The 1998 British Crime Survey, England and Wales*. Home Office Statistical Bulletin 21/98. Home Office, London

Oh H L, Scheuren F 1983 Weighting adjustment for unit nonresponse. In: Madow W G, Olkin I, Rubin D B (eds.) *Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies*. Academic Press, New York

Pfeffermann D 1993 The role of sampling weights when modeling survey data. *International Statistical Review* **61**: 317–37

Rao J N K 1996 On variance estimation with imputed survey data. *Journal of the American Statistical Association* **91**: 499–506

Rubin D B 1976 Inference and missing data. *Biometrika* **63**: 581–92

Rubin D B 1987 *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York

Schafer J L, Ezzati-Rice T M, Johnson W, Khare M, Little R J A, Rubin D B 1996 The NHANES III multiple imputation project. In: *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 28–37

Schafer J L, Khare M, Ezzati-Rice T M 1993 Multiple imputation of missing data in NHANES III. In: *Proceedings of the Bureau of the Census Annual Research Conference*. US Bureau of the Census, Suitland MD, pp. 459–510

StataCorp 1999 *Stata Statistical Software: Release 6.0*. Stata Corporation, College Stations, TX

Statistics Canada 1997 *Survey of Labour and Income Dynamics Microdata User's Guide*. Catalog no. 75M0001GPE. Statistics Canada, Ottawa

Voss D S, Gelman A, King G 1995 The polls—a review—preelection survey methodology: Details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly* **59**: 98–32

Wang S T, Yo M L, Lin Li Y 1997 Consequences of analyzing complex survey data using inappropriate analysis and software computing packages. *Public Health* **111**: 259–62

Winship C, Radbill L 1994 Sampling weights and regression analysis. *Sociological Methods and Research* **23**: 230–57

Wolter K M 1985 *Introduction to Variance Estimation*. Springer-Verlag, New York

Zaslavsky A 2000 Summary of survey analysis software. American Statistical Association Survey Research Methods Section Webpage. http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html

E. Zanutto and A. Gelman

Lashley, Karl Spencer (1890–1958)

It is probably fair to say that Karl Lashley was the most important and influential American scientist in the first half of the twentieth century in the development of modern biological (and even experimental) psychology. His work and writings had a profound influence on such fields as brain substrates of learning and memory, the study of serial behaviors, complex cognitive processes, and our understanding of basic brain organization, particularly of the cerebral cortex.

Karl Lashley was born in Davis, West Virginia on June 7, 1890. His parents were themselves successful—his father was variously a storekeeper, postmaster, banker, and mayor of Davis and his mother, energetic and intellectual with literary and artistic interests. Lashley grew up in Davis, except for a rather extraordinary trek; the family took to the gold rush in the Klondike in 1898–1899.

Lashley was an only child and acknowledged a great debt to his mother for nurturing his intellectual interests. He had learned to read well by the age of four. His mother taught in a country school until Karl was born but maintained a center for adult education in her home where women (and Karl) came for instruction in such subjects as civics, Greek poetry, and drama.

As a boy Karl read widely from the extensive family library but also displayed great interest in plant and animal life, wandering through the local fields and woods, collecting butterflies, snakes, frogs and snails, and keeping pet mice and raccoons.

Karl entered a private school at the age of four and later attended public schools in Davis, where he graduated from high school at the age of fourteen. He entered the University of West Virginia but did not qualify as a freshman because his high school was not accredited. He spent a preparatory year which he felt was wasted. He initially planned to study Latin but was instead assigned English as a major. To fill a vacancy in his schedule he enrolled in a course in Zoology taught by John Black Johnston, a neurologist and the only teacher of zoology at the university. This course and Johnston had a profound influence on Karl. He later wrote that 'within a few weeks in his class I knew that I had found my life's work.'

Lashley received his BA at West Virginia in 1910. He then entered the master's degree program at the University of Pittsburgh as a teaching fellow. During his year there in biology, he attended a course in experimental psychology by Karl Dallenbach, a teaching fellow in psychology. To quote Dallenbach: 'Lashley's laboratory was on the floor above mine. Though he had never taken a course in psychology, he was permitted, because he was a fellow, to elect my laboratory course in experimental psychology. The class was small and we worked intimately together upon the various experiments. Lashley was intensely interested and was the outstanding student in the class—as one might expect being a graduate student in an undergraduate class, but he was more than a run-of-the-mine graduate student. He showed in that course the promise that he later fulfilled.' (Beach 1961, p. 170).

The following summer Lashley worked at the Carnegie Laboratories at Cold Spring Harbor, where he conducted basic studies on a ciliate. A very distinguished zoologist at Johns Hopkins, H. S. Jennings, was so impressed with this work that he offered Lashley a fellowship. At Hopkins, Lashley worked with Jennings on paramecia and completed a Ph.D. thesis on inheritance in the asexual reproduction of Hydra (1914). Lashley took a minor in psychiatry with Adolf Meyer, a pioneering figure in biological psychiatry and another minor (in psychology) with John B. Watson. The influence of Watson on Lashley was profound.

In a letter written much later to Ernest Hilgard at Stanford, Lashley described taking a seminar with Watson in 1914. Watson called their attention in the seminar to the writings of Bechterev on conditioned reflexes, which they translated. 'In the spring I served as an unpaid assistant and we constructed apparatus and did experiments, repeating a number of their

experiments. Our whole program was then disrupted by the move to the lab in Meyer's Clinic. There were no adequate animal quarters there. Watson started work with infants as the next best material available. I tagged along for awhile but disliked the babies and found me a rat lab in another building. We accumulated a considerable amount of experimental material on the conditioned reflex that was never published. Watson sought the basis of a systemic psychology and was not greatly concerned with the reaction itself.' (Hilgard, personal communication, 1984).

The conditioned reflex thus came to form the basis of Watson's behaviorism. Lashley, on the other hand, had become interested in the physiology of the reaction and the attempt to trace conditioned reflex paths through the central nervous system. At the end of his career, Lashley was reputed to have said that anyone who knows American psychology today (Circa 1958) knows that its value derives from biology and from Watson.

During the period at Hopkins, Lashley worked with Shepherd Ivory Franz at St. Elizabeth's Hospital in Washington. Together, they developed a new approach to the study of brain mechanisms of learning and memory and published landmark papers on the effect of cortical lesions on learning in rats. From this time (1916) until 1929 Lashley systematically used the lesion method in the attempt to localize memory traces in the brain. Following Watson (and Pavlov), Lashley conceived of the brain as a massive reflex switchboard, with sequential chaining of input-output circuitries via the cerebral cortex as the basis of memory. This work culminated in his classic monograph *Brain Mechanisms of Intelligence* (1929). He was also President of the American Psychological Association that year. His presidential address, and his 1929 monograph, destroyed the switchboard reflexology theory of the brain function and learning as it existed at that time.

In complex mazes, rats were impaired in learning and memory in proportion to the degree of cerebral cortex destroyed, independent of locus. Lashley employed the terms 'mass action' and 'equipotentiality' more to describe his results than as a major theory. He did not deny localization of function in the neocortex; indeed, his work on visual discriminations strongly supported localization (see Lashley 1942). Using anatomical and physiological methods, he was perhaps the first to demonstrate the point-to-point projection from the lateral geniculate nucleus to the visual cortex. Rather, he argued that the neural substrate for higher order memorial functions, he often used the term 'intelligence', as in complex maze learning in rats, was widely distributed in the cerebral cortex. He did attempt to develop a theory of cortical field-like interference patterns as a mechanism to account for equipotentiality in perception and learning, but he found it wanting. 'He remarked that he had destroyed all theories of behavior, including his own' (Hebb 1959, p. 149). Lashley was perhaps the most formidable critical thinker of his time, successfully demolishing all major theories of brain-behavior extant at the time, from Pavlov to the Gestalt psychologists, to his own views.

Lashley's influence was such that the field concerned with brain substrates of learning and memory went into decline for a period of years. The studies by Lashley had critics. Walter Hunter, for example, noted that the more cortex destroyed, the more sensory areas are destroyed and the fewer sensory cues are available to the animal. Part of the problem is that maze learning is a very complex task; the notion that a simple localized engram or memory trace could serve to code the memory for a maze, seems unrealistic today. In fact, simple memories do appear to have engrams that are quite localized in the brain: conditioning of discrete responses like eyeblink in the cerebellum and conditioned fear in the amygdala. The brain structure most involved in spatial learning and memory is not the neocortex but rather the hippocampus, at least in rats. However, Lashley's view is still held today in the sense that some cognitive scientists argue that complex memories are widely distributed in the brain. The jury is still out on this matter, no distributed memories having been found.

Near the end of his career, looking over his lifetime of research on memory, Lashley (1950) concluded that: 'This series of experiments yielded a good bit of information about what and where the memory trace is not. It has discovered nothing directly of the real nature of the memory trace. I sometimes feel, in reviewing the evidence of the localization of the memory trace, that the necessary conclusion is that learning is just not possible. It is difficult to conceive of a mechanism that can satisfy the conditions set for it. Nevertheless, in spite of such evidence against it, learning sometimes does occur.' (pp. 477–8).

Lashley's positive contributions were extraordinary, including an insightful analysis of 'instinctive' behaviors, analysis of sexual behaviors, characterization of the patterns of thalamocortical projections, a rethinking of cortical cytoarchitectonics, etc. Lashley's first professional position was as instructor of psychology at the University of Minnesota (1917). Robert Yerkes was to become the chair of the psychology department there and arranged Lashley's hire. However, Yerkes became much involved in the organization of the National Research Council in Washington and did not move to Minnesota. Lashley left after one year, returning to Baltimore where he resumed his work with John Watson. Lashley returned to Minnesota in 1920, where he was promoted from instructor to full professor in three years. Lashley remained at Minnesota until 1926, when he moved to the University of Chicago.

In 1936 James B. Conant, then the new president of Harvard, appointed an ad hoc committee to find the best psychologist in the world. Lashley, then at the University of Chicago, was chosen and hired in 1937.

He became Director of the Yerkes Laboratory of Primate Biology in Florida but maintained his Chair at Harvard, traveling to Cambridge once each year to give his two week graduate seminar. The roster of eminent psychologists and neuroscientists who worked with Lashley is without parallel.

Lashley was an extraordinary teacher, albeit most of his teaching was individual. He was of the view that those who need to be taught can't learn and those who can learn don't need to be taught. Following is a description of his teaching at Minnesota by a former student: 'I saw Karl for the first time when I was taking a General Psychology course. An assembly lecture was given to all psychology classes once every two weeks, with different professors giving one lecture to about 200 students. Karl, tall, lean, sardonic held up a frog, made the muscles jump, grinned at us and talked informally in his rather precise voice. We learned more in 40 minutes from him than we had from the other eight professors and we all filed out of the hall wanting to know when we could take a course from this thin, splendid lecturer, with the pince-nez on a black string. We quickly learned that we couldn't. He hated teaching formally and was available for only a few graduate students.' (Beach 1961, p. 182).

Although Lashley is not widely recognized as one of the sources for the development of the current field of cognitive neuroscience, his extraordinary paper on the problem of serial order in behavior clearly anticipated this field (see Bruce 1994).

'This is the essential problem of serial order: the existence of generalized schemata of action which determine the sequence of specific acts, acts which in themselves or in their associations have no temporal valence' (Lashley 1951, p. 122).

Lashley never wrote directly about the nature of consciousness. According to his student Roger Sperry (personal communication) he was interested in the problem of consciousness but refused to write about it, considering it to be perhaps something of an epiphenomenon. He did, however, speculate about the nature of mind: 'Mind is a complex organization, held together by interaction processes and by time scales of memory, centered about the body image. It has no distinguishing features other than its organization… there is no logical or empirical reason for denying the possibility that the correlation (between mental and neural activities) may eventually show a complete identity of the two organizations' (Lashley 1958, p. 542).

Lashley received essentially all the honors available to a psychologist: President of the American Psychological Association; election to membership in the Society of Experimental Psychologists; British Psychological Association; Harvey Society; US National Academy of Sciences; Royal Society, London (Foreign Member); American Philosophical Society; American Academy of Arts and Sciences; granted the Howard Crosby Warren Medal by the Society of Experimental Psychologists for outstanding scientific contributions to psychology; the Daniel Girard Elliot Medal in Zoology by the National Academy of Sciences and the William Baly Medal in Physiology from the Royal College of Physicians. About the only honor he did not receive is the Nobel Prize, and many felt he should have done so. Sir Francis Walshe, a most distinguished neurologist, gave the following tribute: '… one of the most brilliant exponents of those problems that include psychology and physiology of the brain. He moved in this difficult territory with an experimental flair, an intellectual poise, and a competence he shared with few. His contributions to science were of a major order, and lesser men than he have been Nobel Prize men' (Beach 1961, p. 191).

Frank Beach (1961) offered a favored tribute: 'Eminent psychologist with no earned degree in psychology. Famous theorist who specialized in disproving theories, including his own. Inspiring teacher who described all teaching as useless' (Beach 1961, p. 163).

Karl Lashley died in Poitiers, France on August 7, 1958.

See also: Comparative Psychology; Experimentation in Psychology, History of; Implicit Learning and Memory: Psychological and Neural Aspects; Learning and Memory, Neural Basis of; Lesion and Behavior Approaches in Neuroscience; Neural Plasticity of Spinal Reflexes; Neural Systems: Models of Behavioral Functions; Occipital Lobe (Visual Cortex): Functional Aspects; Psychologism, History of; Representation of Serial Order, Cognitive Psychology of; Topographic Maps in the Brain; Visual System in the Brain; Watson, John Broadus (1878–1958)

Bibliography

Beach F A 1961 A biographical memoir [of Karl Lashley]. *Biographical Memoirs of the National Academy of Sciences* **35**: 162–204

Bruce D 1994 Lashley and the problem of serial order. *American Psychologist* **49**: 93–103

Hebb D D 1959 Karl Spencer Lashley, 1890–1958. *The American Journal of Psychology* **72**: 142–50

Lashley K S 1929 *Brain Mechanisms and Intelligence*. University of Chicago Press, Chicago

Lashley K S 1942 The problem of cerebral organization in vision. *Biographical Symposia* **7**: 301–22

Lashley K S 1950 In search of the engram. *Society of Experimental Biology*, *Symposium* **4**: 454–82

Lashley K S 1951 The problem of serial order in behavior. *Cerebral Mechanisms in Behavior*. John Wiley & Sons, New York, pp. 112–36

Lashley K S 1958 Cerebral organization and behavior. *Proceedings of the Association for Research in Nervous and Mental Diseases* **36**: 1–18

R. F. Thompson

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Latent Structure and Causal Variables

A 'latent variable' in a statistical model is unmeasured, although not necessarily unmeasurable. Many types of statistical models contain latent variables, including factor analytic models, item response models, some structural equation models, Rasch models, and finite mixture models. Section 1 describes latent structural equation models, which are representative examples of latent variable models. (Latent structural equation models include factor analytic models as a special case. See *Factor Analysis and Latent Structure: Overview*, and *Factor Analysis and Latent Structure, Confirmatory*.) Section 2 describes the purpose of introducing latent variables into a model. Section 3 describes methods of constructing latent variable models, and problems that arise with the introduction of latent variables. Section 4 is a brief summary, and Section 5 gives pointers to the literature.

1. Linear Structural Equation Models

There are several different methods of parameterizing structural equation models. The one described here is essentially the same as the one used in Bentler (1985). The variables in a linear structural equation model (SEM) can be divided into two sets, the 'error variables' or 'error terms,' and the substantive variables. The error terms are latent, and some of the substantive variables may be latent as well. A structural equation model consists of a set of structural equations, one for each substantive variable, and the distributions of the error terms; together these determine the joint distribution of the substantative variables. The structural equation for a substantive variable X_i is a linear equation with X_i on the left-hand side of the equation, and the direct causes of X_i plus an error term ε_i on the right-hand side of the equation. (In some representations, some substantive variables occur only on the right-hand side of structural equations and do not have a corresponding error variable.) A SEM contains information about both the joint probability distribution over the substantive variables, and about the causal relations between the substantive variables.

Figure 1 contains an example of a latent variable SEM that is a model of mathematical test scores discussed in Spirtes et al. (2001). The original data set came from Mardia et al. (1979). The test scores for 88 students in five subjects (Mechanics, Vector Algebra, Algebra, Analysis and Statistics) are the measured variables. The latent substantive variables are Algebra Skill, Vector Algebra Skill, and Real Analysis Skill. The distribution of the test scores is approximately normal. In Model M of Fig. 1, the free parameters are the linear coefficient a , b, c, and d, and the variances and means of the error terms ε_M, ε_V, ε_{Al}, ε_{An}, ε_S, δ_{Al}, δ_{An} and δ_V. (The coefficients in the structural equations for Mechanics, Algebra, and Statistics have been fixed at 1 to ensure identifiability, which is explained below.)

Structural equation models are a subset of graphical models (see *Graphical Models: Overview*). Each SEM is associated with a graph (known as a path diagram) that represents the causal structure of the model and the form of the linear equations, where there is a directed edge from X to Y ($X \rightarrow Y$) if the coefficient of X in the structural equation for Y is nonzero (i.e., X is a direct cause of Y), and there is a bi-directed edge between the error terms ε_X and ε_Y if and only if the covariance between the error terms is nonzero. In path diagrams, latent substantive variables are enclosed in ovals, and measured variables are enclosed in rectangles.

1.1 SEMs and Probability Distributions

Let θ be a set of parameter values for SEM N that contains just the parameter values for the linear coefficients that have not been fixed, and the covariances of the error terms (i.e., the means of the error terms are not included in θ). In a SEM N, θ determines a covariance matrix over the substantive variables which will be denoted by $\Sigma(\theta)$ (Bollen, 1989). A covariance matrix Σ is represented by a SEM N if there is a set of parameter values θ for the parameters in N such that $\Sigma(\theta) = \Sigma$. From the joint distribution over the variables, it is also possible to calculate conditional distributions, e.g., the conditional covariance matrix is just the partial covariance matrix if the error terms are normally distributed.

If there are no two distinct sets of parameter values θ and θ' for a model N such that $\Sigma(\theta) = \Sigma(\theta')$ then the model is *identified*. If a model is identified then there are a variety of methods for estimating the parameters, and a number of programs that perform the estimation (Bentler 1985, Jöreskog and Sörbom 1998). Also, if a model is identified, it is possible to evaluate it by testing it against a nested model using a χ^2 test with the appropriate degrees of freedom, or by assigning one of a variety of kinds of scores to the model (Bentler 1985, Jöreskog and Sörbom 1998).

In addition to implying a distribution over the substantive variables in the model, a SEM implies the total effect of each variable on each other variable (Bollen, 1989). A SEM is *acyclic* when its path diagram contains no directed path from a variable to itself. For acyclic SEMs, there is a simple rule for calculating the total effect of A on B. Suppose that the edge $X \rightarrow Y$ in the path diagram is labeled by the linear coefficient of Y in the structural equation for X (e.g., the edge from *Real Analysis Skill* to *Analysis* is labeled by b.) Then the total effect of A on B is the sum over all directed paths from A to B, of the product of the edge labels along each path (e.g., the total effect of *Algebra Skill* on *Real Analysis* is $b \times d$).

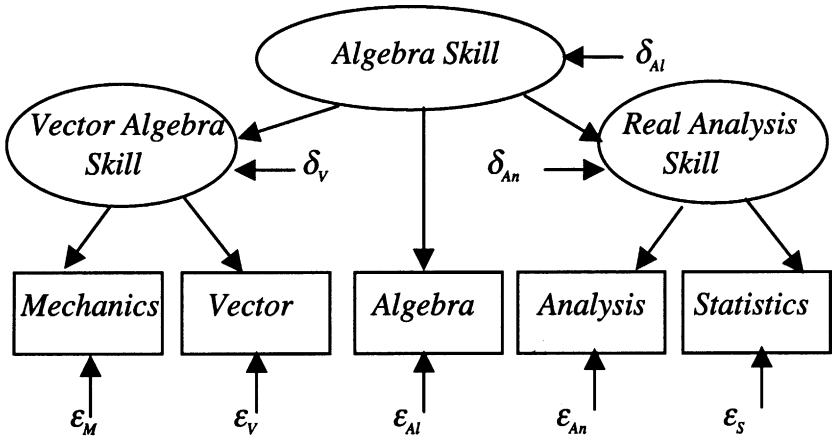

$$\textit{Mechanics} = 1 \times \textit{Vector Algebra Skill} + \varepsilon_M$$
$$\textit{Vector} = a \times \textit{Vector Algebra Skill} + \varepsilon_V$$
$$\textit{Algebra} = 1 \times \textit{Algebra Skill} + \varepsilon_{Al}$$
$$\textit{Analysis} = b \times \textit{Real Analysis Skill} + \varepsilon_{An}$$
$$\textit{Statistics} = 1 \times \textit{Real Analysis Skill} + \varepsilon_S$$
$$\textit{Vector Algebra Skill} = c \times \textit{Algebra Skill} + \delta_V$$
$$\textit{Real Analysis Skill} = d \times \textit{Algebra Skill} + \delta_{An}$$
$$\textit{Algebra Skill} = \delta_{Al}$$

Figure 1
Model M

The meaning of the total effect of A on B can be described as follows. In contrast to conditioning, a manipulated probability distribution is not a distribution in a subpopulation of an existing population, but is a distribution in a (possibly hypothetical) population formed by externally forcing a value upon a variable in the system. For example, among the subgroup of students who scored above 65 on the *Analysis* test (which was quite a high score), the mean *Statistics* score was above the mean for the whole class. On the other hand it is possible to manipulate the scores of all of the students by teaching them the answers to the questions on the *Analysis* test before they take it. In that case the mean *Statistics* score would remain the same for the whole class. So the mean *Statistics* score conditional on the *Analysis* score being above 65 would not equal the mean *Statistics* score, given that the score of the *Analysis* test was manipulated to be above 65. The total effect of A on B is just the change in B resulting from a manipulation that forces a unit change in A.

Of course, how an actual manipulation would affect other variables would depend upon how the system is manipulated. If the *Analysis* test score were changed by practice and studying to improve *Real Analysis Skill*, then one might expect that the *Statistics* test score might change as well. In that case, the manipulation has changed *Statistics*, but not through the mechanism of changing the *Analysis* score. The total effect of *Analysis* on *Statistics* is the resulting change in *Statistics* due to a kind of ideal manipulation of *Analysis*, in which the only variable in the system directly affected by the manipulation of *Analysis* is *Analysis* itself. Strotz and Wold (1960) suggested that an ideal manipulation of a variable X to a fixed value x could be represented by replacing the structural equation for X with a new equation $X = x$. The manipulated distribution is then just the distribution for the new structural equation model. Figure 2 shows the model that results from the manipulation of *Analysis*. The path diagram of the manipulated population is formed by breaking all of the edges into the manipulated variable.

Unlike a conditional distribution which is completely determined by the population joint distribution, a manipulated distribution depends upon both the population joint distribution and the true path diagram (or structural equations). In particular, two models with different path diagrams can represent the same set of joint probability distributions, but differ in

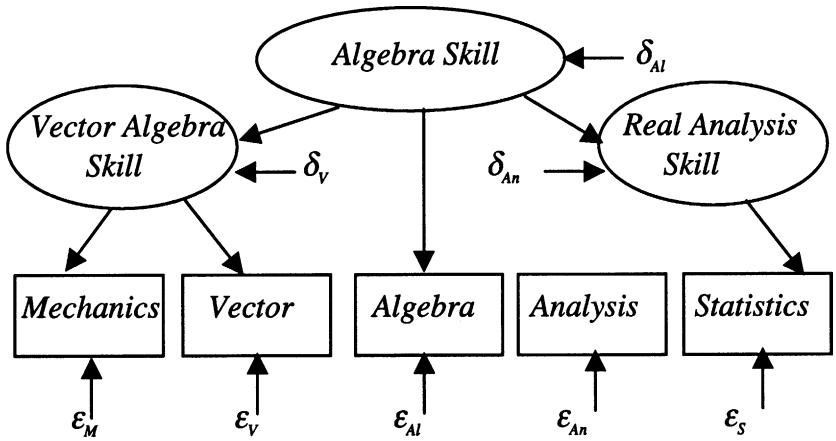

$$Mechanics = 1 \times Vector\ Algebra\ Skill + \varepsilon_M$$
$$Vector = a \times Vector\ Algebra\ Skill + \varepsilon_V$$
$$Algebra = 1 \times Algebra\ Skill + \varepsilon_{Al}$$
$$Analysis = 100$$
$$Statistics = 1 \times Real\ Analysis\ Skill + \varepsilon_S$$
$$Vector\ Algebra\ Skill = c \times Algebra\ Skill + \delta_V$$
$$Real\ Analysis\ Skill = d \times Algebra\ Skill + \delta_{An}$$
$$Algebra\ Skill = \delta_{Al}$$

Figure 2
Model M after *Analysis* is manipulated to 100

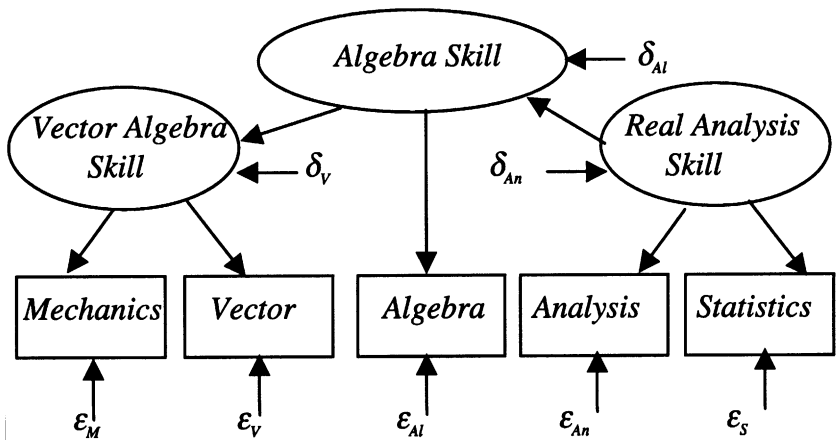

Figure 3
Model M

their predictions of the total effect of manipulating a variable. Figure 3 shows an alternative to Model M of Fig. 1, where the edge between *Real Analysis Skill* and *Algebra Skill* has been reversed. In Model M, the total effect of *Algebra Skill* on *Real Analysis* is nonzero, while in Model M′, the total effect of *Algebra Skill* on *Real Analysis* is zero. However, both M and M′ represent the same set of distributions over the substantive variables. If Model M′ were considered to be compatible with background knowledge, the fact

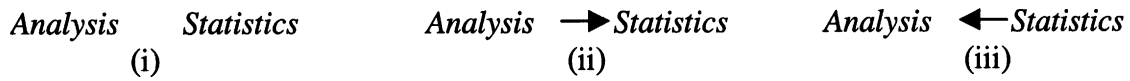

Figure 4
SEMs without latent variables

that M′ passes a statistical test or is assigned a high score is not sufficient evidence that M′ can be used for calculating total effects; it is also important to know whether there are any other plausible models compatible with background knowledge (such as M) that would also pass a statistical test or receive a high score, and predict a different total effect.

2. *Introducing Latent Variables*

There are three main reasons for introducing latent variables into a model. The first is that there may be variables of interest that either were not measured, or cannot be measured without error. The second is that a latent variable model may be necessary in order to represent correctly both the probability distribution over the measured variables, and the total effects that the measured variables have on each other. The third is that a latent variable model may be a more parsimonious representation of the distribution over the observed variables than any model without latent variables.

2.1 *Measurement Error*

Whenever there are measurement errors, the actual measurements are observed, but the quantity of interest that is being measured is a latent variable. For example, in Fig. 1, *Algebra Skill*, *Real Analysis Skill*, and *Vector Algebra Skill* are measured indirectly by test scores. The test scores are not perfect measures of the corresponding latent quantities, but are also dependent on the error terms, i.e., there is measurement error in this model. Hence *Algebra Skill*, *Real Analysis Skill*, and *Vector Algebra Skill* are latent variables. If the goal is to estimate the total effect of *Algebra Skill* on *Statistics*, and the variance of ε_{Al} is not zero, then regressing *Statistics* on the measured variable *Algebra*, instead of on the latent variable *Algebra Skill*, gives a biased estimate of the total effect. However, under the assumption that the latent variable Model M is correct, estimations of the parameters of M can be used to calculate an unbiased estimate of the effect of *Algebra Skill* on *Statistics*.

2.2 *Representing Causal Relationships Between Variables*

The second main reason for introducing latent variables into a model is to represent correctly the total effects of the variables upon each other. Suppose that Model M is the correct representation of the causal relations between the variables in the path diagram. In that case, the total effect of *Analysis* on *Statistics* is zero, the total effect of *Statistics* on *Analysis* is zero, but *Analysis* and *Statistics* are correlated. The possible SEMs (with no correlated errors) when there are no latent variable and only two measured variables *Analysis* and *Statistics*, are shown in Fig. 4. None of these nonlatent variable models correctly represents both the correlation between *Analysis* and *Statistics*, and the zero total effects of each variable on the other. So in this case, a latent variable model is needed for the correct representation.

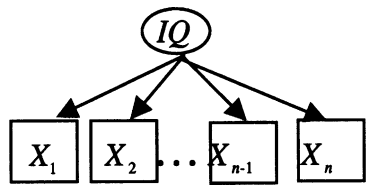

Figure 5
A parsimonious latent variable model

A bi-directed edge between the ε_{Al} and ε_S also correctly models the probability distribution and the total effects and does not explicitly introduce a latent variable that is a direct cause of both *Analysis* and *Statistics*; however, it does introduce such a latent variable implicitly. Assuming that a correlation between ε_{Al} and ε_S is produced either by ε_{Al} causing ε_S, ε_S causing ε_{Al}, or some third latent variable causing both of them, a bi-directed edge between ε_{Al} and ε_S also implies that there is some latent variable (possibly ε_{Al} or ε_S) that is a direct cause of both *Analysis* and *Statistics*.

2.3 *Parsimonious Models*

In some cases a latent variable model provides a more parsimonious (i.e., fewer parameters and higher degrees of freedom) representation of a distribution than any nonlatent variable model that represents the same distribution. Consider the model in Fig. 5, where one latent variable (e.g., *IQ*) causes many measured indicators (e.g., various intellectual tasks.) It follows from results in Spirtes et al. (2001) that the latent variable model is more parsimonious than any other SEM nonlatent variable model that represents the same distribution (assuming all of the parameters of the

nonlatent variable model are not constrained). Other things being equal, from a small sample size, the more parsimonious latent variable model could be used to provide a prediction of x_n (e.g., grades in first year in college) from $x_1,\ldots, x_{n-1}$ (e.g., scores on many different kinds of tests) that has a smaller mean squared error than any nonlatent variable model.

3. Construction of Latent Variable Models

Construction of SEMs often proceeds through a combination of the application of background knowledge with either nonautomated or automated searches through the space of alternatives. The searches look for a model or a set of models with high posterior probabilities (see *Bayesian Statistics*), or high scores, or that pass a statistical test. Even without latent variables, the problem is a difficult one because the space of models to be searched grows exponentially with the number of variables. The search space may be constrained by background knowledge, but it is still typically enormously large.

Allowing latent variables into models makes the problem of constructing good models even more difficult. This is because estimating, testing, scoring, and searching for latent variable SEMs is fundamentally more difficult than the corresponding operations for acyclic directed graph SEMs (i.e., SEMs without latent variables, cycles, or correlated errors). The space of possible acyclic directed graph SEMs is finite, but the space of latent variable models is infinite. An acyclic directed graph SEM is always identified. A SEM with latent variables may not be identified, and the only known algorithm for determining if it is identified has a running time that grows exponentially with the number of variables. It is always easy to calculate the number of degrees of freedom for an acyclic directed graph SEM, but it is often difficult to calculate the number of degrees of freedom for a SEM with latent variables (Geiger and Meek 1999).

While the estimation, testing, and scoring problems with latent variable models have been solved for special cases, and there are heuristic rules for other cases (Bollen 1989), practical general solutions are not available.

4. Conclusion

There are three main reasons for introducing latent variables into models: using latent variables can lead to a more parsimonious model that improves prediction; when there is measurement error the natural representation of the quantities of interest is as latent variables; in some cases the only correct representation of the joint distribution of the observed variables and of the total effects of variables on each other is with models that contain latent variables (or bi-directed edges which imply the existence of latent variables).

While latent variable models have some advantages over nonlatent variable models, construction of a latent variable model is often more difficult than construction of nonlatent variable models. It is generally more difficult to estimate, test, score, and search for latent variable models than for nonlatent variable models.

5. Pointers to Literature

Sewell Wright (1921) introduced a form of structural equation model called path models. Bollen (1989) provides an overview and a brief history of linear structural equation models with latent variables, and Bartholomew and Knott (1999) describe a wide variety of latent variable models. Rubin (1977) discusses manipulations, which are also discussed in Spirtes et al. (2001), Pearl (1995), and Pearl (2000). Heckerman et al. (1999) and Spirtes et al. (2001) discuss the causal interpretation of latent variable models and describe a variety of automated search techniques. Acyclic directed graph structural equation models are special cases of Bayesian networks and graphical models (see *Graphical Models: Overview*).

See also: Causal Inference and Statistical Fallacies

Bibliography

Bartholmew D, Knott M 1999 *Latent Variable Models and Factor Analysis*. Charles Griffin, London, UK

Bentler P 1985 *Theory and Implementation of EQS: A Structural Equations Program*. BMDP Statistical Software, Los Angeles, CA

Bollen K A 1989 *Structural Equations with Latent Variables*. John Wiley, New York

Geiger D, Meek C 1999 Quantifier elimination for statistical problems. *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, pp. 226–35

Heckerman D, Meek C, Cooper G 1999 A Bayesian approach to causal discovery. In: Glymour C, Cooper G (eds.) *Computation, Causation and Discovery*. MIT Press, Cambridge, MA

Jöreskog K, Sörbom D 1998 *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language* Scientific Software, Mooresville, IN

Mardia K, Kent J, Bibby J 1979 *Multivariate Analysis*. Academic Press, New York

Pearl J 1995 Causal diagrams for empirical research. *Biometrika*. **82**: 669–709

Pearl J 2000 *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK

Rubin D 1977 Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* **2**: 1–26

Spirtes P, Glymour C, Scheines R 2001 *Causation, Prediction, and Search*, 2nd edn. MIT Press, Cambridge, MA

Strotz R, Wold H 1960 Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* **28**: 417–27

Whittaker J 1990 *Graphical Models in Applied Multivariate Statistics*. Wiley, New York
Wright S 1921 Correlation and causation. *Journal of Agricultural Research* **20**: 557–85

P. Spirtes

Latin American Studies: Economics

The economics of Latin America and the Caribbean is a study of structure, performance, and policy. It is also a study in ideas, since these play a central part in the story. By convention, Latin America is defined as the 19 republics that gained their independence from Portugal or Spain together with Haiti (a former French colony). The Caribbean includes all other islands from the Bahamas southwards together with the mainland countries of Belize and the Guyanas. Although separated by history and geography, the barriers between Latin America and the Caribbean are breaking down and the region will be considered as a whole in this article.

1. Economic Structure

The countries of Latin America and the Caribbean (LAC), from the largest to the smallest, have all been shaped by their common experience of European imperialism. Starting with the revolution in Haiti in 1791, the independence movement gathered pace and by the end of the 1820s all mainland countries except Belize and the Guyanas had become separate states. Decolonization in the Caribbean was much more complicated and, indeed, has never been completed. Several countries remain overseas territories of either Holland or the United Kingdom, others are an integral part of mainland France, while the United States retains responsibility for defense and foreign affairs in Puerto Rico and the US Virgin Islands.

All the available evidence suggests that at the time the United States secured its independence, the difference in living standards between North America and the rest of the hemisphere was modest (see Coatsworth and Taylor 1998, Chap. 1). During the nineteenth century the gap, as measured by Gross Domestic Product (GDP) per head, steadily widened and by the end of the century GDP per head in Latin America was only 12.5 percent of the comparable US figure (see Thorp 1998, p. 353). This differential did not increase in the twentieth century, but this is largely due to the rapid growth of Brazil—the largest country by far in LAC—from 1900 to 1980. If Brazil is excluded from the comparison, the gap has in fact widened over the last 100 years.

The LAC region has therefore lagged behind North America (the US and Canada) despite the fact that many countries have been independent for nearly 200 years. The region is (correctly) classified as 'less developed' by international agencies and a handful of countries (e.g., Haiti) are so poor that they are categorized as 'least developed.' Within each country poverty is widespread and the most recent estimates suggest that over one-third of households are poor; this amounts to some 200 million people (see CEPAL 1999, p. 36).

While poverty is generalized, there are also pockets of wealth for two separate reasons. First, a handful of countries in the Caribbean have enjoyed rapid economic growth and now have standards of living comparable to those in the less wealthy countries of the European Union. These include some (but not all) former British or Dutch colonies, such as the Bahamas and Aruba, the French *Départements d'Outre Mer*, and the territories linked to the United States. There are no rich Latin American states, but the regions around São Paulo in Brazil, Buenos Aires in Argentina, and Monterrey in Mexico have many of the characteristics of developed countries and a standard of living comparable to southern Italy or Portugal.

Second, income distribution is highly unequal in almost all LAC countries (the main exception is Cuba where socialist economics after 1958 narrowed wage differentials and eliminated private income from capital) and it is not uncommon for the top 10 percent to receive 40 to 50 percent of household income. This social group—nearly 20 million people in Brazil, for example—is almost a state within a state and is largely isolated from the rest of society through education, housing, transport, and international ties. The consumption pattern of this decile is similar to its counterpart in developed countries, while its propensity to save is much lower than the equivalent group in Asia. Thus, wealth—not just poverty—is an important component of the LAC region at the beginning of the twenty-first century.

Ideology has always played a central role in the LAC countries. Economists and policy-makers have by tradition identified explicitly with one or other school of economic thinking since the time of independence. There have been intense debates over the economic model considered most appropriate for each country; new ideas have also spread quickly, helped by an absence of cultural and linguistic barriers. The adoption of an apparently successful new policy in one country is invariably followed by a similar experiment in other parts of the region. It has been relatively rare for intellectuals to argue in favor of the uniqueness of their country's economy (as happens frequently in the United States and the United Kingdom). Part of the reason for this is the exposure (some would say vulnerability) of each country to similar external influences.

The role of ideas is illustrated by the enthusiasm with which LAC countries have adopted since the mid-1980s a new paradigm based on market-friendly policies and export-led growth. Based to some extent

on a false reading of the Asian miracle, the new paradigm also owes a great deal to the experience of Chile, widely recognized as the star performer in the LAC region from 1985–2000. Indeed, Chile could be said to have initiated many of the reformist policies—in social security, health, education, and environmental management—that are now being adopted not only in the rest of LAC, but also in many European countries.

2. *The Inward-looking Model*

The new paradigm has replaced the inward-looking model of development based on import-substituting industrialization (ISI), which became increasingly popular after the 1929 depression. The slump in commodity prices and the collapse in capital flows, after widespread bond defaults, hit LAC countries hard in the 1930s. The region's integration into the world economy based on the export of primary products and the import of manufactures, although it had contributed to the rapid rise in living standards in a few countries (notably Argentina), was criticized for its vulnerability to external shocks and its failure to take account of protectionism in developed countries.

The ISI model was slow to reach maturity. The political elite remained close to the pressure groups associated with export-led growth and policy reform in the 1930s was piecemeal rather than wholesale. World War II gave a strong boost to the region's incipient efforts at industrialization, spurred on by the absence of competing imports. In the late 1940s, and in some cases even later, tariffs and non-tariff barriers were set explicitly to favor the protection of industry. This process gathered pace and by the 1960s Latin American countries had some of the highest tariffs in the world coupled with a system of quotas and licenses that made the import of consumer goods almost impossible in the larger countries.

The shift to a fully-fledged model of import substitution after World War II owed much to the influence of the United Nations Economic Commission for Latin America and the Caribbean (usually known by its Spanish acronym, CEPAL) and especially its second Director (Raúl Prebisch). No doubt influenced by his experience at the Central Bank of Argentina in the 1930s, Prebisch was persuaded that specialization in primary products did not reflect Latin America's long-run dynamic comparative advantage. This idea, put forward by Prebisch for the first time in 1948, was also being developed independently by Hans Singer at the United Nations in New York and for that reason it has come to be known as the Prebisch-Singer Hypothesis (PSH).

The PSH claims—on both empirical and theoretical grounds—that there is a secular tendency for the price of primary product exports to fall in relation to the price of manufactured imports. Since LAC countries by tradition export primary products and import manufactures, this amounts to a claim that Latin America's Net Barter Terms of Trade (NBTT) are set to deteriorate over the long-term. The policy implication, according to the PSH, is to shift resources out of primary product exports and into manufacturing for the home market, i.e., ISI. Over time, both Prebisch and Singer recognized the need to adopt additional measures that would promote the export of manufactured goods and this was the rationale behind the move towards regional integration in LAC countries in the 1960s and 1970s.

The empirical basis for the PSH was always stronger than the theoretical. Prebisch in particular was forced to switch between different—not always consistent—theoretical explanations for the hypothesis, including the claim that different 'institutional arrangements' made it possible for factors of production to reap the benefits of technical change in the 'north' (i.e., developed countries) while forcing the 'south' to transfer the benefits to the rest of the world through lower prices for exports.

The PSH spawned a huge debate inside and outside the region and produced a series of theoretical innovations. One of these was dependency theory, popularized by Fernando Henrique Cardoso in the 1960s nearly three decades before he became President of Brazil (see Cardoso and Faletto 1979). Dependency theory went beyond the PSH by claiming that Latin America's links with the rest of the world had resulted in a distortion of development from which the region had derived only modest benefits. In its extreme form (the theory of unequal exchange), dependency theory provided a justification for almost any policy that promoted national autonomy through inward-looking development and discriminated against imports, multinational companies, and other forms of foreign investment.

The ISI model was not without its successes. The manufacturing sector grew rapidly and increased its share of GDP. Industrial activity became increasingly sophisticated, including capital goods production in the larger countries. Direct foreign investment by multinational companies was promoted by legislation offering tariff reductions on imported inputs and tax holidays on profits. Employment opportunities in the large industrial conurbations, where manufacturing was concentrated, encouraged rural–urban migration and by the end of the 1970s almost four out of five Latin Americans were living in cities.

However, the timing of the ISI model was most unfortunate. It reached its apogee just as the postwar secular boom in world trade was unfolding. With its anti-export bias and discrimination against imports, the model led to a rapid withdrawal of the LAC region from international trade. The share of world trade of the 20 Latin American republics, over 10 percent in 1950, had fallen to under five percent in 1980. The LAC region was in large part deprived of the stimulus

from foreign trade: so important for the transformation of much of South-East Asia after 1960.

Withdrawal from world trade did not reduce the region's vulnerability to external shocks, contrary to what had been claimed by the architects of the ISI model in CEPAL. ISI proved to be very import-intensive, while the profitability of the traditional export sector was undermined by overvalued exchange rates, tariffs on inputs, and (in many cases) taxes on exports. Manufactured goods, profitable to produce behind a tariff wall, were too high cost for the world market. Thus, many of the region's governments faced an uncomfortable balance of payments constraint that often required standby agreements with the International Monetary Fund (IMF).

The ISI model was also associated with high rates of inflation. Not all LAC countries suffered from high rates of inflation and some (smaller) countries used a nominal exchange rate anchor with great success to keep inflation at or below the US rate. However, all the larger countries (with the exception of Mexico) suffered from chronic inflation, which occasionally resulted in hyperinflation. This inflationary experience led to a heated debate between monetarists, associated with the IMF, and structuralists linked to CEPAL. The monetarists, with their emphasis on the budget deficit rather than structural weaknesses such as an alleged food supply bottleneck, had the better of the intellectual argument, but they were no more successful than their rivals in eliminating inflation.

The ideological conflict between monetarists and structuralists led to an abrupt change of policy in the 1970s in the southern cone (Argentina, Chile, Uruguay) after military coups in each country. The southern cone model was a conscious attempt to reverse the inward-looking policies in each country that had brought low growth, high inflation, and successive balance of payments crises. The most extreme version of this model was adopted in Chile under General Pinochet, where it was led by a group of technocrats dubbed the 'Chicago Boys' in honor of their association with the University of Chicago's Department of Economics where Milton Friedman had kept the flame of monetarism burning during the long period of Keynesian dominance.

The southern cone model was radical and excited much attention. Tariffs were lowered sharply, state-owned enterprises were privatized, and market forces were given free reign. However, the model was undermined by the application of inconsistent policies at the macroeconomic level and the excessive speed with which financial liberalization was adopted. This led to financial crises in all three countries at the beginning of the 1980s and the (temporary) abandonment of the model.

The exhaustion of the ISI model was apparent to many as early as the late 1960s. However, the move towards syndicated lending and flexible interest rates in the international capital market, coupled with the 1973/4 and 1978/9 oil crises, gave Latin American governments and state-owned enterprises access to commercial bank loans in almost unlimited quantities. This loosened the balance of payments constraint, but at the expense of a rapid build up of external public debt. The result was the Latin American debt crisis, triggered in August 1982 by a threatened Mexican default.

3. *The New Economic Model*

It took nearly 10 years to work out a solution to the debt crisis, culminating in many cases in the replacement of commercial bank loans with Brady bonds (named after Nicholas Brady, the US Treasury Secretary at the time) offering longer maturities and lower interest rates. By that time the ISI model was completely discredited and Latin America had begun its transition to the new paradigm. Labeled variously neo-orthodoxy, neo-liberalism, or the New Economic Model (NEM), the new paradigm revolved around a series of policies that were favored by international financial institutions. For that reason, it was also known as the Washington Consensus.

3.1 *Trade Liberalization*

The first step towards the new paradigm was trade liberalization. Those—numerous—countries that were not yet members of the General Agreement on Tariffs and Trade (replaced in 1995 by the World Trade Organization, WTO) applied to join and tariff and non-tariff barriers were reduced unilaterally. For those countries doing so in the mid-1980s (e.g., Mexico), this was a bold step, as the balance of payments had once again become a major constraint on growth. The consequence was a sharp depreciation of the nominal exchange rate, leading in some cases to an increase in inflation.

Trade liberalization was eventually adopted all over the LAC region; in many cases (e.g., Argentina and Brazil) it coincided with the renewal of net capital inflows at the start of the 1990s. This eased the adjustment problem of unilateral trade liberalization, but made it more difficult to promote nontraditional exports. One result was a revival of interest in regional integration schemes. These had been launched in the region in the 1960s as part of the ISI model, but had collapsed with the debt crisis. The new schemes were less concerned with the promotion of industry and more concerned with providing a platform for reducing costs of production and thereby gaining access to the world market for a broad range of goods and services. It was therefore named 'open regionalism' by CEPAL in order to distinguish it from the 'closed regionalism' of the 1960s and 1970s.

Regional integration has been a mixed success. The efforts to revive previously moribund schemes—the Caribbean Community (CARICOM), the Central American Common Market (CACM), and the Andean Pact (now known as the Andean Community)—have been disappointing with many of the old problems surfacing again. The new schemes—the North American Free Trade Agreement (NAFTA) and the *Mercado Común del Sur* (MERCOSUR)—have been much more promising. Mexico's membership of NAFTA has given economic policy a credibility it lacked before and has contributed to a rapid increase in exports to the US and Canada. MERCOSUR, made up of Argentina, Brazil, Paraguay, and Uruguay together with Bolivia and Chile as associate members, has not only brought about a major increase in intra-regional trade since its launch in 1991, but has also stimulated direct foreign investment and the overhaul of the region's antiquated infrastructure.

3.2 Fiscal Reform and Financial Liberalization

The increase in imports made possible by trade liberalization and the renewal of capital inflows forced governments to address seriously the problem of inflation stabilization. At the center of these efforts has been the reform of fiscal policy, including an overhaul of the tax system, a revision of spending plans, and the privatization of state-owned assets. This process is far from complete in most countries, but budget deficits have been reduced and monetary policy has become more responsible. Argentina, under the *Ley de Convertibilidad* in 1991, established an extreme version of the new orthodoxy, which succeeded in bringing the rate of inflation below that in the United States.

The renewal of capital inflows to the LAC region since 1990 has been a mixed blessing. While it has allowed countries to raise their investment rates above what would have otherwise been possible, it has been subject to great volatility and generated considerable uncertainty. It has also led in many cases to an appreciation of the real effective exchange rate (i.e., the nominal rate adjusted for inflation), undermining the efforts to promote nontraditional exports. Some countries, notably Chile, have tried to reduce the inflow of short-term capital through the system of taxation, but none have resorted to capital or foreign exchange controls. It is fair to say that LAC countries are still searching for policies that will maximize the benefits of capital inflows while minimizing the costs.

The other elements of the NEM have all involved microeconomic reforms: health, education, social security, and the labor market to name but a few. There has been a learning-by-doing process with Chile in many cases as the pioneer. The privatization of the pension system in Chile has been adapted, rather than copied, in many parts of the region while the reform of the health system in Colombia has been much admired. The reform program in general has benefited from the priority given to it by international funding agencies: notably the Inter-American Development Bank and the World Bank. Inevitably, political constraints have slowed down or distorted the process of reform in some countries, but the overall picture is impressive with very few sacred cows left untouched.

4. The Performance of the LAC Economies

It will be many years—perhaps decades—before the new paradigm can be judged satisfactorily. The rate of growth of GDP per head in the 1990s was certainly faster than in the 'lost decade' of the 1980s, but it has not yet returned to the rate achieved during the golden age of the ISI model in the 1950s and 1960s. A small number of countries appear to have achieved a structural break with their past, leading to a realistic prospect of sustainable long-run growth rates of GDP per head faster than those in the United States and other developed countries. Chile and Mexico are the most important cases, but some of the smaller countries—Costa Rica, for example—may also be included in this list.

The disappointing rate of growth of the LAC region in the first decade of the NEM is in part due to the sheer weight of Brazil. With over 30 percent of the population and over 35 percent of GDP, the performance of Brazil is crucial for any regional indicator. While Brazil has succeeded since the launch of the 1994 *Plano Real* in eliminating chronic inflation, it has been only partially able to implement other elements of the NEM. As a result, exports remain less than 10 percent of GDP and the country has yet to enjoy export-led growth. There is no doubting the commitment of successive administrations in Brazil to the need for reform, but the size of the country, the complexity of decision-making, and the depth of the inherited problems have made the application of the new model particularly difficult.

The new paradigm emphasizes the need for LAC countries to promote nontraditional exports. There has been some progress in this area. Mexico has rapidly expanded its exports of manufactures and oil now accounts for only a small proportion of the total. Chile has developed new markets for agro-industrial products and non-oil mineral exports are emerging in Argentina for the first time since the collapse of the Spanish empire. Even Cuba, a reluctant reformer denied access to the US market, has been able to diversify exports with tourism replacing sugar as the major source of foreign exchange. However, traditional primary products continue to account for almost half of all exports and a much higher proportion in the smaller countries. Prices remain volatile for these products and the reforms to world trade in agricultural products under the WTO have not been able to eliminate the many obstacles that exporters face.

Diversification of products has not been matched by diversification of markets. Despite the revival of regional integration, the United States is still the main market and absorbs almost 50 percent of all LAC exports (this proportion tends to fall as one goes south so that for MERCOSUR the European Union is the principal trade partner). The United States is also the leading source of direct foreign investment in most countries and provides much of the equity flows to the region. The bulk of new bond issues are placed in New York and the commercial bank links through financial centers in the US (now including Miami) are very strong. However, far from being resented, US economic influence is generally welcomed. This is partly because of the cultural changes associated with the NEM in the LAC region, but also because international economic relations have become more rule-based with arbitration (in the WTO, for example) occasionally favoring LAC countries. Furthermore, the proposal by President Bush in 1990 for a Free Trade Area of the Americas (FTAA) was taken up enthusiastically by the rest of the continent and negotiations involving all countries except Cuba were formally launched in 1998.

The NEM has yet to make much impact on equity in the LAC region. The proportion of households classified as poor has fallen in the 1990s, but it is still no lower than before the debt crisis. Income distribution remains stubbornly unequal; indeed, inequality has increased in many countries from 1990–2000 despite the fall in inflation (widely seen as a tax on the poor). This is partly due to the widening of wage differentials as firms compete for the skilled labor needed to meet the quality thresholds demanded by foreign buyers. It is also due to downsizing by companies in the private and public sector looking for cost reductions; this has swelled the numbers seeking work in the informal sector where wages and hours of work are flexible. The drop in household income has also forced many women and children to seek work, further adding to the supply of labor and putting downward pressure on unskilled or semi-skilled wage rates.

The launch of the NEM has coincided with a rise in concern for the impact of economic growth on the environment both inside and outside the region. Some of these concerns are focused on pollution in cities, where inadequate legislation and lax regulation have combined to produce environmental hazards and health problems. The northern border of Mexico, home to most of the country's *maquiladoras*, is a good example of this problem and this was one of the main reasons for the introduction of a side agreement in NAFTA obliging Mexico to enforce its environmental laws. However, the LAC region is also home to some of the most sensitive rural environments, notably the Amazon basin, where damage can have spillover effects on the whole planet. International agencies now insist on Environmental Impact Assessments before approving loans in these areas, but the region is still far from achieving a pattern of growth that does not involve some deterioration of the environment.

None of the LAC countries, with the possible exception of a handful of small island states in the Caribbean, has yet fulfilled the potential that has been promised for so long. On the world stage, the LAC region occupies an intermediate position between developed countries and the poorest countries of Asia and Africa. Narrowing the gap with the rich countries has proved both difficult and frustrating. Neither orthodoxy nor heterodoxy has yielded the benefits claimed by their proponents. There have been some successes, but no major LAC country has been able to sustain a strong performance over a long enough period to break out of the straitjacket of underdevelopment. Now that equity and the environment are part of any definition of development, the task has become even harder since these are areas where LAC countries are seriously deficient. The sense of frustration will therefore remain and the current consensus on economic management is therefore likely to prove fragile.

That consensus owes much to the belief among Latin American elites that globalization will remain the dominant external influence on the region and that the task of policymakers is to help the private and public sectors adapt to the new reality. The economic theory underpinning globalization is generally accepted among the elites (President Chavez in Venezuela and Fidel Castro in Cuba are the main exceptions), leaving little or no room for a specifically Latin American theory of development. This is partly a reaction to the failures of inward-looking development, but is also a reflection of the influence of higher education institutions in North America and Europe where so many of the Latin American elites have been trained. Among mass organizations, however, including many trade unions, political parties, and churches, there is much greater willingness to think in terms of a specific Latin American reality that demands a different approach. Thus, the balance of power between mass and elite organizations—a critical part of any democracy—is likely to have a strong influence in the future on the extent to which LAC countries reassert the specificity of their region.

See also: Caribbean: Sociocultural Aspects; Dependency Theory; Latin American Studies: Geography; Latin American Studies: Politics; Latin American Studies: Society; South America: Sociocultural Aspects; World Systems Theory

Bibliography

Bulmer-Thomas V 1994 *The Economic History of Latin America Since Independence*. Cambridge University Press, New York

Cardoso F H, Faletto E 1979 *Dependency and Development in Latin America*. University of California Press, Berkeley, CA

CEPAL 1999 *Panorama Social de América Latina*. Naciones Unidas, Santiago, Chile
Coatsworth J, Taylor A (eds.) 1998 *Latin America and the World Economy Since 1800*. Harvard University Press, Cambridge, MA
Edwards S 1995 *Crisis and Reform in Latin America*. Oxford University Press, New York
Ffrench-Davis R, Muñoz O, Palma G 1994 The Latin American economies, 1950–90. In: Bethell L (ed.) *The Cambridge History of Latin America, Vol. VI Part 1*. Cambridge University Press, New York
Freckleton M, Lalta S (eds.) 1993 *Caribbean Economic Development: The First Generation*. Ian Randall, Kingston, Jamaica
Hirschman A O 1963 *Journeys Towards Progress: Studies of Economic Policy Making in Latin America*. Twentieth Century Fund, New York
Mitchell B 1993 *International Historical Statistics: The Americas, 1750–1988*. Macmillan, Basingstoke, UK
Stallings B, Peres W 2000 *Growth, Employment and Equity: The Impact of the Economic Reforms in Latin America and the Caribbean*. Brookings/ECLAC, Washington, DC
Thorp R 1998 *Progress, Poverty and Exclusion: An Economic History of Latin America in the 20th Century*. Inter-American Development Bank, Washington, DC
Tulchin J, Rudman A (eds.) 1991 *Economic Development and Environmental Protection in Latin America*. Westview Press, Boulder, CO

V. Bulmer-Thomas

Latin American Studies: Education

Major trends in Latin American education research reflect wider developments in academia, including its increased responsiveness to the concerns of powerful actors. The eclectic present includes echoes from once-dominant modernization and dependency perspectives on education's role in society. Substantive interests continue to encompass such matters as expansion, inequality, identity, quality, and efficiency. Despite this continuity, the past overarching political emphasis on education for social transformation has yielded ground to research that aspires to more modest improvements through education policy. In these newer approaches, while social science methods and theories are widely invoked, most work is written by and for those interested in advocacy and reform rather than in social scientific theory *per se*. Several tendencies have been evident in research on compulsory primary, noncompulsory secondary, and higher education. While these levels are usually discussed separately, their investigations have much in common.

1. Major Trends and Tendencies

Three key tendencies in research on Latin American education are internationalism, intra-national localism, and a humbled but hopeful sense of progress through the application of research to practice.

There has been a widening international exchange of scholarship among those concerned with the *commonality* of educational problems in Latin America and beyond. In the early 1970s, Ernesto Schiefelbein and his colleagues established the REDUC database (now available on the worldwide web), summarizing country-specific projects (Schiefelbein 1982). Numerous applications of research from one area of Latin America have become available to other nations, although there is doubt about how faithfully researchers read one another's work (see Briones 1993 for a critical review). The Inter-American Dialogue projects and the Program to Promote Educational Reform in Latin America (PREAL) illustrate how cosmopolitan educational research has become. So does the impact of women's movements in Latin America. United Nations conferences on population and women's status have led to evidence for gender as an independent dimension of educational stratification (Bourque and Conway 1993). The new internationalization has blurred boundaries between research concerns and methods across countries of Latin America and other areas of the world, where common questions are posed and pursued. Nor does Latin American research today serve only to illuminate problems in other lower- or middle-income countries. For example, both advocates and critics of vouchers in England and the USA read evaluations of Chilean school vouchers (see Gauri 1998). Alongside the *expansionist* tendency of internationalization, however, is a tendency to localism, at least at the compulsory education level. With the end of the Cold War and decline of socialist–capitalist debate, many commentators have noted a resurgence of intra-national tensions. In education research, this has been accompanied by the emergence of research questions that were long left on the back burner. Often shifting the focus from the national and legal to the postmodern construction of the group, this research has legitimated a new emphasis on grassroots change agents who, though formally without authority, use education to pursue their ends. One growing view (at least for the compulsory level) is that not only individuals but groups have rights, and this view has advanced research into the status of linguistic and other ethnic minorities (see Stavenhagen 1990 for examples in English).

The third research tendency holds that even the best social scientific research usually has limited utility for policy-makers. Some (Farrell 1995) argue that North American researchers offer little guidance for Latin America, given the intractability of Canadian and US problems, and the disagreement among academicians in wealthy countries about how to improve education. The new humility among researchers is exemplified by the work of Harbison and Hanushek (1992), in one of the most sustained and rigorous social science experiments with interventions to promote achievement, student flows, and access among the poor (in north-

eastern Brazil). These researchers concluded that they were unable to say definitively whether their interventions worked as intended, though their findings suggested some positive impact. Nonetheless, the tenacity of educational problems has not led researchers to abandon explorations of how schools and universities promote progress in students or society. Rather, one narrowed focus aims to improve students' learning, and as a consequence, to transform the societies in which they live. Nor does all contemporary reform research aspire only to incremental change or increased efficiency (though much does). Whether or not it is related to a neoliberal agenda, such research pays considerable attention to such matters as policies to promote private institutional growth, private funding of public institutions (including tuition), vouchers, and substantial decentralization of public-sector activity (on higher education, see Balán 1999, Levy 1986). One underlying impetus for policy-oriented research is the near consensus that educational performance is poor. While negativism is exaggerated in many comparisons, international tests do show Latin American countries trailing counterparts at similar levels of economic development.

2. *Theory and Practice in Social Science Research*

A major wave of study on Latin American education came with the modernization school that dominated social research in developing countries after World War II. Central to modernization theory was the idea that significant positive change was possible. This change could be promoted, indeed directed and accelerated, through research. Most of all, it could be accelerated through education itself. Education was often taken as the key to shaping development. Such beliefs encouraged inquiry into education in the developing world (Coleman 1965.) The main social process studied was expansion. To bring rapid and broad development, education needed to achieve greater coverage. 'More' generally meant 'better.' Primary schooling needed to achieve universal coverage and literacy; higher education needed to move from a few percent of the cohort group to 15 percent and more. Expansion at all levels was considered crucial to social mobility and decreased inequality. Another concern was education's effect on spreading modern outlooks and norms, building identity beyond the narrower social units of village and family. Education was to contribute both to the productivity of individual students and their societies' growth through increased skills and knowledge. Human capital and 'rates of return' were considered important, if crude, indicators of this growth. Politically, education was expected to promote democratic values, participation, tolerance, and national identity. A policy concern to expand schooling to the masses guided much research into the reasons why schooling was slow to take off, and how much culture and poverty impeded access. A structural–functionalist assumption in much of this work was that, properly constituted, schools and universities would naturally prepare members of society for enhanced production and creativity. However, neo-Marxists and conflict theorists emphasized the competing interests that underlay the expansion of schooling. By definition, they concluded, state actors could not play a disinterested role in providing schooling to their citizens. Dependency theory argued that the interests of the elites and governments were too often subjugated to the interests of powerful extranational forces. Some on both sides attacked extremes or stereotypes of the other side. But, over time, contributions in certain perspectives and findings from each side gained acceptance with the other.

Today, economists and sociologists continue to be concerned with the importance of schooling for national progress or national dependency. However, if the goal of education is not certification but learning, then growth needs to be accompanied by quality. One emergent agenda for researchers has been how to measure quality. Educational psychologists have developed several international applications of student achievement tests for this purpose, building on the experience of Chile, which has long included mandatory testing as a measure of school quality. Earlier optimism that 'alternatives' to the school could work to the interests of the poor gave way in the 1970s to the unenthusiastic understanding that formal education is here to stay. The question is whether the system can be reformed by the state (as in Colombia's 'New School') or perhaps devolved to the constituencies (as in Chile) in order to increase the participation of stakeholders. The 1980s have commonly been depicted as a 'lost decade' for educational as well as other socioeconomic progress. Some researchers portray declining quality as the outcome of decreased per pupil spending (as percentages of GDP) at both basic and higher education levels, lamenting such matters as larger class size, and lower teacher salaries and professionalism. In this context, social psychologists have explored the subjective meaning of education to children and families, while anthropologists have looked at the integrating function of these beliefs in fragmented societies (Ansión et al. 1997). Others put more emphasis on public policies that fail to promote professional and competitive incentives for improved performance; in higher education studies this includes the deleterious consequences for quality, equality, and efficiency of overly standardized subsidization, inadequate cost recovery, and insufficient functional and inter-institutional differentiation, autonomy, and accountability (Tyler et al. 1997, Castro and Levy 2000).

A related stream of research into reform investigates the distribution of publicly financed benefits across the levels of education. Some argue that massive public

universities direct resources away from primary schools, where the returns to education are greater, and where there is also greater participation by the poor. Debates over the means of higher education finance have borrowed fruitfully from a North American concern with the equality of opportunity. Will decentralized or devolved responsibilities for education tend to promote or to close off the opportunities for upward mobility? Will public universities become more accountable to business, elected government, and the general public (Brunner 1995)? How do efforts to decentralize or even privatize schools and universities expand 'voice' as well as 'exit' opportunities? Is there greater citizen participation and ownership in local or private institutions?

With regard to curricular reform, vocational training, and adult education, there is no possibility for the 'manpower planning' exercises seen in the early 1960s, when central government sought, and accepted, all responsibility for education. Some researchers may wish that an international market for skills and goods did not govern the educational decisions made by families and nations; but even these researchers now acknowledge the reality. Consequent questions include how foreign investment opens or closes educational possibilities, and whether technology transfer to Latin America creates in any predictable way a greater demand for science and math skills and expanded technological higher education and research. This does not mean that government planning is unconcerned with longer-term consequences. Today, how do planners and parents train children and young adults for a future that is impossible to foresee, much less to control? What types of public or private investments will prove most beneficial in the future?

3. *Quality and Modes of Production*

The balance of this article focuses on the relationship between the substantive concerns considered above and the modes through which they are explored: how well, by whom, and within what institutional structures?

The quantity of research on Latin American education has increased notably in recent decades. Most of that which reaches an international readership comes from the larger countries led by Brazil and Mexico. More US scholars contribute research on primary–secondary levels than on higher education in Latin America. Research quality has also improved and there is a substantial tier, as noted throughout this article, of solid work. At the same time, too much investigation documents problems that have already been established (such as inequality in access, treatment, or results in the educational system). Too much still relies on limited data, and too little offers fresh empirical analysis. Much is based on the work of colleagues in the same unit or school of thought. Such weaknesses in educational research are hardly confined to Latin America. Further improvement could come where leading work would simultaneously consider the basic and higher education levels, where subject matter or policy reform is overlapping (as in social stratification or pressures for evaluation and accountability).

There is an increasing tendency for consultancy, contract research, and project evaluation, financed directly or indirectly through international agencies, led by the Inter-American Bank and the World Bank, as well as UNESCO. The results produced under these arrangements contrast with literature that concentrates on social critique. There is less often an assumption that central states must finance education at every level or for every individual. The efficacy of a strong state role is considered an empirical question rather than an article of faith; there is a suspicion that states are naturally poor agents of progress. National scholars sometimes react to perceived dictamens in policy or research agendas by denying the general applicability of positivism to the needs of the region, seeing structural adjustment as a dogma aimed at weakening the state, and seeing 'North American' approaches as neo-imperialism.

Today, however, most striking is how far these 'alien' assumptions have penetrated the curriculum and the national funding agencies. Even severe critics of penetration recognize its local appeal and institutionalization, sometimes even its pertinence, no longer associating empiricism with imperialism. For example, *Imperialism and Education in Latin America*, the 1980 classic by Argentine writer Adriana Puiggrós, appeared translated and revised in 1999 as *Neoliberalism and Education in the Americas*, with a postscript not attributing the neoliberal movement to the interests of any particular nation.

Sociological perspectives continue to remain prominent in Latin American education research, but economics is ascendant, given the current concern with both finance and feasible change within the system more than with social transformation. Although there are many policy pieces by practitioners, political science *per se* remains limited (despite the fact that Latin American research seldom assumes that education is apolitical). Exceptions include the spurt of work on university student activism in the 1960s, on the state later on, or on the similarities and differences in educational policy when regimes change, as in Chile, Cuba, and Nicaragua. Historical works have treated such matters as the development of national systems and universities, but much more common are historical sketches within works primarily about the present, and history has often been marginalized in today's preoccupation with policy study. (On disciplinary mixes and other matters of research production, see Myers 1981.)

For many, the policy preoccupation also comes from the conviction that analysis can lead to im-

provement. This obviously contrasts to most dependency, neo-Marxist, and other grand theory in that it concentrates not on blaming broad contextual realities but on education factors that can be manipulated. Yet it also differs from modernization's earlier optimism in that the current 'can-do' approach concentrates on feasible change within an accepted (if lamented) reality of the prevailing social structure and political balance of power nationally and internationally. The poor will always be with us, girls will long continue disadvantaged, but progress can nonetheless be made if researchers improve financial efficiency, incentives, teacher qualifications, curriculum, textbook distribution, and the like. These policy trends substantiate this article's theme of internationalization. Educational agendas (for policy and research both) come largely from abroad. Much of Latin Americans' best work becomes like others' best work. Along with the influence of funding for educational research is the influence of advanced academic training in US graduate schools, as well as increased collaboration between Latin American and US scholars. Moreover, as in the wider worlds of both policy and scholarship, internationalization has been led by the USA at the higher education level, notwithstanding growing regional contact, and numerous accords with European-based NGOs and universities. Although there is still little empirical scholarship that directly compares Latin American education with that of other countries, increasingly Latin American cases enter broad international volumes and conferences (Kempner et al. 1998).

The policy and international emphases have meant major changes in who does what types of educational research within Latin America. It also could be interpreted as part of the general decline of Latin America's left. Although a large body of work attacks international neo-liberal research-policy trends, this work appears defensive because neo-liberal projects have achieved greater currency, and funding, for their preparation and dissemination. Criticisms notwithstanding, there is a new age of interaction and even integration with US scholarship in orientation, norms, methods, and substance. Much of the internationally recognized and contract work is done in special research centers (Levy 1996). This contrasts to work done where the greatest number of the region's professors remain: in the administrative faculties of universities, which try to combine professional training and teaching with research. In research centers, by contrast, investigators need do little teaching and are buffered from the mass student movements, labor conflicts, and political ferment common in public university faculties. Centers have usually been receptive to international ideas, money, and networks. This is true especially of freestanding private centers, but also of certain public centers or government agencies, as well as many centers inside universities. Some longstanding centers are devoted to education, whereas other work on education comes from centers more widely engaged in social analysis; respective and prominent Chilean examples have included CIDE (Center for Educational Research and Development) and FLACSO (Latin American Faculty of Social Science). Downplaying basic and theoretical research in favor of projects aimed at policy, centers often combine research and action on education. Government funding, including that from research councils, tends to be allocated in ways supportive of internationalized orientations. The combination of governmental and international funding for educational research and graduate study abroad is hardly counterbalanced by research spending in university faculties or through teachers unions. In sum, the centers illustrate many of the tendencies identified in this article about both the substance and the modes of production of educational research in Latin America.

Several patterns can be identified in the expanding studies of Latin American education. These include internationalism, intra-national localism, and a reconceptualization of the relationship of scholarship to practice. The patterns reflect wider tendencies in social science research, in both substantive and methodological focuses. Notwithstanding considerable appropriate debate about these focuses, and about research purposes and quality, a fair overall assessment is that the field offers more abundant, sophisticated, and practical findings today than in the past.

See also: Dependency Theory; Educational Innovation, Management of; Educational Institutions and Society; Educational Research for Educational Practice; Latin American Studies: Economics; Latin American Studies: Religion; Latin American Studies: Society; Modernization, Sociological Theories of

Bibliography

Ansión J, Lazarte A, Rodríguez J 1997 *Educación, la Mejor Herencia: Decisiones Educativas y Expectativas de los Padres de Famma* (Education, the Best Inheritance: Educational Decisions and Expectations of Parents). Catholic University, Lima, Peru

Balán J (ed.) 1999 *Politicas de Reforma en la Educación Superior*. UNAM-CRIM, Mexico City

Bourque S, Conway J (eds.) 1993 *The Politics of Women's Education: Perspectives from Asia, Africa, and Latin America*. University of Michigan, Ann Arbor, MI

Briones G (ed.) 1993 *Usos de la Investigación Social en Chile* (Uses of social research in Chile). FLACSO, Santiago, Chile

Brunner J J 1995 *Educación Superior en América Latina: Una Agenda Para el Ãno 2000* (*Higher Education in Latin America: An Agenda for the Year 2000*). Universidad Nacional de Colombia, Bogotá, Colombia

Castro C M, Levy D 2000 *Myth, Reality, and Reform: Higher Education Policy in Latin America*. IDB/Johns Hopkins, Baltimore

Coleman J S (ed.) 1965 *Education and Political Development*. Princeton University Press, Princeton, NJ

Farrell J 1995 Educational cooperation in the Americas: A review. In: Puryear J, Brunner J J (eds.) *Education, Equity and Economic Competitiveness in the Americas*. Organization of American States, Washington, DC
Gauri V 1998 *School Choice in Chile*. University of Pittsburgh Press, Pittsburgh, PA
Harbison R, Hanushek E 1992 *Educational Performance of the Poor: Lessons from Northeast Brazil*. Oxford University Press, New York
Izquierdo C M 1994 *La Contribución de la Educación al Cambio Social* (*The Contribution of Education to Social Change*). Universidad Ibero Americana, Mexico City
Kempner K, Mollis M, Tierney W (eds.) 1998 *Comparative Education*. Simon and Schuster, New York
Levy D 1986 *The State and Higher Education in Latin America*. University of Chicago Press, Chicago
Levy D 1996 *Building the Third Sector: Latin America's Private Research Centers and Nonprofit Development*. University of Pittsburgh Press, Pittsburgh, PA
Myers R 1981 *Connecting Worlds: A Survey of Developments in Educational Research*. International Development Research Center, Ottawa, Canada
Puiggrós 1980 *Imperialismo y Educación en América Latina*. Editorial Nueva Imagen, Buenos Aires, Argentina
Schiefelbein E 1982 *Educational Networks in Latin America*. IDRC, Ottawa, ON
Stavenhagen R 1990 *The Ethnic Question: Conflicts, Development, and Human Rights*. United Nations University Press, Tokyo
Torres C A, Puiggrós A (eds.) 1997 *Latin American Education: Comparative Perspectives*. Westview Press, Boulder, CO
Tyler L, Mahilhaes Castro M, Courard H, Levy D, Mollis M, Navarro J 1997 *Higher Education in Latin America*. Garland, New York

D. Post

Latin American Studies: Gender

Scholarship on women and gender in Latin America contributes much to global gender studies. It has long been informed by attention to social structures, world systems, and the hierarchy of economic class privilege within countries. Scholar activists have strategically moved their work into the Women in Development (WID) and Gender and Development (GAD) areas. WID and GAD work has also informed and enriched Latin American gender studies. The current challenge to Latin American gender studies comes from border analysis: both borders constructed around nation-states and metaphorical borders of cultural and linguistic lines. As such, borders challenge the very scope of 'area' in Latin America. Does the Latin American border end at the northern Mexican border, the old Mexican border before the 1848 Treaty of Guadalupe in which Mexico lost half its land to the USA, or as far as the Mexican diaspora to the north?

This essay incorporates studies from the traditional Latin American region: South America (including the Southern Cone of Argentina, Chile, and Uruguay and the largely Portuguese-speaking Brazil; Central America and Mexico; and the south-to-north diasporas in the Americas). The area is largely a Spanish-speaking one, which overlays indigenous populations who undergo assimilation into the dominant language regimes. The Caribbean English-, Creole-, Dutch- and French-speaking populations fall outside the geographic space addressed in this essay.

1. *Gender: Grounded in Area Studies Approaches*

The study of women and gender in Latin America is informed by the history of the region, its nation-states, especially their centralized (often Spanish) colonial experience and incorporation into the capitalist world system. Gender scholars are firmly grounded in analyses of countries with firmly defined class systems and economic 'dependency' on the global economy (Safa 1976, Latin American Perspectives 1977). Compared to other world regions, the study of women and gender has been more fully accepted and legitimized in Latin American area studies, thus transforming regional studies as a whole (Staudt and Weaver 1997, p. 84). Why has this happened? Gender scholars acquired legitimacy from their firm grounding in the social structural, class, and dependency/global perspectives. Besides that, though, leadership of the Latin American Studies Association has been in women's hands not just once, but several times, including the hands of strong women/gender scholars such as Carmen Diana Deere, Helen Safa, and Jane Jaquette.

2. *Global and Social Structural Perspectives*

Beginning with the global and connecting it with the national and local, Latin Americanists have analyzed public policy agendas imposed by foreign investment, dependency on outside capital, and specific state downsizing strategies from international banks in the name of 'structural adjustment' (Léon 1982, Deere and Léon 1987, Benería and Feldmann 1992, Bose and Acosta-Belén 1995). Gender scholars have influenced WID and GAD with respect to the need to contextualize analysis in the global economic context.

This structural approach, in both Latin American studies and women/gender scholarship, created an early scholarly climate of attention to the social construction of male and female, the power relations between men and women, and the complex intersection of economic class. Thus, we might say 'gender analysis' was pioneered in Latin America, even before this once uncommon term became the millennial codeword for what was once known as women's studies. Latin Americanists analyzed the social construction of men and women, even though they, too, titled their studies with the word 'women.' With the strength of social constructionism, an essentialist or biologically determinist approach never took hold among Latin Americanists. The analytic breakthroughs and predominance of social constructionism

preceded those of other world regions, including the USA (often regarded as a region unto itself).

Despite an overwhelming scholarly approach to the social construction of male and female, popular widespread perceptions until recently projected a deep and essential difference between men and women. Such internal cultural constructions are at odds with scholarly interpretations and worthy of study in themselves. These male–female differences were viewed as both temperamental and behavioral, and the expectation of difference deeply permeated the economic, political, and religious institutions imposed upon people in their everyday lives. The counterpart to the (now contested) *machismo*, or ideology of virulent men, is *marianismo*, a parallel ideology that idealizes women in terms of saintly mothers who sacrifice themselves for family. This Madonna image offers a dichotomous polar-opposite outcast female ('whore'), unworthy of protection/reverence. Perceptions, inaccurate or blurred as they may be, are used to justify interpersonal violence against women in homes and on streets. The *marianismo* writings (Stephens 1973) gave rise to insightful analyses of women in politics, from formal governance to everyday politics that crossed the public–private borders which historically defined female as apolitical. *Marianismo* also aids in understanding interpersonal relations within households.

Supermadre is the title of a pathbreaking book that draws on interviews with Peruvian and Chilean women in government (Chaney 1979). These 'super mothers' worked in institutions with deeply differentiated expectations for female and male civil servants and politicians. Idealized images of women as uncorrupted, honest, caretakers of the household allowance and the public purse led to some interesting female niches in government: accounting, personnel, and budgets. Deep prejudice persisted even into these *marianismo*-legitimized roles, as women described their work with the saying: whereas a man may be made of silver, a woman must be made of gold (Chaney 1979, p. 110).

3. *Political Structural Perspectives*

As Latin American countries moved toward revolution or military regimes, so also would analyses move to women's actions therein. Under military dictatorships in many Central American and Southern Cone countries, women took the risk of moving politically in the streets with the motherhood mantle as (semi-) protective cover. Mothers of disappeared children and husbands circled central city squares with pictures of their loved ones hanging around their necks. *Las Madres de la Plaza de Mayo*, the mothers in Buenos Aires, Argentina, symbolized to the country, region, and world, their grief, anguish, and willingness to seek accountability on life and death from dictatorships.

'[T]he concept of politics expanded to include the daily universe that had been invaded by the dictatorship. The dividing line between the public and the private became increasingly hazy as the repressive policies of the regime affected domestic unity and as its economic policies pushed women into the workforce' (Schmukler et al. 1998, p. 628).

In Argentina, Chile, and elsewhere, scholars analyzed 'social motherhood' and the ways it extended women's activism, participatory process, and public space itself. But women also assumed militant, revolutionary actions in civil wars throughout the hemisphere and in revolutions that triumphed. Many scholars have analyzed women's involvement in the Cuban and Nicaraguan revolutions especially, as well as the consequences of revolution for women and gender equality. The social imposition of motherhood on women, coupled with economic desperation for the majority of the population, usually casts shadows on women's ability to gain more power in and from formal politics and policies.

Under the Sandinista regime in Nicaragua, a revolution in which women participated in critical-mass numbers, the social differentiation of male and female continued to diminish women's ability to gain from real gender equality. Economic crisis also challenged the regime's ability to respond to people's wage and employment demands.

Women's everyday responsibilities for children, plus their significant numbers as heads of households, lead to social activism to push for 'practical' versus 'strategic' gender interests (Molyneux 1985). If and when women gained successes, their practical needs were met, but the social and legal differentiators that institutionalized male privilege remained in place. Molyneux's language would soon be popularized and spread globally with the prolific WID and GAD scholarly-activist writings.

Cuban women also confronted the deep hierarchy of gender privilege and hierarchy after the 1959 revolution. Women never gained hold in formal power exercised from the top; however, more than half of adult women belonged to the *Federacíon de Mujeres Cubanas* by 1970, and official policy legitimized 'strategic' gender claims. Cuba's Family Code of 1975, its pioneering feature on interpersonal relations that became famous worldwide, called for equality within marriage, establishing equal responsibility between husbands and wives for household work and childcare. Although the policy is not enforced (as governments generally retreat from policies that affect 'private' space), judges incorporate the Family Code language in civil marriage ceremonies (Hernández and Esperanza 1998, pp. 623–4).

4. *Toward Democracies?*

Despite attention to revolutions, 'transition' was the norm, with the common claim of transition to democracy. By democracy, claimants meant civilian rule, with people's participation in elections and civil

organizations, known as nongovernmental organizations (NGOs). Would women be part of this transition or its eventual consolidation, and might their participation create a new way of doing politics (*una nueva manera de hacer política*) or more gender-equal policy outcomes? These are the questions that inform edited collections on Latin America (Jaquette 1994, Jelin 1987) and comparative or global collections on movements and political institutions with chapters on Latin American countries (Nelson and Chowdhury 1994, Basu 1995). Statistical profiles paint a grim picture of gender equality in the region (Valdes and Gomariz 1993, UNDP 1990–2000), but women fare better in education and health indicators in Latin America than their counterparts in absolute and gender relative terms in Asia and Africa.

While women are gaining ground in the formal political process, their engagement with the state offers long-term challenges for achieving gender equality. One collection, with a section on *invadiendo las instituciones políticas* (invading political institutions) makes it clear that women must address the details of institutional mechanisms if they are to gain political space and use it for gender-equal outcomes (Tarrés 1998). Among institutions are the multiple political parties themselves, their policy platforms, and the electoral lists that privilege names at the top of lists in the common Latin American Proportional Representation electoral system. 'Women, Vote for Yourselves!' was a Peruvian feminist campaign, but the United Left listed women at the bottom (Vargas 1986). Yet women in Argentina have gained nearly a third of legislative seats (a level only reached in Western Europe) since early 1990s laws that required party quotas for women (Jones 1996). In Mexico, women's political work in multiple parties and civil organizations has begun to produce gender-equality gains (Rodríguez 1998).

Women's political advocacy work inside government institutions would also get attention. In Brazil, shortlived but influential women's commissions would engage women's NGOs and influence government budgets (Alvarez 1997). Their niche in the Chilean state did not go so far (Waylen 1997, Goetz 1998). Again, the Latin American experience would consolidate and inform knowledge about women's bureaus and gender in bureaucracy worldwide, predating the Australian-labeled 'femocrat' writing.

Once-autonomous radical feminist organizations struggle over the compromises they make with liberal feminist proposals that engage the still male-dominant legislatures (Staudt 1998, pp. 19–31). In Mexico, new policies to reduce interpersonal violence and sexual assault against women reflect these compromises (Lamas et al. 1995); the lingering shroud of death hovers over large Mexican cities like Juárez with its 200-women/girl murder record since 1993 (Benitez et al. 1999). Some women are reluctant to compromise, as with selected mothers' groups and their moral stance.

5. *Everyday Lives*

Yet does increased female space in political institutions, with occupants from economically privileged backgrounds on the whole, have consequence for working class women and those earning little or nothing in informal economies? And what about the women working the *doble jornada*, or double day of unpaid work inside the home and paid work outside the home? Although globalization and structural adjustment may have heightened women's work burdens (Moser 1997), even women less enveloped in the national political structure and global capitalism experience subordination in the indigenous gender hierarchy (Bourque and Warren 1981). Gender hierarchy is complex, with its double decked 'modern' subordination overlaying internal subordination. Cultural constructions of gender difference offer rationales for hierarchical continuity.

In the real economic options of the free-trade regimes of late-twentieth century Latin America, the majority of women work in export-processing factories (Fernández-Kelley 1983), home-based contracting (Benería and Roldan 1987), domestic service (Chaney and García Castro 1989), and informal labor (Berger and Buvinic 1989, Rakowski 1994, Staudt 1998a). Women's struggle to maintain economic livelihoods is constrained in a savage capitalism that provides little social support from political regimes, themselves constrained in the global economic contexts.

Women's efforts to reform or transform interpersonal inequality and violence have met only modest successes, as noted above. From multiple-country anti-violence measures to the Cuban Family Law, the cultural construction of gender difference, notably unequal difference, helps to sustain female subordination.

The borders of women's realities are moving northward, challenging the traditional northern border of Latin America: Mexico. These borders move just as the supposed 'free trade' sheds commercial borders.

6. *Borders: Moving North*

Border studies have been in place for a quarter century, but confined to the narrow zone of space around national territorial lines, particularly the USA and Mexico. Borders as metaphors for cultural lines and linguistic code switching (mixing two or more languages, especially Spanish and English) emerged in feminist literary work (Anzaldúa 1987, García 1992). Literary work has always expanded the intellectual borders of political economy to include the cultural and interpersonal relations of everyday life.

Economic crises in Latin America have prompted many to seek employment *en el norte*, to the north in the USA. With the US Bracero Program and the

amnesty provisions of the 1986 immigration law reforms, people of Mexico have become the largest category of 'foreign' residents and naturalized citizens. This Latin diaspora into the USA is redefining the boundaries of area studies in the western hemisphere. Informal economies, so common in Mexico and countries to the south, are alive and well on the US side of the border (Staudt 1998a). Perhaps informal economies flourish elsewhere, though the scholarly questions of Latin Americanists rarely emerged in US studies.

From international and comparative studies, scholars have coined terms like transnational, globalization, deterritorial, and debordering to understand people's movements that carry cultural, linguistic, and identity baggage in that migration (Staudt and Spener 1998). Through migration, both men and women bring experiences and expectations that are not gender neutral (Pessar and Grasmuck 1991). Latin Americanists should confront and engage the challenge to move from national-centered analyses (albeit in global contexts) to transnational attention to border spaces, regionalism, migration, and movements.

7. Conclusions

Latin American gender studies, which inform the study of women/gender worldwide, draw on the structural approach and particulars of this area. Gender is situated in the relations between women and men of different classes in the global capitalist economy. Yet gender studies necessarily draw on internal cultural constructions of everyday male–female relations that range from reverence to brutal violence, reflecting the *marianismo–machismo* constructs. Women's activism brings these agendas into political struggles toward policy change and democratic transition.

See also: Borders, Anthropology of; Catholicism and Gender; Dependency Theory; Land Rights and Gender; Latin American Studies: Politics; Latin American Studies: Religion; Latin American Studies: Society; Social Movements and Gender

Bibliography

Alvarez S 1997 Contradictions of a 'women's space' in a male-dominant state: The political role of the commissions on the status of women in postauthoritarian Brazil. In: Staudt K (ed.) *Women, International Development and Politics: The Bureaucratic Mire*, 2nd edn. Temple University Press, Philadelphia, pp. 59–101

Anzaldua G 1987 *Borderlands/La Frontera: The New Mestiza*. Spinsters/Aunt Lute Press, San Francisco, CA

Basu A 1995 *The Challenge of Local Feminisms*. Westview Press, Boulder, CO

Beneria L, Feldman S 1992 *Unequal Burden: Economic Crises, Persistent Poverty, and Women's Work*. Westview Press, Boulder, CO

Beneria L, Roldan M 1987 *The Crossroads of Class and Gender: Industrial Homework, Subcontracting, and Household Dynamics in Mexico City*. University of Chicago Press, Chicago

Benitez R, Candia A, Cabrera P, de la Moa G, Martinez J, Velázque I, Ortiz R 1999 *El silencio que la voz de todas quiebra: Mujeres y Victimas de Ciudad Juarez*. Azar, Chihuahua, Mexico

Berger M, Buvinic M (eds.) 1989 *Women's Ventures: Assistance to the Informal Sector in Latin America*. Kumarian Press, West Hartford, CT

Bose C E, Acosta-Belen E (eds.) 1995 *Women in the Latin American Development Process*. Temple University Press, Philadelphia

Bourque S, Warren K 1981 *Women of the Andes: Patriarchy and Social Change in Two Peruvian Towns*. University of Michigan Press, Ann Arbor, MI

Chaney E 1979 *Supermadre: Women in Politics in Latin America*. University of Texas Press, Austin, TX

Chaney E, Garcia Castro M (eds.) 1989 *Muchachas No More: Household Workers in Latin America and the Caribbean*. Temple University Press, Philadelphia

Deere C D, Leon M (eds.) 1987 *Rural Women and State Policy: Feminist Perspectives on Latin American Agricultural Development*. Westview Press, Boulder, CO

Fernández-Kelly M P 1983 *For We are Sold, I and My People: Women and Industry in Mexico's Frontier*. State University of New York Press, Albany, NY

García C 1992 *Dreaming in Cuban*. Ballantine, New York

Goetz A M 1998 Mainstreaming gender equity to national development planning. In: Miller C, Razari S (eds.) *Missionaries and Mandarins: Feminist Engagement With Development Institutions*. Intermediate Technology Publications, London

Hernández T, Esperanza B 1998 Women in contemporary Cuba. In: Stromquist N (ed.) *Women in the Third World*. Garland, New York/London, pp. 618–25

Jaquette J S 1994 *The Women's Movement in Latin America: Participation and Democracy*. Westview Press, Boulder, CO

Jelin E 1987 *Ciudadanía e identidad: Las mujeres en los movimientos sociales latino-maericanos*. United Nations Research Institute for Social Development, Geneva

Jones M 1996 Increasing women's representation via gender quotas: The Argentine *Ley de Cupos*. *Women and Politics* **16**(4): 75–98

Lamas M, Martinez A, Tarres M L, Tunon E 1995 Building bridges: The growth of popular feminism in Mexico. In: Basu A (ed.) *The Challenge of Local Feminisms*. Westview Press, Boulder, CO, pp. 302–23

Latin American Perspectives 1997 Women and class struggle. Winter and Spring, IV, Nos 1 and 2

León M (ed.) 1982 *Debate sobre la Mujer in América Latina y el Caribe*. ACEP, Bogotá, Colombia

Molyneux M 1985 Mobilization without emancipation? *Feminist Studies* **11**: 227–54

Moser C 1997 *Household Responses to Poverty and Vulnerability*, vol. I. World Bank Urban Management Programme, Washington, DC

Nelson B, Chowdhury N 1994 *Women and Politics Worldwide*. Yale University Press, New Haven, CT

Pessar P R, Grasmuck S 1991 *Between Two Islands: Dominican*

International Migration. University of California Press, Berkeley, CA
Rakowski C (ed.) 1994 *Contrapunto: The Informal Sector Debate in Latin America*. State University of New York Press, Albany, NY
Rodríguez V E (ed.) 1998 *Women's Participation in Mexican Political Life*. Westview Press, Boulder, CO
Safa H 1976 *Sex and Class in Latin America*. Praeger, New York
Schmukler B, Valenzuela M E, Garcia S M, Sapriza G, di Marco G 1998 The women's movements in the Southern Cone and Brazil. In: Stromquist N (ed.) *Women in the Third World*. Garland, New York/London, pp. 626–36
Staudt K 1998a *Free Trade? Informal Economies at the US–Mexico Border*. Temple University Press, Philadelphia
Staudt K 1998b *Policy, Politics and Gender: Women Gaining Ground*. Kumarian Press, West Hartford, CT
Staudt K, Spener D 1998 The view from the frontier: Theoretical perspectives undisciplined. In: Spener D, Staudt K (eds.) *The US–Mexico Border: Contesting Identities, Transcending Divisions*. Lynne Rienner Publishers, Boulder, CO
Staudt K, Weaver W 1997 *Political Science and Feminisms: Integration or Transformation*? Twayne/Macmillan, New York
Stephens E 1973 Marianismo: The other face of Machismo in Latin America. In: Pescatello A (ed.) *Female and Male in Latin America*. University of Pittsburgh Press, Pittsburgh, PA
Tarrés M L (ed.) 1998 *Género y Cultura en América Latina*. El Colegio de Mexico,
United Nations Development Programme (UNDP) 1990–2000 (annual editions). *Human Development Report*. Oxford University Press, New York
Valdez T, Gomariz E 1993 *Mujeres Latinoamericanos en Cifras*. Instituto de la Mujer, Madrid, Spain, FLASCO, Santiago, Chile
Vargas V 1986 Women, vote for yourselves! Reflections on a feminist election campaign. *The Latin American Women's Movement: Reflections and Actions*. Isis International, Santiago, Chile
Waylen G 1997 Women's movements, the state and democratization in Chile: The establishment of SERNAM. In: Goetz A M (ed.) *Getting Institutions Right for Women in Development*. Zed, London

K. A. Staudt

Latin American Studies: Geography

Geographers have made major contributions to the study of urban and regional development, environmental resource management, and cultural patterning in Latin America. Geographic research on Latin America during the 1970–2000 period was dynamic and innovative, but often marked by clear continuities with earlier research and indeed has shown a renewed interest in the collection of data (including the reading and interpretation of texts) from earlier periods. This is consistent with geography's nature as a discipline that combines methodologies from the sciences, social sciences, engineering, and the humanities, and that relies on both historical and contemporary data for its analyses, synthesis, and relevance to contemporary problems.

Before 1970, geography had established several strong traditions of research in Latin America, often focused on productive and charismatic scholars. Carl Sauer at Berkeley had trained numerous students in the 'Berkeley School,' emphasizing field work, the study of the cultural landscape, attention to long-term processes and native Americans, and a critique of modernization in cultural and ecological terms. Karl Troll in Germany had focused on mountain geoecology of the Andes, Preston James at Syracuse (among others) had developed regional geography, and Raymond Crist in Florida had focused on migration issues. There were many national schools of geography in Latin America, focused around such figures as Javier Pulgar Vidal in Peru, and Horacio De Frieri and Francisco de Aparacio in Argentina, for example.

The remainder of this article will consider more recent developments in regional geography and planning, political geography and ethnogeography, other varieties of human geography including human–environment relations, physical geography, and mapping. It will conclude with a discussion of recent institutionalization of geography within Latin America.

1. *Regional Geography, Planning, and Development*

By 1970, regional geography and planning were well represented in national geographic institutes in Latin America (often associated with the military), and were also stressed in the work of Latin Americanist regional geographers in the United States, Canada, the United Kingdom, and France. Much of this traditional regional geography was qualitative and oriented toward the needs of geographic education and journalism, as well as the needs of national administration and foreign policy.

The 'quantitative revolution' in Anglo–American geography between 1965 and 1978 absorbed theories and methods from other social sciences, and resulted in new approaches to the comparative study of migration, regional inequalities, innovation diffusion, urbanization, and regional development in Latin America. This emphasis resulted in work exploring the spatial patterns and processes of economic modernization. Large-scale projects such as the Trans-Amazon highway and the Ciudad Guyana growth pole created a sense of the importance of regional planning involving geographic analysis. A key Latin American center for development research was Santiago, Chile, where the United Nations had a think tank at

CEPAL/ILPES. Scholars included John Friedmann, who had studied geography as well as planning at the University of Chicago, and who published important works on development planning in Venezuela and Chile. Peter Odell (at the London School of Economics), Walter Stöhr, and Poul Ove Pedersen were other significant geographers working and publishing in this context. Much of the funding for geographic research on these themes came from the United Nations, Ford and Rockefeller Foundations, and United States programs such as the Alliance for Progress.

The 1973 Pinochet coup acted to disperse the Santiago group, while the 1973 Yom Kippur war in the Middle East led to the energy crisis and a period of economic problems in Latin America. Together, these events acted to slow and, at least temporarily, demoralize the integrated geographic study of development as modernization in Latin America. Key intellectual figures in Anglo-American Latin Americanist development geography either moved to the political left (Friedmann), to other regions of the world, or to other areas of geography, including urban studies. However Odell supervised British Latin Americanist dissertations by Alan Gilbert, David Slater, and Allen Lavell, all of whom were to continue to work on development issues. David Preston, Arthur Morris, Peter Ward, and Ray Bromley and Rosemary Bromley continued the British tradition of development geography into the 1980s and 1990s, while Christoph Stadel, Bromley and Ward moved to the United States, helping to support interest in the geography of development in that country. Ohio geographer, Lawrence Brown, and his students, among others, also supported the study of contemporary development from bases in the United States. All of these geographers have tended to emphasize the role of social factors and globalization in conditioning such spatial processes as migration, transportation development, and innovation diffusion.

Although over time the 'big project' focus of Latin American development came to be discredited intellectually and economically, somewhat less affordable, major government agencies such as ONERN in Peru and SIP in Mexico continued to dedicate funds to the production of planning documents, maps, and regional surveys for a variety of purposes. These provided employment for cartographers, remote sensers, and a variety of consultants, many of them geographers. French geographers such as Jean-Paul Deler not only helped support international research and development projects for mapping, resource analysis, and urban and regional planning, but also produced important original research on a variety of topics, including all branches of physical geography, agricultural and resource geography, and distinctive studies of the 'management of space' in various countries, regions, and time periods. Generous French support for geographic training (with funding from ORSTOM, now IRD—Institut de Recherche pour le Développement) has led to a growing base of French-trained geographers in many Latin American countries.

2. *Marxism, Political Ecology, and Ethnogeography*

Marxist and structuralist approaches to understanding the geography of Latin America built on both Latin American and extraregional sources. André Gundar Frank's *Development of Underdevelopment* was an interdisciplinary stimulus to this work, but the Brazilian geographer Milton Santos' (1975) work on the two circuits of the urban economy (published in French in 1975 and later translated into English) also had a major impact. John Friedmann and David Slater contributed to critical and radical geography of Latin America during the later phases of their careers. In much of Latin America, as well as in some circles in Europe and the United States, many geographers came to accept at least some of the tenets of dependency approaches as helping to explain the persistence of spatial inequality within and between Latin American regions; and some also saw political economy as a useful tool for explaining the environmental impacts of development. Susanna Hecht and Alexander Cockburn's book *Fate of the Forest* (Hecht and Cockburn 1990) provided a long-term examination of the role of political and economic factors in Amazonian deforestation. Radical geography continued into the 1990s with the diffusion of the work of the Colombian scholar Alberto Escobar, who was a frequent attendee of geography meetings in the United States. Escobar's work provided a radical critique of the notion of development itself, including its 'sustainable' variant, and argued for a 'postdevelopment' attention to local cultural and political processes of autonomy in a decentralized 'socioepistemological space.' His work has influenced such recent scholars as Juanita Sundberg, who has developed the critical concept of an 'NGO landscape' through study of the Petén in Guatemala.

The emergence of nongovernmental organizations in the 1980s attracted the attention of geographers who analyzed the rise of 'social capital' and civil society in the creation of a sustainable context for local livelihoods in particular places and regions. The Chicago-trained geographer Sheldon Annis was a pioneer in this work, helping to found the seminal journal *Grassroots Development* published by the Inter-American Foundation. More recently, Anthony Bebbington has conducted numerous studies of the new, more diverse social fabric of development in Latin America, and has encouraged development organizations to be more flexible in dealing with a greater variety of local actors at different spatial and social scales.

Movements for indigenous rights gained international attention in the 1990s. Many of the indigenous struggles revolved around claims to land and even territoriality. The Berkeley geographer Bernard Nietschmann (1973), working in Nicaragua, was an early advocate of the salience of indigenous claims for enabling a culturally and environmentally sustainable development. His earlier work was highly controversial, however, as it was seen by many social scientists as undermining the legitimacy of the Sandinista regime. The availability of GPS technology allowed for the development of innovative participatory mapping projects, especially in Central America, led by geographers such as Peter Herlihy; these have had a major impact on national policy affecting indigenous rights, although some have begun to critique some of the limitations of the technology. Geographers have also continued their earlier intellectual tradition of studying and mapping spatial patterns of indigenous groups over time, using archival and archaeological evidence, as well as documenting indigenous environmental practices, past and present. For example, William Denevan (1992) studied the distribution of native population prior to the Spanish conquest, suggesting high population counts with correspondingly high impacts on the environment. Geographers such as Karl Offen have worked on the more theoretical issues of the generation of ethnic identity and its relation to space, place and territory; their work is strikingly relevant to contemporary ethnic politics.

Geographers have also begun to redress the neglect of the role of women in Latin America, as resource managers, migrants, and workers in the space-economy of towns and cities, as well as beginning to examine their role in the historical geography of the region. For example, Diane Rocheleau has developed the method of 'counter-mapping' using women's perspectives on resources in the Dominican Republic.

3. *Varieties of Human Geography*

Since 1960, cities have experienced explosive growth, and increasingly geographers have been attracted to study them. Ernest Griffen and Larry Ford developed an influential schematic and qualitative model of Latin American urban structure which has subsequently been amended and critiqued as perhaps underestimating the diversity of influences on urban form. Urban geographers have been influenced by a variety of Marxist and other social science approaches in developing tools for dealing with such urban issues as the informal economy, migration, and the impacts of globalization. Influential scholars have included the Spanish-born and French-trained sociologist Manuel Castells, who authored numerous works on society and economy in urban places. The British-trained geographer, Peter Ward (1998), has published extensively on housing and urban geography of Mexico. Lawrence Brown and his student Victoria Lawson have provided innovative analyses of urban migration and economy. The distinction between urban and rural has increasingly become blurred as geographers see similar sociospatial development processes at work in both realms.

Geographers have worked on the political geography of Latin America, which, although not as dynamic as that of Eastern Europe, has exhibited considerable tensions. Carlos Reboratti (1987) has analyzed the geopolitical myths surrounding proposals for a new Argentine capital, while Cesar Caviedes (1995) and Jack Child (1985) have examined the salience of geopolitics in guiding policy in South America. Outside of academic geography, geopolitics has also been a popular subject for military officers, and even dictators such as Augusto Pinochet.

Geographers have also been concerned with studying how Latin America has been perceived and represented by both insiders and outsiders, by different interpretive communities, and how these spatial or locational perceptions interact with systems of power and knowledge. For example, Karl Zimmerer (1996) has examined how different groups interpret soil erosion in Andean Bolivia, thus creating potential for confusion in conservation planning. This is an exciting area with considerable growth potential.

4. *Human–Environment Relations*

The interest in global environmental change has led to the creation of projects to study the impact of climate change on Latin American environments and societies. British geographer Martin Parry, working with teams in Ecuador and Brazil, coordinated a pioneering project modeling the impacts of projected climatic changes. His methodology involved using specialists to model various systems (social, economic, environmental), and then to model the interactions between these systems. Many geographers have worked to document human dimensions of environmental change at a variety of scales, from the local to the regional (Steve Walsh and his colleagues in the Ecuadorian Amazon, Bill Turner and his colleagues in Yucatan, Diana Liverman and Karl Butzer (1992) in Mexico).

Latin Americanist geographers have continued to do research on the relationships between populations and their social and environmental resources, following the earlier tradition of Carl Sauer, James Parsons (1989), and Robert West. This has included research in the very rich (and still largely underexplored) archives of geographical data, and maps in the Americas and Spain. Colonial patterns and processes have been studied by David Robinson (1990), Linda Newson, George Lovell, Karl and Elisabeth Butzer, and other historical geographers. Many geo-

graphers, including William Denevan (1992) and his students, have continued to study the implications of the Conquest for regional population collapse, either using evidence from historical records and abandoned settlements, and agricultural fields, or mathematical models (Thomas Whitmore). The collapse in turn has implications for environmental change; William Denevan has challenged the 'pristine myth' of an American environment primarily impacted by post-Colombian processes, pointing out that the post-Colombian population collapse led to a resurgence of native forests.

Investigations into agricultural intensification and agricultural landforms have been pursued by many geographers. These studies have documented and reinterpreted the rationales of agricultural terracing (Robin Donkin), irrigation (William Doolittle), raised fields, shifting cultivation, floodwater farming, and other practices. William Woods and Joseph McCann have shown that 'black earth' soils are widespread in the Amazon, and indicate the importance of sedentary settlement away from the rivers in that region. Geographers such as Zimmerer and Knapp have amended earlier environmental and ecological models by Holdridge, Joseph Tosi, and Javier Pulgar Vidal to better characterize the 'strategically relevant environment' for groups with multiple crops and complex field systems. Much of this work combines contemporary ethnographic field interviews with research into the archives and attention to archaeological data of fields, settlements, and road patterns. Progress has been made in documenting the rich ethnogeographic lore and perceptions of environment by local peoples, whether of soils (Barbara Williams) or of traditional maps or quasi-maps (William Gartner).

5. *Physical Geography and GIS*

Physical geographers have increasingly been attentive to human–environment relationships, helping to break down the barriers between social, earth, and biological science. Ken Young and Blanca Leon, for example, have documented contemporary biodiversity at a variety of spatial scales, in the context of human activities. Tom Veblen and his colleagues have studied disturbance regimes along environmental gradients in northern Patagonia, and shown that both human-historical and natural disturbances need to be studied in tandem to explain vegetation patterns at the scale of landscapes. Carol Harden's work on soil erosion has shown that abandoned land and roads/trails are especially critical in accelerating erosion in Latin America's mountain regions. Work by Karl Butzer, Paul Hudson, and others is showing that Latin American watersheds need to be studied as units, and in the context of both human and natural impacts, to fully understand river behavior and aid in water resource management decisions.

By 1970 the survey and mapping of Latin America was well underway, with the use of the tools of air photography and (increasingly) remote sensing. The recent rapid expansion of Geographic Information Systems and other technologies, in conjunction with remote sensing, gives geographers a set of powerful tools for addressing environmental and planning problems. Numerous GIS facilities have been established in Latin America to analyze a wide range of social and environmental problems.

6. *Academic Geography within Latin America*

Within Latin America, there has been a vigorous development of academic geography, especially since the 1970s. Most academic programs in geography, as well as geographic institutes, in Latin America have tended to have a strong focus on contemporary urban and development issues, physical geography, and geographic techniques. Brazilian geography is strong, influenced both by the critical tradition of Milton Santos (1975) and by the vigorous national planning infrastructure. Mexican geography has been best developed at the Autonomous University of Mexico (UNAM) which has a variety of geography programs and an active publication series focusing on contemporary applied and urban geography, but also including physical geography. In Argentina, a numerous and lively younger generation operating out of centers in Buenos Aires, Mendoza, Tucuman, and elsewhere provides a diverse geographical scene. Colombia, Costa Rica, Cuba, the Dominican Republic, Puerto Rico, and Chile also have strong local geographic communities. In Peru, Ecuador, and other countries, local geographic societies and small academic communities are supplemented typically by a strong representation of international geographers resident or doing research in these nations. Historically, Latin Americans have found it easier to network with colleagues in Europe or North America than with other Latin American countries, but this is changing as Latin Americans develop their own international meetings and contacts.

Most Latin American nations have geographic societies with local publication activities and meetings. Geographers, especially European geographers, have been strong participants of the International Congress of Americanists. Geographers also participate in the national Latin Americanist organizations, such as the American Latin American Studies Association. The major organization of Latin Americanist geographers in the US has been the Conference of Latin Americanist Geographers, founded in 1970. It publishes a refereed journal, conducts international meetings, and provides support for international contacts and special projects. It includes a significant Latin American and European membership. More recently the Latin American Specialty Group of the Asso-

ciation of American Geographers has served the needs of the (mostly American) members of that association. Until recently, many Latin American geographers have found it difficult to attend international meetings, but this is changing.

7. Conclusion

At the time of writing, Latin Americanist geography is healthy, with a variety of methodological and theoretical approaches, new computer techniques, and an enhanced focus on environmental, cultural, and gender issues. The environmental focus is particularly noteworthy as helping to bridge the gap between the natural and human sciences. There is increased cross-fertilization between different national traditions, more collaboration between geographers of different specialties, and more stress given toward participatory research and empowerment. At the same time, Latin Americanist geography's historic strengths in regional and historical analysis have never been more in evidence. The next ten years promise to exhibit more large, interdisciplinary projects focusing on socially significant problems of environmental change and global–local interactions, in which the geographer's skills at studying the interaction between different systems on a variety of scales will be even more crucial.

See also: Dependency Theory; Latin American Studies: Economics; Latin American Studies: Politics; Latin American Studies: Society; Remote Sensing and Geographic Information Systems Analysis; Sauer, Carl Ortwin (1889–1975)

Bibliography

Butzer K W (ed.) 1992 The Americas before and after 1492: current geographical research. *Annals of the Association of American Geographers*, Special issue **82**(3)

Caviedes C, Knapp G 1995 *South America*. Prentice-Hall, Englewood Cliffs, NJ

Child J 1985 *Geopolitics and Conflict in South America*. Praeger, New York

Deler J-P 1981 *Genese de l'Espace Equatorien: Essai sur le Territoire et la Formation de l'Etat National*. Paris, Editions ADPF; Lima, Institut Francais d'Etudes, Andines

Denevan W M (ed.) 1992 *The Native Population of the Americas in 1492*, 2nd edn. University of Wisconsin Press, Madison, WI

Friedmann J 1968 *Regional Development Policy: The Case of Venezuela*. MIT Press, Cambridge, MA

Harrison P, Turner B L II (eds.) 1978 *Pre-Hispanic Maya Agriculture*. University of New Mexico Press, Albuquerque, NM

Hecht S B, Cockburn A 1990 *Fate of the Forest*, 2nd edn. HarperCollins, New York

Nietschmann B 1973 *Between Land and Water: The Subsistence Ecology of the Miskito Indians, Eastern Nicaragua*. Seminar Press, New York

Parsons J J 1989 In: Denevan W M (ed.) *Hispanic Lands and Peoples: Selected Writings of James J. Parsons*. Westview Press, Boulder, CO

Preston D (ed.) 1996 *Latin American Development: Geographical Perspectives*. Longman, Harlow, UK

Reboratti C 1987 *Nueva Capital, Viejos Mitos*. Sudamericana/Planeta, Buenos Aires, Argentina

Robinson D J (ed.) 1990 *Migration in Colonial Spanish America*. Cambridge University Press, Cambridge, UK

Santos M 1975 *L'espace Partage: Les Deux Circuits de l'Economie Urbaine des Pays Sous-developpes*. M.-Th. Genin, Paris

Ward P M 1998 *Mexico City*, 2nd edn. Wiley, Chichester, UK

Zimmerer K 1996 *Changing Fortunes: Biodiversity and Peasant Livelihood in the Peruvian Andes*. University of California Press, Berkeley, CA

G. W. Knapp

Latin American Studies: Health

Despite overall improvements in health, the persistence of inequality in health outcomes for the poorest segments of most societies and inequitable access to healthcare resources has provided rich territory for social scientists working in Latin America. Health studies inevitably attract cross-disciplinary researchers, given the multiple roles that ill health and disease play within society. In addition to parent disciplines, several hybrids of the social sciences have focused on health, including medical anthropology, social demography, political economy, medical history, medical ecology, social epidemiology, and medical geography. This article examines the state and trends within social science research in Latin America over the past few decades, selecting the most common research themes tackled by these disciplines. The importance of the early 1990s for health and social science is captured by two dates. First, the quincentennial year of 1992 symbolized and energized the study of the contact between European and indigenous, New World populations, and the health ramifications and depopulation that followed. Second, a major outbreak of cholera, beginning in the early 1990s, which crossed the South American continent from east to west, shook the confidence of public health officials and scientists alike in disease control measures. Other themes include globalization, changing trade relations and complex migration flows between and within countries, rapid agrarian transitions, deteriorating environments, urbanization, and the changes in health delivery systems. At the micro level, medical anthropologists and sociologists have examined traditional health systems including folk medicine, healers, and the cultural contexts for multiple illnesses and treatments. These studies have helped enrich another bio-behavioral focus of study: ethnobotany and its counterpart ethno-pharmacopoeia. Between these macro and micro approaches lie sub-

national or community studies where themes of privatization, decentralization, sustainability, and the role of civil society organizations are investigated in the light of (mainly deteriorating) public healthcare delivery systems.

1. Definition of Health

'Health studies' incorporate the analysis of health status as an outcome, where health is defined as 'a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.' Health status is also used in many social science studies as a determinant, for example, where poor health is seen as a push factor for migration to cities. But health studies also include the study of the organization of healthcare services, whether they are formal public or private services modeled on Western biomedical systems, or more informal, also private, traditional health services.

2. Contact Studies

Much research dating from the 1930s (Cook 1937) has investigated the depopulation of the New World following the European contact. The general consensus is that, by the late 1700s, levels had declined to 10 percent of the original population (Whitmore 1992). Critics of the reconstructed figures argue that the estimates are inflated for the pre-Colombian population (e.g., Henige 1998, Rosenblat 1992). The causes of depopulation are also disputed, though most agree that the primary cause was the introduction of Old World diseases to which the native populations had little resistance. Recent studies of the genetic constitution of these vulnerable populations suggest that high homogeneity meant that their low resistance to new pathogens was also equally homogeneous and can probably account for high mortality rates from smallpox, chickenpox, measles, influenza, and other infectious diseases.

The notion of a 'health revenge' by the New World in the form of syphilis has also been a fertile field of study in biosocial science. Medical geographers have traced the path of syphilis as it entered an apparently susceptible population (in Europe from the Caribbean and/or Meso-America) where the pathogen is hypothesized to have converted itself from a relatively harmless skin disease into a virulent sexually transmitted infection. The latest studies (using genetic markers) suggest that the infectious agent for syphilis existed on both sides of the Atlantic, but for reasons that are still unexplained the transfer of the New World version into Europe produced a more dangerous pathogen.

The term 'Colombian Exchange' has been commonly used since first popularized by Crosby (1972), to symbolize the biological, cultural, and sociopolitical consequences of the contact. For example, the impact of a changed diet has been examined by historical demographers and anthropologists in relation to survival, health, and nutrition. The changed diet came about through altered trade relations, the export of food items such as corn, potato, and tomato into Europe, and the import to Latin America of wheat, soya, sugar, rice, and many other foods, and massive ecological disruption in the New World as a result of pig importation (Settipane 1995).

3. Ethnomedicine

The idea of 'exchange' and the consequences for the health of native populations has also affected the way ethnomedicine has been documented and interpreted by anthropologists. Ethnomedicine has provided rich research areas including: the study of healers and the cured, the incorporation of Mediterranean medicine into healing traditions, the role of traditional African medicine and practices, patient–provider relationships, community-level integration between Western medicine and traditional healing systems, and studies of ethnobotany and ethno-pharmacopoeia. One contested area has been humoral medicine (e.g., Foster 1994, Tedlock 1987), a medical system based on concepts of fluid, temperature, and temperament balance. Humoral medicine encompasses the diagnostic methodology of defining illnesses as either hot or cold (or hotter or colder—i.e., a continuum), and thence prescribing hot or cold herbal or other treatments to (counter) balance the condition.

One ongoing controversial aspect has centered on the historical and cultural origins of humoral medicine. Did this classification originate from the contact period as Greco–Hellenic–Arab notions reflected in Spanish medicine were adopted by indigenous cultures (Foster 1994)? Was this development a coincidence, and the hot/cold categorization of illness and cures simply an easy generalization that is found in an enormous array of different cultures around the world (Leatherman 1998)? The notion of syncretism is generally understood to describe the coalescence or joining together of traditions from different cultures. As applied to the contemporary debate over ethnomedicine's origins, we can define it as meaning the adoption of traditions from one culture to another where the form remains the same, but the meaning and interpretation of the tradition changes.

Research has also been conducted around whether or not particular cures work, or whether they may actually, in some cases, do more harm than good (Greenway 1998). This has required extensive documentation and evaluation of folk medicines in a variety of settings. The Andean nations, Guatemala, Mexico, and Brazil have been the sites of most research. A few analysts have compared treatments using traditional

medicine to equivalent treatments in biomedicine for certain illnesses (e.g, Bastien 1998). But much remains to be done in this field of comparative medical efficacy.

Some of these challenges in ethnomedical research center on the problems of scale and unit of analysis (Lock and Scheper-Hughes 1990). Ethnographic research, using interpretist frameworks, treats the health outcome as integral to understanding the meaning behind the particular treatments. Others rely on a more positivist framework, requiring 'objective' standards by which to judge the outcome of treatments (see Greenway 1998, Briceño'León et al. 1998). Put another way: social science has met a major challenge as it tries to answer questions of biomedical efficacy, since biomedical is *only* one form of efficacy, and the meaning of the use of particular herbs often extends beyond the mechanistic understanding of 'fixing' a single sick body part.

Controversy also centers on how to understand the medical pluralism that is almost ubiquitous in most societies, and specifically how traditional medicine should be integrated into biomedical health systems (Bastien 1992, Leatherman 1998). Arguments focus on the meaning of 'integration,' as well as the continuing debate of efficacy and safety of both modern and older pharmaceutical regimes in contexts where, apart from a few urban settings, drug distribution and control is largely unregulated. Some researchers and practitioners even see the pressure to include folk medicine into modern health systems as a way of justifying the limited introduction of modern healthcare in the remoter rural areas (Foster 1994). Meanwhile, Cuban health officials, in reaction to economic shocks from the US embargo and the removal of the Soviet-era subsidies in the early 1990s, and from the incentive to treat chronic illnesses with a wider range of treatments, have developed one of the most complete examples (in the Americas) of traditional medicine being officially incorporated within the biomedical system (Feinsilver 1993).

4. *Macro Forces and Healthcare*

Health and political economists have examined policy debates that encompass the relationships between the state, economic development, the health systems, and health and nutrition outcomes (Ugalde 1985, PAHO 1998). Debate revolves around two interrelated questions: first, what is the nature of the relationship between inequality (persistent and increasing poverty among the lower income groups), health, and economic growth? Second, what is the nature of the relationship between political participation, health, and economic growth (Frenk 1989, WHO 2000), and what is the role of the state in healthcare? Since Weil and Scarpaci's (1992) analysis of the health in Latin America in the 'lost decade' (the 1980s), two streams of research have emerged. One stream has followed health services research paradigms examining payment mechanisms in different contexts, the fiscal instability of social security systems, and problems associated with health providers and health delivery systems.

A second school challenges the so-called 'Americanization' of healthcare systems. Of particular concern has been the impact of wide-scale privatization, the role of corporations in the delivery of healthcare, and the increase in health maintenance organizations. An array of reforming systems is being studied across the continent—from the municipalization with universal care in Cali, Colombia, to the competition with choice in Argentina. Costa Rica and Cuba stand out as examples of primary care models that can be regarded as classic system-level initiatives in health delivery. The critical examination of this 'Americanization' is an interesting twist on research that a few years ago was calling for more Westernization of health services (Waitzkin 1998, Stocker et al. 1999).

5. *Health Transitions*

Great efforts have gone into trying to explain why there is so much heterogeneity among Latin American populations, and why inequalities of health have been exacerbated over the past 50 years—years of increasing economic, if not social, development (Weil and Scarpaci 1992, PAHO 1998). A useful framework for describing the recent demographic experience is the demographic transition framework. Most Latin American nations have followed a common pathway moving from high to low mortality and fertility rates.

For example, overall in the last 15 years of the twentieth life expectancy rose from about 66 to 72 years, with a range of between 78 to 55 between the best and worst national performers. These shifts have changed the population structure from a typically wide-based pyramid, with a majority of its population below 15 years), to a more evenly distributed population structure. Concurrent with (or as a consequence of) the demographic is an epidemiological transition, where the common diseases of a population change from acute infectious to chronic illnesses such as coronary heart disease, diabetes, and cancer—the so-called life-style diseases. Deaths from these diseases outnumbered deaths from infectious diseases by 10 to 1 in 2000. This was a doubling of the ratio since 1985. These transitions, collectively known as the health transition, might make a modern-day cynic conclude that for many, life in much of Latin America is still nasty and brutish, but for a multitude of reasons, it is no longer all that short.

6. *Conclusions: Continuity and Change*

In general terms, there is a recognition among social and medical scientists of the need to substantially broaden the frame of analysis for health, development, and environmental change to include psychosocial

and biocultural dimensions. But despite calls for this more inclusive approach, there is little follow-through yet. In terms of specific health problems, much more research is needed to reflect the changes within Latin societies; for example there have been few studies involving occupational health or chronic illnesses such as cardiovascular disease, diabetes, and cancer. Dependency problems associated with changing lifestyles such as alcohol, tobacco, and other drug addictions, or their consequences such as depression and other mental health challenges are beginning to attract some research (Good 1997). The consequences of natural disasters have also begun to attract attention in social epidemiological studies, but this has not yet been reflected in further interest within other social sciences.

Special populations such as children, adolescents, and poor women have energized a broad range of studies: from specific studies of family planning access and use, to integrated studies of reproductive health and development, to vulnerability to tuberculosis and AIDS. Many of these sub-population studies can trace their origins to the mainstreaming of feminist theory, to development studies that implicate women's low status in many societies, and to the empowerment of indigenous populations and indigenous movement in general and the study of marginalized groups such as street children. The most recent developments in this trend is the study of the violation of human rights in the context of health. In the case of adolescent health studies, much recent work stems from the perception of adolescents being increasingly vulnerable to sexually transmitted infections. Social science may yet play a significant role in improving our understanding of the social networks of these children, and help policy makers provide better protective environments for them.

Other new themes include the application of ecological concepts to the study of complex relationships between humans, their landscapes and the pathogens and vectors that cause disease. An investigation of the relationship between ecology and disease was incorporated into social science research (especially in geography) beginning in the 1960s, but the variables were generally limited to the impact of altitude on health and the weight/height adaptations to different ecological zones. There has always been an interest in the ways in which human behavior and human health may vary by place, but the field of disease ecology sets out to investigate the synergistic relationship between a specific disease (both pathogen and vector), a given ecology and the human factors and behaviors that affect that environment. Disease ecologists and human behavioral scientists are beginning to collaborate, incorporating multi-level frameworks (from micro to community level) to enhance our understanding of specific disease and human environments for diseases such as chagas, dengue fever and DHF, malaria, and other infectious parasitic and vector-borne diseases. Human adaptation of the landscape, particularly in the form of urbanization and widespread monocropping, has provoked changes in disease patterns that can only be understood from a perspective that integrates various disciplines, and moves outside that traditional realm of treating the patient in order to control disease.

Although it is too early to tell how long-lasting a tradition this may be, it would seem that the AIDS epidemic has succeeded in overcoming some of the more rigid boundaries between disciplines (Farmer et al. 1996). There are signs that as the social and biomedical fields begin to recognize each other's contributions, social epidemiologists, psychologists, infectious disease clinicians, and medical anthropologists have begun to collaborate to advance our still very basic understanding of the interactions between social, political, and cultural structural constraints and the power of individuals.

Bibliography

Bastien J W 1992 *Drum and Stethoscope: Integrating Ethnomedicine in Bolivia*. University of Utah Press, Salt Lake City, UT

Bastien J W 1998 *The Kiss of Death: Chagas' Disease in the Americas*. University of Utah Press, Salt Lake City, UT

Bobadilla J L, Frenk J, Lozano R, Freijka T, Stern C 1993 The epidemiological transition and health priorities. In: Jamison D, Moseley W H, Measham A, Bobadillo J L (eds.) *Disease Control Priorities in Developing Countries*. Oxford University Press, New York

Briceño-León R, Bronfman M, Bonilla Castro E 1998 Research in the social sciences and health. Working Paper, June, PAHO/ACHR/98.11

Cook S F 1937 *The Extent and Significance of Disease Among the Indians of Baja California, 1697–1773*. University of California Press, Berkeley, CA

Crosby A W 1972 *The Columbian Exchange: Biological and Cultural Consequences of 1492*. Greenwood, Westport, CT

Farmer P, Connors M, Simmons J 1996 *Women, Poverty and AIDS: Sex, Drugs and Structural Violence*. Common Courage Press, Monroe, ME

Feinsilver J 1993 *Healing the Masses: Cuban Health Politics at Home and Abroad*. University of California Press, Berkeley, CA

Foster G 1994 *Hippocrates' Latin American Legacy: Humoral Medicine in the New World*. Gordon and Breach, Langborne, PA

Frenk J 1989 Health transitions in middle income countries: New challenges for health care. *Health Policy and Planning* **4**: 29–39

Good B 1997 Studying mental illness in context: Local, global, or universal? *Ethos* **25**(2): 230–48

Greenway C 1998 Landscapes of health in the Andes. *Social Science and Medicine* **47**(8): 989–92

Henige D 1998 *Numbers from Nowhere: The American Indian Contact Population Debate*. University of Oklahoma Press, Norman, OK

Leatherman T 1998 Biocultural perspectives on health in the Andes. *Social Science and Medicine* **47**(8): 1031–41

Lock M, Scheper-Hughes N 1990 A critical–interpretive ap-

proach in medical anthropology: Rituals and routines of discipline and dissent. In: Johnson T M, Sargent C F (eds.) *Medical Anthropology: Contemporary Theory and Method*. Praeger, New York, pp. 47–73
PAHO 1998 *Health in the Americas*. Pan American Health Organization, Washington, DC
Pendersen D, Baruffati V 1989 Healers, deities, saints and doctors: Elements for the analysis of medical systems. *Social Science and Medicine* **29**(4): 487–96
Pendersen D 1996 Disease ecology at a crossroads: Man-made environments, human rights and perpetual development utopias. *Social Science and Medicine* **43**(5): 745–58
Rosenblat A 1992 *The Population of Hispaniola at the time of Columbus. The Native Population of the Americas in 1492*, 2nd edn. University of Wisconsin Press, Madison, WI
Scheper-Hughes N 1992 *Death Without Weeping: The Violence of Everyday Life in Brazil*. University of California Press, Berkeley, CA
Settipane G A 1995 *Columbus and the New World: Medical Implications*. Oceanside Publications, Providence, RI
Stocker K, Waitzkin H, Iriart C 1999 The exportation of managed care to Latin America. *New England Journal of Medicine* **340**(14): 1131–6
Tedlock B 1987 An interpretive solution to the problem of humoral medicine in Latin America. *Social Science and Medicine* **24**(12): 1069–83
Ugalde A 1985 Health and social science in Latin America. *Social Science and Medicine* **21**(1): 1–3
Waitzkin 1998 Is our work dangerous? Should it be? *Journal of Health and Social Behavior* **39**(1): 7–17
Weil C, Scarpaci J 1992 *Health and Health Care in Latin America in the Lost Decade: Insights for the 1990s*. Pan American Health Organization, Washington, DC
Whitmore T M 1992 *Disease and Death in Early Colonial Mexico: Simulating Amerindian Depopulation*. Westview Press, Boulder, CO
WHO 2000 *World Health Report. Health Systems: Improving Performance*. World Health Organization, Geneva

K. Macintyre

Latin American Studies: Politics

The study of politics under the rubric of 'Latin American studies' bears two important features that distinguish it from the more general practice in US university departments of political science: it has been approached from an interdisciplinary perspective and it has been influenced importantly by Latin Americans themselves. Through these methods, 'Latin Americanists' have contributed to some of the most important and influential theories and debates in political science in recent history, from dependency to democratization, from studies on the state to research on social movements. This article examines the intellectual trajectory of the study of politics within Latin American studies, noting how it has been influenced by real-world events and discussing its relationship, intellectual and political, to the discipline of political science.

1. *Emergence of the Subfield*

While some US scholars devoted themselves to the study of Latin America, including the region's politics, in the late nineteenth and early twentieth centuries, it was not until the 1960s, after the Cuban Revolution, that Latin American studies got its big boost. With important support from the Joint Committee on Latin American Studies of the Social Science Research Council and the American Council of Learned Studies, politics became a major scholarly focus for regional specialists. By 1979, articles by political scientists made up the largest percentage of submissions to the leading journal devoted to the area, the US-based *Latin American Research Review*, a trend that continued in the 1980s and 1990s. Studies on politics also composed an important percentage of articles in the *Journal of Latin American Studies*, published in England.

Notwithstanding the Cold War 'national interest' incentives for funding US scholarship on Latin America, many tensions arose over the years between the Latin American studies community and the foreign policy, defense, and intelligence circles of the US government. Beginning with the 'Operation Camelot' scandal in 1964, in which a sociologist working in Chile exposed a US Army-funded initiative to use area scholars to gather information for the US government's counterinsurgency program, scholars of the hemisphere criticized a wide array of US policies, including those toward Cuba, multinational corporations, Brazil, Central America, and especially Chile. The Latin American Studies Association, founded in 1964, has issued frequent and strong criticisms of US policy in the hemisphere. This passionate engagement of US and European Latin Americanists with policy issues in the region is one outcome of their genuine collaboration with their counterparts to the south, most of whom have been sharply critical of US imperialism, interventionism, capitalism, conservatism, and association with dictators.

2. *From Modernization to Dependency*

The mutual influence of Latin American and North American/European scholars has been evident in the evolution of political science inquiry on the region. In the 1950s and early 1960s, 'modernization' was the predominant paradigm in the field of comparative politics. Modernization theory arose in the context of decolonization in Africa and Asia and the early years of the Cold War. It grew out of efforts to understand how recently independent nations and other 'Third World' countries might achieve economic and political development similar to that of the US and northern Europe, which were viewed as the products of a linear and potentially universal process of rationalization and progress. The theory was developed largely by specialists on Africa and Asia, but Latin Americanists fell in line to offer supporting evidence and concepts.

Following the spread of US interest in the region from the Caribbean basin towards the larger, industrializing countries, scholarly attention turned towards Mexico and the Southern Cone of Latin America.

The main argument of modernization theory was that industrialization and economic growth, and the value orientations associated with them, were the engines of social and political progress. This was a vision rooted in classic, Western liberal economic and political thought. In order to develop, Third World societies needed to embrace ideas, values, techniques, and organizations commensurate with urbanization, a complex division of labor, increased social mobility, and a rational–legal, impersonal economic and political system. As countries overcame feudal, semi-feudal, precapitalist, or at least inefficient behavior patterns and institutions from the past, new urban social groups, particularly the middle classes, would emerge, and these groups would in turn push for social equality and political democracy. Thus, political scientists focused on issues such as elite and mass education, mobilization of the popular classes, interest articulation, and institutional development. Some of the pre-eminent works of this period were Johnson's (1958) *Political Change in Latin America*, Lambert's (1967) *Latin America: Social Structure and Political Institutions*, and Lipset and Solari's (1967) edited volume, *Elites in Latin America*.

Quickly, however, critics assailed the main tenets of modernization theory based on evidence and perspectives from Latin America itself. In the 1960s and 1970s, economic growth in most countries did not meet expectations, social inequalities were rarely reduced, and military dictatorships became the norm in the region. Scholars thus began to take issue with many of the underlying assumptions of modernization theory. They challenged the idea of a linear, evolutionary developmental continuum, the conception of preindustrial societies as homogeneous and static, the assumption of the Western European, capitalist experience as generalizable and desirable, the faith that new urban social sectors would be democratic and progressive, and the neglect of constraining factors exogenous to Third World societies.

The primary challenge to modernization theory was the dependency approach. The dependency school accepted modernization's economic determinism, but turned it on its head: the adoption of US and European-style economic policies had not and would not lead to healthy economic and political development, but rather to skewed and highly limited development, or 'underdevelopment.' Rather than the cure for underdevelopment, capitalism was seen as the cause. The dependency approach, developed mainly by Latin Americans but also by foreign Latin Americanists, and highly influential outside the region, melded insights from a variety of disciplines to explain the major features of Latin American societies: economic underdevelopment, social inequality, political instability, and authoritarianism. It emphasized the need to go beyond the examination of individual societies to understand the international historical process of development.

Within the dependency paradigm, economists, sociologists, and political scientists argued that a country's position in the international system is determinant of internal class behavior. Because Latin American countries occupied an inferior position in the international division of labor, producing mainly raw materials and cheap workers, they were the victims of unequal terms of trade and exploitation by foreign investors. Local entrepreneurial classes and political leaders were captives of the international market and had only limited opportunities to steer the development of their own economies and societies. In addition, the copying of consumption patterns characteristic of the advanced industrialized countries led to severe distortions within Latin American economies. According to the more radical dependency writers, foreign and national capitalists siphoned off Latin America's surplus, leaving the vast majority of people sunk in poverty and oppressed by authoritarian regimes. Some of the most influential books on dependency were Frank's (1967) *Capitalism and Underdevelopment in Latin America*, and Cardoso and Faletto's (1969) *Dependencia y Desarrollo en América Latina*.

In political science, the most important theoretical contribution to come out of the dependency paradigm was bureaucratic–authoritarianism, articulated by O'Donnell in (1973) *Modernization and Bureaucratic–Authoritarianism*. This theory argued that the bureaucratic–authoritarian state that emerged in the most economically advanced Latin American countries was a structurally determined phenomenon, produced by the alliance of local political forces and foreign capital with the objective of overcoming economic stagnation and deepening industrialization. Otherwise put, the authoritarian regimes in the region emerged because continued economic growth depended on the repression of the working classes, whose demand for higher wages and other guarantees would otherwise fuel inflation and drive out foreign investment.

Dependency arguments held sway into the late 1970s, when Latin American countries began a wave of transitions to formally democratic regimes. Already, dependency-related economic theories had been displaced by international monetarism in many countries, increasingly so as the 1980s unfolded. Both import substituting industrialization and socialism seemed to have failed to overcome underdevelopment. In addition, the examples of export-led development in East Asia, namely South Korea and Taiwan, cast heavy doubt on the pessimistic tenets of the *dependentistas*. Latin American governments thus began slashing trade barriers and encouraging comparative advantage, while they pruned the bloated

public sector. Moreover, foreign investment and loans were welcomed, particularly in the wake of the debt crisis of the 1980s. Along with the imperative for economic liberalization came pressures, from both within and without the region's authoritarian regimes, for political liberalization. Suddenly, scholars who had argued that authoritarianism in the region was structurally determined had to account for transitions to democracy.

3. *Corporatism*

A major alternative theory produced during the 1970s to explain Latin American economic and political development patterns, and one that was also challenged and swept aside with the focus on democratization in the 1980s, was corporatism. Corporatist theory had two main, sometimes overlapping, strains. One strain was institutional corporatism, fully developed first by Schmitter (1971) in *Interest Conflict and Political Change in Brazil*. Schmitter and others called attention to a pattern of interest group politics in Latin America in which the state played a dominant and controlling role. In an effort to harmonize relations among classes and other social groups, the state often sponsored the formation of functional groups, offering them a monopoly of representation but maintaining a significant degree of control over their internal organization and demand making. This pattern did not fit the pluralist model based on the American cases, but bore certain similarities to—as well as important differences from—interest group politics in some Western European countries (Schmitter 1974).

Proponents of the second version of corporatist theory, cultural corporatism, took the presence of corporatist institutions as evidence of a distinct political tradition, rooted in the Iberian and colonial past. Scholars such as Wiarda (1974), editor of *Politics and Social Change in Latin America*, held that Latin American societies had inherited a political culture that featured feudalistic master–man relations, anticapitalist preferences and incentives, patrimonial extended families, hierarchical Roman Catholic religious affinities, corporatist and organic links between the state and society, and authoritarian, verticalist governing structures. According to proponents of this point of view, such characteristics were not necessarily undesirable or destined to vanish with economic development, as modernization theorists would have it, but were part of Latin America's unique developmental path. These arguments gained currency as authoritarian regimes swept the hemisphere in the 1970s.

Cultural corporatism was criticized strongly on a number of fronts. Some viewed it as an apology for authoritarian regimes that emerged in Latin American countries. Others (including Schmitter 1974) questioned the theory's blurring of cultural and institutional elements, arguing that forms of governmental corporatism can be found in many non-Latin, developed countries, serving to organize functional groups as links between the state and society. They also pointed out that many other developing countries are plagued with poverty and authoritarianism without any Iberian heritage.

The most telling critique of cultural corporatism, however, was launched against its sweeping determinism. Despite their common colonial past, Latin American countries exhibit an extremely varied economic and political record. Some countries have made impressive strides in terms of industrialization and the institutionalization of democratic practices. Even within individual countries, people have experienced radically changing forms of economic, social, and political organization without suddenly transforming their cultural heritage. Contrasting levels of modernity exist between different cities and regions. Moreover, cultural attributes vary widely by ethnicity, race, class, and gender. Cultural corporatism could not account for this diversity, nor could it predict or explain the wave of democratization that swept Latin America beginning in 1978.

4. *Democratization and Beyond*

In the 1980s, then, real-world events forced analysts of the region to turn their attention to the politics of liberalization and transition from authoritarian regimes. Gradually they abandoned the grand theorizing and structural determinism that had characterized both the modernization and dependency eras to focus instead on the dynamics of agency, the role of ideology, the issue of political will, and the application of game theory. In this effort, scholars could draw on the work of Linz and Stepan (1978), who had emphasized the importance of agency in their important edited work, *The Breakdown of Democratic Regimes*. O'Donnell et al. (1986) produced the four-volume set, *Transitions from Authoritarian Rule*, and Przeworski (1991) wrote *Democracy and the Market*, both of which were hugely influential in the field of political science as a whole.

The 'transitology' literature, which focused almost exclusively on the interests and actions of political elites, soon gave way to analyses of the different types of democratic regimes that emerged in the region, debates about the meaning and possibility of 'democratic consolidation,' and greater attention to civil society. Scholars who, early on, were wary of upsetting the delicate equilibria achieved in the region's democratic transitions gradually became more critical. Some took strong stands in debates over institutional reform, arguing that with strong presidentialism or

poorly designed electoral rules, Latin American countries courted instability. Others, pointing to a crisis of representation in many of the region's polities, focused on the need to strengthen parties or analyzed the important role of grassroots organization.

This does not mean that either modernization or dependency was swept definitively into the dustbin of history. Like most good theories in the social sciences, both bequeathed a legacy of important lessons and middle-level hypotheses, shorn of their more grandiose pretensions. Both theories contributed an abiding concern with underlying structural conditions, especially with dependency's emphasis on historical structuralism, although most political scientists now insist that we must focus on institutions, agents, identities and/or choices as well as structures. Dependency left its mark in terms of general awareness of the important role of external factors to the internal economic and political systems of Latin America, and political economists specializing in Latin America moderated and adapted the theory to explain persuasively instances of 'dependent development,' not only in Latin America but in other regions of the globe as well (e.g., Evans 1979). Remnants of modernization theory are evident in some recent analyses which attempt to establish prerequisites, economic and/or cultural, of political democracy (e.g., Malloy and Seligson 1987). Now some argue that it just took longer than expected in the 1950s for Latin America, let alone Southern and Eastern Europe and many other parts of the globe, to achieve the socioeconomic conditions to support democratic government.

Notwithstanding this legacy, in the 1980s and 1990s, Latin American studies has been characterized by the absence of a prevailing paradigm. To be sure, analysts of Latin American politics have not been unaffected by the more recent trend toward theory-driven analysis, whether shaped by world-systems theory, rational choice theory, or postmodernism. For example, some study the politics of migration flows among peripheral, semi-peripheral, and center countries in the context of the globalization of capitalism; others examine how politicians' desire to maximize their career success affects policy; and still others analyze the political perspectives and resistance strategies of historically trampled or ignored ('subaltern') groups. For the time being, however, studies within these paradigms must share the intellectual terrain with an abundance of middle-level theories. Such theories have emerged on topics ranging from the political effects of different approaches to incorporating labor into the polity, to the sources and political effects of different institutional structures, to the emergence and effectiveness of social movements, to changing constructions of gender and citizenship. Among the most significant of these recent works are Collier and Collier's (1991) *Shaping the Political Arena*, Mainwaring and Scully's (1995) *Building Democratic Institutions: Party Systems in Latin America*, Escobar and Alvarez's (1992) *The Making of Social Movements in Latin America*, and Linz and Stepan's (1996) *Problems of Democratic Transition and Consolidation*.

5. *Political Science vs. Area Studies*

Despite all of the noteworthy contributions discussed here, Latin American area specialists are still sometimes accused, as are other area specialists, of 'ghettoizing' themselves within the discipline of political science. Few articles on Latin America (or on other 'developing regions') have appeared in the pages of the leading disciplinary journals in the past 40 years (Martz 1990). Instead, most work on Latin America has been published as chapters in edited volumes, as collaborative multiauthored books, or as articles in area or alternative thematic journals.

Perhaps as a result, an old debate has re-emerged in the pages of major political science journals and newsletters regarding the quality of contributions by area specialists in general. The nature of work on Latin American politics, as most work in 'area studies,' has tended to be more qualitative than quantitative, and generally more oriented towards a transdisciplinary audience. Latin American studies in general is characterized by methodological diversity, a fact which partially may be explained by the significant influence of Latin American scholars, whose disciplinary norms and boundaries differ from those of their US counterparts. The work of Latin American specialists is thus sometimes viewed with suspicion by colleagues at the 'hard' end of the social scientific spectrum. Most recently, a Harvard political scientist and noted Africanist (Bates 1996) has argued that comparative political scientists should follow the lead of many specialists in US politics who use rational choice and game-theoretic models to produce testable hypotheses and strive for universalizable conclusions. He and those who share his convictions view area studies in much the same way their behavioralist predecessors did: as primarily descriptive, largely atheoretical, and, above all, methodologically soft and hence unanalytical or unscientific.

The view of Latin American studies as unscientific is, however, simply inaccurate. Regional specialists, including political scientists, have helped to test the validity of discipline-specific theories and to identify new empirical puzzles that require theoretical explanations and that generate hypotheses. It was the familiarity of Latin Americanists, for example, with particular historical and structural features of Latin American societies that allowed the universalist assumptions of modernization theory to be challenged, and produced the highly influential dependency approach in the 1970s. As dependency itself came under fire, it was a close empirical analysis by Latin America area experts that produced the concept of 'dependent

development' and led to the elaboration of theories of state-led development around the world. In the fields of democratization and social movement theory, it has been transcontinental and transdisciplinary cooperation by Latin America experts that have produced some of the most significant recent works.

6. The Bigger Challenge: Globalization

Notwithstanding the critique from within the academy, the greater challenge to specialists on Latin American politics comes from the broader international political context. The end of the Cold War has meant the demise of the general 'national interest' justification in the US for funding area studies programs. The issues that affect US security interests are increasingly understood as global problems, better handled by issue experts rather than area experts. In such a scenario, Latin American studies may be particularly vulnerable, given the association of many Latin American specialists with causes often at odds with those of the Washington policy community. Relatedly, the global expansion of US power in the 1990s, both economic and political, as well as the great leaps in communications technology of that same decade, have fed the notion that the world is becoming increasingly homogeneous. English has become the *lingua franca* of the international business and political worlds, and more and more countries have accepted and even embraced the 'Washington consensus' on neoliberal economics and procedural democracy.

The challenge for those who study Latin American politics in the twenty-first century is thus to forge cross-regional ties and build a global research agenda without losing sight of the factors that have allowed for fruitful intellectual production in the past: a commitment to deep knowledge of cases through language proficiency and fieldwork, openness to interdisciplinary cross-fertilization, and collaboration with foreign scholars on an equal basis. While globalization may signify that many different countries face similar problems, it does not mean that similar solutions will work everywhere. Local and regional traditions and politics will continue to influence events and outcomes in all parts of the world, and knowledge of those traditions and politics will continue to be essential for policy makers and academic theorists.

See also: Area and International Studies in the United States: Intellectual Trends; Corporatism; Democratic Transitions; Dependency Theory; Latin American Studies: Economics; Latin American Studies: Geography; Latin American Studies: Religion; Latin American Studies: Society; Minimum Winning Coalition, in Politics; Modernization, Sociological Theories of; Political Science: Overview

Bibliography

Bates R H 1996 Letter from the President: Area studies and the discipline. *APSCA-CP: Newsletter of the APSA Organized Section in Comparative Politics* **7**(1): 1–2

Cardoso F H, Faletto E 1969 *Dependencia y Desarrollo en América Latina: Ensayo de Interpretación Sociológico*. Siglo Veintiuno, Mexico City [1979 *Dependency and Development in Latin America*. University of California Press, Berkeley, CA]

Collier D, Collier R B 1991 *Shaping the Political Arena*. Princeton University Press, Princeton, NJ

Escobar A, Alvarez S E (eds.) 1992 *The Making of Social Movements in Latin America: Identity, Strategy, and Democracy*. Westview Press, Boulder, CO

Evans P 1979 *Dependent Development: The Alliance of Multinational, State, and Local Capital in Brazil*. Princeton University Press, Princeton, NJ

Frank A G 1967 *Capitalism and Underdevelopment in Latin America*. Monthly Review, New York

Johnson J J 1958 *Political Change in Latin America*. Stanford University Press, Stanford, CA

Klarén P F, Bossert T J (eds.) 1986 *Promise of Development: Theories of Change in Latin America*. Westview Press, Boulder, CO

Lambert J 1967 *Latin America: Social Structure and Political Institutions*. University of California Press, Berkeley, CA

Linz J J, Stepan A (eds.) 1978 *The Breakdown of Democratic Regimes*. Johns Hopkins University Press, Baltimore, MD

Linz J J, Stepan A (eds.) 1996 *Problems of Democratic Transition and Consolidation: Southern Europe, South America, and Post-communist Europe*. Johns Hopkins University Press, Baltimore, MD

Lipset S M, Solari A (eds.) 1967 *Elites in Latin America*. Oxford University Press, New York

Mainwaring S, Scully T R (eds.) 1995 *Building Democratic Institutions: Party Systems in Latin America*. Stanford University Press, Stanford, CA

Malloy J M, Seligson M A (eds.) 1987 *Authoritarians and Democrats: Regime Transition in Latin America*. University of Pittsburgh Press, Pittsburgh, PA

Martz J D 1990 Political science and Latin American studies. *Latin American Research Review* **25**(1): 67–86

O'Donnell G 1973 *Modernization and Bureaucratic–Authoritarianism: Studies in South American Politics*. Institute for International Studies, Berkeley, CA

O'Donnell G, Schmitter P C, Whitehead L (eds.) 1986 *Transitions from Authoritarian Rule*. Johns Hopkins University Press, Baltimore, MD

Przeworski A 1991 *Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America*. Cambridge University Press, New York

Schmitter P C 1971 *Interest Conflict and Political Change in Brazil*. Stanford University Press, Stanford, CA

Schmitter P C 1974 Still the century of corporatism? *Review of Politics* **36**(1): 85–131

Valenzuela A 1988 Political science and the study of Latin America. In: Mitchell C (ed.) *Changing Perspectives in Latin American Studies*. Stanford University Press, Stanford, CA

Wiarda H J (ed.) 1974 *Politics and Social Change in Latin America: The Distinct Tradition*. University of Massachusetts Press, Amherst, MA

P. W. Drake and L. Hilbink

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Latin American Studies: Religion

Prior to the 1960s, religion in Latin America was an understudied phenomenon. Most observers simply assumed that the Latin American Catholic church was a conservative institution that reinforced an antidemocratic, hierarchical, elite rule in what was generally viewed as a 'Catholic continent.' To the extent that the scholars studied the Latin American Catholic church at all, they tended to focus on formal church–state relations. The essay below reviews the major developments in Latin American Catholicism since the 1960s, as well as the rapid growth of Pentecostalism in the region, and notes the central debates among academic observers of Latin American religion.

1. *Catholicism*

Several developments in the 1960s signaled to scholars that the conventional wisdom regarding Latin American religion needed reassessment. Vatican II (1962–5) launched far-reaching reforms within the international Catholic church in an effort to reconnect the faithful to the immediate and worldly relevance of Catholic belief. In Latin America, many younger priests participated in new initiatives such as worker-priest movements that resulted in a heightened social and political awareness of poverty. By the 1968 meeting of the Latin American Bishops Conference (CELAM) meeting in Medellin, Colombia, the winds of a new progressivism were blowing through the Latin American Catholic church.

The CELAM meeting in Medellin was a turning point for the Latin American Catholic church. Using the model 'see, judge, act' from Vatican II, Latin American bishops called on social scientists to paint a picture of Latin American 'reality' in order to better assess how the church might renew its relevance to people's lives. The overwhelming image of Latin American society offered by Latin American economists, sociologists and political scientists was one of chronic and grinding poverty resulting from Latin America's economic dependence on the United States, and backed up by military repression and US imperialism.

At the same time, progressive bishops, inspired by the writings of Brazilian educator Paulo Freire (1970) and liberation theologians such as Gustavo Gutierrez (1973), embraced a radical agenda of social justice, political democracy and economic redistributive policies. Liberation Theology, like other historical Christian movements, is based on the Old and New Testament prophesies of the coming of the Kingdom of God, at which time all humanity will live in peace, justice and equality. According to liberationists, traditional Catholicism supports the status quo of oppression and dehumanization by calling on the poor to bear the travails of this life in expectation of eternal salvation, and exhorting the wealthy to appease their consciences by giving to charity. Drawing on Old Testament examples of God's willingness to reach into human history on behalf of His people (for example, the Exodus), liberationists call on the church to cast its lot against the 'institutionalized violence' of the status quo. From the New Testament, liberationists argue that Jesus chose to reveal his message to the poor and meek, linking Christian faith with solidarity with the poor and their temporal struggle for social justice. Unlike traditional Catholicism that relies on charity to the poor, liberationists call on Christians to embrace political reform, even the overthrow, if necessary, of oppressive and repressive regimes.

Influenced by the contributions of both academics and theologians, the Latin American church committed itself to a 'preferential option for the poor' at the Medellin CELAM meeting. The church also adopted a pastoral strategy centered on base Christian communities (or CEBs, the acronym for *comunidad eclesial de base*). CEBs are parish-level neighborhood groups in which clergy and/or pastoral agents attempt to raise social awareness through group discussions regarding the relevance of weekly Bible readings to members' own everyday lives (Azevedo 1987). According to many observers, the 1968 CELAM meetings in Medellin signaled a change within the Latin American Catholic church from the traditional alliance with the landed elite to an activist and progressive agenda (Lernoux 1982).

The first scholar to examine the implications of changes within the Catholic church was sociologist Ivan Vallier. Vallier's interest in the Latin American Catholic church was shaped by the concern for economic, political and social modernization that dominated US social sciences in the 1960s and 1970s. Unlike others who continued to view Catholicism as a hindrance to Latin American modernization, however, Vallier (1970) looked at different ways in which the Catholic church had affected historical development patterns across countries rather than the effects of a monolithic 'Catholicism.' Vallier's work opened the way for a number of new country-specific studies examining the historical role of national Catholic hierarchies in twentieth century political and social development (Mainwaring 1986, Smith 1982, Williams 1989).

During the 1970s and 1980s, the dominant focus in US scholarship shifted from the challenges of economic and political modernization to the crises associated with military dictatorship throughout Central and South America. Between the 1960s and 1980s, democracy had failed throughout Latin America. In a number of countries (notably excepting Argentina and Colombia), the Catholic church's commitment at Medellin (reaffirmed at the 1979 CELAM meeting in Puebla, Mexico) to serve as the 'voice of the voiceless' came to mean serving as a central conduit for

opposition to authoritarianism and military rule. Under the progressive leadership of a number of highly visible and vocal bishops, national hierarchies in countries such as Brazil, Chile, and Peru supported a wide range of grassroots organizations whose purposes ranged from ensuring the survival of the poor, to protecting human rights victims of the state, to organizing massive prodemocracy mobilizations. The prodemocracy leadership on behalf particularly of the poor and political dissidents provided by bishops, priests, nuns and lay leaders made the church, itself, a target of military repression. The assassination of El Salvador's Archbishop Oscar Romero, as well as the assassination and/or torture of thousands of priests, nuns and lay workers, were the price of church leadership during this period. Not surprisingly, the majority of studies during this period focused on church progressives and the host of new forms of social organization that grew under the protection and encouragement of progressive church structures (Levine 1992, Mainwaring and Wilde 1989).

Beginning in the 1980s, Latin American redemocratization allowed a reassessment of progressive trends in the Latin American Catholic church. On the one hand, the Vatican, under the leadership of Pope John Paul II and Cardinal Joseph Ratzinger (head of the Vatican's Congregation for the Doctrine of the Faith), has orchestrated a conservative reversal of much of the progressive political commitment made by Latin American hierarchies during the dictatorships. Pope John Paul II, a populist well-beloved by Latin America's poor and dispossessed, nevertheless is a staunch opponent of church political activism on behalf of the poor. Of particular concern to the Pope and Cardinal Ratzinger is liberation theology's use of Marxist categories of analysis for understanding Latin American conditions and liberationists' calls to concrete political action, even to the point of joining forces with Marxist movements in some cases. Cardinal Ratzinger has moved to silence several well-known Latin American theologians, including Brazilians Leonardo Boff and Ivone Gebara, a noted feminist theologian. In addition, the Vatican has filled vacancies for new bishops and archbishops with strong church traditionalists who have withdrawn much of the church's previous support for grassroots organizations and political organizing in favor of a traditional pastoral focus on individuals' private morality, charity and individual sin. In addition, many have noted a certain desire, even among many progressives, for the church to withdraw from explicitly political, even partisan, activities and positions in favor of the spiritual and moral work that is the proper work of the church. In the 1990s, many younger clergy are drawn to a more traditional view of the role of the church in society than their counterparts during the dictatorships (Cleary and Stewart-Gambino 1992). Although the previous focus on progressives created the illusion of a new consensus within Latin American hierarchies, the return of democracy allowed longstanding tensions among bishops to reemerge, dividing national hierarchies and thus reducing their visibility in public debates and controversies.

The renewed debate within national churches regarding the proper role of the Catholic church in contemporary society shifted scholarly focus away from changes within the institutional Catholic church and their effects on Latin American societies, to the varying experiences of the faithful themselves. Recent studies, often authored by anthropologists and sociologists, focus less on national-level institutional behavior and more on local groups and how their religious identities shape their interaction with larger economic and political forces. Individuals and local groups differ substantially in terms of the way and degree to which their religious beliefs are influenced by official church programs and pronouncements. In turn, individuals and local groups differ in terms of how their religious beliefs shape their own political behavior, depending on such factors as local opportunity, both material and symbolic resources, and the like. Gender also affects the intersection of religious belief and individual religious, social or political behavior (Mariz 1994, Burdick 1998).

Also in the 1990s, the fall of the Soviet bloc returned US scholars' attentions to perennial questions concerning the nature and requisites for democratic institution building. Like new post-Soviet states, many newly democratic Latin American nations lack a dense or complex civil society. Consistent with the recent focus on the neo-Toquevillian benefits of a densely organized civil society for democracy, some observers of Latin American religion have began to examine how religious life can foster democratic values and/or effective checks on the nondemocratic impulses of new regimes (Oxhorn 1995). With the decline of progressivism in the Latin American Catholic church, much of the search for religion's contribution to civil society is found in studies of the continued growth and expansion of Latin American Pentecostalism.

2. *Pentecostals*

Perhaps the most dramatic change in Latin American religions in the latter half of the twentieth century is the explosion of Pentecostalism in the region. Pentecostalism is a twentieth century phenomenon, the roots of which date to a prayer meeting of Bible school students in Topeka, Kansas on January 1, 1901. The group prayed for an extraordinary sign of the Holy Spirit, whereupon one young woman began to speak in tongues. The intense euphoria and physical manifestations of the 'baptism of the Holy Spirit' spread to the other students, leading them to believe that they had experienced the Pentecost, the immediate experience of the Holy Spirit described in Acts 2. Five years later, a Pentecostal mission was founded on Azusa

Street in Los Angeles, where thousands began to convert, beginning the rapid spread of Pentecostalism not only in the United States, but also around the globe. By 1916, Pentecostal churches were recorded in eight Latin American countries (Cleary and Stewart-Gambino 1997). Latin American Pentecostal churches, always known for a central commitment to proselytism, steadily attracted converts throughout the 1920s and 1930s through the combination of public preaching, spirited and emotional worship, and claims to supernatural healing power of the Holy Spirit. The Assemblies of God, one of the largest Latin American Pentecostal churches, grew at an average of 23 percent per year from 1934 to 1964. By 1962, Brazilian Pentecostals alone numbered over one million members. At the time of writing, non-Catholic Christians in Latin America are estimated to number between 60–80 million, representing between 13–17 percent of the total population. Of these, between 75–90 percent are believed to be Pentecostal. Pentecostal growth has exceeded population growth for decades, although Pentecostal membership remains concentrated among the poor, and to a lesser extent the middle classes.

As early as the 1960s, a few academic observers pointed to the dramatic growth in Pentecostal membership. Like observers of Latin American Catholicism in the 1960s and 1970s who focused on larger processes of modernization, early observers examined Pentecostalism in terms of its potential contribution to democracy. Christian Lalive d'Epinay (1969) argued that Pentecostalism replicated the authoritarian values and social organizations of rural society, thus attracting peasants relocated to mushrooming urban centers. According to this view, Pentecostalism reinforces non-democratic political forms and hinders modernization. On the other hand, Emilio Willems (1967) argued that the lack of hierarchy or formal theology, coupled with individuals' free access to the Holy Spirit, makes Pentecostalism an inherently democratic belief structure that attracts the poor as a symbolic protest against the traditional social order.

Debates about Pentecostalism's implications for larger social and political developments were highlighted again with the 1990 publication of two studies, by David Martin (1990) and David Stoll (1990). Like Lalive d'Epinay, Stoll's examination of Pentecostal growth in countries such as Guatemala, where the greatest growth occurred during the brutal dictatorships and civil war of the 1970 and 1980s, leads him to the conclusion that Pentecostalism reinforces political authoritarianism and repression. Moreover, Stoll argues that overt and covert support for Pentecostal missionaries from both US churches and government demonstrates the insidious imperialism at the core of Latin American Pentecostalism. On the other hand, reminiscent of Willems' more optimist interpretation, Martin stresses the egalitarian and participatory nature of Pentecostal congregations that can contribute, although indirectly, to strengthened democratic values and empowerment of the poor.

Pentecostal membership appears to be particularly powerful for the most vulnerable; for example, the addicted, ill, or other social outcasts. Yet, recent studies show that the stereotypes of Pentecostals as *simply* the poor or vulnerable are false. For example, while scholars agree that Pentecostalism attracts primarily the poor, some in the middle class are drawn to the doctrines of hard work, sobriety and clean living that bring God's spiritual and, especially, temporal rewards. Solid analysis is replacing other stereotypes as well. In the past, Pentecostals' tendency to eschew 'politics' was interpreted as an inherently conservative, even authoritarian bent; however, with the return to formal democracy throughout much of the region, Pentecostals have shown themselves to be as divergent in political persuasion as the population at large.

Over time, as churches strengthen and institutionalize, many Pentecostal pastors and congregations are demonstrating an increasing willingness to enter the political arena, in spite of a general distaste for 'temporal' commitments outside the missionary zeal common to all Pentecostal groups. The assumption that Pentecostal belief, with its emphasis on traditional family patterns with strong male leadership, damages the cause of women's rights also has been questioned. Pentecostal congregations typically do exhort women to submit to their husbands and dress in a plain, conservative manner; however, several observers find that women whose husbands convert feel that their lives have improved dramatically. Pentecostalism's model for male behavior is a sober, responsible and faithful husband and father. To the extent that Pentecostal men forsake drinking, womanizing and domestic violence often associated with traditional, machista cultural values, women find that their domestic incomes increase and their marital and family burdens lighten (Garrard-Burnett and Stoll 1993, Cleary and Stewart-Gambino 1997, Smith and Prokopy 1999).

3. *Other Religions*

Although the Catholic church still continues to dominate the Latin American religious landscape, the region has always housed a range of non-Catholic, non-Christian and/or syncretist religious forms, especially those of African descent in areas such as the Caribbean and Brazil. Among the better known are Brazilian *umbanda* and *candomble*, Afro-Caribbean *santeria*, Haitian voodoo, and others. Given the Catholic church's historical inability to provide even the most basic of pastoral services adequately to the majority of urban and rural poor, Latin American popular religious forms always have flourished at the margins of institutional religion. So-called 'popular religion,' or religious forms that lie outside of official,

institutionalized religion, are expressions of the desire of common people to find and recreate religious meaning in their everyday lives. Popular religion arises out of people's own oral traditions, ancestral roots (native American, African, Iberian or other), class identities, community memories, and individual and collective creativity. Not surprisingly, a vast range of forms of popular religion endures, thrives, and mutates over time in Latin America (Burdick 1998).

The persistence and even growth of different forms of Latin American popular religion, in addition to the extraordinary rise of Pentecostalism in the region, has diminished the once hegemonic position of the Catholic church. There is, now in the twenty-first century, a far greater degree of religious pluralism in Latin America than ever before. Religious pluralism, in turn, will cause further changes within Catholicism and Pentecostalism as institutional churches struggle to maintain their memberships and appeal to new converts in the context of a broadened Latin American religious marketplace.

4. Conclusion

Future research on Latin American religious institutions and belief likely will head in several directions. The apparent retreat of many national Catholic hierarchies from political activism in the wake of recent transitions to democracy masks the fact that bishops throughout the region are trying to clarify the proper role of the church in changed national circumstances. How national Catholic churches will contribute to or undermine Latin American democracy remains a live question. Tensions between bishops' view of the proper role of the church and grassroots religious experience further complicate church–society relations, as do issues of gender, class and race. Examinations of both national Catholic churches and local-level case studies will contribute to our understanding of the evolving role of the Catholic church in Latin America. At the same time, the continued growth of Pentecostal and other non-Catholic religions will strengthen Latin American pluralism, to which all religious institutions will have to respond. Not only will Catholic hierarchies have to adapt to a more competitive religious 'marketplace,' but differing strains in Pentecostalism also will become clearer as some groups strengthen and institutionalize, while continuing schism fragments others.

We need to know a great deal more about individual Pentecostal churches, however, before adequate national or theological generalizations can be made. In addition, comparatively little is known about a wide range of popular religious forms found throughout Latin America. Beyond institutionalized religion and religious elites' decisions, we need to know more regarding people's lived religious lives and how their religious belief intersects with the workplace, family, and community. Moreover, how do changes in national economic strategies that profoundly affect people's daily lives intersect with individual religious belief? It is often said that religious belief and sentiment are widely held and deeply rooted in Latin American cultures. What this means in changing political, economic and cultural conditions will help shape future trends in Latin America in ways that we do not yet fully understand.

See also: Catholicism; Christianity: Evangelical, Revivalist, and Pentecostal; Latin American Studies: Politics; Latin American Studies: Society; Religion: Mobilization and Power; Religion: Nationalism and Identity

Bibliography

Azevedo M 1987 *Basic Ecclesial Communities in Brazil*. Georgetown University Press, Washington, DC

Burdick J 1998 *Blessed Anastacia: Women, Race, and Popular Christianity in Brazil*. Routledge, New York

Cleary E L, Stewart-Gambino H (eds.) 1992 *Conflict and Competition: Latin American Catholicism in a Changing Environment*. Lynne Reinner, Boulder, CO

Cleary E L, Stewart-Gambino H (eds.) 1997 *Power, Politics and Pentecostals in Latin America*. Westview Press, Boulder, CO

Freire P 1970 *Pedagogy of the Oppressed*. Continuum, New York

Gutierrez G 1973 *A Theology of Liberation*. Orbis Books, Maryknoll, NY

Eagleson J, Scharper P (eds.) 1979 *Puebla and Beyond*. Orbis Books, Maryknoll, NY

Garrard-Burnett V, Stoll D (eds.) 1993 *Rethinking Protestantism in Latin America*. Temple University Press, Philadelphia, PA

Lalive d'Epinay C 1969 *Haven to the Masses: A Study of The Pentecostal Movement in Chile*. Lutterworth Press, London

Lernoux P 1982 *Cry of the People: The Struggle for Human Rights in Latin America—The Catholic Church in Conflict with U.S. Policy*. Penguin Books, New York

Levine D H 1992 *Popular Voices in Latin American Catholicism*. Princeton University Press, Princeton, NJ

Mainwaring S 1986 *The Catholic Church and Politics in Brazil: 1916–1985*. Stanford University Press, Stanford, CA

Mainwaring S, Wilde A (eds.) 1989 *The Progressive Church in Latin America*. University of Notre Dame Press, Notre Dame, IN

Mariz C 1994 *Coping with Poverty*. Temple University Press, Philadelphia, PA

Martin D 1990 *Tongues of Fire: The Explosion of Protestantism in Latin America*. Basil Blackwell, Oxford, UK

Oxhorn P D 1995 *Organizing Civil Society: The Popular Sectores and the Struggle for Democracy in Chile*. Pennsylvania State University Press, University Park, PA

Smith B H 1982 *The Church and Politics in Chile: Challenges to Modern Catholicism*. Princeton University Press, Princeton, NJ

Smith C, Prokopy J (eds.) 1999 *Latin American Religion in Motion*. Routledge, New York

Stoll D 1990 *Is Latin America Turning Protestant? The Politics of Evangelical Growth*. University of California Press, Berkeley, CA

Vallier I 1970 *Catholicism, Social Control, and Modernization in Latin America*. Prentice-Hall, Englewood Cliffs, NJ
Willems E 1967 *Followers of the New Faith: Culture, Change, and the Rise of Protestantism in Brazil and Chile*. Vanderbilt University Press, Nashville, TN
Williams P A 1989 *The Catholic Church and Politics in Nicaragua and Costa Rica*. University of Pittsburgh Press, Pittsburgh, PA

H. W. Stewart-Gambino

Latin American Studies: Society

1. *The West and Latin America: Similar Yet Distinct*

Latin America refers broadly to the whole of the American Continent south of the United States. It owes its name to the Latin root of the two dominant colonial languages spoken: Spanish and Portuguese. Colonized by the Portuguese and the Spanish, most of the region fought for and won independence by the mid-nineteenth century. Clearly set apart from Europe and the English white settler colonies that later became part of the group of developed countries, Latin America has stood for most of its history at the middle of the international system of stratification and (with the exception of Cuba) squarely in the capitalist route of development. While highly heterogeneous both in the past and the present, Latin America remains a laboratory for innovative policy creation and, as a region, it has also constituted a unique case for groundbreaking comparative studies. While most of Africa came to independence only in the twentieth century, and Asia chose different development routes well into the twentieth century, what is peculiar about Latin America is that very early on it embraced Western goals of development and also bore the closest resemblance to the countries that were the basis of Western development in Europe and the USA. Yet neither the will nor the tools of Latin American development were sufficient and the Western ideal has never been fully achieved.

Economically, Latin America chose capitalism and sought growth and industrialization; politically, though less consistently, it sought liberal republics and later mass democracy; and, socially, it believed in urban and modern societies. To put it more strongly, Latin American elites, and thinkers that discussed the Latin American case, believed that this region would one day look very much like the industrialized West, and that it would achieve that goal by adopting programs based on European history (retold in a stylized fashion). Yet in all of these endeavors, Latin America has proven that the notion of a continuous, linear, and consistent development route is largely wishful thinking. While other authors drawing on the European experience had challenged early on this linear evolutionary model, no region has, at the same time, both embraced and challenged this belief. Thus the fertility of Latin America, both for policy innovations and for cutting-edge research and theory is a function of the power that comes from differences among similars. The promise, realizations, and failures of Latin American development constitute a fascinating story for the social sciences and have consistently been food for thought and building stones of major academic debates in fields as diverse as economic theory, political science, and sociology.

2. *Latin American Development: Reality, Ideas, and Theories*

2.1 *The Colonial Past*

Officially Europeans arrived in what today is Latin America in 1492. During the sixteenth and seventeenth centuries, the Spanish, the Portuguese, the English, and the Dutch fought and settled in the Caribbean, Central America, and South America. The Spanish and the Portuguese Empires came to control all of South America and parts of the Caribbean and Central and North America. The Portuguese settled mostly in the Caribbean and especially in the far western part of South America and expanded eastwards, gaining dominions formerly belonging to the Spanish Empire until they controlled almost half of South America. Yet occupied with their expansion over the South Atlantic and monsoon Asia, the Portuguese were slow to consolidate their rule over these newly colonized lands. Spanish colonists, by contrast, settled into the Americas and expanded rapidly, battling the native population and creating military outposts and cities. By 1521 the conquest of Mexico and the Aztec Empire was completed with the fall of Tenochtitlan. Cuzco, center of the Inca Empire, was conquered in 1533, and Santiago de Chile was founded as early as the 1540s. The Spanish organized the new territories in the form of viceroyalties, resembling the monarchical structures of Spain. Cities were founded across the continent and organized with church, military, and *cabildos* (town council). At first gold and later silver were the major prizes that the Spanish sought in the New World. Extraction was controlled by the Spanish and performed by enslaved natives. Soon famine, poor working conditions, and European illnesses (especially smallpox and measles) decimated the native population.

Silver, gold, and, to a much lesser extent, tobacco and sugar became the central economic activities of the major cities and regions of South America in the sixteenth and seventeenth centuries. In the eighteenth and early nineteenth centuries, the expansion of the sugar economies, together with cacao, tobacco, salt, and later cattle (coffee as a staple came later in the

nineteenth century) combined with the slave trade to enable the fully-fledged entry of the region into the world economy. These developments also shaped a more complex economic and social system in which private enterprise and empire, local and international economic elites, collided, bargained, and cooperated. The 'hacienda' and the 'plantations' in which native and African labor was coerced became central ingredients of the South American and Caribbean economies and left an indelible mark on the region's elites and its people's power, status, and ideologies.

Towards the end of the eighteenth century, local elites developed both separate interests from those of the Empire and political and philosophical views that increasingly saw in Europe a limit to their growth and development. Napoleon's invasion of the Iberian Peninsula in 1808 was the spark needed for the ignition of a continent-wide movement for independence from Spanish dominion. Portugal, in turn, moved the center of its empire from Lisbon to Rio de Janeiro, and eventually, in 1822, the heir to the Portuguese throne, Pedro I, became Emperor of the independent Empire of Brazil. The colonial past of more than three centuries gave way to a new continent of independent nations. Yet for many comparativists, Latin Americanists, and historians of the region and the world, the colonial past was to be the key to understanding Latin America's future.

In addition to the socioeconomic legacies that mining, haciendas, and plantations left in Latin America, a centralized political organization permeated by patrimonialism and the hegemony of the counter-reform movement (that had been defeated in Europe but triumphed in Spain and Latin America) added to Latin America its Iberian, Catholic, and anti-positivist flavor. Historiography and later comparative studies from political science, sociology, and economics have sought to explain the roots of Latin American backwardness and the obstacles to the success of Western-type development in terms of its socioeconomic, political, and cultural legacies. The failure of liberal constitutionalism and capitalism in Latin America has, indeed, been accounted for by a black legend that places the blame squarely on the region's colonial heritage (for an exposition of this view, see Paz 1961, Stein and Stein 1970; it is reviewed and critiqued in Gibson 1971; see also Adelman 2000 for a review and discussion of this debate).

2.2 The Nineteenth Century: Independence, Civil War, Staples, and the Making of Latin America

While, strictly speaking, independence from Spanish and Portuguese dominion (with short interludes of French occupation in Mexico and English attempts in the River Plate area) ranged from Haiti's black revolution of 1804 to Panama's independence in 1903, the bulk of the independence process took place in a short period between 1810 and the Battle of Ayacucho in 1824. Most of the nineteenth century was dominated first by wars of independence and later by civil wars, wars among new nations, 'caudillo' struggles, and in a few cases the conquest of the frontiers (as in the Argentinean war of the desert). Yet an account of the nineteenth century that simply stresses the armed struggles and a sense of anarchy in the region misses the point. These were times of major intellectual debates that would have a long-lasting influence on Latin American scholarly debates and studies, and these were also times in which the oligarchic regimes and the export-oriented model of development took root in Latin America, only to be overcome in the first half of the twentieth century.

Cattle in Argentina and Uruguay; coffee, cacao, sugar, and tobacco in Brazil and Central America; minerals, coffee, salt, and fruits along the Pacific coast—such is a rough geography of the staples produced in Latin America throughout the nineteenth century. While nonlabor-intensive staples were required in the Rio de la Plata, in most of Latin America, production of staples was labor-intensive and based on haciendas and plantations. Coerced or semi-coerced labor dominated the latter, while forms of free or market-regulated labor dominated the former. Scholars that draw on the tradition of comparative work spurred by Barrington Moore came to view this original distinction between labor-intensive and non-labor-intensive regimes as a key to understanding democratic experiments (Stephens 1989, Rueschmeyer et al. 1992). On the economic side of comparative studies, the importance of staples and their forms of production are highlighted in Hirschman's unsurpassed work on development, and especially in his notions of forward, backward, fiscal, and consumption linkages. He proposes a simple yet powerful idea for understanding the nineteenth and twentieth centuries in developing countries. Staples create, or have the potential to create, forward and backward productive linkages as well as tax possibilities and consumption expansion. Depending on the staple and on the intensity, extensivity, and control (foreign or domestic) of those linkages, development will be more robust and balanced or less so (Hirschman 1958, 1973, 1981).

Regarding the politics of Latin America, this was a period of upheaval and intense ideological debates. Liberals and conservatives, landed elites and incipient modern capitalists, and disputes over territory or between warlords all combined to create a violent political and ideological melting pot. In particular, on the ideological side of the debate. a continent-wide movement of intellectuals was calling for Latin America's second independence (or what they called mental independence). Sarmiento and Alberdi in Argentina, Caballero in Cuba, Varela in Uruguay, Altamirano in México, to name a few, presented positivism and an unabashed faith in public primary education as the

tools for mental emancipation and political and economic development.

Out of the nineteenth century a continent evolved that was scarred with civil wars, squarely placed in the world economy as a producer of staples, oligarchic in nature, Iberian in flavor, but also carrying the seeds of positivism and liberalism, and the recognition of the Anglo-Saxon giant of the north as both a threat and an ideal.

2.3 Searching for Development in the Twentieth Century: Promises, Achievements, and Failures

As the period of civil wars came to a close, the future of Latin America was laid open to the world, and the region (or so its elites believed) became ready to embrace the path and gather the fruits of Western progress. Capitalism, republican order, and modern cities and citizens would fulfill Sarmiento's dream of moving from 'barbarie' to civilization. Reality proved more stubborn and, in reflecting on its failures and obstacles, Latin American scholars, comparativists from all over the world, and students of development found in the region a case study that supported, challenged, modified, or even created paradigms for understanding society, politics, and economic development.

2.3.1 Economic promise and failure: adapting, creating and recreating development economics. In the early twentieth century the per capita income of the working population in Argentina was similar to that in the UK; in Uruguay to France; in Chile to Norway; in Brazil to Italy; and in Mexico to Portugal, Finland, or Greece (quoted in Halperin 1997). Growth in gross domestic product (GDP) was higher in most of Latin America than it was in Europe, and the value of its exports per capita, higher than or similar to most of Europe. This, of course, obscures the structure of their respective economies with respect to industrialization, social development, and types of exports. Yet, these data provide a basis for understanding the widespread belief that an outward-oriented model of development based on primary exports, import of manufactures, free commerce, and little involvement by the state was the right path, and why classic economics with its Ricardian beliefs in comparative advantages was the dominant economic paradigm.

Still, if the comparison shifts from Europe to the United States the rate of growth and the riches accumulated in Latin America appear less than satisfactory. While there are different estimates, the USA had in the 1930s about four to six times Latin America's GDP per capita and had grown since independence at an average rate of 2 percent per annum while Latin America had grown at no more than 1 percent (Ramos 1993) Furthermore, while the data for Europe and Latin America suggest similar starting points, the points of arrival were starkly different. In the 1950s, Prebisch showed that, contrary to the hypothesis of convergence that classic and then neoclassic economic thought championed, Latin America had grown further apart from rather than closer to the developed countries during the period of its first orthodox model from early independence to roughly the 1930s.

When World War One dried up the flow of imports to Latin America, its economic elites moved, albeit shyly, to manufacturing and substituting some of the imports. But only with the world economic crisis of 1929 did the region make a drastic U-turn in its economic model. Import substitution was fact before it was theory, yet it eventually gave rise to the first homegrown theory of economic development that would have a lasting impact on the region and on comparative studies in the social sciences.

In 1948 the Economic Commission for Latin America (*Comisión Económica para América Latina y el Caribe* or CEPAL) was created and it brought together a number of innovative economists led by the already famous Raul Prebisch(for a synthesis of his work and a founding document of CEPAL, see Prebisch 1950). In Montecinos and Markoff's (2001) view, CEPAL's theoretical framework sought the ambitious project of establishing a new school of economic thought that would offer the basis for specific development policies for non-industrialized countries. Departing from the neoclassical economic vision of a single economic truth for all countries and criticizing some assumptions from Keynesianism (although also drawing ideas from it), CEPAL and Prebisch put forward the idea that core and periphery countries, advanced and developing countries, should seek different economic and development policies. The Ricardian comparative advantage did not benefit Latin America because agricultural products were subject to declining terms of trade. Thus the need to industrialize and to protect infant industries from outside competition was crucial in order to advance in terms of economic development. Active state involvement through industrial subsidies, tariff protection, and planning became central strategic devices for development. As CEPAL and 'cepalino' thought became hegemonic in the region, most countries adopted a model of development that has been summarized as Import Substitution Industrialization (ISI) and embraced an ideology that may be loosely defined as developmentalism (or '*desarrollismo*'). CEPAL argued that, through state protection and active planning, Latin America would overcome traditional agrarian relations and productivity and claim its place among modern industrial societies.

Yet by the 1960s what was called the easy phase of import substitution was over. Light manufactures had been substituted, yet the production of capital goods

and heavy machinery did not come about. Growth stagnated, inflation soared, and fiscal deficits became increasingly unmanageable. The most famous Latin American offspring for the social sciences was born out of the crises of import substitution and became known as dependency theory or '*dependentismo*' (Cardozo and Faletto 1979; see also the works of André Gunder Frank 1970, Theotonio Dos Santos 1970, and James Petros 1970). While *dependentismo* owed a great debt to CEPAL's idea of core and periphery and to its notion of declining terms of trade, it went far beyond these ideas. For dependency theorists Latin American underdevelopment was not simply a problem of core–periphery relations, but of the interaction of domestic and international capital and the internal class structure in the region. The monopolistic character of multinationals, the weakness of Latin America's domestic bourgeoisie, the analyses of enclave economies in which domestic resources were exploited by outward-looking elites and international capital, and the internationalization of the domestic markets were factors that helped explain a peculiar development, one that was unique to the developing nations and that rested on a domestic class structure of large rentist capitalists and a weak industrial bourgeoisie.

A group of scholars and universities in the US and in Latin America, while less visible, was also rethinking development in radical ways. In Chicago, in Mexico's *Instituto Tecnológica Autónoma de México* (ITAM), and in the Catholic University of Chile, a new brand of orthodoxy was spreading its wings: neoclassical economic thought. In radical disagreement with *dependentismo*, this group did not believe that CEPAL's shortcomings lay in its shyness to push forward state led development. CEPAL and ISI were a problem, but they had been a problem from the start. The moment Latin America had embraced state intervention, planning, and protection was the moment of the region's mistake. While dependency saw the need to deepen state involvement and to destroy agrarian elites, neoclassical thought saw ISI as the brake on Latin American development because of state intervention and because it punished those economic sectors that had a real comparative advantage: landowners who grew and sold staples.

If the Peru of Velazco Alvarado and the Chile of Allende embodied the radicalization of ISI and dependency theory, the coups d'etat of the 1970s in the southern cone of Latin America opened the door to yet another, and maybe the most radical, experiment in social and economic change. Chile, in particular, came to be seen as the first laboratory in the world for neoclassical economic revival and for the final demise of both ISI and Keynesianism. Through tariff bashing, massive privatization, labor market deregulation, and liberalization of financial markets Chile became, before Thatcherism and Reaganism, the darling of the Chicago Boys (a group of Chilean economists who had studied at the University of Chicago between about 1955 and 1963). The end of ISI and the defeat of the dependency agenda gave rise to a third phase in economic development: neoliberalism or what has more neutrally been called the New Export Model (NEM). While the jury is still out regarding the merits and shortcomings of this second drastic development U-turn, the experiments in Latin America since the mid-1970s have fed economic and social debates and these scholarly debates have contributed to a rethinking of development and developmentalism, and comparative studies in general.

2.3.2 The search for democracy and the lost republic. Orderly republics and democratic politics was another promise that Latin America sought to capture as it entered the twentieth century. Yet democracy and order have proven elusive in its twentieth-century history. While Argentina (1912) and Uruguay (1917) embraced mass democracy early, most other countries in the region experimented with oligarchic, military, and populist regimes with short and unstable periods of democracy throughout most of the twentieth century. Even Argentina, the most advanced and most 'European' nation reverted to authoritarianism in the 1930s, not regaining stable democracy until the 1980s. Chile, Uruguay, and Costa Rica (the last one especially after the 1940s) were the only Latin American countries with a reasonable, but by no means perfect, democratic record. A number of authors and scholars drew on the region to unveil the clues to democratic success and failure. Some of these scholars were specifically studying Latin America, others had larger samples, but almost all gave a prominent space to Latin America as a particularly promising puzzle.

Modernization theory was, as we shall see later, a dominant paradigm in regard to social change and development in Latin America, but its political version was also of importance and Seymour Lipset one of its major proponents. In his classic article, a simple yet powerful proposition was laid out: it was not simply capitalism that led to democracy, but also growth and social development (Lipset 1959). Following this hypothesis, Latin America was not democratic, because it was not rich. As economies developed, so democracy would flourish. By increasing the riches, creating a middle class and expanding education, countries that successfully traveled towards industrialized and modern economies would also increasingly become democratic. From another perspective, Robert Dahl and Samuel Huntington also proposed a hypothesis that would help explain Latin America's democratic failures: the lack of order (Huntington 1957, 1968, 1991, Dahl 1971). Dahl's strongest claim implied that countries that incorporated the masses before institutionalizing rules of the game would have frail democracies. For Huntington, praetorian

societies in which no order could be constructed together with the dislocation that rapid economic and social change brought about helped explain the role of the military as a constructor of order and institutions and as a major player in Latin American politics. Neither Lipset nor Huntington nor Dahl constructed their theories for Latin America. And, in contrast to the homegrown economics of Latin America, the region only produced one truly unique body of work that not only grew from the specific study of the region but that also questioned the former theories.

Guillermo O'Donnell (1979) asked why the richest nation in Latin America was under authoritarian rule in the 1960s and 1970s. Lipset could not answer that, nor could Dahl (as a matter of fact, in Dahl's classic work, he offers an *ad hoc* explanation for Argentina). In O'Donnell's view, a new form of authoritarianism was emerging in Latin America—one that grew out of the tensions and political bottlenecks that Import Substitution Industrialization had created—he called the governments that exemplified it, bureaucratic-authoritarian regimes. These regimes were meant to neutralize the demands of the middle and working classes in order to deepen Import Substitution. In contrast to Lipset, O'Donnell suggested that in rich economies, dictatorship was indeed a possibility. Even more, when ISI met dependency, authoritarianism was the natural outcome. The spread of authoritarian regimes in the 1970s suggests that O'Donnell was not far off the mark, even though these new authoritarian regimes would eventually bury, rather than deepen, import substitution (this was the case in Chile, Uruguay, and the last dictatorship in Argentina, but not in Brazil and Argentina in the 1960s).

In the 1980s and 1990s the region returned to democracy and a series of studies looking at the key to successful transitions from authoritarianism to democracy was to become a whole new field of study for comparativists. These studies renounced structural and long-term explanations of democracy and focused on the concrete choices and strategic devices that in short periods of time were used to move countries from dictatorship to open competitive regimes. The name adopted for this new field, which had as its major input the study of Latin American societies and of Southern Europe, is 'transitology.' Its major contribution to political science was the recognition of contingency and strategic action as critical factors to understand large political outcomes.

As democracies have endured in Latin America since the mid-1980s, a new body of literature has emerged regarding the issues of democratic quality. To the old notions of order and governability once championed by Huntington, this literature adds the problems of accountability and rule of the law. In addition, and as never before, the debates over party systems and executive-legislative relations have become a normal field of study for Latin Americanists. At the same time, Latin America has become a relevant region in these debates from which it was previously absent.

Finally, and after 'transitology,' a return to macro-narratives and structural interpretations of Latin American democracies has taken place. Collier and Collier in *Shaping the Political Arena* (1991) and Rueschemeyer et al. in *Capitalist Development and Democracy* (1992), return to the historical questions of democratic rule in Latin America's history. The former concentrate on the region and introduce the idea of critical historical junctures attempting to understand the stability and openness of regimes as a product of the forms and shapes of popular incorporation. The latter revisit Barrington Moore's classic work on *Democracy and Dictatorship* and construct probably the best neo-Marxist account of democracy and class struggle in the world and in Latin America to date.

2.3.3 Inequality, dual societies, poverty, and uneven modernity. According to Kuznets (1959) and his famous curve, countries increased inequality as they took off economically, moderated inequality as they entered a second industrial phase, and lowered it as industrialization and economic development further advanced. Latin America has in this area the dubious merit of being the most unequal region in the world. Indeed, the region constitutes for most scholars debating the issues of growth and inequality, a category apart: the bad Latinos, who increased inequality with economic take off but then never lowered it, and in many cases further increased it.

Besides the fact that inequality is extremely high in Latin America, other aspects of modern social development that were expected did not take place in the region or occurred in very uneven fashion. Modernization theory, with its faith in linear evolution and continuous development, envisaged countries moving from backward agrarian societies to modern industrial ones with all that this implied: urbanization; demographic transition; more complex social structures with dominant middle classes; and societies that, while stratified, had some sort of continuity and mobility chances for all the population. As the country modernized, the people would achieve and secure modern statuses and roles, moving from peasant to worker, from rural dweller to urban citizen, form caste to class, and very especially from uneducated to educated. Parsons' pattern variables characterized this fate in development. The adscriptive would give way to the elective, the particular to the universal, the affective to the neutral, the general purpose institutions to the specific purpose institutions, and so on.

A tiny part of Latin America looks, at first sight, precisely as modernization theory would expect, but Latin American societies remain largely 'dual,' combining aspects and regions of modernity with others of blatant backwardness. Incorporation into a modern

industrial society and its rules has occurred for less than half of its population. Truly, only the ISI model implied a major push towards de-segmentation, and then only in some countries. Latin America is thus unequal, dual and segmented and, for many, modernity is a curse rather than a possibility.

Among modernization theorists more sophisticated versions were developed in Latin America in order to cope and confront this reality. Gino Germani and Peter Heintz writing in the 1960s are perhaps the best exponents of this critical yet loyal twist on classic modernization theory (Heintz 1971, Germani 1962, 1971). Germani represents the most loyal representative of modernization theory and of structural functionalism as the bedrock of this theory. In his work, Germani recognizes the unevenness and the incomplete nature of social modernization in Latin America. Yet, in his opinion, this is due mainly to the asynchronicity with which the region has moved in different dimensions from traditional society to modern society. Critical to his understanding of Latin America's shortcomings is the limited capacity of the economy to incorporate migrants fully into modern labor markets and the political tension that such a reality creates. As labor, education, urban status, political citizenship, and modern consumption patterns converge, full and even incorporation of the population will take place. For Heintz the problem is less simple and has a tougher structural basis. As the upper classes or elites allow and even promote modernization, Heintz claims, they also endanger their own basis of power and privilege. Thus, elites are willing to expand and incorporate the masses only into certain areas of modernity: urban statuses and, later, education are the least threatening. Yet they are not willing to let go of their political monopoly and their economic basis of power, thus closing incorporation into modern labor markets and consumption patterns and into the state and the political system to a large part of the population. For Heintz this unbalanced development has no single solution and the point of arrival for Latin America need not be the Western ideal that modernization championed.

As import substitution and its promise of industrial modernity entered its crisis and eventually came to a close, social development was rethought from two distinct perspectives: dependency and neoclassic economics. For the former, social and national revolutions were the only route to egalitarian and robust social development; for the latter, a retreat from the state, and a return to free markets and open economies (that relied yet did not punish those sectors with comparative advantages) was the basis for social development. The late 1980s and the 1990s were under the hegemony of neoliberal faith, and a locomotive version of development in which market-led growth was the key became uncontested. Yet in the late 1990s the previous work of CEPAL and its notion of egalitarian productive transformation (*transformación productiva con equidad*), and the United Nations Development Program with its idea of Human Development reclaimed an autonomous space for the problems of poverty and inequality, and have forced scholars to rethink development as far more complex game than that of economic development (CEPAL 1990; see also the annual *Human Development Reports* published since 1990). The recent scholarly work coming from the IADB, American universities and Latin American scholars have contributed to reshape development and comparative studies all over the world (Birdsall et al. 1998).

3. *In Closing*

Latin America has remained a laboratory for the social sciences throughout its history. This is due both to the richness of its experiences and to its common traits. The Developmentalist experiment in Chile with Frei, the Socialist experiment of Allende, and the Neoliberal experiment of Pinochet are just a tiny part of the mosaic that includes the democratic exceptionality of Costa Rica and Uruguay; the reversal of development in Argentina; the resilience of democracy in violence-ridden Colombia; the oil promise and failure of Venezuela; the uniqueness of development and origins in Brazil; the Alvarado experiment in Peru; the Mexican Revolution and its unique legacy of a non-socialist one-party system; and the revolutionary guerrillas of Central America. There is thus more than enough material for comparativists searching for deviant and test case scenarios, and much more for those searching for the exception that confirms the rule. On top of that, Latin America remains a unit. A region that chose Western ideals and capitalism (again with the exception of Cuba, an interesting case in its own right) and that has always come short in its accomplishment, will always be a puzzle to solve, given our still faithful, yet neurotic dependence on a simplified version of Western development.

Two excellent recent books that adopt a similar metaphor illustrate the point. Sandra Halperin's *In the Mirror of the Third World: Capitalist Development in Modern Europe* (1997) claims that what is wrong with development theory is precisely its attempt to understand a different route of development for the Third World. Europe, she claims, was very much like the oligarchic, rentist, authoritarian Third World until the end of World War Two. While it is true that she makes her claim of similarity for all nonindustrialized countries, her major examples of the Third World come from Latin America. The region is again the test case scenario for evolutionary models of development. On the other hand, *The Other Mirror*, edited by M. A. Centeno and F. López-Alvez (2000), attempts to trace the usefulness of grand theory for Latin America. While they claim that Latin America has been absent from the grand theory debates, they recognize that

Africa and Asia have been even more absent. As one reads the book, and all its superb chapters, one is slowly convinced that their premise of Latin American absenteeism from major debates is blatantly wrong. Latin America has been at the center of comparative and development debates throughout most of the second half of the twentieth century. All the grand authors debated have had a lasting influence on comparativists and scholars that later left their mark in their different fields through the study or the privileged attention of Latin America. Some of those grand authors have, indeed, themselves, drawn on the Latin American experience of development.

See also: Authoritarianism; Dependency Theory; Latin American Studies: Economics; Latin American Studies: Politics; Latin American Studies: Religion; Modernization, Political: Development of the Concept; Modernization, Sociological Theories of

Bibliography

Adelman J 2000 Institutions, property, and economic development in Latin America. In: Centeno M A, López-Alvez F (eds.) *The Other Mirror, Grand Theory through the Lens of Latin America*. Princeton University Press, Princeton, NJ, Chap. 1

Birdsall N, Graham C, Sabot R (eds.) 1998 *Beyond Tradeoffs: Market Reforms and Equitable Growth in Latin America*. Brookings, Washington, DC

Cardozo F H, Faletto E 1979 *Dependency and Development in Latin America*. [first published in Spanish in 1969] University of California Press, Berkeley, CA

Centeno M A, López-Alvez F (eds.) 2000 *The Other Mirror, Grand Theory through the Lens of Latin America*. Princeton University Press, Princeton, NJ

CEPAL 1990 *Transformación Productiva con Equidad*. CEPAL, Santiago de Chile

Collier R B, Collier D 1991 *Shaping the Political Arena: Critical Junctures, the Labor Movement and Regime Dynamics in Latin America*. Princeton University Press, Princeton, NJ

Dahl 1971 *Polyarchy: Participation and Opposition*. Yale University Press, New Haven, CT

Dos Santos T 1970 The structure of dependence. *American Economic Review* **60**: 235–46

Frank A G 1970 The development of underdevelopment. In: Rodhes R I (ed.) *Imperialism and Underdevelopment*. Monthly Review Press, New York

Germani G 1962 *Política y Sociedad en una Sociedad en Transición*. Editorial Paidós, Buenos Aires, Argentina

Germani G 1971 *Sociología de la Modernización*. Editorial Paidós, Buenos Aires, Argentina

Gibson C 1971 *The Black Legend: Anti Spanish Attitudes in the Old World and the New World*. Knopff, New York

Halperin S 1997 *In the Mirror of the Third World: Capitalist Development in Modern Europe*. Cornell University Press, Ithaca, NY

Heintz P 1971 *Un paradigma Sociológico del Desarrollo*. Instituto Di Tella, Buenos Aires, Argentina

Hirschman A O 1958 *The Strategy of Economic Development*. Yale University Press, New Haven, CT

Hirschman A O 1973 *A Bias for Hope: Essays on Development and Latin America*. Yale University Press, New Haven, CT

Hirschman A O 1981 *Essays in Trespassing: from Economics to Politics and Beyond*. Cambridge University Press, New York

Huntington S 1957 *The Soldier and the State: The Theory and Politics of Civil-Military Relationships*. Belknap Press, Cambridge, MA

Huntington S 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, CT

Huntington S 1991 *The Third Wave: Democratization in the Late Twentieth Century*. University of Oklahoma Press, Norman, OK

Kuznets S 1959 Economic growth and income inequality. *American Economic Review* **45**: 1–28

Lipset S M 1959 Some social requisites of democracy. *American Political Science Review* **53**: 69–105

Montecinos V, Markoff J 2001 From the power of economic ideas to the power of economists. In: Centeno M A, Lopez-Alvez F (eds.) *The Other Mirror. Grand Theory Through the Lens of Latin America*. pp. 105–50

O'Donnell G 1979 *Modernization and Bureaucratic Authoritarianism*. Institute of International Studies, University of California, Berkeley, CA

Paz O 1961 *Labyrinth of Solitude*. Grove, New York

Petras J 1970 *Politics and Social Forces in Chilean Development*. University of California Press, Berkeley, CA

Prebisch R 1950 *The Economic Development of Latin America and its Principal Problems*. United Nations, New York

Ramos J 1993 Growth, crises and strategic turn-arounds. CEPAL Review no. 68

Rueschemeyer D, Stephens J, Stephens E H 1992 *Capitalist Development and Democracy*. The University of Chicago Press, Chicago

Stein S, Stein B 1970 *The Colonial Heritage of Latin America: Economic Dependency in Historical Perspective*. Oxford University Press, New York

Stephens E H 1989 Capitalist development and democracy in South America. *Politics and Society* **17**: 282–352

F. Filgueira

Law and Aging

A notable demographic characteristic of the twentieth century, particularly in industrialized nations, has been the increase in the number and percentage of persons aged 65 and older. In the near future, 20 percent or more of the populations of Japan, the United States, and Western Europe will be aged 65 or older. This singular shift in the age of populations produces profound societal changes and presents significant questions about appropriate benefit programs for the elderly (Cockerham 1997).

Aging presents a number of challenges for law and policymaking. The elderly inevitably experience a decline in their physical abilities, which poses problems in employment and housing. Many older people suffer from chronic physical conditions that will limit their ability to care for themselves. Some suffer a significant loss of mental capacity due to dementia and other afflictions of the brain (Cox 1998). In part because of declining physical skills or the onset of a mental

condition, most elderly are retired. Even those who continue to work often experience diminished incomes because they are less productive or discriminated against in the work force. Surviving spouses, in particular women, often have reduced incomes either because their deceased spouse was employed at the time of death or because the couple's pension either ceased or was reduced at the death of the retired spouse. Because of these physical and mental changes, the elderly have special needs that nations often seek to address through social benefit programs and laws. Interestingly, law generally identifies the elderly through the use of chronological age (such as those over age 65), even though it might better identify the characteristics of aging through factors such as functional capacity, or physical and mental health (Arnold et al. 1998). Although chronological age is only a rough indicator of aging, it has the advantage of providing an unambiguous test that makes identification easier and more definitive. For example, if the nation provides an old age pension, it is much simpler to grant it to all above a certain age, rather than to identify those eligible by a more subjective categorization.

1. *Income Benefits for the Elderly*

The need to provide financial support for the elderly has spawned a number of social programs and laws. Most countries provide some sort of public pension, such as the social security program in the United States. Laws regulating the provision of pensions by employers to retired employees are also common. Some nations, including France, Germany, Italy, Sweden, and the United States, provide favorable tax treatment to encourage employers to provide pensions to retired employees or to increase the value of the pension to the employee; other nations, such as Canada and Japan totally or partially exempt employee pensions from income tax (Forman 1997). Laws mandating private pensions or altering the tax structure to encourage pensions often involve complex enforcement or monitoring procedures. Thus, the aging population has created a host of new laws and litigation in pursuit of adequate public and private pensions (Bodie et al. 1996).

Countries also address poverty among the elderly by outlawing age discrimination in employment (see *Discrimination*). Many elderly want to work, but they are denied the opportunity because of social attitudes that equate being old with being unproductive. While it is true that some elderly cannot perform certain jobs that require great physical strength or stamina, many are quite capable of being productive workers. Some want to retire from their current employment but wish to reenter the labor force, perhaps in a less demanding job. Unfortunately, custom and prejudice can be powerful barriers to elderly employment, and, while the United States has enacted laws that prohibit the arbitrary use of age in hiring and bar mandatory retirement on account of age, in most countries age discrimination against older workers is legal.

1.1 *Family Support Laws*

Some countries expect adult children to support their parents. In many countries such as Japan, older persons traditionally moved into the households of their children, who supplied care and financial support. However, many countries now have laws that require adult children to support their destitute parents or reimburse the state for the cost of their support. As might be expected, these laws are controversial and not necessarily enforced. The laws often lack support because it is unclear whether their purpose is to reduce the financial burden on government or simply to endorse the ideal that children should support their parents.

As a practical matter, family support laws seem antiquated in light of governmental welfare programs that provide income, medical care, and other forms of support for the elderly. For example, in the United States, the Medicaid health and long-term care program for the poor, which is jointly funded by the federal and state governments, specifically prohibits states from enforcing family support laws against the children in search of reimbursement for benefits paid to older Medicaid recipients.

There are also ideological objections to family support laws. Some believe that the support of the elderly ought to be a governmental function rather than a family obligation. Many commentators point out that in an intact family, the adult children will almost always assist their parents. If they choose not to, perhaps they have good reasons, such as financial instability. Family support laws cannot arbitrarily set a fair level of support in light of varying financial conditions, nor can they recognize in-kind assistance by children. Even more difficult are instances in which the older person may have been neglectful, abusive, or emotionally harmful to the child.

2. *Health Care Benefits*

Although many nations offer health care benefits for the elderly, payment systems vary greatly. Some countries favor tax-funded health care systems, while others rely on social insurance-based health care (Hennessy 1997). Most countries, including Canada, Great Britain, and Germany, have a single healthcare system, but in a few countries there is a separate program for the elderly. The Medicare program in the United States is an example of the latter.

Because of chronic health conditions and mental disabilities such as dementia, many elderly require long-term custodial care. To meet these needs, most economically advanced countries have enacted laws creating programs to subsidize long-term care for the

elderly. These include mandatory long-term care insurance, innovative housing to shelter dependent elderly, family support programs to encourage voluntary care of the elderly, and economic need-based assistance programs. All of these approaches spur activity in the private sector, as entrepreneurs seek to take greatest advantage of government aid. In turn, the government must regulate the private activity so that it does not place an undue burden on the government or otherwise pervert the aims of the government program. In short, the governmental response to a social problem creates a complex legal pattern of response and reaction.

3. *Housing Needs*

Housing is yet another area in which an aging population, experiencing declining mental or physical health and often financial problems as well, creates new challenges for the law. Some nations, such as Denmark, Sweden, and the United States, provide subsidized housing for the elderly. Sometimes housing for the elderly creates legal problems because it conflicts with laws that preclude discrimination on the basis of age or family status. Although age-segregated housing may discriminate against younger people, it also can create a safer environment and render more efficient the delivery of supportive services to the elderly (Frolik 1997).

4. *Mental Incapacity and the Law*

Aging populations force nations to confront the problem of an increasing number of mentally incapacitated elderly. Neurologic disorders such as dementia, depression, and strokes are common for the elderly and often cause loss of mental capacity (Beers and Berkow 2000). Because the legal world operates on the assumption that adults can make rational decisions, the legal system allows the creation of substitute decision makers, known as guardians, for individuals who lack sufficient mental capacity. Laws specify the process for establishing guardianship and the decisions that guardians may make for an incapacitated person.

5. *Legal Aspects of Health Care Decision-making*

Health care decision-making by the elderly can create complex legal issues with respect to who decides what is appropriate care (the patient, the physician, the family, or occasionally a non-marital partner). While the United States permits the patient to decide, other nations give the physician and the family more rights in the final decision. The most perplexing legal issue is whether health care can be terminated for a terminally ill patient when there is no further hope of recovery. There is considerable variation among countries with respect to the discontinuation of life support and who can make that decision if the patient is mentally unable to do so (Meisel 1995). While the United States has adopted laws that enable the individual to create a living will or an advance healthcare directive (documents that state under what conditions to terminate care or identify who should decide when to terminate care), almost no other countries have approved living wills or advance directives.

The law must also respond to individuals who demand the right to terminate their lives because they are terminally ill and do not want to undergo a painful process of dying. Whether patients have the right to terminate their lives is controversial. The right of a patient to terminate an intolerable life is gaining adherents and is legal, for example, in the state of Oregon in the United States. Most jurisdictions, however, do not permit assisted suicide.

Even more questionable is euthanasia, the legal right of someone to terminate the life of another, usually terminally ill individual. Though not officially legal, in The Netherlands physicians are not prosecuted for practicing euthanasia. Many contend that demands for physician-assisted suicide and euthanasia would greatly diminish if patients, particularly the elderly, were given adequate pain relief. It is not that older patients want to hasten death so much as that they fear pain. If assured that they would be given sufficient pain relief medications, even if that were to hasten death, they would not desire assisted suicide (Meisel 1995).

6. *Conclusion*

The dramatic growth throughout the world in the number of older persons has raised significant legal issues in many countries. The economic, health, and social needs of the elderly differ enough from the larger population to require societal and legal responses. Although the national responses are somewhat congruent, each country will necessarily evolve policies and practices that reflect its culture and laws. Thus while rising numbers of elderly persons is a worldwide phenomena, how the law reacts varies across the globe.

See also: Age Policy; Age, Sociology of; Aging and Health in Old Age; Caregiving in Old Age; Differential Aging; Discrimination; Population Aging: Economic and Social Consequences; Psychiatry: Informed Consent and Care; Retirement and Health; Retirement, Economics of; Social Cognition and Aging

Bibliography

Arnold R, Graetz M, Munnell A 1998 *Framing the Social Security Debate*. National Academy of Social Insurance, Washington, DC

Baker D, Weisbrot M 1999 *Social Security: The Phony Crisis*. University of Chicago Press, Chicago
Beers M, Berkow R 2000 *The Merck Manual of Geriatrics*, 3rd edn. Merck, Whitehouse Station, NJ
Bodie Z, Mitchell O, Turner J 1996 *Securing Employer-based Pensions: An International Perspective*. University of Pennsylvania Press, Philadelphia, PA
Cockerham W 1997 *This Aging Society*, 2nd edn. Prentice-Hall, Upper Saddle River, NJ
Cox H (ed.) 1998 *Aging*, 12th edn. Duskin Publishing Group, Guilford, CT
Forman B 1997 The tax treatment of public and private pension plans around the world. *The American Journal of Tax Policy* **14**: 299–360
Frolik L 1997 *Residence Options of the Older or Disabled Client*. Warren, Gorham, & Lamont, Valhalla, NY
Hennessy P 1997 The growing risk of dependency in old age: What role for families and for social security? *International Social Security Review* **50**: 23–39
Lynch E 1997 Late-life crisis: A comparative analysis of the social insurance schemes for retirees of Japan, Germany, and the United States. *Comparative Labor Law Journal* **14**: 339–84
Meisel A 1995 *The Right to Die*, 2nd edn. Wiley, New York

L. A. Frolik

Law and Democracy

1. *The Relationship Between Law and Democracy*

The relationship between law and democracy is complex and dynamic, with philosophical, political, legal, economic, and socio-cultural dimensions. Simplifying matters greatly, one may define democracy as rule by the *demos* (i.e., by equal human beings who resolve their conflicts on the basis of procedures and laws agreed upon by themselves), and the law as more or less systematically ordered, formalized, and with rational rules of conduct backed by the threat of force by those in political power when these rules are violated. What is decisive in the deep links between democracy and the law is the fact that the *demos*—'the people'—is the political source of the law to which it, in turn, subordinates itself. In modern liberal societies it is also important to devise such a system of autonomous and general rules whose content would enable the protection of individuals and social groups struggling for their rights and the fulfillment of their interests against a misuse of authoritative power, and to enable them to participate in the political decision-making process.

Although issues related to the relationship between democracy and law have been debated in the West since ancient times, contemporary processes of democratization, and the growing recognition of cultural pluralism and differentiation, greatly contribute to this debate. Of special importance here is the collapse of totalitarian and authoritarian regimes, which has been decisive in the reconsideration of classic concepts and principles, such as the principle of the rule of law, the concept of rights, and the principle of majority rule, as well as in the formation of new democratic rights, such as the civic right to truth about atrocities committed by the functionaries of a totalitarian or authoritarian regime (see *Civil Liberties and Human Rights*; *Rule of Law*; *War Crimes Tribunals*).

2. *Key Areas of Interest and Disagreement*

Many scholars and politicians, ancient and modern, have maintained that democracy cannot exist without law; without law, democracy would become anarchy. It is also argued that democracy is fundamental to the development of the law and contributes to the compatibility of the legal order to changing political and social reality. Democracy also contributes to the validity and legitimacy of the law: to democratic procedure, legitimacy, and to political coercion. A key issue is the potential misuse of law by a democratic power—government may pose a threat to individual sovereignty and creativity.

Another key area of inquiry is whether citizens have an obligation to obey the law. Scholars take different positions on this issue depending upon the model of democracy and on the type of society that they study.

3. *Democracy and Law in Classic Debates: From the Ancient* Polis *to Liberal Democracy*

3.1 *The Ancient* Polis

Thinking on democracy and law in the West has been shaped by the issues and questions raised by ancient philosophers and politicians with reference to the aims, ideals, and the constitution of the *polis*, the ancient Greek city-state.

Three issues are most relevant for deliberations on democracy and law: (a) the civic and political culture (civic virtues); (b) the rule of law, due process, and constitutional government; and (c) the idea of superior argument in lawmaking, and its enforcement. The latter idea demonstrates a foundation of the most influential contemporary theory on law and democracy, i.e., the theory of procedural democracy.

Among the ancient thinkers and politicians, were Pericles, Aristotle, and Plato. Pericles emphasized that respect for the law results from a law making process in which people participate and in so doing, simultaneously subject themselves to an obligation to obey the law. The citizen's law, and respect for such law, lies in contrast to the arbitrary will of tyranny and subjugation to it through the use of brute force.

In his *Politics*, Aristotle (384–22 BC) observed that, if liberty is the first defining principle in the model of the democratic city state, it had to be closely related to an equality of opportunity to hold office: only equal

participation in legislation protects the freedoms and liberties of citizens. Therefore, liberal democracy relies on the principle that every citizen has an equal opportunity to rule, which in turn justifies obedience to laws made by a democratic majority, even if free citizens want to live 'as they like' (Aristotle 1981, pp. 195–8).

The most important criticism of democratic rule was raised by Plato. In one of the famous dialogs Plato (427–347 BC) challenges democratic rule by raising the question: Why is the tyranny of the majority better than that of an oligarchy or of an individual? This question returns in every debate on the issue of majority democratic institutions protecting individuals and minorities from a potential majority abuse of power (Aubert 1983, pp. 6, 7).

Despite the intellectual potential and contemporary relevance of issues and arguments raised by ancient thinkers, the institutional forms and ideas linked with the ancient *polis* have their obvious limitations. First and foremost, equality and freedom were restricted to those inhabitants of the city-state who were regarded as citizens, i.e., adult, male, and Athenian-born. Women, immigrants and, above all, slaves, were excluded from this category. These ideas lacked the modern liberal fundamental conviction that all human beings are 'individuals,' and bearers of inalienable rights. Ultimately, their empirical terms of reference were based on a relatively small and structurally simple, homogeneous society whose economy was predominantly based on slave labor, and whose members could participate directly in legislative and juridical proceedings because of the small numbers involved and their affinity with one another. Nevertheless, the issues raised by ancient thinkers and politicians have set a framework for debate, as well as an ideal for institutional arrangements in modern times.

3.2 Liberal Democracy

Questions and arguments concerning the relationship between democracy and the law have been especially important under conditions of rapidly growing social differentiation, of changes in social integration accompanied by diffusion of ownership, and of a developing free-market economy. This last condition entails the formation of a relatively broad class of producers and merchants on the one hand, and the growing complexity and centralization of an impersonal, bureaucratic state apparatus on the other. These conditions characterize liberal democracy in the West. Although classic, these questions and arguments—informed by processes of modernization—are reflected in basic institutional arrangements, predominantly in the constitutions of contemporary liberal democracies.

There are two important issues connecting law and democracy within the set of institutional arrangements defined as liberal democracy. One is presented by conflicting interests, the other by protection of individual rights.

In a rapidly changing, diffused society characterized by conflicting interests, relations between citizens themselves, as well as relations between citizens and the state, are developing new forms of the concept of the social contract. These relations also require specific forms of law—a medium through which compromises among conflicting interests can be struck. Moreover, they require a law with attributes of generality and autonomy in order to reflect changing social values and power relations. In the words of Unger (1976, pp. 68, 69), in liberal societies the law somehow accommodates antagonistic interests, and its procedures are such that anyone might find it in his or her own interests to subscribe to them regardless of the ends one happens to be seeking. Therefore 'the liberal type of social organization generates, and is reinforced by, a style of consciousness whose substance is the image of society as an arena of conflicting subjective interests.'

Thoughts concerning the idea of a social contract as a foundation for political institutional arrangements which accommodate antagonistic interests, and the delegation of legal sovereignty crucial for the validity and legitimacy of law and its observance are to be found in the writings of Thomas Hobbes (1588–1679). These ideas were developed and supplemented by John Locke (1632–1704), Charles Louis de Secondat, Baron de Montesquieu (1689–1755), Jean-Jacques Rousseau (1712–78), James Madison (1751–1836), Jeremy Bentham (1748–1832), James Mill (1773–1836) and John Stuart Mill (1806–73).

Generally, according to proponents of the classic model of liberal democracy, the activities of a democratic government are to be limited and controlled, since the sole aim of the state is to preserve the 'life, liberty, and estate' of its citizens (Locke 1960), or, as emphasized by John Stuart Mill, to protect the free development of the individuality of every citizen. This is possible in the form of a limited, constitutional government, subordinated to the law, whose activities are also subject to the mutual checks and balances of autonomous branches of government—in the highly developed doctrine of Montesquieu, these are the checks and balances of the legislative, executive, and judiciary powers (Montesquieu 1952).

In the public sphere of relations between citizens and the state, the idea of a contract concluded by free and formally equal citizens, together with the 'extended' franchise (initially understood as applying to male property-owing citizens), legitimizes the political and legal sovereignty of the government. These ideas are closely linked both to the notion of liberty, i.e., freedom within legal limits, and to the notion of natural, universal, and inalienable public rights of citizens whose activities are governed by reason, and who participate, directly or through their representatives, in the legislative process.

Classical thinkers located the source and foundation of individual rights and liberties in the idea of natural law as a 'higher' than man-made, positive law (see *Natural Law*). Natural laws were defined as those discovered by reason, i.e., by rational citizens acting under conditions of freedom according to the arguments of the German philosopher Immanuel Kant. However, they were also understood to be of divine origin (Locke 1960, p. 311). For Montesquieu, natural laws discovered by reason were to be interpreted in the context of local customs and traditions (Montesquieu 1952, p. 32). The utilitarians Jeremy Bentham and James Mill enunciated clear formulas of natural laws as the maximization of individual satisfaction whose source was the natural tendency of human beings to pursue individual happiness and interests, and to avoid suffering (Bentham, *Fragment on Government*; James Mill, *An Essay on Government*).

In the classic model of liberal democracy (based on the primary importance of protecting individual property rights and the utilitarian maximization of individual profits), the 'haves' are protected first and foremost, and individual goals and interests dominate the constitutional, democratic state. The concept of equality proposed by the classic thinkers, and written into the first liberal constitutions, was therefore initially still very restrictive. Only male property-owners had a right to vote or to hold public office. However, John Stuart Mill, in discussing the right to the development of individuality referred also to women and minorities, and connected this right with a postulated broad franchise (John Stuart Mill, *On Liberty*, 1982). Those postulates became enshrined in liberal constitutions.

Different ideas on the contract, the legitimacy of the law, civic virtues, civic commitment, and public government were proposed by the first representatives of the republican version of classic democracy. First, the notion of contract was understood not as a legal fiction, but as a real contract, closed by the people, in public debate and referenda on the most important public issues. The issue of civic virtue and of civic duties, and the notion of a citizen as an unselfish, committed participant in public affairs, was emphasized by Jean-Jacques Rousseau. Advocating direct participation in legislation, and defining the equality of citizens as economic independence and material security, Rousseau opted for the model of a discursive democracy (Rousseau 1973).

4. *Liberal Democracy and Law: Contemporary Issues*

'Liberal democracy was the first great secular religion in the Western history' (Berman 1983, p. 32). However, new problems in the relationship between democracy and the law have been created by the large-scale, complex, densely-populated, culturally-pluralistic, and differentiated modern societies which developed after the industrial revolution. Three different issues are most important for the current debates on liberal and democratic institutional arrangements. The first of these concerns the principle of equality which has an impact on public debate and civic engagement in the democratic process. The second issue is presented by the growing differentiation of law because of its adaptation to a complex and pluralistic, modern democratic society, and efforts to interpret universal legal standards and principles in light of local experiences. The third issue involves the growing concern with human rights protection.

One of the main questions raised in contemporary debates on liberal democracy primarily concerns the concepts of equality and justice which are no longer defined only in terms of formal justice, but rather in substantive terms such as equal opportunity, and even affirmative action and positive discrimination designed to achieve social justice. Debates about social justice point to the growing scope and range of state intervention and its consequences for law, on the one hand, and the civic engagement of the democratic electorate on the other.

It is argued that the political organization of the democratic welfare state challenges the universality and autonomy of the law, as well as its democratic legitimacy. The law becomes an instrument of state intervention, characterized by a growing number of particular substantive goals. Authors such as Unger in the United States, or Luhmann in Germany, have pointed to the specific consequences of state interventionism for democracy and law. According to Luhmann, political inclusion—i.e., the subordination of an ever-growing range of social relations to the intervention of the political democratic system of the welfare state—imposes new public tasks on the state which is considered responsible for the well-being and quality of life of its citizens. This results in rising social demands for new standards and new regulations of ever more broad and particular issues, from a minimum wage to the functioning of kindergartens, and to the regulation of public space, expressed not in the form of citizens' rights, but of citizens' claims on the state (Luhmann 1990).

This phenomenon is described as the 'socialization of the law,' on the one hand, and 'juridification' or 'legalization' of social relations on the other. These new developments led to changes in the form of legal regulation and legal intervention, and in relations between the legislative, judiciary, and executive branches of government. The integrity and autonomy of the legal system is challenged by increasingly specific demands for interventionist legal regulation and the necessity of *ad hoc* reactions to changing social needs. Moreover, instrumental legislation is produced to an ever increasing extent by governmental agencies in the form of administrative prescriptions, based on a concern for technical efficiency in the fulfilling of state

tasks of social protection—not exclusively on democratic legitimacy, and certainly not on some deeper transcendental or cultural reasons why it could be very difficult to agree in a complex, pluralistic, and culturally deeply differentiated society. A consequence is the technical nature of legal regulation; the standards on which it is based preclude public debate on the contents of the law, and could result in the growing passivity of a democratic electorate.

According to Unger, post-liberal, policy-oriented legal reasoning and discretionary justice challenge legal formalism and lead to the development of more flexible legal institutions which reflect and enhance a strong sense of community (Unger 1976, pp. 194, 196).

In the light of totalitarian and authoritarian experiences, it is, however, important to note that, in modern and complex societies, a strong sense of community may be imposed from above by a highly centralized government, and by political propaganda. Discretionary justice, based on some politically-supported standards could easily turn into repression and violation of human rights, if universal standards and autonomous law are not strong enough. This was well-illustrated by Nazi Germany, and by the so-called 'people's democracies' in Eastern Europe and the Soviet Union. In the latter, 'socialistic legality' was understood, on one hand, as the conformity of instrumental regulation to an ideology interpreted by the higher ranks of party functionaries, and, on the other, on the observance of rules by society and the lower echelons of authority (Damaska 1986, p. 197).

Increasingly technical legal reasoning, together with the growing complexity of democratic institutional arrangements, is challenged by new developments and ideas on democracy and law. One observes the growing importance of judicial review—above all, of constitutional review—as controlling not only the application of law in individual cases, but also the conformity of governmental decisions and parliamentary legislation to the constitution. Therefore, instead of a clear separation of powers, a functional interdependence of the legislative, judicial, and executive branches of government characterizes modern democracies.

Another important feature of late modernity is the impact on law-making of lobbying groups, NGOs, or spontaneous social movements representing civil society. These new developments seem to be both inevitable and benign, but the pluralism of values and lifestyles and the increasing differentiation of collective interests raise the question of how a social contract is possible.

In an attempt to solve this problem German philosopher and social theoretician Juergen Habermas postulates that discursive and procedural legitimacy of legal and political institutions should be justified not in terms of common values, but in procedures regulating the free and unconstrained public debate on values and norms, open to every citizen at any time (Habermas 1986, 1992). According to this view, norms are considered justified only if they can stand up to an unlimited and unrestricted debate.

The model of procedural democracy and the proceduralization of law leaves some questions open. The legal and political systems of modern democratic societies are not only flexible but also fragile. Their institutional arrangements and interpretations of norms depend on the results of debates and deliberations, or are left unresolved because common understanding or compromise cannot be reached. Questions remain about (a) legal certainty and legal security as components of the rule of law principle, (b) more stable and reliable foundations for the adjudication and legitimation of the democratic state ruled by law, and (c) constraints on political and judicial deliberations.

In contemporary democracies, stable foundations for the legal as well as political system are sought in the concept of rights. This increasingly more important approach to law—called 'new personalism' (Friedman 1990, p. 44–46)—refers to inalienable human rights as a higher standard for the evaluation of democratic and legal systems; these would serve as a final test of democratic and legal order.

To conclude: the relationship between democracy and law in modern, complex, and pluralistic societies must be analyzed and evaluated with reference to universal standards—encompassed by human rights—interpreted and applied in the changing social and differentiated cultural conditions of the modern world.

See also: Balance of Power, History of; Balance of Power: Political; Communication and Democracy; Delegation of Power: Agency Theory; Democracy; Democracy, History of; Democracy: Normative Theory; Democratic Theory; Democratic Transitions; Justice: Political; National Security Studies and War Potential of Nations; Power in Society; Power: Political; Rechtsstaat (Rule of Law: German Perspective); Society/People: History of the Concept; State and Society; State, Sociology of the

Bibliography

Aristotle 1981 *Politics*. Penguin, Harmondsworth

Aubert V 1983 *In Search of Law. Sociological Approaches to Law*. Robertson, Oxford

Berman H J 1983 *Law and Revolution. The Formation of the Western Legal Tradition*. Harvard, Cambridge, MA

Damaska M R 1986 *The Faces of Justice and State Authority: A Comparative Approach to the Legal System*. Yale, New Haven, CT

Friedman L B 1990 *The Republic of Choice: Law, Authority and Culture*. Harvard, Cambridge, MA

Habermas J 1986 *Law as Medium and Law as Institution*. In: Teubner G (ed.) *Dilemmas of Law in the Welfare State*. de Gruyter, Berlin
Habermas J 1992 *Faktizität und Geltung*. Suhrkamp, Frankfurt
Held D 1987 *Models of Democracy*. Blackwell, Cambridge
Locke J 1960 *Two Treatises on Government*. Cambridge University Press, Cambridge, UK
Luhmann N 1990 *Political Theory in the Welfare State* [trans. Bednarz J Jr.]. de Gruyter, Berlin
Mill J S (ed.) 1982 *On Liberty*. Penguin, Harmondsworth, UK
Montesquieu C 1952 *The Spirit of Laws*. Benton, Chicago
Rousseau J J 1973 The social contract. In: *The Social Contract and Discourses*. Everyman, London
Unger R M 1976 *Law in Modern Society. Toward a Criticism of Social Theory*. Free Press, New York

G. Skąpska

Law and Development

1. *Law and Development in the Beginning*

'Law and development' is a term usually used to describe legal assistance programs for developing countries and the related academic work. The movement as such, began in the 1960s with overseas projects funded by various foreign foundations and development agencies. Legal scholars at leading American law schools, as well as in some European institutions soon adopted the topic as an academic field. Following foreign aid patterns of the time, the movement primarily focused on Latin America and Africa. It sought to aid developing countries by promoting changes in laws and legal systems to make them more 'modern.' Paralleling the then popular and now defunct modernization theory, this law and development paradigm presumed that as these countries developed, they would construct legal institutions and cultures similar to those that spurred economic and political development in the West. The law and development movement assumed that convergence on the Westernized notion of the rule of law could be hastened by external assistance.

American scholars sought to develop a theory of the role of law in state and market development that could be integrated into general modernization theory. To that end, they drew on Max Weber's historical explanation of the rise of capitalist societies in Western Europe (Trubek 1972). Weber's theory asserted that in addition to the 'Protestant ethic,' the rational, formal, and logical legal system was necessary to provide the stability and predictability in economic exchange, and the protection of property rights, which facilitated economic innovation. Weber's work suggested that law would best contribute to capitalism if it was based on universal rules uniformly applied, contract law guaranteeing the security of future exchanges, and property law capable of protecting the fruits of labor.

However, this model of legal development did not fit squarely with the spirit of the 1960s and the needs of the development agencies. In an era in which development policy favored a strong and proactive role for the state, a legal theory was needed that would promote the use of law as an instrument to modernize society. In such a vision, lawyers, and judges were seen as social engineers, who would cultivate the legal norms and the changes in the formal system of law needed to promote progressive political and economic development. The emphasis on law as an instrument of a modernizing state and the Weberian idea of a set of formal rules that guaranteed predictability did not sit easily together. To bridge the gap, law and development scholars produced a hybrid theory, dubbed by some 'liberal legalism' that sought to unite a faith in the rule of law with a commitment to aggressive social engineering. It was assumed that this hybrid was the prevailing ethos of US legal culture, and if exported to the developing world it would hasten progress towards modernity. As a result, the emphasis of foreign aid was on legal education, and training was done by Western scholars and at universities in the US and other developed 'first world' countries. It was assumed that if the ideas taught in developing country law schools could be changed, then the behavior of lawyers would also change. If the right laws were created, they would be enforced and if the positive benefits of the Westernized 'rule of law' could be demonstrated, then it would be embraced.

In the mid-1970s, however, a series of articles appeared that challenged the original ideas of the law and development movement (e.g., Trubek and Galanter 1974). These articles reflected a growing unease among scholars regarding the assumptions and prescriptions of legal liberalism. A decade of experience with the original approach showed there were serious problems with the strategy of transplanting foreign legal models through an elite-driven approach that did not coincide with the cultural, social, and political contexts of the developing countries (Burg 1977, Merryman 1979). The failure of legal transplants and the stigma of 'legal imperialism' (Gardner 1980) convinced many that the movement had got off on the wrong foot.

As experience accumulated, scholars began to realize that the enterprise had not been guided by a robust theory of the relationship between law, and economic and political development. Critics pointed out that what passed for law and development theory was a somewhat idealized projection of the self-understanding of US and other developed country lawyers, concerning the role of law in their societies. This was combined with a simplistic model of history that presupposed causal and linear relationships between the evolution of a 'modern' legal system and market democracy. Needless to say, no empirical evidence was ever produced to support this theory which was little more than a romantic and ethnocentric projection of a

developed country legal culture's ideal of itself. It became clear to some that the movement suffered from an entrenched ethnocentrism and naïveté (Trubek and Galanter 1974).

This critique was intended to strengthen law and development scholarship by distancing the academic project from immediate activist policy agendas and encouraging the development of more systematic empirical research and conceptualization. The hope was that in the long run the academy could come up with better and more effective ideas for legal assistance to developing countries. But that was not the result: rather, this critique helped kill the nascent academic project. The critique was taken by many to be a denunciation of the movement. Moreover, it came at a time when the aid agencies were losing interest in law for other reasons, and when US intellectuals had become especially sensitive to charges of meddling in foreign countries as a result of the Vietnam War protests. As a result, the law and development movement lost its momentum. Comparative legal scholars turned their attention to other pursuits, funds for law and development programs dried up, and after a 'short, happy life' the movement was presumed dead.

2. *Law and Development Reawakened*

The 1990s revealed, however, that the law and development movement was not dead, but merely in hibernation (Tamanaha 1995). While some law reform activities continued throughout the 1970s and 1980s (such as USAID judicial assistance programs and foundation support for public interest law in developing countries), it was the end of the Cold War and the collapse of communism in Eastern Europe and the former Soviet Union which incited a 'rule of law revival' and gave new life to external assistance in legal and judicial reform (Carothers 1998). At the twentieth century's end, if all the current 'rule of law' portfolios of the World Bank, regional banks, major nongovernmental organizations and foundations, the US Agency for International Development (USAID), and other bilateral aid institutions and local counterparts were added up, it would equal several billion dollars worth of assistance funds being spent on law projects.

The current law and development movement emerged from the confluence of two broad forces that came to prominence in the post-Cold War era. The first of these could be called 'the project of democracy' and came out of the human rights movement of the 1970s and 1980s. The international community made great progress in specifying human rights norms, creating machinery for international action to enforce them, and ensuring that internationally recognized human rights became a part of the discourse of domestic politics in many countries. Events such as the Helsinki process drew attention to the lack of protection for human rights in domestic institutions. This highlighted the need for legal change at the national level to ensure that human rights would not only be nominally secured in the letter of the law, but would also be protected in practice. These universal human rights, in addition to the 'Third Wave' of democratization (Huntington 1991) drew attention to the need for constitutional protections for civil liberties, the independence of the judiciary from potential corrupt and weak states, and the guarantee of equitable access to justice for all of civil society.

The second—and stronger—force propelling the 'new law and development' was based on economic reform. While the burgeoning international human rights regime and democracy assistance sustained the notion of foreign legal aid in the name of justice throughout the 1980s, the dramatic revival of the law and development movement in the 1990s took on a decidedly economic character. The catalyst was the collapse of communism in Eastern Europe and the former Soviet Union, and the subsequent failure of shock therapy to produce a successful transition to the market through a tripartite plan of liberalization, stabilization, and privatization. Development specialists began to claim that initial reform strategies had failed due to a lack of attention to the 'institutional preconditions' of a capitalist economy. Foremost among these were the appropriate legal institutions.

The empirical lessons of the post-Communist transition coincided with the increasing popularity of new institutional economics, based on the Nobel prize-winning work of Douglass North, among others. In his *Institutions, Institutional Change and Economic Performance*, North (1990) argues that as economies develop from 'traditional' to 'modern,' rational economic actors seek to construct institutions that among other things will protect property rights and reduce transaction costs by providing credible and efficient means of contract enforcement. These legal institutions (both formal and informal) create incentives for rational economic actors to engage in increasingly complex modes of exchange conducive to furthering economic growth. In North's theory, it is the absence of such modern legal institutions that explains the lack of economic development in the Third World (North 1990, pp. 64–5).

Consistent with the 'Washington Consensus,' embedded in the development ideology of the Bretton Woods Institutions, this economic theory reaffirmed the role of law in creating and protecting an enabling environment for the development of the market. Unlike the previous law and development movement, when state-led economic development theories were in vogue, in this approach the state plays a largely passive role and the law assumes an autonomous and proactive role in providing the 'legal foundations of capitalism' (Commons 1924) while at the same time monitoring and constraining the arbitrary use of power or the excessive use of regulation by the state. In the context of globalization, this role of law and the

formal legal system is deemed critical to establishing the 'good governance' necessary for attracting foreign and domestic sources of private investment (Shihata 1997). In this approach, we are back to Weber, shorn of the activist, social engineering gloss of the 1960s 'liberal legalist' hybrid.

The strategies of the new law and development movement, aimed at the dual projects of human rights and markets, reflect a widened scope of legal and judicial reform. While much of external assistance continues to target legislative drafting, legal education, training, and the professionalization of the bar (including exchange trips), reform activities place more emphasis on the effective administration of justice. This includes the efficient management of cases, increased access to justice through the construction of alternative dispute resolution mechanisms, enhanced means of enforcing judicial decisions, and the promotion of judicial independence. Moreover, broader legal reform projects seek in certain contexts to create a demand for a rule of law where one is perceived to be absent (as in much of the former Soviet Union) by constructing public awareness campaigns and building domestic consensus and support for reform efforts.

3. *Comparing the 'Old' and 'New' Law and Development: Lessons Learned?*

The revived legal reform agendas do not avoid the ethnocentric, 'imperialistic' nature of the old law and development movement. There is an over reliance on legal technical assistance, which tends to neglect the deeper political and cultural environments shaping the legal systems. There is still a faith in legal transplants, a belief that Western lawyers can design systems that will work in emerging markets and new democracies, and a stress on education. However, there are also signs of dissent from this orthodoxy. For example, some activists profess to avoid the mistakes of the past by engaging in more participatory work with civil society groups, build upon existing grassroots legal reform efforts, and generally working as much as possible within the existing legal infrastructure. Critics have also rediscovered the old critique of legal transplants. It is increasingly recognized that in the absence of a clear demand from key societal actors and reform strategies that build upon (rather than against) the rooted legal culture of a country, 'supply-side' legal reform efforts focused on constructing new laws and institutions, will be insufficient for the task of creating and consolidating a rule of law (Hendley 1996).

While these breaks in the law and development orthodoxy are welcome, they represent a minority voice. The faith in legal change through top-down reforms supported by Western legal expertise remains strong. The idea that there is one best way to build the legal institutions for markets and democracy is still pervasive, as is the faith that this path has already been trod by the rich countries of the West which stand as models and as sources of expertise. People still believe that Western laws can be transplanted *tout court* to countries as diverse as Brazil and Bangladesh.

The continuing faith in law and development orthodoxy comes as no surprise. The leaders of the West have decided that the rule of law is essential for both emerging markets and new democracies. As a result, the international and bilateral development agencies are committed to legal reform on a global scale. To implement these programs in all the countries they deal with, these agencies need relatively simple models that can be easily supported by the traditional tools of foreign assistance. The last thing they want to hear is that each country is unique, that there is no easy answer to what laws and what legal system reforms are needed, or that Western expertise is of dubious value.

Billions of dollars have been spent on reforms that are guided more by intuition and hope than by any systematic knowledge or empirical evidence. Projects still rest on assumptions criticized in the 1970s. Will history repeat itself? Will the cry for a more independent and critical academic examination of 'law and development' still go unheard? While the academic side of the 'new law and development' remains modest at least in contrast to the scale of the assistance effort, at the beginning of the twenty-first century there are signs that a critical and autonomous endeavor may be taking shape in the academy.

If so, there are several questions that must be dealt with. First, scholars should study the microfoundations of the so-called 'rule of law.' The old law and development movement was a very top-down effort. It was assumed that law was a universal boon for all classes. It assumed that if the elite said they favored the rule of law, they meant it. It was assumed that most people would see law in a positive light so that legal reform would enjoy popular support. Needless to say, none of these assumptions were correct in the 1960s, yet they still pervade policy circles at the turn of the century. Scholars should investigate how various social interests benefit from law and legal order, and how interests can be mobilized to build and embed the rule of law while ensuring that legal orders do not just favor the strong and the rich.

Second, scholars must ask to what extent can there be a true trans-national discourse on law and development. Is it possible to create useful and authentic knowledge about law and development that is based on theory and has relevance beyond a single country or even a single field of law in one country? One of the cardinal accusations levied at the old law and development movement was that it tried to impose an ethnocentric bias for Western law in the guise of academic knowledge. As we approach the task of refining the new law and development, we have to be careful to avoid that trap. It is tempting to see the legal institutions of our own society as a model for

emulation by developing nations, albeit with a lag to allow for evolution to do its work. This is even more alluring when to do so happens to foster our national interest in shaping foreign markets in ways that will be congenial to Western capital and our own national firms in particular. To avoid this fallacy, the new law and development should be tentative, situational, and dialogic. We should be careful not to project our own past on others and idealize our own society. We should seek to develop knowledge only in conjunction with those directly affected by our 'science' in a truly transnational environment.

Finally, the academy needs to address the normative vision that guides scholars and activists working on law and development. The law and development movement has multiple and potentially conflicting goals. On the one hand, the promotion of democracy and human rights through legal and judicial reform promises greater access to justice and a means of enforcing accountability and transparency. On the other hand, legal reform efforts aimed at creating a market-friendly environment attractive to foreign and domestic investors (such as the growing attention to alternative dispute resolution mechanisms and the funding of economic legislative drafting) can overshadow or crowd out the social and political dimensions of reform. Even more dangerous is the possibility that market-driven legal reform that benefits an economic elite minority may run counter to democracy-oriented reform goals that are geared towards a politically powerful but impoverished majority, with the resulting potential for ethnic tensions or conflict (Chua 1998). The academic community must always be sensitive to potential conflicts, and ensure that the enterprise promotes fairness as well as efficiency, and democracy as well as markets.

See also: Development and the State; Development, Economics of; Development: Social; Development: Social-anthropological Aspects; Human Rights, Anthropology of; Human Rights, History of; Human Rights in Intercultural Discourse: Cultural Concerns; Justice and Law; Law and Democracy

Bibliography

Burg E 1977 Law and development: A review of the literature and a critique of scholars in self-estrangement. *American Journal of Comparative Law* **25**: 492–530

Carothers T 1998 The rule of law revival. *Foreign Affairs* **77**(2): 95–107

Chua A L 1998 Markets, democracy, and ethnicity: Toward a new paradigm for law and development. *Yale Law Journal* **108**(1): 4–107

Commons J R 1924 *Legal Foundations of Capitalism*. The Macmillan Company, New York

Gardner J A 1980 *Legal Imperialism: American Lawyers and Foreign Aid in Latin America*. University of Wisconsin Press, Madison, WI

Hendley K 1996 Law and development in Russia: A misguided enterprise? In: *American Society of International Law, Proceedings of the 90th Annual Meeting*, March 20–27, 1996, Washington, DC

Huntington S P 1991 *The Third Wave: Democratization in the Late Twentieth Century*. University of Oklahoma Press, Oklahoma City, OK

Merryman J H 1979 Comparative law and social change: On the origins, style, decline and revival of the law and development movement. *American Journal of Comparative Law* **25**: 457–91

North D C 1990 *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, UK

Shihata I F I 1997 *Complementary Reforms: Essays on Legal, Judicial, and Other Institutional Reform Supported by the World Bank*. World Bank, Washington, DC

Tamanaha B Z 1995 The lessons of law-and-development studies. *American Journal of International Law* **89**: 470–86

Trubek D M 1972 Toward a social theory of law: An essay on the study of law and development. *Yale Law Journal* **82**(1): 1–50

Trubek D M, Galanter M 1974 Scholars in self-estrangement: Some reflections on the crisis in law and development studies in the United States. *Wisconsin Law Review* 1062–102

D. M. Trubek

Law and Economics

Under the economic approach to the analysis of law, two basic questions about legal rules are addressed: descriptive questions, concerning the effects of legal rules on behavior and outcomes; and evaluative questions, concerning the social desirability of the effects of legal rules. In answering these questions, the method employed is that used in economic analysis generally. Namely, individuals and firms are ordinarily presumed to be forward-looking and rational, and the framework of welfare economics is adopted to assess the social desirability of outcomes. The field of economic analysis of law may be traced significantly to Bentham (1789), but lay essentially dormant until the contributions of Coase (1960), Becker (1968), Calabresi (1970), and Posner (1972). The field is now rapidly growing, although it is far from mature (one indication being lack of empirical work). To illustrate the approach, this article first focuses on accident law, briefly considers other areas of law, and concludes with a section on basic foundations and criticisms of the economic approach.

1. Economic Analysis of Law Illustrated: Accident Law

By accident law is meant the law governing liability for accidents, that is, the rules determining when a party who causes an accident must pay for the harm done. Economic analysis of this branch of law centers

on three issues: incentives toward safety; insurance and compensation of accident victims; and litigation-related costs.

1.1 Incentives Toward Safety

A major effect of the liability system is that it fosters the taking of precautions against accidents. Suppose that a precaution will lower accident-caused harm for which a party would be liable from $3,000 to $2,000, that is, by $1,000, but costs less than this amount, say $500. The party would then be likely to take the $500 precaution, as it would reduce the party's liability expense by $1,000. Such logic underlies the conclusion that, under many forms of liability, parties will be led to take socially desirable precautions.

An important qualification about the general point that liability creates incentives toward safety applies to accidents caused by firms' products. In particular, a firm's interest in its reputation may lead it to take proper precautions even in the absence of liability. If a firm's water heaters tend to fail and consumers know about this, they will not pay as much for the heaters, and thus the firm will have a motive to reduce the risk of heater failure in order to avoid having to accept a lower price. Note, however, that this argument depends on consumers' obtaining information about product risk.

1.2 Insurance and Compensation

The role of insurance in the context of accidents and liability is of substantial importance due to the widespread ownership of insurance. A major form of insurance is liability insurance, which provides coverage against legal liability. Because liability insurers pay for much of the losses for which injurers are found liable, it might initially be thought that liability insurance largely negates the incentives toward safety inherent in liability. However, some such incentives are preserved under liability insurance because insurers often link premiums or conditions of coverage to the adequacy of precautions, raise premiums on the basis of accident history, and offer only partial coverage against liability.

In addition to liability insurance, standard insurance for victims, that is, their (private or public) medical, life, disability, and property insurance, is of significance. The prevalence of victims' insurance limits the need for the liability system as a means of assuring victims compensation for accidents.

1.3 Litigation-related Costs

The litigation-related costs of the liability system are the legal fees and associated expenses (including litigants' time and effort) borne by parties in resolving disputes that arise when harm occurs. Litigation-related costs are high; for every dollar received by a victim, it appears that over a dollar is spent delivering the dollar to the victim.

1.4 Evaluation Under the Economic Approach

Under the economic approach, the liability system is considered to be socially worthwhile where its social benefits exceed its litigation-related costs. The social benefits of the liability system do not lie significantly in compensation of victims, because standard forms of insurance for victims are a cheaper means of compensation than the liability system. Rather, the social benefits of the liability system reside largely in its influence on accident rates. If this accident reduction effect is sufficient to outweigh litigation-related costs, the liability system is socially worthwhile. For example, the liability system might be socially worthwhile in the area of industrial pollution—perhaps the desire to avoid liability induces firms to reduce substantially polluting activity. However, the liability system might not be worthwhile in the area of automobile accidents—perhaps the prospect of liability does not affect driving behavior, which is mainly influenced by fear of personal injury in accidents.

1.5 Traditional Analysis Contrasted with Economic Analysis

The traditional view of the liability system is that its primary effect, and its major warrant for existence, is the compensation of victims of harm. This view is at odds with the economic view. On one hand, as just emphasized, compensation is to a great extent accomplished by standard forms of insurance; and, because insurance is a less expensive means of compensation than the liability system, it would be economically unwise to employ the liability system for the purpose of achieving compensation. On the other hand, the inducement of safer behavior is usually not paid serious attention under the traditional view of the liability system. Hence, the prescriptions for use of liability under the two views may conflict. Under the economic view, but ordinarily not under the traditional view, the recommendation is that the scope of liability be reduced where liability has little influence on accident frequency, and that the scope of liability be increased where liability would substantially reduce accident frequency.

A second, related aspect of the traditional view is that liability is intended to ensure the attainment of corrective justice in the classic sense that a wrongdoer should be punished by being made to pay the victim for harm done. From the economic perspective, this

conception of the purpose of the liability system is problematic even in a descriptive sense. As stressed above, a party who is found liable usually does not pay a judgment himself but has his liability insurer pay; if there is punishment, it might be that the party is not fully covered or that his liability insurance premiums may rise. Hence, the degree to which a wrongdoer is punished owing to liability is not direct; it is attenuated and translated in character by liability insurance. Additionally, the victim often does not receive the payment made by the liability insurer; rather it is the victim's insurer that frequently obtains the payment (as reimbursement for the payment that the insurer made to the victim earlier). In all, then, the liability system does not achieve corrective justice in the way contemplated by traditional analysts, because they overlook the effects of liability and of victims' insurance.

2. *Economic Analysis of Other Areas of Law*

In this section, the contours of economic analysis of other major areas of law will be indicated.

2.1 *Property Law*

A fundamental topic in economic analysis of property law is the justification for the very existence of property rights. From the economic point of view, these rights are said to exist because they promote incentives to work, to maintain and improve things, and to trade; and, as well, because the rights reduce problems of wasteful and destructive efforts to take things and to prevent takings. Some economic literature traces historical instances of the emergence of property rights (especially in land) and certain recent property rights developments (for example, in the broadcast spectrum) to these social advantages. Property rights are seen as beneficial due to their salutary effects, not to any intrinsic belief that a person should own the fruits of his labor.

Given the general basis for the existence of property rights, many issues pertaining to property law have been addressed from the economic perspective. One concerns acquisition of property rights in things not yet owned, such as fish in the sea and undiscovered mineral deposits. Here, a theme is that the 'finders-keepers' rule has the advantage that it creates incentives to find things. Yet these incentives may be socially excessive: individuals may engage in duplicative activities to find things (a large number of fishermen may compete for fish that a smaller number could easily catch; many parties may drill oil wells, even though a small number of wells would be sufficient to extract the oil from a reservoir). When this is so, it is suggested that regulation (limiting the catch of fish, 'unitizing' ownership of an oil reservoir) may be desirable to curb the problem of excessive effort.

Another issue of interest concerns external effects associated with the use of property. These 'externalities' may be detrimental, such as pollution or noise creation, or beneficial, such as beautification of land or spraying to kill mosquitoes. Bargaining among affected parties may sometimes resolve externality problems. Suppose that a factory causes annual pollution harm of $10,000 to its neighbor but can prevent the pollution by installing scrubbers at an annual cost of $1000. In the absence of legal rules preventing pollution or imposing liability for it, the victim of the pollution might be expected to pay the factory to obtain scrubbers in order to avoid suffering harm. The victim should be willing to pay more than $1000 (perhaps $2000) for the factory to obtain scrubbers. Hence, a legal rule requiring the factory to install scrubbers, or a rule imposing liability for harm, may not be needed to achieve the result that the factory obtain scrubbers. The general possibility that externality problems may be avoided through bargaining, and that legal rules may not be necessary for their resolution, is known as the Coase Theorem, and was advanced in Coase (1960). However, much economic writing discusses reasons why bargaining may not cure externality problems: costs of bargaining (especially significant when the number of affected parties is large), breakdowns in negotiation (often arising when one side misgauges the other's situation), victims' lack of knowledge of danger (suppose the pollution is not apparent). Because bargaining will often fail to solve externality problems, attention has been paid to the use of legal rules (particularly liability rules, regulatory requirements, property rights) to accomplish their amelioration.

A very different issue in economic analysis of property law involves public property, which is to say, property like roads and libraries that the government needs to supply because the private sector will not provide them, or not in appropriate quantity. In this connection, the legal system sometimes allows the government to take land through its powers of eminent domain, but requires it to pay compensation for its takings. According to economic analysis, the power of eminent domain may be socially desirable where government would have undue difficulty in making purchases. A standard example is that, in trying to purchase land for a road, government would be likely to be stymied by even a few landowners on the road's planned path (landowners might be tempted to hold out for strategic reasons). The requirement that government compensate for takings cannot be justified as it sometimes is, as an implicit form of insurance against takings; for, in the absence of a compensation requirement, individuals doubtlessly would purchase insurance against takings. A possible advantage of the requirement of compensation for takings is that it may

discourage government from ill-advised takings, though that argument has been criticized.

Another area of property law concerns intellectual property rights: the law of patents, copyright, and trade secrets. The main theme of economic analysis of this body of law is that intellectual property rights have the beneficial effect of spurring innovations, but the detrimental effect associated with high prices and lower than socially desirable sales of goods incorporating the innovations. For example, economic thinking suggests that the copyright doctrine of 'fair use,' permitting for instance limited excerpting from a book for inclusion in teaching materials, might be socially desirable. This excerpting probably does not reduce book sales or the financial motive to author books (indeed, limited excerpting might increase sales of a book because it serves as a form of advertising), whereas the excerpting benefits teaching, because it means that teaching materials can be rapidly assembled without added cost. Most rules of intellectual property law are viewed against the background of their influence on incentives to innovate and on the sale and dissemination of goods embodying innovations.

2.2 Contract Law

A primary issue addressed in economic analysis of contract law is that of contract formation. The basic rule of contract formation is that a contract is legally recognized when and only when both sides have given explicit assent, such as by signing a document. This rule is said to be desirable for two reasons. First, it obviously enables parties to make a contract. Second, it protects parties from becoming bound against their wishes due to their having engaged in negotiations; this protection against unwanted obligations is beneficial because, without it, the negotiations that lead to the making of contracts would be inhibited.

Another issue surrounding contract formation concerns legal duties to disclose information. Economic analysis emphasizes that the social desirability of disclosure depends on the situation. For example, disclosure of a material defect (such as a leaky basement) in the home that the owner seeks to sell tends to be beneficial; for if the buyer knows about the defect, he can take steps to avert harm (avoid storing valuables in the basement or repair it). But should an oil company that learns, through costly investigative effort, that oil lies under a parcel of land, be required to disclose this information to the seller of the parcel? Perhaps not: the effect of such a disclosure obligation would be to make oil companies pay substantially more for parcels of land that they learn are romising, and thus to discourage expensive investigation of the location of oil deposits. Note from the foregoing that the economic analysis of legal rules about disclosure obligations concerns their effects on outcomes and does not derive from a possible moral call to tell the truth.

The most developed aspect of economic analysis of contract law deals with enforcement of contractual agreements. Enforcement is accomplished mainly by requiring parties who commit breach to pay the victims of breach for harm, to pay them 'damages.' One effect of the requirement to pay damages is that it induces contractual performance, which tends to raise the value of contracts to the parties and to society. A less obvious advantage of damage payments is that they constitute an escape hatch that parties can use when contractual performance becomes difficult, for they can breach and pay damages rather than bear very high costs to perform. This escape hatch element of damage payments also raises the value of contracts, as it makes parties more willing to assume contractual obligations. The escape hatch feature of damages for breach is also socially advantageous—it is not socially desirable for parties to perform when the cost of so doing outstrips the benefit to the recipient of performance. These points and others (notably, concerning risk allocation, and incentives to invest) regarding the virtues of payment of damages for breach have been analyzed intensively in the economic literature on contracts.

The orientation of economic analysis of contractual enforcement, through damage payments for breach, is very different from that of traditional legal analysis. Under the latter, damage payments for breach tend not to be regarded as incentives toward performance or as implicit escape hatches. Damage payments are seen primarily as compensation for harm or as proper desert for the wrong of breaking a promise. It should be added that, under the economic view, breach of a contract should not necessarily be identified with breaking a promise. The contracts that are written are not interpreted as detailed promises that parties truly want to be kept, but rather as incomplete promises that are only rough guides for behavior, and that the parties do not want to govern when performance would be very difficult.

Economic analysis of contracts has also been concerned with specific classes of contracts, including principal and agent contracts, insurance contracts, financial contracts, and donative contracts; the literature on some of these contractual contexts is now highly refined.

2.3 Litigation

One aspect of the economic analysis of litigation describes the motive to bring suits in terms of the potential plaintiff's costs of suit, the likelihood of success at trial, and the amount that would be obtained in the event of success.

Another element of the analysis of suit concerns the issue of whether the number of suits is socially

excessive (is there a litigation explosion?) or perhaps socially inadequate. In this regard, it is observed that when a person considers suit, he does not factor in, as a cost to himself, the defendant's legal costs or the state's costs. This indicates that plaintiffs' incentives to bring suit may be socially excessive and thus that suit should be curtailed or barred in some domains. Yet a person considering suit will not usually take into account the deterrent value of the suit—the message that suit will send, which will affect the behavior of others in the future—as well as any additional, wider social benefits. This divergence between the private and the social benefits from suit suggests that in some contexts the number of suits brought might be inadequate, and that public promotion of suit might be desirable.

Given that suit has been brought, the question arises whether the parties will settle their dispute or proceed to trial. The thrust of economic analysis of this question is that settlement is likely when the beliefs of the two sides about the trial outcome are similar, but that trial is likely when the plaintiff is much more optimistic than the defendant. For example, suppose that the beliefs of both sides are identical—each thinks the plaintiff would definitely obtain $100,000 at trial—and that the trial expenses of each would be $10,000. Then the plaintiff should be willing to accept as little as $90,000 (that is, $100,000 minus his litigation costs) in a settlement, and the defendant should be willing to pay as much as $110,000 ($100,000 plus his litigation costs). Hence, there should be room for settlement (any amount in between $90,000 and $110,000). If, however, the plaintiff's estimate of his winnings is much higher, say $200,000, than the $100,000 the defendant expects to pay at trial, the plaintiff would demand at least $190,000, which exceeds the $110,000 the defendant would be willing to pay. Accordingly, trial would be likely.

Economic analysis of litigation has also begun to address topics beyond the general ones of the bringing of suits and of settlement decisions. Among the topics considered are disclosure of information before trial, appeal of trial outcomes, class actions, and alternative dispute resolution.

2.4 Law Enforcement and Criminal Law

An additional area of economic analysis concerns public enforcement of law: the use of enforcement agents (such as police, safety inspectors, auditors) to detect violations of law; and the imposition of penalties for violations.

A theme of the literature on law enforcement is that the magnitude of penalties should be inflated from the level that would be appropriate were detection of violations certain. Inflation of penalties is needed to maintain deterrence, in effect to compensate for the possibility that a violator will not be detected. For example, if a polluting firm that causes harm of $10,000 is detected only a third of the time, then the fine that is imposed when the firm is detected should be not $10,000, but this amount multiplied by three, or $30,000. For, if the fine when the firm is detected is $30,000, the firm's probability-discounted, or average, fine is one-third of $30,000, or $10,000, providing it with incentives not to pollute similar to those that would exist if it paid a certain fine equal to the $10,000 harm.

Another point of emphasis in enforcement literature is that, because law enforcement is expensive, it will often be desirable for society not to spend so much on enforcement as to detect violations with high probability: it may be best, all things considered, to conserve on enforcement resources even though this means that many violators will escape detection. To combat the problem of inadequate deterrence that might accompany a low level of enforcement, multiplied penalties can be applied, as just discussed. Thus, there is appeal in employing an enforcement strategy that involves significant chances of escaping punishment combined with high levels of penalty. However, a problem with this low probability-high penalty strategy is that high penalties may be infeasible. Monetary penalties cannot exceed the assets of violators, which may be quite modest. Also, jail sentences can only be so long, and sanctions that conflict with retributive justice—sanctions that are out of proportion to the gravity of a bad act—might be resisted by the public.

A further issue examined in the economic literature on enforcement is the socially desirable choice between fines and imprisonment as forms of penalty. Here, it is generally said that fines are preferable to prison sentences, because fines do not themselves deplete social resources (but rather transfer command over resources from violators to the state). Whereas, imprisonment does diminish social resources, for prisons are expensive to operate, deprive individuals of their liberty, and prevent individuals from participating in the labor force. Therefore, the economic prescription is that fines be employed as the form of penalty when possible. But when fines cannot be used to deter, because the appropriate fines exceed the assets of violators (the typical robber could not be deterred by the threat of fines, given his level of assets), imprisonment should frequently be employed as the form of penalty. A closely connected point is that when fines cannot be used to deter, imprisonment may be useful as a penalty not only because it may deter, but also because it will incapacitate, that is, prevent individuals from doing further harm while in prison.

The foregoing conclusions and associated ones have been applied to criminal law. It has been suggested that many of the undesirable acts that are punished under criminal law (robbery, murder, rape) have the feature that civil suit and fines would not suffice to achieve adequate prevention of the acts. Therefore,

imprisonment is often necessary to deter and to incapacitate those who commit the acts. Further, the magnitude of sanctions under criminal law has been related to the likelihood of detection of the acts, among other elements. Additionally, various doctrines of criminal law have been interpreted as desirable from the economic perspective. For example, that the sanction for murder committed in the heat of passion is less than that for premeditated murder is said to be rational: the ability to deter acts carried out in the heat of passion is relatively low, implying that it would be a mistake for society to incur the costs of the higher level of sanctions that are imposed for planned acts for which deterrence is more effective. Economic analysis of criminal law is concerned generally with the efficacy and the social costs of enforcement and the imposition of sanctions, and does not usually view punishment as a means of achieving retributive justice or other ideas of desert.

2.5 *Additional Areas of Research*

Economic analysis has been brought to bear on a host of other areas of law. These include many business-related areas of law, such as corporate law, tax law, antitrust law, and bankruptcy law, as well as, increasingly, diverse other areas, such as family law, anti-discrimination law, and constitutional law. Moreover, economic analysis has addressed questions surrounding the role of legislative bodies in formulating and enacting legal rules and the role of courts and regulatory agencies in applying legal rules.

3. *Foundations of, and Criticism about, Economic Analysis of Law*

Economic analysis of law is premised on the general assumptions of the discipline of economics. These assumptions and their relationship to economic analysis of law in particular are sketched here. Certain commonly encountered criticisms of economic analysis of law are also mentioned.

3.1 *Basic Assumptions*

With regard to prediction of behavior, the usual assumption made in economics is, as noted at the outset, that parties are forward-looking and rational. This assumption is sometimes criticized as unrealistic. However, the assumption is usually made with the understanding that, although it is best for predicting central tendencies in behavior, various psychological and cognitive biases also influence behavior, and that these sometimes should be taken into explicit account. For example, the propensity to underestimate certain classes of risk, and thus for individuals not to take proper precautions or to insure adequately against them, has been recognized in economic analysis of law.

With regard to the evaluation of outcomes under the economic framework, the well-being or 'utility' of a person is basic. Economists' conception of utility is completely general and reflects not only the material pleasures of life to a person, but also, for example, the influence on his/her happiness of the treatment of others. From the utilities of individuals, a measure of social welfare is constructed, but there is no single, objective measure of social welfare that analysts study. Thus, utilitarianism is just one among a continuum of measures of social welfare that could be examined. The only significant presumption that is ordinarily made is that the measure of social welfare depends exclusively on the utilities of individuals. This assumption is consistent with concerns for equity in the distribution of utility and wealth, and economists have studied the implications of such concerns in depth.

3.2 *Notions of Fairness and the Law*

As has been seen in Sects. 1 and 2, classic notions of fairness, such as corrective and retributive justice, typically are omitted from the evaluation of legal rules under the economic framework, whereas these notions are traditionally viewed as of great significance to the assessment of the law. The essential reason that, under welfare economics, the notions of fairness are not accorded intrinsic importance is the assumption that they do not directly enter into individual's wellbeing. For instance, whether punishment is in proportion to the seriousness of a crime is ordinarily assumed not to affect individuals' utilities *per se*; rather, punishment may affect individuals' wellbeing through its deterrent or incapacitative effects. Because satisfaction of notions of fairness is presumed not to raise individuals' wellbeing in a direct manner, granting these notions independent weight in the evaluation of outcomes would tend to alter social decisions in ways that lower individuals' wellbeing.

Nevertheless, several qualifications to the last paragraph should be made. First, the assumption that notions of fairness do not matter to individuals may not always be apposite: individuals may have tastes for adherence to notions of fairness (individuals might feel happier if punishments fit crimes). To the extent that that is so, satisfaction of a notion of fairness properly enters into individual wellbeing and thus into social welfare, just as satisfaction of a taste for a material good does. (Notice that the importance of notions of fairness as personal tastes, being contingent on what the tastes of individuals happen to be, is different from the intrinsic importance accorded to conceptions of what is fair and right under deontological philosophical views.) Second, notions of fairness tend to have a desirable functional role (punishment only in proportion to the gravity of bad acts tends to dis-

courage bad acts at relatively low social cost). Thus, advancing notions of fairness and inculcating them in the population (perhaps partly through adoption of legal rules that embody them) may serve to promote social welfare.

3.3 Income Distributional Equity and the Law

It may have been noted in Sects. 1 and 2 that the income distributional effects of legal rules were not mentioned as relevant to their evaluation under welfare economics, even though, as noted in Sect. 3.1, distributional equity does enter into the assessment of social welfare. The reason that the income distributional effects of legal rules are usually not considered in their evaluation is that economic analysis suggests that the income tax system (combined with income transfer programs) is a better means of achieving distributional objectives than the legal system. The income tax system is overtly redistributive, reaches all individuals, and is relatively cheap to administer. The legal system is not well designed to redistribute: it directly affects only those individuals who are involved in litigation; and even among them, the legal system is difficult to employ to effect redistribution, for a given class of litigants is often comprised of individuals with widely varying incomes (consider the class of victims of automobile accidents). Furthermore, the legal system is a very expensive device for transferring income. Hence, according to economic analysis, the legal system should not be used as a tool to achieve distributional goals, and if legal rules turn out to have undesirable distributional effects, these can be remedied through adjustment of the income tax system.

3.4 Economic Explanation of the Law

Finally, a strand of economic analysis should be mentioned claiming that the legal rules that are observed can be explained as those which best advance social welfare. This hypothesis seems attractive at a very gross level of description (for instance, that liability is imposed for causing harm, rather than for doing good, is explainable in the sense that such liability discourages harmful acts); and sometimes the hypothesis is appealing at a fairly detailed level. However, many, if not most, economic analysts hold a nuanced view of the economic rationality of the law, for many legal rules do not have obvious economic rationales, and a number undoubtedly reduce social wellbeing.

See also: Criminal Law and Crime Policy; Disability: Sociological Aspects; Income Distribution; Insurance; Insurance and the Law; Intellectual Property: Legal Aspects; Law and Economics: Empirical Dimensions; Legal Insurance; Litigation; Property: Legal Aspects; Property: Legal Aspects of Intergenerational Transmission; Property Rights; Safety, Economics of; Social Insurance: Legal Aspects; War Crimes Tribunals; Social insurance

Bibliography

Becker G S 1968 Crime and punishment: An economic approach. *Journal of Political Economics* **76**: 169–217

Bentham J 1789 An introduction to the principles of morals and legislation. In: Bentham J (ed.) 1973 *The Utilitarians*. Anchor, Garden City, New York

Bouckaert B, De Geest G (eds.) 2000 *Encyclopedia of Law and Economics*. Elgar, Cheltenham, UK

Calabresi G 1970 *The Costs of Accidents*. Yale University Press, New Haven, CT

Coase R 1960 The problem of social cost. *Journal of Law and Economics* **3**: 1–44

Cooter R, Ulen T 1997 *Law and Economics*, 2nd edn. Addison-Wesley, Reading, MA

Kaplow L, Shavell S 2001a Economic analysis of law. In: Auerbach A, Feldstein M (eds.) *Handbook of Public Economics*. Elsevier, Oxford, UK

Kaplow L, Shavell S 2001b Fairness versus welfare. *Harvard Law Review* **114**: 961–1388

Miceli T 1997 *Economics of the Law*. Oxford University Press, New York

Newman P (ed.) 1998 *New Palgrave Dictionary of Economics and the Law*. Stockton, New York

Polinsky A M 1989 *An Introduction to Law and Economics*, 2nd edn. Little Brown, Boston

Posner R A 1972 *Economic Analysis of Law*, 1st edn. Little Brown, Boston

Posner R A 1998 *Economic Analysis of Law*, 5th edn. Aspen, New York

Shavell S 1987 *Economic Analysis of Accident Law*. Harvard University Press, Cambridge, MA

S. Shavell

Law and Economics: Empirical Dimensions

Since the 1960s, economics has extended its reach beyond the marketplace to study new subjects such as the family, education, discrimination, politics, and even the law. Empirical work plays an essential role in this undertaking, both to assess the predictive power of a given economic model with respect to the examined question, and to provide guidance as to how to amend the economic model to improve future analyses. Empirical work has often served as a touchstone for multidisciplinary discourse on law because the social sciences share common statistical methods for evaluating empirical data, and because such work has commonly suggested the utility of amending the neoclassical economic model to address issues studied in other disciplines, including sociology, political science, criminology, and psychology.

1. Economic Analysis

Economics first developed as a discipline in the seventeenth century, in close association with the accompanying social sciences of political science and sociology. During this early period of development, the social sciences enjoyed considerable overlap in examined problems and methodology. Although some works from this time can be identified as being distinctly 'economic,' early social scientists such as Adam Smith, Karl Marx, and John Stuart Mill hypothesized about economic, social, and political institutions with equal skill and aplomb. This close association between the social sciences continued into the beginning of the twentieth century with the work of institutional economists such as John R. Commons, and economically oriented sociologists such as Emile Durkheim and Max Weber. Indeed, Weber, one of the most distinguished sociologists of his time, held several academic positions as an economist.

However, early in the twentieth century, economics began to separate from the other social sciences in two important ways. First, economics began to diverge in terms of its examined subject matter. Economists began to focus more on studying the machinations of the market, leaving subjects such as social relationships and political economy to the other disciplines. Since the work of Gary Becker (1968, 1976), however, this trend has been reversed, so that now economists study a wide variety of social and political phenomena, including the law. Second, economics began to diverge in terms of its adopted methodology. Although all of the social sciences adopted modern statistical methods for analyzing data as the primary focus of empirical work, economists also undertook to develop a 'positive science' of economics under the neoclassical model.

1.1 Positive Economics

Since early in the twentieth century, most economists have endeavored to develop economics as a positive science analogous to the physical sciences. As the name suggests, the ultimate goal of a 'positive science' is to produce theories that describe observed phenomena accurately, rather than offer a normative evaluation of the desirability of the phenomena. Such theories are devised by making assumptions about the world that abstract the essential features of the examined problem, and then using logic to deduce predictions about the as yet unobserved events. The validity of the theory is then confirmed through empirical work, which examines the predicted events and tests the statistical significance of results consistent with the theoretical model. This deductive process, working from theory to data, can be distinguished from the inductive process used by many other social scientists, who tend to analyze the data before inductively deriving their theories. Accordingly, for the majority of the twentieth century, most economists aspired to conduct economics as a positive science that deduces empirical predictions from theory and then tests those predictions against data.

1.2 The Neoclassical Model

In the development of economics as a positive science, the dominant paradigm has been the 'neoclassical model.' The neoclassical model consists of a set of assumptions that economists make in analyzing problems. This set of assumptions achieved paradigmatic dominance in economics because they proved useful in analyzing market behavior, and because they were tractable mathematically. Under the neoclassical model, economists assume generally that: the appropriate unit for analysis is the individual consumer or firm, individuals and firms rationally choose among available alternatives according to their preferences to select the option that maximizes utility or profits, people's preferences are exogenous to the model, and transaction and information costs are zero. Economists have a well-defined notion of 'rationality'; they assume that people have preferences with respect to all possible states of the world, that these preferences are reflexive and transitive, and that the people will select the option they most prefer. The assumption of free information is usually referred to as the assumption of 'perfect information.' It has deceptively broad-ranging implications in that it implies that each person knows the quality of each good and the price of all alternatives, and indeed has the knowledge necessary to enter the market and produce the good if the market for the good enjoys supercompetitive profits. The assumption of zero transaction costs also has potentially far-reaching implications since, as Ronald Coase (1960) has pointed out, in a world of zero transaction costs, everything happens at once.

The conduct of economics as a positive science using the neoclassical model tends to separate the work of economists from that of other disciplines. The adoption of the perspective of the individual as the unit of analysis, and the assumption that preferences are exogenous, preclude consideration of the effect of social norms and the process of socialization that are so important to many sociological analyses. The construction of the concept of individual utility based merely on an ordinal preference ranking was undertaken with the purpose of avoiding some of the stickier problems in psychology, thus making the concept 'psychology-free.' The economists' well-defined concept of rationality is much narrower than that used in any other discipline. Moreover, since the other social sciences rely more on inductive logic, working from data to theories, they generally do not make any *a priori* assumptions about what constitutes rational choice. Accordingly, some sociologists have argued that they study rather than assume human rationality.

1.3 The Importance of Empirical Work

The assumptions of the neoclassical model are clearly unrealistic, and this fact raises the importance of empirical work in economics. Every analysis of a problem is necessarily an abstraction from reality, ignoring some of the complexities of the real word to focus on others. The art of modeling is to know which simplifying assumptions one can make and still capture the essence of the problem. Of course, an important test of how well a theory captures the essence of a problem is how well it fits the available data. Milton Friedman (1953) has argued that the realism of a model's underlying assumptions is of little importance as long as the model yields good predictive results. Indeed, Friedman has gone so far as to argue that, paradoxically, the most important hypotheses have the most unrealistic assumptions, because a hypothesis is important if it abstracts the crucial elements of a phenomenon in such a way as to explain much of the phenomenon by assuming little. Other economists have argued that there are purposes to modeling beyond mere prediction, such as explaining the phenomenon or providing a catalog of knowledge for future work. As a result, they have argued that there are reasons for considering the realism of economic assumptions that go beyond the simple predictive value of the model.

2. The Economic Analysis of Law

Beginning with the pioneering work of Gary Becker (1968, 1973), economists have sought to broaden the scope of their inquiry to all aspects of society. This movement to broaden the scope of economic analysis has brought economics into a variety of subject matters previously studied by other disciplines, including the family, education, discrimination, politics, and even the law.

2.1 The Modern Law and Economics Movement

The modern law and economics movement had its origin in the work of Gary Becker, Ronald Coase and Guido Calabresi. Becker (1968) used simple principles of labor economics to analyze the criminal's decision of whether to engage in crime, and derived implications for optimal penalties and optimal enforcement of the law. Coase's (1960) work established the proposition known as the 'Coase Theorem', showing that, in a world of zero transaction costs, the distribution of resources is invariant to the rule of law. However, Coase later made clear that he considers his work concerning how to minimize the effects of transaction costs in assigning the rule of law under the more realistic assumption of positive transaction costs to be more important. Calabresi (1970) wrote a classical study analyzing the costs of accidents and suggesting ways in which tort law might be formulated to minimize those costs. After these initial contributions, a notable and extensive body of work in the economic analysis of law was, and continues to be, authored by Richard Posner, who has applied the principle of wealth maximization to a plethora of legal problems, explaining either how the law can be understood as promoting wealth maximization, or how it might be amended to better achieve that objective (see, for example, Posner 1998). As will be discussed in detail below, in more recent years, work in the economic analysis of law has taken a variety of forms, from the traditional neoclassical analyses to socioeconomic and behavioral economic analyses.

2.2 Empirical Work in Law and Economics

Empirical work in the economic analysis of law has been slow to follow these theoretical advances. The reasons for the relative paucity of empirical work are several and varied, including a lack of training in empirical methods among legal academics, a lack of resources dedicated to such work, a lack of prestige and incentives attached to such work, and perhaps even some fear of exposure to falsification through replication on the part of law and economics scholars.

Despite these impediments to empirical work in law and economics, a number of notable studies have been done directly on issues important to the economic analysis of law. One of the earliest significant empirical works in law and economics was Isaac Ehrlich's (1975) work showing a deterrent effect from capital punishment. Ehrlich's work prompted enormous academic and public debate on the subject, and his results were later disputed by other scholars. A number of scholars have done empirical work in an attempt to verify or disprove the fundamental implication of Coase's work that, when transaction costs are negligible, the distribution of resources is invariant with respect to the rule of law. Among these works, perhaps the most notable for the purposes of this essay is Robert Ellickson's (1986) on the distribution of cattle in Shasta County, California, under alternative rules of free range and non-free range. Ellickson found that Coase was correct that changes in the rule of law did not affect the distribution of cattle; however, Ellickson attributed this finding more to the sociological reason that neighbors had informal means of dealing with animal trespasses that were not influenced by changes in the law, than Coase's reasoning that neighbors would just bargain around inefficient assignments of entitlements. Kip Viscusi (1983, 1992) has also made important empirical contributions in the economic analysis of legal issues through his analysis of the valuation of life and limb and the relative magnitude of various risks. Viscusi's work is important in

evaluating the efficiency of tort law, workers' compensation laws, and government regulation of various workplace and environmental risks.

3. *Empirical Work in Law and Economics as a Touchstone for a Multidisciplinary Discourse on Law*

It has become apparent that empirical work in law and economics often serves as a touchstone for interdisciplinary discourse on law. This is true for several reasons. First, economics shares the same accepted statistical methods for analyzing data with most other disciplines. As a result, as economists have expanded the scope of their research to areas (such as law) that traditionally have been studied by other disciplines, they have found that their empirical work is relevant to, and can be usefully informed by, empirical work already done in the subject area by other disciplines. Second, as economists have done empirical work in the economic analysis of law, or as they have borrowed from the empirical work of other disciplines, many have found reasons to relax some of the assumptions of the neoclassical model in their analysis of the law. The traditional economic model of individual rational choice does not always capture adequately the essence of legal problems in the way that it did in the study of atomistic markets, and as a result economists have sought to relax some of the assumptions of the neoclassical model in order to produce theoretical models with greater predictive or explanatory power.

By relaxing some of the assumptions of the neoclassical model, economists have found that their work becomes less distinct from that of theoretical work in other disciplines. This is true not only for the reason that such economic work is inductive rather than deductive in nature, working from the data to theory, but also because relaxation of some of the assumptions of the neoclassical model allows for the consideration of work from other disciplines. In particular, some law and economics scholars have found connections between their work and the work of sociologists who study law, while other law and economics scholars have made use of empirical work out of the behavioral sciences for their work in the economic analysis of law.

3.1 *Sociological Influences in the Economic Analysis of Law*

Some economists have drawn on the work of sociologists in their economic analysis of law. Indeed, in recent years there has been a growing interest among law and economics scholars in the influence of social norms and status production on people's behavior, and the role of law in the socialization process. In order to examine these phenomena, the economist usually has to relax one or more of the traditional assumptions of the neoclassical model. For example, to examine the influence of social norms or status production on people's behavior, the economist might adopt a perspective of analysis that takes into account group or community interests, rather than just individual interests. Similarly, in examining the role of law in the socialization process, some economists have found it useful to relax the assumption that people's preferences are determined exogenously. With the relaxation of these assumptions, it is also necessary generally to adopt an objective function other than individual utility, wealth or profit for maximization in the economic model—for example, net social benefits or social welfare.

The intensity of economists' interest in social norms in the 1990s has been rather remarkable. No less than four symposia on the subject have been published in major US law reviews since 1996 (Ellickson 1998). Although some very interesting work in the study of norms from the economic perspective has been done by Richard McAdams (1997) and Eric Posner (1996), perhaps the seminal work on the subject was the work previously alluded to by Robert Ellickson (1986). As previously mentioned, Ellickson undertook an empirical study of the effect of changes in the legal rule from open range to closed range on the allocation of cattle and fencing resources to test Coase's famous hypothesis that, in low transaction cost cases, the allocation of resources would be invariant to the rule of law because the affected parties would just bargain around any inefficient allocation of entitlements. To study this problem, Ellickson surveyed land owners in Shasta County, California, as to the number of animal trespass cases they had experienced under different legal regimes, and how those incidents were resolved. Ellickson found that the allocation of resources was indeed invariant to the legal rule, as predicted by Coase.

However, Ellickson found that this result, obtained not because transaction costs were low and the parties had bargained around the law to reach an efficient allocation of resources, but instead because resort to the law was costly, and as an alternative the affected parties were committed to a norm of cooperation with their neighbors, regardless of the rule of law. As a result, Ellickson concluded that work in law and economics would benefit from consideration of previous work on such questions in sociology, and the law and society movement. Taking his own suggestion to heart, Ellickson has used empirical work from other disciplines on norms of 'fairness' and 'civic responsibility' to analyze a variety of legal problems including rent control, the doctrine of unconscionability, the implied covenant of good faith and fair dealing, the duty to rescue, political and judicial processes, and the First Amendment.

Law and economics scholars have also demonstrated an interest in using models of individual and

group status production to analyze the efficacy of antidiscrimination laws. The seminal work in this regard is that of Richard McAdams (1995). Although Gary Becker's (1957) neoclassical analysis of discrimination based on a distaste for association has also been used to analyze antidiscrimination laws, this model has never been empirically satisfying, since it predicts that firms that discriminate will be put at a competitive disadvantage and driven from the marketplace, even in the absence of antidiscrimination laws. Drawing on work in sociology, McAdams developed a model of discrimination based on people's desire to produce inter and intragroup status. According to McAdams, whites discriminate against lower black group status and raise white group status relative to blacks. Intragroup status is used by whites to compensate individuals for costs they incur in raising whites' group status relative to blacks, for example by forgoing profits they would make in nondiscriminatory associations with blacks. McAdams' model is empirically more satisfying than Becker's, because it explains how discrimination can persist in the marketplace despite individual costs associated with discrimination. McAdams' model also explains nicely the existence and function of Jim Crow laws in the old South to reinforce group status production and prevent individual defection from this strategy. The status production model also yields very different policy prescriptions from those associated with Becker's neoclassical model since, under the status production model, antidiscrimination laws are no longer a superfluous addition to market pressures, but instead can be a useful tool to discourage wasteful competition in group status production.

A final area in which the economic analysis of law has shown connections to work in sociology is the analysis of the role of law in the process of socialization. In this area, the present author has written an economic analysis of the criminal law as a preference-shaping policy (Dau-Schmidt 1992). If one relaxes the neoclassical assumption that preferences are exogenous to the model, and instead assumes that the criminal law is intended to discourage preferences for criminal behavior, as well as to raise the costs of engaging in such behavior, then one can explain many of the fundamental doctrines of the criminal law. Under this analysis, the requirement of intent in the criminal law can be understood as part of the attempt to identify deviant preferences that should be punished to be discouraged. After all, it is only when someone intends the harm that they cause that we know that they desired to bring it about. Similarly, we criminalize attempts (in the absence of actual harm) to ideality and punish deviant preferences in need of modification. Indeed, the preference-shaping model of criminal law allows one to distinguish criminal liability from ordinary tort liability, on the basis that criminal liability is intended to modify both preferences and rewards, whereas ordinary tort liability is intended merely to modify rewards. The preference-shaping model lends itself naturally to integration with work from sociology, criminology, and psychology, because it raises questions concerning how we determine which preferences are socially deviant, and how best to discourage such preferences in people.

3.2 Behavioral Law and Economics

Other economists have drawn on the empirical work of behavioral psychologists in their analysis of law. In 'laboratory' experiments, behavioral psychologists such as Amos Tversky and Daniel Kahneman (Tversky and Kahneman 1982) have established that the human decision-making process typically varies in predictable ways from the simple model of rational choice assumed in the neoclassical economic model. In particular, they have established that people commonly exhibit 'bounded rationality' in that they can only process a finite amount of information and thus must rely on rules of thumb to help them make decisions, 'bounded willpower' in that they sometimes do things that are clearly not in their long-term interests, and 'bounded self-interest' in that they care about other people and whether the treatment in their relationship is reciprocally 'fair.' Tversky and Kahneman have also established that people evaluate outcomes based on the change they represent from an initial 'reference point,' weighting potential losses from the reference point more heavily than potential gains. This finding is referred to as 'prospect theory' in the literature, and it can be exploited to manipulate people's decision-making process by suggesting the initial reference point to 'frame' possible choices as gains or losses.

As with the influence of sociology on the economic analysis of the law, there has been a boom in the number of scholars who seek to use 'behavioral economics' in their analyses of the law. Early proponents of such work include Thomas Ulen (1989) and Robert Ellickson (1989), but they have since been joined by others including Russell Korobkin (see Korobkin and Ulen 2000), and Christine Jolls, Cass Sunstein, and Richard Thaler. These scholars argue that amendment of the assumptions of the neoclassical model concerning individual rational choice according to the empirical findings of behavioral economics is necessary to improve the predictive power of the economic model. Their critique is fundamental, in that it goes back to basic assumptions about the purposes of law, and how people will react to and enforce the law. For example, Jolls et al. (1998) have argued that the Coase theorem is often empirically false, because empirical work in behavioral economics suggests that people sometimes develop attachments to objects they possess, and this 'endowment effect' may mean that people must sometimes be paid more to part with an entitlement than they would pay to possess the

entitlement. Accordingly, people will not bargain around the assignment of entitlements in the way described by Coase. Similarly, Korobkin has argued that information heuristics will affect the efficacy of tort damages in providing efficient incentives to take care. Because work in behavioral economics shows that people are often overly optimistic in that they undervalue the chances of something bad happening, and overvalue the chances of something good happening, this suggests that, even if tort damages are set at efficient levels, people will not account adequately for the chance of accident and so will not take efficient care. Paradoxically, Jolls et al. (1998) have argued that, although potential tort feasors may be overly optimistic, the juries that later try them will suffer from 'hindsight bias' and overestimate the probability of accident, merely because an accident has in fact happened, thus too readily finding tort defendants liable. Of course, whether potential tort feasors anticipate hindsight bias on the part of later jury members, and which effect dominates in their decision-making process, are questions for future empirical work.

See also: Behavioral Economics; Decision and Choice: Economic Psychology; Decision Biases, Cognitive Psychology of; Decision Making: Nonrational Theories; Decision Making, Psychology of; Decision Research: Behavioral; Economics, History of; Law and Economics; Problem Solving and Reasoning, Psychology of; Tversky, Amos (1937–96)

Bibliography

Becker G 1957 *The Economics of Discrimination*. University of Chicago Press, Chicago

Becker G 1968 Crime and punishment: An economic approach. *Journal of Political Economy* **76**: 169–217

Becker G 1976 *The Economic Approach to Human Behavior*. University of Chicago Press, Chicago

Calabresi G 1970 *The Costs of Accidents: A Legal and Economic Analysis*. Yale University Press, New Haven, CT

Coase R 1960 The problem of social cost. *The Journal of Law and Economics* **3**: 1–44

Cotterell R 1995 Sociological interpretations of legal development. *European Journal of Law and Economics* **2**: 347–59

Dau-Schmidt K G 1991 Relaxing traditional assumptions and values: Toward a new multidisciplinary discourse on law. *Syracuse Law Review* **42**: 181–95

Dau-Schmidt K G 1992 An economic analysis of the criminal law as a preference shaping policy. *Duke Law Journal* **1990**: 1–38

Dau-Schmidt K G 1997 Economics and sociology: the prospects for an interdisciplinary discourse on law. *Wisconsin Law Review* **1997**: 389–419

De Geest G 1995 Toward an integration of economic and sociological approaches. *European Journal of Law and Economics* **2**: 301–8

Ehrlich I 1975 The deterrent effect of capital punishment: A question of life and death. *American Economic Review* **81**: 397–417

Ellickson R C 1986 Of Coase and cattle: Dispute resolution among neighbors in Shasta County. *Stanford Law Review* **38**: 623–87

Ellickson R C 1989 Bringing culture and human frailty to rational actors: A critique of classical law and economics. *Chicago-Kent Law Review* **65**: 23–55

Ellickson R C 1998 Law and economics discovers social norms. *The Journal of Legal Studies* **27**: 537–52

Friedman M 1953 The methodology of positive economics. In: Friedman M (ed.) *Essays in Positive Economics*. University of Chicago Press, Chicago pp. 1–38

Heise M 1999 The importance of being empirical. *Pepperdine Law Review* **26**: 807–34

Jolls C, Sunstein C R, Thaler R 1998 A behavioral approach to law and economics. *Stanford Law Review* **50**: 1471–550

Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263–92

Korobkin R, Ulen T 2000 Law and behavioral science: Removing the rationality assumption from law and economics. *California Law Review* **88**: 1051

Macneil I R 2000 Other sociological approaches. In: Bouckaert B, De Geest G (eds.) *Encyclopedia of Law and Economics*. Edward Elgar, Cheltenham, UK

McAdams R 1995 Cooperation and conflict: The economics of group status production and race discrimination. *Harvard Law Review* **108**: 1005–89

McAdams R 1997 The origins, development and regulation of norms. *Michigan Law Review* **96**: 338–433

Posner E 1996 The regulation of groups: The influence of legal and nonlegal sanctions on collective action. *University of Chicago Law Review* **63**: 133–97

Posner R 1998 *Economic Analysis of Law*, 5th edn. Aspen Law and Business, New York

Tversky A, Kahneman D 1982 Judgment under uncertainty: Heuristics and biases. In: Tversky A, Slovic P, Kahneman D (eds.) *Judgment Under Uncertainty*. Cambridge University Press, Cambridge, UK

Ulen T 1989 Cognitive imperfections and the economic analysis of law. *Hamline Law Review* **12**: 385–410

Viscusi W K 1983 *Risk by Choice: Regulating Health and Safety in the Workplace*. Harvard University Press, Cambridge, MA

Viscusi W K 1992 *Fatal Tradeoffs: Public and Private Responsibilities for Risk*. Oxford University Press, New York

K. Dau-Schmidt

Law and Everyday Life

Law and everyday life have been defined historically in opposition to one another. Everyday life has typically denoted that sphere of social action characterized by being familiar, routine, pragmatic, and particular: the world of work, play, family, and leisure. By contrast, law has been conceptualized as formal, timeless, general, and transcendent. In the context of this opposition, law was understood to radiate downward toward everyday life to resolve conflict and to impart order. Everyday relationships, interactions and practices were thought to constitute the objects upon which law operated with varying degrees of success.

Alexis de Tocqueville articulated such a view of law and the everyday over two hundred years ago:

> The language of the law thus becomes, in some measure, a vulgar tongue; the spirit of the law, which is produced in the schools and courts of justice, gradually penetrates their walls into the bosom of society, where it descends to the lowest classes, so that at last the whole people contract the habit and tastes of the judicial magistrate. (de Tocqueville 1959, pp. 2–20).

During the twentieth century, the boundary separating law from everyday life has become increasingly blurred. With the blurring of this boundary, the presumed direction of effect from law ('the schools and courts of justice') downward to 'society' has likewise been challenged. Increasingly, sociolegal scholars understand law and everyday life to be mutually constituted in a dynamic and emergent process wherein law is produced and applied in the everyday domain of social action. At the same time, the practices, relationships and meanings that make up the everyday are themselves, at least in part, constructed by law, or legality.

1. *Legal Realism: The Everyday Life of the Law*

In the early twentieth century, a perspective called legal realism rejected the prevailing formalist perspective on law (see *Legal Realism*; *Legal Formalism*). Whereas legal formalism saw the law as having an internal logic that led to determinate outcomes, legal realism posited that law is more than the rules and pronouncements contained in judicial decisions or statutes. Legal realists understood that legal rules were not simply and mechanically applied by judges, lawyers, police and regulators. They proposed to account for the gap between 'law on the books' and 'law in action'; that is, between the law as envisioned by its makers, and law as it is actually enforced on a daily basis.

Legal realism taught us that, in entering the realm of lived experience, legal rules are articulated within particular social and historical contexts that shape law's effects. This proposition authorized an examination of those contexts and constraints, thus creating an opening to the everyday in the study of law.

1.1 *Law First Perspective*

Legal realism—even while it acknowledged the life, and not merely the spirit, of the law—assiduously maintained the boundary between the law and everyday life, retaining what has been called a 'law first' orientation (Sarat and Kearns 1993). Karl Llewellyn and E. Adamson Hoebel, in their classic book *The Cheyenne Way* (Llewellyn and Hoebel 1941), proposed that law performed a dual function in ordering everyday life. Frequently, the law's prescriptions operated beneath the level of notice or consciousness 'to make [people] go round in more or less clear ways.' In this regulative function, the law created routines and norms that structured people's practices and expectations. At times, however, the so-called drift of everyday life produced divergent interests, expectations, and norms that upset the orderliness of social life. Llewellyn and Hoebel referred to these instances as 'trouble cases' and proposed that a central function of the law is to process these cases to restore social order. 'Law,' they wrote, 'exists for the breach of law and has a major portion of its essence in the doing of something about such a breach' (1941, p. 20). How law responds to such breaches—how it selects cases of trouble, in what way it redresses these situations and with what consequences—ultimately comes to comprise the law. Accordingly, Llewellyn and Hoebel recommended empirically examining the selection, processing, and resolution of trouble cases as a way of elucidating and understanding the law. Research in this tradition focused on the everyday work of legal professionals, examining such processes as judicial decision-making, prosecutorial discretion, or police arrest practice within specific organizational and political contexts.

When the everyday world was examined within this tradition, it was only at the boundary that separated law from society, in those settings where citizens brought their disputes, claims and grievances to the attention of legal professionals. Social scientific research examined, for instance, the sorts of claims and awards made in small claims courts (Yngvesson and Hennessey 1975), or the advice given in lawyer's offices (Sarat and Felstiner 1995). In short, the everyday was included in these analyses of law only in those instances when its familiar routines broke down to the point the law was mobilized to restore order, to redress an injury or to protect a right.

2. *Everyday Life vs. the Law*

This 'law first' perspective shifted somewhat as scholars began to consider the social processes and institutions that preceded law. Whereas Llewellyn and Hoebel depicted the law as intervening in social life to create and restore order, Paul Bohannon (1968) presented law as emerging out of social life. He claimed that law resulted from a double institutionalization of customary practices. This process of formalization granted to custom a force and legitimacy that eventually led to law having a life and dynamic of its own. Hence, according to Bohannon, law did more than simply reflect the everyday; it interacted with social institutions. These interactions were a source of social change.

The law first perspective was also displaced in research that examined citizens' decisions to turn to the law. Studying the daily enforcement practices of legal agents revealed the extent to which most Western

legal systems are reactive; that is, they rely on citizens to mobilize the law. Obviously, this is true in the case of civil law, where citizens seek legal remedies for private disputes. It is also true in regard to criminal law, however. In a classic study of police arrest decisions, for instance, Donald Black (1971) found that approximately 87 percent of all police–citizen encounters were initiated by citizens, either in the role of victims or witnesses. The crucial role of citizens in mobilizing both civil and criminal law directed attention to how citizens interpret situations and make the decision to turn to law.

Felstiner et al. (1980) developed a model of the interpretive process that leads up to a citizen's decision to seek legal remedies. To begin, problems and situations (potential 'trouble cases') are not self-evidently injurious or legal. Therefore, Felstiner et al. refer to the first stage in this process as 'naming,' or defining a situation or event as injurious. 'Blaming' refers to the identification of a party who is responsible for the problem. Finally, 'claiming' entails bringing the grievance or complaint to a third party (which may or may not be a legal body) to obtain a remedy. Although the model postulates an individual decision maker, it stimulated interest in the cultural meanings and values that might compete with interpretations favoring legal action at any point in the process. In considering these matters, sociolegal research extended its analyses to the communities and cultures that exist beyond the legal arena.

Researchers were particularly interested in those instances where the law was rejected and legal protection or remedy was not sought. For instance, in a classic article entitled 'Non-contractual Relations in Business: A Preliminary Study' (Stewart Macaulay 1963), found that law suits for breach of contract are relatively rare between businessmen. The paramount reason for this included the availability of effective non-legal sanctions that operated within the business community such as gossip, and the giving or withholding of advice or expertise. In general, the businessmen's reluctance to use the law in lieu of these alternatives was due to the mutual desirability of maintaining ongoing relationships, and the sense that invoking the formal law would interfere with that goal. More recently, Robert Ellickson (1991) found that neighboring cattle ranchers in California rarely invoke the law in disputes involving trespass. Like Macaulay's businessmen, the ranchers, according to Ellickson do not want, or necessarily need, the law to achieve order.

Carol Greenhouse (1986) studied a very different community and found a similar pattern of law avoidance. She conducted an in-depth ethnography of a small Baptist community in Georgia and found that the religious commitments of the members of this community tended to suppress disputes and steer members away from secular legal processes in the event that disputes did occur. Robert Kidder (1990) studied other 'law-avoiding' societies, including the Amish and the Japanese, and similarly discovered in each case an incompatibility between the community's ideological or spiritual commitments and law. The research on law avoidance thus served to correct the sense that law was paramount in ordering relationships in modern societies. By proposing that law is only one of variously competing normative orders available to citizens for achieving orderly relationships, this research further underscored the need for a serious consideration of everyday lives, relationships, and beliefs of citizens.

Although this research abandoned an exclusive or primary focus on formal law—considering informal community norms and practices as well—it maintained a sense that law and everyday life are distinct and relatively autonomous from one another. Indeed, by focusing on law avoidance, the relationship between the formal law and the informal norms of everyday life was depicted as not merely distinct but oppositional. Subsequent research and theory has challenged the idea that law and everyday life are autonomous or oppositional, suggesting instead that the two are integrated.

3. *Law in Everyday Life*

Since the 1980s, sociolegal scholars have begun to examine everyday life not simply as an object of legal action or a reservoir of potential legal cases, but as a domain in which law circulates, is appropriated, invoked, and produced. This view of the law rejects the notion of law as a mere instrument to be used or avoided by citizens for achieving social or private ends. Law and the everyday are seen as mutually constitutive; that is, literally 'making up' each other in an ongoing dynamic process.

For instance, Barbara Yngvesson (1989) has argued that the invention of law is not a one-way process. Law, she argues, is invented by citizens in local struggles over the meaning of fundamental cultural symbols such as 'community,' 'neighborhood,' or 'rights.' As citizens act on their understandings, law 'emerges in the exchanges and the struggles of everyday people, in neighborhood fights and lovers' quarrels, in the framing of claims and complaints to the police, the court, neighborhood justice and other agencies, and in the response to these agencies to such claims (p. 1709).

Citizens do not, in other words, simply mobilize the police or go to court with their preconstituted legal cases. Before the formal legal apparatus has been invoked, the cases have been fashioned out of the popular legal culture circulating in everyday life. The results of that fashioning come to constitute which cases are processed, decided, and contested in formal legal settings. Rather than imagining law as existing apart from social relations, or conceiving of it as

produced solely by groups of powerful law 'makers' (such as legislatures), in this constitutive view, law is understood as emerging from the 'bottom up' as a continuing production of practical reason and action. As David Engel (1993a) writes,

> Law, despite its apparent claim to "self-totalization," is dependent on everyday life to give meaning to its central concepts (what is "reasonable," "customary," "excessive," or "appropriate"?), to root its abstract rules and principles in human understanding, and to produce implementation, compliance, and judgment. It is misleading to regard law as capable of existence apart from or in opposition to everyday life. (p. 125)

Of course, everyday life is fundamentally dependent upon law as well. Even when we act in situations that seem devoid of legal meaning, or that involve a self-conscious avoidance of law, we do not escape the constitutive power of law. The law creates roles, relationships, and obligations. It 'names the world.' To routinely enact the roles of wife, doctor, or consumer, or any social role, one tacitly implicates the law. In fact, the appearance of routine is in no small measure a consequence of those unspoken legal meanings.

This view of law and everyday life would take issue with the conclusions of research such as Macaulay's and Ellickson's. This research, in demonstrating the absence of law in the ongoing relationship between businessmen or cattle ranchers, fails to note that the very roles and relationships in which their subjects act are legal creations. Moreover, the concepts of private property or profit, for instance, draw upon legally produced and sanctioned meanings. They also fail to take into account the effect of both parties' knowing that the law exists as an alternative if extra-legal sanctions fail. As Sarat and Kearns (1993) have written, 'one is apt to catch more than a fleeting glimpse of law's presence in the putatively alegal responses' (p. 47).

Attempting to catch such a fleeting glimpse of law has led to a broadening of subjects and sites of research. One example of research that has systematically explored the presence of law in everyday life is Hendrik Hartog's analysis (Hartog 1993) of the eighteenth-century diary of Abigail Bailey. Hartog's reading of Abigail Bailey's diary reveals how this woman, over the course of many years, struggled to make sense of her marriage, her husband's sexual abuse of their daughter, and their separation and eventual divorce, as well as her own religious beliefs regarding her duties as a wife and mother.

Recently, the recognition that law and everyday life are mutually constitutive has led to an interest in legal consciousness, ideology, and resistance (see *Legal Culture and Legal Consciousness*). The study of these topics is animated by an interest in how citizens imagine and use law. This would include how and when citizens mobilize formal law, but also how they may invoke, interpret, avoid, and deploy law and legal concepts in daily interactions even when no formal or official actors or rules are involved.

For example, David Engel's work on the intersection of law and everyday life reflects this interest in legal consciousness, ideology, and resistance. In his study of the Education for All Handicapped Children Act of 1975 (Engel 1993b), he rejects a traditional analysis that would have traced the ways in which this legislation unilaterally constrained or defined the educational experiences of children, the meaning of their physical conditions, or the interactions that occurred between parents and educators. Instead, he examines how this legislation and everyday life within this community are interdependent and mutually shaped this social domain. At times, Engel notes, the law reinscribed 'commonsense' images of disability and difference, constituting the meaning and content of those social categories as they operated within the community.

In examining the interpenetration of law and everyday life, Ewick and Silbey (1998) have suggested that the concept of legality be used to augment the concept of formal law. By legality they refer to the meanings, sources of authority and cultural practices that are commonly recognized as legal, regardless of who employs them, or for what ends. In this rendering, people may invoke and enact legality in ways neither approved nor acknowledged by the law. Legality operates as both an interpretative framework and a set of resources with which and through which the social world (including that part known as the law) is constituted. In sum, they conceive of legality as an emergent structure of social life that manifests itself in diverse places, including both formal institutional settings and the everyday lives of groups and communities.

See also: Conflict and Conflict Resolution, Social Psychology of; Dispute Resolution in Economics; Disputes, Social Construction and Transformation of; Justice and its Many Faces: Cultural Concerns; Legal Realism; Legal Systems: Private; Problem Solving (Everyday), Psychology of

Bibliography

Black D 1971 The social organization of arrest. *Stanford Law Review* **23**: 1087–111

Bohannon P 1968 Law and legal institutions. *Encyclopedia of the Social Sciences* 9. Macmillan, New York

de Tocqueville A 1959 *Democracy in America*. Oxford University Press, London

Engel D M 1993a Law in the domain of everyday life. In: Sarat A, Kearns T R (eds.) *Law in Everyday Life*. University of Michigan Press, Ann Arbor, MI

Engel D M 1993b Origin myths: Narratives of authority, resistance, disability and law. *Law and Society Review* **27**(4): 785–826

Ellickson R C 1991 *Order Without Law: How Neighbors Settle Disputes*. Harvard University Press, Cambridge, MA
Ewick P, Silbey S S 1998 *The Common Place of Law: Stories From Everyday Life*. University of Chicago Press, Chicago, IL
Felstiner W L F, Abel R L, Sarat A 1980–81 The emergence and transformation of disputes: Naming, blaming and claiming … . *Law & Society Review* **15**(3–4): 631–54
Greenhouse C J 1986 *Praying for Justice: Faith, Order, and Community in an American Town*. Cornell University Press, Ithaca, NY
Hartog D 1993 Abigail Bailey's coverture: Law in a married woman's consciousness. In: Sarat A, Kearns T R (eds.) *Law in Everyday Life*. University of Michigan Press, Ann Arbor, MI
Holmes O W 1897 The path of law. *Harvard Law Review* **10**: 457–78
Kidder R L, Hostetler J A 1990 Managing ideologies: Harmony as ideology in Amish and Japanese societies. *Law and Society Review* **24**(4): 895–922
Llewellyn K N, Hoebel E A 1941 *The Cheyenne Way; Conflict and Case Law in Primitive Jurisprudence*. University of Oklahoma Press, Norman, OK
Macaulay S 1963 Non-contractual relations in business: A preliminary study. *American Sociological Review* **28**: 55
Sarat A, Felstiner W L F 1995 *Divorce Lawyers and Their Clients: Power and Meaning in the Legal Process*. Oxford University Press, New York
Sarat A, Kearns T R (eds.) 1993 *Law in Everyday Life*. University of Michigan Press, Ann Arbor, MI
Yngvesson B 1989 Inventing law in local settings: Rethinking popular legal culture. *Yale Law Journal* **98**: 1689–709
Yngvesson B, Hennessey P 1975 Small claims, complex disputes: A review of the small claims literature. *Law and Society Review* **9**(2): 219–74

P. Ewick

Law and Geography

Law and geography: the analysis of the manner in which law, as a set of processes, texts, and practices, is shaped by the geographic dimensions of social and political life, and the ways in which the geography of social life is in turn structured by law. Defined thus, the systematic study of the relationship between law and geography has a long, though uneven pedigree. Over the last century, scholars have defined both 'law' and 'space,' and the relationship between the two, in opposed ways, sequentially foregrounding space and then law. The last decade, however, has seen considerable interest in transcending the binary divide between law and space.

1. *Regional Studies of Law*

The earliest 'legal geographies' were concerned with the description and explanation of the regional diversity of legal systems. Thus defined, eminent eighteenth century scholars such as Montesquieu (1748/1949) could be classified as early 'legal geographers.' Regional studies of law, however, were to come into their own in the early years of the twentieth century, with the development of comparative law and the sociology of law. The American comparative legal scholar John Wigmore, in particular, undertook an extensive three-volume survey entitled *A Panorama of the World's Legal Systems* (1928) He also co-edited a collection on legal evolution in 1918 (see Picard 1918), including papers written by lawyers, sociologists, and geographers, which perhaps is the definitive summary of the regional field.

The central claim among these authors appears to be that spatial variations in law are a function of local environmental factors. For some, the latter is defined as the physical environment. Picard (1918), for example, points to variations in climate, soil, forest cover, and distance from the sea as important in explaining legal diversity: 'All these and other variations of land and atmosphere show their traces in the jural system … Brazil and Russia, because they are different geographic regions, will therefore have many differences in law' (p. 170). In keeping with the assumptions of the day, other scholars point to 'race' as an explanatory factor in legal variation. Similar regional schools also seem to have developed in interwar France and Germany. While regional studies have fallen into disfavor, at least within geography, Grossfeld (1984) proves that the determinism of this sensibility is not dead. He argues in uncompromising terms that 'geography is fate … not only for a country but also its culture and the law' (pp. 1512–3).

2. *Impact Analysis*

If regional studies read from geography to law, an influential paper by American geographer Whittlesey (1935) was to invert this relationship in his call for the exploration, as he put it, of the 'impress of effective central authority on the landscape.' Trade tariffs or embargoes, land tenure systems and the regulation of resource exploitation were all identified as laws that result in 'landscape modification.' Thus, in understanding a particular region, geographers were urged to consider those 'phenomena engendered by political forces' (p. 97). The geographer's task, it seems, was now to focus on law's spatial 'impacts,' while taking law itself as an external 'given.'

Such a definition of the relation between law and geography (which Clark 1989 terms 'impact analysis') proved influential until the late 1980s. The tendency has been to treat space as a surface upon which laws instrumentally impact. For Wije (1990, pp. 26, 30), for example, 'Law is a means of achieving social control by regulating behavior or by altering behavioral incentives while geography can study the spatial ramifications of this behavior…. Law can be considered as a geographic… variable.'

Both traditions offer specific insights. A regional analysis is useful to the extent that it reminds us of the ways in which law, despite appearances of homogeneity, can in fact be different and diverse. There are similarities to Clifford Geertz' (1983) fascinating discussion of Balinese law, which proves unsettling to the extent that it alerts us to the internal coherence and logic of systems which, from a Western perspective, can appear entirely alien. In reminding non-lawyers of the undeniable importance of laws on social life and spatial forms and patterns, impact analyses are also useful. We only need to reflect on the direct and indirect ways in which, for example, the urban landscape is a product of countless legal decisions, rulings, and mediations, relating to real property, administrative law, and so on. Yet some commentators have found these analyses partial and limited in terms of the way in which they theorize law, space, and their inter-relation (cf. Clark 1989, Blomley 1994).

3. *Critical Legal Geography*

The last decade has seen the emergence of a third school of legal-geographic enquiry which, in building upon but departing from the above, attempts to redefine both law and space, and also moves beyond the dualistic thinking that relates one to the other. This school of 'critical legal geography,' which one can trace within geography, law, socio-legal studies, anthropology, sociology, and elsewhere, is far from unified. Yet it is probably fair to identify a number of foundational claims concerning law, space, and their mutual relationship.

First, in keeping with much recent scholarship, law itself is defined as more than a set of positive, external commands, but is understood as implicated in an array of social relationships. For example, law can be productive of many social and political effects, being able to define relationships, confer status, and selectively empower, or is argued to define dominant social meanings. There is clear social selectivity to this for many legal scholars. Far from being socially neutral, law is necessarily rooted in, and protective of, prevailing social inequalities around class, race, and gender. Second, to that extent, law is 'socially produced,' rather than being transcendent and a-social; that is, law is formed under particular historic conjunctions by social actors, informed by conditional understandings of society, the State and the individual. Critical legal scholars, or those informed by feminism, for example, have clarified the ways in which broadly liberal understandings of society have informed the common law legal systems of the UK, the USA and Canada in ways that while potentially liberating (Unger 1983) can also prove confining and oppressive, naturalizing hierarchy and limiting progressive alternatives (Peller 1985). Yet, third, law is also 'socially produced' to the extent that it provides an important vocabulary for everyday life. Socio-legal writers consider the manner in which, as Geertz (1983) notes, law can become a form of 'local knowledge' that is drawn upon by people in their relationships with other individuals and the wider social and political world. Rose (1994), for example, explores the manner in which property discourse provides a vital vocabulary by which not only powerful interests, but also individual property holders communicate with one another. Again, this can prove enabling, for example, when oppressed individuals are able to lay claim to a legal right (Laclau and Mouffe 1985), yet can also prove confining. Nedelsky (1990) argues that the tendency to regard others as a threat to our individual autonomy rather than a potential source for freedom is a product of the liberal-legal imagination. Finally, debate has turned on the broader question of legal meaning. Most immediately, the issue is one of textual interpretation—what is the meaning of statute or precedent? Poststructuralist legal theorists have argued that the quest for an essential or stable meaning (which is implicit, for example, to impact analyses) is not only chimerical but also politically dangerous, given the manner in which interpretation is context specific and meaning itself is always referential and endlessly deferred (Tushnet 1984). The 'meanings' of law, from this perspective, must be sought in both dominant ideologies and representations, and also in everyday social practice, where oppositional legal meanings may possibly be discerned.

What then of space? Recent geographic scholarship on law has been critical of both regional perspectives and impact analyses for their tendency to reduce geography to a limited set of variables (for example, by thinking of space as purely a matter of regional diversity) or in presenting space as an objective, autonomous, and neutral variable that, like law, is set apart from society. In tandem with recent legal scholarship, 'critical' geographers have insisted that the effect is not only analytically misguided, given the claim that space is both socially produced and socially constitutive, but also politically problematic, given the claim that space 'is political and ideological. It is a product literally filled with ideologies' (Lefebvre 1979, p. 31).

Examples of the social and political dimensions of space can be traced at various levels of inquiry. For example, simple reflection reveals the ways in which interpersonal spatial categories such as 'near' and 'far' are not abstract concepts, but socially contingent, being very different for men and women, for example (as the concept of 'personal space' reveals), as well as being culturally diverse. Straight Italian men may publicly embrace in a way that would be frowned upon amongst men in Northern England, for example. However, the categories are also more complicated. The workings of racism, for example, depend in part upon spatial arrangements that place racialized groups identified as unlike 'us,' at a spatial remove from those

we hold 'near and dear.' Or, for another example, we need only think about the ways in which Western urbanites find it relatively easy to engage in economic practices that can prove destructive to the everyday lives of 'distant strangers,' such as *campesinos* displaced from their lands by multinational fruit growers, perhaps because of the ways in which distance hides the ethical consequences of our actions from us.

With the claim that both law and space are therefore socially and politically implicated, 'critical legal geographers' have sought to think through the ways in which the relationship between the two might be understood. Rather than treating law and space as autonomous categories, and then tracing the effect of one upon the other, both are understood as socially produced and socially productive. To that extent, the dualism between the two is recast as one of complicated recursive relations; put simply, law can shape space, yet at the same time can itself be shaped by the geographic structures of social and political life. The effect, ultimately, is to dissolve the binary split between the two. This scholarship is 'critical' not only in its attempt to go beyond the traditional categories of more mainstream social science, but also in its more engaged political and ethical stance, given the claim that prevailing legal and spatial arrangements and understandings are associated with hierarchical forms of oppression and domination, including those that center on class domination or patriarchy (Chouinard 1994).

It is hard to generalize about this research, which is still diverse and often unconnected. Scholarship has explored all spatial scales, whether particular localities (e.g. Clark 1985), national legal systems (Engel 1990), and international law (Darian-Smith 1995). Substantive areas covered include crime and policing (Herbert 1997), racism (Ford 1994), the city (Cooper 1998), gender (Sanger 1995), and property (Coombes 1995). That said, some broad areas have received attention.

3.1 Law and Place

The relationship between law and place is considered in a number of ways. For example, some writers point to the ways in which modern Anglo-American law, at least, has historically sought to distance itself or detach itself from local contexts. Blomley (1994) traces the ways in which English common law moved very consciously away from the local communities that had traditionally been the source of its meanings and interpretations to emerge in the sixteenth century as that law which was 'common' to all. This can have direct political consequences. Frug (1980), for example, explores the implications for the legal disempowerment of the city, and raises broader issues concerning local democracy. More generally, Pue (1990) condemns judicial interpretation as 'anti-geographical' in its refusal to recognize local contextual conditions. The insistence on the spatial contextuality of law is not only important in analytical terms. Such a claim can be a direct challenge to what has been termed 'legal closure,' that is, the dominant assumption that legal knowledge and interpretation is detached from the vagaries and politics of social and political life, operating according to its own closed logic. As noted, critics argue that legal closure is not only nonsensical, given the social nature of all discourse and meaning, but also ideological, naturalizing what are fundamentally political claims about society.

That said, in at least two ways place retains its legal importance. First, there has been a number of attempts to argue that formal legal interpretation is, in fact, not as detached from the locality as it claims. American constitutional law, for example, wavers between centralization and a recognition of localized contextual conditions and the integrity of community-based forms of law making and interpretation (Briffault 1990).

Second, in thinking through the relationships between law and place, others recognize that law is much more than formal, judicial decision making. As noted, law emerges as a vital medium and resource for daily life, being called upon in formal and informal ways. Given that, generally speaking, we live our lives out in particular places, that process of legal interpretation could also be said to be localized in some interesting and important ways. In a fascinating series of studies in three US towns, for example, Greenhouse et al. (1994) reveal 'the place of law and the court in the construction of community and hierarchy,' noting the ways in which local as well as extra-local conceptions of law and rights are central to the ways in which community and place are constructed. Similarly, Cooper (1998) explores the way in which an appeal to 'place and belonging,' as she puts it, is important to legal struggles in particular areas. Thus, an attempt by orthodox Jews in Barnet, London, to build an 'eruv,' allowing them to conform to Jewish legal obligations concerning activity on the Sabbath, becomes the spark for conflict with local non-Jewish residents as well as non-orthodox Jews more generally. Cooper traces the way in which this struggle turned on legal and spatial understandings of 'insiders' versus 'outsiders,' rights to place, and the construction of 'Englishness.'

3.2 Geographic Dimensions of Legal Discourse

Closely linked is the study of the geographic claims and representations contained within legal discourse, it being noted that, despite its suspicion of local contingencies, law nevertheless draws upon and helps define a range of spatial patterns, divisions, and associations. These legal representations touch all aspects of legal life, including property, constitutional law, contract, crime, and inter-governmental law. Don Mitchell (1996), for example, outlines the ways in

which judges and other legal actors help define the boundaries between public and private space in US cities, and regulate the activities that are deemed permissible in such spaces. This, he notes, has had particularly oppressive consequences for those people who are effectively denied private space — the homeless. Legal action and interpretation in this sense can actively create certain spatial arrangements, practices, or representations. So, for example, David Delaney (1998) analyzes legal decisions concerning the legality of restrictive covenants in the USA, designed to exclude black residents from white neighborhoods. Necessarily, this entailed various construals of urban space on the part of the judiciary. For Delaney, the effect of such apparently objective mappings was to position the judge as an active participant in the construction of ethnic spatial segregation.

The manner in which such legal constructions of space are contested are locally reworked by marginalized groups is also receiving attention; see, for example, Azuela's (1987) discussion of the ways in which squatters in Mexico City redefined concepts of property. Such legal understandings and representations of space, in other words, are not simply the provenance of formal legal actors. Blomley (1998) explores the way in which struggles over gentrification in Vancouver, Canada, entailed competing conceptions of property rights and land that, in turn, entailed specific representations of the landscape of a local area. For groups representing the poor of the area, opposed to gentrification, the buildings, parks, hotels, and streets of a local area were imbued with legal and political meanings, serving as a physical monument to past acts of oppression and displacement, but also a means by which a collective claim to a neighborhood could be figuratively staked out. In other words, to understand the conflict it was necessary to think about both the importance of contesting claims to property and at the same time, to explore the ways in which those claims were expressed in, and dependent upon local spaces, that were both material and representational.

3.3 *The Politics of the Law–Space Relationship*

As already noted, critical scholarship in law and geography begins with the claim that the two are deeply political. That is, legal–spatial relationships can serve to reproduce prevailing inequalities in power and resources, yet can also serve as a site for struggle and resistance to such forms of domination. Santos (1987), for example, borrows from critiques of cartography to argue that state law deploys particular scales, projections, and symbols that can serve to advance dominant interests. Similarly, Peters (1997) has documented the ways in which law has been used to geographically 'place' (and hence 'displace') First Nations in Canada —both by spatially locating Native peoples in marginal reserves and by socially and politically locating them within Canadian society. In a related vein, Borrows (1997) documents the manner in which dominant forms of Canadian legal knowledge have no place for indigenous understandings of place and nature. Others seek to turn the political tables. Chouinard (1989) focuses on local law clinics in Ontario as a site for class-based resistance while Kobayashi (1990) opposes legal closure in relation to racist legal practices.

4. *Future Research*

While suggestive, the critical legal geography literature is still in a state of formation. To date, it has (perhaps fortunately) avoided a clear self-identity. While the work of many scholars is legal and geographic, most would not necessarily identify themselves in those terms. This is thought to be a good thing, encouraging the sort of truly interdisciplinary inquiry that is called for. That said, these writings have managed to begin to pose some suggestive and novel questions. Yet many issues still await careful inquiry. More generally, the theoretical basis for critical inquiry into law and geography is still somewhat undeveloped, despite some attempts (e.g., Clark 1989, Blomley 1994, Chouinard 1994). Some have embraced the postmodern turn, focusing attention on discourse and the construction of meaning. Others have worried at the degree to which this constitutes a rejection of political–economic and socialist–feminist insights (Chouinard 1994). Similarly, a focus on the meanings and discourses of law has meant the neglect of its more practical dimensions. To that extent, careful engagement with 'impact analyses' might be productive. Detailed, empirical accounts of the workings of law within particular locales are also still needed. Revisiting the regional school might again be useful here, if only as a counterpoint. The relationships between law, space, and nature, another traditional area of geographic concern, have only just begun to be explored. The ethical dimensions of law—for example, the complicated politics of rights—are also worth thinking about, particularly given recent geographic scholarship in the area. While some work has been done, the manner in which social and spatial differentiation and inequality along axes such as class, 'race' and gender are expressed in, facilitated by, or reconstituted through law also calls out for careful analysis. The geographic relationships between law and other socially and politically powerful forms of knowledge, such as planning, or medicine, could also be profitably considered. More generally, the manner in which law represents the monopoly of organized violence has not been thought through, particularly from a geographic perspective.

See also: International Justice: Philosophical Aspects; International Law and Treaties; Law and Democracy;

Law and Development; Law and Economics; Law and Society: Sociolegal Studies; Law: Anthropological Aspects; Law as an Instrument of Social Change; Law: Overview; Law, Sociology of

Bibliography

Azuela A 1987 Low income settlements in Mexico City. *International Journal of Urban and Regional Research* **11**(4): 522–41

Blomley N K 1994 *Law, Space and the Geographies of Power*. Guilford, New York

Blomley N K 1998 Landscapes of property. *Law and Society Review* **32**(3): 567–612

Borrows J 1997 Living between water and rocks: First Nations environmental planning and democracy. *University of Toronto Law Journal* **47**: 417–68

Briffault R 1990 Our Localism, I: The structure of local government law. *Columbia Law Review* **90**(1): 1–15

Chouinard V 1989 Transformations in the capitalist state: The development of legal aid and legal clinics in Canada. *Transactions of the Institute of British Geographers* **14**(3): 329–49

Chouinard V 1994 Geography, law and legal struggles: Which ways ahead? *Progress in Human Geography* **11**(5): 415–40

Clark G L 1985 *Judges and the Cities: Interpreting Local Autonomy*. University of Chicago Press, Chicago

Clark G L 1989 The geography of law In: Peet R, Thrift N (eds.) *The New Models in Human Geography*. Unwin Hyman, London, UK, pp. 310–37

Cooper D 1998 *Governing Out of Order: Space, Law and the Politics of Belonging*. Rivers Oram, London, UK

Coombes R 1995 The cultural life of things: Anthropological approaches to law and society in conditions of globalization. *American University International Journal of Law and Policy* **10**: 791–835

Darian-Smith E 1995 Law in place: Legal mediations of national identity and state territory in Europe. In: Fitzpatrick P (ed.) *Nationalism, Racism and the Rule of Law*. Dartmouth, Aldershot, UK pp. 27–44

Delaney D 1998 *Race, Place and the Law—1836–1948*. University of Texas Press, Austin

Engel D M 1990 Litigation across space and time: Courts, conflicts and social change. *Law and Society Review* **24**(2): 333–44

Frug G E 1980 The city as a legal concept. *Harvard Law Review* **93**(6): 1059–154

Ford R T 1994 The boundaries of race: Political geography in legal analysis. *Harvard Law Review* **107**(8): 1841–1921

Geertz C 1983 *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York

Greenhouse C J, Yngvesson B, Engel D M 1994 *Law and Community in Three American Towns*. Cornell University Press, Ithaca

Grossfeld B 1984 Geography and the law. *Michigan Law Review* **82**: 1510–9

Herbert S 1997 *Policing Space: Territoriality and the Los Angeles Police Department*. University of Minnesota Press, Minneapolis, MN

Kobayashi A 1990 Racism and the law in Canada: A geographical perspective. *Urban Geography* **11**(5): 447–73

Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy: Towards a Radical Democratic Politics*. Verso, London

Lefebvre H 1979 Reflections on the politics of space. *Antipode* **8**(2): 30–7

Mitchell D 1996 Political violence, order and the legal construction of public space: Power and the public forum doctrine. *Urban Geography* **17**(2): 152–78

Montesquieu B de 1748/1949 *The Spirit of the Laws*. Hafner Press, New York

Nedelsky J 1990 Law, boundaries and the bounded self. *Representations* **30**: 162–89

Peller G 1985 The metaphysics of American Law. *California Law Review* **73**: 1151–290

Peters E 1997 Challenging the geographies of 'Indianness': The Batchewana case. *Urban Geography* **18**(1): 56–61

Picard E 1918 Factors of legal evolution. In: Kocourek A, Wigmore J H (eds.) *Formative Influences of Legal Development*. Little Brown, Boston, pp. 163–81

Pue W W 1990 Wrestling with law: (geographical) specificity vs. (legal) abstraction. *Urban Geography* **11**(6): 566–85

Rose C 1994 *Property as Persuasion: Essays on the History, Theory and Rhetoric of Ownership*. Westview Press, Boulder, CO

Sanger C 1995 Girls and the getaway; cars, culture, and the predicament of gendered space. *University of Pennsylvania Law Review* **144**: 705–56

Santos B 1987 Law: a map of misreading. Toward a postmodern conception of law. *Journal of Law and Society* **14**(3): 279–302

Tushnet M 1984 An essay on rights. *Texas Law Review* **62**(8): 1363–404

Unger R 1983 The critical legal studies movement. *Harvard Law Review* **96**(3): 320–32

Whittlesey D 1935 The impress of effective central authority upon the landscape. *Annals of the Association of American Geographers* **25**: 85–97

Wigmore J H 1928 *A Panorama of the World's Legal Systems* (3 Vols). West Publishing Company, St Paul, MN

Wije C 1990 Applied law and applied geography. *Operational Geographer* **8**(1): 26–31

N. Blomley

Law and Morality: A Continental-European Perspective

Under the general rubric of the concept and nature of law, one of the most fundamental problems is the relation between law and morality. Two positions have been locked in combat for over 2000 years: positivism and nonpositivism. All positivists share the separability thesis; all nonpositivists contest it in at least one respect.

1. *The Separability and Separation Theses*

The separability thesis says that there is no necessary connection between law and morality. This implies that all connections between law and morality are merely contingent in character. There exists no neces-

sary relation between what the law is and what morality or justice requires, or between the law as it is and the law as it ought to be. The great positivist Hans Kelsen has put it this way: 'Any content whatever can be law' (Kelsen [1934] 1992).

The separability thesis defines the softest version of legal positivism. Its only claim is that it is possible to have law with any content whatever, independently of any and all demands of justice. It does not rule out the possibility that the positive law of a certain community includes moral principles, for instance, through constitutional provisions that transform human rights into positive law. Where the concept of law is understood positivistically, the only requirement stemming from the concept of law itself is that the inclusion of moral content is considered as a matter of fact, that is, as merely contingent and not in any way necessary.

A stronger form of positivism is expressed by the separation thesis. The separation thesis necessarily presupposes the separability thesis, and adds to it that there are good normative reasons for defining the concept of law in a way that excludes all moral elements. This means that the separability and the separation theses refer to two different kinds of arguments. The separability thesis, along with its opposite, the nonseparability thesis, is concerned with what is necessary, analytic, or *a priori*. The separation thesis, too, depends of course on these arguments, for there cannot be separation without separability. But it claims more. It does not confine itself to determining what definitions of law are possible. It also tries to identify the best definition. For this purpose, normative arguments are indispensable. There is, therefore, a division of labor between the separability and the separation theses. Arguments supporting the first thesis attempt to show that it is possible to exclude all moral elements from the concept of law; arguments on behalf of the second thesis seek to establish that this exclusion is preferable.

If the separation thesis applies, only two elements remain as candidates for the definition of law: appropriate enactment and social efficacy. The concept of social efficacy refers to such matters as habits, orders backed by threats, and to beliefs, emotions, and attitudes. Authors of a legal realist persuasion have tried, following this lead, to reduce law to social or psychic facts conceived as natural facts (Olivecrona 1939). The concept of appropriate enactment is more refined. It is impossible to distinguish between what is appropriately, and inappropriately, enacted without reference to norms conferring and defining the power of enactment. The most famous power-conferring norm is Kelsen's Basic Norm (*Grundnorm*). Introducing a basic power-conferring norm, Kelsen aims at normativity without morality (Paulson and Litschewski Paulson 1998).

The elements of social efficacy and appropriate enactment can be interpreted and combined in very different ways. For this reason, a great many variations on the theme of legal positivism are conceivable. What is common to all of them is that the concepts of social efficacy and/or appropriate enactment are sufficient for the definition of law. Morality as such is neither a necessary element nor an optimal third element of the concept of law. Precisely this is contested by nonpositivism. All nonpositivists share the view that the concept of law must or ought to include moral elements. If the inclusion of moral elements is conceived as necessary, one can speak of a strong connection thesis. The strong connection thesis is the contradictory of the separability thesis. If the inclusion is conceived only as preferable, not necessary, one can speak of a weak connection thesis. The weak connection thesis is the contradictory of the separation thesis.

The most radical nonpositivist concept of law consists in the complete substitution of social efficacy and appropriate enactment by moral correctness. This would be a pure natural law theory. In practice, such a radical pure natural law theory would amount to anarchism. For this reason, serious natural law theorists such as Thomas Aquinas stress the necessity of positive law as defined by appropriate enactment and social efficacy (Aquinas 1988). The crucial question, therefore, is not whether law is to be defined either by social efficacy together with appropriate enactment or by moral correctness but rather whether or not social efficacy and appropriate enactment are to be connected with moral correctness in one way or another. There are three ways in which such a connection can be conceived. Morality can be connected with positive law, first, by including moral principles and arguments in the law, second, by limiting the possible content of law, and, third, by providing an appropriate foundation for the obligation to obey the law. One can, therefore, distinguish three different problems concerning the relation between law and morality: the inclusion problem, the limitation problem, and the problem of an appropriate foundation.

2. *The Inclusion Problem*

The logical basis of the inclusion problem is the open texture of law. Positivists and nonpositivists agree, first, that law has an open texture (Hart 1994) and, second, that cases falling within the open sphere of positive law are often decided on moral grounds. Such a purely factual connection is fully compatible with legal positivism. What is more, if one now goes on to assume that moral principles are incorporated by virtue of their correctness into the law through a rule of recognition *qua* 'conventional normative practice' (Coleman 1996), even then one remains inside the positivist camp. A conventional practice is a practice that either exists or does not exist. Whether or not it

exists is a matter decided by the positive legal system at hand. The classification as 'inclusive positivism' is, therefore, correct (Coleman 1996).

Positivism yields to nonpositivism only if it is maintained that the inclusion of moral principles and arguments into law is necessary and not merely contingent. A central argument for the necessity of this inclusion is the argument of correctness. It proceeds in two steps. In a first step, the attempt is made to show that law necessarily raises a claim to correctness. The second step consists in explaining that this claim implies a necessary connection between law and morality.

Implicit claims can be made explicit by showing that their explicit negation involves absurdity. One has only to imagine a judge who hands down the following judgement: 'The defendant is hereby sentenced to life imprisonment, albeit wrongly, because valid law was interpreted incorrectly.' The absurdity of this judgment stems from the contradiction between the claim to correctness implicit in acts of law and its explicit negation. This contradiction can be avoided only if the claim to correctness is given up and some sort of power claim is introduced in its place. But this would be to abandon law. A social practice claiming nothing but power or force would not be a legal system. In this sense, the claim to correctness is necessarily connected with law.

A positivist can agree with this and nevertheless maintain that the necessary connection between law and the claim to correctness does not imply that there exists a necessary connection between law and morality. All he needs to say is that the claim to correctness has a purely legal content and that this legal content has no moral implications.

This leads to the second step of the correctness argument. That a case falls within the open sphere of positive law means that positive law does not determine the answer. If the judge were bound only by positive law, he could, whenever the positive law reasons run out, decide the case by appealing to his own personal preference or by tossing the dice. This, however, would be compatible with the claim to correctness only if there existed no reasons for the correctness of a legal decision outside the class of positive law reasons. Now legal decisions are answers to practical questions, and there exists a variety of reasons as answers to practical questions that are found outside the class of positive law reasons. The spectrum comprises consideration of utility, familiar ideas from the tradition about what is good and bad, and principles of justice.

Justice is correctness of distribution and balance, and questions of justice are moral questions. Legal decisions essentially concern questions of balance and distribution. Therefore, legal decisions essentially concern moral questions. This, together with the claim to correctness necessarily raised by legal decisions, leads to a necessary connection of legal and moral argument. This necessary connection of arguments does not mean that morally mistaken legal decisions cannot be legally valid. It does mean, however, that they are not only morally but also legally mistaken. In this way, the idea of justice is incorporated into the concept of law. This affects the picture of law fundamentally (Alexy 1998).

3. *The Limitation Problem*

If the correctness argument is sound, legal norms and legal decisions that for moral reasons do not fulfil the claim to correctness are legally mistaken, but they do not necessarily lose, on this basis alone, their legal character or their legal validity. Anarchy would be the consequence if each and every moral defect were sufficient by itself—that is, without any institutional act or decision—to undermine legal validity or even the very legal character of the norm or the decision. So far, there exists a broad consensus. Highly contested, however, is the question of whether gross injustice undermines legal validity or legal character. This question was the topic of Hart's well-known critique (Hart 1958) of Radbruch's famous formula (Radbruch 1946). The shortest conceivable formulation of this formula runs as follows: extreme injustice is not law. This formula, which was applied by German courts after the defeat of National Socialism in 1945 and after the collapse of the German Democratic Republic in 1989, does not require any sort of complete fit between law and morality—as the claim to correctness does. Rather, it allows that appropriately issued and socially effective norms are valid law even if they are severely unjust. It is only in cases of extreme injustice that the formula gives priority to material justice over legal certainty. In this way, it builds into law an outermost limit. This limit is substantially defined by the core meaning of human rights.

The acceptability of the Radbruch formula essentially depends on whether it is judged from the observer's point of view or from that of a participant. It is not a problem for the nonpositivist to grant that an observer, wanting merely to describe the law of a wicked legal system, can and should use a positivistic concept of law based exclusively on appropriate enactment and social efficacy. The controversy over Radbruch's formula begins once the participant's perspective is considered. The participant's perspective engages one who, from within the legal system (for instance, as an official who has to apply the law or as a citizen who has to abide by it), asks what the law requires, forbids, and permits.

The dispute about whether or not an outermost moral limit ought to be built into the concept of law cannot be decided merely by means of conceptual arguments. The meaning of the expression 'law' excludes neither Radbruch's formula nor its negation. The decision can only be grounded on normative

arguments. In other words, the inclusion or exclusion of an outermost moral limit cannot be based on a conceptual necessity, but only on a normative necessity.

There are good reasons for both positions. The main argument of the positivists is the argument of legal certainty. They claim that the Radbruch formula leads to hidden retroactivity which, in penal law, boils down to a violation of the principle *nulla poena sine lege*. This is the case because Radbruch's formula invalidates statutes of an iniquitous legal system that permit extreme injustice. This is said to be unacceptable for, it is argued, the principle *nulla poena sine lege* protects everybody, even the henchmen of an iniquitous regime, so long as their atrocities are covered by the positive law of that system. The nonpositivist responds that legal certainty and, in particular, especially the principle *nulla poena sine lege* are, indeed, lofty values, but that they are not the only values at stake. They are in conflict with the material justice to which past and future victims of iniquitous regimes have a claim. In the end, these problems are to be resolved by weighing the principles at issue. In normal situations, the *nulla poena sine lege* principle must have priority. In cases of extreme injustice, however, the situation is more difficult. There are good arguments for giving priority to the protection of the rights of past and future victims over the protection of those who, implicated in the acts of an unjust state, have relied on a legal justification of their deeds on the basis of a 'legal positivization' of injustice (Dyzenhaus 1999).

4. *The Problem of an Appropriate Foundation*

The third problem of the relationship between law and morality concerns the question of whether there is a moral obligation of all addressees of the legal system to obey the law simply because it is the law, independently of its content. This is the question of a general moral obligation to obey the law (Raz 1979). If such a general moral obligation exists, morality offers an appropriate foundation of law.

The problem of an adequate foundation is related to legal positivism in a way that is completely different from the relation of the inclusion and limitation problem to legal positivism. Positivism is compatible with the assumption of a general moral obligation to obey the law as well as with a complete denial of such an obligation. The first version can be called moral positivism, the second, neutral positivism. Moral positivism is the strongest form of positivism. It combines a moral obligation to obey even the most immoral law with the thesis that law can be as immoral as possible as long as it does not lose social efficacy. Neutral positivism expresses a far weaker form of positivism. Law imposes only legal obligations. These legal obligations need not be in conflict with moral obligations, but they can be. Neutral positivists can even say that in case of conflict, moral obligations enjoy priority. They remain positivists so long as they say that the priority of a moral obligation *qua* moral obligation does not undermine the legal validity of the conflicting legal obligation. The situation of nonpositivists is more complex. Not all, but some conflicts between law and morality are already solved for them before the question of a moral duty to obey the law arises. Cases of extreme injustice are conceptualized by nonpositivism not as conflicts between valid law and morality but as instances of the limits of law. Thus, from the nonpositivist point of view, there is no confrontation of a general duty to obey the law with the problem of extreme injustice. Beneath the threshold of extreme injustice, the problem of a general duty to obey the law is, however, the same for a positivist and a nonpositivist. Nonpositivism mitigates the problem but is not able to solve it. Its solution poses a moral question.

The most general reasons for the general duty to obey the law are the values of nonhostile conflict resolution and social cooperation. Neither of these can be achieved without law, and both ought to be achieved, for rights are endangered where social conflicts are resolved by brute force, and general welfare and well-being are not possible without social cooperation. This is the classical argument for a general moral obligation to obey the law as it is found—with respect to the protection of rights—in Kant (Kant [1798] 1991).

There are two kinds of objections to a general moral obligation to obey the law. The first attempts to undercut such an obligation, the second, to rebut it. The undercutting objection maintains that there are cases in which transgressing the law does not affect the values of nonhostile conflict resolution and social cooperation at all—or at any rate, so minimally that the effect can be neglected. Two examples, frequently given, are not stopping at a red traffic light in a remote area late at night where no one is around to notice it and small tax offences that no one will ever discover. If it can be taken for granted in these cases that no one will take notice of the offences, no bad example is set that might endanger the general willingness to abide by the law and, through this, law's social efficacy. Where this point is concerned, the two cases are indeed on the same footing. There is, nevertheless, a fundamental difference. If the conditions of the traffic light case prevail—no danger, and no observer—this case is indeed a case in which violating the law has no negative effects. The decision here turns on the solution to a general problem of moral philosophy, namely, the question of whether acts or rules are the genuine objects of moral assessment. This is the focus of the debate between act and rule utilitarianism (Rawls 1955). If one follows the dictates of act utilitarianism, the general duty to obey the law will be undercut in cases such as the stoplight case. The practical impact

of this solution, however, is small, for the occasions on which one can be completely confident that the two conditions—no danger, and no observer—are fulfilled, are rare.

Whilst it is possible to conceive the general obligation to obey the law as being undercut in the red traffic light case, this is impossible in the tax case. Even if only a small amount of money is at stake, it is clear that the offence will affect negatively the overall public financial situation. The advantage on the side of the offender corresponds to a disadvantage on the side of the community. Taking advantages by failing to contribute to a common enterprise where others contribute is—at least *prima facie*—unfair. In the tax case, the moral obligation to obey the law is, therefore, not undercut—even if, where no one knows anything about the offence, the general willingness to abide by the law is not endangered.

The bad example problem must be distinguished from the problem of whether there is a moral obligation to obey immoral law. What is at stake here is the question of rebutting the general obligation to obey the law. Essentially this problem is a problem of balancing legal certainty against moral correctness. Only those who give absolute priority to the moral value of legal certainty will never allow that the general moral obligation to obey the law may be rebutted by moral reasons. All the others have to solve difficult moral dilemmas case by case. For the others, too, there exists a general obligation to obey the law, but this obligation is not indefeasible. This is true not only for positivists, who have to take up the weighing procedure with an unlimited concept of law, but also for nonpositivists, who with their extreme-injustice clause are able to solve the most serious—if by no means all—moral dilemmas posed by law.

See also: Justice and its Many Faces: Cultural Concerns; Justice and Law; Law and Morality: An Analytic Perspective; Moral Reasoning in Psychology; Moral Sentiments in Society; Morality (East and West): Cultural Concerns; Religion: Morality and Social Control

Bibliography

Alexy R 1998 Law and correctness. In: Freeman M D A (ed.) *Legal Theory at the End of the Millennium*. Oxford University Press, Oxford, UK, pp. 205–21

Aquinas T 1988 *Summa Theologiae*. Editiones Paulinae, Milan

Coleman J 1996 Authority and reason. In: George R P (ed.) *The Autonomy of Law: Essays on Legal Positivism*. Clarendon Press, Oxford, UK, pp. 287–319

Dyzenhaus D (ed.) 1999 *Recrafting the Rule of Law: The Limits of Legal Order*. Hart Publishing, Oxford, UK

Hart H L A 1958 Positivism and the separation of law and morals. *Harvard Law Review* **71**: 593–629

Hart H L A 1994 *The Concept of Law*, 2nd edn. Clarendon Press, Oxford, UK

Kant I [1798] 1991 *The Metaphysics of Morals*. Cambridge University Press, Cambridge, UK

Kelsen H [1934] 1992 *Introduction to the Problems of Legal Theory*. Clarendon Press, Oxford, UK

Olivecrona K 1939 *Law as Fact*. Munksgaard, Copenhagen

Paulson S L, Litschewski Paulson B (eds.) 1998 *Normativity and Norms: Critical Perspectives on Kelsenian Themes*. Clarendon Press, Oxford, UK

Radbruch G 1946 Gesetzliches Unrecht und übergesetzliches Recht. *Süddeutsche Juristen-Zeitung* **1**: 105–8

Rawls J 1955 Two concepts of rules. *The Philosophical Review* **64**: 3–32

Raz J 1979 *The Authority of Law: Essays on Law and Morality*. Clarendon Press, Oxford, UK

R. Alexy

Law and Morality: An Analytic Perspective

As complementary and competing systems of normative social order, law and morality have much in common. Not only is there substantial overlap in the content of the moral and legal norms of all societies, but law may readily be seen as both an expression of and an influence on a society's morality. Judges are expected to have regard to moral considerations when interpreting, applying, and developing law and, in most democratic countries, they have the power to override legislation which conflicts with those basic moral rights, such as freedom of expression, which are enshrined in the constitution of their polity. Further, legislatures are expected to make law in the light of the moral beliefs of the population, particularly those which relate to peace, well-being, and social justice.

Moreover, legal systems depend, in varying degrees, on securing the conformity of citizens, which in turn depends, in part, on satisfying their moral expectations as well as their self-interest. The moral constraints which citizens impose on each other and on themselves through the informal processes of social approval and disapproval are crucially reinforced by the sanctions of criminal law and the remedies for such matters as breach of contract and negligence, which are enforced by the civil law. The ethics of public officials is underpinned by administrative as well as criminal law. Moreover, legal and moral discourse shares a whole range of basic concepts. Justice, fairness, rights, utility, equality, and freedom are the common vocabulary of both normative systems. In less developed societies there is little or no institutional divide in the normative structure between law and morality.

These commonalities of law and morality must be kept in mind when examining attempts to distance law and morality. They suggest that the social functions of law and morality are similar and overlapping, part of a wider system of normative order arising out of the requirements of an ordered human society. That said, however, there are important differences between law and morality which call for analysis, explanation, and evaluation.

1. *Critical and Positive Morality*

Focusing on apparent differences between law and morality, several key factors emerge: (a) law is said to be mandatory, even coercive, whereas morality is voluntary; (b) law is said to be an institutionalized system of norms, with specialized officials and procedures, whereas morality is informal and diffuse in its operation; (c) law is public and associated with the state, morality is private, a matter for the individual to determine for herself; (d) the identification of law requires reference to past events and previous authoritative decisions, whereas moral determinations are not so restricted; (e) laws and legal systems may or may not be just or justified depending on their content and procedures but there is something incongruous in the notion of an unjust moral principle.

Some of these points can be clarified by distinguishing between two senses of morality, the first referring to the collection of norms of conduct which are generally observed and used to criticize conduct in a given social group. This is sometimes called 'positive morality' (Austin 1954, p. 122). The second sense of morality is the activity of criticizing and evaluating the norms which make up positive morality. It is critical morality (Hart 1961, pp. 176–80) in which the content and form of positive moralities are assessed and judgments are made about what actually is right and wrong that give rise to most of the differences between law and morality. Making critical moral judgments cannot be coerced or forced on others and it makes no sense to come to the critical conclusion that it is right to act unjustly, unless injustice is defined by reference to norms of positive morality. Critical morality may be contrasted to both legal and positive moral systems of norms, with the law and positive morality sharing the characteristics which typify ongoing cultural structures and critical morality standing outside and above these phenomena.

Of course, positive systems may encourage and include critical attitudes towards prevailing norms so that critical morality can itself be viewed as part and parcel of positive morality, but the importance of the distinction between positive morality and critical morality has more to do with the perspectives involved, in particular the distinction between describing the moral beliefs and actions of an individual or group and evaluating those beliefs and actions from the moral rather than the sociological point of view. Precisely what that moral point of view may be is a controversial issue within moral philosophy, but we may take it to be the point of view which seeks to determine the ultimate criteria for assessing the worth of human conduct from an impartial or universal perspective (Baier 1958).

Having distinguished positive and critical morality, we can frame the questions which arise in relation to law and morality in a more sophisticated way. Thus we may raise, as a matter of critical morality, whether (a) law ought to incorporate the positive morality of the citizens of that jurisdiction, (b) the critical moral principles which we adopt to criticize positive morality ought to be the same as those we use to criticize law, and (c) there is such a thing as critical as opposed to positive law, similar to the distinction between critical and positive morality.

1.1 *Legal Positivism and Natural Law*

The relationships between law and morality have long been prime sites for the articulation of competing legal philosophies, particularly classical natural law and legal positivist approaches to the understanding and evaluation of law and legal systems. However, as we have seen, the relationships involved are varied in type and substance and it is by no means clear that the traditional packages of theses associated with the classical theories still hang together. Indeed there are real prospects for conceptual realignments which render some of the historical debates on law and morality outmoded.

Legal positivism, which received its classical formulation in the work of Jeremy Bentham and John Austin during the eighteenth century, is the legal theory which emphasizes the separation of law and morality, and the subordination of morals to the will of the sovereign (Austin 1954). Following on from Hobbes' analysis of political authority as based on the requirements of mutual survival in a predatory social world, the state is seen as imposing a morally arbitrary rule of order so as to secure peace between warring factions, religions, and moralities. This characteristically modern theory of the state replaced the medieval European synthesis of government, religion, and morality which received its classical formulation in the *Summa Theologica* of Thomas Aquinas whose definition of law as 'a rational ordering of things which concern the common good; promulgated by whoever is charge with care of the community' (Aquinas 1966, Art. 4) sums up the fusion of law and morality in the natural law tradition.

Austin, through whose work legal positivism became the reigning orthodoxy in Anglo-American legal theory, presents his position in a series of definitions designed to identify law in empirical terms, such as

command, sanction, and habitual obedience. Law is defined as the general commands of a sovereign, that is, the person or group who is habitually obeyed in a certain territory and does not himself habitually obey anyone else. A command is the expression of a wish relating to the conduct of another which is backed by threats of force, legal obligation being the likelihood that failure to comply with the a command will be followed by the imposition of sanctions. Each state has one such sovereign (which may comprise an individual or a collectivity) who is at the giving but not the receiving end of such commands. The existence of this unitary sovereign is the core of the positivist analysis of the modern sovereign state.

It is a defining feature of this mode of legal positivism that there is no necessary connection between law and morality (Hart 1961, p. 181). The classical positivist separation of law and morality is intellectually based on Hume's distinction between 'is' (statements about what is and is not the case) and 'ought' (statements about what ought or ought not to be the case). In Austin's ringing words 'the existence of law is one thing, its goodness or badness another' (Austin 1954, p. 184). Ever since, legal positivists have insisted that it is possible to separate out the description of law from our moral evaluation and criticism of that law: legal argument is not moral argument, legal authority is not moral authority, and a law is a law whether it is just or unjust, good, or evil. All this applies only to what is called 'positive law,' that is, the law as it is actually experienced in human societies.

Such a dichotomy depends on our capacity to identify law in an empirical manner and this Austin does with his analysis of law as the command of a political sovereign. In the terminology of a more recent and equally famous legal positivist, H. L. A. Hart, the law of a jurisdiction can be identified by means of its 'rule of recognition' (Hart 1961, pp. 92–3), which lists the criteria used by officials to determine legal validity, that is, what is and what is not part of the law of that jurisdiction. In Austin's case that rule nominates the commands of the sovereign as the criterion which is necessary and sufficient for the determination of legal validity.

A similar approach is taken by the classical legal positivists to positive morality. The rules of conduct which are widely followed in a given society and are backed by the sanctions of public approval and disapproval are seen as a system of identifiable norms analogous to positive law. Both of these are distinguished from 'natural law,' a term which Austin is happy to apply to what is actually morally right and wrong in human conduct, which he himself, following Bentham and perhaps Hobbes, identifies with the utilitarian standard of acting or refraining from acting so as to maximize human happiness and minimize human pain. While such natural laws can be regarded as the commands of the ultimate sovereign, God, they feature in the critique of positive law and not in its identification. The classical natural law position, in contrast, is that knowledge of right and wrong as manifest in God's law speaking through human reason or conscience plays an essential part in the identification and understanding of all laws, including positive laws.

For some natural lawyers this means that there can be no such thing as an 'unjust law.' However, John Finnis, one of the most influential natural law theorist currently writing, denies that natural lawyers have always or have needed to deny the existence of unjust laws. He also points out that natural law has a developed view on the role of positive law as a social institution that is necessary to give concrete application to the general moral principles which derive from natural law (Finnis 1980). Further, there are powerful arguments to the effect that, while individual laws may or may not be just, the concept of law, as distinct from the content of particular laws, is necessarily connected with justice, for law is an institution that must not only claim to be just but seek to be (Beyleveld and Brownsword 1989).

Put in these terms, it appears that the disagreement between legal positivists and natural lawyers is about the meanings of the terms 'law' and 'morality,' in which case the apparently major theoretical division between legal philosophies is in the nature of an uninteresting irrelevance. Stipulative definitions of law and morality may give rise to a logical separation between the terms, but there are no essential and unchanging meanings which we are required to adopt, so that it becomes a matter of choice, individual or cultural, whether or not we conceive of law and morality as conceptually tied.

A crucial matter of interpretation is whether the role of natural law in both classical approaches is an exemplar of positive or critical morality. In the case of legal positivism there seems no doubt that the utilitarian morality used to assess legal systems is presented as a critical morality. Indeed, Bentham gives this activity the title 'censorial jurisprudence.' In the case of natural law theory, the situation is less clear. This is because of the moral methodology of natural law which depends on reasoning about 'nature,' which includes human behavior, particularly normative behavior, that is conduct which involves the control of natural appetites in a way which makes social life and individual spiritual development possible. In effect this means that universal positive morality has a key evidential role in determining what is in accordance with right reason. However, it is clear that classical natural law theory presents as a critical morality, and modern 'natural law' theories, which are standardly uncoupled from theological backgrounds or moral deference to what is 'natural,' hold that it is critical not positive morality that is central to the very idea of law.

This does not mean, however, that the classical theories are simply presenting competing schemes of

critical morality with the utilitarianism of legal positivism pitted against the theological naturalism of natural law. Indeed, both theories have outgrown their original moral affiliations. Exactly what is now at stake between the two approaches is best explained by reference to more recent controversies over the 'separability thesis.'

1.2 Separation and Separability

With the introduction of Hart's rule of recognition and, before that, the impact of Kelsen's 'pure theory of law' which includes the idea of a presupposition about the authority of law from which all legal validity derives, the classical debate about law and morality has come to be formulated in terms of what is known as the 'separability thesis.' This goes behind the positivist idea that positive law can all be traced to social sources, such as commands, or enactments, or court decisions, to an examination of the content of the rule of recognition that is used in the processes which make up those human sources of law. In other words, the focus has shifted to the sources of law as understood by the officials whose pronouncements constitute the social origins of law, namely the criteria given in the operative rule of recognition.

The separability thesis, as formulated by Jules Coleman, is the view that a rule of recognition need not include moral criteria, such as justice or utility, but can consist entirely of empirical identifiables such as human commands or legislative enactments. Because this thesis is compatible with all actual rules of recognition including moral criteria, this minimalist form of legal positivism is called 'soft', or sometimes 'inclusive' legal positivism (Waluchow 1994), as opposed to the 'hard' or 'exclusive' variety which admits only nonmoral criteria into anything which counts as a rule of recognition. According to soft legal positivism it is a contingent matter how far a legal system officially incorporates moral considerations into its criteria of legal validity. Such a view is regarded as positivist because it requires that morality enters into legal reasoning through an empirically identifiable rule of recognition which happens to include moral criteria and not, as it were, in its own right. Hard positivism also accepts that moral criteria are, indeed ought to be, utilized in the making of law through legislation, but insists that what makes that morality or anything else into law is its adoption by authorized law makers. The difference between soft and hard legal positivism seems to come down to whether legal officials, such as judges, can make law through the direct application of their own moral views or by reference to the moral views of others or whether they must identify law in much the same way as a social scientist seeking to describe the law of a particular jurisdiction. It can be argued that if soft positivism allows that some legal systems require judges actually to make correct moral determination in determining what the law is, then it is really a soft natural law theory.

Important empirical issues then arise for social science as to whose positive morality finds its way into the law, how that law in turn affects morality in general, which elements of morality most commonly find legal expression, and which are more often excluded from law. It is also a matter for empirical study how far the morality within the law stems from legislative or judicial activity.

Viewed as a descriptive thesis, soft legal positivism is clearly superior as all legal systems appear to allow some role for moral judgments in the legal reasoning which determines legal validity. Common law systems permit judicial development of law in the light of broad moral principles, albeit subject to the theoretical right of legislators to amend the common law. More and more legal systems incorporate a bill of rights with general moral principles which it is the court's duty to interpret, something which can be done only by reference to the moral preferences of those who give concrete application to such vague moral standards as equality and freedom. All this seems to accord more with a natural law approach, at least with respect to the idea of fundamental rights as the point where law and morality are inseparable (Dworkin 1996).

It is true that the moral reference in rules of recognition often serve only to incorporate positive morality or some identified aspect of it within the law; often this is done by the idea that the law must enforce reasonable expectations. However, if only because there is seldom any indisputable agreed content to positive morality on any topic, in effect these approaches call for the judicial exercise of critical moral judgment. For the soft positivist to claim that this need not be the case seems a relatively minor point to make about law and morality.

The retreat into soft positivism is explicable because of the emphasis that legal positivists have historically placed on providing a descriptive theory of all developed legal systems. However, there are good reasons for seeing legal positivism in a more normative light as an approach which is actually setting forth an idea of a 'good' system of law, so turning the separability thesis into a prescriptive claim about what a rule of recognition ought to contain. Thus, hard positivism may be viewed as a more or less disguised prescription which is recommending that a rule of recognition ought not to include moral criteria and that all positive law should be framed in such a way that it can be understood and applied without recourse to the moral opinions of citizens or judges (Campbell 1996, p. 72).

This normative approach to the separation of law and morality derives from a complex political philosophy that points to the role of authoritative rules in human society (Raz 1979). Social cooperation and coordination, dispute resolution and maintenance of

public order, it is alleged, require that there be rules which are followed and enforced according to an authorized formulation that leaves as little room as is possible for doubt as to their content and implications. These arguments often relate to the utility of rules as social devices which give effect to conventions that make it possible for individuals and groups to anticipate each others conduct with considerable reliance. The utility of these rules is relatively independent of their precise content since what matters is that there is some generally recognized convention rather than precisely what that convention may be (Postema 1982).

While these considerations give us reason to see why it may be socially important for some laws to be clear and applicable without controversy as to their meaning, other laws, such as those forbidding clearly harmful conduct, are more dependent on the moral acceptability of their content. Even here, however, there is a role for an official determination of which conduct is punishable and which is merely open to social criticism. The need for definitive rulings in morally salient cases derives from the fact of moral disagreement, the fact that there is rarely any one coherent system of positive morality, particularly where the culture encourages critical approaches and individuals have some autonomy in making up their own minds as to what they believe ought to be punished or recompensed. The political need to reach a working agreement on matters with such potential to generate conflict and disorder gives a role for the positivist idea of law as a system of mandatory rules whose authority depends on their adoption as law and not on their specific content.

Other moral arguments in favor of such a formalist conception of law are powerfully presented by Lon Fuller (1964) who demonstrates the connection between what he calls governance of rules, rules which are clear, promulgated, prospective, practicable, and stable, and a society in which there is respect for autonomous moral agents who can reasonably be held responsible for their actions and accept that their obligations are part of a mutually beneficial political system. Fuller calls his theory a procedural form of natural law. Its formal requirements of 'good' law are part of a morality of mutual respect. He goes further than this and argues that a government which abides by the requirements of good procedural law will thereby tend to produce good substantive law. In a classic exchange with Fuller, H. L. A. Hart pointed out that Fuller's procedural natural law is 'compatible with great iniquity' (Hart 1957). However, the general thesis that there are moral reasons for government in accordance with rules has clear affinities with prescriptive legal positivism, and it can be argued that Hart himself includes elements of this prescriptive approach (in a largely descriptive analysis) when he describes the benefits which arise for social order when the informal 'law' of simple societies is supplemented by 'secondary rules' which enable official determinations about what the primary rules of a society are and who has the right to interpret and change them (Hart 1961, pp. 77–96).

1.3 Enforcing Morality

Issues of law and morality as often framed in terms of whether or not law should enforce morality. It is generally accepted that there is a sense in which morality cannot be enforced because truly moral conduct involves the willing acceptance of moral obligations. Thus, Immanuel Kant defined moral goodness as doing what is morally right because it is morally right and not, for instance, because the law requires it of us.

Further, it is rarely argued, as a matter of critical morality, that the law should require conformity to all and only conduct that is morally correct. Natural law theory tends to identify justice and the common good as that part of morality which is binding on human beings in a way which makes it appropriately enforced through positive law. The theory also provides for governments to fill out positive law with lots of moral neutral ways which are required to give specific application to natural law values.

Classical legal positivists apply the general principle of utility to all moral questions, but point out that choosing a law is not making the same moral choice as choosing how to behave as an individual. Law involves sanctions and restricts liberty of action, which human beings enjoy and use to achieve their own happiness. Therefore, there are always moral reasons against enacting a law. This approach was developed by John Stuart Mill (1910) into the famous harm principle according to which only harm-producing conduct may be prohibited and then only when the negative effects of enforcing conduct are outweighed by the positive effects of reducing harm. He confined this harm to that which is caused to others, an anti-paternalist position which is a dominant strand in all liberal political philosophies.

This highly individualistic approach encouraged by the utilitarian moral theory is standardly countered by various forms of communitarian analyses of society, including natural law theory. This conflict is well illustrated in a classic debate between Lord Devlin (1965) and H. L. A. Hart (1963) on the matter of decriminalizing homosexual conduct between consenting adults. Devlin argues that the basic institutions of a society, such as marriage, have to be protected in the interests of that society as a whole, therefore, it is treasonable to allow forms of conduct which undermine such foundational social institutions. It follows that society has a right to enforce a particular type of sexual organization. This was a response to H. L. A. Hart, who takes the view that sexual conduct of this

sort is a purely private matter so that any disgust which the majority might feel towards such conduct has no bearing on the moral question of whether or not homosexual conduct should be a criminal offence.

Debates of this sort may be approached in a number of ways. Sociologically, Devlin's communitarian point can be investigated to see whether it is the case that the preservation of monogamous marriage as the dominant social institution with respect to sex and reproduction does require prohibition of incompatible forms of sexual conduct. Other arguments center on the concept of toleration, the idea that it is morally desirable to allow people to follow out their chosen way of life even when you think that they are wrong to behave in this way. The arguments for toleration are complex (Mendus 1988). For some people toleration is an acknowledgment of the significance of moral goodness. For others it is simply morally right not to interfere with the conduct of others when this does not harm our interests. For others it is a response to moral disagreement and the difficulty of reaching an objective decision about which competing form of life is morally superior. A more extreme form of this position is that choices between competing ways of life cannot themselves be moral choices, since all moral choices have to assume values which arise from a particular set of cultural circumstances. All these approaches neither permit nor encourage a pluralism of positive moralities which make it impossible for the law to enforce morality without a choice being made as to whose morality or which morality is to be enforced.

Responses to the problem of choosing which morality to enforce can be broken down into at least three categories.

(a) Those which aspire to a neutral principle which is not moral at all. Mill's harm principle has been presented in this light, but it is clearly part of the utilitarian maxim which values pain negatively. Indeed, it is clear that many theories smuggle moral ingredients into the concept of 'harm,' identifying it with the approved interests of those affected by the actions of others.

(b) Those who seek to distinguish between private and public spheres, arguing that the morality which pertains to law is public morality, that is, morality about political and state relationships, while private has to do with the individual and her freely chosen associates. This shifts the argument to what is public and what is private, a principal battlefield being the sphere of family relationships. Feminists have powerfully pointed out that to classify the family as private is to leave this male dominant sphere unregulated. Ultimately 'private' comes to mean simply those matters which the law ought not to deal with, making it uselessly circular as a principle for determining the proper sphere of the law.

(c) Those who seek to identify a certain number of basic rights which while not morally neutral, allow for an extensive range of different forms of life to flourish as each individual and group pursues its own conception of what is good. This is the position developed by Rawls in *A Theory of Justice* (1972) and enunciated by Dworkin in a series of works which represent the most sustained recent attempts to identify a morality which the law ought to enforce (1985, 1996). His position allows us to say that judges should, within the confines of inherited legal decision, exercise their own critical morality in determining a point of law. These theories do not however limit law to the enforcement of basic rights, rather they permit law to enforce conduct as long as this enforcement does not violate these rights. If they succeed then we can have a form of natural law theory which makes it a necessary condition of legality that a mandatory rule does not violate a human right, leaving considerable scope for the generality of law to be determined on democratic lines, the right to self-determination being itself a human right.

Such theories do not however enable us to achieve actual agreement as to what these rights are or should be. We are therefore left with the problem of determining whose system of rights to enforce. In effect this comes down to the question of who has the right to decide which rights are human rights. This takes us back to constitutional questions which are crucial to contemporary arguments about law and morality. If we believe that there is a moral method which enables us to identify human rights in an objective manner, then it is reasonable to enact by plebiscite or legislative processes a general Bill of Rights which is then filled out and applied by courts in their determination as to whether the law produced by elected assemblies do or do not violate these rights. This may be regarded as a natural law position since it assumes an identifiable law-like morality particularly related to fundamental justice. It therefore makes sense to have rules of recognition which list these rights and leave it to the critical morality of the judiciaries to give them concrete application.

If we are sceptical about our capacity to solve disagreements about fundamental rights in this way and give high priority to the right of self-determination, then we can draw on a prescriptive form of hard legal positivism and suggest that it is for elected assemblies and popular plebiscites to determine the content of all law in the light of the critical moral judgments of the people as a whole, and the duty of the courts to apply such laws without reference their own moral views, in which case the morality which is in the law is determined independently of and externally to the process of law application.

See also: Causes and Laws: Philosophical Aspects; Common Law; Injustice: Legal Aspects; Justice and Law; Law and Morality: A Continental-European Perspective; Law: Anthropological Aspects; Law: History of its Relation to the Social Sciences; Law: Overview; Natural Law

Bibliography

Aquinas T 1966 [1269] *Summa Theologica*. Blackfriars, London
Austin J 1954 [1832] *The Province of Jurisprudence Determined*. Noonday Press, New York
Baier K 1958 *The Moral Point of View*. Cornell University Press, Ithaca, NY
Bentham J 1970 *An Introduction to the Principles of Morals and Legislation*. Athlone, London
Beyleveld D, Brownsword R 1989 *Law as Moral Judgment*. Sweet & Maxwell, London
Campbell T 1996 *The Legal Theory of Ethical Positivism*. Dartmouth, Aldershot, UK
Coleman J L 1982 Negative and positive positivism. *Journal of Legal Studies* **11**: 139–64
Devlin P 1965 *The Enforcement of Morals*. Oxford University Press, Oxford, UK
Dworkin R 1977 *Taking Rights Seriously*. Duckworth, London
Dworkin R M 1985 *A Matter of Principle*. Harvard University Press, Cambridge, MA
Dworkin R 1996 *Freedom's Law: The Moral Reading of the American Constitution*. Harvard University Press, Cambridge, MA
Finnis J 1980 *Natural Law and Natural Rights*. Clarendon Press, Oxford, UK
Fuller L L 1958 Positivism and the separation of law and morals: a reply to Professor Hart. *Harvard Law Review* **71**: 597–639
Fuller L L 1964 *The Morality of Law*. Yale University Press, New Haven, CT
Hart H L 1957 Positivism and the separation of law and morals. *Harvard Law Review* **71**: 593–629
Hart H L 1961 *The Concept of Law*. Clarendon Press, Oxford, UK
Hart H L 1963 *Law, Liberty and Morality*. Stanford University Press, Stanford, CA
Kelsen H 1967 *The Pure Theory of Law*. Trans. by Max Knight, University of California Press, Berkeley, CA
Lyons D 1984 *Ethics and the Rule of Law*. Cambridge University Press, Cambridge, UK
MacCormick D N 1989 The ethics of legalism. *Ratio Jurisdiction* **2**: 184–93
Mendus S 1988 *Justifying Toleration: Conceptual and Historical Perspectives*. Cambridge University Press, Cambridge, UK
Mill J S 1910 [1859] *Utilitarianism, On Liberty and Representative Government*. Dent, London
Postema G J 1982 Coordination and convention at the foundations of law. *Journal of Legal Studies* **11**: 165–203
Rawls J 1972 *A Theory of Justice*. Clarendon Press, Oxford, UK
Raz J 1979 *The Authority of Law*. Clarendon Press, Oxford, UK
Waluchow W L 1994 *Inclusive Legal Positivism*. Clarendon Press, Oxford, UK

T. Campbell

Law and People with Disabilities

Understanding the relationship between law and disability requires an analysis both of the idea of disability and of the role and function of law. Law is primarily concerned with the regulation of social behavior and is implicated in the orderly function of society; it constitutes, constructs, and applies social values. Law relevant to people with disabilities can be found in legislation and policy developed at a national or local level, as well as in international instruments. The complexity of the law stems from the fact that the law relating to people with disabilities consists not just of laws specifically directed at people with disabilities, but also of the full range of laws.

1. The Scope and Function of Law

Laws targeting people with disabilities include anti-discrimination law, laws relating to competence, laws relating to the provision of services, guardianship law, and the regulation of mental health. They also include laws of general application which are relevant to people with disabilities such as family law, criminal law, social security law, and international human rights law. In each case the law may play a role in the lives of people with disabilities by defining disability, by excluding people with disabilities, by regulating activities relating to disability, or by providing remedies for wrongs suffered by people with disabilities. Law also has the potential to promote an inclusive society by articulating the rights of people with disabilities and by constructing a legal framework for the achievement of substantive equality (Jones and Marks 1999).

2. Approaching Disability

The inconsistency in the values contained in law is a reflection of the different views held about disability at any given point in time, as well as over time. For many years the dominant approach to disability centered around what has become known as the 'medical model' of disability. More recently the social model has come to the fore, with minority rights and identity politics being adopted as strategies for change. A human rights model related, but not identical, to the universal approach to disability is beginning to emerge as an alternative mechanism for improving the position of people with disabilities. These models coexist and each model has different implications for the relationship between law and disability.

The 'medical model' of disability invokes a 'common-sense' definition of disability (Bickenbach 1993; Shakespeare 1999). Disability is conflated with impairment and incapacity and is seen as an individual deficit. Disability is therefore viewed as a personal tragedy and people with disabilities are to be regarded with pity and concern. Medicine, rehabilitation, and social welfare are mobilized to overcome or alleviate the consequences of impairment and to provide a basis for identifying those in need of support. Adopting a

negative view of disability and of people with disabilities legitimizes strategies such as recommending abortion of disabled fetuses, relegating people with disabilities to institutions, sterilizing them for eugenic purposes, or denying them citizenship rights.

The 1980 World Health Organization International Classification of Impairment, Disability, and Handicap (ICIDH) exemplifies the medical approach to disability. Although it attempts to separate the person from the biomedical condition by defining impairment as 'any loss or abnormality of psychological, physiological, or anatomical structure or function,' disability as 'any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being,' and handicap as 'a disadvantage for a given individual, resulting from impairment or a disability, that limits the fulfilment of a role that is normal (depending on age, sex, and social and cultural factors) for that individual,' it nonetheless fails to locate any aspect of disability in the social and political environment. It is recognition of the limitations of the medical approach that has led to the current revision of the classification schema.

The challenge to the medical model has come from those claiming that reducing the person to their 'abnormality' has the effect of legitimizing the exclusion of people with disabilities from the benefits of society. The social model of disability (Oliver 1990) assumes that disability is not internal to the individual, but is the external response of society to bodily or behavioral difference. As such, disability is considered to be socially produced and what is disabling is not so much the impairment as the complex array of conditions, activities, and relationships created by the social and political environment. Disempowerment and exclusion of people with disabilities are therefore viewed as the direct result of the exercise of power, through political and cultural practices, which bears little relationship to the reality of physical or intellectual difference (Barnes 1998, cf. Shakespeare 1999). The politics of disablement locates the source of the disadvantage in the failure of the social environment to accommodate difference rather than in the failure of the body (Thompson 1997).

While the medical model of disability focuses solely on the impaired body and ignores the barriers to participation arising from legal, social, economic, and political forces, the social model suggests that it is these forces alone which are responsible for the disadvantage arising from disability. This has led to a call for theorists to rethink the social model so as to 'include all of our lives' (Crow 1996, Wendell 1996). While the social model discloses the barriers to full participation, it offers no insight into the means of resistance to exclusion, nor strategies for change. The minority rights approach to disability (Zola 1994), on the other hand, provides a framework for social action.

Drawing on the experience of the civil rights movement, disability activists have invoked identity politics in order to encourage people with disabilities to see themselves as members of a minority social group (Hahn 1987). His civil rights strategy encourages people with disabilities to work together to challenge their common experience of oppression and recommends legal strategies similar to those used in other civil rights movements. Antidiscrimination laws are seen as an appropriate response to discrimination experienced, while the role of education in achieving equality is not ignored. However, the minority rights model is problematic because it attempts to equate all disabilities/impairments with each other. Yet the experience of those, for example, with visual impairments is not the same as the experience of a person with mental illness or an intellectual disability. Further, the historical treatment of different disability types has been inconsistent. Hence, the analogy between people with disabilities and women or racial minorities does not hold (Rioux 1994). Moreover the minority rights model ignores the reality of multiple identities and suggests that people need to focus on their difference in order to achieve equality and the difference to be demonstrated is 'abnormality' (Corker 1999).

The human rights model starts from the point of view of the universal entitlement of all people to rights, equality, and justice (Rioux 2000). People with disabilities, like all other people, are entitled to have their rights recognized and to be treated with dignity and respect. An inclusive theory of human rights takes the celebration of diversity and difference to be fundamental and equality to relate to both the process itself and the outcome of that process. An inclusive human rights model does not accept formal equality as an appropriate standard, nor does it accept that equalization of opportunity will result in inclusion. People with disabilities can only have their human rights respected when society makes accommodations for difference, not just when asked to do so, but at every step along the way. This model does not elevate people with disabilities to a special status nor require that the primary identity of the person be that of a person with a disability. Rather, the human rights model proposes a fluidity of treatment and a flexibility in policy and practice. The needs of all people are to be equally acknowledged and respected such that accommodations will be made independent of status of the person.

3. *Legal Responses to Disability*

There is more than one perspective of the relationship between law and disability. One view is that law has the potential to empower people with disabilities by fostering an inclusive society. This is the objective for

law sought by people with disabilities who see law as an important strategy for achieving justice and equality. Another perspective focuses on the way in which law is implicated in the construction of values around disability. Law is a significant institution of state power and has a hegemonic force. The reflection of values in law has a legitimating effect over and beyond the particularities of the legal issue in question (Jones and Marks 1999).

Two examples of the importance of interrogating law's relationship with disability will suffice. First, while the medical model is considered to be disempowering by social theorists, it is the model most frequently adopted in law. Where the law provides goods and services to people with disabilities, including pensions and other financial support, it will usually be contingent upon classifying the person with a disability. In this way, the law has a gate-keeping function and is designed to ensure that only specified individuals benefit. The process adopted by law will often be disempowering to the individual concerned, as he or she will be required over and again to prove 'abnormality' and to focus negatively upon his/her difference before any support will be forthcoming. In many circumstances medical practitioners will be called as experts to provide evidence about the disability.

The second example of the role of law in constructing disability relates to the many ways in which law imports ideas of competency in order to include or exclude individuals. Rules of competency come into play in matters as diverse as voting rights, access to justice, rules of evidence, medical treatment, criminal responsibility, rules relating to sexuality, the age of consent, and the care and custody of children. A true understanding of the relationship between law and disability requires a thorough analysis of the operation of law in these areas. Further, attitudes toward people with disabilities and the role of the law in determining values in relation to disability can be seen through an assessment of competency in areas such as mental health law and the rules relating to guardianship of people with disabilities.

These two examples are indicative of the analysis that needs to be undertaken if we are better to understand the relationship between law and disability. It would be possible to do the same sort of analysis focusing on the way in which people with disabilities are differentiated from those without disabilities, such as with respect to education, or areas of law where people with disabilities are simply left out of the equation. Further, the differential impact of supposedly neutral rules on people with disabilities is largely hidden. Evidence of the way in which law values or devalues people with disabilities can also be found in an analysis of any law which, consciously or unconsciously, excludes a disability perspective.

There are two basic strategies invoked by those who believe that law has the potential to achieve equality for people with disabilities. The first of these is the antidiscrimination model, which is tied closely to the civil rights/minority rights perspective on disability. The alternate approach can be found in international law. This latter approach has the potential to be interpreted in line with the human rights model.

4. *Antidiscrimination Law*

Antidiscrimination law is designed to redress historic social and economic disadvantage of certain groups in the society. It has been used as a strategy to achieve equality with respect to race and gender, and more and more jurisdictions are adopting antidiscrimination law with respect to disability. Antidiscrimination law is an individual complaints based law, where action to challenge the discrimination may be brought in a court of law or, in many jurisdictions, to an administrative body which may have the power to use alternative dispute resolution techniques of arbitration conciliation and mediation. The assumption of antidiscrimination law is that it is possible to level the playing field by outlawing discriminatory practices. The focus is on equalization of opportunity in the hope that, once historic disadvantage has been remedied, the status of the individual will cease to be relevant in the application of law.

Each piece of antidiscrimination legislation outlaws certain types of discrimination, but in no case is all discrimination considered unlawful. The legislation is always limited to a specified target group, in the case of disability law to those defined by the legislation to be 'people with disabilities.' Antidiscrimination laws involve the action of an individual complainant (and in the case of some laws, groups of complainants) who brings an action against a specified alleged discriminator. The sorts of discrimination covered by legislation include both direct and indirect discrimination, as well as harassment. Direct discrimination involves a situation where a complainant is discriminated against because of his or her disability and can be understood in terms of a 'but for' test. Direct discrimination, then, occurs in circumstances where, if not for the status of disability, the person would have been treated differently. The concept of indirect discrimination, sometimes referred to as 'disparate impact,' involves discrimination where an apparently neutral law has a differential impact on people with disabilities. Indirect discrimination, then, is an attempt to respond to systemic discrimination which poses barriers to full inclusion. The prohibition of harassment is an important addition to direct and indirect discrimination, for the actions of an harasser may not constitute discrimination *per se* but may create an environment in which a person with a disability is unable to function. Some legislation also includes provision for affirmative action.

While a number of jurisdictions have introduced antidiscrimination law, the contrast between the United States and Australian legislation is perhaps the most telling (Quinn et al. 1993). The *Americans with Disabilities Act* (ADA) is based on the minority rights approach to disability and is designed to respond to the needs of a class of qualified individuals. In most cases under the ADA, court decisions revolve around the question of whether or not the applicant fits within the class of people with disabilities protected by the legislation. As a result, there is little jurisprudence on the question of what constitutes discrimination and when discrimination is unlawful. The Australian *Disability Discrimination Act* (DDA) approaches the question of disability discrimination from an entirely different standpoint. More in line with the human rights approach than the minority rights model, the Australian legislation defines disability so broadly that almost any member of the community could qualify as a person with a disability. This has allowed the courts to focus on the action of the alleged discriminator and to consider whether or not the discriminatory conduct is unlawful. As a result of the broad definition of disability in the DDA, there is no need to focus on the 'abnormality' of the applicant, or to prove that he or she falls within a particular classification. What is important is the action of the alleged discriminator (Jones and Marks in press).

Whatever form the antidiscrimination law takes, this approach to empowering people with disabilities is limited. First of all, antidiscrimination law perpetuates the status quo and assumes that the primary form of disadvantage confronting people with disabilities is discrimination. However, antidiscrimination law can neither deal with systemic issues nor bring about the fundamental change in values and attitudes which is required for a truly equal society. This is the case even where affirmative action is encouraged through the use of, for example, taxation law. Further, the technical requirements of antidiscrimination law limit its ability and potential success (Redman and O'Connell 2000). Antidiscrimination law depends upon there being an individual complainant and a named alleged discriminator. As such, many instances of discrimination are unable to be challenged by antidiscrimination law. The limitations of antidiscrimination law are well known and the issues for people with disabilities are the same as those experienced by other disadvantaged groups.

5. *Human Rights Law*

Human rights law provides an alternative legal approach for achieving equality for people with disabilities. The rights-based approach is found in international law and sometimes is incorporated into constitutions or statutes. For example, the Canadian Constitution specifically guarantees the equality rights of people with disabilities. Section 15 of the *Constitution Act* 1982 provides

> (1) Every individual is equal before and under the law and has the right to equal protection and equal benefit of the law without discrimination and in particular without discrimination based on … physical or mental disabilities
> (2) Subsection (1) does not preclude any law, program or activity that has as its object the amelioration of conditions of disadvantaged individuals or groups including those that are disadvantaged because of … or physical disability.

This guarantee applies to all Canadian law, at all levels, and covers both substantive and procedural rights and provides for affirmative action. The Canadian Supreme Court has recognized that section 15 provides protection against discriminatory impact as well as intent. In *Andrews* vs. *Law Society of British Columbia* 1989 154, the Supreme Court takes a contextual, effects-based approach which recognizes disadvantage as central to the analysis of discrimination. In addition, in Canada there is a federal *Human Rights Act* as well as provincial and territorial Human Rights Codes, which provide protection from discrimination, harassment, and reprisal, and extend into all realms of public and private activity (Rioux and Frazee 1999). The advantage of adopting a human rights approach is that any disadvantage, whether or not it takes the form of discrimination, can be addressed.

The international community has developed a number of international human rights instruments which have relevance for people with disabilities. People with disabilities can make claims under the International Bill of Rights—the *Universal Declaration of Human Rights*, the *International Convention on Civil and Political Rights*, and the *Convention on Economic, Social and Cultural Rights*. Equally, people with disabilities may take advantage of the *International Convention on the Elimination of All Forms of Racial Discrimination*, the *Convention Against Torture and Other Cruel, Inhuman, or Degrading Treatment or Punishment*, the *Convention on the Elimination of All Forms of Discrimination against Women*, the *Refugee Convention*, and the *Convention on the Rights of the Child* (Degener and Koster-Dreese 1995). While each of these international law instruments is said to have universal application, none but the last convention makes any specific reference to its applicability to people with disabilities. However, each of the treaties contains a nondiscrimination provision which requires the human rights instruments to be applied independent of the 'status' of the individual concerned and there is jurisprudence to the effect that 'disability' is included within the category 'other status' (e.g., *Vienna Declaration and Programme of Action*, para 63 (A/CONF.157/24, 25 June 1993)).

International covenants have greater legal force than other instruments of international law. Instru-

ments such as General Assembly resolutions, declarations, and statements of human rights bodies are nonetheless important both for their support for people with disabilities and for their value in the interpretation of the treaties themselves. There are a range of 'soft' international instruments including the *Declaration on the Rights of Disabled Persons*, *Declaration on the Rights of Mentally Retarded Persons*, and the *Standard Rules on the Equalization of Opportunities for Persons with Disabilities*. The last mentioned resolution of the United Nations General Assembly is by far the most significant and far reaching international instrument regarding the rights of people with disabilities. The *Standard Rules* adopt a social constructionist perspective of disability and call upon governments to provide for the equalization of opportunity for people with disabilities in all aspects of their lives. However, the *Standard Rules* do not fully incorporate a rights perspective as they are not directed at resolving systemic disadvantage.

There is no specific convention or treaty dealing with the rights of people with disabilities. There is some dispute about whether a convention on the rights of people with disabilities is strategically appropriate, for it is argued that by separating people with disabilities from the mainstream one falls into the trap of differentiating between people when the primary objective of a rights perspective is to include outsiders, rather than to treat their needs as distinct from the needs of all other members of the community. However, in the absence of a convention on the rights of people with disabilities, it has been extremely difficult to ensure that these rights are included on the general human rights agenda and to develop mechanisms to ensure a disability perspective is maintained with respect to the operation of rights in practice. A disability convention would not guarantee, in itself, the implementation of rights for people with disabilities. Such a convention would, however, offer a starting point for those seeking equality for all members of the community.

6. Conclusion

On one level, international law could be considered to be only as valuable as its enforcement mechanisms and the implementation measures of ratifying States. This could equally be the case with respect to domestic law. Even where enforcement and implementation measures are less than ideal, it could be argued that the very existence of law sets standards for the treatment and inclusion of people with disabilities in the community. Aspirational instruments have the strategic possibility of empowering people with disabilities. They can be used as leverage in negotiation, both with governments and with service providers, and can be held up as a blueprint for acceptable behavior. International law, on its own, cannot bring about equality for people with disabilities. Equally, the possibilities of domestic law are limited, but adopting a human rights approach creates the greatest opportunity for moving towards an inclusive society. While law has a part to play in providing rights for people with disabilities, it must not be forgotten that law also may be implicated in the creation of disadvantage. An appreciation of the relationship between law and disability therefore is essential before confidence can be placed in any legal strategy and before law should be invoked as a solution for the problems associated with disability.

See also: Disability, Demography of; Disability: Psychological and Social Aspects; Disability: Sociological Aspects; Discrimination; Discrimination, Economics of; Human Rights, Anthropology of; Human Rights, History of; Human Rights in Intercultural Discourse: Cultural Concerns

Bibliography

Barnes C 1998 The social model of disability: A sociological phenomenon ignored by sociologists. In: Shakespeare T (ed.) *The Disability Studies Reader: Social Science Perspectives*. Cassell, London and New York, pp. 65–78

Bickenbach J 1993 *Physical Disability and Social Policy*. University of Toronto Press, Buffalo

Bickenbach J 1994 Voluntary disabilities and everyday illnesses. In: Rioux M, Bach M (eds.) *Disability is not Measles: New Research Paradigms in Disability*. Roeher Institute, Toronto Canada, pp. 109–126

Corker M 1999 Differences, conflations and foundations: The limits to 'accurate' theoretical representation of disabled people's experience. *Disability and Society* **14**(5): 627–42

Crow L 1996 Including all of our lives. In: Morris J (ed.) *Encounters with Strangers: Feminism and Disability*. The Women's Press, London

Degener T, Koster-Dreese Y (eds.) 1995 *Human Rights and Disabled Persons: Essays and Relevant Human Rights Instruments*. Martinus Nijhoff, Dordrecht, The Netherlands

Hahn H 1987 Civil rights for disabled Americans: The foundations of a political agenda. In: Gartner A, Joe T (eds.) *Images of the Disabled/Disabling Images*. Praeger, New York

Jones M, Marks L A B 1999 Law and the social construction of disability. In: Jones M, Marks L A B (eds.) *Disability, Diversability, and Legal Change*. Martinus Nijhoff, The Hague, The Netherlands, pp. 1–24

Jones M, Marks L A B in press A bright new era of equality, independence, and freedom: Casting an Australian gaze on the ADA. In: Francis L P, Silvers A (eds.) *Americans with Disabilities*. Routledge, New York

Oliver M 1990 *The Politics of Disablement*. Macmillan, London

Quinn G, McDonnagh M, Kimber C 1993 *Disability Discrimination Law in the United States, Australia, and Canada*. Oak Tree Press, Dublin, Ireland

Redman R, O'Connell K 2000 Achieving pay equity through human rights law in Australia. *Australian Journal of Human Rights* **6**(1): 107–22

Rioux M 1994 Towards a concept of equality of well-being: Overcoming the social and legal construction of inequality. *Canadian Journal of Law & Jurisprudence* **7**: 127–47

Rioux M 2000 In: Blanck P D (ed.) *Employment, Disability, and the Americans with Disabilities Act: Issues in Law, Public Policy, and Research*. Northwestern University Press, Chicago

Rioux M, Frazee C 1999 The Canadian framework for disability equality rights. In: Jones M, Marks L A B (eds.) *Disability, Divers-ability, and Legal Change*. Martinus Nijhoff, The Hague, The Netherlands, pp. 171–88

Shakespeare T 1999 What is a disabled person? In: Jones M, Marks L A B (eds.) *Disability, Divers-ability, and Legal Change*. Martinus Nijhoff, The Hague, The Netherlands, pp. 25–34

Thompson R 1997 *Extraordinary Bodies: Figuring Physical Disability in American Culture and Literature*. Columbia University Press, New York

Wendell S 1996 *The Rejected Body: Feminist Philosophical Reflections on Disability*. Routledge, New York

Zola I 1994 Towards inclusion: The role of people with disabilities in policy and research issues in the United States: A historical and political analysis. In: Rioux M, Bach M (eds.) *Disability is Not Measles: New Research Paradigms in Disability*. Roeher Institute, Toronto, Canada, pp. 49–66

M. Jones and L. A. B. Marks

Law and Risk

Contemporary society can be described as a 'risk society' (Beck 1992). This term indicates a fundamental change in the relationship between society and nature. As in the nineteenth century, when modernization meant the dismantling of agrarian society, the 'second modernity,' as Beck calls the period at the end of the second millenium, has broken down the nucleus of industrial society. The emerging new forms of technology challenge society with a new type of risk (Perrow 1984). Social behavior creates new hazards in various situations but it is not able to cope with them adequately. Scientific and technological development mean new risks for every individual—for social groups, for minorities, for society as a whole, and for future generations. This is true, for instance, with respect to nuclear energy, information and communication technology, modern biotechnology, and genetic engineering. Moreover, social change, reflected in growing individualism, patchwork biographies, or changes in the organization of labor, can be understood as an expression of the risk society.

Usually the law is considered to be an instrument of risk regulation. Any damage done to another person can be judged in terms of the legal questions of guilt, liability, or compensation, for example. Criminal law, civil law, and public law provide individuals, companies and administrative bodies with a number of instruments designed first and foremost to prevent damage but also to recover compensation for its consequences. The principle of causality and the instrument of absolute liability can be seen as legal reactions to the typical forms of risk encountered in modern society. However, under the conditions posed by the 'risk society,' this simple notion of regulation begins to appear doubtful. Legal instruments fail to work when confronted with cognitive uncertainty. Scientific experts are no longer expected to produce valid knowledge for every single case; more often they demonstrate deep-rooted dissent over the issues of risk and inherent danger. In many cases, the economic as well as the technological capacity to manage risk is restricted. Therefore, in theory and practice, the question arises whether the law is really an adequate instrument for regulating risk (Jasanoff 1995).

1. *Information and Decision—Two Concepts of Risk*

The term 'risk' has two dimensions: information and decision. Both are legally relevant. Historically, the information-based understanding of risk dates back to the maritime trading companies of the early modern age. In order to minimize economic loss on long journeys overseas, merchants were forced to calculate possible damage. Therefore, the idea of risk was understood to be based upon sufficient information about the circumstances of an individual case. Risk was defined as the probability of the occurrence of some event causing damage, multiplied by the amount of the damage anticipated. The risk problem, therefore, was to forecast both quantities: the probability of the event and the projected financial loss. This concept of risk is fundamental to every insurance business; it is also the basis for legal risk management. The connection between the law, legal regulation, and risk is knowledge generation and certainty as the bases for such regulation. Where the law is used to prevent damage or to recover compensation for its consequences, we need risk assessment, that is, an estimate of probable risk upon which a legal judgment or decision will be based. This calculation provides the necessary certainty underlying legal attribution of events and their consequences. Let us illustrate the case with some examples. For instance, standards fixing maximum daily intake or exposure to potentially toxic substances, or determining the duration of exposure to x-rays, or setting maximum values for the hazardous emissions, give the law the certainty to adjudicate—to affix blame, guilt, liability, or degree of compensation. Combined with scientific expertise, the law creates predictability, that is, we know what to expect and we know what should or should not be done. One could say, then, that the law absorbs the risk inherent in everyday activities, as well as in economic ventures or administrative decision making.

Decision-based understanding of risk shows, on the other hand, that decisions are inevitably afflicted with

the problem of uncertainty and insufficient knowledge, and that there is never enough information to rid us of all risk altogether. Risk, then, is understood to be the tension resulting from the necessity to make a decision and the impossibility of attaining perfect justification for that decision. This means that, given a social situation where adjudication is necessary—where a decision is expected—any possible variation or deviation in behavior under these circumstances will be understood to be a result of that decision. In this case, even doing nothing constitutes deliberate behavior, and any possible consequence that could turn up subsequently will be attributed to the decision made. Here, the risk is the tension between not knowing completely and perfectly what events will result from the causal chain that begins with the decision and, simultaneously, being compelled to decide. In other words, risk is a characteristic dilemma between the temporal dimension and the social dimension of behavior. In particular, the decision aspect shows that risk is closely connected to the distribution of responsibility and the attribution of consequences to a specific decision. It is precisely at this point that regulation of risk becomes the focal point, because the law is not necessarily a perfect means to manage risk. The law says what ought to be done to somebody who has damaged somebody else but it does not lend much support to forecasting, to deciding which of two alternatives would lead to negative outcomes. In other words, under the condition of uncertainty and imperfect knowledge, it is doubtful whether the law can really contribute much to the prevention of risk. Under this supposition, the question is whether risk regulation is indeed part of the function and performance of the law in modern society (and, if it is, what form should it take).

2. *Stabilizing Expectations vs. Regulating Risk—The Function and Performance of the Law*

The function and performance of the law are distinct: function refers to the type of tasks the law must fulfill with respect to society as a whole; performance is used to describe the relationship between the law and other social subsystems, such as the political and policy arena, the economy, or the educational system.

The social function of the law is designed to stabilize social expectations. In general, we can distinguish between two modes of expectation, the cognitive and the normative. Cognitive expectations are open for correction; they are the basis for learning processes. For example, one may expect the Dow Jones to fall during the next 24 hours, but if that expectation proves false, one does not hold stubbornly and steadfastly to the same hypothesis: rather, one adapts one's estimate accordingly and corrects one's behavior. Cognitive expectations are not stable; they are open for modification or amendment (learning). The contrary is true for normative expectations. Normative expectations are upheld even in the case of deviant behavior. For example, one expects others to respect one's property. If others behave deviantly, in a manner disrespectful of one's property, one does not abandon this expectation: rather, the principle of 'respect for property' will be reiterated all the more emphatically. Normative expectations remain unchanged even *contra factum*, and it is counterfactual stabilization of expectations that reveals their normative character. With respect to society as a whole, the law is a means to stabilize such expectations. The law is usually guaranteed by a state authority with the power to enforce legal norms through sanctions. This function of the law, stabilization, makes its very structure relatively resistant to flexibility or to the modification of expectations that would seem appropriate for the decision aspect of risk. A negative outcome that results from a decision, disappointed or unfulfilled expectations inherent in the decision itself, can best be coped with through learning, not by clinging to false expectations. Some theorists argue that this demonstrates the impossibility of regulating risk with legal instruments (see Sects. 4 and 5).

The performance of the law in relation to other societal subsystems (that is, what the law does for them) is fivefold:

(a) Conflict resolution: competing social expectations can generate internal conflicts in societal subsystems, such as in science, the economy, or in the educational system. The law provides these subsystems with the ability to identify conflicts and to export them to the legal context. If a risky decision has harmful consequences, then science, research, and development can learn—theories can be modified or technical equipment revamped. The law takes over the task of conflict resolution and thus frees science to operate without normative interference.

(b) Regulation of conduct: the law provides the political system with a means to regulate individual behavior.

(c) Legitimation of the social order: the law defines the framework or sets the parameters for political action. Under the rule of law, this framework is commonly laid out in the form of constitution.

(d) The law permits planning and prevention of risk.

(e) The law can act as a means of surveillance.

Under each of these five aspects of performance, the law absorbs risks by providing rules for decision making for the other societal subsystems, or relieves others of the burden of justifying a decision. The question is how and to what extent the law is really able to provide this service to the other subsystems under the condition of risk. As pointed out above, under the condition of risk, the function of the law *qua* stabilizer of expectations over time tends to run contrary to its intended performance for the other subsystems.

3. Different Perspectives on Regulation—Theories of Risk in the Social Sciences and Consequences for the Law

The degree to which risk regulation may be deemed successful or not will depend upon the theoretical perspective one takes. Three main approaches can be distinguished: psychological risk theory, cultural theory, and systems theory.

3.1 Psychological Risk Theory

The psychological, cognition-oriented approach is based upon the difficulties arising from the dilemma of decision. If it is true that the information basis for a valid risk assessment is generally inadequate, then the question of how individuals come to make certain decisions about risk becomes increasingly significant (Jungermann and Slovic 1993). Psychological risk analysis seeks to determine how individuals judge the level of risk and decide upon some behavior in a given situation, and psychological risk research has generated some interesting findings. Voluntary risk is more readily accepted than involuntary risk. Acceptance of risk depends upon the amount of control one believes one has. The risks associated with new technologies are estimated to be much higher than those from older technologies. Risks are also judged to be more acceptable, the further away one perceives the consequences of a decision to be. After all, acceptance of risk depends upon a believable estimate of how easy or difficult it will be to repair damage. In terms of psychological risk theory, whether or not regulation of risk is possible will depend upon the impact of the law on individual behavior. Insofar as the law generates commitment, it may be able to influence or even to regulate individual behavior. Under this supposition, therefore, the law can be a suitable means to regulate risk within the scope of known risks. In such cases, the law can be used to program conduct with reference to science-based information. A basic problem nevertheless remains: how to reconcile the complex relationship between attitude, legal incentives, and sanctions on the one hand, and behavior on the other.

3.2 Cultural Theory

The decision problem shows that risk assessment cannot be solely understood as deliberate behavior. On this premise, cultural theory introduces social actors and conceptualizes risk as a collective construction. The concrete form of a risk concept depends on certain qualities of the social field in which this concept is shaped. Cultural theory devises a matrix of social fields based upon two dimensions (Douglas and Wildavsky 1982); 'Group' describes the external boundaries between a social collective and the outside world, while 'Grid' means all the other social distinctions determining how individuals behave towards one another. We can distinguish four types of social interest groups with distinct perceptions of risk and, correspondingly, four regulatory models:

(a) A hierarchical culture is characterized by high grid group scores. This means that it will handle risks through routine procedures and that it will probably use legal instruments to establish such procedures.

(b) An individualistic market culture is characterized by low grid and low group scores. It will incorporate risk into the domain of individual concern; calculated risk is accepted and the law is used to provide a framework for individual, competitive, and risky activities.

(c) An egalitarian (sectarian) culture is characterized by low grid and high group scores. Normally it has high resistance to risk and it is the culture most sensitive to possible hazards. The law is seen as an impediment to the right way to live, because it allows risky behavior in many cases.

(d) A fatalist culture is characterized by high grid and low group scores. Risks are perceived as imposed from the outside by others and they provoke passive resentment. On this view, legal regulation can hardly make any difference whatsoever.

3.3 Systems Theory

The most radical version of the decision dilemma can be found in systems theory, especially in the work of Niklas Luhmann (1993). Risk is understood as the tension between the temporal dimension and the social dimension of decision making. In modern society, the introduction of the concept of risk means that our relationship to the environment, to the future, and to social disparity have been redefined. First, risk communication indicates that damage previously perceived as a consequence of some natural phenomenon is increasingly attributed to social conduct (decisions). The question of accountability becomes a core issue together with the law as a means to distribute that accountability. Second, following on from that, the future becomes increasingly relevant for justifying social activities. It is no longer sufficient or appropriate to disregard the future; it must be treated as a calculable element. The issue of risk assessment and risk forecasting begins to infiltrate every part of the social system. The law does not seem able to deal with this kind of cognitive orientation towards the future. Third, risk communication indicates a social disparity based upon a division between those who decide and those who are affected by the consequences of the decision. This kind of disparity means that the concept of risk is most suitable for generating social movements where individuals place themselves on the side of the affected parties. They challenge policy makers and confront the law with their interests. On this view,

the law can be seen either as a means for political action or as a political obstacle.

In sum, according to all the theories, the problems of legal risk regulation will occur in the areas of information and behavior, the cultural determinants of behavior, and the temporal structure of legal communication. According to psychological and cultural theory, in certain situations and with respect to certain social fields, the law will be able to resolve conflicts, regulate conduct, practice surveillance, and, through all of these, underscore the legitimation of a given social order. Finally, according to systems theory, it is doubtful whether the law is adapted to and adequate for planning and prevention insofar as these require cognitive orientation.

4. *Options, Limits, and Risks of Legal Risk Regulation*

In all legal fields a dense network of risk regulation and control can be found (Bora 1999). Civil law provides us with instruments mainly to prevent or to compensate for damage. Contract law, liability law (including rules of evidence), and business law are continuously adapted upon the basis of new information and possible risks. The same is true for public law, especially administrative law, which is traditionally concerned with issues of technological risk. Administrative law has developed to become the standard instrument for dealing with environmental problems and policies. Technology assessment and environmental impact assessment have been instituted in numerous organizations, industry, and governmental or administrative agencies. In the field of biotechnology, for instance, a whole set of institutions concern themselves with regulatory risk and safety issues, including environmental protection agencies, food and drug administrations, and animal and plant health inspection services. In the USA, for instance, consider the Environmental Protection Agency (EPA), Food and Drug Administration (FDA), United States Department of Agriculture (USDA), and the Animal and Plant Health Inspection Service (APHIS), among others. In Germany, comparable organizations are the *Umweltbundesamt* (the German Federal Office of Environment), the *Biologische Bundesanstalt* (the national agency for biological affairs), the Robert Koch Institute, the *Bundessortenamt* (the national regulatory agency for plant and seed types), and the various *Pflanzenschutzämter* (plant protection agencies). Many forms of regulation can be observed, from the traditional—legal intervention, proscription and sanction—to voluntary agreements.

This highly developed system of legal risk regulation also has its limits. First, there is no absolute benchmark for risk regulation; the law always depends upon a comparison of risks, typically the comparison between a traditional, conventional technology and a new one. Second, legal regulation needs the criterion of causality, the rule of law stands in opposition to regulation based upon mere suspicion. Third, the task of regulating unknown risks confronts the law with the problem of reversal of the burden of proof. This, in turn, generates empirical and logical problems, because it is impossible to demonstrate empirically the nonexistence of a fact. Fourth, the law always requires concrete damage as a material fact. The politically motivated aim to guarantee environmental quality is insufficient. Fifth, there is typically no check-and-balance system for social needs within the scope of the law; legal norms are in fact the result of political disputes over social needs. Sixth, the law will therefore not be a suitable political instrument to curtail innovation, nor can it function as an instrument of innovation control in general.

Whenever we speak of limits, we must also address the special risk of the law itself. To an increasing extent the other societal subsystems (politics, the economy, etc.) dump their risks onto the law. Unless the law can somehow redistribute the risks, it runs the risk itself of being overloaded with risky decisions. This is the case, in particular when programming of the law via political decision fails—a frequently observed phenomenon. The citizenry expects even the most hypothetical dangers to be regulated or managed by the state, a fact that undermines the usual means of programming the law through new information, established knowledge, or causal attribution. Thus, the law will always be condemned to promise more prevention than it will be able to deliver. We can say, then, that as the regulation of risk becomes increasingly necessary, it becomes increasingly risky itself at the same time.

5. *Is Global Regulation Possible? Future Developments in Theory and Research*

The empirical and theoretical challenge for the future relationship between law and risk is globalization. Globalization is characterized by the worldwide perceptibility of communication, the worldwide relevance of events, and the plurality of world views. Under these circumstances, risk takes on new dimensions. Global warming, deforestation, and biological safety are problems that reach far beyond the range of regulatory instruments available to the classic nation-state. The law tends in some respect to become global law without a state. Is the law in the global age a reliable means to regulate worldwide risk, or will it, as some theories predict, be replaced increasingly by nonlegal instruments which focus more on cognitive expectation and the mechanisms of learning? The future of the law in the global risk society seems to be open.

See also: Administrative Law; Environment Regulation: Legal Aspects; Globalization: Legal Aspects;

Guilt; Law: Overview; Liability: Legal; Risk: Empirical Studies on Decision and Choice; Risk, Sociological Study of; Risk, Sociology and Politics of; Risk: Theories of Decision and Choice; Rule of Law; Scientific Evidence: Legal Aspects

Bibliography

Beck U 1992 *Risk Society: Towards a New Modernity*. Sage, London

Bora A (ed.) 1999 *Rechtliche Risikoregulierung. Form, Funktion und Leistungsfähigkeit des Rechts in der Risikogesellschaft*. Duncker and Humblot, Berlin

Douglas M, Wildavsky A 1982 *Risk and Culture: An Essay on the Selection of Technical and Environmental Dangers*. University of California Press, Berkeley, CA

Hiller P, Krücken G (eds.) 1997 *Risiko und Regulierung: Soziologische Beiträge zu Technikkontrolle und präventiver Umweltpolitik*. Suhrkamp, Frankfurt, Germany

Japp K P 1996 *Soziologische Risikotheorie. Funktionale Differenzierung, Politisierung und Reflexion*. Juventa, Munich, Germany

Jasanoff S 1995 *Science at the Bar: Law, Science, and Technology in America*. Harvard University Press, Cambridge, MA

Jungermann H, Slovic P 1993 Characteristics of individual risk perception. In: Bayerische Rück (ed.) *Risk is a Construct: Perceptions of Risk Perception*. Knesebeck, Munich, Germany, pp. 89–107

Luhmann N 1993 *Risk: A Sociological Theory*. De Gruyter, Berlin

O'Riordan T, Wynne B 1987 Regulating environmental risks: A comparative perspective. In: Kleindorfer P R, Kunreuther H C (eds.) *Insuring and Managing Hazardous Risks. From Seveso to Bhopal and Beyond*. Springer, Berlin, pp. 389–410

Perrow C 1984 *Normal Accidents: Living with High-risk Technologies*. Basic Books, New York

Starr C 1969 Social benefit vs. technological risk. *Science* **165**(19): 1232–8

A. Bora

Law and Society: Sociolegal Studies

Scholars who have sought to integrate law and social science have united under various banners, and the most important of these are 'law and society,' which refers mainly to the USA (Garth and Sterling 1998), and 'sociolegal studies' (Campbell and Wiles 1976), which refers mainly to Great Britain.

There are scholars in many parts of the world who share these identities, but the field, even in Britain, owes much to origins in the USA. The emergence in the USA relates to the fact that the legal profession has long held a dominant professional position in economic and state power, and social science disciplines in the USA developed very differently from their counterparts elsewhere. In the USA, the disciplines became more differentiated, more empirical, and more positivistic than in other settings. Those trained in the disciplines could challenge the hegemonic position of law in the USA, and this continuing challenge helped to produce the field of law and society. The relative success of this field in the USA made it attractive to others, including most notably the British. The two largest organizations are the Law and Society Association (LSA) in the USA and the Socio-Legal Studies Association in Britain. By 2000, the LSA had more than 1,500 members and the SLSA about 750. Both groups contain a diverse array of scholars, interests, and approaches. Common frames of reference include work on the legal profession, issues of gender and race, state regulation, courts and dispute resolution, and crime and criminal justice. These research topics define the mainstream of the field, which includes the full range of social science methods.

Recognizing the importance of the USA in defining the field of law is not inconsistent with seeing these fields as part of an international phenomenon. First, there are many scholars in a number of countries, including India, Japan, Korea, Europe, and Latin America who have been influenced strongly and have contributed to this field. The influence can be traced in large part to the law and development movement in the 1960s and 1970s, which promoted educational interchange and a range of interdisciplinary work on the legal profession and the courts in particular. Notable examples include the work of Rogelio Perez Perdomo in Venezuela, and Upendra Baxi and Rajiv Dhavan in India. Much work in the law and society tradition continues to be produced by these individuals and others.

A second international influence came from the fact that participants in the law and development movement from the USA, such as Lawrence Friedman, Marc Galanter, and David Trubek, took their insights from abroad and used them to develop theories and models that could be applied to the USA. The best-known of such approaches is Galanter's famous article on 'Why the haves come out ahead' (1974), which was in large part based on his experience in India.

More generally, a related but distinguishable field of 'legal sociology' took shape at roughly the same time that 'law and society' developed in the USA, and it continues to be important in Europe and elsewhere. The major institution is the Research Committee on the Sociology of Law, which is affiliated with the International Sociological Association and which also meets with the US Law and Society Association every five years or so. One of the pioneers of legal sociology in the USA, William Evan, joined with Adam Podgorecki of Poland—then visiting in the USA—to take the initiative in 1962 to form the Research Committee. The first president was Renato Treves of

Italy, who built the Research Committee as a distinct entity devoted to legal sociology. Again, while there have long been overlaps with law and society, the scholarship associated with the Research Committee is more 'European,' which can be defined as more theoretical and less empirical than that typically associated with law and society or sociolegal studies.

The development of 'law and society' in the USA occurred at almost the same time as the founding of the Research Committee, but the story begins with the Legal Realists, who used the growing prestige of social science and especially sociology in the 1920s to mount an attack against the formalism of the generation that controlled the legal academy at the time. Centered at Yale and Columbia, the Realists carried out empirical studies of the legal system and challenged the assumptions of those they labeled as 'formalists'—those who believed that the law was a self-contained system such that legal precedents actually dictated the results in legal cases. Among the prominent legal realists were Charles Clark, William O. Douglas, Jerome Frank, Underhill Moore, and Walter Wheeler Cook. Although the Legal Realist movement as such died by the end of the 1930s, its leaders had been absorbed into the mainstream of the legal profession, and the interest in social science had seemingly disappeared. The orientation of the Realists, however, helped set the stage for later developments (see *Legal Realism*; *Legal Formalism*).

A gradual revival of the interest in law and social science after World War II generated the people and momentum to create the Law and Society Association (LSA) in 1964. During this period the Ford Foundation invested substantially in empirical projects that focused on the legal system. One major grant was to the newly established American Bar Foundation to fund the Survey of the Administration of Criminal Justice, which was directed by Frank Remington of the University of Wisconsin. Another important grant was to the University of Chicago to support a variety of research projects, the best known of which was by Harry Kalven and Hans Zeisel on the American jury.

In the 1950s, one of the promoters of the legacy of Legal Realism, David Cavers, used his position as Associate Dean at Harvard Law School to influence a prominent alumnus, Walter E. Meyer, to provide in his will for the establishment of the Walter E. Meyer Research Institute of Law, which was to encourage law-trained researchers to use the tools of social science in their research. Once established in the late 1950s, the Institute made numerous grants over a 20-year period to help develop the field of law and society.

Even more than the Ford Foundation and the Meyer Institute, the Russell Sage Foundation's grants served to institutionalize these collaborative ventures in the 1960s. Prior to the time of the establishment of the LSA, itself promoted by the Russell Sage Foundation, the Foundation had already invested more than a million dollars in law and social science. The major investment began with a grant to the Center for the Study of Law and Society at the University of California at Berkeley, and other grants were made to establish centers at Wisconsin (1962), Northwestern (1964) and Denver (1964). Later in the 1960s, after the establishment of the LSA, the Meyer Institute and Russell Sage Foundation funded six years (1967–73) of summer institutes in Denver for legal academics in a program entitled SSMILE: Social Science Methods in Legal Education. The National Science Foundation program in law and social science began in 1972 and reinforced the idea that this set of topics comprised a distinctive field.

The 1960s and early 1970s represented a kind of golden age for law and social science. There was a strong relationship between what was happening in the production of knowledge and what was happening in the streets and in the civil rights movement outside of the academy. Social science provided a new professionalizing expertise that offered ways to manage the new social agenda. Sociological knowledge thus gained in value as a new generation helped to define a political agenda of concern with urbanization, race relations, poverty, and crime. This social science learning could be used to challenge the postwar complacency of law schools and the training and expertise of lawyers for governing the state.

Remaining followers of Legal Realism, including such figures as Cavers and Willard Hurst of the University of Wisconsin, used this rising prestige of the social sciences to renew their own attacks against legal traditionalism and formalism. As both a cause and a reflection of this challenge, a few philanthropic foundations made it their missions to build institutions that would fly the banner of this new sociological expertise. Their investments in learning and in building institutions succeeded in bringing a new expertise to law and to the activist state.

The period of relatively high prestige for scholars bridging law and social science was in retrospect quite brief. The 1980s again brought a period of rapid social change, but social science generally and 'law and society' in particular declined in relative prestige. Indeed, just as law and society helped to build and legitimate the activist state (and the role of law in its construction), the competing movement of law and economics provided much of the learning and legitimacy for the later turn away from social welfare and state activism.

The way that law and society linked with the activist state, and the tensions within the law and society field, can be seen in the stories of two of the Russell Sage centers—Berkeley and Wisconsin. The Russell Sage Foundation grant to establish the Law and Society Center in Berkeley in 1962 began a process that continued through the closing of the School of Criminology and the establishment of the Jurisprudence and Social Policy program at the law school in the late 1970s. Phillip Selznick was the key individual

in developing the idea of the Center. Selznick's early work focused on organizations as institutions (Selznick 1949), but he became increasingly interested in law and its relation to the normative lives of organizations. In 1969, Selznick et al. published *Law, Society, and Industrial Justice* (1969), written in collaboration with Philippe Nonet and Howard Vollmer and published by the Russell Sage Foundation. Drawing on his studies of institutions, Selznick and his collaborators examined the general institutional setting of labor management relations, asking how 'emergent law'—including due process—could be brought to and realized in the industrial setting as 'industrial justice.' Paralleling Selznick's own move from institutionalism toward legal and even jurisprudential issues, the book asked 'what it means to *legalize* an institution, that is, to infuse its mode of governance with the aspirations and constraints of a legal order' (Selznick et al. 1969, p. 8).

Others who joined the Center followed in Selznick's footsteps, advancing the study of law as an institution; these scholars included Sheldon Messinger, Jerome Skolnick, and Jerome Carlin. Within the law school, there was relatively limited but growing support for the Center and for sociolegal studies, in particular from Geoffrey Hazard and Caleb Foote. A new cohort of law professors, led by Sanford Kadish, began to move much closer to the Center later in the 1960s. Carlin initially decided to focus his research on solo practitioners, and the result was a pioneering account of that world, entitled *Lawyers on their Own* (Carlin 1962). He moved to the Bureau of Applied Social Research at Columbia, where he produced another major work, *Lawyers' Ethics* (Carlin 1966). He then moved to Berkeley to work at the Center and the sociology department. Beginning in 1963 at the Center, his first project explored the relationship between the legal system and the class system. The project, which also involved Messinger and Jan Howard, resulted in the well-known article in the first issue of the *Law and Society Review* on 'Civil justice and the poor' (Carlin et al. 1966), and a book on the same topic. The conclusions emphasized the unequal access to law by the poor and helped make the case for the redistribution of legal services. It fitted very well with the 'war on poverty' and the ferment that was building around new legal services programs. The research connected closely to the social and legal activism of the time.

Also in Berkeley, but less oriented toward the Center, was Laura Nader, one of the few women involved in the early years of LSA and also a creator (or re-creator) of the subfield of legal anthropology. Nader's Harvard dissertation was based on fieldwork in Mexico with a focus on law cases. She joined the Berkeley anthropology department in 1960 with a specialty in the anthropology of law, which she then helped to develop. She organized the first major conference on law and anthropology, which resulted in a book that helped define this emerging subfield of legal anthropology. She also organized her students into the Berkeley Village Law Project, which focused on villages and 'disputing,' and became very influential in law and society as well as in legal anthropology. Legal anthropology continued to be a very live field into the 1970s, contributing to the emerging interest in alternative dispute resolution.

Eventually the Center moved into the law school, in part because of the Berkeley administration's decision to close the School of Criminology. What was left of the criminology school was merged in 1977 with the Center for Law and Society under the auspices of the new Jurisprudence and Social Policy (JSP) program. The social scientists moved to the law school. The law school, for the most part, remained dominated by the legal mainstream, but the legacy of the origins in sociology remained evident—even in the jurisprudential component of the JSP program.

Wisconsin, in contrast to Berkeley and the other centers, already had a thriving interdisciplinary community in the 1950s. It centered in the law school and was built on Willard Hurst and the legacy of Legal Realism. The Hurst contribution can be traced to Lloyd Garrison, the Dean at Wisconsin Law School during the 1930s. At Wisconsin, Hurst worked enthusiastically with Garrison to develop first-year teaching materials for a course on 'law in society,' and he also invested substantially in the cohort of young professors in the law school. The group built an identity that helped to constitute law and society.

Following Hurst's lead, Stewart Macaulay also moved toward the sociology of law and the attack on legal formalism. Hurst also encouraged Lawrence Friedman, and many others, to write about law in the specific context of the state of Wisconsin, which led to the publication of *Contract Law in America* in 1965 (Friedman 1965). Macaulay and Friedman teamed up to produce one of the early sets of materials on 'Law and the behavioral sciences' (Friedman and Macaulay 1969). This Hurst-inspired group then expanded in the 1960s with the law school at the center.

One of the most important individuals was Harry Ball, a sociologist recruited initially to work on the ABF Survey of the Administration of Criminal Justice. Not only was he a key person in the Russell Sage grant, but also he was one of the leaders with Robert Yegge (Dean at the University of Denver College of Law) and Richard (Red) Schwartz (the first editor of *Law and Society Review*) in establishing the Law and Society Association in 1964. Along with Ball, Friedman, Hurst, and Macaulay, the law and society contingent grew to include Joel Grossman (Grossman 1965, Grossman and Grossman 1971), Joel Handler (Handler 1967, Handler and Hollingsworth 1971), Herbert Jacob (e.g., Jacob 1965), Stanley Katz (Katz 1971), Jack Ladinsky, Frank Remington, and others.

The Wisconsin tradition built around Hurst, which can be characterized as that of 'law as the dependent

variable' or, perhaps more accurately, 'law in context,' provides one enduring approach in the field of law and society. It is characteristic of the work of Friedman and Macaulay, for example, and others from the same group such as Joel Handler. To take an example from the early days of the 'war on poverty,' Handler criticized the proposition, identified with Yale's Charles Reich, that the legalization of welfare and the provision of due process hearings would be sufficient to deliver concrete gains for welfare recipients. Criticizing the faith in legal formalism, he suggested that a careful examination of social context could make sure that the law would deliver better substance. In the writings of Hurst and other followers, the idea that law is a dependent variable was not inconsistent with idealism or an active role for law. Educated lawyers aware of social and economic contexts could make law a much more effective instrument for social progress.

Wisconsin was able to maintain a critical mass within the law school, nourished by the other departments of the university. It was able to sustain a law and society tradition that self-consciously offered a different model than that found in the elite law schools, which remained outside of the LSA orbit. The group in Madison identified its approach with the LSA organization, with Friedman, Macaulay, Handler, Jacob, and Marc Galanter, who came later, all serving as presidents, and Joel Grossman, serving also as editor of the *Review*.

This University of Wisconsin approach—built around law professors—both complemented and competed with the more sociological and jurisprudential orientation in Berkeley. Wisconsin's center was dominated by the law school and the personality of Hurst, while Berkeley's center was initially dominated by sociology—especially Selznick. The law professors at Wisconsin developed their approach by emphasizing in their struggles with the legal establishment that law was the dependent variable. This strategy allowed them to use social scientists and social scientific tools to attack legal formalism and the legal establishment on behalf of the concept of 'law in context.' The idea, especially in Hurst's works, was that the context was necessary to make the law more effective and better able to solve social problems.

The Berkeley sociologists, in contrast, emphasized jurisprudence and legal philosophy, which they—in particular Selznick—could see that sociology lacked. Scholars of institutions, as Selznick argued, needed to bring law to those institutions. The sociologists also could emphasize their superior training for social research. The law professors, they suggested, were incapable of making serious contributions to important theoretical questions of concern to social science, which was necessary to solve social problems. When the Berkeley sociologists and the Wisconsin lawyers found themselves on similar terrain, however, the opposition tended to turn into one between law and social science. Depending on what position one took in this competition, therefore, the sociologists at Berkeley could be seen as 'merely sociologists,' lacking any real legal expertise, or they could be characterized as 'real scholars,' in contrast to 'atheoretical' or dilettante lawyers.

The relative positions can be seen in the approaches to legalization and procedural due process. While Handler, from the perspective of law and context, based his criticism of law professors such as Charles Reich on their neglect of the social factors that would limit the impact that welfare rights and hearings would have on the lives of the poor (Handler 1966), Selznick's *Law, Society, and Industrial Justice* (Selznick et al. 1969) emphasized due process as 'the primary source of concepts and doctrines used in bringing the rule of law to new settings' (Selznick et al. 1969, p. 274). The different approaches relate to the fact that Handler was initially using social science to challenge legal formalism, while Selznick was initially using law to show its relevance to the study of social institutions.

The same basic contrast between law and sociology was also evident in the approaches of the Walter Meyer Institute, which was close to law, and the Russell Sage Foundation, which was closer to sociology. The Law and Society Association was initially the creation of people who, in terms of this relative opposition, were much closer to Russell Sage and to sociology. Yegge was the only law professor among the major founders, and the Russell Sage Foundation provided the initial funds. The contrasting structural positions, however, should not obscure the close proximity of these players on the emerging field. These were the lawyers closest to social science and the social scientists closest to lawyers. Even at the personal level, the protagonists were all friends and collaborators in building interest in the field of law and society. They shared a position against legal and sociological traditionalism, but they disagreed about how to define the new field of law and social science—and the relative positions of lawyers and social scientists within that field. They tended naturally to want to define the field in terms of their own expertises and approaches.

By the mid-1960s, this general field of law and social science had not only developed a number of institutions and adherents, but also served to define and build much of the expertise of the state as the 'war on poverty' began and added further resources to the field. Many of the people intersected with the new set of state programs, and indeed the relative success of the field is reflected in the career successes of those who developed this field. At the time of the development of the LSA, in short, the expertise of law and social science—including but not limited to the LSA—was gaining legitimacy and moving rapidly into the state.

The field that was constructed and the LSA itself gradually tilted in favor of law. By the late 1960s law had incorporated enough social science to stave off the social science challenge. This shift in the center of gravity of the field of law and social science generally

was also evident in the 1970s in the relative power within the LSA of law professors and in particular professors linked to Madison. The presidents of the LSA were almost all law professors linked to Madison. A later indicator of the shift in the center of gravity occurred when the Berkeley Center for Law and Society was brought into the domain of the law school. Put another way, much of the investment in the field of law and social science in the 1960s ultimately found its way into the LSA and the law schools. Indeed, for a short period of time in the early 1970s, this movement seemed to have found a place at the core of the law.

The relative value of social science expertise declined in the 1980s. A know-how that served to help construct an activist state seemed to lose some of its relevance. Academic investment in economics by the mid to late 1970s tended to pay richer legal career dividends than investment in the disciplines that were built more into the LSA. Just as bright and ambitious people were drawn to social science in the 1950s, many were drawn to economics in an era where inflation and the state were considered the great enemies of progress. Economics seemed to define the problems and the solutions for the 1980s just as sociology did for the 1960s.

Nevertheless, once established, the LSA developed its own organizational dynamic. A second generation brought up within the LSA came to define the scholarship around the LSA as a 'field' distinct from law and from the social science disciplines. The link of the LSA to the progressive politics of the 1960s also helped to give the LSA an enduring progressive political imprint.

The next group of leaders in the LSA, followed an approach similar to that of the Wisconsin pioneers. Rather than emphasizing 'law and context' as an antidote to optimistic progressivism, however, the new generation—many educated at Wisconsin by the pioneers—took aim at the politically leftist legal scholars of the Critical Legal Studies movement and suggested that they needed more social science to understand and substantiate their leftist critiques. The 'Amherst seminar,' which produced much of the LSA leadership in the 1980s (including Austin Sarat, Susan Silbey, Sally Merry, Christine Harrington, Barbara Yngvesson, and Kristin Bumiller), showed how the 'constitutive' nature of law prevented the social change that had been sought in the 1960s.

As an ongoing organization with both established scholarly tools and an openness to those looking for an identity to challenge the legal mainstream, the LSA continues to thrive as an organization many decades after it was founded. The tensions and hierarchies present at its creation, however, continue to play a prominent role in shaping the organization and the field that it represents.

As the field entered the new millennium, it renewed the commitment to internationalization that characterized it in the 1960s and 1970s. There is a new law and development movement anchored in places such as the World Bank that adds momentum to the internationalization of the field. With unprecedented educational interchange both in the social sciences and in law, in addition, the US-centered field is opening up to many scholars from outside the USA. Much of the growth of the LSA in recent years, for example, has come from its international membership. It remains to be seen whether the field will continue to be a vehicle for primarily US-based scholarly approaches—dominated by the domestic tensions that have characterized it since its formation—or will become a more broadly based and cosmopolitan scholarly domain.

See also: Critical Legal Studies; Law and Development; Law and Economics; Law as an Instrument of Social Change; Law as Constitutive; Law: History of its Relation to the Social Sciences; Law, Sociology of; Legal Formalism; Legal Process and Social Science: United States; Legal Realism; Legal Scholarship

Bibliography

Campbell C M, Wiles P 1976 The study of law and society in Britain. *Law and Society Review* **10**: 547

Carlin J 1966 *Lawyers' Ethics*. Russell Sage Foundation, New York

Carlin J 1962 *Lawyers on their Own*. Rutgers University Press, New Brunswick, NJ

Carlin J, Howard J, Messinger S 1966 Civil justice and the poor: issues for sociological research. *Law and Society Review* **1**: 9

Friedman L 1965 *Contract Law in America*. University of Wisconsin Press, Madison, WI

Friedman L, Macaulay S 1969 *Law and the Behavioral Sciences*. Bobbs Merrill, Indianapolis, IN

Galanter M 1974 Why the 'haves' come out ahead: speculations on the limits of legal change. *Law and Society Review* **9**: 95

Garth B, Sterling J 1998 From legal realism to law and society: reshaping law for the last stages of the social activist state. *Law and Society Review* **32**: 409

Grossman J 1965 *Lawyers and Judges: The ABA and the Politics of Judicial Selection*. Wiley, New York

Grossman J, Grossman M 1971 *Law and Change in Modern America*. Goodyear Publishing, Pacific Palisades, CA

Handler J 1966 Controlling official behavior in welfare administration. *California Law Review* **54**: 479

Handler J 1967 *The Lawyer in his Community*. University of Wisconsin Press, Madison, WI

Handler J, Hollingsworth E J 1971 *The Deserving Poor: A Study of Welfare Administration*. Markham Publishing Co, Chicago

Jacob H 1965 *Justice in America: Courts, Lawyers, and the Judicial Process*. Little Brown, Boston

Kalven H, Zeisel H 1966 *The American Jury*. Little Brown, Boston

Katz S 1971 *Colonial America: Essays in Politics and Social Development*. Little Brown, Boston

Reich C 1964 The new property. *Yale Law Journal* **73**: 733

Selznick P 1949 *TVA and the Grass Roots: A Study in the Sociology of Formal Organization*. University of California Press, Berkeley, CA
Selznick P, Nonet P, Vollmer H 1969 *Law, Society, and Industrial Justice*. Russell Sage Foundation, New York

B. Garth

Law: Anthropological Aspects

Anthropologists approach law from a comparative perspective and examine it in social context. Law refers broadly to the rules and processes that regulate social life. Although there are frequently institutionalized systems for making decisions and enforcing them, legal anthropology does not restrict its attention to formal systems of courts nor to the law of the nation state. It also examines the legal ordering of villages, bands, and urban neighborhoods—places, which usually lack specialized legal institutions and codified rules. The anthropological study of law began from nineteenth-century evolutionary theories in which the form of law revealed the stage of social development. Twentieth-century legal anthropology examines processes of resolving disputes and maintaining social order. Law is understood as a form of power. Recent scholarship focuses on the relationship between law and culture, law and language, and the intersections among plural legal orders rooted in the community, the state, and the region. It examines the contemporary legal complexity produced by postcoloniality and globalization.

1. *Historical Developments in the Anthropology of Law*

The great nineteenth-century theorists of social evolution used law as a measure of civilization and savagery. Sir Henry Maine, for example, theorized that with the shift from kin-based to territorially-based societies, there was a change in the law affecting persons from one based on status to one based on contract. In Emile Durkheim's evolutionary scale, societies ranged from 'primitive' societies held together by mechanical solidarity—the sameness of all constituent members—to complex societies organized by an organic solidarity between different and specialized but interdependent members. Societies with mechanical solidarity are characterized by repressive, punitive law that maintains cohesion, while those with organic solidarity have restitutive law and damages arranged through more specialized tribunals and functionaries. Although Durkheim's theory is empirically wrong its tendency to link civilization with the form of law affected subsequent debates within anthropology about the definition of law and its link to 'civilization.' In the early twentieth century, some anthropologists argued that law was universal, not characteristic only of people labeled 'civilized.' Bronislaw Malinowski, for example, argued that the Trobriand Islanders had no law in the sense of a state institution but governed their relationships with one another through reciprocity (1926). But the assertion that all societies had law raised the question of what 'law' was. During the early twentieth century, two major schools of thought developed. One school defined law in terms of institutions, requiring socially authoritative mechanisms for enforcing rules through the imposition of sanctions. Radcliffe-Brown pioneered this perspective, defining law as 'social control through the systematic application of the force of politically organized society' (1933, p. 202). Under this definition, all societies have customs and sanctions, but not all societies have law. E. A. Hoebel defined law more broadly, but also within an institutional context, as 'the legitimate use of physical coercion by a socially authorized agent' (1954, p. 26). Critics of this view of law argue that his definition is based on Western conceptions of law (Comaroff and Roberts 1981, pp. 6–7).

A second school of thought, following Malinowski, defined law in functional terms, identifying it with a broad range of social processes that produce social conformity through social pressure and inducement (1926). Reciprocity was a fundamental mechanism for maintaining social control. With Malinowski's definition, every society has law. Critics of this view contend that defining law as social control is too broad and includes practices such as gossip, ostracism, and nicknaming that are not law-like. By the 1970s, the debate over the definition of law became increasingly sterile and was largely abandoned in favor of understanding law as a social process (see Nader 1969).

2. *Law and Dispute Processes*

Between the 1930s and 1960s anthropologists sought to describe the nature of rules of whole societies. Yet they confronted the paradox that there were frequently discrepancies between these rules and everyday forms of behavior. In the early 1970s anthropologists turned to studying processes by which conflicts are handled and resolutions achieved (see Moore 1978). Laura Nader and her students researched how litigants pursued disputes, and how these disputes were resolved in a wide variety of societies (Nader and Todd 1978). This approach used extended case study analysis and emphasized the perspective of litigants rather than judges. It broke dramatically with the earlier focus on rules and looked instead at the processes by which individuals pursued grievances. At the same time, it built on a tradition of analyzing extended case studies in the context of social relationships. For example, Max Gluckman argued that close-knit villages tended to have compromise settlement processes, while more

loose-knit ones favored more adversarial processes (Gluckman 1955). The focus on dispute processes remains a basic framework for the anthropology of law. By the early 1980s, however, it became clear that a focus on the processes of resolving conflicts could not be separated from a consideration of rules and how they were formed. Comaroff and Roberts found in their study of the Tswana that the meaning and relevance of rules is negotiated in disputes (1981, pp. 5–17). Arguments among disputants are framed by rules, but the content of rules are negotiated as disputes are argued and resolved (1981, p. 19). This more interactive understanding of rules and processes has proved valuable for examining the operation of law in communities as well as in more formal settings, such as courts. It offers a way of understanding how different systems of rules can develop in different settings, a situation described as legal pluralism.

3. *Legal Pluralism*

In the contemporary world, societies typically have more than one legal system. Legal pluralism refers to situations of legal coexistence in a single social field. Initially described in colonial situations, in which the colonial power superimposed a European legal system over an existing indigenous system, legal pluralism is now understood as a fundamental characteristic of all legal systems. Advanced capitalist states as well as colonial and postcolonial states have multiple legal systems. Some are rooted in semiautonomous social fields that are self-regulated but also subject to regulation by the larger society, such as commercial networks (Moore 1978) or criminal communities. Many institutions such as schools and corporations develop forms of private governance. Informal but rule-governed relationships are characteristic of families, communities, and voluntary organizations, although there is considerable debate about whether such forms of normative ordering should be called law.

Early descriptions of plural legal systems examined each in isolation, but more recent work emphasizes the way plural systems shape one another. Thus, one legal system may offer the right to appeal decisions from another, or a litigant may be able to choose among several competing systems. Plural legal systems typically differ in power, so that the more powerful has the capacity to trump the actions of the less powerful. In colonial situations, the colonial authority often established one legal system for the subjugated and another for the colonial authorities (Moore 1986, Merry 2000). Local social groups such as schools, communities, businesses, and even gangs, have their own systems of law-like ordering, but these function within the framework of nation state law. The term 'interlegality' has been used to describe the interactions between multiple legal systems.

Globalization increases legal pluralism in two ways. First, it fosters the circulation of legal forms and practices from one nation state to another. Constitutions, codes, and legal institutions are commonly transplanted, much as they were during the earlier period of colonialism. These transplants introduce codes and procedures quite different from those of the receiving society, often with the intention of reform. Colonial transplants characteristically were phrased as furthering civilization; mid-twentieth-century transplants as modernizing institutions; and late-twentieth-century ones as promoting democracy and the rule of law.

A second way in which globalization increases legal pluralism is by constructing a global legal order. This includes international commercial law, fostered by the accelerating flow of transnational commercial activity. In recent years, global institutions which regulate international business, such as the International Monetary Fund and the World Bank, have become increasingly central to this legal order, along with a group of legal experts who help to manage international business transactions. The human rights system, made up of rules and sanctioning systems embedded in international institutions such as the United Nations, is another facet of the new global legal order (see *Human Rights, Anthropology of*). Although premised on a system of sovereign states, the pressure for compliance with human rights by states and transnational nongovernmental organizations is redefining the nature of state sovereignty itself. Local actors, such as indigenous peoples, make claims for legal and political sovereignty within the nation state, using international principles of self-determination and rights to protection of culture. The creation of new regional legal entities such as the European Union is also exploding the autonomy of local legal orders (Darian-Smith 1999).

4. *Law and Power*

Since the 1980s, the anthropology of law has become concerned increasingly with the relationship between law and the exercise of power (see Starr and Collier 1989). Inspired by Marxist and Foucauldian theory, and by the Critical Legal Studies Movement in law schools which argued that law reflects the interests of dominant groups rather than simply the logic of legal reasoning, anthropologists examined the way law supports relations of power in a wide range of social contexts (see Lazarus-Black and Hirsch 1994). But this research also demonstrated how law challenges dominant groups. It showed law to be a double-edged sword, expressing the rules and sanctioning authority of established political authorities and at the same time offering opportunities and justifications for resisting political authorities. For example, social movements that call on civil rights, such as the US civil

rights movement against racial discrimination in the 1950s and 1960s, challenge existing power hierarchies.

Feminists have used the criminal law to challenge male privilege by demanding penal sanctions for those who use violence against women in intimate relationships. The discourse of human rights has been deployed by indigenous peoples, women, victims of war atrocities, prisoners of war, and many other groups to claim protection from state authorities. Thus, law usually reinforces existing power arrangements, but can provide an avenue of resistance through its institutions and the discourse of rights.

Attention to the way legal forms and institutions exercise power, legitimate power, and resist established forms of power leads to a concern with the way laws are produced and enforced as well as with the linguistic and social exchanges which take place in courtrooms, police stations, and lawyer's offices (e.g., Conley and O'Barr 1990). Law constitutes social identities and understandings of social relationships through its regulations and through the spectacular as well as the mundane processes of arrest, trial, and mediation through which these regulations are enforced and failure to comply is punished. Even when laws do not impinge directly on social life, they exercise power through the ways in which they shape individuals' legal consciousness. This refers to the way people see themselves defined by and within law, the rights they can assert, and their entitlement to seek legal remedies for grievances (Merry 1990).

Law is a fundamental aspect of the exercise of power in colonial situations, since colonizing nations characteristically imposed a metropolitan legal code on to societies with very different legal orders. Such legal transplants were usually legitimated by goals of reforming family and community life. But they also privatized land ownership and created the basis for wage labor on plantations and mines. Several historically grounded studies have explored the complex relationships between this new European law and the shifting legal terrain of the colonized peoples of Africa (e.g., Moore 1986), Latin America (e.g., Collier 1973), and the Pacific (e.g., Merry 2000), among many others (see *Colonialism, Anthropology of*).

5. *Law, Culture, and Language*

Law is embedded in culture. The legal system of a society reflects its fundamental values and its normative order (Rosen 1989, Geertz 1983). At the same time, law constitutes cultural practices and defines basic identities such as citizen and criminal. Moral and religious understandings of self and community determine when and how people use law (Greenhouse 1986). The law also provides a rich vocabulary for describing and legitimating relationships of unequal power. It contributes in significant ways to the hegemony of a political system by engendering consent and fostering legitimacy for its exercise of power. Law is essentially a language and legal procedures are fundamentally linguistic. A body of recent work on law and language combines sociolinguistic analysis and law and society scholarship to look at the way law creates meanings and exercises power through linguistic exchanges in courtrooms and other legal settings (Conley and O'Barr 1998). This work provides a way to look at the detailed interactions that determine how the law exercises power. For example, many features of the kinds of talk which are persuasive in American courtrooms are more characteristic of male than female speakers, and of powerful rather than powerless speakers, so that male litigants in these courtrooms have an advantage (Conley and O'Barr 1998). But legal language and institutions also offer avenues for resistance. In particular, the language of rights has proved a powerful framework for social justice movements, especially since the end of the Cold War. Rights talk has been central to labor movements, indigenous peoples movements, the anti-apartheid movement in South Africa, and movements for women's rights, to name a few examples. Local groups around the world now draw on global conceptions of human rights for inspiration. Much of contemporary social justice talk and activism takes place within the domain of law.

6. *Conclusion*

The anthropology of law changed dramatically over the twentieth century. While earlier theories saw law as a structural feature of a whole society and used it to classify societies, more recent anthropological studies examine dispute processing, the linguistic and cultural features of legal settings that shape the way power is exercised, and the extent to which legal categories and concepts are embedded in individual consciousness and local culture. At the same time, the notion of culture as an integrated, homogeneous, and consensual system has been replaced by models of law in society which focus on the multiple and contradictory nature of rules and procedures. Contests among these systems of power and symbolism account for social change. Some of the competing frameworks are the product of alternative systems of legal regulation developed through colonialism or globalization. Contemporary legal anthropology views law as the product of contestation about meanings and identities in social fields constituted by multiple legalities. It examines the way law exercises power and sees this process as mediated by legal procedures, linguistic forms and exchanges, and cultural practices that shape legal settings and provide individuals access to these settings.

See also: Colonialism, Anthropology of; Human Rights, Anthropology of; Law and Society: Sociolegal Studies; Law, Sociology of

Bibliography

Collier J F 1973 *Law and Social Change in Zinacantan*. Stamford University Press, Stamford, CT
Comaroff J L, Roberts S 1981 *Rules and Processes: The Cultural Logic of Dispute in an African Context*. University of Chicago Press, Chicago
Conley J M, O'Barr W 1990 *Rules vs. Relationships*. Univ-ersity of Chicago Press, Chicago
Conley J M, O'Barr W 1998 *Just Words: Law, Language, and Power*. University of Chicago Press, Chicago
Darian-Smith E 1999 *Bridging Divides*. University of California Press, Berkeley, CA
Geertz C 1983 *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York
Gluckman M 1955 *The Judicial Process among the Barotse of Northern Rhodesia*. Manchester University Press, Manchester, UK
Greenhouse C J 1986 *Praying for Justice*. Cornell University Press, Ithaca, NY
Hoebel E A 1954 *The Law of Primitive Man: A Study in Comparative Legal Dynamics*. Harvard University Press, Cambridge, MA
Lazarus-Black M F, Hirsch S F (eds.) 1994 *Contested States: Law, Hegemony, and Resistance*. Routledge, New York
Llewellyn K-N, Hoebel E A 1941 *The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence*. University of Oklahoma Press, Norman, OK
Maine H 1917 *Ancient Law. Everyman's Library*. J. M. Dent and Sons, London
Malinowski B 1926 *Crime and Custom in Savage Society*. Routledge & Kegan Paul, London
Merry S E 1990 *Getting Justice and Getting Even: Legal Consciousness Among Working-Class Americans*. University of Chicago Press, Chicago
Merry S E 2000 *Colonizing Hawaii: The Cultural Power of Law*. Princeton University Press, Princeton, NJ
Moore S F 1978 *Law as Process: An Anthropological Approach*. Routledge & Kegan Paul, London
Moore S F 1986 *Social Facts and Fabrications: Customary Law on Kilimanjaro, 1880–1980*. Cambridge University Press, Cambridge, UK
Nader L (ed.) 1969 *Law in Culture and Society*. Aldine, Chicago
Nader L, Todd H F Jr (eds.) 1978 *The Disputing Process—Law in Ten Societies*. Columbia University Press, New York
Radcliffe-Brown A 1933 Law, primitive. In: *Encyclopedia of the Social Sciences*, Vol. 9, pp. 202–6
Rosen L 1989 *The Anthropology of Justice: Law as Culture in Islamic Society*. Cambridge University Press, Cambridge, UK
Starr J, Collier J F (eds.) 1989 *History and Power in the Study of Law: New Directions in Legal Anthropology*. Cornell University Press, Ithaca, NY

S. E. Merry

Law as an Instrument of Social Change

The relationship between law and social change is highly controversial among legal and sociological theories, between different philosophical orientations, and between legal cultures. There is no consensus regarding the role of law in past, present, or future developments. Consequently, opinions as to the impact of law in the processes of social reform differ according to the areas of law under scrutiny.

For some theorists law is only or mainly a dependent variable in society. Class structures, levels of economic development, culture, gender and so forth explain the specific characteristics of legal institutions and their changes. The opposite position conceives law as a tool for steering society in any possible direction. Human behavior is in their view largely determined by the existence of legal rules and sanctions. Intermediate positions which recognize both influences on the law and from the law are found in interdisciplinary studies, in particular in the field of the sociology of law (see *Law and Society: Sociolegal Studies*).

1. Law as Social or Supernatural

Western approaches share an understanding that law is part of the social world in its broadest sense. It emerges from social life or is intentionally created by social actors to steer society in intended directions. Law in the books is as much a product of society as is law in action. Depending on how law is defined, it is regarded as part of the culture or of the structure of society. But this view was never shared by all legal cultures and is at present mainly challenged by Islamic cultures. Muslim law—the *Shari'a*—is an integral part of Islamic religion. Divine revelation is its source. No authority in the world is qualified to create law or to change it, and not to obey Muslim law is a sin leading not only to criminal sanctions but also to punishment in another world. Hence state regulation through legislation encounters serious obstacles in some of the more orthodox Islamic countries and law contributes little to much needed social reform.

2. Differences Between Law and Society

Law is then necessarily different from society in Islamic logic. Western thinking also observes a necessary incongruence of law and society. In this case it does not follow from the divide between the secular and the divine but from the normative element characteristic for legal methodology. Although legal doctrine and legal practice in great part describe prevailing opinions and customs, they also develop new rules and always add value judgements (see *Legal Reasoning and Argumentation*). Normative expectations are by definition (Luhmann 1985) different from actual social behavior and a certain tension between them both exists in all societies. The gap between law and social practice is getting wider with increasing legislative activities observable since the middle of the nineteenth century, first in Europe and then in most other parts of the world. On closer inspection there are a number of

gaps (see *Legal Realism*) in every legal system: between higher courts and daily legal practice, between legislative acts and administrative routines, between lawyers and their clients, and between legally relevant behavior in different strata of the society (see *Law and Society: Sociolegal Studies*).

3. *Social Change*

Societies are complex networks of patterns of relationships constantly influenced and changed by new situations, new ideas, new social values, new technologies, and new innovations. Social change is a product of factors such as conflict and structural strains, economic pressures, political and ideological preferences, technology or the natural environment. However, none of these factors deserve undue weight as unique or prevailing 'cause.' Disregarding exceptional events like natural disasters, wars, or revolutions, societies change on the basis of a multitude of influences and do so permanently. Although 'modern' economic structures, cultural patterns, and technological innovations accelerate the social dynamic, traditional societies have always been able to adapt and to develop to a greater or lesser degree. Current social theories believe that change and disorder are normal and what needs sociological explanation is not disorder but exceptional phenomena like order, stability, and stagnation.

4. *Social Change and the Law*

Law is an element of society, and as such, is continuously changing with its surrounding social environment (see *Law: Change and Evolution*). More static societies like tribal communities and those societies which define law as divine show a smaller degree of permanent legal change, whereas 'modernity' is characterized by rapid change in most areas of law. Market economies create their legal support structures, grass root groups strive for civil rights or environmental regulation, and organized crime is a constant challenge for the administration of justice, creating pressures for more sophisticated legal devices. But the opposite is also observable: law influences or even creates social change. Most reforms initiated by state agencies use law as a tool for their policy aims.

The tension between law as a result of changing social demands or as a mover of social change is a much debated subject in the sociology of law (Vago 1997, Macaulay et al. 1995, Roach Anleu 2000). At issue is not so much the social control function of the law and even less the role of the law as a support structure for contractual relations. Rather, there is an apparent consensus that law is a key instrument for those purposes. However, it seems highly questionable how much law can change established traditions like discrimination against women, colored people, foreigners, or outcasts, or to change habits like the use of the automobile, or to establish a market economy in post-communist countries.

5. *Potentials for the Instrumental Use of the Law*

Since the rise of the nation-state, law was employed as one of the major instruments of political power, either criticized as a means of brutal oppression or welcomed as a democratic element in close connection with the rise of the *Rechtsstaat* or rule of law (see *Rule of Law*; *Rechtsstaat (Rule of Law: German Perspective)*; *Critical Legal Studies*). More and more areas of social life, from public affairs to intimate relationships, from local issues to global problems, are structured by laws emanating from municipal, regional, federal, international, or supranational legislative bodies. Courts intervene in virtually all social settings. Societies under the Civil Law complain about the 'juridification of social spheres' (Habermas 1984, 1987), whereas in the Common Law countries the litigation explosion (Olson 1991, Posner 1985) has become a major issue (see *Legalization*). These observations have led to a debate of a 'regulatory crisis' in Western societies pointing to inflationary use of legal tools, to predominant ineffectiveness of law in reaching its goals, and to wrong perceptions of law's 'realm' in society (Teubner 1992, Rottleuthner 1989).

6. *Theoretical Assumptions*

Although policy makers and legal scholars tend to assume that law is an effective instrument of social change and that the only relevant debate is about the way the law is modeled, sanctioned, or implemented (see *Compliance and Obedience: Legal*), social scientists struggle for the best explanation of the extent of behavior in conformity with the law (Fisahn 1999). Hobbes, as the forefather of utilitarianism and in particular of current economic theories of law, emphasized rational choice. Law is instrumentally effective if law-abiding behavior is more beneficial than deviant behavior (Posner 1992). Current sociology is heavily influenced by Durkheim (1964 [1893]) and Parsons (1951) who both explain legal behavior as expressing a value structure shared by all members of society. Legal rules are followed where they are supported by such values and rejected where they are not. The internalization of new normative patterns takes place in accordance with an existing structure of generalized expectations and of specific roles for specific social situations. Equally influential is Weber's (1954, originally 1913–19) theory of legitimation. In order to guide social behavior, law requires acceptance of charismatic, traditional, or legal authority. In 'modern' governmental structures, 'legality' prevails as the major legitimation mechanism: legal rules are accepted if formal procedures are followed in the legislative process. This emphasis on 'formal ration-

ality' is a dynamic but highly problematic aspect of the law in modern societies. Luhmann (1969) extends this idea to claim that the legitimation of law and legal decisions depends on participation in the decision-making process—rather than the outcome of that process. Potential protest is absorbed by participation as a voter in democratic elections or as a party in a court case.

Subsequently, Luhmann (1995) and Teubner (1993) rejected the very idea of law as a tool. According to an 'autopoietic' systems theory (see *Law, Autopoiesis in*), the study of effectiveness of law contracts into a study of legal discourses with those of other social systems. Giddens (1984) modified the prevailing sociological paradigm by explaining society as a reflexive process whose societal structures are less 'imperialistic' continuously produced and reproduced by individual actors. Daily routines not only follow existing rules but also create their own normative order. In this view rules are an instrument for achieving social change 'for everybody'—not only for those in power. Bourdieu (1990) develops similar assumptions about the individual actor as a structured and a structuring element—both being part of the 'habitus' which enables the actor relative freedom in a societal 'field.' The instrumental power of state law is, according to these 'postmodern' social theories, severely limited.

A universally applicable theory of the instrumental use of law is difficult to imagine since law is created in too many forms and is implemented (or not) in a great variety of situations. But further research can lead to situation-specific theories which may help to differentiate between both sociological and legal variations in legal regulation. Examples of research are reported in Daintith (1988) (law as an instrument of economic policy) and Gunningham and Grabosky (1998) (role of law in environmental policy). Each publication emphasizes the importance of a regulatory mix that includes nonlegal instruments as complementary or even dominant devices of policy implementation. Law is a useful and even necessary element in social reform processes, but more radical social changes cannot be accomplished solely by legal means. Economic incentives, structural and institutional support and educational programs are required for successful implementation strategies.

7. *Empirical Studies*

Measurement of legal regulation is difficult (see *Regulation and Administration*). Traffic signal behavior is observable but the impact of antidiscrimination law is not. Most legal initiatives have some direct effects and many indirect consequences which are hard to imagine and impossible to verify. Welfare laws regulate benefits for the poor but at the same time they have an impact on family structures, on the labor market, on the economy, and on all other goals of public policy which have to share with welfare a common public budget. Even direct effects are difficult to ascertain in complex societies where the new law might respond to a need in one sector and confound the interests of another. Some examples from different countries and legal cultures to show the kinds of information they can deliver:

(a) The impact of *Brown* vs. *Board of Education*, a US Supreme Court case decided in 1954, was studied by Rosenberg (1991). The judgment outlawed segregation in elementary and secondary schools which at that time was mandatory in 17 states and allowed in four others. According to Rosenberg's data, although the use of courts in the civil rights movement was considered the paradigm of a successful strategy for social change, courts had virtually no direct effect on ending discrimination in the field of education. Only when the Civil Rights Act was promulgated in 1964, and the executive branch acted forcefully to implement it, was segregation effectively ended.

(b) Like race discrimination, protection of the environment requires new and creative forms of state regulation. Two studies examine in great detail the regulatory efforts of German and English authorities to prevent pollution of open watercourses. Winter (1975) observed a single case in which a number of legal devices were used over a period of 25 years in order to stop the discharge of waste water into the river Rhein and lake Konstanz by three adjacent factories. The regulatory goals were reached only partially and after considerable delay and public investment. Winter explains this in terms of specific characteristics of German law and also the economic and political interests of the affected actors such as administrative agencies, business units, environmental groups, experts, and the local press. Hawkins (1984) was less interested in legal devices and structural influences on law enforcement than in the interaction processes between administrative bodies and industry. His research on two Water Authorities in England and Wales, based on interviews with field officers and representatives of industrial plants, found a distinctively conciliatory approach. Field staff, instead of initiating formal legal (criminal) proceedings against the polluter, enlist active cooperation. Decisions either to prosecute or to negotiate are made on the officers' own discretion and his/her own moral judgments. Indirect forms of control confer on field staff an authority which, for practical purposes, is in most cases ultimately persuasive.

(c) Independent India embraced equality as a cardinal value against a background of a culturally and religiously established caste system. The programs legislated after independence include compensatory discrimination (preferential treatment) for scheduled classes, scheduled tribes, and other backward castes as well as the opportunity for untouchables to use the courts in civil rights complaints. Galanter (1984, 1989) has evaluated their success. Although preferential treatment for victimized groups was carried out

forcefully by a complex system of reservations (granting of educational facilities, sharing of political power), Galanter observes considerable unfairness and negative side effects (not only for the upper strata of Indian society). A detailed study of court cases reveals an ambivalent picture. On the one hand, the High Courts where compensatory discrimination litigation takes place are highly respected and trusted institutions. The proceedings are inexpensive and quick. But litigation is initiated not by the beneficiaries of protective discrimination but by those complaining of schemes which affect their interests. The beneficiaries fail to get the benefits the law seems to promise. One of the reasons is the lack of adequate legal services providing beneficiaries with continuing and skilled support. After half a century of independence, India displays the effects of the limits of law in achieving social change: the gap between the economic opportunities of upper classes and upper castes on the one hand and scheduled classes on the other remains wide.

8. *Legal Exports to Defeated, Colonized, Developing, and Postcommunist Countries*

Historically, law has often been used as an instrument for achieving change in economic and social conditions in foreign countries (see *Law: Imposition, Reception, and Colonial*). Legal exports were part of colonial dominance and of postwar influence on defeated nations. In the 1960s and 1970s a US foundation and government programs attempted to use law to combat underdevelopment (see *Law and Development*). Similar programs have been initiated in the 1990s in post-communist countries by US scholars and also by European institutions. Since the fall of the Berlin Wall, many French, German, and Italian universities have been heavily involved in legal assistance activities; the Netherlands are exporting their new Civil Code to Eastern Europe and the European Community supports legal and educational reform programs in countries interested in becoming members of the European Union.

The assumption is that the setting up of an 'adequate' (Western) legal system including legal education and an efficient court structure is a precondition for rapid social and economic change (see *Markets and the Law*). Legal scholars are sent abroad, governments in developing countries employ Western legal experts; sometimes entire law codes are prepared by Western institutions or individuals and promulgated in the target country. Experience shows, however, that success is very limited and that the models have had at most a modest and incremental impact. In Latin America as well as in Eastern Europe, law and development programs proved to be counterproductive, leading to greater class differences, to distorted markets due to monopolistic structures, and to legal cultural rejection of imported models. Some authors talk of legal imperialism and the use of legal assistance as a Trojan colt (Gardner 1980). Oftentimes, the Western models are imported and legislated without the required legal infrastructure such as a well-educated legal profession, an honest and efficient court system, or some simple but necessary institutions like public registers. Positive examples of law and development activities (Japan in the nineteenth century, Germany after World War II, Poland and Hungary after 1990) point to a decisive intermediate factor in the instrumental use of law for development, namely the existence of a strong legal scientific community. Local legal scholars are able to filter foreign imports, adapt them to local needs, and teach the new law in the local cultural context. Since very few developing countries have such a strong academic community, law and development activities usually fail or remain incomplete.

9. *Instrumental Use of State Law*

Lawmakers have become aware of the contribution that empirically oriented social scientists can make to improve the instrumental use of law. Implementation and evaluation research accompanies many legislative initiatives. Critical studies in the fields of sociology of law, administrative science, sociology of deviance, and criminology influence regulative measures and judicial behavior. Concerns about the 'regulatory crisis' focus on dismantling forms of state intervention or on improvements in regulatory techniques including private ordering. The traditional focus on sanctions as a major resource in the use of law has been largely abandoned in favor of more participative methods. Lawmaking has sometimes been replaced by self-regulation (see *Legal Systems: Private*). Increased discretion has been given to administrative bodies as well as to judicial decision making. The overall result is a 'civil society' with state law as only one of many mechanisms to achieve social order and welfare. Social regulation takes place 'in the shadow of the law,' but it is not monopolized by the law.

10. *Law and Global Social Change*

The phenomena of globalization seem to support the argument that social change is a social rather than a legal process. Law only follows this change without the ability to take the lead and define the goals. Financial transactions are no longer limited by national borders, mergers of capital tend to monopolize global markets, communication is facilitated by the Internet, and environmental damage is caused on a global scale. The perspectives of law as an instrument of global change are rather somber. Most devices developed on the level of the nation-state for improving the implementation of regulatory law simply do not work on the global level where legislative

processes are deficient, common enforcement institutions and procedures are absent, and judicial institutions with global competencies are scarce (see *Globalization: Legal Aspects*).

There are legal initiatives which should nevertheless be mentioned (see *International Law and Treaties*). The GATT/WTO agreements have changed the conditions for global economic transactions and have created a quasi-judicial institution for handling disputes. In the area of Human Rights, a number of institutions such as the International Criminal Court in the Hague, the European Court of Human Rights in Strasbourg, and the International Labor Organisation in Geneva have been able to make some modest steps toward better protection of the individual. A number of international agreements attempt to prevent cross-border pollution, waste trafficking, global warming, and protect endangered species. Experiences in these fields might in the future provide useful know-how for global regulatory efforts.

Much more problematic are the effects of harmonization of regulatory law. One problem is the 'race to the bottom' which leads to regulation on the lowest levels of intervention in order to attract the greatest number of participants. Another problem is caused by differences in the application of harmonized or unified law. Even the EU, with its highly sophisticated regulatory structure, suffers from political or legal-cultural obstacles when EU law is transposed into law of the Member States (Duina 1999).

The globalization of law also takes place in the activities of nongovernmental actors (Gessner 1998, Appelbaum et al. 2001) who create what is called *lex mercatoria* (see *Lex Mercatoria*), who use their own private tribunals for the settlement of cross-border disputes (see *International Arbitration*), who make use of international law firms (see *Law Firms*) in order to set up new business structures and open new markets, or who draft codes of conduct, for example, concerning child labor or timber from tropical forests. Although expectations are high regarding these nongovernmental activities, few cases are reported where these activities have led to social change in a more narrow understanding—a change that replaces existing social structures and privileges with new ones and defines new priorities.

See also: Law and Society: Sociolegal Studies; Law: History of its Relation to the Social Sciences; Law, Sociology of; Social Change: Types; State and Society

Bibliography

Appelbaum R, Felstiner W F, Gessner V (eds.) 2001 *Rules and Networks—The Legal Culture of Global Business Transactions*. Hart, Oxford, UK

Bourdieu P 1990 *The Logic of Practice*. Polity, Cambridge, MA

Daintith T (ed.) 1988 *Law as an Instrument of Economic Policy: Comparative and Critical Approaches*. Walter de Gruyter, Berlin, New York

Duina F G 1999 *Harmonizing Europe—Nation States Within the Common Market*. State University of New York Press, Albany, NY

Durkheim E 1964 [1893] *The Division of Labor in Society*. Free Press, New York

Fisahn A 1999 *Natur-Mensch-Recht—Elemente einer Theorie der Rechtsbefolgung*. Duncker & Humblot, Berlin

Galanter M 1984 *Competing Equalities—Law and the Backward Classes in India*. University of California Press, Berkeley, CA

Galanter M 1989 Missed opportunities: the use and non-use of law favourable to untouchables and other specially vulnerable groups. In: Galanter M (ed.) *Law and Society in Modern India*. Oxford University Press, Delhi, India

Gardner J 1980 *Legal Imperialism—American Lawyers and Foreign Aid in Latin America*. University of Wisconsin Press, Madison, WI

Gessner V 1998 Globalization and legal certainty. In: Gessner V, Budak A C (eds.) *Emerging Legal Certainty—Empirical Studies on the Globalization of Law*. Ashgate, Aldershot, UK

Giddens A 1984 *The Constitution of Society—Outline of the Theory of Structuration*. University of California Press, Berkeley, CA

Gunningham N, Grabosky P 1998 *Smart Regulation: Designing Environmental Policy*. Clarendon Press, Oxford

Habermas J 1984 *The Theory of Communicative Action, Vol. 1, Reason and the Rationalization of Society*. Beacon Press, Boston

Habermas J 1987 *The Theory of Communicative Action, Vol. 2, Lifeworld and System: A Critique of Functionalist Reason*. Beacon Press, Boston

Hawkins K 1984 *Environment and Enforcement*. Clarendon Press, Oxford, UK

Luhmann N 1969 *Legitimation durch Verfahren*. Luchterhand, Neuwied

Luhmann N 1985 *A Sociological Theory of Law*. Routledge & Kegan Paul, London

Luhmann N 1995 *Das Recht der Gesellschaft*. Suhrkamp, Frankfurt, Germany

Macaulay S, Friedman L, Stookey J 1995 *Law & Society—Readings on the Social Study of Law*. Norton, New York

Olson W 1991 *The Litigation Explosion*. Dutton, New York

Parsons T 1951 *The Social System*. Free Press, New York

Posner R A 1985 *The Federal Courts: Crisis and Reform*. Harvard University Press, Cambridge, MA

Posner R A 1992 *Economic Analysis of Law*. Little, Brown & Co., Boston

Roach Anleu S 2000 *Law and Social Change*. Sage, London

Rosenberg G 1991 *The Hollow Hope*. The University of Chicago Press, Chicago

Rottleuthner H 1989 The limits of law: The myth of a regulatory crisis. *International Journal of the Sociology of Law* **17**: 273

Teubner G 1992 Regulatory law: Chronicle of a death foretold. *Social and Legal Studies* **1**: 451

Teubner G 1993 *Law as an Autopoietic System*. Blackwell, Oxford, UK

Vago S 1997 *Law and Society*, 5th edn. Prentice-Hall, Upper Saddle River, NJ

Weber M 1954 *Law in Economy and Society*. Harvard University Press, Cambridge, MA

Winter G 1975 *Das Vollzugsdefizit im Wasserrecht. Ein Beitrag zur Soziologie des öffentlichen Rechts*. Erich Schmidt, Berlin

V. Gessner

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Law as Constitutive

The idea that law not merely regulates but also constitutes or creates social phenomena has been a powerful component of recent sociologically oriented approaches to the study of law. Yet it is an old idea, resurfacing in social science and legal studies after having been marginalized through much of the modern development of sociology of law in the second half of the twentieth century. In its recent forms, the idea of law as constitutive focuses particularly on law's power to shape fundamental moral or political assumptions, and general conceptions of the nature of social relations and social institutions. In this sense, law is seen as a mechanism shaping consciousness, establishing frameworks of thought and evaluation in terms of which social life is structured.

1. *Classic Conceptions*

Classic legal sociology, represented especially by the work of the Austrian jurist Eugen Ehrlich at the beginning of the twentieth century, confronted the constitutive role of law as an urgent political question: to what extent can the law of the nation state define the nature of social relations in the face of powerful competing moral forces? (Ehrlich 1975). In a multicultural society (such as the disparate and fragmented Austro-Hungarian empire of Ehrlich's time) in which diverse religious and ethnic influences on social life remained strong and sometimes antagonistic to the values enshrined in state law, Ehrlich concluded that state law's constitutive power is limited and it must learn from, as much as seek to control, informal norms of social life. Many other theories in the formative era of modern legal sociology carried similar messages: law was either rooted in existing culture and established collective understandings, or was enfeebled by its separation from them. The idea is expressed in Sumner's (1960) view in 1906 that effective legislation must be consistent with social mores, in Pound's (1917) insistence on the 'limits of effective legal action' and in persistent echoes of von Savigny's (1975) claim in 1814 that law must be rooted in a 'common consciousness' (*Volksgeist*) of the people, which it expresses but cannot change.

According to all such views, because law is strong when it represents popular understandings and weak when it runs counter to them, its power to constitute social relations is circumscribed. Law has a degree of independent social power only in so far as it avoids intruding on matters of belief or deeply rooted traditional practice. Technically-oriented law can be effective, while expressive (ideologically charged or value-infused) law is relatively weak in effect unless the values and understandings it presupposes are already in place in social life.

Thus, an important agenda for sociology of law in the second half of the twentieth century became the exploration of the conditions under which law's regulatory techniques could be used to bypass the assumed inertia of culture, or to shape specific kinds of social relations in ways that might indirectly encourage more general social change. In an era (especially from the 1960s) in which urgent attempts were made to harness law to tasks of social engineering (for example, in preventing racial and sexual discrimination), the idea that law could only mirror established systems of values and beliefs symbolized legal sociology's practical task: to confront cultural obstacles that stood in the way of the successful use of law as an instrument of social change (see *Law as an Instrument of Social Change*). Law as an instrument of social control or social change was limited by its cultural dependence and by its general lack of power to constitute popular understandings, values, and beliefs.

Classical social theory supported this view of law obliquely only. Émile Durkheim's sociology treats law as merely one dimension of a larger moral universe. Law neither constitutes nor is constituted by moral understandings and moral bonds, but is an aspect of them; it is these bonds and understandings that constitute society (Cotterrell 1999). Max Weber, associating modern law with a particular kind of rationality, sees the development of law as running parallel to other kinds of rational reordering of Western experience (Kronman 1983), though his sociology remains ambiguous as to the nature of law's specific contribution to a more general modern 'rationalization' of life. Karl Marx's conception of law as part of the superstructure of a mode of production is also ambiguous about law's 'independent' constitutive role, but suggests that this role will, at best, be limited. Thus, while law may provide indispensable support (ideological and repressive) for a mode of production, it cannot, according to standard Marxist analyses, constitute such a mode, or change it in any historically fundamental way (see *Marxism and Law*. Nevertheless, law constitutes, in the sense of formalizing and defining, relations and understandings of contract and property. To that extent it contributes powerfully to ideology in capitalist societies.

2. *Modern Critical Perspectives*

The Marxist-derived idea of law framing consciousness (by shaping ideas of responsibility, ownership, freedom, consent, agreement, causality, authority, etc.) has been a very important basis of recent ideas of the constitutive power of law, especially as a result of its advocacy as a central tenet of the modern critical legal studies movement. Critical legal studies (CLS), developed in many Western countries and especially in the United States from the 1970s, had a major influence on the destruction of old ideas of law as a

passive reflection of other social forces (see *Critical Legal Studies*). Its effect on sociology of law has been to help to turn attention away from questions about how law might be made more effective as an instrument of social change, and towards studies of law's power to reinforce established social practices and institutions, primarily by influencing collective understandings. Thus, because of the continual invocation of basic legal concepts, especially those of property and contract, in everyday regulatory practice, law, in the CLS view, contributes importantly to the 'reification' of concepts, that is, the turning of debatable ideas into settled, widespread, usually unchallenged understandings.

Marxist legal theory in its most sophisticated forms (e.g., Pashukanis 1978) and modern critical legal theory emphasize that law shapes ideas of personality and subjectivity which influence much wider currents of social thought. Thus, law defines who can be responsible for actions and who can engage in transactions that will be legally recognized. In particular societies at particular times, children, married women, mentally handicapped people, prisoners, and aliens of the state (foreigners) have been treated as wholly or partly lacking legal capacity (for example, to own property directly). Thus, for certain purposes, law constitutes human capacities or denies them (see *Legal Personality*). Correspondingly, law routinely gives personality and capacity to nonhuman entities. In various societies these have included religious idols and animals, but in modern law by far the most important example of the construction of nonhuman personality is the attribution of legal personality to the business corporation. This attribution constitutes the corporation as an entity entirely separate from the human individuals who work within it. In recent critical legal scholarship, the concern with law's constitution of personality has broadened to focus on the wider range of effects of law on the construction of identity.

Feminist theory has emphasized that law has typically constituted the human person in terms of male characteristics. In the past this legal stance was often explicit, with legal doctrine's denial of women's legal personality in certain contexts (for example, as property owners). Modern feminist theory (see *Feminist Legal Theory*) has shown how legal thought and practice ignores differences between women's and men's experience, or covertly assimilates the former to the latter, applying a 'male standard' to the definition of all social relations, capacities, and responsibilities. It has been claimed that law constructs women (in its typical rhetoric and evaluations) as 'terrorized, sexualized and maternalized' (Frug 1992). In a host of studies, critical socio-legal scholarship has also focused on the legal construction or denial of the social identities of minority groups and, more generally, on law's power to define and exclude 'the other' in sharp opposition to an image of the 'normal' subject of law (see *Critical Race Theory*; *Sexual Orientation and the Law*; *Law and People with Disabilities*).

The effect of this cumulation of critical work is a widespread rejection among socio-legal scholars of the doubts about law's social importance that long haunted legal sociology. In their place is a confidence that legal ideas and practices are central matters for sociological study. The older legal sociology sometimes slipped into the self-defeating position of holding that law 'mirrored' society (thus implying not only law's lack of social scientific importance but also its 'non-social' identity as 'merely' technical regulation and so its substantial immunity from conceptualization as a social phenomenon in itself). The newer emphasis on law as constitutive asserts law's power to shape fundamental collective understandings and practices that define the nature of social life. Taken to extremes, this position can suggest that law constitutes the very idea of society, rather than society constituting law (Fitzpatrick 1995). A further step might be to discard all reified conceptions of 'society' or even of 'the social' as a category (on the basis that the social picture is now too fluid and complex to allow such encapsulations or to allow these concepts to retain analytical power). The aim then would be to focus specifically on the particular kinds of social conceptions (for example, of identity, community, culture) which legal doctrine itself proposes, assumes, or sustains. In a sense, such an approach entirely reverses the conception of law as mirror of society; the social (whether or not it remains possible to conceptualize this as any kind of totality) is seen as the construction of law.

3. *Limitations of Constitutive Approaches to Law*

The danger of these approaches is partly that of greatly exaggerating law's constitutive power (how can we know that *law* does the constituting and, even if it does, under what conditions and constraints?). Jurists have long been tempted in this way, some even speculating that law largely determines morality by contributing to socialization (Olivecrona 1939). Philosophers and jurists note law's 'word-magic,' the power of legal words as 'performatives' to bring about social effects, such as inaugurating marriages, creating personal statuses, or transferring property (Austin 1982). But these formal effects raise rather than answer questions about law's social character. The risk is of attempting to analyze law in isolation from other social forces, treating it (as the earlier 'mirror' conceptions tended to do) as separate from other aspects of social life, with a causal relationship existing between the two. This danger has been especially serious for CLS which, being essentially a law school movement rather than a movement in social science,

has tended to assume law's social significance and power (in conformity with the main trend of lawyers' thinking throughout modern times) rather than seeking to assess it systematically and empirically in a wider social context.

Equally, it is not always recognized sufficiently that law's constitutive power may vary greatly in different societies. In some societies, its power to shape ways in which politics is conducted, social movements are formed and their campaigns or collective strategies are pursued, may be clear if complex in its effects (Brigham 1996). In other societies, complex cultural factors make law much less capable of constituting the form and character of social relations. Indeed, they might make it largely irrelevant.

It is also significant that one influential and sophisticated theoretical approach in contemporary legal sociology entirely rejects the idea of law as constitutive of society, because it denies that law as a system can communicate with, and so influence, social life in any direct way (Teubner 1993). Autopoiesis theory (see *Law, Autopoiesis in*) sees modern law as a discourse that normatively constitutes (more specifically produces, reproduces, and observes) only itself. One consequence is that its conceptions of agency, responsibility, personality, subjectivity, etc. are produced solely for its own purposes and processes and have no necessary relation to anything existing outside these. Far from constituting economic, scientific, political, or cultural life, law produces decisions to which these other social systems may respond on their own terms. Autopoiesis theory denies all conceptions of law as a mirror of society. At the same time, its assertion of law's radical discursive autonomy deliberately provokes endless questions as to how law can engineer social effects that are other than intolerably unpredictable.

4. *The Plurality of Legal Experience*

Some recent sociolegal approaches have discarded both the idea of law as mirror of society and that of law as a more or less independently constitutive force in social life, and have replaced these with a view of law as social mediation (e.g., Mertz 1994). Law thus appears as one way (or perhaps a cluster of ways, if it is not seen as a unitary discourse or practice) in which social experience is mediated, conceptualized, defined, and evaluated.

Such an approach presents its own difficulties as a sociological conception of law. If law mediates the social in some way, what enables it to do so, and from where does its power or authority to mediate derive? What is the nature of law if it is, in some way, separate from the social life that it mediates? These questions are ultimately no different from those which haunt the idea of law as constitutive. (Where does law derive its power or authority to constitute social life? And how, in what way, is law separated from social life in order to constitute it?)

In answering these questions the temptation is often to dissolve one term into the other. Thus, law may be seen as, for all practical purposes, constituting the social, so that the latter ceases to exist (at least for purposes of sociolegal analysis) as something independent of legal categorizations and evaluations. According to this view, law may certainly be studied in a social context, but its significance as an object of study requires no validation by social science. Alternatively, law can be seen merely as a mirror of social forces and social structure, meriting no special, independent sociological analysis; juristic concerns are socially insignificant unless reducible to general sociological categories and concepts.

A productive approach may be available to avoid both of these positions. Law is sometimes conceptualized sociologically as an aspect or field of social experience (e.g., Cotterrell 1995). In this sense, it is a component of social consciousness, a normative patterning of experience or a framework of understandings or practices. As an aspect of social life, it is to be separated from other social phenomena only provisionally and contingently for purposes of study. Important antecedents of this approach exist in legal sociology (Gurvitch 1947), but they have been obscured, to some extent, by approaches that treat law entirely as an instrument of the state. Viewed in this state-focused way, law's constitutive power appears problematic or mysterious, given (a) the state's frequently noted remoteness from the local milieux it regulates and (b) the increasing production of law by agencies other than the state. Law may, however, be understood pluralistically (see *Legal Pluralism*) as existing not only in the practices and procedures of state agencies, but in the everyday circumstances, sites, and settings of a wide variety of social groups and networks. Legal experience, according to this view, is the subjective experience of individuals and groups confronting, in these diverse environments, 'local' issues of governmental regulation and social control. An approach along these lines will see law as constituting social life insofar as it is itself embedded in social life, in expectations about security, responsibility, and other matters. At the same time, state law (the legal practice of lawyers and courts serving the centralized state) needs not merely to be politically constituted, according to criteria of political legitimacy, but also (from a pluralist perspective) morally constituted by its relation to the social experience of the populations it purports to regulate. From this perspective, the issue of law's moral authority identified by Ehrlich a century ago still remains crucial for modern law.

See also: Critical Legal Studies; Feminist Legal Theory; Law and Everyday Life; Law and Society: Sociolegal Studies; Law as an Instrument of Social Change; Law, Autopoiesis in; Law: History of its Relation to the Social Sciences; Law, Sociology of;

Legal Culture and Legal Consciousness; Marxism and Law

Bibliography

Austin J L 1982 *How To Do Things With Words*, 2nd edn. Oxford University Press, Oxford

Brigham J 1996 *The Constitution of Interests: Beyond the Politics of Rights*. New York University Press, New York

Cotterrell R 1995 *Law's Community: Legal Theory in Sociological Perspective*. Clarendon, Oxford

Cotterrell R 1999 *Émile Durkheim: Law in a Moral Domain*. Edinburgh University Press, Edinburgh, Stanford University press, Stanford, CA

Ehrlich E 1975 *Fundamental Principles of the Sociology of Law*. Arno, New York

Fitzpatrick P 1995 Being social in socio-legal studies. *Journal of Law and Society* **22**: 105–12

Frug M J 1992 *Postmodern Legal Feminism*. Routledge, New York

Gurvitch G 1947 *Sociology of Law*. Routledge and Kegan Paul, London

Kronman A 1983 *Max Weber*. Edward Arnold, London

Mertz E 1994 A new social constructionism for sociolegal studies. *Law and Society Review* **28**: 1243–65

Olivecrona K 1939 *Law as Fact*, 1st edn. Einar Munksgaard, Copenhagen, Humphrey Milford, London

Pashukanis E B 1978 *Law and Marxism: A General Theory*. Ink Links, London

Pound R 1917 The limits of effective legal action. *International Journal of Ethics* **27**: 150–67

Savigny F C von 1975 *Of the Vocation of Our Age for Legislation and Jurisprudence*. Arno, New York

Sumner W G 1960 *Folkways: A Study of the Sociological Importance of Usages, Customs, Mores and Morals*. New American Library, New York

Teubner G 1993 *Law as an Autopoietic System*. Blackwell, Oxford

R. Cotterrell

Law, Autopoiesis in

Legal autopoiesis refers to the idea that the legal system presupposes and reproduces itself. Modern sociological systems theory employs the concept of autopoiesis to emphasize the self-constituting nature of social systems. In analyses of the legal system it constitutes a hybrid form of a legal theory and a sociological account of law.

1. *Autopoiesis and the Functionally Differentiated Society*

Autopoiesis describes the capacity of an entity to reproduce itself. As a concept it was first introduced in theoretical biology to explain cognition and the essence of life (see Maturana and Varela 1980, 1987 and was then further developed in general systems theory (for example, von Förster 1984). It has been widely applied in mathematics, in the study of cognition, and in studies of the nervous system as well as in information systems, cognitive science, and artificial intelligence (see Mingers 1995).

In general systems theory, the following definition is widely accepted: autopoietic systems 'are defined as unities and as networks of production of components that recursively, through their interactions, generate and realize the network that produces them and constitute, in the space in which they exist, the boundaries of the network as components that participate in the realization of the network' (Maturana 1981, p. 21). Maturana's definition operates with a notion of recursive closure of the system. An autopoietic system is not only able to produce and eventually to change its structures, but is also capable of reproducing its components, and in particular its undecomposable elements, through recursive self-reference of these elements. Thus, its self-reproduction relies exclusively on internal operations and is not dependent on the environment.

Niklas Luhmann (1927–98) (see *Luhmann, Niklas (1927–98)*) is responsible for introducing the concept of autopoiesis to social systems theory (Luhmann 1986. His approach transcends the Parsonian version of sociological systems theory, which operates with an input–output model of the social system and its environment. A crucial element of Luhmann's concept is the idea of communications as the ultimate, non-decomposable elements of an autopoietic social system. Social systems are capable of reproducing themselves because they are communication systems (Luhmann 1995, 1997).

Autopoietic social theory distinguishes between society and functionally differentiated social systems. Society is seen as a first-order social system, whereas the legal system is viewed as a second-order social system or societal subsystem. Society as a first-order social system differs from second-order social systems insofar as it has no other social system as an environment; society's environment consists of natural and psychic systems. In fact, human beings do not form part of society or any social system but belong to their environment; the individuality and the consciousness of human beings are features of highly complex psychic systems that operate separately from social systems (Luhmann 1995, Chap. 7).

2. *Autopoiesis of the Legal System*

In accordance with general systems theory, autopoietic legal theory assumes that law is a special communication system that reproduces itself in self-referential processes. Legal communications, such as legal norms, legal decisions, or legal doctrine, are central to the

autopoiesis of the legal system (Teubner 1993, Luhmann 1993). In order to determine whether a communication is part of the system, the legal system operates with a specific binary code: law/nonlaw or legal/illegal (Luhmann 1990).

However, the code itself is not sufficient for the creation of law. The code must be translated into legal programs, which are rule-bound. Furthermore, the function of law is to deal with expectations, and in particular with disappointments of expectations. The legal system operates with a specific combination of normative closure and cognitive openness (Luhman 1987). It upholds the existing norms and procedures but it is also capable of learning in response to disappointments, and provides mechanisms for its own change. An important feature of autopoietic legal theory is the radicalization of the notion of legal autonomy. It is the peculiar character of autopoietic law that 'only the law can change the law. This is not to deny that the environment (in particular the political system) has an impact on the legal system. But the legal system reproduces itself through legal events and only by legal events' (Luhmann 1990, p. 229). On this basis, the modern legal system becomes positive and adopts a reflexive identity.

A further consequence of radical legal autonomy is the emergence of a special legal understanding of the external world. Like other function systems, the autopoietic legal system develops an epistemology according to its own criteria (Teubner 1989). It is this specific worldview that guides its internal operations and thereby influences legal development.

3. *Legal Evolution and Structural Coupling*

The legal system forms part of the functionally differentiated modern society (other functional systems include the economic, the political, and the religious systems). Each system within society evolves separately along its own trajectory, with its own internal structures and processes determining its development.

However, the evolution of social systems is not determined exclusively by internal conditions. Specific external relations that derive from structural couplings also have an effect. These structural couplings are constant sources of mutual irritations and perturbations of the coupled systems. Structural coupling can neither overcome the identity and autonomy of the coupled systems nor rank functional subsystems in an asymmetrical hierarchical fashion. Structural coupling allows only selective exchanges between the systems. Structural coupling produces irritations inside the system which are implemented by the system through its network of operations into further operations.

These perturbations do not actually cause changes; rather, they serve as triggers that may initiate internally controlled operations in operationally closed systems. Thus, structural coupling of social systems does not contribute directly to the reproduction of the systems. It is simply the specific form in which the system presupposes specific states or changes in its environment and relies on them.

The constitution provides an example of the role of structural coupling between the legal and the political systems. When constitutions emerged in the eighteenth century, they helped to recombine the functionally differentiated political and legal systems. From an autopoietic perspective, constitutions were adopted in response to problems of self-reference, paradoxes, and asymmetries in the legal and the political systems of the respective countries. Factors such as independence from colonial domination in the USA, or the formation of political sovereignty in a revolutionary situation in France, reflected only specific historical constellations. In both cases, the functionally differentiated legal and political systems had to realize that law completely determines what is legal or illegal, and that the political system had to design its own self-binding sovereign power.

The concept of the constitution offered the solution for both systems because it separated the political and the legal systems, yet provided for their structural coupling. The legal system made reference to the political system by acknowledging that the constitution derived from the political will of the people, and the political system made reference to the legal system by endowing the latter with the power to determine the constitutionality of positive law.

4. *Reflexive Law and Legal Regulation*

The concept of reflexive law, which arose from a debate of a new form of reflexive legal rationality in modern law, accompanies autopoietic legal theory. In particular, Gunther Teubner has argued that the modern legal system is characterized not only by formal and substantive rationality but also, increasingly, by reflexive legal rationality (Teubner 1983). The reflexive law concept was originally intended to combine the social theories of Jürgen Habermas and Niklas Luhmann. However, as the construct of reflexive law became more integrated with autopoietic legal theory (Teubner 1993), the influence of Habermas' theory decreased and Luhmann's systems theory became dominant.

The reflexive law concept emphasizes the limits of legal regulation and processes of self-regulation in other social systems. It analyzes the regulatory capacities of the legal system, both in relation to the regulation of other systems and in relation to the legal systems capacity to regulate itself. Law is confronted with a regulatory dilemma. It encounters ineffective-

ness when the regulated system simply does not respond to the regulatory attempt. Furthermore, if law imposes its own solution, it risks legalizing politics; but if law yields to the political imperatives, it risks becoming politicized. Teubner (1985) suggests that by facilitating self-regulation by the regulated system, law can sometimes overcome the dilemma.

There have been a number of attempts to use the reflexive law concept to analyze developments in particular areas of law. For example, Collins (1999) sees modern contract law as a form of reflexive regulation. Reflexive environmental law (Orts 1995) and reflexive labor law (Rogowski and Wilthagen 1994) have been suggested as concepts that can provide new insights in legal developments of legal subdisciplines. There have been attempts to combine reflexive law and postmodern legal theory in constitutional law (Ladeur 1992, 1997).

5. *Law in the World Society*

An 'unavoidable consequence of functional differentiation' (Luhmann 1982[1990, p. 178]) is that society becomes global. By concentrating on one function and excluding other functions, each subsystem gains a capacity to include all function specific communicative behavior, unlimited by territorial boundaries. Traditional subsystems that are constituted by specific communications can spread over the globe. Society as a whole can no longer integrate its subsystems by common territorial frontiers. Indeed, a modern, functionally differentiated society is a world society. It is this dimension with which the reflexive and autopoietic law concept fundamentally challenges traditional legal analyses.

Autopoietic legal theory predicts that legal globalization is unavoidable, and that a world legal system will emerge. This global law develops without support by any particular state (Teubner 1997). Despite considerable confusion about the meaning of globalization, it seems clear that there is not one trajectory that law follows within the world society. It is thus a further feature of reflexive law that it observes and compares legal developments in different parts of the world. For autopoietic legal theory, the study of global law requires a reflexive comparative law that is both empirically informed and theoretically aware of the autopoietic nature of world law (see *Globalization: Legal Aspects*).

See also: Chaos Theory; Evolution, Natural and Social: Philosophical Aspects; Evolution: Self-organization Theory; Globalization: Legal Aspects; Law: Change and Evolution; Legal Systems, Classification of; Luhmann, Niklas (1927–98); Social Evolution: Overview; Social Evolution, Sociology of

Bibliography

Collins H 1999 *Regulating Contracts*. Oxford University Press, Oxford, UK

Förster H von 1984 *Observing Systems*. Intersystems Publications, Salinas, CA

Ladeur K-H 1992 *Postmoderne Rechtstheorie*. Duncker and Humblot, Berlin, Germany

Ladeur K-H 1997 Post-modern constitutional theory: A prospect for the self-organising society. *Modern Law Review* **60**: 617–29

Luhmann N 1982 The world society as a social system. *International Journal of General Systems* **8**: 131–8 [reprinted in Luhmann N 1990 *Essays on Self-reference*. Columbia University Press, New York, pp. 175–90]

Luhmann N 1986 The autopoiesis of social systems. In: Geyer F, Zouwen J van der (eds.) *Sociocybernetic Paradoxes*. Sage, London [reprinted in Luhmann N 1990 *Essays on Self-reference*. Columbia University Press, New York, pp. 1–20]

Luhmann N 1987 Closure and openness: On reality in the world of law. In: Teubner G (ed.) *Autopoietic Law. A New Approach to Law and Society*. De Gruyter, Berlin, pp. 335–48

Luhmann N 1990 The self-reproduction of law and its limits. In: Luhmann N (ed.) *Essays on Self-reference*. Columbia University Press, New York, pp. 227–45

Luhmann N 1993 *Das Recht der Gesellschaft*. Suhrkamp, Frankfurt am Main, Germany

Luhmann N 1995 *Social Systems* [original German title *Soziale Systeme* 1984]. Stanford University Press, Stanford, CA

Luhmann N 1997 *Die Gesellschaft der Gesellschaft*. Suhrkamp, Frankfurt am Main, Germany

Maturana H 1981 Autopoiesis. In: Zeleny M (ed.) *Autopoiesis: A Theory of Living Organization*. North-Holland, New York, pp. 21–33

Maturana H R, Varela F J 1980 *Autopoiesis and Cognition: The Realization of the Living*. Reidel, Dordrecht, The Netherlands

Maturana H R, Varela F J 1987 *The Tree of Knowledge: The Biological Roots of Human Understanding* [original Spanish title: *El árbol del conocimiento* 1984]. New Science Library, Boston

Mingers J 1995 *Self-producing Systems: Implications and Applications of Autopoiesis*. Plenum, New York

Orts E W 1995 Reflexive environmental law. *Northwestern University Law Review* **89**: 1227–340

Rogowski R, Wilthagen T (eds.) 1994 *Reflexive Labour Law. Studies in Industrial Relations and Employment Regulation*. Kluwer, Deventer, The Netherlands

Teubner G 1983 Substantive and reflexive elements in modern law. *Law and Society Review* **17**: 239–85

Teubner G 1985 After legal instrumentalism? Strategic models of post-regulatory law. In: Teubner G (ed.) *Dilemmas of Law in the Welfare State*. De Gruyter, Berlin Germany, pp. 217–41

Teubner G (ed.) 1988 *Autopoietic Law. A New Approach to Law and Society*. De Gruyter, Berlin, Germany

Teubner G 1989 How the law thinks: Toward a constructivist epistemology of law. *Law and Society Review* **23**: 727–57

Teubner G 1993 *Law as an Autopoietic System* [original German title *Recht als autopoietisches System* 1989]. Blackwell, Oxford, UK

Teubner G (ed.) 1997 *Global Law Without a State*. Dartmouth, Aldershot, UK

R. Rogowski

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Law: Change and Evolution

1. Legal Evolution

There has been a good deal of speculation on the question whether legal systems 'evolve,' that is, is there some kind of natural, automatic process or direction in which legal systems move and change over time. A distinction may be made between macroevolution—a theory of long-term change, involving whole systems or groups of systems of law; or law in general—and microevolution, which would concern single systems of law, and short-term changes. Of course, there might be many gradations in between. An interesting theory, more micro than macro, but not entirely, is the theory that common law doctrines evolve over time toward economic efficiency; the theory assumes a process, somewhat like Darwinian natural selection, that winnows out wasteful or irrational rules (for a discussion, see Elliott 1985, pp. 62–71). But microevolution merges with the theory of legal change in general. This essay will focus first on macroevolutionary change; then, briefly, on more general theories of legal change.

A (macro) evolutionary theory of law would mean, first, that it is possible to speak about stages in the growth or development of legal systems. It would imply a patterned, orderly progression: legal systems begin in stage one, then move to stage two, then three, and so on. Few if any legal scholars believe in 'evolution' in the sense of an inevitable progression from stage to stage. Practically speaking, historians and legal scholars have thought of evolution in a weaker sense: not all systems will reach the highest stage (whatever that might be), but the order will always be one, then two, then three (Friedman 1975).

The whole concept of evolution is, in a sense, distinctly modern, for the simple reason that many or most traditional legal systems regarded themselves as essentially immutable. Law was divine or magical or the product of natural reason, hence basically timeless and beyond human power. Change was, if anything, dramatic intervention by the gods. On the other hand, insofar as evolution implies progress, it is as unfashionable an idea in contemporary times as is the idea of progress in general.

Any theory of legal evolution has to posit some source of evolutionary change, that is a force that impels legal systems to move along given paths. Essentially, there are three general answers to the question of the source. One is that the evolution of society drags legal systems along with it, so that stages of legal development correspond to stages of social development. Most of the theorists discussed below fall into this category. A second is that law is independent of society, but that it contains within itself some principle of growth or evolution, which unfolds gradually over time.

Such an idea is at least implicit in legal formalism, both of the common law and civil law varieties; development either occurs automatically, or depends on the craftsmanship of judges or jurists; in either event, the law unfolds like the flower from the bud. Among modern authors, Alan Watson perhaps most strikingly illustrates this view: law, for Watson, is 'largely autonomous and not shaped by societal needs'; it 'evolves from the legal tradition' and not from the forces and demands of the surrounding society (Watson 1985, p. 119.) Still a third view would be that changes in law are basically random or accidental, following no special path, and tending in no special direction. This view does not seem to have any serious support in modern legal thought.

Many classical works in legal sociology and in social thought generally have at least touched on the subject of legal evolution, and most of them fall squarely within the first type mentioned above (Raiser 1995, pp. 318–334, Elliott 1985). Marx expressed an evolutionary theory, after all, in his views of the relationship between ownership of the means of production, and the organization of society; and this theory had definite implications for the legal order. Sir Henry Maine, in his important book, *Ancient Law* (1861), expressed the view that there was a definite pattern in the way legal systems grew and changed. Ancient law began in a patriarchal stage, in which the family, not the individual, was the unit of legal analysis; moreover, the father had basically the key rights and authority within the family. Birth order, gender, and age determined one's legal rights and duties. Society was based on status. The movement of 'progressive societies,' he stated, has gone in a 'uniform' direction; the individual replaces the family; voluntary, arrangements replace ascriptive ones; legal systems, to use his famous phrase, evolved 'from Status to Contract.'

Implicit evolutionary theories appeared also in other classical theories of law and society, Durkheim, for example, drew a distinction between primitive and modern societies. In *The Division of Labor in Society*, Durkheim distinguished between 'mechanical' and 'organic' solidarity. In traditional societies (societies of 'mechanical' solidarity), rules and norms were held in common. Violations threatened social cohesion; and had to be punished. The law was therefore mostly concerned with punishments. On the other hand, in modern society (characterized by 'organic' solidarity), there is a tremendous amount of complexity and interdependence; and legal rules are mostly civil and 'restitutive.' In Durkheim's scheme, as in Maine's, there is strong emphasis on contract, that is, societies as they enter the modern world are characterized more and more by voluntary agreements.

The most subtle, and influential, of the classic writings on legal evolution is Max Weber's (Rheinstein, 1954). Weber was interested in the question: what makes modern law different from older and traditional systems? The distinctive quality he saw was

what he called rationality—essentially, the process of making and interpreting law on the basis of general principles and rules. Oracles, ordeals, law-making by charismatic leaders—these are 'irrational' forms of making or finding law. Modern legal systems, in developed countries, tend (according to Weber) to replace the irrational with the rational. In particular, modern legal systems are characterized by formal rationality, a 'peculiarly professional, legalistic, and abstract approach to law.' A formally rational system uses logic and deductive reasoning to derive results from rules, concepts, and principles. Irrational systems do not produce predictable results—results that can be controlled by the intellect, and therefore results that a citizen can rely on. Modern law, the law of the modern capitalist state, has a natural preference, therefore, for Weberian rationality.

Weber falls short of arguing that this progression, from the irrational to the rational, is inevitable or predetermined. His discussion of legal history cannot be summed up in this single formula, since Weber describes, in rich detail, at least four stages of development (Sterling and Moore 1987). Charismatic legal revelation is succeeded by formally irrational 'kadi' justice—*ad hoc* decision making based on general social norms; then comes a stage of substantive rationality; in this stage a systematic body of rules is imposed on society, but less in terms of juristic reasoning than in terms of religious, ethical, or economic considerations. The stage of formal rationality comes last; in this stage the jurists create a more or less self-contained, highly logical and systematic body of law, in which abstract legal concepts are arranged in some definite and principled order. But this is merely a more or less idealized picture of the growth of law in the West. Weber does not claim that this is bound to happen elsewhere; and he never quite argues that the movement he describes is irreversible. In particular, modern law has tended to drift away from formal rationality in the direction of what Weber called substantive rationality, making or finding law in accordance with principles and policies (economic, social, cultural) that are not strictly 'legal.'

Richard Schwartz and James C. Miller (1964, see Wimberley 1973) tried to put the concept of evolution on a systematic basis, They drew material from 51 different societies, and tried to 'scale' the development of legal institutions. Their scales were in terms of legal institutions: the least advanced societies had no police, lawyers, or forms of mediation. Societies develop mediation first, then some form of policing, and a legal profession last of all. Most of the societies that Schwartz and Miller discussed were preliterate.

In a sense, Weber and Maine (although they wrote long before Schwartz and Miller) pick up where Schwartz and Miller leave off, since their main concern was with advanced societies. Galanter (1966), in an important essay, tried to define what makes a legal system 'modern,' which was, of course, Weber's chief interest as well. He suggests a list of 11 traits, which, in his view, set modern law apart from pre-modern law. The scheme, then, is implicitly evolutionary. For Galanter, modern law is bureaucratic, professional, rational, and universalistic; it is also 'amendable,' that is, it lacks 'sacred fixity.' Friedman (1994) returns to this theme and lists six traits which, in his view, together go to make up 'modern legal culture.' Modern law is committed to change, reform, adaptation; it lacks the conception of timelessness of traditional legal systems. It is pervasive and ubiquitous, and is frankly instrumental. At the same time, it does contain strong notions of fundamental rights. Rehbinder (2000) also lists a group of 'tendencies' which characterize contemporary law: a strain toward universalism (this was also one of Galanter's traits), but at the same time specialization, bureaucratization, and an increase in the sheer amount of legal material.

Most of these discussions of legal change focus, then, on 'modernity,' and they are descriptive and explanatory. There is a strong normative strain, however, in the work of Selznick (1969). Selznick was concerned, above all, with the evolution of 'legality.' In a later work, Philippe Nonet and Selznick discuss a tendency for laws to evolve toward 'responsiveness' (Nonet and Selznick 1978). These ideas have influenced the work of Gunther Teubner (1983), who described a form of evolution from a kind of dry, rule-bound bureaucratic system, to a more 'reflexive' one, in which the law, particularly within administrative systems, would be responsive to the needs and wants of the governed, and which would develop rules and procedures that would be just and adaptive. Teubner, like Nonet and Selznick, adds a distinctively normative flavor to the more descriptive tone of Weber. Teubner's views touched off a lively debate (see Blankenburg 1984); some of the empirical issues raised in this line of work remain to be definitively examined.

2. *Legal Change*

The subject of the theory of legal change has to be dealt with rather briefly, not because it is less important than the theory of legal evolution, but precisely because it is so totally pervasive. One of the most obvious aspects of modern legal systems is the fact that they are constantly changing. Every session of every legislature in every country makes at least some amendments or modifications of the law. The public expects change: every political party in modern Western countries has to have a program of legislation. Legal systems are very much dynamic systems, and it is part of the legal culture to expect them to be such.

Thus all modern legal systems accept change as normal. Not all legal systems, historically, have done so. In many classical systems, and in non-literate systems, the body of laws, or at least crucial aspects of them, were felt to be immutable—or, in any event,

beyond, the reach of ordinary humans. Since circumstances in fact changed, and the needs and wants of society changed, these legal systems were in fact changeable, and did in fact change with changing times. But legal change typically came in disguised or muted form; such systems often resort to devices such as legal fictions to conceal what had taken place.

Theories of legal change—theories about what produces legal change—depend, in the first instance, on the answer to larger questions, very notably, the question of the 'autonomy' of the legal system. A fully autonomous legal system would be one which operates according to its own inner logic or program, independent of the forces of the outside society; and changes within that system would accordingly proceed according to that inner logic or inner program. A totally non-autonomous system would, on the contrary, faithfully reflect changes in the outside society, and mirror them almost mechanically. Needless to say, most jurists and legal scholars are not found at either pole of this continuum, but somewhere in between. Most orthodox legal scholars cluster closer to the autonomous end of the continuum; social scientists who study law cluster closer to the other end.

In discussing legal change, there are obviously many factors to be taken into account (for a general discussion, see Ferrari 1997, Chap. 6). Not every legal system will change in the same way. This is true both for internal and external reasons. Societies undergoing rapid change will naturally process more legal changes than relatively static societies. Legal systems, particularly advanced ones, which legitimate change, will contain rules and procedures and institutions for making and processing change. Modern parliamentary democracies put the burden of change in formal law on legislatures, for example, though administrative rule-making is an extremely important vehicle of change as well; and the role of courts as change-agents is important in most of these societies. Much change also originates in customs, practices, and norms, which then influence the change-making institutions.

External forces lead to legal change, but the exact causal mechanism varies from instance to instance. A major event—war, plague, revolution—or a major technological or social development (the Industrial Revolution, or globalization, or the so-called sexual revolution) will inevitably lead to legal change, sometimes on a massive scale; but the exact form of the change is never precisely predictable. The 'outside' event or situation will impact a society which has a particular pattern or structure, and that existing pattern or structure will bend and mediate the influences of the 'outside' forces prismatically. In any event, these forces will create their impact through making changes in the general legal culture—that is, in the pattern of demands, expectations, and values that bear on the legal system—and legal change will result from this change in legal culture, rather than flowing immediately from the 'outside' event or situation. Moreover, once a legal structure, doctrine, or institution is in place, it can exert its own force on the legal culture; it affects legal ideology, and it may come to seem natural or inevitable. The process of legal change, then, is dynamic and interactive and forms a never-ending chain of causes and effects.

In theory, one can distinguish four types of legal change, depending on the point of origin of the change, and its final point of impact (Friedman 1975): (a) some change originates inside the legal system, and also spends whatever impact it might have within that system; (b) some change begins inside the system, but in fact does have an impact on society; (c) some change originates outside the legal system, but affects only the legal system, and ends there, with no further impact on society; and (d) some change originates outside the legal system, but ultimately has some impact on society, that is, it is more or less effective.

The first two are examples of technical change, and include many projects of 'law reform' or codification, within legal systems—technocratic work of jurists, which has little or no effect on the actual operation of the legal system. Many examples can be mentioned. Codification often falls into this category. But sometimes, wittingly or unwittingly, what begins as purely technical change has unforeseen consequences in the larger society. When this occurs, it is likely that the technical change simply furnishes a convenient 'hook' or formal excuse for some group in society which uses the technical change as a way to advance its goals. The last two are instances of change that begin with social forces or social movements. Legal change that results from demands arising out of social forces or social movements should, in theory, produce an impact in society. Where this does not happen, it is usually because some other social force or social movement is able to resist change, and the result is a kind of stalemate: a 'law' is passed, for example, after a bitter struggle between interest groups, but the 'losing' group has enough power to prevent effective enforcement.

The fourth type is, in many ways, the most interesting and pervasive, and it ranges all the way from revolutionary changes in the law—such as the changes that took place after the Russian or the Chinese revolutions; or, contrariwise, that took place in Russia after the collapse of the Soviet Union—to relatively small, incremental changes, which produce only a minor effect on some portion of the population.

The discussion above has mostly assumed change in formal, overt aspects of law. But legal systems are complex, and there are many instances where one aspect of the system changes while others remain (on the surface) unchanged. For example, the formal or official law might remain unchanged, while the operating system changes, perhaps quite drastically. One example would be divorce law and practice in the United States, in the late nineteenth and early twentieth centuries. The formal law of some states (New

York, for example) preserved rules that made divorce difficult and expensive; but at the operating level, collusion and other devices made it possible for citizens of New York to get divorces that in theory they were not entitled to have. What prevented formal change, and led to this example of a dual system, was apparently a political situation which for various reasons was deadlocked. Enactment of reform was politically impossible. The growing demand for divorce had to seek other channels, which it did quite successfully. Legal systems very commonly change in this ragged and complex manner.

One should also distinguish between long run change and short run change (and various gradations in between). In the long run, legal systems are almost infinitely plastic, and change to fit the shape of the society they are embedded in. A feudal system will generate a legal system that deals with the problems and institutions of feudal society and reflects its norms and structures; a capitalist, or socialist, or tribal society will behave in a similar manner. But in the very short run, random, accidental, or 'internalist' factors may condition or shape forms of legal change. It is this fact in modern societies that gives the impression of legal autonomy—an illusion that the legal system has a tough and impermeable outer skin. Another factor is the strength and power of the legal profession itself, which may have a vested interest in resisting certain kinds of change.

The rate and forms of change also depend on whether the society is relatively open and democratic, or relatively closed and aristocratic or totalitarian. In an open society, 'public opinion' is an important factor in shaping the law. Public force would be a more accurate way of putting the matter. 'Public opinion' as such—whatever it is that polls and surveys measure—has no role in itself in bringing about legal change, except insofar as it influences legislators or judges, or is translated into public action 'Public opinion,' therefore, as a formative element in legal change, cannot be equated simply with the wishes of the majority, as measured by some survey instrument. Salience of opinion is of crucial importance, because it affects whether or not attitudes are translated into behavior. Moreover, in no society, no matter how democratic, can one assume equality of power and influence; quite the contrary. As an axiom of legal change, one can start with the notion that a legal rule, doctrine, practice or institution will change if the social force exerted against it is greater than the force generated to preserve it intact.

This is a very abstract, and perhaps not very helpful, formulation. The interesting questions are more specific: in any given situation, what *are* the social forces that are in play, and what lies behind the exertion of social force?

In any society, some people have more actual or potential power than others; and more of a propensity to use it. A wealthy corporation simply counts for more than (say) an impoverished and uneducated slum-dweller. Moreover, it can be assumed that at any given time, the actual social force exerted on an issue is much less than the potential social force, since most people and institutions, even the powerful ones, most of the time, are indifferent to or ignorant of the majority of the legal issues and problems within a given society. For this reason, scandals, incidents, and 'horror stories' are particularly important sources of legal change in open societies: they can arouse the sleeping giant and turn potential force into actual exerted force. Hence the role of the media—which communicates these scandals, incidents, and 'horror stories'—is a particularly powerful, but insufficiently studied, catalyst for socio-legal change.

Particularly in common law systems, there is much discussion of the role the courts have played in bringing about social change. To what extent, for example, did *Brown vs. Board of Education* (1954) make a difference to American race relations? This is a much discussed, and somewhat controversial question (Rosenberg 1991). That the courts are an important institution in American society and play some role in bringing about social change cannot be denied; the question is how much. There is a small literature on the impact of judicial decisions, but it seems impossible to generalize, since the amount of change, if any, which a judicial decision brings about seems to vary with the circumstances; no two cases seem alike. The issue is not confined to common law countries. Modern courts, in many countries, have the power of judicial review, and their potential for pushing society along certain, paths has certainly increased. There are some signs of dramatic change in judicial culture in the Western world in general. Crusading judges in Italy led to the collapse of the party system; and an activist Spanish judge touched off a legal and political crisis when the judge issued an order directed against General Pinochet, the former dictator of Chile. With regard to judicial decisions, and to legal innovations in general, impact depends on many factors. One obvious factor is whether and how the innovation has been communicated to its audience—a factor we have already mentioned, with regard to scandals and incidents. One should also distinguish between immediate changes (on the litigants, or on a particular group which is the subject of legislation), and more remote, indirect, and long-term types of change. Many legal commands have effects which ripple out into society. As is generally the case, the research on all such issues is thin, and it is difficult so far to generalize.

See also: Change: Methods of Studying; Cultural Evolution: Overview; Durkheim, Emile (1858–1917); Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Law and Society: Sociolegal Studies; Law, Sociology of; Legal Systems: Private; Social Change: Types; Social Evolution, Sociology of; Weber, Max (1864–1920)

Bibliography

Blankenburg E 1984 The poverty of evolutionism: A critique of Teubner's case for 'Reflexive Law.' *Law and Society Review* **18**: 273

Elliott E D 1985 The evolutionary tradition in jurisprudence. *Columbia Law Review* **85**: 38

Ferrari V 1997 *Lineamenti di Sociologia del Diritto: I. Azione Giuridica E Sistema Normativo*. Laterza, Rome

Friedman L M 1975 *The Legal System: A Social Science, Perspective*. Russell Sage Foundation, New York

Friedman L M 1994 Is there a modern legal culture? *Ratio Juris*. **7**: 117

Galanter M 1966 The modernization of law. In Weiner M (ed.) *Modernization* Basic Books, New York, p. 153

Maine H 1861 *Ancient Law*. J. Murray, London

Nonet P, Selznick P 1978 *Law and Society in Transition: Toward Responsive Law*, Octagon Books, New York

Raiser T 1995 *Das lebende Recht: Rechtssoziologie in Deutschland*. Nomos, Baden-Baden

Rehbinder M 2000 *Rechtssoziologie*, 4th edn. Walter de Gruyter, Berlin

Rheinstein M (ed.) 1954 *Max Weber on Law in Economy and Society*. Harvard University Press, Chicago

Rosenberg G N 1991 *The Hollow Hope: Can Courts Bring About Social Change?* University of Chicago Press, Chicago

Schwartz R, Miller J C 1964 Legal evolution and societal complexity. *American Journal of Sociology* **70**: 159

Selznick P 1969 *Law, Society, and Industrial Justice*

Sterling J C, Moore W E 1987 Weber's analysis of legal rationalization: A critique and constructive modification. *Sociological Forum* **2**: 67

Teubner G 1983 Substantive and reflexive elements in modern law. *Law and Society Review* **17**: 239

Watson A 1985 *The Evolution of Law*. Johns Hopkins University Press, Baltimore, MD

Wimberly H 1973 Legal evolution: One further step. *American Journal of Sociology* **79**: 78

L. M. Friedman

Law: Defense of Insanity

1. United States

The United States criminal law presumes that a person must generally act in accordance with the acceptable range of norms of conduct adopted by society. A person exercising free will should be held accountable for his actions, because individuals do in fact have at least a minimum capacity for making the voluntary and rational choices required for criminal responsibility. Thus, the ordinary criminal defendant is viewed as 'culpable,' 'blameworthy,' or 'responsible' because that person could have chosen to abide by the dictates of the law. But there are a few individuals who cannot be held accountable. A mental disability or disease deprives them of even the minimal capacity for rational and voluntary choices on which the law's expectation of responsibility is predicated. Because of their inability to comply with the law, such persons are not held culpable, and are not criminal sanctions invoked or applied consequent to their conduct.

The insanity defense, therefore, is the exception that 'proves' the rule of free will. Supporters of the defense view it as vital to a healthy society, which uses its criminal law to build and buttress self-reliant action on the part of its citizens. Eliminating the insanity defense would remove from criminal law and public conscience the vitally important distinction between 'illness' and 'evil.' Throughout history, it is this vision of law that has rallied supporters to resist efforts to abolish the insanity defense. There is a deeply entrenched human feeling, as old as recorded history, that persons who are grossly disturbed (whether called 'madmen,' 'lunatics,' 'insane,' or 'mentally disordered') should not be punished as ordinary criminals would be.

The insanity defense also conforms to our ideas about the purposes of punishment. For the few who cannot be held accountable because of mental disability or disease, social control may be best served by confinement in a secure hospital setting (especially in the case of those who are dangerously insane). Similarly, the general deterrence theory underlying punishment suggesting that a person's awareness and fear of unpleasant consequences will restrain him or her from engaging in criminal behavior will not apply to the insane. Deterrence is effective only with persons who can understand the signals directed at them by the criminal code, who can respond to warnings, and who can understand the significance of sanctions imposed on violators (see *Punishment: Social and Legal Aspects*; *Deterrence*). Thus, punishment is not likely to deter seriously disturbed individuals from future antisocial conduct.

Finally, the insanity defense is seen by supporters as quite consistent with the notion of specific deterrence or restraint. If the defense is invoked successfully, the defendant is not merely incarcerated for a fixed period of time, but instead can or should be committed until such time as he or she is no longer dangerous.

Those who favor abolishing or severely limiting the insanity defense have advanced the following arguments: (a) the key terms in the various insanity tests are so vague as to invite speculation and intuitive moral judgments in the guise of factual determinations; (b) there is little or no basis in psychiatry for allowing expert witnesses to testify—as they often do, in conclusory terms—concerning the differentiation between persons who are personally blameworthy and those who are not; (c) it is therapeutically more desirable to encourage treatment of persons as actors responsible for their conduct, rather than as involuntary victims 'playing a sick role,' (d) the insanity defense discriminates against persons who commit

crimes because of influences on their personalities other than mental disease or defect. In addition, opponents of the defense argue that, if it be therapeutically desirable to provide a medical-custodial disposition, then such a decision should be made directly—immediately following the defendant's conviction, rather than indirectly during the trial. Finally, opponents of the insanity defense claim that, in practice, it is a 'rich person's defense' because usually it is only the wealthy who can afford the array of experts needed to mount a convincing defense. These scarce psychiatric resources, they argue, should be spent in treatment of those who have been committed or imprisoned.

Of the 47 states in the United States that allow a defense of insanity (Idaho, Montana, and Utah have abolished it), about half apply the American Law Institute (ALI) standard, first adopted in 1962; the others continue to apply the M'Naghten standard originally adopted in England in 1843 and then in federal and most state courts in the US in 1851. The ALI test states: 'A defendant is not responsible for criminal conduct if at the time of such conduct as a result of mental disease or defect he lacks substantial capacity either to appreciate the criminality of his conduct or to conform his conduct to the requirement of the laws.' The M'Naghten rule states: a person cannot be convicted if, at the time the criminal act was committed, that person was laboring under such a defect of reason (from a disease of the mind) as not to know the nature and quality of the act he or she was doing—or, if that person did know it, as not to know that the act was wrong.

In 1984, largely in response to the public outcry at the jury's acquittal of John Hinckley, Jr., the attempted assassin of then President Ronald Reagan, the US Congress enacted the Insanity Defense Reform Act. It represents the first federal codification of the insanity defense. The new standard provides for the affirmative defense that 'the defendant, as a result of a severe mental disease or defect, was unable to appreciate the nature and quality or the wrongfulness of his acts.' This formulation does not allow an insanity defense based on a defendant's inability to conform his conduct to the requirements of the law, as provided by the American Law Institute standard.

The act also shifts the burden of proof from the prosecution to the defendant, who must prove insanity by clear and convincing evidence. It changes the verdict form to 'not guilty only by reason of insanity' and establishes a federal procedure for commitment of persons who have been found not guilty only by reason of insanity. The act also limits the role of experts: No expert witness testifying with respect to the mental state or condition of a defendant in a criminal case may state an opinion or inference as to whether the defendant did or did not have the mental state or condition constituting an element of the crime charged or of a defense thereto. Such ultimate issues are matters for the trier of fact alone (Simon and Aaronson 1988).

Although the insanity defense is introduced in fewer than one percent of all criminal trials, it has provoked on intense amount of interest and attention. Allen said of it: 'The issue of criminal responsibility has attracted more attention and stimulated more controversy than any other question in the substantive criminal law' (Allen 1964).

The defense of insanity attracts so much attention because it touches on ultimate social values. It was founded on the concept of individual responsibility as a prerequisite for criminal punishment. In the words of Stone: 'The defense of insanity purports to draw a line between those who are morally responsible and those who are not, those who are blameworthy and those who are not, those who have free will and those who do not, those who should be punished and those who should not, and those who can be deterred and those who cannot' (Stone 1975).

2. *European Countries*

The following paragraphs provide a brief review of the legal criteria and procedures surrounding the defense of insanity in Western European and selective other societies (Simon and Aaronson 1988).

2.1 *Great Britain*

Great Britain still uses M'Naghten as the criterion of nonresponsibility but the defense is invoked very rarely. More popular is the partial defense of 'diminished responsibility' for persons charged with first-degree murder. Mental illness (which does not involve use of the M'Naghten right–wrong criterion) is introduced as a partial defense—to reduce a charge of first degree homicide to manslaughter—manslaughter does not result in mandatory life in prison anywhere. This is true for murder, not manslaughter.

A convicted criminal defendant who is considered mentally ill, although not legally insane, can be sent either to a prison or to a hospital. If a person is sent to a mental health hospital, it is usually for an indeterminate period. The court can and usually does impose a 'restriction order' on such a person if it considers him or her dangerous. If a person is not subject to a restriction order, he or she can be released by the hospital staff, subject to a determination that he or she can safely return to the community. But if a person is subject to a restriction order the Home Secretary must concur in the decision to release. Thus, the release of a mentally ill offender from a psychiatric facility or from a prison in England is usually a combined medical–political decision. Courts have virtually no control over the decision.

2.2 France

Article 64 of the French Penal Code provides: 'If the person charged with the commission of a felony or misdemeanor was then insane or acted by absolute necessity, no offense has been committed.' The defense need not establish any causal relationship between the mental illness and the behavior.

If insanity is raised as an issue, the court appoints an attorney as an officer of the court to conduct an inquiry into the state of the defendant's mental health at the time of the offense. The court-approved psychiatrists make their evaluations as to the sanity of the defendant and report their findings to the judge. The judge has the final say as to the defendant's sanity or insanity. The defendant found to be insane then undergoes further examination to determine whether he or she is mentally ill and dangerous at the present time. The defendant so found is sent either to a prison or to a hospital, where he or she remains until both the medical experts and the civil authorities give approval for release.

2.3 Germany

The German Penal Code drafted in 1972 reads much like the criteria spelled out in the ALI test. It states: 'Whoever is incapable, at the time of committing the offense, of appreciating the wrongfulness of the offense or of acting according to such appreciation because of a pathological mental defect, a seriously disturbed consciousness, feeblemindedness, or another serious form of mental deviance, acts without blameworthiness.' The code also allows for diminished responsibility: it states: 'If the actor's capability to appreciate the wrongfulness of the offense or to act according to such appreciation is, at the time of committing the offense, significantly reduced for one of the reasons named in Article 20, the punishment can be reduced in accordance with paragraph 49.'

If the defendant meets the criteria defined in Articles 20 or 21, the court will order an evaluation of his or her act and a determination of his or her dangerousness to the public. If the offender is found to be dangerous, he/she will be committed to a psychiatric hospital.

The court can, at any time, examine whether the further execution of detention is to be suspended on parole. Such examination would occur after the offender has been in a psychiatric hospital for one year.

2.4 Switzerland

Switzerland—like Germany—applies a criterion for determining insanity that is similar to the ALI test. It states: 'If a defendant was mentally ill at the time of the offense and had no capacity to appreciate the wrongfulness of his act or to conform his conduct to the requirement of the law, he has a personal excuse which precludes his being punished for his conduct.'

When the insanity issue is raised, the court appoints an expert to examine the defendant. The judge invariably follows the opinion of its own appointed expert.

If the defendant is declared not guilty by reason of insanity, he or she may be released or committed to a prison or a hospital. Release is usually a medical decision, which needs approval by an executive authority.

2.5 Spain

The Spanish code drafted in 1982 has the breadth and lack of specificity noted in the French code. It states that a person who is considered mad or in a state of temporary mental confusion, and who has committed an act that the law has sanctioned as a criminal offense—is exempt from criminal responsibility. A tribunal will decree the offender's committal into a mental hospital, where he or she will remain until otherwise authorized by the same tribunal.

3. Comparison between US and Europe

Many of the West European Countries are less concerned about attributing specific individual blame and responsibility than the US, which emphasizes the relationship between specific characteristics of the defendant's mental state and the act committed. The European systems tend also to focus more on treatment of the offender and on protection of society. In the US, the system works harder to establish blame or the lack of it, and expends less energy on treatment.

The West European countries are also more likely to defer to medical opinion and to extract medical testimony from the partisan system. Experts' views have greater influence in European courts than in US proceedings—not only on the issue of the defendant's responsibility, but also in decisions concerning commitment and subsequent release.

4. Selective Other Societies

4.1 Argentina

Argentina's statute concerning the defense of insanity adopted by Argentina in 1963 reads much like the ALI standard. Article 34 of the penal code states that the following persons are not criminally liable: 'Anybody who at the time of the commission of the crime could not appreciate the unlawfulness of the deed or control his actions by reason of insufficiency or diseased disturbance of his mind, or by unconsciousness, or by error of fact or ignorance for which he is not responsible.'

A defendant found not guilty by reason of insanity is automatically committed to an insane asylum. Release is dependent on a judicial hearing in the presence of a public prosecutor and on the presentation of a medical expert's report that, in his or her view, the patient no longer constitutes a danger to self or others.

4.2 Israel

The Israeli Penal Law updated in 1983 reads as follows: 'A person shall not bear the criminal responsibility for an act that he has committed if, by reason of a mental illness or defect, he is incapable of choosing between performing the act and refraining from doing so.'

4.3 Kenya

Kenya's standard also reads much like M'Naghten: 'Where an act or omission is charged against a person as an offence, and it is given in evidence in the trial of that person for that offence that he was insane so as not to be responsible for his acts or omissions at the time when the act was done or the omission made, then if it appears to the court before which the person is tried that he did the act or made the omission charged but was insane at the time he did or made it, the court shall make a special finding to the effect that the accused was guilty of the act or omission charged but was insane when he did the act or made the omission.'

If a defendant is found to be insane, the case is reported to the president; the accused is placed in custody—in either a mental hospital or a prison—and remains under the jurisdiction of the court.

4.4 Japan

The Japanese Criminal Code—amended in 1954—states simply: 'An Act of a person of unsound mind is not punishable. Punishment shall be reduced for acts of weak-minded persons.'

4.5 South Korea

The South Korean Criminal Code—amended in 1960—reads as follows: 'A person is not punishable if, because of mental disorder, he is unable to pass rational judgment or to control his will.'

South Korea also allows for diminished responsibility: 'The punishment of a person who, due to a mental disorder, is deficient in the capacity mentioned in the preceding Section, shall be mitigated.'

See also: Crime and Class; Criminal Defenses; Criminal Law and Crime Policy; Deterrence: Legal Perspectives; Expert Testimony; Guilt

Bibliography

Allen F 1964 *The Borderline of Criminal Justice*. University of Chicago Press, Chicago, p. 105

Simon R J, Aaronson D E 1988 *The Insanity Defense*. Praeger, New York, p. 22

Stone A 1975 *Mental Health and Law: A System in Transition*. DHEW pub. no. (ADM) 75–176. NIMH, Rockville, MD, p. 218

R. J. Simon

Law: Economics of its Public Enforcement

In this article we consider the economic theory of the public enforcement of law—the use of public agents (inspectors, tax auditors, police, prosecutors) to detect and to sanction violators of legal rules. Economically-oriented analysis of public law enforcement dates from the eighteenth century contributions of Montesquieu (1748), Beccaria (1767), and, especially, Bentham (1789). Curiously, after Bentham (1789), the subject of enforcement lay essentially dormant in economic scholarship until the late 1960s, when Becker (1968) published a highly influential article that has led to a voluminous literature. In Sects. 1 through 3, we present the basic elements of the theory of public enforcement. Our concern is with the probability of imposition of sanctions, the magnitude and form of sanctions, and the rule of liability. In Sects. 4 through 14 we then examine a variety of extensions of the central theory, including accidental harms, costs of imposing fines, mistake, marginal deterrence, settlement, self-reporting, repeat offenses, and incapacitation. (A more expansive treatment of the subject of this article is contained in Polinsky and Shavell 2000.)

1. The Basic Framework

An individual (or a firm) chooses whether to commit an act that for simplicity is assumed to cause harm with certainty. If he commit the act, he obtains some gain, and also faces the risk of being caught, found liable, and sanctioned. The rule of liability could be either strict—under which he is definitely sanctioned; or fault-based—under which he is sanctioned only if his behavior fell below a fault standard. The sanction

that he suffers could be a monetary fine, a prison term, or a combination of the two.

Whether an individual chooses to commit a harmful act is determined by an expected utility calculation. He will commit the act if that would raise his expected utility, taking into account the gain he would derive and the subsequent probability of being caught and sanctioned. We will usually first examine the assumption that individuals are risk neutral with respect to sanctions, that is, that they treat an uncertain sanction as equivalent to its expected value; but we will also consider alternative assumptions.

Social welfare is presumed to equal the sum of individuals' expected utilities. An individual's expected utility depends on whether he commits a harmful act, on whether he is a victim of someone else's harmful act, and on his tax payment, which will reflect the costs of law enforcement, less any fine revenue collected. If individuals are risk neutral, social welfare can be expressed simply as the gains individuals obtain from committing his acts, less the arms caused, and less the costs of law enforcement.

We assume, as is conventional, that fines are socially costless to employ because they are mere transfers of money, whereas imprisonment involves positive social costs because of the expense associated with the operation of prisons and the disutility due to imprisonment.

The enforcement authority's problem is to maximize social welfare by choosing enforcement expenditures, or, equivalently, a probability of detection, the level of sanctions and their form (a fine, prison term, or combination), and the rule of liability (strict or fault-based).

2. *Optimal Enforcement Given the Probability of Detection*

We consider here optimal enforcement given the assumption that the probability of detection is fixed (the probability will be treated as a policy instrument in the next section). Thus, we ask about the optimal form and level of sanctions under strict and fault-based liability, and about how the two liability rules compare.

2.1 *Strict Liability*

Assume initially that fines are the form of sanction and that individuals are risk neutral. Then the optimal fine is the harm h divided by the probability of detection p, that is, h/p; for then the expected fine equals the harm (observe that $p(h/p) = h$). If, for example, the harm is \$1,000 and the probability of detection is 25 percent, then the optimal fine is \$4,000, and the expected fine is \$1,000. This fine is optimal because, when the expected fine equals the harm, an individual will commit a harmful act if, and only if, the gain he would derive from it exceeds the harm he would cause. Essentially this basic formula was noted by Bentham (1789, p. 173) and it has been commented upon by many others since.

If individuals are risk averse with regard to fines, the optimal fine would tend to be lower than in the risk-neutral case for two reasons. First, reducing the fine reduces the bearing of risk by individuals who commit the harmful act. Second, because risk-averse individuals are more easily deterred than risk-neutral individuals, the fine does not need to be as high as before to achieve any desired degree of deterrence.

Next assume that imprisonment is the form of sanction. In this case, there is not a simple formula for the optimal imprisonment term. The optimal term could be such that there is either underdeterrence or overdeterrence, compared to socially ideal behavior. On one hand, a relatively low imprisonment term, implying underdeterrence, might be socially desirable because it means that imprisonment costs are reduced with respect to those individuals who commit harmful acts. On the other hand, a relatively high term, implying overdeterrence, might be socially desirable because it means that imprisonment costs are reduced due to fewer individuals committing harmful acts.

Now consider the combined use of fines and imprisonment. Here, the main point is that fines should be employed to the maximum extent feasible before resort is made to imprisonment. In other words, it is not optimal to impose a positive imprisonment term unless the fine is maximal. (The maximal fine might be interpreted as the wealth of an individual.) The rationale for this conclusion is that fines are socially costless to impose, whereas imprisonment is socially costly, so deterrence should be achieved through the cheaper form of sanction first. This point is noted by Bentham (1789, p. 183) and Becker (1968, p. 193); see also Polinsky and Shavell (1984).

2.2 *Fault-based Liability*

Assume again that fines are the form of liability. Then the same formula for the fine that we said was optimal under strict liability—namely, h/p, the harm divided by the probability of detection—will lead to compliance with the fault standard (assuming that the fault standard is optimally selected).

If individuals are risk averse, they are deterred more easily than if they are risk neutral, so the fine does not need to be as high to induce compliance with the fault standard. Moreover, assuming that compliance occurs, no one actually is sanctioned because no one is found at fault (provided that there are no mistakes). Thus, fault-based liability has the attractive feature that it can accomplish desired deterrence of harm-creating conduct without imposing risk on risk-averse individuals (Shavell 1982).

Next, consider imprisonment as the sanction; see Shavell (1987a). Here, for essentially the reasons given in the case of fines, any sanction above a threshold level will ensure compliance with the fault standard, and the minimum sanction necessary to induce compliance is higher the lower is the probability of detection. Also (Shavell 1985), fault-based liability can accomplish deterrence without the actual imposition of costly imprisonment sanctions.

Finally, consider the joint use of fines and imprisonment. In this case, it does not matter what the combination of sanctions is, provided that the sanctions achieve compliance with the fault standard.

2.3 Comparison of Liability Rules

Because sanctions are not imposed under fault-based liability (in the absence of mistakes), this form of liability has an advantage over strict liability when the sanction is a fine and individuals are risk-averse, or when the sanction is imprisonment. However, fault-based liability is more difficult to administer. Namely, to apply fault-based liability, the enforcement authority must have more information than under strict liability: it must be able to calculate optimal behavior to determine the fault standard and it must ascertain whether the fault standard was met. Under strict liability, the authority need only ascertain harm. (Moreover, for reasons we discuss in Sect. 6 below, strict liability encourages better decisions by injurers regarding their level of participation in harm-creating activities.)

3. Optimal Enforcement Including the Probability of Detection

We now consider the optimal system of enforcement when expenditures on enforcement, and hence the probability of detection, are allowed to vary. Consideration of this issue originated with Becker (1968).

3.1 Strict Liability

Assume first that the sanction is a fine and that individuals are risk neutral. Then the optimal level of the fine is maximal and the optimal probability is low (in a sense to be described). The basic explanation for this conclusion is that if the fine were not maximal, society could save enforcement costs by simultaneously raising the fine and lowering the probability without affecting the level of deterrence. Suppose, for example, that the fine initially is \$4,000 and that the probability of detection is 25 percent. Now raise the fine to \$10,000, presuming that the maximal fine is at least this high, and lower the probability of detection to 10 percent. Then the expected fine remains equal to \$1,000, so that deterrence is maintained, but expenditures on enforcement are reduced significantly, implying that social welfare rises. This process can be continued, and social welfare augmented, whenever the fine is below the maximal level f_m. Becker (1968) suggested this result; Carr-Hill and Stern (1979, pp. 280–309) and Polinsky and Shavell (1979) note it explicitly.

The optimal probability is low in the sense that there is some underdeterrence; that is, the optimal p is such that the expected fine pf_m is less than the harm h. The reason for this result is that if pf_m equals h, behavior will be ideal, meaning that the individuals who are just deterred obtain gains essentially equal to the harm. These are the individuals who would be led to commit the harmful act if p were lowered slightly. Lowering p will be socially beneficial because these individuals cause no net social losses (their gains essentially equal the harm), but reducing p saves enforcement costs.

If individuals are risk averse, the optimal fine generally is less than maximal, as first shown in Polinsky and Shavell (1979) (and elaborated upon in Kaplow 1992). This is because the use of a very high fine would impose a substantial risk-bearing cost on individuals who commit the harmful act.

Next, assume that the sanction is imprisonment and that individuals are risk neutral in imprisonment, that is, the disutility of imprisonment is the same for each additional year. Then the optimal imprisonment term is maximal. The reasoning behind this result parallels that used to show that the optimal fine is maximal when individuals are risk neutral in fines. Specifically, if the imprisonment term is raised and the probability of detection lowered so as to keep the expected sanction constant, neither individual behavior nor the costs of imposing imprisonment are affected (by construction, the expected prison term is the same), but enforcement expenditures fall.

Suppose instead that individuals are risk averse in imprisonment. In other words, the disutility of each year of imprisonment grows with the number of years in prison, perhaps because imprisonment becomes increasingly difficult to tolerate. In this case there is a stronger argument for setting the imprisonment sanction maximally than when individuals are risk neutral (Polinsky and Shavell 1999). This is because, when the imprisonment term is raised, the probability of detection can be lowered even more than in the risk-neutral case without reducing deterrence. Thus, not only are there greater savings in enforcement expenditures, but also the social costs of imposing imprisonment sanctions decline because the expected prison term falls.

Last, suppose that individuals are risk preferring in imprisonment, that is, the disutility of each year of imprisonment falls with the number of years in prison. This assumption seems particularly

important: the first years of imprisonment may create special disutility, due to brutalization of the prisoner, or due to the stigma effect of having been imprisoned at all. Additionally, the fact that individuals discount the future disutility of imprisonment makes earlier years of imprisonment more important than later ones. If individuals are risk preferring in imprisonment, the optimal sanction may be less than maximal (Polinsky and Shavell 1999). In particular, the type of argument used above does not necessarily apply. When the sanction is raised, the probability that maintains deterrence cannot be lowered proportionally, implying that the expected prison term rises. Because the resulting increased cost of imposing imprisonment sanctions might exceed the savings in enforcement expenditures from lowering the probability, the optimal prison term might not be maximal.

When the probability of detection is set optimally, together with the imprisonment term, underdeterrence may well result, not only to save enforcement expenditures, but also to reduce the costs of imposing imprisonment sanctions.

3.2 Fault-based Liability

The least expensive way to accomplish compliance with the fault standard is to use the highest possible sanction and, given this sanction, the lowest probability of detection that deters individuals who would be at fault. The reason is that, if all individuals who would be at fault are deterred, the only cost incurred is associated with the setting of the probability; this cost is minimized by using the maximal sanction and a correspondingly low probability. This is true regardless of whether the sanction is a fine or imprisonment and regardless of individuals' attitudes toward the risk of fines or of imprisonment.

3.3 Comparison of Liability Rules

As we emphasized earlier, under fault-based liability sanctions are not actually imposed (in the absence of mistakes), which often is an advantage over strict liability. However, this advantage of fault-based liability would have to be weighed against the disadvantages of this rule that we mentioned at the end of Sect. 2.

4. Accidental Harms

Until now, we have assumed that individuals decide whether or not to commit acts that cause harm with certainty, that is, they decide whether or not to cause intentional harms. In many circumstances, however, harms are accidental—they occur only with a probability. For instance, if a firm stores toxic chemicals in a substandard tank, the firm only creates the probability of a harmful spill.

Essentially all that we have said above applies in a straightforward way when harms are accidental. If parties are risk-neutral, sanctions are monetary, and the expected sanction equals harm, then induced behavior will be socially optimal; further, the optimal magnitude of sanctions is maximal if individuals are risk neutral because this allows enforcement costs to be saved; and so forth.

There is, however, an additional issue that arises when harm is uncertain. A sanction can be imposed either on the basis of the commission of a dangerous act that increases the chance of harm—storing chemicals in a substandard tank—or on the basis of the actual occurrence of harm—only if the tank ruptures. In principle, either approach can achieve optimal deterrence. If liability is based on the dangerous act, the expected fine should equal the expected harm, while if liability is based on actual harm, the expected fine should equal the actual harm.

Several factors are relevant to the choice between act-based and harm-based sanctions (Shavell 1993). First, act-based sanctions need not be as high to accomplish a given level of deterrence (because expected harm is less than actual harm), and thus are more likely to be able to be paid. Second, because act-based sanctions can be lower, they tend to be preferable when parties are risk averse. Third, act-based and harm-based sanctions may differ in the ease with which they can be applied. Act-based sanctions may be simpler to impose (it might be less difficult to determine whether an oil shipper properly maintains its vessels' holding tanks than to detect whether one of the vessels leaked oil into the ocean); or harm-based sanctions may be easier to impose (a driver who causes harm might be caught without difficulty, but not one who speeds). Fourth, it may be hard to calculate the expected harm due to an act, but relatively easy to ascertain the actual harm if it eventuates, thereby favoring harm-based liability.

5. Costs of Imposing Fines

Now suppose that there are costs borne by enforcement authorities in imposing fines. Our principal observation is that such costs should raise the level of the fine. To see why, suppose for simplicity that the probability of detection is fixed, that liability is strict, and that individuals are risk neutral. In this setting, recall from Sect. 2 that the optimal fine is h/p, the harm divided by the probability of detection. Now let there be a public cost k of imposing a fine. The optimal fine then becomes $h/p+k$; in other words, the cost k should be added to the fine that would otherwise be desirable (Becker 1968, p. 192, Polinsky and Shavell 1992). The intuition behind this result is that, if an individual commits a harmful act, he causes society to bear not only the immediate harm h, but also, with probability p, the cost k of imposing the fine—that is, his act results in an expected total social cost of $h+pk$.

If the fine is $h/p+k$, the individual's expected fine is $p[(h/p)+k] = h+pk$, leading him to commit the harmful act if and only if his gain exceeds the expected total social cost of the act. Hence, he will behave in a socially appropriate way.

6. Level of Activity

We have been assuming that the sole decision that an individual makes is whether to act in a way that causes harm when engaging in some activity. In many contexts, however, an individual also makes a choice about his activity level—that is, whether to engage in that activity, or, more generally, at what level to do so. For example, besides deciding how to behave when driving (whether to exercise care in changing lanes), an individual also chooses how many miles to drive; the number of miles driven is the individual's level of activity.

The socially optimal activity level is such that the individual's marginal utility from engaging in the activity just equals the marginal expected harm caused by the activity. Thus, the optimal number of miles driven is the level at which the marginal utility of driving an extra mile just equals the marginal expected harm per mile driven.

Will parties' choices about their activity levels be socially correct under the two major forms of liability? The answer is that under strict liability, their choices about activity levels will be correct, but under fault-based liability, they generally will participate in activities to a socially excessive extent. Under strict liability, parties will choose the optimal level of activity because they will pay for all harm done. Under fault-based liability, however, parties generally do not pay for the harm they cause because, as we have discussed, they will tend to behave so as not to be found at fault. Consequently, when deciding on their level of activity, they will not take into account the harm that their participation in the activity causes, and therefore they will participate too much.

The interpretation of the preceding points in relation to firms is that under strict liability, the product price will reflect the expected harm caused by production. Hence, the amount purchased, and thus the level of production, will tend to be socially optimal. However, under fault-based liability, the product price will not reflect harm, but only the cost of precautions; thus, the amount sold, and the level of production, will be excessive.

The tendency of parties to choose an excessive level of activity under fault-based liability, but not under strict liability, was first emphasized in Shavell (1980) and Polinsky (1980).

7. Mistakes

An individual who should be found liable might mistakenly not be found liable—a Type I error. Alternatively, an individual who should not be found liable might mistakenly be found liable—a Type II error. For an individual who has been detected, let the probabilities of these errors be ε_1 and ε_2, respectively.

Given the probability of detection p and the chances of Type I and Type II errors, an individual will commit the wrongful act if and only if his gain g net of his expected fine if they do commit it exceeds his expected fine if they do not commit it, namely, when $g-p(1-\varepsilon_1)f > -p\varepsilon_2 f$, or, equivalently, when $g > (1-\varepsilon_1-\varepsilon_2)pf$.

The first point to note is that, as emphasized in Png (1986), *both* types of error reduce deterrence: the term $(1-\varepsilon_1-\varepsilon_2)pf$ is declining in both ε_1 and ε_2. The first type of error diminishes deterrence because it lowers the expected fine if an individual violates the law. The second type of error, mistaken liability, also lowers deterrence because it reduces the difference between the expected fine from violating the law and not violating it—in effect, making a violation less costly to the individual.

Because mistakes dilute deterrence, they tend to reduce social welfare. Specifically, to achieve any level of deterrence, the probability p must be higher to offset the effect of errors.

If individuals are risk averse, the possibility of mistakes may increase the desirability of lowering the fine because, due to Type II errors, individuals who do not violate the law are subject to the risk of having to pay a fine (Block and Sidak 1980).

When liability is based on fault, an important implication of mistake is that some individuals will bear sanctions even if they comply with the fault standard, tending to make fault-based liability operate like strict liability. Moreover, as stressed by Craswell and Calfee (1986), individuals will often have a motive to take excessive precautions in order to reduce the chance of erroneously being found at fault.

8. General Enforcement

In many settings, enforcement may be said to be general in the sense that several different types of violations will be detected by an enforcement agent's activity. For example, a police officer waiting at the roadside may notice a driver who litters as well as one who goes through a red light or who speeds. To investigate such situations, suppose that a single probability of detection applies uniformly to all harmful acts, regardless of the magnitude of the harm.

When enforcement is general in this sense, the optimal sanction rises with the severity of the harm and is maximal only for relatively high harms; this point was first made in Shavell (1991) (Mookherjee and Png 1992 is closely related). Suppose, for example, that liability is strict, the sanction is a fine, and injurers are risk neutral. Let $f(h)$ be the fine given harm h.

Then, for any given general probability of detection p, the optimal fine schedule is h/p, provided that h/p is feasible; otherwise the optimal fine is maximal. This schedule is optimal given p because it implies that the expected fine equals harm, thereby inducing ideal behavior, whenever that is possible. The result that, when enforcement is general, sanctions should rise with the severity of harm up to a maximum also holds if the sanction is imprisonment and if liability is fault-based.

9. *Marginal Deterrence*

In many circumstances, a person may consider which of several harmful acts to commit, for example, whether to release only a small amount of a pollutant into a river or a large amount. In such contexts, the threat of sanctions influences which harmful acts individuals choose to commit. Deterrence of a more harmful act because its sanction exceeds that for a less harmful act is sometimes referred to as marginal deterrence (apparently so named by Stigler 1970).

Other things being equal, as observed by Beccaria (1767, p. 32) and Bentham (1789, p. 171), it is socially desirable that enforcement policy creates marginal deterrence, so that law violators have a reason to moderate the amount of harm they cause. This suggests that sanctions should rise with the magnitude of harm and, therefore, that sanctions generally should not be maximal. Note that marginal deterrence also can be promoted by increasing the probability of detection for more severe harms. For formal analyses of marginal deterrence, see Shavell (1992), Wilde (1992), and Mookherjee and Png (1994).

10. *Principal–Agent Relationship*

Although we have assumed that an injurer is a single actor, injurers often are more appropriately characterized as collective entities, and specifically as a principal and the principal's agent. For example, the principal could be a firm and the agent an employee.

When harm is caused by the agent of a principal, many of our prior conclusions are not fundamentally altered; they simply carry over to the sanctioning of principals. Notably, if a risk-neutral principal faces an expected fine equal to harm done, he will in effect be in the same position vis-à-vis his agent as society is vis-à-vis a single violator of law (see Newman and Wright 1990 on a closely related point). Consequently, the principal will behave socially optimally in controlling his agent, and in particular will contract with him and monitor him in ways that will give the agent socially appropriate incentives to reduce harm (but see Arlen 1994).

The allocation of a financial sanction between the principal and the agent would not matter if, as would be the natural presumption, the parties can reallocate the sanction through their own contract. The allocation of the sanction would matter, however, if it would allow the parties to reduce their total burden, for example, if the agent is unable to pay a fine because his assets are less than the fine; see Sykes (1981) and Kornhauser (1982). Then, the fine should be imposed on the principal.

The imposition of imprisonment sanctions on agents may be desirable when their assets are less than the harm that they can cause; see Polinsky and Shavell (1993). Because an agent's assets are limited, the principal may be unable to control him adequately through use of contractually determined penalties, which can only be monetary. It may be socially valuable to use the threat of personal criminal liability and a jail sentence to remedy this problem.

11. *Settlements*

We consider here how settlements affect deterrence and the optimal system of public enforcement, and whether settlements are socially desirable.

There are two general reasons why parties might prefer an out-of-court settlement to a trial (see generally Cooter and Rubinfeld 1989, and regarding settlement in a criminal context, see, for example, Reinganum 1988 and Miceli 1996). First, a trial is costly in terms of time and/or money. Second, settlements eliminate the risks inherent in the trial outcome. For these reasons, settlement tends to be socially valuable.

But a complicating factor is that settlements dilute deterrence: for if injurers desire to settle, it must be because the expected disutility of sanctions is lowered for them (see generally Polinsky and Rubinfeld 1988). The state may be able to offset this effect by increasing the level of sanctions; if so, settlements are socially desirable for the reasons mentioned in the previous paragraph.

12. *Self-reporting*

We have assumed that individuals are subject to sanctions only if they are detected by an enforcement agent, but in fact parties sometimes disclose their own violations to enforcement authorities. For example, firms often report violations of environmental and safety regulations. There are two basic reasons why self-reporting is socially advantageous; see Kaplow and Shavell (1994b). First, self-reporting reduces enforcement costs because the enforcement authority does not have to identify and prove who the violator was. Second, self-reporting reduces risk. For example, drivers bear less risk because they know that if they cause an accident, they can report this to the police and suffer a lower and certain sanction, rather than face a substantially higher sanction (for hit-and-run

driving) imposed only with some probability. Self-reporting can be induced by lowering the sanction for individuals who disclose their own infractions. Although this will tend to reduce deterrence, the reward for self-reporting can be made small enough that deterrence is only negligibly reduced.

13. Repeat Offenders

In practice, the law often sanctions repeat offenders more severely than first-time offenders. This may be beneficial for two reasons. First, as developed in Polinsky and Shavell (1998), raising the sanction for repeat offenders may create additional deterrence: if getting caught violating the law implies not only an immediate sanction, but also a higher sanction for any future violation, an individual will be more deterred from committing a violation presently. Second, as studied, for example, in Rubinstein (1979) and Polinsky and Rubinfeld (1991), making sanctions depends on offense history and allows society to take advantage of information about the propensity of individuals to commit offenses: individuals with offense histories may be more likely than average to commit future violations, which might make it desirable for purposes of deterrence to impose higher sanctions on them. It also may be desirable to incapacitate such individuals by imprisoning them (see the next section).

14. Incapacitation

We have focused on the use of sanctions to reduce harm by deterring individuals from causing harm. However, an entirely different way to reduce harm is by incapacitating individuals so that they cannot cause harm. Imprisonment is the primary incapacitative sanction; on the economic theory of incapacitation, see Shavell (1987b).

If the sole function of sanctions were to incapacitate, the optimal sanction would be determined by comparing the expected harm, net of gains, an offender would cause if not in prison to the private and public costs of imprisonment. Jail should only be used to incapacitate individuals whose net harm is relatively high.

See also: Crime: Sociological Aspects; Crime, Sociology of; Deterrence; Law and Economics; Law, Sociology of; Norms

Bibliography

Arlen J A 1994 The potentially perverse effects of corporate criminal liability. *Journal of Legal Studies* **23**: 833–67

Beccaria C 1995 [1767] *On Crimes and Punishments, and Other Writings*. In: Bellamy R (ed.) [trans. Davies R et al.] Cambridge University Press, New York

Becker G S 1968 Crime and punishment: An economic approach. *Journal of Political Economy* **76**: 169–217

Bentham J 1789/1973 An introduction to the principles of morals and legislation. In: Bentham J (ed.) *The Utilitarians*. Anchor Books, Garden City, New York

Block M K, Sidak J G 1980 The cost of antitrust deterrence: Why not hang a price fixer now and then? *Georgetown Law Journal* **68**: 1131–39

Carr-Hill R A, Stern N H 1979 *Crime, the Police and Criminal Statistics: An Analysis of Official Statistics for England and Wales Using Econometric Methods*. Academic Press, London

Cooter R D, Rubinfeld D L 1989 Economic analysis of legal disputes and their resolution. *Journal of Economic Literature* **27**: 1067–97

Craswell R, Calfee J E 1986 Deterrence and uncertain legal standards. *Journal of Law, Economics, and Organization* **2**: 279–303

Kaplow L 1992 The optimal probability and magnitude of fines for acts that definitely are undesirable. *International Review of Law and Economics* **12**: 3–11

Kaplow L, Shavell S 1994 Optimal law enforcement with self-reporting of behavior. *Journal of Political Economy* **102**: 583–606

Kornhauser L A 1982 An economic analysis of the choice between enterprise and personal liability for accidents. *Californian Law Review* **70**: 1345–92

Miceli T J 1996 Plea bargaining and deterrence: An institutional approach. *European Journal of Law and Economics* **3**: 249–64

Montesquieu C 1748/1977 *The Spirit of the Laws*. University of California Press, Berkeley, CA

Mookherjee D, Png I P L 1992 Monitoring vis-à-vis investigation in enforcement of law. *American Economic Review* **82**: 556–65

Mookherjee D, Png I P L 1994 Marginal deterrence in enforcement of law. *Journal of Political Economy* **102**: 1039–66

Newman H A, Wright D W 1990 Strict liability in a principal-agent model. *International Review of Law and Economics* **10**: 219–31

Png I P L 1986 Optimal subsidies and damages in the presence of judicial error. *International Review of Law and Economics* **6**: 101–5

Polinsky A M 1980 Strict liability vs. negligence in a market setting. *American Economic Review* **70**: 363–7

Polinsky A M, Rubinfeld D L 1988 The deterrent effects of settlements and trials. *International Review of Law and Economics* **8**: 109–16

Polinsky A M, Rubinfeld D L 1991 A model of optimal fines for repeat offenders. *Journal of Public Economics* **46**: 291–306

Polinsky A M, Shavell S 1979 The optimal tradeoff between the probability and magnitude of fines. *American Economic Review* **69**: 880–91

Polinsky A M, Shavell S 1984 The optimal use of fines and imprisonment. *Journal of Public Economics* **24**: 89–99

Polinsky A M, Shavell S 1992 Enforcement costs and the optimal magnitude and probability of fines. *Journal of Law and Economics* **35**: 133–48

Polinsky A M, Shavell S 1993 Should employees be subject to fines and imprisonment given the existence of corporate liability? *International Review of Law and Economics* **13**: 239–57

Polinsky A M, Shavell S 1998 On offense history and the theory of deterrence. *International Review of Law and Economics* **18**: 305–24

Polinsky A M, Shavell S 1999 On the disutility and discounting of imprisonment and the theory of deterrence. *Journal of Legal Studies* **28**: 1–16
Polinsky A M, Shavell S 2000 The economic theory of public enforcement of law. *Journal of Economic Literature* **38**: 45–76
Reinganum J F 1988 Plea bargaining and prosecutorial discretion. *American Economic Review* **78**: 713–28
Rubinstein A 1979 An optimal conviction policy for offenses that may have been committed by accident. In: Brams S J, Schotter A, Schwödiauer G (eds.) *Applied Game Theory*. Physica-Verlag, Wűrzburg, Germany, pp. 406–13
Shavell S 1980 Strict liability versus negligence. *Journal of Legal Studies* **9**: 1–25
Shavell S 1982 On liability and insurance. *Bell Journal of Economics* **13**: 120–32
Shavell S 1985 Criminal law and the optimal use of nonmonetary sanctions as a deterrent. *Columbia Law Review* **85**: 1232–62
Shavell S 1987a The optimal use of nonmonetary sanctions as a deterrent. *American Economic Review* **77**: 584–92
Shavell S 1987b A model of optimal incapacitation. *American Economic Review* **77**: 107–10
Shavell S 1991 Specific versus general enforcement of law. *Journal of Political Economy* **99**: 1088–1108
Shavell S 1992 A note on marginal deterrence. *International Review of Law and Economics* **12**: 345–55
Shavell S 1993 The optimal structure of law enforcement. *Journal of Law and Economics* **36**: 255–87
Stigler G J 1970 The optimum enforcement of laws. *Journal of Political Economy* **78**: 526–36
Sykes A O 1981 An efficiency analysis of vicarious liability under the law of agency. *Yale Law Journal* **91**: 168–206
Wilde L L 1992 Criminal choice, nonmonetary sanctions, and marginal deterrence: A normative analysis. *International Review of Law and Economics* **12**: 333–44

A. M. Polinsky and S. Shavell

Law Firms

The overwhelming majority of lawyers in private practice in most countries around the world practice alone or in very small and relatively informal groupings (Abel and Lewis 1988–95). Beginning in the 1960s, however, there has been a steady trend for lawyers who represent business interests to practice in larger and more bureaucratically organized law firms. Although lawyers for individuals have also attempted to organize larger firms, these efforts have been largely unsuccessful (Van Hoy 1997). As a result, at the end of the twentieth century the large corporate law firm is the dominant organizational model for private practitioners.

1. *The Evolution of the Large Law Firm: From Cravathism to Megalawyering*

The American law firm of Cravath, Swaine, and Moore is widely credited with creating the blueprint for the modern large law firm in the first decade of the twentieth century. The 'Cravath System' consists of four interrelated practices. First, lawyers, typically called 'associates,' are hired directly out of law school, paid a fixed salary, and required to work exclusively for the firm. Second, these new recruits are hired for a probationary period during which they are trained by the senior lawyers in the firm, called 'partners,' to handle matters of increasing responsibility for the firm's clients. Third, at the end of the probationary period, the partners select only the best associates for promotion to partnership, requiring those who are not selected to leave the firm. Finally, partners share in both the firms' profits and management, including, most importantly, the selection of new partners (Swaine 1946).

Firms patterned on the Cravath model grew dramatically in size and geographic scope from 1960 to the end of the twentieth century. In 1960, there were 38 US law firms with more than 50 lawyers, with the largest consisting of 125 attorneys (Smigel 1969). Forty years later, there were 250 law firms in the USA with more than 100 lawyers, with several exceeding 1,000 attorneys (National Law Journal 1999). In 1960, virtually all firms consisted of a single office. By 2000, most large firms had multiple offices, with many practicing in several jurisdictions. In 1960, only a handful of US firms had foreign offices. Forty years later, foreign offices of the top 250 American law firms were located in 72 foreign cities and employed almost 5,000 US and foreign lawyers.

During this same period, 'megafirms' (Galanter and Palay 1991) patterned on the Cravath model began to emerge in the UK, Europe, Asia, and other commercial centers. England was the first to develop comparable firms to those found in the USA (Flood 1989). By the end of the century, these firms were among the largest and most globalized firms in the world. Lawyers in the Netherlands (Blankenburg and Bruinsma 1994), Germany (Gerber 1999), Spain (Stewart 1991), Canada (Arthurs et al. 1988), Australia (Galanter and Palay 1991), China (Alford 1995), and Venezuala (Perez Perdomo 1988) have also developed their own versions of the large law firm. Although the structure and culture of these firms reflect important national and regional differences (Trubek et al. 1994), the Cravath System continues to exert a powerful influence on the development of large law firms around the world. As a result, theories about how US law firms are structured are likely to have important implications for firms in other countries that are patterned on the American model.

2. *Explaining Law Firm Growth*

Theorists offer three related explanations for law firm growth: tournament theory, portfolio theory, and demand theory. Although each theory highlights important truths, none captures the dynamic interaction between markets, institutions, and lawyer

entrepreneurship that has fueled the expanding market for corporate legal services.

2.1 The Tournament of Lawyers

Tournament theory posits that law firms must promote a fixed percentage of their entering associates to partnership in order to overcome the mutual monitoring problems of partners and associates inherent in the Cravath System (Galanter and Palay 1991). Partners, on this view, must protect their human capital (i.e., reputation, skill, and relationships with clients) from associates who might be tempted to damage partners by shirking on their assignments or grabbing the firm's clients and leaving. Associates, on the other hand, must protect themselves against partners who appropriate their hard work during the probationary period only to renege on their promise to promote those who perform well to partnership. Promoting a fixed percentage of every entering class of associates to partnership, tournament theorists assert, solves both problems by giving associates the incentive to maximize their chances for one of these prized slots by working hard with relatively little supervision, and at the same time providing associates with an objective measure (the number of associates promoted to partnership in any given class) of whether the partners are keeping their promise to reward the best associates. Since firms must hire at least one new associate for every new partner they promote (to leverage the new partner's human capital), tournament theorists conclude that the Cravath System's 'promotion-to-partner tournament' creates an internal growth engine that causes firms to grow exponentially in size.

Tournament theory contributes to understanding—but does not fully explain—the growth of large law firms. For the reasons outlined in Sect. 3.3, the internal labor markets of large law firms are not structured as rank order tournaments. Although the competition for partnership continues to play an important role in these institutions, two aspects of contemporary law firms limit its usefulness as an explanatory variable in law firm growth. First, many associates joining large law firms at the beginning of the twenty-first century have no intention of staying long enough to become partners in the firm (Wilkins and Gulati 1998). Second, rather than growing primarily by promoting associates who begin their careers with the firm, many of the largest and most successful law firms have grown through lateral hiring of senior associates and partners and by mergers with other firms (Johnson 1991). Each of these developments demonstrates that something other than the quest for promotion is driving law firm growth.

2.2 Portfolio Theory

Portfolio theory offers a potential explanation for why firms have increased their lateral hiring, particularly at the partner level. According to this view, law firms exist primarily to help lawyers manage and spread risk (Gilson and Mnookin 1985). A solo practitioner who loses her biggest client will face catastrophic consequences. A lawyer working in a law firm—particularly one that shares profits equally among the partners—is partly protected from this risk by the presence of a portfolio of firm clients. The more diversified the portfolio, the greater the protection e.g., firms can do initial public offerings when the economy is strong and bankruptcy work when the economy is weak. As a result, firms have strong incentives to grow by acquiring new partners (who bring new clients to the firm) and by developing new practice specialties.

Portfolio theory also fails to provide a full explanation for law firm growth. There has been a steady shift toward full service law firms. But this development does not appear to be fueled by the desire to minimize risk. Plaintiff's lawyers, who in the USA receive the bulk of their income from contingent fees that are only awarded at the end of successful litigation, are subject to far more downside risk than corporate lawyers who are typically paid by the hour. Yet plaintiffs firms generally are small and comprised almost exclusively of lawyers who are doing similar work (Coffee 1986). Moreover, as firms have grown larger they have tended to switch from lock step compensation for partners to 'eat what you kill' arrangements that reward partners for their individual contribution to the firm's profits. Marginal productivity compensation systems defeat the entire purpose of risk-spreading among partners.

2.3 Demand Theory

By concentrating on the internal dynamics of law firms, both tournament and portfolio theory neglect one obvious influence on the size of law firms—the demands of clients. Since 1970, there has been a tremendous increase in demand for all legal services in the USA (Sander and Williams 1989). Demand is strongest in the corporate sector, which in 1992 purchased more than $60 billion in legal services from American law firms (Galanter 1999). Law firms have obvious incentives to hire new lawyers, open new offices, and acquire new specialties to meet this demand.

Although client demand undoubtedly spurs law firm growth, it too is only a partial explanation. Demand side theories treat lawyers as passively responding to client pressure. American lawyers, however, have always played an active role in promoting the demand for their services. Thus, the lawyers who encouraged their corporate clients to use litigation as a tool for resolving business disputes are partly responsible for the increase demand for litigators. Similarly, law firms such as Skadden, Arps,

Meagher, Slate and Flom, helped fuel the mergers and acquisition boom by developing new strategies for hostile takeovers. It is this entrepreneurial style of lawyering that is spreading around the world.

2.4 Toward a Dynamic Theory of Growth

A complete theory of law firm growth must take account of the dynamic interaction among institutional structures, client demands, and lawyer entrepreneurship. Large law firms in the USA initially developed at the turn of the twentieth century as the large industrial concerns and financial institutions consolidated and formalized their power. The lawyers who sought to be the trusted advisers to these new economic enterprises adapted their practices—including developing expertise in specialties such as finance, reorganization, and antitrust, and creating internal practices (such as *de facto* tenure for partners, associates who owe their entire loyalty to the firm, and locating their offices in the buildings of their largest clients) that created symbiotic and durable relationships with their clients' owners and managers. As corporations reorganized and globalized a century later, it is not surprising that the law firms that serve these multinationals followed suit, expanding in size and geographic scope to better serve the competitive needs of their clients, while at the same time driving their clients to seek out and cultivate new markets. As the revolution in information technology opens the door for new forms of industrial and financial organization, we should not be surprised to see law firms creating new products and services, including ancillary businesses or joint ventures with nonlawyers to deliver nonlegal business and consulting services, designed to persuade companies to retain law firms—as opposed to other professionals—to guide them into the new economy (National Law Journal 1992).

3. Large Law Firms and the Public Interest

In addition to explaining the growth of large law firms, the dynamic interaction between the entrepreneurial initiatives of lawyers in these institutions and the expanding demands of their corporate clients has important implications for the distribution of legal services, the integrity of the public legal framework, and the gender and racial integration of the bar.

3.1 Why the 'Haves' Continue to Come Out Ahead

Corporations consume an increasingly large share of the legal services produced by US lawyers (Heinz et al. 1998). As a result, individuals have a more difficult time finding advocates willing and able to protect their legal rights (Hadfield 2000). Moreover, because access to law is a relative, as well as an absolute good, the more legal resources corporate interests employ the harder it will be for individuals to challenge corporate power in those areas—tort litigation, labor negotiations, health and safety regulation—where the interests of the two kinds of clients are likely directly to conflict (Galanter 1974).

3.2 The Paradox of Professional Independence

The American bar's ethical rules mandate that lawyers must be both zealous advocates for the interests of their clients and officers of the legal system who uphold the public purposes of the legal framework. To fulfill these dual obligations, lawyers are expected to be independent from both the state and their clients. Large law firms have always posed a potential threat to the public responsibilities of the bar. As early as the 1930s, critics complained that the emerging 'law factories' threatened the profession's independence by fostering a business perspective and an obsession with money that undermined the bar's commitment to autonomy and public service (Berle 1933). For the next 50 years, corporate lawyers were largely successful in rebuffing these charges by pointing to the important public service performed by high profile partners in leading law firms and by asserting that lawyers in large firms maintained the highest standards of professionalism. Although law firm partners continue to assert that they are the ultimate independent professionals, studies of the actual practices of large firms underscore the degree to which corporate lawyers are increasingly beholden to the interests of their powerful clients (Nelson 1988).

This fact has important implications for the rule of law and the integrity of the public legal framework. As corporate lawyers work longer and more diligently to attract and retain corporate business, they have less time for the public service and *pro bono* legal work that in the past helped to mitigate the imbalance of legal resources between corporate and individual clients (Kronman 1993). More important, corporate lawyers who come to identify strongly with their clients' interests are more likely to adopt an instrumental view of legal rules that treats regulatory constraints as little more than the probability of detection discounted by the price of enforcement (Gordon 1988). As the American savings and loan crisis in the 1990s amply demonstrates, such an approach to corporate counseling can have disastrous consequences for the public fisc in circumstances where the clients who receive this kind of instrumental advice are in a position to subvert important public regulatory protections (Wilkins 1993).

3.3 Winners and Losers in the Tournament of Lawyers

Many who are skeptical that tournament theory explains the growth of large law firms nevertheless

continue to use the tournament model to describe the internal labor markets of these institutions. The hiring, promotion, and compensation practices of law firms, however, differ from those assumed in tournament theory in four important respects (Wilkins and Gulati 1998). First, as indicated above, not everyone who 'enters' the tournament as an associate is competing for partnership. Second, unlike a sporting event, associates do not compete on a level playing field. Instead, the high associate-to-partner ratios of large law firms ensure that only a few associates will receive the training and mentoring that they need to develop the legal skill and business contacts that will transform them into credible partnership candidates. Third, far from being neutral umpires who seek to promote only those associates who will best serve the interests of the entire firm, partners are actively competing with each other to promote associates who will further their personal interest in servicing their own clients. Fourth, firms award partnership prizes to those they believe will be the best partners—i.e., ranking associates on the basis of their perceived ability to bring in business and other revenue generating tasks rarely performed by associates—as opposed to rewarding those who have billed the most hours or otherwise demonstrated the most loyalty in their past work for the firm.

These differences from standard tournament theory affect who is likely to succeed in a large law firm. US law firms have abandoned most of the overt discriminatory policies that until the 1960s excluded most women, racial minorities, Jews, and even many Catholics. Nevertheless, 40 years later these institutions have made only modest progress in hiring, and most especially promoting to partnership, women and racial minorities (Chambliss 2000). The real rules of the tournament of lawyers help to explain this continuing lack of progress. The well-documented tendency for people to see potential in protégés who remind them of themselves makes it more difficult for minority associates to receive scarce training and developmental opportunities from white male partners (Wilkins and Gulati 1996). Women face the additional burden of being perceived as not as committed as male associates and therefore a riskier bet for partners who want to invest their time in associates who they believe will have long-term careers with the firm (Epstein et al. 1999). And, should they surmount these barriers, women and minorities both face important obstacles to developing the kind of client relationships with major corporations that would enable them to become powerful partners in firms that are increasingly driven by the bottom line (Epstein et al. 1999, Wilkins 1999).

4. *The Future of the Large Law Firm*

Large law firms face increasing competition from accounting firms, consultants, investment bankers, and technology companies. This competition is proceeding on three fronts. First, nonlegal professional service firms are forming multidisciplinary partnerships with lawyers that are capable of 'practicing law' and giving integrated advice at the intersection of law, business, and competitive strategy. Although prohibited in the USA, such partnerships are legal (albeit with certain restrictions) in many European countries, including France, Germany, and the Netherlands as well as in England, Canada, and Australia (New York State Bar 2000). Second, even in jurisdictions where multidisciplinary practice is prohibited, nonlegal professional service firms continue to advise corporate clients how to manage and reduce legal risk. Third, nonlegal professional service firms are increasingly competing with law firms for legal talent, both by hiring lawyers directly out of law school and raiding associates, and sometimes even partners, from large firms.

Large firms have responded to these competitive pressures by: (a) invoking professionalism norms to condemn multidisciplinary firms as unethical, (b) merging with mega law firms from other jurisdictions (or entering into global legal alliances or networks) for the purpose of creating law firms that can successfully compete with global nonlegal professional service firms, (c) creating new categories of workers (e.g., nonequity partners, senior counsels, permanent associates, and contract lawyers) to enable firms to grow and provide more services without decreasing partner profits, and (d) dramatically increasing starting salaries for entering lawyers (New York State Bar 2000). It remains to be seen whether any of these measures will be sufficient to maintain the dominant position that large law firms achieved in the twentieth century in the more competitive and knowledge-driven economy of the twenty-first century.

See also: Firm Behavior; Lawyers; Lawyers, Regulation of; Legal Professionalism; Political Lawyering; Public Interest Law

Bibliography

Abel R L, Lewis P S C (eds.) 1988–1995 *Lawyers in Society*, 4 Vols. University of California Press, Berkeley, CA

Alford W P 1995 Tasseled loafers for barefoot lawyers: Transformation and tension in the world of Chinese legal workers. *The China Quarterly* **147**: 27–38

Arthurs H, Weisman R, Zemans F 1988 Canadian lawyers: A peculiar professionalism. In: Abel R L, Lewis P S C (eds.) *Lawyers in Society: The Common Law World*. University of California Press, Berkeley, CA, Vol. 1

Berle A A 1933 Modern legal profession. *Encyclopedia of the Social Sciences* **5**: 340–5

Blankenburg E, Bruinsma F 1994 *Dutch Legal Culture*. Kluwer Law and Taxation Publishers, Deventer and Boston

Chambliss E 2000 *Miles to Go 2000: Progress of Minorities in the Legal Profession*. American Bar Association, Chicago, IL
Coffee J 1986 Understanding the plaintiff's attorney: The implications of economic theory for private enforcement of law through class and derivative actions. *Columbia Law Review* **86**: 669–727
Epstein C F, Serron C, Oglensky B, Saute R 1999 *The Part-Time Paradox: Time Norms, Professional Life, Family and Gender*. Routledge, New York
Flood J 1989 The growth of large law firms and it's effect on the legal profession and legal education: Megalaw in the UK: Professionalism or corporation? A preliminary report. *Indiana Law Journal* **64**: 569–92
Galanter M 1974 Why the 'haves' come out ahead: Speculations on the limits of legal change. *Law and Society Review* **9**: 95–160
Galanter M 1999 Old and in the way: the coming demographic transformation of the legal profession and its implication for the provision of legal services. *Wisconsin Law Review* **6**: 1081–113
Galanter M, Palay T 1991 *Tournament of Lawyers: The Transformation of the Big Law Firm*. University of Chicago Press, Chicago, IL
Gilson R, Mnookin R 1985 Sharing among the human capitalists: an economic inquiry into the corporate law firm and how partners split profits. *Stanford Law Review* **37**: 313–92
Geber A 1999 *Juvenhanbuch 1999/2000*, 2nd edn. Frankfurt, Germany
Gordon R 1988 The independence of lawyers. *Boston University Law Review* **68**: 1
Hadfield G K 2000 The price of law: How the market for lawyers distorts the justice system. *Michigan Law Review* **98**: 953–1006
Heinz J P, Lanmann Eo, Nelson R L, Michelson E 1998 The changing character of lawyer's work: Chicago in 1975 and 1995. *Law and Society Review* **32**: 751
Johnson V R 1991 On shared human capital, promotion tournaments, and exponential law firm growth. *Texas Law Review* **70**: 537
Kronman A T 1993 *The Lost Lawyer: Failing Ideals of the Legal Profession*. Harvard University Press, Cambridge, MA
National Law Journal 1992 Ancillary businesses of the nation's 250 largest law firms. *National Law Journal* **December 21, 1992**: 31
National Law Journal 1999 National law journal's annual listing of the nation's largest law firms. *National Law Journal* **December 13, 1999**: C5–C20
Nelson R L 1988 *Fastners with Power: Social Transformation of the Large Law Firm*. University of California Press, Berkeley, CA
New York State Bar 2000 *Preserving the Core Values of the Legal Profession: The Place of Multidisciplinary Practice in the Law Governing Lawyers*. New York State Bar, Albany, NY
Perez Perdomo R 1988 The Venezuelan legal profession: Lawyers in an inegalitarian society. In: Abel R L, Lewis P S C (eds.) *Lawyers in Society, The Common Law World*. University of California Press, Berkeley, CA, Vol. 2
Sander R H, Williams E D 1989 Why are there so many lawyers? Perspectives on a turbulent market. *Law and Social Inquiry* **14**: 431–79
Smigel E O 1969 *The Wall Street Lawyer: Professional Organizational Man?*, 2nd edn. Indiana University Press, Bloomington, IN and London
Stewart P 1991 Is the siesta over for Spanish lawyers? *International Financial Law Review* **Feb**: 20–2
Swaine R 1946 *The Cravath Firm and It's Predecessors, 1819–1947*. Ad Press, New York
Trubek D, Dezelay Y, Buchanan R, Davis J 1994 Global restructuring and the law: Studies of the internationalization of legal fields and the creation of transnational arenas. *Case Western Reserve Law Review* **44**: 407–98
Van Hoy J 1997 *Franchise Law Firms and the Transformation of Personal Legal Services*. Quorum Press, Westpost, CT
Wilkins D B 1993 Making context count: Regulating lawyers after Kaye, Scholer. *Southern California Law Review* **66**: 1145
Wilkins D B 1999 Partners without power? A preliminary look at black partners in corporate law firms. *Journal of the Institute for the Study of Legal Ethics* **2**: 15
Wilkins D B, Gulati G M 1996 Why are there so few black lawyers in corporate law firms? An institutional analysis. *California Law Review* **84**: 493
Wilkins D B, Gulati G M 1998 Reconceiving the tournament of lawyers: Tracking, seeding, and information control in the internal labor markets of elite law firms. *Virginia Law Review* **84**: 1581–681

D. B. Wilkins

Law: History of its Relation to the Social Sciences

Any attempt to discuss the history of the relationship between law and social science encounters a number of problems of definition. For the purposes of this contribution, it is most important to draw attention to some of the problems which can flow from the range of meanings attached to or associated with the term or idea of law. Deriving from this range of meanings, the variety of projects associated with the idea of a legal science can be problematic. As a result of both of these elements, when an attempt is made to formulate the boundaries of the question about the relation between law and the social sciences, indeterminacy results.

1. *Two Images of Law*

At its simplest, the idea of 'law,' at least in Occidental or even Indo-European culture, conjures up two broad images. One is of legislation or rule making. The other is of adjudication or judgment. Cognate distinctions, for example between law and policy or law and politics, and principled statements of the supposedly modern requirements of a democratic order (for example, the doctrine of the separation of powers), all play with or draw upon this much more

deeply rooted distinction, tension, and ambiguity in the idea of law. And even this primordial distinction becomes more complicated with the emergence of a self-thematized modern era, in which the idea of law as rule-making falls apart or splits into two broad patterns. On the one hand, the idea of making highly general, abstract, decontextualized rules which establish the framework to govern the operations of society (the so-called European codes; for this see Caenegem 1987). On the other hand, the idea, which persists in Anglo–American or 'common law' jurisdictions, that rule-making is a matter of detailed, specific rule-making targeting specific problems which come to the attention of the legislator, whether, in the traditional model, through deficiencies brought to light in the course of adjudication, or, in the modern model, because of problems demanding policy adjustments which come to light in, for example, governmental bureaucracies through the growth in knowledge of the shapes and characteristics of society and its pathologies. Some have argued that, at least in the common law world, the increasing volume and significance of legislation so overwhelms our idea of law that it comes to be projected back on to the activity of adjudication, which comes to be understood as a kind of inferior lawmaking, rather than a distinct process with its own specific forms of intelligibility (Simmonds 1984). Since Bentham (and to some degree Montesquieu before him), the idea of a link between a science of legislation and a science of society has seemed obvious. More obscure to social science is the idea of adjudication as such a science. Paradoxically, perhaps, the view has long been generated from within the discipline of law that it is in the domain of adjudication or adjudication-related activities that a knowledge of society is both required and developed, and from this perspective acts of legislation are seen as arbitrary or at least as the product of political will rather than an understanding of society. This kind of lawyer-knows-best attitude carried over, especially in England, into the new encounter between adjudicators and social scientists, especially in the sphere of criminal justice and criminology.

2. *Law and Social Science*

The reason why it is necessary to highlight these divergent emphases is simple: there persists an enthusiasm in many quarters, and by no means just among lawyers, to thematize the idea that law is the oldest social science, that law was and/or remains a 'science of society,' and, in some versions, that the modern social sciences represent merely a continuation of this science or in some cases a derogation from it, a diminution in what society knows about itself.

The question of the relation between law and social science and the history of this relation is further confused or complicated by many of the uncertainties involved in the specification of what is or is not within the ambit of social science. Confining this enormous issue to a narrow focus for our present purpose, we need to separate three kinds of question. The first is the relation between law and the empirical or positivistic branches of the social sciences. The second is its relation to the interpretive and/or phenomenological branches of the social sciences. The third, which cannot neatly be separated from the second, is the relationship between law—or, to switch gear, legal theory—and social theory, which becomes relevant if we assume that there is some kind of connection or link between the conceptualizations of social theory and the problematizations which emerge in the diverse fields of the empirical social sciences.

It is perhaps unsurprising that the closest affinities and even influences between law and social science seem to emerge in the domain of social theory. It is increasingly acknowledged that the writings of some of the founding fathers of social theory, and, indirectly perhaps, of sociology, had a legal background and that their theoretical constructions and projects bear the imprint of legalistic or law-inspired concerns. If, however, this is true of both Durkheim and particularly of Weber (Turner and Factor 1994), then this observation does suggest that legal influences and origins could lead late-nineteenth-century aspiring social theorists to formulate very different questions and very different theories of society, assuming that the construction of 'social theory' was what they sought to achieve. A legal background, in other words, did not seem to lead to any particular kind of theoretical orientation. In something of the same spirit, one of the leading post-war texts in legal theory describes itself as an exercise in 'descriptive sociology' (Hart 1961). This focus makes it possible to see strong continuities between law and (the theory underpinning) social science. Some contemporary political philosophers (Rose 1984), lawyer/psychoanalysts (Goodrich 1997), and historians of political thought (Kelley 1990) have not hesitated to drive home the point, with the polemical claim that modern social theory masquerades as different or new, whereas the basic framework of Western social thought is already contained in the law and indeed that the tradition of law already contains a more sophisticated understanding of the central issues than much modern social theory.

Alternatively, some influential theorists adopt what is more or less the opposite position (e.g., Foucault 1976, Luhmann 1981). On this view, there is a sharp break between an era of social thought in which law provides the basic conceptual scheme for understanding society, and the modern or late modern era in which this vision of society is undermined or displaced by the growth of positivistic social sciences. Hacking (1990) encapsulates what is perhaps the central dimension which underlies this view of historical change

in contrasting the role played by the concept of the 'normal' in the projects of Durkheim and Galton respectively. Durkheim's notion of the normal is still legalistic; in Galton it becomes entirely probabilistic. And in his view it is Galton's version in the twentieth century which wins—in science, at least.

The second area is the relation between law and the interpretive social sciences. Here the connections are in one sense much clearer and visible, especially if we have the 'European' law-as-code model of law in mind. At first sight this seems to be a terrain on which law, history, and interpretive social science meet in the sense that they are engaged in the same kind of endeavor. This is because each is engaged in an exercise of interpretation, and the paradigmatic interpretive situation is the 'presence' of a text which is in some sense 'other' to the person required to interpret it. In hermeneutical science, for example, the enterprise of interpretation from this starting point is characterized as precisely the bringing together, via the text of author/creator and reader/interpreter and as the quest for a fusion of horizons, a merging of perspectives which endeavors not to transcend temporality (as does analytical philosophy) but to make it work, productively, at both ends of its spectrum. It is apparent, viewed historically, that this style of science (whether it is 'scientific' is not a question for this contribution) owes much to law and just as much to theology, and to the study of texts in each discipline in the universities of medieval Europe (Smalley 1983). Already in these disciplines interpretive protocols had been formalized and their shapes were largely carried forward into the emergence of more explicitly philosophical projects by the eighteenth century. Many have traced an unbroken genealogy from that point, in the writings of, for example, Boekh and later Schliermacher through to Gadamer (1965) and contemporary practitioners of this approach to 'truth' (Rorty 1979). Much was made of 'meaning' in these narratives of the history of a partial discipline-under-construction, and therefore in many versions Weber or Verstehenssociologie were annexed to the narrative as part of the same story. 'From the actor's point of view,' a slogan variously popularised (at least within the academic world) by Malinowski and Geertz (1973, 1988), could lay claim to being the heir to this perspective or 'methodology' and claims that this involved a direct succession to the lawyer's perspective that, for example, '*actus non sit reus nisi mens sit rea*' (the doctrine of *mens rea* which seemingly involves investigation of inner mental states and 'intentionality' and its degrees) possessed and still possess a certain plausibility.

But it is precisely at this point in the history of law and the social sciences that continuities may need to be balanced by discontinuities. First, the once innovative proposition that social science can proceed on the basis that culture is a text has been much debated and criticized (e.g., Clifford and Marcus 1986). Secondly, political and legal anthropology in particular has raised a number of questions about the tendency in earlier studies to posit a 'jural perspective' when investigating acephalous societies, which presented a particular challenge because no visible structures of government or law in the familiar western sense were available for outside observers to observe (which is not to say that they could not find them virtually present) (Comaroff and Roberts 1981). Under the influence in part of some of this work, some historians came to question the top-down model of social integration and/or development which has had a habit of permeating social as well as political history.

The net result of these overlapping concerns has been to move away from or be suspicious of the deployment of a 'juridical' model in the interpretive social sciences, and this has served to highlight the difference in the concerns or preoccupations of lawyers and social scientists. For lawyers, the key question is what beliefs, assumptions, values, intentions it is reasonable or rational to impute to actors or legal subjects—therein lies the craft, one might say, of the law, and there is nothing new about this. Social scientists making use of interpretive procedures are increasingly concerned to understand or uncover the logic of individual and social action, including motivations, aspirations, intentions, and use concepts like culture, horizon, context, situation, to help render these intelligible. These social scientific ambitions are not new, but a combination of background factors (postcolonialism, for example) and increasing sophistication in formulating these components, means that interpretation has lost what innocence it ever possessed, and that the kinds of protocols employed by lawyers whenever they need to concern themselves explicitly with actors rather than texts are increasingly divergent from the highly reflexive methods now required of social scientists. Indeed, legal self-understandings have come to look to these interpretive disciplines for some corrective to what are increasingly perceived to be deficiencies in epistemic capacity and practical deliverability within the legal world itself. The sharpest symptom of this (though appearances can sometimes deceive) is the rapid rise to prominence in policy-making, especially in the Anglo-American common law world, regarding the adjudicative process of Alternative Dispute Resolution (ADR), an approach to conflict resolution which has unambiguous roots in 1950s colonial and post-colonial anthropology. One example can perhaps serve as a sharp reminder of how far apart in some respects the preoccupations of the law and of modern interpretive social sciences have become. For the law, an axiom of prudence and indeed an operational condition of possibility has more or less for ever been the postulate that '*ignorantia juris haud excusat*.' Contrast this with the role which '*docta ignorantia*' plays in Bourdieu's not-quite-Marxist sociology of domination (Bourdieu 1980). Axiomatically the law disregards what people claim they do not know (based on the suspicion that

such disclaimers are easy to make and dubious in the first place—a protocol rooted in scepticism, in other words). By contrast, Bourdieu perhaps leads the contemporary field in focusing on a modern version of Marxist false consciousness, detached from economic determinism as such, but emphasising the prison-house of cultural schemes which prevent people from knowing what they do. What for the law is something to be ignored is for the social sciences something to be explored. For the law ignorance is sterile; for social sciences it is productive.

3. *Law and the Empirical Social Sciences*

What is surely unambiguous is the relationship between law and the development of the empirical social sciences, especially to the extent that these sciences are underpinned by the burgeoning of statistical knowledge and growing sophistication in statistics-gathering methods and techniques. Yet this contains within itself new ambiguities. Specifically, at least at the level of images of truth and the supporting rhetoric or legitimation strategies available to science, it is noticeable how a number of writers from the side of science—including natural science—seek to invoke the image of the law court or some of its components—notably the jury—as the central basis for the image of truth or of what it means to claim validity for scientific statements. Trial by your peers has much ideological resonance, especially in North America; but whether it is more than a cultural theme or idiom is a more intricate question which requires explicit if not detailed discussion here.

The metaphor of scientific truth as something resolved in a court of appeal in which the judges or jury decide following adversarial confrontation is a powerful and longstanding one in Western thought, albeit one which is by no means universal (see Lloyd 1996, Needham 1969). While it is not the same idea, it dovetails with the notion of truth as a matter of gentlemen's agreement (Shapin and Schaffer 1985, Shapin 1994) which in turn rests and relies upon the presence of implicit paradigms which are not themselves usually open to inspection or discussion (Kuhn 1970, Foucault 1966). Yet all of this perhaps amounts to little more than the recognition that truth is contestable and the claim that scientific practice has developed—or has always possessed—procedures for resolving epistemic conflicts which arise in the natural course of scientific discovery and invention.

These metaphors of judgment (whether by judges or juries) draw upon the experience of the world of law but do not in reality replicate it. It is interesting to note that they appeal rather more to the common law adversarial style of legal argumentation in the courtroom than the so-called inquisitorial style of the European codified tradition. But whatever precise analogy is used, it runs the risk of obscuring the important and significant gulf which opens up in the modern world between legal science—or the practice of adjudication—and empirical social science.

The most significant here is revealed in the distinction between the level of the individual (for lawyers, a 'subject,' for many social scientists, an 'actor') and the aggregate or collective ('society'). The law has been and remains concerned with disputes between individuals and with the attribution of liability to individuals and with the language of individual responsibility. All of this requires a focus, much of the time, on individual mental states, on postulates about what these can be taken to be or how they can be sensibly understood—or on appeals to 'ordinary people' to use their own experience of life to decide what they believe or disbelieve in the claims about such matters made by the parties in court. Judges operate within a similar framework, although they bring to their decision making the added rigmarole of professional training and/or experience of previous kinds of problems or disputes in a forensic setting. In other words, legal schemes combine two types of individualism—the individual as decision maker relying on his or her experience and capacities to draw analogies from this reservoir of experience in order to make decisions, and the individual as the object of decision, as a person who can be assumed to be like or not like the person asked to make a judgment, in matters relating to responsibility for events or actions. The law therefore struggles with supra-individual entities like corporations, states, nations, cultures. Can a corporation be made liable for manslaughter or culpable homicide? This individualism of the law extends beyond persons and into events. A typical legal scenario, for example in the field of tax avoidance, is the problem of deciding whether a chain of events is to be understood as a multiplicity of individual events or as a composite ('quasi-individual') event. For lawyers, these are not simple questions.

Empirical social science, by contrast, is preoccupied, at its most subtle, with the often perplexing relationship between 'micromotives' and 'macrobehavior' (Schelling 1978). Correspondingly, individualization of events is of much less if any interest than patterns, tendencies, logics, trends. Critical to these perspectives are perhaps two modern capacities—measuring and modeling. And underpinning these new epistemic possibilities are modern notions of probability and of statistical thinking which diverge sharply from the traditional idea of probability still largely at work in the ideational schemes of the law (Hacking 1975, Porter 1986, Gigerenzer et al. 1989, Murphy 1997).

These developments challenge legal schemes of reality. They create a reality through practices and procedures which are remote from the mechanisms through which law constructs 'its' truth. Surveys, data analysis and number-crunching, the quantification of the qualitative (and the very emergence of an explicit

distinction between these two and the consequent problematization of the relationship between them), probabilistic judgment—all these epistemic practices are quite different from the argumentative, judgmental, and commonsensical manner in which law decides what is true. Even before this became obvious, there were pressures upon law to acknowledge what was changing and to incorporate into itself the findings or the protocols of social science (most famously in the so-called 'Brandeis Brief' in the USA, in which statistical information was introduced into court as part of the case). In modern US and UK law for example, a concept of indirect discrimination has been introduced, which makes it unlawful to discriminate not just intentionally but by applying to, for example, job applicants a requirement or condition with which members of the protected group—principally race or sex but these definitions can be expected to expand over time through new legislation—are disproportionately unable or less able to comply. This kind of formulation inevitably invites the introduction of statistical data into the forensic setting. Other outputs of modern social and behavioral sciences increasingly find their way into such settings, notably in the criminal law and family law. Social inquiry reports for offenders before sentencing, or all manner of inputs on child welfare, have come to play a role in the way cases are disposed of in the areas of sentencing and child custody disputes. It is also increasingly the case that the output of the social sciences plays a role in the evaluation of current law, and generates critique, proposals for new legislation, and arguments for changes in the outcomes of judicial decision making.

These kinds of developments lead some observers and commentators to discern an historical process at work in which the legal sphere steadily succumbs to the influences of the social and behavioral sciences. Foucault's thesis about normalization and disciplinary power (Foucault 1975), for example, can be read in this way (see also Donzelot 1977). Others stress rather the rise of differentiation. Among many legal theorists if not social theorists, one of the most influential has been the function-systems theory of Niklas Luhmann (1984). At its simplest, what is at issue here is whether the 'worlds' of law and of social science are converging or merging or pulling further apart. The answer is probably that both tendencies are simultaneously visible. Convergence of course now means the 'colonization' of law by social science, and this is most likely to occur in the area of individual and especially cognitive psychology. In the future, we can anticipate that genetics will play a greater role here. But pulling in the opposite direction are strong tendencies towards the assertion of legal autonomy, in the international as well as national and sub-national spheres. As often in the previous history of law, legal thought often proceeds ahead of successful institutionalization, while without the latter the former remains stuck in the ebb and flow of opinion and fashion. However, to the extent that this tendency towards autonomy is becoming a 'reality', it potentially means that the pre-social-scientific epistemic schemes of the law will succeed in extending their life and even consolidating their position into the future, disregarding the findings of social science on the grounds of increasingly abstract principle.

See also: Critical Legal Studies; Hermeneutics, History of; Law and Society: Sociolegal Studies; Law: Anthropological Aspects; Law as an Instrument of Social Change; Law: Overview; Lawyers; Legal Process and Social Science: United States; Paradigms in the Social Sciences; Social Science, the Idea of

Bibliography

Bourdieu P 1980 *Le Sens Pratique*. Les Editions de Minuit, Paris. [1990 *The Logic of Practice*. Polity, Cambridge, UK]

Caenegem R C van 1987 *Judges, Legislators and Professors: Chapters in European Legal History*. Cambridge University Press, Cambridge, UK

Clifford J, Marcus G E (eds.) 1986 *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA

Comaroff, J, Roberts S 1981 *Rules and Processes: The Cultural Logic of Disputes in an African Context*. Chicago University Press, Chicago

Donzelot J 1977 *La Police des Familles*. Les Editions de Minuit, Paris [1979 *The Policing of Families: Welfare Versus the State*. Hutchinson, London]

Foucault M 1966 *Les Mots et les Choses: Une Archéologie des Sciences Humaines*. Editions Gallimard, Paris [1970 *The Order of Things: An Archaeology of the Human Sciences*. Tavistock, London]

Foucault M 1975 *Surveiller et Punir: Naissance de la Prison*. Editions Gallimard, Paris [1977 *Discipline and Punish: The Birth of the Prison*. Pantheon, New York

Foucault M 1976 *Histoire de la Sexualité i: La Volonté de Savoir*. Editions Gallimard, Paris [1979 *The History of Sexuality Vol I: An Introduction*. Allen Lane, Harmondsworth, UK]

Gadamer H-G 1965 *Wahrheit und Methode*, 2nd edn. J C B Mohr, Tünbingen [1979 *Truth and Method*, 2nd edn. Sheed and Ward, London]

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Geertz C 1988 *Works and Lives: The Anthropologist as Author*. Stanford University Press, Stanford, CA

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L 1989 *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, Cambridge, UK

Goodrich P (ed.) 1997 *Law and the Unconscious: A Legendre Reader*. St. Martin's Press, New York

Hacking I 1975 *The Emergence of Probability*. Cambridge University Press, Cambridge, UK

Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK

Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford

Kelley D R 1990 *The Human Measure: Social Thought in the Western Legal Tradition*. Harvard University Press, Cambridge, MA

Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago
Lloyd G E R 1996 *Adversaries and Authorities: Investigations into Ancient Greek and Chinese Science*. Cambridge University Press, Cambridge, UK
Luhmann N 1981 *Politische Theorie im Wohlfahrtsstaat*. Olzog, Munich, Germany [1990 *Political Theory in the Welfare State*. Walter de Gruyter, Berlin]
Luhmann N 1984 *Soziale Systeme: Grundriss Einer Allgemeinen Theorie*. Suhrkamp, Frankfurt am Main, Germany [1995 *Social Systems*, Stanford University Press, Stanford CA]
Murphy W T 1997 *The Oldest Social Science? Configurations of Law and Modernity*. Oxford University Press, Oxford
Needham J 1969 *The Great Titration: Science and Society in East and West*. Allen and Unwin, London
Porter T M 1986 *The Rise of Statistical Thinking 1820–1900*. Princeton University Press, Princeton, NJ
Rorty R 1979 *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ
Rose G 1984 *Dialectic of Nihilism: Post-Structuralism and Law*. Blackwell, Oxford, UK
Schelling T C 1978 *Micromotives and Macrobehavior*. Norton, New York
Shapin S 1994 *A Social History of Truth: Civility and Science in Seventeenth-century England*. Chicago University Press, Chicago
Shapin S, Schaffer S 1985 *Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life*. Princeton University Press, Princeton, NJ
Simmonds N E 1984 *The Decline of Juridical Reason: Doctrine and Theory in the Legal Order*. Manchester University Press, Manchester, UK
Smalley B 1983 *The Study of the Bible in the Middle Ages*, 3rd edn. Blackwell, Oxford
Turner S P, Factor R A 1994 *Max Weber: The Lawyer as Social Thinker*. Routledge, London

W. T. Murphy

Law in Latin America

Latin America is usually taken to comprise the 20 independent countries in America that were Spanish, Portuguese, or French colonies before the nineteenth century. (The countries are Argentina, Brazil, Bolivia, Chile, Columbia, Costa Rica, Cuba, Ecuador, El Salvador, Guatemala, Haiti, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Dominican Republic, Uruguay, and Venezuela.) It is doubtful that other former Latin colonies now integrated in Canada and the United States (Quebec, Puerto Rico, Louisiana) should be considered part of Latin American legal culture. This article concentrates on the former Spanish colonies and Brazil.

There is no such a thing as Latin American law, since the nations each have completely independent legal systems. The issues to be discussed here are whether there are general features of these legal systems and if their legal culture could be distinguished from that of the colonizing countries.

Latin America is usually placed in the civil law, or Roman law, tradition. The main reason for this inclusion is the colonial link to the Latin countries of Europe. In fact, the way law is taught and practiced, as well as the place of judges and legal scholars in society, are quite similar in Latin Europe and Latin America. But there are significant differences due mostly to the more heterogeneous character of Latin American societies. The history of Latin American countries is usually divided into the pre-Columbian, the colonial, and the national periods. The pre-Columbian goes up to the sixteenth century; the colonial, from the sixteenth to the early nineteenth centuries, and the national covers the nineteenth and twentieth centuries. The pre-Columbian period can hardly be called Latin, but is included because it is a part of the history or prehistory of each country. The Spanish kings decreed that those indigenous customs not contrary to the Catholic faith should be applied to their Indians subjects and they should keep their 'principals.' In fact, very little pre-Columbian political organization and law survived the colonization.

During the colonial period, Castilian law was applicable in the Spanish dominions to the White and Mestizo population. There were special royal regulations that applied only to the colonies—the so-called the *Indiano* law. The main official compilation is the *Recopilación de Leyes de Indias* (1680). There was not an equivalent compilation for Brazil. Modern historians prefer to call *derecho indiano* the law as it was applied in the colonies. The legislation tended to be uniform for all Spanish colonies, as it was for the different parts of Brazil. The structure of colonial administration (King—colonial council—viceroys—*audiencias*) and the mechanisms of administrative and judicial control (*visita* and *residencia*) were similar for Spain and Portugal.

Universities and law schools were established from the early sixteenth century onwards in several Spanish American cities (Santo Domingo, 1538; Mexico, 1551; Lima, 1551; Quito, 1586; Cuzco, 1598). By the early nineteenth century there were law schools in the main cities and several thousand lawyers were part of the political elite in the different countries. Portugal followed a different path and no university or law school was established in colonial Brazil, but wealthy colonists educated their scions in Coimbra. Two were established in 1826, shortly after independence. By 1830 there were near 500 *bachareis em direito* in the country.

Lawyers were extremely important during the independence period in the early nineteenth century. They drafted constitutions and were conspicuous members of congresses and cabinets. Their political and social importance is parallel to the proliferation of constitutions in the early nineteenth century. The Spanish and Portuguese legislation survived independence, but there were strong movements for a national law that usually led to national codes in the second

part of nineteenth century. In fact, the codes are quite similar to the European codes. A scholarly tradition dating from the colonial period in the most important capitals extended to most countries during the nineteenth and twentieth centuries.

The number of lawyers grew slowly in the nineteenth and early twentieth centuries, but they kept a high political profile. By mid-twentieth century the growth pace increased substantially due to the generalization of higher education and increased legal activity related to the economic growth. The codes became relatively less important due to the proliferation of regulatory legislation. A substantial effort of judicial reform and an increased political and social importance of judges characterize the late twentieth century. Since the mid-1980s there have been important deregulation efforts.

Latin American legal culture has been characterized in many different ways. Legalism, formalism and, at the same time, lack of penetration of the formal legal system, are usually noted (for example, Karst 1965). Many Latin American countries have an important native population, mostly rural, and in all of them the urban population has a high proportion of poor and very poor people. Cultural and economic factors lead them to avoid the formal legal system and have serious difficulties when they encounter it. This is the root of the lack of penetration of the formal legal system, or more appropriately, of legal pluralism in Latin America. This distinguishes the Latin America legal culture from that of Latin Europe.

The importance of law and lawyers has coexisted with dictatorships, local *caudillos*, and political turbulence. The heterogeneous character of society and the fear of disorder or anarchy are at the root of strong men government and the sometimes massive violations of human rights. The primary difference between Latin American and European dictators is that the former have paid more attention to constitutional formalities. They usually change the constitution to accommodate to their needs. Nevertheless, the tendency in the last 25 years of the twentieth century has been for democratization and rule of law. This new trend is closely related to the increased importance of judges and judicial reform efforts.

See also: African Legal Systems; Colonization and Colonialism, History of; Folk, Indigenous, and Customary Law; Latin American Studies: Politics; Latin American Studies: Society; Law and Society: Socio-legal Studies; Law: Anthropological Aspects

Bibliography

Bastos A W 1998 *O ensino juridico no Brasil*. Lumen Juris, Rio de Janeiro

Guzman Brito A G 2000 *La codificación civil en Iberoamérica. Siglos XIX y XX*. Editorial Jurídica de Chile, Santiago

Junqueira E B 1993 *A sociologia do direito no Brasil*. Lumen Juris, Rio de Janeiro

Karst K 1965 *Latin America Legal Institutions. Problems for Comparative Study*. University of California Press, Los Angeles

López Ayllón S 1997 *Las transformaciones del sistema jurídico y los significados sociales del derecho en México: la encrucijada entre tradición y modernidad*. Universidad Nacional Autónoma de México, México

Merryman J H 1985 *The Civil Law Tradition*, 2nd edn. Stanford University Press, Stanford, CA

Merryman J H, Clark D S 1978 *Comparative Law: Western European and Latin American Legal Systems*. Bobbs-Merrill, Indianapolis, IN

Pérez-Perdomo R 1981 *Los abogados en Venezuela*. Monte Avila, Caracas

Tau Anzoátegui V 1997 *Nuevos horizontes en el estudio histórico del derecho indiano*. Instituto de Investigaciones de Historia del Derecho, Buenos Aires

de Trazegnies F 1979 *La idea del derecho en el Perú republicano del siglo XIX*. Pontificia Universidad Católica, Lima

R. Pérez-Perdomo

Law: Imposition, Reception, and Colonial

Conceptually, legal imposition is set against the demotic acceptance or creation of law. The colonial situation is usually seen as the extreme of legal imposition, a situation where an alien law is brought to bear on a subjugated people without their consent. Reception is commonly and sharply distinguished from that situation. Although the law received is still alien, it is accepted voluntarily by the people to whom it applies—something of a self-imposition. There is indeed a sustainable truth to this straightforward distinction between the voluntary and the involuntary application of law, but that truth has to be complicated in two respects: one concerns an ambivalence in the relation between law and the demotic, and the other responds to the variety of colonizations and receptions. Each complication will now be dealt with in turn and, in the process, legal imposition, reception and colonial law will be refined and given content.

1. Imposition

All law is imposed, but never entirely so. Law's operative mode is the decision—the decision of the judge, the legislator, or the legal subject. The decision cannot be reduced to, or rendered merely in terms of, some prior determination or mandate, demotic or otherwise. If it could, there would be no 'call' for the legal decision, no demand for 'fresh judgment' (Derrida 1992, p. 26). Law, then, always comes from beyond any existent situation, and is an imposition from without. In this movement from beyond into 'fresh' determination, however, the effective legal decision cannot be imposed purely. If law is not to be

a 'dead letter,' it has to assume an engaged responsiveness to the demotic and to anything else involved in its making or application.

This is the point at which criticism of the idea of imposition usually enters. Summarily, this is a criticism of the idea of 'imposition as a fait accompli rather than an interactional process affected by power differentials' (Kidder 1979, p. 291). The criticism could be said to take a weak and a strong form. The weak variant has it that imposition is an inadequate conception, because indigenous legal modes persist and are even necessarily recognized by the would-be imposer. The system of so-called indirect rule in British colonies provides one of the more explicit types of such recognition. Imposition can, however, be more nuanced. This very 'recognition' of indigenous law will often transform that law in ways conformable to the exogenous imposition (Fitzpatrick 1984).

This penetrative force of imposed law is countered in the second, and strong, criticism of the idea of imposition. Such a criticism would extend the efficacy of the indigenous into the very domains of imposition. Here the epistemic arrogation of imposed law is 'got around,' and such law operatively subordinated to indigenous contexts and purposes (Fitzpatrick 1989). This process is facilitated by what is often a rigidity, an unresponsiveness in imposed law, something which particularly affects modern colonizations with their myopic claims to a completeness or, at least, an exemplarity of being. The premonitory quality of this myopia and the limits of simple imposition can be discerned in a hiatus vexing the colonial governor of Bombay in the middle of the nineteenth century when he remarked on 'the perilous experiment of continuing to legislate for millions of people, with few means of knowing except by rebellion, whether the laws suited them or not' (see Thornton 1965, p. 181).

This ambivalence of law's imposition—the ambivalence between its determinant fixity and its responsiveness—will now be returned to the initial distinction between the voluntary and the involuntary imposition of law, and both the ambivalence and the distinction will be used to provide instances and some ordering of the varieties of legal reception and of colonial law. At the risk of being over-schematic, the combination of these two oppositions can provide four types of reception/colonial law: the involuntarily determinant, the voluntarily determinant, the involuntarily responsive, and the voluntarily responsive. Each of these will now be illustrated in turn.

2. *Colonial Law*

It would be difficult to imagine a scene more adapted to the involuntarily determinant type than that offered by the modern colonial situation. To the imperial eye, law was pre-eminent among the gifts of an expansive civilization, one which could extend in its abounding generosity to the entire globe. The supreme justification of colonial law in its 'civilizing mission' was that it subjected the colonized to a new and beneficent order. It was solely the colonist who was to provide civil and civilized order. There were no rights for the colonized in this scheme, apart from 'rights' to have things done to them so as to bring them within the ambit of civilization (Thornton 1965, p. 158). To sustain this scheme of things, however, the colonized were also labeled and maintained as intractably apart from that civilization. They had to become the same as the civilized colonists yet remain unalterably different from them. That unsettled ambiguity permeated the whole of colonial law. The colonist took on the 'burden' of pervasive legal powers in the cause of an inclusive civilization, only to use them to exclude, dissipate, and generally 'hold down' the colonized as incorrigibly deviant. Despite this, some effective action had to be allowed to the colonized because legal imposition could not be simply or totally determinate, an inhibition heightened by what was usually called the limited penetration of colonial rule. Various systems of 'indirect rule' and of 'recognition' of native modes necessarily proliferated. Nonetheless, colonial law always sought to circumscribe effective action on the part of the colonized. Not only was it characterized as static, repetitive, and mimetic but, for good measure, such action was held or attempted to be held within a supervisory system of administration. Indigenous custom, for example, was 'recognized' solely in subordination to the law of the colonists and denied recognition where it was 'repugnant to natural justice, equity and good conscience,' or contrary to 'the general principles of humanity,' to take two standard and indicative formulations.

The legal mechanics of this colonial type of imposition both underline its monadic quality and set a significant distinction. Here one must accommodate, in very broad brush, a variety of the laws of colonizing nations and an immensely convenient international law, so-called, which reflected their practices and interests (see, for example, Roberts-Wray 1966, Chaps. 2, 11). All the legal modes of acquiring territory and imposing laws upon it were founded on the distinction between the civilized and the uncivilized, or on something akin to that distinction. According to what was, with an indulgent exactitude, called the 'doctrine' of discovery, a 'European' nation could lay claim to territory it had 'discovered' and have its laws applied there, only if that territory were uninhabited or considered insufficiently settled, and it was rare for any uncivilized settlement to be taken as sufficient. With territorial acquisition by conquest or cession, if the territory belonged to a European country, generally the law of that country continued in force, but the position was usually otherwise if the territory was occupied by the uncivilized. When territory was acquired by way of colonial settlement, the colonists were most often taken to have brought their law with

them at the expense or subordination of the law or custom of the colonized. It was the case, however, that with all these modes of acquiring territory and of imposing law, there was often a vague requirement that the metropolitan law would not apply where it was not suited to local conditions. One important area of unsuitability corresponded to the civil rights frequently found in metropolitan constitutions. These were only very rarely extended to the colonized.

3. *Reception*

If the colonial situation exemplifies the involuntarily determinant, the first in the quartet of law's imposition, then the reception of law is usually taken to exemplify the second in the quartet, the voluntarily determinant. This seeming self-imposition of law is associated typically with the European reception of Roman Law. 'The Reception,' as its chroniclers often put it, was a long and many-stranded process (Ibbetson and Lewis 1994, Stein 1999). This reception is tied conventionally to the new learning, which in most accounts comes onto the scene from the late eleventh century, eventually forming one of a momentous trio of 'Renaissance, Reformation and Reception.' This company perhaps renders reception more historionic than it was predominantly. It did play a large part in telling political conflicts but, in the main, its mode was one of quiet infiltration into existing legal processes. That infiltration was fueled, however, by a spectacular increase in the number of lawyers educated at the great medieval universities, most notably Bologna. Despite its operative adaptability, this was a reception of a collection of laws temporally remote and contrary to much of the then existing law—a reception of the Digest, Code, and Institutes of Justinian. This distanced quality of the reception was softened by a massive accumulation of scholarly annotations of these texts which, although purportedly exegetical, were not without their contemporary connections. And such connections also qualify the voluntary quality of this reception. Roman Law was responsive to new or emergent economic and political interests, and it would not have been uniformly acceptable. This supranational impact of Roman Law terminates with the national codifications of the late eighteenth and the nineteenth centuries, but these codes themselves exhibit some commonality influenced by Roman Law, a commonality reflected in the designation of 'civil law countries.'

4. *Variations*

It is these codes which have continued the history of this type of reception most conspicuously. Perhaps the most noted example comes from the adoption by Turkey in the nineteenth century, and after the declaration of the Republic in the twentieth century, of a range of codes from a variety of European countries (Güriz 1978). This was part of a project of full-scale modernization and, in line with this, these laws were received as determinant. Another notable example is provided by the purposively modernizing Meiji Constitution of 1889, in which Japan drew greatly on the constitutions of several European states (Fujikura 1996). This type of reception has spread widely of late, through what has most compendiously been described as 'structural adjustment' or 'conditionality,' modes of reception which can be described with some paradox as involuntarily responsive. This is a putatively contractural mode of domination in which the requirement to enact or enforce certain laws is a condition attached by the International Monetary Fund or by the World Bank to loans made to nation states. Formally, these institutions operate as banks, and the relevant conditions are attached to loans to enable countries to deal with financial exigencies, or to assist in their development. More recently, conditions have also become attached to qualifying for the relief of debt. These conditions, which have by now assumed enormous and seemingly enduring significance, usually entail the promoting of the rule of law and the 'transplanting' of occidental laws, especially those of a commercial variety. The unabashed aim in relation to the 'new nations' is 'to reform the legal system and make it more like those in the core countries' (Trubek et al. 1994, pp. 481–2).

Doubtless it should not be overdrawn, but the contrast between such 'conditionality' and the earlier reception of law under the labels 'law and development' or 'law and modernization' is quite marked, and can serve to introduce the final member of the quartet, the voluntarily responsive. Although this earlier transplantation of laws to the 'new nations' could be accompanied by a considerable degree of outright imposition, it did nonetheless have some regard to indigenous elements and it did aim, ostensibly at least, to impart or enhance their own capacity to achieve development or modernization. Still, the responsive element here was usually quite attenuated and this particular type of reception is much more exuberantly present at the beginning of the twenty-first century in 'Eastern Europe,' in East Asia, and in China, even if this presence is not always without an admixture of the dependent compliance which has to accompany conditionality. This is sometimes a matter of the revival of earlier borrowings from the codes of Western Europe, but with the new transfers Anglo-American law does tend to predominate. In this situation, it is not only the case that the imported law responds to what elites in these societies want, but there is usually some effective adaptation of the laws by these elites in the process of their reception (Sharlet 1998).

Sweeping as this conspectus has been, it has only touched on some significant aspects and instantiations

of legal imposition, reception, and colonial law. These legal phenomena, perhaps because of the focus in modern legal studies on the singular nation, have not been given the attention which their enormous impact and influence would suggest they deserve. Being graphically situated between the voluntary and involuntary purchase of law, and being situated between law's determinant and responsive qualities, these phenomena raise fundamental questions about the constitution of law which have for too long been obscured within the national domain.

See also: Colonialism: Political Aspects; Colonization and Colonialism, History of; Japan: Sociocultural Aspects; Legal Systems, Classification of; Postcolonial Law

Bibliography

Derrida J 1990 Force of law: The 'mystical foundations of authority'. [Trans. Quaintance]. In: Cornell D, Rosenfeld M, Carlson D G (eds.) *Deconstruction and the Possibility of Justice*. Cardozo Law Review, New York

Fitzpatrick P 1984 Traditionalism and traditional law. *Journal of African Law* **28**: 20–7

Fitzpatrick P 1989 Crime as resistance: The colonial situation. *The Howard Journal* **28**: 272–81

Fujikura K (ed.) 1996 *Japanese Law and Legal Theory*. New York University Press, NY

Güriz A 1987 Sources of Turkish law. In: Ansay T, Wallace B (eds.) *Introduction to Turkish Law*. Oceana Publications, New York

Ibbetson D, Lewis A 1994 The Roman law tradition. In: Lewis A, Ibbetson D (eds.) *The Roman Law Tradition*. Cambridge University Press, Cambridge, UK

Kidder R L 1979 Toward an integrated theory of imposed law. In: Burman S B, Harrell-Bond B (eds.) *The Imposition of Law*. Academic Press, New York

Roberts-Wray K 1966 *Commonwealth and Colonial Law*. Stevens and Sons, London

Sharlet R 1998 Legal transplants and political mutations: The reception of constitutional law in Russia and the newly independent states. *East European Constitutional Review* **7**: 59–68

Stein P 1999 *Roman Law in European History*. Cambridge University Press, Cambridge, UK

Thornton A P 1965 *Doctrines of Imperialism*. John Wiley, New York

Trubek D M, Dezalay Y, Buchanan R, Davis J R 1994 Global restructuring and the law: Studies of the internationalization of legal fields and the creation of transnational arenas. *Case Western Reserve Law Review* **44**: 2

P. Fitzpatrick

Law, Mobilization of

Theories of mobilizing the law try to explain under which conditions problems come to be defined and socially accepted as legal issues, and under which conditions they are taken to lawyers and courts. They look at the process of (a) moving from perception to motivation, (b) the interaction of parties and their strategies, and (c) finally at courts as institutions which define the frontiers of law.

1. Access to Courts as a Process

'Naming, blaming, and claiming' are the catchwords by which Felstiner et al. (1981) describe the steps of framing social conflicts in legal terms. They discuss how experiences come to be felt as injurious (naming), how a party might be singled out as responsible (blaming), and under which conditions claims might be formulated on the basis of these experiences. They turn around the assumption of market researchers who identified 'unmet legal needs' (Curran 1977, ABA 1994) by surveying people for injurious experiences which in the minds of lawyers would give rise to a legal claim. In fact, claiming does not automatically follow upon naming and blaming (Genn 1999). A representative British survey documents the various strategies of self-help people pursue before mobilizing the the law (Blenn 1999). It shows the importance of informal access to law for which Britain with its citizens' advice bureaux and legal aid provisions offers a wide range. People make choices: the greater the interest in a continuous social relation, the higher social costs would be. In a survey of everyday injurious situations which people encounter in a German modern city (such as Frankfurt), Hanak et al. (1989) show that only a tiny fraction of all potential legal disputes actually reach lawyers or appear in court. 'Naming' and blaming are everyday reactions to disappointment and injury, but 'claiming' is a far rarer phenomenon. To preserve ongoing relations, it can be entirely 'rational' to avoid engaging in legal steps of any kind. Consequently, Hanak et al. declare people who directly frame their social problems in legal terms as 'socially incompetent.'

An empirically well-documented example is given by the selectivity with which people claim compensation for accidents: Harris et al. (1983) sampled 3,600 people in England who had been absent from work for more than two weeks during the previous 12 months: while 60 percent said that they had been seriously ill, 34 percent blamed their absence on an accident, of which 90 percent occurred at home or during their free time, one tenth during work, and every fifth as traffic accidents. The point of the study is that 'blaming' others for causing an accident is largely independent of whether victims proceeded to 'claiming' compensation. Accidents at home or in free time might just as well be caused by others, but they would be seen as mishaps without legal consequences, while accidents at work which might be seen as 'one's own fault' would nevertheless give rise to claims for compensation. The explanation lies in the everyday logic of the social relation: while a claim for injury which remains in the

Table 1

Fault?	Claim compensation?		
	Yes	No	
Yes, the same who has to pay	20%		
		10%	
Yes, somebody else	8%		
No	10%	52%	
			100%

sphere of personal social relations is avoided because it would outweigh the social costs of alienating that relation, a legal claim is easily made if covered by employers or other kinds of insurance. Often the claim consciousness works back into the perceptions: framing in terms of legal liability colors the perception of social fault. See Table 1.

Accidents are easily identifiable as a basis for damage compensation. Other potential sources of liability claims are less obvious: damages in consumer protection, medical malpractice, and tobacco litigation were discovered (and as liability claims 'invented') by pioneering lawyers and consequently advertised to help people climb over the barriers of mobilizing the litigation process. From the point of view of the providers of legal services it becomes a marketing problem that only a tiny fraction of all potential legal disputes actually reach lawyers or appears in court. However, people do not always have a choice but to enter into legally defined handling of situations. Frequently it is the other party who invokes the law. In dealing with them, people mobilize the law defensively rather than proactively (Black 1973), and they are forced to do that the more they are interacting with formal organizations and in anonymous situations. Also, not all social relations are designed for continuity. The more urban and mobile people get, the more they handle incidental and anonymous social relations. After a traffic incident or as clients of a big organization, parties are confronted with predefined ways of dealing with conflicts. Especially formal organizations and government agencies pattern their relations in legal terms to begin with. The bigger they are, the more likely they contract with employees, partners, or clients in written terms, and these in turn are bound by general regulations. Therefore there is no choice but to take legal steps. Black (1973) generalizes these findings into the proposition 'the greater the social distance, the more law.'

Thus, if asking why the frequency of litigation is rising in all modern societies, the explanation has to move from micro-sociology to macro-sociological conditions. A number of trends come together to increase 'judicialization' of many walks of life, be they world-wide such as urbanization, or specific to certain societies and to lower social classes, such as welfare state systems which bind people's life chances into dependency. In other countries and in the middle classes it may not be the state but rather private insurance systems which increase the dependence of formal bureaucracies—the trend remains that the coverage of life risks is less provided by informal care and personal ties but rather contracted and insured with large formal organizations. Culture critics (like Habermas 1981) have called this trend of modernity 'legal terms patterning the modern lifeworld.'

2. *Strategies of Mobilizing the Law: Enforcing, Mediating, Adjudicating*

The perspective of social actors, taken in the above studies, can be contrasted with that of legal institutions. Government, by definition, regulates by statutory law, but if state regulation is compared with the contract terms of big organizations, there appears to be not much difference between being a citizen of the regulative state or a client of organized enterprises. For individual actors, the difference from state regulation lies in the initial choice and the chances for exit which markets might offer, but once chosen many long term contracts turn out to be just as bureaucratic and regulatory, while for multinational actors even the regulatory regimes of states appear as a matter of choice. The likelihood of social relations being defined in legal terms depends on the size of organizations and the degree of their formalization.

However, within a more or less formal relationship actors do differ in the way they bring grievances forward. Disputes are seldom purely adversary; often both parties have at least some interests in common. That is even the case when parties battle furiously as in a divorce suit: typically the parties have at least one goal in common—the desire to separate. In fact, most divorces are settled; in only a minority is there a battle about the terms of separation. The same is true for other constellations where courts are invoked to set the terms of separation. Labor dismissal and landlord–tenant conflicts seldom try to regulate the terms of an ongoing relationship; they typically end with a settlement on how to part. As most countries require a court decision for divorce, parties regularly appear with a settlement negotiated in advance; they need the court to function as quasi-notary. The largest number of

Table 2
Goals and outcomes: functions/goals of law suits

	Outcome predictable	Outcome unpredictable
Parties have some interests in common	Quasinotarial act	Settlement
Interests form zero-sum	Enforcement decree	Adversary decision

Table 3

Naming /Blaming Framing	Claiming		Litigating	Implementing
------->	---------->		---------->	
Social-distance		Capability of handling legal institutions	Burden of proof	Decision to enforce a legal degree
	Burden of invoking courts		Duration	
Formalization–of social relations		Infrastructure of access to law: lawyers, legal aid	Costs	Infrastructure for executing a legal degree
	<---------- anticipation: chance of success / chance of implementation			

court cases, however, concern undisputed claims: the plaintiff simply needs a court decree to force the defendant to comply; the defendant may not even show up and the case ends with a default judgment. In most courts the dispute that is fought all the way to an adversary decision forms the minority of cases (Blankenburg 1980/95) (see Table 2).

The distribution of court functions varies with the issue at stake. Miller and Sarat (1980/81) depict the different distributions as pyramids. Discrimination grievances are very frequent, but rarely is a lawyer consulted by the aggrieved; only 8 out of 1,000 appear in court, thus the top of the pyramid gets extremely thin. Tort more often develops into a dispute, but still not as often as postdivorce cases which are regularly fought all the way to an adversorial decision—thus the top of the pyramid covers 45 percent of the baseline.

However, court decisions are not the end of the conflict story. Implementation studies after litigation have shown that parties who won in court do not necessarily make use of the entitlement they have fought for, either because the losing party is incapable of paying or satisfying the claim, or because the winner in court is using the entitlement for out-of-court negotiations in a wider social context than that of the court verdict. Invoking a law suit means entering an interaction system with the roles of several actors being determined by procedure: plaintiffs and defendants are determined by their proactive, respectively defensive strategies; decision makers by the requirements of due process which implies a highly dramatic presentation of impartiality. Actually, the strategic game starts before the suit itself: a decisive factor bearing on the likelihood of whether legal institutions are invoked at all is the rule on who has the burden of invoking the law. Litigation costs time and money: the procedural losses may outweigh the gain of even a favorable verdict. Plaintiffs have to assess their evidence, anticipate the strategy of the defendant, and evaluate the chances of being able to implement a legal title, should they get it. Basic preconditions for involving a court are already prescribed in the legal construction underlying social relationships: if the party which has the 'burden' of invoking the court has insufficient legal expertise or hesitates for motivational reasons, the other party can exploit the situation until challenged. The asymmetric party constellation in the majority of court cases suggests such exploitation of mobilization thresholds on a massive scale: most plaintiffs before civil courts are business firms and other organizations, while the majority of defendants are private parties. Organizations that routinely deal with legal problems can often turn the differential access conditions to their advantage.

Anticipation of the potential costs and the possible gains of a procedure adds to the advantages of the repeat player in court (Galanter 1974). Good lawyers and experienced court users will have better foresight than others will. Whether parties go all the way to a court decision or even an appeal depends on continual re-assessment, anticipating not only the chances of success, the costs of the procedure and the opportunity costs, but also the assessments of the other party likewise anticipating its chances. The scheme in Table 3 helps to organize the variables which are relevant in the process.

Assessing the juridical configuration in different social fields is daily work for lawyers. Their strategies and those of their clients interact within the frame of the given legal order. Their assessments vary with national jurisdictions and possibly local regimes. Comparing litigation frequencies of similar disputes between countries and states (or even court districts) reveals the overriding impact of legal regimes. To compare legal cultures, researchers look at similar grievances, like the regulation of injury and damage

after accidents. The most obvious example might be traffic accidents: traffic tort in some legal cultures frequently entangles the victims in legal procedures, while in others insurance companies regulate largely outside the courts. Comparison of American with Japanese accident regulation demonstrates that it is not cultural mentality or personal skill which explains how the Japanese avoid legal costs, but rather the legal and institutional conditions which frame their avoidance of litigation (Haley 1978). The same can be said of the infrastructure of avoidance that explains why the Dutch litigate considerably less frequently than the neighboring Germans: accident victims in these countries hardly differ in conflict mentality nor does their tort law differ. The explanation for different litigation frequencies lies in institutional conditions of the legal regime: burdens of proof, lawyer fees, legal cost insurance, and the availability of alternative modes of regulation (Blankenburg 1994). Similar comparisons can be drawn with respect to other comparable disputes: divorce, employment dismissal, landlord–tenant disputes, or debt enforcement need to be dealt with by courts in some countries while others treat them as routine regulation, reserving litigation in court as a last resort for a few, serious cases.

3. *Frontiers of the Law: Test Cases and Strategies of Mobilizing Rights*

If in the eyes of the occasional litigant 'serious' refers to the stakes which are at dispute in a particular case, to the legal profession and the legal academy, serious implies that there are principle questions of law to be resolved. The greater chances of *repeat players* (*RPs*) to win in court (Galanter 1974) are amplified in the long term: repeat players can pursue strategies over a series of cases, and they can aim at precedents and high court decisions which change the interpretation of the law. The strategic advantage enables big players in courts, such as insurance, banking, and manufacturing corporations to set the terms of statutes and standard contracts in their favor. Occasional litigants (one shooters (OSs) in Galanter's (1974) terms can only fight the legal battle by organizing as interest groups and getting together for class actions. Interest groups implicitly threaten with a potential of future litigants. Parties with especially diffuse interests, such as consumers, parties, or victims of environmental damage, whose single case might not be worth fighting a legal battle, can see their strategic chances in mobilizing courts rather than trying to win in the legislative process.

A number of institutional features favor such public interest strategies among American lawyers and in American courts: the tradition of case law, the instruments of class action, the possibility of contingency fees, and the chances for popular support by juries have made American lawyers especially inventive in thinking up novel sorts of claims, mobilizing collective expectations, and framing them in legal terms. Media coverage of pioneering cases like the asbestos and tobacco liability litigation have made their way around the world. These claims have been spectacularly successful in changing the perception and the legal liability of producers and service providers in North America and, consequently, also in continental legal systems. Another feature of the American legal system which has turned courts into a law-making forum has been the possibility of Supreme Courts to set aside statutory law by the application of human rights and constitutional jurisprudence. Many continental systems have introduced similar possibilities of overturning legislative codes by establishing constitutional courts—the collapse of fascist, communist, and apartheid regimes, with their experience of abuses of legal institutions, have facilitated such institutions. Mobilizing the global forum of human rights has opened a wave of innovation challenging traditional codes of law, and turning courts into law making rivals of parliamentary legislation (Shapiro 1980).

Of course, strategies of mobilizing courts for pioneering cases are even more demanding than the mobilization of courts by occasional litigants. The selectivity of invoking legal institutions on all levels makes us aware of a general feature of any kind of law: so long as actors comply with legal norms we cannot know whether conformity reflects the impact of law or is motivated by custom, by their present interests, or mere inertia. Under conformity, law rests dormant; it manifests itself only under special circumstances of deviation and disagreement when law is mobilized explicitly (Geiger 1947). Paradoxically, one might conclude that law works best if it does not need to be mobilized.

See also: Constitutional Courts; Disputes, Social Construction and Transformation of; Judicial Review in Law; Law and Everyday Life; Law, Sociology of; Legal Culture and Legal Consciousness; Litigation; Mediation, Arbitration, and Alternative Dispute Resolution (ADR)

Bibliography

ABA(American Bar Association) 1994 *Legal Needs and Civil Justice*. ABA, Chicago

Black D 1973 The mobilization of law. *Journal of Legal Studies* **2**: 125

Blankenburg E 1980/95 *Mobilisierung des Rechts*. Springer, Berlin

Blankenburg E 1994 The infrastructure of avoiding civil litigation. *Law and Society Review* **28**: 789–808

Blankenburg E, Voigt R, (eds.) 1987 Implementation von berichtsentscheidungen *Jahrbuch für Rechtstheorie und Rechtssoziologie* **11**

Curran B 1977 *The Legal Needs of the Public*. Chicago

Felstiner W, Abel R, Sarat A 1981 The emergence and transformation of disputes. *Law and Society Review* **15**: 631
Galanter M 1974 Why the haves come out ahead. *Law and Society Review* **9**: 95
Geiger T 1964 *Vorstudien zu einer Soziologie des Rechts*. Luchtenband
Genn H G 1999 *Paths to Justice*. Hart, Portland, Oxford, UK
Habermas J 1981 *Theorie des kommunikativen Handelns*. Frankfurt, Germany, Vol. 2, Chap. 6
Haley J O 1978 The myth of the reluctant litigant. *Journal of Japanese Studies* 359–90
Hanak G, Stehr J, Steinert H 1989 *Ärgernisse und Lebenskatastrophen*. AJV, Bielefeld
Harris D et al. 1984 *Compensation and Support for Illness and Injury*. Oxford, UK
Shapiro M 1981 *Courts: A Comparative and Political Analysis*. University of Chicago Press, Chicago

E. Blankenburg

Law: New Institutionalism

Although institutionalist approaches have seen a resurgence in many of the social sciences (sociology, political science, economics, and economic history), the 'new institutionalism' in the study of law is primarily an outgrowth of organizational sociology. The new institutionalists have focused especially on the process of institutionalization—the process by which a practice or structure is diffused, adopted by an increasing proportion of the organizations in a field, becomes a prerequisite for legitimacy, and comes to be taken for granted as a normal and expected part of organizational life. In organizational sociology, neo-institutionalism emphasizes symbolic over instrumental reasons for adopting policies and practices and organizations' attempts to gain legitimacy with key members of their environments. Researchers who import this approach into the study of law tend to emphasize the symbolic effects of bringing organizational practices into conformity with the law.

1. *Old and New Institutionalisms*

The 'old institutionalism' in sociology, as represented in the work of Selznick (especially 1957), distinguished between organizations as tools to accomplish particular purposes and organizations as entities to which people formed commitments. In pointing to processes by which an organization, originally formed for instrumental purposes, garnered the loyalty of participants who then became committed to keeping the organization alive, researchers argued that some organizations became institutions. Moreover, Selznick (1969) identified legalization as an important part of this transformation of instrumental organizations into institutions. When workers acquire rights in the workplace, he argued, they become members of an emerging polity and cease to be simply subordinates in a hierarchy.

In the classic studies of the old institutionalists, the evolving relationship between an organization and its environment was a central focus; this interest was retained by the new institutionalists. But where the old institutionalists might have looked at the process by which an organization itself (e.g., a university, a government agency) became institutionalized, the new institutionalists (starting with Meyer and Rowan 1977) instead look at the ways in which performing a broad set of tasks becomes institutionalized in particular areas of social life (what Bourdieu 1985, calls 'fields'). They investigate the processes by which particular practices come to be regarded as the 'usual' ways of, for example, providing medical care (e.g., fee for service), organizing higher education (e.g., the conventional division of university training into disciplines and departments), or choosing among potential suppliers of goods and services (e.g., the bidding practices that link government agencies with contractors).

Developed as correctives to utilitarian or economistic approaches, the old and new institutionalist perspectives contend that (a) organizational phenomena cannot be explained solely by efficiency considerations or by the performance goals of organizations, and (b) the relationship between organizations and their environments is central to institutionalization. Institutionalists also share an interest in (c) the staying power of social phenomena, and (d) the process by which social phenomena come to be taken for granted and regarded as a normal, enduring part of the social landscape. Old and new institutionalists differ, though, in (a) what social phenomena they scrutinize (the institutionalization of organizations vs. the process by which institutionalized practices diffuse throughout a field), (b) the emphasis placed on normative forces (such as commitments to the organization) as compared with cognitive factors (such as scripts and schemas), and (c) whether they believe that organizational processes are shaped deeply or more superficially by institutionalization.

2. *Law and Institutionalization*

When new institutionalists ask about the relationship between organizations and their environment, they are especially interested in the common elements of organizational environments. Laws, legal actors, and regulatory bodies are important common elements of organizational environments. Even when organizations in the same industry interact with different suppliers or market to somewhat different consumers,

they are likely to be governed by the same laws. Further, the legal environment probably has more capacity to coerce conformity than do other members of the social landscape, although conformity is also induced by the capacity of law to enunciate values and expectations and to resolve uncertainties by offering models for action. The legal environment, then, is an important source of uniformity and an important pressure for institutionalization (DiMaggio and Powell 1983).

This insight has been developed in a series of investigations of when and how specific statutes (particularly US employment law) have affected the practice of organizations. In order to gain legitimacy, researchers have argued, organizations adopt policies that conform to legal requirements (e.g., about equal opportunity hiring, wrongful discharge, or due process), develop procedures (e.g., grievance procedures) that build legal requirements into organizational routines, create positions for affirmative action officers, and charge human resources departments with the duty of implementing the law. Researchers often use event history techniques to study the diffusion of these procedures and positions that bring an organization into compliance with the law (e.g., Edelman 1992, Sutton et al. 1994, Dobbin 1998). Other researchers investigate the metamorphosis of law into organizational routine, the diffusion of legal innovations, and the role of the professions in spreading, constructing, and implementing law (e.g., Edelman 1990, Edelman et al. 1993, Carruthers and Halliday 1998, Grattet et al. 1998, Heimer 1999; for a general discussion of new institutionalist research on law, see Suchman and Edelman 1996).

3. *Symbolic and Instrumental Effects of Law in Organizations*

A key argument of the new institutionalists is that policies are adopted for symbolic purposes. In adopting a policy, an organization signals its willingness to play by the rules endorsed by other actors in its environment. It claims membership in a world in which organizations doing roughly the same thing agree on a core set of rules that make the game 'fair' and a style of interaction with regulators, suppliers, customers, and other publics. But the adoption of a policy need not alter organizational practice in any fundamental way, particularly when members of the environment are unable to monitor how deeply organizational practice is transformed.

When organizations adopt policies that bring them into conformity with the law, is organizational practice altered? Or are practices decoupled from ceremony as institutionalist theory would predict? Researchers have discovered that the answer is more complex than anticipated. Although organizations often seem to adopt conforming policies for the expected symbolic reasons, and although such policies seem to have mainly symbolic effects at the start, over time the policies may have deeper effects. For instance, affirmative action officers appointed to send a signal of conformity without disrupting organizational practices may be more deeply committed to affirmative action than those who hired them. Organizations are not unitary actors and new hires bring with them links to new constituencies. Affirmative action officers may augment an organization's ties to associations of personnel professionals. They may also feel some loyalty to minority communities as potential consumers or as labor pools.

But even symbolic effects are more complex than anticipated. For instance, in affirmative action, equal economic opportunity has more often been translated into a right to make a complaint than a right to a job. Likewise, medical organizations have responded to informed consent statutes by introducing consent forms that often do not ensure that patients and their families are informed about and have consented to a particular treatment.

However, researchers have also discovered that going through the motions of conforming to the law may lead ultimately to changes in the culture of organizations or modifications in the values of key employees. Decreases in pressure from the legal environment to adopt affirmative action policies had little effect on the adoption or implementation of such policies in American businesses, Dobbin (1998) argues. These policies had in the meantime taken on positive associations in the corporate world and been incorporated into a new story about how 'diversity' contributes to efficiency. In medicine, where physician autonomy remains an important value, scholars nevertheless concede that although in the 1960s, the vast majority of American physicians did not inform patients of cancer diagnoses, by the late 1980s, the vast majority of them claimed that they did inform their patients. Initially resistant to informed consent procedures and still inclined to implement them in a perfunctory way, physicians have nevertheless come to believe in patient rights (Heimer 1999). Symbolic effects may in the long run lead to instrumental effects, and initial symbolic gestures should perhaps be seen as the first phases in a protracted engagement whose ultimate outcome depends on whether new constituencies are created and new rationales developed that transform an imposed policy into an indigenous project.

New institutionalists have made legal scholars more cognizant of the importance of thinking about organizations as sites for the implementation of law. Because laws often are implemented by organizations, such as through the adoption and enforcement of new procedures or the creation of new departments and positions, the efficacy of laws depends deeply on what goes on inside organizations. If the creation of new posi-

tions leads to new affiliations, builds constituencies, or creates a forum for voices that have previously not been heard, then a policy adopted only to mollify a government agency may ultimately have some of the effects that the government agency intended. As often happens as a new perspective matures, the novel observations of the new institutionalists about the symbolic aspects of organizational policy have been recombined with earlier streams of thought. The new synthesis acknowledges that policies may have both instrumental and symbolic effects, and asks when each is especially likely to occur. Interestingly, the very scripts and schemas that have been identified as part of the process of institutionalization are now seen as routes to instrumental effects. Some policies remain window dressing; others are written into routines. Some organizational routines restructure work, others do not.

4. *Institutions, Uniformity, and Innovation*

Institutionalist arguments tend to focus on the production of uniformity. As institutionalized practices are more widely adopted in a field, it becomes deviant to do things differently: not to have an affirmative action officer, not to have an informed consent form, not to include a statement about diversity in advertisements for jobs, not to have rules about hate crimes. But these observations about legitimacy and uniformity raise questions about whether institutionalization produces stagnation and how innovation is possible.

Institutionalist arguments about decoupling of ceremony from day-to-day practices provide a partial answer. If legitimacy can be secured by superficial conformity, organizational handbooks may be brought into conformity with law while instrumental practices remain as before. Institutionalization may produce organization handbooks that portray organizations as rigid, and in particular as scrupulously conforming to the law, when they are in fact quite supple and quite innovative in devising ways to redefine the law. Resistance in decoupling ceremony from practice and in devising policies that satisfy the letter if not the spirit of the law, then, is one form of innovation. Edelman et al. (1993), for instance, discovered that in the hands of human resources officers responsible for processing discrimination complaints, civil rights law moved into the managerial realm. As a result, a right to complaint resolution replaced a more political right to a nondiscriminatory workplace.

We should also remember, though, that the standard practices of one field can be significant innovations when imported into another. Thus, increasing uniformity of some sorts may bring with it increasing diversity of other sorts. For instance, uniformity in strategies for solving a category of problems may bring with it diversity in the problems to which an institutionalized solution is applied. Thus Grattet et al. (1998) found that, in the United States, the variety of legal solutions used to combat hate crimes (e.g., criminalizing interference with civil rights, creating new categories of crime, increasing penalties for preexisting categories of crime) has decreased over time. At the same time, the number of 'protected statuses' (e.g., race, religion, ethnicity, sexual orientation) has increased.

These arguments that innovation occurs when actors invent ways to signal compliance without altering mundane organizational practice, transform unpalatable laws into more circumscribed routines, or transport institutionalized practices to new settings, all refocus attention on the innovators. In a variety of settings—employment law, corporate bankruptcy law, medical regulation, to name a few—scholars have found that professionals have played important roles. Rather than being identified solely with the organization, professionals retain commitments associated with their professions. They may dampen or amplify the impact of the law depending on whether legal tools are useful in advancing professional interests and achieving professionally valued goals (Edelman 1992, Heimer 1999). In their investigation of corporate bankruptcy law in the USA and the UK, Carruthers and Halliday (1998) argue that the role of professionals in legal innovation varies with their status and autonomy. Members of strong professions are not limited to devising ways of decoupling ceremony from practice or molding the routines that implement the law. Instead, high status professionals shape more directly the laws that govern an organizational field and frame their own work. They originate the practices that will eventually become the institutions that others adopt. Similar arguments have been made about the creation of international legal regimes (Dezalay and Garth 1996).

5. *Institutional Competition and Legitimacy*

In addition to the questions about efficacy and innovation discussed above, new institutionalists' work on law also raises questions about the difficulty of gaining legitimacy with one portion of an organizational environment without alienating others. Initial formulations of the institutionalist argument suggested that the divergence of interest between internal and external audiences would often lead to a decoupling of ceremony from instrumental practice. Policies conforming to law might be adopted formally but ignored in practice. But because organizations' environments cannot be expected to have unitary interests, and because environments actively compete for the right to construct the institutions that will govern a field (Bourdieu 1985), signals must be crafted carefully to appeal to, or at least not to alienate,

multiple audiences. This fine-tuning of organizational policy often takes into account the relative power of members of the environment, the differential capacity of environmental elements to discern the difference between superficial and deeper adjustments to their interests, and the ease with which people and ideas pass between the organization and some parts of its environment.

In medical settings, for instance, the adoption of legally mandated 'patient friendly' policies may make medical practitioners nervous and so decrease the chances of recruiting prestigious departmental heads or the most attractive residents. At the same time, though, reluctance to bring hospital practice into conformity with new statutory requirements (e.g., ethics committees, medical care power of attorney, or advance directives) may increase state scrutiny. How, then, do hospitals simultaneously curry favor with both the state and with professionals? Organizational boundaries are, to varying degrees of course, porous, and some members of the environment have more capacity to cross such boundaries than others.

In medical settings, for instance, legal regulators are more completely outsiders than are members of medical professional societies. Those who span the boundary (e.g., professionals working in a hospital who are also members of professional associations) will be able to detect when policy and practice are decoupled. They will also be more able to ensure that day-to-day practice is more deeply shaped by the institutionalized policies to which they are committed. When the interests of medical practitioners and legal actors coincide, as when laws create tools that increase the power and autonomy of medical professionals, organizational routines that implement these laws can be expected to have substantial effects on practice.

Questions of legitimacy inevitably lead back to concerns about efficacy, then. Legitimacy can be purchased with ceremony only when powerful outside audiences cannot tell the difference. When organizational boundaries are more permeable and outsiders are able to form alliances with key insiders, an organization will be regarded as legitimate only to the degree that symbol and substance coincide. For legal actors, that can mean forming alliances with other professionals who are deputized to represent state interests inside organizations, or even in hospitals, moving court sessions inside the physical boundaries of organizations.

6. *Conclusion*

New institutionalist investigations of legal phenomena have led to important findings about the metamorphosis of statutes as they are transformed into organizational routines, the role of these routines in creating new constituencies, and the location of innovation. These studies have modified the agendas both of institutionalists, by altering their understanding of the place of legal actors in the process of institutionalization, and of sociolegal scholars, by clarifying thinking about the efficacy of law.

See also: Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives; Institutional Economic Thought; Institutionalism; Institutionalization; Institutions; Legitimacy; Organizational Culture, Anthropology of; Organizations and the Law; Organizations, Sociology of

Bibliography

Bourdieu P 1985 The social space and the genesis of groups. *Theory and Society* **14**: 723–44

Carruthers B G, Halliday T C 1998 *Rescuing Business: The Making of Corporate Bankruptcy Law in England and the United States*. Clarendon Press, Oxford, UK

Dezalay Y, Garth B G 1996 *Dealing in Virtue: International Commercial Arbitration and the Construction of a Transnational Legal Order*. University of Chicago Press, Chicago

DiMaggio P J, Powell W W 1983 The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review* **48**: 147–60

Dobbin F 1998 The strength of a weak state: The rights revolution and the rise of human resources management divisions. *American Journal of Sociology* **104**: 441–76

Edelman L B 1990 Legal environments and organizational governance: The expansion of due process in the workplace. *American Journal of Sociology* **95**: 1401–1440

Edelman L B 1992 Legal ambiguity and symbolic structures: Organizational mediation of civil rights law. *American Journal of Sociology* **97**: 1531–76

Edelman L B, Erlanger H S, Lande J 1993 Internal dispute resolution: The transformation of civil rights in the workplace. *Law and Society Review* **27**: 497–534

Grattet R, Jenness V, Curry T R 1998 The homogenization and differentiation of late-crime law in the United States, 1978 to 1995: Innovation and diffusion in the criminalization of bigotry. *American Sociological Review* **63**: 286–307

Heimer C A 1999 Competing institutions: Law, medicine, and family in neonatal intensive care. *Law and Society Review* **33**: 17–66

Meyer J W, Rowan B 1977 Institutionalized organizations: Formal structure and myth and ceremony. *American Journal of Sociology* **83**: 340–63

Powell W W, DiMaggio P J (eds.) 1991 *The New Institutionalism in Organizational Analysis*. University of Chicago Press, Chicago

Selznick P 1957 *Leadership in Administration*. Row, Peterson, Evanston, IL

Selznick P 1969 *Law, Society, and Industrial Justice*. Russell Sage Foundation, New York

Suchman M C, Edelman L B 1996 Legal rational myths: The new institutionalism in organizational analysis. *Law and Social Inquiry* **21**: 903–41

Sutton J R, Dobbin F, Meyer J W, Scott W R 1994 The legalization of the workplace. *American Journal of Sociology* **99**: 944–71

C. A. Heimer

Law: Overview

Law refers to systems of rules and standards connected to social institutions that pronounce and apply them. Law is not just a discipline of inquiry but a set of practices, roles, and ideas embodied in distinctive institutions and reflected in general culture. There are two kinds of knowledge about law. The first 'internal' kind of legal learning is generally found among the professional groups that operate legal institutions (judges, lawyers, academics, administrators, and so forth) (see *Legal Scholarship*; *Legal Reasoning and Argumentation*; *Legal Formalism*). These actors carry and expound a body of authoritative normative learning that regulates and instructs both officials and citizens. This 'internal' kind of legal learning is entwined with an emergent 'external' kind of learning about law and legal institutions most often found among academics. The external kind of learning seeks explanation rather than justification, emphasizes social process rather than rules, and appreciates the dynamics of law as part of more inclusive patterns of social life.

Although these two kinds of knowledge about law are entwined and not entirely distinct, the articles about law and legal institutions in this encyclopedia draw primarily upon the second 'external' kind of inquiry, known variously as 'law and society,' sociolegal studies, the sociology of law, or simply legal studies. This social inquiry into law is itself varied. The late twentieth century saw a great proliferation of rival and overlapping 'schools of thought about law' (see *Legal Positivism*; *Natural Law*; *Legal Realism*; *Law and Economics*; *Critical Legal Studies*; *Critical Race Theory*; *Feminist Legal Theory*; *Law and Everyday Life*; *Law, Autopoiesis in*; *Law: New Institutionalism*; *Law as Constitutive*; *Law and Society: Sociolegal Studies*), many affiliated with various disciplines in the social sciences and humanities, and with movements in the wider political culture. Although there were distinguished precursors and ancestors (e.g., Maine, Durkheim), systematic social inquiry into law was envisioned only in the early years of the twentieth century, and the emergence of a community of scholars actively and continuously pursuing such inquiry, involved in a cumulative discourse across disciplinary boundaries, dates from the middle of the century (see *Law and Society: Sociolegal Studies*).

With the emergence of 'external' scholarship about law, the focus of scrutiny has broadened and decentralized, moving downward from peak decision makers to the field level, from Supreme Courts to trial courts, from professionals to the law's users and customers, from formal process to backstage bargaining, from comprehensive policy making to strategic maneuvering, from distinctive legal institutions to understandings of law in everyday activities. Attention has moved backward from lawsuits to disputes, grievances and injuries, and forward from judgments to impacts and wider effects, and outward from legal doctrine to perceptions and understandings of law among citizens. As these new external discourses proliferate and infiltrate mainstream 'internal' legal discourse, internal discourse itself becomes more receptive to facts and theories that would have been rejected as 'extra-legal' a generation earlier. For example, American courts cite law reviews less frequently, and newspapers more (Schauer and Wise 2000), while appellate briefs routinely marshal empirical arguments. By the end of the twentieth century, the discourses of and about law had become less distinct, hermetic, and self-contained.

Various entries in this encyclopedia describe some of the most prominent legal systems (see *Common Law*; *Civil Law*; *Hindu Law*; *Jewish Law*; *Chinese Law*; *Socialist Law*; *Lex Mercatoria* and others). Other entries describe the social institutions in which such rule systems are embodied (e.g., *Courts and Adjudication*, *Supreme Courts*, *Constitutional Courts*, *Juries*, *Law Firms*, *Prosecution*, *Policing*, *Punishment: Social and Legal Aspects*, *Prisons and Imprisonment*, *Liability: Legal*), the distinctive social roles associated with those institutions (*Judges*; *Lawyers*), and the distinctive values to which they are oriented (see *Rule of Law*; *Rechtsstaat (Rule of Law: German Perspective)*; *Due Process in Law*; *Legitimacy*; *Equity*; *Justice and Law*). Another set of entries take up some of the characteristic features and processes of legal institutions in complex societies (see *Procedure: Legal Aspects*; *Discovery: Legal Aspects*; *Expert Testimony*; *Scientific Evidence: Legal Aspects*; *Trials: Cultural*; *Appeals: Legal*; *Remedies and Damages: Legal Aspects*; *Statutes of Limitations*) and some of the fields of legal regulation (see *Property: Legal Aspects*; *Intellectual Property: Legal Aspects*; *Torts in Law*; *Contracts: Legal Perspectives*; *Bankruptcy*; *Family Law*; *Administrative Law*). Other entries address the attempts of legal institutions to regulate various spheres of activity (see *Markets and the Law*; *Regulation Theory in Geography*; *Environment Regulation: Legal Aspects*; *Employment and Labor, Regulation of*; *Occupational Health and Safety, Regulation of*; *Race and the Law*; *Affirmative Action: Empirical Work on its Effectiveness*; *Affirmative Action: Comparative Policies and Controversies*; *Law and Aging*; *Law and People with Disabilities*) and the impact of these other spheres on the law and its institutions (see *Class and Law*; *Crime and Class*; *Crime and Ethnicity (Including Race)*; *Crime and Gender*; *Organizations and the Law*).

Finally, other entries focus on the ways the legal complex is changing over time (see *Law: Change and Evolution*; *Globalization: Legal Aspects*; *Legalization*). This article attempts to distill some of the major themes and findings of social inquiry into law, and to suggest their interrelationships.

1. *The Gap*

Much scholarship in the sociology of law has focused on what is seen as a gap between 'law on the books' and 'law in action.' 'Law on the books' refers not simply to laws codified in statute books or judicial decisions, but more broadly to the formal legal system and to assumptions about how that system ought to operate. In contrast, 'law in action' refers to how law in fact operates in society, taking into account power, discretion, social roles, deviance, and other social realities. For example, Macaulay's seminal article on contract law showed that businesses often preferred informal agreements to formal contracts, and rarely invoked formal contractual remedies when contracts were breached (Macaulay 1963). A good deal of law and society research examines aspects of this gap in the contexts of courts, legislation, administrative justice, lawyers, organizations, law enforcement, and the behavior of parties with legal complaints (see, e.g., *Courts and Adjudication*; *Lawyers*; *Rights: Legal Aspects*; *Race and the Law*; *Class and Law*; *Civil Rights*; *Injustice: Legal Aspects*). The focus on the gap has also led to inquiries into the conditions under which law can bring about social change (e.g., Rosenberg 1991, McCann 1994), the circumstances under which people obey law (Lind and Tyler 1988, Tyler 1990), and the role of law in social control.

Although the idea of the gap has motivated some important inquiries into the workings of law in society, late-twentieth-century scholarship found the idea of the gap to be flawed theoretically. Feeley (1976) suggests, for example, that the characterization of the gap depends on an overly narrow conception of the formal law and inability to identify or measure the goals of that law. Because formal law is often quite ambiguous, in part because its creators had diverse interests and objectives, it is misleading to characterize the actual operation of law as a deviation from the clear goals of formal laws. Rather, law can only be understood as embedded in its social context, where it is both the product and the source sociolegal practices.

2. *Law as a Symbolic System*

The pronouncement and implementation of legal norms are attended invariably by indirect and unanticipated consequences. Law rarely works as anticipated by lawmakers. The rules and rights that are promulgated by legislatures and courts are usually broad and ambiguous, in large part because lawmakers agree more on broad symbolic meanings than on detailed constructions. Especially because legal measures are only partially funded by the state and depend on inputs from interested parties, the way that rules and rights are interpreted and enforced depends on the contention among interested parties, who vie for resources and attention and negotiate the meaning of compliance according to their interests or institutionalized ways of thinking. Even though laws are promulgated by the state and intended to penetrate downward through society, law is in many ways driven from below by litigants, clients, and other intended subjects of regulation. Citizens are not law's passive subjects; rather, they are active interpreters of legal ideas and symbols and, at least sometimes, strategic actors who mobilize the law selectively to pursue their own visions of advantage and justice.

Law operates not simply as a body of regulatory controls and public edicts, but also as a set of symbols, with different meanings for different social groups. The symbolism of law, by invoking ideals and principles such as due process or equal opportunity, may ermpower citizens to take political action or to make stronger claims in informal negotiations (see Rights). However, the symbolism of law may also diminish the substantive role of law. When contests over morality are fought on legal terrain, parties may care more about the symbolic victory than about the actual effects of law (Gusfield 1963). And political actors benefit from laws that provide symbolic victories to powerless parties without disturbing the status quo (Edelman 1964). Critical legal scholars and neoMarxists emphasize the hegemonic role of law: Laws that appear to provide rights irrespective of class, gender, or race help to solidify the power of the elite (see *Critical Legal Studies*; *Marxism and Law*).

Law may engender symbolic forms of compliance as well. In response to civil rights law, for example, work organizations have created an array of structures designed to symbolize attention to civil rights: These include affirmative action or equal employment opportunity programs, plans, procedures, offices, and functionaries. Over time, these 'symbolic structures' become institutionalized forms of compliance with law. Once in place, however, symbolic structures may enhance employees' rights consciousness and make it more difficult for employers to deny employees' law-related claims, thus enhancing indirectly the impact of law (Edelman 1992; see also *Organizations and the Law*). Studies of law must therefore be attentive not only to the direct and substantive role of law but also to its symbolic role in society.

Law as a symbolic system also has important effects on 'legal consciousness.' Although there is no consensus as to the meaning of this construct, it refers broadly to an institutionalized set of ways in which individuals come to think about law and to employ legal ideas and principles in their daily lives. Over time, certain ideas about law acquire stable and objectified

meanings, which in turn limit and constrain future meaning-making vis-à-vis law and legal rights (see *Legal Culture and Legal Consciousness*;). Legal consciousness scholarship seeks to decentralize the study of law, emphasizing the role of law in everyday life rather than the behavior of distinctively legal institutions and actors. Scholars seek to show how legal constructs acquire meaning through everyday social interaction, contest, and routine. Although somewhat inspired by neoMarxist scholarship, this work is more influenced by studies of culture and institutions. It does not generally see legal consciousness as false consciousness, but rather as a site of resistance to, or engagement with, law and legal institutions.

Legal consciousness has been studied within particular social arenas such as families, religious communities, or welfare offices. Other work focuses on legal settings. For example, Conley and O'Barr (1990) studied how litigants in small claims courts identify and analyze legal problems, and observed a difference between experienced players, who adopt a rule-oriented style, and less experienced (especially female, poor, and minority) players, who focus on relationships and narrative. Some authors emphasize shared experience as a source of legal consciousness (suggesting, for example, that law would have similar meanings to people on welfare, or Hispanic women) while others focus on the development of legal consciousness in a community (so that, for example, people in a particular geographic locale would share a legal consciousness irrespective of their economic status, race, or sex). Others emphasize that legal consciousness is multiple and varied: Ewick and Silbey (1998) suggest that legal consciousness is not a direct reflection of legal structures or personal experience, but that the same individuals in their daily lives can manifest mutually 'inconsistent' relations to law, with their view of law alternating between a remote and majestic authority, a game, and an oppressive power.

3. *The Distributive Tilt of the Legal Processes*

The mobilization of law is a social process, one that is dependent on access to information, resources, legal representatives, and organization (see *Mobilization of Law*). Mobilization requires that a potential claimant recognize an injury, attribute blame to a legally relevant party, and pursue a claim against that party (Felstiner et al. 1981). The very social characteristics that make people more likely to suffer legal injuries, such as poverty or minority identity, also act as barriers to mobilization (Bumiller 1988). Those whose lives are characterized regularly by discrimination or rights denials are less likely to perceive a particular injury as worthy of legal redress, and are more likely to view the legal system as inaccessible or unjust (see *Injustice: Legal Aspects*; *Equality and Inequality: Legal Aspects*). As a result, social reform laws—especially to the extent that they depend on victims to mobilize their rights—infrequently to effect major social change (see *Law as an Instrument of Social Change*).

Although modern adjudication is infused with ideals of equality before the law, and the isolation of the dispute from the identity, connections and resources of the parties, this isolation is always realized imperfectly. Those possessed of wealth, power and other resources can hire better lawyers, invest in marshaling proofs and arguments, use delay to their advantage, and adopt strategies that deflect unfavorable rules and reap the full advantage of favorable ones (see *Injustice: Legal Aspects*; *Equality and Inequality: Legal Aspects*). The advantages of stronger parties who are recurrent users of the legal system is even more pronounced in nonjudicial forums such as legislatures and administrative agencies. So legal regimes that accord a central place to equality norms may at the same time enable 'the "haves" [to] come out ahead' (Galanter 1974).

In response to this paradox legal systems have incorporated various devices to overcome the disparity in legal resources (see *Justice, Access to: Legal Representation of the Poor*; *Class Actions: Legal*; *Public Interest Law*). The achievements and deficiencies of these devices have been a focus of law and society research. Although the presence of lawyers tempers the disparities in legal resources, it does not overcome them fully, for legal services are typically purchased in markets where those with superior resources can attract the most proficient practitioners. Apart from differences in quality and quantity, legal representation is infected with agency problems. Lawyers have goals and needs that do not align perfectly with those of their clients. It is not simply that financial incentives may induce lawyers to do too much or too little, but that even the lawyer who eschews financial gain may be deflected by appetite for fame, honor, or for 'interesting' work. For example, the low monetary stakes and repetitiveness of the legal needs of the poor run against the powerful pull of craft satisfactions and the aspiration to heroic distinction.

4. *Regression Toward Bargaining*

One of the 'discoveries' of law and society research is the pervasiveness and centrality of bargaining throughout the legal world. Processes of authoritative decision-making turn out in fact to revolve around negotiation. Ostensibly adjudicative processes decompose into negotiative ones; decision makers clothed with arbitral powers are transformed into mediators. The negotiated outcome in the shadow of the law turns out to be the master pattern of disputing in American courts and administrative agencies.

Bargaining often takes place in the 'shadow of the law' (Mnookin and Kornhauser 1979). The rules of formal law confer on the parties what Mnookin and Kornhauser call a 'bargaining endowment,' that is, a set of 'counters' to be used in bargaining between disputants (Ross 1970).

The bargaining endowment that the formal law bestows on parties derives not only from the substantive entitlements conferred by the legal rules but also from the adjectival rules that enable those entitlements to be vindicated. But rules are only one source of the endowment conferred by the legal process: The delay, cost and uncertainty of eliciting a favorable determination also confer bargaining counters on the disputants. Delay, cost and uncertainty may themselves be the product of rules, but they also result from such nonrule factors as the number, distribution, fees, and work patterns of courts and lawyers. The meaning of the endowment bestowed by the law is not fixed and invariable, but depends on the relative characteristics of the disputants: their preferences, negotiating skill, aversion to risk, ability to bear costs and delays, to respond to deadlines and emergencies, and so forth. A different mix of disputant capabilities may make a given endowment take on a very different significance.

Knowledge about the legal system is often incomplete or inaccurate, however, so that social, business, and professional norms have a greater impact on how parties bargain than do their putative legal bargaining positions (Ellickson 1991, Galanter 1986, Edelman, Erlanger and Lande 1993). Parties may be influenced less by what would happen in court than by how their neighbors or business partners or co-workers think they ought to behave. The preservation of relationships often trumps the assertion of legal entitlements.

5. *Plurality*

Some legal systems openly yoke together several bodies of law from different origins. For example, some ancient and modern societies have had a system of 'personal law' in which the incidence of particular rules depends on the religious or tribal identity of the parties (see *Legal Pluralism*; *Folk, Indigenous, and Customary Law*). In most modern societies, the law presents itself as unitary, but nonetheless consists of multiple systems of regulation that coexist and intertwine. In addition to primary rules, the law contains various sets of secondary rules about the recognition and application of its own rules (Hart 1961) and the recognition and reinforcement of regulatory activity in other institutions. Religious bodies, universities, trade groups, and associations of all kinds devise and broadcast norms, regulate their affairs, and impose sanctions—often by means of proceedings that mimic, if imperfectly, those of state law.

Regulation is not an official monopoly; there are multiple arenas of normative enterprise, innovation and interpretation. The state's law does not preside over a landscape barren of regulation, but over a thick tangle of rivals and companions. Its effects depend on the way that it interacts with the various sorts of indigenous ordering that surround it; the core official institutions themselves are the scene of persisting and pervasive local variation. These various departures from unity and uniformity have been studied under the rubrics of private government, indigenous law, semi-autonomous social fields, and local legal culture. Pluralism is very much with us: The centralist view of a monolithic integrated legal order turns out not to be a description of modern law, but part of its ideology.

Even though law regulates increasing areas of social life, most social regulation goes on within various institutional and relational settings; only a small and unrepresentative minority of disputes proceed to a formal legal institution for resolution. To a large but perhaps diminishing extent, when formal legal institutions devise rules to regulate conduct in institutions and relationships, they refine and restate norms that already prevail in those settings (Bohannan 1965); legal norms in turn transform and feed rather than eradicate the regulatory activity indigenous to these settings. The continuance of *de facto* pluralism is fostered by various ideological currents, including movements to preserve and extend cultural diversity, promote local control, and de-emphasize adjudication in favor of 'alternative dispute resolution.'

6. *Convergence and Difference*

For several centuries, there has been a worldwide diffusion of legal norms, materials and devices from Western Europe, lately joined and surpassed by North America as a source of legal inspiration. Voluntary reception by settler societies, the imposition of colonial rule, and importations by independent regimes seeking a badge and instrument of advancement have been followed by a worldwide spread of constitutionalism and continued influence of Western legal models to promote new forms of economic activity stimulated by increased in global trade (see *Postcolonial Law*; *Law and Development*). Diffusion of Western models was accelerated by the demise of communism and the subsequent importation of Western experts to write constitutions and oversee the democratization of formerly communist regimes. Comparative analyses of constitutional borrowing, adaptation, and hybridization have become prominent in sociolegal scholarship (see *Constitutionalism, Comparative*).

Notwithstanding some convergence of formal legal doctrines and institutional structures, there remains broad variation in the amount and character of litigation, the number and roles of lawyers, and the styles of dispute resolution across nations and social

contexts. Even when constitutions and statutes are imported directly from one country to another, their interpretation and significance reflect persisting differences in social, political, and legal settings.

7. The Legalization of Society

Law has become more pervasive, both as a set of social structures and as a component of consciousness, and this process is likely to continue. The latter half of the twentieth century saw a steady increase in the amount and complexity of regulation; the frequency of litigation; the number, coordination, and productivity of lawyers; the number of legal actors and the resources they devote to legal activity; the amount of information about law, and the velocity with which it circulates; and a concomitant rise in consciousness of law and awareness of legal rights (Galanter 1992) (see *Rights: Legal Aspects*).

An increasingly global economy together with new communications technologies have fueled the rise of legal activity, in areas such as intellectual property (see *Property: Legal Aspects*). The number of lawyers increased dramatically in many countries and became markedly more diverse (in terms of gender, ethnicity, and class origin) (see *Law Firms*).The large law firm devoted to chaperoning business clients, born in the United States in the late nineteenth century, spread throughout the common law world and began making inroads into civil law jurisdictions as well (see *Law Firms*). Regulation has increasingly taken on the features of adjudication (Teubner 1987). The role of courts has been enlarged by the spread of constitutionalism and judicial review, together with an increase in multi-tier regulation deriving from federalism or from participation in international institutions. Increasingly, courts are accepted as the final arbiters of many public issues (Cappelletti 1985).

In many quarters, these changes have engendered a recoil against law and regulation, not least among elites who are best protected by the legal order. Concerns about excessive legalization, 'juridification,' the 'litigation explosion,' and the dominance of lawyers have led to the advocacy of deregulation, delegalization, and embrace of alternatives to traditional litigation. The term 'alternative dispute resolution,' virtually unheard of prior to 1970, has become part of the standard policy lexicon. With the rise of forums such as mediation, mini-trials, and other private or court-annexed modes of dispute processing (see *Mediation, Arbitration, and Alternative Dispute Resolution (ADR)*) the locus of some legal activity has shifted away from adjudication and into nonpublic forums, but it remains doubtful that the tide of regulation has been reversed.

As discussed below, the prospect is that the legal system has an ever-expanding agenda. But the process of remedying injustice or simply assuaging grievances consumes resources and society gets an increment of justice instead of medicine or missiles. In addition to the costs of operating legal institutions and costs borne by the participants in the legal process, the use of law to address injustice imposes another kind of cost, as it accumulates into a thick layer of rules to be learned, permissions to be obtained, forms to be filled out, and records to be kept, adding to expense and burdening activity. These costs of law are visible and palpable, while the benefits of legalization are less so: Victims may enjoy vindication or honor or a heightened sense of self-worth—goods that are not subject to the same zero-sum distributional rules as money. The legal process may generate public goods by crystallizing and broadcasting norms, preventing injustice by deterrence and education, and providing channels for interaction and coordination (e.g., driving on the right or left, as the case may be). The more successful, the less visible..

8. The Organizational Incline

As law has become more complex, technical, expensive, and pervasive, it has become increasingly a realm in which organizational actors flourish and into which natural persons enter only in life's exigencies (although their vicarious experience of it through the mass media has increased). An increasing portion of lawyers' effort is devoted to work for organizations. Law firms, courts, and other legal institutions become more managerial, more rationalized, and more bureaucratic (Heydebrand and Seron 1990), and organizations have become legalized (Edelman 1990).

Organizational life is increasingly influenced by law. (Edelman and Suchman 1999; see *Organizations and the Law*). As the amount of law regulating organizational life grows, organizations respond by creating internal rules that specify the meaning of ambiguous legislation in terms of corporate interests; hiring knowledgeable in-house counsel who have organizational objectives at heart; using private police forces who answer to corporate rather than public concerns; and creating internal dispute resolution forums in which both procedural and substantive rules reflect managerial training and values. The internalization of law makes organizations ever more powerful legal actors in the public realm as well, and they drive up the ante by dominating the dockets, playing for favorable statutes and judicial interpretations of those statutes, and employing strategies and experts that tend to increase the cost and complexity of the legal game. The result is that law is mediated increasingly through organizations (Edelman, Uggen and Erlanger 1999) and becomes less accessible to individuals.

9. Law's Moving Frontier

The scope of potential legal intervention expands with the growth of knowledge and the development of technology. As the risks of everyday life have declined dramatically there is a sense that science, technology, and government can produce solutions for the remaining problems. As more things are capable of being controlled by human institutions, the line between what is seen as unavoidable misfortune and what is seen as imposed injustice shifts: The realm of injustice is enlarged. (Shklar 1990). Hurricanes are misfortunes, but bungled relief efforts are injustices. Once having an incurable disease was an unalterable misfortune; now a perception of insufficient vigor in pursuing a cure can give rise to a claim of injustice. As the scope of possible interventions broadens, more and more terrible things become defined by the incidence of potential intervention. Thus poverty, disease, and disability are not unalterable matters of fate but a matter of appropriate interventions. Our consciousness of injustice—and the undone work of law—increases not because the world is a worse place, but because it is in important ways a better, more just place.

Advances in human capability and rising expectations result in a moving frontier of injustice. Like medicine, law enlarges both the supply of answers and the supply of unanswered questions. For the most part, the advances in human capability and control that drive the justice frontier are located in or managed by corporations and governments. New claims of injustice arise in response to this management. These organizations produce new injustices not because they are bad guys, but because they are good guys doing their thing. We rely on law to respond to some of these claims by inducing actors (through education and threats) into better (less unjust) practices.

Whatever the return on legal initiative, the supply of injustice is endless, and there are attractive competitors to investment in justice-seeking, so that doing justice inevitably entails questions of rationing, priorities and efficiency. Whose grievances are addressed, and which of these get substantive remedies and which symbolic succor, is affected by differential access to law and capacity to mobilize the law.

Bibliography

Bohannan P 1965 The differing realms of law. *American Anthropologist* **67**(6) Part 2: pp. 33–42

Bumiller K 1988 *The Civil Rights Society: The Social Construction of Victims*. Johns Hopkins University Press, Baltimore, MD

Cappelletti M 1985 Repudiating Montesquieu? The expansion and legitimacy of constitutional justice. *Catholic University Law Review* **35**: 1–32

Conley J, O'Barr W 1990 *Rules versus Relationships: The Ethnography of Legal Discourse*. University of Chicago Press, Chicago, IL

Edelman L B 1990 Legal environments and organizational governance: The expansion of due process in the workplace. *American Journal of Sociology* **95**(6): 1401–40

Edelman L B 1992 Legal ambiguity and symbolic structures: Organizational mediation of civil rights law. *American Journal of Sociology* **97**(6): 1531–76

Edelman L B, Suchmann M C 1999 When the 'haves' hold court: The internationalization of disputing in organizational fields. *Law and society Review* **33**: 941–91

Edelman L B, Uggen C, Enanger H S 1999 The endogeneity of legal regulation: Grievance procedures as cational myth. *American Journal of Sociology* **105**: 406–54

Edelman M J 1964 *The Symbolic Uses of Politics*. University of Illinois Press, Urbana, IL

Ellickson R C 1991 *Order Without Law: How Neighbors Settle Disputes*. Harvard University Press, Cambridge, MA

Ewick P, Silbey S S 1998 *The Common Place of Law: Stories from Everyday Life*. University of Chicago Press, Chicago, IL

Feeley M M 1976 The concept of laws in social science: A critique and notes on an expanded view. *Law and Society Review* **10**(4): 497–524

Felstiner W L F, Abel R L, Sarat A 1981 The emergence and transformation of disputes: Naming, blaming, claiming. *Law and Society Review* **15**(3–4): 631–54

Friedman L M 1985 *Total Justice*. Russell Sage Foundation, New York

Galanter M 1974 Why the 'haves' come out ahead: Speculations on the limits of legal change. *Law and Society Review* **9**: 95–160

Galanter M 1986 Adjudication, litigation and related phenomena. In: Lipson L, Wheeler S (eds.) *Law and the Social Sciences*. Russell Sage Foundation, New York, pp. 152–257

Galanter M 1992 Law abounding: Legalisation around the North Atlantic. *The Modern Law Review* **55**: 1–24

Gusfield J R 1963 *Symbolic Crusade: Status Politics and the American Temperance Movement*. University of Illinois Press, Urbana, IL

Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK

Heimer C A, Staffen L R 1998 *For the Sake of the Children: The Social Organization of Responsibility in the Hospital and the Home*. University of Chicago Press, Chicago, IL

Heydebrand W, Seron C 1990 *Rationalizing Justice: The Political Economy of the Federal District Courts*. State University of New York Press, Albany, NY

Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum, New York

Macaulay S 1963 Non-contractual relations in business: A preliminary study. *American Sociological Review* **28**: 55–76

McCann M W 1994 *Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization*. University of Chicago Press, Chicago, IL

Merry S E 1990 *Getting Justice and Getting Even: Legal Consciousness among Working Class Americans*. University of Chicago Press, Chicago, IL

Mnookin R H, Kornhauser L 1979 Bargaining in the shadow of the law: The case of divorce. *Yale Law Journal* **88**: 950–97

Rosenberg G N 1991 *The Hollow Hope: Can Courts Bring About Social Change?* University of Chicago Press, Chicago, IL

Ross H L 1970 *Settled Out of Court: The Social Process of Insurance Claims Adjustments*. Aldine, Chicago

Sarat A 1990 The law is all over: Power, resistance, and the legal consciousness of the welfare poor. *Yale Journal of Law and the Humanities* **2**: 343

Schauer F, Wise V J 2000 Nonlegal information and the delegalization of law. *Journal of Legal Studies* **29**: 495–515
Shklar J N 1990 *The Faces of Injustice*. Yale University Press, New Haven, CT
Teubner G (ed.) 1987 *Juridification of Social Spheres*. Walter deGruyter, Berlin
Tyler T R 1990 *Why People Obey the Law*. Yale, New Haven
Tyler T R 1989 The quality of dispute resolution processes and outcomes: Measurement problems and possibilities. *Denver University Law Review* **66**: 419

L. B. Edelman and M. Galanter

Law, Sociology of

Sociology of law is a distinctive form of sociology—the nexus of two disciplines within the social sciences. As such, it is essentially interdisciplinary by nature, and therefore of interest to both jurists and sociologists. Although law historically is considered one of the oldest institutions structuring society, its definition still remains problematic. For the purposes of this article, 'law' will be defined restrictively—as the norms, rules, and normative principles established by a political authority (legislator, courts) in order to secure social order and cooperation among members of a given society. A more extensive and more sociological definition would, however, include the institutions and professionals vested with the legitimacy to produce, interpret, and apply the rules.

Sociology of law has been developed as an empirical knowledge, with theoretical foundations, aimed at describing and explaining the mutual relationships between law in its various forms, and social structures, social institutions, culture, ideologies, and values. Its objective is to unravel the various social and cultural influences that shape the law, whether law is produced by the legislator, the courts, or professionals of law and, conversely, to throw light on the action exerted by law in social life, social structures, and social change.

1. *The Origins of Legal Sociology*

Two sources can be identified as the origins of legal sociology. The first is the questioning of legal formalism (by American legal realism and Roscoe Pound's 'sociological jurisprudence'), of legal positivism (by the German Free School of Law), and of the dominant theory of natural law (Weber). Maybe more than any other school of thought, American legal realism has exerted a longlasting influence upon Anglo-Saxon sociology of law (US, Canadian, Australian) through its emphasis on 'living law' (i.e., law as it is currently interpreted and applied by the courts) and through the mixture of scientific and pragmatic orientation that characterized an approach largely inspired by the experience that judges had gained of the judiciary decision-making process (Hunt 1978).

The second source of legal sociology can be assigned to the importance attributed to law by the founders of sociology in understanding and explaining society, especially modern society. Accordingly, the origins of legal sociology go back to the origins of sociology itself. Montesquieu's *L'Esprit des lois* (1748) was the first treatise of general sociology as well as of legal sociology. The founding fathers of sociology in the nineteenth century and the beginning of the twentieth century (Marx, de Tocqueville, Weber, Durkheim, Tarde, Ross) were all aware of the central position occupied by law and the legal professionals in the structuring of social life. Law appeared to them as one major source of social control, and as being variously implied in social change. After being so predominent in the works of the founders of sociology, the interest in law declined among succeeding generations of sociologists (almost to the point of its disappearance) —with the exception of criminology. It is mainly since the 1970s that a renewed interest in legal sociology has become evident and that it has been officially reintegrated into sociology.

Modern sociology of law extends in various directions: it is a building with many rooms. But the building lacks a blueprint. The insufficiency of any broad general theoretical scheme to unite the scattered pieces of the whole is often lamented. This article can do no more than reflect this state of affairs.

2. *Law and Social Order*

The founders of sociology were mesmerized by the fact that, despite the individualistic and egotistical tendencies of members of society and the potential reservoir of conflicts between them, there existed something resembling a general social order, to which they turned for some kind of fundamental explanation. In the production and reproduction of the 'conscience collective,' which he regarded as the basis of social order, Durkheim assigned a dominant role to religion and, to some extent, law. Law is partly responsible for the constraint that society imposes upon the individuals. Even though it is generally said that Durkheim did not produce a systematic sociology of law, law is found in practically all his writings (Lukes and Scull 1983) and he is responsible for the place that law occupied in the *Année sociologique* and in the *École française de sociologie* under his intellectual leadership.

In answer to his quest for a concrete symbol and index for the broad concept of social solidarity, Durkheim turned to the law. The two types of social solidarity he identified were represented by two types of law: the mechanical solidarity of traditional societies was associated with repressive law, while the

organic solidarity of the advanced division of labor in industrial societies was associated with restitutive law. Within the realm of repressive law, Durkheim explored the social functions of crime and its punishment through the permanent reproduction of the essential values of the 'conscience collective' and society. Contract, a legal form typical of restitutive law, also received Durkheim's attention. But here again Durkheim insisted on the social dimensions of contract—society being the 'third party' with the rituals, symbols, and general rules it provides to the individual parties.

Because of his views, Durkheim is regarded as the founder of the functionalist approach to law. Through its various functions, law plays a central role in the production and reproduction of different types of social order; the presence of law is an essential factor to explain how society perpetuates itself over time (see *Functionalism in Sociology*).

Durkheim's functionalist approach to law was subsequently transformed into the systemic scheme developed by Parsons. It is the integrative function of law as part of the social system that Parsons emphasized. The systemic approach, however, has also been used to explore in greater detail the relative autonomy of the legal subsystem in its relations with other subsystems within the social system.

3. *Law, Social Class, and Capitalism*

Marxism has exerted a strong influence on modern legal sociology, giving way to various interpretations of law. Beyond its variations, Marxist approach to law is founded on Karl Marx's materialistic model of society and historical evolution. Marx himself had been trained in law and philosophy of law and although, like Durkheim, he did not provide a systematic sociology of law, he addressed law throughout both his work and that in which he collaborated with Engels. In class societies, he argued, law and legal institutions reflect and serve the interests of the dominant class. This is even more the case in capitalist societies where social classes are polarized and class struggle has come to its extreme intensity. Marx foresaw that the proletarian revolution would eventually result in the withering away of both law and the state.

After the Soviet Revolution of 1917, an intense debate took place in the new USSR between those who adhered strictly to Marx's views about the future of law and those who saw law as playing an essential 'parental' role in educating the new socialist citizen (Berman 1963). Ultimately, Marx's predictions were not retained. But his analysis of law in terms of class dominance, class struggle, the production and reproduction of economic and social inequalities has remained at the core of modern Marxian sociology of law. Marxian legal sociology has taken pains to emphasize the role of law in structuring and reproducing the capitalist economy based on the alienation and exploitation of the labor class, in generating the ideological superstructure for justifying the class society and the interests of the dominant bourgeoisie, and in formulating an abstract legal definition of the citizen detached from his or her real life conditions (Cain and Hunt 1979, Michel 1983). Marxian legal sociology has also had a more indirect, but nonetheless real, influence on the various orientations taken by critical approaches to law, whether in European legal sociology (Miaille 1976) or in the US movement called 'critical legal studies' (Unger 1983) (see *Class: Social*; *Marxism in Contemporary Sociology*).

4. *Law, Rationality, and Capitalism*

Among the founders of sociology, Max Weber was the first to elaborate an explicit and quite detailed sociology of law, and to treat law, legal professionals, and legal institutions in a rigorous and systematic way. In what is considered his major work, *Wirtschaft und Gesellschaft*, an entire section was entitled 'Sociology of law' (Weber 1954). Moreover, Weber often referred to law in his great comparative analysis of religion, in his *Economic History*, and in his history of economy and society in antiquity.

Weber made both theoretical and empirical contributions to legal sociology. His theoretical contribution lies in his typology of legal forms of thinking along two axes: the irrational/rational one and the formal/substantive one. These forms produced four types of legal systems: the formal irrational, the substantive irrational, the formal rational, and the substantive rational. In the increasing rationalization that he identified as unique to the Western world, Weber used this theoretical scheme to underline empirically the dominant role played by law. Law evolved toward more and more rational justice, which was an outcome as well as a factor in the process of Western rationalization. At the same time, more rational law contributed to ease the emergence of the capitalist economy by the greater predictability it provided, as well as by the connection established between contractual freedom and the expansion of the market. Weber also emphasized the part played by various types of legal professionals (legal *honoratiores*) in the structure and evolution of law throughout the ages and in the progressive divorce of law from religion, politics, and morals.

It has been argued that a full understanding of Weber's sociology of law requires interpretation within the broader framework of his *Wirtschaft und Gesellschaft*, as well as that of his sociology of religion (Kronman 1983, Coutu 1995). Such is the case because

the roots of the notions of rationality and rationalization and their relationships to capitalism, which are central to his sociology of law, are to be found in his comparative historical analysis of the various types of economic ethics. Max Weber's contribution to the sociology of law is the most systematic and, at the same time, the most broadly and deeply sociological.

5. *Law and Inequalities: The Feminist Sociology of Law*

Feminist studies that have questioned the legal assumption of 'equality of all before the law' and of a value-free justice represent the most vibrant branch of critical legal studies. Feminist legal sociologists, more than any others, have been critical of what has been called 'the official version of law' which they describe as a view of law as being rational, objective, impartial, and neutral (Naffine 1990). It is this view of law that feminist legal studies hold to be mostly responsible for the sexist dominant legal ideology, essentially dictated by a male view of law and humankind (i.e., mankind) (Comack 1999). Feminist legal sociology has been very efficient in pointing out the male stereotyped definition of the subject of law and the unequal treatment of women by the law and its myriad authorities.

Being both theoretically and pragmatically oriented, feminist legal sociologists have documented many of the problems faced by women when dealing with the law and the courts. They have especially emphasized how judges and lawyers tend to use a discourse that dichotomizes reason and emotions, with the consequence that reason is inevitably identified with masculinity and emotions with feminity: male reason becomes opposed to female unreasonableness (Lees 1994). The analysis of how courts have been dealing with cases of rape and homicide of women has served more than any other to highlight the sexist interpretation currently resorted to by police, lawyers, juries, and judges in defense of male offenders (Busby 1999).

More than any other branch of critical legal sociology, the feminist approach has been activist and militant—involved in lobbying to change the law and to influence the courts in favor of a more just treatment of women.

6. *Sociology of the Legal Profession and the Courts*

Following Weber's emphasis on legal professionals, sociology of law has actively investigated the legal profession. Legal sociology joins here with the sociology of professions (see *Professions, Sociology of*). Several studies have dealt specifically with various aspects of the legal profession. Far from being unified and homogeneous, the legal profession has been found to be highly differentiated. The increasing complexity of modern law has given rise to a growing number of specialties within the profession. The variety of the client interests has also been a major factor in differentiating between, for instance, legal practice with corporate clients and that with individual clients, as well as between lawyers working in public services and those working in private offices, in large or small offices, in criminal law or civil law. With differentiation comes a prestige scale within the profession, concomitant with the social prestige of the client (Heinz and Laumann 1982).

The lawyer's relations with capital, their contribution to the evolution of Western capitalism, the mediating functions they take up between state and capital, as well as the role they play in the defense of powerless and socially excluded groups are among the main topics of a more critically oriented sociology of the legal profession. This approach has emphasized the various forms of power enjoyed by the legal profession (Cain and Harrington 1994). An alternative line of research has brought to light the contributions of legal professionals to the emergence of civil society in Western democracies, of political liberalism, and of liberal policies (Halliday and Karpik 1997).

Because judges are members of the legal profession, the functioning of the courts and the jury system has also been the focus of legal sociology. Much of what has been written on this topic has dealt with the judiciary decision-making process, with the explicit objective of improving it (Cramer 1981, Devlin 1979).

7. *Legal Culture and Legal Socialization*

Law is especially susceptible to constructivist interpretations. From this point of view, law is essentially a collective normative symbolic universe of linguistically conditioned mental representations, which are continuously reconstructed through their current use in active life. Social construction and reconstruction is part of law's very nature. Two varieties of such constructions can be distinguished. The first legal culture is that of legal professionals who have been trained in knowledge of the law of a given land and who share the ideological universe of the profession as well as its corporatist interests. The other is the legal culture of lay people and the relative and (more or less) approximate knowledge they have of the law as well as their attitudes and behavior toward the law, its institutions and professionals. The interaction between these two legal cultures varies both in terms of distance and intensity; both are greatest in modern society, following the professionalization of law.

The socialization process into a legal culture is an important chapter in legal sociology. Two major

theories of the psychology of learning have been tested: the cognitive development paradigm based on Piaget's and Mead's works and the social learning paradigm that emphasizes the role of affectivity and environmental context. Legal socialization is one way to explain why a particular individual is either a law-follower or a law-breaker (Cohn and White 1990). At the same time, compliance with the law has proved to be determined or conditioned by other contextual factors, such as rewards and sanctions, regulating agencies, and social environment (Friedland 1990).

It is especially in relation to culture and socialization that the indirect symbolic power of the law (based on nonlegal values and sanctions) can be appreciated, as opposed to its direct instrumental efficiency.

8. *Law, Law Making, and Social Change*

The association between law and social change is complex and many faceted. It is often said that law always lags behind changes taking place in society. A good number of examples can, however, be brought forward to illustrate the fact that law is quite often used by the legislator and sometimes by the courts to bring about various (sometimes radical) kinds of social changes, and that different segments of society must progressively adjust to changes generated by the law. Law can launch social change as well as block or inhibit it (see *Alienation, Sociology of*).

The law-making process in modern democracies has been one vantage point from which to explore the role of law in social change. An important distinction has been made between the instrumental and the symbolic function of law making (Tomasic 1985). It sometimes happens that the legislative process is part of a political strategy aimed more at symbolically exhibiting the legislator's good will than implementing the law.

On the other hand, 'legislative sociology' has emerged as a special branch of legal sociology, which is used to help the legislator enhance the effectivity of the law (Carbonnier 1978). Within this branch, the legal sociologist is explicitly invited to get actively involved in contributing to social change. This approach has also been interrogated as part of what has been called the political sociology of law—an analysis of interstices of the legal and political regulation of society (Commaille 1994).

9. *Legal Pluralism*

An important current of thought which has recently developed and expanded, emphasizes the fact that law is but one of the regulatory agents in society. In this pluralistic approach, law is theorized and empirically analyzed as part of a broader normative system that includes ethics as well as various kinds of technical, professional, and administrative normativities.

A major source of legal pluralism has been anthropological analyses of the meeting of indigenous and colonial laws (Rouland 1988). But even in modern societies, where the legal system has proliferated and become dominant, many other types of social control and social regulation exist alongside the state legal system. This is especially the case with the formal and informal disciplinary procedures currently operating in industrial and bureaucratic organizations (Henry 1983, Arthurs 1985).

Legal pluralism has also served to highlight the ways in which law is produced by the legislator and interpreted by the courts via the incorporation and legalization of rules and norms from other normative systems external to the law itself. It has also underlined how the effectivity of the law is often dependent upon the interaction between formal legal rules and nonlegal principles—between rules and norms.

10. *Conclusion*

Law is never created, applied, or used without some kind of relationship with the various powers in any given society. The most obvious case is, of course, the role played by political power (in the form of 'the state') in any matter involving the law. Political power is the main producer of law, both as legislator and judge. It is political power that guarantees the legitimacy of the law at the same time that the law grounds the legitimacy of the political power: modern society is especially characterized by a fundamental circularity of the foundations of legitimacy between state and law.

Economic power has a less visible, but nonetheless effective, influence upon the law. Big business is generally well organized and has the expertise to lobby legislators and to successfully bring its case before the courts.

In modern democracy, a growing number of interest groups promoting ideological (as well as other) interests have learned how to both influence and use the law. Finally, the media—supposedly the voice of public opinion—also play an active role in influencing legislators and, sometimes, the courts.

In sum, it can be said that in most social conflicts, law is an issue dividing parties as well as a tool they use to promote their interests.

See also: Control: Social; Crime, Sociology of; Justice and Law; Law and Economics; Law and Society: Sociolegal Studies; Law: Anthropological Aspects; Law as an Instrument of Social Change; Law: History of its Relation to the Social Sciences; Legal Culture and Legal Consciousness; Legal Positivism; Legal

Professionalism; Legal Scholarship; Legitimacy, Sociology of; Norms; Political Lawyering

Bibliography

Arthurs H W 1985 *Without the Law: Administrative Justice and Legal Pluralism in Nineteenth-Century England*. University of Toronto Press, Toronto, ON, Canada

Berman H J 1963 *Justice in the USSR. An Interpretation of Soviet Law*. Harvard University Press, Cambridge, MA

Busby R 1999 'Not a victim until a conviction is entered': Sexual violence prosecutions and legal 'truth.' In: Comack E (eds.) *Locating Law: Race/Class/Gender Connections*. Fernwood Publishing, Halifax, Nova Scotia, Canada

Cain M, Harrington C B (eds.) 1994 *Lawyers in a Postmodern World*. New York University Press, New York

Cain M, Hunt A 1979 *Marx and Engels on Law*. Academic Press, London

Carbonnier J 1978 *Essais Sur les Lois*. Répertoire du notariat Defrénois, Paris

Cohn E S, White S O 1990 *Legal Socialization. A Study of Norms and Rules*. Springer-Verlag, New York

Comack E (ed.) 1999 *Locating Law: Race/Class/Gender Connections*. Fernwood Publishing, Halifax, Nova Scotia, Canada

Commaille J 1994 *L'Esprit Sociologique des Lois*. Presses universitaires de France, Paris

Coutu M 1995 *Max Weber et les Rationalités du Droit*. Presses de l'Université Laval, Québec, Canada

Cramer J A 1981 *Courts and Judges*. Sage, London

Devlin P 1979 *The Judge*. Oxford University Press, Oxford, UK

Friedland M L (ed.) 1990 *Securing Compliance. Seven Case Studies*. University of Toronto Press, Toronto, ON

Halliday T, Karpik L 1997 *Lawyers and the Rise of Western Political Liberalism: Europe and North America from the Eighteenth to Twentieth Centuries*. Clarendon Press, Oxford, UK

Heinz J P, Laumann E O 1982 *Chicago Lawyers. The Social Structure of the Bar*. American Bar Foundation, Chicago

Henry S 1983 *Private Justice: Towards Integrated Theorising in the Sociology of Law*. Routledge and Kegan Paul, London

Hunt A 1978 *The Sociological Movement in Law*. Macmillan, London

Kronman A T 1983 *Max Weber*. Stanford University Press, Stanford, CA

Lees S 1994 Lawyers' work as constitutive of gender relations. In: Cain M, Harrington C B (eds.) *Lawyers in a Postmodern World*. New York University Press, New York

Lukes S, Scull A (eds.) 1983 *Durkheim and the Law*. Martin Robertson, Oxford, UK

Miaille M 1976 *Une Introduction Critique au Droit*. François Maspéro, Paris

Michel J 1983 *Marx et La Société Juridique*. Publisud, Paris

Naffine N 1990 *The Law and the Sexes: Explorations in Feminist Jurisprudence*. Allen and Unwin, Sydney, Australia

Rouland N 1988 *Anthropologie Juridique*. Presses universitaires de France, Paris

Tomasic R 1985 *The Sociology of Law*. Sage, London

Unger R M 1983 The critical legal studies movement. *Harvard Law Review* **96**(3): 561–675

Weber M 1954 *Max Weber on Law and Economy* [trans. Shils E, Rheinstein M]. Harvard University Press, Cambridge, MA

G. Rocher

Laws, Conflict of

1. Definitions

1.1 Conflict of Laws

The term 'conflict of laws' describes generally the body of law that seeks to provide solutions to international or interstate legal disputes between private or public persons or entities other than countries or states *as such*. A dispute is considered international, interstate, or multistate if one or more of its constituent elements is connected with more than one country or state. For brevity's sake, the word 'state' is used hereafter to denote any country or any territorial subdivision of such country, such as a state or province, that has its own system of private law. The law of conflict of laws deals with both international (country-to-country or country-to-state) and intranational (state-to-state) cases, so long as they have one or more foreign elements. These elements may relate to the parties' nationality, citizenship, domicile, residence, or other affiliation with another state, to the location of the events that give rise to the dispute, or to the location of the object of the dispute. Thus, a contractual dispute between citizens of different countries or domiciliaries or residents of different states, or a property dispute between residents of one state regarding assets situated in another state, or a tort resulting from conduct occurring in one state and causing injury in another state are all examples of disputes that fall within the scope of this subject.

1.2 Private International Law

The term 'conflict of laws' was coined in Europe by the Dutch jurist Ulricus Huber (1624–1694) and is the term primarily used in the United States, Canada, and more recently in England. In the rest of the world, the corresponding term is 'private international law,' which was coined in the United States by Justice Joseph Story (1779–1845). Neither term is fully descriptive of the subject matter.

The term 'conflict of laws' seems to assume that, in all cases that have contacts with more than one state, each involved state has an active or passive desire, claim, or 'interest' to apply its law, and that these claims 'conflict' in the sense of being of roughly equal validity and intensity. Both of these assumptions have been questioned throughout history. Indeed, even accepting the propriety of using anthropomorphic terms to describe state objectives, both the hypothesis that a state is 'interested' in the outcome of disputes between private persons and the concomitant assumptions that such an interest is implicated in each multistate case or that the application of a state's law indeed effectuates that interest are subject to intense debate.

The term 'private international law' seems less loaded and more descriptive, but also more idealistic than the term 'conflict of laws.' The word 'international' describes an important aspect of the disputes that fall within the scope of this subject—they are international (or interstate) in that they have contacts with more than one state. The word 'private' echoes the civil-law division between private and public law and signifies that only private-law disputes fall within the scope of this subject, that is, disputes arising from relations between persons other than a state or its subdivisions acting in the exercise of governmental authority. In contrast, public-law disputes of an international character, such as those between sovereign countries or other international-law persons, fall within the scope of the public international law, also known as the law of nations (see *International Law and Treaties*).

1.3 The Source of Conflicts Law

The term 'private international law' can be misleading to the extent it gives the impression that there exists a distinct body of law universally observed by most nations which provides solutions to multistate private-law disputes. Reality is much different. Except for a few international conventions that avoid or resolve conflicts through substantive or conflicts rules, international law has little to say on the subject. Thus, for the most part, the task of resolving multistate private-law disputes is left to individual countries, subject to certain mild restraints imposed by international law. Accordingly, conflicts law is essentially *national* law. Similarly, within the United States, conflicts law is—*de facto* and for the most part—*state* law rather than federal law. Although the United States Constitution provides the power to federalize the law of interstate conflicts, this power has not been exercised in any appreciable degree. Thus, the initiative for resolving interstate and international conflicts remains with the states, albeit subject to some federal constitutional, treaty, and statutory constraints.

2. The Source of Conflicts of Laws

2.1 Legal Diversity

Conflicts of laws occur because each state has its own system of private law and because human activity often is not confined to a single state. If either one of these conditions is missing, conflicts do not arise. One who lives in complete isolation and avoids contact with nonlocal residents and foreign products need not be concerned with conflicts of laws. Similarly, one whose contacts with nonlocals are confined to residents of states that have identical laws would not need to be concerned with problems caused by legal diversity. Needless to say, human activity transcends national borders and countries do have different laws. For example the contracts law of France differs from that of England, which in turn differs in some respects from that of Scotland, and the torts law of the Mexican province of Sinaloa differs from that of the State of Arizona, which in turn differs from that of California. Indeed, differences exist not only between countries that belong to different legal traditions, such as France and England which belong to civil-law and common-law traditions, respectively (see *Legal Systems, Classification of*), but also between countries that partake of the same legal tradition, such as France and Italy, or England and the United States, respectively. Moreover, such differences also exist as between or among the component units of a single country, such as the states of the United States, or the provinces of Mexico or Canada, or in the United Kingdom between England and Scotland (see *Legal Pluralism*).

2.2 Interpersonal or Intertemporal Conflicts Law

This legal diversity is usually coextensive with a distinct territory that is administered by a governmental unit with authority over that territory. In some instances, however, the existence of a distinct body of law, such as family law or the law of personal status, is not attached to a particular territory but rather to a particular ethnic or religious group, regardless of territory. Conflicts among these laws are referred to as interpersonal or interreligious conflicts and are not discussed here. Also not discussed here are the so-called intertemporal conflicts of laws, namely conflicts that arise when a law purports to regulate or is sought to be applied to events or transactions that occurred before its enactment (retroactive application).

3. The Three Parts of Conflicts Law

While many problems of conflict of laws are resolved through negotiation, mediation, or arbitration, most of them end up in litigation. In planning for this eventuality, or when forced to confront it, parties involved in multistate activity must keep in mind three major questions: (a) Where can or should litigation be initiated? (b) Which law will the court apply? and (c) Where can the resulting judgment be enforced? These three questions correspond to the three consecutive phases that comprise the process of judicial resolution of most conflicts problems, namely: (a) jurisdiction; (b) choice of law; and (c) recognition and enforcement of judgments. In the United States and other common-law systems, these are also the three major divisions of the law of conflict of laws. In contrast, in civil-law systems, jurisdiction and judgment-recognition form

part of 'international procedural law' rather than of private international law, which deals primarily with choice-of-law and secondarily with the law of nationality and citizenship and the condition of aliens.

The jurisdictional question and the choice-of-law question are analytically distinct questions although in many cases they are practically interrelated. Indeed, the two questions are oriented in different directions and seek to satisfy different ends. The jurisdictional question focuses on the relationship between the defendant and the state of litigation ('forum state') and asks whether that relationship is sufficiently close as to justify subjecting the defendant to litigation in that state and utilizing that state's judicial resources to adjudicate the particular dispute. The choice-of-law question focuses on both parties and their dispute and seeks to identify that state, be it the forum state or another state, whose relationship to the dispute is such as to render most appropriate the application of its law to the merits of the dispute. In many cases, the same relationship that forms the basis of a state's jurisdiction to adjudicate the case can also be the basis for applying that state's law to the merits. This, however, is by no means true in all cases.

As said above, the law of conflict of laws does not deal with public-law disputes, namely, disputes involving a state in the exercise of governmental authority, such as in the enforcement of its criminal laws. Indeed, in such disputes, all three of the above questions are merged into one—the question of jurisdiction. When a state has jurisdiction to try a criminal case, that state applies its own law and enforces the resulting judgment. It does not apply the criminal laws nor enforce the criminal judgments of another state. Cases in which more than one state have concurrent jurisdiction, such as cases involving transborder crimes, are addressed and resolved as jurisdictional conflicts—for example through extradition—rather than as choice-of-law conflicts.

4. *Jurisdiction*

In its broadest sense, jurisdiction is the power of a decision-maker to resolve a dispute in a binding manner. When, as in the context pertinent here, the decision-maker is a court or other tribunal, the dispute is resolved through adjudication and the power to do so is called judicial or adjudicatory jurisdiction. It can be contrasted with legislative or prescriptive jurisdiction, which is the power to enunciate norms governing the merits of the dispute. The source of judicial jurisdiction is the law that constitutes the particular court and thus it is a national or state law rather than a supranational or suprastate law. However, in some countries the exercise of jurisdiction is restricted by international treaties, such as the Brussels-Lugano conventions among several European countries, while in some federal systems like the United States, the exercise of jurisdiction by the constituent states must conform to certain outer limits imposed by federal law, primarily the due process clause of the federal constitution. Although these super-arching sources impose a certain degree of uniformity on the affected states, the jurisdictional laws of the various countries or states continue to differ widely in both nomenclature and substance.

Thus, each state delineates for itself the bases on which its courts are authorized to adjudicate cases with nonlocal elements (jurisdiction in the international sense) and then allocates internally that authority among the various geographical subdivisions or judicial districts within that state (internal or local jurisdiction or venue) and among its various general or specialized courts, tribunals, or other adjudicating agencies in each locality (competence or subject-matter jurisdiction). To render a valid judgment, a court must, in the absence of acquiescence by the defendant, satisfy all three types of jurisdiction.

In the United States, for historical reasons, a state court's interstate or international jurisdiction is divided into personal jurisdiction (*in personam*) and jurisdiction over things (*in rem*). Personal jurisdiction, which is the power to adjudicate claims against a defendant and to bind him or her personally, is further subdivided into general and specific jurisdiction. General jurisdiction exists when the defendant is domiciled, incorporated, has its principal place of business, or conducts 'continuous and systematic' business in the forum state. Historically, the service of process on the defendant within the territory of the court was also a basis of general jurisdiction but the constitutionality of this basis is now questionable, at least when it is not supported by other defendant contacts. When a court has general, jurisdiction, the court may adjudicate any claims against the defendant, even claims unrelated to the defendant's contacts with, or activity in, that state. In contrast, when a court has only specific jurisdiction, the court may only adjudicate claims against the defendant that arise from the defendant's contacts with, or activity in, the forum state.

Jurisdiction over things is based on the presence of a thing within the territory of the forum state and is subdivided into jurisdiction *in rem* and *quasi in rem*. When the lawsuit seeks to determine the parties' rights in, or relating to, that thing, the court has jurisdiction *in rem* and may adjudicate those rights even if it does not have personal jurisdiction over the defendant. In contrast, when the lawsuit seeks to vindicate claims against the defendant that are unrelated to that thing, the court is said to have *quasi in rem* jurisdiction. Since 1977, this basis of jurisdiction is—in the absence of other defendant contacts—constitutionally insufficient unless the claim has been previously reduced to a judgment rendered by another court that had jurisdiction over the defendant.

In other countries, jurisdictional bases are in some respects broader and in other respects narrower than

those described above. For example, some European countries theoretically allow the exercise of what might be called *quasi in rem* general jurisdiction, namely, the adjudication of claims against a defendant who owns property within the forum state, even if the claims are unrelated to the property or vastly exceed its value. On the other hand, while most countries agree that the defendant's domicile, incorporation, seat, or principal place of business in the forum state is a basis for general jurisdiction, they reject general jurisdiction based on either service of process alone or 'doing business' within the forum state.

Beyond this, most countries and states agree that the defendant's actual or imputed consent is a sufficient basis of jurisdiction even when jurisdiction is otherwise lacking, that contractual claims can be adjudicated in the place where the contract was to be performed, and that claims arising from torts can be adjudicated in the place where the wrong was committed or the injury foreseeably occurred.

Because of the differences between the jurisdictional bases recognized by the various states, but also because, even when they do not disagree, these states may authorize more than one jurisdictional basis for certain categories of cases, a given case may often be subject to the concurrent jurisdiction of more than one state. In such a case, the plaintiff, who always has the choice of where to file the lawsuit, may shop for the most advantageous forum, a phenomenon known as forum shopping. One of the mechanisms devised by common-law systems to discourage this phenomenon is the doctrine of forum *non conveniens* which allows a court to decline to adjudicate the lawsuit if, despite the existence of jurisdiction, litigation in that state would be seriously inconvenient and another more convenient forum is available to the plaintiff, Another mechanism which is sanctioned in varying degrees by most systems but which is available only when the disputants are parties to a pre-existing contractual relationship, is a choice-of-forum clause, namely a clause by which the parties agree that disputes arising from that relationship will be litigated in the courts of the agreed upon state.

5. *Choice of Law*

When a court exercises its jurisdiction to adjudicate a case with foreign elements, the court applies its own *procedural* law to the conduct of the proceedings before it. With regard to the merits of the case, however, the court may or may not apply its own *substantive* law. This is the choice-of-law question and it is answered by the forum's choice-of-law rules, which may be established legislatively as in most civil-law systems or through judicial precedent as in most common-law systems including those of the United States. These rules may point to the law of either the forum state or to another state, depending, *inter alia*, on each state's pertinent contacts with the case. For example, in tort cases, these rules may point to the state where the tort was committed or the injury occurred (*lex loci delicti*), in contract cases to the state where the contract was made (*lex loci contractus*), and, in cases involving immovable property, to the state where the property is situated (*lex rei sitae*).

The above three rules are typical of what is usually referred to as a traditional choice-of-law system. For much of the twentieth century, these rules and others like them were more or less uniformly followed in most countries. These rules place a high premium on providing certainty and predictability in the choice of the governing law, leaving little room for judicial discretion. As the above examples illustrate, these rules give no preference to the forum state and indeed they aspire to produce international or interstate uniformity, that is, make it more likely that each multistate case will be governed by the same law regardless of where the case is litigated. The choice of the governing law is based not on the content of the laws of the involved states but rather on the territorial or other contacts of those states to the case at hand. Once a state is found to have the predesignated contact, e.g., the place of the injury, that state's law is applied almost automatically and—save for some limited exceptions—regardless of its content, its underlying policy, or the substantive quality of the result that such application produces. Indeed, the objective of the traditional choice-of-law process is not supposed to be to ensure a substantively 'just' result (material justice), but rather to ensure the application of the spatially appropriate law (conflicts justice).

Traditional choice-of-law rules like the ones described above continue to be followed in many countries, including some parts of the United States. However, in the 1960s, a movement emerged in the United States which has been characterized as a conflicts revolution, and which appeared to reject both the rules and the goals of the traditional choice-of-law process. At least in tort and contract conflicts which are the most numerous, preconceived choice-of-law rules were abandoned in favor of various 'approaches' which, although different in many respects, reject the notion of pre-selecting the applicable law on the basis of a single territorial contact and instead leave the selection to be made by the court on a case-by-case basis. The choice is to be based on multiple contacts and factors, such as the content of the conflicting substantive laws and their underlying policies, the presumed 'interests' or claims of the involved states to apply their respective laws, and the substantive quality of the result that the chosen law will produce in the individual case.

Parallel movements have also appeared in other countries, though nowhere with the same intensity. For example, European systems now appear more willing than in the past to entrust courts with a certain

degree of flexibility in choosing the applicable law. Although most of these systems still rely on statutory choice-of-law rules, the rules enacted in the last quarter of the twentieth century are 'softer' than those of the preceding era. For example, some of these rules base the choice of the applicable law not on a single territorial contact but rather on multiple contacts, usually described by the phrase 'strongest' or 'closest connection,' while other rules contain clauses that authorize courts to deviate from the predesignated law in appropriate circumstances, depending on the needs of the particular case.

As a result of the above movements, the choice-of-law process at the beginning of the twenty-first century is more flexible, less predictable, and more ethnocentric than it was only one generation earlier. The possibility that the forum state will apply its own law to most multistate cases adjudicated by its courts is now greater than in much of the previous century, especially in the United States. In turn, this increases the possibility that a given case will be governed by a different law depending on where it is litigated, and this in turn reinforces the incentive for forum shopping described earlier.

Yet, while national conflicts systems have been gradually diverging in most respects, they have also been converging in at least one important respect. Most of them now recognize the principle of party autonomy, namely the notion that parties to a contractual relationship should have the power to agree in advance on the law that will govern disputes arising from that relationship. Although this is an ancient principle whose origin can be traced as far back as ancient Greece, it has had a checkered history through the centuries. By the second half of the twentieth century, however, this party autonomy acquired the status of a universal principle and is now sanctioned in one form or another by most countries in the world. In fact, some of these countries have extended this principle beyond ordinary contracts and have made it available to matrimonial agreements, other family-law agreements, and even unilateral juridical acts such as testaments.

The use of a choice-of-law clause, especially when combined with a choice-of-forum clause or an arbitration clause, can provide parties with a modicum of predictability. This is a good antidote to the uncertainty generated by the movements described above. This antidote, however, can only work in some cases. For example, the principle of party autonomy is not available in cases such as torts in which the disputants are not parties to a pre-existing relationship. Second, while this principle is recognized in most countries, it is also subject to several limitations and exceptions that differ from country to country. Thus, the possibility that choice-of-law clauses can be enforced in some countries but not in others remains, and with it the possibility of a different outcome depending on where the case is litigated.

6. *Recognition and Enforcement of Foreign Judgments*

One of the many differences between local and multistate cases is the increased possibility that in multistate cases a judgment may have to be enforced in a state other than the one in which it was rendered. This will be the case if the defeated litigant (judgment debtor) does not voluntarily comply with the judgment and does not have sufficient assets in the forum state against which the successful litigant (judgment creditor) can execute the judgment. In such a case, the creditor will seek to have the judgment recognized and enforced in another state that has jurisdiction over the debtor and in which the debtor has sufficient assets. This section discusses the requirements under which the second state will recognize such a judgment.

As between states of the United States, these requirements are prescribed by federal law, primarily the Full Faith and Credit Clause of the Constitution. As interpreted by the Supreme Court, this clause requires each state to give to a judgment of a sister state the same effect that the judgment had in the state of rendition, provided that the judgment is 'final,' it is 'on the merits,' and it was rendered by a court that had jurisdiction under its own law which must conform with federal due process standards. Once these requirements are met, the judgment must be recognized and may not be reviewed on the merits, even if it applied the wrong law or reached a result that contravenes the public policy of the second state. The reason for this strict regime is the national policy of finality embodied in the doctrine of *res judicata*, which seeks to conserve judicial resources and protect party expectations by prohibiting re-litigation of disputes that have been finally decided.

The Full Faith and Credit Clause does not apply to the judgments of foreign countries and thus American courts are not constitutionally compelled to enforce such judgments. Nevertheless, as a matter of comity, American courts recognize foreign-country judgments that meet the requirements for recognizing a sister-state judgment, but also reserve the right to refuse recognition on certain grounds that are not available with regard to sister-state judgments. For example, recognition may be refused if the judgment does not satisfy American jurisdictional standards, even if it does satisfy the foreign-law standards; if the judgment is one for taxes or penalties; or if it was rendered in a system that does not provide impartial tribunals or does not satisfy certain minimum standards of procedural fairness, or contravenes the public policy of the recognizing court.

In many other countries foreign judgments are recognized under similar but somewhat less hospitable standards. For example, in some countries recognition depends on whether the rendering court had applied the law that would have been chosen under the choice-of-law rules of the recognizing court or had reached a

substantially equivalent result. In other countries, recognition is usually denied on public policy grounds to judgments awarding 'excessive' sums of money, especially American judgments imposing punitive damages. Finally, some countries condition recognition on reciprocity, that is, on whether the country of rendition would recognize similar judgments from the second country, while other countries refuse recognition outright in the absence of a treaty requiring such recognition.

One such treaty of increasingly growing importance is the Brussels/Lugano Convention on Jurisdiction and the Recognition and Enforcement of Judgments among the member states of the European Union and the European Free Trade Association. This Convention establishes a comprehensive regime for jurisdiction and judgment-recognition that approximates in important respects the regime that exists within the United States. With regard to defendants domiciled in a member state, the Convention defines the permissible bases of jurisdiction and then provides that a judgment rendered by a court that satisfies those bases must be recognized in the other member states without review of the merits and subject to only limited exceptions such as public policy, lack of proper and timely notice, or conflict with a prior judgment entitled to recognition under the Convention or a judgment of the recognizing state. As of the time of this writing, an international convention on jurisdiction and judgment recognition is being negotiated under the auspices of the Hague Conference of Private International Law which includes countries from all continents.

See also: International Law and Treaties; Legal Systems: Private

Bibliography

Aguillar Navarro M 1982 *Derecho Internacional Privado* [*Private International Law*]. Universidad de Madrid, Madrid, Spain

Audit B 1997 *Droit International Privé* [*Private International Law*], 2nd edn. Economica, Paris

Batiffol H, Lagarde P 1993 *Traité de droit International Privé* [*Treatise on Private International Law*], 8th edn. Librairie générale de droit, Paris

Boggiano A 1993 *Curso de Derecho Internacional Privado* [*Private International Law*], 3rd edn. Abeledo-Perrot, Buenos Aires, Argentina

Castel J G 1994 *Canadian Conflict of Laws*, 3rd edn. Butterworths, Toronto, ON, Canada

Cheshire G, North P 1992 *Private International Law*, 12th edn. Faucett J, Butterworths, London

Dicey A, Morris J 2000 *Conflict of Laws*, 13th edn. Collins Sweet & Maxwell, London

Kegel G, Schurig K 2000 *Internationales Privatrecht* [*Private International Law*], 8th edn. C.H. Beck, Munich, Germany

Leflar R, McDougal L, Felix R 1986 *American Conflicts Law*, 4th edn. Michie, Charlottesville, VA

Mosconi F 1996 *Diritto Internazionale Privato e Processuale* [*Private International and Procedural Law*]. UTET, Turin, Italy

Nygh P E 1995 *Conflict of Laws in Australia*, 6th edn. Butterworths, Sydney, Australia

Rigaux F 1987 *Droit International Privé* [*Private International Law*], 2nd edn. Larcier, Brussels, Belgium

Scoles E, Hay P, Borchers P, Symeonides S 2000 *Conflict of Laws*, 3rd edn. West Group, St. Paul, MN

Symeonides S 2000 *Private International Law at the End of the 20th Century: Progress or Regress?* Kluwer Law International, The Hague, The Netherlands

Vischer F 1982 *Internationales Privatrecht* [*Private International Law*], 2nd edn. Helbing & Lichtenham, Basel, Switzerland

S. Symeonides

Lawyers

Lawyers are defined by function—manipulating rules (usually governmental)—and structure—a profession with credentials based on education, examination, and experience. In the common law world (England and its former colonies) the profession is dominated by private practitioners (sometimes divided into solicitors or attorneys, who do everything but appear in the higher courts, which is reserved for barristers or advocates), though more lawyers are becoming civil servants or corporate counsel. In the civil law world (the rest of Europe and the countries it colonized or influenced) jurists (endowed with a university law degree) are divided into private practitioners (*avocats, abogados, avvocati, Rechtsanwälte*), the magistrature (judges and prosecutors), and lawyers employed by business and the state; these roughly equal categories diverge in qualifications, collective interest, and professional association (Abel and Lewis 1989–95). Movement between them is much less frequent in civil law jurisdictions. These differences articulate with the relative roles of judicial precedent and legislative code as sources of law, judges and legal academics in rationalizing and reforming law, and lawyers and judges in litigation. Most socialist regimes enlarged the role of lawyers in the state (jurisconsults) at the expense of private practitioners. With the fall of communism and globalization of capital, common law professional practices are colonizing the world.

1. Commodifying Law

Lawyers, like law, have social preconditions. There must be a state: in acephalous societies, where power is dispersed among kinship groups and age grades, parties typically represent themselves. Secular power must be differentiated from religious: the priesthood performs legal functions in theocracies. There must be a market: administrators replace most lawyers in command economies. Power must be exercised

through general rules: totalitarian states deploy force rather than law. The amount of law and lawyering vary with the proportions of state and nonstate social control, secular and religious authority, market and plan, legality and autocracy, as well as the distribution of wealth, power, and status (dispersed or concentrated, equal or stratified) and the nature of the economy (goods or services; land, natural resources, capital, and intellectual property).

But even if law is differentiated from informal social control, religion, plan, and diktat, lawyering must be commodified. Most societies value meaning and salvation (religion), health (medicine), physical comfort (engineering) even financial security (accounting) over justice (which is inherently ambiguous—one's justice is another's outrage). Knowledge occupations advance their professional project by capturing vital arenas: church, hospital, factory, stock market, tax office, court, document registry. (This advantages advocacy over transactional lawyering.) Potential clients must be persuaded to buy another's services rather than represent themselves, as we do whenever we comply with or evade, mobilize, or resist law (Kritzer 1998). The division of labor makes generalist consumers dependent on specialist producers (although some societies may resist this in the name of revolutionary ideology or religion, e.g., Wahabi Islam). Lawyers draw upon the devices of other professions—ritual, dress, esoteric language—though they lack some of the most effective (technology, scientific validation). Legal knowledge can effloresce (like science) and decline (like religion). Although law is proliferating, its techniques have been challenged by economics and its legitimacy by legal realism and critical legal studies (including feminism, antiracism, and queer theory). Knowledge, to be commodified, must balance technicality and indetermination (Jamous and Peloille 1970): too technical and it becomes an algorithm consumers perform for themselves (ATMs replacing bank tellers); too indeterminate and the only performance criterion is artistic taste. This balance between theory and craft, symbolic and practical mastery, is related to the tension between university and apprenticeship (Schön 1987). Information technology is restoring some lawyering functions to more educated consumers (both individual and commercial). Constant change in the knowledge base influences the careers and turnover of personnel, professional fission into subspecialities, and the relationship between knowledge producers (teachers and researchers) and consumers (practitioners).

2. Controlling Supply

Once consumers have been persuaded that they need legal services, producers seek to dampen competition by limiting supply. In other markets producers control access to natural resources (South African diamonds, Zanzibari cloves, Chinese silkworms), capital, or technology (patents). Because services are incorporeal, control must be exercised over producers (Larson 1977).

2.1 Production of Producers

Professions use guilds to control the production of producers (Krause 1996). Recruitment processes differ in who selects, when, and by what criteria. Many European law faculties admit all secondary school graduates but choose few to become judges; elite American and British law schools are intensely selective but place most graduates in prestigious firms. In the civil law world the state is the central actor, regulating the number of practitioners (there is still a *numerus clausus* for French *notaires*) and university places (the sole means of qualification), the difficulty of examinations, and employment in the magistracy and civil service. (Controlling supply has become increasingly important to the growing number of private practitioners produced by expanding universities and unable to obtain employment in the public or private sectors.) In the common law world, professional associations successfully erected numerous hurdles: apprenticeship (articles for solicitors, pupilage for barristers, entrants paying substantial premiums to work unpaid for years, barred from other sources of income), through which practitioners themselves limit competition; formal education (secondary school, liberal arts and law degrees, and accreditation of educational institutions); size and length of professional courses; number and difficulty of professional examinations; citizenship; residency; character examinations; even the number of starting positions (tenancies in the Inns of Court, assistant solicitorships in firms). The mass of lawyers sought shelter from competition; the elite were more concerned to elevate collective status. Functionalist social scientists, following professional apologetics, portrayed these barriers as quality assurance (Parsons 1964). Economists (Friedman and Kuznets 1945) and critical social scientists beginning with Weber (1978, see also Freidson 1970, Johnson 1972, Parkin 1979, Collins 1979) accused professions of preventing consumers from trading price for quality while failing to validate the hurdles against performance. Each national legal system insists its unique qualification process is essential for lawyers to perform tasks that vary little across national boundaries. But lawyers never mobilize most of the formal knowledge they must demonstrate to qualify (Genn 1987, Sarat and Felstiner 1995).

Until recently, ascribed characteristics dominated selection. Qualifying was very expensive. Apprenticeship and hiring allowed superiors to indulge class, race, gender, and religious biases. The French *notariat* could be inherited and bought. Indeed, traditional

professions invoked the warrant of character, transmitted by lineage and acculturation and reinforced by collegiality (Inns of Court dinners) and example (during apprenticeship). Entrants conferred status on the profession rather than receiving it from the profession.

Because of its centrality to the state, the legal profession is under particular pressure to reflect societal demographics. Character has been largely supplanted by the warrant of expertise (associated with science), acquired through formal education and demonstrated by examination. Apprenticeship and employment are increasingly meritocratic. But that, too, reproduces inequality, leading some jurisdictions to adopt gender and race targets. Women are half of entrants in many jurisdictions, but they may have accommodated to legal practice rather than transforming it.

Supply control, like any cartel (e.g., OPEC), is constantly eroding. The monopoly rents it extracts increase pressure to enter and undermine the legitimacy of barriers. Novice lawyers, like doctors, engage in conspicuous production, partly to justify future privilege. All meritocratic hurdles can be surmounted. Both the university—synonymous with expertise—and private courses for professional examinations have incentives to increase production. It is difficult to deny mandatory apprenticeships to university graduates (if easier for the 'market' of legal employers to deny them jobs). Globalization undermines barriers to transnational migration. State funding for education becomes the new regulator (alternatively, graduates become virtually indentured to repay heavy debt). The rise of women entrants to numbers equal with men virtually doubled entry (though childrearing leaves created periodic shortfalls) (Epstein et al. 1999). Because of the lengthy qualification process, professions oscillate between over- and under-supply (Freeman 1975). The ratio of lawyers to population, constant or falling for the first half of the twentieth century (partly from exogenous causes like the Depression and world wars), everywhere rose dramatically in the second half (to about 1:300 in the United States). Countries where less than 5 percent of aspirants succeeded (Japan, Korea, and Taiwan) have significantly relaxed their examinations.

Professions not only control the number and characteristics of entrants but also allocate them to roles. Although professions proclaim egalitarianism by virtue of their common qualification, all are stratified (Brint 1994). Civil lawyers traditionally accorded greatest respect to the magistracy, civil service, and professoriat, common lawyers to private practitioners (and advocates in divided professions) and the higher judiciary. Independent practitioners typically have enjoyed higher status than house counsel. With increasing individual inequality and economic concentration, private practitioners fission into two hemispheres, differing in the size and structure of the productive unit, clientele (individual or corporate, poor or wealthy, publicly or privately funded), and rewards (money, status, and power) (Heinz and Laumann 1982). Although the profession invokes meritocracy and 'choice' to justify stratification, parallels with gender and race tend to delegitimate the hierarchy (Epstein 1981, Thornton 1996, Sommerlad and Sanderson 1990).

2.2 *Production by Producers*

Once a profession limits entry it can manage competition among producers in order to enhance both economic rewards and (perversely) status by feigning indifference to material concerns (Abbott 1988). Lawyers secure state-protected monopolies or conclude market-division agreements with other occupations (concerning legal advice, land transactions, probate, authentication of documents, and advocacy) which they defend against 'unauthorized practice.' They allocate markets among legal professions—advocacy to barristers, conveyancing to solicitors. They stratify professional categories (Queen's Counsel and junior barristers). They prohibit foreign lawyers from practicing domestic law or with domestic lawyers. In federal polities, they restrict practice to one jurisdiction. They limit firms to one office. They mandate overstaffing (Queen's Counsel requiring a junior, barristers accepting briefs only from solicitors; conflict of interest rules requiring a lawyer for each party). They prohibit lawyers from seeking business—sometimes even from representing another lawyer's client without permission—and require referral schemes to distribute business equally. They outlaw contingent fees and regulate fees (at least minima). They limit the size of partnerships or forbid them (barristers). They prohibit partnerships with or capital investment by nonlawyers and curtail the activities of lawyers employed by nonlawyers. That some jurisdiction rejects each of these rules argues that none is essential to preserve quality or protect clients.

Like controls on production of producers, these restrictions contain the seeds of their own destruction. Cocooned in the conveyancing monopoly (which generated more than half their income), English solicitors ceded taxation to accountants; proudly litigating, French avocats allowed *conseils juridiques* to capture commercial work. Lured by lawyers' monopoly rents, claims adjusters, immigration advisers, and will writers encroach on legal work. Consumers frustrated by price, inaccessibility or inefficiency look elsewhere: to commercial arbitration or alternative dispute resolution (Dezalay and Garth 1996), accountants or other business consultants (Dezalay and Sugarman 1994), tax and administrative scriveners in Japan, offshore transactions or fora. Large consumers (guided by in-house lawyers) make

firms participate in 'beauty contests'; third-party payors (government, insurers, banks) dictate contractual terms. The erosion of control over production of producers intensifies competition among them. Fractions that had accepted subordination (employed barristers, *conseils*, *juridiques*, *Syndici*) demand equality. Disadvantaged lawyers may evade professional rules (chasing ambulances, advising an employer's clients) or attack them frontally (test cases overturning minimum fee schedules, advertising bans, and limits on number of offices). Although judges are lawyers, their interests and perspectives differ from those of private practitioners.

The conjunction of a movement of consumers (enlarged and emboldened by postwar prosperity), who make demands like price quotations, and the growing hegemony of market ideology undermined restrictive practices. A successful challenge to one erodes the legitimacy of all others. English solicitors, forced to share land transactions with licensed conveyancers and mortgage lenders, challenged the barristers' monopoly of rights of audience. The latter retaliated by extending direct access to professional and even lay clients, bypassing solicitors. When solicitors accused banks of tying conveyancing to mortgage lending, barristers charged solicitors with tying litigation to advocacy. The global economy (embodied in the European Union (EU), NAFTA, and the World Trade Organization) and communication technology have limited the ability of national professions to obstruct foreign competition (notwithstanding differences between the attitudes of exporters, like the common law world, and importers, notably France and Japan).

Restrictive practices are falling ever faster. The professional monopoly has contracted. Internal divisions have diminished or disappeared. Foreign lawyers can practice in more countries, form partnerships with domestic lawyers, and requalify to practice domestic law. Intranational barriers have been lowered or razed. Lawyers can advertise, and many do so aggressively, cutting prices to attract a mass clientele (Seron 1996). Jurisdictions abolished minimum fees, and more permit contingent fees. Public funding encourages competition between private practitioners and employed lawyers. Vertical integration is increasing, e.g., among estate agents, mortgage lenders, conveyancers, surveyors, and even moving companies. Multidisciplinary partnerships intensify competition between lawyers, accountants, and other financial advisers.

Lawyers have invoked new warrants—consumer choice and independence—against the state's power as employer and paymaster. Practitioners discourage competition by withdrawing referrals from foreign lawyers or accounting firms who threaten their turf; banks respond by boycotting solicitors who engage in mortgage lending. Lawyers create subspecialities, defined by additional credentials and endowed with new monopolies (rights of audience in specialized courts, legal aid eligibility). Nevertheless, the real incomes of many lawyers have declined (though those of elite lawyers continue to rise, further differentiating the hemispheres).

3. *Demand Creation*

Producers can respond to market pressures by both curtailing supply and increasing demand. Although law's adversarial nature means that for every action there tends to be an opposite (if unequal) reaction (like the escalation of offensive and defensive military strategies), lawyers displayed little collective interest in stimulating demand until the postwar era. Professional associations initially were suspicious of public subsidies, although significant fractions quickly became dependent on them, and professions studied lawyer use to demonstrate 'unmet legal need.' The rights revolution (in national constitutions and legislation and international treaties) conferred entitlement to legal representation in criminal, if not civil, matters. The erosion of supply control motivated some lawyers to seek clienteles through newly legalized forms of advertising.

Demand for legal services is highly elastic, influenced by the economy (size, concentration, proportions of goods and services, relative importance of natural resources, finance capital, and intellectual property), demographics (age, class composition), governmentality (informal social control, state regulation), globalization, frequency of legal events (divorce, crime, accidents, mergers, bankruptcies, housing sales), and rights consciousness. Lawyers can invent new legal claims (palimony) and transactions (premarital agreements, poison pills against hostile takeovers). But lawyers have never matched doctors and dentists in selling individuals preventive services. Demand creation can provoke backlash against litigiousness, waste, fraud, and lawyer greed. And the substitution of state payment for lawyer philanthropy (services rendered *pro bono publico*, or *pro deo*) can erode the profession's tenuous symbolic capital.

Just as solicitors were seduced into the monoculture of conveyancing, so barristers became unduly dependent on legal aid and thus on a fickle, and increasingly parsimonious English state. Whereas professional efforts to control supply tend to unite the profession (or its subdivisions), efforts to create demand are either collective and largely futile or individual and divisive. They also accelerate concentration among providers. Although lawyers who charge private clients astronomical fees may be envied, 'fat cats' feeding at the public legal aid trough are reviled. Partly because that budget is large and growing, partly because lawyer-bashing elicits populist cheers, the state energetically seeks to control legal

costs: fixing fees, capping budgets, and compelling price competition; shifting the burden to the parties (through contingent fees); substituting salaried lawyers for private practitioners, paralegals for lawyers, and parties for representatives (small claims court); and delegalizing entire domains of social life (divorce, accidents).

There are sharp disagreements about the consequences and desirability of extending legality. Some lawyers persuade clients to virtue, but others create legalistic obstacles and loopholes. Some see legality as essential to economic activity (Weber 1978) or a natural bulwark against state power (Halliday and Karpik 1998). But capitalism antedated commercial lawyers and thrives in countries where they remain marginal (Japan, Korea, Taiwan). Clientelism, corruption, and terror coexist with equal or greater rates of economic growth, and entrepreneurs and their lawyers tolerate, even welcome, autocracy. While some observers extol the formal equality of law, epitomized by the icon of blindfolded justice, others show how formal equality amplifies substantive inequalities (Galanter 1974). Lawyers can aggregate the unorganized (personal injury victims, especially through class actions), but they also engage in union-busting. Some observers applaud the proliferation and enforcement of rights; others blame the victim culture and nanny state for economic stagnation and political correctness.

4. *Relations of Production*

Whereas Weberian and economic theories of professions analyze the distribution of services, Marxist theories focus on relations of production. Extrapolating from the fate of nineteenth-century craftsmen, some have predicted the progressive deskilling of professionals. Certainly lawyers delegate much work to paraprofessionals (Van Hoy 1997), and information technology increasingly substitutes capital for labor (Katsh 1995, Susskind 1996). As multinational firms have grown to thousands of lawyers they also have extended the internal hierarchy among professionals: executive committee, equity and salaried partners, senior, junior, and permanent associates, and contract lawyers. If partnership no longer entails a financial investment, it does require human capital in the form of rainmaking (business-getting). Remuneration is changing from lock-step seniority to 'eat-what-you-kill.' Some firms have hired nonlawyer managers (beginning the separation of ownership and control, whose logical conclusion would be capital investment by nonlawyers). But, as Derber (1982) argues, the knowledge base of all professionals (including lawyers) constantly expands, protecting them from technical proletarianization while increasingly subjecting them to an ideological proletarianization: others determine why they will deploy their expertise, not how. Common law professions have seen the proportion in private practice decline relative to those employed by government and corporations (Spangler 1986). The large-firm practitioner's heavy dependence on a few clients seriously compromises the claim to 'independence' (Nelson 1988). Monopsonistic state consumers make legal aid firms bid for contracts and then sweat profits out of their workers. Elite lawyers also have seen work conditions worsen and hours lengthen. That associates bill approximately three times what they are paid, while few become the partners who profit from this arrangement, has been explained as a tournament that compensates for the difficulty of monitoring the quantity and quality of work, partners leasing surplus reputational capital in a precarious pyramid scheme (Galanter and Palay 1991) and capitalist extraction of surplus value.

Theorists have speculated about the political leanings of the 'knowledge class,' but lawyers, like other professionals, exhibit more differences among themselves than with other occupational groups (Brint 1994) and tends to sing the tune of their paymasters (Derber et al. 1990). Although lawyers are greatly over-represented in most legislatures, they behave little differently from other occupational groups.

5. *Collective Action*

The core of all professions is the claim to collective disinterest (Durkheim 1933, 1957). The information asymmetry engendering professionals make them uniquely capable of protecting consumers. As lawyers multiplied and diversified, informal social control lost efficacy, but professional associations were slow in promulgating and lax in enforcing ethical rules. Lawyers are perpetually torn between client loyalty, personal morality, and obligations to the legal system (Luban 1988, Simon 1997). The dissatisfaction of increasingly assertive and organized consumers is forcing lawyers to improve their complaint mechanisms and exhibit independence from professional associations (perhaps adding lay members) under threat of losing self-regulation altogether. At the same time, external regulation proliferates: malpractice claims (and insurers), courts (supervising litigation), ombudsmen, competition authorities, supranational bodies (EU, NAFTA), and administrative agencies with jurisdiction over the host of ancillary services lawyers are beginning to offer (such as investment, mortgages, estate agency, and insurance). Within the profession, factions struggle over substantive rules (advertising, confidentiality) and the cost of regulation and mutual insurance funds.

Governance is similarly contested. Geographic and functional divides (e.g., barrister–solicitor) are being supplanted by specializations. The refusal of alienated lawyers to join associations forces them to choose

between declining numbers and influence (aggravating free rider problems) and compulsion (limiting their capacity to represent members) (Halliday 1987, Powell 1989). Participation rates are low, belying democratic pretensions. The growing proportion of lawyers in public or private employment and large firms sees associations as irrelevant or hostile. Special interest groups proliferate, based on gender, race, religion, ideology, age, employment, firm size, clientele, dependence on public funds, and functional or substantive specialization. As associations aggressively pursue self-interest (in battles with regulators and oligopsonistic consumers) they abandon the disinterest grounding their claim to self-regulation. Furthermore, the usual forms of industrial action (strike, boycott, go-slow) are ineffective and impair professional status.

The changes sketched above have forced the legal profession to re-evaluate its collective ideology. Large consumers, even individuals represented by third-party payors, disdain professional paternalism. Professional rules have been exposed too often as restrictive practices swelling monopoly rents to retain much public credibility (for instance, contemporary opposition to multinational and multidisciplinary partnerships, or employed advocates). The obscene wealth of highly visible lawyers (which provokes greater resentment than that of entertainers, athletes, or entrepreneurs) has, perversely, undermined the profession's collective status, as shown by the proliferation of lawyer jokes and demeaning media portrayals. Self-regulation appears to favor the accused lawyer over the aggrieved client. Consequently, legal professions increasingly base their legitimacy on 'independence.' But this is an ambiguous value. Lawyers opposing state power on behalf of unpopular clients risk accusations of uncritical loyalty and indifference to crime victims. Few lawyers actually engage in criminal defense, especially of political dissidents, and those lawyers tend to be outsiders, not the establishment (Sarat and Scheingold 1998). If lawyers instead claim independence to pursue justice, clients understandably fear their representatives will seek outcomes (tort damages, divorce settlements, plea bargains) optimizing lawyer, not client, interest. Lawyers are derided as ideological when they invoke personal morality to reject clients or tactics. Legal professionals proclaim their independence by reforming law and legal institutions; but as they move beyond technicalities (where expertise is visibly relevant) to broader social issues—judicial appointments, the death penalty, discrimination, abortion, drug abuse—they risk appearing political. Paradoxically, just as earlier forms of legal professionalism nostalgically invoked aristocratic ideals of *noblesse oblige* and material indifference (Weber's *honoratiores*) just when these were being rendered increasingly anachronistic by the market, so lawyers now proclaim their independence just when it is being increasingly threatened by employment, bureaucracy, subordination to large consumers, and external regulation.

6. *The Future*

Having relinquished control over supply to universities, subcategories are trying to reclaim it, but through credentialing more than licensure. As these specialisms are increasingly differentiated and stratified, legal professions move toward fission. Lawyers find it more difficult to justify restrictive practices, especially as consumers become more concentrated and sophisticated, resorting instead to flexing their own market power. Private efforts to stimulate demand provoke a backlash against litigiousness. States seek to control and even contract their legal aid budgets. Within law firms, productive relations are increasingly capitalist, as differences in control and reward widen. The divisions and hierarchy among lawyers, often mirrored by ascribed characteristics, greatly complicate both self-regulation and governance—the hallmarks of a profession. The professional constellation, which lawyers successfully struggled to attain in the nineteenth and twentieth centuries, is gradually being rendered obsolete by the combined forces of market and state.

See also: Counsel, In-house; Government Lawyers; Judges; Lawyers, Regulation of; Legal Education; Legal Professionalism; Para-lawyers: Other Legal Occupations; Professionalization/Professions in History

Bibliography

Abbott A 1988 *The System of Professions: An Essay on the Division of Expert Labor*. University of Chicago Press, Chicago

Abel R L, Lewis P S C (eds.) 1988–95 *Lawyers in Society: an Overview*. University of California Press, Berkeley, CA, 4 Vols.

Brint S 1994 *In an Age of Experts: The Changing Role of Professionals in Politics and Public Life*. Princeton University Press, Princeton, NJ

Collins R 1979 *The Credential Society: An Historical Sociology of Education and Stratification*. Academic Press, New York

Derber C 1982 *Professionals as Workers*. G. K. Hall, Boston

Derber C, Schwartz W A, Magrass Y 1990 *Power in the Highest Degree: Professionals and the Rise of the New Mandarin Order*. Oxford University Press, New York

Dezalay Y, Garth B G 1996 *Dealing in Virtue: International Commercial Arbitration and the Creation of a Transnational Legal Order*. University of Chicago Press, Chicago

Dezalay Y, Sugarman D (eds.) 1994 *Professional Competition and the Social Construction of Markets*. Routledge, London

Durkheim E 1902 *De la division du travail social*. F Alcan, Paris [1933 *The Division of Labor in Society*. Free Press, New York]

Durkheim E 1957 *Professional Ethics and Civic Morals*. Routledge and Paul, London

Epstein C F 1981 *Women in Law*. Basic Books, New York
Epstein C F, Seron C, Oglensky B, Sauté R 1999 *The Part-time Paradox: Time Norms, Professional Life, Family, and Gender*. Routledge, New York
Freeman R B 1975 Legal cobwebs: a recursive model of the market for new lawyers. *Review of Economics and Statistics* **57**: 171–9
Freidson E 1970 *Profession of Medicine*. Dodd Mead, New York
Friedman M, Kuznets S 1945 *Income from Independent Professional Practice*. National Bureau of Economic Research, New York
Galanter M 1974 Why the 'haves' come out ahead: speculations on the limits of legal change. *Law & Society Review* **2**: 95–160
Galanter M, Palay T 1991 *Tournament of Lawyers: The Transformation of the Big Law Firm*. University of Chicago Press, Chicago
Genn H 1987 *Hard Bargaining: Out of Court Settlement in Personal Injury Actions*. Clarendon Press, Oxford, UK
Halliday T C 1987 *Beyond Monopoly: Lawyers, State Crises, and Professional Empowerment*. University of Chicago Press, Chicago
Halliday T C, Karpik L (eds.) 1998 *Lawyers and the Rise of Western Liberalism*. Oxford University Press, New York
Heinz J P, Laumann E O 1982 *Chicago Lawyers: The Social Structure of the Bar*. Russell Sage, New York and American Bar Foundation, Chicago
Jamous H, Peloille B 1970 Professions or self-perpetuating systems? Changes in the French university-hospital system. In: Jackson J A (ed.) *Professions and Professionalization*. Cambridge University Press, Cambridge, UK
Johnson T J 1972 *Professions and Power*. Macmillan, London
Katsh M 1995 *Law in a Digital World*. Oxford University Press, New York
Krause E A 1996 *Death of the Guilds: Professions, States, and the Advance of Capitalism, 1930 to the Present*. Yale University Press, New Haven, CT
Kritzer H M 1998 *Legal Advocacy: Lawyers and Non-lawyers at Work*. University of Michigan Press, Ann Arbor, MI
Larson M S 1977 *The Rise of Professionalism: A Sociological Analysis*. University of California Press, Berkeley, CA
Luban D 1988 *Lawyers and Justice: An Ethical Study*. Princeton University Press, Princeton, NJ
Nelson R L 1988 *Partners with Power: The Social Transformation of the Large Law Firm*. University of California Press, Berkeley, CA
Parkin F 1979 *Marxism and Class Theory: A Bourgeois Critique*. Tavistock, London
Parsons T 1964 The Professions and Social Structure. In: Parsons T (ed.) *Essays in Sociological Theory*. Free Press, New York
Powell M 1989 *From Patrician to Professional Elite: The Transformation of the New York City Bar Association*. Russell Sage Foundation, New York
Sarat A, Felstiner W L F 1995 *Divorce Lawyers and Their Clients: Power and Meaning in the Legal Process*. Oxford University Press, New York
Sarat A, Scheingold S 1998 *Cause Lawyering: Political Commitments and Professional Responsibilities*. Oxford University Press, New York
Schön D A 1987 *Educating the Reflective Practitioner*. Jossey-Bass, San Francisco
Seron C 1996 *The Business of Practicing Law: The Work Lives of Solo and Small Firm Attorneys*. Temple University Press, Philadelphia, PA
Simon W 1998 *The Practice of Justice: A Theory of Lawyers' Ethics*. Harvard University Press, Cambridge, MA
Sommerlad H, Sanderson P 1998 *Gender, Choice, and Commitment: Women Solicitors in England and Wales and the Struggle for Equal Status*. Ashgate, Aldershot UK
Spangler E 1986 *Lawyers for Hire: Salaried Professionals at Work*. Yale University Press, New Haven, CT
Susskind R 1996 *The Future of Law: Facing the Challenges*. Oxford University Press, New York
Thornton M 1996 *Dissonance and Distrust: Women in the Legal Profession*. Oxford University Press, Melbourne, Australia
Van Hoy J 1997 *Franchise Law Firms and the Transformation of Personal Legal Services*. Quorum, Westport, CT
Weber M 1956 *Wirtschaft und Gesellschaft: Grundriss der verstehenden Soziologie*. J C B Mohr, Tübingen [1978 *Economy and Society: An Outline of Interpretive Sociology*. University of California Press, Berkeley, CA]

R. Abel

Lawyers, Regulation of

The 'regulation of lawyers' includes the regulation of criteria for admission to practice (such as training and testing requirements and criteria for the admission of foreign lawyers); the regulation of lawyers' professional conduct (such as rules governing confidentiality, conflicts of interest, fees, advertising, and so on); and the regulation of lawyers' monopoly over certain types of work (for instance, through interprofessional treaties and unauthorized practice legislation).

Traditionally, such regulation has been written and enforced on a local (state or national) basis, such that the rules varied significantly from jurisdiction to jurisdiction. Regulators and scholars have focused primarily on the regulation of lawyers' individual conduct by way of professional ethics codes.

Recently, however, rapid changes in the legal services market have strained this local, individualistic approach to lawyer regulation. Lawyers increasingly practice in firms, rather than alone or in two-person partnerships; thus their professional conduct increasingly is shaped by firm-level policies, procedures, and culture.

Further, firms themselves have grown exponentially and expanded geographically, such that lawyers increasingly practice across state and national borders. The emergence of a global economy and the growth and globalization of law firms have made the facilitation of 'transnational' practice a central regulatory issue.

Finally, lawyers increasingly practice in firms staffed by other types of professionals, such as accountants and management consultants; and accounting firms increasingly are seeking formal and informal alliances with law firms. The emergence of 'multidisciplinary' practice and its associated organizational forms has created a host of regulatory issues outside of the traditional regulatory paradigm.

Bar associations and other regulators are in the process of responding to the many challenges presented by the changing conditions of law practice. In 1988, the Council of the Bars and Law Societies of the European Community (CCBE) adopted a common Code of Conduct for Lawyers to facilitate transnational practice with in the European Union (Goebel 2000, Terry 1993). State and national bar associations have been hotly debating the regulatory issues raised by multidisciplinary practice—particularly whether lawyers should be permitted to form partnerships with other professionals (e.g., see ABA Commission on Multidisciplinary Practice 2000).

Scholars, too, have begun to call for a rethinking of the traditional regulating paradigm. Rather than focusing only—or primarily—on professional self-regulation by way of ethics codes, scholars have called for increasing attention to other sources of regulation, such as judicial and administrative action, malpractice liability and risk management, and constitutional and international law (Cone 1996, Davis 1996, Schneyer 1998, 1999, Wilkins 1992). In the USA, the field of study that used to be called 'legal ethics' is increasingly referred to more broadly as 'the law governing lawyers.'

Scholars have also called for increasing attention to the organizational contexts in which lawyers practice, and the effects of organizational structure and culture on lawyers' individual conduct (Chambliss 2000, Nelson 1988, Wilkins 1998). Given that lawyers increasingly practice and are socialized within firms, both regulators and scholars have begun to focus on how such firms are managed and what firm-level regulatory controls, if any, should be required (Schneyer 1991, 1998). The emergence of new organizational forms, such as multinational and multidisciplinary partnerships, has only underscored the need to develop a strategy for firm-level (or 'entity') regulation (Chambliss 2000).

Thus, while the 'regulation of lawyers' traditionally has focused on the local regulation of lawyers' individual conduct, both the regulatory system and regulatory scholarship are in the midst of dramatic development and change. As lawyers increasingly practice in firms, across jurisdictional borders, and with other professionals, the regulation of lawyers increasingly will demand collaboration between nations, professions, and regulators. This essay provides a brief overview of regulatory debates for the beginning of the new century.

1. *A Threshold Issue: Defining 'Lawyer'*

Comparing the regulation of lawyers from different jurisdictions requires a shared definition of 'lawyer.' Yet the Anglo-American, common law concept of 'lawyer' has no counterpart in civil law countries because most civil law countries do not have a unified legal profession. Instead, civil law countries recognize multiple legal occupations, such as judge, prosecutor, civil servant, and private practitioner, with private practitioners typically constituting a relatively small proportion of the total (see *Lawyers*). Further, in civil law countries the list of 'legal' occupations includes jobs that in common law countries would be performed by those without legal training, such as notaries (everywhere), police chiefs (in Brazil and Norway), and bailiffs and process servers (in France). None of these groups sees itself as part of a unified legal 'profession'; typically, each group has its own occupational associations and licensing criteria (Abel 1988). Thus, while it is relatively easy to compare 'lawyers' from the USA, the UK, Canada, Australia, and New Zealand, there is no comparable professional group in most other countries.

This article focuses on the regulation of private practitioners in the USA, Europe, and the financial centers of the Pacific Rim (e.g., Australia, Japan). It focuses specifically on regulatory issues arising from the increasing globalization and consolidation of professional services, such as the regulation of transnational and multidisciplinary practice.

Although only a small fraction of all lawyers practice in large, global firms, transnational practice is one of the fastest growing sectors in the legal services market (Abel 1994, Cone 1996, Terry 1999). Likewise, though only a small fraction of lawyers practice in multidisciplinary partnerships, the emergence of multidisciplinary practice raises fundamental questions about the definition of 'legal' services and the boundaries between law and other occupations and professions (e.g., see Abbott 1988, Daly 2000). Regulatory developments regarding transnational and multidisciplinary practice therefore promise to have a significant impact on the organization and regulation of lawyers around the world.

2. *An Overview of the Regulatory System at the End of the Twentieth Century*

Theoretically, the power of self-regulation is a defining feature of the 'professions.' Because there is an 'asymmetry of expertise' between professional and client, clients and nonprofessional regulators may be ill-equipped to monitor professional standards or the quality of professional service. As a result, professionals must regulate themselves—ideally, in view of the public interest. (This justification for self-regulation is associated with 'functional' theories of the professions, see *Lawyers*.)

In the USA and other common law countries with strong national professions, private practitioners historically have enjoyed substantial powers of self-regulation. US lawyers are regulated primarily by way of ethics codes promulgated by state bar associations

and enforced by professional disciplinary authorities. Professional associations control the criteria for admission to practice, including the accreditation of law schools and the substance of bar examinations; and professional ethics codes (though they try to avoid it) shape the standards for malpractice liability. Likewise, in England and Wales, the regulation of solicitors is controlled substantially by the Law Society, a voluntary association formed by solicitors in the 1820s.

In civil law countries, private practitioners are regulated primarily by the state, which controls the number of university places and the difficulty of qualifying exams (see *Lawyers*). Though bar associations may play an important role in promulgating ethical rules, they tend to be subordinate to the state in the disciplinary process. In Denmark, Norway, and Sweden, for instance, professional associations have no coercive authority; disciplinary sanctions must be enforced by the state. In Germany, Italy, and Switzerland, most disciplinary bodies are state agencies or contain only a minority of private practitioners (Abel 1988).

In both common law and civil law countries, the professional regulatory system was established during a time when most lawyers practiced alone and within the boundaries of a single jurisdiction. As a result, much professional regulation is out of step with the conditions of modern law practice—especially the conditions of large firm practice in the global legal services market. Meanwhile, external regulation (by government agencies, malpractice insurers, and international trade organizations) is proliferating. It remains to be seen whether bar associations can adjust to the changing conditions of law practice in time to prevent their own marginalization in the regulatory process.

3. *The Growth of Transnational Practice*

The export of US legal services increased from $97 million in 1986 to $1.9 billion in 1996, while the US import of legal services increased from $40 to $516 million (Terry 1999 [citing US Department of Commerce Statistics]). In both the US and Europe, large law firms increasingly are hiring foreign lawyers, establishing foreign offices, and merging with foreign law firms to create multinational partnerships (Abel 1994, Terry 1999). In 1999, UK firm Clifford Chance merged with US firm Rogers & Wells and German firm Punder, Volhard, Weber and Axster to become the largest law firm in the world (Terry 1999).

The rapid growth of 'transnational' practice has created enormous pressure on bar associations to relax local restrictions on the admission and practice rights of foreign lawyers. In general, however, local regulation has not kept pace with market demands. While some jurisdictions have relaxed restrictions on foreign lawyers in recent years, many continue to impose significant barriers to cross-border practice, and there continues to be significant jurisdictional variation. Lawyers have responded to the complexities of local regulation in part by non-compliance (Abel 1994, Bevernage 1995) and many scholars and regulators urge deregulation (e.g., Goebel 2000). In the coming years, this promises to be an important arena for regulatory reform.

3.1 *Regulation at the End of the Twentieth Century*

All countries regulate the admission of lawyers from other countries. The rules vary enormously, but in general there are two categories of 'admission.' First, foreign lawyers may seek admission to the local profession. Such admission confers full practice rights but typically is difficult to attain. Members of the European Community are required to admit lawyers from other member states subject only to an aptitude test or three years experience, but some countries (France, Spain) have been slow to comply. Outside of the European Community, most countries require foreign lawyers to re-qualify completely—that is, to satisfy all of the criteria for the admission of new lawyers—unless there is a reciprocity agreement between the home and host countries (Abel 1994).

Foreign lawyers also may seek admission to practice as 'foreign legal consultants.' Typically, such admission requires foreign lawyers to: (a) register as foreign legal consultants and pay a registration fee; (b) have three to eight years of experience in their home countries (most countries also demand reciprocity with the home country); and (c) comply with local rules of professional conduct, especially rules governing advertising and fees (Abel 1994). Most countries limit the areas in which foreign legal consultants can practice, reserving for local lawyers such areas as litigation, advocacy, conveyancing, domestic relations, wills and trusts, and advice on local law. In some jurisdictions, foreign lawyers can practice only the law of their home country (Abel 1994).

In general, the European Community offers the most liberal regulatory regime. EC lawyers and law firms may establish a permanent office in any member state and practice most areas of law under their home title. This rule eliminates the need for EC lawyers to become members of the local profession. EC lawyers also may form multinational partnerships with other EC lawyers in any member state (Goebel 2000).

Some EC countries also have liberal rules regarding the admission of non-EC lawyers. England allows non-EC foreign lawyers to set up permanent offices and practice under their home title, as long as they do not practice local law. In 1995, there were over 120 foreign law firms practicing under their home title in London (including over 60 US law firms) (Adamson 1995). Both England and Belgium allow multinational

partnerships between local lawyers and foreign lawyers from outside the EC (Adamson 1995, Bevernage 1995).

Other countries are more restrictive. Until 1980, most US states treated any activity by foreign lawyers as the unauthorized practice of law (Abel 1994) and as of 1998, only 21 states had rules recognizing foreign legal consultants (Needham 1998). New York's rules are the most liberal, allowing foreign lawyers to set up offices under their home firm name and practice local as well as national and international law (Cone 1995). California, on the other hand, prohibits foreign legal consultants from giving advice on anything other than the law of their home country. State-by-state variation makes it difficult for US representatives to negotiate reciprocity agreements with representatives from other countries (Needman 1998). State-by-state variation also makes it difficult for US lawyers who engage in multistate practice domestically.

In Japan, foreign legal consultants must have five years of experience in their home country and that country must offer reciprocity. Foreign legal consultants are prohibited from practicing under the name of their home firm, prohibited from advising on any law other than the law of their home country, and prohibited from hiring or entering into partnerships with local private practitioners. In response to these restrictions, most foreign law firms handle their Japanese business from home (Abel 1994, Myerson 1995). France allows foreign law firms to hire local lawyers and practice local law, but in 1992, France introduced more stringent requirements for non-EC lawyers, which has obstructed the establishment of US law firms (Abel 1994).

3.2 *Likely Twenty-first Century Developments*

The primary justification for restricting the admission and practice rights of foreign lawyers is to insure the quality of professional services. Because 'foreign' lawyers by definition practice in jurisdictions other than the jurisdiction in which they were trained, foreign lawyers may be less competent than local lawyers in matters involving local law. Because clients are presumed to lack the training to judge the quality of professional service, they must be protected, so the argument goes, by professional and state regulation.

Restrictions on foreign lawyers also serve an obvious anti-competitive function, because they restrict foreign competition with the local profession. To the extent that foreign lawyers seek to practice foreign or international law—and to the extent that sophisticated clients manifest a demand for foreign lawyers—restrictive practices are difficult to justify (especially when they are more restrictive than the rules of competing jurisdictions). Restrictive practices that are more burdensome than necessary to ensure the quality of professional services may violate international trade agreements, such as the General Agreement on Trade in Services (GATS) and the North American Free Trade Agreement (NAFTA) (Terry 1999).

In 1999, representatives from 29 national and international bar associations met in Paris to discuss regulatory issues raised by transnational practice, and to consider various strategies for reform. In general, participants in the 'Paris Forum' urged greater cooperation between national bar associations, and greater uniformity in the rules governing foreign legal consultants. In 2000, the American Bar Association established a Commission on Multijurisdictional Practice to reconsider the state-by-state system for licensing lawyers in the US. Concerned in part with remaining competitive with other professionals providing multijurisdictional service (such as accountants), the Commission is considering a variety of proposals to facilitate multistate and transnational practice (including a proposal to establish a national licensing system for foreign legal consultants).

Eventually, the current decentralized system for the regulation of foreign lawyers promises to give way to a more centralized system governed in part by external regulators (such as the World Trade Organization). Meanwhile, however, bar associations are scrambling to respond to the increasing demand for transnational legal services.

4. *The Emergence of Multidisciplinary Practice*

Since the mid-1990s, the 'Big-Five' accounting firms (Arthur Andersen, Deloitte Touche Tohmatsu, Ernst & Young, KPMG, and PricewaterhouseCoopers) have been aggressively expanding their capacity to provide integrated professional services—including legal services—on a transnational basis. They have done so, first, by employing lawyers to perform 'consulting' services: as of 1998, the Big Five collectively employed over 5000 lawyers in the area of tax consulting and approximately 5500 lawyers in areas other than tax (Terry 1999). Ernst & Young alone employs approximately 3300 tax lawyers worldwide (Farrell 2000). Currently, the Big Five accounting firms are among the largest employers of lawyers in the world.

The Big Five also have led recent efforts to establish 'multidisciplinary partnerships' (MDPs) between accountants and other professionals, such as lawyers, investment bankers, and management consultants. MDPs differ from other forms of multidisciplinary practice (including the employment of lawyers as consultants), in that they explicitly offer *legal* services and therefore implicate lawyer regulation (Daly 2000).

The move to integrate legal and other professional services has prompted heated debate among lawyers in the US and elsewhere. Opponents argue that lawyers' participation in multidisciplinary partnerships threatens core values of the legal profession, such as lawyer

independence, competence, confidentiality, and loyalty to clients (NSBA Special Committee 2000). Opponents argue, further, that some 'consulting' services currently offered by the Big Five—such as litigation support, estate planning, and some forms of business planning—constitute the unauthorized practice of law.

Nevertheless, client demand for multidisciplinary services is increasing. Regulation that inhibits lawyers' participation in—and ownership of—firms providing such services threatens to compromise lawyers' competitiveness in this growing and lucrative market.

4.1 Regulation at the End of the Twentieth Century

Only a few jurisdictions expressly allow multidisciplinary partnerships (Germany; Switzerland; New South Wales, Australia; Spain; and, to some extent, the UK), though some countries (Canada, France) appear to be on the verge of allowing them (ABA Commission on Multidisciplinary Practice 2000). Meanwhile, however, various forms of (informal) multidisciplinary practice are proliferating and for the most part have escaped regulation (Daly 2000). The tension between bans on multidisciplinary partnership and the evident increase in multidisciplinary practice threatens to undermine the authority of professional regulation.

In the US, for instance, all states prohibit the formation of multidisciplinary partnerships, if any of the activities of the partnership involve the 'practice of law.' (The District of Columbia is an exception: DC permits MDPs if the *sole* purpose of the partnership is the delivery of legal services.) Yet the enforcement of such regulation depends on a distinction between what lawyers in law firms do and what lawyers in non-law firms do—a distinction that often is difficult to maintain (Daly 2000). When does the 'practice of tax' become the practice of tax *law*? When does business consulting that includes strategies for legal compliance (for instance, with environmental regulation or anti-discrimination law) become the 'practice of law'? Though state bar associations have challenged accounting firms for engaging in unauthorized practice, such challenges have been unsuccessful.

In 1998, the American Bar Association established a commission to study the emergence of multidisciplinary practice in Europe and make recommendations regarding the recognition and regulation of MDPs in the US. After a year of study, the Commission unanimously recommended that MDPs be allowed; but that non-lawyer controlled MDPs be subject to annual certification and audits by state bar associations (ABA Commission on Multidisciplinary Practice 1999). The ABA House of Delegates rejected this proposal and directed the Commission to engage in further study.

In 2000, the Commission again recommended the recognition of lawyer-controlled MDPs, accompanied by the elaboration of firm-level controls (to be defined over time by the states). The Commission noted the increasing demand for multidisciplinary services, especially in the areas of tax, estate planning, and business and management consulting (ABA Commission on Multidisciplinary Practice 2000). The House of Delegates again rejected the Commission's recommendation—and disbanded the Commission. As a result, the US debate over MDPs has shifted to the states.

4.2 Likely Twenty-first Century Developments?

Proponents argue that MDPs offer unique benefits to clients, such as the benefits of one-stop shopping and integrated professional advice; and that clients should have the opportunity to choose multidisciplinary service. Proponents note, further, that many of the ethical challenges associated with MDPs are handled effectively in other settings (such as when lawyers are employed as house counsel); as well as in other jurisdictions that allow MDPs (e.g., see ABA Commission on Multidisciplinary Practice 2000). Most pointedly, perhaps, from the perspective of lawyers' collective economic interests, proponents argue that multidisciplinary practice, in some form, is here to stay; and that the main challenge for lawyers is to remain competitive in a multidisciplinary market.

One potential benefit to lawyers of forming MDPs is that MDPs provide lawyers with greater access to capital for technological development and geographic expansion. Most jurisdictions forbid non-lawyer investment in law firms (though legislation was introduced in New South Wales, Australia that would allow passive investment in law firms, and allow firms to sell shares on the Australian Stock Exchange). Some argue that restrictions on non-lawyer ownership limit law firm productivity and the efficiency of law firm management (Adams and Matheson 1998). MDPs also offer opportunities to 'cross-sell' professional services, and to capitalize on investments in professional reputation (Ribstein 1998).

Abbott (1988, p. 2) argues that the professions 'make up an interdependent system' whose 'boundaries are perpetually in dispute.' The dispute over the recognition and regulation of multidisciplinary partnerships promises to be a fascinating moment in the history of the legal profession. Its resolution will shape the organization of law practice and the boundaries of lawyers' monopoly in both local and global markets for years to come.

See also: Ethical Codes, Professional: Business Codes; Lawyers; Legal Process and Social Science: United States; Legal Professionalism; Professionalization of Social and Behavioral Scientists: United States; Professionalization/Professions in History; Regulation and Administration; Regulatory Agencies

Bibliography

Abbott A 1988 *The System of Professions: An Essay on the Expert Division of Labor*. University of Chicago Press, Chicago

Abel R L 1988 Lawyers in the civil law world. In: Abel R L, Lewis P S C (eds.) *Lawyers in Society: The Civil Law World*. University of California Press, Los Angeles

Abel R L 1994 Transnational law practice. *Case Western Reserve Law Review* **44**: 737–825

Adams E S, Matheson J H 1998 Law firms on the big board?: A proposal for nonlawyer investment in law firms. *California Law Review* **86**: 1

Adamson H 1995 The English system. In: Daly M C, Goebel R J (eds.) *Rights, Liability and Ethics in International Legal Practice*. Transnational Juris Publications, Irvington-on-Hudson, NY

American Bar Association Commission on Multidisciplinary Practice 1999 *Recommendations, Report and Reporter's Notes on the ABA Commission on Multidisciplinary Practice www.abanet.org/cpr/mdpfinalreport.html*

American Bar Association Commission on Multidisciplinary Practice 2000 *American Bar Association Commission on Multidisciplinary Practice Report to the House of Delegates www.abanet.org/cpr/mdpfinalrep2000.html*

Bevernage C 1995 The Belgian system. In: Daly M C, Goebel R J (eds.) *Rights, Liability and Ethics in International Legal Practice Practice*. Transnational Juris Publications, Irvington-on-Hudson, NY

Chambliss E 2000 Professional responsibility: Lawyers, a case study. *Fordham Law Review* **69**: 817–57

Cone S M 1995 The American system. In: Daly M C, Goebel R J (eds.) *Rights, Liability and Ethics in International Legal Practice*. Transnational Juris Publications, Irvington-on-Hudson, NY

Cone S M 1996 *International Trade in Legal Services: Regulation of Lawyers and Forms in Global Practice*. Little Brown, Boston

Daly M C 2000 Choosing wise men wisely: The risks and rewards purchasing legal services from lawyers in a multidisciplinary partnership. *Georgetown Journal of Legal Ethics* **13**: 219–88

Davis A E 1996 Professional liability insurers as regulators of law practice. *Fordham Law Review* **65**: 209–32

Farrell E D 2000 Accounting firms and the unauthorized practice of law: Who is the bar really trying to protect? *Indiana Law Review* **33**: 599–629

Goebel R J 2000 The liberalization of interstate legal practice in the European Union: Lessons for the United States? *International Lawyer* **34**: 307

Green B A 2000 *Assisting Clients with Multi-state and Interstate Legal Problems: The Need to Bring the Professional Regulation of Lawyers into the 21st Century. www.abanet.org/cpr/mjp-bruce_green_report.html*

Myerson T S 1995 The Japanese system. In: Daly M C, Goebel R J (eds.) *Rights, Liability and Ethics in International Legal Practice*. Transnational Juris Publications, Irvington-on-Hudson, NY

Needham C A 1998 The licensing of foreign legal consultants in the United States. *Fordham International Law Journal* **21**: 1126–51

Nelson R N 1988 *Partners with Power: Social Transformation of the Large Law Firm*. University of California Press, Berkeley, CA

New York State Bar Association Special Committee on the Law Governing Firm Structure and Operation 2000 *Preserving the Core Values of the American Legal Profession: The Place of Multidisciplinary Practice in the Law Governing Lawyers*. New York State Bar Association, Albany, NY

Ribstein L E 1998 Ethical rules, agency costs, and law firm structure. *Virginia Law Review* **84**: 1707–59

Schneyer T 1991 Professional discipline for law firms? *Cornell Law Review* **77**: 1–46

Schneyer T 1998 A tale of four systems: Reflections on how law influences the 'ethical infrastructure' of law firms. *South Texas Law Review* **39**: 245–77

Terry L S 1993 An introduction to the European Community's legal ethics code Part I: An analysis of the CCBE Code of Conduct. *Georgetown Journal of Legal Ethics* **7**: 1–87

Terry L S 1999 An introduction to the Paris Forum on Transnational Practice for the Legal Profession. *Dickenson Journal of International Law* **18**: 1–33

Wilkins D B 1992 Who should regulate lawyers? *Harvard Law Review* **105**: 801–87

Wilkins D B 1998 Everyday practice is the troubling case: Confronting context in legal ethics. In: Sarat A, Constable M, Engel D, Hans V, Lawrence S (eds.) *Everyday Practice and Trouble Cases*. Northwestern University Press, Evanston, IL

E. Chambliss

Lazarsfeld, Paul Felix (1901–76)

Paul Felix Lazarsfeld was born in Vienna, Austria, on Feb.13, 1901 to a middle-class professional couple—his father was a lawyer and his mother an Adlerian psychologist—and died in New York City on Aug 30, 1976. During his Vienna youth he was active in the political and intellectual milieu of Austrian socialism and was a leader in a number of socialist student organizations. He earned a Ph.D. in mathematics from the University of Vienna in 1925 and became Charlotte Bühler's assistant at the Psychological Institute. With the Bühler's backing he established the Wirtschaftspsychologische Forschungsstelle (Institute of Economic Psychology). Its best-known accomplishment was the study of an industrial village with long-term, massive unemployment (Jahoda et al. 1971).

As antisemitism was intensifying in Austria, Lazarsfeld arrived in the US in 1933 with a Rockefeller Foundation fellowship and decided to remain when it was extended for a second year. From the late 1930s, in a succession of positions at the universities of Newark, Princeton, and Columbia, and in research institutes and projects he led and/or founded, the Office of Radio Research and the Columbia Bureau of Applied Social Research, Lazarsfeld and his associates virtually created the field of mass communications research together with new methodologies and techniques in empirical social research (1940). In 1940, he was appointed to the faculty at Columbia University where he served until his retirement in 1969. At Columbia he initiated a program in political sociology marked by two pioneering longitudinal studies of the

1940 (*The People's Choice* 1944) and the 1948 US presidential elections (*Voting* 1954), of opinion leadership (1955), and of the effects of McCarthyism on higher education (1959). Methodological innovations such as panel (longitudinal) analysis, snowball sampling, and contextual analysis were pioneered and developed in this research. During this time the Bureau of Applied Social Research became the premier academic social research institute in the US.

In the 1950s Lazarsfeld and his collaborators turned to the codification and systematization of these methods which resulted in many journal articles and edited volumes (*Language of Social Research* 1955; *Continuities in the Language of Social Research* 1972). Later still he focused on the uses and applications of social research (*Introduction to Applied Sociology* 1975). He was convinced that training, methodology, implementation, and utilization of social research in a professional manner were best carried in organizational settings such as those he had created at the Bureau rather than as solo pursuits by individual scholars, and that these organizations should be aided financially and administratively attached to universities (1962). He himself participated in creating such institutes elsewhere in Washington DC, Oslo, and Vienna.

Although known primarily as a methodologist, Lazarsfeld had a theoretical vision and program he called the empirical analysis of action (1955). It was inspired by Max Weber's (see *Weber, Max (1864–1920)*) action theory, methodological individualism, and the operation of '*verstehen*,' framed in a realistic, empirical social psychology (*Max Weber* 1965). Its foundation was the choice studies of Lazarsfeld and his associates, on buying, voting, occupational choice, choice of friends, moving residence, and others (1959). Lazarsfeld's studies in the history of social research found precedents for a sociological theory built from the ground up on empirical foundations rather than top down abstract philosophyzing (1961).

Throughout his life, he pursued other interests and lines of work, notably mathematical sociology (*Mathematical Thinking in the Social Sciences* 1954), scaling techniques, in particular latent structure analysis which he invented (*Latent Structure Analysis* 1967), and multivariate statistical analysis with attribute variates. He had a close professional relationship with Samuel Stouffer. At Columbia, he became in 1962 the first Quetelet Professor of Social Sciences. He had a unique 30-year working and professional relationship with his colleague the theorist Robert K. Merton (Merton 1998). Many leading sociologists in successive cohorts of students at Columbia were educated, trained by, and collaborated with these two outstanding scholars. Lazarsfeld's personal style of doing social science was to recruit capable students and colleagues to work with him and to co-author books and articles. During his career, he received many honors, including honorary degrees from the universities of Columbia, Chicago, Vienna, and the Sorbonne. He was president of the American Sociological Association and the American Association for Public Opinion Research. He was a member of the National Academy of Sciences and the National Academy of Education. He was an institution builder who forged transatlantic ties by inviting European scholars to Columbia, and who served frequently as a visiting professor in Europe, helping create research institutes and programs in France, Poland, Norway, and Austria.

1. *Lazarsfeld's Methodology*

He distinguished philosophizing about social knowledge and how-to-do-it techniques from methodology, which is a systematic analysis and codification of what social scientists were doing on data collection, measurement, and drawing inferences from findings, i.e., explication and critique of the work of social scientists. Methodology was a creative act: 'intuition reconsidered in tranquility.' He identified four steps in measurement:

(a) imagery, or the sifting through intuition and disparate findings for an articulate idea. The goal was 'depth of conceptualization' (1959);

(b) concept specification and dimensionalizing;

(c) selection of indicators for each dimension; and

(d) the combination of indicators into an overall measure or index.

Indicators have a probabilistic relationship to an underlying dimension. With several indicators, however, the probability of describing a respondent correctly increases; thus, multiple indicators should be used (1958). Though the choice of indicators from a large pool is somewhat arbitrary, and though some respondents' answers will change depending on specific indicators, findings in the aggregate will remain the same as will the correlation between variables. This principle of interchangeability of indices removed an important challenge to internal validity in survey research.

Other advances in method included the 'accounting scheme' and 'reason analysis' for getting at an assessment of why people make certain choices by measuring dispositions and attributes, exposure to various influences, and the respondents assessment of their importance (*Analysis of Reasons* 1942).

In Lazarsfeld's approach to survey research much attention was expended on getting these four steps right and on creating a questionnaire that captured the process accurately. During the research that led to *The Academic Mind* (1958), the questionnaire went though eight versions over six months as the central 'apprehension' variable was sorted into a 'worry' and a 'caution' dimension in pretests, and indicators and wording appropriate for a faculty population ranging from small religious colleges to major research universities were carefully crafted. Lazarsfeld liked to end up

with a 'pivotal' variable that he said was 'less than a theory but more than a label for just one or two measurements' (1959). It organized a range of findings into a key concept or typology that provided deep insight into the topic and linked the research to other social science research. In Jahoda et al. (1971), the pivotal variable was apathy or resignation.

Although identified with quantitative techniques, Lazarsfeld favored combining information from a variety of sources. In Jahoda et al. (1971), the researchers made use of participant observation, census statistics, election results, library circulation records, newspaper circulation, school essays, family records, life histories, psychological tests, expert reports, even measurements of the walking speed of villagers. Mindful of critics who argued that survey research forced an individualist straightjacket on sociological topics, Lazarsfeld created contextual analysis. Groups, social units, organizations, and social milieus were characterized by global, relational, and aggregated variables called collective properties. These enabled hypotheses to be researched on social influences, climates of opinion, and other sociological processes (1961, *Continuities in the Language of Social Research*, 1972). Other methodological innovations that came out of the Lazarsfeld–Columbia tradition were deviant case analysis (*Language of Social Research* 1955), concept of property space and substruction for systematizing typologies, and snowball sampling (*Personal Influence* 1955).

2. Data Analysis and Statistical Inference

Lazarsfeld presented findings in a simple, understandable form to a wide public: two and three variable contingency tables and bar charts. He favored data collected in natural settings to controlled experiments carried out in a laboratory. Discrepancies between findings in survey research and in controlled experiments in communications research he attributed to selection effects, i.e., in natural settings people exposed themselves selectively to information and influences they tended to agree with. To compensate for the lack of control groups in natural settings, he developed a multivariate analysis derived from the English mathematical statistician G. Udny Yule, a pioneer in the analysis of covariance for categorical data (*Continuities in Social Research* 1950). The complex effect of three variables on one another is decomposed into a weighted sum of two variable effects. Starting with a hypothetical two variable relationship, a third 'test' variable is introduced to check whether the original relationship is spurious, contingent on the level of the test variable, or exhibits other patterns called specification and interpretation (*Language of Social Research* 1955, Sect. 2, pp. 111–66). Survey analysis is the art of probing hypothesis with test variables for ascertaining the conditions, contexts and domains, and extent to which they are true. In the 1940s and 1950s, computers were not available and statistical operations were performed on IBM countersorters using IBM data cards. For quantitative variables, desk calculators took much time and concentration to grind out even the simplest correlations. Thus, survey analysis seldom went beyond the simultaneous study of three variables. Although analysis of findings for causal patterns is nowadays done with multivariate regression techniques with statistical packages in computers in a matter of minutes, the logic of explanation in quasi-experimental design has remained the same.

In the Lazarsfeld's quasi-experimental designs, the time order of the variables was crucial because it ruled out some patterns of causation. Repeated observation on the same units, called 'panel' analysis and known today as longitudinal sample design pinned down a specific time order and became central in major Lazarsfeld studies. The 1940 election campaign study (*People's Choice* 1944) was based on five waves of interviews with the same respondents over several months. The method allowed a deep understanding of how various groups and types of voters formed political opinions and choices as a result of campaign, media and personal influences. Lazarsfeld (1955) realized that the study of long term effects, e.g., the cumulative impact of television viewing on children's attitudes and behavior, was beyond both the experimental and quasi-experimental methods then current.

3. Mass Communications and Political Sociology

Lazarsfeld and associates undertook two major election campaign studies, in 1940 (*People's Choice* 1944) and in 1948 (*Voting* 1954), both using a longitudinal design of repeated interviews with potential voters over several months in typical towns, and a study of opinion leadership among 800 women about marketing, fashion, movies, and public affairs (*Personal Influence* 1955) which used snowball sampling enabling influentials repeatedly named by the initial sample to be interviewed in turn. The findings challenged the conventional wisdom on democratic theory and on the mass media.

Political opinions and voting are only in part based on informed, calculated interest reflecting the arguments of parties, platforms, and candidates. Voting and opinion are social validations of group membership by conformity to social milieus and family tradition. Campaigns reinforce existing views rather than change them, reactivate those who were disengaged from politics, and foster turnout and participation. Much of the public have their minds made up early in the campaign. The least informed and those under cross-pressures from multiple group memberships delay making choices, are the most likely nonvoters, and are the most likely to switch their vote

between parties. Lazarsfeld's discoveries challenged the common notion that democracy would only work with an informed and enlightened public.

The impact of the mass media on voting is modest. Voters expose themselves selectively to views they tend to agree with. Moreover, in what he called the two-step flow of communication, people seek political guidance from opinion leaders in proximate social milieus—at work, in neighborhoods, in families—who are a transmission belt for public affairs and attitudes from the political class to the citizenry. Interpersonal communication is flexible and casual, from a trusted source, which makes it more effective than mass communications. Opinion leaders are found in every group and social stratum, and reinforce the dominant views of these social milieus. Thus, there is no large political influence from elites and top strata down to the lower strata. The impact of the mass media is limited and interpersonal communication has a conserving and differentiating effect on political opinion, attitudes, and choices. These findings challenged the views of many intellectuals, culture critics, and mass society theorists.

Lazarsfeld was open to criticism and was his own critic as well. On political sociology, he knew that he had neglected the institutional side of politics—political parties as organizations, platforms, choice of agendas. In mass communications, he regretted not studying the 'inside of the communications industry.' There were also missed opportunities: the pioneering campaign studies were not followed with studies of national campaigns that the University of Michigan's Survey Research Center undertook. And when television became the dominant mass medium, Lazarsfeld and the Bureau failed to develop a research program that was both acceptable to and supportable by the foundations and the networks.

4. *Lazarsfeld's Influence*

His greatest influence and lasting legacy is on social science methodology, on survey analysis, mass communications theory and research, and political sociology and political science, in which he created entirely new fields of scholarship, as has been described above. Public opinion polling and research has diffused worldwide. In a broad sense his research highlighted the importance of small groups and proximate social milieus, and personal influences on attitudes, opinions, and behaviors which sheltered people from macrosociological forces. Much of what Lazarsfeld accomplished has been incorporated thoroughly into social science and is not even referred to or footnoted. This is true not just in academic disciplines and university settings but in journalism, business schools, communications departments, market research, policy studies, and practitioners and professionals in related fields. Though he thought of himself as a marginal man in academia, both in Vienna and in the US, which he overcame with an entrepreneurial style of action combining intellectual and administrative innovation, he became internationally influential and recognized. His legacy and thinking live on in the pursuits and accomplishments of his colleagues, collaborators, associates, and students. At one time or another, and sometimes over many years, he was working with Robert Merton, Marie Jahoda, Hans Zeisel, Bernard Berelson, Patricia Kendall, Samuel Stouffer, Alan Barton, Elihu Katz, Ernest Nagel, Mira Komarovsky, Robert Lynd, C. Wright Mills, Arthur Kornhauser, Theodor Adorno (see *Adorno, Theodor W (1903–69)*), Morris Rosenberg, William McPhee, Frank Stanton, David Sills, Edward Suchman, Herbert Menzel, Leo Lowenthal, Joseph Klapper, Marjorie Fisk, Neil Henry, Sam Sieber, Wagner Thielens, Jr., Seymour Lipset, James Coleman, (see *Coleman, James Samuel (1926–95)*), Juan Linz, Herbert Hyman, Raymond Boudon, Bernard Lecuyer, Anthony Oberschall, and others. He was a man of broad learning and enormous energy who, in addition to social science was well-versed in history, the humanities, and the philosophy of science.

See also: Campaigning: Political; Large-scale Social Survey, Analysis of; Longitudinal Data; Mass Communication: Empirical Research; Mass Media: Introduction and Schools of Thought; Media and Child Development; Multivariate Analysis: Overview; Panel Surveys: Uses and Applications; Quasi-Experimental Designs; Sample Surveys: Methods; Sample Surveys: The Field; Voting, Sociology of

Bibliography

Berelson B, Lazarsfeld P F, McPhee W 1954 *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago University Press, Chicago

Jahoda M, Zeisel H, Lazarsfeld P F 1971 *Die Arbeitslosen von Marienthal*. Hirzel, Leipzig, 1933, Frankfurt, 1961 (new introduction by Lazarsfeld P English trans. Marienthal: The Sociography of an Unemployed Community). Aldine, Chicago

Katz E, Lazarsfeld P F 1955 *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Free Press, Glencoe, IL

Lautman J, Lecuyer B P (eds.) 1998 *Paul Lazarsfeld (1901–1976)*. L'Harmattan, Paris

Lazarsfeld P F 1940 *Radio and the Printed Page*. Arno Press, New York

Lazarsfeld P F 1942 Statistical analysis of reasons as a research operation. *Sociometry* **5**: 29–47

Lazarsfeld P F 1955 Why is so little known about the effects of television on children and what can be done. *Public Opinion Quarterly* **19**: 243–51

Lazarsfeld P F 1958 Evidence and inference in social research. *Daedalus* **87**(4): 99–130

Lazarsfeld P F 1959 Methodological problems in empirical social research. *Proceedings of the Fourth World Congress of*

Sociology. International Sociological Association, Louvain, Belgium, Vol. 2
Lazarsfeld P F 1959 Reflections on business. *American Journal of Sociology* **65**(July): 1–31
Lazarsfeld P F 1961 Notes on the history of quantification in sociology—trends, sources and problems. *ISIS* **52**(2): 147–203
Lazarsfeld P F 1962 The sociology of empirical social research. *American Sociological Review* **27**(6): 757–768
Lazarsfeld P F 1970 Sociology *Main Trends of Research in the Social and Human Sciences*. UNESCO, Paris, Chap. 1, 61–165
Lazarsfeld P F, Barton A 1951 Qualitative measurement in the social sciences: classifications, typologies and indices. In: Lerner D, Lasswell H D (eds.) *The Policy Sciences*. Stanford University Press, Stanford, CA, pp. 155–92
Lazarsfeld P F, Berelson B, Gaudet H 1944 *The People's Choice*. Columbia University Press, New York
Lazarsfeld P F, Fleming D, Bailyn B (eds.) 1969 An episode in the history of social research: A memoir. In *The Intellectual Migration*. Cambridge University Press, Cambridge, MA, Chap. 6
Lazarsfeld P F, Henry N 1968 *Latent Structure Analysis*. Houghton Mifflin, New York
Lazarsfeld P F, Menzel H 1961 On the relation between individual and collective properties. In: Etzioni A (ed.) *Complex Organizations: A Sociological Reader*. Henry Holt, New York, pp. 424–40
Lazarsfeld P F, Merton R 1950 *Continuities in Social Research*. Free Press, Glencoe, IL
Lazarsfeld P F, Oberschall A 1965 Max Weber and empirical social research. *American Sociological Review* **30**(Spring): 185–99
Lazarsfeld P F, Pasanella A, Rosenberg M 1972 *Continuities in the Language of Social Research*. Free Press, New York
Lazarsfeld P F, Reitz J, Pasanella A 1975 *An Introduction to Applied Sociology*. Elsevier, New York
Lazarsfeld P F, Rosenberg M 1954 *Mathematical Thinking in the Social Sciences*. Free Press, Glencoe, IL
Lazarsfeld P F, Rosenberg M 1955 *The Language of Social Research*. Free Press, Glencoe, IL
Lazarsfeld P F, Sills D 1979 In *International Encyclopedia of the Social Sciences*. Free Press, New York, vol.18
Lazarsfeld P F, Thielens W Jr. 1958 *The Academic Mind*. Free Press, New York
Merton R K, Coleman J, Rossi P (eds.) 1979 *Qualitative and Quantitative Social Research, Papers in Honor of Paul F. Lazarsfeld*. Free Press, New York
Merton R K 1998 Working with Lazarsfeld: notes and contexts. In: Lautman J, Lecuyer B P (eds.) *Paul Lazarsfeld (1901–1976)*. L'Harmattan, Paris

A. R. Oberschall

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Lea

Leadership in Organizations, Psychology of

The primary focus of leadership in organizations is on managerial leadership as opposed to religious, political, and educational leadership, or informal leadership in peer groups.

1. *Leadership in Organizations*

Leadership in organizations is commonly defined as having and being seen to have the ability to influence, motivate, and enable others to contribute toward the effectiveness and success of a working group or an organization of which they are members. This conception is an essential component of most of the numerous definitions of organizational leadership that have been proposed in the literature (Yukl 2001). The same conception also emerged as a universal definition from discussions among scholars from about 60 countries worldwide as part of the Global Leadership and Organizational Effectiveness (GLOBE) cross-cultural research project (House et al. 1997).

The many definitions of leadership reflect serious disagreement, especially about the identification of leaders and leadership processes. Researchers who differ in their conception of leadership are likely to investigate different phenomena and to interpret empirical results differently. Yukl (2001) argues that at this point, it is neither feasible nor desirable to attempt to resolve the controversies on the appropriate definition of leadership. Leadership research should be designed to provide information relevant to the entire range of definitions.

Of equal importance is that particular initial definitions of leadership should not predetermine the answer to the central questions investigated in organizational leadership research, such as 'How to identify leaders?,' 'What are leaders doing?,' and 'What makes a leader effective?' Taking the identification of leaders as an example, a major proportion of leadership research was concerned with identifying personality dispositions, individual values, motives, and behavioral skills that would differentiate leaders from nonleaders or effective leaders from ineffective leaders. Smith (1995), speaking of leadership from a social psychological perspective, defines leadership as 'a quality attributed to people as a result of their interrelations with others.' This definition implies that leadership is not inherent in people (e.g., personality characteristics like intelligence or strength of power motives) or in the positions that they occupy (e.g., role scripts and specifications of rights and duties attached to a particular leadership position). Personal characteristics and peculiarities of positions should be seen as conditions that facilitate or inhibit the expression of leadership and its effectiveness, and as such, they are matters for scientific discovery rather than defining constituents of the terms 'leader' and 'leadership.' An example in the field of leadership processes is that some theorists argue that individuals who rely on coercion and manipulation for influencing followers are not really 'leading' them if leadership is understood in terms of exercising influence which results in fervid commitment by followers. This conception of leadership, however, is too restrictive in situations when force and manipulation are recognized as legitimate means of exerting beneficial influence on followers due to, for example, corresponding cultural values, or the followers' later assessment that the use of coercion and manipulation in particular cases was the best option.

2. *Context of Inquiry*

Throughout history, proper identification of effective leaders has frequently determined the survival or demise of groups, organizations, and entire societal cultures. Thus, it is no wonder that leadership has been of concern to the foremost thinkers and social philosophers in history (e.g., Plato, Machiavelli). Today, identifying and developing effective leaders are major concerns of industry (and the military).

The scientific research on leadership did not begin until the twentieth century. It is informed by conceptions and methods of science, that is, terms are defined in observable and measurable ways, theoretical assumptions are tested with objective evidence and the desire to find general principles guides theory development as well as the design of empirical studies. Leadership research mainly focused on factors that determine how well a leader is able to influence followers and to accomplish group and organizational objectives (i.e., leadership effectiveness). Also of interest were the reasons why certain persons emerge as leaders, the determinants of the way leaders act, and the perception of leaders and leadership in the 'eye of the beholders.'

3. Approaches to the Study of Leadership

Three major categories of leadership research can be distinguished based on where they place their emphasis for explaining leadership effectiveness:

(a) leader oriented trait and behavior approaches;

(b) situational or contingency approaches that incorporate situational factors; or

(c) approaches that address power and influence processes.

3.1 Trait and Behavior Approaches

The trait approach, focusing on a variety of relatively stable individual dispositions of personality, temperament, motives, and values, was dominant in early leadership research. However, the early trait approaches did not reveal a particular set of universally relevant traits to be successful (Stogdill 1948). Recently, personality traits have regained credibility in leadership research through advance in personality assessment. Today, traits such as high energy level, high self confidence, internal locus of control, emotional maturity, and integrity appear to be positively related to leadership effectiveness (cf. Yukl 2001).

A widely accepted research program on managerial motives was introduced by McClelland (1965) who suggested identifying unconscious individual motives with the projective Thematic Apperception Test (TAT) (see *Projective Methods in Psychology*). The optimal pattern of managerial motives, namely a moderately high need for achievement, a strong (socialized) power orientation, and a low need for affiliation, was shown to correspond with managerial advancement in various studies and somewhat less clearly with managerial effectiveness.

The research focus has changed from abstract traits to particular personal attributes that relate more directly to specific behaviors required for effective leadership, termed 'skills,' that is, the ability to do something in an effective manner. Leadership skills can be classified into technical skills concerned with things, conceptual skills concerned with ideas and abstractions, and interpersonal skills concerned with people (cf. Yukl 2001). Particular leadership traits and skills increase the likelihood of success but they don't guarantee it. For developing sound theories about how traits and skills affect leadership effectiveness, we need to understand how traits and skills are expressed in actual behavior.

Leadership behavior has been more widely researched than any other aspect of leadership. Throughout the 1950s and 1960s descriptions of leadership behavior and styles and their relationship to followers' attitudes and performance were investigated. Research at Ohio State University sought to classify relevant aspects of leadership behaviors and found that subordinates perceived the behavior of their leaders primarily in terms of two independent dimensions, initiating structure (i.e., task-oriented behaviors) and consideration (people-oriented behaviors). The questionnaire used in this research, called the Leader Behavior Description Questionnaire (LBDQ), is a hallmark in the history of leadership research. Various theoretical models capitalized on these two conceptions of leadership style, however, they perceive them to be two antagonistic poles of one dimension (bipolar models), by contrasting for example, theory X vs. theory Y, democratic vs. autocratic, participative vs. directive, or task vs. group maintenance oriented leadership behaviors.

A taxonomy of a variety of specific leadership behaviors was suggested by Yukl and associates (cf. Yukl 2001) distinguishing between task behaviors (e.g., clarifying role expectations), relations behaviors (e.g., mentoring), and transformational behaviors (e.g., empowering people to implement new strategies). Specific leadership behaviors seem to be better predictors of leadership effectiveness than abstract dimensions. For example, the cumulative empirical evidence speaks for 'clarifying role expectations,' in particular, the setting of specific challenging but realistic goals, as strongly predicting leadership effectiveness on followers' motivation and performance (Locke and Latham 1990).

Trait and behavioral approaches have the tendency to look for simple answers to complex problems. Particular leadership styles and behaviors may be effective only in certain situations or at some point in the leadership process but not at others. For example, monitoring behavior is useful for discovering problems, however, unless something is done to solve problems when discovered, it will not contribute to leadership effectiveness. By looking at situational conditions and patterns of behavior rather than at traits and isolated behaviors it is possible to predict leadership effectiveness in organizations to a higher extent.

3.2 Situational and Contingency Approaches

Situational approaches emphasize the role of contextual factors and how they either influence leadership behavior directly or moderate the relationship between leadership and measures of leader effectiveness (i.e., contingency). Several theories have been proposed, each of which stresses the importance of a particular array of situational factors (for reviews, see Bass and Stogdill 2001, Yukl 1998). There is no unified theory from which the most critical situational factors for leadership can be derived.

Fiedler's *Contingency Theory* concentrates on three situational factors (task structure, quality of leader-member relations, power inherent in the leader's position) that affect the leader's ability to influence followers depending on whether the leader prefers a task-oriented or a people-oriented leadership style.

Although the conceptions of Fiedler's theory and the empirical evidence are controversial, overall his work suggests that people-oriented leadership is more influential in moderately favorable situations than task-oriented leadership which works best in very favorable (high task structure, high quality of leader–member relations, high power) or very unfavorable situations.

Path-goal Theory by Robert House concentrates on an array of situational factors (task ambiguity, characteristics of followers, and the work environment) and suggests a situation-sensitive use of either directive, supportive, participative, or achievement oriented leadership styles in order to clarify for subordinates the path between performance and reward. The assumption is that the clearer the path is laid out, the higher is the followers' task motivation and performance. Despite inconclusive research results and conceptual deficiencies, path-goal theory provides a valuable conceptual framework for identifying situational factors relevant to leadership effectiveness.

Vroom and Yetton's *Contingency Theory* focuses on characteristics of short-term decision-making tasks. Their normative model prescribes the use of different leadership behaviors, ranging from autocratic, through consultative, to group decision-making, depending on the extent to which there is an objectively 'good' task solution available and the extent to which the followers' acceptance is required for proper enactment of the decision made. This approach received considerable empirical support and has been successfully used for leadership training purposes.

Implicit Leadership Theory developed by Lord and Maher (1991) is concerned with the cognitive processes of the led when confronted with leaders and leadership attempts. The theory posits that when a certain set of conditions and events is observed (or not), then individuals attribute that leadership has occurred (or not). Thus, leadership is regarded as a subjectively perceived rather than an objective construct. Individuals not only hold conceptions of prototypical leaders and evaluate actual leaders accordingly, but they also use their implicit leadership theories to judge the degree of leadership effectiveness. In so doing, they derive judgments about behavior that they actually have not observed. Implicit leadership theory poses a problem for the assessment of leadership through questionnaires such as the LBDQ because it underlines the long-ignored fact that we don't know whether such questionnaires measure the actual behavior of the leader or the leadership conceptions of the raters. Implicit leadership theory is to be viewed as a contingency approach because it suggests that the better the match between perceived leader attributes and behaviors, and the leadership concept held by the perceivers (the led), the more likely it is that the perceivers actually 'see' an individual as a leader (or an excellent leader) and allow the leader to exert leadership influence on them.

One basic message of these and many other situation-oriented approaches is that leaders must be able to flexibly adapt and cope with different and changing circumstances, otherwise they lose their influential status. However, adaptation and coping do not mean accepting situational factors as unchangeable conditions. Thus, we finally turn to power and influence approaches that address the issue of transforming attributes of the led, and the issue of transcending values and concepts so as to create an appealing vision (or a narrow-minded ideology) via leadership influence.

3.3 Power and Influence Approaches

Power approaches seek to explain leadership effectiveness in terms of the amount and type of power possessed and exercised by leaders. A classic distinction of forms of power was presented by French and Raven (1960): reward and coercive power, that is, the capability to offer incentives and to make use of organizational sanctions; legitimate power, when followers believe that the organizational or positional power over them is rightful; expert power, when high experience, knowledge, or ability are attributed; and referent power, when a person is referred to, or group norms are identified with, due to appealing personal qualities or values systems. Yukl (2001) describes three qualitatively different outcomes for the followers that result from employing these sources of power: commitment is most likely to be associated with referent and expert power, compliance with legitimate and reward power, and resistance with coercive power. The link between the use of power and behavioral approaches of leadership influence has been established by research on so-called influence tactics, for example, rational persuasion, consultation, ingratiation, exchange, coalition building, or pressure (cf. Yukl 1998).

Leader-member Exchange Theory (LMX) by Graen and his associates focuses on the development and the quality of the mutual relationship between leader and follower. Leaders are assumed to differentiate their followers according to their competence, trustworthiness, and motivation to assume ever more responsibility. Followers whom the leader perceives to display these attributes are categorized as so-called in-group members and in exchange are given more attention, support, and sensitivity by their leaders. The other, so-called out-group members, attract the more routine tasks and maintain a formal relationship with their leaders who in return exert influence with formal authority. Three 'currencies of exchange' between follower and leader, namely personal contribution, loyalty, and affect, have been identified. A meta-analytic review of LMX has been presented by Gerstner and Day (1997).

Transformational Leadership (Bass 1998) describes leaders to transform followers' attitudes and values, to activate their higher order motives and to stimulate them to transcend the organizations' higher order goals by their self-interests. The four components of transformational leadership (idealized influence, inspirational motivation, intellectual stimulation, and individualized consideration) are measured in the Multifactor Leadership Questionnaire (MLQ) and were found to be empirically distinguishable and to relate to work unit effectiveness. Transformational leadership is often contrasted with transactional leadership which is based primarily on compliance and norms recognized through leader–follower exchanges, that is, reward and praise are given by the leader for task completion and loyalty given by the follower.

Charismatic leadership is defined more narrowly than transformational leadership. It specifies an idolized leader's characteristics as perceived and attributed by followers. Behaviors typical of charismatic leaders are for example, the articulation of appealing visions, communication of high expectations, and expression of high confidence in followers. Some evidence of the principally positive relationship between personal charisma and effectiveness, especially in situations of crisis, was presented by House et al. (1991) in an investigation of US American presidents and various measures of national effectiveness during each president's term.

There is also a 'dark side' to charismatic leadership. Problems that can occur with 'negative' charismatics are, for example, that they seek to induce commitment to narrow-minded ideological goals or to themselves, start projects with often unrealistic premises, omit properly investing in the implementation of their visions, and fail to develop competent successors.

4. *Issues of Concern for the Future*

The current body of leadership research includes over 5,000 studies and we know much more about leadership than is usually recognized (cf. Yukl 2001). However, there is a lack of research integrating the different aspects outlined above (leadership traits, skills, behaviors, situational factors, power, and influence processes). Today, a general theory of leadership is not available, but some authors have described the basic conceptual components that should be part of it (e.g., Yukl 2001).

On the way to developing general theories of leadership, two issues seem to be of particular importance. First, the psychological theories of personal leadership which dominate the field should be complemented by social psychological and sociological theories of 'shared' leadership which integrate situational factors within the social context and consider the social interactive nature of leadership processes. Second, the prevailing 'western' style leadership theories should be complemented by culture-specific and cross-cultural theories which are more appropriate for different societal cultures and for informing expatriate managers and a multicultural work force.

4.1 *Personal vs. Shared Leadership*

Most theories of leadership favor conceptions of unilateral influence by a single leader who is perceived as a person with outstanding characteristics. In order to understand how leadership is embedded in the dynamics of the social systems that constitute an organization, the unrealistic expectation of an 'heroic leader' who is more informed and confident than anyone else in an organization needs to be complemented by the view that leadership in organizations is a reciprocal influence process (cf. Smith 1995) involving many people from the various social systems in an organization. Bradford and Cohen (1984), speaking of 'shared' leadership, contend that the predominant conception of an 'heroic leader' undermines the principally positive effects of shared responsibility for leadership functions and empowerment of followers on leadership effectiveness. The questions of how power and leadership can be shared in organizations, and of identifying the conditions that increase the effectiveness of shared leadership deserve more research. However, this kind of research needs to employ social psychological and sociological conceptions of social interaction processes, leadership, and power as well as more elaborate research methodologies for investigating the complex nature of leadership within social systems that have been developed for studying the unilateral relationship between a single leader on followers.

4.2 *Cross-cultural Leadership Research*

The quantity of leadership research performed in cultures other than 'western' societies (e.g., USA, Canada, Western Europe) is very limited. This leaves it open as to whether the currently available leadership theories are universally applicable. We have good reasons to assume that there are some similarities but also marked differences in conceptions about leadership across cultures (House et al. 1997), even within particular regions that share a common market, like Europe (Brodbeck et al. 2000). For example, the GLOBE research program found that differences in the implicit leadership theories (Lord and Maher 1991) held by middle level managers are linked to the degree of societal cultural differences of the nations studied (Brodbeck et al. 2000, House et al. 1999). To move beyond a formal role in influencing others, one must first be perceived as a leader (or an effective or a trustworthy leader etc.). It is unlikely that someone

not perceived as a leader can exercise the influence on others necessary for performing effectively. With regard to culturally endorsed implicit leadership theories, it is expected that the less they overlap in terms of cross-cultural leader–follower relationships, the less likely it is that the leader will be accepted and the respective interpersonal relationships will be characterized by trust, motivation, and high performance.

The removal of trade barriers and the growth of global markets increases the permeability of organizational and national cultures. Managing a culturally diverse workforce and the transitions it entails requires that insights derived from research inform participants in a multicultural workforce, as well as the trainers and consultants facilitating them in accommodating values, skills, and behavior and adjusting the managerial context in suitable ways. Thus cross-cultural leadership research is clearly an important issue for future research.

See also: Educational Leadership; Hierarchies and Markets; Leadership in Organizations, Sociology of; Leadership, Psychology of; Organizational Behavior, Psychology of; Organizations, Sociology of; Psychological Climate in the Work Setting; Trust: Philosophical Aspects; Well-being and Burnout in the Workplace, Psychology of

Bibliography

Bass B M, Stogdill R M 1990 *Handbook of Leadership: Theory, Research and Managerial Applications*. Free Press, New York

Bass B M 1998 *Transformational Leadership*. Erlbaum, Mahwah, NJ

Bradford D L, Cohen A R 1984 *Managing for Excellence: The Guide to Developing High Performance in Contemporary Organizations*. John Wiley, New York

Brodbeck F C, Frese M, Ackerblom S, Audia G, Bakacsi G, Bendova H et al. 2000 Cultural variation of leadership prototypes across 22 European countries. *Journal of Occupational and Organizational Psychology* **73**: 1–29

French J, Raven B H 1960 The bases of social power. In: Cartwright D, Zander A F (eds.) *Group Dynamics*. Row and Peterson, Evanston, IL

Gerstner C R, Day D V 1997 Meta-analytic review of leader–member exchange theory: Correlates and construct issues. *Journal of Applied Psychology* **82**(6): 827–44

House R J, Spangler W D, Woyke J 1991 Personality and charisma in the US presidency: A psychological theory of leadership effectiveness. *Administrative Science Quarterly* **36**: 364–96

House R J, Wright N S, Aditya R N 1997 Cross-cultural research on organizational leadership: A critical analysis and a proposed theory. In: Earley P C, Erez M (eds.) *New Perspectives on International Industrial/Organizational Psychology*. New Lexington Press, San Francisco

House R J, Hanges P, Ruiz-Quintanilla S A, Dorfman P W, Javidan M, et al. 1999 Cultural influences on leadership and organizations: Project GLOBE. In: Mobley W, Gessner M J, Arnold V (eds.) *Advances in Global Leadership*. JAI Press, Stanford, CA, Vol. 1

Locke E A, Latham G P 1990 *A Theory of Goal Setting and Task Performance*. Prentice-Hall, Englewood Cliffs, NJ

Lord R G, Maher K J 1991 *Leadership and Information Processing: Linking Perceptions and Performance*. Unwin-Hyman, Boston

McClelland D C 1965 Achievement and entrepreneurship: A longitudinal study. *Journal of Personality and Social Psychology* **1**: 389–92

Smith P M 1995 Leadership. In: Manstead A S R, Hewstone M (eds.) *The Blackwell Encyclopedia of Social Psychology*. Blackwell, Oxford, UK

Stogdill R M 1948 Personal factors associated with leadership: A survey of the literature. *Journal of Psychology* **25**: 35–71

Yukl G 2001 *Leadership in Organizations*, 5th edn. Prentice-Hall, Upper Saddle River, NJ

F. Brodbeck

Leadership in Organizations, Sociology of

1. A Definitional Confusion

The Anglo-Saxon etymological origin of the words lead, leader, and leadership is *laed*, which stands for 'path' or 'road.' The verb *laeden* means 'to travel.' Thus a leader is one who shows fellow travelers the way by walking ahead. This metaphor of the leader as helmsman is still very much on the mark. Unfortunately, the clarity of leadership's etymology is rarely matched with clarity of meaning. Papers, books, and articles claiming to delineate leadership proliferate, yet their conclusions can be confusing and even conflicting. Among the more popular are descriptions in terms of traits, behavior, relationships, and follower perceptions.

The proliferation of literature on leadership in recent years is amply reflected by the increase in the number of articles listed in the latest edition of the *Handbook of Leadership* (Bass 1990). Reading through this gargantuan tome, however, is a sobering and often bewildering experience. The content of the book demonstrates that the popularity of leadership research is not always equaled by its relevance. Rather than concentrating on what key decision makers at the strategic apex of their organization are doing in the context of their work environment, researchers all too frequently draw their major conclusions from laboratory experiments, observations of leaderless groups, or the activities of lower-level supervisors. If leadership is to be a viable area of study—and that study is to be of service to a constituency of executives—its research focus needs to be closely tied to observations of the behavior and actions of individuals in leadership positions.

2. A Proliferation of Theories

Broadly speaking, two extreme positions can be identified in leadership research. On one side of the spectrum are the 'personalists'—researchers who argue that specific personality variables determine leadership effectiveness. On the other side of the spectrum are the 'situationists'—those who deny the influence of individual differences and attribute all variations in leadership effectiveness to environmental constraints. While personalists views leaders as heroic helmsmen, in control of whatever situation they find themselves, situationists turn leaders into figureheads—puppets manipulated by the forces of the environment.

These opposing positions set the stage for a cornucopia of theories, each backed by strong defenders. We can find 'great man' theories, trait theories, situational theories, psychoanalytic theories, political theories, humanistic theories, cognitive theories, leader-role theories, reinforced change theories, path-goal theories, contingency theories, multiple linkage theories, vertical dyad linkage theories, exchange theories, behavioral theories, and attribution theories. This lack of apparent convergence has caused some scholars to abandon the subject altogether.

One of the problems in dealing with the subject of leadership is that it can be looked at as both a property and a process. As a property, leadership is seen as a set of characteristics—role behaviors and personality attributes—that make certain people more effective in attaining a set of goals. As a process, it is seen as an effort by a leader, drawing on various bases of power, to influence members of a group to direct their activities toward a common goal. Taking the property–process debate as a point of departure, let us turn first to the property perspective.

2.1 Order Out of Chaos

The trait theory (or great man theory) offered one of the earliest conceptual ways of looking at leadership. This theory holds that there is one best way to lead and that deeply seated personality variables allow certain people to master that best way. According to this theory, there are a number of universal characteristics of personality that determine a leader's effectiveness, without regard to behavior in a given situation. Because leadership is viewed as a set of relatively stable and enduring personal traits or physical properties, specific personality characteristics distinguish effective from ineffective leaders.

The initial search for these universal traits applicable to any setting was not overly successful, however. The results was conflicting, with methodological problems in research design cited as the major reason. Disappointed by the results of these studies, many scholars interested in leadership abandoned this line of research altogether, turning to other approaches to leadership. Recently, however—after a long hiatus in trait research (and with the help of better measuring techniques)—a revival of trait theory has been observed. Those studies that have gone beyond the simplistic, atomistic approach of previous trait studies have identified a number of personality characteristics that consistently emerge, differentiating leaders from nonleaders—dimensions of character that can be mapped into the Big Five model of personality structure (Hogan et al. 1994). These various dimensions can be described in terms of surgency (a broad term that embraces competitiveness, achievement orientation, self-assuredness, and dominance), agreeableness, conscientiousness, emotional stability, and intelligence (including emotional intelligence). In addition, these studies regularly list factors such as physical energy and extraversion.

Another group of leadership scholars espouses a behavioral theory. Distinguishing between the technical actions of a leader and the human actions, these behaviorists emphasize a set of observable role behaviors rather than traits. Like trait theorists with their individual characteristics, these scholars see certain role behaviors as being universal—that is, as producing leadership effectiveness regardless of the setting. Generally, however, the constructs employed by these scholars have been too rudimentary. This approach often looks at behaviors via dimensions such as consideration versus initiation, or task orientation versus relationship orientation (Bales 1958, Fleishman et al. 1955). Another popular typology with a behavioral slant on leadership contrasts autocratic and democratic approaches (Tannenbaum and Schmidt 1958).

From this approach to leadership behavior, we move on the continuum to contingency theory. Instead of taking the position that leadership traits or behaviors are applicable to any situation, those supporting the contingency theory claim that the emergence of any one style is contingent on the environment in which the leader is operating. According to this point of view, the most effective leader is the one who is able to adapt his or her actions depending on the situation. In the model of one of the most prominent advocates of this point of view, the effectiveness of task- or relationship-oriented leaders depends on the favorableness of the situation as defined by the power of their position, the task structure, and the quality of the leader-member relationship (Fiedler 1967).

2.2 The Importance of Attribution to Leadership

Another group of scholars espouses the attribution theory of leadership. According to these situationists, leadership is not a viable scientific construct; it is a mere label given to behavior. Only people's inferences about and reactions to leaders are viable (Calder 1977). Because individuals have an inherent need to

explain events that surround them, they assume that certain types of behaviors and actions can be attributed to the leader. Thus leadership is a perceptional issue, an illusion: individuals infer causation from observed behavior. The knowledge of the outcome causes individuals to attribute certain qualities to a leader.

This more situational point of view has been reinforced by a number of scholars of leadership who doubt whether leaders affect organizational performance. Advocates of this line of thinking contend that there are powerful external forces that shape organizational activities. Each leader is embedded in a social system—a system in which other actors not only have expectations regarding appropriate behavior but also make efforts to modify the leader's behavior—that places serious constraints on leader behavior. Leadership turns into a set of myths encouraging social construction of meaning, creating an illusionary causality (Pfeffer and Salancik 1978).

The symbolic role of leadership has been further explored by other leadership scholars (Meindl et al. 1985). In an extension of the attributional school of thought, these researchers call attention to the 'romanticized' conception of leadership. Ironically (given the attributional school's situational roots), these scholars suggest that leaders can play an important role through the manipulation of symbols in the management of meaning—activities that can be highly effective in influencing others. Using the vehicle of symbolism, many advocates of the situational point of view have been inching toward an interactionist approach to leadership studies, positioning the leader/led relationship clearly at central stage.

2.3 Leadership as a Charismatic Process

Like the attributional school of thought, the contingency approach has also set the stage for a more relational slant on the study of leadership. Believing a leader cannot be studied meaningfully in isolation from his or her surroundings, this approach views leadership as an interactive process between the leader, the followers, and the situation. This orientation, transcending earlier, more naive approaches to leadership behavior, is an important step forward.

The problem with many relational theories in the past is that their point of convergence was too narrow. Initiation vs. consideration, social orientation vs. task orientation, autocracy vs. democracy—such dimensions are overly simplistic in describing leadership in its context. Furthermore, these earlier relational studies focused far too much on exclusive superior–subordinate relationships, ignoring (or slighting) the various constituencies of the leader: the industry environment, the national culture, and the culture that characterizes the organization.

This shortcoming has opened the door for a fresh look at charismatic leadership. Another impetus for this line of research has been the prevalence of a business climate of uncertainty and unpredictability—a breeding ground for the emergence of charismatic leadership. In our competitive, global world, where the transformation and revitalization of organizations holds a central position, the leader is increasingly seen as a crucial agent of change.

The new focus, then, is on the inspirational role of leaders. Researchers are turning to the study of leaders who by force of their personality have an extraordinary effect on their followers. The challenge for leaders of organizations becomes how to affect the mind-set of the organizational participants through value creation, through influencing the organization's culture, and through building commitment to the organization's mission, objectives, and strategies to obtain well-above-average organizational performance.

The first person to take up this new challenge was political scientist MacGregor Burns (Burns 1978). In his writing, he extends Max Weber's reflections on charisma (Weber 1947), making a distinction between transactional and transformational leadership. While transactional leadership can best be viewed as a mundane contractual exchange based on self-interest (often described in the literature as the manager's role), transformational leadership seeks to satisfy the higher needs of followers—to engage in a process of mutual stimulation and elevation whereby followers will transcend their own self-interests for the good of the group (Bennis and Nanus 1985, Conger and Kanungo 1998).

A number of researchers have built on Burns' notion of transformational leadership, using observed behavior of leaders to break the concept down into various components, in an effort to broaden early charismatic conceptualizations. For example, Bass and Avolio (1993), who view charisma as a subset of transformational leadership, list four behavioral components in the context of transformational leadership: (a) charisma or idealized influence, (b) inspiration, (c) intellectual stimulation, and (d) individualized consideration. According to them, charisma alone is insufficient to put in place a successful transformation process. Shamir et al. (1993), building on earlier work done by House (House 1977), contend that charismatic leadership affects followers' self-concepts and has motivational consequences due to (a) changing follower perceptions of the task that has to be accomplished, (b) offering an attractive vision of the future, (c) creating a group identity, and (d) heightening individual and collective feelings of self-efficacy.

These various offshoots of a focus on the inspirational role of leadership contribute to a rich description of what the leadership mystique is all about. Researchers who view leadership as a charismatic or transformational process give proper attention to the

contextual and cultural dimensions that are part and parcel of leadership dynamics. They are sensitive to the impact of the environment on leaders and on their behavior. Furthermore, they reject narrow instrumentalism in favor of a perspective whereby the leader is seen as the transformational agent of change. Some scholars have made the point, however, that the transactional role of leadership should not be ignored. They suggest that the most effective leaders take on two roles: a charismatic role (consisting of envisioning, empowering, and energizing) and an architectural role (designing the organization, setting up structures, and formulating control and reward systems (Kets de Vries and Florent-Treacy 1999)).

Some scholars of leadership argue, however, that in spite of the new, richer color given to leadership research, additional steps need to be taken to deepen our understanding of the leader's relational interchanges. And the challenge is formidable. In spite of the various rational ways in which researchers attempt to deconstruct leadership and charisma, charismatic leadership is not rational in the traditional sense of the word. By its very nature, it is unstable, in that it exploits what can be interpreted as irrational processes. We need now to find ways to explore the forces that transcend rationality. Critics also argue that the study design of many researchers evaluating inspirational leaderships treats all leaders and all followers as amorphous, interchangeable groups of people; in other words, they fail to attend to differences in personality style. To rectify these shortcomings in leadership research, deeper insight into people's desires, wishes, and needs is needed; and that insight can be provided by a clinical focus.

2.4 The Clinical Paradigm

The clinical orientation to leadership research uses findings from psychoanalysis, cognitive theory, developmental psychology, and family systems theory to arrive at a richer understanding of personality and leadership. In the deconstruction of the dynamics of leadership, this orientation looks to the triangle of mental life consisting of emotion, cognition, and behavior. While in other approaches to leadership the focus is generally on cognition and behavior, in the clinical approach emotions enter the equation. Research on how people alter has revealed that cognition alone does not create change; cognition needs to be complemented by emotion. The clinical paradigm also factors in unconscious processes.

The clinical orientation toward the study of leadership has helped achieve greater understanding of the leader–follower interchange. Research into the dyadic relationships created by leaders (and acquiesced to by followers) suggests that failure in leadership can also be considered failure in 'followership': just as influence moves down from the leader to followers, so also does it move up. Understanding the impact of transferential processes such as 'mirroring' and 'idealizing'—processes characterized by confusion of time and place between leader and led—helps researchers clarify otherwise inexplicable phenomena.

The clinical paradigm, with its belief that every thought and action has a reason, sheds new light on irrational behavior in organizations. Processes such as projection, projective identification, splitting, collective regression, identification with the aggressor, *folie à deux*, the fear of success, scapegoating, narcissism, vindictiveness, and containment—along with other elements of the clinical paradigm—can help researchers better understand the leader–follower exchange. The clinical paradigm can provide insights into the dynamics of group behavior and the role of the leader; illuminate male–female differences in leadership through the analysis of fantasies around gender; and provide an understanding of the positive and negative effects leaders have on the corporate culture, structure, and decision-making processes (Kets de Vries and Miller 1984). By looking at a leader's inner theater, scholars can better appreciate the reasons why that leader derails and his or her company fails.

3. Future Concerns

Although, at first step, venturing into the domain of leadership research may seem like walking on quicksand, this brief overview demonstrates the considerable advances that have been made over the last decades. Due to promising new research directions, especially those working toward a gradual convergence of situationist and personalist positions, the prevailing attitude of disillusionment with leadership studies is experiencing a turnaround. Most researchers of leadership now perceive the importance of a relational, interactionist point of view that looks at actual leaders in their 'natural' setting. As a result, there is now considerable agreement that less laboratory and more field studies of leadership are needed. Moreover, it is no longer difficult—given the rapid changes of our era—to convince researchers of the relevance of the transformational side of leadership.

Although leadership research has come a long way, more work needs to be done. One key factor in the selection of topics that require further investigation should be relevance of a topic to leadership's various constituencies. But 'relevance' needs to be broadly construed: the subject of leadership should be seen as applicable not only to a few highly exceptional individuals at the top of the organization but to a much broader audience.

Meeting these criteria, a number of issues feature prominently on the research agenda of the future. Many of them address challenges growing out of the exponential rate of change in this age of transformation:

(a) The role of leaders as catalysts of change needs further exploration. All too many organizational transformation efforts fail, at great cost to people and society.

(b) As global mergers, acquisitions, and strategic alliances become increasingly common, more attention needs to be given to the cross-cultural dimensions of leadership. To what extent does effective leadership behavior vary from culture to culture?

(c) Given the cross-cultural focus of business today, we need to attempt identification of the required competencies for leaders who run global organizations.

(d) Likewise, we need to investigate the leadership implications for running cross-cultural teams.

Many issues centered around corporate governance are begging for further investigation:

(a) Top-executive role constellations need to be further explored. What new insights can be provided about complementarity of roles? What kind of underlying dynamics can be identified?

(b) Better methods of selecting people with leadership potential need to be developed. What can be said about early identification of leadership potential?

(c) Given the costly dramas that frequently surround top-management succession, how can we better understand the emerging psychological processes that go with it?

Those establishing a future research agenda must remember that leaders do not work alone. They need followers, and (with rare exceptions) they themselves are also followers; they also have many colleagues. Future research needs to address that collaboration:

(a) Followership needs to be given a more prominent place in leadership research. How, for example, do followers manage their leaders?

(b) Research money also needs to be devoted to the question of how leadership can be distributed throughout the organization. What can be done to 'stretch,' to get the best out of the most people? Given the increased reliance of organizations on creativity and innovation, what can leaders do to stimulate this process?

A final research category includes issues related to leaders' self-understanding:

(a) To guide others effectively, leaders must—absolutely must—know themselves. As part of that process, they need to recognize and explore their blind spots. Leaders who fail to take their irrational side into account are like captains who blindly plow their ships into a field of icebergs: the greatest danger is hidden below the surface.

(b) How we can develop leaders' emotional intelligence? Recognizing the limits of rationality, leaders must become more sensitized to the irrationality in themselves and in others.

Leaders fulfill many different roles in people's imagination. They are catalysts of change; they are symbols; they are objects of identification; and they are scapegoats when things go wrong. Leaders are also prone to hubris. As Napoleon (an expert on the topic of hubris as well as leadership) once said, 'Glory is fleeting, but obscurity lasts forever.' All leaders are susceptible to the darker sides of power. The most effective leaders, however, are the ones who know how to balance action with reflection by using self-insight as a restraining force when the sirens of power are beckoning.

See also: Attributional Processes: Psychological; Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Charisma and Charismatic; Charisma: Social Aspects of; Elites: Sociological Aspects; Leadership: Political; Leadership, Psychology of; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture; Organizational Decision Making; Organizations and the Law; Weber, Max (1864–1920)

Bibliography

Bales R F 1958 Task roles and social roles in problem-solving groups. In: Maccoby E E, Newcomb T M, Hartley E L (eds.) *Readings in Social Psychology*, 3rd edn. Holt, New York

Bass B M 1990 *Bass & Stogdill's Handbook of Leadership: Theory, Research, and Applications*, 3rd edn. The Free Press, New York

Bass B M, Avolio B J 1993 *Transformational Leadership: A Response to Critiques in Leadership Theory and Research Perspectives and Directions*. Academic Press, New York

Bennis W, Nanus B 1985 *Leaders: The Strategies for Taking Charge*, 1st edn. Harper & Row, New York

Burns J M 1978 *Leadership*, 1st edn. Harper & Row, New York

Calder B J 1977 An attribution theory of leadership. In: Straw B M, Salancik G R (eds.) *New Directions in Organizational Behavior*. St. Clair Press, Chicago, pp. 179–204

Conger J A, Kanungo R N 1998 *Charismatic Leadership in Organizations*. Sage, Thousand Oaks, CA

Fiedler F E 1967 *A Theory of Leadership Effectiveness*. McGraw-Hill, New York

Fleishman E A, Harris E F, Burt H E 1955 *Leadership and Supervision in Industry*. Ohio State University Press, Columbus, OH

Hogan R, Curphy G J, Hogan J 1994 What we know about leadership-effectiveness and personality. *American Psychologist* **49**(6): 493–504

House R J 1977 A 1976 theory of charismatic leadership. In: Hunt J G, Larson L L (eds.) *Leadership: The Cutting Edge*. Southern Illinois University Press, Carbondale, IL, pp. 189–207

Kets de Vries M F R, Florent-Treacy E 1999 *The New Global Leaders: Richard Brandson, Percy Barnevik, and David Simon and the Making of the International Corporation*, 1st edn. Jossey-Bass, San Francisco

Kets de Vries M F R, Miller D 1984 *The Neurotic Organization*, 1st edn. Jossey-Bass, San Francisco

Meindl J R, Ehrlich S B, Dukerich J M 1985 The romance of leadership. *Administrative Science Quarterly* **30**: 78–102

Pfeffer J, Salancik G R 1978 *The External Control of Organizations: A Resource Dependence Perspective*. Harper & Row, New York

Shamir B, House R J, Arthur M B 1993 The motivational effects of charismatic leadership: a self-concept based theory. *Organization Science* **44**: 577–94
Tannenbaum R, Schmidt W H 1958 How to choose a leadership pattern. *Harvard Business Review* **36**: 95–101
Weber M 1947 *The Theory of Social and Economic Organization*, 1st American edn. Oxford University Press, New York
Zaleznik A 1977 Managers and leaders: Are they different? *Harvard Business Review* **55**: 67–78

M. F. R. Kets de Vries

Leadership: Political

Students of political leadership are faced with a paradox. On the one hand, the subject has been studied widely. There are many books and articles which deal with the concept from a theoretical perspective (e.g., Blondel 1987, Burns 1978, Edinger 1975, 1990, Paige 1977, Tucker 1981) and there is even more work which examines the actions of political leaders from an empirical point of view (e.g., Blondel and Müller-Rommel 1988, Elgie 1995, Jones 1991, and Rose and Suleiman 1980). On the other hand, though, the concept remains poorly understood. There is still no agreed definition of the term 'leadership.' Indeed, there are thousands of competing definitions (Rost 1991, pp. 37–98). In addition, over the years the study of political leadership has been associated with an almost bewildering variety of approaches based on radically different assumptions about the fundamental nature of the political process. The result is a situation where the main strength of leadership studies is also its principal weakness. That is to say, there is at once a tremendous richness to the study of political leadership *grosso modo*, but at the same time it remains an ill-defined and sometimes even slightly esoteric subfield of the political science discipline as a whole.

1. The Two Meanings of the Terms 'Leader' and 'Leadership'

The terms 'leader' and 'leadership' are understood in two particular ways. First, the terms are used in an institutional sense to mean the equivalent of office-holding. Here, they are merely shorthand ways of referring to people who hold particular positions of authority. So, for example, reference might be made to the president as the leader of the United States (US) or to a person having been elected to the leadership of the British Conservative Party. Second, the terms are also used in a behavioral sense to mean a certain type of relationship which occurs between the various members of a particular group. In this case, the 'leader' is the person who is able to convince the other members of the group to follow a particular course of action and 'leadership' refers to the process of interaction between the leader and the followers. In this sense, leadership is a reciprocal and essentially noncoercive relationship.

The fact that the terms 'leader' and 'leadership' can be used in these two different ways means, first, that leadership does not have to be exercised by leaders and, secondly, that there can be leaders who do not lead. The first of these initially perplexing statements refers to the situation where leadership in the behavioral sense is exercised by someone who does not occupy a formal position of authority. For example, social movements are often formed by individuals who are able to articulate demands which are latent within a community. By creating a sense of mutual purpose towards a particular goal, political leadership is present. However, it is often exercised in this case by unelected people who do not hold any formal office. Such people are, as Tucker (1981) puts it, non-constituted leaders, Martin Luther King being a prime example. The second statement refers to the situation where people who occupy formal positions of authority are unable to convince others that a particular course of action should be followed. These people are still office-holders and, thus, are leaders in the institutional sense of the term, but they are unable to influence others and exercise leadership in the behavioral sense of the word. One writer has suggested that 'headship' rather than 'leadership' might best describe the relationship between the leader and the other members of the group in this type of situation (Kellerman 1984).

2. The Study of Political Leadership

The study of political leadership reflects the two basic understandings of the term.

The first type of leadership studies examines the roles and powers of office-holders. The study of political leadership in this sense is intrinsically bound up with the fields of constitutional law, public administration, and government generally. Thus, the focus of inquiry is very broad. Amongst other things it includes country-specific or comparative studies of government figures at the central and subcentral level, the powers of state officials and party leaders, the profiles of elected representatives and political appointees, as well as the conduct of electoral campaigns and political contests. As befits the generality of this type of leadership studies, the subject may also be approached from a variety of methodological standpoints, behavioral, rational choice, and/or institutional. The result is a rich and very varied literature indeed. Classic examples include Neustadt (1965) and Linz and Valenzuela (1994).

The second theme in the work of political leadership concerns the study of how and why leadership in the behavioral sense comes about, if at all. This aspect of

the literature is somewhat more specialized. It is also more informed by work in other disciplines, especially social psychology and business and management studies.

The earliest studies of this type suggested that leaders were born not made. They were said to possess innate qualities which set them apart from others and which meant that they had a natural gift for bringing about followership and, thus, for creating history (Carlyle 1907—written in 1840). This approach, however, subsequently was opposed. Leaders, it was argued, were imprisoned by social forces with events being determined by the impersonal interplay of social and cultural forces over which individuals had little or no control (Spencer 1884). These two mutually exclusive approaches have long since been superseded. There is now general agreement that the study of political leadership should be approached from an interactionist perspective (Greenstein 1992). That is to say, the extent to which individuals are able to influence the political process is considered to be contingent upon the interaction between individuals and the environment in which they find themselves. Whether or not people who hold official positions of authority are able to affect the behavior of others depends on the qualities of the people concerned and the circumstances with which they are confronted at any given time.

Contemporary examples of this type of leadership studies have concentrated overwhelmingly on the first element of the interactionist approach, namely the motivations of individual political leaders. One aspect of this work concerns the long-standing issue of whether or not certain personality traits (such as intelligence or appearance) are positively correlated with successful leadership. The findings suggest that the importance of traits is situationally specific (Van Fleet and Yukl 1989). Another aspect relates to the concept of leadership styles, meaning the rigorous classification of the different ways in which leaders tackle the tasks before them. The most well-known example of this sort is Barber's study of the US presidential character (Barber 1972). One of the main reasons for the interest in Barber's thesis was because it appeared to have predicted that Richard Nixon would not be a success as president. So far, though, it has yet to be applied systematically outside the context of US politics. A further aspect relates to the work on political psychology. The literature in this subfield is very broad, but of particular interest to leadership studies is the work on psychobiography, or the explanation of political events in terms of an explicit theory of personality. A fine example of this approach is the study of Woodrow Wilson by George and George (1956).

All told, the study of political leadership encompasses a wide variety of work which approaches the subject from many different perspectives. This is, of course, a strength. That said, one of the weaknesses of the literature as whole is the absence of a unifying theory, never mind contending theories. The nature of the subject matter is such that it scarcely lends itself to such a grand project.

3. *The Future of Leadership Research*

In one sense the study of political leadership has gone somewhat out of vogue. The behavioral studies of the 1960s and 1970s proved to be less than compelling in their attempts to identify the essence of the concept. Indeed, it is perhaps telling that one of the most influential books of this era (Burns 1978) appears to have had a greater impact among students of management than within the political science community itself. In another sense, though, the study of political leadership is central to some of the most basic puzzles of political science: the relationship between structure and agency, between individual and collective action, between free will and determinism. Thus, there is still great scope for the advancement of this subfield. In this context, there are some specific research agendas to be promoted. Rejai and Philips (1997) note seven such areas: agreement about the definition of leadership; classifications of types of leaders; studies of women leaders; refining the distinction between leaders and elites; studies of minority and third-world leaders; the nature of crisis leadership; and the determinants of leadership duration. More generally, though, there is perhaps an even more urgent need to integrate the leadership factor into both rational choice analysis and the 'new' institutionalism. In terms of the former, to what extent does leadership affect preference formation and the articulation and definition of political choices? In terms of the latter, to what degree does leadership affect the creation of institutions and the process of institutional change? In these ways, the study of political leadership has the potential to continue to be at the heart of the political science project for the foreseeable future.

See also: Authoritarianism; Authority: Delegation; Authority, Social Theories of; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Leadership, Psychology of; Participation: Political; Political Elites: Recruitment and Careers; Political Parties; Weberian Social Thought, History Of

Bibliography

Barber J D 1972 *The Presidential Character. Predicting Performance in the White House*. Prentice-Hall, Englewood Cliffs, NJ

Blondel J 1987 *Political Leadership. Towards a General Analysis*. Sage, London

Blondel J, Müller-Rommel F (eds.) 1988 *Cabinets in Western Europe*. Macmillan, London
Burns J M 1978 *Leadership*. Harper and Row, New York
Carlyle T 1907 *On Heroes, Hero-Worship, and the Heroic in History*. Educational Publishing, Boston
Edinger L J 1975 The comparative analysis of political leadership. *Comparative Politics* **7**(2): 253–69
Edinger L J 1990 Approaches to the comparative analysis of political leadership. *Review of Politics* **52**(4): 509–23
Elgie R 1995 *Political Leadership in Liberal Democracies*. Macmillan, London
George A L, George J L 1956 *Woodrow Wilson and Colonel House: A Personality Study*. Day, New York
Greenstein F I 1992 Can personality and politics be studied systematically? *Political Psychology* **13**(1): 105–28
Jones G W (ed.) 1991 *West European Prime Ministers*. Cass, London
Kellerman B 1984 Leadership as a political act. In: Kellerman B (ed.) *Leadership: Multidisciplinary Perspectives*. Prentice-Hall, Englewood Cliffs, NJ, pp. 63–89
Linz J J, Valenzuela A (eds.) 1994 *The Failure of Presidential Democracy*. Johns Hopkins University Press, Baltimore, MD
Neustadt R 1965 *Presidential Power*. Wiley, New York
Paige G D 1977 *The Scientific Study of Political Leadership*. Free Press, New York
Rejai M, Philips K 1997 *Leaders and Leadership. An Appraisal of Theory and Research*. Praeger, Westport, CT
Rose R, Suleiman E N (eds.) 1980 *Presidents and Prime Ministers*. American Enterprise Institute, Washington, DC
Rost J C 1991 *Leadership for the Twenty-first Century*. Praeger, New York
Spencer H 1884 *The Study of Sociology*. Appleton, New York
Tucker R C 1981 *Politics as Leadership*. University of Missouri Press, Columbia, MI
Van Fleet D D, Yukl G A 1989 A century of leadership research. In: Rosenbach W E, Taylor R C (eds.) *Contemporary Issues in Leadership*. Westview Press, London, pp. 65–90

R. Elgie

Leadership, Psychology of

Leadership is a process of social influence in which one person is able to enlist the aid and support of others in the accomplishment of a task or objective. *Effective* leadership requires that the person in the leadership role establish trust and credibility to enlist the support of followers; build relationships with those followers that motivate them to contribute their energy and resources to the collective effort; and manage, direct, and apply those collective resources to accomplish the group's mission or task. The study of leadership has focused on the characteristics of leaders, followers, and situational contexts. It has sometimes taken an individualistic focus and sometimes a group orientation. It has concerned itself both with the 'realities' of leader and follower behavior and situational parameters, as well as with the perceptual constructions and biases surrounding the leadership relationship. The paragraphs which follow will present the major empirical and theoretical trends in the leadership literature and will end with an integrative summary.

1. Early Attempts: Blind Alleys and Dead Ends

1.1 Traits

The first 50 years of scientific leadership research (roughly 1910–60) were characterized by failed attempts to find simple explanations for a complex phenomenon. The earliest research on leadership was influenced by the successes in the development of the first intelligence tests and pursued a search for a leadership trait, i.e., a stable, personal characteristic associated with leadership status or effectiveness. A wide range of obvious traits was chosen for examination, including intelligence, social sensitivity, dominance, physical height or bearing, and so on. Stogdill (1948) reviewed over 125 leadership trait studies and concluded that no single trait or collection of traits was consistently related to leadership status, nor, he concluded, should one expect them to be. He based his conclusion on two points of analysis. First, leadership was such a complex process and occurred in so many different contexts with different tasks and goals (e.g., combat leadership vs. academic leadership) that no one trait should predict leadership in all situations. Second, he believed that a truly useful theory of leadership would require the integration of leadership traits and situations and specify the nature of their interaction.

1.2 Behaviors

After traits, research focused on the description and measurement of leader behavior. A broad range of methods included interviews with workers describing their supervisors; observations and categorization of the behavior of student groups in the laboratory; and questionnaire ratings of military and industrial leaders. The most extensive and influential of these projects was the development of the Leader Behavior Description Questionnaire (LBDQ) (Hemphill 1950). The LBDQ included 150 descriptions of leadership behavior grouped into a questionnaire that allowed for the rating of leader behavior by subordinates, peers, superiors, or trained observers. Subsequent analyses of these ratings indicated that leader behavior fell into two general categories. The first, labeled *Consideration*, included leader attempts to create a positive group climate through attentiveness and consideration of the feelings and concerns of subordinates. The second factor, labeled *Initiation of Structure*, reflected leadership efforts to accomplish the group's assigned task by organizing, directing, and structuring the work of subordinates. These two

general behavioral orientations—a concern with task and a concern with followers—was common to many measures and observations of leader behavior, but did not, unfortunately, seem to have much predictive power to distinguish effective from non-effective leadership.

1.3 Situations

Although not at the mainstream of early leadership research, a few investigations of the effects of situational variables on leadership were done. While no common theory of situations emerged, it was found that situational factors such as the nature of the group's task, the structure of communication networks, and the number of people in the group could affect leadership patterns. The failure of the trait, behavior, and situational approaches highlighted the cogency of Stogdill's (1948) suggestion that a good theory must integrate these elements and set the stage for the emergence of the 'contingency theories.'

2. The Contingency Theories

The notion underlying contingency theories is that the relationship of a leader's characteristics (e.g., personality, behavioral style) to group performance is moderated by (i.e., *contingent* upon) aspects of the leadership situation or environment.

2.1 The Contingency Model of Leadership Effectiveness

The first and most extensively researched and validated of the contingency theories is the Contingency Model (Fiedler 1967, Fiedler and Chemers 1984). The leader characteristic included in the model is the leader's 'motivational orientation' which determines the priority of task accomplishment relative to interpersonal harmony. So-called 'task-motivated' leaders who emphasize structure, order, and clear directions are differentiated from 'relationship-motivated' leaders who emphasize cohesion, flexibility, and participative decision-making. The situational variable in the Contingency Model is the degree of predictability and orderliness in the task environment which is, in turn, determined by the supportiveness of followers, the clarity and structure in the task, and the amount of formal authority accorded the leader. Task-motivated leaders are most effective in situations where their emphasis on clear direction and structure fits well with existing levels of orderliness (i.e., highly predictable situations) or might contribute what is lacking and needed (as in highly unpredictable situations). Relationship-motivated leaders fare best when their more flexible and responsive style fits with the ambiguities of moderately predictable situations where some elements contribute to orderliness, but others might detract from it (e.g., situations requiring creativity or sensitivity to potential conflict).

Elaborations and extensions of the Contingency Model have shown that the fit between leader and situation was related to the leader's experienced stress and stress-related illness (Chemers et al. 1985), and that such stress affected the ability of the leader to make effective use of cognitive resources such as intelligence and job-related experience (Fiedler and Garcia 1987).

2.2 The Path–Goal Theory of Leadership

The basic premise of Path–Goal Theory (House 1996) is that the leader's most important function is to enhance subordinate motivation by clarifying the paths to the subordinate's goals and making it easier for the subordinate to traverse those paths. The empirical research on the model focused on how situational factors (including the subordinate's personality, needs, and task-relevant ability and the nature of the task environment), moderated the relative impact of the leader's 'structuring' or 'considerate' behavior on subordinate motivation, satisfaction, and performance. Path-goal theory predicts (with moderate to good empirical support) that leader structuring and direction will be most helpful and have their most positive effects when subordinates lack ability or are confronted with situations of task demand or complexity beyond their capability. Conversely, considerate and supportive leader behavior will be most satisfying and motivating for subordinates who are confronted with boring or aversive situations. In other words, leader behavior which provides what the subordinate needs most to cope with the situation will be most well received.

2.3 Normative Decision Theory

Vroom and Yetton (1973) presented a theory of group decision-making that argues that the most accurate and effective decisions will be made when the leader employs a decision strategy that is compatible with the characteristics of the decision situation. Decision strategies vary with the degree to which the leader allows subordinate participation in decision-making. Situational variables reflect the clarity of the decision environment (i.e., how much structured, interpretable information is available) and how cooperative and supportive the subordinates are. Autocratic decisions are most efficient when information is clear and followers are supportive. As decision-making ambiguity increases, because of follower resistance or equivocal information, more participative, democratic decision strategies offer the potential for greater

follower involvement leading to better information processing and higher motivation.

These three major contingency approaches, as well as a number of less prominent theories, share an emphasis on leader variables related to structuring vs. consideration and power-sharing and on situational variables related to clarity, predictability, and order. They all agree that the most effective leadership style depends on the fit or match to the situational contingencies.

3. Exchange Theories: From Transaction to Transformation

Leadership is a relationship between leader and follower, and like all relationships, must be grounded in both parties' perceptions that the relationship is fair and mutually beneficial. The quality of the relationship between leader and follower can vary widely.

3.1 Legitimacy

The act of choosing to follow a leader is based on the perception that the surrendering of some degree of personal autonomy will be advantageous to the follower. Hollander (1964) found that the legitimacy of an individual's status as a potential leader was based on followers' perceptions that the potential leader possessed task-relevant competence and loyalty to collective values and goals. A competent and trustworthy leader holds the promise of facilitating the goal attainment of the followers.

3.2 Quality of Exchange

Relationships between leader and follower can range from poor to excellent. Graen and Scandura (1987) report that high quality exchanges are characterized by a leader's willingness to afford the follower greater participation in decision-making, wider job scope, and more opportunities for personal development, and lead to greater follower motivation and commitment.

3.3 Transformational Leadership

The leader–follower relationship can move beyond a self-interested, *quid pro quo* transaction between leader and follower to an interaction which transforms the follower into a fully committed, mission-dedicated team member. House and Shamir (1993) argue that such transformative relationships are made possible by attaching the follower's personal identity and self-esteem to the collective mission. Bass (1998) has shown that transformative relationships are characterized by the perceptions of the leader as (a) possessing exceptional task-related competencies, (b) defining the group's mission in inspirational ways, (c) showing personal and individualized consideration for follower needs, and (d) providing a work environment for followers that is intellectually challenging and growth promoting. Teams led by such transformational leaders evidence much higher levels of motivation, commitment, satisfaction, and performance than average groups.

4. Perception and Judgment

Relationships depend on the perceptions of the parties involved—both perceptions of leaders by followers and other observers and judgments of followers by leaders.

4.1 Perceptions of Leadership

If leaders gain legitimacy by being perceived as competent and trustworthy, how are such judgments made? Extensive research by Lord and his associates (Lord and Maher 1991) reveals that judgments of leaders are influenced by two processes, labeled 'recognition' and 'inference.' People have expectations or 'prototypes' (i.e., sets of traits and behaviors associated with a social role) for leaders which provide the basis for recognizing a good leader. When a person behaves in a way that is seen as sufficiently similar to the leadership prototype, followers and other observers make the judgment that the person is, indeed, a leader or a good leader. Subsequent perceptions and memories of the leader are guided by the prototype, meaning that an individual already judged as a leader is more likely to be perceived and remembered as leaderlike in the future. Observers also infer a leader's competency by association with the outcomes of the group led. When groups are highly successful or unsuccessful, inferences are made to the leader. This tendency to attribute almost all group outcomes to the effects of leadership can be so exaggerated that it has been labeled the 'romance of leadership' (Meindl 1990).

4.2 Judgments of Followers

Leader actions toward subordinates are heavily determined by judgments of follower ability and motivation which are based on observations of follower behavior and task-related success and failure. Research indicates that leaders use a variety of contextual cues to understand the causes of a subordinate's success or failure, including the consistency of the performance, its uniqueness compared to the performance of other subordinates, and situational factors that might have affected performance (Mitchell and Wood 1980). However, the success and failure of subordinates have important implications for the

success or failure of the leader—a fact which makes leaders prone to ego-defensive judgments that can lead to placing the blame for failure on the individual subordinate while other possible causes (e.g., poor training, inadequate organizational support) are not considered (Brown 1984).

5. *An Integrative Summary*

The extensive empirical literature on leadership effectiveness can be summarized into three pervasive elements: image management, relationship development, and resource deployment. Effective leaders must be able to project an image of competency and honesty that legitimates the validity of their authority. They must establish motivating and rewarding relationships with followers by understanding follower needs and capabilities, and providing coaching and guidance that both facilitate good current performance and provide the possibility for personal growth. Finally, effective leaders must use the energies and resources of groups efficiently by matching the processes of communication, problem-solving, and decision-making to the task and organizational environment, i.e., by using directive and structuring process when the environment is clear and predictable and by using more flexible and participative process when the situation is equivocal and unpredictable.

See also: Attributional Processes: Psychological; Educational Leadership; Group Processes in Organizations; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Intrinsic Motivation, Psychology of; Leadership in Organizations, Psychology of; Leadership: Political

Bibliography

Ayman R 1993 Leadership perception: The role of culture and gender. In: Chemers M M, Ayman R (eds.) *Leadership Theory and Research: Perspectives and Directions*. Academic Press, San Diego, pp. 137–66

Bass B M 1998 *Transformational Leadership: Industrial, Military, and Educational Impact*. Lawrence Erlbaum Associates, Mahwah, NJ

Brown K A 1984 Explaining group poor performance: An attributional analysis. *Academy of Management Review* **9**: 54–63

Chemers M M 1997 *An Integrative Theory of Leadership*. Lawrence Erlbaum Associates, Mahwah, NJ

Chemers M M, Murphy S E 1997 Leadership and diversity in groups and organizations. In: Chemers M M, Oskamp S, Costanzo M A (eds.) *Diversity in Organizations: New Perspectives for a Changing Workplace*. Sage, Thousand Oaks, CA, pp. 157–88

Chemers M M, Hays R B, Rhodewalt F, Wysocki J 1985 A person–environment analysis of job stress: A contingency model explanation. *Journal of Personality and Social Psychology* **49**: 628–35

Fiedler F E 1967 *A Theory of Leadership Effectiveness*. McGraw-Hill, New York

Fiedler F E, Chemers M M 1984 *Improving Leadership Effectiveness: The Leader Match Concept*, 2nd edn. Wiley, New York

Fiedler F E, Garcia J E 1987 *New Approaches to Effective Leadership: Cognitive Resources and Organizational Performance*. Wiley, New York

Graen G B, Scandura T A 1987 Toward a psychology of dyadic organizing. *Research in Organizational Behavior* **9**: 175–208

Hemphill J K 1950 *Leader Behavior Description*. Ohio State University Personnel Research Board, Columbus, OH

Hofstede G 1984 *Culture's Consequences: International Differences in Work-Related Values*, abridged edn. Sage, Beverly Hills, CA

Hollander E P 1964 *Leaders, Groups, and Influence*. Oxford University Press, New York

House R J 1996 A path-goal theory of leadership: Lessons, legacy and a reformulated theory. *Leadership* **7**: 323–52

House R J, Shamir B 1993 In: Chemers M M, Ayman R (eds.) *Leadership Theory and Research: Perspectives and Directions*. Academic Press, San Diego CA, 81–108

Lord R G, Maher K J 1991 *Leadership and Information Processing: Linking Perceptions and Performance*. Unwin Hyman, London

Meindl J R 1990 On leadership: An alternative to the conventional wisdom. In: Staw B A (ed.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 12, pp. 159–203

Mitchell T R, Wood R E 1980 Supervisor's responses to subordinate poor performance: A test of an attribution model. *Organ Behaviour Human Decision Processes* **25**: 123–38

Stogdill R M 1948 Personal factors associated with leadership: A survey of the literature. *Journal of Psychology* **25**: 35–71

Vroom V H, Yetton P W 1973 *Leadership and Decision-Making*. University of Pittsburgh Press, Pittsburgh, PA

M. M. Chemers

Learned Helplessness

Learned helplessness refers to the maladaptive passivity shown by animals and people following experience with uncontrollable events. Learned helplessness also refers to the cognitive explanation of this phenomenon. The individual learns in one situation that responses and outcomes are independent, represents this learning as an expectation of helplessness, and then generalizes this expectation to other situations to produce passivity even if outcomes objectively can be controlled.

Researchers first described learned helplessness in the 1960s. Dogs were immobilized and exposed to a series of electric shocks. Regardless of what the dogs did or did not do, the shocks went on and off. Twenty-four hours later, these animals were placed in a long box with a barrier in the middle. Shocks were delivered periodically through the floor of the box, and the dogs could escape them by jumping over the barrier and

running to the other end of the box. Researchers found that animals previously exposed to uncontrollable shock acted passively when placed in a shuttle box. They simply sat there and endured the shocks. In contrast, animals without prior experience with uncontrollable shocks had no difficulty learning to escape the shocks.

These investigators proposed that the dog had learned to be helpless. When originally exposed to uncontrollable shock, it learned that nothing it did mattered because the shocks came and went independently of the dog's behaviors. They hypothesized that this learning of response-outcome independence was represented cognitively as an expectation of future helplessness that was generalized to new situations to produce a variety of deficits: motivational, cognitive, and emotional.

The deficits that follow in the wake of uncontrollability have come to be known as the *learned helplessness phenomenon*, and their cognitive explanation as the *learned helplessness model*. Learned helplessness in animals continues to interest experimental psychologists, in large part because it provides an opportunity to investigate the interaction between mind and body.

1. The Learned Helplessness Model

The learned helplessness model is at odds with a stimulus-response (S-R) view of learning because it proposes that the helpless animal learns not specific responses but rather general expectations affecting responses across a variety of situations. The helplessness model is a cognitive account of learning, and in the 1960s, it was a radical theory in a field long dominated by strict behaviorism.

Accordingly, much of the early interest in learned helplessness stemmed from its clash with the tenets of traditional S-R theories. Alternative accounts of learned helplessness were proposed by theorists who saw no need to invoke mentalistic constructs. Different alternatives were proposed, many of which emphasized an incompatible motor response learned when animals were first exposed to uncontrollable shock. This response was presumably generalized to the second situation where it interfered with performance at the test task. Said another way, the learned helplessness phenomenon is produced by an inappropriate *response* learned in the original situation rather than an inappropriate *expectation* (of response-outcome independence). For example, perhaps the dogs learned that holding still when shocked somehow decreased pain. If so, then they hold still in the second situation as well, because this response was reinforced previously.

Maier and co-workers conducted a series of studies testing between the learned helplessness model and the incompatible motor response alternatives. Several lines of research implied that the expectations emphasized by the learned helplessness model were operative. Perhaps the most compelling argument comes from the so-called *triadic design*, a three-group experimental design that shows that the uncontrollability of shocks is responsible for ensuing deficits. Animals in one group are exposed to shock that they are able to terminate by making some response. Animals in a second group are yoked to those in the first group, exposed to the identical shocks, the only difference being that animals in the first group control their offset whereas those in the second do not. Animals in a third group are exposed to no shock at all in the original situation. All animals are then given the same test task.

Animals with control over the initial shocks typically show no helplessness when subsequently tested. They act just like animals with no prior exposure to shock. Animals without control become helpless. Whether or not shocks are controllable is not a property of the shocks *per se* but rather of the relationship between the animal and the shocks. That animals are sensitive to the link between responses and outcomes implies that they must be able to detect and represent the relevant contingencies. A cognitive explanation of this ability is more parsimonious than one phrased in terms of incompatible motor responses.

Also arguing in favor of a cognitive interpretation of helplessness effects were studies showing that an animal can be 'immunized' against the debilitating effects of uncontrollability by first exposing it to controllable events. Presumably, the animal learns during immunization that events can be controlled, and this expectation is sustained during exposure to uncontrollable events. Learned helplessness is thus precluded.

Along these same lines, other studies show that learned helplessness deficits could be undone by forcibly exposing a helpless animal to the contingency between behavior and outcome. In other words, the animal is forced to make an appropriate response at the test task, by pushing or pulling it into action. After several such trials, the animal notices that escape is possible and begins to respond on its own. Again, the presumed process at work is a cognitive one. The animal's expectation of response-outcome independence is challenged during the 'therapy' experience, and hence learning occurs.

The studies described so far exposed animals to aversive stimuli (i.e., shocks). But it is also possible to produce helplessness by providing uncontrollable appetitive stimuli: food or water regardless of responses. Here an explanation in terms of incompatible motor responses becomes quite difficult to support. What is gained by holding still in response to food or water? Certainly not the reduction of pain.

Psychologists interested in humans, and particularly human problems, were quick to see the parallels between learned helplessness as produced by uncontrollable events in the laboratory and maladaptive passivity as it exists in the real world. Thus began

several lines of research looking at learned helplessness in people.

In one line of work, helplessness in people was produced in the laboratory much as it was in animals, by exposing them to uncontrollable events and seeing the effects on their motivation, cognition, and emotion. Unsolvable problems usually were substituted for uncontrollable electric shocks, but the critical aspects of the phenomenon remained. Following uncontrollability, people showed a variety of deficits.

Other studies further attested to the similarity between the animal phenomenon and what was produced in the human laboratory. Uncontrollable bad events made anxiety and depression more likely. Previous exposure to controllable events immunized people against learned helplessness. Similarly, forcible exposure to contingencies reversed helplessness deficits.

In another line of work, researchers proposed various failures of adaptation as analogous to learned helplessness and investigated the similarity between these failures and learned helplessness on various fronts. Especially popular was Seligman's proposal that reactive depression and learned helplessness shared critical features: causes, symptoms, consequences, treatments, and preventions.

2. *The Attributional Reformulation*

As these lines of work were pursued, it became clear that the original learned helplessness explanation was an oversimplification when applied to people. Most generally, it failed to account for the range of reactions that people displayed in response to uncontrollable events. Some people indeed showed pervasive deficits, as the model hypothesized, that were general across time and situation, whereas others did not. Further, failures of adaptation that the learned helplessness model was supposed to explain, such as depression, were sometimes characterized by a striking loss of self-esteem, about which the model was silent.

As used to explain human behavior, the learned helplessness model was therefore revised by Abramson et al. (1978) to emphasize the role of the person's interpretation of the cause of the original events. According to this reformulated theory, depression follows uncontrollability when a person explains an uncontrollable event with causes that are internal ('it's me'), stable ('it's going to last forever'), and global ('it's going to undermine everything I do'). Explanatory style refers to the causal explanations that people habitually offer for negative events. An internal, stable, and global style for explaining negative events is often described as a pessimistic explanatory style. Research shows that pessimistic causal explanations and explanatory style relate to depression as proposed by the reformulated learned helplessness model.

Explanatory style has since been extensively studied in its own right. Here is some of what has been learned. Explanatory style seems to take form as a coherent individual difference around 8 years of age, when children's cognitive abilities allow them to think in causal terms. Twin studies suggest that explanatory style is moderately heritable, but there is no reason to believe that there is a specific optimism-pessimism gene. Rather, such characteristics as intelligence, physical prowess, and attractiveness—which are themselves heritable—set the stage for success or failure in a variety of domains, and it is these experiences that lead the individual to entertain optimistic or pessimistic explanations for important outcomes. Early loss, such as death of a parent, or early trauma, such as sexual abuse, makes an individual more pessimistic. Early success makes an individual more optimistic.

Social learning is also implicated in the origins of explanatory style. Studies show convergence between the explanatory styles of parents and their children, although the fine detail of how an explanatory style is transmitted across generations has yet to be explored. Messages from schoolteachers, peers, and the media about the causes of events may also be critical.

Explanatory style has a vast array of consequences other than depression. Pessimistic explanatory style has been linked to various negative emotions and poor morale, to giving up and ineffective problem solving, to academic, athletic, military, occupational, and political failure, to social estrangement, to poor health, and even to untimely death. The mechanism by which explanatory style leads to these disparate outcomes has yet to be fully specified, but it is likely that the pathway is overdetermined, involving biological, psychological, and interpersonal processes.

3. *Conclusions*

Learned helplessness has become a popular line of investigation because it sheds light on humans' ills involving passivity. Helplessness theory, along with the attributional reformulation, suggests a ready explanation of such instances of passivity as well as interventions to prevent or remedy them. There is a downside to the availability of powerful theory. Some applications have been too zealous, overstating the similarity between learned helplessness on the one hand and a given failure of adaptation on the other. When applied carefully, however, helplessness theory can be highly useful, suggesting how maladaptive passivity can be changed or prevented.

Investigations of learned helplessness are ongoing. Laboratory studies with animals have delved into the biochemical basis of the helplessness phenomenon, suggesting that learned helplessness may have as much to do with anxiety as with depression. Furthermore, animal studies suggest that experience with uncontrollable events can sometimes disrupt the immune system, making illness more likely. Survey researchers

have started to look at how explanatory style and its consequences differ across cultures. What is identified in the Western world as a pessimistic style may be more common in Asian cultures that emphasize the collectivity and individual modesty. Finally, community intervention studies are exploring the use of cognitive-behavioral strategies for immunizing children against helplessness and its negative consequences.

See also: Adulthood: Dependency and Autonomy; Caregiving in Old Age; Classical Conditioning and Clinical Psychology; Conditioning and Habit Formation, Psychology of; Explanatory Style and Health; Nursing Homes; Operant Conditioning and Clinical Psychology

Bibliography

Abramson L Y, Seligman M E P, Teasdale J D 1978 Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology* **87**: 49–74

Buchanan G M, Seligman M E P (eds.) 1995 *Explanatory Style*. Erlbaum, Hillsdale, NJ

Chang E C 2001 *Optimism and Pessimism: Implications for Theory, Research and Practice*. American Psychological Association, Washington, DC

Gillham J E (ed.) 2000 *The Science of Optimism and Hope: Research Essays in Honor of Martin E P Seligman*. Templeton Foundation, Philadelphia, PA

Maier S F, Seligman M E P 1976 Learned helplessness: Theory and evidence. *Journal of Experimental Psychology: General* **105**: 3–46

Mikulincer M 1994 *Human Learned Helplessness: A Coping Perspective*. Plenum, New York

Overmeier J B, LoLordo V M 1998 Learned helplessness. In: O'Donohue W T (ed.) *Learning and Behavior Therapy*. Allyn and Bacon, Boston, MA, pp. 352–73

Peterson C M, Maier S F, Seligman M E P 1993 *Learned Helplessness: A Theory for the Age of Personal Control*. Oxford University Press, New York

Seligman M E P 1975 *Helplessness: On Depression, Development, and Death*. Freeman, San Francisco

M. E. P. Seligman and C. Peterson

Learning and Instruction, Social-cognitive Perspectives

1. Learning and Development in the Nature–Nurture Debate

Development and learning were for decades a matter of distinct ways of research and theoretical discussion, with a switch from the primacy of one of the poles, for example, from development as a basis for learning, to learning as a cause of development, with some radical positions, for example, learning without development. But the development–learning dilemma and the larger nature–nurture debate, even in the more recent forms of genetics and experience, has long hindered theoretical and methodological attention to the communicative social and cultural framing of the learning and testing situations considered. In the last decades of the twentieth century, important shifts in the research paradigms under consideration brought forward data that could not be interpreted anymore in terms of social factors as external variables affecting the learning of biopsychological individuals from the outside. It also required the consideration of the critical role played by social interaction in learning and instruction and the social construction of cognition within cultural activities.

This research has been built on a long tradition (e.g., the seminal works of scholars like Baldwin, Piaget, Wallon, Vygotsky, or Luria, and more recently, Bronfenbrenner). However, although it is one of the fundamental tenets of the social and behavioral sciences that human behavior and development is, at least in part, a creation of social forces, social interaction, and social appropriation, it is also true that theory and method do not always coincide in developmental work. Thus, there is a continuing search for the role of the 'social-interactive' in developmental theory and research.

One persuasive example is the enrichment and transformation of social-learning theory by cognitive dimensions to get a better understanding of the social and collective foundation of action and thought. Another example of this continuous struggle for a better match between theory and method is the revisitation of early twentieth century social constructivist scholars, such as Baldwin or Piaget . A third example is the ongoing mandate of interactionism, or the recent stream of 'postmodern' efforts at articulating reality as social construction. Finally, the recent debates about collectivism versus individualism deserve mentioning in this context.

On the surface, the fundamental role of social forces and social processes in the evolutionary and ontogenetic origins of behavior and the human mind seems clearly and widely acknowledged. Yet the translation of the intellectual agenda into scientific evidence continues to be incomplete. The non-social person-centered research paradigms seem to persist in the study of cognition, even when the fundamental role of context is obvious.

Many have tried to achieve a greater match between theory and empirical methods. At the turn of the century, several reviews have appeared that aim at similar goals (e.g., Baltes and Staudinger 1996, Carugati 1999, Resnick et al. 1991, Van der Linden et al. 2000, Perret-Clermont et al. in press).

These shifts in the perspectives on learning have been accompanied by changing views on the nature of

psychological development, the interdependencies between cognition, language, socialization, motivation and identity; and also on the nature of knowledge, no longer tied to individuals but relative to a community and subject to change. All of these approaches could still gain by devoting more attention to the critical role played by the communication and the construction of meaning in learning and instruction. This could enhance a revisitation of the current understanding of cognitive development, and contribute landmarks to the current educational endeavors that aim at encouraging learning together (not as isolated individuals), with relevant semiotic mediations on substantive issues.

2. *Gradual Shifts of the Paradigms: from the Individual to the Societal*

Studies on learning have operated gradual shifts of paradigms, from the overemphasis of the biopsychological functioning of the individual learner, where failure is necessarily seen as a sign of incompetence, to taking into account other social factors, either in a perspective inspired by Lewin's notion of field, or in a more macrosociological approach. Notably, the latter has called attention to major social problems in discrepancies in learning performances in students in schools, according to their social position (socioeconomic background, gender, ethnicity), educational methods and their hidden curriculum, the streaming structure of school institutions, etc.; and differences between the same subjects' performances in and out of schools. At the psychosocial level, evidence has pointed to field effects on performances. Since Orne's pioneering work in 1962, many studies have shown that the assessment of the subjects' competencies is affected by social factors such as the linguistic, motivational, attributional, identity characteristics of the subjects; teachers' expectations; individual and public dimensions of the situation; and the climate of cooperation or competition. The rather controversial results of these studies have called for a more integrative theoretical framework, articulating the various levels of analysis from the individual to the societal, in order to take into account the ways in which these factors interact. This relies upon precise descriptions of the networks of sociocognitive processes within which subjects think, behave, and learn (e.g., in the laboratory, at home, at school, in the street), and observations of the impact these experiences have on individual and collective cognition.

This has led to a new shift from determinism to constructivism, and from product-oriented studies to process-oriented ones. Studies were made on the interdependency between successful formal learning in schools and family socialization. Closer attention was paid to the teacher's role in defining precise expectations and giving adequate feedback to the students, on the one hand, and to the communication modes and patterns elicited in the learning and testing situations, on the other.

3. *Learning as a Sociocognitive Activity in Which Partners have an Essential Role*

It has been established that interaction, in order to be instructive, does not always have to be between a child (or novice) and an adult (or teacher/expert), but can be just as effective if it takes place between equally novice peers, given a certain number of conditions and for certain tasks. But this called for detailed studies of the influence of social interaction on the appropriation of cognitive tools. Some authors have illustrated the effects of the destabilization of the previous subjects' procedures (e.g., in problem-solving tasks). Destabilization triggers the questioning of, and the change in procedures and representations concerning, a given task (e.g., reasoning by recurrence; the notions of left and right; hypothetico-deductive reasoning in a fictitious weighting task). The control function takes on various forms: control can be exerted by triggering a verification of a statement or solution proposed by the partner, or by simply agreeing with or reformulating the partner's proposal. There appears also to be an interdependency between the type of problem to be solved, the cognitive functioning of the individual and the sociocognitive functioning of the two or more partners.

Recovering the importance of conflict, both in social and in developmental psychology, the notion of 'sociocognitive conflict' was introduced in order to stress the effect of the simultaneous confrontation of different individual perspectives, or points of view, during social interaction that necessitates and gives rise to their integration within a new cognitive organization. These experiments have explored under which conditions such confrontations with other points of view induce learning.

A further claim is made in studies on 'social marking' that show that the social regulations that govern a given interaction constitute an important factor for the building of a new cognitive coordination, because they bring a symbolic order that conveys meaning to the tasks.

4. *Learning as Participation in Social Discourse*

The problems of the meaning of a cognitive task, and the expectations activated in the subjects questioned both in experimental and in a more concrete school situation, are well-known in developmental psychology. Rose and Blank's studies (1974) on the role of double questioning in Piagetian tasks and Donaldson's (1978) study of tasks which 'make human sense' for young children, are only several examples of well-established studies that have had an important

influence on the study of communication in learning and testing situations, bringing forth evidence for the inextricability of meaning and expectations in every cognitive activity. A series of studies whose common ground is often the notion of 'contract' (inspired by Rommetveit's communication contract) have renewed the general claim that experimental and pedagogical situations are interactional settings governed by specific contracts, made up of explicit and implicit rules on the basis of which social actors adjust their behavior.

Failures in learning are reappraised as failures to communicate, and to convey meaning. Attention is then paid to the interactional patterns between teachers and students, and to the nature of the regulations that permit interlocutors to establish joint attention as a common object of discourse and maintain this common ground and mutuality. Learning is understood as participation in a social process in which meaning takes form (Bruner 1990), scaffolded by language and culture. And on the instructional side, authors explore the role of discourse in teaching.

5. *Studies of Learning Have Encouraged a Reconsideration of the Idea of Development*

Gradually, learning is appearing as the sociocultural experience of a person who is experiencing his or her agency within a history of social encounters, joint activity or *post hoc* individual or social meta-reflection. The basic research paradigm can no longer take the simple form of 'pretest–treatment–post-test,' as this chain of events itself gets its meaning partly from the setting and social context in which it is embedded. The pretest is already an interaction and in some ways a treatment from which the subject will draw information and experience on how to behave at the next step. What is it then that develops or that is learned along the different moments of a person's encounters with his or her social, semiotic and physical environment? The metaphors of children's developmental growth have usually been that of ascending lines, ranging from straight lines and scales to curves and spirals, all having in common, some vector towards fixed endpoints, with a general flavor of biological or cultural absolutes (such as formal reasoning, autonomous thinking, generalized and transferable competencies, etc.). The cognitive competency of the adult (teacher, mother, etc.) was usually taken for granted.

When the learner is considered as being the partner of joint activities and conversations in which he or she has stakes and continuously unravels and renegotiates rules, norms, and expectations, in changing social environments, then the form of development loses its determinism. New metaphors have been emerging to designate this sociocognitive activity in terms of gradual participation in socially shared cognition, and moves from peripheral participation to more central social positions in communities of practice, apprenticeship and guided participation, distributed cognition, the socialization to discourse in activity settings. It is becoming clearer that not only individual novices and tutors learn, they learn to become social learners, but also social entities.

Hence the shifts in the research paradigms have brought about very different understandings: learning as nurturing, a stage (or step-by-step) process of individual development; or learning as apprenticeship, via guided participation into the skills, discourses, and competencies of given communities of practices; or learning as co-construction and negotiation of activities, discourses, understandings, and identities.

See also: Cognitive Styles and Learning Styles; Cooperative Learning in Schools; Instructional Psychology; Piaget, Jean (1896–1980); Piaget's Theory of Human Development and Education; Teaching and Learning in the Classroom; Vygotskij, Lev Semenovic (1896–1934); Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Baltes P B, Staudinger U M (eds.) 1996 *Interactive Minds*. Cambridge University Press, Cambridge, UK

Bruner J S 1990 *Acts of Meaning*. Harvard University Press, Cambridge, MA

Carugati F 1999 From Piaget and Vygotsky to learning activities: A long journey and an inescapable issue. In: Hedergaard M, Lompscher J (eds.) *Learning Activity and Development*. Aarhus University Press, Aarhus, Denmark, pp. 211–32

Donaldson M 1978 *Children's Minds*. Fontana, Glasgow, UK

Orne M T 1962 On the social psychology of ten psychological experiments: With particular reference to demand characteristics and their implications. *American Psychologist* **17**: 776–83

Perret-Clermont A-N, Pontecorvo C, Resnick L, Zittoun T, Burge B (eds.) in press *Joining Society: Social Interaction and Learning in Adolescence and Youth*. Cambridge University Press, New York

Resnick L B, Levine J M, Teasley S D (eds.) 1991 *Perspectives on Socially Shared Cognition*. American Psychological Association, Washington, DC

Rose A, Blank M 1974 The potency of context in children's cognition. An illustration through conservation. *Child Development* **45**: 499–502

Van der Linden J, Erkens G, Schmidt H, Duffy T 2000 Collaborative learning. In: Simons P R J, Van der Linden J (eds.) *New Learning*. Kluwer Academic, Dordrecht, Netherlands, pp. 37–54

A. N. Perret-Clermont and F. Carugati

Learning and Memory: Computational Models

Models of neural basis of learning and memory (also known as computational models or neural networks) are computer programs that simulate the perform-

ance of various brain structures in learning and memory, based on the theorized function of those brain structures. These models are able to produce results that match the behavioral results from human and animal studies. These models test the ability of existing theories to simulate behavioral results and also make novel predictions which lead to new experiments.

The basic structure of neural networks is a set of units (or nodes) which represent neurons in the brain. The connections between these units represent the synaptic connections between neurons. The connections are plastic and can change their strength based on some learning algorithm. The nodes are grouped into layers that have both forward and backward connections. These backward connections may be recurrent, whereby active units can feedback their activity onto other active units and form a pattern of activity for each input. In this way, inputs can enter the network, cause changes in connections within the layers, and produce an output that represents the learned behavior.

Models can be applied to learning and memory at different levels of specificity. At a general level, psychological models deal with a mathematical formula which is theorized as a mechanism for learning but do not apply this mechanism to any particular brain region. At a more complex level, qualitative models consist of a central concept or metaphor that attempts to capture the essence of the function of a brain region. Finally, neural networks apply mechanisms to particular brain regions based on the known circuitry of the brain region thought to be responsible for a particular learning task.

1. *Psychological Models*

The most influential psychological model of learning and memory is the Rescorla–Wagner (1972) model. This model proposed a mathematical equation which could account for learning simple associative tasks such as classical conditioning. In classical conditioning, a neutral stimulus (the conditioned stimulus or CS) is paired with a response-evoking stimulus (the unconditioned stimulus or US). Early in training, a reflexive response is elicited by the US while the CS elicits no response. Following repeated pairings of the CS and US, the CS begins to elicit a learned response (the conditioned response or CR).

The Rescorla–Wagner (1972) model assumes that the change in association between the CS and the US is a function of the difference between the US and an animal's expectation of the US given all CSs present on the trial. Because the discrepancy, or 'error,' between the animal's expectations and what actually occurs drives learning in this theory, the theory is referred to as an 'error-correcting' learning rule.

The Rescorla–Wagner model with this single simple error correction mechanism can account for much learning in classical conditioning including simple acquisition, discrimination learning, blocking, and conditioned inhibition. While the Rescorla–Wagner model can account for many learning phenomena it does not address where in the brain their proposed mechanism acts.

2. *Neural Network Models*

Psychological models such as the Rescorla–Wagner rule have been incorporated into subsequent neural network models. Therefore the mechanism of error correction can be solved within the context of specific neural circuits.

Neural network models have been put forth which address various kinds of learning and memories such as motor learning in the cerebellum, declarative memory, spatial learning, classical conditioning in the hippocampus, and conditioned fear in the amygdala.

3. *Motor Learning and the Cerebellum*

Motor learning is best understood through work on classical eyeblink conditioning in the rabbit. In classical eyeblink conditioning, a tone CS is paired with a corneal air puff US. The US elicits a reflexive eyeblink. Following repeated pairings of the tone and air puff, the tone begins to elicit a conditioned eyeblink response. Not only is a conditioned eyeblink elicited, but it is timed so that maximal eyelid closure occurs at about the time of the air puff arrival. In other words, not only is it learned that the tone predicts the air puff, but that the tone predicts the air puff's arrival at a specific point in time.

The acquisition of a conditioned eyeblink has been found to involve the cerebellum and brainstem structures (for review see Anderson and Steinmetz 1994). All of the cerebellar models of eyeblink conditioning to be discussed have a foundation in the earlier theories of Marr (1969) and Albus (1971). They proposed theories for a role of the cerebellum in motor learning and the error correction of ongoing movements. The cerebellum has two distinct input pathways: the mossy and parallel fiber system, and the climbing fiber system. These two inputs converge both on Purkinje cells in the cerebellar cortex and on neurons in the deep cerebellar nuclei. Marr and Albus proposed that these two systems are involved in motor learning. The convergence of these two inputs was thought to alter Purkinje cell responsivity to parallel fiber inputs. Marr (1969) proposed that the Purkinje cell responsivity would be facilitated by conjunctive activity in the parallel and climbing fiber systems while Albus (1971) proposed that the Purkinje cell responsivity would be reduced. Albus's (1971) prediction was demonstrated by Ito and Kano (1982) who discovered the phenomena of long-term depression or LTD, in which sim-

ultaneous activity in the parallel and climbing fiber systems caused a reduction of Purkinje cell responsivity to parallel fiber inputs. This change in Purkinje cell responsivity is the basis for many computational models of cerebellar function in motor learning.

As mentioned previously, the cerebellum and brainstem structures are necessary for eyeblink conditioning. The tone information is transmitted to the cerebellum via the mossy and parallel fiber pathway while the air puff information is transmitted via the climbing fiber pathway. The convergence of tone and air puff information in the cerebellum causes an increase in activity in the cerebellar deep interpositus nucleus which is associated with the conditioned response. Models of eyeblink conditioning have attempted to simulate the action of the cerebellum through a variety of mechanisms.

Some cerebellar models have been put forth which work on the error-correction principle of the Rescorla–Wagner (1972) model (Gluck et al. 1995). In this model, the cerebellar CR output is fed back to the inferior olive, which is the source of the climbing fiber input that conveys air puff information to the cerebellum. This feedback from the cerebellum is inhibitory. Therefore, as learning occurs, the cerebellum is inhibiting the air puff information from entering the cerebellum. In this way, the error signal of the air puff is reduced as learning to respond to the tone occurs.

Other cerebellar models are more concerned with the real-time aspects of the conditioned response. Not only is a conditioned eyeblink learned, but this response is well timed with regard to the arrival of the US air puff. Many models have been put forth which can account for the well-timed nature of the conditioned response through various mechanisms. Some cerebellar models simulate the real-time characteristics of the conditioned eyeblinks by way of a feedback loop that is similar in nature to the error-correction feedback loop to the inferior olive (Gluck et al. 1995). The feedback loop that is responsible for the well-timed nature of the conditioned eyeblink returns cerebellar CR output to the pontine nuclei, which are the source of the mossy fiber inputs. Therefore, there is a convergence of CS and CR information that allows for the proper timing of the CR. Another recurrent cerebellar model produces well-timed CRs due to recurrent connections within the cerebellar cortex between granule, Golgi, and Purkinje cells (Buonomano and Mauk 1994). Another type of real-time cerebellar model is based on tapped delay lines (Desmond and Moore 1988). In a tapped delay line model, the well-timed CR is due to a series of synapses arranged so that the CS arrival to the Purkinje cells is delayed until the proper time to coincide with the US.

A more physiologically accurate model of the cerebellum was put forth by Fiala et al. (1996) who modeled cerebellar timing of the conditioned eyeblink response using a model of the metabotrophic glutamate receptor in cerebellar Purkinje cells. The metabotrophic glutamate receptors elicit slow responses by way of phosphoinositol hydrolysis and calcium release, which bridges the time interval between the CS and US onsets. Temporal correlation of the metabotrophic responses and climbing fiber inputs produce phosphorylation of AMPA and calcium-dependent potassium channels. This phosphorylation leads to long-term depression (LTD) of AMPA receptors. The phosphorylation of calcium-dependent potassium channels reduces baseline membrane potential and also reduces Purkinje cell firing rates in the CS–US interval. The decrease in Purkinje cell firing rates releases interpositus cells from inhibition and allows them to fire at the proper time. Overall, the Purkinje cells are responsible for timing the conditioned response while the interpositus is responsible for calibrating the strength of the response.

All of these cerebellar models are based on the same anatomy and connections and can simulate most of the same set of trial-level and real-time characteristics of eyeblink conditioning through various mechanisms. The question which remains is whether all of these mechanisms are physiologically realistic and may be operating in parallel in the cerebellum, or whether only one of these proposed mechanisms is correct. In this way, computational models are driving future empirical studies that should further advance the understanding of the exact mechanisms of motor learning in the cerebellum.

4. *Hippocampal Models*

Another brain region which has been modeled by many investigators is the hippocampal region. The hippocampal region consists of the hippocampus, dentate gyrus, subiculum, and entorhinal cortex. The hippocampus is thought to be involved in many forms of learning including declarative and episodic memory (the remembering of facts and events), spatial learning, and some forms of classical conditioning.

One of the earliest and most influential models of hippocampal region function was proposed by Marr (1971). Based on the hippocampal anatomy and physiology, Marr sought to infer the information processing capacity of the region. Marr's basic idea was to distinguish separable roles in memory for the archicortex (including the hippocampus) from the neocortex. He assumed that the chief role of the neocortex was to store large complex event memories. The role of the hippocampus was to be a separate processor which could rapidly store event memories and then allow gradual transfer of this pattern to the neocortex.

Marr's model consisted of two layers of cells. Inputs cause activity in the first A layer of cells which project onto the second B layer of cells. The B cells in turn project back to the A cells. All connections between

cells are modifiable, but they are simplified to allow only binary on or off values. A stored pattern can be retrieved if, when part is presented to the A cells, the evoked activity on the B cells feeds back to complete the original firing pattern on the A cells.

The network described by Marr is a form of autoassociator. An autoassociator network learns to associate an input pattern with an identical output pattern. An autoassociator requires three basic features: (a) a high degree of internal recurrency among the principal cells; (b) strong, sparse synapses from external afferents, which could function as forcing synapses; and (c) plasticity at the synapses between co-active cells. These requirements are satisfied by the known physiology of the CA3 region of the hippocampus.

Fast temporary storage in an autoassociator is an important component of an episodic or declarative memory system (Alvarez and Squire 1994, Hasselmo et al. 1996, Hasselmo and Wyble 1997, Murre 1996). It is generally assumed in these models that a relatively small temporary store in the hippocampus interacts with a relatively large long-term storage neocortical system.

For example, in a network model of hippocampal function in episodic memory (Hasselmo and Wyble 1997), a stimulus enters the neocortex via the sensory system and subsequently activates cells in the entorhinal cortex. Entorhinal inputs funnel multimodal sensory inputs to various portions of the hippocampus by forming a very compressed or fused representation of the inputs. The dentate gyrus serves as a pattern separator and forms less overlapping representations of the entorhinal activity patterns. These differentiated patterns of activity are sent to the CA3 region of the hippocampus. The CA3 region serves as a recurrent autoassociator and encodes and retrieves the features. The CA1 region of the hippocampus receives inputs directly from the entorhinal cortex and forms representations of these compressed inputs. These CA1 representations are then compared to the recalled CA3 representations. Through these interactions between the entorhinal cortex, dentate gyrus, CA3, and CA1, specific episodic memories can be formed and accurately retrieved. Presentation of a portion of the original input can lead to retrieval of the entire memory by retrieval of the correct episodic memory representation by CA3.

The hippocampus in turn feeds back to the neocortex and initiates activation patterns there. It may activate new cell populations, which are then added to the representation, or it may allow connections to form between active cells in the neocortex. The hippocampus may be required to present memories to the neocortex repeatedly, over some period of time, to allow the neocortex to integrate new knowledge without overwriting the old. This process is termed memory consolidation. Over time, as consolidation occurs, the sensory input is able to activate these neocortical cells directly, without hippocampal intervention, in order to retrieve the stored memory.

Spatial memory is a hippocampal-dependent task in rats (O'Keefe and Nadel 1978), and many models of hippocampal processing in spatial learning have been based on autoassociative models of the hippocampal region. One strategy is to define spatial maps as composed of sets of complex associations representing places (McNaughton and Nadel 1990). In one place, there would be many views, depending on which direction the animal is facing. The hippocampal autoassociator would be able to map from one of these views to the full representation of the current place. Given this interpretation, spatial learning need not differ from other types of hippocampal learning.

The hippocampus is also thought to be involved in classical conditioning tasks that require learning about nonreinforced configurations of stimuli, contextual information, or relationships that span short delays. Several recent computational models have focused on possible information-processing roles for the hippocampal region (e.g., Eichenbaum et al. 1992, Gluck and Myers 1993, Myers et al. 1995, Schmajuk and DiCarlo 1991, Sutherland and Rudy 1989). Most of these models assume that while the hippocampus is required for some complicated forms of stimulus association, the neocortex and cerebellum are sufficient for simpler stimulus–stimulus associations such as those that underlie classical conditioning.

The key idea of one of these models, the Gluck and Myers (1993) cortico-hippocampal model, is that the hippocampal region is able to facilitate learning by adapting representations in two ways. First, it is assumed to compress, or make more similar, representations of stimuli that co-occur; second, it is assumed to differentiate, or make less similar, representations of stimuli that are to be mapped to different responses.

This model fits nicely with the hypothesized function of the entorhinal and hippocampal function previously mentioned. That is, entorhinal activity patterns are compressions of all the sensory inputs to the hippocampal system and the dentate gyrus separates these inputs into different patterns of activity.

More recent computational models have not only dealt with the hippocampal region in general, but attempted to differentiate the function of the hippocampus from the entorhinal cortex (Eichenbaum et al. 1994, Myers et al. 1995). In this way, these models have served to drive new empirical studies that test the predictions of the effects of different types of selective lesions within the hippocampal region.

It is very interesting that while the hippocampus is thought to be involved in a variety of apparently diverse memory types, the application of the autoassociator mechanism can account for these different memory functions. In this way the modeling work has served to form a more united understanding of the function of the hippocampus which is consistent

for spatial learning, episodic memory, and stimulus representations that may not have been apparent from the diverse empirical data.

5. Conditioned Fear and the Amygdala

Another brain region which has been modeled is the amygdala. The amygdala is thought to be involved in emotional memories, especially conditioned fear responses. The amygdala, specifically the lateral nucleus, receives sensory inputs from all modalities as well as highly processed sensory information from the association cortices and hippocampus. The fear circuit is best understood for auditory conditioned stimuli. Sensory information reaches the amygdala from the thalamus via two parallel pathways: a direct connection from the extralemniscal areas of the thalamus, and an indirect connection via primary and secondary auditory cortices.

Armony et al. (1995) implemented a model of the conditioned fear network. The architecture of the network is modular; relevant structures are represented by interconnected modules based on the anatomical connectivity of the amygdala. In the model the input patterns represent a range of tone frequencies. As a result of training with the input patterns, each unit develops a receptive field, that is, it responds only to a subset of input patterns centered around a best frequency. For simulating conditioning, all the input patterns are presented with one pattern chosen as the CS and paired with the US. This pairing resulted in some units showing frequency-specific retuning of their receptive fields, which matches with empirical results.

6. Future Physiological Models

A criticism of the computational models described above is that while the network nodes capture some of the response characteristics of real neurons, they are too simplistic in their input–output functions to represent the full range of physiological complexity in real neurons. Obviously, many factors such as neurotransmitter types, different receptor types, and membrane dynamics play an essential role in the responsivity and plasticity of neurons. The computational models previously described do not represent these features. Another class of models addresses these issues by modeling the biophysical, structural, and synaptic characteristics of individual neurons (e.g., Fiala et al. 1996). While these physiological models can simulate the action of individual neurons, they do not deal with large numbers of neurons in network connectivity. Future computational models should attempt to account for both network properties of learning and memory while maintaining physiologically precise mechanisms.

See also: Amygdala (Amygdaloid Complex); Artificial Neural Networks: Neurocomputation; Cerebellum: Associative Learning; Classical Conditioning, Neural Basis of; Eyelid Classical Conditioning; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Marr, David (1945–80); Neural Networks: Biological Models and Applications; Neural Networks, Statistical Physics of

Bibliography

Albus J S 1971 A theory of cerebellar function. *Mathematical Biosciences* **10**: 25–61

Anderson B J, Steinmetz J E 1994 Cerebellar and brainstem circuits involved in classical eyeblink conditioning. *Reviews in Neuroscience* **5**: 251–73

Alvarez P, Squire L 1994 Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences* **91**: 7041–5

Armony J L, Servan-Schreiber D, Cohen J D, LeDoux J E 1995 An anatomically constrained neural network model of fear conditioning. *Behavioral Neuroscience* **109**: 246–57

Buonomano D V, Mauk M D 1994 Neural network model of the cerebellum: Temporal discrimination and the timing of motor responses. *Neural Computation* **6**: 38–55

Desmond J E, Moore J W 1988 Adaptive timing in neural networks: The conditioned response. *Biological Cybernetics* **58**: 405–15

Eichenbaum H, Cohen N J, Otto T, Wible C 1992 Memory representation in the hippocampus: Functional domain and functional organization. In: Squire L R, Lynch G, Weinberger N M, McGaugh J L (eds.) *Memory Organization and Locus of Change*. Oxford University Press, Oxford, UK

Eichenbaum H, Otto T, Cohen N 1994 Two functional components of the hippocampal memory system. *Behavioral and Brain Science* **17**: 449–518

Fiala J, Grossberg S, Bullock D 1996 Metabotropic glutamate receptor activation in cerebellar Purkinje cells as a substrate for adaptive timing of the classically conditioned eye-blink response. *Journal of Neuroscience* **16**: 3760–74

Gluck M A, Myers C E 1993 Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus* **3**: 491–516

Gluck M A, Myers C E, Thompson R F 1995 A computational model of the cerebellum and motor-reflex conditioning. In: Zornetzer S F, Davis J L, Lau C (eds.) *An Introduction to Neural and Electronic Networks*. Academic Press, San Diego, CA

Hasselmo M E, Wyble, B 1997 Simulation of the effects of scopolamine on free recall and recognition in a network model of the hippocampus. *Behavioral and Brain Research* **89**: 1–34

Hasselmo M E, Wyble B, Wallenstein G 1996 Encoding and retrieval of episodic memories: Role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus* **6**: 693–708

Hinton G 1989 Connectionist learning procedures. *Artificial Intelligence* **40**: 185–234

Ito M, Kano M 1982 Long-lasting depression of parallel-fiber-Purkinje cell transmission induced by conjunctive stimulation of parallel fibers and climbing fibers in the cerebellar cortex. *Neuroscience Letters* **33**: 253–8

Marr D 1969 A theory of cerebellar cortex. *Journal of Physiology* **202**: 437–70

Marr D 1971 Simple memory: A theory for archicortex. *Proceedings of the Royal Society, London, Series B* **262**: 23–81
McNaughton B, Nadel L 1990 Hebb–Marr networks and the neurobiological representation of action in space. In: Gluck M, Rumelhart D (eds.) *Neuroscience and Connectionist Theory*. Lawrence Erlbaum, Hillsdale, NJ
Murre J 1996 Tracelink: A model of amnesia and consolidation of memory. *Hippocampus* **6**: 675–84
Myers C E, Gluck M A, Granger R 1995 Dissociation of hippocampal and entorhinal function in associative learning: A computational approach. *Psychobiology* **23**: 116–38
O'Keefe J, Nadel L 1978 *The Hippocampus as a Cognitive Map*. Clarendon Press, Oxford, UK
Rescorla R, Wagner A 1972 A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In: Black A, Prokasy W (eds.) *Classical Conditioning II: Current Research and Theory*. Appleton-Century-Crofts, New York
Schmajuk N A, DiCarlo J J 1991 A neural network approach to hippocampal function in classical conditioning. *Behavioral Neuroscience* **105**: 82–110
Sutherland R J, Rudy J W 1989 Configural association theory: The role of the hippocampal formation in learning, memory and amnesia. *Psychobiology* **17**: 129–44

M. T. Allen

Learning and Memory, Neural Basis of

1. *Basic Learning- and Memory-related Neural Changes*

Much of our current understanding of basic memory-related neuronal changes and modifications stems from the pioneering research of Eric Kandel and his co-workers (Kandel 1998) who largely used an invertebrate, the marine mollusc *Aplysia* for their studies. Initially, they investigated simple memory mechanisms such as sensitization and habituation and found that there are both short- and long-term forms hereof and that the physiological correlates of the short-term forms are facilitated neuronal responses (increases in the magnitude of the excitatory post-synaptic potentials formed between sensory and motor neurons), and that of the long-term forms (repeated stimulus applications) required new ribonucleic acid (RNA) and protein synthesis (see *Protein Synthesis and Memory*), involving the growth of new synaptic connections between sensory and motor neurons. Later, they extended their work to gene expression and transcription factors which provided the genetic switch for memory consolidation and storage (see *Memory, Consolidation of*). A specific protein, named CREB, which stands for cAMP-responsive element binding protein, became a new standard label within this kind of research. cAMP stands for cyclic-adenosine monophosphate, a substance important for providing energy to cells. Other workers from both the invertebrate and vertebrate side demonstrated, for example, that adenylyl cyclase is an important enzyme for short-term memory (Zars et al. 2000) and long-term potentiation (LTP) as well as long-term depression (LTD). These are important mechanisms for memory consolidation (Lynch 2000) (see *Long-term Potentiation (Hippocampus)*; *Long-term Potentiation and Depression (Cortex)*; *Long-term Depression (Hippocampus)*). There is evidence that LTP and LTD lead to changes in the morphology of nerve cells, particularly to so-called synaptic plasticity (Coussens and Teyler 1996) and to changes in dendritic spine density (Engert and Bonhoeffer 1999).

The crucial questions are whether these findings can be directly transmitted to mammals and whether they hold for different kinds of memories (Eichenbaum et al. 1999). Here is less safe ground from current research (see, e.g., the findings of Zamanillo et al. [1999] against LTP as an essential prerequisite for memory formation) and the focus lies more strongly on metabolic changes which can be related to brain circuitries.

2. *Memory Representation Within the Mammalian Nervous System*

Amnesias (see *Amnesia*) in humans occur after a wide range of disease conditions (Markowitsch 2000). This fact alone implies that information processing in the mammalian brain is not controlled by a single center but, instead, by a distributed network. Furthermore, memory is also dependent on a large number of sensory, perceptual, attentive, emotional, and motivational processes each with its own anatomical substrates. The investigation of brain-damaged individuals consequently is of only limited predictive value for the action of the normal brain, as it demonstrates only what the brain can or cannot do without a particular region (Sometimes, a brain with a damaged region is compared with a chain in which a link is missing; this implies that with the lesioning approach the functions of the damaged area are, in a way, never attainable, as the respective region is missing). Therefore, the complementary approach of studying the intact brain with neuromonitoring (EEG, evoked potentials, magnetoencephalography) and the functional brain imaging methods methods (positron emission tomography, functional magnetic resonance imaging, single photon emission tomography) provides the necessary completion for inferences on the action of the brain (see *Functional Brain Imaging*).

In spite of the present-day range of methodological approaches for studying learning and memory representations in the brain, the old dichotomy between localizationists and antilocalizationists remains and can be perceived as one of ideology. The localizationists have a more mosaic-like view of information representation, they speak of 'centers' and 'modules,'

the antilocalizationists prefer a Gestalt-like view and emphasize interactions and networks. A bridge between the two can be built by assuming a mixture of hierarchical and parallel processes of information representation (McClelland and Rumelhart 1986; Barnden and Pollack 1991).

3. *Memory Systems and their Functional and Anatomical Correlates*

The study of mammalian and in particular of human memory has to acknowledge the existence of several memory systems and their different neural representation. Aside from short-term memory (see *Short-term Memory: Psychological and Neural Aspects*, there are at least four major long-term memory systems: (a) the perceptual representation system (PRS), (b) procedural memory, (c) the knowledge system (or semantic memory), and (d) episodic memory (Schacter et al. 2000, see also *Episodic and Autobiographical Memory: Psychological and Neural Aspects*). The PRS refers to the fact that nonconscious presentation of a stimulus may lead to a higher probability of re-identification of it at a later time point; procedural memory is usually skill based, examples are riding a car, skiing, or playing the trumpet—that is, nonconscious, automatic processing of information. Both of these systems, but sometimes also the knowledge system, are termed implicit memory (see *Implicit Learning and Memory: Psychological and Neural Aspects*). The knowledge system (semantic memory) refers to general facts ('world knowledge') of the kind: 'Madrid is the capital of Spain' or '12^2 is 144.' While the knowledge system consists of context-free facts, the episodic memory system (see *Episodic and Autobiographical Memory: Psychological and Neural Aspects*) represents context- (time, locus) embedded information. These episodes allow a mental time travel. One of the important distinctions between semantic and episodic memory is therefore that the latter allows to remember the episode and the former simply to know the fact.

3.1 *Short-term and Working Memory*

The concept of working memory derived from the earlier one of short-term (and long-term) memories. While, however, short-term memory is seen as the on-line holding of information for an initial time period after first confrontation with it, working memory refers to a multicomponent, active processing of information which also includes the transmission of already long-term stored information in a temporary buffer prior to retrieval, and which is composed of a controlling central executive system and a number of subsidiary slave systems (Baddeley 2000).

Numerous experiments using monkeys, cats, and rats have reported that bilateral damage to the prefrontal cortex interferes with the short-term storage of information and that single units in this region display working memory related neuronal firing (Goldman-Rakic et al. 2000, see also *Working Memory, Neural Basis of*). Results from evoked-potential recordings and functional imaging studies have confirmed the dominant role of prefrontal regions for working memory in the human brain. Some reports found that damage to parietal regions impairs the on-line processing of information. Consequently, a frontoparietal network is seen as engaged in the control of attention and in processes of on-line working with memory.

3.2 *The PRS and Procedural Memory*

The two principal memory systems, which function independent of limbic system structures and of conscious reflections, are procedural memory and priming. There are two basic forms of priming, perceptual and conceptual—perceptual priming refers to the direct match between the presently and the previously perceived stimulus, while conceptual priming refers to the same category or class of the present and the previously perceived stimuli. The distinction between these two forms is important as conceptual priming apparently is more complexly represented in the brain, recruiting more or other neural structures than perceptual priming. Consequently, conceptual priming deteriorates in age-related brain diseases and possibly even in patients with limbic-system related amnesias, while perceptual priming does not (see *Amnesia*). Conceptual priming has been related to prefrontal and even partly to medial temporal regions, while perceptual priming has much stronger affinities to unimodal areas of the posterior neocortex.

The procedural memory system is closely related to motor regions of the brain—to the basal ganglia, to portions of the frontal lobe, particularly of the premotor and motor cortex, and to cerebellar structures (some authors question, however, a specific role of cerebellar structures in procedural memory). Therefore, diseases such as Parkinson's disease are taken as models—at least during the initial disease process—for selective procedural memory deficits (Thomas-Ollivier et al. 1999).

While not much is known about the processes from information encoding to information retrieval, there is at least preliminary evidence that principally the same structures activated during the initial stages of information transfer are also relevant for the late ones and even hold for storage of this implicitly learned information.

3.3 *Brain Regions Relevant to Episodic Memory and the Knowledge System*

Episodic memory and the knowledge system are sometimes combined and then named declarative

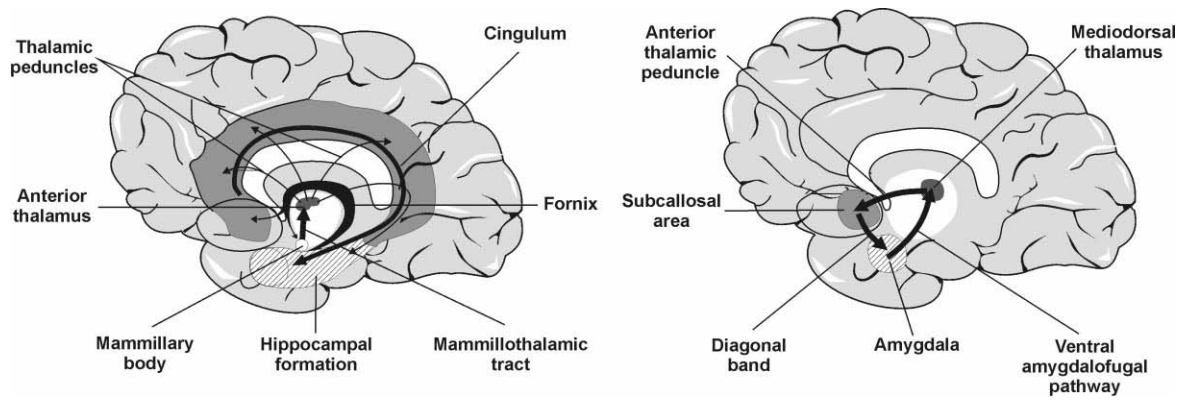

Figure 1
The medial (Papez) (right) and the basolateral limbic circuits (left). The Papez circuit is composed of the hippocampal formation, mammillary bodies, anterior thalamus and (conventionally) the cingulate cortex. These regions are interconnected via the fornix, mammillothalamic tract, thalamocortical peduncles and the cingulum. The components of the basolateral limbic circuit are the amygdala, the thalamic mediodorsal nucleus and the subcallosal area (basal forebrain); here interconnections are the ventral amygdalofugal pathway, the inferior thalamic peduncle, and the bandeletta diagonalis. It is assumed that both circuits are relevant especially for encoding of information associated with the episodic and knowledge memory systems and that the medial circuit is more strongly engaged in the cognitive and the basolateral one in the affective side of information processing

memory (Squire and Knowlton 1999, see also *Declarative Memory, Neural Basis of*). Both systems have in common that they—at least during the stages of memory encoding and consolidation—rely on structures of the limbic system (Mesulam 2000), the episodic memory system most likely in addition on prefrontal structures (see *Limbic System*). For retrieval of information, there seems to be a distinct lateralization; it appears that primarily right-hemispheric regions are engaged in episodic and primarily left-hemispheric ones in semantic-memory retrieval (Tulving et al. 1994, Markowitsch 2000).

The structures of the limbic system have long been discussed as important for transferring information from short- to long-term memory (see *Limbic System*). More recently, components of this system have been combined under the headings 'medial temporal lobe' and 'medial diencephalic memory system' and have been named bottlenecks for information transmission from short- to long-term memory (Markowitsch 2000). For the first system, the hippocampal formation was (and still is) seen as its hub for the second, medial thalamic nuclei and the hypothalamic mammillary bodies are core structures and Korsakoff's syndrome is one of the most well-known diseases associated with it (see *Hippocampus and Related Structures*; *Temporal Lobe*; *Hypothalamus*; *Korsakoff's Syndrome*). However, it is more a matter of philosophical preferences whether one sees both systems as separate entities or as one. Likewise, there has been a waxing and waning discussion of the contributions of individual regions to memory formation (Tulving and Markowitsch 1998). Aside from the hippocampal region, the amygdala and paralimbic cortices have to be listed for the medial temporal system, and certain medial and anterior thalamic nuclei, the mammillary bodies, and interconnecting fiber bundles for the medial diencephalic one; structures of the basal forebrain and the ventral striatum may be added (Mesulam 2000, Markowitsch 2000).

There is another division of the limbic system which receives support from a number of sources, namely the distinction between two interacting circuits: (a) the medial (or Papez-) circuit centered around the hippocampus, and (b) the basolateral limbic circuit (amygdaloid circuit). While Papez (cited in Markowitsch 2000) considered his proposed circuit (hippocampal formation → fornix → mammillary bodies → mammillothalamic tract → anterior thalamus → thalamocortical peduncles → cingulate cortex → cingulum → hippocampal formation) as principally engaged in the analysis of emotions, subsequent work showed that it plays a critical role for the transfer of information into long-term memory. The basolateral circuit is more closely related to emotional processing, but is also relevant for encoding the emotional valence of experiences. It includes the amygdala, the mediodorsal nucleus of the thalamus and associated regions, largely situated in the basal forebrain and in its extensions to anterior insular and temporopolar cortical areas plus interconnecting fibers such as the ventral amygdalofugal pathway, the anterior thalamic peduncle and the diagonal band (Fig. 1).

Bilateral damage to structures within the medial circuit is usually followed by severe memory disturbances—a fact which led researchers to speculate that

amnesia might arise from an interruption of interconnections within this circuit and that the structures of this circuit are necessary for consolidation and for the transfer from short- to long-term storage (see *Amnesia*). Bilateral damage to structures within the basolateral limbic circuit, on the other hand, seem to degrade memory processes by disconnecting the emotional from the mnestic properties of a given set of stimuli. For example, bilateral damage to the amygdaloid region may induce the Klüver–Bucy syndrome characterized by hyperorality, hyperphagia, agnosia, amnesia, hypersexuality, tameness/placidity, and hypermetamorphosis. Since information encoding is, especially for episodic (autobiographic) information, dependent on emotional content, damage to the amygdala is likely to have detrimental effects on memory. This indeed was shown in patients with Urbach–Wiethe disease; such patients manifest a selective bilateral damage (mineralization) of the amygdalae. Two such patients were presented stories containing emotionally charged items and asked to recall them. They were significantly impaired because they could not show the preferential encoding of emotionally-laden vs. neutral items (Cahill et al. 1995).

Lesions in the basal forebrain (e.g., medial septum, basal nucleus of Meynert) may also disturb memory processing, though the deteriorations on the memory side are usually less severe and enduring than those after medial temporal or medial diencephalic damage; personality disturbances and tendencies to confabulate, on the other hand, are more common.

The similarity between the amnesic conditions arising after bilateral damage to medial temporal and medial diencephalic structures led Squire et al. (1993) to conclude that these structures might be components of a common memory system and to speak of 'medial temporal–diencephalic amnesia.' While this seems likely, there may also be regional functional specializations. Patients with bilateral diencephalic damage seem to have less awareness of their memory, a higher tendency to confabulate, and a larger span of retrograde amnesia than those with medial temporal lobe damage.

3.3.1 Encoding and consolidation. While encoding refers to the more early stage of information acquisition, there are different views on the beginning and end of the consolidation phase. Sometimes, it is seen as directly succeeding the encoding phase, sometimes it is viewed as a process occurring after information has already been stored long term. As both encoding and consolidation of episodic and semantic memories engage structures of the limbic system, they can be perceived as closely related.

There are nevertheless single cases of patients with amnesias of various origins showing a wide range of unstable memory processing periods (Markowitsch 2000) so that it has to be concluded that there are several states and stages of memory consolidation which depend on complex anatomico–behavioral interdependencies. A comparison with repeated attempts of producing a complex painting may be apt. Still, medial temporal and diencephalic regions contribute substantially to both the stages of encoding and consolidating by triggering, accompanying, and assisting the steps from initial binding to the final assignment. Especially the hippocampus has been proposed to index (previously encoded) information and assign it to specific cortical loci. Amygdaloid and septal regions may evaluate and filter for importance—that is, they screen the biological and social significance of perceived stimuli.

Aside from the more 'classic' limbic regions, functional brain imaging results have pointed to a strong role of prefrontal cortical areas in both encoding and retrieving of information. Among the first studies were those of Tulving and co-workers (1994) who condensed their findings in what they termed the 'hemispheric-encoding-retrieval-asymmetry' (HERA) model of information processing. This model considers prefrontal regions as important for information processing, but assigns different roles to the left and right frontal lobes: the left one being necessary for encoding of semantic and episodic information and the right one for the retrieval of episodic memories. While this model has sometimes been questioned or modified, there is still accumulating evidence for its principal validity. Current research tries to specify whether there are differential contributions of dorsolateral, polar, and inferolateral prefrontal regions and to what degree the prefrontal contribution can be regarded as a process of working memory.

3.3.2 Storage of information. The ways in which information is—frequently life-long—stored in the brain, is one of the biggest riddles of the neurosciences. Most scientists agree that the nervous system contains or forms the medium for storage, the more concrete mechanisms vary between species—invertebrates with their more 'rudimentary' nervous system may indeed have adopted other representations than mammals. Genetic and biochemical processes may nevertheless be common denominators (Kandel 1998). For mammalian nervous systems, information may very likely be distributed along widespread cerebral cortical networks (McClelland and Rumelhart 1986) with or without auxiliary engagement of subcortical limbic structures (Markowitsch 2000). The expanded cortical association areas with their large number of neurons and neuronal connections are primary candidates for information representation. Especially for emotionally colored memories—such as our biographical episodes—the

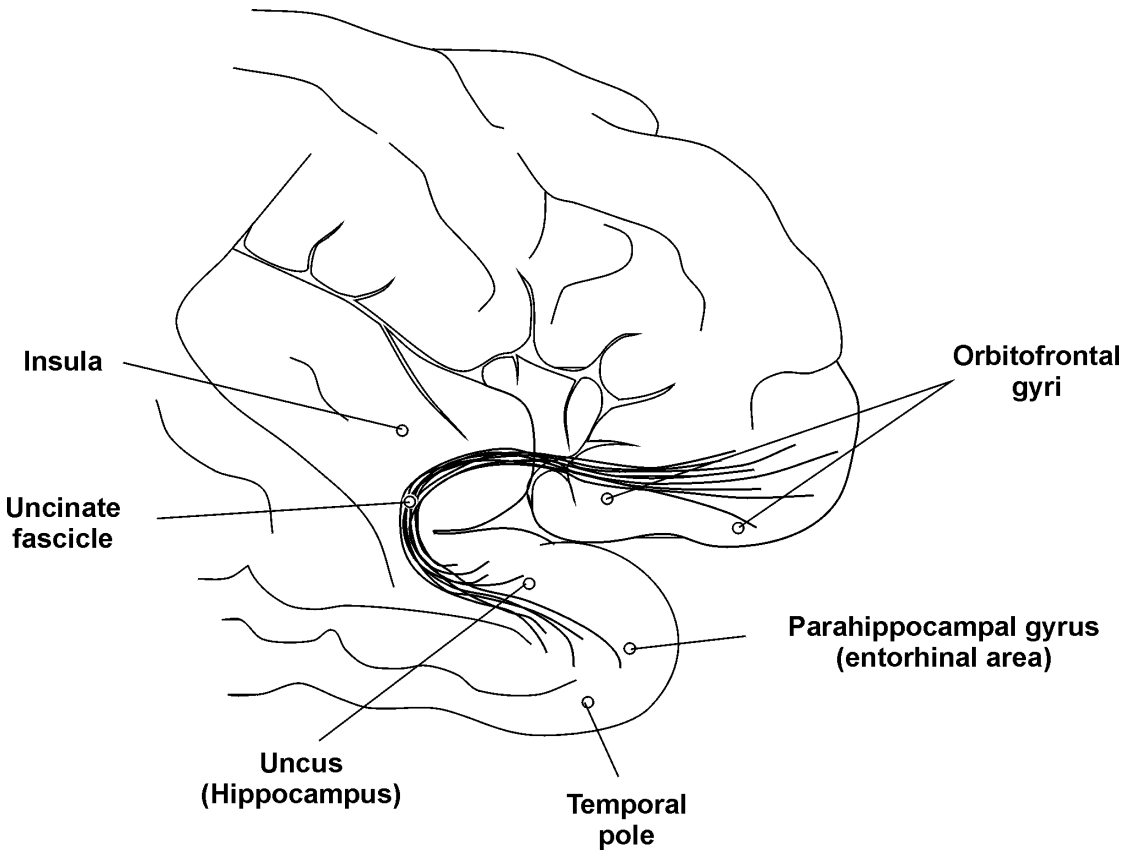

Figure 2
The regions of the temporal pole and the inferior lateral prefrontal cortex which are interconnected via the uncinate fascicle. This regional constellation of the right hemisphere is assumed to control the retrieval of episodic memories, and the left-sided one that of the knowledge system

cortical nets may engage and trigger affect-coding limbic regions for their rerepresentation or 'ecphory' (Markowitsch 2000, Sara 2000, Tulving and Markowitsch 1998). (Tulving reintroduced the term ecphory to describe the process by which retrieval cues interact with stored information so that an image or a representation of the information in question appears.)

There are, however, a few scientists who are of the opinion that most of our memories are represented outside our nervous system. Romijn (1997) considers a submanifest order of being as the principal storage place of memories. He is of the opinion that the whole universal in its spatial–temporal configuration would be permanently present in the submanifest order of being. Of this, a small portion would continuously materialize and thereby become available for psychic processing. Declarative memory would be found somewhere in the individual area of the submanifest order of being. Similar views are held by Sheldrake (1988) and Hameroff et al. (1998).

3.3.3 Retrieval of information. The conventional view is that retrieval activates similar (namely mainly limbic) brain structures as does encoding. This view may, however, be false or at least only partly correct: As mentioned above in connection with the HERA model of Tulving et al. (1994), the prefrontal cortex is a primary region for information retrieval. Converging evidence from brain-damaged patients and from studying normal subjects with functional imaging methods points to the necessity of (inferolateral) frontal and anterolateral temporal cortical regions for triggering the network-like stored information for successful retrieval (Markowitsch 1995, 2000) (Fig. 2). While a few authors even assume the temporal cortex to be the actual storehouse of information, here the view is favored that it provides the most direct access to the cortical network, while the prefrontal cortex is responsible for providing the effort necessary for a successful overall neural activation. Limbic structures of the medial diencephalon and the medial temporal lobe may have supporting roles in

context-rich information retrieval (Markowitsch 2000).

Retrograde amnesia—the failure of ecphory—may occur as a consequence of a progressively decaying cortical network—as is the case in patients with cortical dementia (e.g., Alzheimer's disease) or, sometimes, after a condition of severe hypoxia (e.g., after a heart attack and coma state). In this group of patients, the information is lost permanently. When the damage is restricted to the hippocampus and related structures in the medial temporal lobes, the retrograde amnesia may be temporally limited (covering a period of weeks in rodents, up to years or even decades in humans, leaving remote memories intact), suggesting to several authors that the hippocampus may play a role in the initial formation of new memories, and may somehow contribute to the gradual integration of new memories into neorcortical memory circuits (Squire et al. 1984, McClelland et al. 1995).

Alternatively, retrograde amnesia may be the consequence of a disrupted access to the net. Here, the expression 'mnestic block syndrome' has been coined to describe this phenomenon (Markowitsch 2000) which can be found both after manifest brain damage in the frontotemporal cortex and after so-called functional amnesia which is seen in patients with dissociative personality states (see *Amnesia: Transient and Psychogenic*). Such functional amnesias especially appear as a consequence of stress and trauma experiences (see *Stress, Neural Basis of*). In this second category of patients, the mnestic block is usually selective, affecting primarily emotion-laden personal episodes, and it is assumed that the information—the engram—is in principle still manifest in the cortical nets.

Furthermore, static and functional imaging results from both neurologic and psychiatric patients point to alterations in especially the right temporofrontal regions when autobiographic information retrieval is disrupted and to corresponding changes in the left temporofrontal regions when semantic retrieval is impaired (Markowitsch 1995, 2000).

4. *The Learning and Memory Networks—Conclusions*

Learning and memory engage a wide range of brain regions—nevertheless, a conclusion that it is the total mass of the nervous system which codes information for storage would be a false one. Instead, the principles of convergence and divergence of information transmission hold throughout the nervous system and there are both modality- and material-specific channels for information encoding, storage, and even retrieval and there are bits of information which are rich in affect and consciously reflected (and consequently require wide neural nets), while others are processed only subliminally and may never gain conscious representation. Furthermore, information is not represented in a static manner within the brain – mood, the environment, focused attention, internal associated cues, hormones, and many additional factors influence, modify, and rearrange our knowledge. The phenomena of state-dependent learning and retrieval and of false memories are examples for the susceptibility and vulnerability of our memories. The concept of plasticity has gained considerable support from invertebrate to human research and also the concept of life-long learning, proposed by the social sciences, demonstrates that information representation is much less hard-wired and rigid than formulated by traditional frameworks.

See also: Amnesia; Amnesia: Transient and Psychogenic; Declarative Memory, Neural Basis of; Episodic and Autobiographical Memory: Psychological and Neural Aspects; Functional Brain Imaging; Hippocampus and Related Structures; Hypothalamus; Implicit Learning and Memory: Psychological and Neural Aspects; Korsakoff's Syndrome; Limbic System; Memory, Consolidation of; Protein Synthesis and Memory; Short-term Memory: Psychological and Neural Aspects; Stress, Neural Basis of; Temporal Lobe; Working Memory, Neural Basis of

Bibliography

Baddeley A 2000 Short-term and working memory. In: Tulving E, Craik F I M (eds.) *The Oxford Handbook of Memory*. Oxford University Press, New York, pp. 77–92

Barnden J A, Pollack J B (eds.) 1991 *Advances in Connectionist and Neural Computation Theory: High-Level Connectionist Models*. Ablex, Norwood, NJ, Vol. 1

Cahill L, Babinsky R, Markowitsch H J, McGaugh J L 1995 The amygdala and emotional memory. *Nature* **377**: 295–96

Coussens C M, Teyler T J 1996 Long-term potentiation induces synaptic plasticity at nontetanized adjacent synapses. *Learning and Memory* **3**: 106–14

Eichenbaum H B, Cahill L F, Gluck M A, Hasselmo M E, Keil F C, Martin A J, McGaugh J L, Murre J, Myers C, Petrides M, Roozendaal B, Schacter D L, Simons D J, Smith W C, Williams C L 1999 Learning and memory: Systems analysis. In: Zigmond M J, Bloom F E, Landis S C, Roberts J L, Squire L R (eds.) *Fundamental Neuroscience*. Academic Press, San Diego, CA, pp. 1455–86

Engert F, Bonhoeffer T 1999 Dendritic spine changes associated with hippocampal long-term synaptic plasticity. *Nature* **399**: 66–70

Goldman-Rakic P S, Scalaidhe S E O, Chafee M V 2000 Domain specificity in cognitive systems. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 733–42

Hameroff S R, Kaszniak A W, Scott A C 1998 Toward a science of consciousness II: The second Tucson discussions and debates. The Tao of consciousness. *Journal of Cognitive Neuroscience* **11**: 132–4

Kandel E R 1998 A new intellectual framework for psychiatry. *American Journal of Psychiatry* **155**: 457–69
Lynch G 2000 Memory consolidation and long-term potentiation. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*, 2nd edn. MIT Press, Cambridge, MA, pp. 139–58
Markowitsch H J 1995 Which brain regions are critically involved in the retrieval of old episodic memory? *Brain Research Reviews* **21**: 117–27
Markowitsch H J 2000 Memory and amnesia. In: Mesulam M-M (ed.) *Principles of Cognitive and Behavioral Neurology*. Oxford University Press, New York, pp. 257–93
McClelland J L, McNaughton B L, O'Reilly R C 1995 Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–57
McClelland J L, Rumelhart D E 1986 A distributed model of human learning and memory. In: McClelland J L, Rumelhart D E, and the PDP Research Group (eds.) *Parallel Distributed Processing. Explorations in the Microstructure of Cognition: Psychological and Biological Models*. MIT Press, Cambridge, MA, Vol. 2, pp. 171–215
Mesulam M-M 2000 Behavioral neuroanatomy. Large-scale networks, association cortex, frontal syndromes, the limbic system, and hemispheric specializations. In: Mesulam M-M (ed.) *Principles of Cognitive and Behavioral Neurology*. Oxford University Press, New York, pp. 1–120
Page M 2000 Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences* **23**: 443–512
Romijn H 1997 About the origin of consciousness. A new, multidisciplinary perspective on the relationship between brain and mind. *Proceedings van de Koninklijke Nederlandse Akademie van de Wetenschappen* **100**: 181–267
Sara S J 2000 Retrieval and reconsolidation: Toward a neurobiology of remembering. *Learning and Memory* **7**: 73–84
Schacter D L, Wagner A D, Buckner R L 2000 Memory systems of 1999. In: Tulving E, Craik F I M (eds.) *The Oxford Handbook of Memory*. Oxford University Press, New York, pp. 627–44
Sheldrake R 1988 *The Presence of the Past: Morphic Resonance and the Habits of Nature*, 1st edn. Times Books, New York
Squire L R, Cohen N J, Nadel L 1984 The medial temporal region and memory consolidation: A new hypothesis. In: Weingartner H, Parker E (eds.) *Memory Consolidation*. Erlbaum, Hillsdale, NJ, pp. 185–210
Squire L R, Knowlton B J 1999 The medial temporal lobe, the hippocampus, and the memory system of the brain. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*, 2nd edn. MIT Press, Cambridge, MA, pp. 765–79
Squire L R, Knowlton B, Musen G 1993 The structure and organization of memory. *Annual Reviews of Psychology* **44**: 453–95
Thomas-Ollivier V, Reymann J M, Le Moal S, Schück S, Lieury A 1999 Procedural memory in recent-onset Parkinson's disease. *Dementia and Geriatric Cognitive Disorders* **10**: 172–80
Tulving E, Kapur S, Craik F I M, Moscovitch M, Houle S 1994 Hemispheric encoding/retrieval asymmetry in episodic memory: Positron emission tomography findings. *Proceedings of the National Academy of Sciences of the USA* **91**: 2016–20
Tulving E, Markowitsch H J 1998 Episodic and declarative memory: Role of the hippocampus. *Hippocampus* **8**: 198–204
Zamanillo D, Sprengel R, Hvalby O, Jensen V, Burnashev N, Rozov A, Kaiser K M, Koster H J, Borchardt T, Worley P, Lubke J, Frotscher M, Kelly P H, Sommer B, Andersen P, Seeburg P H, Sakmann B 1999 Importance of AMPA receptors for hippocampal synaptic plasticity but not for spatial learning. *Science* **284**: 1805–11
Zars T, Fischer M, Schulz R, Heisenberg M 2000 Localization of a short-term memory in *Drosophila*. *Science* **288**: 672–5

H. J. Markowitsch

Learning by Occasion Setting

1. *Simple Conditioning and Occasion Setting*

In classical conditioning, organisms learn to predict the presence or the absence of an unconditioned stimulus (US) upon presentation of a conditioned stimulus (CS). Whereas prediction of the presence of the US seems to be mediated by excitatory CS–US associations, prediction of the absence of the US seems to be mediated by inhibitory CS–US associations. When the CS predicts that the US will be present, a conditioned response (CR) is generated. When several CSs are presented together they can combine their predictions and determine whether the CR is generated.

Although CSs are generally considered to only act in the role of 'simple CS' as described in the previous paragraph, experimental data show that a CS can also play as an 'occasion setter.' Whereas as a simple CS a CS forms excitatory or inhibitory associations with a US, as an occasion setter, it modulates the CR generated by another CS.

These alternative roles played by a CS were established in experiments that included different temporal arrangements of feature positive (FP) and feature negative (FN) discriminations. In a FP discrimination, the 'feature' (X) and the 'target' (A) are presented together with the US, and these presentations alternate with presentations of A alone. In a FN discrimination, XA compounds are presented in the absence of the US while A is presented with the US. FP and FN discriminations are simultaneous when X and A are presented at the same time, and serial when the presentation of X precedes the presentation of A.

Using simultaneous and serial FP and FN discriminations, Holland (1983, 1992) offered the following three criteria to distinguish simple conditioning from occasion setting roles of a CS.

1.1 *Response Form*

Under food deprivation, the form of the CR evoked by a light and tone compound reveals its associative origins: if the CR is rear behavior, then responding is the consequence of light-food associations, but if the

CR is head jerk behavior, responding reflects tone-food associations. Taking advantage of this fact, Ross and Holland (1981) reported that in the simultaneous FP, the form of the CR during XA presentations was characteristic of X–US associations, whereas in the serial FP, the rats displayed CRs characteristic of A–US associations. Thus, Ross and Holland's observations suggested that in a simultaneous FP discrimination X acts as a simple CS, but with a serial FP discrimination X acts as an occasion setter. A more complicated pattern of results for simultaneous FP discriminations as a function of the relative intensity of the feature and target stimuli was later reported by Holland (1989c).

1.2 Extinction and Counter-conditioning

If performance on FP and FN discriminations is based on simple associations between X and the US, then elimination of those associations after discrimination training should abolish discriminative performance. Therefore, to the extent that rats solve FP discriminations by forming simple excitatory associations between X and the US, then repeated presentation of X alone after training should extinguish those associations, and consequently eliminate responding to the XA compound. Similarly, to the extent that rats solve FN discriminations by forming simple inhibitory associations between X and the US, then repeated reinforced presentations of X after training should establish excitatory X–US associations, and hence eliminate X's simple inhibitory powers. Accordingly, in simultaneous FP and FN discriminations, Holland (1989b, 1989d) found that post-training extinction of X produced severe decrements in performance on FP discriminations, and post-training counter-conditioning of X abolishes performance on FN discriminations. In contrast, in serial FP and FN discriminations, thought to establish occasion setting, post-training extinction or counter-conditioning of X had minimal effects (Holland 1989b, 1989d, Rescorla 1985, 1986).

1.3 Transfer

The result of combining simple CSs and occasion setters with other CSs has been studied in transfer experiments. For example, if a feature X previously trained in a simultaneous FP discrimination is paired with a trained-then-extinguished CS, B, the simple CS–US associations of the feature will immediately generate responses on XB trials (Holland 1986a, 1986b, Holland and Lamarre 1984). By contrast, if a feature X previously trained in a serial FP discrimination is paired with a trained-then-extinguished CS, B, the occasion-setting properties of the feature might (Davidson and Rescorla 1986, Jarrard and Davidson 1991) or might not generate responses on XB trials (Holland 1986a).

2. A Real-time Model of Pavlovian Conditioning

The alternative roles played by a CS can be best understood when described in terms of a formal model of classical conditioning presented by Schmajuk and DiCarlo (1992) and Schmajuk et al. (1998). The model is a real-time neural network in which CS inputs are connected to different CR outputs both directly and indirectly through a hidden-unit layer that codes configural stimuli. In the network, a CS acts as a simple stimulus through its direct connections with the output units and an occasion setter through its indirect

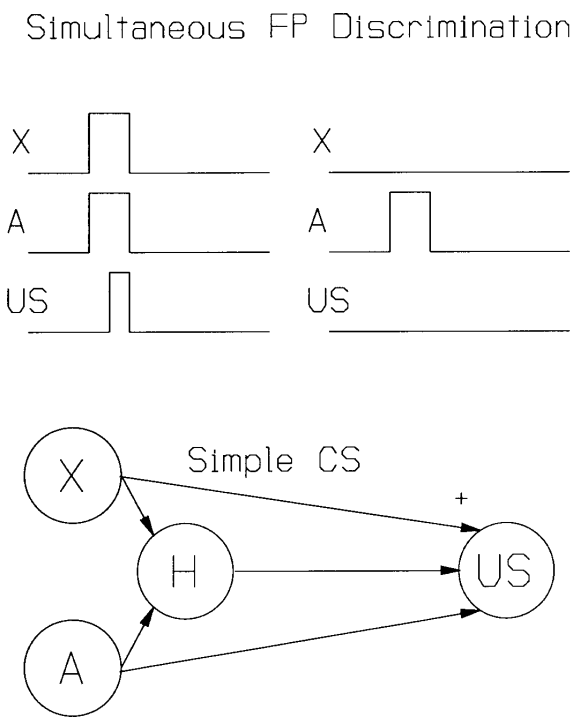

Figure 1
Simultaneous feature-positive discrimination. Top panel: real-time representations of X and A, their traces, and of the US on an XA+ and a A− trial. Bottom panel: simplified diagram of the Schmajuk and DiCarlo (1992) network. When A and X are of similar intensities, X acquires a direct excitatory association with the US. The hidden units do not acquire a strong association with the US. Associative strengths are represented by plus or minus signs placed next to the arrows connecting the nodes. Only significant associative strengths are indicated

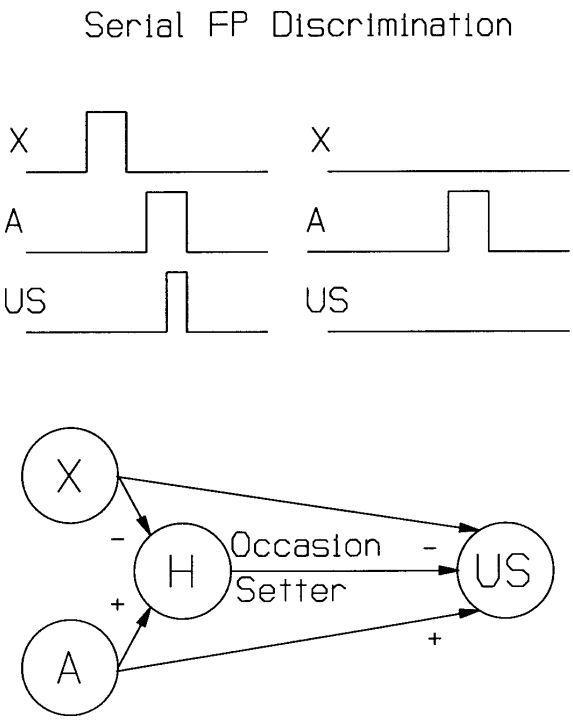

Figure 2
Serial feature-positive discrimination. Top panel: real-time representations of X and A, their traces, and of the US on an X6A+ and a A− trial. Bottom panel: Simplified diagram of the Schmajuk and DiCarlo (1992) network. X acquires an inhibitory association with the hidden units (H). A acquires an excitatory association with the US and an excitatory association with the hidden units (H). The hidden units acquire an inhibitory association with the US. Associative strengths are represented by plus or minus signs placed next to the arrows connecting the nodes. Only significant associative strengths are indicated

configural connections via the hidden units. Schmajuk et al. (1998) showed through computer simulations that the network accounts for a large part of the data on occasion setting. Two examples for FP discriminations are shown in Sects. 2.1 and 2.2, (a) the conditions in which CS behaves as an occasion setter, and (b) how occasion setters are impervious to extinction.

2.1 Conditions in which a CS Behaves as an Occasion Setter

The model explains why a CS behaves (a) as a simple CS in a simultaneous FP discrimination; or (b) as an occasion setter in a serial FP discrimination.

Simultaneous FP discrimination in Fig. 1 (top panels) illustrates the real-time values of X, A, and the US during reinforced and non-reinforced trials in a simultaneous FP discrimination. Figure 1 (bottom panel) summarizes the model's solution for a simultaneous FP discrimination. Whereas X acts as a simple CS because it acquires a strong direct association with the US, A acquires a very weak association. Because direct associations of A and X with the US determine the response form, the model generates mostly CRs characteristic of X–US associations during compound XA presentations.

Serial FP discrimination in Fig. 2 (top panels) illustrates the real-time values of stimuli during reinforced trials in a serial FP discrimination. Figure 2 (bottom panel) shows a simplified depiction of the associative structure established in the network after serial discrimination trials. Because of their different temporal relations with the US, A accrues strong A–US associations but X only accrues weak X–US associations. Since the direct associations of X and A with the US determine the form of responding, responding on XA presentations comprises mostly CRs characteristic of A–US associations. Furthermore, responding on A-alone trials is small because it is strongly inhibited by the hidden units (which are activated by A), and responding on XA presentations is large because the inhibitory influence of the hidden units is ameliorated by the inhibitory action of X on the hidden units.

2.2 Occasion Setters are Impervious to Extinction

Since simultaneous FP discriminations are solved by the model almost completely by direct X–US associations (see Fig. 1), the effect of X extinction is dramatic, abolishing CR responding on both X and XA test trials. By contrast, compound responding for serial FP is controlled by A–US associations which are attenuated by the inhibitory H–US associations (see Fig. 2). Since X–H and H–US associations are only minimally affected by X extinction, very little decrement is observed in XA compound responding after extinction.

3. Conclusion

A CS can behave either as a simple CS or as an occasion setter. In the first case, the CS acquires excitatory or inhibitory associations with the US, in the second case the CS might enhance or depress the effect of the simple association accrued by another CS. The properties of simple CSs and occasion setters have been thoroughly studied (for a review see Swartzentruber 1995).

A neural network model presented by Schmajuk et al. (1998) can accurately describe most of the available data. Other learning models that also include direct

and indirect associations between CS inputs and CR outputs (e.g., Bouton 1994) also provide correct depictions of the phenomena. Neural networks without direct connections (e.g., Gluck and Myers 1993, Wagner 1992) can only describe some of the experimental results. Some of the physiological mechanisms of occasion setting have been addressed in terms of neural networks (Gluck and Myers 1994, Schmajuk and Buhusi 1997). A recent volume, published by Schmajuk and Holland (1998) presents the latest data and theoretical approaches to the subject.

See also: Artificial Neural Networks: Neurocomputation; Autonomic Classical and Operant Conditioning; Avoidance Learning and Escape Learning; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Connectionist Models of Concept Learning; Eyelid Classical Conditioning; Fear Conditioning; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of Models of Neural Basis of Learning and Memory.

Bibliography

Bouton M E 1994 Conditioning, remembering, and forgetting. *Journal of Experimental Psychology: Animal Behavior Processes* **20**: 219–31

Davidson T L, Rescorla R A 1986 Transfer of facilitation in the rat. *Animal Learning and Behavior* **14**: 380–6

Gluck M A, Myers C E 1993 Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus* **3**: 491–516

Gluck M A, Myers C E 1994 Context, conditioning, and hippocampal representation in animal learning. *Behavioral Neuroscience* **108**: 835–47

Holland P C 1983 Occasion-setting in Pavlovian feature positive discriminations. In: Commons M L, Herrnstein R J, Wagner A R (eds.) *Quantitative Analyses of Behavior: Discrimination Processes*. Ballinger, New York, Vol. 4, pp. 183–206

Holland P C 1986a Temporal determinants of occasion setting in feature positive discriminations. *Animal Learning and Behavior* **14**: 111–20

Holland P C 1986b Transfer after serial feature positive discrimination training. *Learning and Motivation* **17**: 243–68

Holland P C 1989a Acquisition and transfer of conditional discrimination performance. *Journal of Experimental Psychology: Animal Behavior Processes* **15**: 154–65

Holland P C 1989b Feature extinction enhances transfer of occasion setting. *Animal Learning and Behavior* **17**: 269–79

Holland P C 1989c Occasion setting with simultaneous compounds in rats. *Journal of Experimental Psychology: Animal Behavior Processes* **15**: 183–93

Holland P C 1989d Transfer of negative occasion setting and conditioned inhibition across conditioned and unconditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes* **15**: 311–28

Holland P C 1992 Occasion setting in Pavlovian conditioning. In: Medin D L (ed.) *The Psychology of Learning and Motivation*. Academic Press, New York, Vol. 28, pp. 69–125

Holland P C, Lamarre J 1984 Transfer of inhibition after serial and simultaneous feature negative discrimination training. *Learning and Motivation* **15**: 219–43

Jarrard L E, Davidson T L 1991 On the hippocampus and learned conditional responding: Effects of aspiration versus ibotenate lesions. *Hippocampus* **1**: 103–13

Rescorla R A 1985 Conditioned inhibition and facilitation. In: Miller R R, Spear N E (eds.) *Information Processing in Animals: Conditioned Inhibition*. Erlbaum, Hillsdale, NJ, pp. 299–326

Rescorla R A 1986 Facilitation and excitation. *Journal of Experimental Psychology: Animal Behavior Processes* **12**: 325–32

Ross R T, Holland P C 1981 Conditioning of simultaneous and serial feature-positive discriminations. *Animal Learning and Behavior* **9**: 293–303

Schmajuk N A, Buhusi C 1997 Stimulus configuration, occasion setting, and the hippocampus. *Behavioral Neuroscience* **111**: 235–58

Schmajuk N A, DiCarlo J J 1992 Stimulus configuration, classical conditioning, and the hippocampus. *Psychological Review* **99**: 268–305

Schmajuk N A, Holland P C 1998 Occasion setting. In: *Associative Learning and Cognition in Animals*. American Psychological Association, Washington, DC

Schmajuk N A, Lamoureux J, Holland P C 1998 Occasion setting and stimulus configuration: A neural network approach. *Psychological Review* **105**: 3–32

Swartzentruber D 1995 Modulatory mechanisms in Pavlovian conditioning. *Animal Learning and Behavior* **23**: 123–43

Wagner A R 1992 Some complexities anticipated by AESOP and other dual-representation theories. Paper abstracted in Kimmel H (Chair), Symposium on Pavlovian Conditioning with Complex Stimuli, XXV International Congress of Psychology. *International Journal of Psychology*: 101–2

N. A. Schmajuk

Learning Curve, The

Most tasks get faster with practice. This is not surprising because we have all seen this and perhaps know it in some intuitive sense. What is surprising is that the rate and shape of improvement is fairly common across tasks. Figure 1 shows this for a simple task plotted both on linear and log–log coordinates. The pattern is a rapid improvement followed by ever lesser improvements with further practice. Such negatively accelerated learning curves are typically described well by power functions. Thus, learning is often said to follow the 'power law of practice.' Not shown on the graph, but occurring concurrently, is a decrease in variance in performance as the behavior reaches an apparent plateau on a linear plot. This plateau masks continuous small improvements with extensive practice that may only be visible on a log–log plot where months or years of practice can be seen. The longest measurements suggest that for some tasks improvement continues for over 100,000 trials.

There are some related regularities. There is evidence to suggest that standard deviation and skew in

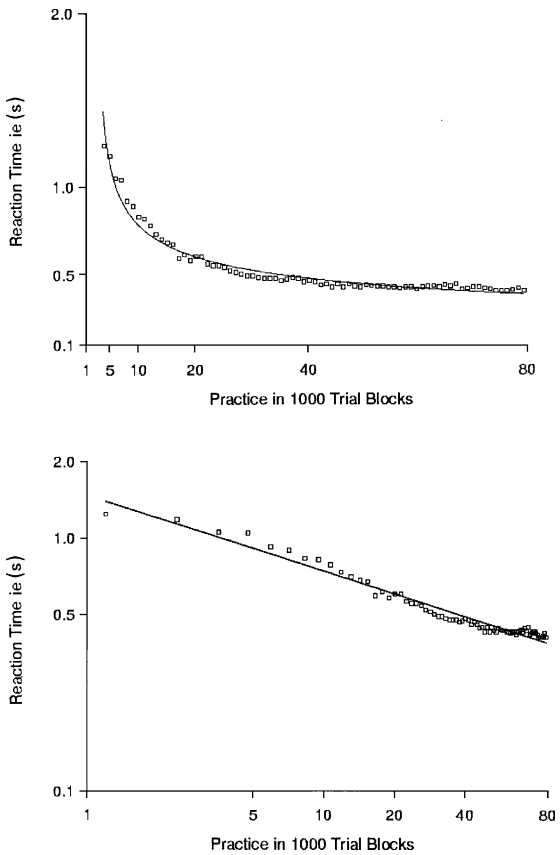

Figure 1
Time to perform a simple task on a linear and log–log plot as well as a power law fit to the data (adapted from Seibel 1963)

performance time also decrease according to a power law, but with worse correlations. Indeed, in some cases the decrease in standard deviations appears to cause the improvement, because the minimum time to perform a task does not change (Rabbitt and Banerji 1989).

The power law of practice is ubiquitous. From short perceptual tasks to team-based longer term tasks of building ships, the breadth and length of human behavior, the rate that people improve with practice appears to follow a similar pattern. It has been seen in pressing buttons, reading inverted text, rolling cigars, generating geometry proofs, and manufacturing machine tools (cited in Newell and Rosenbloom 1981), performing mental arithmetic on both large and small tasks (Delaney et al. 1998), performing a scheduling task (Nerb et al. 1999), and writing books (Ohlsson 1992). Further examples are noted in reviews (e.g., Heathcote et al. 2000). In manufacturing this curve is called a progress function. It can be demonstrated by taking a task—nearly any task—and timing how long it takes to complete over 10 trials, or better over a hundred trials. An example might be to try reading this article upside down. The time per paragraph will generally decrease, but with some differences caused by the different words and paragraph lengths.

In general, the more averaging, the smoother the curve. The learning curve appears smoother when the data is averaged across subjects, across tasks, or both. When the tasks are known to vary in difficulty, such as different complex mental arithmetic problems (e.g., 27×5 and 23×28), the learning curve only appears when averaging is performed because the different problems naturally take different times. Even when problems are of comparable difficulty, subjects may use different strategies. For example, arithmetic problems can be solved by two at least strategies, retrieval and calculation. The power law applies across strategies, but the fit is better to each strategy (Delaney et al. 1998), or even an individual's strategies (Heathcote et al. 2000).

Averaging can mask important aspects of learning. If the tasks vary in difficulty, the resulting line will not appear as a smooth curve, but will bounce around. Careful analysis can show that different amounts of transfer and learning are occurring on each task. For example, solving the problem 22×43 will be helped more by previously solving 22×44 than by solving 17×38 because there are more multiplications shared between them. Where subtasks are related but different, such as sending and receiving Morse code, the curves can be related but visibly different (Bryan and Harter 1897).

The learning curve has implications for learning in education and everyday life. It suggests that practice always helps improve performance, but that the most dramatic improvements happen first. Another implication is that with sufficient practice people can achieve comparable levels of performance. For example, extensive practice on mental arithmetic (Staszewski reported in Delaney et al. 1998) and on digit memorization have turned average individuals into world class performers.

To sum up, the learning curve is visible with enough aggregation of dissimilar tasks or across similar tasks down to the level of individual subject's strategies.

1. *Mathematical Definitions*

The shape of the curve is negatively accelerated—further practice improves performance, but with diminishing returns. Power laws and exponentials are both functions that provide this shape. Mathematical definitions are given in Table 1. The exact quality of the fit depends on innumerable details of averaging, the precise function used, and the scale. For example, the full power law formula is the most precise, but it has additional terms that are difficult to compute, and

Table 1
Functions that fit practice data

Time = MinTime + $B \cdot (N+E)^{-\beta}$ [Full power law formula]
Time = $B \cdot N^{-\beta}$ [Simple power law formula]
Time = $B \cdot e^{-\alpha N}$ [Simple exponential formula]
B is the range of learning N is the trial number E is the number of previous practice trials α,β are the learning rate parameters

the asymptote is usually only visible when there are over 1,000 practice trials (Newell and Rosenbloom 1981). The typical power law formula is simpler, but leaves out previous practice. When using this formula, the coefficients of the power law for a set of data can be easily computed by taking the log of the trial number and log of task time and computing a linear regression. That is, fitting a regression in log–log space.

In general, the power function fit appears to be robust, regardless of the methods used (Newell and Rosenbloom 1981). However, recent work (Heathcote et al. 2000) suggests that the power law might be an artifact arising from averaging (Anderson and Tweney 1997), and that the exponential function may be the best fit when individual subjects employing a single strategy are considered. Distinguishing between the power and exponential functions is not just an esoteric exercise in equation fitting. If learning follows an exponential, then learning is based on a fixed percentage of what remains to be learnt. If learning follows a power law, then learning slows down in that it is based on an ever decreasing percentage of what remains to be learnt.

Regardless of the functional form of the practice curve, there remain some systematic deviations that cause problems, at the beginning and end of long series. The beginning deviations may represent an encoding process. For example, it may be necessary to transform a declarative description of a task into procedures before actual practice at the task can begin (Anderson and Lebiere 1998); the residuals at the end may represent approaching the minimum time for a given task as defined by an external apparatus. These effects appear in Fig. 1 as well.

2. *Process-based Explanations of the Learning Curve*

The power law of learning is an important regularity of human behavior that all theories of learning must address. For example, Logan (1988) suggests that the retrieval of memory traces guides responding in speeded tasks. Response times represent the fastest retrieval time among the memory traces racing to

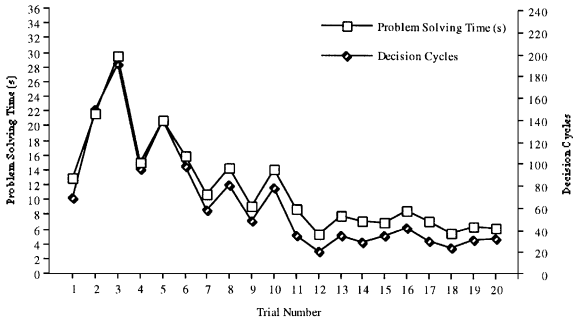

Figure 2
The time to perform a trouble-shooting task by both a cognitive model and a subject (after Ritter and Bibby 2001)

guide the response. The power law falls out of this analysis, because with practice the number of redundant memory traces increases, which, in turn, increases the chances of observing fast retrieval times. Olhsson's (1996) theory notes how the learning curve could arise out of error correction. Different assumptions (e.g., how fast errors are caught) give rise to different curves including the exponential and power law. The learning curve has also been demonstrated for connectionist models as well.

ACT-R (Anderson and Lebiere 1998) and SOAR (Newell 1990), two cognitive architectures, generally predict a power law speedup, but for different reasons. ACT-R does this because rules and memory traces are strengthened according to a power law based on the assumption that the cognitive system is adapted to the statistical structure of the environment (Anderson and Schooler 1991). Several models in SOAR (see *Cognitive Theory: SOAR*) have been created that model the power law (e.g., Nerb et al. 1999, Newell 1990). These models explain the power law as arising out of mechanisms such as hierarchical learning (i.e., learning parts of the environment or internal goal structure) that initially learns low-level actions that are very common and thus useful, and with further practice more useful larger patterns that occur infrequently are learned. The SOAR models typically vary from human learning in that they learn faster than humans, and they do not learn for as long a period. SOAR and ACT-R also predict variance in the improvement on all tasks due to different amounts of transfer across problems and learning episodes. Figure 2 shows how such a model can predict differential transfer as well as continuous learning as it appears in human data.

3. *Summary*

The learning curve is a success story for cognitive psychology, which has shown that learning is ubiquitous and has provided mathematical accounts of the

rate. The learning curve is also a success story for cognitive modeling, which has explained many aspects of the curve and the noise inherent in it partly as differences in transfer of knowledge between tasks and how the curve arises out of mechanisms necessary for processing. The multiple explanations also suggest that there may be multiple ways that the curve can arise.

See also: Categorization and Similarity Models; Cognitive Theory: ACT; Cognitive Theory: SOAR; Mathematical Learning Theory; Mathematical Learning Theory, History of; Memory Models: Quantitative; Transfer of Learning, Cognitive Psychology of

Bibliography

Anderson J R, Lebiere C 1998 *The Atomic Components of Thought*. Erlbaum, Mahwah, NJ

Anderson J R, Schooler L J 1991 Reflections of the environment in memory. *Psychological Science* **2**: 396–408

Anderson R B, Tweney R D 1997 Artifactual power curves in forgetting. *Memory and Cognition* **25**: 724–30

Bryan W L, Harter N 1897 Studies in the physiology and psychology of the telegraphic language. *Psychological Review* **4**: 27–53

Delaney P F, Reder L M, Staszewski J J, Ritter F E 1998 The strategy specific nature of improvement: The power law applies by strategy within task. *Psychological Science* **9**(1): 1–7

Heathcote A, Brown S, Mewhort D J K 2000 The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review* **7**: 185–207

Jones G, Ritter F E, Wood D J 2000 Using a cognitive architecture to examine what develops. *Psychological Science* **11**(2): 1–8

Logan G D 1988 Toward an instance theory of automatization. *Psychological Review* **95**: 492–527

Nerb J, Ritter F E, Krems J 1999 Knowledge level learning and the power law: A Soar model of skill acquisition in scheduling. In: Wallach D, Simon H A (eds.) *Kognitionswissenschaft* [*Journal of the German Cognitive Science Society*] Special issue on cognitive modeling and cognitive architectures, pp. 20–9

Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

Newell A, Rosenbloom P S 1981 Mechanisms of skill acquisition and the law of practice. In: Anderson J R (ed.) *Cognitive Skills and their Acquisition*. Erlbaum, Hillsdale, NJ, pp. 1–51

Ohlsson S 1992 The learning curve for writing books: Evidence from Asimov. *Psychological Science* **3**(6): 380–2

Ohlsson S 1996 Learning from performance errors. *Psychological Review* **103**(2): 241–62

Rabbitt P, Banerji N 1989 How does very prolonged practice improve decision speed? *Journal of Experimental Psychology: General* **118**: 338–45

Ritter F E, Bibby P A 2001 Modeling how and when learning happens in a simple fault-finding task. *Proceedings of the 4th International Conference on Cognitive Modeling*. Erlbaum, Mahwah, NJ

Rosenbloom P S, Newell A 1987 Learning by chunking, a production system model of practice. In: Klahr D, Langley P, Neches R (eds.) *Production System Models of Learning and Development*. MIT Press, Cambridge, MA, pp. 221–86

Seibel R 1963 Discrimination reaction time for a 1,023-alternative task. *Journal of Experimental Psychology* **66**(3): 215–26

F. E. Ritter and L. J. Schooler

Learning from Text

In many everyday situations we need to acquire information from texts, both oral and written, such as newspaper articles, manuals, talks, or telephone conversations. Learning from text is a ubiquitous occurrence. Of course, the exceptional interest that the topic has received is driven by the need to improve learning from text in formal educational settings. Research on this issue has been motivated by different, but related, goals. Educational research is mostly concerned with the development and evaluation of teaching methods (Hynd et al. 1998), linguistic theory is concerned with the description and analysis of the texts used as learning materials, and the psychological literature focuses more on the cognitive processes subserving comprehension and learning (e.g., Kintsch 1998). In the late twentieth century, the advancement of text comprehension theories has made it possible to apply concepts from psycholinguistic research to educational issues, so that there is now more and more overlap between these areas.

1. Definitions

How do we define learning from text? Learning from text is often equated with learning *of* text, that is, with memory for the text information proper (e.g., the memorization of a poem). In contrast, Kintsch makes a clear distinction and defines learning *from* text as 'the ability to use the information acquired from the text productively in novel environments' (Kintsch 1998, p. 290).

Using concepts from psycholinguistic theories of text comprehension, these definitions can be further sharpened: van Dijk and Kintsch (1983) distinguish three levels of representation. First, the surface level encodes the verbatim wording of the text information. On this level, the syntactic and lexical realization of the text is represented. The second level is a semantic representation, called the textbase, closely corresponding to the propositional, or meaning-based content, but independent of the exact wording. In addition, the textbase contains some information not explicitly mentioned in the text, such as associative elaborations, certain inferences, and macropropositions. Finally, and most importantly, the situation model is an integration of the reader's prior knowledge and experiences with the information currently encountered. This representation of 'what the text is all about'

does not need to be in a verbal or propositional format, and it depends both on the text information, and on the individual background the comprehender contributes.

These three levels of representations are differentiated not only by their content, but also by their retention characteristics: The surface level is forgotten quickly, the textbase might be available over several hours, but the situation model is what survives in the long term. Thus, learning *of* text is located at the surface and textbase levels, whereas learning *from* text needs to be conceptualized as the construction, retention and utilization of a situation model representation (Ferstl and Kintsch 1999).

2. Empirical Methodology

For dissociating memory for text and learning from text, we need measures that tap these levels of representation differentially. Traditional measures for the textbase are recall and recognition, as well as the factual comprehension questions, that is, questions about information explicitly mentioned at a particular point in the text. Moreover, those inference questions that probe information on presumably automatic inferences (e.g., bridging inferences) are also located on the textbase level.

In contrast, for assessing the situation model, tasks requiring the 'productive use' of the acquired information in 'novel environments' are needed. When a manual is read, for instance, we can directly test the nonverbal ability to operate the device explained. For testing learning from a physics textbook we can ask students to perform problem-solving tasks requiring the transfer or the application of text information. However, even when the text information cannot be tested nonverbally, we can devise 'novel situations' going beyond the explicit text information. From recall protocols, we can score those statements that were not mentioned in the text, but were inferred or elaborated by the learner. An evaluation of their plausibility within the ideal situation model provides a score of the creative, reconstructive use of the information. In recognition tasks we can include words or statements that are consistent with the situation described, although they were not directly included in the text. In designing questions, the situation model is probed when a combination of prior knowledge with text information or the integration of statements from different parts of the text is needed.

Another approach attempts empirically to describe learning from text as the change of the reader's prior knowledge structure (Ferstl and Kintsch 1999). After identifying key concepts of the text's domain, each learner's internal organization of these concepts is estimated based on relatedness ratings, results from sorting tasks, or cued associations. The resulting associative network is considered an approximation of the learner's knowledge structure. After reading the text, the same task is repeated, and the resulting network is considered an approximation of the learner's situation model. The changes from before to after reading are taken as an index of learning from text. Moreover, the knowledge networks after reading can be compared with an ideal structure, obtained, for instance, from an analysis of the text content, or from an expert's assessment of the same domain.

Many inconsistencies in the literature on learning from text might be explained by evaluating in more detail the methods used for assessing learning. However, it is important to note that, although there are empirical dissociations, in most cases building a good textbase representation goes hand in hand with establishing an elaborate situation model. Therefore, in describing learning from text, the vast knowledge about text memory accumulated to date should not be neglected.

3. Text Characteristics

An obvious answer to the question of how we can improve learning from text is to just write better texts. Consequently, there have been numerous suggestions for writers on how to increase the readability of their texts (e.g., Taylor and Taylor 1983), most of which seem to consist of simplifying the style. On the word level, short and familiar words are recommended, and on the sentence level, syntactically complex constructions are to be avoided.

On the text level, the notion of coherence has been shown to be crucial for readability, that is, for comprehensibility. Local coherence refers to the connection between subsequent sentences, while global coherence refers to the connection between parts of the text, such as paragraphs. For instance, when moving back from a concrete example to the general topic of the text, global coherence is preserved, while local coherence might be violated. Coherence facilitates comprehension because no additional inferences are required for deriving the text structure. Situation model construction can be facilitated by adding explicit cohesive ties, that is, lexical information signaling the type of connection between successive sentences.

A second important factor is the macrostructure of a text, corresponding roughly to the outline or a summary of the text. The macrostructure has to be constructed by using macrorules, such as marking some propositions as particularly important or subsuming others under a more general one. Writers can facilitate the application of macrorules, for instance, by using a canonical, expected structure. The macrostructure of an existing text can be clarified by adding explicit information in the form of subheaders, introductory or structuring statements (e.g., 'the following three examples … '), or underlining of important information. Using these macrostructure cues

helps to set up the textbase, and in turn frees resources for processing on a situation model level (for an early approach to text structure, see Meyer 1975).

4. Adjunct Information

Another way of improving learning from text is to provide additional materials that extend the text information. Very influential suggestions, in particular in educational research, are advance organizers and illustrations.

Outlines or summaries given before reading explicitly provide the macrostructure of the text to be read. Similar to headers or explicit organizing statements within the text, this additional information facilitates comprehension. An advance organizer, as proposed by Ausubel (1968), goes beyond a simple outline or summary. Rather, it uses a different level of abstraction and has the function to 'provide ideational scaffolding for the stable incorporation and retention of the more detailed and differentiated material that follows in the learning passage' (Ausubel 1968, p. 148). Although this concept seems compelling, there was considerable debate about the efficacy of advance organizers. While some studies showed a beneficial effect, others failed to replicate. One reason for these inconsistencies was that early studies did not have available the tools for analyzing the content and structure of both text and advance organizer, and that the levels of representation were not carefully distinguished. As an example of a more recent study controlling these factors, Mannes and Kintsch (1987) studied the interaction between organizer structure and representational level. When the organizer and the text were of a similar structure, readers formed a better textbase representation, because the advance organizer provided the macrostructure for the text read later. Different structures, on the other hand, improved the verification performance on situation model statements. In this case, the macrostructure of the target text had to be derived and the availability of the information from the advance organizer made a more elaborate situation model likely.

One of the most intriguing demonstrations of the effect of illustrations on text comprehension is the study by Bransford and Johnson (1972). They demonstrated that looking at a picture showing an unusual serenade substantially improved comprehension and recall of a vaguely worded text describing the scene on the picture. Since that early study, numerous experiments attempted to delineate the possibilities and limits of illustrations for improving learning processes (Mandl and Levin 1989). The kind of picture used depends on the text information to be illustrated. For instance, a narrative can be embellished by a portrait of the protagonist, a manual might contain a schematic depiction of the device, and the description of a hiking trail is facilitated by a map. Illustrations aid learning from text when they are available before or during processing, and when the information provided by the picture is relevant to and redundant with text information. Purely decorative illustrations do not have a beneficial effect. One theory for explaining the facilitative effect of illustrations was proposed by Paivio (1986), who postulated that dual coding of the information, in a verbal and in a visual representation, increases the likelihood of the information becoming available through one of the codes. However, illustrations tend to have a smaller effect on recall of text, but often facilitate situation-based inferences and elaborations. Rather than merely doubling the amount of information, as the dual coding theory suggests, the illustrations provide background knowledge, and they encourage the mapping of the text information to the visually presented information. Thus, illustrations and other nonverbal sources, such as the demonstration of a procedure, are utilized for constructing the situation model.

5. The Learning Process

As we have seen, the external information not only enhances the reader's background knowledge, but also has a direct impact on the type of processing. How a text is being read determines the kind of representation being built up. More careful, effortful reading, with more time spent on the task, increases performance. In addition, there is no doubt that instructions can be used to induce effectively certain reading goals. For instance, when told to memorize a text for later recall, readers focus more on the construction of a good textbase representation than when told to read for comprehension. When instructed to attend to a certain type of information, encoding takes place selectively and irrelevant information gets neglected. When provided with an illustration, the reader processes the text information overlapping with the picture more thoroughly.

A more direct way to influence processing is to ask the right questions (Graesser et al. 1994). Questions are asked by the learner to clarify points that are not readily comprehensible. Questions are asked by the teacher to test comprehension, or lead students' processing into a specific direction. When the effect of this type of question is to be evaluated, the questions are interspersed in the text or added afterwards. These so-called adjunct questions have two functions. First, they can induce comprehension monitoring, an activity that readers often neglect. An important strategy during learning from text is to amend comprehension failures, for example, by rereading passages, or by activating additional background knowledge. Difficulties with answering an adjunct question, in particular, if feedback is provided, can signal problems that might otherwise go unnoticed.

The second function of adjunct questions is to support directly situation model construction. Ques-

tions can be formulated so that they point to particularly relevant factual information (e.g., definitions), or so that they promote inferences needed for deep understanding. Most studies document a benefit of adjunct questions mainly for the information probed, but little transfer to other information in the text. When the questions are not provided, but are to be generated, learners often face considerable difficulties. Without extensive training, effective questioning is more likely to be the result of an already established situation model, rather than an aid to constructing it.

6. *Individual Differences*

In his account of meaningful learning—which encompasses learning from text—Ausubel (1968) stresses the role of the learner. 'Meaningful learning presupposes that the learner manifest a meaningful learning set, that is, a disposition to relate the new material nonarbitrarily and substantively to his cognitive structure' (Ausubel 1968, p. 37). In this section we look at research on individual differences, and consider some specific factors influencing whether a 'meaningful learning set' is available.

Learning studies encompass groups as varied as children just barely able to read, high-school students, and adults. Hence it is important to take into account the learner's age. One important finding is that younger children tend to rely more on verbatim encoding of text, and that strategies needed for situation model construction, such as structure building, analogical reasoning, or active integration with previous knowledge, become available only at a later stage during development.

Studies grouping adult learners according to their reading skills find that it is exactly those strategies that less skilled readers tend to use for compensating their weaknesses on the textbase level. For instance, poor readers rely more on contextual constraints and causal structures. Poor readers have lower recall scores, and benefit from good writing, but these textbase differences are less noticeable on the situation model level—at least as long as the situation model of the particular text is consistent with expectations based on general world knowledge.

Another learner characteristic concerns motivational factors, such as the interest the learner brings to the task at hand. Within a text, interesting information is more likely to be recalled, but embellishing a text with interesting tidbits does not necessarily improve comprehension for the remainder of the text. Thus, few 'seductive' details do not increase the overall motivation to become more involved in text processing. Because it is unavoidable that some content areas are not inherently interesting for each student, strategies for increasing the learners' motivation are important (Wade 1992).

One of the most crucial individual differences is the learner's background knowledge. In the serenade example, for instance, the illustration provided the necessary background knowledge for constructing a situation model, and in this case, even for constructing a coherent text representation. Some advance organizers also have the function of providing necessary background knowledge. Often, however, it does not suffice to look at a picture or read a second text. Instead, the reader needs more extensive domain knowledge. Experts have better vocabulary, they know more facts, but most importantly, these facts are closely interrelated and well structured. When readers are experts in the text's domain, they recall more text information, they can distinguish important information from details, they can bridge coherence gaps, and they are more likely to derive goal structures.

The disadvantage caused by not being able to relate the current information to already existing cognitive structures (Ausubel 1968) or to 'hang it on the right hooks' in general world knowledge (Kintsch 1998) can be overcome by the better writing. As we have seen before, explicit coherence cues facilitate understanding. And indeed, when learners with high or low knowledge read an explicitly coherent text, their problem-solving performance and card-sorting results are comparable. Only for less coherent text do these situation model measures show a clear benefit for the high knowledge participants (McNamara et al. 1996). A particularly interesting result in the same study was an interaction between knowledge and coherence. High-knowledge readers developed better situation models when they read the less coherent text than when they read the better, more explicit text. An explanation for this somewhat counterintuitive result is that the coherence gaps elicited more active processing. Each coherence gap asks an adjunct question, 'how does this fit together?' The high-knowledge readers generated the necessary inferences on their own, but they needed some processing effort for doing so, which in turn helped them in developing deep understanding (for replications see Kintsch (1998)). Thus, just by providing the right text for the right learner, the learning success was increased for both high- and low-knowledge groups.

7. *Teaching Strategies*

An important driving factor of this area of research is the possibility of developing teaching programs incorporating the educational and psycholinguistic findings. The facilitating factors mentioned before can be provided by the teacher or within the text. Alternatively, the use of any comprehension strategy can be taught to students. For instance, underlining helps the reader distinguish important information from details, summarization puts a focus on the macrostructure, providing background knowledge is helpful, and prior

activation of a relevant domain can foster integration of text information with available background knowledge. Some programs incorporate several of these techniques. For instance, the well-known PQRST technique recommends an elaborate sequence of processing: preview, question, review, state, and test (Robinson 1970). Of course, such a technique can only succeed when the learners are able to take advantage of it. As we have seen before, though, it is a skill to ask the appropriate questions for facilitating deep processing. On the other hand, explicitly teaching learning techniques provides a benefit to the learners only if they do not yet employ self-initiated processing strategies. Therefore, empirical confirmations of the efficacy of such programs, that almost everybody would intuitively agree upon, are difficult to obtain. Not just the right text for a given learner, but also just the right teaching technique for that learner is what is needed.

8. *Conclusions and Outlook*

While it may seem that empirical studies too often produce the answer 'it depends' to the question of how we can improve learning from text, this should not discourage further experimental research. We now have the theoretical background for analyzing text characteristics and for assessing learner-centered factors. In addition, there are some promising attempts to automate these (Kintsch 1998). Of course, sometimes, the intuitions of good educators are worth more than a multitude of controlled studies. Nevertheless, we should strive to objectify and delineate these intuitions, in order to support the learning process of many student generations to come.

The way printed media have shaped our intellectual world for centuries accounts for the exceptional interest in the issue of learning from text. This influence of books is gradually changing with the advent of new electronic media. Traditional textbooks or newspapers may soon be a relict of the past, but research on learning from text will become even more important. The large majority of information presented will remain verbal. Existing research on the interactions between text information and other, mostly visual, presentation forms can help to understand and improve modern communication. The usual way of writing texts as coherent wholes that are best read in their sequential order has also changed with the advent of technologies such as the Internet and other hypertext media. Using these tools provides the freedom to select, review, or enhance the learning material on a step-by-step basis. Thus, the learners' active role in shaping their individual learning environment will become more and more important.

See also: Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Literary Texts: Comprehension and Memory; Memory for Text; Text Comprehension: Models in Psychology; Textbooks

Bibliography

Ausubel D P 1968 *Educational Psychology: A Cognitive View*. Holt, New York

Bransford J D, Johnson M K 1972 Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior* **11**: 717–26

Ferstl E C, Kintsch W 1999 Learning from text: Structural knowledge assessment in the study of discourse comprehension. In: van Oostendorp H, Goldman S R (eds.) *The Construction of Mental Representations During Reading*. Erlbaum, Mahwah, NJ, pp. 247–77

Graesser A, McMahen C L, Johnson B K 1994 Question asking and answering. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA, pp. 517–38

Hynd C R, Stahl S A, Carr M, Glynn S M (eds.) 1998 *Learning from Text Across Conceptual Domains*. Erlbaum, Mahwah, NJ

Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK

Mandl H, Levin J R (eds.) 1989 *Knowledge Acquisition from Text and Pictures*. Elsevier, Amsterdam

Mannes S M, Kintsch W 1987 Knowledge organization and text organization. *Cognition and Instruction* **4**: 91–115

McNamara D S, Kintsch E, Songer N B, Kintsch W 1996 Are good texts always better? Interactions of text coherence, background knowledge and levels of understanding in learning from text. *Cognition and Instruction* **14**: 1–43

Meyer B J F 1975 *The Organization of Prose and its Effects on Memory*. Elsevier, Amsterdam

Paivio A 1986 *Mental Representations: A Dual Coding Approach*. Oxford University Press, New York

Robinson F B 1970 *Effective Study*, 4th edn. Harper and Row, New York

Taylor I, Taylor M M 1983 *The Psychology of Reading*. Academic Press, New York

van Dijk T A, Kintsch W 1983 *Strategies of Discourse Comprehension*. Academic Press, New York

Wade S E 1992 How interest affects learning from text. In: Renninger K A, Hidi S (eds.) *The Role of Interest in Learning and Development*. Erlbaum, Hillsdale, NJ, pp. 255–77

E. C. Ferstl

Learning: Organizational

Organizational learning is a metaphor that matches two concepts—learning and organization—and enables exploration of the organization as if it were a subject which learns, deals with knowledge, reflects on

experiences, and is endowed with a stock of knowledge, skills, and expertise. A metaphor works by matching what is distant with what is close, similar features with dissimilar ones, and it is a cognitive tool that develops creativity and social imagination and reveals the importance of language and symbols in the construction of reality and in the formulation of theory. In this way, cognitive transpositions come into operation that make it possible to imagine and to talk about the object in question (the organization) as if it were another already-known object (learning).

Therefore, study of organizations as if they were systems of knowing and learning may be of assistance in the design of knowledge intensive organizations and of networks of organizations that create and transfer knowledge. At a theoretical level, the representation of the learning organization may stand as a valid theoretical alternative to the rational organization because it replaces the classical concept of the organization that acts according to the principles of rational choice with that of the organization which acts according to principles of experimentation, of trial and error, of success and failure, of discovery and invention.

The following sections outline the history of the concept of organizational learning, the divergences between two bodies of literature—organizational learning and learning organization—as they grew in the 1990s, and the further development in the field.

1. *A Brief History of the Concept of Organizational Learning*

The first mentions of the concept can be traced back to March and Simon (1958), but for long it was peripheral to organizational theory. March has always alluded to learning and to the link between learning and decision-making processes, although only few of his works are specifically devoted to this topic.

The work generally recognized as pioneer in the field is by Argyris and Schön (1974) and a brief historical excursus into the literature on organizational learning will illustrate how researchers interested in knowledge have examined specific organizational phenomena in terms of learning defined in variety of ways:

(a) *As learning curves* (Yelle 1979), a concept which probably stands at the basis of the most long-standing interest in learning. Learning curves rest on the principle that the man-hours required to produce a unit of output decrease in proportion to the experience accumulated. Productive processes display characteristic learning curves in relation to a set of organizational factors. Hence derives the idea that there is something more than mere individual learning and that experience constitutes a transferable body of knowledge that may become institutionalized and incorporated into situations different from the original ones.

(b) *As behavioral change through trial and error*. Learning is a self-regulating process which detects errors and corrects them or, if preferred, which checks actions against their outcomes and makes appropriate adjustments to subsequent actions. This conception may be applied to the organization as an information processor or as a producer of knowledge on the nexuses between input and output and their reception by the environment.

(c) *As adaptation to the environment*, and therefore as the capacity to 'fit in,' to adjust to external stimuli which must be discovered opportunely and rapidly. Learning is, therefore, synonymous with structural change and acts as a signal of successful evolution. Moreover, learning as systemic wisdom does not come about solely within organizations or within competitive relationships among organizations. It also appears in the networks of informal learning (Trist 1983) established among numerous organizations by means of collaborative exchanges.

(d) *As economy of search*. The economizing of cognitive energies in problem solving has been demonstrated by Newell and Simon (1972). New solutions are first looked for in familiar settings and then, because the costs of the search increase proportionately, in progressively more unfamiliar ones. Those solutions that have proved valid in the past are more likely to be the ones adopted in the future. They are learnt and retained as long as they continue to function. Learning takes place through the incorporation of experience into search rules, attention rules, and organizational goals.

(e) *As the modification of mental maps*. Since Argyris and Schön (1974) pointed out the gap between what organizations say they do (exposed theory) and what they actually do (theory-in-use), learning resides in the ability to bring explanations and the ideas guiding what is done into alignment with the organizational changes introduced. These authors make explicit reference to the cybernetic analogy; but at the same time they stress that organizational change requires modification of the way in which organizations represent reality to themselves. Organizational change and learning proceed in parallel.

The diverse knowledge interests of researchers have located learning at various positions along a continuum with adaptation to the environment at one extreme (together with the analogy of the cybernetic system) and the autonomous elaboration of thought at the other (together with the analogy of the organization as a mind).

In order to show the polysemy of the term 'learning', I shall refer to four systematic surveys of studies on organizational learning (Argyris and Schön 1978, Shrivastava 1983, Fiol and Lyles 1985, Huber 1991) that were published when the aim was to integrate the field and which comprise approximately 200 titles.

Argyris and Schön (1978) identify six conceptualizations of learning, according to how the organization is conceived:

(a) *if it is a group*, then learning takes place at the level of the interactions among individuals performing a task;

(b) *if it is a collective actor*, it learns from the experience stored in organizational maps and action-plans;

(c) *if it is a structure*, it learns by changing in accordance with the external and internal environment;

(d) *if it is a system*, it learns through error-regulating mechanisms;

(e) *if it is a cultural system*, learning is a process of socialization, of transformation of cognitive and judgmental patterns, as well as deliberate transformation of reality;

(f) *if it is a political arena*, the subject-matter of learning is the strategies—competitive and co-operative—employed to control organizational dynamics.

Shrivastava (1983) identifies four conceptualizations of organizational learning: as institutionalized experience, as adaptation, as knowledge of action-outcome relationships, as assumption sharing. In the first conception, by organizational learning is meant only that learning embedded in standard operating procedures, methods of communication and co-ordination, and shared understanding about tasks that have a persistent effect. In the adaptive approach, organizational learning is a complex process of mutual intraorganizational regulation triggered by the combination of different kinds of perceived stress. From the perspective of organizations as systems of knowledge of action-outcome relationships, organizations are purposeful systems or information processors relying on available inputs to confront the turbulence of their environment. From the assumption-sharing perspective, organizations are artifacts based on the cognitive maps that their members use to orient themselves in their interactions, and organizational learning consists of the process whereby these maps are modified.

The third typology (Fiol and Lyles 1985) seeks to prescribe what can be legitimately defined as organizational learning. Fiol and Lyles distinguish between the development of behavior and the development of cognition, so that changes of behavior constitute adaptation while learning only takes place within cognition where it is hierarchized into low-level (single-loop) and high-level (double-loop) learning. For Fiol and Lyles, organizational learning consists in the improvement of actions through better knowledge and understanding, while cognition is the development of reflections, knowledge and associations between past actions, their efficacy and future actions. But if we examine the definitions of learning and cognition more closely we discern the profound embarrassment, shared by all inquiries in the field, over a too-sharp distinction drawn between thinking and doing, over the difficulty of establishing a logical priority between thought and action, between the self-constitution of the subject (individual or organization) and its self-relationing to the environment.

Although complex, Fiol and Lyles (1985) distinguish three areas of agreement among studies on learning:

acknowledgment of the importance of adjustment to the environment;

a distinction drawn between learning by individuals and learning by organizations;

identification of four contextual factors which create and reinforce learning, and which are created by learning in turn: culture, strategy, structure, and environment.

Another influential survey of the literature is by Huber (1991), who proposes closer integration among studies of organizational learning. Huber identifies four constructs correlated with learning: knowledge acquisition, information distribution, the interpretation of information, and organizational memory. There is an abundant and diversified literature on the first two constructs, but the last two have been relatively ignored and require a great deal of further empirical research.

Huber concentrates on knowledge acquisition, which he divides into five subprocesses by which an organization acquires knowledge. It should be stressed, in fact, that Huber views learning as a process which he defines as follows: 'An entity learns if, through its processing of information, the range of its potential behaviors is changed' (Huber 1991). Unlike other authors, therefore, Hubert does not assume that learning immediately translates into either change or increased learning or into competitive advantage.

2. *The Growth of the Field and the Growing Divergences within It*

At the same time that Huber's review was published another authoritative voice remarked that: 'people in organization theory began to talk about learning … just about the time psychologists began to desert the concept' (Weick 1991). One of the main liabilities of the concept of organizational learning is that the traditional psychological models of learning (based on stimulus-response theory) have been transferred acritically to the study of organizational learning and with them a realist ontology, notwithstanding that the growth of the field began in those years since the field attracted the attention of scholars from different disciplines.

The growth of the field can be explained by at least three external factors: the speed of technological change, the advance of globalization, and growing corporate competition (Easterby-Smith et al. 1998).

The speed of technological change means that there is continual pressure on firms to reduce the time from conception to launch of new products, and communications innovations mean that companies and knowledge workers within them may have to adapt radically and in unforeseen directions. As a consequence, firms cannot rely on established practices, they need to learn new ways of doing things. The advance of globalization provides an additional source of pressure. It started with the opening up of labor markets in Asia in the early 1980s, and in Central Europe in the late 1980s, and it has accelerated with the establishment of global financial markets and with the establishment of global consumer markets nourished by the spending power of new middle classes in emerging economies such as India and China. To compete on a global scale, companies need to increase radically their scales of operation in order to keep up with the evolution of new technologies and businesses anywhere in the world. The preceding factors lead to a widely-held view that the most important contributor to corporate competitiveness in the future is going to be the ability of each organization to learn faster than its rivals.

In a scenario of fast economic and technological changes, knowledge came to be seen as an invisible asset and many consultants and firms caught onto the commercial significance of organizational learning. Two concepts express this move: the economics of knowledge and the leaning organization. What they have in common is the managerial interest in bringing knowledge under control at the service of production or innovation.

The exploitative ethos is particularly evident in the economics of knowledge and in knowledge management, whose starting point has been the identification of knowledge as a production factor distinct from the traditional ones of capital, labor, and land. The resource-based theory of the firm has conceptualized knowledge as 'core competencies' or 'core capabilities,' naturalizing the relationship by means of the metaphor of the tree of knowledge: the trunk and major limbs are core products and the root system is the core competence (Prahalad and Hamel 1990). The reification of knowledge has grown more overt with the 'objectified transferable commodity' envisaged by the knowledge-management approach, which treats knowledge as practically synonymous with information created, disseminated, and embedded in products, services, and systems. A number of theorists have focused on making practical interventions in organizations in order to help them to become 'learning organizations.' Some models of learning organizations have been based on observations of organizations that appeared to be good at learning and elements of good practice have been extracted and synthesized. Although the idea of a leaning organization is drawn on insights from organizational learning, nevertheless the two literatures have developed along different tracks and around them two different scientific communities are gathered. The divergences concern the questions of the why, what, and how of organizational learning, and also dilemmas about how to conduct research into it (Easterby-Smith et al. 1998):

(a) *Why organizational learning?* Two bodies of literature have developed according to whether the aim is analytic and descriptive or action oriented and prescriptive. The former focuses on 'How does an organization learn?' The latter focuses on 'How should an organization learn?'

(b) *What is organizational learning?* This question debates ontologies and the choice between a realist or a constructionist approach. It is about whether organizational learning is best seen as a rational and technical process or whether humanistic and political perspectives provide the best insights. A focus on data and information processing is likely to lead to assumptions of rationality which can be aided with appropriate learning technology, whereas the focus on human behavior inevitably leads to concern about hidden motives, emotions, and the possibility of political calculation.

(c) *Which methods of inquiry?* This area of divergence involves epistemologies and the preferred ways of inquiring into the nature of organizational learning. The major divide is between positivist methods with the emphasis on measurement and constructionist methods which emphasize language and the uniqueness of meaning systems. In general, there is a preference for quantitative methods (and for an individual/cognitive framework) in the USA due to the influence of academic journal house styles and research training in the management field; this contrasts with the strong development of qualitative methods (and a social/cultural framework) in European countries.

(d) *How to implement organizational learning?* A debate is open between incrementalists and radicals. Incremental learning is not only seen to be most common in organizations but also rather useful. A more radical perspective is represented by those who argue that because organizations normally operate on single-loop learning principles, it is more important for them to develop the capability of double-loop learning. A second debate is between those in the organizational development community who employ stage theories and those who use cyclical models. The former would suggest variously that there are qualitative differences between organizational processes and that these form progressive hierarchies. Those who propose cyclical models prefer to see the learning organization as a continual process of striving and improvement, that incorporates successive stages of data gathering, reflection, conceptualization, and experimentation.

The dilemma—as March (1996) expressed it—is between exploration and exploitation in the use of

knowledge. The same dilemma is reflected also in the way knowledge is conceived and situated in the relationship between the knowing subject and the object of knowledge.

3. Looking Forward to Knowing in Practice

The utopia of unifying and integrating the field that characterized the 1980s failed in the 1990s and gave way to the encouragement of the existing diversity of perspectives. Within diversity, a 'bandwagon' of theorists started to move behind the concept of practice (or activity) and the idea of community of practice as the place of learning.

Practice is the concept that allows the processes of 'knowing' at work and in organizing to be articulated as historical processes, material and indeterminate. In practice-based theorizing, different traditions of studying learning merge without losing their identity: situated learning theory offers the concept of community of practice, actor-network theory brings the idea of knowledge grounded in materiality, activity theory offers a pragmatic theory of action, and the ethnographic study of knowledgable communities brings a research method which is anti-individualistic.

The specific contribution of practice-based theorizing to the study of organizational learning consists in the following propositions:

Knowing is situated (Suchman 1987), resilient but provisional.

Learning is acquired through participation in communities of practice (Wenger 1998).

Knowledge and action are located in ecologies of social-material relations (Star 1995).

Organizing can be seen as a 'activity system' which reveals the tentative nature of knowledge and action (Blackler 1993). Incoherences, inconsistencies, paradoxes, and tensions are integral parts of activity systems.

Practice involves the accomplishment of alignments across human and nonhuman elements (Law 1994) from particular positionings at a particular times, within a network of relations.

The field of organizational learning has grown very rapidly, attracting the attention of scholars from disparate disciplines such as psychology, sociology, economics, and information systems. As a consequence, the field has become conceptually fragmented but there may be considerable value in distinguishing between groups of researchers in order to sharpen the clarity with which elements of the field are understood and may be developed in a noncompetitive way.

Bibliography

Argyris C, Schön D A 1974 *Theory in Practice*. Jossey-Bass, San Francisco, CA

Argyris C, Schön D A 1978 *Organizational Learning*. Addison-Wesley, Reading, MA

Blackler F 1993 Knowledge and the theory of organizations: Organizations as activity systems and the reframing of management. *Journal of Management Studies* **30**(6): 863–84

Easterby-Smith M, Snell R, Gherardi S 1998 Organizational learning and learning: Diverging communities of practice? *Management Learning* **29**(3): 259–72

Fiol C M, Lyles M A 1985 Organizational learning. *Academy of Management Review* **10**(4): 803–13

Huber G 1991 Organizational learning: The contributing processes and the literatures. *Organization Science* **2**(1): 88–117

Law J 1994 *Organizing Modernity*. Blackwell, Oxford, UK

March J G 1996 Exploration and exploitation in organizational learning. In: Cohen M, Sproull L (eds.) *Organizational Learning*. Sage, Thousand Oaks, CA pp. 101–23

March J C, Simon H A 1958 *Organizations*. Wiley, New York

Newell A, Simon H 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Prahalad C K, Hamel G 1990 The core competence of the corporation. *Harvard Business Review* **May–June**: 79–91

Shrivastava P 1983 A typology of organizational learning systems. *Journal of Management Studies* **20**: 8–28

Star S L 1995 *Ecologies of Knowledge*. State University of New York Press, Albany, NY

Suchman L A 1987 *Plans and Situated Action*. Cambridge University Press, Cambridge, UK

Trist E 1983 Referent organizations and the development of inter-organizational domains. *Human Relations* **36**(3): 269–84

Weick K 1991 The nontraditional quality of organizational learning. *Organization Science* **2**(1): 116–24

Wenger E 1998 *Communities of Practice. Learning, Meaning and Identity*. Cambridge University Press, Cambridge, UK

Yelle L 1979 The learning curve: Historical review and comprehensive survey. *Decision Science* **10**: 302–78

S. Gherardi

Learning Theories and Educational Paradigms

Theories of learning changed dramatically during the 1980s and 1990s. New ways of thinking about learning and knowledge have emerged to challenge traditional views of what learning is and how people acquire new knowledge and skills. The various theories, however, are not necessarily incompatible with one another. Rather, they provide different perspectives on the complex phenomena of learning and, in general, are relevant to different situations in which learning occurs.

Traditional theories of learning focus on changes in behavior. During the 1970s and 1980s, these theories gave way to cognitive theories that focus on mental processes and the understanding of complex material. In the 1990s, these cognitive theories were challenged by theories emphasizing social interaction and the sociocultural context. Many psychologists and educators no longer consider learning to be a process that occurs strictly within an individual. Rather, they see

learning as something distributed among several individuals (and/or environmental affordances—e.g., tools such as calculators) or existing within a 'community of practice' or a community of learners.

This evolution from behavioral to cognitive to social theories of learning was accompanied by new conceptions of knowledge, and how it is acquired. Traditional theories conceive of knowledge as a commodity that is transmitted, more-or-less intact, to individuals who learn it in a manner that permits them to reproduce the knowledge. According to these theories, knowledge is something an individual acquires.

More recent theories, however, view knowledge as something the learner constructs rather than assimilates. Thus, each person's knowledge is, to some extent, unique. According to these theories, truly objective knowledge, as we normally think of it, does not exist. Rather, knowledge resides in the community of learners (individuals) that creates it. This knowledge is distributed among members of that community and various environmental characteristics that support it.

The various theories of learning can be summarized in several ways. The most common is to classify the theories with regard to the type of theoretical explanation each provides: i.e., behavioral, cognitive, and social. Each of these three general views of learning and knowing is associated with a particular philosophical and methodological perspective: empirical, rational, and pragmatic-socio-historical, respectively (Greeno et al. 1996).

These and other theoretical approaches to learning are discussed in more detail below. First, however, the meaning (definition) of 'learning' and several general issues will be considered.

1. *Definitions of Learning*

Understanding any theory requires a clear idea of what the theory is trying to explain. As new theories of learning emerged, the conception of learning changed. In short, 'learning' was defined in new ways. These fresh perspectives add to our understanding of learning, but to the extent the theories explain different aspects of the learning process, fitting the pieces together is not a simple task.

Traditionally, learning has been defined in terms of a change in a person's behavior, knowledge, or ability to perform some task (including intellectual tasks). These definitions usually stipulate that the change must result from the learner interacting with his or her environment, although most researchers would include the mental, as well as the physical, environment.

Change is the main criteria in these definitions, although changes that result from physiological maturation and temporary changes (e.g., those that result from drugs or fatigue) are explicitly excluded from the definition. This definition remained pretty much the same with cognitive theories, although as might be expected, the focus shifted to changes in mental processes, knowledge structures, and understanding.

With the advent of constructivism and sociocultural theories of learning, knowledge no longer was seen as a commodity to be transmitted. Knowledge was now seen as being distributed among several individuals and 'affordances' (tools, artifacts, etc.) provided by the environment. The definition of learning changed to interacting with other learners (including individuals with more experience) on a meaningful task. Some psychologists might question whether this is truly a definition of 'learning,' but it does describe the phenomena in which these investigators are interested.

Lave and Wenger (1991), for example, are explicit about redefining learning, thinking, and knowing as 'relations among people in activity in, with, and arising from the socially and culturally structured world' (p. 51). From this perspective, learning is largely a matter of becoming attuned to both the constraints and affordances (what the situation makes possible) of participating (learning) in this social setting (Greeno et al. 1996, 1998). As Greeno et al. (1996) suggest, 'When knowing is viewed as practices of communities and the abilities of individuals to participate in those practices, then learning is the strengthening of those practices and participatory abilities' (p. 23).

Thus, several different definitions of learning are apparent in the literature. If a phenomenon is defined in different ways, it is reasonable to expect that substantial and qualitative differences will exist in the theories proposed to explain the phenomenon. It sometimes seems as if a single theory should be able to explain learning in all situations. The various theories of learning often compete with one another as to which is 'best,' but actually they complement rather than contradict one another. One theory may provide a better explanation than another theory for a particular type or aspect of learning, but there is plenty of room for more than one theory, for learning is multifaceted.

2. *General Issues*

Several general issues provide a context for interpreting what is presently understood about the nature of learning, how it occurs, and how it can be facilitated.

2.1 *More than one Type of Learning*

As noted previously, there are several different types of learning. Those working within a particular theoretical perspective often ignore (occasionally deny) the importance or existence of other types of learning. To fully understand learning, one must realize that a particular theory may provide a good explanation of

learning in one situation but a poor explanation of learning in a different situation.

There is little agreement on exactly how many different types of learning exist. Nevertheless, several different kinds of learning are evident in the following examples: (a) learning to tie one's shoes; (b) becoming fearful of the dark after being scared by a loud clap of thunder after the lights went out during a storm; (c) explaining the causes of World War II; and (d) negotiating the meaning of 'learning' with a person holding a different theoretical perspective. Different theories are good at explaining each type.

When evaluating the validity of a particular theory, it is wise to consider what the person is learning and what is taken as evidence that learning has occurred. For example, a teacher might assume that a student has gained an understanding of a topic when he or she has only memorized certain facts.

2.2 *What is Learned? Learning vs. Understanding vs. Participation*

Theories of learning differ with regard to what they assume a person is learning. Traditional theories suggest that individuals acquire associations between stimuli and responses. These associations are reflected in specific behaviors, including verbal behavior (words), and various skills.

Cognitive theories, on the other hand, emphasize the acquisition of knowledge and mental processes rather than behavior. Highly organized knowledge structures, rather than isolated facts, are seen as the outcomes of learning. The focus of these theories is on understanding, problem solving and conceptual change rather than the memorization of inert knowledge.

'Understanding' is difficult to define in a rigorous manner. Nevertheless, certain characteristics of the concept can be identified. Knowledge must be structured and organized in order for it to be meaningful and capable of being understood. Few people would suggest, for example, that one could understand a particular telephone number, although most would agree it is possible to learn, know, or remember one. Understanding a body of knowledge involves establishing relationships among the concepts and facts that comprise that body of knowledge. Such understandings can be assessed through paraphrasing, summarizing, or answering questions about the material and/or by performing a transfer task.

Situated and sociocultural theories are concerned with individuals participating in a community of learners. Thus, rather than acquiring behavior or knowledge, learners are seen as developing social practices appropriate for functioning in such communities, as well as the ability to use tools (such as computers and calculators) that enhance one's 'intelligence.'

3. *Traditional Approaches to Learning*

Traditional theories of learning view learning within a stimulus-response framework—that is, as something that happens to the learner from the outside in. A stimulus occurs in the environment (perhaps something done or said by another person), and the individual responds. Depending on what happens next (the occurrence of reinforcement or punishment), the probability that the learner will make the same response in the future either increases or decreases. For instance, the frequency of a student's hand raising either increases due to the teacher's positive statement following a correct response or decreases due to the teacher consistently calling on other students or the embarrassment that results from providing an incorrect answer.

These theories provide good explanations for certain kinds of behavior, especially the acquisition of factual information, skills, and behavior patterns. In these cases, the focus is on the performance of behavioral tasks rather than adding components to the learner's cognitive structure. Behavioral theories, however, provide poor explanations for the way individuals come to understand complex ideas and phenomena. The importance of classical conditioning is frequently dismissed as relevant only to dogs salivating (Pavlov's initial research paradigm). This type of learning, however, provides by far the best explanation of how and why humans respond emotionally to a large variety of stimuli and situations. Fear of the dark (or water), a strong dislike (or hatred) for a particular person or group of people, phobias of various kinds, and infatuation with another person are only a few examples of the types of emotional reactions acquired through classical conditioning.

4. *Contemporary Approaches to Learning*

Most contemporary theories of learning assume that learning arises out of the internal conditions or states of the learner (that is, from the inside out rather than the outside in). In general, these internal factors are considered more important than external environmental factors. The learner does not merely 'record' or remember the material to be learned. Rather, he or she constructs a unique mental representation of the material and the task to be performed, selects information perceived to be relevant, and interprets the information on the basis of his or her existing knowledge and current needs. During this process, the learner adds information not explicitly provided by an outside source, whenever he or she needs such information to make sense of the material being studied. Most of these theories, to varying degrees, share three sets of beliefs. First, meaningful learning (i.e., learning for understanding) is an active, self-regulated, constructive, cumulative, and goal-oriented process. Second, learning is dependent on (situated in) the

particular context in which it occurs. Third, learning is fundamentally a social, cultural, and interpersonal process; a process governed as much by social and situational factors as by cognitive factors. Each of these characteristics will be discussed.

4.1 Learning is Active and Self-regulated

It often is said that traditional theories of learning view the learner as passive, although it is more accurate to say they characterize the learner as reactive—i.e., the learner reacts to a stimulus in the environment, and that response is reinforced (or punished). Although behavioral psychologists usually recognize the need for the learner to be active in practicing associative bonds of various kinds, this type of activity is far different from the activity suggested by more recent theorists.

Current theories of learning view the learner as mentally active, even proactive, in his or her efforts to understand, for instance, the use of metaphors or the concept of photosynthesis. The learner carries out various cognitive and metacognitive operations on the information being learned, a process that results in the material being acquired in a meaningful manner.

During this process, the learner makes many decisions about what to do next (e.g., rehearse a particular piece of information, seek an answer to a question that comes to mind, look for similarities among various pieces of information). Effective learners also monitor the learning process, making periodic checks of how well the material is understood. As learning proceeds, the learner regulates his or her activities, making any adjustments that are needed. Among the factors involved in this process of self-regulated learning are metacognition, self-efficacy, and studying.

4.2 Learning is Constructive

The constructive, rather than reproductive, nature of learning is an integral part of most contemporary theories of learning. Understanding is achieved, according to these theories, through an active process of construction rather than by the passive assimilation of information or the memorization of facts. These theories discard the idea that knowledge is an entity that can be passed from one person (teacher, book, etc.) to another (the learner). Rather, each learner (individually or as a member of a group) perceives and interprets new information in a unique manner, based on factors such as prior knowledge, interest, motivation, and attitude toward self. The learner then elaborates this interpretation by relating it to his or her existing understanding of the topic and/or other aspects of the material being learned. As a result, the learner's understandings are reorganized and refined, and he or she may experience growth of general cognitive abilities such as problem solving and metacognitive processes.

Since no two people interpret the same information in the same way, no two people end up with the same understanding of the concepts and facts being studied. In the final analysis, the manner in which the learner processes new material and the type of cognitive processing in which he or she engages is the single most important determiner of what the individual learns.

There is not, however, a single constructivist theory of learning. Considerable differences exist among constructivist theories with regard to the relative influence of the mind versus the environment, private versus public knowledge, and individual versus social factors (Phillips 1995, 2000). Radical constructivism, cognitive constructivism, social constructivism, and situation cognition represent different types of constructivist theories that occupy different positions on these dimensions.

4.3 Learning is Cumulative

All learning builds upon the individual's prior knowledge and experiences, which can either facilitate or inhibit the new learning. Research on schema theory and the difficulties involved in overcoming one's prior conceptions (sometimes referred to as misconceptions when they differ from established conceptions in a field such as science) illustrate the potent influence that prior knowledge has on learning.

4.3.1 Schema Theory. Schema theory, an important aspect of cognitive theories of learning, suggests learners acquire complex, organized structures of information about a topic, such as eating in a restaurant or playing music (Derry 1996, Winn and Snyder 1996). When someone encounters information (perhaps by hearing a story), it is interpreted according to a schema considered appropriate by the learner. The learner will fill in details missing from the material by making inferences based on information contained in the schema.

For example, a person hears a story about four friends having dinner in a restaurant. Unless the story mentions that the group left an unusually small or large tip, the listener will assume with considerable confidence that a normal tip was left. Leaving a tip is part of the schema for eating in a restaurant, although the amount of the tip may vary, depending on the complexity of the person's schema with regard to tipping in different types of restaurants and different countries.

4.3.2 Prior Conceptions. When learners engage in a learning activity, they bring with them a set of prior

conceptions of the topic being studied (e.g., Rezaei and Katz 1998). These initial conceptions make sense to the learner, but frequently they are considered incorrect by the adult community. Because they are so well established, these prior conceptions are very difficult to change, probably far more difficult than most people realize.

In order to understand photosynthesis, for instance, many students must overcome erroneous conceptions about 'food' developed from their experiences with food for people. They 'must abandon their assumptions about the metabolic similarities between plants and humans and restructure their thinking about the nature of food … [In this regard, they must] learn that [certain of their] beliefs about food do not generalize from humans to plants while [other beliefs must] be clarified, expanded, and given new prominence' (Anderson and Roth 1989 p. 278).

Students who score well on classroom tests and receive high grades in a course often maintain inaccurate conceptions of key concepts they had before the topic was studied. These students' true level of understanding is often concealed by tests and assignments that permit them to do well without understanding the material (e.g., by memorizing information). Students must be asked to apply their understandings to a wide range of problems and situations before can one have confidence that they possess an appropriate and adequate understanding of the topic.

4.4 Learning is Goal-oriented

Instructional goals can facilitate learning, although the use of instructional objectives is rejected by most constructivist and social theories of learning. Narrowly defined objectives are reminiscent of behavioral theories and are seen by many as stifling of meaningful learning. Yet, meaningful learning is more likely to be successful if the learner has at least a general idea of the goal being pursued and holds appropriate expectations for achieving the desired understanding.

There are many ways in which learning goals can be established. In many instructional situations, it is appropriate for students to develop or discover their own goals. Identifying appropriate learning goals (which may change as learning progresses) can become an important part of self-regulated learning and metacognitive control. In any case, it is the student'*s* goals that are critical for learning. The statement of objectives or goals by a teacher is not sufficient. Unless the learner adopts the goals as his or her own, they will have little if any impact on the learning process.

4.5 Learning is Situated and Distributed

According to situated theories, knowledge and learning are situated in a particular social context and distributed across various individuals, artifacts and tools (books, calculators, etc.), and the norms and practices of the group in which the learner is participating. The interactive nature of these various systems is considerably broader and more complex than the behavior and cognitive processes of the individual learner (Greeno et al. 1998). As Greeno (1997) notes, 'the situative view [of learning] focuses on practices in which individuals have learned to participate, rather than on knowledge that they have acquired' (p. 6). The role of the individual learner, a hallmark of traditional theories, is no longer a central focus in many current theories.

The suggestion that knowledge and learning are distributed is readily apparent in many common situations. For instance, individuals in a class (or book club) studying 'To Kill a Mocking Bird' will have varying degrees of knowledge about (and understanding of?) the plot, the various characters, the issues addressed, etc. Discussing, analyzing, and critiquing these various perspectives increase each person's understanding of the book and the issues that are raised. But this understanding is partial and incomplete. The individual learner will never have the full understanding that is present in the group. In fact, some of the person's knowledge and understanding may depend on (be available only during) the interaction of various agents (people, artifacts, etc.) that are part of that particular situation.

4.6 Learning is Social, Cultural, and Interpersonal

One of the most important differences between traditional and contemporary theories of learning is the role 'participation' plays in most current theories. Learning is seen as a social, cultural, and interpersonal activity. A community of learners, however, does not refer only to students. Such a community might consist of professional practitioners, a religious organization, or a political enclave.

Many present-day theorists agree with Lev Vygotsky that cultural development and learning occur initially in the social (interpersonal) realm. Social interaction and language play critical roles during this phase. As ideas, understandings, and ways of interacting become established on the social plane, they begin to move to the psychological (intra-personal) plane. As Weade (1992) notes:

> The structure and meaning of an evolving academic discourse is embedded within an evolving social structure. Simply put, the social structure mediates who can talk to (or act toward) whom, when, where, in what ways, for what purposes, under what conditions, and with what tangible or imagined outcomes (p. 95).

5. Mechanisms for Learning

The various theoretical perspectives differ in the way they conceptualize the factors responsible for learning

to occur. Traditional theories suggest learning is the result of reinforcement or feedback. Practice also is important, although practice will not produce learning unless it is accompanied by reinforcement. In the case of classical conditioning, contiguity (proximity) of two stimuli is the primary cause of learning. Cognitive theories generally postulate that learning occurs as the result of elaboration, thinking, problem solving, and/or reflection. It is the mental processing of information, rather than external reinforcement, that produces learning. Shuell (1996) suggests that 12 'learning functions' need to be engaged for meaningful learning to occur: (a) establishing expectations; (b) establishing motivation; (c) activating prior knowledge; (d) focusing attention on relevant features of the material being studied; (e) encoding information; (f) comparing information; (g) generating hypotheses about possible relationships and solutions; (h) repetition of information until it is integrated into higher-level concepts; (i) receiving feedback on adequacy of one's understanding; (j) evaluating one's hypotheses and the feedback received; (k) monitoring the learning/understanding process; and (l) combining, integrating, and synthesizing available information to form a new (revised) understanding of the topic being studied.

Constructivist, situated, and sociocultural theories suggest that learning results from social interaction, discourse, negotiation, and activity (physical and/or mental) in the performance of 'authentic' activities. Presumably, different types of discourse and negotiation, for example, result in different types of learning, but very little research on these factors is presently available.

6. *Educational Paradigms*

Just as there is not a single type of learning, neither is there a single best educational paradigm or instructional practice. Schools and educational programs are far more likely to be based on philosophical presumptions than on empirical and theoretical foundations of learning. Schools differ in their philosophical beliefs about human nature, children, locus of authority, and knowledge, as well as beliefs about the nature of teaching and learning. Every educational system and instructional program contains a theory of learning, although this theoretical foundation usually is implicit.

Some of the philosophical and theoretical differences have endured for centuries, and the debate is likely to continue. For example, the 'factory model' of schooling, based on production and management procedures proven successful during the industrial revolution, dominated education in the USA for many years. This perspective on teaching and learning, however, stands in sharp contrast to the voices of Thoreau, Dewey, and others who advocate discovery, social reform, and freedom as the appropriate means of education (Shuell 1993).

These differences are evident in the various theories of learning discussed previously. Traditional classrooms, for example, are based on traditional theories of learning and knowledge. Knowledge is seen as an entity that can be passed from one person to another (in this case, from teacher to student). Consequently, the curriculum consists primarily of factual information. Since learning is viewed as a process of memorization and practice, an appropriate classroom is one in which the teacher tells students what they need to know through the use of expository and didactic methods. The consistency between the underlying theory of learning and instructional practices is quite apparent.

On the other hand, if one believes that knowledge is something created afresh by students and that learning occurs from working on authentic tasks in a social environment, a very different type of classroom emerges. A visit to this classroom would find students working in groups, very likely discussing how best to solve a problem in order to complete a task. Or they might be found negotiating the meaning of a particular concept in an effort to understand and perhaps apply it to the solution of some problem. Once again, consistency exists between a theory of learning and the instructional practices occurring in the classroom.

Schools and classrooms are organized in different ways, have different learning practices, and assess students in different ways. Students learn more than what they are explicitly taught. They 'also develop patterns of participation and identities that are shaped by these different practices in which they learn' (Greeno et al. 1998: 14). Classroom learning involves social, emotional, and participatory—as well as cognitive—factors. Several educational paradigms based on contemporary theories of learning are discussed.

6.1 *Cognitive Apprenticeship*

The cognitive apprenticeship model of teaching grew out of situated theories of learning. Apprenticeships (both formal and informal) are a common way of learning in non-school settings. Although there are differences between traditional models of apprenticeship and cognitive-apprenticeship models of teaching, the latter were proposed in a belief that the knowledge and skills learned in school have become too abstracted from their use in the world outside of school (Shuell 1996).

Generally, cognitive apprenticeship models of teaching involve a series of six teaching procedures. Students first observe an expert (usually the teacher) model the desired performance in an environment similar to the ones in which the performance is to occur. Second, coaching (hints, feedback, modeling, reminders, etc.) is provided. Next, conceptual scaffold-

ing is provided with the student performing as much of the task as possible, although this external support is gradually faded as the student gains proficiency. Fourth, students are asked to articulate their knowledge and understanding of the task, followed by a request for them to reflect upon their understanding and reasoning. Sixth, students are encouraged to explore new ways in which the knowledge or skill can be used. Cognitive apprenticeship programs have been developed in the areas of reading, writing, and mathematical problem solving.

7. *Reciprocal Teaching*

Reciprocal teaching is a cognitive-apprenticeship model for teaching reading comprehension. The procedure involves a structured dialogue in which the teacher and students take turns (hence the name) leading a discussion that includes four types of strategic activities: predicting, questioning (making up a question on the main idea), summarizing (self-review), and clarifying. Throughout the instructional session, the teacher provides guidance and feedback (scaffolding) tailored to the needs of the current student discussion leader in an attempt to improve the level of his or her performance. A more extensive summary of reciprocal teaching and a list of references are available in Shuell (1996).

7.1 *Conceptual Change Teaching*

A large body of literature, primarily in science and mathematics education, has evolved on overcoming the misconceptions that many students have prior to the start of classroom instruction. As noted previously, these pre-existing conceptions are difficult to change. Changing them involves more than students gaining 'insight,' and the process is a slow one. Four conditions are generally considered important for conceptual change to occur in students (Posner et al. 1982): (a) the student must become dissatisfied with his or her existing conceptions; (b) the new conception must be intelligible to the student; (c) the new conception must initially appear plausible and capable of solving the problems generated by its predecessor; and (d) the student must see the new conception as being useful for understanding a variety of situations.

8. *Summary*

New conceptions and theories have expanded our understanding of human learning by focusing on factors ignored in earlier research. As theories progressed from behavioral to cognitive to situated to sociocultural explanations of learning, the emphasis shifted from changes in behavior to acquisition of knowledge to social participation. The corresponding mechanisms for learning also changed from reinforcement and practice to thinking and other mental processes to participation in a community of learners. To a large extent, however, the various theories compliment one another, for each focus on a different aspect of the learning process or a different type of learning. Generally speaking, contemporary theories suggest that meaningful learning is an active, self-regulated, constructive, cumulative, goal-oriented, context-dependent, social, culturally, and interpersonal process. Various educational paradigms have emerged from these theories, including cognitive apprenticeships, reciprocal teaching, and teaching for conceptual change. With regard to education, these theories suggest that meaning does not reside in the material being learned or the manner in which it is presented to students. Information may possess a potential for being meaningful, but it is the learner who makes it meaningful by processing it in a meaningful manner. It is not sufficient for the teacher to explain or demonstrate how various concepts and facts are related. This activity can occur without the occurrence of meaningful learning on the part of the student.

See also: Educational Learning Theory; Self-regulated Learning; Situated Learning: Out of School and in the Classroom; Teaching and Learning in the Classroom

Bibliography

Anderson C W, Roth K J 1989 Teaching for meaningful and self-regulated learning of science. In: Brophy J (ed.) *Advances in Research on Teaching*. JAI Press, Greenwich, CT, Vol. 1, pp. 265–306

Derry S J 1996 Cognitive schema theory in the constructivist debate. *Educational Psychologist* **31**: 163–74

Greeno J G 1997 On claims that answer the wrong questions. *Educational Researcher* **26**(1): 5–17

Greeno J G, Collins A M, Resnick L B 1996 Cognition and learning. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Simon and Schuster Macmillan, New York, pp. 15–46

Greeno J G 1998 The Middle School Mathematics Through Applications Project Group. The situativity of knowing, learning, and research. *American Psychologist* **53**: 5–26

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, New York

Mayer R E 1992 Cognition and instruction: Their historic meeting within educational psychology. *Journal Educational Psychology* **84**: 405–12

Phillips D C 1995 The good, the bad, and the ugly: The many faces of constructivism. *Educational Researcher* **24**(7): 5–12

Phillips D C (ed.) 2000 *Constructivism in Education: Opinions and Second Opinions on Controversial Issues* (99th yearbook of the National Society for the Study of Education, Pt. I). University of Chicago Press, Chicago

Posner G J, Strike K A, Hewson P W, Gertzog W A 1982 Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education* **66**: 211–27

Rezaei A, Katz L 1998 A cognitive model for conceptual change in science instruction with a focus on educational software development. *Journal of Educational Computing Research* **19**: 155–74
Sfard A 1998 On two metaphors for learning and the dangers of choosing just one. *Educational Researcher* **27**(2): 4–13
Shuell T J 1993 Toward an integrated theory of teaching and learning. *Educational Psychologist* **28**: 291–311
Shuell T J 1996 Teaching and learning in a classroom context. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Simon and Schuster Macmillan, New York, pp. 726–64
Weade G 1992 Locating learning in the times and spaces of teaching. In: Marshall H H (ed.) *Redefining Student Learning: Roots of Educational Change*. Ablex, Norwood, NJ, pp. 87–118
Winn W, Snyder D 1996 Cognitive perspectives in psychology. In: Jonassen D H (ed.) *Handbook of Research for Educational Communications and Technology*. Simon and Schuster Macmillan, New York, pp. 112–42

T. J. Shuell

Learning to Learn

The quest for understanding and harnessing an individual's ability to identify, reflect upon, improve, and control his or her learning processes and outcomes, began in ancient times with the study of mnemonics (memory aids) and Socratic teaching methods. However, it is only since around 1975 that incredible progress in understanding 'learning to learn' and the methods that can be used to improve it have occurred. This entry will present a definition of learning to learn followed by a brief history of the development of the study of learning to learn processes. A model summarizing much of the current research in this area will then be presented as well as a summary of the results of current research examining the effectiveness of methods for improving learning abilities and skills. The entry concludes with a discussion of suggestions for future research and conceptual development in the area of 'learning to learn.'

1. *Learning to Learn Strategies*

'Learning to learn' strategies include any thoughts, behaviors, beliefs, or emotions that facilitate the acquisition, understanding, or later application and transfer of new knowledge and skills in different performance contexts. They range from active rehearsal to help remember word lists, to the use of elaboration and organization to encode, integrate, and later recall or apply knowledge across several content areas. Learning to learn strategies help generate meaning for the new information that is to be learned. Repeating the names of the planets in order, organizing the discoveries of the great explorers by creating timelines and mind maps, comparing and contrasting the causes of World War I and World War II, are all examples of learning strategies. They all are designed to help the learner generate meaning and store the new information in memory in a manner that will facilitate integration with related knowledge and increase the probability of later recall and use, particularly in transfer contexts.

A learning strategy is also a plan for orchestrating cognitive resources to help reach a learning goal. 'Learning to learn' strategies have several characteristics in common. First, they are goal-directed. Learning strategies are used to help meet a standard of performance or to reach a learning goal. Second, learning strategies are intentionally invoked, which implies at least some level of active selection. The selection of one or more of these strategies is determined by a number of factors, such as a student's prior experience with the strategy, his or her prior experience with similar learning tasks, his or her ability to deal with distractions, and the student's commitment to his or her goals. Third, cognitive learning strategies are effortful; they require time and often involve using multiple, highly interactive steps. Because of the effort required, a student must be motivated to initiate and maintain strategy use (e.g., see *Motivation, Learning, and Instruction*). In addition, the student must believe that the strategy will be effective and that he or she can be successful using the strategy. Finally, cognitive learning strategies are not universally applicable—they are situation-specific. The student's goals, the task requirements, the context, and other factors all interact to help determine which strategy may be best. To be successful in selecting and using a strategy, a student must understand under what circumstances a given strategy is, or is not appropriate.

The following section will discuss the evolution of learning to learn as a field of study and a source of applications for many students experiencing difficulty in an educational setting.

2. *Historical Overview*

In ancient times the great Greek orators were known for giving long speeches that sometimes lasted more than two or three hours. They managed this task through the use of mnemonic, or memory devices that helped them to remember both the ideas and the order of the ideas for their presentations. The most famous of these techniques is the Method of Loci. Basically, using the Method of Loci involves remembering a series of locations in order. This could be the buildings and landmarks the orator passed on the way to the arena or the rooms in a large villa. As the ideas were

generated for the speech, the orator would 'store' an image of the idea at a specific location. When the time came to give the speech, the ideas would be recalled as the speaker 'went' from location to location.

For years the study of mnemonics lay dormant until interest renewed in the second half of the twentieth century. Psychology has gone through major upheavals in the twentieth century. The rejection of philosophy and introspection as primary tools for the study of psychology led to the rise of behaviorism and the rejection of 'mental' phenomena. In the rush to be scientific in developing theories and methods, researchers lost their focus on the mind and cognitive processes. It was not until the late 1960s and the 1970s that studies of perception, memory, and cognitive processes once again gained prominence (see *Educational Learning Theory*, *Learning Theories and Educational Paradigms*). Some of the earliest studies of memory re-examined the classical mnemonic systems in an attempt to identify and understand the cognitive processes involved in creating and using them. In the 1970s this literature was enriched by the work of Flavell and others on what he called 'metacognition' (Flavell 1971). Metacognition refers to thinking about thinking. It focuses on knowledge, reflection, and analyses of how an individual thinks and learns (see *Metacognitive Development*: *Educational Implications*). Building on both of these themes, researchers increasingly turned their attention to cognition, the variables that impact cognition, and the degree to which cognition can be influenced through educational interventions. This is the research that led most directly to the study of learning to learn.

3. *Types of Learning to Learn Strategies*

Weinstein and Mayer (1986) summarized much of the research and findings in 'learning to learn' and developed a taxonomy to describe the six major categories of learning strategies. The following is a summary of the categories in the taxonomy.

3.1 Rehearsal Strategies for Basic Learning Tasks

Many instructional tasks require simple recall. The acquisition of basic knowledge, however, is often just a first step in the creation of a more extensive and integrated knowledge base in an area. An example of a task in this category is memorizing the order of the colors in the spectrum. A useful strategy for this type of learning might be to repeat, in sequential order, the names of the colors.

3.2 Rehearsal Strategies for Complex Learning Tasks

Learning tasks in this category are more demanding and tend to involve knowledge and skills extending beyond the superficial learning of lists or unrelated bits of information. Examples of strategies in this category would include highlighting class notes or copying key ideas from the class readings. Like the rehearsal strategies for simple tasks, these methods are often most appropriate in the early stages of building a knowledge base in an area.

3.3 Elaboration Strategies for Basic Learning Tasks

Meaningful learning often involves building bridges between what is already known and what a student is trying to learn. Elaboration strategies help the student build these bridges by using previous knowledge or experiences to add to what he or she is trying to learn as a way to make the new information more meaningful and memorable. Examples of strategies in this area include generating a mental image of a scene described in a novel and relating a scientific principle to everyday experience. The facilitating effect of using these strategies is usually attributed to both the processing that is involved in creating the elaboration and the product itself.

3.4 Elaboration Strategies for Complex Learning Tasks

Strategies in this category help the learner to be more active in building bridges between prior knowledge or experience and more complex tasks. Examples of strategies in this category include paraphrasing and summarizing, applying a problem-solving strategy to a new problem, creating analogies, and using everyday experience to help understand a new concept.

3.5 Organization Strategies for Basic Learning Tasks

Organization strategies involve translating or transforming information into another form and creating some sort of scheme to provide structure for this new way of characterizing the information. Examples of organization strategies for basic tasks include grouping foreign language vocabulary words by their parts of speech and grouping the battles of World War II by their geographic location. Like elaboration strategies, the facilitating effect of organization strategies is usually attributed to both the processing that is involved in creating the organization and the product itself.

3.6 Organization Strategies for Complex Learning Tasks

Complex tasks can often be made more meaningful and more manageable by using organization strat-

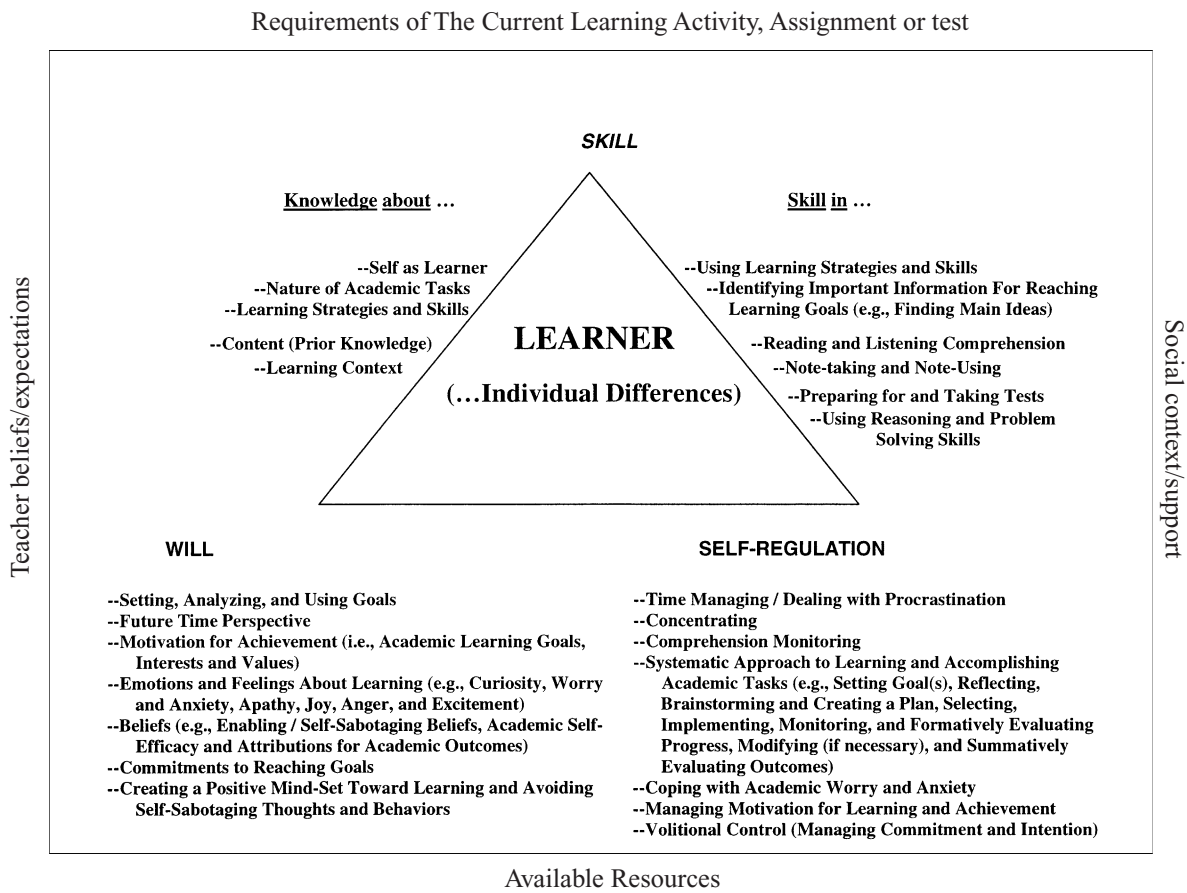

Figure 1
Model of strategic learning.

egies. Examples of strategies in this category include creating a hierarchy of main ideas to use in writing a term paper, generating a force diagram in physics, or creating an outline for the class notes from a section of a course. Like all of the strategies in the elaboration and organization categories, it is both the process and the outcome that appear to help in both knowledge acquisition and later recall and performance.

4. *Model of Strategic Learning*

As research in 'learning to learn' expanded in the 1980s and 1990s, it became clear that a number of different factors contribute to, and influence both choice and type of strategy used by a student (see *Self-regulated Learning*). For example, it is not enough for a student to know about different types of learning strategies. He or she must also know how to use them and under what conditions it is appropriate to use them. Another factor that affects the strategies selected and used by a student is the social contexts of both the educational setting, in general, and the individual classroom. There are also support variables that affect a student's attitudes and beliefs about learning and his or her motivation to succeed. These variables help to generate a context in which effective learning can take place. For example, a student must be clear about his or her long-term and short-term academic goals in order to use them to aid his or her motivation to initiate and maintain use of effective learning strategies.

In the past researchers and educational program developers have usually focused on one or a subset of topics within learning to learn, such as cognitive elaboration strategies or student motivation. Current work is purposefully examining the interaction among two or more components related to the acquisition and use of learning strategies. This change is a result of the increasing understanding of the nature of student learning and school achievement at all educational levels. Researchers have found that it is the interaction

among varying factors that results in successful learning and transfer of new knowledge and skills. The components that seem to have the greatest impact on a student's acquisition and use of learning strategies are summarized in a model developed by Weinstein and her colleagues at the Cognitive Learning Strategies Project located at the University of Texas at Austin in the USA (see Fig. 1).

This model is an extension of the earlier model developed by Weinstein and Mayer (1986). It focuses on variables impacting strategic learning; learning that is goal driven. Briefly, the core of the model is the learner's long-term learning goals and self-system variables. Goals are considered to be critical components that provide direction for self-regulating thoughts and actions and for generating and sustaining the motivation necessary to carry out those thoughts and behaviors. A hallmark of a strategic learner is that he or she sets specific and challenging, yet realistic, learning goals and actively pursues them. These learning goals are outcomes to be achieved in and of themselves, but are also subsumed within larger and much longer-term goals and thereby have a future utility value. However, setting specific learning goals is necessary but not sufficient to ensure successful learning outcomes. A strategic learner must also have the skills, motivation and self-regulatory thoughts, beliefs, and behaviors necessary to successfully pursue such goals. The core of the model is surrounded by these three components: (a) the learner's 'skill' levels in relation to the learning task at hand; (b) the learner's 'will' or motivation to accomplish the desired outcome; and (c) the learner's 'self-regulating thoughts, beliefs, and actions' (sometimes referred to as executive control processes in earlier literature). It is the properties that emerge from the interactions among the learner and these components that are at the core of 'learning to learn.' For any given learning task, the student must take into account variables from each of these components.

5. *Current Research, Methodological Issues, and Directions for Future Research*

Much of the current research in learning to learn has focused on improving student learning outcomes, primarily in college or tertiary settings. Researchers have demonstrated the effectiveness of teaching learning to learn strategies to students, often with quite dramatic results. These results have also been robust across content domains, individual classes, and years of college.

Given the importance of learning to learn for student achievement and retention to graduation, additional research is needed to address some of the methodological and conceptual issues in this area. First, there is a need for the development of consistent terms for the various aspects of learning to learn. Currently, each individual or group of researchers uses somewhat different terms to often describe very similar ideas and processes. Second, there is a need for the further development of models of learning to learn at the global level and at more task-specific levels. Third, more knowledge is needed about the nature of transfer of 'learning to learn' strategies. Fourth, measurement of 'learning to learn' strategies and processes needs to be refined. Current instruments like the 'Learning and Study Strategies Questionnaire' and the 'Motivated Strategies for Learning Questionnaire' are good general diagnostic screening measures, but more precise measurements of the components of 'learning to learn' and how they interact in a given learning context are needed. Fifth, additional developmental work to create instructional paradigms to teach learning to learn in both traditional learning contexts and web-based applications is needed.

'Learning to learn' is crucial for success in any learning context. As society moves toward a world where lifelong learning is a fundamental reality, the importance of understanding and using learning to learn strategies will only increase.

See also: Cognitive Development: Learning and Instruction; Cognitive Styles and Learning Styles; Educational Learning Theory; Mastery Learning; School Learning for Transfer; Self-regulated Learning; Transfer of Learning, Cognitive Psychology of

Bibliography

Boekaerts M, Pintrich P, Zeidner M (eds.) 2000 *Handbook of Self-Regulation*. Academic Press, New York, NY

Flavell J H 1971 First discussant's comments: What is memory development the development of? *Human Development* **14**: 272–8

Schunk D H, Zimmerman B J (eds.) 1998 *Self-Regulated Learning: From Teaching to Self-Reflective Practice*. Guilford, New York

Weinstein C E, Hume L M 1988 *Learning and Study Strategies*. Academic Press, San Diego

Weinstein C E, Mayer R E 1986 The teaching of learning strategies. In: Wittrock M C (ed.) *Handbook of Research on Teaching*, 3rd edn. Macmillan, New York, pp. 315–27

C. E. Weinstein

Legal Culture and Legal Consciousness

The phrases 'legal culture' and 'legal consciousness' are used analytically to identify the understandings and meanings of law that circulate in social relations. The prefix *legal* characterizes an aspect of the general culture that is associated with law, legal institutions,

legal actors and behaviors. Legal *culture* refers to an aggregate level (macro or group) phenomenon; legal *consciousness* usually refers to micro level social action, specifically the ways in which individuals interpret and mobilize legal meanings and signs.

Culture and consciousness are constructs debated and contested regularly, and the adoption of these terms in legal scholarship has also been fraught with controversy. The meaning of these terms is unstable, both theoretically and empirically. This may be due to a long history in social science as well as common parlance, but it may also be, as Raymond Williams writes, because 'culture is one of the two or three most complicated words in the English language' (Williams 1983, p. 87). Some confusion derives from intermingling two common meanings of culture: as an analytic tool of social analysis referring to the meaningful semiotic aspects of human action, and as a concrete world of beliefs and practices, ways of doing and thinking that are associated with a particular group or society. In the latter use, referring to the distinctive customs, opinions and practices of a particular group or society, the term is often used in the plural, as in the legal cultures of Japan and China, or in reference to African or Latin cultures. In the former, analytic sense, the word is used in the singular, as in legal culture, or the culture of academia. These distinctions need to be specified clearly when using these terms.

Despite often theoretically abstruse debate, many scholars find the concepts of legal culture and consciousness hard to avoid when they attempt to focus on aspects of law that are not confined to formal legal materials (such as doctrines, statutes, cases, constitutions) or offices and institutions (such as lawyers, judges, courts). Legal culture and consciousness are terms used to emphasize analytically ways in which formal legal institutions and everyday social relations intersect and share cognitive resources.

1. *The Concept of Culture Generally*

The concept of culture as an aspect of social life has been invoked in numerous ways. Referring primarily to learned behavior as distinct from that which is given by nature, or biology, culture has been used to designate everything that is humanly produced (habits, beliefs, arts, and artifacts) and passed from one generation to another. In this formulation, culture is distinguished from nature, and distinguishes one society from another. A narrower conception of culture refers to a particular set of social institutions that is devoted specifically to the production of signs and meanings. In this usage, cultural institutions include, for example, art, music, theater, fashion, literature, religion, media, and education. While the first definition is overly broad, including just about all of human life, the second is too specific: the meanings produced and circulating through the other institutions and 'noncultural' spheres of life are ignored or devalued.

Contemporary cultural analyses have moved beyond these extreme conceptions of culture as either everything human or as only that which calls itself culture, to conceive of culture as a system of symbols and meanings and their associated social practices. This formulation of culture treats as inseparable signs and performances, meanings and actions. It highlights by abstracting 'the meaningful aspect of human action out of the flow of concrete interactions … [by disentangling], for the purpose of analysis, the semiotic influences on action from the other sorts of influences—demographic, geographical, biological, technological, economic, and so on—that they are necessarily mixed within any concrete sequence of behavior' (Sewell 1999, p. 44). This attention to linguistic and symbolic aspects of social action rejects any notion that culture is uniform, static, or shared ubiquitously. Moreover, this conception of culture goes beyond a focus on language alone. Drawing from sociological theories of action and practice (Bourdieu 1990), culture is conceived as an arena or field of practical daily interaction, competition, and struggle, deploying repertoires and strategies of action (Swidler 1986). Here, culture is not a coherent, logical and autonomous system of symbols, but a diverse collection of resources that are deployed in the performance of action. Variation and conflict concerning the meaning and use of these symbols and resources is expected. Although many of these cultural resources are discrete, local, and intended for specific purposes, it is possible to observe patterns so that we are able to speak of a culture, or cultural system.

> System and practice are complementary concepts: each presupposes the other … The employment of a symbol can be expected to accomplish a particular goal only because symbols have more or less determinate meanings—meanings specified by their systematically structured relations to other symbols. Hence practice implies a system. But it is equally true that the system has no existence apart from the succession of practices that instantiate, reproduce, or—most interestingly—transform it. Hence a system implies practice. System and practice constitute an indissoluble duality or dialectic: the important theoretical question is thus not whether culture should be conceptualized as practice or as a system of symbols and meanings, but how to conceptualize the articulation of system and practice. (Sewell 1999, p. 47)

2. *The Concept of Legal Culture Introduced and Debated*

Writing in 1975, Lawrence Friedman introduced the concept of 'legal' culture as a means of emphasizing the fact that law was best understood and described as a system, a product of social forces, and itself a conduit of those same forces. Although law is com-

monly understood as 'a set of rules or norms, written or unwritten, about right and wrong behavior, duties and rights,' (1975, p. 2) according to Friedman this conventional notion attributes too much independence and efficacy to the law on the books, effacing the power and predictability of legal practices. To advance a social scientific study of law in action, Friedman adopted the model of a system—a set of structures that processes inputs (demands and resources) from an environment to which it sends its outputs (functions) in an ongoing recursive feedback loop (see *Luhmann, Niklas* (1927–98)). He identified three central components of the legal system: (a) the social and legal forces that, in some way, press in and make 'the law'; (b) 'the law' itself—structures and rules; and (c) the impact of law on behavior in the outside world. 'Where the law comes from and what it accomplishes—the first and third terms—are essentially the *social* study of law' (Friedman 1975, p. 3).

Although the social scientific study of law was, with minor but notable exceptions, marginalized in American universities—law schools and social science departments—Friedman was working in a tradition that had strong European roots. In jurisprudence, for example, Friedrich Karl von Savigny (1831), a nineteenth-century German statesman and jurist, had described law as one of the most important expressions of the 'spirit of a people' (*Volksgeist*), a continuous thread in an evolving culture. Oliver Wendell Holmes (1881), the American judge and jurist, wrote from a very different, distinctly realist and pragmatic rather than romantic perspective, but nonetheless expressed a similar conviction that the law was a great anthropological document. Such notions of the cultural character and significance of law challenged positivist jurisprudence that defined law as an autonomous system of officially sanctioned and logically derived rules and procedures (e.g., see *Legal Scholarship*). Classical works of European sociology also placed law at the center of social life rather than at the margins. Both Durkheim and Weber offered analyses of law both as an expression of broader social forces in the transformations toward modernity, and as a channel for developing social sensibilities, interests, and actions (e.g., see *Durkheim, Emile (1858–1917)*; *Weber, Max (1864–1920)*).

Friedman chose the phrase legal culture to name the 'social forces … constantly at work on the law,' 'those parts of general culture—customs, opinions, ways of doing and thinking—that bend social forces toward or away from the law' (1975, p. 15). As an analytic term, legal culture emphasized the role of taken for granted and familiar actions that operated on and within the interactions of the legal system and its environment. As a descriptive term, it identified a number of related phenomena: public knowledge of and attitudes toward the legal system, as well as patterns of behavior with respect to the legal system. These included judgments about the law's fairness, legitimacy, and utility. To the extent that patterns of attitudes and behaviors are discernible within a population and vary from one group or state to another, it was possible, he said, to speak of the legal culture(s) of groups, organizations, or states (Friedman 1975, p. 194). As an example of variations within legal cultures, Friedman distinguished the internal legal culture of professionals working *in* the system from the external legal culture of citizens interacting *with* the system. As the 'ideas, values, expectations, and attitudes toward law and legal institutions, which some public or some part of the public holds,' legal culture was meant to name a range of phenomena that would be, in principle, measurable (Friedman 1997, p. 34). Although never theorized elaborately and reformulated several times over the years as a general concept more than a set of quantifiable indicators, the concept of legal culture was useful as a way of 'lining up a range of phenomena into one very general category' (1997, p. 33).

Following its introduction, researchers began using the concept of legal culture in a range of empirical projects including studies of: children's knowledge of and attitudes to law (Tapp and Levine 1974), rights consciousness among Americans (Scheingold 1974), the practices of criminal courts (Eisenstein et al. 1988; but see Kritzer and Zemans 1993), and comparative analysis of different groups and nations (Kidder and Hostetler 1990, Hamilton and Sanders 1992, Sanders and Hamilton 1992, Bierbrauer 1994, Chanock 2001). For those who attempt to measure variations in legal culture, the indicators include such diverse phenomena as litigation rates and institutional infrastructures (Blankenburg 1994, 1997); crime rates (Miyazawa 1987), or assumptions about who should use the legal system and when (Ferrarese 1997).

Predictably, debates have arisen among researchers who have attempted to use the concept in empirical projects (Nelken 1997). The most persistent divide seems to be among those who use culture as an analytic concept within a more developed theory of social relations, and those who view legal culture as concrete, measurable phenomena. Those who attempted to use the concept as a focus for comparative research moved quickly toward measurement and a more limited concept. For some of these researchers, when the concept of legal culture is used with insufficient specificity, the distinction between all of culture and legal culture is unclear, and what constitutes the legal seems too often assumed and undefined (Blankenburg 1997). Some researchers insist that legal culture is that which is produced and studied most effectively among professional legal actors, while others insist that such a narrow definition belies the theoretical utility of the concept of legal culture as a way of marking the inescapable connection between law and everyday life, the feedback loop that Friedman posited in his notion of a legal system.

Cotterrell has produced one of the most sustained critiques of the concept (1997). He insists that 'every-

thing about law's institutions and conceptual character needs to be understood in relation to the social conditions which have given rise to it. In this sense law is indeed an expression of culture.' Nonetheless, Cotterrell is unwilling to accept a concept of legal culture that makes it indistinguishable from other forms of social control or normative ordering; instead, he seeks a middle ground that recognizes the cultural influences on and from law but yet retains a recognition of the distinctiveness of legal forms and doctrines.

To some extent, the sociolegal debates reproduce the controversies among anthropologists and sociologists concerning the concept of culture generally. The most important issues are less empirical than theoretical. The measurement problems derive from the theoretical arguments. How is legal culture evident and measurable, and yet diffuse and abstract; what is the relative importance of causal explanation as against description and interpretive understanding in producing a social scientific study of law; how central is formal legal doctrine in understanding participation in, support for, and consequences of law? Contemporary research on legal consciousness addresses these questions.

3. *Legal Consciousness*

If research on legal culture focuses attention on the myriad ways in which law exists within society generally, the study of legal consciousness traces the ways in which law is experienced and interpreted by specific individuals as they engage, avoid, or resist the law and legal meanings. This empirical attention to popular understandings of law reformulates some of the theoretical debates expressed in the study of legal culture.

3.1 *Legal Consciousness as Attitude*

Some scholars conceptualize consciousness as the ideas and attitudes of individuals that, when taken together, determine the form and texture of social life. An expression of the liberal tradition in political and legal theory, this conception of consciousness suggests that social groups of all sizes and types (e.g., families, peer groups, corporations, communities, nation states and societies) emerge out of the aggregated actions of individuals. In this individualist conception, consciousness consists of both reason and desire. According to liberal ideology, inevitable human variation will always assure that humans will desire different things, even if they reason similarly.

Relying on this attitudinal conception of legal consciousness, researchers have attempted to document the variation in beliefs, attitudes, and actions of individuals as a means of explaining the shape of legal institutions and practices. Some assess the degree to which the rule of law is part of the socialization and habits of population (Stalans and Kinsey 1994). Others, Lind and Tyler (1988) for example, document Americans' attachment to ideas of fairness and due process, describing a deep normative consensus on what they call procedural justice (cf. Thibault and Walker 1975, Tyler 1990). The research suggests that Americans commonly evaluate their legal experiences in terms of the processes and forms of interaction rather than the outcomes of those interactions (see *Compliance and Obedience*: *Legal*). In contrast, a study of legal attitudes among Europeans observes considerable variability in citizens' commitments to, and alienation from, law and legal values (Gibson and Caldeira 1997). Studies in states transitioning from authoritarian rule to democracy also report notable intra-nation as well as international variation (Kourilsky-Augeven 1997).

3.2 *Legal Consciousness as Epiphenomenon*

In other formulations, some scholars regard consciousness as a by-product of the operations of social structures rather than the formative agent in shaping institutions and history. Structural anthropology, for instance, describes social actors located in complex webs of patterned social relationships that determine their perceptions and actions. Similarly, some Marxist structuralism treats ideas, including cultural symbols and narratives, as a superstructural residue of material conditions that serve the interests of elites. Following this perspective, legal consciousness is an epiphenomenon, because a particular social and economic structure is understood to produce a corresponding legal order with similarly constituted legal subjects. Individuals seem to lack agency, portrayed as the bearers rather than authors of social relations. Work in this tradition often describes how the needs of capitalist production and reproduction mold legal behavior and consciousness. Studies focus on the production and practice of law, its accommodation to class interests, and the inequities that result.

Some research from this structuralist perspective complicates the proposed relationship between law and institutional structures by suggesting that the legal order develops in response to conflicts and inconsistencies generated by the capitalist mode of production rather than as a direct instrument of particular class interests. 'To legitimize the inconsistencies and irrationalities born of the contradictions of the economy, the legal order constitutes myths, creates institutions of repression, and tries to harmonize exploitation with freedom, expropriation with choice, inherently unequal contractual agreements with an ideology of free will' (Chambliss and Seidman 1982, p. 70). By focusing on the legitimating functions of law, research describes the ways in which law helps people see their worlds, private and public, as both natural and right.

Balbus, for example, argues that certain features of liberal law, such as the highly prized claims to formal equality, what others describe as procedural justice (see *Compliance and Obedience*: *Legal*, buttress and legitimate the inequality of the existing economic order. Due process and formal equality help convince 'the "propertyless" that they have the legal right and, hence, the real opportunity of rising into the bourgeoisie' (Balbus 1973, p. 6). Some of this literature has focused on 'false consciousness,' or the inability of subjects, especially members of the working class, to perceive their true interests or to recognize opposing interests.

Other structuralist and Marxist theorists, however, do not treat cultural symbols as superstructural residues. Althusser (1971), for example, suggests that cultural symbols, or ideologies, are themselves material practices and relatively autonomous; hence culture and ideology are more than just residues. In this sense, the distinction between structure and consciousness is overdrawn. Rather than by-products of material practices, culture and consciousness participate actively in the production of material practices and social realities (see *Structuralism*; *Marxist Social Thought*, *History of*).

4. *Constitutive Theories of Legal Culture and Consciousness*

Constitutive analyses work to resolve these debates concerning causality, determinism, structure, and agency in studies of culture and legal consciousness. Research from a constitutive perspective emphasizes the roles of consciousness and cultural practice as communicating factors between individual agency and social structure rather than expressions of one or the other. Defining legal consciousness as participation in the processes of social construction, research has focused on the ways in which local, concrete action accumulates into systemic institutions and structures (Ewick and Silbey 1998).

Working within a Gramscian framework (e.g., see *Gramsci*, *Antonio* (1891–1937)), analysts try to describe how the taken-for-granted aspects of social relations, including the legal aspects, are produced, enacted, and reproduced. They document situations in which local processes reproduce macro social structures and institutions recursively, and at the same time provide openings for creativity in reshaping those structures. In particular these constitutive/cultural analyses of legal consciousness describe the processes by which law contributes to the articulation of meanings and values in daily life. Attention is directed to the local contests to create controlling meanings from competing discourses within most aspects of ordinary life including families (Merry 1990, Yngvesson 1993), religious communities (Greenhouse 1986), medical care (Heimer and Stassen 1998, Heimer 1999), engineering (Espeland 1998), gender, race or age discrimination (Bumiller 1988), management of handicapped identities (Munger and Engel 1996) or poverty and public welfare (Sarat 1990, Munger 1999) (e.g., see *Law and Everyday Life*). In these analyses, researchers observe both the orchestration of the local contest and the systematic outcome, in this way mediating micro and macro perspectives. Within this framework, consciousness is understood to be part of a reciprocal process in which the meanings given by individuals to their world become patterned, stabilized and objectified. These meanings, once institutionalized, become part of the material and discursive systems that limit and constrain future meaning-making. In this constitutive conception of legal consciousness, law does more than reflect or encode what is otherwise normatively constructed. Law enables as well as constrains the possibilities of social interaction.

In one study, Conley and O'Barr observed litigants in small claims courts to determine how ordinary 'people identify and analyze legal problems, how they decide when and in what form to bring a problem to the legal system, and how they respond to the demands that the system makes of them' (1990, p. ix). The authors observed two distinct discursive styles. One style is rule oriented, adopted primarily by experienced players who see law as a system for assessing and distributing responsibility; this style is valued by the judges and court personnel. The second style focuses on and emphasizes relationships, sequences of events, and the temporal unfolding of a narrative; this style is preferred by minority and group litigants as well as women and poorer citizens. Although having one's day in court is an American ritual, highly prized as an essential human right, many citizens end up frustrated, the researchers claim, because legal authorities prefer and often demand something other than a heart-rending narration of pain or an impassioned moral claim (Merry 1990).

In a recent study of the legal consciousness of ordinary Americans, Ewick and Silbey (1998) construct three accounts of law that encompass the range of cultural materials with which people produce and experience legality as a structure of social action. The stories incorporate alternative normative bases for legal claims to authority, varying constraints that define action as legal, varying sources for legal capacities, as well as varying temporal and spatial locations for law. In one account, the law is remote, impartial and objective, something to be invoked for solemn and collective purposes that transcend the messiness and partiality of individual lives. The law itself resides in times and spaces separate from everyday places and, while enacted by legal functionaries, exists apart from the words or deeds of particular persons. In a second account, legality is understood to be a game of skill, resource, and negotiation, where persons can seek their own interests

legitimately against others. Law in this rendering appears as a defined arena for strategic interactions, sometimes engaged playfully and sometimes with deadly seriousness, but always simultaneously alongside and within everyday life. A third story describes the law as an arbitrary power against which people feel virtually incapacitated. Often the only means of deflecting the legal power is to employ various subterfuges and evasions. These minor forms of resistance typically leave the law unchallenged and unchanged. Their employment and effectiveness is premised, however, upon a potentially subversive recognition of the structure and organization of legality in everyday life.

Ewick and Silbey demonstrate the connection between the micro phenomenon of legal consciousness and the macro institutions of law by showing how the multiple forms of legal consciousness expressed in the three stories of law exist simultaneously, and together sustain legality as a durable structure of social action. Although the dual depictions of law as god-like (remote, transcendent, objective and magisterial) and game-like (legitimate, rule bound, and resource dependent) seem to challenge one another, they are actually complementary. Although each view emphasizes different normative values and provides a different account of the social organization of law, as an ensemble they cover the range of conventional experiences of legality. Any particular experience can find expression within the heterogeneity of the whole. The law is neither rendered irrelevant for everyday life (by virtue of being remote) nor subsumed by it (by virtue of being familiar).

These constitutive studies reject the dualisms in recurrent debates about the relative roles of structure and agency in shaping the world and focus instead on the role of consciousness (participation in social interaction) in (re)producing the social world (e.g., see *Law as Constitutive*). Moreover, these constitutive analyses of culture and consciousness specify how myriad discrete actions accumulate into institutional forms through processes of reification in which the contingent human production of social life is effaced by attending to its continuous, durable features. This attention to systematic aspects make the social interactions and practices appear to be independent of just that human action of which it is composed. Thus, studies of legal culture and consciousness try to emphasize the daily, locally repeated enactment or practice of social structure. Some cultural accounts, however, go so far as to reject explicitly any distinction between structure and agency, insisting that the analytic distinction itself contributes to the ongoing processes of reification constituting social reality. Finally, in these constitutive studies of legal culture and consciousness, law is understood to be part of a complex totality in which 'it constitutes as well is constituted, shapes as well as is shaped' (Kairys 1982, p. 5).

See also: Civil Liberties and Human Rights; Cultural Psychology; Disputes, Social Construction and Transformation of; Human Rights, History of; Law and Society: Sociolegal Studies; Law: Anthropological Aspects; Law: History of its Relation to the Social Sciences; Law, Sociology of; Legal Process and Social Science: United States; Lifelong Learning and its Support with New Media: Cultural Concerns; Norms

Bibliography

Althusser L 1971 Ideology and ideological state apparatuses. In: Althusser L (ed.) *Lenin and Philosophy*. New Left Books, London

Balbus I 1973 *The Dialectics of Legal Repression: Black Rebels Before the American Criminal Courts*. Transaction Books, New Brunswick, NJ

Bierbrauer G 1994 Toward an understanding of legal culture: Variations in individualism and collectivism between Kurds, Lebanese, and Germans. *Law and Society Review* **2**(2): 243–64

Blankenburg E 1994 The infrastructure of legal culture in Holland and West Germany. *Law and Society Review* **28**(4): 789–809

Blankenburg E 1997 Civil litigation rates as indicators for legal cultures. In: Nelken D (ed.) *Comparing Legal Cultures*. Dartmouth, Brookfield, VT, pp. 41–68

Bourdieu P 1990 *The Logic of Practice*, trans. Nice R. Stanford University Press, Stanford, CA

Bumiller K 1988 *The Civil Rights Society: The Social Construction of Victims*. Johns Hopkins University Press, Baltimore, MD

Chambliss W, Seidman R 1982 *Law, Order, and Power*. Addison-Wesley, Reading, MA

Chanock M 2001 *The Making of South African Legal Culture 1902–1936: Fear, Favour and Prejudice*. Cambridge University Press, Cambridge, UK

Conley J M, O'Barr W M 1990 *Rules Versus Relationships: The Ethnography of Legal Discourse*. University of Chicago Press, Chicago, IL

Cotterrell R 1997 The concept of legal culture. In: Nelken D (ed.) *Comparing Legal Cultures*. Dartmouth, Brookfield, VT, pp. 13–32

Durkheim E 1933 [1902] *The Division of Labor in Society*. Free Press, Glencoe, IL

Eisenstein J, Flemming R B, Nardulli P F 1988 *The Contours of Justice: Communities and Their Courts*. Little, Brown, Boston, MA

Espeland W N 1998 *The Struggle for Water: Rationality, and Identity in the American South West*. University of Chicago Press, Chicago, IL

Ewick P, Silbey S 1998 *The Common Place of Law: Stories From Everyday Life*. University of Chicago Press, Chicago, IL

Ferrarese M R 1997 An entrepreneurial conception of the law? The American model through Italian eyes. In: Nelken D (ed.) *Comparing Legal Cultures*. Dartmouth, Brookfield, VT, pp. 157–82

Friedman L M 1975 *The Legal System: A Social Science Perspective*. Russell Sage Foundation, New York

Friedman L M 1997 The concept of legal culture: A reply. In: Nelken D (ed.) *Comparing Legal Cultures*. Dartmouth, Brookfield, VT, pp. 33–40

Gibson J L, Caldeira G A 1997 The legal cultures of Europe. *Law and Society Review* **30**(1): 55–85

Greenhouse C J 1986 *Praying for Justice: Faith, Order, and Community in an American Town*. Cornell University Press, Ithaca, NY
Hamilton V L, Sanders J 1992 *Everyday Justice: Responsibility and the Individual in Japan and the United States*. Yale University Press, New Haven, CT
Heimer C 1999 Competing institutions: Law, medicine and family in neo-natal intensive care. *Law and Society Review* **33**(1): 17–66
Heimer C, Stassen L 1998 *For the Sake of the Children*. University of Chicago Press, Chicago, IL
Holmes O W 1881 *The Common Law*. Little, Brown, Boston, MA
Kairys D (ed.) 1982[1990] *The Politics of Law*. Pantheon Books, New York
Kidder R L, Hostetler J A 1990 Managing ideologies: Harmony as ideology in Amish and Japanese societies. *Law and Society Review* **24**: 895–922
Kourilsky-Augeven C (ed.) 1997 *Socialisation Juridique et Conscience du Droit: Attitudes Individuelles, Modeles Culturels et Changement Social*. Reseau Europeen Droit et Société 2
Kritzer H M, Zemans F K 1993 Local legal culture and the control of litigation. *Law and Society Review* **27**(3): 535–7
Lind A E, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum Press, New York
Merry S E 1990 *Getting Justice and Getting Even: Legal Consciousness Among Working-class Americans*. University of Chicago Press, Chicago, IL
Miyazawa S 1987 Taking Kawashima seriously: A review of Japanese research on Japanese legal consciousness and disputing behavior. *Law and Society Review* **21**: 219–41
Munger F 1999 Immanence and identity: Understanding poverty through law and society research. *Law and Society Review* **32**(4): 931–67
Munger F W, Engel D M 1996 Rights, remembrance, and the reconciliation of difference. *Law and Society Review* **30**(1): 7–53
Nelken D 1997 *Comparing Legal Cultures*. Dartmouth, Brookfield, VT
Sanders J, Hamilton V L 1992 Legal cultures and punishment repertories in Japan, Russia, and the United States. *Law and Society Review* **26**(1): 117–38
Sarat A 1990 'The law is all over': Power, resistance, and the legal consciousness of the welfare poor. *Yale Journal of Law and Humanities* **2**(2): 343–79
Savigny F K von 1831[1975] *Of the Vocation of Our Age for Legislation and Jurisprudence*, trans. Hayward A. Arno Press, New York
Scheingold S A 1974 *The Politics of Rights: Lawyers, Public Policy, and Political Change*. Yale University Press, New Haven, CT
Sewell W H Jr 1999 The concepts of culture. In: Bonnell V E, Hunt A H, Biernacki R (eds.) *Beyond the Cultural Turn: New Directions in the Study of Society and Culture*. University of California Press, Berkeley, CA, pp. 35–61
Stalans, L J, Kinsey K A 1994 Self-presentation and legal socialization in society: Available messages about personal tax audits. *Law and Society Review* **28**(4): 859
Swidler 1986 Culture and action: Symbols and strategies. *American Sociological Review* **51**: 273–86
Tapp J L, Levine F L 1974 Legal socialization: Strategies for an ethical legality. *Stanford Law Review* **27**: 1–72
Thibaut J, Walker L 1975 *Procedural Justice: A Psychological Analysis*. L. Erlbaum Associates, New York
Tyler T R 1990 *Why People Obey the Law*. Yale University Press, New Haven, CT
Weber M 1947 *The Theory of Social and Economic Organization*, trans. Henderson A M, Parsons T. Oxford University Press, New York
Weber M 1954 *On Law in Economy and Society*, trans. Rheinstein M. Simon and Schuster, New York
Williams R 1983[1976] *Keywords: A Vocabulary of Culture and Society*, rev. edn. Fontana, London
Yngvesson B 1993 *Virtuous Citizens, Disruptive Subjects: Order and Complaint in a New England Court*. Routledge, New York

S. S. Silbey

Legal Education

Legal education, in any systematic sense, is a relatively new phenomenon in most jurisdictions. In ancient Greece and Rome there was some formal training in advocacy (moral and legal) supporting the establishment of a professional class of persons with legal expertise, but this class was not a profession. Early forms of legal education were delivered primarily through observation and training acquired by serving as an apprentice to a legal consultant (*jurisprudent*) rather than through formal academic study. For many centuries legal texts were quite rare, which meant that the art of rhetoric inevitably formed the 'core' of early legal education. Gradually, however, as texts on both law and custom became more common and records of court decisions were kept, this dominant oral tradition was supplanted with a new written tradition created by full-time law teachers dedicated to establishing and communicating a literature on, about, and for law. Whether this literature was written to serve the goals of the academy or the professional bodies has never been clear. To the present day, the ultimate goals of legal education have been strongly contested by its various stakeholders.

1. The Establishment of Organized Legal Education

Legal education began by asking universal questions linking legal rules, morals, and justice. Only much later were these elements separated, and some were discarded altogether. The focus on national legal systems correlates with the rise of nation states and legal professions. Early legalists were certainly not parochial in their outlook and did not draw sharp distinctions between law and morality, nor did they fail to consider the relationship with other professions and disciplines. If one draws a distinction between legal *education* and *training*, it could be said that, with the passage of time, the latter has steadily displaced

the former. As a result, the boundaries of legal thought have been shrinking and the legal mind has become more introverted and absorbed with technical ephemeral rules rather than fundamental values.

Roman law, canon law, and international law were the focal points of legal study in the medieval universities on the continent of Europe, and these legal sources formed the basis of a common transnational legal language facilitating trade between commercial classes (see *International Law and Treaties*). The English common law was an exception to this trend where the impact of this triad was greatly reduced by the relatively early development of an indigenous and quite sophisticated legal culture insulated from Roman law. Legal education in medieval England was provided by the Inns of Court, which organized apprenticeships involving both shadowing practitioners and practical exercises within a closed, and almost monastic, community. With the advent of printing, legal texts started to supplant, but not to replace, these informal methods of socialization. Organized legal education for solicitors did not commence until Sir William Blackstone started to give his famous lectures on English law at Oxford in the 1750s. It was not until the Victorian era, with visionaries such as John Austin, Jeremy Bentham, and Sir Henry Maine, that university liberal legal education became established at University College London and beyond, including the Council of Legal Education which was set up in 1852 (see Cocks 1983, 1988). In the United States, legal education is generally seen to have begun with the foundation of Harvard Law School in 1817 (Sutherland 1967, Stevens 1983).

2. *The Aims of Legal Education*

Throughout the world toward the close of the last century, several national reviews were set up to report on the state of legal education (Ormrod 1971, Arthurs 1984, Pearce 1987, MacCrate 1992, Franken 1995 ACLEC 1996). From these national reviews, it is apparent that legal education has increasingly diverse aims serving concurrently both academic and professional goals. The precise mix balancing scholarship and practice varies between, and within, particular jurisdictions. And even within particular institutions there are heated debates about theory, substance, and method. Scholarly approaches have become pluralistic and fall into a number of subcategories: doctrinal ('black letter'); sociolegal; contextual; critical; jurisprudential; interdisciplinary ('law and'); comparative; clinical; and skills-oriented (see Twining 1994). At the most general level almost all would agree that legal education should introduce students to what it means to 'think like a lawyer' (Ormrod 1971, p. 102). Law is an intellectual discipline with distinctive concepts, institutions, and methods of reasoning, and a primary responsibility of all law schools is to create, develop, and transmit this 'core' knowledge. In England and Wales, as elsewhere, the professional bodies take an active interest in defining this core but their involvement is often resented, or resisted, by the academy which claims jurisdiction over the academic stage. The ACLEC Report sought to balance the interests of various stakeholders by concentrating on articulating an 'outcome statement' which specified the broad skills, knowledge, and values law students should possess by the time they graduate without dictating to law teachers precisely how their students should be taught (ACLEC 1996, p. 72).

Interdisciplinary approaches, initially on the periphery of the curriculum, have become increasingly part of the mainstream. Anglo-American legal education has embraced the study of 'law and society,' while law schools in continental Europe have often made non-law subjects, such as legal history, economics, philosophy, or sociology, compulsory for undergraduate law students. In the civil law world, legal education continues to be dominated by the study of codes that can be traced back to Roman law. Codes provide the focal point for formal instruction, which tends to discourage or displace critical reflection. Within the common law systems doctrinal analysis still prevails, although it is now generally accepted that law students should also be exposed to ideas about the context of law.

By the time of their graduation, law students are expected to have laid the foundation for a career which is law-related but not necessarily within the legal profession. Although the majority of law students in England and Wales still expect to qualify as either a solicitor or barrister, this is not true in continental Europe, the United States, or Japan, where the emphasis is on producing lawyers who will disperse into a range of occupations in the public and private sectors. Increasingly it is being recognized that law schools not only prepare students for a career in law but also enable people who can handle legal sources to cope with change in a wide variety of working environments, which could be described as the legal services industry. This industry is being transformed by globalization, new technology, and interdisciplinary tendencies common to all modern professions. Legal education continues to teach students how to 'think like a lawyer,' but today this thinking requires much more than a passive understanding of primary sources, principles, and concepts. It must include a critical awareness of what is happening outside the world of law. Because so much law has a built-in obsolescence, it is vital that law students possess the skills and confidence to interpret new rules and situations. University legal education must therefore provide a solid foundation for the education taking place at the vocational stage and during continuing professional development, as well as a framework for understanding the uncertainties which lie ahead.

3. Trends in Legal Education

In demographic terms, university legal education has been steadily expanding in most jurisdictions since World War II. This has created a mismatch between the supply of law graduates and their demand from prospective employers, at least from within the legal profession. Today most lawyers are middle class university graduates, though they are no longer necessarily law school graduates, or male, or Caucasian. Methods of delivering legal education are constantly evolving, but a university degree which acts as a qualifying examination recognized by the profession has replaced apprenticeship, which was for centuries the normal entry route to a legal career. Women have become increasingly visible in law schools and the profession, and their presence is challenging traditional male-dominated attitudes toward both the law and legal work (see *Feminist Legal Theory*). Moreover, it is possible to enter the profession having studied another degree, and indeed some law firms express a clear preference for those with non-law or mixed degrees (Economides 1999, p. 400).

The professions have always sought to control entry, partly to maintain standards and to protect what they see as a finite amount of legal work that can be shared (Abel 1999). On the other hand, because universities (and to some extent government policy) decide who enters law school—and in most countries these have been expanding in both size and number—law students continue to flood the market. In Japan, which produces 40,000 law graduates a year, recent reforms have increased the total number of places on the professional training course to still only around 1,000 places. What some may describe as the overproduction of law graduates who fail to gain full professional qualifications has also brought about new forms of legal service delivery through so-called 'para-legals.'

The increase in the numbers of those entering law school has not always been matched by a corresponding increase in resources, resulting in adverse staff–student ratios or an imbalance in the social and ethnic composition of the law student population. In order to cope with this expansion in numbers, new teaching methods have evolved, whereby students participate in more independent, active learning, and teachers are under pressure to explore the potential of computer-assisted learning (CAL). Distance-learning in legal education remains uncommon, though is not unknown, and could have significant implications not only for citizens, access to legal education but also for citizen access to legal services more generally. Progress with the notion of community legal education (Cooper and Trubek 1997, p. 233) could be accelerated with increasing access to the Internet and the development of expert systems in law. However, these developments threaten the power base of the legal profession and its exclusive control over legal knowledge and expertise (Susskind 1998).

Law schools tend to deploy a variety of methods to deliver teaching and these are determined by national traditions and available resources. Perhaps as a consequence of increased communication and awareness of different approaches, there is convergence resulting from the exchange of ideas and experimentation. The casebook approach, which was pioneered by Christopher Columbus Langdell at Harvard over a century ago, and the Socratic method in which students are forced to explain the logic of legal argument through a question and answer session conducted in a large classroom, have been exported to other common law countries. Most undergraduate law teaching is delivered through a combination of formal lectures (normally optional) backed up by tutorials, workshops, and/or seminars. Postgraduate legal studies is more library and seminar based. Teaching legal skills has now become an important part of the informal law curriculum in Anglo-American law schools, which have moved beyond mooting (mock trails) to include client-interviewing and clinical approaches sometimes involving live clients. However, the fact that European Union law is now part of the mainstream of the curriculum means that both civilian and common law methodologies are understood by law students throughout Europe (East and West) and European law students can observe the synthesis of these hitherto contrasting legal methodologies in the courts of Luxembourg and Strasbourg.

A further recent trend in legal education common to many modern law schools concerns the revival of interest in the ethics and morality of law. In the United States, the Watergate scandal led to compulsory courses in professional responsibility. In many continental legal systems, deontology is taught, but usually in a highly formal, uncritical manner which does not depart far from the code of professional conduct. For most of the past two centuries positivism has prevailed in the teaching of law, with law students being taught to focus almost exclusively on studying legal rules divorced from their moral, political, or social context. Contextual and critical tendencies in legal scholarship have challenged legal formalism since the 1970s, but most recently there has been an attempt to (re)introduce into the undergraduate curriculum ethical perspectives in order to enrich and deepen the education of law students (Economides 1998). By exposing law students to ethics and values, educators seek to promote a more responsible and less cynical approach to both legal study and practice. Evidence from several jurisdictions in both the common law and civilian legal systems suggests that the modern law school inculcates egotistical cynicism rather than an altruistic sense of public duty or any strong commitment to justice (Cooper and Trubek 1997, 1999, Economides 1998, 1999, Junqueira 1999, Erlanger et al. 1996, Granfield 1992). The need to introduce the

teaching of legal ethics arises not only at the initial stage of legal education but is also necessary at other stages of the educational continuum (see *Legal Positivism*; *Legal Formalism*; *Law in Latin America*).

4. The Educational Continuum

Legal education should be seen as part of lifelong learning and need not be confined to what happens inside law schools or preparation for legal practice. The fact that law schools are expected to deliver lawyers ready for practice often results in narrow, cramped, and overcrowded curricula with an unhealthy obsession with examination grades. This obsession with grades (which determine entry and status within the job market) arguably undermines the educational experience and substitutes competitiveness for collegiality. It also inhibits intellectual freedom by forcing the law student to memorize rules rather than explore ideas, with the result that legal education tends toward 'training for hierarchy' (Kennedy 1982). While the education of lawyers clearly falls into distinct phases, it is perhaps preferable to regard the acquisition of legal skills and knowledge as part of ongoing processes, which are continually refined throughout a legal career, and in this way to reduce some of the pressure and expectations invested in particular rites of passage such as graduation and professional qualification. Ideally each stage should prepare the student for what follows next without being too prescriptive about the content or methods deployed at any particular stage.

4.1 The Initial Stage

Legal education normally commences with an undergraduate degree, though in some countries law may be taught in secondary schools along with instruction in civics or citizenship. In the United States most law schools require graduate entry and this is also the case in India and Pakistan. The undergraduate law degree—in which the foundations of legal knowledge are laid—is thus the normal initial or academic stage of legal education, though most countries permit a measure of flexibility with regard to entry to (and exit from) professional qualification. Postgraduate degrees serve to prepare students for specialist areas of legal work or a career in academic law; students may follow a Masters course or read for a higher degree leading to an MPhil or PhD.

In the Anglo-American system, and in many other countries in the common law world, students are expected to pass examinations in four or five subjects at the end of each year. It is also increasingly common for law schools themselves to be ranked according to their research and teaching quality; such rankings are organized and publicized by government or leading newspapers. The US *News and World Report*, for example, ranks law schools in the United States; in the United Kingdom, the Higher Education Funding Council (HEFCE) administers the Research Assessment Exercise (RAE). Examinations also feature in many continental European countries, but here the university law degree invariably takes much longer; in Germany, the norm is five years (though this is currently under review) and concludes with a single comprehensive examination (the First State Examination) at the end of those five years. Other European countries combine the system of yearly examinations in each course with a comprehensive examination at the end of the five-year period of study (Lonbay, 1990). The examination system has been accused of producing compartmentalized thinking, in which students are unable to think across subject boundaries; for this reason, it has become common to introduce general topics such as Public Law and Private Law which are spread over two years. Apart from the written examination, most law schools offer students the opportunity to write a long essay or dissertation, sometimes in lieu of a written examination; and continental law schools in Italy and France make use of oral examinations and a thesis to be written in the students' final year. In English law schools, the oral examination is usually confined to resolving problematic or borderline results, but increasing use is made of various forms of continual assessment. Students can sometimes also gain credit for working on law reviews, but this is quite rare outside the United States.

4.2 The Vocational Stage

In England and Wales lawyers qualify as barristers or solicitors by taking either the Bar Vocational Course (BVC) or the Legal Practice Course (LPC) (see Economides and Smallcombe 1991). Both courses have introduced a new skills element and can now be taken at various franchised institutions including some university law schools. Following the Courts and Legal Services Act 1990, which blurred the distinction between barristers and solicitors by allowing certain of the latter to gain higher courts rights of audience, the arguments in favor of common training for both branches have become much stronger (see ACLEC 1996, Chap. 5). Currently, barristers' training is focused on the four Inns of Court (Lincoln's Inn, Inner Temple, Middle Temple, Gray's Inn) and falls under the jurisdiction of the Senate of the Inns of Court and the Council on Legal Education, while solicitors' training is governed by the Law Society. University graduates may obtain partial or total exemption from the first part, depending on their degree, provided it satisfies the Joint Announcement on Qualifying Law Degrees. To qualify as a barrister, one must pass the BVC and serve a period of

apprenticeship (pupilage) which involves six months in chambers attached to a barrister. The pupil barrister cannot practice unsupervised until s/he has been a pupil for one year. A major obstacle for students coming from poorer backgrounds has been finding a place in chambers, but the profession has initiated reforms to assist the fair distribution of the limited training places which become available. To qualify as a solicitor, one must pass the LPC and get a training contract in which one serves an apprenticeship as a trainee solicitor, normally attached to a firm of solicitors, for two years. Scotland and Ireland have similar arrangements, though both jurisdictions have progressed farther with the notion of common training. Judicial training, which usually follows a successful career at the Bar, is now overseen by the Judicial Studies Board, while employed lawyers, for example in the Crown Prosecution Service, are subject to further training in-house. Vocational legal education in recent years has been dominated by the large city firms which have invested heavily in in-house training (and toyed with the idea of setting up their own vocational course provider), while many small firms in the provinces and rural areas, even though willing, have been unable to afford to take on trainee solicitors. Common training exists in the United States in that passing the state bar examination (which do vary) both grants admission to the bar and qualifies one for all categories of legal work.

In civil law countries, specialization takes place following graduation. In Japan, Spain, and France there are special national schools to train judges, entry to which is highly competitive and based on performance in an examination. In Germany and Scandinavia law graduates intending to pursue a judicial career also take further vocational examinations and are then attached to a court while serving a probationary period, at the end of which tenure for life is granted. Those wishing to qualify as lawyers follow training geared to particular forms of legal practice. In France, the choice is between qualifying as an advocate, an *avoué*, or a notary. In Germany, law graduates wishing to practice follow a program of practical training as a *Referendar* in diverse legal settings across the public and private sectors at the end of which they must pass the state bar examination (*Assessorexamen*) (see further Lonbay 1990).

4.3 Continuing Professional Development

Most modern professions prefer lawyers to follow practical courses post-qualification in order to refresh their knowledge of law and keep abreast of new developments, but few have gone so far as to make continuing professional development (CPD) compulsory. In the United States CPD is mandatory in only 16 states; in Australia, only New South Wales requires general mandatory CPD (Nelson 1993). There is no CPD in Austria, Belgium, France, the Nordic countries, Germany, Greece, Ireland, Italy, or Luxembourg, although a compulsory CPD regime has been operative in the Netherlands since 1996 for all Dutch lawyers (*advocaten*). Spain and Portugal have been reviewing CPD and increasing the provision of voluntary courses. In England and Wales both branches of the profession now require CPD (see ACLEC 1997).

See also: Expert Witness and the Legal System: Psychological Aspects; Lawyers, Regulation of; Legal Professionalism; Professionalization/Professions in History; Professions in Organizations; Professions, Sociology of

Bibliography

Abel R 1999 The politics of professionalism: The transformation of English lawyers at the end of the twentieth century. *Legal Ethics* **2**(2): 131–47

Advisory Committee on Legal Education and Conduct (ACLEC) 1996 *First Report on Legal Education and Training*. ACLEC, London

ACLEC 1997 *Continuing Professional Development for Solicitors and Barristers: A Second Report on Legal Education and Training*. ACLEC, London

Arthurs H 1984 *Law and Learning. Report to the Social Sciences and Humanities Research Council of Canada by the Consultative Group on Research and Education in Law*. Social Sciences and Humanities Research Council, Ottawa, Canada

Cocks R 1983 *Foundations of the Modern Bar*. Sweet & Maxwell, London

Cocks R 1988 *Sir Henry Maine. A Study in Victorian Jurisprudence*. Cambridge University Press, Cambridge, UK

Cooper J, Trubek L (eds.) 1997 *Educating for Justice: Social Values and Legal Education*. Dartmouth, UK

Cooper J, Trubek L (eds.) 1999 *Educating for Justice: Legal Education, Legal Practice and the Community*. Dartmouth, UK

Economides K (ed.) 1998 *Ethical Challenges to Legal Education and Conduct*. Hart, Oxford, UK

Economides K 1999 Learning the law of lawyering. In: Freeman M D A (ed.) *Current Legal Problems 1999, Vol. 52*. Oxford University Press, Oxford, UK, pp. 392–418

Economides K, Smallcombe J 1991 *Preparatory Skills Training for Trainee Solicitors*. Law Society Research Study No. 7. The Law Society, London

Erlanger H, Epp C, Cahill M, Haines K 1996 Law student idealism and job choice: Some new data on an old question. *Law & Society Review* **30**(4): 851–64

Franken H 1995 *Een eigen richting voor hel recht … Eindrapport van de Verkenningscommissie Rechtsgeleerdheid* [*A sense of direction … The future of legal research in the Netherlands*. Final Report of the Legal Education Review Commission]. Leiden, The Netherlands

Granfield R 1992 *Making Elite Lawyers: Visions of Law at Harvard and Beyond*. Routledge, New York

Junqueira E B 1999 *Faculdades de Direito ou Fábricas de Ilusões?* [*Faculties of Law or Factories of Illusion?*]. Letra Capital/IDES, Rio de Janeiro, Brazil

Kennedy D 1982 Legal education and the reproduction of hierarchy. *Journal of Legal Education* **32**: 591

Lonbay J 1990 *Training Lawyers in the European Community*. The Law Society, London
MacCrate R 1992 *Legal Education and Professional Development: An Educational Continuum*. Report of the Task Force on Law Schools and Profession: Narrowing the Gap. American Bar Association, Chicago
Nelson J W 1993 *A Study of the Continuing Legal Education Needs of Beginning Solicitors*. Centre for Legal Education, Sydney, Australia
Ormrod Report 1971 *Report of the Committee on Legal Education*. HMSO Cmnd. 4595, London
Pearce D 1987 *Australian Law Schools: A Discipline Assessment for the Commonwealth Tertiary Commission*. Australian Publishing Service, Canberra, Australia
Stevens R 1983 *Law School: Education in America from the 1850s to the 1980s*. University of North Carolina Press, Chapel Hill, NC
Susskind R 1998 *The Future of Law. Facing the Challenges of Information Technology*. Oxford University Press, Oxford, UK
Sutherland A E 1967 *The Law at Harvard. A History of Ideas and Men 1817–1967*. Belknap Press, Cambridge, MA
Twining W 1994 *Blackstone's Tower: The English Law School*. Sweet & Maxwell, London

K. Economides

Legal Formalism

'Legal formalism' is an important category in the history of law, the sociology of law, comparative law, and the cultural study of law, as well as in the philosophy of law and the interdisciplinary field currently called 'legal theory.' It is used in different senses in these different fields, and within each field it is a contested concept, rather than a well-established term with a clear meaning. This entry presents a catalogue of different usages and a brief introduction to the modes of contestation of the meaning of the term.

The modern usages of the word derive from the work of leading legal theorists of the late nineteenth and early twentieth centuries who were much concerned with two historical phenomena that play little role in the late twentieth-century discussion. One of these was 'primitive formalism,' meaning the practice of deciding disputes through devices such as oracles and trial by battle, regarded as 'irrational.' The other was the ancient Roman and medieval English system of 'formulary justice' or 'strict law,' in which a claimant could get redress through the legal system only by fitting his case into a closed class of 'actions.' No overarching principles were available, at least according to the theory, to deal with cases that fell outside the class, but within generally held ideas of moral responsibility. Modern law, in the nineteenth-century view, was characterized by its movement beyond both primitive formalism and formulary justice, but had to find a way to preserve some of the virtues of these earlier systems. (Maine 1917, von Jhering 1869, Holmes 1881, Pollock and Maitland 1898, Weber 1954).

1. *Formalism as a Descriptive Category*

Greater or lesser formalism is one of the dimensions along which we compare legal regimes and assess internal change, whether at the level of detail, of a large ensemble of rules, or of a system as a whole. One might also just say 'more or less formal,' or call it the dimension of 'formality.' In this usage, no positive or negative evaluation is indicated by categorizing a body of legal rules as more or less formal or formalist. A system is procedurally formalist to the extent that it makes the success of a substantive legal claim depend on following procedural rules (Schauer 1988). It is transactionally formalist to the extent that it requires specific formalities for transactions such as contracts or marriages (von Jhering 1869, Demogue 1911, Fuller 1941). It is administratively formalist to the extent that it surrounds the exercise of state power with procedural and transactional formalities. Rule formalism is a general preference for rules over standards.

Two important dimensions of formality are the degree of insistence on compliance with formalities (what exceptions are permitted?), and the degree of absoluteness of the sanction of nullity for failure to comply (what remedies, if any, for a person who fails to comply?). The degree of formality in each dimension varies within systems, between systems, and over time. What binds the types together is the willingness of the formalist to sacrifice substantive justice (or 'equity') in the particular case. Western legal discourse provides a highly developed set of arguments in favor of and against adopting more or less formal rule systems, and a rich literature on the moral and political meaning and practical consequences of the choice. Every argument for greater formality has a symmetrical opposite urging less formality (Kennedy 1976).

A different descriptive use of the term legal formalism refers to a range of techniques of legal interpretation based on the meaning of norms (whether established privately, as in contracts, or publicly, as in statutes), and refusing reference to the norms' purposes, the general policies underlying the legal order, or the extrajuristic preferences of the interpreter. Textual interpretive formalism decides by identifying a valid norm applicable to the case and then applying it by parsing the meanings of the words that compose it. Textual formalism is literalist to the extent that it refuses to vary meaning according to context, and originalist to the extent that it finds meaning only through the context at the time of enactment (Schauer 1988).

Conceptual interpretive formalism 'constructs' general principles thought necessary if the legal system is to be understood as coherent. It uses the principles to resolve uncertainty about the meaning of extant valid norms, and applies the principles according to their meaning to fill apparent gaps (Geny 1899). Precedential interpretive formalism interprets according to the meaning of norms derived as the holdings of prior cases (Grey 1983).

Interpretation positing gaplessness requires the interpreter to apply in every case, according to their meanings, the legal norms he or she can derive textually, conceptually, or through precedent; it categorically forbids reference to purposes and policies (Weber 1954). A final descriptive use of the term formalism in legal discourse refers to theories that purport to derive particular rules of law, or prohibitions on adopting particular rules, from a small group of internally consistent abstract principles and concepts (e.g., corrective justice, fault) understood as morally binding on legal actors (Weinrib 1988).

As with the discourse of the appropriate degree of formality of norm systems, there is a wide range of arguments for and against the adoption of each of these types of interpretive formalism. There is also disagreement as to their conceptual coherence and practicability (Kennedy 1997).

2. *Formalism as a Critical Category*

The critical use of the category of formalism was developed by the sociological jurists around the turn of the twentieth century, in their attack on mainstream late nineteenth-century legal thought. According to the critics, the mainstream saw law as having a strong internal structural coherence based on the two traits of 'individualism' and commitment to legal interpretive formalism. These traits combined in 'the will theory.'

In the sociological jurists' version, the will theory was that the private law rules of the 'advanced' Western nation states were well understood as a set of rational derivations from the notion that government should help individuals realize their wills, restrained only as necessary to permit others to do the same. In its more ambitious versions, the will theory made public as well as private law norms follow from this foundational commitment (for example, by generating theories of the separation of powers from the nature of rights).

The will theory presupposed consensus in favor of the goal of individual self-realization. It was not a political or moral philosophy justifying this goal; nor was it a positive historical or sociological theory about how this had come to be the goal. Rather, the theory offered a specific, will-based, and deductive interpretation of the interrelationship of the dozens or hundreds of relatively concrete norms of the extant national legal orders, and of the legislative and adjudicative institutions that generated and applied the norms (Pound 1917, Kennedy 2000).

The sociological jurists critiqued the individualist premises of the will theory in the name of 'social law' (Gurvitch 1932). They also critiqued its methodology, on two grounds. First, they argued that in practice it involved the widespread abuse of deduction, meaning that jurists habitually offered deductive justifications for interpretations that were in fact logically underdetermined (von Jhering 1877–83, Holmes 1897, Geny 1899).

Second, they argued that the will theory falsely assumed the possibility of constructing the legal order in such a way that it would be gapless in fact, and therefore susceptible to exclusively meaning-based interpretation. The sociological jurists claimed that particular instances of the abuse of deduction, and the theory of meaning-based gaplessness generally, disguised the biases of interpreters, and prevented the consideration of the 'legislative' element in interpretation. They advocated interpretation on the basis of 'scientifically' established social desiderata, as well as or instead of according to meaning (von Jhering 1877–83, Geny 1899, Cardozo 1921).

It is a matter of dispute in legal historiography to what extent late nineteenth-century legal thought was well characterized by its critics, and to what extent the will theory and formal methods caused it to have a politically conservative substantive content (contrast Kennedy 1980, Grey 1983, Horwitz 1992). It is nonetheless clear that sociological jurisprudence dramatically changed mainstream legal academic discourse through its critiques of the abuse of deduction and of the possibility of meaning-based gaplessness. However, structurally analogous conflicts, involving the same elements, preoccupied legal thinkers throughout the twentieth century. These reflect a historical dialectic of critique and reconstruction, in which new positions are only uncannily similar to, not identical with, previous ones.

Whether in the form of Geny's '*libre recherche scientifique*' (Geny 1899) or Cardozo's 'method of sociology' (Cardozo 1921), the sociological jurists' proposed alternative method was rejected during the 1930s by a new legal theoretical avant-garde, including both the American legal realists and Kelsenian neopositivists on the Continent, on the ground that its proponents confused facts and values, scientific and normative judgments. The critics argued that the social purposes or functions the sociological jurists used to base their means/ends rational derivations of legal rules were either vague or conflicted in particular cases, so that the claim to a scientific method was no more than a screen for a new form of politically progressive natural law (Llewellyn 1930, Kelsen 1934). In the US, the progressive academic elite abandoned sociology in favor of two (arguably contradictory) normative projects: in private and administrative law, 'policy analysis,' understood as the process of reconcil-

ing or balancing diverse legal desiderata on the basis of information about the social context (Fuller 1941, Kennedy 2000); and in constitutional law, civil libertarianism.

Critics of the Warren Court's progressive civil libertarian reform of American public law charged the Court with formalism, that is, with an abuse of deduction from personal rights exactly analogous to the earlier conservative abuse of deduction from property and contract rights (Hand 1958). Critical legal scholars critiqued the policy-based post-New Deal technique of private and administrative law as 'social conceptualism' or 'policy formalism,' because it selected policies arbitrarily, underestimated the conflicts among them, and offered no defense of balancing as a rationally determinate procedure. (Klare 1978). The post-1960s reconstruction projects that grounded adjudication either in the efficiency norm or in concepts like autonomy or commutative justice were vulnerable to similar critiques.

Modern legal theory (with the exception of natural law theorists) is antiformalist, in the sense of denying the possibility of strictly meaning based gaplessness. Positivism, whether in its Hart or its Kelsen variant, affirms that discretion in adjudication is inevitable, while limiting it to the 'penumbra' or to the area inside the 'frame' provided by the norm in question (Hart 1994, Kelsen 1934). American legal theory, from Cardozo to Dworkin, is antipositivist, and affirms gaplessness, but on the basis of policy, purpose, rights, or principles, rather than on the basis of textual, conceptual, or precedential formalism (Kennedy 1997). In the debate between H. L. A. Hart and Lon Fuller, Hart charged Fuller with formalism because Fuller believed in a gapless order, and Fuller charged Hart with formalism because Hart believed there were easy cases where norm application was truly meaning-based.

Since World War II, there have been a variety of theories that challenge not meaning-based gaplessness, but all meaning-based interpretation, in so much as it claims to be a matter of legal 'truth.' In the common law world, a first critique of precedential interpretive formalism is that policy argument is always necessary in order to determine the relevance of a precedent for a new case (Brewer 1996). A second is that common law theory, if it is to be coherent, must authorize the creation of an exception to a precedentially established rule in any case where an exception would serve the policies animating the system as a whole (Schauer 1991).

On the Continent, rhetorical and hermeneutic theories of legal interpretation assert that outcomes are always relative to horizons, no matter how superficially deductive (and then reground interpretation in the supposedly shared horizon of liberal faith) (Herget 1996). Postmodern theorists take it for granted that there are critiques of meaning-based and policy-based legal reasoning sufficiently conclusive that the interesting questions concern the interpretation of the subjective experience of legal certainty and the juristic status of concepts like 'justice' (Goodrich and Carlson 1998). Skeptical theories gain support from comparative law scholarship showing diametrically opposed interpretations of identical code provisions, and identical case law derived from contradictory code provisions (Sacco 1991).

3. *Formalism as a Category in the Sociology of Law*

Treves usefully distinguishes between sociologies that include law as one of the elements in an integrated representation of society, and sociologies that take law as their object, bringing to bear on it the variety of instruments of sociological investigation (Treves 1995). The concept of legal formalism has been deployed critically against the first type of theory and descriptively within the second type.

In both descriptive sociologies of modernity and prescriptive political philosophical accounts, it is common for law to figure as an important, sometimes an essential, building block. The theory is not about law, but if its representation of law is inaccurate, the theory fails. The charge of formalism, in this context, means that the general theory represents law as having a gapless, meaning-based internal structure, responsive to outside imperatives of some kind. In fact, according to the critics, the contradictory internal structure of Western legal systems leaves adjudicators and other legal administrators great discretion in the interpretation of norms. It follows that neither the legitimacy of the legal order, nor its content, nor the effects of legal institutions, can be inferred from the external imperatives the theorist imagines animate them.

In this vein, critical legal scholars attacked the Marxist sociology of law as no more than the will theory, with the 'logic of the commodity' in place of natural rights (Kennedy 1997), and then turned the same critique on American functionalist legal sociology (Gordon 1984), and finally against the Habermasian attempt to distinguish between discourses of justification and application (Michelman 1996). In each case, a formalist treatment of law was the weak link in the general sociological construction.

The field-defining work in Treves's second mode—the multifaceted investigation of law as a social phenomenon—has been Max Weber's *Economy and Society*. Weber there offers a descriptive typology of forms of legal rationality. The 'highest' type—a logically formal rational system—is a collection of norms that are internally consistent and that officials apply to particular fact situations according to textual and conceptual formalism premised on gaplessness, as described above. In Weber's phrase, interpretation is

the 'logical analysis of meaning.' Despite the superficial resemblance, this ideal type is not just the late nineteenth-century will theory under another name, because it is not intrinsically associated with either the concept or the social reality of individual freedom, and because it is a contingent product of Continental legal history, rather than implicit in a larger normative conception. Weber dismissed the critiques formulated by the sociological jurists, described above, as self-serving resistance to the inevitable rationalizing trend of modernity (Weber 1954).

For this reason, his ideal type today defines a problematic rather than offering a powerful description. First, it is a matter for investigation how an order claiming any of Weber's types of legal rationality operates through officials at the level of practice (Sarat 1985), how state law interacts with other normative orders, and what effects are plausibly linked to a type (Trubek 1972). Second, it is a question for further study whether it is ever plausible, given the critique of gaplessness and the ever-present possibility of the abuse of deduction, that a legal order operates in practice in a way usefully described as logically formal rationality. Third, as the above discussion shows, the legal/bureaucratic mode of legitimation no longer relies on the claim of logically formal rationality, but rather on a complex mixture of claims of local meaning-based closure, claims of policy rationality (substantively rational, in Weber's terminology), and claims of democratic procedural legitimacy. These have not precluded a 'legitimation crisis.'

Paradoxically, given Weber's starting point, the modern sociology of law typically takes claims of closure in interpretation as motivated errors requiring explanation, rather than as the accurate self-description of legal modernity. Since 1900 critics have argued that the claim of closure masks conscious or unconscious 'legislative' or ideological agendas, and, since the 1930s, that the same claim is in one way or another linked to the Freudian superego and the Father (Holmes 1897, Frank 1949, Kennedy 1997, Goodrich and Carlson 1998).

4. Conclusion

Formalism, in its various senses, has been an epithet for so long that it is a plausible act of rebellion to embrace it (Scalia 1989), and many of the doctrines of the late nineteenth-century US Supreme Court that provoked the sociological jurists' critique have been revived in recent American constitutional law. These developments have produced interesting new critiques, but have not revised the terms of the discussion. Toward the end of Karl Llewellyn's classic article, 'What price contract?' (1931), he writes: 'One turns from contemplation of the work of contract as from the experience of Greek tragedy. Life struggling against form, or through form to its will—"pity and terror".' One might respond, against Llewellyn's romanticism, that descriptively formal law—both formal rule systems and formal techniques of interpretation—involves the morally delicate refusal to respond to the call for justice in the particular case, for reasons that may be good or bad according to the circumstances. Formal law is part of the drama of governance, the trivial or murderous drama of breaking eggs to make omelettes. The critical use of the term formalism, against the abuse of deduction and the fantasy of gaplessness in legal discourse, is part of the twentieth-century battle between those who have wanted to depoliticize the drama as much as possible, through reason, and those who have seen it as inevitably a dangerous improvisation.

See also: Disputes, Social Construction and Transformation of; Justice and Law; Law and Society: Sociolegal Studies; Law as an Instrument of Social Change; Law as Constitutive; Law, Mobilization of; Law: New Institutionalism; Law, Sociology of; Natural Law; Procedure: Legal Aspects; Rights: Legal Aspects

Bibliography

Brewer S 1996 Exemplary reasoning: Semantics, pragmatics, and the rational force of legal argument by analogy. *Harvard Law Review* **109**: 923–1028

Cardozo B 1921 *The Nature of the Judicial Process*. Yale University Press, New Haven, CT

Demogue R 1911 *Les notions fondamentales du droit privé: Essai critique*. Arthur Rousseau, Paris

Frank J 1949 *Law and the Modern Mind*. Coward-McCann, New York

Fuller L 1941 Consideration and form. *Columbia Law Review* **41**: 799–824

Geny F 1899 *Methode d'interpretation et sources en droit prive positif: Essai critique*. A. Chevalier-Marescq, Paris

Goodrich P, Carlson D G 1998 *Law and the Postmodern Mind*. University of Michigan Press, Ann Arbor, MI

Gordon R 1984 Critical legal histories. *Stanford Law Review* **36**: 57–125

Grey T C 1983 Langdell's orthodoxy. *University of Pittsburgh Law Review* **45**: 1–53

Gurvitch G 1932 *L'idée du droit social*. Recueil Sirey, Paris

Hand L 1958 *The Bill of Rights*. Harvard, Cambridge, MA

Hart H L A 1994 *The Concept of Law*. Clarendon, Oxford, UK

Herget J E 1996 *Contemporary German Legal Philosophy*. University of Pennsylvania Press, Philadelphia, PA

Holmes O W 1881 *The Common Law*. Little, Brown, Boston, MA

Holmes O 1897 The path of the law. *Harvard Law Review* **10**: 457–78

Horwitz M J 1992 *The Transformation of American Law, 1870–1960: The Crisis of Legal Orthodoxy*. Oxford: New York

von Jhering R 1869 *Der Geist des romischen Rechts auf den verschiedenen Stufen seiner Entwicklung*. Breitkopf und Hartel, Leipzig [1880 *L'esprit du droit romain dans les diverses phases de son développement*. A. Marescq, Aîné, Paris]

von Jhering R 1877–83 *Der Zweck im Recht*. Breitkopf & Hartel, Leipzig, Germany [1913 *Law as a Means to an End*. Boston Book Company, Boston, MA]
Kelsen H 1934 *Reine Rechtslehre*. F. Deuticke, Leipzig, Germany [1992 *Introduction to the Problems of Legal Theory*. Clarendon, Oxford, UK]
Kennedy D 1976 Form and substance in private law adjudication. *Harvard Law Review* **89**: 1685–1778
Kennedy D 1980 Toward an historical understanding of legal consciousness: The case of classical legal thought in America, 1850–1940. *Research in Law and Sociology* **3**: 3–24
Kennedy D 1997 *A Critique of Adjudication* (fin de siècle). Harvard, Cambridge, MA
Kennedy D 2000 From the will theory to the principle of private autonomy: Lon Fuller's 'Consideration and Form'. *Columbia Law Review* **100**: 94–175
Klare K 1978 Judicial deradicalization of the Wagner Act and the origins of modern legal consciousness, 1937–1941. *Minnesota Law Review* **62**: 265–339
Llewellyn K 1930 A realistic jurisprudence—The next step. *Columbia Law Review* **30**: 431–75
Llewellyn K 1931 What price contract?—An essay in perspective. *Yale Law Journal* **40**: 704–51
Maine H S 1917 *Ancient Law*. Dutton, New York
Michelman F 1996 Family quarrel. *Cardozo Law Review* **17**: 1163–77
Pollock F, Maitland F 1898 *The History of English Law*. Cambridge, Cambridge, UK
Pound R 1917 The end of law as developed in juristic thought II: The nineteenth century. *Harvard Law Review* **30**: 201–46
Sacco R 1991 Legal formants: A dynamic approach to comparative law. *American Journal of Comparative Law* **39**: 1–43, 343–79
Sarat A 1985 Legal effectiveness and social studies of law: On the unfortunate persistence of a research tradition. *Legal Studies Forum* **9**: 23–31
Scalia A 1989 The rule of law as a law of rules. *U. Chi. L. Rev.* **56**: 1175–88
Schauer F 1988 Formalism. *Yale Law Journal* **97**: 509–48
Schauer F 1991 *Playing By the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*. Clarendon, Oxford, UK
Treves R 1995 La sociologie du droit: Un debat. In: Andrini S, Arnaud A-J (eds.) *Jean Carbonnier, Renato Treves et la sociologie du droit: Archéologie d'une discipline*. L.G.D.J., Paris
Trubek D 1972 Max Weber on law and the rise of capitalism. *Wisconsin Law Review* **1972**: 720–53
Weber M 1954 *Max Weber On Law In Economy and Society*. Harvard, Cambridge, MA
Weinrib E J 1988 Legal formalism: On the immanent rationality of law. *Yale Law Journal* **97**: 949–1016

D. Kennedy

Legal Insurance

1. The Term 'Legal Insurance'

Legal insurance is a voluntary private insurance which covers the costs of lawsuits. It may also be called *legal cost insurance*, *legal expenses insurance*, or *legal protection insurance*. The German expression is *Rechtsschutzversicherung*, the French *l'assurance de défense*. The European Union directive EU 87/344/EWG uses the term *legal expenses insurance*, which is defined in Art. 2 as follows: 'Such consists in undertaking, against the payment of a premium, to bear the costs of legal proceedings and to provide other services directly linked to insurance cover, in particular with a view to:

(a) securing compensation for the loss, damage or injury suffered by the insured person, by settlement out of court or through civil or criminal proceedings,

(b) defending or representing the insured person in civil, criminal, administrative or other proceedings or in respect of any claim made against him.'

The expressions *Prepaid Legal Service* and *Group Legal Plan*, which are used in the USA, illustrate the particular economic context and judicial functions of the service in this country. In Asian and African countries, legal insurance is practically unknown.

2. Development and Actual Situation

Although the origins of legal insurance can be found in the nineteenth century, it grew into an economically and politically relevant insurance branch in Western Europe only after World War II, albeit with regional differences. In Great Britain, for example, until 1967 it was forbidden and punishable to financially support another party in litigation if one had no personal interest in the outcome of the trial ('champerty'). Thus, the first legal insurance policy entered the British market as late as 1974. The spread of legal insurance has since been hindered by the high costs of litigation in Great Britain and the resulting low tendency of the general population to go to court (Prais 1995, p. 433). The French have also shown little interest in legal insurance policies. Legal insurance has, in contrast, achieved its greatest importance in German-speaking countries. In 1989 average per capita spending on legal insurance was 23.5 ECU in Germany, 10.8 ECU in Switzerland, and 17.4 ECU in Austria. Other countries follow at some distance. Average spending in Belgium was 8.9 ECU, in the Netherlands 5.3 ECU, in France 1.7 ECU, and in Great Britain a mere 1.1 ECU per capita (Blankenburg 1994, p. 295). Since 1989 this branch of insurance has expanded rapidly in all above-mentioned countries, although at different rates. In 1997 the total revenues from legal insurance amounted in Germany to 2.44 billion ECU, or an average 28.4 ECU spent per capita. Swiss revenues reached a level of 17.16 ECU, Austria 30.6 ECU, Belgium 18.9 ECU, the Netherlands 11.9 ECU, and France 5.8 ECU (Comité Européen des Assurances 1999). Over 50 percent of all German

households are covered by at least one legal insurance policy.

Several factors are responsible for the different development of legal insurance coverage in European countries. Most important is the difference in national systems of calculating legal counseling and court costs, in particular because contingency fees for lawyers are illegal in many countries. In calculating the risks of this type of insurance, the insurance company must be able to estimate the future costs of counsel and litigation. The German method of calculating legal counseling fees for both parties according to the sum in dispute (including legal counseling, paid by the party that loses the case), facilitates these estimates.

3. *Types of Insurance*

Insurance for motoring and traffic accidents has been the main factor behind the expansion. The second important field for legal insurance is family insurance, which covers claims arising out of the private sphere of the insured party, his or her spouse, and underaged children. The most common insurance package includes protection from claims arising out of diverse types of contracts, social insurance, and defense in criminal and misdemeanor cases. It is often combined with automobile and traffic insurance. Generally excluded are counseling for legal matters concerning divorce and inheritance, as well as for the stipulation of contract such as real estate purchases and testaments, since the latter are deliberately chosen legal interactions. In this sense, the insurers remain true to the concept of insurance as protection from unattended and unforeseeable risks.

A third field of legal insurance offers commercial legal protection, a policy for businesses and professionals. The coverage includes labor litigation for the employer on the one side, defense in misdemeanor cases and criminal procedures on the other. Some cover taxation matters as well. Generally, coverage also includes insurance for the company's motorized vehicles.

Legal insurance for real estate matters or for commercial leasing are often offered as special policies. A further policy insures farmers.

Legal insurance policies are offered both by specialized companies and by normal insurance companies, for which it is only one, often minor, business area. In some countries, legal counseling is offered by employees of the insurance company, who may not necessarily have legal training; in others the insured party is given complete freedom in the choice of attorney. In Germany, legal counseling by a lawyer is mandatory. Insurance companies allow free choice of attorney, but may offer recommendations. The EU directive 87/344/EWG has brought a certain degree of standardization to Europe, regulating conflicts between insured and insurer and protecting freedom in the choice of legal counsel.

4. *The Situation in the United States*

The American *prepaid legal services* began as a service offered by trade unions and professional groups to their members. These organizations employ attorneys, who advise members not only on labor law, but on all types of legal matters, and represent them in court. This service was suppressed in several states until the US Supreme Court ruled in 1971 that the fundamental constitutional right of 'meaningful access to the courts' would be denied 'if courts could deny associations of workers or others the means of enabling their members to meet the cost of legal representation' (US Transportations Union vs. State Bar of Michigan, 401 US 576 (1971), 585).

The steady development of welfare policies then led to a liberalization of laws regarding legal counsel. Labor unions, employers' associations, and the American Bar Association agreed that legal protection can legally be the subject of a wage settlement. Since this agreement, many funds have been established and financed by workers, unions, and employers. These funds are favorably taxed for nonprofit-organizations, which usually do not cooperate with commercial insurance companies (Pfennigstorf and Schwartz 1986, p. 52). The most common type is the 'group legal service plan,' which covers workers in a particular branch and is financed substantially by the employers.

Through these legal service plans, workers and their families have gained access to legal counseling and, when necessary, to legal support in court. The services extend to all types of legal matters including consultations with and financing of counsel in private concerns (such as real-estate transactions, rental contract, divorce proceedings, and the writing of testaments). These funds thus cover legal counsel needs which are excluded from the European legal insurance policies.

The differences are due to the noncommercial character and socially motivated goals of the US programs. These intend to raise the awareness of legal rights, especially in a population with modest financial means, and to offer them the possibility of asserting these rights (Pfennigstorf and Schwartz 1986, p. 47, see also *Law, Mobilization of*). In order to control the costs and the quality of the rendered legal services, prepaid legal service plans offer counseling only through their own employees, who may be attorneys bound by exclusive contract but in many cases have little or no official legal training. Alternative dispute regulation (ADR) is one very important field of action for these plans.

With this background, commercial legal insurance has only developed slowly. A few insurance companies succeeded with special programs such as consultation by telephone or a limited number of hours of legal

consultation (Pfennigstorf and Schwartz 1986). Since the premium must be relatively low to attract clients, legal insurance tends not to be a profitable commercial venture in the US.

5. *Consequences and the Success of Legal Insurance*

The effects of the expansion of legal insurance have been considered primarily from two perspectives. For some, the growing number of covered persons is itself a success, since ideally everyone, including the financially weaker members of a population, should have access to legal counseling and protection. Others focus on the readiness to litigate, which is facilitated by legal insurance, and the growing burden placed on the courts.

This conflict forms the basis of many discussions of legal insurance in sociological literature and in the public arena, especially in Germany. While most agree that equal access to the courts is a fundamental right under the rule of law, there is also a consensus that unnecessary litigation should be avoided. Critics of legal insurance claim that insured parties will be more likely to use the courts to pursue their rights or presumed rights, since their own financial risks have been minimized, in order to profit from their insurance coverage and the paid premium. Attorneys are purported to recommend more often that a client go to court if that client is insured. This not only increases the growing numbers of lawsuits but also the financial opportunities for the legal professions.

Rational choice theorists argue that legal insurance results in a 'torrent of lawsuits,' 'wasting a huge amount of resources,' and thereby leads to an 'inefficient allocation in civil law procedures' (Adams 1981, p. 125, 1986). According to this analysis, legal insurance must necessarily lead to an increase in the number of lawsuits, since there is no force to counter this expansion. Attorneys have no economic interest in reducing the costs and numbers of lawsuits, since this directly affects their income. Insurance companies are also interested in an increase in lawsuits, since this expands their markets, providing, of course, that they can pass the increase in costs on to their policy holders. Courts, on the other hand, are obliged by law to accept and to decide on every claim and have therefore no means of controlling the number of cases. Meanwhile, the general population seems to react to the increasing complexity of social relationships and the legal system by looking to legal insurance policies as a form of protection against possible future lawsuits (Blankenburg 1994).

Empirical research has, in contrast, shown that legal insurance has caused only a mild increase in lawsuits in Germany (Blankenburg and Fiedler 1981, Jagodzinski et al. 1994). Insured plaintiffs litigate only 5–10 percent more often and more persistently than uninsured plaintiffs. In some fields of litigation, the difference is even smaller. Considerable differences of 10 percent to 27 percent between the two groups have been observed only in the contestation of traffic misdemeanors. In these particular cases, the insured tend to be less successful than the uninsured, since insured plaintiffs are more willing to go to court in less promising cases or to claim excessive sums which can not be awarded by the court (Jagodzinski et al. 1994, p. 141).

These empirical studies show that it is incorrect to assume that a 'torrent of lawsuits' have been caused by legal insurance. Legal insurance does not lead to a significant amount of abuse or negative developments. The research suggests that attorneys adequately differentiate between founded and unfounded claims before beginning litigation. Statistics show that when insured plaintiffs litigate more persistently than uninsured plaintiffs, they tend to achieve favorable verdicts (Jagodzinski et al. 1994, p. 143).

Legal insurance has had a definite financial effect on the legal profession, particularly in Germany, where 70 percent of the turnover goes to attorneys. Every attorney received an average of 44,000 DM from legal insurance in 1989. In other countries, it has proved to be far less important for the bar (Blankenburg 1994).

According to Blankenburg, different national regulations concerning legal counseling are the main reason for these discrepancies. While only lawyers are allowed to give legal advice in Germany, insurance companies in the Netherlands rely on the less expensive alternative of legal counseling through company employees. Attorneys are only consulted in complicated cases. The existing competition for clients between legal insurance and lawyers has led to a downward trend in legal counseling fees in this country, and has also resulted in lower premiums for the insurance policies. In Germany, legal insurance counseling and services quickly become unaffordable for the financially weak members of the population. For this reason, legal insurance has not replaced state financed legal aid programs (Blankenburg 1994, p. 296). This observation is also true for GB (Prais 1995, p. 444).

As these discussions show, the effects of legal insurance are varied. It is useful to protect clients' rights but tends to become more and more expensive, since neither lawyers nor the insurance companies have an interest in limiting the costs of legal counseling. As long as there is no effective competition on the market, premiums will continually rise. In this respect, the US system of prepaid legal services is more economical, but it also closes the market for commercial legal insurance, thereby hindering competition. According to European experiences, such a system of legal counseling tends in the long run to increase costs, as long as there is no counteracting force.

See also: Civil Society/Public Sphere, History of the Concept; Economics: Overview; Insurance; Insurance and the Law; Law and Economics: Empirical Dimensions; Law and Risk; Liability: Legal; Risk: Empirical Studies on Decision and Choice; Risk, Sociological Study of; Risk, Sociology and Politics of; Risk: Theories of Decision and Choice; Torts in Law

Bibliography

Adams M 1981 *Ökonomische Analyse des Zivilprozesses* (*Economic Analysis of Civil Procedure*). Athenaeum, Königstein

Adams M 1986 Der Zivilprozeß als Folge strategischem Verhaltens (Civil law procedures as the result of strategic behavior). *ZfRSoz* **2**: 212–25

Blankenburg E 1994 Rechtsschutzversicherung als Alternative zur sozialen Rechtshilfe? (Is legal insurance an alternative to legal aid?). *Zeitschrift für Rechtspolitik* **8**: 294–7

Blankenburg E, Fiedler J 1981 *Die Rechtsschutzversicherung und der steigende Geschäftsanfall der Gerichte* (*Legal Insurance and the Growing Demands on the Courts*). Mohr, Tübingen

Cerveau B, Tribondeau D 1991 *L'Assurance de Protection Juridique* (*Legal Protection Assurance*). Argus, Paris

Comité Européen des Assurances 1999 (Legal Expenses Insurance Committee) Report 1998 Appendix to document PJ 8029 (12/98). Paris

Jagodzinski W, Raiser T, Riehl J 1994 *Rechtsschutzversicherung und Rechtsverfolgung* (*Legal Insurance and the Recourse to the Court*). Bundesanzeiger, Köln

Pfennigstorf W, Schwartz A 1986 *Legal Protection Insurance*. American Bar Foundation, American Prepaid Legal Services Institute, Chicago

Prais V 1995 Legal expense insurance. In: Zuckerman A, Cranston R (eds.) *Reform of Civil Procedure*. Clarendon Press, Oxford, UK

T. Raiser

Legal Issues: Public Opinion

Two guiding principles of Western democratic thought are that effective government requires the consent of the governed, and public consent requires the fair and just rule of law. Thus research on public opinion toward the law and the legal system has preoccupied mainstream social science for much of its history, reappearing in various guises, including the 'fair press vs. free trial' debate, seemingly paradoxical links between actual crime and fear of crime, public disdain for 'judicial leniency,' and theories of perceptual deterrence.

This short review will of necessity be selective in its coverage. With respect to substantive legal issues, it will emphasize criminal justice, providing only passing references to tort litigation and no coverage of public opinion in the domains of contract law, family law, intellectual property law, and so on. With respect to methodology, it will emphasize quantitative public opinion surveys, statistical modeling of field data, and laboratory experiments, neglecting the rich but idiosyncratic literature using qualitative interviews and qualitative analyses of textual sources. Finally, the article—like the English-language empirical literature—gives disproportionate emphasis to North American and British research, though some lines of cross-cultural research are noted.

1. Public Evaluations of the Legal System: An Overview

1.1 Evaluating the Police

In the aggregate, Americans and Europeans evaluate the police favorably. For example, most Americans report confidence in the police as an institution, and about half have confidence in the ability of the police to protect them from violent crime (Bureau of Justice Statistics 2000). Such confidence is not inevitable. For example, in 1993, positive views of the police were expressed by a majority of French, Belgian, Dutch, and West German citizens but by fewer than half of East Germans (Reif and Melich 1996), and within the United States, African Americans consistently report more negative views of police conduct than the white majority; they are about twice as likely to report low confidence in the police or to express the view that the police have low ethical standards.

1.2 Evaluating Attorneys

Most adults know, or at least recognize, a few of the multitude of 'lawyer jokes' long in circulation (see Galanter 1998). We rarely tell jokes about physicians, whom we hold in high regard, or molesters, whom we loathe. This suggests that we are ambivalent about lawyers; we decry their ruthless mercenary adversarialism—except when they are representing us in a conflict. Thibaut and Walker (1975) asked Americans to imagine various hypothetical conflicts and various mechanisms for resolving them—including two-party mediation, third-party investigation and resolution by a neutral inquisitor, and a third-party decision based on investigation and arguments presented by advocates for each side. Importantly, the descriptions were stripped of labels like 'lawyer' or 'judge' and other overt references to actual legal systems. A strong majority identified the adversarial system as the fairest mechanism. Subsequent studies (see Lind and Tyler 1988) have

replicated this finding in European nations with inquisitorial systems, suggesting that it is not an artifact of American socialization or familiarity with adversarial justice. (However, respondents in Asian countries tend to divide their support between the adversarial model and two-party mediation.)

1.3 Evaluating the Courts

A majority of Americans—of all ethnicities—rate the honesty and ethical standards of judges as high or very high (Bureau of Justice Statistics 2000). But this high regard does not translate into satisfaction with judges' decisions. For several decades in the twentieth century, American public opinion polls showed consistent majority support for the view that the courts are too lenient in their treatment of criminal defendants, (Bureau of Justice Statistics 2000). Interestingly, African Americans are only slightly less likely to endorse this view, and the gap between racial groups has been closing.

Taken at face value, these statistics imply that the courts are either 'out of touch' with public views, or intentionally ignoring them. There is some basis for the latter interpretation, of course; Western systems of rights and protections evolved in large part to protect individuals from group tyranny. But a number of clever experiments reveal these citizen complaints are largely based on distorted perceptions of the justice system and its caseload (see Roberts and Stalans 1998). Citizens who view courts as lenient in the abstract have been found to recommend criminal sentences that fall *below* statutory mandatory minimum sentences or the discretionary sentences of judges. And citizens' views in actual cases become significantly more lenient when they read lengthy trial descriptions rather than short trial summaries.

Most available survey evidence indicates that Americans hold much more favorable opinions of the jury system than of the courts more generally (Hans 1993). MacCoun and Tyler (1988) found that citizens strongly preferred trial by jury to trial by judge, and the traditional 12-person unanimous jury to smaller or nonunanimous juries. Relative to trial by judge and to smaller or nonunanimous juries, the traditional jury structure was seen as fairer, more accurate, more thorough, and more representative of community viewpoints. This was not blind enthusiasm; citizens preferred more efficient approaches (trial by judge or small nonunanimous juries) for trivial cases like shoplifting.

Proponents of alternative dispute resolution (ADR) mechanisms such as arbitration and mediation argue that traditional trials are too costly, too time-consuming, too complex, and too alienating. Indeed, many (but not all) ADR programs receive high marks from all parties (plaintiffs, defendants, and attorneys), but contrary to the ADR movement's rhetoric, disputants who actually participate in trials tend to view the trial process favorably, and to a surprising extent, this is true of 'losers' as well as those who prevail at trial (MacCoun et al. 1992).

2. Sources and Mechanisms of Influence on Public Views

2.1 Direct Experience

The procedural justice literature documents the importance of direct personal experience in shaping the public's views of law and legal authorities. Most citizens have had direct contact with the police; some have had contact with lawyers; fewer have participated in trials; and only a minority have direct experience with the penal system.

For many citizens, jury service is their major (or only) direct experience with the court system. Diamond (1993) has reviewed survey evidence on the effects of jury service on public attitudes, finding that a majority of jurors provide positive evaluations of the jury system, and of the legal system, after their service. Most jurors say their attitudes became more favorable than before their service, and attitudes are more favorable among those who are selected for a jury than among those who serve but are not seated.

2.2 Mass Media Effects

Clearly, citizens without direct contact with the legal system must base their views on other sources. But even those with direct experience are routinely bombarded with indirect, vicarious information. There is an enormous literature demonstrating that the mass media promote distorted views of crime, and the process and outcomes of the justice system (e.g., Roberts and Stalans 1998, Warr 1994). The presumed causal chain is ACTUAL EVENTS → MEDIA FILTERING AND DISTORTION → BIASED PUBLIC BELIEFS AND OPINIONS, with these public views feeding back to the media via consumer choices.

The first part of the chain is well supported. Numerous content analyses over several decades show a consistent pattern: Media coverage provides a highly skewed picture of crime and the legal system, and this is true of newspapers, television news, 'reality-based' police shows, and fictional portrayals on television and in movies. In the criminal domain, reported offenses are disproportionately violent, defendants are disproportionately African American, police officers are disproportionately European American, victims are disproportionately female, and offenses disproportionately involve strangers (see Warr 1994). In the personal injury domain, reported lawsuits are

much more likely to result in a jury trials and to produce a plaintiff victory than the typical or average lawsuits, and reported jury awards are 14 times larger than the average award in state courts, and 34 times the size of the median award (Bailis and MacCoun 1996).

There is also evidence for the second half of the causal chain. The prejudicial effects of pretrial publicity on criminal juror judgments are well documented (see Steblay et al. 1999). And numerous studies indicate that citizens' views of court leniency are shaped by the mass media (see Warr 1994). Citizens tend to overestimate the severity of typical crimes and criminals (relative to official statistics), and these misperceptions reliably predict the propensity to see the courts as excessively lenient; e.g., citizens given actual newspaper accounts of trials supported harsher criminal sentences than citizens who read actual court documents from the same cases (see Roberts and Stalans 1998).

2.3 Hindsight Bias

Hindsight bias is the *ex post* tendency to overestimate the *ex ante* likelihood of an outcome, relative to what one would have actually guessed before the event. Because most legal judgments are made *ex post*, they are vulnerable to this bias, as documented in a variety of experimental studies (Rachlinski 1998). Hindsight bias also influences citizens' *ex post* reactions to newsworthy legal decisions. Trial evidence appears more incriminating to people who believe a defendant has been convicted than for those who believe the defendant was acquitted (Bodenhausen 1990).

A pernicious example of this phenomena was the elite and popular reaction to the acquittal of actor/athlete O. J. Simpson at his criminal homicide trial in the USA. Because 9 of the 12 jurors were African American, a large number of commentators, including the lead prosecutor, 'explained' the verdict as a case of 'racial nullification'—the notion that African Americans are more lenient when judging other African Americans. In an October 1995 Gallup poll, 48 percent of whites vs. 18 percent of African Americans said they had 'less confidence' that 'jurors can reach a verdict in a trial without letting their racial attitudes affect their judgment' (Moore and Saad 1995). In fact, the jury literature provides little support for 'racial nullification' as a general phenomenon (Kerr et al. 1995). The reasons for the acquittal will never be known with certainty, but it seems clear that pronouncements of 'racial nullification' reflected hindsight bias. Despite widespread knowledge of the composition of the jury, surveys during the trial showed that few trial attorneys and citizens expected an acquittal; indeed *Sporting Index* reported that most American gamblers were betting on a conviction (*Los Angeles Times* 1995).

2.4 Victim Blaming

Cognitive accessibility and hindsight effects are 'cold' cognitive mechanisms, but there is little doubt that 'hot' emotional and motivational factors influence public opinion as well. According to the just world theory (see Montada and Lerner 1998), it is psychologically threatening to believe that undeserving victims can suffer a horrible fate. Experiments show that we often alleviate this distress by searching for reasons why victims 'got what they deserved.'

3. Underlying Dimensions of Public Evaluations of the Legal System

The history of theory and research on lay justice judgments involves a continuous branching process in which simple, unidimensional models have given way to more complex multidimensional models (see Tyler et al. 1997).

3.1 Outcome Satisfaction, Relative Deprivation, and Distributive Justice

Economics, behaviorist psychology, the social exchange tradition of sociology, and the public choice tradition of political science share a common emphasis on the explanatory power of outcomes and incentives. The mass media and the legal literature tend to perpetuate the view that outcomes—especially monetary outcomes—drive legal behavior, legal judgments, and evaluations of the legal system. An early refinement of this view was relative deprivation theory (see Tyler et al. 1997)—the notion that what matters to citizens are relative outcomes (mine vs. yours or theirs), rather than absolute outcomes. As a predictive theory, the relative deprivation approach has had only mixed success, in large part because the theory *overpredicts* resentment and rebellion. One problem is that the theory fails to predict which of many possible comparison standards the citizen will use; citizens do not invariably choose the source that provides the most invidious comparisons. Another weakness is the failure to predict whether outcome discrepancies will be seen as justified or not.

Descriptive (rather than normative) theories of distributive justice address this latter gap. In the 1970s, the most influential model was equity theory (e.g., Walster and Walster 1975), which models justice judgments as a function of the relative ratio of inputs to outcomes for each actor in a relationship. Indeed, in reactions to job-related conflicts, citizens in Eastern and Central Europe (Bulgaria, Hungary, Poland, Russia) are at least as likely to apply this equity standard as are citizens in traditional Western market democracies (France, Spain, the United States) (Cohn

et al. 2000). But it has become clear that this model works less well in nonmarket contexts and relationships (Fiske 1992). Across various settings, studies have found support for allocation by equality, by need, or by more complex multidimensional decision rules (see Mellers and Baron 1993). In addition, judgments of distributive justice (and their behavioral consequences) vary by social level of analysis; for example, perceiving that one's group has received unfair outcomes is more likely than purely individual deprivation to motivate anger and political action (see Tyler et al. 1997).

3.2 Procedural Justice

Outcomes are clearly linked to public evaluations of justice, but there is growing evidence that scholars have overestimated the importance of this linkage. Thibaut and Walker's (1975) influential experimental research program provided considerable evidence that the processes by which outcomes are reached matter profoundly to citizens. Since the 1970s, this 'fair process effect' has been replicated across a wide variety of nations (in North America, Europe, and Asia), domains (criminal justice settings, tort litigation, private organizations, and the political sphere) and dependent variables (satisfaction with the legal system, acceptance of third-party decisions, compliance with legal authorities, and decisions to file lawsuits) (see Lind and Tyler 1986, MacCoun et al. 1992, Tyler et al. 1997). In a surprising number of studies, procedural justice judgments emerge as a stronger independent predictors than either objective or subjective measures of the outcomes citizens received.

Thibaut and Walker (1975) contended that procedures matter to citizens because fair procedures produce fair outcomes. The key parameters of their model were decision control and process control; the latter is important because it is rarely possible to resolve conflicts without relinquishing some decision control. The predominant determinant of process control in the Thibaut–Walker program was 'voice'—the opportunity to control the presentation of one's arguments to a third-party decision-maker. Even among tort litigants, popularly portrayed by the media as greedy plaintiffs and stingy defendants, the 'opportunity to tell my story' is the most commonly cited goal of both sides. Indeed, this voice motive helps to explain why ADR mechanisms have generally failed to streamline court dockets; because ADR provides a relatively cheap way to 'tell my story,' it diverts more cases from private settlement than from trial (see MacCoun et al. 1992).

Thibaut and Walker's control-based account of voice effects has fared less well. Lind and Tyler (1988) have provided the most empirically successful account of voice effects. Their group value theory argues that voice effects are 'relational' rather than instrumental; they reflect citizens' motivation to achieve a favorable and secure social identity. In numerous studies, relational concerns—the sense that authorities acted with benevolent and impartial motives, and that one was treated with dignity and respect—emerge as more important determinants than either outcome evaluations or perceptions of control (Lind and Tyler 1988, Tyler 1990, Tyler et al. 1997). Within the United States, these relational factors are found in judgments across gender, ethnic, educational, and occupational categories, and these findings have been replicated in Europe and even in more collectivist Asian societies (Tyler et al. 1997).

There are moderators and boundary conditions on the model. Procedural justice effects are weaker for citizens who identify strongly with a subordinate group, but weakly with a superordinate group (e.g., minority 'separatists' as opposed to 'assimilationists'). Another is that procedures matter more when citizens lack clear objective criteria for evaluating outcomes (van den Bos et al. 1997).

Nevertheless, fair process effects are so robust that scholars fear they pose a threat of 'false consciousness'—the notion that authorities can use the appearance of fair procedure (dignity, respect, voice) as an inexpensive way to coopt citizens and to distract them from outcomes that by normative criteria might be considered substantively unfair or biased (e.g., MacCoun et al. 1992). For example, an insurance executive requested a meeting with two procedural justice researchers to ask how he might increase the formality of meetings between clients and insurance adjustors, in order to reduce the rate of contested claims.

3.3 Retributive Justice

It is important to note that citizens' departures from pure economic self-interest are not merely warm, sentimental, and benevolent. The behavioral game theory literature has provided ample evidence that people will often forgo personal gain in a spiteful attempt to harm others, and it is tempting to interpret the strong retributive streak in American public opinion in a similar way. During the 1990s, the USA has seen a dramatic escalation in the size of the prison population, especially relative to that of other industrialized nations. In an era when even liberal politicians decry 'big government,' the American public seems blithely unconcerned by staggering increases in the total share of the state and federal budgets accounted for by the prison complex.

Public misperception of crime rates is surely a partial explanation, but several lines of evidence implicate retributive motives, rather than an instrumental attempt at crime control. For example, many citizens overtly endorse a deterrence rationale for the death

penalty, believing that 'it will prevent crimes.' Yet most do not change their views when asked how they would feel if there were unequivocal evidence that execution provided no marginal deterrence above and beyond life imprisonment (Ellsworth and Gross 1994). Indeed, attitudes toward capital punishment are relatively impervious to research evidence, showing little or no deterrent effect; proponents and opponents each interpret the evidence in a manner that preserves their pre-existing opinion (Lord et al. 1979) (see *Death Penalty*).

3.4 Fear of Crime

It is a cliché among academicians that fear of crime bears no relation to actual crime rates. In fact, between 1973 and 1987, American fear ratings moved in tandem with the violent crime rate (Warr 1994). But perceptions of crime trends are not very accurate; a growing minority (from 4 percent in 1993 to 35 percent in 1998) recognize that US crime fell in the late 1990s, but throughout that period, a majority (from 87 percent in 1993 to 52 percent in 1998) contended that 'there is more crime in the US than a year ago' (Bureau of Justice Statistics 2000).

In many surveys, the elderly and women express the most fear of crime; these patterns are called 'paradoxical' (because young males are the group at greatest risk of victimization) but only under the peculiar notion that fear should solely reflect relative statistical risk. It appears that these patterns reflect differences in the way women and the elderly conceptualize the question (Ferraro 1995, Warr 1994), but also the influence of media depictions that exaggerate the proportion of crimes in which the victims are white females (Chiricos et al. 1997). But the fear literature is plagued by inconsistent findings, perhaps due to a frequent reliance on crude, single-item 'fear' measures that may not measure the experience of fear (i.e., events in the amygdala and autonomic nervous system) and that surely tap other constructs, like perceived risk, attitudes toward the police and the courts, attitudes toward stereotypic offender groups (youth, males, ethnic minorities), and so on.

3.5 Fear of Legal Sanctions

The general public tends to exaggerate the risks of arrest and punishment for many crimes, relative to actual risks and to the more accurate risk perceptions of active criminal offenders (see Paternoster 1987). These facts are consistent with either a deterrence interpretation (perceived legal risk deters crime) or a learning interpretation (offenders learn that legal risks are low). But quasi-experimental studies of legal interventions and panel self-report surveys suggest that the certainty of sanctions does deter crime, and that the effects of sanction severity are weak and unreliable (Paternoster 1987) (see *Deterrence*).

3.6 Legitimacy

In a series of experiments, Robinson and Darley (1997) identified numerous discrepancies between the Model Penal Code and lay intuitions about legal liability. They hypothesize that a congruence between lay views and the formal law enhances both legitimacy and compliance. Tyler's (1990) panel survey of Chicago citizens suggested that those who perceived laws and legal authorities to be fair were less likely to report having committed criminal offenses when surveyed at a later point in time. A re-analysis of data from the Milwaukee Domestic Violence Experiment found that the perceived unfairness of police interventions was a stronger predictor of subsequent domestic violence than was the initial decision to arrest or not to arrest the offender (Paternoster et al. 1997).

4. Conclusion

Public opinion research on legal issues has benefitted enormously from the increasing use of advanced multivariate analyses, panel surveys tracking individuals' views over time, and factorial experiments. But there are far too few cross-national studies. The bulk of the English language literature describes the views of North American and British citizens in the late twentieth-century. Thus we still know far too little about how people judge laws and legal authorities outside of common-law systems in wealthy, advanced industrial nations. We do not know whether these Anglo-American views reflect fundamental psychological principles or simply the particular context in which they were expressed.

See also: Criminal Justice, Sociology of; Death Penalty; Disputes, Social Construction and Transformation of; Juries; Justice and its Many Faces: Cultural Concerns; Justice and Law; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Law, Sociology of; Liability: Legal; Opinion Formation; Public Opinion: Microsociological Aspects; Public Opinion: Political Aspects

Bibliography

Bailis D S, MacCoun R J 1996 Estimating liability risks with the media as your guide: A content analysis of media coverage of tort litigation. *Law and Human Behavior* **20**: 419–29

Bodenhausen G V 1990 Second-guessing the jury: Stereotypic and hindsight biases in perceptions of court cases. *Journal of Applied Social Psychology* **20**: 1112–21

Bureau of Justice Statistics 2000 *Sourcebook of Criminal Justice Statistics Online, 1998*. (Various tables) http://www.albany.edu/sourcebook/
Chiricos T, Eschholz S, Gertz M 1997 Crime news and fear of crime: Toward an identification of audience effects. *Social Problems* **44**: 342–57
Cohn E S, White S O, Sanders J 2000 Distributive and procedural justice in seven nations. *Law and Human Behavior* **24**: 553–79
Diamond S S 1993 What jurors think: Expectations and reactions of citizens who serve as jurors. In: Litan R E (ed.) *Verdict: Assessing the Civil Jury System*. Brookings Institution, Washington, DC
Ellsworth P C, Gross S R 1994 Hardening of the attitudes: Americans' views on the death penalty. *Journal of Social Issues* **50**: 19–52
Ferraro K F 1995 *Fear of Crime: Interpreting Victimization Risk*. State University of New York Press, Albany, NY
Fiske A P 1992 The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychology Review* **99**: 689–723
Galanter M 1998 The faces of mistrust: The image of lawyers in public opinion, jokes, and political discourse. *University of Cincinnati Law Review* **66**: 805–45
Hans V P 1993 Attitudes toward the civil jury: A crisis of confidence? In: Litan R E (ed.) *Verdict: Assessing the Civil Jury System*. Brookings Institution, Washington, DC
Kerr N L, Hymes R W, Anderson A B, Weathers J E 1995 Defendant–juror similarity and mock juror judgments. *Law and Human Behavior* **19**: 545–67
Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum Press, New York
Lord C G, Ross L, Lepper M R 1979 Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* **37**: 2098–109
Los Angeles Times 1995 They bet on it. *Los Angeles Times*, 25 October 1995, A6
MacCoun R J, Tyler T R 1988 The basis of citizens' perceptions of the criminal jury: Procedural fairness, accuracy, and efficiency. *Law and Human Behavior* **12**: 333–52
MacCoun R J, Lind E A, Tyler T R 1992 Alternative dispute resolution in trial and appellate courts. In: Kagehiro D K, Laufer W S (eds.) *The Handbook of Psychology and Law*. Springer Verlag, New York
Mellers B A, Baron J (eds.) 1993 *Psychological Perspectives on Justice*. Cambridge University Press, New York
Montada L, Lerner M J (eds.) 1998 *Responses to Victimizations and Belief in a Just World*. Plenum Press, New York
Moore D W, Saad L 1995 No immediate signs that Simpson trial intensified racial animosity. *Gallup Poll Monthly* (October): 2–8
Paternoster R 1987 The deterrent effect of the perceived certainty and severity of punishment: A review of the evidence and issues. *Justice Quarterly* **4**: 173–217
Paternoster R, Brame R, Bachman R, Sherman L W 1997 Do fair procedures matter? The effect of procedural justice on spouse assault. *Law and Society Review* **31**: 163–204
Rachlinski J J 1998 A positive psychological theory of judging in hindsight. *University of Chicago Law Review* **65**: 571–625
Reif K, Melich A 1996 *Eurobarometer 40: Poverty and Social Exclusion*. European Commission, Cologne, Germany
Roberts J V, Stalans L J 1998 Crime, criminal justice, and public opinion. In: Tonry M (ed.) *The Handbook of Crime and Punishment*. Oxford University Press, New York
Robinson P H, Darley J M 1997 The utility of desert. *Northwestern University Law Review* **91**: 453–99
Steblay N M, Besirevic J, Fulero S M, Jimenez-Lorente B 1999 The effects of pretrial publicity on juror verdicts: A meta-analytic review. *Law and Human Behavior* **23**: 219–35
Thibaut J, Walker L 1975 *Procedural Justice*. Erlbaum Associates, Hillsdale, NJ
Tyler T R 1990 *Why People Obey the Law*. Yale University Press, New Heaven, CT
Tyler T R, Boeckmann R J, Smith H J, Huo Y J 1997 *Social Justice in a Diverse Society*. Westview Press, Boulder, CO
van den Bos K, Lind E A, Vermunt R, Wilke H A M 1997 How do I judge my outcome when I do not know the outcome of others? The psychology of the fair process effect. *Journal of Personality and Social Psychology* **72**: 1034–46
Walster E, Walster W G 1975 Equity and social justice. *Journal of Social Issues* **31**: 21–43
Warr M J 1994 Public perceptions and reactions to violent offending and victimization. In: Reiss A, Roth J A (eds.) *Understanding and Preventing Violence: Volume 4. Consequences and Control*. National Research Council, Washington DC

R. MacCoun

Legal Personality

In order to be able to participate in any legal or political system an actor must have a legal personality, that is, a set of rights and duties recognized by that system. In theory a legal system may be premised on the belief that certain actors have an inherent legal personality, or that legal personality is entirely a matter of the grant of rights and placement of duties by the sovereign, or that some actors have inherent legal personalities and some are granted a personality. In practice, however, every legal system both recognizes rights and duties in certain actors and insists that certain legal personalities are a matter of positive law. As a historical matter, the content of an actor's legal personality has waxed and waned depending on the legal system in question, instrumental questions about the utility of legal personality, and cultural presuppositions about who and what might have the benefit of rights or the burden of duties. Especially because a legal actor may be neither human nor animate, as when a corporate body has a legal personality, discussions of both the legitimacy and the content of the actor's legal personality may become quite controversial. Thus, the jurisprudential basis for an actor's legal personality is a matter of theoretical and historical concern.

Every legal system must decide which individuals and collectivities, whether formally or informally constituted, or anything else, have legal personalities and what the content of those legal personalities might be. Some legal systems assume that certain actors have

rights (see *Rights*) within the system based on inherent features of the actors, such as their humanity. Every legal system also grants certain actors not assumed to be able to act within the system the right, sometimes limited, to pursue their legal aims. This ascription of rights and duties upon an entity not presumed to be able to act within the system is its legal personality. Some legal systems, notably those premised on natural law (see *Natural Law*), but certainly not limited to such systems, have recognized that some actors have inherent legal personalities. In systems of positive law (see *Legal Positivism*), however, no quantum of rights inheres in any actor, at least in theory. Nonetheless, the history of legal personality is replete with examples of legal systems recognizing inherent rights in individuals, groups, and entities, whether their right to participate in the system is presupposed or granted as a matter of positive law.

The most prominent twentieth century example of an entity granted legal personality is the business corporation, though the history and theory of legal personality long antedate the business corporation. Religious, charitable, and governmental bodies are among the entities that have historically been granted legal personalities, especially insofar as legal systems have recognized distinctions between the rights of an office as distinct from the rights of the person occupying the office. Moreover, in legal systems characterized by hierarchies of actors not all human beings have been assumed to have full legal personalities. Therefore, women (see *Gender and the Law*), children (see *Children and the Law*), slaves (see *Slaves/Slavery, History of*), and foreigners, among others, have had to have their legal personality defined. Recently, the possibility of ascribing legal personality to animals and other entities has entered jurisprudential debates, largely in the hope that in doing so the legal system would necessarily find that they have rights that might be vindicated.

1. Legal Personality, Inherent Rights, and Positive Law

Two theoretical difficulties attend the concept of legal personality. The first is the class or classes of entities which have personality in the eyes of the law. The second is whether legal personality inheres in any entity and what the content of inherent legal personality might be. Systems premised in natural law of some variety, for example, assume both that individual human beings have an irreducible quantum of legal personality and that the content of that personality provides each individual with a certain level of protection and dignity. Such systems need not, however, treat all human beings as having equivalent legal personality. An aristocrat and a slave may each possess a legal personality, but they need not be equal. Nor do political systems have to limit their vision to an individualist conception of personality. Theories of organic polities, for example those premised in republican (see *Republicanism: Philosophical Aspects*) or communitarian ideologies, maintain that group identity is fundamental to social organization and should be recognized at law. Even societies with a largely individualist politics recognize the legitimacy of some group activity. Legal personality may therefore be presumed to inhere in collectivities within society, such as tribes, governments, churches, and even offices, such as governmental or religious offices. The difficulty in any such system is, of course, that human beings are required to discover both who and what might have a legal personality and then to discover the content of that personality.

Systems of positive law appear to solve the problem of legal personality present in systems which maintain beliefs in inherent legal personality. Under positive law legal personality does not inhere in anything or anybody but is rather the result of a grant by the sovereign authority. Hence, legal personality is a fiction created by the sovereign for instrumental purposes. The sovereign may decide to grant the right to act within the legal system to a person, to a class of persons, or to groups of people, constituted either formally or informally within the social structure. Indeed, the sovereign may grant the right to act to inanimate objects or to entities which are themselves the creation of the sovereign, such as corporate or subsidiary political entities. The difficulty presented by systems of positive law is, however, one of rights consciousness within a culture. That is, cultural conceptions of the dignity of human beings, either individually or collectively, the naturalness of human associations, or even conceptions of the inherent rights of, for example, animals, may war with the sovereign's willingness to grant legal personality, no matter how minimal.

In practice virtually no system is either purely one in which legal personality inheres in people or things or in which the sovereign jealously guards the prerogative of the grant of personality. Systems premised on an understanding of inherent legal personality find themselves constantly revising the classes of individuals and entities covered and constantly discovering new aspects of their legal personalities. Moreover, sovereigns within such systems also indulge in the power to grant legal personality. Systems of positive law regularly discover that they must recognise aspects of legal personality that are not spelled out within the grant of personality. Couched sometimes as inferences from the grant and sometimes as necessary components of any grant of personality, these redefinitions of legal personality depend upon changing cultural, especially changing political and economic, circumstances for their content, even though the content of the changes must be recognized by some arm of the sovereign, for example, the courts. The history of legal personality demonstrates, moreover, that conceptions of legal

personality have not neatly evolved from concepts of inherent personality to positive grants of personality. Instead, concepts of legal personality both depend upon and in turn influence surrounding circumstance.

2. Legal Personality in Jurisprudential History

The conflict between inherent and fictive conceptions of legal personality has a long history. Among the first instances in which the distinction between inherent and fictive legal personalities arose in western culture came during the thirteenth century during the papacy of Innocent IV. The Pope, attempting to consolidate the power of the Church and to harmonize society under its auspices, used the concept to delegitimate the authority of institutions inferior to the Church. As the Church was the agent of God, all collective human activity was subordinate to it. Thus, other institutions were inferior to the Church and owed their existence not to their independent authority but existed at the behest of the Church or by its grace. They were, therefore, legal fictions. The Church was not alone in its attempt to subordinate other human institutions by fictionalizing their existence.

The early modern state and its successors have also tried to centralize power by subordinating other human institutions in the name of sovereignty. For example, guilds could only be legitimate to the extent that they acknowledged their derivative status. Similarly, corporate bodies could maintain corporate status only by virtue of charters. Under the Bubble Act collective corporate activity in England could be prosecuted unless done under charter, though such prosecutions were evidently quite rare. In the United States the power of cities, towns, and other local entities derives from the individual states. Indeed, even the sovereign authority of the states and the other sovereignty recognized within the United States, that of the Indian nations, have their own tortured histories. The reason they are tortured, of course, is that modern political consciousness accepts Innocent's positivist precept, that authority to act within a legal system depends on a grant from the sovereign authority, turning the concept of a sovereign existing within another sovereign into a first-rank problem of political logic.

By way of contrast, the concept that legal personality is inherent stems both from aspects of natural law and political conceptions of an organic society. The religious traditions that informed natural law suggested that individuals possessed a minimum of dignity that others, the sovereign included, were bound to respect. Failure to respect human dignity imperilled one's soul. The history of the law of slavery is, for example, in many ways dependent on the distinction between the recognition of a slave's human dignity and the consideration of the slave's status as mere property. Parallel considerations animating organic politics legitimate recognition of groups and group activity. Some political theorists have long argued that social groups and collective activity are fundamental aspects of human life and therefore the recognition by a legal system of collective identity and the legal right of collective activity is not a function of sovereign grace but rather a function of the human condition. Such recognition has taken a variety of forms throughout history. Recognition of social and political class divided parliamentary bodies in England and France, whereas, at least in the pre-Civil War United States, one's loyalty to a state, defined formally as a geographic matter and informally as a matter of culture and politics, in part grounded the legitimacy of the federal system. Belief in the inherent legitimacy of group activity has informed spirited intellectual and political activity, such as Frederic Maitland's late nineteenth- and early twentieth-century desire to undermine certain individualist premises in the common law and to personify at law labor unions in an attempt to protect concerted activity by the working class in England.

In modern jurisprudence, however, no form of collective activity has caused more ink to be spilled in the debate over the nature of legal personality than the business corporation. Nowhere was the debate over the meaning of corporate personality as keenly debated in such a concentrated period as in the United States from the end of the second decade of the nineteenth century through the first third of the twentieth century. No mere academic dispute, the debate over corporate personality was framed by the twin considerations of society's power to promote and regulate economic activity on the one hand and the nature of collective activity, in this case entrepreneurial activity, on the other.

While English and American discussions of corporate personality existed before the nineteenth century, both in the common law and elsewhere, the first definitive exploration of the American position came in the Supreme Court's decision in *Dartmouth College vs. Woodward*, 17 US 518 (1819). In that case Chief Justice John Marshall wrote that a corporation 'is an artificial being, invisible, intangible, … possess[ing] only those properties which the charter of its creation confers upon it, either expressly, or as incidental to its very existence. These are such as are supposed best calculated to effect the object for which it was created.' The declaration of the Chief Justice appeared to put the United States squarely in the camp of positive law insofar as corporate personality was concerned. Considering that elsewhere in American jurisprudence the legal personality of white male property-owning citizens was assumed, it is, however, quite clear that neither a belief in inherent legal personality nor a purely positivist vision animated overall American legal thought. Indeed, even within the realm of corporate personality both the case generally and Marshall's own opinion left room for both visions.

Despite the clear import of Marshall's language, the case held that the sovereign could not unilaterally alter the charter of its own creation when that creation was a private entity, which the college was. Even within the realm of positive law, therefore, Innocent's vision was subtly altered. The sovereign could, of course, reserve the power to alter charters it granted, and the chartering states quickly proceeded in that direction by legislative act. Nevertheless, even without recourse to the legislature, Marshall had left open the potential expansion of corporate personality by recognition of the necessary incidents of corporate existence.

Dartmouth itself was, of course, an educational and not a business institution. In the United States, however, the expansion of the economy and the entrepreneurial fervor of the citizenry meant that individuals looking for vehicles under which to organize their capital for business purposes quickly adopted the corporate form as their preferred mode. States, eager to accommodate, granted corporate charters in ever greater numbers. As the corporate form grew to be the natural mode for organizing capital, Marshall's supposition that corporations were 'artificial being[s]' grew ever more suspect. Rather than beings created by the state, they seemed creations of entrepreneurs. These creations, moreover, did not have simply the rights granted by the sovereign, but had whatever properties necessary to accomplish their object. Business existed in environments which required properties of great flexibility in order to operate and to achieve their object. They needed, among many other things for example, the right to own property, both real and personal, including intellectual property; they needed the right to operate across jurisdictional lines in order to better exploit markets for their products; they needed to be able to exploit resources in different jurisdictions, be they natural resources, pools of labor, sources of power or transport, or even pockets of capital. In addition to finding the elementary aspects of legal personality, such as the right to sue and to be sued, courts began to find other rights incidental to corporate existence. What is equally interesting is the tension between the legislature, which had the right to set the terms of a corporation's legal personality, and the courts, upon which fell the interpretive burden of discovering rights which were incidental and through their discovery legitimating those rights. To be sure, legislatures often resisted granting certain rights and American courts moved in fits and starts in recognizing incidental rights, but the direction was clear. Marshall's qualification was, therefore, one of the ways in which the American jurisprudential image of the corporation reversed itself.

Well before the end of the nineteenth century, corporations asserted and courts recognized their changing legal personality. Though the object was clear, the protection of the property interests of the owners of corporations, the corporation's malleable legal personality was deployed as a jurisprudential weapon. At the time of the Dartmouth College judgment, corporations were rare, the state's imprimatur was thought to be useful in aggregating capital for risky and large ventures, and legislative consideration of corporate charters was real. A fictional understanding of the corporation's legal personality made sense, both to the state and to the individuals who sought a charter. By the last quarter of the nineteenth century, however, none of those conditions obtained. Corporations, especially business corporations, were common. Organizing entrepreneurial activity under corporate auspices was increasingly the norm. Capital markets had advanced to the point where the state's grant was largely irrelevant to the capacity to obtain capital and, as important, the state's stamp of approval was no longer felt to be necessary. Regulation of the behavior of corporations through their charters was largely nominal and the granting of charters was becoming an administrative rather than a legislative process. Indeed, the politics of regulation, always present, grew more contentious as the century progressed, especially as corporate activity made state and local regulation problematic. Regulation of behavior, regardless of legal form, increasingly became the norm and owners of corporations, in turn, sought to protect their property through new conceptions of corporate legal personality.

American legal treatises concerned with corporate law at first edged in the direction of abandoning the conception of the corporation as an entity with a legal personality of its own and instead made a brief attempt to see the corporation as the aggregation of the property of individuals and protected as would be the individual property of those individuals. Decades of jurisprudence dealing with corporate personality, the separation of owner and entity powerfully reinforced by the limited liability of corporate shareholders, and the willingness of courts to find an ever greater sphere of corporate rights suggested a different path. By the end of the century American jurisprudence, while still paying verbal homage to Marshall's understanding, had, in fact, begun to assume that many aspects of a corporation's legal personality were inherent. Ironically, the American jurisprudence of corporate personality, which proved to be politically conservative, fed on many aspects of progressive jurisprudence. Maitland's writings on personification of the corporation, for example, though designed to help the aspirations of English labor as part of a larger jurisprudential vision, abetted an understanding of the naturalness of corporate rights in the United States. Indeed, the use of the term legal person to describe the corporation carried with it strong implications of a legal personality the equivalent of other persons.

In American jurisprudence the debate over the personification of the corporation exhausted itself in the years immediately prior to the Second World War. In part the business corporation had achieved a legal

personality sufficient for its needs and the needs of its owners. Further dispute on the reality and inherent content of a corporation's legal personality seemed irrelevant. Corporations existed and their legal rights and duties gave them sufficient flexibility to operate, flourish, and protect themselves. Simultaneously, attacks on general conceptions of rights by legal realists and others in favor of concepts of legal interests that society ought to recognize turned out to be just as hospitable to corporations in an environment that valued economic productivity more highly than much else. Nonetheless, the instrumental value of asserting an inherent legal personality was far from lost. In the last 1980s, for example, the United States Supreme Court recognized free speech rights of corporate bodies premised on the First Amendment of the Constitution, going so far as to characterize as 'extreme' the position that corporations have only the rights granted to them by states.

Legal personality, as the history of corporate personality demonstrates, is a jurisprudential tool with an independent significance. Without a legal personality, no person, entity, or thing may actively participate in a legal system. Once having obtained a legal personality, either by the assertion of an inherent personality and the recognition of that personality by the appropriate political bodies, or by the grant of personality from the sovereign, the content of the personality is plastic but the capacity to participate, even if limited, is clear. By right of participation, the avenue to enlarging the content of that personality opens itself up to assertions of right and instrumental value.

See also: Children and the Law; Children, Rights of: Cultural Concerns; Corporate Law; Gender and the Law; Slaves/Slavery, History of

Bibliography

Dewey J 1926 The historic background of corporate legal personality. *Yale Law Journal* **35**: 655–73

Gierke O 1900 *Political Theories of the Middle Ages*. Cambridge University Press, Cambridge, UK

Hallis F 1930 *Corporate Personality*. Oxford University Press, London

Hurst J 1970 *The Legitimacy of the Business Corporation in the Law of the United States*. University Press of Virginia, Charlottesville, VA

Laski H 1916 The personality of associations. *Harvard Law Review* **29**: 404–26

Maitland F 1911 Moral personality and legal personality. *The Collected Papers of Frederic William Maitland*. Cambridge University Press, Cambridge, UK

Mark G 1987 The personification of the business corporation in American Law. *University of Chicago Law Review* **54**: 1441–83

Nekam A 1938 *The Personality Conception of the Legal Entity*. Harvard University Press, Cambridge, MA

Vinogradoff P 1924 Juridical persons. *Columbia Law Review* **24**: 594–604

Wolff M 1938 On the nature of legal persons. *Law Quarterly Review* **54**: 494–521

G. A. Mark

Legal Pluralism

The concept of legal pluralism refers in the most general sense to the situation in which the 'law' that obtains in a social field consists of more than one set of binding rules, whose behavioral requirements are different and sometimes conflicting. The concept plays a key role in modern sociology and anthropology of law (and to a lesser degree in related disciplines such as the economic analysis of law). It is one of the most important contributions that the sociology and anthropology of law have made to legal theory, demanding attention to unresolved questions concerning the concept of law itself and its relation to the nation-state, and affording an analytic link between 'sociological jurisprudence' (a recurrent theme in legal scholarship) and sociological analysis of the place of rules and social control in social life. It is also an important although marginal (because of the limitations of Western legal scholarship) concept in legal analysis (e.g., in connection with the role of 'custom' in liability in tort and contract).

The concept has two related but distinct meanings (Griffiths 1986); 'Juridical legal pluralism,' refers to a particular sort of legal arrangement, a way in which a legal system that is in principle unitary can deal with cultural and other sorts of diversity in its population by providing (or recognizing) distinct rules and procedures for specific groups (indigenous peoples, ethnic and religious groups, local communities, merchants and other economic groups, etc.). 'Empirical legal pluralism' refers to a social state of affairs in which an actor is confronted by different and possibly conflicting behavioral expectations (as when the norms among the participants in a particular economic activity differ from the applicable legal norms). Since in the case of juridical legal pluralism there are different legal rules for different situations, but only one rule for a given actor in a particular situation, empirical legal pluralism is not entailed; nor does empirical legal pluralism entail juridical legal pluralism, since the competing systems may themselves be monistic. The remainder of this article is devoted largely to legal pluralism in the empirical sense.

1. History of the Concept

Part of the ideological baggage of the modern nation-state is the idea that the state is the source of all law, properly so-called, and that law is (or at least ought to

be) exclusive of other forms of regulation and is uniform for all persons. When Western European expansion began to take the form of colonization, this conception of law confronted the existence of indigenous peoples whose existing law could not be ignored. Whether for reasons of international law (the right of conquered peoples to retain their own law) or of administrative convenience ('indirect rule'), the idea of legal pluralism as a form of governance was born: colonial law (q.v.) incorporated by reference (some parts of) the pre-existing 'laws and customs' of subjected peoples, rejecting others as 'repugnant' to morality and public order. Colonial legal pluralism brought forth a considerable body of law (legislation, case law, legal doctrine) and an extensive legal literature (Hooker 1975). See Van den Bergh (1986) for the prehistory of the concept of legal pluralism in Roman law, medieval and canon law, and early colonial law.

Attention to legal pluralism in the empirical sense was for a long time restricted largely to legal anthropologists. Anthropology of law (q.v.) arose under the wings of colonial administrations, producing a number of early classics on 'the law of the …,' conceived of in the first place as contributions to colonial administration. On the whole, until the second half of the twentieth century legal anthropology treated the 'law' of the societies with which it dealt as an autarkic phenomenon, 'recognized' but not influenced by, or interacting with, the colonial state and its 'law.' The emancipation of legal anthropology from its subservience to the concerns of colonial administration led to a fresh start in which the legal order of a society, at all levels, is seen as a dynamic and complex whole.

The intellectual roots of modern, empirical legal pluralism can be traced back to Ehrlich (1913) who insisted that the 'rules for behavior' that actually govern social interaction are the norms that constitute the 'inner order of the associations' of which a society consists, whereas the law of the state consists of 'rules for decisions' addressed to courts and other legal decision-makers. But it was only in the second half of the twentieth century that the fact that the behavior of social actors is generally subject not just to one set of more or less binding rules ('law'), but to a variety of them, came to be generally accepted among sociolegal scholars as an essential theoretical and methodological insight. Scholars made sensitive by the colonial experience and the process of decolonization to the possibility of legal diversity began to treat the idea as fundamental in sociolegal theory and research.

In the 1960s, the concept of legal pluralism served as the rallying cry for a new approach to the study of law in society. Vanderlinden's essay '*Le pluralisme juridique: essai de synthèse*' marked the transition from preoccupation with the relationship between colonial law and local customary law to the realization that legal pluralism is characteristic of modern, industrialized societies too. Vanderlinden defined legal pluralism as '*L'éxistence, au sein d'une société déterminée, de méchanismes juridiques différents s'appliquant a des situations identiques*' (1971, p. 19). Armed with such a definition, and with an increasingly loose, encompassing conception of 'law,' legal pluralism came to seem ubiquitous. If one looked carefully it could be found in the business world, bureaucratic organizations, schools, neighborhoods, and so forth. Recently, the supranational arena (the European Union, international organizations, international trade and economic globalization) has also come to be recognized as an important locus of legal pluralism.

In the 1980s, scholarly interest in legal pluralism became institutionalized. The Commission on Folk Law and Legal Pluralism of the International Union of Anthropological and Ethnological Sciences was founded in 1978. It organizes regular international symposia, and publishes books, proceedings, and a *Newsletter*. There are two international scholarly journals devoted to the field: the *Journal of Legal Pluralism and Unofficial Law*, founded in 1981 as a continuation of *African Law Studies*, and *Law and Anthropology*.

The usual variety of approaches can be found among those who use the concept of legal pluralism in the empirical sense. These include Marxism (e.g., Fitzpatrick 1980, Santos 1977, Snyder 1981a), feminism (e.g., Griffiths 1997), interpretive, constructivist and 'postmodern' approaches, and eclecticism that straddles the line between social science and a normative, legal approach. Historians have made some important contributions (e.g., Thompson 1991). Most writing on legal pluralism has been from a Western perspective, but Chiba (1989) is a notable exception.

2. *Critique of the Empirical Concept of Legal Pluralism*

The juridical concept of legal pluralism is generally accepted and has never given rise to any significant disagreement. The empirical concept, by contrast, despite the fruitful role it has played in social scientific research and theory, has given rise to some controversy (Tamanaha 1993, Woodman 1998). The objection is in particular to use of the word 'legal,' which, it has been argued, is irredeemably normative, carrying with it etymological associations (with 'legality' and, in languages such as German and Dutch, with the idea of justice) that inevitably give rise to confusion.

While well-taken as far as it goes, the criticism seems to require us to formulate the interaction captured in the concept of legal pluralism as one between two different things: 'law' and non-'legal' social control. Such an opposition entails just the misleading legal-positivist (q.v.) assumption that invention of the unifying concept of legal pluralism was intended to overcome: that 'law' can be considered as theoretically

and empirically distinct from other forms of social control. Use of the concept of legal pluralism reflects the conviction that the differences between 'legal' and other forms of social control are matters of continuous variation and that the factors that affect the impact of rules on behavior are the same in both cases. To accomodate the criticism of the concept of legal pluralism without returning to a sterile opposition between 'law' and other social control, one could choose, for empirical purposes, to speak of normative pluralism or 'rule-system pluralism' (Tamanaha 2000, p. 306). Another line of criticism, that the concept of legal pluralism ignores the importance of differences in power (see Starr and Collier 1989), seems in light of the literature plainly mistaken.

3. *Use of the Empirical Concept in Empirical Theory and Research*

From the outset, the most active area of empirical research and theory in which legal pluralism plays a central role has concerned dispute processes. Holleman's (1973) insistence that the social scientific study of law must concern itself not only with 'trouble cases' but also with 'trouble-less cases' of law use, together with the interest in the 1970s in 'law and development' (q.v.), led to the growth of a second major area of interest: the social working of law under circumstances of legal pluralism. The effects of a situation of legal pluralism on the functioning of state and nonstate institutions such as courts, is a third area of interest. Finally, attention had been paid to the ecclectic and opportunistic way that people make use of a situation of plural legal resources in constructing relationships and complex social institutions such as 'property,' 'social security,' and 'international trade.'

3.1 *Dispute Processes (Litigation and Adjudication)*

Legal pluralism is an important factor in understanding what people have disputes about, how disputes emerge in social life, what choices are made about how to deal with them, how they progress, what ultimately happens with them, and how the disposition of a particular dispute affects the way other people deal with similar problems. At all stages of a (potentially) 'legal' dispute, the existence of non-'legal' norms, procedures, and institutions profoundly influences what happens. Building on the general literature on dispute processes, or litigation (q.v.), a number of scholars have devoted particular attention to the normatively plural context in which disputes often take place. Exemplary studies are Ellickson (1991) on cattle-trespass disputes between neighbors in rural California; Todd 1978, on disputing relationships and social status in a Bavarian village; Santos (1977) on informal law and disputing in a Brazilian squatter settlement; and Abel (1979) on the introduction of Western courts in the non-Western setting of British Africa. The ways in which adjudicatory institutions adjust to the choices their clientele make has attracted particular interest K. von Benda-Beckmann (1984), in particular, has used the 'extended case' method of legal anthropology to good effect, showing how 'forum-shopping' litigants make differential use of state and nonstate law in different phases of a dispute, so that the execution of a court judgment, for example, may be regulated in practice by nonstate norms. Galanter (1981) and Griffiths (1983) have sought to integrate the perspective of legal pluralism into general formulations of litigation theory.

3.2 *Social Working*

Research, in both Western and non-Western settings has shown repeatedly that legal pluralism can be the key to understanding the effectiveness of a legal rule. Moore (1973) argued that the ways in which legal rules do or do not influence behavior can only be understood if one considers them from the perspective of the actors to whom they are addressed. These actors are generally exposed not only to the behavioral expectations expressed in legal rules but also to the often quite different expectations of more immediate sources of regulation to which they are exposed ('semi-autonomous social fields').

The idea that the relationship between local law and state law is that the former frustrates the effectiveness of the latter is, however, too simple. In the first place, as Moore observes, when external law is effective in a social field this is often thanks not to the enforcement capacities of the state but to the support legal rules receive from local social control (see Kagan and Skolnick 1993 for the example of rules regulating smoking). In the second place, state law often adapts and builds upon forms of regulation that have earlier emerged in, and are enforced by, local social fields. The Dutch experience with regulation of medical practice surrounding death (Griffiths 2000) is a good example of the process by which such self-regulation (in this case, by the medical profession) can later be adopted by legal institutions and enforced as 'law.'

3.3 *Interaction of State and Nonstate Law and Institutions*

A phenomenon of special interest concerns the 'creation of customary law': the fact that when local law is 'recognized' by state law (as in the case of colonial legal pluralism), it must be ascertained and applied by noncustomary tribunals. In the process, a great deal of transformation takes place, so that the 'customary law' applied in state courts may have only a distant relationship to its supposed sources (Snyder 1981a, Woodman 1988).

A related phenomenon is the effect on local political authority when it is incorporated into a (post) colonial system of government. Such incorporation brings with it notorious errors in the identification of indigenous authorities, the witting or unwitting attribution to them of various nonindigenous powers, the manipulation of indigenous processes of succession to high office, and so forth. The fundamental bases of legitimacy tend to be corrupted and indigenous mechanisms of control over political authority eroded. This is a central theme in a number of studies of local political authority (e.g., Holleman 1969, Owusu 1970). The more recent literature emphasizes that legal pluralism can be the idiom in terms of which political struggles are fought. Asserting or rejecting the validity of a rule of particular provenance, or attributing particular virtues or defects to particular sorts of law (customary, religious, or state law), can be ways of claiming political authority (e.g., Benda-Beckmann 1989).

3.4 Eclectic Use of Normative Resources

Colonial legal pluralism was characterized by a 'choice of law' approach: either state law or customary law applied to a given case, and if customary law was applicable, then it was either the law of group A or that of group B. It took some time before scholars working with the newer empirical conception of legal pluralism came to appreciate that social reality is generally not either/or but both/and: actors with access to multiple sources of law construct legal relationships ecclectically, fashioning a patchwork composed of bits and pieces of local law, religious law, state law, and international law. Thus property and inheritance relationships in Minangkabau combine elements of Indonesian law, local adat law, and Islamic law (cf. Benda-Beckman 1979).

Eclecticism in the choice and promotion of legal sources is a well-known phenomenon in the case of dispute processes (see Sect. 3.1 above). A good example of eclecticism in the structuring of relationships concerns social security. The social security arrangements that people in developing countries construct for themselves include elements of state law and elements of customary and religious law, and if they migrate to a Western country its formal social security arrangements get included in the mixture (Benda-Beckman et al. 1988). Since intra-family income transfers continue to be very important in industrialized countries, it seems safe to assume that here, too, normative eclecticism is a crucial feature of the actual social security many people depend on.

In the area of economic globalization, Snyder has argued that 'global economic networks are governed by the totality of strategically determined, situationally specific, and often episodic conjunctions of a multiplicity of institutional, normative, and processual sites throughout the world.' A patchwork of what he calls 'global legal pluralism' results from the behavior of 'strategic actors [whose choices have] been absolutely fundamental in determining which … sites have seen the light of day, which have flourished and developed, and which have withered and even died for lack of clients' (Snyder 1999, p. 371).

See also: Aboriginal Rights; African Legal Systems; Conflict and Conflict Resolution, Social Psychology of; Disputes, Social Construction and Transformation of; Federalism; Feud and Internal War: Legal Aspects; Folk, Indigenous, and Customary Law; Hindu Law; Jewish Law; Law and Development; Law and Society: Sociolegal Studies; Law as an Instrument of Social Change; Law, Mobilization of; Law: Overview; Laws, Conflict of; Legal Culture and Legal Consciousness; Legal Positivism; Legal Systems, Classification of; Legal Systems: Private; Legalization; Lex Mercatoria; Litigation; Pluralism; Pluralism and Toleration; Relativism, Pluralism, and Skepticism (Philosophical Perspective): Cultural Concerns; Remedies and Damages: Legal Aspects; Value Pluralism

Bibliography

Abel R 1974 A comparative theory of dispute institutions in society. *Law and Society Review* **8**: 217–347

Abel R 1979 Western courts in non-western settings: Patterns of court use in colonial and neo-colonial Africa. In: Burman S, Harrell-Bond B (eds.) *The Imposition of Law*. Academic Press, New York, pp. 167–200

Arthurs H W 1985 *Without the Law. Administrative Justice and Legal Pluralism in Nineteenth-Century England*. University of Toronto Press, Toronto, Canada

Benda-Beckmann F von 1979 *Property in Social Continuity: Continuity and Change in the Maintenance of Property Relationships Through Time in Minangkabau, West Sumatra*. Martinus Nijhoff, The Hague, The Netherlands

Benda-Beckmann F von 1989 Scapegoat and magic charm: Law in development theory and practice. *Journal of Legal Pluralism* **28**: 129–48

Benda-Beckmann F von 1998 Citizens, strangers and indigenous peoples: Conceptual politics and legal pluralism. *Law and Anthropology* **9**: 1–42

Benda-Beckmann F von et al. 1988 *Between Kinship and the State: Social Security and Law in Developing Countries*. Foris, Dordrecht, The Netherlands

Benda-Beckmann K von 1984 *The Broken Stairways to Consensus: Village Justice and State Courts in Minangkabau*. ICG, Dordrecht, The Netherlands

Benda-Beckmann K von 2000 Transnational dimensions of legal pluralism. In: Fikentscher W (ed.) *Begegnung und Konflikt: eine kulturanthropologische Bestandsaufnahme*. C. H. Beck, Munich, Germany

Bergh G C J J van den 1986 The concept of folk law in historical context: A brief outline. In: Benda-Beckmann K von, Strijbosch F (eds.) *Anthropology of Law in the Netherlands: Essays on Legal Pluralism*. Foris, Dordrecht, The Netherlands, pp. 67–89

Chiba M 1989 *Legal Pluralism: Toward a General Theory Through Japanese Legal Culture*. Tokai University Press, Tokyo

Ehrlich E 1913 *Grundlegung der Soziologie des Rechts*. Duncker & Humblot, Munich, Germany
Ellickson R C 1991 *Order Without Law. How Neighbors Settle Disputes*. Harvard University Press, Cambridge, MA
Fitzpatrick P 1980 *Law and State in Papua New Guinea*. Academic Press, London
Galanter M 1974 Why the 'haves' come out ahead: Speculations on the limits of legal change. *Law and Society Review* **9**: 95–160
Galanter M 1981 Justice in many rooms: Courts, private ordering and indigenous law. *Journal of Legal Pluralism* **19**: 1–47
Griffiths A M O 1997 *In the Shadow of Marriage: Gender and Justice in an African Community*. University of Chicago Press, Chicago
Griffiths J 1983 The general theory of litigation—a first step. *Zeitschrift für Rechtssoziologie* **5**: 145–201
Griffiths J 1986 What is legal pluralism? *Journal of Legal Pluralism* **24**: 1–50
Griffiths J 2000 Self-regulation of medical behavior that potentially shortens life. In: Krabbendam H, Ten Napel H-M (eds.) *Regulating Morality: A Comparison of the Role of the State in Mastering the Mores in the Netherlands and the United States*. Maklu, Antwerp/Apeldoorn, The Netherlands, pp. 173–190
Holleman J F 1969 *Chief, Council and Commissioner*. Van Gorcum, Assen, The Netherlands
Holleman J F 1973 Trouble-cases and trouble-less cases in the study of customary law and legal reform. *Law and Society Review* **7**: 585–609
Holleman J F 1981 *Van Vollenhoven on Indonesian Adat Law: Selections form Het Adatrecht van Nederlandsch-India*. Nijhoff, The Hague, The Netherlands
Hooker M B 1975 *Legal Pluralism: An Introduction to Colonial and Neo-colonial Laws*. Clarendon Press, Oxford, UK
Kagan R, Skolnick J 1993 Banning smoking: Compliance without enforcement. In: Rubin R, Sugarman S D (eds.) *Smoking Policy: Law, Politics and Culture*. Oxford University Press, Oxford, UK, pp. 69–94
Moore S F 1973 Law and social change: The semi-autonomous social field as an appropriate subject of study. *Law and Society Review* **7**: 719–46
Owusu M 1970 *The Uses and Abuses of Political Power. A Case Study of Continuity and Change in the Politics of Ghana*. University of Chicago Press, Chicago
Santos B de Souza 1977 The law of the oppressed: The construction and reproduction of legality in Pasagarda. *Law and Society Review* **12**: 5–126
Sawyerr G F A 1972 The 'choice of law' approach and the application of law in Ghana. *University of Ghana Law Journal* **9**: 173–81
Snyder F 1981a Colonialism and legal form: The creation of 'customary law' in Senegal. *Journal of Legal Pluralism* **19**: 49–90
Snyder F 1981b *Capitalism and Legal Change: An African Transformation*. Academic Press, New York
Snyder F 1999 Governing economic globalization: Global legal pluralism and European law. *European Law Journal* **5**: 343–74
Starr J, Collier J (eds.) 1989 *History and Power in the Study of Law*. Cornell University Press, Ithaca, New York
Tamanaha B 1993 The folly of the social scientific concept of legal pluralism. *Journal of Law and Society* **20**: 192–217
Tamanaha B 2000 A non-essentialist version of legal pluralism. *Journal of Law and Society* **27**: 296–321
Teubner G (ed.) 1997 *Global Law Without a State*. Dartmouth, Aldershot, UK
Thompson E P 1991 *Customs in Common*. Merlin Press, London
Todd H F 1978 Litigious marginals. Character and disputing in a Bavarian village. In: Nader L, Todd H F (eds.) *The Disputing Process—Law in Ten Societies*. Columbia University Press, New York, pp. 86–121
Vanderlinden J 1971 Le pluralism juridique: essai de synthése. In: Gilissen J (ed.) *Le Pluralisme Juridique*. Université de Bruxelles, Brussels, Belgium, pp. 19–56
Woodman G R 1988 How state courts create customary law in Ghana and Nigeria. In: Morse B W, Woodman G R (eds.) *Indigenous Law and the State*. Foris, Dordrecht, The Netherlands, pp. 181–22X
Woodman G R 1998 Ideological combat and social observation: Recent debate about legal pluralism. *Journal of Legal Pluralism* **42**: 21–59

J. Griffiths

Legal Positivism

Legal positivism is the label attached to a widely accepted, though hotly disputed, theory about the nature of law. Within jurisprudence, legal positivism is most often discussed by way of contrast to natural law theories of law. The contest between these rival traditions has long dominated the discussion in legal philosophy, though there are signs that the debate may have run its course. Because legal positivism originated as a reaction to natural law theories about law, the best way to understand legal positivism is through this debate. The following section will describe the nature of the debate; the section thereafter will set out the core elements of legal positivism; the final section will discuss the reaction to, and significance of, legal positivism, from the standpoint of social scientific approaches to law.

1. The Debate between Legal Positivism and Natural Law

Ironically, Hobbes, who is widely credited to be the early progenitor of legal positivism, built his positivistic theory of law on a natural law base (Bobbio 1993, Fuller 1940). According to Hobbes, the necessity for peace and order is the fundamental precept of natural law. Given that the state of nature is, in his immortal phrase, 'solitary, poore, nasty, brutish, and short,' people join together and give over law-making and enforcing power to a Sovereign for the purpose of maintaining peace and order (Hobbes 1946, Chaps. XIV, XV). Thereafter, law consists of the commands of the Sovereign. Although the Sovereign is to be guided by reason in the establishment of law, even unjust commands are valid laws; to suggest otherwise would invite the disorder that will result if people were

allowed to apply their own judgment to evaluate the law. Under Hobbes' formulation, while natural law provides the initial foundation for positive law, once that foundation is established, natural law becomes irrelevant. It does not provide a continuing standard or test for the validity of the law. (See *Natural Law*; *Justice and Law*; *Injustice: Legal Aspects*.)

This conclusion raises the issue that has long set legal positivists and natural law theorists at odds. Thomas Aquinas, citing Augustine, asserted that 'A law that is not just, seems to be no law at all' (Q. 96, A. 4, Obj. 3). 'Consequently every human law has just so much of the nature of law, as it is derived from the law of nature. But if in any point it deflects from the law of nature, it is no longer a law but a perversion of law' (Q. 95, A. 2, Obj. 4). A similar statement was made by William Blackstone. '[N]o human laws are of any validity, if contrary to this [law of nature]' (Boorstin 1996, p. 49). More recently, Lon Fuller asserted that there is a minimum moral element to law, without which it is 'not simply bad law, but not law at all.' (Boorstin 1999, p. 197).

Legal positivists are adamantly opposed to this position. Austin, who produced the first widely influential articulation of the legal positivist position, wrote a scathing response to Blackstone. 'Now, to say that human laws which conflict with the Divine law are not binding, that is to say, are not laws, is to talk stark nonsense. The most pernicious laws, and therefore those which are most opposed to the will of God, have been and continually are enforced as laws by judicial tribunals' (Austin, 1995, p. 158). 'The existence of law is one thing; its merit or demerit is another.' What law *is*, and what law *ought* to be, legal positivists insist, are separate questions.

Aside from the general confusion it engenders, the failure to maintain the separation between *is* and *ought* raises two particular political concerns for legal positivists (Hart, 1958, pp. 597–98). The first concern involves the persons who conclude, based upon his or her understanding of natural law, that the offensive positive law is invalid and, therefore, can be freely disregarded. An example of this is the radical anti-abortionist who invokes natural law to justify illegal actions—including the execution of doctors who perform abortions—on the grounds that laws allowing abortion are invalid because they condone the killing of innocent lives. The second concern is that the identification of law with morality will lead people to be less critical of the law, their vigilance against it softened by the seductive assumption that the law *is* what it *ought* to be. Many legal positivists believe that the fact that a norm happens to have the status of 'law' does not of itself give rise to an obligation of fidelity (Schauer 1996). An obligation arises only after the law passes moral muster. Thus, legal positivists assert, maintaining the *is*/*ought* distinction allows us to more critically evaluate the law from a moral standpoint. These two concerns, it should be observed, while not contradictory, are not entirely consistent, in so far as the first worries about people choosing to disregard the law on the grounds that it is invalid, while the second worries about people slavishly following the law because they assume it is necessarily morally worthy. These two concerns, pointing as they do in different directions, simultaneously expose legal positivism to the charge of being excessively conservative as well as the charge of being excessively radical.

The second concern, more so than the first, animates many legal positivists in their dispute with natural law theorists, and their concern is not a purely abstract one. In the aftermath of World War II, some natural law theorists blamed the legal positivist separation of law and morality for the failure of the German legal establishment to resist atrocities committed by the Nazi government through law (Hart 1958, pp. 615–21, Rommen 1998). The view that 'law is law' regardless of its immorality, according to Gustav Radbruch, led the German legal profession to be completely subservient to state legal power. The legal positivist response, besides questioning whether there is any empirical support for the asserted causal connection, is that this view fails to appreciate that the separation of law from morality has the effect of reducing the prestige of law and thereby exposing it to greater moral critique. Under this view, the fact that a pronouncement is valid law carries little weight, and an immoral law (valid as it may be) should be resisted.

When viewed in these terms, the longstanding debate between legal positivists and natural law theorists is a narrow one. Legal positivists argue that a positive law is law, regardless of whether it is inconsistent with morality. Nevertheless, they immediately assert, a valid law can and should be evaluated from a moral standpoint, and not considered obligatory if found wanting. Many natural law theorists have asserted that consistency with natural law is a condition of the validity of positive law. An immoral law is, therefore, not 'law' at all, and carries no obligation (except perhaps is instances when a greater good might still be served by adherence to a bad law). However, anything that qualifies as law pursuant to this understanding automatically generates an obligation to comply. Whichever direction one approaches from, in the end both sides agree that an immoral positive law (whether or not characterized as 'law') generally does not give rise to any obligation to obey.

After generations of disagreement, there are now signs of a rapprochement. Finnis, the most influential modern exponent of natural law theory, argued that Aquinas was actually an early legal positivist, in the sense that he recognized that immoral laws were still in fact 'laws,' albeit deviant or deficient forms of law (Finnis 1980, 1996). A leading legal positivist, MacCormick (1992), has acknowledged that there are certain respects in which law or legality contains moral qualities, though he emphasized that this does not

vitiate the separation thesis. Indeed, at base, only the separation thesis separates these two schools. If natural law theorists, like Finnis, are willing to acknowledge that an immoral law is still 'law' in positive law terms—a point that a number of natural law theorists have not actually contested—there are few reasons to maintain the mutual opposition. There is no inherent incompatibility between these two approaches. A legal positivist may, without contradiction, believe in natural law moral theory, and apply the moral precepts produced by this theory to evaluate the law.

To be sure, not everyone agrees that the debate is over, but the bulk of the discussion of late has shifted to other matters, away from the issue of validity. One aspect of this classical debate remains unresolved, however. Although it has attracted relatively little attention, it is of continuing relevance and bears mention. The primary focus of the debate has been on whether immoral laws are invalid; the reverse question this raises is whether natural law principles have the status of law even if not affirmatively recognized by the positive law. If the answer is yes, then the Nuremberg Trials, and more recently the trials of the German border guards prosecuted for shooting people attempting to escape former East Germany, were routine applications of pre-existing natural law principles with direct effect. If the answer is no, then (ignoring weak arguments that there were pre-existing applicable positive law prohibitions based upon general declarations of human rights) these trials can be condemned as extralegal impositions of power by the ultimate victors in a political battle. Furthermore, if the answer is no, a strong argument can be made that these trials themselves violated the natural law prohibition against the *ex post facto* imposition of law, criminalizing and punishing behavior that was not illegal at the time it was committed. The issue of the legitimacy and legality of these kinds of actions, with serious unresolved moral as well as practical implications, will continue to arise as long as there is an attempt to hold government officials responsible for putatively legal conduct committed while in office.

2. *The Core Elements of Legal Positivism*

Awareness of the context within which legal positivism arose and developed is essential to an understanding of the theory, but it does not capture nor exhaust the insights contained within the theory. As with every theoretical tradition, there are many varieties of legal positivism. Of late, as the dispute between legal positivism and natural law has moderated, internal divisions within legal positivism have taken center stage in the theoretical discussion. The nuances of the internal debate need not be explored. It takes place within shared parameters that characterize most legal positivists, and these parameters are what matter. Before describing these parameters, it should be noted that there have been two dominant versions of legal positivism in the twentieth century, the first by Kelsen and the second by Hart. The essential difference between the theories lies in their orientation (Bix 1996, Chap. 6): Hart had an empirical bent, whereas Kelsen constructed an abstract system consisting of a hierarchy of norms (Kelsen 1967). The focus in this discussion will be on Hart's version, because, owing to its greater clarity and persuasiveness, it has eclipsed Kelsen's version substantially.

Hart's *The Concept of Law* (1961) is without question the most important legal positivist text, and probably the most influential book in the field of jurisprudence in the second half of the twentieth century. Hart's project was to produce 'a theory of what law is which is both general and descriptive' (Hart 1994, p. 239). His theory was 'general' in the sense that he intended to clarify the nature of law as a social phenomenon found in many different cultures and societies, by way of contrast to a 'particular' jurisprudence, which is focused on a specific legal system. His text was 'descriptive' in the sense that he eschewed evaluative analysis, limiting himself to describing practices and clarifying concepts related to law. He characterized his standpoint as that of an external observer describing the ways participants engage in and understand law from an internal point of view. To emphasize his empirical orientation, Hart described his book as 'an essay in descriptive sociology' (Hart 1961, p. v).

According to Hart, a legal system consists of the union of what he called primary and secondary rules. Primary rules are the rules that govern relations within society. Secondary rules are the rules legal officials follow to identify, apply, and change the primary rules. Hart famously labeled the core secondary rules the 'rules of recognition,' 'which specify some feature or features possession of which by a suggested rule is taken as conclusive affirmative indication that it is a rule of the group to be supported by the social pressure it exerts' (Hart 1961, p. 92). According to Hart, the emergence of secondary rules is what separates prelegal (or primitive) societies from societies with law. Although simple societies with only primary rules can exist, they are static and inefficient in responding to new situations or bringing about desired change. The availability of secondary rules cures these defects.

Thus, secondary rules identify the primary rules; the question that remains is how to identify the secondary rules. Hart's answer to this—in contrast to Kelsen's abstract identification of the ultimate *Grundnorm* as the source of all other norms—is that the rules of recognition can be identified by observing the *actual social practices* of the legal officials when they identify law. The existence and nature of the rules of recognition in a given legal system is thus 'a matter of

fact.' Based upon this resort to existing social practices, theorists have labeled Hart's approach to law as 'conventionalist.' Law is whatever the legal officials indicate by their social behavior they recognize as law, with no content restrictions on what that might be. Hart added one further requirement to the secondary rules. He was concerned to distinguish law from mere habits of obedience, and from the 'Gunman Writ Law,' which was the impression left by Austin's command theory of law. Hart felt that Austin's coercive view of law failed to recognize that law contains a normative aspect. This normative aspect, according to Hart, can be recognized only if one takes into consideration the internal point of view of legal officials. From this perspective, it is evident that legal officials *accept* the secondary rules as 'common standards of official behavior and appraise critically their own and each other's deviations as lapses' (Hart 1961, p. 113). Consistent with his nonevaluative stance, Hart specifies the necessity for the officials' acceptance of the secondary rules without suggesting whether this acceptance is merited (which can only be determined by examining particular legal regimes).

Hart concluded that there are two minimum conditions necessary and sufficient for the existence of a legal system. The first condition is that the primary rules must be generally obeyed by the populace, though it is not essential that the populace normatively accept the primary rules. It does not appear to make much sense to say that a legal system exists if the bulk of the populace generally disregards it. The second condition is that the legal officials accept the secondary rules as critical standards for their behavior. Whenever these two conditions obtain, a legal system will exist.

With its simplicity, clarity, elegance, and scope, Hart's reconstruction revitalized legal positivism, and spawned an enormous response, in support and against. As a consequence of his work, virtually all legal positivists (in the Hartian tradition) adhere to two core theses about the nature of law: the Separation Thesis and the Social Sources Thesis. The Separation Thesis holds that there is no necessary relation between law and morality (Coleman 1982, Coleman 1996). The genesis of this thesis has already been identified in the earlier described insistence of legal positivists that the fact of legality alone does not carry with it the imprimatur of moral rectitude. It is important to recognize that this is a limited, negative thesis, a denial of the natural law position, rather than an affirmative statement of the relationship between law and morality. Legal positivists are not insisting that there are no connections between law and morality, only that there is no *necessary* connection. They recognize that the 'law of every modern state shows at a thousand points the influence of both the accepted social morality and wider moral ideals.' (Hart 1961, p. 199). This influence comes in a myriad of ways. Law and legal norms often reflect (or claim to reflect) the morality of its community. Legal norms often expressly refer to and incorporate moral principles like justice, fairness, reasonableness, and due process. Law often has positive moral consequences, especially in its role in maintaining order in society. Law also often helps perpetuate moral norms, and participates in the moral education of members of the community, when it adds a legal sanction to support moral norms that have been recognized by law. Recognition of these intimate connections between law and morality is consistent with the Separation Thesis. The key qualifier in the above list is *often*. The Separation Thesis insists only that it is not necessarily the case that a moral norm, by virtue of that fact, automatically becomes a legal norm, and that it is not necessarily the case that an immoral law is, by virtue of that fact, not valid law, although both possibilities may be the case in a given legal system if the positive law of that system so determines.

The Social Sources Thesis holds that what is or isn't valid law in a given system, as determined by the rules of recognition, is a social fact. By observing the practices of the legal officials within a legal system, it will be possible to identify the criteria applied by the legal officials to determine what is valid law. These practices often include, for example, treating legislative enactments and judicial precedents as valid sources of law.

Although these two theses are shared by legal positivists, the broad consensus in support of them conceals a substantial internal divergence among legal positivists about what these theses represent and allow or disallow. Much of the internal disagreement has been generated by attempts by Hart and other legal positivists to respond to the sustained critique of legal positivism by Dworkin. This complex internal and external debate cannot be recounted here, though its core themes can be articulated summarily.

Dworkin has attacked Hart at several different levels and points (Dworkin 1977, Dworkin 1985, Dworkin 1986). He denies the very possibility or value of general, external, purely descriptive jurisprudence. Dworkin insists that jurisprudence itself is an aspect or extension of the law of a particular legal system and must, therefore, be conducted from the particular, internal perspective, which is inherently evaluative in orientation. Against the substance of Hart's argument, Dworkin denies the Separation Thesis by arguing that law and morality are inextricably intertwined, with moral principles providing the unseverable backdrop for the interpretation and application of legal rules. Based upon this view of the intimate connection between morality and law, Dworkin also denies Hart's assertion that there are points at which the law runs out. According to Dworkin, law incorporates a large body of moral norms and principles, and, understood as such, law always provides, a single correct answer to legal problems. Dworkin further argues that the social practice aspect of Hart's rules of recognition fails to accurately describe our own legal system, which

expressly incorporates moral norms (like due process and fairness) in a manner that renders the identification of the primary rules dependent upon difficult and controversial questions about the truth of moral norms. He identifies two specific problems with Hart's notion of rules of recognition. First, rules of recognition are incapable of identifying all of the many and varied moral norms and principles incorporated into the law. Second, Dworkin insists that legal positivists cannot, consistent with their theory of law, allow the incorporation of such norms and principles because the uncertainty which would result—making the identification of law contingent upon the resolution of complex moral questions—defeats legal positivists' claim that the value of secondary rules is their ability to identify primary rules in a way that provides guidance to the populace.

This abbreviated recitation is a partial selection of Dworkin's criticisms of Hart and legal positivism, focusing on the themes described earlier, and inevitably does not do justice to his arguments. A number of the issues separating Hart and Dworkin can be resolved by the acknowledgment on both sides that general (Hart) and particular (Dworkin) jurisprudence are two different enterprises which need not necessarily conflict. The most significant remaining controversy relates to the extent to which the rules of recognition can refer to open-ended moral questions that turn the issue of what law is on questions of content (rather than just social practices). Hart insists that there are no restrictions on the rules of recognition that would prohibit a particular legal system from turning the identification of law on the resolution of moral questions, as long as that is what the actual social practice constituted by the applicable rules of recognition requires. Hart labels his view 'soft positivism' (Hart 1994, pp. 250–4, Coleman 1998). With regard to whether this defeats the point of legal positivism, Hart (1994, p. 251) responds that positivism tolerates a certain degree of uncertainty in the identification of the primary rules, and furthermore that the resort to moral standards would not necessarily result in an excessive amount of uncertainty.

In theory, soft positivism would still be a correct description of a legal system that has only a single rule of recognition which says: do justice (Coleman 1998). Once this 'inclusive' view of legal positivism is taken, it is no longer evident that, as with the debate between natural law and legal positivism, anything of great moment rests upon the resolution of the debate over legal positivism between Hart and Dworkin. However, contrary to Hart, there are 'hard' or 'exclusive' legal positivists, led by Raz, who do not concede Dworkin's position about the inclusion of open questions of morality within law (Raz 1979). Now much of the debate is an internal one between hard and soft legal positivists (Bix 1999, Leiter 1998, Coleman 1998). Although hard positivists recognize that the law often refers to such norms, they assert that at that moment the law runs out and the judge's decision is based on extralegal considerations. The moral considerations to be resolved are not a part of the pre-existing law before the issue is decided, though they can thereafter become law if recognized and treated as legal precedent (consistent with the applicable rules of recognition). Hard positivists believe it is essential to take this position if legal positivism is to satisfy its claims that the function of law is to provide guidance to the public on the rules of conduct, and if law is to claim independent authority, apart from morality.

Once again, it would appear that the issue that divides the disputants is a relatively narrow one. Neither side denies that often the law refers to moral norms (like justice and fairness) which turn on the resolution of issues of substance or content. At that point the soft legal positivist says the judge is still doing law, while the hard legal positivist says that the law has run out. This is not to suggest that the debate is insignificant. When considering these issues, legal philosophers are attempting to come to an understanding of the nature of law. It is important to recognize that, as with several of the debates discussed above, the resolution will not be found in resort to the facts, since there is not substantial disagreement about the facts, but rather over how these facts are to be described from a theoretical standpoint. Despite the avowedly nonnormative stance of legal positivists, it would appear that the resolution of many of these debates is contingent on the outcome of evaluative questions, like which concept of law and its relation to morality is not just descriptively accurate but also morally optimal.

3. *Legal Positivism and Social Science*

Legal positivism generally, and Hart's theory in particular, has an ambivalent relationship with the social sciences. Even what might seem like an obvious connection—use of the label 'positivism'—is not free of ambiguity. The term positivism literally refers to the fact that law is 'posited,' a human social creation. Some legal theorists assert that there is no direct connection between legal positivism and scientific positivism (Patterson 1953, p. 15, Posner 1990, p. 20); others suggest that the connection is a general one which lies in the shared insistence on the *is/ought* distinction (Schauer 1996, p. 49). Still others insist that there is a strong connection, grounded in a shared empiricist narrowing of reality, need for certainty (Rommen 1998, pp. 109–13), and a shared opposition to theological and metaphysical modes of thought (Lee 1989, pp. 16–18). Hart himself invoked connections to social science. As noted earlier, he described his book as an essay in descriptive sociology, though this claim did not endear him to social scientists. Furthermore, in his discussion of the internal aspect of

rules, Hart acknowledged Winch's *The Idea of a Social Science* (1958), and he has been lauded for importing the hermeneutic perspective from social science into jurisprudence.

These connections aside, Hart's theory has been criticized by many social scientists who study law. A standard social scientific criticism is that Hart made a host of putatively descriptive sociological claims without having first done the requisite sociological investigation (Cotterrell 1989, pp. 94–106). An anthropological critique of Hart focuses on the fact that his concept of law was abstracted from state legal systems, which consequently leads to the conclusion that certain societies—those lacking in secondary rules—did not have law, suggesting that they are primitive or less civilized (Roberts 1979, pp. 23–5). This smacks of ethnocentric privileging of Western legal systems. Social scientists who hold the belief that 'no society is without law' (Moore 1978, p. 215) must reject Hart's concept of law. A third common criticism, though one aimed at legal positivism more generally, is that legal positivist views of law tend to have the effect of masking the social reality that there are a plurality of 'legal' systems operative in society, not just the state legal system that legal positivism emphasizes (Griffiths 1986).

Despite these criticisms, Hart's concept of law has had a significant, if mostly underground, influence in social scientific approaches to the concept of law. Bohannan's widely followed anthropological definition of law as the reinstitutionalization of norms bears a strong resemblance to Hart's notion of primary and secondary rules (Bohannan 1965). Galanter, one of the leading theorists of legal pluralism, adopts Hart's concept of law to distinguish legal from non-legal forms of social regulation (Galanter 1981, pp. 18–19, n. 26). Hart's concept of law is attractive because it comports with our intuitions about law; it comports with our intuitions because, in the modern world, it is the form of law with which we are most familiar. Since law is a social institution, it is inevitable that, whichever concept of law one formulates, it will be derived from a preconception about the nature of law (Tamanaha 1993, 1995).

Legal positivism, especially Hart's version of it, regardless of its limitations, holds a great deal of promise for the social scientific study of law, though that promise has yet to be fully realized. The legal positivist ambition to construct a general and descriptive approach to law is one the social sciences should embrace. Social scientists can conduct the sociological studies Hart failed to complete, using his model of law as the union of primary and secondary rules to guide their research, without necessarily accepting the separate propositions that this is the best or only way to characterize and understand law. Hart's primary mistake from the standpoint of a descriptive sociology came when he built normative acceptance by legal officials into his concept of law. Whether or not legal officials accept the secondary rules is an open question that can only be resolved following investigation (Tamanaha 1997, pp. 131–7). With these and other adjustments, a social scientific effort to continue and extend Hart's project could apply his conventionalist approach to identify, abstract from, and describe other forms of law operative in society. This sociological or 'sociolegal' version of positivism has the potential to provide information about the relationship between law and society and to establish a bridge between jurisprudence and the social scientific study of law (Tamanaha 2001).

See also: Logical Positivism and Logical Empiricism; Natural Law; Positivism, History of; Positivism: Sociological; Rechtsstaat (Rule of Law: German Perspective)

Bibliography

Acquinas T 1987 *Treatise on Law*. Regnery Gateway, Washington, DC

Austin J 1995 *The Province of Jurisprudence Determined*. Cambridge University Press, Cambridge, UK

Bix B 1996 *Jurisprudence: Theory and Context*. Westview Press, Boulder, CO

Bix B 1999 Patrolling the boundaries: Inclusive legal positivism and the nature of the jurisprudential debate. *Canadian Journal of Law and Jurisprudence* **12**: 17–32

Bobbio N 1993 *Thomas Hobbes and the Natural Law Tradition*. University of Chicago Press, Chicago

Bohannan P 1965 The differing realms of the law. *American Anthropologist* **67**: 33–42

Boorstin D J 1996 *The Mysterious Science of Law*. University of Chicago Press, Chicago

Coleman J L 1982 Negative and positive positivism. *Journal of Legal Studies* **11**: 139–64

Coleman J L 1996 Authority and reason. In: George R P (ed.) *The Autonomy of Law: Essays on Legal Positivism*. Oxford University Press, Oxford, UK

Coleman J L 1998 Incorporationism, conventionality, and the practical difference thesis. *Legal Theory* **4**: 381–425

Cotterrell 1989 *The Politics of Jurisprudence*. Butterworth, London

Dworkin R 1977 *Taking Rights Seriously*. Duckworth, London

Dworkin R M 1985 *A Matter of Principle*. Belknap Press of Harvard University Press, Cambridge, MA

Dworkin R M 1986 *Law's Empire*. Belknap Press of Harvard University Press, Cambridge, MA

Finnis 1980 *Natural Law and Natural Rights*. Clarendon Press, Oxford, UK

Finnis J 1996 The truth in legal positivism. In: George R P (ed.) *The Autonomy of Law: Essays on Legal Positivism*. Oxford University Press, Oxford, UK

Fuller L 1940 *The Law in Quest of Itself*. Foundation Press, Chicago

Fuller L 1969 *The Morality of Law*. Yale University Press, New Haven, CT

Galanter M 1981 Justice in many rooms: Courts, private ordering, and indigenous law. *Journal of Legal Pluralism* **19**: 1–47
Griffiths J 1986 What is legal pluralism? *Journal of Legal Pluralism* **24**: 1–55
Hart H L A 1958 Positivism and the separation of law and morals. *Harvard Law Review* **71**: 593–629
Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK
Hart H L A 1994 *The Concept of Law*, 2nd edn. Clarendon Press, Oxford, UK
Hobbes T 1946 *Leviathan*. Blackwell, Oxford, UK
Kelsen H 1967 *The Pure Theory of Law*. University of California Press, Berkeley, CA
Lee K 1989 *The Positivist Science of Law*. Avebury, Aldershot
Leiter B 1998 Realism, hard positivism, and conceptual analysis. *Legal Theory* **4**: 533–47
MacCormick N 1992 Natural law and the separation of law and morals. In: George R (ed.) *Natural Law Theory*. Oxford University Press, Oxford, UK
Moore S F 1978 *Law as Process: An Anthropological Approach*. Routledge and Kegan Paul, Boston
Patterson E W 1953 *Jurisprudence: Men and Ideas of the Law*, 1st Printed edn. Foundation Press, Brooklyn, NY
Posner R A 1990 *The Problems of Jurisprudence*. Harvard University Press, Cambridge, MA
Raz J 1979 *The Authority of Law: Essays on Law and Morality*. Oxford University Press, Oxford, UK
Roberts S 1979 *Order and Dispute: An Introduction to Legal Anthropology*. St. Martin's Press, New York
Rommen H 1998 *The Natural Law: A Study in Legal and Social History and Philosophy*. Liberty Fund, Indianapolis
Schauer F 1996 Positivism as pariah. In: George R P (ed.) *The Autonomy of Law: Essays in Legal Positivism*. Oxford University Press, Oxford, UK
Tamanaha B Z 1993 The folly of the 'social scientific' concept of legal pluralism. *Journal of Law and Society* **20**: 192–217
Tamanaha B Z 1995 An analytical map of social scientific approaches to the concept of law. *Oxford Journal of Legal Studies* **15**: 501–35
Tamanaha B Z 1997 *Realistic Socio-legal Theory*. Clarendon Press, Oxford, UK
Tamanaha B Z 2001 Sociolegal positivism and a general jurisprudence. *Oxford Journal of Legal Studies* **21**: 1–32
Winch 1958 *The Idea of a Social Science and its Relation to Philosophy*. Humanities Press, Atlantic Highlands, NJ

B. Z. Tamanaha

Legal Process and Social Science: United States

Introduction

There are two complementary orientations in the field of 'law and social science.' One is usually termed 'law and society' and addresses theories of how law comes into being and changes in a culture, and how law relates to other cultural forces, such as religion, or the family. The second is called 'social science in law' and deals with more applied questions of how courts and lawyers make use of social science information in reaching decisions in actual legal cases. This article is concerned with the latter orientation.

1. Legal Realism

Since the rise of American legal realism, early in the twentieth century, courts have, to varying extents, been open to using social science research when that research could help resolve an empirical issue that arose in litigation. Social science research was first urged upon a court in 1908 by Louis Brandeis in *Muller vs. Oregon* (208 US 412), a landmark case dealing with the constitutionality of social welfare legislation limiting to 10 hours the work day of any woman or girl employed in a factory or laundry. Brandeis assembled a substantial body of social science research tending to show the debilitating effect on women and girls of working long hours, and presented this material to the United States Supreme Court in a brief defending Oregon's statute. In the opinion in *Muller* upholding the legislation, after referring to the social science materials in a footnote, the Court stated that although they 'may not be, technically speaking, authorities,' the studies would nonetheless receive 'judicial cognizance.'

2. Three Uses of Social Science in Law

Although the use of what came to be known as 'Brandeis briefs' became common in the years after *Muller*, legal commentators were hard pressed to explain an apparent anomaly: How was it that Brandeis could present research to an appellate court in a written brief, when 'facts' were supposed to be introduced at the trial level by the oral testimony of witnesses? This conundrum persisted until Kenneth Culp Davis, in a seminal article published in 1942, distinguished two types of 'fact.' The first type, legislative fact, referred to facts that were used by courts to help decide broad questions of law or policy, as Brandeis had used research to help decide the constitutionality of social welfare legislation. The second type, adjudicative fact, referred to facts that were used to decide questions of interest only to the specific parties to a lawsuit, such as whether a particular traffic light was red or green. Davis' distinction soon became widely accepted and now forms the traditional scheme used by courts and commentators to describe the judicial uses of social science. There are indications that a third use of social science in law has developed, however, and that Davis' distinction between different types of 'fact' may have outlived its usefulness.

2.1 Social Science Used to Make Law

Legislative facts are facts that courts use when they make law (or 'legislate'), rather than simply apply settled doctrine to resolve a dispute between particular parties to a case. While the determination of adjudicative facts affects only the litigants before the court, the determination of legislative facts influences the content of legal doctrine itself, and therefore affects many parties in addition to those who brought the case. It was Davis' position in 1942 that '[t]he rules of evidence for finding facts which form the basis for creation of law and determination of policy should differ from the rules for finding facts which concern only the parties to a particular case.' In this manner, Davis justified the Supreme Court's practice in *Muller* of disregarding the rules of evidence by accepting 'facts' on appeal and in a brief. The rules of evidence did not apply to Brandeis' social science materials: they were 'legislative' facts.

Judicial acceptance of social science research as a form of legislative fact became commonplace throughout American law after *Brown* vs. *Board of Education* (347 US 483 (1954)). In that case, the United States Supreme Court cited the published research of numerous social scientists to support its empirical assertion that the segregation of public schools instills in an African American child 'a sense of inferiority [that] affects the motivation of a child to learn.' In the five decades since *Brown*, research has frequently been invoked by courts to demonstrate the validity of empirical assumptions made in the process of modifying existing law or creating new law. In *United States* vs. *Leon* (468 US 897 (1984)), for example, the United States Supreme Court considered the question of whether the rule excluding evidenced produced by an illegal search should be modified to admit evidence that the police seized while they were relying in good faith on a search warrant that later proved to be defective. The Court held that whether the rule should be modified to allow such evidence 'must be resolved by weighing the costs and benefits' of creating an exception to the rule when the police act in good faith. In the process of weighing these costs and benefits, the Court reviewed six social science studies concerning the effects of the exclusionary rule on the disposition of felony arrests and concluded that the benefits of creating an exception, in terms of increasing the number of guilty criminals brought to justice, outweighed any costs involved.

Criticism of the concept of legislative fact has been intense. It has been claimed to be largely a negative concept: the usual rules of evidence should not apply to facts introduced for the purpose of making new law. The concept provides little positive guidance for courts as to how they should deal with this type of social science information. Improvements in the manner that courts use social science information to create rules of law may be possible only by abandoning the notion of 'legislative fact' and developing a new concept that alters the way in which courts view social science materials. 'Social authority' has been proposed by Monahan and Walker (1986) as one alternative to legislative fact as an organizing principle for courts' use of social science to create or modify a rule of law. Under this rubric, courts would treat social science research relevant to creating or modifying a rule of law as a source of 'authority' rather than a source of 'facts.' More specifically, courts would treat social science research much as they treat legal precedent under the common law.

2.2 Social Science Used to Determine Facts

Adjudicative facts—also called 'social facts'—are facts that apply only to the particular parties before the court. They are used to determine (or 'adjudicate') what happened in a specific case, and not for some larger purpose, such as to argue that the law should be changed.

One of the most frequent uses of social science research as adjudicative fact involves trademarks. The Lanham Trademark Act (15 USCA Sec. 1127) states that a new trademark will not be registered if it so resembles a trademark already registered to another person 'as to be likely … to cause confusion.' A person who sells a product that is likely to cause confusion with an already trademarked product is liable for trademark infringement. Social science research in the form of surveys of consumers or potential consumers to ascertain the degree of confusion between products has been admitted in court for many decades. Initially, such evidence was often successfully challenged as contravening the prohibition against 'hearsay' evidence, since the respondents to the surveys were not present in court to testify. In *Zippo Manufacturing Co.* vs. *Rogers Imports* (216 F. Supp. 670 (1963)), however, the hearsay objection was definitely laid to rest: 'The weight of case authority, the consensus of legal writers, and reasoned policy considerations all indicate that the hearsay rule should not bar the admission of properly conducted public surveys.'

A wide variety of social science methodologies are now routinely used by both plaintiffs and defendants in trademark cases to determine the presence of consumer confusion. *In Processed Plastic* vs. *Warner Communications* (675 F.2d 852 (7th Cir. 1982)), for example, Warner licenced several toy companies to produce replicas of a unique automobile that figured prominently in a television series that it owned. The Processed Plastic Company, which had no agreement with Warner, began to sell a toy car that Warner claimed looked like the car in its television series. To demonstrate this fact, Warner introduced in court a social science survey of random groups of children questioned at shopping malls that found that 82 percent of the children shown a Processed Plastic car

identified it as the car that appeared in Warner's television series. The court decided that these data showed that Processed Plastic has likely infringed Warner's trademark rights by creating consumer confusion.

The key to understanding how courts deal with social science research used to determine an adjudicative fact is that, unlike 'legislative' facts, adjudicative facts are susceptible to proof by the usual rules of evidence. Evaluating the research for the purpose of admissibility is limited to determining whether it is material, whether it has any probative value, and whether its admission violates any other evidentiary rule. The ultimate weight to be placed on the social science evidence is a matter for the fact-finder.

2.3 Social Science Used to Provide Context

Most of the judicial uses of social sciences fall into one of the two categories originally described by Davis. In recent decades, however, courts have begun to confront uses of social science research that do not conform to the established classification. There are increasing indications that a new, third use of social science in law is emerging. Examples can be found in cases concerning eyewitness identification and sexual victimization.

In *State* vs. *Chapple* (660 P.2d 1208 (1983)), the only evidence that the defendant had committed murder was the word of two eyewitnesses. At trial, the expert testimony of a research psychologist was offered by the defendant to counter the testimony of the eyewitnesses. In a proffer of the evidence, the psychologist described published studies on factors such as the speed with which memory decays over time, the effects of stress on eyewitness accuracy, and the relationship between the confidence of a witness and the accuracy of identification. The trial court held the research on the accuracy of eyewitness identification was not a 'proper subject' for expert testimony and excluded the expert. However, the Arizona Supreme Court reversed and remanded the case for a new trial, stating that 'there were a number of substantive issues of ultimate fact on which the expert's testimony would have been of significant assistance.'

Similarly, the defendant in *State* vs. *Myers* (359 N.W.2d 604 (Minn, 1984)) was found guilty of criminal sexual conduct involving a child. The trial court had allowed the prosecution to introduce a social scientist as an expert witness to describe the behavioral traits 'typically' observed in abused children, traits that were also seen in the child complainant. The defense objected to this evidence at trial and appealed. The Minnesota Supreme Court upheld the conviction, however, stating that 'background data providing a relevant insight into the puzzling aspects of the child's conduct and demeanor which the jury could not otherwise bring to its evaluation is helpful and appropriate.'

In neither of these cases was social science research being used to provide legislative facts. No creation or modification of a rule of law was sought. Rather, in both cases the parties offering the research accepted the applicable legal rules and sought to show that the research would help the jury to decide the specific factual issues being litigated. Yet in neither of these cases was social science research being used to provide adjudicative facts either. The parties to the cases were not involved in the research at all: the experts relied on research studies published before the events that gave rise to the litigation took place.

The way social science was used in *Chapple* and *Myers*, however, reflects the defining characteristics of both legislative fact and adjudicative fact. In each case, the research demonstrated the critical component of legislative fact—generality. Just as the research used in *Leon* focused on the 'collective' costs and benefits of changing the exclusionary rule, the research used in *Chapple* focused on 'general factors' affecting all eyewitnesses. Similarly, the research in *Myers* concerned 'general characteristics' that were 'typically observed' in groups of abused children. In each case, the research used also possessed the critical component of adjudicative fact—specificity. Just as the research used in *Processed Plastic* spoke only to whether consumers were confused between two given products, the studies in *Chapple* were used only to challenge the testimony of eyewitnesses to a particular crime, and the research in *Myers* was used only to show that the named victim was, in fact, abused.

The research used in these cases and many like them is thus neither wholly legislative nor wholly adjudicative fact but has the essential elements of both of the conventional categories. Walker and Monahan (1987) have proposed a third category, termed social framework, to denote the use of general conclusions from social science research to determine factual issues in a specific case.

3. Future Directions

At the end of beginning of the twenty-first century, social science in the legal process is being transformed in two ways. First, the law of scientific evidence is very rapidly developing. In 1993, the United States Supreme Court, in *Daubert* vs. *Merrell Dow Pharmaceuticals* (509 US 579), for the first time decided the legal standard for the admissibility of scientific evidence.

The Rules of Evidence … assign to the trial judge the task of ensuring that an expert's testimony both rests on a reliable foundation and is relevant to the task at hand. Pertinent evidence based on scientifically valid principles will satisfy those demands.

The Court offered 'some general observations' on how judges should evaluate the validity of scientific research. Among other things, judges should consider whether a scientific theory or technique has been 'tested,' whether it has been 'subjected to peer review and publication,' whether it has a 'known or potential rate of error,' and whether it is 'generally accepted' in the scientific community. In a subsequent case, *Kumho Tire Company* vs. *Carmichael* (119 S. Ct. 1167 (1999)), the Court held that these concerns apply not just to 'scientific' expert testimony, but to all forms of expert testimony.

The second development in this area is the rapid rise in the availability of material to train lawyers and judges in the uses of scientific evidence, including social science evidence. Shortly after *Daubert*, the Federal Judicial Center (1994) published a *Reference Manual* that provided guidance for the admissibility of statistical evidence and the findings from social surveys. Subsequently, a three-volume work, *Modern Scientific Evidence* (Faigman et al. 1997), offered detailed analyses, from both legal and social scientific perspectives, of all of the issues discussed above. As the authors of this latter work note, '*Daubert*, perhaps represents nothing more, or less, than that henceforth the legal culture must assimilate the scientific culture.'

See also: Expert Testimony; Justice and its Many Faces: Cultural Concerns; Law and Society: Sociolegal Studies; Law as an Instrument of Social Change; Law: History of its Relation to the Social Sciences; Legal Realism; Scientific Evidence: Legal Aspects

Bibliography

Davis K C 1942 An approach to problems of evidence in the administrative process, *Harvard Law Review* **55**: 364–425

Faigman D L, Kaye D H, Saks M J, Sanders J 1997 *Modern Scientific Evidence*. West group, St. Paul, MN

Federal Judicial Center 1994 *Reference Manual on Scientific Evidence*. West Group, St. Paul, MN

Monahan J, Walker L 1986 Social authority: Obtaining evaluating and establishing social science in law. *University of Pennsylvania Law Review* **134**: 477–517

Walker L, Monahan J 1987 Social frameworks: A new use of social science in law. *Virginia Law Review* **73**: 559–598

J. Monahan

Legal Professionalism

The elaborate body of law that governs public and private life in most states today is not self-executing. Legal personnel and institutions must give it effect, as when judges adjudicate cases in the courts. Since ancient Rome, experts have represented or advised clients in legal matters (Pound 1953). They now exist in ratios of up to one per 300 citizens, the rate in the United States (Abel 1989). Most represent clients before tribunals, advise them about the legal consequences of their actions, draft their wills and contracts, or negotiate the terms of their transactions with others. Some do all these things; others specialize by task or field of law.

In the United States, the experts authorized to do this work are called lawyers or attorneys. (But legal tasks often overlap with tasks that non-lawyers perform, as where lawyers and accountants alike may prepare clients' tax returns.) They are trained, licensed, and regulated as a unitary profession, though many specialize. Comparable experts in other states may be divided into several legal professions, but the distinctions may change over time (Abbott 1988). In France, *avocats* go to court, *conseils juridiques* advise, and *notaires* draft documents. In England, barristers once held exclusive 'rights of audience' in the courts while solicitors provided other services, but the lines are now blurring. For brevity's sake, this article refers to all authorized legal service providers as lawyers and members of the legal profession or bar.

'Legal professionalism' refers to a cluster of values (see Sect. 1)—an ethic or ideology—that lawyers arguably share and that arguably influences their conduct. The concept has led two lives: as a folk concept for lawyers and as an analytic tool for sociologists. Lawyers find the concept normatively attractive as well as rhetorically useful, where they are politically diverse, for reaching a consensus in intra-professional policy debates. Bar leaders *ascribe* the ethic to lawyers, hoping to justify the profession to the laity, or foster professional solidarity, or avoid external regulation (see Sect. 1.7). They also *prescribe* the ethic for lawyers, hoping to inspire them to serve clients well while also promoting legality or the rule of law. Sociologists, for their part, have cited professionalism to explain or predict lawyers' conduct, to compare law practice in different times or places, or to compare lawyers with other occupational groups. They emphasize what legal professionalism has in common with the prevailing ethic or ideology in other professions (Friedson 1986), while lawyers see it as a 'learned' distinguishing characteristic.

Professionals, as opposed to amateurs, work to make money. Yet legal professionalism, according to *scholarly* defenders, implies pursuing a learned art as a common calling in the spirit of public service—'no less a public service because it may incidentally be a means of livelihood' (Pound 1953). So understood, the ethic appears to have precapitalist roots and might be expected to lose sway in today's entrepreneurial world (Krause 1996, but see Larson 1977).

Just as lawyers' work, clientele, and status vary from one legal system to another (Rueschemeyer 1973), so professionalism's impact on lawyers appears

to vary. Political or religious ideology can diminish its influence, as happened in the former Soviet Union. And yet, since its values are only weakly specified, professionalism may spur lawyers in dictatorships to resist state intrusions into lawyer–client relations even as it inspires lawyers in constitutional democracies to avoid such intrusions by forbearing to help clients commit illegalities (Friedman and Zile 1964).

1. *Core Values of Legal Professionalism as Lawyers Understand It*

Legal professionalism urges lawyers, individually and collectively, to pursue or maintain these conditions: competence, client trust, independent judgment, a just distribution of legal services, the legitimacy of the legal process, a professional community, and self-regulation.

1.1 *Professional Competence*

Seen through the lens of professionalism, lawyer competence takes on a distinctive gloss. Lawyers must be learned in law and skilled in applying it. Since their knowledge and skills are esoteric, the quality of their work is to be judged by professional, not lay, standards. Of course, craftsmen also possess esoteric knowledge, viewed in their case as competence gained through experience. Lawyers in some states view their expertise in this way; lawyers elsewhere consider it grounded on theory, a 'science' to be studied at university (Weber 1954). Either way, incompetence so threatens vital interests that would-be lawyers should not practice law, or call themselves lawyers, without obtaining a license, which usually entails long study, exams, and/or apprenticeship. Once licensed, lawyers should continue to hone their skills. They do not simply work, they practice.

Lawyer competence exalts quality over cost. Lawyers should not so limit the scope of their engagements that clients must give up their right to competent service, even if the limits make services more affordable. In the name of ensuring competence, some bars have even forbidden lawyers to charge fees below floors promulgated in minimum-fee schedules. Recent attacks on these schedules illustrate the tensions between professionalism and market values. Doubting that the floors ensure quality, and suspicious of restrictions on price competition, some states have removed a 'learned-profession' exemption from antitrust laws and now treat the schedules as illegal restraints on *trade* or *commerce*. More generally, critics disparage many bar policies that pit professionalism against market values. They see professionalism as a policy-making bias, if not a mask for lawyer self-interest. Scholars fuel the fire by likening minimum fees to the practices of medieval guilds (Krause 1996). In return for setting and enforcing just prices and quality standards, princes gave guilds exclusive rights to make and sell their goods.

1.2 *Promoting Client Trust Through Ethical and Fiduciary Duties*

Competence is a cognitive achievement. But effective lawyering also depends on maintaining client trust, which is more a function of ethical commitment than of cognition. Professionalism urges lawyers to shoulder fiduciary duties to clients that exceed other service providers' duties to customers. Consider the duty of confidentiality. Clients must reveal sensitive or compromising facts to their lawyers but they fear that such disclosures could be used against them. The duty of confidentiality and the related attorney–client 'privilege' counter this fear. If lawyer or client is summoned to testify in court about their private communications, the client may be privileged to suppress the testimony. Believing that clients need such protection in order to be candid, lawyers see no tension between keeping clients' secrets and the societal goal of getting at the truth through adjudication.

Since information disparities enable lawyers to manipulate client demand, lawyers should also promote trust by giving clients disinterested advice about the services they need and not soliciting clients so aggressively that disinterestedness is put at risk. Lawyers, moreover, owe each client undivided loyalty. Ethical norms discourage engagements in which personal or third-party interests may conflict with client interests and thereby cloud a lawyer's judgment on the client's behalf.

Lawyers often enjoy considerable social status. Max Weber (1954) found the source of lawyers' status in their knowledge base, while Emile Durkheim (1958) stressed their ethical or fiduciary commitments, which he saw as a product of professional socialization. But how well professionalization instills these commitments is unknown and in the past states relied instead on class, ethnicity, or family prominence to ensure lawyer trustworthiness. Before World War II, for example, only aristocrats and gentlemen became English barristers.

1.3 *Independent Professional Judgment*

Most engineers are not partners in engineering firms; they are employees in public or private entities run by others. Engineers are thus a learned, but not a 'free,' profession as lawyers aspire to be. Legal professionalism exhorts lawyers to remain free to exercise independent professional judgment, both in public life and in law practice. Yet, with 20 percent of the bar in North America and Western Europe now employed

by businesses, public agencies, or state-subsidized offices serving low-income clients (Spangler 1986), lawyers worry that independent judgment is becoming unsustainable.

Like Edmund Burke, legal professionalism views public policy-making as a process of reasoned deliberation, not a mere reflection of power or toting up of preferences. On this view, the bar can facilitate policy-making in states prone to political gridlock by recasting some issues as apolitical questions of law reform and debating them in the idiom of legalism, which focuses less on substantive aims than on whether proposals are constitutional and cohere with existing legal principles (Halliday 1987). Professionalism also urges the organized bar to address lawmakers on behalf of a palpable public interest, not the parochial interests of lawyers or powerful clients.

In law practice itself, independence implies sufficient discretion to exercise sound judgment on behalf of clients. Professional norms warn lawyers not to defer to the wishes of third-party payers in deciding how to serve clients and, in a dwindling number of states, also forbid lawyers to take non-lawyers such as accountants as partners. English barristers and French *avocats* have even been forbidden to form partnerships with others of their own kind lest their own capacity to exercise independent judgment be impaired.

But independence is not valued solely for the client's sake. Lawyers must also withstand clients' pressures to disregard third-party rights or the legal system's institutional needs. Lawyers call themselves 'officers of the court' or 'ministers of justice' to stress this reason for maintaining independent judgment. Fearing that lawyers who work solely for one client (often a company) will be too client-dependent, some European bars forbid them to designate themselves publicly as lawyers, or represent their client in court, or shelter client secrets under the attorney–client privilege. On similar grounds, English barristers are not retained and instructed by their clients, but by their clients' solicitors, who, as go-betweens, may shield barristers from overbearing clients.

1.4 Maintaining a Legitimate Legal Order Through a Just Distribution of Legal Services

State licensing gives lawyers market power, while professionalism urges them not to leave the distribution of legal services solely to market forces, even as tempered by state subsidies for clients of modest means. Because legal services often affect non-clients as well as clients, client ability to pay is not a valid measure of the resources that a just legal order would devote to those services. For example, since the public has a general interest in the integrity of the criminal justice system, access to criminal defense lawyers should not turn on a defendant's ability to pay, for indigents would then have no one to defend them and the validity of the state's charges would be left in doubt. Subsidies are an imperfect solution, not just because they are often minimal, but because the state may so restrict the ways in which lawyers may use the funds that lawyers think they can no longer exercise independent judgment. Professionalism therefore urges lawyers to accept some engagements *pro bono publico*, i.e., for little or no pay.

1.5 Maintaining the Procedural Legitimacy of the Legal System

Professionalism encourages lawyers to promote a legitimate legal order by preserving the integrity and rationality of legal processes. Lawyers should support the bar's campaigns for law reforms that rationalize procedures, and should not abuse procedures in order to pursue clients' aims. The tension between client interests and system-maintenance is a central theme in bar discourse.

1.6 Maintaining a Professional Community Through Bar Associations

As noted in Sect. 1.2, Durkheim thought professional associations could have a positive influence on their members. This is also a tenet of legal professionalism as lawyers understand it. Julius Henry Cohen, an American bar leader who was Durkheim's contemporary, argued that

> Lawyers must subordinat[e their] inclination, appetites, and ambitions to an organization which has as its object to promote the performance of a social function (Cohen 1924, quoting R. H. Tawney).

Since the bar must play superego to the lawyer's instincts, Cohen believed unaffiliated lawyers cannot be wholly professionalized. He saw no risk that greater bar solidarity might lead to harmful restrictions on professional competition.

1.7 The Elaboration and Promotion of a Professional Ethic Through Self-regulation

'Self-regulation' can refer to the ideas that lawyers should be governed by internalized values rather than external rules, or regulated by the judiciary rather than the 'political' branches of government. More often it refers to the idea that the bar itself should provide regulatory oversight by formulating ethical norms to govern lawyers and maintaining a disciplinary system to censure, fine, or disbar violators. Formal sanctions become important once lawyers become so numerous, diverse, and unknown to one another that concern for their reputations no longer ensures compliance. Legal

professionalism favors self-regulation largely as a way to obviate any need for regulation by the 'political' branches of government, which might not grasp or accept the lawyer's role in society and might be too quick to sanction those who challenge the legality of state action.

2. *The Uses and Contested Value of Legal Professionalism in Sociology*

Since the late-nineteenth century, scholars, notably sociologists, have found professionalism a useful analytic concept, though few have used it exclusively in studying lawyers. They have invoked professionalism, often as a variable, to (a) characterize an occupational mentality, (b) account for the conduct or relative status of occupations, (c) explain how occupations avoid outside regulation, (d) construct occupational taxonomies, or (e) trace the evolution of occupation toward an ideal type (Pound 1953). Until 1970, most scholars who deployed the concept were functionalists, such as Durkheim or Talcott Parsons (1954), who stressed the societal benefits of training lawyers to embrace the communal and anticommercial values of professionalism-as-ethic. But since then, critical theorists have questioned its analytic value or validity, at least as used by the functionalists. The critical school views professionalism as a modern conceit or ideology that occupational groups use, wittingly or not, to mobilize themselves or to justify their autonomy, status, and market power to the laity (see, e.g., Larson 1977).

3. *The Murky Prospects for Legal Professionalism as Ethic and as Ideology*

Whether one views legal professionalism as ethic or ideology, it is clear that powerful forces are arrayed against its continued vitality (Krause 1996). Lawyers, especially those with business clients, cannot be insulated from the entrepreneurial values ascendant in much of the world today. Clients press them to form practice alliances with accountants and consultants, alliances that may be cost-effective but undermine intra-professional solidarity. Relentless specialization in response to law's penetration into all aspects of life is generating specialty bars that threaten the viability of more inclusive bar associations. Specialization is also turning lawyers into technicians whose narrow skills lay competitors can often match (Kritzer 1998). In dynamic economies, moreover, lawyers cannot quench the demand for legal services; again inviting lay competition (Abbott 1988). Cost-cutting efforts to 'unbundle' or fragment traditional legal tasks, coupled with the transformation of some legal services into standardized products, are also eroding the distinctiveness of legal work. Finally, the proliferation of large, bureaucratically-controlled worksites for lawyers—mega-law firms, corporate law departments, public agencies, and subsidized legal aid offices—is in obvious tension with professional independence.

However, the cluster of ill-specified values summed up in the term 'legal professionalism' is not necessarily doomed to extinction. One can imagine counter forces that might sustain professionalism as ethic or ideology. Renewed respect for the rule of law and for legal dispute resolution techniques may accompany the ascendancy of commercial values, and specialization may be tempered by renewed demand for ethically committed lawyers who provide wise counsel based on wide understanding of law, legal traditions, and the virtues of legality.

See also: Ethical Codes, Professional: Business Codes; Ethics in Organizations, Psychology of; Government Lawyers; Judges; Lawyers; Lawyers, Regulation of; Public Interest Law

Bibliography

Abbott A 1988 *The System of Professions*. University of Chicago Press, Chicago

Abel R L 1989 *American Lawyers*. Oxford University Press, New York

Cohen J H 1924 *The Law: Business or Profession?* G A Jennings, New York

Durkheim E 1958 *Professional Ethics and Civic Morals*. The Free Press, Glencoe, IL

Friedman L, Zile Z 1964 Soviet legal profession: developments in law and practice. *Wisconsin Law Review* 32–7

Freidson E 1986 *Professional Powers*. University of Chicago Press, Chicago

Halliday T C 1987 *Beyond Monopoly: Lawyers, State Crises, and Professional Empowerment*. University of Chicago Press, Chicago

Krause E A 1996 *Death of the Guilds: Professions, States, and the Advance of Capitalism, 1930 to the Present*. Yale University Press, New Haven, CT

Kritzer H M 1998 *Legal Advocacy: Lawyers and Nonlawyers at work*. University of Michigan Press, Ann Arbor, MI

Larson M S 1977 *The Rise of Professionalism: A Sociological Analysis*. University of California Press, Berkeley CA

Parsons T 1954 A sociologist looks at the legal profession. In: Parsons T (ed.) *Essays in Sociological Theory*. Free Press, Glencoe, IL

Pound R 1953 *The Lawyer from Antiquity to Modern Times*. West Publishing, St Paul, MN

Rueschemeyer D 1973 *Lawyers and Their Society: A Comparative Study of the Legal Profession in Germany and in the United States*. Harvard University Press, Cambridge, MA

Spangler E 1986 *Lawyers for Hire: Salaried Professionals at Work*. Yale University Press, New Haven, CT

Weber M 1954 *Max Weber on Law in Economy and Society* (M Rheinstein ed.). Simon & Schuster, New York

T. J. Schneyer

Legal Realism

'Legal Realism' is the name generally given to a movement in American legal thought between World War I and World War II. However, any attempt to identify precisely the intellectual content of the Realist movement is made difficult by the fact that its salient tenets, and thus its meaning, were widely contested during its heyday and are still widely contested.

1. *Difficulties in Defining 'Legal Realism'*

The Realist movement has been seen as an outgrowth of the Sociological Jurisprudence of Rudolf von Jhering as expounded by Roscoe Pound (1911–12), as well as a reaction to Pound's work. In addition, many have understood Realism as an outgrowth of the thought of Oliver Wendell Holmes, Jr. (1897), although it is not clear that such an identification is warranted, given Holmes' relentless pessimism and lack of faith in democracy. Also, while all agree that Realism was a reaction to legal formalism, sometimes referred to as 'conceptualism'—the notion that legal rules can be correctly deduced from a proper understanding of the nature of a limited set of important, narrowly legal premises—there nevertheless is disagreement as to whether any significant American legal thinker held such a position and thus what an opposition to formalism might signify (see *Legal Formalism*). The most that can be said is that all scholars agree that some portion of the activities of individuals commonly called Realists was directed at critique of the doctrinal results of legal formalism, at understanding judicial decision-making, at empirical research in law, at reform of legal education, and at participation in the activities of the New Deal administration of Franklin D. Roosevelt.

There is, however, nothing like agreement among scholars about the relative importance of these various activities for understanding Realism; moreover the weighing of relative importance in part determines who is included in a list of The Realists. Thus, while it is likely that most scholars would agree that Thurman Arnold, Charles E. Clark, Walter Wheeler Cook, William O. Douglas, Jerome N. Frank, Walton H. Hamilton, Karl N. Llewellyn, Underhill Moore, Herman Oliphant, and Hessel Yntema belong on the list, the inclusion or exclusion of men such as Adolph A. Berle, Walter Bingham, Felix Cohen, Morris R. Cohen, Arthur L. Corbin, John Dewey, Leon Green, Robert L. Hale, Robert M. Hutchins, Edwin Patterson, Thomas Reed Powell, Edward S. Robinson, Max Radin, Fred Rodell, Wesley Sturges, and Leon Tulin, all plausible figures, will change one's understanding of Realism, as well as be dictated by that understanding. What follows, therefore, is only one of several possible expositions of American Legal Realism. Other expositions can be found in the works of Duxbury (1995), Horwitz (1992), Hull (1997), Kalman (1986), Purcell (1973), Rumble (1968), Stevens (1983), Twining (1973) and Verdun-Jones (1979).

2. *A Contestable Definition*

'Legal Realism' is best understood as the reaction of a group of individuals within the American legal academy to the collapse of the Progressive movement after World War I, the decisions of the United States Supreme Court invalidating economic regulation on grounds either of substantive due process or of interference with vested contractual rights, and the general shift of the USA to the political right in the Harding, Coolidge, and Hoover administrations. The important words in this definition are 'within the American legal academy,' for the movement largely took place at three schools—Columbia, Johns Hopkins, and Yale. Obviously, this definition does not succinctly describe the political orientation of the Realist movement. On this topic the best that can be said is that, taken as a whole, the Realists were hardly radicals. It would be astonishing to learn belatedly that any of these men were members of the Communist Party, and it is doubtful whether any even voted Socialist. Many served in the Roosevelt administration, but doing was an indication of socialist leanings only to the most reactionary members of the business elite and their Republican Party allies. Thus, Realists are better seen as mainstream liberals caught in a very conservative time.

The relationship of the Realist movement to the Progressive movement can be seen both positively and negatively. Positively, there is the similarity in the political causes supported (organized labor), and opposed (big business). Also, the empirical research engaged in by many of the Realist scholars is similar to research undertaken by Progressive reformers (although not progressive law professors) before World War I. Indeed, both groups shared an 'all men of good will' ethics, as is evident from the notion underlying their empirical research that once the facts about a social problem are known, then reform will follow. In this sense neither group was committed to what we would now see as an ideological view of law that, in contrast, would suggest that exposure of the 'facts' about a social practice might bring embarrassment, but would also bring denial and resistance

to change. For both the Progressives and the Realists evil existed in the world, but it existed in individuals, not in classes or other affinity groups.

The Realists' relationship to the Progressive movement can be seen negatively in the reaction of the members of the movement to Pound as Dean of the Harvard Law School. Pound was a mainstay of the legal wing of Progressive reform movements before World War I. His condemnation of 'Mechanical Jurisprudence' (Pound 1908) was a rallying cry for Progressives, as was his exposition of 'Sociological Jurisprudence' (Pound 1911–12). Pound's jurisprudence was designed to be a response to the first wave of antiprogressive constitutional decision-making by the Supreme Court. It seemingly encouraged empirical research by emphasizing the difference between 'the law on the books and the law in action' (Pound 1910), in order to urge courts to avoid a formal, mechanical application of law, and instead to recognize, balance, and thus adjust the interests of the parties in the light of contemporary social and economic conditions.

Yet, for the Realists, Pound was also part of the problem that their activities were designed to address. For them, all of Pound's talk about the gap between the law on the books and the law in action and about the need for balancing of interests had never translated into the activity of exploring these gaps or identifying the interests in question through empirical research, or into actually suggesting how particular interests might be balanced. Moreover, Pound's law school was little changed from the school of Langdell and Ames, seen by Realist scholars to be among the arch formalists. It reflected none of what they saw as the implications of Progressivism and Sociological Jurisprudence for legal education—a revision of category systems, an emphasis on factual investigation, and explicit work on social policy.

3. *What Realists Did*

What the above definition pointedly lacks is some explicit statement of what the Realists believed, and therefore what Realism is. This omission is intentional. Realism is what Realists did and it is only with that aphorism in mind that one can make sense of the disparate materials that were published by what must be seen as a network of friends, not a coherent intellectual position. Start, then, at the beginning.

Although some pre-War work by Bingham has a particularly Realist cast, Realism largely started at Columbia with pieces by Cook and Moore, who respectively argued for a more empirically based understanding of the real, as opposed to verbal, activities of appellate courts in the field of conflict of laws and of the real, as opposed to verbal, activities that made up the institutional practices that the law defined, such as property and contract. From their work, one can see that Legal Realism is not even a distant relative of philosophical realism, the doctrine that the concepts of daily life are ontologically real, but, as Leiter (1997) has suggested, is much closer to naturalism in art or literature, a commitment to see and describe the realities of daily life scientifically, that is without idealization or the avoidance of the ugly, although in Legal Realism's concern with democracy and the activities of the middle class, there is some affinity with literary and artistic realism as well.

Curiously, the group of Realists who taught at Columbia, including Oliphant, Llewellyn, and Douglas, was not particularly interested in constitutional law, as might have been expected given the group's affinities to Progressivism, but in corporate and commercial law. These men argued that, because of antiquated ideas about judicial decision making, in areas of law related to business, just as in constitutional law, the law was lagging behind changes in commercial life, particularly the growth of the large corporation employing thousands of workers and the change in methods of production and distribution of goods that had broken older patterns of finance and trade. This focus had much in common with the Institutional Economics of John Commons, Richard T. Ely, and Thorstein Veblen that emphasized the way that the institutional, by which they meant regular, patterns of economic behavior determined the growth or decline of industries, attitudes toward labor, and practices in the pricing of goods. Thus, Moore worked on the law of banking, Oliphant on labor and antitrust, Llewellyn on the sale of goods, and Douglas on corporations and commercial bankruptcy (see *Markets and the Law*).

Even more curiously, the empirical research that was implied by the work at Columbia was not undertaken there, but rather at Yale and Johns Hopkins, mostly in the years after the Columbia law faculty blew up over the results of a Dean search. At Yale, pioneering empirical work, superintended by Clark, Douglas, and Moore, focused on the activities of the federal courts in civil, criminal, and bankruptcy cases, on the activities of bankers in handling negotiable instruments, on the operation of the tort system in auto accidents, on the effect of parking regulations in altering the behavior of parkers, and on the legal needs of citizens of modest means. At Johns Hopkins, equally pioneering work, superintended by Oliphant, Yntema, and their colleague at Columbia, Leon C. Marshall, was done on the operation of state courts in civil, criminal, divorce, and justice of the peace cases.

Much of the rest of the scholarship of the Realists, though not empirical in the narrowest sense, was still within artistic and literary naturalism's desire to describe the realities of daily activities. Cook's continuing work on what courts were really doing in conflict of laws doctrine and Green's parallel work on torts doctrine, fit this mold, as does the work of Arnold and Robinson on the symbolic uses of law, Llewellyn's work on commercial practice, Hutchins'

work on the psychological underpinnings of evidence law, and Berle's work on corporate governance. All of this work, and most of the more narrowly empirical work as well, was directed at critiquing legal, particularly judge-made rules, but Realism is more famous for its related, but more frontal assault on contemporary understandings of judicial decision-making.

Realists attacked formalist understandings of judicial decision-making in three ways: logically, psychologically, and sociologically. The logical attack argued that legal concepts that the formalists, especially the formalists on the Supreme Court, saw as necessarily determining decisions were not in fact determinative. Here Realists relied on Wesley N. Hohfeld, who argued that legal concepts inappropriately treated as unitary various discrete legal relationships, any one of which could not be implied from the existence of another; on Judge Benjamin N. Cardozo, who asserted that there were four different methods for deciding a case and so allowed the judge the 'sovereign prerogative of choice'; on Cook, who claimed that legal rules 'hunt in pairs,' so that for every rule there is an exception or counter rule whose application would lead to a different result; and on Llewellyn, who eventually demonstrated that there is a great range of acceptable techniques for dealing with a legal precedent. Together, these observations suggested that legal concepts did not produce necessary results by the use of logic alone, but rather permitted judges a choice of results, a choice that was made surreptitiously, as it were, without explicitly considering the function that the rules played in society or the policy behind those rules.

The psychological attack argued that judicial opinions, traditionally presented in a seemingly syllogistic form, hid a decision that was reached in very different ways. Here Realists relied on the work of another judge, Joseph Hutchinson, who argued that most decisions were decided on the basis of a 'hunch' about the appropriate result, and of Frank, who argued strenuously that a judicial opinion was but a rationalization of a decision reached in another way, that formalism was motivated by an infantile search for paternal authority and that a truly adult judge would recognize and expose the many factors, particularly the idiosyncrasies of his response to the facts proven at trial, that lead to the choice that was made in an individual case. Together these observations suggested that judicial decisions obscured behind a web of false necessity the real factors that operated to generate results in particular cases.

The sociological attack, which was more muted than either of the other two, attempted to identify the factors at work behind the facade of the judicial opinion. Here the most common argument implicitly drew on the work of William J. Ogburn, who argued that the structure of social understandings of the world commonly lagged behind the practices that these understandings purported to describe. Versions of this concept of 'cultural lag' were offered by the Realists to explain how it was that judges utilized an 'outdated' economic theory to understand labor conditions or commercial transactions, an 'outdated' psychological theory to understand witness testimony, or an 'outdated' political theory to limit governmental action to address the conditions of modern life. It was here that the Realists' jurisprudence, as that word is traditionally used, met up with their interest in empirical research. Faced with the facts gathered from empirical research, it was assumed that judges and other legal actors would come to abandon their outdated theories.

The Realists' remedy for the inadequacies of formalism was functionalism. The problems with formalism were seen to be the lack of relationship between the concepts and categories of the law and the 'realities' of modern life. Closer attention to these realities would allow one to focus the law's attention on institutions, such as the organized activities of businessmen in a trade association, and on the functions that these activities represented, and so see to it that the law addressed directly the problems in that trade as a matter of effective and appropriate social policy. Such an analysis would disclose just when the law needed to be updated because of its lag behind social practice and also would expose circumstances where law was dysfunctional, where a body of doctrine was ineffective, as, for example, when Arnold argued that the antitrust laws provided the comfort of a myth that law was doing something to address public fears about big business, while ignoring the anticompetitive practices that made for monopoly.

For all the talk about functionalism as an antidote to formalism as a jurisprudence, indeed for all that Realism is taken to be a jurisprudence, it is important to understand that much of the impetus behind realism was educational. The Realists were mostly law professors for whom the question of how to present the materials of law study was a daily concern. It is with respect to legal education that functionalism and empirical research might have been expected to have some real bite. Functionalism raised questions about what were the appropriate categories under which to group the legal materials, and a study of the entire curriculum at Columbia was undertaken to examine how the curriculum could be organized more functionally. Similarly, empirical research was designed to answer the questions about the efficacy of the institutions and rules of law in the aid of understanding what social policy was appropriate in each functionally defined area.

Although almost no part of the curriculum was rearranged functionally, indeed, for example, the course in trusts is still taught despite the absolutely withering critique that Arnold leveled against it, and though law professors do little more empirical research today than was the case before World War I, it is in

the area of education that Legal Realism had its greatest impact. The critique of formalism utterly changed the way that law was justified in the law school classroom. From a formalism based on assertedly logical derivation from the nature of a given legal category the justification of legal rules has shifted to a realism based on assumed knowledge of social conditions to which law is applied in the pursuit of supposedly agreed upon social policy. The law professors now talk the language of the early twentieth century political science professors, as well as of the bureaucrats in Washington and in scattered state capitals. Whether this is an advance is for others to say (see *Legal Education*).

See also: Law and Society: Sociolegal Studies; Legal Formalism; Legal Positivism; Legal Scholarship

Bibliography

Duxbury N 1995 *Patterns of American Jurisprudence*. Clarendon Press, Oxford, UK

Holmes O W Jr. 1897 The path of the law. *Harvard Law Review* **10**: 457–78

Horwitz M J 1992 *The Transformation of American Law 1870–1960: The Crisis of Legal Orthodoxy*. Oxford University Press, New York

Hull N E H 1997 *Roscoe Pound and Karl Llewellyn: Searching for an American Jurisprudence*. University of Chicago Press, Chicago

Kalman L 1986 *Legal Realism at Yale 1927–1960*. University of North Carolina Press, Chapel Hill, NC

Leiter B 1997 Rethinking legal realism: Toward a naturalized jurisprudence. *Texas Law Review* **76**: 267–315

Pound R 1908 Mechanical jurisprudence. *Columbia Law Review* **8**: 101–16

Pound R 1910 The law on the books and the law in action. *American Law Review* **44**: 12–36

Pound R 1911–12 The scope and purpose of sociological jurisprudence. *Harvard Law Review* **24**: 591–619, **25**: 140–68, **25**: 489–516

Purcell E A Jr. 1973 *The Crisis of Democratic Theory: Scientific Naturalism and the Problem of Value*. University of Kentucky Press, Lexington, KY

Rumble W E Jr. 1968 *American Legal Realism: Skepticism, Reform and the Judicial Process*. Cornell University Press, Ithaca, NY

Schlegel J H 1995 *American Legal Realism and Empirical Social Science*. University of North Carolina Press, Chapel Hill, NC

Stevens R 1983 *Law School: Legal Education in America from the 1850s to the 1980s*. University of North Carolina Press, Chapel Hill, NC

Twining W 1973 *Karl Llewellyn and the Realist Movement*. Weidenfeld and Nicholson, London

Verdun-Jones S 1979 Cook, Oliphant and Yntema. *Dalhousie Law Journal* **5**: 3–44

J. H. Schlegel

Legal Reasoning and Argumentation

1. Reasoning in Adjudication

It is sometimes said that the main aim of legal education is to develop skills of 'thinking like a lawyer,' including skills of legal reasoning. This phrase is often taken to imply that all lawyers think alike, that they only think about questions of law in the context of adjudication; that there is a single correct way of thinking about such questions; and that this way of thinking is unique or special to lawyers. An alternative view is that lawyers' reasoning extends far beyond binary questions of law in adjudication to a wide variety of legal contexts and operations; that what constitute valid, cogent, and appropriate modes of reasoning in each kind of context, and how far rationality is attainable, is contested; and that the relevant skills of reasoning involved are not a lawyers' monopoly. Special considerations apply in particular legal contexts, such as rules of precedent or evidence or procedure, but the basic criteria of validity and cogency for all of these operations can be subsumed under one or other general theory of practical reasoning. This article proceeds on assumptions that are closer to the second view but, solely for reasons of space, it focuses on reasoning about questions of law and questions of fact in adjudication. It does not deal with the psychological processes by which adjudicators reach decisions.

Reasoning in adjudication is a focal point for some perennial problems of legal philosophy, including contested questions about epistemology, rationality, and skepticism; differing conceptions of law and justice; the relations between law and morality; formalism; the role of judges in a democracy; official discretion; and transparency in decision-making.

According to one view, a question of law is easy or straightforward if a justification for a decision can be rationally reconstructed in the form of a simple syllogism:

Major premise: Whenever X happens, then Y ought to happen (Rule (R)).

Minor premise: X happened (Facts (F)).

Conclusion: Therefore Y ought to happen (Judgment of guilt, liability etc.).

This model of legal reasoning has often been dismissed as 'mechanical' jurisprudence, because most problems of interpretation arise in relation to doubts about the formulation and precise interpretation of the major premise (R) and the categorization of the facts in the minor premise (F). However, this is precisely the form in which justifications for decisions in 'easy' or 'clear' cases can be rationally reconstructed.

What if the major premise or the minor premise requires justification? There is quite widespread, but not universal, agreement among jurists that deduction has only a limited role in such second order justifi-

cation of R (legal propositions) and in inferential reasoning from evidence about disputed issues of fact (F). For example, the standard alibi defense can be reconstructed in the form of two linked syllogisms:

No person can be in two different places at the same moment of time.

A was in a different place when this crime was committed.

Therefore A was not physically present when this crime was committed.

It is necessary for a person to be physically present to be guilty of this crime.

A was not physically present when this crime was committed.

Therefore A is not guilty of X.

Insofar as deduction plays a limited role in justifying R or F, the reasoning involved does not lead to necessary conclusions ('open system reasoning').

2. *A Rationalist Model of Reasoning in Adjudication*

Orthodox rationalist views about reasoning in adjudication can be reconstructed as a model or 'ideal type,' which the views of many leading jurists approximate. This model represents a set of basic assumptions about the aims and nature of adjudication and what is involved in valid reasoning about disputed questions of law and of fact in this context.

Aim and Nature of Adjudication

(a) The direct end of adjudication is rectitude of decision, that is the correct application of valid laws to true material facts (facts in issue).

(b) The logic of justification of judicial decisions involves an application of general principles of practical reasoning in a specific context.

(c) A judicial decision is legitimate if and only if it is justified by sound arguments that satisfy the moral requirement of formal justice that like cases should be treated alike.

(d) In clear cases a sound justification satisfies the deductive form: if R, then C; F is a case of R; therefore C.

(e) Doubts can arise about R (e.g., about its validity, its identity, its scope) or about its application to F, or about whether F has been established to the relevant standard of proof, or a combination of these.

(f) Doubts about R give rise to questions of law; doubts about F give rise to questions of fact.

(g) In adjudication both questions of law and questions of fact are typically binary (e.g., liable/not liable; guilty/not guilty).

Questions of Law

(h) Resolution of questions of law in hard cases requires second-order justification, which may involve different forms of reasoning, notably deduction, induction, reasoning by analogy, or a combination of these.

(i) A number of independent material reasons may be advanced in a single argument for or against a decision of a question of interpretation or application of R. These reasons may be of different kinds.

(j) The validity and cogency of different kinds of material reasons are system-specific; they depend on the rules and conventions of a given system with respect to authoritative sources of law and appropriate modes of argumentation.

(k) Second order justification of R may involve two kinds of reasons: (a) authority reasons which relate to the compatibility or fit of a line of argument with the rules and principles of the system and the authoritative sources recognized by that system (arguments about validity, consistency, and coherence); (b) substantive reasons, which relate to values, goals, or institutional requirements of the system (Summers 1992).

Questions of Fact

(l) Rational determination of the truth of alleged facts (F) in adjudication is typically a matter of probabilities, falling short of absolute certainty.

(m) Judgments about the probabilities of allegations about particular past events can and should be justified by reasoning from relevant and admissible evidence considered by the decision-maker.

(n) Judgments about probabilities, generally speaking, have to be based on the available stock of knowledge about the common course of events.

(o) Rational determination of F is a matter of ordinary common-sense reasoning, as modified by the rules of evidence and procedure of a given system.

2.1 *Questions of Law*

Theorists of reasoning about questions of law can be roughly categorized into *rationalists* and *critics*. A reasonably representative sample of leading rationalists might include Aulis Aarnio (Finland), Robert Alexy (Germany), Ronald Dworkin (USA/Britain), Torstein Eckhoff (Norway), Neil MacCormick (Scotland), Joseph Raz (Israel/Britain) and Robert Summers (USA). Although they come from different legal traditions and quite varied intellectual backgrounds, rationalists are in broad agreement on most of the first ten points in the model. Most focus mainly on second-order justification of adjudicative decisions on questions of law in hard cases, because these are perceived to be the most interesting theoretically. These theorists may differ about which kinds of substantive reasons are valid, about their relative weight or cogency, and whether substantive reasons are an integral part of the law or are extraneous to it. Some jurists in the civil law tradition maintain that all second-order justification

can be reconstructed in deductive form, but the predominant view among common lawyers is that this kind of justification involves mainly open system reasoning.

Among rationalists there is a variety of views on material reasons in second-order justification (k). Some differences are matters of emphasis or detail or conceptual refinements, but others are seen as significant. For example, Ronald Dworkin (1977, 1986) has advanced three theses that are regularly disputed: (a) that there is a single right answer to almost every question of law even in very hard cases; (b) that principles of political morality underlying a legal system are an integral part of the law, and are not external to it; (c) that principles of political morality are the only valid substantive reasons for justification in adjudication; policy reasons and other arguments external to the law are not valid. These theses are part of a general theory of adjudication that maintains that legal reasoning is essentially a moral enterprise; the duty of judges in a liberal democracy is to uphold the law and to vindicate rights on the basis of interpretations that make 'it the best it can be.'

At first glance, Dworkin's theory is in conflict with standard tenets of legal positivism, especially that (a) a sharp distinction must be drawn between law as it is and law as it ought to be; (b) when the law 'runs out,' judges may invoke non-legal reasons, including consequentialist reasons; and (c) in such circumstances judges can, may, and do make law interstitially. However, the differences between Dworkin and other rationalists are narrow. Most rationalists agree that moral arguments play a valid role in legal justifications. If 'one right answer' means a solution justified by the strongest conceivable argument, most would accept that it is meaningful to talk of stronger and weaker arguments, and that *ties* are a rarity even with respect to open-system reasoning (see *Legal Positivism*).

The rationalist model has been subject to several challenges. First, it is narrow. It does not address modes of reasoning in earlier times or in non-western cultures or in the variety of other legal contexts in which practical reasoning is involved. Interpretation by an upright judge may provide a general model of correct interpretation for other officials and for good citizens, but what constitutes valid, cogent, and appropriate reasoning in adjudication will not be exactly the same for other interpreters. For example, a factory or a tax inspector may appropriately take into account different considerations, and interpret the same law differently from a judge.

Second, skeptics may doubt the possibility of rationality in adjudication or challenge the idea that this represents the only possible, or the most appropriate, model of rationality or that a single model is applicable across cultures. Suggestions that judicial opinions are merely *ex post facto* rationalizations have been shown to be based on confusion between the logic of justification and the psychology of decision (Wasserstrom 1961). The requirement or expectation that decision-makers should justify their decisions publicly may serve as an institutional constraint or steadying factor that limits arbitrariness or corruption or pure subjectivity (Llewellyn 1962, Raz 1978). Recently, critical and postmodern challengers have argued that rationalist models conceal the fundamental indeterminacy of legal decision-making. For example:

> In all the Western systems, the discourse that judges, legal authorities, and political theorists use to legitimate the application of state power denies (suppresses, mystifies, distorts, conceals, evades) two key phenomena: (a) the degree to which the settled rules (whether contained in a Code or in common law) structure public and private life so as to empower some groups at the expense of others, and in general function to reproduce the systems of hierarchy that characterize the society in question; (b) the degree to which the system of legal rules contains gaps, conflicts, and ambiguities that get resolved by judges pursuing conscious, half-conscious, and unconscious ideological projects with respect to these issues of hierarchy. (Kennedy 1997) (see *Critical Legal Studies*)

The scope of disagreement between rationalists and their critics is easily exaggerated, for few jurists subscribe to extremes of strong philosophical skepticism or of radical indeterminacy on the one hand or of strict rationalism and objectivity on the other.

A third challenge suggests that the distinction between clear and hard cases is problematic. Are cases inherently hard or easy? Or is it that some cases are merely perceived or treated as easy and others as hard? Cannot a creative lawyer construct arguments that raise doubts about what was previously perceived to be settled? Such skepticism receives *prima facie* support from a consideration of the many different kinds of doubt that can arise in relation to the interpretation of both legal and non-legal rules (Twining and Miers 1999, Chap. 6). Even in a seemingly simple situation, a single factor can be a starting-point for an argument that raises a doubt about what previously may have been treated as clear or settled or simple; but it does not follow from such indications of *potential* complexity that in practice every routine or seemingly clear case can easily be treated as hard. There are indeed cases in which a significant development of the law took place because what was previously taken as settled was successfully challenged. But these are 'leading cases' just because they were exceptional.

A fourth type of challenge suggests that normative theories of legal reasoning tend to be too abstract and simplistic to catch the complexities and subtleties of actual legal arguments. Judicial opinions that run to many pages are 'reconstructed' in just a few propositions. Furthermore, it is often unclear whether the reconstruction is of *actual* or of *possible* arguments. Is a reconstruction of a formalistic, ambiguous, or opaque judicial opinion, or of a decision not backed

by any public justification (such as jury verdicts at common law), to be regarded as a reconstruction of the actual justification presented, or of the best justification that might have been advanced, or of something else? For example, the discursive style of appellate judges at common law is often contrasted with the more succinct style of French judgments which, it is said, are typically expressed in deductive form. If hard cases require a second order justification that can only exceptionally be deductive, this suggests that French courts typically do not indulge in explicit second order justification.

2.2 *Questions of Fact*

The study of evidence and inference has roots in the long history of rhetoric and of probabilities, but interest in reasoning about questions of fact in adjudication has been intermittent. Evidence has received more sustained attention in common law countries than in the tradition of civil law, but there the main focus has been on the Law of Evidence. Notable exceptions include Bentham's *Rationale of Judicial Evidence* (1826) and Wigmore's *Science of Judicial Proof* (1937). Since 1970 'The New Evidence Scholarship' has been the forum for lively debates about the competing claims of different theories of probability; the thesis that the paradigm case of probabilistic reasoning in adjudication is in principle inductive (Baconian) rather than mathematical (Pascalian) (Cohen 1977); and whether 'atomistic' models of rationality can accommodate 'holistic' ideas of narrative coherence on which, according to psychological research, decision-making by jurors is predominantly based (e.g., Pennington and Hastie 1991). Such legal concerns are relevant to the study of evidence and inference in other disciplines (Schum 1994).

Here, three points deserve emphasis. First, Anglo-American ideas in this area have recently spread to Continental Europe (e.g., Malsch and Nijboer 1999), but there has been almost no sustained comparison between conceptions of evidence and inference in law in modern Western and other traditions. Second, as with reasoning about questions of law, a high proportion of legal scholarship about evidence and inference in law shares a set of assumptions that approximates to the rationalist model of reasoning in adjudication. Third, these assumptions are similarly open to various skeptical or critical challenges including (a) strong philosophical skepticism about the possibility of knowledge or of rational argument or of objectivity in this context; (b) ideological skepticism about the coherence and tenability of 'liberal legalist' ideas about adjudication; (c) skepticism about claims that the main end of adjudication is implementation of law and vindication of rights; (d) legal fact-skepticism about the feasibility of designing institutions, procedures and rules that will, in practice, regularly maximize the probability of achieving correct decisions (Frank 1949); and (e) contextual skepticism about the sense of studying adjudicative decisions outside the broader context of total legal processes and broader social processes. To these might be added further challenges associated with post-modernism, feminism, cultural relativism, and developments in the philosophy of science (Twining 1994).

3. *Law and Fact*

At first sight, the distinction between law and fact seems simple. 'What is the scope of this statutory provision?' is a question of law; 'What happened on that particular occasion?' is a question if fact. However, the line is often difficult to draw, especially with regard to the application of rules and the appropriate categorization of a particular fact situation. 'Was X dishonest?' or 'Is this a case of murder?' could be interpreted as a question of law or a question of fact, or a mixed question of law and fact. How such questions are treated in particular cases varies between systems and among different contexts within the same system for different practical purposes. For example, in English jury cases, questions of law are for the judge while questions of fact are for the jury; decisions on questions of law can have precedent value, but findings of fact do not. Questions relating to appeals, mistake in contract, or the obligation to give reasoned justifications turn on the distinction between law and fact, but where exactly the line is drawn varies according to context. Furthermore, legal cultures differ significantly in the framing of issues in particular disputes as questions of law or of fact (e.g., Dezalay and Garth 1996).

Reasoning about issues of law and of fact have stimulated two largely separate bodies of literature, despite the intimate connection and similarities between them. Both are perceived to be species of practical reasoning with structures that are binary (guilty/not guilty, liable/not liable) and dialectical (all reasons tend to support or to negate an ultimate proposition or conclusion). The ultimate decision is a normative judgment (e.g., liable/not liable). Many questions of fact have normative elements: for example, 'did X behave reasonably in the circumstances?' is a question of fact, which involves both evaluative and empirical elements. Categorization of an issue as being of fact or of law can have important practical consequences, and the kinds of reasoning that are deemed appropriate differ; for example, no one has seriously suggested that Bayes' theorem applies to questions of law, but nevertheless it might be interpreted as providing a suggestive metaphor regarding the combination of judgments based on different kinds

of reasons within a single argument. Where decisions on fact, law, and sanctions are the responsibility of a single tribunal, there may be some freedom to consider what is a good solution to the problem of the case as a whole, an option not generally available where responsibility for decisions is divided between judge and jury.

4. Argumentation

Theories of legal reasoning are primarily normative theories. 'Argumentation' here refers to the actual discourses used in advancing arguments, including reasonings that are explicitly or implicitly embodied in such discourses, nonrational means of persuasion, and the strategy, tactics and styles of argument. Argumentation can be studied from a wide range of empirical, interpretive, and critical perspectives that apply to social discourse generally including rhetoric, conversation analysis, semantics, and semiotics. See *Ethnomethodology: General*; *Logic and Linguistics*; *Rhetoric*. There has been little sustained research into actual discursive practices of legal actors and, as with the normative literature, most studies have focused on argumentation in adjudication and advocacy.

Most normative theorists claim that their accounts have a more or less close relation to how judges and advocates actually argue. For example, Ronald Dworkin (1986) boldly claims that his theory not only prescribes an ideal, but also describes 'best practice' in common-law courts; however this claim is not backed by evidence. MacCormick (1978) uses actual cases from several jurisdictions as illustrations, but his sample is not representative. Some analysis of judicial styles has been based on fairly extensive random samples (e.g., Llewellyn 1962, Wetter 1960), but these studies have not been replicated.

Explicit links between normative theories of reasoning and actual discourses have been both diverse and intermittent. In social theory the most important connection has been Max Weber's thesis that formal rationalization of law is a concomitant of the rise of capitalism and modernity (see *Weber, Max* (1864–1920)). Critiques of rationalist theories of adjudication are often part of a more general critique of 'liberal legalism' and the Rule of Law (e.g., Kennedy 1997). Closest to the orthodox juristic literature is the neo-Aristotelian 'New Rhetoric' of the Brussels School (e.g., Perelman and Olbrechts-Tyteka 1969, and in Germany, Viehweg [1953] 1993), whose central concern is to describe the starting-points (*topoi*), argumentation schemes, and material factors that can be deployed to try to convince a specific audience (such as a court) or, more speculatively, a universal audience, to do or to decide something. Sharing some intellectual ancestors with the New Rhetoric but being more empirically oriented, the Amsterdam School of Pragma-dialectics aims to develop a model for the rational analysis and evaluation of legal argumentation as a specific, institutionalized form of argumentation in general (e.g., van Eemeren et al. 1996). This is part of a general effort to develop a model and code of conduct for rational discussants in a conversation intended to resolve disputes. Applied to law, this model sits somewhat awkwardly with robust adversarial argument. The self-imposed limitation of analyzing what is actually said ensures that the approach is quite concrete, but makes it difficult to catch unspoken conventions.

5. Conclusion

There is a rich body of normative theories about reasoning in modern Western adjudication, but more work is needed to integrate the literature on reasoning about questions of law and questions of fact and to extend the focus of attention to other legal operations and other legal traditions. With respect to argumentation, one can agree with two leading theorists that 'what is needed are reflective hermeneutic studies of the current actuality of reasoning practice in modern legal systems' (Aarnio and MacCormick 1992).

Bibliography

Aarnio A, MacCormick D N (eds.) 1992 *Legal Reasoning*. Ashgate, Aldershot, UK (a useful anthology)

Alexy R, *A Theory of Legal Argumentation* (transl. Adler R, MacCormick N). Oxford

Cohen J 1977 *The Probable and the Provable*. Oxford

Detmold M 1989 Law as practical reason. *Cambridge Law Journal*. **436**

Dezalay Y, Garth B 1996 *Dealing in Viture: International Commercial Arbitra and the Construction of a Transnational Legal Order*. University of Chicago Press, Chicago

Dworkin 1977 *Taking Rights Seriously*. Duckworth, London, UK

Dworkin R 1986 *Law's Empire*. Fontana, London

van Eemeren F H, Grootendorst R, Henkemans F S 1996 *Fundamentals of Argumentation Theory*. Erlbaum, Mahwah, NJ

Frank J 1949 *Courts on Trial*. Princeton University Press, Princeton, NJ

Jackson B 1985 *Semiotics and Legal Theory*. Routledge, London

Kennedy D 1997 *A Critique of Adjudication*. Harvard University Press, Cambridge, MA

Levi E H 1948 *Introduction to Legal Reasoning*. University of Chicago Press, Chicago

Llewellyn K N 1962 *The Common Law Tradition: Deciding Appeals*. Little Brown, Boston

MacCormick D N 1978 *Legal Reasoning and Legal Theory*, 2nd edn. Oxford University Press, Oxford, UK

MacCormick D N, Summers R (eds.) 1997 *Interpreting Precedents*. Dartmouth, MA

Perelman C, Olbrechts-Tyteca L 1969 *A New Rhetoric: A Treatise on Argumentation* (transl. Wilkinson J, Weave P). University of Notre Dame Press, London

Raz J 1978 *Practical Reason and Norms*. Hutchinson, London

Samuel G 1994 *The Foundations of Legal Reasoning*. Metro, Maklu
Schum D 1994 *Evidential Foundations of Probabilistic Reasoning*. Wiley, New York
Summers R S 1992 *Essays on the Nature of Law and Legal Reasoning*. Duncker and Humbolt
Susskind R 1987 *Expert Systems in Law, A Jurisprudential Inquiry*. Oxford
Twining W 1994 *Rethinking Evidence*. Northwestern University Press, Evanston IL
Twining W, Miers D 1999 *How To Do Things with Rules*, 4th edn. Butterworth, Oxford, UK
Viehweg T 1993 *Topics and Law: A Contribution to Basic Research in Law* (transl. Durham W. C. Jr.). Lang, Frankfurt am Main, Germany
Wasserstrom R 1961 *The Judicial Decision*. Stanford University Press, Stanford, CA
Wetter J G 1960 *The Styles of Appellate Judicial Opinions*. Sythoff, Leiden, The Netherlands
Wigmore J H 1937 *The Science of Judicial Proof*, 31st edn. Little Brown, Boston, MA

W. L. Twining

Legal Reasoning Models

1. Introduction

Research in legal reasoning models investigates formal and computational theories of how legal experts analyze problems, create arguments, and make decisions. The formalization of law and legal reasoning is not a new enterprise, but rather one that has challenged philosophers of law throughout the twentieth century. Wesley Hohfeld (1913), a legal positivist, proposed that all legal states of affairs can be described in terms of four fundamental relations: rights, duties, powers, and liabilities (with inverse relations: no-rights, privileges, disabilities, and immunities). In contrast, legal realists such as Jerome Frank (1931) argued that law cannot be characterized as a formal system, because the infinite variety of particular facts on which legal decisions turn make legal reasoning inherently unpredictable. The problem of open-textured concepts (i.e., legal concepts that do not have clear definitions to determine their applicability, but which depend on experience and common sense, such as the concept of recklessness) is the focus of a famous debate between H. L. A. Hart and Lon Fuller (Hart 1958, Fuller 1958), in which a legal rule stating 'No vehicles are permitted in the park' is used to demonstrate that even an apparently clear concept such as 'vehicle' can be legally indeterminate, as in the case where a veterans group wants to drive a tank through the park as part of a Memorial Day parade.

With the advent of computers and artificial intelligence, both the range of techniques available and the practical motivation for creating legal reasoning models have grown. Rule-based systems using logical inference techniques have been developed and used successfully in a number of legal domains, especially those dominated by complex regulations such as taxation and social benefits administration (Johnson and Mead 1991). A unique challenge in developing legal rule-based systems is the need to model the relationship between the logical representation and its textual sources (Bench-Capon and Coenen 1992). This need arises first, because legal experts (and legal expert systems) must be able to cite the authoritative texts on which their conclusions are based, and second, to enable legal knowledge bases to be efficiently updated when laws and regulations are amended.

Despite the successful adaptation of rule-based reasoning models to the legal domain, it is the view of most researchers that the essence of legal reasoning is its open-textured and indeterminate nature (Berman and Hafner 1987). Two essential cognitive abilities for legal reasoning models include:

(a) reasoning with cases to interpret open-textured or conflicting legal rules and concepts

(b) adversarial reasoning: the ability to create persuasive arguments for both sides of an issue.

2. Case-based Legal Reasoning Models

Case-based reasoning (CBR) research in artificial intelligence (see also *Problem Solving and Reasoning: Case-based*) is concerned with the use of remembered prior experiences in analyzing new situations and solving problems. The distinguishing characteristic of case-based reasoning is that prior experiences are stored as distinct histories or 'cases,' in contrast to other AI reasoning frameworks where knowledge (however acquired) is integrated into a unified structure of logical rules, frames, neural nets, or combinations of these.

Research in case-based legal reasoning has a more restricted meaning: formal judicial decisions (i.e., legal cases) play the role of remembered experiences, and the problem-solving goal is the application of a prior case to answer legal questions about a new case. In legal CBR models, a set of domain-dependent, legally-relevant features is defined, and a prior case is considered more or less *on point* depending on the degree of match between the features of the prior case and those of the new case. Computational models of legal CBR are concerned with the following problems: representation of cases, including their facts, issues, *holdings* (the judicial conclusions they contain), and results; organization of the case database, including the indexing structure and retrieval algorithms; ranking of retrieved cases by their similarity to a new case; and creation of case-based legal arguments.

The first legal CBR model that addressed all of these problems was embodied in HYPO (Rissland and Ashley 1987, Ashley 1990), a computer program for analyzing cases and constructing legal arguments in

the domain of trade secrets law. In HYPO, legally relevant features called *dimensions* are used to index and retrieve cases. HYPO's legal domain knowledge includes, for each feature, which side of a dispute would be favored by the feature's having particular values (e.g., a *true* value for predicates, or a larger value for number-valued features), and how the feature's value is related to a case's underlying facts. For example, in the trade secret domain, one such feature is *Generally-Known*, indicating that the alleged secret information is generally known within the industry. When a new case is presented to HYPO, the features applicable to the new case are computed. Then, HYPO constructs a 'claim lattice'—a similarity network that has the new case as its root node, and prior cases that share a maximal set of feature-values with the new case as its immediate successors. These immediate successors are taken to be the *most-on-point cases*.

Later research extended the use of HYPO-style CBR in several ways. In the GREBE system (Branting 1991), a representation of the *ratio decidendi* (the rationale or justification) of cases was used to identify portions of a prior case's holdings that were relevant to a new case, even when other portions of the prior case were not on-point. Branting's model addressed a well-known weakness of the pure feature-based approach to CBR: its failure to represent the reasoning that connects the features with the results of a case. The logic underlying a legal decision is implicitly encoded in GREBE by grouping the features of a precedent according to the holdings they influence. Berman and Hafner (1993) addressed another aspect of legal CBR: the need for legal reasoning models to include the societal goals and policies that motivate legal rules and decisions. They proposed an extended representation of cases which incorporates knowledge about legal purposes.

Several researchers have developed models of legal reasoning in which rules and cases are used in combination. Gardner (1987) used a logic tree to model the top-down, rule-based analysis of a legal situation. At the leaves of the tree are undefined concepts, linked to cases providing positive and negative examples. In the CABARET system, Skalak and Rissland (1991) described a mixed strategy for legal argumentation using both rules and cases. When a rule-based strategy produces a result favoring the opposing side, CABARET uses precedents to support argument moves such as the broadening of a rule's interpretation or the discrediting of a rule.

3. Models of Legal Argumentation

Argumentation is considered by many law teachers and legal scholars to be the quintessential element of a lawyer's cognitive skill. Thus, many models of legal argumentation have been developed, falling into three broad categories: case-based, logic-based, and legal discourse models.

Case-based models of legal argumentation focus on the process of arguing with precedents, using techniques such as analogizing, distinguishing, and trumping. A case-based model of legal argumentation was developed as part of the HYPO system described above. In HYPO, after the claim lattice is created to determine the most-on-point precedents, HYPO constructs a three-ply argument for the new case. Letting Side 1 represent the party on whose behalf HYPO is arguing, first a *Point for Side 1* is produced making a claim that Side 1 should win, next a *Response for Side 2* is produced, countering the claim, and then a *Rebuttal for Side 1* is produced to answer the response. To create the initial point for Side 1, HYPO identifies the most on point case whose result favored Side 1, and makes an argument by analogy, noting the similarity between the new case and the prior case. Responses and rebuttals use the legal argument techniques of *distinguishing* by focusing on differences rather than similarities, and *trumping* by presenting more on-point counterexamples (Ashley 1990, p. 71).

Logic-based models of legal argumentation begin by expressing arguments as formal proofs within a system of logic, and then focus on the common situation in law where two logical arguments exist which reach contradictory results. In order to resolve such contradictions, it is necessary to adopt a framework of defeasible logic (Reiter 1980), where rules may have exceptions, and beliefs may be defeated by the acquisition of new, conflicting knowledge. For example, a legal rule 'No vehicles are permitted in the park' may co-exist with another legal rule 'Emergency vehicles are permitted to violate traffic laws.' Two logical arguments can then be constructed regarding an ambulance that entered the park: one concluding that the ambulance violated the law, the other concluding that the ambulance acted legally.

A logical argumentation framework is described in Prakken (1993), where two conflicting arguments, represented as formal proofs, can be compared using one or more ordering principles applied to the legal premises of the two arguments. (Traditional ordering principles found in legal texts include *Lex Superior* (e.g., federal laws trump state laws), *Lex Specialis* (more specific laws trump more general laws), and *Lex posterior* (newer laws trump older laws); however, Prakken's framework allows for any criterion whatsoever that can prefer one argument to another.) Given a set of ordering principles, Prakken's framework classifies legal arguments into three categories: justified arguments (which cannot be defeated), defeated arguments, and defensible arguments (which are neither justified nor defeated).

Legal discourse models are concerned with the structural and semantic characteristics that give a legal argument legitimacy. A number of argumentation models have been developed based on Toulmin's

(1958) theory of justified arguments. In Toulmin's theory, an argument's conclusion is called the *claim*; the *grounds* are the facts asserted to support the claim; the *warrant* is a bridge-like statement (e.g., a legal rule) that provides the logical connection between the grounds and the claim; and the *backing* is the authority for the warrant (e.g., a statute or regulation).

Loui and Norman (1995) have developed a discourse-based model of legal argumentation that focuses on defining categories of rationales (i.e., warrants) used in adversarial legal arguments, and the forms of attack appropriate for each. They identify several types of legal rationales: for example, the *compression* rationale justifies a simple and general rule as a normal-case simplification of a more complex and specific one. An example of a compression rationale described by Loui and Norman (p. 161) is the rule 'vehicles are not allowed in the park,' which (one might argue) is a normal-case simplification of the more complex rule: 'vehicles used for private transportation are not allowed in the park,' since normally vehicles are used for private transportation. A straightforward strategy for attacking the 'no-vehicles' rule (as in the case of the Memorial Day tank display) is to restate the rule in its uncompressed form, and argue that the case at hand falls outside the normal category of private transportation.

4. Conclusions and Future Directions

The field of law offers a fertile domain for the study of important problems in artificial intelligence and cognitive modeling, such as using prior experiences to solve new problems, reasoning with open-textured concepts, and making persuasive arguments. Promising new directions for research include the development of formal legal ontologies and their use in knowledge-based systems (Visser and Bench-Capon 1998) and the use of explicit legal concepts and models (such as contracts) to govern the behavior of systems for electronic commerce.

See also: Decision Support Systems; Judges; Legal Reasoning and Argumentation; Problem Solving and Reasoning: Case-based; Rules in the Legal Process

Bibliography

Ashley K D 1990 *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. MIT Press, Cambridge, MA

Bench-Capon T J M, Coenen F P 1992 Isomorphism and legal knowledge based systems. *Artificial Intelligence and Law* **1**: 65–86

Berman D H, Hafner C D 1987 Indeterminacy: A challenge to logic-based models of legal reasoning. In: *Yearbook of Law Computers and Technology*, Butterworths, London, Vol. 3, pp. 1–35

Berman D H, Hafner C D 1993 Representing teleological structure in case-based legal reasoning: the missing link. *Proceedings of the 4th International Conference on Artificial Intelligence and Law*. ACM Press, New York, pp. 50–60

Branting K 1991 Reasoning with Portions of Precedents. *Proceedings of the 3rd International Conference on Artificial Intelligence and Law*. ACM Press, New York, pp. 145–54

Frank J 1931 *Law and the Modern Mind*. Brentano, New York

Fuller L 1958 Positivism and fidelity to law—a reply to Professor Hart. *Harvard Law Review* **71**: 640

Gardner A van der Leith 1987 *An Artificial Intelligence Approach to Legal Reasoning*. MIT Press, Cambridge, MA

Hart H L A 1958 Positivism and the Separation of Law and morals. *Harvard Law Review* **71**: 593

Hohfeld W N 1913 Fundamental legal conceptions as applied in judicial reasoning. *Yale Law Journal* **23**: 16

Johnson P, Mead D 1991 Legislative knowledge-based systems for public administration—some practical issues. *Proceedings of the 3rd International Conference on Artificial Intelligence and Law*. ACM Press, New York, pp. 108–17

Loui R P, Norman J 1995 Rationales and argument moves. *Artificial Intelligence and Law* **3**(3): 159–85

Prakken H 1993 A logical framework for modelling legal argument. *Proceedings of the 4th International Conference on Artificial Intelligence and Law*. ACM Press, New York pp. 1–9

Reiter R 1980 A logic for default reasoning. *Artificial Intelligence* **13**: 81–132

Rissland E L, Ashley K D 1987 A case-based system for trade secrets law. *Proceedings of the 1st International Conference on Artificial Intelligence and Law*. ACM Press, New York, pp. 61–7

Skalak D B, Rissland E L 1991 Argument moves in a rule-guided domain. *Proceedings of the 3rd International Conference on Artificial Intelligence and Law*. ACM Press, New York, pp. 1–11

Toulmin S E 1958 *The Uses of Argument*. Cambridge University Press, Cambridge, UK

Visser P R S, Bench-Capon T J M 1998 A comparison of four ontologies for the design of legal knowledge-based systems. *Artificial Intelligence and Law* **6**: 27–57

C. D. Hafner

Legal Scholarship

Legal scholarship can be distinguished from scholarship about law, such as the history of law or the sociology of law, by virtue of its internal perspective (Dworkin 1986, Hart 1961). This means that it treats law as a set of norms that can be analyzed according to their internal coherence and moral significance, rather than as one component of a social system whose contours and content are determined by the same forces that affect other aspects of the society. Within the boundaries of this internal approach, there has traditionally been a division between jurisprudence, which concerns the theory of law, and doctrinal studies, which concern specific legal provisions. In recent years, however, both the distinction between internal and external perspectives, and between jurisprudence and doctrinal studies, have been significantly eroded by intellectual developments such as law and economics, critical legal studies, critical race theory,

feminist legal theory, gay legal studies and sociolegal studies. These developments have deployed the methodologies of other disciplines in the analysis of the internal structure and moral significance of law. At the same time, they have made the study of particular legal doctrines more theoretical and rendered legal theory less unified and more subject-specific. While the distinctions between legal scholarship and scholarship about law, and between jurisprudence and doctrinal studies, remain useful, they now refer to bodies of scholarship with generalized family resemblances, rather than distinct methodologies with clearly-defined boundaries.

1. *Jurisprudence and Doctrinal Studies Between 1900 and 1970*

1.1 *The Attack on Formalism*

In the early years of the twentieth century, doctrinal studies in both the continental and Anglo-American worlds were dominated by an approach that can be described as legal formalism, or conceptualism (see *Legal Formalism*). Its essential premise was that the law is an autonomous system, based on general principles, that is accessible to reason and that enables judges to reach uniform and definitive decisions in contested cases (Schauer 1988). Formalism did not necessarily advance the natural law idea that this system represented transcendental truth (see *Natural Law*), but it did claim that law was animated by general principles of justice that transcended particular political or policy positions. In continental Europe, the source of this autonomous system was regarded as each nation's legal code; in Great Britain and the United States, it was the common law. By virtue of the claim that law is autonomous, and the rather close focus on judicial decisions, both versions of formalism were somewhat insulated from major jurisprudential movements such as Marxism (see *Marxism and Law*) and sociological jurisprudence.

The attack on formalism came from two sources. The first, generally described as legal realism, asserted that judicial decisions and legal concepts are indeterminate or contextual, that they are the product of social attitudes, rather than reasoned analysis, and that they are subordinated to the political process, rather than transcending it (see *Legal Realism*). American legal realism emphasized the indeterminacy of judicial decisions and their dependence upon the particular views of the judge. Continental realism, which was centered in Scandinavia, emphasized the contextual nature of legal concepts and their dependence upon more general social attitudes. The second source of the attack was sociological jurisprudence, which often did not attack formalism *per se*, but rather attempted to develop an alternative conception of law as determined by, and contributing to, more general social forces (Ehrlich 1975). This was connected with the work of anthropologists and sociologists who treated law as an aspect of social systems, particularly those without formal legal institutions. A number of leading American realists saw the sociological approach as an alternative to doctrine, and a few tried to study legal behavior as a sociological phenomenon (see Schlegel 1995).

Apart from this sociological approach, the dominant theme in early twentieth-century jurisprudence was legal positivism (see *Legal Positivism*). According to Kelsen, law is simply the set of threats promulgated by public officials whose authority is established by a single, agreed upon principle, or *Grundnorm* (Kelsen 1967) As such, it is completely separate from morality; a particular normative statement is a law if, and only if, it has been duly promulgated by a recognized public official, or by a private party who has been delegated to act by such a recognized official. Legal positivism was motivated partially by larger philosophic trends (such as logical positivism) and partially by the effort to refute natural law jurisprudence. While its connection with doctrinal studies is not immediately self evident, it seems consistent with American realists' rejection of general common law, that is, the idea that legal rules can be derived from general principles and need not correspond to the decisional law of any specific jurisdiction, and with the Scandinavian realists' insistence that there are no overarching legal concepts, but only concepts specific to particular societies.

1.2 *The Rehabilitation of Law*

The years following World War II saw renewed interest in the concept of law in both jurisprudence and doctrinal studies. In jurisprudence, scholars began searching for a way to re-infuse some sense of social morality into the term law, so that one would not be required to apply that term to the rules promulgated by a generally immoral regime such as Nazi Germany. This led to a revival of natural law (Finnis 1980), but the appeal of this work was limited by its continuing religious overtones (see *Natural Law*). More influential was Fuller's effort to articulate a purely procedural version of natural law (Fuller 1964), which asserted that any declarations that we would be willing to regard as law must have certain universal characteristics such as generality, constancy, and nonretroactivity. Still more influential was Hart's recharacterization of legal positivism in light of Wittgenstein's linguistic theory (Hart 1961). Law, in Hart's view, is a complex set of social rules that not only states the sanctions that public officials will impose, but also the manner in which: (a) they impose these sanctions; (b) new sanctions are developed; (c) private parties are empowered to take actions that will be enforced by public authorities; and (d) the authorities who carry out these various functions are identified.

Doctrinal scholarship rejected legal realism more definitively than jurisprudence rejected legal positivism. In the United States, the legal process school conceded that the legal system was subordinate to politics, thus acknowledging positivism's essential claim, but insisted that the concept of law retained an important role within this more limited sphere of action. Fuller argued that reasoned decision making according to rules established in advance is the best way to adjudicate disputes between parties, although a poor way to formulate new social policy (Fuller 1978). Other members of this school explored the rationales by which the courts could use the US Constitution to evaluate, and, if necessary, strike down legislation, without substituting its political judgment for that of the legislature (Bickel 1962, Wechsler 1959). Central to this inquiry was the belief that doctrine possessed an inherent logic or coherence, so that judges could reach generally accepted conclusions by means of reasoned argument. Implicit in this belief was that public officials, particularly judges, generally acted in the public interest, and that their decisions harmed social welfare only when they strayed beyond the limits of their institutional competence.

In Europe, theories of judicial decision making focused on statutory interpretation, which generally (although not in Britain) meant the interpretation of a comprehensive civil or criminal code. Many scholars asserted, or assumed, that judges could reach generally accepted conclusions by means of reasoned argument if they followed and extrapolated the inherent logic of the code (MacCormick and Summers 1991). As in the case of the American legal process school, this work was premised on the idea that doctrine was coherent within a politically defined area of proper operation. Because it concerned interpretation of a legal code, rather than common law or a vaguely worded constitution, it was more closely allied to legal positivism than its American equivalent. Both the American and European versions of this postwar approach were sufficiently focused upon law's internal rules to minimize the influence of social science studies of law.

2. *The Critique of Law, 1970 to 1985*

2.1 *Law and Economics*

The attack on this rehabilitated concept of law was initiated primarily in the United States, and came from two sources that were associated with opposite ends of the American political spectrum. The first was the Law and Economics Movement, which was derived from the Chicago School of microeconomics and shared its basic premise that all social action is a rational effort to maximize the actor's material self-interest (see *Law and Economics*). According to this analysis, legislators are motivated solely by their desire to be re-elected (assuming that effective sanctions exist against outright bribery); thus, the laws they produce have no inherent logic, and no necessary relation to the public interest. While these laws might correspond to the preferences of voting citizens, they are more likely to reflect the distorting influence of special interest groups that can organize and fundraise more effectively than the general public (Olson 1965). A judge who is interpreting such laws cannot be guided by their internal logic; all she can do is enforce the bargain that the interest groups reached, or interpret the law narrowly in deference to public values (Easterbrook 1983, Macey 1986). Posner originally asserted that the common law, as opposed to statutory law, represented a coherent body of doctrine that implemented the social policy of efficiency (Posner 1972). Other law and economics writers soon abandoned this implausible contention and devoted their efforts to analyzing the self-interested motivations that lead private parties to engage in or to avoid litigation, and that lead lawyers to encourage or discourage them from doing so, to settle cases, or to adopt specific trial strategies (Priest and Klein 1984). This work suggests that common law, like statutory law, lacks any conceptual coherence, and is merely the product of economic forces.

In addition to its analysis of doctrine, the Law and Economics Movement advances the jurisprudential position that the only proper purpose of law, or of government in general, is to achieve economic efficiency, that is, to maximize the nation's wealth (Posner 1972). When combined with the economic analysis of law, this leads to the conclusion that regulatory law is generally undesirable, and that the social control of economic relations should be left to the market. The argument seems more political than legal, in that it addresses the purpose of law rather than its nature, and it is clearly related to the conservative politics of most of the early law and economics scholars. But it also represents a theory of law, in that it advances a general rule that identifies improper laws and improper roles for legal actors.

2.2 *Critical Legal Studies*

The Critical Legal Studies Movement was derived, albeit rather loosely, from the critical theory of the Frankfurt School. It revives the realist claim that law, both statutory and decisional, is not an internally coherent system but a collection of unrelated rules. Unlike legal realism, however, critical legal studies asserts that the reason for this lack of coherence is that law is a device by which the political elite maintains its hegemony, or social control (Horwitz 1977, Unger 1975). It is a particularly effective device, moreover, because the widely-accepted claim that law is coherent or logical masks its class-oriented character; in effect, it creates a false consciousness among the disadvantaged that leads them to accept their subordinated status (see *Critical Legal Studies*). By means of doctrinal analysis, critical legal studies scholars seek

to disprove the claim that legal decisions are logical or coherent, and thereby to reveal the hegemonic purposes of these decisions. In undertaking this analysis, they went beyond the legal realists' demonstration of the internal inconsistencies of doctrine, invoking various aspects of continental thought such as deconstruction and postmodernism to demonstrate that any claim to linguistic coherence is essentially fraudulent.

This analysis of doctrine is linked to the jurisprudential position that the social dominance of the elite, and its use of law to secure its position, is profoundly immoral. A just society achieves equality among its members—if not complete equality, then certainly a much higher degree of equality than contemporary society displays. Thus, modern society needs to be emancipated from its existing social structures and from ideologies, including law, that maintain those structures and preach the legitimacy of those structures to the disadvantaged. Again, this is essentially a political argument, with an obvious connection to the progressive or radical politics of critical legal studies scholars. But again, it constitutes a general theory of law by distinguishing between the improper, hegemonic purposes of law and its proper, emancipatory purposes.

Despite their obvious differences in political orientation, law and economics and critical legal studies rest on closely related insights, and exercise similar impacts on the character of legal scholarship. Both movements renew the realist attack on law's doctrinal coherence, and go beyond realism by grounding this attack on a political analysis. In both cases, moreover, that analysis asserts that law is produced by political and economic elites for the purposes of securing their positions. By thus attributing the incoherence of doctrine to political causes, both law and economics and critical legal studies join doctrinal analysis to larger jurisprudential questions in a manner that was not true for legal formalism, legal realism, or the legal process school. The two movements are further allied in their heavy dependence upon nonlegal methodologies. For the first time, two large groups of legal scholars were prepared to abandon the idea that law is an autonomous or quasi-autonomous system that provides its own analytic tools, and look to other disciplines for the main elements of their analytic framework.

3. *The Topography of Legal Scholarship at the Beginning of the Twenty-first Century*

3.1 *Late Twentieth-century Law and Economics*

Law and economics remains a robust tradition within legal scholarship, and much work continues to be done according to the Chicago School model that assumes rational, self-interested actors and smoothly functioning markets. But the criticisms of this approach by legal scholars, combined with the internal evolution of the field, have led to an increasing body of post-Chicago Law and Economics scholarship. One theme of this work, reflecting developments in economics itself (Williamson 1985), is that institutional arrangements exercise powerful effects in structuring behavior. Since legal actors generally function in an institutional context, this has proven to be a productive line of inquiry. Positive political theory, which treats institutions themselves as rational actors, has also been utilized in current legal scholarship to take institutional factors into account (Eskridge and Frickey 1994, Georgetown Law Journal 1992). A second, and related, theme is that internalized norms exercise powerful effects that conflict with the self-interest, or at least the immediate self-interest, of the actor. Some scholars argue that legal rules can be regarded as internalized norms that reflect varying degrees of long-term rational self-interest (University of Pennsylvania Law Review 1996); others assert that particular communities, such as merchants in a given trade (Bernstein 1996), or farmers in a given area (Ellickson 1991), engage in heavily norm-determined behavior that strongly affects the impact of official law on these communities. The combined effect of focusing on institutions and on norms has been to modify, without eliminating, the external perspective that law and economics generally adopts. Law is not being treated as presumptively coherent, but its internal structure is being given a role that followers of the Chicago School tend to deny.

Still another influence on law and economics has been the attack on human rationality by cognitive psychologists (Kahneman et al. 1982). This work, which is heavily empirical, has revealed that people, at least in laboratory settings, are irrationally optimistic, irrationally attached to the status quo, or irrationally influenced by the way that choices are presented. Legal scholars have applied these results to a variety of settings, such as jury decisions, contracting behavior, and policy making, and have also begun conducting their own empirical research to study behavior that is relevant to specific legal situations (Jolls et al. 1998, Vanderbilt Law Review 1998). Thus, the effort to develop an account of human behavior that is more realistic and more modulated that the rational actor model of law and economics has lead to the extensive incorporation of a second discipline, psychology, into legal scholarship.

3.2 *Outsider Scholarship*

Critical legal studies has been succeeded by a series of movements that tend to refer to themselves as Outsider Scholarship. They include critical race theory, feminist theory, and gay legal studies (see *Critical Race Theory*; *Feminist Legal Theory*; *Sexual Orientation and the Law*). While it cannot be said that critical legal studies itself is entirely moribund, its role within the academy

has been largely replaced by these latter movements, which share its critical approach but contest its particular conclusions. Outsider scholarship, like critical legal studies, asserts that law is an instrumentality of hierarchy and social domination, but it rejects economic class analysis as an explanation for this phenomenon. Instead, it argues that more specific prejudices are involved, namely, the prejudice of the racial majority against minorities like African-Americans, Latinos, and Native Americans, the prejudice of men against women, and the prejudices of heterosexuals against gays and lesbians (Bell 1987, MacKinnon 1989, Eskridge 1994).

One result of this approach is to create a critical scholarship with a more direct connection to American politics; while class has never been a salient political issue in the United States, race, gender, and sexual orientation are all subjects of active political movements. Another, and not unrelated result, has been to decrease the breadth of the critical attack on law. Critical legal studies tends to regard law in its entirety as an oppressive device that hides its true effect behind a facade of objectivity. Outsider scholarship, because it concerns social phenomena that are less comprehensive than class, views law as oppressive because it has denied outsiders the benefits it grants to the dominant group. Its prescription, therefore, is not to abandon or even transform law, but to provide law's benefits to the outsiders, fulfilling the promise that it has thus far denied.

The effect of this outsider scholarship is thus to modify the external perspective of its predecessor. Law is still viewed with suspicion, but its role as a system with internal values is affirmed. For example, instead of arguing that rights are illusory, or an impediment to genuine equality because they can only be validated through litigation, outsider scholarship tends to argue that the disadvantaged cannot afford to wait for such comprehensive transformations, but should use the concept of rights to achieve a more limited but more accessible equality within the existing system (Williams 1991).

Outsider scholarship shares the critical legal studies view that legal concepts are social constructions that serve specific purposes, often those of dominant elites, and that these concepts are radically indeterminate. But it has not relied as heavily on the indeterminacy argument. As many scholars have pointed out, the argument is itself too indeterminate, since it lacks the criteria to judge how determinate the law could be, and how much determinacy is desirable (Kress 1989). In addition, most legal scholars recognize that the indeterminacy argument does not necessarily lead to the political conclusions that critical legal studies advanced. Thus, outsider scholarship analyzes legal rules in fairly traditional terms. Its methodological innovation has been its extensive use of narrative, or story telling, as a means of communicating the personal experience of oppression that racial minorities, women, or gays and lesbians have experienced (e.g., Bell 1987, Eskridge 1994). While this has a literary quality (when done well, at least), it is not related to literary criticism or to the methodology of any other discipline. Rather, it is a technique that is intended to appeal to the values of the reader within the framework of legal scholarship and the legal system.

3.3 Standard Legal Scholarship

Throughout the period when law and economics and critical legal studies were articulating their critique of law and of the legal process school, standard legal scholarship continued to be written. At that time, and at present, it constituted the largest proportion of the legal scholarship being produced. It is standard because of its quantity, but also because it is not explicitly identified with any particular school of thought or methodology. In fact, this scholarship can be loosely regarded as a continuation of the legal process school. Like legal process, it acknowledges that law is subordinate to politics, but treats legal rules as a body of doctrine that has enough coherence to generate analytic standards. In addition, it shares the legal process view that the participants in the legal system, particularly judges, but also legislators and administrators, are public-oriented enough so that scholarly prescriptions for improving their decisions can be addressed to them (e.g., Choper 1980, Dworkin 1986, Ely 1980). The standard scholarship that is currently being produced differs from its predecessors because it has at least partially absorbed the lessons of both law and economics and critical legal studies, although without accepting their critique of law's legitimacy (Rubin 1996).

Most legal scholars now acknowledge that viewing legal actors as public-oriented is problematic. They recognize that these actors are sometimes motivated by their material self-interest, and sometimes intent on preserving existing hierarchies and class relations. Thus, they do not see the legal system as a realm of reasoned, public-oriented decision making that transcends politics, as legal formalism did, nor as a realm of reasoned, public-oriented decision making that is circumscribed by politics but insulated from it, as legal process did. Rather, politics both sets the limits of the legal system and intercalates itself into particular legal decisions. But standard scholarship rejects the notion that decision makers are motivated by any uniform value system, whether it is self-interest, prejudice or public orientation; the prevailing attitude seems to be that some degree of conscientiousness can be assumed, but that the actual motivations must be sorted out in particular situations.

Methodologically, standard scholarship has assimilated the idea that law is not a quasi-autonomous system that provides its own analytic tools. Scholars

regularly rely on informal law and economics or public choice insights in analyzing the behavior of public officials, and, to a lesser extent, on the insights of critical theory. More importantly, these scholars regularly incorporate data from other fields, such as political science, sociology, economics, and psychology, in framing recommendations. The criterion for stating that a particular legal decision, whether by a judge, legislator, or administrator, is incorrect, or could be improved, is less likely to be the internal coherence of legal doctrine, and more likely to be some judgment about social welfare. Such judgments will frequently involve empirical information generated by another discipline and will almost always acknowledge, implicitly if not explicitly, that such information is relevant to the conclusion.

3.4 Contemporary Sociolegal Studies and Legal Knowledge

These developments in law and economics, outsider scholarship, and standard legal scholarship represent a process, extending over the entire course of the twentieth century, by which the distinctions between internal and external scholarship, and between jurisprudence and doctrinal studies, have been partially effaced. This process has important implications for the prevailing conception of legal knowledge, implications which can be illustrated by considering a fourth aspect of contemporary scholarship, namely, sociolegal studies.

The sociology of law, as developed by nineteenth-century scholars such as Maine and Savigny, was originally a jurisprudential position, articulated in opposition to the universalist claims of both natural law and the Napoleonic Code. Its basic insight was that law is the product of particular cultures. With the growth of academic sociology and anthropology, this became translated into detailed studies of legal behavior that were more specific, but also more external to the legal system that they studied. Rather than focusing on the internal coherence of the system, or the experience of its participants, sociolegal work treated law as a phenomenon to be determined by examining the behavior of relevant actors. Legal scholars typically ignored such studies because they regarded law as a self-contained, meaningful system that generated its own criteria of analysis.

In response to the attack on the entire concept of law by law and economics and critical legal studies, most legal scholars now recognize that law possesses only a provisional coherence. Its value is no longer determined by its internal logic, but by its ability to achieve political and social goals. For most scholars, however, law is not simply a set of unrelated provisions that achieve these goals, but a system with both institutional and conceptual features. To have legal knowledge, one must be familiar with the features of this system; to use this knowledge, one must make predictions, or prescriptions, about the way this system achieves independently determined goals. Thus, data from fields such as economics, sociology, political science, anthropology, and psychology becomes directly relevant to legal scholarship. Some of this data comes from work that is not specifically concerned with law, such as organization theory or cognitive psychology, but a good deal comes from modern sociolegal studies, where law is the social scientist's explicit concern. But this work, however useful, does not displace legal scholarship. The legal scholar still has a crucial role in determining how such data will be used by the institutional and conceptual elements of the legal system. In fact, legal scholars have recently begun to re-emphasize these conceptual elements, arguing that, within a richly described social context, law possesses a wide range of symbolic and constitutive roles (Habermas 1996, Lessig 1995, Luhmann 1995) (see *Law, Autopoiesis in*; *Law as Constitutive*). Sociolegal scholars have come to recognize this as an empirical fact, and have thus taken advantage of legal scholarship's greater receptivity to social science data by engaging more fully with the internal features of the legal system. The result is a partial effacement of the distinction between external and internal scholarship, without its complete elimination.

The challenge to the concept of law, the multiplicity of responses, and the greater methodological range of the resulting scholarship has also partially effaced the distinction between jurisprudence and doctrinal studies. Basic questions about the nature and purpose of law, and the nature of legal knowledge, must now be resolved by scholars engaged in detailed doctrinal analysis. Such questions can be ignored, of course, in the interest of simplicity, but the work will be open to challenge by those who question its basic premises. Again, the distinction has not been entirely eliminated; as a practical matter, many scholars content themselves with an analysis of doctrine within one or another subdisciplines of legal scholarship, and let others fight the more theoretical battles on their behalf. But most recognize that the status of their work as legal knowledge depends on theoretical arguments in a more immediate way than was true a century before.

These developments can be illustrated by considering some of the principal topics of modern sociolegal studies. One notable social science approach to judicial decision making in the early postwar era were the judicial attitude studies that treated court decisions as a pure reflection of the judge's ideology, uninfluenced by legal doctrine (Schubert 1959, Ulmer 1960). While work of this sort continues to be produced, often by the same authors, more recent social science studies of judicial decisions tend to emphasize the interplay between ideology and doctrine (Perry 1991, Shapiro 1981), a process which enables legal scholars to utilize

their findings and extend them into jurisprudential areas (Feeley and Rubin 1999, Kennedy 1997, Tamanaha 1997). A second distinctive strand during the early postwar period emphasized the discontinuity between the law on the books and the law in action. Modern sociolegal studies often recharacterize this issue in terms of the implementation process, which involves an analysis of the interplay between statutory or regulatory language and the realities of administrative enforcement (Bardach and Kagan 1982, Coglianese 1997). Once again, such an approach yields jurisprudential implications more readily than its predecessor, since it replaces an inevitable abyss between written and practiced law with a dynamic interchange where the practice suggests a reconsideration of the legal structure (Ayres and Braithwaite 1992, Kagan 2000).

A related development is the sociolegal analysis of contracts. The seminal work in the field emphasized that contract law, as reflected in the decided cases that were the exclusive focus of most law school courses, did not exercise the dominant effect on contractual practices; rather, the relationship between the parties and their underlying cultural understandings were the primary determinants (Macaulay 1963, MacNeil 1980). Building on this insight, and incorporating the lessons of contemporary law and economics, legal scholars have been able to analyze the practical operation and effect of specific contractual terms (Gilson 1984, Hadfield 1990). In effect, modern sociolegal scholars are now engaging with the specific contractual language, just as they are engaging with the internal aspects of legal doctrine and statutory provisions.

See also: Critical Legal Studies; Law and Society: Sociolegal Studies; Law: History of its Relation to the Social Sciences; Law, Sociology of; Legal Education; Legal Formalism; Legal Pluralism; Legal Positivism; Legal Realism

Bibliography

Ayres I, Braithwaite J 1992 *Responsive Regulation*. Oxford University Press, New York

Bardach E, Kagan R 1982 *Going by the Book*. Temple University Press, Philadelphia, PA

Bell D 1987 *And We Are Not Saved: The Elusive Quest for Racial Justice*. Basic Books, New York

Bernstein L 1996 Merchant law in a merchant court: Rethinking the code's search for immanent business norms. *University of Pennsylvania Law Review* **144**: 1765

Bickel A 1962 *The Least Dangerous Branch*. Bobbs-Merrill, Indianapolis, IN

Choper J 1980 *Judicial Review and the National Political Process*. University of Chicago Press, Chicago

Coglianese C 1997 Assessing consensus: The promise and performance of negotiated rulemaking. *Duke Law Journal* **46**: 255–1349

Dworkin R 1986 *Law's Empire*. Belknap Press, Cambridge, MA

Easterbrook F 1983 Statute's domains. *University of Chicago Law Review* **50**: 533

Edelman L B Legality and the endogeneity of law. In: Kagan R A, Krygier M, Winston K (eds.) *Legality and Community: On the Intellectual Legacy of Philip Selznick*. Institute of Governmental Studies Press/Rowman & Littlefield, Berkeley, CA

Ehrlich E 1975 *The Fundamental Principles of the Sociology of Law*. Russell & Russell, New York

Ellickson R 1991 *Order Without Law*. Harvard University Press, Cambridge, MA

Ely J 1980 *Democracy and Distrust*. Harvard University Press, Cambridge, MA

Eskridge W 1994 Gaylegal narratives. *Stanford Law Review* **46**: 607

Eskridge W, Frickey P 1994 The Supreme Court, 1993 term: Foreword: Law as equilibrium. *Harvard Law Review* **108**: 26

Feeley M, Rubin E 1999 *Judicial Policy Making and the Modern State: How the Courts Reformed America's Prisons*. Cambridge University Press, New York

Finnis J 1980 *Natural Law and Natural Rights*. Clarendon Press, Oxford, UK

Fuller L 1964 *The Morality of Law*. Yale University Press, New Haven, CT

Fuller L 1978 Forms and limits of adjudication. *Harvard Law Review* **92**: 353

Georgetown Law Journal 1992 Symposium: Positive political theory and public law. *Georgetown Law Journal* **80**: 457

Gilson R 1984 Value creation by business lawyers: Legal skills and asset pricing. *Yale Law Journal* **94**: 239

Hadfield G 1990 Problematic relations: Franchising and the law of incomplete contracts. *Stanford Law Review* **42**: 927

Habermas J 1996 *Between Facts and Norms* (trans. Rehg W). MIT Press, Cambridge, MA

Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK

Horwitz M 1977 *The Transformation of American Law*. Harvard University Press, Cambridge, MA

Jolls C, Sunstein C, Thaler R 1998 A behavioral approach to law and economics. *Stanford Law Review* **50**: 1471

Kagan R 2000 The consequences of adversarial legalism. In: Kagan R, Axelrad L (eds.) *Regulatory Encounters*, pp. 372–413

Kahneman D et al. 1982 *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York

Kelsen H 1967 *Pure Theory of Law* (trans. Knight M). University of California Press, Berkeley, CA

Kennedy D 1997 *A Critique of Adjudication: Fin de Siecle*. Harvard University Press, Cambridge, MA

Kress K 1989 Legal indeterminacy. *California Law Review* **77**: 283–337

Lessig L 1995 The regulation of social meaning. *University of Chicago Law Review* **62**: 943

Luhmann N 1995 *Social Systems* (trans. Bednarz J). Stanford University Press, Stanford, CA

Macey J 1986 Promoting public-regarding legislation through statutory interpretation: An interest group model. *Columbia Law Review* **86**: 223

Macaulay S 1963 Non-contractual relations in business: A preliminary study. *American Sociological Review* **28**: 55

MacCormick N, Summers R (eds.) 1991 *Interpreting Statutes*. Dartmouth, Aldershot, UK

MacKinnon C 1989 *Toward a Feminist Theory of the State*. Harvard University Press, Cambridge, MA
MacNeil I 1980 *The New Social Contract*. Yale University Press, New Haven, CT
Olson M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
Posner R 1972 *Economic Analysis of Law*. Little, Brown, Boston
Priest G, Klein B 1984 The selection of disputes for litigation. *Journal of Legal Studies* **13**: 1
Rubin E 1996 The new legal process, the synthesis of discourse, and the microanalysis of institutions. *Harvard Law Review* **109**: 1393–438
Schauer F 1988 Formalism. *Yale Law Journal* **97**: 501–48
Schlegel J H 1995 *American Legal Realism and Empirical Social Science*. North Carolina Press, Chapel Hill, NC
Schubert G 1959 *Quantitative Analysis of Judicial Behavior*. Free Press, Glencoe, IL
Shapiro M 1981 *Courts: A Comparative and Political Analysis*. Chicago University Press, Chicago
Simon H 1957 *Administrative Behavior*.
Tamanaha B 1997 *Realistic Socio-legal Theory*. Clarendon Press, Oxford, UK
Ulmer S 1960 The analysis of behavior patterns in the Supreme Court of the United States. *Journal of Policy* **22**: 429
Unger R 1975 *Knowledge and Politics*. Free Press, New York
University of Pennsylvania Law Review 1996 Symposium: Law, economics and norms. *University of Pennsylvania Law Review* 144–1643
Vanderbilt Law Review 1998 The legal implications of psychology: Human behavior, behavioral economics and the law. *Vanderbilt Law Review* **51**: 1495–788
Wechsler H 1959 Toward neutral principles of constitutional law. *Harvard Law Review* **73**: 1
Williams P 1991 *The Alchemy of Race and Rights*. Harvard University Press, Cambridge, MA
Williamson O 1985 *The Economic Institutions of Capitalism*. Free Press, New York

E. Rubin

Legal Systems, Classification of

'Law' and 'legal systems' as concepts are closely related. Since there have been innumerable definitions of law, ideas about legal systems have been similarly diverse. A system involves regular interactions among elements that together make up an entity with boundaries. Thus, lawyers, judges, legislators, administrators, the police, and legal scholars all work with rules in regularized ways that involve cultural expectations about their roles and the legal institutions with which they interact. In this view a legal system is greater than the rules themselves (Merryman et al. 1994). But a legal system may also refer to only the rules of a tribe, city, nation, or the international order, or to the natural rules for humankind itself. The question of classification then consists of looking for similarities and differences among the relevant legal systems.

1. *The History of Legal System Classifications*

The task of classifying legal systems has been tied to the development of comparative law as a practical and scholarly discipline. More recently, the differentiation of the behavioral and social sciences into discrete subjects has added diverse perspectives on the nature of law and legal systems.

1.1 *Classical Classifications*

Classical Greeks made the earliest recorded efforts explicitly classifying legal systems as a result of their philosophical speculation based on the comparative study of law. For instance, Plato (429–348 BC) in his *Laws* discussed the rules of several Greek and other *poleis* in formulating his ideal code and legal institutions for Magnesia. Aristotle (384–322 BC) examined many legal structures in his *Politics* before settling on the three preferred categories of kingship, aristocracy, and constitutional government (*politeia*) with their deviant siblings tyranny, oligarchy, and democracy.

In Rome, jurists' interest in foreign law was more practical, particularly with the creation of the office of *praetor peregrinus* in 242 BC. This praetor supervised jurisdiction over disputes involving non-Romans. It was here that the mixture of a less formalistic Roman *ius civile*, Greek law, and other foreign legal norms led to an internationalized *ius gentium*. Roman jurists did not have a philosophical interest in studying and classifying foreign legal systems, since they believed in the superiority of their own law. An exception that makes the point was the fourth-century *Collatio legum Mosaicarum et Romanarum*, in which excerpts from Roman jurists were paired with the laws of Moses to improve the stature of Christianity (Zweigert and Kötz 1996, p. 48).

1.2 *Early Modern Classifications*

In the early modern era some natural law jurists such as Hugo Grotius (1583–1645), Samuel Pufendorf (1632–1694), and Baron Charles Louis de Montesquieu (1689–1755) used the comparative law method to provide empirical backing for natural law principles. Others such as William Blackstone (1723–1780) had the additional objective of providing prestige for their national law: Blackstone's four-volume *Commentaries on the Laws of England* (1765–1769) (Blackstone 2001) emphasized the similarities in English common law, Roman law, and natural law.

Montesquieu's extensive speculation about the relation of law to natural and social forces and explicit classification of legal systems emphasized the distinct

manners (*moeurs*) of a particular people in a particular era. His *Lettres Persanes* (1721) (Montesquieu 1964), published anonymously, used the device of two Persians traveling in Europe who satirized the customs of French and European society. They pointed to the abuses and cruelties of criminal justice systems and argued for a rational theory of punishment adjusted to the settled sensibilities of the people. Letter 80 could have been written to critique the turn-of-the-millennium criminal sentencing and prison policy of the USA:

> [O]bedience to the laws of a state does not correspond with the greater or lesser degrees of cruelty in punishment. In countries where penalties are moderate, they are as much feared as those in which they are tyrannical and dreadful … . Our imagination adapts itself to the customs of the country in which we live (Montesquieu 1964, p. 136).

Montesquieu's relativist view of humans' ability to achieve justice was further developed in *De l'esprit des lois* (1748) (Montesquieu 1989), which drew upon causative factors such as climate, soil, population size, a people's morals and customs, religion, and commerce. He classified governments along with the appropriate principle for their constitutive societies. Thus democracy required civic virtue; aristocracy, moderation against the people; monarchy, honor for the law and intermediate institutions; and despotism, fear to maintain order. He found different mechanisms for lawmaking and adjudication in the first three types of legal system, but pointed out that law was not necessary for despotism.

1.3 Nineteenth- and Twentieth-century Classifications

This set the stage for much more elaborate classifications of local and national legal systems beginning with nineteenth-century evolutionary theorists. Law in these schemes was part of the larger social system that changed over time. For instance, Sir Henry Maine (1822–1888), interested in the striking parallels between the development of English common law and classical Roman law, posited in *Ancient Law* (1861) (Maine 1977) stages of social development common to different peoples that could be correlated with particular instruments of legal growth. Legal systems could be characterized by whether social change was primarily dealt with by legal fictions, equity or natural law, or conscious creative legislation.

Max Weber (1864–1920) (Weber 1968) used many of the same historical sources as legal evolutionists, but in *Wirtschaft und Gesellschaft* (1922) constructed ideal types to facilitate the comparison of actual social and legal systems in trying to explain the rise of industrial capitalism. His four-cell typology for lawmaking and law application considered the rationality or irrationality of legal thought, as well as whether the legal norms and decision making were highly differentiated from religion, ideology, or emotion. He argued that the category of formal (autonomous) rationality with a consistent body of general legal norms provided the predictability facilitating modern capitalism.

In an effort to make US lawyers and political scientists less parochial, John Henry Wigmore in 1928 published the three volume *A Panorama of the World's Legal Systems* (Wigmore 1928). He covered 16 historic and contemporary legal systems by the pictorial method, using between 20 and 50 pictures accompanied by text for each legal system to enliven its justice buildings, principal legal actors, and characteristic legal materials. Organized historically, these systems included the Egyptian, Mesopotamian, Chinese, Hindu, Hebrew, Greek, maritime, Roman, Celtic, Germanic, canon, Japanese, Mohammedan (Islam), Slavic, Romanesque (civil law), and Anglican.

2. Legal System Classifications by Disciplines Related to Law

Drawing on these earlier studies and with a common concern for the relationship between law and society, anthropologists, sociologists, and political scientists generally believe that formal legal rules should be deemphasized when classifying legal systems.

2.1 Legal Anthropology

Anthropologists have also observed that some societies operate without government but nevertheless have social order and law-like institutions. In modern nations there may be distinct societies, so that it becomes appropriate to speak of legal pluralism. Classification of legal systems then depends on the characteristics of different societies or their subgroups. Leopold Pospíšil (1974, pp. 106–26), for example, described this situation as a patterned mosaic of subgroups that belong to certain identifiable types with different memberships. Each subgroup owes its existence to a legal system that regulates the behavior of its members. These legal systems then form a hierarchy reflecting the degree of inclusiveness of the corresponding subgroups, and can be categorized by legal level such as the family, lineage, community, state, or nation. People are simultaneously members of several subgroups of different inclusiveness, which may lead to contradictory legal loyalties.

2.2 Legal Sociology

Lawrence Friedman (1977, pp. 78–86), a legal sociologist, argued that legal systems can be classified

according to their internal legal culture, which includes the attitudes and values of judges, lawyers, and other legal actors toward a task such as lawmaking or adjudication that relies on reasoning or some other basis of legitimacy. He considered two variables. First, when a legal system requires reasoning to be premised on legal propositions it is a closed system; when propositions may be nonlegal the system is open. Second, some systems regard legal propositions as fixed and timeless; others accept change as normal.

Using these two variables, Friedman derived four types of reasoning associated with law. First, some systems such as those based on sacred law are closed and reject the idea of change. What change does occur must be based on legal fictions. Second, a legal science system such as that in Germany is closed, but it permits change by simply adding legal rules or deriving new legal principles. Third, traditional or customary law systems do not accept explicit change but are open to new norms based on shifting behavioral patterns. Fourth, a system permitting change by looking outside the stock of legal propositions includes what Weber called substantive rationality. This could be revolutionary legality such as that in socialist Russia, or welfare legality that looks to ethical imperatives or utilitarian principles.

2.3 Political Science

The 1950s comparative politics revolution shifted that field from one that tended to emphasize public law and to describe the legal institutions of foreign governments to the explanation of political behavior in structural and functional terms. Gabriel Almond and Bingham Powell (1978), for instance, used the developmental variables of cultural secularization and structural differentiation, together with the low or high level of political subsystem autonomy associated with interest groups and political parties, to classify political systems. Modern systems could be premobilized, mobilizing, or penetrative, and within each of these categories either authoritarian or democratic.

3. Current Classifications in Comparative Law

A useful classification of legal systems should make the criteria of division clear and allow scholars to assimilate, simplify, and comprehend the great wealth of legal detail existing in individual systems. The dominant approach to classifying national legal systems in contemporary comparative law is to identify families, styles, or traditions. Beyond the aim of simplification, this idea could permit one to understand more clearly pluralistic legal systems with an official, dominant legal tradition, but including indigenous communities with their own law or communities lacking a sovereign political organization that use canon law, Muslim law, Hindu law, or Jewish law.

3.1 The Influence of René David

Comparative lawyers have disagreed about the criteria to use for classification but the results have been surprisingly similar, differing primarily in the number of categories emerging. René David, in *Traité élémentaire de droit civil comparé* (1950), used two criteria that were further refined in his 1964 *Les grands systèmes de droit contemporains* (David and Brierley 1985). He believed that the primary criterion should be the legal system's ideology, its 'philosophical, political or economic principles,' the 'social objectives to be achieved,' and 'the place of law itself within the social order.' A secondary criterion should be a system's legal technique, 'the law's conceptual structure or … the theory of sources of the law.' On the basis of these guides, David found three prominent legal families: Romano-Germanic, common law, and socialist law. Some elements from one or more of these families have influenced all national legal systems. In addition, there are lesser legal systems, such as those for religious law (Muslim and Hindu) and for East Asia and sub-Saharan Africa (David and Brierley 1985, pp. 20–31).

This typology controlled the organization of the *International Encyclopedia of Comparative Law*'s volume two (1984), entitled 'The Legal Systems of the World: Their Comparison and Unification,' of which René David was the chief editor (David 1984). The only variation from David's earlier formulation occurred in Chap. 1, where Romano-Germanic (civil) law and common law were combined into the Western conception of law (see Clark 1990).

3.2 Legal Traditions

David argued that it

> is a superficial and indeed false view to see law as being composed of the totality of [legal] rules … [T]he juridical phenomenon which they represent … constitutes a *system*: it has a vocabulary used to express concepts, its rules are arranged into categories, it has techniques for expressing rules and interpreting them, it is linked to a view of the social order itself which determines the way in which the law is applied and shapes the very function of law in that society (David and Brierley 1985, p. 19).

John Henry Merryman in 1969 proposed a similar idea of legal traditions, of which three were most influential in the modern world: civil law, common law, and socialist law:

> A legal tradition … is not a set of rules … . Rather it is a set of deeply rooted, historically conditioned attitudes about the nature of law, about the role of law in the society and the polity, about the proper organization and operation of a legal system, and about the way law is or should be made, applied, studied, perfected, and taught. The legal tradition relates the legal system to the culture of which it is a partial expression (Merryman 1985, pp. 1–2).

In teaching and research on comparative law in the USA, scholars have invoked considerations of geography, history, and legal culture as secondary criteria to legal traditions. This enlarges on the concentration on German and French examples by adding attention to Mediterranean European, Latin American, and East Asian legal systems (Merryman et al. 1994). Even at the end of the twentieth century David's tripartite division of families, including socialist law and other categories for East Asia and sub-Saharan Africa, still seemed to hold substantial validity for some research topics, such as the role and importance of the legal profession in society (Clark 1999).

3.3 Legal Styles

About the time of David's revised formulation of legal families Konrad Zweigert (1961) wrote about his theory of *Rechtskreisen*. It looked at the styles (*Stile*) of legal systems to divide them into legal spheres or families. He critiqued earlier attempts at classification as too one-dimensional, turning primarily on one criterion. Zweigert, as updated by Hein Kötz, found five factors to be crucial to classify the style of a legal system or family: (a) historical background and development; (b) predominant mode of legal thought; (c) distinctive legal doctrine or institutions; (d) sources of law; and (e) ideology. Using these factors, Zweigert and Kötz settled on eight categories: Romanistic family, Germanic family, Nordic family, common law family, Peoples Republic of China (replacing the earlier socialist family), Japan, Islamic law, and Hindu law (Zweigert and Kötz 1996, pp. 62–73).

3.4 Critique of Current Classifications

There are a number of problems with these classification schemes, which their proponents often recognize. First, a classification is intended for a specific purpose, which may make it unsuitable for another purpose. Thus Zweigert and Kötz intended their divisions to serve as chapters in their comparative law treatise for students and jurists. An African legal family was omitted because it is still in the process of emerging (1996, pp. 65–6).

In addition, almost all comparative law classifications are based on a private law emphasis, which weakens their utility for public law comparisons. For instance, a study of the judicial review of legislation as an aspect of constitutionalism could put England and France in one category, but Germany and the USA in another, confounding the civil law–common law dichotomy. Mirjan Damaška (1986) rejected the usual classification of legal families in his comparative study of procedure and the administration of justice. He formed a Weberian four-cell model that considered the structure of state authority on one axis (hierarchy or coordinate) and the type of proceeding on the other (conflict solving or policy implementing). Illustrations include the active judge in continental European civil procedure in the conflict-solving hierarchy cell, English and USA's civil and criminal procedure in the conflict-solving coordinate cell, Soviet civil and criminal procedure in the policy-implementing hierarchy cell, and USA's public interest litigation in the policy-implementing coordinate cell.

Second, no matter how logical a classification, there will always be hybrid legal systems that do not fit easily within one category or another. Well-known hybrids in legal family classifications include, for example, Scotland, Louisiana, Puerto Rico, Québec, Israel, South Africa, Sri Lanka, and the Philippines that share civil law and common law elements; the People's Republic of China with East Asian and socialist elements; India with Hindu and Islamic law for family and inheritance matters but common law for obligations and commercial issues; and Arabian countries with Islamic law for family and inheritance issues but French-influenced law for the commercial sector (Schlesinger et al. 1998, pp. 291–304, Zweigert and Kötz 1996, pp. 64, 72–3).

Third, related to the problem of hybrid systems is the situation of legal pluralism, associated with the personality principle for choice of law, where an official, dominant legal tradition coexists with an indigenous or religious community that de facto regulates itself according to its own norms (Merryman et al. 1994, pp. 656–703).

Fourth, Ugo Mattei (1997) suggested a new tripartite classification to reduce the Eurocentricism of earlier approaches. His families depend on the dominant pattern of legal decision-making in a system, which can be characterized as professional law, political law, or traditional law. In the professional law family (including west European civil law nations and the common law United Kingdom, Oceania, United States, and Canada), the idea of law autonomous from politics and religion predominates. The political law family includes ex-socialist and developing countries (eastern Europe, sub-Saharan Africa, and Latin America) in which the notion of autonomous law is not strong enough to prevent politics from supplanting legal rules for important legal decisions. Finally, the traditional law family consists of those legal systems that may have absorbed substantial elements from the civil law or common law (Asia and northern Africa), but in which the dominant culture controlling legal

decision-making is rooted in Confucianism, Hinduism, or Islam.

4. Convergence or Divergence of Legal Traditions

John Merryman et al. (1994, pp. 16–25) identified three strategies of convergence that could bring the civil law and common law traditions closer together. First, programs for the unification of law are gaining support. These include the Hague Conference on Private International Law, the International Institute for the Unification of Private Law (UNIDROIT), and the United Nations Commission on International Trade Law (UNCITRAL). Within Europe, working groups of jurists from both common law and civil law countries are drafting rules with the hope to return to a kind of Roman-canonic *jus commune*.

A second strategy is legal transplantation, which can follow military conquest or occur by conscious choice. Developing countries often select legal rules and institutions as evidence of modernization. Japan in the late nineteenth century provides an interesting example of explicit legal Europeanization. Countries interested in law reform may also look abroad for inspiration.

Third, the tendency for countries with similar social and economic conditions to develop similar legal systems is part of what may be called natural convergence. This similarity is encouraged by greater international communication, travel, and business. In response to the notion that focusing on convergence of the civil law and common law is too Eurocentric, one could respond that developing nations and East Asian countries have themselves been consciously involved in a process of convergence by supporting rules and institutions associated with either of these two worldwide legal traditions.

Although the beginning of the twenty-first century seems to be characterized by an overarching process of globalization in political, economic, and social life, there are in fact two opposing forces in international society. One is the universalizing force expressed in the increased importance of the United Nations, the World Trade Organization, and regional supranational legal systems such as the European Union and NAFTA.

Besides universalizing pressures there are particularizing forces within some nations that reflect elements of diversity such as religion, ethnicity, language, and social class. Everywhere there are movements for political decentralization, economic privatization, and social recognition that seem to complicate the legal universe. Examples include the Scottish and Welsh in the UK, the Basques and Catalans in Spain, the Kosovars in Yugoslavia, and the Québecois in Canada. Thus while it is true that the principal legal traditions are converging, they are also at the same time on some levels diverging.

See also: Chinese Law; Civil Law; Common Law; European Union Law; Hindu Law; Jewish Law; Law in Latin America; Socialist Law

Bibliography

Almond G A, Powell G B, Jr 1978 *Comparative Politics: System, Process, and Policy*, 2nd edn. Little, Brown, Boston, MA

Blackstone W 2001 In: Morrison W J (ed.) *Blackstone's Commentaries on the Laws of England*. Cavendish, London

Clark D S 1990 The idea of the civil law tradition. In: Clark D S (ed.) *Comparative and Private International Law: Essays in Honor of John Henry Merryman on His Seventieth Birthday 11–23*. Ducker and Humblot, Berlin

Clark D S 1999 Comparing the work and organization of lawyers worldwide: The persistence of legal traditions. In: Barceló J J III, Cramton R C (eds.) *Lawyers' Practice and Ideals: A Comparative View* 9–155. Kluwer Law International, The Hague

Damaška M R 1986 *The Faces of Justice and State Authority: A Comparative Approach to the Legal Process*. Yale University Press, New Haven, CT

David R (ed.) 1984 *The Legal Systems of the World: Their Comparison and Unification*. J C B Mohr (Paul Siebeck), Tübingen, Germany, Vol. 2

David R, Brierley J E C 1985 *Major Legal Systems in the World Today: An Introduction to the Comparative Study of Law*, 3rd edn. Stevens and Sons, London

Friedman L M 1977 *Law and Society: An Introduction*. Prentice-Hall, Englewood Cliffs, NJ

Glendon M A, Gordon M W, Carozza P G 1999 *Comparative Legal Traditions in a Nutshell*, 2nd edn. West Group, St. Paul, MN

Maine H S 1997 *Ancient Law*. Dent, London

Mattei U 1997 Three patterns of law: Taxonomy and change in the world's legal systems. *American Journal of Comparative Law* **45**: 5–44

Merryman J H 1985 *The Civil Law Tradition: An Introduction to the Legal Systems of Western Europe and Latin America*, 2nd edn. Stanford University Press, Stanford, CA

Merryman J H, Clark D S, Friedman L M 1979 *Law and Social Change in Mediterranean Europe and Latin America: A Handbook of Legal and Social Indicators for Comparative Study*. Oceana Publications, Dobbs Ferry, NY

Merryman J H, Clark D S, Haley J O 1994 *The Civil Law Tradition: Europe, Latin America, and East Asia*. Michie, Co, Charlottesville, VA

Montesquieu C 1964 *The Persian Letters* Healy, trans. G R. Bobbs-Merrill, Indianapolis, IN

Montesquieu C 1989 *The Spirit of the Laws* [trans. Cohler A M et al.]. Cambridge University Press, Cambridge, UK

Pospíšil L 1974 *Anthropology of Law: A Comparative Theory*. HRAF Press, New Haven, CT

Schlesinger R B, Baade H W, Herzog P E, Wise E M 1998 *Comparative Law: Cases, Text, Materials*. Foundation Press, New York

Weber A 1968 In: Roth G, Wittich C (eds.) *Economy and Society: An Outline of Interpretive Sociology* [trans. Fischoffe et al.]. Bedminster Press, New York, Vols. 1–2

Wigmore J H 1928 *A Panorama of the World's Legal Systems*. West Publishing, St. Paul, MN, Vols. 1–3

Zweigert K 1961 Zur Lehre von den Rechtskreisen. In: Nadelmann K N, von Mehren A T, Hazard J N (eds.) *XXth Century Comparative and Conflicts Law: Legal Essays in Honor of Hessel E. Yntema*. Sythoff, Leyden, pp. 42–55

Zweigert K, Kötz H 1996 *Einführung in die Rechtsvergleichung auf dem Gebiete des Privatrechts*, 3rd edn. Mohr J C B (Siebeck P), Tübingen [1992 *Introduction to Comparative Law*, Weir T (trans). 2nd edn. Clarendon Press, Oxford, UK]

D. S. Clark

Legal Systems: Private

Private legal systems are nonstate institutionalized means for obtaining compliance with the norms or rules of a social order. They comprise the numerous forms of social control that are part of group or organizational life, but that are not formally part of state law. They include, for example, the practices of disciplinary bodies, boards and councils of industrial organizations, the tribunals of professional or trade associations and labor unions, and the disciplinary committees of universities and colleges. They also include the peer sanctioning of relatively amorphous voluntary associations, such as self-help and mutual aid groups.

1. *Mechanisms of Control*

Private legal systems bring about conformity to the normative expectations of their collective setting through one of three mechanisms for administering either positive or negative sanctions: (a) coercion, being the threat or use of negative sanctions that reduces a person's social standing; (b) persuasion or ideological control that uses positive inducements to conformity; and (c) suggestion, that reminds people of their obligations, responsibilities and duties in relations with one another.

In some private forms, such as the disciplinary bodies of universities or professional associations, the administration of justice is elaborate, formalized and with considerable history and established precedent, and whose sanctions may involve fines, loss of privileges, or suspension of licensure. In other cases, such as the community or neighborhood association or the local meeting of Alcoholics Anonymous, the administration of justice may be minimal, implicit, or nonexistent, comprising simply of the exercise of sanctions such as ostracism, shaming, gossip, or ridicule. For discussion of related concepts readers should see *African Legal Systems*; *Legal Pluralism*; *Folk/Indigenous/Customary Law*; *Conflict of Laws*.

2. *Significance*

Nonstate private legal systems and their administration of justice are important because more people are directly governed by and controlled through them on a daily basis than they are by state law, which remains relatively distant from most people's day-to-day life. However, while state law and its various agencies and operations are publicly accountable, private systems of justice often escape accountability. While numerous refinements and endless energy are invested in public systems of law to protect the individual's right to due process, to guard against excessive or abusive use of state power, and to minimize arbitrary decisions, rarely are such limitations to potential abuse exercised on the day-to-day working of private legal systems and their forms of justice. Insofar as the ideal of reducing arbitrariness is the hallmark of legality, this is 'a variable achievement' (Selznick 1969).

3. *Private Legal Systems as Types of Law*

Consistent with Weberian analysis, private legal systems are types of law insofar as they are guaranteed by the probability that physical or psychological coercion to bring about conformity to the order or avenge violation is applied by a staff of people holding themselves especially ready for that purpose. Yet, Selznick (1969) argued that if law is seen only in terms of coercive rules, its moral idealism, its progressive striving to reduce arbitrariness in the administration of justice, is overlooked. For Selznick (1969, p. 7) law is 'generic' in that it is 'endemic in all institutions that rely for social control on formal authority and rule making.' This means that the norm of legality is not confined to the state, but occurs in every group that develops rational rules to administer justice. However, while this approach accommodates the concept of private legal systems, it restricts legality to the formal rational, and discounts the informal, which in Weber's terms may be both substantive and irrational, but nonetheless is a form of legal authority over those whom it exercises control.

Building on Gurvitch's (1947) more inclusive analysis, all legal systems, whether rational or irrational, can be considered as existing in two planes. On the horizontal plane is the range of legal systems differentiated by the scale of organizational complexity within which they are constituted and through which they primarily function. This horizontal continuum would range from the global to the group, with a variety of intermediate organizational fields including society, community, and organization. Within any organizational field there are many different types of organization and many types of groups. Moreover, each type of organization may exist in a wider context, but each also has a degree of separation from its

Table 1
Law, private legal systems and levels of formality

	Global	Societal	Communal	Organizational	Group
Type of legal system	International law	Different national state legal systems; Different kinds of state law	Private legal systems; Norms and controls established by communities	Private legal Systems; Rules and controls established by corporations and organizations	Private group norms
Formal (examples)	International law on human rights; Law on genocide	US statutory law; US common law, e.g., criminal law	Neighborhood rules and ordinances affecting housing, noise, pollution	Rules on gross misconduct, cheating, employee theft, absenteeism, property damage	Norms of Alcoholics Anonymous
Informal (examples)	Norms governing smuggling	Police discretion; Plea bargaining	Neighborhood norms on street crime gay rights racial tolerance	Norms of indulgence; Levels of acceptable violation	Norms of employee work team; Norms of street drug gangs

encompassing social form, based on its own semi-autonomous existence in relation to the whole, of which it is a part. Then, for analytical purposes, consider the degree of formality of the legal system in a vertical plane, as ranging from formal to informal. For Gurvitch, the formal is characterized by being organized, written, fixed, and planned in advance, whereas the informal is unorganized, flexible, spontaneous, and intuitive (see Table 1.) Clearly it is not necessary for informal systems of private law to be spontaneous, since they can be established over time and build up considerable institutionalized history as in the case of street gang initiation rituals.

This schema is helpful in that it shows us that informal social control is not the exclusive control mechanism of the group. While developed organizational forms, such as the state in capitalist industrial society, appear to be dominated by formal social control, such as criminal law, within law's organization, and reflective of the groups that make it up, exist numerous levels of informal private justice working within, and often against the explicit principles of its wider system. Examples would include police discretion and plea-bargaining. Conversely, formal social control is not the exclusive mechanism of the state or formal organization. While levels of organizational development often preclude small groups from having highly formal private legal systems, they sometimes develop rudimentary formal controls (Henry 1983). Thus, each type of collectivity has depths of formal and informal operation. For example, an organization, such as a manufacturing company is governed by the regulations of formal state law, but also has its own non-state private legal system for administering internal discipline. Insofar as this relies on written formal rules, procedures, and penalties, it would be considered a type of formal private law. However, within the same company there may exist a range of groups of employees, of different professions, skills or trades, different unions, work teams and social cliques, etc. Each may exercise its own private collective control over its members. Arguably, the most effective system of control is that where the multiple kinds of social control, and each of their levels of operation from formal to informal, complement each other in bringing about conformity. Indeed, rather than using only one type of legal system, a collectivity typically uses a combination of types and of levels of legal system, simultaneously (Henry 1983).

4. The Interrelationship Between Private Legal Systems and State Law

Although private legal systems by definition, exist outside the state, they have a semi-autonomous relationship with state law (Moore 1978), being subject to its broad framework, yet within these constraints being free to create and enforce their own rules. That they exercise control consistent with the broad requirements of the state means that they can reaffirm state law, and so function as state surrogates. As such, private legal systems are a subset of the more generic notion of social control, which refers to broad societal, community, or organizational level processes such as socialization, education, social welfare, etc., that bring

about conformity to societal social order. In this broader sense, private legal systems are a reaffirmation of the relations of order, part of the ongoing construction of a collectivity's social reality. It is in this sense that private legal systems merge with ongoing daily interaction as an integral part of the constitution of group process to develop the appearance of stability, predictability, and order. This point is at the heart of Ehrlich's (1913) observation that 'living law' is that which dominates life itself even though it has not been posited in legal propositions. Yet the semi-autonomy of private legal systems also allows for a contradictory and subversive role in relation to the state. This web of 'legal pluralism' (Pospisil 1971) may exist in harmony or conflict, reinforcing or undermining the social control of other private legal systems and even the formal law of the state, such that they may condone and even approve of behaviors that the overarching state system would condemn. For example, consider the toleration of some level of pilfering (theft) by employees in a food manufacturing company.

Clearly, any consideration of nonstate private legal systems, therefore, must also involve an examination of the theory of 'legal pluralism' and must address the varying interrelationships between the different types of social control and between their varying levels of formality. Further, insofar as private legal systems exist in an ambivalent relationship to state law, sometimes supporting it, at other times contradicting it, and even undermining it, consideration must also be given to how these forms relate to, and even constitute, the totality of control in a society.

4.1 The Limits of Traditional Legal Pluralism

Recognition of the significance of nonstate forms of law has led sociolegal theorists to revise their thinking about the law and society relationship. The suggestion that state law and its institutions for administering justice are but one form of social control has been a core feature of legal pluralism from the foundations of theorizing about law. However, the legal pluralist tradition has focused on the character, source, and hierarchical place of nonstate normative orders in relation to state law. Until the mid-1970s, much of their argument explored the relationship between indigenous tribal custom and European colonial law (Pospisil 1971). This tradition saw indigenous orders as autonomous, independent, yet subordinate to colonial law. By the mid-1980s, the emphasis had shifted to non-colonial societies but retained its concern with the hierarchical power relations between dominant and subordinate law forms. Thus, many critical theorists take 'informal justice institutions' as the subordinate normative order and examine how they serve the ideological function of blurring the power of the state so that it appears to be a benign part of the social fabric. They have shown that this ideological subordination is accomplished by the co-optation and exploitation of a human desire for informal, localized, community justice, and that the episodic tendency toward an 'informal,' decentralized state control serves a dual legitimating and net-widening control function for the state (Cohen 1985). However, these criticisms seem more applicable to the growth of state sponsored dispute settlement institutions, community policing, and restorative justice, than to explaining the state's interrelationship with other, established, private legal systems, which have largely been ignored. During the late-1980s, attention turned to the contribution that the social relations of these nonstate systems of law make in constituting state law.

4.2 Private Legal Systems and Constitutive Legal Theory

Critical contributions to the legal pluralist debate argued that recognition of the interpenetrating role of non-state normative orders requires a new conception of law in society. This has been variously referred to as the 'relational theory of law,' 'constitutive theory,' and 'postmodernist' or 'deconstructionist' theory.

Constitutive theory is based on the idea that law is, in part, social relations and social relations are, in part, law (Moore 1978, Fitzpatrick 1984, Henry and Milovanovic 1996). It is the movement and tension whereby these are socially constituted, 'the way "society" is produced *within* "law"' as David Nelken has described it and how law is produced within society, rather than how they interact, that is crucial to understanding the law–society interface. Thus, instead of assuming that state law is the hub of social control whose spokes radiate as unidirectional pathways of influence to other social and normative orders, constitutive theory also directs our attention to the reverse process: forms and mechanisms whereby legal-relations are penetrated by extra-legal social relations, often constituted as non-state private legal systems. A constitutive approach examines both the presence and the source of other social forms, which as Alan Hunt says, 'are not simply variant forms of legal reasoning but derive their significance and their legitimating capacity from the forms of social relations from which they originate' (see Hunt 1993).

It is not, then, just that law is created by classes or interest groups, whether instrumentally or symbolically, to maintain or increase their power. Rather, some of the relations of these groups, particularly their rules and procedures are, and indeed become, the relations of law, as relations of law similarly become social relations. This dialectical analysis suggests that 'any site of social relations is likely to be traversed by a variety of state and non-state legal networks' and that 'what constitutes "the law" in any specific context will depend upon which legal networks (or more

precisely which parts of which networks) intersect in that context, how these orders are mobilized, and how they interact' (O'Malley 1991, p. 172).

Thus, instead of treating law as an autonomous field of inquiry linked only by external relations to the rest of society, or investigating the way 'law' and 'society' as concrete entities 'influence' or 'affect' each other, constitutive legal theory takes law as its subject of inquiry, but pursues it by exploring the interrelations between legal relations and other social forms. As Hunt has observed, constitutive theory explores how law or legal relations are generated within non-legal social relations and how legal relations and social relations interpenetrate (Hunt 1993).

Peter Fitzpatrick (1984, 1992) incorporated several contributions (Foucault 1977, Moore 1978, Macaulay 1963) about the notion of interrelationships between law and the larger social matrix within which it is set, into the concept of 'integral plurality.' He says that the reason state law is, in part, shaped by the plurality of other social forms, while these forms are simultaneously being shaped by it, is because 'elements of law are elements of other forms and vice versa.' While law incorporates other forms, transforming them into its own image and likeness, the process is not unilateral but mutual, such that 'law in turn supports the other forms but becomes in the process, part of the other forms.' As such, 'state law is integrally constituted in relation to a plurality of other social forms' and 'depends on social forms that tend to undermine it' (Fitzpatrick 1984).

Fitzpatrick's theory of integral plurality is a considerable advance, both over earlier legal pluralism, and over critical legal theory, for it demonstrates that there is not so much a unilinear relationship with other social forms but rather that 'law is the unsettled product of relations with a plurality of other social forms. As such, law's identity is constantly and inherently subject to challenge and change.' Fitzpatrick's fundamental insight was to recognize that state law obtains some of its identity from its interrelationship with non-state forms and vice versa; that without this connection each would be constitutively different.

Fitzpatrick (1988, 1992) argues that the interrelations between state law and nonstate normative orders forms new entities, as a common discourse and set of practices is worked out between participating arenas of power. This is particularly evident in the context where law is being synthesized from other existing sets of rules and norms. He says that in the process of synopsis of existing rules and practices, the participating networks retain their own relative autonomy, but corresponding structures are formed which merge selected elements of the component networks into an emergent whole that becomes the new law. Pat O'Malley, developing this idea further, describes these attempts at legal synthesis as 'synoptic projects' which, are characterized by the emergence of a common and integrating discourse and set of practices worked out between interacting agencies. Such negotiation involves suppression of incompatible elements of the different participating agencies, knowledges and practices, translation of other elements into more compatible forms, and the integration of all into a workable whole, albeit often inconsistent, labile, and conflicting. In this process emergent, synthetic, or synoptic social practices and knowledges may appear (O'Malley 1991, pp. 172–3)

O'Malley claims that such 'synoptic projects' are most likely to occur where changing conditions, such as the emergence of obstacles to the continued effectiveness of existing arrangements, make their continued operation problematic. For others, such as Santos (1987) and Teubner (1992, pp. 1453–4), the process is one of 'interdiscursivity' where intra-organizational legal discourse 'productively misunderstands,' and misread (through their rereading, reinterpreting, reconstructing, and reobservation) 'organizational self-production as norm production and thus invents a new and rich "source" of law.' A similar misreading occurs, says Teubner, when the organization reincorporates legal rules developed and refined in disciplinary proceedings and makes use of them to restructure its organizational decision making.

In summary, developments during the 1990s in constitutive legal theory have taken a postmodernist stance arguing that law is mutually constituted through social relations and discursive misreading. The discursive processes of nonstate private legal systems with which state law is interrelated and interwoven provides a significant context of synoptic projects wherein old power is molded into new forms. Although critical of this new legal pluralism from the perspective of legal autopoiesis, Teubner (1992, p. 1443) accurately observes, that the relations between the legal and the social are characterized by 'discursive interwovenness,' 'are highly ambiguous, almost paradoxical: separate but intertwined, closed but open.' He argues that this new legal pluralism is 'no longer defined as a set of conflicting social norms in a given social field but as a multiplicity of diverse communicative processes that observe social action under the binary code of legal/illegal' (Teubner 1992, p. 1451). It follows, as O'Malley (1991), Hunt (1993), and Santos (1987) have argued, that to understand the form and process of state law, it is crucial to establish the social sources of the nonstate forms and private legal systems with which it is interrelated and thereby partially constituted.

Although much has been said about how law shapes social forms, and about how interests shape law, until mid-1980s, little research had been done on the way that semi-autonomous non-state private legal systems interrelate with and constitute state law while also being constituted by it. However, as Sally Merry has observed, the study of the way nonstate normative orders constitute state law is in its infancy. Import-

antly, the few studies exploring this interrelationship tend to show how the human agents who penetrate state law through introducing normative content from private non-state arenas, are drawn to reproduce state law more than they are to transform it. Neglected is the extent to which their agency contributions, as carriers for the private normative systems, transform that law in the process of constituting it. A critical research question becomes how private legal systems, such as professional self-regulatory bodies, industrial discipline and dismissals rules and procedures, and the variety of other nonstate forms of social control are constitutive of state law.

In considering the theoretical significance of constitutive legal theory to understanding the relationship between law and private legal systems it is worth stating the obvious. Law and particularly statutory law creation, is often a deliberate effort to incorporate and synthesize custom and practice, norms and rules prevailing in other settings into a general legal framework. The point, however, is that by recognizing the relative autonomy of these nonstate systems of private law, and simultaneously recognizing their ability to interrelate with state law, constitutive theory allows for the possibility that law creation will have an unpredictable outcome. In this process the eventual product is always unintentionally different from what its drafters will have intended. This is so, precisely because of the relative autonomy of the relations of non-state private legal systems.

See also: Classification of legal systems; Justice, social; Justice, social psychological perspectives in; Lex mercatoria; Sanctions, in political science

Bibliography

Cohen S 1985 *Visions of Social Control*. Polity Press, Oxford, UK

Ehrlich E 1913 *Fundamental Principles of the Sociology of Law*. Harvard University Press, Cambridge, MA

Fitzpatrick P 1984 Law and societies. *Osgoode Hall Law Journal* **22**: 115–38

Fitzpatrick P 1988 The rise and rise of informalism. In: Matthews R (ed.) *Informal Justice?* Sage, London

Fitzpatrick P 1992 *The Mythology of Modern Law*. Routledge, London

Foucault M 1977 *Discipline and Punish*. Allen Lane, London

Gurvitch G 1947 *The Sociology of Law*. Routledge and Kegan Paul, London

Henry S 1983 *Private Justice: Toward and Integrated Sociology of Law*. Routledge and Kegan Paul, London

Henry S, Milovanovic D 1996 *Constitutive Criminology: Beyond Postmodernism*. Sage, London

Hunt A 1993 *Explorations in Law and Society: Toward a Constitutive Theory of Law*. Routledge, London

Macaulay S 1963 Non-contractual relations in business: A preliminary study. *American Sociological Review* **28**: 55–66

Moore S F 1978 *Law as Process*. Routledge and Kegan Paul, London

O'Malley P 1991 Legal networks and domestic security. *Law, Politics and Society* **111**: 170–90

Pospisil L 1971 *Anthropology of Law*. Harper and Row, New York

Santos B S 1987 Law: A map of misreading. Toward a postmodern conception of law. *Journal of Law and Society* **14**: 279–302

Selznick P 1969 *Law, Society and Industrial Justice*. Russell Sage Foundation, New York

Teubner G 1992 The two faces of Janus: Rethinking legal pluralism. *Cardozo Law Review* **13**: 1443–62

S. Henry

Legalization

Legalization is a Janus-faced concept that betokens both freedom to act and subjection to law. In the first sense, it means being freed by law to act in ways formerly forbidden—for example, the repeal of Prohibition by the US Constitution's Twenty-first Amendment that legalized drinking, and the 'decriminalization' of so-called 'victimless' crimes such as prostitution, and use and possession of certain drugs. In earlier ages in Western Europe (and in claims being pressed in other parts of the globe in the early twenty-first century) legalization in this sense is exemplified in acts of tolerance that freed worshipers from the straits of official religion. In the second sense, legalization signifies transferring control of an action or behavior to the legal system. It is this second sense that constitutes the subject of this article. (Acts of legalization may embrace both senses of the term—for example, the enactment of divorce laws in some countries has served, on the one hand, to make divorce possible, while on the other hand, subjecting the divorcing parties to legal requirements.)

1. Legalization Generally

Taken over long periods, legalization as a process has been part of the development of all but the smallest of human societies. Placing responsibility for the control of many aspects of human behavior in the legal system is central to civilization itself. In the long run, legalization replaces the unfettered decision of rulers, the bonds of custom, and religious authority with the rule of law, at least considering the rule of law in its broadest constitutional sense of applying predictable and stable legal rules to the citizenry equally, and avoiding the imposition of arbitrary power. A familiar example is the growth of the common law in England

over several centuries, as legal principles and policies absorbed rules and traditions of various institutions that once held sway over most aspects of life, including family, work, property, religions, and political relations.

But legalization can be understood in narrower senses in which the spread of law and law-like institutions need not be synonymous with the rule of law or constitutional order. Throughout the twentieth century, communist societies and certain others, most notably fascist Germany, dispensed with constitutionalism in fact while pretending to it in theory, even as legal decrees and rules infiltrated numerous activities, including the economy and social relations, that had been much freer of control under the prior regime.

Nor does legalization necessarily imply any particular content to laws, which may be as liberal or repressive as the regimes that promulgate them. Laws regulating travel across political borders, which were instituted as early as the sixteenth century, have varied considerably over the centuries, and from country to country, in the strictness of the compulsory passport system that most nations have created. Except during the American civil war, no passports were required to enter or stay in the United States until World War I. Quite the opposite prevailed in France, beginning in the 1790s, and until the middle of the nineteenth century in much of Europe. Likewise, as a means of regulating internal travel, many countries adopted internal passport requirements, while others, notably the United States, have maintained steadfastly a constitutional right to travel without hindrance or papers.

In modern times, the process of legalization can be discerned most clearly in three primary developments: (a) the increasing legal regulation of activities and behaviors previously unregulated or regulated by other means; (b) the growth of private rights of action by which individuals may press legal claims against other individuals and the state; and (c) the spread of legal relations into the international sphere. The first two developments are closely connected, most especially in the United States.

2. *Increasing Legal Regulation*

In 1881, Oliver Wendell Holms Jr, the future US Supreme Court justice, declared in his classic book *The Common Law* that 'the general principle of our law is that loss from accident must lie where it falls, and this principle is not affected by the fact that a human being is the instrument of misfortune' (Holmes 1963, p. 76). But at that very time, American law was undergoing a sea change: the law of negligence would transform many accidents into legally compensable events, and the courts were pulling down a barricade of rules that they had constructed laboriously during much of the nineteenth century against defendants' liability. The growth in American law followed closely on the rise of industrialization, the explosion of technological development, and the spread of the market economy. Not all increase in the number of statutes and judicial decisions can be counted as legalization, since refinement and improvement in law is not necessarily the same as the spread of law and legal relations to new areas. Nevertheless, much of the outpouring of new law was part of the legalization of the country, a fitful though eventually sweeping movement toward 'total redress' (Lieberman 1981, p. 31) or 'total justice' (Friedman 1985). Throughout the twentieth century, the state has endeavored to refashion the theory of harm, and to encompass innovative ways of deterring and redressing the manifold injuries that came to be recognized as products of changing economic and social relations. This legalization of harm has unfolded in four principal ways.

First, from the 1880s to the time of writing, the legislatures of the states and the US Congress have enacted a tidal wave of statutes, dwarfing all previous eras of legislative enactment in all other countries. Since this is not a history of law-making, the list that follows is not intended as a chronology but rather as a snapshot of the kinds of activities that have been enveloped by law during the twentieth century: labor and employee relations (including working conditions, pay, health, safety), environmental pollution, consumer concerns (including banking and credit, housing, product safety, and advertising), the market mechanism (including antitrust, corporate structure, securities regulation, and the like), communications, elections, transportation, education, discrimination (by a host of institutions, public and private, on a variety of bases, including race, sex, ethnicity, religion, and disability), family relations (including marriage and divorce, child custody, and welfare), public assistance (including unemployment insurance, welfare benefits, and old-age security), and personal relations (including sexual relations and procreation), to name only the most prominent of the areas that have come under legal sway. Statutes operate in four ways: (a) they occasionally declare specific rules that require changes in behavior (e.g., a reduction of pollution by a certain percentage no later than a specific year); (b) even when they are not specific, statutes put decisions formerly private at risk of being second-guessed by public decision makers, including courts, prosecutors, administrators, and adversaries; (c) statutes aimed at altering complex business and other practices almost invariably wind up being enforced by bureaucracies with considerable discretion to define the law on their own terms: and (d) frequently, legislators empower individual private litigants, as well as public prosecutors, to enforce statutes in court.

Second, as just noted, legalization has been enshrined in administrative bureaucracies created to manage a host of private relations. Although some statutes spell out bright-line rules (for instance, setting the drinking age at 21 years, or placing a ceiling on interest rates that may be charged), legislatures routinely delegate not merely enforcement power to administrative agencies to carry out major regulatory programs but also the very power to define the law. One measure of the scope of administrative law in the United States at the close of the twentieth century is contained in the pages of the *Federal Register*, the US government's official digest of proposed and promulgated federal regulations; it runs to around 60,000 pages annually. The impulse to legalize has thus created a vast bureaucracy and undercut the rule of law as understood by one of its leading proponents, A. V. Dicey, who, speaking of the British legal system in the nineteenth century, said 'there can be with us nothing really corresponding to the 'administrative law' (Dicey, 1959, p. 203).

Third, alongside the legislatures and administrative agencies courts themselves have multiplied legal rights. The common-law understanding of harmful activity has expanded to encompass, for example, strict liability in tort, so that manufacturing operations that were once legally unapproachable in the absence of statutes have been regular fare at the judicial feast. Likewise, the courts in the United States have expanded vastly the panoply of public rights, particularly in the arena of the Fourteenth Amendment's command that the states provide due process and equal protection of the laws.

Fourth, as a direct consequence of the preceding three developments, individual litigants have gained the right to take their multiplying grievances to court, as discussed in the next section.

3. *Growth of Private Rights of Action*

The widely observed penchant of Americans to sue springs from a rich tradition. As long ago as 1835, Alexis de Tocqueville observed, in an often quoted passage in *Democracy in America*, that 'there is hardly a political question in the United States which does not sooner or later turn into a judicial one' (de Tocqueville 1966, p. 248). Nevertheless, during the last quarter of the twentieth century, it became the conventional wisdom of many observers that the United States was suffering from something new: a 'litigation explosion' (Olson 1991). Though the term is obscure, and such statistics as are available are subject to different interpretations (Galanter 1983, 1993, 1996, 1998), it is clear that the *types* of actions that plaintiffs may bring to court were greatly extended during the twentieth century. Many activities once entrusted to other domains—for example, politics—or left wholly unregulated and open to the whims of individuals, have become objects of legal oversight.

The most notorious example of the trend toward legalizing what was once purely political was the investigation of US presidents and high-ranking executive officials by so-called 'independent counsel.' Although in form a public prosecutor, the independent counsel was in effect a combination of a private lawyer and a client, since the executive branch was precluded from effective oversight of the prosecution. The independent counsel movement had its origins in the Watergate scandal in 1973, when the US attorney general authorized the appointment of a private lawyer as a special prosecutor, and promised to keep the government's hands off the investigation. President Nixon's discharging of Archibald Cox, the first Watergate special prosecutor, led to a federal statute providing for the appointment of independent counsel by the federal court. Ultimately, the independent counsel system ensnared President Bill Clinton in a series of investigations, culminating in the squalid inquiry into his sex life and subsequent impeachment in 1998. The independent counsel law was allowed to lapse in 1999, stemming from widespread revulsion against the legalization of such questions, and the inept tactics of Kenneth Starr, the independent counsel.

Other notable examples: The overthrow of racial segregation laws in the United States primarily began with cases filed in individual courtrooms in the 1950s. Apportionment of legislative districts, once solely a matter for political resolution (even in the face of constitutional commands widely disregarded by the legislators themselves), has become a cottage industry of litigation in the United States since the 1960s. In the 1970s, manufacturing and other practices were subject to sweeping oversight in court efforts to clean up the environment. Sexual harassment, a still evolving concept, became justiciable as the result of statutes and court decisions in the 1980s, overturning a centuries-long practice of remitting women to their private suffering in the face of it. In the 1990s, the tobacco industry was forced to agree to pay hundreds of billions of dollars in health-related claims, largely through settlements in the wake of legal claims, and a similar effort to control the distribution and sale of guns was mounted in the late 1990s. At the beginning of the twenty-first century, animal rights activists are attempting to prevent medical experimentation on animals by seeking to persuade courts to declare animals not merely property but possessors of rights.

Less generally remarked are the legal devices that serve as a spur to further legalization. In the Sherman Antitrust Act, a federal law enacted in 1890, Congress authorized private individuals to sue for restraints of trade and monopolization by business interests, and the act permitted winning plaintiffs to recover both treble damages and their lawyers' fees from losing defendants. Other devices include contingent fees in

personal injury suits; the class action, which permits the consolidation of small claims arising from the same event into a single suit, thus giving the plaintiffs' lawyers a financial incentive to proceed with the case; and the relaxation of so-called 'standing' rules that determine who is the proper party to file a suit—The US courts have occasionally been receptive, for example, to members of Congress who file suits against the President and the executive branch to try to reclaim in court what they lost in the give-and-take of legislative voting.

Legalization through private rights of action is not limited to the lawsuit itself. Understanding that lawsuits are always possible as the final outcome of legal relationships has led people to redefine personal relationships as legal ones. Since innumerable examples are possible, this one will serve as emblematic: a magazine report in 1999 declared that 'some celebrities have asked wedding guests to sign nondisclosure agreements as a condition of attending a ceremony that might attract the interest of the tab[loid]s' (Beam 1999, p. 64).

4. *Legalization in and of the International Arena*

Historically, legalization is an internal affair of each nation. Relations between nations have been governed by diplomatic and political process and, failing that, by war. But following World War II, the prosecutions of Nazi war criminals laid the groundwork for the principle that government officials are not legally free to do whatever they wish as leaders of sovereign nations, and may not always claim immunity when finally caught. The Nuremberg trials established the concept that war crimes and crimes against humanity are justiciable, and in modern times have promoted the increasingly popular notion that the shield of diplomatic immunity may be penetrated by the courts under certain circumstances. Although large numbers of war crimes since then have doubtless gone unpunished (only one major Vietnam war crimes case, concerning the massacre at My Lai, was prosecuted in the United States), by the 1990s a new sentiment led to the tentative creation of the International Criminal Tribunal, situated in the Hague, with jurisdiction to investigate and prosecute acts of genocide and other war crimes in the former Yugoslavia and Rwanda. In 1999, it indicted then President Slobodan Milosevic of Serbia, the first sitting president to be indicted by an international tribunal, and other high officials of the regime, and in 2000 it began a series of prosecutions that garnered international attention.

Other portents include successful civil suits against the estate of Ferdinand Marcos, the Philippine dictator, for human rights violations, and the British proceedings, in 1998, against former Chilean dictator Augusto Pinochet, who was arrested in London on a warrant issued by Spain, charging him with numerous civil rights violations during his 17-year rule in Chile, including murder, torture, and kidnapping. The British courts denied his claim to diplomatic immunity, although he was released in 2000 for health reasons. In the United States, the courts have been amenable increasingly to civil suits by victims of terrorism against their attackers. In 1998, US federal courts awarded more than $300 million in two suits against Iran for its involvement in supporting terrorists involved in killing and kidnapping, and in 2000 Iran was ordered to pay $300 million in damages to one of the hostages taken in Teheran in 1980. Also in 2000, thousands of Filipino slave laborers conscripted by the Japanese Army during World War II filed a class action under California law against Japanese companies or their successors that forced them to labor in mines and plantations. In that same year, Congress enacted legislation enabling victims to collect damage awards from assets in the United States of nations that sponsor terrorism.

5. *Legalization outside the United States*

Although the trend toward legalization is more advanced in the United States than in other countries, there is evidence that the direction of legal change is toward more law, not less; the increase in law is occurring in both industrialized Western nations and less developed nations worldwide, and for the same reasons that the rule of law emerged in the West. A globalized economy demands stability, and globalized communications demonstrate to peoples everywhere the gap between stable and unstable legal systems. Not every nation fits this abstraction: despite its emerging markets, communist China maintains a strong political hold on its courts, just as religiously fundamentalist countries prefer clerical decrees to secular legal rules. In the West, what few comparative studies exist suggest that litigation rates in certain European countries, adjusted for differences in costs, are not appreciably different from rates in the United States (Galanter 1992, p. 10). Two articles from the *New York Times* in 2000, chosen at random, suggest that the impulse toward legalization may be growing in radically different societies. A Chinese citizen sued Canon, Inc., a Japanese company, for $12 million for mental distress, because a computer disk promoting Canon's ink-jet printers said that China, Hong Kong, and Taiwan are separate countries, thus allegedly undermining China's sovereignty (Smith 2000). In Canada, the four major Christian Churches (Anglican, Roman Catholic, Presbyterian, and United Church) may be bankrupted by thousands of suits filed against them for sexual, physical, and 'cultural' abuse by Indian former boarding school students (Brooke 2000) .

6. Measuring Legalization

No good measures of legalization exist, in part because the topic is not studied conventionally, and in part because legal data particular to individual countries and cross-national comparative data are compiled only haphazardly. Even in the United States, a well-defined measure of litigation rates does not exist. One measure sometimes propounded is the gross number of lawyers or lawyers per capita. In the United States, lawyers numbered about 285,000 in 1960, and more than tripled to nearly one million in 2000; per capita, the ratio rose from one lawyer for every 631 Americans in 1960 to one lawyer for every 272 in 2000. By comparison, in Japan, lawyers number around 100,000, for a per capita ratio of one lawyer for every 1260 Japanese. Whether these differences are meaningful, and if so, in what way, remains controversial. The number of lawyers may be explained at least as convincingly by the number and nature of transactions and interpersonal contacts as by some tendency to greater legalization. Quantifying legalization is a research task that remains to be undertaken.

See also: Drugs, Decriminalization of: Cultural Concerns; Law as an Instrument of Social Change; Regulatory Agencies; Rule of Law

Bibliography

Beam A 1999 Tabloid law. *The Atlantic Monthly* August: 55–68

Brooke J 2000 Indian lawsuits on school abuse may bankrupt Canada churches. *New York Times*, November 2, p. A1

de Tocqueville A 1966 In: Mayer J P, Lerner M (eds.) *Democracy in America*. Harper & Row, New York

Dicey A V 1959 *The Law of the Constitution*. 10th ed. Macmillan, London

Friedman 1985 *Total Justice*. Russell Sage Foundation, New York

Galanter M 1983 Reading the landscape of disputes: What we know and don't known (and think we know) about our allegedly contentious and litigious society. *University of California at Los Angeles Law Review* **31**: 4–71

Galanter M 1992 Law abounding: Legalisation around the North Atlantic. *The Modern Law Review* **55**: January 1–24

Galanter M 1993 News from nowhere: The debased debate on civil justice. *Denver University Law Review* **71**: 77–113

Galanter M 1996 Real world torts: An antidote to anecdote. *Maryland Law Review* **55**: 1093–160

Galanter M 1998 An oil strike in hell: Contemporary legends about the civil justice system. *Arizona Law Review* **40**: 717–52

Holmes O W Jr 1963 In: Howe M de W (ed.) *The Common Law*. Little, Brown, Boston, MA

Lieberman J K 1981 *The Litigious Society*. Basic Books, New York

Olson W 1991 *The Litigation Explosion*. Dutton, New York

Smith C R 2000 Miffed Chinese sue Japan companies. *New York Times*, August 7, p. A6

J. K. Lieberman

Legislatures: United States

Legislatures are key governmental institutions performing the function of representation and lawmaking in the political system. Legislatures vary enormously in governmental power, from the weakest in parliamentary systems where principal power resides with the executive, to the strongest in separation-of-power systems, where the legislative and executive branches of government have roughly coequal status. Most of the world's democratic systems have parliamentary governments (see *Parliamentary Government*). Among separation-of-power systems, legislative bodies at the national (see *Congress: United States*) and state levels in the United States are the most highly developed. The present discussion deals with the legislatures of the 50 American states, in terms of the many functions they perform and the main elements that characterize them.

1. What Functions Do Legislatures Perform?

Legislatures perform three major functions in the political systems of their states: representing constituencies, lawmaking, and balancing power.

1.1 Representing Constituencies

The predominant responsibility of legislatures is that of representation, primarily of individuals and groups within the states. Individual legislators see themselves as representatives of their districts. While districts within states have the same number of people, districts vary in size among the 50 states. Legislators may represent as many as 796,956 people, as the 40 senators in California do, or as few as 3,924 as some of the 150 representatives in Vermont do.

One of the important jobs of senators and representatives is to provide service to constituencies and constituents—responding to people's requests for help, doing case work, and taking care of the district's interests with respect to state-aid formulas and local projects. Representation requires that organized groups and individual citizens have access to members, to committees, and to the legislative process generally to make known their views. The legislature is attentive, and on major issues, responsive to what people in a district and in the state want. On very few issues do a majority of constituents feel strongly on one side or the other. On those issues it is likely that the legislator feels similarly. Whatever the legislator's views, where there seems to be a mandate in the constituency, legislators will be inclined to go along with it as 'delegates' elected to carry out the will of the people. On the large majority of issues, however, relatively few

constituents have views and legislators respond mainly to organized interest groups in the constituency as well as to other influences. On such issues legislators can act as 'trustees,' according to their own views of the issue and political situation.

1.2 Lawmaking

Overlapping representation are a number of activities the focus of which is primarily internal. The legislative role in formulating, reviewing, and adopting a state budget has special significance in the process of making laws, since the budget is probably the most important bill (or bills) that a legislature passes. The lawmaking process itself is characterized by organized interest groups and their lobbyists and by scattered individuals, who offer diverse perspectives and positions on the issues in question.

Deliberation is an important feature of lawmaking in legislatures. It involves a give-and-take and an exchange of information and ideas. It is a vital element of committee activity, but it also goes on in leadership conferences, and informal exchanges among members. Building consensus on a bill requires more than the efficacy of deliberation; it also requires the willingness of opposing sides to enter into a bargaining process whereby they try to resolve their differences—through dealing, trading, and compromise. Sometimes one side wins decisively and the other loses decisively, but the overwhelming majority of laws enacted by a legislature are settled by means of building consensus—within a legislative party, within a legislative chamber and with the other legislative house.

1.3 Balancing Power

Constitutionally, legislatures are designated as the first branch of government and the executive as the second branch. American state legislatures, unlike parliamentary bodies or many American city councils, are relatively independent of executives and participate coequally in lawmaking for the state. Although most budgets are formulated by the executive branch and nearly all governors can veto line-item appropriations, legislatures are instrumental in deciding on spending priorities for the state and allocating funds for state agencies and state programs. Governors have a lot to say over the legislature's agenda of major issues and can veto legislative enactments; yet, the legislature's cooperation is necessary if a major law is to be adopted.

2. What Factors Play a Role in the Performance of Legislatures?

American state legislatures can be differentiated on a number of dimensions, each of which helps shape the legislature's performance of its three major functions.

2.1 Legislative Structures

With the exception of Nebraska, each state legislature is bicameral, with a senate and a house. Bills must be enacted by both bodies before they are sent to the governor for signature, or possibly veto. Senates are smaller in size than houses, normally with only one-third to one-half as many members. Senates range from Minnesota's 67-member body to Alaska's 20-member body. The largest house, in New Hampshire, has 400 members, while the smallest, in Alaska, has 40. The principal working groups of state legislatures are standing committees, which have responsibility for particular policy domains—such as education, health, welfare, and transportation—and for revenues and expenditures. Committees enable a senate or house to divide labor, allow members to specialize, and provide for a more thorough review of legislative proposals than could be done on the floor of the chamber.

2.2 Legislative Capacity

The quantity and quality of resources available to the legislature to perform its functions are important. A number of elements signify a legislature's capacity. The amount of time, both during the formal session and during the period between sessions (the interim), matters. With 44 of the 50 state legislatures meeting annually, sessions vary in length from those that run only 20 or 30 days to those that run for most of the year. California, Illinois, Massachusetts, Michigan, New York, and Pennsylvania are virtually full time, while Arkansas, North Dakota, and Wyoming are very part time. Another aspect of capacity is professional staff. The organization and size of legislative staffs range greatly. In some states staff tend to be centralized in one non-partisan agency (as in Kentucky and Ohio); in other states staff is decentralized—divided between the senate and house and among legislative parties and standing committees. The most populous states with full-time legislatures (like California and New York) have the largest staffs, while the least populous states with very part-time legislatures (like Delaware and Wyoming) have the smallest staffs. Additional resources employed by legislatures include facilities, technology, and the amount of money budgeted for the legislative operation as a whole. Whether a legislature is more or less professional depends largely on its capacity to do its job.

2.3 Legislative Members

The total number of legislators in the 50 states is 7,425. Members of senates generally serve four-year terms, while members of houses generally serve two-year terms. In 19 states legislators are limited to six, eight, or twelve years of service in the senate or house. Term

limits were adopted by the initiative process (or by threat of the initiative) in 20 states and by the legislature itself in another state, but were thrown out by the courts in three. Most members in most states are 'citizen' legislators, who supplement modest public salaries with income earned in private employment—as attorneys, business persons, teachers, farmers, etc. Large proportions of legislators aspire to careers in politics, working their way through leadership chairs in the state senate or house or running for Congress or statewide office. Relatively few incumbents leave voluntarily (excluding those who run for higher office) and relatively few who stand for re-election are defeated.

2.4 Legislative Parties

Parties are of little importance in the Nebraska legislature, which is not only unicameral but is also officially nonpartisan, and in legislatures that are controlled overwhelmingly by one party or the other. Otherwise, parties play a significant role. About two-thirds of the 99 legislative chambers are competitive, in that both the Democrats and the Republicans have held the majority in recent years and/or have a prospect of winning a majority in the near future. (This competition at the chamber level exists despite the fact that anywhere from two to four out of five senate and house districts are reasonably safe for either Democrats or Republicans.) The majority party in a legislative body normally exercises predominant control. In the lawmaking process, the minority can choose either to work with the majority in an effort to shape the legislative product or challenge the majority in an effort to create issues on which to contest the forthcoming election. Where the legislative parties are most competitive partisanship is likely to be strongest. The legislative parties employ caucuses or conferences of members, which meet during the session on an occasional basis, weekly, or every day just before the senate or house convenes.

In about three-quarters of the states the legislative parties are not only involved in governance through their influence on the lawmaking process, but are also involved in campaigns for the election of candidates to the legislature. The parties and legislative party leaders raise funds and allocate them to incumbents and challengers, who face contests in the most competitive districts. Although campaigns remain candidate-centered and local, the legislative parties exercise a centralizing influence. All of this increases partisanship in the legislature.

2.5 Legislative Leaders

The top leaders in a legislative body are the speaker of the house and the president, president pro tem, or in a few cases majority leader of the senate. The presiding officers of the body are elected by the entire membership, although the majority party candidate normally wins. Other leaders are appointed by top leaders or are selected by the party caucuses.

Legislative leaders have a number of responsibilities. First, they build consensus within their party or in the chamber as a whole in order to pass legislation and serve the interests of their partisan or chamber members. Second, they negotiate with the other house and with the executive so that legislation can proceed from passage in one chamber to enactment in law. Third, they organize their chamber, appointing members to committees and designating committee chairs. Fourth, they exercise control of the agenda and the flow of bills to the floor. Fifth, they preside over floor sessions. Sixth, they manage, or oversee the management of, the senate or house and of joint facilities. Seventh, they minister to members who need help with their bills, assistance for their districts, appointments for constituents, or other kinds of assistance. Eighth, they represent the chamber in dealings with the media. Ninth, they help their party's incumbents get re-elected and support their party's challengers to the opposition.

Today few leaders exercise the power of their predecessors. Leaders no longer have a monopoly on information or staff and members are less willing to follow leadership. Legislative leadership nowadays is more a matter of helping the membership develop consensus on where it wants to go than of getting the membership to go where their leaders want to go.

2.6 Legislative Institutionalism

A combination of factors define the institutionalism of legislatures. First is concern, which has to do with a sense of, identification with, or dedication to the legislature. Members who care about the legislature's well being and who engage in institution-building activities (and do not engage in activities that are institutionally harmful) foster institutionalism. Second is community, which encompasses the culture and norms of the legislature and requires some level of agreement on the need for civility and some manifestation of collegiality. Third is continuity, which not only provides for greater knowledge and skill on the parts of lawmakers but also promotes institutional values. Fourth is autonomy, which requires that legislatures be insulated, at least to some extent, from the environment and that they have substantial control over essentially internal processes.

See also: Balance of Power, History of; Balance of Power: Political; Congress: United States; Democratic Party; Governments; Law and Democracy; Parliamentary Government; Political Parties; Political

Representation; Presidential Government; Representation: History of the Problem; Republican Party

Bibliography

Copeland G W, Patterson S C 1994 *Parliaments in the Modern World*. University of Michigen Press, Ann Arbor, MI
Ehrenhalt A 1991 *The United States of Ambition*. Times Books, New York
Jewell M E 1982 *Representation in State Legislatures*. University Press of Kentucky, Lexington, KY
Jewell M E, Whicker M L 1994 *Legislative Leadership in the American States*. University of Michigen Press, Ann Arbor, MI
Lijphart A 1984 *Democratics*. Yale University Press, New Haven, CT
Pitkin H F 1967 *The Concept of Representation*. University of California Press, Berkeley, CA
Przeworski A, Stokes S C, Manin B 1999 *Democracy, Accountability and Representation*. Cambridge University Press, Cambridge, UK
Rosenthal A 1998 *The Decline of Representative Democracy: Process, Participation, and Power in State Legislature*. CQ Press, Washington, DC
Silbey J H 1994 *Encyclopedia of the American Legislative System*, 3 Vols. Scribner's Sons, New York
Wahlke J C, Eulau H, Buchanan W, Ferguson L C 1962 *The Legislative System: Explorations in Role Behavior*. Wiley, New York

A. Rosenthal

Legitimacy

How legal order and change are produced is a central question of law and social science inquiry. Similar questions about how social and political order—or conversely, change—come about are central to sociology and political science. Max Weber designed his concept of legitimacy in part to answer these questions. Broader discussions of legitimacy remain motivated by these themes.

Weber (1978) distinguished between obedience to commands achieved by applying force and obedience to commands achieved through legitimation. When domination is legitimated, that is, when power becomes transformed into authority, by definition subordinates consider the exercise of authority valid. Commands are obeyed because subordinates believe they are appropriate. Force, which Weber regards as an ineffective guarantor of order in the longer term, becomes unnecessary. Social order is assured through voluntary compliance.

According to Weber, there are three general types of legitimated social orders. Traditional authority stems from 'an established belief in the sanctity of immemorial traditions and the legitimacy of those exercising authority under them' (Weber 1978, p. 215). Charismatic authority stems from subordinates' belief in the singular and exceptional personal qualities of the leader. Since it inheres in attribution of extraordinary, magical qualities to an authority, it does *not* depend on any particular institutional arrangements.

In contrast, rational–legal authority is exercised in bureaucratic settings. Bureaucracy is characterized by general, written rules, a hierarchy of positions constituting both a chain of command and a career ladder, universalistic allocation of positions based on certified expertise, and rule-application in an impartial manner by bureaucratic officials. In rational–legal authority, strictly personal characteristics of leaders are irrelevant to subordinates' belief in the validity of leaders' commands. Instead, rational–legal authority stems from 'a belief in the *legality* of enacted rules and the right of those elevated to authority under such rules to issue commands' (Weber 1978, p. 215). A rational–legal system is made up of an inter-related set of formal rules specifying what must, may, or must not be done—and where relevant, by whom and under what conditions. Included are formal procedural rules that authorities must follow so that any substantive commands they issue will be valid. Under principles of legality, there also are rules that specify how authority figures come to occupy their positions, and thus to have the right and duty to issue any commands at all. Under conditions of legality, authorities are subject to the rules just as are those they command. Behavior of subordinates and superordinates alike is circumscribed, taking place according to rules.

In this view, *legal* rules are self-legitimating by definition. To better understand what characteristics of rules or rule-production can legitimate law, and what are the further consequences of law's legitimation or delegitimation, inquiry must be broadened and deepened. Nonetheless, Weber's discussion of legitimacy has been seminal for three reasons.

First, conflicts based on the divergent instrumental interests of haves and have nots are endemic in stratified social orders. Weber's concept of legitimacy suggests that people's *cognitive* and *normative* orientations to some legitimacy source—whether tradition, charisma, or legality—can counteract the tendency of interest-based conflicts to undermine social order. When legitimacy results from collective orientation to, and belief in, a particular set of 'rules of the game,' people will support and comply with decisions that run counter to their socioeconomic, political, or legal interests. Even those who consistently are disadvantaged will not rebel (Della Fave 1986, Scharpf 1999, Stryker 1994, Walker et al. 1991). Viewed this way, legitimacy may be immensely powerful in maintaining social order. Theorists and empirical researchers alike are encouraged to inquire into mechanisms producing legitimation and delegitimation of decisions, decision-making authorities, and decision-making institutions.

Second, Weber's discussion is seminal because it points toward the potential centrality of *legal order* for the legitimacy, and hence stability, of the broader

political system. Based on Weber's concept of rational–legal authority, law should play a central role in legitimating any political system that is organized bureaucratically with rules formally enacted and applied by a specialized administrative staff. As Weber's (Weber 1946, 1978) own notion of the historically increasing rationalization of modern life suggests, political systems meeting these criteria encompass much of the globe at the start of the twenty-first century. Code law and common law systems both meet the basic criteria. We most often associate democracy with law-based legitimation, but many nondemocratic systems also meet basic criteria for law to help produce political legitimacy.

Third, consistent with Weberian premises, pervasive *legalization* is a well-documented fact of contemporary economic and social life. New institutionalist research on organizations shows how *state law* penetrates the workplace (Edelman 1992, Dobbin and Sutton 1998). These scholars also document increasing organizational diffusion of *nonstate*, *but law-like* formal-rational governance systems characterized by written company rules and quasijudicial, due process-oriented grievance mechanisms (e.g., Sutton et al. 1994). Such diffusion stems in part from organizations' search for legitimacy among diverse constituencies (Stryker 2000).

In short, whether scholars focus on the formal polity or on other institutional arenas, they gain by inquiring about causes and consequences of law's legitimacy. They also should inquire about the nature and impact for the political-economy and society of any collective 'belief in legality.' Such issues are related, but not identical to, those addressed in the voluminous literature on the *rule of law* (Walker 1988, Dworkin 1985). The rest of this article reviews current concepts of legitimacy. It then discusses causes and consequences of legitimation or delegitimation.

1. Legitimacy: Conceptual Debates and Clarifications

Contemporary approaches to legitimacy contain three major themes: legitimacy as cognitive orientation to a set of binding rules of the game for political and legal acts, authorities and institutions; legitimacy as attitudinal approval of those rules; and legitimacy as behavioral consent to those rules (Stryker 1994, 2000). Implicit in each definition is a different explanation for legitimacy. A normative mechanism—internalized belief in the rightness of political or legal acts, authorities, or institutions, and a consequent internalized obligation to obey their commands—predominates for scholars who view legitimacy as attitudinal approval (e.g., Tyler 1990, Della Fave 1986, Lipset and Schneider 1983). An instrumental mechanism—ordinarily provision of material resources and rewards—predominates for scholars who view legitimacy as behavioral consent. Consent can vary along a continuum from active participation to passive acceptance to sullen obedience (compare, e.g., Habermas 1975 with Przeworski 1980). A constitutive mechanism—cognitive recognition of '*the* way things are'—predominates for scholars who view legitimacy as cognitive orientation (e.g., Walker et al. 1986), Walker et al. 1991, p. 6: emphasis in original). The research on legitimacy processes by Thomas et al. (1986, p. 378) distinguished between *validity*—'collective approval of a binding rule'—and *propriety*—'an individual's approval of the rule.' Validity does not require consensus or uniformity of belief, but only that there is 'a known institutionalized order within the collective' (Thomas et al. 1986, p. 380). Valid rules have constitutive effects when actors conform because they view the rule as binding, even and perhaps especially if they do not personally approve of the rule or if they receive adverse outcomes when the rule is applied.

Adding to the proliferation of diverse concepts and explanations for legitimacy, some scholars emphasize the formal or procedural basis for producing cognitive orientations or internalized beliefs in legal or political acts or in entire legal or political orders, such as constitutional democracy. Other scholars suggest that unless the content of legal and political commands reflects a population's substantive values, the commands will not be perceived as legitimate (see, e.g., Friedman 1981, Hyde 1983). The sum total of empirical research suggests that no matter what one's definitional starting point, all the extant approaches to legitimacy processes have produced useful insights and findings (Tyler 1990, Stryker 1994, 2000). Now these insights must be put together. As Stryker (1994, p. 85, citations omitted) noted:

> [Theory and research] [s]eparating recognition of binding rules from their approval is important because it suggests how cognitive awareness of a set of institutionalized rules can reproduce these rules by causing behavioral consent and attitudinal approval. Existence of valid rules logically presupposes that people presume peer and authority figures will behave according to, and in support of, those rules. Peer support feeds back to reinforce a person's orientation to the rules, as does that person's expectation that sanctions can be applied effectively and with peer support. Empirical findings suggest that increased collective orientation to rules lessens the mobilization of collective action against rules that are detrimental to actors' material interests. Conversely, if peers or authority figures do not play by and support the rules, collective orientation to those rules should weaken and diminished behavioral consent should follow.

2. Legitimacy: Causes and Consequences

Attitudinal support for political or legal authorities or institutions has been a more frequent operationalization of the concept of legitimacy than has been perceived obligation to comply with the decisions of

these authorities or institutions (Tyler 1990). One important line of research focusing on both types of concepts and indicators is that of Tom Tyler and diverse colleagues (Tyler 1990, 1994, et al. 1989; Lind et al. 1993).

To study the relationship between legitimacy and lawbreaking in 'the range of laws people deal with in their everyday lives,' for example, drunk driving, speeding, littering, and minor shoplifting. Tyler (1990, p. 30) conducted research on a random sample of Chicago residents. To test the hypothesis that the legitimacy of law leads to voluntary compliance, creating either an internalized obligation to obey and/or affective support for legal authorities, Tyler analyzed data from a subsample that had recent personal experiences with the courts or police. Perceived obligation to obey the courts and police was higher than was attitudinal support for the courts and police. Both perceived obligation to obey and attitudinal support directly and statistically significantly influenced reported behavioral compliance. In an earlier study focused on political legitimacy, Rasinski et al. (1985) measured institutional legitimacy by the degree to which respondents perceived that a political official's right to make decisions depended on his/her position or legal authority. Higher institutional legitimacy lessened the degree to which negative evaluations of former US President Reagan's policy performance translated into a generalized lack of support for the entire US political system. Clearly, political and legal legitimacy are both consequential and inter-related (see also Lipset 1994, Somers 1994).

As to sources or causes, as well as consequences, of legitimacy, Tyler (1990) found that a 'process perspective' on procedural justice was a more powerful predictor of law's legitimacy than was an 'outcome perspective.' The outcome perspective presumes that people assess the fairness of procedures based on the degree to which they feel they can control the *content* of the decision. The process perspective presumes that people focus on more formal criteria like neutrality, impartiality or lack of bias. To some extent Tyler's (1990) findings echo Weber's assumptions about how rational–legal authority is legitimated, since in Tyler's data, people who perceived they had been treated neutrally, impartially, honestly, politely, and with respect exhibited enhanced support for police and courts. Perceived control over outcomes also contributed to enhanced support for police and courts, but the effect of process control was greater than that of decision control. An important caveat is that neither perceived process control nor outcome control was directly related to the measure of legitimacy with the most impact on compliance with law: perceived obligation to obey.

Subsequent studies by Tyler and others continue to stress the formal procedural bases for legitimacy, whether assessed for legal or political referents. These studies investigate a wider range of legal and political acts, authorities or institutions. Tyler (1994) empirically examined the effect that evaluations of the fairness of government law-making procedures had on subjective assessments of the legitimacy of the US Congress. Evaluations of procedural fairness proved much more important for legitimacy than whether or not policies adopted by Congress matched respondents' self-interested preferences. Tyler (1994) viewed this as strong support for the role formal procedures play in maintaining public support for political institutions, and hence political stability, even and especially when citizens do not agree with the policies adopted.

Lind et al. (1993) extended empirical research from the more everyday criminal law matters investigated by Tyler (1990) to monetarily consequential federal contract and tort litigation involving both individual and corporate litigants. These lawsuits required that litigants use arbitration, but litigants could reject the arbitrator's decision, in which case the lawsuit went to trial. The authors found that litigants' acceptance of the nonbinding arbitration awards was higher for both corporate and individual actors who positively evaluated the procedural justice they received.

As in Tyler's (1990) study, positive evaluations of arbitration outcomes and of arbitration processes both significantly and positively affected perceived procedural justice, but the impact of the process evaluations was greater than that of the outcome evaluations. This points again to the role of formal procedures as a basis for legitimation. However, Lind et al. (1993) also constructed an objective indicator of how favorable or unfavorable were the outcomes their litigants received through arbitration awards. The authors assessed the direct effect of this measure and also of litigants' subjective evaluations of their outcomes on litigants' acceptance of arbitration awards, controlling for perceived procedural justice. Since a significant positive effect of favorable outcomes on voluntary award acceptance remained, instrumental mechanisms were, in part, at work. Similarly, an instrumental mechanism is suggested by Tyler et al. (1989) who found that felony defendants' evaluation of judges was affected not only by defendants' perceptions of procedural justice, but also by their perceptions of distributive justice and (negatively) by the length of sentence they received.

Finally, with respect to the role of constitutive mechanisms in producing legitimacy, researchers who distinguished conceptually between recognition of a valid rule and affective support for that rule have shown that, over time, a process of rule internationalization *does* occur. In laboratory experiments conducted by Walker and co-workers (Thomas et al. 1986; Walker et al. 1986; Walker et al. 1991), collective orientation to a set of rules produced, over time, approval of those rules by many individuals in the collective. In addition to this internalization effect, the researchers found that the constitutive aspect of

legitimacy *directly* affected mobilization of collective action. In an experiment in which an experimenter-established reward structure that disadvantaged subjects was *not* meaningfully justified by reference to some group purpose, about 80 percent of subjects tried to initiate collective action to change that structure. When the experimenters legitimated the reward structure by clearly and explicitly linking it to the purpose of the investigation, a much smaller percentage of subjects initiated collective action, even though it remained in their material interests to do so (Walker et al. 1991).

Also inspired by Weber are inquiries into the impact that scientific–technical rationalization may have on the legitimacy of government and law. For Weber, both legal and scientific rationalization are key aspects of a broader societal rationalization trend. Weber did not always distinguish clearly between legal and scientific rationalization, but later scholars analytically distinguished between formal–legal and scientific–technical rationality (Offe 1975, Therborn 1977, Stryker 1994). This makes it possible to ask how law's increasing reliance on cause-effect reasoning in place of formal, rule-oriented reasoning and law's increasing incorporation of scientific expertise will affect its legitimacy (Stryker 1994, 2000). Neo-Marxist critical scholarship presumed that increasing governmental reliance on science would legitimate government by converting political issues to be resolved by popular debate into technical issues removed from popular control (Habermas 1970, 1975). The empirical literature is mixed. Political and legal authorities do adopt scientific–technical expertise intending to increase their legitimacy (Eisner 1991, Stryker 1990). But as Stryker (1994, citations omitted) noted: 'empirical work also shows that, no matter what proponents of increased reliance on science intend, under some conditions, scientific–technical rationality increases politicization and delegitimates processes and outcomes of welfare state [political and] legal institutions.'

By analytically separating constitutive, normative, and instrumental mechanisms linking law's enhanced reliance on science to law's legitimacy, Stryker (1994, p. 853) deduced the consequences of each mechanism for 'legitimacy processes, including mobilization and conflict over legal acts, authorities and institutions; and [she theorized] the conditions under which and groups for whom the three mechanisms reinfor[ced] or, conversely undercut each other in legitimating or deligitimating law.' Stryker's (1994) theory is consistent with diverse illustrative evidence, but the theory has not been tested directly. Still, it specifies exactly how the stability-threatening 'legitimacy crises' that are the focus of so much writing on government legitimacy may get created and diffused. At the same time, and consistent with conflict theory's focus on order *through* change, Stryker (1994, pp. 890–1) suggests:

> [The] dynamics of legitimation, delegitimation, mobilization, countermobilization and conflict are an integral part of the everyday, ordinary workings of legal institutions … Democratic politics of welfare states provide institutionalized political and legal rules for changing legal acts, authorities and institutions. Individual and collective protest can and does take place through such institutionalized channels as voting, political advertising, electoral fund-raising and campaigning, organized interest group lobbying, litigation and threatened litigation. Legally circumscribed strikes, pickets, marches, boycotts, and rallies provide potentially more disruptive but still legally institutionalized methods of collective protest … when the dynamics of legitimation and delegitimation play themselves out through these institutionalized means, they help achieve political order through legally and politically induced legal change.

See also: Democracy: Normative Theory; Durkheim, Emile (1858–1917); Law as an Instrument of Social Change; Law: Change and Evolution; Norms; Rule of Law; Trust: Philosophical Aspects; Trust, Sociology of; Weber, Max (1864–1920)

Bibliography

Della Fave L R 1986 Toward an explication of the legitimation process. *Social Forces* **65**: 476–500

Dobbin F 1998 The strength of the weak state: The rights revolution and the rise of human resources management divisions. *American Journal of Sociology* **104**: 441–76

Dworkin R 1985 *A Matter of Principle*. Harvard University Press, Cambridge, MA, and London

Edelman L 1992 Legal ambiguity and symbolic structures: Organizational mediation of civil rights law. *American Journal of Sociology* **97**: 1531–76

Eisner M A 1991 *Antitrust and the Triumph of Economics: Institutions. Expertise and Policy Change*. University of North Carolina, Chapel Hill, NC

Friedman K V 1981 *Legitimation of Social Rights and the Western Welfare State: A Weberian Perspective*. University of North Carolina, Chapel Hill, NC

Habermas J 1970 *Toward a Rational Society*. Beacon Press, Boston

Habermas J 1975 *Legitimation Crisis* (trans. McCarthy T). Beacon Press, Boston

Hyde A 1983 The concept of legitimation in the sociology of law. *Wisconsin Law Review* **1983**(2): 379–426

Lind E A, Kulik C T, Ambrose M et al. 1993 Individual and corporate dispute resolution: Using procedural fairness as a decision heuristic. *Administrative Science Quarterly* **38**: 224–51

Lipset S M 1994 The social requisites of democracy revisited. *American Sociological Review* **59**: 1–22

Lipset S M, Schneider W 1983 *The Confidence Gap: Business, Labor and Government in the Public Mind*. Free Press, New York

Offe C 1975 The theory of the capitalist state and the problem of policy formation. In: Lindberg L N, Alford R, Crouch C, Offe C (eds.) *Stress and Contradiction in Modern Capitalism*. Lexington Books, Lexington, VI, pp. 125–44

Przeworski A 1980 Material bases of consent: Economics and politics in a hegemonic system. In: Zeitlin M (ed.) *Political Power and Social Theory*. JAI, Greenwich, CT, Vol. 1, pp. 21–66

Rasinski K, Tyler T R, Fridkin K 1985 Exploring the function of legitimacy: Mediating effects of personal and institutional legitimacy on leadership endorsement and system support. *Journal of Personality and Social Psychology* **49**: 386–94
Scharpf F W 1999 *Governing in Europe: Effective and Democratic?* Oxford University Press, Oxford, UK
Somers M R 1994 Rights, relationality and membership: Rethinking the making and meaning of citizenship. *Law and Social Inquiry* **19**: 63–112
Stryker R 1990 Science, class and the welfare state: A class-centered functional account. *American Journal of Sociology* **96**: 684–726
Stryker R 1994 Rules, resources and legitimacy processes: Some implications for social conflict, order and change. *American Journal of Sociology* **99**: 847–910
Stryker R 2000 Legitimacy processes as institutional politics: Implications for theory and research in the sociology of organizations. *Organizational Politics: Research in the Sociology of Organizations* **17**: 179–223
Sutton J R, Dobbin F, Meyer J W, Scott W R 1994 The legalization of the workplace. *American Journal of Sociology* **99**: 944–71
Therborn G 1978 *What Does the Ruling Class Do When It Rules?* New Left Books, London
Thomas G M, Walker H A, Zelditch M Jr 1986 Legitimacy and collective action. *Social Forces* **65**: 378–404
Tyler T R 1990 *Why People Obey the Law*. Yale University Press, New Haven, CT
Tyler T R 1994 Governing amid diversity: The effect of fair decision-making procedures on the legitimacy of government. *Law & Society Review* **28**: 809–31
Tyler T R, Casper J D, Fisher 1989 Maintaining allegiance toward political authorities: The role of prior attitudes and the use of fair procedures. *American Journal of Political Science*. **33**: 629–52
Walker G de Q 1988 *The Rule of Law: Foundation of Constitutional Democracy*. Melbourne University Press, Carlton, Victoria
Walker H A, Rogers L, Thomas G M, Zelditch M Jr 1991 Legitimating collective action: Theory and experimental results. *Research in Political Sociology* **5**: 1–25
Walker H A, Thomas G M, Zelditch M Jr 1986 Legitimation, endorsement and stability. *Social Forces* **64**: 620–43
Weber M 1946 *From Max Weber: Essays in Sociology* (trans. and eds. Gerth H H, Mills C W). Oxford University Press, New York
Weber M 1978 *Economy and Society*, 2 Vols. (eds. Roth G, Wittich C). University of California Press, Berkeley, CA

R. Stryber

Legitimacy: Political

Legitimacy is a critical but often vexing concept in politics and political science. It is critical because it goes to the very heart of any normative claim made by a government, a state, or a power that it should be willingly obeyed or respected. As such, the concept of legitimacy is closely interwined with a network of other normative and empirical concepts in political science—power, authority, rights, obligation, sovereignty, consent, institution, and the state. Legitimacy is vexing because of the difficulty of disentangling the normative bases of compliance or respect from any number of other possible motivations for obedience or deference. As with most critical and vexing concepts, the definition of legitimacy has itself been the subject of extensive debate and discussion. No single and universally acceptable definition of legitimacy exists. However, it is possible and useful to provide a discussion of its various meanings and the consequences of adopting one of them.

Perhaps the most fundamental divergence over the meaning of legitimacy is between a normative and an empirical approach to the concept. The normative approach is used by political philosophers to identify the standards by which a regime or action must be judged if it is to be regarded as legitimate. Political philosophers and the general public, for instance, might be interested in trying to answer the question: Why should the government be obeyed? And thus they might want to identify those conditions under which an authority is legitimate—the reasons why citizens *ought* to obey. A Lockean would argue, for instance, that consent is the central criterion that makes government authority legitimate (Simmons 1979). Without consent, there is no moral compulsion for citizens to obey.

In contrast, the empirical approach to legitimacy is not concerned with normative standards *per se*. Instead, it wants to explain why or when people do obey, respect, or show allegiance to a particular government, regime, state, policy, or institution. Or, conversely, why or when do they revolt, disobey, or act disloyally? Thus, the empirical approach is more concerned about whether the normative standards people hold—however defined—are met, than in debating what standards should hold. Political philosophers have sometimes criticized the social scientific approach to legitimacy as relativistic, and consequently, as potentially complicit in tyranny (Schaar 1981). The Nazi regime, so the argument might go, is deemed legitimate as long as its subjects regard it as such. As both Rodney Barker and David Beetham argue, however, the normative and the empirical approaches to legitimacy simply have different purposes and should not be regarded as antithetical (Barker 1990, Beetham 1991).

1. *Max Weber's View of Legitimacy and its Legacy*

The German sociologist Max Weber provided social science with its most enduring empirical approach to legitimacy. Weber linked legitimacy to the willingness to comply with a system of rule ('legitimacy orders') or

to obey commands ('imperative control'). An administrative staff, for instance, may obey commands because of custom, affectual ties, or material incentives. He argued, however, that compliance or obedience also typically requires a belief in the legitimacy of the system of rule or command. Every 'system of authority,' he argued, 'attempts to establish and to cultivate the belief in its 'legitimacy' (Weber 1947, p. 325). Compliance or obedience based on custom or expediency are unlikely, according to Weber, to be stable. It is worth noting here the similarity between Weber's concept of legitimacy and Marx's notion of ideology (Barker 1990, p. 59).

Weber identifies three types of 'pure' legitimate authority: *rational-legal authority* rests 'on a belief in the "legality" of patterns of normative rules and the right of those elevated to authority under such rules to issue commands'; *traditional authority* rests 'on an established belief in the sanctity of immemorial traditions and the legitimacy of the status of those excercising authority under them'; and *charismatic authority* rests 'on devotion to the specific and exceptional sanctity, heroism, or exemplary character of an individual person, and of the normative patterns or order revealed or ordained by him' (Weber 1947, p. 328). At other points in his analysis, Weber also describes a *value-rational* legitimacy, which holds 'by virtue of a rational belief in its absolute value' (Weber 1947, p. 130). Barker argues that value rationality ought to be included as a fourth type of legitimacy (Barker 1990, p. 49).

Legitimacy is at the heart of Weber's sociology of domination. Its centrality grows out of his fundamental claim that social scientific explanation ought to be based on the way in which individual action is subjectively and meaningfully oriented towards the actions of others. By arguing that different modes of domination depend on different patterns of authority, and that authority depends on whether people believe it to be justified (i.e., legitimacy), Weber sought to make legitimacy central to our understanding of social institutions and political order. This is readily apparent in his famous definition of the state, which Weber argued was 'a human community that (successfully) claims the *monopoly of the legitimate use of physical force* within a given territory' (Weber 1946, p. 78). Moreover, the concept of legitimacy is central to Weber's broader analysis of social and political change. His well-known argument about the emergence of modernity in the West claims that traditional and charismatic authority is progressively replaced by rational-legal authority.

Those following in the Weberian tradition have continued to emphasize the importance of legitimacy while refining it as an empirical tool. David Beetham has sought to reformulate Weber's description of legitimacy while maintaining its role at the heart of social scientific inquiry. The problem with Weber's definition of legitimacy, Beetham argues, is that Weber conflated belief and legitimacy. 'A given power relationship,' he writes, 'is not legitimate because people believe in its legitimacy, but because it can be justified in terms of their beliefs' (Beetham 1991, p. 11). Consequently, argues Beetham, we can analyze the legitimation of power in terms of three criteria. Power can be said to be legitmate to the extent that: (a) it conforms to established rules, (b) the rules can be justified by reference to beliefs shared by both dominant and subordinate actors, and (c) there is evidence of consent by the subordinate to the particular power relation. Conversely, a breach of rules leads to *illegitimacy*, a discrepancy between rules and supporting beliefs or the absence of shared beliefs leads to a *legitimacy deficit*, and a withdrawal of consent leads to *delegitimation*. (Beetham 1991, pp. 18–20). Beetham argues that legitimacy, by helping to promote compliance and cooperation, can enhance the order, stability and effectiveness of regimes (Beetham 1991, p. 35).

Rosemary O'Kane has challenged Beetham's claim that legitimacy is a critical concept in explaining regimes and regime change. While she approves of Beetham's attempt to refine Weber, she poses three general kinds of objections to his reformulation: (a) the definition of legitimacy in terms of moral beliefs is circular, (b) it is difficult in practice to disentangle acquiescence or compliance attributable to belief from that deriving from other factors, and (c) the focus on legitimacy leads to misleading assessments of the direction of causality (O'Kane 1993). She suggests that we limit criteria for judging legitimacy to 'legal validity' (the easiest criterion to evaluate empirically) and otherwise shift our focus from legitimacy to support. 'Investigation directly into the nature of support,' she argues, 'might prove the more rewarding for an understanding of the workings of political systems' (O'Kane 1993, p. 476). In fact, 'support' approximates a rational choice conception of legitimacy put forward by Ronald Rogowski, who defines a 'rationally legitimate' government as one that maximizes the expected utility of its citizens (Rogowski 1974, p. 43).

Beetham has replied to O'Kane that the distinction between legitimacy and support is essential if we are to grasp different logics of regime crisis and transformation (Beetham 1993). Thus, a military regime might enjoy temporary support based upon its performance, but it might have a high probability for instability in the longer term because of its lack of legitimacy. Beetham's argument recalls Seymour Martin Lipset's distinction between legitimacy and effectiveness. Legitimacy, according to Lipset, is 'the capacity of the system to engender and maintain the belief that the existing political institutions are the most appropriate ones for the society.' Effectiveness, in contrast, 'means actual peformance, the extent to which the system satisfies the basic functions of government as most of the population and such powerful groups within it as

big business or the armed forces see them' (Lipset 1963, p. 64). While legitimacy is an 'evaluative' concept, he argues, effectiveness is an 'instrumental' one.

2. From Legitimacy to Legitimation

Part of the problem with the term 'legitimacy' is that it suggests a static property of a regime or institution. Yet, it is obvious that many regimes and institutions are constantly striving to shore up their claims to legitimacy. In this sense, the terms 'legitimation' and 'delegitimation' imply the more active processes by which legitimacy is created and maintained or eroded and lost. Focusing on the process calls attention to the two-sided nature of legitimacy. On one side, legitimacy is about beliefs in the moral rightness or goodness of a regime or institution. On the other side, regimes or institutions make claims about their moral rightness or goodness. Barker has emphasized this two-sidedness, arguing that legitimacy '… is both a belief held by subjects, or by some subjects, and a claim made by rulers' (Barker 1990, p. 59).

In their treatise on the sociology of knowledge, Peter Berger and Thomas Luckmann have emphasized that legitimation is a process of explaining and justifying the validity of an institutional order. They argue that four different levels of legitimation can be distinguished. The first ('incipient') level of legitimation is the creation of language that is implicitly legitimating. The second level of legitimation entails the articulation of a set of rudimentary claims or maxims. At a third level of legitimation, an explicit 'theory' based on an autonomous body of knowledge is developed. Finally, a 'symbolic universe' develops that integrates disparate symbolic meanings into a 'totality' (Berger and Luckmann 1967, p. 96). Thus, legitimacy claims can be seen as varying in terms of their explicitness and integration.

Obviously, regimes and institutions may make highly explicit and integrated claims and still be regarded as illegitimate. Yet it is also plausible that the justificatory claims of a regime will significantly influence the beliefs of the regime's subjects, particularly if coercion or instrumental incentives are effectively used to transform these beliefs. This point brings us back to the political philosopher's fear that power might be regarded as legitimate because it is powerful. While this article has only briefly explored the concept of legitimacy, the tour has hopefully been sufficient to convince the reader that legitimacy is indeed a critical but vexing concept.

See also: Authority, Social Theories of; Legitimacy; Legitimacy, Sociology of; Political Sociology; Power in Society; Power: Political; Weber, Max (1864–1920)

Bibliography

Barker R 1990 *Political Legitimacy and the State*. Clarendon Press, Oxford, UK

Beetham D 1991 *The Legitimation of Power*. Humanities Press International, Atlantic Highlands, NJ

Beetham D 1993 In defence of legitimacy. *Political Studies* **XLI**: 488–91

Berger P, Luckmann T 1967 *The Social Construction of Reality*. Anchor Books, Garden City, NY

Lipset S M 1963 *Political Man: The Social Bases of Politics*. Anchor Books, Garden City, NY

O'Kane R H T 1993 Against legitimacy. *Political Studies* **XLI**: 471–87

Rogowski R 1974 *Rational Legitimacy: A Theory of Political Support*. Princeton University Press, Princeton, NJ

Schaar J H 1981 *Legitimacy in the Modern State*. Transaction Books, New Brunswick, NJ

Simmons A J 1979 *Moral Principles and Political Obligations*. Princeton University Press, Princeton, NJ

Weber M 1946 Politics as a vocation. In: Gerth H H, Mills C W (eds.) *From Max Weber: Essays in Sociology*. Oxford University Press, New York

Weber M 1947 *The Theory of Social and Economic Organization*. Free Press, New York

C. K. Ansell

Legitimacy, Sociology of

Legitimacy may be defined as the formula by which individuals accept a power and consider their obedience as a just commitment. Thus, it implies consent and stresses the sociological dimension of the free acceptance of a power. As such, legitimacy emphasizes the symbolical aspect of obedience. Power is not exclusively based on coercion, even in the most authoritarian systems; totalitarianism itself includes an adherence which is conceived of as justified by the people. So, the sociology of legitimacy would argue that (a) power is a social relation, (b) this social relation should be at least partly meaningful for people who are involved in it, (c) this system of meanings justifies the power and its ability to make use of violence, and (d) this symbolic relation will be manipulated by the power-holders. Thus, free acceptance does not postulate the liberty of social actors but the social construction of a free adherence to a political order.

1. The Components of Legitimacy

Therefore legitimacy is one of the main topics of every research in political science. In a Weberian perspective, legitimacy is cosubstantial to politics. As Max Weber said, 'like the political institution his-

torically preceding it, the state is a relation of men dominating men, a relation supported by means of legitimate (i.e., considered to be legitimate) violence' (Weber 1970, p. 78). Reinhard Bendix (1977, p. 294) pointed out that this 'study of domination stresses the importance of group formation and of beliefs.' Thus legitimacy should be analyzed by taking into account the formation of the social structure (i.e., social history, social stratification, divisions, etc) and the nature of the beliefs which are used in a given sequence of history. Consequently, legitimacy is concretely made of various historical and cultural references and is currently analyzed by Weber as a specific and volatile combination of types of legitimization.

Thus, beliefs in the legitimacy of a system of domination are a philosophical, a political, and a cultural matter. In a philosophical perspective, it is clearly implied by the social contract philosophy. The obedience to the law stems from the adherence of everyone to the social order which is contractually defined. In a political perspective, 'men in power want to see their position as "legitimate" and their advantages as "deserved," and to interpret the subordination of the many as the "just fate" of those upon whom it falls' (Bendix 1977, p. 294). In a cultural perspective, the political relations of subordination would be conceived as 'politics of meanings' (Geertz 1973): The intensity and the efficiency of the subordination will depend on the appropriateness of the political myths that the rulers tend to promote. 'For a state to do more than administer privilege and defend itself against its own population, its acts must seem continuous with the selves of those whose states it pretends it is' (Geertz 1973, p. 317). In other words, a sociology of legitimacy must combine the analysis of the sources of legitimacy, of its political manipulation, and of its cultural congruence.

In his famous typology, Max Weber listed three sources of legitimization used to justify the power of command. Traditional domination is the authority of the 'eternal yesterday,' that is, 'the mores sanctified through the unimaginably ancient recognition and habitual orientation to conform' (Weber 1970, pp. 78–9). In this construction a command will be obeyed, and conceived as just, only if it is strictly in accordance with a learnt tradition. Charismatic domination is legitimated by 'the extraordinary and personal gift of grace (charisma), the absolutely personal devotion and personal confidence in revelation, heroism, or other qualities of individual leadership' (Weber 1970, pp. 78–9). This is the legitimacy of the prophet, the hero, the warlord, the plebiscitary ruler, but also the demagog. Charismatic leaders have to prove that they possess charisma and thus gain the privilege to break with tradition and free their command from accordance with tradition. From this point of view, this kind of legitimacy inserts a dynamic of change since it makes it possible to get the people's support without reproducing the traditional order. However, this source of legitimacy is particularly fragile and volatile since it may be called into question by a failure or a decreasing performance of the leader. In legal domination, people obey the law rather than a person. The command has to be in accordance with the legal system and those who hold power get their authority from their institutional role. As such, they are also subject to rules which determine their competence.

This Weberian typology has an evolutionist propensity. Legal domination is conceived as rational and superior to the others, on account of its universal orientation. The law is supposed to be equal for everybody, it is clearly differentiated from the person, and has to be written in order to be valid. As such, legal domination is not only fairer than the others, it is also supposed to be more efficient. Moreover, it tends to a homogeneous vision of legitimacy throughout the world and is supposed to transcend the cultural pluralism affecting humankind. Weber has usefully nuanced this construction by conceiving three types of legitimacy as 'pure types,' contrasting with a reality made of a complex combination of all of them. The sources of legitimacy have to be found in the propensity of social relations to be based on loyalty to tradition, the attachment to a charismatic leader, and respect for the law.

2. *Strategies of Legitimization*

This complexity stresses one of the strategic dimensions of legitimization, that is to say the process by which those who hold power strive to make their commands accepted as such. The strategy implies an optimal combination of the three sources of legitimacy and a skillful attempt at persuasion toward the ruled people. The use of traditional legitimacy is a major method for mobilizing different kinds of structures, like communitarian groups, family, tribes, and clans, or for fueling social relations, like clientelism or kinship. Therefore, it is one of the main components of patrimonial domination and is an important device for balancing weakness, crisis, or even the collapse of the state. This is one of the major factors of tribalization where the state is less developed, just as it contributes to explain religious fundamentalism where the political institutions are less efficient. On the other hand, when traditional legitimacy dominates in a political action, the autonomy of power holders tends to be limited, like their capacity to innovate and to create new kinds of commitment.

Charismatic legitimization, however, creates opportunity for innovation, inventing new ideologies, producing new kinds of institution, as it does with prophetic mobilization or during revolutions. Instead of traditional institutions, charismatic legitimacy mobilizes followers and disciples who are swift to trust

in their leader. However, the more domination is purely charismatic, the more it risks being fragile, threatened by a failure or a defeat, the more it lacks permanent structure and institutional order. That is why this kind of legitimacy is relieved by a 'routinization of charisma,' by which the new order turns to a new tradition which hampers the free initiative of the leader.

In the same way, legally sanctioned procedures will limit and even encapsulate political action. Even in modern states, the power holders strive to circumvent these hindrances by reviving traditional legitimacy and by making use of their own charisma. This practice is one of the main ways of by-passing institutional rationality in modern political systems, and is also a source of differentiation among the Western democracies.

The power holder manipulates the legitimizing formulas by achieving a persuasive power on those ruled and striving to create the conditions of a consensus mobilization, that is, a downward process that increases congruence within a society. This consensus mobilization implies a socializing process by which the myths on which legitimacy is based are shared among the member-units. It implies also a propaganda by which political practices are described as conforming to these myths (Etzioni 1968, pp. 361, 470). The myths that are learnt include a world vision and a system of values on which legitimacy is built. Universalism, equality, democracy, human rights, social contract, but also modernity, progress, and rationality are the main components of legal legitimacy which reflects the evolutionist postulate of the convergence among modernizing societies.

3. *Legitimacy and History*

This is the main point stressing the cultural dimension characterizing the sociology of legitimacy. Since it is value-oriented, legitimacy obviously has a cultural orientation, implied by its own history and by the social context of its construction. Historical sociology has shown how legal domination, as it is described by Max Weber, was strongly bound to a religious and cultural situation. Joseph Strayer (1970) pointed out how the Western nation-state authority was transferred from the Roman Church's vision of authority. Reinhard Bendix (1978, p. 247ff.) showed the religious bases of royal authority in Western, Islamic, and Chinese civilization, while he emphasized that the transformation from the rule of king to the government in the name of people supposed an 'intellectual mobilization' which is itself rooted in the social structure and the cultural background of each society. Therefore, the intellectual mobilization has varied from one country to another, both for strategic reasons and according to cultural factors. These differences are real gaps in societies which are clearly far from the Western model of legal authority as the model is built in the cultural context of securalism, Christian delegation of authority and representation, Roman pre-eminence of law and institution.

In this perspective, the sociology of legitimacy should be strongly comparative. It will mobilize close comparison in order to interpret the different ways of legitimization of similar institutions, as de Tocqueville did between France and the USA (Tocqueville 1981). It will also use 'huge comparisons' and stress 'dramatic contrast' for evaluating how the Western legitimacy formulas are able to be imported into Islamic, Asiatic, or African countries (Bendix 1978, Eisenstadt and Rokkan 1973, Badie 1992). The crisis stemming from the impossible transfer and pressure of globalization and homogenization results in a chronic deficit of legitimacy in non-Western countries but also in a dangerous inversion by which power holders are constantly less legitimate than the antisystem political entrepreneurs, who are mobilizing on behalf of revivalism and a true return to endogenuous values. That is why religious fundamentalism or ethnic identity have the monopoly of protest in many developing countries while they tend to embody a strong legitimacy. Consequently, this rift between state power and legitimacy and this dramatic weakening of the consensus mobilization results in a split inside the society, promoting either a quasi-state (Jackson 1990), a 'lame Leviathan' (Callaghy 1987), or a plural society in which legitimacy is entirely encapsulated in the subnational communities (Young 1976).

Thus the sociology of legitimacy also encompasses a sociology of change and a sociology of crisis. Values do not perpetuate themselves without any change among the different social groups. For instance, the increasing influence of postindustrial or postmaterial values implies transformations in the ways of legitimating social institutions (Bell 1960, Inglehart 1977). It will also result in new kinds of social movement (Offe 1985). This intermingling of value systems and mobilizations tends to weaken the consensual formation of legitimacy in modern and postmodern societies. It implies, in them, a competition of visions which are promoting new divisions, instead of furthering a convergence toward a homogeneous modern society; thus legitimacy is nowadays jeopardized by increasing structural oppositions between different kinds of culture: materialism vs. postmaterialism, nationalism vs. transnationalism, liberalism vs. statecentrism, open society vs. closed society. These new changes are cutting across traditional ones and are directly staking out the consensual basis of modern societies in which different collective imaginations of what power should be and how it can be accepted are competing.

In the meantime, turbulence in world politics (Rosenau 1990) is destabilizing the main references of nation-state legitimacy: By promoting transnational relations, it challenges the territorial dimension of

politics and the sovereignty claim of the nation (Biersteker and Weber 1996); by encouraging globalization, it promotes common goods and common global interests and redefines the postulates of democracy (Held 1995). In short, the new dimensions of international relations emancipate legitimacy issues from the nation-state structure and transfer them to broader references: humanity, the global world, world integration, human responsibility (Deng 1997). Therefore, this renewal of the sociology of legitimacy entails new reflections on ethics and international politics (Hassner 1998), on human rights (Mullerson 1997, Sellers 1996), on intervention (Mayall 1996), and on the transformation of international law and the creation of an international criminal court.

At the same time, the discovery and the rise of these new legitimacy issues are balanced by the development of a more positivistic vision of the public policies issues which are understood through the governance paradigm (Rosenau and Czempiel 1992). According to this, the functions of government are monopolized by the state and do not exclusively come within a vertical and hierarchical construction of authority. This last point tends to take political action away from the legitimacy debate: Political action should be appreciated more and more as a technical rather than as a just action. This trend can be analyzed as an ideological attempt at depoliticization and a striving to understate the importance of legitimacy issues in modern societies (Habermas 1975).

Actually, the sociology of legitimacy has also to take into account the intensity of the conscious and willful adherence of each individual to the political and social order. Legitimacy is not only the opposite of coercion, but has also to be distinguished from the passive agreement that political obedience is currently based on. As such, legitimacy is also one of the major components of civic culture.

See also: Authority, Social Theories of; Charisma and Charismatic; Charisma: Social Aspects of; Democracy; Institutions; Legitimacy; Legitimacy: Political; Organizations: Authority and Power; Political Culture; Political Sociology; Power in Society; Power: Political; Tradition, Anthropology of; Traditions in Sociology; Traditions: Social; Weber, Max (1864–1920)

Bibliography

Badie B 1992 *L'Etat importé*. Fayard, Paris

Bell B 1960 *The End of Ideology: On the Exhaustion of Political Ideas in the Fifties*. Free Press, Glencoe, IL

Bendix R 1977 *Max Weber. An Intellectual Portrait*. University of California Press, Berkeley, CA

Bendix R 1978 *Kings or People. Power and the Mandate to Rule*. University of California Press, Berkeley, CA

Biersteker T, Weber C 1996 *The State Sovereignty as Social Construct*. Cambridge University Press, Cambridge, UK

Callaghy T 1987 The state as lame Leviathan. In: Ergas Z (ed.) *The African State in Transition*. Macmillan, London, pp. 87–116

Deng F (ed.) 1996 *Sovereignty as Responsibility*. Brookings Institution, Washington, DC

Eisenstadt S, Rokkan S (eds.) 1973 *Building States and Nations*. Sage, Beverly Hills, CA

Etzioni A 1968 *The Active Society: A Theory of Societal and Political Processes*. Free Press, New York

Geertz C 1973 *The Interpretation of Cultures*: *Selected Essays*. Basic Books, New York

Habermas J 1975 *Legitimization Crisis*. Beacon Press, Boston

Hassner P 1998 De la crise d'une discipline à celle d'une époque. In: Smouts M C (ed.) *Les nouvelles relations internationales: pratiques et theories*. Presses de la Fondation Nationale des Sciences Politiques, Paris, pp. 377–96

Held D 1995 *Democracy and the Global Order: From the Modern State to Cosmopolitan Governance*. Polity Press, London

Inglehart R 1977 *The Silent Revolution: Changing Values and Political Styles Among Western Publics*. Princeton University Press, Princeton, NJ

Jackson R H 1990 *Quasi-States: Sovereignty, International Relations and the Third World*. Cambridge University Press, Cambridge, UK

Mayall J (ed.) 1996 *The New Interventionism, 1991–94: United Nations Experience in Cambodia, Former Yugoslavia and Somalia*. Cambridge University Press, Cambridge, UK

Mullerson R 1997 *Human Rights Diplomacy*. Routledge, London

Offe C 1985 New social movements: Challenging the boundaries of institutional politics. *Social Research* **52**(4): 817–68

Rosenau J 1990 *Turbulence in World Politics: A Theory of Change and Continuity*. Princeton University Press, Princeton, NJ

Rosenau J, Czempiel E O 1992 *Governance Without Government: Order and Change in World Politics*. Cambridge University Press, Cambridge, UK

Sellers M N S 1996 *The New World Order: Sovereignty, Human Rights and the Self Determination of People*. Berg, Washington, DC

Strayer J P 1970 *On the Medieval Origins of the Modern State*. Princeton University Press, Princeton, NJ

Tocqueville de A 1981 *De la démocratie en Amérique* (2 vol.) Garnier-Flammarion, Paris

Weber M 1970 *From Max Weber: Essays in Sociology* (Gerth H, Wright Mills C eds.). Routledge, London

Young C 1976 *The Politics of Cultural Pluralism*. University of Wisconsin Press, Madison, WI

B. Badie

Leisure and Cultural Consumption

Most definitions of leisure treat it as activity (or inactivity) opposite to or the counterpoint of work. Leisure can be considered as a reward for work, as the purpose for work or as 're-creation' to recover from work. In a postmodern society in which opposites are

increasingly blended, we see attempts by individuals and institutions both to make work more leisurely and to make leisure more work-like. Thus, in their work time, people dress more casually, take sabbaticals or attend retreats; while in their 'harried leisure' away from work they 'work out,' act as purchasing agents for their households, carry laptops into the wilderness or master new work skills.

A classic definition of leisure speaks of it as 'freedom from the necessity of labor.' In the economic sense of that definition, America has become a leisure society *par excellence*, since unprecedented proportions of its citizens no longer need to earn money from their work labors to obtain adequate provisions of food, clothing and shelter. While no one has come forward with a satisfactory estimate of the proportion of the population at that level of economic comfort, it is hard to imagine that it has not increased significantly since the 1950s—probably to include a healthy majority of the US adult population (not to deny the significant minorities who remain below the poverty line and need to work to make ends meet, allowing them minimal opportunity for leisure). Yet, those Americans who have no need to care for minimal creature comforts continue to be drawn to work despite any lack of its necessity for many of them.

Indeed, economists like Schor (1991) argue that 'overworked' Americans are working longer hours and enjoying less leisure than earlier generations—at a time when their counterparts in Europe have much longer vacations and shorter working weeks. Some recent economic reports proclaim that Americans now work even longer hours per year than Japanese workers.

1. Measurement of Work and Free Time

Such arguments are limited methodologically by the data sources involved, in that they are based on estimates made by workers of their work hours rather than more detailed measurement procedures, using on-site observation, electronic pagers or time diaries. One value of time diaries is that workers report on *all* their daily activities, and not just their work time. In accounting for all time, diary figures are less prone than estimates to problems of respondent memory loss, self-projection and double counting of time.

Time-diary figures compiled over the last 35 years show a surprisingly opposite trend from figures from government surveys that ask workers to estimate their working hours (see Sect. 5.4 for some telling biases in these estimates). In diaries, notable declines in time spent at work are recorded rather than the increases or constancies in workers' estimated work hours (Robinson and Godbey 1999). Nor do the diary figures support arguments that there has been increasing variance in work (or free time).

Declines in diary work hours mean that there is more opportunity to use such freed time for 'free time' activities like socializing, fitness activity and mass media use. That opportunity could be lost, however, should the workers in question use that reduced work time for family care or personal care activities.

However, these same diary studies indicate that family care activities have also declined and that personal care has stayed about the same since the 1950s. The surprising result from detailed diary data, then, is that Americans during their prime working years (ages 18–64) now have more free time than in the 1950s or 1960s (Robinson and Godbey 1999).

To some extent, that increased free time has come about because of structural changes in the workforce and population, in that:

(a) workers enter the workforce later in life, mainly because they receive more years of college education;

(b) workers retire at age 62 or earlier, largely due to increased benefits; and

(c) workers past age 65 live longer and thus can enjoy more free time during those retirement years—when most of them are truly freed from the necessity of labor.

The end result is that, when aggregated across a lifetime, a typical person's free time hours will notably outweigh time spent at the workplace (Ausubel and Grubler 1994).

The diary studies suggest that further gains in Americans' free time are also being made during work years in the form of unofficially shorter working weeks, increased holidays and longer vacations. Nonetheless, there remains the question of why Americans continue to work more (and thus forego free time) than they need to, and whether they are incapable of appreciating leisure, as Riesman (1958) worried in the late 1950s.

2. Cultural Consumption

One answer lies in Rainwater's (1979) argument that the costs of the 'survival market basket' keep increasing as community norms and expectations about possessions keep rising. Another is the increased media emphasis on the consumption of goods and services, as characterized by Schor's (1999) 'Overspent American' and by Ritzer's (1999) depiction of 'Cathedrals of Consumption'—the ever-more elegant shopping malls, vacation cruises and home entertainment systems on which Americans spend the incomes they make from their hours at work. The hours invested in these goods and locales suggest that the 'culture' being consumed with these work dollars is rather superficial, redundant and spiritually uninspiring in contrast to classic leisure. On the other hand are arguments from some postmodernists who accept people's 'commodified' popular culture standards as more suitable

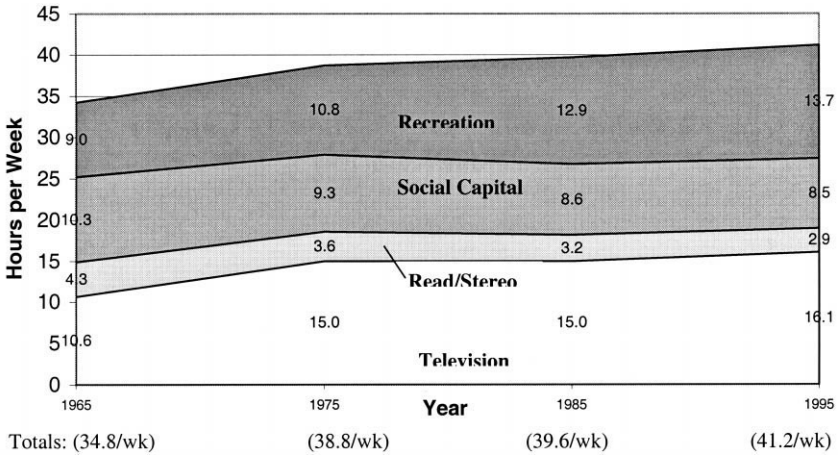

Figure 1
Trends in free time in America: 1965–95

for them than those of elitist intellectuals (as argued further below).

Similar arguments arise from examination of the free-time activities chosen by Americans during the hours they are not working. As shown in Fig. 1, close to half of their free time is spent watching television programs that are undemanding and/or uncritical mainly of materialism or environmental degradation. Television hours dwarf time spent reading, and even that reading is rather undemanding (Zill and Winglee 1990). Further concern has arisen over the decreased time spent on 'social capital' free-time activities that provide the important social 'glue' that allows societies to flourish (Putnam 2000). The optimistic finding that fitness activity has almost tripled since the 1970s (as part of other 'healthier' life-style changes, such as decreased smoking or red meat consumption) is offset by suggestions that even these activities are governed by values from work and efficiency experts.

3. *Free Time vs. Leisure*

One political philosopher who foresaw many of these developments was Sebastian DeGrazia (1962). He lamented how modern democracies had overcome the 'tyranny of man' only to submit to the 'tyranny of the clock,' exemplified by the giving of gold watches to retiring employees for whom structured time was now meaningless. DeGrazia made a sharp distinction between free time, which can be quantified, and leisure, which cannot (although if DeGrazia could reliably 'know it when he saw it,' it was measurable in the social science sense). In ancient societies, the purest leisure emerged when individuals pursued activities (music, poetry, literature) from a personal sense of commitment, quality and fulfillment. In modern societies, DeGrazia saw such leisure in decline, even as free time continued to increase.

Perhaps the closest modern equivalent to classic leisure is evident in Csikszentmihalayi's (1990) studies of 'flow,' except that here individuals seek optimal experience and joy in doing. Two ironic features of flow episodes are that (a) the sense of time in a time-oriented society largely disappears and (b) most flow episodes the author cites occur during work time rather than free time.

Like leisure, flow remains unquantified, so it is not clear whether such deep cultural experiences are more or less prevalent. One set of 'high culture' indicators that has been monitored since the 1980s concerns Americans' participation in the arts. Table 1 shows the extent to which these arts activities are engaged in as spectators vs. producers in the most recent national study (Bradshaw 1998). Although level of respondent education is the best predictor of arts participation, and these levels have been rising, trend studies indicate no notable increase in arts participation across time. (The same conclusion emerges from trend studies of other cultural indicators, such as information levels, political participation and of community participation/social capital. Delli Carpini and Keeter 1997; Putnam 2000). At the same time, some scholars have argued that it is remarkable that participation in live arts events has not declined, given the competition from easily accessible media. What is to be gained from attending a live performance of a Beethoven symphony one has heard hundreds of times on the radio, or a live performance of a Shakespeare play, when an episode of *Hill Street Blues* is easier to view and understand? Why visit an art museum when the paintings can be arrayed more meaningfully on an Internet site?

Table 1
Proportions of Americans engaging in various arts activity in prior year (1998 data; ages 18 and older)

	Attend performance	Perform	Perform publicly	Mass media		
				TV	Radio	Records
Jazz	12%	2%	1%	31%	39%	29%
Classical	16	11	1	32	41	34
Opera	5	2	0.3	15	11	11
Musical	25	8	2	25	5	12
Plays	16	N/A	N/A	23	N/A	6
Ballet	6	0.5	0.3	39	N/A	N/A
Other dance	12	13	2	N/A	N/A	N/A
Art museum	35	16	3	45	N/A	N/A
Literature	63 (Read)	12 (Write)	1	N/A	N/A	N/A

N/A Question not asked

The time-diary studies indicate that less than one percent of Americans' free time is devoted to such arts participation activities, too small to be detectable in the 'recreation' block in Fig. 1. Even the Table 1 entries total less than five episodes, or less than 20 hours, of high culture participation per year.

Nonetheless, these data sources do point to several important conclusions about the nature of such high culture activities, in that they are engaged in more by (a) better-educated individuals (also true for many popular culture activities), and (b) busy individuals who have a long working week and are heavily involved in other activities.

Both findings fit in to a larger 'Newtonian' pattern of behavior, in which 'Bodies in motion stay in motion' that applies to other leisure behavior (a corollary to Parkinson's Law that 'If you want something done, give it to a busy person').

4. *Correlates of Free Time*

Who, then, has more of the resource of free time, which at least affords the opportunity or leisure, even if rarely used for that purpose?

4.1 *Gender*

While American women are generally thought to have much less free time than men, that is only true for women who are employed (Robinson and Godbey 1999). Because many women (aged 18–64) remain out of the labor force, however, the free time of women and men is surprisingly similar (as well as what they do in that free time). The gender inequality that is found is between working and non-working women, and that gap has been increasing over time. Much the same gender patterns have been found in most other Western countries (Goldschmidt-Claremont and Pagnossin-Aligisakis 1995).

4.2 *Age*

As expected, free time is U-shaped across the life-cycle, reaching more than 40 hours a week for teens and young adults, dropping to 30 hours or less for those raising families during the peak career years of their 30s and 40s, then rising to 40 hours again in the 'empty nest' years and approaching 50 hours in the retirement years.

The main reason for this curve is not age *per se* but the various role obligations that peak in middle age—namely work, marriage, and children. The more hours of work, obviously the less free time; however, as noted above, there is far from a 1:1 correspondence. Moreover, workers who put in long hours are more active in the arts, in fitness activity, in civic participation, and in sexual activity (Robinson and Godbey 1999; Putnam 2000). Getting married usually means less free time as well, as does having children, so that married women with children have 10–15 hours less free time a week than single childless women of equivalent work status.

4.3 *Status*

While social status may be a most powerful predictor of one's power, resources, attitudes or behavior, it has little to do with how much free time people have. Part of that is due to the greater enjoyment of work (and other leisurely features) among higher status workers and the need for lower status workers to work longer hours to offset their lower wages. The rich 'leisure class' of which Veblen (1899) wrote so critically over a century ago seems to have given way to rich people who work 80–90-hour weeks today. Indeed, working long hours has become something of a status symbol in its own right, and Hochschild (1997) has documented how certain workers treat work as a consumption activity.

4.4 Ecology

Surprisingly little difference in free time can be found by region, and in presumably less hectic rural areas; nor are notable differences found in different seasons of the year.

5. Subjective Aspects of Free Time

Since leisure has been argued to be a fundamentally subjective state, it should require more subjective measures to reflect it. Four types of subjective measures that have been proposed as reflective of greater leisure include are discussed next.

5.1 Greater Enjoyment

When asked to rate the enjoyment levels of different activities in their time diaries, people also rate free-time activities higher than work, but the differences are not great; moreover, the lowest-rated daily activities usually involve housework or medical treatment (not work), and the personal care activities of eating and sleeping are rated higher than most free-time activities. Indeed, when asked to rate them side-by-side, most workers rate their enjoyment of work and free-time activities equally. Saturday and Sunday activities, when free time is 50 percent higher per day than during the week, are rated as no more enjoyable than weekday activities, further supporting the indistinguishability of work and free time.

5.2 Less Time Pressure

Despite the increased free time in society, more people now describe their lives as 'always rushed' than they did 35 years ago, with less discretionary 'time on their hands.' About half of Americans describe their lives as at least moderately stressful, and continue to report that they would give up a day's pay to get a day off from work. At the same time, more Americans say they would prefer to work longer hours in general—presumably to earn extra money.

5.3 Less Fatigue

Fewer than 25 percent of workers say they are 'very tired' at the end of their work day—much the same as found in 1975. More educated workers are less likely to feel tired, as are male workers in general.

5.4 Appreciation of Quality

The arts surveys *do* find increasing proportions of Americans saying that they enjoy more 'sophisticated' forms of music like classical music or jazz. There is evidence as well of consumption of more 'gourmet' foods and beverages. Instructive data may come from patterns of beer consumption, in which American microbreweries have taken the lead in developing classic and innovative new styles, but these have yet to capture more than 3 percent of the market; the mass-produced counterparts still claim over 90 percent of the market and there is evidence that the mass-produced product is becoming less complex over time.

Further clues to how Americans feel about their time and leisure emerge from comparing their estimated activity times with their time-diary activity times. As noted above, work activities are overestimated (by 10–15 percent) but not as much as housework and other family care activities (40–50 percent) or volunteering (200–500 percent). Sleep is underestimated by an hour a day, while free time is underestimated by 50 percent or more (18 hours estimated vs. close to 40 diary hours). (TV time, surprisingly, is accurately estimated.)

The lives that Americans think they are leading, in other words, seem to match Weber's 'Protestant ethic' more than the behavior respondents report in their time diaries.

6. International Comparisons

The conclusions drawn above refer mainly to studies conducted in the USA. Many of them, however, have been replicated in other Western societies, particularly studies of time use. Using data from the 20-country archive of time-diary studies at the University of Essex in England, for example, Bittman (1998) found average cross-countries gains in free time in the period from the mid-1960s to the mid-1990s. These gains largely mirrored those reported by Robinson and Godbey (1999) for the USA. Bittman found most weekly free time reported in diaries dating from about 1978 from the Netherlands (43 hours) and Denmark (47 hours) and least in France (31 hours) and Hungary (25 hours), with the USA (34 hours) about midway between these extremes.

Based on the Essex archive database, however, Gershuny and Fisher (2000) found that British respondents in 1996 did not report more free time than they did in the 1970s. He also found an international trend toward more work hours for upper middle-class respondents across time, as work hours for the working class declined. Again, the extent that such workers experienced more leisure or pleasure from that work suggests the need for more qualitative studies. In examining specific countries in more detail, Robinson and Godbey (1999) found both objective and subjective measures of time pressure to be rather similar in the USA and in Canada. In Japan, free time had increased from 1975 to 1991, much as in the USA over the same period, with surprisingly similar overall amounts of free time in the two countries.

However, on various measures of time pressure, Japanese respondents reported lower levels of stress and subjective pressure than US respondents. In Russia, free time also increased between 1965 and the 1990s, but was still well below US levels. Reports of time pressure in Russia, in contrast, were higher than in the USA.

In all these countries, the correlations of free time and time pressure with demographic factors were rather similar to those in the USA. Women, the more highly educated, and those in their 30s reported most time problems.

7. *Summary*

The American way of life, which is increasingly emulated in other countries, has created more free time and thus far more opportunity for leisure. However, people describe their lives as increasingly hectic and remain committed to working longer hours than necessary, presumably to keep up with the culture of consumerism. Work seems to provide more enjoyment for people than is commonly believed, even if it may be far from the ideals of leisure identified by philosophers and leisure scholars.

What is needed to advance our understanding and levels of argument is more qualitative study of the meaning of various activities in people's lives. There are several excellent individual studies of this type (e.g., Viditch and Bensman 1959), but they, too, rarely connect with one another, address issues raised by quantitative research, or are conducted with representative samples. The meaning of free time and leisure could be the ideal focus for studies using the 'representative ethnography' approach described in Robinson and Meadow (1982).

What is it, then, that would represent uses of free time to satisfy criteria for leisure, or the improved quality of life implied by the term 'leisure'? Various philosophers and social critics offer alternative criteria and rare unanimity, but one notes some of the following features of experiences or activities that involve the individual developing:

(a) a (non-work) skill or talent for its own sake;
(b) a more discriminating palate or set of tastes;
(c) a personal cosmology or philosophy of life;
(d) spiritual regeneration or insight;
(e) stronger ties to some larger community;
(f) greater self-actualization; and
(g) elements of play, curiosity, and wonder.

What seems to be called for at this point is the operationalization of a set of reliable and valid 'leisure indicators' based on such concepts, that could be regularly monitored to reflect whether progress is being made on the non-material, as well as the material, quality of non-work time.

See also: Art and Culture, Economics of; Art, Sociology of; Consumer Culture; Consumer Economics; Consumer Psychology; Consumption, Sociology of; Entertainment; Leisure, Psychology of; Leisure, Sociology of; Time Use: Research Methods.

Bibliography

Ausubel J, Grubler A 1994 Working less and living longer: Long-term trends in working time and time budgets. Working Paper 94–99. International Institute for Applied Systems Analysis, Laxenberg, Austria

Bradshaw T 1998 *Survey of Public Participation in the Arts 1997*. National Endorsement for the Arts Research Division Report, Washingington, DC

Bittman M 1998 The land of the long lost weekend: Trends in free time among working age Australians, 1974–1992. *Society and Leisure* **21**(2): 353–79

Csikszentmihalay M 1991 *Flow: The Psychology of Optimal Experience*. Harper Collins, New York

DeGrazia S 1962 *Of Time Work and Leisure*. Twentieth Century Fund, New York

Delli Carpini M, Keeter S 1996 *What Americans Know about Politics and Why It Matters*. Yale University Press, New Haven, CT

Gershuny J, Fisher K 2000 Leisure in the UK across the 20th century. In: Halsey A (ed.) *Twentieth Century British Social Trends*. Macmillan, London, pp. 620–49

Godbey G 1994 *Leisure in Your Life: An Exploration*. Venture, State College, PA

Goldschmidt-Claremont C, Pagnossin-Aligisakis E 1992 Measures of unrecorded activities in fourteen countries. In: *Human Development Report #20*. United Nations Development Office, New York

Hochschild A 1997 *The Time Bind*. Henry Holt, New York

Putnam R 2000 *Bowling Alone*. Simon and Schuster, New York

Rainwater L 1979 *What Money Buys*. Basic Books, New York

Reisman 1958 Leisure and work in post-industrial society. In: Larabec E, Meyersohn R (eds.) *Mass Leisure*. Free Press, Glencoe, IL

Ritzer G 1999 *Enchanting a Disenchanted World*. Sage Publications, Pine Forge

Robinson J, Godbey G 1999 *Time for Life*. Penn State Press, Philadelphia, PA

Robinson J, Meadow R 1982 *Polls Apart*. Seven Locks Press, Cabin John, MD

Schor J 1991 *The Overworked American*. Basic Books, New York

Schor J 1999 *The Overspent American*. Basic Books, New York

Veblen T 1899 *Theory of the Leisure Class*. Heubsch, New York

Viditch A, Bensman J 1958 *Small Town in Mass Society*. Princeton University Press, Princeton, NJ

Zill N, Winglee M 1990 *Who Reads Literature?* Seven Locks Press, Cabin John, MD

J. P. Robinson

Leisure, Psychology of

The boundaries between the everyday and the scientific understanding of leisure are blurred, and the ambiguity of the term's usage in everyday language is reflected in the lack of a standard definition in the

social sciences. The first and broadest definition is based on a narrow understanding of work as paid work. Here leisure is understood to mean time free from paid work, the entire nonoccupational domain of life. In the same way as paid work reflects the obligation to secure the necessities of life and is heteronomous, leisure has connotations of freedom and autonomy. The main problem with this definition is that housework and reproductive work in the family are also accomplished in 'free' time. Such activities, for which women bear greater responsibility than men, are thus not considered to be 'real' work and are devalued in relation to paid work.

This problem does not arise in the second, narrower definition. Here, not only paid work, but all other 'obligatory' activities, and phases of passivity, such as housework, periods of mobility such as travel between the home and the workplace, time spent asleep, and so forth, are subtracted from the total time. In contrast to such obligatory activities and passive periods, the implicit view of leisure is now more that of an autonomous domain at the individual's free disposal. However, this disregards the possibilities that occupational activities may involve autonomy or, conversely, that leisure may entail external control, constraints, and duties.

In the even narrower third definition, leisure is limited to the domain which individuals describe as such, the area they see as their realm of personal freedom. As autonomy is often linked to hedonism, subjective definitions of leisure tend to overlook self-imposed duties and restrictions.

1. *The Change in Meaning of Work and Leisure*

The conception of leisure as a domain of individual autonomy evident in all definitions only developed in the nineteenth century and in contrast to the meaning of work. Prior to this, work, residence, and family life all occurred in one place, in the same environment across the lifespan. In the course of industrialization an increasingly distinct segmentation of work and leisure began. The understanding of freedom and individual autonomy could no longer be reconciled with restrictive work (in the sense of industrial labor), and instead became associated with the modern concept of leisure as the 'free' time outside of contractually defined working hours. While the working week of industrial laborers was more than 80 hours in the middle of the nineteenth century, most employees at the beginning of the twenty-first century work less than half of this, and paid work is no longer characterized by restrictions alone. Nevertheless, the understanding of a segmentation of life has survived in cases where restrictive work is still strictly separated in time and space from all activities with greater scopes of action.

The modernization process of the twentieth century was characterized by both differentiation of the spectrum ranging from restrictive labor to the independent professions and differentiation of the partly autonomous, partly heteronomous areas between work and leisure. In recent decades, there has also been a general change in values, which has implied a shift—rather than a loss—in the meaning of work. Paid work is again more strongly determined by the demand for autonomy and self-realization, and leisure time is increasingly characterized by constraints and duties.

At the beginning of the twenty-first century, a complete dissolution of the work–nonwork segmentation can frequently be observed. This is the case wherever autonomy is required in the increasingly flexible regulation of the times, places, contents, and forms of work, and wherever unpaid work in leisure time is increasing. One can almost speak of an obligation to autonomy or controlled autonomy. This can also be said for leisure in the narrowest sense of the definition. Time at the individual's free disposal is increasingly seen as time for 'work' on oneself, for example, on one's body. The differentiation of each domain of life and the dissolution of the clear spatial and temporal boundaries between them mean that new, autonomous forms of coordination and integration are required: activities spread over the course of the days, weeks, and years must be cognitively related to one another, screened for interactions and priorities, planned in time and space, negotiated with partners, and so forth. Here again, the experience of freedom is linked to that of constraint.

2. *Leisure Research in Psychology and the Social Sciences*

Against this background, the research traditions can now be enumerated: first, leisure research across the disciplines is based on the narrow definitions and investigates leisure without reference to work. Second, psychological research takes theoretically based approaches with rather different themes, and studies subjectively significant, autonomous goals, thus indirectly dealing with leisure in the narrowest sense. Third, work–nonwork research with a sociological and psychological orientation is explicitly directed at the relationships between the domains of life. Here, the broad view of leisure—that is, time free from paid work—is taken, and insufficient attention is paid to the domains of housework and family life, for which women bear greater responsibility.

The first, applied field of leisure research does not take theoretically original approaches, and some of the many studies on, for example, time budgets or attitudes borrow theories from sociology (with respect to lifestyles or changes in values) or from developmental psychology (concepts of the stages of life). The studies deal with the use of various media, communication technology, cultural programs and

consumer goods, with travel aspirations, favorite hobbies, sports, and so forth. Particular areas of leisure time are brought to the fore, as are intragroup and intergroup comparisons on the basis of age, cohort, social status, gender, and so forth.

The second field comprises theoretically based psychological approaches to 'personal projects,' 'personal goals,' and 'personal strivings,' intersecting with self-concept and identity research as well as motivation and action theory (e.g., Little 1993, Beck 1996). Here, it is not directly a question of leisure, but of the subjective structuring of complex, longer term fields of action in everyday life and across the lifespan. However, it is noticeable that such projects often fall under leisure in the narrowest sense of the definition. Above all, when a number of action situations can be clustered together in a longer-term project with an overarching goal, the associated experiences of autonomy and heteronomy, and self-efficacy or control, are constitutive for self-concept and identity.

3. *Research on the Relationships between the Domains of Life*

In the third field of work–leisure research, individuals are not asked directly about the relationships between their thoughts, feelings, and actions in the two main domains of life. Instead, the features of each domain are surveyed separately and then correlated. Hence, indicators of experiences and actions at work are gathered on the one hand, and indicators of leisure behavior (e.g., time budgets for hobbies, membership of associations) and leisure satisfaction on the other. Results can be classified according to the central research hypotheses (Wilensky 1962). The 'generalization hypothesis' assumes a transfer of subjective positive or negative cognitions, emotions, and behavioral patterns from one domain to the other. Hitherto, the 'the long arm of the job' (Meissner 1971) has been emphasized, and the findings of many studies conducted in the USA have been interpreted in the sense of a rather one-sided influence of work on leisure. The 'compensation hypothesis' implies that it is possible to compensate for negatively valued experiences and behavioral patterns at work by engaging in positively valued ones in leisure time. This hypothesis is supported by weaker findings in subdomains and subpopulations. The 'neutrality hypothesis' postulates that thoughts, feelings, and actions in the two domains of life are unrelated. Weak or zero-order correlations in most of the studies conducted in the German-speaking countries seem to support this hypothesis.

Numerous objections can be raised to the majority of the studies. Cross-sectional studies do not show whether one-sided influences, such as that of leisure on work, or interactive processes are operating. Furthermore, the processes of generalization and compensation are not necessarily mutually exclusive, and positive correlations in certain subgroups can be canceled out by negative ones in others, resulting in zero-order correlations. In this context, however, it should be noted that compared with the original work–nonwork research (Staines 1980), later studies were more likely to differentiate between occupational groups (e.g., Kabanoff and O'Brien 1986) and emphasize the impact of the increase in women's employment on their work-free domain and associated life satisfaction (e.g., Tait et al. 1989).

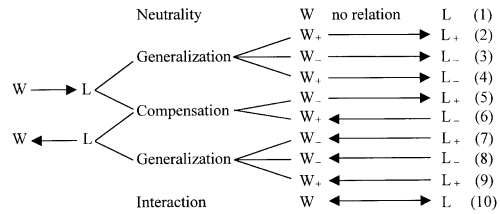

Figure 1
Classification of subjective theories. The relationships between work (W) and leisure (L) are symbolized by connecting lines; the directions of determination are indicated by arrows; the valences of the domains of life are marked as positive (+) or negative (−)

One objection to the neutrality hypothesis opens up a new perspective: if the assumption that there is no consistency of actions between the two domains of life (or that these are determined by different personality traits) is rejected on the basis of an interactional paradigm, zero-order correlations must be interpreted differently. These may well signify a subjective conception that one's own thoughts, feelings, and actions are independent. This conception could then be interpreted as a compensatory strategy operating whenever extremely negative experiences at work threaten to overshadow leisure time.

When adults are directly asked whether they see a relationship between the two domains of life, the classification of subjective 'theories,' that is, conceptual patterns, shown in Fig. 1 is obtained (Hoff 1992). Thus, not only conceptions in which work is seen as the cause of leisure experiences and actions (forms 2–5), but ones which postulate a reversed influence of leisure on work (forms 6–9) are conceivable. It is evident that private problems can be compensated in working life (form 6), or that thoughts about favorite nonwork activities can detract from work (form 7). The final form (10) consists of a reciprocal interaction of the thoughts, feelings, and actions of occupational and private life.

This classification has proved its worth in first empirical investigations. Subjective conceptual patterns can also be formulated in a generalized form and in view of personal past and future. They develop over the course of the occupational and nonoccupational

biography in connection with real constellations of restrictions and scopes of action.

4. Final Remarks on Future Research

As mentioned above, insufficient attention has been paid to housework and family life in work–leisure research. While the significance of this field has decreased since the early 1990s, there has been an expansion in research investigating the relationships between occupation and family life. Here, equally little attention has been paid to leisure in the narrow sense. Topics of investigation include conflicts due to the discrepancy of occupational and private goals, the double burden placed on working mothers, and integration problems both in everyday life and in the planning of the life course. In the future it would be useful to relate the research traditions of work and leisure to those of career and family—this would be of interest primarily where new forms of work such as tele-homework (cf. Büssing and Aumann 1996) are concerned. Occupational and gender comparisons would then be possible, and the significance of both leisure (in relation to paid work and the family) and family life and housework (in relation to paid work and leisure) could be investigated. Further research topics could include partners and their shared organization of leisure, their shared projects, and the ways in which they manage to integrate the shared domains across the life course.

See also: Leisure and Cultural Consumption; Leisure, Sociology of

Bibliography

Beck P 1996 *Persönliche Projekte—eine empirische Annäherung an komplexes Handeln*. Asanger, Heidelberg, Germany

Büssing A, Aumann S 1996 Telearbeit und das Verhältnis von Betrieb, Familie und Freizeit: Eine Bestandsaufnahme. *Zeitschrift für Arbeitswissenschaft* **50**: 225–32

Hoff E-H 1992 *Arbeit, Freizeit und Persönlichkeit. Wissenschaftliche und alltägliche Vorstellungsmuster*. Asanger, Heidelberg, Germany

Kabanoff B, O'Brien G-E 1986 Stress and the leisure needs and activities of different occupations. *Human Relations* **39**: 903–16

Little B R 1993 Personal projects and the distributed self: Aspects of a conative psychology. In: Suls J (ed.) *Psychological Perspectives on the Self*. Erlbaum, Hillsdale, NJ, Vol. 4, pp. 157–85

Meissner M 1971 The long arm of the job: A study of work and leisure. *Industrial Relations* **10**: 239–60

Staines G L 1980 Spillover versus compensation: A review of the literature on the relationship between work and nonwork. *Human Relations* **33**: 111–29

Tait M, Padgett M Y, Baldwin T-T 1989 Job and life satisfaction: A re-evaluation of the strength of the relationship and gender effects as a function of the date of the study. *Journal of Applied Psychology* **74**: 502–7

Wilensky H L 1962 Labor and leisure: Intellectual traditions. *Industrial Relations* **1**: 1–12

E. H. Hoff

Leisure, Sociology of

Leisure studies, paradoxically, are as much about work as about leisure. Indeed, central issues in the sociology of leisure revolve around leisure activities' emergence from historical changes in the structure and organization of work, the effects of work on workers and their 'need' for leisure, and the 'consumption' of leisure made possible by employed labor in the West. There are widely differing social science views on the character and meaning of leisure in the modern day, and they reflect both disciplinary and ideological stances of the researchers. This article characterizes recent approaches to the study of leisure by sociologists. It reviews the history of leisure as a sphere of activity in the West and looks at two themes that organize much of the literature on leisure—the time and place of leisure, and leisure as a reflection of culture. Finally, the article suggests trends that will likely inspire changes in leisure patterns and, hence, sociological research.

1. Approaches to Leisure

Although leisure is an increasingly organized and economically important social activity in the developed world, sociologists have rarely made leisure a central focus of study. Rather, leisure is typically treated as ancillary to the study of work, and sometimes to the study of culture. Work, sociologists often assume, defines the character, need, type, and meaning of leisure, making leisure a consequence, or a residual category in studies of work.

Sociologists are not the only social scientists to study leisure. Neoclassical economists typically defined leisure as 'nonwork,' the time 'left over' after work, with workers' preferences determining how they allocate their time. Gary Becker (1965) revised this presumption by arguing that leisure, as well as work, is an economic activity. Leisure is the time used for consuming goods purchased through wages. Recent economic approaches follow on Becker's ideas and often concern the ways in which people make tradeoffs

in their expenditure of time working or consuming. For example, economists have examined the ways in which culture (Scitovsky [1976] 1992) and regulation affect this tradeoff (Schor 1991).

Anthropologists and some sociologists have treated leisure as a form of cultural activity that includes artistic expression, play, and community building (Fine and Holyfield 1996). Anthropologists have been concerned with the impact of the recreation activities of the affluent of the developed world on the culture of those who inhabit underdeveloped regions being transformed by tourism.

Social psychologists treat leisure as an activity that can contribute to the mental and social health of actors, a sense of 'well-being.' For example, those who work too much and leisure too little can become stressed and suffer social pathologies such as isolation. Social psychologists look at the role of leisure in the life course, trying to ascertain how organized leisure contributes to social integration and psychological well-being. This approach underlies much of the curricula of professional recreation management programs and marketing studies examining the impulse for the consumption of leisure (Zins 1998).

Sociologists have tended to view leisure from the perspectives of either social class or culture, or sometimes both (Biggart 1994). Class theorists, largely rooted in British sociology, focus on the conditions that create different life circumstances for workers and see leisure variously as an opiate for harsh work conditions or as ameliorating those conditions. Culture theorists focus on leisure as either an expression of values—including gender, class, and racial values—or as an autonomous sphere that produces cultural values.

2. *Work and Leisure in Historical Perspective*

Sociologists and social historians have conducted a long-running debate on whether leisure, as a separate sphere of activity, is a modern concept in the West (nearly all studies of leisure are concerned with leisure in Western civilization). The historical record certainly shows that people have always had leisure activities. In the Greece of Aristotle's era, leisure, not work, was considered the highest form of occupation. Sports and cultural activities were important elements in the ethic of self-cultivation held by elite Athenians who did not need to work; productive labor was largely performed by slave labor and agricultural territories under military control.

In the feudal period of European history elites continued to live lives of leisure when not conducting military activities. Chivalric games and hunting served both as recreation and practice for military endeavors. Peasants, too, had ample time off from work when compared against modern standards. The workday was dictated by the amount of daylight and the demands of agriculture, and was as long as 16 hours in the summer. Tradition, however, conferred time for three meals a day, naps, and refreshments. Historians have also noted that workers sometimes 'played' while working, for example by singing and dancing as they labored. The number of days free from work varied in different parts of Europe, but probably averaged about one-third of the year. The religious tenor of the time assured observances for holy days of various sorts including long Christmas and Easter breaks and observances for the birthdays of saints. The word 'holiday' comes from 'holy day.'

Although there was ample free time in premodern Europe, was it leisure in the modern sense of time away from work (see Marfany and Burke 1997). Clearly, both elites and the lower orders amused themselves at fairs and tournaments, and had substantial free time, but they did not 'go to work' or 'go on vacation.' Life was lived more holistically with little separation between labor, leisure, religion, and secular life.

The ascent of industrial capitalism gave rise to the idea of work as a separate sphere of life and, therefore, to leisure as an alternative set of activities. The newly affluent ownership classes emulated the traditional aristocracy in leisure activities such as travel to spas and exotic locales. Workers displaced from traditional rural roots found leisure in drinking establishments, and a new commercial leisure industry rose to provide workers with travel and space for taking holidays at newly constructed seaside resorts and sporting arenas. The owners of new transportation technologies, such as trains, promoted their use for leisure excursions.

3. *Time and Place*

Paradoxically, it was the industrialization and urbanization of work that led to leisure as it is known in the modern world, the recreation (literally 're-creation') of the worker. Removed from the land and forced to find work in urban factories, laborers were also removed from traditional land-based activities such as hunting, fishing, and outdoor games. The time devoted to work grew dramatically with industrialization, too, from less than 2,000 hours in 1600 for an English male, to more than 3,500 hours in 1840, often under brutish factory conditions (Schor 1991, p. 45).

The idea of the 'workday' as fixed time for labor started in the fourteenth century in textile mill towns (Le Goff 1980). Employers rationalized work hours and pay, and instituted 'work bells' to note the beginning and end of a standardized day of employment. Although vigorously opposed by workers, the idea of a workday and the temporal control of labor were well established by the eighteenth century. There was relatively little free time for leisure in the eighteenth century, however, in either Europe or the

USA, but the time away from work was indisputably 'owned' by the worker. The separation of private and work spheres was a firmly incorporated understanding by then.

The factory system and other technological advances greatly improved worker productivity during the nineteenth century, and labor hours gradually shortened. Workers had more time in which to leisure and by the end of the century a 'leisure ethic' developed. This ethic encouraged people to use spare time for self-cultivation and civic activities. It was an era of the development of clubs—for gardening, sports, hobbies, and all manner of scientific and literary pursuits—and physical education programs in schools and colleges. It was also the period in which governments began to develop park and recreation facilities, and to see their management as a government responsibility.

World War I and its aftermath quickly ended the leisure ethic and the movement toward a shorter workday in the USA. The industrial complex that developed during the War spewed out more goods than a peacetime economy would consume. Concerned about economic growth, retailers and manufacturers promoted consumption through advertising and credit sales. A consumption ethic—'to consume is to be modern'—was widely promoted by business.

Labor practices established in the twentieth century have greatly influenced the availability of leisure for many workers. In the early decades of the century managers began to receive paid time off as a benefit, but by 1955, 90 percent of US industrial workers enjoyed paid vacations (Editors of *Fortune* 1958). After the depression of the 1930s, welfare state policies more firmly regulated work hours and conditions in developed nations. In the USA the eight-hour day was designed in part to spread available work to more laborers. Coupled with a consumption ethic, workers had both the time and justification for consumption of leisure products and services after World War II. Leisure as cultural and social practice, and as an industry, began to take its modern form.

This brief account obscures important variations. Leisure time has varied by race, social class, gender, and other social categories. African-Americans, for example, never had the same rates of regular employment as the white majority. Social and economic exclusion historically created forced 'leisure,' with some arguing that this time gave birth to cultural creativity in music and other art forms (La Rose 1996). Time budget studies of North America and Europe show that women do more childcare and housework and have less time to leisure than men (Hochschild 1997, Pronovost 1998). Europeans leisure considerably more than North Americans. Some argue that this is because Americans value consumption of material goods more than time off, but others argue that Americans have little choice about the hours they work (see Schor 1991).

Leisure practices vary by cohort and place in the life cycle, too (Pronovost 1998). Cohorts are groups of people born at about the same time. Leisure researchers note, for example, that adults born after World War II—the 'baby boomers'—have higher rates of cultural participation (e.g., museum going, theater attendance) than cohorts born earlier. In part this is due to the 'period effect' of being born during an era of prosperity with an emphasis on education. Cohorts tend to change their leisure habits depending on where they are in the life cycle, too. For example, young people consume more mass media such as films and popular music, and are more likely to participate in physically active sports than those over the age of 40.

Older people in the developed West are expected to—and expect to—retire. Retirement, a period of leisure and consumption, is expected to follow a period of work and production. Retirement is an idea that emerged at the end of the nineteenth century (Graebner 1980) and is supported by legislation, government welfare programs, the investment industry, and social expectations. Retirees are important targets of the leisure industry as retirement has increasingly been defined as a period of activity for the consumption of travel, education, and sports such as golf and organized exercise.

4. *Culture and Leisure*

The sociology of leisure increasingly has been entwined with cultural studies. Social historians and sociologists have noted that leisure practice, indeed the idea of leisure, is always imbricated with cultural values. For citizens of Ancient Greece, leisure was highly valued as a time for discussion and contemplation because elites placed a high value on self-perfection (the original meaning of 'scholar' is *to leisure*). In religious premodern Europe, holy days taken off from work were intended in part as time for devotion. In contrast, the Puritan colonists of America frowned on leisure as time taken away from the performance of God's work; they saw leisure as performing the work of the devil and prohibited dancing and card playing. In an era of rising productivity at the end of the nineteenth century, US values shifted toward the *proper* use of leisure as good for civil society and good for the development of the self. The Young Men's Christian Association (YMCA) dates from this period. Its secularized vision promoted physical fitness as part of the development of moral character. Leisure ideologies and practices are a useful lens through which to understand the character of a society.

Recent culture studies scholars have continued the tradition of understanding social order through leisure practices. The mass production and commodification

of leisure goods and services have given people of modest means and social status access to recreation activities that once were limited to the truly affluent. Cruises, time-share condominiums in luxury resort areas, user-friendly museum exhibits, and foreign cuisine are within the means and understanding of many employed workers. Global trade in media images and cultural goods transform traditional authority relations and cultural experiences once segregated by class and ethnicity.

5. Trends

Leisure is now firmly established as an acceptable and even expected part of a good life. Social practice and government regulation to limit work hours, and rising affluence in rich countries have given firm footing to the development of the leisure industry. Indeed, the scholarly recognition of leisure as an economically significant part of the economy is overdue. Spending on travel, entertainment, and vacations in the early 1900s represented about 2–3 percent of GDP in the USA, but grew to an extraordinary 10 percent of GDP by the end of the twentieth century. The influence of international travel and leisure activities on local income is substantial for a number of tourist destination cities and regions. International travel affects the balance of payments of nations. The leisure industry is increasingly large, professionalized, and globalized, and is both cause and effect of cultural and economic shifts.

The movement of people around the world for reasons of recreation, whether through travel or vicariously through film and television, has made distant events take on an immediacy and relevancy not true for earlier generations. Politics is no longer merely local, and causes such as environmentalism and ethnic conflicts take on new meaning for well-traveled populations and consumers of media images. Adventure travel and ecotourism to Nepal and the Galapagos can become the foundation for activism at home. The political implications of contemporary leisure have barely been noted.

The unification of work and play in preindustrial society was transformed into separate geographic and temporal spheres under capitalism. There is evidence that the relationship between work and play is shifting again, at least in some industries. Information technology, especially, is making it possible to work at all times of the day and night in any location. The eight-hour day has little meaning for many knowledge workers, the vanguard laborer of the twenty-first century. Many work from their residences and the rigid boundaries between the workplace and home have little meaning. Leisure practices will remain entwined with the organization of work, but in new configurations.

See also: Age, Sociology of; Consumer Culture; Consumer Economics; Consumer Psychology; Consumption, Sociology of; Dance, Anthropology of; Entertainment; Industrial Society/Post-industrial Society: History of the Concept; Industrialization; Leisure and Cultural Consumption; Play, Anthropology of; Retirement, Economics of; Work, History of; Work, Sociology of

Bibliography

Becker G 1965 A theory of the allocation of time. *Economic Journal* **75**: 493–517

Biggart N W 1994 Labor and leisure. In: Smelser N J, Swedberg R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ and the Russell Sage Foundation, New York, pp. 672–90

Fine G A, Holyfield L 1996 Secrecy, trust, and dangerous leisure: generating group cohesion in voluntary organizations. *Social Psychology Quarterly* **59**: 22–38

Fortune Editors of 1958 $30 billion for fun. In: Larabee E, Meyersohn R (eds.) *Mass Leisure*. Free Press, Glencoe, IL, pp. 161–72

Graebner W 1980 *A History of Retirement*. Yale University Press, New Haven, CT

Hochschild A 1997 *The Time Bind*, 1st edn. Metropolitan, New York

La Rose J 1996 Unemployment, leisure and the birth of creativity. *The Black Scholar* **26**: 29–31

Le Goff J 1980 *Time, Work, and Culture in the Middle Ages* [trans. Goldhammer A]. University of Chicago Press, Chicago

Marfany J-L, Burke P 1997 The invention of leisure in early modern Europe. *Past and Present* **156**: 174–98

Pronovost G 1998 The sociology of leisure. *Current Sociology* **46**(3): 1–156

Schor J B 1991 *The Overworked American*. Basic Books, New York

Scitovsky T [1976] 1992 *The Joyless Economy*, rev. edn. Oxford University Press, Oxford, UK

Zins A H 1998 Leisure traveler choice models of theme hotels using psychographics. *Journal of Travel Research* **36**: 3–15

N. Woolsey Biggart

Lesbians: Historical Perspectives

A discussion of lesbian history requires, first, the definition of the term 'lesbian.' Contemporary popular understandings of the concept rest in part on the sexual identities asserted explicitly by individuals, groups, and communities. Herein lies the difficulty for historians. How does one talk about sexual identity, especially cross-culturally and during periods when such identities were not necessarily linked to particular kinds of sexual behavior? How does one talk about sexual identity cross-culturally or during periods in

which the meanings of such behaviors were dramatically different than they are in the early twenty-first century? Can we use the term lesbian to describe only those women whom we can 'prove' had genital sexual contact with one another? Or should we broaden our definition to encompass multiple kinds of passions between women, including intimate friendships that might or might not have been sexual? Alternatively, should we talk only about lesbians historically when women identify themselves explicitly as 'lovers of women' or at the points when self-conscious communities formed around shared sexual identities? The answers to and debates over these questions have served as points of departure for scholars seeking to illuminate lesbian history. Such answers also require locating historical records that speak of female same-sex erotic ties—a challenge that is compounded by the fact that women generally left few sources behind. The records become even scarcer when the focus is on women from subordinate racial or ethnic groups or the lower classes.

Like the field of women's history, initial scholarly forays into lesbian history concentrated on individual women such as Anne Lister. Lister was an early nineteenth-century English 'gentlewoman' who, in coded diary entries, recorded the details of her sexual involvement with other women. Anne Lister understood herself as a woman who was attracted to and enjoyed sex with other women and she developed what one historian has characterized as an elaborate sexual 'vocabulary' to define her desire (Vicinus 1996). That she hid her records of her sexual relationships from potentially prying eyes through coded language suggests that she also understood that others might judge her sexual desires and actions as 'deviant.' With Lister we have all the needed ingredients to make a clear judgement that she was a lesbian—written discussion of genital sexual contact with other women available for scholarly perusal, a self-awareness, and an explicit statement of her desire for other women. Legal records provide similar unequivocal evidence of sexual contact between women. During the same historical period Amsterdam court records demonstrate a flurry of prosecutions of women for 'tribadism.' Defined as sexual acts between two women that approximated the sexual interaction between a man and his wife, these cases also seem to provide indisputable evidence of 'lesbianism' (Van der Meer 1991).

But what about other relations between women that were more ambiguous? One of the key paradigms for researching lesbian history has been that of 'romantic friendships.' The term was used to describe intimate relationships between women from the seventeenth through the twentieth centuries, characterized by declarations of undying love for one another expressed in poetry and passionate letters replete with references to kissing, cuddling, and sharing a bed. Some historians have defined these relationships, especially during the eighteenth and early nineteenth centuries in Western Europe and the United States, as marking almost a lesbian 'golden age.' According to this perspective, during this period women's abiding affection for other women, especially in the emerging bourgeoisie, was not perceived as threatening to either heterosexual marriage or dominant standards of sexual morality (Faderman in Smith-Rosenberg 1985). In fact, the cultural norm of intense female friendships among white women of the middling and upper classes was well established in England and the United States by the eighteenth century and rested on constructions of women from the dominant race and class as by nature virtuous and predominantly asexual. Thus, the acceptability of these relationships resided in their ostensibly nonsexual nature, women in these relationships characterized their feelings for one another as 'love,' yet did not proclaim, and often disclaimed any erotic attachment. There are numerous examples of 'romantic friendships' that conform to this platonic model, while others suggest relationships of a more explicitly sexual nature. Mary Grew and Margaret Burleigh were lifelong companions in eighteenth-century America.

In writing of their friendship, Grew characterized it as a 'closer union than that of most marriages' while also describing her love for Burleigh as 'spiritual' not 'passionate' nor 'sexual.' Almost a century later, however, two other 'romantic friends,' Molly and Helena, implicitly acknowledge the passion that was central to their relationship. In a letter just prior to Helena's marriage, Molly wrote to Helena's betrothed that she and Helena had loved each other 'almost as girls love their lovers' (Smith-Rosenberg 1985). At the same time Frances Willard, a white middle-class female activist and creator of the Women's Christian Temperance Union, agonized in her diaries about her love for Mary, her brother's fiancée. She understood this love to be 'abnormal' and reproached herself endlessly for the desires she felt toward her friend (Gifford 1995). In another example, two free African-American women, Addie Brown and Rebecca Primus, were involved for nine years in a relationship that was both 'highly visible' within their Hartford, Connecticut, community and explicitly sexual (Vicinus 1996). Moreover, even in relationships that generally were perceived as 'above reproach,' popular suspicions of sexual deviance abounded. Jane Pirie and Marianne Woods were 'romantic friends' who opened a boarding school for young girls in Edinburgh in 1809. In 1810 the grandmother of one of the girls withdrew her granddaughter from the school for 'very serious reasons'—her granddaughter, Jane Cumming, alleged a sexual relationship between the two women. Pirie and Woods sued the grandmother, Lady Cumming Gordon, for libel and eventually won their case. It remained nonetheless a pyrhhic victory as they had lost all of their material means during the course of the nine-year litigation. For Pirie and Woods, the allegation itself might not have been sufficient to stir the

countryside, but their long companionship, lack of male suitors, and commitment to reside together combined to heighten fears of sexual deviance (Moore 1992 in Vicinus 1996).

Characterizing these diverse examples as lesbian turns on the question of how we define the term. Is spiritual love between women sufficient for such a definition? Is explicitly sexual love sufficient? Must the women themselves or the culture in which they reside perceive such relationships as 'deviant' from heterosexual norms of behavior in order to speak of them as lesbian?

Romantic friendships would continue as one dominant framework for relationships between women into the twentieth century in both Western Europe and the United States. By the late nineteenth and early twentieth centuries, however, the infant 'sciences' of sexology and psychology were beginning to identify such relationships as sexually deviant. Sexologists like Magnus Hirschfeld (Germany), Richard von Kraft Ebbing (Germany), and Havelock Ellis (England) characterized same-sex relations between both men and women as 'sexual inversion.' While 'romantic friendships' generally were depicted as the 'least deviant' form of such 'inversion,' their formal inclusion in the category nonetheless resulted in heightened public scrutiny of close female 'friendships.' At a point historically when some white, middle- and upper-class women were increasingly able to survive economically outside of marriage, the classification of their intimate friendships with other women as potentially deviant was a serious blow to such involvements as well as to the emerging female public activism that these relationships engendered and maintained.

According to sexologists the 'most deviant' form of same-sex relations were those that involved a 'masculine' or 'mannish' woman and a 'feminine' female counterpart.

Scholars' attention to the 'mannish' woman and female cross-dresser has been another mode of inquiry for historians of lesbians. Cross-dressing, in the literal sense, sometimes, though not always, signified sexual desire for another woman. In the Western tradition, the cross-dresser also was symbolic of the potential female usurpation of male social, economic, and sexual power and privilege. In indigenous North American cultures prior to European colonization, in contrast, the female assumption of a male gender identity and vice-versa ritually was scripted and provided for women to engage in male work, adopt male dress, attire, and social standing and also allowed for such women to take wives. The Navajo used the term *nadle* to refer to a woman who had been transformed ritually into a man. This man was honored by the culture. In other cultures like the Lakota this change involved a woman becoming a third category or 'man/woman.' These ritual alterations institutionalized same-sex sexual relations and the Mohave had a particular word to describe the sexual relations between a transformed woman and her female sexual partner or wife (Roscoe 1988, Rupp 1999). In contrast, female cross-dressing in Western Europe and the United States was an explicit transgression of gender boundaries and from the seventeenth through the twentieth centuries carried with it the presumption of sexual deviance. That said, European and Euro-American women donned male attire for a variety of reasons, including accessing the greater economic, social, and sexual freedoms that were available to men. There were untold numbers of women who cross-dressed and served in the military as sailors and soldiers. We have the stories only of those whose identity was exposed for some reason. Once discovered cross-dressing women often faced public censure and some were tried under fraud and impersonation laws. Catherine Margaretha Linck was a cross-dressing soldier in eighteenth-century Germany. After leaving the army Linck married a woman. Her/his masquerade was discovered when his/her wife told her mother of Linck's secret, resulting in the mother bringing Linck to court for the deception (Faderman 1981, Rupp 1999).

While Linck was sentenced to death for her/his transgression, other cross-dressing soldiers who were 'discovered' faced lesser punishments for their acts and some were even lauded publicly. Deborah Samson was a cross-dressing soldier who served in the Continental Army during the American Revolution. Samson's 'exploits' were celebrated as epitomizing the patriotic fervor of the colonists in their war against England. Although Samson married a man after leaving military service, the presumption that sexual deviance would accompany her actions was made apparent by her subsequent apology for her 'masquerade,' and assurances to public audiences that she had not engaged in any sexual transgressions during her cross-dressing years (Meyer 1996, Wheelwright 1989).

Although Catherine Linck married a woman after her military service, there is no way of knowing whether she cross-dressed in order to engage in sexual relations with another woman. In contrast, during the late nineteenth and early twentieth centuries in the United States some women's cross-dressing was a means of explicitly claiming such a sexual desire. Prescriptive sexual norms of the period depicted women as sexually passionate, but mandated that passion be directed at men and contained within marriage. In such a context, some women dressed as men to make explicit their sexual desires for and relations with other women (Newton 1984). During the twentieth century, in North America and Western Europe, the cross-dresser and the mannish woman became archetypes of the sexual invert or lesbian. This connection between 'mannishness' and female sexual inversion or lesbianism was supported by sexologists' and psychologists' arguments that gender inversion,

defined by female cross-gender behavior or appearance, signaled potential sexual inversion. In 1934 the German sexologist Magnus Hirschfeld went even further when he drew a direct connection between women's desire to enter the military and their potential homosexuality.

Hirschfeld believed that female soldiers, because they engaged in the 'most extreme' cross-gender behavior, were the most 'confused' sexually (Meyer 1996, Wheelwright 1989). By the 1920s psychiatrists and psychologists would include sexual object choice in their definition of homosexuality, however, gender inversion remained a powerful cultural marker of female homosexuality in Western Europe and the United States throughout the twentieth century.

The popular conflation of mannishness and lesbianism was articulated through assumptions that masculine occupations and pursuits like military service or sport would both attract and produce female homosexuals. Women entering the military in England and the United States during World War II, for example, found that they had to engage with popular presumptions that their military service was an indicator of their potential sexual deviance. Similarly, the participation of women in sports during the twentieth century was framed by the figure of the mannish lesbian athlete. The institutional answer to such presumptions was to present both female athletes and soldiers as 'feminine' and avowedly heterosexual (Meyer 1996, Cahn 1994). While institutional responses were meant to decrease lesbian visibility, such reactions also made the association between lesbians and gender inversion more culturally salient. Therefore efforts to repress sexual deviance resulted in heightened public consciousness of particular sexual identities and the community spaces that supported them (Foucault 1990). Individual servicewomen and female athletes positioned themselves within these discussions by defining themselves or 'others' as lesbians. Thus, while many women who were heterosexual found themselves accused of sexual deviance because of military service or sports activity, others flocked to these all-female cultural spaces precisely in order to find others 'like themselves,' allowing for the common identification among lesbians.

Cross-culturally and historically many same-sex institutions provided space and cultural approbation of female enclaves in which some women could explore their sexuality. From the medieval period forward, convents and monasteries provided another same-sex environment within which women might develop intimate relationships with one another. The early seventeenth-century Italian records of an ecclesiastical investigation into the case of Benedetta Carlini, Abbess of the Convent of the Mother of God, document her sexual relationship with another sister (Brown 1982). Catholic female religious orders in Western Europe warned routinely against 'particular friendships' that might develop between members of a sisterhood. Although the sexual relations were not explicitly proscribed, the presumption remained that emotional intimacy could lead to sexual intimacy. In the early twentieth-century United States, prisons for women and girls came under heightened scrutiny because of reports that inmates engaged in sexual relationships with one another. The single-sex environment commentators saw as facilitating such behavior augmented suspicions of other such venues like female boarding schools and colleges. In both England and the United States the publicly accepted 'crushes' that schoolgirls developed for one another were increasingly viewed with alarm rather than the amused tolerance that had greeted such relations in the nineteenth century.

While female same-sex communities provided the possibility for some women to develop sexual relationships with one another, such communities were not themselves necessarily 'lesbian communities.' The investigation of communities created around shared sexual identities marks another investigative route for scholars of lesbian history. Communities of women identifying themselves by their sexual desire for other women have been documented as early as the eighteenth century in London and in such unexpected places as Salt Lake City, Utah, in the United States. In the 1920s and 1930s a group of white, middle-class women of predominantly Mormon backgrounds created a community for themselves based on their common identity as lesbians. This group included several women who were married to men, others who set up households with women, and some who were 'unattached.' The members of this community confined their social activities to their homes, downplayed sexual expression as a factor in their relationships, and tried to lead otherwise 'respectable' lives. Their lack of visibility as sexual vs. social actors and the protections offered them by their class and racial status mitigated against their encountering any hostility because of their sexual identities. In contrast, by the 1920s working-class lesbians in the United States began to develop communities that made visible a lesbian sexuality symbolized by and articulated through a 'masculine' woman and her 'feminine' partner. While butches, recognizable by their masculine attire, haircuts, and mannerisms were the targets of harassment throughout the mid-twentieth century, their *femme*-partner's attire, haircuts, and mannerisms were more conventionally feminine. These working-class, urban, lesbian communities and the butch/*femme* cultures at their centers, formed around their members, shared identities as lesbians. Most significant, individuals within these communities were understood to be and articulated themselves explicitly as sexually desiring other women. In cities like Buffalo, New York, United States, working-class lesbians carved public spaces for themselves in bars and restaurants and resisted police raids and street violence to create for themselves a community of women who engaged in sexual relations

with other women. This resistance prefigured the Gay Liberation Movement of the 1970s.

Defining communities and individuals as lesbian requires some indication of how they understood and presented themselves. Returning to the example of Anne Lister, we see a woman who described herself not only as engaging in sexual relations with other women, but as part of a community of women who desired women. Similarly, members of working-class butch/*femme* communities through the 1960s saw their sexual relationships with other women as essential to their identity as lesbians. The Salt Lake City community, in contrast, consisted of women who downplayed any sexual aspect of their identities, although they understood themselves as lesbians. Likewise, lesbian feminists in the United States and Great Britain during the 1970s and 1980s denigrated what they characterized as the 'role-playing' evident in butch/*femme* relationships and emphasized the importance of women's relationships with other women *sans* sex. Some defined lesbian as a continuum that encompassed sexual relations between women and also included nonsexual female friendships.

Scholars have been divided equally on the question of how to characterize lesbian identities historically. Some have argued against a focus on sexual activity as the defining characteristic of lesbian lives. These historians contend that focusing instead on the variety of 'passions' between women or behaviors and lives that were 'lesbian-like' will tell us more about the context within which women made sexual and social choices. Other historians have asserted that we must acknowledge the ways in which women's same-sex love has always challenged male power and 'sexual prerogatives.' As such, understanding lesbian history is essential to understanding women's history and the history of gender relations.

While there is no one answer to the queries posed at the beginning of this article, what is clear is that, in the end, sexual identity has been both self-disclosed and externally imposed or 'assumed' from various cultural markers. There are many lessons to draw from the exquisite difficulty of illuminating lesbian lives historically. Narrow the definition of lesbian too much and you risk erasing the experiences of many who loved women—draw the circle too broadly and you risk rendering any definition of lesbian that emerges from historical investigation meaningless. This tension frames 'lesbian history' and the historical search for lesbian lives including women's passionate friendships, sexual desire for other women, and the creation of communities. This tension continues as we enter the twenty-first century at a moment when lesbian and gay is being replaced by lesbian/gay/bisexual/transgender, is being replaced by 'queer.' The question remains—is this simply a move to a greater pluralism? Does it mark a sea change in contemporary understandings of sexuality? Or does it represent the erasure of the particularities of sexual identity and the material and ideological circumstances informing such particularities? The answer to these questions is the next challenge for scholars of same-sex or lesbian sexuality.

See also: Feminist Movements; Feminist Theory: Radical Lesbian; Foucault, Michel (1926–84); Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender, Feminism, and Sexuality in Archaeological Studies; Homosexuality and Psychiatry; Lesbians: Social and Economic Situation; Masculinities and Femininities; Queer Theory; Rationality and Feminist Thought; Regulation: Sexual Behavior; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sexual Behavior: Sociological Perspective; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Preference: Genetic Aspects; Sexuality and Gender; Social Movements and Gender; Transsexuality, Transvestism, and Transgender

Bibliography

Bennett J M 2000 Lesbian-like and the social history of lesbianisms. *Journal of the History of Sexuality* **9**(1–2): 1–24

Brooten B J 1996 *Love Between Women: Early Christian Responses to Female Homoeroticism*. University of Chicago Press, Chicago

Brown J C 1982 Lesbian sexuality in Renaissance Italy: the case of Sister Benedetta Carlini. *Signs: Journal of Women in Culture and Society* **9**(4): 271–8

Brown J C 1986 *Immodest Acts: The Life of a Lesbian Nun in Renaissance Italy*. Oxford University Press, New York

Cahn S K 1994 *Coming on Strong: Gender and Sexuality in Twentieth-century Women's Sport*. The Free Press, New York

D'Emilio J, Freedman E B 1988 *Intimate Matters: A History of Sexuality in America*, 1st edn. Harper & Row, New York

Donoghue E 1995 *Passions Between Women: British Lesbian Culture, 1668–1801*, 1st US edn. Harper Collins, New York

Faderman L 1981 *Surpassing the Love of Men: Romantic Friendship and Love Between Women from the Renaissance to the Present*, 1st Morrow Quill paperback edn. Morrow, New York

Faderman L 1991 *Odd Girls and Twilight Lovers: A History of Lesbian Life in Twentieth-century America*. Columbia University Press, New York

Foucault M 1990 *The History of Sexuality*, Vol. I. Vintage Books, New York

Gifford De Swarted (ed.) 1995 *Writing Out My Heart: Selections from the Journal of Frances E. Willard, 1855–96*. University of Illinois Press, Urbana, IL

Kennedy E L, Davis M D 1993 *Boots of Leather, Slippers of Gold: The History of a Lesbian Community*. Routledge, New York

Meyer L D 1996 *Creating G.I. Jane: Sexuality and Power in the Women's Army Corps During World War II*. Columbia University Press, New York

Nestle J 1987 *A Restricted Country*. Firebrand Books, Ithaca, NY

Nestle J (ed.) 1992 *The Persistent Desire: A Femme–Butch Reader*, 1st edn. Alyson Publications, Boston, MD

Newton E 1984 The mythic mannish lesbian: Radclyffe Hall and the new woman. *Signs: Journal of Women in Culture and Society* **9**(4): 557–75
Roscoe E (ed.) 1988 *Living the Spirit: A Gay American Indian Anthology*. St. Martin's Press, New York
Rupp L J 1999 *A Desired Past: A Short History of Same-sex Love in America*. University of Chicago Press, Chicago
Smith-Rosenberg C 1985 *Disorderly Conduct: Visions of Gender in Victorian America*, 1st edn. A. A. Knopf, New York
Trumbach R 1991 London's sapphists: From three sexes to four genders in the making of modern culture. In: Epstein J, Straub K (eds.) *Body Guards: The Cultural Politics of Gender Ambiguity*. Routledge, New York
van der Meer T 1991 Tribades on trial: Female same-sex offenders in late eighteenth-century Amsterdam. *Journal of the History of Sexuality* **1**(3): 424–45
Vicinus M (ed.) 1996 *Lesbian Subjects: A Feminist Studies Reader*. Indiana University Press, Bloomington, IN
Wheelwright J 1989 *Amazons and Military Maids: Women Who Dressed as Men in Pursuit of Life, Liberty, and Happiness*. Pandora, London

L. D. Meyer

Lesbians: Social and Economic Situation

Lesbians, or women who primarily have sexual and emotional relationships with other women, face a distinctive set of social, legal, and cultural challenges and stigmas in many parts of the world that are not faced by heterosexual women. Although lesbians have proven relatively difficult to study, increasing social science attention has illuminated key features of lesbians' social and economic situation, including workplace outcomes, employment discrimination, and family formation. An overview of this body of research reveals that lesbians generally face economic and social disadvantage, both in the workplace and in organizing their family life.

1. *Legal and Normative Challenges*

In almost all countries, lesbians' economic and social conditions are determined within a social, legal, and cultural context that is at best negligent in responding to lesbians' particular needs and at worst hostile to the existence of lesbians (Rosenbloom 1996). As of this writing, lesbians are treated differently from heterosexual women in every known legal system. In many countries, same-sex sexual practices between women are punishable by prison terms (although apparently this is enforced rarely against lesbians compared with gay men). States also actively oppress lesbians through arrests and police harassment, discrimination in government employment, exclusion from military service, and censorship of lesbian voices and organizations. Political action by lesbians and gay men has resulted in some expansions of rights for lesbians in some countries, however, such as the explicit bans on sexual orientation discrimination contained in the constitutions of South Africa, Ecuador, and Fiji.

In many societies, norms about gender roles and sexuality stigmatize women who form primary and intimate relationships with other women, and lesbians are often denied the protection of law against the social, physical, and economic abuse that accompanies social stigma. Some evidence of the existence of norms against lesbianism (or homosexuality more generally) comes from public opinion surveys from Russia, Serbia, Romania, and the United States (see Rosenbloom 1996). Reports from 30 countries provide examples of physical violence against lesbians, including rape, murder, forced conversion therapies, incarceration, and suicide.

The potential economic consequences of social stigma include being cut off from family support, being denied housing or access to important public resources, and facing discrimination in employment. While human rights laws in most countries ban many forms of sex discrimination and offer recourse when discrimination occurs, human rights laws rarely protect lesbians from discrimination based on sexual orientation. Firing or refusing to hire a lesbian because of her sexual orientation is legal in most of the world.

Legal distinctions and social stigma also greatly influence lesbians' family position. No country allows two women (or two men) to marry legally. In a few countries (the Netherlands, Sweden, Norway, Denmark, Hungary, and Iceland) two women are allowed to form and register a state-recognized relationship but are excluded from some rights pertaining to marriage, such as adoption rights or the right to a state church wedding. In some other countries or parts of countries (as in certain cities in the United States), more limited recognition is accorded same-sex couples for purposes of certain social programs, inheritance rights, or employment benefit provision, for example. Discriminatory family policies also often deny lesbians access to reproductive services, child custody, parental status, and public and private resources available to married couples.

Overall, behavior motivated by antilesbian norms about sexuality and gender and the laws that permit such behavior place lesbians in an unequal position vis-à-vis heterosexual women.

2. *Methodological Issues in Studying Lesbians*

Since the 1970s, researchers have begun to study the economic and social position of lesbians. Research on lesbians must first address the tricky issue of defining 'lesbian.' A study of sexuality in the United States reveals the lack of concordance between several possible definitions (Laumann et al. 1994). Women

who are attracted sexually to other women will not necessarily act on those attractions. Women who have had sex with women may not refer to themselves as 'lesbians.' Researchers have focused generally their attention on women who have sexual relationships with other women or who identify themselves as lesbian, depending on the historical period or on the culturally appropriate and legally relevant definition (see *Sexual Orientation: Biological Influences*; *Sexual Orientation: Historical and Social Construction*).

Because lesbianism, however defined, is not a readily identifiable physical trait, and because of the social stigma attached to being a lesbian, lesbians have often kept their sexual identities and intimate lives hidden from a hostile society. This protective strategy makes research on lesbians difficult, and researchers continue to find recruitment of subjects challenging. Many researchers have addressed this situation by using 'convenience samples' drawn from bars or organizations frequented by lesbians, for instance. Others have used 'snowball sampling' methods to recruit study participants through personal networks of other lesbians. These nonprobability sampling strategies do not allow researchers to generalize a study's findings to the broader population of lesbians, unfortunately, since respondents might differ in important ways from the typical lesbian. Probability sampling allows generalizations but faces the problem of encouraging accurate responses to survey questions on sexual orientation, and only a few such (quite expensive) studies have been attempted. (For an excellent survey of these methodological issues, see Solarz 1999.)

As a result of these problems, existing research on lesbians' economic and social conditions must be interpreted carefully. Generalizations about all lesbians based on a nonrandom sample are inappropriate, and even generalizations from probability samples might be inappropriate if disclosure of sexual orientation on a survey is biased in some way. Furthermore, without a larger random sample of lesbians than has yet been constructed, attention to differences among lesbians based on race, ethnicity, class, or nationality, for instance, is impossible.

In addition to the methodological challenges, research on lesbians has been hampered by the stigma associated with such research. Some political scientists and sociologists in the United States and lesbian researchers from 14 countries, for instance, have reported difficulties in doing research on lesbian and gay-related topics, including job discrimination, professional isolation, and lack of funding (see Solarz 1999).

Despite these roadblocks, researchers have persevered, spurred on by social scientists' recognition that the study of lesbians (and gay men) can shed light on larger social and economic dynamics (e.g., Blumstein and Schwartz 1983) and by the need for knowledge created by political debates arising from an international movement for gay and lesbian rights (see *Gay/Lesbian Movements*). A growing cadre of social scientists in sociology, anthropology, psychology, economics, history, political science, and other fields has analyzed lesbians' situation with respect to employment and other economic issues and to family life.

3. *Lesbians' Position in Employment*

Social scientists, particularly economists and sociologists, have become increasingly interested in how social identities, such as race, class, and gender, influence an individual's labor market experience and economic position. Researchers have also begun to study discrimination based on sexual orientation, since lesbians are legally vulnerable to employment discrimination. Survey data is available from several countries which suggests that discrimination is not a rare occurrence, although most of these surveys are nonrandom and particular statistics cannot be generalized. (Also, in some cases, lesbians and gay men were lumped together for reporting purposes in the original studies.)

A review of surveys conducted over time in different parts of the United States found that between 11 percent and 44 percent of lesbians and gay men reported experiences of employment discrimination at some point in their lives (Badgett 1997). In 1993, 8 percent of British lesbians surveyed believed that they had been fired because of their sexual orientation, and almost half (48 percent) had been harassed at work (Rosenbloom 1996). Almost one-quarter (23 percent) of Irish lesbians and gay men surveyed had experienced discrimination in training, hiring, or promotion (GLEN and NEXUS 1996). No independent confirmation of these experiences exists, in general, but the credibility of these claims is supported by the fact that employers surveyed in some countries, such as the United States (Badgett 1997) and Zimbabwe, plainly state to researchers that they would not hire or promote lesbians (or gay men).

Not surprisingly, then, concealing sexual orientation is a dominant strategy for lesbians to avoid employment discrimination. In surveys of gay men and lesbians in the United States, anywhere from one-half to four-fifths of respondents reported hiding their sexual orientation at work. Over two-thirds (68 percent) of the British lesbians surveyed concealed their sexual orientation from some or all of their co-workers. Only 42 percent of Irish lesbians and gay men considered themselves to be completely 'out of the closet' at their workplaces. Concealment might provide some protection from direct discrimination, but that strategy also imposes costs on lesbians, who may face increased workplace stress and the sacrifice of career-enhancing social contact with co-workers.

The evidence of self-reported discrimination is obviously suggestive of a serious economic issue for

lesbians. One standard independent test for the presence of discrimination looks for lower group earnings as evidence of discrimination. Evidence from random samples of lesbians and gay men working full-time in the United States (the only country with appropriate data) show that gay men earn significantly less than heterosexual men after controlling for productive characteristics, but lesbians' and heterosexual women's earnings are basically indistinguishable (Badgett 1995, Klawitter and Flatt 1998).

While the earnings evidence appears to contradict the claims of discrimination against lesbians, they might experience discrimination without earning less than heterosexual women for several reasons. First, lesbians might be crowded into the same female-dominated, relatively low-wage occupations as heterosexual women in which wages are closer to (or equal to) the minimum wage and thus cannot be reduced much for lesbians. Second, since lesbians will not have a male partner's earnings to fall back on, lesbians might actually have more labor market experience, skills, or training. Support for this possibility comes from the fact that some studies show that lesbians have higher labor force participation rates than do heterosexual women. But if lesbians are more skilled or experienced, they ought to earn more than heterosexual women, and a finding of earnings equality is evidence of discrimination. Unfortunately, available datasets do not have explicit measures of training or of labor market experience, so researchers cannot control for any sexual orientation-related differences in these characteristics in earnings comparisons. Third, it is possible that predominantly male supervisors are more prejudiced against gay men than against lesbians. Future research will be necessary to distinguish between these possibilities.

A clearer form of discrimination that reduces lesbians' economic well being relates to the interacting effects of labor market practices and family law. In some countries, such as the United States, significant parts of employee compensation come in the form of non-wage benefits, such as pensions and health care coverage. Employers often incorporate family members in such benefits, providing healthcare coverage for employees' legal spouses and children as well as for the employee. Since lesbian workers cannot marry their same-sex partners, those partners are ineligible for spousal benefits, reducing lesbians' effective total compensation. A small but growing number of employers in the United States and Canada recognize employees' same-sex domestic partners for benefits purposes.

4. *The Position of Lesbians and Their Families*

Social scientists long have studied and theorized about families as social and economic units, and some attention has focused on lesbians and their families. Work by anthropologists and sociologists demonstrates that lesbians form relationships with other adults that would typically be recognized as kinship relationships. Lesbians often form primary relationships with one other woman, in some cases entering into state-sanctioned partnerships where those are allowed or going through religious or personal union ceremonies without state recognition (see Blumstein and Schwartz 1983 or Lewin 1998 for studies of the United States). Other kinship structures, such as the 'families we choose' described in the San Francisco Bay Area by Weston (1991), extend beyond household boundaries and include individuals tied to one another through emotional and sometimes material bonds. No statistically legitimate estimates of the proportion of lesbians forming such relationships exists in any country, unfortunately.

Many lesbian families include children. Several surveys based on probability samples of the United States suggest that 20–32 percent of lesbians have children in their households. Many of those children, if not most of them, are from past heterosexual relationships, and an unknown number are probably the result of donor insemination or adoption in the so-called 'lesbian baby boom' of the 1980s and 1990s. Research in the US and UK has found that children raised by lesbians are well cared for and have scores on standard psychological tests that indicate healthy psychological adjustment (see surveys by Patterson 1995, 1998; see also *Nontraditional Families and Child Development*).

Studies of the inner workings of lesbian families find some similarities in the types of tasks and pressures faced by those families. Lesbians in families raise children, divvy up housework, take jobs to support their families, and care for the family's adult members. As with heterosexual families, lesbians appear to arrive at family decisions through bargaining and appeals to norms, individual preferences, individual skills, and efficiency.

Although the processes within families show some similarities to heterosexual families, other contextual factors are quite different and result in marked differences. Studies of nonrandom samples of lesbian and heterosexual couples in the US and Britain show that lesbian couples have more equal distributions of housework and labor market work than do heterosexuals and that their power relations are less affected by which person has higher earnings than is true of heterosexual or gay male couples (Blumstein and Schwartz 1983, Kurdek 1995, Dunne 1997).

Social scientists have offered varying explanations for these patterns, with the hypotheses rooted in different theories about the division of family labor. Lesbians might have different and more egalitarian norms related to the contributions of partners and the economic role of women. Lesbians' expectations about likely labor market earnings for each partner might be more equal than those for married opposite-sex

couples, reducing the advantage of having one partner do all the housework and the other partner earn all the money income. Lesbians do not have access to marriage, an institution that encourages stricter specialization of household labor. Or two lesbians' bargaining power might be more equal as a result of more equal earnings potential. All of those hypotheses would explain more egalitarian divisions of labor within the household, and additional research will be necessary to identify which explanations are most appropriate.

The legal and institutional conditions within which lesbians form and organize families are also quite different from heterosexual women. As noted earlier, lesbians cannot marry their same-sex partners. And outside of the few countries giving lesbians the option of a registered partnership, lesbians are excluded from the kind of legal framework and public resources that support other families. Lesbian couples lack legal ties to each other that assign inheritance rights, allocate property after a breakup, provide access to income security programs, and make couples eligible for spousal employment benefits, for example.

In most countries lesbian mothers face the prospect of losing custody of their children to heterosexual fathers (Rosenbloom 1996). Lesbian mothers usually cannot become a legal parent of a child born to or adopted by her lesbian partner, and a nonlegal parent has no legal rights to visitation or custody of the child if the couple splits up. These legal disadvantages for lesbians and their families imply that lesbians face the same family-oriented tasks and challenges as heterosexual women and their families but do so with additional uncertainty and without legal and economic support.

5. *Future Research Directions*

As noted, many important questions about lesbians' position have provisional answers and await further research. Knowing more about lesbian's workplace conditions—including their experiences of discrimination and occupational choices, for instance—and family lives—including alternative family forms, the division of household labor, and childbearing decisions—will require the development and use of large-scale probability samples that allow generalizations. With larger samples, more attention can be paid to differences among lesbians, such as those related to race, culture, class, and nationality. In addition, more extensive use of qualitative methods will be necessary to research the often subtle processes at play in workplaces and families. Researchers in the United States and Canada and in Western Europe are already moving in these directions.

Studies of lesbians and gay men might also present 'natural experiments' for researchers interested in the impact of policy changes and in social and cultural change. As norms evolve and if collective action by lesbians and gay men results in political victories, lesbians and their families will likely respond to cultural and legal changes in ways that will help to illuminate the underlying social processes that interest social scientists.

See also: Feminist Movements; Feminist Theory: Radical Lesbian; Foucault, Michel (1926–84); Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender and Feminist Studies; Gender, Feminism, and Sexuality in Archaeological Studies; Heterosexism and Homophobia; Homosexuality and Psychiatry; Lesbians: Historical Perspectives; Masculinities and Femininities; Placebo Studies (Double-blind Studies); Queer Theory; Rationality and Feminist Thought; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Safety, Economics of; Sexual Behavior: Sociological Perspective; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Preference: Genetic Aspects; Sexuality and Gender; Social Movements and Gender; Transsexuality, Transvestism, and Transgender

Bibliography

Badgett M V L 1995 The wage effects of sexual orientation discrimination. *Industrial and Labor Relations Review* **48**: 726–39

Badgett M V L 1997 Vulnerability in the workplace: Evidence of anti-gay discrimination. *Angles* **2**: 1–4

Blumstein P, Schwartz P 1983 *American Couples*: *Money*, *Work*, *Sex*, 1st edn. Morrow, New York

Dunne G A 1997 *Lesbian Lifestyles: Women's Work and the Politics of Sexuality*. University of Toronto Press, Toronto, ON

Gay and Lesbian Equality Network (GLEN) and NEXUS Research Cooperative 1996 *Poverty: Lesbians & Gay Men, The Economic and Social Effects of Discrimination*. Combat Poverty Agency, Dublin, Eire

Klawitter M, Flatt V 1998 The effects of state and local antidiscrimination policies for sexual orientation. *Journal of Policy Analysis and Management* **17**: 658–86

Kurdek L 1995 Lesbian and gay couples. In: D'Augelli A R, Patterson C J (eds.) *Lesbian*, *Gay*, *and Bisexual Identities Over the Lifespan*. Oxford University Press, New York

Laumann E O, Gagnon J H, Michael R T, Michaels S 1994 *The Social Organization of Sexuality: Sexual Practices in the United States*. University of Chicago Press, Chicago

Lewin E 1998 *Recognizing Ourselves: Ceremonies of Lesbian and Gay Commitment*. Columbia University Press, New York

Patterson C J 1995 Lesbian mothers, gay fathers, and their children. In: D'Augelli A R, Patterson C J (eds.) *Lesbian*, *Gay*, *and Bisexual Identities Over the Lifespan*. Oxford University Press, Oxford, UK

Patterson C J 1998 The family lives of children born to lesbian mothers. In: Patterson C J, D'Augelli A R (eds.) *Lesbian, Gay, and Bisexual Identities in Families*. Oxford University Press, New York

Rosenbloom R (ed.) 1996 *Unspoken Rules: Sexual Orientation and Women's Human Rights*. Cassell, London

Solarz A L (ed.) 1999 *Lesbian Health: Current Assessment and Directions for the Future*. National Academy Press, Washington, DC

Weston K 1991 *Families We Choose: Lesbians, Gays, Kinship*. Columbia University Press, New York

M. V. L. Badgett

Lesion and Behavior Approaches in Neuroscience

One of the earliest and most frequently employed experimental research methods used to investigate how the brain is involved in controlling behavior and behavioral processes involves ablation (removing) or lesion (destruction or functional disruption) of brain structures or areas. While considerable information has been obtained regarding brain-behavior relations from clinical cases where the damage is a result of either head injury, tumor, vascular problems, brain surgery, or disease (see *Cognitive Neuropsychology, Methodology of*), the nature and extent of the damaged and affected areas in humans are difficult to determine and often are not known. By employing laboratory animals and experimental lesion approaches, it is possible to obtain more precise and localized damage to the brain. Further, with the use of animal-testing techniques any changes in behavior can be carefully determined, the brains can then be obtained for detailed histological analysis, and the actual damage to the brain can be subsequently correlated with changes in behavior. Thus, by using the lesion approach together with experimental animals to study brain function, meaningful conclusions can be drawn regarding the relationship that exists between specific areas of the brain and how they are involved in behavior (e.g., localization of function).

1. *Localization of Function*

Issues surrounding the extent to which control of various behaviors is localized in specific brain areas have been around for many centuries (see Corse 1991). It was the work of Franz Josef Gall (1758–1828) and Johann Casper Spurzheim (1776–1832) that established the theory of phrenology, i.e., the theory that particular parts of the brain controlled separate mental faculties. Specifically, these early investigators proposed that personality types are related to bumps located in specific places on the skull. Phrenological theory was spelled out in sufficient detail that some components of the theory could be tested experimentally, and not many years passed before Pierre Flourens (1794–1867) developed the technique of damaging the brains of animals to study the changes produced in their behavior. The result was not only the refutation of phrenology, but justification for the use of the experimental lesion method in studying and attempting to understand brain function. Localization of function continued to receive considerable attention during the nineteenth century with the discovery by Paul Broca (1824–1880) that a discrete area of damaged cortex interfered with speech, and the demonstration by Gustav Theodor Fritsch (1838–1929) and Eduard Hitzig (1838–1907) that movements of particular parts of the body could be elicited by applying electrical stimulation to specific regions of the animal and human brain.

One of the most systematic early attempts to use a lesion approach to determine where in the brain a complex behavioral process is controlled can be found in the research of Karl Lashley (1890–1958). Lashley devoted a lifetime of research to what he described as 'the search for the engram,' that is, the physical representation of learning or the site in the brain where the trace of the memory of a learned thing is stored. The approach he used involved the training and testing of rats to perform tasks of various levels of difficulty, and ablating or using knife cuts to various parts of the cortex either before or after learning. Rather than finding that memory for learned tasks is localized in specific areas, Lashley found that the amount of damage to the cortex was more important than which part was removed (Lashley 1950).

2. *Lesion Techniques and Procedures*

A variety of techniques and procedures have been used to produce selective damage to the brain. In studying the functions of the cortex, early researchers removed large areas on the external surface of the brain by first removing a flap of skull, directly exposing the brain, and then excising brain tissue with a knife or with vacuum suction. A similar technique is often used in current research, especially research involving nonhuman primates. While this approach is feasible for the cortex, damaging structures and/or areas that are deep within the brain creates special problems. Systematic exploration of subcortical areas had to wait until 1908, when Sir Victor Horsley and Robert H. Clarke developed the stereotaxic instrument, a device that permits the precise positioning of electrodes deep within the brain. Thus, by carefully implanting electrodes in known locations, investigators could for the first time damage, stimulate, or record from areas below the external surface of the brain. One lesion technique that is still used in research involves passing direct current through an electrode that is insulated except at the tip, thus creating a localized area of

damaged tissue. A similar technique involves electrocauterization using high frequency current (e.g., radiofrequency).

A major problem associated with the use of conventional lesions (aspiration, electrolytic, radiofrequency) is that the resulting damage is nonspecific. Thus, there is frequently loss of cells in adjacent areas, axons that pass through the area of the lesion are interrupted, and the vasculature can be compromised. For this reason, there has been an increase in recent years in the use of intracerebral injections of neurotoxins. By employing stereotaxically placed focal injections, and procedures that minimize the spread of the toxin to adjacent areas, one can limit the damage to the intended structure. Further, most neurotoxins do not damage fibers of passage, afferents that terminate in the area are minimally affected, and the vasculature is spared. Since the resulting damage is specific, the investigator is in a better position to attribute changes in behavior to damage to a specific structure.

It is especially helpful that some chemicals are selectively toxic to specific kinds of neurons. For example, 6-hydroxydopamine is a chemical that resembles the catecholamine neurotransmitters dopamine and norepinephrine. Because of this resemblance, when the toxin is injected into an area it is selectively taken up by the axons and the result is death of the cell. Investigators have used this approach to produce an animal model of Parkinson's disease.

It was discovered in 1971 by Olney and his colleagues that glutamate, a putative excitatory neurotransmitter, had an excitatory effect on neurons and that high concentrations caused cell death. Several analogues of glutamate have been found that are especially toxic to specific types of cells. For example, when kainic acid is injected peripherally, the result is behavioral convulsions accompanied by the loss of specific cell populations within the brain. Focal injections of kainic acid directly into some brain structures also result in a loss of cells, not only of neurons at the site of injection, but distant damage that is thought to be mediated by seizure activity. Because one population of cells that is especially sensitive to kainate damage is found in the hippocampus (see *Hippocampus and Related Structures*), the toxin has been used in a number of behavioral experiments concerned with learning and memory.

A number of other glutamate analogues have been found that produce restricted damage when injected directly into brain structures. Foremost among these is ibotenic acid, a toxin extracted from the mushroom *Amanita muscara*. This toxin not only produces discrete and circumscribed lesions at the site of injection, but it does this in most structures without causing seizures and without damaging either axons of passage or afferents that terminate in the area (Jarrard 1991). Because of the axon-sparing properties, both ibotenic acid, and a similar analogue n-methyl-d-aspartate, have been used to produce restricted damage in a number of experiments involving both rodents and monkeys.

A recent approach used to produce brain lesions with greater selectivity involves suicide transport and immunolesioning techniques (Wiley 1992). In both of these techniques cytotoxins are targeted to restricted groups of neurons. Suicide transport involves the axonal retrograde transport of toxic lectins from the site of injection back to the cell body with the resulting destruction of the neuron. Monoclonal antineuronal antibodies are used in immunolesioning techniques to deliver cytotoxins selectively to only those neurons that have surface characteristics recognized by the antibodies. An advantage of immunolesioning is that the toxins not only destroy a specific type of neuron at the site of injection but the toxin is transported in a retrograde direction back to the cell body from the injection site. While these newer approaches offer considerable promise, there are a number of potential problems associated with using these extremely toxic chemicals in the laboratory.

Still another lesion approach that has been developed only recently involves the gene-knockout approach. By using biochemical methods to direct a mutation at a particular gene, one can destroy or alter widespread systems of cells or subsystems within cells.

All of the lesion approaches described above produce permanent damage to the brain. A technique that offers considerable promise is the focal injection of drugs that temporarily inactivate brain cells. The most direct way to produce a temporary, reversible brain lesion is to inject a local anesthetic (such as lidocaine) into the area of interest. However, local anesthetics not only inactivate brain cells at the point of injection but they also disrupt the activity of axons passing through the area. Thus, the overall effect is similar to that found with the use of conventional lesion techniques, except that the effect with local anesthetics is reversible. It was reported recently that injection of a newly-developed glutamate receptor antagonist temporarily inactivated the cells at the point of injection without disrupting fibers of passage (Riedel et al. 1999). With the development of new drugs that have temporary effects on specific transmitter systems the use of temporary inactivation as a reversible lesion will no doubt increase.

3. *On Interpreting the Behavioral Effects of Lesions*

Correctly interpreting the effects on behavior of damage to specific brain structures requires an understanding and appreciation of a number of potential problems. Since a change in behavior following damage to the brain can be caused by unwanted effects on sensory system functioning (e.g., vision or olfaction),

the motor system, attention, motivation, etc. it is important to eliminate as many of these explanations as possible by carefully designing the experiment. Over the years, scientists have devised a number of control procedures that take into account these interpretative problems. Thus, the typical lesion experiment includes not only a group of unoperated animals to serve as controls, but also a group of animals that have sham lesions, e.g., that undergo the same surgical intervention as animals in the experimental lesion group except the structure of interest is not damaged. If it is observed in subsequent testing that the behavior of the animals with the brain lesion is different from that of the sham-operated and control animals, then one can be more confident that the structure of interest caused the observed behavioral deficit.

In order to eliminate other possible explanations for an observed behavioral change following brain damage, investigators often employ an experimental design to test for double dissociation of function. In this type of design lesion 1 impairs behavior A but not behavior B, while lesion 2 (in a different region of the brain) impairs behavior B but not behavior A. An example of a double dissociation is when lesion 1 impairs an animal in learning a spatial task (say, learning to swim to a safe platform consistently located just under the surface of the water in a circular tank) but has no effect on learning to discriminate between two visual stimuli, whereas, lesion 2 impairs discrimination learning but has no effect on learning the location of the safe platform. With this type of finding, it can be concluded that the impairment in learning the spatial task is not caused by changes in attention, visual processes, motor performance, motivation, or several other behavioral processes.

Interpretation of lesion studies is also complicated by the fact that many regions of the brain are interconnected, so that damage in one area may result in morphological, biochemical, and/or physiological changes in areas of the brain not damaged directly by the lesion. Further, following damage to the brain there is often sprouting (or new growth), sparing of function, spontaneous recovery, and indirect effects on other behaviors (see Kolb and Whishaw 1996) that can complicate any interpretation.

While there are potential problems in interpreting the results of lesion studies, there are several advantages that this approach has over other available procedures. First, by developing animal models using lesions, one can attempt to understand better the nature and extent of the brain damage found with various diseases and injury in humans. Second, the specifics of the lesion can be determined by careful histological study, thus, there is a permanent record of the actual brain damage. Third, by using different procedures to study the brain (e.g., recording, functional MRI) one can often determine that several different regions seem to be involved in a particular behavior but the question arises regarding whether the structures operate serially or whether they provide separate inputs to a common structure. In this case, lesion studies are required to determine which of these structures are actually necessary for the behavior in question.

4. *Behavioral Assay of Brain Function*

Since the brain is designed to produce and control behavior, a careful analysis of changes in behavior following brain lesions is the ultimate assay of brain function. Early investigators, using lesion techniques and behavioral testing procedures available at the time, made important discoveries regarding regions of the brain that are involved in controlling such important behaviors as emotions, motivation, and learning and memory. For example, it was discovered in 1939 that bilateral ablation of the anterior temporal lobes in monkeys led to dramatic behavioral changes where previously wild and aggressive monkeys failed to display normal fears and anxieties. It is now known that the amygdala (see *Amygdala (Amygdaloid Complex)*) plays a crucial role in such emotions, especially fear and anxiety. By using more sophisticated lesion techniques it has been possible to identify the underlying brain circuitry (see *Fear: Psychological and Neural Aspects*; *Anxiety and Fear, Neural Basis of*). In 1940 Hetherington and Ranson showed that destruction of a region of the hypothalamus (see *Hypothalamus*) using electrolytic lesions produced animals that ate voraciously until they become extremely obese. With the use of selective lesion techniques, it is now known that the hypothalamus is not as crucially involved in the control of eating as these early findings indicated (see *Hunger and Eating, Neural Basis of*; *Appetite Regulation*).

The lesion approach has been used frequently to attack the most basic and challenging problem in the neurosciences, e.g., an understanding of the neural bases of learning and memory. Since the early research of Karl Lashley and 'the search for the engram,' the development of more sophisticated lesion and behavioral testing procedures has permitted major advances in this important area of research. In a recent search for the location of the engram, Richard F. Thompson and colleagues (Thompson and Krupa 1994) used a simple behavioral task (*Eyelid Classical Conditioning*) and a combination of recording and lesion techniques (both conventional lesions and temporary inactivation). By determining which structures changed activity during learning, then selectively damaging or temporarily inactivating inputs and outputs of the suspected crucial structure, the investigators were able to show that the learning (engram) occurred in a single nucleus within the cerebellum (see *Cerebellum: Associative Learning*).

Following a report in 1957 of a case of severe anterograde amnesia (e.g., inability to new material

and events) following ablation of structures deep within the medial temporal lobes, investigators attempted to understand the neural bases by developing animal models (monkeys, rats). Using monkeys, it was found that aspiration of medial temporal structures resulted in a severe memory problem that was similar in many ways to that found in the clinical case. The structure that was suspected of being responsible for the deficit in both human and monkey was the hippocampus. Research using the rat model generally failed to provide support for a global memory problem that encompassed all types of material, especially when an axon-sparing lesion technique was employed to remove the hippocampus (Jarrard 1993). Subs-equent research with monkeys employing more selective lesions has indicated that the major impairment that was attributed to the hippocampus is, in fact, a result of damage to cortical areas that are adjacent to the hippocampus that were inadvertently damaged during surgery.

5. *Future Directions*

With the development of new procedures that permit more selective lesions limited to particular structures, types of neurons, or specific proteins, it is now possible to have a level of lesion control that has been lacking in the past. A similar increase in the level of sophistication in the analysis of behavior and understanding behavioral processes has occurred as described in other entries in this volume (see *Fear Conditioning*; *Learning by Occasion Setting*; *Blocking, Neural Basis of*; *Hippocampus and Related Structures*; *Hunger and Eating, Neural Basis of*; *Thirst and Drinking*). While a more complete understanding of brain-behavior relations will necessarily depend on a convergence of results obtained from a variety of approaches used to study the brain, the careful use and interpretation of lesion studies combined with sophisticated behavioral testing procedures will make major contributions to this understanding.

See also: Behavioral Neuroscience; Blocking, Neural Basis of; Brain Damage: Neuropsychological Rehabilitation; Cerebral Cortex: Organization and Function; Cognitive Functions (Normal) and Neuropsychological Deficits, Models of; Memory: Genetic Approaches; Neuropsychological Functioning, Assessment of; Psychophysiology

Bibliography

Corsi P (ed.) 1991 *The Enchanted Loom*. Oxford University Press, Oxford, UK

Jarrard L E 1991 Use of ibotenic acid to selectively lesion brain structures. In: Conn P M (ed.) *Methods in Neuroscience, Volume 7*. Academic Press, San Diego, CA

Jarrard L E 1993 On the role of the hippocampus in learning and memory in the rat. *Behavioral and Neural Biology* **60**: 9–26

Kolb B, Whishaw I Q 1996 *Fundamentals of Human Neuropsychology*, 4th edn. W. H. Freeman, New York

Lashley K S 1950 In search of the engram. *Society of Experimental Biology, Symposium* **4**: 454–82

Riedel G, Micheau J, Lam A G M, Roloff E V L, Martin S J, Bridge H, de Hoz L, Poeschel B, McCulloch J, Morris R G M 1999 Reversible neural inactivation reveals hippocampal participation in several memory processes. *Nature Neuroscience* **2**: 898–905

Thompson R F, Krupa D J 1994 Organization of memory traces in the mammalian brain. *Annual Review of Neuroscience* **17**: 519–49

Wiley R G 1992 Neural lesioning with ribosome-inactivating proteins: Suicide transport and immunolesioning. *Trends in Neurosciences* **15**: 285–90

L. E. Jarrard

Letter and Character Recognition, Cognitive Psychology of

Letters are basic constructional units in alphabetic scripts that are used to form words; in writing these words are entities which are marked by spaces. How has letter recognition been studied? What are the main results obtained and the major models proposed? These questions are considered in the first part of this article. In contrast to those scripts constructed on an alphabetic principle where symbols (i.e., letters) are made to correspond to sounds in speech, other scripts, such as Chinese, have been developed following a logographic principle, in which symbols mainly and directly encode meaning. Characters are distinct orthographic units in Chinese. Although there are notable differences between letters and characters, they share certain crucial properties, among them that both characters and letters have names, are used to form words, and are the first things that users of the corresponding script start to learn when learning how to read and write. The second part of the article addresses how characters are processed in recognition.

1. *Letter Recognition*

Different alphabetic scripts, such as Hebrew and English, make use of different sets of letters. We shall confine the discussion here to the English letters, because a good deal of research is available on them.

1.1 How has Letter Recognition Been Studied and What are the Main Results?

Early research on letter recognition aimed at identifying a set of features used in distinguishing among letters. But what is a feature? A feature (e.g., horizontal straight, closed curve, etc.) can be considered as a

specific component which can appear in combination with other components to form different letters. One way to find a possible feature set for letters is to observe systematically and analyze patterns of responses when asking people to make matching or same–different judgments on serially or simultaneously presented letters. It is typically found that considerable confusions are made between certain pairs of letters such as N and M, M and W, E and F, P and R, and Q and O, presumably because these pairs of letters share certain features (for an example of the features of capital letters and a review of related research see, e.g., Gibson 1969).

A widely used paradigm in studying letter recognition is to ask people to identify letters that are displayed very rapidly (e.g., for 50 ms or less) on a tachistoscope (i.e., a device that allows researchers to show visual stimuli for a short, fixed amount of time). A common finding in the studies using the tachistoscopic procedure is that people are more likely to confuse letters with similar features, such as E and F, than those with distinctive features, such as E and Q (e.g., Rumelhart 1970). This result suggests that people are possibly engaged in such activities as feature detection and feature analysis in letter recognition and that these perceptual activities play an important role in recognizing a letter.

Letter recognition has also been studied by using a visual search task. In Neisser's (1963) classic research, for instance, he asked people to search an array of letters for a particular target and found that search performance for the target (e.g., Z) was slower when the array was made up of letters with features similar to the target (e.g., X and V, consisting of straight lines) than when it consisted of letters with features distinctive from the target (e.g., O and G, containing rounded features). This result confirms the idea that features are used in recognizing letters.

Apart from studying how individual letters are recognized, researchers have also investigated the way in which letters are processed in context. This topic has been investigated since Cattell's study (1886) and has attracted much research interest over several decades for at least three reasons. First, letters are more often seen in a word context than in isolation. Second, there is good evidence to show that different processes are involved in the two situations. Finally, this topic is related to significant theoretical issues, such as the nature of the processes involved in reading and the relationship between the processing of letters and words. Perhaps the most well-known work on investigating letter recognition in context is Reicher's (1969) study, in which he adopted the tachistoscopic procedure previously mentioned to show three types of stimuli: four-letter words (e.g., WORK), four-letter nonwords (OWRK), or single letters (K). Immediately following the display of a stimulus item, a forced-choice recognition test (e.g., - - - $\frac{D}{K}$ or $\frac{D}{K}$) was given to the participants to judge which one of the two alternative letters had appeared in the specified position in the original stimulus. Reicher discovered that letters were recognized better in the context of a word than in the context of a nonword letter string or in no context at all (i.e., when presented individually), indicating that (a) letter recognition can be affected by context-dependent processes, and (b) word identification may not need to be based on letter-by-letter recognition. This *word superiority effect* is well known, for it provides an important, empirical basis in the development of modern models of letter recognition, a topic which will be covered in the next section.

It is worth noting here that the contextual effects in letter perception have also been investigated by researchers using a letter detection paradigm in situations more closely resembling the conventional reading environment. Healy and Drewnowski (1983), for example, asked people to read a prose passage and to detect every instance of a target letter, such as *t* or *T*. They found that people made many more letter detection errors on words (e.g., *the*) than on misspelled, consequently meaningless words (e.g., *thd*). This *word inferiority effect* not only confirms the idea that the basic encoding units in skilled conventional reading are not likely to be individual letters, it also indicates, in conjunction with the word superiority effect previously mentioned, that the effect of familiar context on letter recognition depends on task and can either be facilitative or inhibitory.

1.2 Major Models Proposed

A classic model for letter recognition is the one proposed by Selfridge (1959), called 'pandemonium.' This model involves four levels of 'demons,' which detect and analyze the properties of a visual input. The stimulus is first scanned and encoded by image demons to produce the mental representation of the stimulus (i.e., an image). The image is then processed by feature demons, each of which searches for a specific feature. Once a feature demon finds the searched-for feature, it calls out to cognitive demons at the next level of the system. Each of these cognitive demons corresponds to one single letter in English, and it pays special attention to a particular set of feature demons. Cognitive demons call out when they hear the yelling from the corresponding features demons. Finally, a decision demon listens to the noises produced by the cognitive demons and makes the final decision regarding the identity of the stimulus based on which cognitive demon is making the loudest noise, indicating that all necessary component features of the cognitive demon have been detected.

Feature-analysis models of letter perception, such as the pandemonium model proposed by Selfridge, were popular in the early days of studying letter recognition. This type of model emphasizes the role of features in recognition but de-emphasizes the pro-

cessing roles played by certain factors, such as context, expectations, and perceptual experience. The recognition and correction of this problem has stimulated the development of contemporary models of letter recognition that incorporate both bottom-up processes, such as feature detection and analysis, and top-down processes, such as expectations and context. A well-known example of such models is the interactive activation model of McClelland and Rumelhart (1981).

The interactive activation model has the following major assumptions. First, there are processing units at three levels: feature, letter, and word. Second, recognition processing occurs simultaneously at different levels (i.e., parallel processing). Third, there can be bidirectional information processing between two neighboring levels (i.e., interactive processing). For instance, top-down influences can be passed along the word level down to the letter level. Fourth, there are both excitatory and inhibitory connections between levels, whereas only inhibitory connections exist within a level.

This type of interactive activation model can easily account for the word superiority effect previously mentioned by allowing interactive activation between levels. In other words, word units that have been activated can increase activation for their constituent letters at the letter level. Consequently, under the tachistoscopic task of Reicher (1969) where only a single letter string or a letter is shown at one time, letters embedded in words can in principle receive feedback from the activated words at the word level and have a better chance to be activated than those in nonword strings or in isolation.

The word inferiority effect aforementioned can also be understood with the multilevel, parallel processing mechanism proposed in the interactive activation model. The word inferiority effect is typically demonstrated in conventional reading situations where many words are shown at once. In order to read for comprehension, readers not only need to identify each individual word in a text, they also need to integrate the meanings of those words to form a coherent understanding. Thus, a good strategy is to start to process the next word in the text as soon as possible once a given word has been recognized even if some of its constituent letters have not been processed to the level of identification. Consequently, letter detection can be disrupted by word context in conventional reading, as demonstrated by Healy and Drewnowski (1983).

2. Character Recognition

Thus far we have examined what has been done over the past decades in understanding the processes involved in letter recognition. An interesting question one might ask at this point concerns whether (or the extent to which) we can apply what we have learned about the processing of sub-word units to other non-alphabetic scripts, such as Chinese. In this section, we will consider how distinctive orthographic units in Chinese (i.e., characters) are processed in recognition.

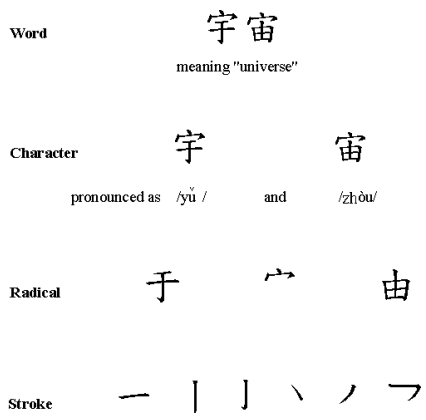

Figure 1
An illustration of constructional units in written Chinese

2.1 Main Properties of Chinese Characters

Each Chinese character is made up of strokes, radicals (i.e., certain combinations of strokes), simple characters (i.e., characters that cannot be further segmented into radicals), or a combination of these components in a constant, square-shaped region (see Fig. 1 for an illustration of constructional units in written Chinese). Hence, different characters can differ in number of strokes (e.g., some characters, such as 一 and 丁, have only one or two strokes, but others, such as 鬱 and 鑿, are formed by more than two dozen strokes), in number of radicals (e.g., 月, meaning *moon*, and 朋, *friend*), or in manner of construction (e.g., 柑, *tangerine*, is made up of two simple characters 木 and 甘 along a horizontal dimension, while 某, *unspecified*, is formed by the same components but in a vertical manner).

Chinese characters are typically monosyllabic. Since there are far fewer syllables (about 1,300 in Mandarin Chinese) than there are characters that are commonly used (about 5,000), the degree of homophony in the language is very high. Moreover, about 80 percent of Chinese characters consist of compound characters made up of two components, one of which (the signific) provides a categorical cue to the meaning of the character; and the other component (the phonetic) may cue in an all-or-none fashion the pronunciation of the character. However, the phonetic parts in these compound characters are themselves characters, having their own pronunciation (thus providing pho-

netic cues) and meanings (typically different from the meanings of the characters in which they are embedded). The cues they provide to pronunciation are neither constant, as some may function as the phonetic part in certain characters and the signific in others, nor reliable, in that less than 30 percent of the phonetic parts provide the correct pronunciation for commonly used characters. In other words, a notable property of the Chinese script is that is has an opaque script-speech relationship.

Furthermore, being constructed following the logographic principle, Chinese characters are basically morphemes, and more than 60 percent of Chinese words are comprised of two or more characters. Since the majority of characters can join others to form multiple-character words with drastically different meanings, the meaning of a character can be highly context-dependent (for relevant research and discussion on this issue see, e.g., Chen 1992). For example, the character 生 can be used to form such distinctive words as 衛生 (health), 花生 (peanut), 生活 (life), 生長 (to grow), 接線生 (operator), and so on. In fact, we can use about 2,500 characters to form some 16,000 different words which can account for about 99 percent of the material in a Chinese corpus of about one million words (Liu et al. 1975).

2.2 Cognitive Processing of Characters

Research on character recognition has had a relatively short history compared to the study of letter recognition. In fact, although some progress has been made over the past two decades, what we have learned about the processing of characters is still very limited in comparison with what we know about letter recognition. The research on the cognitive processing of characters has usually been conducted using methods and techniques similar to those used in English research, aiming to understand the processes underpinning character recognition and related language activities.

Since Chinese characters are constructed of strokes and radicals, an important question for character recognition in Chinese is whether these constructional units are the most crucial perceptual bases used in distinguishing among different characters. Interestingly, there is some evidence to indicate that this may not be the case. For instance, Yeh et al. (1997) asked people to sort characters into groups so that those that looked alike were together. These researchers applied multidimensional scaling to the sorting data and discovered that three configurational properties of characters (i.e., arrangement of components: vertical vs. horizontal, closeness of the configuration: open vs. close, and squarishness of the configuration: square vs. nonsquare) were the most powerful dimensions in explaining the results of sorting. Not among these were the basic patterns of strokes (e.g., horizontal, vertical, oblique, or twisty lines, dots, and so forth), Gibson's (1969) set of critical features, or radicals. In other words, the global configuration of a character seems to play a more salient role in character classification than do local properties of strokes or radicals.

It is worth noting here that although individual radicals may not provide the most effective information in character discrimination, there is evidence to show that they are processed and used in the course of recognition. Indeed, it has been shown that properties of radicals (e.g., their type frequency, their location in a character, and their lexicality—whether they themselves can be used as independent characters) can exert significant effects on performance in such tasks as recognition (deciding whether a stimulus character appeared in a previous display), character decision (deciding whether a stimulus item is indeed a character used in daily life), and naming (for relevant results and/or discussions see, e.g., Li and Chen 1997, Zhou and Marslen-Wilson 1999). Thus, it appears that radicals are segmented from the base characters and their related information is utilized within the other processes resulting in the recognition of the global characters.

Another question that one might ask regarding Chinese characters is the processing role of characters in reading Chinese. There is clear evidence to indicate that individual characters are important encoding and processing units in Chinese. For instance, using Reicher's (1969) tachistoscopic paradigm, Cheng (1981) demonstrated a character superiority effect whereby a radical component could be better detected when it was embedded in a real character than in a pseudocharacter (i.e., one that was constructed following the orthographic rules of the script) or a noncharacter (i.e., one that was formed against the orthographic rules of the script). Moreover, Chen (1984) reported a character inferiority effect in reading Chinese text using a component detection task, by which more detection errors were found when a target component was embedded in genuine characters than in pseudocharacters. These results clearly indicate that characters can be better perceived than radicals and that the fundamental unit in processing written Chinese is the character, and not the radical. This raises another interesting issue concerning the role of the character in reading. Although it is the word, rather than the morpheme, that is commonly considered to be the basic unit of language processing and used as the unit of comprehension analyses, there is evidence to indicate that the character still has a significant role to play even in the case of reading text. For example, using a character detection task, Chen (1987) found that there was a clear word superiority effect in reading Chinese text (i.e., a character was better detected when it was part of a two-character word than when it was part of a two-character nonword). If the word were the most important unit of comprehension analyses, one would expect to find a

word inferiority effect rather than a word superiority effect, because the word-level processing should enjoy a higher priority than the character-level processing, and the word context should create an effect disruptive to the character detection task. This word superiority effect in reading Chinese stands in interesting contrast to both the character inferiority effect in Chinese and the word inferiority effect in English (e.g., Healy and Drewnowski 1983) and suggests that (a) characters serve as important encoding and processing units in reading Chinese text, and (b) characters play a more salient role in reading Chinese than letters do in reading English.

Finally, because different writing systems present linguistic information in distinctive forms, whether similar or different processes are activated and used in processing lexical items in the various orthographies has attracted considerable interest over the years. For instance, it has been argued that phonology plays a crucial role in the process of identifying a visually presented lexical item (e.g., Van Orden 1987). However, most research on this topic has been done using materials in alphabetic scripts. Hence, one way to verify the universality of this phonological hypothesis is to investigate whether phonology significantly influences the recognition of logographic characters (for recent research and discussion on this issue see, e.g., Chen and Zhou 1999). It turns out that there is good evidence to indicate that orthography rather than phonology plays a more crucial role in the recognition of Chinese characters, although phonology can also be activated in the course of recognition (e.g., Shen and Forster 1999). On the other hand, orthographic processing seems to play a less important role for alphabetic materials than phonological processing (e.g., Van Orden 1987). Similar evidence has also been shown in a memory research by Chen and Juola (1982) where Chinese and English were directly compared. A therefore conceivable hypothesis is that specific properties of a given script can affect related recognition and memory processes.

2.3 Research Models Proposed

At present, models of Chinese character recognition are quite scarce. Nevertheless, there are a few such models available. The existing models typically have structures similar to those developed from English materials, but with modifications to capture the properties of the Chinese language. Thus, for instance, there is a hierarchical interactive-activation model, in which there are orthographic and phonological representations structured in levels corresponding to multi-character words, characters, radicals, and strokes (Taft et al. 1999). Similar attempts have been made using parallel-distributed processing models (e.g., Seidenberg and McClelland 1989) where orthographic, phonological, and semantic representations are connected in a triangular form and interact continuously (e.g., Zhou and Marslen-Wilson 1999). However, this type of modeling work is usually not explicit and should only be considered as preliminary for the time being, since the systematic study of Chinese character processing has only just begun.

3. Conclusions

The results obtained from the cognitive studies of character recognition complement the letter recognition studies that have probed the perceptual and cognitive mechanisms that are involved in the processing of written English. Together these studies on distinctively different scripts can provide valuable insights into the formation of a genuinely comprehensive theory of human language processing in general, and the cognitive processing of subword units in particular. Although there has been in recent years an upsurge of interest in conducting cognitive research across different scripts, much work remains to be done. Identifying universal principals and script-specific processes associated with the reading of different scripts should continue to be a challenge to researchers who are interested in the psychology of language in the new century and beyond.

See also: Reading, Neural Basis of; Reading Non-alphabetic Scripts; Reading Skills; Speed Reading; Writing, Evolution of; Writing Systems; Writing Systems, Psychology of

Bibliography

Cattell J M 1886 The time it takes to see and name objects. *Mind* **11**: 63–5

Chen H-C 1984 Detecting radical component of Chinese characters in visual reading. *Chinese Journal of Psychology* **26**: 29–34

Chen H-C 1987 Character detection in reading Chinese: Effects of context and display format. *Chinese Journal of Psychology* **29**: 45–50

Chen H-C 1992 Reading comprehension in Chinese. In: Chen H-C, Tzeng O J L (eds.) *Language Processing in Chinese*. North-Holland, Amsterdam, pp. 175–205

Chen H-C, Juola J F 1982 Dimensions of lexical coding in Chinese and English. *Memory & Cognition* **10**: 216–24

Chen H-C, Zhou X (eds.) 1999 *Processing East Asian Languages*. Psychology Press, London

Cheng C-M 1981 Perception of Chinese characters. *Chinese Journal of Psychology* **23**: 137–53

Gibson E J 1969 *Principles of Perceptual Learning and Development*. Prentice-Hall, New York

Healy A F, Drewnowski A 1983 Investigating the boundaries of reading units: Letter detection in misspelled words. *Journal of Experimental Psychology: Human Perception and Performance* **9**: 413–26

Li H, Chen H-C 1997 Processing of radicals in Chinese character recognition. In: Chen H-C (ed.) *Cognitive Processing of Chinese and Related Asian Languages*. The Chinese University Press, Hong Kong, pp. 141–60

Liu I-M, Chuang C J, Wang S C 1975 *Frequency Count of 40,000 Chinese Words*. Liuguo Chubanshe, Taipei, Taiwan

McClelland J L, Rumelhart D E 1981 An interactive activation model of context effects in letter perception: Part 1. An account of basis findings. *Psychological Review* **88**: 375–407

Neisser U 1963 Decision-time without reaction-time: Experiments in visual scanning. *American Journal of Psychology* **76**: 376–85

Reicher G M 1969 Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology* **81**: 275–80

Rumelhart D E 1970 A multicomponent theory of perception of briefly exposed visual displays. *Journal of Mathematical Psychology* **7**: 191–218

Seidenberg M S, McClelland JL 1989 A distributed, developmental model of visual word recognition and naming. *Psychological Review* **96**: 523–68

Selfridge O G 1959 Pandemonium: A paradigm for learning. In: (eds.) *Mechanisation of Thought Processes: Proceedings of the 10th Symposium of the National Physical Laboratory*. HM Stationery Office, London, pp. 511–29

Shen D, Forster KI 1999 Masked phonological priming in reading Chinese words depends on the task. *Language and Cognitive Processes* **14**: 429–59

Taft M, Liu Y, Zhu X 1999 Morphemic processing in reading Chinese. In: Wang J, Inhoff AW, Chen H-C (eds.) *Reading Chinese Script: A Cognitive Analysis*. Lawrence Erlbaum, Mahwah, NJ, pp. 91–113

Van Orden GC 1987 A ROWS is a ROSE: Spelling, sound and reading. *Memory & Cognition* **15**: 181–98

Yeh S-L, Li JL, Chen I-P 1997 The perceptual dimensions underlying the classification of the shapes of Chinese characters. *Chinese Journal of Psychology* **39**: 47–74

Zhou X, Marslen-Wilson W 1999 Sublexical processing in reading Chinese. In: Wang J, Inhoff AW, Chen H-C (eds.) *Reading Chinese Script: A Cognitive Analysis*. Lawrence Erlbaum, Mahwah, NJ, pp. 37–63

H.-C. Chen

Lewin, Kurt (1890–1947)

1. Biography

Kurt Lewin was born on September 9, 1890 in Mogilno, a small town in the (then Prussian) province of Posen, now Poland. He was the second of the three children of Leopold and Recha Lewin who owned a general store and a small farm. Kurt was sent to a *Gymnasium*, a secondary school, in Posen until, in 1905, the whole family moved to Berlin, as did many ascending Jewish families living in the eastern provinces of the Reich and in Eastern Europe.

In Berlin, Kurt Lewin attended the humanistic *Kaiserin-Augusta-Gymnasium*. In 1909 he took up university studies of medicine and biology in Freiburg and Munich, but in 1910 he was back in Berlin to take courses in philosophy and science. His special interest soon became philosophy of science, in which field the (neo-Kantian) philosopher Ernst Cassirer and the psychologist Carl Stumpf were most influential with regard to Lewin's intellectual development. Stumpf, the director of the Psychological Institute and the patron of what was to become the Berlin school of gestalt psychology (see *Köhler, Wolfgang (1887–1967)*), became Lewin's doctoral supervisor.

Since in 1914 Lewin volunteered for the army and soon did service on the Western Front, he was awarded his doctoral degree as a soldier in 1916. While on furlough from the front in 1917, he married Maria Landsberg, a fellow student from prewar days. The same year, 1917, was also the year of Lewin's first two publications, which, in their dissimilarity, indicate the scope of Lewin's interests: a paper on mental activities in the inhibition of processes of volition, based on his doctoral thesis, and an almost phenomenological paper on the 'war landscape,' foreshadowing Lewin's later involvement in ecological problems.

Also predictive for Lewin's further intellectual as well as social development were his extracurricular activities and associations as a student, as from early on he had engaged himself in socialist student groups which took an active interest in the social and, mainly, educational problems of workers in agriculture and industry. From this enduring interest resulted two early publications in 'applied psychology,' one on rationalization in agriculture, one on the socialization of the Taylor system. Many later publications followed as a result of his and other Jewish psychologists' forced emigration to the USA.

Before the war ended Lewin was severely wounded and hospitalized. Decorated with the Iron Cross and promoted to the rank of lieutenant, he was discharged from military service.

In the following years, Lewin, who had returned to the Psychological Institute of Berlin University, and received his *Habilitation* in 1921, became assistant in the new department for applied psychology, a position which he held until his emigration in 1933. In 1927 he was appointed *außerordentlicher Professor*—a title, not a tenured position.

Although in close spatial and intellectual proximity with Köhler's group of gestalt psychologists, Lewin, who concentrated on studies of will, affect, and action, soon gathered a group of young talented students around him, 'whose background was more often than not female, Jewish, and eastern European' (Danziger 1990, p. 173). Some of Lewin's Berlin women students gained international recognition by their association with Lewin, and some shared his later fate as refugees.

But it is fair to say that Lewin's rapidly growing scientific reputation was also due to the works of his doctoral students, which Lewin represented at conventions and in journals, preferably in his own series in the journal *Psychologische Forschung*.

An important date in Lewin's vita was 1929. In this year Lewin, who had separated and divorced from his wife Maria, remarried. His second wife, Gertrud, had been a friend of the family for several years. Professionally, 1929 was the year for which he received an invitation to attend the ninth International Congress of Psychology at Yale. This was shortly after J. F. Brown, an American who had studied with Lewin, had published a paper on 'The methods of Professor Lewin' in the 1929 *Psychological Review*, which had attracted the attention of US psychologists to the work of Lewin's Berlin group. In Yale Lewin impressed his audience with a short paper on environmental forces (read in German) and an equally short, but self-made motion picture of an eighteen-month-old girl trying to sit on a stone for the first time in her life—one of the many films that Lewin found useful for the graphic representation of the life space (see below). According to G. W. Allport, who had witnessed Lewin's performance in Yale, 'This ingenious film was decisive in forcing some American psychologists to revise their own theories of the nature of intelligent behavior and of learning' (quoted in Marrow 1969 50f.). It was his first encounter with a US audience that resulted in Lewin being invited as visiting professor by Stanford University in 1932.

It turned out that after this stay in Stanford Lewin would only briefly return to Berlin, as in January 1933 the Nazis seized power in Germany. On his return trip from California via Japan and the USSR Lewin decided that he and his family could not live in Nazi Germany. Fortunately, shortly after his return he received a cable informing him that Cornell was inviting him to join the faculty staff, which he gladly accepted—just in time.

From 1933 to 1935 Lewin then worked at the Cornell School of Home Economics. In 1935 he was offered a position at the Iowa University Child Welfare Research Station. For eight years he worked in Iowa as a professor of child psychology. It was in these years that Lewin taught excellent students who became his disciples and, later, colleagues; for example, Barker, Bavelas, Cartwright, Festinger, French, Lippitt, White, and Zander. It was also during this period that Lewin's research interests shifted toward social psychological problems. In 1937 he became one of the founders of the Society for the Psychological Study of Social Issues, whose president he became in 1941.

This social engagement reflects Lewin's increasing concern with the interaction between research and practice which led to his conception of 'action research,' and to the idea and foundation of the Research Center for Group Dynamics at MIT in 1945, whose main research task was to be 'the development of scientific methods studying and changing group life and the development of concepts and theories of Group Dynamics' (quoted in Patnoe 1988, p. 9).

This institution framed the final period of Lewin's life. In 1947, in the middle of a rather hectic spell of activity, he died of a heart attack, on February 12 in Boston, at the age of 56. In the same year the Research Center moved to the University of Michigan, Ann Arbor, where it still exists as a prominent place of social research.

2. *Research Topics and Fields*

Social interaction, mainly in small face-to-face groups, was such a driving force in Lewin's professional life that it is difficult, often impossible, to single out and identify his strictly individual contribution to what the Berlin or Cornell or Iowa or MIT group produced. Conversely, psychological knowledge today owes a great deal to names such as Barker, Cartwright, Dembo, Festinger, French, Hoppe, Karsten, Kelley, Lippitt, Ovsiankina, Pepitone, Thibaut, Zeigarnik (an alphabetical, hence arbitrary, order of students and colleagues who worked with Lewin). From their oeuvre, whether published individually or as co-authors, whether during Lewin's lifetime or thereafter, it would be difficult to do an auctorial analysis of variance. But it would be futile, too, since Lewin, at least, believed firmly in the creativity of dialogical thinking. Hence, Lewin is not only an author's name, it is also a brand name for products of the interaction between Kurt Lewin and his associates.

If one is looking for a name designating the unity of products from Lewin's groups the choice, after all, would be 'field theory,' which was also Lewin's final preference although for many years the favorite was 'topology.' It has been stated frequently but is useful to repeat that 'field theory is probably best characterized as a method: namely a method of analyzing causal relations and of building scientific constructs' (Lewin 1997).

The conceptual tools of this metatheory and methodology (rather than theory and method) are found mainly in two monographs (Lewin 1936, 1938) and two anthologies (Lewin 1997, 1999). Lewin adopted the philosophical foundation of Ernst Cassirer's (1953) distinction between 'substance' and 'function,' his theory of a historical transition from 'thing-concepts' to 'relation- (or function-) concepts' and from an abstraction- to a construction-mode of concept formation. Lewin applied this conception to psychology in his 1931 paper on the transition from an Aristotelian to a Galileian mode of thought (in Lewin 1935). For him psychology deals with relations or functions (rather than thing-like entities) which are basically dynamic i.e., forces.

In accordance with gestalt theoretical principles, every individual psychological instance (experience or

behavior) is the result of an interaction of several features of a given whole or situation; ultimately, each behavior is determined by the total situation in which it occurs. For Lewin the total situation was the 'life space' organized by interdependent forces. In other words, the life space is the 'totality of facts which determine the behavior (B) of an individual at a certain moment. The life space (L) represents the totality of possible events. The life space includes the person (P) and the environment (E). '$B = f(L) = f(P, E)$' (Lewin 1936, p. 216).

It is this formula, which binds person and environment inextricably together, which has become a kind of shorthand for field theory: It is important to note that, owing to the principle of interdependence, neither person nor environment is granted independence. The field, which is structured into regions, paths, and barriers, is basically a field of forces (tendencies to change or to resist change). Some regions are attractive to the person (they have a positive 'valence'). Others are repulsive (they have a negative 'valence'). The person is also considered to be a differentiated region with similar structural and dynamic properties.

These formal characteristics of the life space may be sufficient to explain why Lewinian psychology, whether of learning, of memory, of decision, of action, or of social interaction and relations, is intrinsically dynamic and deals with tension systems (as in anger, aggression, regression, substitution, flight into irreality, etc.), with forces affecting the level of aspiration, conflict, decisions, pressure, resistance, and with styles of leadership or the social-emotional atmosphere of small groups. It was mainly the latter type of studies, their designation (of 'authoritarian' vs. 'democratic' vs. 'laissez-faire' styles of leadership), and, last but not least, the fact that they were published in the early war years, that attracted wide and lively public interest, curiosity, and critique. From these studies, in which things 'happened' that were not designed to happen (such as 'laissez-faire' in lieu of a 'democratic' style), moving toward the use of research for policies of social change, what Lewin has called 'action research,' was but a small, though consequent and innovative step. It was final confirmation for the interdependence that Lewin had postulated between theory and practice, basic research and applied science.

3. *Assessment*

A critical assessment of what in the first five decades after Lewin's death has variously been called the Lewin tradition, legacy, or heritage is difficult because this is a multidimensional and multiperspectival task.

It is multidimensional because of the many research fields in which Lewin was engaged, beginning with the general experimental psychology of learning, memory, affect, motivation, and action, through child and personality psychology, to social, mainly small-group psychology, education, and social change, as well as, from the start, metatheory and methodology. It is multiperspectival since, depending on which experts or resources are consulted at any given time, Lewin's influence on present psychology will be differently evaluated.

It may be easy to neglect the temporal or epochal factor. While, in the first two decades after his death, Lewin's or 'Lewinian' theory could rightly be named a 'contemporary systematic framework,' and 'field theory' occupied sizeable chapters in handbooks of social psychology, this situation had begun to change in the 1980s to such an extent that quite a few of the centennial retrospectives due in 1990 were celebrations with reservations (cf. Graumann 1992, Schönpflug 1992, Wheelan et al. 1990). Field theory was no longer one of the grand theories of (social) psychology (if indeed there were any great ones left). Topology and vector psychology were going out of fashion. Progress in mathematical psychology no longer included topology and hodology as Lewin (1936) had envisaged and practiced them. If Lewin identified himself and has been identified with topology and vector psychology, his significance for contemporary psychology has noticeably decreased. None of his influential disciples, who all remember Lewin drawing 'bathtubs and eggs' in class, has seriously pursued topology as a theoretical and methodological system. But, as outlined above, many of the 'psychological concepts' listed in Lewin's 'Topology' of 1936 have not only survived but become common psychological knowledge.

What is more important for Lewin's heritage and his presence in contemporary psychology is the diversity with which his *oeuvre* has been handed down by his disciples and followers. To name but two traditions which may both rightly be called Lewinian, but in fact have little in common, there is, on the one hand, the prominent group of experimental social psychologists, going back to Lewin's Iowa period (1935–44) and the beginnings of the MIT Research Center for Group Dynamics through Lewin's death in 1947. On the other hand, there is the tradition in which group dynamics and action research have been seen essentially as means of changing human relations, individuals, groups, and societies. The journal *Human Relations*, founded by Lewin in cooperation with the Tavistock Clinic in London, has become an outlet for publications on applied group dynamics from which so-called 'sensitivity training' and resulting 'encounter groups' have gained the broadest notoriety by growing into a movement (cf. Back 1973).

But even if one stays within Lewinian experimental social psychology, the variance of derivations from Lewin's original ideas is considerable. If one looks for the common element, most of Lewin's senior disciples testify that it was Lewin's 'style' and the 'atmosphere' he created around him—a thoroughly Lewinian re-

mark since 'Lewinian style included an appreciation and sensitivity to context—the interplay of person with environment' (Patnoe 1988, p. 88, cf. Graumann 1992). What Patnoe (1988, p. 15) concluded from the testimonies of Lewin's contemporaries, namely that 'Lewin created his own microculture,' has been taken by historiographers of psychology as an enduring characteristic of Lewin's research style. Mitchell Ash (1995), who as a historian of gestalt psychology (and of other émigré psychologists after 1933) is interested in the role of milieu changes, states that Lewin, who had already established 'a scientific microculture around him' in Berlin (Ash in Schönpflug 1992, p. 20), accomplished the same in Iowa. The (sometimes very intense) interpersonal interaction in Lewin's research group was not just different from the methods of the (other) gestalt theorists (Ash 1995, p. 272). For Kurt Danziger, who has specialized in the history of psychological research, Lewin developed an investigative practice that 'departed fundamentally from the natural scientific model of experimentation' (Danziger 1990, p. 175). Considering that Lewin's Berlin experiments were not in social but in personality psychology, it is a striking innovation that even here (and in principle) the 'investigative situation is recognized as intrinsically social, and in research on personality it must form part of the object of investigation' (Danziger 1990, p. 175).

In anticipation of and in accordance with Lewin's later theorizing, the experimental subject is not a decontextualized individual, but a person-in-a-situation, and an important interacting part of the situation is the person of the experimenter. It is here that the Galileian thought and postulate which gave birth to field theory have become instrumental in an innovative, psychologically autonomous research practice—a legacy of Lewin's that psychologists still have to appropriate.

See also: Behavior Analysis, Applied; Contextual Studies: Methodology; Heider, Fritz (1896–1988); Human Development, Bioecological Theory of; Interactionism and Personality; Köhler, Wolfgang (1887–1967); Psychology: Historical and Cultural Perspectives; Psychology: Overview; Situation Model: Psychological; Social Psychology; Social Psychology: Research Methods; Social Psychology, Theories of

Bibliography

Ash M 1995 *Gestalt Psychology in German Culture, 1890–1969: Holism and the Quest for Objectivity*. Cambridge University Press, Cambridge, UK

Back K W 1973 *Beyond Words: The Story of Sensitivity Training and the Encounter Movement*. Penguin, Baltimore, MD

Cassirer E 1953 *Substance and Function*. Dover, New York

Danziger K 1990 *Constructing the Subject. Historical Origins of Psychological Research*. Cambridge University Press, Cambridge, UK

Graumann C F 1992 Lewin 1990. In: Schönpflug W (ed.) *Kurt Lewin. Leben, Werk und Umfeld*. Lang, Frankfurt, Germany, pp. 223–30

Lewin K 1935 *A Dynamic Theory of Personality*, 1st edn. McGraw-Hill, New York

Lewin K 1936 *Principles of Topological Psychology*, 1st edn. McGraw-Hill, New York

Lewin K 1938 *The Conceptual Representation and the Measurement of Psychological Forces*. Duke University Press, Durham, NC

Lewin K 1981 *Kurt Lewin Werkausgabe* [ed. Graumann C F]. Huber, Switzerland

Lewin K 1997 *Resolving Social Conflicts and Field Theory in Social Science*. American Psychological Association, Washington, DC

Lewin K 1999 *The Complete Social Scientist*, 1st edn. [ed. M. Gold]. American Psychological Association, Washington, DC

Marrow A J 1969 *The Practical Theorist. The Life and the Work of Kurt Lewin*. Basic Books, New York

Patnoe S 1988 *A Narrative History of Experimental Social Psychology: The Lewin Tradition*. Springer Verlag, New York

Schönpflug W (ed.) 1992 *Kurt Lewin. Leben, Werk und Umfeld*. Lang, Frankfurt, Germany

Stivers E, Wheelan S A (eds.) 1986 *The Lewin Legacy-Field Theory in Current Practice*. Springer Verlag, Berlin

Wheelan S A, Pepitone E A, Abt V (eds.) 1990 *Advances in Field Theory*. Sage, London

C.-F. Graumann

Lex Mercatoria

Lex Mercatoria, often translated as law merchant, consists of rules that are private, autonomous, international, a-national, or transnational. Lex Mercatoria is not enacted by state courts or a state legislative body. Its practical impact, however, has been undisputed since ancient Roman times and is discussed now in relation to cyber law.

1. Historical Development and Current Definitions

In Roman law, Lex Mercatoria was part of *ius gentium*, that is, a kind of international law more flexible than the Roman *ius civile* (civil law). Lex Mercatoria comprised national customary law in international trade. Historically, it is the major root of the consensual contract (without special form). However, it is disputed whether Lex Mercatoria as part of *ius gentium* offers a precedent for Lex Mercatoria as a legal system apart from national legal systems (De Ly 1992).

The medieval *ius commune* had for centuries constituted a legal framework in Europe. Legal scholars cultivated a common understanding based on Justinian's *Corpus Iuris* despite the division of medieval Europe into many more distinct political entities

than today, all with their various local laws. Similarly, medieval Lex Mercatoria constituted a cosmopolitan mercantile law based upon transnational customs and deeds copied for similar transactions. Disputes were settled by market tribunals or special courts. The English *Carta Mercatoria* of 1303 provided certain foreign merchants with a fast dispute-settlement mechanism and application of the law merchant (as opposed to feudal law or city law) (De Ly 1992). The English admiralty courts relied on the best available sources, be they domestic or foreign, in elaborating a worldwide maritime law (Juenger 1995).

In the English common law tradition, common law courts in the seventeenth century started incorporating customary rules developed by traders and usually applied in arbitration or special courts. 'Law Merchant' was defined as

> a number of usages, each of which exist among merchants and persons engaged in mercantile transactions, not only in one particular country, but throughout the civilized world, and each which has acquired notoriety, not only amongst those persons, but also in the mercantile world at large, that courts of this country will take judicial notice of it (Lethulier's Case (1692) 2 Salk 443 (per *Holt, CJ*)).

This 'Law Merchant' was less rigid and formal than the common law. As the common law courts became the forum for commercial litigation, 'Law Merchant' lost its international character and became national commercial law. Later codifications of commercial law, as for instance the English Sale of Goods Act 1893 and the American Uniform Commercial Code (UCC), are to a large extent restatements of law merchant principles as represented in common law. Today, 'Law Merchant' is still used to refer to national common law relating to commercial transactions.

Modern usage understands Lex Mercatoria as systems of standardized clauses and general legal principles in international trade and commerce. In order to set it off from law merchant incorporated in national common law some authors prefer 'New' or 'Modern Lex Mercatoria' or 'New Law Merchant' and similar terminology when referring to modern (informal) codifications of transnational commercial law.

Systems of standardized clauses are developed by a number of intergovernmental and nongovernmental organizations through extensive comparative efforts. Such rules have often served as a model for national lawmaking or international conventions. An outstanding example of a formulating agency is the International Institute for the Unification of Private Law (*Institut international pour l'unification du droit privé*, UNIDROIT), set up in 1926 as an auxiliary organ of the League of Nations. UNIDROIT is now an independent intergovernmental organization with 58 member states. Its purpose is to examine ways of harmonizing the private law of states, and to prepare gradually for the adoption by states of uniform rules of private law. Many of UNIDROIT's drafts have resulted in international instruments and Conventions, like the Convention on a Uniform Law on the Formation of Contracts for the International Sale of Goods adopted at a diplomatic Conference of the member states in The Hague in 1964. The UN Commission on International Trade Law (UNCITRAL), formed by the United Nations General Assembly in 1966, has similar topics on its agenda. A well known example is the Vienna Convention on the International Sale of Goods (CISG). Another system of clauses referred to worldwide are the INCOTERMS, specific trade terms promulgated by the International Chamber of Commerce (ICC).

Lex Mercatoria also refers to unwritten or, at least, not systematically collected principles and clauses which are considered international trade law, usage or custom (Goode 1997). Such rules are invoked in arbitration and applied by arbitration tribunals, for example, the ICC. This kind of Lex Mercatoria is compiled and discussed in legal literature. UNIDROIT prepared a well researched collection of contract law used in international trade, the 'UNIDROIT Principles of International Commercial Contracts' first published in 1994. These principles are intended to enunciate rules which are common to many existing legal systems and are selected to accommodate the special requirements of international trade. The UNIDROIT principles do not claim binding force; they are not intended as uniform law or for an international convention and will be applied in practice by reason of their persuasive value only. In many other fields of law, general principles, guidelines, or best practices are set forth by professional bodies, trade associations, interest groups, legal scholars, governmental, or intergovernmental agencies. In some instances, like the UNIDROIT project, the objective is to establish an unbiased and well-balanced restatement; in other instances, a political or policing impact is intended.

2. *Main Characteristics and Practical Impact*

Based on the historic development and present day definitions, Lex Mercatoria can be characterized as transnational and a-national. In the absence of a 'world legislator,' international trade and commerce have developed very functional and sophisticated rules based on practice and persuasive value. Parties to transnational contracts tend to avoid recourse to domestic law. Relying on well established Lex Mercatoria-clauses, contracting parties have a chance to overcome language barriers and all kinds of biases. Furthermore, they can draw from the condensed experience in their special field of business. When writing a transnational contract, parties or their counsel rarely invoke Lex Mercatoria explicitly. Rather they make use of specific clauses and principles.

Therefore, Lex Mercatoria is not as well established in the terminology of international trade as in its practices.

Usually, substantive references to Lex Mercatoria occur in agreements of arbitration or other means of alternative dispute resolution that avoid state courts. Arbitration rules are often part of Lex Mercatoria, too, like UNCITRAL's, ICSID's (World Bank International Center for the Settlement of Investment Disputes), or ICC's rules on international commercial arbitration. In the case of Scherk vs. Alberto Culver Co (417 US 506), the US Supreme Court pointed out the purpose of such arbitration agreements:

> A contractual provision specifying in advance the form in which disputes shall be litigated and the law to be applied is…an almost indispensable pre-condition to achievement of the orderliness and predictability essential to any International transaction. Furthermore, such a provision obviates the danger that a dispute under the agreement might be submitted to a forum hostile to the interests of one of the parties or unfamiliar with the problem area involved.

This case from 1974 is in line with the 1958 United Nations Convention on the Recognition and Enforcement of Foreign Arbitral Awards, which has been ratified by over 50 countries.

3. *Theoretical Analysis*

In legal theory, it is disputed whether Lex Mercatoria is 'law.' As law, it would form a third category between national and international law. Proponents of the theory of Lex Mercatoria as an independent body of law (mainly French, English, Dutch, and German scholars and practitioners experienced in arbitration) argue on the basis of a functional or sociological concept of law. This view acknowledges the existence of an autonomous body of law governing international commerce without a source in state power: *Il existe un droit dont la source est ailleurs que dans l'Etat* (Kahn 1961).

On the other hand, theorists, especially legal positivists, who see all law based on the powers of the (nation) state, deny the existence of an autonomous Lex Mercatoria. They argue that the absence of jurisdiction cannot be offset by private agreement, for even contracts need national law in order to be enforceable. Constitutionalism, positivism, and the concept of the sovereign state are irreconcilable with rule making by private governance. A contract is only conceivable within a legal order; there is no '*contrat sans loi*'. Additionally, Lex Mercatoria is considered much too vague, difficult to ascertain, and incomplete to qualify as a legal system.

Practically, the dispute is of only limited impact. The theoretical question is resolved whenever model laws or other products of formulating agencies have been incorporated into national law. Then, we have state enacted law applicable in courts and suitable for choice of law clauses. In other cases, when the state is not a signatory state or the respective set of clauses is a private convention, courts can refer to Lex Mercatoria clauses only as incorporation of standard business conditions into the contract at hand and will test the validity of such clauses according to national law. A choice of law clause that calls for the application of Lex Mercatoria in general is certainly very rare. In such a case, the objection of vagueness would be well argued. However, the parties to a contract tend to be much more specific, and the research necessary to deal with a particular Lex Mercatoria clause does not seem more burdensome than the application of foreign law mandated by national conflict of law rules. As already mentioned, arbitration clauses are prevalent in international trade. An arbitration tribunal is free to apply Lex Mercatoria rules and reach its decision on the equities of the case. The validity of arbitration clauses is generally recognized by state courts, especially if the clause inserted is in accordance with the 'Rules of Conciliation and Arbitration of the International Chamber of Commerce,' the UNCITRAL rules on international commercial arbitration, or a similarly well-established procedure.

A third approach is the advancement of guidelines, recommendations, codes of conduct, best practices, etc. with the clear understanding that these instruments are not binding. Sometimes, such sets of rules are called 'soft law.' Nevertheless, the practical impact and even legal relevance of nonbinding principles can be considerable. The OECD Guidelines for Multinational Enterprises, for example, are recommendations addressed by governments to multinational enterprises; they provide voluntary principles and standards for responsible business conduct consistent with applicable laws (Horn 1980). Similarly, the OECD Principles of Corporate Governance are intended to assist member and nonmember governments in their efforts to evaluate and improve the legal, institutional, and regulatory framework for corporate governance in their countries, and to offer guidance and suggestions for stock exchanges, investors, corporations, and other parties that have a role in the process of developing good corporate governance.

Because 'soft law' is not binding, governments and international institutions discuss sensitive policy issues freely in an open manner and have fewer reservations to enter into agreements that are just recommendations (Dryden 1999). Seen from the perspective of the individual, guidelines, recommendations, and best practices leave private autonomy formally untouched. However, the deviation from widely recognized standards has an impact on the standing of the enterprise with respect to business partners or investors. The effect of accepted best practices is described as 'comply or explain.' The informal rule is honored by reasoned comment. Legally, noncompliance with informal standards does not lead to liability. However, the con-

struction of technical legal concepts as, for instance, the business judgment rule, may very well draw upon best practices.

Another important effect of private governance is control of access to certain activities and ensuing ramifications. A stock exchange, organized in a private form (corporation), may require stock corporations to establish audit committees in order to have their stock listed. Then, audit committees will be established even if state corporation law does not require committees as an organizational structure of the board of directors. Legislative activities, as pointed out above, quite often follow suit when well established formulating agencies propose a systematic set of rules.

4. Perspectives

Globalization of commerce, information, and other areas of life brings about challenges similar to those ancient Roman merchants and English tradesmen faced and met with the creation of Lex Mercatoria. In global electronic commerce, the realities of the borderless world of cyberspace runs headlong into geographically delimited national jurisdictions of sovereign states. The attainments of constitutionalism, the nation state and the rule of law are finite in a global market. Private régimes are gaining ground much in the same way Lex Mercatoria established itself as reality, be it legal, commercial or sociological. Self-regulation is found in the Internet as 'netiquette' and other attempts to create an environment of trust and stability that are propitious for e-commerce. The question of global regulation is not moot, though. Few governments would consider self-regulation to be a suitable approach for taxation of electronic business. Given the extreme difficulties of global regulation, the OECD, based on a broad consensus among the member countries, wants to facilitate the development by the private sector of effective market-driven self-regulatory mechanisms. Regulation and self-regulation is not seen as a dichotomy anymore. Rather, a complementary or integrated approach is favored and the dominating issue is the right mix (Dryden 1999).

The historical experience of Lex Mercatoria shows that transnational, nongovernmental régimes can be efficient and productive. Modern Lex Mercatoria includes harmonization of specific fields of law, setting standards by codes of conduct, and developing self-regulation in the new communication environment. The basic structure of the phenomenon that in the absence of an international legislator the interested parties resort to self-regulation, assisted by specialists and transnational associations, is found in other areas, too. In sports, we are used to regulation by transnational nongovernmental associations. World championships and similar events follow elaborate procedures on the basis of rules which are not law in a formal sense that requires a national law maker. Nevertheless, the private governance of a worldwide network of sports associations exercises undisputed jurisdiction over the rules of the various games.

See also: Administrative Law; Business Law; Common Law; Contracts: Legal Perspectives; Globalization: Legal Aspects; International Arbitration; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; Law and Development; Law: Overview; Law, Sociology of; Laws, Conflict of; Legal Positivism; Legal Systems, Classification of; Legal Systems: Private; Organizations and the Law

Bibliography

Bonell M J 1997 *An International Restatement of Contract Law*, 2nd edn. Transitional Publishers, Irvington-on-Hudson, NY

Carbonneau T (ed.) 1990 *Lex Mercatoria and Arbitration*. Transnational Juris Publications, Dobbs Ferry, NY

De Ly F 1992 *International Business Law and Lex Mercatoria*. North-Holland, Amsterdam

Dezalay, Bryant 1995 Merchants of law as moral entrepreneurs. *Law & Society Review* **29**: 12–27

Dryden J 1999 *The Work of the OECD on Electronic Commerce*. http://www.oecd.org/subject/e_commerce/Ottawa_speech.pdf

Goode 1997 Usage and its reception in transnational commercial law. *International and Comparative Law Quarterly* **46**: 1

Horn N 1980 Codes of conduct for MNEs and transnational Lex Mercatoria: An international process of learning and law making. In: Horn N (ed.) *Legal Problems of Codes of Conduct for Multinational Enterprises*. Kluwer, Deventer, The Netherlands

Juenger 1995 American conflicts scholarship and the new law merchant. *Vanderbilt Journal of Transnational Law* **28**: 487

Kahn P 1961 *La vente commerciale internationale*. Paris

Lando 1994 Assessing the role of the UNIDROIT principles in the harmonization of arbitration law. *Tulane Journal of International and Comparative Law* **3**: 129

OECD-material: http://www.oecd.org

Osman F 1992 *Les principes généraux de la lex mercatoria. Contribution à l'étude d'un ordre juridique anational*. Libr. générale de droit et de jurisprudence, Paris

Stein U 1995 *Lex Mercatoria: Realität und Theorie*. V. Klostermann, Frankfurt am Main, Germany

Teubner G (ed.) 1997 *Global Law without a State*. Dartmouth, Aldershot, UK

UNIDROIT-material: http://www.unidroit.org/english/presentation/pres.htm

C. Windbichler

Lexical Access, Cognitive Psychology of

In order to produce and understand language, whether it be spoken or written, an individual must have available some form of cognitive representation of the

words of their language. This pool of mentally stored information is called *lexical memory* or the *mental lexicon*, while *lexical access* refers to the retrieval of such information. Initial research into the mechanisms of lexical access focused primarily on visual word recognition and, indeed, the term 'lexical access' was originally coined in relation to this domain (Forster and Chambers 1973). However, in more recent years, the term has been more commonly applied to language production rather than to language recognition. The reason for this change in the pattern of its usage arises from the fact that it is not a theoretically neutral term and is no longer concordant with the currently dominant theoretical approaches to visual word recognition (see also *Word Recognition, Cognitive Psychology of*). This will be detailed below.

1. Use of the Term 'Lexical Access'

1.1 Accessing a Lexical Entry

The earliest attempts to explain how information is retrieved from lexical memory equated the stored representation with a dictionary entry. Thus, the task of word recognition was seen to involve access to a discrete *lexical entry*, either via a search procedure (e.g., Forster and Chambers 1973) or by activating it to some threshold on the basis of features extracted from the stimulus (e.g., Morton 1969). Thus, for a visually presented word, a match is made between a coded sensory representation of the stimulus and the representation of the form of the word as stored in lexical memory, that is, there is a *sensory-to-lexical* match. This is the process that has been labeled 'lexical access.' Once the match is made, the functional characteristics of the word become available (i.e., its meaning and its grammatical characteristics) because these are associated with the lexical entry, and any processing that makes use of this information is therefore said to be *postaccess* or *postlexical*.

In contrast *preaccess* or *prelexical* processes refer to any processing that takes place prior to the sensory-to-lexical match being achieved, e.g., the conversion of specific font information into some abstract letter code. The topic of prelexical processing is notably germane to the issue of *phonological recoding* (see later).

1.2 Connectionist Models of Lexical Processing

In the 1980s, the advent of connectionist modeling of word recognition processes led to a conceptualization whereby lexical information does not reside in a discretely defined entry.

In *localist* connectionist models (e.g., the Interactive-Activation account of McClelland and Rumelhart 1981), although there may be discrete units of activation that represent the words of the language, there are also units representing subword (i.e., *sublexical*) entities (e.g., letters). Sublexical activation is as integral to the recognition of the word as is lexical activation because there is an interaction between the sublexical and lexical levels in the determination of the output. In such a model, the process of matching the stimulus with a memory representation of the word involves not only the accessing of lexical information, but also sublexical information. Indeed, the whole word need not be represented at all, because its meaning could be activated solely via sublexical units (Taft 1991). So it is somewhat misleading, within this framework, to use the term 'lexical access' to refer to the actual matching process because it may not be based on lexical information, at least not exclusively. However, the term could be appropriately used to refer to the outcome of the matching process, namely the point at which information about the whole word is activated to some criterion of acceptability and is therefore 'accessed.'

In *distributed* connectionist models (e.g., the Parallel Distributed Processing model of Seidenberg and McClelland 1989), the presented word activates a set of input units that produces a pattern of activation in a set of output units (via an intermediate set of hidden units) with no explicit lexical representation (see *Cognition, Distributed*). This again obscures the idea of lexical access as a process of finding a sensory-to-lexical match. Nonetheless, at some point in processing, the system must settle on a particular output as being the most relevant to the input and, because this means that information about the word has become available for response, it could be argued that this is when 'lexical access' has occurred.

So, within connectionist accounts of word recognition, 'lexical access' refers most appropriately to the final outcome of processing rather than to the processing itself. The use of the term has therefore waned, because the central interest of cognitive investigations into word recognition is the nature of the actual processes involved in identifying a word and not the mere fact that the word is recognized. For this reason, the more general term 'lexical processing' tends to be preferred.

1.3 The Scope of the Term 'Lexical'

Even when the term 'processing' replaces the term 'access,' there still remains an issue of the scope of the term 'lexical,' which is simply the adjective for the noun 'word.' A word has a form (both phonological and orthographic), and also a function (both semantic and syntactic). Is the processing of the form of a word sufficient to say that lexical access has occurred, or is access to the meaning and syntactic class of the word required as well? This is rarely, if ever, specified in the literature. Certainly when the semantic (or syntactic)

information about a word participates in processing at the sentence level, this must be occurring 'postaccess,' but it is still debatable whether the actual extraction of this information should be also seen as a 'postaccess' process or part of lexical access itself.

1.4 Language Production

When language is generated, meaning is being translated into form. The term 'lexical access' is still frequently used in relation to language production (where most of the research focuses on speech rather than writing). The continued use of the term perhaps reflects the fact that the interface between the meaning and form of a word is largely arbitrary and, therefore, the processing of the form-level representation (called the *lexeme*) does not involve the gradual increment of activation based on sublexical information. However, theorists in speech production propose that a lexical item consists not only of the lexeme, but also of a functionally defined representation, the *lemma*, that is selected on the basis of the desired meaning (see *Speech Production, Psychology of*). The term 'lexical access' is used to refer to retrieval of the lemma as well as the lexeme, with the main theoretical issue being the relationship between the two (see, e.g., Dell and O'Seaghdha 1991, Levelt 1992).

1.5 A Lemma in Word Recognition?

If the notion of an abstract lexical unit (i.e., the lemma) mediating between meaning and form is accepted for word production (but see Caramazza 1997), then the same should hold for word comprehension. The involvement of such a unit in the word recognition system has only recently been considered (see, e.g., Baayen et al. 1997). A lemma-level would represent units that link form and function and, as such, could be considered the core lexical representations, that is, units that allow access to all the information about a word. In terms of word recognition, then, lexical access occurs when a lemma is sufficiently activated via the form-level representations that are appropriate to the incoming sensory information. Thus, as in word production, lexical processing in word comprehension (be it visual or auditory) can be defined in terms of access to two types of representation, one defined in terms of form and the other in terms of function.

2. Issues in Lexical Processing

For issues related to lexical processing in word production, see *Speech Production, Psychology of*. The following will concentrate solely on issues concerning lexical processing in word recognition. The focus will be on the processing of visually presented words, given the dominance of this domain as a research area. However, issues specific to lexical processing in speech recognition will be also be considered.

2.1 Modality of Access

When a word is visually presented, it is obviously a logical possibility that the form-based representation that is accessed is an orthographically coded one (i.e., one that is compatible with the visual characterization of the word). However, for a long time it has been suggested that when words are to be read, they are recoded in terms of their sound, that is, they are recoded phonologically (see Taft 1991, Chap. 4), even though they do not need to be overtly pronounced. After all, recognizing spoken language is an ecologically more fundamental skill than recognizing written language. The idea is that the form-based lexical match is made in the phonological modality, and this means that the visually presented item must be recoded into its sound prior to accessing any lexical information, i.e., prelexically. This is achievable via a set of rules that convert letters (i.e., graphemic information) into sounds (i.e., phonemic information).

However, there are languages, such as English, for which such *grapheme-to-phoneme conversion* (GPC) rules will not always be successful (e.g., there would be no way of knowing that the EA of GREAT and the EA of TREAT are pronounced differently). Therefore, the phonological representation of an irregular word (such as GREAT) must be generated on the basis of information about the whole word. So, there must be a means of accessing such information other than via a phonological recoding of the visual stimulus. That is, if phonological information about a word could only be accessed via GPC rules, GREAT would always be recoded into the same pronunciation as GREET and therefore confused with that word. Clearly, the form-level access of an irregular word such as GREAT must be achieved on the basis of orthographic information, and the associated phonology of the word can then be made available postlexically.

Saying this, however, assumes that 'lexical access' refers only to form-based access and not to access of meaning or of a lemma. 'Postlexical,' in this context, would not refer to processing that occurs after the meaning is accessed, that is, it is unlikely that the phonology of a visually presented word would only become available after its meaning does. Indeed, to the contrary, it has been proposed that the primary route to meaning is via the phonological representation (see Frost 1998). There is a large body of research that suggests that the pronunciation of a word is activated in the course of reading silently for meaning. However, it still remains controversial whether this entails phonological mediation between orthographic processing of the stimulus and access to its meaning (Frost 1998), or whether meaning is primarily accessed via orthographic information in silent reading and that

automatic activation of phonology occurs independently, perhaps for the purposes of holding the material in working memory (Taft and van Graan 1998).

2.2 Use of Internal Word Structure

Words are more than just a linear concatenation of letters (or phonemes). There is an internal structure whereby some letters go together more closely than others. In the word PICNICKING, for example, the *morphological* structure is PICNIC (stem) and ING (affix), while the stem itself has the *syllabic* structure PIC + NIC. In turn, each syllable has the same vowel plus consonant unit (i.e., IC), called the *rime* or *body*, with a different consonantal *onset* (P and N). There is a growing body of research that suggests that analysis of this internal structure takes place during the course of lexical processing. Studies in a wide range of languages indicate the importance of morphological analysis (see, e.g., Feldman 1995), while the existence and nature of syllabic processing seems to be determined by the characteristics of the language involved (e.g., Cutler et al. 1986). Onset/rime analysis has also been clearly demonstrated, but so far only in English (e.g., Treiman et al. 1995).

In theoretical terms, the importance of such units of analysis means either that they are explicitly represented as sublexical entities within the processing system (see, e.g., Taft 1991) or that they implicitly emerge as a reflection of the statistical characteristics of the relationship that exists between orthography, phonology, and meaning (e.g., Seidenberg and McClelland 1989).

2.3 Lexical Processing in Context

When a word is presented in isolation, the only basis for recognition is the sensory features extracted from the stimulus, i.e., *bottom-up* information. When presented in a sentence context, however, there is the additional possibility of anticipating what the word is on the basis of semantic or syntactic expectations (i.e., *top-down* information), thus reducing the importance of bottom-up processing. So, is it the case that lexical processing qualitatively changes when the word is placed within a sentence context?

Although the recognition of a word (e.g., MIRROR) might be facilitated by the presence of a highly predictive context, like SHE LOOKED AT HERSELF IN THE, research suggests that there is no advantage in having a compatible, but not so predictive, context, such as SHE LOOKED FOR A (Taft 1991, Chap. 3). The most robust effect of context actually appears to be its inhibitory impact on word recognition when the word is incompatible with the context. For example, the time taken to recognize MIRROR is delayed by the immediately prior presentation of SHE ATE THE DELICIOUS. It has therefore been concluded that context has its major influence at a postaccess stage and that its main role is to suppress inappropriately accessed candidates (e.g., Forster 1981).

If a word is being accessed in the same way whether it fits the context or not, the two meanings of an ambiguous word (e.g., PORT) should be accessed in the same way regardless of any contextual bias (e.g., AFTER DINNER HE DRANK A GLASS OF PORT versus THE SHIP CAME INTO PORT). The contextual bias should only come into play by suppressing the inappropriate meaning at a postaccess stage, rather than by preventing the inappropriate meaning being accessed at all. In fact, there have been a number of studies supporting this position, although there may be some impact of the appropriateness of the context at an early stage as well (see, e.g., Simpson 1984).

Note that an ambiguous word has a single representation at the form level, but two representations at the lemma or meaning level. Therefore, if context is said to have its selective effect postlexically rather than prelexically, then 'lexical access' must be used here to refer to the function-based level of access (because there is nothing at the form-based level to select). So, we see that there is potential for confusion in the understanding of lexical access because, in relation to the issue of phonological mediation in silent reading, the terms 'prelexical' and 'postlexical' were used to refer to form-based, rather than function-based access. When using such terminology, then, one needs to be aware of its scope of reference within the context within which it is being used.

2.4 Lexical Processing in Speech Comprehension

There are aspects of lexical processing that are specific to the recognition of spoken words as opposed to written words (see *Speech Perception*). First, while most alphabetically written scripts physically separate each word within a sentence, the speech signal does not provide any reliable indication of word boundaries. Therefore, in the recognition of spoken words there is no guarantee that the sensory information being processed corresponds to a whole word: It may even be that what one initially processes as the beginning of a word is actually the end of the previous word. Unlike in reading, though, the physical signal contains prosodic features (notably patterns of stress) and these have been shown to be important cues in determining word boundaries (e.g., Norris et al. 1995).

A second feature of the speech signal is that its presentation to the sensory system is not under the control of the recipient: the acoustic features of the speech signal are exposed to the auditory system fleetingly and in a fixed temporal order. Therefore, these features must either be processed continuously

on-line or accumulated within a memory buffer for later processing. When reading a text, on the other hand, one has the freedom to move one's eyes to wherever one chooses and it is possible to take in information from anywhere within a word. Furthermore, while processing a word in the foveal region of the eye, it is possible to begin parafoveal processing of the upcoming information, which can potentially influence subsequent foveal processing of that information (see *Eye Movements in Reading*). So, in speech comprehension, the temporal characteristics of the signal are likely to have a major impact on lexical processing and this is indeed a feature of models of spoken word recognition (see, e.g., Marslen-Wilson 1989).

3. *Measuring Lexical Processes*

Again, the primary focus here will be on visual word recognition. For experimental tasks in the areas of language production and speech recognition, see *Speech Production, Psychology of* and *Speech Perception*, respectively.

3.1 *Experimental Tasks*

In order to gauge the processes involved in accessing a word in lexical memory, a task is required that cannot be performed on the basis of prelexically derived information. One commonly used task is where the presented word must be overtly named and the latency to begin the pronunciation is measured. While this *naming task* provides a good measure of the processes involved in generating sound from print, it is not necessarily a good measure of lexical processing because it is possible at least to begin pronouncing many words on the basis of units smaller than the word (i.e., the initial letters).

The task that is the most widely used is one that makes use of the obvious fact that lexical memory only contains words that have been encountered before. The item FLINK could potentially be a word in English, but is actually a *nonword* or *pseudoword*. Unlike an unstructured letter-string such as LKNFI, the only way to know that FLINK is a nonword is to establish that it does not exist in the lexicon. Therefore, if asked to classify items as 'word' or 'nonword,' the only way to do so is to consult lexical memory. So the time taken to classify a word as a word in this *lexical decision task* is taken as a measure of the ease of lexical access. The sort of experiments that are undertaken using this task tend to compare two sets of words that differ in one characteristic while being as similar as possible in all other potentially important characteristics. If the two sets differ in their lexical decision response times, it is concluded that the manipulated characteristic plays a role in lexical processing. For example, two sets of words of the same length will produce different lexical decision times if they differ in their frequency of occurrence in the language and, therefore, any account of lexical processing must consider the role of word frequency (see Taft 1991 for other examples).

The characteristic that is manipulated need not be a feature of the word itself, but instead a feature of the presentation conditions. The most common presentation manipulation is to precede the target item with another item that is related in some way to it. This other item is called the *prime*. If the prime has an impact on responses to the target, it indicates that the characteristic on which they are related plays a role in lexical processing. For example, faster lexical decision times to DOG when preceded by CAT (compared with when preceded by CAP), suggests that lexical access involves a semantic level of analysis (see *Priming, Cognitive Psychology of*).

3.2 *The Cognitive Neuropsychological Approach*

Conclusions have also been drawn about lexical processing by studying acquired dyslexics. These are people who were able to read normally before a trauma to the brain (e.g., through accident or stroke) resulted in a dysfunction in processing. Research into the reading behavior of such dyslexics has been influential in the development of accounts of lexical processing. If it is found that an individual has one area of functioning disrupted, but another intact, it suggests that those two aspects of processing are independent of one another. For example, some dyslexics (*surface dyslexics*) can read aloud nonwords and regular words, but not irregular words, while others (*phonological dyslexics*) only have a problem with nonwords. Such a pattern of dysfunction is taken to support the existence of two routes from print-to-sound (Coltheart et al. 1993), namely a lexical route where the phonology associated with the whole word is available (disrupted for surface dyslexics) and a rule-based GPC route (disrupted for phonological dyslexics).

3.3 *Direct Measures of Processing*

Lexical access can also be examined by looking directly at the organs involved in the process, namely the eyes and the brain. By examining eye movements, an analysis can be made of the timing and placement of fixations during the course of normal reading. Manipulations of lexical information in both the fovea and parafovea have provided a rich source of data regarding the processing of words (see *Eye Movements in Reading*).

Direct observations of the activity in the brain when words are presented (e.g., using ERP, fMRI, and PET techniques) is also beginning to provide fertile ground for an understanding of lexical processing, although

the contribution of this research is likely to be more in determining the physical locus of aspects of processing than in providing a fine-grained analysis of this processing.

See also: Literacy, Cognitive Psychology of; Reading Nonalphabetic Scripts; Reading Skills; Writing Process, Psychology of

Bibliography

Baayen R H, Dijkstra T, Schreuder R 1997 Singulars and plurals in Dutch: Evidence for a parallel dual-route modes. *Journal of Memory and Language* **37**: 94–117

Caramazza A 1997 How many levels of processing are there in lexical access? *Cognitive Neuropsychology* **14**: 177–208

Coltheart M, Curtis B, Atkins P, Haller M 1993 Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review* **100**: 589–608

Cutler A, Mehler J, Norris D, Segui J 1986 The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language* **25**: 385–400

Dell G S, O'Seaghdha P G 1994 Stages of lexical access in language production. *Cognition* **42**: 287–314

Feldman L (ed.) 1995 *Morphological Aspects of Language Processing*. Lawrence Erlbaum, Hillsdale, NJ

Forster K I 1981 Priming and the effects of sentence and lexical context on naming time: Evidence for autonomous lexical processing. *Quarterly Journal of Experimental Psychology* **A33**: 465–95

Forster K I, Chambers S M 1973 Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior* **12**: 627–35

Frost R 1998 Towards a strong phonological theory of visual word recognition: True issues and false trails. *Psychological Bulletin* **123**: 71–99

Levelt W J M 1992 Accessing words in speech production: Stages, processes and representations. *Cognition* **42**: 1–22

Marslen-Wilson W 1989 Access and integration: Projecting sound onto meaning. In: Marslen-Wilson W (ed.) *Lexical Representation and Process*. MIT Press, Cambridge, MA, pp. 3–24

McClelland J L, Rumelhart D E 1981 An interactive activation model of context effects in letter perception. Part 1. An account of the basic findings. *Psychological Review* **88**: 375–407

Morton J 1969 Interaction of information in word recognition. *Psychological Review* **76**: 165–78

Norris D, McQueen J M, Cutler A 1995 Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology*: *Learning and Memory Cognition* **21**: 1209–28

Seidenberg M S, McClelland J L 1989 A distributed, developmental model of word recognition and naming. *Psychological Review* **96**: 523–68

Simpson G B 1984 Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin* **96**: 316–40

Taft M 1991 *Reading and the Mental Lexicon*. Lawrence Erlbaum, Hove, UK

Taft M, van Graan F 1998 Lack of phonological mediation in a semantic categorization task. *Journal of Memory and Language* **38**: 203–24

Treiman R, Mullennix J, Bijeljac-Babic R, Richmond-Welty E D 1995 The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology*: *General* **124**: 107–36

M. Taft

Lexical Functional Grammar

1. Introduction

Lexical Functional Grammar (LFG) is a theory of language structure dealing with the syntax, morphology, and semantics of natural languages. It is different from other theories in having several parallel representations for sentences, each with its own architecture and vocabulary, subject to its own organizational constraints, and linked by principles of correspondence (mappings), not derived from one another. LFG is thus a constraint-based non-derivational unification model, unlike other generative grammar models such as Principles and Parameters or Relational Grammar. There is an extensive international community of scholars working within an LFG framework; it has been used to describe an impressive array of languages from various families and typological groups. Current research in LFG includes the role of morphology in language structure, semantic representations, and computational implementation; as well as an Optimality Theory-based model of LFG.

2. Background

From its inception in the late 1970s LFG has been concerned to be a model that is typologically grounded, computationally implementable, and consistent with psycho-linguistic understanding of language acquisition and comprehension. Bresnan (1982) sets out the original model and describes individual languages.

LFG emphasizes analysis in 'lexical' and 'functional' terms, rather than purely as phrase structure configurations (and movement of elements from one configurational position to another). Arguments in favor of separating constituency and functional representations (see Bresnan 2000, Carnie 2001, p. 339, Falk 2001) are discussed in Sects 2.1 and 2.2.

2.1 Structure–Function Mismatches

These mismatches occur where there is an imperfect correspondence between constituency and function, e.g., 'talked about' requires a Noun Phrase (NP) object and cannot have a tensed complement clause (CP) object:

(1) We talked about [$_{NP}$ the fact that he was unhappy $_{NP}$] for weeks

(2) *We talked about [$_{CP}$ that he was unhappy $_{CP}$] for weeks

However, if the CP occurs as a topic in the initial position the sentence is perfectly grammatical:

(3) [$_{CP}$ That he was unhappy $_{CP}$] we talked about for weeks

Transformational models would be required to move the CP from a post-predicate position where it cannot actually occur, while LFG is able to capture the difference in grammaticality in terms of functional differences (topic versus object—see Sect. 2.2).

2.2 *Non-configurational Languages*

In these languages word order is free and sentences have a 'flat' structure: there is no evidence for phrasal constituency asymmetries that distinguish grammatical functions. Non-configurational languages are typically 'dependent-marking' (case morphology encodes functional information), or 'head-marking' (agreement morphology distinguishes functions such as subject and object). Jiwarli (Western Australia) is a dependent-marking non-configurational language: case marking codes function; word order is free and semantically related items can appear discontinuously:

(4) *Juma-ngku* *ngatha-nha* *nhanya-nyja*
child-erg 1sg-acc see-past
ngulu *walhirrkura-lu*
that-erg naughty-erg
'That naughty child saw me'

Here the subject appears split at the beginning and end of the clause (all elements are marked for the same ergative (transitive subject) case); any other ordering or reordering of words is perfectly grammatical in Jiwarli, thus:

(5a) *Ngulu* *juma-ngku* *ngatha-nha*
nhanya-nyja *walhirrkura-lu*
(5b) *Ngatha-nha* *walhirrkura-lu* *nhanya-nyja*
ngulu *juma-ngku*
(5c) *Nhanya-nyja* *ngatha-nha* *ngulu*
juma-ngku *walhirrkura-lu*

all mean 'That naughty child saw me' (see Austin and Bresnan 1995, Austin 2001). In non-configurational languages constituency and function must be separately represented, unlike configurational languages (such as English) where the Verb Phrase constituent linking Verb and Object (but not Subject) allows function to be expressed in constituency (and word order) terms.

3. *Parallel Structures*

LFG analyses sentences in terms of (at least) four parallel structures:

- c-structures represent *constituency* facts (word order and phrasal grouping) in the form of context-free phrase structure trees of the usual X-bar theory type;
- f-structures represent *functional* information (grammatical functions such as SUBJect and OBJect, and discourse functions like TOPic) in the form of matrices of attribute-value pairs. Attributes may be morphosyntactic features like tense or number, or grammatical functions;
- a-structures represent *predicate-argument* information such as the number and type of arguments and their thematic role (such as Agent, Patient or Location), stated as arrays of predicates and argument slots;
- σ-structures represent *semantic* information through deductive assemblage of meanings based on reasoning subject to constraints; σ-structures use a linear logic building on a-structure information.

There may be other parallel structures, such as *thematic-structure* representing discourse pragmatic information, or *phonological-structure* representing sounds; they have not been elaborated in LFG work to date.

The parallel representations are linked by *correspondence principles* where information of one type can be *mapped* to other structures. Thus c-structure nodes in a configurational language map to grammatical function attributes in the corresponding f-structure (see below).

3.1 *C-structure*

Constituency information is represented as context free labeled phrase markers (trees). LFG c-structures adopt the X-bar model of head-dependent relations; 'functional' elements such as Determiners, Complementizers and Inflections are co-heads of lexical elements such as Nouns and Verbs. C-structures are subject to the *lexical integrity principle*: the minimal c-structure elements are whole words, not parts of words or empty categories. No movement of c-structure constituents (such as V to I movement) is allowed, unlike other syntactic theories. Syntax cannot see into the internal composition of words. An example of a c-structure is:

(6)

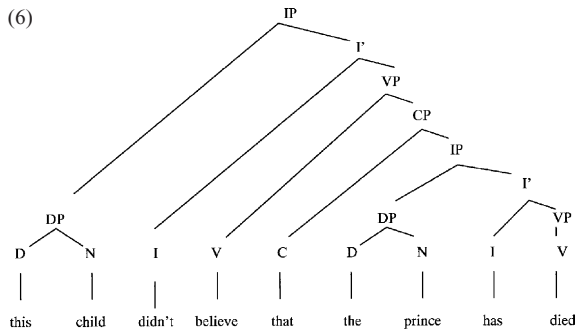

LFG uses lexical category information to deal with similarities in distribution that are captured by move-

ment in transformational models, e.g. in English auxiliaries occur in the I c-structure position, however if there is no auxiliary then the verb may appear there, as in:

(7) This child *didn't* believe that the prince has died.
(8) This child *can* believe that the prince has died.
(9) This child *believes* that the prince has died.

In Principles and Parameters the inflected verb is moved from its V position at underlying structure to I position in a later structure. In LFG the distribution is captured by classifying an inflected verb like *believes* as an I category marked for TENSE, but an untensed form like *believe* as V. A consequence is that in sentences like (9) the c-structure VP contains no head V (a principle of *economy* on c-structures says that nodes should only be expanded when necessary).

In configurational languages c-structure nodes are annotated with functional equations to pass information up from daughter to mother nodes, using ↑('information about my mother') and ↓('information about me') symbolizations. Thus, in English the DP immediate daughter of IP ('this child' in example (6)) is the subject of the sentence and annotated as ↑SUBJ = ↓. A DP immediate daughter of VP will be annotated ↑OBJ = ↓. Lexical heads like V and N simply have ↑ = ↓ annotations (i.e. they pass all their lexical information up to their mother nodes).

(10)

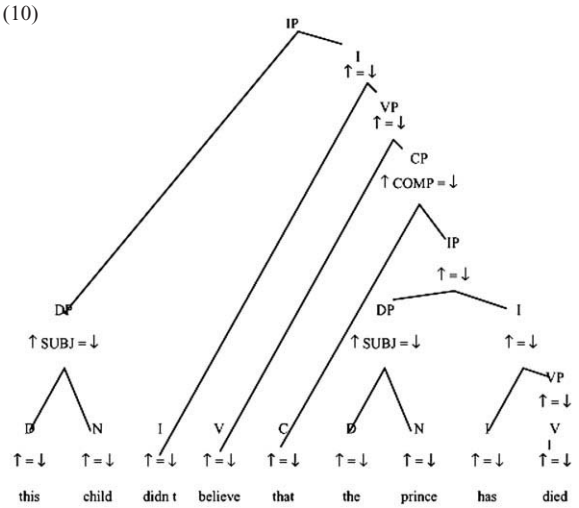

C-structures in configurational languages are generated by context free phrase structure rules with information-passing annotations, as in:

IP→ DP I'
↑SUBJ = ↓ ↑ = ↓

In non-configurational languages morphology plays a major role in function assignment.

3.2 A-structure

LFG includes a richly articulated lexicon expressing all morpho-syntactically relevant information for individual lexical items; for predicational elements the number and types of argument and complement slots that they subcategorize for, as in:

(11) *gives* I (↑PRED) = 'give ⟨Agent, Theme, Recipient⟩'
(↑TENSE) = present
(↑SUBJ NUMBER) = sing
(↑SUBJPERSON) = 3

A-structure and functional information (such as the Agent being the SUBJECT in simple transitive clauses) are related by lexical mapping correspondence principles (see Sect. 3.4.1).

3.3 F-structure

F-structures capture functional information and are sets of paired attributes and values in an *attribute-value matrix*. Attributes are morpho-syntactic features (derived from lexical entries) such as TENSE or NUMBER, or grammatical functions such as SUBJECT and OBJECT. LFG has a richly articulated typology of grammatical functions, including argument and complement functions subcategorized for by predicates, and adjunct functions. It also has discourse functions like TOPIC and FOCUS. In configurational languages correspondence principles map from c-structure positions to f-structure functions (see Sect. 3.4.2).

An f-structure for the sentence in (7) above would be:

(12)

SUBJ	DEF +		
	DIST proximate		
	PRED 'child'		
TENSE	past		
NEG	+		
PRED	'believe < SUBJ, COMP>'		
COMP	SUBJ	DEF +	
		PRED 'prince'	
	MOOD	perfect	
	PRED	'die <SUBJ>'	

There are three layers of f-structure: the outer f-structure (corresponding to the whole sentence) has five attribute names: SUBJ, TENSE, NEG, PRED and COMP. Three attributes are features with simple

values (PRES, +, 'believe ⟨SUBJ, COMP⟩'), while SUBJ and COMP are functions containing subordinate f-structures as values. These inner f-structures also have attributes and values – the SUBJ function of the embedded COMPlement takes a lower f-structure with grammatical features.

F-structure has an entirely different organization and vocabulary from c-structure (trees with constituents like N, DP and VP). F-structure attribute-value matrices are subject to well-formedness conditions of unification, i.e. the information within an f-structure must be:

- *unique*—attributes must have only one value
- *complete*—an f-structure must contain all the grammatical functions that a given predicate requires
- *coherent*—all the grammatical functions must be required by some predicate within the local f-structure

LFG assumes f-structures have a degree of universality: translationally equivalent sentences in different languages have identical f-structures, even if their c-structures are quite different (as they would be for English and Jiwarli, for example).

3.4 Correspondence Principles

Central to LFG are *correspondence* or *mapping principles* relating the several parallel and independent representations of sentences.

3.4.1 A-structure to f-structure mapping. Lexical mapping theory establishes correspondences between a-structure thematic roles and f-structure grammatical functions. LFG has four core grammatical functions: SUBJECT, OBJECT, OBJECT_θ (thematically restricted 'second object' in sentences like 'Bill gives John *money*') and OBLIQUE_θ (obliques with a range of thematic roles), decomposable into two features:

- $[\pm r]$ for *semantically restricted* functions; OBJECT_θ and OBLIQUE_θ will be $[+r]$
- $[\pm o]$ for *object-like* functions, OBJECT and OBJECT_θ will be $[+o]$

This gives:

$\text{SUBJ} = [-r, -o]$ $\text{OBJ} = [-r, +o]$
$\text{OBL}_\theta = [+r, -o]$ $\text{OBJ}_\theta = [+r, +o]$

Grammatical functions are hierarchically ranked from least marked to most marked:

$\text{SUBJ} > \text{OBJ}/\text{OBL}_\theta > \text{OBJ}_\theta$

Lexical mapping assumes a Universal Thematic Hierarchy (see Bresnan and Kanerva 1989) reflecting a scale of thematic prominence:

Agent > Beneficiary > Goal > Instrument > Patient/Theme > Locative

and serves to link the hierarchy of thematic roles to that of grammatical functions in a predicatable fashion (thus Agent links to SUBJ in unmarked constructions, Theme to OBJ_θ and Goal to OBJ with verbs like 'give,' etc.).

Argument mappings can be affected by *lexical rules* which change a-structures. So, *passive* is a lexical rule which removes the highest argument and makes it unavailable for mapping (hence Agent is no longer able to be SUBJECT, but Patient or Recipient can be).

3.4.2 C-structure to f-structure mapping. In configurational languages mapping from c-structure to f-structure is mediated by phrase structure annotations (see (10)) passing information from lexical items and structural positions. Information is mapped from phrase structure nodes to a *variable* (corresponding to a pair of brackets in the f-structure matrix). Variables are resolved by an *f-description*, a set of *functional equations* for all the nodes in the tree. Consider just the subject DP in (7) above with variables indicated in (13).

(13)

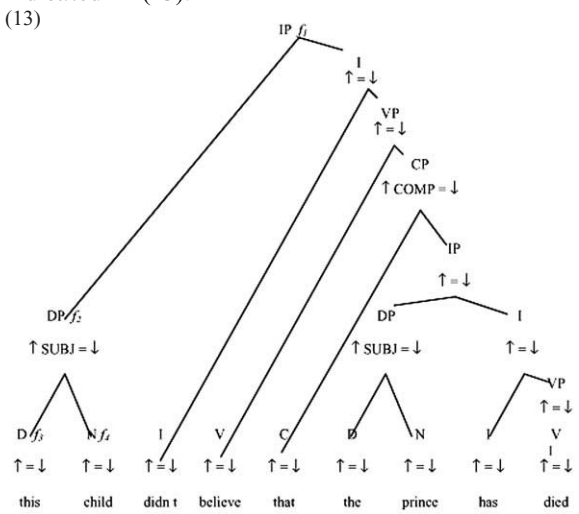

This gives the following functional equations:

(14) $(f_1\ \text{SUBJ}) = f_2$
$(f_2\ \text{DEF}) = f_3$
$(f_2\ \text{DIST}) = f_3$
$(f_2\ \text{PRED}) = f_4$

These are resolved as:

$$\begin{bmatrix} \text{SUBJ } f_2 f_3 f_4 & \begin{bmatrix} \text{DEF} & + \\ \text{DIST} & \text{proximate} \\ \text{PRED} & \text{'child'} \end{bmatrix} \end{bmatrix}$$

and so on for the rest of the nodes on the tree. Mapping is from c-structure configurations plus lexical information to f-structure matrices.

In non-configurational languages structural position does not define grammatical functions; this information must be derived from other sources. Nordlinger 1998 suggests that morphologically inflected forms are annotated to construct their local f-structure environment. Thus, in Jiwarli the ergative case form would construct a partial f-structure where the nominal bearing this case is SUBJECT (regardless of its c-structure position), accusative constructs an OBJECT function, etc. In head-marking languages agreement morphology is involved in resolution of functional equations to construct f-structures directly from morphological (lexical) information, without mapping from c-structure (annotated) positions.

4. *Functions in Complex Sentence Constructions—Raising and Control*

In non-finite complement clause constructions the subordinate verb may be missing a subject:

(15) Harry doesn't like to speak Sumbawan

(16) Harry doesn't seem to speak Sumbawan

The main clause subject 'Harry' is also understood as the subject of the subordinate predicate 'speak Sumbawan,' however:

- in (15) (often called a *control construction*) two thematic roles are involved: Harry bears a thematic relation to both 'like' and 'speak';
- in (16) (often called a *raising construction*) only one thematic role is involved: Harry bears a thematic relation to 'speak,' but not to 'seem.'

In transformational grammars these two constructions are analyzed as having an *empty category* (or *gap*) as the subordinate subject; for control this is PRO, which obtains its reference from the main clause subject, and for raising it is a DP trace that remains after the subordinate subject has been moved up to the main clause:

Referential control

Harry doesn't like [PRO to speak Sumbawan] (17)

Harry doesn't seem [t to speak Sumbawan]

Movement (18)

Since LFG does not permit empty categories in c-structure, the constituency of these two sentences types is identical; both have a main clause verb taking a VP' (to VP) complement with no subject DP in c-structure:

(19)

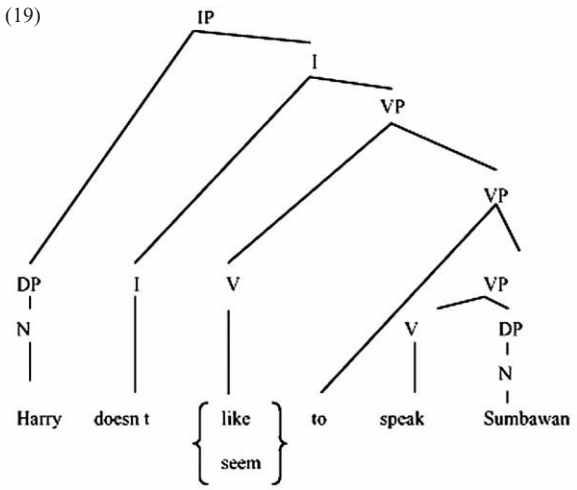

The grammatical function of VP' is XCOMP, an *open function* complement missing a subject function in c-structure. The identity of this subject function in f-structure is resolved by *functional control*. The lexical entries for *like* and *seem* are annotated to indicate that their XCOMP's SUBJect is informationally equivalent to their own SUBJect (for *like* the main clause subject is semantically selected by the verb (inside the 〈 〉 brackets in the verb's a-structure); for *seem* it is not (outside the 〈 〉 brackets):

(20) *like* V (↑ PRED) = 'like 〈 (↑ SUBJ) (↑ XCOMP) 〉'
(↑ SUBJ) = (↑ XCOMP SUBJ)

(21) *seem* V (↑ PRED) = 'seem 〈 (↑ XCOMP) 〉 (↑ SUBJ)'
(↑ SUBJ) = (↑ XCOMP SUBJ)

The functional control equation is mapped to f-structure as a link between the main subject's f-structure and the subordinate subject's f-structure (the controller *f-commands* the controllee, i.e. is part of a less embedded f-structure):

(22)

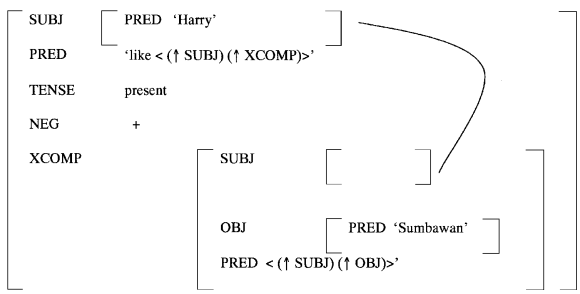

5. Long Distance Dependencies

In sentences like the following there is a dependency relationship between an element in sentence-initial position and a grammatical function later in the clause:

(23) What doesn't Harry like to speak?

(24) Which analysis does Joan seem to prefer to write about?

(25) That person, I've never seen you talk to.

The distance between initial position and the site of the grammatical function assignment is potentially unlimited; these are referred to as *long distance* or *unbounded dependencies*. Transformational analyses have movement (called wh-movement or A-bar movement) with the sentence-initial *filler* being *extracted* and leaving a *gap* (or wh-trace empty category). LFG has no movement operations; long distance dependencies must be captured in another way.

There are several accounts of long distance dependencies in the LFG literature, the most recent use *functional uncertainty* (see Kaplan and Zaenan 1989, Bresnan 2000). Essentially a functional equation identifies the initial element bearing a *discourse function* (DF) such as TOPIC or FOCUS with a *grammatical function* (GF) such as SUBJECT or OBJECT later in the sentence. The path of this identification can be indefinitely long and passes through (in English) any number of COMPlement clauses—function assignment is thus *uncertain*. There are two approaches to the functional equation: *outside-in functional uncertainty* (Kaplan and Zaenan 1989) or *inside-out functional uncertainty* (Bresnan 2000). For the former the initial topic or focus position in a language such as English has a functional equation associated with it:

(26) $(\uparrow \mathrm{DF}) = (\uparrow \mathrm{COMP}^* \ \mathrm{GF})$

This says: in f-structure identify the f-structure of an element bearing a discourse function with the f-structure of a grammatical function, passing through indefinitely many COMPlements (indicated by the Kleene star operator *). This identification is signaled by a link in f-structure, and any restrictions on the long distance dependency (so-called *island constraints*) can be stated in terms of the path from filler to gap through the f-structure. (There is also a link for raising and control structures, however this is strictly local and lexically determined.) LFG claims that c-structure information is irrelevant for long distance dependencies, contrary to transformational accounts. Evidence this is correct comes from *category mismatches* between the filler and gap (see example (3)); it is *functional* identification which is relevant.

6. Current Issues

Among areas currently of concern in LFG are:

- the role of morphology in language structure, including the contribution of morphology to functional structure independently of c-structure, the relationship between inflectional morphology and syntax (and the notions of functional category, competition and blocking), and the interface between morphology and syntax (including concepts of what it is to be a word, and how to analyze predicates that are morphologically complex, such as serial verb or main verb-classifying verb constructions);
- the nature of semantic representations and σ-structure, especially the development of a 'glue language' to relate semantics and f-structures;
- computational implementation of LFG grammar models on large data sets, computer architectures (including grammar-translation) using a data-oriented probabilistic approach to LFG-parsed corpora (called DOP-LFG)
- the relationship between Optimality Theory (OT) and LFG models, exploring how violable constraints can be added to LFG analyses to give a preference ranking between them, and how OT approaches to syntax can be informed by research on parallel constraint-based models such as LFG.

7. Conclusion

Lexical Functional Grammar is a powerful approach to linguistic analysis and theory building that has been practiced for over twenty-five years. It is characterized by parallel independent but related linguistic structures that each contribute information in their own constrained fashion. It has been applied to a wide range of languages and is being actively developed for linguistic analysis, theory construction, and computational modeling with the goal of analyzing and understanding the nature of human language.

8. Further Reading

Falk (2001) is a very readable introduction; see also Chapter 13 of Carnie (2001). Bresnan (2000) is a more advanced state-of-the-art presentation of the theory and several new ideas. Bresnan (1982) and Dalrymple et al (1995) contain classic articles that remain relevant, though now somewhat dated.

Much material is available on the internet at www-lfg.stanford.edu/lfg/ and clwww.essex.ac.uk/LFG/, and through the International Lexical Functional Grammar Association at: www-lfg.stanford.edu/lfg/ilfga/. There is also an LFG e-mail and discussion list: access is described on the web sites.

Bibliography

Austin P 2001 Word order in a free word order language: the case of Jiwarli. In: Simpson J, Nash D, Laughren M, Austin P, Alpher B (eds.) *Forty Years On*. Pacific Linguistics, Canberra

Austin P, Bresnan J 1995 Non-configurationality in Australian Aboriginal languages, *Natural Language and Linguistic Theory* **14**: 215–68

Bresnan J (ed.) 1982 *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, MA
Bresnan J 2000 *Lexical Functional Syntax*. Blackwell, London
Bresnan J, Kanerva J 1989 Locative inversion in Chichewa: A case study of factorization in grammar. *Linguistic Inquiry*, **20**(1): 1–50
Carnie A 2001 *Syntax*. Blackwell, London
Dalrymple M, Kaplan R M, Maxwell J T III, Zaenen A (eds.) 1995 *Formal Issues in Lexical-Functional Grammar*. Center for the Study of Language and Information, Stanford University
Falk Y 2001 *Lexical-Functional Grammar: An Introduction to Parallel Constraint-based Syntax*. Center for the Study of Language and Information, Stanford
Kaplan, R M. and Zaenen A 1989 Long-distance dependencies, constituent structure, and functional uncertainty. In: Baltin M, Kroch A, eds. *Alternative Conceptions of Phrase Structure*. Chicago University Press, Chicago, pp. 17–42. [Reprinted 1995, 137–165]
Nordlinger, R 1998 *Constructive Case: Dependent Marking Nonconfigurationality in Australia*. CSLI Publications, Stanford, CA

P. K. Austin

Lexical Processes (Word Knowledge): Psychological and Neural Aspects

He doth not neglect the profit of Lexicons (wherein all sayings and speeches are numbred)
E. Topsell, *The historie of foure-footed beastes*, 1607

If one looks in any psychology or psycholinguistics textbook for a definition of the mental *lexicon*, the superficial description given does not differ to one of the first uses of the word quoted above from the Oxford English Dictionary. Most modern researchers would agree that the lexicon refers to our knowledge of the words in our vocabulary. There is, however, a very broad range of opinion if one asks the next obvious questions: what is this knowledge and how is it represented in the brain?

As with many areas of higher brain function, studies of the mental lexicon generally follow one of two levels of explanation: descriptive and computational. There are currently no theories of word knowledge at the level of individual neurons. Descriptive explanations try to outline the nature of the lexicon in detail, attempting to establish both what information is represented by the lexicon and how this information is accessed. Computational explanations provide more explicit descriptions of word knowledge and lexical processes through either formal mathematical treatments or in the form of computer-based models that take a specific input and transform it to the appropriate output.

Studies on the lexicon can be grouped into four categories, each employing a number of different research methodologies:

(1) *Normal, intact subjects*: most research here relies principally on reaction time, and to a lesser extent accuracy, as the critical dependent measures. Participants are typically required to give some form of response to spoken or written words (word recognition) or to give a spoken, single name (word production). A large number of different tasks are employed in each of these modalities. These include lexical decision (deciding whether a spoken or written stimulus is a real word or not), word monitoring (listening to spoken input and indicating when a target word has been presented), picture naming, and word naming (single word reading aloud). All these techniques aim to capture variations in reaction times in terms of differences between words of different types and/or with different characteristics. Alternatively the experimenter aims to manipulate a baseline reaction time through different presentation conditions (for example, the experimenter may investigate the effect of semantic priming—presenting a target word soon after a semantically related item).

(2) *Patients with brain injury*: the critical assumption underpinning this area of research is that different aspects of language, or in this case word knowledge, are supported by different areas of the brain (concentrated almost entirely in the left, dominant hemisphere). Therefore by studying patients with damage to one of these areas, it should be possible to reveal the underlying characteristics of how the human brain supports language, including word knowledge. Unlike studies with intact, normal participants, most research in this area relies most heavily on accuracy rather than reaction times. Studies are conducted on groups of patients (comparing to normal subjects and/or different types of patient group) or individual cases. A number of recent studies have also used a technique known as case-series methodology in which patients are selected on the basis of having the same type of deficit but at different levels of severity.

(3) *Functional neuroimaging*: for those techniques that concentrate more on spatial than temporal resolution, the underlying assumption for functional neuroimaging is the same as for neuropsychological studies with aphasic patients. In this way, functional magnetic resonance imaging (fMRI) or positron emission tomography (PET) compare words that vary according to a specific characteristic either using a parametric, subtraction methodology (looking for areas of the brain that are more active for one set than another) or a correlational technique (looking for areas in the brain where changes in activation are correlated with a changing characteristic of the presented words). Other imaging techniques are able to investigate how the time-course of activation com-

pares for words of different type. Such methods include event-related potentials (ERP) and magnetoencephalography (MEG).

(4) *Computational modeling*: this is a formal technique designed to instantiate a proposed computational description of the processes that underlying a specific set of behaviors. Most computational models do this by specifying an arrangement of simple processing units (known as a *network architecture*). The architecture describes how simulated activation can potentially be passed between these units. Information is either represented in localist form (the activation of a single unit corresponds to a specific piece of information—e.g., a word) or in a distributed manner (a number of pieces of information are represented by different patterns of activation across the same set of units—therefore, the activation of a particular unit is not unique to a specific datum). These computational models also differ along another dimension. In some models, the strength of the connection between units (that governs how much activation is passed between them) is specified explicitly by the modeler. Other networks use mathematical algorithms such that the model can extract its own set of weights to performance optimally. In doing so, these latter types of network can also be used to compare patterns of development as well as mature performance between the model and human subjects. Both types of model can be used to simulate disordered behavior (e.g., through brain injury) by various means including the removal of units and/or their connections, or by adding noise to the activation of units.

These four methods have given rise to a vast literature on the nature of the lexicon and its interaction with the rest of the language system. Some of the major themes arising from this work are summarized below.

1. *The Traditional Conception of a Lexicon*

Many theories of lexical representation are not that different from Topsell's definition of 1607: the lexicon is a list of all the words in a person's vocabulary. By some form of look up, all or part (in the case of models with multiple lexicons, see below) of our knowledge about a word is retrieved. An early functional and mathematical description of this process was given by Morton with his influential logogen theory (1969), which was originally applied to explain word recognition but was extended later to address spoken and written recognition and production. The key elements to this approach are as follows. Each word is represented by a separate unit, or logogen, which gathers evidence for the presence of that word in the input (in the form of sensory input for recognition or semantic memory for production). Each logogen does this relatively independently of the others. A word is recognized or produced when there is sufficient positive evidence for that word in the nature of the input (e.g., the letters in the written input match those for the target word). At this point, the logogen makes other information about that word available—such as the word's meaning or its pronunciation. With reference to signal detection theory, Morton also included the notion of a 'threshold' which governs how much positive evidence is required before the logogen makes a response. In this way, word frequency effects in recognition and production are described in terms of different thresholds for different logogens (high frequency words have lower thresholds and thus require less positive evidence to respond).

2. *How Many Lexicons are there?*

If one considers the literate adult, the sum total of the acquired knowledge can be classified into three forms: the meaning of words; the sound of each word; and, the written form or orthography. With reference to the traditional view of the lexicon (see Sect. 1), some researchers have argued that these three forms of information are represented in separate lexicons: lexical semantics, a speech lexicon, and an orthographic lexicon. Although most models have interconnections between these different lexicons (at some level they need to combine to reflect all aspects of word knowledge), the notion is that they are discrete entities both at a functional and neural level. Functionally, each lexicon can then have its own dedicated form of representation and associated processes. At the neural level (both for neuropsychological and neuroimaging studies) this means there are distinct regions in the brain that can be damaged/activated relatively independently of each other. The cognitive neuropsychology literature, in particular, has also encouraged a further division in the mental lexicon—for input and output processes. The notion here is that there might be four separate lexicons to cover the four combinations of recognition and production for spoken and written words (for an overview, see Ellis and Young 1988). The same division is not proposed for meaning, which in a sense would be nonsensical anyway (for effective communication, the meaning driving the speaker/writer's output must be the same as that decoded by the listener/reader). The need to have separate input and output representations is still contested, however. Some neuropsychological data would seem to favor this position—for example, there are patients with word-finding difficulties (anomia) that do not have trouble understanding words (word deafness), and other patients with the opposite pattern. However, more studies are needed to rule out the possibility that the patterns reflect modality-specific nonlexical deficits or the fact that expression is always more vulnerable to impairment than receptive abilities.

Two separate strands of research have argued for another lexicon—the semantic lexicon or lemmas, that mediate between meaning and other language processes and representations (neuropsychology—e.g., Butterworth et al. 1984; speech production—e.g., Levelt 1992). In terms of neuropsychology, an impaired separate lexicon that normally acts as a gateway between meaning and language processes explains the pattern found in a number of patients who have impaired production and comprehension of speech and writing but much better understanding of nonverbal material (again, this position is contended, for a recent overview, see Lambon Ralph and Howard 2000). In speech production, Levelt (1992) has argued for a two-stage process within lexical access. First the meaning of the intended utterance is used to select a single abstract lexical form, a lemma (that also codes for syntactical markings such as gender and number), and then this in turn activates a lexemme, which makes the sound of the word available both in terms of the constituent phonemes, their arrangement and associated stress pattern. There is, however, a general concern with this direction of theorizing in that a proliferation of multiple lexicons means the potential for redundancy of representations is much higher, and tends to make the theories unfalsifiable.

3. *An Alternative Conception of the Nature of Lexical Knowledge?*

A quite different approach has been adopted by some parallel distributed processing (PDP) computational models in which there is knowledge of the words in our vocabulary but not a lexicon—at least, in the traditional sense of that word. A landmark publication in this respect was the word reading and recognition model of Seidenberg and McClelland (1989). In this framework word knowledge is represented as distributed patterns of activation across three 'primary' systems corresponding to orthography, phonology and semantic memory. Through a gradual learning process, in which the weights of the connections within and between these representational layers are adjusted, models of this type extract word 'knowledge' from the vocabulary they are trained on. Note this is quite different from the traditional metaphor of looking up an entry in dictionary or Morton's logogen theory. The Seidenberg and McClelland model does not have a list of all the known words (i.e., a lexicon) so the terms 'access' and 'retrieval' are redundant. Instead the model takes one form of input (e.g., the written word) and then computes the appropriate representation reflecting the meaning and/or the pronunciation. In this way the model can be said to have word knowledge but not a lexicon. This is further reinforced by the fact that the model processes novel words (nonwords that, by definition, cannot be in a lexicon) through the very same set of units corresponding to the written or spoken form. Another potential allure of this approach to word knowledge is that the formulation may help when future research (see Sect. 5) attempts to bridge the gap between brain and behavior. The neural basis for the primary systems of semantics, phonology, and orthography is much more obvious. It is possible, for example, that orthography reflects a stable set of representations extracted through visual perception (Behrmann et al. 1998) and semantic memory, or conceptual knowledge, reflects high-level abstraction based on all forms of verbal and nonverbal experience, perhaps supported primarily by regions within the temporal lobes (Tranel et al. 1997).

This alternative approach has also been used to address some of the data that has lead researchers, using a traditional view of the lexicon, to posit the need for multiple lexicons (see Sect. 2). For example, patients with both poor comprehension and production of words (whether spoken or written) are described from a traditional viewpoint in terms of damage to a semantic lexicon that mediates between conceptual knowledge and the language system. The same data are adequately captured by PDP computational principles in terms of the nature of the mapping between representations for words vs. pictures/objects (Lambon Ralph and Howard 2000). The fact that there is an arbitrary mapping between phonology/orthography and conceptual knowledge, unlike the mapping for pictures or objects, means that verbal comprehension and production are highly vulnerable to damage to the system that represents the meaning of both words and pictures.

4. *How Does Word Knowledge Give Rise to Specific Language Activities?*

Within the repertoire of language activities there are some that are assumed to depend directly on word knowledge or lexical information. For example, understanding words or producing words is assumed to proceed with reference to lexical information. So, as described above, within a logogen framework speech production occurs by the target meaning acting as an input to a set of logogens, which selects the correct word and then makes the spoken form available. There is also a direct use of word knowledge in PDP computational models as well—the knowledge extracted through learning the mapping between semantics and phonology is used to produce the correct spoken form in speech production. Other language activities are assumed to depend upon a combination of lexical knowledge and other nonlexical processes. So, for example, Levelt (1992) proposes that speech production is based upon a 'formulator' that specifies the syntactical frame for a sentence into which lexical representations (lemmas) and their associated knowledge are fitted. Two other well-known examples are reading aloud and the past tense transformation of

English verbs. Researchers studying the process of reading aloud have classified words into three basic types: (a) most words, known as regular or consistent words, are pronounced in line with their spellings (examples of these are YARD and MINT); (b) a smaller but still substantial number of irregular or inconsistent words deviate from the typical spelling-to-sound translations (examples of these are YACHT and PINT); and (c) we are also able to read new or novel words without extra training (e.g., EMAIL or WUG). There is a very similar classification for describing the past tense of an English verb. There are many regular verbs, all of which are turned into the past tense by the addition of -ED (e.g., *talk* → *talked* or *need* → *needed*), and a smaller number of irregular verbs which mark the past tense in another way (e.g., *speak* → *spoke* or *think* → *thought*). We are also able to produce the past tense for new or novel verbs (e.g., *email* → *emailed*, *wug* → *wugged*), again without any specific training on these novel forms. The traditional theories of verb inflection (the dual-mechanism account, Pinker 1999) and reading aloud (the dual-route theory, Coltheart et al. 1993) have both used this tripartite division of word type to argue that each language activity is underpinned by two distinctly different processes. The two theories are sufficiently similar that they can be summarized together. Irregular words/verbs (e.g., PINT → 'pint' for reading aloud and 'bring' → 'brought' for verb tense) are said to be translated by reference to a lexicon—or indeed, in the case of reading, two lexicons: one orthographic and the other phonological—that stores the whole specific word form for every irregular item in the language. The ability to read or inflect novel forms is explained in these theories by positing a task-specific, generative process encapsulated by a set of rules: three rules for the three allomorphs required for the inflection of regular verbs (/t/, /d/, /@d/) and an extensive series of rules for the translation of letters/graphemes into phonemes. Neuropsychological investigations of impairments to both reading aloud and past tense production have highlighted two contrasting patterns that have been taken to support these dual-mechanism theories (some patients have specific difficulties with irregular forms and others with novel forms: see Ullman et al. 1997 for verb morphology; an overview of acquired dyslexia can be found in Ellis and Young 1988).

Again, PDP computational models have taken a different approach (verb morphology, Joanisse and Seidenberg 1999; reading aloud, Plaut et al. 1996). In these frameworks word knowledge is not reflected in a lexicon and thus the strict division between the lexicon and the generative rules does not apply. Instead the interaction between and within the primary systems (semantics, phonology, and orthography) not only give rise to word knowledge, the models are also capable of transforming one form of representation into another (e.g., reading is orthography → phonology) for all three word types (irregular, regular, and novel). Critically, however the division of labor amongst the primary representations is not even for the different word types, and dissociations in performance can occur if one of the primary systems is damaged (for an overview of this perspective on acquired dyslexias, see Patterson and Lambon Ralph 1999).

5. *Future Directions*

Research on the psychological aspects of word knowledge will, no doubt, continue in the future. One potential area for new research must be on the neural aspects of word knowledge. There are already signs that neuropsychological and computational investigations of people with acquired language disorders (aphasia) are being used to address both psychological and more neural aspects of word knowledge (examples can be found in Ullman et al. 1997, Patterson and Lambon Ralph 1999). While there has been some use of positron emission tomography (PET) and functional magnetic resonance imagine (fMRI) to look at word knowledge (for an overview, see Price 1998), there is much more for these techniques to tell us about the neural aspects of word knowledge in the future.

See also: Lexical Access, Cognitive Psychology of; Lexical Semantics; Lexicon; Word Meaning: Psychological Aspects; Word Recognition, Cognitive Psychology of

Bibliography

Behrmann M, Plaut D C, Nelson J 1998 A literature review and new data supporting an interactive account of letter-by-letter reading. *Cognitive Neuropsychology* **15**: 7–52

Butterworth B, Howard D, McLoughlin P 1984 The semantic deficit in aphasia: The relationship between semantic errors in auditory comprehension and picture naming. *Neuropsychologia* **22**(4): 409–26

Coltheart M, Curtis B, Atkins P, Haller M 1993 Models of reading aloud: Dual-route and parallel-distributed- processing approaches. *Psychological Review* **100**(4): 589–608

Ellis A W, Young A W 1988 *Human Cognitive Neuropsychology*. Lawrence Erlbaum, London

Joanisse M F, Seidenberg M S 1999 Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 7592–7

Lambon Ralph M A, Howard D 2000 Gogi aphasia or semantic dementia? Simulating and assessing poor verbal comprehension in a case of progressive fluent aphasia. *Cognitive Neuropsychology* **17**: 437–66

Levelt W J M 1992 Accessing words in speech production: Stages, processes and representations. *Cognition* **42**: 1–22

Morton J 1969 The interaction of information in word recognition. *Psychological Review* **76**: 340–54

Patterson K, Lambon Ralph M A 1999 Selective disorders of reading? *Current Opinion in Neurobiology* **9**: 235–9

Pinker S 1999 *Words and Rules: The Ingredients of Language*. Weidenfeld & Nicolson, London
Plaut D C, McClelland J L, Seidenberg M S, Patterson K 1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* **103**(1): 56–115
Price C J 1998 The functional anatomy of word comprehension and production. *Trends in Cognitive Sciences* **2**: 281–8
Seidenberg M, McClelland J L 1989 A distributed developmental model of word recognition and naming. *Psychological Review* **96**: 523–68
Tranel D, Damasio H, Damasio A R 1997 A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia* **35**(10): 1319–27
Ullman M T, Corkin S, Coppola M, Hickok G, Growdon J H, Koroshetz W J, Pinker S 1997 A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience* **9**(2): 266–76

M. A. Lambon Ralph

Lexical Semantics

The 'core' areas of linguistics are generally agreed to be: phonology, the study of the sounds and sound systems of language; syntax, the study of grammar; and semantics, the study of linguistic meaning. Words figure in all three areas, but it is in their semantic aspect that they offer the most distinctive field of inquiry. This article presents a survey of the topics which form the principal foci of interest within the domain of lexical semantics.

1. Introduction

To the layman, words are the prime bearers of meaning in language, and formulating a message is essentially a matter of choosing the right words. This view is of course an oversimplification, and seriously understates the roles of grammar and of context, but there is more than a germ of truth in it. The branch of linguistics whose domain is the systematic study of word-meanings is lexical semantics. But before proceeding any further, it is necessary, first, to mark off an area of the slippery notion of 'meaning' which is appropriate for our purposes, and second, since other people, notably lexicographers, also concern themselves with the meanings of words, something needs to be said about what is distinctive about the activities of lexical semanticists.

If John says to Mary *Have you been to the bank and seen the manager?* and Mary replies *Yes*, there is an obvious and intuitively valid sense in which Mary's reply 'means' 'I have been to the bank and seen the manager.' Yet it is also fairly obvious that 'I have been to the bank and seen the manager' would not be a satisfactory definition of the meaning of the word *Yes*. It is therefore necessary to make a distinction between the 'inherent meaning' of a word and its 'meaning-in-context.' (The inherent meaning of *Yes* might be glossed: 'The proposition you present to me for confirmation is true.') A similar point (less dramatic) can be made about *The child ran to his mother*. Here it is clear that *the child* refers to a boy; but this, too, is meaning-in-context—the inherent meaning of *child* is something like 'immature human being.' As a first approximation, it can be said that lexical semantics focuses on inherent meanings of words: meanings-in-context are the concern of pragmaticians.

Lexicographers also are interested mainly in the inherent meanings of words. So what is the difference between a lexicographer's interest in word meaning and that of a lexical semanticist? Basically, this is a matter of theoretical vs. practical concerns, parallel in many ways to the relation between physicist and engineer, or botanist and farmer/gardener. A lexicographer aims to impart the maximum of information useful to the dictionary-user with maximum economy; a lexical semanticist wants to understand the principles underlying lexical semantic phenomena. In an ideal world there would be considerable cross-fertilization between the two fields: in reality, the cross-domain traffic has been somewhat disappointing.

2. Approaches to Word Meaning

There are numerous ways of approaching the study of word meaning, depending on certain basic assumptions one makes about it.

2.1 The Componential/Atomistic Approach

One influential approach, much criticized but constantly reborn, is to think of the meaning of a word as being a more or less complex structure, built up out of combinations of simpler, or more primitive, units of meaning, a simple example would be the analysis of *woman* as [ADULT] [HUMAN] [FEMALE]. Such 'componential' or 'atomistic' approaches come in a variety of types, depending on what is claimed regarding the nature of the analysis, the identity and nature of the units, how they are combined, and what aspects of word meaning are to be explained by them.

Some componential analyses (probably the majority) aim at a reduction in the number of combinatorial elements, that is to say, they aim to explain the large number of word meanings in terms of combinations of units drawn from a smaller set, parallel to the way that the phonological forms of all

words can be described in terms of combinations from a much smaller set of phonemes. To achieve a reduction, the basic elements must be highly reusable.

Not all analyses are reductive in this sense. One version aims only to distinguish the meaning of a word from that of all other items in the vocabulary, so each contrast potentially creates a new component. An example is the analysis of *chair* as [FURNITURE] [FOR SITTING] [FOR ONE PERSON] [WITH BACK]. This is certainly distinctive, but there is no guarantee that one will end up with fewer components than lexical items analyzed.

In the earliest versions of componential analysis, the components were the meanings of words, and the aim of the analysis was to extract a basic vocabulary, in terms of which all nonbasic meanings could be expressed. Generally speaking, the features recognized by earlier scholars had no pretensions to universality, and indeed were often avowedly language-specific. Later scholars aimed at uncovering universals of human cognition, a finite 'alphabet of thought.' Accessible introductions to componential analysis can be found in Nida (1975) and Wierzbicka (1996).

2.2 Holistic/Contextual Approaches

An atomistic approach to word meaning is predicated on the assumption that the meaning of a word can in principle be specified in isolation, without taking into account the meanings of other words in the language's vocabulary. Holistic approaches, on the other hand, maintain that words do not have meaning in isolation, but only in relation to the rest of the vocabulary.

A typical holistic approach is that of Lyons (1963, 1968, 1977), who espouses and develops the basic Saussurean notion that the 'value' of any linguistic element, whether phonological, syntactic or semantic, is essentially contrastive, that is to say, its identity is constituted by its difference from other elements with which it potentially contrasts. So, for instance, the essence of the meaning of *dog* is that it is 'not-cat,' 'not-wolf,' 'not-horse,' and so on. An important consequence of this is that it is not possible to learn the meaning of any word in isolation from other members of its contrast set; or, to put it another way, it is not enough to be able, for instance, to assign the label *dog* to any canine animal—one must also know to withhold the label from nondogs, such as foxes, wolves and so on. Lyons expounds an extended version of this, making it truly holistic, and defines the sense of a word as its position in a network of sense relations, direct or indirect, with all other items in the vocabulary. This network includes not only contrastive relations like that between *dog* and *cat*, but also relations of inclusion like that between *animal* and *dog*, and *dog* and *collie*.

2.3 Conceptual Approaches

Conceptual approaches locate meaning in the mind, rather than in language or in the extralinguistic world. Much discussion has had to do with the meanings of words in relation to conceptual categories (for general discussion, see Taylor 1989, and Ungerer and Schmid 1996).

The currently dominant view of what natural conceptual categories (concepts) are like is that they cannot in general be defined by any finite set of necessary and sufficient features, but are somewhat indeterminate in nature. One important theory, prototype theory, holds that natural categories are organized around ideal examples (prototypes), and that other items belong to the category to the extent that they resemble the prototype. There is normally no definite cut-off point dividing members from nonmembers: category boundaries are typically fuzzy, or vague.

A major difference of opinion exists over whether words map directly onto conceptual categories, or whether there is an intermediate level of semantic structure which is purely linguistic in nature. Those, like the cognitive linguistics, who argue for direct mapping, claim that there is no theoretical work for a linguistic level of semantics to do: all semantic phenomena can be given a conceptual explanation. Those who believe in an independent linguistic level of semantics argue that the fuzziness of conceptual distinctions contrasts with the sharpness of linguistic distinctions, and indicates a separate level of structure. (Think of the uncertain dividing line between 'alive' and 'dead' in 'real life' in relation to the fact that *The rabbit is dead* logically entails *The rabbit is not alive*.) They also point to cases like *book*, which (it is claimed) has a single sense (at the linguistic level), but corresponds to two different concepts, a concrete physical object (*The book fell from the shelf*) and an abstract text (*It is a very difficult book*).

A conceptual approach is not necessarily incompatible with a componential approach. Many prototype theorists characterize categories/concepts in terms of a set of prototype features, which may or may not have the status of semantic primitives. Such features are not of the necessary-and-sufficient sort, however, but have the property that the more of them something has, the better an example of the relevant category it is.

3. Core Problems of Lexical Semantics

Certain areas of investigation can be regarded as the defining core of the discipline.

3.1 Possible Words

An important question in lexical semantics is whether there are any restrictions on possible word meanings:

is there anything which can be conceived and described, but which could not constitute the meaning of a word? This is a difficult question, and there are a number of ways of approaching it.

It might be thought that there could not possibly be a verb which meant, for instance, 'to eat popcorn on a Wednesday morning while standing facing west,' or a noun denoting, collectively, a computer mouse, a used condom, and a map of the world. These are admittedly unlikely, but not logically impossible, provided they can be conceptually unified, and there is no apparent limit to what disparate events, properties, or things can be conceived as a unity in some belief system or other. (The first example above, for instance, could conceivably be part of the ritual practices of an eccentric religious cult.) Of course, prototypically, a noun denotes some spatiotemporally continuous entity, and prototypical events have a similar unity, and such tendencies are an important part of the semantic nature of words.

An apparently universal constraint on word meaning is 'structural continuity.' A simple illustration of this will have to suffice, as a precise statement of the constraint is quite complex. Take the phrase *a very old man*. A word meaning 'very old,' or 'old man,' or 'very old man' would occasion no surprise. But one meaning 'very—man,' such that any adjective applied to it would be automatically intensified (e.g., *an old very-man* = 'a very old man,' *a poor veryman* = 'a very poor man,' etc.) is inconceivable.

There are certain language-specific constraints on what can be packaged inside a single word. A famous example is the inability of, for instance, a French verb to simultaneously encode movement and manner information. In the English sentence *John ran up the stairs*, movement and manner are encoded in *ran* and direction in *up*. However, a French translation of this must package the information differently: *Jean monta l'escalier en courant*. Here, movement and direction are encoded in *monta*, and manner separately in *en courant*. The extent of such restrictions is at present insufficiently researched.

3.2 Polysemy

The meanings of words are notoriously affected by context: it is only a slight exaggeration to say that the semantic contribution a word makes is different for every distinct context in which it occurs. But the difference between two contextually-induced readings can range from relatively slight, as in *My cousin married a policeman* and *My cousin married an actress* ('male cousin' vs. 'female cousin'), to rather major, as in *elastic band* and *wind band*, with perhaps *a coat of paint* and *a coat of many colors* as intermediate examples.

A distinction can be made between cases where a word has more than one meaning and different contexts select different meanings, and cases where only one meaning is involved, which can be enriched in different ways in different contexts. The first type of case in general involves ambiguous words. These have a number of distinctive properties. Think of *bank* in *We finally reached the bank*. The two readings of *bank* are in competition, and only one can win: it is not possible to 'mean' both readings at once, except when punning. Nor is it possible to remain uncommitted with respect to the two meanings: with *cousin*, for instance, whether the cousin in question is male or female can be left open, but with *bank* one meaning or the other must be intended. In the case of *cousin* it can be said that there is a general meaning which covers both of the more specific possibilities; with *bank* there is no such 'superordinate' meaning. Contextual enrichment without polysemy may involve a type-to-subtype specification as with *cousin*, or *fly like a bird* and *sting like a bee*, where reference is to a prototypical instance of the mentioned category (i.e., not an ostrich or a drone), or it may be whole-to-part, as in *My car has a puncture*.

The distinct readings of ambiguous words may be totally unrelated, as with the two senses of *bank* and *band*: this is called *homonymy*. Lexicographers normally treat such cases as separate words that happen to have the same phonetic form. In other cases, there may be an intelligible semantic relation between the readings, in which case we speak of *polysemy*: lexicographers treat polysemy as involving a single word (with one main entry) which has two or more meanings. It is important to realize that there is no sharp dichotomy between 'related' and 'unrelated' meanings, but a continuous scale of relatedness. Some relationships are intuitively clear, such as the metaphorical relation between *branch of a tree* and *branch of a bank*, or the metonymic relation between *the old oak tree* and *It's made of solid oak*. Speakers are less sure whether they would mark *try* in *try a jacket on* and *try someone for murder* as related. Lexicographers often use common etymological origin as a criterion for polysemy, but this may run counter to current speakers' intuitions (usually more important for linguistics. Famous cases are *ear of corn* and *human ear*, which do not have a common origin, but are nonetheless felt by most people to be related, and a *glass of port* and *the ship entered port*, which have a common origin, but no relationship accessible to unaided intuition.

Some polysemic relationships are 'one-off'; others are highly recurrent, in which case the term *regular polysemy* may be used. Many metonymic relations are at least to some extent recurrent: for instance, the 'tree'/'wood' example mentioned above, or the difference between *Mary planted some roses* and *Mary picked some roses*. Metaphorical relations are typically not recurrent. For instance, both *head* and *foot* are often metaphorically extended, but the fact that one occurs is no guarantee that the other does. Hence, for instance, *foot of mountain* occurs but not *head of*

mountain, *head of school* but not *foot of school*, and so on.

Where there is a set of polysemous readings, it is often possible to identify one which is an intuitively plausible source from which the others can be derived by extension. For instance, given the three meanings of *position* in: *The house is in a wonderful position*, *John holds an important position in the company* and *What is your position on the euro*?, it is fairly clear that only the first is a plausible metaphorical source for the other two. This is commonly called the *literal* meaning. (For a general discussion of polysemy and related matters, see Cruse 2000, Chap. 6).

3.3 Specifying Word Meanings

How (or whether) one goes about the enterprise of specifying the content of a single sense depends very much on one's favored theoretical approach, and also to some extent on the purpose of the exercise. (Only a few linguistic semanticists have made anything beyond very small-scale attempts to provide full descriptions of real lexical senses.) A componentialist will seek to provide a finite set of necessary and sufficient semantic features; most meaning descriptions by cognitive linguists are also in the form of a set of features, but these are often weighted for importance, and are not necessary and sufficient. For any sort of holist, the full specification of a sense will necessarily be somewhat cumbersome. Lexicographers usually try (with varying degrees of success) to provide a watertight definition of the necessary-and-sufficient variety. This is one area where there is little in the way of consensus.

3.4 Sense Relations

The observation that the vocabularies of natural languages are at least partially structured has led to a lively interest in the underlying structuring principles—in particular, recurrent relations between meanings (usually referred to as '*sense relations*' or '*lexical relations*'). These are of two main sorts, paradigmatic and syntagmatic. It is the former that have been the most studied. (Cruse 1986 discusses sense relations in some detail.)

3.4.1 Paradigmatic relations. Paradigmatic sense relations represent sets of choices at particular points in the structure of sentences, and reflect the way a language articulates extra-linguistic reality. A limited number of sense-relations have figured prominently in discussions; they are: synonymy, hyponymy, meronymy, incompatibility, and oppositeness.

(a) *Synonymy*. Synonymy, or sameness of meaning, is not as simple as it appears. Absolute synonymy, which we can define as having the same degree of normality in all contexts, is extremely rare (if it occurs). Propositional synonyms, such as *begin* and *commence*, which have the same logical properties, are more common, but would not serve to fill a 'Dictionary of Synonyms.' Lexicographers' synonyms, such as *brave* and *heroic*, have an adequate intuitive basis, but are hard to pin down: they may be used contrastively (*He was brave, but I wouldn't say he was heroic*) but their differences must be of a minor nature.

(b) *Hyponymy*. Hyponymy is one of two important relations of inclusion. Examples are: *alligator: reptile, aunt: relative*, *truck: vehicle*. In the last pair, *truck* is the hyponym and *vehicle* the superordinate: the hyponym denotes a subclass of the class denoted by the superordinate (alternatively, the meaning of the superordinate is included in that of the hyponym). Notice that the terms *hyponym* and *superordinate* are relational, not absolute: *dog* is a hyponym of *animal* but a superordinate of *collie*.

(c) *Meronymy*. The second important inclusion relation is meronymy, the 'part–whole' relation, exemplified by *finger: hand*. (Hyponymy and meronymy must not be confused: a finger is not a type of hand, nor is a dog a part of an animal.) True parts are interestingly different from 'mere' pieces, in that they prototypically have clear boundaries, a characteristic form recognizable in other wholes of the same type, and a determinate function with respect to their wholes. (Think of the difference between the parts of a teapot—handle, spout, lid, and so on—and what typically results if a china teapot is dropped onto a hard floor.)

(d) *Incompatibility*. Incompatibility is related closely to hyponymy, being the typical relation between sister hyponyms of the same superordinate—like *dog*, *cat*, (*animal*); *apple*, *peach* (*fruit*). Incompatibility is not simply difference of meaning (sometimes termed '*heteronymy*'): incompatibles have a relation of mutual exclusion. For instance, *dog: cat* and *dog: pet* both represent 'difference of meaning.' But while it follows logically from the fact that something is a dog, that it is not a cat, hence *dog* and *cat* are incompatibles, it does not follow that it is not a pet, so *dog* and *pet* are not incompatibles. There is a parallel relationship between sister parts of a whole (co-meronymy): sister incompatibles denote classes with no members in common; co-meronyms denote sister parts that do not overlap physically (like *upper arm*, *lower arm*, and *hand*).

(e) *Opposites* (*Antonyms*). Opposites can be briefly characterized as binary incompatibles, that is, incompatibles that necessarily occur in two-member sets, several types can be recognized.

(i) Complementaries divide some domain into two, with no 'sitting on the fence.' For instance, organisms must be either *dead* or *alive*; those that are not dead are alive, and those that are not alive are dead. Other examples are *open: shut*, *true: false*, *stationary: moving*.

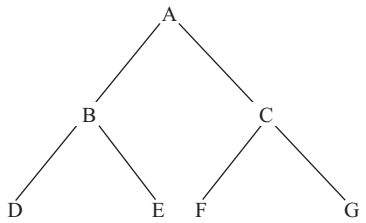

Figure 1
The structure of a hierarchy

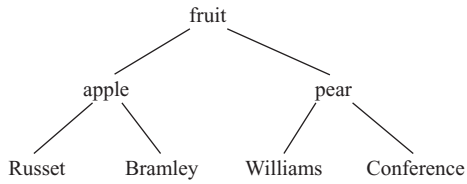

Figure 2
A taxonomic lexical hierarchy (fragment)

(ii) Gradable contraries (in some systems, simply *antonyms*) differ from complementaries, first, in that they are gradable, that is, they occur normally with intensifiers such as *very*, *slightly*, *a bit*, etc., and in the comparative degree, and second, they are contraries, in that denying one term of a pair does not commit one to the other. A typical example is the pair *long*:*short*.

(iii) *Reversives* are opposites which denote counter-directional processes or actions; examples are *rise: fall*, *open: close*, *dress: undress*.

(iv) *Converses* present the same situation or event from different viewpoints: *The painting is above the photograph/The photograph is below the painting*; *John sold the car to Bill/Bill bought the car from John*.

3.4.2 Syntagmatic relations. Syntagmatic relations hold between items in the same larger unit, such as phrase, clause or sentence. They represent not choice but restrictions on choice, and control discourse cohesiveness. The effects of putting two words together in a grammatical construction can be described as 'normality' (*The cat purred*), 'clash' (*The cat barked/broke*), or 'pleonasm' (*He kicked it with his foot*). Syntagmatic relations are described usually in terms of 'selectional restrictions,' which specify the conditions for the normality of a sequence. So, for instance, for *a pregnant X* to be normal, *X* must denote a female mammal; for *Mary drank the Y* to be normal, *Y* must denote a liquid, and so on.

3.5 Lexical Structures

Sense relations form the basic structuring principles of various lexical configurational patterns in the vocabulary of a language. An important type is the lexical hierarchy, structured on the following pattern (which in principle can continue indefinitely in a downward direction, and can have indefinitely many items at each level except the first; see Fig. 1).

The two main types of lexical hierarchy are the classificatory, or taxonomic, variety, and the part–whole, or meronomic type. An important feature of a hierarchy is that, as we move down the structure, the branches proliferate, but never converge. In a taxonomic hierarchy, the successive levels represent successive subdivisions into mutually exclusive classes of items: each item is related to items higher up the hierarchy by the relation of hyponymy (as shown by the lines), and to items at the same level by incompatibility (see Fig. 2).

The structuring relations for a meronomic hierarchy are meronymy (vertical) and co-meronymy (horizontal) (see Fig. 3).

Other common lexical structures are nonbranching chains of various sorts:

Monday—Tuesday—Wednesday (etc.)
puddle—pond—lake—sea—ocean
freezing—cold—cool—lukewarm—warm—hot—scorching
millimetre—centimetre—metre—kilometre
second—minute—hour—day—week—month—year—decade—century—millennium

and 'grids' or correlated series:

sheep	mutton	mare	stallion
pig	pork	ewe	ram
cow	beef	doe	buck
deer	venison	lioness	lion

3.6 Historical Change of Meaning

Words are being used continually in ways which transform, stretch, broaden, or narrow their established senses. This flexibility is one of the strengths of language, enabling it to adapt and remain an effective tool of communication through ever changing circumstances. If a word is used repeatedly with a particular extended meaning, this may well, over time, become established, either as a new polysemous sense, or it may be incorporated as a result of a shift in the boundaries of the old sense. Many new senses can be found in the domain of personal computers: for example, *mouse*, *scroll*, *icon*, *alias*, *tool*, *save*, *back-up*. An example of changing boundaries is the incorporation of the legs of tables, chairs, etc., into the general category of LEGS, without any intuition of

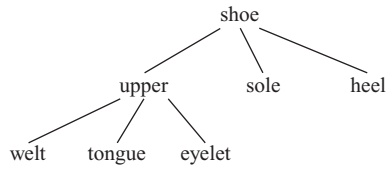

Figure 3
A meronomic lexical hierarchy

figurative use. Most speakers regard the current use of *pet* in *cyberpet/electronic pet* as metaphorical extension, but in time this may well go the way of *leg*. The creation of a new polysemic sense is a sudden event (although the process of establishment is gradual). Some diachronic changes, however, are gradual. If one accepts a prototype account of meaning, then the prototype of, say, *car* has changed gradually over the last century or so, as cars themselves have changed.

4. General Questions

4.1 Dictionary vs. Encyclopedia

Many theorists (but by no means all) draw a distinction between an (ideal) dictionary of a language, and an encyclopedia. The former purports to contain information about the meaning of a word, while the latter incorporates general knowledge about the word's referents. The general assumption is that only the former is truly linguistic in nature; the latter concerns not semantics, but pragmatics. The distinction has some intuitive support: on the basis of our experience with dictionaries and encyclopedias, we have fairly clear expectations of what to find in the two types of publication. However, making the step from the likely contents of familiar publications to the contents of a 'lexical entry,' either in a theoretical model of word-semantics, or in a psychologically real 'mental lexicon,' is far from straightforward, and it has in fact proved difficult to provide a sound theoretical basis for the distinction. One proposal is that only truth-conditional features of meaning should form part of the lexical entry—that is, features which determine whether a statement containing the word in question is true or false of a particular situation.

However, this assumes that word meanings can in general be specified in terms of truth-conditional features, an assumption which is challenged by prototype theorists. Also, it is not intuitively clear why, for instance, 'can bark' should not be part of the meaning of 'dog,' even though *X is a dog* does not entail *X barks*. However, a claim that everything we know about dogs is legitimately part of the meaning of *dog* would also be intuitively suspect. One solution for a conceptualist is to say that the dictionary/lexical entry simply indicates which concept a word maps onto, but does not specify the contents of the concept.

4.2 Lexical Semantics and Syntax

An important and controversial topic is the relation between the meaning of a word and its grammatical properties. Historically, a wide spectrum of views can be noted, from proponents of the autonomy of syntax, to those who claim that all syntactic properties are semantically motivated. Probably the majority of linguists would take up some sort of intermediate position on this question: recent work by, for instance, Levin (see, e.g., Levin and Rapaport Hovav 1992), reveals a considerable amount of semantic determination of syntactic patterning; on the other hand, examples of differing syntactic behavior like the following are hard to see as semantically motivated:

I've finished the job.
I've completed the job.
I've finished.
I've completed.

A typical example of semantically motivated syntactic behavior is the following alternation:

John broke the stick.
The stick broke.

Verbs which show this behaviour must (a) denote a change of state in their transitive object/intransitive subject (*John saw the dog/* The dog saw*), and (b) must not encode any information other than the change of state, such as the nature or manner of the action performed by the causing agent (*John washed the car/* The car washed*).

5. Future Prospects

There is no doubt that considerable advances in the understanding of lexical semantics have been made in recent decades. But there remain nonetheless, areas where understanding is far from complete, and in which future progress can be anticipated. One such area is in a way the most fundamental of all, namely, what is it that distinguishes lexical semantics from phrasal or clausal semantics? An understanding of this issue might be expected to lead to a proper typology of word meanings, understood both interlingually and intralingually. This is currently in its infancy. Another vital area is that of polysemy, or more generally, variation in context. The varieties and mechanisms of contextual variation have received a great deal of attention; what is less well understood is how and why a reading crystallizes as a new, autonomous sense, with independent relational, combinatorial, and truth-conditional properties. Mention should also be made

of the eternal problem of how to specify what an individual sense of a word 'means.'

The study of lexical semantics is less 'autonomous' than that of, say, phonology, or syntax. Especially if one takes a cognitive linguistic view, there is no clear dividing line between lexical semantics and the study of conceptual categories within cognitive psychology; and advances in one field tend to have repercussions in the other. Advances in certain areas of psycholinguistics can also be expected to throw light on word meaning. For instance, there is currently a developing body of work on the time course of semantic activation. The meaning of a word is not activated all at once when a word is recognized, and the details of the activation process cannot fail to have relevance to our understanding of the internal structure of a word's meaning.

One area of practical concern, which is poised for a major take-off, but is currently held back by lexical semantic problems, is the automatic processing of natural language by computational systems. The main problems are the complexity of natural meanings and their contextual variability. The work currently being done in this area can be expected to spill over, not only into general lexicography, but also into the linguistic study of word meanings.

See also: Dementia, Semantic; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Lexicology and Lexicography; Lexicon; Semantic Knowledge: Neural Basis of; Semantics

Bibliography

Cruse D A 1986 *Lexical Semantics*. Cambridge University Press, Cambridge, UK

Cruse D A 2000 *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press, Oxford. UK

Levin B, Hovav Rappaport M 1992 Wiping the slate clean: A lexical semantic exploration. In: Levin B, Pinker S (eds.) *Lexical and Conceptual Semantics*. Blackwell, Oxford, UK, pp. 123–52

Lyons J 1963 *Structural Semantics*. Cambridge University Press, Cambridge, UK

Lyons J 1968 *Introduction to Theoretical Linguistics*. Cambridge University Press, Cambridge, UK

Lyons J 1977 *Semantics*. Cambridge University Press, Cambridge, UK

Nida E A 1975 *Componential Analysis of Meaning: An Introduction to Semantic Structures*. Mouton, The Hague, The Netherlands

Taylor J R 1989 *Linguistic Categorization: Prototypes in Linguistic Theory*. Clarendon Press, Oxford, UK

Ungerer F, Schmid H-J 1996 *An Introduction to Cognitive Linguistics*. Longman, London

Wierzbicka A 1996 *Semantics: Primes and Universals*. Oxford University Press, Oxford, UK

D. A. Cruse

Lexicology and Lexicography

1. Introduction

Each language has a lexicon and a grammar, i.e., a set of elementary expressions and a set of rules according to which complex expressions are constructed from simpler ones. Some of these rules form complex words; others operate beyond the boundaries of the word, thus producing phrases and sentences. These distinctions, familiar from the days of the Greek grammarians, are not always clear cut, for at least two reasons. First, the notion of 'word' is not very well-defined (see also *Word Classes and Parts of Speech*). Second, there are complex expressions, whose meaning is more or less predictable from the meaning of its components, whereas this is not true for other complex expressions. The former are said to be 'compositional,' whereas the latter are 'lexicalized'; slightly different terms to characterize this opposition are 'productive' vs. 'idiomatic,' and 'free' vs. 'fixed'; in each case, the distinction is gradual. Lexicalization is rarely observed for inflected words (a possible exception are 'participles' such as *crooked* in *a crooked street*), but very frequent for compound words, such as *landlord* or (*to*) *withdraw*, or phrases such as *to kick the bucket*, which has a compositional as well as a lexicalized reading. Do lexicalized expressions belong to the lexicon of a language or to its grammar? There is no straightforward answer; their form is complex and rule-based, their meaning is not. Therefore, it is useful to take the term 'lexicon' in a somewhat broader sense; it contains all elementary expressions (lexicon in the narrower sense) as well as those expressions which are compound in form but not accordingly in meaning (see also *Lexicon*). The scientific investigation of the lexicon in this sense is usually called lexicology; it includes, for example, the historical development of the lexicon, its social stratification, its quantitative composition or the way in which some subfield is encoded in lexical items (e.g., 'terminology of hunting,' 'verbs of movement'). Lexicography, by contrast, deals with the compilation of dictionaries. There is considerable overlap between both disciplines, and in fact, not all authors make such a terminological distinction.

2. The Lexicon

The lexicon of a language is stored primarily in the head of its speakers, and for most of the history of mankind, it was only stored there. We do not know what form the 'mental lexicon' has (see also *Psycholinguistics*: *Overview*). There is agreement, however, that it consists of individual lexical units which are somehow interrelated to each other. There is no generally accepted term for lexical units. The familiar

term 'word' is both too broad and too narrow; one would not want to consider *goes* as a lexical unit, although it is a word, whereas expressions such as (*to*) *cut up* or *red herring* are lexical units but consist of several words. Other terms occasionally found are 'lexeme,' 'lemma,' or 'lexical entry,' but since these are also used in other ways, it is probably best to speak of lexical units.

It is important to distinguish between a lexical unit and the way in which it is named. The word *house* in a dictionary, followed by all sorts of explanations, is not the lexical unit—it is a name for such a unit. The lexical unit itself is a bundle of various types of properties. These include:

(a) phonological properties, which characterize how the lexical unit is pronounced; they include sounds, syllabic structure, lexical accent and, in some languages, lexical tone;

(b) graphematic properties, which characterize how the lexical unit is written (see also *Spelling*);

(c) morphosyntactic properties, which characterize how the unit can become part of more complex expressions; typically, they concern inflectional paradigm, word class, government relations, and others;

(d) semantic properties, which concern the 'lexical meaning' of the unit, i.e., the contribution which it makes to the meaning of the construction in which it occurs.

Some of these properties may be absent. This is most obvious for graphematic properties, since not all languages are written. There are a few lexical units without lexical meaning, such as the expletive *there* in English. Many linguists also stipulate 'zero elements,' i.e., units with morphosyntactic and semantic properties but without phonological properties (such as 'empty pronouns'); but these are normally treated in the grammar rather than in the lexicon.

Whereas these four types of properties are the defining characteristics of a lexical unit, other information may be associated with it, for example, its etymology, its frequency of usage, its semantic counterpart in other languages, or encyclopedic knowledge (thus, it is one thing to know the meaning of *bread* and a different thing to know various sorts of bread, how it is made, its price, its role in the history of mankind, etc.).

The lexical units of a lexicon are in many ways interrelated. They may share some phonological properties (for example, they may rhyme with each other), they may belong to the same inflectional paradigm, they may have the opposite meaning ('antonyms,' such as *black* and *white*), approximately the same meaning ('synonyms,' such as *to begin* and *to start*), or when complex in form they may follow the same construction pattern. Lexicological research is often oriented towards these interrelations, whereas lexicography tends to give more weight to the lexical unit in itself. In general, there is much more lexicographical than lexicological work (for a survey of the latter, see Schwarze and Wunderlich 1985); in fact, if there is any piece of linguistic description for some language, it is probably an elementary bilingual dictionary. The depth of this work varies massively not only across languages, but also with respect to the particular lexical properties. Whereas the phonological, graphematic and morphosyntactic features of the lexicon in Latin, English, French, and some dozen other languages with a comparable research tradition are fairly well described, there is no theoretically and empirically satisfactory analysis of the semantics of the lexicon for any language whatsoever. This has three interrelated reasons. First, there is no well-defined descriptive language which would allow the researcher to represent the meaning of some lexical unit, be it simple or compound; the most common practice is still to paraphrase it by an expression of the same language. Second, there is no reliable and easily applicable method of determining the lexical meaning of some unit; the most common way is to look at a number of occurrences in ongoing text and to try to understand what it means. Third, the relation between a particular form and a particular meaning is hardly ever straightforward; this is strikingly illustrated by a look at what even a medium-sized English dictionary has to say about the meaning of, for example, *on*, *sound*, *eye* or (*to*) *put up*. As a rule, there is not just one lexical meaning, but a whole array of uses which are more or less related to each other. This is not merely a practical problem for the lexicographer; it also casts some doubt on the very notion of 'lexical unit' itself (see also *Lexical Semantics*).

3. *Making Dictionaries*

Lexicographers often consider their work to be more of an art or a craft than a science (see, e.g., Landau 1984, Svensén 1993). This does not preclude a solid scientific basis, but it reflects the fact that their concrete work depends largely on practical skills such as being 'a good definer,' on one hand; and that it is to a great extent determined by practical, often commercial, concerns, on the other. Dictionaries are made for users, and they are intended to serve specific purposes. Their compilation requires a number of practical decisions.

3.1 *Which Lexical Units are Included?*

Languages are neither well-defined nor uniform entities; they change over time, and they vary with factors such as place, social class, or area talked about. A great deal of this variation is lexical. It is not possible nor would it be reasonable to cover this wealth in a single dictionary. Large dictionaries contain up to 300,000 'entries'; since idiomatic expressions are usually listed under one of their components (such as

to kick the bucket under (*to*) *kick*), they contain many more lexical units, perhaps up to 1 million. But, even so, they are by no means exhaustive. The second edition of the *Deutsches Wörterbuch* (see Sect. 5), the largest dictionary of German, covers less than 25 per cent of the lexical units found in the sources, and these sources are quite restricted themselves.

3.2 Which Lexical Properties are Described?

Just as it is impossible to include all lexical units of a language in a dictionary, it is neither possible nor desirable to aim at a full description of those which are included. Since a dictionary is normally a printed book, the graphematic properties of the unit (its 'spelling') are automatically given. Among the other defining properties, meaning is traditionally considered to be most important. Samuel Johnson's dictionary from 1755 (see Sect. 5) defines 'dictionary' as 'A book containing the words of any language in alphabetical order, with explanations of their meaning.' But Johnson also noted which syllable carries the main stress, and he gave some grammatical hints. In general, however, information on phonological properties was rare up to the end of the nineteenth century, and information on grammatical properties is usually still very restricted in nonspecialized dictionaries. But there are, of course, also dictionaries which specifically address these properties as well as some of the nondefining properties associated with a lexical entry, such as its origin (etymological dictionary) or, above all, its equivalent in other languages ('bilingual dictionary').

3.3 What is the Description Based Upon?

Usually, two types of sources are distinguished: 'primary sources' are samples of text in which the unit is used, 'secondary sources' refers to prior work of other lexicographers (and lexicologists). In fact, there is a third source, normally not mentioned in the theory of lexicography (sometimes called 'meta-lexicography'): this is the lexicographer's own knowledge of the language to be described, including his or her views on what is 'good' language. In practice, the bulk of a new dictionary is based on older dictionaries. This is always immoral and often illegal, if these are simply copied; but on the other hand, it would be stupid and arrogant to ignore the achievements of earlier lexicographers.

3.4 How is the Information Presented?

A dictionary consists of lexical entries arranged in some conventional order. Normally, an entry combines several lexical units under a single 'head word'; thus, all lexical units which include the word *put* may be listed under this head word, forming a kind of nest with an often very complex microstructure. We are used to alphabetically-ordered dictionaries; but there are other possibilities, for example, by thematic groups or by first appearance in written documents. Languages without alphabetic writing require different principles; in Chinese, for example, entries are usually arranged by subcomponents of the entire character and by the number of strokes.

These four questions can be answered in very different ways, resulting in very different types of dictionaries (see the survey in Hausmann et al. 1991, pp. 968–1573).

4. History

The first lexicographic documents are lists of Sumerian words (up to 1400) with their Akkadian equivalents, written in cuneiform script on clay tablets about 4,700 years ago. The practice compiling such word lists was continued throughout Antiquity and the Middle Ages; thus, the oldest document in German, the *Abrogans* (written around 765), is an inventory of some Latin words with explanations in German. Usually, these 'glossaries' did not aim at a full account of the lexicon; they simply brought together a number of words which, for one reason or another, were felt to be 'difficult,' and explained them either by a more familiar word in the same language or by a translation. Words were ordered alphabetically, by theme, or not at all. But there are also more systematic attempts, such as the *Catholicon*, a mixture of encyclopedia and dictionary which, compiled around 1250, was the first printed lexical work in Europe (Mainz 1460).

In the sixteenth century, two developments led to major changes. The first of these was the invention of printing by Gutenberg. By 1500, virtually all classical authors were available in print, thus offering a solid basis for systematic lexical accounts of Latin and Greek, such as Calepinus' *Dictionarium* (1502), soon to be followed by two early masterpieces: Robert Etienne's *Dictionarium seu Latinae Linguae Thesaurus* (Paris 1531) and Henri Etienne's *Thesaurus Graecae Linguae* (Paris 1572). The second major development was the slow but steady rise of national languages. Since early Italian, French, English, or German were hardly codified, a major aim of the first dictionaries in these languages was to give them clear norms. In some countries, national Academies were founded to this end. The outcome were dictionaries with a strongly normative, often puristic, stance, such as the *Vocabulario degli Academici della Crusca* (Venice 1612), the *Dictionnaire de l'Académie Française* (Paris 1694) and the *Diccionario de autoridades publicado por la Real Academia Espanola* (1726–1739).

The bulk of lexicographic work, however, was always done by enterprising publishers and engaged

individuals, such as Dr Samuel Johnson. Helped by six assistants, he produced *A Dictionary of the English Language* (London 1755), the first scholarly description of the English vocabulary, in less than eight years. It surpassed all its predecessors, including Bailey's *Dictionarium Britannicum* from 1736, which Johnson took as his point of departure, by the systematic use of quotations, taken from the best writers, and by his brilliant, sometimes somewhat extravagant, definitions (not everybody would dare to characterize patriotism as 'the last refuge of a scoundrel'). Less known, much less witty, but broader in coverage is the first comprehensive dictionary of German, Johann Christoph Adelungs *Versuch eines vollständigen grammatisch-kritischen Wörterbuchs der hochdeutschen Mundart* (Leipzig 1774–86).

The rise of historical-comparative linguistics in the early nineteenth century led to an enormous increase in grammatical and lexical knowledge. The first dictionary which tried to cover this knowledge was the *Deutsches Wörterbuch* by Jacob Grimm and (to a much lesser extent) his brother Wilhelm Grimm. Its first fascicle appeared in 1852, after about ten years of preparatory work, in which the Grimms were helped by about 100 scholars providing excerpts ('covering my desk like snowflakes,' Jacob Grimm). At that time, it was already clear that the original plan of 6–7 volumes, to be finished within 10–12 years, was unrealistic. The Grimms finished only letters A–(most of) F, and the final folio volume (of altogether 32) appeared in 1960. This long duration, as well as the varying talents and preferences of the contributors, has led to many inconsistencies; some entries got out of balance (no less than 60 pages are devoted to the single word *Geist*); still, it is an incommensurable source of lexical information.

The work of the Grimms inspired a number of similar ventures, such as Emile Littré's masterly *Dictionnaire de la langue française* (1863–1873), which is much shorter, but also much more consistent: Matthias de Vries and his numerous successors' voluminous *Woordenboek der Nederlandsche Taal* (1864–1998), and finally *A New English Dictionary on a Historical Basis* (1884–1928), generally referred to as the 'Oxford English Dictionary' (OED). It was initiated in 1857 by the philologist and churchman Richard Trench; in 1860, members of the Philological Society started to collect excerpts; in 1879, the Clarendon Press appointed James Murray as the Principal Editor. The first fascicle appeared in 1882, and the whole work was completed in 1928, 13 years after Murray's death. More than 200 scholars were involved in its production, more than 2,000 people are known to have contributed excerpts. The OED is not without flaws, even in its revised edition, which appeared in 1989 in print and in 1992 on CD-ROM; but among all attempts to describe the lexicon of a language, it comes closest to falsify what Dr Johnson stated in the preface to his own dictionary: 'Every other author may aspire to praise; the lexicographer can only hope to escape reproach.' (For a comprehensive survey of lexicographic work across languages, see Hausmann et al. 1991, pp. 1679–2710, 2949–3119).

5. *The Use of Computers*

We tend to think of dictionaries as the normal, if not the only possible way to compile and to present lexical information. But the dawn of the computer has provided us with a very different and in many ways more efficient tool. Computers can be used in at least three ways in lexicography. It is possible to transfer an existing dictionary to a computer, as has often been done over the last 20 years. Such a transfer offers several advantages: search is faster and more exhaustive; it is easier to revise and update the dictionary; and it is possible to add information not available in book format, for example, spoken sound instead of phonetic transcriptions. But essentially, the format of the printed book is maintained. Next, computers are a powerful tool in the production of a new dictionary. Rather than having a number of people read through books and newspapers and make excerpts of all occurrences which look interesting, it is now possible to compile huge text corpora that cover all varieties of a language, to scan these texts for all occurrences of words or word combinations, to sort these occurrences by various criteria, to link them to other occurrences, to add as much context as needed, etc. (see *Corpus Linguistics*).

The OED, is based on about 5 million excerpts, mostly handwritten on paper slips. A computer can easily process corpora of several hundred million words, i.e., several hundred million occurrences; new sources can rapidly be added. This allows a much broader and much more representative coverage of a lexicon than ever. But electronic corpora only provide the raw material; it still awaits lexical analysis. This analysis can be facilitated by computer tools, also; but no computer can tell us what a word means in a particular context. But even if only one minute is devoted to each occurrence in a one-hundred million corpus, it would take 10 lexicographers 100 years to go through it. This means that printed dictionaries can never reflect the wealth of information accessible in large corpora, since they presuppose that the lexicographer has finished the analysis. Therefore, the only way to make full use of large corpora is by means of lexical retrieval systems. They consist of (a) a computer-accessible and expandable corpus, (b) a set of tools, which allow, for example, not only the search for certain items but also statistical analysis or the determination of the first occurrence, and (c) a selective but steadily proceeding lexical analysis of the corpus. Thus, it is possible to add spoken forms in various dialects, information about word classes, or

the semantic analysis of some subset of lexical units, say all prepositions or all morphologically simple verbs. Similarly, translation equivalents can be added. Unlike printed dictionaries, such a lexical retrieval system will never come to an end, it is steady work-in-progress to which many can contribute and which will give us a deeper and broader understanding of the lexicon than any other method.

See also:; Corpus Linguistics; Lexical Access, Cognitive Psychology of; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Lexical Semantics

Bibliography

Hausmann F-J, Reichmann O, Wiegand H E, Zgusta L (eds.) 1991 *Wörterbücher—Dictionaries—Dictionnaires*. De Gruyter, Berlin

Landau S A 1984 *Dictionaries: The Art and Craft of Lexicography*. Scribner, New York

McArthur T 1986 *Worlds of Reference. Lexicography, Learning and Language from the Clay Tablet to the Computer*. Cambridge University Press, Cambridge, UK

Schwarze C, Wunderlich D (eds.) 1985 *Handbuch der Lexikologie*. Athenäum, Königstein, Germany

Svensén B 1993 *Practical Lexikography. Principles and Methods of Dictionary Making*. Oxford University Press, Oxford, UK

W. Klein

Lexicon

1. The Notion of Lexicon

The lexicon is standardly viewed as a listing of all the morphemes of a language, with information indicating how each morpheme behaves in the components of grammar involving phonology, syntax, and semantics. In no small part, the shape and character of grammar is determined by what the lexicon contains for these other grammatical devices. Nevertheless, both historically and conventionally, the lexicon has been seen as the passive module in the system of grammar.

More recently, the model of the lexicon has undergone significant revision and maturation. In particular, two trends have driven the architectural concerns of lexical researchers: (a) a tighter integration of compositional operations of syntax and semantics with the lexical information structures that bear them; and (b) a serious concern with how lexical types reflect the underlying ontological commitments of the grammar. In the process, the field has moved towards addressing more encompassing problems in linguistic theory, such as those below:

(a) How can we explain the polymorphic nature of language?

(b) How can we capture the creative use of words in novel contexts?

(c) How can semantic types predictably map to syntactic representations?

(d) What are the 'atoms' of lexical knowledge, if they exist at all?

In this article, we first review the conventional view of the lexicon and then contrast this with the theories of lexical information that have emerged since around 1990.

By all accounts, the conventional model of the lexicon is that of a database of words, ready to act in the service of more dynamic components of the grammar. This view has its origins squarely in the generative tradition (Chomsky 1955) and has been an increasingly integral part of the concept of the lexicon ever since. While the 'Aspects' model of selectional features restricted the relation of selection to that between lexical items, work by Jackendoff (1972) and McCawley (1968), showed that selectional restrictions must be available to computations at the level of derived semantic representation rather than at deep structure. But where did this view come from? In order to understand both the classical model of the lexicon as a database and the current models of lexically encoded grammatical information, it is necessary to appreciate the structuralist distinction between 'syntagmatic processes' and 'paradigmatic systems' in language. The lexicon has emerged as the focal point communicating between these two components, and can be seen as a hook which links the information at these two levels. One can go further still and view the elements of the lexicon as not just the building blocks for the more active components of the grammar, but also as actively engaging the building principles themselves.

While syntagmatic processes refer to the influence of horizontal elements on a word or phrase, paradigmatic systems refer to vertical substitutions in a phrasal structure. Syntagmatics evolved into the theory of abstract syntax while paradigmatics was all but abandoned in generative linguistics. In an early discussion of syntagmatic dependencies, Hjelmslev (1943) uses the term 'selection' explicitly in the modern sense and notes the importance of integrating paradigmatic systems with the syntagmatic processes they participate in. For Hjelmslev, there are two possible types of relations that can exist between elements in a syntagmatic process: 'interdependence' and 'determination', the latter of which is related to the notion of selectional restriction as developed by Chomsky (1965). As Cruse (1986) notes 'One reason that selectional restrictions were not integrated into mechanisms of grammatical selection and description in the 1970s and 1980s is that, if they are imposed correctly, the grammar is forced to model two computations:

(a) the entailment relations between selectional

restrictions as features must be modeled formally, in order to contribute to the computation of a syntactic description;

(b) the manner in which selectional features or constraints contribute to the determination of the meaning of expressions must be enriched in order to exploit these very features.'

Recently, with the convergence of several areas in linguistics (lexical semantics, computational lexicons, and type theories) several models for the determination of selection have emerged which actively integrate these central syntagmatic processes into the grammar, by making explicit reference to the paradigmatic systems which allow for grammatical constructions to be partially determined by selection. Examples of this approach are Generative Lexicon Theory (Bouillon and Busa 2001; Pustejovsky 1995), and to a certain extent, Construction Grammar (Goldberg 1995), CCG (Steedman 1997), and HPSG (Pollard and Sag 1994).

These recent theoretical developments have led to a new direction in lexical design. Rather than restricting the scope of the lexicon to that of a passive database, current frameworks have re-architected the relationship between syntactic and semantic representations and the underlying recurrence relations that generate them. These developments have helped to characterize the approaches to lexical design in terms of a hierarchy of semantic expressiveness. There are at least three such classes of lexical description, which are defined (cf. Pustejovsky 1995 for discussion):

(a) Sense Enumerative Lexicons: lexical items have a single type and meaning, and ambiguity is treated by multiple listings of words.

(b) Polymorphic Lexicons: lexical items are active objects, contributing to the determination of 'meaning in context' under well-defined constraints.

(c) Unrestricted Sense Lexicons: meanings of lexical items are determined mostly by context and conventional use. Few, if any, restrictions are imposed on how a word may refer.

Although there have been proponents for each class of lexical description defined here, the most promising direction seems to be a careful and formal elucidation of the polymorphic lexicons, and this will form the basis of our subsequent discussion of both the structure and content of lexical entries.

2. *The Structure of a Lexical Entry*

As mentioned in Sect. 1, it is generally agreed that there are three components to a lexical item: phonological, syntactic, and semantic information. In this entry, we mainly focus on the manner in which syntactic and semantic representations are encoded in the lexical entry.

There are two types of syntactic knowledge associated with a lexical item: its 'category' and its 'subcategory.' The former includes traditional classifications of both the major categories, such as noun, verb, adjective, adverb, and preposition, as well as the minor categories, such as adverbs, conjunction, quantifier elements, and determiners (see *Word Classes and Parts of Speech*).

Knowledge of the subcategory of a lexical item is typically information that differentiates categories into distinct, distributional classes. This sort of information may be usefully separated into two types, 'contextual features' and 'inherent features.' The former are features that may be defined in terms of the contexts in which a given lexical entry may occur. Subcategorization information marks the local syntactic context for a word. It is this information that ensures that the verb 'devour,' for example, is always transitive in English, requiring a direct object; the lexical entry encodes this requirement with a subcategorization feature specifying that an NP appear to its right. Another type of context encoding is collocational information, where patterns that are not fully productive in the grammar can be tagged. For example, the adjective 'heavy' as applied to 'drinker' and 'smoker' is collocational and not freely productive in the language (Mel'čuk 1988). 'Inherent features,' on the other hand, are properties of lexical entries that are not easily reduced to a contextual definition, but rather refer to the ontological typing of an entity. These include such features as count/mass (e.g., 'pebble' vs. 'water,' abstract, animate, human, physical, and so on.

Semantic information can also be separated into two categories: 'base semantic typing' and 'selectional typing.' While the former identifies the semantic class that a lexical item belongs to (such as entity, event, property), the latter class specifies the semantic features of arguments and adjuncts to the lexical item.

2.1 *Word Classes and Typing Information*

There are two major approaches to classifying lexical items by their type: syntactic and semantic. (Another influential tradition for verb classification, which we will not discuss in detail here, is a more descriptive approach to word classes, where membership is defined on the basis of grammatical behavior and verbal valency alternation, such as that elaborated on and compiled in Levin (1993)

One obvious way to organize lexical knowledge, be it syntactic or semantic, is by means of lexical inheritance mechanisms. In fact, much recent work has focused on how to provide shared data structures for syntactic and morphological knowledge (Flickinger et al., 1985). Evans and Gazdar (1990) provide a formal characterization of how to perform inferences in a language for multiple and default inheritance of linguistic knowledge. The language developed for that purpose, DATR, uses value-

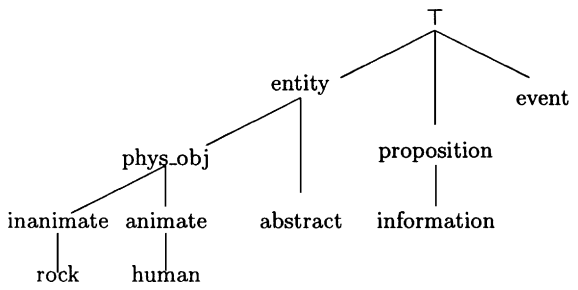

Figure 1
Fragment of a type hierarchy

terminated attribute trees to encode lexical information. Briscoe et al. (1993) describe a rich system of types for allowing default mechanisms into lexical type descriptions.

Type structures can express the inheritance of syntactic features (Sanfilippo 1993) as well as the relationship between more conventional taxonomic information, such as that shown in Sect. 2.2 (cf. Copestake and Briscoe 1992, Pustejovsky 1995, Pustejovsky and Boguraev 1993) and Fig. 1.

Given a semi-lattice of types such as this, the lexical items in the language can be associated with a much richer system of differentiated semantic classes. Verbs may also be structured in hierarchical relations, as done in HPSG, LFG, and other lexical frameworks (cf. Pollard and Sag 1994, Alsina 1992, Koenig and Davis 1999).

2.2 Argument Structure

Once the base syntactic and semantic typing for a lexical item has been specified, its subcategorization and selectional information must be encoded in some form. There are two major techniques for representing this type of knowledge:

(a) associate 'named roles' with the arguments to the lexical item;

(b) associate a logical decomposition with the lexical item; meanings of arguments are determined by how the structural properties of the representation are interpreted (cf. Hale and Keyser 1993, Levin and Rappaport 1995, Pustejovsky 1995).

One influential way of encoding selectional behavior is the theory of thematic relations (cf. Gruber 1976, Jackendoff 1972). Thematic relations are typically defined as partial semantic functions of the event being denoted by the verb or noun, and behave according to a predefined calculus of roles relations (e.g., Carlson 1984, Chierchia 1989, Dowty 1989). For example, semantic roles such as agent, theme, and goal can be used to partially determine the meaning of a predicate, when they are associated with the grammatical arguments to a verb.

The theory of argument structure as developed by Williams (1981), Grimshaw (1990), and others can be seen as a move towards a more minimalist description of semantic differentiation in the verb's list of parameters. The argument structure for a word can be seen as the simplest specification of its semantics, indicating the number and type of parameters associated with the lexical item as a predicate. For example, the verb 'build' can be represented as a predicate taking two arguments, while the verb 'give' takes three arguments.

(1) (a) build (x, y)
(b) give (x, y, z)

What originally began as the simple listing of the parameters or arguments associated with a predicate has developed into a sophisticated view of the way arguments are mapped onto syntactic expressions. Williams' (1981) distinction between 'external' (the underlined arguments) and 'internal' arguments and Grimshaw's proposal for a hierarchically structured representation (cf. Grimshaw (1990) provide us with the basic syntax for one aspect of a word's meaning. Similar remarks hold for the argument list structure in HPSG (Pollard and Sag 1994) and LFG (Bresnan 1994).

The interaction of a structural argument list and a rich system of types, such as that presented, provides a mechanism for semantic selection that overcomes the difficulties mentioned in Section 2. The most direct impact of semantic type systems on syntactic subcategorization can be seen with the analysis of a simple example.

(2) (a) The man/the rock fell.
(b) The man/*the rock died.

Returning to the example in (2), consider how the selectional distinction for the feature [+/− animacy] is modeled. For the purpose of illustration, the arguments of a verb will be represented in a list structure, where each argument is identified as being typed with a specific value.

(3) (a) $\begin{bmatrix} \text{fall} \\ \text{ARGSTR} = [\text{ARG}_1 = \textbf{x:phys_obj}] \end{bmatrix}$

(b) $\begin{bmatrix} \text{die} \\ \text{ARGSTR} = [\text{ARG}_1 = \text{x:animate}] \end{bmatrix}$

In the sentences in (2), it is intuitively clear how rocks can't die and men can but it is still not obvious how this inference is computed, given what we would assume are the types associated with the nouns 'rock' and 'man' respectively. What accomplishes this computation is a rule of subtyping, Θ, that allows the type associated with the noun 'man' (i.e., human) to also be accepted as the type animate, which is what the predicate 'die' requires of its argument as stated in (3b)

(4) $\Theta\ [\textit{human} \sqsubseteq \textit{animate}]: \textit{human} \rightarrow \textit{animate}$

The rule Θ applies, since the concept human is subtyped under animate in the type hierarchy. Parallel considerations rule out the noun 'rock' as a legitimate

argument to 'die' since it is not subtyped under animate. Hence, one of the concerns given for how syntagmatic processes can systematically keep track of which 'selectional features' are entailed and which are not is partially addressed by such lattice traversal rules as the one presented here.

Selection can also be employed to solve the problem of polymorphism, where the same lexical item can appear in multiple syntactic contexts, as illustrated in (5) below.

(5) (a) Mary began to read the novel.
(b) Mary began reading the novel.
(c) Mary began the novel.

Generative Lexicon Theory handles such examples by the use of semantic selection (cf. Pustejovsky 1995) and canonical syntactic mapping rules, specifying how semantic types correspond to syntactic expressions. Here, the verb 'begin' is typed as taking an event description as its internal argument, begin (Ind, $\mathscr{E}$), which can be realized in one of the three syntactic forms shown. This illustrates how a generative system of type operations can account for polymorphic behavior of selection in the syntax.

2.3 Decomposition and Event Structure

The second major approach to the specification of lexical knowledge is that taken by decompositional theories. Most of this research has focused on the lexical specification for verbs, their arguments, and their syntactic behavior. Some recent work, however, has been done on noun semantics as well (Busa 1996). In this section, we examine the motivations for lexical decomposition in linguistic theory, and the various proposals that have emerged on how to encode lexical knowledge as structured forms. We then relate this to the manner in which verbs refer to events, since this directly impacts the nature of the lexical decomposition structure.

Since Davidson (1967), events have played an increasingly important role in the determination of verb meaning. While early researchers on decompositional models (Dowty 1979, Lakoff 1965, McCawley) made no ontological commitments to events in the semantics for verbs, a new synthesis has emerged in recent years which attempts to model verb meanings as complex predicative structures with rich event structures (cf. Hale and Keyser 1993, Parsons 1990, Pustejovsky 1991, Tenny 1994). This research has developed the idea that the meaning of a verb can be analyzed into a structured representation of the event that the verb designates, and has furthermore contributed to the realization that verbs may have complex, internal event structures. Recent work has converged on the view that complex events are structured into an inner and an outer event, where the outer event is associated with causation and agency, and the inner event is associated with telicity (completion) and change of state. Under this view, a canonical accomplishment predicate as in 'John sliced the bread,' for example, can be represented as composed of an inner and an outer event. The inner event is the telic event in which the bread undergoes a change of state, and the outer event is the event in which John acts agentively (to do whatever is involved in the act of slicing). Since the outer event causes the inner one, it is associated lexically with causation.

Although there is a long tradition of analyzing causation as a relation between two events in the philosophical (cf. Davidson 1967) and psychological literature (cf. Miller and Johnson-Laird 1976, Schank 1973) in contemporary models of natural language semantics this idea has only recently been adopted. For example, Carter (1976), one of the earlier researchers in this area, represents the meaning of the verb 'darken' as follows:

(6) x CAUSE ((y BE DARK) CHANGE))

The predicate CAUSE is represented as a relation between a causer argument x and an inner expression involving a change of state in the argument y, although there is an intuition that the cause relation involves a causer and an event. Carter does not make this commitment explicitly. Levin and Rappaport (1988) follow a similar strategy, with a CAUSE predicate relating a causer argument and an inner expression involving a change of state in the argument y. The change of state is represented with the predicate BECOME:

(7) wipe the floor clean:
x CAUSE [y BECOME (AT) z] BY [x 'wipe' y]]
x CAUSE [floor BECOME (AT) clean] BY
[x 'wipe' floor]

The work of Levin and Rappaport, building on Jackendoff's Lexical Conceptual Structures, has been influential in articulating the internal structure of verb meanings (see Levin and Rappaport 1995).

Jackendoff (1990) develops an extensive system of what he calls 'Conceptual Representations' which parallel the syntactic representations of sentences of natural language. These employ a set of canonical predicates including CAUSE, GO, TO, and ON, and canonical elements including Thing, Path, and Event. These approaches represent verb meaning by decomposing the predicate into more basic predicates. This work owes obvious debt to the innovative work within generative semantics, as illustrated by McCawley's (1968), analysis of the verb 'kill.' Recent versions of lexical representations inspired by generative semantics can be seen in the Lexical Relational Structures of Hale and Keyser 1993, where syntactic tree structures are employed to capture the same elements of causation and change of state as in the representations of Carter, Levin and Rapoport, Jackendoff, and Dowty.

Pustejovsky (1988, 1991) extends the decompositional approach presented in Dowty (1979) by explicitly reifying the events and subevents in the

predicative expressions. Unlike Dowty's treatment of lexical semantics, where the decompositional calculus builds on prepositional or predicative units (as previously discussed), a 'syntax of event structure' makes explicit reference to quantified events as part of the word meaning. Pustejovsky further introduces a tree structure to represent the temporal ordering and dominance constraints on an event and its subevents. For example, a predicate such as 'build' is associated with a complex event such as that shown in (8):

(8) $[_{\text{transition}}[e_1\text{: PROCESS}]\ [e_2\text{: STATE}]]$

The process consists of the building activity itself, while the state represents the result of there being the object built. Grimshaw (1990) adopts this theory in her work on argument structure, where complex events such as 'break' are given a similar representation. In such structures, the process consists of what x does to cause the breaking, and the state is the resultant state of the broken item. The process corresponds to the outer causing event as discussed, and the state corresponds in part to the inner change of state event. Both Pustejovsky and Grimshaw differ from the authors above in assuming a specific level of representation for event structure, distinct from the representation of other lexical properties. Furthermore, they follow Higginbotham (1985) in adopting an explicit reference to the event place in the verbal semantics.

2.4 Qualia Structure

Thus far, we have focused on the lexical information associated with verb entries. All of the major categories, however, are encoded with syntactic and semantic feature structures that determine their constructional behavior and subsequent meaning at logical form. How this is accomplished, of course, varies from theory to theory.

In Generative Lexicon Theory, it is assumed that word meaning is structured on the basis of four generative factors, or 'qualia roles,' that capture how humans understand objects and relations in the world and provide the minimal explanation for the linguistic behavior of lexical items (these are inspired in large part by Moravcsik's (1975, 1990) interpretation of Aristotelian *aitia*).

formal: the basic category that distinguishes the object within a larger domain;

constitutive: the relation between an object and its constituent parts;

telic: its purpose and function;

factors involved in the object's origin or 'coming into being.'

Qualia structure is at the core of the generative properties of the lexicon, since it provides a general strategy for creating new types. For example, consider the properties of nouns such as 'rock' and 'chair.' These nouns can be distinguished on the basis of semantic criteria which classify them in terms of general categories such as natural_kind, artifact_object. Although very useful, this is not sufficient to discriminate semantic types in a way that also accounts for their grammatical behavior. A crucial distinction between 'rock' and 'chair' concerns the properties which differentiate natural_kinds from 'artifacts': functionality plays a crucial role in the process of individuation of artifacts, but not of natural kinds. This is reflected in grammatical behavior, whereby 'a good chair,' or 'enjoy the chair' are well-formed expressions reflecting the specific purpose for which an artifact is designed, but 'good rock' or 'enjoy a rock' are semantically ill-formed since for 'rock' the functionality (i.e., TELIC) is undefined. Exceptions exist when new concepts are referred to, such as when the object is construed relative to a specific activity, such as in 'The climber enjoyed that rock'; 'rock' itself takes on a new meaning, by virtue of having telicity associated with it, and this is accomplished by integration with the semantics of the subject NP. Although 'chair' and 'rock' are both physical_object, they differ in their mode of coming into being (i.e., AGENTIVE): artifacts are man-made, 'rocks' develop in nature. Similarly, a concept such as 'food' or 'cookie' has a physical manifestation or denotation, but also a functional grounding, pertaining to the relation of 'eating.' These apparently contradictory aspects of a category are orthogonally represented by the qualia structure for that concept, which provides a coherent structuring for different dimensions of meaning.

For relations, the qualia act in a similar capacity to thematic relations, but where the individual qualia are possibly associated with entire event descriptions, and not just individuals. For discussion, see Pustejovsky (1995).

3. Consequence of Lexical Design

Considering the potential range of information that can be represented lexically results in a reconceptualization of what a lexicon is, so that the very design of the grammar is significantly impacted. Furthermore, our current understanding of psychological and computational properties of language processing suggests that the resources available for lexical storage and access are considerably higher than originally imagined by early grammatical theorists. The consequence is that, what had originally been accomplished in syntax, because of the combinatoric properties inherent in production rules, can be handled by the lexicon itself. The various approaches to lexical encoding can be analyzed in terms of two parameters.

(a) Precompiling the information into lexical items forms;

(b) Computing or generating new forms or senses during the compositional process.

Typically, idioms are presented as examples of precompiled lexical entries. But some theories have adopted this idea as fundamental for the entire

compositional operation. The best example of this is Combinatory Categorical Grammar (CCG) (Steedman 1997). CCG has recently been articulated in enough detail to handle most of the major linguistic phenomena using a library of precompiled lexical types, together with the combinatoric rules of categorical syntax. If the grammar utilizes representations with such non-local dependencies, then there must be additional mechanisms for unifying these representations; these are provided in the form of function composition rules and lexical rules (Schabes et al. 1988).

Lexical rules have been invoked in HPSG, as well, to explain the relationship between the various senses for lexical items, from grinding and packaging operations (such as that relating the animal and food senses of 'chicken' and 'lamb,' to the relation between logically polysemous items, such as 'book' information and physical object), and 'lecture' (information and event) (cf. Copestake and Briscoe 1992). In Generative Lexicon such relations are explicitly represented in context (cf. Johnston 1995). It is very likely, however, that language makes use of both types of devices, namely complex types such as dot objects, as well as the application of lexical rules. Regardless, both types of devices must be seriously constrained by the grammar in order not to overgenerate unwanted forms and interpretations.

Considered independently of the issue of precompiled vs. generated forms and senses, there is no question that the mental lexicon is large, containing arguably up to 400,000 lexical entries. This is based on fairly conservative estimates of speaker competence with active and passive vocabularies. For example, an average speaker lexicon might contain at least 5,000 distinct verbs, 30,000 distinct nominal forms, and over 5,000 adjectives. Combine this with an additional 10,000 compound forms and at least 300,000 distinct proper names. Obviously, the psychological (and hence computational) demands of these classes are quite distinct. There are two dimensions that can help us distinguish these classes: (a) the degree of combinatoric (functional) complexity of the lexical item; and (b) whether the lexical item is part of active or passive lexical knowledge. Most closed class items, for example, are functionally complex, as are many open class verbs and relational nouns. The majority of the open class items will also involve a fair amount of information regarding combinatoric possibilities. The class of names, however, is unique in that, although it is by far the largest class of lexical items, it is the least demanding in terms of computational resources.

In conclusion we see that the lexicon is neither a mere listing of morphemes in the language, nor a database of items passively waiting in the service of grammatical processes. Rather, the lexicon is a dynamic and active system of grammar, incorporating as well as dictating essential components of syntactic and semantic composition and interpretation.

See also: Generative Grammar; Lexical Access, Cognitive Psychology of; Lexical Functional Grammar; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Lexical Semantics; Lexicology and Lexicography; Syntax–Semantics Interface

Bibliography

Alsina A 1992 On the argument structure of causatives. *Linguistic Inquiry* **23**(4): 517–55

Baker M 1988 *Incorporation: A Theory of Grammatical Function Changing*. University of Chicago Press, Chicago

Bresnan J 1994 Locative Inversion and the architecture of universal grammar. *Language* **70**(1): 2–31

Boguraev B, Briscoe E 1989 *Computational Lexicography for Natural Language Processing*. Longman, Harlow and London

Boguraev B, Pustejovsky J 1996 *Corpus Processing for Lexical Acquisition*. Bradford Books/MIT Press, Cambridge, MA

Bouillon P, Busa F 2001 *The Syntax of Word Meaning*. Cambridge University Press

Briscoe T, de Paiva V, Copestake A (eds.) 1993 *Inheritance, Defaults, and the Lexicon*. Cambridge University Press, Cambridge, UK

Busa F 1996 *Compositionality and the Semantics of Nominals*. Ph.D Dissertation, Brandeis University

Carlson G 1984 Thematic roles and their role in semantic interpretation. *Linguistics* **22**: 259–79

Chomsky N 1955 *The Logical Structure of Linguistic Theory*. University of Chicago Press, Chicago

Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Copestake A, Briscoe E 1992 Lexical operations in a unification-based framework. In: Pustejovsky J, Bergler S (eds.) *Lexical Semantics and Knowledge Representation*. Springer Verlag, New York

Cruse A 1986 *Lexical Semantics*. Cambridge University Press

Davis A R, Koenig J P 2000 Linking as constraints on word classes in a hierarchical lexicon. Language **76**, no. 1

Dowty D R 1979 *Word Meaning and Montague Grammar*. D. Reidel, Dordrecht, The Netherlands

Dowty D R 1989 On the semantic content of the notion 'Thematic Role'. In: Chierchia G, Partee B, Turner R (eds.) *Properties, Types, and Meaning* Vol. II. Dordrecht

Dowty D 1991 Thematic proto-roles and argument selection. *Language* **67**: 547–619

Flickinger D, Pollard C, Wasow T 1985 Structure-sharing in Lexical Representation. In: *Proceedings of 23rd Annual Meeting of the ACL*. Chicago, IL, pp. 262–7

Goldberg A E 1995 *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago

Grimshaw J 1979 Complement selection and the lexicon. *Linguistic Inquiry* **10**: 279–326

Grimshaw J 1990 *Argument Structure*. MIT Press, Cambridge, MA

Gruber J S 1976 *Lexical Structures in Syntax and Semantics*. North Holland, Amsterdam

Guthrie L, Pustejovsky J, Wilks Y, Slator B 1996 The role of lexicons in natural language processing. *Communications of the ACM* **39**: 1

Hale K, Keyser J 1993 On argument structure and the lexical expression of syntactic relations. In: Hale K, Keyser J (eds.) *The View from Building 20*. MIT Press, Cambridge, MA

Jackendoff R 1990 *Semantic Structures*. MIT Press, Cambridge, MA
Johnston R 1995 Semantic underspecification and lexical types: Capturing polysemy without lexical rules. *Proceedings of ACQUILEX Workshop on Lexical Rules*, August 9–11, 1995, Cambridgeshire, UK
Levin B 1993 *Towards a Lexical Organization of English Verbs*. University of Chicago Press, Chicago
Levin, B, Rappaport M H 1995 *Unaccusativity: At the Syntax-Semantics Interface*. MIT Press, Cambridge, MA
Lyons J 1968 *Introduction to Theoretical Linguistics*. Cambridge University Press, Cambridge, UK
McCawley J 1968 Lexical insertion in a transformational grammar without deep structure. *Proceedings of the Chicago Linguistic Society* **4**
Mel'čuk I A 1988 Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography* **1**: 165–88
Miller G WordNet: An on-line lexical database. *International Journal of Lexicography* **3**: 235–312
Miller G 1991 *The Science of Words*. Scientific American Library
Pinker S 1989 *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, MA
Pollard C, Sag I 1994 *Head-Driven Phrase Structure Grammar*. University of Chicago Press and Stanford CSLI, Chicago
Pustejovsky J 1991 The syntax of event structure. *Cognition* **41**: 47–81
Pustejovsky J 1992 Lexical semantics. In: Shapiro S (ed.) *Encyclopedia of Artificial Intelligence*, 2nd edn. Wiley, New York
Pustejovsky J 1995 *The Generative Lexicon*. MIT Press, Cambridge, MA
Pustejovsky J, Boguraev P 1993 Lexical knowledge representation and natural language processing. *Artificial Intelligence* **63**: 193–223
Sanfilippo A 1993 LKB encoding of lexical knowledge. In: Briscoe T, de Paiva V, Copestake A (eds.) *Inheritance, Defaults, and the Lexicon*, Cambridge University Press, Cambridge, UK
Schabes Y, Abeille A, Joshi A 1988 Parsing strategies with lexicalized grammars: In: *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary
Steedman M 1997 *Surface Structure Interpretation*. MIT Press, Cambridge, MA
Talmy L 1985 Lexicalization patterns: Semantic structure in lexical forms. In: Shopen T (ed.) *Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon*. Cambridge University Press, Cambridge, pp. 57–149
Weinreich U 1972 *Explorations in Semantic Theory*. Mouton, The Hague, The Netherlands
Williams E 1981 Argument structure and morphology. *Linguistic Review* **1**: 81–114

J. Pustejovsky

Liability: Legal

Legal liability may be defined most generally as any legally enforceable obligation to act or refrain from acting in a particular way. Some preliminary distinctions are in order.

First, a legal liability may be potential, in the sense that a legally defined duty exists whose violation by the duty-holder would trigger liability, or it may be current, in the sense that the duty-holder has already violated it and the fact of violation either has been authoritatively determined or is not really in dispute.

Second, a legal liability may be criminal, civil, or administrative. It is criminal if violations are prosecuted by the government as an offense against the public and may subject violators to imprisonment or criminal fines under the special rules and procedures applying to criminal prosecutions, such as the 'guilt beyond a reasonable doubt' standard. It is civil if the violation is enforceable by private parties (or by the government acting in a proprietary or other analogously private capacity) seeking relief in the form of money damages, an injunction, or a declaratory judgment. It is administrative if the violation is enforceable by an administrative agency pursuant to a regulatory statute.

Third, legal liabilities may be distinguished according to the substantive law that prescribes the duty in question. This distinction is most commonly applied in connection with civil liability, where the duty may be a creature of tort law, contract law, fiduciary law, property law, inheritance law, family law, and so forth.

Fourth, the process by which liability is determined differs significantly between common law and civil law systems (Damaska 1986). This article focuses on a common law system, that of the USA, but notes several points on which the civil law systems adopt a different approach.

The discussion here is confined to general tort liability. At its most general and schematic level, the system for imposing tort liability ('tort system') consists of six elements: goals; substantive rules; procedural rules; remedies; choice of law; and institutions. A seventh element, which defines by exclusion, consists of enclaves in which tort law has been limited or eliminated (see *Torts in Law*).

1. Goals of Tort Law

Considerable controversy exists over the goals that tort law should serve and how those goals should be weighted (Calabresi 1970). According to most tort law scholars, the dominant—and for some, the only—legitimate purpose of tort law was traditionally thought to be 'corrective justice.' This is an Aristotelian notion defining justice as the restoration of the *status quo ante* by repairing wrongfully inflicted losses and annulling wrongfully obtained gains. Some scholars see corrective justice as requiring a system of tort law that enforces a direct, moral relationship between the victim and the injurer, obliging the latter to compensate the former for the injury that he or she has wrongfully inflicted. The purpose of compensation, in this view, is simply to affirm this moral

relationship and to effectuate this duty of repair, correction, and restoration (Coleman 1992).

Increasingly, however, the corrective justice conception of tort law has been augmented, and to some extent supplanted, by more functional goals. They seek to enforce not a particular moral relationship between injurer and victim for its own sake, but rather a variety of social goals to which the individual parties to the tort case are merely instrumental. The most important of these social goals are (a) the deterrence of excessive risk creation, and (b) the distribution of injury losses through the compensation of victims.

The functional rationale for compensation is that it distributes the costs of accidents more broadly; at the limit (social insurance), it spreads those costs from the individual victim to all members of society, who can better bear it. But a functionalist tort law might—depending on the dictates of benefit–cost analysis, effective deterrence, administrative efficiency, insurability, and other policy criteria—require that an entity other than the injurer (e.g., private or social insurance) compensate the victim, or might deny the victim compensation altogether. In the functional view, morality is at most a policy constraint; it is not constitutive of justice in tort law. Some scholars oppose functional justifications for tort liability because they believe that judges' use of private tort disputes to pursue broad social policy goals is institutionally illegitimate, incompetent, or both (Sugarman 1989).

2. *Substantive Rules*

Tort law prescribes a body of substantive legal doctrine (i.e., rules or principles) to govern behavior. In common law systems, these rules are derived largely by courts from general principles of justice, previously decided cases, policy considerations, and other sources, though statutes and regulations are becoming more important sources of liability rules. In civil law systems, statutes (or codes) are the principal sources, but considerable judicial interpretation, guided by general principles of justice, is often required.

Whatever their sources, these rules or principles define the legally enforceable duties that an individual (or other entity) owes to others, and the circumstances in which he or she may violate those duties (the 'standard of care'). For example, tort law prescribes the legal duties that a physician owes to his or her patient and the standard of care that he or she must satisfy in order to avoid liability for injuries to the patient caused by the physician's acts or omissions. The standard of care is usefully understood in terms of the risks of injury that the physician is permitted to create in the course of treating the patient. Ordinarily—and in the case of healthcare—tort law defines this standard of care as 'reasonable behavior or care under all of the relevant circumstances.' A failure to meet this standard of care is called 'negligence' or 'fault' (or, in the specific case of medical treatment, 'medical malpractice') (Prosser and Keeton 1984).

For certain types of activities, such as the use of explosives or the sale of defective products, the standard of care is not negligence but 'strict liability.' If an activity is subject to a strict liability standard, the actor who causes an injury to another person is liable for that injury even if he or she was exercising reasonable care (i.e., was not negligent). In common law countries such as the USA, negligence is the default standard of care and strict liability the exception. Indeed, even strict liability rules usually incorporate fault-type balancing considerations that make the liability less strict than it might appear on its face.

In addition to negligence and strict liability, there is a third liability standard: intentional infliction of harm. Although courts have defined 'intention' in a number of different ways, it is not limited to harming another with malice, ill-will, or an affirmative desire to injure, but also includes a risky action undertaken with the knowledge (or in some cases, reason to know) that it is likely to injure another. A fourth liability standard, no liability, is discussed in Sect. 7.1 under 'immunities.'

In order to establish a defendant's tort liability, the plaintiff must prove that the defendant owed him or her a legal duty of care, that the defendant breached this duty by violating the applicable legal standard (usually negligence), and that the defendant's breach of duty caused harm to the plaintiff of a kind that tort law recognizes as compensable.

Even if the plaintiff is able to establish duty, breach, causation, and damage, however, tort principles may provide the defendant with a defense. If the defendant can prove that a particular defense applies, the defense is usually a 'complete' one barring any liability. Examples include assumption of the risk (i.e., the plaintiff knowingly and voluntarily consented to the risk), statute of limitations (i.e., the plaintiff did not bring suit within the requisite period of years), and immunity (i.e., the law bars liability for certain categories of actors). The plaintiff's own negligence ('contributory negligence') traditionally was a complete defense, but this rule has been replaced by various versions of 'comparative negligence (or fault)' under which the plaintiff's negligence does not bar his or her recovery completely but instead reduces it according to the parties' relative fault (Prosser and Keeton 1984).

3. *Procedural Rules*

Procedural rules specify how claims must be presented to the court, which party bears the burden of producing evidence in the first instance, which party bears the burden of persuading the trier of fact (judge or

jury), and which standard of proof must be satisfied (usually 'a preponderance of the evidence', also described as 'more likely than not'). These rules also prescribe how jurisdiction over a party must be obtained, what kinds of notice must be provided, how discovery (i.e., disclosure of information possessed by a party or witness) is conducted, how other parties may be brought into the lawsuit, the powers that judges may lawfully exercise, how parties should conduct motion practice (i.e., the rules governing requests for judicial orders), how decisions may be appealed, and the like. Another set of quasi-procedural rules govern the admissibility of evidence, although such rules are much fewer in civil law systems where there is no need to insulate lay jurors from inadmissible evidence, thus making for much quicker and less costly and cumbersome trials (American Law Institute 2000).

4. *Remedies*

Once a defendant's tort liability is established, essentially two kinds of remedies (i.e., forms of relief) may be available (see *Remedies and Damages: Legal Aspects*). The most common remedy by far is money damages, which in turn can be subdivided into compensatory and punitive damages. Compensatory damages are designed to restore the plaintiff to the *status quo ante* insofar as money can do so. Nevertheless, under the so-called 'American rule,' each party bears its own counsel fees and costs unless a statute specifically provides for fee-shifting. The 'English rule,' which requires the losing party to defray some or all of the prevailing party's fees, is also more common in civil law countries. In those countries, however, counsel fees tend to be much lower because much of the investigatory work done by lawyers in the USA is instead done by magistrates, and also because civil law proceedings tend to be far less complex and protracted.

In the USA, compensatory damages may cover physical injuries, lost wages, medical expenses, other out-of-pocket costs, loss of profit opportunity, emotional distress, pain-and-suffering, hedonic loss, damage to intimate relationships ('consortium'), reputational loss, loss of privacy, and so forth. Punitive damages may be assessed against a defendant whose tortious conduct was particularly egregious and socially abhorrent. Tort reform statutes in many states, however, now impose a variety of limits on punitive damages (Blatt et al. 1991).

Tort law ordinarily applies to relatively sudden, unintentional, and isolated events, so the only meaningful relief is post-accident money damages. In cases where the wrongful conduct is of a continuing nature, however, injunctive relief—a judicial order requiring the defendant to act, or refrain from acting, in a particular way, on pain of being held in contempt and other sanctions—may be available. This occurs most commonly in 'nuisance' actions against polluters or others whose conduct interferes with a plaintiff's use and enjoyment of her own property (Prosser and Keeton 1984).

5. *Choice of Law*

In the USA, tort liability is largely adjudicated by state judges applying state law to disputes. State tort law mainly takes the form of common law principles, although an increasing number of statutes or regulations affect tort adjudication. Judges in the forum state (i.e., where the suit is filed) ordinarily apply the forum state's own procedural, substantive, and remedial law to the dispute. If, however, the accident and wrongful conduct occurred in another state, the judge consults the forum state's 'choice of law' rules in order to determine which state's law governs which issues in inter-state disputes of this kind. These choice of law rules may dictate that the judge apply the law of the other state (*lex delicti*) rather than the forum state's law (*lex fori*) to some or all of the issues in the case.

Many American tort cases are filed in federal rather than state court. This may be because they are 'diversity' cases (i.e., the plaintiff and defendant are from different states), which the US constitution permits federal courts to hear. Even in diversity cases, however, the federal court is required to apply the state tort and choice-of-law principles of the state in which it sits, while it must follow federal procedural rules. A plaintiff may also file tort cases in federal court if the claim arises under federal law—a federal statute, federal agency regulation, a limited body of federal common law, or (very occasionally) the US Constitution.

A number of federal statutes create 'federal causes of action' (i.e., rights to seek tort-like relief under federal law). For example, the Federal Tort Claims Act creates a federal court remedy for wrongs committed by federal officials. (Even this remedy, however, is ultimately based on state tort law; the plaintiff must show that the official's wrong would be a tort under the law of the state where the injury occurred.) A federal civil rights statute known as Section 1983 creates a remedy, in either state or federal court, for violations of federal law committed by state or local governments. The Consumer Product Safety Act creates a federal remedy for people injured by products that violate regulations issued by a federal product safety agency.

6. *Institutions*

Tort law develops and operates in a distinctive institutional context. Two of the most important of these institutions are the jury (in the USA) and insurance.

Tort litigants in the USA generally have a right to a jury trial under the federal and most state constitutions, although they may waive this right and opt for a bench trial (i.e., by the judge alone). In principle, the jury is the trier-of-fact, applying the law that the judge gives it to the factual evidence presented in the case. Because of the open-ended nature of many tort law standards (e.g., 'reasonable care'), however, jurors inevitably enjoy some leeway or discretion that enables them to introduce into jury decisions their own values and attitudes toward the parties and the law. The subjectivity and discretion of juries, together with other factors that may be closely related to the jury system (e.g., lawyers' rhetorical skills), produce greater legal uncertainty in the USA than in other countries' liability systems. Juries also increase the cost and duration of trials and often make settlements more difficult to reach (Litan 1993) (see *Juries*).

Juries usually render only a 'general' verdict (i.e., 'we find for defendant' or 'we find for plaintiff in the amount of $X'); they neither state their reasons nor make explicit findings of fact. Accordingly, their verdicts—unlike the judge's rulings on points of law—carry little or no precedential value and convey little or no information, other than the verdict itself. (In a particularly complex case, the judge may ask them to render a 'special' verdict with specific findings of fact in response to factual questions that the judge has put to them.) Juries are selected through a lawyer-orchestrated process ('voir dire') that tends as a practical matter to produce many jurors who are relatively ignorant and unsophisticated about the kinds of technical issues increasingly arising in tort cases. Much of the unwieldy and costly apparatus of the trial, including many of the rules of procedure and of evidence, are necessary only because of the risk that juries will be improperly influenced by the trial lawyers on both sides. Proceedings in Britain and in Commonwealth countries are simpler, faster, and more predictable than those in the USA because those countries have largely abolished the lay jury in civil cases (Kagan 2001) (see *Courts and Adjudication*).

Tort liability in all countries is profoundly shaped by the existence of insurance (Abraham 1995) (see *Insurance and the Law*). Many if not most of the risks of injury that tort law regulates are insurable, in the sense that third parties (i.e., insurers) would be willing to bear those risks *ex ante* in exchange for some premium. Many of these insurable risks are in fact insured, either on a first-party or third-party basis. In first-party insurance, a person who is subject to a risk of injury purchases insurance to cover some or all of that risk, as when one buys a health insurance policy. First-party coverage is usually on a no-fault basis; the insurer indemnifies the insured merely on the occurrence of the insured event, without regard to how it occurred (except in the case of the insured's misconduct). In third-party or 'liability' insurance, a person who fears that he or she may injure others purchases insurance to cover some or all of that liability risk, as when a driver buys an auto accident liability policy. Here, the insurer only pays if the insured has incurred a legal liability to compensate a third party.

The tort system, at least with respect to accidents between strangers (i.e., those who have not contracted over who will bear this risk), is in effect a system of mandatory third-party insurance. It is mandatory because even one who agrees to opt out of the tort system often cannot do so. This is especially true in the area of healthcare, where courts usually will not enforce contracts whereby a patient waives his or her right to sue the physician in tort in exchange for a lower fee or for a service that the physician would not otherwise be willing to provide. It is a third-party insurance system because the insurer is insuring a risk that its insured imposes on a third party who is a stranger to both of them. In effect, every healthcare consumer is required to purchase an insurance policy along with the physician's services. This policy gives the patient a right to seek damages in the tort system according to the principles of medical malpractice law, in the event that the physician negligently injures him or her. The cost of this policy, which is reflected in the physician's fee, is equal to the physician's potential tort liability to the patient, which in turn is equal to the premium that the physician must pay for medical malpractice insurance (or put aside for self-insurance) to cover the liability risks the physician incurs by treating the patient.

This third-party insurance feature of tort law leads to important theoretical and policy debates (Priest 1987, Croley and Hanson 1995). Third-party insurance is often criticized as being both less efficient and less equitable than first-party insurance, which permits an insured to purchase the kind and amount of coverage that he or she prefers, and permits the insurer to demand the kind and amount of information it thinks it needs in order to underwrite the particular risk at a profit. In any event, judges applying an increasingly functional tort law take insurance into account, sometimes explicitly, in framing liability rules, and they often allocate the costs of accidents to the party they think is in the best position to insure against (distribute the risk of) the particular risk that caused the plaintiff's injury (Calabresi 1970).

7. *Non-tort Enclaves*

Tort liability in the USA expanded enormously during the twentieth century, most radically and dramatically during the 1970s and 1980s, while in other countries it has been marginalized by social insurance schemes. Most American states have sought to stem this expansion by adopting so-called 'tort reform' statutes, which attempt in a variety of ways to limit the tort liabilities of healthcare providers, product manu-

facturers, governments and public officials, employers, and other categories of potential injurers. Although some legal scholars have claimed to detect some slowing, or even a reversal, of this expansion (Henderson and Eisenberg 1990), there is no question that the scope of tort liability today is still much larger than it was even as recently as the early 1960s (Schwartz 1992). This expansion is the subject of much political debate and controversy. Still, important enclaves remain in which tort principles do not apply or apply only to a limited extent. Two such enclaves are immunities and no-fault schemes.

7.1 Immunities

Traditionally, US courts declined to apply tort principles to accidents arising from certain categories of actors or relationships. Charitable institutions were generally immune from tort liability, and suits between between spouses and between parents and children were also barred. The general expansion of tort liability and the availability of liability insurance, however, have encouraged courts to narrow these particular immunities (Prosser and Keeton 1984).

Governments and their officials also traditionally enjoyed broad protection from tort liability, loosely referred to as 'sovereign immunity.' Like charitable and intrafamily immunities, courts and legislatures have narrowed the scope of sovereign immunity but it nevertheless remains quite broad. At the federal level, the Federal Tort Claims Act waives the sovereign immunity of the USA, but this waiver is subject to a number of exceptions and limitations—for discretionary decisions and many intentional torts, for example—that continue to protect the federal government from liability for many of its torts. The Eleventh Amendment to the US constitution also generally immunizes the states from being sued for money damages without their consent in federal court. Individual states are also generally immune from suits brought against them in their own courts, although all have waived this immunity to some degree. Sometimes, plaintiffs can circumvent federal or state sovereign immunity by suing individual officials rather than the government itself, but those officials are often able to claim an absolute or qualified (good-faith) immunity nonetheless, and a variety of other special rules limit suits against governments and/or their officials (Schuck 1983).

7.2 No-fault Plans

Dissatisfaction with how the tort system allocates the costs of accidents has led governments to create some schemes for compensating victims of certain kinds of accidents without requiring them to prove an injurer's wrongdoing. Workplace accidents are the most important area in which no-fault compensation has replaced the tort system, with workers compensation statutes in every state and the federal government. No-fault systems are more prominent in other developed countries, with New Zealand's accident compensation law being the most comprehensive in the world (Franklin and Rabin 2001).

Workers' compensation effects an important trade-off. Under such a scheme, an injured worker's sole remedy against his or her employer is to file a claim with the employer's workers' compensation insurance carrier, whose decision on the claim is then subject to appeal to an administrative board and ultimately to the courts. In order to receive an award, the injured worker need not prove the employer's negligence but must only show that the accident occurred in the course of his or her employment. On the other hand, the award is limited to scheduled amounts for pecuniary losses (e.g., wages, medical expenses, therapy) for a specified number of weeks; the employee is barred from recovering for the intangible losses (e.g., pain and suffering, hedonic loss, emotional distress) that he or she could have sought had she sued in tort. (The worker may still be able to bring a tort action against a non-employer such as the manufacturer of the machine that caused the injury; such third-party actions in tort are very common in certain settings, especially industrial accidents.)

Many American states have also adopted no-fault compensation systems applicable to property damage and personal injuries sustained in automobile accidents (Schwartz 2000). Almost all of these statutes, however, do not really supplant the tort system; instead, they preserve the right to sue in tort once plaintiff's losses exceed a specified threshold, which tends to be quite low. Indeed, some commentators contend that automobile no-fault schemes actually encourage tort litigation by supplying plaintiffs with some immediate no-fault payments that they can then use to finance their tort actions (Bell and O'Connell 1997). Politicians' interest in these schemes, essentially ended in the 1970s. Few, if any, states have adopted them since then, and although the US Congress has recently considered a variant of the idea, it seems unlikely to gain the necessary political support.

Some narrowly targeted no-fault laws have been adopted. Congress, for example, has enacted no-fault compensation schemes for certain vaccine-related childhood diseases, for certain swine flu-related illnesses, and for nuclear power plant accidents, and a few states have enacted schemes relating to certain perinatal malpractice claims. In each instance, however, victims may under certain circumstances reject the statutory no-fault alternative and instead proceed to sue in the tort system (Rabin 1993).

8. Conclusion

The future of tort law in various societies is difficult to predict both because each nation's legal system is

different and because the scope of tort law is linked, perhaps inversely, to that of social insurance. Although the political impulses favoring a wider distribution of risk through social insurance are always strong, new experiments with or expansions of social insurance will be impeded by global competition, the liberalization of national economies, and tax reduction pressures. The kind of individualistic, fault-based, juried adjudication that characterizes the US tort law system, however, is one that few other societies can afford or would find congenial. In the US itself, the most recent generation of tort reform statutes, which are largely designed to limit adjudication costs and liability in the interests of corporate and professional defendants (at least in the first instance), are perhaps the best evidence of what the future holds.

See also: Administrative Law; Courts and Adjudication; Deterrence; Deterrence: Legal Perspectives; Deterrence Theory: Crime; Insurance and the Law; Juries; Punishment: Social and Legal Aspects; Social Insurance: Legal Aspects; Torts in Law

Bibliography

Abraham K S 1995 *Insurance Law and Regulation: Cases and Materials*, 2nd edn. Foundation Press, Westbury, NY

American Law Institute 2000 Transnational Rules of Civil Procedure, Preliminary Draft No. 2 (March)

Bell P A, O'Connell J 1997 *Accidental Justice: The Dilemmas of Tort Law*. Yale University Press, New Haven, CT

Blatt R L, Hammesfahr R W, Nugent L S 1991 *Punitive Damages: A State-by-State Guide to Law and Practice*. West Publishing Co., St Paul, MN

Calabresi G 1970 *The Costs of Accidents*. Yale University Press, New Haven, CT

Coleman J L 1992 *Risks and Wrongs*. Cambridge University Press, Cambridge

Croley S P, Hanson J D 1995 The nonpecuniary costs of accidents: pain-and-suffering damages in tort law. *Harvard Law Review* **108**: 1785

Damaska M 1986 *The Faces of Justice and State Authority: A Comparative Approach to the Legal Process*. Yale University Press, New Haven, CT

Franklin M, Rabin R L 2001 *Cases and Materials on Tort Law and Alternatives*, 7th edn. Foundation Press, New York

Henderson J, Eisenberg T 1990 The quiet revolution in products liability: an empirical study of legal change. *UCLA Law Review* **37**: 479

Kagan R A 2001 *Adversarial Legalism: The American Way of Law*. Harvard University Press, Cambridge, MA

Litan R E (ed.) 1993 *Verdict: Assessing the Civil Jury System*. Brookings Institution, Washington, DC

Priest G L 1987 The current insurance crisis and modern tort law. *Yale Law Journal* **96**: 1521

Prosser W L, Keeton W P 1984 *Prosser and Keeton on the Law of Torts*, 5th edn. West Publishing Co., St Paul, MN

Rabin R L 1993 Some thoughts on the efficacy of a mass toxics administrative compensation scheme. *Maryland Law Review* **52**: 951

Schuck P H 1983 *Suing Government: Public Remedies for Official Wrongs*. Yale University Press, New Haven, CT

Schwartz G T 1992 The beginning and the possible end of the rise of modern American tort law. *Georgia Law Review* **26**: 601

Schwartz G T 2000 Auto no-fault and first-party insurance: advantages and problems. *Southern California Law Review* **73**: 611

Sugarman S D 1989 *Doing Away with Personal Injury Law: New Compensation Mechanisms for Consumers, Victims, and Business*. Quorum Books, New York

P. H. Schuck

Liaison and Consultation Psychiatry

Consultation-liaison psychiatry delineates the work of psychiatrists with patients who are in medical treatment primarily for somatic reasons in nonpsychiatric departments of a general hospital, or in ambulatory care of a nonpsychiatric physician. More specifically, consultation psychiatry describes a practice where psychiatrists see patients treated by other nonpsychiatric physicians only when called upon to give advice about diagnostic and treatment issues to consultees, who in turn follow through on these recommendations in the treatment of patients. In liaison psychiatry, however, psychiatrists are involved more directly in the treatment of physically ill patients with psychiatric comorbidity, assuming more responsibility in joint patient care, (e.g., having the prerogative of seeing patients admitted to an internal or surgical ward without being specifically endorsed to do so in every given case) as well as starting and following through psychiatric or psychotherapeutic treatment. As in everyday clinical care, boundaries between psychiatric and liaison approaches are not clear-cut, a worldwide consensus has developed to use the phrase consultation-liaison psychiatry (or c-l psychiatry), with the notable exception of the UK, where liaison psychiatry is the only term used for both kinds of approaches.

1. Psychiatric Comorbidity in Physically Ill Patients

The need for c-l psychiatric services stems from the high comorbidity of psychiatric illnesses in patients suffering from physical diseases.

1.1 Psychiatric Comorbidity in the General Hospital

It is estimated that between 30 and 50 percent of physically ill patients in general hospitals show comorbid psychiatric disorders, especially organic psychoses (*ICD-10 F0*), psychological and behavioral disturbances due to psychotropic substances (*ICD-10 F1*), and acute stress reactions (*ICD-10 F4*), with percentages of these diagnoses ranging as high as 16 percent, 8 percent and 10 percent in medical, and 17

percent, 8 percent and 13 percent in surgical patients, respectively (Arolt 1997). Special patient populations—such as those in organ transplantation programs, intensive care units, or oncological patients—show an even higher psychiatric comorbidity. The presence of psychiatric comorbidity does not always call for psychological treatment at the same time as the underlying or concomitant physical disease is being treated by physicians or surgeons. Neverthless, conservative estimates point to a figure of some 10 percent of general hospital inpatients who should receive psychiatric care alongside somatic care.

While the evaluation of patients committed to Accident and Emergency Departments after attempting to harm themselves is an important task for psychiatrists, and will often be performed by c-l services, their care is only one part of a larger picture of care for patients suffering from medical and psychiatric comorbidity, with the latter impacting negatively on course and outcome of the former. A prerequisite for psychiatric services to achieve an impact on this group of patients is the integration of psychiatry into general hospitals. Thus, in the USA, where the foundation of psychiatric departments as integrated parts of general hospitals started as early as the 1920s, c-l psychiatry has a long tradition, whereas in the Federal Republic of Germany, for instance, where state mental hospitals have given way to general hospital psychiatric departments only since the end of the 1970s, c-l psychiatry started to gain shape only toward the end of the twentieth century. Overall, international interest in c-l psychiatry increased outside the USA during the 1980s and 1990s with special emphasis in the UK, the Netherlands, Australia and New Zealand, where c-l psychiatric sections exist within national psychiatric organizations. The history of c-l psychiatry shows that the acceptance of psychiatric care for patients with medical–psychiatric comorbidity is dependent largely on the regular availability of psychiatrists in general hospitals, which provides the opportunity to demonstrate the added value of psychiatric diagnosis and treatment to other medical disciplines in the care of comorbid patients.

1.2 Psychiatric Comorbidity in the Primary Care Sector

While c-l psychiatry developed primarily in the general hospital setting, there was growing emphasis worldwide on diagnosis and treatment of psychiatric comorbidity of patients in the primary care sector toward the end of the twentieth century. Results from the *National Institute of Mental Health Epidemiological Catchment Area Program* show that about 16 percent of the community-dwelling adult population fulfil criteria for at least one psychiatric disorder. This figure increases to at least 21–26 percent of primary care patients. Studies show that most cases of adjustment disorders and about three-quarters of depressive disorders are treated in primary care. In most cases, these patients present with physical symptoms that are not caused by physical disease (see *Somatoform Disorders*). The proportion of undetected and untreated psychiatric (co-)morbidity in this patient group is high (APM 1997).

1.3 Psychiatric Morbidity in Nursing Homes

The percentage of the population older than 65 years of age is increasing. It is estimated that 20–50 percent of these people will eventually live in nursing homes. In the USA, it is estimated that the total number of nursing home residents will exceed 4 million by the year 2040. The very high prevalence of psychiatric-related illness among those in nursing homes is documented in a number of studies, showing that up to 70 percent of nursing home residents may meet diagnostic criteria for dementia, and 90 percent may have at least one form of behavioral disturbance, such as agitation, delusions, or depression. Depression is largely undetected and undertreated in this population, despite the fact that the majority of cases treated, e.g., with antidepressants, will recover from a depressive episode. Only recently, there has been growing awareness that people with mental retardation (or learning disability) show a significantly higher percentage of psychiatric comorbidity than the average population. While attempts are made to achieve a close to normal living situation for a large part of this clientele, people with severe mental retardation—which is often complicated not only by psychiatric comorbidity but also by other physical diseases such as epilepsy—will receive adequate care only in nursing home settings (Reichman and Katz 1996).

In sum, there has been increasing awareness of hidden mental comorbidity in the general hospital, primary care and nursing home sectors, and of the need for specialist psychiatric treatment in a significant percentage of patients.

1.4 Effects of Psychiatric Comorbidity in Physically Ill Patients

Studies show a significant association between psychiatric comorbidity and more complicated illness courses, both for patients in the primary care sector, and for general hospital inpatients. Patients with somatization disorder use nine times the healthcare services of the general population, panic disorder patients have 10 times the number of visits to hospital emergency rooms than nonanxious patients, and 70 percent will see more than 10 physicians before an accurate diagnosis is made. Medically ill patients with depression have a poorer prognosis than nondepressed patients with equal somatic comorbidity. For ex-

ample, postmyocardial infarction patients with depression are at significantly greater risk of death over a period of 6 months following the infarction, and depressed stroke patients show poorer results in rehabilitation programs with more persistent disability than nondepressed stroke patients. It has been shown that a small percentage of patients are responsible for the use of a major portion of healthcare, with, for example, 10 percent of patients of a large health maintenance organization in the USA consuming approximately one-third of all primary care visits, 52 percent of all speciality visits, and 40 percent of in-hospital days. Fifty percent of those high utilizers of healthcare showed some form of psychiatric comorbidity, e.g., 20.2 percent somatization disorder, 33.6 percent generalized anxiety disorder or panic disorder, and 40 percent major depression or dysthymia (APM 1997).

In general hospital inpatients it has been shown that medically ill patients with psychiatric comorbidity show longer lengths of stay in hospitals, more frequent re-admissions, and an overall increase in costs for treatment as compared to patients without psychiatric comorbidity. Adjustment for other factors associated with longer length of stay and rehospitalizations, such as severity of underlying medical disease, did not reduce the significance of the negative impact exerted by psychiatric comorbidity (APM 1997). The elderly are exposed to special risks, with about 10 percent of elderly patients admitted to a general hospital suffering from delirium. About 30 percent of elderly medical inpatients have dementia, which makes more complex management necessary, with problems arising from, for example, diminished compliance with medication, and in following through with rehabilitation programs.

2. *The Practice of Consultation and Liaison Psychiatry*

The scope of c-l psychiatric services spans primary and hospital care settings, the care of elderly residents in nursing homes, and of clients with learning disability and accompanying psychiatric disorders.

2.1 C-l Psychiatric Services in the General Hospital

Despite the high percentage of psychiatric comorbidity in the general hospital, referrals to psychiatric c-l services amount at most to only a small number of physically ill inpatients, approaching an average number of 1 percent of general hospital inpatients worldwide. Usually, the majority of patients are referred by internal medicine departments, with a large percentage being older than 65 years of age. Despite all differences in service delivery that are due to local necessities in the provision of care for given populations, c-l psychiatrists most often diagnose delirious states and other disorders of cognition, followed by depressive syndromes and adjustment disorders, fitting well to the above mentioned prevalence figures of psychiatric comorbidity in general hospital patients. In contrast, despite their obvious frequency, patients with alcohol-use disorders are referred much less by their treating physicians. The low rate of referrals to psychiatric c-l services has led to the assumption that nonpsychiatric physicians do not recognize psychiatric disorders sufficiently, or, if they do, do not think of psychiatric interventions as being helpful to their patients. Liaison psychiatry proper takes its vantage point from this discrepancy: other than the usual psychiatric consultation with its mainly emergency psychiatric character that by default is only able to care for the tip of the iceberg of psychiatric comorbidity, i.e., patients having passed the filter of recognition of a nonpsychiatric physician. Liaison psychiatric approaches aim at the denominator (Strain 1999) of psychiatric comorbidity, triggering specialist psychiatric care, e.g., by screening patients on hospital admission, or identifying problem behavior in joint rounds with nonpsychiatric physicians, or by educating the latter in how to better identify patients that might profit from referral to specialist psychiatric care. Thus, liaison psychiatric approaches try to go well beyond the 'firefighter function' of typical psychiatric consultation, aiming at changing referral patterns of nonpsychiatric departments, e.g., enabling physicians to deal with and/or prevent the development of delirious states, or increase referral rates of depressed patients or patients with alcohol-use disorders by demonstrating the efficacy of even time-limited psychiatric interventions (Guthrie and Creed 1996).

Both in consultation, and to a somewhat lesser extent in liaison psychiatric approaches, psychiatrists have to be aware of the stages a psychiatric consultation will typically go through. In the initial phase, and due to uncertainty about how to deal with a patient the consultant's referral diagnosis might be rather imprecise. In such cases it must be the task of the c-l psychiatrist to help in working out what is in fact the problem. The ensuing intervention must then be, other than in a typical psychiatric office situation, directed not only toward the patients, but also take into consideration the whole system, including the ward doctor, nurses, and relatives. Otherwise, c-l psychiatric interventions, when performed in isolation and/or out of context, might be unsuccessful (Diefenbacher 2000).

2.2 C-l Psychiatric Services in Primary Care and in Nursing Homes

Along with the growing interest in psychiatric (co)morbidity in primary care patients and nursing home residents, international c-l psychiatry shows a remarkable shift from the general hospital setting to collaborative care practices in the primary care setting,

and, though still in its beginnings, to liaison psychiatric services with nursing home populations. For primary care settings, studies show that inclusion of a psychiatrist in primary care treatment teams, e.g., on the basis of regularly performed visits of a psychiatrist, leads to a marked improvement in the symptoms of patients with somatization disorder or depression than provision of services by primary care physicians alone (APM 1997). For nursing home populations, the need to consider psychopharmacological as well as nonbiological interventions such as talking therapies, recreational therapies, and so on, in combined treatment plans tailored to specific residents and specific nursing home settings, stress the need for psychiatrists to oversee the implementation of such integrated approaches. Hence, liaison models with more intense involvement of psychiatrists working with other nonpsychiatric physicians, as well as administrative and nursing staff in addition to care for residents and their relatives will be developed in the near future. The latter holds especially true for people with mental retardation (or learning disability) living in nursing homes, as diagnostic and treatment approaches in this clientele will require special expertise not necessarily familiar to primary care physicians, or even psychiatrists themselves. Especially in patients that are not able to communicate verbally, 'cooperative' and 'uncooperative' behavior may reflect an underlying psychiatric disorder, which in turn may be treatable, as in the case of hostile behavior in demented patients with dysphoric depression that may well respond to pharmacological treatment with, for example, selective serotonin reuptake inhibitors. C-l psychiatrists working in such settings are ideally placed to teach nursing staff and primary care physicians that expert assessment of behavioral and psychological symptoms in nonverbal patients may lead to more specific management and better patient care. Thus a patient with dementia whose delusional experiences lead to bizarre behavior should be treated differently from a demented patient who suffers from pain, or a depressed nonverbal patient who refuses aggressively to accept food (Reichman and Katz 1996).

Despite the increasing interest in models of collaborative care in the outpatient, as well as nursing home sectors, it has to be borne in mind, that, at least at this point in time, many patients with somatopsychic comorbidity are seen for the first time by a psychiatric specialist after having been admitted to a general hospital. Thus, psychiatrists as well as planners of health services will have to bear in mind the importance of the general hospital as a pathway to specialist psychiatric care.

2.3 Evaluation of C-l Psychiatric Interventions

Psychiatric care leads to improved management of medically ill patients with psychiatric comorbidity. This is the rationale of the practice of c-l psychiatry, whose benefits can be demonstrated within a value matrix, taking into account treatment outcomes in medically ill patients with psychiatric comorbidity, as well as economic savings by adding specialist psychiatric treatment to standard medical treatment. Cost–benefit issues of consultation psychiatry service delivery have been demonstrated initially in general hospital inpatients, with primarily US studies showing a trend to reduced length of hospital stay with corresponding reductions in cost of treatment in medical patients after a psychiatric liaison service was inaugurated, compared with control groups without such an intervention program (APM 1997). However, as other studies failed to show such an effect, the overall conclusion was that this research approach had to be judged as inconclusive in the face of a sharp decrease in lengths of hospital stays worldwide due to secular trends in shifting the provision of healthcare into the outpatient sector in order to reduce the escalation of overall healthcare costs. Consequently, research on the efficacy of psychiatric intervention strategies in patients with psychiatric (co-)morbidity in the primary care sector has increased. Studies showed that integrated c-l psychiatric approaches, but not those that only offered screening strategies to identify patients with psychiatric disorders, has led to improved outcomes in patient care (Katon and Gonzales 1994). In the inpatient setting, models of integrated medical and psychiatric care also are being proposed.

3. Future Developments in C-l Psychiatry

Future developments will see an increased emphasis on the education of psychiatrists in collaborating with primary care practitioners in the treatment of medically and psychiatrically ill patients, the development of models of integrated care, and the need to develop and implement guidelines and standards of care in quality assurance programs.

3.1 Education

The education of physicians working at the interface of medical–surgical and psychiatric medicine has to be twofold. Psychiatrists have to learn how to treat medically ill patients with psychiatric comorbidity. Psychiatric training cannot remain restricted to gaining experience predominantly with patients with schizophrenia or manic-depressive illness, but has to take into account the needs of patients suffering from medical illnesses and comorbid psychiatric illness, for example, depressive patients suffering from cardiac disease or stroke, who require somatic as well as

psychological treatment. To achieve this goal, c-l psychiatric curricula are being developed within specialist psychiatric training, e.g., in Europe and the USA. Psychiatrists will have to become proficient in mastering diagnostic and treatment problems arising from somatopsychic comorbidity, e.g., address problems of diagnosing depressive disorders in the medically ill, and adjusting psychological as well as psychopharmacological treatment issues for patients suffering form both medical and psychiatric disorders.

On the other hand, c-l psychiatrists will have to expand the education of primary care physicians, in order to broaden nonpsychiatric doctors' knowledge of and experience in diagnosis, referral, and treatment of patients with psychiatric (co-)morbidity, in order to improve the care of patients with somatopsychic disorders in primary and hospital care settings. Psychiatrists, as well as nonpsychiatric physicians, will have to adopt behavioral medicine treatment strategies for patients with for example, somatization disorders and chronic pain syndromes, and will have to learn treatment strategies other than psychotropic drug prescribing, strategies such as cognitive behavioral treatment for patients with adjustment disorders, anxiety disorders, and the milder depressive disorders (so-called subthreshold disorders), as these disorders make up the bulk of psychiatric comorbidity treated in the primary care sector.

3.2 Guidelines and Quality Assurance

As the awareness for an improved psychological care of medical and surgical patients increased toward the end of the twentieth century, one step has been the development of guidelines in order to define standards of care. In the UK, conjoint working parties of physicians, surgeons and psychiatrists (Royal Colleges of Physicians, and Psychiatrists 1995, Royal Colleges of Surgeons and Physicians 1997) published documents with practical suggestions regarding methods of identifying potentially harmful psychological components of patients' illnesses, and recommendations for the provision of care as well as for referral for psychiatric help when appropriate. As only a minority of patients will need psychiatric care in addition to medical–surgical treatment, these guidelines underline that normally physicians and surgeons should be able to make some assessment of psychological factors. The guidelines propose simple screening mechanisms in order to improve detection of psychological disorders in the general hospital, for example, the *CAGE Questionnaire* for screening for alcohol misuse. Overall, the implementation of guidelines is thought to help in a number of areas, such as in reducing the number of investigations performed for physical symptoms that result from underlying stress, as, for example, in patients with somatization; help in relieving symptoms of stress and improving the quality of life of patients with serious physical illness; and, from an economic perspective, reducing the length of stay in hospital caused by psychiatric comorbidity.

It is assumed that the implementation of guidelines within quality assurance programs will lead eventually to improved patient care. However, measures against which to gauge c-l psychiatric interventions have yet to be established. Given the low referral rate for psychiatric consultation as compared with the high prevalence of medical–psychiatric comorbidity, and the need for specific psychiatric treatment in general hospital inpatients, as a first step to improving care, the C-l Psychiatric Section of the Royal Australian and New Zealand College of Psychiatrists proposes to define the benchmark for psychiatric referrals to be 5 percent of all admissions to medical–surgical wards of a general hospital. If this percentage is not reached, an audit should be performed to establish whether there are identifiable reasons for not providing psychiatric support in an adequate number of patients in medical–surgical departments of a given general hospital.

3.3 Research in C-l Psychiatry

Basic requirement for research in c-l psychiatry is the identification of high-risk patient groups, such as those suffering from a significant amount of psychiatric comorbidity. Failure to identify such high-risk groups, along with the overstated assumption that just one psychiatric consultation might lead to improved patient outcomes may have been the basis of the ambiguous results of general hospital inpatient intervention studies performed in the USA in the 1980s. C-l research in the primary care setting, on the other hand, has demonstrated that restriction to a high-risk group with identifiable psychiatric comorbidity, e.g., high utilizers in primary care suffering from a psychiatric disorder, such as somatoform disorder, may that demonstrate a psychiatric intervention can lead to improved patient outcomes, if the intervention is intense enough, i.e., consisting of an eight-session block of group therapy (Katon and Gonzales 1994).

Problems that need to be solved in order to improve care of patients with medical–psychiatric comorbidity are:

(a) To develop a more precise description of the phenomenology of psychiatric illnesses comorbid with particular physical illnesses, e.g., as in the case of the diagnosis of depression in cancer or stroke patients (Guthrie and Creed 1996).

(b) To study the natural course of psychiatric comorbidity in relation to an underlying or con-

comitant physical disease, e.g., in which patients suffer from a depression not expected to be resolved spontaneously once the physical disease has been treated successfully.

(c) To further study efficacy and effectiveness of psychotherapeutic and psychopharmacological treatment strategies in patients with medical comorbidity. Meta-analyses point out that antidepressants are as effective in depressive patients with medical comorbidity as in those without (Gill and Hatcher 1999).

The effectiveness of different models of care for patients with medical–psychiatric comorbidity with regard to different outcome measures such as clinical improvement, quality of life, or cost effectiveness, will require further investigation, especially in inpatient settings for different patient populations, e.g., acute vs. chronic care patients, with regard to differential effects of more traditional consultation or liaison approaches vs. models of integrated medical–psychiatric care.

3.4 Integrated Medical and Psychiatric Services

A model for integrated medical and psychiatric treatment programs within the general hospital are so-called medical–psychiatric units. Such units received increasing interest in the USA during the 1990s. Medical–psychiatric units are run jointly by physicians, mainly internists, and psychiatrists. They care for high-risk patients, for example, delirious patients who cannot be managed on a medical ward, or suicidal patients in need of acute medical treatment that cannot be provided on a psychiatric ward. It seems possible that such conjoint models of care in the inpatient setting lead to improved patient outcome while saving costs, as compared with traditional models of sequential care with referrals from psychiatric to medical departments and vice versa (Rundell and Wise 1996). They may be the new wave of the future in a changing health care environment that will be looking for more effective ways of treatment beyond the boundaries of traditional departmental structures of general hospitals.

4. Epilogue

Readers interested in widening their knowledge of the topics dealt with in this article are referred to the *Consultation-Liaison Psychiatry Database*, compiled and in part commented upon, by international experts in c-l psychiatry, which contains approximately 2,500 citations of articles, book chapters, and books delineating the state of the art in the field (Strain 1999).

See also: Caregiving in Old Age; Comorbidity; Geriatric Psychiatry; Health Care Organizations: For-profit and Nonprofit; Hospitals and Health: Psychosocial Aspects; Managed Care; Mental Illness, Epidemiology of; Nursing Homes; Program Evaluation; Psychiatry, History of; Psychosomatic Medicine; Public Psychiatry and Public Mental Health Systems; Somatoform Disorders

Bibliography

APM (Academy of Psychosomatic Medicine) 1997 *Mental Disorders in General Medical Practice: Adding Value to Healthcare through Consultation-Liaison Psychiatry*. Kendall/Hunt, Dubuque, IA

Arolt V 1997 *Psychische Störungen bei Krankenhauspatienten* (*Psychiatric Disorders in General Hospital Inpatients*). Springer, Berlin

Diefenbacher A 2000 Consultation and liaison psychiatry. In: Helmchen H, Henn F, Lauter H, Sartorius N (eds.) *Contemporary Psychiatry. Vol 1, Part 2: General Psychiatry*. Springer, Berlin, pp. 253–67

Gill D, Hatcher S 1999 Antidepressant drugs in depressed patients who also have a physical illness (Cochrane Review). The Cochrane Library, Issue 2. Update Software: Oxford, UK (www. cochrane. org/cochrane/revabst/ab001312.htm)

Guthrie E, Creed F (eds.) 1996 *Seminars in Liaison Psychiatry*. Gaskell, London

Katon W J, Gonzales J 1994 A review of randomized trials of psychiatric consultation-liaison studies in primary care. *Psychosomatics* **35**: 268–78

Reichmann W E, Katz P R (eds.) *Psychiatric Care in the Nursing Homes*. Oxford University Press, New York

Royal Colleges of Physicians and Psychiatrists 1995 *Joint Working Party Report: The Psychological Care of Medical Patients*. London

Royal Colleges of Surgeons and Physicians 1997 *Report of the Working Party on the Psychological Care of Surgical Patients*. London

Rundell J R, Wise M G (eds.) 1996 *The American Psychiatric Press Textbook of Consultation-Liaison Psychiatry*. American Psychiatric Press, Washington, DC

Strain J J 1999 Consultation-liaison psychiatry database (2000 update). *General Hospital Psychiatry* **21**(6): 402–502

A. Diefenbacher

Liberalism

Liberalism is a family of political philosophies, and a set of associated institutions and policies. All liberal philosophies give primacy to the protection of basic liberty. Competing versions of liberalism—such as democratic liberalism and libertarianism—differ in their answers as to what basic liberty is, and what institutions best protect it. Since its origins, liberalism itself has been an object of controversy. At the same

time, liberalism has been committed to defending the set of freedoms—of speech, press, conscience, and association—that support the rights of all parties to carry on such controversies in public.

1. Origins

Although the term 'liberalism' was not used until the early nineteenth century, major elements of liberal political philosophy originate in the seventeenth century. The first systematic defense of a politics that gives priority to individual liberty was John Locke's *Two Treatises of Government* ([1690] 1988). Against Thomas Hobbes's defense of absolute sovereignty (*Leviathan* 1651), Locke argued that individuals have a set of natural rights to life, liberty, and property that they do no alienate to any government, and they therefore have a right to revolt against a government that violates those rights. The American Revolution of 1776 has Lockean liberal roots, as does the political thought of James Madison and other framers of the US Constitution.

Adam Smith offered the first systematic exposition of liberal economic theory (*Wealth of Nations* 1776). Social wealth is generated not by direct state action but by the invisible hand of free competition among producers and consumers. Smith presents a liberal plan of equality, liberty, and justice that links commercial freedom to a constitutional political order protective of civil and political liberties. Although Smith uses 'liberal' in its then everyday sense of open-mindedness of spirit, he links open-mindedness to a political program that grants individuals the greatest freedom from interference compatible with a like freedom for all.

On the Continent, liberal political theory also developed throughout the eighteenth century: The major liberal thinkers of the French Enlightenment criticized the *ancien regime*; Voltaire championed individual freedom of thought and religion (*Lettres philosophiques* 1734); and Montesquieu defended a system of checks and balances and separation of powers as means of protecting individual freedom (*L'Esprit des Lois* 1748). In Germany, Wilhelm von Humboldt defended limited government as a means to facilitate the development of human powers to their highest level in *Ideen zu einem Versuch, die Grenzen der Wirksamkeit des Staats zu bestimmen (Ideas Toward an Investigation to Determine the Proper Limits of the Activity of the State* 1792).

Not until 1812, however, was the term 'liberalism' actually used in politics. The Spanish political party, the *Liberales*, mounted a Lockean defense of constitutional monarchy, parliamentary government, and human rights against absolute monarchy. Ironically, Tory politicians in England called their domestic opponents *'liberales'* rather than 'liberals' in order to suggest that liberal politics were un-English. Before long, however, the term 'liberal' became widely associated throughout the Anglo-Saxon world with a defense of individual liberty and a critique of absolutist government.

2. Development of Conceptions and Controversies

As the influence of liberalism increased throughout the nineteenth and twentieth centuries, different conceptions of liberalism developed, as did controversies between them, and between liberals and their critics. Three major controversies, each discussed below, are the debate between liberalism and libertarianism; conservative, radical, and communitarian critiques of liberalism; and the emergence of political liberalism as an alternative to comprehensive liberalism.

2.1 Democratic Liberalism vs. Libertarianism

The early nineteenth century French philosopher Benjamin Constant contrasted 'the liberty of the moderns'—freedom of speech, religion, and association, and other legal protections of the individual from outside interference—with 'the liberty of the ancients'—the freedom of citizens who collectively exercise popular sovereignty (1819/1988). All forms of liberalism affirm the liberty of the moderns, and therefore give primacy to protecting basic individual liberties against the tyranny of any political authority, no matter how popular. But democratic liberalism puts fewer constraints on popular government than libertarianism.

Two related concerns—about how to protect individual freedom and how to realize the full value of individual freedom—lead many liberal thinkers to ally liberalism closely with democracy. Individuals who are preoccupied with their own private lives may neglect the need to guard their political guardians and they also may be prevented by their government from developing some important part of their own potential. An antidote to these dangers, according to Constant, is political liberty, which will have the effect of enlarging the vision and spirit of citizens, and opening up more public offices and activities to ordinary citizens.

Democratic liberalism condemns the tyranny of the majority but defends representative democracy (John Stuart Mill, *On Liberty 1859* and *Considerations on Representative Government 1861*). Along with *Alexis de Tocqueville* (*Democracy in America 1835*), Mill argues that without effective opportunities to participate in political and civic associations, ordinary people will be easily manipulated by demagogues and public opinion, and individual freedom will become a hollow ideal. Democratic liberalism therefore includes both political and personal freedoms among the most basic individual liberties.

The modern model of liberal democracy differs in two important ways from ancient Athenian democracy. First, most citizens do not directly participate in policy making but are represented by public officials (who are to be held accountable by frequent, free, and fair elections). Second, government is constitutionally required to respect individuals' personal liberties of individuals, not only their political freedoms.

Contemporary conceptions of democratic liberalism support taxation for many public purposes, not only to provide for the common defense and the education of children, but also to maximize the worth of liberty for the least advantaged citizens, as John Rawls argues in *A Theory of Justice* (1971). The power of government is limited for the sake of protecting basic individual liberties, but the limits on government are viewed as compatible with regulating the economy in order to protect the worth of liberty for all individuals. Equality and democracy are regarded as allies of liberty. The overarching aim of liberalism, according to Ronald Dworkin's 'Liberalism' (1978), is to treat free individuals *as equals*, but not equally. Judith Shklar (1998) argues that liberalism requires democracy, because political rights and freedoms are needed to defend freedom. Other liberals defend an increasingly popular conception called deliberative democracy, which defends a politics that respects the basic freedoms and opportunities of all individuals and also encourages publicly accountable deliberation to help resolve political disagreements and promote mutual respect among citizens amidst the disagreements that remain unresolved (Gutmann and Thompson 1996).

Libertarianism, by contrast, downplays both democracy and equality. The classical libertarian Herbert Spencer (1894/1981) defends the unfettered economy on pragmatic grounds—as a way of maximizing social wealth—but also on principled grounds as a reward for individual desert, which he argues is best measured by a free market. Libertarians accept the need for a minimal level of taxation to provide necessary public goods such as a common defense. But taxation for the purposes of redistribution, as contemporary libertarians like Robert Nozick argue (1974), is on a par with forced labor, a clear violation of basic liberty. Libertarians begin with a premise of self-ownership. If individuals own themselves, then they have a right to their natural assets. More controversially, libertarians argue that individuals also have a right to whatever flows from the politically unfettered use of those assets in society.

Democratic liberals and libertarians are united in defending a broad set of civil liberties (freedom of speech, press, association, conscience, and religion), a market economy, and the right to own personal property, but they part company over taxation for social welfare purposes. Whereas democratic liberals subscribe to the American revolutionary slogan 'no taxation without representation', libertarians—who often identify as 'conservatives' in US politics—defend freedom from taxation except when absolutely necessary to support the minimal—'night watchman'—-state. This dispute can be viewed as a family quarrel, and US politics is often characterized as an arena of conflict among competing versions of Lockean liberalism (Hartz 1955).

2.2 *Conservative, Radical, and Communitarian Critiques*

Conservative, radical, and communitarian political perspectives each pose a different challenge to the liberal emphasis on the priority of individual liberty.

The disagreement between Edmund Burke and Thomas Paine ([1992]/[1792] 1995) over the French Revolution reveals the basic tension between conservatism and liberalism. Whereas the American revolutionaries, according to Burke, were fighting to preserve their traditional birthright (the rights of Englishmen), the French revolutionaries were trying to remake their society, a treacherous task, which would 'like a palsy' attack 'the fountain of life itself' (Burke, *The Revolution in France* 1790). Paine avidly supported both revolutions in the name of the kind of liberal philosophy—a universalistic defense of human rights–that is anathema to conservatives like Burke and more recently Michael Oakeshott (1991 *Rationalism in Politics and Other Essays*) who develops a more philosophical version of the conservative critique of liberal rationalism. Liberalism today is defended in theory as in practice on both particularistic and universalistic grounds.

Whereas conservatives argue that liberalism goes too far in defending individual rights, radicals argue that it does not go far enough. According to the radical critique, inspired by the early Marx, liberalism's defense of individual freedom is sorely insufficient to the aim of actually liberating individuals from the oppressive power relationships that are endemic to a capitalistic economy, to which liberalism is committed. Liberal egalitarians (Gutmann 1980) meet radicals on their own ground by arguing that liberalism supports whatever economic system will best secure the worth of liberty for the least advantaged citizens, not just the most advantaged. Other radical critics argue that the coercive and manipulative powers of the state apparatus, modern media, and mass public opinion combine to undermine any possibility of individual liberation. Radical feminists mount a more focused attack on the liberal right to publish pornography, arguing that it furthers the oppression of women.

Drawing upon a conception of humans as historically conditioned beings, communitarians attack liberalism at its foundations for giving priority to individual freedom over the communal values—such as a shared religious faith—that give meaning to

human lives. Liberalism defends the priority of individual rights over the preservation of communal values. It accepts the idea of human beings as historically conditioned and still consistently defends individual liberty as a primary value that politics should protect. The enforcement of individual rights, liberals argue, is necessary to protect people against the contemporary equivalents of witch hunting. Speculation that communal bonds of fraternity can take the place of justice in a modern pluralistic society has not been strong enough to undermine the liberal defense of individual rights (Gutmann in Avineri and de-Shalit 1992).

2.3 Political Liberalism as an Alternative to Comprehensive Liberalism

John Rawl's *Political Liberalism* (1993) marked an important new development in liberal theory. Other conceptions of liberalism—such as the perfectionist and autonomy-based theory of Joseph Raz (1986)—offer comprehensive philosophies of life, which include ideals of personal character, friendship, and association. By contrast, political liberalism is a moral conception concerned only with the basic structure of a society, and presented as a view that is independent of any comprehensive doctrine. Its fundamental terms—such as the idea of individuals as free and equal citizens—are derived from the public political culture of a liberal democracy. Political liberalism recommends its principles as a way of fairly resolving disagreements in a pluralistic society; it does not expect free and equal citizens to agree on one comprehensive conception. Political liberalism instead locates an overlapping consensus in the midst of reasonable pluralism. Like liberalism itself, political liberalism is a family of conceptions, which share the aim of finding fair terms of social cooperation among free and equal members of a democratic society.

Accompanying the development of political liberalism are other conceptions of liberalism, including feminist liberalism and multicultural liberalism. These build on the basic premises of either political or comprehensive liberalism, and explicitly address the concern that women and cultural minorities should not be excluded from enjoying the full benefits of liberal freedoms and opportunities by virtue of their gender, race, ethnicity, or cultural identity. The family of different conceptions of liberalism continues to grow rapidly at the same time as various forms of liberal democracy multiply around the world.

See also: Bill of Rights; Communitarianism: Political Theory; Constitutionalism; Deliberation: Political Aspects; Democracy; Egalitarianism: Political; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Libertarianism; Locke, John (1632–1704); Montesquieu, Charles, the Second Baron of (1689–1755); Smith, Adam (1723–90); Tory; Whig

Bibliography

Avineri S, de-Shalit A (eds.) 1992 *Communitarianism and Individualism*. Oxford University Press, Oxford, UK

Coustant B 1819/1988 The liberty of the ancients compared with that of the moderns. In: *Political Writings*. Cambridge University Press, Cambridge, UK

Dworkin R 1978 Liberalism. In: Hampshire S et al. (eds.) *Public and Private Morals*. Cambridge University Press, Cambridge, UK, pp. 113–43

Gutmann A 1980 *Liberal Equality*. Cambridge University Press, Cambridge, UK

Gutmann A, Thompson D 1996 *Democracy and Disagreement*. Belknap Press of Harvard University Press, Cambridge, MA

Hartz L 1955 *The Liberal Tradition in America*. Harcourt Brace Jovanovich, New York

Locke J [1690] 1988 *Two Treatises of Government*. Cambridge University Press, Cambridge, UK

Mill J S 1975 *John Stuart Mill: Three Essays*. Oxford University Press, Oxford, UK

Nozick R 1974 *Anarchy, State and Utopia*. Basic Books, New York

Oakeshott M 1991 *Rationalism in Politics and Other Essays*. Liberty Press, Indianapolis

Paine T [1792] 1995 *The Rights of Man*. Oxford University Press, Oxford, UK

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Rawls J 1993 *Political Liberalism*. Columbia University Press, New York

Raz J 1986 *Morality of Freedom*. Clarendon Press, Oxford, UK

Ruggiero G de 1959 *The History of European Liberalism*. Beacon Press, Boston

Shklar J 1998 The liberalism of fear. In: Shklar J (ed.) *Political Thoughts and Political Thinkers*. University of Chicago Press, Chicago, pp. 3–19

Spencer H 1884/1981 *The Man versus the State*. Liberty Press, Indianapolis

A. Gutmann

Liberalism and War

Liberalism is the label attached to the philosophies, doctrines, and political movements that are concerned with the relationship between the exercise of political authority and the pursuit of welfare by individuals and groups governed by that authority. No other political and philosophical tradition has sustained as intense and long-standing an effort to eradicate war from modern political life, or at least mitigate its affects on the welfare of societies. Liberal ideas continue to attract thinkers and activists around the world as they struggle to overcome deadly conflicts within societies and the possibility of large-scale war across the globe.

1. The Liberal Tradition

With precedents running back at least to the seventeenth century, liberalism became an established intellectual and political movement from the last

decades of the eighteenth century onward. Figures as famous as Thomas Paine, Immanuel Kant, John Stuart Mill, Richard Cobden, L. T. Hobhouse, Woodrow Wilson, and Thomas Dewey are among the ranks of public intellectuals associated with the liberal tradition.

Liberals have addressed a diverse range of social issues from economic life to education. Common to that diversity is the self-understanding of liberals that their goal is to maximize the well-being of members of society through increased freedom, equity, or security. Political and civil rights, economic and market freedoms, and provisions for the redistribution of wealth have been turned to as means to achieve this well-being. The varying and sometimes contradictory nature of the values—freedom, equity, and security—within the liberal tradition means that liberalism is not reducible to any fundamental principle or doctrine, despite common concerns with political authority and human welfare. Indeed, states and societies that can be counted as liberal have taken quite different forms depending on whether individual economic rights are stressed over the redistribution of wealth or the social security of families.

Liberals have naturally been focused on the political and social life of individual states and societies, since the perfection of civil and political society has been so central to their aspirations. But many liberals, recognizing that domestic life is shaped by forces and processes external to national boundaries such as war, have focused their attention on the nature of international life.

In the international realm, liberal concerns have been concentrated in a number of areas. Most famously, multilateral economic exchange and the relatively free movement of goods and capital across state boundaries (or international 'free markets') have preoccupied liberals. Open economic relations tie states, civil societies, and domestic market economies into complex networks of international economic relations. These relations and networks are taken by liberals to expand the economic freedom of market actors and limit the restrictions to that freedom inherent in state-defined borders. From Adam Smith in the late eighteenth century and the Manchester School of free traders in the nineteenth century, to the architects of Bretton Woods during World War II, liberals have generally assumed that economic openness is essential to well-being within and across national borders.

A second area of liberal concern has been the advancement of liberal forms of government across the globe. From Kant and Woodrow Wilson through Francis Fukuyama (1993), liberals to varying degrees have considered the global spread of representative government, the rule of law, constitutional limits to executive power, and democratic processes as essential to their social vision. Viewing political liberalization as an international process, spreading from nation-state to nation-state, has been compelling for liberals most basically because any discussion of liberalism as a real political and social force in the world has little meaning without the presence of at least some liberal social spaces within it.

Liberals, especially in the twentieth century, have also made the international spread of human rights—political, civil, and economic—a process critical to the promotion of liberalism. Indeed, rights are essential for all areas of liberal concern, from markets to political representation. While the struggle for rights has mostly taken place through political movements operating within individual nation-states, there has been a quite robust international effort to advance a wide range of rights. In the twentieth century human rights were encoded in myriad international declarations at the UN and addressed by many nongovernmental organizations operating on an international basis. Precedents in the nineteenth century include the transnational struggle against the slave trade and protections for minorities and citizens of other countries living abroad through international treaties.

A fourth area of focus in the international realm is national self-determination, which involves the legitimation of claims made in the name collectivities to establish and maintain independent states for self-government. It is the only area of social action in the international realm where there occurred anything like the sort of struggle against authority and power that liberal political and social movements undertook within individual domestic spheres. John Stuart Mill (1991) expressed the liberal belief that without independence it is not possible to fashion 'free institutions.' Liberals in the nineteenth century often supported nationalist movements for independence within Europe at the same time that they overlooked the contradictions posed by the empires their own states and societies pursued outside of Europe and its immediate periphery. Nationalist leaders emerging from the colonial world were, however, able to exploit the principles of self-determination to make claims against these very empires in the mid-twentieth century.

Also, some liberals have believed that human welfare can be advanced by the formation of an international community of states linked for the common purposes of peace, economic progress, or the strengthening of systems of international law. Others have gone further in advocating the construction of a transnational or universal community of humankind within which rights, entitlements, and human security are realized.

2. *War and Society*

A sixth area of concern to liberals has been war. For over 200 years advocates of liberal principles and practices in government, the media, the academy, and

public life have treated most war as inimical to human welfare. At the same time that liberal movements and practices became more widespread across the nineteenth and twentieth centuries the range and extent of destruction associated with war expanded. By the first half of the twentieth century states had devised new strategies to extract massive levels of material resources and human capital from society for war, and to apply them more effectively on the battlefield and ultimately against civilians living in rural and urban settings.

Of course, by almost any account, war is a disruptive force. While the loss of life and resources due to war can be staggering, liberals have had other reasons to be troubled by its destructive effects having to do with their interest in the construction and maintenance of a liberal commonwealth. Their interest in effective civil society, representative government, economic and political freedoms, operational markets, and systems of wealth redistribution has made the stakes of disruption especially high for them. War can lead to the suspension of rights (such as *habeas corpus*); the restraint of activities in the public sphere (such as in media 'black outs'); the rechanneling of economic life (such as in wartime command economies); the concentration of state power (such as in the enactment of martial law); and the restriction of individual access to basic necessities for physical survival (such as in the wartime rationing of food).

In addition, an issue for liberals regarding the effects of war that was prominent prior to World War I is the fear that the occurrence of organized, interstate violence reinforces the dominance of belligerent aristocratic classes over state and society. War and a militant aristocratic culture, in this view, provided an opportunity for the nobility to maintain their dominant positions within the war-making state and to prevent other forms of political and social organization attractive to liberals from becoming more prevalent. It was these very classes, forming an ancien regime of hereditary privilege, that liberals saw themselves struggling against in their pursuit of more diverse forms of political representation and more open economic relations based on robust markets.

While concerns with the impact of war on life inside individual states and societies have dominated the agenda of liberals, they have not ignored the effects of war directly on practices and structures in the international realm deemed relevant to liberalism. In the international realm war could cause disruptions and lead to injustices across a wide range of concerns. Established structures of international commerce and investment can be halted through attacks on trade routes; human rights can be violated across national boundaries through invasion; national aspirations to self-determination and sovereign independence can be undermined through military occupation; bodies of international law can be weakened through blatant disregard; and the international spread of norms of democratic governance can be constrained by subversion and conquest. For liberals concerned with the building of an international or universal community, war is seen as markedly debilitating since states typically take sides in exclusive alliances and wartime pacts that starkly define enemies and friends.

Alternatively, wars can stimulate significant forms of political and social change relevant to liberals. Most directly, liberal causes have been advanced by large-scale, organized violence in the name of liberty, from the American War for Independence onward. Wars can also provide political actors with new conditions and historical contexts within which to operate. For example, out of World War II emerged the Declaration of Human Rights, new international structures for trade and investment, a renewed and more broad-based commitment to democratic governance, a greater legitimacy granted to claims to self-determination by colonized peoples, and new forms of international political community among states in the West. Governments also experimented with new systems of income redistribution through welfare, progressive taxation, and social security. Some liberals like Immanuel Kant (1970) have argued that war can help establish—and strengthen existing—republics. Others have argued that war can be justified in the name of self-determination against tyranny and civil and political rights against oligarchies. But, as recognized by historian Michael Howard (1986), liberals historically have never really come to terms with the fact that war can be a productive force for social change and not just a practice automatically condemned.

3. *Liberal Reforms for Peace*

The long-standing critical attitude of liberalism to war has carried some force in public life in part because thinkers in the liberal tradition have typically combined their condemnations of war with broad-ranging and complicated programs for actually counteracting the conditions and forces that lead to it. These include changes in the political regimes of states; the structure of societies; the nature of the international economy; the character of international life; and the mechanisms available to foster peaceful interactions.

3.1 *Domestic Regime Change*

In 1795, Immanuel Kant made famous in his essay *Perpetual Peace* (1970) the proposition that certain forms of government are more prone to war than others. Specifically, he argued that if a regime could be organized along liberal lines (with a constitution providing for full-fledged representation, a separation of powers, and civil rights) it would be less likely to go to war based on the will of a monarch and more likely

to enter into long-standing peaceful relations with other liberal states. Liberal states, Kant argued, allow for public discussion of the purposes and costs of war and require popular consent for it to be waged.

Although early-twentieth-century leaders like Woodrow Wilson and Franklin D. Roosevelt endorsed versions of the Kantian perspective, it would take 200 years for the proposition 'democracies do not fight one another' to become a central tenet of students of international relations and policymakers in the US government (see Doyle 1997). Convinced of its empirical reality, scholars are still searching for the exact factors that produce the 'democratic peace,' from the acceptance of norms of peaceful conflict resolution in democratic societies to the existence of robust forums for public opinion. Upon its strength, policymakers extol the virtues of promoting and sustaining democracy around the world as a means to help eradicate war. It is generally recognized by scholars, however, that while democratic states are less likely to wage war on one another, they hold out no special pacific tendencies when it comes to war with non-democratic states.

3.2 Transformations in Social Structure

Among the conditions that made the establishment of a liberal regime so crucial for Kant was its power to displace a system of absolutist monarchy that relied on the aristocratic war-making classes described above. For liberals such as Thomas Paine, the noble classes of Europe had not only developed an interest in war as an honorable pursuit that secured their positions within their own countries, it also provided the basis for a Europe-wide transnational aristocratic culture of diplomacy and war. This exclusive culture was seen as being based on secrecy in negotiations and private considerations when it came to decisions about war and peace. Although the revolutions in Europe that occurred from the late eighteenth to the mid-nineteenth centuries had replaced many absolutist with constitutional regimes, aristocrats remained a powerful force in state and society. Thinkers like Joseph Schumpeter (1989) could look back and blame imperialist wars on the continued influence of the aristocracy on the foreign policy of European states.

The answer for most liberals was to advance the political and social fortune of the middle classes within each state and society. Not only did they not carry with them values and codes of war-making, their interests and security were seen as anchored in commercial life and the possession of a wide range of rights for which large-scale conflict was inimical. To the extent that the middle class—or 'bourgeoisie'—could become the predominant political class across Europe there would emerge an entirely new 'peace interest' and culture of public diplomacy.

As the aristocratic influence waned, from the twentieth century on, liberals such as Norman Angell (1972) began to focus more intently on another powerful class: large military industrialists who had a deep commercial interest in the profits of war. And once again, the peace interest of the middle classes was also considered an antidote.

3.3 The Development of International Trade

In much liberal thinking, the rise of commercial classes was to be complemented in the international realm by the emergence of a robust system of international free trade for them to participate in. An overriding concern of liberals in this area has been to counteract mercantilist tendencies in the foreign policies and relations of states. Mercantilism was associated with the pursuit by autocratic states of exclusive economic access to trade routes and special concessions for the extraction of resources—the economic benefits of access for one state would be closed off to other states. Liberals feared that gaining access or enforcing exclusivity constituted grounds for war.

Even as mercantile practices became less prominent in the nineteenth and twentieth centuries as compared with the seventeenth and eighteenth centuries, liberals still worried that protectionist policies could reinforce the sense of rivalry between states. Indeed, colonialism, which had reached new heights in European policy at the end of the nineteenth century, was mostly understood by liberals to be a mercantilist phenomenon that fed competition (annexation brings a set of people and places into the orbit of one state and not another). Liberals generally believed that states were willing to go to war over colonies to protect their interests.

The effectiveness of international trade as an antidote to these practices in the international realm rests on the prospect that, instead of seeking colonies or exclusive access, states would simply promote trade between their citizens and the rest of the world. A vested state interest in securing open access to resources, peoples, and places would develop and concerns with facilitating market competition between individual economic actors would shape policies. The connections between individuals across national boundaries would deepen and widen and rivalry would take form in the productive field of the market rather than the interstate system. This ultimately would reinforce, as John Stuart Mill (Silberner 1946) believed, an interest in peace rather than war (an activity that could be highly disruptive in the international economy).

3.4 The Emergence of New Forms of International Life

According to many liberals, the development of more direct transboundary connections between individuals would make possible the worldwide cosmopolitan

community mentioned above. Commerce, as well as travel and international media, would help create a global body of public opinion, a shared identity as *humanity*, and a common understanding of human rights and justice. This cosmopolitan community of individuals would undermine the belligerent ambitions of governments by making it difficult to mobilize national citizenries for war.

Not all international communities of interest to liberals were grounded in a universal humanity. In the nineteenth century, liberals like the Italian nationalist leader Giuseppe Mazzini and the British Prime Minister William Gladstone argued that Europe constituted a community of nations. If each nation, satisfying its desires for national self-determination, could devote itself to its own national destiny and recognize other nations as part of a common family of nations, then the motives for war would dissipate.

Democracy was a third ground for community besides cosmopolitanism and nationalism. As mentioned above, liberals like Kant and Woodrow Wilson believed that a league of democratic states could form a community capable of preventing war and advancing peaceful means of conflict resolution.

Finally, some liberals have argued that states could form a world federation, bounded by international law and mutual political obligations, within which war would be an unthinkable practice.

3.5 The Establishment of New International Mechanisms

Liberal ambitions regarding international community have gone hand-in-hand with liberal expectations that international law could become a powerful force shaping the action of states within a community of nations. The execution of treaties and pacts and the development of legal theories since the seventeenth century have fueled the hope that normative proscriptions against unilateral decisions to wage non-defensive wars could effectively constrain officials and compel them to pursue peaceful means of conflict resolution and avoid violent intervention in the affairs of foreign countries (which is also contrary to the principle of self-determination).

Other mechanisms pursued by liberals to promote peace besides law include collective security. The basic premise is that states forming a league should agree to avoid war with one another and be willing to punish with sanctions or even collective military action any state that violates international peace and security. This principle was built into both the League of Nations formed after World War I and the UN of today.

Two other relevant mechanisms are arbitration and disarmament. Most liberals have argued that states with disputes should pursue their resolution through the mediation of an institution such as the UN or other third parties such as neutral governments. If arbitration became an accepted part of international relations, then war would become the last rather than the first recourse of states.

Disarmament has been an enduring liberal goal that reached its greatest prominence in the second half of the nineteenth century. Transnational peace movements grew in number on both sides of the Atlantic around the hope that states would abandon their readiness to prepare for and wage war. Many of these associations called for special congresses and treaties, and petitioned governments like Great Britain to take the lead in promoting disarmament.

4. Conclusion

The contrast between the liberal record in the domestic and international spheres with regard to peace and security is sharp. Throughout the nineteenth and twentieth centuries, liberals and their supporters struggled and often succeeded at constructing liberal states and societies in the West, helping to secure individuals and groups against arbitrary violence by states and factions. Liberals, as shown above, also tried to refashion the relations between states along more peaceful lines. However, relative to the domestic sphere they had far less impact. It is true that nineteenth- and twentieth-century thinkers and policymakers associated with the liberal tradition helped to advance the commitments of states to arbitration, disarmament, collective security, and the building of international community. Nonetheless, to this day no institutions exist to tie states and societies in a common security framework that is binding and that limits their ability to wage war.

Liberals certainly changed the discourse of international security. They also helped establish channels of negotiation. However, liberals did not alter the way states organize themselves regarding military power and national security. These practices have remained firmly grounded in the classic model of the war-making state. And while phenomena such as treaties to limit the proliferation of weapons, UN peacekeeping forces, and liberal zones of peace in the West are potential alternatives that currently challenge the classic war-making model, none of these are comparable to the transformation of absolutist into liberal, constitutional states in the West.

There are two important factors to bear in mind when considering the limitations of liberal efforts to promote reforms leading to peace. First, attempts to reform the security sphere were mostly undertaken by a rather narrow range of thinkers, citizens, and statesmen such as Richard Cobden working through relatively small antiwar organizations. If it typically took broad-based, popular political challenges to authority (including the revolutions of 1848) to change domestic spheres, we certainly should not expect to see the reformation of security practices based simply on

the decision of government officials to heed their own conscience or the call to reason by liberal publicists and advocates.

Second, liberals have simply invested most of their intellectual and political capital in the pursuit of more perfect states and societies. Thus, international life has not received the same level of attention. And while liberals have been concerned with interactions between groups and individuals across national boundaries (e.g., trade) and the possibilities of new forms of international community (political union), liberal international political thought has been focused on the relations of states. Indeed, this might be expected since, in the absence of international government, there is no legal or binding political basis to pursue and secure claims important to liberals regarding citizenship, economic rights, or equality, within the international realm.

That there are limits to the liberal approach to war is a claim that has been made by critics of the tradition, who fall broadly into two camps. Realists like Kenneth Waltz (1965) have argued that liberalism offers a misplaced faith in the prospects for and affects of change and reform, both domestic and international. They also are skeptical of the ultimate willingness of states to abandon their own management of their security and capacity to wage war. And they generally fault liberals for failing to recognize the truly enduring presence of war in a world of ambitious and security-minded states.

A second line of critique has come from thinkers associated with the political left who fault liberalism for positions on war that obscure the exclusions and violence liberal states, in practice, underwrite through their foreign policies (see, for example, Robinson 1996). For instance, liberal concerns with building a union of liberal democracies have been condemned for justifying militarized alliances such as NATO that exclude non-liberal states and societies in the developing world and which emphasize differences and security threats rather than commonalities and community across the globe. Likewise, the pursuit of a global liberal economic order has been criticized for leaving the developing world behind in poverty and subjecting unprepared governments to policies that, in the name of privatization, place too much faith in the market and undermine public authority. The shrinking of public authority is often linked to increases in internal conflict within the national boundaries of states no longer capable of ensuring public security. And while liberal principles hold out the hope of a more peaceful world, liberal states have in practice been denounced for helping to perpetuate significant programs of overt and covert violence against individuals and political groups and aiding military build-ups in the developing world, most notably during the Cold War. This violence was sometimes small scale; other times it took on the dimensions of a full-scale civil war (e.g., in Guatemala and Cambodia).

While for some it is easy to separate the pacific values of liberalism from the violent policies of liberal states griped in global conflict, critics, both realist and leftist, remain concerned that the failure to draw out the connections between practice and principle undermines the ultimate impact of the latter.

See also: Area and International Studies in the United States: Intellectual Trends; International Relations, History of; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; National Security Studies and War Potential of Nations

Bibliography

Angell N 1972 *The Great Illusion*. Garland, New York

Doyle M 1997 *Ways of War and Peace: Realism, Liberalism and Socialism*. Norton, New York

Fukuyama F 1993 *The End of History and the Last Man*. Avon, New York

Howard M 1986 *War and the Liberal Conscience*. Rutgers University Press, New Brunswick, NJ

Kant I 1970 *Kant's Political Writings*. Cambridge University Press, Cambridge, UK

Mill J S 1991 *Considerations of Representative Government*. Prometheus Books, Amherst, NY

Robinson W 1996 *Promoting Polyarchy: Globalization, US Intervention, and Hegemony*. Cambridge University Press, Cambridge, UK

Schumpeter J 1989 *Imperialism and Social Classes*. A. M. Kelley, Fairfield, NJ

Silberner E 1946 *The Problem of War in 19th Century Economic Thought*. Princeton University Press, Princeton, NJ

Waltz K 1965 *Man, the State and War*. Columbia University Press, New York

R. Latham

Liberalism: Historical Aspects

Liberalism is a sociopolitical idea about political order that has contributed substantially, since the late eighteenth century, to the modern age. It aims at creating a civil society, but adjusts its concrete demands to the existing situation of society and the state. This readiness to adjust has made liberalism willing to compromise as well as open to development. The core principles of liberalism endure, but the details of what it understands as civil society have changed in the course of history and have always been debated among liberals. For this reason, the history of liberalism as an idea and as political practice can be adequately understood only as a continuing struggle about interpretation, but as controversial as the definition of liberalism has always been, liberalism has always aimed to make responsible citizenship possible. This is true for all societies, even if the paths liberalism has

taken have differed in Europe, the USA, and other parts of the world. Europe is where liberalism originated and formed its liberal ideas of political order. For this reason, we will concentrate on European liberalism.

Since its beginnings, liberalism has demanded three kinds of citizenship rights: (a) equality before the law, guaranteed by the rule of law; (b) equal opportunities for political participation, made possible by the right to vote and to participate in public life, and (c) the provision of the basic prerequisites for opportunities in life. The demand for equality before the law was made and fulfilled earliest. Then came political equality, though in practice liberals, in contrast to democrats, long favored giving different weights to the votes of members of different classes. Whether and to what extent the state should guarantee social rights has remained controversial among liberals to this day.

1. *Liberalism as an Idea of Political Order*

Liberal thinking centers on the individual, whose freedom of decision is to be defended against state and social constraints. All demands for individual freedom can thus be considered part of the liberal tradition. This would make liberalism as old as human memory, but a liberal doctrine of the state and society first arose in seventeenth century England. Its most influential thinker was John Locke (*Two Treatises of Government*, 1690). The European Enlightenment's attempts to limit Absolutism and to enable 'man's exit from self-caused lack of self-responsibility' (Immanuel Kant) and finally the impetus of the American and French Revolutions (human rights, popular sovereignty, constitutionality) are further important stages of liberal thinking.

As political terms, 'liberalism' and 'liberal' were coined in Spain and did not establish themselves until after 1800. The nineteenth century was then the century of liberalism. No other sociopolitical idea of political order changed the state and society as radically. With the idea of the citizen, liberalism sketched a program for the future that was egalitarian in principle and that was directed against everything that limited individual freedom. At that time, this revolutionary program, which liberals nevertheless wanted to realize evolutionarily in step-by-step reforms, was directed primarily against absolutism and the rule of the nobility, and thus also against class and denominational privileges.

The liberals' central political demand was for a constitutional state. It was to guarantee the continuity of law, secure individual property, and enable the citizen to participate in the state. The liberals saw the absolutist state as their antagonist and wanted to tame it with constitutional law. The individual was to have free access to the public political arena and, as voters, exert influence on the state's decision making process, but the state was not to be stripped of its executive powers, only its functionings to be liberalized. The liberals did not call for a weak 'night watchman' state, but for freedom in and through the state. Liberalism's specific political program varied substantially from country to country and over time. What critics condemned as programmatic weakness or even as a betrayal of principles was precisely the strength of the liberal movement. Thanks to its flexibility, it was able to adjust its idea of political order to very disparate conditions of action. Liberals argued everywhere for the parliamentary form of government, but they also accepted compromises to avoid endangering the stability of state and social order.

Among the large European countries, the liberals' will to establish the parliamentarian system of government was weakest in autocratic Russia and in pre-World War I, Germany. Liberals did not insist upon a particular form of state, but their ideal was parliamentary monarchy; they wanted to preserve the monarchy as a reserve constitution for states of emergency. This sharply distinguished them from the democrats, whose ideal was the republic. The European liberals of the nineteenth century were willing to accept the republican form of state (first in France and Italy), but, since the French Revolution of 1789, most liberals saw the republic as burdened with the odium of social revolution and the annihilation of civil society. The liberals' fear of social revolution put them at a distance from the idea of the republic, which could prevail only through revolution wherever monarchical states existed. In republican North America, this was not an issue. Limiting the power of the state and securing private property were the foci of American property-oriented, individualistic liberalism, which emphasized self-help and self-organization. Franklin D. Roosevelt's New Deal liberalism created the USA's first instruments for state intervention in society and economic policy.

In practice, the liberal idea of political order focusing on a society with egalitarian citizenship was compatible with inequality. The liberals of the nineteenth century wanted immediate implementation only of equality before the law; political equality could wait. They saw freedom of opinion and the right to vote as the most important means of apportioning participatory rights in the state and the community. By the early twentieth century, liberals had accepted democratic equal voting rights, but they never advocated or demanded the establishment of such rights. Rather, the majority of liberals accepted them only when they could no longer be avoided. The core of their idea of civil society was the economically self-sufficient, educated individual.

Liberalism offered an educational program for those who had not yet achieved this social status; the individual himself was to be responsible for its future

fulfillment. This program was directed against a society that apportioned different political rights in accordance with membership in specific groups; but in the century of liberalism, the struggle against the historical and traditional society of privileges did not yet aim at the 'mass democracy' of the twentieth century. When the latter appeared on the horizon in the second half of the nineteenth century, liberalism had to change fundamentally if it was to survive party-political competition for votes. European liberalism was by and large able to cope with this adaptation until the beginning of World War I.

Emancipation to become a citizen with full rights not only developed at different paces for different classes, but also for the two sexes. Only a few nineteenth-century liberals could imagine woman as a citizen; one of these was John Stuart Mill (1859) *On Liberty and Other Essays*, an exceptional liberal figure precisely in the radicality of his emancipatory demand for gender parity. Not until the twentieth century did the majority of average liberals come to accept and recognize woman as having equal political rights. Until then, they used the idea of a naturally given eternal inequality of the sexes to avoid the consequences of their egalitarian model of law. Their constitutional idea of civil society was egalitarian, because it was based on the principle of contractual freedom for all. But they did not want to see this principle applied to the family.

Their refusal to politically equate man and woman turned out to have great consequences for their model of society, because they saw the family as the foundation of the state and civil society. In the long run, the liberals could not suppress the contradiction between their political model of an egalitarian civil society and their attempt to construe the family as fundamentally gender-hierarchical. The egalitarian philosophy of liberalism forged an intellectual weapon against the practice of liberal politics, which tolerated or even affirmed certain forms of political inequality. The golden age of liberalism ended with World War I, although its chances for development seemed more favorable than ever before. After all, the war ended in a victory for the Western democracies (USA, Great Britain, and France), which were regarded as the primary representatives of liberalism, while the monarchies of Central Europe and Russia did not survive the war.

Political liberalism did not sweep the field after the fall of these anti-liberal bulwarks. Instead, the signature of the new epoch was rivalry and finally struggle among the liberal-democratic and the new Bolshevist and fascist ideologies, movements, and states. The National Socialist dictatorship, in whose parlance 'liberalistic' was perverted into a life-endangering, denunciatory pejorative, finally threatened to extinguish forever the central idea of liberalism, the dignity of the individual, first in Germany and then in Europe's conquered states. The disintegration of liberal parties, including in Great Britain, the mother country of liberalism, continued after World War I and even after World War II, but at the same time, the constitutional orders and individual behavioral norms in the Western world have been characterized by fundamental liberal values more than ever before. Their validity is no longer tied to the strength of liberal parties. Liberalism's power to form parties appears to decline when its political ideas prevail and are recognized even by opposing political parties.

The collapse of the Soviet empire seems to confirm this observation. The successor states adopted liberal principles, or what they took to be liberal principles, in their new political and economic orders, but this triumph of liberalism over its communist antagonist has at the same time triggered a new international debate over the central principles of a liberal system of state and society (Ackermann 1992). This debate focuses on the old problem of whether a liberal society can trust solely in the regulative miracle of the market or if it should strive to create a state that, through a 'combination of citizenship rights and chances for welfare' (p. 196), empowers all people 'to take part as citizens in the life of the society' and 'to enjoy the achievements of their time' (Dahrendorf 1987, pp. 196, 219).

2. *Ideas of Social and Economic Order*

Historical liberalism arose in the pre-industrial age. It strove for a civil, rather than bourgeois society, in both its ideas of social and economic order as well as in its model of the state. With the idea of the citizen, European liberalism sketched a vision of the future of a society of citizens without great property-ownership differences that would have contradicted the liberal ideal of a middle-class society with a broad distribution of assets (Schapiro 1958). This socially harmonious guiding idea was directed against the class/feudal order, but did not initially field an industrial view of the world against it. Liberalism comprised a plurality of social ideas that, depending on a country's state of social development, ranged from agrarian models of the future (as in Hungary, Spain, and Italy) to industrial capitalistic models (as in the USA, Great Britain, Belgium, and, in part, France and Germany). The liberals demanded social and economic 'progress,' but not an industrial–capitalistic class society. Only after the mid-nineteenth century did pre-industrial liberal ideas of order pale in favor of industrial capitalistic ones.

This shift, pushed forward by the wave of revolutions of 1848, marked the transition from early liberalism to liberalism in the age of industrialization. Only now, when dynamic industrialization ended the pauperism of the first half of the century and sociopolitical control mechanisms seemed renouncable, did political liberalism fuse with *laissez-faire* economics.

The main current of liberalism now became more bourgeois than ever before, without narrowing to a bourgeois economic liberalism, even in this phase of Manchester Liberalism. Even now, the liberals were not characterized by convinced affirmation of a free, unregulated market. The liberal demand for free trade among states did not rule out a willingness to accept and call for regulation of the market within the state.

Since the last third of the nineteenth century, the trend to eliminate social-welfare aspects of ideas of societal order in favor of *laissez-faire* liberalism was tempered by a renaissance of social-liberal guiding ideas. Like early liberalism, the New Liberalism, an international phenomenon, displayed a marked social responsibility, but it assumed the conditions of an industrial society in which the freedom of the individual would be ensured by a policy of social welfare. Liberals had originally contributed to the political climate that, since the 1880s, gave birth to social security laws, first in Germany and then in rapid succession in other countries. This launched a development that led to the modern type of social welfare state. In the United States, it was not until Roosevelt's New Deal that political theory allowed for the state's substantial involvement, in contrast to the traditional liberalism of interest groups (Lowi 1979).

No less important was community liberalism, which played a key role in the building of infrastructure in the enormously expanding cities, a precondition of urban life. This community welfare policy ranged from improved hygiene through municipal provision of gas and electricity to the new parks, museums, theaters, and mass education facilities. The achievements of community liberalism show how intensely liberals participated in the communal protection of individual life chances when, in the face of social and economic developments, individual provisions were no longer adequate or possible for increasing segments of the population.

But the individual, not the collective, was still the focus of liberal norms and politics. The conviction that the individual must take priority over the collective distinguishes even the decidedly social liberalism of the late nineteenth and twentieth centuries from the other progressive movements, such as socialism. But the dwindling numbers of liberal voters since the early twentieth century, teaches us that the variant offered by the liberal parties as a mediation between collective welfare security and individual freedom to decide no longer had the persuasive power that had characterized the liberal guiding ideas of the nineteenth century.

3. *Liberalism and the Nation*

Beginning in the late eighteenth century, the demands for popular sovereignty and national self-determination became two sides of the same coin. Only states that were recognized as national states could prevail. Liberals took a leading role in making the nation the new supreme principle of legitimacy. Like liberalism, nationalism developed a model opposed to the existing order. Both demands aimed at a society ordered in accordance with egalitarian values, constituted as a state and employing a collective, and thus equally egalitarian concept of sovereignty. This was why nationalism historically arose as, and can still present itself as an ideology of liberation and a movement of emancipation, despite all nationalistic atrocities. Anyone who wanted to ally with nationalism had to accept this basic egalitarian stance. Liberals did so. In the nineteenth century, they became the spokesmen of national movements, formulated their goals, and provided their elites.

This symbiosis of liberalism and nationalism strengthened and shaped both. Liberalism's influence expanded far beyond its narrower primary bourgeois clientele, and nationalism developed a reform program promising more social as well as political democracy. The vision of a national civil society that expands citizenship rights into the social realm and promises to allow all to share in the resources produced by a society were visions of the future common to liberalism and nationalism. These promises of participation for all, regardless of social status, are what made these movements so unrivaled in attractiveness. Liberalism was one of the creators of this vision that, here too, uncoupled itself from liberalism. Socialists and conservatives both adopted it.

The task of mediating between the central liberal principle of the responsible individual and the collective claims of the nation was nowhere more urgent and difficult than in the multinational Habsburg monarchy (Judson 1996). The attempt to de-escalate national conflicts through constitutional guarantees of the right to national cultural individuality failed, but it indicated a path that would be taken with the help of the League of Nations after World War I in Europe's new multinational nation-states. This second attempt failed as well, but fundamental principles based in the guiding liberal idea were formulated: the will of the individual must be central, and no collective has the right to disregard this individual will. Making cultural assimilation the precondition for national minorities to enjoy equal citizenship rights in a society with a different national majority contradicts the central convictions of historical liberalism as it had developed since the late eighteenth century, but liberals have often violated this principle in political practice.

4. *Organization and Milieus*

Even before the emergence of lastingly organized modern parties, liberals tried to penetrate the public

sphere with their ideas and to influence state politics. Liberalism had the most favorable conditions for influence where parliamentary institutions already existed.

Secret organizations with revolutionary intentions, on the other hand, were not part of the liberal arsenal for action, because liberal politics did not aim for revolution, but rather for institutionalized, constitutionally regulated cooperation in state politics. Liberals sought to avoid revolutions or to transform them as rapidly as possible into parliamentarily legitimated reforms and thus to put them in a legal framework.

After parliaments became more significant and male voting rights expanded, political movements that sought a voice in state politics could no longer renounce party organizations. Beginning in the second half of the nineteenth century, the liberals, too, formed parties, though this was more difficult for them than for their competitors on the right and left. The guiding liberal idea of the responsible individual, who would determine the 'common good' in free discussion, continuously acted as a certain obstacle to hierarchically organized decision-making procedures. Among the liberals, the party apparatus generally had less importance than the parliamentary groups and local dignitaries.

The guiding image of civil society that has been associated with liberalism since its beginnings is also reflected in its social profile: its focus is on citizens, not the bourgeoisie. The social origins of liberal spokesmen displayed a broad spectrum, especially in the early phase. Nowhere was it limited to the bourgeoisie. Even in Great Britain, the pioneer of industrialization, from 1859 to 1874, only about 21 percent of the liberal delegates to the House of Commons were entrepreneurs. In Italy and Spain, liberalism found a bulwark among the nobility, and in Hungary the nobility dominated completely. In France, Germany, and Belgium, the bourgeoisie took a more prominent place in leading liberal circles. Everywhere here, liberal opinion-makers were highly respected men from respected families. They stood at a distance from tradition, but did not break with it. This membership in 'good society' was an important precondition for the power of liberal ideas to prevail even in those states that denied liberals access to state offices. A strictly bourgeois liberalism could not have achieved this.

Since the second half of the nineteenth century, the social 'catchment field' of liberal parties narrowed. They became more bourgeois. Two lines of conflict increasingly limited the liberal milieu:

(a) One of the liberal parties' heaviest losses was that the workers, whose number greatly increased during the process of industrialization, turned away from them. This happened earliest in Germany (beginning in the 1860s), while in Great Britain the first workers' party still worked together with the liberals in elections (Lib–Lab coalition), but the British liberals, too, lost their worker votes when the Labor Party rapidly grew after World War I.

(b) The secularity of the state was always a central liberal credo. In political practice, this goal could be tied to religious tolerance, but also to a struggle against Church competencies and claims. Intense conflicts between the state and the Church broke out in many countries on the European continent in the second half of the nineteenth century. This greatly limited the extent of liberal milieus. Catholics loyal to the Church generally turned away from liberal parties, which they saw as representatives of laicism, but the situation was different in Great Britain. Until World War I, the Irish question tied Catholic voters to the liberals, whose strongest support came from nonconformist congregations.

See also: Christianity: Liberal; Citizenship, Historical Development of; Citizenship: Political; Civic Culture; Civil Society, Concept and History of; Civil Society/Public Sphere, History of the Concept; Egalitarianism: Political; Egalitarianism, Sociology of; Equality and Inequality: Legal Aspects; Equality: Philosophical Aspects; Freedom: Political; Liberalism: Impact on Social Science; Nationalism: General; Nationalism, Historical Aspects of: The West; Women's Suffrage

Bibliography

Ackermann B 1992 *The Future of Liberal Revolution*. Yale University Press, New Haven, CT

Arblaster A 1984 *The Rise and Decline of Western Liberalism*. Blackwell, Oxford, UK

Bramsted E K, Melhuish K-J (eds.) 1978 *Western Liberalism. A History of Documents from Locke to Croce*. Longman, London/New York

Dahrendorf R 1987 *Fragmente eines neuen Liberalismus*. Deutsche Verlags-Anstalt Stuttgart, Germany

Freeden M 1982 *The New Liberalism*. Clarendon Press, Oxford, UK

Gall L, Koch R (eds.) 1981 *Der europäische Liberalismus im 19. Jahrhundert*. Suhrkamp, Frankfurt am Main, Germany, 5 Vols

Hartz L 1955 *The Liberal Tradition in America*. Harcourt, Brace and World, New York

Judson P M 1996 *Exclusive Revolutionaries. Liberal Politics, Social Experience and National Identity in the Austrian Empire, 1848–1914*. University of Michigan Press, Ann Arbor, MI

Langewiesche D 2000 *Liberalism in Germany*. Princeton University Press, Princeton, NJ

Langewiesche D (ed.) 1988 *Liberalism im 19. Jahrhundert. Deutschland im europäischen Vergleich*. Vadenhoeck and Ruprecht, Göttingen, Germany

Langewiesche D 1993 Liberalism and the middle classes in nineteenth-century Europe: A comparative study. In: Kocka J, Mitchell A (eds.) *Bourgeois Society in Nineteenth-Century Europe*. Berg, Oxford, UK, pp. 40–69

Lowi T J 1979 *The End of Liberalism. The Second Republic of United States*, 2nd edn. Norton, New York
Marshall T H 1950 *Citizenship and Social Class*. Cambridge University Press, Cambridge, UK
Ruggiero G de 1959 *The History of European Liberalism*. Beacon Press, Boston
Schapiro J S 1958 *Liberalism. Its Meaning and History*. Van Nostrand, New York
Sheehan J J 1978 *German Liberalism in the Nineteenth Century*. University of Chicago Press, Chicago

D. Langewiesche

Liberalism: Impact on Social Science

A doctrine with roots in eighteenth century thought, liberalism emerged in the nineteenth century as the prevailing ideology of Western capitalist societies and democracies. Though much transformed by the social and political developments of the twentieth century and seriously challenged by the various movements associated with them, especially fascism and socialism in the first half of the century and feminism, ecologism, and multiculturalism in the second, liberalism has retained its dominant place. It combines philosophical, social, and political elements. Philosophically, liberalism consists of a commitment to the ideals of equality, liberty, individuality, and rationality. However, liberals have combined these concepts in numerous and often opposed configurations, which reflect different social assumptions. Many liberals have believed these values to be inscribed within the very fabric of modern societies, reflecting the innovative technologies, social differentiation, and free movement of labor, goods, and capital characteristic of the commercial and industrial age and its aftermath. Others have thought the links between the two are fortuitous and contingent, and that these and other related social changes, such as the growth of bureaucracy, potentially threaten liberalism. Politically united against traditional and hierarchical societies, liberals have differed, therefore, over whether liberalism requires various kinds of political support, such as the state regulation of markets, or merely the absence of any interference by either the state or others to the spontaneous and autonomous social activities of individuals. Meanwhile, critics of liberalism have either blamed it for all contemporary social ills, or regarded it as a transitory phase to be replaced by superior forms of social organization. Liberalism has featured in social science in correspondingly divergent ways: as the 'natural' value system of modern societies, as a historically and culturally specific phase of social development, as either pernicious or expressive of basic human goods.

1. Liberalism Defined

For analytical purposes it is convenient to break liberalism down into its philosophical, social, and political components, though these form a package that different liberals have interpreted in diverse ways.

1.1 Liberal Philosophy

At a philosophical level, liberals seek the maximal equal liberty possible for all individuals to rationally and freely choose and realize their various projects and preferences without interfering with each other. This position draws together the concepts of equality, liberty, individuality, and rationality in a distinctive way. Liberals value equality in the sense of denying that anyone is the natural subordinate of others. This view does not entail regarding people as the same, merely that everyone has equal moral worth. Instead of seeking equality of outcome like certain socialists, liberals simply desire that all should have an equal opportunity to deploy what talents they do possess on the same basis as everyone else. Their attachment to liberty follows from this interpretation of equality and its implicit endorsement of the moral primacy of the individual. Among moral equals, individuals should be allowed to pursue their own good in their own way to the extent that is compatible with a like liberty for others. Finally, as rationalists, liberals believe that individuals should behave in a coherent and consistent manner and that decisions in the public domain should be open to critical scrutiny and amendment and capable of attaining the reasoned assent of those they affect (see *Freedom/Liberty: Impact on the Social Sciences*).

Though liberals share a moral ethos and broadly agree on the conceptual relationships between liberalism's core values, they offer very different interpretations of them reflecting a wide variety of epistemological and ontological perspectives. Thus, liberals include methodological individualists such as Adam Smith or Herbert Spencer and holists such as Emile Durkheim, materialists such as John Locke and idealists such as T. H. Green or Benedetto Croce, utilitarians such as J. S. Mill and natural rights theorists such as Tom Paine, abstract universalists such as Kant, and historicist relativists such as Herder, and so on. Moreover, their main practical differences stem less from these considerations than from contrasting social visions and political judgements. Though individual thinkers naturally relate these last to their particular metaphysical and methodological views, the connections between them are far from clear cut. Similar theoretical perspectives often generate quite different practical stances, while thinkers of quite diverse theoretical positions may often be in substantial practical agreement.

1.2 Liberal Society

Historically, the most significant social influence on the formation of liberalism was the passage from feudalism to capitalism from the sixteenth to the nineteenth century. This fuelled the liberal attack on ascribed status and its commitment to equality of opportunity. The closed feudal order of inherited rights, with its hierarchical system of aristocratic privileges, was to give way to an open and egalitarian capitalist order, in which an individual's social position and success supposedly mirrored his or her ability and effort—a way of life they felt was best realized in a free market economy.

Some contemporary liberals and many of liberalism's critics regard the ethos resulting from such social origins as essentially materialist, instrumental, and egoistic. Traditionally, however, liberals have been extremely wary of such attitudes and have sought to avoid them. The liberal conception of economic behavior emerged from the experience of the entrepreneurs, traders, shopkeepers, and artisans of early capitalism, and invokes the image of an idealized market order in which rewards are closely linked to effort. This linkage was codified in the labor theory of value and, in the British and North American context at least, drew additional support from the religious beliefs of certain Protestant sects that encouraged industriousness as a mark of salvation. The market also acts as a discovery mechanism, with competition and the profit motive weeding out bad practices and poor products whilst stimulating the search for new processes and commodities, and so promoting the progress of society.

Liberals have disagreed, however, over how far a policy of *laissez-faire* can secure these advantages, and doubts have grown as capitalism has taken on a more corporate form. Most have accepted the need for some regulation to protect the market against itself and maintain its freedom—such as anti-trust and monopolies legislation. Others have sought to distinguish between the economic and social disadvantages that individuals are responsible for and those they are not, supporting quite extensive unemployment insurance and state-supported training schemes, for example, to get the unintentionally out of work back into jobs. Finally, liberals have always acknowledged certain public goods will be badly provided or even neglected by market mechanisms, even if some have believed state provision could employ market elements such as competitive tendering—defense and certain infrastructural goods, such as light-houses, were the classic cases. Clean air, education, and a health service serve as more recent and contentious examples.

1.3 Liberal Politics

The political face of liberalism follows on from the philosophical and social elements outlined above but was additionally influenced by the wars of religion and the rise of modern science in the sixteenth and seventeenth centuries, which fuelled calls for religious toleration and the related belief in rationalism and the moral equality of individuals. Along with the shift from social structures based on status to contractual relations between civic equals, these commitments led to the demand that politics reflect liberal principles. The state, like society as a whole, came to be seen as a voluntary association held together by the mutual consent of its members rather than the ties of deference to social superiors—a conception of political order summed up in the theory of the Social Contract.

The basic components of a liberal political system were established in the wake of the British, American, and French revolutions of 1688, 1776, and 1789, respectively. All three gave rise to new constitutional settlements and bills of rights. Two criteria informed these arrangements. First, laws were to be framed and applied impartially and universally. There were to be no special exemptions for particular groups, such as clerics and nobles, as in the past. Second, the laws were to guarantee to the greatest extent possible the equal right of each individual to pursue their own plan of life. For the majority of liberals, the most basic rights in this respect were traditionally the rights to private property and freedom of belief. These rights were essential to their understanding of the virtues of the market mechanism and of toleration, which they regarded as the embodiments of the new ethos.

At the heart of liberal politics is a distinction between the state and civil society (see *State and Society*). The purpose of the state is to facilitate social interaction rather than to substitute for individual initiative through state control of social institutions. The latter is the mark of dictatorships in liberal eyes. However, social liberals nevertheless believe economic intervention and state provision of certain social goods can have an enabling role, though market liberals contend such steps undermine liberalism. Liberals are also often said to favor a related distinction between the public and private sphere, allowing behavior in the latter they would ban or restrict in the former. However, this distinction has never come out cleanly. To the extent liberals do distinguish between the two it is as a corollary of their views of the state's role in regulating society. The marriage contract, for example, offers a quintessential liberal device of a public character with a clear impact on private behavior. Just as liberals of different hues dispute the degree to which workers are free to sell their labor and the extent to which a valid contract implies certain standards of pay and conditions, so they have also debated whether and in what ways the marriage contract constrains the behavior of couples and grants women, in particular, rights to common property and protection against rape and abuse.

We have already disputed the view of liberalism as promoting possessive individualism in the socioecon-

omic sphere. A related objection contends that liberals advocate subjectivism, skepticism, and relativism in the political sphere, with all views treated as little more than personal opinions. However, the liberal insistence on the individual's right to follow his or her own path does not entail regarding all ways of life as equally valuable. Drawing on the experimental method in modern science and the Protestant defense of freedom of conscience, liberals contend that truth and morality emerge through individuals being allowed to call into question accepted doctrines and try out new and occasionally eccentric or potentially offensive pursuits. Ideas have to be tested for their validity while personal virtue only comes from our taking responsibility for our lives. For the state or any group to paternalistically impose its views on the rest of society risks falling into dogmatic error, demoralizes the population, and produces stagnation. Individuals must be able to make mistakes so that they can learn from them and emerge morally and materially improved.

Nevertheless, liberals have seen no need to tolerate the 'intolerable,' defined as views that are incompatible with treating others as moral equals. Some critics have regarded this position as inconsistent with liberalism's claims to openness. Yet certain limits are clearly implicit in liberalism and would otherwise render it incoherent. Where exactly these boundaries lie and their legitimacy and justifiability are more contentious matters, however. Liberalism has been accused of a gendered, imperialist, and Euro-centric bias, and historically this has often been the case, albeit to differing degrees and with certain important if qualified exceptions, such as J. S. Mill's attack on the *Subjection of Women*. How far gender and multiculturalism remain a blind spot is less clear. Many contemporary liberals maintain that multicultural, gay, and feminist demands for recognition and respect, like other appeals against unjust discrimination, implicitly invoke liberal norms of equality and tolerance. They have simply revealed these principles to have a broader application than certain past liberals had appreciated.

The complexities of the liberal position emerge when we consider their ambivalent view of democracy. Liberals have recognized that democracy offers a means for holding governments to account, obliging policies to be justified and improved through critical debate, and allowing politicians to be peacefully removed when they prove incompetent, biased, or corrupt. However, past liberals sought to limit political citizenship to those with the intellectual attainment and economic resources they believed were required for 'independent' rational decision making—criteria that usually had the effect of ruling out workers and women. Even so, liberals worried about the 'tyranny of the majority' as represented by an irrational and dependent mass, whose votes could be bought by unscrupulous populists willing to pander to their prejudices and interests. As a result, liberals have tended to insist on the need for constitutional safeguards to protect liberal values from being overturned by democratic decisions.

2. *Liberalism and Modern Society*

Social scientists have seen liberalism as either an inherent and defining characteristic of modernity or as threatened by it. Amongst the first camp, moreover, there are those who see it as symptomatic of modernity's ills rather than its advantages, identifying liberalism with a possessive individualism that is not only antisocial but anti sociological as well.

2.1 *Liberalism as the Ethos of Modernity*

The view of liberalism as the ethos of modernity was best expressed by British social scientists such as Adam Smith, Herbert Spencer, J. S. Mill, T. H. Green, and L. T. Hobhouse, and to a lesser extent French thinkers such as Benjamin Constant, Francois Guizot, and Emile Durkheim. For these thinkers, liberalism's philosophical commitment to expanding equal individual liberty was underwritten by a theory of social evolution whereby potential conflicts between different liberties would be harmonized. This latter thesis idealized the market relations between small-scale entrepreneurs characteristic of early capitalism. These were taken as a model of a meritocratic society of self-reliant and responsible citizens, who freely contract with each other for mutual advantage. Liberal politics was conceived as the superstructural products of these social and economic changes. Freedom, reason, morality, and progress came in this way to be identified, with an optimistic account of social development supporting an implicitly ethical naturalist philosophical belief in the compatibility of different forms of self-realization.

Adam Smith's doctrine of the 'invisible hand' of the market and Herbert Spencer's characterization of the 'industrial' form of organization as realizing the 'Law of Equal Freedom' offer classic examples of this position. The first a product of the commercial society of the eighteenth century, the second of the early industrial age of the 1840s, by the late nineteenth century the concentration of capital and the rise of organized labor had made liberals rather less sanguine about the inevitability of their vision. To the dismay of Spencer and other adherents of the earlier 'classical' liberalism, liberals began to envisage a more proactive role for the state in sustaining favorable social conditions. T. H. Green and L. T. Hobhouse, for instance, saw policies such as factory hours regulation and unemployment insurance as mechanisms for promot-

ing freedom of contract and freeing individuals from certain structural constraints. Importantly, though, both saw the state as reducing the mutual interferences between individuals and fostering autonomy. In a similar manner, Durkheim argued that the liberal 'cult of the individual' was inherent to industrial societies and that the modern division of labor fostered an organic solidarity between autonomous agents. However, he had to confront various commercial crises, bankruptcies, and conflicts between capital and labor that suggested a less harmonious picture. He attributed these disturbances to contemporary societies being in a 'transitional phase' in which people had yet to fully assimilate liberal social norms. He too saw a role for the state in fostering these attitudes, with various 'intermediary institutions' serving to inform state policies as to social needs and act as a counter weight to state oppression of the individual (see *Social Question: Impact on Social Thought*).

Critics of liberalism sometimes portray this supposed relationship between liberalism and modernity in negative terms. They associate liberalism with a naturally instrumental and individualist model of human behavior, a disposition they believe is encouraged by the liberal conception of the market and democracy. They also claim it produces a view of social science as following the physical sciences by elaborating general laws of social development on the basis of various constant human passions. Some commentators have even argued that it was only radical and conservative opponents of liberalism who appreciated the distinctiveness of the human sciences by noting, respectively, how human behavior was modified by social structures and by culture, and the capacity of individuals to interpret and evaluate circumstances in different ways. At best, liberalism results in a positivistic bourgeois ideology that ignores the impact on human thought and action of material conditions and productive forces, historically evolving communal traditions and conventions, and the possibility for reflective criticism. Thus liberalism is associated with certain behaviorist versions of political science and economics but not with sociology or anthropology. These arguments result from a highly selective and tendentious reading of the liberal tradition whereby, for example, De Tocqueville becomes a 'conservative,' Durkheim either a 'conservative' or a 'socialist,' and Herbert Spencer, read as an advocate of egoistic possessive individualism, becomes the exemplar of liberal social science. The liberal tradition, however, is much more diverse than this caricature, both methodologically and substantively. The vast majority of liberals distinguished respect for 'individuality' from 'egoism' and even positivist behaviorists, such as J. S. Mill, appreciated that autonomy only flourished given certain material conditions, cultural values, and economic structures and played a part in its turn in how these were fashioned in the future. This awareness of the social and historically contingent character of liberalism is no more evident than in those liberals who saw it as being undermined by modernity.

2.2 *Liberalism as Threatened by Modernity*

The threat posed to liberalism by elements of modernity was especially evident in countries such as Italy and Germany, where a modernizing economy and society went along with fragile and incomplete liberal regimes which ultimately succumbed to fascism. The challenge these developments posed for liberal social science is powerfully illustrated by the 'frustrated' liberalism of the Italian Vilfredo Pareto and the 'disenchanted' liberalism of Weber. Though commonly categorized as sociologists, their work ranges over the whole gamut of the social sciences, covering economics, politics, and, in Pareto's case, social psychology. Pareto was in many respects a sophisticated exponent of the instrumentally rational actor model of liberal individualism and was initially a strong advocate of both *laissez-faire* economics and a libertarian and democratic politics. However, he became convinced that what he called 'logico-experimental' thinking was the exception rather than the rule. Instead, most conduct was 'nonlogical' and arose from certain common psychological dispositions of either a conservative or more inventive kind. Politics simply consisted of the manipulation of these psychic states, a process that had been greatly aided by contemporary mass democracy. The electorate was too large and poorly informed and the decisions needed too numerous and diverse for rational debate to be possible. Democracy simply allowed populist leaders to legitimize their rule via emotive appeals to a popular will that was largely of their own making. He even briefly supported fascism as a possible way of breaking free from this system and establishing a liberal market order from above.

Weber, by contrast, offered a more historically and sociologically nuanced account of similar phenomena. He disputed that economic development *per se* fostered liberalism or that liberal values could be justified *a priori*. Liberalism's apparent basis in natural law had been historically contingent, reflecting the peculiar social and political experiences of early modern Europe outlined above. Not only was this constellation of factors unlikely to reappear, but even within established liberal regimes liberalism was threatened by the spread of economic instrumentalism, bureaucratic rationality, and corporate organization resulting from the further development of capitalism. The chances of liberal democracy being established in societies such as Russia, which had not benefited from the earlier favorable historical experiences, were minimal, therefore. Quite different cultural values were likely to shape the modernizing processes of such

societies, as indeed has occurred throughout Asia where capitalism has by and large taken a nonliberal form. Meanwhile in already existing liberal societies, the autonomous action characteristic of liberalism was possible only for the political leaders and entrepreneurs who controlled and gave direction to the various political and economic administrative structures that govern our lives (see *Modernity*).

Most contemporary liberal thinkers regard Weber as too pessimistic and many believe liberal values can be justified on grounds distinct from their historical and social origins. However, the possible disjunction between late modernity and liberalism, on the one hand, and the search for liberal foundations that appeal to reasons that ought to be accessible to all and so compatible with a variety of different cultures and ideals, on the other, remain the prime challenges a viable liberal theory has to meet.

See also: American Revolution, The Bureaucracy and Bureaucratization; Civil Liberties and Human Rights; Civil Society, Concept and History of; Conservatism: Historical Aspects; Conservatism: Theory and Contemporary Political Ideology; Constitutionalism; Dependency Theory; Freedom: Political; French Revolution, The; Individualism versus Collectivism: Philosophical Aspects; Liberalism; Liberalism and War; Liberalism: Historical Aspects; Mill, John Stuart (1806–73); Modernity; Modernity: Anthropological Aspects; Modernity: History of the Concept; Smith, Adam (1723–90); Tolerance; Traditions in Sociology

Bibliography

Arblaster A 1984 *The Rise and Decline of Western Liberalism*. B Blackwell, Oxford, UK

Bellamy R P 1992 *Liberalism and Modern Society*. Polity Press, Cambridge, UK

Bellamy R P 2000 *Rethinking Liberalism*. Continuum, London and New York

De Ruggiero G 1927 *The History of European Liberalism*. Oxford University Press, Oxford, UK

Gray J 1986 *Liberalism*. Open University Press, Milton Keynes, UK

Seidman S 1983 *Liberalism and the Origins of European Social Theory*. B Blackwell, Oxford, UK

R. P. Bellamy

Libertarianism

Predominantly an American political movement of the twentieth century, libertarianism is a political doctrine whose deepest commitment is to liberty. Libertarians believe that the freedom of the individual is the most important value that should guide political reflection; and that the good society is one in which that freedom is respected. Yet libertarians disagree about what a commitment to individual liberty entails politically; and about the basis of this commitment. To understand libertarianism, then, it is necessary to come to terms with the variety within its tradition; and with its relationship with other political ideas.

Libertarians approach liberty by asking what are the legitimate functions of the state. Broadly, they offer two kinds of answer. The libertarian anarchist answer is that no state is legitimate; for minimal state libertarians, however, the state is legitimate to the extent that its activities are confined to a limited range of functions, including national defense, the protection of individual rights, and the enforcement of contracts. Even within these alternatives, however, further divisions are to be found. Libertarians differ over the proper scope of government, but also over the foundations of individual freedom. Any consideration of libertarianism, then, would have to account not only for the fundamental principles or commitments which give it shape, and for the differences between minimalist and anarchist libertarianisms, but also for the varying doctrines advanced by libertarians from Robert Nozick to Murray Rothbard.

1. *Fundamentals of Libertarianism*

1.1 *Origins*

Libertarianism is a species of liberalism. Liberalism is a tradition of political thought which understands the good society as one in which individuals are free to pursue their own ends rather than one which enforces one particular understanding of what constitutes the good life. Central to liberalism is a commitment to individual liberty. For the liberal, liberty is a good which ought to be honored and protected; and doing so requires recognizing that each and every individual is entitled to a domain within which he or she may act without interference from others. Liberalism, however, describes a broad tradition which includes thinkers as diverse as John Locke in the seventeenth century and John Rawls in the twentieth century.

Libertarianism is that side of liberalism which emphasizes the importance of individual freedom from the state, and which is generally wary of the operation of government. It is skeptical about power, but especially about governmental power. Modern libertarians see themselves in this regard as the heirs of a long tradition in European thought which champions the rights of the individual against the state. While they insist that this tradition stretches back into medieval times, when kings could not create law, or

legislate, but only find or declare law, it is to writers of the seventeenth century that they most commonly turn when looking for the first expressions of libertarian thinking. In particular, it is the ideas of the Levellers—John Lilburne, William Walwyn, and Richard Overton—and later of John Locke, which provide the inspiration for much of libertarianism.

Libertarians take two fundamental ideas from these thinkers. The first is that each individual is entitled to live as he or she chooses provided he or she does not try to force others to live as he or she thinks right. The second is that each individual is entitled to acquire and keep property. If government is necessary, it is to protect these rights by securing individuals against those who would invade or aggress against their property and their persons. But any government that itself violates these rights by coercing them or taking their property without their consent is tyrannical and may rightly be resisted.

Libertarianism has its origins in these ideas, and its heroes are thinkers who have, in different ways, asserted the rights and liberty of the individual against the state. However, libertarianism as an intellectual and political movement did not emerge until after World War I, in America. It arose largely as a response by some intellectuals to the growth of government and the rise of socialism. The most important writers who criticized socialism, central planning, and the growth of government were the Austrian economists Ludwig von Mises and F. A. Hayek who, throughout the 1930s and 1940s, debated the advocates of central planning over the feasibility of socialism. Although based in Europe, their work received a more enthusiastic reception in the USA among intellectuals hostile to the politics and economics of the New Deal. These included H. L. Mencken, Albert Jay Nock, Garet Garret, Rose Wilder Lane, Isabel Paterson, and Henry Hazlitt—journalists, writers, and economists who wrote warning of the dangers to constitutionalism posed by the growth of the state, touting the merits of free-market economics, and defending the virtues of American individualism.

It was in the 1950s that the term 'libertarianism' first came into use to describe this political stance, when Leonard Read, who established the Foundation for Economic Education, began calling himself a libertarian. While the term was rejected by writers like Ayn Rand and her followers, it came into use in the 1960s, and in 1971 the Libertarian Party was founded.

1.2 Principles

Libertarians share two sets of convictions. First, all individuals are entitled to live as they choose, free from interference by other persons or by the state. Second, in the absence of deliberate design by government or the state, order will prevail: granted the freedom to contract and exchange, markets will coordinate the production and distribution of goods better than any other available institutional mechanism. Yet while most libertarians see these two convictions as consistent and mutually supportive, they differ over which is the fundamental basis of libertarian thinking. For Mises the commitment to individual freedom rests on straightforwardly utilitarian considerations: capitalism produces greater human welfare than socialism or the mixed economy. Yet for Murray Rothbard, the rights of the individual are defensible independently of any consequentialist argument.

Both of these convictions are intimately associated with libertarian thinking; however, it is the first which is more important philosophically, since it better identifies what is theoretically distinctive about libertarianism. For libertarians are particularly concerned to defend the freedom of individuals to live without being subject to aggression by others—against their persons or the property they may rightly acquire. An early statement of the first aspect of this claim is was offered by the Leveller Richard Overton, in *An Arrow Against All Tyrants*: 'To every individual in nature is given an individual property by nature, not to be invaded or usurped by any: for everyone as he is himself, so he hath a self-propriety, else could he not be himself, and on this no second may presume to deprive any of, without manifest violation and affront to the very principles of nature …' (Overton 1646, p. 68). This is the thesis of self-ownership. And for libertarians, out of this thesis comes the contention that individual freedom includes the freedom to own private property. The most famous formulation of this claim is John Locke's in his *Second Treatise of Government*: 'Though the earth, and all inferior creatures be common to all men, yet every man has a property in his own person. This nobody has any right to but himself. The labor of his body, and the work of his hands, we may say, are properly his. Whatsoever then he removes out of the state that nature hath provided, and left it in, he hath mixed his labor with, and joined to it something that is his own, and thereby makes it his property.' The centrality to libertarianism of self-ownership, and the corollary right to own property, has led many to conclude that, perversely, the fundamental value for libertarians is not 'liberty' but the right to private property (including, but by no means only, one's property in one's person). All individuals must enjoy liberty; but only that liberty that is consistent with a respect for the property rights of others.

This understanding of individual rights is the basis of the libertarian view of the justification, and proper role, of the state. If individuals own themselves and have a right not to be aggressed against, no government is legitimate save that which enjoys the consent of the people. And the purpose of government is nothing more or less than protecting individuals against those who would violate their rights by

invading their persons or their property. Libertarianism, in this regard, is profoundly antipaternalistic. It is no part of the purpose of government to improve people or society, or even to provide for their welfare. It should not busy itself with human needs or deserts. Its only concern is individual rights.

1.3 The Market

Although its philosophical principles are what make libertarianism distinctive, the emphasis its adherents place on free markets and the virtues of market society is no less important. Libertarians defend capitalism and the market order. For many libertarians, capitalism is simply that form of economic organization which grows out of the freedom of individuals to exchange private property. For others, market society should be defended because it has important virtues of its own.

For thinkers like Hayek and Mises, the defense of the market was crucial to the defense of freedom because they saw the challenge to freedom as coming from modern socialism. Socialism, seemingly being put into practice in the former Soviet Union, and gaining intellectual adherents in the West, criticized the market for being, at once, insensitive to the poor, who fared badly under capitalism, and less efficient than a properly planned economy. Markets were wasteful in production and inequitable in distribution. Socialism's attack on the market made the defense of the market necessary.

In the 1930s, Hayek and Mises devoted their energies to showing the impossibility of socialism, understood as a form of economic organization which eschewed markets. In the absence of the signals sent out by market prices, economic coordination could not happen, and central planners could not acquire the knowledge needed to effect such a coordination, for that knowledge was dispersed throughout society, and often available temporarily and locally. Coordination without markets would produce shortages of some goods and an excess of others: supply and demand would never be in step. Hayek further argued that the deeper danger of socialist central planning was not economic but political. In order to pursue the goal of production without markets, planners would have to resort increasingly to coercion, to curb people's natural proclivities to trade. Central planning required the creation of monopolies; and this would require suppressing those whose activities, intentionally or not, competed with monopoly suppliers. In the long run, Hayek argued, this posed a threat to political freedom, for it would undermine the rule of law by granting to some privileges under the law.

This argument resonated particularly with libertarians in the USA in the 1950s. Here it was deployed less against socialism than against the expanding regulatory state. Government intervention in economic life, libertarians claimed, was neither necessary nor benign. More often than not it upheld privileges—notably those enjoyed by producers and professionals seeking protection from competition which might challenge their monopolies. The most notable proponent of this argument in popular forms was Milton Friedman, notably in his book, *Capitalism and Freedom* (Friedman 1962). However, libertarians, unlike conservatives, also challenged the defensibility of government intervention not only in the economy but also in other areas of society. Friedman and others took issue with the draft, the legal prohibition of drug use, licensing, and various other forms of compulsion by the state.

Another important target of the libertarian challenge was the welfare state. The state was not needed to provide what markets clearly could: health care, disability and unemployment insurance, pensions, among other things. Government, libertarians argued, would, like any monopoly supplier, provide such services at greater cost and lower quality than would be found under competitive conditions. In part they objected to the limiting of individual liberty: compulsory retirement insurance, for example, compelled people to invest old age savings with the state, which provided only one of a range of possible retirement contracts—and not a particularly favorable one at that. The reply that such institutions were needed to ensure that people were protected should they fail to make provision for themselves was rejected as paternalistic.

Nonetheless, libertarians have always disagreed over whether or not it is part of government's role to provide a welfare safety-net. Hayek maintained that government needed to supply some forms of welfare because markets and charity would not be able to meet the demand. Friedman argued for a negative income tax to ensure that no-one dropped below some minimum threshold. Insofar as libertarians have accepted the need for a public policy response to such problems as welfare, however, they have sought to remove the discretionary power of governments, which they think more likely to be captured by particular interests than to become the guardian of the general good.

2. Varieties of Libertarianism

While libertarians share certain general commitments or attitudes, they often differ both in the specific content and in the justification of their philosophical positions. The post-War period has seen the development of numerous attempts to state the case for, and the basis of, libertarianism.

2.1 Nozick's Minimal State Libertarianism

The most important mainstream libertarian is Robert Nozick, whose 1974 book, *Anarchy, State and Utopia*

(Nozick 1974), has been widely discussed and criticized. Beginning wit the presupposition that 'Individuals have rights, and there are things no person or group may do to them (without violating their rights)'; and the conviction that 'So strong and far-reaching are these rights that they raise the question of what, if anything, the state and its officials may do,' Nozick set out to examine, from a libertarian point of view, three of the fundamental questions of political philosophy: can the state be justified?; what principles of justice must govern such a state?; and what would a good society look like if it ran along such principles?

Nozick's first concern is the repudiation of anarchism. He is a minimal state libertarian, and the task of the first part of his book is to show how such a state can, and must necessarily, arise out of a prepolitical state without violating individuals' rights. For Nozick, individuals have rights to their persons and their property; and these rights act as 'side-constraints' on the actions of others, forbidding such others to act in any way that violates rights—even if violating rights would improve the general welfare or, indeed, help reduce the number of rights violations. Rights are absolute; they are also fundamental: rights are good in themselves, and are not derived from some other, more basic, good. For this reason, rights cannot be weighed in the balance with other, competing, goods.

In these circumstances, how could a state ever be justified? Nozick tries to show that, without rights being violated, a minimal state would emerge out of a state of nature in a four-stage process. First, individuals would form mutual protection associations in that condition, so a number of protection agencies would appear. Second, one agency or federation of agencies would come to dominate any given area—through its superior economic efficiency, and without any violation of rights. Third, the dominant agency will rightly be entitled to protect its clients against independent agencies because individuals have the right to protect themselves. Finally, however, the dominant agency will have to extend its protection to those independents by way of compensation for the violation of their (and their clients') rights. Emerging out of all this is a minimal state—one which claims a monopoly on the use of force, while offering protection to all.

What remains doubtful in all this is whether such a state is indeed consistent with libertarian principles, since it must deny persons the right to enforce their rights against others, and require them to pay taxes. And Nozick has duly been criticized for this, particularly by libertarian anarchists. No less controversial, however, is Nozick's insistence that only a minimal state is justifiable: it is no part of the state's right or duty to do more than offer protection. Nozick's theory of distributive justice explains why.

According to Nozick, justice is a matter of how people have come to have the goods they have rather than a matter of the pattern of distribution of people's holdings. Unlike the theory of John Rawls, Nozick's theory is historical rather than structural. There are three principles of justice in holdings: a principle of justice in the acquisition of property; a principle of justice in transfer; and a principle of justice in rectification of past injustice. According to this view, if property is justly acquired, and justly transferred, the resulting distribution of property is, necessarily, just—however small or great the resulting differences of wealth or income among people.

What, however, would make an original appropriation of property just? Can the first person to come across unowned property simply acquire it all? In Nozick's theory, original acquisition is constrained by the 'Lockean proviso': that each is allowed to appropriate as much as he or she desires provided 'enough and as good [is] left in common for others.' The importance of this proviso lies not only in what it forbids by way of future appropriation, however, but also in its implications for all existing holdings. How a person may transfer property will also be governed by the Lockean proviso; and an earlier violation of the proviso may cast doubt on the worth of a particular property right.

Nozick's insistence on the importance of the proviso is consistent with his view of justice as historical rather than structural. His theory is as much an attack on patterned or end-state theories as a defense of his own 'entitlement' theory. End-state theories think of justice in terms of some particular distributive outcome, with which principles of distribution must conform. For Nozick, this is incompatible with a proper understanding of justice and of rights. But he has a further objection: the attempt to secure preferred distributive patterns necessarily requires interference with the voluntary activities of individuals. Liberty upsets patterns, so liberty will have to be compromised if patterns are to be preserved.

Nozick's theory of justice has been vigorously criticized, however, by those who reject his libertarian starting point as well as by others who want to resist his conclusions. The thesis of self-ownership, and Nozick's account of its implications, have been subject to searching examination by G. A. Cohen, in particular.

2.2 *Rand's Objectivism*

While Nozick has dominated academic philosophy, however, no libertarian has reached a wider audience than the Russian-born American writer Ayn Rand. In a series of novels, most famously *The Fountainhead* and *Atlas Shrugged* (Rand 1943, 1957), essays, and newsletter articles, she offered a vigorous defense of capitalism, and articulated a comprehensive philosophy—'Objectivism'—which claimed to set individual rights on a firm epistemological and metaphysical basis.

Capitalism, according to Rand, 'is a social system based on the recognition of individual rights, including property rights, in which all property is privately owned.' On this account, capitalism is understood, or at least defined, not as an economic system but as a social arrangement consistent with the principles of ethics. Capitalist society is one which respects individual rights. The question, however, is: what rights do individuals have, and why do they have these (and not any other) rights? Rand defines the rights an individual has by saying that 'for every individual, a right is the moral sanction of a *positive*—of his freedom to act on his own judgment, for his own goals, by his own *voluntary*, uncoerced choice. As to his neighbors, his rights impose no obligations on them except of a *negative* kind: to abstain from violating his rights.' Rights are thus freedoms to act, but not assurances of anything more than forbearance on the part of others. An important corollary of this, however, according to Rand, is that the individual has a right to acquire property, since human beings need property if they are to survive. The basis of individual rights, she argues, is the right to life. 'The right to life is the source of all rights—and the right to property is their only implementation. Without property rights, no other rights are possible. Since man has to sustain his life by his own effort, the man who has no right to the product of his effort has no means to sustain his life.' But none of this means that an individual is entitled to any particular share of property. All it means is that noone may rightly be prevented from seeking to acquire property noncoercively, and that no one should be prevented from using as he or she wishes the property he or she has rightly acquired.

Rand, in her extensive writings, attempted to defend this stance by showing that it followed from a proper understanding of humans' nature and a grasp of the value of life. This philosophy is called 'objectivism.' Beginning with assumptions about the existence of an 'objective reality,' the centralilty of life as a value, and the importance of reason for humans' survival, she claims to have shown that the only ethically defensible social system is capitalism. Even within libertarianism these arguments have proven controversial. Nozick, for example, finds them unconvincing. But there has been no shortage of philosophers, such as Douglas Den Uyl and Douglas Rasmussen, for whom the construction and defense of a Randian position continues to be an important task.

Although Rand, unlike Nozick, receives little attention from mainstream academic philosophers, the extent of her influence, and the interest of her thought, should not be underestimated. Not only do her books continue to reach a large popular audience, making her one of the best-selling authors of the twentieth century, but she has also exerted a significant influence on academic philosophers, economists, and policymakers (the most famous being Alan Greenspan, Chairman of the Federal Reserve Bank, and co-contributor to Rand's *Capitalism: The Unknown Ideal* (Rand 1967)).

2.3 Murray Rothbard's Libertarian Anarchism

For Ayn Rand, a proper understanding of human rights could yield only a conception of government whose role was limited to the protection of individual rights. Even then, taxation would have to be voluntary; for government, as policeman, could not violate individuals' rights by taking by force what individuals were not prepared to give up. For many libertarians, however, individual rights are in no way compatible with tohe existence of the state. The most important libertarian defender of the anarchist position was Murray Rothbard.

Rothbard has been libertarianism's great polymath: an economist in the Austrian tradition, he was also a historian and social philosopher, as well as a pamphleteer of the highest order. His works ranged from a treatise on economics, *Man, Economy and State*, to a history of *America's Great Depression*, to a philosophical examination of *The Ethics of Liberty* (Rothbard 1970, 1972, 1982).

According to Rothbard, the 'crucial axiom' of the libertarian creed is this: 'no man or group of men have the right to aggress against the person or property of anyone else.' Aggression is defined as 'the initiation of the use or threat of physical violence against the person or property of someone else.' Central to libertarianism, then, is 'the absolute right to private property of every man: first, in his own body, and second, in the previously unused natural resources which he first transforms by his labor.' In Rothbard's view, these two axioms, 'the right of self-ownership and the right to "homestead," establishes the complete set of principles of the libertarian system.'

In common with Nozick and Rand, Rothbard holds that individuals have rights to do as they wish for so long as they do not invade the rights of others. Thus, individuals should be free to use or dispose of their property as they see fit, and live their lives—as well or as badly—as they wish, provided they do not violate anyone's rights. However, Rothbard also maintains that the existence of the state is incompatible with libertarianism, for the state, by definition, must infringe these basic rights. Nozick's 'immaculate conception' of the state, he argues, is unsuccessful both because the logic of its derivation is flawed, and because, even if it could be shown that a state could arise without the violation of rights, this would not serve to legitimize that state (nor, *a fortiori*, any existing state). Nozick, however, has at least offered a philosophical argument attempting to rebut anarchism. Rand, on the other hand, according to Rothbard, does little more than evade the difficulty posed by her own libertarian principles for her claim that that there

is role for government, and for her allegiance to the American state. Rand's patriotic defense of America during the Vietnam War was, for Rothbard, untenable given the fact of military conscription—an unambiguous violation of the libertarian principle of nonaggression and its prohibition of involuntary servitude.

For Rothbard, minimal state libertarianism must either abandon libertarianism's core commitment to nonaggression or lapse into incoherence. None of the extant nonanarchist forms of libertarianism have come to terms with this problem, since none have been able to explain how the most fundamental need of the state—to raise revenue by taxation—could be accounted defensible in libertarian terms. More than any other libertarian, Rothbard has tried to push the principles of libertarianism to their logical conclusion, to show that if individuals have rights, no form of authority which might invade their persons or take their property without their consent is morally legitimate.

2.4 *Hillel Steiner's Left-wing Libertarianism*

In rejecting as flawed the libertarianisms of Nozick and Rand, Rothbard assumes that individuals not only have property in their persons but also can acquire just title to property in the external world. Indeed, for Rothbard, the right to acquire property is not constrained by anything like a Lockean proviso that enough and as good be left over for others. Should all land, for example, be appropriated by those who come to it first, leaving nothing for the johnny-come-latelies, that is simply the way of the world. Thus he rejects the idea, associated in the nineteenth century with Henry George, that the earth itself cannot be owned, even if individuals can own the product of their labor. If the original land is nature- or God-given, he argues, then so are all the people's talents, health, and beauty; if land must be owned only by society, so must everything else. Rothbard therefore rejects any idea that we start by assuming that everyone has an equal claim to land, or to the earth or its resources.

Yet this is itself a highly controversial issue. And some libertarians, notably Hillel Steiner in his *An Essay on Rights* (Steiner 1994), have argued that the correct starting point is an assumption of an equal right of all to the earth and its resources. Persons have, on this account, two original rights: to self-ownership and to an equal share of initially unowned things. Each 'person' is entitled to ownership of himself, for to be owned by others is to be a slave; and so only 'things' may be owned by others. But no one is entitled to appropriate every unowned thing. Indeed, no one is entitled to appropriate more than the Lockean proviso allows. In a world of fully appropriated things, 'each person's original right to an equal portion of initially unowned things amounts to a right to an equal share of their total *value*.' Moreover, and correlatively, any person's possession of a just title to any thing encumbers him or her with a duty to pay every person an equal share of its value. Since there is, Steiner argues, no moral right of inheritance—for only existing, and not past or future persons have rights—the upshot of all this is that a just distribution of things must reflect our original rights, and not the accidents of history or geography. Our just entitlements do not depend upon where, or to whom, we were born. The implication of this libertarian view is that the recognition of libertarian rights is fully consistent with, and indeed requires, a significant measure of property redistribution among persons, and that means all persons, regardless of nationality.

From such libertarian premises, some, like Philippe Van Parjis, who would not describe themselves as libertarians ideologically, go on to argue that justice requires recognizing that everyone is entitled to a form of unconditional basic income. As he points out in *Real Freedom for All*, the core problem for libertarians is to find a non-arbitrary criterion of original appropriation. Homesteading, or an application of the principle of first come, first served, is not, on this view, morally defensible, for it would be a serious error to suggest that anyone could be free living under a regime in which he or she has no property rights, except in his or her person, since all property has already been appropriated and bequeathed. This left-wing libertarianism takes much more seriously the importance of equality as a moral consideratin, just as the other variants of libertarianism reject its significance altogether as a substantive value.

3. *Libertarianism as Ideology*

Politically, libertarianism is often identified as a 'right-wing' ideology but, in fact, its position on the ideological spectrum is problematic. While some libertarians, like Rand, describe themselves as opponents of 'the left' or as right-wing, others, like Rothbard, see themselves as having much more in common with the left as opponents of the state—which they tend to see as the captive or instrument of big business. Moreover, most libertarians try to distinguish (if not dissociate) themselves from conservatives, since they reject the conservative's interest in the regulation of personal morality. To the extent that they align themselves with any other ideological viewpoint, libertarians are inclined to describe themselves as the heirs of classical liberalism. Libertarians as diverse as Hayek, Rothbard, and Steiner have been ready to describe themselves as classical liberals.

In public affairs, however, libertarianism has found itself allied, more often than not, with the conservative side of politics. Over the short history of modern libertarianism, some conservatives have tried to transform it by incorporating its commitment to liberty, but rejecting its hostility to authority and tradition. A notable advocate of such a 'fusionist' view was the American conservative Frank Meyer. To date, however, libertarianism has resisted such attempts. It remains a more radical political philosophy than conservatism can countenance.

See also: Civil Liberties and Human Rights; Freedom/Liberty: Impact on the Social Sciences; Freedom: Political; Liberalism; Republicanism: Philosophical Aspects

Bibliography

Barry N P 1986 *On Classical Liberalism and Libertarianism*. Macmillan, London

Boaz D (ed.) 1997a *The Libertarian Reader: Classic and Contemporary Writings from Lao-Tzu to Milton Friedman*. Free Press, New York

Boaz D 1997b *Libertarianism: A Primer*. Free Press, New York

Cohen G A 1995 *Self-ownership, Freedom and Equality*. Cambridge University Press, Cambridge, UK

Friedman M 1962 *Capitalism and Freedom*. University of Chicago Press, Chicago

Hayek F A 1960 *The Constitution of Liberty*. Routledge and Kegan Paul, London

Hospers J 1971 *Libertarianism*. Nash, Los Angeles

Narveson J 1988 *The Libertarian Idea*. Temple University Press, Philadelphia

Nozick R 1974 *Anarchy, State and Utopia*. Blackwell, Oxford, UK

Overton R 1646 An arrow against all tyrants. Aylmer G E (ed.) 1975 *The Levellers in the English Revolution*. Thames and Hudson, London

Rand A 1943 *The Fountainhead*. Bobbs-Merrill, New York

Rand A 1957 *Atlas Shrugged*. Random House, New York

Rand A 1967 *Capitalism: The Unknown Ideal*. Signet, New York

Rothbard M N 1970 *Man, Economy, and State: A Treatise of Economic Principles*. Nash, Los Angeles

Rothbard M N 1972 *America's Great Depression*. Sheed and Ward, Kansas City, KS

Rothbard M N 1978 *For a New Liberty: The Libertarian Manifesto*, 2nd edn. Macmillan, New York

Rothbard M N 1982 *The Ethics of Liberty*. Humanities Press, Atlantic Highlands, NJ

Steiner H 1994 *An Essay on Rights*. Blackwell, Oxford

Wolff J, Nozick R 1991 *Property, Justice and the Minimal State*. Polity Press, Cambridge, UK

C. Kukathas

Libraries

Although libraries as collections of 'documents' have existed from ancient times, the modern institutional library dates from about the mid-nineteenth century when, in the United States and Europe—the 'West'—a developing print culture accompanied industrialization and the growth of the nation state. An explosion in publishing, not only of books, newspapers, and magazines, but also of administrative and commercial genres—catalogues, manuals, regulations, technical and business reports—prompted the development of libraries to select and organize these print resources. What major Western libraries chose to collect, and how they structured these collections, subsequently became a blueprint for institutional libraries in large and small communities worldwide. In the late twentieth century, an explosion in digital materials had a radical affect on library collecting and organizing practices, while reinforcing their domination by the West.

1. *The 'Modern' Library—a Typology*

Some European countries founded 'national' libraries as early as the seventeenth century, others in the eighteenth and nineteenth centuries, as preservers of their cultural heritage. During the nineteenth century, in North America and parts of Europe, an expanding public sphere included community-supported 'public' libraries. At the same time, colleges and universities multiplied, and 'academic' library collections garnered prestige. Primary and secondary schools also began to support the curriculum through library collections. Public and private organizations established 'special' libraries, as research institutes, religious foundations, and commercial corporations also sought to organize rapidly growing document collections.

1.1 *National Libraries*

A growing national consciousness encouraged the organization of state-funded libraries that acted as both depositories and sources of national bibliography. France's *Bibliothèque Nationale* dates from the fourteenth century, but only in the late nineteenth century did a single published catalogue give unified access to its many collections. In 1841, Antonio Panizzi proposed a set of cataloguing rules for the British Museum Library (now the British Library). In the United States, the Library of Congress (founded in 1800), though not strictly a national library, assumed that role. Librarian Ainsworth Spofford (1864–1897) combined the functions of copyright registration and comprehensive collection of the United States's intellectual product, while under Herbert Putnam

(1899–1939) the Library developed shared cataloging and classification procedures and services that extended its influence—even hegemony—to libraries throughout North America, and from there to other countries worldwide.

1.2 Public Libraries

No single model describes public libraries worldwide. However, many countries have drawn upon British and American experience in establishing public library systems. In nineteenth-century North America and Britain, industrialization spurred the spread of public education. From around 1850, some private libraries began to acquire characteristics typical of public libraries: local tax support, board governance, and accessibility to all members of the community. However, universal literacy was seen by some as a threat, by others as a benefit fostering self-improvement. In the United States and Canada, for example, a faith in 'wholesome' literature for the 'masses' led Progressive Era women's clubs to boost public libraries, as did philanthropist Andrew Carnegie's building program. Public libraries commonly provided free reading materials—and the space in which to read—to children, women, and men of all races and classes, but their collections lacked the diversity of their patrons.

Elsewhere, free public libraries were largely a twentieth-century phenomenon. In China from 1905 the work of American missionary Mary Elizabeth Wood inspired their growth. After 1917, the Russian Revolution spurred the development of public libraries in the USSR so that by 1941 one public library existed for every 2,000 people. In general, late twentieth-century public library services include supporting independent learning and research, maintaining collections of popular adult and children's materials, answering reference questions, providing Internet access, and maintaining space for community activities. However, what—and what not—to collect remains a central issue.

1.3 Academic and School Libraries

As a research model of higher education replaced the classical college, and academic libraries became an essential curricular component, university collections grew dramatically. From mid-nineteenth century on, Harvard's collection expanded from 154,000 volumes to over 12 million, Oxford's Bodleian collection to over 5 million, and the University of Göttingen's to over 3 million. In 1924, S. R. Ranganathan's appointment as librarian at the University of Madras initiated an expansion of Indian academic libraries. By the mid-twentieth century, Moscow State University Library headed an extensive network of academic library collections, and by the mid-1960s, Beijing University's collection held nearly 4 million volumes. Some Western elementary and high schools also developed libraries in the late nineteenth century, many more following suit between the two world wars. After World War II, other countries, including newly-independent Asian, African, and South American states followed suit. Typically, school librarians combine teaching with librarianship to provide materials in various formats and teach students library skills and information literacy.

1.4 Special Libraries

'Special libraries' refers to mission-oriented libraries that range from publicly-funded scientific and technical research libraries to private-sector corporate information centers. Some originated in preindustrial trade, legal, and religious organizations; others—especially those focusing on science, technology, and business—began during the nineteenth century. After World War II, the Cold War expansion of state-funded science and technology fueled a documentation movement that in turn encouraged the growth of corporate and research libraries. The recent rapid increase in digital information resources has prompted another surge in resources devoted to special, especially corporate, libraries.

2. Librarianship

As libraries grew in collection size, librarians' activities diversified. No longer simply guardians of materials repositories, American and British public librarians sought to create methods of locating books in open stacks, and services that targeted new groups of 'users': children, immigrants, rural residents, and the urban poor. Collection management, reference, and cataloging became core components of a new profession: librarianship.

In the late twentieth century, digitization, computer miniaturization, and networking have simultaneously offered solutions and created problems for libraries. Librarians use electronic technology to enhance traditional functions and as new electronic services have proliferated, their reach has expanded far beyond the local institution or community. However, during fiscal stringency, resources assigned to electronic projects detract from those available for print collection, walk-in reference, and youth services. The digital, virtual, or electronic library, sometimes called the library 'without walls' inevitably threatens the traditional library.

2.1 Skills and Techniques

In an expanding print universe, collection management involved selecting materials to build collections and controlling quality by 'weeding' unwanted materials. However, in a mixed print and digital era and with

increasing commodification of information, collection managers face new issues. For example, as commercial publishers entered the scholarly journal publishing market, many serials' prices soared—medical journals by over 90 percent in the 1990s. Another 1990s development—the American 'commercial' university—provides no library services, forcing students to purchase materials or borrow from publicly supported libraries. Materials sharing is, thus, an unequal partnership in which some institutions systematically absorb resources from others. The digital environment is also transforming reference (the personal assistance provided by librarians). Traditional walk-in reference, long augmented by telephone service, is now supplemented, and sometimes replaced, by email and web-based reference, thus reducing services to those without computers.

In the late nineteenth century, librarians developed tools for control of the new but vast array of printed publications. Subject classification and cataloging systems that owed much to contemporary developments in lexicography and the natural sciences helped solve administrative and epistemological problems of book, report, and record description and location. In 1876, the American Melvil Dewey published his *Decimal Classification*, while between 1904 and 1907, Belgians Henri LaFontaine and Paul Otlet produced the *Universal Decimal Classification*. Commercially published periodical indexes also appeared, but it was the mid-twentieth century increase in 'documents' (scientific and technical reports not amenable to traditional classification and cataloging) that prompted a major expansion of professional indexing. The second half of the twentieth century saw attempts to bring scientific, business, and technical 'information' under bibliographic control, including coordinate indexing and the computer analysis of natural language. Combined, issues of intellectual and physical access to documents formed the basis of a new field: information science.

In the 1960s, initiatives to automate information storage and retrieval had far-reaching consequences for libraries worldwide. In 1964, the American corporation Lockheed began making data produced by the National Aeronautics and Space Administration (NASA) commercially available in searchable online databases through DIALOG; similarly, the National Library of Medicine's *Index Medicus* was reborn as MEDLINE in 1971. Online databases multiplied during the 1970s as new vendors created an international market dominated by the United States. Librarians included online searching in their skills repertoire, and struggled to add online database access to an increasingly diverse roster of services.

Librarians were also developing their own electronic databases. American libraries have shared the costs of original cataloging since 1901, when the Library of Congress began to offer ready-cataloged cards for purchase. However, a dramatic increase occurred in the mid-1960s, when the Library of Congress developed the Machine Readable Cataloging (MARC) format for creating, storing, and sharing bibliographic data in what has become an international standardized form. In 1967 a bibliographic utility, the Ohio College Library Center (OCLC), made MARC format available to its members, and in the 1970s, extended its shared cataloging services outside Ohio and to nonacademic libraries. Calling itself the Online Computing Library Center, OCLC later included fee-based public access not only to its catalog of worldwide library holdings but also to journal databases, thus rivaling older commercial database vendors.

As Internet access became widely and internationally available, spreading from universities and other research institutions to businesses and individuals, libraries were quick to make use of it. By setting up networked computers in public areas, they provided direct user access. They made databases of library materials available over the Internet and created gateways to indexes at other Web sites. Libraries also began to digitize documents and graphics materials and make these available over the Web. The Library of Congress began the National Digital Library Program, a cooperative venture with private corporations, to digitize part of its collection, emphasizing special collections of unique items in need of preservation. Conservation and preservation—whether by protection, repair, or copying—have long been core library activities. For much of the twentieth century, conservationists focused on paper deterioration resulting from over-bleaching and the use of acidic alum. One answer was to convert paper documents to microform. But while conversion to digital format can solve libraries' twin problems of space and conservation, it raises new issues, especially the rapid obsolescence of hardware need to 'read' the digital materials.

2.2 Education and Values

For much of the twentieth century, analysis of libraries as cultural and social institutions saw libraries as neutral institutions that could improve people's lives by providing reading materials and educational opportunity. Recently, more critical perspectives have spurred an analysis that highlights the relationship of print and power, and points to ways in which reading is channeled by a social infrastructure shaped, for example, by class and gender relations, as well as ways by which reading itself shaped those relationships.

In North America, librarianship was largely feminized by 1900. The Carnegie-funded Williamson Report (1923) lamented that this caused the profession's low status, and prescribed two-tier staffing: mainly female clerks would perform routine filing and typing, managed by a predominantly male administrative cadre educated in specialized, graduate-level library schools. In 1926, the Carnegie Foundation also helped fund the

University of Chicago's Graduate Library School, stimulating librarians to adopt a research agenda strongly influenced by the empirical, positivist social sciences—'library science.' Library schools awarded doctoral degrees, mostly to men who in turn became library school faculty and large research library administrators, thus reinforcing the gender gap between library managers and clerks.

In the Anglophone West, Victorian values of utilitarian individualism and collective responsibility competed as sources of justification for libraries accessible to the public. Carnegie's systematic support of public libraries in English-speaking countries helped perpetuate a myth of libraries as the foundation of the 'self-made man.' But others, including some North American women's club members, argued for library access to print as essential to an informed citizenry—itself seen as the foundation of republican democracy, resulting in a continuing tension between individualist and collectivist legitimations. Other tensions revolved around collecting 'high' and 'low' cultural items, definitions of 'improving' as opposed to 'pernicious' literature, and its concomitant—'censorship' vs. the 'freedom to read.' Moreover, special libraries constituted a new phenomenon: librarians who defined their mission in private, for-profit terms seen by some public-sector librarians as antithetical to their beliefs, thus fueling a new tension around the opposition of 'public' and 'private.'

In the 1930s, while some librarians supported a value-free library science, others debated what was for many to become a defining professional value: intellectual freedom. Early in the twentieth century, the American Library Association (ALA) regularly blacklisted controversial titles, while World War I librarians banned books by German authors, and books criticizing the war. However, during the 1920s and 1930s, mainstream public opinion turned against intrusive censorship. Exiles from Nazi Germany called for civil liberty; Chinese librarians appealed for American help against Japanese censorship. In 1939, ALA issued the Library Bill of Rights, a policy statement favoring intellectual freedom, that has endured as an expression of a core value in American librarianship. Late twentieth-century political and technological developments have raised new debates about intellectual freedom. In the former Soviet Union, libraries and reading rooms systematically discriminated among readers, making materials defined as anti-Soviet available only in certain restricted locations and under restricted conditions. The end of the Cold War loosened many restrictions, and libraries began to make available the works of previously banned writers. At the same time, electronic transmission has made readily available materials previously controlled through the publishers' and librarians' cultural practices. Availability of sexually explicit material has prompted new calls for censorship in the guise of Internet filtering software. Thus, the late twentieth-century explosion in digital information renews questions among librarians of what to 'collect,' encouraging them to create new means of physical and epistemological access. Traditional professional skills and activities, like providing bibliographic control and user services are in increasing demand, albeit adapted to new technologies, as organizations of all types struggle to meet their changing information needs.

See also: Computer Networking for Education; Information, Economics of; Libraries: Digital, Electronic, and Hybrid; Literacy Education

Bibliography

Black A 1996 *A New History of the English Public Library: Social and Intellectual Contexts, 1850–1914*. Leicester University Press, Leicester, UK

Crawford W, Gorman M 1995 *Future Libraries: Dreams, Madness & Reality*. American Library Association, Chicago

Hahn T B, Buckland M (eds.) 1998 *Historical Studies in Information Science*. Information Today, Medford, NJ

Harris M H, Hannah S A, Harris P C 1998 *Into the Future: The Foundations of Library and Information Services in the Post-industrial Era*, 2nd edn. Ablex, Greenwich, CT

Kent A, Lancour H, Daily J E (eds.) 1968 *Encyclopedia of Library and Information Science*. Marcel Dekker, New York

Lerner F 1998 *The Story of Libraries: From the Invention of Writing to the Computer Age*. Continuum, New York

Lin S C 1998 *Libraries and Librarianship in China*. Greenwood Press, Westport, CT

Rubin R E 1998 *Foundations of Library and Information Science*. Neal-Schuman, New York

Shera J H 1972 *The Foundations of Education for Librarianship*. Becker and Hayes, New York

Thompson J 1980 *University Library History: An International Review*. Saur, New York and Bingley, London

Wiegand W A, Davis D G (eds.) 1994 *Encyclopedia of Library History*. Garland, New York

C. Pawley

Libraries: Digital, Electronic, and Hybrid

With the integration of electronic resources of all kinds, traditional libraries (see *Libraries*) have had to expand their collection policy and holdings to accommodate items not physically present within the walls of the library building especially those available in electronic form. Similarly, various library functions have had to be extended and new services are evolving, corresponding to different levels of integration of external, electronic resources into existing services, library catalogs, databases, and electronic full text offerings. The expansion of data transfer and search and retrieve (S&R) protocols have changed the means by which libraries are fulfilling their information

mediatory role. Furthermore, use of the information obtained through the library has changed with respect to new regulations and usage conditions, dictated not by the library administration, but by the content providers themselves. In addition, new forms of information resources needed by scholars and students can no longer only be made available via the institutional library: External services and virtual information centers or portals are fulfilling various information demands, though still without adequate quality control and validity checks.

The term 'electronic library' evolved in the 1980s to describe the increasingly automated aspects of library functions, as well as all aspects of creating an online public access library catalog. Individual services, including access to scholarly information stored on CD-ROMS and electronic information resources available through external hosts, such as databases and electronic document delivery, became integrated into the library offerings. Libraries provided these and other services in local area networks (LANS) on the same campus and in wide area networks (WANS), such as library union catalog coalitions and regional library networks.

In the 'electronic library,' resources were often based on digitized copies of the library's own holdings, and digitized surrogates of physical resources (such as photographs of museum and art objects) including items located in other institutions or databases, and these were made available specifically through the user's own library. As technology evolved, electronic resources also include digitized images (both still and moving), sound, and data sets. By the mid-1990s, many libraries had discovered the advantages of a hypertext-based platform for offering the library holdings, access to information resources, and other services, to their internal users via library intranet and to the public via the World Wide Web. This was especially important for libraries, as previously far too often the mixture of formats made total integration of all available electronic information resources impossible.

The combination of networked, distributed electronic resources via the Internet, often completely extraneous to the library resources, as a supplement to or even integrating the two as much as possible led to the designation 'virtual library' for the (often subject-oriented) collections of predominately electronic information resources. Thus, the 'virtual library' is not a library as such, but rather a site in the Internet representing a collection of links or electronic metadata available for individual digital and nondigital items which are usually distributed among an indistinguishable number of web sites produced by heterogeneous institutions. 'Virtual libraries' were usually compiled in a manner and structure which resemble traditional library collections and subject bibliographies. In connection with such subject-oriented collections of active hypertext links to distributed web sites throughout the Internet, the term 'clearinghouse' or 'subject gateway' became quite popular in the last half of the 1990s.

The 'digital library' evolved out of the research focus and technological evolution of these two forms. The digital library represents a more complex integration of digital and nondigital objects in distributed, often digitally noncompatible formats but for which appropriate metadata in a standardized electronic form exist. Metadata are terms used to describe the bibliographic, formal, physical, content, and contextual characteristics of digital and nondigital resources. In addition to achieving standards for metadata, so as to provide precise resource discovery within various digital library collections, digital library architecture of necessity also requires interoperability standards to enable integrated processes of information searching and retrieval in multiple, differently structured collections of digitized objects and digital metadata. Interoperability involves creating data exchange structures which allow a web-based search engine or similar mechanism to gather or 'harvest' metadata from distinctly differently structured data formats and collections into an index structure and reprocess it according to the user's needs. Thus, a digital library makes possible the processing of metadata and other information about collections of diverse digital objects which are in different formats, different internal structures, and different data processing codes. Furthermore, the digital library offers services to process the information resulting from the level of resource discovery and retrieval which range from further search and retrieval precision to document ordering and various other aspects of information processing.

The concept of the 'digital library' evolved out of the potential for digitizing of all forms of media: print, image, sound, complex data sets, and computer simulations of nontime-bound sequences, otherwise not demonstrable events, etc., and retrieving them (or the metadata describing them) in context of multiple, networked information systems. Endres and Fellner (2000) have indicated that the 'digital library' is really a metaphor for the electronic knowledge market and therefore also a metaphor for the provision of knowledge for the society of the future. The digital library often serves as a multidimensional (library) catalog or index to bring together the resource discovery information of every resource which was topically, creatively, or by reason of other connections generatively related. Hence, a search in a digital library could produce not only books and journals articles, but also art works, images, videos, sound files, data sets, complex data simulations, etc. This concept requires, however, search machines which can transcend and work with different data storage formats, different physical mechanisms, and operating codes of various types of computers, and still process diverse kinds of data to produce the desired results.

Hence, a digital library does not only consist of digitized and digitally born information resources, but also of electronic metadata on nondigital objects. Its fundamental components are a robust search engine, an interoperability standard which has taken all heterogeneous collections to be integrated into account, as well as metadata for each information unit in compliance with internationally accepted metadata guidelines. Instrumental in defining and forming the field of digital library research and implementation were a series of digital library research projects beginning with the Defense Advanced Research Projects Agency (DARPA), National Aeronautics and Space Administration (NASA), and National Science Foundation (NSF) jointly sponsored Digital Library Program 1994–98 and a second phase from 1998–2002. In the first phase, six major US universities were awarded grants for research and implementation of six aspects of digital library functions: interoperation mechanism among heterogeneous services (technical architecture and info bus at Stanford University), spatially referenced map information and digital imaging (the Alexandria project at the University of California at Santa Barbara), federating repositories of scientific literature (University of Illinois at Urbana and Champaign), developing the informedia digital video library (Carnegie Mellon University), intelligent agents for information location (University of Michigan-Ann Arbor), and environmental planning and geographic information systems (University of California at Berkeley).

In the second phase of the NSF Digital Library Program, the focus of research was expanded to include test bed development of digital objects, research on scalability of retrieval mechanisms and workflow models, study of use and usability of globally distributed, networked information resources, human–computer interaction, next-generation operational systems, evaluation of digital libraries and the use of digital library technology in the learning environment. In addition, the 'digital libraries life cycle from information creation, access and use, to archiving and preservation' was to be addressed (http://www.dli2.nsf.gov). This second phase was also marked by allied research projects selected under the same program but funded by the Library of Congress, the US National Institute for Health, the National Endowment for the Humanities, and the National Library of Medicine, as their focal points fit more closely to these specialty areas. Not only was a greater subject orientation allowed in the second phase, as opposed to the interdisciplinary approach in phase one, but also the social, behavioral, and economic implications of and effects of new digital libraries in the areas of education, commerce, defense, health services, and recreation were also added as parts of the initiative.

Parallel to these developments in the USA, the UK's higher education boards commissioned a comprehensive study of information provision and electronic information resources in the UK (Follett Report) and various recommendations were realized in the ensuing Electronic Library (eLib) program of the Joint Information Systems Council (JISC) which from 1995–99 supported over 60 research projects in aspects of the digital library structure. During the latter part of this same period, the university system in Australia supported the development of a system of subject-oriented virtual libraries and the Australian government decided officially to supply all government publications and information on an electronic basis with metadata which was closely modeled after the Dublin Core Metadata.

In the USA, Sweden, Finland, Germany, Russia, Japan, Hong Kong, and Australia further digital library research and project implementation (including bilateral international research grants by the US NSF, the European Union and corresponding research funding bodies in other countries), further development of various metadata initiatives, electronic publishing initiatives, efforts towards international standardization, building harvesting robots for standardized metadata and supplementary digital library services such as navigation tools, reference linking and context sensitive linking (SFX), open source codes, the digital object identifier (DOI), other naming methodologies for digital objects (including namespace identification), and a variety of innovative electronic scholarly publishing initiatives (see *Communication: Electronic Networks and Publications*) have been started. Furthermore, the movement towards the digital library and comprehensive navigation and access to electronic resources worldwide has also opened a wide area of research and discussion on the role of the library, the publishers' role and the change of economics in scientific publishing. Economic and business models for electronic publications and content management, preservation of digital objects, archiving of electronic resources, and content management have been additional areas of focus for various workshops, seminars, and large-scale digital library projects worldwide.

In addition, distributed networks enabling shared information on digital objects belonging to national cultural heritage are being implemented by libraries and other institutions, including the Library of Congress in its American Memory Project (http://memory.loc.gov/), California Institute of Technology, and many more with the goal of digitizing unique resources and making them publicly available for research and education. In 1995, the Digital Library Federation (http://www.clir.org/diglib/) was founded to coordinate and guide the implementation and maintenance of the digital library and assure digitization of cultural heritage in the USA. In 2001 there are 25 institutional members and four allied members including consortia and the Association of Research Libraries. The Federation is dedicated to exploring

digital library architectures, systems, and services; investigating standards and best practices; development of digital collections, digital preservation; study of use, users, user support, and user services; monitoring the digital library's changing institutional roles and responsibilities; and coordinating funding strategies for digital library projects.

In 1998, Chris Rusbridge, then head of the eLib Programme, described the next phase in digital library research and characterized this phase as the development of the 'hybrid library' (Rusbridge 1998, Knight ?) Arguing that a pure digital library could not exist because of legacy collections and services, Rusbridge suggests that instead of a 'digital library,' 'hybrid libraries' will provide a 'one-stop shop' for digital and nondigital objects and services. Within the concept of the hybrid library, information resources from various databases, online library catalogs, CD-ROMs, publishers' full text servers, e-print archives and servers, as well as other services join in-house services including nondigitized services, and will all be searchable under the same user interface. Ideally, the user interface will support a robust and precise navigation tool which will function using metasearch mechanisms and harvesting protocols which will enable simultaneous searching in completely differently structured data services. The user can define specific data sets and databases for the search and for different subject areas a set of subject-oriented databases and full text journals online will be suggested. The aspect of customized access to subject-specific resources and personalized information searching mechanisms are being researched and discussed, also in terms of content management and subject portals. The 'hybrid library' thus combines all manner of digital and nondigital information.

The developments surrounding digital, electronic, virtual, and hybrid libraries are symptomatic of impending changes in scholarly publishing, the entire academic realm, and in the sociological structures of education, research, and scholarly activities as a whole. The significance of information management in context of university structures, commercial enterprise, and research endeavors is increasing rapidly. Among the technological and structural changes in library services within the context of the digital, electronic, and hybrid libraries and an information provision structure for educational and research institutions, the professions of librarian and information specialist are being challenged profusely by all the other players in the arena. Researchers publish their own papers in the Net on individual homepages and do information seeking and purchasing directly from their desktops often bypassing library services; publishers index electronic full texts with simple but often sufficient metadata and—through contractual agreements with the author—gaining full rights over distribution, reproduction, and dissemination of the author's intellectual property.

Traditional subscriptions to scientific literature are being made subject to access only through licensing agreements, as opposed to ownership, as was possible with printed versions of the same item. Archiving and permanent accessibility, however, are not guaranteed. Users search using publicly available Internet search engines, believe they have obtained all relevant titles to a given topic, and purchase individual articles directly from the publisher. To survive the massive challenges to the profession, the librarian must become an information specialist, continuously remaining on the cutting edge of information technology, on the various forms of information offered by the content providers, on the cost-effectiveness of physical and electronic document delivery options, and lastly continuously update the users, to train and motivate them to use the full range of adequate information resources. Aspects of free access to information through digital and hybrid libraries, staff competencies in information technology and content management, archiving schemes for consortia sharing of digital information, and evaluation of networked information resources are still areas of the digital, electronic, and hybrid libraries which have not been sufficiently clarified with regard to the responsibility for maintaining access, currency, validity, sustainability and future migration of electronic information sites.

See also: Computer-assisted Instruction; Computers and Society; Digital Computer: Impact on the Social Sciences; Discrete State Models of Information Processing; Information Society; Information Technology; Internet: Psychological Perspectives; Libraries; Publishing as Medium

Bibliography

Annotated Bibliography of Digital Library Related Sources http://robotics.stanford.edu/users/ketchpel/annbib.html

Arms W Y 2000 *Digital Libraries*. MIT Press, Cambridge, MA

Borgman C L 2000 *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. MIT Press, Cambridge, MA

Caplan P 2000 Oh what a tangled web we weave: Opportunities and challenges for standards development in the digital library arena. *First Monday* **5** http://www.firstmonday.org/issues/issue5_6/caplan/index.html

Endres A, Fellner D W 2000 *Digitale Bibliotheken: Informatik-Lösungen für globale Wissensmärkte*. dpunkt-Verlag, Heidelberg, Germany

IFLA Electronic Collections *DIGITAL LIBRARIES: Resources and Projects*. http://www.ifla.org/II/diglib.htm

Kaplan N R, Nelson M L 2000 Determining the publication impact of a digital library. *Journal of the American Society for Information Science* **51**(4): 324–39

Knight J The hybrid library: Books and bytes, The Knight's Tale. ***Ariadne*** **11** http://www.ariadne.ac.uk/issue11/knight/

Lesk M 1997 *Practical Digital Libraries: Books, Bytes, and Bucks*. Morgan Kaufmann, San Francisco

Rusbridge 1998 Towards the hybrid library. ***D-Lib Magazine*** July/August http://www. dlib. org/dlib/july98/rusbridge/07rusbridge. html

D. D. Rusch-Feja

Life Course in History

The history of the life course, of age categories, and of aging is essential to understanding how any society organizes itself over time. The notion of life cycle was first used by social science in the 1930s, but has now been superceded by the more flexible notion of life course. Historians had long been interested in the aging of individuals, but were slower to appreciate the social dimension. Today, however, notions of life course and aging are as essential to historical practice as they are to the social and behavioral sciences.

1. *Foundations of the Field*

Representations of the ages of man had previously attracted the interest of historians of art and religion, but it was not until the 1960s that the life course became a focus of social and cultural history. Beginning with Philippe Ariès' *Centuries of Childhood* (1962), there began a veritable avalanche of studies which have explored the history of every dimension of the Western life course from the Greeks to the present. Today there is a substantial body of literature on the history of childhood, youth, old age, and death, focused mainly on Europe and North America, but also touching on the non-Western world.

Ariès was concerned mainly with representations of aging, but his successors were influenced by the social sciences to direct their attention to behavior. The first generation of new work in the 1960s and 1970s was heavily influenced by historical demography, psychology, biology, and quantitative sociology, seeking to define general patterns of historical development and understand the structures and functions of age groups.

Today, the influence of positivist notions of human development are less evident. In the most recent studies of the life course it is the variability and instability of the life course and age categories that have gained the most attention.

2. *Development of the Field*

The historical study of the life course began with childhood, moved subsequently to youth, old age, and death, and, most recently, has focused attention on mid-life. This sequence is less the effect of research design than historical contingency. The youth movements of the 1960s accounted for the interest in the history of adolescence; the evident importance of older age groups as a consequence of the longevity revolution focused interest on them from the 1970s onwards. Death and dying became important topics in the same period, while the recent work on mid-life reflects the uncertainties experienced by adults in recent decades. What began as a field concerned mainly with the middle-class male life course has now become much more sensitive to gender and ethnic differences, yet another reflection of the ways in which the studies in this area reflect contemporary concerns.

The study of childhood has undergone dramatic changes since it was introduced by Ariès. His notion that there was no concept of the child before the early modern period was immediately challenged by medieval and Renaissance historians who produced linguistic, artistic, and medical evidence of concern with children in their respective periods. The idea that children were subject in pre-modern times to an indifference bordering on neglect has also been modified in recent decades, so that we now have a much more nuanced view of earlier eras of child rearing and childhood, but one that preserves the essence of Ariès' original insight into the malleability of notions of childhood over the centuries. We have come to understand that practices such as abandonment, which may seem cruel to us, made sense in other social and cultural settings.

Ariès' insight into the mutability of age categories also influenced the first generation of studies of youth. It became clear that in pre-modern Europe youth was more a function of social context than biological age. Numerical age was less important than work and marital status in defining the boundary between youth and adulthood. It was not until the later nineteenth century that entry and exit into this age category became standardized, and then mainly for middle-class males who constituted the first group to be called 'adolescents.'

It was not until the twentieth century that everyone was seen as being entitled to both a childhood and an adolescence as these were defined by the emerging medical and psychological sciences. As the understanding of childhood and youth became ever more universal and precise, there developed a notion of a normal life course, leading through adulthood to old age. By following the proper rules of aging, individuals could expect to arrive at what Erik Erikson, the major theorist of the linear life course, deemed a coherent understanding of themselves and life more generally. Failure to move through the prescribed stages of life in a timely manner was often portrayed as a cause of personal and social disorder.

By the 1980s, attention had become focused on old age and death. As more and more people began to live longer, healthier, more active lives, many stereotypes

of older people began to be challenged. Old age could no longer be equated with poor health, debility, and passivity. These changes contributed to new interest in old age in earlier periods, and a more nuanced view of aging and 'retirement' in the past. It turned out that the modern period has no monopoly on notions of good aging or, for that matter, of the good death.

Even adulthood, that part of the life course that had previously been neglected by historians because it seemed so universal and unproblematic, has come in for its share of attention. Awareness of midlife crises, first among men and later among women, prompted a number of recent studies focused on the changing nature of adulthood over time. Studies of fatherhood and motherhood demonstrate just how historically variable this dimension of the life course has also been, and how changing definitions of masculinity and femininity, and of sexual orientation, have affected notions of adulthood over time.

3. *Pre-modern and Modern Life Courses*

Historical work from 1960–2000 has complicated our understanding of the life course and aging in the present as well as in the past. It is now generally agreed that prior to the modern era, age boundaries were more fluid and the life course less rigidly segmented. The identity of pre-modern peoples was determined less by how old they were numerically than by the status or place they occupied in the social order. In a society where heading a household conferred adulthood regardless of age, a propertyless man or woman would remain a 'boy' or 'girl' all his or her life. Schooling was not age graded and there was no set age of retirement.

The nature of the agricultural economy, the absence of centralized institutions, and high levels of mortality and fertility all contributed to this condition. So too did a pre-scientific culture, which had no notion of universal quantifiable laws of nature, but was prepared to accept sudden, nonlinear qualitative change as quite normal.

All this changed in the course of the nineteenth century. For the first time numerical age came to define status for a growing part of the population. In a process that Martin Kohli has called 'chronologization,' age became the criterion for schooling, entry into work, and military service. By the end of the century, it was also associated with entitlements such as retirement and welfare benefits.

Chronologization proceeded fastest among urban populations, affecting middle-class men first, but by the early twentieth century it was shaping the lives of a majority of the European and North American populations. Not only was aging standardized, but it was normalized, producing an acute sense of being too old or too young for certain categories of activities. The life course was organized into a linear sequence consisting of childhood/youth, adulthood, and old age, each with its appropriately gendered roles and expectations. By the mid-twentieth century, norms of successful aging held sway throughout Western society and were being exported to the non-Western world as integral to what it meant to be modern.

In retrospect, we can see that chronologization was the product of an urban industrialized society concerned to maximize productivity through the rationalization of time, including age. The newly developed nation-state was equally interested in maximizing the resources of its citizens by organizing them into age cohorts. These efforts were reinforced by the new biological, medical, and psychological sciences, which took over the responsibility for assigning meaning to age and aging from religion. By the beginning of the twentieth century, each age had its own body of expertise. Childhood was now covered by the newly invented field of pediatrics; adolescent psychology was recognized; and a new discipline of gerontology would soon develop. In our own time, death has become yet another area of expertise, thanatology.

The process of chronologization reached its apogee in the post-World War II period, but by the 1980s some were foreseeing the arrival of the 'age-irrelevant society.' Since 1980 there has been a noticeable dechronologization of the Western life course. Not only has numerical age ceased to be such a powerful determinant of status, but the boundaries between ages have blurred and new age categories such as the 'young-old' or what Peter Laslett called the 'Third Age.'

The timing and sequencing of the life course have become less uniform and normative. The ages of entry into and leaving work and schooling show greater variation. The same is true of marriage and parenting. Given new reproductive technologies, the biological clock is no longer as determinative of women's life course. There is no longer concensus on what constitutes a child, a youth, or an old person. Adulthood, previously that part of the life course that could be taken for granted, has also become problematic.

All this has led to a reconsideration of the linearity of the life course. In the late twentieth century, the institutions that once enforced age norms no longer did so. Schools, employers, and the state all now acknowledge, even encourage, flexibility with respect to age. More concerned with consumption than with production, Western industrial society has encouraged both younger and older age groups to enter into roles in the marketplace previously reserved for adults. Children and the youth have become major consumers, while older persons are encouraged to lead active lives in the consumer society.

The nation-state has also relaxed its age criteria. Its high-tech armies are no longer composed exclusively of young men, but now include not only women but a much wider range of age groups. Today boundaries between age groups blur. The legal distinctions be-

tween childhood and adulthood have become less clear. Legal ages of drinking, driving, and voting have all been revised in recent decades.

While most people still maintain some normative notion of the life course, this gets less reinforcement than it once did from the medical and behavioral sciences. Recent studies have challenged the predictive power of numerical age in everything from learning capacity to physical fitness. It is no longer clear that life follows a uniform, predictable sequence. Gender differences and variation by class and ethnicity have complicated the contemporary picture of aging to the point that generalizations become increasingly suspect.

Today, age is seen as more subjective, more contingent. We seem to be returning to a situation similar to that which existed before the process of chronologization began. But now it is not one's status that determine's one's age, but one's own effort. We are told that how we age is up to us, dependent on how we eat, exercise, and medicate ourselves. Notions of successful aging and the good death were present in earlier periods, but were largely defined by religion. Today, religion competes with the prescriptions of science and medicine. As recent studies show, factors of class, gender, generation, and ethnicity still have a powerful effect on the way we think about age. However subjective age categories have become they do not exist outside history or culture.

4. *Current State of the Field and Future Directions of Research*

Current interest in the history of the life course reflects contemporary uncertainty about aging and the desire to know more about its contingent nature. The history of the life course and aging is no longer so influenced by the behavioral sciences as it once was, while, at the same time, psychologists and sociologists have become more aware of the historicity of their subject. Psychologists now question the notion of a linear life course and the assumption of universal laws of development that underlay it. It is now generally acknowledged that gender, class, race, and cultural background have enormous effect on the life course and life chances.

The relationship between history and the behavioral sciences has been profoundly altered. There is now much greater appreciation of the variability not only of conceptions of age, but of age relations themselves. Aging has ceased to be seen as some invariable law of nature, to be discovered and obeyed. It is no longer a problem to be formulaically solved, but a challenge that confronts every generation and to which there is no one answer, only an ongoing dialogue with ourselves and others about the meaning of age.

No period in history has ever achieved agreement on the question of when life begins and ends, or how the life span should be divided or sequenced. Furthermore, our conceptions of life are not limited to the actual life span. We have the capacity to imagine life before birth and after death, and conceptions like the fetus as 'unborn child' or notions of life after death have very powerful effects on the way we lead our lives.

Over time the perceived number of ages of man has varied enormously. Some cultures have assigned the highest value to the oldest ages; others have valorized childhood or youth over age. The so-called 'prime of life' has been equally variable over time and space. Even within a single society in a given period, people have thought about the life course in very different ways. Religion, which has traditionally been the source of ideas about age and aging, has always been a point of division. Catholics traditionally assigned meaning to particular moments in life, such as the deathbed, while Protestantism insisted that every moment was equally significant in the eyes of God.

Today, intra-religious battles over the meaning of the life course pale in comparison with the struggles between religion and science over the definition of the beginning and ending of life.

Different class groups have also been at odds when thinking about the life course. And there are obvious differences in the ways that different ethnic groups think about age and aging. Gender is also a source of variation. The notion of the linear life course has been closely associated with a certain kind of middle-class masculinity, which places great value on the development of a career. Until recently, few women have had access to careers, their lives having been more contingent on family responsibilities. They have tended to have a more flexible notion of the life course and to be more adaptable than men. In the late twentieth century, when so many people lived beyond all previous societal expectations, this kind of flexibility proved to be a great asset.

It is clear from recent research that the life course is as much the product of culture as of biology. Life is not so much a script we follow, as one we as individuals and as societies write as we go along. In pre-modern times that script was a religious one. In the nineteenth century, the natural and behavioral sciences took over its authorship, producing a grand narrative of the life course sufficiently consistent with the requirements that the dominant economic and political institutions endowed it with until the fourth quarter of the twentieth century.

Today, the traditional naturalistic grand narrative no longer suffices. The global economy demands flexibility; the nation-state ceases to police age boundaries and moves to divest itself of the entitlements (public schooling, mandatory retirement, social security) that were once associated with age. Individuals are encouraged to become more flexible and self-reliant, charged with the responsibility not only for defining but narrating their own life courses. This is reflected in the explosion of personal memoirs and family histories in the last few decades. In the absence

of grand narratives of the life course, everyone seems bent on telling his or her own story.

This new development represents both a freedom and a burden to individuals, and a challenge to experts in the area of life course studies. It requires a new sensitivity to what people are telling us about their lives, new methods that go beyond the old quantitive measures of age to assess subjective experience of aging. It also calls for new theories of aging that take account of the interaction of cultural and biological factors. Historians as well as social scientists now acknowledge that we inhabit not only a physical world but also a symbolic world, and that it is the interplay between the two that determines the history as well as the sociology of the life course and aging.

Historians as well as behavioral scientists have begun to acknowledge that the sources of data they have traditionally relied on to measure aging, mainly physical and behavioral data, tells us only part of what we need to know. They are beginning to explore further the stories that people have told themselves about the aging process and how different subcultures have imagined the life course. For it is through these that we grasp the ways in which people not only think about age but think with age, and how the meanings that are constructed through and around the life course affect the way we live. For we not only live with certain historically defined notions of life course and aging, but live by them. These constitute the source of the meaning as well as the structure of any social system.

Historical studies of the life course and aging are beginning to turn to those humanistic disciplines such as literature and art history traditionally devoted to close analysis of language, narrative, and images. This does not mean that they are discarding the alliances with the behavioral and social sciences that have been so productive to this point, but rather that all the disciplines concerned with the life course and aging have become more aware of the need to examine their epistemological and ontological assumptions.

See also: Aging and Health in Old Age; Aging, Theories of; Childhood: Anthropological Aspects; Family and Kinship, History of; Family as Institution; Family Systems in Europe; Household in Anthropology; Kinship in Anthropology; Life Course: Sociological Aspects; Midlife Psychological Development; Youth Culture, Anthropology of; Youth Culture, Sociology of

Bibliography

Ariès P 1960 *L'enfant et la vie familiale sous L'Ancien Regime*. Plon, Paris

Ariès P 1962 *Centuries of Childhood: A Social History of Family Life*. Knopf, New York

Ariès P 1981 *The Hour of Our Death*. Knopf, New York

Ariès P, Duby G 1987–91 *A History of Private Life*. Belknap Press of Harvard University Press, Cambridge, MA

Chudacoff H 1981 *How Old are You? Age Consciousness in American Culture*. Princeton University Press, Princeton, NJ

Cole T R 1992 *The Journey of Life: A Cultural History of Aging in America*. Cambridge University Press, Cambridge, UK

Demos J 1986 *Past, Present, and Personal: The Family and Life Course in American History*. Oxford University Press, New York

Erikson E 1963 *Childhood and Society*, 2nd edn. Norton, New York

Gillis J 1974 *Youth and History: Tradition and Change in European Age Relations, 1770—present*. Academic, New York

Hareven T K (ed.) 1996 *Aging and Generational Relations in the Life Course: A Historical and Cross-cultural Perspective*. de Gruyter, New York

Kohli M 1985 Die Institutionalisierung des Lebenslaufes. *Vierteljahresschrift fuer Soziologie und Sozialpsychologie* **1**: 1–29

Laslett P 1991 *A Fresh Map of Life: The Emergence of the Third Age*. Harvard University Press, Cambridge, MA

Modell J 1989 *Into One's Own: From Youth to Adulthood in the United States*. University of California Press, Berkeley, CA

Pollock L A 1983 *Forgotten Children: Parent–Child Relations from 1500–1900*. Cambridge University Press, Cambridge, UK

J. R. Gillis

Life Course: Sociological Aspects

Human lives represent an enduring interest in the social sciences that reflects important social changes over the twentieth century. Most notably, developments after World War II called for new ways of thinking about lives, society, and their relationship. Longitudinal studies of children before the war became studies of adult development in the postwar era. Demographic change also assigned greater priority to the study of adults and aging. The scientific challenge presented by these priorities focused attention on the social pathways of life. This article presents the life course as both concept and theory for studies of lives, human development and aging.

1. Elementary Concepts

In concept, the life course refers to a sequence of age-graded events and social roles that are embedded in social structure and historical change. The life course also defines a field of investigation with a theoretical orientation that guides research in terms of problem identification and formulation, variable selection, rationales of design, and explanatory analysis. Building on advances since the 1960s, life course theory

(Elder 1998) has uniquely forged a conceptual bridge between developmental/aging processes, the social life course, and ongoing changes in society, one based on the premise that age places people in the social structure and in particular birth cohorts.

The life course evolves over an extended period of time, a trajectory of marriage, work, or earnings; and it also takes form within a short time-span, marked by the transition between statuses, such as leaving home, entering a new job, and divorce followed by remarriage. Social transitions are always embedded in trajectories that give them distinctive form and meaning. Life trajectories are constructed by linking states across successive years. Each work trajectory, for example, is marked by a sequence of jobs and transitions between jobs, along with occasional spells of unemployment. A work transition may entail little change or produce a turning point—a redirection of life through changes in situation, meaning, and/or behavior, whether antisocial or conforming. Entry into military service has served as a turning point for men who grew up in poverty and were involved in an adolescent life of crime (Sampson and Laub 1993).

Each of these concepts is applicable to different levels of the life course: (a) institutionalized pathways may be established on the macrolevel by laws and social policies of the state or firm (Mayer and Müller 1986); (b) the life course of the individual is worked out at the microlevel in terms of choices and lines of action; and (c) the developmental or aging trajectory of the individual as defined, for example, in terms of intellectual development. An example of these multiple levels is provided by Spilerman's (1977) analysis of work careers. 'Career line' refers to pathways defined by the aggregated work histories of individuals. They are patterned by industry structures and the labor market. An employee's worklife, one aspect of the life course, varies by the career requirements of the firm and marketplace. At the developmental level, positive and negative changes in worklife have predictable consequences for emotional health. Each level of the life course represents a potential entry point for research. However, contemporary studies are increasingly extended across levels of analysis.

Three concepts have been used interchangeably with the life course—life cycle, life history, and life span. Life cycle generally describes a sequence of life events from birth to death, though it more accurately refers to stages of parenthood over the life course, from the birth of a first child to the departure of children from the home to the childbearing of the next generation. The life cycle is repeated from one generation to the next, but only within the framework of a population. Some people do not have children and consequently are not part of an intergenerational life cycle.

Life history typically refers to a lifetime chronology of events and activities that variably combine data records on social events and sometimes also psychological processes. These records may be obtained from archival materials or from interviews with a respondent. Some life history interviews are prospective and focus on the present and future, while others are retrospective and enable the investigator to obtain information that was not collected in the past. Retrospective life histories (Giele and Elder 1998) record the age (year and month) at which transitions occur in each activity domain, and thus depict an unfolding life course in ways uniquely suited to event history analyses and the assessment of time-varying causal influences.

Life span specifies the temporal scope of inquiry and specialization, as in life span developmental psychology or life span sociology. A life span study extends across a substantial period of life and generally links behavior in two or more life stages. Instead of limiting research to social and developmental processes within a specific life stage, as at middle age, a life span design favors studies of antecedents and consequences that extend beyond a single life stage. Each of these concepts, from life span to life cycle, has an important place in studies of the life course.

2. *The Emergence of Life Course Theory*

Life course ideas coalesced in a theoretical orientation during the 1960s and 1970s. One of the crystallizing forces came from longitudinal studies of children that were extended into adulthood during the 1940s—examples include the Oakland and Berkeley Growth Studies, and the Guidance Study at the University of California, Berkeley. The life-span extension of these studies gave fresh momentum to the study of adult development at a time when few longitudinal studies were underway. Three challenges confronted investigators in the social and behavioral sciences: (a) to replace child-based, growth-oriented accounts of development with concepts that apply to development and aging across the life course; (b) to think about how human lives are socially organized and evolve over time; and (c) to relate lives to an ever-changing society.

The first challenge led to the formulation of life-span concepts of development, especially within the expanding field of life-span developmental psychology (see *Lifespan Development, Theory of*). Distinctive principles include the relative plasticity and agency of the aging organism, the life-long interaction of person and social context, and the multidirectionality of life-span development. Differentiating and cumulating experiences across the life span tend to generate complementary aging dynamics (Dannefer 1987): (a) greater heterogeneity between individuals over time, and (b) enhanced continuity within individuals. Sociological concepts of life-span development include trajectory, transition, and turning point, as well as the

cumulation of advantages and disadvantages over the life course.

Life-span thinking on human development and aging occurred initially with little attention to a well-established 'role theoretical perspective' on human lives, one that dates back at least to Thomas and Znaniecki's (1918–20) *The Polish Peasant in Europe and America* . Role and relationship theories provided a way of thinking about socialization, generational succession, and social networks into the 1960s. The central concept at the time was life cycle; it depicted life organization in terms of social relationships and generational succession through the kinship system. The missing element was temporality. Life cycle models did not specify the timing of role entries or exits or the duration of time-in-role. Also, generational membership failed to locate people with precision in historical time and thus according to social change.

These limitations were addressed by theory and research on age and time. The path-breaking work of Neugarten (1968) documented substantial life variations in social roles by age. Contrary to established views (Eisenstadt 1956), she found that people of the same age varied significantly in the pace and sequencing of life transitions. Ever since this research, the differential timing and order of events has been among the most active topics of life course study, especially in the transition to adulthood (Heinz 1999).

Another age-based contribution to life course theory came from an appreciation of historical variations in people's experiences and lives (Ryder 1965, Riley et al. 1972). Ryder proposed the term 'cohort' as a concept for studying the life course in relation to social change. Cohort refers to the age at which people enter the system; thus, a birth cohort locates people in history according to their birth year. Ryder's life stage principle states that the impact of historical change on the life course reflects the life stage at which the change was experienced (see *Age, Sociology of*).

The three streams of life course theory (social relations/life cycle, age, and life-span concepts of development) came together in a study of children who were born in the early 1920s, grew up in the Great Depression, and then entered service roles in the Second World War (Elder 1974 (orig.), 1998 (new edn.)). The study began with ideas from studies of social relations, such as generation, socialization, and social roles, but soon turned to the analytic meanings of age for ways of linking family and individual experience to historical change, and for identifying age-graded trajectories across the life course. The project tested Ryder's life-stage hypothesis by comparing the effects of drastic income loss in the Great Depression on the life experience of the Oakland cohort members (born 1920–1) with that of a younger Berkeley cohort, born at the upper end of the 1920s. Consistent with the life-stage hypothesis, the younger boys in particular were more adversely affected by family hardship, when compared to the older boys, though such differences faded in the adult years through the impact of military service, family support, and higher education.

By the 1990s, the life course had become a general theoretical framework for the study of lives, human development, and aging (Binstock and George 1996). This development was coupled with the growth of longitudinal studies and the emergence of new methodologies for the collection and analysis of life history data (Giele and Elder 1998, Bryk and Raudenbusch 1992).

3. *Principles of Life Course Theory*

Life course theory is defined by a set of core principles. Most basic is the premise that behavior cannot be fully understood by focusing solely on the specific life stage in question, defined by age or social role. The origins of adolescent and young adult behavior lead back to childhood and the consistency of nurturance and discipline. Likewise, behavioral adaptations at mid-life are not influenced solely by current circumstances, but by developmental trajectories that extend back to childhood and by anticipations of the future. Accordingly, the principle of life-span development states that: human development and aging are lifelong processes.

The early pioneering studies in California became life-long enterprises without an initial life-course vision. More ambitious projects have recently adopted this perspective, such as the national longitudinal studies of birth cohorts in Great Britain (birth dates of 1946, 1958, and 1970). They are scheduled to be followed into the later years of life. Studies across the life course are changing our understanding of poverty (Leisering and Liebfried 1999), work, and achievement, while identifying pathways of resilience and increasing vulnerability.

Preventive measures for health in the later years require knowledge that can only come from programmatic life-course studies of early influences and behavior. However, particular research questions and matters of cost can make compelling arguments for panel studies that are launched with older people, such as the Berlin Aging Study (Baltes and Mayer 1999) of men and women between the ages of 75 and 100. The study includes a retrospective life calendar and draws upon a variety of disciplines in assessing the differences between old age and old old age, and the potential benefit of intervention.

People and their situations change across the life course, raising the probability that influences also differ by stage of life. Empirical evidence documents a SES gradient in emotional depression, but there is reason to expect the mechanisms to vary as people become older. In fact, Miech and Shanahan (2000)

find that education and mastery beliefs in the early adult years offer protection against depressed feelings. By later life, these feelings are mainly a product of low income and poor physical health. With these contrasts in mind, the principle of timing states that: The developmental antecedents and consequences of life transitions, events, and behavior patterns vary according to their timing in a person's life.

Temporal variation across the life course also takes the form of change in the meaning and implications of particular life transitions, such as the loss of a spouse immediately after marriage vs. the more expectable event of death in old age. In theory, the normative concept of social timing specifies an appropriate age for events, such as sexual experience, childbearing, and retirement. Age at retirement has declined in Western societies along with normative expectations concerning the duration of work (Kohli et al. 1991). The timing of lives refers to change and stability in event timetables, to the scheduling of transitions, and to the timing of developmental events, such as puberty.

Social timetables may not correspond with biological change, as illustrated by early maturing girls. They are older in physique than their age and frequently develop associations with older boys that accelerate their onset of sexual experience (Stättin and Magnusson 1990). Neugarten (1968) refers to a social clock by which socially approved age norms are superimposed on the biological organism, as in the relation between sexual norms and the timing of menarche. In these and other ways, the principle of timing is fundamental to the role of 'interdependent lives' in the life course and the principle of linked lives: Lives are lived interdependently and that social-historical influences are expressed through this network of relationships.

The principle of linked lives moves beyond the conventional model of the life cycle by referring to all forms of relationships, not simply those having to do with family and reproduction of the generations. Human lives are typically embedded in and structured by social relationships with kin, friends, and co-workers across the life span. Kahn and Antonucci (1980) refer to these lifetime associates as convoys of social integration, social support, and obligation. Social relationships also represent a vehicle for transmitting and amplifying the effects of stressful change, as in families under the pressures of change. The prevalence of dual career families has centered research on issues of linked lives and timing, including the challenge of synchronizing lives at points of transition, e.g., job offers, retirement.

Issues of timing and linked lives are played out in specific historical contexts, whether adolescence in the prosperous 1920s, depressed 1930s, or war mobilized early 1940s. Especially in societies undergoing rapid change, a different birth year exposes individuals to different historical worlds, with their constraints and options. Historical effects take the form of a cohort effect when social change differentiates the life course of successive cohorts, such as older and younger men before the Second World War. The aforementioned Great Depression study of two birth cohorts (Elder 1974, 1998) illustrates this cohort effect. The principle of lives in time and place asserts that: The life course of individuals is embedded in and shaped by the historical times and places they experience over their lifetime.

History also takes the form of a period effect when the influence of a social change is relatively uniform across successive birth cohorts. Secular trends in the scheduling of marriages and first births across the twentieth century are largely an expression of powerful period effects. A third type of effect occurs through maturation or aging. The heterogeneity of historical experience within birth cohorts calls for inquiry of the change process itself, such as the actual decline of family income. Other examples of historical change in lives include research which highlights a pattern of resilience among urban youth who were 'sent down to the countryside' during the Chinese Cultural Revolution (1966–76) (Zhou and Hou 1999).

Though structured by historical circumstance, the life course is also shaped by the choices people make—a young girl's pregnancy, the decision on whether to carry the child to term, and the issue of who cares for the child and its consequences for the daughter's education and future. During China's Cultural Revolution, the choices of young people were often made by their work unit, but some youth were sent to the countryside and managed to continue their education, postpone marriage, and return to their city of origin. The principle of human agency asserts that: Individuals construct their own life course through the choices and actions they have taken within the constraints and opportunities of history and social circumstances.

Human agency has been a defining theme of biographical studies that extend back to the early Chicago School of Sociology. Within this intellectual tradition, Clausen's *American Lives* (1993) found evidence supporting the hypothesis that competent young people who think about the future with a sense of personal efficacy are more effective in making sound choices and in implementing them. This evolving life course generally reflects a 'loose coupling' between social stages and transitions. Age grades and loose coupling (Elder 1998) can be viewed as opposite sides of the adult life course—its social regulation on the one hand, and the actor's initiative on the other, with its element of disorder.

4. *Challenges and New Directions*

Since the 1970s, the life course has become a lively interdisciplinary field of inquiry, focused around studies of human lives and their social pathways, with

emphasis on the imprint of historical change on trajectories of human development and aging. This time period has witnessed an unparalleled growth of longitudinal samples for life course research along with statistical and methodological innovations for the study of change and multilevel contexts. Life course studies continue to draw upon widely diverse specialties, from developmental science to the sociology of the state. Advances in the field have led to new challenges and directions, as noted below.

First, the task of linking early and later experiences in the life course remains a challenge. Studies have traced the persistent effect of early events and experiences to subsequent outcomes, such as the effect of a teenage birth and military service. But the chain of linking mechanisms is largely uncharted.

Second, studies of the individual life course are beginning to cross levels, but more work is needed to assess the ongoing influence of social structure and change, the impact of cohort life patterns on social structure, and the effect of changing lives on human development and aging.

Third, interlocking trajectories, such as work and family, or work and psychological health, represent a basic distinction of life course theory, but only recently have statistical techniques enabled research to assess this dynamic.

Fourth, biological as well as social events have been linked to life patterns, as in studies of physical maturation. This work and other factors (e.g., sibling samples) are prompting more thought and research on genetic influences.

Fifth, cumulating research on social transitions across the life course suggests that general models of life transitions and their effects are possible at this time.

All of these research challenges and new directions will benefit from a greater investment in comparative study, both within and between societies. More than in the past, the initiation of life course projects is informed by cross-national collaboration and comparative options (see *Age Stratification*).

See also: Age Structure; Aging, Theories of; Biography and Society; Erikson, Erik Homburger (1902–94); Life Course in History; Lifespan Theories of Cognitive Development; Midlife Psychological Development; Youth Culture, Anthropology of; Youth Culture, Sociology of

Bibliography

Baltes P B, Mayer K U (eds.) 1999 *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, Cambridge, UK

Binstock R H,, George L K (eds.) 1996 *Handbook of Aging and the Social Sciences*, 4th edn. Academic Press, San Diego, CA

Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA

Clausen J A 1993 *American Lives: Looking Back at the Children of the Great Depression*. Free Press, New York

Dannefer D 1987 Aging as intracohort differentiation: Accentuation, the Matthew effect, and the life course. *Sociological Forum* **2**: 211–36

Eisenstadt S N 1956 *From Generation to Generation: Age Groups and Social Structure*. Free Press, Glencoe, IL

Elder, Jr. G H 1974 *Children of the Great Depression: Social Change in Life Experience*. University of Chicago Press, Chicago

Elder, Jr. G H, 1998 The life course and human development. In: Damon W (ed.) *Handbook of Child Psychology, Vol. 1: Theoretical Models of Human Development*, 5th edn. Wiley, New York

Giele J Z, Elder, Jr G H (eds.) 1998 *Methods of Life Course Research: Qualitative and Quantitative Approaches*. Sage, Thousand Oaks, CA

Heinz W R (ed.) 1999 *From Education to Work: Cross National Perspectives*. Cambridge University Press, Cambridge, UK

Kahn R L, Antonucci T C 1980 Convoys over the life course: Attachment, roles, and social support. In: Baltes P B, Brim, Jr O G (eds.) *Life-span Development and Behavior: Vol. 3*. Academic Press, New York

Kohli M, Rein M, Guillemard A-M, van Gunsteren H (eds.) 1991 *Time for Retirement: Comparative Studies of Early Exit from the Labor Force*. Cambridge University Press, Cambridge, UK

Leisering L, Leibfried S 1999 *Time and Poverty in Western Welfare States*. Cambridge University Press, Cambridge, UK

Mayer K U, Müller W 1986 The state and the structure of the life course. In: Sørensen A B, Weinert F E, Sherrod L R (eds.) *Human Development and the Life Course*. Erlbaum, Hillsdale, NJ

Miech R A, Shanahan M J 2000 Socioeconomic status and depression over the life course. *Journal of Health and Behavior* **41**: 162–76

Neugarten B L 1968 *Middle Age and Aging: A Reader in Social Psychology*. University of Chicago Press, Chicago

Riley M W, Johnson M E, Foner A (eds.) 1972 *Aging and Society: A Sociology of Age Stratification Vol. 3*. Russell Sage Foundation, New York

Ryder N B 1965 The cohort as a concept in the study of social change. *American Sociological Review* **30**: 843–61

Sampson R J, Laub J H 1993 *Crime in the Making: Pathways and Turning Points Through Life*. Harvard University Press, Cambridge, MA

Spilerman S 1977 Careers, labor market structure, and socioeconomic achievement. *Americam Journal of Sociology* **83**: 551–93

Stättin H, Magnusson D 1990 *Pubertal Maturation in Female Development*. Erlbaum, Hillsdale, NJ

Thomas W I, Znaniecki F 1918–20 *The Polish Peasant in Europe and America*. University of Chicago Press, Chicago, Vols. 1–2

Zhou X, Hou L R 1999 Children of the Cultural Revolution: The state and the life course in the People's Republic of China. *American Sociological Review* **64**: 12–36

G. H. Elder, Jr.

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Life Expectancy and Adult Mortality in Industrialized Countries

This article is concerned with trends, sex differences, and socioeconomic differences in life expectancy since 1950. The focus is on the so-called established market-economy (EME) countries, including Western Europe, North America, Oceania, and Japan (see *Infant and Child Mortality: Central and Eastern Europe*).

1. *Trends in Life Expectancy in Main Groups of Industrialized Countries*

There were some marked differences in life expectancy among groups of the (currently) industrialized countries in the early 1950s. Average life expectancy in the less developed countries (Southern and Eastern Europe, and Japan) was more than four years lower than in the more developed countries in Western Europe, North America, and Oceania (Fig. 1). In the 1950s and 1960s the differences diminished, as life expectancy increased more rapidly in the high mortality countries.

The period from 1945 to 1970 has been called the 'era of antibiotics' (Kannisto et al. 1999). During this period antibiotics proved effective against a wide range of diseases and infections. The increase in life expectancy was therefore fastest in the less developed areas, where infectious diseases were still common and infant and child mortality high (e.g., Southern Europe and Japan).

This convergence of life expectancies proved short-lived, however; the differences between the main groups of countries started to increase during the late 1960s. By 1980 the differences among the regions were far greater than in 1950. By the early 1990s, life expectancy in Japan was more than 12 years higher than in the countries of the former Soviet Union. This was due to the exceptionally rapid increase in life expectancy in Japan and the stagnation and even decline of life expectancy in the (former) Soviet Union. Life expectancy also increased in the market-economy countries other than Japan, but less rapidly so. In the (former) socialist countries of Eastern Europe the increase in life expectancy was slow during the 1970s and 1980s, and these countries lagged far behind the other industrialized countries in the 1990s. This development is discussed in more detail in the article on mortality in Eastern Europe.

Because infant and child mortality was low as early as in the 1970s in the EME countries, decline in adult mortality became a more important and later the dominant cause of the increase in life expectancy in these countries. For example, in England and Wales about one-third of the increase in life expectancy was attributable to the decline in mortality in the age group 0–14 years from 1970 to 1980, but less than one-tenth

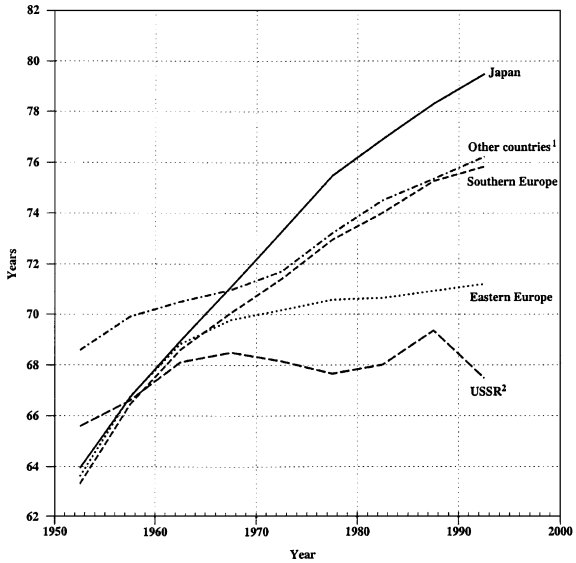

1 Weighted average of Northern Europe, Western Europe, North America and Oceania (UN classification)
2 For the years 1991-1995 weighted average for countries of the former Soviet Union
Source: Dataset of World Population Prospects 1950-2050, United Nations 1998

Figure 1
Life expectancy at birth (both sexes) in groups of industrialized countries 1991–95

from 1980 to 1990. More than half of the increase in the 1980s was due to the decline in mortality among persons over 65 years of age (Meslé 1996).

In populations with a low life expectancy, infant and child mortality have a strong influence on the level of life expectancy at birth. In the EME countries discussed here, this is no longer the case: death rates in the youngest age groups have dropped to such a low level that they have a minor influence on trends differentials in life expectancy at birth. They are determined almost entirely by mortality in adult ages.

The increase in the importance of adult age groups has been mainly due to the decline of mortality from cardiovascular diseases (coronary heart disease, cerebrovascular diseases, and other diseases of the circulatory system). The decline in cardiovascular disease mortality was slow in the 1950s and 1960s, and in many countries male mortality from these diseases actually increased. The decrease in all-cause mortality in older age groups was modest during this period. Around 1970 mortality among the elderly began to decline in most Western countries as never before in recorded history, and gains in survival were observed even at the highest ages (Kannisto et al. 1999).

An analysis of the decline in adult mortality in the United States led Olshansky and Ault (1986) to conclude that the United States had entered a 'fourth stage in epidemiologic transition.' This new stage is characterized by rapid mortality declines in older people resulting from a postponement of the age at

Table 1
Life expectancy at birth) (both sexes) in 1995 and change from 1966–70 to 1995 in established market economy countries (years)

Country	Life expectancy 1995	Change from 1966–70 to 1995
Japan	79.8	8.7
Sweden	79.1	5.0
Switzerland[1]	78.8	6.6
France	78.7	7.2
Italy	78.4	7.4
Canada	78.3	6.3
Australia[1]	78.2	7.4
Spain	78.1	6.5
Iceland	78.0	4.6
Norway	77.9	4.1
Greece	77.8	6.8
Netherlands	77.7	4.1
Luxembourg	77.4	7.3
Austria	77.1	7.3
Belgium[1]	76.9	6.0
United Kingdom	76.9	5.5
Finland	76.8	7.2
Germany	76.8	6.0
New Zealand[1]	76.4	5.1
USA	75.8	5.4
Ireland	75.5	4.4
Denmark	75.4	2.5
Portugal	75.0	8.9

Sources: Dataset of World Population Prospects 1950–2050, United Nations 1998 WHO Europe: Health for all, Statistical Database (HFA-DB) United Nations 2000 Demographic Yearbook 1996–1997, United Nations.
[1] The estimate for 1995 is based on earlier period in the following cases: Australia 1994–96; New Zealand 1993–95.

which degenerative diseases, particularly cardiovascular diseases, tend to kill. They refer to this new stage as the Age of Delayed Degenerative Diseases (see *Mortality, Epidemiological, and Health Transitions*).

Life expectancy increased in all EME countries from the late 1960s to the 1990s. The increase was greatest in Portugal (see Table 1), which had lagged far behind the other countries in the 1960s. Despite this rapid increase Portugal still had the lowest life expectancy in the EME group in 1995. In general, the increase in life expectancy was below average in countries with a high life expectancy in the 1960s (e.g., Norway, Sweden, and The Netherlands).

Denmark is a special case. Danish life expectancy was relatively high in the 1960s, but since then progress has been clearly slower than in the other EME countries. In 1995, life expectancy in Denmark was second lowest after Portugal in the EME group. Japan is another special case. By the late 1960s life expectancy had reached the average level of advanced industrialized countries, by the mid-1990s Japan had the highest life expectancy in the world at 79.8 years.

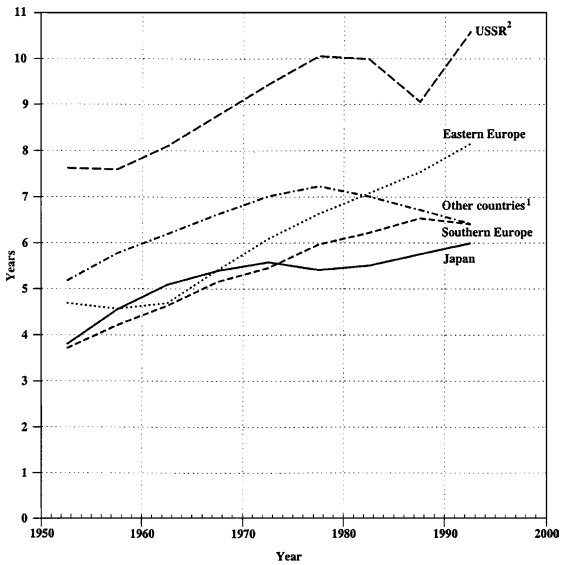

1 Weighted average of Northern Europe, Western Europe, North America and Oceania (UN classification)
2 For the years 1991-1995 weighted average for countries of the former Soviet Union
Source: Dataset of World Population Prospects 1950-2050, United Nations 1998

Figure 2
Female–male difference in life expectancy at birth, industrialized countries by region 1951–95

Cross-national differences in life expectancy among developed countries are not related to standard of living or the level of gross national product (Wilkinson 1996). For example, rich countries such as the US and Germany have lower life expectancies than Greece and Spain. No systematic analyses have been done of the causes of cross-national differences in life expectancy, but several studies have addressed specific questions. These studies are based on the analysis of differences in mortality by cause of death and in the prevalence of known risk factors for specific causes of death. For example, a partial explanation for the high life expectancy in Japan is the low-fat Japanese diet, which contributes to the very low mortality from heart disease. The low mortality from coronary heart disease in southern Europe has also been explained by reference to dietary factors, more specifically to the so-called Mediterranean diet which includes olive oil, fresh fruits, and relatively little animal fat. The exceptionally slow recent increase in life expectancy in Denmark seems to be related to the high prevalence of smoking- and alcohol-related diseases.

2. *Sex Differentials in Life Expectancy*

Women live longer than men in all industrialized countries, but the difference between female and male

Table 2
Life expenctancy at birth by sex in selected EME countries 1966–70 and 1995

	1966–70			1995		
Country	Males	Females	Difference	Males	Females	Difference
Japan	68.5	73.9	5.4	76.4	82.9	6.5
Sweden	71.9	76.5	4.6	76.3	81.8	5.5
Switzerland	69.4	75.1	5.7	75.3	82.1	6.8
France	67.9	75.4	7.4	74.6	82.8	8.2
Italy	68.2	73.9	5.7	75.0	81.6	6.6
Canada	69.0	75.3	6.3	75.3	81.3	6.0
Australia[1]	67.6	74.3	6.7	75.2	81.1	5.9
Spain	69.1	74.3	5.2	74.4	81.8	7.4
Iceland	70.7	76.3	5.6	75.9	80.1	4.2
Norway	71.1	76.7	5.6	74.9	81.0	6.1
Greece	69.3	72.8	3.5	75.2	80.5	5.3
Netherlands	71.0	76.4	5.4	74.7	80.6	5.9
Luxembourg	66.8	73.2	6.4	73.6	81.1	7.5
Austria	66.5	73.4	6.9	73.7	80.3	6.6
Belgium	67.8	74.3	6.5	74.0	80.8	6.8
United Kingdom	68.3	74.6	6.3	74.1	79.5	5.4
Finland	65.9	73.5	7.6	72.9	80.4	7.5
Germany	67.8	73.6	5.8	73.4	80.0	6.6
New Zealand[1]	68.3	74.4	6.1	73.7	79.1	5.4
USA	66.9	74.1	7.1	72.5	78.9	6.4
Ireland	68.9	73.4	4.5	72.8	78.3	5.5
Denmark	70.6	75.3	4.7	72.8	77.9	5.1
Portugal	6.31	69.3	6.2	71.3	78.6	7.3

Sources: Dataset of World Poulation Prospects 1950–2050, United Nations 1998 WHO Europe: health for all, Statistical Database (HFA-DB) United Nations 2000 Demographic Yearbook 1996–1997, United Nations.
[1] The estimate for 1995 is based on earlier period in the following cases: Australia 1994–96; New Zealand 1993–95.

life expectancy varies. At the beginning of the twentieth century the differential in several countries (Bulgaria, Ireland, Italy, and Japan) was small (United Nations Secretariat 1988). The differential was largest (3.9 years) in England and Wales. From the 1930s to the 1970s, increases in the sex differential in life expectancy were a universal feature because women benefited more than men from the decline of mortality.

Figure 2 describes the development of the sex differential in life expectancy for groups of industrialized countries from 1950 to 1995. By the 1950s the average differential in North-Western Europe and North America exceeded five years. The differential was smaller in the less industrialized regions (Eastern and Southern Europe and Japan). In the relatively un-industrialized Soviet Union, however, the difference was clearly largest, almost eight years.

Until the late 1970s the gap between female and male life expectancy grew almost constantly in all regions. This growth was most rapid in the Soviet Union and Eastern Europe. In North-Western Europe, North America, and Oceania the gap started to diminish in the late 1970s and early 1980s, and in Southern Europe a decade later. In Eastern Europe the difference continued to grow rapidly. In the Soviet Union there was a curious temporary decrease from 1976–9 to 1986–9. This was associated with the increase in life expectancy brought about by President Gorbatchev's antialcohol campaign, which had a greater impact on male than female mortality.

Both the sizes of and the trends in the sex difference in life expectancy have varied markedly between countries. For example, France and the United States had approximately the same difference in 1970. In the United States the difference was relatively constant until 1979, and then decreased by more than two years. In France, the sex difference continued to increase into the second half of the 1980s, and diminished only slightly in the early 1990s. In the United Kingdom the decline started earlier than in other countries. In Italy the first signs of decrease did not appear until the 1990s. In 1995 France, Finland, Portugal, and Luxembourg had differentials of more than seven years, whereas the differential was less than five and half years in Sweden, Ireland, Greece, the United Kingdom, Ireland, and Denmark (Table 2).

The lower life expectancy of men in developed countries is due to higher male death rates in all age groups and for most causes of death. An analysis of the contribution of broad groups of causes of death in 26 developed countries in the early 1980s showed that the higher mortality of men from diseases of the

circulatory system, particularly from ischaemic heart disease, made the largest contribution to the differential in life expectancy. Diseases of the circulatory system accounted on average for 40 percent of the total sex differential. Neoplasms (cancer) accounted for 18 percent, more than half of which was due to lung cancer. External causes (accidents, suicide, and violence) accounted for 19 percent. The contributions of these causes of death varied between different countries, however, widely (United Nations Secretariat 1988).

The causes of the higher mortality of men from circulatory diseases, cancer, and other causes of death are only partly known. The size of the sex difference in mortality from each specific cause of death is influenced by a wide variety of environmental and genetic factors, and the contribution of particular factors varies greatly depending on socioeconomic and cultural conditions (Waldron 1986). For example, when maternal mortality was high, that reduced the female–male difference in life expectancy. The decline in maternal mortality has contributed to the increase in the life expectancy gap during this century, but the present very low rate of maternal mortality has hardly any effect on sex differences in life expectancy.

There is evidence to support the hypothesis that female sex hormones reduce women's risk of ischaemic heart disease, in part due to their favorable effects on serum lipids, while men's higher testosterone levels have unfavorable effects on serum lipids. The higher male mortality from ischaemic heart disease is therefore probably due in part to genetic differences between men and women. In countries (e.g., Finland or the United Kingdom) where mortality from ischaemic heart disease is high, these genetic differences contribute more to the overall sex difference in life expectancy than in countries with low mortality from ischaemic heart disease (e.g., Japan or Greece). Waldron (1995) suggests that changes in diet, particularly in the consumption of animal fat, have had greater effects on male than female ischaemic heart disease mortality. It is likely that changes in diet have contributed both to the earlier increase and to the more recent decrease in sex differentials in life expectancy.

A major well-established cause of the increase in the sex differential since the 1930s is that men adopted cigarette smoking much earlier than women. Male cigarette smoking has contributed to the increase in male excess mortality from cardiovascular diseases (particularly ischaemic heart disease), cancer (particularly lung cancer), and diseases of the respiratory system. In her review of 12 studies, most of which were based on samples drawn from the US population from the 1950s to the 1980s, Waldron (1986) estimated that approximately half of the sex difference in adult all-cause mortality was attributable to smoking.

A study on the contribution of smoking to sex differences in life expectancies in four Nordic countries and The Netherlands in 1970–89 (Valkonen and van Poppel 1997) estimated that, on average, more than 40 percent of the average sex difference in these countries in 1970–4 was attributable to more prevalent smoking among men. By 1985–9 the contribution of smoking dropped to approximately 30 percent of the total difference. In all five countries the loss of life expectancy due to cigarette smoking increased among women. In Denmark the loss was 1.6 years in 1985–9, but less than six months in the other countries.

Alcohol use is another important factor contributing to the lower life expectancy of men. There are no cross-national studies, but the contribution of this factor obviously increases in line with rising alcohol-related mortality and growing sex differences in drinking men and women. Mäkelä (1998) estimated that at least one-fifth of the sex difference in life expectancy in Finland in 1987–95 was attributable to the higher male mortality from alcohol-related causes.

Death rates for accidents, suicide, and other violent causes have been found to be higher for males than females in almost all available historical and international data, although the magnitude of the male excess varies. This difference is partly due to the heavier alcohol consumption among men. Social norms and expectations that favor more aggressive and risk-taking behavior among men also play a role. There are also hypotheses that there is a genetic basis for the differences in accident-prone and violent behavior.

There are other factors apart from those discussed above that affect or may affect sex differences in mortality. These include differences in the use of health services, in the prognosis or survival rates for various diseases, economic activity, vulnerability to disruption of social relationships including marriage, etc. However, in the light of the evidence available, these are less significant than behavioral factors.

3. *Socioeconomic Differences in Mortality*

Differences in mortality between subpopulations within countries have been studied by several variables, such as region of residence, rural–urban residence, marital status, and socioeconomic position. These intracountry differences are sometimes as large or larger than cross-country differences, and they are extremely important from the point of view of health and social policy planning. Differences between socioeconomic groups are particularly important because they tend to be greater than other differences between subpopulations.

Figures on mortality by socioeconomic category are not normally presented in regular statistics. The data for England and Wales published by the Registrar General (later the Office of National Statistics) since 1851 in the so-called Decennial Supplements provide the only long-term series of mortality statistics by

occupational and social class (OPCS 1978, Drever and Whitehead 1997). Most of the knowledge about socioeconomic mortality differences is based on specific studies, the results and reliability of which depend on several factors, such as the source and coverage of data, the socioeconomic indicators used, and the way in which mortality and size of mortality differentials are measured.

Despite data problems, study after study in different populations have shown that persons in lower socioeconomic positions die on average at a younger age than those in higher positions. For example, in England and Wales the estimated gap in life expectancy at birth between men in Classes I/II (professional, managerial, and technical occupations) and men in Classes IV/V (partly skilled and skilled workers) was five years in 1987–91; for women the differential was three years (Drever and Whitehead 1997).

Cross-national variations in the extent of socioeconomic differences in mortality by level of education and/or occupational class have been studied in a large-scale project covering 11 Western European countries, the United States, and three former socialist countries in Central and Eastern Europe (Kunst 1997). The main part of the data covered middle-aged men from ca. 1980 to ca. 1990.

In all countries the mortality rate of manual classes was higher than that of nonmanual classes. The relative excess mortality of the manual classes was remarkably similar for most Western European countries and the United States. France and Finland had somewhat greater mortality differences than other Western European countries but smaller than Hungary. The results did not support the view that inequalities in health are smaller in countries where social, economic, and health care policies are to a greater extent influenced by egalitarian principles such as Sweden and Norway (Mackenbach et al. 1997).

It has been observed that male mortality in manual classes is higher than in nonmanual classes for nearly all causes of death. Relative differences are particularly large and systematic in mortality from respiratory diseases and from accidents and violence. The comparative study by Mackenbach and colleagues showed one interesting exception: hardly any class differences were found in France, Switzerland, Italy, and Spain in mortality from ischaemic heart disease. In Portugal mortality rates were higher in the nonmanual than manual classes. On the other hand, socioeconomic gradients in mortality from causes other than ischaemic heart disease were steeper in Southern than in Northern European countries and the United States.

Discussing the reasons for the small differences in mortality from ischaemic heart disease in Southern Europe, Kunst (1997) suggests that this is probably related to the low national levels of ischaemic heart disease. In these countries men seem to be protected against ischaemic heart disease by their frequent consumption of fresh vegetables, fruits, and vegetable oil. Diet may have protected lower socioeconomic groups in particular.

In all countries for which reliable data are available, socioeconomic differences in mortality are smaller among women than men, but there have been no comparative analyses to explore the causes of this phenomenon.

Koskinen and Martelin (1994) found that the smaller socioeconomic differences among women in Finland were mainly due to the different distribution of causes of death among women and men. Causes of death with small (most cancers) or reversed (breast cancer) socioeconomic differentials were more common among women than men. Causes with large differentials (accidents and violence, lung cancer, coronary heart disease), on the other hand, were more common among men.

Data on life expectancy by occupational class for England and Wales as well as for Finland show a clear widening in the gap between manual and nonmanual classes both among men and women since the 1970s (Drever and Whitehead 1997, Valkonen 1999). Results on changes in life expectancy by socioeconomic variables are not available for other countries, but other types of data show that at least relative socioeconomic differences in mortality among men increased in the 1980s in all countries for which some data are available (e.g., the United States, the Nordic countries, and France).

It has been suggested that these trends are a consequence of the general increase in social inequality and income differentials. Although this may be true in some cases, it cannot be generalized to all countries, because mortality differences have increased both in countries where income differentials have increased (e.g., Britain) and in countries where they have decreased (e.g., Finland).

The main common factor contributing to the increase in socioeconomic mortality differentials in several countries is the decline in mortality from cardiovascular diseases, which has been more rapid in the nonmanual and better educated classes than in other population groups. The upper classes have benefited more from the decline in cardiovascular mortality, probably because they have been quicker to adopt recommended health behaviors (concerning diet, smoking, physical activity, etc.), and because they may have had better access to new medical treatments, such as coronary artery bypass grafting and angioplasty.

Causes of socioeconomic differences in mortality have been discussed extensively and many different hypotheses have been offered. Some hypotheses are based on the causal effect of socioeconomic position on mortality. According to these hypotheses differences between classes in material working and living conditions, in health-related behaviors (e.g., smoking, alcohol use, diet) or in the prevalence of psychosocial

stressors cause differences in mortality. Another set of hypotheses is based on assumptions of direct or indirect selection of persons into social classes. According to the direct selection hypothesis, the higher mortality of lower classes is due at least in part to the negative effect of poor health on socioeconomic position. The indirect selection hypothesis assumes that certain characteristics of individuals (e.g., social background, intelligence, or psychosocial traits) affect both their socioeconomic position and their health and mortality.

See also: Adult Mortality in the Less Developed World; Mortality, Biodemography of; Mortality Differentials: Selection and Causation; Mortality of the Oldest Old; Mortality: The Great Historical Decline; Socioeconomic Status and Health; Socioeconomic Status and Health Care

Bibliography

Caselli G, Lopez A D (eds.) 1996 *Health and Mortality among Elderly Populations*. Clarendon Press, Oxford, UK

Daugherty H G, Kammeyer K C W 1995 *An Introduction to Population*. The Guilford Press, New York, London

Department of International Economic and Social Affairs 1982 *Levels and Trends of Mortality since* 1950. ST/ESA/-SER.A/74, United Nations, New York

Drever F, Whitehead M (eds.) 1997 *Health Inequalities. Decennial Supplement*. Office of National Statistics, Series DS No. 15, The Stationery Office, London

Kannisto V, Turpeinen O, Nieminen M 1999 Finnish life tables since 1751. *Demographic Research* 1: 1–26, http://www.demographic-research.org/Volumes/Vol1/1

Koskinen S, Martelin T 1994 Why are socioeconomic mortality differences smaller among women than among men? *Social Science Medicine* **38**: 1385–96

Kunst A 1997 Thesis Erasmus University Rotterdam, Cip-Gegevens Koninklijke Bibliotheek, The Hague, The Netherlands

Lopez A, Caselli G, Valkonen T (eds.) 1995 *Adult Mortality in Developed Countries. From Description to Explanation*. Clarendon Press, Oxford, UK

Mackenbach J P, Kunst A E, Cavelaars A E J M, Groenhof F, Geurts J J M The EU 1997 Working Group on Socioeconomic Inequalities in Health 1997 Socioeconomic inequalities in morbidity and mortality in western Europe. *The Lancet* **349**: 1655–9

Mäkelä P 1998 Alcohol-related mortality by age and sex and its impact on life expectancy. *European Journal of Public Health* **8**: 43–51

Meslé F 1996 Mortality in Eastern and Western Europe: A widening gap. In: Coleman D (ed.) *Europe's Population in the* 1990*s*. Oxford University Press, Oxford, UK

Office of Population Censuses and Surveys 1978 *Occupational Mortality. The Registrar General's Decennial Supplement for England and Wales* 1970–72 . Series DS no. 1, Her Majesty's Stationery Office, London

Olshansky S J, Ault A B 1986 The fourth stage of the epidemiologic transition: The age of delayed degenerative diseases. *Milbank Memorial Fund Quarterly* **64**: 355–91

Trovato F, Lalu N M 1996 Causes of death responsible for the changing sex differentials in life expectancy between 1970 and 1990 in thirty industrialized nations. *Canadian Studies in Population* **23**: 99–126

Uemura K, Piša Z 1988 Trends in cardiovascular disease mortality in industrialized countries since 1950. *World Health Statistics Quarterly* **41**: 155–78

United Nations 1998 *World Population Prospects*. The 1996 Revision. Department of Economic and Social Affairs, Population Division. ST/ESA/SER.A/167, United Nations, New York

United Nations Secretariat 1988 Sex differentials in life expectancy and mortality in developed countries: an analysis by age groups and causes of death from recent and historical data. *Population Bulletin of the United Nations*, Department of International Economic and Social Affairs, No. 25-1988, ST/ESA/SER.N/25, United Nations, New York

Valkonen T 1999 The widening differentials in adult mortality by socio-economic status and their causes. In: Chamie J, Cliquet R L (eds.) *Health and Mortality. Issues of Global Concern*. Proceedings of the symposium on Health and Mortality, Brussels, 19–22 November 1997. Population Division, Department of Economic and Social Affairs, United Nations Secretariat and Population and Family Study Centre, Flemish Scientific Institute, Leuven, Belgium

Valkonen T, van Poppel F 1997 The contribution of smoking to sex differences in life expectancy. Four Nordic countries and The Netherlands 1970–1989. *European Journal of Public Health* **7**: 302–10

Vallin J, D'Souza S, Palloni A (eds.) 1990 *Measurement and Analysis of Mortality. New Approaches*. Clarendon Press, Oxford, UK

Waldron I 1985 What do we know about causes of sex differences in mortality? A review of the literature. *Population Bulletin of the United Nations* No. 18-1985, ST/ESA/SER.N/18, United Nations, New York

Waldron I 1986 The contribution of smoking to sex differences in mortality. *Public Health Reports* **101**: 163–73

Waldron I 1995 Contributions of biological and behavioral factors to changing sex differences in ischaemic heart disease mortality. In: Lopez A, Caselli G, Valkonen T (eds.) *Adult Mortality in Developed Countries*. Clarendon Press, Oxford, UK

Wilkinson R G 1996 *Unhealthy Societies. The Afflictions of Inequality*. Routledge, London and New York

T. Valkonen

Life Management, Developmental Psychology of

1. Brief Historical Perspective on the Concept of Life Management

The term life management is not frequently used in the psychological literature. One of the reasons might be that the active role of the individual in managing his or her life has only entered psychological models after the 'cognitive revolution.' In the following section, a brief historical overview of the emergence of the concept of

life management is given, concluding with a tentative definition of the term 'life management.'

1.1 Emergence of the Recognition of the Active Role of the Individual in Shaping One's Life

To introduce the notion of 'life management' into the description and explanation of human development implies the assumption that people actively contribute to their developmental pathways and are not entirely under the control of external stimuli or biological processes. Although hardly controversial nowadays, it took a revolution—the 'cognitive revolution'—to introduce this notion into psychology in the 1960s that was dominated by behaviorism of the Watsonian or Skinnerian type. The importance of self-regulation for describing and predicting human behavior was stimulated by research by Mischel (1968) and Bandura (1969). Their research has convincingly shown that situational, external stimuli are but one factor influencing behavior and that, when attempting to understand human behavior, it is necessary to take sources of behavior regulation into account that lie within the individual (i.e., self-regulatory processes).

Developmental psychology has long been dominated by the view that development is best described either in terms of biologically determined maturation (i.e., unfolding of a biological program or blueprint) or as the result of a specific learning history in the sense of classic and operant conditioning. The latter view is exemplified in Watson's famous statement that if one were to give him a dozen babies he could—by arranging the environmental conditions—produce whatever adults one would wish to see (Watson 1925). Watson also noted that he was overstating the role of the environment but he wanted to juxtapose the introspective approach to psychology focusing on inner states such as feelings and thinking with the behaviorist position. Although the interaction of biological and environmental factors for development has also been stressed early on, for instance by Piaget, the proactive role of the individual in creating his or her development has only been recognized fairly recently (Lerner and Busch-Rossnagel 1981). In this perspective, development is viewed as the co-constructive interplay of biology, culture/society, and the individual (Baltes 1997). On the one hand, individuals proactively shape their environment and place themselves into certain contexts. On the other hand, individuals react to their environment by continuously adapting and reacting to their physical, cultural, and social environment.

One important way in which individuals play an active role in their development is by choosing, committing to, and pursuing a set of (life-)goals such as wanting to become a teacher or deciding to have a child. The kind of goals a person develops and commits to is, again, embedded in sociocultural, biological, and phylogenetic factors. The less behavior is under control of these factors, the more degrees of freedom a person has to choose what personal goals he or she wants to pursue and how. In other words, the less societal or biological influences structure and determine which, when, and how to pursue goals, the more influential are self-regulatory processes for life management.

1.2 An Attempt to Define 'Life Management'

The term 'life management' is only rarely used in the psychological literature. When searching the currently most comprehensive database for publications in psychology (PsychInfo 11/2000) for 'life management,' only 72 publications are listed for the past century, mostly referring to coping with stress in various life-domains such as work, family, or health. In this context, 'life management' is mostly used more or less synonymously with 'coping.' Taking a developmental perspective, however, a focus on coping appears too narrow. I propose to use the term life management in a broader sense to refer to how people actively take part in and shape their developmental trajectories across the lifespan.

2. Central Approaches to Life Management in Developmental Psychology

Currently, there is no unified model of life management in psychology. Instead, there are a number of approaches to life management that can loosely be clustered into three groups: (a) self-regulation models, (b) models focusing on age-related expectations, and (c) models stressing the role of personal goals. Despite their differences, these approaches have in common that they stress motivational, goal-related processes as central processes of life management across the lifespan.

2.1 Self-regulation Theories

In contrast to 'life management,' self-regulation is a popular concept with more than 1,300 entries listed for the past decade in the PsychInfo database (PsychInfo 11/2000). Mostly, self-regulation is used as an explanatory construct intervening between situational/external factors and individual behavior. In these publications, the term self-regulation refers to a variety of different processes located within the individual regulating behavior, such as social comparison, perceived control, emotional control, delay of gratification, or self-efficacy. Despite the heterogeneity of processes that are subsumed under self-regulation, they have one function in common: they are all concerned with internal processes regulating thoughts, feelings, or actions. Very broadly, self-regulation can be defined as 'the extent to which people influence, modify, or control their own behavior' according to

goals or standards (Baumeister and Heatherton 1996, p. 1).

What goals and standards do people adopt? How do people initiate, execute, and end goal-related actions? Feedback-loop models of self-regulation attempt to answer these questions by referring to cybernetic control processes. Carver and Scheier (e.g., 1998) applied cybernetic models to personal goals. Personal goals can be defined as states a person wishes to achieve, avoid, or maintain. In this context, self-regulation comprises the following three aspects. (a) *Setting of goals and standards*: the setting of goals and standards creates a negative discrepancy of the actual and the desired state that the organism tries to overcome by employing goal-relevant means. (b) *Monitoring*: the comparison of the actual and desired state indicates whether the right means are chosen for achieving one's goals/standards or whether other means need to be employed. (c) *Operation* (goal-pursuit): the initiation, maintenance, and ending of actions for attaining one's goals/standards. There is a large empirical basis supporting the importance of each of these processes for describing goal-directed behavior (see Carver and Scheier 1998, for an overview). One of the features that make this model particularly attractive is that it elaborates general processes of self-regulation that are not content specific and that can be applied equally well to goals of varying levels of abstractness, including short-term goals (e.g., wanting to dine at a nice restaurant) and to long-term goals (e.g., wanting to become the owner of a nice restaurant). Taking a developmental approach to life management, however, this model has one severe limitation: it starts when goals or standards are already set, and does not explain the process of goal-setting itself. How do people develop and select the goals that direct their behavior? This issue is central for the concept of developmental tasks and age-related expectations that will be discussed in the following section.

2.2 Age-related Expectations as Guidelines for Managing One's Life Across the Lifespan

The question, what kinds of goals a person selects at a certain age, is tightly linked with theories about normative development. These theories suggest that individual development is embedded in an age-sequential order of developmental tasks. In contrast to the very general feedback-loop model of self-regulation, Erikson (1968) took a highly content-specific approach to life management. According to Erikson's developmental model, there is a sequence of eight developmental stages covering the entire lifespan, each of which is characterized by a specific theme or, in Erikson's terms a 'psycho-social crisis.' According to Erikson, progression to a higher stage can only occur after having successfully solved the previous crises. Examples for such crises are 'identity vs. identity diffusion' in adolescence and 'generativity vs. stagnation' in middle adulthood. In this framework, life management can be viewed as successfully coping with a sequence of salient themes in each life-phase.

Taking a similar approach to life management, Havighurst introduced the concept of 'developmental tasks' referring to 'a task which arises at or about a certain period in the life of the individual, successful achievement of which leads to his happiness and to success with later tasks, while failure leads to unhappiness in the individual, disapproval by society, and difficulty with later tasks' (Havighurst 1972, p. 2). Havighurst proposes a set of about eight developmental tasks for each of the six phases into which he divides the lifespan from 'infancy' to 'later maturity.' Examples of developmental tasks are 'starting a family' for young adulthood, and 'adjustment to retirement' for old age (see *Adulthood: Developmental Tasks and Critical Life Events*).

The notion of a universal and unidirectional hierarchy of psychosocial crises or developmental tasks has been criticized because it neglects variability between cultures, between individuals, and within individuals (i.e., there is no uniform development across domains of functioning). There is high agreement, however, that within most societies there exist age-related expectations that have evolved from and reflect biological, societal, and historical influences. Such age-related expectations provide information about a timetable of being 'on-time' or 'off-time' (Heckhausen 1999, Neugarten 1968), and a frame of reference regarding the kinds of goals that should be pursued at a given age.

Age-related expectations reflect in part age-graded social opportunity structures (i.e., the amount of resources, support, and reinforcement a society provides for pursuing certain goals for people of a specific age). Complementarily, certain goals are more difficult to strive for when deviating from age-related expectations because necessary resources and support might be less accessible or even negatively sanctioned. In other words, age-related expectations contain information about the likelihood of goal attainment for a given point in the lifespan. Motivational psychology has shown that the expectation to succeed is one of the key determinants for selecting a goal. In this sense, age-related expectations function as a guideline for adaptive goal selection because they indicate which goals should best be pursued at which point in the lifespan, and which goals should better be abandoned. Empirical evidence supports the notion that individuals select personal goals that are congruent with the age-related structure of the lifespan.

Taken together, age-related expectations provide a reference standard regarding the kinds of goals that should be pursued at a given point in the lifespan. Moreover, as they reflect age-related opportunity structures, age-related expectations help to guide an

individual towards goals for which resources are available and that, therefore, have a higher likelihood of being achieved than 'off time' goals. In this sense, age-related expectations are important guidelines for life management across the lifespan.

3. *Personal Goals as Building Blocks for Life Management Throughout the Lifespan and into Old Age*

Action-theoretical approaches view goals as central building blocks of life management because goals structure and direct attention, emotions, and behaviors into particular pathways (Emmons 1996). Recently, a number of models have stressed the importance of goal-related processes for developmental regulation and life management. In the author's view, the currently most prominent models are the model of selection, optimization, and compensation (SOC-model), the model of optimization in primary and secondary control (OPS-model), and the model of assimilative and accommodative coping. In this section, how these complementary models—that all have received good empirical support—conceptualize life management is briefly presented.

The SOC-model (Baltes and Baltes 1990, Freund and Baltes 2000) posits that, throughout the lifespan, individuals continually seek to successfully manage their lives through the orchestration of three processes of developmental regulation: selection, optimization, and compensation.

The importance of selection of goals is based on the assumption that, throughout the lifespan, biological, social, and individual opportunities and constraints specify a range of alternative domains of functioning. The number of options, usually exceeding the amount of internal and external resources available to an individual, needs to be reduced by selecting a subset of these domains on which to focus one's resources. Selection (i.e., developing, elaborating, committing to goals) gives direction to development by guiding and focusing resources on certain domains of functioning (specialization). Moreover, goals organize behavior into action sequences and reduce the complexity of a given situation as they guide attention and behavior.

According to the SOC-model, by delineating the range of goals on which to focus one's resources, selection is the first important step for successfully managing one's life. To actually achieve higher levels of functioning in the selected domains (goals), however, optimization needs to take place. Optimization denotes the process of acquiring, refining, coordinating, and applying the means or resources that are necessary to reach the goals envisioned. The importance of optimization is illustrated in the research on expertise. This literature has demonstrated the central role of the deliberate practice of skills, encompassing the acquisition of new skills, the refinement of existing skills, and the integration and automatization of skills for achieving peak-performance. The process of optimization addresses the growth-aspect of development.

The third component of the SOC-model is compensation (see also Dixon and Bäckman 1995). Compensation is aimed at counteracting losses in previously available goal-relevant means and involves substitution of lost means by activating unused means or by acquiring new means. The objective is to maintain functioning in the face of losses of means. A second form of managing losses is to restructure one's goals by focusing resources on the most important goals, adapting goal standards, or substituting no longer achievable goals (loss-based selection). The SOC-model posits that loss-based selection is an important process of life management as it allows one to focus or redirect resources adaptively when other means for the maintenance of functioning substituting the loss (compensation) are not available or would be invested at the expense of more promising goals.

According to the SOC-model, although all of these processes have positive developmental functions, it is their orchestration that contributes mostly to adaptive life management in the face of changing developmental opportunities, demands, constraints, and personal needs.

Building on the SOC-model and focusing on control, Heckhausen and Schulz (see Heckhausen 1999; *Control Behavior: Psychological Perspectives*) propose in their model of optimization in primary and secondary control (OPS-model) two modes of control jointly contributing to life management across the lifespan, primary and secondary control. According to the OPS-model, people have an innate need to control their world and try to maximize their control potential. Primary control encompasses instrumental efforts to modify the environment according to one's goals. Secondary control is employed when primary control strategies directed at the external world are not available or fail. Secondary control focuses on internal aspects such as changing one's goals, engaging in self-protective attributions, or social comparisons. Heckhausen and Schulz distinguish between selective and compensatory forms of primary and secondary control strategies. Selective primary control refers to the investment of means for goal achievement. Compensatory primary control denotes the employment of external resources (e.g., help of others) or the acquisition of alternative means when available internal resources are insufficient. Selective secondary control refers to enhanced selectivity of resource investment into existing goals. Compensatory secondary control denotes cognitive reframing of goals such as devaluation of unobtainable goals. According to the OPS-model, life management is only optimized when the four types of control are orchestrated according to an individual's developmental ecology.

A similar view was developed by Brandtstädter and his colleagues (e.g., Brandtstädter and Renner 1990) in their model of assimilation and accommodation. According to this model, people use two complementary forms of coping to achieve a match between the actual and the desired developmental state, assimilation, and accommodation. Assimilation denotes an actional mode of coping that is characterized by tenaciously pursuing one's goals by modifying the environment according to personal goals. Accommodation refers to a nonintentional, unconscious mode of coping, namely flexible goal adjustment. Accommodation comprises changing, downgrading, or discarding personal goals or lowering one's level of aspirations to match the environment. Following Brandtstädter, people will first use assimilative strategies to close the mismatch between the actual state and personal goals. When assimilative attempts fail, a gradual shift to the accommodative mode occurs, first by investing additional goal-related activities, then by engaging external aids or support systems, and, finally, by adjusting standards or even disengaging from goals.

The three models reported above converge in the importance they place on goals and motivational processes as building blocks of life management. Moreover, these models suggest that key processes of life management across the lifespan lie in the adaptation to developmental ecologies by opportunity-oriented selection, focused investment of resources, and compensation of losses. Age-related changes in the availability and efficiency of resources necessitate the flexible use of goal related strategies. Starting out to be highly positive in childhood through early adulthood, the ratio of gains to losses becomes less favorable with age. Thus, with increasing age, individuals need to allocate more resources into the maintenance of functioning and compensation of losses than into growth (Staudinger et al. 1995). Life management, following these models, encompasses optimizing developmental potential and simultaneously compensating for losses.

4. Outlook

Different lines of research inform us that one important way in which individuals manage their lives is by developing, committing to, and pursuing a set of (life-) goals. What goals a person selects and pursues is embedded in sociocultural, biological, and phylogenetic factors.

An interesting question awaiting empirical investigation is the relative importance of these factors across the lifespan. During childhood, adolescence, and young adulthood, biological constraints and social expectations regarding the kinds of goals to be pursued are relatively strong. In late adulthood and old age, there are fewer normative age-related expectations. This relatively greater social freedom might strengthen the impact of personal goals for life management. At the same time, in old age resources needed for shaping one's environment according to personal goals decline. Therefore, it seems particularly crucial to focus on goals that are both personally very important and for which resources are still available. Future investigations of the interaction of the various factors influencing the selection, pursuit, and maintenance of goals across the lifespan might increase our understanding of how people manage their lives.

See also: Aging, Theories of; Childhood and Adolescence: Developmental Assets; Life Course: Sociological Aspects; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development; Personality and Adaptive Behaviors; Personality and Conceptions of the Self; Self-monitoring, Psychology of; Self-regulation in Adulthood; Self-regulation in Childhood

Bibliography

Baltes P B 1997 On the incomplete architecture of human ontogeny: selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80

Baltes P B, Baltes M M 1990 Psychological perspectives in successful aging: the model of selective optimization with compensation. In: Baltes P B, Baltes M M (eds.) *Successful Aging: Perspectives from the Behavioral Sciences*. Cambridge University Press, New York, pp. 1–34

Bandura A 1969 *Principles of Behavior Modification*. Holt, Rinehart, and Winston, New York

Baumeister R F, Heatherton T F 1996 Self-regulation failure: an overview. *Psychological Inquiry* **7**: 1–15

Brandtstädter J, Renner G 1990 Tenacious goal pursuit and flexible goal adjustment: explication and age-related analysis of assimilative and accommodative strategies of coping. *Psychology and Aging* **5**: 58–67

Carver C S, Scheier M F 1998 *On the Self-regulation of Behavior*. Cambridge University Press, Cambridge, UK

Dixon R A, Bäckman L (eds.) 1995 *Compensating for Psychological Deficits and Declines: Managing Losses and Promoting Gains*. Erlbaum, Mahwah, NJ

Emmons R A 1996 Striving and feeling. Personal goals and subjective well-being. In: Gollwitzer P M, Bargh J A (eds.) *The Psychology of Action: Linking Cognition and Motivation to Behavior*. Guilford Press, New York, pp. 313–37

Erikson E H 1968 *Identity: Youth and Crisis*. Norton, New York

Freund A M, Baltes P B 2000 Selection, optimization, and compensation: an action-theoretical conceptualization of processes of developmental regulation. In: Perrig W J, Grob A (eds.) *Control of Human Behaviour: Mental Processes and Consciousness*. Erlbaum, Mahwah, NJ, pp. 35–58

Havighurst R J 1972 *Developmental Tasks and Education*, 3rd edn. Davis McKay, New York

Heckhausen J 1999 *Developmental Regulation in Adulthood: Age Normative and Sociostructural Constraints as Adaptive Challenges*. Cambridge University Press, New York

Lerner R M, Busch-Rossnagel A (eds.) 1981 *Individuals as Producers of Their Development: A Life-span Perspective*. Academic Press, New York
Mischel W 1968 *Personality and Assessment*. Wiley, New York
Neugarten B L 1968 The awareness of middle age: a reader in social psychology. In: Neugarten B L (ed.) *Middle Age and Aging*. University of Chicago Press, Chicago, pp. 22–8
Staudinger U M, Marsiske M, Baltes P B 1995 Resilience and reserve capacity in later adulthood: potentials and limits of development across the life span. In: Cicchetti D, Cohen D (eds.) *Developmental Psychopathology*. Vol. 2: *Risk, Disorder, and Adaptation*. Wiley, New York, pp. 801–47
Watson J B 1925 *Behaviorism*. Norton, New York

A. M. Freund

Life Table

The life table is perhaps the oldest tool developed for the analysis of survival patterns in human and other populations. Its roots go back at least to the famous book on the English Bills of Mortality that John Graunt published in 1662, a book that is sometimes cited as the beginning of systematic statistical science. Life table techniques are described in detail in most introductory textbooks on the methods of actuarial statistics, biostatistics, demography, and epidemiology (see, e.g., Chiang (1984), Elandt-Johnson and Johnson (1980), Manton and Stallard (1984), Preston et al. (2001). In this article the basic notions of life-table construction are explained with a focus on issues that are sometimes underemphasized in other accounts. More of the underlying mathematical theory can be found in Hoem (1998) and in the textbooks. For the history of the topic, consult Seal (1977), Smith and Keyfitz (1977), and Dupâquier (1996).

1. *A Cohort Life Table*

Classical presentations of life-table techniques typically start with data in a format like that of the first four columns in Table 1. These particular data come from the Eritrean Demographic and Health Survey of 1995 and show the survival up through their fourth birthday of 7,098 girls whose births and deaths were recorded in the survey. This data set is different from normal textbook examples in two ways. First, the data are given for intervals of single-*month* segments instead of the usual single-*year* intervals. Second, the present data are subject to strong heaping in the reported age at death (for the girls that died). In part this is due to a typical inherent inaccuracy of age reporting in societies like the data source at the time of data collection, and in part to the fact that (probably for that very reason) only integer years of age attained was asked except for very young ages at death. By contrast, an individual child's age when observation ended (*age at censoring*) could be computed accurately to the month, for in these data a child's record was censored only because the child's parent was interviewed; the age of a live child could then be calculated from its date of birth. This unconventional data set is used in order to illuminate how a number of practical issues that confront a demographer can be handled in a first analysis of mortality. A more complete investigation would subsequently use intensity-regression methods from event-history analysis (see *Event History Analysis: Applications*), as was done by Woldemicael (1999).

If there had not been any heaping in the reported ages at death, one would calculate $R_x = S_x - \frac{1}{2}(D_x + W_x)$ as an approximate number of person-months of exposure at age x. Here S_x is the number of children who survive to exact age x months, D_x is the number of deaths recorded between exact ages x and $x+1$ months, and W_x is the corresponding number of records that were censored during this interval. The approximation consists in assigning deaths and interviews to in the middle of the month on average. It has become known as the *actuarial method*. The next step would be to compute occurrence/exposure rates D_x/R_x (see *Demographic Techniques: LEXIS Diagram*). The age heaping is abundantly evident in the plot of the 'raw' death rates (converted to 'deaths per 1,000 person-years') in Table 1.

Such age heaping must have been caused by a shift of the real age at death to a reported near-by 'round' *anchor age* like three months, six months, or multiples of 12 months. The deaths in some interval $[x_0 - a, x_0 + b]$ around each anchor age x_0 have largely been assigned to the anchor age instead of to each age in the interval, but the total number of deaths $\sum_{t=-a}^{b} D_{x_0+t}$ in such an interval should be largely correct. Similarly, the total exposures $\sum_{t=-a}^{b} R_{x_0+t}$ in the interval should be about right, even though the individual term in the latter sum may be wide off the mark. Under these circumstances it is more sensible to compute death rates for groups of reported ages and list them in the row for the corresponding anchor age. By this reasoning,

$$m_3 = \sum_{x=2}^{4} D_x / \sum_{x=2}^{4} R_x$$

is entered for age three months,

$$m_6 = \sum_{x=5}^{7} D_x / \sum_{x=5}^{7} R_x$$

is entered for age six months, and so on, as indicated in column 7 of Table 1. For the intervening ages x, rates m_x have been computed by linear interpolation. The corresponding curve of rates is given with the label 'grouped' in Table 1.

The rest of the life-table computations are based on the following simple mathematical theory. Each individual child i has a lifetime T_i with some distribution function $F(t) = P\{T_i \leqslant t\}$ (with $F(0) = 0$), a probability density $f(t) = \mathrm{d}F(t)/\mathrm{d}t$, and a *force of mortality* or *death intensity* $\mu(t) = f(t)/\{1-F(t)\}$. The various lifetimes T_i are independent of each other and they have the same distribution $F(t)$. Simple integration gives $F(x) = 1-\exp\{-\int_0^x \mu(t)\,\mathrm{d}t\}$. The death probability is

$$q_x = P\{T_i \leqslant x+1 \mid T_i > x\} = \{F(x+1)-F(x)\}/F(x)$$

from which is derived

$$q_x = 1-\exp\{-\int_0^1 \mu(x+t)\,\mathrm{d}t\} \tag{1}$$

If one approximates the function $\mu(x+t)$ by the constant m_x for $0 < t < 1$ (the assumption of *piecewise constancy*), then (1) gives $q_x = 1-e^{-m_x}$, which has been used to compute the next-to-last column in Table 1. Finally, a life-table *survival function* is defined as $\ell(x) = \ell(0)\{1-F(x)\}$, with $\ell(0) = 100\,000$. This is the expected number of survivors to exact age x out of an original cohort of $\ell(0)$ initial individuals. Simple manipulation with the definition of q_x gives

$$\ell(x+1) = \ell(x)(1-q_x), \quad \text{for } x = 0, 1, \ldots \tag{2}$$

This is a practical recursive formula by which we have computed the final column in Table 1. From the final entry in the $\ell(x)$ column in Table 1 we see that more than 15 percent of the original cohort of girl babies would die before age 4 (years) according to our data, because $1-0.84859 = 0.15141$.

The mean number of months lived between birth and age n is

$$e^o_{0:\,n} = E\min(T_i, n) = \int_0^n \ell(t)\,\mathrm{d}t/\ell(0) \tag{3}$$

If one partitions the integral here into a sum of n terms $\int_{x-1}^x \ell(t)\,\mathrm{d}t$ for $x = 1, 2, \ldots, n$, use the trapezoidal rule of numerical integration to see that this integral is approximately equal to $\frac{1}{2}\{\ell(x-1)+\ell(x)\}$, and collect terms suitably, we get the numerical approximation

$$e^o_{0:\,n_n} \simeq \left\{ \sum_{x=0}^{n} \ell(x) - \frac{1}{2} - \frac{1}{2}\ell(n) \right\} / \ell(0) \tag{4}$$

The girls in the Eritrean data set lost five months of life on average before they reached their fourth birthday, for according to Eqn. (4) with $n = 48$ months they experienced a mean lifetime of 43 months.

2. Other Life-table Procedures

Life-table construction consists in the estimation of parameters and tabulation of functions like those above from empirical data. The data can be for age at death to individuals, as in the illustrative example, but they can also be observations of duration until recovery from an illness, of intervals between births, of time until breakdown of some piece of machinery, or of any other positive duration variable. In general, a T_i is the duration until occurrence of some event that ends individual survival in a given status. The method of estimation depends on the character of the data. If accurate (ungrouped) individual-level data are available, then the Kaplan–Meier estimator (see *Event History Analysis: Applications*) can be used to estimate ℓ_x for all relevant x, and estimates of the other life-table functions can then be computed subsequently. Alternatively a segment of the Nelson–Aalen estimator can be used to estimate each integral $\int_0^1 \mu(x+t)\,\mathrm{d}t$, Eqn. (1) can then be used to estimate q_x for each integer x, and the rest of the computations follow suit. Sometimes the force of mortality is represented by some function $h(x; \theta)$, where θ is a vector of parameters. For instance, actuaries often use the classical Gompertz–Makeham function $h(x; a, b, c) = a+bc^x$ for the force of mortality in their life tables (see *Demographic Models*; *Demographic Analysis: Probabilistic Approach*).

From any given schedule of death probabilities $q_0, q_1, q_2, \ldots$, the $\ell(x)$ table is easily computed using the recursive Eqn. (2). Much of the effort in life-table construction is, therefore, concentrated on providing a schedule $\{q_x\}$. So far, this account has assumed that the data come from a group of independent individuals who have all been observed in parallel, essentially a cohort that is followed from a significant starting point (namely from birth in our empirical example) and which is diminished over time due to *decrements* (*attrition*) caused by the risk in question and also subject to reduction due to censoring. The empirical example displayed data only for the first four years of life of such a cohort. One would have had to follow the cohort until the end of the cohort's total life to produce a *complete cohort life table*. It is more common to compute a q_x schedule from data collected for the members of a population during a limited time period and to use the mechanics of life-table construction to produce a *period life table* from the q_x values. If real mortality patterns are tied to cohorts, individuals who live at widely differing ages in the observational period do not normally have the same risk structure, and the period life table is taken to reflect the patterns of a

Table 1
Life table computations for Eritrean girls aged 0 through 4 years

Age x in months	Observed survivors	Deaths D_x	Censored W_x	Exposures R_x	Raw rate D_x/R_x per 1000	Age grouping (months)	Grouped rates m_x	Death probab. $1000q_x$	Life table survivors ℓ_x
0	7098	208	15	6986.5	357.26	0	357.26[a]	29.33	100000
1	6875	48	39	6831.5	84.32	1	84.32[a]	7.00	97067
2	6788	36	39	6750.5	64.00		70.01[b]	5.82	96387
3	6713	43	36	6673.5	77.32	2–4	55.70[a]	4.63	95826
4	6634	14	29	6612.5	25.41		56.00[b]	4.66	95383
5	6591	21	22	6569.5	38.36		56.30[b]	4.68	94939
6	6548	50	42	6502	92.28	5–7	56.60[a]	4.71	94494
7	6456	21	23	6434	39.17		54.98[b]	4.57	94050
8	6412	26	31	6383.5	48.88		53.35[b]	4.44	93620
9	6355	19	26	6332.5	36.00		51.73[b]	4.30	93204
10	6310	16	23	6290.5	30.52		50.11[b]	4.17	92803
11	6271	10	37	6247.5	19.21		48.48[b]	4.03	92417
12	6224	144	30	6137	281.57	8–17	46.86[a]	3.90	92044
13	6050	6	34	6030	11.94		45.45[b]	3.78	91685
14	6010	7	31	5991	14.02		44.05[b]	3.66	91339
15	5972	3	15	5963	6.04		42.64[b]	3.55	91004
16	5954	5	35	5934	10.11		41.24[b]	3.43	90681
17	5914	3	31	5897	6.10		39.83[b]	3.31	90370
18	5880	25	34	5850.5	16.40		38.43[b]	3.20	90071
19	5821	5	24	5806.5	10.33		37.02[b]	3.08	89783
20	5792	2	28	5777	4.15		35.62[b]	2.96	89506
21	5762	0	16	5754	0.00		34.21[b]	2.85	89241
22	5746	0	29	5731.5	0.00		32.81[b]	2.73	88987
23	5717	2	18	5707	4.21		31.40[b]	2.61	88744
24	5697	148	25	5610.5	316.55	18–30	30.00[a]	2.50	88512
25	5524	0	34	5507	0.00		29.28[b]	2.44	88291
26	5490	0	32	5474	0.00		28.57[b]	2.38	88076
27	5458	0	31	5442.5	0.00		27.85[b]	2.32	87866
28	5427	0	27	5413.5	0.00		27.14[b]	2.26	87663
29	5400	0	35	5382.5	0.00		26.42[b]	2.20	87465
30	5365	0	33	5348.5	0.00		25.71[b]	2.14	87272
31	5332	0	39	5312.5	0.00		24.99[b]	2.08	87086
32	5293	0	28	5279	0.00		24.27[b]	2.02	86904
33	5265	0	36	5247	0.00		23.56[b]	1.96	86729
34	5229	0	24	5217	0.00		22.84[b]	1.90	86559
35	5205	0	28	5191	0.00		22.13[b]	1.84	86394
36	5177	109	22	5111.5	255.89	31–42	21.41[a]	1.78	86235
37	5046	0	38	5027	0.00		20.44[b]	1.70	86081
38	5008	0	20	4998	0.00		19.48[b]	1.62	85935
39	4988	0	32	4972	0.00		18.51[b]	1.54	85795
40	4956	0	31	4940.5	0.00		17.54[b]	1.46	85663
41	4925	0	30	4910	0.00		16.57[b]	1.38	85538
42	4895	0	30	4880	0.00		15.60[b]	1.30	85420
43	4865	0	24	4853	0.00		14.63[b]	1.22	85309
44	4841	0	26	4828	0.00		13.66[b]	1.14	85205
45	4815	0	18	4806	0.00		12.70[b]	1.06	85108
46	4797	0	20	4787	0.00		11.73[b]	0.98	85018
47	4777	0	16	4769	0.00		10.76[b]	0.90	84935
48[c]	4761	46	20	4728	116.75	43–54	9.79[a]	0.82	84859

a: Computed as sum of deaths/sum of exposures for age groups indicated
(Used for rows where age groups are indicated) b: Computed by linear interpolation
c: Additional rows for ages 49–60 months have been deleted

synthetic (fictitious) *cohort* exposed to the risks of the period at their various ages.

Whichever way the life-table survivor column has been produced, Eqn. (4) can always be used to calculate mean numbers of time-units lived under the risk for which the life-table was constructed. If the time unit is a year and n is chosen so large that no one lives more than n years (symbolized by letting $n = \infty$ and $\ell(\infty) = 0$), $e^{o}_{0:\overline{\infty}|}$ is called the *life expectancy* of the life table and is normally denoted e^{o}_{0}. The subscript 0 indicates that the calculation is made for a newborn individual. Similar computations can be made at any age x and one gets expected remaining lifetimes of the form $e^{o}_{x:\overline{n}|} = \int_0^n \ell(x+t)\,\mathrm{d}t/\ell(x)$, with approximation formulas similar to Eqn. (4).

3. *Multiple-decrement tables*

When suitable data are available, the risk of death (or force of mortality) may be partitioned according to mutually exclusive causes of death. Let $\mu_k(t)$ be death intensity or force of mortality for cause k at age t, and let the total force of mortality be $\mu(t) = \sum_k \mu_k(t)$. Then the probability that an individual who is alive at age x will die from cause k before age $x+1$ is

$$q_x^{(k)} = \int_0^1 {}_tp_x \mu_k(x+t)\,\mathrm{d}t \tag{5}$$

where ${}_tp_x$ is the probability of surviving to age $x+t$ for an individual who is alive at age x. For given risk intensities $\{\mu_k(\cdot)\}$, $q_x^{(k)}$ can be computed by numerical integration in Eqn. (5). A corresponding column of $q_x^{(k)}$ values may then be added to the life table for each k, and other cause-specific life table functions may also be computed. Note that $q_x^{(k)}$ is influenced by risk intensities other than $\mu_k(\cdot)$, for

$${}_tp_x = \exp\left\{-\sum_j \int_0^t \mu_j(x+s)\,\mathrm{d}s\right\}$$

where the exponent depends on all $\mu_j(\cdot)$, including those for $j \neq k$.

Several further life-table functions can be defined by formal reduction or elimination of one or more of the risk intensities in formulas like Eqn. (5). In particular, a *single-decrement life table* can be computed for each cause k in order to show what the life table would look like in the hypothetical case where this cause were the only one operating in the study population and where it did so with the risk function estimated from the data. The purpose is to see the '*pure*' effect of the risk in question on a fictitious cohort without interference from other causes. (The decrement probabilities and other features of the single-decrement life table do not depend on any other intensities than the one for which the table is computed.) This does not mean that one should believe that the total risk intensity can actually be reduced to the risk in focus or that this risk operates independently of other causes.

Most single-decrement life-table functions have straightforward interpretations. If, for instance, $\ell_k(x)$ denotes the corresponding survival function, then $\ell_k(x)/\ell_k(0)$ is the probability of surviving to age x in the fictitious table. Suppose, for example, that the study population is a group of unpartnered individuals subject to the competing risks of marital and non-marital union formation as well as to the risk of death. One can then compute a single-decrement life table based on the risk of transition into a nonmarital union alone and find the probability that an individual would end up in such a union by age 50, say, if marriage were no alternative and death did not operate, even though in reality marriages continue to be formed by group members and unpartnered individuals continue to die. In such a table one may get $\ell_k(\infty) > 0$ because the event never occurs to some individuals, even if it operates alone. (We will all die but some individuals never enter a nonmarital partnership even if they live to age 100.)

Life expectancies $e^{o}_{15:\overline{n}|}$ computed in the three-decrement life table would be interpreted as the mean number of years (say) lived in the unpartnered state between ages 15 and $15+n$. In the single-decrement life table for nonmarital-union formation, one possibility is to compute the mean age of entry into a nonmarital union only for those who do enter such a union by age n, that is, the conditional mean

$$\int_0^n \ell_k(x)\,\mathrm{d}x/\{\ell_k(0)-\ell_k(n)\} \tag{6}$$

Conversely one can compute a *cause-deleted life table* by eliminating one or more of the cause intensities in the formulas. To 'remove' cause k, introduce $\mu_{-k}(t) = \mu(t)-\mu_k(t)$ and construct a 'normal' life table by replacing $\mu(t)$ by $\mu_{-k}(t)$ everywhere. This would show what the life table would look like *if* it were possible to eliminate cause k without changing the risk of any other cause. It could be used for example to see how partnering probabilities would appear if the partnering process were undisturbed by mortality. One would then remove the force of mortality from the life-table formulas and operate only with the sum $\mu_{-k}(t)$ of the intensities of marital and nonmarital union formation in the computations. Note that this is again a purely hypothetical construct, in fact it is counterfactual for mortality cannot of course be completely removed.

Life expectancies can of course be calculated for a cause-deleted life table as for any other life table. If $\ell_{-k}(x)$ is the table's survival function and $\ell_{-k}(\infty) = 0$, then a formula for the cause-eliminated life expectancy is $e_0^{(-k)} = \int_0^\infty \ell_{-k}(x)\,\mathrm{d}x/\ell_{-k}(0)$. The classical applications were to cause-of-death removal, and the difference

$e_0^{(-k)} - e_0^o$ was interpreted as the (fictitious) gain in life expectancy produced by removal of cause k. For example, Westergaard (1907) found that according to the English life tables of 1881–90 life expectancy for females would increase by 2.7 years if tuberculosis deaths could be eliminated, by a year if cancer deaths could be removed, and by 1.1 years if deaths due to diarrhea and dysentery could be deleted separately. The effects were not additive, and if all three cause-of-death groups could be removed, the gain would be 5.6 years and not 4.8. The life expectancy for females in the original table was $e_0^o = 47.2$ years.

See also: Age Structure; Life Course: Sociological Aspects

Bibliography

Chiang C L 1984 *The Life Table and its Applications*. Krieger, Malabar, FL

Dupâquier J 1996 *L'invention de la table de mortalité*. Presses universitaires de France, Paris

Elandt-Johnson R C, Johnson N L 1980 *Survival Models and Data Analysis*. Wiley, New York

Graunt J 1662 *Natural and Political Observations Made upon the Bills of Mortality*. Roycroft, London

Hoem J M 1998 The life table. In: Armitage P, Colton T (eds.) *Encyclopedia of Biostatistics*. Wiley, Chichester, pp. 2235–9

Manton K G, Stallard E 1984 *Recent Trends in Mortality Analysis*. Academic Press, Orlando

Preston S H, Heuveline P, Guillot M 2001 *Demography: Measuring and Modeling Populations*. Blackwell, Oxford, UK

Seal H 1977 Studies in history of probability and statistics, 35: Multiple decrements or competing risks. *Biometrika* **64**(3): 429–39

Smith D, Keyfitz N 1977 *Mathematical Demography: Selected Papers*. Springer-Verlag, Heidelberg, Germany

Westergaard H 1907 The horoscope of the population in the twentieth century. *Compte-Rendu de la XI Session de l'Institut international de Statistique á Copenhague du 26 au 31 aout 1907*

Woldemicael G 1999 Infant and child mortality in Eritrea: Levels, trends, and determinants. Ph.D. thesis, Stockholm University

J. M. Hoem

Lifelong Learning and its Support with New Media: Cultural Concerns

1. Introduction

Learning needs to be examined across the lifespan because traditional notions of a divided lifetime—education followed by work—are no longer tenable (Gardner 1991). Professional activity has become so knowledge-intensive and fluid in content that learning has become an integral and irremovable part of work activities. Learning is a new form of labor, and working often is (and needs to be) a collaborative effort among colleagues and peers. In the emerging information society, an educated person will be someone who is willing and able to consider learning as a lifelong process. More and more knowledge, especially advanced knowledge, is acquired well past the age of formal schooling, and in many situations through educational processes that do not center on traditional schools (Illich 1971). Lifelong learning is more than adult education, which often is restricted to providing people with opportunities to engage in (school-like) learning activities during their adult life. The challenge for lifelong learning is to fundamentally rethink learning, teaching, and education for the information age in an attempt to change mindsets. It involves and engages learners of all ages in acquiring and applying knowledge and skills in the context of authentic, self-directed problems, and it exploits the possibilities offered by new media.

Lifelong learning has emerged as one of the major challenges for the worldwide knowledge society of the future. A variety of recent events supports this claim: (a) 1996 was the 'European Year of Lifelong Learning,' (b) UNESCO (United Nations Educational, Scientific and Cultural Organization) has included 'Lifetime Education' as one of the key issues in its planning, and (c) the G7/G8 group of countries has named 'Lifelong Learning' as a main strategy in the fight against unemployment. Despite this great interest, there are few encompassing efforts to tackle the problem in a coherent way. Lifelong learning is comprehensive; it cannot be investigated in isolation by looking just at one small part of it, such as K-12 education, university education, or worker re-education.

2. Problems in the Information Age

2.1 Lack of Creativity and Innovation

Societies and countries of the future will be successful not 'because their people work harder, but because they work smarter.' Creativity and innovation are considered essential capabilities for working smarter in knowledge societies (Drucker 1994); thus, an important challenge is how these capabilities can be learned and practiced. An implicit assumption is made that self-directed and lifelong learning can influence the creativity and innovation potential of individuals, groups, organizations, and countries (Dohmen 1999).

2.2 Coping with Change

Most people see schooling as a period of their lives that prepares them for work in a profession or for a change of career. This view has not enabled people to

Table 1
Transcending Skinner and Taylor

Skinner–Taylor	Beyond Skinner and Taylor
There is a scientific, best way to learn and to work	Real problems are ill-defined and wicked
Separation of thinking, doing, and learning	Integration of thinking, doing, and learning
Task domains can be completely understood	Understanding is partial, tacit and situated; total coverage is impossible
Objective ways to decompose problems into standardizable actions	Subjective, situated personal interests; need for iterative explorations
All relevant knowledge can be explicitly articulated	Much knowledge is tacit and relies on tacit skills
Teacher–manager as oracle	Teacher–manager as facilitator or coach
Operational environment: mass markets, simple products and processes, slow change, certainty	Customer orientation, complex products and processes, rapid and substantial change, uncertainty and conflicts

cope well with the new realities of our world that (a) most people change careers several times in their lives, even though what they learned in school was designed to prepare them for their first career; and (b) that the pace of change is so fast that technologies and skills to use them become obsolete within 5–10 years.

2.3 School-to-work Transition is Insufficiently Supported

If the world of working and living relies on collaboration, creativity, and framing of problems; deals with uncertainty, change, and distributed cognition; and augments and empowers humans with powerful technological tools, then the world of schools and universities needs to prepare students to function in this world. Industrial-age models of education and work (based on B. F. Skinner and F. W. Taylor; see Table 1) are inadequate to prepare students to compete in the knowledge-based workplace.

2.4 The 'Gift Wrapping' Approach Dominates Educational Reform

Information technologies have been used to mechanize old ways of doing business, rather than fundamentally rethinking the underlying work processes and promoting new ways to create artifacts and knowledge. In learning, these technologies have been used primarily as add-ons ('gift-wrapping') to existing practices (Fischer 1998b) rather than a catalyst for fundamentally rethinking what education should be about in the next century. Conventional frameworks, such as instructionism, fixed curricula, memorization, decontextualized rote learning, etc., are not changed by technology itself. 'Computer literacy' has often been equated with a bag of superficial and transitory knowledge rather than with what really matters about computers: namely, that all people can use them for their own purposes. New technologies and new media (such as the World Wide Web, multimedia design, etc.) are necessary to achieve certain learning objectives and to provide foundations for people to change their mindsets, but they by themselves are not sufficient.

3. Changing Mindsets

Cultures are substantially defined by their media and their tools for thinking, working, learning, and collaborating. A large number of new media are designed to see humans only as consumers. Television is the most obvious medium that promotes this mindset and behavior (Postman 1985). Unfortunately, a consumer mindset does not remain limited to television, but in many cases is a model dominating our culture. In our educational institutions, learners are often treated as consumers, creating a mindset of consumerism for the rest of their lives (Illich 1971). Citizens often feel left out in the decisions by policy makers, denying them opportunities to take an active role. The current mindset about learning, teaching, and education is dominated by a view in which teaching is often fit 'into a mold in which a single, presumably omniscient teacher explicitly tells or shows presumably unknowing learners something they presumably know nothing about' (Bruner 1996). A critical challenge is a reformulation and reconceptualization of this impoverished and misleading conception. Although this model may be more realistic for the early grades in schools, it is obviously inadequate for learning processes as they occur in lifelong learning, where knowledge is distributed among many stakeholders and 'the answer' does not exist or is not known. Learning is more than being taught (Illich 1971). Learning new skills and acquiring new knowledge cannot be restricted to formal educational settings. By integrating working and learning, people learn within the context of their work on real-world problems. Learning does not take place in a separate phase and in a separate place, but is integrated into the work process. The direct usefulness of new knowledge for actual problem situations greatly improves the motivation to learn

Table 2
Comparison of different conceptualizations of school/university and lifelong learning

	School–University	Lifelong learning
Emphasis	'Basic' skills; exposure; access	Education embedded in ongoing work activities; informed participation
Potential drawbacks	Decontextualized, not situated	Relevant concepts are not encountered
Problems	Given	Constructed
New topics	Defined by curricula	Arise incidentally from work situations
Structure	Pedagogic or 'logical' structure	Work activity
Roles	Expert–novice model	Asymmetry of knowledge; reciprocal learning
Teachers–coaches	Expound subject matter	Engage in work practice
Mode	Instructionism (knowledge absorption)	Constructionism (knowledge construction)

new materials because the time and effort invested in learning are immediately worthwhile for the task at hand—not merely for some putative long-term gain.

Many conventional frameworks are grounded in the behaviorist learning theory of Skinner and the models of industrial work of Taylor. Table 1 contrasts these approaches with new goals and new objectives for the information age.

The conceptual framework outline in Table 1 provides the foundation to differentiate between school and lifelong learning, as illustrated by Table 2. Schools often create mindsets about learning, teaching, and collaboration that would require a 'big switch' at some point in people's lives: seeing themselves as active contributors rather than as passive consumers. To create different mindsets, we need to change our formal educational institutions to environments in which knowledge is (collaboratively) created, externalized and shared (Arias et al. 2000; http://www.cs.colorado.edu/~gerhard/papers/tochi99.pdf.). Skills and processes that support learning as a lifetime habit must be developed, and efforts must be undertaken for learners to develop a positive attitude, enjoyment, and motivation toward learning—because no one will engage in processes and attitudes during their whole life for goals and objectives that they do not like.

4. *Emphasizing New Forms of Learning from a Lifelong Learning Perspective*

Lifelong learning refers to a society in which learning possibilities exist for those who want to learn. Table 3 summarizes four innovative forms of lifelong learning, and addresses their contributions toward the creation of mindsets and the media requirements generated by them.

5. *The Impact of Lifelong Learning Themes on Mindsets*

Impacting mindsets is a socio–technical design problem (Fischer 1999). Research and practice should be conducted as an instrument of change to alter and improve the way people learn and think. Creating innovative media and technologies supporting a 'new' mindset requires the following co-evolution: we must rethink and reinvent learning, working, collaboration, and education in the context of new media, and simultaneously we must invent and create new media in the context of learning, working, collaboration, and education.

5.1 *From Consumers to Active Contributors*

To create designer mindsets, one of the major roles for new media and new technologies is not to deliver predigested information to individuals, but to provide the opportunity and resources for social debate, discussion, and collaborative knowledge construction. In many human activities, learning cannot be restricted to finding knowledge that is 'out there.' If nobody knows the answer, we have to create new knowledge. The knowledge to understand, frame, and solve many problems does not exist, but is constructed and evolved during the process of solving these problems (Schön 1983). From this perspective, *access* to existing information and knowledge (often seen as the major advance of new media) is a limiting concept. Many social and technological innovations are limited to provide primarily better access, leading to 'consumer' cultures (Fischer 1998a). Other research focuses on and creates support for lifelong learning activities grounded in informed participation (Arias et al. 2000; http://www.cs.colorado.edu/~gerhard/papers/tochi99.pdf.). This goal creates the challenge to develop open, evolvable systems, allowing end-users themselves to extend their tools and information spaces, and thereby become independent (if they desire) of professionals.

5.2 *New Media in Support of Lifelong Learning*

The most promising way to provide opportunities for an active contributor mindset is to allow learners and workers to engage in design activities by creating environments supporting them in making external

Table 3
Overview of new forms of learning contributing to lifelong learning

Form	Complementing form	Contribution towards mindset creation	Major challenges	Media requirements
Self-directed learning	Prescribed learning	Authentic problems	Problem framing	Understanding evolving tasks
Learning on demand	Learning in advance	Coverage is impossible; obsolescence is guaranteed	Identifying the breakdown leading to the demand; integration of working and learning	Critics; supporting reflection-in-action
Informal learning	Formal learning	Learning by being in the world	Larger, purposive activities provide learning opportunities	End-user modifiability
Organizational learning	Individual learning	The individual human mind is limited	Shared understanding	Externalizations understandable by all stakeholders

artifacts that they can reflect upon and share with others. Over the last 10 years, design environments have been created which support the different forms of learning described in Table 3. These environments are not only used to instruct and assist novice designers, but they are able to support designers as lifelong learners. A specific example of a domain-oriented design environment is the *Envisionment and Discovery Collaboratory* (Arias et al. 2000; http://www.cs.colorado.edu/~gerhard/papers/tochi99.pdf.). This system (for details see: http://www.cs.colorado.edu/~13d/systems/EDC/) not only provides access to existing information, but allows all participants to engage in self-directed learning by supporting informed participation and collaborative knowledge construction. It combines physical and computational simulations to allow organizational learning by supporting group decision-making, mutual learning processes, and the creation of shared understanding. The incremental design and construction of an externalized world provides numerous opportunities for all involved stakeholders for self-directed learning and learning on demand.

The 'open source' movement (O'Reilly 1999) that is currently emerging as a new paradigm for software development represents an exciting perspective for a society of lifelong learners. 'Open source' approaches regard software and other cognitive artifacts not as a commodity to be consumed, but one that is collaboratively designed and constructed—providing a model for the knowledge society of the future.

6. *Evaluation*

Self-directed learning, learning on demand, informal learning, and organizational learning are fundamentally different from the traditional classroom learning dominated by curricula and tests. Evaluation of these forms of learning is an important, unresolved research topic in itself; there are no off-the-shelf assessment techniques available for these new forms of learning. These new approaches require alternatives to standard tests and the experimental methods of psychology, which measure how people perform on predetermined tasks undertaken by subjects in a laboratory or in an instructionist classroom. Approaches from anthropology, sociology and ethnography that study people 'in the wild,' as they go about their everyday activities in offices, homes, schools, laboratories, or wherever they live their lives are needed to evaluate these forms of learning. It is all too easy to collect data and statistics about those things that are easiest to identify and count or measure but have little or no connection with the objectives in which we are really interested. We have to develop evaluation techniques based on performance-based examinations or execution of large-scale projects and portfolios (Gardner 1991). It is obvious that a change

of mindsets cannot be evaluated by a test; it requires assessment of motivation, interest level, and participation in communities of learners combined with techniques for long-term longitudinal assessment.

7. Conclusions

Peter Drucker argued that 'There is nothing so useless as doing efficiently that which should not be done at all.' Adding new media and new technologies to existing practices will not change the mindsets that learners will acquire in our formal educational institutions. These mindsets are the major stumbling blocks preventing people from engaging in lifelong learning. Moving beyond the 'gift-wrapping approach' requires that we explore fundamentally new possibilities and limitations of computational media on how we think, create, work, learn, and collaborate. New tools will not just help people do cognitive jobs more easily but in the same way they used to, but they will also lead to fundamental alterations in the way problems are solved. A lifelong learning perspective requires that we change mindsets. This will include, for example, that teachers understand their roles not only as truth-tellers and oracles, but as coaches, facilitators, and mentors; and that knowledge is not presented as a commodity to be acquired or delivered, but as a human struggle to understand and as a source to deal with personally meaningful problems. The future of how we live, think, create, work, learn, and collaborate is not out there to be 'discovered'—it has to be invented and designed. Mindsets grounded in seeing learning as an important part of human lives will be an integral part of the future.

See also: Adult Cognitive Development: Post-Piagetian Perspectives; Adult Development, Psychology of; Adult Education and Training: Cognitive Aspects; Cognitive Development: Learning and Instruction; Discovery Learning, Cognitive Psychology of; Education in Old Age, Psychology of; Educational Learning Theory; Environments for Learning; Explanation-based Learning, Cognitive Psychology of; Intelligence, Prior Knowledge, and Learning; Learning from Text; Learning Theories and Educational Paradigms; Lifespan Theories of Cognitive Development; Motivation, Learning, and Instruction; Self-regulated Learning; Technology-supported Learning Environments; Transfer of Learning, Cognitive Psychology of

Bibliography

Arias E G, Eden H, Fischer G, Gorman A, Scharff E 2000 Transcending the individual human mind—creating shared understanding through collaborative design. *ACM Transactions on Computer–Human Interaction* **7**(1): 84–113

Bruner J S 1996 *The Culture of Education*. Harvard University Press, Cambridge, MA

Dohmen G 1998 *The Future of Continuing Education in Europe*. Federal Ministry of Education and Research, Bonn, Germany

Drucker P F 1994 The age of social transformation. *The Atlantic Monthly* 53–80

Fischer G 1998a Beyond 'couch potatoes': From consumers to designers. In *IEEE: Asia-Pacific Computer and Human Interaction, APCHI'98*. IEEE Computer Society, Los Alamitos, CA, pp. 2–9

Fischer G 1998b Making learning a part of life—beyond the 'gift-wrapping' approach of technology. In: Alheit P, Kammler E (eds.) *Lifelong Learning and Its Impact on Social and Regional Development*. Donat Verlag, Bremen, Germany, pp. 435–62

Fischer G 1999 Lifelong learning: Changing mindsets. In: Cumming G, Okamoto T, Gomez L (eds.) *7th International Conference on Computers in Education on 'New Human Abilities for the Networked Society' (ICCE'99, Chiba, Japan)*. IOS Press, Omaha, NE, pp. 21–30

Gardner H 1991 *The Unschooled Mind*. Basic Books, New York

Illich I 1971 *Deschooling Society*. Harper and Row, New York

O'Reilly T 1999 Lessons from open source software development. *Communications of the ACM* **42**: 32–7

Postman N 1985 *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. Penguin Books, New York

Schön D A 1983 *The Reflective Practitioner: How Professionals Think in Action*, Basic Books, New York

G. Fischer

Lifespan Development: Evolutionary Perspectives

From an evolutionary perspective, reproductive success can be understood as the driving force for specific developmental pathways which represent coherent answers to contextual demands. The sequence and structure of developmental stages across the lifespan constitute reproductive styles that promise best inclusive fitness within the respective context. Although the prolongation of the lifespan in humans compared with all other species is certainly a result of selective forces, senescence seems not to be shaped by natural selection, but rather a by-product of earlier neutral or beneficial gene expression.

1. Evolutionary Assumptions

Evolutionary theorizing considers striving for optimal reproductive success as the ultimate goal also of human development. This assumption not only specifies that the proximate regulation of physical and biological systems, but also social behavior and the complex psychology of human beings, follow a reproductive logic. Reproductive success is expressed in terms of inclusive fitness which comprises the individual's own procreation (Darwinian fitness) and the procreation of relatives with whom the individual shares genes (indirect fitness), thus establishing the

principle of kin selection (Hamilton 1964). Therefore, the evolutionary perspective focuses on the gene as the unit of analysis. However '… it is not the naked gene, that is exposed to selective forces directly' (Mayr 1994, p. 206), but rather the 'realized animal, that lives or dies, breeds or helps the relatives' (Daly and Wilson 1983, p. 32). Although individuals are considered as 'vehicles' of their genes (Dawkins 1976), the importance of the phenotypical individual for determining reproductive success has recently been (re)emphasized (Laland et al. 2000). This view has implications for the metatheoretical perspective on the nature of humans. It can be assumed that genetic 'egoism' is correlated with respective psychological characteristics, thus leading to the consequence that humans are not regarded as altruistic and cooperative *per se*, but rather as a result of—possibly implicit and non-intentional—selfish cost–benefit calculations.

However, it is not the adult individual who represents the product of evolution; rather, the whole lifespan with the different developmental phases is shaped by selective forces with the eventual exception of senescence, which itself seems not to be a product of selection. Aging is a concomitant of sexually reproducing individuals because the generational change increases the speed of evolution. The inevitability of death due to environmental stressors during human history has resulted in a decrease of individuals with increasing age, so that the power of natural selection decreased also (Medawar 1952).

Life-history theory (Alexander 1990) regards the developmental course of the individual lifespan as the result of a trade-off between investment into one's own growth and development (somatic effort) and investment into reproduction, comprising the functional systems of mating and parenting (reproductive effort; Chisholm 1996). Somatic efforts accumulate reproductive potential whereas reproductive efforts decrease the residual reproductive value. Evolution has thus shaped humans into being reproductive strategists who constantly have to make 'decisions' about the best allocation of investment (Voland 1998). The heuristics of reproduction are contingent upon environmental demands, which comprise material and ecological forces as well as social complexity (see Dunbar 1996). Especially the evolution of intelligence as a necessary prerequisite for exploring environmental resources and regulating the social behavior of expanding groups favored learning as a second asset of adaptation. Yet learning is not random, but follows (genetically imposed) biases and constraints ('central tendencies,' MacDonald 1997; 'epigenetic rules,' Lumsden and Wilson 1981) to acquiring specific environmental information during specific phases of development allowing easy learning at these points in time (Boyd and Richerson 1985).

Psychological development across the lifespan can simultaneously be interpreted as a structured sequence of developmental tasks in order to improve social and cognitive skills as well as the formation of a reproductive style. Evolutionary socialization models have described the ontogeny of individuals' reproductive styles as a meaningful passage through the developmental tasks of childhood and adolescence. In particular, the resource situation and the family climate during early childhood are considered to be predictive of attachment formation. The quality of attachment in turn influences psychological adjustment during late childhood. Psychological development also influences neurophysiological, maturational processes, leading to the somatic consequence of defining the beginning of puberty. The beginning of puberty channels the reproductive style, especially with regard to the age at which first giving birth, the spacing, and number of children with the respective implications for parental investment (Belsky et al. 1991, Keller 1996). Empirical studies from different parts of the world have supported the view of different developmental trajectories as specified by these models. The new perspective of evolutionary theorizing is thus the assumption that neurophysiological and psychological regulations of behavioral development (proximate causes) are influenced by adaptive, i.e., phylogenetically shaped, mechanisms which serve the ultimate function of optimizing reproductive success. The postreproductive phase is therefore not subject to the same degree of selection pressure as the prereproductive and reproductive life phases (Baltes 1997, Geary and Bjorklund 2000). Accordingly, old age is not a distinct phase in stage conceptualizations of the life-history approach (Bogin 1997).

In the following, the functional significance of the different developmental phases for reproductive success will be analyzed. The developmental phases of infancy, childhood, adolescence, adulthood, and old age will be briefly characterized with respect to their possible evolutionary roots. Old age is included in this analysis, since the assumption of structural continuity implies that earlier experiences shape later ones, so that the quality of old age could be regarded as a consequence of the earlier reproductive life (Keller 1991).

2. *Developmental Phases Across the Lifespan*

2.1 *Infancy*

The extreme helplessness ('altriciality') of the human newborn can be considered as part of a reproductive strategy to becoming a better adult ('better adult hypothesis,' Alexander 1990), since it allows the baby to invest all available resources in growth and development. This view is in contrast with the common interpretation of infants' helplessness as a consequence of a physiological preterm birth, due to hominid brain growth. Early development is shaped by the invest-

ment that parents allocate to each child individually, depending on their own and the child's reproductive values (e.g., age of parent, social support, sex, and birth position of child). Early experiences form the basis of the psychological adaptation to the respective physical and social environment in prompting phenotypic development. Accordingly, infants adopt internal representations of social relationships (see *Attachment Theory: Psychological* and *Adult Psychological Development: Attachment*) and a primary conception of the self, focusing on the expectations of reciprocity as the balance of give and take that can be expected in future social relationships.

2.2 Childhood

At about 2 years of age, a developmental transition with a new focus on the extension of the physical and social range of children's activities can be observed in many cultural contexts. This marks the beginning of children's apprenticeship which involves the adoption of promising strategies for acquiring resources in their particular environment. Mediated through the culturally primed beliefs and values ('ethnotheories') of their social environment, children acquire and construct their environmental and social competence through participatory learning as well as by adopting and creating developmental niches ('niche picking') which include the implicit and explicit reflection of sex differences as concomitants of adult reproductive life.

2.3 Adolescence

Adolescence constitutes the transition from childhood to adulthood. The beginning of adolescence is universally defined by entry into puberty. The duration of adolescence, however, differs substantially across cultures, varying from transitional rituals during a few weeks to a stage which last several years, depending mainly on economic constraints and educational provisions (see Schlegel and Barry 1991). The attainment of sexual maturity is a biological marker variable with a hereditary component and it is contingent upon ecological factors, such as nutritional status and physical condition, as well as previous socialization experiences, especially the security and stability of the early home environment. At the same time, this marker variable is believed to be predictive of future reproductive decisions (e.g., age at first sexual intercourse and age at first childbirth, duration of interbirth intervals). Although similar developmental trajectories for males and females have been empirically identified, sex-specific reproductive strategies with their respective costs make it likely that environmental factors influence female fecundity more strongly than male. Therefore, and owing to the greater accessibility of menarche as an empirical datum, female developmental trajectories are overrepresented in the literature. Psychosocial stress in the family of origin, especially marital conflict and father absence, has been documented as decreasing the age of menarche in different contextual systems. Particularly the divorce of parents during a child's first five years of life could reliably be related to early menarche (Chasiotis et al. 1998). The perception of this environmental scenario may support the development of a more quantitative oriented reproductive style with lesser parental investment whereas marital stability could be considered as part of a qualitative style with higher maternal and paternal investment. Thus, low ecological stress (i.e., good nutritional status and a balanced physical condition) as well as high psychosocial stress (e.g. divorce of parents before the age of five years) can result in earlier maturation, whereas high ecological stress and/or low psychosocial stress can lead to later maturation. The main reproductive function of the variability of sexual maturity is to reduce the costs of premature reproduction. If an individual 'waits' for sexual maturity, the resources can be allocated to growth which will increase the reproductive value at maturity. On the other hand, the delay might decrease the probability of surviving until the reproductive age (Charnov 1989).

The onset of reproduction varies with the duration of adolescence across cultural contexts. Especially in traditional rural societies, reproduction begins soon after sexual maturity and seems to be more quantitatively oriented, whereas reproduction in urban educated societies is delayed until the end of an educational moratorium and seems to be more qualitatively oriented.

2.4 Adulthood

In evolutionary terms, adulthood is the phase of reproduction with the challenge to resolve various allocation conflicts. First, the conflict between current and future reproduction (i.e., investment in growth or reproduction) has to be addressed, culminating in the decision of when to start reproduction, and the related issue of nepotistic effort, i.e., investing in one's own offspring or temporarily (or even permanently) adopting the role of a 'helper in the nest' (Voland 1998). The second allocation conflict concerns the question of investment in quality or quantity of offspring, which determines the number and spacing of children. Since these decisions not only have to be made generally as an expression of reproductive style, but also individually for each offspring separately, parent–child conflict is an inevitable consequence (Trivers 1974). Survival and psychological adaptation are facilitated when not only maternal, but also paternal investment is spent on the infant. It is assumed that partner

bonding and family cohesion evolved as psychological mechanisms to facilitate the development of 'better adults.' This view seems to confirm the relationship between father's absence and (increased) infant mortality in traditional societies, e.g., among the Ache Indians in South America (Hill and Hurtado 1996).

However, the determinants of reproductive success differ for males and females. Females *per se* make a higher investment owing to intrauterine pregnancy and nursing, supporting a qualitative parenting style with higher investment in fewer offspring. They have to decide whether they want to continue to look for a better mate or start reproducing. The prototypical male strategy, on the other hand, should be more quantitative, that is, to invest less in more offspring, due mainly to the fact that it is difficult for a male to know precisely whether he is the father of an infant ('paternity insecurity'). Moreover, males show greater variability in their reproductive rates, from having no children to having many children. Men have to decide between the 'cad versus dad' strategies, i.e., maximizing fertilization or parental investment (Voland 1998). The trade-off that individuals have to accept therefore entails the balance of mating and parenting efforts. Studies on partner selection have documented universal preferences for the two sexes: females prefer partners with characteristics indicative of economic success, whereas males prefer females with characteristics that indicate reproductive capacity (youth, health). In addition, there is preference by both males and females for mates who are similar to themselves in a variety of characteristics ('assortive mating'). Parental investment constitutes the intergenerational link through which the offspring experiences the socialization style that the parent has developed as a result of their own past as well as current experiences. In the case of human beings, not only time and energy as parameters of investment decisions have to be taken into consideration, but also psychological care systems which prepare the psychology of the phenotypic adult (Keller 2000).

2.5 *Old Age*

Old age covers the postgenerative phase of the lifespan. Yet it has been proven meaningful to differentiate a third young–old age span from about 60 to 75 years where the vital capacity has improved to compensate for about five years due to cultural evolution, from a fourth age span from around 80 to 100 years (cf., Baltes 1997) which has emerged due to increasing life expectancy, but where senescence has predominantly detrimental effects. These might originate as by-products from possibly two processes: (a) selective neutrality of the genome by no effect at early life and deleterious effect at later life and (b) antagonistic pleiotropic gene action that supports reproductive success during the reproductive life stages (Martin et al. 1996). One detrimental factor that has been identified across species may be metabolic oxidative damage operating in macromolecules. On the other hand, it has been argued that the extension of the lifespan serves individual fitness by supporting the development of the child's and grandchild's generation. Although this 'grandmother' hypothesis (Peccei 1995) is controversial, researchers nevertheless report that being a grandparent increases life satisfaction in old age. The physical deterioration especially of the old–old age requires extra societal cultural resources which at the same time become more difficult to be exploited effectively. Accordingly, the regulation and compensation of losses become a major mechanism for ontogenetic processes (Baltes 1991). Relating interindividual differences in regulatory competencies to the earlier reproductive life histories might be a challenge for lifespan psychology.

3. *Outlook*

Although essential questions such as the interplay of genetic and social influences on phenotypic development remain open, the evolutionary view on the human lifespan offers a stimulating perspective on many controversial issues of psychological development, e.g., the nature of sex differences or the normative character of secure attachment relationships. During their ontogenetic development, individuals, based on universal genetic programs (e.g., for language acquisition, partner preferences, parenting styles), acquire the contextually relevant ecological and social and cultural information (e.g., mortality rate of their population, economic constraints, cultural values) both intentionally and intuitively to promote their psychological development so that it is adaptive for the reproductive style that promises best inclusive fitness in the given situation. Old age challenges the cultural nature of humans in a very special way, since trade-offs are no longer associated with epigenetic rules or central tendencies. Humans therefore have to use the accumulated knowledge of the previous life stages in order to redefine and redirect the allocation of resources and balance universally evolved behavioral heuristics, ontogenetically acquired styles, and situational regulations, making each life history unique. The question of assumed ultimate causes, 'how do these behavioral regulations or these motivational systems contribute to reproductive success,' might allow us to reframe empirical findings and adapt research designs accordingly.

See also: Emotions, Evolution of; Evolution of Cognition: An Adaptationist Perspective; Evolutionary Social Psychology; Evolutionism, Including Social Darwinism; Human Cognition, Evolution of; Intelligence, Evolution of; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development; Psychological Development: Ethological and Evolutionary Approaches

Bibliography

Alexander R D 1990 Epigenetic rules and Darwinian algorithms—the adaptive study of learning and development. *Ethology and Sociobiology* **11**: 241–303

Baltes P B 1991 The many faces of human aging: Toward a psychological culture of old age. *Psychological Medicine* **21**: 837–54

Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80

Belsky J, Steinberg L, Draper P 1991 Further reflections on an evolutionary theory of socialization. *Child Development* **62**: 682–5

Bogin B 1997 Evolutionary hypotheses for human childhood. *Yearbook of Physical Anthropology* **40**: 63–89

Boyd R, Richerson P J 1985 *Culture and Evolutionary Process*. Chicago University Press, Chicago

Charnov E L 1989 Natural selection on the age of maturity in shrimp. *Evolutionary Ecology* **3**: 236–9

Chasiotis A, Scheffer D, Restemeier R, Keller H 1998 Intergenerational context discontinuity affects the onset of puberty: A comparison of parent–child dyads in West and East Germany. *Human Nature* **9**: 321–39

Chisholm J S 1996 The evolutionary ecology of attachment organization. *Human Nature* **7**: 1–38

Daly M, Wilson M 1983 *Sex, Evolution and Behavior*, 2nd edn., William Grant Press, Boston

Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford, UK

Dunbar R 1996 *Grooming, Gossip and the Evolution of Language*. Faber and Faber, London

Geary D C, Bjorklund D F 2000 Evolutionary developmental psychology. *Child Development* **71**: 57–65

Hamilton W D 1964 The genetical evolution of social behaviour (I + II). *Journal of Theoretical Biology* **7**: 1–52

Hill K, Hurtado A M 1996 *Ache Life History. The Ecology and Demography of a Foraging People*. Aldine, New York

Keller H 1991 A perspective on continuity in infant development. In: Lamb M E, Keller H (eds.) *Infant Development: Perspectives from German Speaking Countries*. Erlbaum, Hillsdale, NJ, pp. 135–50

Keller H 1996 Evolutionary approaches. In: Berry J W, Poortinga Y H, Pandey J (eds.) *Handbook of Cross-cultural Psychology, Volume 1: Theory and Method*, 2nd edn. Allyn and Bacon, Boston, pp. 215–55

Keller H 2000 Human parent–child relationships from an evolutionary perspective. *American Behavioral Scientist* **43**: 957–69

Laland K N, Odling-Smee J, Feldman M W 2000 Niche construction, biological evolution and cultural change. *Behavioral and Brain Sciences* **23**(1): 131–75

Lumsden C J, Wilson E O 1981 *Genes, Mind and Culture: The Coevolutionary Process*. Harvard University Press, Cambridge, MA

MacDonald K 1997 Life history theory and human reproductive behavior. Environmental/contextual influences and heritable variation. *Human Nature* **8**: 327–59

Martin G M, Austad S N, Johnson T E 1996 Genetic analysis of ageing: Role of oxidative damage and environmental stresses. *Nature Genetics* **13**: 25–34

Mayr E 1994 Evolution—Grundfragen und Mißverständnisse. *Ethik und Sozialwissenschaften* **5**: 203–209 [Evolution—Basic questions and misunderstandings]

Medawar P B 1952 *An Unsolved Problem of Biology*. Lewis, London

Peccei J S 1995 The origin and evolution of menopause: The altriciality–lifespan hypothesis. *Ethology and Sociobiology* **16**: 425–49

Schlegel A, Barry III H 1991 *Adolescence*. Maxwell Macmillan International, New York

Trivers R L 1974 Parent–offspring conflict. *American Zoologist* **14**: 249–64

Voland E 1998 Evolutionary ecology of human reproduction. *Annual Review of Anthropology* **27**: 347–74

H. Keller

Lifespan Development, Theory of

1. Lifespan Development, Psychological Aspects

In psychology, the study of human development from infancy until old age has been conducted under the auspices of lifespan psychology (see also *Life Course: Sociological Aspects*). In this article, first, a brief historical sketch of the development of lifespan psychology as a field is discussed; second, central constructs and theoretical tenets are presented and illustrated by empirical results; third, methodological issues related to the study of lifespan development are examined; and finally concluding commentary and future perspectives are offered.

2. Historical Background

Johann Nicolaus Tetens (1777) is considered the founder of the field of developmental psychology in Germany. From this beginning, in Germany development was studied as a lifelong phenomenon. In the USA and in other European countries, such as England, developmental psychology emerged around the turn of the nineteenth century with a strong emphasis on child development. In recent decades, however, a lifespan approach has become more prominent in North America as well, for several reasons. Interdisciplinary synergism is one reason: especially in sociology, which is related to scholars such as Bernice Neugarten or Glen Elder, life course sociology took hold as a powerful intellectual force. A second factor was the emergence of gerontology as a field of study including the search for lifelong precursors of aging. And finally, a third reason may have been the coming of age of several longitudinal studies of child development that had started in the 1920s and 1930s (Baltes et al. 1998).

The first West Virginia Conference on lifespan development—which turned into a series of conf-

erences as well as a book series titled *Life-span Development and Behavior* (Baltes and Brim, first eds.)—probably marks the emergence of lifespan psychology as a modern psychological theory and field of study (Goulet and Baltes 1970, Thomae 1959). The establishment of lifespan psychology is illustrated by the 1980 *Annual Review of Psychology* article. In the late 1990s, the continued importance of lifespan psychology is illustrated by contributions to the new *Handbook of Child Development* and another *Annual Review of Psychology* article.

3. *Lifespan Psychology: Definitions, Metatheoretical Perspectives, and Theoretical Claims*

Lifespan developmental psychology or lifespan psychology deals with the study of individual development (ontogenesis) as it extends across the entire life course. Influenced by evolutionary perspectives, neofunctionalism, and contextualism, lifespan psychology defines development as selective age-related change in adaptive capacity (Baltes 1997). In particular, this adaptive capacity involves the acquisition, maintenance, transformation, and attrition in psychological functions and structures. The focus on selection and selective adaptation highlights that development neither is a uniform nor integrated phenomenon across different domains of functioning and across time. This implies that development is conceived as a developing system comprising a multidimensional and multifunctional dynamic. Owing to selective adaptation, different parts of the developing system develop at different rates, in different directions, for different purposes, and may show continuities as well as discontinuities. A further consequence of this definition is that at any point in the lifespan, development is considered as being constituted by gains and losses. With increasing age, the proportion of gains to losses changes in favor of losses. Criteria of what constitutes a gain and what a loss can be of a subjective and objective nature. Losses, in the sense of selection and of crises, are even considered crucial motors of development (Montada et al. 1992, Riegel 1976).

Considering one domain of psychological functioning as an example may help to illustrate this terminological maze. Take for instance, the lifespan development of intellectual functioning. Two main components (multidimensionality) with very different developmental trajectories (multidirectionality) have been distinguished. The two components are the mechanics and the pragmatics of the mind or fluid and crystallized intelligence (e.g., Baltes et al. 1998). The mechanics refer to cognition as an expression of the neurophysiological architecture of the mind as it evolved during biological evolution. The cognitive mechanics are assessed by tasks of reaction speed or inferencing. In contrast, the pragmatics of cognition are associated with the bodies of knowledge available from and mediated through culture. The cognitive pragmatics are operationalized using tests of verbal ability or knowledge. While losses emerge in the mechanics of the mind quite early in development (after 25 years of age), the pragmatics are characterized first by gains and later by stability until quite late in adulthood. The mechanics and pragmatics of the mind also illustrate the notion of the developing system, as they do not operate in isolation but complement each other to support proactive, selective adaptation.

Besides multidimensionality and multidirectionality, the concepts of multifunctionality, equifinality, and multicausality are crucial when taking a dynamic-system view on lifespan development. Multifunctionality relates to the fact that one and the same developmental change can serve multiple purposes. Research on the concept of dependency provides impressive examples for how in old age, dependency not only implies the loss of autonomy but also the gain of social contact (Baltes 1996). Equifinality in turn refers to the notion that multiple roads can lead to the same developmental outcome. The example of intellectual development illustrates that one and the same behavioral outcome, such as a given cognitive performance, can be attained by using aspects of the mechanics of the mind if one is not used to that cognitive task or it can be accomplished by accessing one's experience with the task that is by referring to pragmatic aspects of the mind.

Multidimensionality and multidirectionality of development highlight the fact that within one individual there is variability in functioning across different domains. Lifespan psychology, however, is also interested in the intraindividual variability of functioning within one domain across time. This is an aspect of behavior that other subdisciplines of psychology often have devaluated as error variance. Lifespan psychology takes intraindividual variability seriously and considers it as an indicator of the plasticity of development. The notion of plasticity implies that any given developmental outcome is but one of numerous possible outcomes. The search for the conditions, range and limits as well as age-related changes in plasticity is fundamental to the study of lifespan development (Lerner 1984).

Over the years, systematic work on the concept of plasticity necessitated further differentiation. One involved the distinction between baseline reserve capacity and developmental reserve capacity (e.g., Kliegl et al. 1989). Baseline reserve capacity refers to the current level of plasticity available to individuals. For instance, how many words from a list of 20 can a person remember. Developmental reserve capacity specifies what is possible in principle given optimizing interventions. That is, how many words can a person remember after learning a mnemonic technique and practicing this technique for extended periods of time. Such training studies have found that there are impressive intellectual training gains well into old

age. Training studies comparing young and old participants, however, have also demonstrated that the training efficiency or developmental reserve capacity is much reduced in old age (Lindenberger and Baltes 1995).

Developmental reserves decline with age. But it is not only the amount of reserves that changes but also the functions that they serve. With increasing age, reserves are less used for growth and more and more for maintenance, recovery and eventually also for the management of loss (Staudinger et al. 1995).

The concepts of plasticity and reserve capacity also highlight the contextual interdependencies of development. Ontogenetic and historical contextualism is another key element of lifespan psychology (Riegel 1976). Contextualism stands in contrast to mechanist or organismic models of development. Lifespan contextualism is related to ecological–contextualist perspectives as well as action-theoretical positions that emphasize the importance of both individual and social–contextual factors in the regulation of development (Smith and Baltes 1999). According to lifespan contextualism, individuals exist in contexts that create and limit opportunities of individual pathways. But individuals also select and create their own contexts.

Contexts evolve according to at least three different logics (Baltes et al. 1980). One is the normative age-graded logic, the second is the history-graded logic, and finally third there is an idiosyncratic or non-normative logic. The age-graded logic refers to those biological and environmental aspects that, because of their dominant age correlation, shape individuals in relatively normative ways. Examples are developmental tasks such as starting school or retirement, or the age-based processes of physical maturation (puberty, menopause). The history-graded logic concerns those variations in ontogenetic development that are due to historical circumstances. Take for instance, the historical evolution of the educational system or the effect of war on ontogenetic development. Finally, the non-normative logic reflects individual–idiosyncratic events of a biological or environmental nature, such as winning the lottery or losing a leg in an accident. All three logics also interact in their shaping of ontogenetic development. In order to understand or predict development, age-graded, history-graded, and person-specific factors have to be taken into account. However, besides contributing to similarities in development, these logics, as Dannefer has argued, also contribute to systematic interindividual variations owing to, for instance, social class.

4. *Outlines of an Architecture of Human Development*

Lifespan contextualism with its three logics leads to the discussion of the two main influences shaping human development: biology and culture. It is in the theoretical conceptualization and study of these two driving developmental forces that lifespan psychology has made the most progress in recent years by formulating the overall biological and cultural architecture of lifespan development (Baltes 1997). This architecture also illustrates the interdisciplinary nature of lifespan development, and how lifespan psychology profits from advances in biology or cultural anthropology. Figure 1 illustrates the three pillars of this architecture. The specific form (level, shape) is not crucial, what is critical is the overall direction and the reciprocal relation between the three functions.

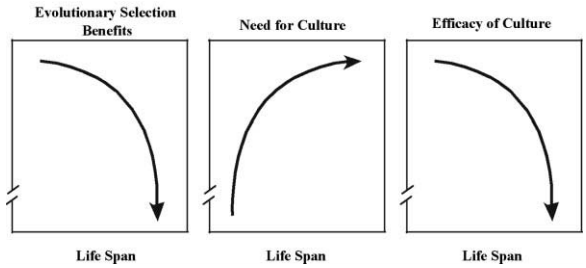

Figure 1
Schematic representation of the basic assumptions about the average dynamics between biology and culture across the life span (after Baltes 1997)

The first element of the tripartite argument derives from an evolutionary perspective. The older the organism, Finch has argued, the less the genome benefited from the genetic advantages associated with evolutionary selection. This assertion is in line with the idea that evolutionary selection has been tied to the process of reproductive fitness and its location in the first half of the life course. This general statement holds true even if indirect positive effects are carried into old age, for instance, though grand-parenting, coupling and exaptation. This age-associated diminution of evolutionary selection benefits was further affected by the fact that in earlier times few people reached old age. Thus, most people died before negative genetic attributes were activated or possible negative biological consequences of earlier developmental events became manifest. The negative consequences of the relative neglect during evolution of the second half of life is amplified by other aspects of biological aging. For instance, biological aging involves wear and tear as well as the accumulation of errors in genetic replication and repair.

The middle panel of Figure 1 summarizes the overall role of culture in lifespan development. Culture, in this context, refers to all psychological, social, material, and symbolic (knowledge-based) resources that humans have produced over the millennia. Among these cultural resources are literacy, cognitive skills, motivational dispositions, physical structures, the world of economics, as well as medical and physical technology. The age-related increase in the need for

culture is defined by two aspects. First, the farther we expect human ontogenesis to extend into old age, the more necessary it will be for particular cultural resources to emerge to make this possible. Take, for instance, the increase in life expectancy or educational status during the twentieth century. The genetic make-up did not change during this time. It is primarily through the medium of more advanced levels of culture that people are able to develop throughout longer spans of life. The second part of the argument relates to the age-related biological weakening of the body and the related need for cultural support (material, social, economic, psychological) to generate and maintain high levels of functioning.

The third pillar of the architecture of lifespan development is the decrease in efficacy of cultural resources, that is how much effect a cultural resource has. Despite the acquisition of a variety of bodies of knowledge, the efficiency of cultural factors is reduced, even though large interindividual differences exist in onset and rate of these losses in efficiency. Consider cognitive training interventions as an example of a cultural resource. Research has demonstrated that the older the adult the more time and practice it takes, for instance, to attain the same learning gains. Moreover, in some domains of cognitive functioning and at high levels of performance, older adults may not be able to reach the same level of functioning as younger adults even after extensive periods of training (Baltes et al. 1998). In other words, the efficacy of the cognitive intervention is reduced.

5. *Methodological Issues Related to Lifespan Psychology*

The search for methods adequate for the study of developmental processes across the lifespan has been a continuing part of the agenda of lifespan researchers (e.g., Lindenberger and Baltes 1995, Magnusson 1998, Nesselroade 1991). One example of methodological innovation is a strategy to examine the scope and limits of plasticity. It is a perennial question of cognitive-aging researchers, for instance, whether aging losses in cognitive performance reflect age-related experiential deficits rather than biological decline. The testing-the-limits paradigm has been developed to answer this question (Kliegl et al. 1989). The testing-the-limits paradigm compresses time by providing for high-density developmental experiences, and by doing so to identify asymptotes of performance potential. These asymptotes, obtained under putatively optimal conditions of support, are expected to estimate the upper range of developmental reserve capacity. Again, this paradigm has been most widely used in research on cognitive development and its plasticity. For instance, in the area of memory functioning, a testing-the-limit approach has been applied. Given young and old participants undergo the same intensive training procedure in a memory technique originally not known to any participant, it was found that after long periods of training and practice, older adults were not able to reach the levels of performance of young adults. These age differences have been much more pronounced than those identified in standard training studies.

Another example is the development of methods that allow the measurement of development including its interindividual differences and contextual variations. Traditionally, the main designs used in developmental psychology were cross-sectional and longitudinal methods. Both methods are characterized by serious flaws, however, when it comes to understanding development. Cross-sectional designs, in which different age groups are compared at one historical point in time, confound chronological age and generational membership. In addition, in a cross-sectional design it is impossible to track individual change trajectories. Longitudinal designs, on the other hand, while providing insight into individual change trajectories, do so only with regard to one single cohort. The challenge to track individual and historical change within one design resulted in the formulation of so-called sequential methods (for history see Baltes et al. 1998). The sequential methods combine sequences of cross-sectional studies (cross-sectional sequence) with successions of longitudinal studies (longitudinal sequence). When both sequences are applied at once exhaustive information about interindividual differences in developmental trajectories, about retest effects, as well as historical change is obtained. As Schaie has demonstrated (1996), cohort effects can be positive or negative, and can concern different age periods to different degrees.

6. *Conclusion and Future Developments*

Since its formation as a theoretical orientation for the study of human development between infancy and old age, lifespan psychology has advanced our knowledge about human development. A couple of examples may illustrate this point. For instance, we owe it to lifespan psychology that it is now known in which way and to which degree cohort effects can mask age differences (Schaie 1996), or that we have learnt about the reserves of human development, be it in terms of cognitive functioning, self and personality, or social functioning (Staudinger et al. 1995). With its orientation towards understanding developmental plasticity, lifespan psychology has opened new vistas on possibilities of intervention, and has eventually succeeded in making intraindividual variability of functioning a construct of interest to many researchers outside of the field of development (Nesselroade 1991). Finally, it has been lifespan psychology characterized by an orientation towards understanding human development as a dynamic system with functional consequences that has

started to further our understanding of the interplay between different areas of psychological functioning, such as cognition, emotion, and motivation, as well as our insight into ecologically relevant and complex constructs such as wisdom and art of life (e.g., Baltes et al. 1998, Labouvie-Vief 1994, Staudinger 1999).

Needless to say, a number of issues remain to be explored. Lifespan psychologists, for instance, have only started to integrate the study of functional aspects of human development with the investigation of microgenetic processes. Projects along these lines have been to ask which are the regulatory processes that allow the aging person to maintain comparatively high levels of well-being and fulfilling lives (e.g., Baltes and Baltes 1990, Brandtstädter and Greve 1994, Heckhausen and Schulz 1995, Staudinger et al. 1995). Within the field of cognitive aging more and more research is oriented towards understanding the microanalytic processes that may underlie the age-related decline in the cognitive mechanics.

Another issue that needs further theoretical and empirical work is the study of the ecology of human development. There is too little systematic work on classifying environments in their effects on human development. And finally the dynamic systems approach borrowed from physics has not yet been explored to its fullest when it comes to understanding human development.

Lifespan psychology does not prescribe the goals of development. Rather, it has taken on the mission to accumulate and disseminate knowledge about which processes and characteristics contribute under which circumstances to the optimization of development. Eventually, it will be this kind of knowledge that every individual may use to compose his or her life in a fulfilling manner.

See also: Developmental Psychology; Lifespan Development: Evolutionary Perspectives; Lifespan Theories of Cognitive Development; Plasticity in Human Behavior across the Lifespan

Bibliography

Baltes M M 1996 *The Many Faces of Dependency in Old Age*. Cambridge University Press, New York

Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80

Baltes P B, Baltes M M 1990 Psychological perspectives on successful aging: The model of selective optimization with compensation. In: Baltes P B, Baltes M M (eds.) *Successful Aging: Perspectives from the Behavioral Sciences*. Cambridge University Press, New York, pp. 1–34

Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn., Wiley, New York, Vol. 1, pp. 1029–1143

Brandtstädter J, Greve W 1994 The aging self: Stabilizing and protective processes. *Developmental Review* **14**: 52–80

Goulet L R, Baltes P B (eds.) 1970 *Life-span Developmental Psychology: Research and Theory*. Academic Press, New York

Heckhausen J, Schulz R 1995 A life-span theory of control. *Psychological Review* **102**: 284–304

Kliegl R, Smith J, Baltes P B 1989 Testing-the-limits and the study of adult age differences in cognitive plasticity of a mnemonic skill. *Developmental Psychology* **25**: 247–256

Labouvie-Vief G 1994 *Psyche and Eros: Mind and Gender in the Life Course*. Cambridge University Press, New York

Lerner R M 1984 *On the Nature of Human Plasticity*. Cambridge University Press, New York

Lindenberger U, Baltes P B 1995 Testing-the-limits and experimental simulation: Two methods to explicate the role of learning in development. *Human Development* **38**: 349–60

Magnusson D 1998 The logic and implications of a person-oriented approach. In: Cairns R B, Berman L R, Kagan J (eds.) *Methods and Models for Studying the Individual*. Sage, Thousands Oaks, CA, pp. 33–64

Montada L, Sigrin-Heide F, Lerner M-J (eds.) 1992 *Life Crises and Experiences of Loss in Adulthood*. L. Erlbaum Associates, Hillsdale, NJ

Nesselroade J R 1991 Interindividual differences in intraindividual change. In: Collins L M, Horn J L (eds.) *Best Methods for the Analysis of Change*, 1st edn. American Psychological Association, Washington, DC, pp. 92–105

Riegel K F 1976 The dialectics of human development. *American Psychologist* **31**: 689–700

Schaie K W 1996 *Intellectual Development in Adulthood: The Seattle Longitudinal Study*. Cambridge University Press, New York

Smith J, Baltes P B 1999 Life-span perspectives on development. In: Bornstein M H, Lamb M E (eds.) *Developmental Psychology: An Advanced Textbook*, 4th edn. L. Erlbaum Associates, Mahwah, NJ, pp. 47–72

Staudinger U M 1999 Social cognition and a psychological approach to an art of life. In: Hess T M, Blanchard-Fields F (eds.) *Social Cognition and Aging*. Academic Press, New York, pp. 343–75

Staudinger U M, Marsiske M, Baltes P B 1995 Resilience and reserve capacity in later adulthood: Potentials and limits of development across the life span. In: Cicchetti D, Cohen D J (eds.) *Developmental Psychopathology: Risk, Disorder, and Adaptation*. Wiley, New York, Vol. 2, pp. 801–47

Thomae H 1959 *Entwicklungspsychologie* [Developmental psychology]. Verlag für Psychologie, Göttingen, Germany

U. M. Staudinger

Lifespan Theories of Cognitive Development

Most lifespan theories of cognitive development (LTCD) provide an integrative framework for the description, explanation, and optimization of cognitive development across different functional domains and age periods. Arguably the first LTCD dates back to Johann Nicolaus Tetens (1736–1807), a philosopher and psychologist of the Enlightenment

era. Tetens (1777) noted that well-trained skills are less likely to decline with advancing adult age than the basic abilities underlying their acquisition, and this observation and conceptual distinctions related to it have remained at the heart of LTCD.

1. Central Assumptions of LTCD

Most LTCD are based on two central assumptions (Baltes et al. 1998):

(a) Cognitive development reflects the operation of two intertwined components, one biological and the other cultural. The specific definition and terminology regarding the two components varies across different LTCD (see Sect. 3). In general, the biological component is construed as an expression of the neurophysiological architecture of the mind as it evolved during biological evolution and unfolds during ontogeny. In contrast, the cultural component refers to bodies of knowledge available from and mediated through culture.

(b) During ontogeny, potential related to the biological component is *invested into* various cultural domains, thereby leading to the acquisition of culturally transmitted bodies of knowledge. A good example is the acquisition of reading and writing skills. Hence, at any point in time during development, two types of cognitive capacities can be distinguished: (i) the capacity to invest (i.e., to acquire new knowledge of various sorts) and (ii) the capacity to think and act on the basis of acquired knowledge.

2. Theoretical Background: the Lifespan Dynamics Between Biology and Culture

Cognitive development participates in the general lifespan dynamics between biology and culture (see also *Lifespan Development*, *Theory of*). According to Baltes (1997), these dynamics can be characterized by three interrelated lifespan trends: (a) a decrease in evolutionary selection benefits; (b) a concomitant increase in the need for culture; and (c) a decrease in the efficacy of culture in reaching adaptive outcomes.

2.1 Decrease in Evolutionary Selection Benefits

This proposition speaks to the biological component of cognition, and derives from an evolutionary perspective on the nature of the genome and its age-correlated changes in expression and biological potential. During evolution, the older the organism, the less the genome benefited from the genetic advantages associated with evolutionary selection. After maturity, and with age, the expressions and mechanisms of the genome lose in functional quality. This holds true even though some indirect positive evolutionary selection benefits may be carried into old age (e.g., through grandparenting). The age-associated diminution of evolutionary selection benefits and its implied association with an age-related loss of biological potential is further accentuated by the fact that fewer people reached old age in earlier times.

2.2 Increase in Need for Culture

The second proposition summarizes the overall perspective on lifespan development associated with culture and culture-based processes. Among these cultural resources are physical structures, the world of economics as well as that of medical and physical technology, but also cognitive skills such as formal logic, literacy, and written documents. The argument for an age-related increase in the 'need' for culture has two parts. First, for human ontogeny to have reached increasingly high levels of functioning across historical time, whether in physical or psychological domains, there had to be a conjoint evolutionary increase in the richness and dissemination of the resources and 'opportunities' of culture. The second argument for the proposition relates to the biological weakening associated with age. That is, the older we are, the more we need culture-based resources to generate and maintain high levels of functioning. For instance, in old age, it generally takes more time and practice to attain the same amount of learning gains than in early adulthood (see next).

2.3 Decrease in the Efficiency of Culture

The efficiency of cultural factors and resources in producing desired outcomes decreases with age, especially during the second half of life. There are three major causes for this age-related reduction in cultural efficiency. The first is age-related loss in biological potential. The second is related to the shape of *learning curves*, which express the process of knowledge acquisition in specified domains. Such curves generally start out steep and then reach an asymptote. Given that older individuals, on average, are more likely to find themselves in later phases of learning than younger individuals (i.e., closer to the asymptote), increasing amounts of effort and technology are needed to produce further gains. The third cause refers to negative consequences of earlier learning on later learning (i.e., *negative transfer*).

2.4 Consequences for LTCD

In combination, the two core assumptions of LTCD and the general lifespan dynamics between biology and culture predict the presence of multidirectionality in cognitive lifespan development: the biological component of cognition is expected to decline after maturity, reflecting the decrease in evolutionary selec-

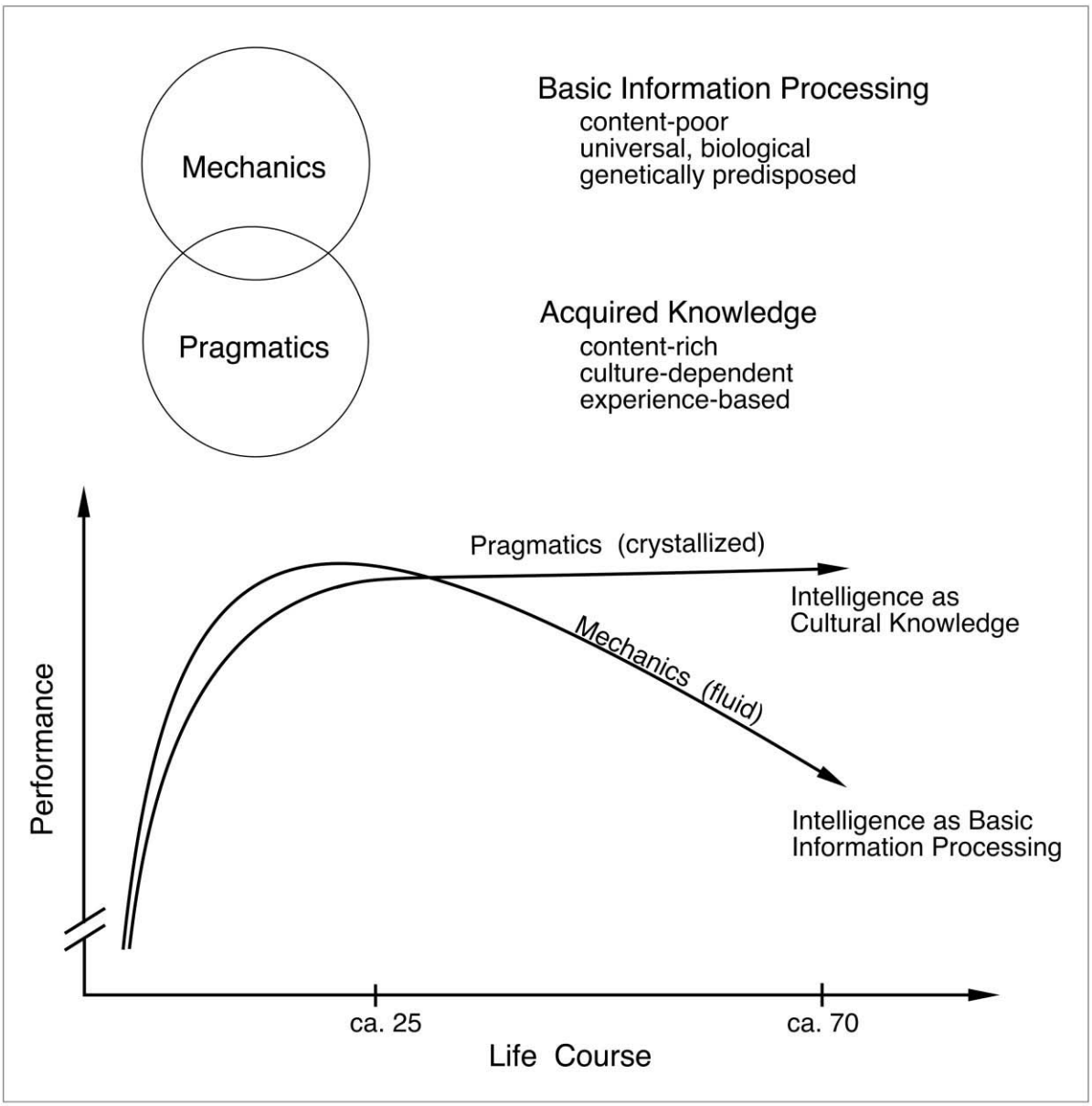

Figure 1
Lifespan theories of cognitive development posit two-component models of cognition. The top section defines the categories, the bottom section illustrates postulated lifespan trajectories (after Baltes et al. 1998; cf., Hebb 1949, Cattell 1971). In advanced old age, the trajectories show less and less differentiation (Lindenberger and Baltes 1997)

tion pressures with age. In contrast, the cultural component of cognition is expected to increase with age as long as knowledge maintenance and knowledge acquisition outweigh age-based losses in biological potential (see Fig. 1).

Moreover, any given cognitive performance can be portrayed as the combined outcome of biological and cultural systems of influence (Lindenberger and Baltes 1997). Age differences in skilled chess performance (Charness and Bosman 1990) illustrate this point. The mean age at which a world championship is first won is about 46 years for correspondence chess but about 30 years for tournament chess. In correspondence chess, players are permitted three days to deliberate a

move; in tournament chess, deliberation averages three minutes per move. Conceivably, the difference in peak age reflects differences in the relative importance of task-relevant mechanisms with a strong biological component, such as speed of perception and working memory, and mechanisms with a strong cultural component, such as procedural and declarative knowledge about chess.

3. *Variation in Terminology and Scope Among LTCD*

3.1 *Two-component LTCD*

Two-component LTCD vary in terminology and scope (Baltes et al. 1998). As mentioned, Tetens (1777) was the first to propose an LTCD. His distinction between biology-based *absolute capacities* and culture-based *relative capacities* anticipated central ideas and findings regarding lifespan cognitive development, such as the investment relation and the divergence in age gradients between the two components.

Within the psychometric research tradition, the most influential LTCD is the theory of *fluid* and *crystallized intelligence* (*Gf–Gc theory*) introduced by Raymond B. Cattell (see *Cattell, Raymond Bernard (1905–98)*) and modified by John Horn (Cattell 1971, Horn 1989). Fluid intelligence represents the biological component; it is called fluid because it can be invested into various cultural domains. In contrast, crystallized intelligence represents the cultural component; it is called crystallized because it has solidified into knowledge. Gf–Gc theory summarizes a vast amount of empirical evidence obtained through *standardized testing* and *factor analysis*.

More recently, Paul Baltes (1987) also proposed a model with two components, the biology-based *mechanics* and the culture-based *pragmatics* of cognition. In comparison with the Gf–Gc distinction, the two-component model introduced by Baltes is broader in scope. For instance, in addition to the emphasis on individual differences and measurement precision inherited from Gf–Gc theory, the model seeks to integrate findings from cognitive psychology, evolutionary psychology, cultural psychology, cognitive neurosciences, and developmental biology (Baltes et al. 1998). Hebb's (1949) distinction between intelligence A (intellectual power) and intelligence B (intellectual products) resembles the mechanics–pragmatic distinction. Additional LTCD are mentioned in Baltes et al. (1998).

3.2 *Structuralist (NeoPiagetian) LTCD*

Two-component LTCD are functionalist in character. For instance, they emphasize the local nature of developmental adaptations during knowledge acquisition, selective specialization, and transfer. In contrast, *structuralist* LTCD posit that lifespan intellectual development follows a stage-like movement towards higher (more mature) forms of reasoning and thought (Labouvie-Vief 1982, Blanchard-Fields 1989). Building upon the theory of cognitive development formulated by Jean Piget, structuralist LTCD search for the emergence of one or more 'post-formal' or 'dialectical' stages of cognitive development after the advent of formal operations, the highest stage in Piaget's theory (see *Piaget, Jean (1896–1980)*). The conceptual description of these post-formal stages often connects personality development (e.g., generativity in the *Eriksonian* sense) with logical considerations (e.g., awareness and acceptance of contradiction). As a consequence of this particular linkage, the emergence of post-formal stages is expected to be accompanied by increments in reflexivity and general awareness for the human condition.

Generally, two-component, functionalist LTCD seem more amenable to operational definition and empirical investigation than structuralist LTCD (Baltes et al. 1998). For instance, from a functionalist perspective, the search for higher forms of reasoning can be redirected into a search for bodies of acquired knowledge with a high degree of generality and meaning.

The following sections focus on two-component LTCD. First, the mechanics–pragmatics distinction is elaborated in some detail. Select findings and hypotheses will be summarized thereafter.

4. *Two-component Models of Cognition: the Mechanics–Pragmatics Distinction*

4.1 *The Mechanics of Cognition*

In the mechanics of cognition, the predominant age-graded ontogenetic pattern is one of maturation, stability, and aging-induced decline. Age-based changes in this component early and late in ontogeny are assumed to be strongly influenced by genetic and other brain status-related factors, albeit in fundamentally different ways. Early in ontogeny (i.e., during embryogenesis, infancy, and early childhood), age-based changes in the mechanics primarily reflect the unfolding and active construction of more or less domain-specific and predisposed processing capabilities (Wellman and Gelman 1992). In contrast, negative changes in the mechanics of cognition late in life result from brain-related consequences of less effective phylogenetic selection pressures operating during this period.

The cognitive mechanics, then, reflect fundamental organizational properties of the central nervous system. In terms of psychological operations, these

properties are indexed by the speed, accuracy, and coordination of elementary processing operations as assessed in tasks measuring the quality of information input, sensory and motor memory, discrimination, categorization, and selective attention, as well as reasoning ability in highly overlearned or novel domains.

4.2 The Pragmatics of Cognition

In contrast to the mechanics, the pragmatics of the mind reveal the power of human agency and culture. Developmental changes in this component reflect the acquisition of culturally transmitted bodies of *declarative* and *procedural knowledge* that are made available to individuals in the course of *socialization*. Some of these socialization events are normative but specific to certain cultures (e.g., formal schooling), others are more universal (e.g., mentoring), and still others are idiosyncratic or person-specific (e.g., specialized ecological and professional knowledge). In all cases, the corresponding bodies of knowledge are represented both internally (e.g., *semantic networks*) and externally (e.g., books). Pragmatic knowledge systems are acquired during ontogeny but may build on evolutionarily prestructured, domain-specific knowledge (Siegler and Crowley 1994).

Compared with the mechanics, the pragmatics of cognition direct the attention of lifespan researchers towards the increasing significance of acquired knowledge in cognitive behavior (Ericsson and Lehmann 1996). Typical examples include reading and writing skills, educational qualifications, professional *expertise*, and varieties of everyday problem-solving, but also knowledge about the self and the meaning and conduct of life (Staudinger and Pasupathi 2000) (see *Wisdom, Psychology of*).

4.3 The Relation Between Mechanics and Pragmatics

A common misunderstanding about two-component LTCD is that biological and cultural components are assumed to operate separately. As illustrated by the intersecting circles in Fig. 1, LTCD recognize that cognitive behavior reflects the joint operation of biological and cultural systems of influence, and posits that biological (mechanic) and cultural (pragmatic) development interact in complex and changing ways throughout life. During infancy and early childhood, the unfolding of the mechanics is dependent upon the acquisition of pragmatic knowledge. For instance, the ability of newborns to prefer the language spoken by their mothers during pregnancy over other languages has a strong biological component, but this component needs cultural specification (e.g., one specific language is preferred over another). Later in life, the acquisition and maintenance of pragmatic knowledge increasingly serve the function to buffer (circumvent) the negative consequences of mechanic decline. Moreover, at least in some cases, pragmatic knowledge is acquired as a compensatory reaction to losses in the mechanics (Freund et al. 1999). In agreement with Sect. 2, the buffering and compensatory functions of pragmatic knowledge increase in importance but decrease in efficiency with advancing age.

Another misunderstanding about the two-component model concerns its relation to the issue of *domain generality* versus *domain specificity*; that is, to the question whether a given set of cognitive mechanisms applies to a large or a small number of task domains. Here, the mechanics are often identified with *domain generality* in cognitive development, and the pragmatics with *domain specificity*. At first sight, this seems reasonable because acquired bodies of knowledge tend to be specialized (e.g., knowledge about chess), whereas mechanisms associated with basic learning capabilities tend to generalize across domains (e.g., the *law of reinforcement*). However, at closer inspection, the mapping is not so clear. First, certain components of the mechanics are dedicated to, or at least particularly important for, specific domains (e.g., infants' preference for human faces). Second, some bodies of acquired knowledge, such as *knowledge about the fundamental pragmatics of life* (with *wisdom* being its highest expression; see *Wisdom, Psychology of*), are general in the sense that their range of applicability is very broad. Therefore, the distinction between domain generality and domain specificity is correlated but not identical with the mechanics–pragmatics distinction.

5. Select Findings and Hypotheses Associated with LTCD

LTCD attempt to integrate and generate research findings and hypotheses across different content domains and age periods. A few central examples may suffice for illustration; more can be found in Baltes et al. (1998).

5.1 Divergence of Mechanic and Pragmatic Age Gradients

The most longstanding evidence in support of the mechanics–pragmatics distinctions is the difference between maintained and vulnerable intellectual abilities observed in lifespan studies of psychometric intelligence (Jones and Conrad 1933). Abilities that critically involve the mechanics, such as reasoning, memory, spatial orientation, and perceptual speed, generally show a pattern of monotonic and roughly linear decline during adulthood, with some further acceleration of decline in very old age. In contrast,

predominantly pragmatic abilities, such as verbal knowledge and certain facets of numerical ability, remain stable or increase up to the sixth or seventh decade of life. They only start to evince some decline in very old age, when the mechanics fall below a certain threshold in most individuals.

5.2 Differentiation–Reintegration Hypothesis of Interindividual Differences in Intellectual Functioning Across the Lifespan

As a general rule, the degree of covariation across different intellectual abilities, often represented as the prominence of a general factor (*g*) of intellectual functioning, appears to decrease from childhood to middle adulthood, and to increase again in very old age (Baltes and Lindenberger 1997, Lindenberger and Baltes 1997). This empirical regularity is well captured by assuming curvilinear age-based changes in the importance of domain-general mechanic constraints on cognitive performance, as formulated by the *differentiation–reintegration hypothesis* of intellectual lifespan development (for its origins, see Baltes et al. 1998; also *Intelligence: Central Conceptions and Psychometric Models*).

5.3 Malleability and Plasticity

A central theme in *lifespan psychology* concerns the malleability, and especially optimization, of psychological functioning and behavior throughout all phases of life. The effects of cohort, historical period, and environmental change on age changes and age differences in cognitive performance are one indication of malleability (Schaie 1996). Experimental *testing-the-limits* procedures offer another, more direct way to explore age differences in upper limits of performance potential (Baltes and Kliegl 1992). In adulthood, cognitive plasticity continues to be present in cognitively healthy older adults, albeit to a lesser degree than in young adults. In fact, when pushed to the upper limits of performance in tasks which critically involve the mechanics, the performance distributions of young and old adults are close to non-overlapping (Baltes and Kliegl 1992). For the most part, cognitive plasticity reflects the acquisition of novel cognitive tools (e.g., the investment of existing mechanic potential), rather than an amelioration of the mechanic component itself.

5.4 Search for Basic Determinants of Mechanic Development

Attempts have been made to identify domain-general mechanisms acting as pacemakers for lifespan changes in the mechanics of cognition. Among the most prominent constructs are *information processing rate*, *working memory capacity*, and the ability to suppress irrelevant information (*inhibition*). Supportive evidence has been gathered for each of the three constructs at both ends of the lifespan, especially with respect to information processing rate (Salthouse 1996). At the same time, all three constructs suffer from a lack of direct evidence at the physiological level and from difficulties in making differential predictions. A closer link to cognitive neurosciences and formal modeling approaches is expected to foster progress in this area of research (Li and Lindenberger 1999).

5.5 Relative (e.g., Rank Order) Stability and Level Stability in Intellectual Functioning Across the Lifespan

Generally, relative stabilities computed after identical lapses of time show a strong increase in magnitude from middle childhood to adolescence into middle adulthood and early old age (Hertzog and Schaie 1986), followed by a probable decrease in old and very old age. According to LTCD, these age-based changes in relative stability need to be interpreted in conjunction with age-based changes in level (Molenaar et al. 1991). The specific prediction is that interindividual differences change more rapidly early in development because the intellectual repertoire is smaller but growing faster than at later points in ontogeny, thereby giving room for larger amounts of new variance per unit time (both environmental and genetic). Similarly, the preponderance of age-based losses in very old age may not only lead to decrements in level but also to a reshuffling of individual differences. Both predictions specify relations between changes in covariance and changes in mean level over time, and call for the use of complex statistical modeling techniques. In this respect, they point to the more general phenomenon of a close connection between conceptual and methodological advances in the context of LTCD.

6. Conclusion

Compared with the large number and great diversity of cognitive-developmental theories about specific age periods (e.g., infancy), content domains (e.g., faces), and processes (e.g., recognition), LTCD form a small and relatively homogeneous set. One characteristic feature of this set is to articulate disparate bodies of knowledge about cognitive development with general propositions about the overall architecture of human ontogeny. The resulting bird's eye view on cognitive lifespan development stresses the variability and modifiability of human development at the inter- and intraindividual level, but it also stresses its invariant, age-dependent properties. The future role played by LTCD will depend on their utility for scholars working on more specific aspects of cognitive development, and on their ability to link the field of cognitive lifespan development to neighboring fields and disciplines.

See also: Adult Cognitive Development: Post-Piagetian Perspectives; Aging Mind: Facets and Levels of Analysis; Aging, Theories of; Cattell, Raymond Bernard (1905–98); Cognitive Aging; Education and Learning: Lifespan Perspectives; Lifespan Development, Theory of; Memory and Aging, Cognitive Psychology of; Wisdom, Psychology of

Bibliography

Baltes P B 1987 Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* **23**: 611–26

Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80

Baltes P B, Kliegl R 1992 Further testing of limits of cognitive plasticity: Negative age differences in a mnemonic skill are robust. *Developmental Psychology* **28**: 121–5

Baltes P B, Lindenberger U 1997 Emergence of a powerful connection between sensory and cognitive functions across the adult life span: A new window to the study of cognitive aging? *Psychology and Aging*. **12**: 12–21

Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology: Vol. 1. Theoretical Models of Human Development*, 5th edn. Wiley, New York, pp. 1029–143

Blanchard-Fields F 1989 Postformal reasoning in a socio-emotional context. In: Commons M L, Sinnott J D, Richards F A, Armon C (eds.) *Adult Development: Comparisons and Applications of Developmental Methods*. Praeger, New York, pp. 77–93

Cattell R B 1971 *Abilities: Their Structure Growth and Action*. Houghton Mifflin, Boston

Charness N, Bosman E A 1990 Expertise and aging: life in the lab. In: Hess T H (ed.) *Aging and Cognition: Knowledge Organization and Utilization*. Elsevier, Amsterdam, pp. 343–85

Ericsson K A, Lehmann A C 1996 Expert and exceptional performance: Evidence of maximal adaption to task constraints. *Annual Review of Psychology* **47**: 273–305

Freund A M, Li K Z H, Baltes P B 1999 Successful development and aging: the role of selection, optimization, and compensation. In: Brandtstädter J, Lerner R M (eds.) *Action and Self Development: Theory and Research Through the Life Span*. Sage, Thousand Oaks, CA, pp. 401–34

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Hertzog C, Schaie K W 1986 Stability and change in adult intelligence: 1. Analysis of longitudinal covariance structures. *Psychology and Aging* **1**: 159–71

Horn J L 1989 Models of intelligence. In: Linn R L (ed.) *Intelligence: Measurement Theory and Public Policy*. University of Illinois Press, Urbana, IL, pp. 29–73

Jones H E, Conrad H 1933 The growth and decline of intelligence: a study of a homogeneous group between the ages of ten and sixty. *Genetic Psychological Monographs* **13**: 223–98

Labouvie-Vief G 1982 Dynamic development and mature autonomy: a theoretical prologue. *Human Development* **25**: 161–91

Li S-C, Linderberger U 1999 Cross-level unification: A computational exploration of the link between deterioration of neurotransmitter systems and dedifferentiation of cognitive abilities in old age. In: Nilsson L-G, Markowitsch H (eds.) *Cognitive Neuroscience of Memory*. Hogrefe and Huber, Toronto, pp. 104–46

Lindenberger U, Baltes P B 1997 Intellectual functioning in old and very old age: Cross-sectional results from the Berlin Aging Study. *Psychology and Aging* **12**: 410–32

Molenaar P C M, Boomsma D I, Dolan C V 1991 Genetic and environmental factors in a developmental perspective. In: Magnusson D, Bergman L, Rudinger G, Törestad B (eds.) *Problems and Methods in Longitudinal Research: Stability and Change*. Cambridge University Press, Cambridge, UK, pp. 250–73

Salthouse T A 1996 The processing-speed theory of adult age differences in cognition. *Psychological Review* **103**: 403–28

Schaie K W 1996 *Adult Intellectual Development: the Seattle Longitudinal Study*. Cambridge University Press, New York

Siegler R S, Crowley K 1994 Constraints on learning in nonprivileged domains. *Cognitive Psychology* **27**: 194–226

Staudinger U M S, Pasupathi M 2000 Life-span perspectives on self, personality, and social cognition. In: Craik F I M, Salthouse T A (eds.) *Handbook of Aging and Cognition*, 2nd edn. Erlbaum, Mahwah, NJ, pp. 633–88

Tetens J N 1777 *Philosophische Versuche über die menschliche Natur und ihre Entwickelung*. Weidmanns Erben und Reich, Leipzig

Wellman H M, Gelman S A 1992 Cognitive development: foundational theories of core domains. *Annual Review of Psychology* **43**: 337–75

U. Lindenberger

Likelihood in Statistics

A statistical model for phenomena in the sciences or social sciences is a mathematical construct which associates a *probability* with each of the possible outcomes. If the data are discrete, such as the numbers of people falling into various classes, the model will be a discrete probability distribution, but if the data consist of measurements or other numbers that may take any values in a continuum, the model will be a continuous probability distribution. When two different models, or perhaps two variants of the same model differing only in the value of some adjustable parameter(s), are to be compared as explanations for the same observed outcome, the probability of obtaining this particular outcome can be calculated for each and is then known as the *likelihood* for the model or parameter value(s) given the data. This article enlarges on the definition of likelihood, discusses its properties and uses, and relates it to other approaches to statistical inference.

1. Probability and Likelihood Distinguished

Probabilities and likelihood are easily (and frequently) confused, and it is for this reason that in 1921 R.A. Fisher introduced the new word: 'What we can

find from a sample is the likelihood of any particular value of [the parameter], if we define the likelihood as a quantity proportional to the probability that, from a population having that particular value, the [observed sample] should be obtained. So defined, probability and likelihood are quantities of an entirely different nature' (see *Fisher, Ronald A (1890–1962)*).

The first difference to be noted is that the variable quantity in a likelihood statement is the hypothesis (a word which conveniently covers both the case of a model and of particular parameter values in a single model), the outcome being that actually observed, in contrast to a probability statement, which refers to a variety of outcomes, the hypothesis being assumed and fixed. Thus a manufacturer of dice may reasonably assert that the outcomes 1, 2, 3, 4, 5, 6 of a throw each have 'probability' 1/6 on the hypothesis that his dice are well-balanced, whilst an inspector of casinos charged with testing a particular die will wish to compute the 'likelihood' for various hypotheses about these probabilities on the basis of data from actual tosses.

The second difference arises directly from the first. If all the outcomes of a statistical model are considered their total probability will be 1, since one of them must occur and they are mutually exclusive; but since in general hypotheses are not exhaustive—one can usually think of another one—it is not to be expected that the sum of likelihoods has any particular meaning, and indeed there is no addition law for likelihoods corresponding to the addition law for probabilities. It follows that only 'relative' likelihoods are informative, which is the reason for Fisher's use of the word 'proportional' in his original definition.

2. *The Likelihood Function*

The most important application of likelihood is in parametric statistical models. Consider the simplest binomial example, such as that of the distribution of the number of boys r in families of size n (an example which has played an important role in the development of statistical theory since the early eighteenth century). The probability of getting exactly r boys will be given by the binomial distribution indexed by a parameter p, the probability of a male birth. Denote this probability of r boys by $\mathrm{P}(r|p)$, n being assumed fixed and of no statistical interest. Then we write

$$\mathrm{L}(p \parallel r) \propto \mathrm{P}(r|p)$$

for the likelihood of p given the particular value r, the double vertical line $\parallel$ being used to indicate that the likelihood of p is not conditional on r in the technical probability sense. In this binomial example $\mathrm{L}(p\|r)$ is a continuous function of the parameter p and is known as the 'likelihood function.' When only two hypotheses are compared, such as two particular values of p in the present example, the ratio of their likelihoods is known as the 'likelihood ratio.'

The value of p which maximizes $\mathrm{L}(p\|r)$ for an observed r is known as the 'maximum-likelihood estimate' of p and is denoted by p^; expressed in general form as a function of r it is known as the 'maximum-likelihood estimator.' Since the pioneering work of Fisher in the 1920s it has been known that maximum-likelihood estimators possess certain desirable properties under repeated sampling (consistency and asymptotic efficiency, and in an important class of models sufficiency and full efficiency). For this reason they have come to occupy a central position in repeated-sampling (or 'frequentist') theories of statistical inference (see *Estimation: Point and Interval*). However, partly as a reaction to the unsatisfactory features which repeated-sampling theories display when used as theories of evidence, coupled with a reluctance to embrace the full-blown Bayesian theory of statistical inference, likelihood is increasingly seen as a fundamental concept enabling hypotheses and parameter values to be compared directly.

3. *The Likelihood Principle*

The basic notion, championed by Fisher as early as 1912 whilst still an undergraduate at Cambridge but now known to have been occasionally suggested by other writers even earlier, is that the likelihood ratio for two hypotheses or parameter values is to be interpreted as the degree to which the data support the one hypothesis against the other. Thus a likelihood ratio of 1 corresponds to indifference between the hypotheses on the basis of the evidence in the data, whilst the maximum-likelihood value of a parameter is regarded as the 'best-supported value', other values being ranked by their lesser likelihood accordingly. This was formalized as the *Law of Likelihood* by Ian Hacking (1965). Fisher's final advocacy of the direct use of likelihood will be found in his last book *Statistical Methods and Scientific Inference* (1956).

Such an approach, unsupported by any appeal to repeated-sampling criteria, is ultimately dependent on the primitive notion that the best hypothesis or parameter-value on the evidence of the data is the one which would explain what has in fact been observed with the highest probability. The strong intuitive appeal of this can be captured by recognizing that it is the value which would lead, on repeated sampling, to a precise repeat of the data with the least expected delay. In this sense it offers the best statistical explanation of the data.

In addition to specifying that relative likelihoods measure degrees of support, the likelihood approach requires us to accept that the likelihood function or ratio contains all the information we can extract from

the data about the hypotheses in question on the assumption of the specified statistical model—the so-called 'Likelihood Principle.' It is important to include the qualification requiring the specification of the model, first because the adoption of a different model might prove necessary later and second because in some cases the structure of the model enables inferences to be made in terms of 'fiducial probability' which, though dependent on the likelihood, are stronger, possessing repeated-sampling properties which enable confidence intervals to be constructed (see *Fiducial and Structural Statistical Inference*).

Though it would be odd to accept the 'Law of Likelihood' and not the 'Likelihood Principle', Bayesians necessarily accept the Principle but not the Law, for although the likelihood is an intrinsic component of Bayes's Theorem, Bayesians deny that a likelihood function or ratio has any meaning in isolation (see *Bayesian Statistics*). For those who accept both the Law and the Principle it is convenient to express the two together as:

> *The Likelihood Axiom*: Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data, and the likelihood ratio is to be interpreted as the degree to which the data support the one hypothesis against the other (Edwards 1972).

The likelihood approach has many advantages apart from its intuitive appeal. It is straightforward to apply because the likelihood function is usually simple to obtain analytically or easy to compute and display. It leads directly to the important theoretical concept of 'sufficiency' according to which the function of the data that is the argument of the likelihood function itself carries the information. This reduction of the data is often a simple 'statistic' such as the sample mean (see *Statistical Sufficiency*). Moreover, the approach illuminates many of the controversies surrounding repeated-sampling theories of inference, especially those concerned with ancillarity and conditioning. Birnbaum (1962) argued that it was possible to derive the 'Likelihood Principle' from the concepts of sufficiency and conditionality, but to most people the Principle itself seems the more primitive concept and the fact that it leads to notions of sufficiency and conditioning seems an added reason for accepting it.

Likelihoods are multiplicative over independent data sets referring to the same hypotheses or parameters, facilitating the combination of information. For this reason 'log-likelihood' is often preferred because information is then combined by addition. In the field of genetics, where likelihood theory is widely applied, the log-likelihood with the logarithms taken to the base 10 is known as a LOD, but for general use natural logarithms to the base e are to be preferred, in which case log-likelihood is sometimes called 'support.' Most importantly, the likelihood approach is compatible with Bayesian statistical inference in the sense that the posterior Bayes distribution for a parameter is, by Bayes's Theorem, found by multiplying the prior distribution by the likelihood function. Thus when, in accordance with Bayesian principles, a parameter can itself be given a probability distribution (and this assumption is the Achilles' heel of Bayesian inference) all the information the data contain about the parameter is transmitted via the likelihood function in accordance with the 'Likelihood Principle.' It is indeed difficult to see why the medium through which such information is conveyed should depend on the purely external question of whether the parameter may be considered to have a probability distribution, and this is another powerful argument in favor of the Principle itself.

4. *Parameter Estimation*

In the case of a single parameter the likelihood function or the log-likelihood function may easily be drawn, and, if it is unimodal, limits may be assigned to the parameter, analogous to the confidence limits of repeated-sampling theory. Calling the log-likelihood the 'support', m-unit support limits are the two-parameter values astride the maximum at which the support is m units less than at the maximum. For the simplest case of estimating the mean of a normal distribution of known variance the 2-unit support limits correspond closely to the 95 percent confidence limits which are at ± 1.96 standard errors. In this normal case the support function is quadratic and may therefore be characterized completely by the maximum-likelihood estimate and the curvature at the maximum (the reciprocal of the radius of curvature) which is defined as the 'observed information.' Comparable definitions apply in multiparameter cases, leading to the concept of an m-unit support region and an observed information matrix. In cases in which the support function is not even approximately quadratic the above approach may still be applied if a suitable transformation of the parameter space can be found.

The representation of a support function for more than two parameters naturally encounters the usual difficulties associated with the visualization of high-dimensioned spaces, and a variety of methods have been suggested to circumvent the problem. It will often be the case that information is sought about some subset of the parameters, the others being considered to be nuisance parameters of no particular interest. In fortunate cases it may be possible to restructure the model so that the nuisance parameters are eliminated, and in all cases in which the support function is quadratic (or approximately so) the dimensions corresponding to the nuisance parameters can simply be ignored. Several other approaches are in use to eliminate nuisance parameters. Marginal likelihoods rely on finding some function of the data which does

not depend on them; notable examples involve the normal distribution, where a marginal likelihood for the variance can be found from the distribution of the sample variance which is independent of the mean, and a marginal likelihood for the mean can similarly be found using the t-distribution. Profile likelihoods, also called maximum relative likelihood, are found by replacing the nuisance parameters by their maximum-likelihood estimates at each value of the parameters of interest. It is easy to visualize from the case of two parameters why this is called a profile likelihood.

Naturally, a solution can always be found by strengthening the model through adopting particular values for the nuisance parameters, just as a Bayesian solution using integrated likelihoods can always be found by adopting a prior distribution for them and integrating them out, but such assumptions do not command wide assent. When a marginal likelihood solution has been found it may correspond to a Bayesian integrated likelihood for some choice of prior, and such priors are called neutral priors, to distinguish them from so-called uninformative priors for which no comparable justification exists. However, in the last analysis there is no logical reason why nuisance parameters should be other than a nuisance, and procedures for mitigating the nuisance must be regarded as expedients.

5. *Relation to Significance Testing*

All the common repeated-sampling tests of significance have their analogues in likelihood theory, and in the case of the normal model it may seem that only the terminology has changed. At first sight an exception seems to be the χ^2 goodness-of-fit test, where no alternative hypothesis is implied. However, this is deceptive, and a careful analysis shows that there is an implied alternative hypothesis that allows the variances of the underlying normal model to depart from their multinomial values. In this way the paradox of small values of χ^2 being interpreted as meaning that the model is 'too good' is exposed, for in reality they mean that the model is not good enough and that one with a more appropriate variance structure will have a higher likelihood. Likelihood ratio tests are based on the distribution under repeated-sampling of the likelihood ratio and are therefore not part of likelihood theory.

When likelihood arguments are applied to models with continuous sample spaces it may be necessary to take into account the approximation involved in representing data, which are necessarily discrete, by a continuous model. Neglect of this can lead to the existence of singularities in the likelihood function or other artifacts that a more careful analysis will obviate.

It is often argued that in comparing two models by means of a likelihood ratio, allowance should be made for any difference in the number of parameters by establishing a 'rate of exchange' between an additional parameter and the increase in log-likelihood expected. The attractive phrase 'Occam's bonus' has been suggested for such an allowance (Edwards 1969). However, the proposal seems only to have a place in a repeated-sampling view of statistical inference, where a bonus such as that suggested by Akaike's information criterion is sometimes canvassed.

The major application of likelihood theory so far has been in human genetics, where log-likelihood functions are regularly drawn for recombination fractions (linkage values) (see Ott 1991), but even there a reluctance to abandon significance-testing altogether has led to a mixed approach. Other examples, especially from medical fields, will be found in the books cited below.

6. *History and Further Reading*

Although historically the development of a likelihood approach to statistical inference was almost entirely due to R.A. Fisher, it is interesting to recall that the Neyman–Pearson approach to hypothesis testing derives ultimately from a remark of 'Student's' (W. S. Gosset) in a letter to E. S. Pearson in 1926 that 'if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability … you will be very much more inclined to consider that the original hypothesis is not true', a direct likelihood statement (quoted in McMullen and Pearson 1939). Indeed, it has been remarked that 'Just as support [log-likelihood] is Bayesian inference without the priors, so it turns out to be Neyman–Pearson inference without the ′errors′′ (Edwards 1972).

The literature on likelihood is gradually growing as an increasing number of statisticians become concerned at the inappropriate use of significance levels, confidence intervals and other repeated-sampling criteria to represent evidence. The movement is most advanced in biostatistics as may be seen from books such as Clayton and Hills (1993) and Royall (1997), but general texts such as Lindsey (1995) and Sprott (2000) exist as well. Amongst older books Cox and Hinkley (1974) contains much that is relevant to likelihood, whilst Edwards (1972, 1992) was the first book to advocate a purely likelihood approach, and is rich in relevant quotations from Fisher's writings. The history of likelihood is treated by Edwards (1974; reprinted in Edwards 1992).

See also; Probability and Chance: Philosophical Aspects; Probabilistic Thinking

Bibliography

Berger J O, Wolpert R L 1988 *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, CA

Birnbaum A 1962 On the foundations of statistical inference. *Journal of the American Statistical Association* **57**: 269–326

Clayton D G, Hills M 1993 *Statistical Models in Epidemiology*. Oxford University Press, Oxford, UK
Cox D R, Hinkley D V 1974 *Theoretical Statistics*. Chapman and Hall, London
Edwards A W F 1972 *Likelihood*. Cambridge University Press, Cambridge, UK
Edwards A W F 1974 The history of likelihood. *International Statistical Review* **42**: 9–15
Edwards A W F 1992 *Likelihood*. Johns Hopkins University Press, Baltimore, MD
Edwards J H 1969 In: Morton N E (ed.) *Computer Applications in Genetics*, University of Hawaii Press, Honolulu, HI
Fisher R A 1912 On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41**: 155–60
Fisher R A 1921 On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron* **1**(4): 3–32
Fisher R A 1956 *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, UK
Hacking I 1965 *Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK
Lindsey J K 1995 *Introductory Statistics: A Modelling Approach*. Clarendon Press, Oxford, UK
McMullen L, Pearson E S 1939 William Sealy Gosset, 1876–1937. *Biometrika* **30**: 205–50
Ott J 1991 *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, MD
Royall R 1997 *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall, London
Sprott D A 2000 *Statistical Inference in Science*. Springer, New York

A. W. F. Edwards

Limbic System

1. Introduction

In this article current ideas about the anatomical structures and functions of the limbic system will be described. The information will be limited to the limbic system as found in most mammals with greatest emphasis placed on structures prominent in the human brain. From the start it should be noted that there is less than universal agreement both about the structures that should be included in the limbic system and also about the roles played by them in mental functions and behavior. Some of the reasons for the differences of opinion among scientists arise from the many different approaches used to investigate the limbic system in both clinical and research settings. The simplest description of the limbic system is a set of highly interconnected anatomic areas lying below the cortical surface but above the most central regions of the forebrain, namely the thalamus, the hypothalamus, and areas of the ventral forebrain called the basal ganglia. Suggestions for other approaches to understanding the limbic system will be presented.

2. History of the Term, Limbic System

2.1 Early History

In the middle and later parts of the 1800s anatomists became interested in a region of tissue that surrounds the most medial parts of the brain, i.e., the thalamus, hypothalamus (together, the diencephalon), and the basal ganglia (caudate, putamen, nucleus accumbens, and associated regions). This medial band of tissue lies underneath the neocortical mantle and largely is separated from it by various sulci (fissures or indentations). In 1878 the French scientist Paul Broca described this deep band of tissue as a separate 'lobe' of the brain containing phylogenetically older cellular structures as well as the tracts that interconnected its various regions. Because it lies below the overlying neocortex it is hidden from external viewing. Broca argued that this hidden lobe was a common denominator of all mammalian brains and gave it the name 'the great limbic lobe.' The word limbic, which means border or fringe, was used because it is a fringe region around the deeper, more medial brain regions. Over time, many cortical regions, nuclear groups, and tracts have come to be considered components of the limbic lobe. Figure 1 shows a simplified diagram of the limbic lobe in the otter as drawn by Broca in 1878. The white area in this drawing represents the limbic lobe. From this illustration the great size of the limbic lobe is apparent, as is its strong connection with the large olfactory tracts in this animal. While in most mammals the limbic system is connected with the olfactory apparatus, the strength of the connections varies from species to species. Nevertheless the limbic lobe was originally thought to be that part of the brain associated with the sense of smell. As a result, it

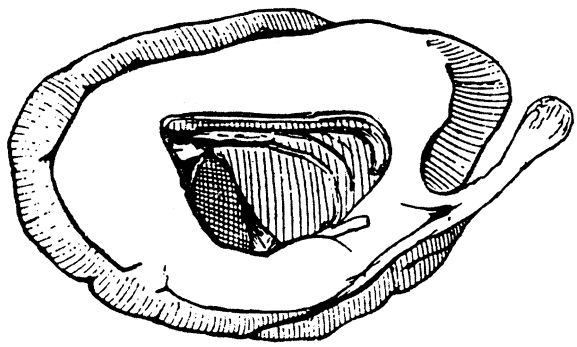

Figure 1
A modified version of Broca's drawing of the limbic system of the otter. In this sketch the brain is shown as if it had been cut along the longitudinal axis near the midline. The white area extending from the olfactory bulbs and surrounding the most medial parts of the brain represents the 'limbic lobe'

became known as the 'smell brain' or the rhinencephalon. Since the human has a relatively small olfactory system and smell was considered to be of little importance to people, it was relatively neglected by anatomists for many years.

2.2 More Recent History

It was in the late 1930s that some scientists began to think of the limbic system as being involved with emotional experiences, learning, and memory. Comparative anatomists were finding that the limbic system and its basic connections with other brain regions were quite similar in all vertebrates. This was most obvious in mammals. The great American anatomist C. Judson Herrick demonstrated similarities in the connections of limbic structures with other brain regions from the salamander through mammals (see Herrick 1956). In the salamander the most medial portions of the brain contain areas that are the primitive versions of at least three of the major cell groupings of the limbic system in all species, namely the septal area, the hippocampus, and the amygdala. Moreover, he was able to demonstrate that all of these areas had strong connections to the hypothalamus and the medial forebrain bundle. Each of these structures is discussed in detail in other sections of this encyclopedia.

2.3 The Medial Forebrain Bundle

In the salamander Herrick (1948) demonstrated a medial forebrain bundle that runs through the body of the hypothalamus. This bundle is found in all mammals and probably all vertebrates. It contains fibers running in both directions from the olfactory apparatus to brain stem nuclear groups, interchanging fibers with many nuclei of the hypothalamus along the way. These connections remain extensive throughout all mammalian brains. The medial forebrain bundle contains a number of powerful pathways connecting one end of the forebrain to the hindbrain. These pathways allow the forebrain and limbic system to modulate both somatic and visceral activity and provide routes whereby the internal organs and bodily actions can influence the activity of higher brain functions. Through the medial forebrain bundle and through other neural routes information from every sensory system reaches the limbic system.

2.4 The Papez Circuit

Research into the functions of the limbic system was greatly stimulated by a paper by Papez (1937). In it he argued that a particular set of limbic structures comprised the neural mechanism for the expression of the emotions. The pathways involved were extensive but the final projection leading to the subjective experience of emotion led from the hypothalamus to the anterior nucleus of the thalamus and then to the cingulate cortex. MacLean further championed the involvement of the limbic system with emotional behavior in a series of articles in the 1940s and 1950s (e.g., MacLean 1949, 1958). Both Papez and MacLean emphasized the importance of understanding the roles played by the cortical components of the limbic system.

3. Cortical Limbic Regions

In neuroanatomy the term cortex simply refers to tissue with neurons arranged in a layered pattern. In the human the neocortex is a thin sheet of layered neural tissue on the surface of the cerebral hemispheres. In animals with a large number of fissures about two-thirds of the total neocortex is buried within the walls of fissures. In the main, the cells of the neocortex are organized into six fairly recognizable layers. A considerable amount of cortical tissue in the forebrain is not neocortical but rather of a simpler design and with different patterns of cellular organization. Much of this non-neocortical tissue surrounds the limbic system, as seen by the white area in Fig. 1. The cingulate cortex (derived from the Latin word for belt or girdle) is one example of this more primitive cortex. It seems reasonable to consider the primitive cortical areas surrounding almost all of the limbic system structures as 'transitional cortex' in the sense that it comes between limbic structures and this tissue acts, in many if not all cases, as routes of entry and exit of neuronal information for the limbic structures. Other transitional cortices often mentioned are the entorhinal cortex, the periamygdaloid cortex, the pyriform cortex, and the subgenual cortex, to name but a few.

4. Nuclear Groups and Major Structures

The major nuclear structures usually considered as cornerstones of the limbic system are the septal area, the amygdala (see *Amygdala (Amygdaloid Complex)*), the hypothalamus (see *Hypothalamus*), the anterior nuclear group of the thalamus, and the hippocampus (see *Hippocampus and Related Structures*). The last region is actually a primitive forebrain cortical mass (archicortex). These nuclei and some of the neural pathways between them are shown in Fig. 2. They represent what many people think of as the limbic system. But, based on both the degree of anatomical

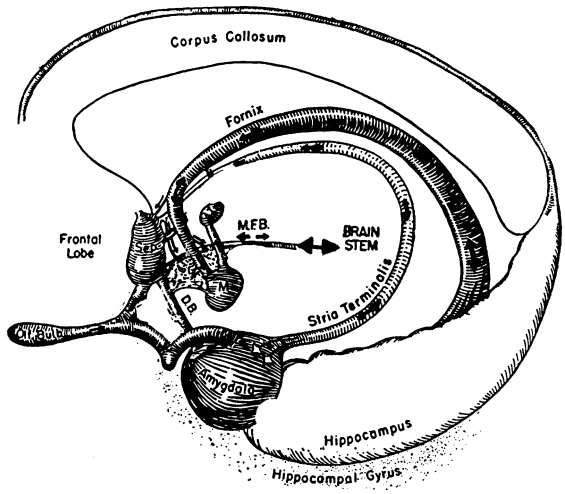

Figure 2
A schematic representation of many of the most commonly cited cell groups and fiber tracts of the limit system as found in the brains of primates. The hippocampus, amygdala, and frontal lobe region are identified as such. The fornix (connecting the hippocampus, septal areas, and the mammillary bodies, M,) is also labeled clearly. The stria terminalis also is designated. This is a major pathway interconnecting portions of the amygdala with the hypothalamus and other forebrain regions. M.F.B.: medial forebrain bundle. D. B.: diagonal band of Broca. In this drawing the diagonal band is shown as a fiber tract. However, most scientists consider it to be a nerve cell rich area continuous with the medial septal area. Sep: septal area. Olf. Bulb: olfactory bulb. Modified from MacLean 1949

interconnections and functional considerations, many other nuclear groups have to be included to represent a reasonably complete limbic system in terms of knowledge at the start of the twenty-first century. For example, the nucleus accumbens and the ventral and medial portions of the caudate nucleus share strong neural interconnections with several major limbic regions. Functionally, they have attributes associated with limbic activities. One such attribute is their participation in the generation of feelings of pleasure.

4.1 *Fiber Tracts*

The major tracts of the limbic system include the fornix (hypothalamus–septal area–hippocampus), the stria terminalis (amygdala–hypothalamus), the mammillothalamic tract (mammillary bodies of the hypothalamus to the anterior nuclei of the thalamus), and the cingulum bundle (running through the cingulate cortex from one end to the other). Although this last tract is sometimes thought to be a private association pathway for the limbic lobe, actually it carries fibers of neocortical origin forward and backward through the brain.

Of course, as described above, these nuclei, cortical areas, and tracts represent but a rough approximation of the true complexity of the regions and pathways. All of these limbic categories are far more complex than have been described above. The interested reader is referred to Paxinos (1995) for the anatomic details of the rat brain and to Crosby et al. (1962) for the human.

5. *Ever-increasing Complexity*

5.1 *The Consequences of Increased Knowledge*

In general, the complexity of the limbic system arises from the ever-increasing subdivisions of the larger limbic units and pathways. As new methods of identifying subdivisions of larger regions became available through advances in anatomical, physiological, and behavioral methods, finer differentiations of limbic structures were necessary. For example, as many as 22 distinct nuclear divisions within the main body of the amygdala have been proposed. But anatomical differentiation does not imply functional specificity or homogeneity within the subdivisions of anatomical areas. Even the most important, or at least best known, tracts within the limbic lobe have multiple components. For example, the large group of fibers connecting the septal area and the hippocampus has at least three separate components, that is, a large bundle just below the corpus callosum, a fiber bundle that travels for long distances in the *corpus callosum*, and a fiber bundle traveling over the top of the corpus callosum (the *fornix longus*).

5.2 *Contacts with Areas Beyond Traditional Limbic Regions*

As more and more has been learned about the connections of the various structures of the limbic system, it has become impossible to think of any nuclear group as an isolated entity. For example, the amygdala has strong connections with the prefrontal cortical areas, both the lateral and orbital regions (see *Prefrontal Cortex*), and with various neocortical and transitional cortical and neocortical regions of the temporal lobe (see *Temporal Lobe*). No unit of the limbic system operates on its own but rather in close collaboration with these other regions of the forebrain mantle. Every component of the limbic system has multiple associations with areas and nuclei below the

forebrain regions, including dopamine-containing cell groups of the ventral tegmentum and the substantia nigra. The fibers from these nuclei pass through the medial forebrain bundle to the hypothalamus and the anterior extremes of the forebrain, as well as the nucleus accumbens.

5.3 Extensions of Limbic Structures

Several years ago, the concept of an 'extended amygdala' was proposed (Alheid et al. 1995). The reason for the creation of this concept was that many areas, pathways, and nuclei associated with these pathways seem closely tied to specific functions of certain regions of the amygdala. They seemed to be regions that were true extensions of systems originating in the body of the amygdala, forming a unified system. Some of these associated regions were not located close to the amygdala itself. The same sort of approach could be used with the hippocampus. Some distant regions of the brain can produce similar effects to those obtained from stimulation or lesions of the hippocampus. They mimic one or more of the functions of the hippocampus. Such areas are strongly connected with the hippocampus through known neuronal pathways despite distant locations. One example would be the hindbrain sites involved in food intake and blood sugar regulation (Ritter et al. 2000). Thus, there is a basis for the concept of an 'extended hippocampus.' The process of creating 'extended' limbic areas, or for that matter 'extended' visual or auditory systems, would soon make every part of the brain an extension of many, perhaps all, of what are thought to be the major functional groupings of the brain. The use of 'extended' systems in the brain may not turn out to be a useful approach. It should not be forgotten that at all times many parts of the brain are working together to accomplish specific goals.

6. A Different Approach

The problem faced by scientists studying the limbic system is that while most have a general idea of the history and the traditional functions associated with a 'visceral' and emotion-related set of neural structures, there are no firm boundaries between limbic and nonlimbic forebrain regions. Evidence can be cited to implicate every area of the brain, including the neocortical mantle, in visceral, autonomic, and emotional behavior and experience. Given this situation, it is appropriate to reject an all-or-none, black or white, approach to the limbic system. It is time to recognize that every area of the brain may have involvement with 'limbic functions' at certain times or in certain situations.

This can be accomplished by adopting a Fuzzy Logic approach (Klir and Folger 1988) to the limbic system in which every area of the brain has some degree of limbic involvement (see Isaacson 1993). The question then becomes not whether an area is or is not a part of the system but to what degree the area is associated with limbic functions. The amount of this association may, of course, change under different circumstances. The cortical regions as well as the nuclei and pathways found in Figs. 1 and 2 would certainly have high degrees of 'limbic-ness' but this would not exclude their participation in other types of brain functions since they would also have varying degrees of associations with other brain systems. At the same time it would not exclude other brain areas from participating in limbic functions to some degree and at certain times. This approach would also fit with the fundamental principle of the human brain, namely that under all but the most unusual conditions, the brain's systems operate together as a harmonious whole.

See also: Amygdala (Amygdaloid Complex); Emotion, Neural Basis of; Hippocampus and Related Structures; Hypothalamus; Olfactory System; Prefrontal Cortex; Taste and Smell, Psychology of; Temporal Lobectory Discrimination.

Bibliography

Alheid G F, de Olmos J S, Beltramino C A 1995 Amygdala and extended amygdala. In: Paxinos G (ed.) *The Rat Nervous System*, 2nd edn. Academic Press, San Diego, CA, pp. 495–578

Broca P 1878 Anatomie comparée des circonvolutions cérébrales. Le grand lobe limbique et la scissure limbique dans la série des mammifères. *Review of Anthropology* **1**(Ser.2): 385–498

Crosby E C, Humphrey T, Lauer E W 1962 *Correlative Anatomy of the Nervous System*. Macmillan, New York

Herrick C J 1948 *The Brain of the Tiger Salamander, Ambystoma tigrinum*. University of Chicago Press, Chicago

Herrick C J 1956 *The Evolution of Human Nature*. University of Texas Press, Austin, TX

Isaacson R L 1993 A 'fuzzy set' perspective of the limbic system: implications for cognition and memory. *Neuroscience Research Communication* **13**(suppl.): S15–S18 (2nd edn. Academic Press, San Diego, CA, pp. 495–578)

Klir G J, Folger T A 1988 *Fuzzy Sets, Uncertainty, and Information*. Prentice Hall, Englewood Cliffs, NJ

MacLean P D 1949 Psychosomatic disease and the 'visceral brain.' Recent developments bearing on the Papez theory of emotion. *Psychosomatic Medicine* **11**: 338–53

MacLean P D 1958 Contrasting functions of limbic and neocortical systems of the brain and their relevance to psychophysiological aspects of medicine. *American Journal of Medicine* **25**: 611–26

Papez J W 1937 A proposed mechanism of emotion. *Archives of Neurology and Psychiatry* **38**: 725–43

Paxinos G (ed.) *The Rat Nervous System*, 2nd edn. Academic Press, San Diego, CA

Ritter S, Dinh T T, Zhang Y 2000 Localization of hindbrain glucoreceptive sites controlling food intake and blood glucose. *Brain Reseach* **856**: 37–47

R. L. Isaacson

Liminality

The concept of 'liminality,' as developed and introduced into anthropological discourse by Victor Turner in the 1960s and 1970s, derives from the work of Arnold van Gennep, a French ethnologist and folklorist. In his protostructural analysis of transition rituals, *Les rites de passage*, published in 1909, but not translated into English until 1960, van Gennep (1960) argued that 'regeneration' was the law of life, and that it was accomplished through rites of passage which have three major phases: 'separation, transition or margin (limen), and incorporation.' This article examines Turner's elaboration and application of 'liminality' in ritual and symbolic studies, its primary dimensions of 'communitas' and 'reflexivity,' the distinction between 'liminal' and 'liminoid,' and the popularity and significance of these concepts throughout the human sciences.

1. *'Liminality' and African Ritual*

Like other anthropologists associated with the Rhodes-Livingstone Institute in the 1950s, Turner read and found relevant to his fieldwork Henry Junod's use of van Gennep's interpretative apparatus for understanding Thonga ritual in *The Life of a South African Tribe* (Junod 1927). During their layover in Cape Town, before returning to Manchester, England, Turner's wife and academic partner, Edie, attended Monica Wilson's seminar on van Gennep and *rites de passage*.

Turner's first published use of van Gennep's theory came in 1960 in a conference paper on Ndembu circumcision ritual (Turner 1962). This was one of four essays analyzing rites of passage published by mentor Max Gluckman in *Essays on the Ritual of Social Relations*. Gluckman's own essay 'Les rites de passage,' lays out van Gennep's theory regarding transitions, and his three-part model of such rituals. Like van Gennep, Turner was especially interested in the threshold or 'liminal' phase with its dance of old and new and promise of regeneration, and with the classificatory ambiguity embodied in, for example, being no longer a boy but not yet a man.

In 1964, Turner reformulated and developed van Gennep's concept and introduced the term 'liminality' in 'Betwixt and Between: The Liminal Period in rites de passage,' a paper presented at the American Ethnological Society meetings, and published that same year in the AES Proceedings. This model of 'processual analysis' was reprinted as the key Chapter 4 in *The Forest of Symbols: Aspects of Ndembu Ritual* (Turner 1967), and 'liminality,' as not only a phase or space in a rite of passage, but as the condition or state of being 'betwixt and between,' soon became commonplace in ritual studies. 'Liminality' also began to extend its reach throughout academic disciplines, referring to places, persons, and processes far afield from tribal ritual.

2. *The Ritual Process*

The Ritual Process: Structure and Anti-structure (Turner 1969) was based on the Lewis Henry Morgan Lectures that Turner delivered at the University of Rochester in 1966. As he wrote in 1977 in the foreword to the paperback edition of this book, 'it was in the course of these lectures that I crossed the threshold between the study of ritual in an African tribal context and the analysis of processual symbols in cross-cultural and transtemporal terms.' In the 'incursive nomadism' of his own work, Turner was to use 'liminal' to describe everything from millenarian movements and pilgrimage processes in industrial society, to cross-cultural play phenomena, to cultural performances and the written genres of literate cultures, and finally to experimental postmodern theater and to the functioning of the human brain. Cultures and societies, he argued, were processes not systems, and all sociocultural processes were, of necessity, characterized by a dialog between structure and antistructure, which was used synonymously both with 'liminal' or 'liminality' and with 'communitas.'

3. *'Liminal' to 'Liminoid'*

By the early 1970s, the adjective 'liminal' had been coupled with so many diverse phenomena, and Turner had been so criticized by his anthropological colleagues, that he felt compelled to introduce the term 'liminoid' to refer to the secular leisure genres and optional play forms of complex industrial society, reserving 'liminal' for sacred, prescribed, ritualized transitions between states in small-scale, stable societies (Turner 1974b). In answer to criticism that he had overgeneralized and overextended the concept of 'liminality,' he now insisted that 'liminoid' genres are similar to but not identical with the ritually 'liminal.' The distinction never really caught on—in part because people found the word 'gratingly neologistic'; in part because Turner himself had already coupled 'liminal' with many phenomena other than tribal ritual; but in largest part because the characteristics, if not the consequences, of the situation being described, whether a Ndembu circumcision ritual, a carnival in

Rio, a Pueblo clown performance, or a Dostoevski novel, are much the same.

Turner discovered, as he wrote in 1974, that the concept of liminality, of social antistructure was 'the theoretical operator' necessary to processualism, to a view of society as dynamic and dramatic (Turner 1974a); to getting beyond what he regarded as 'the dualistic rigor mortis of the system of binary oppositions' (1974b), and to a 'comparative symbology' method. And, like the interstitial ritual situation that 'liminality' described, the concept offered him 'the possibility of standing aside and formulating alternatives' and 'the freedom to meditate and speculate' (1974a).

4. *Liminality and Communitas*

In *The Ritual Process*, Turner (1969) also argued that the phenomenon of 'communitas' epitomized antistructure and that this relation of full, unmediated communication was one of the fundamental aspects of 'liminality.' These transient, undifferentiated states of relatedness are 'seedbed[s] of cultural creativity,' both the source of structures and at the same time their critique (Turner 1974b). As he repeatedly argued (1969), 'no society can function without a dialectic between structure and "communitas."' In response to criticisms that his formulations and applications of this concept were, like 'liminality,' too general and universal, Turner (1974a) was to distinguish among 'spontaneous,' 'ideological,' and 'normative' 'communitas.'

5. *Liminality and Reflexivity*

Like 'communitas,' the idea of 'reflexivity' is implicit in his earliest formulations of liminality. In 'Betwixt and Between' (1967) Turner suggests that 'liminality may be partly described as a state of reflection.' Drawing upon William James's 'law of dissociation,' he argues that Ndembu figures such as the man-lion monsters who appear to neophytes 'break the cake of custom and enfranchise speculation.' Subsequently, he was to argue that major 'liminal' and 'liminoid' situations are 'occasions on which a society takes cognizance of itself,' of 'plural reflexivity' (1974a), and that 'social drama is the major form of plural reflexivity in human social action' (in MacAloon 1984).

Turner's genius was to recognize not only that conflict and ambiguity provoke thought, but further that only when there is a breach in social relations, a discontinuity in customary process, a gap or aporia in discourse, or the deliberate paradox and disordering of liminal symbols, do we truly reflect upon the order of things. The betwixt and betweenness that he described as 'liminality' is a paradoxical and universal human experience. Like all paradoxes—whether in life, literature, or religion, and whether rhetorical, logical, or epistemological—the paradoxes of the 'liminal' situation embody and engender 'reflexivity.'

Turner also had a genius for generating variations on a theme, and for organizing colleagues around these topics. Between 1977 and 1984, he engendered and participated in two interdisciplinary symposia and one Wenner-Gren international conference focused on 'reflexivity.' All were published as books or special issues of journals: Babcock (1980), Ruby (1982), and MacAloon (1984). In the last years of his life, Turner became increasingly interested in a 'multiplicity of desacralized performative genres' that 'have assumed the task of plural reflexivity' (Turner 1985), in 'the ways in which both "liminal" and "liminoid" phenomena constitute metalanguages (including nonverbal ones), devised for talking about the various languages of everyday' (Turner 1992).

6. *A Critical Commonplace*

As a consequence of the formulations that Turner made in lectures and publications between 1962 and 1982, 'liminality' has become virtually a household word, not only in anthropology and the social sciences generally, but in literary and religious studies as well. Like 'irony' in the New Literary Criticism of several decades ago, 'androgyny' in early feminist theory, or 'deconstruction' in post-structuralist theory, Turner's formulation of 'liminality' gave 1970s scholars across the disciplines a model of and a word for talking about central incertitudes—the ambiguity, paradox, and disorder as well as the order of the human condition.

With the more recent interdisciplinary developments of cultural studies, postcolonial theory, resistance studies, and queer theory, and the increasingly positive valuation of the interstitial, the hybrid, the *mestizaje* in the last two decades of the twentieth century, 'liminality' has become even more popular and visible. In addition to the critical terms already mentioned, 'liminality' is also the paradigm for the recognition and valuing of borderlands and borderland epistemology, contact zones, heteropias, undecidability, instability, indeterminacy, disorder, gynesis, impurity, dialogism, heteroglossia, the carnivalesque, the pharmakon, transgender, and third sexes and genders, etc. Largely because of Turner's championing of the *via negativa*, such ambiguous states and spaces are now regarded as fundamental, necessary, and desirable rather than problematic.

Few concepts in the social sciences produce their own journal. In 2000, *LIMEN: A Journal for the Theory and Practice of Liminal Phenomena* was founded in the politically liminal space of Croatia. 'Liminal' is no longer a marked adjective. In scholarly publications, it is commonly used without quotation marks

and without references. 'Liminality' is a crossover term not only among disciplines, but between academic and popular discourse. Novels, movies, actors, musicians all talk about 'that space in between,' for which 'liminal' has become the adjective of choice.

See also: Borders, Anthropology of; Play, Anthropology of; Reflexivity in Anthropology; Ritual; Ritual and Symbolism, Archaeology of; Symbolic Boundaries: Overview; Symbolism in Anthropology; Queer Theory

Bibliography

Babcock B A 1980 Signs about signs: The semiotics of self-reference. *Semiotica* **30**: 1–2

Junod H A [1912/13] 1962 *The Life of a South African Tribe*. University Books, New Hyde Park, NY

MacAloon J J (ed.) 1984 *Rite, Drama, Festival, Spectacle: Rehearsals Toward a Theory of Cultural Performance*. ISHI Press, Philadelphia, PA

Ruby J (ed.) 1982 *A Crack in the Mirror: Reflexive Perspectives in Anthropology*. University of Pennsylvania Press, Philadelphia, PA

Turner V W 1962 Three symbols of passage in Ndembu circumcision ritual: An interpretation. In: Gluckman M (ed.) *Essays on the Ritual of Social Relations*. Manchester University Press, Manchester, UK

Turner V W 1967 *The Forest of Symbols: Aspects of Ndembu Ritual*. Cornell University Press, Ithaca, NY

Turner V W 1969 *The Ritual Process: Structure and Anti-Structure*. Aldine, Chicago

Turner V W 1974a *Dramas. Fields, and Metaphors: Symbolic Action in Human Society*. Cornell University Press, Ithaca, NY

Turner V W 1974b Liminal to liminoid in play, flow, and ritual: An essay in comparative symbology. *Rice University Studies* **60**(3): 53–92

Turner V W 1977 *The Ritual Process: Structure and Anti-structure*. Cornell University Press, Ithaca, NY

Turner V W 1985 Process, system, and symbol: A new anthropological synthesis. In: Turner V W *On the Edge of the Bush: Anthropology as Experience*, Turner E (ed.) University of Arizona Press, Tucson, AZ

Turner V W 1992 Variations on a theme of liminality. In: Turner V W (ed.) *Blazing the Trail: Way Marks in the Exploration of Symbols*. University of Arizona Press, Tucson, AZ

Van Gennep A [1909] 1960 *The Rites of Passage*. Routledge and Kegan Paul, London

B. A. Babcock

Linear Algebra for Neural Networks

Linear algebra is particularly well suited to analyze the class of neural networks called associators. These quantitative models learn to associate input and output patterns adaptively with the use of learning algorithms. When the set of input patterns is different from the set of output patterns, the models are called heteroassociators. When input patterns and output patterns are the same, the models are called autoassociators. Associators consist of layers of elementary units called neurons. Information flows from the first layer (the input layer), to the last one (the output layer). Some architectures may include intermediate layers (the hidden layers). Typically neurons from a given layer are connected to the neurons of another layer. Linear algebra operations describe the transformations of the information as it flows from one layer to another one.

1. Prerequisite Notions and Notations

Vectors are represented by lower case boldface letters (e.g., $\boldsymbol{x}$), matrices by upper case boldface letters (e.g., $\boldsymbol{X}$). The following elementary notions are supposed to be known (see Abdi et al. 1999, for a neural network approach): the transpose operation, (e.g., $\boldsymbol{x}^{\mathrm{T}}$), the norm of a vector (e.g., $\|\boldsymbol{x}\|$), the scalar product (e.g., $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}$) and the product of two matrices (e.g., $\boldsymbol{AB}$). We will use also the elementwise or Hadamar product (e.g., $\boldsymbol{A} \circledast \boldsymbol{B}$).

2. Projection

2.1 Cosine Between Two Vectors

The *cosine* of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the cosine of the angle made by the origin of the space and the points defined by the coordinates of the vectors. Thus,

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\| \; \|\boldsymbol{y}\|}. \tag{1}$$

The cosine reflects the *similarity* between vectors. When two vectors are proportional to each other (i.e., when they have the same direction), their cosine is equal to 1; when they are orthogonal to each other, their cosine is equal to 0.

2.2 Distance Between Vectors

Among the large family of *distances* between vectors, the most popular by far is the *Euclidean* distances. It is closely related to the cosine between vectors and is defined as

$$\begin{aligned} \mathrm{d}^2(\boldsymbol{x}, \boldsymbol{y}) &= (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}-2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \\ &= \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \\ &\quad -2[\cos(\boldsymbol{x}, \boldsymbol{y}) \times \|\boldsymbol{x}\| \times \|\boldsymbol{y}\|] \\ &= \sum (x_i - y_i)^2. \end{aligned} \tag{2}$$

2.3 Projection of One Vector onto Another Vector

The (orthogonal) projection of vector $\boldsymbol{x}$ on vector $\boldsymbol{w}$ is defined as

$$\mathrm{proj}_{\boldsymbol{w}}\boldsymbol{x} = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}\boldsymbol{w} = \cos(\boldsymbol{x}, \boldsymbol{w}) \times \frac{\| \boldsymbol{x} \|}{\| \boldsymbol{w} \|}\boldsymbol{w}. \tag{3}$$

The norm of $\mathrm{proj}_{\boldsymbol{w}}\boldsymbol{x}$ is its distance to the origin of the space. It is equal to

$$\| \mathrm{proj}_{\boldsymbol{w}}\boldsymbol{x} \| = \frac{|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}|}{\| \boldsymbol{w} \|} = |\cos(\boldsymbol{x}, \boldsymbol{y})| \times \| \boldsymbol{x} \| . \tag{4}$$

2.4 Illustration: Hebbian and Widrow–Hoff Learning Rules

A neural network consists of cells connected to other cells via modifiable weighted *connections* called *synapses*. Consider a network made of I input cells and only one output cell. The information is transmitted via the synapses, from a set of external input cells to the output cell which gives a response corresponding to its state of activation. If the input pattern and the set of synaptic weights are given by I-dimensional vectors denoted $\boldsymbol{x}$, and $\boldsymbol{w}$, the activation of the output cell is obtained as

$$a = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}. \tag{5}$$

So the activation is proportional to the norm of the projection of the input vector onto the weight vector. The *response* or *output* of the cell is denoted o. For a linear cell, it is proportional to the activation (for convenience, assume that the proportionality constant is equal to 1). *Linear heteroassociators* and *autoassociators* are made of linear cells. In general, the output of a cell is a function (often, but not necessarily, continuous), called the *transfer function*, of its activation

$$o = f(a). \tag{6}$$

For example, in *backpropagation networks*, the (nonlinear) transfer function is usually the logistic function

$$o = f(a) = \mathrm{logist}\, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} = \frac{1}{1+\exp\{-a\}}. \tag{7}$$

Often, a neural network is designed to associate, to a given input, a specific response called the target, denoted t. *Learning* is equivalent to defining a rule that specifies how to add a small quantity to each synaptic weight at each learning iteration. Learning makes the output of the network closer to the target.

Learning rules come in two main flavors: *supervised* (e.g., Widrow–Hoff) which take into account the error or distance between the response of the neuron and the target and *unsupervised* (e.g., Hebbian) which do not require such 'feedback.' The Hebbian learning rule modifies the weight vector at iteration $n+1$ as

$$\boldsymbol{w}_{[n+1]} = \boldsymbol{w}_{[n]} + \eta t \boldsymbol{x} \tag{8}$$

where η is a small positive constant called the learning constant. So, a Hebbian learning iteration moves the weight vector in the direction of the input vector by an amount proportional to the target.

The Widrow–Hoff learning rule uses the error and the derivative of the transfer function (f') to compute the correction as

$$\boldsymbol{w}_{[n+1]} = \boldsymbol{w}_{[n]} + \eta f'(a)(t-o)\boldsymbol{x}. \tag{9}$$

So, a Widrow–Hoff learning iteration moves the weight vector in the direction of the input vector by an amount proportional to the error.

For networks with several cells (say J) in the output layer, the pattern of activation, output, and target become J-dimensional vectors (denoted $\boldsymbol{a}$, $\boldsymbol{o}$, and $\boldsymbol{t}$, respectively), and the synaptic weights are stored in an $I \times J$ matrix $\boldsymbol{W}$. The learning equations become

$$\boldsymbol{W}_{[n+1]} = \boldsymbol{W}_{[n]} + \eta \boldsymbol{x}\boldsymbol{t}^{\mathrm{T}} \text{ (Hebbian) and}$$

$$\boldsymbol{W}_{[n+1]} = \boldsymbol{W}_{[n]} + \eta(f'(\boldsymbol{a}) \circledast \boldsymbol{x})(\boldsymbol{t}-\boldsymbol{o})^{\mathrm{T}}$$

$$\text{(Widrow–Hoff)} \tag{10}$$

(the derivative of the transfer function is applied elementwise).

In general, several (say K) input/target associations are to be learned. Then the set of input patterns is stored in an $I \times K$ matrix denoted $\boldsymbol{X}$, the activation and target patterns respectively are stored in $J \times K$ matrices denoted $\boldsymbol{A}$ and $\boldsymbol{T}$ respectively. The activation and learning iterations can be computed for all the patterns at once (this is called *batch* mode). The output matrix is computed as

$$\boldsymbol{O} = f(\boldsymbol{A}) = f(\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}) \tag{11}$$

(f is applied elementwise). The learning equations become

$$\boldsymbol{W}_{[n+1]} = \boldsymbol{W}_{[n]} + \eta \boldsymbol{X}\boldsymbol{T}^{\mathrm{T}} \text{ (Hebb) and} \tag{12}$$

$$\boldsymbol{W}_{[n+1]} = \boldsymbol{W}_{[n]} + \eta(f'(\boldsymbol{A}) \circledast \boldsymbol{X})(\boldsymbol{T}-\boldsymbol{O})^{\mathrm{T}}$$

$$\text{(Widrow–Hoff).} \tag{13}$$

3. Eigenvalues, Eigenvectors, and the Singular Value Decomposition

Eigenvectors of a given square matrix, $\boldsymbol{W}$ (resulting from its *eigendecomposition*) are vectors invariant under multiplication by $\boldsymbol{W}$. The eigencomposition is best defined for a subclass of matrices called *positive semi-definite* matrices. A matrix $\boldsymbol{X}$ is positive semi-definite if there exists another matrix $\boldsymbol{Y}$ such that $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$. This is the case for most matrices used in neural networks and so we will consider only this case here.

Formally, a (nonzero) vector $\boldsymbol{u}$ is an eigenvector of a square matrix $\boldsymbol{W}$ if

$$\lambda \boldsymbol{u} = \boldsymbol{W}\boldsymbol{u}. \tag{14}$$

The scalar λ is the *eigenvalue* associated with $\boldsymbol{u}$. So, $\boldsymbol{u}$ is an eigenvector of $\boldsymbol{W}$ if its direction is invariant under multiplication by $\boldsymbol{W}$ (only its length is changed if $\lambda \neq 1$). There are, in general, several eigenvectors for a given matrix (at most as many as the dimension of $\boldsymbol{W}$). They are in general ordered by decreasing order of their eigenvalue. So, the first eigenvector, $\boldsymbol{u}_1$ has the largest eigenvalue λ_1. The number of eigenvectors with a non-zero eigenvalue is the rank of the matrix.

The eigenvalues of positive semi-definite matrices are always positive or zero (a matrix with strictly positive eigenvalues, is *positive definite*). Also, any two eigenvectors with different eigenvalues are orthogonal, i.e.,

$$u_l^{\mathrm{T}} u_{l'} = 0 \quad \forall l \neq l'. \tag{15}$$

In addition, the set of eigenvectors of a matrix constitutes an orthogonal basis for its rows and columns. This is expressed by defining two matrices: the eigenvector matrix, $\boldsymbol{U}$, and the diagonal matrix of the eigenvalues: $\boldsymbol{\Lambda}$. The eigendecomposition of $\boldsymbol{W}$ (with rank L) is

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} = \sum_{l}^{L} \lambda_l u_l u_l^{\mathrm{T}},$$
$$\text{or equivalently: } \boldsymbol{\Lambda} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{U}. \tag{16}$$

The *singular value decomposition* (SVD) generalizes the eigendecomposition to rectangular matrices. If $\boldsymbol{X}$ is an $I \times K$ matrix, its SVD is defined as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^{\mathrm{T}} \quad \text{with} \quad \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}$$
$$\text{and } \boldsymbol{\Delta} \text{ being a diagonal matrix.} \tag{17}$$

($\boldsymbol{I}$ being the identity matrix). The diagonal elements of $\boldsymbol{\Delta}$ are real positive numbers called the singular values of $\boldsymbol{X}$. The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are the left and right matrices of singular vectors (which are also eigenvectors, see later). The SVD is closely related to the eigendecomposition because $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{\Delta}$ can be obtained from the eigendecomposition of matrices $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ as

$$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}, \quad \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}, \quad \text{and } \boldsymbol{\Delta} = \boldsymbol{\Lambda}^{\frac{1}{2}} \tag{18}$$

(note that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ have the same eigenvalues).

The eigen- and singular value decompositions are used in most fields of applied mathematics including statistics, image processing, mechanics, and dynamical systems. For neural networks, they are essential for studying the dynamics of linear autoassociators and heteroassociator.

3.1 Iterative Processes

A linear heteroassociator using the Widrow–Hoff rule, learning modifies only the eigenvalues of the weight matrix. Specifically, if the patterns to be learned are stored in an $I \times K$ matrix $\boldsymbol{X}$, with singular value decomposition as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^{\mathrm{T}}$, then the learning equation (see Eqn. 12) becomes [because for a linear heteroassociators, $\boldsymbol{O} = \boldsymbol{A}$, and $f'(\boldsymbol{A}) = \boldsymbol{I}$]

$$\begin{aligned}\boldsymbol{W}_{[n+1]} &= \boldsymbol{W}_{[n]} + \eta \boldsymbol{X}(\boldsymbol{T} - \boldsymbol{A})^{\mathrm{T}} \\ &= \boldsymbol{U}\{\boldsymbol{\Delta}^{-1}[\boldsymbol{I} - (\boldsymbol{I} - \eta\boldsymbol{\Delta}^2)^{n+1}]\}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{I}^{\mathrm{T}}\end{aligned} \tag{19}$$

(see Abdi 1994, p. 54).

The Hebbian weight matrix corresponds to the first iteration of the algorithm, i.e.,

$$\begin{aligned}\boldsymbol{W}_{[1]} &= \boldsymbol{U}\{\boldsymbol{\Delta}^{-1}[\boldsymbol{I} - (\boldsymbol{I} - \eta\boldsymbol{\Delta}^2)^1]\}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{T}^{\mathrm{T}} \\ &= \boldsymbol{U}\{\eta\boldsymbol{\Delta}\}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{T}^{\mathrm{T}} = \eta\boldsymbol{X}\boldsymbol{T}^{\mathrm{T}}.\end{aligned} \tag{20}$$

Equation 19 characterizes the value of η that allow the iterative process to converge. Denoting by $\delta_{\max}$ the largest singular value of $\boldsymbol{X}$, if η is such that

$$0 < \eta < 2\delta_{\max}^{-2} \tag{21}$$

then it can be shown that (see Abdi 1994)

$$\lim_{n \to \infty} (\boldsymbol{I} - \eta\boldsymbol{\Delta}^2)^n = 0, \quad \text{and}$$
$$\lim_{n \to \infty} \boldsymbol{W}_{[n]} = \boldsymbol{U}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{T}^{\mathrm{T}} = \boldsymbol{X}^{+}\boldsymbol{T}. \tag{22}$$

The matrix $\boldsymbol{X}^{+} = \boldsymbol{U}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^{\mathrm{T}}$ is the *pseudo-inverse* of $\boldsymbol{X}$. It gives a least-square optimal solution for the association between the input and the target. Thus, the linear heteroassociator is equivalent to *linear multiple regression*. If η is outside the interval defined by Eqn. 21, then both the singular values and the elements of the weight matrix will grow at each iteration. In practice, because neural networks are stimulated by digital computers. The weight matrix will eventually reach the limits of the precision of the machine.

When the target vectors are the same as the input vectors (i.e., when each input is associated to itself), the linear heteroassociator becomes a linear autoassociator. The previous approach shows that, now, the Hebbian weight matrix is the cross-product matrix

$$\boldsymbol{W}_{[1]} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}. \tag{23}$$

With Widrow–Hoff learning, when convergence is reached, all nonzero eigenvalues of the weight matrix are equal to 1. The weight matrix is then said to be *spherical*; it is equal to

$$\boldsymbol{W}_{[\infty]} = \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}. \tag{24}$$

Because the statistical technique of *principal component analysis* (PCA) computes the eigendecomposition of a cross-product matrix similar to $\boldsymbol{W}$, the linear autoassociator is considered as the neural network equivalent of PCA.

4. Optimization, Derivative and Matrices

Neural networks are often used to optimize a function of the synaptic weights. The *differentiation* of a function is the major concept for exploring optimization problems and, for neural networks, it involves differentiating vectors or matrix functions. In this context, we need to consider the transfer function as being a function of the weight vector. This is expressed by rewriting Eqn. 6 as

$$o = f(\boldsymbol{w}). \tag{25}$$

The derivative of $f(\boldsymbol{w})$ with respect to the $I \times 1$ vector $\boldsymbol{w}$ is denoted by $\nabla_{f(\boldsymbol{w})}$. It is also called the gradient of f, i.e.,

$$\nabla_{f(\boldsymbol{w})} = \frac{\partial f}{\partial \boldsymbol{w}} = \left[\frac{\partial f}{\partial w_1}, \ldots, \frac{\partial f}{\partial w_i}, \ldots, \frac{\partial f}{\partial w_I} \right]^{\mathrm{T}}. \tag{26}$$

For example, the derivative of the output of a linear neuron is

$$\begin{aligned} \frac{\partial f}{\partial \boldsymbol{w}} &= \left[\frac{\partial \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}{\partial w_1}, \ldots, \frac{\partial \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}{\partial w_i}, \ldots, \frac{\partial \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}{\partial w_I} \right]^{\mathrm{T}} \\ &= [x_1, \ldots, x_i, \ldots, x_I]^{\mathrm{T}} = \boldsymbol{x}. \end{aligned} \tag{27}$$

When a function is twice differentiable, the second order derivatives are stored in a matrix called the *Hessian* matrix of the function. It is often denoted by $\boldsymbol{H}$ or ∇_f^2 and is formally defined as

$$\boldsymbol{H} = \nabla_f^2 = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 w_I} \\ \frac{\partial^2 f}{\partial w_2 w_1} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 w_I} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_I w_1} & \frac{\partial^2 f}{\partial w_I w_2} & \cdots & \frac{\partial^2 f}{\partial w_I^2} \end{bmatrix}. \tag{28}$$

4.1 Conditions for Minimum

A standard problem is to show that a given learning rule finds an optimum solution in the sense that a function of the weight vector (or matrix) called the *error function* reaches its minimum value when learning has converged. Often, the error function is defined as the sum of the squared error over all patterns.

When the gradient of the error function can be evaluated, a necessary condition for optimality (i.e., either minimum or maximum) is to find a weight vector $\tilde{\boldsymbol{w}}$ such that

$$\nabla_{f(\tilde{\boldsymbol{w}})} = 0. \tag{29}$$

This condition is also sufficient provided $\boldsymbol{H}$ is positive definite (see Haykin 1999).

4.2 Taylor Expansion

The Taylor expansion is the standard technique used to obtain a linear or a quadratic approximation of a function of one variable. Recall that the Taylor expansion of a continuous function $f(x)$ is

$$\begin{aligned} f(x) &= f(a) + (x-a)\frac{f'(a)}{1!} + (x-a)^2\frac{f''(a)}{2!} \\ &\quad + \ldots (x-a)^n \frac{f^{[n]}(a)}{n!} + \ldots \\ &= f(a) + (x-a)\frac{f'(a)}{1!} + (x-a)^2\frac{f''(a)}{2!} + \mathscr{R}_2. \end{aligned} \tag{30}$$

(Where $\mathscr{R}_2$ represents all the terms of higher order than 2, and a is a 'convenient' value at which to evaluate f).

This technique can be expected to matrix and vector functions. It involves the notion of gradient and Hessian. Now a vector function $f(\boldsymbol{x})$ is expressed as:

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + f(\boldsymbol{x}-\boldsymbol{a})^{\mathrm{T}}\nabla_{f(\boldsymbol{a})} + f(\boldsymbol{x}-\boldsymbol{a})^{\mathrm{T}}\nabla^2_{f(\boldsymbol{a})}f(\boldsymbol{x}-\boldsymbol{a}) + \mathscr{R}_2. \tag{31}$$

4.3 Iterative Minimization

A learning rule can be shown to converge to an optimum if it diminishes the value of the error function at each iteration. When the *gradient* of the error function can be evaluated, the gradient technique (or *steepest descent*) adjusts the weight vector by moving it in the direction opposite to the gradient of the error function. Formally, the correction for the $(n+1)$-th iteration is

$$\boldsymbol{w}_{[n+1]} = \boldsymbol{w}_{[n]} + \boldsymbol{\Delta} = \boldsymbol{w}_{[n]} - \eta\nabla_{f(\boldsymbol{w})} \tag{32}$$

(where $\nabla_{f(\boldsymbol{w})}$ is computed for $\boldsymbol{w}_{[n]}$).

As an example, let us show that for a linear heteroassociator, the Widrow–Hoff learning rule minimizes iteratively the squared error between target and output. The error function is

$$\begin{aligned} e^2 &= (t-o)^2 = t^2 + o^2 - 2to \\ &= t^2 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - 2t\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}. \end{aligned} \tag{33}$$

The gradient of the error function is

$$\frac{\partial e}{\partial \boldsymbol{w}} = 2(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})\boldsymbol{x} - 2t\boldsymbol{x} = -2(t - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})\boldsymbol{x}. \tag{34}$$

The weight vector is corrected by moving it in the opposite direction of the gradient. This is obtained by adding a small vector denoted $\boldsymbol{\Delta}_{\boldsymbol{w}}$ opposite to the gradient. This gives the following correction for iteration $n+1$:

$$\begin{aligned} \boldsymbol{w}_{[n+1]} &= \boldsymbol{w}_{[n]} + \boldsymbol{\Delta}_{\boldsymbol{w}} = \boldsymbol{w}_{[n]} - \eta\frac{\partial e}{\partial \boldsymbol{w}} \\ &= \boldsymbol{w}_{[n]} + \eta(t - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})\boldsymbol{x} \\ &= \boldsymbol{w}_{[n]} + \eta(t-o)\boldsymbol{x}. \end{aligned} \tag{35}$$

This gives the rule defined by Eqn. 9.

The gradient method works because the gradient of $\boldsymbol{w}_{[n]}$ is a first order Taylor approximation of the gradient of the optimal weight vector $\tilde{\boldsymbol{w}}$. It is a favorite technique in neural networks because the popular error backpropagation is a gradient technique.

Newton's method is a second order Taylor approximation, it uses the inverse of the Hessian of $\boldsymbol{w}$ (supposing it exists). It gives a better numerical approximation but necessitates more computation. Here the correction for iteration $n+1$ is

$$\boldsymbol{w}_{[n+1]} = \boldsymbol{w}_{[n]} + \boldsymbol{\Delta} = \boldsymbol{w}_{[n]} - (\boldsymbol{H}^{-1})(\nabla_{f(w)}) \tag{36}$$

(where $\nabla_{f(\boldsymbol{w})}$ is computed for $\boldsymbol{w}_{[n]}$).

See also: Artificial Neural Networks: Neurocomputation; Hebb, Donald Olding (1904–85); Neural Networks, Statistical Physics of; Perceptrons; Statistical Pattern Recognition

Bibliography

Abdi H 1994 *Les Réseaux de Neurones*. PUG, Grenoble, France

Abdi H, Valentin D, Edelman B 1999 *Neural Networks*. Sage, Thousand Oak, CA

Bishop C M 1995 *Neural Network for Pattern Recognition*. Oxford University Press, Oxford, UK

Ellacott S, Bose D 1996 *Neural Networks: Deterministic Methods of Analysis*. International Thomson Computer Press, London

Hagan M T, Demuth H B, Beale M 1996 *Neural Networks Design*. PWS, Boston

Haykin S 1999 *Neural Networks: A Comprehensive Foundation*, 2nd edn. Prentice Hall, Upper Saddle River, NJ

Reed R D, Marks R J 1999 *Neural Smithing*. MIT Press, Cambridge, MA

Ripley B D 1996 *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, MA

Rojas R 1996 *Neural Networks*. Springer-Verlag, New York

H. Abdi

Linear and Nonlinear Programming

Linear and nonlinear programming refers, broadly speaking, to the area of applied mathematics dealing with the following problem: find numerical values for a given set of variables so that they are *feasible* (i.e., they satisfy certain *constraints*, typically given by equalities or inequalities) and also a certain criterion, called *objective function*, which depends on such variables, is optimized, that is it attains its minimum value among all the combination of feasible variables. As an example, we describe in Section 3 the *diet problem*, consisting of finding a combination of ingredients which satisfy nutritional requirements at minimum cost. Unlike classical problems of applied mathematics, most of which originate in physics, linear and nonlinear programming problems generally lack solutions given by closed formulae, and must be solved through numerical procedures, called *algorithms*, performed on computers.

Different features of the problem data (e.g., linearity or nonlinearity of the objective function and the constraints) call for different methods in order to achieve efficient solution of the problem. The design of such algorithms, and the analysis of their performance,

are the backbone of the theoretical side of the area. On the practical side, the ability of these methods to solve very large problems (i.e., with a large number of variables) has allowed for the modeling of highly realistic and detailed real life situations, so that nowadays these methods are routinely applied to the day-to-day execution of complex tasks in a wide range of activities, like oil refineries, power stations, airlines and many others.

1. Linear Programming: History and Relevance

Linear Programming (LP) was one of the first achievements of Operations Research, which can be described as an interphase between mathematics, engineering, and economics, or as the formulation and implementation of mathematical models of complex systems. When the American army was confronted with the logistics involved in the World War II effort, it became clear that traditional allocation procedures were insufficient. On the other hand, computers, the first of which were being developed at the same time, offered the promise of actual solutions of many previously intractable problems. The US Air Force, which needed to deliver in time up to 20,000 different items to bases all around the world, created an Operations Research office, staffed with several bright and young mathematicians, many of whom continued their research on the same topic after the war in the Rand Corporation. One of them was G. Dantzig, who chose the LP problem, as formulated in Section 2, as an object deserving careful study. This choice was particularly succesful. Several special cases, such as transportation problems, or minimum cost flow problems, which correspond to LP problems where the matrix A of Section 2 has a particular structure (e.g., each a_{ij} is either 0 or 1) had been or were being studied at the time, and efficient methods for their solution were being introduced. For the transportation problem, the work in 1939 by the Russian mathematician L. V. Kantorovich (1960) clearly anticipated Dantzig's LP theory, but it became known in the West much later. The methods for these special instances were, however, rather different from each other and hardly generalizable, while, on the other hand, the LP format comprehended all of them, and seemed in fact rather universal, in the sense that most problems in economics could be cast in this setting, at the cost of approximating nonlinear functions by linear ones (see Section 4). At the same time, the LP problem was not too general, in the sense that it was possible to devise an efficient method for the numerical solution of all its instances, provided their data were not too large, using the then available computational equipment. This procedure, called the SIMPLEX method (see Section 3), was introduced by Dantzig in 1947 (Dantzig 1949), and it had an enormous impact in several research fields, among them economics.

At that time, the mathematization of economics was at a rather initial stage, and most mathematical results offered only new qualitative insights, rather than effective procedures for finding numerical solutions of economic problems, because of the complicated nature of the modeling tools, nonlinear in general. LP and SIMPLEX allowed actual computation of solutions for any problem where an economic agent (a firm, for instance) has to choose the most efficient among many possible courses of action, provided that all the relevant variables could be connected through linear relations, such as the input-output matrices proposed in the 1930s by V. Leontieff, which enter the LP formulation as the matrix A of Eqn. (5) below. This situation gave rise to a new area of economic research, called activity analysis (Koopmans 1951, Gale 1960) which involved many of the most distinguished economists at the time, such as T. Koopmans, K. Arrow, D. Gale, and others. The influence of LP en economics was reinforced by the introduction of the *duality theory* of LP. Shortly after Dantzig obtained his first results on LP, he showed them to one of the leading mathematicians of the time, von Neumann, who was then developing another new branch of applied mathematics, namely game theory (Von Neumann and Morgenstern 1944). Observing immediately that a certain class of games fitted the LP format, Von Neumann suggested the extension to LP of a game theoretic result, giving birth to LP duality, which, given an LP problem (e.g., Eqns. (2)–(3) below), associates with it another problem, called *dual*, which shares exactly the same data c, b, and A, but arranged in a different way. Duality plays a very significant role in economic models, where the variables of the dual problem, called shadow prices, have an interesting economic interpretation. The Dantzig–Wolfe decomposition method (Dantzig and Wolfe 1960), which was a SIMPLEX-based method for solving certain special instances of LP problems, but which could also be seen as a model of a central planning procedure combined with autonomous decision by lower level agents, reinforced this impact of LP on economics, which produced perhaps, as a negative side effect, exaggerated attempts to force the modeling of nonlinear economic phenomena as linear ones.

The impact of LP on mathematics was also significant, and contributed to reinforcing the status of the new computational mathematics viz-a-viz the classical branches (e.g., analysis, geometry, algebra). Its success resulted from its ability to offer SIMPLEX as an efficient computational procedure and the LP duality theory as a theoretical counterpart. It should be mentioned that the mathematical tools required by the LP theory are rather elementary, and the theory could easily have been developed more than a century before it appeared. In fact, for the cases in which no inequalities are present (e.g., Eqns. (4)–(5) without Eqn. (6), but with a nonlinear f instead of cx) the theory for the resulting equality constrained mini-

mization problem had been started by Lagrange at the beginning of the nineteenth century, and thousand of papers had been written on it, while only a handful of them had been devoted to the subject of linear inequalities, like Eqn. (3) or Eqn. (6), before 1945 (Dantzig 1963). A very likely reason for this oversight lies in the fact that the motivation for the study of minimization problems had been, for centuries, mainly research in physics, studying systems which move along deterministic trajectories, which are usually expressed by equalities. Inequalities, on the other hand, are typical of economic situations where agents are constantly confronting options for possible action rather than following fixed trajectories determined by initial conditions. The low level of formalization of economics up to the 1940s, combined with the impossibility of developing at that time numerical methods, in the absence of computers, kept the field of linear inequalities outside mainstream mathematics for a century and a half.

2. *Linear Programming: Definition and Example*

The LP problem is formulated as follows. The data consists of

(a) a string c and n real numbers: $c_1, \ldots, c_n$ (an n-vector c);

(b) a string b of m real numbers: $b_1, \ldots, b_m$ (an m-vector b), and

(c) a rectangular array A of mn real numbers (an $m \times n$-*matrix* A):

$$\begin{array}{l} a_{11}, a_{12}, \ldots, a_{1n} \\ a_{21}, a_{22}, \ldots, a_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}, a_{m2}, \ldots, a_{mn}. \end{array}$$

The unknowns to be determined form an n-vector $x = (x_1, \ldots, x_n)$. The *objective function* f is defined as

$$f(x) = c_1x_1 + c_2x_2 + \cdots + c_nx_n.$$

The *feasible region* F consists of all the n-vectors $x = (x_1, \ldots, x_n)$ such that

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leqslant b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \leqslant b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leqslant b_m. \end{array} \tag{1}$$

The problem consists of finding an n-vector $x^* = (x_1^*, \ldots, x_n^*)$, belonging to F, which minimizes $f(x)$ in F, meaning that the m inequalities, Eqn. (1), are satisfied with x_i^* instead of x_i $(1 \leqslant i \leqslant n)$, and that $c_1x_1^* + c_2x_2^* + \ldots c_nx_n^* \leqslant c_1x_1 + c_2x_2 + \ldots c_nx_n$ for all n-vector $x = (x_i, \ldots, x_n)$ which also satisfies the m inequalities in Eqn. (1). In short mathematical notation, this is written as

$$\min cx \tag{2}$$

$$\text{s.t. } Ax \leqslant b. \tag{3}$$

An LP problem presented as Eqns. (2)–(3) is said to be in *canonical* form. There are other forms which are *equivalent* to it, in the sense that a solution of the canonical form allows one to obtain in an immediate way a solution of the equivalent form and conversely. One such form consists of inverting all the inequalities in (1) ('$\geqslant$' instead of '$\leqslant$'). The most important equivalent form is the *standard* one, where the unknowns are required to be nonnegative $(x_1 \geqslant 0, \ldots x_n \geqslant 0)$ and the inequalities in Eqn. (1) are changed into equalities, which is denoted in short notation as

$$\min cx \tag{4}$$

$$\text{s.t. } Ax = b \tag{5}$$

$$x \geqslant 0. \tag{6}$$

We present next as an example a particular instance of the so called *diet problem*, which is a case of the LP in canonical form with reversed inequalities. The manager of a poultry ranch has three alternative chicken food products with prices c_1, c_2, and c_3 per pound. Let a_{11}, a_{12}, and a_{13} be the protein content of a pound of products 1, 2, and 3 respectively, and a_{21}, a_{22}, and a_{23} the corresponding mineral contents per pound. If b_1 and b_2 are the minimal amounts of protein and minerals required for every chicken in the ranch, then the inequalities

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \geqslant b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \geqslant b_2 \end{array}$$

mean that an amount of x_1 pounds of product 1, x_2 pounds of product 2, and x_3 pounds of product 3 satisfy such minimal nutritional requirements, and a solution $x^* = (x_1^*, x_2^*, x_3^*)$ of the diet problem

$$\min cx$$

$$\text{s.t. } Ax \geqslant b_1$$

indicates the amounts of each product which satisfy the nutritional requirements at the lowest possible cost.

3. Linear Programming: SIMPLEX and other Computational Methods

The first significant result in LP theory concerns the feasible region F, which, for a problem in standard form, consists of those n-vectors x which satisfy Eqns. (5)–(6). When $n = 3$, such set is a polyhedron in the usual sense. For larger values of n, it is still what mathematicians call an n-dimensional polyhedron, and the result says that if the LP problem has solutions, then at least one *vertex* of the polyhedron F is a solution. Also, assuming that $m < n$ and that nothing in Eqn. (5) is superfluous, vertices of F are precisely the n-vectors obtained as follows: choose $n-m$ components of the vector x and give them the value 0. Look at Eqn. (5) as a system of m equations but now in m unknowns, namely the components of x which have not been set to 0, or, in other words, form an $m \times m$-matrix B with the m colums of the matrix A correspoding to the nonzeroed variables, and look at the system $Bx = b$. If such a system has a solution $\tilde{x}$ with nonnegative components, the resulting vector x (formed by the components of $\tilde{x}$ plus the zeroes in the remaining components) is a vertex of F. These facts provide immediately the following 'brute force' procedure for solving LP problems: consider consecutively each one of the possible choices of m out of the n original colums of A. Check whether each choice gives raise to a vertex $\bar{x}$, and if it does, evaluate the objective function at $\bar{x}$, that is, compute $f(\bar{x}) = \sum_{j=1}^{n} c_j \bar{x}_j$. Compare all the values so obtained and pick up the vertex x^* with the lowest value for $f(x^*)$. The drawback is that the number of choices of m objects out of n grows very fast with m and n. For instance, if we take an LP problem with n unknowns and the following $2n$ inequalities

$$x_i \leqslant u_i \quad (1 \leqslant i \leqslant n) \tag{7}$$

$$-x_i \leqslant -l_i \quad (1 \leqslant i \leqslant n) \tag{8}$$

meaning just that each variable x_i must be between a lower bound l_i and an upper one u_i, the resulting polyhedron F has 2^n vertices, resulting from all the possible choices of $x_i = l_i$ or $x_i = u_i$ for each of the n variables. Now, 2^n grows very fast: for $n = 100$, checking all the vertices exceeds the capabilities of any existing computer. The SIMPLEX methods, instead of checking all the vertices, proceeds as follows. It constructs one vertex, say $\bar{x}$, and checks whether it is or not a solution (which is a rather 'easy' task). If it is not, it does not proceed to just any other vertex, but to one which is *adjacent* to $\bar{x}$, which mathematically means that it shares with $\bar{x}$ all but one of its $n-m$ zero components. There are at most n adjacent vertices (corresponding to all possible choices of the nonshared column) and at least one, say $\hat{x}$, has an objective value $f(\hat{x}) < f(\bar{x})$. Then $\hat{x}$ is taken as the new candidate, and, if it is not a solution, then the process is repeated with vertices adjacent to $\hat{x}$, and so on, until a solution is found, which is guaranteed to happen. How long does this procedure take? More than 50 years of practice show that, almost always, the number of visited vertices is not too high; generally less that $m \log n$, which is negligible as compared to, for example 2^n (rigorous versions of this statement on the average performance of SIMPLEX can be found in Borgwardt 1987 and Smale 1983). However, Klee and Minty (1972) found a highly artificial example, with a carefully chosen objective function f, and a feasible set F obtained as a slight variation of the one given by Eqns. (7)–(8), for which SIMPLEX goes through all the 2^n vertices before finding the solution, which is just the last one.

Despite Klee and Minty's example, SIMPLEX was absolutely superior to any other known procedure for solving LP problems, in terms of computational performance. Thus, the following theoretical issue was posed: does it exist a method for LP which is *never* too slow, for instance such that the number of arithmetic operations required to solve any particular LP problem does not exceed a fixed power of the size of the problem (understood as some measure of the size of c, b, and A)? SIMPLEX is not up to this requirement, because it needs more than 2^n operations to solve Klee and Minty's example, and 2^n exceeds any fixed power of n, which can be taken as a measure of the size of this specific LP problem. Methods with such property were called *polynomial*. The fact that no polynomial method for LP had been found in the first 30 years after the introduction of SIMPLEX, made most mathematicians think that no such method could ever be devised, and therefore it was a surprise when Kachyan presented such a procedure in 1979 (Kachyan 1979). Unfortunately, Kachyan's method, called the *elipsoidal* one, turned out to be far worse than SIMPLEX in almost all problems (though, of course, much better in the very rare cases in which SIMPLEX takes almost forever, like in Klee and Minty's example). Thus, the elipsoidal method aroused only theoretical interest, and did not replace SIMPLEX in any 'real life' application. Nevertheless, existence of a polynomial method (albeit inefficient in the average) encouraged research attempting to find a procedure which would be both polynomial and efficient in practice (i.e., 'good' both in the average cases and in the worst ones). Such a method was discovered by Karmarkar in 1984 (Karmarkar 1984), and extensive progress has been achieved in this class of methods for LP (called *interior point* ones) since the mid-1980s (Roos et al. 1997). They differ from SIMPLEX in the fact that, instead of considering only vertices, which lie in the boundary of the feasible polyhedron F, all the intermediate vectors, which are tested as candidates for solutions, are taken in the interior of this polyhedron. Thus, these methods fully neglect the combinatorial nature of LP problems (i.e., the facts that for an LP problem with solutions there exists a vertex which is a solution, and that the number of vertices is finite). Interior point methods are

quite efficient, but none of the many of those which have been proposed turned out to be universally superior to all others (nor to SIMPLEX, for that matter). Currently, the situation in LP, regarding computational methods, is similar to most other applied mathematics areas: instead of a method universally superior to all others, as was the case for LP between 1947 and 1984, a variety of methods are available, some of which are clearly superior, in terms of computational performance, for some specific cases of LP problems, but inferior to at least another one for other types of LP problems.

4. Nonlinear Programming: Theory and Relevance

As soon as LP was born, it was clear to its inventors and practitioners that linearity of the data severely limited the ability to model most real life systems, and therefore, already in the late 1940s, attention was given to the nonlinear programming problem (NLP), which generalizes LP through the admission of rather arbitrary functions, instead of linear ones, both in the objective and in the constraints, so that its mathematical formulation can be given as

$$\min f(x_1, \ldots, x_n) \tag{9}$$

$$\text{s.t. } g_1(x_1, \ldots, x_n) \leqslant 0, \ldots, g_m(x_1, \ldots, x_n) \leqslant 0. \tag{10}$$

Here, both f and the g_i's ($1 \leqslant i \leqslant m$) are functions which assign to each n-vector $x = (x_1, \ldots, x_n)$ a real number denoted as $f(x_1, \ldots, x_n)$ and $g_i(x_1, \ldots, x_n)$ respectively. This assignment is generally nonlinear, in the sense that it might involve, for example powers of the variables, $x_1, \ldots, x_n$ larger than 1.

The first theoretical difference with LP arises from the fact that generally it is impossible to devise procedures able to produce, after a finite number of arithmetic operations, an exact solution of the problem. Thus, the general feature of an NLP method is that it generates a sequence of n-vectors (to be denoted as $x^0, \ldots, x^k, \ldots$), starting from a given n-vector x^0. In general, none of the x^k's will be a solution, but for large enough k, each component of the vector x^k should be very close to the corresponding component of a solution of the problem. An exception to this situation is the special case of quadratic programming (Boot 1964), in which all the g_i's are linear (i.e., as in Eqn. (1)) and the objective function f is quadratic, that is

$$f(x_1, \ldots, x_n) = \sum_{j=1}^{n} \sum_{k=1}^{n} d_{jk} x_j x_k + \sum_{j=1}^{n} c_j x_j \tag{11}$$

where the real numbers d_{jk} form an $n \times n$-matrix D. For this case there exist several extensions of the SIMPLEX method for LP, which produce exact solutions, but with the following limitation: each one of these methods works for a certain class of matrices D; in fact, no known method is able to provide exact solutions for an arbitrary quadratic programming problem.

Another very important theoretical issue refers to necessary and/or sufficient conditions on an n-vector x^* to be a solution of Eqns. (9)–(10). In the case of equality constraints, that is, when all the inequalities in Eqn. (10) are replaced by equalities—such a condition was known in the early nineteenth century—if we denote as $\nabla f(x)$ the n-vector whose j-th component is the derivative of f with respect to its j-th variable at the n-vector x, and we use the same definition for $\nabla g_i(x)$ ($1 \leqslant i \leqslant m$), then the so called *Lagrangian* condition states that

$$\nabla f(x) = u_1 \nabla g_1(x) + \cdots + u_m \nabla g_m(x) \tag{12}$$

where the real numbers $u_1, \ldots, u_m$ are called Lagrangian multipliers. In fact, if x solves Eqns. (9)–(10), then Lagrange multipliers do exist. Conversely, existence of Lagrange multipliers for a given n-vector x^*, together with an additional condition involving second order derivatives of f and the g_i's at x^*, guarantee that x^* is a *local solution* of the problem, meaning that $f(x^*) \leqslant f(x)$ for all x satisfying Eqn. (10) and close enough to x^* (i.e., such that all the differences $x_i^* - x_i$ are small enough). The extension of the Lagrangian condition to the case of inequality constraints was the first theoretical achievement of NLP. The resulting conditions, called Karush–Kuhn–Tucker ones, were published by Kuhn and Tucker (1951), and had been anticipated in Karush's MSc dissertation. They state that the Lagrange multipliers, besides satisfying the Lagrangian condition, Eqn. (12), must be nonnegative, and also that $u_i = 0$ if $g_i(x) < 0$. Also, an additional condition, called *constraint qualification*, must be imposed on the feasible region, that is, on the set of n-vectors which satisfy Eqn. (10). The 150 years between the establishment of the conditions for the equality and the inequality constrained cases is again just a consequence of classical mathematicians' lack of interest in inequalities, as discussed above. A third important theoretical issue was the extension to NLP of the duality theory of LP. This was achieved only in the *convex* case (Rockafellar 1970) that is, when both f and all the g_i's are convex. We recall that a function h is convex when

$$h(\alpha x + (1-\alpha)y) \leqslant \alpha h(x) + (1-\alpha) h(y)$$

for all n-vectors x and y and all real number α between 0 and 1.

Though NLP is used extensively in many real-life applications, it is much less widespread than LP. This is due to the facts that NLP lacks an universal method, like SIMPLEX for LP, so that the procedure to be used must be chosen according to the nature of the specific problem, and that NLP methods in general do not work as black boxes, fed with the problem data and handing over the solution, but rather must be fine

tuned, choosing appropriate parameters, properly interpreting the output of the method, etc. Thus, an expert in Operations Research, unavailable in most moderate sized institutions, is required for successful implementation and solution of NLP problems. On the other hand, the impact of NLP on mathematics itself has been much deeper than the one of LP. Important results in several classical areas, like convex analysis and variational analysis, have been drastically improved as a consequence of the development of NLP theory (see Rockafellar 1970, Rockafellar and Wets 1998).

5. Nonlinear Programming: Methods

The simplest case to be considered is the *unconstrained* one, where the conditions in Eqn. (10) are absent, and the problem consists just of minimizing the function f. Most methods for this case generate sequences of n-vectors $x^0, \ldots, x^k, \ldots$ which approximate a local solution of the problem. The simplest ones, which were known in the nineteenth century, though they only became significant with the advent of computers, are *Cauchy's* methods, or *steepest descent*, and *Newton*'s method. Newton's method generally requires less iterations than Cauchy's for attaining a certain precision (i.e., a sufficiently good approximation x^k will be achieved with a lower value of k) but each of its steps is more time consuming (see Mangasarian 1969). More modern methods for unconstrained minimization include quasi-Newton ones, which combine the best features of Cauchy's and Newton's methods (Dennis and Moré 1977), and trust region methods (Dennis and Schnabel 1983).

In the *constrained* case, the constraints $g_i(x) \leqslant 0$ $(1 \leqslant i \leqslant m)$ are present, and most methods reduce the original problem Eqns. (9)–(10) to a sequence of simpler problems, each one of which has to be solved by another appropriate procedure. Most methods can be seen as belonging to one of the three following families:

(a) Feasible methods: each step consists of two stages. In the first one the constraints are ignored and an unconstrained method is used to reduce the objective f; in the second one only the constraints in Eqn. (10) are taken into account (see Avriel 1976).

(b) Sequential quadratic programming: in these methods, at each step the problem is approximated by a quadratic programming problem, which can then be solved with one of the exact methods mentioned in section 5 (see Gomes et al. 1999).

(c) Penalty methods: In this case, the original problem Eqns. (9)–(10) is replaced by a sequence of unconstrained problems, each one of which is then solved with some method for unconstrained optimization. Within this family of methods, special attention has been given to the so called augmented Lagrangian methods (see Bertsekas 1982), devised to find at the same time approximate solutions to the problem and approximate Lagrange multipliers, as defined by (12).

Efficient implementations of any of the methods above usually require knowledge of the first and second derivatives of f and the g_i's. Also, excepting in the convex case, the generated n-vector approximate in general just local solutions, as discussed in Section 5. Recent research in NLP has attempted to overcome these two limitations. Nonsmooth optimization is devoted to devising methods which work for problems whose data functions do not have derivatives (Clarke 1983), while global optimization addresses the issue of finding *global* solutions, that is, n-vectors x^* satisfying Eqn. (10) and such that $f(x^*) \leqslant f(x)$ for all other n-vector x which also satisfies Eqn. (10), and not only for n-vectors close to x^* (Horst and Tuy 1993).

See also: Linear Algebra for Neural Networks; Linear Hypothesis

Bibliography

Avriel M 1976 *Nonlinear Programming, Analysis and Methods*. Prentice Hall, Upper Saddle River, NJ

Bertsekas D P 1982 *Constrained Optimization and Lagrange Multipliers*. Academic Press, New York

Boot J C G 1964 *Quadratic Programming*. North Holland, Amsterdam

Borgwardt K H 1987 *The Simplex Method: A Probabilistic Analysis*. Springer, Berlin

Clarke F H 1983 *Optimization and Nonsmooth Analysis*. Wiley, New York

Dantzig G B 1949 Programming of independent activities II, mathematical model. *Econometrica* **17**: 200–11

Dantzig G B 1963 *Linear Programming and Extensions*. Princeton University Press, Princeton, NJ

Dantzig G B, Wolfe P 1960 Decomposition principle for linear programs. *Operations Research* **8**: 101–11

Dennis J E, Moré J J 1977 Quasi-Newton methods: Motivation and theory. *SIAM Review* **19**: 46–89

Dennis J E, Schnabel R B 1983 *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall, Upper Saddle River, NJ

Gale D 1960 *The Theory of Linear Economic Models*. McGraw-Hill, New York

Gomes F M, Maciel M C, Martínez J J 1999 Nonlinear programming algorithms using trust regions and augmented Lagrangians with nonmonotone penalty parameters. *Mathematical Programming* **84**: 161–200

Hillier F S, Lieberman G J 1980 *Introduction to Operations Research*. Holden Day, San Francisco

Horst R, Tuy H 1993 *Global Optimization. Deterministic Approaches*. Springer, Berlin

Kantorovich L V 1960 Mathematical methods in the organization and planning of production. *Management Science* **6**: 366–422

Karmarkar N A 1984 A new polynomial time algorithm for linear programming. *Combinatorica* **4**: 373–95

Khachiyan L G 1979 A polynomial algorithm in linear programming. *Doklady Akademiia Nauk SSSR* **244**: 1093–6

Klee V, Minty G J 1972 How good is the simplex algorithm. In: Shisha O (ed.) *Inequalities-III*. Academic Press, New York pp. 159–75

Koopmans T (ed.) 1951 *Activity Analysis of Production and Allocation*. Wiley, New York
Kuhn H W, Tucker A W 1951 Nonlinear programming. *Econometrica* **19**: 50–1
Mangasarian O L 1969 *Nonlinear Programming*. McGraw-Hill, New York
Rockafellar R T 1970 *Convex Analysis*. Princeton University Press, Princeton, NJ
Rockafellar R T, Wets R J-B 1998 *Variational Analysis*. Springer, Berlin
Roos C, Terlaky T, Vial J-Ph 1997 *Theory and Algorithms for Linear Optimization: An Interior Point Approach*. Wiley, New York
Smale S 1983 On the average number of steps of the simplex method of linear programming. *Mathematical Programming* **27**: 241–62
Von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

A. N. Iusem

Linear Hypothesis

1. Introduction

The term 'linear hypothesis' is often used interchangeably with the term 'linear model.' Statistical methods using linear models are widely used in the behavioral and social sciences, e.g., regression analysis, analysis of variance, analysis of covariance, multivariate analysis, time series analysis, and spatial data analysis. Linear models provide a flexible tool for data analysis and useful approximations for more complex models.

A common object of linear modeling is to find the most precise linear model that explains the data, to use that model to predict future observations, and to interpret that model in the context of the data collection. Traditionally, analysis of variance models have been used to analyze data from designed experiments while regression analysis has been used to analyze data from observational studies but the techniques of both analysis methods apply to both kinds of data. See also *Experimental Design: Overview* and *Observational Studies: Overview*.

2. Definition

The linear hypothesis is that the mean (average) of a random observation can be written as a linear combination of some observed predictor variables. For example, Coleman et al. (1996) provides observations on various schools. The dependent variable y consists of the average verbal test score for sixth-grade students. The report also presents predictor variables. A composite measure of socioeconomic status x_1 is based on father's and mother's education, family size and intactness, home items, and percent of fathers who are white collar. Staff salaries per pupil is x_2. The average score on a verbal test given to the school's teachers is x_3. Denoting different schools using the subscript i and the mean of y_i by m_i, a linear hypothesis states that for some unknown numbers (parameters) β_0, β_1, β_2, and β_3,

$$m_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \beta_3 x_{i3}$$

Mosteller and Tukey (1977, pp. 326, 566) and Mosteller et al. (1983, pp. 408–20) give excerpts and analysis of the data.

In other applications, the predictors only identify whether an observation is in some group. For example, in 1978 observations y_{ij} were collected on the age at which people in Albuquerque committed suicide, see Koopmans (1987, p. 409). Here i is used to identify the person's group membership (Hispanic, Native American, non-Hispanic Caucasian), and j identifies individuals within a group. The three categories are taken to be mutually exclusive for the present discussion (although the US government now allows for individuals to identify themselves with multiple races in various surveys and the decennial census). We can define group identifier predictor variables. Let δ_{1i} take the value 1 if an individual belongs to group 1 (Hispanic) and 0 otherwise, with similar predictors to identify other groups, say δ_{2i} and δ_{3i} for Native Americans and non-Hispanic Caucasians. Note that the predictor variables do not depend on the value of j identifying individuals within a group. Denoting the mean of y_{ij} by m_{ij}, a linear hypothesis states that for some unknown parameters μ, α_1, α_2, α_3,

$$m_{ij} = \mu + \alpha_1 \delta_{1i} + \alpha_2 \delta_{2i} + \alpha_3 \delta_{3i}$$

Since two of the δs are always zero, this model is often written more succinctly as

$$m_{ij} = \mu + \alpha_i$$

A linear hypothesis is usually combined with other assumptions about the observations y. Most commonly, the assumptions are that the observations are independent, have the same (unknown) variance σ^2, and have normal (Gaussian) distributions. For the two examples, these assumptions are written

$$y_i \text{ indep. } N(m_i, \sigma^2) \quad \text{and} \quad y_{ij} \text{ indep. } N(m_{ij}, \sigma^2)$$

where, for example, $N(m_i, \sigma^2)$ indicates a normal distribution with mean m_i and variance σ^2. Incorporating these additional assumptions, the linear

hypothesis becomes a full-fledged linear model, traditionally written

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,\ \varepsilon_i \text{ indep. } N(0, \sigma^2) \tag{1}$$

and

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ \varepsilon_{ij} \text{ indep. } N(0, \sigma^2) \tag{2}$$

where the εs are unobservable random errors.

Both theoretical and computational work with linear hypotheses and linear models is facilitated by the use of matrices. If $i = 1, 2, 3, 4$, the model (1) can be written as the equality of two 4×1 matrices

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \varepsilon_1 \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{33} + \varepsilon_2 \\ \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \varepsilon_3 \\ \beta_0 + \beta_1 x_{41} + \beta_2 x_{42} + \beta_3 x_{43} + \varepsilon_4 \end{bmatrix}$$

Using properties of matrix algebra, this can be rewritten as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

Similarly, model (6) for $i = 1, 2, 3$, $j = 1, 2$ can be written using δ as

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

In general, a linear model is traditionally written using matrices as

$$Y = X\beta + e \tag{3}$$

where Y is a $n \times 1$ matrix containing the observable random variables y, X is a $n \times p$ matrix containing observed predictors, β is a $p \times 1$ matrix containing unobservable parameters, and e is a $n \times 1$ matrix containing unobservable random errors. To have a full-fledged model, assumptions must be made about the distribution of e. One standard assumption is that the mean value of each element of e is 0. Denote the $n \times 1$ matrix of mean (expected) values as $\mathrm{E}(e)$. This assumption is written

$$\mathrm{E}(e) = 0$$

The corresponding linear hypothesis involves the mean values of the entries of Y, i.e., $\mathrm{E}(Y)$. The linear hypothesis is that

$$\mathrm{E}(Y) = X\beta$$

for some matrix β. Another standard assumption about e is that the covariance matrix, the $n \times n$ matrix of variances and covariances between the ε_is or equivalently between the y_is is

$$\mathrm{Cov}(e) = \mathrm{Cov}(Y) = \sigma^2 I$$

where I indicates the $n \times n$ identity matrix that has 1s on the diagonal and 0 elsewhere.

An alternative way to write model (3) in terms of each y_i using some matrices is

$$y_i = x_i'\beta + \varepsilon_i$$

$i = 1, \ldots, n$ where β is as before and x_i' is the ith row of the X matrix.

See also *Analysis of Variance and Generalized Linear Models* and *Linear Hypothesis: Regression (Basics)*.

3. *Estimation*

The unknown parameters of the linear model are the elements of β, the βs, and the variance of an observation, σ^2; they must be estimated from the data. The β_js are generally estimated using least squares estimation. For given data Y, the least squares estimates are the values of β_j that minimize

$$\sum_{i=1}^{n} (y_i - x_i'\beta)^2 \quad \text{or equivalently} \quad (Y - X\beta)'(Y - X\beta)$$

where the prime on $(Y - X\beta)'$ indicates the matrix transpose of $(Y - X\beta)$. (The transpose of a matrix turns the columns of the original matrix into the rows of a new matrix.)

Under the standard assumptions, least squares estimates have optimal statistical properties. If the ε_is are independent $N(0, \sigma^2)$, then the least squares estimates provide maximum likelihood estimates (MLEs) and minimum variance unbiased estimates (MVUEs). A MVUE is an estimate that has less variability than any other unbiased estimate. If we drop the assumption of normality but retain independence (actually 0 covariances) and equal variances, least squares

estimates are best linear unbiased estimates (BLUEs). A BLUE is an estimate based on a linear combination of the observations that has less variability than any other unbiased estimate based on a linear combination of the observations.

In matrix terms, the least squares estimates, say, $\hat{\beta}$, are obtained as solutions to the normal equations

$$X'X\beta = X'Y$$

Typically, in regression problems, the unique solution is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

In general, if $X'X$ is singular, so that $(X'X)^{-1}$ does not exist, an infinite number of solutions exist, giving an infinite number of least squares estimates. Any one of these estimates is just as valid as any other, but fortunately, for all practical purposes, the estimates are unique. This means that whenever a linear parameter $\lambda'\beta$ is identifiable in the model (estimable), the least squares estimate of the parameter is uniquely defined. See also *Statistical Identification and Estimability* and *Analysis of Variance and Generalized Linear Models*.

The other unknown parameter to be estimated is σ^2, the variance of an observation. To estimate the variance, first rearrange the model into

$$\varepsilon_i = y_i - x_i'\beta \quad \text{or} \quad e = Y - X\beta$$

Then estimate β to get the residuals

$$\hat{\varepsilon}_i = y_i - x_i'\hat{\beta} \quad \text{or} \quad \hat{e} = Y - X\hat{\beta}$$

The sum of the squared residuals is called the sum of squares error,

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

The degrees of freedom error (dfE) is n minus the rank of the matrix X, i.e.,

$$df\text{E} = n - r(X)$$

The rank of X is the number of free parameters that can be fitted to the model by least squares. The variance σ^2 is estimated by the mean squared error,

$$MSE = \frac{SS\text{E}}{df\text{E}}$$

See also *Estimation: Point and Interval*.

4. Testing

The basic idea of testing is to create some model for the process of generating the data, and evaluate whether the observed data are consistent with that model. For example, such a model might be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

One can think of this model has having four parts: (a) ε_i are independent, (b) ε_i have $\text{E}(\varepsilon_i) = 0$, i.e.,

$$\text{E}(y_i) = m_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

(c) ε_i have common variance σ^2, (d) ε_i are normally distributed. Data that are inconsistent with the model may occur whenever any of these four parts is inappropriate.

In the context of testing the linear hypothesis, it is assumed that these four specifications are correct and a further refinement of the model is then imposed by placing an additional restriction on

$$\text{E}(y_i) = m_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

The test is usually thought of as a test of this additional restriction, generally called the null hypothesis. For example, the additonal restriction could be the null hypothesis H_0: $\beta_1 = \beta_3 = 0$. In the sixth grade verbal test example, this hypothesis indicates that the variables x_1, socioeconomic status, and x_3, teachers mean verbal test score, have no ability to help predict sixth graders' mean verbal test scores. The test of whether the data are consistent with a model that incorporates $\beta_1 = \beta_3 = 0$ is based on assuming that the null hypothesis and items (a) through (d) are all correct. If assumptions (a) through (d) are correct, then data that are inconsistent with the model suggest that either $\beta_1 \neq 0$ or $\beta_3 \neq 0$ or both, i.e., that the null hypothesis is not true. However, in reality, data that are inconsistent with the model can result if any of the assumptions are violated, i.e., any of the four itemized assumptions or the null hypothesis. Because of this, if one is interested in testing only the null hypothesis, it is crucial to do everything possible to validate the assumptions embodied in items (a) through (d). See also *Hypothesis Testing in Statistics* and *Significance, Tests of*.

There are two ways to think about testing a linear hypothesis. One is to think in terms of testing models and the other is to think in terms of testing parameters.

4.1 Models

In testing models one assumes the validity of a (full) model and tests to see if some smaller (reduced) model

is consistent with the data. For example, with the verbal test score data a full model might be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

incorporating all three predictors. The reduced model may incorporate the idea that the socioeconomic and mean teacher verbal score variables x_{i1} and x_{i3} are not important, yielding

$$y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

In this context, the null hypothesis is that the reduced model provides an adequate explanation for the data. In any particular test, the full model need not incorporate all of the variables that have been measured. It need only be a model that is plausibly correct.

Both the full model and the reduced model have sums of squares error, degrees of freedom error, and mean squared errors associated with them. The reduced model is rejected as being inconsistent with the data if the following F statistic is too large,

$$F = \frac{SSE(\text{Red}) - SSE(\text{Full})}{dfE(\text{Red}) - dfE(\text{Full})} / MSE(\text{Full})$$

If the reduced model is true, then both the full model and the reduced model are true, so both MSE(Red) and MSE(Full) are estimates of σ^2. It can then be shown that $(SSE(\text{Red}) - SSE(\text{Full}))/(dfE(\text{Red}) - dfE(\text{Full}))$ is also an estimate of σ^2, so both the numerator and the denominator of F are estimates of σ^2, and the ratio should be about 1. Of course, there is sampling variability associated with the F statistic. The theoretical F distribution describes the usual distribution of values one will encounter when making such calculations. It depends on two parameters: $dfE(\text{Red}) - dfE(\text{Full})$ degrees of freedom for the numerator and dfE(Full) degrees of freedom for the denominator.

On the other hand, if the reduced model is an inadequate explanation of the data, it can be shown that $(SSE(\text{Red}) - SSE(\text{Full}))/(dfE(\text{Red}) - dfE(\text{Full}))$ estimates σ^2 plus a positive number, so that if the reduced model is bad, the F statistic should be larger than 1. If the observed value of the F statistic is so much larger than 1 as to be inconsistent with the theoretical F distribution, one can reasonably conclude that the reduced model is an inadequate explanation of the data, i.e., one rejects the null hypothesis.

4.2 Parameters

The other way to perform tests in the linear hypothesis is to specify a null hypothesis in terms of the parameters. For example, if we assume the verbal test score model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

we might be interested in testing the null hypothesis H_0: $\beta_1 = \beta_3 = 0$ in which socioeconomic status and teachers mean verbal test score have no ability to predict. One way to test this hypothesis is to identify the reduced model associated with it, i.e.,

$$y_i = \beta_0 + 0x_{i1} + \beta_2 x_{i2} + 0x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

or

$$y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

For a more complicated example, suppose two previous studies had suggested the hypothesis H_0: $\beta_1 + \beta_3 = 5$ and $\beta_2 = 0$. We can rewrite the hypothesis as H_0: $\beta_1 = 5 - \beta_3$ and $\beta_2 = 0$, and create a reduced model by substitution,

$$y_i = \beta_0 + (5 - \beta_3)x_{i1} + 0x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i, \text{ indep. } N(0, \sigma^2)$$

Rearranging terms gives a reduced model

$$y_i - 5x_{i1} = \beta_0 + \beta_3(x_{i3} - x_{i1}) + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

This is just a new linear model with a new dependent variable $y_i - 5x_{i1}$ and one predictor variable $(x_{i3} - x_{i1})$. One can compute SSE(Red) and dfE(Red) in the usual way for this new model, and simply substitute these quantities into the usual formula for the F statistic to get the test.

If the data are consistent with the reduced model, we have a more precise model than the one we started with. While testing provides no assurance that the reduced model is correct, the data are at least consistent with it. This more precise model can be investigated for the validity of its predictions and its usefulness in explaining the data collection process.

In addition, (estimable) parametric hypotheses can be tested directly using matrices, without identifying the reduced model. For example, the hypothesis H_0: $\beta_1 + \beta_3 = 5$ and $\beta_2 = 0$ can be written in matrix form by creating a matrix Λ' that, when multiplied by β, isolates $\beta_1 + \beta_3$ and β_2. The hypothesis is

$$\Lambda'\beta \equiv \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_4 \end{bmatrix}$$

$$= \begin{bmatrix} \beta_1 + \beta_3 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix} \equiv d$$

In matrix terms the F statistic for a null hypothesis $H_0 : \Lambda'\beta = d$ is

$$F = \frac{(\Lambda'\hat{\beta} - d)'[\Lambda'(X'X)^{-}\Lambda]^{-}(\Lambda'\hat{\beta} - d)}{r(\Lambda)MSE} \tag{4}$$

where $\hat{\beta}$ is a least squares estimate of β, $r(\Lambda)$ is the rank of the matrix Λ, and A^{-} is a generalized inverse of the matrix A. If A^{-} exists, it is the unique generalized inverse. The reduced model test and this direct test give the same F statistic and use the same degrees of freedom.

Confidence regions can be constructed using the F distribution. For the example, we can construct a simultaneous confidence region for $\beta_1 + \beta_3$ and β_2 as follows. Let $F(0.95, r(\Lambda), n - r(X))$ be the number so that with probability 0.95

$$\frac{(\Lambda'\hat{\beta} - \Lambda'\beta)'[\Lambda'(X'X)^{-}\Lambda]^{-}(\Lambda'\hat{\beta} - \Lambda'\beta)}{r(\Lambda)MSE} \leqslant F(0.95, r(\Lambda), n - r(X))$$

Here, $MSE(X'X)^{-}$ is an estimate of the covariance matrix of $\hat{\beta}$ and $MSE[\Lambda'(X'X)^{-}\Lambda]$ is an estimate of the covariance matrix of $\Lambda'\hat{\beta}$. After substituting the observed values of $\Lambda'\hat{\beta}$ and MSE into the inequality, the only thing that is unknown in this inequality is the value of $\Lambda'\beta$. A confidence region for $\Lambda'\beta$ consists of all the values that can be substituted for $\Lambda'\beta$ while maintaining the truth of the inequality. A precise definition of the 95 percent confidence region is as the collection of all values d in (4) that would not be rejected by an $\alpha = 0.05$ level F test. If Λ' has only one row, the confidence region simplifies to a confidence interval. See also *Estimation: Point and Interval.*

5. *Bayesian Estimation*

An alternative to least squares estimation is Bayesian estimation. In Bayesian estimation the parameters of the linear model are considered to be random and probability distributions reflecting the investigator's prior knowledge about the parameters are specified. Using Bayes theorem, the prior probability distributions are updated into posterior distributions that reflect the information embodied in the data. All statistical inferences are then based on the posterior distributions. Sometimes, a noninformative (improper) prior is used in which case, interval estimates arrived at using Bayesian methods are numerically equal to the interval estimates obtained from classical methods, only the interpretations of the intervals change. Alternatively, real prior information is often specified by taking the prior distribution of β given σ^2 to be a multivariate normal distribution and the prior distribution of σ^2 to be an inverse gamma distribution. This leads to modified least squares estimates that are shrunk towards the mean of the prior distribution β with corresponding changes to interval estimation. See also *Bayesian Statistics*; *Distributions, Statistical: Special and Continuous*; *Elicitation of Probabilities and Probability Distributions.*

6. *Model Checking*

Valid testing of a null hypothesis requires that all of the four itemized assumptions embodied in the model be true. Confidence intervals and regions are collections of parameter values that cannot be rejected by tests, so they also require the validity of the basic assumptions in the model as do Bayesian and non-Bayesian estimation. Model checking is considered an integral part of all statistical inference. A variety of methods have been developed for model checking. Most are based on plotting the residuals, see *Linear Hypothesis: Regression (Graphics)*. These methods include checking the normality assumption by plotting the ordered (standardized (Studentized)) residuals against (some approximation to) the expected order statistics of a standard normal distribution. Such a plot should be approximately linear if the data are normal.

The (standardized (Studentized)) residuals are also plotted against their corresponding predicted variables or any other variable (that does not depend on y) for which a measurement is available for each observation. In these plots, any identifiable systematic structure is an indication of problems with the assumptions.

Heteroscedasticity refers to having observations with different variances. In some models, formal tests for equal variances are available, e.g., Bartlett's test and the model-based tests discussed in Carroll and Ruppert (1988).

Lack of fit refers to the problem of having specified an incorrect linear mean structure, i.e., an incorrect linear hypothesis. Formal tests for lack of fit are often based on identifying clusters of observations that have similar predictor variables, i.e., similar rows of the X matrix, see Christensen (1996a, Sect. 6.6). Miller et al. (1998) have provided optimal methods for grouping observations into near replicate clusters.

The independence assumption is perhaps the hardest to evaluate. If the observations are taken at equally spaced time intervals, standard time series methods are often applied to the residuals. More generally, Christensen and Bedrick (1997) proposed creating rational subgroups (clusters) of the observations in which one suspects the observations may be more similar within clusters than between clusters. The methods of near replicate lack of fit tests can then be applied to test the independence assumption.

7. Weighted Models and Covariance Estimation

The standard assumption in linear models is that the observations all have the same variance and that the observations are all independent. If either or both of those assumptions is invalidated, different methods of estimation and testing are required. The usual linear model has

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 I$$

Writing a model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 V$$

where $\sigma^2 V$ is the variance–covariance matrix of the observations (and also of the unobservable errors), V can incorporate unequal variances for observations as well as correlations between observations. Assume that V has an inverse matrix V^{-1}.

Rather than using least squares estimates of β, we use weighted least squares estimates that minimize

$$(Y - X\beta)' V^{-1} (Y - X\beta)$$

for given Y and X. The estimates can be obtained by solving the normal equations

$$X' V^{-1} X\beta = X' V^{-1} Y$$

Typically, in regression problems, the unique solution is

$$\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} Y$$

These estimates are also BLUEs. Moreover, under the assumption that the observations have a multivariate normal distribution, the weighted least squares estimates are MLEs and MVUEs. The variance σ^2 is estimated by $MSE = SSE/dfE$ as before, except now

$$SSE = (Y - X\hat{\beta})' V^{-1} (Y - X\hat{\beta})$$

Testing a full model and a reduced model, works as before, with the F statistic still being

$$F = \frac{SSE(\mathrm{Red}) - SSE(\mathrm{Full})}{dfE(\mathrm{Red}) - dfE(\mathrm{Full})} / MSE(\mathrm{Full})$$

Note that incorporating the more general matrix V has no effect on the process of identifying reduced models—which only depend on the mean structure and not on the variance–covariance structure of the data.

All of these methods presume that V is known. In practice, V is often unknown and must be estimated from the data. This greatly complicates the entire process of data analysis. Typically, when V is estimated, weighted least squares estimates are not BLUEs or MVUEs, nor does the F statistic have an F distribution under the null hypothesis that the reduced model is true. Nonetheless, the ideas behind weighted regression are fundamental to many applications of the linear hypothesis.

A common application of weighted regression is when the observations are independent but have different variances. If the relative sizes of the variances are known, then V is known. Often the variance of each observation is modeled as a function of some predictor variables, see Carroll and Ruppert (1988).

Split plot models involve groups of observations that all have the same correlation. For example, in an experiment that involves different ways of scheduling classes in schools and different ways of assigning homework to students, the behavior of the schools might be considered independent of each other, but a group of five students taken from the same school should have similarities with each other, while being independent of students from other schools. This correlation structure must be incorporated into V. If the same number of students are measured from each school, i.e., if the groups of observations having the same correlation are all of the same size, many linear models still allow one to use least squares estimates and their associated F tests.

Mixed models involve the use of random effects, Partition X and β into two parts, so as to write

$$Y = X_1\beta_1 + X_2\beta_2 + e$$

Instead of considering β_1 and β_2 as fixed unknown parameters, treat β_2 as random. Assume that $\mathrm{E}(\beta_2) = 0$ and $\mathrm{Cov}(\beta_2) = \sigma^2 D$ with, say, e and β_2 independent. Consolidate all of the random parts into $\xi = X_2\beta_2 + e$ and write a linear model

$$Y = X_1\beta_1 + \xi, \quad \mathrm{E}(x) = 0, \quad \mathrm{Cov}(\xi) = \sigma^2 (X_2 D X_2' + I)$$

where $V = X_2 D X_2' + I$. Typically, D has to be estimated. To estimate β_1, simply use the estimate of D to get an estimate of V which is then treated as the true V.

Mixed models get much more complicated. X_2 is often partitioned into submatrices $Z_1, \ldots, Z_r$ and β_2 is correspondingly partitioned into γ_k, so as to write

$$Y = X_1\beta_1 + \sum_{k=1}^{r} Z_k \gamma_k + e$$

where $\mathrm{E}(\gamma_k) = 0$, $\mathrm{Cov}(\gamma_k) = \sigma_k^2 I$, and the γ_k are independent. The σ_k^2 are unknown parameters, and various methods such as residual (restricted) maximum likelihood (REML) and minimum norm

quadratic unbiased estimation (MINQUE) are used to estimate them. See also *Hierarchical Models: Random and Fixed Effects*; *Longitudinal Data*.

Another application of weighted estimation is in the analysis of spatial data. With spatial data, observations that are taken close together spatially are usually more correlated than observations taken far apart. V reflects the spatial correlations of the observations and X may reflect the locations of the observations. Models for V are developed based on the distances between observations. A common application in geographic information systems is map making, in which predictions are made for a very fine grid of locations. An individual best linear unbiased predictor for an observation on the grid, say, y_0 corresponding to location x_0' is

$$\hat{y}_0 = x_0'\hat{\beta} + V_{oY}V^{-1}(Y - X\hat{\beta})$$

where V_{0Y} is a matrix of covariances between y_0 and the original observations in Y. In practice, both V and V_{0Y} must be estimated from the data. See also *Spatial Statistical Methods*.

Multivariate linear models involve creating linear models for several variables at once. For example, y_{i1} might measure verbal ability and y_{i2} mathematical ability. Obviously, for any individual i, y_{i1} and y_{i2} may be correlated. The multivariate linear model fits

$$y_{i1} = x_i'\beta_1 + \varepsilon_{i1} \quad \text{and} \quad y_{i2} = x_i'\beta_2 + \varepsilon_{i2}$$

simultaneously. See also *Multivariate Analysis: Overview*.

Time series data involve observations $y_1, y_2, y_3, \ldots$, taken at regular time intervals. The correlation between two observations depends on the amount of time between them. One simple linear model for explaining the behavior of such data is a first order autoregression model.

$$y_i = \beta_0 + \beta_1 y_{i-1} + \varepsilon_i$$

in which the observation at time $i-1$ is used to predict the observation at time i. See also *Time Series: General* and *Time Series: ARIMA Methods*

Generalized linear models involve transforming the mean before making the linear hypothesis. For example, if observations y_i have means m_i one might make the hypothesis that

$$\log(m_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

This hypothesis together with assumptions about the distribution of y_i form a generalized linear model. In particular, since the linear hypothesis is based on the log transformation, this would be called a log-linear model. Parameter estimates are typically found by performing a series of weighted least squares analyses. Generalized linear models allow for the analysis of a variety of nonnormally distributed data. See also *Analysis of Variance and Generalized Linear Models*; *Multivariate Analysis: Discrete Variables (Loglinear Models)*; *Multivariate Analysis: Discrete Variables (Logistic Regression)*

8. Conclusion

Models based on the linear hypothesis, and their generalizations provide tools for analyzing a wide variety of data. Linear hypotheses allow one to address numerous interesting questions about the mean structure of the data. Moreover, the models can adjust for very general forms of dependence among the data. Linear models and their generalizations continue to be the most important and widely used models for the statistical analysis of data.

See also: Analysis of Variance and Generalized Linear Models; Bayesian Statistics; Distributions, Statistical: Special and Continuous; Elicitation of Probabilities and Probability Distributions; Estimation: Point and Interval; Experimental Design: Overview; Hierarchical Models: Random and Fixed Effects; Hypothesis Testing in Statistics; Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Longitudinal Data; Multivariate Analysis: Discrete Variables (Logistic Regression); Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Overview; Observational Studies: Overview; Significance, Tests of; Spatial Statistical Methods; Statistical Identification and Estimability; Time Series: ARIMA Methods; Time Series: General

Bibliography

Carroll R J, Ruppert D 1988 *Transformation and Weighting in Regression*. Chapman and Hall, New York

Christensen R 1991 *Linear Models for Multivariate, Time Series, and Spatial Data*. Springer, New York

Christensen R 1996a *Plain Answers to Complex Questions: The Theory of Linear Models*, 2nd edn. Springer, New York

Christensen R 1996b *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Chapman and Hall, New York

Christensen R 1997 *Log-linear Models and Logistic Regression*, 2nd Edn. Springer, New York

Christensen R, Bedrick E J 1997 Testing the independence assumption in linear models. *Journal of the American Statistical Association* **92**: 1006–16

Coleman J S, Campbell E Q, Hobson C J, McPartland J, Mood A M, Weinfeld F D, York R L 1996 *Equality of Educational Opportunity*. US Dept. of Health, Education, and Welfare, Office of Education, Washington, DC
Cook R D, Weisberg S 1994 *An Introduction to Regression Graphics*. Wiley, New York
Cressie N A C 1993 *Statistics for Spatial Data*, rev. edn. Wiley, New York
Koopmans L H 1987 *Introduction to Contemporary Statistical Methods*, 2nd edn. Duxbury Press, Boston, MA
McCullagh P, Nelder J A 1989 *Generalized Linear Models*, 2nd edn. Chapman and Hall, London
Miller F R, Neill J W, Sherfey B W 1998 Maximin clusters for near-replicate regression lack of fit tests. *Annals of Statistics* **26**: 1411–33
Mosteller F, Fienberg S E, Rourke R E K 1983 *Beginning Statistics with Data Analysis*. Addison-Wesley, Reading, MA
Mosteller F, Tukey J W 1977 *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading, MA
Searle S R 1971 *Linear Models*. Wiley, New York
Searle S R, Casella G, McCulloch C E 1992 *Variance Components*. Wiley, New York
Seber G A F 1977 *Linear Regression Analysis*. Wiley, New York
Seber G A F 1984 *Multivariate Observations*. Wiley, New York
Shumway R H, Stoffer D S 2000 *Time Series Analysis and its Applications*. Springer, New York

R. Christensen

Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox)

The study of relationships between two quantities has been important since earliest times: from the connection between weather and the hunted animals, down to cigarettes causing lung cancer. In the laboratory sciences it is possible to determine the relationship by controlling the conditions, changing one quantity and observing the effect of the change on the other. In the social and behavioral sciences this determination is not possible for three, related, reasons. First, conditions can rarely be controlled. Second, it is often not possible to set a quantity at an assigned value. Third, the relationship may not be exact but possess a random element, as with the heights of parents and their children. As a result, it is not easy to determine the nature of any possible relationship between quantities. Often an apparent connection has been found, on further study, to be spurious. We begin by looking at one of the more bizarre possibilities, usually known as Simpson's paradox (Simpson, 1951), but known to earlier writers. A more recent reference is Lindley and Novick (1981). Chapter 6 of Pearl (2000) contains a fine discussion in terms of causation. It is explained using some illustrative, medical data.

Patients were given either a treatment T, or a placebo T', to alleviate a condition. After a fixed time, it was observed whether they had recovered, R, or not, R'. Table 1 gives the data in the form of a 2×2 contingency table: thus 16 patients recovered, even with the placebo. The table also provides the marginal totals and the two recovery rates. The treatment appears to have been effective in increasing the recovery rate over the placebo by 10 percent. (As an aside, the numbers here are small, 80 patients in all; but the phenomenon could persist with 8,000. It is not the effect of small samples that is at issue.) There is an apparent relationship between treatment and recovery, but it has a random element and outside conditions have not been controlled.

One of these conditions is the patient's sex. Table 2, in the same form, gives the data for the men who participated in the study. Again the treatment has an effect but in the opposite direction, reducing the recovery rate by 10 percent. Common sense might suggest that a treatment which was overall beneficial, but harmful to men, would be highly beneficial for women. The data for them can be found by subtracting the cell entries in Table 2 for men from those for all patients in Table 1. Thus no new information is needed. The results are given in Table 3. Common sense misleads: the recovery rate for women also drops by 10 percent as a result of the treatment, just as with the men. This is a treatment which is bad for men, bad for women, but good for all of us. This is the paradox. It is here in an extreme form of complete reversal when an additional factor is included. It often arises in the form where there is some change in relationship when extra information is included. How can this happen?

It is easy to see what has gone wrong in the medical example. The condition is more serious for the women, with lower recovery rates than the men. Yet the treatment has been mainly given to the men—30 of them, against only 10 women. Perhaps the doctor was suspicious of the treatment and felt it was too

Table 1

	R	R′		Rate
T	20	20	40	50%
T′	16	24	40	40%
	36	44	80	

Table 2

Males	R	R′		Rate
T	18	12	30	60%
T′	7	3	10	70%
	25	15	40	

Table 3

Females	R	R′		Rate
T	2	8	10	20%
T′	9	21	30	30%
	11	29	40	

dangerous to give to the susceptible women. The consequence is that overall the treatment appears beneficial because it has been mainly given to the men, who were much more likely to recover in any case than were the women. Once this is appreciated, it is clear, from the data for the two sexes separately, that the treatment is harmful. Or is it clear? Here are two considerations that raise doubts:

(a) First, suppose that another quantity in addition to sex is included: for example, whether the patient came from a rural or urban environment. There would then be four contingency tables; one for rural males, one for urban females, and so on. We do not give the data, but it could happen that for each of these four tables there was a 10 percent improvement in the recovery rate, as with the overall data in Table 1. For this to happen, the allocation of patients would need to depend heavily on sex and environment, but it could happen. Furthermore, it could happen for some factor that was unrecognized by the scientists responsible for the experiment. The lesson is that, without some care in the consideration of quantities beyond the two that are the prime object of study, it is easy to be misled about the relationship between them.

(b) Second, suppose the numbers in the three tables remain unaltered but they apply to an agricultural, rather than medical, example. The treatment and placebo are replaced by two varieties of maize; recovery is now high yield as against low; and being male becomes growing tall as opposed to short. Thus the phenomenon of males having a higher recovery rate than females, corresponds to tall plants giving a greater proportion of high-yielders. Finally, the recovery rate is replaced by the proportion of plants with a high yield. In the medical example, it was suggested that the treatment was harmful. This translates, in the agriculture case, to T' giving the better yield. Is that really right? Surely T gave the better yield (Table 1) and grew taller (Tables 2 and 3). The suggestion is that the conclusions, from the same data, are quite different in the two contexts. The reason is that, whereas the medical treatment has no influence on sex, which maize variety is grown does influence the heights of the plants. The causal patterns in the two cases are not the same.

There is another way of appreciating the distinction between the agricultural and medical examples that avoids the difficult concept of causation. Suppose, on the evidence of the tables, and no other, you had another patient with the condition: would you give them the treatment? The answer is No, because the chance of recovery for treated males is less than for males having the placebo, and similarly for females, so that sex (which is outside your control) is irrelevant, and not influenced by the treatment applied. But when it comes to which variety of maize to plant, the replacement for sex, growing height, is influenced by the variety chosen. Hence the chances referred to above are irrelevant; only the chances of high yield for each variety are relevant and the growing height does not matter since it is determined by the choice of variety. Thus T is the better maize.

Notice that in all these examples, one of the quantities, T, could be set at a chosen value. Often this is not possible and the scientist merely passively observes whether T or T' obtains. In this case, transferring from the data to a new patient, in the manner of the last paragraph, becomes even more difficult. There can be a substantial difference between observing that an event occurs and arranging that it occurs. A person who is naturally a nonsmoker can behave quite differently from one who has been pressured into giving up smoking.

In order fully to appreciate what is going on, it is necessary, as always, to argue abstractly; to escape from the contexts of medicine or agriculture and to examine what principles are involved. This requires some mathematics and here the specific tool of probability. We write $p(A/B)$ for the probability that event A occurs, when it is known, or supposed, that event B has occurred. AB denotes the event which occurs if, and only if, both A and B occur. Then a reasonable interpretation of Table 1 is that $p(R/T) = 0.5$, $p(R/T') = 0.4$ and that T is preferred to T'. From now on think of R, T, and M (for male) as any three events and R', T', M' as their negations. The medical and agricultural examples provide illustrations. A third, social example with racist overtones is provided by the case, T is white (T' black), R is intelligent, and M upper, social class. All data are hypothetical.

Tables 2 and 3 refer, in this abstract form, to probabilities like $p(R/TM)$, differing from those in Table 1 by the inclusion of the conditioning event M or M'. These are related to the probabilities of Table 1 by the standard results

$$p(R/T) = p(R/TM)p(M/T) + p(R/TM')p(M'/T) \quad (1)$$

and

$$p(R/T') = p(R/T'M)p(M/T') + p(R/T'M')p(M'/T') \quad (2)$$

These introduce two extra probabilities, $p(M/T)$, and $p(M/T')$. Table 1 compared the probabilities on the left-hand sides of the two equations. Tables 2 and 3 compared the probabilities vertically above one another on the right. Those on the left were larger in the upper equation; those on the right were smaller

there. The reason for the paradox is that, on the right-hand sides of the two equations, the weights attached to those probabilities in order to combine them to produce those on the left, namely $p(M/T)$ in the upper, and $p(M/T')$ in the lower, are different. The difference was noted earlier with the treated patients being predominately male.

The paradox can be avoided if the weights are the same; namely $p(M/T) = p(M/T')$, for then any differences on the right will be reflected on the left. The general conclusion is that in studying a relationship involving T, any extraneous factor, M, should, if possible, arise with the same probability whether T is true or not. M and T should be independent. Statisticians say M is confounded with T if the probabilities are not equal. The sound recipe is to avoid confounding. Much effort has been put into designing experiments with little or no confounding.

We referred above to the difference between the medical and agricultural examples, in that the former resulted in a preference for T', the latter for T. The reason is that in the medical case, the scientist has control over $p(M/T)$ and, in particular, can avoid confounding by having the same proportion of males amongst treated patients as among those receiving the placebo. With the two varieties of maize, the proportion growing tall, M, is outside the scientist's control, so $p(R/T)$ is what matters. The reader may like to consider the politically-sensitive example mentioned above where the same data refer to the relative intelligences of black and white people.

The discussion so far has referred to events occurring or not. An event can be thought of as a quantity which takes only two values, 1 when it occurs, 0 when it does not. The probability that it occurs is then the expectation of the quantity. Looked at from this viewpoint, probability is a special case of expectation. The ideas above, including Simpson's paradox, extend to general quantities, X, Y, and Z, replacing T, R, and M, whenever the expectations are linear, as they necessarily must be when only two numbers, 1 and 0, are involved. Suppose then that the expectation of X, given Y and Z, is linearly dependent on their values,

$$E(X/Y,Z) = \alpha + \beta Y + \gamma Z \tag{3}$$

If, still retaining the dependence on Y, expectations are taken over Z, the result is

$$E(X/Y) = \alpha + \beta Y + \gamma E(Z/Y)$$

If $E(Z/Y)$ is also linear, equal to $\lambda + \mu Y$, then

$$E(X/Y) = \alpha' + \beta' Y \tag{4}$$

where

$$\alpha' = \alpha + \gamma\lambda \text{ and } \beta' = \beta + \gamma\mu$$

The contrast between Eqns. (4) and (3) is similar to that between Tables 1 and 2. For example, it could happen that β was positive but β' negative. If so, Eqn. (3) suggests that an increase in Y would be expected to produce an increase in X. Equation (4) similarly suggests a decrease in X.

More careful analysis modifies these statements. Changing γ in Eqn. (3) provides a measure of the effect on X, whilst Z is constant. No such restraint on Z is present in Eqn. (4). This is similar to Simpson's paradox, where Tables 2 and 3 held Z, now M or M' constant, whereas Table 1 did not. It explains the difference between the agricultural and medical examples. Z, height, cannot be held constant; whereas Z, sex can.

It is common for Eqn. (3) to be included in a statement like

$$X = \alpha + \beta Y + \gamma Z + \varepsilon$$

where ε is a random error with zero mean. Similarly Eqn. (4) becomes

$$X = \alpha' + \beta' Y + \varepsilon'$$

Statements in this form, unlike Eqns. (3) and (4), disguise the fact that the real distinction between them lies in the conditions under which X is being studied. Just as with Eqns. (1) and (2), the important subtleties arise because of the conditions. Thus ε' hides the fact that the variation of Z is included within it; ε does not. It is typically advantageous, when thinking about an expectation, or probability, to recognize the conditions that are assumed to hold, as well as the quantity being expected. In particular, probability depends on two arguments, not one; the uncertain event whose probability is being evaluated, and the events that are supposed to have occurred, the conditions. Distinction is sometimes made between probability and conditional probability. In reality, all probabilities are conditional. Care is needed in working with conditional expectations.

The main conclusion is that it is difficult to establish whether or not there exists a connection between two quantities and, if there does, to determine its precise form. The difficulty principally arises from the presence of other quantities that may disturb any apparent relationship. If controlled experimentation is possible, then consideration of interactions and the avoidance of confounding can be helpful in producing sound conclusions. With observational studies, in contrast, great care is necessary in avoiding inferences that the data do not fully support. An important reason why the social and behavioral sciences are less advanced than the physical ones is that controlled experimentation is more often possible in the latter than in the former.

These considerations do not deny the possibility of establishing relationships in society; merely that more

care is required. Emphasis needs to be placed on other quantities and the influence they might have on the relationship. These should be measured and the data incorporated into the analysis. On the other hand, there will be factors, like height in the agricultural example, which, although influential, can be left out of the analysis. All the time it is necessary to bring to the fore the conditions under which the basic relationship obtains. In consideration of that between X and Y, it is essential to take into account Z.

See also: Linear Hypothesis; Linear Hypothesis: Regression (Basics)

Bibliography

Lindley D V, Novick M R 1981 The role of exchangeability in inference. *Annals of Statistics*. **9**: 45–58

Pearl J 2000 *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK

Simpson E H 1951 The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, *Series B* **13**: 238–41

D. V. Lindley

Linear Hypothesis: Regression (Basics)

Regression analysis is the study of how a response variable depends on one or more predictors, for example how crop yield changes as inputs such as amount of irrigation or type of seed are varied, or how student performance changes as factors such as class size and expenditure per pupil are varied. In this article, we present the basics of the *linear regression model*, in which (a) the dependence of the response on the predictors is through the mean and (b) the mean of the response given the predictors is a linear function of the predictors. This is probably the most important and frequently encountered regression model.

1. Introduction

Regression analysis is the study of *dependence*: how does a response variable change as the values of one or more predictor variables are changed? In the general case, we have a response variable y, a set of predictor variables x, and we seek to understand the *conditional distributions* given by $F(y|x)$ as the predictors x are varied. Although the study of dependence probably began even earlier, two principal ideas lead to modern regression analysis (Stigler 1986). At the beginning of the nineteenth century, Legendre in France and Gauss

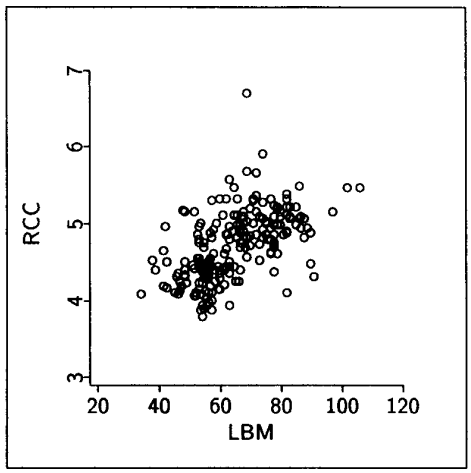

Figure 1
Scatterplot for the regression of *RCC* on *LBM*

in Germany studied the mean of a response as the predictors are varied using parametric models with only a few parameters. They suggested estimating the parameters by minimizing the sum of squared errors, which is the foundation for the least squares methods commonly used today. Toward the end of the nineteenth century in England, Galton, and later Pearson and Yule, provided many of the modern notions of correlation that provide a language for understanding dependence.

An example of a regression problem with one predictor is given graphically in Fig. 1, where the response variable $y = RCC$ is the red cell count and $x = LBM$ is the lean body mass in kilograms for 202 elite athletes that trained at the Australian Institute of Sport in the early 1990s (from Cook and Weisberg 1999). This scatterplot serves two general purposes in a regression context. First, it can help the viewer find separated or otherwise unusual points. In Fig. 1, one of the athletes had an unusually high value of *RCC*, about 7, with a relatively modest value of *LBM*. Such separated points are often of interest in their own right, and in any case they may have substantial effects on any computed summaries. The second use for this graph is to visualize how the conditional distribution $F(RCC|LBM)$ changes as the value of *LBM* is varied. Conditioning on *LBM* means looking at narrow vertical slices in the plot. The points within a slice represent approximately a sample from the distribution of *RCC* given that *LBM* is 'fixed' at the midpoint of the interval. We can visually integrate the contents of many slices to summarize changes in conditional distributions as we move from left to right. In Fig. 1, for example, we can see that the *mean function*, $E(RCC|LBM)$ that joins the mean in each slice appears to increase as we move from left to right, and furthermore that the increase appears to be

approximately the same for all values of *LBM*. In addition, we have enough points to say something about the conditional variability by examining the *variance function* $\mathrm{Var}(RCC \mid LBM)$. The variability appears to be approximately the same across the plot, or apparently $\mathrm{Var}(RCC \mid LBM = 50) \approx \mathrm{Var}(RCC \mid LBM = 80)$, or any other value.

2. Simple Linear Regression

The simple linear regression model makes assumptions concerning both the mean function $\mathrm{E}(y \mid x)$ and the variance function. In particular, we assume that for some transformation $u = u(x)$, we can write

$$\mathrm{E}(y \mid u) = \beta_0 + \beta_1 u \tag{1}$$

This equation requires some careful explanation. First, following Cook and Weisberg (1999), we allow the predictor x to be replaced by any monotonic transformation of it, such as the logarithmic, or inverse transformation. Doing so considerably expands the range of problems that can be handled using simple regression. Given this transformation, Eqn. (1) says that the expected increase in y for a unit increase in u is always equal to β_1 units, regardless of the initial value of u. Thus in a graph of y vs. u, the mean function specified by Eqn. (1) is a straight line with intercept β_0 and slope β_1.

The mean function of Eqn. (1) concerns average behavior, but says nothing about the variation. Suppose we have observed data $(x_i, y_i), i = 1, \ldots, n$, observed on n units, and suppose further that we have computed $u_i = u(x_i)$ for the known transformation u. Of course the points (u_i, y_i) will not fall exactly on the line (they certainly do not do so in Fig. 1). The unobservable differences

$$\varepsilon_i = y_i - \mathrm{E}(y_i \mid u_i)$$

are called *errors*. By construction, if the model is correct the errors are random variables with mean zero and variance given by a positive constant σ^2. This amounts to assuming that the variance function $\mathrm{Var}(y \mid x)$ is constant. We assume that the ε_i are independent random variables. They are often assumed to be normally distributed, but the normality assumption is required only for inference, not for estimation.

The model given by Eqn. (1) and the assumptions concerning variability just described greatly simplify the regression problem by reducing the mean function to only two parameters, and the variance function to only one parameter. When this model is appropriate, it gives a very elegant summary to a regression problem.

2.1 Fitting

The most common method of estimating the unknown parameters in Eqn. (1) based on data uses ordinary least squares. Find the values $\hat{\beta}_0$ and $\hat{\beta}_1$ of (β_0, β_1) that minimize the residual sum of squares function $RSS(\beta_0, \beta_1)$, given by

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \mathrm{fit}_i)^2 \tag{2}$$

where $\mathrm{fit}_i = \beta_0 + \beta_1 u_i$. Finding the estimates is an elementary calculus problem, and in general the estimates are unique and are given by a formula. The estimate of σ^2 is just $\hat{\sigma}^2 = RSS/(n-2)$, where *RSS* is the value of the residual sum of squares function evaluated at the estimates, and $n-2$ is the degrees of freedom, which is equal to the number of observations n minus the number of parameters estimated in the mean function, which is 2 for simple linear regression in Eqn. (1).

Table 1 gives typical computer output for fitting the simple linear regression model to the athletes data. The column marked Estimate gives the estimated coefficients, so the intercept is estimated to be about 3.47. The estimated slope of 0.019 suggests that we

Table 1
Simple linear regression of *RCC* on *LBM*

	Coefficient estimates		
Label	Estimate	Std. error	*t*-value
Constant	3.46616	0.136820	25.334
LBM	0.0193060	0.00206768	9.337
R squared:		0.303574	
Sigma hat:		0.383145	
Number of cases:		202	
Degrees of freedom:		200	

expect *RCC* to increase on the average by about 0.019 units for each 1 kg increase in *LBM*. The column marked Std. error gives the estimated standard errors of these values, which are computed as $\hat{\sigma}$ times a constant that depends on the values of the u_i but not on the response. The t-values are the ratio of the estimate to its standard error. These are appropriate test statistics for the hypotheses that each of the coefficients in turn is equal to zero. Assuming normal errors, these values can be compared with the t_{n-2} distribution to get significance levels. In the athletes data, tests concerning the intercept are of little interest. The test that the slope is equal to zero vs. the general alternative has value $t = 9.34$ and a corresponding tiny significance level, suggesting clearly that *RCC* increases with *LBM*.

A common summary statistic for the fit of a regression is the quantity R^2, which is defined to be the *fraction of the variability in the response explained by the model*, so

$$R^2 = \frac{SYY - RSS}{SYY} \tag{3}$$

where $SYY = \mathrm{E}(y_i - \bar{y})^2$ is the total variability in the response ignoring the predictors. If $R^2 = 1$, then $RSS = 0$ and the mean function matches the data perfectly. If $R^2 = 0$, then $SYY = RSS$, and knowledge of the predictors tells us nothing about the response. For the athletes data, $R^2 = 0.30$, so about 30 percent of the variability in *RCC* is explained by *LBM*. R^2 is a most appropriate measure when the corresponding scatterplot appears to be an elliptical swarm of points which in turn suggests bivariate normality of both the response and the predictor. This is a very strong assumption that appears to be appropriate for this example.

2.2 Diagnostics

The residuals are defined to be the differences between the observed values y_i and the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 u_i$, so $e_i = y_i - \hat{y}_i$. Unlike the errors ε_i, the residuals e_i are observable, and so they can be used in graphs and other procedures as a check on the assumptions of the fitted model. Imagine drawing a scatterplot of the residuals vs. u. If the model is correct, then we should expect that the conditional distribution of $e \mid u$ should be the same as we move from left to right: the mean function should be constant and the variance function should be constant. If the model is not correct, then either of these two characteristics may not hold. The residual plot for the athletes data is shown in Fig. 2. As one would have expected, the model for the mean function matches the data fairly well, as both the mean and variance functions appear to be constant. The other interesting feature is the poor fit for one athlete with very high *RCC*. Deleting this case and refitting can tell us about the impact of this point on the fitted equation; doing so results in only minor changes in the estimated coefficients or in $\hat{\sigma}$. Other separated points can cause very large changes in estimates, and such points are called influential.

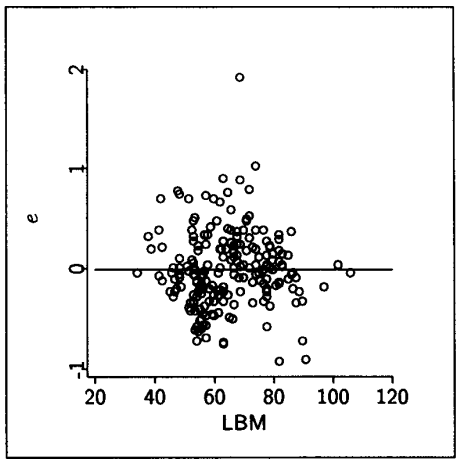

Figure 2
Residual plot from simple regression of *RCC* on *LBM*

3. Multiple Regression

Many regression problems require consideration of more than one predictor, and it is required to understand how the response y depends simultaneously on the predictors $x_1, x_2, \ldots, x_p$. For example, the dependence of *RCC* on *LBM* discussed so far ignores the fact that these data consist of both male and female athletes, and one might reasonably assume that gender differences are important. This will require either fitting separate models for the two populations of males and females, or more generally including gender effects in the mean function. Other problems might require additional predictors; in the athletes data, for example, including height, weight, or some blood chemistry variables might be appropriate.

We begin with p predictors $x_1, \ldots, x_p$, and from these we will build a set of k *terms* $u_0, u_1, \ldots, u_{k-1}$, to get the multiple linear regression mean function

$$\mathrm{E}(y \mid u_1, \ldots, u_{k-1}) = \beta_0 u_0 + \beta_1 u_1 + \cdots + \beta_{k-1} u_{k-1} \tag{4}$$

Including terms, which are functions of the predictors, makes this model very general. For example, a model with one predictor x might include the terms $u_0 = 1$ for the intercept, $u_1 = x$, $u_2 = x^2$ and $u_3 = x^3$, so polynomial models are included in the multiple linear regression class. In addition, the u terms can include interactions (e.g., $u = x_1 x_2$), transformations (e.g., $u = \log(x_1)$), and other functions of the predictors. Quantitative variables like gender can be included in a model by defining *factors*, which are sets of indicator

variables (often having the values 0 or 1) that correspond to the levels of the factor. For example, the variable *Treatment* with three levels can be included in a model with two indicator variables, the first with the value 1 for units that received treatment T1 and 0 otherwise; the second with the value 1 for units getting treatment T2, and 0 otherwise; T3 is indicated if both indicators are equal to 0. Since gender has only two levels, only one indicator variable is required.

Suppose that *Gender* has the value 1 for females and 0 for males. Figure 3 shows a fit of the mean function

$$\mathrm{E}(RCC \mid LBM, Gender) = \beta_0 + \beta_1 Gender + \beta_2 LBM + \beta_3 LBM \times Gender \quad (5)$$

to the athletes data. This is a multiple regression model with two predictors, *LBM* and Gender, and four terms, an intercept, an interaction, and terms for *LBM* and *Gender*. For males, *Gender* = 0 and (5) becomes

$$\mathrm{E}(RCC \mid LBM, Gender = 0) = \beta_0 + \beta_2 LBM$$

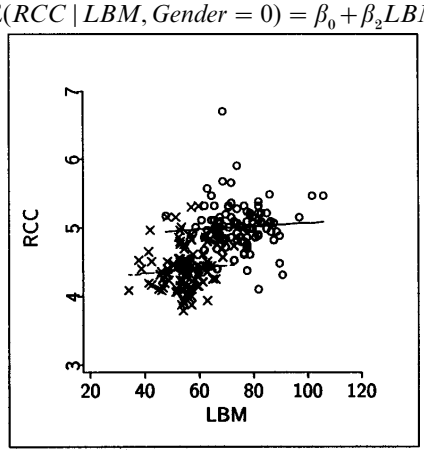

Figure 3
Regression of *RCC* on *LBM* controlling for gender. Points for females are marked with a X and have a fitted line given by the dashed line. Points for males are marked with a 0 and the fitted line is the solid line

while for females, *Gender* = 1 and

$$\mathrm{E}(RCC \mid LBM, Gender = 1) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) LBM$$

We recognize that this model fits straight lines separately for each gender with intercept and slope β_0 and β_2 for males, and $\beta_0 + \beta_1$ and $\beta_2 + \beta_3$, for females. We see from the fitted least squares lines shown in Fig. 3 that within gender the regression have different intercepts (males appear to have higher *RCC* than females), but the two groups appear to have approximately equal slopes, and those two slopes may well be zero. Thus, by conditioning on group, the relationship between *RCC* and *LBM* disappears. Table 2 summarizes the ordinary least squares fit based on model (5). The fitting method is a direct generalization of the method for simple regression given at (2), except that fit_i is now defined using (5). We now see that the estimates for *LBM* and *LBM* × *Gender* are very small both in absolute value, and small relative to their standard errors. All this suggests that, given gender, *RCC* and *LBM* are independent.

This example illustrates that choice of predictors in a regression problem can be very important. The initial analysis, ignoring gender, suggested that *RCC* increased with increasing *LBM*. This spurious dependence appears to be due to gender differences: women have lower values of *LBM* than men, on the average, and they also appear to have lower values of *RCC* than do men (at least for elite athletes). The initial analysis reflected gender differences, not a meaningful relationship between the response and the predictor.

4. *Interpreting Parameters*

Weisberg (1985) reported the results of a regression to study per capita fuel consumption for the United States as a function of four predictors: *Tax* rate, in cents per gallon, *Dlic*, the fraction of adults with drivers licenses; *Inc*, the per capita income in thous-

Table 2
Multiple linear regression of *RCC* on *LBM* and *Gender*

	Coefficient estimates		
Label	Estimate	Std. error	*t*-value
Constant	4.85114	0.255403	18.994
Gender	−0.666013	0.372175	−1.790
LBM	0.00234984	0.00339170	0.693
LBM × *Gender*	0.00164645	0.00595363	0.277
R squared:		0.466583	
Sigma hat:		0.337009	
Number of cases:		202	
Degrees of freedom:		198	

ands of dollars, and *Road*, the number of miles of roadway in the state. The fitted linear regression model was:

$$Fuel = 377 + 35Tax + 13Dlic - 66Inc - 2.4Road$$

Assuming that this mean function is appropriate for these data (a question that should always be considered in trying to understand a regression problem), how are the parameter estimates to be interpreted? For example, the coefficient of -66 for *Inc* suggests that for every \$1,000 increase in per capita income, fuel consumption will decrease by about 66 gallons per person, assuming that all other predictors could be held fixed at their current values. Similarly, the value of $+35$ for *Tax* suggests that for every penny increase in fuel tax, consumption goes *up*, contrary to the common sense result that increasing cost through taxes should decrease consumption, assuming all other factors are held fixed. How can this happen?

These data are observational: tax rates are determined by state legislatures in response to many other factors, possibly including the per capita fuel consumption. Thus the observed positive relationship may be due to high consumption causing an increase in taxes, rather than the other way around. There is no way to sort out an issue like this one without actually experimentally changing variables and observing what happens; this is surely not possible in setting tax rates.

As another example, Weisberg (1985) reported on the regression summary of the dependence of accident rate per million vehicle miles on speed limit *Slim*, number of access points *Acpt* and length *Len* of a segment for 39 segments of major Minnesota highways. The fitted model was

$$Rate = 9.32 - 0.77Len - 0.10Slim + 0.10Acpt$$

We get the sensible result that increasing the number of access points in a segment, while keeping the other two variables fixed, will increase the accident rate by 0.1. The other two coefficients are less obvious. The negative sign for *Slim* suggests that increasing speed limits will lower accident rates (these data predate the original reduction of the maximum speed limit to 55 mph), but once again we cannot tell if higher speeds cause fewer accidents, or if dangerous segments of road encourage the highway department to lower the speed limit; both would give rise to the negative coefficient.

The coefficient for *Len* is also curious; why should accident rate be related to the length of the segment, since the rate is already adjusted for the miles driven? If most accidents occur at isolated dangerous spots that are relatively rare, then increasing the length of a segment by a small amount is unlikely to include another dangerous spot. The longer segment will then have about the same number of accidents, but a lower value for accidents per million vehicle miles. Consequently, the negative association between the rate and the length, adjusted for the other variables, is an artifact of the way the data are collected. It has no intrinsic interest on its own, but it must be included in all regression models to remove the effect of this artifact.

5. Conclusion

Regression methodology generally seeks to understand the dependence of a response variable on a set of predictors as the predictors are varied. This very general methodology finds application in almost every branch of human inquiry. The approach to regression briefly outlined here is expanded at a relatively low mathematical level in Cook and Weisberg (1999). Fox (1997) provides another entry to the literature on regression, at about the same level. Draper and Smith (1998) requires a modestly higher level of mathematical sophistication.

Almost any of the popular statistical computing package will include programs for fitting the type of models discussed in this article. The larger programs like SAS offer several seemingly different procedures with names like 'proc reg' and 'proc glm' that in reality are just fitting the regression models described here. Newer approaches to regression require extensive use of graphics, and several programs are specifically designed for graphics, for example, the free program Arc that is described in Cook and Weisberg (1999).

See also: Linear Hypothesis; Linear Hypothesis: Regression (Graphics); Multivariate Analysis: Discrete Variables (Logistic Regression)

Bibliography

Cook R D, Weisberg S 1999 *Applied Regression Including Computing and Graphics*. Wiley, New York

Draper N, Smith H 1998 *Applied Regression Analysis*, 3rd edn, Wiley, New York

Fox J 1997 *Applied Regression Analysis, Linear Models, and Related Methods*. Sage, Thousand Oaks, CA

Stigler S M 1986 *The History of Statistics*. Belknap, Cambridge, MA

Weisberg S 1985 *Applied Linear Regression*, 2nd edn. Wiley, New York

S. Weisberg

Linear Hypothesis: Regression (Graphics)

Regression analysis is the study of how a response variable depends on one or more predictors. For example, as a part of an ecological study of New Zealand mussels, we might wish to study the dependence of muscle mass Y (g), the edible portion of a mussel, on the size of the shell as represented by the

shell height X_1, shell length X_2, and shell width X_3 (mm). We later use a data set from this study containing observation on 82 mussels.

Graphical displays can be quite useful in regression analysis because they can be used to visualize dependence, to discover unexpected relationships, to guide the choice of a model, and to check and refine plausible models. In this article, we present the basic ideas and methods of regression graphics.

1. Introduction

The overarching goal of a regression analysis is to understand how the conditional distribution of the response variable Y given a vector of p predictors $\mathbf{X} = (X_j, j = 1, \ldots, p)$ depends on the value assumed by $\mathbf{X}$. While attention is often restricted to the mean function $\mathrm{E}(Y|\mathbf{X})$ and perhaps the variance function $\mathrm{Var}(Y|\mathbf{X})$, in full generality it is the conditional distribution of $Y|\mathbf{X}$ that is of interest. The population of New Zealand mussels can be viewed as consisting of several subpopulations, one for each distinct shell size as represented by the value of $\mathbf{X}$. In the regression we wish to study how the distribution of muscle mass changes as we move between subpopulations.

Linear regression is a restrictive but common form of analysis in which $\mathrm{E}(Y|\mathbf{X})$ is expressed as a linear function of the predictors and the data are used to estimate the unknown coefficients of the predictors (see *Linear Hypothesis: Regression (Basics)*). For example, we might begin an analysis of the mussel regression by fitting the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma\varepsilon \qquad (1)$$

where ε is an unobservable random error with mean 0 and variance 1, the βs are unknown regression coefficients, and σ is the unknown standard deviation. In this particular model, the distribution of $Y|\mathbf{X}$ depends on $\mathbf{X}$ only through the mean function

$$\mathrm{E}(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \qquad (2)$$

which is linear in $\mathbf{X}$. The variance function is constant $\mathrm{Var}(Y|\mathbf{X}) = \sigma^2$ under the model.

The choice of a model can be problematic, and is often the most difficult part of a regression analysis. Beyond the general notion that linear models may serve as useful approximations, there is little prior information to support model (1) for the mussel regression and data-based modifications may well be required. Graphics can be useful in the choice of a first model. If the mussel regression had only one predictor X_1, a two-dimensional scatter plot of Y vs. X_1 could be used to visualize how the conditional distribution $Y|\mathbf{X}_1$ changes with the value X_1. Similarly, if the regression had only two predictors, say X_1 and X_2, we could use

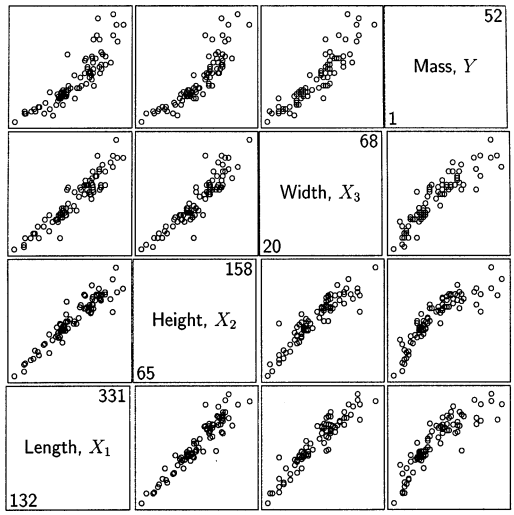

Figure 1
Scatterplot matrix for the mussel regression.

a computer-generated three-dimensional plot of Y vs. (X_1, X_2) to visualize the conditional distributions. But the mussel regression has three predictors, requiring a four-dimensional plot to view the data in full. Four-dimensional plots can not be viewed in their entirety in our three-dimensional world, so we must find indirect low-dimensional methods of viewing regression data when there are more than two predictors.

For instance, a scatterplot matrix of the mussel data is shown in Fig. 1. The diagonal cells in the scatterplot matrix serve to name the axes for the individual plots and show the observed range for each variable. The top row contains plots of the response vs. each of the predictors, and the predictors are plotted in the lower left 3×3 array. This scatterplot matrix contains much information: we see that each of the mean functions for the marginal regressions in the top row is curved and that the variance functions are increasing. There are fairly strong linear relationships between the predictors, indicating that collinearity may be an issue. This type of information from low-dimensional plots can be useful but unfortunately may have little to do with the regression of Y on $\mathbf{X}$. The fact that the marginal mean functions $\mathrm{E}(Y|X_j)$, $j = 1, 2, 3$, are curved does not necessarily imply that the full mean function $\mathrm{E}(Y|\mathbf{X})$ is a nonlinear function of $\mathbf{X}$. Similarly, we cannot safely conclude that the variance function $\mathrm{Var}(Y|\mathbf{X})$ is not constant.

Plots of convenience such as the scatterplot matrix of Fig. 1 may not portray the regression accurately and can misdirect the analysis. Low-dimensional plots must be constructed and interpreted carefully if we are to avoid such possibilities. In regression graphics we pursue low-dimensional *sufficient summary plots* which contain all the information on Y that is available from

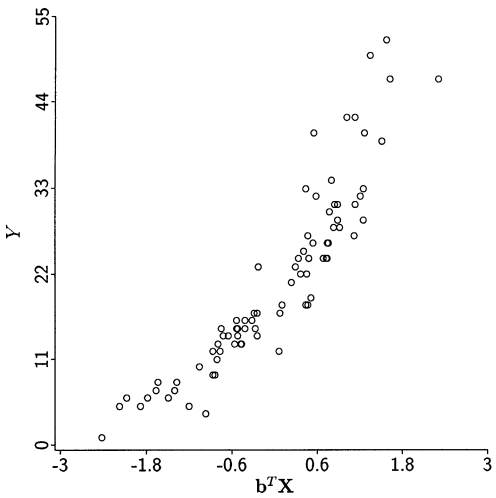

Figure 2
Summary plot for the mussel regression.

X, and which avoid the distortion that can result from simple convenient displays of the data.

Figure 2 shows a summary plot of muscle mass Y vs. the linear combination of the predictors

$$\mathbf{b}^T\mathbf{X} = 0.50X_1 + 0.08X_2 + 0.86X_3 \tag{3}$$

Methods of regression graphics discussed later in this article enable us to infer that this plot is *sufficient* and thus contains all the information on muscle mass Y that is available from the three predictors in **X**, namely shell height, length, and width. Because the plot is inferred to be sufficient and because no model was used in its construction, we can use it to guide the remaining analysis including the choice of a first model. We conclude from the plot that the mean function is curved and that the variance function is very likely not constant. Both conclusions contradict model (1).

2. *Sufficient Summary Plots*

Dimension reduction is a leitmotif of statistics. For instance, starting with a sample $z_1, \ldots, z_n$ from a univariate normal population with mean, μ and variance 1, we know that the sample mean $\bar{z}$ is sufficient for μ (see *Statistical Sufficiency*). This means that we can replace the original n-dimensional sample with the one-dimensional mean $\bar{z}$ without loss of information on μ.

In the same spirit, dimension reduction without loss of information is a dominant theme of regression graphics. The goal of a regression study is to infer about the conditional distribution of the response Y given the $p \times 1$ vector of predictors **X**. A central goal of regression graphics is to reduce the dimension of **X** without loss of information on the conditional distribution of $Y|\mathbf{X}$ and without requiring a model. We call this sufficient dimension reduction, borrowing terminology from classical statistics. Sufficient dimension reduction leads naturally to sufficient summary plots which contain all of the information on the regression that is available from the sample.

In regression graphics we search for the smallest number $d \leqslant p$ of linear combinations of the predictors, say $\mathbf{b}_1^T\mathbf{X}, \ldots, \mathbf{b}_d^T\mathbf{X}$, that contain all the information about Y that is available from **X**. To formalize this idea, we re-express it in the form of a conditional independence statement

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T\mathbf{X} \tag{4}$$

Here columns of the $p \times d$ matrix **B** are the vectors $\mathbf{b}_1, \ldots, \mathbf{b}_d$ so that $\mathbf{B}^T\mathbf{X}$ is the $d \times 1$ vector of linear combinations $\mathbf{b}_1^T\mathbf{X}, \ldots, \mathbf{b}_d^T\mathbf{X}$. The notation '$\perp\!\!\!\perp$' means independence. The statement is, thus, that Y is independent of **X** given any values for the d linear combinations $\mathbf{B}^T\mathbf{X} = (\mathbf{b}_1^T\mathbf{X}, \ldots, \mathbf{b}_d^T\mathbf{X})^T$. The value of d is called the structural dimension of the regression of Y on **X**, and we refer to a regression as having kD structure for $k = 0, 1, \ldots, p$. We can always find a value of d and a corresponding matrix **B** so that Egn(4) is true. For example, (4) is always true when $d = p$ and $\mathbf{B} = \mathbf{I}$.

If a regression has 0D structure, then $Y \perp\!\!\!\perp \mathbf{X}$, and the predictors contain no information about the response. In such cases, a simple histogram or other graphical representation of the marginal distribution of Y is a sufficient summary plot. If $d = 1$ then all the information about Y that is available from **X** is contained in a single linear combination $\mathbf{b}_1^T\mathbf{X}$, and a plot of Y vs. $\mathbf{b}_1^T\mathbf{X}$ is a sufficient summary plot for the regression. Under 2D structure all the information about Y that is available from **X** is contained in two linear combinations $(\mathbf{b}_1^T\mathbf{X}, \mathbf{b}_2^T\mathbf{X})$, and a three-dimensional plot of Y vs. $(\mathbf{b}_1^T\mathbf{X}, \mathbf{b}_2^T\mathbf{X})$ is a sufficient summary plot for the regression.

Each of the following regression models is an example of a regression with 1D structure.

$$Y = b_0 + \mathbf{b}^T\mathbf{X} + \sigma\varepsilon \tag{5}$$

$$Y^{(\lambda)} = b_0 + \mathbf{b}^T\mathbf{X} + \sigma\varepsilon \tag{6}$$

$$Y = f(\mathbf{b}^T\mathbf{X}) + \sigma\varepsilon \tag{7}$$

$$Y = f(\mathbf{b}^T\mathbf{X}) + \sigma(\mathbf{b}^T\mathbf{X})\varepsilon \tag{8}$$

Model (5) is the linear regression model. In model (6) we have a linear regression model after some response transformation represented by the unknown parameter λ. The mean function in model (7) is $\mathrm{E}(Y|\mathbf{X}) = f(\mathbf{b}^T\mathbf{X})$, where f may be unknown or known, but it depends on only one linear combination of the predictors and thus the regression still has 1D structure. Similarly, in model (8) both the mean and standard deviation functions may depend on $\mathbf{b}^T\mathbf{X}$. Generalized linear models with corresponding link functions also have 1D structure. The class of re-

gression models with 1D structure is quite large and covers many models used in practice. A regression model with 2D structure can be considerably more complex that a regression with 1D structure. The model

$$Y = f(\mathbf{b}_1^T\mathbf{X}, \mathbf{b}_2^T\mathbf{X}) + \sigma(\mathbf{b}_1^T\mathbf{X}, \mathbf{b}_2^T\mathbf{X})\varepsilon$$

is one example of a regression with 2D structure.

The structural dimension d of a regression and the corresponding linear combinations $\mathbf{B}^T\mathbf{X}$ will generally be unknown. In the next section we discuss methods for estimating these quantities, leading to estimated sufficient summary plots.

3. *Estimating Sufficient Summary Plots*

There are several methods for estimating sufficient summary plots or properties thereof. While there are no restrictive constraints placed on the foundations of regression graphics as represented by Eqn. (4), some constraints are necessary for estimation. The most notable is that there should be no strong nonlinear relations among the predictors. This requirement constrains the marginal distribution of $\mathbf{X}$ and not the conditional distribution of $Y|\mathbf{X}$ as is usual when modeling in practice. It is part of the cost of insuring that certain low-dimensional plots contain useful information on the regression. There are several methods for insuring that this requirement holds in practice, including the possibility of using power transformations of the predictors to remove extreme nonlinearities.

Returning to the scatterplot matrix in Fig. 1, we see that the mean functions in the predictor plots are all strongly linear, and thus we can proceed to infer about properties a sufficient summary plot. A first inference is available from the *inverse response plots* of X_j vs. Y in the last column of the scatterplot matrix. Since all of these plots have the same shape we infer that $d = 1$ and thus that there is a two-dimensional sufficient summary plot. The justification for this conclusion is based on the vector of inverse mean functions $\mathrm{E}(\mathbf{X}|Y)$. If all inverse mean functions are linear related so that for some function $m(Y)$ of the response $\mathrm{E}(X_j|Y) = \mathrm{E}(X_j) + \alpha_j m(Y)$, $j = 1, \ldots, p$, then 1D structure is indicated. Next, if $d = 1$ as we have just inferred then there is a $p \times 1$ vector $\mathbf{b}$ such that (a) $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{b}^T\mathbf{X}$ and (b) a plot of Y vs. $\mathbf{b}^T\mathbf{X}$ is a sufficient summary of the regression.

The next issue in our discussion of the mussel data is how to estimate $\mathbf{b}$. If there are no strong nonlinear relations among the predictors then $\mathbf{b}$ can be estimated as the coefficient vector of $\mathbf{X}$ from the ordinary least squares regression of Y on $\mathbf{X}$. And the estimated sufficient summary plot can be taken as a plot of Y vs. the fitted value $\hat{Y}$ from this same regression. This is how the summary plot of Fig. 2 was produced. No model is being assumed in this use of ordinary least squares. Indeed, we saw previously that the multiple linear regression model (1) is not sustained by the data. Nevertheless, the estimated coefficient vector $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^T$ still provides a consistent estimate of $\mathbf{b}$. Estimation methods other than least squares can be used as well, provided the objective function is convex.

There are several more comprehensive methods for inferring about sufficient summary plots. All are based on transforming the original vector $\mathbf{X}$ of p predictors to p estimated linear combinations

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \rightarrow \mathbf{Z} = \begin{pmatrix} \hat{\mathbf{b}}_1^T X \\ \hat{\mathbf{b}}_2^T X \\ \vdots \\ \hat{\mathbf{b}}_p^T X \end{pmatrix}$$

The p estimated linear combinations $\hat{\mathbf{b}}_j^T\mathbf{X}$ are ordered on the likelihood that they will be needed in a sufficient summary plot. If it is inferred that $d = 0$ then none of the linear combinations is needed. If it is inferred that $d = 1$ then only the first is needed and a plot of Y vs. $\hat{\mathbf{b}}_1^T\mathbf{X}$ is the estimated summary plot. Similarly, if the structural dimension is inferred to be $d = 2$ then only the first two estimated linear combinations are needed and a three-dimensional plot of Y vs. $(\hat{\mathbf{b}}_1^T\mathbf{X}, \hat{\mathbf{b}}_2^T\mathbf{X})$ is the estimated summary plot.

There are three primary numerical methods for producing the vector $\mathbf{Z}$ of linearly transformed predictors, each with its own advantages and disadvantages

(a) Sliced Inverse Regression (SIR)
(b) Sliced Average Variance Estimation (SAVE)
(c) Principal Hessian Directions (pHd)

SIR and SAVE use the column of inverse response plots in a scatterplot matrix of the predictors and the response, as discussed previously in connection the mussel data. They measure the shape of the inverse response plots by first replacing the response Y with a discrete version $\tilde{Y}$ constructed by partitioning the range of Y into K *slices*. SIR then gauges the shapes of the inverse mean functions $\mathrm{E}(\mathbf{X}|\tilde{Y})$, while SAVE gauges the shapes of the inverse functions $\mathrm{E}(\mathbf{X}|\tilde{Y})$ and the inverse variance functions $\mathrm{Var}(\mathbf{X}|\tilde{Y})$ simultaneously. The number of slices is a tuning parameter in both approaches. pHd is based on third moments between the response and the predictors, and does not have a tuning parameter. Each method comes with an inference procedure for the structural dimension d of the regression. In particular, beginning with $\mathbf{Z}$ predictors, graphical regression is a method for visually estimating the structural dimension of a regression and the corresponding summary plot.

SIR is good at estimating sufficient summary plots with linear trends, but it does not generally work well for finding plots with curvature and little linear trend. pHd works best for finding quadratic trends. SAVE is

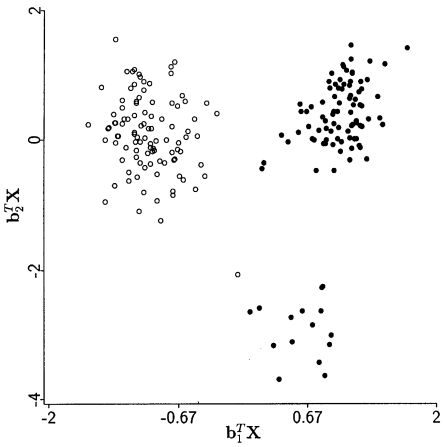

Figure 3
Estimated sufficient summary plot from SAVE for the bank note data.

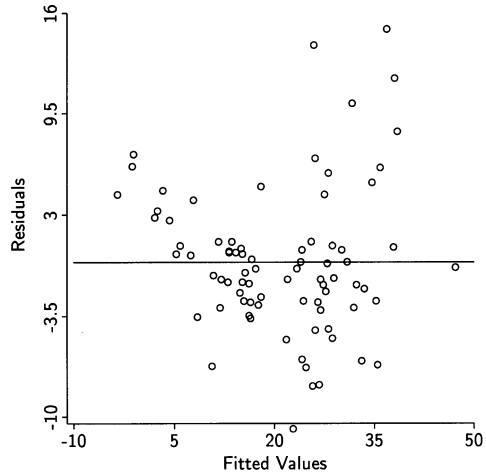

Figure 4
Scatterplot of residuals versus fitted values from the ordinary least squares fit of model (1).

the most comprehensive, but may miss information when it is concentrated in linear trends where SIR works well.

We use a data set on 200 Swiss bank notes for illustration. The response is binary with $Y = 1$ indicating a counterfeit note and $Y = 0$ indicating a genuine note. The methods of regression graphics discussed previously apply equally to regressions with continuous and binary responses. There are six predictors, each giving a different aspect of the size of a note: X_1 = length of bottom edge, X_2 = diagonal length, X_3 = left edge length, X_4 = length at center, X_5 = right edge length, and X_6 = top edge length.

Application of SAVE to the bank note data indicated 2D structure. Shown in Fig. 3 is a binary response plot of the first two linear combinations $(\hat{\mathbf{b}}_1^T\mathbf{X}, \hat{\mathbf{b}}_2^T\mathbf{X})$. We infer that this plot contains all the information about a note's authenticity that is available from the six predictors. The summary plot shows an outlying genuine note, good separation between the genuine and counterfeit notes, and a bimodal distribution within the counterfeit notes. The outlying genuine note could be a mislabeled counterfeit note, and the bimodal distribution within the counterfeit notes might reflect two different production processes.

4. *Special Purpose Graphics*

In addition to sufficient summary plots, there are a variety of special purpose graphics that can be used to address specific issues. These visualization methods mostly center on assessing or improving a target model.

Perhaps the most common graphical methods of model assessment are based on searching for dependence between residuals from the fitted model and the predictors. The various methods for estimating sufficient summary plots can be used to detect such dependence by simply using residuals in place of the response. In this setting the model is sustained when no information is found to contradict 0D structure.

A plot of residuals vs. fitted values from the ordinary least squares fit of model (1) is shown in Fig. 4. The horizontal line intersecting zero on the vertical axis was added as a visual reference. If the model was correct, this plot should leave the impression that the residuals are independent of the fitted values. Because the residuals and fitted values in the plot are clearly dependent, we have information to contradict 0D structure and the model is called into question.

CERES plots, including the class of partial residual plots can be used to study the need for predictor transformations in a multiple regression model such as model (1). Similar graphical methods are available for studying the need for response transformations. Net-effect plots can be used to study the contribution of a selected predictor to the regression. Versions are available for both model-based and model-free regression studies. Marginal model plots can be used to assess model adequacy.

5. *Further Reading*

The approach to regression graphics based on sufficient summary plots and many special purpose graphics were discussed at a relatively low mathematical level by Cook and Weisberg (1999). A more advanced treatment that requires knowledge of linear models and linear algebra was given by Cook (1998a). Sliced inverse regression and the method of principal Hessian directions were proposed by Li (1991, 1992). Sliced average variance estimation was proposed by Cook and Weisberg (1991). Extensions of these

methods were studied by Cook (1998b), Ferré (1998), Filliben and Li (1997), Hsing and Carroll (1992), Schott (1994), Velilla (1998), and Zhu and Fang (1996). Graphical regression was proposed by Cook (1998a). The mussel data were taken from Camden (1989). The bank note data were taken from Flury and Riedwyl (1988); the analysis follows Cook and Lee (1999) who studied regressions with a binary response.

Bibliography

Camden M 1989 *The Data Bundle*. New Zealand Statistical Association, Wellington, New Zealand

Cook R D 1998a *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York

Cook R D 1998b Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association* **93**: 84–94

Cook R D, Lee H 1999 Dimension reduction in binary response regression. *Journal of the American Statistical Association* **94**: 1187–200

Cook R D, Weisberg S 1991 Sliced inverse regression dimension reduction comment. *Journal of the American Statistical Association* **86**: 328–32

Cook R D, Weisberg S 1999 *Applied Regression Including Computing and Graphics*. Wiley, New York

Ferré L 1998 Determining the dimension of sliced inverse regression and related methods. *Journal of the American Statistical Association* **93**: 132–40

Filliben J J, Li K C 1997 A systematic approach to the analysis of complex interaction patterns in two-level factorial designs. *Technometrics*. **39**: 286–97

Flury B, Riedwyl H 1988 *Multivariate Statistics: A Practical Approach*. Chapman and Hall, London

Hsing T, Carroll R J 1992 An asymptotic theory for sliced inverse regression. *The Annals of Statistics* **20**: 1040–61

Li K C 1991 Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**: 316–27

Li K C 1992 On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association* **87**: 1025–39

Schott J R 1994 Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association* **89**: 141–8

Velilla S 1998 Assessing the number of linear components in a general regression problem. *Journal of the American Statistical Association* **93**: 1088–98

Zhu L X, Fang K T 1996 Asymptotics for kernel estimates of sliced inverse regression. *The Annals of Statistics* **24**: 1053–68

R. D. Cook

Linear Hypothesis: Regression (Quantile)

Quantile regression is a statistical technique intended to estimate, and conduct inference about, conditional quantile functions. Just as classical linear regression methods based on minimizing sums of squared residuals enable one to estimate models for conditional mean functions, quantile regression methods offer a mechanism for estimating models for the conditional median function, and the full range of other conditional quantile functions. By supplementing the estimation of conditional mean functions with techniques for estimating an entire family of conditional quantile functions, quantile regression is capable of providing a more complete statistical analysis of the stochastic relationships among random variables.

Quantile regression has been used in a broad range of application settings. Reference growth curves for childrens' height and weight have a long history in pediatric medicine; quantile regression methods may be used to estimate upper and lower quantile reference curves as a function of age, sex, and other covariates without imposing stringent parametric assumptions on the relationships among these curves. Quantile regression methods have been widely used in economics to study determinents of wages, discrimination effects, and trends in income inequality. Several studies have modeled the performance of public school students on standardized exams as a function of socioeconomic characteristics like their parents' income and educational attainment, and policy variables like class size, school expenditures, and teacher qualifications. It seems rather implausible that such covariate effects should all act so as to shift the entire distribution of test results by a fixed amount. It is of obvious interest to know whether policy interventions alter performance of the strongest students in the same way that weaker students are affected. Such questions are naturally investigated within the quantile regression framework.

In ecology, theory often suggests how observable covariates affect limiting sustainable population sizes, and quantile regression has been used to directly estimate models for upper quantiles of the conditional distribution rather than inferring such relationships from models based on conditional central tendency. In survival analysis, and event history analysis more generally, there is also often a desire to focus attention on particular segments of the conditional distribution, for example survival prospects of the oldest-old, without the imposition of global distributional assumptions.

1. Quantiles, Ranks, and Optimization

We say that a student scores at the τth quantile of a standardized exam if he performs 'better' than the proportion τ, and 'worse' than the proportion $(1-\tau)$, of the reference group of students. Thus, half of the students perform better than the median student, and half perform worse. Similarly, the quartiles divide the population into four segments with equal proportions of the population in each segment. The quintiles divide the population into five equal segments; the deciles into

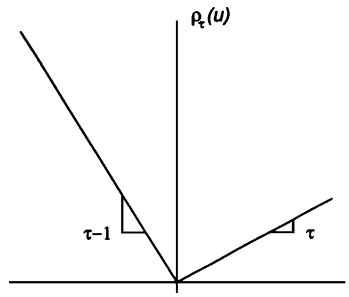

Figure 1
Quantile regression ρ function

10 equal parts. The quantile, or percentile, refers to the general case.

More formally, any real valued random variable, Y, may be characterized by its distribution function:

$$F(y) = \text{Prob}(Y \leqslant y)$$

while for any $0 < \tau < 1$:

$$Q(\tau) = \inf\{y : F(y) \geqslant \tau\}$$

is called the τth quantile of Y. The median, $Q(1/2)$, plays the central role. Like the distribution function, the quantile function provides a complete characterization of the random variable, Y.

The quantiles may be formulated as the solution to a simple optimization problem. For any $0 < \tau < 1$, define the piecewise linear 'check function,' $\rho_\tau(u) = u(\tau - I(u < 0))$ illustrated in Fig. 1.

Minimizing the expectation of $\rho_\tau(Y - \xi)$ with respect to ξ yields solutions, $\hat{\xi}(\tau)$, the smallest of which is $Q(\tau)$.

The sample analogue of $Q(\tau)$, based on a random sample, $\{y_1, \ldots, y_n\}$, of Y's, is called the τth sample quantile, and may be found by solving,

$$\min_{\xi \in \mathrm{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi)$$

While it is more common to define the sample quantiles in terms of the order statistics, $y_{(1)} \leqslant y_{(2)} \leqslant \ldots \leqslant y_{(n)}$, constituting a sorted rearrangement of the original sample, their formulation as a minimization problem has the advantage that it yields a natural generalization of the quantiles to the regression context.

Just as the idea of estimating the unconditional mean, viewed as the minimizer,

$$\hat{\mu} = \text{argmin}_{\mu \in \mathbb{R}} \sum (y_i - \mu)^2$$

can be extended to estimation of the linear conditional mean function $E(Y|X = x) = x'\beta$ by solving;

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum (y_i - x_i'\beta)^2$$

the linear conditional quantile function, $Q_Y(\tau \mid X = x) = x_i'\beta(\tau)$, can be estimated by solving;

$$\hat{\beta}(\tau) = \text{argmin}_{\beta \in \mathbb{R}^p} \sum \rho_\tau(y_i - x'\beta)$$

The median case, $\tau = 1/2$, which is equivalent to minimizing the sum of 'absolute' values of the residuals has a long history. In the mid-eighteenth century Boscovich proposed estimating a bivariate linear model for the ellipticity of the earth by minimizing the sum of absolute values of residuals subject to the condition that the mean residual took the value zero. Subsequent work by Laplace characterized Boscovich's estimate of the slope parameter as a weighted median and derived its asymptotic distribution. F. Y. Edgeworth seems to have been the first to suggest a general formulation of median regression involving a multivariate vector of explanatory variables, a technique he called the 'plural median.' The extension to quantiles other than the median was introduced in Koenker and Bassett (1978).

2. *An Example*

To illustrate the approach, we may consider an analysis of a simple first order autoregressive model for maximum daily temperature in Melbourne, Australia. The data are taken from Hyndman et al. 1996. In Fig. 2 we provide a scatter plot of 10 years of daily temperature data: today's maximum daily temperature is plotted against yesterday's maximum. Our first observation from the plot is that there is a strong tendency for data to cluster along the (dashed) 45 degree line implying that with high probability today's maximum is near yesterday's maximum. But closer examination of the plot reveals that this impression is based primarily on the left side of the plot where the central tendency of the scatter follows the 45 degree line very closely. On the right side, however, corresponding to summer conditions, the pattern is more complicated. There, it appears that either there is another hot day, falling again along the 45 degree line, or there is a dramatic cooling off. But a mild cooling off appears to be more rare. In the language of conditional densities, if today is hot, tomorrow's temperature appears to be bimodal with one mode roughly centered at today's maximum, and the other mode centered at about 20 degrees Celsius.

Several estimated quantile regression curves have been superimposed on the scatterplot. Each curve is specified as a linear B-spline. Under winter conditions these curves are bunched around the 45 degree line, however, in the summer it appears that the upper quantile curves are bunched around the 45 degree line and around 20 degrees. In the intermediate temperatures the spacing of the quantile curves is somewhat greater indicating lower probability of this temperature range. This impression is strengthened by

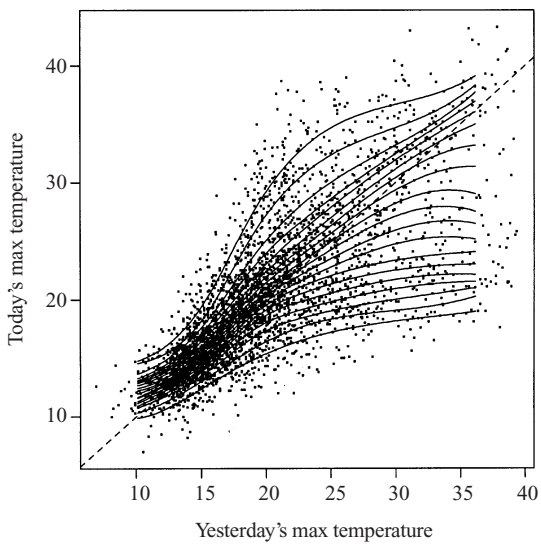

Figure 2
Melbourne maximum daily temperature: the plot illustrates 10 years of daily maximum temperature data (in degrees Celsius) for Melbourne, Australia as an AR(1) scatterplot. The data is scattered around the (dashed) 45 degree line suggesting that today is roughly similar to yesterday. Superimposed on the scatterplot are estimated conditional quantile functions for the quantiles $\tau \in \{.05, .10, \ldots, .95\}$. Note that when yesterday's temperature is high the spacing between adjacent quantile curves is narrower around the 45 degree line and at about 20 degrees Celsius than it is in the intermediate region. This suggests bimodality of the conditional density in the summer

considering a sequence of density plots based on the quantile regression estimates. Given a family of reasonably densely-spaced estimated conditional quantile functions, it is straightforward to estimate the conditional density of the response at various values of the conditioning covariate. In Fig. 3 we illustrate this approach with several density estimates based on the Melbourne data. Conditioning on a low previous day temperature we see a nice unimodal conditional density for the following day's maximum temperature, but as the previous day's temperature increases we see a tendency for the lower tail to lengthen and eventually we see a clearly bimodal density. In this example, the classical regression assumption that the covariates only affect location of the response distribution, but not its scale or shape, is violated.

3. *Interpretation of Quantile Regression*

Least squares estimation of mean regression models asks the question, 'How does the conditional mean of Y depend on the covariates X?' Quantile regression asks this question at each quantile of the conditional distribution enabling one to obtain a more complete description of how the conditional distribution of Y given $X = x$ depends on x. Rather than assuming that covariates only shift the location or scale of the conditional distribution, quantile regression methods enable one to explore potential effects on the shape of the distribution as well. Thus, for example, the effect of a job-training program on the length of the participants' current unemployment spell might be to lengthen the shortest spells while dramatically reducing the probability of very long spells. The mean treatment effect in such circumstances might be small, but the treatment effect on the shape of the distribution of unemployment durations could, nevertheless, be quite significant.

3.1 *Quantile Treatment Effects*

The simplest formulation of quantile regression is the two-sample treatment-control model. In place of the classical Fisherian experimental design model in which the treatment induces a simple location shift of the response distribution, Lehmann (1974) proposed the following general model of treatment response:

> Suppose the treatment adds the amount $\Delta(x)$ when the response of the untreated subject would be x. Then the distribution G of the treatment responses is that of the random variable $X + \Delta(X)$ where X is distributed according to F.

Special cases obviously include the location shift model, $\Delta(X) = \Delta_0$, and the scale shift model, $\Delta(X) = \Delta_0 X$, but the general case is natural within the quantile regression paradigm.

Doksum (1974) shows that if $\Delta(x)$ is defined as the horizontal distance between F and G at x, so

$$F(x) = G(x + \Delta(x))$$

then $\Delta(x)$ is uniquely defined and can be expressed as

$$\Delta(x) = G^{-1}(F(x)) - x$$

Changing variables so $\tau = F(x)$ one may define the quantile treatment effect,

$$\delta(\tau) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau).$$

In the two sample setting this quantity is naturally estimable by

$$\hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_m^{-1}(\tau)$$

where G_n and F_m denote the empirical distribution functions of the treatment and control observations, based on n and m observations, respectively.

Formulating the quantile regression model for the binary treatment problem as,

$$Q_{Y_i}(\tau \mid D_i) = \alpha(\tau) + \delta(\tau) D_i$$

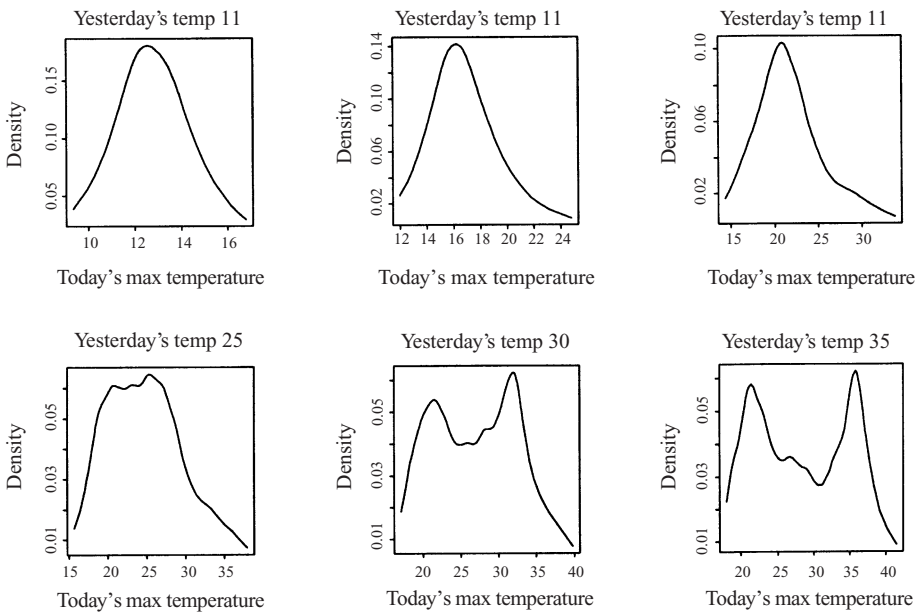

Figure 3
Conditional density estimates of today's maximum temperature for several values of yesterday's maximum temperature based on the Melbourne data. These density estimates are based on a kernel smoothing of the conditional quantile estimates as illustrated in the previous figure using 99 distinct quantiles. Note that temperature is bimodal when yesterday was hot

where D_i denotes the treatment indicator, with $D_i = 1$ indicating treatment, $D_i = 0$, control, then the quantile treatment effect can be estimated by solving,

$$(\hat{\alpha}(\tau), \hat{\delta}(\tau))' = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \delta D_i)$$

the solution $(\hat{\alpha}(\tau), \hat{\delta}(\tau))'$ yields $\hat{\alpha}(\tau) = \hat{F}_n^{-1}(\tau)$, corresponding to the control sample, and

$$\hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_m^{-1}(\tau)$$

Doksum suggests that one may interpret control subjects in terms of a latent characteristic: for example in survival analysis applications, a control subject may be called frail if he is prone to die at an early age, and robust if he is prone to die at an advanced age. This latent characteristic is thus implicitly indexed by τ, the quantile of the survival distribution at which the subject would appear if untreated. And the treatment, under the Lehmann model, is assumed to alter the subjects control response, $\alpha(\tau)$, making it $\alpha(\tau) + \delta(\tau)$ under the treatment. If the latent characteristic, say, the propensity for longevity, were observable *ex ante*, then one could view the treatment effect $\delta(\tau)$ as an explicit interaction with this observable variable. In the absence of such an observable variable, however, the quantile treatment effect may be regarded as a natural measure of the treatment response.

It may be noted that the quantile treatment effect is intimately tied to the two-sample QQ-plot, which has a long history as a graphical diagnostic device. The function $\hat{\Delta}(x) = G_n^{-1}(F_m(x)) - x$ is exactly what is plotted in the traditional two sample QQ-plot. If F and G are identical then the function $G_n^{-1}(F_m(x))$ will lie along the 45 degree line; if they only differ by a location scale shift, then $G_n^{-1}(F_m(x))$ will lie along another line with intercept and slope determined by the location and scale shift, respectively. Quantile regression may be seen as a means of extending the two-sample QQ plot and related methods to general regression settings with continuous covariates.

When the treatment variable takes more than two values, the Lehmann–Doksum quantile treatment effect only requires minor reinterpretation. If the treatment variable is continuous as, for example, in dose-response studies, then it is natural to consider the assumption that its effect is linear, and write,

$$Q_{Y_i}(\tau \mid x_i) = \alpha(\tau) + \beta(\tau) x_i$$

We assume thereby that the treatment effect, $\beta(\tau)$, of changing x from x_0 to $x_0 + 1$ is the same as the

treatment effect of an alteration of x from x_1 to x_1+1. Note that this notion of the quantile treatment effect measures, for each τ, the change in the response required to stay on the τth conditional quantile function.

3.2 Transformation Equivariance of Quantile Regression

An important property of the quantile regression model is that, for any monotone function, $h(\cdot)$;

$$Q_{h(T)}(\tau \mid x) = h(Q_T(\tau \mid x)).$$

This follows immediately from observing that;

$$\text{Prob}(T < t \mid x) = \text{Prob}(h(T) < h(t) \mid x).$$

This equivariance to monotone transformations of the conditional quantile function is a crucial feature, allowing one to decouple the potentially conflicting objectives of transformations of the response variable. This equivariance property is in direct contrast to the inherent conflicts in estimating transformation models for conditional mean relationships. Since, in general, $E(h(T) \mid x) \neq h(E(T \mid x))$, the transformation alters in a fundamental way what is being estimated in ordinary least squares regression.

A particularly important application of this equivariance result, and one that has proven extremely influential in the econometric application of quantile regression, involves censoring of the observed response variable. The simplest model of censoring may be formulated as follows. Let y_i^* denote a latent (unobservable) response assumed to be generated from the linear model;

$$y_i^* = x_i'\beta + u_i \quad i = 1, \ldots, n$$

with $\{u_i\}$ iid from distribution function F. Due to censoring, the y_i^*'s are not observed directly, but instead one observes;

$$y_i = \max\{0, y_i^*\}$$

Powell (1986) noted that the equivariance of the quantiles to monotone transformations implied that in this model the conditional quantile functions of the response only depended on the censoring point, but were independent of F. Formally, the τth conditional quantile function of the observed response, y_i, in this model may be expressed as:

$$Q_i(\tau \mid x_i) = \max\{0, x_i'\beta + F_u^{-1}(\tau)\}$$

The parameters of the conditional quantile functions may now be estimated by solving;

$$\min_b \sum_{i=1}^{n} \rho_\tau(y_i - \max\{0, x_i'b\})$$

where it is assumed that the design vectors x_i, contain an intercept to absorb the additive effect of $F_u^{-1}(\tau)$. This model is computationally somewhat more demanding than conventional linear quantile regression because it is nonlinear in parameters.

3.3 Robustness

Robustness to distributional assumptions is an important consideration throughout statistics, so it is important to emphasize that quantile regression inherits robustness properties of the ordinary sample quantiles. The estimates and the associated inference apparatus have an inherent distribution-free character since quantile estimation is influenced only by the local behavior of the conditional distribution of the response near the specified quantile. Given a solution $\hat{\beta}(\tau)$, based on observations, $\{y, X\}$, as long as one doesn't alter the sign of the residuals, any of the y observations may be arbitrarily altered without altering the initial solution. Only the signs of the residuals matter in determining the quantile regression estimates, and thus outlying responses influence the fit in so far as they are either above or below the fitted hyperplane, but how far above or below is irrelevant.

While quantile regression estimates are inherently robust to contamination of the response observations, they can be quite sensitive to contamination of the design observations, $\{x_i\}$. Several proposals have been made to ameliorate this effect.

4. Computational Aspects of Quantile Regression

Although it was recognized by a number of early authors, including Gauss, that solutions to the median regression problem were characterized by an exact fit through p sample observations when p linear parameters are estimated, no effective algorithm arose until the development of linear programming in the 1940s. It was then quickly recognized that the median regression problem could be formulated as a linear program, and the simplex method employed to solve it. The algorithm of Barrodale and Roberts (1974) provided the first efficient implementation specifically designed for median regression and is still widely used in statistical software. It can be concisely described as follows. At each step, we have a trial set of p 'basic observations' whose exact fit may constitute a solution. We compute the directional derivative of the objective function in each of the $2p$ directions that correspond to removing one of the current basic observations, and taking either a positive or negative step. If none of these directional derivatives are negative the solution has been found, otherwise one chooses the most negative, the direction of steepest descent, and goes in that direction until the objective function ceases to decrease. This one dimensional search can be formulated as a problem of finding the

solution to a scalar weighted quantile problem. Having chosen the step length, we have in effect determined a new observation to enter the basic set, a simplex pivot occurs to update the current solution, and the iteration continues.

This modified simplex strategy is highly effective on problems with a modest number of observations, achieving speeds comparable to the corresponding least squares solutions. But for larger problems with, say $n > 100{,}000$ observations, the simplex approach eventually becomes considerably slower than least squares. For large problems recent development of interior point methods for linear programming problems are highly effective. Portnoy and Koenker (1997) describe an approach that combines some statistical preprocessing with interior point methods and achieves comparable performance to least squares solutions even in very large problems.

An important feature of the linear programming formulation of quantile regression is that the entire range of solutions for $\tau \in (0, 1)$ can be efficiently computed by parametric programming. At any solution $\hat{\beta}(\tau_0)$ there is an interval of τ's over which this solution remains optimal, it is straightforward to compute the endpoints of this interval, and thus one can solve iteratively for the entire sample path $\hat{\beta}(\tau)$ by making one simplex pivot at each of the endpoints of these intervals.

5. Statistical Inference for Quantile Regression

The asymptotic behavior of the quantile regression process $\{\hat{\beta}(\tau): \tau \in (0, 1)\}$ closely parallels the theory of ordinary sample quantiles in the one sample problem. Koenker and Bassett (1978) show that in the classical linear model;

$$y_i = x_i\beta + u_i$$

with u_i iid from dfF, with density $f(u) > 0$ on its support $\{u \mid 0 < F(u) < 1\}$, the joint distribution of $\sqrt{n}(\hat{\beta}_n(\tau_i) - \beta(\tau_i))_{i=1}^m$ is asymptotically normal with mean 0 and covariance matrix $\Omega \otimes D^{-1}$. Here $\beta(\tau) = \beta + F_u^{-1}(\tau)e_1$, $e_1 = (1, 0, \ldots, 0)'$, $x_{1i} \equiv 1$, $n^{-1}\sum x_i x_i' \to D$, a positive definite matrix, and

$$\Omega = (\omega_{ij} = (\min\{\tau_i, \tau_j\} - \tau_i\tau_j)/(f(F^{-1}(\tau_i)f(F^{-1}(\tau_j)))$$

When the response is conditionally independent over i, but not identically distributed, the asymptotic covariance matrix of $\xi(\tau) = \sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$ is somewhat more complicated. Let

$$\xi_i(\tau) = x_i\beta(\tau)$$

denote the conditional quantile function of y given x_i, and $f_i(\cdot)$ the corresponding conditional density, and define;

$$J_n(\tau_1, \tau_2) = (\min\{\tau_1, \tau_2\} - \tau_1\tau_2)n^{-1}\sum_{i=1}^{n} x_i x_i'$$

and

$$H_n(\tau) = n^{-1}\sum x_i x_i' f_i(\xi_i(\tau))$$

Under mild regularity conditions on the $\{f_i\}$'s and $\{x_i\}$'s, we have joint asymptotic normality for vectors $(\xi(\tau_i), \ldots, \xi(\tau_m))$ with mean zero and covariance matrix

$$V_n = (H_n(\tau_i)^{-1}J_n(\tau_i, \tau_j)H_n(\tau_j)^{-1})_{i=1}^m$$

An important link to the classical theory of rank tests was made by Gutenbrunner and Jurečková (1992), who showed that the rankscore functions of Hájek and Šidák (1967) could be viewed as a special case of a more general formulation for the linear quantile regression model. The formal dual of the quantile regression linear programming problem may be expressed as;

$$\max\{y'a \mid X'a = (1-t)X'1,\ a \in [0, 1]^n\}$$

The dual solution $\hat{a}(\tau)$ reduces to the Hájek and Šidák rankscores process when the design matrix, X, takes the simple form of an n vector of ones. The regression rankscore process $\hat{a}(\tau)$ behaves asymptotically much like the classical univariate rankscore process, and thus offers a way to extend many rank based inference procedures to the more general regression context.

6. Extensions and Future Developments

There is considerable scope for further development of quantile regression methods. Applications to survival analysis and time-series modeling seem particularly attractive, where censoring and recursive estimation pose interesting challenges, respectively. For the classical latent variable form of the binary response model where;

$$y_i = I(x_i'\beta + u_i \geqslant 0)$$

and the median of u_i conditional on x_i is assumed to be zero for all $i = 1, \ldots, n$, Manski (1985) proposed an estimator solving,

$$\max_{\|b\|=1} \sum (y_i - 1/2)I(x_i'b \geqslant 0)$$

This 'maximum score' estimator can be viewed as a median version of the general linear quantile regression estimator for binary response;

$$\min_{\|b\|=1} \sum \rho_\tau(y_i - I(x_i'b \geqslant 0))$$

In this formulation it is possible to estimate a family of quantile regression models and explore, semi-parametrically, a full range of linear conditional quantile functions for the latent variable form of the binary response model.

Koenker and Machado (1999) introduce inference methods closely related to classical goodness of fit

statistics based on the full quantile regression process. There have been several proposals dealing with generalizations of quantile regression to nonparametric response functions involving both local polynomial methods and splines. Extension of quantile regression methods to multivariate response models is a particularly important challenge.

7. Conclusion

Classical least squares regression may be viewed as a natural way of extending the idea of estimating an unconditional mean parameter to the problem of estimating conditional mean 'functions,' the crucial step is the formulation of an optimization problem that encompasses both problems. Likewise, quantile regression offers an extension of univariate quantile estimation to estimation of conditional quantile functions via an optimization of a piecewise linear objective function in the residuals. Median regression minimizes the sum of absolute residuals, an idea introduced by Boscovich in the eighteenth century.

The asymptotic theory of quantile regression closely parallels the theory of the univariate sample quantiles; computation of quantile regression estimators may be formulated as a linear programming problem and efficiently solved by simplex or barrier methods. A close link to rank based inference has been forged from the theory of the dual regression quantile process, or regression rankscore process.

Recent nontechnical introductions to quantile regression are provided by Buchinsky (1998) and Koenker and Hallock (2001). Most of the major statistical computing languages now include some capabilities for quantile regression estimation and inference. Quantile regression packages are available for R and Splus from the R archives at http://lib.stat.cmu.edu/R/CRAN and Statlib at http://lib.stat.cmu.edu/S, respectively. Stata's central core provides quantile regression estimation and inference functions. SAS offers some, rather limited, facilities for quantile regression.

Bibliography

Barrodale I, Roberts F D K 1974 Solution of an overdetermined system of equations in the l_1 norm. *Communications ACM* **17**: 319–20

Buchinsky M 1998 Recent advances in quantile regression models: a practical guide for empirical research. *Journal of Human Resources* **33**: 88–126

Doksum K 1974 Empirical probability plots and statistical inference for nonlinear models in the two sample case. *Annals of Statistics* **2**: 267–77

Gutenbrunner C, Jurečková J 1992 Regression quantile and regression rank score process in the linear model and derived statistics. *Annals of Statistics* **20**: 305–30

Hájek J, Sidák Z 1967 *Theory of Rank Tests*. Academia, Prague, Czech Republic

Hyndman R J, Bashtannyk D M, Grunwald G K 1996 Estimating and visualizing conditional densities. *Journal of Computational. and Graphical Statistics* **5**: 315–36

Koenker R, Bassett G 1978 regression quantiles. *Econometrica* **46**: 33–50

Koenker R, Hallock K 2001 Quantile Regression. *Journal of Economic, Perspectives*, forthcoming

Lehmann E 1974 *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco

Manski C 1985 Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics* **27**: 313–34

Portnoy S, Koenker R 1997 The Gaussian hare and the Laplacian tortoise: computability of squared-error vs. absolute-error estimators, with discussion. *Statistical Science* **12**: 279–300

Powell J L 1986 Censored regression quantiles. *Journal of Econometrics* **32**: 143–55

R. Koenker

Linguistic Anthropology

'Linguistic anthropology' is an interdisciplinary field dedicated to the study of language as a cultural resource and speaking as a cultural practice. It assumes that the human language faculty is a cognitive and a social achievement that provides the intellectual tools for thinking and acting in the world. Its study must be done by detailed documentation of what speakers say as they engage in daily social activities. This documentation relies on participant observation and other methods, including audiovisual recording, annotated transcription, and interviews with participants.

As an interdisciplinary field, linguistic anthropology has often drawn from and participated in the development of other theoretical paradigms. Some of its own history is reflected in the oscillation often found among a number of terms that are not always synonyms: linguistic anthropology, anthropological linguistics, ethnolinguistics, and sociolinguistics. Its main areas of interest have changed over the years, from an almost exclusive interest in the documentation of the grammars of aboriginal languages to the analysis of the uses of talk in everyday interaction and throughout the life span (Duranti 1997, Foley 1997). This article provides a brief historical account of linguistic anthropology, and highlights important past and present issues, theories, and methods.

1. Linguistic Anthropology within the Boasian Tradition

In the holistic tradition established by Franz Boas (1858–1942) in the USA at the beginning of the twentieth century, anthropology was conceived as

comprising four subfields: archaeology, physical (now 'biological') anthropology, linguistics (now 'linguistic anthropology'), and ethnology (now 'sociocultural anthropology'). This vision of anthropology differs from the one found in the European tradition, where linguistics and social anthropology remained rigidly separate disciplines for most of the twentieth century, despite the emphasis on the use of native languages in fieldwork among UK anthropologists, and the theoretical and methodological influence of Bronislaw Malinowski (1884–1942), who wrote about the importance of linguistic research for an anthropological understanding of human societies. In the 1950s, the adoption of the term 'ethnolinguistics' (reflecting the European preference for 'ethnology' over '(cultural) anthropology') for those studies that merged linguistic and anthropological interests signaled the intellectual recognition, at least in some European academic circles, of the importance of an 'ethnological side' of linguistic studies (Cardona 1976), but the institutional recognition of such a discipline within European anthropology has been slow to come. European scholars with research interests similar to those of North American linguistic anthropologists are thus more likely to be found in departments of linguistics, foreign languages and literatures, folklore, communication, sociology, or psychology.

To understand the special role given to the study of languages in the Boasian tradition, we must go back to the time when anthropology became a profession in the USA, in the period between the last decades of the nineteenth and the first decades of the twentieth century. In that time, the study of American Indian languages emerged as an essential part of anthropological research. John Wesley Powell (1834–1902), the founder of the Bureau of Ethnology, later renamed Bureau of American Ethnology (BAE), supported, through grants from the US government, linguistic fieldwork, in the belief that by collecting vocabularies and texts from American Indian languages, it would be possible to reconstruct their genetic relations and thus help in the classification of American Indian tribes. Boas himself had become fascinated by the grammatical structures of Chinook and other languages of the American northwest coast early on in his fieldwork, and seized the opportunity to work for the BAE and edit the *Handbook of American Indian Languages* (1911).

Although Boas, a diffusionist, was quite skeptical of the possibility of using languages for reconstructing genetic relations between tribes—and was against any correlation between language and race—he tried to transmit to his students a passion for the details of linguistic description and the conviction that languages were an important tool for (a) fieldwork, and (b) the study of culture, especially because the categories and rules of language are largely unconscious and thus not subject to secondary rationalizations. Furthermore, Boas was committed to what later became known as 'salvage anthropology,' that is the documentation of languages and cultural traditions that seemed on the verge of disappearing. This enterprise—a struggle against time due to the great damage already done, by the end of the nineteenth century, to the indigenous cultures of the Americas by European colonizers—produced valuable information on Native American traditions, but it had its methodological and theoretical drawbacks, the most flagrant of which was the inability to see or accept the effects of cultural contact and colonization.

Through his writing and teaching, Boas brought scientific rigor to linguistic description and helped demolish a number of unfounded stereotypes about the languages that were then called 'primitive.' In an 1889 article entitled 'On alternating sounds,' Boas argued that the commonly held view that speakers of American Indian languages were less accurate in their pronunciation than speakers of Indo-European languages was false and probably due to the lack of linguistic sophistication of those who had first tried to describe indigenous languages. Consistent with his cultural relativism, Boas believed that each language should be studied on its own terms rather than according to some preset categories based on the study of other, genetically unrelated languages (e.g., Latin). In his 'Introduction' to the *Handbook* (1911), Boas provided an overview of the grammatical categories and linguistic units necessary for the analysis of American Indian languages and argued against overgeneralizations that would obscure differences across languages. He identified the sentence (as opposed to the word) as the unit for the expression of ideas, and listed a number of grammatical categories that are likely to be found in all languages, while pointing out that the material content of words (the meaning of lexical items) is language-specific and that languages classify reality differently. One language might express the semantic connections among words pertaining to the same semantic field by modifying one basic stem, whereas another language might have words that are etymologically completely unrelated. As examples of the latter type, Boas (1911) mentioned the different words that are used in English for concepts centered around the idea of 'water'—'lake,' 'river,' 'brook,' 'rain,' 'dew,' 'wave,' 'foam'—and four different words for concepts based on 'snow' in Eskimo. These examples were later taken out of context and the number of words for 'snow' in Eskimo (languages) grew larger and larger over the next decades in both academic and popular publications.

It was Edward Sapir (1884–1939) who, more than any other of Boas's students, further developed Boas's interest in grammatical systems and their potential implications for the study of culture, and trained a new generation of experts of American Indian languages (e.g., Mary Haas, Morris Swadesh, Benjamin Lee Whorf, Carl Voegelin). Unlike Boas, however, Sapir was not a four-field anthropologist. He wrote and

lectured on culture and personality, but had no interest in archaeology or physical anthropology, which he saw as being more appropriately housed in museums than in anthropology departments. Furthermore, whereas Boas was skeptical of genetic reconstruction and tended to favor acculturation as a cause of similarities between languages, Sapir was a strong believer in the power of the comparative method, which he used to reconstruct Proto-Athabascan and posit the Na-Déné linguistic group (comprising Athabascan, Tlingit, and Haida).

While he was at Yale, Sapir encouraged his students to go into linguistics rather than anthropology, and it is no accident that they ended up calling themselves 'anthropological linguists.' They thought of themselves primarily as linguists, as demonstrated by their dedication to the study of the grammatical structures of American Indian languages (and other previously undocumented languages). Their emphasis on fieldwork and their preference for historical and descriptive linguistics kept them separate from the new linguistics of the 1960s, Chomsky's generative grammar. To anthropological linguists such as Mary Haas the advent of this new paradigm was a threat because (a) it seemed more committed to linguistic theory than to languages, and in fact devalued grammatical description *per se*; (b) it was (especially at the beginning) almost exclusively based on English—Chomsky argued that one could posit universals of language by working on one language; and (c) it predicated the need for linguists to work on their own intuitions instead of working with native speakers or inferring grammatical rules on the basis of a corpus of elicited texts (Haas 1987). This last point was particularly problematic for those students of Native American languages who often had only one or two old speakers to work with and could not find younger speakers to train in linguistic theory and methods.

1.1 Linguistic Relativity in the History of Linguistic Anthropology

Linguistic relativity is a general term used to refer to various hypotheses or positions about the relationship between language and culture (see *Sapir–Whorf Hypothesis*). Although Sapir and Whorf differed in their discussion of the relationship between language and culture, and never produced a joint formulation of what is meant by linguistic relativity, there is no question that the themes and issues often identified as linguistic relativity are the continuation of the Boasian paradigm. First, Sapir and Whorf followed Boas's intellectual curiosity for the indigenous languages of North America as a way of channeling a more general fascination for alternative ways of being in the world and the desire to make sense of those ways. Second, to the extent to which it started from an emphasis on human diversity, linguistic relativity was related to cultural relativism, if not a corollary of it. It was accompanied by a concern for the proper representation of grammatical systems that could not be described using the categories of European languages. Third, the same antiracist attitude that characterized Boas's views on human diversity seemed to motivate the lack of value judgment associated with linguistic diversity.

For Sapir, linguistic relativity was a way of articulating what he saw as the struggle between the individual and society (Mandelbaum 1949). In order to communicate their unique experiences, individuals need to rely on a public code over which they have little control. Linguistic rules are usually unconscious, and it is difficult for individual speakers to enter the logic of the linguistic system and alter it to their liking. In this perspective, linguistic relativity becomes a way of exploring the power that words have over individuals and groups. It is thus a precursor to more recent topics in linguistic anthropology, such as language ideologies (see Sect. 4.3).

Sapir never developed the conceptual framework or methodology for testing the implications of these intuitions about the language faculty. This task was left to another important figure in the history of linguistic anthropology, Benjamin Lee Whorf (1897–1941), a chemical engineer who worked as an insurance inspector, taught himself linguistics, and after 1931 entered into contact with Sapir and his students at Yale. Although Whorf started out sharing several of the basic positions held by Boas and Sapir on the nature of linguistic classification, he developed his own conceptual framework, which included the distinction between *overt* and *covert* grammatical categories, and an important analytical tool for understanding what kinds of categorical distinctions speakers are sensitive to—this issue was later further developed in the work on metapragmatics (Whorf 1956). Contrary to popular belief, Whorf was not so much concerned with the number of words for the same referent (e.g., 'snow') in different languages, but with the implications that different grammatical systems and lexicons have for the way in which speakers make inferences about the world. He believed that ways of thinking may develop by analogy with 'fashions of speaking,' a concept that was later revived by Hymes's notion of 'ways of speaking.'

Whorf's work was harshly criticized in the 1960s and 1970s, especially after the publication of Berlin and Kay's (1969) study of color terminology, in which they claimed that lexical labels for basic color terms are not arbitrary but follow universal principles. But more recent studies have given support to some of Whorf's ideas (Lucy 1992), and even the universality of basic color terminology and its innate perceptual saliency have been questioned (e.g., Levinson 2000). Sapir and Whorf's ideas about the unconscious aspects of linguistic codes continued to play an important part in the history of linguistic anthropology, and re-

appeared in the 1980s in the context of a number of research projects, including the study of language ideology (Kroskrity 2000).

2. Linguistic Anthropology in the 1960s

In the same years in which Chomsky's generative grammar was becoming popular in the USA, two other important programs were also launched: the ethnography of communication and urban sociolinguistics (see *Sociolinguistics*).

Through his own qualitatively oriented work and his collaboration with Charles Ferguson on multilingualism in South Asia, John Gumperz criticized the notion of 'language' as used by linguists, and introduced the notions of variety, repertoire, and linguistic community or speech community. In the same years, Dell Hymes launched a call for a comparative study of communicative events to capture the ways in which speaking is a cultural activity and should be studied as such. His collaboration with Gumperz produced a new paradigm in linguistic anthropology, one in which researchers were expected to study both knowledge and use of languages through ethnographic methods (Gumperz and Hymes 1964). The object of inquiry was no longer the grammars of indigenous languages, but communicative events and contextual variation within and across speech communities. Although Hymes stressed the need to see linguistics as part of anthropology, and insisted on the name 'linguistic anthropology' over 'anthropological linguistics,' he also helped defined a domain of inquiry that was in many ways independent of both linguistics and the other subdisciplines within anthropology. Whether intended or not, this effort resulted in a type of linguistic anthropology that was much less preoccupied with grammatical description and linguistic reconstruction than the one practiced by the previous generations, and more focused on the performance of language and consequently its aesthetic and political dimensions. This emphasis on the actual use of language allowed for great progress in the understanding of the cultural organization of speaking, but left scholars in other fields worrying about linguistic issues that were still central to anthropology, such as language evolution. While Chomsky was leading an antibehavioristic 'cognitive revolution,' Gumperz and Hymes were trying to develop a paradigm in which language behavior could be fully explored as social activity. In this perspective, the notion of communicative competence was central (see *Communicative Competence: Linguistic Aspects*).

2.1 From Communicative Competence to Performance

The documentation of variation and the observation of the social life of speech had taught fieldworkers that Chomsky's notion of ideal speaker-hearer was problematic for many reasons, including the fact that not all members of a speech community have access to the same verbal resources or the same opportunities to use them. In his writings, Hymes shifted the focus from Universal Grammar to linguistic diversity. This change had methodological implications. From eliciting grammatical patterns and speech genres (e.g., myths, traditional stories, proverbs, riddles) from native speakers, linguistic anthropologists shifted their attention to the actual use of language in social interaction. Gumperz and Hymes's students went to the field with the goal of studying how language was being used in different events (e.g., ceremonies, village councils, classrooms) and different places (e.g., at a street corner, market, dinner table). Linguistic performance—the use of language (or its significant absence)—became the starting point of any investigation (Bauman and Sherzer 1974).

Hymes and others reconceptualized performance as a realm of social action, which emerges out of the interaction with other speakers and as such is not reducible to the use of the linguistic knowledge controlled by one individual. The focus on performance allowed fieldworkers to recognize the creative dimensions of any act of speaking, the role of individuals and groups in the reproduction and transformation of linguistic codes and the institutions they support, the responsibility associated with any display of one's verbal skills, and the interactive construction of messages and meanings.

3. Context

In an ethnographic approach, language is seen as permeable to social situations and social roles while at the same time helping to define those situations and roles. This means that linguistic anthropologists need analytical tools and units of analysis to define the context of speech. With this goal in mind, in the late 1960s Hymes expanded Jakobson's model of the speech event, and proposed the 'SPEAKING Model,' with each letter of the word 'speaking' standing for a cluster of dimensions of speech events that ethnographers of communication should explore: Situation, Participants, Ends, Act Sequences, Key, Instrumentalities, Norms, and Genre (Hymes 1967 reprinted in Hymes 1974). This list was meant to provide an etic grid for comparative purposes, and although it was rarely used in its entirety or formally adopted by researchers, it helped many to identify dimensions of verbal communication that had been left out of previous investigations. In the meantime, Gumperz refined his analysis of multilingualism and language contact, and became interested in the analysis of face-to-face interaction. He expanded the concept of linguistic repertoire to include the range of resources that speaker-hearers use to make inferences about the ongoing context. This line of inquiry resulted in the

notion of contextualization cues, linguistic features through which 'speakers signal and listeners interpret what the activity is, how semantic content is to be understood and how each sentence relates to what precedes or follows' (Gumperz 1982, p. 131). Typical examples of contextualization cues are intonational patterns, paralinguistic features (e.g., tempo, pausing), choice of code (e.g., English vs. Spanish), use of key words, and formulaic expressions. They can be studied in order to make sense of both successful and unsuccessful communication—'crosstalk' was the name with which miscommunication between people with different cultural backgrounds came to be known (thanks especially to a BBC program centered around Gumperz's research on South Asian speakers of English in the UK).

In the late 1970s and early 1980s, the notion of context was also being revisited by a number of other researchers within linguistic anthropology. Michael Silverstein (1976) proposed a model of indexicality that could be adapted to the sociocultural study of language in and as context. He identified linguistic signs along a continuum from highly presupposing (i.e., they can be interpreted only on the basis of an existential connection with some independently established aspect of context) to highly creative (i.e., constitutive of their own context). Deictic terms like 'this' in 'this room is cold' are presupposing because it is assumed that the room must exist for the utterance to be interpretable, whereas second-person pronouns (e.g., 'you' in English, *tu* in Spanish) are creative, given that they establish the identity of the addressee/recipient while simultaneously creating the role of addressee/recipient in the ongoing speech event. Languages that have socially differentiated second-person pronouns (e.g., the classic T/V type of distinction of many European languages, French *tu/vous*, Spanish *tu/Usted*, German *du/Sie*, and Italian *tu/Voi* or *tu/Lei*) are more extreme examples of systems in which words are used to activate or establish the relevant social coordinates of equality/inequality, solidarity/power. The study of indexicality has become a major focus of interest in contemporary linguistic anthropology, and has been accompanied by a renewed interest in the role of the human body in the establishment of the referential grounding of most communicative acts (Hanks 1990).

A more recent effort toward the definition of context is Elinor Ochs's (1996) model for the construction of social identities, which is based on a number of situational dimensions established through language use: social acts, activities (a sequence of two or more acts), and affective and epistemic stances.

The 1990s saw a rethinking of the concept of context in part due to: (a) a renewed awareness of the role of theory and methodology in defining the difference between the message and its context (Duranti and Goodwin 1992); (b) the influence of a number of theorists from other disciplines (e.g., Pierre Bourdieu, Anthony Giddens, Mikhail Bakhtin, Michel Foucault, Clifford Geertz, Paul Ricoeur); (c) the use of the concept of ideology in trying to understand how speakers conceptualize what constitutes appropriate and interpretable language; and (d) the introduction and wider adoption of new recording technologies (e.g., video, digitized images), and their implications for the definition of what constitutes an empirically adequate representation of speaking or, more broadly, communication.

4. *New Trends*

The last two decades have seen the development of several new projects involving the interface between language and other cultural resources. Three of the most important ones are language socialization, multilingualism, and the study of the linguistic dimensions of power and control.

4.1 *Language Socialization*

Although the acquisition of communicative competence was always meant to be an important part of the program in the ethnography of communication, the field of language socialization did not fully develop until the mid-1980s, when Elinor Ochs and Bambi B. Schieffelin (1984) defined it as (a) the process of getting socialized *through* language and (b) the process of getting socialized *to* language, and offered some specific directions for research. By applying an anthropological reading to prior work on language acquisition, they reframed it as embedded in culturally specific expectations about the role of children and adults in Western societies and, particularly, in white middle-class families. Using their discovery that neither one of the two speech communities they had studied (i.e., in Papua New Guinea and in Samoa) have a register corresponding to what is known as 'baby talk' or 'motherese,' Ochs and Schieffelin not only demonstrated that simplification in talking to infants, contrary to what was suggested by some linguists and psycholinguists, is not universal, but also, and more importantly, that simplification in talking to infants correlates with other forms of accommodation to children, and local conceptualizations of children and their place in society.

Although related to child language acquisition studies, language socialization studies examine the cultural implications of what is being done with, to, around, and through talk to children, with the theoretical assumption that learning is a two-way street, and that both experts and novices may come out of routine social encounters with new ways of thinking, acting, and feeling.

Language socialization is conceptualized as a never-ending process because speakers never stop learning

new ways of using language, for example in school, at work, in church, at play. Among the various forms of 'secondary socialization,' literacy has occupied an important role in linguistic anthropology. The work of Shirley Brice Heath (1983) has been very influential in showing the benefits of an ethnographic study of literacy practices in the home, where children are exposed to literacy in ways that may or may not be precursors of the type of activities they will encounter in school.

4.2 *Multilingualism: New Perspectives on Contact and Change*

In the 1980s the writings of Mikhael Bakhtin (and one of his alter egos, Valentin Voloshinov) were particularly influential for their conceptualization of meaning as a joint activity, the attention given to coexisting styles and voices within the same 'text,' and the identification of both centripetal (toward unity and standardization) and centrifugal forces (away from unity and standardization) in language use. The notion of a unitary language then becomes both empirically and ideologically suspect because it hides from us the inequality inherent in any linguistic system, as well as the potential and actual aesthetic effects of the juxtaposition of multiple voices and coexisting language varieties. Bakhtin's work inspired a number of linguistic anthropologists including Jane and Kenneth Hill (1986), who introduced the notion of syncretic language to describe the mixing of grammars that takes place in contemporary Mexicano (Nahuatl) (see *Code Switching: Linguistic*). Even when speakers are no longer considered bilingual, some aspects of their 'lost' language and its cultural contexts are maintained, sometimes in occasional code switches and in a variety of other hybrid constructions. It is the task of the researcher to find out what survives of the old code and under what conditions it reappears in spoken or written discourse. Communities differ in the extent to which they recognize the presence of alternative ways of speaking. Ideological positions based on linguistic purism, enforcement of national identity, and control over ethnic boundaries play an important role in the types of language varieties that are supported or oppressed (Schieffelin et al. 1998).

4.3 *Power and Control*

The study of speaking as a cultural practice cannot be made without encountering the issue of how language can be used to control the action of others. A number of contributions within linguistic anthropology have dealt with this issue.

Building on Goffman's notion of 'face work' and Grice's conversational maxims, Penelope Brown and Stephen Levinson (1987, first published in 1978) presented a theory of politeness as a set of strategies based on rational principles used to mitigate 'face threatening acts.' (see *Politeness and Language*). In this theory, language plays a crucial role in mediating differences in power between speakers.

Other approaches have stressed the lack of control that speakers have over their linguistic resources. For example, the analysis of traditional oratory by Maurice Bloch defined it as a coercing system within which speakers could only reproduce the existing power relations. Earlier work on language and gender also uncovered some of the implications in linguistic codes and linguistic routines that are responsible for defining female speakers not only as different, but as weak, unassertive, or submissive (see *Language and Gender*). More recent work has questioned some of these findings, at least as generalized statements about women, and stressed the importance of careful analysis of face-to-face encounters. Marjorie H. Goodwin (1990) found the girls in her study as assertive and confrontational as boys. But she also discovered that there were some differences in the interactional strategies used by boys and girls. Among girls, offenses were constructed out of reported deeds, and especially reported speech, by absent parties (in the so-called 'he-said-she-said' sequences).

As researchers improve their understanding of the subtle functioning of different language varieties (e.g., codes, dialects, registers, genres, styles) in the definition of social identities (e.g., gender, race, ethnicity), they unveil the active roles that speakers play in adapting existing linguistic resources to their interactional goals, and their ability simultaneously to index multiple social worlds and their associated identities (Hall and Bucholtz 1995, Zentella 1997).

The power of new technologies in the definition of persons and their rights was a central theme of Michel Foucault's historical analysis of the development of asylums and other institutions that dealt with health in France. In a similar vein, but using detailed analysis of face-to-face encounters in which participants communicate through talk, gestures, and the use of material artifacts, Charles Goodwin (1994) identified a series of interpretive procedures (e.g., 'coding,' 'highlighting') which use particular types of inscription techniques to constitute what he calls 'professional vision.'

5. *Methods*

In the last few decades of the twentieth century, there was considerable improvement in the tools used in documenting language use. Whereas descriptions of verbal activities such as greetings, proverbs, insults, and speechmaking used to be based on participant observation or on work with native speakers, today researchers are expected to have recordings of exchanges in which the phenomena they describe are occurring spontaneously. As the technology for visual documentation improves and becomes more access-

ible, we are able to notice phenomena (e.g., synchronization between talk and gestures) which used to be missed in past analysis of verbal communication. At the same time, audiovisual documentation has also increased the level of intrusion into people's lives. This means that researchers must be ever more aware of the social and ethical dimensions of fieldwork. The relationship between researchers and their subjects is as delicate and as important as any other human relationship and as such requires care, mutual respect, and honesty. If the goal of our study is a better understanding of the role played by language in the human condition, we must be guided in our efforts by the desire to improve our communication across social and cultural boundaries. This must also apply to our fieldwork situation and our relationship with the speech communities we want to study.

Fieldworkers' participation in the social life of the community must be recorded as systematically as possible. This is done by writing field notes and by transcribing recordings of social encounters, activities, and events. Field notes are important because they provide researchers with a chance to document important information (which is soon forgotten if not written down) and reflect on what they have just experienced. Transcription is equally important because it allows researchers to fix on paper (or on a computer screen) salient aspects of interactions that can then be interpreted, translated, collected, and compared. Transcription is thus a particular type of what Ricoeur called 'inscription,' that is an abstraction and a fixing of something that by nature is or was moving across time and space. Linguistic anthropologists strive to produce rich transcripts by relying on native speakers who have the necessary cultural background to provide the information necessary to make sense of what is being said. There are many different ways of transcribing speech and nonverbal communication, and it is important for researchers to become familiar and experiment with more than one way before choosing the one that better fits their research goals and needs. For example, those who are interested in grammatical analysis must provide word-by-word glosses; for those who are interested in the relation between speech and the spatial organization of the event, visual representations of the settings become crucial; a transcript that utilizes phonetic symbols is appropriate when writing for linguists, but would be too hard to decipher for anyone else. Similarly, a transcript that tries to cover most of the information available to the participants at the time of speaking would be too cumbersome and equally hard to interpret. More generally, a transcript is always work in progress. It constitutes a first analysis of the data collected. It forces us to make important decisions about what is salient in an interaction and, at the same time, while being produced or once completed, it can reveal phenomena that we might otherwise have missed.

See also: Anthropology, History of; Area and International Studies: Linguistics; Boas, Franz (1858–1942); Discourse, Anthropology of; Ethnography; Ethnology; Foucault, Michel (1926–84); Historical Linguistics: Overview; Linguistic Fieldwork; Linguistics: Overview; Malinowski, Bronislaw (1884–1942); Sapir, Edward (1884–1939); Sapir–Whorf Hypothesis; Sociolinguistics

Bibliography

Bauman R, Sherzer J 1974 *Explorations in the Ethnography of Speaking*. Cambridge University Press, Cambridge, UK

Berlin B, Kay P 1969 *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA

Boas F 1911 Introduction. In: Boas F (ed.) *Handbook of American Indian Languages*. Smithsonian Institution and Bureau of American Ethnology, Washington, DC, Vol. BAE-B 40, pt. 1, pp. 1–83

Brown P, Levinson S C 1987 *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK

Duranti A 1997 *Linguistic Anthropology*. Cambridge University Press, Cambridge, UK

Duranti A, Goodwin C (eds.) 1992 *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge University Press, Cambridge, UK

Foley W A 1997 *Anthropological Linguistics: An Introduction*. Blackwell, Malden, MA

Goodwin C 1994 Professional vision. *American Anthropologist* **96**: 606–33

Goodwin M H 1990 *He-Said-She-Said: Talk as Social Organization among Black Children*. Indiana University Press, Bloomington, IN

Gumperz J J 1982 *Discourse Strategies*. Cambridge University Press, Cambridge, UK

Gumperz J J, Hymes D E 1964 The ethnography of communication. *American Anthropologist* **66**(6)

Hall K, Bucholtz M (eds.) 1995 *Gender Articulated: Language and the Socially Constructed Self*. Routledge, New York

Hanks W F 1990 *Referential Practice: Language and Lived Space Among the Maya*. University of Chicago Press, Chicago

Heath S B 1983 *Ways with Words: Language, Life and Work in Communities and Classrooms*. Cambridge University Press, Cambridge, UK

Hill J, Hill K C 1986 *Speaking Mexicano: Dynamics of a Syncretic Language in Central Mexico*. University of Arizona Press, Tucson, AZ

Hymes D 1974 *Foundations in Socioliguistics: An Ethnographic Approach*. University of Pennsylvania Press, Philadelphia

Kroskrity P V (ed.) 2000 *Regimes of Language: Ideologies, Politics and Identities*. School of American Research Press, Santa Fe, NM

Levinson S C 2000 Yéli Dnye and the theory of basic color terms. *Journal of Linguistic Anthropology* **10**(1): 1–53

Lucy J A 1992 *Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis*. Cambridge University Press, Cambridge, UK

Mandelbaum D G (ed.) 1949 *Selected Writings of Edward Sapir in Language, Culture, and Personality*. University of California Press, Berkeley, CA

Ochs E 1996 Linguistic resources for socializing humanity. In: Gumperz J J, Levinson S C (eds.) *Rethinking Linguistic*

Relativity. Cambridge University Press, Cambridge, pp. 407–37
Ochs E, Schieffelin B B 1984 Language acquisition and socialization: Three developmental stories. In: Shweder R A, LeVine R A (eds.) *Culture Theory: Essays on Mind, Self, and Emotion*. Cambridge University Press, Cambridge, UK, pp. 276–320
Schieffelin B B, Woolard K, Kroskrity P (eds.) 1998 *Language Ideologies*. Oxford University Press, New York
Silverstein M 1976 Shifters, linguistic categories, and cultural description. In: Basso K H, Selby H A (eds.) *Meaning in Anthropology*. University of New Mexico Press, Albuquerque, NM, pp. 11–56
Whorf B L 1956 *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* [ed. Carroll J B]. MIT Press, Cambridge, MA
Zentella A C 1997 *Growing Up Bilingual: Puerto Rican Children in New York*. Blackwell, Oxford, UK

A. Duranti

Linguistic Fieldwork

1. History

The history of linguistic fieldwork began in the nineteenth century when linguists started to explore the dialects of European languages. The most famous fieldworker of this time was Edmond Edmont who helped the Swiss dialectologist Jules Gilliéron to collect data for the *Atlas linguistique de France*. Cycling to 639 locations, he interviewed 700 dialect speakers (Chambers and Trudgill 1980, p. 20). Non-European unwritten languages were first systematically investigated on a large scale by the anthropologist Franz Boas (1858–1942), who started to do fieldwork on American Indian languages at the end of the nineteenth century. Boas was also the teacher of a number of influential twentieth century anthropologists and linguists, e.g., Margaret Mead, Edward Sapir, and Leonard Bloomfield.

Until recently linguistic fieldwork on unwritten non-European languages did not attract much attention in mainstream linguistics. It is only since the early 1990s that the linguistic departments of universities and the professional societies have become increasingly aware of the need for researching these languages, about 90 percent of which are endangered (Grenoble and Whaley 1998, Krauss 1992).

2. Learning and Analyzing Unresearched Languages

Fieldwork on languages which have not yet been described differs from other kinds of linguistic fieldwork in that the linguist comes to the speech community with no or only marginal knowledge of the language (called field language). The process of analyzing its structure and use is therefore closely tied to the process of language learning. This means that the choice of research topics, the sequence in which they are studied, and the research methods and techniques depend on the fieldworkers' progress in learning. The better they understand and speak the language, the higher the quality of their data will be, because they will have to rely less on translation. Although the progress of learning a field language depends on the structure of the language and the talent of the researchers and their teachers, experience shows that the methods which have been developed for learning and analyzing a field language work well for very different types of languages and people.

Since languages are never homogenous systems, but show regional, societal, and situational variation, the linguists and the speech communities they work with have to decide which variety should be the object of their first investigation. Only when they have a good command of one variety can they endeavor to study another.

3. Time Plan

Before leaving for fieldwork, the researchers gather as much information as possible about the field language, and the political and sociocultural situation of the speech community they are going to visit. If no information of the structure of the language is available, they should try to access information about related languages in order to get an idea of what to expect. For the work in the field itself, the ideal time plan consists of at least three field trips. The first trip is used for establishing contact and collecting the first samples of data. On the second trip the main body of data is collected, while the last trip will be used for filling gaps and checking what has remained unclear after the analysis and description of the results of the previous trip.

Data collection during the fieldwork is not confined to audio or video recording of spoken language and taking notes during interviews. Rather, all recordings need to be transcribed and translated with the help of native speakers who understand exactly the context of the speech situation and what the people are talking about. In addition, the fieldworker should at least draft a rough analysis of the data, as only analysis and description reveal gaps and unclear items in the data. If for practical and financial reasons it is not possible to visit the fieldwork site several times, the fieldworker has to structure her/his stay into comparable phases of work, so that phases of interviews and recording, translation, and transcription alternate with phases where the researcher steps back from data collection in order to concentrate on analysis and description.

The question of how much time has to be planned for doing fieldwork with the purpose of writing a grammar, compiling a dictionary, or editing a reason-

able number of texts cannot be answered as it depends on the situation of the fieldwork, how the fieldworkers and the speech community cooperate, the availability, talent, and the training of the people they work with, the complexity of the language and culture, etc. In other words, the course of fieldwork is influenced by numerous things that cannot be predicted, but which nevertheless have to be planned. In order to cope with all kinds of unexpected difficulties, the research plan has to be structured in such a way that even in the event of a premature abandonment of the project, the data are sufficiently processed to yield at least some useful output. With regard to dictionary work this means, for instance, that you do not start with the letter A and then work through the alphabet letter by letter, because if the project stops for whatever reasons, a dictionary comprising only a limited number of letters is completely useless.

4. *The Relationship between Linguists and Informants*

The outcome of fieldwork always depends on how well the researchers are accepted by the community, and on which people they can work with. The ideal fieldwork situation is to live in a family or similar kind of group, and be exposed to the field language most of the day. This, however, presupposes that the researchers can adjust to a different standard of food and hygiene, accept their rules of behavior, and can tolerate less privacy than most academics are used to. After having settled, the most difficult task is to find people who can work with the researcher on a regular basis. Especially at the beginning, these people, traditionally called informants, must be able to communicate with the linguist in a second language; but they should also be good speakers of their native language and enjoy studying it together with the linguist. Which people whom the speech community considers to be suitable informants do not exclusively depend on what kind of skills are required for the work, but may also be determined by the linguist's age and sex, by family relations, prestige, or political matters (Grinevald Craig 1987, pp. 9–13)

As in Western societies, work relationships can be of very different kinds. But what makes it more difficult is that fieldworkers do not know beforehand how the speech community will define the roles of the linguist and the informants. None of the recently suggested terms for 'informant,' e.g., 'language helper,' 'native expert,' or 'consultant,' can cover all these types of relationships. Therefore, instead of inventing seemingly politically correct terms, the traditional term 'informant' is given preference in this article.

The relationship between linguists and speech communities should be based on a collaborative approach in which the linguists not only research indigenous languages, but also become involved in the training of native linguists and in language maintenance projects (Grinevald Craig 1997, Wilkins 1992). Apart from their work relations, it is desirable that fieldworkers cultivate friendly relationships within the speech community; in an atmosphere of mutual trust, native speakers will be more ready to advise them on culturally appropriate behavior, and allow them to gain deeper insights into the culture and language.

5. *Types of Method*

The methods used in linguistic fieldwork depend on the research topic, the progress in the linguist's ability to communicate in the then field language, and the kind of work that informants regard as being worthwhile. The best way to ensure the interest and cooperation of informants is to explain to them what one is doing and why. The more the informants understand the linguist's work, the more they can contribute. Thus linguistic fieldwork should be understood as a process of mutual learning in which the linguist learns something about the structure and the use of the field language, and the informants learn something about linguistic methods and the structure of their language.

In particular, linguists employ three kinds of methods: participant observation, interviews, and the recording of narratives, conversations, and other kinds of verbal interaction. Participant observation means that linguists take part in the activities of the speech community so that they can observe how the people communicate and, if they are allowed to play an active role, can also learn to communicate themselves. In the early stages of fieldwork, for instance, games are a good occasion for active participant observation, as the vocabulary used in games is restricted and repetitive, and the grammatical structure of what is said is usually simple.

Participant observation is an excellent means of learning the rules of verbal conduct and the appropriate speech formulas in various situations. It also helps to extend one's vocabulary, as it is easy to memorize new words and their contextually bound meaning when they can be associated with a particular activity. But for the analysis of grammatical structure it is less useful; in order to capture various types of grammatical constructions, one needs to write down hundreds of examples, which is hardly possible while being engaged in activities.

Inexperienced fieldworkers might think that interviews are the easiest and fastest way to obtain all the information you need for grammatical analysis and the compilation of a dictionary. This is an error. Before even thinking of an interview, the linguists need to study 'the local ecology of questioning,' i.e., 'to find out who is allowed to question who, when and how'

(Duranti 1997, p. 104). If it is all right to ask questions, the researcher must be careful only to ask those questions the informants can easily answer; not being able to answer a question about their mother tongue can be so embarrassing that they will try to avoid further interviews, or invent a false answer in order to save face. Questions must be worded in such a way that informants can spontaneously answer without being influenced by what they think the linguist wants to hear.

Two types of interview can be distinguished: the metalinguistic interview and the direct elicitation of language data such as words and sentences. The main purpose of the metalinguistic interview is to help with the transcription and translation of recorded texts, and to provide the necessary background information for understanding their content. Before recording the first texts, however, the linguists should learn the basic vocabulary and grammatical structure of the language. Usually they will do this by elicitation, although there are reportedly cases where the speech community appoints one or two people to teach the linguist in their own way (Wilkins 1992).

All grammatical descriptions must be based on texts which have been recorded, transcribed, and translated, because the function of many grammatical constructions only becomes clear when the wider context of their use is known. In interviews the informants only produce isolated sentences; how these would fit into the natural flow of narratives or other genres of speaking will remain unclear. Nevertheless, elicitation remains necessary in all phases of fieldwork, because even a very large corpus of texts does not contain all wordforms and grammatical constructions which belong to a systematic grammatical description. Similarly, the lexicography of a field language relies on elicitation and the recording of texts.

6. *Fieldwork for Grammatical Descriptions*

If the field language is an unwritten language, the very first thing that needs to be done is to investigate the sound system and develop an orthography. A good method which has been practiced by many linguists starts with a minimum dictionary (Foley 1991, Kibrik 1977, pp. 35–9, Nida 1949, pp. 175–91). This minimum dictionary contains 200–500 words of a maximum length of two syllables which denote different kinds of people, bodyparts, animals, concrete objects, every day activities, states of being, and properties (e.g., words for 'child,' 'dog,' 'house,' 'eat,' 'walk,' 'sleep,' 'sick,' 'big,' 'small'). As these words can be combined to form meaningful utterances, they do not only serve for the investigation of the sound system, but will also, later on, be used to form the first phrases and sentences (e.g., 'the small dog,' 'the child is sick,' 'the child is sleeping,' etc.).

In order to avoid translations from the second language into the field language, the informants are asked to answer questions like 'Could you tell me the names of the most common animals here? What are the basic food items?' and so on. If the word is not too long, the linguist writes it down in a phonetic transcription and asks the informant to translate it or describe its meaning. After at least 200 words have been recorded, the informant's pronunciation of these words are recorded on a tape or disk. First the linguist says the translation equivalent, or the paraphrase, of each word, then the informant repeats the word in question two or three times in the field language. Afterwards the linguist makes a more refined phonetic transcription which is used for a phonological analysis and the design of a practical orthography, both of which will probably have to be revised later on. If there are people in the community who know to read and write in the regional lingua franca, they should help with the development of the orthography to make sure that it is also acceptable and practicable from a technical point of view.

Once the orthography has been designed, the fieldworker can start with the elicitation of short utterances. Again translations should be avoided, as they could result in interferences from the second language and, for example, show an unnatural word order. One strategy to overcome this problem is to give the informant only the idea of the meaning and present it in a way that it cannot be directly translated. For instance, instead of asking to translate the sentence 'Eat the rice!,' one can ask: 'Tell me, what a mother would say to her little son if she wants him to eat up his rice?' In order to elicit utterances of the most natural wording, the linguist should help the informant to imagine a realistic scenario. The linguist writes down the sentences given by the informant and later, preferably not on the same day, reads the sentences to informants and asks them to correct the pronunciation and translate the sentences into the second language. These short sentences should also be recorded; while the recording helps the informant to get used to the tape recorder, the linguist can use the recordings for learning the language.

As for the various components of grammar, it is most difficult to elicit the inflectional paradigms of verbs, since paradigms like 'I ate, you ate, he ate, …' do not make sense to informants who are not trained in linguistics. But later on, after they have become more familiar with language work, at least some of them will understand what such paradigms are. Since it is impossible to identify all morphemes of complex verb forms in the very beginning, the elicitation should start with plain imperatives as they are used by adults when addressing little children. Probably these are, as in many languages, the least complex forms.

After having captured the basic sentence structure, one should start to record, transcribe, and translate short texts, and use them as the basis for further

elicitation by asking the informant to paraphrase certain expressions or to transform sentences. If, for example, the text is about a third person, one can ask the informant to repeat it in the first person: 'How would you say this sentence if it was you who did this?'

7. The Collection of Texts

Which kinds of texts fieldworkers collect depends not only, of course, on the focus of their research, but also on what the speech community regards as both worthwhile recording and suitable for publication. The language a speech community uses for their various communicative interactions is not a homogeneous system, but consists of many styles differing in their phonology, grammar and lexicon. Since it is impossible to handle extensive variation when investigating a language from scratch, the fieldworker should start with one well-defined style and at first, if possible, stick to speakers of the same generation and background. As for transcription, translation, and grammatical analysis, the easiest types of text are narratives of personal experiences. Narratives are usually told at a moderate speed, and contain fewer elliptical sentences and a larger variety of grammatical constructions than casual conversations. Furthermore, the recording of conversations is often impaired by several people talking at the same time, which is very difficult to transcribe. Legends and folk tales, on the other hand, are less suitable than personal narratives early in research, as they may contain archaic words and proverbs which require a deeper knowledge of the language and culture. If, however, members of the speech community ask for a particular text to be recorded and written down, the linguist should never refuse.

In order to guarantee that the informants speak as naturally as possible, and do not simplify their language in reaction to the linguist's low degree of communicative competence, the linguist should ask them to tell the stories in the presence of other native speakers. Ideally, the topic should get the informants so emotionally involved that they forget the tape recorder and the linguist. Since our ethics forbid recording people without their knowledge, emotional involvement can help to overcome the 'observer's paradox,' formulated by Labov (1972, p. 113) 'To obtain the data most important for linguistic theory, we have to observe how people speak when they are not being observed.'

For the representative documentation of a language, the editing of narratives is not sufficient. In order to give a comprehensive picture of how the language is used across a large variety of speech events, it must also contain casual conversations, ceremonial events, political speeches, jokes, poems, and whatever else might count as a distinct genre in this culture (Duranti 1997, pp. 96f). The recording of such speech events is, however, only useful if the linguist and her or his team of native speakers are able to transcribe and translate them. Sometimes it is difficult to get the informants to speak on the recorder because they cannot find a topic which they think would be suitable for recording and publication. In this case it can be helpful to show them a film or a picture story and ask them to describe what they have seen. A short film which was especially designed for linguistic purposes is the so-called Pear-Story (Chafe 1980). Employing the same film for the recording of a number of people can, amongst other things, reveal recurrent patterns of discourse structure which are significant for the respective language and culture. Linguists interested in cognitive anthropology have designed fieldwork kits for highly specialized cross-linguistic investigations such as the conceptualization and expression of spatial relations and the structure of events. Besides questionnaires and video clips, these fieldwork kits also contain materials for language-related games in which native speakers have to fulfill certain tasks while being recorded.

8. Lexicographic Fieldwork

The dictionaries of field languages are usually bilingual dictionaries in which the meaning of the words of the field language are translated into an international language. Some dictionaries are trilingual and have, in addition, translations into the regional lingua franca. Both the fieldworker and the informants must be aware that writing a good comprehensive dictionary takes years, if not decades. If their time is limited, it is more reasonable to plan a glossary to accompany the edition of texts and smaller specialized dictionaries on selected topics such as the terms for food, parts of the house, animals, or speech-act verbs.

The work on a dictionary starts with compiling lists of those words which become the headwords of dictionary entries. There are two complementary methods of compiling word lists: the extraction of words from texts, and elicitation. The extraction of words from texts has the advantage that the words are embedded in sentences which can be used as natural examples in the dictionary. The disadvantage is that even the most common words may be missing. Therefore this method has to be complemented by elicitation. Elicitation begins with the identification of subject areas which should be covered by the dictionary and which are most suitable to start with because the meaning of the words is easy to understand and explain.

The biggest problem of lexicographic fieldwork is that informants are usually not fluent speakers of the second language, whereas the linguist only has a limited knowledge of the field language and culture. Furthermore, many words cannot be translated by single words of a European language because they denote very different concepts. In such cases a translation should be accompanied by a definition. If space

does not matter, the definition can be given both in the indigenous and the second language so that the semantics of the headword as understood by native speakers is clearly shown.

A good example for the usefulness of definitions are animal and plant names. A dictionary which only gives their translations would not show the taxonomy of animal and plant names in the indigenous language and culture, as the semantic relations between generic and specific terms would not be explained. In Samoan, for instance, all fishes, dolphins, whales, and turtles belong to a single category of animals which is called *i'a*. A bilingual dictionary would translate the word *laumei* as 'turtle,' but not explain that it belongs to the class of *i'a*, which is mostly misleadingly translated as 'fish.'

See also: Code Switching: Linguistic; Dialectology; First Language Acquisition: Cross-linguistic; Language Endangerment; Language Obsolescence; Lexicology and Lexicography; Pragmatics: Linguistic; Sociolinguistics; Space: Linguistic Expression

Bibliography

Chafe W L 1980 *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Ablex, Norwood, NJ

Chambers J K, Trudgill P 1980 *Dialectology*. Cambridge University Press, Cambridge, UK

Craig C 1987 Jacaltec: Fieldwork in Guatemala. In: Shopen T (ed.) *Languages and Their Speakers*. University of Pennsylvania Press, Philadelphia, PA

Duranti A 1997 *Linguistic Anthropology*. Cambridge University Press, Cambridge, UK

Foley W A 1991 Fieldmethods. In: Malmkjaer K (ed.) *The Linguistic Encyclopedia*. Routledge, London

Grenoble A, Whaley L J 1998 *Endangered Languages. Current Issues and Future Prospects*. Cambridge University Press, Cambridge, UK

Grinevald Craig C 1997 Language contact and language degeneration. In: Coulmas F (ed.) *The Handbook of Sociolinguistics*. Blackwell, Oxford, UK

Kibrik A E 1977 *The Methodology of Field Investigations in Linguistics*. Mouton, The Hague

Krauss M 1992 The world's languages in crisis. *Language* **68**: 4–10

Labov W 1972 Some principles of linguistics methodology. *Language in Society* **1**: 97–120

Nida E 1949 *Morphology: The Descriptive Analysis of Words*. University of Michigan Press, Ann Arbor, MI

Samarin W 1967 *Fieldlinguistics*. Holt, Rinehart and Winston, New York

Wilkins D 1992 Linguistic research under Aboriginal control. A personal account of fieldwork in Central Australia. *Australian Journal of Linguistics* **12**: 171–200

U. Mosel

Linguistic Presupposition

Presuppositional information is information that is taken for granted by the participants in a discourse or a conversation. As such it has a different status from information the participants present as new or relevant. The latter information is said to be asserted. This distinction between presupposition and assertion is reflected in the linguistic structure of sentences. For example, by his use (1) a speaker typically indicates that he takes for granted that Adam ate the apple, and asserts that Thomas regrets that.

Thomas regrets that Adam ate the apple. (1)

The information that Adam ate the apple is called a presupposition of (1). This presupposition is said to be linguistically induced by the use of the factive verb 'regret.'

This way of introducing the notion of presupposition calls for a distinction. And the basic distinction that should be made is the distinction between presupposition as information taken for granted and presupposition as induced, invoked, or triggered by a linguistic expression. The first notion is the notion of presupposition as background information, that is information which is already given or taken for granted in a conversation. Stalnaker and Karttunen call it the common ground. The second notion is the notion of presupposition as information that is conventionally associated with linguistic expressions or syntactic constructions. It is found in the literature under a variety of names, potential or pre-supposition (Gazdar 1979), conventional implicature (Karttunen and Peters 1979), or elementary presuppositions (Van der Sandt 1988). Though the two notions of presupposition are fundamentally different there is a straightforward connection. If a linguistic element induces a presupposition, the sentence containing the inducing element will typically be appropriate in a context which already contains the presuppositional information. Thus (1) will normally be judged appropriate in a context in which the information that Adam ate the apple is somehow established or taken for granted. If not, one should rather convey this information in an assertoric way as in (2).

Adam ate the apple and Thomas regrets that. (2)

1. Linguistic Encoding of Presuppositions

Linguistic presuppositions can thus be seen as pieces of information triggered or induced by certain linguistic structures. Such triggers can be identified by using specific tests. These tests rely on the fact that presuppositional information tends to escape from environments which block ordinary entailments. The

traditional test is constancy under negation. Other tests are embedding under epistemic modals, embedding in the antecedent of a conditional construction and questioning. Of these the negation test is the least reliable since negative sentences can in many cases be interpreted in a metalinguistic way so as to defeat the presuppositional inference (see *Negation, Linguistics of*).

Applying these tests to the verb *stop* identifies it as a presupposition inducer.

Hubert stopped smoking (at t). (3)
Hubert used to smoke (during some period preceding t). (4a)
Hubert did not smoke (after t). (4b)
Hubert didn't stop smoking. (5a)
It is possible that Hubert stopped smoking. (5b)
If Hubert stopped smoking, he must be saving a lot of money. (5c)
Did Hubert stop smoking? (5d)

The presuppositional content of (3) is given in (4a), the non-presuppositional content in (4b). Only the presuppositional information is preserved as an inference from the embeddings in (5a) through (5d).

The following linguistic structures are widely recognized as presupposition inducers. The presuppositional information that is triggered by the relevant items is given in the b-sentences.

(a) Definite NPs (definite descriptions often including proper names and indexicals; they induce a presupposition to the effect that there is some entity satisfying the description)

The King of France is bald. (6a)
There is a King of France (6b)

(b) Cleft constructions (including 'that-' and pseudo-clefts; they trigger an existential presupposition that there is some entity satisfying the VP)

It was Adam who ate the apple (7a)
Someone ate the apple. (7b)

(c) Focal stress (induces the same existential presupposition as clefts, though many authors want to account for this inference by an independent theory of focus)

ADAM ate the apple (8a)
Someone ate the apple. (8b)

(d) 'Wh'-questions (the presuppositional material is obtained by existential quantification over the variable introduced by the 'wh'-expression)

Who ate the apple? (9a)
Someone ate the apple. (9b)

(e) Quantifiers (the strong quantifiers 'all,' 'most,'… ; weak quantifiers like 'some' are considered to be ambiguous between a presuppositional and a non-presuppositional reading; the presupposition induced is that the domain of quantification is non-empty (see *Quantifiers, in Linguistics*)

All of John's children are bald. (10a)
John has children. (10b)

(f) Factive verbs ('realize,' 'discover,' 'regret,'… ; they induce a presupposition that the presuppositional complement is true)

Thomas regrets that Adam ate the apple. (11a)
Adam ate the apple. (11b)

(g) Temporal clauses ('before,' 'while,' 'since,' 'after,'… ; as in (f) they induce information of a propositional nature.

After Adam ate the apple, he was expelled from Paradise. (12a)
Adam ate the apple. (12b)

(h) Sortal restrictions on predicates ('bachelor,' 'close,'… ; the presuppositional information consists of the requirement that the arguments of the predicate are of the appropriate sort)

John is a bachelor. (13a)
John is male, adult,… . (13b)

(i) Implicative verbs ('manage,' 'succeed,'… ; put a presuppositional requirement on the action denoted by the VP)

John managed to open the door. (14a)
It took some effort to open the door. (14b)

(j) Verbs of transition ('start,' 'stop,' 'continue,'… ; see (3) above)

(a) through (d) are examples of existential presuppositions. They all involve an existential quantification over some individual- or set-variable. In (f) and (g) the presuppositional information is of a propositional nature. (h) through (j) involve requirements on the applicability of a predicate. This is only a small sample of the linguistic structures that have been identified as presupposition inducing. Further examples include focus particles ('only' and 'even,' adverbs of repetition ('again,' 'too,' 'back'), but also, more controversially, counterfactual conditionals, non-restrictive relative clauses, and discourse particles like 'because' and 'although.'

2. *Projection*

The presuppositions induced by simple sentences tend to survive under embedding, that is whenever a subsentence of a compound φ induces the presuppositional information χ this information will surface as an inference of φ. Karttunen (1973) was the first to note and catalogue numerous exceptions to this generalization. Here are some examples. In the a-

sentences the presupposition that John has children escapes from its embedded position, in the b-sentences the same presupposition does not surface as an intuitive inference.

If baldness is hereditary, John's children are bald. (15a)
If John has children, John's children are bald. (15b)
Either baldness is not hereditary, or John's children are bald. (16a)
Either John has no children or John's children are bald. (16b)
It is possible that baldness is hereditary and that John's children are bald. (17a)
It is possible that John has children and that John's children are bald. (17b)

In the terminology prevalent in the 1970s and 1980s it is said that, in the a-sentences, the presupposition that John has children is projected to the main sentence, whereas it is said to be filtered or cancelled in the corresponding b-sentences.

3. *Semantic, Pragmatic, and Discourse Presupposition*

We may distinguish between three main types of theories that deal with presupposition and projection: semantic, pragmatic, and discourse theories.

3.1 *Semantic Theories*

The semantic view originates in Frege and was revived by Strawson. It construes presupposition as a binary relation between sentences (or propositions). A presupposition is a requirement that has to be fulfilled for a sentence to have a truth-value. Thus ψ is taken to be a presupposition of φ just in case ψ has to be true in order that φ has a truth value. This requires a semantics that allows truth-value gaps. Assuming the standard definition of negation as an operator which maps truth onto falsity and vice versa, semantic presupposition can then be defined as follows:

φ semantically presupposes χ if φ entails χ and not-φ entails χ.

The defining characteristic is constancy under negation. The projection behavior of presuppositions in other embeddings follows from the definitions of the logical connectives and other operators. One problem for such theories is that they cannot handle the presuppositions triggered by adverbs like 'too' and 'even.' Here presupposition failure arguably does not affect the truth-conditions of the sentence, but the appropriateness of the utterance. It also turns out that no definition of the logical connectives can be given which accounts for all the projection data. Defenders of the semantic account are thus forced to either postulate an ambiguity in the logical connectives or transfer the burden of explanation to pragmatics.

3.2 *Pragmatic Accounts*

Pragmatic accounts (see *Pragmatics: Linguistic*) view presupposition not as a semantic relation between sentences. Instead presupposition is viewed as a property of utterances. The basic notion is that of a speaker presupposing a proposition in a context of utterance. This allows for certain variability of presuppositions with the context of utterance. The first formally developed theory of this kind is found in Gazdar (1979). The basic idea here is that a presupposition of a sentence is preserved unless it conflicts with contextual information or with an implicature invoked by its utterance. In the case of (15) through (17) this amounts to the following. A typical utterance of one of the b-sentences will invoke the implicature that it is possible that John has no children. This implicature conflicts with the presuppositional assumption that he does and so the latter is cancelled. Since no such implicature is invoked for the a-sentences, they will be interpreted as presupposition preserving.

3.3 *Dynamic Theories*

A different style of modeling is found in the dynamic approaches. The two main variants are the so-called satisfaction theory and the anaphoric or binding theory. The first derives from Stalnaker (1974) and Karttunen (1974) and has been developed in particular by Heim (1983) and Beaver. The anaphoric account is found in van der Sandt (1992) and Geurts (1999). Comparisons and different assessments of their relative merits are given in Beaver (1997) and Geurts (1999).

Both accounts start off from the idea that an utterance changes the context in which it is made. The basic requirement of the satisfaction account is that the presuppositional information should be entailed by the context of utterance. If so, the input context is incremented with the information contained in the inducing sentence. An utterance of (1) thus requires a context which entails that Adam ate the apple. If this requirement is not met, the input context is not a suitable one. In this case a mechanism of accommodation is invoked. This means that the input context will (under appropriate conditions) be revised so as to incorporate the relevant presuppositional information thus restoring interpretability.

The anaphoric or binding account is formulated as an extension of discourse representation theory. The basic tenet is that presuppositions are anaphoric expressions (see *Anaphora*). Presuppositions only differ from pronouns and other attenuated anaphors in that they have internal structure and semantic content. The basic requirement is not that the presuppositional information is entailed by the context of utterance, but that the presuppositional anaphor should find a suitable antecedent. If no such antecedent is available, the semantic material carried by the presuppositional anaphor will be accommodated in the discourse structure so as to provide an antecedent after all.

The context-change semantics of the satisfaction theory generally yields weaker predictions than are found in the anaphoric account. This is due to the different status that presuppositional material has in both theories and to the different ways in which this material is processed. For example, for a conditional of the form $\varphi \rightarrow \chi_{\langle \psi \rangle}$ where ψ is triggered by χ the satisfaction theory will predict $\varphi \rightarrow \psi$ as presupposition for the whole sentence. This yields for (15a) not the presupposition that John has children, but the presupposition that he has children *provided* that baldness is hereditary. Just like the classic semantic account, the satisfaction theory has thus to appeal to some pragmatic strengthening mechanism to derive the stronger presupposition that John has children. On the anaphoric account, presuppositional information is never transformed into a new semantic construct. Consider (15a) again. The presuppositional anaphor associated with the consequent cannot be bound to some preestablished antecedent. So some entity representing the set of John's children will be accommodated in the main context, thus yielding the stronger prediction that the utterance of (15a) presupposes that John has children.

See also: Comprehension, Cognitive Psychology of; Constructivism in Cognitive Psychology; Information Processing Architectures: Fundamental Issues; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Linguistics: Overview; Pragmatics: Linguistic; Sentence Comprehension, Psychology of; Semantics; Text Comprehension: Models in Psychology

Bibliography

Beaver D I 1997 Presupposition. In: Benthem J A van, Meulen A ter (eds.) *Handbook of Logic and Language*. Elsevier, Amsterdam, pp. 939–1008

Gazdar G 1979 *Pragmatics: Implicature, Presupposition, and Logical Form*. Academic Press, New York

Geurts B 1999 *Presuppositions and Pronouns*. Elsevier, Oxford, UK

Heim I 1983 On the projection problem for presuppositions. In: Barlow M, Flickinger D P, Wescoat M T (eds.) *Proceedings of the West Coast Conference on Formal Linguistics Volume 2*. Stanford Linguistics Association, Stanford, CA, pp. 114–25 [Reprinted in Davis S (ed.) *Pragmatics*. Oxford University Press, Oxford, UK, pp. 397–405]

Karttunen L 1973 Presuppositions of compound sentences. *Linguistic Inquiry* **4**: 167–93

Karttunen L 1974 Presupposition and linguistic context. *Theoretical Linguistics* **1**: 181–94

Karttunen L, Peters S 1979 Conventional implicature. In: Oh C K, Dineen D (eds.) *Presupposition*. Academic Press, New York pp. 363–85

Sandt R V van der 1988 *Context and Presupposition*. Croom Helm/Routledge, London

Sandt R A van der 1992 Presupposition projection as anaphora resolution. *Journal of Semantics* **9**: 333–77

Stalnaker R 1974 Pragmatic presuppositions. In: Munitz M K, Unger P K (eds.) *Semantics and Philosophy*. New York University Press, New York, pp. 197–214

R. A. van der Sandt

Linguistic Turn

'The linguistic turn' generally designates in the social sciences a movement away from positivist or historicist accounts of social and historical phenomena toward analyses of the structures of representation that are understood to have helped shape those phenomena. The phrase is frequently (and sometimes loosely) used to characterize contemporary forms of the critique of ideology and the critical 'demystification' or 'demythologization' of historical and socio-political representations that offer themselves as transparent or natural accounts. Since such critical work often proceeds by way of attention to the rhetorical and conceptual articulations of representations that tend to efface their very character as representations, and since it frequently draws from philosophical and even literary theories of language and representation, it is understood to participate in a turn from socio-historical 'realities' to their linguistic presentation.

The phrase is used so broadly, however, that its strong meaning for theoretical methodology is frequently lost. Frequently, the critique of representational structures proceeds without a consequent attention to the question of language, and even some of the strongest versions of 'linguistic constructivism' fail to interrogate their own presuppositions. Thus, the linguistic turn will be defined here fairly strictly from the basis of its foundations in order to recover its meaning and implications for the social sciences. It will be defined as beginning when language is no longer understood as an instrument in social practice, but rather as the site or 'ground' for such practice. It

will be understood to have been fully undertaken when such a fundamental understanding of language shapes the procedures of the sciences themselves.

The beginnings of the linguistic turn may easily be traced to the eighteenth century and seen to be lurking in the very rich debate concerning the origins of language. It is clearly occurring in the research of Wilhelm von Humboldt and in the writings of the German Romantics, and then actively recognized—some would say, practiced—in the later works of Friedrich Nietzsche. Thus, when the official occurrence of a linguistic turn is registered in Vienna for philosophy of language, a more 'fundamental' turning is already long underway in the tradition of German philosophy that Walter Benjamin and Martin Heidegger brought to the fore in their remarkably convergent speculative writings on language. Indeed, if Ludwig Wittgenstein himself was ready to grasp in 1929 the import of Heidegger's still largely latent meditation on language and the significance of Heidegger's own linguistic practice, it is because the philosophical ground was well prepared for Heidegger's subsequent turn (his famous *Kehre*) from an existential philosophy still marked by the metaphysics of the human subject to a thought of what is given of Being in and by language.

Historians of the social sciences will provide ample testimony of methodological tremors linked to the developments noted here in philosophy of language—in short, a displacement of the human subject or human consciousness as the grounding source of meaning in signifying practice, and the requirement of rethinking the subject/object relation and relation itself. However, it may well be argued that the strong turn realized by authors such as Benjamin and Heidegger is not assumed in the social sciences in Europe until the period of the structuralist movement (a movement that explicitly unseated human consciousness from its founding role in favor of broader structures of determination: kinship systems, linguistic structures, and the laws of sexuality or desire). Even here, however, the turn to language remained frequently at no more than the level of analogy, and systems of myth or kinship were shown to be formalized equally effectively with recourse to mathematical logic (another form of appeal to a *logos*, some would argue). Thus, while structural linguistics functioned as a kind of 'point science' for the initiatives throughout the human sciences in Europe for over a decade, only the endeavors of 'post-structuralist' authors such as Michel Foucault, Jacques Lacan, Jacques Derrida, or Gilles Deleuze actually carried the notion of the linguistic constitution of being to its full import for the social sciences. Only from the time of post-structuralism did it become possible to entertain thinking the social relation from language.

The linguistic turn has thus occurred, for example, when Michel Foucault, in his methodological treatise *The Archaeology of Knowledge* (1969), announces that the fundamental element for his study of discursive formations will be the *énoncé*, the statement. In an effort to displace the governing categories of the history of thought (totalizing categories that establish continuity and that all lead back, in Foucault's view, to the sovereignty of the human consciousness), Foucault attempts a theoretical elaboration of a discursive 'object' that is always relational and always to be thought from its material emergence. 'Object' must be left here in quotation marks because the statement, as Foucault understands it, is the 'modality' or 'function' of existence of discursive facts, what 'gives' them or enables them to occur as language and thus as (a) a relation to a domain of objects; (b) a set of possible positions for an enunciative subject; (c) an element in a coexisting discursive field; and (d) a repeatable materiality. A statement, in Foucault's definition, cannot be understood as a sentence that lends itself to hermeneutic investigation or logical analysis, but must rather be seen as the material, always relational, condition for something like a sentence or a proposition. Historical analyses that seek the 'signified' of any group of signs (reading any document for the meaning it bears, any text for its 'allegorical' sense), like linguistic studies that seek rules of formation, invariably efface the irruptive, eventful character of the fact of language, returning all linguistic events to their organizing conditions. Thus they participate in what Foucault calls, in his famous *Discourse on Language*, the 'ancient elision of the reality of discourse.' Foucault's effort in thinking the *énoncé* as the material function of existence of any discursive act or organization is directed at honoring the 'anarchic' character of language—a potential for disruption which 'disciplinary' and institutional formations seek to channel and control. At the heart of his project is an assumption that the orders of knowledge and the social institutions in which they are embedded are organized by a fear of the undermining, or even violent dimensions of discourse for the orders in which it emerges (undermining for the mere fact that it can reveal the inherent lack of foundations for the social order, violent for the fact that it can carry the irruption of desire):

> 'There is undoubtedly in our society, and I would not be surprised to see it in others, though taking different forms and modes, a profound logophobia, a sort of dumb fear of these events, of this mass of spoken things, of everything that could possibly be violent, discontinuous, querulous, disordered even and perilous in it, of the incessant, disorderly buzzing of discourse.'

Of course, Foucault's own study, fiercely abstract in its unfolding design, is not without its own violence. Indeed, his *Archaeology* sometimes reads as a veritable storm of theoretical elaboration; its winds tear unrelentingly at the foundations of the history of

thought. But a glimpse at the virulence of reaction on the part of proponents of rational norms of 'communicative transparency' and/or pragmatic 'consensus' provide a clue to the explosive potential of Foucault's effort to honor the disruptive dimensions of language.

Numerous authors have pursued Foucault's 'archeological' analyses of discursive formations; in fact, significant currents in the growing field of cultural and 'post-colonial' studies principally draw upon Foucault for their understanding of the discursive constitution of social reality. Theories of social construction are deeply indebted to his analyses of 'disciplinary' knowledge and mechanisms of control. Far fewer authors have carried forward his understanding of the subversive dimensions of the enunciative act, though one sees its impact in the fields of gender studies and sexuality or in a topic such as the study of 'the practices of the everyday' by Michel de Certeau. It is for this reason that one might regard the 'linguistic turn'—which many would undoubtedly anchor in the widespread acceptance of the theory of linguistic or social 'constructionism'—as incomplete and even faltering. Theories of social construction, and even theories that stress the necessary 'performative iteration' of social norms, almost inevitably turn short of the engagement with language that marks Foucault's work (or that of the other 'post-structuralist' authors to which I referred). They fail, in particular, to attend to the materiality of discursive usage in that dimension that involves what I might term an enunciative experience with language. This avoidance is easily attributable to Foucault's own wariness regarding the 'ontologizing' and 'transcendentalizing' sources upon which he silently draws (principally Heidegger), not to speak of his strictures against anthropomorphism and humanism. But Foucault's own turn to ethical questions in his late work point firmly to the necessity of an exploration of the subject's engagement with the fact of language at the fundamental level evoked in the *Archaeology*—with the fact and manner of the way the *énoncé* is given in its material singularity. What dimension of experience is engaged with this 'gift,' and what engagements (to evoke a motif pursued throughout Jacques Derrida's late work) does it presuppose if the relations offered by the *énoncé* must be thought from the way they are brought into existence in singular instances of usage?

Heidegger had attempted in his later meditations on language to name with precisely the term 'usage' the human relation to the fact of language. To avoid turning language into a kind of transcendental horizon (a trap into which many 'linguistic constructionists' inevitably fall), and to hold to a thought of its finitude, Heidegger attempted to think through the necessity of a human 'accession' to language—i.e., the conditions for any 'interpellation' of the human subject by language, and any human response. He did not avoid a certain anthropocentrism in this undertaking, but he also pointed beyond the purview of the metaphysics of the human subject. Unfortunately, few authors have followed through Heidegger's thinking on this crucial point, though among them we may count the French psychoanalyst Jacques Lacan (whose thought of the category of the 'real' presupposes this understanding of the 'human relation to the signifier') and the French writer Maurice Blanchot, whose enormous influence on 'post-structuralist' thought still remains to be measured. But their work, taken together with that of the distinguished Jewish philosopher Emmanuel Levinas and the more recent work of authors such as Jean-François Lyotard and Jean-Luc Nancy, make it possible to pursue a thought of the manner in which the social relation opens in and with language. From a notion of the human engagement with language, they elaborate an understanding of the event by which human speech communicates a form of 'finite transcendence': a transcendence of determined orders of meaning that is first of all an exposure to the other human being. They thus make it possible, in effect, to return to a thought of the social relation like that approached by Hannah Arendt in *The Human Condition* (1998) in her effort to rethink the notion of public space and the political act itself.

This latter thought of 'communication' (not a transparent communication of signified meaning, but rather, as Walter Benjamin argued so powerfully, a communication of the very possibility of relation) clearly carries with it the obligation of defining the conditions for its engagement in critical or theoretical practice. For it is clear that the pertinence of a traditionally 'theoretical' mode of inquiry must come into question once language is engaged at levels that exceed that of its representational or instrumental function. Once a 'saying' that bespeaks the irreducible presence of an enunciating subject becomes the object of a meditation on the grounds and possibility of community (but this can be no 'object' for description), theory can no longer claim an exclusive purchase on the question of the social relation. Or it must rethink its own discursive modality and its relation to others (to the language of literature, for example). The challenge of the linguistic turn for the social sciences thus bears upon the discursive practices of the sciences themselves. It is not a challenge that requires the social sciences to abjure their commitment to clarity and rigor in analysis, but it does require them to entertain the question of their own discursive usage and the possibility of translation between discursive modes or between their own usage and that of a singular event of language. Here, it becomes clear that a fundamental reflection on linguistic usage is the very condition for a consequent engagement of the social relation—a relation that is not reducible to its linguistic determinations, but also not thinkable apart from the site of its emergence and articulation. The 'linguistic turn' is not a methodological 'given,' a path already secured; it is a required turn to the grounds of meaning that must

be repeatedly undertaken if social analysis is to achieve concreteness in its effort to regain the social itself.

See also: Constructivism/Constructionism: Methodology; Foucault, Michel (1926–84); Linguistic Turn and Discourse Analysis in History; Logical Positivism and Logical Empiricism; Representation: History of the Problem; Romanticism: Impact on Social Thought

Bibliography

Arendt H 1998 The Human Condition, University of Chicago Press, Chicago

Butler J 1993 *Bodies That Matter*. Routledge, New York

Benjamin W 1972 *Gessamelte Schriften*. Tiedemann R, Schweppenhäuser H (eds.). Suhrkamp, Frankfurt am Main, Germany [Demetz P (ed.) 1986 *Reflections*. Schocken Books, New York]

De Certeau M 1984 *The Practice of Everyday Life*, trans. Rendall S F. University of California Press, Berkeley, CA

Foucault M 1969 *L'Archéologie du savoir*. Editions Gallimard, Paris. (1972 *The Archaeology of Knowledge and The Discourse on Language*. Pantheon Books, New York, 1972)

Fynsk C 1996 *Language and Relation … That There is Language*. Stanford University Press, Stanford, CA

Heidegger M 1985 *Unterwegs zur Sprache, Gesamtausgabe*, von Hermann F (ed.) Klostermann, Frankfurt am Main, Germany, Vol. 12. [trans. Hertz P D 1971 *On the Way to Language*. Harper and Row, New York]

Lacan J 1986 *Le Séminaire, L'Éthique dans la psychanalyse*. Seuil, Paris

Lyotard J-F 1983 *Le Diffrend*. Editions de Minuit, Paris

Rorty R (ed.) 1992 *The Linguistic Turn: Essays in Philosophical Method*. University of Chicago Press, Chicago

Wittgenstein L 1965 Lecture on ethics. *Philosophical Review* **74**: 8–15

C. Fynsk

Linguistic Turn and Discourse Analysis in History

1. The Linguistic Turn and the Post-modern Epistemological Crisis

The term 'linguistic turn' emerged as common currency in discussions among historians during the 1980s and 1990s to designate a sweeping, but vehemently contested, shift in the framework within which both the object or content of historical investigation (the 'past') and the methods of achieving historical knowledge of this content were defined. In its weaker, least contentious, form the linguistic turn referred to a more intense and focused attentiveness to the opacity and complexity of the linguistic forms through which all evidence of the past was mediated, and the 'truth' of the past represented, reconstructed, and communicated. The term itself was appropriated from discussions among philosophers, where it had been used as a general description of the turn in twentieth-century philosophy toward reformulating the epistemological problem of the relationship between thought and reality in terms of a relationship between language and reality. In the stronger and more contentious form that surfaced in debates among historians, however, the linguistic turn came to designate a shift away from the conventional view that language was a passive, potentially transparent instrument for transmitting content, for organizing, representing or expressing the truth of a reality outside of itself, and toward a view of language as the dense, fertile and autonomous site in and through which the objective reality of the world as well as the subjective reality of language users was actively produced or constructed. From this latter perspective, the act of taking the linguistic turn was synonymous with the claim that the production and communication of historical knowledge was inextricably entangled in the epistemological crisis associated with the transition from a modern to a 'postmodern' cultural condition.

As a particular disciplinary expression of this epochal transition to postmodernism the linguistic turn in history was grounded on three major premises. First, it accepted the ultimate futility of the historical quest to reconstruct a single, coherent narrative of human history whose pattern or structure could somehow be grasped and mapped in its totality from a transcendent position outside of the particular, contingent meanings that human beings in specific times and places had imposed on the material traces of the past, a quest most recently articulated in the national liberation narratives, global modernization theories, and Marxian historical dialectics of the Cold War decades. Second, the linguistic turn appropriated from poststructuralist theory the sweeping critique of residual beliefs in the existence of noncontingent stable identities (whether natural kinds or metaphysical essences) whose fates might be traced historically through paths of accident, evolution, repression, resistance, and liberation. National, ethnic, racial, sexual, gender, and class identities were not treated as material or metaphysical realities expressed in language and culture, but contingent, constantly changing, constructions produced in language and culture. Third, this dissolution of essential identities implied a transformation in the methods of historical representation. Access to a reality not already shaped and constructed by the linguistic medium in which it was always presented to the knower, either as subjectively lived experience or objectively perceived external object or event, was considered impossible. The historian could only interpret interpretations, could only redescribe and thus reconstruct in his or her own language the linguistic constructions of others.

Various elements of the epistemological crisis associated with the linguistic turn had been present in muted and sometimes more extreme forms within the Western historical tradition since the beginnings of the modern professionalization of history in the early nineteenth century. During the post-World War II era the disconcerting relativist implications of self-consciousness about the density and relative autonomy of the linguistic medium that mediated reality as meaningful was reined in by a broadly shared disciplinary consensus concerning the reality of the past and the methods that provided access to its truth. The more radical linguistic turn of the 1980s, with its implications that the historian's past was a construction grounded on ethical and political choices, rather than on epistemological criteria, was associated with the breakdown of this consensus, as the increasingly effective voices of previously excluded perspectives, particularly those of women, ethnic minorities, and former colonial subjects, mounted a powerful critique of established professional conventions. The linguistic turn was embedded deeply in the feminist, multiculturalist, and postcolonial 'turns' among dissident members of the disciplinary community, and connected to a new series of interdisciplinary liaisons with literary theory, cultural anthropology, and philosophy.

A major difficulty standing in the way of providing a concise description of the linguistic turn in history stems from the diverse meanings assigned to the term 'language.' The everyday conventional meaning of language as a native or vernacular language is rarely what contemporary historians have in mind. In fact, the use of the term has been extended broadly beyond the rules and practices of verbal communication in speech and writing to include nonlinguistic structures of exchange and all signifying practices through which meaning is produced, transmitted, and communicated. As the totality of meaningful action or signifying practices, 'language' becomes virtually indistinguishable from 'culture.' The favored model for organizing discussions of language by proponents of the linguistic turn during the 1980s and 1990s was the semiotic model derived from the linguistic theories of Ferdinand de Saussure and Charles Sanders Pierce, in which language is conceived as a structured system of 'signs' whose meanings are determined by their relations to each other rather than their relation to a 'transcendental' object or subject. Most influential for historians has been the development of this semiotic model in cultural anthropology and especially the French structuralist and poststructuralist theory of Louis Althusser, Roland Barthes, Jacques Derrida, Jacques Lacan, and Michel Foucault. Although 'language' and 'discourse' are still often used interchangeably in debates concerning the linguistic turn, the attempt to provide some kind of theoretical focus on the analysis of language in history has increasingly been organized around the term 'discourse.' Historians who have taken the linguistic turn tend to define the distinctiveness of their own signifying practice as 'discourse analysis.'

2. *Discourse Analysis: Enabling Structures and Communicative Practices*

Despite the widespread use of 'discourse' as an organizing concept within a variety of the humanistic and social scientific disciplines, its meanings have remained fluid, flexible, and often indefinite. However, the two general meanings of the term in common usage—(a) a formal speech or treatise that presents a theme or object in a systematic and authoritative manner, and (b) a conversation or discussion involving a communicative exchange among a number of speakers—provide a framework for the more technical, disciplinary uses within current historical work.

2.1 *Discourse as the Historical* a priori *of Thought, Speech, and Action*

The more formal, systematic, and authoritative dimension of discourse is articulated in historical studies in the attempt to outline the formal structures of rules, codes, categories, and procedures (the relatively stable aggregate of conventions, usages, idioms, tropes, vocabularies) that define a particular communicative space, delimiting the field of possible statements and actions, enabling what can be meaningfully and effectively articulated. Historical discourses in this sense are impersonal formal structures operating anonymously behind the backs of speakers and actors, constituting the world of objective reality and subjective agency by making certain types of speech or behavior possible and other types impossible, by recognizing certain speakers and actors as legitimate or effective, and excluding others as illegitimate or ineffective, by defining certain objects and events as real or objective, and others as fictional and subjective. Different discourses constitute the world of experience in different ways, and are thus incommensurable with each other. They produce homogeneity (within their field of power) and heterogeneity in relation to other discursive fields. Specific statements or actions are only meaningful as instances of the rules and regularities governing the discourse in which they occur. Discourses become markers of the stable, synchronic, self-contained, incommensurable, discontinuous dimension in historical systems of meaning or cultural formations. For the historian, mapping the structures or codes of particular discourses defines historical action and historical actors as determined by those structures, as events and agents in a foreign world that has its own criteria for determining what is true or real, who is an effective agent, whose speech is meaningful. Discourses may thus, in Foucault's terms, create

specific historical 'regimes' of truth and systems of power.

Discourse analysis can be a way for the historian to establish the otherness of the past, as it has been by practitioners of the new cultural history who imagine the past as a 'foreign country.' Textual traces from the past can be provided with their own reality, differentiated from the realities of other times and places, by inserting them into the discourses that originally defined their significance. By recognizing the strangeness, the alien character of the 'voices' or 'traces' from the past deposited in the archival record, the present also comes into focus as a particular historical moment with its own sets of systematic enabling and limiting conditions. In this sense the deployment of an ahistorical, formal conception of discourse can produce a very concrete historical consciousness, an awareness of the contingent historicity of present existence. It provides few guidelines for reconstructing the developmental processes that lead from the past to the present.

2.2 Discourse as Communicative Process

As a constitutive system of categories and codes that differentiates historical worlds or cultures from each other, discourse becomes a tool for the reconstruction of past as different, for marking historical distance and avoiding anachronism. But it also freezes systems of meaning into static cross-sectional snapshots that allow analysis of the self-relating structures of individual systems, but do not illuminate the interaction among systems or the processes of transformation from one system to another. Moreover, the employment of the structural definition of discourse as a closed semiotic system seems to dissolve historical agency. The identity of agents is called forth or constituted by the discourse itself and their statements and actions remain mere instances of the discursive rules that produce the reality in which they operate. A second dimension of discourse analysis, however, breaks apart the apparent stasis and structural determinism of this stance. Conceived on a conversational or dialogical model in which signifying practices operate in continuous processes of confrontation, exchange, and transformation, discourse becomes a term for conceptualizing historical change, communicative interaction, and the dynamics of historical agency. From this point of view, linguistic or signifying actions are not authorless events in language but actions on language by individual and collective agents who appropriate inherited possibilities contained in existing discursive structures in order to both modify those structures and to transform themselves as agents. But how could the agents, the individual, and collective identities contingently and historically constructed within the structures of discourse, become themselves the creative transformers of discourse? One possible answer to this question could be constructed from within discourse theory itself, but ultimately it does raise the issue: what is the 'other' of discourse? and how is discourse changed and transformed in its relationship to its other?

3. Discourse and Discourses: Pluralism and Heterogeneity

In an early manifesto of discourse analysis, Foucault (1972) distinguished between a general domain that included all meaningful or signifying actions and their formal or structural conditions, and more individualized, singular groups of signifying practices organized in a focused disciplinary fashion around the construction of a coherent regime of truth. Historians have not been completely disinterested in the more sweeping, universalizing or totalizing notions of discourse that define the historical *a priori* of whole epochs or cultures, like the discourse of 'man' or secular historical humanism that rules over the 'modern' epoch, but they have tended to gravitate toward a practice of discourse analysis that emphasizes both the plurality of discourses and the conflict among discourses within a general discursive domain. Three forms of such heterogeneity in discourse analysis are particularly obvious. First, discourses can be differentiated in terms of the social and cultural dimensions or spheres in which they operate or are effective. Thus, one might differentiate among public, official, popular, literary, elite, urban, or religious variants of discourses of class or madness or justice or gender at any particular time or in any particular geographic area. Second, discourses obviously are differentiated in terms of the specific domains of meaning and action—the objects, subjects, and rules of relationship—which they construct. Thus, orientalist discourse constructs 'the east' as well as the 'orientalist'; psychiatric discourse constructs the psyche and the psychiatrist. Differentiations of this type are often imagined as analogous to academic disciplines that produced their putative objects of analysis in the way that the discipline of history constructs 'the past,' or sociology constructs 'society' or 'the social.' Third, historians tend to differentiate competing or alternative discourses that operate within these various dimensions and domains. Within the dimension of public discourse, for example, there may not only exist separate discourses pertaining to the spheres of political authority, citizenship, social hierarchy, or legal structures, but also dominant and oppositional, liberal and conservative, patriarchal and feminist, bourgeois and working class discourses that delineate the boundaries of this sphere and organize and define its contents and relations in competing and conflicting ways.

This proliferation of difference in the discursive realm is one of the major conditions of creativity and historical change. At any particular moment in time

an historical actor will inhabit a heterogeneous compound of discourses, dissolving the apparent formal disassociation of discourses from each other, allowing for creative performances in which the inner logics of different discourses are brought in contact with each other in transformative ways. Just as a scholar can inhabit an interdisciplinary space in which various disciplinary discourses can be brought into interaction in order to transform conventional, dominant discourses about a particular issue or problem, an actor in the public sphere might exploit the plurality of existing constructions of citizenship in various inherited discourses to imagine and act out innovative forms of membership and responsibility in a political community. The first task for the historian engaged in discourse analysis is to identify a discourse within a given body of evidence, to map out its boundaries, and reconstruct its inner logic or structure. A discourse is not simply a synoptic summary of the aggregate of statements that refer to some specific theme, like references to the work of women in France between 1815 and 1848. To identify the discourse in such thematically associated material would be to reconstruct the formal dimensions within the signifying practices informing the evidence. Discourse would be the organizing power that constructs the 'world' the evidence assumes and provides effective power to certain statements. Once such a discourse has been mapped or reconstructed, the map can be used, like an 'ideal type' or 'paradigm' to identify or discover its presence in other texts or bodies of evidence. Such reconstructive mapping of historical *a prioris* establishes historical difference, marks off the conditions of possible action and expression in different worlds. The second task of the historian of discourse, however, would be to cut through or modify the radical difference produced in the first operation—to describe the ways in which the specific actions—linguistic and nonlinguistic performances with communicative intent—within the formal conditions of a discourse create changes in that discourse and reveal the porous, contingent nature of discursive boundaries. This model of discourse analysis as a constant oscillation between the reconstructions of formal contexts and concrete, innovative performances in a dialogical, communicative context has been highly developed within specific realms of intellectual history—like the history of political theory, public opinion, and ideology, but its expansion into a broader analysis of the whole range of historical practices has highlighted a number of problems, problems that have become the center of the controversies dividing historians about the validity of discourse analysis *per se*.

4. *Discourse and its 'Others'*

Discourse and discourse analysis only attain distinctive, specifiable definition if they are placed in relation to something which they are not. What are the 'others' of discourse? This question is directly relevant to two questions that are of particular interest to historians: How can one explain the stability and/or hegemony of certain discursive formations among specific populations over extended periods of time, and how can one explain changes within specific discourses, the shifting relationships among discourses, and the emergence of new discourses in time?

4.1 *Discourse and the Material World*

The polemical thrust of discourse analysis when it emerged as an articulated position in debates among historians during the 1980s was directed primarily against dominant conventions of historical analysis which portrayed the realm of the production, reproduction and transmission of meaning in linguistic and other signifying actions as a product, effect or expression of the structures and transformations of material practices. Despite the complexities within and obvious differences among class oriented British cultural Marxism, the sociologically grounded modernization narratives prevalent in the US and Germany, or the structuralist 'total histories' of the French Annales school, all were viewed as ultimately reducing the production and transformation of meaning to an effect, however indirect, of changes occurring outside of the linguistic and more broadly cultural sphere. Concepts like 'ideology' or 'collective mentality' were criticized and often rejected outright because of their implication that the production of meaning in language could only be grasped with reference to a reality or truth that was situated in material social practices outside of language. In its initial stages the linguistic turn often appeared as an attempt to resist the absorption of the cultural into the social, to insist on the autonomy of processes of meaning production, to assert the autonomy (no longer 'relative') of intellectual and cultural history. More serious tensions among historians developed, however, when the realm of discourse was conceived as creating the categories and identities through which people lived their social being and defined their social interests. Class identities were not defined as public expressions of the ways in which groups articulated social interests grounded in the way they lived their relation to nature in systems of production, but identities produced in discourse that determined how interests were defined and the ways in which economic and social relations were experienced. Gender differentiation in public discourse, or medical discourse, or class discourse was not conceived as representing the contrasting realities of social experience of physically differentiated bodies but as constructed forms in which bodies were experienced as sexually different. Histories of the experiences of classes, of the sexes, of races, ethnic groups, nationalities, colonial subjects, etc., were transformed into historical reconstructions of the discursive processes through which class, sexual,

racial, ethnic, national, and colonial identities were constructed.

It would be misleading however, to treat this new cultural history, grounded in discourse analysis, as some historians have tried to do, as a new form of idealism based on a belief that language makes the material world, that thought creates existence. First, in discourse analysis, discourse itself is conceived as a dense network of signifying practices, a material activity embedded in the communicative interactive relations between producers and consumers of meaning and articulated in institutional structures. The linguistic turn did not simply exalt the status of language but changed the conception of what language is and how it works in history. Second, the 'other' of discourse analysis is not the empirical analysis of concrete evidence: discourse analysts certainly are as attentive to archival evidence as any social historian. The other of discourse analysis is a belief in the ultimate accessibility of an 'unmediated' meaningful reality on the other side of discourse which determines the truth of the meanings produced within discourse. Discourse analysts restrict the historical task to reconstructing the conflicted process of the cultural mediation of meaningless 'reality,' the practical human activity of meaning construction, transmission, and transformation.Yet the actual practice of discourse analysis may raise doubts about historians' ability to finally extirpate the metaphysical impulse to represent a truth beyond the represented 'truths' of human cultural practices. In the work of Michel Foucault and many of the historians who follow his lead, for example, one finds an assumption that the actual workings of discourse are determined by a prior will-to-power, by a universal drive to impose meaningful order on meaningless 'reality,' to narrativize the disassociated fragments of experience, to stabilize the endless proliferation of contingent identities. Discourse analysis, thus, not only marks a turn from the social to the cultural, but also a turn toward the political in the ubiquitous sense that all practices in the world have a political intent, that the legitimation, expansion, and consolidation of power is not a particular sphere of historical action, but the form of all historical practice. Discourse analysis has also been characterized by a renewed and transformed interest in the history of the body and of nature. Both the human body and the world of external nature are constituted within discourse, but at the same time are primary sites of resistance to any particular discursive construction. Within the framework of discourse analysis human and nonhuman physicality cannot have a positive unmediated reality of their own, a reality capable of self-representation within its own nondiscursive, transcendental 'language.' Bodies may be constantly inscribed and re-inscribed as meaningful in various discursive networks, as laboring organisms, as reproductive mechanisms, as emblems of spiritual content, but their sheer intractable existence also marks the limits of discursive constitution and thus fuels an infinite process of re-inscription and re-constitution.

5. *Discourse Analysis and the Problem of Historical Agency*

Discourse analysis is premised on the claim that effective historical agency, and thus the existence of individual and collective subjects who are capable of productive, creative, transformative action in the world, is produced in history rather than expressed in history. It is not subjects who produce meaning, but the networks of meaning that create subjects. In some earlier forms of discourse analysis, which was heavily tilted toward an analysis of the homogenizing, totalizing regularities in discursive structures, subjects, or authors often appeared as nothing more than anonymous, replaceable bearers (*Träger*) of the structures that defined them. In many ways this conception of discourse as a systematic process of meaning production mirrored the position of social historians who defined individual and collective agents in history as the inhabitants of social functions, or as bearers of the process of material production. Historians who took the linguistic turn in the 1980s and 1990s, however, were concerned with revealing the historical contingency of determining structures, with deconstructing systems of meaning production that allowed certain identities—of class, race, gender, ethnicity, and so on—to be experienced as natural realities or metaphysical essences. Attentiveness to the conflictual, dialogical, communicative dimensions of discourses in which historical identities were constantly constructed, modified, dissolved, and transformed did assume the effective agency of historical subjects (both individual and collective) who were able not only to negotiate among conflicting, multivocal, sometimes contradictory discourses, but also to redefine themselves as subjective agents.

Discursive agency, however, was decentered and contingent, the agency of constantly shifting 'identities' that were the creations of discursive practices themselves. Historical subjects were 'subjected' to the logics of the various discursive networks that they inhabited. But within the limitations defined by historical situation and status they made choices that might ultimately change that situation and status, and thus also their own identities. The question of agency in discourse analysis was not so much whether or not agency was possible within its terms, but what kind of agency was imaginable. Within the complex heterogeneous compounds of any cultural or discursive formation, a particular individual might simultaneously inhabit a whole series of dispersed, fragmented subject positions. Within each of them—as parent, worker, consumer, or as sexual, racial, or national subject—the individual could act, either

individually or by identifying himself with other individuals similarly positioned in various discourses, to shift the hierarchical relations among identities, and to change the identities themselves. In discourse analysis, human agency, the identity and effectiveness of the individual and collective subject, was construed as an always coming-to-be in worldly particularly, as historical through and through.

This historicization of agency was at the same time a politicization of agency. The effective construction of specific collective identities, like nation, class, or race was described as a conflictual political process in which certain subject positions were excluded or privileged, became dominant or were repressed, attained stability through forceful imposition on others, or were negotiated in a more consensual fashion. The historicization and de-centering of human agency in the many historical works devoted to the cultural or discursive constructions of class, gender, national, or racial identities written during the late 1980s and through the 1990s did not fully expunge questions concerning the extralinguistic, culturally 'transcendent' grounding or foundation of human agency among historians. The writings of intellectual mentors of the linguistic turn, like Foucault, Walter Benjamin, or Jacques Lacan (and, in the1990s, Slavoj Zizek), were replete with references to the problematic nature not only of the physicality of the body but of the transcendent autonomy of some form of unmediated subjectivity—that which sought or was given shape and identity in the various 'subject positions' constituted in discourse—which appeared to be a necessary condition for the historical differentiation and transformation of subject identities. Historians might modestly decline professional competency in addressing such questions. But that which was outside discourse analysis inevitably shaped what occurred inside discourse analysis. The metaphysical limits of discourse analysis suggested that discourse analysis was itself a historical, that is, contingent and particular identity, a 'subject position' for historians of a particular place and time.

6. *Discourse Analysis and the Discourse of Historians*

The linguistic turn in history began with, and has continued to be characterized by, an intensive self-reflexive focus. With the publication of *Metahistory*, White (1973) introduced many historians to discourse analysis as an analysis of the discourse of historians. The primary object of concern was not so much the reconstruction of the way past identities were produced in language, as the construction of the individual and collective identities of historians through their discourses about the past. Stories historians told about the past were unveiled as stories in which they fashioned themselves, and possibly their readers, in the present. In the discourse on historical theory and historiography instigated by White's work, historians' texts are analyzed as linguistic artifacts akin to the rhetorical and fictional texts of literature, in which the structures of the linguistic form determine the 'objectivity' of the content. The discourse of objective realistic representation that had conventionally legitimated entry into the disciplinary community devoted to knowledge of the past was displayed as a mystifying, self-deceiving attempt to portray the subjective (that is, based on ethical and political considerations) construction of its object as a true representation of its object. The goal of such analysis was self-recognition and emancipation.

Once the burdens of determination by the putative objective reality of the past were recognized as self-imposed limitations, historians were free to imagine and construct alternative pasts in terms of their ethical and political commitments in the present and their visions of alternative futures. Moreover, as soon as the texts inherited from the past were grasped as historically contingent discursive constructions of meaningful identity in time, the inhabitants of the present could recognize themselves in them, not just constitute themselves through them. History could become a discovery of the 'reality' of the human past which was in itself a constantly reformulated attempt to represent reality as meaningful.

Two major questions continue to fuel the debate instigated by the analysis of the historical discipline as a discursive community. The first concerns the politics of this disciplinary community. How should the boundaries of legitimate historical discourse be determined in relation to other disciplinary discourses, or what are the limits of interdisciplinary crossing if history is to maintain a distinctive identity? Within the terms of discourse analysis this question must be decided through the communicative practices of historians rather than through an appeal to the object of analysis. The question then becomes one of defining the discursive process in which historical reality and truth are represented. Who is recognized as a 'subject' with rights to participate fully in the discourse, and how are the negotiations within this discourse to be regulated? The problem of reconstituting a common past becomes a political issue of extending membership and of democratizing power within the profession. Solidarity and objectivity are inextricably bound up with each other. One of the reasons that the linguistic turn and discourse analysis have become such contentious issues among historians is that they pertain directly to the identity of history and the identity of historians.

The second issue raised by the reflexive moment in historical discourse analysis relates to the more general issue that troubles the whole debate surrounding the linguistic turn. What are the ethical (and political) implications of construing the past as a construction of discourse? Is the creation of an inclusive, demo-

cratically negotiated consensus about the nature of the past adequate to the goal of recreating the reality of the object of historical investigation and defining the criteria of historical knowledge? As the ongoing debates about historicizing the Holocaust have suggested, discourse about the truth of the past must involve the dead as well as the living. To converse with the dead entails a recognition not only of their physical existence, their extradiscursive 'reality' as bodies that transcend all the meanings inscribed on them, but also their status as actual or potential subjects of a discourse of their own that must be taken into account in any discourse we might construct about them. It would seem that an adequate conception of historical discourse would need to include some recognition that historians' discursive production of the past can be interrupted by the voices of the dead, and that a part of the ethical obligation of the historian is to listen to those voices.

See also: Community Studies: Anthropological; Discourse, Anthropology of; Foucault, Michel (1926–84); Historiography and Historical Thought: Current Trends; History: Forms of Presentation, Discourses, and Functions; History: Theories and Methods; Linguistic Turn; Maroons in Anthropology; Postmodernism: Methodology

Bibliography

Baker K M 1990 *Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century*. Cambridge University Press, Cambridge, UK

Canning K 1994 Feminist history after the linguistic turn: Historicizing discourse and experience. *Signs: Journal of Women in Culture and Society* **19**(2): 368–404

Central European History 1989 Vol. 22, nos. 3–4. Special Issue: German Histories: Challenges in Theory, Practice, Technique

Chartier R 1988 *Cultural History: Between Practices and Representations*. Cornell University Press, Ithaca, NY

Eley G 1996 Is all the world a text? From social history to the history of society two decades later. In: McDonald T J (ed.) *The Historic Turn in the Human Sciences*. University of Michigan Press, Ann Arbor, MI, pp. 193–243

Friedlander S (ed.) 1992 *Probing the Limits of Representation: Nazism and the Final Solution*. Harvard University Press, Cambridge, MA

Foucault M 1972 *The Archaeology of Knowledge*. Pantheon, New York

Goldstein J (ed.) 1994 *Foucault and the Writing of History*. Blackwell, Oxford, UK

Hunt L (ed.) 1989 *The New Cultural History: Essays*. University of California Press, Berkeley, CA

Mills S 1997 *Discourse*. Routledge, London

Palmer B D 1990 *Descent into Discourse: The Reification of Language and the Writing of Social History*. Temple University Press, Philadelphia, PA

Schoettler P 1989 Historians and discourse analysis. *History Workshop* **27**: 37–65

Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York

Jones G S 1983 *Languages of Class: Studies in English Working-class History 1832–1892*. Cambridge University Press, Cambridge, UK

Torfing J 1999 *New Theories of Discourse: Laclau, Mouffe and Zizek*. Blackwell, Oxford, UK

White H V 1973 *Metapolitics: The Historical Imagination in Nineteenth-Century Europe*. Johns Hopkins University Press, Baltimore, MD

White H 1978 *Tropics of Discourse: Essays in Cultural Criticism*. Johns Hopkins University Press, Baltimore, MD

Williams G 1999 *French Discourse Analysis: The Method of Post-structuralism*. Routledge, London

J. E. Toews

Linguistic Typology

The term typology in linguistics, as in other fields, is generally understood to mean the classification of languages in terms of structural types. However, from the very beginning in the history of typological investigation, classification itself has never been the ultimate goal. For the grammarians of the early nineteenth century, exhaustive classification of the world's languages was a way to answer the question of what is possible variation in human language. For modern linguists typological classification is part of a method for discovering the constraints governing possible human languages or possible grammatical constructions.

Typological classification is frequently contrasted with the genetic classification of languages and is often characterized as ahistoric in its orientation. This was also not the case in the beginning of the typological studies in the early nineteenth century, when genetic and comparative approaches were not clearly distinguished from the typological one. Indeed, it was in the context of developing a program of comparative-historical research in his 1808 book *Über die Sprache und Weisheit der Indier* (On the Language and Wisdom of the Indians) that Friedrich Schlegel (1772–1829) developed a typological framework most famous in the history of typology. It was mainly due to the structuralist context, in which modern typology arose, that it countenanced a synchronic outlook. In contemporary typological scenes dynamicization of typology, as propounded most forcibly by Greenberg (1969, 1995), is fast obliterating the view that typology is a purely synchronic discipline.

1. Classical Typology

In the above-mentioned book Friedrich Schlegel proposed what is believed to be the first typological framework, according to which the languages of the world are classified into two types—flexional (or

inflectional) type and affixing (or agglutinative) type. Though later typologists, most notably August Wilhelm Schlegel (1767–1845) and Wilhelm von Humboldt (1767–1835), elaborated on his bipartite classification, Friedrich Schlegel, alluding to the ultimate goal of typological classification, made a bold claim that the opposition between *Sprachen durch Flexion* and *Sprachen durch Affixa* represents the two main categories of language, all other cases being mere modifications and variants of these.

The typological studies of the early nineteenth century are also known as morphological typology since the classificatory basis was the morphological structure of words. A beginning linguistics textbook typically introduces four types of language from this early endeavor: (a) inflectional (or fusional) type: relational meaning is expressed by internal modification of the lexical base such that component parts are not easily isolable (e.g., Indo-European languages, English *sing:sang:sung*); (b) agglutinative type: words consist of easily segmentable parts, each indicating a distinct grammatical meaning (e.g.,Turkish *ev-ler-im-de* house-PL-1SG.POSS-LOC 'in my house'); (c) isolating type: words do not give overt expression of relational meaning or else do so by employing to the end the same kind of unit as is used for encoding lexical concepts (e.g., in Chinese the verb *gěi* 'to give' is used also as an indirect object marker); (d) incorporating (or polysynthetic) type: nominal arguments (typically objects) and pronominal forms referring to them are incorporated in the verb form (e.g., Ainu *a-wakka-ta-re*1SG-water-draw-CAUS 'I make someone draw water').

Classical typology embodied system internal inconsistencies and had associated value judgements that caused its decline in the second half of the nineteenth century, when the study of genetically related languages was viewed as the only possible scientific comparative study of languages. Different languages were thought to represent different stages of the development of the languages and language families with the belief that the inflectional type, which fused distinct morphological information into single morphemes, represented the highest degree of linguistic sophistication. This assumption put classical Indo-European languages at the top of the evolutionary ladder and isolating languages of Asia, e.g., Chinese, at the bottom, contradicting the long history of the Chinese language and the advanced civilization associated with Classical Chinese.

The quadripartite classification has shortcomings as a true classificatory scheme in that a language may have morphological characteristics of several types simultaneously. English, for example, shows an inflectional characteristic in some irregular forms (e.g., *bring:brought, man:men*), an agglutinative characteristic in regular verbal and nominal inflections (e.g., *walk:walked, book:books*), an isolating characteristic in certain modal expressions (e.g., *can not walk*), and an incorporating feature in some compound forms (e.g., *to hand-knit a sweater*). Edward Sapir's multidimensional typology developed in his 1921 book *Language* (Sapir 1921) was in part response to the need to accommodate gradient characterizations of linguistic types, as he saw that languages in their entirety could not be neatly pigeon-holed into a given type, the matter being a question of tendency. While classical typology saw the culmination in Sapir's highly elaborate classificatory scheme, his followers have not generally put it into practice. Indeed, the only place where classical, morphology-based typology has been seriously pursued is Prague, where Vladimir Skalička (1909–1991), who recognized five types of languages (the four previous types plus the introflexive type), was a pillar of the so-called Prague School of Typology (see Sgall 1995).

While morphological typology as formulated by the nineteenth century grammarians is no longer considered to be adequate in meeting the ultimate goal of language typology, classical typology saw several developments of contemporary relevance. It is perhaps even fair to say that the followers of classical typology have contemplated most contemporary issues of language typology. For example, August Schleicher (1821–1868) recognized the possible connections between morphological characteristics and the manner in which grammatical relations are expressed, thereby pointing to the current view that a language type is to be defined in terms of a combination of properties that exhibit correlative patterns. The evolutionary chain of isolating-agglutination-inflection envisioned earlier is now recast as a cyclic development and such a model is integrated in the current study of grammaticalization (see Bybee 1985, Hagège 1988). The Humboldtian claim that the structure of a particular language reflects its speakers' *Weltanschauung* has stimulated a line of research on the relationship between language types and the speakers through the nineteenth century leading up to the so-called Sapir–Warf Hypothesis in the early twentieth century. This hypothesis claims a connection between the linguistic structure and the way the speakers perceive the outside world. As typological studies take on more cognitive outlook (see DeLancey 1981), a reappraisal of the Humboldtian tradition may be in order. An interesting development along this line is the recent revival of interest in ethnopsychology and the development of what is dubbed as 'ethnosyntax' by Wierzbicka (1992) and others.

2. *Modern Typology*

A major difference between classical typology and modern typology, developed in the middle of the twentieth century, lies in its scope and methodology. Classical typology sought a global characterization of

the entire language according to a small number of typological characteristics. This holistic approach has proved to be too ambitious, and modern typologists practice partial typology, where specific phenomena or individual grammatical constructions are studied. Methodologically, modern typology heavily relies on cross-linguistic distributional patterns. Implicational universals are drawn from the distribution of existing types and non-attested types.

Roman Jakobson (1896–1982) is generally regarded as the founding father of modern typology. In his seminal paper, Jakobson (1958/1971) recognized that typological analysis leads to the discovery of the dependency relation between two elements such that the presence of one element, X, implies the existence of another element, Y, in the same system (but not vice versa). For example, the parameters of voicing and aspiration in consonants yield the following four types of consonant system (illustrated with dentals); (1) /t/, /d/; (2) /t/, /t^h/, /d/; (3) /t/, /t^h/, /d/, /d^h/; (4) /t/, /d/, /d^h/. Of these four logically possible types, only the first three are actually observed—no language has voiced aspirates in the absence of voiceless aspirates, i.e., the fourth system. From this, it is possible to formulate an implicational statement that the presence of voiced aspirates implies the presence of voiceless aspirates. This statement constitutes a universal on human language if no language indeed violates the pattern implied by this statement—that no consonantal system can contain voiced aspirates in the absence of voiceless aspirates.

Developing Jakobson's insight in a systematic and over-arching manner, Joseph H. Greenberg (1915–2000) has changed the whole perspective of linguistic typology transforming it into a new method of universal research. Greenberg's (1966) study on word order typology drew a large number of implicational universals encompassing both morphological and syntactic levels, and some transcending them. That languages differ in word order is a common place experience of those who are exposed to foreign languages. English is an SVO (Subject-Verb-Object) language, while Japanese is an SOV language. The less frequently observed VSO order is instantiated by Welsh. The impact of Greenberg's work was very strong because of the far-reaching predictions it made on the basis of this kind of simple, basic feature of language. For example, it has been shown that if a language has dominant VSO order, (a) it has prepositions (G's *Universal 3*); (b) the genitive follows the governing noun (G's *Universal 2*, originally formulated in terms of the presence of prepositions); (c) question particles which are specified in position by reference to a particular word in the sentence do not occur (G's *Universal 10*); (d) it always puts interrogative words or phrases first in interrogative word questions (G's *Universal 16*); (e) an inflectional auxiliary precedes the main verb (G's *Universal 16*); and (f) adjectives come after the noun (G's *Universal 17*).

Greenberg's universals have the following two characteristics. They are implicational, such that they are restricted universals. The universals cited above obtain only with regard to VSO languages, and they have nothing to say about those languages having a different word order. Secondly, Greenberg recognizes statistical universals or high cross-linguistic tendencies, which do not obtain absolutely. For example, Greenberg's *Universal 4* is formulated as: 'with overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional,' which allows a language like Persian, which has SOV order with prepositions. Beside implicational and statistical universals, linguists in general recognize unrestricted universals, which are believed to obtain absolutely; e.g., all languages have vowels and consonants; all languages have nouns and verbs.

Implicational statements are also obtainable from hierarchies that grammatical categories exhibit. The most well known grammatical hierarchy is that of grammatical relations formulated by Keenan and Comrie (1977): Subject > Direct Object > Indirect Object > Genitive > Oblique Object > Object of Comparison. This hierarchy, derived from the cross-linguistic relativization patterns, predicts, for example, that if a language allows relativization on an oblique object (i.e., an object of a preposition), then it also allows the same relativization strategy to apply to all those nominals holding the grammatical relations higher in the hierarchy than the oblique object.

Another hierarchy with far-reaching ramifications is the so-called animacy hierarchy: first/second person > third person > pronoun > human noun > animate noun > inanimate noun. In a language such as the Australian language Dyirbal, where an agentive nominal of a transitive clause may receive either ergative or nominative marking—split-ergative marking—those high on the hierarchy tend to be marked nominative, while those low on the hierarchy tend to be marked ergative, such that we can draw implicational statements to the effect that if an agentive nominal X in the hierarchy is marked nominative, all those higher in the hierarchy are also marked nominative. The animacy hierarchy is also at work in voice phenomena such that an agentive nominal high in the hierarchy tends to be expressed in subject position, i.e., in the active voice, while when a patient nominal is high in the hierarchy, passivization tends to occur. A corollary of these tendencies is that those nominals high in the hierarchy are less likely to be expressed overtly as an agentive phrase (as in the English *by* agentive phrase) in the passive clause.

Recent studies of prototype categories have revealed that grammatical categories and constructions form continua rather than exhibiting rigid category boundaries. For example, studies on transitivity have shown that the traditional binary distinction between intransitive and transitive verbs (or constructions) needs to be recast as a multi-valued continuum (see Hopper

and Thompson 1980). A simpler and much coarser form of this continuum yields the following hierarchy: Transitive > Active intransitive > Inactive intransitive, where active intransitive verbs/constructions are those expressing activities (e.g., 'to dance'), and inactive intransitive verbs simple processes (e.g., 'to fall'). An implicational statement can be also derived from this hierarchy; e.g., if a language permits passivization of an inactive intransitive, then it also permits active intransitives and transitive verbs to undergo passivization. Also noun incorporation (see above) basically follows this hierarchy such that the object of a transitive clause is the easiest to incorporate, followed by the subject of an inactive intransitive clause, the subject of an active intransitive clause, and finally the subject of a transitive clause. Thus, if a language incorporates the subject of an active intransitive clause, then it is likely to incorporate the subject of an inactive intransitive clause and the object of a transitive clause.

3. *Functional Typology*

The term typology is used in reference to a theoretical and methodological approach that contrasts with other, especially Chomskyan generative, approaches. The Chomskyan generative approach is said to be formal in two senses. In one sense, Generative Grammar is formal in dealing with the 'abstract formal skeleton' (Chomsky 1975p. 55) of syntactic structures as an autonomous system independent of semantic or functional factors. In another sense Generative Grammar is formal in that its explanations are in terms of the system internal elements and structural relations. The typological approach sharply contrasts with these two characteristics of Generative Grammar. Its basic orientation is functional in that it assumes that the central function of language is to represent cognition and to communicate experience and thought. The task of typology is to determine the range of variation that language permits in achieving this central function. In this framework a language is seen as a problem-solving system; how languages conceptualize a given cognitive domain and express the concepts by linguistic means (see Seiler 1995). Studies dealing with specific constructions such as causative and possessive constructions thus start out with the formulation of the concepts or the cognitive domains to be expressed and typologize the relation between grammatical form and semantic function (see Shibatani 1976 and Seiler 1995). This cognitive-functional stance has a long and venerable history linking the current works with Sapir's and ultimately with the Humboldtian conception of language as an *'enérgeia'* (a creative activity) rather than an *'érgon'* (a product).

The term 'functional typology' is used in recent years in order to underscore the functional orientation the current typological research countenances. This approach also differs from Generative Grammar in its recourse to external explanations. Explanations for cross-linguistic patterns of linguistic structure are sought in functional motivations such as iconic motivation—preference for the structure of language to reflect and the conceptual structure in a parallel manner (see Haiman 1985). Haiman has shown that conceptual distance is reflected in grammatical distance. Thus, inalienable possession, which involves an inseparable or a closely connected whole-part relation, is expressed by a more tightly integrated form than inalienable possession; e.g., *m-pôlu* 'my back' vs. *ŋa pɛ́rɛi* 'my house' (Kpelle). Similarly, in causative expressions, shorter forms tend to express direct causation and the longer forms indirect causation, e.g., *kill* vs. *cause to die*. Another line of research along the functional typological premise has been developed by Hawkins and his associates in which processing mechanisms are invoked as an explanation for observed linguistic patterns (see e.g., Hawkins 1990, Hawkins and Cutler 1988).

4. *Typology and Language Change*

The importance of typology to one of the major concerns of historical linguistics, namely reconstruction of proto-languages, has long been demonstrated. The basic idea here is that where a reconstruction conflicts with the known typological facts, there is a *prima facie* reason to suspect its validity. The most celebrated phonological example of this use of typology as a control on reconstruction is concerned with the Proto-Indo-European consonant system. The standard reconstruction posited a three-fold system of voiced and aspirated consonants (illustrated with dentals) /t/, /d/, /dh/. However, Jakobson (1958/1971) noted that typological investigation revealed that no language would contain /dh/ in the absence of the voiceless counterpart /th/, casting doubt on the validity of the Proto-Indo-European system reached by the standard reconstruction. This has led to either reinstating the Neogrammarians' earlier four-point stop system of /t/, /th/, /d/, /dh/, or developing alternative hypotheses. The so-called glottalic theory, which replaces the traditionally reconstructed voiced stops with ejectives, has also been informed of typological results. Phonological systems comprising a voiceless, a glottalized, and an aspirated series (/t/, /t$^{?}$/, /t^{h}/) conform to synchronic typological findings, but the system arrived at by the conventional reconstruction does not (see Gamkrelidze and Ivanov 1973).

The move to bring typology to the heart of diachronic linguistics, namely linguistic change itself, has been spearheaded by Greenberg (1969, 1995), who advocates the dynamicization of a synchronic ty-

pology. Greenberg offers what he calls the 'state-process model,' which views attested language types as possible stages through which languages can pass and non-attested language types as impossible stages into which a language can change. In other words, just as synchronic typology puts constraints on language types, constraints on typological change can be viewed as limitations on language change.

The connection between synchronic typology and possible change can be illustrated by the typology of number categories in the noun. It has been shown that the category of dual is only found in a system having a singular/plural distinction—no language has a singular/dual system or a plural/dual system. When a language with no number distinction develops a number system, it involves a singular/plural distinction; no language develops a system with a tripartite singular/dual/plural distinction without going through a stage in which a singular/plural distinction has been established. In this way, diachronic typology seeks to identify allowable and non-allowable transition from one state to another so as to determine possible paths of language change.

In the Greenbergian diachronic typology, typological states and transitions between states are characterized by two logically independent factors—stability and frequency. Stability represents a probability that a language, which is in a particular state, will exit from that state, while frequency is the likelihood that a language will enter a particular state. For example, a state of having nasal vowels is characterized as having high frequency but low stability, because a vowel frequently acquires the nasality feature from an adjoining nasal consonant, but once a system acquires nasal vowels, it tends to retain them. The dimensions of frequency and stability of typological states render the state-process model a probabilistic one, which countenances the gradualness of the transition from one language state to the next. In this model synchronic exceptions are accounted for as transitional states arising from gradual diachronic process of change from one state to another.

The two main concerns of diachronic typology—gradualness of historical change and its directionality—have received focused attention in the recent developments in grammaticalization theory (see Bybee 1985, Heine et al. 1991, Hopper and Traugott 1994, Lehmann 1982). Grammaticalization refers to the process by which lexical items come to take on grammatical functions such as marking of tense, aspect, grammatical relations, definiteness, modality, etc. In Chinese the verb *gěi* 'to give' is used as an indirect object marker, and the verb *iru* 'to be/to exist' in Japanese has developed an aspectual usage of indicating an on-going activity, the progressive aspect.

Grammaticalization typically involves synchronized change in several different grammatical dimensions. As a lexical item gets grammaticalized, it tends to undergo phonological reduction, which in turn renders the form less independent, downgrading it to a morphologically dependent status as in the case of enclitics and affixes. Semantically, a lexical item gradually loses its original referential meaning and acquires such relational meanings as those previously noted. As in the case of the Chinese verb *gěi* and the Japanese verb *iru*, Grammaticalization is a gradual change such that these forms co-exist as verbs along with the newly developed grammatical functions. The two main concerns of grammaticalization theory are the relationship between the input and the output of grammaticalization—what kind of lexical items give rise to what kind of grammatical functions?—and the directionality of change—is the change unidirectional from concrete to abstract?

The gradual nature of historical change brought to the fore by diachronic typology and grammaticalization research has an important implication to the matter of definition as well as the synchronic characterization of a language. Many phenomena are in transition from one state to another, and as such they are not amenable to categorical definitions. For example, whether Chinese has prepositions or not is not an easy question to answer, as forms such as the indirect object maker evolving from the verb *gěi* 'to give' can be construed either as verbs or prepositions. The existence of transitional phases of this kind calls into question the traditional distinction between synchronic linguistics and diachronic linguistics. Dynamicization of a synchronic description is the next logical step that diachronic typology suggests.

5. *Typological Differences in Clausal Organization*

As described in Sect. 4, modern typology typically examines specific domains or individual grammatical constructions in order to ascertain possible variation in human language. This leaning toward partial typology stems from the difficulty encountered in attempting to typologically characterize languages in their entirety. Despite this, there is a lingering sense among typologists that a holistic characterization of a language is a desideratum. Georg von der Gabelentz most eloquently expressed this feeling at the beginning of the twentieth century:

> But what an achievement would it be were we to be able to confront a language and say to it: "you have such and such a specific property and hence also such and such further properties and such and such an overall character"—were we able, as daring botanists have indeed tried, to construct the entire lime tree from its leaf. (Gabelentz 1901, p. 481)

Whether or not the ideal state entertained by Gabelentz can ever be reached remains to be seen. However, one area that merits further investigation is

the differences that languages show in the organization of clause elements. In compensation for the emphasis on methodological matters in the preceding *exposé*, the following allows a glimpse of what actual typological investigation or comparison of languages in general involves.

In expressing a transitive situation involving an agent and a patient participant, languages typically distinguish these in some way, so that who is acting on whom is made clear. English, for example, achieves this by using different pronominal forms; e.g., 'He hit him.' In the case of intransitive clauses, its sole nominal argument (S) can be assimilated either to the form of an agentive nominal (A) or that of a patient nominal (P). English assimilates it to the nominal form of A—'He fell, He ran.' A clause that treats S and A alike and distinguishes P from these is said to display the nominative-accusative coding pattern ($S = A \neq P$), and the languages possessing such clauses predominantly are called nominative-accusative type languages.

Some languages organize clauses differently from the nominative-accusative pattern. Many Australian aboriginal languages follow the absolutive-ergative pattern, whereby S is identified with the P nominal a transitive clause, while distinguishing A from these; i.e., the $S = P \neq A$ pattern. Clauses may be organized in yet another way. The American Indian language Lakhota, for example, splits S nominals into an agent-like category (S_a) and a patient-like category (S_p), yielding the coding pattern, known as active-type: $S_a = A \neq P = S_p$.

All of these languages have a single type of dominant transitive clause, where the coding of A and P is fixed. Some languages have more than one type of transitive clause making a unique comparison of intransitive clauses with transitive clauses problematic. Philippine languages are such languages, where a simple transitive clause comes in two variant forms. Cebuano, for example, renders the sentence 'Juan read the book' in two ways; *Ni-basa si Juan sa libro* (AF-read TOP Juan the book), *Gi-basa ang libro ni Juan* (GF-read TOP book ACT Juan). That is, either A or P can be coded as what is called Topic (marked *si* for proper names and *ang* for common nouns) of a transitive clause. The S nominal of an intransitive clause is coded as a Topic as in: *Ni-dagan si Juan* (AF-run TOP Juan) 'Juan ran.' (AF = ctor-focus marking; GF = goal-focus marking; TOP = Topic marker; ACT = Actor marker).

The focus marking in the verb indicates what semantic role is being coded as Topic, such that the prefix *ni-* in Cebuano indexes an agentive Topic, and *gi-* a patient topic. This applies to both transitive and intransitive clauses such that an intransitive clause with a patient S Topic would trigger *gi-* marking; *Gi-kapoy si Juan* (GF-tired TOP Juan) 'Juan is tired.' This split treatment of S Topic marking in the verb is reminiscent of the active-type-coding pattern. However, Philippine languages are not active-type languages in that the nominal marking is consistent for S—it is always realized as *si/ang*-marked Topic—and that there are two competing forms of a transitive clause (see Shibatani 1985 on the Philippine voice system).

The four types of coding pattern examined have important implications for some basic grammatical categories. First, the notion of Subject is problematic in those clause types displaying coding patterns other than the nominative-accusative pattern. Subject, as found in European languages, refers to the category obtained from the $S = A$ grouping in the nominative-accusative pattern. It is a category deriving from the assimilation of S to A, and accordingly is agent-based.

Morphological identification of S and P obtains in an ergative-type language. A long-standing question in typology is whether this grouping of S and P should be identified as Subject. One obvious problem with this move is that whereas Subject is an agent-based category, the $S = P$ grouping is a patient-based category, since it obtains by assimilating S to P. Thus a conservative view is that the $S = P$ grouping forms a category distinct from Subject. Since S and P typically involve zero marking, the term Absolutive suggests itself as the label for this category.

In active-type languages, there is no unique identification between the totality of S and A or P; and therefore, neither Subject nor Absolutive category obtains. In Philippine linguistics, Topic is often likened to Subject, but since it is based on neither A nor P—either can be Topic and S can be identified with either kind of Topic—it is neither an agent-based nor a patient-based category.

While different clausal types of the world's languages may display grammatical relations resembling Subject, they are not entirely identical. Subject is agent-based, whereas Absolutive is patient-based. Philippine Topic is neutral in its alignment with the semantic roles. These different grammatical relations, however, can be identified at a functional level, in that they are the essential relations that a nominal must bear in clause formation. That is, Subject, Absolutive, and Topic are an indispensable nominal category in a clause over which a verb predicates. Predication will not be complete without a nominal category about which the verb says something. The primary status of these grammatical relations is reflected in the primacy they enjoy in such grammatical processes as relativization and focusing. Subject, Absolutive, and Topic, in other words, can be identified as a Primary Relation at the functional level, despite the differences at the substantial level.

The differences in grammatical relations have some additional implications. Firstly, it predicts that those languages that do not have the Subject category do not show the active-passive voice opposition. Active voice (as expressed by 'John hit Bill') is defined in terms of the alignment of A and Subject, while passive voice typically involves the alignment of P and Subject, as in

'Bill was hit by John.' From this it is obvious that languages without the Subject category cannot have the active-passive opposition. Still, other languages may show a different kind of voice contrast. Ergative languages often display ergative-antipassive voice opposition, characterized by the realignment of the absolutive category over P and A.

In both active-passive and ergative-antipassive oppositions, there is a basic voice associated with the basic form of the verb. In the former, active is the basic voice, and in the latter, the ergative voice is basic. Thus, a transitive situation is normally expressed either as an active form or as an ergative form in the languages having these systems. Derived voice forms, the passive and the antipassive, contain derived verb forms, and they are used far less frequently than the unmarked, basic voice forms. Philippine-type languages, in contrast, have no unmarked voice category—the alignment of Topic with A or P is equally marked in the verb. A characteristic of the fluid-voice system of this kind is that both A-Topic form and P-Topic form are used with equal facility such that there is no lop-sided distribution of voice forms as in the other two systems discussed previously. In contradistinction to these types of languages, active-type languages are known to lack voice opposition of the kind discussed here. The three voice mechanisms studied—the active/passive system, the ergative/antipassive system, and the fluid-voice system of Philippine languages—are different in their substance. Yet they are comparable at the functional level in that they all involve diathetic realignment of a Primary argument.

Languages can differ with respect to the nature of grammatical relations and voice mechanisms depending on the way clauses are organized. Besides these, cross-linguistic differences in other grammatical phenomena have been reported, some of which may be attributable to the typological differences outlined here. The question of whether these differences have profound implications for the overall organization of grammar aside, it is becoming increasingly clear that languages are likely to differ at the level of grammatical organization far more greatly than was believed in the latter half of the twentieth century. This suggests that cross-linguistic similarity be sought in the functional equivalence or similarity rather than in the specific manners in which languages define and organize grammatical elements.

See also: Grammar: Functional Approaches; Grammatical Relations; Grammatical Voice; Grammaticalization; Historical Linguistics: Overview; Languages: Genetic Classification; Linguistics: Comparative Method; Linguistics: Iconicity; Linguistics: Overview; Linguistics: Prototype Theory; Morphology in Linguistics; Relative Clauses; Valency and Argument Structure in Syntax; Word Order

Bibliography

Bybee J L 1985 *Morphology: A Study of the Relation between Meaning and Form*. John Benjamins, Amsterdam

Chomsky N 1975 *Reflections on Language*. Pantheon Books, New York

Comrie B 1989 *Language Universals and Linguistic Typology*, 2nd edn. University of Chicago Press, Chicago

Croft W 1990 *Typology and Universals*. Cambridge University Press, Cambridge, UK

DeLancey S 1981 An interpretation of split ergativity and related patterns. *Language* **57**: 626–57

Gabelentz G von der 1901 *Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Tauchnitz, Leipzig, Germany

Gamkrelidze T V, Ivanov V V 1973 Sprachtypologie und die Rekonstruktion der gemeinindogermanischen Verschlüsse. *Phonetika* **27**: 15–56

Givón T 1984/1991 *Syntax: A Functional-Typological Introduction*. 2 Vols. John Benjamins, Amsterdam

Greenberg J H 1966 Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg J H (ed.) *Universals of Grammar*, 2nd edn. MIT Press, Cambridge, MA

Greenberg J H 1969 Some methods of dynamic comparison of linguistics. In: Puhvel J (ed.) *Substance and Structure of Language*. University of California Press, Berkeley, CA

Greenberg J H 1974 *Language Typology: A Historical and Analytic Overview*. Mouton, The Hague, The Netherlands

Greenberg J H 1995 The diachronic typological approach. In: Shibatani M, Bynon Th (eds.) *Approaches to Language Typology*. Clarendon Press, Oxford, UK

Greenberg J H, Ferguson C A, Moravcsik E A (eds.) 1979 *Universals of Human Language*. 4 vols. Stanford University Press, Stanford, CA

Hagège C 1988 Contribution des recherches typologiques. In: Joly A (ed.) *Linguistique génétique. Historie et théorie*. Presses Universitaire de Lille, Lille, France

Haiman J 1985 *Natural Syntax: Iconicity and Erosion*. Cambridge University Press, Cambridge, UK

Hawkins J 1990 A parsing theory of word order universals. *Linguistic Inquiry* **21**: 223–261

Hawkins J, Cutler A 1988 Psycholinguistic factors in morphological asymmetry. In: Hawkins J (ed.) *Explaining Language Universals*. Blackwell, London

Heine B, Claudi U, Hünnemeyer F 1991 *Grammaticalization: A Conceptual Framework*. The University of Chicago Press, Chicago

Hopper P, Thompson S A 1980 Transitivity in grammar and discourse. *Language* **56**: 251–99

Hopper P, Traugott E 1994 *Grammaticalization*. Cambridge University Press, Cambridge, UK

Jakobson R 1958/1971 Typological studies and their contribution to comparative-historical linguistics. In: *Selected Writings*. Mouton, The Hague, The Netherlands, Vol. 1

Keenan, E, Comrie B 1977 Noun phrase accessibility and universal grammar. *Linguistic Inquiry* **8**: 63–99

Lehmann C 1982 Thoughts on grammaticalization. A programmatic sketch. *Arbeiten des Kölner Universalien-Projekts* **48**

Ramat P 1987 *Linguistic Typology*. Mouton de Gruyter, Berlin

Sapir E 1921 *Language*. Harcourt, Brace and World, New York

Seiler H 1995 Cognitive-conceptual structure and linguistic encoding: Language universals and typology in the UNITYP

framework. In: Shibatani M, Bynon Th (eds.) *Approaches to Language Typology*. Clarendon Press, Oxford, UK
Sgall P 1995 Prague school typology. In: Shibatani M, Bynon Th (eds.) *Approaches to Language Typology*. Clarendon Press, Oxford, UK
Shibatani M (ed.) 1976 *The Grammar of Causative Constructions* (*Syntax and Semantics 6*). Academic Press, New York
Shibatani M 1985 Voice in Philippine languages. In: Shibatani M (ed.) *Passive and Voice*. John Benjamins, Amsterdam
Shibatani M, Bynon Th (eds.) 1995 *Approaches to Language Typology*. Clarendon Press, Oxford, UK
Wierzbicka A 1992 *Semantics, Culture, and Cognition*. Oxford University Press, Oxford

M. Shibatani

Linguistics: Comparative Method

The comparative method is the most important of the various techniques used to recover linguistic history. It is a method (or a set of procedures) which compares related languages, descendants from a common ancestral language, in order to postulate, to 'reconstruct,' the ancestral language. It is commonly considered a major event in the history of ideas; inspired by the success of the comparative method in linguistics, numerous other disciplines in the nineteenth century adopted comparative methods of their own. In this article the comparative method is explained and its basic assumptions and limitations are considered.

1. The Comparative Method and Language Families

Language which belong to the same family are said to be genetically related—this means that they derive from a single ancestral language, called a proto-language. Genetic relationships among languages are talked about in genealogical terms: 'sister languages,' 'daughter languages,' 'parent language,' and 'language families' (definitions below). Through linguistic changes, dialects develop in a language, and then as further changes accumulate in time, these dialects become distinct languages, related to one another ('sister languages') and descendants ('daughters') of the original language (the parent or proto-language). By applying the comparative method to related languages which descend from an original ancestor language, linguists postulate what that proto-language was like—that is, they reconstruct that language. For example, a comparison of English with its relatives, Dutch, German, Danish, Swedish, Icelandic, and so on, reveals what the proto-language, in this case called 'Proto-Germanic,' was like. Thus, English is, in effect, a much changed 'dialect' of Proto-Germanic, having undergone successive linguistic changes to make it what it is today, a different language from Swedish and German and its other sisters, each of which underwent different changes of their own. Therefore, every proto-language was once a real language regardless of whether it is successfully reconstructed or not, and all languages which have relatives have a history which classifies them in language families.

The aim of the comparative method is to reconstruct as much as possible of the proto-language by comparing its daughter languages (the related descendant languages), and to determine what changes they underwent as they diversified from their common ancestor. The comparative method serves two purposes: to show that languages are related, and to reconstruct as much as possible of the parent language. If reconstruction is successful, it shows that the initial assumption that the languages are related was warranted, and it reveals the nature of the proto-language.

Of course, success is dependent upon the extent to which evidence of the original traits is preserved in the descendant (daughter) languages and upon how astute the linguist is at applying the comparative method to them. For example, by comparing Romance languages (Spanish, Portuguese, French, Italian, etc.), the linguist reconstructs Proto-Romance; Proto-Romance is equivalent to the spoken language at the time Latin began to split up into its descendants. If successful, what is reconstructed for Proto-Romance by the comparative method should coincide with the Proto-Romance which was actually spoken at the time before it split up into its daughter languages. In this case, since Latin is abundantly documented, it is possible to check to see whether what is reconstructed by the comparative method accurately approximates the spoken Latin that is known from written sources at that time. However, the possibility of checking reconstructions in this way is not available for most language families, for whose proto-languages there are no written records. For example, for Proto-Germanic (from which English, German, Swedish, etc. descend), there are no written records, and this language is known only from comparative reconstruction.

Often attempts to show that languages are genetically related or to reconstruct a proto-language begin with recognition of similarities among the languages being compared. However, similarities among languages can be due to a number of things, for example to borrowing, accident, onomatopoeia, etc., in addition to inheritance from a common ancestor (see Campbell 1998, pp. 311–26). Therefore, to show languages are related or to reconstruct successfully, similarities which are explained by factors other than common inheritance need to be avoided. The comparative method, aided by regular sound correspondences, helps to eliminate these factors.

2. Application of the Comparative Method

To be able to apply the comparative method, one needs to know some technical terms.

Proto-language. (a) The once spoken ancestral language from which daughter languages descend; (b) the language reconstructed by the comparative method which represents the ancestral language from which the compared languages descend. (To the extent that the reconstruction is accurate and complete, (a) and (b) should match.)

Sister language. Languages which are related to one another by virtue of having descended from the same common ancestor (proto-language) are sisters; that is, languages which belong to the same family are sisters to one another.

Cognate. A word which is related to a word in sister languages by reason of these forms having been inherited by these sister languages from a common word of the proto-language from which the sister languages descend.

Cognate set. The set of words which are related to one another in the sister languages because they are inherited and descend from a single word of the proto-language.

Sound correspondence. In effect, a set of 'cognate' sounds; the sounds found in the related words of cognate sets which correspond among related languages because they descend from a common ancestral sound. (A sound correspondence is assumed to recur in various cognate sets.)

Reflex. The descendant in a daughter language of a sound of the proto is said to be a reflex of that original sound; the original sound of the proto-language is said to be reflected by the sound which descends from it in a daughter language. (Cf. Campbell 1998, p. 112.)

In the application of the comparative method, first potential cognates among related languages (or among languages for which there is reason to suspect relatedness) are sought. It is convenient to begin with cognates from 'basic vocabulary' (body parts, close kinship terms, low numbers, common geographical terms), since these tend not to be borrowed from other languages. Ultimately, it is the systematic sound correspondences discovered in the application of the comparative method which demonstrate which are true cognates. Words which are similar due to accident, borrowing, and other non-genetic factors usually do not exhibit the recurring sound correspondences that true cognates do. Some cognates from Greek, Latin, and English (members of the Indo-European family of languages) are illustrated in Table 1.

Table 1
Indo-European cognates

Greek	Latin	English	
pod-	ped-	foot	'foot'
patḗr	pater	father	'father'

Next, the sound correspondences are sought. For example, in the cognates of Table 1, the first sound in each of the cognate sets exhibits the sound correspondence

Greek *p*, Latin *p*, English *f*

Since the sound correspondence recurs in these (and many other) cognate sets, it is unlikely that it could come about by sheer accident.

When the sound correspondences have been determined, the next step is to reconstruct the original sound. Linguists postulate what the original sound in the proto-language most likely was based on the phonetic properties of the sounds in the daughter languages as represented in the sound correspondence set. General guidelines help determine the most realistic reconstruction. For one, the known direction of certain sound changes is a valuable clue. For example, many languages have changed *p* to *f*, but change in the other direction, *f* to *p*, is extremely rare. In this case, it is said that the expected direction is *p* to *f* (and not *f* to *p*). This being the case, for the sound correspondence of Greek *p*-, Latin *p*-, English *f*-, **p* is reconstructed, since a change **p* to *f*, which would be necessary to derive the English reflex of *f*, is in an expected direction. (Note that the asterisk, as here with **p*, symbolizes reconstructed things.) If **f* were postulated for the proto-sound from which this sound correspondence descended, it would be necessary to assume the change of **f* to *p* in Greek and Latin, a change which goes against the expected directionality. Therefore, **p* is reconstructed in this case, together with the assumption that English underwent the common sound change of **p* to *f*.

Another guideline is the principle of 'majority wins': unless there is evidence to the contrary, for the proto-sound one reconstructs a sound coinciding with the particular sound in the correspondence set which is found in the greatest number of daughter languages. In this example, since *p* is found in both Greek and Latin, but *f* is only in English, 'majority wins' selects **p* for the reconstruction. Common sense is behind this; it is less likely that the same change (here **f* to *p*) would take place independently several times (here in Greek and in Latin) than that a single change would take place only once (**p* to *f*, only in English). The interpretation of the majority, however, can be complicated, requiring caution in the application of 'majority wins.'

This process of reconstructing sounds for the various correspondence sets is repeated until all the sounds of the proto-language have been reconstructed and the changes in the daughter languages worked out.

When all the proto-sounds have been reconstructed, it is possible to reconstruct words of the proto-language. For example, in the cognate set Greek *pod*, Latin *ped*, English *foot* 'foot,' **p* has already been reconstructed as the first sound of this word in the proto-language, based on the *p*-, *p*-, *f*- correspondence set just considered. For the second sound (*o*, *e*, *oo*), **e* is reconstructed on evidence not available in this short example; the correspondences for the vowel are complicated. For the last sound of the word, evidence from other cognate sets shows that the sound correspondence, Greek *d*, Latin *d*, English *t*, recurs and that the best reconstruction for this sound is **t*. Putting these reconstructed sounds together, **p* for the first sound, **e* for the second, and **d* for the third, gives the reconstruction **ped* 'foot,' a word in the proto-language. In this way the vocabulary of the proto-language can be recovered.

The reconstruction of a sound, a word, or large portions of a proto-language is, in effect, a hypothesis (actually a set of interconnected hypotheses) concerning what those aspects of the proto-language must have been like. Hypothesized reconstructions can be tested; sometimes they prove wrong; sometimes they must be modified, based on new insights. These insights may involve new interpretations of the data already on hand or new information that may come to light. The discovery of a previously unknown language may provide new evidence, a different testimony of the historical events which transpired between the proto-language and its descendants, which can change how parts of the proto-language are viewed. For example, with the discovery and decipherment of Hittite (and other languages of the Anatolian branch of Indo-European), the whole picture of Proto-Indo-European phonology changed.

There are numerous general treatments of the comparative method; for some of the clearer accounts, see Anttila (1989), Arlotto (1981), Campbell (1998), Crowley (1997), Hock and Joseph (1996), and Trask (1996).

3. *Assumptions and Limitations of the Comparative Method*

There are limitations on what can be recovered of a proto-language via the comparative method. Traditionally, some of these limitations are called 'assumptions,' since they are inherent consequences of how the comparative method is structured and how it is applied.

There are four main 'assumptions.'

(a) The proto-language was uniform, with no dialects or variation.

Clearly this 'assumption' is unrealistic, since all known languages have regional or social variation, different styles, and so on. However, it is not so much that the comparative method 'assumes' no variation; rather, it is just that there is nothing built into the comparative method which would allow it to address variation directly. This means that what is reconstructed will not recover the once spoken proto-language in its entirety, that is, with all its variation. Still, rather than stressing what is missing, linguists are happy that the method provides the means for recovering as much of the original language as it does. This assumption of uniformity is a reasonable idealization; it does no more damage to understanding of the proto-language than, say, modern reference grammars do which concentrate on a language's general structure, typically leaving out consideration of regional, social, and stylistic variation. Moreover, dialectal variation is not always left out, since in some cases scholars do reconstruct dialect differences to the proto-language based on differences in daughter languages which are not easily reconciled with a single uniform starting point.

Assumptions (b) and (c) are related:

(b) Language splits are sudden.

(c) After the split up of the proto-language, there is no subsequent contact among the related languages.

These assumptions are a consequence of the fact that the comparative method addresses directly only material in the related languages which is inherited from the proto-language and has no means of its own for dealing with borrowings, the results of subsequent contact after diversification into related languages. Borrowing and the effects of language contact are, however, not neglected. Rather, it is necessary to call upon other techniques which are not formally part of the comparative method for dealing with borrowing and the results of language contact. It is true that the comparative method contains no means for addressing whether a language gradually diverged over a long period of time before ultimately distinct but related languages emerged, or whether a sudden division took place with no subsequent contact between the parts of the original language community. Usually, however, reconstruction by the comparative method is reasonably successful, regardless of how slow or sudden the split up of the proto-language may have been.

(d) Sound change is regular.

This assumption is true; it is not a limitation, but an asset, extremely valuable to the application of the comparative method. For example, in the example above, Proto-Indo-European **p* regularly changed to *f* in English in all the words of this kind which formerly contained a **p*. Knowing that a given sound in given circumstances changes in a regular fashion permits reconstruction of what the sound was like in the parent language. If a sound could change in arbitrary, unpredictable ways, it would not be possible to determine from a given sound in a daughter language what it may have been in the parent language, and, for a particular sound of the parent language, it would not be possible to determine what its reflexes in its daughter languages would be. If, for example, an

original *p of the proto-language could arbitrarily for no predictable reason become f in some words, y in others, k in still others, and so on, then it would not be possible to reconstruct it. In such a situation, comparing, say a p of one language with a p of another related language would be of no avail, if the p in each could have come in unpredictable ways from a number of different proto-sounds.

4. How Realistic are Reconstructed Proto-languages?

The success of any given reconstruction depends on the material at hand with which to work, and the ability of the comparative linguist to figure out what happened in the history of the languages being compared. In cases where the daughter languages preserve clear evidence of what the parent language had, a reconstruction can be very successful, matching closely the actual spoken ancestral language from which the compared daughters descend. However, there are many cases in which all the daughter languages lose formerly contrasting sounds and lose grammatical categories through changes of various sorts. It is not possible to recover via the comparative method things in the proto-language if the daughters simply do not preserve evidence of them. In cases where the evidence is severely limited or unclear, reconstruction often results in omissions or mistakes. The comparative linguist makes the best inferences possible, based on the evidence available and on everything known about the nature of human languages and linguistic change. It is necessary to make the most of what is available to work with. Often the results are very good; sometimes they are incomplete, occasionally wrong. In general, the longer in the past the proto-language split up and the more linguistic changes that have taken place in the daughter languages, the more difficult it becomes to reconstruct with full success.

A comparison of reconstructed Proto-Romance with attested Latin provides a telling example. It is possible successfully to recover a great deal of formerly spoken Proto-Romance via the comparative method. However, the modern Romance languages for the most part preserve little of the former noun cases and complex verb inflections which Latin had. Subsequent changes have obscured these aspects of the grammar so much that large parts of Proto-Romance are not reconstructible by the comparative method. On the other hand, a great deal of Proto-Romance is recoverable.

An area of speculation and controversy concerns the outer limits of the comparative method. That is, is there a recognizable time depth beyond which the comparative method ceases to be effective? Some suggest that the comparative method cannot reach beyond the oldest currently reconstructed proto-languages; this cut-off date beyond which the comparative method is assumed not to be able to reach is variously speculated to be 6000, 8000, or 10,000 years. Since there is no reliable means of dating languages splits, these dates are rough guesses. The true issue is not really about dates at all, but rather about how much evidence related languages preserve upon which to base a reliable reconstruction. Typically, languages separated by many millennia have had longer in which to undergo changes, making them more different from one another than languages separated only by a few hundred years. Languages which have undergone many changes conceivably could preserve little of their original vocabulary or grammar; however, at the same time, it is also possible for a language to undergo relatively little change over relatively long periods. Thus, the upper limit of the comparative method is a function of whether accumulated changes have obliterated evidence in related languages, and not directly a matter of how long they have been separated.

See also: Areal Linguistics; Historical Linguistics: Overview; Languages: Genetic Classification; Linguistics: Internal Reconstruction

Bibliography

Anttila R 1989 *Introduction to Historical and Comparative Linguistics*, 2nd, rev. edn. Benjamins, Amsterdam

Arlotto A 1981 *Introduction to Historical Linguistics*. University Press of America, Washington, DC

Campbell L 1998 *Historical Linguistics*. MIT, Cambridge, MA

Crowley T 1997 *An Introduction to Historical Linguistics*, 3rd edn. Oxford University Press, Auckland

Hock H H, Joseph B D 1996 *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Mouton, Berlin

Trask R L 1996 *Historical Linguistics*. Arnold, London

L. Campbell

Linguistics: Iconicity

Ever since Plato, the majority view of students of language is that anything can stand for anything, the relationship between signans and signatum being determined by fiat or tradition in every manmade code. Within natural languages, in particular, the universal of double articulation ensures that, at the phonological level, individual signs are associated with no meaning whatsoever. At the morphosyntactic level, a distinction may be made between the lexicon and grammar. Challenges to the doctrine of arbitrariness thus fall into three groups.

At the phonological level, the notion of sound symbolism is such a challenge. Although submorphemic constant correlations between sound and

meaning are marginal in languages like English, it has been argued that they are much more pervasive in many languages of Africa and Southeast Asia. At the word level, the corresponding challenge is the notion of onomatapoeia, which has been acknowledged to be marginal in every language. In the artificial registers of poetry, the sound may indeed be made to be an echo to the sense, but this is not the case in flatfooted prose. At the level of compounds and more complex syntagms, however, diagrammatic iconicity in the sense of Peirce (1932) has always been both pervasive and acknowledged: among the staples of linguistic theory which have been strongly iconic are:

(a) The doctrine of semantic compositionality or Saussure's motivation (Saussure 1916): 'the meaning of a complex expression is in some way the sum of the meanings of its parts.'

(b) Behaghel's first law (Behaghel 1932, p. 4): 'conceptual closeness is reflected in physical closeness.' One of the clearest demonstrations of this is in Bolinger and Gerstman's (1957) discussion of the contrasts between phrases like 'light housekeeper' and 'lighthouse keeper,' in which, as they demonstrated, the perceptual phonetic contrast mirrors the orthographical distinction. Other work in the same vein is Greenberg (1966b)—restatements and elaborations of the traditional insight that derivational affixes are closer to the root than are inflectional ones; Givon (1980)—the degree of embeddedness of a complement clause is a reflection of the degree of control exerted by the subject of the containing clause; and Haiman (1985a)—contrasts between direct and indirect causation and between alienable and inalienable possession, and the semantic consequences of coordination reduction. Related is the idea that the physical incorporation of a string into a larger one corresponds to the degree of its conceptual incorporation in the material into the framework represented by the matrix string. Direct quotation resists incorporation, indirect quotation does not. Relevance conditionals (e.g., 'If you're hungry, there's food in the fridge') resist incorporation, ordinary conditionals do not in that the former do not 'count' as constituents in V/2 languages like German.

(c) 'The same form is used for the same meaning.' This is the basis of grammatical analogy (Anttila 1989, p. 89). The somewhat less familiar corollary, that different form is a reflection of different content, is the basis for Bréal's laws of repartition and specialization (Bréal 1897, p. 26). The 'same is same' heuristic also lies behind all the (never entirely successful) attempts to construct a 'notional' definition for parts of speech and grammatical categories in general (e.g., Jakobson 1936, Wierzbicka 1985), as well as for attempts to find explanations for recurrent polysemy of grammatical constructions.

(d) 'More form reflects more meaning.' This is still the most widespread diagnostic for markedness (see Greenberg 1966a, pp. 26–7). The clearest demonstration in typological research in Benveniste (1946) on the most frequent expression of the third person singular. As a creative dynamic principle of perception and interpretation, the 'more is more' heuristic is identified as what has led to the frequently attested restructuring of verbal paradigms (e.g., Watkins 1962). It is probably also what has led to the excrescence of derivational morphology in some less widely attested cases. In Khmer and in Romanian, for example, a reduced pronunciation of a verb was reinterpreted as the verb, the full form being reinterpreted as a derived nominalization or a derived causative.

(e) The order of clauses (e.g., in 'I came, I saw, I conquered') corresponds to the order of events, other things being equal (Jakobson 1965). Elaborations on this '*post hoc*, *ergo*, *post hoc*' heuristic are to be found in the Prague School notion of 'communicative dynamism,' itself a derivative of Behaghel's recognition (1932, p. 4) that the given precedes the new; Greenberg (1966b) on the order of mention of conjuncts, and of the order of protasis and apodosis; and Tai (1985) on the order of constituents within the clause. Hawkins (1994) and others have proposed that Behaghel's second law (that short constituents precede longer ones, as in Malkiel's 'irreversible binomials' (Malkiel 1959)) may also be reducible to this principle, given the tendency to give reduced expression to what is more familiar or given (frequently old) information.

(f) Conceptual symmetry is reflected in formal symmetry in a syntagm. Thus coordinate conjunctions, reciprocals, comparisons, and '*pari passu*' constructions like 'the more the merrier' tend to be symmetrically expressed in most languages. More subtly, conceptual symmetry is sometimes expressed by paradigmatic symmetry. The failure of reflexivization in sentences like 'I like me' and 'I want ME to stay behind' is a reflection of the loss of 'speaker privilege,' and hence, indirectly, a direct reflection of a divided self (Haiman 1998, p. 77): 'me' is treated no differently from (i.e., symmetrically with) any other participant on the perceptual stage. So, too, conceptual asymmetry may be reflected in formal asymmetry. Langacker (1985) argues that the structure 'there's snow all around (sc. me)' directly reflects the subject's inability to see himself, and thus reflects the asymmetry between the perceiving subject and the perceived object directly.

In spite of the opacity and arbitrariness in much of language, there is a sense in which we do not believe in our bones that it is real. In his short story 'Funes the memorious,' Borges showed what a totally arbitrary code would be:

In place of 7,013, he would say (for example) *Maximo Perez*; in place of 7,014, *The Railroad*; other numbers were *Luis Melian Lafinur*, *Olimar*, *suphur*, *the reins*, *the whale*, *the gas*, *the cauldron*, *Napoleon*, *Agustin de Vedia*. In place of 500, he would say *nine* … (Borges 1963, p. 64).

All students of language, however committed they may be to the idea that 'anything can stand for anything,' will probably find this conception of a language to be very alien. What Borges's imaginative and outrageous reductio forces us to acknowledge is the extreme unlikeliness of such willed original opacity. However opaque the signs and combinations of language may seem to us in any mature grammatical system, this system probably owes its present form to an originally iconic code like a pantomime (Donald 1991), modified over the course of time and use by a number of more or less user-friendly transparent signs and rules which 'seemed to be good ideas at the time.'

Given the acknowledged universality of motivation, analogy, markedness, Behaghel's law, and the extreme implausibility of truly arbitrary 'languages' like the number system of Funes, the interesting question is: what muddies the waters?

One obscuring factor is that limitations of the medium lead to competing (albeit impeccably iconic) motivations. For example, 'same form' is motivated not only by 'same meaning,' but by symmetry or 'lack of speaker privilege.' 'More form' is motivated not only by 'more meaning,' but also by 'less frequency' (Greenberg 1966a, on the relative marking of dual and plural), 'less familiarity,' greater surprise (Haiman 1998, on the marking of coreference), and 'more politeness' ('more form' is 'more wrapping,' cf. Geertz 1960). 'First mention' is motivated not only by being 'most familiar,' but also by Jespersen's (1949, p. 54) 'principle of actuality: what is at the moment uppermost in the speaker's mind tends to be first expressed.' Overburdened with too many semantic distinctions, a form may come to seem more or less arbitrary (Bolinger 1975, p. 110). So far, the most promising demonstration of the reality of iconicity in spite of this limitation is that where there are no competing motivations for the 'limited good' of a particular formal expression, there tends to be agreement (Haiman 1985a on word order).

Opacity or the apparent 'autonomy of syntax' is also the result of a number of diachronic factors, of which erosion, routinization, and systematization are perhaps the most important. We can illustrate some of these tendencies in a number system whose iconic origins are still very much in evidence—the Roman numerals.

The first three numbers are of course as iconic as a tally. The numeral V is apparently opaque, but still barely recognizable as a diagram of a human hand. The motivation for the abbreviation was probably no different from the motivation for the sound changes of truncation or assimilation: it is faster and easier to write V than to draw \\\\/. The numeral X can still be seen as a combination of two V's, one on top of the other. The numbers C and M are also icons, but of a different sort: C (entum) and M(ille) are examples of another kind of abbreviation, and illustrate competing iconic motivations coexisting within the same code.

Implicit in even the simplest compound sign, like II, is a rule which seems so natural that it is hardly recognized at first. It is that the combination XY is to be read as 'X' + 'Y.' This is as true of VI, VII, VIII, …, XXXIII, … MMCCCVIII, as it is of II. At some point, an unknown genius recognized and applied analogical reasoning to this natural rule: if XY is read as 'X' + 'Y,' then perhaps, under some (very restricted) conditions YX could be read as 'X' − 'Y.' The result, apparent in numerals like IV, IX, XC, and the like, is motivated within the system. But these numbers, even when they are drawn in the most painstaking and redundant detail, no longer correspond iconically to the quantities they represent. Analogy has led in this case to something that is the opposite of diagrammatic iconicity. The signs within a system cease to reflect reality and take on a life of their own when they are manipulated by rules internal to the system, as well as by the reality they still attempt to represent.

Within language, aspects of this are recognized as:

(a) the substitution of value for signification within a paradigm (the meaning of signs lies not in their relationship to the world, but in their relationship to each other, cf. Saussure 1916);

(b) grammaticalization, or the routinization of fixed structures beyond the contexts for which they were originally motivated; and

(c) the existence of various kinds of context-sensitivity within a syntagm (Haiman 1998, p. 149).

A compact example of systematization in syntax is the stress pattern of thetic sentences such as 'TRUMAN died' (Lambrecht 1988) in languages like English. A stress pattern should iconically reflect the contrast between old and new information, and a word order pattern should do the same. In English, iconic stress has led to sentences such as 'Truman DIED' where the predicate is new relative to the subject, but routinization has led to a fixed SVO word order, which is emancipated from the motivation of functional sentence perspective. Given that low stress … HIGH stress means 'predicate is new,' perhaps (by an analogical leap similar to the invention of the Roman numeral IV) HIGH stress … low stress means not only 'subject is new' (which it still iconically can do), but also 'everything is new.'

The Roman numeral system is far more iconic than the Arabic system, but it is already fully abstract and digitized. A more primitive system, corresponding to those languages in which there are no numerals higher than some 'highest digit' (frequently 2), resorts to subjective and analogic or iconic means of modelling higher numbers, typically {one, two, … many}. More languages resort to this analogical modelling in the domain of fractions, and present series like {one, a half, … a bit}. The digitization process 'colonizes' the analogic domain, substituting fixed quantities for impressionistic assessments like 'many' and 'a bit.' A roughly analogous 'colonization' process cannot be observed in the lexicon, or in morphosyntax, but it is

observed in the conventionalization of intonation (Léon 1993, Fonagy 1977): from being a purely analogical 'more is more' system, it becomes grammaticalized and transformed into an 'all or nothing' system—although never completely so (Bolinger 1961, 1986).

A future project for evolutionalists is to demonstrate that much of the undisputed arbitrariness of human languages, as of other human institutions, can plausibly be attributed to ritual and system: grammaticalization in the broadest sense.

See also: Language and Poetic Structure; Semantics; Symbolism (Religious) and Icon

Bibliography

Anttila R 1989 *Historical and Comparative Linguistics*, 2nd edn. Benjamins, Amsterdam

Behaghel O 1932 *Deutsche Syntax, Bd. IV*. Carl Winter, Heidelberg

Benveniste E 1946 Relations de personne dans le verbe. *Bulletin de la Société de Linguistique* **43**: 1–12

Bolinger D 1961 *Generality, Gradience, and the All-or-none*. Mouton, The Hague, Netherlands

Bolinger D 1975 *Aspects of Language*, 2nd edn. HBJ, New York

Bolinger D 1986 *Intonation and its Parts*. Stanford University Press, Stanford, CA

Bolinger D, Gerstman L 1957 Disjuncture as a clue to constructs. *Word* **13**: 246–55

Borges J L 1964 Funes the memorious. In: *Labyrinths*. New Directions, New York, pp. 59–66

Bréal M 1897 *Essai de sémantique*. Payot, Paris

Donald M 1991 *Origins of the Modern Mind*. Harvard University Press, Cambridge, MA

Fonagy I 1971 Double coding in speech. *Semiotica* **3**: 189–222

Geertz C 1960 *The Religion of Java*. Free Press, Glencoe, IL

Givon T 1980 The binding hierarchy and the typology of complements. *Studies in Language* **4**: 333–77

Greenberg J 1966a *Universals of Language, with Particular Reference to Feature Hierarchies*. Mouton, The Hague, Netherlands

Greenberg J 1966b Universals of language, with particular reference to the order of meaningful constituents. In: Greenberg J (ed.) 1966 *Universals of Language*, 2nd edn. MIT Press, Cambridge, MA, pp. 73–113

Haiman J 1985a *Natural Syntax*. Cambridge University Press, Cambridge, UK

Haiman J (ed.) 1985b *Iconicity in Syntax*. Benjamins, Amsterdam

Haiman J 1998 *Talk is Cheap*. Oxford University Press, New York

Hawkins J 1994 *A Performance Theory of Constituency and Word Order*. Cambridge University Press, Cambridge, UK

Jakobson R 1936 Beitrag zur allgemeinen Kasuslehre. In: Jakobson R (ed.) *Selected Writings, Vol. 2*. The Hague, pp. 23–72

Jakobson R 1965 Quest for the essence of language. *Diogenes* **51**: 21–37

Jespersen O 1949 *A Modern English Grammar on Historical Principles, Vol. VII*. George Allen and Unwin, London

Lambrecht K 1988 *Information Structure and Sentence Form*. Cambridge University Press, Cambridge, UK

Langacker R 1985 Observations and speculations on subjectivity. In: Haiman J (ed.) 1985 *Iconicity in Syntax*. Benjamins, Amsterdam, pp. 109–49

Léon P 1993 *Traité de phonostylistique*. Nathan, Paris

Malkiel Y 1959 Studies in irreversible binomials. *Lingua* **8**: 113–60

Peirce C 1932 *Philosophical Writings, Vol. 2*. Harvard University Press, Cambridge, MA

Saussure F de 1916 *Cours de linguistique Générale*. Payot, Paris

Tai J 1985 Temporal sequence and Chinese word order. In: Haiman J (ed.) *Iconicity in Syntax*. Benjamins, Amsterdam, pp. 49–72

Watkins C 1962 *Indo-European Origins of the Celtic Verb*. Institute of Advanced Study, Dublin, Ireland

Wierzbicka A 1985 'Oats' and 'wheat'. In: Haiman J (ed.) *Iconicity in Syntax*. Benjamins, Amsterdam, pp. 311–42

J. Haiman

Linguistics: Incorporation

Incorporation is a phenomenon whereby what might be expected to appear as a separate word is in fact part of the word on which it is dependent. The special nature of incorporation has been known to linguists for some time. Although it is known that the phenomenon is not restricted to indigenous American languages, it was first and foremost with respect to these languages that incorporation was, and still is, being discussed. In the following, a three-way distinction between noun incorporation, argument incorporation and grammatical incorporation as a continuum of morphological complexity is introduced.

Traditionally, noun incorporation is understood as a choice of whether arguments are retained on the syntactic level or not. Argument incorporation refers to the realization of the verb's arguments exclusively on the morphological level, i.e., verb internally. Here, no arguments can be identified in syntax, and the basic grammatical relations are realized at the word level. Noun phrases and other supplements have only adjunct status. Going a step further, functional items may be incorporated as well. In such cases of grammatical incorporation, various kinds of grammatical relations are realized verb internally. The term 'incorporation' will be used for all three types, regardless of whether the involved items are independent lexical units, i.e., words, or bound morphemes. Such a distinction becomes instrumental for differentiating the different types of incorporation.

1. Noun Incorporation

Since Kroeber (1909, p. 569) and Sapir (1989/1911, p. 254), noun incorporation has been characterized as stem combination, i.e., the combination of indepen-

dent lexical items into a complex verb. The verb stem determines the newly built form categorically, while the incorporated noun loses its categorical features, as well as any grammatical markings such as plural, if such are employed, see Ex. (3b) below. Nouns do not incorporate as arguments, but as specifiers of the incorporating verb.

(1) (a) Mohawk
/wa?kyvtho? ojì:ja?/ 'I planted a flower'
wa?- aorist modal prefix
-k- 1st person singular agent
-yvtho? 'plant'

(b) Mohawk
/wa?kji?jayvtho?/ 'I flower-planted'
-ji?ja- 'flower'
(Bonvillain 1972, p. 21)

Although the most general features of the construction might at first suggest so, noun incorporation is more than just another variety of compounding. Conspiciously, languages featuring regular and productive compounding, such as German, do not allow for incorporation, and vice versa, incorporating languages do not necessarily allow for compounding.

In many languages, noun incorporation is not a fully productive process, but restricted to certain types of nouns, most often related to the degree of animacy, or it may only be accomplished with suppletive forms. See, e.g., Allen et al. 1984 on Southern Tiwa, Mithun 1984, pp. 863, 876/77.

The incorporating process often entails a loss of referentiality of the incorporated noun and its being backgrounded. From the perspective of the speaker, incorporation represents an option, an alternative way of expression, a strategy to increase the informational content of a proposition, or a focus shift when proceeding with a discourse (cf. Mithun 1984, pp. 869ff. on Mohawk). Thus, the incorporated variant can always be seen in contrast to its nonincorporated counterpart. Being a matter of discourse and style, noun incorporation is expected to be a nonobligatory process. As has been pointed out frequently, it represents an elaborate register and, with bilingual speakers, characteristically is lost with decreasing fluency and competence (see, e.g., Mithun 1984, pp. 879ff.).

It is just this optionality, made possible by the basic lexical independence of the constituents which has motivated the interpretation of incorporation as a syntactic process, as a grammatical function-changing operation (Baker 1988). Incorporation is described as causing the loss of the argument status of the incorporated noun via its vacating its original position in syntax. Such manipulation involves the internal argument (direct object) predominantly.

If at all possible, the incorporation of an external argument typically is restricted to nonagentive noun phrases (Sapir 1989/1911, pp. 252, 266, Mithun 1984, p. 853, Gerdts 1998, p. 87, Cook and Wilhelm 1998).

(2) (a) Chukchee
ŋejə-k ʔəlʔəl təlgə-gʔi
hill-LOC snow.ABS thaw-3SG.S

(b) Chukchee
ŋejə-k ʔələ=lgə-gʔi
hill-LOC snow=thaw-3SG.S
'The snow thawed on the hill'
(Spencer 1995; p. 451)

In its weakest form, noun incorporation simply causes detransitivization, as in examples (1a) and (1b). In its most dramatic version, phrasal constituency is dissolved and the constituents of the former noun phrase become distributed inside and outside the complex verb—the head of the noun phrase is incorporated, while its modifier(s) remain in their external position. This phenomenon is often refered to as 'external modification' or 'classifying incorporation.'

(3) (a) Southern Tiwa
Wisi seuanin bi-mũ-ban
two man:pl 1sB-see-past
'I saw two men'

(b) Southern Tiwa
Wisi bi-seuan-mũ-ban
two 1s:B-man-see-past
'I saw two men'
(Allen et al. 1984, p. 295)

2. *Argument Incorporation*

Baker (1996) extends his syntactic approach of noun incorporation to languages which realize their arguments exclusively verb internally, i.e., which do not exhibit arguments in syntax. Contrary to such an attempt, it must be kept in mind that it is exactly the distinction between syntax and morphology, and the interaction between these two levels, on which the specificity of noun incorporation as compared to other function changing processes, e.g., passives, rests.

The question as to the status of the incorporated noun, i.e., whether it originates as a syntactic argument or not, functions as a crucial and effective criterion for distinguishing noun incorporation from the other types of incorporation. If we consider the implications of the traditional notion 'grammatical,' it is appropriate to distinguish noun incorporation from what was initially called 'pronominal incorporation' by Sapir. Contrary to what the term might suggest, usually no pronoun proper is incorporated. The morphemes under consideration most often cannot even be reconstructed as historically independent forms, but are strictly bound morphemes, obligatory and paradigmatically ordered (Sapir 1989/1911, pp. 250–2). Following Evans (1999) it can also be assumed that their semantic properties are different from those of true pronouns. The more recent term 'pronominal argument' (Jelinek 1984) draws on the relation to syntax again: the arguments of the verb are fully and exclusively realized by bound morphemes, i.e., by

morphological marking. In contrast to the optionality of noun incorporation, argument incorporation does not envisage such a choice. Since the arguments of the verb are realized as bound morphemes joined to the verbal root or stem, no lexical arguments are necessary or even possible outside the verb. The expression of grammatical relations is thus shifted from the syntactic level to the word level, as is illustrated by all examples to come.

As opposed to noun incorporation then, 'pronominal,' i.e., morphological, arguments are not potential syntactic arguments but represent a fundamentally different strategy for the realization of arguments. They fulfill the requirement of representing the fundamental grammatical relations of the sentence, and stand in contrast to inflection, which only reflects the relations constituted on the syntactic level. In order to decide whether a person-marking morpheme has mere inflectional status or whether it actually represents the argument(s) of the verb, careful investigation is needed to decide whether syntactic argument positions are available in a particular language. Such investigations have been carried out only for comparatively few languages, such as the Australian language Warlpiri (Hale 1983, Jelinek 1984, Simpson 1991, see also Austin and Bresnan 1996), the Iroquoian language Mohawk (Baker 1996), Salishan languages (Jelinek and Demers 1994, Jelinek 1995) the Athabaskan language Navajo (Jelinek 2000) and Inuktitut (Eskimo) (Nowak 1996a, 1996b). Discussions of Algonquian languages seem to point in a similar direction, but do not yet cover a broad range of relevant tests. Other languages suggest themselves as promising candidates, but have not been scrutinized, as is, for example, the case with most of the languages discussed by Mithun (1984).

Many morphological argument languages tend to feature a wealth of lexically and/or functionally bound morphemes which have no independent lexical counterparts, and the proportion of roots and stems to bound morphemes is dramatically changed in favor of the bound morphemes (Nowak 2001).

In Inuktitut (Eskimo), whole semantic fields denoting activities, states, events, processes, including 'being' and 'having,' as well as the indication of reported speech and epistemic specifications, are expressed exclusively by bound morphemes which are suffixed to the stem they modify or supplement.

(4) Inuktitut
qimmiqaqtunga
qimmiq-qaq–tunga
dog-have.itr-1s.itr 'I have a dog/dogs.'

(5) Inuktitut
nirijumajara
niri–juma–jara
eat–want–1s.3s.tr. 'I want to eat it.'

In the Salishan languages of the American Northwest Coast, bound morphemes cover the domain of referential expressions and do so exclusively; so-called 'field suffixes' such as -éné, -né, 'ear' in Ex. (6) lack independent counterparts.

(6) Kalispel
na·céné he gets an ear-ring
(Vogt 1940, p. 53)

In cases where items with lexical content are incorporated, no matter whether they are free or bound morphemes, typically the morphological argument and the incorporated lexical items do not exhibit agreement. Yet some cases of such internal agreement relations have been reported. For Southern Tiwa, obligatory incorporation is reported. In this language, person marking on the complex verb also reflects the person, number, and class of the incorporated noun. (Allen et al. 1984, pp. 295–8). The same seems to be the case with Bininj Gun-wok, a language of North Australia.

(7) Bininj Gun-wok
(a) ba-ginje-ng gun-ganj
3/31-cook-PP IV-meat
'(S)he cooked (the) meat.'
(b) Ba-ganj-ginje-ng
3/31-meat-cook-PP
'(S)he cooked (the) meat.'
(Evans 1999, p. 259)

3. *Grammatical Incorporation*

Grammatical relations may also be realized verb internally. Fundamental syntactic operators such as auxiliaries, pronouns, adpositions, complementizers, and other syntactic function words often do not exist at the word level at all.

(8) Inuktitut
(a) niuvirvingmuaqtutit
niuvirvik -muaq–tutit
shop-go.to–2s.itr
'You go to the shop.'

The term *niuvirvik* 'shop' is a lexicalized affix combination, as is *-muaq-* 'go to'; *niuvirvik* is composed of *niuviq-*, 'to acquire' and a nominalizing affix *-vik* 'place where s.th. happens'; *-muaq-* is a highly frequent combination of the (singular) terminalis case *-mut* 'to' and *-aq-* 'go.' Besides these lexicalizations, all affixes are perfectly productive (Nowak 1996a, p. 255).

(8)
(b) niuvirvingmut
niuviq–vik-mut
acquire-place.where-term.s.
'to the shop'

Interestingly enough, the noun marked by the terminalis is not stripped of its plural when involved in such an incorporation.

(8)
(c) niuvirvingnut
niuviq–vik-nut
acquire-place.where-term.p.

'to the shops'

(d) niuvirvingnuaqtutit
niuvirvik -nut -aq- -tutit
shop-term.p-go–2s.itr
'you go to the shops'

Morphological arguments in many instances are fused with other features such as the indication of clause linkage, often referred to as 'switch reference,' or 'anaphoric coreference,' as can be seen in (9).

(9) Inuktitut

tikiniraqtauvutit	piqsiqtillugu
tiki(t)–niraqtau–vutit	piqsiq -tillugu
arrive-say.that.itr–2s.itr	blizzard -4s.vpart.itr

'somebody said that you have arrived while it was blizzarding'

Grammatical incorporation thus can be seen as the morphological realization of functional relators, i.e., as the word internal establishment of relationships which would otherwise demand syntactic spell out. The results are highly complex forms equivalent to full propositions.

(10) Woods Cree
sipwîhcîhkâsow
away = by-foot\AI = pretend\AI = 3I
'he pretends to leave' (Starks 1992, p. 127)

(11) Warlpiri
Pingka-rlipa mata-ma-ninja-kujaku
slow-1pis tired-CAUS-INF-PURP-NEG ya-ni
go-NPST
'We'll go slowly so that it (walking a long distance) doesn't tire us out.'
(Simpson 1991, p. 108)

4. Conclusion

It must be kept in mind that languages hardly ever represent ideal types. This applies to what has been said about syntax and morphology, and the degree to which they interact. The types of incorporation discussed here exemplify an increasing complexity, or configurationality, in morphology, illustrating a unidirectional relationship between morphological complexity and the morphological realization of arguments. Languages of high morphological complexity also feature the morphological realization of arguments, but it is not necessary that all languages realizing their arguments morphologically must be morphologically complex in other respects.

The morphological strategies employed by such languages range from concatenative (Eskimo) through template (Athabaskan) to morphophonemic and prosodic devices (Salishan), and often exhibit mixtures of all these types. A unified account of such complex and multilayered morphological patterns is a desideratum that must be left for future research.

Noun incorporation on the one side, and argument incorporation and grammatical incorporation on the other, are separated by the optionality vs. obligatoriness of incorporation: with the latter, the speakers do not have a choice. The last type, grammatical incorporation, extends morphological configurations to other than argument relations, i.e., to grammatical relations in general. This is the content of the traditional term 'polysynthetic.'

See also: Grammatical Agreement; Grammatical Voice; Morphological Case in Linguistics; Morphology in Linguistics; Morphophonology; Valency and Argument Structure in Syntax; Word, Linguistics of

Bibliography

Allen B, Gardiner D, Frantz D 1984 Noun incorporation in Southern Tiwa. *International Journal of American Linguistics* **50**: 292–311

Austin P, Bresnan J 1996 Non-configurationality in Australian Aboriginal languages. *Natural Language and Linguistic Theory* **14**(2): 215–68

Baker M 1988 *Incorporation*. University of Chicago Press, Chicago

Baker M 1996 *The Polysynthesis Parameter*. Oxford University Press, Oxford, UK

Bonvillain N 1972 Noun incorporation in Mohawk. In: Forster M (ed.) *Papers in Linguistics from the 1972 Conference on Iroquoian Research*. Mercury Series, Ethnology Division Paper No. 10. National Museum of Man, Ottawa, pp. 18–26

Cook E D, Wilhelm A 1998 Noun incorporation. New evidence from Athabaskan. *Studies in Language* **22**(1): 49–81

Evans N 1999 Why argument affixes in polysynthetic languages are not pronouns: Evidence from Bininj Gun-wok. *Sprachtypologie und Universalienforschung* **52**: 255–81

Gerdts D 1998 Incorporation. In: Spencer A, Zwicky A (eds.) *The Handbook of Morphology*. Blackwell, Oxford, UK, pp. 84–100

Hale K 1983 Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory* **1**: 5–47

Jelinek E 1984 Empty categories, case, and configurationality. *Natural Language and Linguistic Theory* **2**: 39–76

Jelinek E 1995 Quantification in Straits Salish. In: Bach E, Jelinek E, Kratzer A, Partee B H (eds.) *Quantification in Natural Languages*. Kluwer, Dordrecht, The Netherlands, pp. 487–540

Jelinek E 2000 Navajo as a discourse configurational language. In: Fernald T, Platero P (eds.) *The Athabaskan Languages: Perspectives on a Native American Language Family*. Oxford University Press, Oxford, UK

Jelinek E, Demers R 1994 Predicates and pronominal arguments in Straits Salish. *Language* **70**: 697–736

Kroeber A 1909 Noun incorporation in American languages. In: Heger F (ed.) *Internationaler Amerikanisten-Kongress XVI*. Hartleben, Vienna, Austria, pp. 569–76

Mithun M 1984 The evolution of noun incorporation. *Language* **60**: 847–94

Nowak E 1996a *Transforming the Images. Transitivity and Ergativity in Inuktitut (Eskimo)*. Mouton de Gruyter, Berlin

Nowak E 1996b Über die organisation komplexer Aussagen. In: Harras G, Bierwisch M (eds.) *Wenn die Semantik arbeitet. Festschrift für Klaus Baumgärtner*. Niemeyer, Tübingen, Germany, pp. 29–56

Nowak E 2001 Lexical categories in polysynthetic languages. In: Cruse A, Hundsnurcher F, Job M, Lutzeier P R (eds.) *Lexikologie-Lexicology. Handbücher zur Sprache und Kommunikation*. Mouton de Gruyter, Berlin
Sapir E 1989/1911 The problem of noun incorporation in American Languages. In: Bright W (ed.) *The Collected Works of Edward Sapir*. Mouton de Gruyter, Berlin, Vol. 5, pp. 27–60
Simpson J 1991 *Warlpiri Morpho-Syntax. A Lexicalist Approach*. Kluwer, Dordrecht, The Netherlands
Spencer A 1995 Incorporation in Chukchee. *Language* **71**(3): 439–89
Starks D 1992 Aspects of Woods Cree Syntax. Ph.D. thesis, University of Manitoba, Winnipeg, Canada
Vogt H 1940 *The Kalispel Language*. Det Norske Videnskaps Akademi i Oslo, Oslo, Norway

E. Nowak

Linguistics: Information Structure

The fundamentally interactive character of natural language, anchored in communication, is reflected by the impact of context (consituation) on sentence structure. Typically, the sentence is divided not only into syntactic subject, predicate, objects, adverbials, etc., but also into what corresponds to cognitively 'given' and 'new' information. This 'information structure' of the sentence (its 'topic-focus articulation,' TFA, or its division into 'theme' and 'rheme' or 'comment') has been established in linguistic writings of the latter half of the twentieth century as one of the constitutive aspects of grammatical structure, rather than just as an issue of the use of language in communication that would be determined only by contextual conditions of the distribution of sentences.

1. *Topic-focus Articulation, Aboutness and Presupposition*

TFA can be illustrated by the following examples (in which capitals indicate intonation center (IC, main sentence stress, falling), cf. Sect. 4):

I was writing this study on the WEEKENDS. (1a)
On the weekends I was writing this STUDY. (1b)
Yesterday John visited BOSTON. (2a)
JOHN visited Boston yesterday. (2b)
John visited Boston YESTERDAY. (2c)
Jim's brother did not insult MARY. (3a)
Mary was not insulted by Jim's BROTHER. (3b)

As (1) documents, the opposition of focus (F, which carries the IC and in the primary case is placed at the end of the sentence) and topic (T, referring to 'given' information) is relevant for the semantic interpretation of the sentence (even for its truth conditions: there are situations, or 'possible worlds' in which (1b) can be used as a true statement, but (1a) would not be true). Thus, although the information structure (TFA) is closely connected with contextual conditions of the use of sentences (e.g., as answers to different questions), it belongs to the information expressed by grammatical means (sentence prosody, word order (WO), and in languages such as Japanese or Tagalog by specific particles) and relevant for the linguistic meaning of the sentence. The WO shift in (1) can be understood as primary, prototypically accompanied by passivization in English, cf. (3) although it occurs also with the active verb in languages with a higher degree of 'free' WO, cf. e.g., Czech, Marii (Accus.) *urazil Jimův BRATR* (Nomin.). In English, a shift of IC is more or less equivalent: 'Jim's BROTHER insulted Mary.'

In declarative sentences TFA corresponds to the 'aboutness' relation in the sense of F being asserted about T. Thus, in (2a) it is asserted about John (x) and yesterday (y) that what x did at timepoint y was to visit Boston. In (2b) it is asserted about yesterday's visit to Boston that it was carried out by John, and (2c) asserts about John (x) and Boston (z) that x visited z precisely yesterday (or, with the verb being included in T, it asserts about John's visit to Boston that it occurred yesterday).

A positive sentence can be interpreted as asserting that F holds about T, i.e., as F(T). The prototypical negative sentences, such as those in (3), are used to assert that F(T) does not hold, which means that what is under negation is identical to F (the expression 'under negation' will be made more precise in Sect. 3); T then is not under negation and a typical definite noun group in T triggers a presupposition that is not present if this group is included in F, as in (3b): while both (3a) and its positive counterpart entail the existence of Jim's brother, this is not entailed by (3b), in which the entailment can be classified as an allegation, rather than a presupposition or a case of 'meaning proper.' It is not exact to speak of a presupposition of existence; in certain cases, especially in sentences directly concerning existence (in the actual world), what is presupposed is not the existence, but rather the referential availability of the object concerned (an abstract object, a mythical being, etc.), or its presence in the part of the universe of discourse that is assumed by the speaker to be accessible to the hearer(s), cf. 'The gifts are from your parents,' 'Santa Claus does not exist,' or also 'White crows would have red eyes, if they existed.'

It is a specific question under which conditions a difference in TFA has a direct impact on the truth conditions of sentences. A similar situation is known from predicate logic, in which, e.g., the order of 'for every x' and 'there exists a y' is relevant, although the order of two variables bound by the existential quantifier is not relevant ('Every child passed an EXAM' is weaker than 'An exam was passed by every CHILD,' but 'A child passed an EXAM' and 'An exam was passed by a CHILD' share their truth conditions.)

It can be argued that there is a difference between contrastive and non-contrastive F; while the sentences in (1)–(3) can occur as typical answers to questions such as 'When were you writing this study?' 'What were you doing on the weekends?' 'Which town did John visit yesterday?,' etc., the cleft construction in (4) would preferably be used in case of contrast, e.g., after 'I am not sure which of your cousins was in Boston yesterday.'

It was JOHN who visited Boston yesterday. (4)

The concept of contrast (based on choice among alternatives) is also relevant for T, as discussed in Sect. 4 below.

2. *Different Traditions*

In the light of the older European tradition, the pioneering insights of H. Weil, G. von der Gabelentz, Ph. Wegener, and others (cf. also the notion of 'logical stress' in Russian language teaching), V. Mathesius introduced an analysis of TFA into structural linguistics. The sentence has been analyzed as divided into a contextually bound and a non-bound part, i.e., into T and F; while the former can be expressed by weak, unstressed or zero pronominal elements, the latter is stressed. This and similar views have then been elaborated especially by Halliday (1967–8), by J. Firbas (1992), as well as by the research group of theoretical and computational linguistics in Prague (Sgall et al. 1986), by D. Bolinger, W. Chafe, T. Reinhart, then (with the division of T or 'ground' into 'link' and 'tail') by Vallduví and Engdahl (1996), and, in Chomskian linguistics, by R. Jackendoff, M. Rochemont, and others. Recently, centering theory (A. Joshi, B. Grosz, E. Prince) attempts to determine which sentence part probably will occur as (or in) the T of the next utterance.

Some of the approaches based on constituency or configurationality face problems connected with the fact that neither T nor F is, in the general case, a single constituent, as can be illustrated by (5) and (6), where the sign '@' indicates the T/F (F/T) boundary:

(How have you spent the vacations?) We went @ for a week to the ADRIATIC (rather than for the whole month to Mallorca, as we had planned). (5)

(What happened to Mary in the valley?) A WASP bit @ her. (6)

It is then necessary to work either with specific rules of 'focus inheritance' (adding more constituents to F, see Selkirk 1995), or with multiple derivations such as those of Steedman's (1996) combinatory categorial grammar, if the authors do not want to multiply the articulations, differentiating, e.g., between an opposition of background vs. focus, and that of topic vs. comment. It may be seen as more advantageous to use dependency syntax (as known from European structural linguistics) and an elaboration of certain notions originally introduced by J. Firbas: as discussed in Hajičová et al. (1998) and in the writings quoted there (a formal treatment was presented by Petkevič 1995), the scale of communicative dynamism (CD) can be handled as an underlying order, which differs from the surface or morphemic order especially in sentences in which IC (carried by the most dynamic item) occupies a secondary position (not at the end of the sentence), as is the case in (2b), (4), and (6) above; contextual boundness (as a grammaticalized feature of every lexical occurrence) can be understood as a primitive notion. This allows for a subtle classification of notions such as 'broad' and 'narrow' F; the T/F boundary can then be specified as determined by a single edge of the dependency tree (which constitutes the core of the sentence structure).

On the other hand, there is a tradition started by J. Jacobs (1983) and continued by M. Krifka, M. Rooth, K. von Fintel, D. Büring, and others, whose approaches are based on the analysis of the positions and functions of focus sensitive particles (such as 'only,' 'also,' 'even'). The foci and backgrounds of the particles and their contextually conditioned properties are systematically studied by methods of formal semantics. An attempt at a synthesis of the two approaches, offered by B. H. Partee and her co-authors (Hajičová et al. 1998), is summarized in the next section.

3. *The Focus of the Sentence and the Focus of a Particle*

As example (3) above illustrates, what is under negation in the primary case is F. In secondary cases, (a) the negation itself, together with the verb, belongs to T, and only the verb is negated, cf. (7), or (b) negation constitutes the whole F, and (a part of) T is under negation, cf. (8):

My brother didn't come to the party since he was ILL. (7)

(e.g., after 'Why didn't your brother come to the party?')

My brother DIDN'T come to the party. (8)

Jane told me that yesterday she did not SUCCEED. (9)

In the secondary cases, and also if negation is present in an embedded clause, not only the articulation of the sentence into F and T is relevant, but also the focus of the negation particle and its scope: in (7) only T, i.e., the main clause, is the scope of negation and 'come to the party' is its focus (i.e., is under negation); in (9) only the 'that'-clause is the scope and its verb is the focus of negation. It is then possible to define the background of the particle as its scope minus its focus.

This concerns also other focus sensitive particles or operators (focalizers), as can be illustrated by examples from German, the WO of which is more perspicuous than that of English. Prototypically, the focalizer is placed at the T/F boundary:

Er ist nicht/auch/eben/sogar/nur/mit großer Freude/nach PRAG gekommen. (10)
[He – has – not/also/precisely/even/only/with – great – pleasure – to PRAGUE – come.]

Also in the secondary cases, other focalizers occur in a position similar to that of negation: the focalizer belongs to T in (11) and constitutes the whole F in (12):

Er ist wegen seiner FRAU nicht/auch/…gekommen.
[He has due – to his WIFE not/also/…come.] (11)

Dein Bruder war dort NICHT/AUCH …
[Your brother was there NOT/TOO/…] (12)

In English, WO is more ambiguous, e.g., in (13) F has several possible ranges (the positions of the T/F boundary on different readings are again marked by '@'):

Martin only @ gave @ a flower @ to a GIRL. (13)

It follows that it is necessary to distinguish between (a) the sentence focus F, typically placed after the topic, marked by IC and empirically specifiable with the use of operational criteria such as the question test or a test with a negative or adversative continuation, and (b) the focus of a focalizer (ff). In the unmarked case ff is identical with F (cf. (3), (10) and (13) above), but this is not the case in secondary examples (such as (11) and (12)).

If negation is one of the focalizers, then in the absence of an overt focalizer the positive modality of the verb can be interpreted as a focalizer (see Jacobs 1984), which is not overtly present in the sentence, but is relevant for its semantico-pragmatic interpretation (i.e., a procedure leading from underlying representations to an account of the content of the expressed statements, questions, etc.). Two studies on issues of the interpretation of TFA by means of post-Montague semantics—B. H. Partee's investigation of the relationship of 'allegation' to accommodation and J. Peregrin's elaboration of the aboutness relation as that of an argument and a predicate—are included in Partee and Sgall (1997).

4. *Means of Expression of Information Structure*

In the prototypical case, the (surface) WO corresponds to the scale of CD: T proper (the least dynamic item) occupies the leftmost position, and F proper (most dynamic) comes last. In sentences with a marked position of IC, this prosodic feature overshadows WO; F proper is expressed by the bearer of IC, cf. J. Pierrehumbert's H*L accent and also the analyses by Ladd (1996), who outlines how TFA varies from one language to another, and Steedman (1996).

Differing both from IC and from lexical stress of individual words, there are other kinds of stress, which are to be classified and specified as to their positions in several hierarchies. Some of these kinds of stress (phrasal, contrastive, or the rising interrogative contour) are immediately relevant for sentence structure, whereas others express different ('pragmatic,' attitudinal) values, which do not directly belong to the system of language (various kinds of emphasis, wonder, aversion, doubts, and so on).

T proper often has a contrastive value (i.e., refers to an entity chosen from a set of alternatives determined by the context); this is expressed by a certain kind of phrasal stress, denoted here by italics:

Paul's *son* has bought a new CAR and his *daughter* has got an APARTMENT of her own. (14)

It is often assumed that in sentences such as (14) a rising contrastive stress on the subject is present, which differs from IC. Example (14) also shows that compound sentences include separate TFA patterns in their individual clauses.

If a focalizer occurs within T, its ff (perhaps always a contrastive part of T) is marked (according to C. Bartels and others) by a specific kind of stress (which perhaps may be identified with the just mentioned contrastive stress), or at least by some other kind of rhythmic prominence, see Rooth (1992). This concerns also the ambiguous position of 'only' in his example reproduced here with a modification concerning the phrasal stress (cf. the discussion in Hajičová et al. 1998, p. 157) as (15a); its preferred reading is identical with that of (15b), with 'only' in T:

People who *grow* rice often only EAT rice. (15a)
People who *grow* rice often EAT only rice. (15b)

5. *Three Layers of Word Order*

Let us have a look at the following examples (with which the symbols for dependents such as Directional or Locative should be self-explaining):

They flew from Chicago to Boston. (Dir. 1–Dir. 2) (16a)
A B
They flew to Boston from Chicago. (16b)
B A
Jim dug a ditch with a hoe. (Objective–Means) (17a)
A B
Jim dug a DITCH with a hoe. (17b)
A B
Ron cannot sleep quietly in a hotel. (Manner–Loc) (18a)
A B
In a hotel Ron cannot sleep quietly. (18b)
B A

If A precedes B in the examples with normal intonation, then A belongs to T on some readings, to F on others; however, if B precedes A (or if A, preceding B, bears the IC), then B belongs to T. A similar relationship can be found with other pairs of arguments and adjuncts (with certain differences concerning groups of verbs, dialects, or generations), so that a basic scale can be established that determines

their CD, with deviations possible only in case B is contextually bound (belongs to T); the differences between CD and WO (such as in (17b), or those concerning the position of the verb in German, etc.) can be handled by low-level rules. The existence and the basic shape of this 'systemic ordering,' known since the 1970s, has been confirmed by several series of tests for Czech, German and to a certain extent for English, see Sgall et al. (1995).

Systemic ordering in English—a preliminary partial scale:

Temp–Actor–Addr–Obj–Origin–Effect–Manner–Dir.1–Means–Dir. 2–Loc

This ordering can be specified in the lexical entries of verbs (and other head words), especially by the order of the obligatory and optional kinds of dependents in the valency frames. Another layer of order of the lexical items in the sentence is the CD (which within T may differ from the systemic ordering), and the third layer is WO.

It can be concluded that at least the core of information structure belongs to grammar and can be described within frameworks with flat patterns, such as dependency syntax; these representations may serve as input for an interpretation using methods of formal semantics. This holds for languages with the most varying degrees of 'free' WO. In most Slavic languages, Latin, etc., in which the basic syntactic relations are expressed by case endings, WO often is 'free' to correspond to CD. English or French, which express these relations by WO rather than by endings, use more syntactic reshaping (passivization, clefting, secondary placement of IC) to achieve such a correspondence between WO and CD.

See also: Communication and Social Psychology; Communicative Competence: Linguistic Aspects; Information Processing and Psychopathology; Information Processing Architectures: Fundamental Issues; Information Theory; Word Order

Bibliography

Firbas J 1992 *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge University Press, Cambridge, UK

Hajičová E, Partee B H, Sgall P 1998 *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer, Dordrecht, Netherlands

Halliday M A K 1967–8 Notes on transitivity and theme in English. *Journal of Linguistics* **3**: 37–81, 199–244; **4** (1968): 179–215

Jacobs J 1983 *Fokus und Skalen: Zur Syntax und Semantik von Gradpartikeln im Deutschen* [*Focus and Scales: On the Syntax and Semantics of Grading Particles in German*]. Niemeyer, Tübingen, Germany

Jacobs J 1984 Funktionale Satzperspektive und Illokutionssemantik [Information Structure and Semantics of Illocution]. *Linguistische Berichte* **91**: 25–58

Krifka M 1994 Focus and operator scope in German. In: Bosch P, van der Sandt R (eds.) *Focus and Natural Language Processing*. IBM Working Paper 7. IBM Deutschland, Heidelberg, Germany, pp. 133–55

Ladd D R 1997 *Intonational Phonology*. Cambridge Studies in Linguistics 79. Cambridge University Press, Cambridge, UK

Partee B H, Sgall P (eds.) 1997 *Discourse and Meaning. Papers in Honor of Eva Hajičová*. Benjamins, Amsterdam

Petkevič V 1995 A New Formal Specification of Underlying Structures. *Theoretical Linguistics* **21**: 7–61

Pierrehumbert J, Hirschberg J 1990 The meaning of intonational contours in the interpretation of discourse. In: Cohen P, Morgan J, Pollock M (eds.) *Intentions in Communication*. MIT Press, Cambridge, MA, pp. 271–312

Reinhart T 1981 Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* **27**(1): 53–94

Rooth M 1992 A Theory of focus interpretation. *Natural Language Semantics* **1**: 75–116

Selkirk E 1995 Sentence prosody: Intonation, stress and phrasing. In: Goldsmith J A (ed.) *Handbook of Phonological Theory*. Blackwell, London, pp. 550–69

Sgall P, Hajičová E, Panevová J 1986 *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, Netherlands

Sgall P, Pfeiffer O, Dressler W U, Půček M 1995 Experimental research on systemic ordering. *Theoretical Linguistics* **21**: 197–239

Steedman M 1996 *Surface Structure and Interpretation*. MIT Press, Cambridge, MA

Vallduví E, Engdahl E 1996 The linguistic realization of information packaging and grammar. *Linguistics* **34**(3): 459–519

P. Sgall

Linguistics: Internal Reconstruction

Internal reconstruction is a method for establishing earlier, unattested forms of languages without reference to 'external,' especially comparative, evidence from other languages. In this it differs from the comparative method (see *Linguistics: Comparative Method*), which is based on such evidence, though the aims of both methods are in practice the same. To signal the difference, forms or languages reconstructed by internal reconstruction are designated 'pre-forms,' 'pre-languages,' etc., as opposed to the 'proto-forms' and 'proto-languages' reconstructed by the comparative method. The method can be applied to individual languages in cases where comparative evidence is not available, for example to isolated languages or to reconstructed proto-languages.

Internal reconstruction is an extremely powerful method, and for this reason some of its results can be both speculative and controversial, though a number of the reconstructions arrived at by this method which were earlier deemed too hypothetical have subsequently become accepted, for example the postulation of a set of 'laryngeal' consonants for Indo-European

(the 'laryngeal theory'), whose existence was corroborated following the discovery of Hittite (Lindeman 1987). As in the case of the comparative method, it evolved in practice before it was formally recognized. However, since the method rests on a quite sophisticated formal analysis of language, it did not appear until the late nineteenth century, much later than the comparative method. In a youthful work, for instance, Saussure used aspects of what would now be regarded as internal reconstruction in his study of the Indo-European vowels. However, the method was not firmly established until the 1940s, principally in the work of American structuralist linguists such as Henry Hoenigswald.

Internal reconstruction is not a single, unitary method; the term embraces a number of related but distinct procedures, which we may designate as follows:

(a) historical morphophonemics;
(b) regularization of systems; and
(c) universal and typological reconstruction.

All of these have in common the attempt to work backwards in time by reversing changes that are assumed to have taken place in the language in question, but they differ in scope and in the factors that are taken into account.

1. *Historical Morphophonemics*

This is the most limited of the methods of internal reconstruction. It applies only to the reconstruction of phonology, and works by identifying phonological alternations between grammatically or lexically related forms resulting from the differential operation of phonological changes in different linguistic contexts. By eliminating these alternations and reducing the set of alternants to a single form, it effectively reverses the changes and establishes the form that existed prior to their taking effect. In practice, it is equivalent to the synchronic process of establishing 'morphophonemes,' or the 'underlying forms' of classical Generative Phonology (see *Phonology*; *Morphophonology*)—abstract phonological entities from which the occurring forms can be derived—though it interprets these morphophonemes as historical antecedents.

Consider, for example, the Latin forms given in (1).

Nominative	Genitive		
consul	consulis	('consul')	
genus	generis	('kind')	(1)
corpus	corporis	('body')	

As we see from *consul*, the genitive is formed by adding the affix -*is* to the stem of the word, but in the case of *genus* and *corpus* the final consonant of the stem is -*s* in the nominative and -*r* in the genitive. In synchronic phonology we might regard these two sounds as realizations of the same element, a 'morphophoneme' which appears as /s/ at the end of the word and as /r/ between vowels. But the alternation can also be interpreted historically, if we assume that this morphophoneme was actually an original phoneme, assumed to be /s/, which was changed to /r/ between vowels. We can therefore postulate earlier forms of the genitive: **genesis* and **corposis*.

This method works, therefore, by attempting to undo the effects of the splitting of phonemes in different contexts. Here, the original phoneme /s/ split into two separate phonemes, /s/ and /r/, depending on the context in which it occurred. This split resulted in a synchronic alternation between the phonemes /s/ and /r/ in different forms of the word. By identifying such alternations, and deriving the alternating phonemes from a single source, we reconstruct the earlier form of the words.

2. *Regularization of Systems*

Like historical morphophonemics, this method attempts to reverse changes by eliminating anomalies, but it extends this principle to cover whole systems of linguistic elements.

Consider, for example, the so-called 'strong' verbs of the Germanic languages. These verbs form their principal parts not by the addition of affixes (*walk–walked*; *love–loved*, etc.), but by modifying the vowel of the stem (*sing–sang–sung*; *ride–rode–ridden*, etc.). Strong verbs fall into a number of classes, according to the pattern of alternation of the vowels; only the first three are relevant here. For these verbs, we can identify three stems, whose vowels can be reconstructed by means of the comparative method, for the Germanic languages as a whole: as in (2).

Class I	i:	ai	i	
Class II	iu	au	u	(2)
Class III	i/e	a	u	

It can be seen that there is a certain regularity here, though also some anomalies. By the simple expedient of regarding the *i*: of the first stem of Class I (the colon indicates length) as *ii*, the first two classes become entirely parallel, with the vowel pattern *i–a–zero*, where the vowel is followed by *i* in Class I and *u* in Class II; in the third stem, the *i* or *u* is left as the only vowel. In Class III verbs, the vowel is followed by a nasal or liquid consonant (*r* or *l*), and the *i*/*e* alternation in the first stem depends on the following consonant: *i* occurs before a nasal and *e* before other consonants. We can therefore eliminate this alternation; the standard reconstruction assumes an *e* here, which is also generalized to the first stem of Classes I and II (*e* became *i* before the *i* and *u* of these classes as well as before the nasal of Class III). Thus, all three classes can be reconstructed with the same pattern (*e–a*) for the first two stems.

There remains the third stem, for which, as we have seen, we assume a 'zero vowel' for Classes I and II. If the pattern is the same for all three classes, Class III

should have no vowel at all here, though in fact we find *u*. This *u* thus alternates with the *zero* of Classes I and II, and we can reconstruct a single source for both, which, following the first two classes, is taken to be *zero*. The nasal or liquid consonant of the stem of this class is assumed to have been 'syllabic' (i.e., it had the role of a vowel, as with the *n* or *l* of such English words as *bitten* [bitn] or *bottle* [botl]). It is presumed that these syllabic consonants later developed *u* before them. The result of these procedures is an entirely regular system, with the pattern *e–a–zero*, the differences among these three classes being in the sound which followed the vowel: *i* in Class I, *u* in Class II, and a nasal or liquid consonant in Class III. Such a regularized system can then be regarded as existing at an earlier stage of the language.

It is legitimate to object, of course, that, carried to its logical conclusion, this procedure results in the elimination of all anomalies and irregularities, and thus assumes that earlier languages had no such irregularities—an unlikely state of affairs. However, it should be borne in mind that all reconstruction, by whatever method, produces an idealization rather than an attested reality.

3. *Universal and Typological Reconstruction*

This method attempts to work backwards by invoking known or assumed general attributes of language or processes of linguistic change. This is the most powerful, but also the most speculative form of internal reconstruction, since the evidence we have of such general processes of change is often of a theoretical kind.

The criteria for reconstruction here may be either universal or typological. In the former case, changes are assumed to have occurred which reflect universals, i.e., characteristics of all languages, though these may be diachronic (universal processes of change) or synchronic (universal characteristics of language states which constrain the processes of change). If we therefore encounter a particular feature in the language under investigation which is known to be the result of a certain universal process, then we may assume this process to have occurred and reinstate the form from which the feature may have derived.

Universal processes or constraints have been recognized for all components of linguistic structure: phonology, morphology, syntax, and the vocabulary. One such process, which is currently receiving considerable attention, is grammaticalization (see *Grammaticalization*), the 'downgrading' of lexical words to grammatical words and affixes (Hopper and Traugott 1993); a similar and related process applies in the lexicon, by means of which words are progressively 'bleached' of semantic content. Recognition of such universals allows us to determine the direction of language change, and thus to reconstruct earlier stages in conformity with this direction. Thus, if we assume that all grammatical affixes are derived from original lexical words, then a reconstruction which posits grammatical morphemes as earlier stages of lexical morphemes rather than the reverse would be ruled out.

Reconstruction based on typological criteria relies on the observation that languages fall into a number of more or less distinct types, characterized by co-occurring or harmonizing features (see *Linguistic Typology*). A standard typological grouping, for example, divides languages into VO and OV types, according to whether the object (O) follows or precedes the verb (V). A number of other features of sentence structure harmonize with this: in VO languages (e.g., French) the adjective typically follows the noun, and prepositions are used; in OV languages (e.g., Japanese) the adjective typically precedes the noun, and postpositions are used.

To apply the method, we identify the language as belonging to a particular type on the basis of some major feature, and, making the further assumption that languages are ideally typologically consistent, bring the remaining harmonizing features into line. If, for example, we find a language (such as English) which is VO but has the order adjective + noun, we will assume that at an earlier stage the language will have either been OV (which is assumed for English) or will have had the reverse order of noun and adjective. Various different typological parameters have been put forward, but their use in reconstruction generally follows the above procedure. Again there are potential weaknesses in the assumptions underlying the method, since most languages are typologically inconsistent to some degree, and it is by no means proven that typological consistency represents an ideal or original state for languages.

4. *Conclusion*

The methods of internal reconstruction are extremely powerful, but they rest on a range of assumptions about both the nature of languages and the nature and direction of language change. The accuracy and reliability of the results inevitably depend on the adequacy of these assumptions. Not surprisingly, the results of the method have often been challenged as the assumptions themselves have been called into question.

See also: Historical Linguistics: Overview; Linguistics: Comparative Method

Bibliography

Campbell L 1998 *Historical Linguistics. An Introduction*. Edinburgh University Press, Edinburgh, Chap. 8

Fox A 1995 *Linguistic Reconstruction. An Introduction to Theory and Method.* Oxford University Press, Oxford, Chaps. 7, 8
Greenberg J H 1963 Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg J H (ed.) *Universals of Language.* MIT, Cambridge, MA
Hock H H 1986 *Principles of Historical Linguistics.* Mouton de Gruyter, Berlin, Chap. 17
Hopper P J, Traugott E C 1993 *Grammaticalization.* Cambridge University Press, Cambridge, UK
Lehmann W P 1992 *Historical Linguistics*, 3rd edn. Routledge, London, Chap. 8
Lindeman F O 1987 *Introduction to the 'Laryngeal Theory.'* Norwegian University Press, Oslo, Norway

A. Fox

Linguistics: Markedness

'Markedness' refers to an asymmetry relation between elements of linguistic or conceptual structure. In a marked–unmarked relation, one term of an opposition, or of a set, is taken to be the dominant one while the other(s) are taken to be secondary. The dominant term is known as 'unmarked' and the secondary one(s) as 'marked.' Linguistic markedness ranges over phonological, grammatical, and semantic oppositions and constructions. The concept has been extended to the analysis of cultural relationships as well, with background cultural categories (right-handedness, sightedness, etc.) being taken as unmarked.

1. *Background in the Prague School*

As Henning Andersen (1989) has noted, the idea of asymmetry in linguistics predated the actual coining of the terms 'marked' and 'unmarked' in the 1930s, going back to work in the early 1800s. The modern concept of markedness originated in the Prague School structuralism of Roman Jakobson and Nikolai Trubetzkoy, and was proposed as a means of characterizing binary oppositions. In Prague School structuralism both sound and meaning were analyzed into systems of binary distinctive features representing sound and meaning as the presence or absence of relational properties or features. Binarism suggests symmetry and equivalence in linguistic analysis; markedness adds the idea of hierarchy. The fact that distinctions were characterized as the presence vs. the absence of some property implied an evaluation of any opposition into a mark and its absence. Thus for example, phonological oppositions such as nasal vs. nonnasal were defined as the presence vs. the absence of nasality; nasality was marked, nonnasality unmarked.

Oppositions manifest asymmetry in several ways, most notably for Trubetzkoy in terms of the diagnostic of neutralization, whereby the unmarked term of a phonological opposition was thought to be that which occurred in positions where the contrast was neutralized (i.e., here only one term of the opposition could occur.) In his 1932 article 'Structure of the Russian Verb,' Jakobson extended the concept of asymmetry to a view of grammatical meaning in which the marked element 'announces the existence of [some meaning] A' while the unmarked element 'does not announce the existence of A, i.e., does not state whether A is present or not.' This meant that an opposition like past vs. nonpast characterized three interpretations: past, present (as the opposite of past), and timelessness (the neutralization of the tense opposition). More generally, any semantic opposition involved three meanings: two unmarked and one marked. The breadth of the unmarked terms was extended by Jakobson to oppositions that were not inherently binary. In a 1972 article, Jakobson described language by saying that 'every single constituent of a linguistic system is built on an opposition of two logical contradictories: the presence of an attribute ('markedness') in contraposition to its absence ('unmarkedness').' Jakobson also suggested a role for markedness in language acquisition and loss, in his monograph *Child Language, Aphasia, and Universals of Language* (1941). Jakobson proposed a relationship between a universal feature hierarchy, the acquisition of sound distinctions by children and the loss of sound distinctions during aphasia. Today few linguists or aphasiologists adopt Jakobson's positions literally, though many view them as useful tendencies (see Caramazza and Zurif 1978).

Semiotically oriented work in the Prague School tradition has viewed markedness as a means to investigate the isomorphism of form and meaning. Henning Andersen and Michael Shapiro have linked markedness to semiotic universals and to the work of Charles S. Peirce. The work of Cornelius van Schooneveld, Edna Andrews, Rodney Sangster, Yishai Tobin, and others foregrounds the notion of 'semantic invariance' (understood as the different general meanings reflected in the contextual specific meanings of features). The most well-known semantic invariants are those in Jakobson's analysis of the Russian case and verb systems which attempt to ground semantics in spatial relations and in the relations among participants in the overall speech event (for recent work on invariance, see Waugh and Rudy 1991).

2. *Markedness in Generative Grammar*

Markedness entered generative linguistic theory through Chomsky and Halle's *The Sound Pattern of English* (1968). Following the Prague School view, phonological features were assumed to be part of a universal phonetic vocabulary. In addition, features

were also incorporated into what was known as the 'evaluation metric,' a central concept in early versions of generative grammar. The evaluation metric was a means of selecting the most highly valued grammar capable of describing the facts of a language. In *The Sound Pattern of English* the value of a grammar was the inverse of the number of features required in that grammar. However, Chomsky and Halle realized that their initial approach to phonological features made implausible rules and segment inventories as highly valued as natural ones. To deal with this, they introduced markedness as a means of encoding the intrinsic content of features. Chomsky and Halle proposed a set of 39 universal marking conventions tying naturalness to evaluation. In their system, an unmarked feature value was cost-free with respect to the evaluation metric, while the marked values were counted by the metric. Segment inventories could also be evaluated according to the number of marked features required to characterize them; sets of lexical items could be evaluated for markedness based on the number of marked features they entailed; and a concept of 'linking' reduced the number of features required to characterize processes corresponding to the markedness conventions. However, the use of phonological markedness as part of the evaluation metric was never able to fully account for the fact that some features are more likely than others or for the fact that phonological system must have a certain minimal complexity and symmetry (see Kean 1980).

In syntax, markedness as feature-evaluation did not receive the same attention that it did in phonology. In Chomsky's work, markedness evolved to a view of unmarked properties as an innate preference structure based first in constraints and later in parameters of universal grammar. Beginning in the 1970s, Chomsky attempted to greatly restrict the class of possible grammars, suggesting a relative rather than an absolute interpretation of conditions on rules. In such an approach, rules apply in accordance with the conditions unless specifically complicated to evade them, with Chomsky (1977) noting in his article 'On *Wh*-movement' that 'The logic of this approach is essentially that of the theory of markedness.' In their 1977 article 'Filters and Control,' Chomsky and Howard Lasnik extended to view markedness as part of a theory of 'core grammar':

> We will assume that [Universal Grammar] in not an 'undifferentiated' system, but rather incorporates something analogous to a 'theory of markedness'. Specifically, there is a theory of core grammar with highly restricted options, limited expressive power, and a few parameters. Systems that fall within core grammar constitute 'the unmarked case'; we may think of them as optimal in terms of the evaluation metric. An actual language is determined by fixing the parameters of core grammar and then adding rules or conditions, using much richer resources, … These added properties of grammars we may think of as the syntactic analogue of irregular verbs. (1977, p. 430)

What is unmarked is the entire core grammar of a language, which is then supplemented by a marked periphery. In later versions of his theories, Chomsky has explored various positions on markedness, for example associating the unmarked with a child's initial hypothesis. In a 1986 work, he summarized his views as follows:

> The distinction between core and periphery leaves us with three notions of markedness: core versus periphery, internal to the core, and internal to the periphery. The second has to do with the way parameters are set in the absence of evidence. As for the third, there are, no doubt, significant regularities even in departures from the core principles (for example, in irregular verb morphology in English), and it may be that peripheral constructions are related to the core in systematic relaxing certain conditions of core grammar. (1986, p. 147)

Other researchers in generative syntax have tried to flesh out aspects of markedness, attempting to operationalize the idea of the cost of a marked syntactic construction or integrate markedness into the theory of parameters. Various approaches have been suggested, including tying markedness to an ordering of parameter values defined by subset relations among sentences, connecting it to types of triggering experiences or to default learning strategies, and using markedness to establish a universal syntactic feature hierarchy. Optimality Theory approaches emerging in the 1990s also incorporate markedness. In such approaches, in which constraints are ranked and candidate outputs may violate them to a greater or lesser degree, markedness is encoded in the constraints and ranking (see Archangeli 1997). Some generative researchers have applied markedness to second language acquisition theory, treating it as an inherent learning hierarchy which reflects the sequence in which constructions are acquired, the difficulty of acquiring certain constructions, and the transferability of rules across languages.

3. *Universals*

Markedness has also been applied to typological and function and approaches to language, influenced by Joseph Greenberg's 1966 book *Language Universals*. Greenberg surveyed the correlates and diagnostics of markedness proposed by Jakobson, Trubetzkoy, and others in an attempt to find parallel diagnostic criteria across levels of language. Noting that unmarked grammatical categories in a language show characteristics correlated to the greater frequency of a category vis-à-vis its opposite, Greenberg took frequency to be the primary determining factor of markedness in grammar and suggested that unmarked categories could be determined by 'the frequency of

association of things in the real world.' In addition to emphasizing the role of frequency in determining markedness within a language, Greenberg applied frequency cross-linguistically, taking unmarked categories to be those that are unmarked in a wide number of languages. However, while frequency is an appealing diagnostic, it is also problematic because a phenomenon may be infrequent cross-linguistically but predominant in a particular language (a much-used example is preposition stranding in English, which appears to be marked cross-linguistically but unmarked in English).

Universals have also been proposed on the basis of implicational laws, with a category taken as generally marked if every language that has the unmarked category also has the marked one but not vice versa. Other approaches to universal markedness relations focus on functional economic and iconic motivations, tying recurring asymmetries to properties of communication channels and communication events. Croft (1990), for example, notes that asymmetries among linguistic elements may be explainable in terms economy of form and in terms of iconism between the structure of language and conceptualization of the world. Similarly, Givón (1990) suggests that markedness related to cognitive complexity—'in terms of attention, mental effort or processing time.' And Mayerthaler, Dressler, and other 'naturalists' view markedness relations in terms of the ways in which extralinguistic principles of perceptibility and psychological efficiency determine what is natural in language. (Mayerthaler (1988), for example, defines unmarked as 'in agreement with the typical attributes of the speaker.')

4. *Cultural Markedness*

A main theme of markedness is the information content and information value of an element. Thus many researchers have seen markedness as an encoding of that which is unusual or informative. Conceptual familiarity with cultural norms provided by familiar categories creates a ground against which marked categories provide a figure, opening the way for markedness to be applied to cultural and social categorization. In the 1930s Jakobson had already suggested applying markedness to any oppositions, explicitly mentioning such pairs as life/death, liberty/bondage, sin/virtue, and holiday/working day. Waugh (1982) extended this idea to note that oppositions like male/female, white/black, sighted/blind, hearing/deaf, heterosexual/homosexual, right/left, fertility/barrenness, clothed/nude, and spoken language/written language can all be treated as unmarked/marked cultural oppositions. And McCawley (1985) suggested that aspects of belief systems and even scientific theorizing could be analyzed as systems in which some assumptions are culturally unmarked. The papers in Myers-Scotton (1998) demonstrates ways in which intentionality is conveyed by adopting stylistically and culturally marked choices. Such approaches hold promise as a descriptive technique for domains beyond language, and recent work has applied markedness to music, myth and facial expression, among other things (see Hatten 1994, Liszka 1989).

As is evident, markedness has been extended and reshaped over the past century and reflects a range of loosely connected theoretical approaches. From emerging in the analysis of binary oppositions, it has become a global semiotic principle, a means of encoding naturalness and language universals, and a terminology for studying defaults and preferences in language acquisition. What connects various approaches is a concern for the evaluation of linguistic structure, though the details of how markedness is determined and what its implications are varies widely.

See also: Generative Grammar; Lexical Functional Grammar; Linguistics: Iconicity; Linguistics: Information Structure; Linguistics: Prototype Theory; Linguistics: Theory of Principles and Parameters; Semiotics

Bibliography

Andersen H 1989 Markedness—The first 150 years. In: Tomic O (ed.) *Markedness in Synchrony and Diachrony*. Mouton de Gruyter, Berlin

Andrews E 1990 *Markedness Theory: The Union of Asymmetry and Semiosis in Language*. Duke University Press, Durham, NC

Archangeli D 1997 Optimality theory: An introduction to linguistics in the 1990s. In: Archangeli D, Langendoen D T (eds.) *Optimality Theory: An Overview*. Blackwell, Malden, MA

Battistella E 1990 *Markedness: The Evaluative Superstructure of Language*. SUNY Press, Albany, NY

Battistella E 1996 *The Logic of Markedness*. Oxford University Press, New York

Caramazza A, Zurif E 1978 *Language Acquisition and Language Breakdown*. The Johns Hopkins University Press, Baltimore, MD

Chomsky N 1977 On *wh*-movement. In: Culicover P, Wasow T, Akmajian A (eds.) *Formal Syntax*. Academic Press, New York

Chomsky N 1986 *Knowledge of Language: Its Nature, Origin and Use*. Praeger, New York

Chomsky N, Halle M 1968 *The Sound Pattern of English*. Harper and Row, New York

Chomsky N, Lasnik H 1977 Filters and control. *Linguistic Inquiry* **8**(3): 425–504

Croft W 1990 *Typology and Universals*. Cambridge University Press, Cambridge, UK

Givón T 1990 *Syntax: A Functional-Typological Introduction*. John Benjamins, Amsterdam, Vol. 2

Greenberg J 1966 *Language Universals*. Mouton, The Hague, The Netherlands
Hatten R 1994 *Musical Meaning in Beethoven: Markedness, Correlation, and Interpretation*. Indiana University Press, Bloomington, IN
Jakobson R 1968 *Child Language, Aphasia and Phonological Universals*. Mouton, The Hague, The Netherlands
Jakobson R 1972 Verbal communication. *Scientific American* **227**: 72–80
Jakobson R 1984 The structure of the Russian verb. In: Waugh L, Halle M (eds.) *Russian and Slavic Grammar Studies, 1931–1981*. Mouton, Berlin
Kean M-L 1980 *The Theory of Markedness in Generative Grammar*. Indiana University Linguistics Club, Bloomington, IN
Liszka J 1989 *The Semiotic of Myth*. Indiana University Press, Bloomington, IN
Mayerthaler W 1988 *Morphological Naturalness*. Karoma, Ann Arbor, MI
McCawley J 1985 Kuhnian paradigms as systems of markedness conventions. In: Makkai A, Melby A (eds.) *Linguistics and Philosophy: Studies in Honor of Rulon S. Wells*. John Benjamins, Amsterdam
Myers-Scotton C (ed.) 1998 *Codes and Consequences: Choosing Linguistic Varieties*. Oxford University Press, New York
Waugh L 1982 Marked and unmarked: A choice between unequals in semiotic structure. *Semiotica* **38**: 299–318
Waugh L, Rudy S (eds.) 1991 *New Vistas in Grammar: Invariance and Variation*. John Benjamins, Amsterdam

E. Battistella

Linguistics: Overview

Linguistics, conventionally defined as the scientific study of language, has generated a wide range of branches, subdisciplines, schools of thought, and applications, many of which are given separate treatment in the encyclopedia. This article provides a perspective within which these and other aspects can be interrelated, insofar as they use the word 'linguistics' as part of their identity.

1. *Linguistics as a Science*

Linguistics achieved its academic identity during the twentieth century, shaped in the first few decades by a series of pioneering and influential personal statements (notably Saussure 1916, Sapir 1921, Bloomfield 1933), consolidated in the middle decades as a university subject at both undergraduate and graduate levels, and reinterpreted in the later decades by a new generation of scholars concerned to give the subject a sound theoretical foundation (notably Chomsky 1957, 1965). In Europe, its origins lay chiefly in a reaction to the philological curiosity about the history and origins of language, which had preoccupied the nineteenth century; in the USA, the primary impetus came from a concern to provide a linguistic dimension to the anthropological investigations of American Indian cultures. But the subject also benefited from an increased general interest in the study of language in relation to human beliefs and behavior (e.g., in philosophy, theology, information theory, literary criticism, communication studies), which manifested itself at both popular and specialized levels. The development of the media, especially broadcasting, further helped to bring language issues to the forefront of national consciousnesses, as did the demands made by foreign-language teachers and other 'language professionals' for more sophisticated information about their subject.

The multiplicity of interests in the subject, each with its own intellectual agenda and methodologies of inquiry, brought a profound appreciation of the range and complexity of linguistic phenomena, and thus of the need to develop a research discipline with its own principles and procedures. Early on, linguistics came to be viewed as a 'science,' and attempts were made to define its axioms and make its investigative procedures explicit (e.g., Bloomfield 1926, Hjelmslev 1943). Linguists emphasized their concern to make their studies systematic and objective, by contrast with the selective impressionism which had characterized much previous literary and pedagogical work on language. A great deal of attention was devoted to making investigative procedures clear and well defined. There was a significant growth, fostered by progress in acoustic technology, of experimental techniques, especially in phonetics. And there was an unprecedented concern for precise definition, clear and comprehensive description, and more powerful explanation. An emphasis in the first half of the century on the procedures for describing individual languages, or the features of language families, changed during the second half into an emphasis on the universal, defining properties of language, and on the nature of language viewed as part of human biology and psychology. Yet there were significant continuities too. Chomsky's focus on the distinction between linguistic competence and performance, and on the status of language as an 'organ of the mind,' echoed Saussure's notions of the collective language system (*langue*) and its concrete manifestation in utterance (*parole*), and of the human language faculty (*faculté de langage*).

Because of its claims to scientific stature, the subject has often been referred to as 'linguistic science,' but sometimes a designation of 'linguistic sciences' will be encountered. The choice hinges on the perceived standing of phonetics. In the 'plural' interpretation, phonetics is seen as a separate discipline, which along with linguistics makes up the linguistic sciences. The reasoning is that the subject-matter of phonetics—which includes acoustics, anatomy, physiology, neurology, and auditory perception—is so different from the 'rest' of linguistics that it cannot sensibly be subsumed under the same heading. Moreover, its

methodology is very different, both in its experimental approach and in its reliance on technology. It is, some argue accordingly, a prelanguage study, and its findings need to be brought into relationship with those of linguistics. Others, however, point to the fact that the bulk of the subject deals only with those aspects of sound production and processing that are relevant to the production of spoken language.

Phonetics focuses on such entities as vowels, consonants, and pitch movements, and not on such noises as whistles, hiccups, and burps, even though these are also made by the vocal apparatus. Its aim is to explicate speech sound, and as it is speech which is the ground of being for all languages (writing being present only in a proportion of instances), this alternative argument sees phonetics as the 'indispensable foundation' (Henderson 1971) for linguistics—one branch of a single linguistic science. (The term 'linguistics,' of course, avoids the need to make a decision.) A similar issue appears when the focus is on speech alone. Speech science (or speech and hearing science) is the study of all the factors involved in the production, transmission, and reception of speech; but speech sciences will also be found, when there is a need to emphasize the constituent disciplines involved.

2. *Classification by Orientation*

These changes of emphasis and direction, which are an inevitable part of the intellectual development of a subject, have given rise to a range of designations for linguistics, each of which captures a particular orientation or 'universe of interest.' The earliest of these, introduced by Saussure, was chronological, distinguishing diachronic linguistics from synchronic linguistics. This terminology reflects his concern to move away from an exclusively historical approach for the study of language to one in which a language is seen as existing at a particular moment (or 'state') in time, 'syn-chronically,' regardless of whatever may have happened to it previously or would happen to it subsequently. Most linguistic research is synchronic, in the sense that linguists are aiming to establish the nature of the system of rules which define a (part of a) language's expressive potential at any one time. For example, the word 'balcony' currently has a stress on its first syllable, and this is what would be noted as part of a synchronic description of modern English stress; the fact that, over a century ago, the stress was on the second syllable, would not be considered relevant. On the other hand, in a diachronic description, the way the stress has shifted is of central relevance—not least because the process is continuing in many other words, such as 'dispute' and 'research.' Saussure saw the two chronological orientations as intersecting—an 'axis of simultaneities' crossing an 'axis of successivities'—and this point of intersection must not be ignored. Uncertainties about usage are one example of the way in which ongoing language change complicates a synchronic analysis. But the point of the distinction is widely appreciated, and Saussure's refocusing has been so successful that it has long been standard practice to interpret the word 'linguistics' as meaning 'synchronic linguistics,' with references to the other orientation requiring a more explicit phrase—usually, these days, 'historical linguistics.'

Another type of orientation relates to the generality of a linguistic inquiry. As already mentioned, a contemporary focus on language is to make statements of maximal generality: the aim is to determine what it is that defines the notion of 'human language.' All languages have a great deal in common in the way they produce sounds, organize their grammars, and construct systems of meaning in words. Identifying the universal principles which govern language, and demonstrating how these principles apply in individual circumstances to produce the languages known to us as French, English, Chinese, Swahili, etc., is the primary aim of theoretical linguistics. The term 'theory' here has its standard scientific application, reflecting the need for explanation of a maximally general kind, and the provision of fruitful models which will generate testable hypotheses about, in this case, linguistic structure and use. Theoretical linguists do not study a language because they are interested in that language *per se*; they study it because of what it can tell them about the nature of language in general. A particular language may be especially 'interesting' because it makes use of a type of sound or grammatical construction that other languages do not. The label 'general linguistics' is often used to capture this breadth of vision. A further consequence of this approach is that, once some linguistic data has been accumulated, a special concern is to find the best way of analyzing it. There are so many variables in pronunciation, grammar, and vocabulary that it is possible to envision an indefinitely large number of ways of showing how these domains are organized. Theoretical linguists spend much of their time searching for plausible and elegant analyses, and identifying criteria for their evaluation.

One of the controversies in the profession relates to this point. According to some linguists, as a result of the above orientation, the subject has become 'too general.' Analyses which try to achieve maximal generality inevitably make statements which are highly abstract, and seem to be far removed from the description of individual languages, and of the way these are used in everyday speech and writing. The emphasis on achieving generality, according to this view, has brought a de-emphasis on individuality—of what it is that makes one language unique, different from all the others. While it is accepted that, in the long term, all the idiosyncrasies of individual languages may be explicable with reference to general rules, there is seen to be no likelihood of this happening

in the foreseeable future. In the meantime, the need to provide descriptions which are as full as possible is as important now as it ever was. Indeed, this focus on the special character of a particular language was given special urgency in the 1990s, when it was realized just how many of the world's languages are near the point of extinction. Of the 6,000 or so languages in the world, it is widely thought that half are likely to die out within the next 100 years (Crystal 2000). The need to make descriptions of the languages most at risk is therefore a priority for the subject, for once a language dies which has never been written down or recorded in any way, it is as if it had never been. This orientation for the subject, where the aim is to make as complete as possible a description of a language, in all its varieties, is usually referred to as descriptive linguistics.

It is not only the endangered languages which benefit from a descriptive approach, therefore. Descriptive accounts of English, German, Japanese, and all the other 'safe' languages are also important. 'Descriptive' here is being seen in contrast with 'traditional,' referring to the tradition of language inquiry which antedated the emergence of twentieth-century linguistics, and which was characterized—especially in the field of grammar—by prescriptive and proscriptive rules about usage, purist condemnation of language change, and the distorted description of modern languages through the use of a model of description derived from Latin. In the field of English, the need for more realistic description was the driving force behind many grammars, pronunciation guides, and dictionaries, especially from the 1950s on (e.g., Jones 1956, Fries 1952, Quirk et al. 1972). A particular consequence of this orientation was the provision, in some countries, of new syllabuses and materials for use in schools and teacher-training courses—a change of direction which in the later decades of the century brought not a little heart-searching, as teachers found themselves engaging in a close encounter with a linguistically based conceptual apparatus and a partly unfamiliar terminology. At a research level, since the 1980s the descriptive orientation of linguistics has been given additional impetus through the use of computational methods of data collection and analysis. Large bodies of data, in some cases consisting of several hundred million words, are now available as sources of evidence for descriptive statement, in a field that has come to be called corpus linguistics.

A considerable conceptual distance separates the description of an individual language from a general account of the properties of human language, so it is not surprising to find orientations which occupy aspects of the 'middle ground.' It is possible to extend the descriptive orientation 'upwards,' bringing individual languages together in specific ways but without the inquiry necessarily bearing on a hypothesis about linguistic universals. This is the motivation behind comparative linguistics, where the purpose is to identify the chief similarities and differences between historical states of a language, between modern languages (e.g., those within a particular family, such as Romance or Germanic) or, in a more ambitious mode, between language families as wholes. This use of 'comparative' in fact antedates modern linguistics, being part of the designation of comparative philology, the subject which grew up at the end of the eighteenth century devoted to the historical comparison of Greek, Latin, Sanskrit, and their hypothetical antecedents, whence arose the concept of an Indo-European family of languages.

Applying the 'comparative method' to other languages which had written records was a natural extension, and philology (as it is today usually called) is practiced still, being distinguished from the orientation of historical linguistics partly by its aims and methods, and partly by the contrasting intellectual background (often from literature or history) of its practitioners. A related designation is typological linguistics, where the aim is to study the structural similarities between languages, regardless of their history, in order to locate them within a classification (or 'typology'). This is the only possible method in cases where there are no written historical records—a situation in which most of the world's languages find themselves.

When linguists talk about the structure of language, they are usually referring to the way a language can be analyzed into several dimensions susceptible of independent study. These dimensions are commonly referred to as 'levels' or 'components,' though terminology varies widely between linguistic models. Most accounts recognize three major levels of linguistic organization: a sound system (or phonology, including its interface with phonetics), a grammatical system (with a division often made between sentence-structure (syntax) and word-structure (morphology)), and a semantic system (typically focusing on vocabulary (the lexicon), though not thereby excluding other ways in which a language's meaning is organized). The study of the written language necessitates the recognition of an additional level: an orthographic system (or graphology). And some linguists recognize other levels of analysis, such as one which extends linguistic inquiry above the sentence (a level of discourse, or text) or one which takes into account the way structure is altered when language is used interactively (pragmatics). Each of these terms has a variety of interpretations, depending on the theoretical approaches involved, but they all have one thing in common: they aim to explicate a domain of language structure, and are thus part of an orientation to the subject that might be legitimately called structural linguistics.

However, this designation has had a controversial position within the history of twentieth-century linguistic thought. In its early use, 'structural' linguistics referred to the particular approaches to phonology

and grammar which were current in the 1940s and 1950s, in the work of such linguists as Bloomfield. There, the emphasis was on finding methods which would establish the contrasting units in a language's sound system (the phonemes), identify the elements of word formation (the morphemes), and analyze sentence constructions into their constituent parts. In a somewhat broader sense, developing the initiatives of Saussure, a 'structuralist' ethos prevailed, which aimed to establish systems of relations between linguistic units—an approach which, under the heading of structuralism, came to influence thinking throughout the arts and social sciences. Within linguistics, it was in due course applied to further areas of linguistic structure (especially semantics), and extended by some linguists (such as Pike 1954–60) to other domains of human communicative behavior. The notion of system continued to play a central role in linguistic theories which emerged during the postwar period, notably in the work of Firth (1957) and his students, and later became central in the systemic linguistics of Halliday (Halliday and Fawcett 1987). It was the perceived limitations of the structuralist orientation, with its focus on the classification of structures and units at a 'surface' level of description within a corpus of data (linguistic 'performance'), which prompted its labeling as 'taxonomic' (with a pejorative implication) by Chomsky, and which motivated his initial formulation of generative linguistics (Chomsky 1957), with its concern to explicate underlying linguistic relations in ways that could account for the creative linguistic abilities (or 'competence') of the individual.

3. *Classification by Interaction*

The generative orientation constituted a revolution in linguistic thought; and although other kinds of linguistic theories and models continued to be devised and developed in the latter part of the century, it was the generative approach, in its various formulations, which provided the subject's center of gravity. At the same time, partly independent of this state of affairs and partly as a consequence of it, several areas of interaction with other academic disciplines emerged. These 'hybrid' subjects were the result of a synergy between two intellectual streams. On the one hand, linguists sensed that they needed to take their subject in fresh directions if they wished to account for everything that influenced the way people spoke and wrote (and also signed, for the analysis of deaf signing was also emerging as a new research field). On the other hand, scholars in other academic disciplines were finding that they needed to provide more sophisticated accounts of language structure and use in order to maintain their own research agendas. The potential for interaction had in some cases been recognized for many years. In the case of the interaction between linguistics and anthropology, it had been there from the earliest days of the subject (Boas 1911). The difference was partly in the way hybrid labels came to be institutionalized as 'recognized' domains for teaching and research (anthropological linguistics, in this case), and partly in the cross-disciplinary climate of the times, which gave rise to nearly two dozen such hybrids by the turn of the century.

There is not always a total meeting of minds in these interdisciplinary encounters: the orientation given to the hybrid subject within a linguistics department is by no means identical with the orientation it is given within the department which constitutes the other half of the hybrid. For example, one of the areas of inquiry in psycholinguistics is the relationship between language and memory. When a linguist investigates how language is affected by memory and a psychologist investigates how memory is affected by language, it might be thought that the two directions of inquiry would 'meet in the middle.' In practice, differences in the researchers' intellectual histories, theoretical aims, and research procedures can produce widely divergent accounts. One need only look at the list of references at the end of journal articles to see how authors, ostensibly investigating the same subject, can come from different directions. There may be a difference of labeling, which may reflect a difference of orientation—as in sociolinguistics (a branch of linguistics) alongside the sociology of language (a branch of sociology), or philosophical linguistics (a branch of linguistics) alongside linguistic philosophy (a particular approach within philosophy). Nonetheless, the shared subject-matter, and a growing mutual familiarization with the intellectual background of the contributing disciplines, produced in the last quarter of the twentieth century a growing number of research collaborations, interdisciplinary conferences, jointly authored textbooks, and shared teaching courses in many of these hybrid subjects, resulting in a growing identity and coherence.

Hybrid subjects can be placed in clusters, on the basis of an overlap in their subject-matter. The earliest clusters to be given clear definition all relate to a concern to see language in its cultural and social context. Anthropological linguistics (or linguistic anthropology) focuses on language variation and use in relation to human cultural patterns and beliefs (in its early years, especially on the Amerindian peoples of North America). Overlapping with this is ethnolinguistics, studying language in relation to ethnic types and behavior, and widely practiced through approaches (variously called the ethnography of communication or ethnography of speaking) which analyze the entire range of variables, extralinguistic as well as linguistic, involved in social interaction. Sociolinguistics, studying all aspects of the relationship between language and society, is the most well established of this cluster of subjects, and is sometimes used as a covering label for all domains of inquiry in which there is a preoccupation with the social func-

tions of language. However, sociological linguistics is sometimes differentiated from it, particularly in mainland Europe, reflecting a concern to see language as an integral part of sociological theory. So too is interactional sociolinguistics, referring to the study of speech in face-to-face communication.

Several other facets of the sociolinguistic cluster have maintained their individual nomenclature. One example is dialectology (sometimes called dialect geography), which focuses on the properties of regional (and more recently social) dialects. Another is areal linguistics, which studies the linguistic properties (languages as well as dialects) of large geographical regions—such as Western Europe or the British Isles. Closely related is geolinguistics (also geographical linguistics), which studies the geographical distribution of languages throughout the world with reference to their political, economic, and cultural status. Pragmalinguistics studies language use from the viewpoint of a language's structural resources; it contrasts with those pragmatic studies which examine the conditions on language use deriving from the social situation (sometimes referred to as sociopragmatics). Finally, there is the study of (linguistic) style, stylistics—a study which, though chiefly concerned with the nature of the personal linguistic identity of individuals (especially in a literary context), has often been extended to include the distinctive linguistic features of group-identifying functions of language, such as are found in advertising, science, and the media. When the focus is specifically on the distinctive language used in social institutions, such as law, medicine, and religion, some linguists use the label institutional linguistics.

Another cluster of hybrid subjects looks at language in relation to human biology. In its broadest characterization, the subject is referred to as biological linguistics (or biolinguisics), focusing on the preconditions for language development and use from the viewpoints of both the history of language in the human race and the development of language in the individual (the latter topic usually being distinguished by the name developmental linguistics or child language acquisition). Its topics include the genetic transmission of language, neurophysiological models of language production, and the anatomical comparison of humans and other species. It also subsumes the study of pathological forms of language behavior (as in aphasia, dyslexia, and language delay), though the complexity of these conditions, and the interdisciplinary nature of their diagnosis and treatment, has prompted the development of separate specialisms. Chief among these is neurolinguistics (sometimes referred to as neurological linguistics), which focuses on the neurological basis of language development and use, in particular aiming to understand the way the brain controls the processes of speaking, listening, reading, writing, and signing. Although its focus is on clinically normal states, much of its evidence comes from the investigation of abnormal language development and language breakdown. It thus overlaps with clinical linguistics, which is exclusively devoted to the diagnosis, assessment, and treatment of all forms of language disability in children and adults.

Any of these domains can of course be studied from the point of view of 'mind' as distinct from 'body,' and psycholinguistics (sometimes called psychological linguistics) has emerged as a major subject, investigating the interaction between language and such topics as memory, perception, attention, and personality, in normal and abnormal states. Depending on the domain of inquiry involved, compound names will be encountered, such as clinical psycholinguistics and developmental psycholinguistics. Psychology shares with education an interest in learning, and there has been special concern to understand how language is dealt with in schools and other educational institutions. Educational linguistics (sometimes pedagogical linguistics) focuses on the way government policies and planning, teacher training, curricula, syllabuses, and pedagogical materials handle the teaching of the mother tongue and foreign languages. The later decades of the twentieth century also saw a growing metalinguistic concern, with researchers investigating the ways in which children become increasingly aware of language and its elements, in a domain which by the 1990s had been labeled language awareness. Within the teaching profession, a corresponding fresh interest in the study of language, largely motivated in the UK by the arrival of the National Curriculum, introduced a new acronym, KAL (knowledge about language).

A further cluster of hybrids relates to the numerical sciences. Mathematical linguistics studies the formal properties of language, usually employing concepts of an algebraic or statistical kind. Its main application has been in the formalization of linguistic theory, as developed in relation to generative linguistics, but several other areas of language study have received mathematical investigation. In particular, statistical (or quantitative) linguistics studies the frequency and distribution of linguistic units in texts, with the aim of establishing general laws concerning the statistical properties of language (such as the relationship between word types and tokens), and identifying the distinctive characteristics of authors (stylostatistics). In the later part of the twentieth century, computational techniques and concepts came to be increasingly applied to the elucidation of linguistic problems, under the heading of computational linguistics. Several research areas have been developed, including natural language processing, automatic speech synthesis and recognition, machine translation, the making of concordances, and the many areas where statistical counts and analyses are required. The provision of large machine-readable texts (computer corpora) was a special development of the 1990s, and gave rise to the domain of corpus linguistics. Linguistic

involvement in the other direction—in relation to such matters as computer programming, information retrieval, document identification, indexing, reference science, and knowledge management—has been limited to isolated projects, and has not yet resulted in a hybrid label, though one imagines that, given the growth in this area, the emergence of Internet linguistics can only be a matter of time.

Not all hybrid terms can be neatly clustered. Some, by their nature, stand outside of classification, because their role is to evaluate aspects of the field as a whole. The study of the history of ideas in linguistics is linguistic historiography. Philosophical linguistics studies the role of language in relation to the understanding and elucidation of philosophical concepts, as well as the philosophical status of linguistic theories, methods, and observations. Critical linguistics emerged in the 1990s with the aim of revealing hidden power relations and ideological processes at work in spoken or written texts. Critical linguists criticize mainstream linguistics for its formalist preoccupations, its lack of adequate social explanations, and its obscuring of ideological and political issues. This 'critical' perspective may be applied to individual branches of the subject. In particular, critical discourse analysis studies the relationship between discourse events and sociopolitical and cultural factors, especially the way discourse is ideologically influenced by, and can itself influence, power relations in society.

4. *Applications and Extensions*

The list of possible hybrid subjects involving linguistics is by no means exhausted by the above itemization, and new ones will doubtless emerge. A similar potential for expansion is also found in the various domains of applied linguistics. In its broadest definition, this is the application of linguistic theories, methods, and findings to the elucidation of language problems which have arisen in other areas of experience. The most established and well-developed branch is the teaching and learning of foreign languages, and sometimes the term is used as if this were the only domain involved. But several other domains of application have emerged since the middle of the twentieth century, including speech pathology and therapy, the teaching of the deaf, mother-tongue education, the writing of dictionaries (lexicography), the analysis of literary texts (literary stylistics), and translation studies. Quite restricted and specialized domains may be encountered. For example, theolinguistics has as its focus the study of the way language is used in theological, biblical, and other branches of religious studies. Forensic linguistics is the use of linguistic techniques to investigate crimes in which language data forms part of the evidence, such as in the use of grammatical or lexical criteria to authenticate police statements. The field of forensic phonetics is often distinguished as a separate domain, dealing with such matters as speaker identification, voice line-ups, speaker profiling, tape enhancement, tape authentication, and the decoding of disputed utterances.

There is an uncertain boundary between applied linguistics and the various interdisciplinary subjects reviewed in Sect. 3, especially as several of those concerns involve practical outcomes (e.g., planning a national language policy within sociolinguistics). On the other hand, as these hybrid subjects develop their own theoretical foundations, the distinction between 'pure' and 'applied' is becoming more evident, and terminological distinctions such as that between (theoretical) psycholinguistics and applied psycholinguistics are now more regularly encountered.

The term 'linguistics' will also be encountered in a wide range of other contexts, where it is often no more than a loosely used synonym for 'school of thought' or 'approach,' or even a stylistic replacement for the word language or grammar. Very general states of mind on the part of the investigator can be given a linguistics designation. For example, mentalist(ic) linguistics is contrasted with behaviorist linguistics. Macrolinguistics, representing a relatively broad frame of reference for linguistic inquiry, is opposed to microlinguistics. In relation to schools, we find such usages as Cartesian linguistics (showing the influence of Descartes), Chomskyan linguistics (from N. Chomsky), Whorfian linguistics (from B. L. Whorf), Saussurean linguistics (from F. de Saussure), Firthian linguistics (from J. R. Firth), and neo-Firthian linguistics (scholars whose teaching derived from Firth's influence, notably M. A. K. Halliday (whose name has, in turn, been given to Hallidayan linguistics).

At this level, there is an appreciable overlap with the relevant aspect of language which is the focus of an approach—for example, the terms Chomskyan linguistics and Chomskyan grammar are often used synonymously, as are Whorfian linguistics and Whorfian semantics. The same point applies to such general approaches as functional(ist) linguistics, which treats the notion of function as central, especially in grammar, and thus allows a terminological overlap with functional(ist) grammar. Systemic and polysystemic linguistics (which treat the notion of system as central) may also be found alongside systemic and polysystemic grammar. Virtually any variety or use of language can be focused in this way, as in creole linguistics (= the linguistic study of creole languages) and paralinguistics (= the linguistic study of paralanguage—tones of voice, facial expressions, etc). But eventually, as increasingly restricted areas of inquiry are encountered, the use of the general term ceases to have a point, and linguists replace it by specific terms such as morphology, phonology, lexicology, semantics, onomastics, and other structural labels best explicated under the heading of 'language.'

See also: Evolution and Language: Overview; Historical Linguistics: Overview; Language and Animal Competencies; Language and Society: Cultural Concerns; Psycholinguistics: Overview; Saussure, Ferdinand de (1857–1913); Sociolinguistics; Structuralism.

Bibliography

Bloomfield L 1926 A set of postulates for the science of language. *Language* **2**: 153–64

Bloomfield L 1933 *Language*. Holt Rinehart & Winston, New York

Boas F (ed.) 1911 *Handbook of American Indian Languages*. Smithsonian Institution, Washington, DC

Bright W (ed.) 1992 *International Encyclopedia of Linguistics*. Oxford University Press, Oxford, UK

Chomsky A N 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands

Chomsky A N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Crystal D 1985 *Linguistics*, 2nd edn. Penguin, Harmondsworth, UK

Crystal D 1996 *A Dictionary of Linguistics and Phonetics*, 4th edn. Blackwell, Oxford, UK

Crystal D 1997 *The Cambridge Encyclopedia of Language*, 2nd edn. Cambridge University Press, Cambridge, UK

Crystal D 2000 *Language Death*. Cambridge University Press, Cambridge, UK

Firth J R 1957 A synopsis of linguistic theory, 1930–55. In: Philological Society (Great Britain), Special Volume, *Studies in Linguistic Analysis*. Blackwell, Oxford, UK, pp. 1–32

Fries C C 1952 *The Structure of English*. Harcourt Brace, New York

Halliday M A K, Fawcett R P (eds.) 1987 *New Developments in Systemic Linguistics*. Pinter, London

Henderson E J A (ed.) 1971 *The Indispensable Foundation: A Selection from the Writings of Henry Sweet*. Oxford University Press, London

Hjelmslev L 1943 *Omkring Sprogteoriens Grundlæggelse*. Lunos, Copenhagen. [Trans. as *Prolegomena to a Theory of Language*, 1961]. University of Wisconsin Press, Madison, WI

Jones D 1956 *An Outline of English Phonetics*, 8th edn. Heffer, Cambridge, UK

Lyons J 1981 *Language and Linguistics: An Introduction*. Cambridge University Press, Cambridge, UK

Pike K L 1954–60 *Language in Relation to a Unified Theory of the Structure of Human Behavior*. Summer Institute of Linguistics, Glendale, CA

Quirk R, Greenbaum S, Leech G N, Svartvik J 1972 *A Grammar of Contemporary English*. Longman, London

Robins R H 1979 *General Linguistics: An Introductory Survey*, 3rd edn. Longman, London

Sapir E 1921 *Language*. Harcourt Brace, New York

Saussure F de 1916 *Cours de Linguistique Générale*. Payot, Paris. [Trans. as *Course in General Linguistics*, 1959. Philosophical Library, New York]

D. Crystal

Linguistics: Prototype Theory

Prototype theory is a theory of categorization. Categorization is a central issue in linguistics, for several reasons. In the first place, words can be regarded as names for categories, and using a word involves an act of categorization (Labov 1973). In stating that *The cup is in the saucer*, a speaker is categorizing the figure object as a 'cup,' rather than as a mug, bowl, vase, or other kind of object; the reference object is being categorized as a 'saucer,' not as a dish or a plate; while the location of the cup vis-à-vis the saucer is being categorized as an 'in'-relationship, rather than as an 'on'-relationship, or whatever. Knowing these words involves knowing how to apply the words appropriately to entities that fall within the categories that they designate.

There is another way in which categorization impinges on linguistic theory. The moment one pursues even the most elementary linguistic analysis, language itself becomes an object of categorization. Like any intellectual discipline, linguistics requires its practitioners to set up categories for the description of the object of study. In any linguistic description, one will encounter such entities as words and morphemes, phrases and clauses, nouns and verbs, subjects and objects, clitics and particles, phonemes and allophones. 'Doing linguistics' consists, in large measure, in setting up categories such as these, defining them, studying their properties and interrelationships, working out criteria for determining their membership, and applying them to the data at hand. The ontological status of the categories themselves may also become an issue. Where do categories such as 'noun' and 'verb' come from? Are they inherent in the object of study? Are they mere descriptive conveniences? Are they constructs of linguistic theory?

1. Two Approaches to Categorization

Two major approaches to categorization are the 'classical' approach and the 'prototype' approach. On the classical approach, a category is defined in terms of a set of necessary and sufficient conditions for category membership. The set of conditions specifies the essence, as it were, of the category. Any entity meeting all the conditions is, by definition, a member of the category; failure to meet any of the conditions excludes an entity from the category. On this view, category boundaries are discrete (an entity either is, or is not, a member), and the category lacks internal structure (all members are of equal status).

The simplicity of the classical view makes it intuitively appealing. A major problem is that many categories are simply not amenable to classical definition. What, for example, are the defining features of 'chair,' 'vegetable,' and 'love,' or, for that matter, of

'cup,' 'saucer,' and 'in'? In fact, classical categorization probably works best only for 'expert' categories, that is, categories that have been created and explicitly defined (in classical terms!) by experts, whether these be scientists, bureaucrats, administrators, or whoever. For most words of everyday speech, the classical theory is quite inappropriate.

On the prototype approach, a category is understood, primarily, in terms of its good examples. Entities are associated with the category to the extent that they resemble the good examples. Often, the boundary of a prototype category will be fuzzy—it might not be clear whether some entities belong in the category or not. Moreover, prototype categories have an internal structure, in that some entities count as more central members than others.

Prototype categorization came to the attention of linguists largely due to the work of the psychologist, Eleanor Rosch. In a series of experiments, summarized in Rosch (1978), it emerged that the meanings of several very ordinary words, such as *bird*, *toy*, *vehicle*, are structured around good examples. People readily agree that a robin is a better example of 'bird' than a penguin; they can determine the truth of *A robin is a bird* faster than *A penguin is a bird*; when asked to name members of a category, they name more central examples before more marginal examples; and when drawing inferences about a category, people tend to draw on knowledge of central members, rather than on knowledge about the category as a whole. Rosch observed prototype effects both with 'natural kind' categories—that is, categories such as 'bird,' which are, as it were, given by nature—and with categories that are understood relative to human concerns, such as 'vehicle,' 'furniture,' and 'toy.'

Rosch's discoveries were eagerly taken up by linguists, and extended by a wide range of linguistic data (Lakoff 1987, Kleiber 1990, Tsohatzidis 1990, Geeraerts et al. 1994, Taylor 1995). An especially fruitful application has been the study of semantic change (Geeraerts 1997). The prototype notion has also been widely applied to the study of the categories of linguistic description (Langacker 1987), Croft 1991).

2. *Problems and Refinements*

Despite its success in lexical semantics and elsewhere, prototype categorization raises a number of problems.

While not denying the reality of prototype effects, Fodor (1996) argues that concepts cannot be equated with prototypes. The fish prototype (a herring?) and the pet prototype (a poodle?) contribute nothing at all to the complex concept 'pet fish,' nor to the prototype of 'pet fish' (a goldfish?). Generally, as a concept gets specified in greater detail, prototype effects not only shift, they also tend to get effaced. Consider the diminishing role of the 'nurse' prototype in one's understanding of *nurse*, *male nurse*, *retired male nurse*, etc. On the other hand, prototype effects show up even on categories that arguably are subject to classical definition, such as 'odd number.' Armstrong et al. (1983) report that subjects readily agreed that 3 was a better example of 'odd number' than 447, even though, obviously, 447 is no less of an odd number than 3.

On one interpretation of these findings, prototype effects are extrinsic to word meanings as such, and arise only when a person applies a concept to the world. This view draws a distinction between the 'essence' of a category (the 'real' meaning of a word), and 'recognition procedures' invoked when speakers attempt to categorize things and eventualities in the world. On this view, the fuzziness of category boundaries results from the fact that eventualities in the world may not lend themselves to clear-cut categorization, even though the words of a language do have clear-cut meanings. Although it might not be clear whether a particular utterance counts as a 'lie,' the meaning of *lie* might nevertheless be clear-cut.

2.1 *Degree of Membership vs. Representativity*

Before endorsing this exclusion of prototypes from word meaning proper, it is important to distinguish two aspects of prototype categories: fuzziness of category boundaries (a question of category membership), and degree of representativity within a category (a question of category structure).

It is uncontroversial that certain categories are inherently fuzzy, in that it is simply not possible to make a clean distinction between what goes into the category and what does not. This is the case with many adjectival concepts, such as 'old,' 'tall,' and 'bald.' While there is little doubt what a (prototypical) *bald head* looks like, also what it means to be *not bald*, it would be futile to draw a boundary between 'bald' and 'not bald,' e.g., in terms of the number of hairs on a person's head.

However, prototype effects also show up on categories that arguably do have clear boundaries. Although turkeys might not be good examples of the bird category, they are birds nonetheless, albeit not very representative ones. Representativity, to be sure, can be understood in different ways—in terms of frequency of occurrence, cultural or environmental salience, or statistical averaging over category instances. Representative examples are typically invoked in default contexts. Given only the mention of 'a bird,' with no further specification, one would assume a representative example. A person who wishes they were a bird would surely be disappointed if they were transformed into a turkey! The internal structure of a category cannot reasonably be excluded from a person's knowledge of word meanings.

2.2 Prototypes and Background Knowledge

Categories, in order to be viable, need to be conceptually coherent (Murphy and Medin 1985). A category is coherent to the extent that it is based on theories of how the world is, and how it links up with a person's encyclopedic knowledge. A category is not simply a conjunction of an *ad hoc* set of features, the category has to make sense.

Lakoff (1987) discusses this aspect in terms of his notion of the Idealized Cognitive Model (ICM). The concept 'bachelor' is understood against an ICM of marriage practices (which in turn is based in ICM's of sexuality, procreation, and family). One aspect of the ICM is the idea that each person, after reaching the 'marriageable age,' sooner or later marries. Within the ICM, *bachelor* may well be definable in classical-like terms, i.e., as 'unmarried,' 'adult,' and 'male.' Prototype effects arise, because the ICM may not fit the world perfectly. The Pope is not a good example of 'bachelor,' not because of uncertainties about whether the Pope is unmarried, adult, or male, but because Catholic clergy are outside the scope of the ICM.

2.3 Prototypes and Features

On the classical theory, categories are assembled, as it were, out of their features; the features pre-exist the category, they are evaluated independently of the category, and must be simpler than the category they define. A prototype, in contrast, can be perceived holistically, as a gestalt; features are no longer the building blocks of categories, they emerge, as dimensions of similarity, through acquaintance with cat egory members. It is unlikely that people construct their bird-category through conjunction of features such as 'having feathers' and 'laying eggs'; knowledge of feathers and egg-laying emerges from acquaintance with birds, not vice versa. Likewise with linguistic categories. One might propose that ability to inflect for tense is a defining characteristic of verbs (for English, at least). Yet, the very notion of tense inflection rests on a prior acquaintance with members of the verb category.

Given this revised understanding of features, a feature-based view of prototype categories becomes possible, through the differential weighting of the features. Some features may be better cues to category membership than others; some features may be virtually indispensable; others are able to be overridden. The prototype will be a representation that maximizes feature weights; an entity exhibiting highly weighted features will automatically count as a category member; an entity exhibiting less highly weighted features might still count as a member, though a more marginal one. Moreover, it is possible for two entities to belong to a category, even though they have no category-specific features in common. The way also opens up for a reconciliation of classical and prototype categorization—a classical category is the limiting case of a prototype category, i.e., a category understood solely in terms of essential features.

3. Prototypes and Polysemy

It is useful to distinguish between the prototype structure of a concept and the polysemy structure of a word. There is a clear difference, in principle, between applying the word *fruit* to different kinds of fruit, and the use of the word in such expressions as *the fruit of his labors*. The latter usage represents a semantic extension of the word, in a way that is different from the use of the word to refer, say, to olives.

Take the preposition *over* (Lakoff 1987). *Fly over the Atlantic*, *the lamp hanging over the table*, *put your hands over your face*, *come over here*, *there's water all over the floor*, *the party's over*, *do it over again*—these exemplify different senses of the word, rather than more or less representative examples of one and the same sense. To the extent that these senses can be clearly individuated, it is possible to construct a network of sense relations, which links up the different senses according to their similarity. Such an array invites a prototype analysis, with one of the senses selected as central, and the others radiating out from it.

There are, however, different criteria for selecting the central sense, and these do not always deliver the same result. On a diachronic perspective, the basic sense is the historically oldest one, from which the others have been derived over time. On a developmental perspective, the basic sense is the one that children acquire first. Alternatively, the descriptive linguist (or lexicographer) might select as basic the most frequent sense, or the one to which the others can be most economically related. The real question, however, concerns the speaker's mental organization of polysemy. Currently, this is a lively topic in polysemy research (Taylor 1995).

4. Prototypes and Linguistic Categories

The prototype concept may be applied, not only to the study of word meanings, but to the very categories of linguistic description. Many linguistic concepts (including the very notion of 'language' and 'a language') turn out to have a prototype structure, in that they exhibit degrees of representativity, and often have fuzzy boundaries. Uncertainty as to what goes into a category and what does not pertains even to such basic notions as what constitutes a word (as opposed to a bound morpheme, clitic, or phrase), the part-of-speech categories (whether a particular item is a noun, verb, adjective, etc.), constituency (whether a particular string of words is a syntactic constituent), and syn-

tactic relations (whether a particular constituent is a modifier or adjunct, the subject of a clause, or whatever).

Some linguistic theories propose to decide such issues by *fiat*. In practice, however, the linguist's strategy often reflects a prototype conception, even though this might not be explicitly acknowledged. When deciding the category status of a linguistic item, it is usual to apply a set of tests (Croft 1991). The tests are set up on the basis of what, intuitively, count as good examples of the category in question, whereby each of the tests diagnoses a property typical of the good examples. An item which passes all the tests is *ipso facto* a member of the category; an item which fails the tests is not a member. Sometimes, an item passes only some of the tests; it will have to be regarded as a marginal example, or even as an item of uncertain status.

Consider the parts of speech, perhaps better called 'lexical categories,' such as 'noun,' 'verb,' 'adjective,' 'preposition' (see *Word Classes and Parts of Speech*). These can be characterized either in semantic or in formal terms. Semantically, nouns prototypically designate categories of things, i.e., concrete, time-stable, spatially bounded objects (*chain*, *cup*, etc.); verbs designate events, involving rapid changes in state (*explore*, *arrive*); adjectives designate fairly stable properties of things (*hot*, *young*); while prepositions designate a relation, typically a spatial relation, between things (*on*, *at*). It is commonly agreed that lexical category cannot be reliably predicted from a word's semantics. In spite of their reference to rapid changes in state, *explosion* and *arrival* are nouns, just as good nouns, in fact, as *chair* and *cup*. This is because the words exhibit the full range of formal properties typical of the noun category, i.e., they pass all the syntactic tests for nounhood—they pluralize, they take the full range of determiners, they can be modified by an adjective, they can head a subject noun phrase, and so on. There are words which fail to exhibit the full range of properties. *Groceries*, in *I bought some groceries*, looks like the plural of a noun *grocery*—yet there is no such noun (**I bought a grocery*), neither can *groceries* take numeral quantifiers (**five groceries*). Syntactically, *groceries* is a somewhat marginal noun (though still a noun). However, there are words whose status is genuinely unclear. Consider *saying*, in *without my saying a word*. Being the complement of a preposition, *my saying a word* looks like a noun phrase, headed by *saying*, and *saying* looks like a noun, in that it takes possessive *my*. Other determiners, though, are not possible (*without *the/*that/*a saying a word*). Moreover, *saying* is followed by what looks like a direct object (*a word*), and the ability to take a direct object is a distinctly verb-like property, not at all a property of nouns.

Both semantic and formal criteria, then, fail to deliver clear-cut lexical categories. Nevertheless, there is an important sense in which the semantic prototypes have priority. Any language has to provide its speakers with the means to refer to 'things' and 'events,' 'properties' and 'relations,' and the semantic prototypes of the lexical categories, in English and other languages, are understood in such terms. Indeed, without the semantic prototypes, there would be no basis for recognizing the categories 'noun' and 'verb' across the different languages of the world.

See also: Lexical Semantics; Lexicology and Lexicography; Lexicon; Syntax; Word Meaning: Psychological Aspects

Bibliography

Armstrong S L, Gleitman L R, Gleitman H 1983 What some concepts might not be. *Cognition* **13**: 263–308

Croft W 1991 *Syntactic Categories and Grammatical Relations*. University of Chicago Press, Chicago

Fodor J A 1998 *Concepts: Where Cognitive Science Went Wrong*. Clarendon Press, Oxford, UK

Geeraerts D 1997 *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Clarendon Press, Oxford, UK

Geeraerts D, Grondelaers S, Bakema P 1994 *The Structure of Lexical Variation: Meaning, Naming, and Context*. Mouton de Gruyter, Berlin

Kleiber G 1990 *La sémantique du prototype: Catégories et sens lexical*, 1st edn. Presses Universitaires de France, Paris

Labov W 1973 The boundaries of words and their meanings. In: Bailey C J, Shuy R (eds.) *New Ways of Analysing Variation in English*. Georgetown University Press, Washington, DC

Lakoff G 1987 *Woman, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago University Press, Chicago

Langacker R W 1987 *Foundations of Cognitive Grammar*. Stanford University Press, Stanford, Vol. 1

Murphy G, Medin D 1985 The role of theories in conceptual coherence. *Psychological Review* **92**: 289–316

Rosch E 1978 Principles of categorization. In: Rosch E, Lloyd B B (eds.) *Cognition and Categorization*. Lawrence Erlbaum, Hillsdale, MI

Taylor J 1995 *Linguistic Categorization: Prototypes in Linguistic Theory*. Clarendon Press, Oxford, UK

Tsohatzidis S (ed.) 1990 *Meanings and Prototypes: Studies in Linguistic Categorization*. Routledge, London

J. R. Taylor

Linguistics: Theory of Principles and Parameters

In generative syntax (see *Generative Grammar*; *Linguistics: Incorporation*) principles and parameters are two related concepts introduced to capture both invariance of language and cross-linguistic vari-

ation. *Principles* express universal constraints on language; *parameters* define the space of cross-linguistic variation.

1. Background: The Problem of Language Acquisition

The goal of generative linguistics is that of developing a model of linguistic knowledge (Chomsky 1981a, 1981b, 1986), i.e., to describe the cognitive system at the basis of the language capacity.

1.1 Knowledge of Language

Linguistic knowledge is represented as a system of constraints, a grammar, which defines all and only the possible sentences of the language (Emonds 1980, Ross 1967, Perlmutter 1971). An adequate grammar of a specific language must not generate ungrammatical sentences, i.e., sentences which are not acceptable to the native speakers of that language. For instance, the grammar of English will generate (1a) and must not generate (1b).

(1a) John always eats chocolate.
(1b) *John eats always chocolate.

1.2 Language Acquisition

The generative linguist also poses the question of language acquisition (see *First Language Acquisition: Cross-linguistic*). Linguistic knowledge is acquired fast and at an early age. This leads to the hypothesis that the system underlying it must be relatively simple. The speed and ease of language acquisition is all the more striking because it is achieved on the basis of an input which is deficient in several respects:

(a) The linguistic input is not homogeneous but varies from one learner to the other.

(b) The input contains ungrammatical material (performance errors, hesitations).

(c) The child acquires linguistic knowledge which he or she cannot infer solely on the basis of the input.

The following illustration of this so-called 'negative evidence' problem is from White (1989). In each sentence, the subordinating conjunction *that* may be omitted.

(2) I think (that) John will leave.
(3) Who do you think (that) John will invite?
(4) The woman (that) you met last week is my sister.

If data such as (2)–(4) constitute the input for the acquisition process, the child will be entitled to infer generalization (5).

(5) The subordinating conjunction *that* may be omitted in English.

However, the subordinating conjunction *that* cannot always be omitted in English (6, 7), and, conversely, there are sentences in which *that* is ungrammatical (8).

(6a) That Mary should be invited is surprising.
(6b) *Mary should be invited is surprising.
(7a) She could not get over the fact that Mary was invited.
(7b) *She could not get over the fact Mary was invited.
(8a) *Who did you say that got there first?
(8b) Who did you say got there first?

How does the native speaker acquire the knowledge that (6b), (7b), and (8a) are not acceptable English sentences, while (6a), (7a), and (8b) are acceptable? All the sentences in (6), (7), and (8), including (6b), (7b), and (8a), are compatible with (5). Being ungrammatical, (6b), (7b), and (8a) are not available in the input. The input contains sentences such as (6a), (7a), and (8b), which are all compatible with (5). For the native speaker to infer that (6b), (7b), and (8a) are incompatible with English grammar, and that (5) is thus an incorrect generalization, the input would have to contain negative evidence, i.e., the information that (6b), (7b) and, (8a) are ungrammatical. Such negative evidence is not available in the input: the non-occurrence of example such as (6b), (7b), and (8a) in the input is not sufficient to establish their ungrammaticality. A pattern may be absent from the input because it is ungrammatical (e.g., (1b)), but it may also be absent because it is stylistically marked.

The deficiency of the input leads to the hypothesis that language acquisition is not based simply on exposure.

2. Universal Grammar: Principles and Parameters

The hypothesis was formulated that a theory of language acquisition must assume a genetically-determined language acquisition device, which mediates between the data in the input and the grammar of the language which the child is to attain. In the *Principles and Parameters Theory* (Chomsky 1981a, 1981b, 1986) abstract universal principles which underlie different linguistic patters are factored out and attributed to Universal Grammar (UG), the innate language faculty. UG guides language acquisition by constraining the possible grammars which can be formulated on the basis of the input. Two types of UG principles are distinguished: invariant principles (see Sect. 3) and parameterized principles (see Sect. 4).

Invariant principles are rigidly fixed and hold universally. For instance, the Empty Category Principle constrains omission of constituents in a sentence. Hence, the application of generalization (5), inferred from the input, is constrained by the Empty Category Principle. Other invariant principles are those which

impose constraints on syntactic structure (see Sect. 3), and the (still controversial) principle that all syntactic variation is driven by morphological variation (see Sect. 5).

Other principles of UG are not fully fixed cross-linguistically; their interaction with the input will be at the basis of cross-linguistic variation. The variable components of UG are referred to as 'parameters.' Parameters are binary: each parameter offers a choice between two options or parameter-settings. Depending on the parameter-setting, a language will or will not display the corresponding syntactic property. For example, the pro-drop parameter determines whether a subject pronoun may be omitted in a given language or not (Sect. 4.1), the verb-movement parameter determines if a verb can be separated from its direct object (Sect. 4.2).

The principles and parameters theory thus captures cross-linguistic invariance (in terms of the universal constraints), and cross-linguistic variation (in terms of the language-specific parameter-settings). The theory reduces acquisition of syntax to parameter-setting on the basis of the linguistic input, the so-called 'triggering experience.'

3. *Invariant Principles: Phrase Structure is Endocentric and Binary Branching*

According to one invariant principle of grammar, syntactic structure is characterized by two properties:

(a) structure is endocentric: a syntactic unit or a constituent is organized around a head; and

(b) constituents are always formed by combining two units.

Based on (a) and (b), all constituents have the binary branching format represented in Fig. 1, where XP is the constituent, X is its head. The head X combines with another constituent, here ZP. ZP is itself built according to the X-bar format. The unit formed by X and ZP (X′) combines with the constituent YP to form XP. The format in Fig. 1 is taken to be universal. Cross-linguistic variation in sentence patterns is determined either by the option of omitting certain constituents (cf. Sect. 4.1) or by moving them (cf. Sect. 4.2).

4. *Parameters*

4.1 *The* Pro-*drop Parameter*

In terms of the X-bar format in Fig. 1 the sentence is conceived of as a constituent whose head is the inflection of the verb (I).

The *pro*-drop parameter (see Fig. 2) determines the variation between languages like Italian, in which the subject pronoun can be omitted, and those like

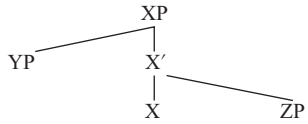

Figure 1
X-bar format of constituent structure

English, in which it cannot be omitted (Jaeggli and Safir 1986, Rizzi 1986).

(9a) *Parla inglese*. (*Italian*)
(9b) *Speaks English.

The *pro*-drop parameter is formulated in terms of the relative richness of the verbal inflection. The inflectional paradigm of the Italian is rich: it has a distinct form for each person/number combination (*parlo*, *parli*, *parla*, *parliamo*, *parlate*, *parlano*, 'speak,' present tense). English inflection is poor, with minimal variation (*speak*, *speaks*). Rich inflection sets the *pro*-drop parameter positively, allowing the language in question to omit subject pronouns.

The concept of parameter should not be taken to mean that each aspect of cross-linguistic variation is related to a distinct parameter. A parameter underlies a cluster of properties. For instance, it is generally held that the *pro*-drop also underlies the contrasts in (10)–(12). As shown by (10), Italian weather verbs such as *rain* cannot have a subject, while English weather verbs must have a subject. While (11) shows that the subject may follow the verb in Italian, an option unavailable in English. Questioning the subject of a subordinate sentence in Italian (12a) is compatible with the presence of the conjunction *che*, while the English analogue in (12b) is unacceptable; (12b) is rendered grammatical by omitting *that* (cf. (8b)).

(10a) *Piove./*Esso piove*.
(10b) *Is raining./It is raining.
(11a) *Arriva Gianni*.
(11b) *Arrives John.
(12a) *Chi credi che abbia telefonato*?
(12b) *Who do you believe that has phoned?

4.2 *Verb Movement*

The distribution of verbs in the clause provides a second illustration for the interaction between verb morphology and syntactic variation. In English (13a), the inflected verb *eats* immediately precedes the direct object *chocolate*. The verb follows the adverb *often*. In French (13b), the finite verb is separated from its direct object *du chocolat* and it precedes the adverb *souvent*.

(13a) John often eats chocolate.
(13b) *Jean mange souvent du chocolat*. (*French*)
*John eats often chocolate.

The verb and its direct object form a semantic unit. While the English verb in (13a) remains adjacent to its

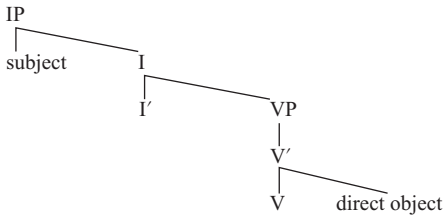

Figure 2
The structure of the clause

object, the French verb *mange* in (13b) is separated from the object. Based on the structure in Fig. 2 it is proposed that French (13b) involves leftward verb-movement from the position V to the inflectional head (I).

The contrast between Icelandic and Danish in (14) is the same as that between English and French in (13). The Danish verb remains adjacent to its object; the Icelandic verb is separated from its object by the adverb *oft*.

(14a) *at Johan often spiser tomater*. (*Danish*)
that John often eats tomatoes.

(14b) *að Jón borðar oft tomata*. (*Icelandic*)
*that John eats often tomatoes.

Verb-movement, as displayed by French (13b) and by Icelandic (14b), is also related to the inflectional paradigm of the verb (Pollock 1989, Holmberg and Platzack 1995). French and Icelandic have a relatively rich inflection, whereas English and Danish have impoverished inflection. In languages with rich inflection, the verb can move away from its direct object.

The kind of paradigm which allows verb movement is related but not identical to the kind of paradigm which triggers a positive setting of the *pro*-drop parameter as discussed in Sect. 4.1. Verb-movement to the inflectional head requires fewer distinct forms in the paradigm than pronoun omission (cf. Vikner 1997, pp. 192ff.). Verb-movement is possible in languages like French or Icelandic which do not allow the subject pronoun to be omitted.

5. *Parameters and Language Acquisition*

The parameter determining verb-movement can be set by the language learner on the basis of inflectional information. However, the inflection is not the only positive evidence that the language learner gets in order to set the parameter. Word-order patterns in the input (i.e., the sequence verb–adverb or adverb–verb) also provide robust positive evidence allowing the learner to determine the parameter-setting. Thus, two types of evidence are available to the language learner for fixing the parameter: syntactic evidence (word-order) and morphological evidence (inflection).

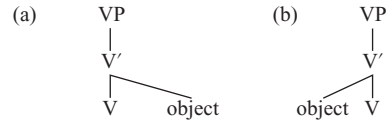

Figure 3
Word-order variation

5.1 *Morphology and the Setting of Parameters*

Both parameters of cross-linguistic variation discussed so far are related to inflection. The question arises whether all parameters can be stated in terms of inflectional variation. One area of cross-linguistic variation whose definition remains subject to controversy concerns the contrast between so-called VO languages such as English, in which the verb precedes its direct object, and OV languages such as Dutch and German, in which the direct object precedes the verb (see *Word Order*):

(15a) I think that John knows this house.

(15b) *Ik denk dat Jan dit huis kent*. (*Dutch*)
*I think that Jan this house knows.

According to the more traditional approach, the variation in (15) is due to a parametric variation in the underlying structure of these languages. In this view, the universal principles of structure (Fig 1) only determine hierarchical relations, leaving the linear order of the constituents to be determined parametrically. Thus the structure of the VP would come out as Fig. 3a for English and as Fig. 3b for Dutch.

The universal base approach, on the other hand, assumes that structure determines linearization and that the verb always precedes the object; (Fig. 3b) is not a possible base structure. In this view, the OV order of Dutch (15b) is derived by a leftward movement of the object. Those adopting the universal base approach define the trigger of the leftward movement of the object in Dutch (15b) in terms of a morphological property of the object. Since the Dutch object does not, at first sight, display any visible properties to distinguish it from its English counterpart, it is proposed that morphological variation need not correspond to an overt manifestation; morphological properties at the basis of cross-linguistic variation may also be abstract.

5.2 *Diachronic Change and the Trigger for Parameter (Re-)setting*

Two types of overt evidence in the input allow the learner to formulate his/her grammar: syntactic evidence (the observed word-order sequence verb–adverb–direct object vs. adverb–verb–direct object, the observed omission of the subject) and morphological evidence (the verbal inflection). The question

arises whether either of these two kinds of evidence is primary in parameter setting.

Inflectional endings in a language get weakened and disappear in the course of time. As a result, the morphological evidence for a given parameter-setting disappears. The development of English (see *Historical Linguistics: Overview*) suggests that the loss of inflectional morphology led to the loss of leftward verb movement (cf. Roberts 1993, Pollock 1989). Changes in the inflection have led to a re-setting of a parameter. Apparently, when the morphological evidence gets lost the syntactic evidence alone is insufficient for triggering the relevant parameter-setting (see Vikner 1997, pp. 205ff. on Scandinavian). The loss of the morphological trigger for a particular word-order does not necessarily lead to an abrupt change in the word-order patterns of the language. A syntactic trigger (i.e., the word-order pattern itself) can maintain the old system for some time, leading to gradual, rather than abrupt change.

See also: First Language Acquisition: Cross-linguistic; Language Acquisition; Linguistics: Incorporation; Second Language Acquisition

Bibliography

Chomsky N 1981a *Lectures on Government and Binding*. Foris, Dordrecht, The Netherlands

Chomsky N 1981b Principles and parameters in syntactic theory. In: Hornstein N, Lightfoot D (eds.) *Explanations in Linguistics*. Longman, London, pp. 123–46

Chomsky N 1986 *Knowledge of Language*. Praeger, New York

Emonds J 1980 Word-order in generative grammar. *Journal of Linguistic Research* **1**: 33–54

Haegeman L (ed.) 1997 *The New Comparative Syntax*. Addison, Wesley, Longman, London

Holmberg A, Platzack C 1995 *The Role of Inflection in Scandinavian Syntax*. Oxford University Press, Oxford, UK

Jaeggli O, Safir S 1986 *The Null Subject Parameter*. Kluwer, Dordrecht, The Netherlands

Lightfoot D 1991 *How to Set Parameters*. MIT Press, Cambridge, MA

Perlmutter P 1971 *Deep and Surface Structure Constraints in Syntax*. Holt, Rinehart and Winston, New York

Pollock J-Y 1989 Verb movement, universal grammar and the structure of IP. *LI* **20**: 365–424

Rizzi L 1986 Null objects in Italian and the theory of pro. *LI* **17**: 501–57

Roberts I 1993 *Verbs in Diachronic Syntax*. Kluwer, Dordrecht, The Netherlands

Ross, J R 1967 Constraints on variables in syntax. Ph.D. thesis, MIT

Vikner S 1997 V-to-I movement and inflection for person in all tenses. In: Haegeman L (ed.) *The New Comparative Syntax*. Addison, Wesley, Longman, London, pp. 189–213

White L 1989 *Universal Grammar and Second Language Acquisition*. John Benjamins, Amsterdam

L. Haegeman

Literacy and Illiteracy, History of

The period from the early sixteenth century up to the end of the eighteenth century saw a veritable revolution in European cultural history. In the Middle Ages, it was taken for granted that only a select few knew how to read and write and that many of them made their living out of these skills as scribes and clerks. By the late eighteenth century, however, in the most developed regions of Europe, about 50 percent of men were able to acquire information from reading the newspapers, and the number of people who could write was steadily rising. The skills of reading and writing and, beyond that, the use of literacy, were rapidly gaining ground, to the detriment of oral communication. The body of written documents increased drastically; literacy made its presence felt more and more strongly throughout society. Printed books grew in number while becoming cheaper, and more things were thought worthy of being committed to paper. Literacy extended its power to ever-newer sectors of a developing state administration and bureaucracy in early modern Europe. Surveys and certificates were drawn up, and the use of written documents gradually squeezed out verbal evidence. Inevitably, people's mentality and ways of thinking underwent a simultaneous change.

Orally transmitted knowledge and the memory of the elderly lost a good deal of their former prestige, while respect for information set down in writing, which increasing numbers of people were able to decipher, strengthened. With the spread of calendars and printed maps, the concepts of space and time altered; numerically articulated, measurable time came to replace an image of a past drawn with vague, uncertain outlines. Acquiring writing skills alone did not necessarily involve a change in mentality. Owing to the gradual spread of literacy in society, however, more and more people were able to read the dates on printed works, and this affected their way of thinking. People knew what the date was and how old they were, and tied events to specific days, months and years. The traditional concept of time had rested on oral tradition and memory, associating the passage of time with events such as wars and plagues, and the changing seasons. This slowly gave way to the new concept based on literacy, supported by written documents. This article tracks the development of literacy, especially in Europe and North America in the early modern period, the most important time in history of literacy, marking the period of transition from restricted to mass literacy in the West.

1. Literacy: Definitions and Mutual Relations

Functional literacy, i.e., the capacity to read a newspaper, sign a check and write a short postcard at least,

was an essential and indispensable precondition of the functioning of nineteenth and twentieth century society. These skills were, however, lacking in the majority of the population in early modern Europe.

Literacy is much more than just the capacity to read and write. It is the constant practical use of writing and of the written record. Literacy represents a different way of seeing the world and storing information from that used by illiterate people.

During the 1990s very important research was carried out on literacy in ancient societies, evaluating in a new light data that had been available for centuries, but not regarded as sources of literacy: descriptions of reading and writing, orthography of mural epigraphs, and so on.

Literacy remains a very real problem in large parts of the world outside the West (and in parts of the Western world where there are poor schools, and where the phone and television have completely replaced letter writing and newspaper reading, etc.). However, literacy played a central part in European cultural and social history in the centuries between 1500 and 1800. Literacy is typically a phenomenon of transition—these centuries saw in Europe the transition from a restricted medieval literacy to modern mass literacy. After the invention of the printing press in the mid-fifteenth century, the quantity of written documents, from books to newspapers and leaflets, grew very quickly, and this encouraged more and more people to learn to read and, to a lesser extent, to write.

Illiteracy is by no means a synonym for lack of culture. The oral tradition and the oral culture of traditional societies is as important and precious as the written culture and literate world of modern societies. The oral tradition did not disappear with the onset of literacy: in early modern Europe there was a coexistence of the two cultures. The oral culture was not, however, an isolated island in the dominating literate world; illiterate people using orality lived very much in the middle of a society where the most important decisions were put into a written form, and where more and more people could read and write. There was a continuous interaction between the two worlds. Illiterates used written documents that they could not decipher, e.g., passes and documents concerning ownership of their houses and land. Often the content of lost written documents was conserved only by orality, e.g. those concerning the migration of population or the origin of settlements. Such information was often kept in memory only by the tales of the old people in the villages who may once have read such lost documents.

General illiteracy influenced the exchange of information, making it difficult for the masses to maintain contact with the world outside their town or village. In the twenty-first century it is natural for a literate man to write a letter and read a notice or a brochure. For the unlettered, though, this causes grave difficulties and requires the assistance of others. Until the widespread use of the telephone, large sections of society practiced writing mainly through corresponding with relatives. Even illiterates tried to exchange letters, which meant sharing their innermost thoughts with those who offered to commit them to paper.

2. *Measuring Literacy: Problems of Methodology*

The best sources for measuring literacy are census records. During a census every citizen was asked if he or she could read and write, or only read, or neither read nor write. However, such censuses were carried out only from the second half of the nineteenth century. As literate people generally came from the upper, richer strata of society, they had a higher general life expectancy than illiterates. For this reason we should not draw retrospective conclusions on literacy from the age-groups of censuses, as it is not absolutely certain that literate and illiterate people died in the same proportion.

We can estimate the rates of literacy before the beginning of censuses by consulting descriptive sources and statistics of crosses and signatures. Descriptive sources such as travelogs, reports of state officials, data on elementary education, book production, and book ownership, cannot furnish reliable quantitative data on the literacy rates. However, these are valuable indirect sources because we can analyze the competence of the writing. It is only exceptionally recorded that someone had authenticated a document 'by alien hand, not knowing the letters'; in most cases, only a cross by way of signature or some characters suspiciously identical with those in other signatures show that somebody was illiterate. In sources such as witness depositions, however, hundreds of nobles, peasants and burghers declared that they could not read, or this fact about them was recorded by travellers, state administrators, etc.

Good statistics on literacy rates in early modern Europe can be drawn up, however, only on the basis of enumeration of people who signed their names or who could only make a cross instead of a signature. For this purpose we need a representative cross-section of the society where the social rank and geographical position (and if possible, age) of the persons in question can be determined. Often, however, only males and only the better-off figure in the documents, or only one professional group is represented.

Statistical data hardly reflect the real state of affairs concerning literacy. The method of counting crosses and signatures rests tacitly on the supposition that all those whose signatures have been found could indeed write, while those who left merely a cross on documents were all lacking the skill of even writing their names. Both suppositions could be erroneous, because examples abound of counterfeit signatutes being

entered on behalf of illiterate people and of crosses made by people well trained in penmanship. This reflects the mentality of a given society: if literacy has already acquired prestige, even illiterates try to fake writing capacity; however, if the great majority in society are illiterate, handwriting has no authentication function and thus even village notaries make only a cross, even if they can write properly. This would not come to light unless we have more than one sample of the signatures or crosses left behind by the same person, which in turn presupposes a recurrence of the same name, whose bearer gave a signature at one time and a cross at another. The statistical data concerning crosses and signatures, therefore, reveal the attitudes of a community toward literacy, rather than the concrete spread of literacy: they speak about the community expectations concerning literacy, the pressures on semi-literate persons to sign the documents themselves, and on the genuinely unlettered to forge a signature. Such mentality only emerged, though, if literacy was on the rise in a society. In their totality, therefore, these statistics more or less do mirror the advance of literacy. There is no doubt that if we divide society strictly into literate and illiterate people, the major part remains concealed: the semi-illiterate layer of society that did learn something of the alphabet and later, in accordance with the community's wishes, either made crosses or wrote their name on official documents. The same applied to reading, too: those who were taught the letters at school but hardly ever read, could profit from the printed prayerbooks whose text they knew by heart (extensive reading). However, they did not fare well when they had to read a handwritten circular whose contents were unknown to them (intensive reading). The unlettered and the literate were in fact never separated as sharply as may be suggested by the rows of crosses and signatures. No clear-cut distinction can be made between fully illiterates, who could neither read nor write, and semi-literates, who could read but not write: both groups left crosses on documents. Yet there is no greater gap in levels of education than that separating those able at least to read, to gather information from written texts, from those to whom the world of letters is literally a closed book.

The greatest number of written sources and of sources on literacy are to be found where the levels of literacy were the highest, and obviously this greatly influences research on literacy overall. In England from the seventeenth century onwards witnesses authenticated their testimony by putting their signature or mark on the records, and people signed oaths of fidelity en masse. In several Western European countries, spouses and their witnesses also confirmed a marriage by signing or marking the marriage register. This explains why the marriage register is a major source for the spread of literacy. In France marriage registers had to be signed from 1686, in England from 1754. The city of Amsterdam, one of the most literate places of the world in this time, required marriage registers to be signed from as early as 1578. In 1630, 43 percent of bridegrooms and 68 percent of brides made a cross; 100 years later in 1730, the proportion is 24 percent and 49 percent respectively. In 1780 only 15 percent of men and 36 percent women could not sign their names at their marriage.

The conquests of Napoleon Bonaparte open up new possibilities for the historians of literacy in Germany and Italy. These two countries, so important for the cultural history of early modern Europe, lacked similar mass sources in the sixteenth to eighteenth centuries for evaluation of literacy levels. After they became part of Napoleon's empire, however, the French system was introduced—thus, around 1800, we have a statistically relevant quantity of crosses and signatures from many regions of Germany and Italy, as they were parts of France at this time. In Central and Southern Europe, however, only the minister conducting the marriage ceremony was expected to write in the register, and it was not compulsory for eyewitnesses to sign their accounts in court. Witnesses confirmed their accounts only by taking an oral oath, without signing the document. This fact in itself is indicative of the lesser role literacy and writing played in these regions compared to Western Europe.

Reading, in contrast to writing, leaves no traces behind; that is why researchers into literacy almost invariably examine the ability to write—or more specifically, to produce a signature—and give up trying to gauge reading abilities as a hopeless venture. In several European countries, however, an extraordinary source lends scope for the examination of reading skills en masse: the Lutheran pastors in those areas made an annual survey of reading skills in their flock from the late seventeenth century. In the history of European literacy, a position of eminence is certainly due to Sweden (with Finland and Latvia, which still belonged to it at the time), Zurich and vicinity in Switzerland, and certain regions in northern Germany (seventeenth and eighteenth centuries) where the Lutheran faith thrived. In the course of Lutheran church visitations in these territories, records were made of, among other things, the reading skills of all adults, even indicating the grades—so an otherwise non-manifest skill can be examined on the basis of a large sample. The Lutheran church believed it to be of great importance for believers to read the catechism and the Bible. Those reluctant to learn to read were denied communion at Easter. The Swedish model is especially interesting, as the church was pushing for better reading skills without enforcing writing capacities. As a result there was a huge gap between reading and writing abilities—in the eighteenth century almost everybody (around 90 percent) was able to read, but only few (around 10 percent) could write. Female literacy was rising, too, and the role of the mother as a teacher at home became very important, especially in the nineteenth century.

The Lutheran church records are, however, not the only sources for examining reading capacities before the age of censuses. In Hungary, when the county administration investigated claims to noble title, it asked the nobleman making the claim whether he had read, or only seen, the lost letters patent that would prove his nobility. In three different regions of eighteenth-century Hungary, 71 percent, 74 percent and 55 percent respectively of the lower noblemen declared that 'as they cannot read at all' they had only seen, but not read, the lost document.

3. *Literacy in Europe*

England together with the Netherlands, was the most literate country in the world during the period. In the mid-seventeenth century, during the stormy years of the Civil War, Englishmen were required to sign oaths of allegiance several times; consequently there are ample data available concerning the state of literacy in various parts of the county, in the form of thousands of signatures and signs. On examining these, David Cressy (1980) has found that in England in the 1640s, some 70 percent of men and 90 percent of women were illiterate. As regards the sixteenth-century situation, only estimates can be made. Cressy avers that in the middle of the sixteenth century, 80 percent of men and 95 percent of women could neither read nor write. The percentages around 1500 may have been 90 and 99 respectively, though this is a mere surmise as even in England no statistical data about literacy can be culled so early. The situation in eighteenth-century Britain is much better known. According to David Cressy's estimates based on the available sources, at the beginning of the century 55 percent of men and 75 percent of women could write their own names. In 1754, it was stipulated by law in Britain that the marriage register had to be signed by both the bridegroom and the bride. From that time onwards, therefore, the development of literacy can be traced on the basis of precise data. By the middle of the eighteenth century, the majority of men knew how to write, only 40 percent being unlettered, while the rate amongst women was 60 percent. In the second half of the eighteenth century, however, this dynamic progress came to a sudden halt. Until the early nineteenth century, illiteracy stopped falling in Britain. The social transformation that was taking place as a consequence of the Industrial Revolution was probably to blame for this—inner migrations and early child labor were not really conducive to literacy. In the second half of the nineteenth century, though, illiteracy in Britain all but disappeared. Between 1850 and 1911, the proportion of illiterate men decreased from 30 percent to one percent, and that of women from 45 to less than two percent.

France is the best researched of all countries in terms of literacy in the early modern age. From the age of Louis XIV, it was compulsory for brides, bridegrooms and witnesses to put their signature in the register at weddings, thus leaving a wide scope for researches into literacy. In France in the eighteenth century the number of literate people almost doubled, even though at the time of the outbreak of the French Revolution they were in the minority, the proportion of literate people in France grew from 21 percent in 1686–90 to 37 percent in 1786–90. Frenchwomen also lagged behind men—literacy among them grew from 14 to 27 percent within 100 years, while amongst the men it rose from 29 to 47 percent. However, averages that apply to entire countries hide what is really relevant—the huge regional differences that were so characteristic of this large country. That imaginary dividing line connecting Geneva and Saint Malo, which in the seventeenth century separated the literate northern and eastern parts of France from the more backward western and southern parts, was still clearly visible on the map of literacy at least a century later, and in the case of women's literacy, as late as 1866. This indicates that regional imbalances in the spread of literacy, as an important index of development, take a long time to change. When the French Revolution broke out, there were huge regional differences in the literacy of men. In Normandy and in the countries north and east of Paris, 90 percent bridegrooms signed the register in Brittany and on the Atlantic coast, their proportion was less than ten percent. In the secluded regions of the Massif Central, the percentage of men who put their signature in the register was 16, of the women a mere four while in Alsace-Lorraine the percentages were 83 and 44 respectively. The differences did not diminish until the second half of the century.

By the end of the nineteenth century, illiteracy had in effect been wiped out in France. While in 1854, 13 percent of bridegrooms were unable to sign the marriage register, the number decreased by 1900 to a mere five percent. Within the same period the proportion of illiterate brides decreased from 46 to six percent. Compulsory schooling put an end to the difference between male and female literacy, which had been a marked feature in earlier centuries.

The newly unified Germany presented a very mixed picture from the point of view of literacy, too. In the Prussian kingdom in 1871 10 percent of men above 10 years could neither read nor write. However, this average was calculated again on the basis of very different regions: the proportion was only 2 percent in Hessen and 3 percent in Bradenburg, while in the Polish-inhabited region around Posen (Wrocław) one-third of the males (32 percent) could not even read.

For lack of relevant researches, there are no statistical data on literacy available from the Austrian provinces of the Habsburg Empire in the sixteenth to eighteenth centuries. After the modern census had

been introduced, the available data testify to huge discrepancies between the provinces, dwarfing the differences between the various French or Prussian regions. In 1900 in the most developed provinces of the Habsburg Empire, in Upper and Lower Austria, and in Bohemia and Moravia, the rate of illiteracy in the population aged 10 years and above was less than 4 percent. In the same time, in Galicia 56 percent, in Bukovina 66 percent and in Dalmatia no less than 73 percent of the population above 10 was illiterate.

There are ample sources in Italy on the spread of education and culture and on schools in the early modern period; however, except for the provinces conquered by Napoleon, no statistics are available about literacy before the introduction of censuses in the second half of the nineteenth century. In the vicinity of Turin in 1710, excluding the town itself, 94 percent of brides and 81 percent of bridegrooms were illiterate; by 1790, the percentages had decreased to 70 and 35 respectively. In the Neapolitan kingdom in Campania province, in the mid-eighteenth century, 76 percent of those submitting tax returns could put only their cross on the document, even though only the well-off, and then only men, were expected to prepare such a return at all. The marked difference between the northern and the southern provinces is evident even from this single piece of information. Censuses reveal, on one hand, that in the whole of Italy in the late nineteenth century, illiteracy, which earlier had reached astonishingly high proportions, was rapidly diminishing; on the other hand, however, the gap between northern and southern Italy was huge. It had not disappeared by the early twentieth century—in 1901 in the northern region of Piedmont 18 percent of the population above six was illiterate, in Lombardy 20 percent, in Veneto 28 percent, while in Sicily the figure was 71 percent, in Sardinia 68 percent and in Calabria 79 percent.

In Poland we can observe huge regional differences, too. In the Voivodeship of Cracow at the end of the sixteenth century, the landed nobility was largely literate: 96.7 percent of the nobles who bore the title *generosus* and usually owned several villages, could sign their name, and the percentages remained unchanged over the next two centuries. Among women, however, literacy rose rapidly in the same period: while in the late sixteenth century, 51 percent were unlettered, by the seventeenth century only 22 percent remained so. Women continued to catch up with men in the eighteenth century, when a mere 15 percent put a cross on documents instead of signing their names. In the lower ranks of landed gentry or *nobilis*, who owned one or at most two villages, the proportion of illiterates made up 24 percent in the sixteenth century and 16 percent in the following century. In another study, the literacy of Polish noblemen was examined on the basis of tax returns submitted in and around Cracow. In 1564–64, 43 percent of the noblemen were illiterate. The vast Polish kingdom, however, comprised regions with very different levels of development. In the Voivodeship of Sandomierz in the seventeenth century we find circumstances very similar to those of the Voivodeship of Cracow: 99 percent of the noblemen bearing the title of *generosus* and 75 percent of the *nobilis* could write. On the other hand, the Voivodeship of Lublin lagged behind: while those titled *generosus* were almost entirely literate, only 65 percent of the women could write. Of the *nobilis*, compared to the 84 percent in and around Cracow, only 70 percent signed the documents, while barely one out of 10 women could write their names. Even this proved to be a highly developed state of affairs when compared to Eastern Galicia, where eight percent of the magnates (*magnifici*), eleven percent of the *generosi* and 72 percent of *nobili* put a cross instead of their names as late as the early eighteenth century. In 1764, two-thirds of the participants of the diet called *sejm* held here could not authenticate documents by signing them.

In Vas county, Hungary, in the eighteenth century, only 2.5 percent of peasants could sign their names, and among the lower strata of the gentry only 26 percent could sign their names with 74 percent putting only a cross on documents. The first census to give an overall picture about the literacy of the population in the Hungarian kingdom was carried out in 1870. At this time three-fifths (58 percent) of the population aged above six in the historical Hungarian kingdom (part of the Habsburg Empire, consisting of Hungary, Transylvania, and Croatia) could neither read nor write. The figure for Hungary alone was 51 percent; 78 percent in Transylvania; and over four-fifths of the population (84 percent) in Croatia and Slavonia. There were huge differences between the various nationalities, however, in 1890 63 percent of Germans and 53 percent of Hungarians could read and write, but only 9.7 percent of Ukrainians living in Hungary, 14 percent of Romanians and 30 percent of Serbians. If Central Europe was lagging very much behind the developed Western parts of Europe, the difference between Central Europe and Russia, and the countries of the Balkan peninsula, was as huge.

No reliable data are available on the measure of illiteracy in Russia for the sixteenth to eighteenth centuries, but Jeffrey Brooks, historian of Russian literacy, gave his 1977 work on the period between 1861 and 1917 a telling title: *When Russia Learnt to Read*. At the 1897 census, only 21–29 percent of men and 13 percent of women could read. Literacy attained the highest proportions in the Polish provinces of the empire, even though this favorable position was relative indeed: 29 percent of men and 23 percent of women had a passive knowledge of the letters. In the European part of Russia, 28 percent of men and 10 percent of women could read, while the relevant percentages in Siberia were only 16 and three respectively. Development in the 50 years preceding World War I was rapid, however while in 1880, four-

fifths of new conscripts did not know the letters, in 1913 only 32 percent of these young men, who were just above school age, were illiterate.

The line of the River Danube and the Carpathian Mountains, which separated Hungary from the Balkan peninsula, was also a demarcation line between two worlds as regards literacy. Missionaries and travelers in seventeenth century Moldavia and Wallachia noticed that hardly anybody knew how to read or write, and even some of the Greek Orthodox bishops could read only with difficulty. 'Of the nobles and the tradesmen, few know the Ruthenian or the Slav script,' and even the majority of priests were illiterate according to the Jesuits, although they may not have been impartial. The population of Serbia was largely unlettered as late as the early nineteenth century. Hardly any of the leaders of the Serbian uprising could read or write; the ruling Prince Milos was probably the only illiterate head of state in Europe at the time. According to contemporaries, when Serbia became independent, neither the priests nor the monks could read and write. There was little improvement in this situation in the following decades: even the 1866 census found that 96 percent of the population was illiterate; at the turn of the century, the number was 83 percent. The Balkan Peninsula remained very much a region of illiteracy even at the turn of the twentieth century: around 1900, of the population above six, 72 percent in Bulgaria, 60 percent in Greece, 78 percent in Romania, and 87 percent in Bosnia were illiterate.

For these parts of Europe, literacy remained very much a problem of the twentieth century—the Bolshevik Party launched huge literacy campaigns after coming to power in Russia. In North America there were very high levels of literacy for New England—male literacy rose here from 70 percent in 1710 to 90 percent at the end of the eighteenth century. However, as these data were calculated on the basis of the wills of the wealthy, doubts must be raised about this very high proportion of literates. The 1840 census showed huge differences within the country: around four percent illiterate in the state of New York, compared with 20–28 percent in Georgia, North Carolina, Tennessee, and Arkansas. However, the African-American population showed high levels of illiteracy even after slavery had been abolished: in 1870 some 80 percent of adult blacks were illiterate, and the figure was 44 percent in 1900. Massive immigration into the USA from Europe before the World War I, imported mass illiteracy: in 1909 99 percent of Danish or Scottish workers in the United States were literate, but only 45 percent of Portuguese and 43 percent of Polish workers could read and write at this time.

In other parts of the world, literacy is a problem of today and tomorrow. Around 1960, the time of dissolution of the colonial empires, 81 percent of the population of Africa, and 55 percent of Asia's population were illiterate. According to various estimates, at this time roughly 40 percent of the world's inhabitants still could not read or write. There remains much to do in 1992 the illiteracy figures for Burkina Faso were 86 percent, in Bhutan 85 percent and in Afghanistan 76 percent. Illiteracy remains a problem for mankind even in the third millennium.

See also: Cultural History; Education, History of; Illiteracy, Sociology of; Literacy and Orality; Literacy Education; Media and History: Cultural Concerns; Social History; Social Inequality in History (Stratification and Classes)

Bibliography

Arnove-Harvey R F (ed.) 1987 *National Literacy Campaigns*. J Graff, London/New York

Brooks J 1985 *When Russia Learned to Read. Literacy and Popular Literature 1861–1917*. Princeton, NJ

Cavallo G (ed.) 1977 *Libri e lettori nel medioevo*. Rome

Chartier Roger (ed.) 1985 *Pratiques de la lecture*. Marseille

Chartier R (ed.) 1987 *Les usages de l'imprimé*. Paris

Cipolla C M 1969 *Literacy and Development in the West*. Harmondsworth, UK

Cressy D 1980 *Literacy and the Social Order. Reading and Writing in Tudor and Stuart England*. Cambridge

Darnton 1971 *Reading, Writing and Publishing in Eighteenth Century France*. Daedalus

Furet F-Jacques O 1977 *Lire et écrire: l'alphabétisation des Français de Calvin à Jules Ferry*. Paris

Goody J (ed.) 1968 *Literacy in Traditional Societies*. Cambridge

Graff H J 1987 *The Legacies of Literacy*. Bloomington, IN

Houston R 1985 *Scottish Literacy and the Scottish Identity. Illiteracy and Society in Scotland and Northern England 1600–1800*. Cambridge

Houston R 1992 *Literacy in Early Modern Europe: Culture and Education (1500–1800)*. London

Johansson E 1981 The history of literacy in Sweden. In: Graff H J (ed.) *Literacy and Social Development in the West*. Cambridge

Marchesini D 1992 *Il bisogno di scrivere. Usi della scrittura nell'Italia moderna*. Rome, Italy

McKitterick R 1992 *The Uses of Literacy in Early Medieval Europe*. Cambridge

Muchembled R 1978 *Culture populaire-culture des élites dans la France*. Paris

Parker G 1980 An educational revolution? The growth of literacy and schooling in early modern Europe. *Tijdschrift voor Geschiedenis* **93**

Pelizzari M R 2000 *La penna e la zappa. Alfabetizzazione, culture e generi di vita nel Mezzogiorno moderno*. Salerno, Italy

Ruwet J Y W (ed.) *L'analphabétisme en Belgique*. Louvain, Belgium

Sanderson M 1972 Literacy and social mobility in the industrial revolution in England. *Past and Present* **56**

Schofield R S 1981 Dimensions of illiteracy 1750–1850. In: Graff H J (ed.) *Literacy and Social Development in the West*. Cambridge

Spufford M 1979 First steps in literacy: The reading and writing experiences of the humblest seventeenth-century spiritual autobiographers. *Social History* **4**

Stone L 1969 Literacy and education in England 1640–1900. *Past and Present* **42**

Tóth I G 2000 *Literacy and Written Culture in Early Modern Central Europe*. Budapest, Hungary
Van de Woude A M 1980 Der alfabetisering. In: *Algemeine Geschiedenis der Nederlanden*. Haarlem, Holland, Vol. 7
Wyczanski A 1974 *Alphabétisation et structure sociale en Pologne au XVIe siecle*. Annales ESC 29

I. G. Tóth

Literacy and Orality

1. Definitions

Notions which are mutually dependent are called dialectical. This label applies to orality and literacy, 'literacy' cannot be conceived of without 'orality,' nor orality without 'literacy.' Hence, their respective meanings depend on how the opposite notion is assessed: if orality is given a positive value, literacy appears as negative, and vice versa.

To give some examples: Marshall McLuhan (1962) started from an oral society which was originally closed, which attached importance to the spoken word, and in which social roles were relatively fixed. He opposed this idealized form of community to Western literate societies which, having transformed oral language into a visible code, were seen as depreciating the spoken word, compelling us at the same time to cope with different social roles. In this context, a particularly negative influence is attributed to the invention of printing. Whereas Jack Goody and Ian Watt (1962) saw the transition from orality to literacy more in the sense of a profit-and-loss account, Walter J. Ong (1982) took an opposite stance underlining the positive aspects of literacy, thus suggesting the positive kind of innovation McLuhan was only willing to attribute to the New Age of television which, in his view, already occupies the place of literacy (Raible 1994).

The rash and overgeneralized judgments to be found in scholarly discussion of the 1990s suggest a more cautious attitude. The first question to be asked in this context is: what do 'orality' and 'literacy' mean? Here we need to distinguish a *medial* from a conceptual or cognitive aspect: the medial one is trivial—something we utter is either spoken or written. The conceptual aspect is a continuum, intertwining orality and literacy. This conceptual side needs some more consideration.

Communication by language is based on social convention. Hence, the utterances we engage in or exchange with others always reflect a socially accepted activity: somebody has to be welcomed; we engage in small talk with her or him, ask questions, give answers, make jokes, engage in gossip, and so on. Some more privileged individuals may pass judgment on somebody, make an injunction, or pronounce a verdict.

The terms in the preceding paragraph describe social activities accompanied—or even only made possible—by the use of language ('speech acts'). They represent a continuum of text types which can best be arranged on a scale, one end of which can be termed 'interactive text,' and the other, 'edited text' (Biber 1988); other labels are 'conceptually oral' and 'conceptually written,' i.e., highly planned and edited text (Koch and Oesterreicher 1994). For a similar distinction, Karl Bühler (1934) had already introduced a pair of Greek terms used by Wilhelm von Humboldt in the nineteenth century: language as *enérgeia* vs. language as *érgon*—i.e., language as an activity or process (involving present partners), as opposed to language as something created, produced, edited, e.g., a book. Increasing from one end of this scale to the other are, among other things, the degree of structuring and, at the same time, the degree of intersubjectivity. Conceptually scriptural texts should be accessible to a large and unspecified public.

The implications of the conceptual scale can best be shown within a three-level approach to language (see Table 1). Each one of these three levels should be conceived of as a scale, reflecting exactly the conceptual and cognitive scale holding between orality and literacy: on level (b), we encounter *text types* with an ever increasing degree of editing and planning. Stylistics and rhetoric belong to this level, rhetoric being, indeed, the art of planning effective texts in subsequent steps. It starts with a general idea and a global strategy: then come linear organization and memorizing; even the delivery is planned.

On level (a), we find the respective realizations, i.e., *utterances* occupying—like small talk—one end of the scale, or, in case of a judgment, a patent specification, the other one. Level (c) is the level of the corresponding *systematic prerequisites*: composing texts belonging to elaborate text types demands a much higher mastery of language than buying bread or joking with our neighbor.

Hence, conceptually speaking, orality and literacy are thoroughly intertwined. The often postulated 'great divide' between orality and literacy, between 'oral' and 'literate' societies, simply does not and cannot exist (Hornberger 1994).

2. Transitional Aspects

This holds all the more as the existence of a script does not change or achieve anything by itself. Beyond doubt there is an immense creative and changing potential in literacy. But in order to unfold it, literacy has to meet the appropriate cultural—e.g., institutional—conditions.

There are, nevertheless, some evolutionary tendencies. One of them is that literate societies are likely

Table 1
The conceptual continuum and its representation on three levels of 'language'

	Cognitive scale		
	Wilhelm von Humboldt		
	enérgeia	↔	*érgon*
	P. Koch and W. Oesterreicher		
	conceptually oral	↔	conceptually written
	Douglas Biber		
	interactive text	↔	edited text
(a) 'language' as utterance	e.g. dialog with neighbor	↔	e.g., The papal encyclical *fides et ratio*
(b) 'language' as speech act types	e.g., small talk, gossip, joke	↔	e.g., judgment, textbook, novel
(c) 'language' as system	makes possible a dialog, etc.	↔	makes possible a judgment, etc.

to develop a greater number of textual genres, along with new—e.g., legal—institutions. This enlarges the conceptual space, shifting considerably the societal acquisition of conceptual scripturality (and augmenting the time it takes us to acquire it). There will be neither essays nor editorials, patent specifications, nor testaments in an oral society.

The augmentation of textual genres and the extension of the conceptual scale automatically means new demands on the systemic side. This is shown, among other things, by the fact that a considerable number of English speech act verbs appear for the first time in Early New English: to acknowledge, to advocate, to assert, to concede, to remind, to apologize, to question, to request; or even only in New English: to remark, to retort, to state, to accept, to guarantee, to volunteer (Traugott 1987). Together with the speech act types already mentioned, these examples enhance institutional aspects of literacy.

One further general tendency, which is illustrated below, is the slowness of all processes bringing about cultural change.

2.1 Alphabetic Writing, Orthography, and the Invention of Printing

There are two kinds of writing systems: those representing formal aspects of speech (syllabic, alphabetic script), and those rendering content (ideographic script). Writing proper names is a task that regularly leads to syllable or even alphabetic elements in ideographic script, whereas alphabetic script tends to become ever more ideographic. By reflecting essential aspects of spoken language, early Greek script is characteristic of early alphabetic writing in general: since we do not hear pauses between words, *scriptura continua* is (with some exceptions) a quite natural outcome. It prevails in Western texts up to the eighth century AD and is linked with reading out loud. By 1200, all the achievements of what most of us would call modern layout are present in scholastic texts: spaces between words, punctuation, capitals at the beginning of a new sentence, paragraph indention, emphasis by means of different colors and different script (all of them ideographic features), chapter headings, short summaries in the margin, footnotes, a table of contents, alphabetic registers. As a result, the reader is not lost in an amorphous text. It is only at this time that the general practice of silent reading can begin (Parkes 1992).

Such developments are due to a feedback loop: the higher the number of readers, the more will their needs be taken into account by the invention of features facilitating the reading process—such as ideographic elements. This explains why such phenomena can be observed at the end of the twelfth century: it was not until then that lay literacy developed to any significant extent.

By this time, we observe an incredible increase in the production of written documents. In the State Archives of the Italian city of Siena, there are, for instance, about 500 documents from the twelfth century, but nearly 17,000 from the thirteenth. This acceleration of text production coincides with major changes in social, commercial, and agricultural organization. Thanks to doings and dealings, the city states of Northern Italy already had a population surplus; this required them to organize life and everyday supply very strictly by written statutes. A similar evolution can be observed in large parts of central Europe (for England, see Clanchy 1993).

As a consequence, there was an immense need for knowledge—how shall I write a letter? make a speech? a treaty? how do I behave well?—in short: 'how to win

friends and influence people?' This was the time when, at Bologna, Roman Law was rediscovered; when the great encyclopedias of the Middle Ages were written; when Italian merchants came to invent paper money as well as the new juridical forms of joint stock companies and insurances. In Art and Science, this was the epoch of Gothic architecture and scholasticism (1100–1450), mediated by contact with Arabic culture, and by the discovery of Aristotle. From this perspective, the invention of printing presents itself much more as a necessary consequence than as an agent of change in the domain of literacy.

There was, nevertheless, at least one topic fostered by printing: discussion on orthography. On the one hand, printing books for a great variety of unspecified readers inspires thoughts on standardization. On the other, it goes without saying that all the achievements made in writing and layout increase the intellectual and the practical effort of the writer. This is why discussions on orthography have tended to be dominated, since the sixteenth century, by elementary school teachers who would like to reduce—for those learning to write—the difficulties brought about by the ideographic tendency under discussion.

2.2 Material Aspects of the Written Text

The evolution of layout has an aspect which merits further attention. Unlike spoken language, written text presents itself in two dimensions. Layout makes use of the respective possibilities in making visible chapters, paragraphs, and so on. There is one further possibility which leads to modern mathematics. Thirteenth-century mathematicians were far from being familiar with the system of notation their successors use today. This system only took shape in a slow process between the thirteenth and the seventeenth centuries. Most of the symbols are the outcome of an intermediate stage presupposing alphabetic literacy: the stage of abbreviation. This led not only to most familiar signs such as & ('et'), but also to the plus and minus (+, −), the symbol for the square root, etc. It was Descartes who, at the beginning of the seventeenth century, introduced the conventions that are still followed for the symbolization of the unknown element; they facilitated the highly momentous discoveries made during the seventeenth century: analytic geometry and calculus, resulting in a new ideographic system.

Mathematical ideograms, consequently exploiting the two dimensions of the page, are accessible to simultaneous rather than to linear perception—provided one knows the system. This enormous progress made mathematics the most important complementary discipline to natural sciences, contributing in a decisive way to their role in our modern world (Krämer 1991). This is why, at the beginning of the seventeenth century, Galileo Galilei tells us that the Book of Nature is written in cipher and that those intending to read it have to master the language of mathematics.

In recent times, this role has partly passed to a new branch of mathematics, informatics. Texts written in programming languages not only are at least as demanding as complex written texts in natural languages (Raible 1999). At the same time, their influence on modern life is pervasive. Retrospectively, the second part of the twentieth century is likely to appear as the age of Claude E. Shannon (1916–2001).

2.3 The Written Text as a Metaphor

The use of written language as a metaphor also has a long history—witness the Book of Nature (Blumenberg 1999). The metaphorical use starts with the early atomists, Leucippus and Democritus. Their basic idea was that the whole complex, manifold world surrounding us is nothing but appearance, whereas in reality all was thought to consist of atoms and the void between them. According to what Aristotle tells us in his *Metaphysics* (985b, 15ff.), their visible model for the invisible structure of matter was alphabetic script. The variety of the visible world would be due to the fact that the atoms are differently shaped—just as an A differs in shape from an N; that their order may be different—as the sequence AN is different from NA; finally, their relative position in space may differ: a rotation of 90 degrees makes an N out of a Z. The central idea behind this conception is the reduction of numerous varieties to a restricted set of elements (here the 20-odd letters of the Greek alphabet combining for a possibly infinite variety of written texts).

While the Book of Nature has been a powerful and influential metaphor since the Middle Ages, the idea of the alphabet as the basis of a code reappears at the end of the nineteenth century in cell biology (Friedrich Miescher, 1844–95). It was again put forward in 1943, this time by the Austrian physicist Erwin Schrödinger, (1887–1961), in a series of lectures entitled 'What is life?' where he suggested a genetic alphabet similar to the Morse code. This idea was confirmed in 1953 by Francis Crick and James Watson who showed that the long strands of DNA (Schrödinger's punched Morse tapes) have the structure of a double helix, and by the series of important discoveries that followed this breakthrough.

Since that time, the metaphor of language in the form of alphabetic script has become omnipresent in molecular biology. The four nucleotide bases abbreviated by A, T, G, and C were called letters of the genetic alphabet. RNA-polymerase is reading DNA-sequences with their reading-frames. This process is called transcription, and it happens thanks to transcription factors; transcription is accompanied by an immediate process of proof-reading. The result is called a copy subject to further editing. The resulting

string of mRNA will be translated into a polypeptide. This is made possible because the triplets of DNA encode or are coding for amino acids. The whole process is called gene expression. It presupposes a grammar of biology. As the genome of humans and of other species is being deciphered, the result is stored in large databases modeling the sequences of nucleotides as sequences of the letters A, T, G, C. In this context, molecular biologists are speaking of gene libraries. Recurrent sequences of nucleotide 'letters' as well as recurrent sequences of amino acids in proteins are called motifs, and the genome itself is seen as an encyclopedia (Raible 2001).

2.4 Some Other Processes of Transition

One of the disadvantages of 'oral' cultures is what was called a 'collective censorship.' Texts the audience dislikes fall into oblivion. In literate societies all written texts can be preserved and rediscovered by future generations. This advantage, however, changes into a serious problem with mass media, including the Internet—this is, above all a true consequence of printing.

Unless the public simply refrains from reading, there are at least three reactions. Long texts are transformed into shorter (and cheaper) versions. A lot of these text types were created in antiquity or in the Middle Ages—e.g., epitomē, summarium, argumentum, catechismus, breviarium, compendium. Other genres make a new text out of parts of others—e.g. anthology. Modern literary scholars cannot but use one of the vernacular versions of Valentino Bompiani's *Dizionario delle Opere di tutti i Tempi e di tutte le Letterature*.

A second reaction is an institutional censorship as embodied by all systems of education. They make use of specific curricula, syllabi, textbooks, introductions, primers in order to transform youth into fully-fledged members of society. Making texts as attractive as possible, at the same time making their content easily accessible, is a third answer to the problem. The most radical changes in writing can be observed in the history of mass media, i.e., in journalistic writing and in the corresponding genres of television. Scientific writing is another good example: the articles published in the respective journals are structured according to an identical model and endowed with tables, schemes, and pictures, whose captions and legends, together with the obligatory abstract of the article itself, swiftly inform the readers. The last point shows one of the problems, though: it holds only for those who are already experts in the subject, with the cognitive demands of such texts increasing at the same time.

The presentation of information in the Internet repeats an evolution that took place in the layout of written text. Witness the plethora of new elements in subsequent releases of hypertext markup language (it corresponds to the third level in the above scheme) and the introduction of meta-tags leading, for web crawling machines, to the equivalent of subject catalogues in our libraries. The web pages of the Internet basically exploit the acquisition of literacy (in traditional texts, 'footnotes,' page references, a table of contents, etc., correspond to 'links.' The only major difference is the ready availability of the objects referred to). New features are the availability of color, sound, and moving items.

3. A Precarious Equilibrium

There is no great divide, then, between 'oral' and 'literate' societies, orality and literacy being intertwined on the conceptual level. In addition to this, there is another factor of cultural variation: a social community can dispense with script, but not with spoken language. Apart from entirely oral cultures, we are thus always confronted with more or less literate societies. Given the dialectical relationship between orality and literacy, this means a steadily changing state of equilibrium.

There are indicators showing that the precariousness of this equilibrium is felt. As the evolution of culture in general may lead to the slogan 'back to nature,' the increase in literacy may make us long for the lost paradise of 'innocent' orality. This desire may materialize in a motto like 'let us write the way we speak' (Bader 1994). First examples were Plato's critique of literacy (*Phaidros* 275 d) and the effectiveness of highly planned Sophist rhetoric which made the public suspicious.

Anyhow, the consequence of such reactions is even more literacy. Since antiquity, theoreticians of rhetoric know that among the genres of speech *ex improviso*, speech is the most artful one. Baldassar Castiglione (1478–1529), whose *Courtier* was a model for Europe from the Italian Renaissance on, tells us that true art means dissimulation of art. The same thing holds for the impression of orality called *skaz* that was discovered and recommended by some of the Russian formalists.

What at first sight might appear as paradoxical is quite clear at a closer look. Given the dialectical relationship, in a literate society orality has to be artificially created—just as seventeenth-century authors conceive of Nature (in literature and in the Fine Arts) as 'the negation of the negation of nature.' Once a culture has adopted literacy, returning to 'authentical' orality is impossible.

See also: Literacy and Illiteracy, History of; Literacy, Cognitive Psychology of; Literacy Education; Media Literacy; Oral and Literate Culture; Reading Non-alphabetic Scripts; Reading Skills; Spelling; Writing Systems

Bibliography

Bader E 1994 *Rede-Rhetorik, Schreib-Rhetorik, Konversationsrhetorik. Eine historisch-systematische Analyse*. Tübingen

Biber D 1988 *Variation across Speech and Writing*. Cambridge

Blumenberg H 1999 *Die Lesbarkeit der Welt*. Frankfurt, Germany

Bühler K [1934] 1990 *Theory of Language: The Representational Function of Language*. Goodwin, Amsterdam

Clanchy M T 1993 *From Memory to Written Record: England 1066–1307*. Oxford

Goody J, Watt I 1962–63 The consequences of literacy. *Comparative Studies in Society and History* **5**: 304–45

Günther H, Ludwig O (eds.) 1994 *Writing and Its Use*. de Gruyter, Berlin, Vol. I

Hornberger N H 1994 Oral and literate cultures. In: Günther H, Ludwig O (eds.) *Writing and Its Use*. De Gruyter, Berlin, Vol. 1, pp. 424–32

Koch P, Oesterreicher W 1994 Schriftlichkeit und Sprache. In: Günther H, Ludwig O (eds.) *Writing and Its Use*. De Gruyter Berlin, Vol. I, pp. 587–604

Krämer S 1991 *Berechenbare Vernunft: Kalkül und Rationalismus im 17. Jahrhundert*,. Berlin

McLuhan M 1962 *The Gutenberg Galaxy. The Making of Typographic Man*. Toronto, Canada

Ong W J 1982 *Orality and Literacy: The Technologizing of the Word*. London

Parkes M B 1992 *Pause and Effect. An Introduction to the History of Punctuation in the West*. Scolar Press, London

Raible W 1994 Orality and literacy. In: Günther H, Ludwig O (eds.) *Writing and Its Use*. De Gruyter, Berlin, Vol. I, pp. 1–17

Raible W 1999 *Kognitive Aspekte des Schreibens*. Heidelberg

Raible W 2001 Linguistics and genetics: Systematic parallels. In: Haspelmath M, König E, Oesterreicher W, Raible W (eds.) *Language Universals and Language Typology*, Berlin, Vol. I, art. 8, pp. 105–23

Traugott E C 1987 Literacy and language change. The special case of speech act verbs. *Interchange* **18**: 32–47

W. Raible

Literacy, Cognitive Psychology of

Once oral language was evolving, humans needed to code and decode it for purposes of storage, transmission, and tradition. Thus, writing systems (scripts) were invented and literacy, that is, competence with a written language, evolved. Cognitive psychology of literacy studies how the cognitive processes underlying (de-)coding oral language function in reading, spelling, and writing and how humans learn to become skilled literates. It uses a variety of cognitive models (e.g., boxological, mathematical, and computational; for a classification see Jacobs and Grainger 1994) and methods (e.g., psychophysics, mental chronometry, brain imaging) to explain how humans (learn to) perform the fundamental processes underlying literacy: (a) code elementary speech sounds as written symbols (e.g., phonemes as letters in alphabetic scripts) and vice versa, (b) transform information about those symbols (e.g., phoneme/letter identities and their relative position) into information about meaning, and (c) transform information about meaning into written symbols.

1. Major Issues

Most of the cognitive research on literacy studies the question how humans learn and apply the alphabetic principle in perception and action, that is, the fact that in many languages letters (or letter groups) symbolize elementary speech sounds. Other major issues concern: (a) the functional units of learning/processing spoken and written language (e.g., letters vs. words; phonemes vs. syllables), (b) the question which role phonological awareness, that is, explicit knowledge of the sound structure of words, plays in acquiring literacy, or (c) the causes of individual differences and deficits in the acquisition of literacy and methods to remedy such deficits.

1.1 Learning and Processing Scripts: A Question of the Right Mapping

Literacy involves the mapping of multiple symbols and mental representations of symbol-related perceptions and actions on to each other. Cognitive psychology provides models for necessary and possible mapping processes and methods for testing those models. Since a single script can represent speech at more than one level of representation and different scripts represent different levels, an elementary task for learners of scripts is to detect the basic (i.e., most fine-grained) level at which visual and phonological representation meet (identity relationship) and to develop efficient cognitive processes for discriminating between them and for generalizing from known representations to new ones. About 10 levels of representations are possible: logogen, morpheme, syllable, onset, rime, nucleus, coda, phoneme, phon, and articulatory feature. By hypothesis, in logographic/morphophonological languages like Chinese, the basic processing level is the word. In syllabic languages (e.g., Japanese kana), the same vowels or consonants (phonetically) are represented by different graphic features, depending on the syllable in which they occur. Therefore, the basic level is the syllable. In alphabetic orthographies (phonographic languages) the basic level is the phoneme, because letters code phonemes, not phones (e.g., [p] vs. [p^h]). In such languages, a necessary condition for literacy is to

detect the alphabetic principle and to develop both stable and flexible representations of the critical spelling-sound/sound-spelling relations.

The task of inducing the right relations is made difficult by a number of learning/processing problems that differ across alphabetic languages. First, language users have to deal with the inconsistency of sound-spelling and spelling-sound mappings: identical visual units can represent different sounds and identical sounds can represent different visual units, as illustrated in the famous 'Beware of HEARD a dreadful WORD, it looks like BEARD and sounds like BIRD.' Thus, acquiring literacy should be easier in languages with more consistent mappings (ideally, one-to-one mappings, as in Serbo-Croatian) and more difficult in languages with more inconsistent ones (many-to-many mappings). Indeed, quantifiable differences in the consistency of languages lead to differences in learning and processing performance. In English, whose spelling system was developed to keep morpheme (root meanings) spelling constant (e.g., heal/health), about 40 phonemes have to be represented by 26 letters: thus, distinctive letter groups (i.e., graphemes) are required to represent all phonemes. Furthermore, about a dozen vowel segments in speech have to be represented by only five letters (in comparison: in Italian seven stressed vowels code five letters, any of which is pronounced uniquely). As a consequence, approximately 31 percent of English words are spelling-to-sound inconsistent. Basically, this inconsistency is due to vowel ambiguity: in contrast to initial and final consonants that have the same pronounciation in about 96 percent to 91 percent of words, in only about 50 percent of the words, vowels are pronounced the same. Learners of English adapt to this fact by using higher units of processing than the basic level: by using onsets and rimes, the spelling-to-sound mapping becomes more consistent (vowels are now pronounced the same in about 80 percent of the words) and performance in a variety of psycholinguistic tests improves.

In contrast to English, there is evidence that learners of languages that are more spelling-to-sound consistent use basic-level representations (i.e., phonemes) in perceptual and motor processing (i.e., reading, spelling, writing): German (in which approximately 90 percent of words are spelling-to-sound consistent) or Italian children detect the alphabetic principle earlier and exhibit better decoding skills than English (Cossu et al. 1995, Wimmer and Landerl 1997). There is also evidence that even in skilled adult readers word recognition performance is affected by both the spelling-to-sound and sound-to-spelling mapping: words that are consistent both ways (e.g., duck) are processed better than words that are inconsistent both ways (e.g., swarm) with words being inconsistent either way yielding intermediate performance levels (note that approximately 80 percent of English and French words are sound-to-spelling inconsistent, as compared to 50 percent of German words). A simple and reliable predictor of skilled adult reading performance in English, as measured by the naming task, is the ratio of consistent pronounciations to all pronounciations of a given orthographic segment. The (type) consistency ratio of the orthographic segment-ave, for example, is $1/(1+12) = 0.08$, because there is only one word (have) with the/Av/pronounciation compared to 12 (monosyllabic) words with a different pronounciation (save, gave, cave, etc.). Taking word frequency into account, the consistency ratio is very much increased to 0.89, because the token frequency of have is very high in relation to the summed frequency of its orthographic neighbors.

This notable asymmetry in the consistency of languages can be used to explain the observable asymmetry in the difficulty of reading and spelling: according to Bosman and Van Orden (1997) spelling is more difficult than reading for speakers of English because the sound-to-spelling mapping is more inconsistent than the spelling-to-sound mapping and because the sound-to-spelling inconsistencies must be resolved by the semantic-spelling dynamic, which is relatively weak in comparison to the semantic-sound dynamic that develops earlier in language learning. In contrast, sound-spelling inconsistencies are resolved by the stronger semantic-sound dynamic. This fundamental spelling-reading asymmetry is enhanced by the fact that reading is a more frequent activity than spelling.

A second problem in acquiring the alphabetic principle is that higher-level representations are more accessible than lower-level ones. English-speaking kindergarteners have better access to the number of syllables in a word than to the number of phonemes in a syllable. They also have better access to rimes than to phonemes.

A third problem is the dominance of meaning over form. An example from children is the observation that common semantic features of two words (e.g., the fact that tiger and clown are associated with a circus) can mask phonological differences between them (e.g., that tiger and clown have different onsets). An example which should be familiar to all adult readers is that one has better mnestic access to semantic features than to form features (e.g., upper- vs. lower-case) of words.

1.2 Functional Units in Orthographic and Phonological Awareness/Processing

Which units are functional in learning and processing scripts? All of the above-mentioned 10 representation levels are possible, in principle, but some have been studied more extensively than others and seem to play a predominant role. As concerns visual units, the majority of studies suggest that reading is letter-based (e.g., Nazir et al. 1998), but some evidence favors the hypothesis that in addition it can also be word-based.

However, intermediate units, such as graphemes (i.e., letter clusters representing a single phoneme) have also been found to play a role in word recognition performance: Identification times are longer for words that have fewer phonemes than letters (e.g., beach is composed of five letters and three phonemes, i.e.,/biJ/) as compared to words with equally many phonemes and letters (e.g., crisp-/krIsp/) suggesting that graphemes can be functional reading units (Rey et al. 1998). In addition, onset and rime can play a role in processing written English, as discussed above.

As concerns the problem of phonological awareness, the standard assumption is that three levels play a role: syllable, onset-rime, and phoneme (Byrne 1998).

As concerns the problem of meaning computation from written words, the debate revolves around the alternative: Phonological recoding of written words (i.e., the translation of visual units into sound codes in reading) as a necessary condition for meaning extraction vs. direct reading, that is, the computation of meaning without involvement of sound representations and processes. A related open question is whether orthographic and phonological processes are interactive or independent, and whether they develop interactively or sequentially (see Sect. 2). Many studies suggest that written words automatically activate their sound representations and there is increasing evidence for strong interactions between orthographic and phonological processes even in skilled adult readers (Ziegler et al. 1997).

1.3 Deficits

Reading, spelling, and writing disorders are complex, as evidenced by the postulated existence of at least seven types of dyslexias (peripheral: neglect dyslexica, attentional dyslexia, and letter-by-letter reading; central: nonsemantic reading, surface dyslexia, phonological dyslexia, and deep dyslexia) and about the same number of dysgraphias.

The standard position is that the basic problem underlying developmental disorders of literacy is a phonological processing deficit both in alphabetic and non-alphabetic scripts (e.g., Chinese; Siegel 1998), as efficiently diagnosed by the pseudoword reading test (see Sect. 3). However, possible deficits in temporal and visual processing including eye movements must also be considered.

2. Methods

A wealth of cognitive methods allows to study the multiple processes involved in the investigation of literacy. Pragmatically, they can be classified on various dimensions, such as modality (auditory vs. visual presentation), cognitive aspect (perceptual, mnestic, motor) and task (detection, discrimination, classification, identification), or dependent variable (error rate vs. simple or choice reaction time).

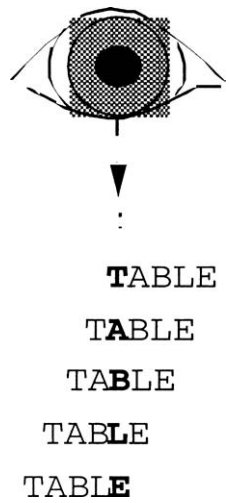

Figure 1
Schematic illustration of the variable viewing position technique: the letter string is briefly presented by the computer such that the eye fixates at a different letter (or viewing zone) on different trials. In general, optimal performance for words is obtained for the viewing zone slightly left of the word middle (e.g., the second to third letter of a five-letter word) with accuracy rapidly decreasing outside this zone

Apart from standard classroom reading and writing tests, tests of phonological awareness are most important for discovering potential developmental disorders in literacy acquisition. Adams (1990) classified the enormous set of tests into five difficulty classes (selected examples are given from difficult to easy): phoneme deletion, addition, or exchange; phoneme tapping; syllable splitting; oddity task, that is, detecting the odd item in a set (e.g., *lot, cot, hat, pot*); and rhyming task.

More sophisticated, computer-controlled tests from psychophysics and mental chronometry (i.e., measurement of reaction times) should complement standard (paper–pencil) classroom reading tests more often in the future. Among the multiple perceptual identification tests, such as the classical Reicher paradigm for properly measuring the word-superiority effect, the grapheme detection (Ziegler et al. 1997) and word fragmentation tests (Ziegler et al. 1998) or the variable viewing position (see Fig. 1; Nazir et al. 1998) and luminance increment procedures (Rey et al. 1998) are recent promising examples for quantifying potential perceptual disorders. For example, children who do not show a word-superiority effect (i.e., the fact that letters in words are processed better than letters in other conditions, such as in pseudowords or in

isolation) in the Reicher task may have a deficit in phonological recoding or lexical processing. Deficits in phonological recoding can be assessed with the grapheme detection test using the absence of a pseudohomophone effect as diagnostic: normal readers yield suboptimal performance for pseudohomophone stimuli (e.g., brane) as compared to control stimuli (e.g., brate), presumably because of an automatic phonological recoding of letter strings. Deficits in lexical processing can be quantified and computationally modeled with the fragmentation test in combination with an interactive activation model of performance in this test. The variable viewing position procedure in combination with the mathematical model of Nazir et al. (1998) allows parametrization of the capacity to integrate letters into words which is an important element for the acquisition of literacy.

The standard alphabetic and lexical decision tasks (i.e., decide as quickly and accurately as possible whether a single sign is a letter or not, or whether a letter string is a word or not, respectively) allow to assess how well alphabetic and lexico-orthographic knowledge can be accessed without a necessary involvement of phonological or semantic processes (Grainger and Jacobs 1996). The naming task (i.e., pronounce a letter string as quickly and accurately as possible) can measure the ease of accessing knowledge about grapheme-phoneme relations. This task is especially interesting for diagnosing problems in acquiring literacy when using pseudowords, because difficulties with reading aloud pseudowords are a reliable indicator of dyslexia (Siegel 1998). For example, to assess whether children have (implicit) knowledge of and access to the rule 'a final e in a one-syllable (English) word makes the vowel long,' pseudowords like rike, fute, or mose can be used.

3. *Models*

As a rule of cognitive psychology a given cognitive test should be accompanied by a falsifiable cognitive (process or task) model that makes explicit the processes and structures underlying performance in the test and allows qualitative and quantitative predictions about performance to assess its validity. Contemporary cognitive models relevant to studies of literacy and reading can be distinguished according to about a dozen criteria (Jacobs and Grainger 1994). Three basic lines that separate them are: (a) prequantitative vs. quantitative, (b) rationalist vs. associationist, and (c) learning vs. performance models.

Prequantitative and quantitative models complement each other: Modeling a given aspect of reality in a prequantitative manner, such as the ubiquitious 'boxological' (i.e., boxes-and-arrows) models of cognition, allows to summarize a wide variety of data obtained with different methods in a static, informal, abstract way. Macroscopic boxological models can be tested at the level of qualitative predictions of off-line data, such as behavioral errors in neuropsychological studies. In order to allow falsifiable predictions, such prequantitative models have to be made more explicit in the form of either mathematical or computational models. Mathematical models use closed-form expressions to describe general functions of dependent vs. independent variables in order to detect lawful relations and are mainly static and very abstract. An example is Nazir et al.'s (1998) model of visual effects in word recognition that can be used to parametrize reading disorders. Computational models allow testing of quantitative predictions of on-line behavioral data that takes into account complex, nonlinear, dynamic interactions between many cognitive processes at different microscopic levels of representation. Prominent examples relevant to research on literacy are interactive activation (Grainger and Jacobs 1996), activation–verification (Paap et al. 1999), dual-route (Coltheart et al. 1993), and parallel distributed processing models of word processing (Zorzi et al. 1998).

Most cognitive models in this field follow the basic rationalism vs. associationism distinction in the philosophy of mind, that is, viewing the mind as a modular system in which rules manipulate symbols vs. the mind as an association machine that detects and acts on correlations between events. Currently open questions addressed by such models are: (a) whether grapheme-phoneme relations are represented by rules or associations or both, (b) whether reading aloud requires access to whole-word representations in a hypothetical mental lexicon or simply involves grapheme-phoneme associations, or (c) whether spelling requires two processes or only one.

A third difference between cognitive models concerns the focus on learning vs. processing/performance. There is a variety of prequantitative models attempting to model the acquisition of literacy, some of which postulate independent stages, while others assume interactive development of orthographic and phonological processes. Increasing evidence favors the latter view (Ehri 1997). As concerns quantitative models, basically, distributed connectionist models allow to study learning problems, while localist connectionist models allow to simulate processing problems (Grainger and Jacobs 1998). Both can complement each other in simulating the main learning/processing problem in acquiring literacy: Analyzing oral speech in a way compatible with the symbols and representations composing the script.

4. *Applications*

For more than a century, teaching literacy has been the battlefield of advocats of different forms of the whole-word method (also known as meaning-based approach) competing with advocats of different forms of the 'phonics' approach. In the former method the

mapping between written and spoken words is taught explicitly at the expense of making explicit letter-phoneme mappings. In the latter approach, children are taught to break words down into their component letters and 'sound them out.' More studies support a superiority of phonics approaches (Byrne 1998), but, of course, an exclusive choice of one or the other method is not necessary: many kinds of mixture approaches are possible. Note, however, that the optimal method depends on the particular spelling-to-sound and sound-to-spelling mappings of a language (Wimmer and Landerl 1997).

5. *Challenges*

In a world full of illiterates and dyslexics, the biggest challenge for cognitive psychology surely remains to provide the applied disciplines with the necessary and sufficient fundamental knowledge allowing to design efficient and reliable methods for diagnosis, teaching, training, and rehabilitation. Given the enormous, heterogeneous set of models and methods in the cognitive literature on literacy, another challenge is theoretical and methodological unification. For example, fundamental research on speech perception still remains relatively disconnected from research on reading and writing; research on implicit learning and working memory is still widely separated from research on literacy; more research on standards for model building and evaluation, as well as for evaluating diagnostic and teaching methods is needed, if we want to fully explain the conditions favoring optimal acquisition and mastering of literacy; finally, so far, too little effort has been spent on developing a formal framework for evaluating the sensitivity, reliability, or validity of cognitive tests relevant to literacy research, such as the ubiquitious lexical decision or the naming tasks.

See also: Literacy and Illiteracy, History of; Literacy and Orality; Literacy Education; Literary Texts: Comprehension and Memory; Oral and Literate Culture; Readability, Applied Psychology of; Reading, Neural Basis of; Reading Skills; Speed Reading; Spelling; Writing Instruction; Writing Process, Psychology of; Writing Systems; Writing Systems, Psychology of

Bibliography

Adams M J 1990 *Beginning to Read: Thinking and Learning about Print*. MIT Press, Cambridge, MA

Bosman A M T, Van Orden G C 1997 Why spelling is more difficult than reading. In: Perfetti C A, Rieben L, Fayol M (eds.) *Learning to Spell: Research, Theory, and Practice across Languages*. LEA, Mahwah, NJ, pp. 173–94

Byrne B 1998 *The Foundation of Literacy*. Psychology Press, Hove, UK

Coltheart M, Curtis B, Atkins P, Haller M 1993 Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review* **100**: 589–608

Cossu G, Shankweiler D, Liberman I Y, Guigliotta M 1995 Visual and phonological determinants of misreadings in a transparent orthography. *Reading and Writing* **7**: 237–56

Ehri L C 1997 Learning to read and learning to spell are one and the same, almost. In: Perfetti C A, Rieben L, Fayol M (eds.) *Learning to Spell: Research, Theory, and Practice across Languages*. LEA, Mahwah, NJ, pp. 237–69

Grainger J, Jacobs A M 1996 Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review* **103**: 518–65

Grainger J, Jacobs A M 1998 On localist connectionism and psychological science. In: Grainger J, Jacobs A M (eds.) *Localist Connectionist Approaches to Human Cognition*. Erlb-aum, Mahwah, NJ, pp. 1–38

Jacobs A M, Grainger J 1994 Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance* **20**(6): 1311–34

Nazir T A, Jacobs A M, O'Regan J K 1998 Letter legibility and visual word recognition. *Memory & Cognition* **26**: 810–21

Paap K R, Chun E, Vonahme P 1999 Discrete threshold versus continuous strength models of perceptual recognition. *Canadian Journal Experimental Psychology* **53**(2): 227–93

Rey A, Jacobs A M, Schmidt-Weigand F, Ziegler J C 1998 A phoneme effect in visual word recognition. *Cognition* **68**: 71–80

Siegel L S 1998 Phonological processing deficits and reading disabilities. In: Metsala J L, Ehri L C (eds.) *Word Recognition in Beginning Literacy*. LEA, Mahwah, NJ, pp. 141–60

Wimmer H, Landerl K 1997 How learning to spell German differs from learning to spell English. In: Perfetti C A, Rieben L, Fayol M (eds.) *Learning to Spell: Research, Theory and Practice across Languages*. LEA, Mahwah, NJ, pp. 81–96

Ziegler J C, Rey A, Jacobs A M 1998 Simulating individual word identification thresholds and errors in the fragmentation task. *Memory and Cognition* **26**: 490–501

Ziegler J, Van Orden G C, Jacobs A M 1997 Phonology can help or hurt the perception of print. *Journal of Experimental Psychology: Human Perception and Performance* **23**: 845–60

Zorzi M, Houghton G, Butterworth B 1998 Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance* **24**: 1131–61

A. M. Jacobs

Literacy Education

Literacy is a relatively recent human invention, and goals of universal literacy within societies are even more recent. Furthermore, although a goal of universal literacy is in place in modern nation-states, it is far from universal, when one considers the vast number of languages for which writing has either not

been developed or for which literacy values have not been firmly established. Studies of the historical contexts of literacy within Western societies make clear that the development of literacy has been embedded in specific cultural 'transactions' (e.g., Resnick and Resnick 1985). The establishment of literacy values in 'developing' non-Western societies occurs within cultural contexts that vary greatly in a number of ways affecting their movement toward literacy (Olson and Torrance 2001).

Where universal literacy goals have been established, the manner of literacy instruction has been subject to change and debate. Although literacy instruction is an issue as inseparable from cultural contexts as the development of literacy itself, it is helpful to focus on the general components of the issue that derive from the distinctive technical nature of literacy—the fact that it has to do with the use of written language. This fact warrants a central place in questions of literacy instruction.

1. *The Definitions of Reading and Literacy*

Reading has been defined both narrowly and broadly. The narrow definition honors an important fact about the dependence of reading on language: *reading is the conversion of written forms into spoken language forms*. With this definition comes the concept of *decoding*, the process of converting printed word forms to spoken word forms. On the narrow definition, then, reading instruction is primarily about the teaching of the code—how the writing system encodes the spoken language system. This can be considered the system-based definition.

By contrast, broader definitions entail literacy as a concept distinct from reading in the narrower sense. In this broader definition, the concept of *literacy* entails the achievement of a broad range of skills embedded in cultural and technological contexts. In this extended culture-centered definition, reading may not even be a critical part of literacy. Thus, a person who can recite religious texts from memory without being able to read them may be included in such a definition (Wagner 1986). Other extended definitions imply functional, rather than cultural, norms. Literacy is 'using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential' (Kirsch and Jungeblut 1986). An extended functional definition is practically useful in specifying expectations for achievement and in helping to make clear the wide range of literacy tasks a society might present to its members. Further extensions of literacy are entailed by such phrases as 'computer literacy,' 'historical literacy,' 'scientific literacy,' etc. In effect, literacy in its broader senses refers to a level of achievement in some subject or specialty, an extension of basic skill to reasoning and discourse in a domain. This amounts to making 'literacy' co-extensive with 'education.'

Intermediate between the narrow definition of reading and the extended definition of literacy is the definition of literacy as reading and writing. This definition may entail the narrower decoding concept, but it adds comprehension, or the attainment of meaning, as the main feature of literacy. In the USA and Europe, reading as obtaining meaning from print has become the most common definition applied in the pedagogy of reading. Thus it may be considered the standard pedagogical definition.

In the face of these various definitions, some clarity is achieved by noting that the common meaning-based definition is *outcome* based, whereas the narrower decoding-based definition is *system*- and *process*-based. Thus to say that reading is the conversion of written forms into language forms, is to appeal to the relationship between *writing systems* and *language systems* and to the process of decoding elements of one system into elements of the other. By contrast, the pedagogical definition is focused on the outcome of comprehension, without regard to the system relationships or the process. The extended literacy definition is neither outcome- nor system-based, but rather is a culturally embedded definition that can complement either the system-based or the outcome-based definition.

In considering the most universal aspects of early reading instruction, it is useful to adopt both the pedagogical definition and the systems definition. The pedagogical definition is useful because it emphasizes a consensus on the major goal of reading, including children's reading. The systems definition is useful because it reflects accurately both the systemic foundation of reading and the related decoding ability that must be acquired by children to learn to read. The treatment of literacy in the sections below center on these two definitions.

2. *Reading and Writing Systems*

Writing systems, which are the result of human invention and refinement, vary according to their design principles even after centuries of convergent development. Conventionally, one distinguishes among alphabetic, syllabic, and logographic systems (Gelb 1952). Thus, English, Italian, Russian, and Korean are examples of alphabetic writing systems, in which graphic units associate with phonemes. Arabic, Hebrew, and Persian represent modified alphabetic systems, in which vowels can be omitted. Japanese Kana is a syllabary system, in which graphic units correspond to spoken language syllables. The logographic system is, at least among current systems, uniquely represented by Chinese (along with the Japanese Kanji adaptation of the Chinese system). In

such a system, the graphic units associate primarily with whole words or lexical morphemes.

Although Chinese approximates this logographic concept, its partial representation of morpheme level pronunciation suggests a different classification, as morpho-syllabic (DeFrancis 1989). Thus, in Chinese the compound character Ñó (pronounced yang 洋 and meaning 'ocean') has two components: the left component signifies that the character meaning has to do with water, while the right component is a character that when it stands alone is also pronounced as yang 羊 (Ñò, meaning sheep). This combination of meaning and pronunciation information inside a character is very far from perfect, but it is frequent enough to challenge the assumption that it is a pure logographic system. Indeed, pure logographic systems seem not to exist among current writing systems.

The significance of writing system analysis for literacy education lies in the following: the fact that writing systems vary in systemic design principles means that children must learn how their writing system works. Do the elementary units correspond to meanings? They do not in an alphabetic system—and not in a syllabary system. The child learning either of these systems must focus some attention on meaningless print units that correspond to meaningless spoken language units. Only in the case of Chinese is there the potential for learning mainly form-meaning mappings. And, as noted above, this mapping is only part of the Chinese system. It is fair to say that there is a universal challenge in learning to read that requires the learning of a relationship between form (written) and form (spoken) rather than between form and meaning.

It is important to note that this challenge is not a consequence of bad writing-system design. It is merely the price to be paid for system efficiency and productivity. Speech-based writing systems are productive in that they make infinite use of finite means (an inventory of syllable signs, an inventory of letters) to express all words and all possible words in a language.

3. *Learning to Read in an Alphabetic Writing System*

In learning an alphabetic writing system, a child comes to learn that letters and strings of letters correspond to speech segments. In principle, this learning might occur implicitly (through the extraction of print-speech correspondences in text) or explicitly (through direct instruction). The problem that has confronted reading instruction is how much success is to be expected from strictly implicit learning.

Complicating the picture is that the consistency of the mappings between letters and phonemes varies within the family of alphabetic systems. The specific manner in which a writing system is instantiated is its orthography, which provided the details of the mapping system. Within the family of alphabetic writing systems, orthographies can be characterized as relatively deep or relatively shallow (Frost et al. 1987). A shallow system provides reliable mappings between letters and phonemes, whereas a deeper system sacrifices some reliability in its mappings, instead preserving more underlying morphology in its spellings. European languages, including Romance, Germanic, Slavonic, and Finno-Ugrian languages, tend to be coded by shallow orthographies. The letters reliably map onto phonemes.

English presents a problem of deep orthography to the learner: many letter patterns in English associate with more than one pronunciation. For example, the letter string 'ow' is read as /ow/ in the words *how* and *cow* and as /o/ in *low* and *tow*. Although the irregularity of English is widely lamented, many of the 'irregularities' turn out to be quite predictable, provided one takes into account more than a single letter. The rime unit—the vowel plus the syllable ending together—turns out to be pronounced very reliably in English (Treiman et al. 1995).

In either a shallow or a deep alphabetic system, the learner needs to grasp the alphabetic principle: letters are associated with meaningless speech segments, or phonemes. Deeper orthographies merely present more context-dependency in these associations.

3.1 *Phonological Sensitivity*

A major link to discovering the alphabetic principle is the child's phonemic awareness, the understanding (more or less explicit) that the speech stream can be segmented into a set of meaningless units (phonemes). Although awareness of phonemes is central to the alphabetic principle, researchers may speak more generally of 'phonological' sensitivity, which entails an appreciation of a broader range of information in the speech signal. Phonemic awareness, variously measured, shows a strong correlation with early reading success. Furthermore, training studies in which children receive explicit instruction in phoneme segmentation produce gains in reading. The research thus supports a strong causal link between explicit knowledge of phonemes and reading in an alphabetic system.

However, the causal relationship between reading and phonemic awareness appears to be reciprocal. First, adults whose cultural context does not include formal schooling may fail to show phonemic awareness (Morais et al. 1979). Second, Chinese children who learned to read Chinese without first learning the Chinese alphabetic (pin-yin) system fail to show phonemic awareness (Read et al. 1986). Third, longitudinal studies of first grade children learning to read

English find a reciprocal relationship, such that initial gains in simple phonemic awareness precede gains in reading, which in turn lead to further gains in phonemic awareness (Perfetti et al. 1987).

For fuller discussions of phonological sensitivity, see Rieben and Perfetti (1991), Liberman and Shankweiler (1991). For a summary of the evidence on the role of phonemic awareness and other factors in learning to read, see the National Research Council Report (1998).

3.2 Theories of Learning to Read

Progress in acquiring reading skills has been viewed typically as a series of stages (e.g., Ehri 1991, Frith 1985). Among the shared ideas across at least some of these theories is the assumption that the earliest stage of reading is characterized by an attempt to learn associations between visual features of graphic forms (not complete orthographic word forms) and spoken words followed by a subsequent stage of graphic-phonological decoding that brings on a truly productive capability in reading. The use of letter names as a bridge to phonology is an important beginning step (Ehri 1991).

Alternative theoretical accounts emphasize the incremental acquisition of decodable lexical representations, rather than discrete stages (Perfetti 1992, Share 1995). Important in these theories is the role of phonology in helping to establish word-specific orthographic representations.

The relationship between spelling and reading comes into the picture quite early. The initial expression of the alphabetic principle appears more often in spelling than in reading (Frith 1985). Indeed, children's attempts at spelling prior to formal reading instruction typically reveal a real understanding that the sounds of a word are to be found in the letters (Read 1971). This early realization forms the basis for later alphabetic reading, even if the earliest stages of reading—as opposed to spelling—appear to be based on partial visual cues or 'logographic' strategies (Frith 1985). Eventually, in nearly all languages, the learner confronts an important fact about spelling: typically, the mapping from pronunciation to spelling is less consistent than the mapping from spelling to pronunciation. Reading is more reliable than spelling. For a collection of papers on spelling in different languages, see Perfetti et al. 1998.

4. Teaching Reading in an Alphabetic System

Given the theoretical understanding of learning to read and the wealth of data that supports central features of these accounts, especially the role of phonological awareness and the need to acquire letter string–phoneme string mappings, the goal of beginning reading instruction for alphabetic systems seems clear. Children need to learn that the letters of their alphabets map onto speech segments of their language; if they lack phonemic awareness, then some instruction in phonemic awareness as part of reading instruction is implied. Phonemic awareness instruction linked to reading instruction seems to be more effective than isolated phonemic awareness instruction.

However, in English-speaking countries—and to some extent in European countries—there has been a history of assertive ambivalence about teaching directly what the child needs to learn. This history is captured in the phrase 'The Great Debate,' the subtitle of a review of teaching methods by Jeanne Chall (1967). The question has been, to what extent—if at all—should beginning instruction focus on teaching directly the correspondences between letters and 'sounds' (phonemes)? The logical answer to this question appears to be that these correspondences, and the alphabetical principle that they instantiate, should be the central initial focus of instruction. However, the tendencies of actual practice have been otherwise. A variety of alternative pedagogies have emphasized instead meaning-focused struction built around story reading, exposure to print, and enhanced language environments. These alternatives are too varied to capture with a single characterization. For example, in the original Great Debate, the alternative to direct instruction was 'whole-word' (also called 'look-say') teaching (Chall 1967), in which basal readers and limited (and later) 'phonics' instruction were typical components. In the last 20 years, the dominant alternative has been 'whole language,' a broad philosophical approach without specified components, except for prohibitions against direct teaching of decoding.

Chall's original conclusion, based on a careful analysis of some 22 programs, classroom observations, and reviews of published studies, was that children who were taught in direct code-based instruction (i.e., decoding-emphasis or 'phonics') classrooms tended to have higher achievement in the first three grades than did children in whole-word classrooms. Although initially whole-word classrooms showed better on measures of comprehension and reading rate, the advantage of decoding-based instruction became highly general, encompassing spelling, word recognition, and comprehension. This conclusion, in its general form, was confirmed in later, less comprehensive reports.

Adams (1990), in a study requested by the US Office of Educational Research and Improvement, provides a review of a number of these research reports and, more generally, an evaluation of teaching methods in the context of a review of the research. Furthermore, Adams (1990) puts the Great Debate in its historical context and helps to explain why there has been so much resistance to the direct teaching of decoding. An emphasis on meaning and comprehension not only

coincides with the main goal of reading, it is also appeals to beliefs that the child's experience in school learning should reflect purposeful learning in authentic contexts. In that spirit, a use of commercially published children's literature as the main or exclusive reading material has become characteristic of whole-language classrooms. Modern advocates of direct instruction point out that there is nothing incompatible between these meaning values and good decoding instruction (often called 'phonics'), which aims to provide the child with the basics of the letter-sound system quickly, while introducing real reading with comprehension at the same time.

The US National Research Council (NRC; the research arm of the National Academy of Sciences) revisited this issue in its report, *Preventing Reading difficulties in Young Children* (National Research Council 1998) The NRC report concluded that research shows that beginning reading 'depends critically on mapping the letters and the spellings of words onto the sounds and speech units that they represent.' Directly counter to the idea that somehow comprehension can proceed on its own, the report adds that 'Failure to master word recognition impedes text comprehension' (p. 321). An additional feature of the NRC report is its emphasis on early language and literacy experience prior to school. It is clear that coming to school with certain relevant skills (e.g., some degree of phonological sensitivity) and dispositions (e.g., an interest in books and stories) eases the burden of school instruction. It is equally clear that if schools are to retain the major responsibility for literacy instruction, they need to organize their efforts to level the variability among children in at-home literacy preparation. Such efforts, organized along the lines supported by research, would aim to ensure that children acquire the ability to decode words and have sufficient opportunity to gain fluency and to read with comprehension.

Although the debate about teaching reading may appear to be about methods of teaching and the research that helps decide which methods are effective, this appearance is probably misleading. Arguments for a whole-language approach (and against the teaching of decoding) have embraced philosophical stances and professional advocacies that have little to do with effective teaching and learning and their research basis. Traces of these philosophical and professional advocacies issues can be seen in the professional teaching publications. The effect of this has been to make research appear irrelevant to the debate. See Stanovich (2000) for a researcher's perspective on this problem.

However, at the close of the twentieth century, there are indications that a stronger consensus has emerged in favor of research-guided practice. There is also an emerging consensus that teaching preparation has been a neglected part of the problem. Strong recommendations to improve both the college education and the 'in-service' training of teachers in the USA have come from recent studies suggesting low levels of teacher preparation in the foundations of reading. Improved teacher training, including more knowledge about the research foundations of reading, was a major recommendation of the NRC report.

5. *Reading Instruction in an International Context*

The questions of how to teach reading are general across languages, writing systems, and national and cultural contexts. However, the English language case is special to some extent. Among alphabetic systems, English is exceptional in its depth and its vocabulary, both its size and its mixed linguistic origins. These two facts are related—a large vocabulary from mixed roots can be expected to produce more spelling variability and hence a less reliable spelling-to-speech mapping. One might wonder whether the facts of English spelling have contributed to the debate on how to teach reading.

In non-English-speaking countries, the debate either has been lacking or appears in a milder form. Among European countries speaking Germanic languages, the tendency is toward the teaching of letter-sound correspondences, although there appears to be variability; e.g., Denmark has placed less emphasis on decoding, the Netherlands and Germany somewhat more. Among Romance language areas, decoding instruction has an important part of beginning reading instruction in Italian and Spanish but not in French.

Indeed, there may be at least a superficial relationship between the reliability of an orthography (its shallowness) and the approach to instruction. Where the mapping from graphic unit to speech unit is highly reliable, whether syllabic or alphabetic, decoding tends to be taught—in Japan, as well as in Korea and most Romance, Germanic, and Slavonic language areas of Europe. In general, children in these language areas are explicitly taught grapheme–phoneme or grapheme–syllable correspondences and are encouraged to use phonological recoding (i.e., 'sounding-out') as their primary word recognition strategy (Feitelson 1988, Mayringer et al. 1998, Reitsma and Verhoeven 1990). The high consistency of these orthographies reinforces the use of phonological coding, allowing most attempts at decoding to produce an accurate word pronunciation.

It is in English-speaking countries, including the USA, the UK, Australia, and New Zealand, where the debate on reading education has been the most heated and where nondecoding methods (particularly whole-language) have been prominent (Turner 1995). This is not an exclusive quirk of English, however. France and French-speaking Switzerland have also tended to emphasize non-decoding approaches. On the orthographic reliability continuum, French orthography

can be characterized as somewhat less reliable than Italian and Spanish, especially in the direction of sound-to-spelling. With these examples, one might be tempted to conclude that the reliability of the spelling-to-pronunciation mapping is a major influence on whether education emphasizes decoding. Certainly, arguments against phonics instruction have sometimes used observations about the 'irregularity' of English spellings as a reason not to teach spelling-sound correspondences. However, it would be misleading to conclude that it is the reduced reliability of an alphabetic orthography that explains failures to teach decoding. Teaching practices are strongly determined by traditions that develop around professional (teacher training), philosophical (child development beliefs), and the marketing of instructional materials. One needs to examine the specific influences that have guided practice contexts within a political region.

In the light of a resistance to teaching decoding, the case of Chinese is especially interesting. Chinese writing presents a high contrast to alphabetic systems—even those with reduced reliability in letter-sound mappings. Even with the presence of phonetic compounds, only a minority of characters give fully accurate information about their pronunciation. Furthermore, for a learner to use this information, he or she must already know the pronunciation of the character that serves as the phonetic cue. Thus, the basic learning procedure is to memorize characters (and the radicals), relating them to spoken words and their associated meanings. This means a logical approach to teaching reading in Chinese would be whole-word or look-say methods. Indeed, until after World War II, this whole-character method was how Chinese children learned to read. Students were taught to recognize individual characters by means of flash card drills and writing practice.

However, in implementing goals of universal literacy, the Chinese added an initial bootstrapping alphabet-based procedure. Children now spend their first eight weeks of school learning a 26-letter alphabet (*Pin Yin* in mainland China, including Hong Kong, and *Zhuyinfuhao* in Taiwan). The instruction focuses on learning grapheme-phoneme correspondences (and phonemic awareness), so that after the first eight weeks of school, the child can read words and sentences written in the alphabet. Then, individual Chinese characters are introduced, accompanied by their Pin Yin counterpart. Gradually, the Pin Yin are removed for familiar characters and then for all characters (Sheridan 1992). A similar system is employed for teaching Japanese children to recognize the Chinese characters (Kanji), which are used together with the syllabic scripts in standard written Japanese. When the kanji are first introduced to children, their pronunciation is written out below the characters in the syllabic script. In both the Japanese and Chinese cases, even with Pin Yin, children spend substantial time with individual characters, including practice at writing characters. One of the most dramatic differences between the educational practices of both China and Japan and those of the North America and Europe is the amount of time children spend on 'homework' (Stevenson and Lee 1990).

It is interesting that the Chinese use an alphabetic system for no other purpose than as an initial teaching device to help children acquire the Chinese system. Not only does this illustrate the value of a speech-based system, its brevity suggests the possibility that preparation for basic decoding can be accomplished in a relatively short time.

6. *International Aspects of Functional Literacy*

Finally, in considering international contexts, it is useful to reconsider the definitions of literacy. The foregoing has assumed the systems and pedagogical definitions of reading. Instruction has been about how to enable children to learn how their writing system maps onto their language. Implicitly, the issue is also about how children can learn to obtain meaning from printed language. The reasonable assumption is that, with the acquisition of decoding skill, a reader's ability to comprehend printed language will be limited only by his or her ability to comprehend spoken language.

However, consistent with broader definitions of literacy, expectations about communicative functioning increase in literate contexts. It is implicit in schooling in Western societies that learning to read enables a broader range of communicative competencies and literate interests. These competencies then become expressed not just in print but in spoken language. Literacy becomes more than being able to read. Its acquisition promotes a range of behaviors, dispositions, and skills that represent broader individual and cultural interests.

These broader aspects of literacy—part of its extended definition—are a source of widespread concern internationally. As local and national expectations for 'functional literacy' become identified, gaps between achievement and performance become visible. Questions of equal opportunities for participation in the economical and social benefits of literacy become salient. Certainly, coming to know the details of how one's writing system represents one's language is the central prerequisite for literacy, even in this broader sense. But both within the technologically advanced countries of Europe and North America and within developing and recently autonomous regions of the world, the development of functional and participatory literacy faces serious obstacles. One shared problem is the target language for reading instruction. Both in developed and developing countries, instruction in a single national language—either the politically privileged or majority language (in the case of Europe, North America, and some countries of South America and Asia) or the language of the former

foreign occupying country (in the case of African countries)—may present cultural and cognitive obstacles to literacy goals.

For additional discussions of functional literacy in this broader sense, see Verhoeven (1994). For papers specifically addressing the obstacles faced in countries without a long history of universal literacy, see Olson and Torrence (2001).

See also: Eye Movements in Reading; Illiteracy, Sociology of; Literacy and Illiteracy, History of; Literacy and Orality; Literacy, Cognitive Psychology of; Literacy Education; Mathematical Education; Reading Skills; Spelling; Writing Instruction; Writing Process, Psychology of; Writing Systems; Writing Systems, Psychology of

Bibliography

Adams M J 1990 *Beginning to Read: Thinking and Learning About Print*. MIT Press, Cambridge, MA

Chall J S 1967 *Learning to Read: The Great Debate*. McGraw-Hill, New York

DeFrancis J 1989 *Visible Speech: The Diverse Oneness of Writing Systems*. University of Hawaii Press, Honolulu, HI

Ehri L C 1991 Learning to read and spell words. In: Rieben L, Perfetti C A (eds.) *Learning to Read: Basic Research and its Implications*. Erlbaum, Hillsdale, NJ, Chap. 5, pp. 57–73

Feitelson D 1988 *Facts and Fads in Beginning Reading: A Cross-language Perspective*. Ablex, Norwood, NJ

Frith U 1985 Beneath the surface of developmental dyslexia. In: Patterson K E, Marshall J C, Coltheart M (eds.) *Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading*. Erlbaum, London, Chap. 13, pp. 301–30

Frost R, Katz L, Bentin S 1987 Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance* **13**: 104–15

Gelb I J 1952 *A Study of Writing*. University of Chicago Press, Chicago

Kirsch I S, Jungeblut A 1986 *Literacy: Profiles of America's Young Adults*. Report No. 16-PL-02. National Assessment of Educational Progress, Princeton, NJ

Liberman I Y, Shankweiler D 1991 Phonology and beginning reading: A tutorial. In: Rieben L, Perfetti CA (eds.) *Learning to Read: Basic Research and its Implications*. Erlbaum, Hillsdale, NJ, Chap. 1, pp. 3–17

Mayringer H, Wimmer H, Landerl K 1998 Phonological skills and literacy acquisition in German. In: Reitsma P, Verhoeven L (eds.) *Problems and Interventions in Literacy Development*. Kluwer, Dordrecht, The Netherlands, pp. 147–62

Morais J, Cary L, Alegria J, Bertelson P 1979 Does awareness of speech as a sequence of phones arise spontaneously? *Cognition* **7**: 323–31

National Research Council 1998 *Preventing Reading Difficulties in Young Children*. National Academy Press, Washington, DC

Olson D, Torrence N 2001 *On the Making of Literate Societies*. Blackwell, London

Perfetti C A 1992 The representation problem in reading acquisition. In: Gough P B, Ehri L C, Treiman R (eds.) *Reading Acquisition*. Erlbaum, Hillsdale, NJ, Chap. 6, pp. 145–74

Perfetti C A, Beck I, Bell L, Hughes C 1987 Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill–Palmer Quarterly* **33**: 283–319

Perfetti C A, Rieben L, Fayol M (eds.) 1998 *Learning to Spell: Research, Theory, and Practice Across Languages*. Erlbaum, Mahwah, NJ

Read C 1971 Preschool children's knowledge of English phonology. *Harvard Educational Review* **41**: 1–34

Read C, Zhang Y, Nie H, Ding B 1986 The ability to manipulate speech sounds depends on knowing alphabetic reading. *Cognition* **24**: 31–44

Reitsma P, Verhoeven L 1990 *Acquisition of Reading in Dutch: Studies on Language Acquisition*. Foris, Providence, RI

Resnick D P, Resnick L B 1985 Standards, curriculum, and performance: A historical and comparative perspective. *Educational Researcher* **14**: 5–20

Rieben L, Perfetti C A (eds.) 1991 *Learning to Read: Basic Research and its Implications*. Erlbaum, Hillsdale, NJ

Share D L 1995 Phonological recoding and self-teaching: *sine qua non* of reading acquisition. *Cognition* **55**: 151–218

Sheridan E M 1992 An analysis of Chinese primary reading and writing textbooks in the People's Republic of China. EDRS No. ED 347 514

Stanovich K E 2000 *Progress in Understanding Reading: Scientific Foundations and New Frontiers*. Guilford Press, New York

Stevenson H W, Lee S-Y 1990 *Contexts of Achievement: A Study of American, Chinese, and Japanese Children*. University of Chicago Press, Chicago

Treiman R, Mullennix J, Bijeljac-Babic R, Richmond-Welty E D 1995 The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General* **124**: 107–36

Turner M 1995 Children learn to read by being taught. In: Owen P, Pumfrey P (eds.) *Emergent and Developing Reading: Messages for Teachers. Children learning to read: International Concerns*. Farmer Press, Bristol, PA, Vol. 1, Chap 6, pp. 80–92

Verhoeven L (ed.) 1994 *Functional Literacy: Theoretical Issues and Educational Implications*. John Benjamins, Amsterdam

Wagner D A 1986 When literacy isn't reading (and vice versa). In: Wrolstad M E, Fisher D F (eds.) *Toward a New Understanding of Literacy*. Praeger, New York, Chap. 14, pp. 319–31

C. A. Perfetti and R. Sandak

Literary Texts: Comprehension and Memory

Nothing seems more effortless and automatic than being swept up in the pages of a good book. In reality, the cognitive processes that underlie the comprehension of literary texts are quite complex, and discourse psychologists are beginning to expose some of these

factors in their theoretical models and empirical work. Some of the relevant variables are common to all types of texts, but others seem to be the special province of literary texts.

1. Literary Texts

In general, literary texts may be differentiated from other types of texts with regard to the identity of the author, the purpose(s) of the author, and the presumed reader. Many literary texts are narratives—their goal is to tell a story. Narratives can be differentiated from expository texts, in which the primary aim is the transmission of information. Encyclopedia articles, or the directions on the back of an aspirin bottle, are narrative texts, and as such lack aesthetic dimensions typically found in literary texts.

1.1 Characteristics of Literary Texts

Many literary texts were originally composed as oral texts—stories transmitted by word of mouth before literacy became widespread. Such stories have features that make them easier for the storyteller to remember. These features include formulaic exposition (e.g., 'Once upon a time … ') and repetition (e.g., the number three in *Goldilocks and the Three Bears*). In addition, literary texts typically have a moral or 'message' for the reader. These morals may be explicit, such as an adage following one of Aesop's fables, or implicit, as in the antiwar message of Heller's *Catch 22*.

Some literary texts, however, contain devices that are designed to sidetrack normal comprehension processes (Zwaan 1996). In poetry, for example, rhyme and alliteration cause readers to focus not only on the content of the poem, but also on the words and their relation to each other. In prose writing, temporal events may be presented out of order, or a whole story might be told in reverse order. Literary writing is also more likely to include figurative language, such as novel metaphors and exaggeration. The purpose of these devices is to achieve the aesthetic effect of defamiliarization—to focus the reader on the linguistic surface structure of the text rather than the situation described by the text. Such texts can be described as 'inconsiderate,' and special control processes may be developed and employed by readers to deal with them.

2. Empirical Study of Literary Texts

The scientific examination of literary texts is a relatively new area of research. Many of the issues that are currently studied in the laboratory were originally raised by literary critics and students of rhetoric. However, these insights were often very general in nature, and could not be related systematically to a large body of texts. Experimental psychologists, however, have been very interested in the processes of reading and comprehension, but have shied away from the complexities of literary texts in favor of simple, experimenter-derived 'textoids' (Graesser et al. 1997). These impoverished texts afforded a great deal of experimental control over variables like sentence length and word frequency, but they cannot shed much light on the aesthetics of literary texts.

Attempts to study literary reading in the laboratory, however, are hampered by the artificiality imposed by the process of measurement. For example, a reader may be asked to read individual sentences on a computer screen, and to press a key after reading each one. In other studies, readers must wear heavy apparatus that allow their eye movements to be tracked. Although much can be learned from such studies, it is difficult to argue that these conditions mimic the process of reading for pleasure.

3. Comprehension of Literary Texts

3.1 Models of Text Comprehension

One influential model of text comprehension suggests that readers combine the textbase (propositions derived directly from the text) with propositions contributed from longterm memory to form a situation model for that text. This construction—integration model has been the focus of much experimentation in cognitive psychology (see van Dijk and Kintsch 1983, Kintsch 1998).

Another cognitive model of comprehension is the CAPS/READER model (Just and Carpenter 1992). This approach employs a production system approach (if ⟨state⟩ then ⟨action⟩) for adding, changing, and removing propositions in both working memory and longterm memory as comprehension takes place. Both the construction—integration and the CAPS/READER model assume that the reader is willing to do the work necessary to make sense of the text, and this process is crucially dependent upon inference generation.

3.2 Inferences

There has been some debate about which inferences are generated as a consequence of processing narratives. Graesser et al. (1994) suggest that there are 13 classes of inferences that might be drawn during narrative processing. Some of these, such as referential inferences, are necessary to establish local coherence,

while others, such as thematic inferences, are required for establishing global coherence. These authors have proposed a constructionist theory of inference generation, which states that readers attempt to construct representations that address the goals of the reader, local and global coherence, and explanations for why particular events, states, and actions appear in the text. As a result of this construction, readers make many inferences on-line (that is, during comprehension), and some inferences off-line (after comprehension).

Although the constructionist model was developed to explain inferential processes in all types of narratives, certain types of inferences may be especially germane for literary texts. These include inferences that involve reader emotion and author intent, such as figuring out the attitude or motivation of the writer. If, for example, a reader encounters a satire or a parody, the reader must make inferences about why the author chose to express his or her ideas in this way.

The constructionist approach makes different predictions than the minimalist hypothesis, proposed by McKoon and Ratcliff (1992). These researchers suggest that the only inferences encoded automatically during reading are those based on information that is explicit, and those that are necessary for establishing local coherence. Strategic inferences, however, are more time consuming to construct. They may be drawn when local coherence cannot be established, and there may be important individual differences that affect whether they are made at all.

Magliano and Graesser (1991) have proposed a three-pronged method for studying inferences generated during the comprehension of literary texts. They suggest that verbal protocols are useful because they reveal inferences generated during the comprehension process. In addition, well-articulated theories of discourse processing allow predictions to be made about which kinds of inferences might be drawn during reading. Finally, behavioral measures, such as reading time or eye movements, may be collected to explicitly test whether certain kinds of inferences are generated.

3.3 Participatory Responses

Gerrig (1993) uses the term participatory responses (or p-responses) to refer to noninferential responses to narrative texts. P-responses can refer to a wide variety of psychological states, such as emotional reactions and imagery. Gerrig's research has suggested, for example, that hopes and preferences are generated as readers process narratives, and that these p-responses can affect decision times to stimuli and the mental representation of the narratives.

Suspense is one type of p-response that is common in literary texts. This reaction will occur in readers when an outcome that they care about is uncertain. Even more interesting is the case of anomalous suspense, or suspense in the absence of uncertainty. For example, when one rereads a favorite novel in which a beloved character dies, it is possible to feel suspense about this outcome even though the fate of the character is not in doubt. Gerrig (1993) suggests that anomalous suspense is the result of an expectation of uniqueness, incorporated within cognitive processes that guide how narrative texts are experienced.

3.4 Control Systems

People read different types of texts with a variety of purposes and goals in mind. Newspaper articles, for example, are often skimmed by readers in order to quickly obtain specific information. The structure of news stories accommodates such an approach: important information appears at the beginning of an article, and background information may appear at the end. Literary texts are very different in their organization, and Zwaan (1991) has suggested that readers have developed different types of control systems for reading different kinds of texts.

Zwaan presented college students with texts that could be read either as news stories or as literary stories. Participants who believed that they were reading a news story read the texts more quickly than those who believed they were reading a literary story. This suggests that a literary-comprehension control system induces slower reading speeds than a news-comprehension control system, and that such control systems influence the on-line processing of texts independently of the features in the texts themselves. It may well be the case that the different control systems exist to accommodate the wide variety of text genres that readers encounter.

4. Memory for Literary Texts

The participants in the Zwaan study described in the previous section also showed differences in their memory for the texts that they read. Subjects who believed that they were reading a literary story had a better surface structure representation (i.e., they more accurately recognized the original wording of the texts) than subjects who believed they were reading a newspaper article. This suggests that the literary-comprehension control system focuses upon the stylistic aspects of a text, at least in comparison to the news-comprehension control system. Since many of the aesthetic effects used by authors occur at the word level (e.g., alliteration, metaphor, and oxymoron), it follows that the literary control system is attuned to such features.

The formulaic nature and repetition of some literary texts can play an important role in being able to

remember them on a later occasion. Freedle (1996), for example, asked college students to recall the folktale of *Goldilocks and the Three Bears*. On average, the participants had last encountered the story 10 years earlier. Recall was greatly aided by the triadic nature of episodes in the folktale (e.g., too hot, too cold, and just right bowls of porridge). The causal connectedness of the propositions in the folktale also predicted recall – events that are linked to many other events in the story are better remembered.

5. Future Directions

Literary texts differ from nonliterary ones in important ways, and researchers are still evolving methodologies that will allow processing and representational differences to be captured. These methodologies will be used to evaluate theories of literary comprehension that have become more finely articulated in the past few years. It also seems likely that physiologically-based measures, such as brain imaging techniques and evoked potentials, will allow even greater temporal precision in the study of on-line comprehension.

See also: History and Literature; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Language and Literature; Memory for Text; Readability, Applied Psychology of; Reading Skills; Text Comprehension: Models in Psychology

Bibliography

Freedle R 1996 Recalling the three bears after ten years: Exploring new vistas for discourse and text-processing models through the study of folktales. In: Kreuz R J, MacNealy M S (eds.) *Empirical Approaches to Literature and Aesthetics*. Ablex, Norwood, NJ

Gerrig R J 1993 *Experiencing Narrative Worlds: On the Psychological Activities of Reading*. Yale University Press, New Haven, CT

Graesser A C, Millis K K, Zwaan R A 1997 Discourse comprehension. *Annual Review of Psychology* **48**: 163–89

Graesser A C, Singer M, Trabasso T 1994 Constructing inferences during narrative text comprehension. *Psychological Review* **101**: 371–95

Just M, Carpenter P 1992 A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* **99**: 122–49

Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK

Magliano J P, Graesser A C 1991 A three-pronged method for studying inference generation in literary text. *Poetics* **20**: 193–232

McKoon G, Ratcliff R 1992 Inference during reading. *Psychological Review* **99**: 440–66

van Dijk T A, Kintsch W 1983 *Strategies of Discourse Comprehension*. Academic Press, New York

Zwaan R A 1991 Some parameters of literary and news comprehension: Effects of discourse-type perspective on reading rate and surface structure representation. *Poetics* **20**: 139–56

Zwaan R A 1996 Toward a model of literary comprehension. In: Britton B, Graesser A (eds.) *Models of Understanding Text*. Erlbaum, Mahwah, NJ

R. J. Kreuz

Lithics and Archaeology

In archaeology, the term 'lithics' refers to all stone artifacts fabricated by humans from the beginning of prehistory until the present time. Stone artifacts have always played a major role in our understanding of prehistory. Stone is the most abundant, and often the only, human-worked material preserved in prehistoric sites. In addition, stone artifacts retain traces on their surfaces that allow us to identify how they were made and used. From an evolutionary perspective, stone-tool making is an essential criterion used by anthropologists to distinguish humans from other hominid species, and to follow the process of human evolution. For the Paleolithic, and along with ceramics and metal for later periods, they are a fundamental source of information concerning past human societies and their economic systems.

1. Historical Aspects

1.1 The Significance of Lithics in the Discovery of Prehistory

At the end of the sixteenth century, Mercati recognized the human origin of polished stone axes, which were previously believed to have been created by lightning bolts. In 1723, Jussieu compared stone axes to ethnographic objects from Canada and the Caribbean. Most notably, as early as the beginning of the nineteenth century, Thomsen classified objects from the Royal Museum of Copenhagen into three successive phases: Stone Age, Bronze Age, and Iron Age. In 1866, Lubbock divided the Stone Age into the Paleolithic, or knapped stone age, and Neolithic, or polished stone age. Meanwhile, the Stone Age, which encompassed all of humanity, and had even been mentioned in biblical and ancient texts, remained undated until the development of absolute dating methods (see *Chronology, Stratigraphy, and Dating Methods in Archaeology*).

However, the notion that stone tools were very ancient was suggested as early as the late eighteenth

century when John Frere identified stone tools in association with the bones of pachyderms at Hoxne (Suffolk, England). Soon after, the first excavations of caves in France, Belgium and England allowed for arguments of a great antiquity of humanity. In 1859, Boucher de Perthes, with support from British geologists, finally succeeded in convincing the intellectual community of the existence of an ancient, 'antediluvian' epoch of humanity. He demonstrated that flint hand axes could be found in situ with the bones of extinct animals deep within gravel quarries in northern France. The idea of very early prehistoric humans equipped with only stone tools was born.

1.2 Lithics as a Criterion for the Definition of Early Humanity?

Chimpanzees, our closest primate relatives, can use natural objects as tools (stone hammers and anvils to break nuts), and even alter natural objects to make tools (removing leaves from twigs to capture termites or ants), but they do not fabricate stone tools in the wild (Berthelet and Chavaillon 1993). The production of stone tools is thus a tempting criterion for distinguishing the earliest species of *Homo* (*Homo habilis*, *Homo rudolfensis*, etc.) from their ancestors and from other contemporary hominids, the australopithicenes, who became extinct between 3 and 1 million years ago. Unfortunately, this criterion remains putative, since it is rare to find stone tools and early hominid fossils in direct association. There is as well no *a priori* reason to believe that some of the australopithecines did not also make stone tools.

To date, the oldest known assemblage of knapped stones, discovered by Helene Roche (1980) at the site of Kada Gona in Ethiopia, and dated to at least 2.6 million years ago, already shows mastery of conchoidal fracture (the basis of controlled stone knapping), perhaps indicating that stone tool technology had already existed for a considerable time.

1.3 The Role of Lithics in the Construction of a Prehistoric Chronology

Once the great antiquity of humankind was finally accepted, researchers put their efforts into developing a chronological classification of the Stone Age. After the first paleontological subdivisions of E. Lartet (1858: Cave Bear Epoch, Reindeer Epoch), classifications were based on stone tools, along with bone tools for more recent periods (G. de Mortillet in 1869, E. Piette in 1889, H. Breuil in 1912, D. Peyrony in 1933). These classifications emphasized the most characteristic tools found in strata of different sites. Following a paleonto-geological perspective, these tools were called 'index fossils'. This procedure led to two excesses that sometimes persist in current research. First, these 'index fossils' were regarded as exclusively representative of distinct time periods and 'cultures,' which were believed to represent different human groups in an ethnic sense, following an overly simplistic historic schematization (see *Ethnic Identity/Ethnicity and Archaeology*). Much work has since been devoted to determining precise dates and geographic origins for these cultures and groups, which are seen to follow a rigid succession of 'evolutionary cycles' based on the appearance and disappearance of a few tool types. Second, these successive cultures, which were originally defined in France following the excavation of numerous stratified sites, were abusively extended to other European regions, thus masking the originality of archaeological data over broad areas.

1.4 A Slowly Developing Technological Perspective

As early as the nineteenth century, some archaeologists were becoming interested in what stone tools could tell us about the daily activities of our ancestors. They asked a first, seemingly accessible, question: how were these tools made? This question led immediately to attempts to reproduce stone tools. However, early experimental procedure lacked rigor and the results remained superficial (Johnson 1978). Numerous prehistorians observed modern gunflint knappers at Meusnes, France and Brandon, Suffolk at the beginning of the twentieth century. However, an extreme production standardization, and exclusive use of iron tools, limited the utility of the comparison. Other prehistorians turned to early accounts of Native Americans manufacturing stone tools. However, many of the descriptions they consulted were second-hand, and thus too imprecise to be useful. Nonetheless, the record in the Spanish Codex of obsidian pressure debitage by the Aztecs permitted D. Crabtree in 1968, and then J. E. Clark in 1982, to reproduce this production process successfully.

Some prehistorians, such as J. Evans, were able to synthesize numerous observations of archaeological collections, ethnographic records, and their own knapping experiments. As a result, the three major knapping operations (shaping, debitage, retouch) and several knapping techniques (direct, hard stone percussion; indirect 'punch' percussion; pressure retouch) were recognized very early (1870). The research of W. H. Holmes (1919) concerning the lithic production activities of prehistoric North Americans was also remarkably modern in its approach to and interpretation of technology. Meanwhile, the functions of stone tools remained hypothetical, except for a few objects such as polished axes and certain arrow heads, the form of which suggested their potential use. Lacking other methods for determining function, the principal categories of stone tools were named after their

similarity of shape with traditional tools (e.g., scrapers, borers, knives, etc.).

2. Recent Methodological Developments and Research Orientations

In conjunction with the genealogical perspective described above, early research was aimed essentially at identifying the major evolutionary stages of the Paleolithic cultures. It was not until the middle of the twentieth century that broader and more analytical perspectives were developed.

2.1 The Typo-Technological Approach

During the 1950s, the French prehistorian, François Bordes, developed a method of studying and comparing assemblages of lithic artifacts, as opposed to single objects, which was both quantitative and qualitative (Bordes 1950). The technological understanding of the objects that he gained through personal knapping experience enriched the methodology he developed. While in the past it had been considered sufficient to discuss only the 'index fossils,' this new method entailed study of all the tools, as well as the technological processes by which they were made. The objective of this broader characterization of lithic industries was the identification of the conceptions of the prehistoric groups that produced them. This became the basis of what is referred to as the 'cultural' approach, which has since been subject to long debates. In particular, Bordes distinguished several different, contemporaneous 'facies' in the Late Mousterian of southwest France. He interpreted these as being representative of different human groups, with different technological traditions, occupying the same region at the same time. L. and S. Binford contested this interpretation, proposing instead that these different facies represented different functional, perhaps seasonal, activities. Other researchers have proposed that these facies could be chronological phases or that they reflect the differential availability of raw materials. None of these interpretations has received unanimous acceptance. In addition, Bordes' method of distinguishing tool types was rather intuitive; some have suggested that several of these tool types actually represent different stages of use and resharpening of a single type of tool (see Renfrew 1973, Mellars 1990, Dibble and Mellars 1992).

Meanwhile, advances in computer technology that have allowed for more elaborate statistical analyses permitted H. L. Movius to confirm the reality of several tool types distinguished by Bordes and D. de Sonneville-Bordes, particularly for the Upper Paleolithic. The 'reality' and meaning of types (especially by Sackett), and the possible use of typological analysis continue to be debated (see *Classification and Typology (Archaeological Systematics)*).

2.2 The Paleoethnological Approach

André Leroi-Gourhan, a French contemporary of Bordes, implemented a paleoethnological approach to prehistoric human groups. This perspective stems from his more general studies of techniques in ethnographic contexts, following a broad anthropological perspective (Leroi-Gourhan 1943, 1945, cf. 1993). Along with other members of his research team, his objective was to reconstruct the spatial dimensions of human activities, which are represented principally by the remains of their technological activities and the organization of the artifacts in relation to various structures (fireplaces, traces of huts or tents, etc.).

Such spatial analysis could be performed only at those rare sites where occupation floors are well preserved. In addition, excavation techniques had to be modified from the primarily vertical, or 'telephone booth' approach that was well adapted to the chronological perspective described above, to a more horizontal approach to allow for exposure of contemporaneous objects on an occupation floor. Under these conditions, this research objective led to two important methodological developments. The first was classification of all lithic remains, including waste-products of production, finished tools, and used and abandoned or broken tools, following the technical process of production and use. This method, which was originally developed in ethnological research, is called *chaîne opératoire* (operational sequence) analysis. The operational sequence is composed of the succession of technical actions beginning at the moment of raw material acquisition and continuing to the abandonment of the tools (cf. Inizan et al. 1999).

The second development was the systematic search for refits between broken objects, retouch flakes and the tool from which they originate, and between lithic objects successively removed from a single core. In this way it became possible to reveal the spatial and chronological relations between isolated objects. Refitting analysis was also developed to detect disturbances within and between archaeological levels, as well as to interpret certain archaeological structures such as different activity areas (cf. Cahen, in Sieveking and Newcomer 1987).

2.3 The Processual Approach

Beginning in the 1960s, L. Binford, and other American archaeologists, developed the approach called

'processual archaeology' (see *Theory in Archaeology*). According to this approach, archaeological artifacts must be considered in terms of the dynamic cultural and natural processes by which they reached the state in which they are found by archaeologists—rather than as static end-products, as they are when a typological approach is used. As with the paleoethnological approach, questions can thus be asked concerning the technological behavior of prehistoric peoples, and we can attempt to situate this behavior within its environmental context. The concept of 'middle-range theory' was borrowed from sociology to provide a link between static archaeological remains and the dynamic processes responsible for them.

This involves the development of explanatory 'models' based primarily on ethnographic references (see *Ethnoarchaeology*) and experimentation. In addition, processual archeologists placed a new emphasis on the natural formation processes (geological, biological, chemical, etc.) of archaeological sites and objects, as well as on the rigorous application of the hypothetico-deductive method to archaeological interpretation. This led to a more critical view of our ability to interpret the past than that espoused by most European archaeologists, and thus to a divergence of research orientations. In terms of lithic studies, American archaeologists became concerned increasingly with the broader contextual aspects of the objects, including the role of lithic technology within the entire technological system of a cultural group. This approach, known as 'technological organization,' addresses issues such as group mobility, tool design needs and constraints, time allocation, risk factors, and degrees of technological investment (see Binford 1983, Odell 1996, Torrence 1989). It is the dominant approach followed by American archaeologists as they enter the twenty-first century.

2.4 Developments in Technological Analysis and Classification

Simultaneously, and until the present time, continental European archaeologists, the French in particular, continued to develop their inductive approach to lithic artifacts. This perspective is based on profound practical knowledge of Old World lithic industries, whose high variability and informational value (due to an abundance of complex and sophisticated technologies) is particularly well adapted to this type of approach.

As early as 1963, Tixier had made the pertinent distinction between method and technique in stone knapping. Technique designates the modes of execution in tool fabrication, including how force is applied (direct percussion, indirect percussion or pressure), the tool employed (e.g., stone cobble, wood, antler or bone hammer, copper point, etc.), and the gestures that are used (e.g., direction of blow, manner of holding the stone and tools, etc.). Method, which is performed according to a systematic and more or less reasoned processes, concerns the organization (order and placement) of series of removals from the object being worked, and thus forms the structure of the operational sequence of production. Systematic experiments are essential to the recognition of techniques, which can be identified based on particular stigmata seen on the objects. In contrast, the knapping method can be identified only through a technological 'reading' of all the objects that entails determination of the direction and order of removals from each piece.

An experienced lithic analyst can perform this technological reading mentally, even in the absence of refit pieces. It is thus possible to classify an assemblage of lithic artifacts according to their original order in the operational sequence of tool fabrication, and distinguish between the products desired by the stone knappers and the waste resulting from their production (Inizan et al. 1999).

2.5 Economic Perspectives

The analytical principles described above, along with the technological classification of lithic artifacts, have revealed that in certain assemblages particular types of 'blanks' were selected for the fabrication of particular types of tools. This realization led to the concept of debitage economy (economy of the products of knapping) (Inizan et al. 1999). This concept was extended to the end of the operational sequence to include tool economy, via which certain blanks were used successively, and/or resharpened in different ways (e.g., work by Binder, and by Perlès). At the beginning of the operational sequence, in a single lithic assemblage, different stone raw materials were often differentially acquired and/or used. This led to the concept of raw material economy (by Perlès, and by Geneste). These perspectives are especially useful for identifying possible choices and decisions made by prehistoric peoples during technological activities (Perlès 1992, Inizan et al. 1999).

Archaeologists also recognized the need for precise determinations of the geographic origins, and modes of acquisition, of the stone materials found in archaeological sites. Following Renfrew's research on the circulation of obsidian in the Near East, surveys and petrographic characterizations were conducted in many regions of the world where diverse workable stone materials are found. The identification of these materials in prehistoric lithic assemblages allows the extent of the territories frequented by foragers to be determined: From several kilometers for the Early Paleolithic human groups, to several hundreds of kilometers during the Upper Paleolithic (Féblot-Augustins 1997). Later, during the European

Neolithic, stone axes and long blades produced in specialized workshops were exchanged over distances up to 1000 km (e.g., Bradley and Edmonds 1993) (see *Trade and Exchange, Archaeology of*).

2.6 The Development of Methods for Determining the Function of Stone Tools

In 1964, with the English publication of his long-standing research, the Russian archaeologist S. A. Semonov (Semonov 1964) demonstrated that the use of stone tools left different macroscopic and microscopic traces on their surfaces. V. E. Shchelinskiĭ and later L. H. Keeley further developed Semenov's results using a high magnification method that allows for identification of the materials worked (e.g., wet or dry skin, bone, antler, meat, plants, etc.) on the bases of comparison with experimental samples. In addition, lower magnification methods can indicate the manner in which the tool was used (e.g., scraping, drilling, sawing, striking, etc.), and how it was possibly hafted, according to the direction of striations and location of polish and micro-removals (Keeley 1980).

Unfortunately, these time consuming methods are not completely reliable. Traces of use are quite often altered, especially on older tools. Another problem is the lack of understanding of some of the physical processes involved in use wear. Nonetheless, certain important results have been obtained, e.g., the recognition of woodworking during the Middle Paleolithic (Anderson, Beyriès; ref. in Dibble and Mellars 1992), and the illumination of harvesting techniques in the Near East (Anderson 1992) (see *Food Production, Origins of*). Other studies, also based on experimentation, have focused on confirming the function of stone and bone points as hunting weapon heads. Such analyses attempt to determine how weapon heads were hafted and used (e.g., thrust or thrown by hand, or with a spear thrower or bow, from long or short distance), and test their durability and performance capabilities (see Knecht 1997).

2.7 The Evaluation of Skill in Lithic Technology

Stone tool fabrication also can be analyzed in terms of cognitive processes (see *Cognitive Archaeology*). Being an acquired ideomotor activity, stone knapping requires a period of apprenticeship, and involves the operation of certain psychomotor capacities. Research questions such as the following can thus be addressed: (a) the development of psychomotor capacities during the course of human evolution: For example, it has been proposed that the production of bifacial handaxes with regular edges and a symmetrical form, by evolved *Homo erectus*, required a propositional level of reasoning (Pelegrin), and/or Piaget's rational degree of intelligence (Wynn); (b) the recognition of different skill levels within a Paleolithic human group, indicating the activities of children and adolescents at different stages of apprenticeship, or a sexual division of activities (Pigeot, Ploux); (c) the recognition of different degrees of technical specialization during the more recent periods of prehistory, from an evaluation of apprenticeship duration based on ethnographic and experimental references (Roux) (cf. Berthelet and Chavaillon 1993, Sinclair and Schlanger 1990, Gibson and Ingold 1993, Renfrew and Zubrow 1994).

3. Conclusion

Since the 1950s/1960s, approaches to lithic artifacts have undergone a genuine revolution. Starting from the early antiquarian perspective, wherein the most 'characteristic' tools were selected *a priori* for the purpose of constructing 'chronocultural' sequences, our approaches to lithic technology have become both more refined and broader. They have been refined by considerable progress in methods of technological analysis of individual objects and of assemblages of objects. The development of practical knowledge through experimental analyses has played a major role in this progress. Approaches have broadened due to the ambition of archaeologists to enhance their field as a human science. From a more ecological and social perspective, researchers are interested in tracing the extraordinary diversity of human behavior since the dawn of humanity, and in understanding its underlying evolutionary mechanisms.

See also: Chronology, Stratigraphy, and Dating Methods in Archaeology; Ethnic Identity/Ethnicity and Archaeology; Ethnoarchaeology; Information Society; Mass Media and Sports; Theory in Archaeology

Bibliography

Anderson P C (ed.) 1992 *Préhistoire de l'agriculture: Nouvelles approches expérimentales et ethnographiques* (Monographies du CRA, No. 6). CNRS, Paris

Berthelet A, Chavaillon J (eds.) 1993 *The Use of Tools by Human and Non-Human Primates* (A Fyssen Foundation Symposium, Oxford Science Publications). Clarendon Press, Oxford, UK

Binford L R 1983 *Working at Archaeology*. Academic Press, New York

Bordes F 1950 Principes d'une méthode d'étude des techniques de débitage et de la typologie du Paléolithique ancien et moyen. *L'Anthropologie* **54**: 19–34

Bradley R, Edmonds M 1993 *Interpreting the Axe Trade: Production and Exchange in Neolithic Britain* (New Studies in Archaeology series). Cambridge University Press, Cambridge, UK

Dibble H L, Mellars P (eds.) 1992 *The Middle Paleolithic: Adaptation, Behavior, and Variability* (The University Museum, monograph 78). University of Pennsylvania, Philadelphia, PA

Féblot-Augustins J 1997 *La circulation des matières premières au Paléolithique*, 2 Vols. (ERAUL No. 75). University of Liège, Liège, Belgium

Gibson K R, Ingold T (eds.) 1993 *Tools, Language and Cognition in Human Evolution*. Cambridge University Press, Cambridge, UK
Inizan M-L, Reduron-Ballinger M, Roche H, Tixier J 1999 *Technology and Terminology of Knapped Stone* (*Préhistoire de la pierre taillée*, 5). CREP, Nanterre, France
Johnson L L 1978 A history of flintknapping experimentation: 1838–1976. *Current Anthropology* **19**(2): 337–72
Keeley L H 1980 *Experimental Determination of Stone Tool Uses*. University of Chicago Press, Chicago
Knecht H (ed.) 1997 *Projectile Technology*. Plenum Press, New York
Leroi-Gourhan A 1993 *Gesture and Speech* (trans. of *Le geste et la parole*, 1964, Albin Michel, Paris). MIT Press, Cambridge, MA
Leroi-Gourhan A 1943 *Évolution et Techniques, I: L'homme et la Matière*. Albin Michel, Paris
Leroi-Gourhan A 1945 *Évolution et Techniques, II: Milieu et Technique*. Albin Michel, Paris
Mellars P (ed.) 1990 *The Emergence of Modern Humans: An Archaeological Perspective*. Cornell University Press, Ithaca, NY
Odell G (ed.) 1996 *Stone Tools: Theoretical Insights into Human Prehistory* (Interdisciplinary Contributions to Archaeology series). Plenum Press, New York
Perlès C 1992 In search of lithic strategies: A cognitive approach to prehistoric chipped stone assemblages. In: Gardin J-C, Peebles C S (eds.) *Representations in Archaeology*. Indiana University Press, Bloomington, IN, pp. 223–47
Renfrew C (ed.) 1973 *The Explanation of Cultural Change: Models in Prehistory*. University of Pittsburgh Press, Pittsburgh, PA
Renfrew C, Zubrow E B W (eds.) 1994 *The Ancient Mind: Elements of Cognitive Archaeology*. Cambridge University Press, Cambridge, UK
Roche H 1980 *Premiers outils taillés d'Afrique*. Société d'Ethnographie, Paris
Semenov S A 1964 *Prehistoric Technology*. Cory and MacKey, London
Sieveking G de G, Newcomer M H (eds.) 1987 The human uses of flint and chert (Proceedings of the Fourth International Flint Symposium held at Brighton Polytechnic 10–15 April 1983). Cambridge University Press, Cambridge, UK
Sinclair A, Schlanger N (eds.) 1990 Technology in the humanities. *Archaeological Review* **9**(1)
Torrence R (ed.) 1989 *Time, Energy and Stone Tools* (New Directions in Archaeology series). Cambridge University Press, Cambridge

J. Pelegrin

Litigation

1. Definition

Litigation is a highly structured process of dispute resolution that invokes the power of the state, or a contractually agreed to private decision maker, to provide a means to authoritatively adjudicate a dispute between two or more parties. In many countries in a large proportion of disputes or cases involving litigation, the parties reach resolutions without completing the full litigation process, either by voluntary settlement, abandonment by one party, or some intermediate authoritative decision that short circuits the full process. Unless there is some legal requirement that the adjudication process be invoked, there is typically an even larger set of disputes that never enter into the formal litigation process but which are resolved with the knowledge that the process could be invoked. Most cases of litigation involve disputes among private parties, either individuals or organizations (businesses, corporations, etc.). Some cases involve the government as a party to the dispute, and, especially in federal systems, the parties on both sides of a case can be governmental entities.

Adjudication involves three key elements: (a) authoritative resolution of a dispute, (b) by a neutral third party, and (c) through the application of pre-existing norms or rules. While the courts customarily are associated with adjudication, there is nothing that limits adjudicatory dispute resolution to courts or other governmental bodies; private adjudication systems are common in many countries and transnationally. Also, while the pre-existing norms or rules most often associated with adjudication comprise what is normally called 'law,' there is nothing that precludes adjudication relying upon any agreed upon set of norms or rules, including those derived from custom or religious texts.

Litigation is a particular form of adjudication that is based on a set of formalized procedures. Formalization is a matter of degree. There are forms of adjudicate dispute resolution that dispense with most aspects of formalized process; examples include small claims courts, qadi courts, or other kinds of community-based courts. On the formality continuum, there is no bright line separating what is appropriately labeled litigation and what is not.

When a governmental entity serves as the adjudicator and oversees the litigation process, it is most often a court. In some governmental systems, the courts are formally independent of the other branches of government, while in other systems the courts are more integrated into the executive. Even in systems where the courts are independent of the executive, there are often specialized adjudicatory forums within the executive, possessing varying degrees of independence. What is labeled a 'court' in one system, may be labeled a 'tribunal' or a 'commission' in another. For example, in Germany, there is a system of 'labor courts' to handle litigation involving many types of disputes arising from the workplace; in Britain some of the cases that would go to labor courts in Germany go to 'Industrial Tribunals' and in the United States the forum is called a Labor Relations or Employment Relations 'Commission' or Workers' Compensation Board. Even within a single country, one finds entities labeled as 'courts' (the US Tax Court) or as something else ('Tax Appeals Commission') where both are

involved in similar adjudicatory activities. Thus, while the 'court' label is usually linked to litigation, that need not be the case.

For a dispute-resolution process to be a form of litigation it only requires that the method of resolution be adjudication as defined above and that there is a formalized, structured process governing the forum's activities. Litigation involves a wide range of issues, and the cases range from the most routine, mundane disputes, to mega-disputes involving large numbers of people or huge sums of money, or both. Because of this range, generalizations are difficult, and it is best to think in terms of multiple worlds of litigation reflecting differing scales of activities and issues.

2. *Theoretical Issues in Understanding Litigation*

The highly structured nature of litigation derives form a combination of rights, procedures, and incentives. Each of these elements influences who uses litigation, when they choose to use it, and the specific nature of the litigation process.

2.1 *Rights*

Rights refer first of all to who is granted access to the mechanism of litigation. In the United States and Canada there are complex rules governing who has 'standing' to bring an action in (federal) court, and under what circumstances they may bring that action; under the US Constitution, the jurisdiction of federal courts is limited to 'cases and controversies,' and only those who are in fact affected by the case or controversy may bring an action. In other countries (e.g., Israel or Germany) many kinds of actions can be brought by persons who have only a tangential connection to harm or the issue in dispute. In still other countries litigation is used regularly as a means of obtaining advisory opinions from judicial bodies, and there need not be any immediate case or controversy in the American sense.

The legal rules and norms within a country or jurisdiction control the kinds of substantive issues subject to litigation. For example, in England tenants often have what amount to a property interest in their tenancy; this can lead to costly disputes and litigation over whether or not a property owner can obtain the use of property held under such a tenancy. In the United States there are very detailed protections afforded certain categories of employees, and there is significant litigation over those protections; in England there are many fewer specific protections, but a more general limitation on an employer's ability to discharge employees without just cause.

In addition to whether or not there are issues that are subject to litigation, the question of whether there are avenues of redress other than litigation is extremely important. In England, consumers may bring claims for defective goods, but they may also turn to authorities to enforce criminal laws vis-a-vis the duties of sellers of consumer goods (i.e., consumer protection can be dealt with under criminal as well as through civil litigation). Despite substantial social and cultural similarities, litigation over torts arising from auto accidents is lower in the Netherlands than in neighboring areas of Germany in part because insurance companies in the Netherlands have institutionalized mechanisms for resolving claims without litigation.

The remedies available also have a great deal to do with whether or not litigation is invoked. In the United States, mass disasters such as fires, airplane crashes, and the like typically result in litigation because American law provides substantial remedies to the victims of such events. In contrast where remedies are severely limited, litigation is less likely. For example, in 1987, an English cross-channel ferry capsized due to crew negligence resulting in almost 200 deaths; English law strictly limited damages, and the ferry company and its insurers quickly resolved the victims' claims. A second example of the impact of available remedies involves persons in the United States claiming job discrimination on some basis other than race. Until the early 1990s civil rights laws limited remedies to lost wages and/or reinstatement; this typically made litigation an unattractive option. In 1991 legislation was passed permitting punitive damages in intentional discrimination cases making redress through litigation more feasible.

2.2 *Procedures*

A set of rules of procedure govern litigation, whether in a court, an administrative tribunal, or a private setting. These rules determine the form by which litigants initiate claims, provide for mechanisms to inform and establish jurisdiction of defendants/respondents, set procedures for gathering the information necessary for the adjudicator to reach a decision, and establish decision-making rules (standards of proof, who makes the decision, etc.). Procedural rules also determine the availability of aggregation procedures (e.g., class actions), who bears the burdens of the cost of litigation (i.e., fee shifting), and who actually serves as the adjudicator of issues of fact and issues of law (professional judges, lay juries, some mixed form). Specific rules may be set by the forum, by the parties, or by some combination of the two, depending upon the structure of the forum.

While there are many variations in specific rules, there are two very broad approaches to litigation which are associated with the two dominant legal traditions: common law systems and civil law systems. The former is built on a model of party adversity, with the adjudicator playing an umpirial role; the parties

and their agents primarily control the flow and process of litigation, with the adjudicator receiving what is offered and forming a decision based on what the parties present. The latter is built on a model variously described as 'inquisitorial' or 'investigatory,' where the primary responsibility for determining and requesting information lies with the adjudicator. Under the common law system, the formal adjudication tends to be concentrated around an event called a trial when the two sides bring in their evidence and witnesses for presentation to the adjudicator. Under the civil law system, the formal proceedings are more episodic with the adjudicator determining what additional evidence he or she would like to see based on what transpires at a particular session. A second major difference between the two traditions is the relative balance placed on oral vs. written processes, with common law procedures involving a heavier oral component while civil law countries emphasizing written materials. Developments in recent years have moved the two systems closer together, with more effort in civil systems to reduce the episodic nature of adjudication procedures and reduce the burden or producing written materials, while common law systems have moved toward increased reliance on written materials.

The specific rules create a vehicle for analyses of litigation by social scientists. These analyses may be process-based, building on the flow of litigation; one approach models litigation as a series of stages and transitions, with litigation progressing from the initial harm through the initiation of a dispute into the formal litigation process culminating in a trail and possibly and appeal. Other analyses rely on tools such as game theory and economic modeling. These latter analyses provide predictions of relative advantages and disadvantages created by the rules, and provide models of how the parties and their agents should behave under assumptions of economic rationality and uncertainty.

2.3 Incentives

Litigation frequently has significant financial implications, in terms of the costs of litigation and/or the resulting remedies. The result is a set of incentives which serve to structure the decision-making of the parties and their agents. The first incentive is the likelihood of success, both in terms of 'winning' and in terms of obtaining the desired redress; those who must bear the cost of litigation are sensitive to whether or not the litigation will succeed. The second incentive is the size of potential remedies, which is a function of the severity of the alleged harm, the law or rules governing the awarding of remedies, and possibly the nature of the adjudicator of the remedy: professional, lay, single, or group, expert, or nonexpert. The third incentive is the costs associated with litigation: institutional fees, costs of representation, and expenses for investigation, experts, etc. The fourth incentive is the nature of the payment system for legal representation: *private payment* based on effort, *payment contingent on success* either on a percentage/commission or some other basis, or *third party payment* such as legal aid, legal insurance, or some other type of group assistance; closely related to payment system is whether or not there is a system of partial or full fee shifting ('loser pays'), and how the level of attorneys's fees is set (relation to size of case, outcome, amount of processing involved).

As with procedure generally, incentive structures lend themselves to economic analyses. Extensive formal analyses have been made of rules governing cost allocation and their expected impacts on decisions to litigate, decisions to go to trial or to settle, the likelihood of frivolous litigation, and the role of attorneys. In addition, to theoretical analyses, procedural rules and the resulting incentives have been the subject of extensive empirical analyses, most heavily in the United States, but also in England.

2.4 The Role of Uncertainty

One of the central features of litigation is uncertainty, and the decisions of actors in the litigation system must be made in light of this uncertainty. Both the degree and source of uncertainty vary depending on rules, the substance of the case, and the general structure of the system. Key elements of uncertainty include: which side will the adjudicator favor, the size or nature of the remedy that the adjudicator will award, the length of time the litigation process will take, and the costs that will be incurred. Part of the uncertainty reflects the strategic nature of all conflict, and part of it reflects the introduction of a third party neutral decision-maker, combined with formalized procedures for influencing that decision-maker. Uncertainty creates risk, and the parties involving in litigation differ significantly in their ability and willingness to bear the burdens created by risk.

3. Empirical Results Concerning Litigation

The most extensive empirical research on litigation has been carried out in the United States, although there are also bodies of empirical litigation focusing on litigation from other common law countries, many European civil law countries, and countries from other regions of the world (Japan, India, the Phillippines, etc.).

3.1 The Likelihood that Disputants will Turn to Litigation

Research from a variety of countries consistently has shown that relatively small proportions of most disputes eventually invoke formal litigation pro-

cedures. The exceptions involve certain types of cases which by law require formal court action (e.g., dissolution of marriage in many countries), but those types of cases frequently involve only *pro forma* court action. Particularly in common law countries, the image of the filtering process has often been cast as a 'Disputing Pyramid,' with 'injuries' lying at the base of the pyramid, and final adjudicatory resolution at the top. The base of the pyramid is broad and the upper areas are narrow; most of the variation whether by type of dispute or country, tends to come in the middle regions of the pyramid, reflecting a combination of incentive structures and alternative avenues of redress.

The initiation of litigation presupposes that an injury is perceived, that responsibility for the injury is externalized, and that a grievant (or an agent of the grievant) seeks redress. Most research in common law countries shows substantial attrition from potential claims to act claims, often reflecting that injured parties are never aware that they have suffered an injury, do not attribute to someone else the responsibility for the injury, or do not initiate a claim because of a belief that redress is not practical or it is not needed given alternative sources of compensation. As suggested above, even if litigation is initiated, most cases end before final adjudication. Settlements occur in what has been dubbed the 'shadow of the law' and they reflect anticipations of what judges or juries will decide.

In civil law countries, the adjudicator plays a more central role and multilevel review is a more integral and usual part of the process. As a result, there are many more adjudicators in civil law countries; the perceived quality of judges making the initial decision is lower in civil law countries compared to common law countries where judges are relatively small in number and are chosen selectively after years of legal practice. In civil law countries, it is relatively easy to get an initial court decision because the judge controls the process; in common law countries, obtaining an initial court decision is more difficult and costly because the parties can fight each other every step of the way. However, in civil law countries, appeals are more common, making the economic advantage of one system over another unclear. In common law countries the emphasis in the civil justice process is on proof provided by the parties; in civil law systems the emphasis is on logical reasoning by the judge.

3.2 Cross-national Patterns in Litigiousness

The United States has a reputation as particularly litigious. This in part reflects the important role of law and courts in the United States from its beginning. However, systematic evidence indicates that the supposed exceptionalism of the United States is either changing or is overstated. Fig. 1 shows estimated litigation rates circa 1990; clearly there is substantial variation among countries in the use of litigation, and the US, while at the higher end of the distribution, does not stand out as exceptional.

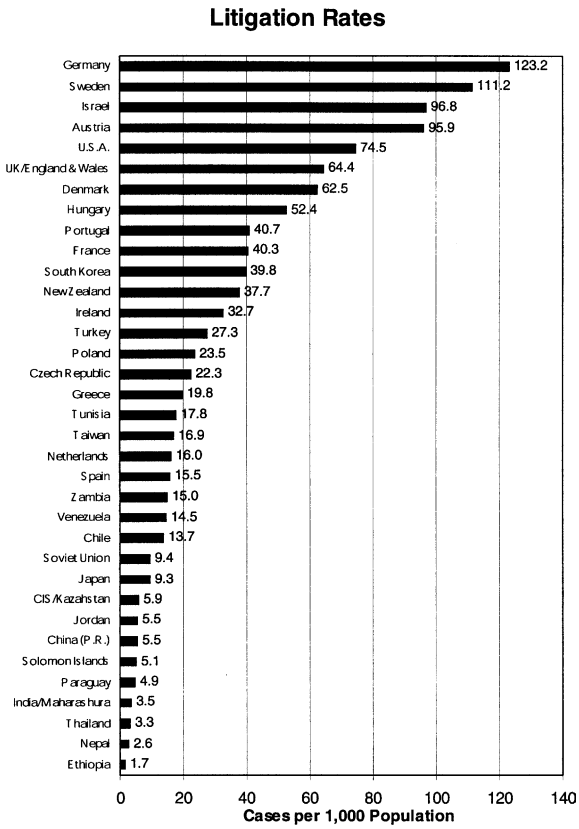

Figure 1
Litigation rates around the world. (Source: Wollschlager C 1998 Exploring global landscapes of litigation rates. In: Brand J, Strempel D (eds.) *Soziologie des Rechts: Festschrift fur Erhard Blankenburg zum 60. Geburtstag*. Nomos, Baden-Baden, Germany, pp. 587–88)

Many explanations for the variation have been advanced. These include stage of economic development, institutional structures, national cultures, alternative remedy systems, and incentive structures. For example, Japan frequently is cited as a country with a very low litigation rate. One explanation for this is that within Japanese culture there is an imperative to avoid conflict, and hence litigation. An alternative explanation is that events such as automobile accidents are dealt with in a way that avoids the need for litigation. A third explanation is that the Japanese elite intentionally has restricted access to litigation as a form of redress.

One general problem with comparing litigiousness is the baseline that should be used. Almost all studies rely on per capita litigation rates. However, litigation arises out of very specific types of events the number of which is not simply a function of population. For example, product liability litigation arises from events involving products that are purchased; a country with a higher rate of material consumption may have a higher per capita rate of product liability litigation simply because the transactions that might give rise to such litigation are more common. The problem of using baselines other than simple population is almost never addressed because of the difficulties of developing such alternative baselines.

3.3 The Role that Litigants and their Agents (Lawyers) Play

Decisions to initiate litigation typically are made by litigants in consultation with their legal advisors. Two elements dominate the decision to litigate. The first is the need to finance the litigation. The second is the likelihood of success, both in terms of whether the desired decision can be obtained, and whether the decision will be enforceable. Research throughout much of the twentieth century has shown consistently that a key element in obtaining redress in tort cases is the availability of a source of payment, either a wealthy defendant or insurance coverage.

Where the potential litigant must directly finance the litigation, the decision to initiate litigation will lie almost entirely with the potential litigant. Lawyers may seek to dissuade potential litigants from pursuing the claim, citing the financial, time, and emotional costs as well as the uncertainty inherent in the process. Where the cost, or at least the risk, of the litigation is to be borne by some other actor, these actors will typically be more involved in the decision to proceed. For example, in the United States where much litigation is funded by the lawyers through the mechanism of a contingency fee, the lawyers' decisions whether or not to accept cases are crucial, and lawyers turn down at least as many cases as they accept. When some third party will cover the litigation expense, such as an insurance company (through litigation expense insurance) or legal aid or a trade union, the third party will need either make the decision whether or not to proceed or require that it be allowed to review and veto any decision to proceed. These decisions are determined by a combination of likelihood of success and the amount at stake.

3.4 Nature of Litigants

The nature of litigants typically varies depending on the type of case. Tort (injury) cases are dominated by individuals as plaintiffs while contract cases more often involve businesses as plaintiffs. Defendants in tort cases are often nominally individuals, but most individual defendants are insured and the insurance company becomes the effective defendant. In contract cases defendants are distributed among individuals and organizations, with variations depending on the exact nature of the issue at stake.

Litigants are not created equal. Litigation typically requires resources and expertise, and analysts have argued that this advantages the wealthy. Galanter (1974) distinguished between 'one shot' and 'repeat players' in litigation, arguing that the repeat players should be expected to win more often than 'one shotters.' While some of the advantages of repeat players reflect resources, other advantages reflect the ability of the repeat player to take a long-run view of litigation. The result is that repeat players are more able to 'play the odds' (i.e., they can act as risk neutral participants), and more able to 'play for rules' (i.e., try to establish precedents that will work to their advantage over the long run).

The importance of the one shot versus repeat player distinction for understanding how litigants behave is further enhanced by dividing the repeat player into the 'government player' which may have even fewer economic constraints than other repeat players, and the 'routine player' for whom litigation is genuinely a part of doing business (e.g., the liability insurance company) and who cares little about the outcome of individual cases as long as it is able to comfortably predict the likelihood of winning and losing. Yet another type of 'repeat player' is the political interest group which uses litigation for explicitly political purposes. While this is most often discussed in the American context, it is by no means unique to the United States; particularly as politics becomes more judicialized, these types of litigants are becoming both more common and more important. Developments such as the increasing role of the European Court of Justice reflect and encourage these changes. Notably, while there has been an increase in 'interest group litigation,' the evidence is inconsistent on whether interest group involvement increases the likelihood of success or whether when successful in court interest groups achieve their ultimate goals.

3.5 Litigation Process

Research on the process of litigation is generally nation specific because the process is so closely tied to the rules governing that process. Some of the research focuses on the handling of specific types of cases, such as injury cases or divorce cases. Much of this research focuses on the role of the lawyer in the litigation process, and how lawyers relate to their clients; there is disagreement over the degree to which lawyers domi-

nate in the decision making process. Other studies focus on specific aspects of litigation, such as the impact of specific rules, the role the judge plays in questions other than adjudication, or problems in aggregating claims. There is no theoretical framework or argument tying together this disparate research which is heavily concentrated in the United States.

3.6 Winning and Losing in Litigation

A natural issue in the study of litigation concerns the outcome: who wins and who loses, and why? The definition of winning and losing is problematic: should outcomes be measured in absolute terms or relative terms? If a plaintiff (the initiator of litigation) obtains some redress from the court, has the plaintiff 'won'? What if that very same result could have been achieved without litigation, and what the plaintiff sought was something more which it failed to achieve? What if the only reason it could have achieved that lesser result was that the option of litigation existed? With relatively few exceptions, studies of litigation outcome ignore these complications. A second problem is that, particularly in common law countries, the final resolution of litigation is most often a private agreement among the parties. Those cases resolved through formal, final adjudication tend to be a small, nonrepresentative subset. Some scholars argue that the cases resolved through trial are those where uncertainty is high or there are strongly conflicting views of the likely outcome. Others argue that some types of parties avoid adjudicatory outcomes intentionally when the result will produce unfavorable results, and concentrate their adjudicatory efforts on cases that will produce favourable precedential outcomes.

Many studies have looked at the relative success of different types of litigants, and while the majority of studies have found that repeat players are more successful than one shot players, the patterns reveal major complications. First, the largest gaps tend to come not between repeat players and one-shot players but between government litigants and nongovernmental litigants. Second, few researchers ask the question of whether one shotters are better off with or without litigation as a means of dispute resolution, that is, litigation must be assessed not just internally but externally vis-a-vis its alternatives. For example, are persons injured through medical negligence better off with litigation as the means of redress, or relying upon the goodwill of medical providers, or with a system of insurance that provides for medical expenses, loss of wages, etc.?

In the United States there has been substantial research on the outcomes of civil trials, particularly those trials decided by juries. Some of these studies focus on civil cases generally, while others focus on single areas such as medical malpractice or products liability. Typically, these studies show a large proportion of verdicts or decisions involving modest amounts of money, but with a distribution showing an extreme tail (i.e., a small number of very large verdicts). While over the broad range of cases, plaintiffs win about half the time and defendants win about half the time, the studies also show a range of plaintiff success depending on the nature of the case reflecting in part whether the uncertainty is over liability or the amount of damages. Research looking at decisions of American trial judges vis-a-vis questions such as liability and damages is much less common; what research there is shows that judges are more likely to find for the plaintiff although the typical judge's award is about 20 percent less than the typical jury's. A danger in all of these comparisons between decisions by judges and decisions by juries is that the cases decided by judges and juries are probably not comparable because litigants themselves choose between the two types of trials.

Another line of research in the US focuses on the substantive decisions of judges in civil cases. This research typically relies upon the examination published decisions by federal trial judges, and distinguishes between decisions described as 'liberal' or 'conservative' (using contemporary American notions of this continuum). The core results of the research demonstrate that there are measurable differences in outcomes depending on whether judges are identified with the Republican or Democratic parties based on either their own affiliation prior to appointment or the affiliation of the President who appointed them. One recognized problem with this research is that published decisions differ from the full population of decisions made by federal district judges in civil cases, and results differ when unpublished decisions are included in the analysis, reducing or eliminating the relationship between decisions and partisanship.

While substantial attention has focused on formal adjudication by judges and juries, less attention has been directed at the outcomes of cases resolved short of adjudication. Part of the reason for this is that determining outcomes is more difficult because they are not part of the public records. A number of empirical studies have focused on the settlement process, both in the United States and elsewhere, generally showing that economic considerations (likelihood of success, who is bearing costs, the magnitude of costs both absolutely and relative to stakes) combined with risk preference are central to understanding settlement. Simply put, cases will tend to settle when the plaintiffs net expected gain (recovery discounted by probability of success *minus* expected costs) is less than or equal to the defendants expected loss (pay out discounted by probability of plaintiff success *plus* costs). When stakes are relatively small, this is almost always true, and it is frequently true for larger cases even when there is substantial disagreement about the likely outcome. It is less clear to what degree the intervention of court officials can influence

the settlement process, in no small part because such a large proportion of cases settle regardless.

See also: Courts and Adjudication; Judges; Lawyers; Legal Insurance; Legal Process and Social Science: United States; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Procedure: Legal Aspects; Rules in the Legal Process

Bibliography

Blankenburg E 1998 Patterns of legal culture: The Netherlands compared to neighboring Germany. *American Journal of Comparative Law* **46**: 1–41

Daniels S, Martin J 1995 *Civil Juries and the Politics of Reform*. Northwestern University Press, Evantson, IL

Galanter M 1974 Why the 'Haves' come out ahead: Speculations on the limits of legal change. *Law & Society Review* **9**: 95–160

Genn H 1987 *Hard Bargaining: Out of Court Settlement in Personal Injury Actions*. Oxford University Press, Oxford, UK

Kritzer H M 1990 *The Justice Broker: Lawyers and Ordinary Litigation*. Oxford University Press, New York

Mnookin R H, Kornhauser L 1979 Bargaining in the shadow of the law: The case of divorce. *Yale Law Journal* **88**: 950–97

Priest G L, Klein B 1984 The selection of disputes for litigation. *Journal of Legal Studies* **13**: 1–55

Shavell S 1982 Suit, settlement, and trial: A theoretical analysis under alternative methods for the allocation of legal costs. *Journal of Legal Studies* **11**: 55–81

Tanase T 1990 The management of disputes—automobile accident compensation in Japan. *Law & Society Review* **24**: 651–91

Zuckerman A A S (ed.) 1999 *Civil Justice in Crisis: Comparative Perspectives of Civil Procedure*. Oxford University Press, Oxford, UK

M. Kritzer

Litigation: Effects on Medical Research

Litigation, broadly defined, is the process where two parties attempt to settle a dispute by court action. The general objective is to obtain a judicial decision about which party's claim is legally stronger. This article considers how litigation has affected medical research. It highlights the concerns of researchers and research subjects about court decisions regarding their rights and liabilities, and how this may have affected their approach to research. The main regions in focus are the USA, Europe, and the Commonwealth countries.

One general area of concern is the nature of dispute resolution in the courts. The ability of parties to argue fairly in an atmosphere that is competitive, stressful, and expensive and the efficacy of legal norms about justice and truth have been widely discussed in critical legal studies literature as well as in media reports. There is a common view that litigation better suits the strategies of powerful, well-resourced parties, and hence has a tendency to determine justice according to their interests. These issues are not addressed in detail in this article as they are not peculiar to the research relationship. This article examines, from a sociolegal perspective, the content of court decisions about medical research and the way that judges have developed legal rules that govern, and hence shape, the experience of research. It considers the claims that judges' decisions have led to legal defensiveness in research and participants' unwillingness to volunteer.

A significant point to note is that there is no clarity about the effect that litigation has had on medical research. On the one hand it is claimed that while court decisions have been aimed at justice, the law in practice has led medical researchers to be anxious and disillusioned, and consequently have limited their innovativeness and medical progress. On the other hand it is claimed that research subjects are concerned about the strength of their legal rights and this makes them reluctant to participate in medical research. There is little empirical data to support either the view that the quality of medical research has diminished, nor whether the effect is due to direct or indirect court-imposed constraints on innovativeness or participants' rights. Thus the competing accounts of the effects of litigation on medical research are inconclusive. Nevertheless it is informative to consider the main areas of law that are reported as being structural evidence for litigation's impact on the relationship between researcher and subject.

1. Compensation for Harm Caused by Medical Research

Medical research inevitably results in some injuries and harm to research subjects irrespective of the care and caution taken. It is a natural consequence of the search for new knowledge. Some researchers have insurance cover, and funding bodies occasionally have a policy of offering limited payment for injuries that were caused by the research, but this is far from universal. However, to receive large sums of compensation, research participants have needed to litigate in the courts.

The compensation cases are usually based on product liability laws, the law of negligence, or insurance contracts. To win their case, the injured person has to prove as a bare minimum that their injury was 'caused' by the medical research and not due to any alternative causal agent like environmental pollutants or underlying disease. 'Causation' has a technical meaning in the law. Suffice to say that it can be difficult to prove legal causation because the typical side effects complained of can be caused by many things and, in therapeutic research, patients are ill to begin with.

Research participants feel that courts require them to prove too much in order to get compensation. They think that their altruism in being a research subject entitles them to easier access to compensation. Researchers feel that they carry the burden of legal liability and that it should instead be spread across the whole community through health insurance or no-fault compensation schemes since medical progress benefits everyone, not just the researcher. They are also alarmed by the variety of legal rules through which they can be called to account, and would prefer more integration and certainty.

This sort of experience could naturally be expected to cause a degree of tension in the research relationship. For example, researchers might be increasingly defensive in their work, and volunteers reluctant to participate. However, the extent to which compensation cases have affected medical research is open to doubt. In the USA, there have been a substantial number of compensation cases about innovative treatment (Siegler et al. 1987), and widespread calls for reform of this area of law. However, there is little evidence of an actual impact on the quality of research, medical progress, or the relationship between research and subject. Furthermore, while liability for compensation has been a significant topic of debate in the USA, it has had a much lower profile in most other countries, which have seen far fewer compensation cases.

2. *Informed Consent*

A shared concern for researchers around the world is the courts' interpretation of 'informed consent.' All research guidelines emphasize the need to obtain prior informed consent from research subjects. Additionally researchers are anxious to avoid liability for assault or bodily trespass for failing to have legally valid consent prior to research involving internal or external body contact.

Court decisions have caused researchers considerable concern because of the demanding levels of communication that they require. A participant must not only be provided with information about the material risks, but steps must also be taken to ensure that they understand that information. Furthermore, consent must be sought without any coercion or inducement, otherwise it will not satisfy the legal requirements of a decision made by free will. The courts have justified these standards on the basis of the importance of individual autonomy. In medical law cases the court has noted that the community expects a less paternalistic relationship with medical professionals than existed in the past.

The medical and research professions complain that this interpretation stifles medical research by imposing impractical and unrealistic standards of consent, and by failing to recognize the utilitarian justifications that exist in the context of medical research. Put briefly, in some cases research participants do not have sufficient scientific understanding to understand the risks, regardless of how much information is provided to them. In other situations, research will inherently have an element of inducement because the mere fact of participating in the study will mean that the patient will receive medical attention that they would not have otherwise had (e.g., where the research methodology involves frequent monitoring or is set in Third World countries).

Researchers are concerned that the effect of court decisions is to rest a heavy burden of potential liability on them. Despite obtaining apparent consent from participants, they may not have met the strict tests of the law. Furthermore, researchers argue that there are certain research studies where obtaining consent would skew research results (e.g., in population studies, or in behavioral studies where subjects' lack of awareness of the research is a critical factor). Therefore, certain exceptions to the consent requirement should be recognized. The exceptions recognized in ethical guidelines are not necessarily recognized by the law.

It is said that the risks of liability for failing to obtain legally valid consent threaten the continuation of research in these areas, albeit they are important research projects with potential benefits for many more people than just the research subjects themselves. Two points should be noted. First, what is a consent burden on researchers is a consent protection for research participants. Thus the effect of litigation about consent cannot be easily categorized as negative or positive. Secondly, there have been very few cases decided where the facts related to medical research rather than consent to treatment (cf. the Canadian case of *Halushka vs. University of Saskatchewan* (1965) 53 DLR (2d) 436). Thus the law relating to consent in medical research is unclear. Researchers could be anxious about a form of liability that the courts will not enforce in view of the public interest in research. This points, though, to one of the general difficulties of court-made law, namely that prior to a binding precedent being set, the law is often ambiguous and uncertain.

3. *Confidentiality and Privacy*

For many years, laws about confidentiality have been largely established by court decisions, as there was very little legislation on this topic. It is clear that legal obligations of confidentiality apply to the doctor–patient and researcher–participant relationship. Researchers have been particularly concerned about the effect of these laws for research that evaluates personal information in medical records. Frequently, a researcher will wish to use information that has already

been collected from individuals for a purpose other than research (e.g., during a general medical consultation). Consent to disclose the information for research purposes will not have been obtained at the time the information was collected. It can be very impractical, sometimes impossible, to work within the courts' interpretation of confidentiality and contact the individuals for their consent to use their record in the research project. For a start the researcher must locate records that belong to a doctor who is willing to spend time identifying and contacting relevant patients, because this process involves looking through confidential details and may not be carried out by an external researcher. Secondly, the contact details will often be long out of date. To proceed without the individuals' express consent leaves the researcher open to liability that is said to be sufficient to deter many.

Research participants, on the other hand, have been getting increasingly concerned that the courts are too lenient about privacy rights beyond nondisclosure. For example, they argue that there ought to be additional standards about data security and aggregation of data held in databanks. In recent years, the privacy movement has successfully motivated governments to enact statutes and protocols with standards for the collection, use, disclosure, security, personal access, and linkage of personal health information. There are limited concessions for medical research in these rules. In many jurisdictions privacy commissioners or data protection commissioners have been established.

4. *Property Rights and Patents*

Court decisions have also had a major impact on the way in which researchers collect tissues and organs for research. One of the most influential cases that has been discussed extensively across the world was *Moore vs. Regents of the University of California* 793 P. 2d 479 (Cal SC1990). In this case a doctor at a university hospital developed a cell line from cells he took from his patient's spleen during an operation. Although the patient had consented to the spleen operation, he had not consented to the use of his cells in research. The cell line was patented and the patient claimed a right to the profits. The court rejected the argument that the patient owned his own body cells. This aspect of the court judgment reinforced the legal norm that patients have no claim to the profits of research despite the fact that they 'invest' their bodies in research. Only those who invest money or professional skill have a claim. This case highlighted the commercial tensions that exist between researchers and participants and dispelled any myths of pure altruistic motivation on the part of either.

The case also had an impact on the relationship between a research subject and a researcher who is also that subject's treating doctor. Although the court said that Mr. Moore did not have property rights in the medical invention, he was nevertheless awarded compensation. The court held that if a doctor treating a patient simultaneously pursues his or her own research for economic gain, the doctor breaches the special legal duties that doctors owe their patients (called 'fiduciary duties'). This decision thus put 'doctor-researchers' on notice that they have a particularly strong obligation to obtain the consent of their patients prior to research, not only to permit interference with their bodies and use of their personal information, but also to keep control of subsequent intellectual property rights and to avoid a serious breach of their professional standards.

The Human Genome Project, which aims to map human genes, has heralded a new wave of controversy about social distribution of biomedical research profits. There have been many court challenges about patents that have been granted by patent offices for human gene sequences. The decisions of US courts have had a global impact. Courts in the USA have upheld the validity of these patents, notwithstanding technical legal arguments about novelty and inventiveness, and submissions that the practice was unethical. This has put considerable economic pressure on other countries to follow suit.

A significant question is whether research participants ought also to profit from a patent that is based on the participant's body but which relies on the researcher's skills and ingenuity. The orthodox position is that research subjects are instruments of research that have rights to human dignity but not to financial profits. It is argued that to change this would radically change the incentives to invest in research to the point of threatening its continuation. Others claim that legal protection of property rights over gene sequences stymies research because new knowledge is disseminated more slowly and small firms are priced out of the market since they cannot afford to license patented knowledge. The empirical evidence is entirely inconclusive.

These landmark court decisions about property rights in human tissue continue to have major reverberations. There have been several recent and highly controversial inquiries in the UK to investigate reports that hospitals have been routinely compiling banks of human tissue and organs from deceased patients (including babies) without the consent of the patients' families and using these for research purposes. The public's shock at the prevalence of this practice has had significant implications for the community's trust in the medical and research professions.

5. *Subpoenas*

Further areas where litigation has generated concern is the ability of parties who are suing or defending a

claim to issue subpoenas for oral and documentary evidence to be given by researchers (Fischer 1994). Not uncommonly the effect of the subpoena is to coerce the researcher to disclose confidential research to a party whose motivation for seeking the information may be questionable. Not only is cross-examination in the courtroom a harrowing experience, but compelled disclosure under subpoena also affects the researcher's ability to release 'fresh' results for publication or intellectual property rights. Some believe that the strategic use of subpoenas to compel release of research has the potential to discourage research in economically or politically sensitive areas. This has been a concern reported in US literature, but less so in other countries despite the similarity of laws on subpoenas.

6. *Putting the Concerns in Perspective*

The discussion above has highlighted the main ways in which it is claimed that court interpretations of the law have affected the practice of medical research. It is important, however, not to form an exaggerated view of the effect of litigation.

There are in fact relatively few court cases about medical research. One US study that systematically reviewed the incidents of medical research litigation in medical faculties and academic medical centers found only 43 cases in 41 years in all US state and federal courts (Helms and Helms 1993). Also the frequency of medical research cases would seem to be considerably less in other countries. Books and articles about legal issues in medical research rarely refer to more than a handful of court decisions.

As noted in the introduction, there is little empirical evidence to sustain the view that the court decisions are actually stifling medical research. The concerns of researchers and patients about the effects of litigation reported above are generally based on anecdotal experience. Naturally, to reinforce their concerns, researchers and patients argue that there is, or is likely to be, a causal link between their concern and the future progress of medical research. However, this has not been established.

Litigation is, in fact, only one part of the regulatory framework for medical research, and its impact cannot be assessed without considering it relative to other forms of regulation. The reality is that medical research is regulated mainly through a system of self-regulation largely controlled by the medical profession with very little involvement by the courts (Moran 1993). The principal avenue of regulation is the process of review by a research ethics committee (sometimes called an institutional review board or human subject committee). This local level of regulation has the most direct and frequent impact on medical research. It determines which projects proceed, which types of patients are recruited, and how the recruitment and research takes place.

The courts are rarely involved for several reasons. The ethics committees are nonstatutory and owe no statutory duties. While they are arguably required to follow rules of administrative law, these relate mostly to procedural breaches rather than substantive rights. Information to establish court cases is also scarce because the committee proceedings are generally held in private. The ethical guidelines enforced by the ethics committees are generally policy documents of professional associations, regulatory agencies, or central government and are not generally reproduced in legislation. Therefore, they too fail to sustain any substantive legal rights and obligations that might be litigated in a court. There are few, if any, sanctions in legislation for breaching the guidelines. Instead, non-compliance is deterred, because it might cause funding to be cut, journals to refuse publication, or give rise to a finding of professional misconduct. However, having noted the limited court involvement in enforcement, it would be an oversight to fail to note the profound influence of one court in propagating the ethical guidelines, namely the Nuremberg Military Tribunal whose decision in the Nuremberg war crime trials in 1946 was transposed into the international treaty on medical research known as the Declaration of Helsinki (adopted by the 18th World Medical Assembly, Helsinki, in 1964 and revised by the Declaration of Tokyo, adopted by the 29th World Medical Assembly, Tokyo in 1975), and subsequently into many self-regulatory research guidelines (Kennedy and Grubb 2000).

The self-regulatory system of medical research has been frequently criticized and may explain (any) disillusionment in medical research more than the effects of litigation. For example, it would seem that the committee review is often cursory and rarely extends beyond examining issues of informed consent because the committees are underresourced. The process rarely requires researchers to submit regular follow-up reports to confirm compliance with the approved research protocol. It is not uncommon for the committees to have too many members, too few, or a pattern of representation that does not accord with government guidelines. The enforcement mechanisms are weak and have limited application to research whose success does not depend on state funding or publication. There can also be an extraordinary degree of bureaucracy for research involving a broad geographical ambit, if the researcher must approach several local ethics committees (Nicholson 1997, Jamrozik 2000).

Although there are frequently calls for reform of the primarily administrative review of research, the proposed alternatives do not include greater court involvement. Most propose a stronger form of self-regulation with legislation, policy guidelines, the backing of state agencies, and more accountable research

review committees. This has also been the case with privacy and confidentiality in medical research (see Sect. 3).

7. Conclusion

In summary there are many reasons to consider that litigation may have had a significant impact on medical research and altered the approach of both researchers and research subjects. This article has explored the principal court decisions that underpin the view that litigation has had such an effect. However, it has also noted that courts are involved in just a few pockets of the mainly self-regulatory framework of research, and have had more particular impact in North America. Further empirical field research is necessary before the claims that litigation is stifling research could be substantiated. In fact this work may well support the contrary view that recourse to litigation is an important safety net for the protection of fundamental human rights and enhances the experience of medical research.

See also: Bioethics: Examples from the Life Sciences; Disputes, Social Construction and Transformation of; Ethical Dilemmas: Research and Treatment Priorities; Intellectual Property Rights: Ethical Aspects; Liability: Legal; Litigation; Parties: Litigants and Claimants; Privacy of Individuals in Social Research: Confidentiality; Research Subjects, Informed and Implied Consent of; Science and Law; Science and the Media

Bibliography

Dingwall R, Lewis P (eds.) 1983 *The Sociology of the Professions: Lawyers, Doctors and Others*. Macmillan, London

Fischer P 1994 Fischer vs. the Medical College of Georgia and the R. J. Reynolds Tobacco company: A case study of constraints on research. *New Directions for Higher Education* **88**: 33

Helms L, Helms C 1993 Litigation involving medical faculty and academic medical centers, 1950–1991. *Academic Medicine* **68**(1): 7

Jamrozik K 2000 The case for a new system for oversight of research on human subjects. *Journal of Medical Ethics* **26**: 334

Kennedy I, Grubb A 2000 *Medical Law*, 3rd edn. Butterworths, London

Moran M 1993 *State, Regulation and the Medical Profession*. Open University Press, Buckingham, UK

Nicholson R 1997 What do they get up to? LREC Annual Report. *Bulletin of Medical Ethics* **129**: 13–24

Siegler M, Toulmin S, Zimring F, Schaffner K (eds.) 1987 *Medical Innovation and Bad Outcomes: Legal, Social and Ethical Responses*. Health Administration Press, Michigan

K. Liddell

Llewellyn, Karl Nickerson (1893–1962)

1. Biography

Karl Nickerson Llewellyn was born in Seattle, Washington on May 22, 1893, though the family soon moved to Brooklyn, New York. At age 16, he went to Germany to study for 2 years, before entering Yale College in 1911. When World War I broke out, he was studying in Paris. Staunchly Germano-philic, he joined up with the 78th Prussian Infantry, was wounded later in 1914, and subsequently received the Iron Cross! Returning to Yale, he completed his undergraduate studies and enrolled in Yale Law School, from which he graduated in 1918. Wartime absences of faculty and students led Llewellyn to stay on at Yale for 1918–19 as editor-in-chief of the *Yale Law Journal* and as a part-time instructor teaching, due to curricular needs, commercial law—the field in which he would later become the most influential American figure of the twentieth century. He began working as a lawyer for a New York City bank in 1920, soon joining the prestigious corporate law firm of Shearman and Sterling. He joined the Yale faculty in 1922, moving to Columbia Law School in 1925; Columbia was then a fertile ground of the new 'Legal Realism,' a movement in which Llewellyn soon emerged as the major figure. He taught at Columbia until 1951, though he spent the academic year 1928–29 once again in Germany, as a visiting professor at Leipzig. During this time, he became better acquainted with the German 'free law' movement, which influenced Llewellyn's conception of Legal Realism. In the 1930s, he worked with an anthropologist doing fieldwork on American Indian law, as well as producing voluminous scholarship in commercial law. From the 1940s until his death, he made his seminal contributions to the Uniform Commercial Code. He moved to the University of Chicago Law School in 1951, which also hired his second wife, Soia Mentschikoff, herself an accomplished legal scholar. He spent the remainder of his career at Chicago, where he died of a heart attack on February 13, 1962.

2. Llewellyn and American Legal Realism

Llewellyn was the most important and famous figure associated with the American Legal Realism of the 1920s and 1930s, a movement that transformed the way both academics and lawyers thought about law. The Realists argued against the dominant 'mechanical jurisprudence' or 'formalism' of the day, which held that judges decide cases on the basis of distinctively *legal* rules and reasons, which justify a unique result in every case. The Realists argued, instead, that a careful look at how judges *really* decide cases reveals that they decide not primarily because of law, but based on their sense of what would be 'fair' on the facts of the case.

Legal rules and reasons figure simply as post-hoc rationalizations for decisions reached on the basis of non-legal considerations.

2.1 Legal Indeterminacy

The Realists argued that the law was 'indeterminate.' By this, they meant that the available class of legitimate legal reasons did not *justify* a unique decision, especially in those cases that reached the stage of appellate review. Llewellyn argued for indeterminacy based on the existence of conflicting, but equally legitimate, canons of interpretation for precedents and statutes. So, for example, courts had endorsed *both* the principle of statutory construction that, 'A statute cannot go beyond its text,' but also the principle that 'To effect its purpose a statute must be implemented beyond its text' (Llewellyn 1950, p. 401). But if a court could properly appeal to either canon when faced with a question of statutory interpretation, then the 'methods' of legal reasoning (including principles of statutory construction) would justify at least two different interpretations of the meaning of the statute. In that case, the question for the Realists was: why did the judge reach that result, given that law and legal reasons did not require him to do so? Llewellyn (1930) made a similar argument about the conflicting, but equally legitimate, ways of interpreting precedent, which he called the 'strict' and 'loose' views of precedent. According to Llewellyn, a judge almost always had the latitude to characterize the decision of a prior case in a highly fact-specific way, so as to distinguish it from the present case, or to characterize its holding in a way that abstracted from the specific facts of the earlier case, so as to make it binding in the present case. Thus, a precedent never really constrained a judge in a later case, because he could construe it 'strictly' to distinguish it, or 'loosely' to treat it as binding.

2.2 Two Branches of Realism

All the Realists agreed that the law and legal reasons are often indeterminate, so that the best explanation for why judges decide as they do must look beyond the law itself. In particular, all the Realists endorsed the following 'core claim': in deciding cases, judges respond primarily to the stimulus of the facts of the case, rather than to legal rules and reasons.

There was, however, a division among Realists over the question of how to explain why judges respond to the underlying facts of the case as they do. The 'Sociological' Wing of Realism—represented by Llewellyn (1930, 1960, 1962)—thought that judicial decisions fell into *predictable* patterns (though *not*, of course, the patterns one would predict just by looking at the existing rules of law). From this fact, these Realists inferred that various 'social' forces must operate upon judges to force them to respond to facts in similar, and predictable, ways. These social forces might include the economic background of the judges and their professional socialization experiences. It was these factors, rather than rules of law, that account for decisions and determine the responses judges have to the underlying facts of the case.

Llewellyn's focus was on the professional socialization and background of judges that would lead them to be sensitive to the norms of ordinary commercial practice. Focussing primarily on commercial law disputes, Llewellyn argued that what judges really do in these cases is try to enforce the uncodified but prevailing norms from the commercial culture in which the dispute arose. Llewellyn (1960, pp. 122–4) gave this famous example: a series of New York cases applied the rule that a buyer who rejects the seller's shipment by formally stating his objections thereby waives all other objections. Llewellyn notes that the rule seems to have been applied rather harshly in the NY cases, since either the buyers may not have known at the time of rejection of other defects or the seller could not have cured the defects anyway. A careful study of the underlying facts, however, revealed that in each case where the rule seems to have been harshly applied what had *really* happened was that the market had gone sour and the buyer was looking to escape the contract. The judge, being 'sensitive to commerce or to decency' (1960, p. 124) applies the unrelated rule about rejection to frustrate the buyer's attempt to escape the contract. Thus, the commercial norm—buyers ought to honor their commitments even under changed market conditions—is enforced by the courts through a seemingly harsh application of an unrelated rule concerning rejection. It is these 'background facts, those of mercantile practice, those of the situation-type' (1960, p. 126) which determine the course of decision.

Llewellyn's seminal work on the Uniform Commercial Code was informed by his Realist perspective (cf. White 1994). For example, the Code imposes an obligation of 'good faith' in all contractual dealings, which is defined as requiring not only honesty but also 'the observance of reasonable commercial standards of fair dealing in the trade.' But for a judge, then, to enforce the rule requiring good faith is just for the judge to enforce the norms of commercial culture! The reliance of the Code throughout on norms of good faith and 'reasonableness' is a constant invitation to the judge to do what he would, on the Realist theory, be likely to do anyway: enforce the norms of the prevailing commercial practice.

The 'Idiosyncracy Wing' of Realism, by contrast—exemplified most prominently by Frank (1930)—claimed that what determines the judge's response to the facts of a particular case are idiosyncratic facts about the psychology of that individual. Thus, Frank notoriously asserted that 'the personality of the judge

is the pivotal factor in law administration' (1930, p. 111). This view is hard to square though with lawyers' ability to predict rather well what judges will do (if they could not do that, litigators, at least, would be out of business). Despite its extreme nature and limited appeal to most Realists, Frank's view has come to dominate the current picture of Realism, to the extent that Llewellyn's far more typical 'sociological' approach is often neglected.

3. *Impact and Current Importance*

By calling attention to the role of nonlegal factors in judicial decision-making, Llewellyn and the Realists initiated the interdisciplinary turn in American legal education, and made clear the need for lawyers to draw on the social sciences in understanding the development of law and what judges do. Much contemporary political science literature on law and courts takes its inspiration from Realism by seeking to explain decisions not by reference to legal reasons (which are assumed to be indeterminate) but by reference to facts about the politics, background, and ideology of the judges (Cross 1997). Recently, it has even been argued that Llewellyn's own approach to commercial law anticipated the now dominant economic analysis of the business law fields (Schwartz 2000). This should hardly be surprising, though, since the economists seek to structure legal rules to maximize efficient outcomes, but efficiency is surely an important norm of all mercantile practice: thus, if judges enforce the norms of mercantile practice (as Llewellyn claims they do and should do), it will turn out that judges will try, among other things, to produce efficient outcomes.

Llewellyn's work on Native American law and dispute resolution practices (Llewellyn and Hoebel 1941) remains an anthropological classic, while his introductory lectures to Columbia law students (Llewellyn 1930) may lay claim to being the most widely read book on law by generations of American law students. While the Uniform Commercial Code (UCC), the single most influential piece of commercial law in the US, bears Llewellyn's imprimatur, it is often lamented in recent years that revisions to the UCC have moved it away from Llewellyn's original vision of a Code sensitive to actual commercial practice. Even so, the lawyers and scholars effecting these changes continually feel the need to grapple with Llewellyn's conception.

Within jurisprudence, Llewellyn's ideas have been subjected to extensive criticism by analytical philosophers (notably H. L. A. Hart), though more recently a philosophical defense has been mounted (Leiter 1997). Among legal scholars, however, Llewellyn's ideas remain influential, both in scholarly movements (like 'Critical Legal Studies' and 'Law and Society') that emphasize law's indeterminacy and the need to look at extra-legal factors to explain law, but also in the work of judges, who now attend to the specific facts of cases (the 'situation-types' as Llewellyn called them) and general nonlegal policy considerations much more openly than before the triumph of Realism in American law.

See also: Courts and Adjudication; Law and Society: Sociolegal Studies; Law: History of its Relation to the Social Sciences; Law, Sociology of; Legal Realism; Legal Scholarship

Bibliography

Cross F 1997 Political science and the new legal realism. *Northwestern University Law Review* **92**: 251–326

Frank J 1930 *Law and the Modern Mind*. Brentano's, New York

Leiter B 1997 Rethinking legal realism: Toward a naturalized jurisprudence. *Texas Law Review* **76**: 267–315

Llewellyn K 1930 *The Bramble Bush*. Oceana, New York

Llewellyn K 1950 Remarks on the theory of appellate decision and the rules and canons about how statutes are to be construed. *Vanderbilt Law Review* **3**: 396–406

Llewellyn K 1960 *The Common Law Tradition: Deciding Appeals*. Little, Brown & Co., Boston

Llewellyn K, Hoebel E A 1941 *The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence*. University of Oklahoma Press, Norman, OK

Schwartz A 2000 Karl Llewellyn and the origins of contract theory. In: Kraus J, Walt S (eds.) *The Jurisprudential Foundations of Corporate and Commercial Law*. Cambridge University Press, New York

Twining W 1973 *Karl Llewellyn and the Realist Movement*. University of Oklahoma Press, Norman, OK

White J J 1994 The influence of American legal realism on Article 2 of the uniform commercial code. In: Krawietz W et al. (eds.) *Prescriptive Formality and Normative Rationality in Modern Legal Systems*. Duncker and Humblot, Berlin

B. Leiter

Lobbying

Lobbying describes the effort of organized interests to inform policy makers and to try to persuade them to choose particular policy choices. The term derives from an earlier time when legislators did not have their own offices. Representatives of organizations would station themselves in the corridors and lobbies to intercept legislators as they walked toward the legislative chamber. On one level lobbying is a communications process, whereby group representatives (lobbyists) or group members send messages that give those in government information favorable to the group's

cause. At another level lobbying is the exercise of a fundamental democratic freedom: the right of citizens to organize and to ask those in government to respond to their preferences. In the US, that right to lobby is enshrined in the First Amendment to the Constitution, which guarantees Americans the freedom to 'petition the government.'

Although lobbying is a basic freedom, it is not without controversy. In no society do all citizens possess equal resources, belong to equally effective organizations, or are equally well represented before government. Interest groups with more resources and higher status constituents are advantaged in the political process. They can lobby on more issues and they can lobby more extensively on the issues they choose to take up.

Citizens look unfavorably upon lobbying because they recognize that interest groups pursue policy goals which are best for their constituents, not necessarily what is best for the nation. Every time an industry group lobbies government for a tax break, that policy change comes at the expense of taxpayers who must make up the shortfall in revenue. Every trade advantage granted to lobbyists works to the detrimient of consumers who must pay the higher costs associated with reduced competition. In one poll of Americans, 64 percent said they regarded lobbying as a threat to democracy in their country. People conveniently ignore that the lobbies they belong to pursue selfish goals as well, believing instead their groups are working for the public good.

1. *Strategies and Tactics*

In studying lobbying political scientists have tried to determine what works best. They have looked at both strategies—broad plans of attack—as well as the tactics—the specific means of communicating with policy makers. Yet in spite of a considerable amount of research, scholars have made limited progress in assessing lobbying effectiveness. Such studies run into three difficult methodological problems. First, policy makers are loathe to admit that they were influenced by lobbyists or interest groups. An affirmative answer diminishes them and could even call into question their integrity. John Kingdon found that legislators he talked to discounted the importance of lobbyists in influencing their voting decisions, but said they were highly influenced by their constituents. On highly visible and important issues, legislators can easily determine how their constituents feel. On complicated and less salient issues, lobbyists help inform legislators and agency officials how their constituents and attentive publics feel. In short, we cannot take this rejection of lobbyists' influence at face value.

Second, much if not most lobbying is directed at policy makers who are already sympathetic to the group's goals. The most critical issue facing lobbyists may be getting sympathizers to actively work on an issue. Lobbying is also directed at fence sitters and those who may have even expressed some doubts about what the group is doing. Sometimes policymakers approach interest groups and lobby them to back a proposal they want to initiate. What is the common metric to measure these different kinds of endeavors, each of which is hard to measure on its own?

Third, the world is a complicated place and lobbying never takes place in a vacuum. Legislators and agency officials are influenced by many different factors including public opinion, political party leadership, the president, and the media. Distinguishing interest group influence from all these other sources has proved to be an intractable methodological problem.

Despite the lack of success in measuring lobbying effectiveness, scholars have succeeded in mapping out the array of tactics used by groups and at developing rich explanations of the underlying strategies. At the risk of oversimplifying the interest group world, lobbying tactics fall within two broad categories. Direct lobbying includes those tactics which involve immediate contact between interest group lobbyists or members and policy makers. This is what we usually think of when we see a reference to lobbying: lobbyists meeting with legislators or administrators. Filing a court suit or an administrative petition also falls within this category as it is a direct effort to persuade policy makers with an argument presented by a group representative. Direct lobbying of policy makers is ubiquitous—98 percent of Washington-based lobbies use it. All lobbyists have their set of the facts and summarizing them in conversation with harried policy makers who have little time to read the substantive literature is seen as critical. Once the European Union began making important decisions, lobbyists began to crowd Brussels. By one estimate there are over 10,000 lobbyists who work the EU parliament and the number is surely growing.

Lobbying the public utilizes a second set of tactics. It is frequently the case that interest group leaders come to the conclusion that policy makers will not do the right thing until the climate of public opinion changes. This is particularly true when key decision makers hold views at odds with the interest group, or don't see the problem as a high priority. In large countries reaching the whole public is difficult and expensive, so efforts are usually tailored to attentive publics—segments of the population that are more likely to respond to tactics such as advertising, direct mail, and membership alerts asking individuals to write to their legislators. Lobbying organizations also make concerted efforts to gain journalists' attention in the hopes of stimulating them to do a story and, thus, generate publicity for the cause. Animal rights groups in the UK, for example, have often resorted to histrionics to generate coverage of their issue.

Interest groups will often mix the two basic strategies in trying to influence government decisions. Members of an organization will receive a letter, fax, phone call, or email asking them to write to their legislators, usually asking them to vote to pass or defeat a piece of legislation before them. At the same time lobbyists for the same group are talking to legislators, reinforcing the message being sent by constituents who are group members. Scholars have long argued that organizations with large memberships are advantaged in the legislative process. This advantage grows when the constituency is one that is easy to mobilize. Memberships that are easy to mobilize are, of course, those which are composed of individuals high in education, status, and income. But it is not just high status individuals who are advantaged. Some lobbying 'groups' are corporations and they have deep enough pockets to fund substantial lobbying operations. Other interest groups are trade associations composed of corporations and are thus able to draw on the substantial resources of many companies within their organization.

2. *Money in Politics*

The advantage of social class in interest group politics goes beyond the help lobbyists get from having members who are cooperative when asked to write to policy makers. Interest groups go to members and donors and ask them for money to be used to strengthen the organization's lobbying efforts. Donations can be translated directly into resources which can be used in their advocacy campaigns. Money can buy more lobbyists, more researchers, more advertising, and more issue alerts sent to members asking them to communicate with legislators.

Money can also be used by interest groups to donate funds to political parties and individual candidates for office. The underlying philosophy here is that the best lobbying tactic is not to try to influence legislators or prime ministers who are already in office, but to put people in office who are already poised to work on behalf of the issues the group cares about. Each country has its own rules governing interest group donations. In Japan, business interests are the principal financial support for various factions within the Diet.

In the USA, interest groups channel contributions in two ways. First, money to candidates for Congress comes from political action committees, most of which are tied to existing lobbies. Second, interest groups can contribute directly to political parties. Despite the controversy over campaign money and its role in elections, political scientists run into the same kind of methodological problems when they try to determine how effective this tactic is. Legislators who receive donations from an interest group are more likely to vote the way the group wants, but that can't be taken as evidence that they were influenced. Most contributions are given to incumbent legislators or candidates who the group has reason to believe are sympathetic to at least some of the goals of the organization. Can we say that labor unions in the UK buy the votes of Labour Party MP's? The unions provide the financial support for the party, but what type of people choose to run for office under the Labour Party banner? In the US scholars have had a particularly difficult time sorting out the complexities of linking PAC contributions to voting behavior by members of Congress. Summarizing the research, political scientist John Wright concluded that 'Empirical evidence about the influence of PAC contributions on congressional voting is filled with ambiguity and apparent contradiction' (Wright 1985).

3. *Stability or Change?*

Although interest groups are an enduring and persistent part of democratic politics, they frequently face significant changes in their political environment. In parliamentary systems with proportional representation, for example, new political movements can quickly metamorphize into new political parties. The lobbying organizations for that cause may have to switch quickly from protest politics to campaign politics. In the American context, with its stable two-party system, the challenge more recently has been the unprecedented growth of interest groups. Looking back on his career, one member of the House remarked 'When I first came to Congress there were five major financial trade groups, but now there are at least five times that. Now if you're trying to satisfy all the trade groups, it's pretty hard to do.'

For present-day lobbyists the most significant tactical decision may be figuring out what to do just to get noticed. With so many lobbyists working on the same issue, how do the lobbyists from a single organization convince legislators and their staffers that their organization is one to consult with and to work up drafts of legislation with? The competition between lobbyists has put an even greater premium on expertise. Although well-connected lobbyists are always helpful, policy expertise is a particularly important trait of effective lobbyists.

An interesting question that has emerged is whether the new technologies are changing interest group politics. The new technologies have certainly made it easier for interest group organizations to reach their members and attentive publics. Today communication can be virtually instantaneous. If events change rapidly lobbyists in the capital can mobilize their members quickly to let policy makers know that voters back home have noticed what's going on and they are upset about it. Interest groups have also taken advantage of the Internet, establishing their own web sites to serve as vehicles for soliciting members, advertising their

views, and facilitating the retrieval of documents by sympathesizers around the country. Another use of the Web is that it facilitates the organization of new groups. Small constituencies with modest resources can use the economies of the Internet to reach those concerned with their issue and to generate communication among themselves.

Yet for all the changes that have taken place in government and in telecommunications, the fundamental nature of lobbying remains relatively stable. At its core lobbyists work assiduously to mobilize their followers and to bring information to the attention of policy makers. Although how those tasks are carried out changes over time, these changes have not revolutionized the process.

As political science has become increasingly sophisticated and the discipline has placed greater emphasis on precise measurement of complex phenomena, lobbying has remained a particularly difficult nut to crack. Assessing the effectiveness of different types of lobbying or different lobbying campaigns continues to be a formidable methodological problem. We know a lot more about many facets of interest group politics, especially in understanding how and why people join organizations. Research on lobbying, however, has made only limited progress.

See also: Campaigning: Political; Influence: Social; Labor Unions; Participation: Political; Policy Knowledge: Advocacy Organizations; Policy Process: Business Participation; Political Communication; Political Machines; Political Money and Party Finance

Bibliography

Baumgartner F R, Leech B L 1998 *Basic Interests*. Princeton University Press, Princeton, NJ

Berry J M 1997 *The Interest Group Society*, 3rd edn. Longman, New York

Berry J M 1999 *The New Liberalism: The Rising Power of Citizen Groups*. Brookings Institution Press, Washington, DC

Cigler A J, Loomis B A 1998 *Interest Group Politics*, 5th edn Congressional Quarterly, Washington, DC

Dalton R J 1994 *The Green Rainbow*. Yale University Press, New Haven, CT

Gais T 1996 *Improper Influence*. University of Michigan Press, Ann Arbor, MI

Heinz J P, Laumann E O, Nelson R L, Salisbury R 1993 *The Hollow Core*. Harvard University Press, Cambridge, MA

Kollman K 1998 *Outside Lobbying*. Princeton University Press, Princeton, NJ

Schlozman K L, Tierney J T 1986 *Organized Interests and American Democracy*. Harper & Row, New York

Wilson G K 1990 *Interest Groups*. Blackwell, Oxford, UK

Wright J R 1985 PACs, contributions, and roll calls. *American Political Science Review* **79**(June): 400–14

J. M. Berry

Local Economic Development

Local economic development is concerned with the well-being of a place, a locality. Distinct from 'economic development' more generally, local economic development focuses on particular contexts, which vary across space. 'Context' refers to the constellation of endogenous behaviors and formal and informal institutions that have evolved in a place and interact with exogenous cross-currents that may be social, cultural, economic, political, ecological. Approaches to local economic development vary substantially, and are themselves context-specific. 'Local' has different meanings relative to the broader context in which places are conceptualized. 'Economic development' also has different meanings; that is, the nature of 'well-being' is variously interpreted, relative to researchers' theoretical perspectives. Accordingly, 'local' and 'economic development' are addressed separately, and approaches to local development as a field of study are elaborated subsequently.

1. The Term 'Local'

'Local' can mean a neighborhood, town, or metropolitan area in the context of a region or nation; it can refer to a region in the context of a nation; and it can signify a nation or even a multinational region in the context of the global economy. 'Local' is relative, referring to scale specificity and relations among processes at different scales, although it is not itself specific to a particular scale. Examples of local development within the USA illustrate the point. In the 1960s and 1970s local development typically entailed subnational, regional units, such as Appalachia, which were recognized as problem areas amid general national prosperity. The US South was compared to a Third World country because it hosted footloose branch plants of multinational corporations (MNCs) that did not invest locally in institutions or in people, via upskilling (Cohen 1977). By the end of the 1980s and in the context of profound economic restructuring nationwide and globally, economic volatility, a large federal deficit, and reduced federal funding to states, 'local' in the USA continued to refer to regions, but also to urban areas, some of which confronted problems of industrial decline and out-migration of people and capital (Rodwin and Sazanami 1989). At the same time, 'localized distress' also referred to problems of defense conversion in cities and towns in several regions in the aftermath of the Cold War (Markusen and Yudken 1992). Further, 'local' referred to slums in cities in all regions of the country (Goldsmith and Blakely 1992).

What is 'local' about each of these situations is the conceptualization of problems specific to a particular place in relation to a broader context. Although the

problems are local, the broader context is critical in understanding the structures, agents, and processes that affect local contexts: for example, MNCs and related corporate strategies and spatial divisions of labor; the pressures of competition in the global economy on places that must survive or thrive via entrepreneurialism and local funding; changes in power relations among the superpowers and the changing geopolitical map that affects defense-based economies; the changing geography of urban employment and residential patterns in relation to changes in migration, interethnic relations, and firm strategies and their spatial divisions of labor.

The localness of problems does not mean that they are unique. Similar problems may emerge in different places during economic upturns and downturns that sweep across space. What varies among places experiencing similar effects of economic change are the political, economic, social, cultural, and ecological processes and (local) state–society relations associated with locally specific configurations of institutions and behaviors, and their histories. A round of investment is likely to affect places differently in light of different histories of labor-management relations, gender relations, industrial foci, and so forth. Massey (1984) used a geomorphological metaphor to convey this idea: each round of investment brings its own division(s) of labor and related distribution of economic activity—a stratum—that is overlaid on previous socioeconomic 'strata' and impacts a locality relative to the characteristics and combinations of previous strata. Massey conceptualized places in terms of their respective histories of successive layering of socioeconomic strata in relation to wider divisions of labor at national and international scales in association with multilocational firms, often MNCs.

While local problems can vary substantially within nations, the parameters of variation may vary from nation to nation in the context of different modes of state regulation and state–society relations, as well as nationally based systems of production and innovation (Porter 1990). Local economic development issues in the USA, for example, rarely involve an interventionist state at the national scale. In contrast, states in Western Europe or East Asia are much more directly engaged in development processes, although in different ways. Western European countries, with long histories of independence in the global economy, tend to focus developmental efforts internally to resolve problems of regional imbalance, whereas East Asian countries, being 'late developers,' have focused more on asserting a national position in the global economy as a matter of survival (Gershenkron 1962). Thinking about national development as 'local' is especially common in discussions about the role of the state and state–society relations in East Asia, and more generally among countries characterized either as late developers, or nations whose relative position of power has declined and which are confronted with regaining and sustaining competitiveness. Similarly but from a somewhat different vantage point, early dependency theory saw former colonial territories as being actively *under*developed in the modern world system (Gunder Frank 1967)—a situation of local underdevelopment (as opposed to development) relative to national or supranational agents such as MNCs.

2. *The Term 'Economic Development'*

What constitutes economic development depends on perspective. Neoclassical approaches measure economic *growth* in terms of aggregate indicators such as population, employment, productivity, and income. Political economy orientations distinguish economic *development* by focusing on qualitative changes in conditions of living and opportunities for work. At issue is not just employment, but *type* of employment and whether work situations offer opportunities for skill improvement, upward mobility, and appropriate compensation. Qualifying principles have, however been interpreted differently.

Especially in the 1970s and early 1980s, researchers focused on the effects of corporate strategies (notably of large, multilocational firms such as MNCs) on conditions of work. Scholars were concerned with people—workers—and how firms' locational strategies for different types of facilities (e.g., branch plants, administration, or research and development—R&D) signified different consequences for wages, type of work, human capital investment, and so on. During this period traditional Anglo-American corporate strategies prevailed worldwide, which competed on the basis of price (as opposed to quality or innovativeness) such that branch plants routinely employed cheap, low, or unskilled labor. Consequently, the places hosting branch plants were generally considered sites of exploitation and underdevelopment.

Later in the 1980s and throughout the 1990s, Schumpeterian-inspired approaches predominated, offering a view of development with reference to innovative capacity. This perspective developed in the USA in the context of growing numbers of small firms and a concern with how to foster entrepreneurialism; the larger context was one of increasing numbers of competitors in the global economy and, relatedly, new corporate strategies emanating from different sociocultural contexts and entailing different systems of human relations. People, from a Schumpeterian perspective, are of interest insofar as they are useful to firms in establishing and sustaining competitiveness, which generally signifies economic development. Schumpeterian thinking tends to be firm-oriented because analysis typically focuses on the competitiveness of firms or industries (measured in terms of technical advance, productivity, and the like),

generally in relation to skill-based production—a departure from prevalent corporate strategies in the 1970s and 1980s. Ironically, Schumpeterian and neoclassical analyses implicitly converge on the assumption that the well-being (competitiveness) of firms and industries signifies positive conditions for workers, that is, it 'trickles down.'

A recent concern has been the exclusionary and politically charged character of development studies, whether people or firm-oriented. Scholars have recognized researchers' own biases that underlie what appear as simple and value-neutral representations of reality. One important byproduct of those concerns is the identification of differentiated processes relative to different types of people, as well as different yet related dimensions of change (e.g., social, cultural, economic, political, ecological). From this perspective, the constituent components of economic development cannot be aggregated into a single, essentialized assessment.

3. *Context and Approaches to Local Economic Development*

Understandings of local economic development have stressed exogenous and endogenous processes or the articulation of the two.

3.1 *Exogenous Processes*

Consider the 'branch plant economy,' which is characterized by a significant percentage of employment in branch plants of MNCs. Critically at issue are corporate strategy and related management principles. During the 1970s and early 1980s, for example, MNCs commonly produced in large volume for mass markets using cheap and low or unskilled labor. The effect on localities was dependency, especially when corporate strategies were focused on price competitiveness (rather than, say, quality control). In this context, firms routinely relocate as wages or union or other pressures increase. The resultant capital mobility renders localities and their employable populations hostage to the impulses of exogenous corporate decision making, which typically is detached, physically and otherwise, from the localities in question. Moreover, places become pitted against each other in a zero-sum game of wage competition, thereby hampering cohesive resistance among workers across space.

Conditions of high capital mobility and exogenous domination of localities have been approached in various ways. Some scholars viewed processes at different scales as interacting—for example, local histories, industrial configurations and production ensembles with national and international corporate strategies, and industrial change (Markusen 1987, Massey 1984, Storper and Walker 1989). Others in different ways saw localities as objects of corporate processes at larger scales. Although there was general consensus that branch plant economies associated with Taylorist management connoted underdevelopment, some scholars focused on the corporate decision-making component with reference to firms' product and innovation cycles (Malecki 1981), whereas others researched the underdevelopment and dependency components in the context of what had become known as a 'new international division of labor' (NIDL) (Froebel et al. 1980).

3.2 *Endogenous Processes*

Not all local economies are dominated by large firms (Taylor and Thrift 1983) and, moreover, spatializing industry and management is not necessarily neatly hierarchical. Whereas under the model of high capital mobility and exogenous domination scholars interpret change with reference to large firms' locational strategies, *in situ* industrial change or development suggests important endogenous processes. Local economies evolve in various contexts that are neither wholly dependent on, nor the object of, exogenous processes in the sense described in Sect. 3.1.

One of these contexts is a localized version of what Schumpeter called 'creative destruction,' in reference to small, innovative firms successfully challenging the pre-existing corporate order in a region and founding a new wave of economic development. The genesis of Silicon Valley is a case in point.

From another perspective, Scott (1988) developed a theory of local development based on corporate restructuring, specifically the termination of many activities in large, multilocational corporations that had been spread out across regions and the move to subcontract previously internal activity to local independent firms. Borrowing from the 'new institutional economics,' Scott suggested that the cost-effectiveness of externalization strategies during times of market instability (and given the condition of a highly segmented local labor market) would engender dense and cost-competitive interfirm interaction in a local economy, thereby signifying local development based on a new production system. More generally, scholars explained uneven development with reference to new production systems (Storper and Walker 1989) or industrial configurations (Markusen 1987) in new industrial spaces.

But what of old industrial spaces? These are not necessarily subject to decline, as well evidenced by the reindustrialization of the US industrial heartland following deindustrialization. In this context, subcontracting to firms from other sociocultural contexts, such as Japan, gradually changes local corporate culture and social relations from Taylorist (where subcontractor–client relations are typically antagonistic and short run) to collaborative (where, in long-

run relationships characterized by trust, client firms help subcontractors develop knowledge and problem-solving skills to enable increasing independence). Regional restructuring occurs through transnational corporations (TNCs) that become locally embedded and committed to training workers in new, upskilling proceses (Florida 1996).

Another type of context departs from hierarchical models of interfirm relations and is focused on lateral or horizontal networks. This type of context entails communities comprised of a large informal economy, or small firms in mature as opposed to new and strategic industries, or communities that are not major hosts to TNC facilities. Localized small-firm networks can evolve either in the absence of policy to achieve this end (as in the case of the Emilia Romagna region in Italy) or as a result of proactive brokering and assistance via local governance (such as communities in the USA in the 1980s and 1990s reacting to lack of federal funds and industrial decline, or countries such as Taiwan that have institutionalized programs for small firms). Constrained by their size, small firms or informal organizations are unable to achieve the efficiency that is possible for large firms. Scale economies can be achieved by pooling resources, that is networking. As in the case of industrial revitalization, competitiveness lies in an appropriate (inter-organizational) human relations system, one that emphasizes collaboration and cooperation—'social capital' (Cooke and Morgan 1998). Some research focused on understanding how networks achieve competitiveness (e.g., Sengenberger et al. 1990), while other research focused, from a local policy perspective, on how to foster such networks in places where they may not already exist (Rosenfeld 1992). Policy-conscious research interfaced with growing interest on the part of local politicians to assist in the development of entrepreneurial economies to foster change via local and regional institutions (Osborne 1990).

At the national scale, among 'developmental states,' endogenous change may occur through state-assisted industrial development in which foreign capital is effectively manipulated to restrict branch plants (and thus dependency) and encourage subcontracting with local, independent firms, sometimes via networks of entrepreneurs (as in the cases of Hong Kong, Taiwan, and South Korea) (Wade 1990).

3.3 Connecting Exogenous and Endogenous Processes

Researchers interested in endogenous development who are attuned to globalization focus on the connections between particular places and global agents. Unlike the model of capital flight (Sect. 3.1), in which global agents are necessarily 'dominant' and capital is mobile, at issue is how local and global competitive forces may be connected to develop entrepreneurial economies (Storper 1997).

Most MNCs, which produce in one country and market products elsewhere, have evolved into TNCs that functionally integrate marketing and production of different components in a wide variety of countries and localities therein. Global production in this context has become more complex than a strategy of establishing branch plants in diverse locations.

Many TNCs withdraw wholly or partially from production itself and focus energies on coordinating a global system of subcontracting via tiers of networks of firms across nations and world regions (Gereffi and Korzeniewicz 1994). This type of production system can result in positive local change, if the state is developmentally oriented and invests in upgrading of production processes and upskilling of workers regarding the local component of a global commodity and knowledge chain. Alternatively, such networks entail uneven power relations across regional divisions of labor, and some regions remain trapped in a Taylorist-oriented low-skilled, low-cost production component (Hart-Landsberg and Burkett 1998).

Other TNCs remain production-oriented and sink costs into localities by establishing plants that become locally embedded (Amin and Thrift 1994). This is a possibilist perspective in which the long-run investment of foreign capital (capital fixity as opposed to capital mobility) is contingent on: the synergy of preexisting local structures (revisiting Massey's geomorphological metaphor), strategies of TNCs specific to facilities in each place of production, and agency (social interactions, leadership). Questionable, however, in this rosy, albeit possibilist picture, is the partiality of so-called 'development.' Embeddedness itself is partial and usually confined to elite circles; workers, if empowered, are only partially empowered and are disempowered in other respects; and workplaces often remain homogeneous and exclusionary (Ettlinger 1999).

4. The Relevance of Reflexivity

The question 'development for whom?' underscores the partiality of all research on local economic development. Poststructuralism calls for reflection on our biases and assumptions with attention to the consequences of our frameworks, and it opens researchers' minds to questions and information precluded by traditional paradigms. A case in point is the assets based community development (ABCD) approach to local development that considers local assets in ways previously considered unthinkable (Kretzmann and McKnight 1993). The unemployed, for example, may be seen as an untapped and extremely valuable local regional resource. Critically at issue is understanding that developing industries and regions are not coincident because people in places outside of firms are as central to local development as those who are fully employed. This highlights the multiple realities in places, requiring a

conceptualization of local development as a hybrid phenomenon that cannot and should not be essentialized (Escobar 1995).

See also: Development and Urbanization; Development, Economics of; Economic Development and Women; Environment and Development; Globalization: Geographical Aspects; Globalization: Political Aspects

Bibliography

Amin A, Thrift, N (eds.) 1994 *Globalization, Institutions, and Regional Development in Europe*. Oxford University Press, New York

Cohen R B 1977 Multinational corporations, international finance, and the Sunbelt. In: Perry D C, Watkins A J (eds.) *The Rise of the Sunbelt Cities*. Sage, Beverly Hills, CA

Cooke P, Morgan K 1998 *The Associational Economy: Firms, Regions & Innovation*. Oxford University Press, New York

Escobar A 1995 *Encountering Development: The Making and Unmaking of the Third World*. Princeton University Press, Princeton, NJ

Ettlinger N 1999 Local trajectories in the global economy. *Progress in Human Geography* **23**: 335–57

Florida R 1996 Regional creative destruction: production organization, globalization, and the economic transformation of the Midwest. *Economic Geography* **72**: 314–34

Froebel F, Heinrichs J, Kreye O 1980 *The New International Division of Labour: Structural Unemployment in Industrialized Countries and Industrialization in Developing Countries* [trans. P Burgess]. Cambridge University Press, Cambridge, UK

Gereffi G, Korzeniewicz, M (eds.) 1994 *Commodity Chains and Global Capitalism*. Praeger, Westport, CT

Gershenkron A 1962 *Economic Backwardness in Historical Perspective*. Belknap Press of Harvard University Press, Cambridge, MA

Goldsmith W W, Blakely 1992 *Separate Societies: Poverty and Inequality in U.S. Cities*. Temple University Press, Philadelphia

Gunder Frank A 1967 *Capitalism and Underdevelopment in Latin America: Historical Studies of Chile and Brazil*. Monthly Review Press, New York

Hart-Landsberg M, Burkett P 1998 Contradictions of capitalist industrialization in east Asia: A critique of 'flying geese' theories of development. *Economic Geography* **74**: 87–110

Kretzmann J, McKnight J 1993 *Building Communities from the Inside Out: A Path Towards Finding and Mobilizing a Community's Assets*. The Assets-Based Community Development Institute, Institute for Policy Research, Northwestern University, Evanston, IL

Malecki E J 1981 Product cycles, innovation cycles, and regional economic change. *Technological Forecasting and Social Change* **19**: 291–306

Markusen A R 1987 *Regions: The Economics and Politics of Territory*. Rowman and Littlefield, Totowa, NJ

Markusen A, Yudken J 1992 *Dismantling the Cold War Economy*. Basic Books, New York

Massey D B 1984 *Spatial Divisions of Labor: Social Structures and the Geography of Production*. Methuen, New York

Osborne D 1990 *Laboratories of Democracy*. Harvard Business School Press, Boston

Porter M 1990 *The Competitive Advantage of Nations*. The Free Press, New York

Rodwin L, Sazanami H (eds.) 1989 *Deindustrialization and Regional Economic Transformation: The Experience of the United States*. Unwin Hyman, Boston

Rosenfeld S A 1992 *Competitive Manufacturing: New Strategies for Regional Development*. Center for Urban Policy Research, Rutgers, State University of New Jersey, New Brunswick, NJ

Scott A J 1988 *Metropolis: From the Division of Labor to Urban Form*. University of California Press, Berkeley, CA

Sengenberger W, Loveman G, Piore M J (eds.) 1990 *The Re-emergence of Small Enterprises: Industrial Restructuring in Industrialized Countries*. International Institute for Labour Studies, Geneva, Switzerland

Storper M 1997 *The Regional World: Territorial Development in a Global Economy*. Guilford, New York

Storper M, Walker R 1989 *The Capitalist Imperative: Territory, Technology, and Industrial Growth*. Basil Blackwell, New York

Taylor M, Thrift N 1983 Business organization, segmentation, and location. *Regional Studies* **17**: 445–65

Wade R 1990 *Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization*. Princeton University Press, Princeton, NJ

N. Ettlinger

Local–Global Linkages, Environmental Aspects

One of the 'grand queries' in science is understanding linkages between macroscale and microscale phenomena and processes. Biologists, for example, seek to understand linkages between molecules, cells, and organisms; ecologists between patches, ecosystems, and biomes; and economists between firms, industries, and economies (Wilbanks and Kates 1999). In the environmental and ecological sciences, broadly defined, this challenge embraces such profound issues as relationships between local ecosystems and global climate change; relationships between local land use and global economic restructuring; and relationships between the sustainability of local environments and global political change.

Pursuing understandings of such linkages is complicated by the fact that processes at one spatial scale tend to be studied by different communities of scholars than processes at another scale. The respective literatures speak different academic languages, arise from different bodies of theory, and are accustomed to very different units of measurement, even when the subject matter is within a single discipline. When, as is often the case, the linkages cross the boundaries of disciplines, the challenge is much greater—and less often attempted as a matter of careful research rather than qualitative assessment by diverse panels of experts.

The reality, of course, is that global changes in climate, environment, economies, populations, institutions, and cultures converge in localities; and their impacts play out at local and regional scales. At the same time, global changes result from an aggregation

of decisions, actions, innovations, mutations, and other changes that arise in specific localities. Examples of the interplay between global and local scales range from AIDS to ecotourism. Macroscale and microscale are linked intrinsically, and neither can be fully understood without considering the other.

1. How Scale Matters

In understanding environmental systems, the issue of global–local linkages would be relatively insignificant if the local scale were merely microglobal, in other words if a phenomenon or process studied at any one scale could be expected to tell the observer essentially the same thing as study at any other scale. There are a number of reasons, however, why scale matters and why understanding linkages between scales is therefore an important part of the search for knowledge (Wilbanks and Kates 1999).

Several of the reasons have to do with how the world works. First of all, the forces that drive environmental systems arise from different domains of nature and society. For example, population ecology operates at a smaller geographic scale than national or global demographic patterns (Clark 1985). Within this universe of different domains, local and regional domains relate to global ones in two general ways: systemic and cumulative (Turner et al. 1990). Systemic changes involve fundamental changes in the functioning of a global system, such as effects of emissions of ozone-depleting gases on the stratosphere, which may be triggered by local actions (and certainly may affect them) but which transcend simple additive relationships at a global scale. Cumulative changes result from an accumulation of localized changes, such as groundwater depletion or species extinction; the resulting systemic changes are not global, although their effects may have global significance. A second reason that scale can matter is that the scale of *agency*—the direct causation of actions—is often intrinsically localized, while at the same time such agency takes place in the context of *structure*: a set of institutions and other regularized, often formal relationships whose scale is regional or global. Land use decisions are a familiar example. This kind of local–global linkage is especially important where environmental impact mitigation actions are concerned, analogous to hazards behavior. A third reason that scale can matter is that the driving forces behind environmental change involve interactions of processes at different locations and areal extents and different time scales, with varying effects related to geographical and temporal proximity and structure. Looking only at a local scale can miss some of these interactions, as can looking only at a global scale. For instance, geographers have shown that processes of change involve patterns of spatial diffusion that can be generalized, and ecological modelers

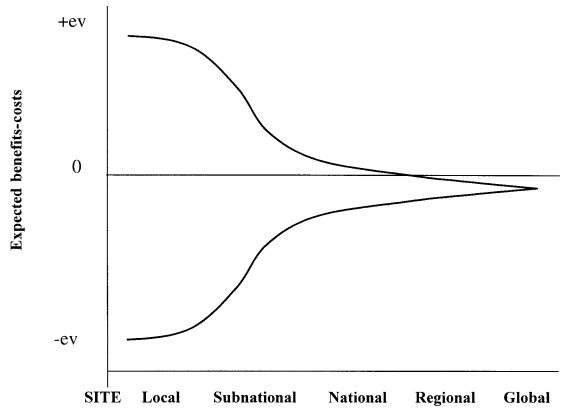

Figure 1
Scale-dependent distribution of impacts of climate change (adapted from Environment Canada 1997)

such as Holling have found that managed biomes are characterized by landscapes with lumpy geometries and lumpy temporal frequencies related to the size and speed of process interactions, shaped by the fact that processes operating at different scales tend to show faster or slower dynamics (Holling 1995).

Several additional reasons have to do with the how we learn about the world. One of the strongest is the argument that complex relations among environmental, economic, and social processes that underlie environmental systems are too complex to unravel at any scale beyond the local. A second reason is that a portfolio of observations at a detailed scale is almost certain to contain more variance than observations at a very general scale, and the greater variety of observed processes and relationships and relationships at a more local scale can be an opportunity for greater learning about the substantive questions being asked. For instance, net benefits/costs from global climate change vary more widely for small areas than large ones (Fig. 1). A third reason is that research experience in a variety of fields tells us that researchers looking at a particular issue top-down can come to dramatically different conclusions from researchers looking at that very same issue bottom-up. The scale embodied in the perspective can frame the investigation and shape the results, which suggests that full learning requires attention at a variety of scales. As one example, Openshaw and Taylor (1979) have demonstrated that simply changing the scale at which data are gathered can change the correlation between variables virtually from $+1$ to -1.

These reasons, of course, do not mean that global–local linkages are salient for every question being asked about environmental and ecological systems. What they suggest is more modest: that examinations of such systems should normally take time to consider linkages between different scales, geographical and

temporal, and whether or not those linkages might be important to the questions at hand. Consider two examples:

(a) *Land use change and climate change*. Obviously, anthropogenic causes of climate change arise from location-specific activities: points, such as a coal-fueled electric power plant; areas, such as a forest or a farming region; and flows, such as interstate highways. Local and regional land use shapes climate change; and reductions in pressures toward climate change are rooted in changes in land uses, although some of the largest contributors occupy relatively small geographical areas. At the same time, climate change affects land use. For instance, it can cause hardwood trees to displace pines or vice versa. It can lead to changes in agricultural cropping patterns. And sea level rise leads to shoreline retreat and coastal land loss. One example of a linkage between the scales is that concerns about impacts at a local level can lead to efforts to promote greenhouse gas emission abatement in order to reduce the likelihood or magnitude of those impacts. An important program of the International Geosphere–Biosphere Program is concerned with land use and land cover change (LUCC), including linkages between very local processes and a variety of larger scales (Turner et al. 1995).

(b) *Biodiversity protection*. Likewise, biodiversity protection has both local and very large-scale dimensions. Most endangered species exist within localized ecological niches with a rather narrow tolerance for change. Other than through zoos or gene banks, their protection requires action at a very local level. On the other hand, many of the forces that endanger them, from environmental pollution to economic and demographic growth, arise at a much larger scale. In addition, where species are migratory, their survival depends not only on a single local biosphere but on a viable migration corridor between seasonal biospheres as well, at the scale of a region or beyond.

2. Tracing out the Linkages

Our understanding of global–local linkages in the environmental and ecological sciences is still poorly developed and is a promising subject for further research. As few hypotheses based on rather fragmentary evidence, however, can suggest some directions for discussion.

It appears, for instance, that whether global change is likely to cause local change depends on how significant and discontinuous the global change is, whether localities are substantially subject to external driving forces, and whether global processes are relatively closely linked with local processes. It appears that whether local change is likely to cause global change depends on the magnitude and force of the local change, whether the local change is focused on a key driving force in a global process, and whether the local change either (a) is associated with structures/processes for dissemination or (b) is associated cumulatively with similar local changes elsewhere. It appears that whether a global–local link is likely to be disruptive depends on existing stress levels in localities and their adaptability and coping capacity. It also appears that divergences between local and global optimality are affected by mobility, which is clearly increasing in many global networks; in some cases the divergence can be greater where mobility and other adaptation mechanisms are limited, while in others it can be magnified by a loss of control at the local level over the nature and speed of change.

3. Linking the Global and the Local

In seeking to improve the understanding of macro–microscale linkages, a common conceptual approach has been to depict the systems operating at different scales as hierarachies: classes of processes in which each smaller unit is nested within a larger unit, which can be associated with bodies of theory about interactions within and between scales (e.g., Allen and Starr 1982). Examples include ecosystems and economic markets. Such approaches have often been fruitful, but they are often criticized for arbitrarily dividing complex systems into a limited number of levels of 'organization.'

Current efforts to improve the capacity to trace out, understand, and if appropriate modify linkages between the global and the local in environmental and ecological systems include attention to five central problems:

(a) Downscaling global and other large-scale models to understand issues at a regional or local scale, such as implications of global environmental change. Recognizing the limitations of top-down paradigms based on global or near-global scale modeling, the research community is increasingly active in moving toward more detailed geographic scales and topical richness, using both numerical (i.e., model-based) and empirical (i.e., statistical-based) approaches. Challenges include data availability at detailed scales, the complexity of process relationships as models become more like the real world, and in some cases computational capacity (although advances in computing have reduced this constraint considerably). Appetites on the part of model users for downscaling appear insatiable; as a result, this trend is expected to continue.

(b) Upscaling data from points or very small areas to understand issues at a regional or global scale. Many kinds of data pertinent to macroscale issues are gathered at specific points, ranging from meteorological observations to crop production; but in many cases these data cannot simply be aggregated to estimate larger-scale values, such as regional agricultural production or climate processes. For instance, the data may fail to meet standards for valid sampling, and they may fail to represent stochastic and geographic variability in representing how processes

work. As one example, it has been shown that an estimated response to an 'average' environment can be a biased predictor of a 'true' aggregate response (Templeton and Lawlor 1981).

(c) Integrating top-down and bottom-up perspectives to arrive at a comprehensive understanding. Where downscaling and upscaling come together is when the two approaches are merged to attain an integrated understanding of processes, benefiting from the strengths of both perspectives. This kind of integration is usually done in one of three ways: (i) by converting both perspectives to a common metric, usually a 'meso' or intermediate scale such as one-half degree of (equatorial) latitude and longitude (e.g., Easterling 1997); (ii) by comparing the two different answers and adjusting the approach used in arriving at each, iterating until a single answer is derived (applied, for instance, in the integrated environment/economy model of the International Institute for Applied Systems Analysis); and (iii) by comparing the two different answers and seeking a meta-understanding, incorporating both perspectives but arriving at a third answer (e.g., the multidisciplinary Susquehanna River Valley study conducted by the Pennsylvania State University: see Abler et al. 2000). In many cases, integration is partly a matter of bridging between analytical styles, for example, top-down models and bottom-up case studies. In addition, it is often a matter of reconciling differences in process assumptions, theoretical foundations, and perceived standards as to what constitutes the best science.

(d) Resolving differences in scales between relevant processes. Because different interacting processes may operate at different scales, for example, between the scale of ecosystems and the scale of governmental units making decisions about them, efforts to incorporate a variety of linkages in a single analysis or action must often confront problems of 'scale discordance.' Among the avenues being investigated are adaptive approaches to analysis and management, which permit modifications of the scale as more is learned about the relevant processes and their interactions.

(e) Accounting for cross-scale dynamics. Because so much data collection and analysis occurs at a particular scale, data are often scarce about cross-scale relationships and interactions. Strategies for reducing this gap are being considered through the establishment of richer information infrastructures, especially regarding longitudinal data sets, but assuring continued financial support for data collection structures is a continuing challenge.

Much of this growing effort to improve the understanding of global–local linkages is motivated not only by scientific curiosity but also by social needs to deal with environmental problems such as reducing environmental pollution, coping with possible impacts of global climate change, and responding to threatened losses of ecosystem quality and variety.

See also: Environment and Development; Environmentalism: Preservation and Conservation; Global Environmental Change: Human Dimensions; Human Ecology; Human–Environment Relationship: Indicators; Human–Environment Relationships; International and Transboundary Accords, Environmental; International Organization; Scale in Geography

Bibliography

Abler D G, Shortle J, Rose A, Oladosu G 2000 Characterizing regional economic impacts and responses to climate change. *Global and Planetary Change* **25**: 67–81

Allen T F H, Starr T B 1982 *Hierarchy: Perspectives for Ecological Complexity*. University of Chicago, Chicago

Clark W C 1985 Scales of climate impacts. *Climatic Change* **7**: 5–27

Easterling W E 1997 Why regional studies are needed in the development of full-scale integrated assessment modelling of global change processes. *Global Environmental Change* **7**: 337–56

Environment Canada 1997 *The Canada Country Study: Climate Impacts and Adaptation*. Adaptation and Impacts Research Group, Downsview, ON

Holling C S 1995 What barriers? What bridges? In: Gunderson L H, Holling C S, Light S S (eds.) *Barriers and Bridges to the Renewal of Ecosystems and Institutions*. Columbia University Press, New York, pp. 3–34

Openshaw S, Taylor P J 1979 A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In: Wrigley N (ed.) *Statistical Applications in Spatial Science*. Pion, London

Templeton A R, Lawlor L R 1981 The fallacy of the averages in ecological optimization theory. *American Naturalist* **117**: 390–3

Turner B L II et al. 1990 Two types of global environmental change: Definitional and spatial scale issues in their human dimensions. *Global Environmental Change* **1**: 14–22

Turner B L II et al. 1995 Land Use and Land-Cover Change: Science/Research Plan. IGBP Report #35, International Geosphere–Biosphere Program, Stockholm

Wilbanks T J, Kates R W 1999 Global change in local places. *Climatic Change* **43**: 601–28

T. J. Wilbanks

Location: Absolute/Relative

Location refers to the relationship of a place to other places in a space. The space can be absolute or relative. Imagine absolute space as a container, or an array of containers, in which elements are found. To move an element means to change its position among the locations available. Locations remain fixed; elements repositioned have changed location. In contrast, relative location is defined by the positions of the places relative to each other. Places are nearest their neigh-

bors where they share a boundary. An intervening nearest neighbor separates a place from its second nearest neighbors. Places are remote from one another where connecting them requires crossing many boundaries. Some boundaries and territories are more easily crossed than others, which means that the intervening places affect the extent of separation or relative location.

The distinction between absolute and relative location has both philosophical and practical importance (Couclelis 1992). At the philosophical level, there are two prevalent views of geographic space: space as a plenum, fixed, and continuous with elements contained within it (absolute); and space as neighborhood, defined by the interaction and exchange between places (relative). The issue will not be resolved in favor of one or the other view, because each is valid and useful in different applications—not unlike the wave and particle views of electromagnetic energy.

In practice, the difference arises in the way geographic space is depicted as either a fixed array of pixels (smallest distinguishable grains) in a raster database, in which spatial elements are distinguished by differences in values of individual pixels, or as topological relationships between spatial objects in a vector database. Satellite images, aerial photographs, newspaper pictures, and art portfolio prints are all raster images, in which spatial elements are distinguishable by differences in value, intensity, or color of individual pixels. The size of the pixels defines the resolution of the images. By contrast, objects in vector models are defined by their edges or end points, and by adjacency and containment relationships with other objects. A road network may be modeled as a set of points (intersections and end points) and lines (road segments). You can travel through the network by moving through a sequence of lines connected by points of intersection. Attributes of the street segments and intersections, such as number of lanes, direction of travel, posted speed limits, signal lights, and turn restrictions, affect the measure of separation between locations. Changes in any of these attributes change the relative location of places. The reference above to spatial resolution is germane, because space has other attributes that affect the notion of location, including scale, distance, and dimension (see *Scale in Geography*; *Place in Geography*).

1. Place or Location: Scale Dependent

A place is a fundamental spatial element, spatially undifferentiated within and positioned relative to other places beyond. The dichotomy *site and situation* establishes this level of spatial abstraction and is scale dependent. It distinguishes what is within a site from what is exterior. The site or place may be any size, such as that of a nation, a city, or a residential neighborhood. Situation or location refers to the external relations of places. Ullman (1956) noted that linkages beyond the borders of a place require exchange or movement between locations. Consequently, as transportation and communication technologies evolve, the relative locations of places change. The familiar expression that the world is shrinking alludes to this trend (see *Access: Geographical*; *Spatial Interaction*).

A sense of place is evocative of an emotional attachment to a locale. Normally this requires repeated and deep experience in a place, be it home or wilderness. In the writings of structuralists, exemplified by works of Sack (1986) and Curry (1998) and by humanists such as Yi-Fu Tuan (1977) and J. B. Jackson (1994), place trumps location in the processes in which society attaches sentiment, value, and power to socially constructed places. Unlike in Ullman's works, maps do not play a prominent role in these traditions, because consideration of place takes president (see *Territoriality: Geographical*; *Humanistic Geography*; *Geographic Information Systems: Critical Approaches*).

Gatrell (1983) defines space as a relation defined on a set of objects. Keeping in mind that places are spatial entities, the set theory view of locations allows consideration of topological relations between places, such as nesting, clustering, and adjacency, without reference to distance (Worboys 1995) (see *Geographic Information Systems: Critical Approaches*; *Spatial Pattern, Analysis of*).

2. Metric Space and Dimensions

Distance is the defining relationship between places in a metric space. The concept of distance in a metric space has very specific characteristics, including: that the distance from a place to itself is zero; that distance to any other place is a positive real number; that distance between places is the same in each direction (symmetry); and that the shortest distance between two places is the direct path between them (triangular inequality). Being able to use a metric space in an analysis is useful because familiar geometry and algebra hold. However, spaces often do not possess all the properties of metric distance. In city street networks with overpasses and one-way streets, neither symmetry nor triangular inequality holds. A finite network of points and lines is an appropriate model for city street space.

Location in a metric space can be specified by a Euclidean coordinate system, in which the number of axes defines the dimensions of the space. In a flat, two-dimensional surface, ordered pairs of real numbers (x, y) specify locations. An ordered triplet of numbers is required to specify location in a three-dimensional space. To complicate matters further, although imbedded in a three-dimensional space, the earth's

surface is itself two dimensional, and two numbers (latitude and longitude) are sufficient to locate a place on it. It is also a closed space, which means there is a finite maximum distance separating places (see *Spatial Data*; *Cartography*).

3. Applications

Location is a central theme in a class of practical location–allocation problems in which the best location for sources and volumes of flows are investigated. Specifying the locations of warehouses, so minimizing the total cost of delivering food to supply a chain of supermarkets, is an example of an application of location–allocation models. Determining the arrangement of emergency vehicle stations in an urban area, such that no place is more than a certain number of minutes from help, is another example (Ghosh and Rushton 1987) (see *Spatial Optimization Models*; *Spatial Interaction Models*).

The notion that space can be defined by an ordered n-tuple of real numbers leads to many interpretations. The spatial properties of a place may be defined by the triplet (x, y, $\mathbf{a}$), where x and y are spatial coordinates and $\mathbf{a}$ is a vector of attributes of the place. When $\mathbf{a}$ is elevation, displaying $\mathbf{a}$ at all locations creates a terrain model. If $\mathbf{a}$ is a collection of attributes, $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3 \ldots \mathbf{a}_n)$, displaying each separately creates a series of overlays that show spatial juxtaposition of the variables. In *Design with Nature*, an early well-known application of this method, Ian McHarg used multiple attributes of places, including physical, social, economic, and historical conditions, to assess the suitability of territories for land use development (McHarg 1967).

Alternatively, each element in the vector $\mathbf{a}$ could be considered a dimension. For example, an n-dimensional ecological space might include soil moisture, temperature, slope, and direction of exposure as dimensions of the space. Places could be located in a moisture/temperature space or in an n-dimensional space that includes all the site variables. Clusters in this space represent multidimensional similarity of places. Studies of spatial behavior—that is, studies of how people perceive, move about, and use space—have employed this type of multivariate analysis (see *Behavioral Geography*).

4. Locational Referencing

There are many practical ways of specifying location. Street addresses and zip code zones are used to locate homes and businesses. Cadastral surveys establish ownership of land parcels for tax purposes. The Census Bureau uses a nested hierarchy of areal units to report demographic data. Police use road intersections and mileposts to locate emergency incidences along roadways. State plane coordinate systems and global referencing systems are used to specify locations referenced to the earth's surface— in particular, the familiar latitude and longitude reference system. When locational referencing systems are based on different models of space, problems arise in the process of comparing or converting locational information between systems, because not all the attributes of spaces are likely to be preserved in the transformation. Providing a Rosetta stone for transferring locational information between locational referencing systems involves using standardized symbols and proper map projections, but semantics and purpose also dictate what is needed.

More broadly, whenever location of places enters into consideration, the nature of the spatial–temporal context rises to importance. The issue is more than a technical and semantic matter. The user's purposes and fundamental views of space as absolute or relative, whether realized or not, will underlie understanding and actions.

Bibliography

Couclelis H 1992 People manipulate objects (but cultivate fields): beyond the raster–vector debate. In: Frank A, Campari I, Formentini U (eds.) *Theories and Methods of Spatial–temporal Reasoning in Geographic Space*. Springer-Verlag, Berlin, pp. 65–77

Curry M R 1998 *Digital Places: Living with Geographic Information Systems*. Routledge, New York

Gatrell A 1983 *Distance and Space: A Geographical Perspective*. Clarendon Press, Oxford, UK

Ghosh A, Rushton G 1987 *Spatial Analysis and Location–Allocation Models*. Van Nostrand Reinhold, New York

Jackson J B 1994 *A Sense of Place, a Sense of Time*. Yale University Press, New Haven, CT

McHarg I L 1967 *Design with Nature*. Wiley, New York

Sack R D 1986 *Human Territoriality: Its Theory and History*. Cambridge University Press, Cambridge, UK

Tuan Y-F 1977 *Space and Place: The Perspective of Experience*. University of Minnesota Press, Minneapolis, MN

Ullman E 1956 The role of transportation and the bases for interaction. In: Thomas W, Jr (eds.) *Man's Role in Changing the Face of the Earth*. University of Chicago Press, Chicago, pp. 862–80

Worboys M 1995 *GIS: A Computing Perspective*. Taylor & Francis, London

J. D. Nystuen

Location Theory

1. The 'Where' Question: Space is Crucial

Location theory is concerned with the intriguing question where economic activity can be found. It addresses the geographic dimension of economic

decision and analyses both the behavior of firms and of households. Thus the locus, the geographic place, is of prime interest in location analysis.

More specifically, the main question in location theory is: which are the motives to choose a particular location and which are the geographic implications? And clearly, there is not a single, conclusive answer. A student who has just left the university will usually have different aims in mind when seeking for a residence than his parents who have just retired. But despite variety in motives and goals, there are many things which are common to anyone seeking a residential location (such as affordability, accessibility, quality of life, etc.).

The same applies to business firms. An ice-cream vendor will not open a shop in a dead-end street, and a diamond seller will not opt for a location in an underground station. Both may have profit motives, but their decisions where to locate may be totally different. But also here, if one observes the actual location patterns of business firms, there are some similarities. The location of households and firms does not exhibit a random pattern worldwide, but displays many regularities. For instance, all over the world we find a population concentration and business concentration in river basin areas and delta areas, a phenomenon already observed by Adam Smith more than 200 years ago.

Location theory does not only investigate the spatial point decision of an individual actor (household, business firm), but looks also into the regularities (patterns, geographic structures) that emerge out of the decision of individual actors. Locational decisions are usually influenced by so-called externalities, which means that the decisions made previously by a group of actors impact on the current decisions of new market entrants. This explains why cities emerge as a spatial conglomerate of many individual actors seeking urbanization economies, or why industrial complexes arise as a consequence of agglomeration advantages (such as a joint use of infrastructure). Thus, location theory does not only have a micro aspect, but also a macro (collective or social) aspect; locational decisions are at the end usually interactive decisions.

Clearly, once a spatial pattern of economic activity has been realized, it will influence other patterns as well. Consumers tend to be attracted by the proximity of shopping facilities, business centers tend to orient themselves towards the presence of a critical mass of clients with a high purchasing power, etc. Thus, there is interactivity in location at both the micro and the meso/macro scales.

The next section offers a very concise presentation of some highlights from traditional location theory, followed by an exposition of some modern developments. After a discussion of firms' location decisions, households' decisions will be looked at. The impact of the modern network economy and of information and communication technology (ICT) will be dealt with as well. The article will be concluded with some prospective remarks.

2. *Conventional Wisdom on Location Decisions of Firms*

Location theory forms a central element in regional economics and economic geography. Location and trade were already regarded by the grandfather of economics, Adam Smith, as two interwoven phenomena: location impacts on trade flows and trade impacts on location decisions.

In the history of economics a great variety of contributions on the nature, motives, and implications of the location of production can be found. We will offer only a few examples here from this rich history.

The location and land use in agricultural production was studied extensively by Von Thünen (1842), by using a simple profit maximization scheme for spatial product choices, as a function of revenues, product types, and transport costs. He was able to demonstrate the existence of a set of concentric rings of agricultural products, based on Ricardian comparative cost theory leading to product specialization. Von Thünen's rent theory has formed the basis for the urban land rent theory, which has become a central focus in urban economics

The location of industrial firms was investigated by Weber (1909) by looking for a cost minimizing location solution of a new firm on the basis of resource inputs (and other material inputs) and product outputs (to be shipped to the market). This intriguing and complex question was solved by Weber (1909) by using a triangular force field of input and output points, within which the optimal location has to be found.

The spatial concentration of economic activity in a city has also been studied intensively, starting from the work of Marshall (1925), who argued that agglomeration benefits are central for city formation. He referred to externalities which caused such benefits, in particular better information and skills, trade growth, specialized equipment, and availability of skilled labor.

The shift towards multiproduct economies and the spatial inter-relationships between multiple products received due attention some 50 years ago in the context of the so-called central place theory, developed by Christaller (1933) and Lösch (1954). The basic idea is that the location of economic activity is subject to agglomeration advantages. As a consequence, economic activities are not spread randomly, but manifest themselves in clusters. Moreover, in the light of product diversity, price elasticity, and the existence of daily and non-daily goods, a spatial hierarchy is a logical outcome of a series of profit-maximizing location decisions of firms. This hierarchy means that a place with a certain ranking order also includes firms

producing goods with a lower order. This theory has become a landmark in location theory and has had far-reaching implications for spatial planning.

In the postwar period, a wide range of new and refreshing contributions has also been offered, such as the growth pole theory, the cumulative causation model, and the forward–backward linkage view on regional development.

The complexity of location decisions in a continuous space or in a discrete network configuration has also prompted a rich research field in the operations research discipline, where many attempts have been made to derive mathematically the optimal site of public facilities or private goods, especially in the context of a multiproduct location decision. It should be noted that in a discrete space with a finite number of location options the number of combinatorial possibilities may be formidable, so that rather advanced mathematical software has to be developed.

In the 1990s much attention was given to a rigorous analysis of the link between trade and locations, especially in the spirit of the monopolistic competition approach advocated by Krugman (1991). In this analysis, spatial accessibility does not only favor transport and trade, but geographical clustering is also positively influenced by a strong critical mass of local demand, by increasing returns to scale in the production sector and by a good transportation system in the area. This will be discussed in more detail in Sect. 3.

3. *Location and Growth*

Locational decisions are usually not stand-alone choices of a firm, but are part of a broader set of firm decisions, e.g., on market areas, marketing channels, technology to be used, image, etc. Against this background transport—and thus trade—plays an important role. This fact was already recognized in the well-known Hecksher–Ohlin theory on international (or interregional) trade flows, where trade intensity between two countries is dependent on the comparative advantage of the use of factor inputs. Clearly, also the degree of homogeneity of products to be traded plays a role here. Based on the assumptions of monopolistic competition, Dixit and Stiglitz (1977) have tried to make the Hecksher–Ohlin theory more realistic, inter alia by allowing for product differentiation. In this way, both inter- and intra-industry trade flows and locational patterns can be analyzed. This integration of trade theory and location theory forms one of the main characteristics of modern growth and trade theory.

In this context there is also a need for a more dynamic perspective to spatial equilibrium phenomena. This presupposes a more integrated view on locational decisions and emerging market patterns, thus necessitating an economic growth perspective. Consequently, trade and location are endogenously determined in modern economic growth theory, as developed inter alia by Krugman (1991). This also means that factor mobility, transportation costs, and transaction costs become an integral part of modern theories on location and urban or regional growth.

In a recent study by Fujita et al. (1999), the authors offer an attempt to re-track economic geography (and regional science), while they also aim to build a 'new economic geography' based on a few vigorous economic principles. The authors signal a regrettable dividing line between mainstream economics and the economics of location. They aim to build a bridge on the basis of a few simple (certainly not universally valid) concepts, in particular the imperfect competition model marked by increasing returns to scale (originating from Dixit and Stiglitz 1977). In an open (multiregion or multicountry) system various types of spatial agglomeration patterns may emerge, depending inter alia on transport costs, forward and backward linkages, and immobility of resources. The authors offer a new perspective in their effort to bring general economics closer to regional economics. They illustrate their arguments by referring to core-periphery phenomena, agriculture, urban systems, city size, transportation, international trade, and industrial clustering. The book forms a clear manifestation of the rigour of solid economic analysis for the explanation of the spatial patterns and evolution of economic activity.

The two pillars of regional economics are certainly formed by agglomeration economies and generalized transportation costs. Much emphasis is made in the book on the economic analysis of urbanization and scale advantages, but less on transportation costs. In our era of ICT development where many economists advocate the 'death of distance,' it would have been necessary to pay more attention to both psychological and virtual distance costs, and their implications for the spatial organization of our world (see Sect. 6).

It is clear that a further integration of location theory, transportation science, trade theory, and industrial organization seems to be a necessary condition for a rigorous way forward.

4. *Where Do People Wish to Live?*

The locational choices of households can be looked at from many perspectives exemplified by a great variety of historical, cultural, social, economic, and geographical views. The most prominent and visible pattern in the spatial distribution of households is the clustering of people in the vicinity of firms that demand workers for their economic activities. For example, the Industrial Revolution, based on new, steam-driven technology, created mass-produced goods at particular geographical locations where the inputs for the indus-

trial production process were found. The workers needed for this new mass production of goods were forced to settle near the factories, given the limited transport facilities at the time and their low financial means.

The ultimate example of spatial clustering of economic agents (firms and households) is observed through city formation and city growth, both in developed (industrialized) countries and in developing countries. Nevertheless, some people do not live in the urbanized areas with the main employment centers, but commute from their home regions to their workplace in the city. This travel distance of households from home to work forms the principal focal point in the classical urban location theories that were developed to explain the spatial structure of agglomerations (see Alonso 1964, Muth 1969, Mills 1972).

The old economic wisdom on urban location decisions of households builds on the functioning of the market for land from which the urban location patterns were derived through the application of standard economic laws of supply and demand. The classical, elementary model for the urban location of households considers the decision-making process of households as a place—utility maximizing process in which a crucial role is played by the travel distance from home to work (commuting distance). For reasons of expositional convenience, company businesses are supposed to be located in the center of the agglomeration (but this was relaxed in later versions of the classical models). In this case, all households would prefer to live in the center, if differences in land prices across the urban space were not yet taken into account (thereby minimizing their commuting costs). However, the excess demand of households for locations in the vicinity of the business center forces the prices for this scarcely available land up, because the households are willing to pay more for central locations. Now, a fundamental locational choice for households emerges in which they have to trade-off the conditions of living near the center (low commuting costs, but high costs for housing) with the countervailing features of living at some further distance of the center (high commuting costs, low housing costs). The difference in the costs for land (housing) will be determined—in the spirit of classical urban models—by an equilibrium state of the land market in which the household's well-being (or utility) is equalized over all possible locations within the urban space. The main economic principle of bidding for those (urban) locations that households consider to be attractive in some sense remains valid, when differences across households (e.g., size, income, preferences), the existence of urban employment subentries, governmental regulations of land and housing markets, and other locational variations or neighborhood characteristics (public and private facilities, environmental amenities, green areas, socioeconomic composition) are taken into account (see Straszheim 1986). This means that besides valuing shorter distances to the city (employment) center, households bid up housing prices in higher valued neighborhoods.

5. *Residential Relocation of Households: Search and Thou Shall Find*

People do frequently move to other locations to obtain better housing during the course of their expected lifetime. There are several reasons for this change of residence and most often the motives are job related (labor migrants) or housing related (size and facilities of the dwelling unit). The residential move could, of course, also be due to health, family moves, change in household composition (marriage or living together, childbirth), or improved transport facilities. Short-distance moves are usually associated with housing related reasons (such as a change in household income), whereas long-distance residential moves are likely to be triggered by job changes (see Clark and Van Lierop 1986). In general, people consider moving when some external shock makes the current location suboptimal, that is, the household becomes dissatisfied with the dwelling itself, the housing environment (social bonds with people living in the neighborhood, availability of parks, recreational facilities, playgrounds and schools), or the employment opportunities that are accessible from the current location through commuting. This shock may originate from the internal household circumstances (e.g., childbirth or a divorce), be related to market forces (rent, interest rate, commuting costs) or be due to externalities such as noise caused by new neighbors or, for example, pollution of the local environment (air, water, soil).

The propensity to move creates a search process in which households try to improve their current situation. At the heart of this search process, households face a lack of information on potential dwellings at alternative locations and are confronted with time and money costs to acquire such information. Moreover, in the search process the household has to decide on the spatial area of search. A residential move will then be made in case the expected gain is higher than the moving costs (this may also include nonmonetary costs related to the psychological value attached to the local social network of contacts).

Accordingly, it is well-known that residential moves are affected predominantly by age which represents the stage in the life cycle. Furthermore, households do not necessarily move in response to small changes in the environment due to transaction (moving) costs. From a methodological point of view, the search behavior of households on the housing market can be investigated appropriately by employing duration modeling techniques in which variations in residential duration are used to analyse housing market dynamics (preferably in a spatial context). Recent theoretical developments along the search-theoretical angle have

also stressed the simultaneity of the decision to move and to change jobs, thereby taking into account commuting costs (see Van Ommeren et al. 1996). One of the central recent issues in this respect is also whether people follow jobs (as outlined in the classical urban models) or jobs follow people (as argued in previous sections on industrial location). Empirical research has established evidence that firms tend to follow—but not to lead—workers to suburban areas (see White 1999).

Last but not least, it should be emphasized that in many Western societies the possibilities to realize the demand for particular locations are limited due to regulatory measures on the housing markets and restrictive spatial planning policies. In many western countries, both housing demand and supply are regulated extensively (by means of price regulation, allocation devices by local communities, etc.).

6. *Modern Residential Location Behavior in a Network Society: The Death of Distance*

Nowadays households live and work in an information society in which economies are characterized by networks and increasingly build on the input of knowledge (and less on the classical production factors capital, labor and land). This new paradigm for the world economy—in which Information and Commutation Technology (ICT) plays a decisive role (see Castells 1996)—has significant consequences for the spatial distribution of activities (at the macro level) and the location decisions of agents (at the individual level). The globalizing world creates footloose agents that are not tied to specific geographical places (either market- or input-related), but at the same time localization trends can be noticed with an increasing importance of local locational factors (such as favorable living conditions, the environmental quality of life and housing facilities).

So what are the main effects of ICT on household location and relocation decision-making? First, ICT greatly influences the search process for new dwellings (and alternative locations), since, search costs become virtually zero due to easy and full access to information (e.g., via the use of Internet). It may, however, be difficult to filter adequately information through these new media and also to obtain reliable information. Moreover, the actual moves still remain to be costly due to monetary and nonmonetary (psychological) moving costs. Second, the ICT-revolution, based on rapidly improving information transport, has initiated a process of internationalization and globalization, and the emergence of the service (network) economy. More specifically, technological progress leads to larger communication flows between firms and households in different regions and cities. In addition, the immaterial character of information removes physical barriers for households in their locational preferences. As a consequence, geographical distance becomes less and less important (due to the use of Internet, e-mail and fax); this seems to be the death of distance because of the enormous decrease in communication costs. At first sight, people may choose to work at any place they prefer, thereby enabling them to live where they want without taking into consideration the transportation (commuting) cost, as was highlighted in the classical view on household location behavior. However, it also appears that local networks of firms are vital to the exchange of so-called tacit knowledge (embodied in people) by means of face-to-face contacts. For example, headquarters of major companies often prefer to be located in large agglomerations with multimodal transport facilities (especially by air). This implies, then, that geographical proximity remains an important factor in the (re-) location decision of particular groups of households. Consequently, the basic trade-off for households between commuting costs and housing costs does not disappear in their locational choice process.

7. *Location of Public Facilities*

As has been argued in the previous sections on household locations, the proximity of private and public facilities will affect the locational decisions of households. From their perspective, the location of facilities is taken for granted. The suppliers of facilities do also have to consider the geographical location of their services (given the location of potential customers). In the real world, one observes that a wide range of services to local inhabitants is controlled by public authorities either at the local or at a regional/national level. The various types of public facilities can be classified as follows (see Johansson and Leonardi 1986).

Travel to facilities: consumers visit services offered such as schools, libraries, concert buildings, parks, hospitals, etc.

Insurance service facilities: consumers are protected against emergencies due to the existence of certain facilities as fire brigades, dikes, police force, army, etc.

Delivered from facilities: consumers obtain benefits (or costs) at the place of residence as a result of certain facilities such as broadcasting, pollution from airports, etc.

In case of the first two classes of facilities, consumers travel to the service concerned and the received benefits will be distance-dependent (due to transportation costs). For the third class, it holds that the consumer obtains the services at home.

In general, facilities are considered to be public if the services supplied by the public authorities also have a 'public good' nature. In short, consumption of a public facility does not hinder the consumption by other consumers of that same good (in technical terms, this means that public goods are nonexcludable,

non-rejectable and allow for joint consumption). Furthermore, the public authorities decide on the particular location of the public facilities. Therefore a major concern in this location decision is given to the distance of consumers to the service at hand. The distance dependence does, however, also imply that the public goods nature is not complete in the sense that some will have easier (cheaper) access to the service than others. To sum up, the authorities first have to select their set of public facilities (that is, they have to estimate the demand for public goods) and next, they have to choose how to locate these facilities, given the spatial distribution of potential consumers.

A well-known hypothesis on the allocation of public facilities across space originates from the early work of Tiebout (1956) who postulated that people reveal their preferences for local (municipal) services by 'voting with their feet.' By this 'feet-voting' behavior, Tiebout meant that households migrate to municipalities that offer the best combinations of preferred services and associated taxes. It also gives municipalities the scope to affect the supply of facilities. As a consequence, local communities can influence the size and composition of their inhabitants (through the in- and outmigration of households) as well as their local tax base to finance these facilities. Of course, households will differ in their spatial ('voting by feet') reactions to the provision of various types of facilities. For example, young households with children will probably be attracted by places of residence offering educational services, playgrounds, and recreational amenities, whereas older households are more likely to move to places with support and care services and social-cultural amenities. One of the most important implications of the concept of 'voting by feet' is that the allocation of goods takes place more efficiently than with majority voting. This conclusion does, however, presuppose that housing search and moving are costless, ideal places can be found for each household, and that local goods do not yield benefits to people living outside the community. This set of conditions is hard to meet in practice.

To conclude, the ultimate challenge to analyse jointly household location and public facility location would require an interactive modeling framework of spatial equilibrium in which household locational choices affect governmental decisions on public facility location and vice versa.

8. Concluding Remarks

Location theory is essentially the heart of economic geography and regional economics, but it is also linked increasingly to industrial organization and trade theory. In particular against the background of the current globalization process, location theory is positioned in a global force field. This force field leads to many rapid responses and behavioral adjustments of business firms, so that stable and robust locations are increasingly replaced by nomadic types of business behavior. The rapidly emerging technological changes (including ICT) encourage location theory to become a battlefield of industrial forces. Clearly, this dynamic evolution has both advantages and disadvantages, and much empirical work would be needed to find out the pros and cons of this dynamic change.

It also ought to be noted that from an economic perspective much attention is given to rational behavior of individual firms or households. In reality, we observe many behavioral responses which are instigated by derived factors, such as neighborhood effects, uncertainty analysis, intra-firm transaction strategies, outsourcing mechanisms, etc. This suggests that also the behavioral basis of location theory certainly needs to be reinforced.

See also: Economic Geography; Firm Behavior; Housing Economics; Industrial Ecology; Industrial Geography; Locational Conflict (NIMBY); Residential Segregation: Geographic Aspects; Residential Segregation: Sociological Aspects; Spatial Labor Markets; Spatial Optimization Models; Spatial Pattern, Analysis of

Bibliography

Alonso W 1964 *Location and Land Use*. Harvard University Press, Cambridge, MA

Castells M 1996 *The Rise of the Network Society*. Basil Blackwell, Malden, UK

Christaller W 1933 *Die Zentralen Orte in Süddeutschland*. Fischer, Jena, Germany [trans. Baskin C W 1966 *Central Places in Southern Germany*, Prentice-Hall, Englewood Cliffs, NJ]

Clark W A V, van Lierop W 1986 Residential mobility and household location modelling. In: Nijkamp P (ed.) *Handbook of Regional and Urban Economics. Regional Economics*. North-Holland, Amsterdam, Vol. I, pp. 97–132

Dixit A K, Stiglitz J E 1977 Monopolistic competition and optimum product diversity. *American Economic Review* **67**: 297–308

Fujita M, Krugman P, Venables A 1999 *The Spatial Economy: Cities, Regions, and International Trade*. MIT Press, Cambridge, MA

Johansson B, Leonardi G 1986 Public facility location: A multiregional and multi-authority decision context. In: Nijkamp P (ed.) *Handbook of Regional and Urban Economics: Regional Economics*. North-Holland, Amsterdam, Vol. I, pp. 133–70

Krugman P 1991 Increasing returns and economic geography. *Journal of Political Economy* **99**: 483–99

Lösch A 1954 *The Economics of Location*. Yale University Press, New Haven

Marshall A 1925 *Principles of Economics*. MacMillan, London

Mills E 1972 *Studies in the Structure of the Urban Economy*. Johns Hopkins University Press, Baltimore, MD

Muth R 1969 *Cities and Housing*. University of Chicago Press, Chicago

Paelinck J H P, Nijkamp P 1985 *Operational Theory and Method in Regional Economics*. Avebury, Aldershot, UK

Smith A 1776 *An Inquiry into the Nature and Causes of the Wealth of Nations*. Liberty Classics, Indianapolis, IN
Straszheim M 1986 Urban residential location. In: Mills E S (ed.) *Handbook of Regional and Urban Economics: Urban Economics*. North-Holland, Amsterdam, Vol. II, pp. 717–57
Thünen J H von 1842 *Der isolierte Staat in Beziehung anf Landwirtschaft and Nationalökonomie*, Léopold, Rostock, Germany [trans. Wartenberg C M 1966 *Von Thünen's Isolated State*. Pergamon, Oxford, UK]
Tiebout C 1956 A pure theory of local expenditures. *Journal of Political Economy* **64**: 416–24
van Ommeren J, Rietveld P, Nijkamp P 1996 Residence and workplace relocation: A bivariate duration model approach. *Geographical Analysis* **28**: 315–29
Weber A 1909 *Über den Standort der Industrien*, Tübingen, Germany [trans. Friedrich C F 1929 *Alfred Weber's Theory of Location of Industries*. University of Chicago Press, Chicago]
White M J 1999 Urban areas with decentralized employment: theory and empirical work. In: Cheshire P, Mills E S (eds.) *Handbook of Regional and Urban Economics: Applied Urban Economics*, Vol. III, pp. 1375–1412

C. Gorter and P. Nijkamp

Locational Conflict (NIMBY)

Locational conflict arises when individuals or groups believing themselves to be negatively affected express opposition to a locational or siting decision made by others. As an expression of protest or opposition within the public sphere, locational conflict represents a political manifestation of a geographical conflict over locational decision-making. Conflict may be initiated by the siting or locational designation of a facility, activity or land use believed by an affected party to be noxious, hazardous, undesirable, stigmatizing, or unwanted for any of myriad other reasons. Such targets of conflict have been dubbed locally unwanted land uses, or LULUs, by Popper (1985).

NIMBY, an acronym for Not In My Backyard, is a colloquial expression of locational conflict, articulating the sentiment that certain activities, facilities or land uses should be located elsewhere (not in my backyard) because of real or perceived negative effects of the sited activity on the surrounding area. NIMBYism, often referred to as the NIMBY syndrome, refers to the common or widespread opposition to change in one's surroundings associated with the introduction of an activity or land use thought to bring negative consequences. The term NIMBY is often used, by extension, to refer both to individuals or groups that habitually or expectedly oppose change in the local environment and to the unwanted land use or facility to be sited. In common usage, NIMBYism pejoratively connotes a reactionary parochialism based on self-interested, and thus possibly biased, perceptions. Locational conflict, in contrast, embraces a broader range of perspectives encompassing structural and systemic as well as perceptual causes and allowing for conflict to be liberating rather than simply reactionary in certain circumstances. This range of approaches, their conceptual origins, and means of conflict resolution are examined in this article, after describing the spatial and temporal dimensions of locational conflict.

1. *Spatial and Temporal Dimensions*

The relatively site-specific character of locational conflict differentiates it from relatively a-spatial forms of political expression such as social movements, protest movements, and even geopolitical territorial or jurisdictional conflicts. Individual sites of locational conflict, however, may seek coalitions with conflict participants in other locations to form aggregate social or protest movements covering wide geographic regions. Local advocacy groups in the USA and elsewhere, opposed to the siting of environmentally hazardous waste treatment facilities, for example, have become adept at forming broad-based coalitions seeking to alter regulatory provisions governing waste production, treatment, and disposal at regional and national scales.

Inherently geographic, locational conflict and NIMBYism are frequently manifested at the local scale of the immediate neighborhood or district in which deleterious effects of a proposed facility or activity are concentrated and experienced. If one conceptualizes the negative effects as an externality generated by the proposed facility (e.g., the smoke and particulates expelled through a factory's smokestack and falling on the surrounding neighborhood), then those effects can be mapped onto a spatial externality field defining the area of impact (Cox and Johnston 1982). Locational conflict is often relatively local in scale since such externality effects usually attenuate sharply with distance. Conflict also becomes manifest at larger spatial scales, often depending on geographical circumstances of adjacency or propinquity. A proposal to site a noxious facility near a municipal or regional border, for example, may generate locational conflict between the abutting jurisdictions. Proposals to site solid waste disposal facilities that will accept waste from multiple jurisdictions often generate opposition to the importation of waste from distant locations, expanding the conflict over a sizable area. Locational conflict has similarly emerged at the national scale over proposals to export hazardous waste from western industrialized nations to less-developed countries and over the migration of noxious effects, such as air or water pollution, across international boundaries. Concerns over the worldwide effects of climate change attributed to material practices in high-consumption nations can be understood as locational conflict at the global scale.

Locational conflict is not a uniquely recent phenomenon nor one exclusively associated with highly industrialized societies. Analysis of newspaper accounts of locational conflicts in nineteenth-century Worcester, Massachusetts (Meyer and Brown 1989) found that the frequency of reported conflicts on a per capita basis was essentially the same in Worcester in the 1870s as occurred in the 1970s in Vancouver, British Columbia (Ley and Mercer 1980), London, Ontario (Janelle 1977), and Columbus, Ohio (Cox and McCarthy 1980). The location of rendering plants, slaughter houses, and saloons generated protest in the nineteenth century, while waste incinerators, homeless shelters, and low-cost housing frequently motivated opposition in the late twentieth century. A vexing new problem for local officials is the location of cellular telephone towers, which are sometimes disguised inside church steeples or building cupolas to avoid protests over visual blight. In general, however, while the targets of conflict have changed to reflect changes in technology, economic activity, and society at large, the intensity of conflict was no less virulent a hundred years ago than it is today.

The perceptual basis of locational conflict exhibits both change and continuity over time. Newly industrializing cities of the nineteenth century provided a context in which proximity to the coal smoke and pollution emitted by mills and factories was perceived as beneficial, no doubt due to an association with steady employment and economic prosperity (Meyer and Brown 1989). In contrast, public awareness of the health hazards of industrial activity is one of the most frequent sources of locational conflict today. A common focus of protest in 1870s Worcester were the millponds which, despite their obvious economic importance, were thought to generate 'miasma,' believed to be a disease-bearing atmospheric poison produced by decaying organic matter (Meyer and Brown 1989). The parallel to today's conflicts over siting waste incinerators is striking. Both the millponds then and the incinerators now are considered necessary adjuncts to essential economic activity but both were and are also perceived to present risks to health and the environment. Both, consequently, are said to produce the NIMBY syndrome: yes, society needs the facility – but Not In My Backyard.

2. *Antecedents and Conceptual Approaches*

Formal analysis of locational conflict by American geographers began in the early 1970s as an extension of classical economic location theory. Classical theories of industrial and residential location presented a model of locational choice by autonomous units (firms or households) seeking to maximize individual utility functions within budget constraints. Political geographers such as Kevin Cox and his students at Ohio State University challenged the assumption of autonomous locational decision-making by recognizing that individual utilities are not independent of the externality effects produced by the locational decisions of others (Cox and McCarthy 1980). Conflict ensues when the utility of a locational choice is diminished by the negative externalities generated by the locational choices of others. This extension of traditional location theory shifted analysis away from individual locational choice towards a focus on the manipulation of the spatial distribution of externalities, that is, on attempts to exclude negative externalities and attract positive ones. While classical location theory addressed individual locational decision-making within a framework of economic rationality, locational conflict theory examines collective strategies for the protection and enhancement of neighborhood quality within a framework of collective political action. As a result, the focus of analysis moved from the economic to the political arena, from individual choice to collective action, and from locational decision-making to locational conflict as a politics of turf.

The view of locational conflict as turf politics, however, retains an axiomatic assumption of traditional location theory. The central assumption in this behavioral approach is that conflict arises when a locational decision perceived as beneficial by some is perceived negatively by others. While focus has shifted to collective action, the impetus for political engagement still resides in the individual's perception of positive or negative consequences associated with a proposed locational decision. The active agent is the autonomous individual whose behavior is guided by a unique calculus of perception, whether locational conflict is situated within an individual choice framework of conflicting utilities or in the pluralist politics of neighborhood change.

Situated at the level of conflicting perceptions, the behavioral approach to locational conflict rarely considers why changes occur in local areas such that conflict arises over their perceived positive or negative consequences. A structuralist approach views locational conflicts as symptomatic of fundamental contradictions situated within the basic structure of society. Locational conflicts, in this view, are merely surface manifestations of deep-seated conflicts inherent in the extant system of social organization (Cox and McCarthy 1980). Within a structural mode of explanation, negative externalities are not viewed simply as evidence of market imperfections subject to correction but rather are understood as necessary and inevitable consequences of class relations within the process of capital accumulation.

Identifying the particular structural contradictions that become manifest as locational conflict, however, is itself a subject of conflict and disagreement. In one view, the inherent contradiction between labor and capital generates contradictory orientations to neighborhoods, such that labor's attachment to place based on the neighborhood's use value conflicts with capi-

tal's investment interest based on the neighborhood's exchange value. Locational conflict, in this framework, arises from the fundamental antagonism between labor and capital with respect to neighborhood change. In another view, locational conflict is situated within the incessant competition between factions of capital, and results when the expansionary interests of mobile capital conflict with the exclusionary interests of capital already fixed in place (Plotkin 1987). For example, existing investors seeking to exclude potential new competitors may resist a shopping center developer's search for new investment sites. Bridging these two approaches is a third view in which labor's attachment to neighborhood erects a barrier to continued accumulation of capital. In response, capital encourages long-running societal processes—the homogenization of space, pervasive ideologies of materialism and consumerism, media manipulation and mass education, and increasing residential mobility, among other contributing factors—which succeed in transforming neighborhoods from communities into commodities and, consequently, transforming labor's attachment to community into an orientation based on protecting neighborhood exchange value (Cox 1981). Now locational conflict occurs when the exchange value interests of capital and labor fail to coincide as, for example, when a land developer's proposal for a shopping center generates traffic and pollution that threaten to reduce the resale value of surrounding homes.

More recently, a post-structuralist approach has emerged that situates locational conflict within the antagonistic discursive or representational strategies of contending groups vying for control over the use of space, where such control is an expression of political power. Rejecting the view of space as simply a container for action, this approach, influenced by the work of Lefebvre (1991) and others, considers the social and political process through which the meaning and use of space are constructed in particular instances. Now locational conflict is not primarily about the spatial distribution of activities or land uses nor is it simply about contrasting preferences for various activities in particular locations. Rather, locational conflict is symbolic conflict over the social distribution of power to assign meaning and uses to space (Mitchell 1992). Locational conflict symbolizes contention over whose values have standing within the political process and whose values, therefore, become expressed in the landscape. Recursively, control over the use of space symbolically constitutes the controlling group as a legitimate actor within the broader political process. Establishment of a squatter settlement within an affluent residential neighborhood despite the opposition of existing residents, for example, provides a site for housing impoverished families. Also, and perhaps more importantly, it discursively represents squatter families as legitimate occupants of space and, therefore, legitimate participants in the political process.

The evolution of locational conflict theory from classical to behavioral, to structural, and post-structural formulations parallels theoretical developments in the field of geography as a whole. Underlying each of these theoretical positions is a set of often incommensurate assumptions regarding the relative autonomy of individual actors, the primacy of market processes, the nature of the public interest, and the respective roles of structure and agency in creating geographic landscapes. In addition, the various theoretical understandings of locational conflict point to substantially different routes for its amelioration or resolution. A behavioral model of conflict based on contrasting perceptions suggests that resolution is attainable through public information and education designed to bring perceptions into agreement. A model of conflict based on the unequal spatial distribution of negative externalities suggests the use of compensation to equalize outcomes. Understanding conflict as symptomatic of deep underlying structural relations requires the elimination of structural contradictions. A focus on locational conflict as the symbolic expression of rights to create the meaning of space requires the deconstruction of opposing claims and their reconstruction in more equitable forms. Far from being merely an academic exercise, one's theoretical approach to locational conflict implicates significantly different avenues for its resolution, as discussed in the following section.

3. *Resolving Locational Conflict*

The contentious character of locational conflict has generated considerable debate over the intervention strategies conducive to its resolution. Approaches to intervention differ according to their proponents' understanding of the sources and dynamics of conflict, as described above, and according to their focus on the varying perspectives of conflict participants. Methods of resolving conflict distinguish between community-centered, user-centered, and state-centered approaches (Takahashi 1998 and DeVerteuil 2000).

3.1 *Community-centered Approaches*

Community-centered approaches focus on community opposition to a proposed facility siting. Depending on the characteristics of the community and, in part, on the perspective of the analyst, local opposition can be interpreted as exclusionary parochialism or as constitutionally protected freedom of speech and dissent (Takahashi 1998). Those taking the latter perspective point to instances in which poor or politically marginalized communities succeed in preventing the siting of an environmentally hazardous or otherwise noxious facility and consider that conflict has been resolved through defeat of the siting proposal (Heiman 1990).

More frequently, and especially when phrased in the NIMBY characterization, locational conflict is viewed as a shortsighted and parochial expression of self-interest that prevents achievement of the greater public good. Their proponents present the facilities whose siting generates locational conflict as both necessary and beneficial for society. A mobile society needs highways and airports. Consumers expect convenient access to shopping. Production of wastes is an inevitable by-product of a consumer society and requires safe and reliable disposal capacity. New prison construction keeps dangerous criminals off the streets and prevents prison overcrowding.

From this vantage point, conflict over siting proposals poses a barrier to attainment of the public good and requires a strategy to overcome, elude, or defuse opposition in order to allow the siting to go forward. Some professional associations of land and real estate developers offer guidelines and training to assist their members to negotiate the hazards of community opposition to proposed developments. Such strategies counsel facility proponents to anticipate opposition, solicit statements of support from potential allies, marshal expert testimony for public presentation, demonstrate the anticipated benefits of the proposed scheme, utilize formal appeals procedures, and the like.

Compensation of affected parties is a central premise within the community-centered approach. If locational conflict arises from the unequal spatial distribution of negative externalities created by the sited facility, then conflict can be resolved through compensation of those bearing a disproportionate share of costs. The theory of compensation assumes that costs are susceptible to translation into a single monetary metric and that a fair system can be devised for identifying affected parties and distributing payment. Such a system is difficult to devise in practice, however. It is difficult to distinguish between payment as compensation of costs and as inducement to accept costs, and communities often reject compensation as synonymous with a bribe. It is relatively easy to compensate for direct costs associated with a newly sited facility, such as additional emergency response equipment, but it is more difficult to assess a value for indirect costs such as stigmatization or economic growth foregone when potential investors avoid proximity to a noxious facility. There is disagreement as to whether costs should be compensated at an equal level in an impoverished and an affluent community. On purely economic grounds, the impoverished community can receive less compensation because poor people are satisfied at a relatively lower cost but this conclusion can easily be challenged on ethical grounds. Compensation as a means of resolving locational conflict is problematic for these and other reasons.

Some practitioners adopting the community-centered approach choose siting strategies based on equity considerations. Such strategies typically distinguish between the goals of procedural equity in decision-making and distributional equity in outcomes. Approaches based on distributive justice assume the necessity of the facilities in question and seek to achieve a fair distribution, where the primary criterion of fairness is to avoid neighborhood concentrations of unwanted facilities. One such method is to devise a point system for desirable and undesirable facilities (parks and libraries in the former category, for example, and waste incinerators and sewage treatment plants in the latter) and to equalize points across neighborhoods or political jurisdictions. Localities might be encouraged to opt for an undesirable facility so as to qualify for a desirable one under this system or might be disqualified from obtaining a beneficial facility due to insufficient points for hosting undesirable ones. The obvious difficulty in assigning points to facility types, however, has prevented this system from being implemented in practice. New York city's legislatively mandated fair share process for distributing unwanted city facilities across neighborhoods similarly has been abandoned in practice due to difficulties of implementation.

Community-centered approaches based on procedural justice argue that public opposition is mobilized not only by unwanted siting outcomes but also by perceived unfairness in the decision-making process. At a minimum, the procedural solution takes the form of a mandated hearing within a public review or permitting process. The public hearing has been criticized as a means of facilitating public participation, however, on the grounds that it is rarely more than advisory and that important decisions, such as the need for the facility, have usually been taken prior to the hearing. More detailed procedural approaches stress the importance of community consultation early in the siting process, inclusion of all stakeholders in negotiations, and the opportunity for sites to withdraw from consideration at any time. The most fully developed approach within this framework allows localities to volunteer themselves as sites for controversial facilities, subject to specified siting criteria and in return for generous compensation. This approach elides difficult ethical questions, however, if localities volunteer for controversial facilities not through choice but due to a lack of alternative means of economic survival.

3.2 User-centered Approaches

The user-centered approach usually applies to conflict over social service facility location. This approach moves beyond the perspectives of siting proponents and opponents and considers the perspective of facility users who often experience extreme poverty and/or physical, social, or mental disability as a barrier to access to needed services (Takahashi 1998). Omission of the users' perspective in community-centered

approaches both reflects and contributes to their stigmatization. While equity in community-centered approaches focuses on community participation in facility siting decisions, equity in the user-centered approach is concerned with facilitating access to needed facilities on the part of service-dependent clients. Residents and users of homeless shelters, drug or alcohol rehabilitation centers, domestic violence shelters, group homes, and similar facilities are often dependent on multiple services and benefit from service clustering that other community residents view as an undue concentration of unwanted facilities.

Social service providers choose from two divergent strategies in countering local opposition (Dear 1992). The collaborative approach seeks to establish a partnership between the service provider and the host community through public outreach and education aimed at improving awareness, tolerance, and acceptance. The autonomous approach considers access to needed services to be a civil right of facility users and relies on legislative and judicial protections to override community opposition.

3.3 State-centered Approaches

The state-centered approach to locational conflict reformulates the traditional question of locational conflict—'Why is the community opposed to this facility?'—and asks instead: 'Why is the state seeking to site this facility in this community?' (Lake and Disch 1992). This rephrased question problematizes the assumption, shared by community- and user-centered approaches, that facilities are needed by society. In the state-centered approach, controversial facilities are understood as needed by capital seeking to externalize costs as a competitive strategy and as an expedient solution for the state seeking to facilitate capital accumulation while maintaining legitimation of the capital-labor relationship (Lake 1993).

Facility siting in this context constitutes a particular problem-solving strategy that is instrumental for the state. By providing a locational solution to a problem of industrial production (waste production or homelessness, for example), facility siting allows production and capital accumulation to continue relatively unimpeded while concentrating costs on host communities. The decision to concentrate costs on communities rather than on capital reflects a political calculation that it is preferable for the state to confront localized political conflict than to risk a challenge to the capital-state relation. The siting strategy deflects political conflict away from a potentially daunting challenge to the capital-state relation and into a relatively benign debate over the merits of alternative facility locations (Lake and Disch 1992).

According to the state-centered approach, the root of locational conflict is situated not in the consequences of hosting a controversial facility but rather in a regulatory approach that inexorably leads to facility siting as a policy solution that concentrates costs on host communities. To the extent that this is the case, the resolution of conflict requires a restructuring of policy assumptions and policy solutions through political debate, redirecting costs from host communities back on to capital, and avoiding the emergence of conflict at its source rather than seeking to ameliorate its occurrence after the fact.

4. Future Directions: Ethical and Social Justice Issues

Recent locational conflict scholarship has increasingly intersected with the literature of environmental racism and environmental justice. This coalescence emerges from a common conceptual focus on equity and on the distinction between distributive and procedural equity, in particular. Influenced in part by development of the state-centered approach, locational conflict scholarship is increasingly shifting attention from siting outcomes to examine political and structural influences on the siting process. This shift closely parallels a redirection within environmental justice scholarship from a focus on distributional equity, concerned with the concentration of environmental burdens in disempowered communities, to a focus on procedural equity, concerned with democratic participation in the gamut of prior decisions affecting the production of burdens and benefits to be distributed (Lake 1996). Recent environmental justice literature explicitly seeks to relinquish its narrow focus on inequity in facility siting, emphasizing instead the pervasive inequity in the broad socio-spatial processes creating geographic landscapes (Pulido 2000). The challenge for these increasingly combined spheres of scholarship is to expand understanding of these processes and to inform the design of institutional structures that expand community participation in their operation.

See also: Behavioral Geography; Citizen Participation; Environmental Justice; Global Environmental Change: Human Dimensions; Location Theory

Bibliography

Cox K 1981 Capitalism and conflict around the communal living space. In: Dear M, Scott A (eds.) *Urbanization and Urban Planning in Capitalist Society*, Methuen, New York

Cox K, Johnston R 1982 Conflict, politics and the urban scene: a conceptual framework. In: Cox K, Johnston R (eds.) *Conflict, Politics and the Urban Scene*. St. Martin's Press, New York

Cox K, McCarthy J 1980 Neighborhood activism in the American city: behavioral relationships and evaluation. *Urban Geography* **1**: 22–38
Dear M 1992 Understanding and overcoming the NIMBY syndrome. *Journal of the American Planning Association* **58**: 288–300
DeVerteuil G 2000 Reconsidering the legacy of urban public facility location theory in human geography. *Progress in Human Geography* **24**: 47–96
Heiman M 1990 From 'Not in My Backyard' to 'Not in Anybody's Backyard!' Grassroots challenge to hazardous waste facility siting. *Journal of the American Planning Association* **56**: 359–62
Janelle D 1977 Structural dimensions in the geography of locational conflict. *Canadian Geographer* **21**: 311–28
Lake R 1993 Rethinking NIMBY. *Journal of the American Planning Association* **59**: 87–93
Lake R 1996 Volunteers, NIMBYs and environmental justice: dilemmas of democratic practice. *Antipode* **28**: 160–74
Lake R, Disch L 1992 Structural constraints and pluralist contradictions in hazardous waste regulation. *Environment and Planning, A* **24**: 663–81
Lefebvre H 1991 *The Production of Space*. Basil Blackwell, Oxford, UK
Ley D, Mercer J 1980 Locational conflict and the politics of consumption. *Economic Geography* **56**: 89–109
Meyer W, Brown M 1989 Locational conflict in a nineteenth-century city. *Political Geography Quarterly* **8**: 107–22
Mitchell D 1992 Iconography and locational conflict from the underside: Free speech, People's Park, and the politics of homelessness in Berkeley, California. *Political Geography* **11**: 152–69
Plotkin S 1987 Property, policy and politics: towards a theory of urban land-use conflict. *International Journal of Urban and Regional Research* **11**: 382–403
Popper F 1985 The environmentalist and the LULU. Environment **27** 7–11, 37–40
Pulido L 2000 Rethinking environmental racism: White privilege and urban development in Southern California. *Annals of the Association of American Geographers* **90**: 12–40
Takahashi L 1998 *Homelessness, AIDS, and Stigmatization: The NIMBY Syndrome in the United States at the End of the Twentieth Century*. Oxford University Press, London

R. W. Lake

Locke, John (1632–1704)

1. *Life and Writings*

Hegel perhaps excepted, John Locke is the most eminent philosopher of the Western world to have addressed himself to political and social issues. Indeed his standing as *the* great thinker amongst English speakers in the eighteenth and nineteenth centuries gave to his writings on such issues, the theories of politics, toleration, economics and education, a standing which they would not have commanded even when taken together. And yet the famous *Essay Concerning Human Understanding* (1975) scarcely provided a general intellectual framework for his theories on individual topics. The action of his mind was like that of the beam of a searchlight shining through the darkness on individual objects but with no linkage made between them or any reflection back from object to source.

John Locke the man can be taken as the archetype of the English country gentleman turned intellectual. He was the first born of the two sons of a lawyer who fought for the Parliament against King Charles I when John was a teenager. The family had a small landed estate in the county of Somerset and Locke was born on August 29 1632 in the village of Wrington, though the family seat was at Belluton some ten miles away. He inherited the family properties and his father's modest position within the community of gentry of the county. He was an absentee landlord, except for brief intervals, for most of his life. This was because his father designed him for a scholarly career, not all that unusual for the gentry but exceptional for an heir (Laslett 1948).

Locke's father set his sights high for his son, whose outstanding intelligence must have been evident from the beginning. In 1647 he sent him to Westminster School, the best in the country, and thereafter, in 1652, to the college of Christ Church in Oxford, where he went on to take office as a teacher between 1658 and 1684. Locke never married, and insisted that he would have liked to have lived his whole life there as a bachelor don. However, it proved otherwise, and within a decade he found himself for much of his time in London at the center of national politics, in the entourage of one of the great political magnates of the time, Anthony Ashley Cooper, Earl of Shaftesbury, who was then in high office but subsequently became leader of the movement of opposition to the succession to the throne of the Catholic James, brother of the childless Charles II, the Exclusion campaign, as it is called. The explanation for this transmogrification, which left his position at Christ Church unaffected, was that Locke had taken up medicine so that he could continue his academic career without becoming a priest. In 1666, on a visit to Oxford, the invalided Shaftesbury met Locke who performed upon his body one of the medical miracles of that age which saved his life. It was as a medical adviser of Shaftesbury, not as an academic, that Locke took up and pursued his career as a writer on the theory of knowledge, and on what we now recognize as social scientific subjects. *Two Treatises of Government*, also published in 1690, was composed while Locke acted as what a close friend called 'assistant pen' to Shaftesbury in the early 1680s during the Exclusion campaign. The Catholic James nevertheless became James II in 1685.

A now outmoded interpretation of *Two Treatises* perceives it as the subsequent rationalization of the so-called Glorious Revolution of 1688/9; it is now accepted that the author's intention was to justify a revolution yet to be brought about by Shaftesbury and

his party, but which did not take place. This makes it easier to understand its usefulness as a persuasive in favour of national revolutions for two hundred years and more after its publication. Without Shaftesbury, it has been said, Locke would not have been Locke at all. Indeed he had composed numbers of unpublished treatises before 1666, mainly of a theological type, though with a political admixture conservative in tone. But the younger Locke cannot be described as wholly backward looking, if only because of his association with Robert Boyle and others in the earliest advances in natural science in England, and in the establishment of the Royal Society, of which he became a fellow in 1668.

As his convictions changed in the liberal direction, Locke suffered for them, both at the hands of his Christ Church colleagues and of state authority. Getting further and further absorbed into Shaftesbury's affairs, taking a job in government, exercising responsibility for various commercial and colonizing undertakings, he undertook the writing out in 1668 of a political constitution of the colony that his master owned on the Atlantic seaboard of North America, the *Fundamental Constitutions of Carolina* (Locke 1794). This was the first written political constitution ever composed, and provided near-absolute toleration for nonconforming belief. Things like this, and failure to conform himself, were unacceptable to the regime and in 1684 Locke was expelled from Christ Church by Royal Order. Shaftesbury had died in exile in Holland, and the beleaguered Locke finally took up residence in that country, the haven of political progressives. He was there in 1689 when William and Mary sailed to England to bring about the Whig revolution. Locke returned with the joint monarchs, and was given minor political offices of a commercial kind. In the fifteen years left for him to live he twisted his fingers round the haft of English intellectual life and got it in so firm a grasp that it began to point in the direction of his choice. This is the period in which he published his books and addressed himself to posterity.

He spent this final, intellectually victorious, period at his London address, but also living as much as he could as a country gentleman once more, not on his own estate but as a lodger in the manor house of Oates in the country of Essex, much nearer his London activities. These activities made him into a founder of the British Empire, especially in relation to the North American colonies which were later to call on his doctrines when they came together as the United States (Laslett 1969). An intimate association grew up with the lady at the house at Oates, Dame Damaris Masham, a devout Christian and a theologian of the Lockeian description, along with her son and stepdaughter. When he died at the house on October 28 1704, he was buried in the churchyard of the nearest village, at High Laver. He should be remembered in this final dual role, as the sovereign of London intellectual society, and as the revered elderly gentleman seated in the garden at Oates—he himself was a passionate gardener—surrounded as often as possible by young people and children.

2. *Publications*

Included here are books he published on the social sciences which have affected his vast readership for three hundred years rather than the treatises left in manuscript, which have begun to influence scholars in recent years.

2.1 *Philosophy*

In *An Essay Concerning Human Understanding*, first published in 1690, Locke (1975) established himself as the first and most important empiricist thinker. The exclusive source of knowledge was experience gained from sensation. Apart from revelation, nothing else was given to the intelligence, and he was particularly insistent that no knowledge was innate. Everything to be known was printed on the blank sheet of the mind through its ideas, simple ideas elaborated systematically into a nexus of complex ideas. Locke went beyond his empirical assumptions and used rational arguments to underwrite his firm belief in a real world behind the ideas which he systematized, and in the existence of God. It is the inconsistency between his rationalism and empiricism which has been seen as his weakness, though by some as also a source of relevance and insight.

2.2 *Political Theory*

Two Treatises of Government (1960), lays it down that the sole source of political power is the consent of the citizens of the society concerned. The target throughout both treatises was the patriarchal theory of Sir Robert Filmer (died 1653) whose posthumous work *Patriarcha* (1680) provided theoretical justification for Shaftesbury's opponents. Filmer and his devotees, the Tories in the nascent English party system, believed that the source of all authority and property, political, social, and personal was fatherhood, on which God had conferred absolute divine right. Extreme, even ridiculous as it may seem, this had evident appeal to landowning gentry families, and it had strong scriptural and ecclesiastical authority behind it. Moreover Filmer displayed a needling effectiveness by exposing all alternative justifications of power and property as even more unrealistic and ridiculous.

This laid Locke's way open to fundamentals. Whatever may have been the patriarchal underpinnings of social and political relationships, governance could

only be exercised by those individuals specifically entrusted with it and elected representatively by citizens, mainly the propertied. Though property could be inherited, its basis was labor: you owned those natural objects on which you had done work, and this property was protected by the state set up in this way as its most important function. Government was in a sense divine, as an emanation of power, but no regime could conceivably rule by divine right, and once it lost the consent of the governed it could be, had to be, overthrown by the property-owning citizenry and replaced by one of their choosing. Both the pre-eminence of property in political society, and the intellectual origins of democratic revolution as a perpetual sanction on all political authority, are plain to see.

2.3 Toleration

Epistola de Tolerantia (1968). Locke flatly denied that any authority had a right to prescribe or proscribe religious belief, or to punish anyone for holding a religious opinion. All such matters were reserved to the Church, defined as 'a voluntary society of men.' The discipline within the Church was limited to exhortation with no other sanction than expulsion. Nevertheless, opinions contrary to civil peace had to be controlled, and atheists also, because they could not be bound by promises. A muted resistance to governmental intolerance was contemplated by Locke, in spite of what he says about civil peace.

2.4 Economics

Some Considerations of the Consequences of the Lowering Interest and Raising the Value of Money—Further Considerations (1991). Locke was the last and most influential mercantilist writer, preoccupied like all the others with the value of money, which was considered to be the spring, instrument, and channel of all economic activity. Money had a price which was determined to a considerable extent by the rate of interest. Maintaining the constancy of the value of money by regulating the coinage and manipulating interest rates was the object of this analysis and of the policy which he recommended. Particularly important were the regular supplements to the supply of money through foreign trade, a crucial activity in the Lockeian system.

2.5 Education

Some Thoughts Concerning Education (Locke 1988). The familiar phrase 'A sound mind in a sound body' are the opening words and the *leitmotiv* of this extraordinarily popular and influential treatise. A great deal of it is given up to habits conducive to health, especially exercise and diet. Strict discipline is needed but not through corporal punishment: the proper instrument is encouragement and strictly consistent practice. The mode of instruction should be as conversational as possible. In short, children should be treated not so much as little adults, but as adults in the making, learning by their own intellectual assent as their capacities developed.

3. Influence and Importance

The huge influence and great historical importance of Locke's political theory should be evident from the account already given. *Two Treatises of Government* was repeatedly printed as the eighteenth century went by; it was translated into French, the international language, in 1691, and was published more often in French than in English; editions also appeared in most of the other European languages (see Locke 1960).

As has been said, it repeatedly came to the fore in the history of nations when issues of despotic government arose, sometimes along with the *Epistola de Tolerantia*. The next most printed and cited work was that on *Education*, though *An Essay Concerning Human Understanding* ran it close. His economic theory had widespread influence up to the time of Adam Smith and has been regarded as the quintessential early expression of the capitalist spirit (MacPherson 1962). His educational doctrine has been continuously influential for at least two centuries and can be regarded as the origin both of companionate upbringing in the home, parents and children on a level, and of the teaching practice of Anglo-Saxon elite schools. In short, the social scientific works discussed here can be claimed to have ushered in empirical–rational social scientific enquiry in their respective areas. The *Essay Concerning Human Understanding* has been aptly called the first modern book, a title which would also sit well on all the treatises we have commented on. There is surely no need to go further to demonstrate Locke's outstanding influence and importance.

See also: Democracy; Democracy: Normative Theory; Democratic Theory; Economics, History of; Educational Philosophy: Historical Perspectives; Empiricism, History of; Political Economy, History of; Tolerance

Bibliography

Cranston M 1957 *John Locke, a Biography*. Longmars, London
De Beer E S (ed.) 1976/83 *The Correspondence of John Locke*. University Press, Oxford, UK

Kelly P H (ed.) 1991 *Locke on Money*. Oxford University Press, Oxford, UK
Laslett P 1948 The gentry of Kent in 1640. *Cambridge Historical Journal* **9**(2):
Laslett P 1969 John Locke, the great recoinage and the Board of Trade. In: J W Yolton (ed.) *John Locke, Problems and Perspectives*. Cambridge University Press, Cambridge, UK
Locke J 1794 *Collected Works*. Vol. 9, pp. 175–99
Locke J 1960 *Two Treatises of Government*. Laslett P (ed.) Cambridge University Press, Cambridge, UK
Locke J 1968 *Epistola de Tolerantia*. Oxford University Press, Oxford, UK
Locke J 1975 *An Essay Concerning Human Understanding*. Oxford University Press, Oxford, UK
Locke J 1988 *Some Thoughts Concerning Education*. Oxford University Press, Oxford, UK
Locke J 1991 *Some Considerations of the Consequences of the Lowering Interest and Raising the Value of Money … 1962 with Further Considerations* … . Kelly P (ed.) Oxford University Press, Oxford, UK
MacPherson C B 1962 *The Political Theory of Possessive Individualism: Hobbes to Locke*. Oxford University Press, Oxford, UK
Rand B 1925 *The Correspondence of John Locke and Edward Clark*. Oxford University Press, Oxford, UK

P. Laslett

Logic and Cognition, Psychology of

All about us, creatures seem more suited to their niches than we humans are: they are more fleet of foot and wing with more finely tuned sensorium. Yet, only humans are able to transcend restricted habitats, develop plans, adapt strategies, and remember not only who said what, but why. In sum, our greatest adaptation across the vicissitudes of time and place is our ability to reason. It has been said that *reasoning* is humans' central intellectual activity (Rips, 1986) that organizes language, culture, and our memories. It allows us to take events, assertions, suppositions, beliefs as input and draw conclusions and create new information. Reasoning allows us to go beyond the given in a situation and see the overall pattern; to find the theme; to see the plan or create one.

1. *Reasoning*

The pervasiveness of reasoning has captured the interests of cognitive psychologists who seek to understand universal principles of human cognition. Early in the twentieth century, psychologists were surprised at how poor people are in solving academic logic problems, yet how perfectly adequate the same people are in their reasoning with everyday materials—and without special training (Wilkins 1928). People tend to feel alienated from the prescriptive logic of the classroom, as well they might since from its very beginnings it had little to do with the rationality of the human mind and very much to do with creating a didactic device for educating students. Aristotle, in an effort to express to his students how to think clearly and how to draw conclusions from data, created a set of idealized conditions (called Figures) from which certain conclusions could be unequivocally drawn; and he contrasted these with conditions from which a conclusion could only be equivocally drawn. From these simple didactic beginnings, the field of philosophy of logic was born. Over the centuries, innovations in understanding human logical or rational abilities have continued to come from efforts at teaching: Euler, who invented the circles to represent sets, did so within the context of tutoring in logic (Euler 1775). Venn (1880), who expanded on the circle representation, introduced the notion of a universe of discourse and analyzed the overlapping of multiple sets/circles and did so within the context of trying to teach logic to his students at Cambridge University. Some of these didactic techniques have inspired modern cognitive psychologists to theorize about natural human inference. Erickson (1978) and Guyote and Sternberg (1981) borrowed Venn diagrams as an approximation to how students may interpret and mentally represent quantified sentences in their combinatorial theories of reasoning. One of the major cognitive accounts of the processes of human reasoning is called 'Mental Models'; it was developed by Johnson-Laird (1983) and had its origins in the efforts at the beginning of the twentieth century by Störring (1908) to teach his students how to reason by drawing pictures. Throughout the centuries, the impoverishment of didactic logic has been so daunting to students that they have argued they should not have to study it. After all, they say that 'we should only study things that are relevant to human life; syllogistic reasoning is not relevant, so we shouldn't be studying syllogisms.' This itself is a type of logical argument that could be rewritten in syllogistic form (the standard form for logic):

We should only study things that are relevant.
Syllogistic reasoning is not relevant.
Therefore: we should not study syllogistic reasoning.

Clearly this argument is self-refuting, but illustrates that even though we may wish to avoid the intricacies of prescriptive logic, some aspects may be descriptive of an underlying system of natural inference processing that becomes manifest even by people untutored in the philosophical tradition. Take a four-year-old, for example. We know that if you tell the child (in the form of a syllogism):

If you brush your teeth, I'll read you a story before bed and
The child brushes her teeth
You can be sure that the child will expect a story.

The structure of this argument is formally described as *If P, then Q*; *P therefore Q* and is termed *modus ponens* (positive mode) and is an argument that is accurately understood by children and adults, without any special instructions (e.g., Taplin et al. 1974). Just as Molière's *Bourgeois Gentilhomme* was surprised that he spoke prose all of his life, so too are the students of the psychology of logic surprised to find that people seem to be inherently rational (at least under specifiable circumstances) and follow a kind of *psycho-logic* which is illustrated below with three different forms of inference: linear, categorical, and conditional reasoning.

1.1 Linear Reasoning

One of the simplest forms of inference compares the extent of things and draws a conclusion about the item that has the most or least of the extent. For example:

Scott is taller than Paul.
Paul is taller than Nancy.
Who's tallest/shortest?

These kinds of syllogisms found their way on to the early measures of intelligence (e.g., Burt 1919) because reasoning skill with these problems seemed to increase with chronological age during adolescence (e.g., Hunter 1957). Interestingly, the pattern of difficulty shown on series problems by children is similar to that shown by adults (Huttenlocher 1968, Clark 1969) which suggests that linguistic processes (and the complexity of the terms: *older–younger*, *taller–shorter*, etc.) rather than inference *per se* contributes to the development of reasoning skill with age (Johnson-Laird and Steedman 1978).

These problems are called 'linear syllogisms' because one way to draw inferences from them is to imagine that the information is placed on a line or on an analogue scale (a continuously varying scale). When such a scale is used, it is possible for the reasoner simply to 'read-off' the mental image to find the answer. Consistent with this idea is students' expectation that the information will be presented in a linear form (left-to-right or right-to-left order), called the *canonical form*. When the information flow fits the canonical form, reasoning is fast and accurate and when it does not fit it, reasoning is labored and error prone (Hunter 1957). A similar effect has been observed for problems with more than just three terms in it. For example, Potts (1974) had students reason with relations among five items embedded in a passage (e.g., *A is greater than B, which is greater than C, which is greater than D, which is greater than E*) and discovered that terms that are further apart on the linear scale are easier for students to evaluate (e.g., *Is A greater than E?* was faster and more accurately judged than *Is A greater than B?*, even though the latter was actually presented in the narrative and the former inference was not). This effect of the way you mentally represent information has been termed the *semantic distance effect* (Holyoak and Mah 1981) and requires primarily that the information be coded on a uniform scale (such as size, intelligence, etc.).

1.2 Categorical Reasoning

Perhaps the simplest form of reasoning with categories is the categorical syllogism:

All men are mortal.
Socrates is a man.
Therefore: Socrates is mortal.

The ability to draw the appropriate inference is nearly universal, does not require knowledge of logical forms, or even to know the meaning of the term 'men' and 'mortal.' It requires only a basic willingness to cooperate to do the task. For example, Cole and Scribner (1974) showed that unschooled adults in Liberia will occasionally decline to draw an inference, but their explanation reveals basic inferential skills:

All men in Kpelle town are farmers.
Joe is a man in Kpelle town.
Therefore: Joe is a farmer.

'I cannot answer the question, because I only speak about people I know and I don't know this Joe.' In the canonical form, the man's explanation looks as follows:

I only speak of people I know.
I don't know Joe.
Therefore: I cannot speak of Joe.

So the justification for failing to do syllogistic reasoning takes the form of a perfectly phrased logical argument and is an example of the universality of this form of reasoning.

The critical factor in the willingness to solve syllogisms seems to be whether a person has had more than 2 years of basic schooling. When asked to solve categorical syllogisms, such educated people, while not necessarily tutored in logic, show the following profile. When they reason about everyday sorts of things, they are more accurate than when they reason about symbolic or abstract relations (Wilkins 1928). Reasoning on symbolic material can be enhanced by having a mental image of how many elements are contained in the sets—the sizes of the sets allows people to represent the relations more like concrete than like abstract materials (Revlin and Leirer 1978). When their beliefs about the world are consistent with the syllogism's conclusion, they are more accurate than when logic and belief are in conflict. When there is a definite unambiguous conclusion to be drawn, people are confident in the conclusion, but when no such conclusion is to be drawn, people are tentative and somewhat capricious in their answers. When reasoners are asked to draw inferences from a multitude of category relations as in the following narrative, they show a kind of semantic distance:

> All members of the Fundala Party are members of the Outcast Coalition. All members of the Outcast Coalition are members of the Hill People Union … members of the Peace Front.

The more distant the relation, the more accurate the judgment—just as with linear relations. A completely different pattern is seen when the categories do not readily fit in a common scale, as in the following (from Griggs 1978):

> All Fundala are Outcasts. All Outcasts are hill people. All hill people are farmers. All farmers are peace loving.

Reasoners view these relations as *divergent*, dealing in succession with a social group, geographic entities, occupational categories, and a political orientation. The very same people who performed optimally with the first passage are at near chance levels when asked to reason about this second narrative (Nguyen and Revlin 1993).

Reasoning with categories shows a developmental sequence somewhat akin to that shown for linear relations. The critical factor seems to be in the interpretation of the terms rather than possession of inferential processes. For example, when seven-year-olds are told that *all of the squares in a display are yellow*, they anticipate that the display will only contain yellow squares and no other kind of object. This is called an 'order-neutral interpretation.' Slightly older children interpret the quantified sentence as implying that all squares are yellow and that all of the yellow things will be squares. However, they are not surprised that there are nonsquares and these will have colors other than yellow. This is called a 'converted interpretation.' Finally, most adults (but not all of them) interpret such quantified sentences as true set-inclusion, that all squares are yellow, but some yellow things may not be squares (Bucci 1978). This ontogeny in the interpretation of quantifiers corresponds to the development of logical competence with categorical relations and suggests that the interpretative processing and not inferential ones may be what is developing.

1.3 Conditional Reasoning

Conditional syllogisms have as their central statement, a conditional expression of the form *If P is true*, *then Q is true*. Although conditional syllogisms are logically equivalent to categorical syllogisms, they have a communicative property that makes them especially interesting to psychologists in studying how people reason with them. The external marker *If* allows for an incredible range for conjecture: the reasoner can interpret it to mean *let's suppose that* … , or *someone believes that* … (e.g., *Paul believes that if P*, *then Q*, even though he may be wrong), or *to perform some action*, *there is a set of conditions that must be followed* (e.g., *If you're drinking alcohol*, *you must be the right age*), and so forth. That is, conditional syllogisms reveal the very property that makes human inference so dynamic: an ability to consider truth at many levels. When adults are presented with conditional syllogisms that have an unambiguous conclusion to be drawn, they are rapid and accurate:

If *P* then *Q*	If *P* then *Q*.
P is true	not *Q* is true.
Therefore: *Q*	Therefore not P.

These problems are called *modus ponens* and *modus tolens*, respectively. The former tends to be more accurately solved than the latter, but people demonstrate real skill in solving these problems without any special training. In cases where the conclusion is ambiguous, students take longer to reach a decision and tend to equivocate and generally draw an incorrect conclusion:

If P then Q	If P then Q.
Not P	Q.
Therefore: ?	Therefore: ?

Just as with other forms of inference, interpretation of the relational terms is critical to accurate judgment. A common error among adults is to assume that a kind of unique causal structure holds when they hear *If P then Q*. That *P* somehow causes *Q* and that it uniquely does this. This typically goes by the name *biconditional* interpretation: *If P is true then Q is true* and *if Q is true*, *so must P* (Taplin 1971). You will notice that this is similar to the 'converted interpretation' on categorical syllogisms. It has been argued that there is a developmental sequence in the ability to solve conditional syllogisms which is linked to the child's interpretation of the conditionals. Very young children treat conditionals as conjunctions, slightly older children treat them as biconditionals, and adolescents and adults are able to treat them as one-way conditionals. Just as with categorical reasoning, conditional inference is sensitive to the content of the material reasoned about. This is particularly striking when people have to test conditional rules (Cox and Griggs 1983). When the rules and situations match a *type* or *schema* (e.g., a permission rule: *If you meet the requirements*, *you can perform an action*), reasoning can be sharply more accurate than if the conditional rule is arbitrary (Cheng and Holyoak 1989, Cosmides and Tooby 1997). The ease and ubiquity of this kind of reasoning suggests it may actually be a part of the human cognitive apparatus, part of our biological heritage that permits us to engage in effective social interactions.

2. Where are the Rules?

These findings from cognitive research in three areas of logical reasoning show that people are far more

rational than is often assumed. Our inability to solve some impoverished symbolic problems does not mean that we are not rational. It does mean that if there is a logic within us—and we must be following some set of rules or procedures to allow us to be rational under normal circumstances—it is probably not the standard logic we are taught in schools. The logic of the head need not be the same as the philosopher's logic because human inference occurs in rich and varied contexts, with cognitive and social constraints operating all around us and directing our attention to critical assertions and providing us with just the right inference procedures to evaluate them. These pragmatic influences may at first sight seem impurities in the crystal clarity of inference, but we must be mindful that even logic rules *can* propagate errors. Rationality and the possession of general reasoning rules do not assure us that the premises are not false, or that situations are not misinterpreted, or that cognitive load will not contribute to erroneous processing. While natural logic may be the basis for rational judgments (Rips 1994, Braine and O'Brien 1998), this does not mean we cannot make errors and do irrational things. An error in the input can be propagated promiscuously throughout our knowledge. Without some pragmatic knowledge to direct our processing or the identification of degrees of truthfulness of propositions, our knowledge base would be completely suspect. To provide a kind of cognitive firewall (Cosmides and Tooby 2000), our inference system relies on the myriad of special representations of information that codes those things that are conjectures, possible truths, fantasies, and what other minds might know and how they may seek to influence us. These not only comprise the fabric of human inference, they are also domains of investigation of modern cognitive psychology. To understand human logical thought, we must also understand our system of knowledge and how it is able to represent the world as it is and as we imagine it to be.

See also: Mental Models, Psychology of; Problem Solving and Reasoning, Psychology of; Rational Theory of Cognition in Psychology; Reasoning with Mental Models

Bibliography

Braine M D S, O'Brien D P (eds.) 1998 *Mental Logic*. Erlbaum, Mahwah, NJ

Bucci W 1978 The interpretation of universal affirmative propositions. *Cognition* **6**: 55–77

Burt C 1919 The development of reasoning in school children. *Journal of Experimental Pedagogy* **5**: 68–77 121–27

Cheng P W, Holyoak K J 1989 On the natural selection of reasoning theories. *Cognition* **33**: 285–313

Clark H H 1969 Linguistic processes in deductive reasoning. *Psychological Review* **76**: 387–404

Cole M, Scribner S 1974 *Culture and Thought: A Psychological Introduction*. Wiley, New York

Cosmides L, Tooby J 1997 Dissecting the computational architecture of social inference mechanisms. In: Bock G, Cardew G (eds.) *Characterizing Human Psychological Adaptations*. Ciba Foundation Symposium No. 208. Wiley, Chichester, UK, pp. 132–61

Cosmides L, Tooby J 2000 Consider the source: The evolution of adaptations for decoupling and metarepresentations. In: Sperber D (ed.) *Metarepresentation*. Vancouver Studies in Cognitive Science. Oxford University Press, New York, pp. 53–115

Cox J R, Griggs R A 1983 The effects of experience on performance in Wason's selection task. *Memory and Cognition* **10**: 496–502

Erickson J R 1978 Research on syllogistic reasoning. In: Revlin R, Mayer R E (eds.) *Human Reasoning*. V. H. Winston, Washington, DC, pp. 39–50

Euler L 1775 *Lettres à une Princesse d'Allemagne sur Divers Sujets de Physique et de Philosophie*. Chez la Société Typographique, Berne, Switzerland

Griggs R A 1978 Drawing inferences from set inclusion information given in text. In: Revlin R, Mayer R (eds.) *Human Reasoning*. Winston, Washington, DC, pp. 163–83

Guyote M J, Sternberg R J 1981 A transitive-chain theory of syllogistic reasoning. *Cognitive Psychology* **13**: 461–525

Holyoak K J, Mah W A 1981 Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory and Cognition* **9**: 197–204

Hunter I M L 1957 The solving of three-term series problems. *British Journal of Psychology* **48**: 286–98

Huttenlocher J 1968 Constructing spatial images: A strategy in reasoning. *Psychological Review* **75**: 550–60

Johnson-Laird P N 1983 *Mental Models*. Cambridge University Press, Cambridge, MA

Johnson-Laird P N, Steedman M 1978 The psychology of syllogisms. *Cognitive Psychology* **10**: 64–99

Nguyen D B, Revlin R 1993 Transitive inferences from narrative relations. *Journal of Experimental Psychology*: *Learning*, *Memory*, *and Cognition* **19**: 1197–210

Potts G R 1974 Storing and retrieving information about ordered relationships. *Journal of Experimental Psychology* **103**: 431–9

Revlin R, Leirer V 1978 Effects of personal biases on syllogistic reasoning: Rational decisions from personalized representations. In: Revlin R, Mayer R (eds.) *Human Reasoning*. Winston, Washington, DC, pp. 51–81

Rips L J 1986 Mental muddles. In: Brand M, Harnish R M (eds.) *Problems in the Representation of Knowledge and Belief*. University of Arizona Press, Tucson, AZ, pp. 258–86

Rips L J 1994 *The Psychology of Proof: Deductive Reasoning in Human Thinking*. MIT Press, Cambridge, MA

Störring G 1908 Experimentelle Untersuchungen über einfache Schlussprozesse [Experimental investigations concerning simple conclusion processes]. *Archiv für die Gesamte Psychologie* **11**: 1–127

Taplin J E 1971 Reasoning with conditional sentences. *Journal of Verbal Learning and Verbal Behavior* **10**: 219–25

Taplin J E, Staudenmayer H, Taddonio J L 1974 Developmental changes in conditional reasoning: Linguistic or logical? *Journal of Experimental Child Psychology* **17**: 360–73

Venn J 1880 On the diagrammatic and mechanical representation of propositions and reasoning. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **9**: 1–18

Wilkins M C 1928 The effect of changed material on ability to do formal syllogistic reasoning. *Archives of Psychology* **16**: 1–83

R. Revlin

Logic and Linguistics

Although the founding fathers of modern logic showed some interest in the analysis of meaning in natural language, solid links between the two disciplines of logic and linguistics had not been established before the middle of the twentieth century. In the late 1950s logical techniques were first applied to the analysis of grammars as theories of linguistic competence. These applications rapidly developed into the field of mathematical linguistics and later became a central part of theoretical computer science. They will be briefly addressed in the second part of this article. The tightest connections between logic and linguistics, however, grew out of the (re-) discovery of logical analysis as a tool of natural language semantics, starting with the advent of Montague Grammar in the early Seventies. The role of logic in the analysis of meaning forms the subject of the first part of this article.

1. *Logic and Natural Language Semantics*

The influence of logic on natural language semantics starts with the key notion of the field: meaning. Following the tradition of analytic philosophy (cf. Ammerman 1965), many linguists take a 'realistic' approach to meaning that makes abundant use of logical apparatus. The basic ideas of the logical analysis of meaning will be explained in Sect. 1.1. Sect.1.2 is concerned with some more specific relations between languages of formal logic and natural languages such as English. (See Barwise and Etchemendy 1990 and Gamut 1991 for introductions to modern logic from a semantic perspective.)

1.1 *The Logical Approach to Meaning*

A primary goal of semantic theory is to describe the *sense relations* that competent speakers perceive between expressions of a language (cf. Lyons 1968, Chap. 10; Larson and Segal 1995, Chap. 1). Thus, e.g., knowledge of the following facts about English phrases is part of native speakers' semantic competence: (a) *brown paper bag* is more specific than, or *hyponymous* to, *paper bag*; (b) *blonde eye-doctor* and *fair-haired ophthalmologist* are *synonymous* in that they necessarily apply to the same persons; (c) *Mary owns a dog* and *Nobody owns a dog* are *incompatible* with each other. Many sense relations can be expressed in terms of *entailments* between sentences: (a) holds because *This is a brown paper bag* entails *This is a paper bag*; (b) holds because *This is a blonde eye-doctor* and *This is a fair-haired ophthalmologist* entail each other; (c) holds because *Nobody owns a dog* entails *It is not the case that Mary owns a dog*. A theory that correctly predicts the entailment relations holding between the sentences of a particular language may therefore be expected to go a long way towards accurately describing its sense relations in general.

In modern logic, the notion of entailment has been successfully formalized by means of *truth conditions*: statement A entails statement B if, and only if, any way of making A true is also a way of making B true. Ways of making statements true are called *models*, which is why the enterprise of describing meaning in terms of truth-conditions is known as *model-theoretic semantics*. More specifically, a model consists of a set of objects, the model's *Universe*, plus an *Interpretation* assigning to every non-logical word a suitable denotation. For instance, the Universe of a specific model might be the set of all numbers, past US presidents, plants, or whatever—any set would do, as long as it is not empty. Similarly, the Interpretation could assign any individual to the name *Mary*—the number 20, Jimmy Carter, my little green cactus—as, long as that individual is in the model's Universe; and it could assign any set (even the empty set) to the noun *ophthalmologist*, as long as that set's members are all in the model's universe, and as long as it is the same set as that which the model assigns to *eye-doctor*. The idea is that any model in which the individual assigned to *Mary* happens to be a member of the set assigned to *ophthalmologist* makes the sentence *Mary is an ophthalmologist* true; and any model in which the set assigned to *ophthalmologist* happens to be empty, is one that makes the sentence *Nobody is an ophthalmologist* true.

In general, then, a model assigns each expression of the language a denotation, thereby deciding the truth and falsity of each (declarative) sentence. In fact, the denotation of a sentence is identified with its *truth value*: the number 1 if the sentence is true, and the number 0 otherwise. Using this numerical convention and writing $[\![X]\!]^{\mathscr{M}}$ for the denotation of an expression X according to a model $\mathscr{M}$, a statement A thus turns out to entail B just in case, for any model $\mathscr{M}$, $[\![A]\!]^{\mathscr{M}} \leqslant [\![B]\!]^{\mathscr{M}}$ whenever A's truth value $[\![A]\!]^{\mathscr{M}}$ according to any model $\mathscr{M}$ is 1, then so is B's truth value $[\![B]\!]^{\mathscr{M}}$ according to the same $\mathscr{M}$.

Traditionally, the statements made true by models are logical formulae. In linguistic semantics, there are two ways of adapting models to the interpretation of natural language: either natural language expressions are translated into logical notation (*indirect interpret-*

ation); or models directly assign denotations to them (*direct interpretation*). On either approach, complex expressions, i.e., anything longer than one word, are usually interpreted following the *Principle of Compositionality* (Janssen 1997): the translation or the denotation of a complex expression consisting of more than one word [like a sentence] is obtained by combining the translations or, respectively, the denotations of its immediate parts [like subject and predicate]. As a consequence, the denotations models assign to (the translations of) natural language expressions tend to be rather abstract. If, for instance, the denotation of the noun phrase *my car*, i.e., a concrete object, is to be obtained by combining the denotations of the noun *car*, which is a set of concrete objects, and that of the possessive *my*, the latter would have to be some 'selection mechanism' that picks out the unique object belonging to the speaker when combined with a set of objects. More abstract denotations turn out to be needed in a compositional treatment of determiners (like *every* and *most*), adverbs (like *quickly*) and prepositions (like *between*). This is one of the reasons why the most common systems of logic, propositional calculus and first-order logic, do not suffice as the basis of indirect interpretation.

The Principle of Compositionality has an important consequence, known as the *Substitution Principle*: if the denotation of a complex expression is completely determined by the denotations of its parts, then one may replace any part by something with the same denotation without thereby changing the denotation of the whole. A case in point is the sentence (*a*) *Mary is an ophthalmologist*. If Mary happens to be the speaker's neighbor, then the name *Mary* and the noun phrase *my neighbor* have the same denotation. Hence, if (*a*) is true and thus denotes the truth value 1, then, by the Substitution Principle, (*b*) *My neighbor is an ophthalmologist* has the same denotation as (*a*), i.e., it is also true. While the Substitution Principle appears to make the right prediction in this case, this is not always so. Together with the assumption that sentences denote truth values, the Substitution Principle immediately leads to what, in the philosophy of language, is known as *Frege's Problem* (Frege 1892): replacing any sentence by another one with the same truth value ought to preserve denotation, which it does not always do. For instance, the denotation, i.e., the truth value, of (*b*) *Nobody doubts that Mary is an ophthalmologist* ought to be preserved when (*a*) is replaced by a sentence with the same truth value. If Mary is an ophthalmologist, the truth value of (*a*) is 1 and hence the same as that of (*c*) *Paris is the capital of France*. However, (*b*) might well be false while (*b′*) *Nobody doubts that Paris is the capital of France is* true. Clausal embedding under so-called *attitude verbs* like *doubt*, *know*, or *tell*, thus turns out to be an *intensional construction*, i.e., one that defies the Substitution Principle.

Various solutions to Frege's Problem have been proposed. Practically all of them employ the idea that, at least in certain constructions, the contribution that a sentence makes to the denotations of larger expressions is not its truth value but something else. According to a widely accepted view (originating with Frege), it is its *informational content*: attitude verbs like *doubt* denote a relation between the individual denoted by the subject and the information conveyed by the embedded clause. A common way of modelling informational content is by means of *possible worlds*—whence the term *possible worlds semantics*. (See Barwise 1989 and Devlin 1991 for alternative approaches.) The basic idea is akin to that of measuring the expectance value of an event by the proportion between favorable and unfavorable cases. The cases become possible worlds, the event is the information (as expressed by a sentence) being correct and instead of the quantitative relation between the sets of cases expressed by a ratio, it is the qualitative difference that is emphasized, i.e., which worlds are favorable and which are not. Following this line of thought, the informational content of a sentence, then, can be represented by a function (in the mathematical sense) pairing worlds with truth values. For example, the informational content of (*a*) above becomes a function assigning to any possible world the truth value 1 if that world is such that, in it, Mary is an ophthalmologist; all other worlds will be assigned the truth value 0. Each possible world is thought of as representing a specific way the world could be, or might have been—to the last detail. So the informational content of a sentence will typically assign 1 to an infinity of possible worlds differing in all sorts of aspects that are irrelevant to the truth of that sentence; and it will typically assign 0 to innumerably many worlds differing in matters that do not affect the falsity of the sentence.

In order to give a full solution to Frege's Puzzle within possible worlds semantics, the informational content of each sentence must be systematically determined, which is again done compositionally, replacing denotations by *intensions*. The intensions of sentences are their informational contents, but the notion of intension is more general in that it applies to all kinds of expressions. The intension of any expression is its denotation (or *extension*) as it varies from world to world, i.e., a function (in the mathematical sense) assigning to each possible world the extension of the expression in that world. As a case in point, one may compare the descriptions (*d*) *the present capital of France* and (*e*) *the largest city in France*: in some worlds, including our actual reality, their extensions coincide, but given a different course of history, as represented by a possible world *w*, Reims, though smaller than Paris, would be the capital of France. So the intensions of (*d*) and (*e*) cannot be the same, because they disagree on *w* (and on many other worlds). In general, then, expressions with the same denotations may still differ in their intensions and hence, given compositionality, in the contribution

they make to the informational contents of the sentences in which they occur.

The method of extension and intension complements, rather than replaces, the model-theoretic approach to meaning. Instead of directly assigning arbitrary denotations to expressions, the models of possible worlds semantics couple each expression with an arbitrary intension, which then, being a function, uniquely determines an extension relative to a given world. Moreover, just like models come with their own Universes, in possible worlds semantics each model is also equipped with its own Logical Space of worlds. This construction originates from *modal logic* (see below) and offers an alternative to the model-theoretic reconstruction of entailment: a sentence *A strictly implies* a sentence *B* if *B* is true in all worlds in which *A* is true, i.e., if the intension of *B* yields the truth value 1 for any possible world for which the intension of *A* yields 1.

1.2 Logic and Language

The basic tools of logically based linguistic semantics—models and possible worlds—had originally been developed for artificial languages designed for specific purposes, like the analysis of mathematical proofs. These logical formulas can usually be paraphrased in colloquial language; the relation between formula and paraphrase is taught in introductory logic classes. Conversely, many constructions in natural language (passive, relative clauses, …) have long been known to be systematically expressible in languages of logic (cf. Quine 1960); *formalization*, i.e., the art of translating ordinary sentences into logical notation, is also something to be picked up in an introduction to logic. As a side-effect of the logical approach to natural language semantics, these relations between natural language grammar and logical notation can be explored in a more rigorous way. We will briefly go through some pertinent examples; more detailed case-studies can be looked up in any textbook or handbook on formal semantics (Dowty et al. 1981, Heim and Kratzer 1998, Gabbay and Guenthner 1984, Vol. 4, von Stechow and Wunderlich 1991, van Benthem and ter Meulen 1997).

The most obvious connections between logic and language concern the meaning of *logical words*, i.e., those lexical items that translate the basic inventory of logical formalism: the ('*Boolean*') *connectives* '¬,' '∧,' '∨,' … translated as *[it is] not [the case that]*, and, or, … and the *quantifiers* '∀' and '∃' corresponding to *everything* and *something*. The logicians' version of the former are combinations of truth-values. For example, just like a sentence of the form '*A and B*' is true just in case both *A* and *B* are, '∧' combines 1 and 1 into 1 and the other pairs of truth values into 0. Similarly, '¬' (glossed as *not*) turns truth (1) into falsity and vice versa; '∨' (∼ *or*) combines 0 and 0 into 0 and everything else into 1, etc. These combinations of truth values can thus be given by means of the *truth tables* in Table 1.

Table 1
Truth tables for Boolean connectives

A	B	¬A [negation; gloss: not *A*]	A ∧ B [conjunction; *A* and *B*]	A ∨ B [disjunction; *A* or *B*]
1	1	0	1	1
1	0	0	0	1
0	1	1	0	1
0	0	1	0	0

Even though truth tables work fine as first approximations to a semantic analysis of the words *not*, *and* and *or* as they are used in mathematical jargon, there is more to the meaning of these words. To begin with, *and* and *or* do not always stand between sentences but may be used to connect almost all kinds of expressions: *every book and a pen*; *see and/or hear*; *true or false*; etc. However, in many cases these non-sentential uses abbreviate coordinated sentences: *John owns every book and a pen* is short for *John owns every book and John owns a pen*, etc. In fact, non-sentential *and* and *or* can frequently be systematically reduced to the relevant truth tables by a *type shift*, a semantic rule that systematically generalizes combinations of truth values to combinations of other denotations (see Keenan and Faltz 1985, Hendriks 1993). However, even this generalization does not cover all non-sentential uses of these words. In particular, there is no obvious reduction of the group reading of *and*, as in *John and Mary are a married couple*, to the truth table of Boolean conjunction. Neither does there seem to be a straightforward reduction of alternative *or*, as in questions like *Shall we leave or do you want us to stay?*, to Boolean disjunction. Certain other semantic aspects that the truth table analysis seems to miss may be captured by pragmatic principles governing efficient use, as opposed to literal meaning (Grice 1989, Levinson 1983). As a case in point, the difference between *John lit a cigarette and Mary left the room* and *Mary left the room and John lit a cigarette* may be explained by a general principle of discourse organization, according to which the order of sentences should match the order of the events they describe. More involved pragmatic reasonings have been given to account for the exclusive sense of *or*, according to which *John missed the train*, *or he decided to sleep late* implies that John did not both miss the train and decide to sleep late.

In addition to the Boolean connectives, standard formal languages of logic contain both a *universal* and an *existential quantifier*, usually translated as *everything* and *something* and expressing properties of predicates: being universal, i.e., true of everything, and being non-empty, i.e., applying to something. Natural

language statements are only expressible in standard ('first-order') logic if they can be rephrased by combinations of Boolean connectives and the two quantifiers. A case in point is the sentence (*N*) *Every boy loves a girl* that can be paraphrased by (*L*) *For everything it holds that either it is not a boy, or for something it holds that it is a girl and the former loves the latter*—which translates into predicate logic as (*F*) $(\forall x)\ [\neg\ B(x) \vee (\exists y)\ [G(y) \wedge L(x,y)]]$. As the logician's paraphrase (*L*) indicates, the internal structure of quantified statements in natural language is quite different from that in logic. In particular, it usually involves *determiners* like *every* and *a* instead of the natural language quantifiers *everything* and *something*. The difference is that a determiner relates *two* sets, whereas a quantifier makes a statement about *one* set. Thus, in (*N*), *every* relates the set of boys with the set of girl-lovers, whereas the quantifier *everything* in (*L*) as well as the universal quantifier '$\forall$' in (*F*) attribute universality to the set of individuals that love girls if they are boys, i.e., the set of those that are either not boys or else love a girl.

Another means of expressing quantification in natural language is by way of certain *adverbs*. In particular, in so-called *donkey sentences* like *If a farmer owns a donkey*, *he never beats it*, the adverb *never* (on one reading) expresses that the set of donkey-owning farmers that beat their donkeys is empty. The very form of this *adverbial quantification* is rather remote from the usual logical formalization; it is not even clear whether in this case, as is usually assumed (see above), the indefinites *a farmer* and *a donkey* correspond to existential quantifiers. Recent work in natural language semantics has developed special logical systems to give a more adequate logical account of indefinites in donkey sentences and related constructions. (See Kamp and Reyle 1993, Chierchia 1995.)

Apart from these structural differences in the formulation of quantified statements, there are also differences in expressive power. In particular, not all natural language determiners can be paraphrased by combining Boolean connectives and logical quantifiers. For instance, it can be shown that a statement like *Most boys are asleep* cannot be expressed in first-order logic (Barwise and Cooper 1981). Adequate formalizations of such sentences, and hence of natural language in general, again call for more powerful and complex languages of logic. (See *Quantifiers, in Linguistics* for more about quantification in natural language and its relation to logic.)

Logicians have devised various extensions of classical propositional and predicate logic as tools of formalizing reasoning involving specific locutions (cf. Gabbay and Guenthner 1984, Vol. 2). For instance, in addition to Boolean connectives, languages of (propositional) *modal logic* (Chellas 1993, Bull and Segerberg 1984) contain a *necessity* operator, usually written as '$\Box$'; a formula '$\Box\varphi$' [read: it is necessarily so that φ] is true *in a world w* of a given model if the formula φ itself is true in all worlds that are possible for (or, technically: *accessible* from) *w*. Combining negation and necessity, it is also possible to express a corresponding notion of possibility as '$\neg\ \Box\ \neg$.' Models of modal logic thus contain domains of possible worlds organized by an accessibility relation. Depending on the precise structure of the latter, modal logic can be used to formalize various uses of modal verbs as in *Someone must have been here* (*epistemic* must) or *You may now leave the room* (*deontic* may). However, detailed studies of modality have revealed that these formalizations are at best approximations of the meanings of modal verbs in natural language (cf. Kratzer 1991). A similar relation is that between *tense logic* (Burgess 1984) and the tenses and other means of reference to time in natural language (see *Tense, Aspect, and Mood, Linguistics of*).

Various phenomena have led semanticists to explore alternatives to classical logic with its model-theoretic interpretation assigning one of two truth values to each formula or sentence (cf. Gabbay and Guenthner 1984, Vol. 3). A sentence like (*1*) *John knows that there will be a party tonight* seems to imply (*P*) *There will be a party tonight*; on the other hand, so does its negation (*2*) *John doesn't know that there will be a party tonight*. But if (*1*) and (*2*) both implied (*P*) in the sense of classical model-theory, then (*P*) would be true throughout the models in which (*1*) is true—as well as in those models in which (*2*) is true, i.e., the ones in which (*1*) is false. But then (*3*) would have to be a *tautology*, i.e., a statement that is true in all models; for any model would have to make (*1*) true or false. Giving up this last assumption, one enters the realm of (certain) *non-classical logics*, the simplest among which are *three-valued logics* with an additional truth value *U* (read: *undefined*) which models assign to sentences like (*1*) if their *presupposition* (*P*) is not true in them. (See *Linguistic Presupposition* for more on this.)

2. The Logic of Grammar

As theories of language structure, grammars themselves have been the objects of logical analysis. The earliest such investigations mark the beginning of generative grammar and were directed to shedding a light on the complexity of human language (see Sect. 2.1). Later approaches focused on the role of grammatical rules in the activity of parsing, which bears some resemblance to the process of deduction in logic; and more recently, grammatical reasoning has been seen as a non-monotonic process, thus reaching beyond the limits of classical logic (see Sect. 2.2).

2.1 Syntactic Complexity

A syntactic theory of a particular language may be seen as a body of theoretical statements describing the

structure of grammatically correct phrases. Formal reconstruction reveals that these theories come in different formats and complexities. For instance, a *context-free* description of a language only contains grammatical rules of the form: 'If a phrase of category B is immediately followed by a phrase of category C, the result will be a phrase of category A.' The standard notation for a rule of this form is: '$A \rightarrow B + C$'. Thus the following rule of English grammar is context-free: *if a preposition* (like *about*) *is followed by a noun phrase* (like *corrupt politicians*), *they form a prepositional phrase* (*about corrupt politicians*). In standard notation, this rule reads: '$PP \rightarrow Prep + NP$.' On the other hand, the following would-be rule is *not* context-free: *after a ditransitive verb* (like *give*), *two noun phrases* (like *every kid* and *a penny*) *form an object-phrase* (*a kid a penny*). Rather, the rule comes under the more complex scheme: 'If a phrase of category B is immediately followed by a phrase of category C, the result will be a phrase of category A, provided that it follows a phrase of category D'; one just has to put A = *object-phrase*, $B = C =$ *noun phrase*, and $D =$ *ditransitive verb*. Rules of this form are called *context-sensitive*. Using standard notation, they are written as: '$A \rightarrow B + C/D__$.'

The distinction between context-free and context-sensitive rules—and similar distinctions between other kinds of grammatical rules—can be used to define an abstract complexity ranking between languages, the *Chomsky Hierarchy* (after Chomsky 1957, 1959). A language is context-free (or, as the case may be, context-sensitive) if it can in principle be described by a body of context-free (or, respectively, context-sensitive) rules. Languages in the sense of this definition are arbitrary sets of sentences, which themselves are arbitrary strings of words. It turns out that, in this abstract mathematical setting (known as *mathematical linguistics* or *formal language theory*), though all context-free languages are trivially context-sensitive, the converse is not the case: one can construct languages that are context-sensitive without being describable by any body of context-free rules whatsoever. A famous example is the 'language' $a^n b^n c^n$ whose 'sentences' consist of arbitrarily many occurrences of the 'word' a followed as many bs, which are again followed by the same number of cs. (A proof and further examples can be found in Martin 1997, Chap. 8.)

In applying these abstract concepts, one must identify a language like English with the set of its grammatical sentences and individuate the latter by their words only—and not, say, by their internal grammatical structure. Given this identification, it is natural to ask for the exact location of human languages in the Chomsky Hierarchy. It may seem that an answer to this question requires complete knowledge of the grammatical sentences of any given language—something that current linguistic research has not attained. However, at least some of these complexity issues can be settled on the basis of partial information about the languages under investigation. In particular, it has been shown that, given the intricate structure of its verb complexes, the Swiss German dialect spoken in Zurich cannot be captured by context-free means (Shieber 1985); a similar result concerning word formation has been obtained for the African language Bambara spoken in Mali (Culy 1985), thus proving that natural languages in general are not context-free.

2.2 Further Issues

The fact that ordinary syntactic notation looks rather different from logical symbolism should not distract from the close connections between the two. For instance, as indicated above, a context-free rule only abbreviates a lengthy statement about the internal structure of phrases; and this statement can be easily (though somewhat clumsily) expressed in standard logical symbolism (= first-order predicate logic). This fact becomes particularly important when it comes to the task of *parsing* phrases of a given language according to a given grammar. As a case in point, parsing the prepositional phrase *about Mary* on the basis of the above context-free rule boils down to making a connection between the following three pieces of information: (*I*) *about* is a preposition; (*II*) *Mary* is a noun phrase; and (*III*) if x is a preposition and y is a noun phrase, then $x\ y$ (i.e., x followed by y) is a prepositional phrase. (*I*) and (*II*) are basic facts about English words to be recorded in a *lexicon* (which may be thought of as part of a syntactic description); (*III*) is just a paraphrase of the above context-free rule: '$PP \rightarrow Prep + NP$.' (*I*)–(*III*) obviously entail that *about Mary* is a prepositional phrase, and this is precisely the kind of entailment that can be formalized by means of elementary logic. Hence it should not come as a surprise that there are close relations between parsing strategies and procedures for verifying logical entailments. This connection has been studied in computational linguistics and in categorical grammar under the heading *Parsing as Deduction* (see Johnson 1991).

The role of logical deduction in grammatical theory is more complex than can be indicated in an article of this length. In particular, certain phenomena like the markedness of grammatical features or the interaction of phonological constraints, seem to call for non-classical logics that violate the *Law of Monotonicity*, according to which a conclusion once drawn remains correct when more premises are added.

See also: Generative Grammar; Lexical Semantics; Logic and Cognition, Psychology of; Logics for Knowledge Representation; Meaning and Rule-following: Philosophical Aspects; Montague Grammar; Semantic Knowledge: Neural Basis of; Semantics

Bibliography

Ammerman R R (ed.) 1965 *Classics of Analytic Philosophy*. McGraw-Hill, New York

Barwise 1989 *The Situation in Logic*. CSLI, Stanford, CA

Barwise J, Cooper R 1981 Generalized quantifiers and natural language. *Linguistics and Philosophy* **4**: 159–219

Barwise J, Etchemendy J 1990 *The Language of First-Order Logic*. CSLI, Stanford, CA

van Benthem J, ter Meulen A (eds.) 1997 *Handbook of Logic and Language*. Elsevier, Amsterdam

Bull R A, Segerberg K 1984 Basic modal logic. In: Gabbay D, Guenthner F (eds.) *Handbook of Philosophical Logic*. Kluwer, Dordrecht, Vol. 2

Burgess J 1984 Basic tense logic. In: Gabbay D, Guenthner F (eds.) *Handbook of Philosophical Logic*. Kluwer, Dordrecht, Vol. 2

Chellas B 1993 *Modal Logic*. Cambridge University Press, Cambridge, UK

Chierchia G 1995 *Dynamics of Meaning*. University of Chicago Press, Chicago

Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands

Chomsky N 1959 On certain formal properties of grammars. *Information and Control* **2**: 137–67

Culy C 1985 The complexity of the vocabulary of Bambara. *Linguistics and Philosophy* **8**: 345–51

Devlin K 1991 *Logic and Information*. Cambridge University Press, Cambridge, UK

Dowty D, Wall R, Peters S 1981 *Introduction to Montague Semantics*. Kluwer, Dordrecht, The Netherlands

Frege G 1892 Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* **100**: 25–50

Gabbay D, Guenthner F (eds.) 1984 *Handbook of Philosophical Logic*, 4 vo1s. Kluwer, Dordrecht, The Netherlands

Gamut L T F 1991 *Logic, Language, and Meaning*, 2 vols. University of Chicago Press, Chicago

Grice P 1989 *Studies in the Way of Words*. Harvard University Press, Cambridge, MA

Heim I, Kratzer A 1998 *Semantics in Generative Grammar*. Blackwell, Oxford, UK

Hendriks H 1993 Studied flexibility: Categories and types in syntax and semantics. Ph.D. thesis, University of Amsterdam

Janssen T M V 1997 Compositionality (with an apppendix by B H Partee). In: van Benthem J, ter Meulen A (eds.) *Handbook of Logic and Language*. Elsevier, Amsterdam

Johnson M 1991 Deductive parsing: The use of knowledge of language. In: Berwick R C et al. (ed.) *Principle-based Parsing: Computation and Psycholinguistics*. Kluwer, Dordrecht, The Netherlands

Kamp H, Reyle U 1993 *From Discourse to Logic*. Kluwer, Dordrecht, The Netherlands

Keenan E, Faltz L 1985 *Boolean Semantics for Natural Language*. Reidel, Dordrecht, The Netherlands

Kratzer A 1991 Modality. In: von Stechow A, Wunderlich D (eds.) *Semantik. Ein internationales Handbuch zeitgenössischer Forschung*. [*Semantics. An International Handbook of Contemporary Research*.]. De Gruyter, Berlin

Larson R, Segal G 1995 *Knowledge of Meaning*. MIT Press, Cambridge, MA

Levinson, S C 1983 *Pragmatics*. Cambridge University Press, Cambridge, UK

Lyons J 1968 *Introduction to Theoretical Linguistics*. Cambridge University Press, Cambridge, UK

Martin J C 1997 *Introduction to Languages and the Theory of Computation*, 2nd ed. McGraw-Hill, Boston

Quine W V 1960 *Word and Object*. MIT Press, Cambridge, MA

Shieber S M 1985 Evidence against the context-freeness of natural language. *Linguistics and Philosophy* **8**: 333–43

Stechow A von, Wunderlich D (eds.) 1991 *Semantik. Ein internationales Handbuch zeitgenössischer Forschung*. [*Semantics. An International Handbook of Contemporary Research*.] De Gruyter, Berlin

T. Zimmermann

Logical Positivism and Logical Empiricism

The philosophy of the Vienna Circle, which revolved around the work of Moritz Schlick, Rudolf Carnap, and Otto Neurath in Vienna in the 1920s and 1930s, was strongly influenced by Ludwig Wittgenstein, Bertrand Russell, Ernst Mach, and Henri Poincaré. It was one of the most important parts of twentieth century analytic philosophy. Together with related tendencies in Poland and Berlin it shaped nearly all serious philosophy of science in the twentieth century, gave rise to many more or less sympathetic developments of its main claims, and to a great deal of opposition, for example the philosophy of science of Karl Popper.

The logical positivism of the Vienna Circle, which came also to be called 'logical empiricism,' contained a number of claims about the nature of empirical science, in particular a strong rejection of the view that the subject-matter, methods, and logic of the natural sciences differ from those of the social and human sciences (the 'unified science' program). Three surveys of the philosophy of logical positivism are Kraft (1968), Barone (1986), and Haller (1993); a very thorough history of the movement and of its influence is provided in Stadler (1997); Ayer (1959) is a representative anthology of work by the logical positivists and associated philosophers.

One remarkable sociological feature of the writings of philosophers belonging to or associated with the Vienna Circle is the militant, polemical tone of the criticisms of alternative accounts of the nature of the nonnatural sciences. Zilsel (1941, p. 131) identified the Circle's philosophical enemies as belonging to the 'anticausal rebellion' in philosophy after 1890 and mentions in this connexion Windelband, Rickert, Dilthey, Weber, Troeltsch, neovitalism, and phenomenology. This polemical attitude went hand in hand with great self-consciousness about the historical, social, and sociological position of the movement. Neurath's ([1936], 1981) monograph, first published in French as 'Le développement du Cercle de Vienne et

l'avenir de l'Empirisme logique' (The development of the Vienna Circle and the future of logical empiricism), which stresses the Austrian context of logical empiricism (Bolzano, Brentano, the near absence of Kantianism in Austria) is exemplary in this connexion (see also Zilsel 1992 and, on Neurath, Hegelsmann 1979).

Among the more important early studies of the foundations of the human and social sciences are Neurath's ([1931], 1981) monograph 'Empirische Soziologie. Der wissenschaftliche Gehalt der Geschichte und Nationalökonomie' (Empirical sociology. The scientific content of history and of political economy), his *Foundations of the Social Sciences* (1944) and two books by Felix Kaufmann (1936, 1944) on the methodology of the social sciences. In addition to their work on the nature of science, members of the Circle such as Edgar Zilsel wrote on the history and sociology of science.

Surprisingly, one of the most important philosophical works to emerge from the Circle, Carnap's ([1928], 1968). *Der logische Aufbau der Welt* (*The Logical Structure of the World*) contains the claim that *geistige* (spiritual, intellectual) objects, i.e., cultural, historical, and sociological objects, must be distinguished from both physical and psychological objects; the autonomy of this class of objects has, Carnap says, too often been overlooked; recent philosophy of history after Dilthey is credited with the discovery of this autonomy. Carnap distinguishes between episodic physical and psychological manifestations of spiritual objects and the enduring physical objects which are documentations of spiritual entities—the buildings that document an artistic style ([1928], 1968, sections 23–4). He explains that he is able to provide only hints as to how cultural objects are to be constructed since 'the psychology (phenomenology) of cultural knowledge' is much less developed than that of perception (Carnap [1928], 1968, Sects. 150–1).

But thanks to Neurath the logical positivists came to reject the view that physical, psychological, and cultural entities all deserve different cages in the philosopher's zoo. The doctrine of physicalism, as expounded in Neurath's ([1931], 1959) paper 'Soziologie im Physikalismus' (Sociology in physicalism) and in his ([1931], 1981) monograph, 'Empirische Soziologie', is the view that:

> every scientific proposition is a proposition about a law-like ordering of empirical facts. All scientific propositions can be connected with one another and form a unified domain, which embraces only propositions about observable facts. For this domain the name Unified Science is suggested. If one wants to emphasise that in this way everything turns out to be physics, one may speak of physicalism. (Neurath [1931], 1981, p. 424)

Since 'observable' means observable by different people (intersubjectivity) psychological entities must be physical; since the facts in question are the spatiotemporal entities of physics, 'propositions about sociological facts' must be about such entities. Psychology, history, and political economy deal only in gestures, words, ways of acting, but not in motives, subjects, personalities, if these cannot be described as spatiotemporal entities (Neurath [1931], 1981, pp. 425, 431). The language of Unified Science is 'derivable by and large from modifications of the language of everyday life' once all 'pre-physicalistic' components have been eliminated (Neurath [1931], 1981, p. 537). Sociology is no *Geisteswissenschaft* but 'social behaviorism' (Neurath [1931], 1981, sections 3–4). Marxism, he thinks, comes closer than any other present-day sociological theory to being a genuinely empirical sociology (Sect. 7).

Whereas interest in the status of sociology looms large in Neurath's formulation of physicalism, it is psychology that is considered at length in Carnap's turn towards physicalism in his paper, 'Die physikalische Sprache als Universalsprache der Wissenschaft' (Physical language as the universal language of science, Carnap 1931, cf. also Carnap 1932/3). But Carnap also claims that:

> if our thesis concerning the translatability of psychological propositions into the physical language is correct, the corresponding thesis for the propositions of (empirical) sociology is readily grasped ... all historical, cultural, economic processes belong here. But we have in mind only the genuinely scientific, logically acceptable propositions of this domain. In the so-called 'Geisteswisenschaften' or 'sciences of culture', as they are at present, logical analysis often reveals pseudo-concepts, concepts which have no correct definition, concepts therefore for which no empirical criteria have been laid down ... (Examples: 'objective Geist (spirit)', 'meaning of history') ... sociology deals with nothing other than states, processes, ways of behaving of groups or individual subjects (humans or other animals), reciprocal reactions, and reactions to environmental processes (Carnap 1931, p. 451)

Carnap rejects all alternatives to observation such as *Verstehen* (understanding), empathy, and feeling one's way into historical products and events in order to grasp their essence or sense (p. 434).

Kaufmann was a regular participant in the meetings of the Vienna Circle, rather than a member of it, and was heavily influenced by phenomenology. But it is his work on a finitistic philosophy of mathematics rather than on the foundations of the social sciences which is phenomenological. In his philosophy of science he argues for a distinction between two sorts of laws: empirical, synthetic propositions, on the one hand, and scientific rules of procedure (theoretical laws) in terms of which 'warranted predictions' are defined, on the other hand. The distinction is applied so as to yield a unified account of scientific methodology or of the logic of scientific procedure, to criticize physicalism and to bring out differences between the subject matters of the natural and the nonnatural sciences. Of particular interest are Kaufmann's analyses of the

conceptual foundations of a number of different social and human sciences, in particular his discussion of what some claim has to be considered as the most important research program in the social sciences in Vienna at the time, Austrian marginal utility theory, in the course of which he rejects the apriorism of Ludwig von Mises (Kaufmann 1936, pp. 255–90, Kaufmann 1944 Chap. XVI).

Although it was the positivists' work on the logical structure of scientific theories that attracted most attention, they also wrote extensively on the history and sociology of science and on the sociology of knowledge. This is particularly true of Neurath, in the publications already mentioned, and of Edgar Zilsel—both heavily influenced by Marx. Neurath's holism and sociological approach to science lead to a view which anticipates some aspects of the views later developed by T. S. Kuhn (see Haller 1986, pp. 130ff.). Zilsel (1942, 1976) investigated in detail the role of craftsmen as pioneers of causal thinking in the period before 1600 and made many contributions to a program he formulated as follows:

> The rise of science is usually studied by historians who are primarily interested in the temporal succession of the scientific discoveries. Yet the genesis of science can be studied also as a sociological phenomenon. The occupations of the scientific authors and of their predecessors can be ascertained. The sociological function of these occupations and their professional ideals can be analyzed. The temporal succession can be interrupted and relevant sociological groups can be compared to analogous groups in other periods.... It is strange how rarely investigations of this kind are made.... there is no reason why the most important and interesting intellectual phenomena should not be investigated sociologically and causally. (Zilsel 1942, p. 560)

Perhaps the most thorough confrontation of views like those of the logical positivists with the views of such founders of the sociology of knowledge as Scheler and Mannheim is given by Kaufmann (1944, ch. XIV).

Logical positivism shaped subsequent philosophy of science. Such classic investigations of the nature of scientific laws and explanation as Ernest Nagel's *The Structure of Science* (1961) and Carl Hempel's (1965) collection of papers *Aspects of Scientific Explanation* emerged from the discussions of the Viennese logical positivists and of related groups in Berlin and Poland.

See also: Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Cognitive Science: Philosophical Aspects; Empiricism, History of; Positivism, History of; Positivism: Sociological; Realisms and their Opponents: Philosophical Aspects

Bibliography

Ayer A J (ed.) 1959 *Logical Positivism*. Free Press, Glencoe, IL

Ayer A J 1963 The Vienna Circle. In: Ayer A J et al. (eds.) *The Revolution in Philosophy*. Macmillan, London

Barone F 1986 *Il Neopositivismo logico*, 2 vols., Laterza, Rome

Carnap R 1931 Die physikalische Sprache als Universalsprache der Wissenschaft. *Erkenntnis* **2**: 432–65

Carnap R [1932/3] 1959 Psychology in physical language. In: Ayer A J (ed.) *Logical Positivism*. Free Press, Glencoe, IL, pp. 165–98

Carnap R [1928] 1968 *The Logical Structure of the World. Pseudoproblems in Philosophy* (trans George R A). Routledge & Kegan Paul, London

Haller R 1986 *Fragen zu Wittgenstein und Aufsätze zur öesterreichischen Philosophie*. Rodopi, Amsterdam

Haller R 1993 *Neopositivismus. Eine historische Einführung in die Philosophie des Wiener Kreises*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany

Hegelsmann R 1979 Otto Neurath—Empiristischer Aufklärer und Sozialreformer. In: Neurath O (ed.) *Wissenschaftliche Weltauffassung, Sozialismus und Logischer Empirismus*. Suhrkamp, Frankfurt am Main, pp. 7–78

Hempel C 1965 *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press, New York

Kaufmann F 1929/30 Soziale Kollektiva. *Zeitschrift für Nationalökonomie* **1**: 294–308

Kaufmann F 1936 *Methodenlehre der Sozialwissenschaften*. Springer, Vienna

Kaufmann F 1944 *Methodology of the Social Sciences*. Oxford

Kraft V 1968 *Der Wiener Kreis. Der Ursprung des Neopositivismus. Ein Kapitel aus der jüngsten Philosophiegeschichte*. Springer Verlag, Vienna

Nagel E 1961 *The Structure of Science: Problems in the Logic of Scientific Explanation*. Harcourt, Brace & World, New York

Neurath O [1931] 1959 Soziologie im Physikalismus. In: Ayer A J (ed.) *Logical Positivism*. Free Press, Glencoe, IL, pp. 282–317

Neurath O [1931] 1981 Empirische Soziologie. Der wissenschaftliche Gehalt der Geschichte und Nationalökonomie, Schriften zur wissenschaftlichen Weltauffassung, Vienna. In: Neurath O (ed.) *Gesammelte philosophische und methodologische Schriften*. Verlag Hölder-Pichler-Tempsky, Vienna, Vol. 1, pp. 423–527

Neurath O [1936] 1981 Die Entwicklung des Wiener Kreises und die Zukunft des Logischen Empirismus. In: Neurath O (ed.) *Gesammelte philosophische und methodologische Schriften*. Verlag Hölder-Pichler-Tempsky, Vienna, Vol. 2, pp. 673–702

Neurath O 1981 *Gesammelte philosophische und methodologische Schriften* (2 vols, Haller R, Rutte H eds). Verlag Hölder-Pichler-Tempsky, Vienna

Stadler F 1997 *Studien zum Wiener Kreis. Ursprung, Entwicklung und Wirkung des Logischen Empirismus im Kontext*.

Zilsel E 1941 Phenomenology and natural science. *Philosophy of Science* **8**: 26–32

Zilsel E 1942 The sociological roots of science. *American Journal of Sociology* : 544–62

Zilsel E 1976 *Die sozialen Ursprünge der neuzeitlichen Wissenschaft*. Frankfurt am Main

Zilsel E 1992 *Wissenschaft und Weltanschauung*. Aufsätze 1929–1933. Böhlau Verlag, Vienna

K. Mulligan

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Logics for Knowledge Representation

Knowledge representation and reasoning play a central role in Artificial Intelligence (AI). Research in AI started off by trying to identify the general mechanisms responsible for intelligent behavior. However, it quickly became obvious that general and powerful methods are not enough to get the desired result, namely, intelligent behavior. Almost all tasks a human can perform which are considered to require intelligence are also based on a huge amount of knowledge. For instance, understanding and producing natural language relies heavily on knowledge about the language, about the structure of the world, about social relationships, etc.

One way to address the problem of representing knowledge and reasoning about it is to use some form of logic. While this seems to be a natural choice, it took a while before this 'logical point of view' became the prevalent approach in the area of knowledge representation. This article describes how the field of knowledge representation evolved and what kind of logical methods have been used. In particular, it argues that the important point about using formal logic is the logical method.

1. *Logic-based Knowledge Representation: A Historical Account*

McCarthy (1968) stated very early on that mathematical, formal logic appears to be a promising tool for achieving human-level intelligence on computers. In fact, this is still McCarthy's (2000) vision, which he shares with many researchers in AI. However, in the early days of AI, there were also a number of researchers with a completely different opinion. Minsky (1975), for example, argued that knowledge representation formalisms should be flexible and informal. Moreover, he argued that the logical notions of correctness and completeness are inappropriate in a knowledge representation context.

While in those days heated arguments of the suitability of logic were exchanged, by the end of the eighties, the logical perspective seemed to have gained the upper hand (Brachman 1990). During the nineties almost all research in the area of knowledge representation and reasoning was based on formal, logical methods as demonstrated by the papers published in the biannual international conference on *Principles of Knowledge Representation and Reasoning*, which started in 1989.

It should be noted, however, that two perspectives on logic are possible. The first perspective, taken by McCarthy (1968), is that logic should be used to represent knowledge. That is, we use logic as the representational and reasoning tool inside the computer. Newell (1982), on the other hand, proposed in his seminal paper on the knowledge level to use logic as a formal tool to analyze knowledge. Of course, these two views are not incompatible. Furthermore, once we accept that formal logic should be used as a tool for analyzing knowledge, it is a natural consequence to use logic for representing knowledge and for reasoning about it as well.

2. *Knowledge Representation Formalisms and Their Semantics*

Saying that logic is used as the main formal tool does not say which kind of logic is used. In fact, a large variety of logics (Gabbay et al. 1995) have been employed or developed in order to solve knowledge representation and reasoning problems. Often, one started with a vaguely specified problem, developed some kind of knowledge representation formalism without a formal semantics, and only later started to provide a formal semantics. Using this semantics, one could then analyze the complexity of the reasoning problems and develop sound and complete reasoning algorithms. This will be called the logical method, which proved to be very fruitful in the past and has a lot of potential for the future.

2.1 *Description Logics*

One good example for the evolution of knowledge representation formalisms is the development of description logics, which have their roots in so-called structured inheritance networks formalisms such as KL-ONE (Brachman 1979). These networks originally were developed in order to represent word meanings. A concept node connects to other concept nodes using roles. Moreover, the roles could be structured as well. These networks permit, for example, the definition of the concept of a bachelor.

Later on, these structured inheritance networks were formalized as so-called concept languages, terminological logics, or description logics. Concepts were interpreted as unary predicates, roles as binary relations, and the connections between nodes as so-called value restrictions. This leads for most such description logics to a particular fragment of first-order predicate logic, namely, the fragment L_2. In this fragment only two different variable symbols are used. As it turns out, this is a decidable fragment of first-order logic.

However, some of the more involved description logics go beyond L_2. They contain, for example, relational composition or transitive closure. As it turns out, such description logics can be understood as variants of multi-modal logics (Schild 1991), and

decidability and complexity results from these multimodal logics carry over to the description logics. Furthermore, description logics are very close to feature logics as they are used in unification-based grammars. In fact, description logics and feature logics can be viewed as members of the same family of representation formalisms (Nebel and Smolka 1990).

All these insights, that is, determination of decidability and complexity as well as the design of decision algorithms (e.g., Donini et al. 1991), are based on the rigorous formalization of the initial ideas. In particular, it is not just one logic that it is used to derive these results, but it is the logical method that led to the success. One starts with a specification of how expressions of the language or formalism have to be interpreted in formal terms. Based on that one can specify when a set of formulae logically implies a formula. Then one can start to find similar formalisms (e.g., modal logics) and prove equivalences and/or one can specify a method to derive logically entailed sentences and prove them to be correct and complete.

2.2 Nonmonotonic Logics

Another interesting area where the logical method has been applied is the development of the so-called nonmonotonic logics. These are based on the intuition that sometimes a logical consequence should be retracted if new evidence becomes known. For example, we may assume that our car will not be moved by somebody else after we have parked it. However, if new information becomes known, such as the fact that the car is not at the place where we have parked it, we are ready to drop the assumption that our car has not been moved.

This general reasoning pattern was used quite regularly in early AI systems, but it took a while before it was analyzed from a logical point of view. In 1980, a special issue of the *Artificial Intelligence* journal appeared, presenting different approaches to nonmonotonic reasoning, in particular Reiter's (1980) default logic and McCarthy's (1980) circumscription approach.

A disappointing fact about nonmonotonic logics appears to be that it is very difficult to formalize a domain such that one gets the intended conclusions. In particular, in the area of reasoning about actions, McDermott (1987) has demonstrated that the straightforward formalization of an easy temporal projection problem (the 'Yale shooting problem') does not lead to the desired consequences. However, it is possible to get around this problem. Once all underlying assumptions are spelled out, this and other problems can be solved (Sandewall 1994).

It took more than a decade before people started to analyze the computational complexity (of the propositional versions) of these logics. As it turned out, these logics are usually somewhat more difficult than ordinary propositional logic (Gottlob 1992). This, however, seems tolerable since we get much more conclusions than in standard propositional logic.

Right at the same time, the tight connection between nonmonotonic logic and belief revision (Gärdenfors 1988) was noticed. Belief revision—modeling the evolution of beliefs over time—is just one way to describe how the set of nonmonotonic consequences evolve over time, which leads to a very tight connection on the formal level for these two forms of nonmonotonicity (Nebel 1991). Again, all these results and insights are mainly based on the logical method to knowledge representation.

3. Outlook

The above description of the use of logics for knowledge representation is necessarily incomplete. For instance, the area of qualitative temporal and spatial reasoning has been left out completely. Nevertheless, one should have got an idea of how logics are used in the area of knowledge representation. As mentioned, it is the idea of providing knowledge representation formalisms with formal (logical) semantics that enables us to communicate their meaning, to analyze their formal properties, to determine their computational complexity, and to devise reasoning algorithms.

While the research area of knowledge representation is dominated by the logical approach, this does not mean that all approaches to knowledge representation must be based on logic. Probabilistic (Pearl 1988) and decision theoretic approaches, for instance, have become very popular lately. Nowadays a number of approaches aim at unifying decision theoretic and logical accounts by introducing a qualitative version of decision theoretic concepts (Benferhat et al. 2000). Other approaches (Boutilier et al. 2000) aim at tightly integrating decision theoretic concepts such as Markov decision processes with logical approaches, for instance. Although this is not pure logic, the two latter approaches demonstrate the generality of the logical method: specify the formal meaning and analyze!

See also: Intelligence: History of the Concept; Knowledge Representation; Knowledge Spaces; Logic and Cognition, Psychology of; Markov Processes for Knowledge Spaces

Bibliography

Allen J A, Fikes R, Sandewall E (eds.) 1991 *Principles of Knowledge Representation and Reasoning: Proceedings of the*

2nd International Conference (*KR-91*). Morgan Kaufmann, Cambridge, MA
Benferhat S, Dubois D, Fargier H, Prade H, Sabbadin R 2000 Decision, nonmonotonic reasoning and possibilistic logic. In: Minker J (ed.) *Logic-based Artificial Intelligence*. Kluwer, Dordrecht, The Netherlands, pp. 333–60
Boutilier C, Reiter R, Soutchanski M, Thrun S 2000 *Decision-theoretic, High-level Agent Programming in the Situation Calculus: Proceedings of the 17th National Conference of the American Association for Artificial Intelligence* (*AAAI-2000*). MIT Press, Austin, TX
Brachman R J 1979 On the epistemological status of semantic networks. In: Findler N V (ed.) *Associative Networks: Representation and Use of Knowledge by Computers*. Academic Press, New York, pp. 3–50
Brachman R J 1990 *The Future of Knowledge Representation: Proceedings of the 8th National Conference of the American Association for Artificial Intelligence* (*AAAI-90*). MIT Press, Boston, pp. 1082–92
Donini F M, Lenzerini M, Nardi D, Nutt W 1991 The complexity of concept languages. In: Allen J A, Fikes R, Sandewall E (eds.) *Principles of Knowledge Representation*. Morgan Kaufman, Cambridge, MA, pp. 151–62
Gabbay D M, Hogger C J, Robinson J A (eds.) 1995 *Handbook of Logic in Artificial Intelligence and Logic Programming*. Oxford University Press, Oxford, UK, vols. 1–5
Gärdenfors P 1988 *Knowledge in Flux—Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, MA
Gottlob G 1992 Complexity results for nonmonotonic logics. *Journal for Logic and Computation* **2**(3): 397–425
McCarthy J 1968 Programs with common sense. In: Minsky M (ed.) *Semantic Information Processing*. MIT Press, Cambridge, MA, pp. 403–18
McCarthy J 1980 Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence* **13**(1–2): 27–39
McCarthy J 2000 Concepts of logical AI. In: Minker J (ed.) *Logic-based AI*. Kluwer, Amsterdam, The Netherlands, pp. 37–58
McDermott D V 1987 A critique of pure reason. *Computational Intelligence* **3**(3): 151–60
Minker J (ed.) 2000 *Logic-based Artificial Intelligence*. Kluwer, Dordrecht, The Netherlands
Minsky M 1975 A framework for representing knowledge. In: Winston P (ed.) *The Psychology of Computer Vision*. McGraw-Hill, New York, pp. 211–77
Nebel B 1991 Belief revision and default reasoning: Syntax-based approaches. In: Allen J A, Fikes R, Sandewall E (eds.) *Principles of Knowledge Representation and Reasoning*. Morgan Kaufmann, Cambridge, MA, pp. 417–28
Nebel B, Smolka G 1990 Representation and reasoning with attributive descriptions. In: Bläsius K-H, Hedtsctück U, Rollinger C-R (eds.) *Sorts and Types in Artificial Intelligence, Vol. 418 of Lecture Notes in Artificial Intelligence*. Springer-Verlag, Berlin, New York, pp. 112–39
Newell A 1982 The knowledge level. *Artificial Intelligence* **18**(1): 87–127
Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco
Reiter R 1980 A logic for default reasoning. *Artificial Intelligence* **13**(1): 81–132
Sandewall E 1994 *Features and Fluents*. Oxford University Press, Oxford, UK
Schild K 1991 *A Corresponding Theory for Terminological Logics: Preliminary Report, Proceedings of the 12th International Joint Conference on Artificial Intelligence* (*IJCAI-91*). Morgan Kaufmann, Sydney, Australia, pp. 466–71

B. Nebel

Lone Mothers in Affluent Nations

Lone mother families are formed by death, divorce, or by a woman having a child outside marriage. Indeed, until the 1970s, there was no single term to describe these families. 'Lone mother' is now the most common term used in English to refer to them, although 'single mother' is still often used. However, this carries specific connotations in respect of marital status, referring as it does only to unmarried mothers, and thus strictly speaking does not cover all types of lone mother family. Much of the international data refers to 'one-parent' or 'lone-parent' families, but only a relatively small percentage (usually below 10 percent) of lone-parent families are headed by men. Countries also define lone mother families differently, for example, in regard to the age under which a child is considered to be dependent, and this makes cross-national comparisons of data difficult.

Lone motherhood is not, of course, new. For example, the UK experienced a sharp increase in unmarried motherhood during the 1840s, at the beginning of the industrial revolution, and at the beginning of the twentieth century, as many families were broken by death as were broken by divorce in the early 1970s. However, widowhood as a pathway into lone motherhood has declined whereas divorce and unmarried motherhood have increased. As lone mother families have become both more numerous and more visible in western countries, so they have also become an issue for public policy (McLanahan and Sandefur 1994), although only in the English-speaking countries have they come to be regarded as something of a social threat as well as a social problem.

1. *The Changing Demography of Lone Motherhood*

There is only a limited amount of statistical information on the proportions of families with dependent children that are lone-parent families. Table 1 shows the position of 12 member states of the European Union at the beginning of the 1980s and the 1990s, together with data for the US. Even within Western Europe there is a good deal of variation, with high proportions of lone parents in Denmark and the UK, intermediate levels in the continental European coun-

Table 1
Lone parent families with at least one child of less than 15 years of age in the EC and 18 years in the USA as percentage of all families with dependent children

Country	1980/1[a]	1990/1[a]
Belgium	9.4	14.6
Denmark	18.1	20.8
Germany	9.8	15.4
Greece	4.4	5.7
Spain	5.4	7.9
France	8.3	10.7
Ireland	7.2	10.8
Italy	7.3	6.4
Luxembourg	9.1	12.2
The Netherlands	7.9	12.2
Portugal	11.6	12.5
United Kingdom	13.7	16.0
USA	19.7 (80)	26.6 (9.2)

Source: Eurostal 1994 Social Europe: the European Union and the Family. Eurostat, Luxembourg; Kamerman A, Kahn A J 1997 Family change and family policies the US. In: Kamerman A, Kahn A J (eds.) *Family Change and Family Policies in Great Britain, Canada, New Zealand and the US*. Clarendon Press, Oxford, p. 323
a Unless otherwise specified.

Table 2
Crude divorce rate: divorces per 1000 average population, 1970–1994

	Year		
Country	1970	1980	1994[a]
Northern Europe			
Denmark	1.9	2.7	2.6
Finland	1.3	2.0	2.7
Iceland	1.2	1.9	1.8
Ireland	–	–	–
Norway	0.9	1.6	2.5
Sweden	1.6	2.4	2.5
United Kingdom	1.1	2.8	3.0
Western Europe			
Austria	1.4	1.8	2.1
Belgium	0.7	1.5	2.2
France	0.8	1.5	1.9
Germany (FDR)	1.3	1.6	2.2
Germany (GDR)	1.6	2.7	1.5
Luxembourg	0.6	1.6	1.7
The Netherlands	0.8	1.8	2.4
Switzerland	1.0	1.7	2.2
Southern Europe			
Greece	0.4	0.7	0.7
Italy	–	0.2	0.5
Portugal	0.1	0.6	1.4
Spain	–	–	0.8
USA	3.5	5.2	4.6 (93)

Source: Council of Europe 1996 *Recent Demographic Developments in Europe*, 1996. Council of Europe Press, Strasbourg; Statistical Abstract of the United States, 1995, Table no. 87
a Unless otherwise specified.

tries, and low levels in Southern Europe with the exception of Portugal.

Tables 2 and 3 show that this pattern of differences within Europe holds good for both divorce and extramarital birth rates. Divorce climbed steadily in most countries from the late 1960s. In the 1990s, it remains high, but the rate has leveled off. Some countries, such as Ireland, have only permitted divorce relatively recently, but the level of informal separation, which is impossible to measure, has probably been quite high. In most European countries the chances of divorce stand at about one in three, with Sweden, Denmark, and the UK leading the league table at around two out of five. However, the chances of divorce in the USA (and the former USSR) are one in two.

Births outside marriage have shown the largest increase in the US and Northern Europe (including the UK). With few exceptions the rise of extramarital childbearing is directly associated with the rise of cohabitation (Prinz 1995), and it is the breakdown of cohabiting unions that has contributed to the recent increase in never-married lone motherhood. This is because a separated cohabitant will be classified *de jure* as single, however there may in practice be little to distinguish the way these women feel from divorcees. There are further differences between countries, which aggregate statistics such as these do not reveal and which may be important from the point of view of policy making. For example, the English-speaking countries have a disproportionately high proportion of young unmarried mothers. The teenage pregnancy rate in the UK is, in the late 1990s, six times that of The Netherlands.

Finally, it should be pointed out that there may well have been some 're-labeling' in respect of the pathways into lone motherhood (Kiernan et al. 1998). In the 1960s, with the advent of the pill and the sexual revolution, sex was increasingly separated from marriage. Large numbers of unmarried women became premaritally pregnant, but the majority married before the baby was born in what were popularly known in English as 'shot-gun marriages.' From what is known about the characteristics of the divorcing population, it is reasonable to suppose that many of these marriages ended in divorce. At the end of the century, pregnant, unmarried women are more likely to have their babies outside marriage and probably to cohabit, which has increased the numbers of unmarried mothers. In addition, cohabitations are more likely than marriages to breakdown. Lone motherhood is the result of many different demographic 'events,' and over time the relative importance of the pathways into it have changed.

Table 3
The proportion of live births outside marriage per 100 live births, 1960–95

Country	Year				
	1960	1970	1980	1990	1995[a]
Northern Europe					
Denmark	7.8	11.0	33.2	46.4	46.9 (94)
Finland	4.0	5.8	13.1	25.2	31.3 (94)
Iceland	25.3	29.8	39.7	55.2	59.6 (94)
Ireland	1.6	2.7	5.0	14.6	22.7
Norway	3.7	6.9	14.5	38.6	47.6
Sweden	11.3	18.4	39.7	47.0	52.9
United Kingdom	5.2	8.0	11.5	27.9	33.6
Western Europe					
Austria	13.0	12.8	17.8	23.6	27.4 (94)
Belgium	2.1	2.8	4.1	11.6	13.6 (93)
France	6.1	6.8	11.4	30.1	34.9 (93)
Germany (FDR)	6.1	5.5	7.6	10.5	12.4
Germany (GDR)	11.4	13.3	22.8	35.0	41.4
Luxembourg	3.2	4.0	6.0	12.9	13.1
The Netherlands	1.4	2.1	4.1	11.4	14.3
Switzerland	3.8	3.8	4.7	6.1	6.8
Southern Europe					
Greece	1.2	1.1	1.5	2.2	2.9
Italy	2.4	2.2	4.3	6.5	7.7
Portugal	9.5	7.3	9.2	14.7	18.7
Spain	2.3	1.4	3.9	9.6	10.8 (93)
USA	5.3	10.7	18.4	28.0	31.0 (93)

Source: Council of Europe 1996 *Recent Demographic Development in Europe, 1996* Council of Europe Press, Strasbourg; UN, Demographic Yearbook, 1960, 1970, 1980; Eurostat Yearbook, 1997
a Unless otherwise specified.

2. *Causes of the Increase in Lone Motherhood*

The causes of the increase in lone motherhood are hotly disputed and in any case differ over time (Laslett 1977). They can only be understood in relation to wider changes in the marriage system. In the case of the rise in the extramarital birth rate in the UK in the 1840s, Shorter (1975) argued that the main reason lay in women's increased independence due to the new possibilities of wage-work in factories. However, others have argued convincingly that women in the traditional occupation of domestic service were particularly vulnerable to extramarital pregnancy. Intercourse seems not to have taken place until marriage was promised, and unmarried motherhood happened when the marriage failed to take place, often because of employment difficulties on the intended husband's part. Thus in this interpretation, traditional practices fell foul of new economic mobility and uncertainty, just as the rise in unmarried motherhood during the Second World War may in large part be attributed to war-time disruption and thwarted marriage.

Debates about the causes of lone motherhood today continue to revolve around the degree to which they can be attributed to women's increased autonomy, in terms of both the exercise of control over fertility and increased economic independence. Both demographers and economists have tended to pay most attention to this variable.

In the postwar period women have acquired the capacity to live alone with a child. This is a very recent phenomenon; until the last quarter of the twentieth century the majority of unmarried mothers found refuge with kin and in countries with little by way of welfare benefits or employment for women they must continue to resort to kin or to charity. A majority of social scientists agree that in western countries the growth in women's employment is more important than the increased availability of state welfare benefits in providing lone mothers with the income that in large measure explains the growth in their numbers. Only in the US has a vigorous case for welfare benefits as the cause been made (especially by Murray 1984), although the 'welfare dependency' of lone mothers has also been deplored in the UK from the late 1980s.

The importance attached to women's earning capacity as an explanatory variable is reinforced by influential economic theories of marriage. The new

'home economics' Becker (1981) and other neoclassical economists has argued that marriage should be understood in terms of the way in which it permits each partner to maximise their welfare. In this view, for biological and social reasons, women are deemed to be better at rearing children than are men. Marriage is seen as involving an efficient specialization of roles, with men acting as breadwinners and women as housewives and carers. But rising female employment and an increase in women's wages means that the parts played by men and women in marriage are less specialized and that there is less incentive for people to be married. While there are many academics who do not subscribe to such a theory of marriage and the family, the idea that women's economic independence has an effect on marital behavior is widespread, possibly because people with very different politics can buy into it.

However, any monocausal explanation of what is an extremely complex phenomenon is unlikely to be satisfactory. The nature of the choices and constraints faced by women who become lone mothers is complicated and involve the position of men. In the first place, it is significant that men's capacity to act as the main earner has been conspicuously eroded. Wilson (1987) was among the first to suggest that more attention should be paid to the deteriorating economic position of black men in the United States. In addition, in the UK the proportion of income contributed to married couple families by men has fallen sharply, eroding their historical position as breadwinners (Harkness et al. 1966). Studies of young unmarried mothers have shown that the putative fathers are all too often unemployed and that there are few economic reasons for the women to marry them. In the case of women seeking divorce, research has shown that women who have suffered violent treatment in their marriage (which some estimate to be the case in 20–25 percent of divorces) are also likely to have been mistreated economically. In Western Europe, these women may find that they are better off living on state benefits.

Men have also been charged with behaving irresponsibly in no longer being prepared to marry the women they get pregnant, and opportunistically in being willing to divorce their wives in favour of—typically—a younger woman. However, for some the increase in female economic autonomy remains the most important cause, because it permits men to behave irresponsibly (Fukuyama 1999). Others stress not so much male irresponsibility but inflexibility. Thus, Hochschild (1989) has documented the tension that results from men's failure to increase their contribution to household work in face of their wives' increased participation in the labor market. This points not to the fact of women's increased employment *per se* as a cause of divorce, but rather to the difficulties couples have in negotiating their domestic roles and, in particular, the problems men have in responding to and coping with change. Burns and Scott (1994) have referred to the increasing 'decomplementarity' between men and women, preferring to ground their explanations in changing structures rather than in individual decision-making, selfish or otherwise. Male inflexibility has undoubtedly exacerbated tensions in modern marriages and may be an important contributing factor to the high rate of divorce.

Research has shown more appreciation of the importance of cultural variables. Mason and Jensen (1995) have pointed out that the microtheory of the neoclassical economists does not recognize or allow for collectively generated or agreed upon norms and sanctions. Yet the meaning that is attached by the majority to marriage may well have changed. For example, women's increased labor market participation may lead the more successful among them to prioritise sexual satisfaction and emotional companionship over men's capacity to provide, and to a greater readiness to divorce if these expectations are not met.

At the broadest level, sociologists have grappled with the changing meaning of personal relationships, identifying a shift towards autonomy and individualism. Giddens (1992) has argued that we have seen the emergence of the 'pure relationship,' a social relationship entered into for its own sake and for what can be derived by each person from it in terms of material, but more usually, emotional exchange. Love has thus become contingent rather than 'forever,' and cohabitation rather than marriage is the logical concomitant. Beck and Beck-Gernsheim (1995) have also pointed out that it is no longer possible to say what family, marriage, parenthood, sexuality, and love mean. All vary from individual to individual and from relationship to relationship. The twin prescriptive norms derived from the male breadwinner model on the one hand, and from an externally imposed moral code (enforced through legislation such as fault-based divorce) on the other have both been eroded. In the vacuum that is left people are left to negotiate their own solutions, which does not necessarily mean that they do so selfishly or without due regard for moral responsibility, but the result will inevitably be more diversity in terms of family structures.

3. *The Problem of Lone Motherhood for Policy*

Lone motherhood has always tended to be defined as a social problem. This is because social policies have worked historically on the assumption that families will broadly adhere to a 'male breadwinner model,' whereby men earn and women care for children (and other dependent adults). The male breadwinner model family is an ideal type. It has never actually existed for all families in any state. Women have always been economically active to different degrees in different

countries and in different socioeconomic classes; however, their earnings are still substantially less than men's because they work fewer hours, and often in female ghetto occupations which pay low wages. Thus, the majority of women are still economically dependent even though their earnings are increasingly crucial to the operation of the family economy. In the context of assumptions regarding the traditional two parent family, the problem of lone mothers becomes essentially that of women-with-children-and -without-men. In the absence of a male breadwinner, the state has to decide to what extent and under what conditions it will step in to fill the place of the missing husband; how far it will exert pressure on an able-bodied absent father to support his child and possibly its mother; and how far it will expect the lone mother to support the child by herself becoming a breadwinner.

In other words, the social problem of lone motherhood has been defined primarily in terms of the costs of financial support. But the solutions favored have changed over time. In early nineteenth-century Britain, the state was unwilling to countenance any claim by an unmarried mother for financial support from a putative father, believing that it would only encourage female immorality. This policy did not last long, but expectations regarding the maintenance of women and children by men remained low until the last decade of the twentieth century in Britain. In contrast, the Swedish authorities made strenuous efforts to enforce the father's obligation to maintain from 1917.

Attitudes towards the behavior expected of lone mothers themselves have undergone more changes still. There have been pendulum swings in their treatment. Broadly speaking in the nineteenth century it was assumed that they should be in employment, working to support as many children (usually one or two) as possible. Where a poor law system operated, as in the UK and in some American states, public relief might be offered in respect of the other children. In other words, in the absence of a male breadwinner, the mother should herself become a worker. By the mid-twentieth century, the policy assumption was rather that she should be a mother and stay at home with her dependent children, drawing a cash benefit from the state. However, the length of time she was deemed to be a mother varied considerably. In the postwar period in The Netherlands and the UK lone mothers were not required to register for work so long as they had a child under age 16. In France, the same rule applied only to mothers with children under age 3. At the end of the twentieth century the pendulum swung again. In the mid-1990s governments in the USA, The Netherlands, and the UK all decided that lone mothers should obtain a greater proportion of their income from earnings, although only the US took the decision also to phase out state benefits (Lewis 1997).

Policy makers have been reluctant to treat all lone mothers the same. Following the logic of the male breadwinner model, widows have long been regarded as inherently more 'deserving' than either separated/deserted or divorced wives, or unmarried mothers. In countries with social insurance provision for those of working age, widows usually have qualified for insurance-based benefits by virtue of their position as wives. The treatment of separated or divorced women historically has been harsher, although once fault-based divorce legislation was abandoned, which happened in the vast majority of western countries during the late 1960s and early 1970s, it became impossible to distinguish between 'guilty' and 'innocent' wives in the settlements made by private law. As lone motherhood, whether by divorce or extramarital childbirth became more common in the late twentieth century, so the stigma attaching to it diminished, and lone mothers became much more ready to claim the social assistance benefits that were available to them in most western countries. This was the case even in the USA, which until 1996 provided for lone mother families via the 'aid to families with dependent children' program (Gordon 1994).

Lone motherhood has served as a focus for all sorts of social concerns. In the US, it seems to have wrapped up fears about the socialization of children, in terms of low levels of educational attainment and rising levels of criminality; anxieties about the role of women, particularly in terms of whether the mothers of young children should go out to work and suspicion of greater female sexual autonomy; concern about the role of men, who appear to be unwilling to shoulder financial responsibility for their children; and anxieties about race. All except the last have also been powerfully expressed in Britain.

However, in other European countries, where the proportion of lone mother families is as high or higher, lone mothers have not become problematized in the same way. The difference is that the difficulties experienced by lone mothers and their children have not been attributed to their status. Thus in Denmark and Sweden, for example, debate has centered on the position of one-earner as opposed to two-earner families. These countries drew virtually all adult women into the labor force from the 1970s onwards, such that the female labor participation rate is almost the same as that of men. Thus, in these countries a choice has not had to be made between defining lone mothers as mothers or workers. All adults share a common worker/citizen status. In this context, one-earner families, whether headed by a lone mother or a male breadwinner (of which there are very few) are recognized as likely to have financial problems. Furthermore, it is recognized that the vast majority of adult women face difficulties in reconciling family and employment responsibilities, but that the problems of lone mothers will be especially acute. The adult worker/citizen model is thus supported by the extensive collective provision of childcare, which is absent in the English-speaking countries, even now

that the pendulum has swung once more towards treating lone mothers as workers. Continental European governments have not subscribed to the view that reconciling paid work and the unpaid work of care is a private responsibility (Hantrais and Letablier 1996). In addition, they provide more generous financial support to all families with children, which is reflected in the much lower poverty rates among lone mothers in these countries than in the US, where they are the highest in the western world.

4. Future Prospects

One of the most striking aspects of the growth in inequalities in the English-speaking countries has been the increase in child poverty. In the UK the number of children living in households with below 50 percent average income trebled between 1979 and 1990. Given that patterns of family change are unlikely to be reversed, the challenge for policy is clear.

See also: Dissolution of Family in Western Nations: Cultural Concerns; Divorce and Gender; Divorce, Sociology of; Family and Gender; Family as Institution; Family Theory and the Realities of Childbearing Behavior; Family Theory: Economics of Marriage and Divorce; Family Theory: Feminist–Economist Critique; Fertility of Single and Cohabiting Women; Lone Mothers in Nations of the South; Partnership Formation and Dissolution in Western Societies; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Repartnering and Stepchildren; Teenage Fertility;

Bibliography

Beck U, Beck-Gernscheim E 1995 *The Normal Chaos of Love*. Polity Press, Cambridge, UK

Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Burns A, Scott C 1994 *Mother-Headed Families and Why They Have Increased*. L. Erlbaum Associates, Hillsdale, NJ

Fukuyama F 1999 *The Great Disruption. Human Nature and the Reconsitution of Social Order*. Free Press, New York

Giddens A 1992 *The Transformation of Intimacy. Sexuality, Love and Eroticism in Modern Societies*. Stanford University Press, Stanford, CA

Gordon L 1994 *Pitied but not Entitled. Single Mothers and the History of Welfare, 1890–1935*. Free Press, New York

Hantrais L, Letablier M-T 1996 *Families and Family Policies in Europe*. Longman, London

Harkness S, Machin S, Waldfogel J 1966 Women's pay and family incomes in Britain, 1979–1991. In: Hills J (ed.) *New Inequalities: The Changing Distribution of Income and Wealth in the United Kingdom*. Cambridge University Press, Cambridge, UK

Hochschild A R 1989 *The Second Shift. Working Parents and the Revolution at Home*. Viking, New York

Kiernan K, Land H, Lewis J 1998 *Lone Motherhood in Twentieth Century Britain*. Oxford University Press, Oxford, UK

Laslett P 1977 *Family Life and Illicit Love in Earlier Generations*. Cambridge University Press, Cambridge, UK

Lewis J (ed.) 1997*Lone Mothers in European Welfare Regimes*. Kingsley, London

Mason K O, Jensen A-M 1995 (ed.). *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK

McLanahan S, Sandefur G 1994 *Growing up with a Single Parent*. Harvard University Press, Cambridge, MA

Murray C 1984 *Losing Ground: American Social Policy, 1950–1980*. Basic Books, New York

Prinz C 1995 *Cohabiting, Married or Single: Portraying, Analyzing and Modeling New Living Arrangements in the Changing Societies of Europe*. Avebury, Aldershot, UK

Shorter E 1975 *The Making of the Modern Family*. Basic Books, New York

Wilson W J 1987 *The Truly Disadvantaged*. Chicago University Press, Chicago

J. Lewis

Lone Mothers in Nations of the South

As in the Northern Hemisphere, women in the poor regions of the 'South' (comprising Africa, Asia, Latin America, and the Caribbean) become lone mothers in various ways, notably through widowhood, divorce, separation, and out-of-wedlock birth (see *Lone Mothers in Affluent Nations*). Aside from this diversity of routes into lone motherhood, it is difficult to draw generalizations about the phenomenon across an area of the world containing myriad cultures and around four-fifths of the global population. Women's identities and circumstances as lone mothers are variegated not only on account of their insertion in different kinship systems, as well as religious, political, and economic contexts, but by individual factors such as age and dependency of children, marital status, and residential organization. Indeed, in several countries of the South, the issue of whether women regard themselves, or are regarded by others, as lone mothers is contingent upon whether or not they head their own households.

Although both 'household' and 'headship' are problematic, contested, and far from universal constructs (Bruce and Lloyd 1992), most official and international statistical sources define households as female-headed when the senior woman in a dwelling and/or consumption unit lacks a co-resident legal or common-law husband. In some cases this also extends to include the physical presence of another adult male such as a father, brother, or grown-up son (United Nations [UN] 1991). Although the senior woman in question is often a lone mother, she may also be childless, or at a stage where her children have left home. Recognizing that female household headship is

a generic category in which lone mothers are only one group, it is also important to note that not all lone mothers head their own households but may live in the homes of others (see below).

Acknowledging these distinctions, households headed by women in general, and lone mothers in particular, may be severely disadvantaged on three main counts. First, they often have higher children to adult ratios than their male-headed counterparts. Second, women routinely face gender discrimination in the labor market and earn less than men. Third, female heads are time- and resource-constrained by their triple burden of employment, housework, and childcare. Notwithstanding that some money may enter female-headed units through remittances from migrant offspring, relatives, and so on, transfer income from other sources may be limited. In most Southern nations there is only weak provision for, and enforcement of, child maintenance payments from absent fathers (especially in cases of nonformalized unions). Added to this, there is little support for lone mothers in the form of social welfare or other state assistance. As such, it is not surprising that 'female household headship' has received more attention than lone motherhood *per se* in the academic and policy literature on the South, not to mention in official statistics.

1. Data on Lone Motherhood and Female Household Headship

Data on lone motherhood in nations of the South are rarely available in their own right and have to be picked up 'by proxy' from data on female household headship. This is problematic because, as noted above, female household headship and lone motherhood are not necessarily synonymous. Moreover, even if figures on female-headed households disclose the marital and fertility status of women, thereby indicating the proportion represented by lone mothers, they do not generally reveal lone mothers who constitute 'embedded units' within larger (often male-headed) households. Young unmarried mothers and their children in the Philippines, for example, frequently form subunits in the households of their parents or senior kin as a means of securing financial support or to make their socially disapproved status less visible (Chant and McIlwaine 1995). This type of 'nesting' arrangement also occurs in countries such as Bangladesh, where widowhood is the predominant route into lone motherhood. Here older women and their younger children are often absorbed into the households of adult sons because they have no independent means of survival and/or because headship of households is culturally and symbolically equated with men (Lewis 1993). In Chile too, one in every five households is estimated to contain a 'sub-family,' and over half of these comprise single mothers with children (Buvinic and Gupta 1997).

In addition to the underestimation arising from 'invisibility,' the difficulties of using data on female household headship to ascertain numbers of lone mothers are compounded by the fact that definitions of female headship are often vague or inexplicit, that they vary across countries, and that they are frequently inaccurate (Chant 1997). Further problems arise from the fact that some female-headed households may be formed through the migration of partners/fathers who, while technically absent from the household, provide support for spouses and children through remittances (see below).

2. Routes to Lone Motherhood

Although it is difficult to generalize about routes into lone motherhood at a national, let alone regional, scale, in areas of the South where there is strong social emphasis on formal marriage, such as in Asia, the Middle East, and North Africa, levels of lone motherhood are low. Here the majority of women enter this state through widowhood, although desertion by men has become more common in places such as southern India and Bangladesh as rates of international male labor migration have risen (Baden and Milward 1995). In the Caribbean, by contrast, where consanguineal links among women, children, and female kin have historically been stronger than marital bonds, out-of-wedlock birth is the principal pathway to lone motherhood. Unmarried motherhood also occurs in Latin America, although separation and divorce are usually more significant. In both these latter regions, lone motherhood is more frequent.

In many instances, high levels of lone motherhood are reflected in high levels of female household headship. This relates not only to the larger pool of potential female heads in the population, but also to the fact that where there are fewer restrictions on women's sexual activity outside marriage, it may be more socially acceptable for women to reside alone with their children. Having said this, it cannot be assumed that lone mothers will necessarily constitute the primary subgroup of female household heads. In many countries of sub-Saharan Africa, for instance, relatively high levels of female household headship are associated not with large numbers of lone mothers, but with male-selective labor migration and/or war, and civil conflict.

Data for the 1980s suggest that female-headed households constitute the highest proportion of households in the Caribbean (17.7 percent), followed by Latin America (14.5 percent), with sub-Saharan Africa intermediate (13.6 percent), and South Asia lowest (5.7 percent) (Folbre 1991). Table 1 shows levels of female headship for selected countries within

Table 1
Levels of female household headship, selected developing countries by region, 1990

Female-headed households as a percentage of all households	
Latin America and the Caribbean	
Dominican Republic	25.0
Costa Rica	20.0
Uruguay	23.0
Chile	21.0
Venezuela	21.3
Panama	22.3
Colombia	22.7
Brazil	20.1
Paraguay	17.0
Nicaragua	24.3
Guatemala	16.9
El Salvador	26.6
Honduras	20.4
Sub-Saharan Africa	
Botswana	45.9
Cameroon	18.5
Burkina Faso	9.7
Zimbabwe	32.6
Swaziland	40.3
Ghana	32.2
Kenya	22.0
Zambia	16.2
Togo	26.4
Tanzania	18.6
Central African Republic	18.7
Uganda	20.6
Liberia	19.1
Burundi	24.7
Middle East and North Africa	
Algeria	11.0
Morocco	17.3
Iraq	12.0
Southeast Asia	
Republic of Korea	15.7
Indonesia	13.0
Philippines	11.3

Source: United Nations Development Programme (UNDP) (1995, pp. 63–5, Annex Table A2.5)
Note: Countries within each region listed in descending order of score on UNDP's Human Development Index (HDI)

the regions of the South in the 1990s, indicating that there is often as much intra- as inter-regional diversity.

3. *Reasons for the Growing Incidence of Lone Motherhood and Female Household Headship*

As in Europe and North America, lone motherhood is by no means a new phenomenon in developing regions. Nonetheless, it is generally agreed that levels have risen throughout the South from the mid-twentieth century onwards. Again, this has tended to be read

Table 2
Recent trends in female household headship, selected developing countries

Female-headed households as a percentage of all households	1980	1990
Costa Rica	17.6	20.0
Republic of Korea	14.7	15.7
Uruguay	21.0	23.0
Chile	21.6	21.0
Venezuela	21.8	21.3
Panama	21.5	22.3
Brazil	14.4	20.1
Paraguay	18.1	17.0
Botswana	45.2	45.9
Dominican Republic	21.7	25.0
Indonesia	14.2	13.0
Cameroon	13.8	18.5
Burkina Faso	5.1	9.7

Source: United Nations Development Programme (UNDP) (1995, pp. 63–5, Annex Table A2.5)
Notes: Countries listed in descending order of score on UNDP's Human Development Index (HDI). Countries included for which data are available for 1980 and 1990

from figures on female-headed households, which, as noted earlier, are problematic. Additional complexities when considering long-term trends are that census bureaus in many developing societies have only recently begun to register gender breakdowns of household headship, and have often been inconsistent in respect of terminology and definitional criteria. Notwithstanding these caveats, for the few countries where national records extend back to the 1960s and 1970s, there appear to have been appreciable increases over past decades. In Ghana, for example, women-headed households grew from 22 to 32.2 percent of the household population between 1960 and 1990, and in Brazil the corresponding proportions were 5.2 percent and 20.1 percent. Even in countries such as the Philippines, where the incidence of female household headship is low, the percentage rose from 9.8 to 11.3 over the same period. More recent data for a variety of countries point to a general maintenance of upward trends between 1980 and 1990 (see Table 2).

Specific reasons for increases in female household headship in the South in the last 30 to 40 years vary in accordance with the particular histories of individual countries, although there are some factors of general significance. One is the way in which processes of postwar capitalist development and 'modernization' have undermined traditional economic and kinship structures, most of which were patriarchal and allowed women limited access to income and resources in their own right. Coupled with the growing demand for female labor brought about by economic globalization and the spread of multinational factories specializing

in the assembly of garments, electronics, and other light industrial goods, waged work for women has been seen as a central factor in women's mounting propensity to head households (Safa 1999). This trend has been exacerbated by declining employment opportunities for men, especially those at the lower end of the occupational hierarchy. Inability to fulfill social expectations to be family breadwinners appears to be making men in some countries less likely to marry, and is also associated with rising rates of conjugal breakdown (González de la Rocha 1995, Moore 1994). Other aspects of economic change linked with female headship include the effects of structural adjustment packages (SAPs) implemented in a range of countries in the South in the 1980s and 1990s at the behest of the International Monetary Fund and the World Bank. Although SAPs aim to restore financial equilibrium in countries with heavy debt burdens, and to promote renewed economic growth, in the short run they have been associated with extreme hardship, especially for poorer groups in society. Typical elements of these programs have included cutbacks in public services, the reduction or removal of government subsidies on basic foodstuffs, and wage freezes. These have tended to increase poverty, which, in turn, has placed greater strain on family relationships, as well as diminishing the prospects for women's (re)incorporation in the households of parents and male relatives following widowhood, divorce, or abandonment. Demographic factors associated with rising female household headship include increased geographical mobility and gender-selective migration. In sub-Saharan Africa, for example, men predominate in most rural–urban migration flows leaving women heading households in the countryside. Declining fertility is another demographic trend with implications for rising rates of female headship. On the one hand, it reduces the child care burdens of lone mothers, thereby increasing their ability to take up paid work. On the other hand, reduced numbers of children mean less likelihood that older lone women will be taken into the households of their offspring in later life.

Last but not least, increases in women's rights in the spheres of sexual and reproductive freedom, access to divorce, and custody of children have also increased female headship. These factors, coupled with changing labor-market trends and poverty, are the most relevant when considering rises in lone motherhood *per se*. Marriage in many societies is not the prerequisite for childbirth that it once was, and divorce rates have increased as the grounds for divorce have widened, as women have become freer to petition for divorce in their own right, and as their entitlements to conjugal property, maintenance payments, and child support have risen. In Costa Rica the number of divorces rose from two per 100 marriages in 1975 to 15.3 in 1991, and in Thailand, a 50 percent rise in divorce during the 1980s meant that by the early 1990s, nearly 25 percent of marriages were ending in divorce (Chant 1997).

4. *Stereotypes and Challenges of Lone Motherhood*

In South and North alike, the married two-parent family is a widespread normative ideal for raising children. Despite the fact that this arrangement provides no guarantee of financial or emotional support, let alone shared parenting (Baylies 1996), most other alternatives are regarded as undesirable. Lone motherhood is wont to be viewed as especially symptomatic of 'family breakdown,' and, as a result, is often marginalized and stigmatized.

The extent to which lone mothers in the South encounter social opprobrium depends on many factors. One is the degree to which women are controlled and/or subordinated in different societies. In strongly patriarchal cultures with marked gender divisions of labor, in which marriage is the major route to social legitimacy and economic survival for women, lone mothers are often shunned in their communities and condemned within political and religious doctrines. Although the most marked disapproval seems to accrue to women who enter lone motherhood through nonmarriage, in some countries, divorced and separated lone mothers are attributed a disproportionate share of the blame for conjugal breakdown. This seems to be particularly the case in South Asia, where even women who have been bereaved can find themselves socially outcast. Symbolically, widows have defied male authority by outliving their husbands, and anxiety surrounds their status as sexually experienced women lacking male guardianship (Buitelaar 1995). By contrast, least stigma seems to attach to lone mothers in areas such as the Caribbean, where historical legacies of slavery and plantation agriculture are argued to have laid the basis for more flexible, informal unions, and substantial social and economic autonomy among women (Ellis 1986). Mothers in the region have also been observed to rely more on their natal kin than on the fathers of their children (Safa 1999). Yet even where there is more societal tolerance of lone mothers, women raising their children alone are commonly regarded in a negative light. One prominent conventional wisdom in government, academic, and media discussions is that rising levels of lone motherhood (and female household headship) are linked with a worldwide 'feminization of poverty,' i.e., that lone-mother households are disproportionately concentrated among low-income and indigent groups, and suffer greater extremes of poverty than male-headed two-parent households (Moghadam 1997). A second and related belief is that lone motherhood gives rise to an 'intergenerational transmission of disadvantage.' Aside from material privation, the life chances of children in one-parent homes are deemed to be severely prejudiced by the psychological problems emanating from 'father absence,' lack of maternal attention, limited parental discipline, and social stigmatization. Influenced by

frameworks such as the 'Culture of Poverty' thesis of the US anthropologist Oscar Lewis, dating from the 1960s, lone motherhood is also regarded as perpetuating a long-term 'vicious circle' of downward mobility and family instability. Children are assumed to have high rates of juvenile delinquency and school drop out, early entry into unskilled segments of the labor market, and premature and fragile sexual unions, which in turn lead to further cohorts of lone-mother families who have few capabilities of breaking out of impoverishment.

Although poverty and intergenerational disadvantage clearly do apply to some lone- mother households in the South, it is also true that these patterns are found in other types of household unit. Moreover, stereotypes about lone motherhood have been criticized as overly crude, generalized, and lacking in substantiation. Feminist analyses in particular have suggested that there is insufficient empirical evidence and that there are too many weaknesses in existing methodologies (for the tracking of family influences on child and youth development, for example) to warrant unilaterally negative conclusions. They also point out that disparaging stereotypical portrayals of lone mothers can be construed as ideologically and politically harmful. To represent a group who is routinely scapegoated for a wide range of social ills as marginal, deviant, and problematic can be even more damaging to the women's public image and self-esteem, and reinforce existing prejudices against them.

A crucial element in calls to arrest the overwhelmingly pathological treatment of lone mothers is the need to recognize their considerable diversity (Baylies 1996). Lone-mother households vary not only in respect of how they come into being, but on several other grounds, such as age and relative dependency of offspring, household composition, socioeconomic status, and access to resources from beyond the household unit (from absent fathers, kinship networks, state assistance, and the like). A common finding from a range of countries, for instance, is that the average age of female heads of household is greater than that of female spouses in male-headed units. As a consequence, they often have children of working age who are able to help out financially or with the care of younger siblings. This may be complemented by incorporating relatives into the home, with lone-mother households showing higher levels of extended composition than their male-headed counterparts. In low-income neighborhoods in urban Mexico, for example, more than one-half of female-headed households are extended, compared with just over one-quarter of male-headed units (Chant 1997).

Detailed comparative research on the economic characteristics of male- and female-headed households has frequently indicated that through the twin strategies of household extension and multiple earning, female-headed units may actually have lower dependency ratios and higher per capita incomes than their male-headed counterparts. Moreover, expenditure within female-headed households is often dedicated to the basic needs of children and is less skewed toward personal consumption by adults. Whereas in two-parent units, considerable portions of the household budget may be spent by male heads on 'nonmerit' goods such as tobacco and alcohol, in lone-mother households, more resources are usually allocated to food and education (particularly that of daughters) (Engle 1995, Hoddinott and Haddad 1991). Contrary to prevailing conventional wisdom, levels of nutrition and educational attainment may thus be higher among children in female-headed units (Chant 1997). Aside from the importance of these issues in deconstructing blanket stereotypes, it is also acknowledged that, in situations where families are affected by male violence or financial neglect, lone motherhood and/or female headship may be important survival strategies in their own right, as well as increasing women's personal power and autonomy (Jackson 1996).

5. *Social Policy and Lone Motherhood*

Comparatively speaking, lone motherhood has not occupied the place in social policy discussions in the South that it has in the North (see *Lone Mothers in Affluent Nations*). This is partly because long-term subsidized programs of social assistance for health, housing, unemployment, and so on are few and far between in Third World countries. Moreover, poverty exists on such a large scale, and public resources have been cut back so severely through debt crises and recession, that the prospects for establishing any comprehensive system of social welfare remains remote. Having said this, a larger share of the spotlight has fallen on female-headed households in the last two decades, as countries undergoing neoliberal economic adjustment have been forced to develop more streamlined programs of poverty alleviation. In the interests of cost savings and 'efficiency,' efforts have been made to reduce public expenditure on universal social programs which benefit all groups in society (albeit with an emphasis on those with lower incomes) in favor of redirecting smaller amounts of resources to target groups of the population. Whether poor households headed by lone mothers should receive priority has become a key question, even if only a handful of countries to date (Chile, Colombia, Bangladesh, India, Honduras, Puerto Rico, and Costa Rica, for example) have actually implemented programs specifically oriented toward female-headed households.

Recognizing the empirical limitations of few test-case evaluations, Mayra Buvinic and Geeta Rao Gupta (1997) have mounted an ambitious review of the potential benefits and drawbacks of targeted assistance. One of their major arguments in favor of targeting is that in societies where there may be

unreliable data on poverty, isolating households headed by women is likely to capture a significant share of the population in need of assistance, especially where there are substantial gaps in male and female earnings, and where subsidized child-care facilities are limited. Another persuasive argument is that targeting assistance to lone mothers may be a very effective means of improving child welfare, given widespread empirical evidence that children fare better where women have resources at their own disposal. Other potential benefits include greater equitability of development resource allocation among men and women (Buvinic and Gupta 1997).

Arguments against targeting include the difficulties inherent in screening processes whereby some female-headed households may not be classified as such due to cultural norms of naming men as heads of household, even if they are largely or permanently absent, or make little contribution to family life and welfare. Many women may not want to be identified as lone mothers given the stigma attached to the status. They may also feel that taking public money will increase antagonism against them. By the same token, female-headed households may become male-headed over time through remarriage or cohabitation, thereby resulting in a leakage of benefits to male-headed households. Another potential slippage of benefits to nonpoor households arises from the fact that not all female-headed households have low incomes, and some may receive support, albeit periodically, from men. Tactics for determining which types of female heads are most in need of help may also be problematic. In Honduras, for example, a food coupon program targeted at primary-school children from female-headed households attempted to ascertain the financial status of mothers through questioning children and neighbors in the community. Aside from the invasive nature of this approach, it did not prevent a number of arbitrary and inappropriate decisions being made (Grosh 1994).

Buvinic and Gupta (1997) also draw attention to the fact that targeting can alienate male-headed households, and thus have high political costs. This is especially likely to be the case when women heads are targeted with interventions that are not perceived as female-specific, such as housing subsidies, food coupons, and so on. Less conflict, alternatively, is likely to occur when female heads receive benefits that are perceived as female-specific, such as skills training for 'female' jobs, or child and maternal health interventions.

Finally, targeting may also produce so-called 'perverse incentives' insofar as programs of support for female household heads may lead to more households adopting this form, which no government in the South (not to mention elsewhere in the world) seems to want. In Costa Rica, for example, the Social Welfare Ministry established its first program for female household heads in 1997. While this offered a relatively minimal temporary stipend for women to enable them to take courses in assertiveness- and skills-training, specific mention was made in the supporting documentation that the government had no intention of encouraging increases in lone motherhood (Chant 1998). Similarly, subsequent programs oriented to adolescent mothers in the country have concentrated not only on helping this group complete their studies, but also on raising awareness among young people generally about the need to prevent early pregnancy and lone motherhood.

An additional consideration in the field of policy is that specialized programs of support might not be necessary where there is substantial public assistance for women and mothers in general. Cuba is a case in point, where although Fidel Castro's revolutionary government has resisted providing special welfare benefits to female heads, policies favoring greater gender equality, higher levels of female labor-force participation, and the provision of support services such as daycare have all made it easier for women to raise children on their own (Safa 1995). A useful counterpoint is presented by Chile, which piloted a Program for Female Heads of Household in 1992–3 that was later extended nationally. Although the program aimed to increase women's access to employment through labor training, access to child care, and so on, success was tempered by the government's failure to address the social and cultural structures underlying gender segregation in the labor market and the perpetuation of poverty among women (Badia 1999).

6. *Lone Motherhood and the Future of 'the Family'*

Given limited programs of state support for mothers in general, and lone mothers in particular, it is unlikely that social policy will play a major role in fuelling increases in lone motherhood in the South in the twenty-first century. However, if recent trends in economic globalization and its accompanying social and demographic corollaries persist, it is entirely conceivable that the widespread ongoing decline in patriarchal family units based on marriage and a 'male breadwinner/female housewife' model will be counterposed by further rises in female-centered arrangements (see also *Lone Mothers in Affluent Nations*). Given the entrenchment of poverty and gender discrimination in many countries of the South, it is also probable that a substantial number of lone-mother households will continue to have to rely on strategies of co-residence and/or cooperation with relatives in order to compensate for the absence of male partners.

See also: Development: Socioeconomic Aspects; Economic Development and Women; Economic Globalization and Gender; Families and Households,

Behavioral Demography of; Families and Households, Formal Demography of; Family and Gender; Family Theory: Competing Perspectives in Social Demography; Gender and Feminist Studies; Household in Anthropology; Lone Mothers in Affluent Nations; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Poverty and Gender in Developing Nations

Bibliography

Baden S, Milward K 1995 *Gender and Poverty Bridge Report No. 30*. Institute of Development Studies, Sussex, UK

Badia M 1999 *The Chilean 'Social Integration' Approach to Poverty Alleviation: The Case of the Programme for Female Heads of Households*. Employment Studies paper 25. University of Hertfordshire Business School, Hertford, UK

Baylies C 1996 Diversity in patterns of parenting and household formation. In: Bortolaia-Silva E (ed.) *Good Enough Mothering? Feminist Perspectives on Lone Motherhood*. Routledge, London and New York, pp. 76–96

Bruce J, Lloyd C 1992 *Finding the Ties that Bind: Beyond Headship and the Household*. Population Council/ International Center for Research on Women, New York

Buitelaar M 1995 Widows' worlds: Representations and realities. In: Bremmer J, Van den Bosch L (eds.) *Between Poverty and the Pyre: Moments in the History of Widowhood*. Routledge, London, pp. 1–18

Buvinic M, Gupta G R 1997 Female-headed households and female-maintained families: Are they worth targetting to reduce poverty in developing countries? *Economic Development and Cultural Change* **45**(2): 259–80

Chant S 1997 *Women-headed Households: Diversity and Dynamics in the Developing World*. Macmillan, Basingstoke, UK

Chant S 1998 Women-headed households: Global orthodoxies and grassroots realities. In: Afshar H, Barrientos S (eds.) *Women, Globalization and Fragmentation in the Developing World*. St. Martin's Press, New York, pp. 91–130

Chant S, McIlwaine C 1995 *Women of a Lesser Cost: Female Labour, Foreign Exchange and Philippine Development*. Pluto, London

Ellis P (ed.) 1986 *Women of the Caribbean*. Zed, London

Engle P 1995 Father's money, mother's money and parental commitment: Guatemala and Nicaragua. In: Blumberg R L, Rakowski C, Tinker I, Monteón M (eds.) *Engendering Wealth and Well-Being: Empowerment for Global Change*. Westview, Boulder, CO, pp. 155–79

Folbre N 1991 Women on their own: Global patterns of female headship. In: Gallin R, Ferguson A (eds.) *The Women and International Development Annual*. Westview, Boulder, CO, Vol. 2, pp. 69–126

González de la Rocha M 1995 Social restructuring in two Mexican cities: An analysis of domestic groups in Guadalajara and Monterrey. *European Journal of Development Research* **7**(2): 389–406

Grosh M E 1994 *Administering Targeted Social Programs in Latin America: From Platitudes to Practice*. World Bank, Washington, DC

Hoddinott J, Haddad L 1991 *Household Expenditures, Child Anthropomorphic Status and the Intra-Household Division of Income: Evidence from the Côte d'Ivoire*. Unit for the Study of African Economics, University of Oxford, Oxford, UK

Jackson C 1996 Rescuing gender from the poverty trap. *World Development* **24**(3): 489–504

Lewis D 1993 Going it alone: Female-headed households, rights and resources in rural Bangladesh. *European Journal of Development Research* **5**: 23–42

Moghadam V 1997 *The Feminisation of Poverty: Notes on a Concept and Trends*. Ilinois State University Press, IL

Moore H 1994 *Is There a Crisis in the Family?* World Summit for Social Development. United Nations Research Institute for Social Development, Geneva, Switzerland

Safa H I 1995 *The Myth of the Male Breadwinner: Women and Industrialisation in the Caribbean*. Westview, Boulder, CO

Safa H I 1999 *Women Coping with Crisis: Social Consequences of Export-led Industrialisation in the Dominican Republic*. North-South Agenda Paper no. 36, North-South Center, University of Miami, FL

United Nations 1991 *The World's Women 1970–1990: Trends and Statistics*. United Nations, New York

United Nations Development Programme 1995 *Human Development Report 1995*. Oxford University Press, Oxford, UK

S. Chant

Longevity, Genetics of

Virtually every aspect of a person's life and well being and, in turn, their survival and longevity is influenced by genes—early development and maturation, metabolic rate and body regulation, resistance and vulnerability to disease, and personality and behavior. Inasmuch as the influence of genes is inextricably intertwined with longevity, it is virtually impossible to consider the broad determinants of longevity without a basic understanding of the underlying genetics. Many of the questions asked by lay persons concerned with longevity are inherently genetic: do long-lived parents have long-lived children? Will it ever be possible to genetically alter the rate of aging? How do genes influence disease susceptibility?

1. Conceptual Framework

1.1 Genetics

The influence of genes on the life span of organisms can be explained at several biological levels including: (a) a phylogenetic level where genes generate growth and form and thus determine the basic body plan, metabolic framework, and the broad details of a species' life history. Thus from a phylogenetic perspective, insects are generally shorter lived than mammals, rodents are usually shorter lived than primates, and chimpanzees are generally shorter lived than humans, *ceteris paribus*; (b) a mechanistic level where genes regulate the basic mechanisms affecting the rate of aging and life span. For example, genes will

influence the life span of individuals through their effects on oxidative damage or replicative senescence in cells; and (c) a disease vulnerability where genes determine the susceptibility of individuals to specific diseases. For example, an individual's risk of contracting breast cancer or Huntington's disease is linked to his or her genes.

1.2 Demographic

Longevity is a shorthand term for a number of different actuarial constructs, the two most common of which are *life expectancy* which is an empirical concept referring to the average age of death in a cohort and *life span* (or *maximum life span*) which is a theoretical concept referring to the highest age attainable by any member of a population. Longevity is the actuarial outcome of individuals subject to the risk of death at each age expressed as *age-specific mortality rate*—the probability that an individual currently alive will die in the next time period (e.g., one year). Longevity can be altered by changing mortality in one of three ways, all of which are under genetic control: (a) shifting the age of maturity—this is one of the most common ways in which fruit fly longevity is increased in population genetics selection experiments; (b) reducing mortality over part of the life course—this can occur through disease prevention or cure and reductions in accidental deaths; and (c) reducing the actuarial rate of aging—this effectively reduces the rate of change (slope) of mortality with age.

2. Genetics of Longevity in Model Systems

Genetic mechanisms that determine longevity and the rate of aging can be grouped into two categories: (a) private—those mechanisms that are idiosyncratic of certain lineages, populations, or species. Most mechanisms of aging associated with mutation accumulation are private mechanisms since the fate of these alleles is largely determined by genetic drift; and (b) public—mechanisms that are generally operational over a large diversity of organisms. By studying aging in different systems it is possible to begin to identify genes that appear in multiple species and thus likely control aging at the most fundamental level. In contrast, those genes that enhance longevity only in the species being studied may point toward genes that are less conserved and affect the mortality of smaller groups of related species.

2.1 Polygenic Approaches

The three main animal models used to study the polygenic (many genes) effects on longevity include:

(a) *Mice*. Studies on mouse strains congenic for the H-2 locus showed significant differences in longevity, that certain H-2 alleles were associated with increased life span, and that genes at the H-2 locus may affect the age of onset of reproductive cyclicity. Quantitative analysis suggests that immune responsiveness can be selected for in mice and that the mode of inheritance is polygenetic with 3–7 independent loci being involved in the response. The genetics of aging has also been studied through the use of senescence-accelerated mice which represents a polygenic progeroid model.

(b) *Drosophila*. Selection studies are polygenic by design—any number of genes which favor long life are retained in laboratory populations subjected to conditions favoring longevity such as allowing only the oldest individuals to breed. Selection studies on *Drosophila* have produced strains whose life spans differ by twofold. Associated with these life span differences include increased resistance to starvation, desiccation, and other environmental stresses and reduced ovary weight early in life.

(c) *Nematodes*. The primary approach for studying polygenic aspects of aging in the free-living soil nematode *Caenorhabditis elegans* has been quantitative trait loci (QTL) mapping. This technique is based on the concept that if one has a dense array of detectable genetic markers spread across each of the chromosomes, it is possible to make appropriate genetic crosses between nematode lines and determine which sets of markers are preferentially associated with the phenotype (i.e., longevity) of interest. The longevity genes thus lie somewhere between and are linked to the known markers. Nematode studies on crosses between two wild-type strains have identified QTLs that are responsible for increased life span located on chromosomes 2, 4, and the X. Estimates suggest that approximately seven QTLs are involved in determining life span of *C. elegans*.

2.2 Mutants

Inasmuch as mutations modify the life span of animals through mechanisms that may be shared by common effects on metabolism and gene expression, the analysis of mutations is the most effective means of identifying genes involved in a biological process because the physiological activity between the mutant and the wild type can be compared. However, one of the problems with this approach in aging research is that it is sometimes difficult to identify mutant strains with altered aging processes.

(a) *Yeast* (*Saccharomyces cerevisiae*). Mutation analysis of aging in yeast has yielded a number of candidate genes including *LAG1* that determines both mean and maximum life span, *LAC1*, a virtual copy of *LAG1* on a different chromosome, and *RAS1* and *RAS2* which have life shortening and life extending

effects, respectively. One of the major distinctions between aging in yeast and aging in other model organisms is that in yeast the scale on which aging is measured in yeast is reproduction (i.e., budding) but in other organisms aging is measured in chronological time.

(b) *Nematodes* (*C. elegans*). Mutational analysis in nematodes has yielded much physiological information on aging and has identified regions of the genome that contain longevity genes including *daf* genes which control a program that results in the formation of a dispersal form, the dauer larva—an alternative waiting stage that worms may enter in response to unfavorable periods (e.g., insufficient food), the gene *age-1* which has been assigned to the *daf* pathway for adult longevity but which increases the resistance to stresses, and *clk-1* which appears to control metabolic adjustments, particularly in response to temperature. Virtually all of the increases in longevity in mutant strains of nematodes have occurred at the post-reproductive ages.

(c) *Drosophila*. The effects of single genes in *Drosophila* that extend longevity have been difficult to find perhaps because of the lack of a waiting stage similar to the nematode dauer or yeast spore stages. However, flies which carry the 'Methuselah gene,' *mth*, displayed a 35 percent increase in life span and exhibited enhanced resistance to stress.

2.3 Transgenics

Transgenics refers to the transfer of genes from one organism to another which, in the context of aging, means transferring into a model organism a gene that extends longevity in another organism. If this so-called 'transgenic' organism then lives longer than the wild type, this then provides the ultimate demonstration that the gene affects the aging process either directly or indirectly. The best example of the use of this concept in aging research is in *Drosophila* using transgenic strains that overexpress several different enzymes and exhibit greater longevity and lower mortality than the wild type. Transgenic strains of both *C. elegans* and mice have been produced but have not yet produced definitive results.

2.4 Gene Expression and Regulation

Although the DNA in all somatic cells is the same and thus contains the same information, each cell expresses only a distinct subset of its genes at any one time. Which genes are expressed differs between cell types and also differs slightly as organisms age. For example, blood cells only expresses the genes needed to produce hemoglobin and other proteins associated with blood; genes capable of producing hair are not expressed in tongue cells. Altered gene expression has been suggested as a central mechanism for causing aging and senescence in which regulatory elements act like on (up-regulated), off (down-regulated), and dimmer switches. Experimental manipulations of life span in both flies and rodents have been shown to cause changes in gene activity including: (a) during temperature-induced shifts in fruit fly life spans, the age schedule for decreased expression of genes in the adult antenna and muscle is shifted in proportion to life span over a threefold range; (b) the potential influence of physiological factors in rat brain aging is demonstrated by the impact of food restriction which slows the age-related increases in transcription; and (c) gene expression is altered in bakers' yeast (*Saccharomyces cerevisiae*) through under- or overexpression of some wild-type genes, a gene that is homologous to the human gene for Werner progeria, and of genes causing a redistribution of telomeric silencing proteins.

2.5 Mitotic Clock

The majority of cell types grown in laboratory cultures have a finite ability to proliferate. After a number of population doublings the cell cultures enter the terminally nondividing state referred to as replicative senescence (so-called Hayflick limit). Many investigations have established a link between aging *in vivo* (live animals) and the proliferative potential of cells in culture. For example, cell cultures derived from one of the longest-lived species of animals, the Galapagos tortoise, doubled up to 130 times, whereas cultures derived from mice with maximum life spans of three years were capable of doubling only 10 or fewer times.

One of the most compelling theories for explaining replicative senescence is that incomplete replication of specialized structures at the ends of chromosomes called telomeres account for the gradual loss of proliferation potential. Telomeres are essential for proper chromosome structure and function including complete replication of the genome—if organisms could not overcome the 'end-replication' problem, they would fail to pass their complete genetic complement from generation to generation. The simplest theory accounting for the Hayflick limit is one in which permanent cell-cycle arrest is due to a checkpoint mechanism that interprets a critically short telomere length as damaged DNA and causes cells to exit the cell cycle. This telomere hypothesis of aging provides a molecular mechanism for counting cell divisions in the normal somatic cell. According to the telomere shortening model of cell senescence, to avoid telomere loss and eventual cell cycle arrest it is necessary for cells to synthesize telomeric DNA by expressing the enzyme telomerase. Therefore, a prediction of this model and one that is borne out in recent

studies is that telomerase activity will be absent in normal somatic cells but present in germline cells and carcinoma cells.

3. *Genetics of Human Longevity*

3.1 *Gerontogenes in Humans*

The term *gerontogenes* is used to describe genes that are involved in aging and usually involves genes whose normal function shortens rather than lengthens life. One of the approaches that researchers use in an attempt to identify genes that influence aging in humans is to study the prevalence of different genes in centenarians relative to younger people (e.g., 20–60 years). One of the most studied genetic polymorphisms involved in human longevity is the apolopoprotein E (APO-E) gene which encodes a protein that is synthesized in the liver, brain, spleen, kidneys, and macrophages and in human populations exists as three common alleles—E2, E3, and E4. Frequencies of these three alleles in centenarians and in noncentenarians revealed that: (a) there was a shortage of the APO-E4 allele among centenarians (5.2 percent) compared to controls (11.2 percent) which suggests a negative effect of this allele on longevity. In fact, individuals who carry the APO-E4 allele are strongly predisposed to developing late-onset Alzheimer's disease; and (b) an excess of APO-E2 allele among centenarians (12.8 percent) relative to the normal controls (6.8 percent) which suggests a positive effect of this allele on longevity.

3.2 *Age-related Genetic Diseases*

3.2.1 Progeroid mutations. Progeroid mutations refer to one of several human genetic syndromes that accelerate some aspects of aging. Although all of the progeroid-like diseases are rare, their effects on aging in young persons are so striking that they have received considerable attention in gerontology and medicine. In this section two of the best known progeroid diseases—Hutchinson–Gilford syndrome and Werner's syndrome—will be described. The outward signs of premature aging are similar in both of these diseases but they differ in their ages of onset and ultimately in the ages at which patients succumb to the disease.

Hutchinson–Gilford (*HG*) *syndrome* is one of the rarest genetic syndromes in humans with only 17 cases reported in the world by 1990. It is caused by a *de novo* gene mutation that is expressed in a dominant manner. At birth HG patients appear normal, but by about one year severe growth retardation is seen followed by rapid loss of subcutaneous fat and hair resulting in a prominent display of the scalp veins. Median age of death in HG patients is 12 years. The most prevalent progeriod disease is *Werner's syndrome* which is an autosomal (i.e., all chromosomes except sex chromosome) recessive disease that is first manifested as a failure to undergo the usual adolescent growth spurt. By the early twenties persons with this disease suffer substantial hair loss, graying of hair, atrophy of the skin, and cataracts as well as an increase in the likelihood of diabetes and cancers. Cytogenetic and molecular genetic studies have suggested that Werner's syndrome is due to a loss in the maintenance of genomic stability due to an unusually high frequency of reciprocal chromosome translocation and of single-gene mutations referred to as a mutator phenotype.

3.2.2 Adult-onset diseases. This section contains a description of three of the best-known adult-onset genetic diseases. The first is *Lou Gehrig's disease* which is formally referred to in the medical literature as amyotrophic lateral sclerosis (ALS). It is a progressive paralytic disorder that strikes middle-aged adults and is usually fatal within five years after diagnosis. Paralysis is due to degeneration of the large motor neurons of the brain and spinal cord. Approximately 10 percent of the ALS cases are inherited as an autosomal dominant disorder while the remaining 90 percent are sporadic. The second of the adult-onset genetic diseases is *Huntington's disease* (HD) which is due to an autosomal dominant gene that typically causes degeneration of the nervous system. HD disease causes memory loss and dementia as well as wrinkled skin, weight loss, and uncontrollable jerky movements. It differs from Alzheimer's disease inasmuch as the average age of onset for this disease occurs at 35–42 years with death occurring within about 15 years after diagnosis. The third and most extensively studied age-related genetic disease in humans is *Alzheimer's disease* (AD). Approximately 2–6 percent of people over 65 show symptoms of AD with the prevalence of symptoms increasing with age—as many as 40 percent of persons over 65 show symptoms of AD. Although the majority of cases of AD appear to be sporadic, several hundred families have been identified in which AD is inherited as an autosomal dominant mode of transmission. Analysis of the familial AD (FAD) has allowed the localization of three distinct genes that are involved in the specification, including: (a) rare mutations in the genes encoding the precursor protein of Alzheimer-associated amyloid on chromosome 21 (2–3 percent); (b) genes on chromosome 14 account for late-onset forms of AD (after age 65); and (c) genes on chromosome 19 are responsible for early-onset forms of AD. One of the major complications of studies of the disease is determining the degree to which FAD is genetic—estimates range from 10 to 100 percent. Individuals homozygous for APOE-ε4 have an al-

most eightfold increase in the probability of getting AD at almost any age when compared with people carrying other alleles of APOE.

3.3 Genealogical Studies

Despite studies demonstrating that genetics account for only 25–35 percent of the total variance in longevity, one of the most enduring and widespread notions in gerontology is the longevity resemblance of kin, particularly between parents and offspring and among siblings. The findings on kinship resemblance of longevity can be grouped into three categories.

(a) *Parent–offspring (familial) resemblance*. Although a modest familial component has also been established in several studies, they are generally plagued by the problem of conflation of shared genetics and environments and so cannot provide conclusive evidence for a genetic component of longevity.

(b) *Twin registries*. This use of twin registries has allowed more quantitative analyses of familial resemblance. Comparing the correlations between monozygotic (genetically identical) twins, fraternal twins, and nontwin siblings begins to disentangle the effects of shared genes from those of shared family environments. Investigations of the relationship between age of death between like-sex fraternal (dizygotic) twins and between identical (monozygotic) twins both reared together and reared apart reveal that longevity is moderately heritable. One twin study of longevity reported that the average absolute difference in age at death between two members of a monozygotic twin pair, two members of a dizygotic twin pair, and two randomly selected same-sex individuals was 14.1, 18.5, and 19.2 years, respectively.

(c) *Comparative ancestry of centenarians and shorter-lived groups*. This comparative approach can elucidate the familial resemblance in extreme longevity. Siblings of centenarians in one study had a four times greater chance of surviving to their early nineties than siblings of a comparison group that died at age 73. Twenty-four percent (13/55) of the immediate ancestors of the oldest person on record, Jeanne Calment from France (died at 122 years 164 days) longer than 80 years, vs. 2 percent in a reference family (1/50). Her immediate ancestors lived an average of 68.2 years vs. 57.7 years for controls in the reference family. An implication of the twin and centenarian ancestry studies is that whereas life span in general is under moderate genetic control, genes play a somewhat larger role for persons attaining the more extreme ages.

4. Evolutionary Biology of Aging

4.1 Theoretical Foundation

Whereas gerontology is concerned with questions concerning the mechanisms of aging, evolutionary biology is concerned with questions in a different biological domain concerned with how senescence evolved. Evolutionary theory of aging postulates that the force of natural selection will always decrease with age with either replacement-level or positive population growth. This provides the conceptual foundation for two population genetic hypotheses of aging. The first is termed *negative pleiotropy* and is based on the concept that alleles that have beneficial effects on one set of components of fitness also have deleterious effects on other components of fitness. The underlying concept is one of tradeoffs—a beneficial effect at young ages may have a deleterious effect at older ages. The declining force of natural selection leads to a tendency for selection to fix alleles that have early beneficial effects, but later deleterious effects. This biases evolution toward the production of vigorous young organisms and decrepit old organisms. Although antagonistic pleiotropy is an important possible mechanism for the evolution of aging, it probably plays a limited role in explaining the persistence of genetic variation in fitness components. The second population genetic hypothesis concerning the evolution of senescence is *mutation accumulation* which arises when the force of natural selection has declined

Table 1
Results of genetic testing in a hypothetical patient in 2010 (from Collins 1999)

Condition	Genes involved[a]	Relative risk	Lifetime risk
Reduced risk			
Prostate cancer	HPC1, HPC2, HPC3	0.4	7
Alzheimer's disease	APOE, FAD3, XAD	0.3	10
Elevated risk			
Coronary artery disease	APOB, CETP	2.5	70
Colon cancer	FCC4, APC	4	23
Lung cancer	NAT2	6	40

a HPC1, HPC2, and HPC3 are genes for hereditary prostate cancer; APOE is gene for apolipoprotein E; FAD3 and XAD are hypothetical genes for familial Alzheimer's dementia; APOB is gene for apolipoprotein B; CETP is gene for choleseryl ester transfer protein; FCC4 is the hypothetical gene for familial colon cancer; APC is the gene for adenomatous polyposis coli; NAT2 is gene for *N*-acetyltransferase 2.

to a point where it has little impact on recurrent deleterious mutations with effects confined to late life.

4.2 Empirical Foundation

A number of empirical studies concerned with understanding the extent to which genes control longevity have been conducted on *Drosophila* in which the ages of reproduction are carefully controlled. A typical design involves the young–old dichotomy of fly strains. Specifically, 'young' lines are created by allowing females to lay eggs at two weeks of age and rearing their offspring whose eggs, in turn, are collected at 14 days. This is repeated for a number of generations. The 'old' lines of flies are created using the same concept except that eggs are collected only from the females who have survived to much older ages. The results of these types of selection studies have revealed that the mean and maximum life spans of cultures selected for early reproduction are lower than those for cultures selected for later reproduction.

5. Longevity Implications of Molecular Medicine

One of the most profound changes in biological research including biomedical and gerontological studies has been the emergence of molecular biology and genomics. These two disciplines are leading to the creation of a universal periodic table of life that will reflect common genetic properties and patterns of ancestral and functional affinities among the genes of both plants and animals. Comparison of related organisms will reveal regulatory regions and key architectural features of proteins that can be used as Rosetta stones for translating and understanding informational pathways and for deciphering biological complexity.

In the very near future, molecular biology and genomics research will have a profound impact on the quantity and quality of human life through disease prevention. One of the best examples of how this emerging science and technology might be applied in the future is provided by the hypothetical example given in a paper by Francis Collins—the Director of the Human Genome Project. The human genome project will provide a baseline of an individual's genome to shed light on the person's risk profile and thus point to prevention strategies if available. The example uses a 23-year-old male college graduate who has smoked for six years and has a strong paternal history of heart disease. This individual agrees to undergo 15 genetic tests that provide risk information (both relative and absolute) for illnesses for which preventive strategies are available. The results of the genetic testing in this hypothetical patient are given in Table 1. The good news is that his risks of contracting prostate cancer and Alzheimer's disease are reduced because he carries low-risk variants of the several genes that contribute to the illnesses. The bad news is the genetic evidence of his increased risks of contracting coronary artery disease, colon cancer, and lung cancer. Confronted with this information, the patient arrives at that 'teachable moment' when a lifelong change in health-related behavior is possible by focusing on reducing specific risks including quitting smoking, reducing his cholesterol level, and having annual colonoscopies starting at age 45.

See also: Aging and Health in Old Age; Aging, Theories of; Differential Aging; Life Expectancy and Adult Mortality in Industrialized Countries; Old Age and Centenarians; Senescence: Genetic Theories

Bibliography

Abbott J H, Abbey H, Bolling D R, Murphy E A 1978 The familial component in longevity—a study of offspring of nonagenarians: III. Intrafamilial studies. *American Journal of Medical Genetics* **2**: 105–20

Bocquet-Appel J-P, Jakobi L 1990 Familial transmission of longevity. *Annals of Human Biology* **17**: 81–95

Chippindale A K, Leroi A M, Kim S B, Rose M R 1993 Phenotypic plasticity and selection in *Drosophila* life-history evolution. I. Nutrition and the cost of reproduction. *Journal of Evolutionary Biology* **6**: 171–93

Collins F S 1999 Shattuck Lecture—Medical and societal consequences of the human genome project. *New England Journal of Medicine* **341**: 28–37

Curtsinger J W, Service P M, Prout T 1994 Antagonistic pleiotropy, reversal of dominance, and genetic polymorphism. *The American Naturalist* **144**: 210–28

Hamilton W D 1966 The moulding of senescence by natural selection. *Journal of Theoretical Biology* **12**: 12–45

Lee C-K, Klopp R G, Weindruch R, Prolla T A 1999 Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**: 1390–3

Lin Y-J, Seroude L, Benzer S 1998 Extended life-span and stress resistance in the *Drosophila* mutant methuselah. *Science* **282**: 943–6

Ljungquist B, Berg S, Lanke J, McClearn G E, Pedersen N L 1998 The effect of genetic factors for longevity: A comparison of identical and fraternal twins in the Swedish twin registry. *Journal of Gerontology* **53A**: M441–6

Masoro E J 1988 Minireview: Food restriction in rodents: An evaluation of its role in the study of aging. *Journal of Gerontology* **43**: B59–64

McGue M, Vaupel J W, Holm N, Harvald B 1993 Longevity is moderately heritable in a sample of Danish twins born 1870–1880. *Journal of Gerontology* **48**: B237–44

Pearl R 1931 Studies on human longevity. IV. The inheritance of longevity. Preliminary report. *Human Biology* **3**: 245–69

Perls T T, Alpert L, Fretts R C 1997 Middle-aged mothers live longer. *Nature* **389**: 133

Perls T T, Bubrick E, Wager C G, Vijg J, Kruglyak L 1998 Siblings of centenarians live longer. *Lancet* **351**: 1560

Ridley M 1999 *Genome*. Harper-Collins Publishers, New York

Rose M R 1991 *The Evolutionary Biology of Aging*. Oxford University Press, New York

J. R. Carey

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Longitudinal Data

Longitudinal data analysis is a subspecialty of statistics in which individual histories—interpreted as sample paths, or realizations, of a stochastic process—are the primary focus of interest. A wide variety of scientific questions can only be addressed by utilizing longitudinal data together with statistical methods that facilitate the detection and characterization of regularities across multiple individual histories. Despite the stated focus on individual histories, most of the analytical strategies either invoke strong assumptions about stochastic dependencies across time and/or use variable-centered methods to study the influence of multiple covariates on process dynamics and its relationship to diverse outcomes. An exception to this pattern is the literature on growth curve analysis, where the full individual histories are modeled and frequently aggregated into categories of similar histories. This set of techniques, as currently constituted, has severe limitations when a study objective is to understand the interrelationships between multiple dimensions of a multivariate time series. Problems arise due to the utilization of interaction terms, as in regression analyses, which represent the imposition of a variable-centered idea in an inherently person-centered question.

The purpose of this entry is to describe some person-centered methods that respond to the challenge of characterizing life histories, ascertained with longitudinal data, placing emphasis on the effective utilization of both cross-time and multiple-life-domain information. We motivate the methods and illustrate their use in the context of explaining how a variety of health outcomes occur. However, the methodological issues are generic and transfer to a wide range of problems in the behavioral and social sciences.

1. *Motivation*

Extensive empirical evidence documents associations between early and mid-life risk factors and later-life chronic disease and disability (Barker and Osmond 1986, Wadsworth and Kuh 1997). These findings generate considerable interest in understanding the subtle nuances of life histories that might explain how such outcomes occur. Broad-gauged interest in successful aging (Rowe and Kahn 1998), with particular emphasis on how some individuals manage to adjudicate effectively multiple difficult challenges, also suggests in-depth comparative studies of life histories. Detailed accounts of whole lives have previously been studied primarily as individual cases, with evidence presented in forms ranging from rich ethnographies to focused interviews and, in a medical context, from open-ended responses to structured inquiries by clinicians.

Complementing such narrative accounts of life histories are structured longitudinal surveys that contain—for each individual—hundreds, or even thousands, of responses to questions about experiences over time and across multiple life domains (Hauser et al. 1993, Power et al. 1991, Wadsworth and Kuh 1997). In principle, those interested in the impact of multiple facets of life histories on a given outcome(s) should be able to construct representations of whole lives from the survey responses. Despite considerable interest in this task across multiple disciplines (Elder 1974, Vaillant 1983), there has been a dearth of analytic strategies for producing representations of whole lives and for aggregating them into meaningful taxonomies that facilitate understanding of how given outcomes occur.

With these observations at hand, our primary objective is to describe top-down and bottom-up methods for constructing taxonomies of life history representations and use them as the basis for explaining how given outcomes occur. We will also emphasize the construction of narratives from numerical survey responses and their utilization as part of an analytical strategy to understand 'whole lives.' We illustrate the ideas and methods in the context of health outcomes. However, the methodological issues are quite general and by no means restricted to this class of outcomes.

2. *Classes of Methods*

The techniques discussed in this entry are called person-centered in the sense that the individual life history is the unit of analysis. Within this broad class of methods are two categories of techniques called 'bottom-up' and 'top-down.' The bottom-up strategies begin with highly idiosyncratic individual histories, and the analytical steps then identify important commonalities and differences across lives, leading to aggregation of histories into relatively homogeneous groups. Optimal matching (Abbott and Hrycak 1990), p-technique integrated with dynamic factor analysis (Shifren et al. 1997), and the integrated qualitative–quantitative aggregation strategy of Singer et al. (1998) are examples of bottom-up strategies.

Top-down techniques sequentially partition a heterogeneous population into progressively more homogeneous sub-groups. Recursive partitioning (Zhang and Singer 1999), Grade-of-Membership models (Berkman et al. 1989), and the clustering methods of Bergman (1998) are generic examples of these methods.

3. *A Bottom-up Example: Understanding Psychological Resilience*

3.1 *The Problem*

Among those who experience major episodes of depression in adult life, or earlier, some manage to

subsequently regain high levels of well-being—that is, they demonstrate resilience. We ask whether those who show recovery have particular features that distinguish them from those who continue to suffer depression. An affirmative answer could suggest prevention strategies which, if applied early in life, might serve to bypass episodes of depression, or minimize their severity and facilitate recovery from them.

3.2 Longitudinal Data

The Wisconsin Longitudinal Study (WLS) began with a random sample of 10,317 men and women who graduated from Wisconsin high schools in 1957. Survey data were collected from the original respondents in 1957, 1975, and 1992/93. Telephone interviews were conducted in 1992/93 with 8,020 of the 1975 respondents, 6,535 of whom also responded to a mail survey. Data have been collected on respondents' family background, starting resources, academic abilities, youthful aspirations, social support, social comparisons, and the timing and sequencing of adult educational and occupational achievements, work events and conditions, family events, and physical and mental health. We focus on a subset of the 5,009 primary respondents who participated in all three waves of the WLS and who responded to questions assessing affective disorder and well-being in the most recent survey. The affective questions were administered to a randomly selected 80 percent of respondents. For each individual, there were responses on over one thousand variables covering all three waves of the survey and information across the multiple life domains indicated above.

The resilience criteria (see Singer et al. 1998) were satisfied by 168 women. There were 59 women who satisfied the depressed/unwell criteria. The analytical task is the development of a taxonomy of life histories for each group such that the characterization of any history is sufficiently nuanced to capture the richness of individual experience, but not so detailed as to represent only very few women. For a more broad-based discussion of resilience, using both narrative and numerical data on individual histories to the study of racial and ethnic inequalities in health, see Singer and Ryff (1997).

3.3 A Person-centered Strategy

Construction of an empirical taxonomy of life histories requires a set of organizing principles—based on prior literature—that constrain the distillation of over one thousand responses for each individual into composite variables. Such composites are the basis for the ultimate representations of categories of histories. The following five principles were used to specify combinations of individual characteristics essential to explain how the outcomes come about.

(a) Adversity and its cumulation over time has negative mental health consequences.

(b) Advantage and its cumulation over time has positive mental health consequences.

(c) Reactions to adversity or advantage influence the impact of life experiences.

(d) Position in social hierarchies has consequences for mental health.

(e) Social relationships influence the impact of life experiences.

These propositions are not new to the social sciences, mental health, or aging fields. What is new is the integration of them into an analytical strategy to guide life-history analysis. The methodology is usefully decomposed into five principle steps:

(1) guided by the organizing principles, life history narratives are constructed for a sub-sample of cases from each outcome group (i.e., resilient and depressed/unwell);

(2) abstracted chronological charts, each one representing the synthesis of several narratives, were constructed;

(3) reduced response vectors, **b**, consisting of composite variables suggested by the chronological charts and individual variables whose rationale derives from the organizing principles are specified;

(4) aggregates (categories) of histories, in terms of response vectors judged to be close to one another, are specified; and

(5) tests of distinguishability are conducted.

Step (5) reduces to ascertaining whether or not the categories of histories characterizing the resilient women occur with low frequency in the depressed/unwell group and, conversely, whether histories characterizing the depressed/unwell women occur with low frequency among the resilient women. If the taxonomies of histories are to be a basis for explanation of the mental health outcomes, there should be distinguishability of categories of histories between the two groups.

3.4 Narratives

A small random sample ($\sim$ 10 cases) is selected from each outcome group, and a narrative biographical story is written about each individual. We adopt Stone's (1979) conception of narrative, defined as 'the organization of material in a chronological sequential order and the focusing of the content into a single coherent story, albeit with subplots.' Our rationale for writing narratives based on survey data (with over one thousand responses per individual) is that the human mind has difficulty processing enormous amounts of information in parallel (i.e., long lists of responses representing a single life), although it is well suited to process a coherent story (i.e., a written narrative about a life) (Turner 1996). The use of narratives here is qualitatively different from their role when they are part of the raw data. Specifically, the construction of

narratives—by the analyst and starting with numerical response data—is part of an analytical strategy. The narratives do not represent empirical data derived from focused interviews, recorded conversations, or text written by the respondents, but stories written by investigators to describe 'whole lives through time.' In generating the narratives, responses selected for inclusion were those expected to impact mental health, based on prior theoretical and empirical literature. The goal is to tell a life story in such a way that omitted information would make little or no difference for understanding the relevant life experiences preceding and predicting midlife mental health outcomes.

3.5 Abstract Chronological Charts

Two life histories can be aggregated and called similar, or even approximately indistinguishable, when the response vectors representing them are similar. At the level of individual cases, represented by raw response data on more than one thousand variables or by narratives, there is so much idiosyncratic detail that meaningful aggregation seems almost hopeless. However, via the organizing principles, it is possible to start with three or four narratives and begin specifying composite variables that ultimately lead to the specification of response vectors, some of which are sufficiently similar that they can be aggregated into a common category.

The composite variables are the initial basis for revised response vectors for all individuals that need to satisfy two criteria. First, the revised response vector for a given individual should provide for a nuanced summary of the person's life history as it relates to mental health outcomes. Second, the response vector should not be too idiosyncratic. The latter condition is essential if we are to aggregate life histories—represented as reduced response vectors—so that individuals with 'similar' response vectors can be grouped into common categories of histories.

3.6 Reduced Response Vectors

Guided by the abstract chronological charts, we specify a provisional reduced vector of variables, whose values for a single individual define the person's reduced response vector. The provisional reduced vector of variables will, clearly, represent all persons for whom we have constructed narratives. However, this is only a small sub-sample of a much larger population. We augment this group with 10–15 new cases for whom we do not have narratives. Their provisional reduced response vectors augmented by responses on additional variables that are clearly linked to the organizing principles are displayed. Visual examination of these augmented response vectors for the new cases AND those for which there are narratives serves to suggest new composite variables and revisions of those specified for the provisional reduced vector of variables. The result of this visual inspection and judgment is a new second-stage reduced vector of variables. We continue bringing in blocks of 10–15 new cases at a time, repeating the above step, and thereby iteratively refining the reduced vector of response variables until all cases for a given mental health outcome group have been examined. The final reduced vector of response variables, **b**, is the basis for initiating the formal specification of a taxonomy of histories.

3.7 Aggregation of Reduced Response Vectors: Pathway Specification

All variables are identified by binary codes in **b**. Next, we identify the longest AND statement common to at least 20 individuals. This identifies a set of response vectors—equivalently, a set of life histories—that exactly match one another on the components of the AND statement. If the components of the AND statement are distributed across the whole lives and they are interpretable as meaningful pathways to a given mental health outcome, then we have succeeded in aggregating a set of life histories into a well-defined group. For the resilient women in WLS, the binary coded version of **b** had 17 components, and the longest AND statement common to at least 20 individuals had seven components.

With an initial aggregate of histories at hand, we eliminate this group from further consideration and identify the longest possible AND statement characterizing at least 20—and hopefully more—of the remaining men. In the WLS resilient women, this did not lead to a long enough AND statement that was sufficiently nuanced to characterize possible routes into episodes of depression, as well as the cumulation of advantage that would promote recovery. This led to specification of more complex Boolean statements that were both consistent with our organizing principles and represented the life histories of a substantial number of women. At least two strategies can be used for this specification. First, visual inspection of the 17-component response vectors revealed a group of 48 women, all of whom grew up in a household with at least one alcoholic parent AND who had a diversity of different forms of compensating advantage in adulthood (represented by logical OR statements). For a narrative description of the resulting Boolean statements, see Singer et al. (1998).

A more formally mechanized approach to this step—using, for example, Ragin's QCA software (1987)—would be to minimize the Boolean expression representing all women who are not in the first group identified via the 'longest possible AND statement criterion.' The resulting minimal Boolean expression would, hopefully, have clearly interpretable components that would represent distinct pathways—

hence, additional members of the desired taxonomy of life histories—to given mental health outcomes. Although the analysis in Singer et al. (1998) proceeded beyond the first specification of an initial group by inspection and judgment, linked to organizing principals, we recommend application of both strategies en route to a final taxonomy. Indeed, as illustrated in Ragin (1987), minimal Boolean expressions that are not readily interpretable on a first pass through reduced response vectors often suggest recoding and construction of alternative composite variables that ultimately lead to a useful taxonomy.

3.8 Tests of Distinguishability

Having constructed a taxonomy of pathways for each outcome category—resilient and depressed/unwell in the present example—it is necessary to demonstrate that the routes to any one outcome differ in important ways from the set of routes to a different outcome. To this end, we first ask whether the aggregate pathways for the resilient women differ from the pathways for the depressed/unwell population. In particular, we tabulate the frequency with which the resilient pathways occur in the depressed/unwell population. Conversely, we take the aggregate pathways representing the depressed/unwell women and tabulate the frequency with which those pathways actually occur among the resilient women. We expect to find that the resilient pathways are rare in the depressed/unwell population and also that the depressed/unwell pathways are rare in the resilient population. For example, 16 percent of the resilient women are on the pathway indicated above by the seven-component logical AND statement. Only 8 percent of the depressed/unwell have the same pathway ($p < .10$). Twenty percent of the resilient women are on the pathway where at least one parent is a chronic alcoholic AND there is a diversity of compensating positive experiences in adulthood. This pathway only occurs for 5 percent of the depressed/unwell population ($p < .05$). There are persistently negative-experience pathways in the depressed/unwell population that have no counterpart in the resilient group. Sharper discrimination occurs depending upon the particular application (see e.g., Zhao et al. 2000). The central point, however, is that assessment of distinguishability is important for making the case that there are, as hypothesized, qualitatively different routes to distinct mental health outcomes.

4. An Alternative Bottom-up Strategy

An analytical strategy that might serve as an alternative to what is put forth in the above example is optimal matching (Abbott and Hrycak 1990, Abbott and Barman 1997). Optimal matching derives from the literature on comparison of DNA sequences. It takes univariate time series for individuals and groups them together into sets of what are regarded as equivalent histories. The grouping requires the availability of a metric on the space of possible sequences (histories); and two sequences are defined to be members of the same group if their distance apart—in terms of the metric—is sufficiently small. To apply the automated optimal matching technique requires that input data be in the form of what we have called 'reduced response vectors.' The analogue of steps leading from more complex longitudinal data to this stage, as we have described above, must still be carried out for univariate time series prior to carrying out the formal aggregation step (Abbott and Hrycak 1990). Although optimal matching is a natural way to proceed in the construction of a taxonomy of histories, the problem with its implementation on cross-time, multi-life-domain data is that there is no obvious metric yet developed that is suited to the complex multidimensional time series corresponding to individual histories, as discussed in the above example. Furthermore, a single measure of distance between two multidimensional histories seems too coarse for decisions about what constitute equivalent histories. The strategy for aggregation of reduced response vectors—in the above example—uses information about multiple features of life histories. This suggests that any automated aggregation technique, analogous to optimal matching, should be based on a multidimensional measure on the space of histories.

5. Top-down Methods

There is a considerable diversity of top-down, person-centered analytical strategies. A valuable review of both concepts and methods is contained in Cairns et al. (1998). One common strategy focuses on clustering techniques, where a critical input to the algorithms is a metric on the space of histories. Not surprisingly, applications of these techniques, which partition a heterogeneous population into relatively homogeneous clusters, have been restricted to low-dimensional data and rather simple summaries of histories. To reach the level of nuance and multi-domain representation of the kind we have discussed above requires a quite different set of measures of similarity of histories to be constructed than heretofore. This is an important research topic for the future.

Recursive partitioning (RP) classification (Breiman et al. 1984, Zhang and Singer 1999) is an automated procedure that fits trees to multi-dimensional longitudinal data such that pathways down the tree to a given outcome (terminal node) identify homogeneous sub-populations with common histories. Trees are

constructed with the objective of maximizing ability to predict membership in the outcome categories. Each pathway down the tree ending in a given terminal node (identified, for example, with a health outcome) corresponds to a logical AND statement incorporating information across multiple life domains. Two or more pathways into the same terminal node are identified with a logical OR statement describing how that particular outcome occurred. The judgment of the investigator is required to determine whether all pathways to the nodes identified with a given outcome should be summarized by a single complex Boolean expression, or partitioned into two or more sets of pathways defining qualitatively different kinds of histories. Recursive partitioning (RP) algorithms could accept either the original raw response vectors or reduced response vectors, as described in the above example.

From our perspective, a limitation of RP for generating a taxonomy of life histories is the fact that the performance criteria and the splitting rules for tree construction are not constrained by substantive knowledge from any field of scientific inquiry (e.g., a set of organizing principles). Thus, pathways down the best predictive tree are not always interpretable within the knowledge of a given field. This disjunction between statistical performance criteria, based solely on numerical goodness-of-fit, and constraints on tree structure and splitting rules derived from extant scientific knowledge have proven problematic in past applications of RP (Levy et al. 1981, 1985), for example, applied RP to neurological assessments of comatose patients to predict which patients were likely to recover with moderate disabilities as opposed to remaining vegetative or dying within one year of hospital admission. The numerically best predictive trees were virtually never neurologically interpretable. Rather, inspection of 20+ nearly optimal trees led to selection, based on scientific judgment, of a prognostic tree that had interpretable pathways. For a start at more interactive—i.e., allowing the investigator to modify trees during the construction process on the basis of organizing principles—and flexible strategies based on RP technology, the reader should consult Zhang and Singer (1999).

6. *Conclusions*

We have described a variety of person-centered methods for the analysis of longitudinal data. Particular emphasis has been given to a relatively new strategy—the bottom-up example—that takes a full complex individual history as the unit of analysis. The principal and novel features of this strategy are:

(a) It includes the writing of narratives at the beginning and terminal stages of the analysis of survey data.

(b) It is person- not variable-centered.

(c) It produces pathways that represent aggregates of individual life histories.

(d) It works simultaneously across multiple life domains.

(e) It is capable of identifying multiple pathways to a given outcome.

(f) It systematizes the decision making en route to delineation of the pathways.

(g) It tests hypotheses about whole lives that are the integration of what is typically studied as separate pieces.

Point (a) underscores the importance of examining whole lives in story form as a means of comprehending the volume of detail that necessarily comprises human lives. Although point (b) is a clear objective in much of the life-history literature (e.g., Elder 1974, Magnusson and Bergman 1990), the lack of person-centered methodological tools has prevented investigators from proceeding to point (c). Attempts to get at aggregates have typically involved the use of variable-centered techniques, since these were available. However, the investigators mentioned above and many others were well aware that the lack of preservation of individuals as units of analysis severely limited the inferences they could draw about the dynamics of individual (or aggregates of individuals) change. Point (d) is conventionally dealt with, if at all, by variable-centered techniques. The bottom-up example strategy is, to the best of our knowledge, the first incorporation of multiple life domain information in a process of aggregation of whole life representations. Many investigators (e.g., Harris et al. 1990, Magnusson and Bergman 1990) have struggled with point (e); however, the absence of tools for aggregating whole life representations has prevented their delineating full pathways. Harris et al. (1990) set forth a template of several pathways to depression, and Magnusson and Bergman (1990) looked for 'patterns' connecting childhood problems with adult maladjustment. Nevertheless, their variable-centered empirical analyses (using logistic regression with interaction terms in the former, and cluster analysis in the latter) did not achieve the desired objective. What was needed was a person-centered methodology. Point (f) emphasizes the critical role of scientific judgment embedded in the progression of methodological steps, something that is an essential part of any data analysis, but is infrequently reported. Finally, point (g) carries hypothesis testing into the uncharted territory of whole lives.

Recommended reading for more detail on the person-centered methods emphasized in this entry and on the integration of numbers and narratives to study whole lives, including narratives as data, see Singer et al. (1998), Singer and Ryff, (2001), and Sampson and Laub (1993). An in-depth exposition of recursive partitioning, including online software, is contained in Zhang and Singer (1999). An instructive example of formal modeling of multiple-episode event history

data is given in Heckman and Walker (1998) and a useful illustration of modeling of multivariate event history data, with both observed and unobserved covariates, is contained in Manton et al. (1987).

See also: Adulthood: Developmental Tasks and Critical Life Events; Factor Analysis and Latent Structure: Overview; Growth Curve Analysis; Life Management, Developmental Psychology of; Lifespan Development, Theory of; Linear Hypothesis: Regression (Basics); Longitudinal Data: Event–History Analysis in Discrete Time; Markov Models and Social Analysis; Person-centered Research

Bibliography

Abbott A, Barman E 1997 Sequence comparison via alignment and Gibbs sampling: A formal analysis of the emergence of the modern sociological article. In: Raftery A (ed.) *Sociological Methodology, 1997*. American Sociological Association, Washington, DC, pp. 47–87

Abbott A, Hrycak A 1990 Measuring resemblance in sequence data: An optimal matching analysis of musician's careers. *American Journal of Sociology* **96**: 144–85

Barker D, Osmond C 1986 Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *Lancet* **I**: 1077–81

Bergman L 1998 A pattern-oriented approach to studying individual development: Snapshots and processes. In: Cairns R, Bergman L, Kagan J (eds.) *Methods and Models for Studying the Individual*. Sage Publications, Thousand Oaks, CA, pp. 83–122

Berkman L, Singer B, Manton K 1989 Black/white differences in health status and mortality among the elderly. *Demography* **26**(4): 661–78

Breiman L, Friedman J, Olshen R, Stone C 1984 *Classification and Regression Trees*. Wadsworth, Belmont, CA

Cairns R, Bergman L, Kagan J (eds.) 1998 *Methods and Models for Studying the Individual*. Sage Publications, Thousand Oaks, CA

Elder G 1974 *Children of the Great Depression: Social Change in Life Experience*. University of Chicago Press, Chicago

Harris T, Brown G, Bifulco A 1990 Loss of parent in childhood and adult psychiatric disorder: A tentative overall model. *Development and Psychopathology* **2**: 311–28

Hauser R, Carr D, Hauser T, Hayes J, Krecker M, Kuo H, Magee W, Presti J, Shinberg, D, Sweeney M, Thompson-Colon T, Uhrig S Warren J 1993 *The Class of 1957 After 35 years: Overview and Preliminary Findings*. Working paper 93-17, Center for Demography and Ecology, University of Wisconsin-Madison, Madison, WI

Heckman J, Walker J 1998 Using goodness of fit and other criteria to choose among competing duration models: A case study of Hutterite data. In: Cloog C (ed.) *Sociological Methodology, 1987*. American Sociological Association, Washington, DC, pp. 247–307

Levy D, Bates D, Caronna J, Cartlidge N, Knill-Jones R, Lapinski R, Singer B, Shaw D, Plum F 1981 Prognosis in non-traumatic coma. *Annals of Internal Medicine* **94**: 293–301

Levy D, Caronna J, Singer B, Lapinski R, Frydman H, Plum F 1985 Predicting outcome from hypoxic-ischemic coma. *Journal of American Medical Association* **253**: 1420–26

Magnusson D, Bergman L 1990 A pattern approach to the study of pathways from childhood to adulthood. In: Robins L, Rutter M (eds.) *Straight and Devious Pathways from Childhood to Adulthood*. Cambridge University Press, Cambridge, UK, pp. 101–15

Manton K, Stallard E, Woodbury M, Tolley D, Yashin A 1987 Grade-of-membership techniques for studying complex event history processes with unobserved covariates. In: Clogg C (ed.) *Sociological Methodology, 1987*. American Sociological Association, Washington, DC, pp. 309–46

Power C, Manor O, Fox A 1991 *Health and Class: The Early Years*. Chapman and Hall, London

Ragin C 1987 *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. University of California Press, Berkeley, CA

Rowe J, Kahn R 1998 *Successful Aging*. Pantheon Books, New York

Shifren K, Hooker K, Wood P, Nesselroade J 1997 Structure and variation of mood in individuals with Parkinson's disease: A dynamic factor analysis. *Psychology and Aging* **12**(2): 328–39

Singer B, Ryff C D 1997 Racial and ethnic inequalities in health: Environmental, psychosocial, and physiological pathways. In: Devlin B, Fienberg S, Resnick D, Roeder K (eds.) *Intelligence, Genes, and Success: Scientists Respond to the Bell Curve*, Springer-Verlag (Copernicus), New York, pp. 89–122

Singer B, Ryff C D 2001 Person-centered methods for understanding aging: The integration of numbers and narratives. In: Binstock R, George L K (eds.) *Handbook of Aging and the Social Sciences*. Academic Press, San Diego, CA, pp. 44–65

Singer B, Ryff C, Carr D, Magee W 1998 Linking life histories and mental health: A person-centered strategy. In: Raftery A (ed.) *Sociological Methodology, 1998*. American Sociological Association, Washington, DC, pp. 1–51

Stone L 1979 The revival of narrative: Reflections on a new old story. *Past and Present* **85**: 3–24

Turner M 1996 *The Literary Mind*. Oxford University Press, New York

Vaillant G 1983 *The Natural History of Alcoholism*. Harvard University Press, Cambridge, MA

Wadsworth M, Kuh D 1997 Childhood influences on adult health: A review of recent work from the British 1946 national birth cohort study, the MRC national survey of health and development. *Pediatric Perinatal Epidemiology* **11**: 2–20

Zhang H, Singer B 1999 *Recursive Partitioning in the Health Sciences*. Springer-Verlag, New York

Zhao H, McLanahan S, Brooks-Gunn J, Singer B 2000 Studying the real child rather than the ideal child: Bringing the person into developmental studies. In: Bergman L, Cairns R, Nilsson L-G, Nystedt L (eds.) *Developmental Science and the Holistic Approach*. Lawrence Erlbaum, Mahwah, NJ, pp. 393–419

B. H. Singer

Longitudinal Data: Event–History Analysis in Discrete Time

The practical motivation for considering discrete-time models of event histories (Allison 1982, Thompson 1977) comes from the fact that such longitudinal data

are often recorded and presented in a discrete-time form (see *Longitudinal Data*). Thus, for example, it is common to collect information on the values of considered individual status variables at regularly spaced time points and to report such measurements directly as the data. A time unit is then fixed, typically to be a day, a week, a month, or a year, thus creating a regular time lattice. A slightly different but related reason why event-history data often appear in a discrete-time form is that the event times themselves are reported in a rounded form. Then the events which happened inside a time interval of unit length are somewhat arbitrarily assigned to either of its end points or, in so-called *actuarial methods*, to the middle. As a consequence, the reported event times are often *tied*, with several individuals sharing the same value. As an example, one can consider the changes of a person's employment status during the work history, typically consisting of spells of employment and unemployment, and finally ending with a permanent transition to outside of the work force. In order to be completely accurate, the reporting might have to done day-by-day. More realistically in practice, it can only be done monthly, and then a number of different alternatives exist.

For example, one could directly report the employment status on the last day of the month. A drawback of such reporting would be that, if the status can undergo frequent changes, the status at the end of the month might be purely accidental and not represent the entire month in an adequate way. An alternative would be to set a threshold x and ask whether the person has been employed for at least x days during the month. Such a solution may not be completely satisfactory either, as a low value of x does not distinguish between people who were employed only briefly from those who were employed during the entire month, with a similar but opposite problem arising for a high value of x. Such problems are of less concern if changes in the status variable are less frequent (such as marital status), or even irreversible.

1. Discrete-time Hazards

The main advantages from linking a statistical model directly with such data are that model construction becomes relatively straightforward, and that no special methods are required for dealing with the ties. For an illustration, suppose for simplicity that there are only two states for each individual, say *alive* and *dead*, and that the chosen time unit is one month. Then the lifetime T of an individual, in a discretized time scale, can be thought of as being determined by a sequence of Bernoulli trials, resulting in a sequence of 'zeros' as long as the individual is alive and ending with a 'one' for the month during which the individual dies. The basic element for setting up a corresponding statistical model is the conditional probability of recording 'one' in a unit interval given that all preceding intervals resulted in 'zeros', that is, the probability of death during the tth month for a person still alive after $t-1$ months (see *Survival Analysis: Overview*). Because of the discrete-time lattice used, such *discrete-time hazards* can be simply expressed as conditional probabilities:

$$h(t) = Pr(T = t | T \geqslant t) \tag{1}$$

By applying the chain multiplication rule of conditional probabilities progressively for increasing values of t, one finds readily that the unconditional probability for the survival is given by the product

$$Pr(T > t) = \Pi^{t}_{s=0}[1 - h(s)] \tag{2}$$

Note that moreover

$$Pr(T = t) = Pr(T > t-1)h(t) \tag{3}$$

These are discrete-time analogues of the well known *exponential formulas* for the survival probability and density, respectively, which are valid in continuous time and express a lifetime distribution in terms of the corresponding *hazard rate* (see, e.g., Andersen et al. 1993, Kalbfleisch and Prentice 1980). In fact, the exponential formulas can be obtained as limits when the time lattice used in the discretization becomes more and more dense, thereby making the discrete-time hazards $h(t)$ numerically small.

2. Regression Models and Likelihood Inference

When considering data consisting of lifetimes of several individuals, the number of such Bernoulli trials equals the total number of person-months at risk, and under natural independence conditions across individuals, a statistical model for the entire data can be formulated in terms of equally many Bernoulli trials, with the corresponding discrete-time hazards depending on age. This gives rise to a simple likelihood expression of a product form, without that tied lifetimes in the data would be a problem. The structure of the likelihood remains essentially unchanged if the discrete-time hazards of the considered individuals depend on covariates, as long as the covariates are fixed in time. One is then dealing with a regression problem, in which the observed development of the response variables (lifetimes) is explained in terms of the corresponding realized covariate values in the data (see *Survival Analysis: Overview*). A convenient framework for considering such regression models, including the possibility of using standard software, is offered by *generalized linear models* (for a general account and references, see Fahrmeir and Tutz (1994), and for asymptotic properties of the estimators, see, e.g., Arjas and Haara (1987)). In the important special

case of a *logistic link* function, if the considered time unit is short enough to make the discrete-time hazards $h(t)$ small and the intercept of the linear expression is allowed to depend on the interval, the model corresponds closely to the proportional hazards model of Cox (1972). If the covariates can themselves change in time, the same logic and structure of the likelihood can be used as long as the covariates are *external* (Kalbfleisch and Prentice 1980) in the sense that their development can be thought of as being exogenous from the point of view of the studied event-history data of individuals.

This is because there is then no need for an explicit modeling of the process governing the covariate process, as the resulting contributions to the likelihood expression do not depend on the actual parameters of interest and can therefore be coalesced into a proportionality constant. The same logic applies if the individual event histories are right censored and the censoring mechanism is assumed to be noninformative. If the model also involves time-dependent *internal* covariates, special care is needed in modeling and inference to make sure that their development in time does not *confound* the regression problem and the corresponding conclusions, which are often given a causal form. The issue of confounding, in the case of sequential designed experiments, has been considered in a series of papers by Robins (e.g., Robins 1997). Sometimes, in order to give useful causal interpretations, it becomes necessary to build the regression model in two or more levels representing, for example, group level and individual level effects. For a recent contribution in this direction, see Barber et al. (2000). The same general ideas apply also in situations in which the considered status or response variables are no longer just binary. For example, the status of an individual could assume more than two possible values, and the same status could repeat itself many times during the follow-up. For example, as noted above, a person's employment status could first alternate several times between the states *employed* and *unemployed*, until it finally reaches the absorbing state *outside of work force* because of retirement. What is needed, however, is a convenient form of conditioning, which carries information about an individual's earlier history and present status and thereby influences the assessment of the discrete-time hazards associated with the future development, always by one time unit ahead. For an interesting case study, applying a variety of inferential techniques, see Fahrmeir and Wagenpfeil (1996).

3. *Structure of the Likelihood*

Let us now make this idea more explicit. Indexing the individuals in the data by $i \in I$ and denoting the value of the corresponding status or response variable at time t by $X_t(i)$, we are led to consider conditional probabilities for a transition from a state $X_t(i)$ at time t to a state $X_{t+1}(i)$ at a later time $t+1$. Denoting by $H_t(i)$ the pre-t-history of individual i, which contains all observations made on i up to and including time t, including the past values of $X_s(i)$ $(s \leqslant t)$ and of possible covariates considered in the analysis, the first task is to specify the conditional probabilities $\Pr[X_{t+1}(i)|H_t(i)]$.

The following natural conditional independence assumptions will then often be valid:

(a) The variables $X_{t+1}(i)$, $i \in I$, are conditionally mutually independent given the collective pre-t-past $\{H_t(i), i \in I\}$ in the data.

(b) Knowledge of the individual's own pre-t-history $H_t(i)$ is sufficient for predicting $X_{t+1}(i)$ in the sense that

$$Pr(X_{t+1}(i)|H_t(j), j \in I) = Pr[X_{t+1}(i)|H_t(i)] \qquad (4)$$

Under conditions (a) and (b), the likelihood expression arising from discrete-time event history data becomes a product of terms of the form $\Pr[X_{t+1}(i)|H_t(i)]$, where the product is taken over all individuals i and all times t. (A tacit assumption here has been that if the covariates registered in $H_t(i)$ are themselves time dependent, they are either exogenous or their evolution is modeled in a way which does not involve the actual parameter of interest.) This makes the statistical inference of the model parameters in principle a straightforward problem, for example, by applying maximum likelihood. If, moreover, the condition

(c) Knowledge of the current state $X_t(i)$ is sufficient in the sense that

$$Pr[X_{t+1}(i)|H_t(i)] = Pr[X_{t+1}(i)|X_t(i)] \qquad (5)$$

holds, the model can be formulated in terms of a discrete-time *Markov chain* (see, e.g., Cinlar 1975, and *Markov Models and Social Analysis*). Information contained in the observed past history $H_t(i)$ can then be stored into a covariate vector of a fixed length, and the likelihood becomes a product of transition probabilities $\Pr[X_{t+1}(i)|X_t(i)]$, again taken over both i and t.

In spite of the seemingly general applicability of such a discrete-time approach to likelihood inference, the following two conventions, both usually involving an approximation, seem necessary in interpreting such models. First, censoring and possible other forms of controlling the number of individuals at risk are always thought to happen at times which are integer multiples of the chosen time unit. Second, if the model involves time-dependent covariates which are used as regressors of the observed status or response variable, their measured values need to be interpreted as 'prevailing (constant) conditions for exposure' during each considered time interval of unit length (see Sandefur and Tuma 1987). In particular, if the response variable can feed back and influence the

values of the covariates, in order to save a causal interpretation of the regression model, the covariate values must be determined already at the beginning of each time interval, thus avoiding the potential problem of conditioning on variables which are intermediate between a contemplated cause and the effect.

4. Embedding into Continuous-time Models

Another, and perhaps a more serious, drawback from a modeling perspective concerns the interpretation of the discrete-time parameter. In almost any conceivable applied context, and corresponding the common human perception, time is viewed as being continuous. Based on this perception, and even in situations in which the data are in a discrete-time form, a more natural alternative for statistical modeling of event histories starts from the idea that there is a true underlying status process in continuous time, and that the recorded event-history (panel) data consist then of a series of measurements from this underlying process taken at a number of typically prescheduled time points. The term *current status data* is sometimes used to emphasize this aspect. The development of a person's true employment status, marital status, support of political parties, or presence or absence of a chronic disease, as functions of time, would be natural examples. Indeed, such individual status histories would be the focus of a statistical analysis if they only could be observed, while discrete-time event history data consist merely of incomplete information, typically drawn from a sequence of cross-sections of such histories. As a consequence of such imprecision, individual event histories in continuous time are generally unidentifiable from discrete-time data. In particular, the exact times of occurrence are then not recorded, and if the status of a person has changed back and forth within a unit interval, this possibility may not even become explicit in the data (see, e.g., Stasny 1986). Nevertheless, it may be possible to embed the data successfully into a larger statistical model in continuous time. The fact that the individual underlying continuous-time developments cannot generally be determined uniquely from the data can often be compensated with some form of *data augmentation*. In other words, we can then consider an entire collection, or distribution, of sample paths which are consistent with the data. The practical solution may involve some form of simulation (multiple imputations), while a theoretical motivation to such techniques comes from Bayesian statistical inference. Exact numerical computations become quickly infeasible in practice, however. This leads to an application of sampling-based *Markov chain Monte Carlo* methods, in which the possible realizations of the individual processes are drawn by simulation (see *Monte Carlo Methods and Bayesian Computation: Overview*). Such data augmentation methods have been applied to data describing recurrent bacterial infections by Auranen et al. (2000), but the same approach applies more generally to problems in which it is natural to relate discrete-event history data to underlying individual developments in continuous time.

5. Conclusion

Depending on the context, modeling of event-history data may emphasize more the *event times* at which some particular events of interest occurred, or the *states* in which some studied process happened to be at the times at which it was measured. To stress the former aspect, one often speaks of survival or duration data. The second aspect is similarly stressed by talking about repeated measurements or longitudinal data. When the time variable in the data and in the model is discrete, it may sometimes be difficult to make a clear distinction between these two basic forms. On the other hand, the likelihood expressions will be of the same canonical product form in both cases, so that likelihood inference will remain the same. Recommended reading on these and closely related issues include Allison (1982), Barber et al. (2000), Diggle et al. (1994), Kalbfleisch and Prentice (1980) and Sandefur and Tuma (1987).

See also: Longitudinal Data; Markov Models and Social Analysis; Monte Carlo Methods and Bayesian Computation: Overview; Survival Analysis: Overview

Bibliography

Allison P D 1982 Discrete-time methods for the analysis of event histories. In: Leinhardt S (ed.) *Sociological Methodology*. Jossey-Bass, San Francisco, CA, pp. 61–98

Andersen P K, Borgan O, Gill R D, Keiding N 1993 *Statistical Models Based on Counting Processes*. Springer, New York

Arjas E, Haara P 1987 A logistic regression model for hazard: Asymptotic results. *Scandinavian Journal of Statistics* **14**: 1–18

Auranen K, Arjas E, Leino T, Takala A K 2000 Transmission of pneumococcal carriage in families: A latent Markov process model for binary longitudinal data. *Journal of the American Statistical Association* **95**: 1044–53

Barber J S, Murphy S, Axinn W G, Maples J 2000 Discrete-time multilevel hazard analysis. *Sociological Methodology* **30**: 201–35

Cinlar E 1975 *Introduction to Stochastic Processes*. Prentice-Hall, Englewood Cliffs, NJ

Cox D R 1972 Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, *Series B* **34**: 187–220

Diggle P J, Liang K-L, Zeger S L 1994 *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK

Fahrmeir L, Tutz G 1994 *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer, New York

Fahrmeir L, Wagenpfeil S 1996 Smoothing hazards and time-varying effects in discrete duration and competing risk models. *Journal of the American Statistical Association* **91**: 1584–94

Kalbfleisch J D, Prentice R 1980 *The Statistical Analysis of Failure Time Data*. Wiley, New York

Robins J 1997 Causal inference from complex longitudinal data. In: M Berkane (ed.) Latent Variable. Modeling and Applications to Causality. *Lecture Notes in Statistics* **120**: 69–117. Springer, New York

Sandefur G D, Tuma N B 1987 How data type affects conclusions about individual mobility. *Social Science Research* **16**: 301–28

Stasny E A 1986 Estimating gross flows using panel data with nonresponse: An example from the Canadian labor force survey. *Journal of the American Statistical Association* **81**: 42–7

Thompson W A Jr 1977 Treatment of grouped observations in life studies. *Biometrics* **33**: 463–70

E. Arjas

Longitudinal Research: Panel Retention

1. Introduction

Panel studies involve the collection of data over time from a baseline sample of respondents. Unlike other forms of longitudinal studies, panels allow for the study of individual behavior change over time. However, because the same individuals are followed over time, there is considerable respondent burden and eventual attrition from waves of interviewing. Since the value of panel surveys is dependent upon the ability to study the same respondents at different points in time, reducing attrition is of major concern. Loss of respondents raises the possibility of bias if those who are lost to follow-up differ from those who remain in the panel on key dependent variables. Therefore panel attrition can affect both the internal and external validity of the study (Cook and Campbell 1979. In this paper, we review the recommended strategies for minimizing panel attrition with general populations. Some of the strategies depend on institutions like the Post Office whose functions vary from country to country. The information provided here is about the USA, but we judge that much of it will transfer to other countries or can be adapted to their particular circumstances. The recommendations discussed here for minimizing panel attrition apply mainly to population-based samples. Panel surveys with more hard-to-locate or high-risk populations (such as the homeless, drug users, or runaway youths will require more specific, tailored strategies (Cottler et al. 1996, Walton et al. 1998, Wright et al. 1995) (see Cottler et al. 1996, Walton et al. 1998, Wright et al. 1995).

2. Factors Associated with Panel Retention

Retaining panel respondents is generally believed to depend on three factors. First is the respondent's willingness to participate in the study at future waves. Even when respondents are located for follow-up, keeping them interested and engaged in participation is a challenge. Second, the respondent's continued availability to participate depends on factors like illness or death. The third, and most important factor in panel maintenance, is the ability to locate respondents for interview at later waves.

Failure to find a respondent at later waves is predominantly due to residential mobility. Based on data from the 1998 Current Population Survey, during which respondents were asked whether they had lived at the same residence one year before the interview, it was estimated that 42.5 million Americans moved (Faber 2000). The majority (64 percent) moved within the same county, 19 percent moved to a different county in the same state, and 16 percent moved to a different state. Age is correlated with moving; the highest rates of residential mobility are found among those aged 20–29 and they decline as age increases. Rates of residential mobility also vary with race and ethnicity and by region of the country. They are lowest for non-Hispanic whites (14.5 percent), higher for Blacks and Asians (18.9 percent) and highest among Hispanics (21.2 percent). People who rent their homes are more likely to relocate than those who own their residences are. Nearly one-third of people living in rental units reported having moved in the previous year compared with less than 10 percent of homeowners. Rates of residential relocation vary by region as well. They are highest in the West (19.4 percent), lower in the South (17.2 percent), Midwest (14.7 percent) and Northeast (11.5 percent), respectively. Understanding these correlates of residential relocation will help investigators plan for the level of tracking they will need to do during follow-up interviews. For example, a panel of Hispanics in California would be likely to have higher attrition than a panel of white householders in Maine. Hence more resources would be required for respondent retention in the California sample than in the Maine sample. Including these variables as indicators at the baseline wave of data collection is important for postsurvey attrition assessment and weighting to adjust for it.

3. Study Design Features to Minimize Panel Bias

There are three main factors that affect attrition: recruiting the respondent to the study, locating the respondent for subsequent interviews, and maintaining the respondent's commitment to the panel.

3.1 Recruiting Respondents to the Panel

Plans for recruitment and retention should be developed at the time the study is designed. They are as important as any other feature of the study design, but often do not get the serious attention (or budget

allocation) they deserve. Later, we discuss the specific recruitment and relocation strategies that might be built into these plans.

3.1.1 Preplanning. Preplanning will ensure not only that there are sufficient resources to successfully recruit and retain the respondents, but also that the information needed to track respondents is collected during the baseline interview. It will also guide the selection of the sample size since this must allow for attrition if locating respondents is anticipated to be a serious problem. It may also affect decisions about the mode of interview because certain segments of the population may not be available by telephone or computer. For example, telephone coverage is related to socioeconomic status and to ethnicity (Belinfante 2000), and computer access is even more strongly so.

3.1.2 Study logo. Part of recruitment and retention is creating a recognizable identity for the study. A study logo and identifiable study title that is used in all correspondence (written and oral) and that appears on self-administered questionnaires, where applicable, have also been suggested as a way to create an identifiable image of the study for the respondent (Given et al. 1990).

3.1.3 Advance letters. Depending on the sample frame (list, random-digit dial generated, or block-listed sample), an advance letter should be sent to respondents at baseline that explains the purpose of the study, and where they can go for further information (a phone number or study web site, depending on the population). Advance letters are most commonly used in studies where the mode of interview administration is face-to-face. Advance letters have also been used in dual frame telephone surveys where telephone interviewing is combined with face-to-face or mail on the theory that they will increase response rates, although the evidence is mixed (Parsons et al. 1999, Traugott et al. 1987). Clearly, however, advance letters require that the investigator have a current address for the respondent. Even if addresses are not possible at the baseline interview wave, they should still be obtained at the baseline interview for follow-up purposes.

3.1.4 Response rates. The most important aspect of recruitment is the baseline survey where respondents have their initial contact with the study. Obtaining an acceptable rate of cooperation to the interview will affect the overall data quality. If the baseline survey is biased due to nonresponse, then the internal and external validity of the panel will always be problematic. Four methods are typically used for ensuring high contact and recruitment rates (Groves and Couper 1998): (a) budget for additional contact attempts with those sample units that cannot be immediately contacted; (b) schedule these contact attempts (whether by telephone or face-to-face) at various time of the day, evenings, and weekends to maximize the opportunity to reach someone at home; (c) plan for lengthier data collection during the follow-up waves to allow for respondent vacations and for locating hard-to-find respondents; and (d) allow interviewers to use local sources, such as neighbors, for information on the respondent or household. These techniques are standard in cross-sectional surveys but are especially important for panel studies to improve response rates.

3.2 Locating and Tracking Respondents for Follow-up Waves

Once respondents are recruited into the panel, several methods can be used to increase the likelihood of contact at designated follow-up waves.

3.2.1 Advance letters. Regardless of whether advance letters were used to recruit subjects, at every subsequent data collection wave, advance letters should be sent before interviewing begins. These letters notify the respondent of the upcoming visit or call and alert the investigator to a possible address change. The 'address correction requested' service available through the post office should be used with each letter.

3.2.2 Secondary sources. During the baseline interview, information should be obtained that will help locate respondents later. One invaluable resource is the secondary contact source. Secondary sources are usually one or two persons, not residing with the respondent, who will know how to contact the respondent at any point in time. At a minimum the relationship of the secondary source to the respondent and that person's telephone number should be collected.

In addition, during the baseline interview all respondents should be asked to provide their dates of birth, social security numbers (for use with credit-based tracking programs), and the names under which the respondents' telephones are listed. Depending on the population, respondents' e-mail addresses should also be collected as they can be used as an inexpensive means of tracking respondents.

3.2.3 In-person location strategies. Even the most conscientiously designed study will not prevent respondents from being lost to follow-up. If all secondary source information does not lead to the whereabouts of the respondent, there are several sophisticated tracking devices that can be used. At the very least, attempts at locating the respondent should persist until a final disposition of the case can be ascertained for the purposes of calculating response rates.

In face-to-face studies, an excellent source of information about respondents comes from people who are connected in some way with the respondent. Roommates, neighbors, property managers, mail carriers, family, and friends might be available to assist in searching for, and contacting, a hard-to-find study participant. When finding the respondent through direct contact does not lead to complete or solid information, personal informants can sometimes provide essential information about how to reach the respondent, what kind of schedules they keep, and whether they have moved.

3.2.4 Mail location strategies. In addition to 'address correction requested' stamped on advance letter envelopes, a number of other useful services are available through the US Postal Service (USPS) to improve mailing addresses and locate respondents who have moved (United States Postal Service 1999). Through the Address List Correction Service, business and resident mailing lists can be submitted for correction. Services provided include cross-listing of names to which mail cannot be delivered or forwarded; correcting ZIP codes, house numbers, rural, or post office box numbers, street names, and suffixes; furnishing new addresses for customers who have moved; and adding building, apartment, suite or room numbers when known. There is a minimum charge for correcting each submitted list, which can be arranged through local post office branches. In addition, the USPS National Change of Address (NCOA) program processes change of address data nationwide. The NCOA service is provided by private sector companies, certified and licensed by the USPS, who charge a fee for the service. These companies receive updated, computerized COA information on a regular basis. To search for lost or unlocatable study respondents, a file of names and addresses is submitted to one of the NCOA vendors, who search for matches in the NCOA. New addresses are provided when an exact match is found. Most other countries will have systems with many of these attributes.

3.2.5 Haines criss-cross directories. In the USA, the Haines Directory alphabetically cross-lists every street in a given community with the associated names, address, and (officially listed) telephone numbers of the people who live on these streets. A second section of the directory lists all published telephone numbers in sequence and cross-references them with the corresponding address and names of the residents. Haines directories can thus be used when only the telephone number or only the address is available. They can be found in local public libraries, which also house the directories for several years in the past.

3.2.6 Credit bureaus and vital statistics agencies. Another source for tracking panel respondents is credit bureau records. Software packages can be purchased from credit bureaus that allow investigators to search the credit bureau databases directly (for updated name and address information). By entering the last known address of the respondent, the file will provide data on whether the respondent has moved and, if so, the new address will be provided. These systems are a valuable investment if you are able to obtain the respondents social security numbers, as credit records are tied to that identifier. However, the social security number is not useful for locating all populations in the USA. For example, low-income or homeless respondents will be unlikely have a credit history that can be referenced or tracked. There are also a variety of Internet search sites that can be used for tracking. While they are free of charge, they are also not very comprehensive or up-to-date. They are often linked to the respondents' e-mail address and if the respondent does not use the Internet, they may not be found via this source.

In other situations, accessing public records is necessary for obtaining missing information about respondents. Public records contain data on births, deaths, marriages, criminal activity, military service, motor vehicle registration, and employment. Of particular use in panel studies that span a long period of time is the National Death Index, maintained by the National Center for Health Statistics. For a fee, the index will match respondents with their file of deceased persons using the name and date of birth. The likelihood of a match increases if the social security number or middle initial is available. In some states it is possible to search a web site from the department of corrections to ascertain if any study participants have been incarcerated. (For searching public records in the USA, see Gunderson 1996, Hoyer and McCann 1988, and Parco 1994.)

3.3 Retaining Respondents Between Interview Waves

Being a member of a panel is burdensome for the respondent. Thus, it is critical to the successful

maintenance of the sample that the relationship be reinforced between contacts.

3.3.1 The importance of contact. Regardless of the duration of time between data collection waves, it is advisable to maintain contact with respondents every 3–6 months as resources allow (Hunt and White 1998). Regular contact keeps the respondents engaged and reminds them of the value placed by the investigator on their participation. It also helps the investigator keep track of the respondent. The extent of the contact need not be more than the advance letter or a thank you note. Besides a friendly greeting and an expression of thanks for the respondents' participation, the content should include a reminder of the time of the next follow-up survey and a request for new contact information if the respondent has moved or will be moving. Investigators can also personalize the contact between interviewing waves by using birthday cards, holiday greetings, or other forms of written correspondence (Walton et al. 1998). These formats reinforce that the study staff is interested in and cares about the participants.

3.3.2 Incentives. Incentives are also critical to effective panel maintenance. Overall, the literature favors the use of incentives to increase response rates. Not only do they represent a concrete expression of appreciation to the respondent, but they also have the effect of keeping the respondent interested and involved. Anecdotal evidence also suggests that respondents are more likely to notify investigators of new addresses when they have moved if they expect an incentive.

Incentives typically take the form of money (cash or checks) or something with cash value (such as gift certificates or gifts), and can be presented to the respondent in advance of the interview, at the time the interview is conducted, or postinterview. The benefits of incentives have been empirically demonstrated in a meta-analysis of their effects on response to mail surveys (Church 1993). This analysis demonstrated that prepaid incentives yielded higher response rates than when the incentive was promised upon return of the survey form, and that cash was more successful than gifts in increasing response rates. Similarly, Singer et al. (1999) reported the results of their study of the effects of incentives in telephone and face-to-face studies and concluded that monetary incentives have a demonstrable effect on response rates. As did Church, they also found that cash incentives are more effective than gifts or tokens. Moreover, the authors found incentives to be effective regardless of the degree of burden on the respondent (e.g., 20 minutes vs. an hour interview).

The work reported by Singer et al. (1999) is also applicable to panel respondents. It suggests that informing the respondent at the start of the study that they will be offered an incentive for participation increases cooperation at later waves. Thus, at the conclusion of each interview, and before collecting the data that will help track the respondent at subsequent waves, it is useful to inform respondents that they will be paid an incentive for their future participation. However, it is important to emphasize that the respondent is not being paid for the data just provided. Rather, the incentive is a token of appreciation for having provided something that is invaluable and therefore cannot be purchased.

4. Summary

Panel retention is an under-appreciated but very important part of longitudinal studies that involve re-interviewing a cohort of respondents over time. While participation in a panel is a burden for the respondent, maintaining an unbiased panel is a challenge for the investigator. Thus, before such a project is undertaken planning is critical and the level of effort and associated costs need to be considered from the outset.

A key aspect of a successful panel is the study design. The sample, questionnaire, and associated budget must all accommodate the need to recruit an unbiased panel and maintain their participation over a lengthy period of time. Recruitment involves attracting individual respondents to agree to participate in the baseline and to be followed for some period of time afterwards. Locating respondents is critical to ensuring the continued quality of the panel data. Finally, recruiting and maintaining the panel will be facilitated if respondents develop a commitment to it. Ensuring that the panel is an unbiased sample of the population of interest requires effort to not only convince individuals to participate but to provide sufficient information to enable the panel to continue.

Achieving a high level of cooperation to the baseline is critical and obtaining cooperation must be done in a manner that does not alienate the respondent. It must also be recognized that overall trends in survey research response rates indicate that collecting any survey data collection is going to become increasingly difficult. In the USA, there has been sufficient decline in response rates to social surveys since the 1990s that the leading statistical agencies and professional associations that rely on survey data convened a special conference devoted to the topic (the International Conference on Survey Nonresponse in Portland, Oregon, October 1999).

In sum, plans for follow-up must be established at the onset of the study so that all the relevant information is collected from the respondents as necessary resources are allocated. Without sufficient data on respondents, when they change residence or

telephone, they are likely to be lost to follow-up, introducing bias into the data. Convincing respondents that they are important to the success of the panel and that the investigator values their participation ultimately helps reduce the cost of locating and following the respondents. Overall, participation in research is a process of exchange. Thus, it is important to show respect and consideration for the sources of the data.

See also: External Validity; Internal Validity; Longitudinal Data; Panel Surveys: Uses and Applications; Selection Bias, Statistics of; Statistical Data, Missing

Bibliography

Belinfante A2000 *Telephone Subscribership in the United States*. Federal Communications ommission, Washington, DC

Church A H 1993 Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly* **57**: 62–79

Cook T D, Campbell D 1979 *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton-Mifflin Company, Geneva, IL

Cottler L B, Compton W M, Ben-Abdallah A, Horne M, Claverie D 1996 Achieving a 96.6 percent follow-up rate in a longitudinal study of drug abusers. *Drug and Alcohol Dependence* **41**(3): 208–17

Faber C S 2000 *Geographic Mobility: March 1997 to March 1998. Current Population Reports P20-520*. US Census Bureau, Washington, DC

Given B A, Keilman L J, Collins C, Given C W 1990 Strategies to minimize attrition in longitudinal studies. *Nursing Research* **39**(3): 184–6

Groves R, Couper M 1998 *Nonresponse in Household Interview Surveys*. John Wiley and Sons, New York

Gunderson T 1996 *How to Locate Anyone Without Leaving Home*. Plume, New York

Hoyer F, McCann J 1988 *Find Them Fast, Find Them Now*. Citadel Press, Secaucus, NJ

Hunt J R, White E 1998 Retaining and tracking cohort study members. *Epidemiologic Reviews* **20**: 57–70

Parco V 1994 *Researching Public Records*. Citadel Press, New York

Parsons J A, Owens L, DeGrush A, Johnson T 1999 Using a dual frame design to increase response rates: Results of two experiments. Paper presented at the. In: *54th Annual Conference of the American Association for Public Opinion Research*. St. Petersburg Beach, FL

Singer E, Van Howeyk J, Gebler N, Raghunathan T, McGonagle K 1999 The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics* **15**(2): 217–30

Traugott M W, Groves R M, Lepkowski J M 1987 Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion Quarterly* **51**: 48–57

United States Post Office 1999 *Address Information Systems: Products and Services*, Publication 40

Walton M A, Ramanathan C S, Reischel T M 1998 Tracking substance abusers in longitudinal research: Understanding follow-up with difficulty. *American Journal of Community Psychology* **26**: 233–53

Wright J D, Allen T L, Devine J 1995 Tracking non-traditional populations in longitudinal studies. *Evaluation and Program Planning* **18**(3): 267–77

R. B. Warnecke and J. A. Parsons

Long-term Depression (Cerebellum)

While long-term depression (LTD) occurs under various conditions and plays various roles (*Long-term Potentiation and Depression (Cortex)*; *Long-term Depression (Hippocampus)*), the cerebellar LTD examined in this article is characterized by its association with a unique structure in the cerebellum, Purkinje cells and climbing fibers, and by its distinct role played in cerebellar function.

1. Neuronal Elements Involved in LTD

Mossy fibers arising from various precerebellar nuclei in the brain stem and spinal cord enter the granular layer of the cerebellar cortex and form synaptic contact with granule cells there. Axons of granule cells ascend through the Purkinje cell layer to the molecular layer and bifurcate to parallel fibers, which extend 2–3 mm to each side of the bifurcation point through the extensive dendritic arbor of Purkinje cells. Estimates of the number of synapses that each Purkinje cell receives from parallel fibers vary from 60,000 to 175,000 (Ecles et al 1967, Palay and Chan-Palay 1974, Ito 2001). While parallel fibers make up the majority of granule cell inputs to Purkinje cells, the ascending axons of granule cells also contribute about 20 percent of the total granule cell inputs, preferentially on the smallest diameter distal regions of the Purkinje cell dendrites.

Another input to Purkinje cells, a climbing fiber, originates from the inferior olive in the medulla oblongata. In contrast to the convergence of numerous granule cell axons to a Purkinje cell, only one climbing fiber comes into contact with each Purkinje cell and forms numerous discrete synaptic junctions on stubby dendritic spines of the Purkinje cell. The number of junctions formed between a climbing fiber and a Purkinje cell has previously been estimated to be approximately 300 in the frog cerebellum. However, in rats, a much larger number (26,000) is suggested because a 100 μm length of PC dendrite contacts 11.45 granule cell axons and 1.7 climbing fibers on the average (Ito 2001).

2. Electrical Signals Involved in LTD

Impulses of granule cell axons evoke fast excitatory postsynaptic potentials (EPSPs) in Purkinje cells as mediated by AMPA receptor. These EPSPs, when sufficiently large, evoke Na^+ spikes in the somatic region, which passively spread into the dendrites. During extracellular recording, these spikes are observed as 'simple spikes,' which are spontaneously discharged *in vivo* at a rate of 50–100 Hz. Impulses in climbing fibers evoke responses composed of large EPSPs and Ca^{2+} spikes evoked by the EPSPs in Purkinje cell dendrites. During extracellular recording, climbing fiber responses take the form of short bursts of spikes, called 'complex spikes,' which are discharged *in vivo* at irregular, low rates around 1 Hz.

During *in vivo* extracellular recording, LTD is represented by the persistent reduction in the probability (firing index) of a Purkinje cell to generate simple spikes in response to the stimulation of granule cell axons. Under slice or tissue culture conditions, LTD is observed as a persistent reduction in the granule cell axon-evoked AMPA-mediated EPSPs or excitatory postsynaptic currents (EPSCs) (see Eccles et al. 1967, Ito 2000).

3. Conditions for Induction of LTD

LTD is optimally induced by conjunctive stimulation of climbing fibers and granule cell axons at 1–4 Hz by 100–300 pulses. LTD can also be induced by a more natural stimulus paradigm such as a series of eight pulses (80 msec, 100 Hz) to granule cell axons and immediately thereafter three pulses to climbing fibers (Ito 2000). LTD is input-specific; it occurs only in those granule cell–Purkinje cell synapses conjunctively stimulated by climbing fibers. However, spread of LTD to synapses in the close neighborhood has recently been reported (Reynolds and Hartell 2000, Wang et al 2000).

Various reduced forms of LTD have been generated by replacing either granule cell axon stimulation or climbing fiber stimulation, or both, with chemical stimuli or electrical currents directly applied to Purkinje cells. In cerebellar slices, climbing fiber stimulation can be replaced by the application of depolarizing pulses, which bring about the entry of Ca^{2+} ions into Purkinje cells through voltage-gated channels. The simplest procedure for inducing LTD is to apply quisqualate or to increase the extracellular K^+ concentration. In cultured Purkinje cells devoid of both granule cells and climbing fibers, a reduced form of LTD is induced by a combination of glutamate (or quisqualate) pulses and membrane depolarization. The LTD is detected as a reduced sensitivity of PCs to glutamate or AMPA. This form of LTD occurs in very simple preparations devoid of dendritic spines, indicating that LTD induction does not require such morphological specialization.

4. Duration of LTD

The observation time for LTD is usually limited to one half to one hour, and occasionally for two to three hours, but in certain special cases, it has been extended to one to two days (Ito 2000). LTD induced in the amplitude of mEPSCs in cultured PCs by conjunctive application of 50 mM K^+ and 100 μM glutamate was observed to last for 36 hours and return to the original level after 48 hours. When the rabbit cerebellum was sliced 24 hours after the rabbit had been trained for eyeblink conditioning (*Eyelid Classical Conditioning*), sequentially applied granule cell axon and climbing fiber stimuli failed to induce LTD, which occurred in slices dissected from control rabbits. This suggests that LTD underlying the eyeblink conditioning persists for at least 24 hours and precludes eliciting another LTD.

5. Signal Transduction for LTD

Diverse chemical reactions have been found to underlie LTD induction (Levenes et al. 1998, Ito 2000). They can be roughly classified into the following seven major pathways:

(a) The glutamate released from granule cell axons reacts with AMPA receptors in Purkinje cells, which open associated cation channels for Na^+ and K^+ ions to generate AMPA-EPSPs. These EPSPs act to open voltage-dependent Ca^{2+} channels.

(b) Granule cell-derived glutamate also reacts with mGluR1s, which in turn activate phospholipase C through G protein (Gαq). Phospholipase C produces diacylglycerol (DAG) which activates protein kinase C (PKC), as well as inositol trisphosphate (IP_3), which acts upon IP_3 receptors to release Ca^{2+} ions from intracellular stores.

(c) mGluR1s activated by granule cell-derived glutamate also stimulate phospholipase A2 via G proteins, eventually producing unsaturated free fatty acids including arachidonic acid and oleic acid.

(d) Granule cell-released NO diffuses into Purkinje cells and activates guanylyl cyclase which produces cyclic GMP (cGMP). cGMP in turn activates protein kinase G (PKG), which, following phosphorylation, converts a protein, G-substrate, into a potent inhibitor of protein phosphatase.

(e) A climbing fiber-derived amino acid transmitter (most likely glutamate, yet unidentified) reacts with receptors and opens associated channels for Na^+ and K^+ ions to evoke a large EPSP, which in turn causes Ca^{2+} influx through voltage-dependent Ca^{2+} channels.

(f) A climbing fiber-derived corticotropin-releasing factor (CRF) reacts with type-1 CRF receptors and associated G proteins, and then yields an as yet unidentified second messenger, which interacts with PKC.

(g) A climbing fiber-derived type-1 insulin-like growth factor (IGF-1) may facilitate internalization of receptors by endocytosis.

There is cross-talk between these pathways. Protein tyrosine kinases (PTKs), mitogen-associated protein kinases (MAPKs), glial fibrillary acid protein (GFAP), an as yet unidentified rapidly turned-over protein(s) and immediate early gene, *Jun*-B, are also implicated in LTD induction. A recent hypothesis is that these complex chemical reactions eventually lead to removal of AMPA receptors by internalization (Wang and Linden 2000, Matsuda et al. 2000).

6. *Error Representation by Climbing Fiber Signals*

What information do climbing fibers convey to induce LTD? In simple situations such as reflexes, climbing fibers convey sensory signals such as pain or loud sounds. These sensory signals suggest a harmful consequence of an inadequately executed movement. In various types of ocular movements, climbing fiber signals are known to represent retinal slips, which reflect a deviation of a realized eye movement from a desired one, that is, an error in controlling eye movement. These observations support the view that, during the repeated exercise of a movement, error-driven LTD reshapes the neuronal circuit of the cerebellum in the direction which minimizes errors consequent to the movement (Ito 1984, 2001).

During locomotion, climbing fiber signals are evoked by a sudden descent of a stepped-on rung at the beginning, but not at the cessation, of the fall of a paw. These signals appear to reflect the discrepancy between command signals descending from the brain stem center to the spinal cord and the signals that the command signals evoke in the spinal cord neuronal circuit. During hand/arm movements, climbing fiber signals arise in three phases of movement: at the beginning, before the end, and after. Climbing fiber signals therefore encode not only the consequence errors detected through sensory systems, but also other types of errors arising from the internal mechanisms of the motor systems.

7. *Involvement of LTD in Motor Learning*

Recent data that the inhibition of LTD following pharmacological or genetic manipulation impairs various types of motor learning supports the view that error-driven LTD is the basis of motor learning (Ito 2001).

(a) Adaptation of the vestibuloocular reflex (233) has been shown to be blocked by superfusing the cerebellum in rabbits and monkeys with a NO scavenger, hemoglobin, or by injection of an inhibitor of NO synthase (NOS) into the goldfish cerebellum. Transgenic mice that selectively express the pseudo-substrate PKC inhibitor in Purkinje cells lacked the vestibuloocular reflex adaptation consistent with the loss of LTD induction in cerebellar slices obtained from these mice. Neural NOS (nNOS)-deficient mice lacked adaptation of the optokinetic eye movement response, which normally occurs under continued rotation of the visual field around a stationary animal. Subdural applications of NO scavenger or NOS inhibitor to the paraflocculus-flocculus depressed the adaptation of smooth pursuit in monkeys.

(b) Rabbits administered a NO synthase inhibitor exhibited learning deficits in the conditioned eyeblink response. GFAP-deficient mutant mice exhibited a loss of LTD and also impaired eyeblink conditioning. Injection of an IGF-1 antisense oligonucleotide in the inferior olive, which reduces IGF-1 levels in the cerebellum, blocked conditioned eyeblink learning in rats.

(c) Postural compensation after unilabyrinthectomy was delayed in rats after intracerebroventricular application of PKC inhibitors. When a decerebrate cat walking on a treadmill experienced a sudden increase in the speed of only the running belt under the left forelimb, regular stable locomotion was restored in 50–100 steps. Injection of a NOS inhibitor into the cerebellum blocked this adaptation. mGluR1-deficient mice walking on a treadmill exhibited abnormally dispersed locomotor cycles of two limbs and did not adapt to an increase in the belt velocity. This sign of impaired interlimb coordination in locomotion diminished when the mGluR1-deficiency was rescued in the cerebellum.

(d) Motor coordination is evaluated by measuring the duration that an animal can remain on a horizontally placed rotating rod. In five types of gene-manipulated mice showing significant reduction in staying time, three types (mGluR1-, GluRδ2-, and Gαq-deficient) lacked LTD. However, the other two (PKCγ- and mGluR4-deficient) retained LTD, and their motor discoordination is explained by the persistent multiple innervations of Purkinje cells by climbing fibers, abnormal transmission in granule cell–Purkinje cell synapses. GFAP-deficient mice and PKC-inhibitor-transfected mice were reported to exhibit motor coordination indistinguishable from wild-type mice in spite of the lack of LTD. These observations are inconsistent with the hypothesis of motor learning. The reasons for this discrepancy need to be investigated.

8. *Comments*

How is LTD eventually converted to permanent memory? One possibility is that LTD as a functional depression is consolidated as a structural change in synaptic contacts and spines. How can the error-

driven LTD-based learning mechanism be expanded to apply beyond motor learning to implicit learning in mental functions such as cognition, language, and thought? This expansion has been achieved at experimental and conceptual levels, but in the future it should be made at computational levels as well.

See also: Cerebellum: Associative Learning; Cerebellum: Cognitive Functions; Long-term Depression (Hippocampus); Long-term Potentiation and Depression (Cortex); Memory: Synaptic Mechanisms; Neural Plasticity

Bibliography

Eccles J C, Ito M, Szentagothai J 1967 *The Cerebellum as a Neuronal Machine*. Springer, Berlin

Ito M 1984 *The Cerebellum and Neural Control*. Raven, New York

Ito M 2000 Cerebellar long-term depression: Characterization, signal transduction and functional roles. *Physiology Review* **81**: 1143–95

Levenes C, Daniel H, Crepel F 1988 Long-term depression of synaptic transmission in the cerebellum: Cellular and molecular mechanisms revisited. *Progress in Neurobiology* **55**: 79–91

Matsuda S, Launey T, Mikawa S, Hirai H 2000 Disruption of AMPA receptor GluR2 clusters following long-term depression induction in cerebellar Purkinje neurons. *EMBO Journal* **19**: 2765–74

Wang Y T, Linden D 2000 Expression of cerebellar long-term depression requires postsynaptic clathrin-mediated endocytosis. *Neuron* **25**: 635–47

Palay S L, Chan-Palay V 1974 *The Cerebellar Cortex*. Springer, New York

Reynolds T, Hartell N A 2000 An evaluation of the synapse specificity of long-term depression induced in rat cerebellar slices. *Journal of Physiology* (*London*) **527**: 563–77

Wang S S H, Khiroug L, Augustine G J 2000 Quantification of spread of cerebellar long-term depression with chemical two-photon uncaging of glutamate. *Proceedings of the National Academy of Science, USA* **97**: 8635–40

M. Ito

Long-term Depression (Hippocampus)

Hippocampal LTD is a form of activity-dependent synaptic plasticity where the cumulative activation of inputs to specific neural pathways in the hippocampus produces a decrease in the excitability of the activated neurons. Several forms of hippocampal LTD have been identified, and much research has examined the mechanisms by which LTD is expressed in the hippocampus. In this article, a characterization of hippocampal LTD and its possible involvement for memory storage in the mammalian brain will be examined.

1. Synaptic Plasticity in the Hippocampus

Synaptic plasticity is a term used to describe several forms of long-lasting, activity-dependent changes in synaptic strength. The importance of synaptic plasticity is due to the presumption that information, or memories, are stored in the brain as complex spatial-temporal patterns of neuronal activity that occur in large ensembles of neurons (see *Memory: Synaptic Mechanisms*; *Neural Plasticity*). Learning, defined as any type of experience-dependent change in sensory processing or experience-dependent change in behavior, is thought to involve the storage of new information via a process that involves long-lasting changes in synaptic strength in the circuits that were activated during the learning process (see *Learning and Memory, Neural Basis of*). Synaptic plasticity is also important for the development of neural circuitry, because an initial activity-independent phase occurs during which axons grow into their target regions and produce an overabundance of synaptic contacts followed by an activity-dependent phase during which some of the synapses are strengthened and consolidated, while other synapses are weakened and pruned away (see *Synapse Formation*).

During the early 1970s, the discovery of long-term potentiation (LTP), a long-lasting enhancement of synaptic activity that follows a brief, high-frequency electrical stimulation, provided strong experimental evidence for a model of synaptic plasticity associated with a number of important functions of the nervous system, including learning and memory (see *Long-term Potentiation (Hippocampus)*). When LTP was first identified, it was most reliably and robustly elicited in the hippocampus, an integral component of the medial-temporal lobe system that mediates a number of very important forms of declarative memory (Squire 1987). LTP can be reliably elicited in intact *in vivo* preparations, as well as in reduced *in vitro* preparations, making it amenable to detailed experimental manipulation and analysis.

If synapses in the hippocampus simply continued to increase in strength as a result of LTP, all synapses would eventually reach some level of maximum efficacy. At this point, it would be difficult or impossible, at least in principle, to encode new information. Moreover, active synapses would be so plentiful that even a small stimulus would excite many cells, resulting in pathologies like epilepsy. To make synaptic strengthening a useful mechanism for encoding information, other processes must selectively

weaken specific sets of synapses. LTD in the hippocampus may be an example of just such a process.

2. Hippocampal Anatomy and Circuitry

The hippocampus provides a useful model system for the study of synaptic plasticity on both anatomical and electrophysiological grounds. The hippocampus is divided into two major subpopulations of neurons: the hippocampus proper, composed primarily of pyramidal cells (subdivided into four subfields: CA1–CA4), and the dentate gyrus, composed primarily of granule cells (see *Hippocampus and Related Structures*).

A series of discrete, anatomical units called lamellae oriented along the transverse axis of the structure compose the hippocampus (Andersen et al. 1971). It is this lamellar organization that allows for the precision in identification of known stimulation and recording sites within the hippocampal trisynaptic circuit. Stimulation of perforant path fibers of the entorhinal cortex, which provides the main input into the lamellae of the hippocampus, results in the activation of granule cells of the dentate gyrus. Stimulation of the granule cells, in turn, activates the pyramidal cells of the CA3 region through the mossy fiber system. The axons of the CA3 pyramidal cells bifurcate, sending an efferent stimulus through their main branch out through the fimbria to the fornix, and a collateral branch (Schaffer collaterals) sweeping upward to synapse on the apical dendrites of CA1 pyramids. The sequential activation of this trisynaptic circuit has been verified both anatomically and electrophysiologically (Swanson et al. 1982).

The Schaffer collateral-CA1 pyramidal cell synapse is one of the best understood synapses in the mammalian brain, and many of the currently known phenomena of synaptic plasticity, including short-term potentiation, LTP, paired pulse facilitation, and LTD, have been found to occur at this synapse.

3. Two Forms of Hippocampal LTD

Two basic forms of hippocampal LTD have been distinguished. Homosynaptic LTD occurs only when the strengths of active synapses are depressed, e.g., homosynaptic LTD can occur at synapses that are given low-frequency electrical stimulation. Heterosynaptic LTD occurs only when the strengths of inactive synapses are depressed, e.g., heterosynaptic LTD can occur at synapses that are inactive during high-frequency stimulation of a converging synaptic input. Heterosynaptic LTD was discovered in the dentate gyrus *in vivo* in 1979, with evidence for homosynaptic LTD in area CA1 reported in 1992. In this article, hippocampal LTD will be described by first focusing on homosynaptic LTD in area CA1, followed by a discussion of heterosynaptic LTD in the dentate gyrus.

3.1 Homosynaptic LTD in area CA1

Homosynaptic LTD has been studied *in vitro* in the excitatory monosynaptic pathway running from the Schaffer collaterals (afferents) to area CA1 cells (see *In Vitro Neurophysiology*; *Perception and Cognition, Single-/Multi-neuronal Recording Studies of*). In the paradigm most often used to induce homosynaptic LTD, the delivery of several hundred electrical stimuli to the Schaffer collaterals at low frequencies produces a sustained depression of modest but significant magnitude in area CA1 cells. The key properties of homosynaptic LTD produced by low- frequency stimulation are summarized below (see Bear and Abraham 1996, Dudek and Bear 1992, Mulkey and Malenka 1992):

(a) Low frequency stimulation-induced LTD is input specific. Only synapses receiving low-frequency conditioning stimulation show LTD; other synaptic inputs converging onto the same population of postsynaptic neurons are unaffected (hence the term homosynaptic LTD).

(b) Low-frequency stimulation-induced LTD is frequency dependent. LTD depends on the frequency of conditioning stimulation. LTD usually results from 0.5–3 Hz stimulation, but the same number of pulses at higher frequencies can produce LTP instead.

(c) Low-frequency stimulation-induced LTD is saturable and reversible. Repeated episodes of low-frequency stimulation in CA1 saturate LTD. Furthermore, synapses that are depressed can be potentiated, and vice versa, indicating that LTD is not a result of lasting damage to the stimulated synapses.

(d) Low-frequency stimulation-induced LTD usually requires *N*-methyl-D-aspartate (NMDA) receptor activation. Under most circumstances, NMDA receptor antagonists reversibly block homosynaptic LTD.

3.2 Heterosynaptic LTD in Dentate Gyrus

Heterosynaptic LTD has been studied *in vivo* in the excitatory monosynaptic pathway running from layer II cells of the entorhinal cortex via the perforant path to the granule cells of the dentate gyrus; it has not yet been observed *in vitro*. Like LTP in the hippocampus, heterosynaptic LTD is saturable and reversible, and, depending on stimulation parameters, can last for days or weeks. In most experiments, the induction of heterosynaptic LTD in one pathway is accompanied by LTP of the path that was tetanized. This fact, plus the failure to observe this form of LTD *in vitro*, has

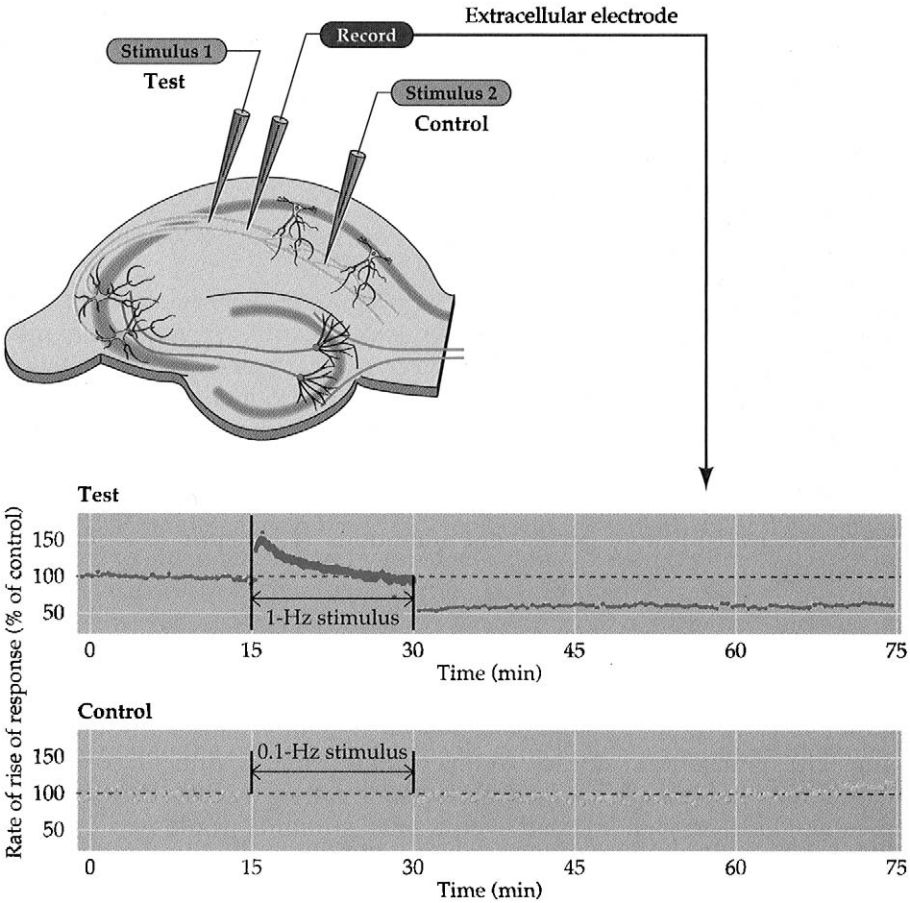

Figure 1
Long-term synaptic depression in the hippocampus (reproduced by permission of Sinauer Associates, Inc. from *Neuroscience*, 1997). In this experiment, two stimulating and one extracellular recording electrode are used. Low-frequency stimulation (1 per second) is applied to the test pathway (stimulus electrode 1). After 15 minutes, the size and slope of the extracellular population response are depressed by about 50 percent (upper panel), while the control pathway (which was stimulated once every 10 seconds) remains unchanged (lower panel). The depression persists for the duration of the experiment

restricted investigations of the mechanisms specifically underlying heterosynaptic LTD.

4. *Mechanisms of Hippocampal LTD Induction*

In most cases, hippocampal LTD requires activation of glutamate receptors and postsynaptic calcium (Ca^{2+}) ion influx at the site of synaptic contact in the dendritic spine (see *Ion Channels and Molecular Events in Neuronal Activity*; *Synaptic Transmission*). Biochemically, hippocampal LTD also appears to require activation of protein phosphatases, which may be preferentially activated by Ca^{2+} influx. Interestingly, the magnitude of synaptically-evoked changes in postsynaptic Ca^{2+} concentration appears to dictate whether LTP or LTD is elicited.

4.1 *Receptor Activation*

Synaptically-released glutamate can activate several types of glutamate receptors, presumably co-localized at individual synaptic contacts at individual dendritic spines (see *Neurotransmitters (Glutamate)*). One of these is the NMDA receptor; another is the α-amino-3-hydroxy-5-methyl-4-isoxazoleproprianate (AMPA) receptor. During basal synaptic transmission, the current that generates hippocampal synaptic responses is due to activation of the AMPA receptor, which is associated with sodium-potassium channels. The channels open briefly, sodium ions rush in, and the excitatory postsynaptic potential develops. This glutamate-AMPA transmitter-receptor is the prototypic fast excitatory synapse in the brain.

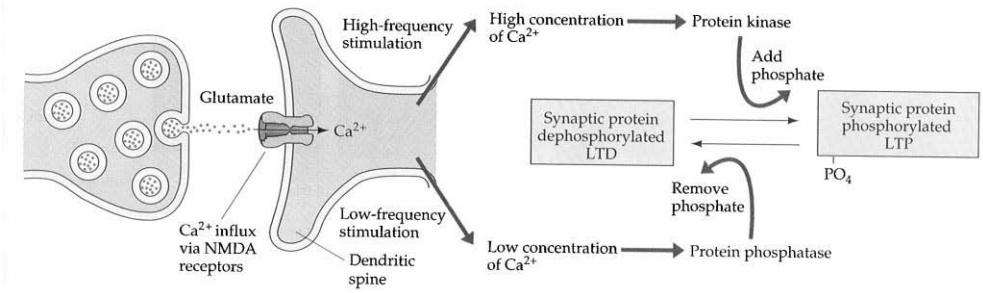

Figure 2
A model for the induction of LTP and LTD (reproduced by permission of Sinauer Associates, Inc. from *Biological Psychology*, 1996). During afferent activity, Ca^{2+} enters dendritic spines through NMDA receptors. High-frequency stimulation causes phosphorylation (via protein kinase) of a synapse protein, and low-frequency stimulation causes dephosphorylation (via protein phosphatase) of the same synaptic protein. The state of phosphorylation of this protein helps to control synaptic strength

Although the NMDA receptor is a glutamate receptor, it is not activated by weak synaptic actions like that of the AMPA receptor, and therefore contributes very little to changes in the membrane potential during basal synaptic transmission. However, when the cell is depolarized by various patterns of synaptic activation, magnesium ions leave the channels, and glutamate activation of the NMDA receptors opens the channels allowing Ca^{2+} to rush into the neuron because unlike AMPA receptors, NMDA receptors are highly permeable to Ca^{2+}. There is strong evidence that the NMDA receptor-mediated change in Ca^{2+} is a necessary, and perhaps sufficient trigger for both LTP and LTD. Hippocampal homosynaptic LTD is reversibly blocked by the NMDA receptor antagonist DL-2-amino-5-phosphonovaleric acid, and is also blocked by strong hyperpolarization during the hyperpolarization protocol or buffering postsynaptic Ca^{2+} (Bear and Malenka 1994, Kemp and Bashir 1997, Mulkey and Malenka 1992). This evidence suggests that NMDA receptor-dependent changes in Ca^{2+} might be the primary mechanism by which hippocampal LTD is mediated.

While much has been learned about the mechanisms of NMDA-dependent hippocampal LTD, there is some additional evidence that another type of glutamate receptor, the metabotropic glutamate receptor (mGluR), might also play a key role in establishing LTD, particularly during early postnatal life (Bolshakov and Siegelbaum 1994). This receptor would be classed as having a slow or modulatory action: it does not directly control ion channels, but controls a second-messenger system via a G protein that acts on the intracellular biochemical machinery of the cell. One of the primary functions of mGluRs in synaptic transmission is to provide activity-dependent autoinhibition through presynaptic suppression of transmitter release (Glaum and Miller 1994). Activation of mGluRs results in reversible depression in the strength of synaptic input to principal neurons in the trisynaptic hippocampal circuitry, and in the presence of mGluR antagonists, LTD is inhibited, as well as mRNA translation, which suggests that protein synthesis, triggered by synaptic activation of mGluR, can modify synaptic transmission (Huber et al. 2000).

4.2 Biochemical Activation

Activity-dependent modulation of synaptic strength in area CA1 pyramidal cells is at least partially controlled by the balance between the activities of critical protein kinases and phosphatases in the postsynaptic cell, with the synaptic glutamate receptor being a potential substrate for this biochemical competition. In one model for the induction of LTP and LTD, during efferent activity, Ca^{2+} enters dendritic spines through NMDA receptors. High-frequency stimulation allows Ca^{2+} to reach high levels, which preferentially activates a protein kinase. Low-frequency stimulation allows Ca^{2+} to reach low levels, which preferentially activates a protein phosphatase. It is then argued that both the protein kinase and phosphatase act on a common synaptic phosphoprotein, the phosphorylation state of which controls synaptic strength (Bear and Malenka 1994).

See also: Hippocampus and Related Structures; In Vitro Neurophysiology; Ion Channels and Molecular Events in Neuronal Activity; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Long-term Depression (Cerebellum); Long-term Potentiation and Depression (Cortex); Long-term Potentiation (Hippocampus); Neurotransmitters; Perception and Cognition, Single/

Multi-neuronal Recording Studies of; Short-term Memory: Psychological and Neural Aspects; Synaptic TransmissionSynaptic Transmission.

Bibliography

Andersen P, Bliss T V P, Skrede K K 1971 Lamellar organization of hippocampal excitatory pathways. *Experimental Brain Research* **13**: 222–38

Bear M F, Abraham W C 1996 Long-term depression in hippocampus. *Annual Review of Neuroscience* **19**: 437–62

Bear M F, Malenka R C 1994 Synaptic plasticity: LTP and LTD. *Current Opinion in Neurobiology* **4**: 389–99

Bolshakov V Y, Siegelbaum S A 1994 Postsynaptic induction and presynaptic expression of hippocampal long-term depression. *Science* **264**: 1148–52

Dudek S M, Bear M F 1992 Homosynaptic long-term depression in area CA1 of hippocampus and effects of N-methyl-D-aspartate receptor blockade. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 4363–67

Glaum S R, Miller R J 1994 Acute regulation of synaptic transmission by metabotropic glutamate receptors. In: Conn P J, Patel J (eds.) *The Metabotropic Glutamate Receptors*. Humana, Totowa, NJ

Huber K M, Kayser, M S, Bear M F 2000 Role for rapid dendritic protein synthesis in hippocampal mGluR-dependent long-term depression. *Science* **288**: 1254–6

Kemp N, Bashir Z I 1997 NMDA receptor-dependent and independent long-term depression in the CA1 region of the adult rat hippocampus in vitro. *Neuropharmacology* **36**: 397–9

Mulkey R M, Malenka R C 1992 Mechanisms underlying induction of homosynaptic long-term depression in area CA1 of the hippocampus. *Neuron* **9**: 967–75

Purves D, Augustine G J, Fitzpatrick D, Katz L C, LaMantia A S, McNamara J 1997 *Neuroscience*. Sinauer Associates, Inc., Sunderland, MA

Rosenzweig M R, Leiman A L, Breedlove S M 1996 *Biological Psychology*. Sinauer Associates, Inc., Sunderland, MA

Squire L R 1987 *Memory and Brain*. Oxford University Press, Oxford

Swanson L W, Teyler T J, Thompson R F 1982 Hippocampal long-term potentiation: Mechanisms and implications for memory. *Neuroscience Research Program Bulletin* **20**: 613–769

M. R. Foy

Long-term Potentiation and Depression (Cortex)

Long-term potentiation (LTP) is a long-lasting increase in postsynaptic response seen at glutamate synapses following afferent activation. In the laboratory, LTP typically is induced by a brief tetanus to afferents and results in a larger postsynaptic response to a constant input that lasts for an hour or more. Long-term depression (LTD) is a long-lasting decrease in the postsynaptic response seen at the same synapses. LTD is induced by a low-frequency tetanus leading to a postsynaptic response depression lasting for an hour or more. Both LTP and LTD are alterations in synaptic gain that display an induction threshold, are specific to the synapses activated, are induced by Ca^{2+} entering the postsynaptic cell, and display associativity—such that the pairing of a strong and weak (unable to induce change alone) stimulus will induce gain change in both inputs. Both forms of synaptic plasticity are thought to be central to the formation of neural networks underlying information storage. The commonly accepted model of the storage of experiential information (e.g., memory) involves the establishment of networks of neurons interconnected in such a way to encode the salient aspects of the experience. These networks—which may be quite extensive and involve thousands of neurons—are themselves the neural representation of experience. LTP and LTD are found in brain regions and at developmental times that suggest that they may encode far more than ongoing experience, perhaps being generic synaptic mechanisms involved in the establishment of fundamental cortical sensory and motor representation, among other things. Cortical LTP is most pronounced during early development coinciding with the critical period for environmentally-induced synaptic plasticity, suggesting that the same mechanism may underlie both phenomena.

1. Cortical Distribution

LTP/LTD are found throughout the neocortex at glutamatergic synapses of pyramidal cells and interneurons. One or both of these forms of synaptic plasticity have been reported in primary sensory areas, primary motor areas, and 'association' cortex of rodent, cat, and primate brain, as well as cortical projections to subcortical structures (thalamus, striatum). While there are some differences between cortical areas, generally both LTP and LTD are seen at all locations, as in hippocampus (but unlike cerebellar Pyrkinje cell synapses where LTD predominates). As opposed to the hippocampus, where most studies of synaptic plasticity are performed, the complex anatomy of cortex makes it difficult to stimulate monosynaptic, homogeneous afferent pathways. Most studies, either *in vitro* or *in vivo*, tetanize the white matter afferents and record the postsynaptic effects in corresponding cortical column, or stimulate and record intracortically within the column (layer IV/V to II/III or layer II to V). In both cases, stimulation activates a heterogeneous population of afferents and antidromically activates efferents traversing the region, making interpretation of the synapses involved

problematic. Similarly, the postsynaptic responses are often combinations of orthodromic mono- and polysynaptic responses and antidromic responses.

2. Mechanisms of nmdaLTP

There are two different forms of cortical LTP: one mediated by activation of the NMDA glutamate receptor subtype, the other mediated by activation of Voltage Dependent Calcium Channels (VDCCs). Both cortical forms appear to have the same characteristics as LTP found in hippocampus, namely: glutamate receptor subtype or VDCC involvement, input specificity, associativity, activation threshold and Ca^{2+} dependency. The induction of the form of LTP mediated by the NMDA receptor (nmdaLTP) requires afferent activity associated with postsynaptic depolarization via the AMPA receptor (a ligand-gated glutamate receptor gating Na^{+}/K^{+}) sufficient to activate the ligand and voltage sensitive NMDA receptor (NMDAR). The active NMDAR gates Ca^{2+} into the postsynaptic cell where it activates several serine–threonine kinases (PKA, PKC, CaMKII) leading to the phosphorylation of the AMPAR—resulting in larger EPSPs, and the activation of incompletely understood signal transduction cascades (involving the CREB and Ras-MAPK cascades) ultimately leading to an alteration in gene expression/protein synthesis. Most of the studies investigating the signaling pathways in LTP have been done in hippocampal neurons using both enzyme inhibitors, and developmental or targeted/inducible genetic knock-out animals. The relatively few neocortical studies indicate that the same mechanisms are present. The activated kinases, some of which are autophosphorylated as a result of their activation (thus ensuring their continued activity in the absence of the calcium transient), serve to phosphorylate their substrates. Many of the receptors, channels, and enzymes present postsynaptically have phosphorylation sites which, when activated, result in increases in postsynaptic responsivity. The expression of LTP immediately following a tetanus is primarily due to phosphorylation of the AMPAR. As will be seen later, some phosphorylation sites on the AMPAR and other substrates are reversible and can be dephosphorylated through the action of phosphatases associated with LTD. It is thought that the persistence of LTP involves synthesis of synaptic proteins, as enzyme inhibitors and protein synthesis inhibitors limit the duration of LTP expression (in hippocampus).

3. Mechanisms of vdccLTP

The form of LTP mediated by VDCCs (vdccLTP) is initiated by depolarizing the postsynaptic cell beyond that required for nmdaLTP—to activate L-type VDCCs and allow an influx of Ca^{2+} into the postsynaptic cell. This means that either the degree of afferent activity must lead to a higher level of postsynaptic depolarization or that other elements must assist in the establishment of appropriate levels of depolarization to induce this form of LTP. There are two known means by which cellular processes contribute to the induction of vdccLTP. The first is through the simultaneous induction of nmdaLTP which contributes depolarization to the postsynaptic cell and thus helps to achieve the conditions for the activation of VDCCs. The second is by reducing the inhibitory drive onto the postsynaptic cell to facilitate the induction vdccLTP. As compared to hippocampus, neocortical LTP appears to be regulated to a greater extent by enhanced inhibitory tone, as GABAa blockers are often required to elicit either form of cortical LTP *in vitro*. The cellular targets of Ca^{2+} gated through VDCCs appears to be different from Ca^{2+} gated through NMDAR channels, in that enzyme inhibitors of serine/threonine kinases that block nmdaLTP have no affect on the induction or expression of vdccLTP. vdccLTP depends upon the activation of tyrosine kinases, and involves the participation of tyrosine receptor kinases (*trk*)—the receptors for neurotrophins. In visual cortex of young rats, the application of the neurotrophin BDNF facilitates the induction of synaptic plasticity to a weak tetanus. As opposed to nmdaLTP which can be reversed

Table 1
Comparison of the features of the three forms of synaptic plasticity observed at neocortical glutamatergic synapses. *Depolarization* and *Calcium levels* refer to approximate values achieved during the initiating tetanus

	nmdaLTD	*nmdaLTP*	*vdccLTP*
Induction tetanus	1–5 Hz	> 25 Hz	> 50 Hz
Depolarization	~ 4 mV	~ 10 mV	~ 18 mV
Calcium levels	~ 6 nM	~ 13 nM	~ 20 nM
Duration	hours	days–weeks	?
Reversability	by LTP	by LTD	no
Signaling	Serine–threonine phosphatases	Serine–threonine kinases	Tyrosine kinases

(depotentiated) by LTD induction, vdccLTP appears to be irreversible. Since both forms of LTP are induced by afferent activity leading to postsynaptic depolarization, sufficient activity will lead to the expression of both forms simultaneously.

The two forms of LTP have somewhat different cortical distribution, with nmdaLTP found at intracortical synapses and associated with white matter activation early in development. In adult cortex, NMDARs decline in number and nmdaLTP seen to white matter stimulation is reduced or absent, leaving vdccLTP dominant. A similar, but less pronounced, effect is seen in hippocampus as well.

4. *Long-term Depression*

LTD is the decremental counterpart of LTP whose mechanism of action also utilizes Ca^{2+} gated through NMDARs. The evidence from Ca^{2+} imaging studies suggests that the amount of Ca^{2+} entering the postsynaptic cell is critical in determining which form of synaptic plasticity will be expressed. LTP induction seen following high frequency afferent activity is associated with greater postsynaptic cytosolic Ca^{2+} levels than that seen following prolonged low frequency activation leading to LTD. These low levels of calcium in the postsynaptic cell bind to high-affinity Ca^{2+} binding sites on protein phosphotases, whose action is to dephosphorylate the substrates of serine/threonine kinases. Thus, LTD acts to dephosphorylate sites that have been phosphorylated by the action of protein kinases (the removal of nmdaLTP by LTD induction is termed depotentiation). Thus, nmdaLTP and LTD (or phosphorylation and dephosphorylation) exist in a opposing regulatory relationship.

5. *Properties of Plastic Cortical Synapses*

Studies in hippocampus have revealed several other forms of synaptic plasticity that have not been reported for neocortex. These include a presynaptic form of LTP at hippocampal CA3 synapses and a form of LTP utilizing opiate transmitters. Given the relative paucity of mechanistic studies of LTP in neocortex, it is unclear if these forms are absent or undiscovered. Two additional aspects of LTP seen in hippocampus are present in at least some regions of neocortex: the influence of the metabotropic glutamate receptor (mGluR) and the presence of the retrograde messenger NO. The mGluRs are a family of G-protein linked glutamate receptors whose different effects are mediated by PKC activation and IP3-mediated release of Ca^{2+} from postsynaptic endoplasmic reticuluum stores. Various mGluR antagonists and knock-outs have demonstrated the requirement for mGlu involvement in LTP in both hippocampus and neocortex, although it's precise role remains elusive. There may be regional cortical differences however, as some mGluR antagonists are ineffective in blocking PI hydrolysis and synaptic plasticity in visual cortex. As in hippocampal pyramidal cell synapses, there is evidence for the involvement of the retrograde messenger NO in neocortical synapses, as inactivation of its synthetic enzyme blocks LTP in auditory and somatosensory cortex.

As changes in synaptic gain are initiated by incremental increases in postsynaptic depolarization (with increasing depolarization, nmdaLTD, then nmdaLTP, then vdccLTP are induced), modulatory synaptic influences onto cortical pyramidal cells can be expected to influence the form and degree of synaptic plasticity induced by glutamatergic activity. Studies primarily in visual cortex have demonstrated: the inhibiting effect of mACh antagonists on LTP, the inhibiting effect of 5-HT1A agonists on visual cortex LTP, and the dopamine facilitation of intracortical prefrontal cortex LTD. As different cortical regions have differential neuromodulatory inputs, it might be expected that they will influence synaptic plasticity in different directions. More detailed studies in hippocampus have revealed that no two synapse types share identical properties. While they are outwardly similar, they differ mechanistically. The same may be found across neocortical synapses.

In sum, neocortical synaptic plasticity is phenomenologically and mechanistically similar to that seen in the better studied archicortex (hippocampal formation). Differences are seen primarily in the regional distribution of forms of LTP between cortical areas and within cortical laminae and the relative dominance of the VDCC form of LTP, particularly its prevalence with maturity. LTP/LTD remain the best candidates for processes that alter the synaptic gain at nodes in neural networks encoding information—and thus defining the architecture of the network. If the forms of synaptic plasticity seen in neocortex underlie information storage, it must be demonstrated that they are specific and necessary to the process they govern. Beyond this, consideration must be given to the nature of the information that is stored, how it is reaccessed, and how it integrates with other brain functions.

See also: Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Long-term Depression (Cerebellum); Long-term Depression (Hippocampus); Long-term Potentiation (Hippocampus); Short-term Memory: Psychological and Neural Aspects; Synaptic Transmission

Bibliography

Aroniadou V A, Keller A 1995 Mechanisms of LTP induction in rat motor cortex in vitro. *Cerebral Cortex* **5**: 353–62

Bear M F 1996 A synaptic basis for memory storage in the cerebral cortex. *Proceedings of the National Academy of Science USA* **93**: 13453–9
Bear M F, Kirkwood A 1993 Neocortical long-term potentiation. *Current Opinion Neurobiology* **3**: 197–202
Kirkwood A, Dudek S M, Gold J T, Aizenman C D, Bear M F 1993 Common forms of synaptic plasticity in the hippocampus and neocortex in vitro. *Science* **260**: 1518–21
Lisman J E, Fallon J R 1999 What maintains memories? *Science* **283**: 339–40
Singer W 1995 Development and plasticity of cortical processing architectures *Science* **270**: 758–64

T. J. Teyler

Long-term Potentiation (Hippocampus)

Since Ramon y Cajal, neurobiologists have been looking for activity-dependent modifications of synaptic efficacy as a potential mechanism for learning and memory. When Bliss went to Per Andersen's laboratory in 1968, he probably did not know he was going to discover a phenomenon that would become one of the most popular, controversial, and debated one in Neuroscience until the end of the twentieth century. The first publication (Bliss and Lomo 1973), reporting the existence of long-term potentiation (LTP) in rabbit brain, was noticed by only a small number of laboratories. Even the demonstration that LTP could be elicited in the *in vitro* hippocampal slice preparation (Lynch et al. 1977, Andersen et al. 1977) remained somewhat of a curiosity until the end of the 1970s. It was not until the mid-1980s and the identification of the unique properties of the N-Methyl-D-Aspartate (NMDA) receptors and of the links between LTP and learning and memory (Morris et al. 1986) that LTP began to attract very large numbers of investigators. This review covers the major advances that have taken place since the 1960s, and gives the reader a sense for what is well understood, what remains speculative, and what remains to be understood.

1. *The NMDA Receptor and LTP Induction*

The cellular events that are necessary and sufficient to produce LTP are now well understood (Malenka and Nicoll 1999). What is required is sufficient postsynaptic depolarization (i.e., the high frequency bursts of stimulation) to relieve the NMDA receptor/channel from a voltage-dependent magnesium blockade and to produce a postsynaptic influx of calcium at the right place and of the right amplitude. The NMDA receptor thus functions as an AND gate (presynaptic release of neurotransmitter and postsynaptic depolarization) and accounts for most features of LTP, such as associativity, calcium dependency, regulation by the degree of inhibition, etc. It should be noted however, that some forms of LTP do not require NMDA receptor activation, and thus there may be multiple forms of LTP as well as multiple ways of triggering the biochemical cascades leading to LTP.

2. *The Pre- vs. Post-synaptic Debate*

It was clear from the outset that LTP could be due to changes in either the release characteristics of the presynaptic terminals or the postsynaptic responsiveness to the neurotransmitter or some combination of both factors. As some reports suggested that glutamate release is enhanced following LTP induction (Bliss 1990), the postsynaptic localization of NMDA receptors imposed the existence of some retrograde signals that could link LTP induction to LTP expression mechanisms. Several such signals were proposed such as arachidonic acid, nitric oxide (NO), or carbon monoxide (CO) (O'Dell et al. 1991). On the other hand, many reports have also indicated several postsynaptic modifications including changes in the properties of another type of glutamate receptors, the α-amino-3-hydroxy-5-methylisoxazole propionic acid (AMPA) receptors, and in the morphology and number of dendritic spines (Maren et al. 1993; Lee et al. 1980). There is now some general agreement that LTP expression is due to several modifications of the distribution of AMPA receptors and the structure of synaptic contacts (Malenka and Nicoll 1999).

3. *The Biochemical Confusion*

Since the 1980s, a very large number of biochemical events have been proposed to participate in one way or another in LTP induction or expression (Sanes and Lichtman 1999). Among these, several calcium-dependent processes have been identified. Activation of the calcium-dependent protease, calpain, results in modifications of the structure of several cytoskeletal proteins and cell adhesion molecules and therefore is ideally suited to promote modifications in the structure of synaptic contacts (Baudry 1991). Activation of several protein kinases, and in particular, calcium/calmodulin kinase type II (CamKII), is also a key process to phosphorylate key proteins in synaptic contacts, including glutamate receptors (Kennedy 1989). Autophosphorylation of CamKII might also provide for a form of short-term plasticity that exceeds the duration of the LTP-inducing stimulus. Finally, activation of phospholipases provides for both the synthesis of lipid breakdown products with various signaling properties as well as for the breakdown of membrane lipids, an essential element for structural

modifications (Massicotte and Baudry 1991). It is interesting to point out that most of these cell biological processes are not unique to neurons but are present in many cell types where they participate in regulation of structure and function.

4. *Gene Expression and LTP Stabilization*

A much-debated issue in the LTP literature concerns the role of gene expression and protein synthesis and the mechanisms linking LTP induction to the protein synthesis machinery of neurons. Numerous studies have attempted to identify genes activated as a result of LTP induction (Abraham et al. 1993, Thomas et al. 1994). In most cases, several transcription factors are induced as a result of trains of electrical activity producing LTP. However, it has been quite difficult to further establish the relationship between gene induction with mechanisms underlying LTP expression and maintenance.

Advances in molecular biology techniques, and in particular, the possibility of knocking out or knocking in specific genes have produced a large number of mutant mice with various 'learning and memory' impairments, in addition to problems with LTP induction or maintenance (Silva et al. 1997). In addition, the mutations that have been generated often affect enzymes, receptor/channels, or other cellular elements, which participate in 'normal' cellular functions, and are likely to modify, even in a subtle way, various aspects of brain structure and function. The limitation on the role of gene expression in synaptic plasticity resulting from the cell-wide modifications it generally produces would be seriously eliminated if there were mechanisms coupling gene expression and local translation of the corresponding mRNAs. For local protein synthesis to take place requires the dendritic targeting of mRNA and there are now a number of examples of such mRNAs. Furthermore, there is also experimental evidence for links between synaptic transmission and regulation of local protein synthesis as activation of glutamate metabotropic receptors in synaptoneurosomes has been shown to rapidly stimulate protein synthesis (Weiler and Greenough 1993). There are therefore two types of mechanisms by which synaptic activity can produce localized modifications of synaptic properties restricted to activated synapses. In the first one, a small population of mRNAs is constitutively present in dendrites, and in response to an appropriate signal, these mRNAs are locally translated, and the newly synthesized proteins are incorporated into the activated synapses. Note that this process is rapid and does not require gene expression, as it only depends on constitutively present mRNAs. In the second one, some locally generated signals are transmitted to the cell nucleus, trigger gene transcription, and the synthesis of mRNAs. These mRNAs are targeted to the dendrites and locally translated at the appropriate site. Note that this requires the existence of a signal or signals that remain present at activated synapses for quite some time as the transcription of genes, and their targeting to the dendrites necessitate a significant period of time. This mechanism is somewhat related to the synaptic tagging described by Frey and Morris (1998). In their model, these authors propose that LTP induction is accompanied by the activation of signals they call 'synaptic tags,' which interact with plasticity-related newly synthesized proteins and stabilize the formation of long-term potentiation. Clearly, much more work is needed to clarify the exact mechanisms involved in this type of regulation, and to determine the types of mRNAs present in the dendrites, the mechanisms involved in targeting them to dendrites and the signals responsible for tagging activated synapses. Finally, the types of proteins involved and their roles in regulating synaptic structure and function need to be further investigated.

It has only recently been recognized that the extracellular matrix plays an equally important role in regulating synaptic morphology and properties. Antibodies to N-CAM and peptide inhibitors of integrin receptors inhibit the formation of stable long-term potentiation (LTP) (Luthi et al. 1994, Lynch 1998). In several current models of LTP, stabilization of the modifications triggered during the induction phase of LTP is accomplished first by the disruption of the adhesive properties of synaptic contacts, possibly mediated through the proteolysis of CAM molecules, followed by the activation of integrins. In this way, synaptic contacts undergo a cycle of destabilization resulting from intracellular as well as extracellular proteolysis of proteins contributing to the morphology of synaptic contacts, followed by restabilization of a new synaptic structure implicating integrin activation and extracellular matrix components (Lynch 1998).

5. *LTP and Learning and Memory*

LTP is the most widely proposed mechanism of memory storage in the hippocampus and neocortex. Although this issue is still debated, evidence supporting this hypothesis comes from a variety of experimental data and theoretical models (Baudry and Davis 1996). LTP is prevalent in hippocampal and cortical networks and exhibits many properties required for a large capacity information storage device: rapid induction, associativity, long duration, links with brain rhythms (in particular, the theta rhythm). Pharmacological manipulations or gene mutations interfering with LTP also interfere with various forms of learning and memory, while pharmacological agents facilitating LTP formation also facilitate learning. Finally, incorporating LTP-based rules in biologically realistic neuronal networks produces large capacity storage devices.

6. Conclusions

Much progress has been made on our understanding of the mechanisms and functions of LTP since its discovery in 1968. At the beginning of the twenty-first century there might still be a few unanswered questions, but there is little doubt that (a) these questions will receive definite answers rapidly and (b) the knowledge gained from the study of this phenomenon will be instrumental in designing new tools to study and improve various forms of learning impairments.

See also: Learning and Memory: Computational Models; Long-term Depression (Cerebellum); Long-term Depression (Hippocampus); Long-term Potentiation and Depression (Cortex); Memory Models: Quantitative

Bibliography

Abraham W C, Mason S E, Demmer J, Williams J M, Richardson C L, Tate W P, Lawlor P A, Dragunow M 1993 Correlations between immediate early gene induction and the persistence of long-term potentiation. *Neuroscience* **56**: 717–27

Andersen P, Sundberg S H, Sveen O, Wigstrom H 1977 Specific long-lasting potentiation of synaptic transmission in hippocampal slices. *Nature* **266**: 736–37

Baudry M 1991 An integrated biochemical model for long-term potentiation. In: Baudry M, Davis J L (eds.) *Long-term Potentiation: A Debate of Current Issues*. MIT Press, Cambridge, MA, pp. 169–82

Baudry M, Davis J L eds. 1996 *Long-term Potentiation*. MIT Press Cambridge, MA, vol. 3

Bliss T V P 1990 Memory maintenance is presynaptic. *Nature* **346**: 698–99

Bliss T V P, Lomo T 1973 Long-lasting potentiation of synaptic transmission in the dentate area of the anesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* **232**: 334–56

Frey U, Morris G M 1998 Synaptic tagging: implications for late maintenance of hippocampal long-term potentiation. *Trends Neurosci*. **21**: 181–88

Kennedy M B 1989 Regulation of synaptic transmission in the central nervous system: long-term potentiation. *Cell* **59**: 777–87

Lee K S, Schottler F, Oliver M, Lynch G 1980 Brief bursts of high-frequency stimulation produce two types of structural change in rat hippocampus. *Journal of Neurophysiology* **44**: 247–58

Luthi A, Laurent J P, Figurov A, Muller D, Schachner M 1994 Hippocampal long-term potentiation and neural cell-adhesion molecules L1 and NCAM. *Nature* **372**: 777–79

Lynch G 1998 Memory and the brain: unexpected chemistries and a new pharmacology. *Neurobiological Learning & Memory* **70**: 82–100

Lynch G S, Dunwiddie T, Gribkoff V 1977 Heterosynaptic depression: a postsynaptic correlate of long-term potentiation. *Nature* **266**: 737–39

Malenka R C, Nicoll R A 1999 Neuroscience long-term potentiation—a decade of progress? *Science* **285**: 1870–74

Maren S, Tocco G, Standley S, Baudry M, Thompson R F 1993 Postsynaptic factors in the expression of long-term potentiation (LTP): increased glutamate receptor binding following LTP induction *in vivo*. *Proceedings of the National Academy of Science* **90**: 9654–58

Massicotte G, Baudry M 1991 Triggers and substrates of hippocampal synaptic plasticity. *Neurological Biobehavioral Review* **15**: 415–23

Morris R G M, Anderson E, Lynch G S, Baudry M 1986 Selective impairment of learning and blockade of long-term potentiation by an N-Methyl-D-Aspartate receptor antagonist, AP-5. *Nature* **319**: 774–76

O'Dell T J, Hawkins R D, Kandel E R, Arancio O 1991 Tests of the roles of two diffusible substances in long-term potentiation: evidence for nitric oxide as a possible early retrograde messenger. *Proceedings of the National Academy of Science* **88**: 11285–89

Sanes J R, Lichtman J W 1999 Can molecules explain long-term potentiation? *Nature Neuroscience* **2**: 597–604

Silva A J, Smith A M, Giese K P 1997 Gene targeting and the biology of learning and memory. *Annual Review of Genetics* **31**: 527–46

Thomas K L, Laroche S, Errington M L, Bliss T V P, Hunt S P 1994 Spatial and temporal changes in signal transduction pathways during LTP. *Neuron* **13**: 737–45

Weiler I J, Greenough W T 1993 Metabotropic glutamate receptors trigger postsynaptic protein synthesis. *Proceedings of the National Academy of Science* **90**: 7168–71

M. Baudry

Lorenz, Konrad (1903–89)

1. Biography

1.1 Family

Konrad Zacharias Lorenz (born November 7, 1903 in Vienna, Austria) grew up within the upper crust of pre-World War I Vienna: his mother Emma (née Lecher) from a family of writers and scientists; his father, Adolf Lorenz, world-renowned orthopedic surgeon, Nobel prize nominee (A. Lorenz 1936). Konrad spent his formative years in Vienna and at his family's country estate in Altenberg on the banks of the Danube. At that time the river was still lined by large tracts of virgin riverine forest, and his inquisitive mind became fascinated by nature. He became a passionate observer and skilled caretaker of any swimming, crawling, or flying creature. These interests did not interfere, however, with his early socialization into the network of learned relatives and their friends (e.g., Karl Popper; see *Popper, Karl Raimund (1902–94)*).

1.2 Education

After visiting a 'progressive' elementary school, he was sent to the esteemed *Schotten-Gymnasium*, where his biology teacher, an enthusiastic Darwinist, introduced

him to evolutionary reasoning. At his father's suggestion, he began to study medicine at Columbia University, New York in 1922. Homesick, especially for his wife-to-be, he continued his studies in Vienna, receiving his Dr.med. degree in 1928. During his first appointment as *Demonstrator* (~ assistant professor) at the department of anatomy at the University of Vienna, he studied zoology (J. Versluys), paleontology (O. Abel), and psychology (K. Bühler, his assistant E. Brunswik); he received his Dr. phil. degree in 1933.

1.3 Scientific Career

Lorenz's meager salary as *Demonstrator* and his wife's (Margarethe née Gebhardt) income as physician at a local hospital provided just enough to raise their family. Lorenz was always short of funds for the private research facility he had established in Altenberg. His negotiations with the (German) Kaiser-Wilhelm-Gesellschaft for funds to create a special institute for behavioral studies date back to 1937, but at the outbreak of World War II this project was shelved. In 1940 E. Baumgarten (a nephew of Max Weber (see *Weber, Max (1864–1920)*) and a pupil of J. Dewey (see *Dewey, John (1859–1952)*) and, hence, a representative of the pragmatist school of philosophy) accepted the famed chair of philosophy at Königsberg University, once occupied by Immanuel Kant (see *Kant, Immanuel (1724–1804)*), and looked for a biologically oriented psychologist who, at the same time, was interested in theory of knowledge, to share with him the prestigious appointment. E. von Holst became the match-maker, and Lorenz was chosen (Lorenz 1974, Leyhausen 2000).

One year later, however, Lorenz was drafted into the German Army as a motorcycle rifleman (in his papers he had not mentioned his medical training because of his limited experience with patients, and stressed only his proficiency in motorcycle racing and repair). When the authorities discovered that he held a medical degree and was a professor of psychology, he was assigned to a unit of the *Heerespsychologie* (Army Psychology) in Posen. Two months later, however, this unit was dissolved, and he was transferred to the Posen Army field hospital for neurology and psychiatry, where he worked as *Unterarzt* (medical noncommissioned officer) in a unit of 285 beds with patients suffering from war injuries, psychoses, and psychosomatic disorders. During a field assignment on the Russian front in 1944 he was captured by the Red Army. After that, he served in various POW camps as field surgeon until his release in 1948.

In post-World War II Austria career prospects for a biologist were worse than ever, while in Britain Lorenz's colleagues became especially interested in the Darwinian approach to behavioral studies. Niko Tinbergen became reader at Oxford University and, in 1950, Lorenz received a similar offer from the University of Bristol. German colleagues, notably E. von Holst and G. Kramer, both meanwhile directors at the newly founded Max Planck Institute for Marine Biology (Wilhelmshaven), were able to revive the old project of an institute for behavioral studies. A matching offer was made, and Lorenz accepted the appointment as head of a small behavior research unit in Buldern, Westfalia, founded in spring of 1951 as precursor to a special Max Planck Institute for Behavioral Physiology. Construction started in Seewiesen, Bavaria in 1955, and the institute opened in 1958: Lorenz was Director of his own division there until he retired in 1973.

1.4 Retirement

Upon retirement, Lorenz returned to his home in Altenberg, Austria. Shortly thereafter, he was awarded, together with Niko Tinbergen and Karl von Frisch, the 1973 *Nobel Prize for Medicine*. During retirement, he tidied up his work on evolutionary epistemology, on the social behavior of greylag geese and reef fishes, and became an eloquent spokesman for environmental conservation (Schleidt 1988). Lorenz died in Vienna on February 27, 1989 at the age of 85 and was buried near Altenberg at St. Andrä-Wördern cemetery (Wuketits 1990).

2. Contribution to Knowledge

Lorenz, cofounding with Niko Tinberger the science of Ethology as 'the study of the biology of organismic behavior', ranks among the most important stewards of Darwin's (see *Darwin, Charles Robert (1809–82)*) heritage of the twentieth century. In the 1930s and 1940s, most psychologists, sociologists, and biologists tried to understand behavior within the framework of Aristotelian (see *Aristotle (384–322 BC)*) philosophy, theology, Marxism (see *Marx, Karl (1818–89)*), or Behaviorism. Lorenz, on the other hand, viewed behavior as playing a central role in the process of evolution, and shaped primarily by 'natural selection', rather than by learning. By applying the methods of comparative morphology and paleontology to the study of 'species specific' behavior, he created a scientific approach which rapidly provided a wealth of data and new insights (Lorenz 1931, 1932, 1950a, 1965b, 1978). Contrary to the claim of his critics, Lorenz ignored neither the conventional schools of thought nor their observations, but was cautious in accepting their theories and conclusions. For example, he never denied the importance of environmental factors in learning, but he challenged the behaviorists' claim that innate behavior was restricted to only a few basic vital reflexes. In fact, he emphasized especially

the interlocking of modular innate and learned behavioral (e.g., as in 'imprinting,' Lorenz 1935, and in his analysis of the channels through which information can enter a living system, Lorenz 1961, 1965b).

2.1 Observation as Research Tool

Fundamental to Lorenz's approach to behavior is his rigorous adherence to the doctrine of *induktive Naturwissenschaft* (inductive natural science; Lorenz 1942, 1959, 1978), i.e., first establish a solid base with meticulous observations, supported by objective records (photos, film), tainted as little as possible by theories, and—only after a solid data base (*Induktions-Basis*) has been established—let patterns emerge from the data.

2.2 Behavioral Modules

Lorenz considered behavior to be composed of basic modules: e.g., classical reflexes, species specific *Instinktbewegung* 'Fixed Action Pattern' (FAP), in later years *Erbkoordination*, *erbkoordinierte Bewegung* (1978) (in English: 'fixed motor pattern'), specific circuits for the environmental control of reflexes and FAPs: 'Innate Releasing Mechanisms' (IRM, *angeborener Auslösemechanismus*), 'appetites and aversions' (*Stimmungen*), etc. Such modules provide each organism with basic behavioral skills to navigate within its environment, detect necessary resources, avoid danger, and interact with conspecifics. Learning can enter in various ways at the level of each of these modules, e.g., in the case of a reflex as 'conditioned reflex.'

This FAP/IRM diad is basically an extension of the classical Stimulus/Response concept, enhanced by several new features: the classical reflex, activated only by a specific external stimulus, cannot account for the observation that 'a healthy animal is up and doing' (as W. McDougall so aptly remarked). Consequently, the spontaneous nature of behavior, in general, and especially in many typical FAPs (locomotion, courtship behavior, certain types of bird song, etc.) became a topic of special interest (FActionP). FAPs constitute much more complex patterns than any of the classical reflexes (FAPatterns) and particular spatiotemporal parameters of FAPs are highly stereotyped within any one species (FixedAP), but different between species. Lorenz exemplified the complex interaction between internal and external variables in his 'psychohydraulic' model (1950a, 1978), and applied this reasoning to explain the control of aggression in animals and humans (1963).

It was basically for this work on behavior modules, which are central to ethology, that Lorenz, Niko Tinbergen, and Karl von Frisch shared the distinction of winning the 1973 Nobel Prize for Medicine 'for their discoveries concerning the organization and elicitation of individual and social behavior patterns.'

2.3 Ethograms

Given the existence of species specific behavior, a behavioral study of individuals of any given species starts with an assessment of the modules of its behavioral repertoire, working towards the open-ended task of a complete 'ethogram' of that species. Lorenz had already touched on this in his first publication (1931), and a decade later published an exemplary comparative study of the courtship modules of ducks (1941). He delayed publishing of his ethogram of the graylag goose until 1988 … as a sort of swan-song.

2.4 The Innate Forms of Possible Experience

This heading is a translation of *Die angeborenen Formen möglicher Erfahrung*, the title of Lorenz's 1943 seminal paper dealing with the evolution of cognition, as a new paradigm. The results of many years of research into 'learning'—from the classical curves of learning and forgetting by I. Pavlov and E. L. Thorndike (see *Thorndike, Edward Lee (1874–1949)*)—indicate that it is a slow process. If an animal's survival depended solely on learning, it would be dead long before it had completed its first set of learning trials. Thus, it should not be surprising that an animal, entering into its world, is equipped with (a) certain 'innate' skills, behavior patterns to satisfy vital needs and escape danger, and (b) '*a priori* knowledge' to detect essential features of resources and danger —innate in the sense that particular patterns of behavior or relevance of particular environmental features are available prior to individual experience. In other words: a reflex, FAP, or IRM not only enables survival enhancing action, but it also constitutes organismic 'knowledge' about the environment and its contingencies, to which that individual's ancestors had been exposed in the evolutionary process (Lorenz 1941, 1943, 1953, 1965b, 1973, 1978).

2.5 Imprinting

This peculiar process of forming an irreversible association between a behavior pattern and a 'stimulus' (object), without any reward other than the presence of that object, has been used for ages in animal breeding, and noted in the literature, but Lorenz was the first to recognized imprinting's far-reaching theoretical significance (1935, 1965b, 1978).

2.6 Nature–Nurture

Much ink has been spilled on this duality, not much because of any intellectual challenge posed by these two concepts, but because of deeply rooted beliefs, religious or political, touching upon free will, accountability, and fate. Lorenz was attacked early on for his use of 'innate' as though he had excluded the possibility of learning, while all he suggested was that innate behavior was essential for the survival of individual organisms. Early on, he pointed to the intertwining of innate and learned behaviors (e.g., *Trieb-Dressur-Verschränkung*, drive-training-interlacing 1935), and with imprinting, provided a splendid example of such a process, which was ignored by the ruling psychologists of his time. He was attacked relentlessly for defending the concept of innateness, despite numerous attempts to respond to his critics (1961, 1965b, 1978, etc.).

2.7 Epistemology

'Philosophy is the academic abuse of those human potentials which have evolved by natural selection for the individual's survival in his struggle with the daily contingencies of his world.' This definition, attributed to Lorenz's mentor O. Heinroth, encapsulates the insight that human ability to acquire knowledge is not merely a matter of the reasoning individual *Homo sapiens* within a literate society, but is based on a long, evolutionary process, preceding the emergence of humanity by millions of years. Lorenz, having been elevated to Immanuel Kant's former chair in Königsberg in 1940, and facing the challenges by contemporary philosophy, and specifically by resident neo-Kantians, took a radically new approach to Kant's concept of *a priori* (Lorenz 1941, 1943), pioneering what is now called Evolutionary Epistemology (Popper 1972, Vollmer 1980). The most detailed discussion of this is contained in his posthumously published *Russian Manuscript* 1944–48 (1992), written with the fresh memories of his encounter with Kant's Königsberg in mind, while a more concise discussion can be found in the English edition of his textbook (1978/81).

2.8 Moral Analogous Behavior, Human Morality, and Environmental Awareness

Lorenz, throughout his life, struggled with his belief that morality could well have an evolutionary basis (as he had shown convincingly for the evolutionary basis of cognition in his evolutionary epistemology). He coined the term 'moral-analoguous behaviour' (moral-analoges Verhalten): special behavior that channels the individual's selfish struggle for its individual existence and reproductive success into an altruistic contribution for the good of its community. Based on his observations of a variety of highly social animals, Lorenz noticed that domesticated forms had lost certain elements of their moral-analoguous social behaviors which their wild ancestors had evolved as essential part of their sociality. Lorenz concluded, by analogy, that human evolution in general, and especially our own modern civilization, has led to comparable domestication effects ('self-domestication'), which have become a mixed blessings. Modern humans have freed themselves from the firm guide of instincts and consequently have more freedom to learn, and to make their own choices, but they also have lost some of their instinctive inhibitions to inflict damage to, and even kill conspecifics to an extent not known in any other species (1940, 1943, 1963, 1983).

In his early papers Lorenz proposed to counter these dangers to humanity with eugenic measures (before the attrocities of Hitler's Germany a widely respected field, especially in England and the USA), especially by encouraging ethically guided (not ethnically, as often misread) mate choice (1940, 1943). Starting in the mid-1950s, Lorenz's proposal of an evolutionary basis of human moral were attacked for various reasons or dismissed as unfounded. As competing hypotheses—especially those claiming sole environmental control of ethical behavior—have failed so far to resist falsification, these issues are, on the scientific level, far from resolved.

Lorenz's concern about human evolution running awry and leading us to extinction by our destruction of the environment was not merely a consequence of the atrocities of World War I, World War II, and of the nuclear demise scenarios of the Cold War, but had been deeply rooted in both his trust in science and in his allegiance to the Hippocratic oath. Thus, when by the vagaries of political and economic interests a nuclear power plant was built near his home in Altenberg, at a location that had been flawed by previously undisclosed risks, he opposed completion of its construction violently. Soon he found himself the revered leader in this fight that ultimately prevented the key ready nuclear power plant from going into service. A few years later, Lorenz was equally successful in a fight to save the last stretch of wetlands along the Danube.

3. Relevance

Lorenz's impact on behavioral and social science is linked closely to his collaboration and friendship with Niko Tinbergen, and it is difficult to measure each one's individual contribution. Both were curious naturalists endowed with outstanding skills: in observing, in detecting regularities—patterns of nature—and in formulating testable hypotheses; and, both were outstanding teachers (e.g., Lorenz 1949, 1950b).

While Tinbergen was more inclined toward an experimental approach, as exemplified in their joint paper (Lorenz and Tinbergen 1938), Lorenz's imprint on ethology is obvious in matters of pure observation, inductive reasoning, and formulating a comprehensive theory: the modular concept of action patterns and releasing mechanisms, these modules being shaped by evolution as well as by learning, and imprinting as a special type of learning.

3.1 Selected Topics

Lorenz achieved his scientific breakthrough in 1936. At an invited lecture before the German scientific elite of the Kaiser-Wilhelm-Gesellschaft in Berlin he proposed to stop arguing about how many drives and instincts one can define and instead start describing and analyzing behavior patterns as though they were anatomical features. In a discussion following the lecture, Max Hartmann, doyen of German biology, remarked that 'Konrad Lorenz's approach opens up to causal analysis a field that was previously just a playground for fruitless philosophical speculation' (Lorenz 1992).

World War II severely restricted research and exchange. But soon after the war had ended, Tinbergen traveled in the UK and the USA, presenting the new, ethological approach to behavioral studies, his effort culminating in his book *The Study on Instinct* (1951). Lorenz, who, up to then, had established an international reputation only among ornithologists, argued the case for a biological approach to the study of behavior and soon became the leading theoretician in the new science of ethology.

The high esteem for ethological research resulted in generous support by the Max Planck Gesellschaft, and subsequently by various national funding agencies, notably National Institutes of Health in the USA. Small informal meetings soon turned into international affairs. The biennial *International Ethological Conferences*, dominated by the Lorenz–Tinbergen School and 'converts' from a variety of related fields, became the central event for discussing new results and theories. The German journal *Zeitschrift für Tierpsychologie*, founded in 1937 with Lorenz as one of its editors, the first journal to favor ethological publications, soon got an English language companion in the Dutch journal *Behaviour*. A few years later *Animal Behaviour* was introduced, as the official organ of Animal Behavio(u)r Societies, founded in the US and UK, as publications in the new science steadily increased. While *animal psychology* and *comparative psychology* had been little known specialties, dealing mainly with the learning abilities of rats, the Lorenz–Tinbergen-inspired new *animal behavior* became a hot topic for lectures in academic departments both in biology and psychology, resulting in a demand for teachers and textbooks (e.g., Eibl-Eibesfeldt 1967/99), and spawning new opportunities and funding for behavioral research.

Since Lorenz, Tinbergen' and their colleagues were not only qualified biologists, but equally well-established in neighboring fields, the popularity of ethology soon led to fruitful cooperation, notably with sensory physiology, brain sciences, and psychiatry (e.g., Lorenz 1953). This process had considerable impact on the emergence of neurobiology, psychobiology, and evolutionary psychiatry, emphasizing a naturalistic, Darwinian approach at a time when the general trend in biology was toward molecular biology and biochemistry.

Subjects of ethological research at first were mainly vertebrates—birds, mammals, and fishes—and insects. The new approach was tested succesfully and applied to all kinds of species, including our own. Even though the magnitude and difficulties of a task like establishing a human ethogram may appear mind-boggling, 'human ethology' becomes workable as a special case of primate ethology (e.g., Eibl-Eibesfeld 1997). A healthy human newborn faces similar tasks as any other primate baby and behaves very similarly.

By applying ethological methods within the framework of natural science and evolutionary reasoning, similarities and differences between human and non-human animals are no longer a matter of poetic license, but causal analysis—not, however, without drawing heavy critique from the old religious and political establishment. *On Aggression* and *The Waning of Humaneness* (Lorenz 1963, 1983) are examples of an ethological approach to the moral strongholds of the humanities.

Human Ethology has become a domain of its own, of broad overlap not only with primatology, but anthropology (physical and cultural), psychology, and sociology, leading toward a new synthesis of knowledge which proves that the old distinction between hard sciences and humanities has become obsolete.

3.2 Politics

Fame has its price. Especially since having been awarded the Nobel Prize, Lorenz has been branded a racist and an anti-Semite and alleged to have provided a pseudoscientific justification for the Holocaust (e.g., Klopfer 1999).

In regard to these allegations, we must consider first that Lorenz grew up within a part of the Viennese society that was enmeshed intimately by friendship, marriage, and professional ties with the wealthy Jewish elite. He maintained these connections throughout his life by hosting many guests and being a guest on many in his frequent travels of old and new friends especially in England and the USA, and was certainly more pro-Semitic than average, and far less racist, especially

than some of those who are now raising their voice. Thus, the charge that he advocated racist and anti-Semitic genocide is unfounded.

With respect to the allegation that Lorenz provided a pseudoscientific justification for the Holocaust, it must be considered that there is a long chain of inferences necessary to connect between his—always controversial—writings on the potentially dangerous effects of human self-domestication on the biological basis or ethical (not, and never ethnical!) behavior or Lorenz's proposed use of eugenics to combat the rise of unethical behavior, and the atrocities of the Third Reich. And, there is no indication—inspite of many inquiries and searches (summarized in Klopfer 1999) that prior to the end of World War II anybody in power was aware of such a link. Lorenz being called to the army early in World War II, his first assignment as a rifleman and later as medical NCO (never promoted to a higher rank) give ample evidence that he lacked political pull (Schleidt 2000).

Lorenz became active in politics only after his retirement. Given all the academic and public honors he had received, people listened when he spoke—as he proudly proclaimed—and he spoke loudly and clearly for environmental preservation and moral responsibility.

4. Concluding Remarks

When in the mid-1970s sociobiology made the headlines and behavioral ecology started riding the new wave of environmental awareness, the glamor of ethology began to fade. In 1999, the crown jewel of ethology, the Max Planck Institute for Behavioral Physiology at Seewiesen, was downgraded to an ornithological research station. However, the ideas conceived and publicized by Lorenz, Tinbergen, and their contemporaries as the foundations of ethology have become so much public property that their original sources are now rarely cited. This may be the best indicator of the importance of the work of Konrad Lorenz, and of ethology.

See also: Animal Cognition; Behaviorism, History of; Birdsong and Vocal Learning during Development; Comparative Psychology; Darwinism: Social; Evolutionary Epistemology; Psychological Development: Ethological and Evolutionary Approaches; Sensitive Periods in Development, Neural Basis of

Bibliography

Eibl-Eibesfeldt I 1999 *Grundriss der vergleichenden Verhaltensforschung—Ethologie*, 8th edn. Piper, Munich/Zürich [1970 *Ethology—the Biology of Behavior*. Holt, Rinehart & Winston, New York]

Eibl-Eibesfeldt I 1997 *Die Biologie des menschlichen Verhaltens*, 4th edn. Piper, Munich/Zürich [1989 *Human Ethology*. Aldine de Gruyter, New York]

Klopfer P H 1999 Konrad Lorenz and the national socialists: On the politics of ethology. *International Journal of Comparative Psychology* **7**: 202–8

Leyhausen P 2000 Das Königsberger [The year in Königsberg. In: Kotrschar K, Müller G, Winkler H (eds.) *Konzepte der Verhaltensforschung—Konrad Lorenz und seine Folgen*. Elander Vertag, Fürth

Lorenz A 1936 *My Life and Work*. Scribner, New York

Lorenz K 1940 Durch Domestikation verursachte Störungen arteigenen Verhaltens. *Zeitschrift für angewandte Psychologie und Charakterkunde* **59**: 2–81

Lorenz K 1941 Kants Lehre vom Apriorischen im Lichte gegenwärtiger Biologie. *Blätter für Deutsche Philosophie* **15**: 94–125 [1962 Kant's doctrine of the *a priori in* the light of contemporary biology. *General Systems* **7**: 23–35]

Lorenz K 1943 Die angeborenen Formen möglicher Erfahrung. *Zeitschrift für Tierpsychologie* **5**: 235–409

Lorenz K 1949 *Er redete mit dem Vieh, den Vögeln und den Fischen*. Borotha Schoeler, Vienna [1952 *King Solomon's Ring*. Methuen, London]

Lorenz K 1950a *The Comparative Method in Studying Innate Behavior Patterns*. Symposia of the Society for Experimental Biology 4, pp. 221–68

Lorenz K 1950b *So kam der Mensch auf den Hund*. G. Borotha-Schoeler, Vienna [1954 *Man Meets Dog*. Methuen, London]

Lorenz K 1953 Über angeborene Instinktformeln beim Menschen *Deutsche Medizinische Wochenschrift* **78**: 1566–9, 1600–4

Lorenz K 1963 *Das sogenannte Böse. Zur Naturgeschichte der Aggression*. G. Borotha-Schoeler, Vienna [1966 *On Aggression*. Harcourt, Brace & World, New York]

Lorenz K 1965a *Über tierisches und menschliches Verhalten. Aus dem Werdegang der Verhaltenslehre. Gesammelte Abhandlungen*. Bd. I (1931, 1932, 1935, 1937, 1938, 1942), Bd. II (1941b, 1950c, 1954, 1959, 1961, 1963b). Piper, Munich [transl. Martin R 1970–71 *Studies in Human and Animal Behavior*. In Vol II, the 1961 paper has been omitted, since an expanded translation by the author was published the same year as Lorenz 1965b. Harvard University Press, Cambridge, MA]

Lorenz K 1965b *Evolution and Modification of Behavior*. Chicago University Press, Chicago

Lorenz K 1973 *Die Rückseite des Spiegels*. R. Piper, Munich [1977 *Behind the Mirror*. Harcourt, Brace & World, New York]

Lorenz K 1974 Konrad Lorenz (autobiographical sketch). In: Frängsmyr T (ed.) *Les Prix Nobel en 1973*. The Nobel Foundation, Stockholm, Sweden

Lorenz K 1978 *Vergleichende Verhaltensforschung: Grundlagen der Ethologie*. Springer Verlag, Vienna/New York [1981 *The Foundations of Ethology*. Springer Verlag, New York]

Lorenz K 1983 *Der Abbau des Menschlichen*. R. Piper, Munich [1987 *The Waning of Humaneness*. Little, Brown, Boston]

Lorenz K 1988 *Hier bin ich-wo bist Du ? Ethologie der Graugans*. Piper, Munich/Zürich [1991 *Here Am I—Where Are You? The Behavior of the Greylag Goose*. Harcourt Brace Jovanovich, New York]

Lorenz K 1992 *Die Naturwissenschaft vom Menschen. Das 'Russische Manuskript*.' Aus dem Nachlaß herausgegeben von Agnes von Cranach. Piper, Munich/Zürich [1996 *The Natural Science of the Human Species*. MIT Press, Cambridge, MA/London]

Lorenz K, Tinbergen N 1938 Taxis und Instinkthandlung in der Eirollbewegung der Graugans I. *Zeitschrift für Tierpsychologie*. **2**: 1–29
Popper K R 1972 *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford, UK
Schleidt W M (ed.) 1988 *Der Kreis um Konrad Lorenz—Ideen, Hypothesen, Ansichten*. P Parey, Berlin
Schleidt W M 2000 Politik gegen und mit Konrad Lorenz. In: Kotrschal K, Müller G, Winkler H (eds.) *Konzepte der Verhaltensforschung—Konrad Lorenz und seine Folgen*. Filander Verlag, Fürth
Tinbergen N 1951 *The Study of Instinct*. Oxford University Press, Oxford, UK
Vollmer G 1980 *Evolutionäre Erkenntnistheorie*. Hirzel, Stuttgart, Germany
Wuketits F M 1990 *Konrad Lorenz—Leben und Werk eines großen Naturforschers*. Piper, Munich

W. M. Schleidt

Lotka, Alfred James (1880–1949)

When A. J. Lotka died on December 5, 1949 his obituary in the *Journal of the American Statistical Association* began 'A brilliant chapter in modern demography came to a close with the death of Alfred James Lotka.' David Smith, translator of his major work in the field (Lotka 1939), calls him 'the father of mathematical demography.' And these were understatements of his contribution, for demography was only a part of Lotka's scientific interests—he was also described as chemist, physicist, biologist, and actuary, everywhere seeking better understanding through the use of mathematics.

Lotka was born on March 2, 1880 in Lemberg (Lvov), a Polish town of the Austro-Hungarian Empire, of American parents who were traveling Moravian missionaries; he died in Red Bank, New Jersey. His training was international: a boyhood mostly in France; a degree at the University of Birmingham (B.Sc. 1901, D.Sc. some 10 years later); graduate work in physics at Leipzig. Coming to the USA in 1902, he worked for the General Chemical Company, briefly for the US Patent Office as a patent examiner, and for the National Bureau of Standards as a physicist. He was editor of the *Scientific American Supplement* (1911–14), finally settling in the Statistical Bureau of the Metropolitan Life Insurance Company (1924–47). He was for long a bachelor, somewhat shy, described as quiet, learned, modest, and gently humorous. In 1935 he married Romola Beattie; they had no children. He was elected President of the Population Association of America (1938–9) and of the American Statistical Association (1942).

In his terminal year of graduate school, at Cornell in 1908–9, he began work with Professor F. R. Sharpe that led to the famous renewal equation—a mathematical description of the replacement of the generations:

$$B(t) = \int_{\alpha}^{\beta} B(t-a)p(a)m(a)da \qquad (1)$$

In the simplest case, rates of birth and death are constant over time, there is no migration to or from the defined area and the female sex only is considered. These conditions make possible a strict accounting of the relation between two generations. For on the right side is $B(t-a)$ the number of births a years earlier ($\alpha \leqslant a \leqslant \beta$ where α is the youngest age of childbearing and β the oldest); $p(a)$, the probability that a child born a years ago survives to time t; $m(a)$ is the birth rate for women aged a; integrating gives current births and so should be comparable to the births $B(t)$ on the left side. The unknown quantity that should make the two sides equal is the trajectory of births, to be provided in terms of the known functions $p(a)$ and $m(a)$. If the trajectory of $B(t)$ in Eqn. (1) is simplified down to a curve of geometric increase, as Lotka initially assumed, then the question posed by the renewal equation is what rate of increase of births would equate the two sides of Eqn. (1). If that rate is r, taken either arbitrarily or from past data, then the age distribution of the population would stabilize over time to numbers proportional to $\exp(-ra)p(a)$ at the several ages a.

In finite intervals of time and age, Eqn. (1) can be expressed naturally as a recurrent series and this was done by Lotka himself late in his career. P. H. Leslie formulated it in matrix terms. To these can be added a fourth version in partial differential equations (Keyfitz and Keyfitz 1997), that also shows and compares the three other formulations.

Like most other discoveries, this one had predecessors who in part anticipated, as well as followers who marginally improved on, Lotka's work. Leonhard Euler (1760) came close on the theory; Edwin Cannan (1895) in the nineteeth century published an arithmetic calculation projecting the population of England and Wales that forecast the (then surprising) fall in births. R. Kuczynski (1931) and Richard Böckh of the Berlin Statistical Office were important predecessors working with the ratio of change per generation, the Net Reproduction Rate (NRR), $R_0 = \int p(a)m(a)da$. Willy Feller (1941), Princeton mathematician, noted that the solution to the renewal equation fell out simply and rigorously if each of the functions involved is translated into its unique Laplace Transform. Golubitsky et al. (1975) published an even more compact and more general solution.

Lotka went on to other demographic questions. He was curious to know why old populations have so few surnames and found the answer in a pure probability accounting for the extinction of family lines (1931a). He ascertained the frequency of orphanhood as a

function of birth and death rates (1931b) and calculated the money loss to dependents and to society when a person dies prematurely (Lotka and Dublin 1946). In an early paper he went beyond human populations on the mode of growth of self-renewing aggregates; then on industrial replacement—for instance in a fleet of trucks of varied ages—he showed how the mathematics fits trucks as well as people.

Lotka's *Elements of Physical Biology* (1925), composed between 1922 and 1924 under a fellowship arranged by Raymond Pearl at Johns Hopkins University, is sharply distinguished from his earlier work—it looks at the whole of the living world from a physicochemical perspective. A first chapter on definition rejects a concept of life that sets it up behind an impassable boundary, such that living matter can never emerge through physical and chemical mechanisms. But it must have so emerged at least once in the history of the planet, and that emergence was of interest to Lotka.

Lotka anticipated modern ecology in his continual search for cyclical processes. One of many examples is the circulation of sulfur, that may be summarized from the *Elements of Physical Biology* as follows.

Sulfur occurs in abundance in inorganic nature in the form of the sulphates of the alkalies and the alkaline earths. Plants assimilate these compounds directly and then form proteins containing from 0.3 to 2 percent of sulfur. In this form it is assimilated by animals, who excrete the element chiefly in the form of sulfates. These, returned to the soil, complete the cycle.

And in the same way, Lotka portrays the cycles of water, iron, chlorine, and other components of the earth's crust. He took the science of these cycles from Swedish chemist Arrhenius and others, but he gets credit for being among the very first to see how crucial these cycles are to life and how they fit into and sometimes exacerbate one another. The number of people in the world more than doubled since his time, and production per person far more than doubled, so we are now forced to attend to such cycles as those of fresh water and carbon, and they enjoy considerable attention in legislative bodies and space in late twentieth-century newspapers and magazines. But all that is a development of the late twentieth century, so we have to remind ourselves that these issues were a preoccupation of Lotka from early in the century when there were very few far-sighted enough to think they might have practical consequences.

Lotka refers to the Swedish chemist Arrhenius who first saw from theoretical considerations the possibility of man-made carbon dioxide leading to global warming, and he cites Arrhenius' finding that the air is becoming richer in CO_2. Moreover plants, especially trees, withdraw CO_2, and the cutting of trees releases CO_2; the oceans could offset the tree-cutting by increasing their absorption of the gas. These are only three of the determinants of free CO_2 in the air and no one has yet shown a way to isolate the man-made CO_2 released by the burning of fossil fuels and cutting of forests from changes in the capacity of the oceans to hold carbon. Seventy-five years ago Lotka had a realistic view of the complexities of global equilibrium and the difficulties of forecasting what forces would dominate within it. The earth is a vast and complex physico-chemical-biological engine in constant motion, and if this is now a commonplace, Lotka was in the forefront of those who made it so.

The Lotka–Volterra equations are much referenced. Vito Volterra, an Italian mathematician, was asked by his biologist son-in-law Humberto D'Ancona whether he could explain the cycles of fish population in the Adriatic Sea. Lotka, also in the first half of the 1920s, wanted to see whether the relation of predator and prey could be assimilated to the autocatalytic processes of chemistry he had studied earlier. A pair that comes to mind as a model for both Lotka and Volterra are foxes ($F(t)$) and rabbits ($R(t)$), in which a carnivorous species is dependent on a grass-eating species. The flow of energy is *sun* → *grass* → *rabbits* → *foxes*. Lotka (1925) and Volterra (1927) seem independently to have noted and explained the same features of predation, especially the cycles it generates. The simplest form of the Lotka–Volterra equations is

$$F'(t) = -\alpha F(t) + \gamma R(t) \tag{2}$$

$$R'(t) = \beta R(t) - \delta F(t) \tag{3}$$

Primed symbols are derivatives, and the four positive numbers $\alpha, \beta, \gamma, \delta$ are fixed parameters. The $-\alpha$ in the first equation tells us that in the absence of the prey (i.e., if $R(t)$ is zero) then the predator declines in numbers, to ultimate extinction. The second equation shows the prey increasing indefinitely in the absence of the predator. That γ has a plus sign in the first equation allows for the fact that with more abundant prey the predator grows faster. With other configurations of signs the same equations can represent competition, symbiosis, etc. They can incorporate limits on the amount of grass and can be extended to three or more species.

Lotka's indifference to what he considered mathematical niceties was matched by a quality found in many other original minds, great sensitivity on priorities. He engaged in a long and indecisive controversy with R. R. Kuczynski. The history of science is pockmarked with such disputes. What puts this one in a special class is that Paul Samuelson undertook a uniquely thorough investigation (1977) to sort out the issues, and credits Lotka, Böckh, and Kuczynski, among others. With access to the Princeton archives where Lotka's private papers, as well as his publications, are deposited, he demonstrates with the help of pages of solid algebra that Lotka's predecessors turned up a good part of the renewal story, but their work was

incomplete. Lotka had both the theoretical interest and the command of data, plus the drive to follow through.

So, Samuelson concludes on this issue, the brighter light has drowned out the earlier light. But now that the dust has died down, it looks as though Lotka's place in history, which is a secure one, is based less on the priorities he claimed so vigorously in controversy than on his anticipation and objective treatment of a complex of problems revolving around population and environment, problems barely suspected when he lived, that we now realize will be with us for a long time. A half-century after his death he is very much alive through his books and papers, and in the demographic profession we can still say, with Frank Notestein (see *Notestein, Frank Wallace (1902–83)*), that 'his is the greatest name.'

See also: Demography, History of; Demography: Twentieth-century History; Statistical Methods, History of: Post-1900; Statistics, History of

Bibliography

Cannan E 1895 The probability of a cessation of the growth of population in England and Wales during the next century. *Economic Journal* **5**: 505

Euler L 1760 A general investigation into the multiplication and mortality of the human species [1970 *Theoretical Population Biology* **1**: 307 (trans. N Keyfitz and B Keyfitz)]

Feller W 1941 On the integral equation of renewal theory. *Annals of Mathematical Statistics* **12**: 243

Golubitsky M, Keeler E, Rothschild M 1975 Convergence of the age structure: Applications of the projective metric. *Theoretical Population Biology* **7**: 84–93

Keyfitz B L, Keyfitz N 1997 The McKendrick partial differential equation and its uses in epidemiology and population study. *Mathematical and Computer Modelling* **26**: 1–9

Kuczynski R 1931 *Fertility and Reproduction: Methods of Measuring the Balance of Births and Deaths*. Macmillan, New York

Lotka A J 1911 A problem in age distribution. *Philosophical Magazine* **21**: 435

Lotka A J 1925 *Elements of Physical Biology*. Williams & Wilkins, Baltimore, MD. Republished, 1956, as *Elements of Mathematical Biology*. Dover Publications, New York

Lotka A J 1931a The extinction of families. *Journal of the Washington Academy of Sciences* **21**: 377–453

Lotka A J 1931b Orphanhood in relation to demographic factors. *Metron* **19**: 37–109

Lotka A J 1939 *Théorie analytique des associations biologiques*. Part I: Principes; Part II: Analyse démographique avec application particulière à l'espèce humaine. *Actualités Scientifiques et Industrielles*, 187, 780. Hermann & Cie, Paris [1998 *Analytical Theory of Biological Populations* (trans. D P Smith and H Bossert). Plenum, New York and London]

Lotka A J, Dublin I 1946 *The Money Value of a Man*, rev. edn. The Ronald Press, New York

Lotka A J, Sharp F 1911 A problem in age distribution. *Philosophical Magazine* Ser. 6, **21**: 435–8

Samuelson P 1977 Resolving a historical confusion in population analysis. In: Smith D, Keyfitz N (eds.) *Mathematical Demography: Selected Papers*. Springer-Verlag, New York

Volterra V 1927 *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*. C Ferrari, Venice, Italy

N. Keyfitz

Love and Intimacy, Psychology of

When people talk about the things that matter most in their lives, they usually place close relationships at or near the top of the list. And when researchers examine the processes that make close relationships satisfying, love and intimacy are usually among the most influential factors. The apparent importance of love and intimacy to human well-being has fueled longstanding interest by lay persons and scholars alike in these fundamental properties of close relationships, yet empirically minded researchers are relatively new to the pursuit, perhaps because of the challenges inherent in studying such subjective concepts (see *Social Relationships in Adulthood*). The sizable literature that has materialized since the 1970s suggests, however, that empirical research into the nature and consequences of love and intimacy can be profitable. Moreover, recent research in this area has been invigorated by methodological advances (e.g., in studying interdependent dyads and in psychophysiological assessment), as well as by increased appreciation of the importance of love and intimacy during human evolution, as noted below.

Although research on love and intimacy evolved as largely independent traditions, a basic connection is now clear. Early on, Berscheid and Walster (1969) distinguished between two types of love. Passionate love refers to a state of intense arousal, characterized by strong emotions, deep longing for the other, and mental preoccupation with that person. In contrast, companionate love, 'the affection we feel for those with whom our lives are deeply intertwined' (1969, p. 9), involves trust, caring, respect, and mutual responsiveness, processes that also characterize intimacy. Both forms of love may arise in any close relationship, although passionate love generally is studied only in regard to romantic relationships. Passionate and companionate love are considered widely to be independent processes, each with its own antecedents, consequences, time course, and underlying mechanisms, although they may co-occur in a given relationship. To be sure, the many variations with which people experience love—maternal love, love for one's fellow humans, courtly love, and so on—suggest that love is a complex and multifaceted phenomenon not easily reduced to common elements. Nevertheless, formal approaches to differentiating its many varieties generally have identified these two factors, labeled as

passion and intimacy, as a first-order distinction (as well as a third factor, commitment; e.g., Sternberg 1986). One such example is Aron and Westbay's (1996) analysis of cognitions, feelings, and behaviors associated with the various experiences of love.

1. *Passionate Love*

Substantial research has investigated the nature and dynamics of passionate love, especially its behavioral and cognitive manifestations. For example, passionate love tends to be associated with high levels of physiological arousal (attributed to the neurotransmitters dopamine, norepinephrine, and serotonin); longing for connection, both physical and emotional, with the loved one; strong, often idealized, positive feelings about the other; mental preoccupation with the loved person, often intrusively so; and emotional sensitivity to the other, such that emotional states may swing from euphoria to despair depending on perceptions of the loved person's response (see Fisher 1998 for a more extensive review). As these and the many other phenomena characteristic of passionate love indicate, but as has not always been recognized, passionate love is properly regarded as one of the emotions (see *Emotions, Psychological Structure of*).

Scholars also have attempted to understand passionate love by considering its functional consequences. Self-theorists see passion arising when relationships, be they imagined, incipient, or ongoing, are perceived to present new or expanded opportunities for meeting the self's most important needs (i.e., for fulfilling personal ideals; for self-enhancement; for mating). In this view, passionate feelings reflect the process of personal growth experienced in the individual's real and/or anticipated involvement with the other. To evolutionarily oriented theorists, in contrast, passionate love serves to redirect the individual's attention and behavior away from routine ongoing activities and instead toward the existence of a potential mating partner, thereby facilitating the initiation and development of mating relationships. This latter approach can help explain one of the more intriguing findings in the literature: that levels of passionate love tend to be a poor predictor of long-term marital satisfaction and longevity. It is thought that this is because passionate love tends to diminish in the early years of marriage, once the function of attracting reproductive partners to each other has been fulfilled (see *Sexual Attitudes and Behavior*).

Intimacy, on the other hand, tends to be better associated with the long-term success of marital and other types of close relationships. It is believed widely that intimacy is essential to human well-being and that its absence can cause significant emotional distress; for example, many studies have demonstrated that the existence of intimate ties and social interactions is associated with higher levels of social and psychological development, and with better mental and physical health. Indeed, to some theorists, 'the capacity to make intimate emotional bonds with other individuals … is regarded as a principal feature of effective personality functioning and mental health' (Bowlby 1983, p. 121). Accounts of laypersons interpersonal experiences, expectations, and goals likewise show a similar appreciation for the qualities that intimate relationships provide, beginning in early adolescence and continuing throughout the lifespan (see *Friendship: Development in Childhood and Adolescence*).

2. *Intimacy*

Diverse phenomena have been investigated under the general heading of intimacy: for example, the concept of intimacy has been used to describe the process of revealing private thoughts and feelings to others; to identify a relatively intense form of nonverbal engagement (e.g., through eye contact, touch, or physical proximity); to refer to the existence of an affectionate, supportive relationship; or to characterize the stage of life in which the primary developmental task is to establish an emotionally close, trusting, and sexual relationship with another person. The variability inherent in these usages, when considered along with yet further everyday-language meanings of the word 'intimate,' that is, as a synonym for closeness, sexuality, privacy, or marriage, has sometimes fostered conceptual ambiguity, making it difficult to compare and integrate different theoretical perspectives and different research traditions. However, this diversity has also provided an extensively documented, multifaceted depiction of a complex phenomenon.

Research on self-disclosure has played a particularly influential role in the development of knowledge about intimacy. Spurred by the pioneering studies of Altman and Taylor (1973), who defined intimacy in terms of the depth and breadth of one person's revelation of nominally private facts and feelings to another person, this work has supported two general conclusions: that self-disclosure moves from superficial to intimate topics as partners become acquainted; and that partners tend to reciprocate levels of self-disclosure. (In other words, too little or too much disclosure, in comparison to personal standards, situational norms, or a partner's behavior, tends to hamper relationship development.) More recent and more sophisticated models have sought to broaden this conceptualization, predicated on the recognition that spoken messages represent only one method of communicating intimacy. Nonverbal communication also contributes to intimacy, in either of two ways: by indicating current affective states, including partners' responses to each other (e.g., facial expressions of sympathy or scorn), and by regulating existing levels

of interpersonal engagement (e.g., behaviors that enhance or diminish immediacy; see *Nonverbal Communication, Social Psychology of*). Intimacy may also be established through behavior (e.g., coming to know another through observation and shared experience and supportive acts that express care and concern for the other).

Other theorists have emphasized the interpersonal process components of intimacy. One such theory stresses the partner's perceived response in the context of self-disclosure (Reis and Shaver 1988, Reis and Patrick 1996). In this approach, intimacy refers to the experience of a partner's response to core (often emotion-laden) aspects of the self as understanding, validating, and caring. This and similar models are helpful in focusing attention on the fundamentally interactive nature of intimacy. Many relationship theorists, in fact, share the perspective of conceptualizing intimacy as a complex process that cuts across two levels of analysis: the psychology of the individual self, and the psychology of interaction and relationships. Intimacy in this sense possesses in common certain features with other intrinsically interdependent processes such as social support, commitment, trust, conflict, cooperation, and shared experience.

2.1 *Individual Differences in Love and Intimacy*

Although it is apparent that passion and intimacy are at their core relationship-specific—for the most part, people feel love for a particular person, not for people in general—considerable research has examined individual differences in love and intimacy, if for no other reason than the fact that people tend to exhibit modest levels of consistency in their responses to various others. For example, people vary in the extent to which they are motivated to pursue intimate relations with others; these tendencies influence their behavior in appropriate social contexts (although situational factors may moderate the extent to which behavior reflects dispositions in particular circumstances). Individual differences reflect both personality (including genetically determined and learned qualities) and past experiences in close relationships. The study of relationship cognition is concerned with how these experiences are represented in the various mental structures that embody knowledge about significant others and relationships (e.g., expectations, beliefs, self- and other-schemas), and how these beliefs influence subsequent relationship-relevant behavior. These effects can be profound. For example, research on attachment theory indicates that early experiences with caregivers may have substantial impact on an individual's feelings of security in adult relationships, as well as a wide variety of behaviors in such relationships (see *Adult Psychological Development: Attachment*).

3. *Future Directions*

Research on love and intimacy seems likely to profit from several recent trends in psychological research. One such trend concerns the recognition that many important constructs are better investigated from an interdependent than from an individual perspective, that is, as noted above, although individual-level variables may contribute meaningfully to love and intimacy, properties such as love and intimacy are most fully illuminated when they are examined within the relationship context in which they emerge and operate (see Reis et al. 2000).

Another important trend concerns the integration of biological and psychological approaches. Thus, recent advances in brain imaging may help researchers better understand the essential nature of love and intimacy, and their broad impact, both behaviorally and physiologically, on the human organism. Also, the rapid growth in evolutionary theorizing holds considerable promise for addressing many as yet unanswered questions about love and intimacy. Although the intrinsic importance of these subjective feelings for human well-being is incontrovertible, recognition of the role played by relationships, and especially close relationships, during human evolution suggests even broader implications. That is, if small groups and nuclear families provided the primary context for human evolution, as is widely believed, then love and intimacy, the processes that attract individuals to mating partners and promote the longterm viability of the relationships that ensue, must be acknowledged as significant contributors to natural selection and sexual selection.

In short, it seems likely that the relevance of these two cardinal processes to a broad understanding of human well-being and functioning will, if anything, expand in future investigation.

See also: Cultural Variations in Interpersonal Relationships; Emotions, Psychological Structure of; Family and Gender; Friendship, Anthropology of; Friendship: Development in Childhood and Adolescence; Interpersonal Attraction, Psychology of; Nonverbal Communication, Social Psychology of; Personality and Marriage; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexual Preference: Genetic Aspects; Social Relationships in Adulthood

Bibliography

Altman I, Taylor D A 1973 *Social Penetration: The Development of Interpersonal Relationships*. Holt, Rinehart, & Winston, New York

Aron A, Westbay L 1996 Dimensions of the prototype of love. *Journal of Personality and Social Psychology* **70**: 535–51
Baumeister R 1999 Passion, intimacy, and time: Passionate love as a function of change in intimacy. *Personality and Social Psychology Review* **3**: 49–68
Berscheid E, Reis H T 1998 Attraction and close relationships. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, Vol. 2, pp. 193–281
Berscheid E, Walster E H 1969 *Interpersonal Attraction*. Addison Wesley, Reading, MA
Bowlby J 1983 *A Secure Base*. Basic Books, New York
Diamond J 1997 *Why is Sex Fun?* Basic Books, New York
Fisher H E 1998 Lust, attraction, and attachment in mammalian reproduction. *Human Nature* **9**: 23–52
McAdams D P 1989 *Intimacy: The Need to be Close*. Doubleday, New York
Reis H T, Collins W A, Berscheid E 2000 The relationship context of human behavior and development. *Psychological Bulletin* **126**: 844–72
Reis H T, Patrick B C 1996 Attachment and intimacy: Component processes. In: Kruglanski A, Higgins E T (eds.) *Social Psychology: Handbook of Basic Principles*. Guilford Press, New York, pp. 523–63
Reis H T, Shaver P R 1988 Intimacy as an interpersonal process. In: Duck S (ed.) *Handbook of Personal Relationships*. Wiley, Chichester, UK, pp. 367–89
Sternberg R J 1986 A triangular theory of love. *Psychological Review* **93**: 119–35

H. T. Reis

Luce's Choice Axiom

1. Context and Notation

Luce's Choice Axiom (hereafter LCA) is a quantitative hypothesis about choice behavior first proposed and analysed by Luce (1959). It envisions a setting in which an individual makes repeated choices from a set A containing N alternatives: $A = (a_1, a_2, \ldots, a_N)$ (e.g., N restaurants). Sometimes all N alternatives are available for selection (all the restaurants are open); on other occasions only subsets of A are available (some restaurants are closed). On each occasion exactly one alternative is chosen. Choice is assumed to be probabilistic: faced with the same set of alternatives on different occasions, the individual may make different choices. $P(i; S)$ denotes the probability that a_i is chosen when the set of available alternatives is S; e.g., $P(1; A)$ is the probability that a_1 is chosen when all N alternatives are available, and if $S = \{a_1, a_2, a_3\}$, $P(1; S)$ is the probability that a_1 is chosen when only a_1, a_2, and a_3 are available. The special case of 2-alternative choice (known as 'paired comparison') arises often and has its own notation: $P(i, j)$ is the probability of choosing a_i when the available set is $\{a_i, a_j\}$.

In this context it is natural to wonder about relationships between choice probabilities for different sets of available alternatives. For example, suppose the full-set probabilities $P(i; A)$ are known for all $i = 1, \ldots, N$. Can those probabilities be used to predict the paired comparison probabilities $P(i, j)$? Alternatively, can the paired comparison probabilities be used to predict the full-set probabilities? LCA allows one to do both things and more—provided, of course, it is empirically valid for the situation under investigation.

2. The Axiom

LCA has two parts. Part 1 is the heart of the matter: it relates the choice probabilities for subsets of A to those for A itself. Part 2 is a housekeeping assumption whose effect is to make all the relevant choice probabilities nonzero. ('$P(S; A)$' below is the probability that when the whole set A is available, the alternative chosen belongs to subset S, i.e., $P(S; A) = \sum_{a_i \in S} P(i; A)$. '$S \backslash \{a_i\}$' is the set S with a_i deleted from it.)

(a) Part 1. If $P(i, j) \neq 0$ for all $a_i, a_j \in A$, $i \neq j$, then for all $S \subseteq A$ and every $a_j \in S$

$$P(i; A) = P(i; S)P(S; A) \tag{1}$$

(b) Part 2. If $P(i, j) = 0$ for some pair a_i, a_j in A, then for all $S \subset A$

$$P(S; A) = P(S \setminus \{a_i\}; A \setminus \{a_i\}) \tag{2}$$

Note that $S = \{a_i\}$ in Eqn. (2) implies $P(i; A) = 0$. The intuition here is that if restaurant i is never chosen in a paired comparison with restaurant j, it will remain unchosen when more restaurants are added to the available set. In Part 1, the assumption $P(i, j) > 0$, combined with Eqn. 1, implies $P(i; A) > 0$ for all i. Consequently $P(S; A)$ is nonzero, so Eqn. (1) can be rewritten as

$$P(i; S) = \frac{P(i; A)}{P(S; A)} \tag{3}$$

The right side of Eqn. (3) is the conditional probability that a_i is chosen from A, given that the choice from A is some member of S. Thus, Part 1 asserts that when the choice alternatives are restricted to any subset S, the probability of choosing a_i from S equals the conditional probability that a_i is chosen when the full set A is available, given that the choice from A belongs to S. Note that this is an empirical hypothesis, not a theorem of probability theory.

3. Implications

First and foremost, LCA implies that there is a function v (a 'scale') mapping the N alternatives in A on to N non-negative real numbers $v(a_1), \ldots, v(a_N)$, with the following useful property: for all $a_i \in S \subseteq A$

$$P(i; S) = \frac{v(a_i)}{\sum_{a_j \in S} v(a_j)} \tag{4}$$

Clearly, from Eqn. (3), $v(a_i) = P(i; A)$ guarantees one such scale, but infinitely many more can be created by multiplying the $P(i; A)$ values by any positive constant (i.e., Eqn. (4) remains true if one substitutes $v'(a_i) = cv(a_i)$ for any $c > 0$). Thus v is a ratio scale: its values for a given set of alternatives are unique only up to multiplication by a positive constant.

From Eqn. (4) it follows that for all a_i, a_j

$$\frac{P(i; S)}{P(j; S)} = \frac{v(a_i)}{v(a_j)} = \frac{P(i; A)}{P(j; A)} \tag{5}$$

(This is called the 'constant ratio rule'.) In particular

$$\frac{P(i, j)}{P(j, i)} = \frac{P(i; A)}{P(j; A)} \tag{6}$$

Thus, when LCA holds, the paired comparison probabilities $P(i, j)$ can be predicted from the full set probabilities $P(i;A)$. Moreover $P(i;A)$ from Eqns. (4) and (6)

$$P(i; A) = \frac{1}{\sum_{a_j \in A} v(a_j)/v(a_i)} = \frac{1}{\sum_{a_j \in A} P(j, i)/P(i, j)} \tag{7}$$

so the full-set choice probabilities can be predicted from the paired comparison probabilities.

Finally, for any triple of indices i, j, k

$$\frac{P(i, k)}{P(k, i)} = \frac{v(a_j)}{v(a_k)} = \frac{v(a_i)}{v(a_j)}\frac{v(a_j)}{v(a_k)} = \frac{P(i, j)}{P(j, i)}\frac{P(j, k)}{P(k, j)} \tag{8}$$

Thus, if any two paired comparison probabilities $P(i, j)$, $P(j, k)$ are known, $P(i, k)$ can be predicted. (This is known as the 'product rule'.) Eqns (6), (7), and (8) all provide ways to test LCA experimentally.

4. Validity

LCA has been tested in both real experiments and thought experiments (Luce 1977). A famous early thought experiment (Debreu 1960) shows that LCA is certainly not universally valid. Consider a choice set consisting of a red mug of coffee (RC), a blue mug of the same coffee (BC), and a green mug of tea (GT). One can readily imagine a person who is equally fond of coffee and tea, and indifferent as to color. In this case we expect that in pairwise choices $P(\text{RC, BC}) = P(\text{RC, GT}) = P(\text{BC, GT}) = 1/2$. Then LCA implies $v(\text{RC}) = v(\text{BC}) = v(\text{GT})$, so in a choice with all three alternatives available, each alternative should be chosen with probability 1/3. But intuitively it is much more plausible that the probability of choosing GT will be 1/2, or very near, on the grounds that the fundamental decision is whether to have coffee or tea, and mug color is irrelevant.

Debreu's thought experiment makes it easy to design real experiments in which LCA will fail, and investigators have done so. However, it is pointless to carry out many such exercises. LCA is arguably the simplest possible relationship between choice probabilities for subsets of a master set of alternatives, and when it is empirically valid, it is extremely useful. The real issue is not its universal validity, but whether there are nontrivial contexts where LCA is valid for practical purposes. And indeed there are; Luce's (1977) review describes several. (In testing whether LCA holds empirically it is important to be aware that because of potential individual differences in v-scale values, every member of a group can satisfy LCA individually and the axiom may still appear to fail when tested against group-average data; Luce 1959.)

5. Relationship to Thurstone's Theory and the Double Exponential Probability Distribution

L. L. Thurstone's (1994) 'Law of Comparative Judgment' is an important early theory of choice behavior covering the same ground as LCA. Here, the alternatives are thought of as having numerical values ('utilities') that vary randomly from one choice occasion to another. If the utilities of alternatives $a_1, \ldots, a_N$ are the random variables $U_1, \ldots, U_N$, the probability of choosing a_i from any set S is the probability that U_i is currently the largest of the random variables $\{U_j: a_j \in S\}$. (Theories of this sort are known as random utility models; see *Decision and Choice*: *Random Utility Models of Choice and Response Time*.) Thurstone assumed that utilities are normally distributed, and applications have generally assumed that these normal U_is are independent and identically distributed except for their means. Then U_i can be written as $u_i + X_i$, where u_i is a real constant (the expected utility of a_i), and X_i is a normal random variable with mean zero and variance one; X_i and X_j are independent for all $i \neq j$. The empirical validity of this theory can be compared to that of LCA using its predictions for paired comparisons vs. those implied by LCA's product rule (Eqn. (8)). When that is done the predictions of the two theories prove to be numerically nearly

identical. This near-identity is explained by E. Holman and A. J. Marley's discovery (cited in Luce and Suppes 1965) that LCA is equivalent to (makes exactly the same predictions as) a Thurstone model in which the independent normal random variables X_i are replaced by independent random variables that have a double exponential probability distribution:

$$P(X_i \leqslant x) = \exp(-\exp(-x)) \quad \text{for } -\infty < x < \infty \tag{9}$$

For paired comparisons, the choice probabilities of any independent Thurstone model depend entirely on the probability distribution of the difference random variables $X_i - X_j$ $(i \neq j)$. For Thurstone's normal model those differences are normally distributed, while for LCA (i.e., the double exponential Thurstone model) they have a logistic distribution:

$$P(X_i - X_j \leqslant x) = 1/(1 + exp\,(-x)).$$

That distribution is numerically very nearly normal, so the pair comparison predictions of LCA and Thurstone's model are understandably nearly identical.

A general family of independent Thurstone models can be formed by allowing the X_i random variables to have various distributions besides the normal and double exponential. It can be shown that within this family, only the double exponential model is equivalent to LCA for choice sets with $N \geqslant 3$ (Yellott 1977, McFadden 1974). That distribution is distinguished in probability theory as one of three forms that can be limit distributions of the maxima (suitably normalized) of sequences of independent random variables (Gumbel 1958). That property in turn can be used to show that LCA is unique in an interesting way. Consider a thought experiment in which there are three basic alternatives—say, the beverages coffee, tea, and milk. On each choice occasion the set of available alternatives consists of n identical cups of coffee, n cups of tea, and n glasses of milk. Intuition suggests that the probability of choosing coffee (i.e., any one of the n identical cups) should be independent of n for all n, no matter how large. Within the family of independent Thurstone models, only the double exponential model has that property (Yellott 1977).

6. *Relationship to Tversky's Theory of 'Choice by Elimination'*

This theory (Tversky 1972) was designed to handle Debreu's counterexample and others like it, doing so in the spirit of LCA (to which it reduces in special cases). Here, the choice alternatives are regarded as sets of attributes or 'aspects.' For example, in Debreu's counterexample, RC has the aspects 'coffee' and 'red,' GT has aspects 'tea' and 'green,' etc. Each aspect has a weight w—a nonnegative real number, e.g., $w(\text{red}) = 1$, $w(\text{tea}) = 0.5$, etc. Choice involves a series of stages. At each stage the individual selects one aspect from the set of all aspects belonging to the available alternatives (ignoring aspects shared by every alternative), and eliminates from further consideration every alternative that does not possess that aspect. The probability of selecting a given aspect is equal to its weight divided by the sum of the weights of all the aspects present in the alternatives still available, omitting aspects that are shared by every alternative. (Thus aspect selection obeys LCA, with the weights as v-scale values. If the alternatives have no aspects in common, choice is a one-stage process that obeys LCA exactly.) The process of aspect selection and elimination continues until only one alternative remains.

To apply this model to Debreu's counterexample, let $w(r)$, $w(b)$, $w(g)$ be the weights of the colors, and $w(c)$, $w(t)$ the weights of coffee and tea. Assuming $w(r) = w(b) = w(g)$, and $w(c) = w(t)$, one easily derives $P(\text{RC, BC}) = P(\text{RC, GT}) = P(\text{BC, GT}) = 1/2$, and

$$P(\text{GT};\{\text{RC, BC, GT}\}) = \frac{1}{2 + \dfrac{w(g)/w(t)}{1 + w(g)/w(t)}} \tag{10}$$

Then Debreu's prediction that $P(\text{GT}; \{\text{RC, BC, GT}\})$ should be near 1/2 can be obtained by making the color weights much smaller than the beverage weights, i.e., $w(g)/w(t) \approx 0$.

7. *LCA in Economics*

Applications of LCA in econometrics (where it is known as the Multinomial Logit model) were pioneered by McFadden (1974), who developed a widely applied extension of LCA called the Generalized Extreme Value model (McFadden 1981).

8. *LCA and Ranking*

An individual's preferences among a set of alternatives can be expressed by rank-ordering them instead of choosing a single one (see also *Ranking Models, Mathematics of*). Rankings can be made in a forward direction ('best,' 'next best,' etc.) or a backward direction ('worst,''next worst' ...). Luce (1959) suggested extending LCA to ranking via a model in which ranking is accomplished by a sequence of covert choices. In forward-ranking the most preferred alternative in the available set is chosen and ranked first,

then a second choice is made from the remaining alternatives (with choice probabilities that obey LCA) and ranked second, and so on until all the alternatives have been rank-ordered. In backward-ranking a least-preferred alternative is selected first and ranked worst, then a second-worst choice is made from the remaining alternatives, etc., again with the choice probabilities obeying LCA. Intuitively this model seems quite plausible, but surprisingly it proves to be mathematically incompatible with the equally plausible idea that when a person ranks the same alternatives from best to worst, and also from worst to best, rankings expressing the same preference ordering should occur with the same probability (Luce 1959). From a Thurstone model perspective, this unexpected incompatibility stems from the asymmetry of the double exponential probability density function. Other (though not all) independent Thurstone models based on asymmetric densities share the same property (Yellott 1997).

See also: Bayesian Theory: History of Applications; Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision and Choice: Paradoxes of Choice; Decision and Choice: Random Utility Models of Choice and Response Time; Heuristics for Decision and Choice; Sequential Decision Making

Bibliography

Debreu G 1960 Review of R. D. Luce, Individual choice behavior: A theoretical analysis. *American Economic Review* **50**: 186–8

Gumbel E G 1958 *Statistics of Extremes*. Columbia University Press, New York

Luce R D 1959 *Individual Choice Behavior*. Wiley, New York

Luce R D 1977 The choice axiom after twenty years. *Journal of Mathematical Psychology* **15**(3): 215–33

Luce R D, Suppes P 1965 Preference, utility, and subjective probability. In: Luce R D, Bush R, Galenter E (eds.) *Handbook of Mathematical Psychology*. Wiley, New York, Vol. 3

McFadden D 1974 Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed.) *Frontiers in Econometrics*. Academic Press, New York

McFadden D 1981 Econometric models of probabilistic choice. In: Manski C F, McFadden D (eds.) *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press, Cambridge, MA

Thurstone L L 1994 [1927] A law of comparative judgment. *Psychological Review* **101**(2): 266–70

Tversky A 1972 Choice by elimination. *Journal of Mathematical Psychology* **9**: 341–67

Yellott J I Jr 1977 The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology* **15**: 109–44

Yellott J I Jr 1997 Preference models and irreversibility. In: Marley A A J (ed.) *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*. Lawrence Erlbaum Associates, Mahwah, NJ

J. I. Yellott, Jr.

Luhmann, Niklas (1927–98)

Niklas Luhmann was born on December 8, 1927, in Lüneburg as the son of a brewer. His mother came from a family related to the Swiss hotel industry. At the end of World War II Luhmann was conscripted to the air defense for a short time. After the war he studied law at Freiburg (1946–9), became an assistant to a local lawyer, and from 1952 to 1962 worked as a public servant in the regional administration of Niedersachsen. In these years Luhmann was an official who after his office hours read lyrics and theory, especially Edmund Husserl and Talcott Parsons. There are two early publications of his in the journal *Verwaltungsarchiv* from 1958 and 1960, but a break in his career came with Luhmann gaining a leave in 1960/61 to study at Harvard with Talcott Parsons. After Harvard Luhmann changed, now as a researcher, to the *Verwaltungshochschule Speyer*, an institution for the continuing education of administrators. Luhmann's early writings were noted by the then very influential German sociologist Helmuth Schelsky at Münster who persuaded Luhmann, who had no academic degrees, to acquire a doctoral degree and a *Habilitation* at Münster in 1966. He became head of a new division Sociological Theory and Sociology of Law at the *Sozialforschungsstelle Dortmund*, affiliated with Münster, from 1966 to 1968. In 1968 he became the first professor of the newly established University of Bielefeld. He stayed there until his retirement in 1993 and continued this relationship until his death on November 6, 1998, in Oerlinghausen.

1. Theory of Social Systems

From its beginnings in the early 1960s Luhmann called his contribution to sociological theory systems theory (Luhmann 1970). Another early self-designation was functional-structural theory, an obvious opposition to Talcott Parsons's structural-functionalism by which Luhmann intended to say that in his writings the concept of function takes a more strategic role than in Parsons. Social reality is thought to consist in problem solutions; the identification of the problem a specific

structure solves is what functionalism is about; and having identified the function of a specific structure one may look for functional equivalents, that is, alternative institutional solutions for the same problem. Methodologically this points to the essentially comparative character of Luhmann's theory; he was always looking at reality as problem solving and compared the solutions identified to institutional alternatives.

What is systems theory? First of all this self-designation establishes system/environment as the guiding distinction for processes of theory building. One possibility is to conceive systems theory as a theory of open systems which implies a focus on exchange processes connecting a system and its environments. In Luhmann's case we have to do with a cybernetic systems theory in which a system is described by the selectivity of its relations to its environments. For this Luhmann invented the formula which made him famous: reduction of complexity. Environments are seen as overly complex and systems constitute themselves by a selective reconstruction of this complexity.

Luhmann primarily was interested in social systems and therefore had to establish their specificity. He did this by a theory of meaning (Habermas and Luhmann 1971, pp. 25–100). Whereas all types of systems produce a surplus of possibilities and then generate selection mechanisms reducing this surplus, social systems are constituted by a specific mode of dealing with reductions. In social systems the possibilities not chosen are not eliminated. They are instead maintained as possibilities to which one can recur at a later point in time.

There exists one other system type which operates on the basis of meaning: psychic systems. As it already was the case in Parsons, social and psychic systems are conceived as two different types, separate from one another but coupled via media such as meaning and language. Luhmann's thesis that human beings (as psychic systems) belong to the environment of social systems was always one of the most controversial aspects of his work. But how does a strict separation of social and psychic systems come about if both operate in the medium of meaning? Luhmann proposes to distinguish the elementary constituents from which both types of systems are built. In the case of psychic systems Luhmann favors a theory related to Husserl (Luhmann 1995a, Chaps. 1–4). Thoughts are elementary constituents in psychic systems which are described as systems of consciousness. Occasionally he experiments with more inclusive concepts of consciousness—including feelings, perceptions, acts of will—and points to intentional acts as a more general element of consciousness.

Which is the constituent element of social systems? In a first approximation one would think of social actions as elements (as unit acts in a Parsonian sense). One finds this understanding in the early Luhmann. He even then complicated the understanding of action by introducing a distinction between action and experience for which there were no antecedents in the sociological tradition. Social systems are conceived as processing selections which are either attributed to one of the systems involved which means they are social actions or to objective circumstances of the world which implies that one only experiences these selective events and can influence them in the present situation. This distinction of action and experience depends on attributions, that is, it is a processual result of social systems themselves.

This already points to the fact that social actions can't function as the constituent elements of social systems. Luhmann opted for communications instead, thereby becoming the first major social theorist to base the understanding of society on communications theory (Luhmann 1984, Chap. 4). This implies a switch in sociological theorizing from focussing on the exchange of resources to accentuating information transfer in social processes. Luhmann's communications theory, closely related to Bühler (Bühler 1934), does not conceive of communication as a single selective event. Instead any communicative event is based on interrelating three selections. (a) An informative event which—in the Batesonian understanding (Bateson 1973)—is a 'difference which makes a difference' to an observing system, but which is not a complete communication in itself. (b) An utterance which means that there exists a system conveying the information in question. (c) An understanding which implies that a second system is involved which projects the difference of information and utterance on the first system and understands thereby. A communication is a unity consisting from these three components. The acceptance or rejection of a communication can be identified as the fourth component of a communicative event, but it already belongs to the next communication, thereby securing the sequential unity of a communication system.

But how are communications elements of social systems? Do they possess the internal stability one should demand of an element? Luhmann theorizes this in terms of the temporality of the elements of social systems. In an essay from 1979 (Luhmann 1981a, pp. 101–25) he proposed that the elements of social systems—be it actions or communications—are events. They are of vanishing duration. Just having begun they are already gone, and new events must arise for the system not coming to an end. The reproduction of events is the most important structural imperative of social systems, and reproduction does not mean invariant reproduction which is inexistent in social systems. Reproduction means that new events connect to previous events and are limited by this. Such reproductive events Luhmann calls operations, and this substitution of the concept of operation for the concept of element is accompanied by further fundamental changes in systems theory.

First among them is a slide toward concepts of self-reference. This is already to be seen in Luhmann's early self-presentation as a proponent of cybernetic systems theory which means that selectivity in the relations of the system to its environment is always related to the self-identity of the system. Self-reference is articulated more strongly in the theory of autopoiesis which Luhmann adapted in the eighties from the Chilean neurobiologists Maturana and Varela (Maturana and Varela 1980, Luhmann 1984). Autopoiesis means that anything which functions as unity in a system—element, operation, structure, boundary—results from the production processes of the system itself. There are no imports to the system, no inputs and outputs. The system is closed operationally and structurally, but it is informationally open which means that from observed differences it projects to its environment the system derives internal differences which influence system processes. Systems which are closed structurally toward one another can be structurally coupled, that is, there exist two sets of structures in two systems which evolve under the pressure of irritations resulting from operations of the other system.

A last conceptual change regards the concept of operation. The theory of meaning and information theory point to the understanding that operations always are controlled by differences. Operations choose a specific option and they exclude another alternative option, that is, one can infer an underlying distinction. In social and psychic systems the other side of the distinction, the possibility not chosen, is often included representationally. If this is the case Luhmann calls the operation an observation (cf. Luhmann 1990b, Chap. 2). Agreeing with numerous social theorists from Garfinkel to Giddens who call this reflexivity, Luhmann postulates that observations and self-observations are basic to modern social systems. Second order observation (von Foerster 1984) is the type of observation characteristic of function systems in modern society. In function systems most observations are reflexively oriented to observations by other observers and only indirectly refer to reality.

Luhmann complements this by a calculus of observations invented by the British logician Spencer Brown (Spencer Brown 1972). Regarding the logical structure of observations Spencer Brown postulates a unity of distinction and indication. In distinguishing the two sides of an observation one always has to indicate one of these two sides and one has to do this in the same moment one uses the distinction. This demonstrates that one can not at the same time observe the observation one just makes use of. One can only do it later by applying another distinction to it. Therefore, insight (by using a distinction) and blindness (by being unable to observe the distinction one uses) always go hand in hand in social systems. Reentries are one strategy of dealing with this. A distinction can enter into the domain which is distinguished by it. A system can observe itself via the distinction of system and environment by which it was demarcated in the first place. Luhmann tried to conceive a theory of reentries as the modern version of a theory of rationality (Luhmann 1984, Chap. XI, 1997, Chap. 1, XI).

2. *Theory of Society*

For Luhmann modern social theory always consisted from two main undertakings. First, the theory of social systems; second, among social systems he identified one which encompasses all other social systems: society. His two main theoretical treatises were, therefore, devoted to the theory of social systems and to the theory of society (Luhmann 1984, 1997).

Luhmann postulates three levels of the formation of social systems: Interaction, Organization, and Society. Interaction is an analogue to what Goffman called encounter or interaction order (Luhmann 1975b, pp. 9–38, Goffman 1965). Interaction systems presuppose the copresence of the participants and their reciprocal perception. If these conditions are fulfilled, a social system will arise because then it is impossible not to communicate as Watzlawick demonstrated (Watzlawick 1967). Interaction systems are limited in their capacity: they can process only one topic at a time. When the participants leave the interaction system comes to an end.

Organizations constitute an intermediary level of system formation being based on membership (Luhmann 1964). Organizational members are bound by rules. Organizations as autopoietic systems consist from decisions which are binding for organizational members as long as they do not conflict with the organizational rules. Society is the encompassing social system including interactions and organizations. Luhmann defines society via communicative attainability and from this he concludes that in the present world there is only one societal system. Therefore, one should speak of World Society (Luhmann 1975, pp. 71–91) which is a stronger claim than what is postulated in globalization theories.

The theory of society is developed in Luhmann's writings in four parallel complexes. The most recent of them arose with the intensified interest in self-reference. Even society as the encompassing macro order is a self-referential system and cultivates this in self-observations and self-descriptions. Luhmann pursued these questions mainly in studies on the historical semantics of modern society (Luhmann 1980, 1981b, 1982, 1989, 1995b, 1997, Chap. 5).

The other three complexes are related to Luhmann's theory of meaning. Luhmann distinguishes a social, a temporal, and a material dimension of meaning and these three terms are connected to his theory of communications media, his theory of evolution, and

his differentiation theory as the core elements of his theory of society. These theories give an answer to the question how social, temporal, and material differences are dealt with in society.

Regarding the social dimension of meaning Luhmann's theory asks: how is it possible to mediate differences of conception and interest among participants in social processes? Language is one medium of communication effective in doing this, and since classical antiquity this function was supported by the art of rhetoric which always concentrated on persuasive effects. In modern society this function was taken over by symbolically generalized communication media such as power, money, love, and truth (Luhmann 1975b, 1988, 1982, 1990b, Chap. 4). This theory is related closely to the theory of generalized exchange media which Parsons invented (Parsons 1969). But whereas in Parsons media are specializing on mediating exchange processes between functional sectors such as the economy and the polity, in Luhmann they operate entirely internal to the function systems. All function systems, such as the economy and polity and science, are in Luhmann conceptualized as communication systems, and communication media are specialized in securing acceptability for communications internal to these systems. Luhmann postulated that in a highly differentiated society the acceptance of communicative suggestions becomes improbable and communication media realize symbolically effective ways of organizing persuasion. Money is one example of this. Why should I accept that someone becomes an owner of goods which I would like to have myself (Luhmann 1988, Chaps. 6–7)? Luhmann explains this by the symbolical effects of money. Money symbolizes the liberty to acquire something if I need to, and therefore makes it acceptable to me that some other one now buys desirable goods. He has to spend money and therefore the liberty of acquiring goods circulates in the economy in a direction opposite to the circulation of goods. Luhmann isolates comparable motivational structures for other communication media, and he formulates a complex technical apparatus for explicating internal structures of communication media and for comparing them. Binary codes (truth/falsity, legal/illegal), inflationary and deflationary phenomena (analogous to money), interrelations with bodily infrastructures of communication (sexuality, perception).

The second complex in Luhmann's theory of society regards structural changes in time. Already in his writings from the 1970s Luhmann did neither opt for modernization or development but instead for a neodarwinist theory of sociocultural evolution, closely related to the writings of Campbell (Luhmann 1990b, Chap. 8, 1993, Chap. 6, 1995c, Chap. 6, 1997, Chap. 3, Campbell 1988). There are two main motives in Luhmann's evolutionary theorizing. First, he was fascinated by the possibility of structure formation in social systems brought about by chance events. For him this was the conceptual core of evolutionary theory. Second, for Luhmann evolutionary theory was a theory of the interplay of evolutionary mechanisms such as variation, selection, and selective retention. Chance, therefore, meant the interruption of interdependencies between evolutionary mechanisms.

In conceptualizing the evolution of society Luhmann identified the variation mechanism in the possibility of saying no. Selection is effected by the binary codes of communication media distinguishing among new meaning components generated by negations. The patterns of meaning arising this way have to be stabilized by building social systems around them. Differentiated and more clearly articulated patterns of meaning enhance the probability of new negations arising, thereby demonstrating the circular interrelatedness of evolutionary processes (Weick 1979). This three-term structure of evolutionary theory mirrors the three-term-structure of Luhmann's theory of society: there is one mechanism which makes plausible how to operate on the base of chance events (temporality); a second mechanism enhances the probability of acceptance for new meaning components (social dimension); and a third which via differentiation orders the material complexity of the world (material dimension).

The third complex is differentiation theory, near to the core of the sociological tradition. The change Luhmann proposes is to interpret differentiation theory as a theory of system formation. Differentiation means that inside of systems new system/environment differences arise (Luhmann 1997, Chap. 4). There is no AGIL logic behind differentiation. New systems are inductive solutions for local or global problems. But Luhmann postulates that there is a limited repertoire of forms of system differentiation. First of all there exists segmentary differentiation. Segments such as kinship units in tribal societies are characterized by equality of status and similarity of internal structure. A second differentiation form Luhmann introduces is center/periphery. Center/periphery distinctions such as town/country always mean institutionalized and systematic differences in the control of resources and information. The third differentiation form is hierarchical differentiation or stratification. Hierarchical differentiation divides society into social systems (estates, castes, strata) which are distinguished by inequality of status. For the persons belonging to these systems they define a total context of life. Modern society, finally, is characterized by functional differentiation. By this is postulated that today it is dominated by macrosystems such as law, science, mass media, the polity, and the economy which are worldwide macrosystems and which are related toward one another by extreme material diversity. No stratification or status order among these macrosystems can be observed. Functional differentiation is the central empirical hypothesis of Luhmann's theory of society. It may be predicted that the future influence of systems

theory on sociology will to a considerable amount depend on the success of functional differentiation as a diagnosis.

3. Profile of Theory Building—Relevance for Contemporary Social Sciences

One of the most conspicuous features of the writings of Niklas Luhmann is a strong belief in the validity and cognitive strength of theoretical thinking. Theory for him primarily meant conceptual work, and in this he had close affinities to the cognitive style of Roman law and of philosophy. Luhmann never would have considered himself a philosopher but for him the philosophical tradition was a treasure trove of conceptual resources which sociology only could neglect at considerable cost. One example is his transformation of the blunt Parsonian concept of contingency which simply meant dependence into the scholastic concept which was defined by the double negation of chance and necessity. This way he created a more flexible conceptual instrument.

Interdisciplinarity was in Luhmann's case a logical corollary of his stance toward the philosophical tradition. In the development of Luhmann's theories he imported concepts continuously from disciplines as heterogeneous as social psychology, cybernetics, linguistics, biology, and logic. And the conceptual structures and theories which he imported were then sociologized completely. A remarkable case in point is autopoiesis, a biological concept well adapted to circular production processes of biological macromolecules in the living cell, which Luhmann transferred to the communication-based function systems of modern society. This is a controversial operation but on one point there can't be any dispute: there are not even traces of materialism or biologism in the sociological theory resulting from the transfer. Marxism was an important context in the development of Luhmann's theories. When Luhmann's scientific work became visible in the late 1960s the intellectual scenery in Germany was dominated by the student movement and its intellectual accompaniments. Luhmann became famous in 1971 when he published a book with Habermas, the main representative of the Frankfurt school (Habermas and Luhmann 1971). In the following years the career of systems theory in Germany was coupled with the demise of Marxism. After the mundane and middle-range sociology of the 1950s and1960s, the intellectual opposition of the late 1960s had re-established the expectation that sociology should deliver complex theories of society. When Marxism and political economy failed, systems theory was the only alternative remaining. Luhmann had no sympathy for the intellectual and political claims of Marxism but systems theory was adapted perfectly to an expectation structure looking for macrotheories of society and their universalizing claims. The structural functionalism of Talcott Parsons is the theory Luhmann started from. From the beginning this was an unusual intellectual relation. On the one hand Luhmann reconstructed point to point, often taking up minor theoretical distinctions hidden in the appendices of the essays on generalized media of exchange. On the other hand, the theory Luhmann built via this reconstructive effort is very different from the Parsonian undertaking: an open, inductive list of function systems supplanting the deductive logic of AGIL; World Society supersedes the solidarity-based national communities Parsons postulates; neo-Darwinist evolutionary theory pushes away the Parsonian thinking in developmental trends; a loose, heterarchical arrangement of different theories is substituted for the hierarchical structure of the Parsonian paradigm. When today systems theory is one of the few elaborated and universalistic paradigms of sociological thinking which will probably have an influence far into the twenty-first century, this is due to the theory-building and theory-reconstructing venture undertaken single-handedly in the writings and teachings of Luhmann.

See also: Communication and Social Psychology; Husserl, Edmund (1859–1938); Meaning and Rule-following: Philosophical Aspects; Parsons, Talcott (1902–79); System: Social; Theory: Sociological

Bibliography

Bateson G 1973 *Steps to an Ecology of Mind*. Paladin, London
Bühler K 1934 *Sprachtheorie*. Gustav Fischer, Stuttgart, Germany
Campbell D T 1988 *Methodology and Epistemology for Social Science*. University of Chicago Press, Chicago
von Foerster H 1984 *Observing Systems*. Intersystems Publications, Seaside, CA
Goffman E 1965 *Encounters*. Bobbs-Merrill, Indianapolis, IN
Habermas J, Luhmann N 1971 *Theorie der Gesellschaft oder Sozialtechnologie?* Suhrkamp, Frankfurt, Germany
Luhmann N 1964 *Funktionen und Folgen formaler Organisationen*. Duncker & Humblot, Berlin
Luhmann N 1970/1975a/1981a/1987/1990a/1995a *Soziologische Aufklärung*. Westdeutscher Verlag, Opladen, Germany, Vols. 1–6
Luhmann N 1975b *Macht*. Ferdinand Enke, Stuttgart, Germany
Luhmann N 1980/1981b/1989/1995b *Gesellschaftsstruktur und Semantik*. Suhrkamp, Frankfurt, Germany, Vols. 1–4
Luhmann N 1982 *Liebe als Passion*. Suhrkamp, Frankfurt, Germany
Luhmann N 1984 *Soziale Systeme*. Suhrkamp, Frankfurt, Germany
Luhmann N 1988 *Die Wirtschaft der Gesellschaft*. Suhrkamp, Frankfurt, Germany
Luhmann N 1990b *Die Wissenschaft der Gesellschaft*. Suhrkamp, Frankfurt, Germany
Luhmann N 1993 *Das Recht der Gesellschaft*. Suhrkamp, Frankfurt, Germany
Luhmann N 1995c *Die Kunst der Gesellschaft*. Suhrkamp, Frankfurt, Germany

Luhmann N 1997 *Die Gesellschaft der Gesellschaft*, 2 Vols. Suhrkamp, Frankfurt, Germany
Maturana H R, Varela F J 1980 *Autopoiesis and Cognition*. Reidel, Dordrecht, The Netherlands
Parsons T 1969 *Politics and Social Structure*. Free Press, New York
Spencer Brown G 1972 *Laws of Form*. Dutton, New York
Watzlawick P, Beavin J H, Jackson D D 1967 *Pragmatics of Human Communication*. Norton, New York
Weick K E 1979 *The Social Psychology of Organizing*. Addison-Wesley, Reading, MA

R. Stichweh

Luria, Aleksander Romanovich (1902–77)

Aleksander Luria was born in Kazan, an old Russian University town east of Moscow on July 16, 1902. He entered Kazan University in 1918 at the unusually early age of 16 owing to the social disorganization (and associated opportunities) associated with the Russian revolution. He obtained his bachelor's degree in 1921 at the age of 19. In 1923 he moved to Moscow where he remained for the ensuing 54 years (with the exception of a period between 1942 and 1945 when he directed a clinic for rehabilitation of soldiers who had suffered brain injuries). During this long career he taught and conducted research at a variety of important institutions including the Institute of Psychology, the Department of Psychology at Moscow State University, the Institute of Defectology, and the Bourdenko Institute of Neurosurgery. He died in Moscow on August 14, 1977.

While an undergraduate student, he established the Kazan Psychoanalytic Association and planned on a career in psychology. His earliest research sought to establish objective methods for assessing Freudian ideas about abnormalities of thought and emotion, as well as the effects of fatigue on mental processes. His dream was to create a unified psychology that would encompass both the intimate personal thoughts and emotions of individuals, as well as the biological and social conditions that supported such processes (Luria 1979). As he summarized his notions about psychoanalysis:

> Here, I thought, was a scientific approach that combined a strongly deterministic explanation of concrete, individual behavior with an explanation of the origins of complex human needs in terms of natural science. Perhaps psychoanalysis could serve as the basis for a scientific *reale Psichologie*, one that could overcome the nomothetic-idiographic distinction (Luria 1979).

In emulation of the psychoanalytical writers, he conducted clinical research on free associations, but he mistrusted the results of such efforts, feeling that any conclusions he tried to reach about the flow of his subjects' thoughts were insufficiently grounded. As he wrote in his autobiography, 'While I was able to fill notebooks with (a patient's) free associations, I was in no position to carry out my plan to use such data to capture the concrete reality of the flow of ideas' (Luria 1979).

In response to this dissatisfaction he created a methodology designed to embody a psychodynamic theory of mind in an objective set of laboratory procedures. The centerpiece of this methodology was an experimental technique that he called the combined motor method, which, he hoped, would provide a way of rendering Freud's clinical methods accessible to experimental treatment.

The fullest existing description of this work is contained in a monograph published in English under the title *The Nature of Human Conflicts: Or Emotion, Conflict and Will* (Luria 1932). Luria explicitly rejected mechanical determinism in psychological analysis, declaring that 'The structure of the organism presupposes not an accidental mosaic, but a complex organization of separate systems…(that) unite as very definite parts into an integrated functional structure' (Luria 1932).

This approach required Luria to solve the following problem. Since this structure is the consequence of a long complicated development, both ontogenetically and cultural-historically, and because the parts are integrated into a whole functional system, how can it be possible to isolate elements in this system for purposes of psychological analysis? Phrased differently, since no two people are constructed alike, how could one possibly obtain valid evidence about the thought processes of another person?

The answer that Luria provided was that other people's thoughts can only be revealed indirectly in so far as they can be reflected in a publicly displayable, voluntary behavior. He phrased his strategy as follows:

> We should on the one hand, produce the central process of the disorganization of behavior; on the other hand, we should try to reflect this process in some [other] system accessible and suitable for examination. The motor function is such a systematic, objectively reflected structure of the neurodynamic processes concealed from immediate examination. And there lies before us the use of the motor function as a system of reflected structure of hidden psychological processes. Thus we proceed along the path we call the combined motor method (Luria 1932).

The first phase in his technique was to induce a well-coordinated, publicly available behavior as the medium of coordination necessary for the psychological analysis to be accurate. He used various devices for this purpose. Often the subject was requested to hold the left hand steady in a device that could record its movements, while simultaneously being asked to press a button or squeeze a bulb in response to verbal stimuli

presented by the experimenter. Once this behavior became stable, the analyst sought to disrupt it selectively in line with his hypothesis about particular internal psychological states.

The combined motor method was applied to psychodiagnosis in a wide variety of real-life circumstances consistent with Luria's goal of demonstrating the possibility of a methodology powerful enough to reach beyond the laboratory to engage the kinds of long-term emotions that typically organize human behavior. For example, it became the basis for early lie detector tests and later for studies of the role of speech in the development of normal and mentally retarded children as well as adults suffering various neuroses and brain damage. It also served as the foundation for a great deal of his subsequent research on the relation of culture and human thought. As Luria (1932) expressed this relationship at the end of *The Nature of Human Conflicts*, 'The analysis of complex cultural mechanisms is the key to the understanding of the simple neurodynamical processes.'

In 1924 Luria had met Lev Vygotsky, a developmental psychologist whose influence was decisive in shaping Luria's future career. Together with Vygotsky and Alexei Leontiev, Luria sought to establish an approach to psychology that would enable psychologists to 'discover the way natural processes such as physical maturation and sensory mechanisms become intertwined with culturally determined processes to produce the psychological functions of adults' (Luria 1979). Vygotsky and his colleagues referred to this new approach variably as 'cultural,' 'historical,' and 'instrumental' psychology. These three labels all index the centrality of cultural mediation in the constitution of specifically human psychological processes, the role of the social environment in structuring the processes by which children appropriate the cultural tools of their society in the process of ontogeny, and the need to study processes of change over time. An especially heavy emphasis was placed on the role of language, the 'tool of tools' in this process: the acquisition of language in early childhood was seen as the pivotal moment in ontogeny when phylogeny and cultural history are merged to form specifically human forms of thought, feeling, and action. From the late 1920s until his death, Luria sought to elaborate this synthetic, cultural-historical psychology in different content areas of the discipline.

During the 1930s he conducted two lines of research aimed at elucidating the relationship between phylogeny and cultural history in children's development (this topic is often referred to as the 'nature–nurture question'). The first project falls within the scope of what is now referred to as 'cross-cultural psychology.' It seeks to hold phylogeny constant while studying the effect of rapid cultural change.

In one implementation of this strategy, Luria led two expeditions to Central Asia where he investigated changes in psychological functioning brought about by changes in economic activity, social organization, and literacy that were then taking place (this work was only published after a long delay resulting from its ideological implications). In this research he presented psychological tasks believed to tap such areas as perception, reasoning, problem solving, and self concept to groups of adults who had, or had not, been incorporated into the Soviet system of collectivized farming or formal system of schooling. He found a marked difference in tested performance as a function of these variations in life experience. The thinking of those people who remained isolated from modern forms of agricultural work or schooling appeared to be dominated by immediate practical concerns bound to particular situations while those who had undergone the modernizing changes sponsored by the state appeared to think in more general and abstract ways that could be applied across a wide variety of situations.

A second line of work on the nature–nurture issue focused on comparisons between identical and fraternal twins exposed to different regimes of deliberate instruction. This work was organized at the Institute of Medical Genetics where hundreds of twins attended a specially organized boarding school. One set of studies involved the performance of identical and fraternal twins of different ages on tasks designed to vary in the extent to which they depended upon cultural knowledge. Luria's basic finding in this work was that the correlation between the performances of identical twins were higher than those of fraternal twins when the tasks they were presented required relatively little specialized cultural knowledge. But the more the tasks required such knowledge, and the older the children were, the closer the correlations between fraternal twins approached the correlations among identical twins. He concluded that as children grow older and acquire more cultural knowledge, the influence of phylogeny over development gives way to the influence of cultural history.

He also conducted a special case study of a pair of identical twins who had developed what appeared to be their own private language before entering the boarding school as young children. For several months one of the twins was provided specialized training in the use of language to guide thinking about a variety of intellectual tasks while the other was not. The twin who received specialized training showed markedly more rapid intellectual development than the untrained twin, although the untrained twin's development rebounded once he, too, was provided specialized training. This result provided converging evidence on the role of culturally organized experience, especially language, on cognitive development (Luria and Yudovich 1959).

In the late 1930s, largely to remove himself from public view owing to the period of political purges initiated by Joseph Stalin, Luria entered medical school where he specialized in the study of aphasia, retaining his focus on the relation between language,

thought, and self-regulation, but now in the context of the politically neutral arena of medical practice. The onset of World War II made his specialized knowledge of crucial importance to the Soviet war effort, and the tragic widespread availability of people with various forms of traumatic brain injury provided him with voluminous materials for developing his theories of brain function and methods for the remediation of focal brain lesions. It was during this period that he developed the systematic approach to brain and cognition that has come to be known as the discipline of neuropsychology.

Several ideas were central to Luria's approach to brain-behavior relationships. First, he conceived of the brain as consisting of three, global 'blocks' which underpin and control three associated kinds of brain processes. The first block, sometimes referred to as the 'primary sensory areas,' controls perceptual and motor integration. The second block, which includes the prefrontal and frontal areas, undergirds programming and executive control of thought and action. The third block, which includes the brain stem and paleocortex, controls the level of arousal and activation of the cortex as a whole. In addition to considering the workings of each of these blocks, he was particularly interested in the ways in which the 'arousal-activation' block interacts with the other two.

Within this general framework, Luria sought to overcome a classical dichotomy between those who believed in the strict localization of brain functions and those who emphasized the 'equipotentiality' of different parts of the brain in subserving various psychological functions. While he accepted the idea that elementary functions such as the registration of sensations from the retina could be localized in particular groups of cells, he insisted that the systemic structure of complex forms of mental activity characteristic of adults humans represent complex functional systems involving the coordinated action of many anatomical locales. These brain systems are themselves mediated by the culturally organized environment, especially its systems of signs. As a consequence, he argued, 'to understand the brain foundations for psychological activity, one must be prepared to study both the brain and the system of activity' (Luria 1979). This insistence on linking brain structure and function to the proximal, culturally organized, environment provides the thread of continuity between the early and later parts of Luria's career. Third, he combined clinical practice with basic research on brain function, taking his ability to create systems of remediation as crucial evidence concerning the adequacy of his theories.

Following the war Luria sought to continue his work in neuropsychology. However, his plans were interrupted for several years in 1948 when he was removed from the Institute of Neurosurgery during a period of particularly virulent antisemitic repression. For the next several years he pursued his scientific interests through a series of studies of the development of language and thought among mentally retarded children at the Institute of Defectology. This work used a variation on the combined motor method to explore how the development of self-control is linked to the development of language, and provided a theoretical basis for the development of special curricula for the mentally handicapped.

In the late 1950s Luria was permitted to return to the study of neuropsychology, which he pursued until his death of heart failure on August 14, 1977. During this period he continued developing the idea of the brain as composed of functional systems closely linked to the conditions of the proximal environment, producing a number of important monographs on the brain bases of language, memory, and cognition. As this work became known outside of Russia, it gave rise to the development of standardized diagnostic tests (often referred to as the Luria Neuropsychological Battery) which are still widely used in clinical work. However, with the advent of highly sophisticated neuroimaging methods the diagnostic function of these tests has decreased in importance. What has remained internationally influential are the basic theoretical principles concerning the organization of higher psychological functions into complex functional systems mediated by sign systems and the use of these principles to guide the development of remedial procedures for brain damaged patients.

In the years just prior to his death, Luria returned to his early dream of constructing a unified psychology that overcame the centuries-old split between natural and cultural science approaches to psychology. He proposed to do this by showing how theoretical principles arrived at through the experimental study of groups of people ('nomothetic' knowledge) could be made relevant to understanding and coping with the concrete life circumstances of an individual human being (idiographic knowledge). He called the approach combining the two psychologies 'romantic science.' He contrasted romantic science with what he called classical science:

> Classical scholars are those who look upon events in terms of their constituent parts. Step by step they single out important units and elements until they can formulate abstract, general laws.... Romantic scholars' traits, attitudes, and strategies are just the opposite. They do not follow the path of reductionism, which is the leading philosophy of the classical group. Romantics in science want neither to split living reality into its elementary components nor to represent the wealth of life's concrete events as abstract models that lose the properties of the phenomena themselves.

He illustrated his concept of Romantic Science in two longitudinal case studies involving people for whom ordinary ways of mediating action in the world is impossible (Luria 1968, 1972). One was a man with a superb, but unusually organized memory. The other was a man who had suffered unusual disorganization

of memory owing to massive destruction of the right-posterior part of his brain. In each case Luria combined information from experimental studies of large groups of subjects with the peculiarities of the individual case. His ideas about how the two kinds of processes combine was evaluated by the therapeutic regimes he prescribed.

These two case studies illustrate in dramatic detail his blend of classical, experimental approaches with clinical and remediational approach, a synthesis that stands as a model for psychological research and theory to the present day.

See also: Aphasia; Cognitive Neuroscience; Comparative Psychology; Cross-cultural Psychology; Emotion: History of the Concept; Experimental Laboratories: Social/Pyschological; Mnemonics, Psychology of; Neuropsychological Functioning, Assessment of; Pavlov, Ivan Petrovich (1849–1936); Psychiatry, History of; Psychoanalysis in Clinical Psychology; Psychoanalysis: Overview; Psychology: Historical and Cultural Perspectives; Psychology: Overview; Psychotherapy, History of: Psychiatric Aspects; Psychotherapy Process Research; Vygotskij, Lev Semenovic (1896–1934)

Bibliography

Cole M, Maltzman I (eds.) 1969 *A Handbook of Contemporary Soviet Psychology*. Basic Books, New York

Goldberg E (ed.) 1990 *Contemporary Neuropsychology and the Legacy of Luria*. Erlbaum, Hillsdale, NJ

Luria A R 1932 *The Nature of Human Conflicts*. Liveright, New York

Luria A R 1966 *Higher Cortical Functions in Man*. Basic Books, New York

Luria A R 1968 *The Mind of Mnemonist*. Basic Books, New York

Luria A R 1970 *Traumatic Aphasia: Its Syndromes, Psychology, and Treatment*. Mouton, The Hague

Luria A R 1972 *The Man with a Shattered World*. Basic Books, New York

Luria A R 1973 *The Working Brain*. Basic Books, New York

Luria A R 1979 *The Making of Mind*. Harvard University Press, Cambridge, MA

Luria A R, Yudovich F A 1959 *Speech and the Development of Mental Processes*. Staples Press, London

M. Cole

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Land Information System

1. History

The concept of the Land Information System was first developed during the 1980s, but its roots go far back to ancient times. Its present-day evolution is a direct result of two simultaneously occurring phenomena of the closing decades of the twentieth century, one being the rapid growth of the world's human population and the pressure that this growth has had and is having on the planet's dwindling land resources. The other continues to be the prodigious progress in the field of computer-based information systems. Before reviewing what a Land Information System could look like today, let us briefly look back at its origins and appreciate that many of the principles maintained at present have come to us from many centuries ago.

There can be little doubt that humans have always in one way or another classified and arranged their immediate and surrounding territorial space. This endeavor is a necessity even for an individual acting alone, and more so for groups or clans working in unison. We can very well imagine prehistoric cave dwellers designating areas for such purposes as cooking, sleeping, storing, and gathering—together within the confines of their individual and shared living space. Over time, as societies grew in numbers and in complexity, accompanied by the conscious setting aside of spaces or units of land within which to perform specific activities or functions, the task of preparing an accounting system designed to establish formally and govern these activities or functions became an inescapable necessity.

A detailed history of the evolution of land recording is not readily available. What we do have is evidence that as a formal activity land accounting or inventorying is several thousand years old. For example, there is proof in the form of a Chaldean tablet prepared some six thousand years ago. The same may possibly represent some form of agricultural inventory, describing the production of different crops within distinct areal units, for a particular harvest period. This singular type of registry, performed over time, could have undoubtedly given the ancient Babylonians a detailed view of their food production in the Tigris Euphrates river valley, which in turn allowed this form of wealth to be planned for and dealt with accordingly.

There is also evidence that the ancient Egyptians, Greeks, Romans, and in America the Incas and the Aztecs, maintained elaborate land records. As for Asia, the Chinese have a millenary tradition in the establishment of formal accounting and controls for the administration of land. The Romans, for example, were particularly active in the preparation of 'cadastral' records. Not only would they prepare and maintain land registers in outlying provincial cities and towns but also copies were remitted to Rome to be consigned to a central registrar. During the reign of Roman emperor Diocletian (AD 284–305) the *jugum*, a tax on a unit of cultivable land, was imposed. Besides requiring extensive surveying to be undertaken and periodic assessments, it established a linkage between the levying of taxes and the ownership and productivity of land. As empires and kingdoms grew in size and complexity, the inventory of land, and thus the capability of exercising controls over this most important source of wealth and power, was indispensable to their survival. But it was not until the end of the eighteenth century and the beginning of the nineteenth that what can be qualified as a 'modern cadastre' began to be implemented, particularly in continental Europe. The driving force behind its development and implementation was the taxation of land and its component buildings, and thus its name 'tax cadastre.'

The modern cadastral survey, with its concomitant cadastral record, is the mainstay of any land information system of the present or near future. What must be kept in mind is that no single cadastral system is suitable for all cases due to the enormous disparities that presently exist between national cultures, jurisprudence, and available material and human resources. It is clear that a cadastral survey must be 'tailored' to particular circumstances, but maintaining universally recognized methodology and content. In short, it is an ongoing process that over time can acquire the additional elements that will eventually result in a modern system with all its inherent advantages.

2. The Cadastral Survey

The term cadastre is believed to have derived from the combination of two late Greek words, *katá* (according to) and *stíchon* (line). At the present time, most land recording activity is undertaken by local or municipal public authorities under the guise of the cadastral survey, performed by the 'cadastral office.' Its primary purpose has been for the most part that of establishing the foundation for the implementation and collection of the property tax, predominantly at the urban level.

The cadastre is a parcel-based system, with the gathered information referenced to unique and well-defined units of land. A land parcel is defined to be 'a continuous area of land (and its components) in which unique and homogenous interests are recognized.' Each parcel is assigned an exclusive parcel number or identifier, generally by means of a geocodification system, and all are graphically described by means of a large-scale cadastral map.

The data collected for each unique parcel, and linked directly to the cadastral map via the parcel number, generally come under four headings: (1) Location attributes; (2) Physical attributes; (3) Economic attributes; and (4) Legal attributes.

2.1 Location Attributes

The location attributes fix the land unit's position on the cadastral map, be it at a small scale (1:20,000) such as a metropolitan area, or with a large scale (1:500) of a city block or the basic unit, the parcel. There are three basic ways of locating a parcel, all of which should, if at all possible, be simultaneously present in the parcel's cadastral record. These are the tangential location, the coordinate location, and the areal location. Each one merits a brief individual description.

Most if not all nations utilize the tangential location for the positioning of individual properties. This age old system is simply the commonplace 'street number or address,' or some other similar locating element or counterpart. The premise is that the parcel's location is a function of the means or method of physically accessing it. Be it for example a street, avenue, or alleyway, along with an additional component, usually a number, it is possible to identify the parcel's singular location. For even greater reference, accompanying political administrative elements are included to give the parcel a global position, example: 1900 Pennsylvania Avenue, Washington, District of Columbia, United States of America, the address of the White House.

Modern cartography has always been dependent on astronomy and the use of geographic coordinate systems. But in spite of being available in one form or another for hundreds of years, the use of coordinate location as a component of the cadastral survey is actually a twentieth century occurrence. Its inclusion must be considered an important advancement in the steady progress and modernization of any cadastre system. The confusion often present in the exclusive use of tangential location, especially in rural areas or regions of difficult access, can most often be satisfactorily resolved by means of coordinate location. Its presence in the cadastral survey can be of two forms, one being a single coordinate reference point within the boundaries of the parcel. The other presence is generally referred to the cadastral map, where the coordinates of each of the existing vertices, or corners of the closed polygon, are clearly labeled.

Areal location refers to the linkage between the total space of the parcel with a technically established identifier or geographic code. The advances in the development of geocodification have occurred simultaneously with the progress in the field of computer technology. Assigning a unique numeric code to each parcel not only enormously facilitates the organization of the data files but also provides for rapid access to both graphic and alpha numeric information, and allows for the cross-referencing of information of a particular parcel from various sources. There is no single unique way of designing what can be called a cadastral location code, but they generally fall into two categories: the technical cadastral code and the simplified cadastral code. The former is based on the use of quadrants of equal size, using as the starting point, for example, maps based on the UTM coordinate system. The advantage of this method is that extensive geographical areas can be associated with a common cartographic reference. The disadvantage is that its complexity renders its use difficult, except to those especially trained. The latter, the simplified cadastral code, has as its main advantage in that it utilizes commonplace geographic elements in establishing the boundaries of the different levels of spatial aggregation, thus making it easy to understand and use. Its disadvantage is that the commonplace is vulnerable to change. Regardless of the method to be chosen, a numeric code of no less than 10 digits can easily link land data with its corresponding location on the cadastral map.

2.2 Physical Attributes

A single universally accepted cadastral register, or dossier, useful for all necessities and circumstances, does not exist. Each governmental authority must select the type and amount of information that it deems necessary and desirable. In order to avoid unnecessary costs and complications, the volume of information should be kept at the necessary minimum. Nevertheless, the design of the register should be modular and flexible so as to allow for adjustments over time, the implementation of positive changes that will not negatively affect the integrity, and usefulness of the existing database.

With respect to the physical attributes of each parcel, the following information should be included:

- Basic characteristics of the land such as size, shape, occupied or vacant, topography, and availability of public services.
- Basic characteristics of existing buildings such as type of occupancy (by owner, rented, unoccupied), type of building (single family home, apartment complex, commercial establishment), use, conservation, and basic structural characteristics.

If possible, it is recommended to include one or more photographs of all parcels and constructions, but especially so of the building if it is of some historical and/or architectural value. In some jurisdictions, the planning office or engineering department stores copies of the building's blueprints. If so, a reference to the location of these instruments should also be included.

2.3 Economic Attributes

For most of its existence, the cadastral survey had as its central and most important function the calculation and collection of the property tax. This tribute represents the primary source of income for many local governments, and therefore the focus of the cadastral undertaking has been very often narrowly confined to this end. But it is now clear that besides being an indispensable instrument for the levying of the property tax, the economic attributes reflected in the cadastral survey can be of significant assistance in the formulation of a multifaceted economic policy, embracing not only the traditional financial aspects but also as an instrument for the planning of growth and development. Examples of the latter include the use of tax incentives and disincentives in order to direct development as a reliable reference for the correct execution of eminent domain, and as a complimentary tool for the calculation of value added to property as a result of investments in public infrastructure.

The financial health of many local governments may very well depend upon an effective implementation of a dynamic, efficient, and fair property tax code. To accomplish this end, the cadastral office must establish a working methodology that will enable it to continuously adjust values to reflect changes in the land market (the inflation and depression in the price of real property), effectively charge and collect the corresponding tax from each property holder, and finally, to ensure that the amount collected is reasonable and in accordance with the law. It is clear that this enormous and complex task cannot be adequately accomplished without automating the land data base.

The initial problem in obtaining the economic attributes of the parcel is in defining the concept of value. There exists no single accepted formula for accomplishing this end, but regardless of the one adopted, it must be flexible enough so as to make adjustments for the purpose of perfecting it over time. Once defined, a value is then assigned for each parcel (individually to the land, and if present, in combination with the building components), and it is then possible to calculate the appropriate tax. Generally, the manner of assigning value to each individual parcel is accomplished by means of a two-tiered process, beginning with a massive appraisal to be followed by an individual appraisal.

Whenever there exists a general backlog in the cadastral record, which may very well be the case in most of the world's municipalities, the recourse to the massive appraisal is unavoidable. Its principal advantage is that it permits an initial calculation of value and the assigning of a tax duty to a large number of properties in the short term. Its main disadvantage is that to ensure fairness to the taxpayer, the massive appraisal should lean towards minimum values so as not to incur the injustice of overtaxing. This in turn represents a lower tax income than would be the case if the survey were reasonably up to date. Consequently, the median-term goal should be to undertake systematic individual appraisals so as to insure that higher valued properties are appraised first, and that eventually each contributor's property tax be in correspondence to the market value of their holdings.

2.4 Legal Attributes

It is often the case that the cadastral record refers legal ownership of each individual parcel to the prevailing jurisprudence governing the Land Registration System. Nonetheless, the cadastral record should include the following basic information concerning tenure and use:

- The name, address, and phone number of the property's presumed owner, and if applicable, the same data as to its administrator. The owner in turn can be of several categories, including private individual, government agency (national, state or provincial, local), community, corporate, condominium, and mixed (national and local government ownership for example).
- The nature of the occupancy of the parcel, for example, whether the same is unoccupied, occupied by owner, rented, entrusted, or squatted.
- The present use given to the property in accordance to an established system of classification of land use activities. These uses in urban areas generally include the following: vacant, residential, commercial, industrial, governmental, religious, transportation, and cultural-recreational. In non-urban or rural areas, uses such as agricultural, mining, and forestry are classified. The legal importance of determining existing use is related directly to the established zoning ordinance and/or approved land use plan, and thus whether or not the use in question is in compliance, as well as the effect the prescribed zoning has on its tax bracket.

2.5 Multipurpose Cadastre

Around the 1930s and 1940s, the term 'multipurpose cadastre' first appeared. The expression 'modern cadastre' could just as easily be used in its place. Regardless of whichever term is employed, they both illustrate the progress the cadastral survey has experienced in the twentieth century. An important aspect of this progress has undoubtedly been the enormous advances in the automating of both the mapping, and the data collection and maintenance, the 'how' aspect. But equally important, if not more so, have been the advances made with respect to 'for what end' the cadastral survey is undertaken. The systematic inclusion of information such as land use,

soil assessment, public services, and historical conditions allow for the cadastral record to be of singular usefulness in the implementation of land management strategies and the preparation of urban and regional plans. The purpose is no longer just fiscal, but managerial and planning as well.

One must not confuse Multipurpose Cadastre with Land Information System. Whereas the cadastre is generally a public sector endeavor, with laws and ordinances governing its purpose and execution, a Land Information System is the outcome of an understanding attained by numerous participants from both the public and the private sector. Of these participants, there can be no doubt that foremost from the public sector is the role played by the cadastral office. But to establish a Land Information System effectively, the presence of another public sector undertaking is required, the one performed by the entity responsible for maintaining the Land Registration System.

3. *Land Registration*

Land Registration is concerned with land tenure, which is 'the nature of the legal estate in land,' or simply stated, the rights by which an individual or groups lay claim to landed property. Internationally, there exist innumerable ways of conceptualizing the character and manner of existing rights over land, but generally they are an expression of a particular combination of 'customary law' and of 'formal law.' In most States, this expression is manifested by means of one of two acknowledged systems of Land Registration, the Land Deed and the Land Title.

A Land Deed is a 'writing or document executed under seal and delivered to effect a conveyance,' with conveyance signifying the transfer of property from one entity (person or party) to another. This in turn can result in the establishment of a Deed Registration System, which 'means that the deed itself, being a document describing an isolated transaction, is registered. This deed is evidence that a particular transaction took place, but it is in principle not in itself proof of the legal rights of the parties to conduct that transaction.' Thus in order to strengthen or guarantee the purchaser's rights of ownership, it is recommendable to trace backwards the property's 'root title.' This can generally be accomplished when a close relationship is established and maintained between a systematic deed registration archive and a functioning cadastral record.

A Land Title entails the registration of the consequence of a transaction, that is, the 'title' or the acquired right in and of itself. It is by means of a publicly accepted registration that the acquired right is produced. The registration of property (land title) is a public institution 'for the purpose of registering the possession of title and the conditions of dominion over a particular landed property, for the exercising of the contractual obligations over the same and as a guarantee for the contracting parties, not only with respect to the property in and of itself, but also to the personal circumstances of the proprietor.' From a juridical point of view, some argue that the relation between deed registration and title registration is comparable to the relation between legal facts and legal consequences.

To be effective, all land registration systems should abide by the following three principles: (1) The Mirror Principle—the land register is the true reflection of the legal status of the property; (2) The Curtain Principle—no further investigation into the root title of the property is required; and (3) The Insurance Principle—the State guarantees to all parties the validity of the information thus subscribed, and is liable to honor any valid disclaimer presented unto the register. In order to best insure the upholding of these three principles, there must be cooperation and coordination of the activities performed by the surveyor, the notary, and the registrar. Each of these parties should have a clear understanding of the role played by the other two if the land recording process is to be effective and reliable.

Almost all land registration systems are exclusively descriptive and legalistic. In order to strengthen the important relation between the 'legal object' (parcel) with its location and physical attributes, modern legislation should require a direct relationship between the register and the cadastral office. Some would argue that the most recommendable direction should be toward an eventual fusing of these two activities. Although the advantages of doing so appear to be obvious, the benefits of having them separate should not be overlooked. Land registration worldwide is fraught with abuse and fraud, due in great measure to the absence of oversight and checks-and-balances between the different actors in the process. Rather than fusing, which would concentrate the decision-making process and thus facilitate possible arbitrariness and corruption on the part of public officials, improved coordination and integration between the register and the cadastral record should improve the detection of irregularities and incongruities when and wherever these may arise. The more open a Land Registration System is to public inspection and scrutiny, the less vulnerable the information it contains.

4. *Public, Private, and Mixed Land Recording Systems*

In Brazil it is called *logradouro*, and it means 'public way, street, public square or garden.' When these spaces, or any nonparcel areal unit is officially surveyed and recorded, it is called *cadastro de logradouro* (public way cadastre). As the management

of urban space has become more and more complex, the importance of having and maintaining information of these and other shared spaces is paramount. But to do so the participation of both the public and private sector, or a combination of the two (mixed), is essential.

The public park or square is incorporated into the cadastral record generally in the form of the 'city block.' Not so the highway, street, or alleyway. Although viewed as a separate entity, these 'spaces of movement' are clearly an important part of the urban landscape. So much so that some studies of major cities in the United States have calculated that streets occupy between 20% and 40% of the land surface of the Central Business Districts. By no means should this significant portion of urban space be excluded from the land recording process.

The preparation of a public way cadastre requires the participation of numerous parties, most of which are involved with the provision of public services. Basic street infrastructure (sewer, water, drainage, pavement, sidewalks, and sidewalk furniture) is generally the responsibility of the public sector. On the other hand, services such as electricity, telephone, and television cable, all of which utilize the public way, are very often provided by private sector companies. But either way, all of them are subjected to some form of public regulation, mostly with respect to the manner the particular service is physically established. Until recent years little importance was given to 'centralizing' the information of these various service providers, a step that would have set the stage for improved planning, coordination, and control over the public way. And even if regulations were in place to create a unified public way cadastre, unless all parties involved agreed to a common geocodification base, the goal of establishing and maintaining a public way cadastre would be seriously hampered.

In the early 1970s, J. Brian McLoughlin addressed the importance of establishing a single cartographic reference base when he stated the following: 'What we wish to reiterate is the need to bring all data to a common system of areal units; unless this is done its utility is very considerably reduced, indeed it may be valueless. This is because of our cardinal need to understand activities within spaces from which it follows that information on all aspects of activity (e.g., the familiar conventional measures of population, employment, land value, ownership, etc.) as well as the spaces which support them must be coincident and congruent; this principle must be maintained strictly.' At the present time this principle of commonality is absent in most societies. National government agencies, local authorities, universities, public service providers, and private land management agencies more often than not design their very own particular geocodification system, with little or no agreement with one another. This veritable Tower of Babel approach to land-related information gathering inevitably results in diseconomies and high opportunity costs when approved land policies lack an up-to-date and integrated Land Information System.

In the case of rural areas the same principle applies. There can be no doubt that the particular attributes of nonurban spaces require a different approach with respect to the type of information to be gathered and the manner in which this is accomplished. Nevertheless the convenience, economy, and usefulness of common areal units for the preparation of 'ecological inventories' (term utilized by Ian McHarg in his landmark work *Design with Nature*) that 'would include human artifacts as well as natural processes,' should not be ignored. It is not uncommon to see government agencies, as for example those responsible for the protection of the environment and for the preparation and implementation of agriculture policy, work in almost complete isolation from one another. The absence of a common land database not only results in important diseconomies when there is duplicity present in the information gathering process, but more importantly, the lack of coordination and information sharing can oftentimes result in the formulation of policies by the different parties that are in direct conflict with one another.

Governments should waste no time in working towards establishing a unified geocodification system, the design of which should include the active participation of all the nonpublic actors involved. But should for some reason or another the creation of a common system of areal units be unattainable, efforts should be made so as to enable the translation, or transformation, of information from all relevant information systems to one another. Clearly, all efforts toward achieving commonality by every agent involved in gathering land information would contribute positively towards the effective implementation of a Land Information System.

5. *Land Information System*

Unfortunately today very few countries have a working Land Information System as has been defined. The required steps that lead towards achieving this goal involve significant economic investments and a determined political will on the part of the society and government. Many would argue, and with some reason, that the larger and more unified an information system is the greater the dangers to individual privacy and liberty will become. But the world we live in today cannot afford to be without well-designed land recording systems and to continue as it has with inadequate land management strategies. With the proper measures in place, it is possible to realize an acceptable balance between individual rights and the public good. Without a working Land Information System, nations individually and collectively will be

ill-adapted to manage future population growth with its ever-increasing demands on the planet's limited land and natural resources. The first quarter of the twenty-first century will be crucial to this end, and should be a period of human history in which governments give land recording systems high priority in order to be better prepared to face the challenges of growth and development in the years ahead.

What does the future hold for Land Information Systems? This question must be addressed under two headings, technological advancements and scope. In the short term, the next 15–25 years will continue to experience advances in computer sciences and their applications in the development of extremely large data base systems. With respect to mapping, which will inevitably become more digitilized and automated, the mission of Space Shuttle Endevour of February 11, 2000 is a harbinger of things to come. Titled *Shuttle Radar Topography Mission*, in a period of eleven days, close to one trillion measurements of the Earth's topography were performed from an orbit of 126 miles. Approximately 80% of the planet's land mass will benefit by the availability of 'unrivaled 3-D images of the Earth's surface.' This information will contribute significantly to improved urban and regional planning, as well as the development of environmentally sound policies.

With respect to scope, the very distant future can very well resemble something like what Isaac Asimov described in one of his novels 50 years ago: 'Biron leafed through the *Standard Galactic Ephemeris* with something like despair. Tens of thousands of stars were listed in detail, with their positions crammed into three figures. There were hundreds of pages of these figures, symbolized by the Greek letters ρ (rho), θ (theta) and ϕ (phi)... .ρ was the distance from the Galactic Center in parsecs; θ, the angular separation, along the plane of the Galactic Lens from the Standard Galactic Baseline (the line, that is, which connects the Galactic Center and the sun of the planet Earth); ϕ, the angular separation from the Baseline in the plane perpendicular to that of the Galactic Lens, the latter measurements being expressed in radians. Given those three figures, one could locate any star accurately in all the vast immensity of space.' No person can foresee how humanity will one day manage the information on the scale of the vastness of the Milky Way Galaxy, but what is clear from Asimov's speculations about the very distant future is that the need to classify and arrange the space in which the human race moves has been, is, and will always be a necessity.

Before we can even seriously speculate on the remote future, it is possible to imagine that over the next century Land Information Systems will most probably evolve in a manner so as to include the whole of the biosphere, with special attention given to the inclusion of detailed data concerning the ocean floor. As land resources continue to dwindle, the oceans will become more important as a source of materials required by human progress. The sound management of the exploitation of the resources under the surface of the oceans will require a well-designed information system, integrated necessarily to all the existing information of the land mass. Once accomplished, this new scope will probably result in the Land Information System, along with the Ocean and possibly Atmosphere Information systems, becoming not the end product of today, but instead a component of a larger information system. This future inventory could very well be titled Biospheric Information System.

Bibliography

Asimov I 1951 *The Stars, Like Dust*, Doubleday, New York

Brown L A 1979 *The Story of Maps*. Dover Publications, New York

Direction Générale des Impôts, Service de l'Administration générale, Sous-Dirction du Cadastre et de la Publicité foncière 1981 *La Documentation Cadastrale Miniaturisée*. Service de la Documentation nationale du Cadastre, France

Fundación para el Desarrollo de la Comunidad y Fomento Municipal (FUNDACOMUN) 1978 *Memoria VII Reunión de Directores de Catastro*, Barquisimeto, Venezuela

Hawerk W 1990 *Cadastral Systems in Developing Countries (Technical Options)*

Henssen J L G, Williamson I P 1990 Land Registration, Cadastre and its Interaction—A World Perspective. First published in *Proceedings of FIG, XIX Congress*, Helsinki, Finland, 10–19 June 1990, pp. 14–43

Institute of Transportation Engineers. 1976 *Transportation and Traffic Engineering Handbook*, Prentice-Hall, New Jersey

McLoughlin J B 1970 *Urban and Regional Planning: A Systems Approach*. Faber, London

NASA Status Report-On Line, Mission: STS-99—Shuttle Radar Topography Mission (SRTM), John F. Kennedy Space Center Space Shuttle Status Report, Friday, Februrary 11, 2000

Ossorio M 1984 *Diccionario de Ciencias Jurídicas, Políticas y Sociales*. Editorial Heliasta, Buenos Aires, Argentina

Quintero J R 1984 *Catastro Integrado y Planificación Metropolitana La Sistematización de los datos urbanisticos de Maracaibo*. III Curso de Administración Metropolitana, Fundação Getulio Vargas, Rio de Janeiro, Brazil

J. R. Quintero

Copyright © 2004 Elsevier Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-07

Law, Diffusion of

1. *Definitional Issues*

The definitional issues adumbrated by the concept of 'Diffusion of Law' need preliminary clarification. So diffuse a concept overlaps with other titles of this chapter; also, particular definitions of the concept may contradict each other, implicitly if not explicitly. 'Diffusion' suggests definitional issues of motivation, of processes, and of outcomes. Especially with respect to the last issue, 'diffusion' also suggests an affinity with a concept much in use recently; namely, 'the convergence of law and of legal systems.' This concept, while in play for over a century, today is principally associated with the globalization or regionalization of national economies, and secondarily with the acceptance of a common standard of international human rights potentially binding the state. Because of these contextual associations, a focus on the 'Convergence of Law' aspect of 'Diffusion of Law' helps clarify the definitional morass inherent in the latter term.

First, this focus serves to differentiate some otherwise overlapping and contradictory neighboring concepts, in particular, those relating to the unification of law and the harmonization of law. The latter are intentional practices, consciously articulated and consciously engaged in (Wiener 1999, 189ff). In the second place, 'Convergence' as a more operational concept of 'Diffusion' highlights that the latter term too broadly suggests both a process or tendency and a result. Convergence focuses more on the latter point, since directly or indirectly it may be the more or less successful outcome—however measured—of the more conscious efforts suggested by 'unification' and 'harmonization.'

At the same time, however, it in turn suggests two other features of particular importance in recent times. First, 'Convergence' captures the underlying polar elements of inevitability and contingency that accompany, facilitate, or hinder those conscious efforts at Unification and Harmonization. Second, it widens our focus beyond the narrow range of formal law-making organs to include the broader range of actors engaged in law production, ranging from individuals through institutions to systems, who or which express the process and shape its outcome. 'Diffusion' of necessity is of a preexisting 'other' law into a host system. To understand it as 'Convergence' more openly and instrumentally faces that reality, while simultaneously accepting the reality that two systems interact in this process, that these may if they do not necessarily influence each other, and that the process may result in a synthesis that differs from each system's starting point.

All of the foregoing facets of gaining an understanding of the concept of 'Diffusion' are displayed in the two principal forms of diffusion of law in the early stages of the modern era, stages that will not be separately discussed herein, though their influences still may be noted in the contemporary era, which is the subject of this entry. The first is the diffusion of law that accompanied the dissolution of the shadow Holy Roman Empire and the arrival of the Napoleonic Period on the European continent at the turn of the nineteenth century. Coing (1989, 56f) describes the situation succinctly:

> The loss of a transnational law [at this time] was viewed as an unnecessary constriction [of legal life]. The way out of this situation, since a ius commune no longer existed, was the study of foreign law. A new discipline, the comparison of laws, emerged.

The motives for the diffusion of law inherent in this new discipline may have differed between the two principal state actors—France and the states of the later Germany. In the former, the instrumental interest in other systems' solution of new problems may have predominated, and in the latter, the need to find a common approach to the search for the applicable law. What was common to both, however, and what therefore is relevant to this discussion, was the outcome: the reciprocal diffusion of law, in the sense of bringing into consciousness the possibility of the convergence of previously disparate law (Buxbaum 1996, 206f).

The second principal form of diffusion in early modern times was that accompanying colonialism. The more or (at later stages) less unreflective introduction of the English common law to lands as varied as Australia, India, and West Africa was an inherent feature of colonial expansion itself, and the same may be said to varying degree of the colonial powers on the Civilian side of the families of law (Zweigert and Kötz 1998, p. 218). Late examples of this wholesale approach, though now marked by more sensitivity to local conditions, would include the (invited and well-considered) introduction of major parts of the Swiss and German code systems in Atatürk's Turkey (Watson 2001, 14ff) and the (invited) French introduction of entire Codes to Ethiopia in the late 1950s (David, 1982).

The principal focus of this entry is on the diffusion of law in the contemporary era. Even to that study, however, a brief overview of some central issues concerning the political and economic substrata of law, including some enduring theoretical issues, is necessary. These issues may have been brought to the fore during the Colonial and Imperial Era one and a half centuries ago, but their currently attenuated forms remain relevant to this era of globalization.

2. *Value Issues*

Law may be seen as an expression of other, more basic forces. The first modern argument in favor of Laws derivation from the cultural seedbed of a linguistically unified nation-state is that of Montesquieu in 1748: 'The political and civil laws of each nation should be adapted in such a manner to the people for whom they are made, as to render it very unlikely for those of one nation to be proper for another ['…doivent tellement propre au peuple pour lequel elles sont faites, que c'est un grand hasard si celles d'une nation peuvent convenir à une autre]' (Montesquieu 1748, p. I/3). And soon after the French Revolution, arguing against law as social engineering, Savigny, in his polemic against Thibaut, restated this view with a power that only he, in his later capacity as Prussian Minister of Justice (and thus of legislation), was able to resist (Savigny 1814, pp. 30, 137):

> The sum, therefore, of this theory is, that all law is originally formed…by custom and popular faith, next by jurisprudence, – everywhere, therefore, by internal silently-operating powers, not by the arbitrary will of a law-giver….[The object of the historical method of jurisprudence] is to trace every established [legal] system to its root, and thus discover an organic principle….

On the other hand, the nineteenth century also saw a considerable amount of conscious shaping of the legal framework of the burgeoning national economies, as these burst the household confines of the classical *oikos*, leading a contemporary scholar to suggest that '[t]he process of economic, social, [and] cultural assimilation or integration…has been accompanied by a process of political differentiation. It is this dual development of cultural and social assimilation and political differentiation which compels us to shift the emphasis of Montesquieu's test in order to find some workable criterion that can be used to determine how far a legal institution is transplantable….'(Kahn-Freund 1974).

These quotations' juxtaposition of Culture and Politics suggests an affinity to the phenomenological juxtaposition of *Lebenswelt* and System, particularly as applied to Law by Habermas (1973/75, 1981/87). The same juxtaposition also suggests an affinity to the tenets of classical Marxism: Law as much as Culture would be superstructural elements, with Economy, recast in Marxist terms more narrowly as control of the means of production, serving as the juxtaposed Basis. To these two approaches, however, Law presents a problem of contradictions. To the former theory, Law not only is the handmaiden of System, legitimating Money and Power. Rather, it may also contain inherent, dare one say essential, attributes (often if inadequately identified as 'procedural regularity') that restrain those expressions of System. Further, if it arose from and depends on Culture/Lebenswelt in Montesquieu's terms, it may carry some of the resistance of Lebenswelt towards System. In later Marxist theory, too, Law is seen to harbor apparent if only partially autonomous power to influence Basis (Renner 1949).

Within Legal Theory's own discourse of recent vintage, autonomously innocent or disparaging of this debate between Social Theories, arguments about the capacity of Law to rise above the vagaries of national cultures—arguments central to the issue of Convergence—have been conducted both at the general level of historical and economic theories, and at the specific level of case studies. In a modern return to the famous Savigny–Thibaut debate that accompanied the Napoleonic era's codification venture, Watson (1974) and Grossfeld (1990, 1986) have pressed their respective claims of Law's autonomy over or embeddedness within Culture largely on the basis of historical evidence. And in the contentious discussions over the inevitability of legal convergence in consequence of economic globalization, some of the Marxist framing of Basis and Superstructure has been reproduced by the Economic-Determinism camp in its own framing of Efficiency as an evolutionary inevitability (Gilson 2001), while the Path-Dependency deniers of necessary convergence (Roe 1994, Coffee 1999) in their turn reproduce both the older Cultural and the newer Habermasian/Lebenswelt arguments.

Preceding and underlying these issues lies the question of whether the very concept of globalization, in its cultural and societal as much as in its economic aspects, is consciously produced or not. While this debate is not directly at issue in the description of Convergence, it influences the perception of that process as undesirable or desirable, resistible or irresistible. The former, or Wallersteinian formulation (Wallerstein 1979), is well expressed by Santos (1995, 250f):

> In view of the hierarchical nature of the world system, it becomes crucial to identify the groups, classes, interests and states that define partial cultures as global cultures, thereby setting the agenda for political domination under the guise of cultural globalization.

The latter formulation finds typical expression in Slaughter, who welcomes Convergence on the assumption that the Thibaut/Watson view of Law's autonomy can be reconciled both with the Montesquieu/Savigny Organicist position and with Habermas' assertion of Law's capacity to tame System (Slaughter 2000, p. 1124):

> Constitutional courts—or any courts concerned with constitutional issues—will be forging a deeply pluralist and contextualized understanding of human rights law as it spans countries, cultures, and national and international institutions.… All this activity…requires recognition of participation in a common judicial enterprise, independent of the

content and constraints of specific national and international legal systems. It requires that judges see one another not only as servants or even representatives of a particular government or polity, but as fellow professionals in a profession that transcends national borders.

A third view of the underlying values and forces supporting the Convergence phenomenon contends that to the extent any conscious will drives it, it is that of the less-favored members of the international community of states and societies—members who see in the end result of Convergence a shortcut to a larger share of aggregate wealth and respect than they enjoy now. This is a theme well expressed by Teubner, who concludes a review of this argument with the assertion that '[g]lobal law will grow mainly from the social peripheries, not from the political centres of nation-states and international institutions' (Teubner 1997; p. 3). Within any of these 'peripheral' states, however, it may—and not paradoxically—be the social forces/classes standing to gain the most from that increase, which in turn drive this effort. As Schneiderman (2000; p. 757) puts it, '[t]his inchoate global law is being developed by economic subsystems ''in relative insulation from politics'' outside the usual centers of gravity of legal development.'

The substantive content of the law that is being diffused or is converging of course is as varied as the substantive content of law as such. Nevertheless, put in substantive and extreme terms, we can identify two principal societal forces converging on Convergence, and each is hinted at in the preceding citations: those involved in the evolution toward a new international order of Human Rights and those involved in the evolution toward a new international order of economic activity (Delmas-Marty 1998). While the drives toward these respective orders share some common features, they also evidence one essential distinction. The movement toward a new Human Rights Order is driven mainly by conscious social forces, often already successful in achieving the embodiment of their values in the formal legal orders of their home states and as a result deriving some of their own power from the political and economic power of those states. The movement toward a new Global Economic Order is driven in the main 'from below,' by market forces that generate the demand for their legal facilitation within those very market processes. This is not to deny that, especially at later stages in the evolution toward transnational markets, these forces evidence conscious use of state power to achieve that result; the role of TRIPS within the post-Uruguay international trade structure of the World Trade Organization alone is ample evidence thereof. And in some cases, of course, the two movements may themselves converge both as to the level of their conscious introduction into legal discourse and in their intended or unintended outcomes: whether the drive to ban the industrial-level fishing that catches dolphins in tuna nets is an issue of ecology or of protectionism may depend on the perspective of the observer.

3. *Diffusion as Osmosis*

3.1 *The Hegemonic Imposition of Law Upon Formal Equals*

The experience of Germany and Japan after World War II is a paradigmatic example of this phenomenon. Two elements mark these legal 'transfers:' their emphasis upon the more political and public branches of the law, especially of constitutional and administrative law; and the degree to which their success depended upon their relatively negotiated as against relatively coerced transplantation. Especially in Germany, but to an extent also in Japan, the vanquished nation's government no longer existed, and the occupation authorities, standing in for these governments, could be said to have negotiated with themselves. In fact, however, these countries' legal experts from the beginning were in essential discourse with those authorities, and their increasingly autonomous negotiation capacity quickly became evident and influential (Symposium 2000, p.1). As a result, one can trace a successful diffusion of these components of public law to that process, as especially the history of the Basic Law of the Federal Republic of Germany attests.

3.2 *The Liberal Economy As the Political Initiator of the Diffusion Process*

The experience, especially of Central European states after 1989, exemplifies this form of diffusion. Their political transformation, once they were out of the shadow of the Soviet Union, was an internal and not an imposed process, nor has the process been either uniform or completed (Mádl 2000). The diffusion of the legal superstructure of a liberal economy into their legal systems, however, has been driven largely if not totally by the purported requirements of the Western European and American economic actors whom these states have invited or been pressured by Western political actors to invite as participants (the title of a major US legislative initiative for the CIS states makes the point: 'The Freedom for Russia and Eurasian Economies and Open Market Support Act of 1992'). In the first stage of this diffusion process, a relatively uncontested acceptance of major legislative packages occurred; as the specifics of national needs became more apparent and the timing of economic liberalization within the larger context of democratization and resource redistribution became more contested, a second stage of partial rejection of these transplants and of more nuanced adaptation came

into being. At this time, it is too early to predict any eventual and in any event temporary equilibrium.

The difference between a diffusion process that is part of a clearly marked larger political transformation, as in the above examples, and one that is part of an incremental change of national markets toward more global units is one of degree as much as it is of kind. A principal differentiation between the two, however, lies in the intermediation, in the latter case, of global or at least supranational legal regimes in this diffusion process. The paradigmatic example of this situation is that of the evolution of the General Agreement on Tariffs and Trade into the World Trade Organization, and especially in the latter's increased scope of substantive coverage from trade through technology transfer (partly correlated with trade in services) to investment.

From the beginning of the postwar (II) period, efforts of goods-, services-, and capital-exporting states to encourage the creation of legal regimes in the importing states facilitative of these processes have been marked by this internationalization of the diffusion process. The first efforts were those of UN-supported agencies such as the abortive International Trade Organization proposed by the Havana Charter, but because of internal conflicts between these exporting states (and the reluctance of the United States to enter an international organization governing trade regimes (Diebold 1952), this ambitious effort was cut back in substance and in institutional power to the GATT regime (Jackson 1969, Jackson et al. 1984). As the subject matter of these efforts suggests, the regimes they produced within municipal legal orders were largely specific elements of public law requiring the reduction of state distortion of a global liberal regime supporting the trade in goods (and later, since the Uruguay Round, in services).

On the investment side, this effort was first internationalized at the private-sector level through the Investment-Code proposals of the OECD and then formalized at the supranational governmental level by the World Bank through its sponsorship of the treaty regime institutionalized in the International Conference on the Settlement of Investment Disputes (ICSID). Here, too, what was introduced into municipal legal regimes were largely specific public-law prohibitions; in this case, of expropriation measures without adequate compensation. In recent years, this international process has doubled back into national venues, with the introduction of networks of so-called Bilateral Investment Treaty regimes. These extend the obligation to provide national and most-favored-nation treatment of incoming direct investment to the point, at least in some of them (Vandevelde 1992), of forcing 'open-door' investment-entry regimes upon the signatory capital-importing countries (Dolzer 1995).

On the other hand, international organizations also play a role in the recent transformative experiences of countries of the former Socialist camp. This is evident in the trade sector (where the desire of these countries to gain the benefits of membership in the WTO in turn requires them to adopt the domestic liberal economic policies and their legal dress that constitute the *acquis mondiale* attributes of membership). It is most pronounced, and in this differs from the immediate post-WWII situation, in the sector of public finance, in which IMF and World Bank assistance is premised on liberal and to some degree politically controversial fiscal and monetary policy regimes.

All of these revolutions or evolutions have in common the relatively unilateral push of prepackaged legal goods onto these host states for relatively nonnegotiated and often nonnegotiable acceptance. This version of diffusion of law may be characterized pejoratively in neo-colonial terms, or it may be characterized as representing and meeting the wish of the recipient nations to gain the benefits of the existing international trade and investment regimes, however legitimated a wish internally along the spectrum of more or less democratic processes in formal and functional terms. In terms of the diffusion of law, however described, it is a unidirectional diffusion. The effort of developing countries in the 1960s and 1970s to negotiate a New International Economic Order, with its implication of a different outcome in the substance of what was diffused, did not succeed at that time.

3.3 The Competition Among the Providers of the Law Being Diffused

While all more or less committed to liberal economic regimes, the states providing the legal goods, in retail or wholesale packages, are not a monolithic legal block; indeed, in their variety they illustrate both Montesquieu's and Kahn-Freund's analyses. Whose offer is accepted, or whose wares are preferred, may depend on preexisting (including in colonial times) affinities of the respective legal systems; may depend upon tying agreements that lead the state proffering desired resources to condition those offers on the condition of acceptance of specific legal regimes; or may depend 'on chance and prestige' (Ajani 1995, 1998).

This competition has had the effect of recreating a degree of bargaining power for the law-importing states, both directly, as a consequence of gaining from this competition, and indirectly, as a consequence of the increase in knowledge of the specific wares and in the sophistication of the tests of their adaptability to the host system (Ajani 1998). This has been helped and to a degree encouraged by the acceptance by the World Bank and the International Monetary Fund of the relevance of appropriate specific legal regimes to the success of their own

primary missions. The policy implications of the IMF's change of approach are illustrative: '[T]he Legal Department…has recently expanded its work to include bankruptcy and foreign investment law….The Legal Department addresses these areas of need in close consultation with the legal staffs of other international organizations…in order to arrive at a reasonable division of technical assistance tasks and avoid duplication of work' (IMF Annual Report 1994, p. 183).

Whether these large-scale legal importations following the 1989 political transformations will be more successful (with success itself perhaps differently defined by the law-exporting and the law-importing states) as a result of this competition and coordination is not yet predictable (Kranjc 1993, World Bank 1995). It may be as much a mark of success as of failure that some initial providers of these materials now pay more attention to the difficulties of the transplantation exercise even in such fields, as Company and Capital Market Law, in which transplantation has historically been seen more as a carburetor transplant between automobiles than a kidney transplant between human beings (to the historical view, Klein 1904, p. 10; to the current self-criticism, compare Black and Kraakman 1996 with Black et al. 2000).

3.4 Convergence of Laws As the Outcome of Globalization of Economies

That in 1945 the United States was the only remaining powerful producer, capable supplier of capital, and large internal market posed a challenge to legal theorists speculating on the future shape and content of legal regimes in the postwar world, at least that industrialized portion thereof not within the Socialist camp. The introduction of American versions of major facilitative structures of law, such as antitrust law and capital-market regulation to nations and systems on their way to or hoping to reach similar levels of wealth production (such as the European Economic Community of 1958) fueled the discussion of legal convergence of legal systems (including of the large systems of the Civil Law and the Common Law (Merryman 1978, p.195) and specifically of convergence onto, even the reception of, American law (Wiegand 1991)). Not only substantive law, but changes in the underpinnings of law production in legal education, in the role of the legal profession, and indeed in legal culture itself often were deemed to occur as the almost inevitable byproduct of the unique position of US law and law-related institutions. Because of its economic situation, US law often had to face problems that would in time arise elsewhere because they occurred in the US first, ranging, e.g., from issues of privacy to issues of product liability in an age of mass production. This first-exposure condition made its law a natural object of attention as the underlying economic and social realities began to be experienced elsewhere.

At the same time, however, the imperfect analogy to the Roman-Law reception of medieval times was also noted, and was predicted to set limits on any one-sided convergence of law. Both tendencies were noted by Wiegand (1991, p. 247):

> American law is not superior to European law in the sense that ius commune was to the European law of that time….The lack of th[is] degree of superiority…, however, is compensated for in other areas; in particular by the political and economic dominance of the United States. This dominance was responsible for the dissemination of the new forms of business transactions as well as for the reception of American legal terms and legal thought….One aspect [of the reception of American law] is of particular importance; namely, the fact that…American legal developments have [first] followed the requirements of the post-industrial era and of the service-dominated society.

At present, we can already see the beginnings of a new set of convergence/diffusion phenomena. One is a convergence on a new European-level substance, driven largely by the unification and harmonization imperatives of the European Community (Markesinis 1994). At its extreme, this implies even the partial move of the core Common-Law regime, Great Britain, towards a European-wide and of necessity at least partially Civil-law system. If the well-established Common-Law system as found in Great Britain is experiencing 'Continental drift,' then the second phenomenon also may be on the way to observable status—the establishing of a mixed but largely Civil-Law European Community legal regime that can rival that of the United States in terms of sophistication as well as in power (Markesinis 2000).

On the level of the global economy and the law particularly relevant to its functioning, one type of tension between the leveling effects of product and service (including above all, financial) markets will arise. This tension may be seen as the collision between global and niche markets. As Ginsburg (2000, 840f) puts it

> The single-dimensional shift toward markets and formal rules implicates the issue of legal convergence. There are valid reasons for doubting that the shift toward rule-based law is inevitable, even in an era of global capitalism. Some economies will enjoy economic niches that would be hindered by greater reliance on legal ordering….From this perspective, globalization produces results that are similar to those of modernization theory, but does so as a result of specific political and economic pressures rather than because of a universal cultural shift. In other words, reliance on law reflects local political outcomes….[M]arket solutions and particular policy responses are still being contested, and any legal convergence will remain incomplete as long as this is the case.

Another way to formulate this tension is to focus on the enduring power of political forces at the national level to immunize national markets and economic institutions from the bite of unified formal legal rules. As one writer illustrates this approach: 'Although Japan and the United States of course have differing institutional structures, if there is one place that theory would predict convergence,…[corporation law] might be it. But convergence has not occurred….Th[is] limited convergence of…rules helps elucidate recent claims in the theoretical literature that convergence of rules may not necessarily lead to convergence of corporate structures' (Symposium 2000, 5ff, Coffee 1999).

Seen this way, the convergence of law on a set of rules comprising a mix of relatively uniform (regulatory) and relatively diverse (facilitative) components may still best be explained, and predicted, by the juxtaposition of the original epigrammatic quotations cited at the beginning of this entry.

Bibliography

Ajani G 1995 By chance and prestige: Legal transplants in Russia and Eastern Europe. *American Journal of Comparative Law* **43**: 93

Ajani G 1998 Transfer of legal systems from the point of view of the 'export countries.' In: Drobnig U, Hopt K J, Kötz H, Mestmäcker E-J (eds.) *Systemtransformation in Mittel- und Osteuropa und ihre Folgen für Banken, Börsen und Kreditsicherheiten* 37. Mohr/Siebeck, Tübingen

Bebchuk LA, Roe M 1999 A theory of path dependence in corporate ownership and governance. *Stanford Law Review* **52**: 127

Black B, Kraakman R 1996 A self-enforcing model of corporate law. *Harvard Law Review* **109**: 1911

Black B, Kraakman R, Tarassova A 2000 Russian privatization and corporate governance: What went wrong? **52**: 1731

Buxbaum R 1996 Die Rechtsvergleichung zwischen nationalem Staat und internationaler Wirtschaft. *RabelsZ* **60**: 201–30

Coffee J C 1999 The future as history: The prospects for global convergence in corporate governance and its implications. *NWUnivLRev* **93**: 641

Coing H 1989 *Europäisches Privatrecht II: 19. Jahrhundert, Überblick über die Entwicklung des Privatrechts in den ehemals gemeinrechtlichen Ländern*. Beck, Munich

David R 1982 *Les grandes systèmes contemporains*, 8th edn. Jauffret-Spinosi C Dalloz, Paris

Delmas-Marty M 1998 *Trois défis pour un droit mondial.* Du Seuil, Paris

Diebold W 1952 *The End of the I.T.O.* Princeton University Press, Princeton

Diedrich F 1998 Internationale Einheitsrecht—Internationalisierung der Rechtswissenschaft? In: Immenhauser M, Wichtermann J (eds.) *Jahrbuch Junger Zivilrechtswissenschaftler—Vernetzte Welt—globales Recht.* Richard Boorberg Verlag, Stuttgart

Dolzer R 1995 *Bilateral Investment Treaties*. M. Nijhoff, The Hague, Boston

Ginsburg T 2000 Does law matter for economic development? Evidence from east Asia. *Law and Society Review* **34**: 829

Grossfeld B 1986 *Kernfragen der Rechtsvergleichung*. Mohr/Siebeck, Tübingen

Grossfeld B 1990 *The Strength and Weakness of Comparative Law.* (Weir A tr.) Clarendon, Oxford

Habermas J 1973/1975 (tr.), *Legitimation Crisis*. (McCarthy T tr.) Beacon Press, Boston

Habermas J 1981/1987 (tr.), *The Theory of Communicative Action* II. (McCarthy T tr.) Beacon Press, Boston

International Monetary Fund (IMF) 1985 Technical Assistance and Training Services of the International Monetary Fund. *IMF Pamphlet Series 43*, Washington DC.

International Monetary Fund (IMF) 1994 *Annual Report*. Washington, DC.

Jackson JH 1969 *World Trade and the Law of GATT*. Bobbs-Merrill, Indianapolis

Jackson JH, Louis J-V, Matsushita M 1984 *Implementing the Tokyo Round: National Constitutions and International Economic Rules*. University of Michigan Press, Ann Arbor

Kahn-Freund O 1974 On uses and misuses of comparative law. *ModLRev* **37**:1

Klein F 1904 *Die neuere Entwicklung in Verfassung und Recht der Aktiengesellschaft.* Berlin

Kranjc J 1993 Probleme der Übernahme ausländischer Rechtssätze in nationale Rechtssysteme. WiRO **2**: 409

Mádl F 2000 State and economy in transformation—Revolution by law in the central and eastern European countries. In: Buxbaum R, Mádl F (eds.) *State and Economy*, Vol. XVII, *International Encyclopedia of Comparative Law.* Mohr Siebeck, Tübingen

Markesinis BS (ed.) 1994 *The Gradual Convergence: Foreign Ideas, Foreign Influences, and English Law on the Eve of the 21st Century*. Clarendon, Oxford

Markesinis BS (ed.) 2000 *The Clifford Chance Millenium Lectures—The Coming Together of the Common Law and the Civil Law*. Hart, Oxford; Portland, Oregon

Merryman J H 1978 On the convergence (and divergence) of the civil law and the common law. In: Cappelletti M (ed.) *New Perspectives For A Common Law of Europe.* Sijthoff, Leyden. Boston

Montesquieu C de S 1748 *De l'esprit des loix*. Barillot, Geneva (1977 *The Spirit of Laws*. Carrithers D W comp., University of California Press, Berkeley) (Book 1 ch 3–12)

Renner K 1949 *The Institutions of Private Law and Their Social Functions*. Schwarzschild A (tr.) Routledge & K. Paul, London

Roe M J 1994 *Strong Managers, Weak Owners: The political Roots of American Corporate Finance*. Princeton University Press, Princeton

Santos B de S 1995 *Towards a New Common Sense.* Routledge, New York

Savigny F K v 1814 *Vom Beruf unsrer Zeit für Gesetzgebung und Rechtswissenschaft*. (ed 3 1840) JCB Mohr, Heidelberg (*Of the vocation of our age for legislation and jurisprudence*. (Hayward A tr.) Littlewood, London; repr. 1975 Arno Press, New York)

Schneiderman D 2000 Investment rules and the new constitutionalism. *Law & Soc. Inquiry* **25**: 757

Slaughter A M 2000 40th Anniversary perspective: Judicial globalization. *Va. J. Int'l. L.* **40**: 1103

Slaughter A M, Bosco D 2000 Plaintiff's diplomacy. *Foreign Affairs* **79**(5): 102–16, II

Symposium 2000 The reception in Japan of the American law and its transformation in the fifty years since the end of World War II. *Law in Japan* **26**: 1

Teubner G 1997 Towards a 'Global Bukowina': Legal pluralism in the world society. In: Teubner G (ed.) *Global Law without a State*. Dartmouth, Aldershot, Brookfield, USA

Twining W 2000 *Gloabalization and Legal Theory*. Butterworths, London

Vandevelde K J 1992 *United States Investment Treaties: Policy and Practice*. Deventer, Netherlands; Kluwer, Boston

Wallerstein I 1979 *The Capitalist World-Economy*. Cambridge University Press, Cambridge

Watson A 1974 *Legal Transplants: An Approach to Comparative Law*. University of Georgia Press, Athens

Watson A 2001 *The Evolution of Western Private Law*. Johns Hopkins University Press, Baltimore

West M D 2000 The puzzling divergence of corporate law: Evidence and explanations from Japan and the United States. *University of Michigan John M. Olin Center for Law & Economics* Paper #00-011

Wiegand W 1991 The reception of American law in Europe. *AmJCompL* **39**: 229

Wiegand W 1998 Europäisierung—Globalisierung—Amerikanisierung. In: Immenhauser M, Wichtermann J (eds.) *Jahrbuch Junger Zivilrechtswissenschaftler—Vernetzte Welt—globales Recht*. Richard Boorberg Verlag, Stuttgart

Wiener J, 1999 *Globalization and the Harmonization of Law*. Pinter, London

World Bank, Legal Dept. 1995 The World Bank and legal technical assistance: Initial lessons. *Policy Research Working Papers 1414*, Washington, DC

Zweigert K, Kötz H 1998 *An Introduction to Comparative Law*, 3rd edn. (Weir T tr.) Clarendon, Oxford

R. M. Buxbaum

Copyright © 2004 Elsevier Ltd.
All rights reserved.

 ISBN: 0-08-043076-07

M

Machiavelli, Niccolò (1469–1527)

Machiavelli became a political writer because of his failure as a practical politician. For this reason his political thought does not centre on the systematic development of principles of worthy life or norms of a just society. Instead it represents an attempt to describe practical problems, with recourse to historically comparable situations as paradigmatic challenges and to develop general rules for successful political activity from these examples. In *Il Principe* (The Prince) Machiavelli developed these rules principally in the perspective of the political activist, while the *Discorsi* (Discourses on Livy) place more emphasis on the long-term development perspectives of political communities. In this respect two differing perspectives, from theories of action and from theories of structural analysis can be found side by side in Machiavelli's political thought. When Machiavelli addresses potential political activists his considerations are developed for the most part from a perspective of action, whereas when he addresses educated political observers, the emphasis is more on structural aspects. In all his writing he emphasised that his reflections were based on two sources; his own political experiences as well as the 'wisdom of the ancients': the works of the classical writers and historians. The affirmative allusion to classical antiquity identifies Machiavelli as a representative of Humanism; however his judgement of classical knowledge measured against his own experiences goes beyond the conventions of Humanism and classifies Machiavelli as the first modern political thinker in early modern Europe.

1. Biography and Work

Machiavelli was born in the Santa Trinità quarter of Florence on May 3, 1469, as son of a jurist and minor political officer. Little is known of his childhood or the early years of his professional life. The only certainty is that he received a humanist education in accordance with his father's wishes which was to enable him a career in the Florentine administration. In 1498, after the execution of the Dominican monk Girolamo Savonarola, who had tried to reform the republic on the basis of a moralization of the city society orientated by Christian ascetic ideals, he was voted secretary of the second chancery (domestic administration) and later additionally secretary of the council of the Ten (foreign and defence matters). Both posts, which Machiavelli held until his dismissal from office in 1512, allocated him the position of a middle-ranking political officer. However, since the members of the council of the Ten were elected only for short periods his political influence was considerably greater than this position would suggest. He undertook a number of important missions, for example, to the French court, to the Holy Roman emperor Maximilian, to the Roman Curia, and to Cesare Borgia, who had conquered a new duchy in Romagna shortly before. His task on these missions was to explain Florentine politics and also to make enquiries to assess the aims and intentions of the people he visited. Through this role a particularly trusting relationship developed between Machiavelli and Piero Soderini, voted *gonfaloniere a vita* (chief magistrate for life), in 1502, who entrusted Machiavelli repeatedly with the most difficult and confidential tasks.

The most important of these was the project to recapture Pisa, which Florence had lost in 1494. After all attempts to retake the port with the help of mercenary soldiers had failed, Machiavelli was commissioned to establish a Florentine militia, which forced Pisa into a capitulation within 3 years. The ensuing victory celebrations in Florence appeared to represent the climax of Machiavelli's political career. His exceptional position within the political administration of Florence was also expressed in the fact that he tried to develop a more long-term strategy in a series of political memoranda. This strategy was based on a lasting stabilization of the republican system on the inner front, and aimed for the preservation of Florence's political sovereignty on the outer front. Through his notorious emphasis on the exemplary nature of Rome he sought to develop an alternative conception of the republic to Savonarola's Christian moralization program. Here he was not concerned with a political program of action, but with a political ideology from which Florence could gain confidence and courage in difficult and threatening situations.

Due to the energy and decisiveness Machiavelli showed in the tasks allotted to him he advanced rapidly to the *spiritus rector* of the Florentine republic and stood out especially in contrast to the hesitations and procrastinations of other Florentine politicians. He always held a strong aversion to all forms of waiting and delaying, which made him a sharp critic of every policy of neutrality. His singular importance in Florentine politics was also recognized by his contemporaries, seen in the fact that he was the only politician

to be removed from office when the Medici returned to Florence with the support of Spanish troops in November 1512, apart from the chief magistrate Soderini. Shortly afterwards Machiavelli was suspected of taking part in a conspiracy against the Medici. He was thrown into prison and questioned under torture, but there was no proof against him. He was freed in the course of a general amnesty on the occasion of Giovanni de' Medici's election to Pope, on condition that he no longer entered the City of Florence and did not hold any political office.

Machiavelli then began composing political writing on his estate in San Andrea in Percussina. At first this was more as a compensation for his forced political inactivity, but it soon aimed to bring his name into conversation among the then ruling groupings in Florence. He gained the reputation of an experienced and well-versed politician that could be entrusted with practical political tasks once again. The Medici appear to have mistrusted him to the end, however, and he received only a few unimportant missions along with permission to enter Florence again. In the meantime Machiavelli had made himself a name as a writer, with the comedy *Mandragola* (1517) among other writing, and was commissioned to write a history of Florence in 1519, which summarized all the existing annals and chronicles. He completed his *Istorie Fiorentine* in 1525, and it was a masterpiece of political history, scanning from the time of the migration of peoples to the death of Lorenzo de' Medici in 1492. Machiavelli had previously written *Il Principe* and the *Discorsi*, which both remained unpublished until after his death, but which circulated as copies among the political intellectual circles of Florence during his lifetime. In 1520 Machiavelli also wrote the *Dialogo dell' arte della guerra*, a dialogue modeled on a humanistic form on the reform of Italian warfare, as well as the historical novel *Vita di Castruccio Castracani*, in which he describes the conditions of political success using the example of a political adventurer from Lucca. Machiavelli's life took a paradoxical turn; it was only his failure as a political actor that forced him into his existence as a political writer, through which he won success and recognition in his lifetime, and fame and infamy after his death.

Machiavelli himself, however, valued practical politics more highly than political writing, which is the reason why he immediately ran for political office after the Medici were successfully ousted in 1527. For many he appeared by now, however, to be one of the Medici's faithful and he lost the vote for the post he had held until 1512. Machiavelli did not survive this disappointment; he fell ill with acute peritonitis and died a few days later on the June 21, 1527 in Florence. Shortly before his death he is reported to have told friends of a dream he had, in which he saw a crowd of miserable and ragged-looking men, who explained that they were on their way to paradise. Then he saw a group of noble and serious-looking men, among them the great classical philosophers and historians Plato, Plutarch, and Tacitus, who were damned and on their way to hell. Machiavelli is reported to have said to his friends that he would rather discuss politics in hell with the great men of antiquity than die of boredom with the blessed and holy in paradise. 'Machiavelli's dream' is not substantiated and most likely expresses the impression which some of his contemporaries and many of his successors had of him. Yet it does illustrate certain characteristics of Machiavelli's life and thinking—his enjoyment of discussing political problems, his veneration of classical philosophy and historiography, and his contempt for lives given over to Christian contemplation; in fact so well that it has been used and abused in research literature on Machiavelli throughout the years.

2. *Machiavelli as Political Scientist*

What allows us to characterize Machiavelli's political writing and his suggestions for successful politics as scientific, and to understand Machiavelli as the founder of a form of political science not derived from philosophical-theological norms, but based on an independent rationality of political action? Neither the sympathy he held all his life with the republic nor the dry irony of his descriptions, neither the cynicism of some of his advices nor the strict internal worldliness of his ideas are sufficient, but only the specific method with which Machiavelli ordered his material and developed his ideas. It is a process of contrasting two terms or options that are so arranged as to include and sort antithetically all possibilities on one theoretical level. States are drawn up either as republics or autocracies, begins the argumentation of *Il Principe* and the *Discorsi*, and everything that falls between these two variants can be ignored. Among the autocracies, according to the further specification in *Il Principe*, there are hereditary and newly acquired forms. Among the newly acquired forms there are in turn those that have come to their conqueror as a result of his virtue (*virtù*), whilst in the other cases fortunate circumstances (*fortuna*) were of the most importance. And the latter is the very subject of *Il Principe*: the government of newly acquired autocracies gained through fortunate circumstances. The claim to validity of Chaps. 15–19 of *Il Principe*, in which Machiavelli advises the use of pretence and lies, deceit and cruelty, relates to this specific situation, and is in no way to be understood as a general rule of political success, a fact which is often overlooked in the literature on Machiavelli.

The process of antithetical contrast not only represents the material ordering principle in Machiavelli's political thought, but also decides the characteristic style of his thinking and arguments: one must decide on one option or the other, and every attempt to avoid this decision or to leave it open is the beginning of political failure. Thus, Machiavelli turns away from

the characteristic thinking and pattern of argument of Humanism of both–and (*et–et*) and replaces it with an either–or (*vel–vel*), which determines a decision. Although Machiavelli takes at the decision its own value as a decision at various points, *Il Principe* and the *Discorsi* can be read over long stretches as giving advice on making the right decision. To this end he adds comparative methods to the process of antithetical contrast, with which he examines the results and prerequisites of examples from his own time and from classical history. One can therefore describe Machiavelli as the founder of comparative political science. Others before him had argued comparatively, yet always within a given normative frame, whereby the comparison had merely the function of testing out political options as to the realisation of the normative given. Machiavelli on the other hand seeks to work without such moral-philosophical or theological-based guidelines and develops his suggestions for action only out of the comparison, whereby the success of the action or the project provides the scale. What success is, is determined for Machiavelli in view of the prevailing conditions of political action, which are a given factor for the political activists and not optional. He compares, for example, Rome and Sparta, which he recognizes as extremely successful political projects, and which can therefore serve as political models. The more exact comparison shows, however, that Sparta existed longer than Rome, but that the latter had a greater capacity for expansion in the confrontation with outer enemies. Since the Italian states were confronted with massive French and Spanish efforts towards expansion during the early sixteenth century, which they had to tackle on the military front in order to retain their political sovereignty, it was in Machiavelli's view imperative that they take Rome as a model. The recent defeats of Venice, in its fight for the Terraferma were his proof that an oligarchic republic was not a successful model for solving Italy's political problems under the conditions of the time.

The comparison process developed by Machiavelli represents an early variant of the fourfold-point scheme used in modern sociology. Here, Machiavelli uses two contemporary cases, and two from classical antiquity, each chosen according to the model of antithetical contrast and in an attempt to find as paradigmatic a description of the case as possible. Thus, Sparta and Venice are examples of oligarchic republics, Rome and Florence on the other hand represent republics in which the citizens are allowed a broader participation (*governo stretto* vs. *governo largo*). The contrast of the two contemporary and historical examples serves, among other things, to analyze contemporary situations, in which the political process is not complete, against the background of the historical cases which are completed and whose effects and side effects, intended or unintentional, can be fully overviewed. Should the historical cases only differ from the contemporary in the fact of complete and incomplete information, classical antiquity would not be a real comparison, but merely an addition to the present. For Machiavelli, however, classical republics differ from contemporary republics universally through their specific civil-religious foundation. He warned repeatedly of the politically negative effects of a Christianity strongly influenced by contemplation, and saw Christianity as responsible for the fact that there were fewer republics in his time than there were in classical antiquity. Since Christianity, or its prevailing interpretation, extolled the virtues of humble citizens sunk in contemplation rather than those of men of action, it had 'rendered the world weak and a prey to wicked men, who can manage it securely, seeing that the great body of men, in order to go to Paradise, think more of enduring their beatings than in avenging them.' In contrast, he praised the religion of the ancient Romans, which had played an important part in 'commanding the armies, in reuniting the plebs, both in keeping men good, and in making the wicked ashamed.' Machiavelli saw religion as an unavoidable element of people's belief systems, and attempted to explain the differences between classical and renaissance republics in view of this political-cultural dimension. With respect to his own time Machiavelli made several references to the example of Savonarola, who carried out his project of moralizing Florence with the help of sermons and prophecies, with a complementary program of widening political participation to the *petty bourgeoisie*. Machiavelli seems to have developed certain sympathies with this project in hindsight, mentioning Savonarola several times in positive comments. His only criticism is that he was an 'unarmed prophet' who did not have the means to constrain those who did not believe in his sermons and prophecies into obedience at the decisive moment, using armed force. He would have had great sympathy for the appearance of an 'armed prophet' as a political reformer or as a founder of a state, as which he saw Moses for example. Machiavelli blamed in the main the pope and the Roman Curia for the fact that Italy had become a plaything for foreign powers, in the first place through the degeneration of morals spread from papal Rome throughout Italy, and secondly through the Curia's vested interest in a political fragmentation of Italy. Both these ideas were taken up in the Italian *Risorgimento* and made fertile for lay anticlerical politics.

3. *Machiavelli's Republicanism*

However, since Machiavelli rather doubted the possibilities of fundamental political renewal through religious reforms, he placed his hopes in political regeneration of Florence and Italy through the reform of the military, which explains why the subject of war and military organization is present throughout the whole of his writing. At first he was naturally concerned with the reform of Italian military organisation

itself, in the course of which he was in favour of replacing the mercenaries with a militia recruited within the area of rule. He also intended to use the increasing importance of drawn-up infantry on the European battle fields to refer back to the Roman model, in as far as the Romans also owed their military successes more to the infantry than to the cavalry. Aside from this he was interested in the political effect of military reform within the state; the duty of citizens to defend their city and their willingness to involve themselves existentially in political projects. He reserved his harshest criticism for mercenary troops, whom he described as cowards and traitors. Machiavelli's ideas on military reforms played a particular role in Europe after the French Revolution and in postcolonial Africa in the 1960s and 1970s.

Machiavelli's preference for Rome over Sparta, which was based on Rome's greater capacity for military expansion, also meant his taking sides in Florence's inner conflict around its type of constitution and the level of political participation. In *principal* there were two opposing parties, one of which was in favor of limiting access to political power to the rich and noble families, in other words the town patricians, while the other favored a stronger involvement of the middle classes, or the urban *petty bourgeoisie*. Since the uprising of the Florentine wool cloth workers in the late fourteenth century the conflicts in Florence revolved increasingly around this problem, which overshadowed the previous political lines of conflict between single families and their clientele groups. As sharply as Machiavelli rejected the power struggle of the clientele groups and factions, he defended the struggles between the social levels and classes for power and influence as a fountain of youth, which represented an irreplaceable form of revitalization of republican energies. He wrote that those who condemned the conflicts between nobility and the common people and attempted to prevent them through legislation not only limited the capacity for expansion of the republic, but robbed it of its most important source of life. We can see in this an early form of party theory, or rather a theory of the integration of political communities through conflict. Machiavelli saw, however, less value in the possibility of defending certain particular interests than in blocking degeneration and corruption by institutionalizing the inner conflict. Thus, he admits to the middle-classes that they have a greater interest in retaining liberty than the town patricians. However, the liberty to which he refers at this point is not the (liberal) freedom to go about their own interests undisturbed, providing these interests are not in conflict with those of the republic, but the (republican) freedom of political participation and civic engagement. With Polybios and Cicero as his starting point, Machiavelli developed a conception of a hybrid constitution, which was not (as in Montesquieu) oriented at a varying control of power to ensure individual freedom, but at the institutionalization of political conflicts to prevent moral degeneration and party spirit. Machiavelli forms the beginning of the constitutional and party theory of modern republicanism. This aspect of his work, to which Spinoza and Rousseaualready referred, has only been rediscovered and examined in the most recent Machiavelli research, particularly through the work of Gilbert (1965) and Pocock (1975).

4. *Reception*

For a long time the Machiavelli reception was dominated by a concentration on Chaps. 15–19 of *Il Principe*, in which Machiavelli discussed breaking one's word and lying, pretence and treason, violence and cruelty as means of ensuring political rule as to their efficiency, and recommended their purposeful use in rejection of the theological moral-philosophical political theory of the Scholastics and of Humanism. The Machiavelli critique of the Antireformation (Cardinal Pole), the Monarchomachs (Gentillet), and in particular of the Enlightenment (Friedrich II of Prussia) concentrated above all on these passages and thereby brought about the dominant perception of Machiavelli, which is of a theorist of political scrupulousness. The popular image of Machiavelli as 'teacher of evil' and 'spoiler of politics' has been decidedly influenced by this. Aside from this, Machiavelli's writing was received in Italy and Germany, where the nation–states were formed at a later date than in other western countries, as a theory of state-building; Fichte's and Hegel's image of Machiavelli in particular show clear signs of this. Under the influence of Fascism and National Socialism there was further intensive debate on Machiavelli in these countries. He was exemplified and celebrated as a forerunner of fascist politics by several writers while others, for example, König (1941) criticized him as a utopian and romantic, who understood little of real politics and, therefore, chased illusions and so represented the prototype of a fascist intellectual. Thus, the discussion of Machiavelli has always served to survey the various contemporary political problems. Like no other theorist in the history of political thought, he provokes his interpreters to deal with the questions of the present in the light of his own ideas, and to project his ideas onto those of the present. This has often led to a careless handling of the material with regard to its interpretation and analysis, but is on the other hand a precondition that a work should provoke such lively and extreme controversies in research literature as that of Machiavelli.

See also: Elites: Sociological Aspects; Leadership: Political; Political Science: Overview

Bibliography

Bock G, Skinner Q, Viroli M (eds.) 1990 *Machiavelli and Republicanism*. Cambridge University Press, UK

Bonadella P E 1973 *Machiavelli and the Art of Renaissance History*. Wayne State University Press, Detroit, MI
Buck A 1985 *Machiavelli*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany
Donaldson P S 1988 *Machiavelli and Mystery of State*. Cambridge University Press, New York
Faul E 1961 *Der moderne Machiavellismus*. Kiepenheuer and Witsch, Cologne and Berlin
Gil C 1993 *Machiavel. Fonctionnaire fiorentin*. Perrin, Paris
Gilbert F 1965 *Machiavelli and Guicciardini. Politics and History in the Sixteenth-century Florence*. Princeton University Press, Princeton, NJ
Hale J R 1961 *Machiavelli and Renaissance Italy*. Macmillan, New York
König R 1941 *Niccolò Macchiavelli. Zur Krisenanalyse einer Zeitwende*. Rentsch, Erlenbach and Zürich, Switzerland
Machiavelli N 1927 *Istorie Fiorentine. Testo critico con introduzione e note per cura di P. Carli*, 2 vols. Sansoni, Florence, Italy
Machiavelli N 1960–1965 *Opere complete, a cura di S. Bertelli e di F. Gaeta*, 8 vols. Feltrinelli, Milan
Machiavelli N 1963 *Opere, a cura di M. Bonfantini*. Salerno, Milan and Naples, Italy
Machiavelli N 1964 *Legazioni e commissarie, a cura di S. Bertelli*, 3 vols. Feltrinelli, Milan
Machiavelli N 1968–1982 *Edizione dell'Opere omnia, a cura di S. Bertelli*, 11 vols. Salerno, Milan and Verona, Italy
Machiavelli N 1979 *Il Principe. Introducione e note di F. Chabod, a cora di L. Firpo*. Einaudi, Turin, Italy
Machiavelli N 1983 *Discorsi sopra la prima deca di Titio Livio. Introduzione di C. Vivanti*. Einaudi, Turin, Italy
Meinecke F 1957 *Machiavellism. The Doctrine of raison d'état and its Place in Modern History*. Yale University Press, Routledge, London
Münkler H 1982 *Machiavelli. Die Begründung des politischen Denkens der Neuzeit aus der Krise der Republik Florenz*. Europäische Verlogsanstalt, Frankfurt
Pocock J G A 1975 *The Machiavellian Moment. Florentine Political Thought and the Atlantic Republican Tradition*. Princeton University Press, Princeton, NJ
Ridolfi R 1954 *Vita di Niccolò Machiavelli*, 2nd edn. Belardetti, Rome
Rubinstein N 1956 The beginnings of Niccolò Machiavelli's career in the Florentine chancery *Italian Studie*, Vol. 11
Sasso G 1958 *Niccolò Machiavelli: Storia del suo pensiero politico*. Nella side deli 'Istituto, Naples
Sasso G 1986 *Machiavelli e gli Antichi e altri saggi*, 3 vols Riccardo. Ricciardi, Milan and Naples
Skinner Q 1978 *The Foundations of Modern Political Thought*, Vol. 1. Cambridge University Press, Cambridge, UK
Skinner Q 1981 *Machiavelli*. Oxford University Press, Oxford, UK
Strauss L 1958 *Thoughts on Machiavelli*. Free Press, Glencoe, IL

H. Münkler

Macroeconomic Data

Macroeconomic data are aggregate data relating either to sectors of the economy, i.e., specified groups of economic units such as households or financial institutions, or to the economy as a whole, i.e., the entire set of economic units resident in a country. They measure the aggregate values of economic flows or stocks associated with economic activities such as production, consumption, investment, and financing. At the level of the economy as a whole, obvious examples are national income and product, or total exports and imports. At a sector level, household income and consumption, and government income and consumption provide examples. Macroeconomic data also include the aggregate values of stocks of assets such as machinery, structures, and land.

Macroeconomic data are used by businesses, the government, financial markets, the press, and the general public to monitor, and assess, changes taking place within economy. They are also used for purposes of economic analysis and econometric modeling and forecasting. They provide the information needed for purposes macroeconomic policy-making and decision taking by governments and central banks. The collection, processing, and publication of macroeconomic data are usually carried out by national statistical offices and other government agencies. In many countries, the central bank is responsible for the collection and publication of a wide range of macroeconomic data, especially financial statistics and balance of payments statistics, and sometimes also the national accounts.

Many flows such as income and consumption can be defined in different ways so that their values depend on the particular definitions and conventions adopted by those responsible for their collection and publication. As the various economic activities taking place within an economy are interdependent, the associated macroeconomic data are also interdependent. It is therefore desirable that individual macroeconomic variables should be defined and measured consistently with each other within an overall statistical framework that uses the same economic concepts, definitions, and classifications throughout.

Macroeconomic data are inevitably subject to error. Complete information is seldom available on all the units concerned. Most macroeconomic data are not based on censuses or surveys specifically designed for the purpose. Many are essentially estimates derived from data collected for other reasons, such as administrative or tax data. One of the arts of the statisticians responsible is trying to reconcile inconsistencies between data drawn from different sources. The appearance of data from a new source may lead to considerable revisions in previously published statistics.

As one of the main uses of macroeconomic data is to provide up-to-date information for economic forecasting and economic policy-making and decision taking by governments and others, there is considerable pressure from users to obtain the data as quickly as possible. There is a trade-off, however, between timeliness and reliability when the underlying basic data only become available gradually over a period of time. First estimates based on incomplete data can

only be provisional. The earlier they are made and released, the greater the risk that they may have to be substantially revised later when more data become available. Revisions may be acceptable to users as the price to be paid for timely data, but occasionally they may be a major source of embarrassment to the government agencies responsible.

1. The Main Macroeconomic Aggregates

1.1 GDP and Gross Value Added

The objective is to measure the output produced by all the economic units resident in a country. Measuring the output of a single enterprise is straightforward, but aggregating the outputs of different enterprises raises a further complication. From the point of view of the economy as a whole, not all outputs are final because some of the outputs produced by some enterprises may be sold to other enterprises to be used up again as inputs into further production. These products are 'intermediate,' not final. They are not available for consumption by households or governments, for augmenting the capital stock, or for export. In order to measure the contribution of each enterprise to the economy's final output it is necessary to deduct from the value of its own output the value of the intermediate products it consumes. The difference is described as 'gross value added.'

Value added can be measured either before or after deducting the depreciation on the capital goods used in production. If before, it is described as 'gross,' if after, it is 'net.' Gross Domestic Product, or GDP, is defined simply as the sum of the gross values added of all the enterprises resident in an economy. GDP is probably the most important single macroeconomic aggregate, being widely used throughout the world as a key indicator of the level of economic activity. It is measured quarterly as well as annually, its movements being factored into their price and quantity components. The price movements measure inflation while the quantity movements are used to measure rate of real economic growth and cyclical fluctuations in the economy.

1.2 From GDP to National Income

The link between aggregate measures of production and income is established by the fact that the incomes generated by production are all paid out of value added. As operating profits are the residual amount remaining after all the other incomes have been paid, the sum of all the incomes generated by production, including operating profits, must be identical with value added. At the level of the economy, the sum of the incomes generated by production is described as Gross Domestic Income, or GDI, which is equal in value with Gross Domestic Product by definition.

Gross National Income, or GNI, is derived from GDI by adding the incomes received by residents from foreigners and deducting the incomes paid to foreigners. (Unfortunately, GNI has traditionally been described as Gross National Product, or GNP, although it is an income rather than a production concept.) Net National Income is obtained by deducting depreciation from GNI. It is often described simply as National Income.

The numerical identity between GDP and GDI means that they can be estimated from two quite different sets of data, namely income or production data. Because income data are often incomplete and unreliable, most countries place more reliance on GDP measured from the production side using data on inputs and outputs collected in industrial inquiries. There is also a third approach that uses expenditures on final outputs, namely household and government consumption, capital formation, and exports less imports. Estimates from the production, income, and expenditure sides provide checks on each other. In practice, there are usually significant statistical discrepancies between them that may be reduced by adjusting the data, but many countries find it impossible to eliminate statistical discrepancies altogether and are obliged to publish them.

2. From National Income to Economic Accounting

Defining and measuring a few aggregates at the level of the economy as a whole is only one small part of macroeconomic measurement. Macroeconomic analysis and policy making requires sets of data covering many economic variables and activities measured at various levels of aggregation.

The development of comprehensive macroeconomic data systems dates back to the 1930s and 1940s. It was stimulated by two factors, the economic depression and the Second World War. The depression and the switch from peacetime to wartime production both led to more active economic policy making and intervention than previously, but such policies also require more information. The relevant macroeconomic theory and models were also being developed at around the same time.

The macroeconomic data systems came to be known first as national accounts and later as economic accounts. The publication of official estimates of national income aggregates for the USA started in the 1930s, but it was not until the Second World War that proper accounts began to be published. In the UK, the first set of official national accounts was compiled in 1941 by Meade and Stone at the request of Keynes. Towards the end of the war, very close links were established between the experts in the UK and the USA, which led to the development of the international systems described below. The development of

these systems was a collective achievement by a number of economists from Europe and North America, but pride of place is generally given to Richard Stone who was awarded the 1984 Nobel Prize in economics for his work in this field. Stone took the leading role in elaborating the international systems published under the auspices of the Organization for European Economic Cooperation, or OEEC (later the OECD), and the United Nations.

Stone focused on transactions between economic units as the elementary items from which macroeconomic data systems have to be constructed. Transactions can take very many forms. For example, the ownership of some goods or assets may be exchanged between A and B, or some service be provided by A to B, accompanied by a counterpart flow, usually of a financial nature, such as the payment of money by B to A. The objective of economic accounting is to organize, classify, and aggregate the countless transactions taking place within an economy in a way that is informative and useful for purposes of economic analysis and policy making. The challenge that this presents is very different from deciding how to define National Income. Stone observed that transactions could be classified according to the type of economic units involved, the type of economic activity with which the transaction is associated, and the nature of the items exchanged. By cross classifying with respect to these kinds of criteria, quite elaborate data systems can be constructed.

In 1953, the United Nations published its '*System of National Accounts and Supporting Tables*,' or *SNA*, intended for use by all countries in the world except the Soviet Union and other Socialist countries who preferred their own Marxian version referred to later. At the same time, the International Monetary Fund, or IMF, was developing closely related international standards covering Balance of Payments Statistics, Financial Statistics, and Government Statistics. Two major revisions and extensions of the UN *SNA* were published in 1968 and 1993. The revision completed in 1993 was conducted jointly by all the major international economic agencies—UN, IMF, World Bank, the Organization for Economic Cooperation and Development (OECD), and the European Union (EU). The 1993 version of the SNA is the system now used by almost all countries in the world as the norm or standard for their own accounts, including former Socialist countries who abandoned their Marxian system at the time the 1993 *SNA* was being finalized.

3. *The International System of Macroeconomic Accounts, or SNA*

The first paragraph of the 1993 edition of the *SNA* reads as follows: 'The System of National Accounts (*SNA*) consists of coherent, consistent and integrated set of macroeconomic accounts, balance sheets and tables based on a set of internationally agreed concepts, definitions, classifications, and accounting rules. It provides a comprehensive accounting framework within which economic data can be compiled and presented for purposes of economic analysis, decision taking, and policy-making. The accounts themselves present in a condensed way a great mass of detailed information, organized according to economic principles and perceptions, about the working of an economy. They provide a comprehensive and detailed record of the complex economic activities taking place within an economy and of the interactions between different economic agents and groups of agents, that take place on markets and elsewhere.'

3.1 *The Structure of the SNA*

The accounts may be compiled at any level of aggregation. The *SNA* divides the economy into a small number of sectors, such as the household, government, and financial sectors, which may in turn be divided into subsectors. While it is appropriate to describe the accounts for the total economy as 'national accounts' this description is not appropriate for the accounts of individual sectors. For this reason, the general expression 'economic accounts' is preferred.

The economic units resident in the rest of the world can also be viewed as a sector. However, an individual country cannot compile sets of accounts for the rest of the world. Instead, it merely records transactions between resident and nonresident units in a special account, a 'rest of the world account.' This account is essentially the same as a Balance of Payments Account, which can thus be seen to be an integral part of the overall system of economic accounts.

3.2 *The Sequence of Accounts*

Each account in the *SNA* relates to a particular kind of economic activity, such as production or consumption. The entries in the accounts consist of the aggregate values of particular kinds of transactions conducted by the units engaged in those activities. For example, a production account records the purchases of inputs and sales of outputs which producers make as a result of engaging in production. The accounts do not record the activities in a physical or technical sense.

Each account has two columns headed 'uses' and 'resources.' As a general guideline, transactions are recorded under 'resources' if they increase the unit's financial resources, for example sales of outputs, and under 'uses' if they reduce them, for example purchases of inputs. The counterpart changes in money or other financial assets are recorded in the financial account. The production, consumption, and capital accounts record real economic activities in the economy while the financial account deals with the financing of those

activities and also transactions of a purely financial nature.

The purpose of compiling economic accounts is not to have tidy book keeping. By comparing the total values of the 'uses' and 'resources' in an account, the outcome of the associated activity can be evaluated from an economic point of view. The differences between the two totals are described as 'balancing items.' Value added, profits, disposable income, and saving are examples of balancing items that can only be measured within an accounting framework. These balancing items are typically the most interesting and important items within the accounts as they encapsulate a great deal of information about the activities in question. Indeed, the most important single reason for compiling economic accounts is to derive the balancing items. GDP itself is a balancing item equal to the difference between the values of total output and total intermediate inputs in a highly simplified production account for the total economy. GDP is an accounting construct. It is not a flow that can be observed directly.

The accounts following the production account of the *SNA* portray the distribution of the incomes generated by production and their subsequent redistribution by government through taxation and social insurance. They are followed by an account showing expenditures on consumption out of disposable income whose balancing item is saving. The remaining accounts of the *SNA* consist of the capital account and a financial account. Detailed financial accounts that track the flows of funds between different sectors and subsectors are an integral part of the *SNA*, specimen flow of funds accounts being given in the 1993 *SNA*. Finally, the *SNA* also includes balance sheets in which the values of the assets owned by the various sectors are recorded together with their liabilities.

A brief sketch of the *SNA* has been given in the above paragraphs. It is a complex sophisticated system. The *SNA* provides a conceptual framework within which macroeconomic data relating to a very wide range of economic activities can be recorded. It imposes a single set of economic concepts, definitions, classifications, and accounting rules throughout so that all the diverse activities are recorded in a mutually consistent way. Both national and international statistical agencies have long recognized the importance of imposing this kind of discipline so that users have macroeconomic data that are mutually consistent conceptually. It also helps to ensure that they are consistent numerically.

4. Socialist Economic Accounting

National Income has a long history. Income has not always been clearly distinguished from wealth. Income is a flow that can only be measured over a period of time, such as a week, month, or year, whereas wealth refers to the value of a stock of assets that can be measured at a point of time. Attempts to estimate the Nation's Income go back over three centuries to the work of William Petty (1662). However, he was more concerned with measuring the wealth of the nation and taxable capacity than income. The same preoccupation with wealth rather than income is found a century later in Adam Smith's *The Wealth of Nations* (1950).

The extent to which macroeconomic data can be influenced by ideas, concepts, and definitions is vividly illustrated by the distinction that Smith introduced between 'productive' and 'unproductive' labor. Smith argued that only workers who produced goods, as distinct from services, should be regarded as productive because only goods could add to the stock of the nation's productive capital equipment. This distinction was followed by many classical economists in the early nineteenth century, including Karl Marx in *Das Kapital* (1867). One far reaching consequence was that it was built into the Marxian based system of national accounting adopted by the former Soviet Union and subsequently by other centrally planned socialist countries. In the international version of that system developed after the second world war under the auspices of the United Nations it is stated that: 'All fields of productive activity are based on material production, which is primary in comparison with the activities rendering services. The global product and national income are produced in the material sphere … The wealthier a community is, the more material goods it produces … *The non-material sphere creates neither product nor income*' (*Basic Principles of the System of Balances of the National Economy*, United Nations, Statistical Office 1971). This makes strange reading when about two thirds of the GDP of most developed countries consists of the production of services. This system of accounting was not abandoned by the countries concerned until the early 1990s.

5. Household Production and GDP

The treatment of household production in GDP is a contentious, almost a political, issue. As GDP is a measure of the aggregate value added created by production, it might be presumed to be a reasonably objective statistic. However, there are two kinds of production: production whose output is destined for the market and nonmarket production whose output is not. The output from nonmarket production is consumed by the same units that produce it. GDP consists mainly, but not entirely, of the value added created by market production. Its absolute size depends on how much nonmarket production is included, this being very much a matter of convention.

In order for any activity to count as productive in an economic sense, the output to be marketable so that the production can be organized on a market basis. Many of the activities undertaken within households

satisfy this basic criterion but the outputs are not actually offered on the market being retained for consumption within the household. For example, households may grow their own food or other agricultural products. In poor countries, a large part of the total production and consumption not only of food but other basic products may consist of nonmarket production for their own consumption.

GDP covers all market production (even if it is illegal) and the SNA keeps track of all the associated income and consumption flows, which are usually monetary. By convention, it has also been agreed that GDP should include the production of any goods produced for own consumption. These include the produce of gardens and other small plots cultivated by households in developed countries and not only the output of subsistence agriculture.

In most countries, however, most of the production for own consumption within households consists of services such as cooking, cleaning, and childcare. By convention, this service production is excluded from GDP (even though the production of housing services for own consumption by owner-occupiers is included). As most of these household services may be produced by women, this exclusion has been criticized on the grounds that the real contribution of women to the economy's production is thereby grossly underestimated. Using the broad criterion of economic production given above, this criticism is valid, as all the household services concerned can be produced on a market basis by paid servants, nurses, teachers, chauffeurs, hairdressers, etc. Unofficial estimates suggest that the unrecorded production of household service for own consumption could equal a half to two thirds of GDP as conventionally defined. If the definition of GDP were changed to include this production, GDP would rise dramatically.

On the other hand, there are some users and economists who would prefer GDP to move the other way by omitting all production for own consumption completely and confining GDP strictly to output from market activities. One reason is the practical difficulty of estimating the quantity of such output and valuing it in an economically meaningful way. The value of the production of household services is speculative and can vary greatly according to the assumptions made about how to value it. A more fundamental reason is that market and nonmarket productions do not have the same economic significance. When goods or services are produced for the market, there may be disequilibria in the form of excess supply or demand precisely because the suppliers and consumers are different economic units. National accounts were developed to provide aggregate macroeconomic data relating to market activities for purposes of fiscal and monetary policy that are mainly concerned with phenomena such as unemployment or inflation that are symptoms of market disequilibria. Some users would therefore prefer accounts based only on market transactions. To move the other way by including large imputed non-monetary flows attributable to nonmarket production within households would greatly reduce the analytical usefulness and policy relevance of the accounts for most users.

The present concept of GDP is a somewhat uneasy compromise resulting from the desire to have a single multipurpose measure of GDP satisfying the demands of different users. Alternatively, it would be possible to define a hierarchy of different GDP's ranging from a strictly market based concept to one that included every kind of own account production, but the existence of alternative GDP numbers might simply create confusion.

6. *Some Specialized Macroeconomic Data Systems*

There are several specialized macroeconomic data systems that focus on particular sectors or types of economic activity: for example, balance of payments statistics, flow of funds statistics, input-output tables, labor force statistics, and capital stock statistics. Although these systems may have originally developed independently of each other and the *SNA*, a common set of concepts and classifications is needed when data from different systems have to be used together.

6.1 *Labor Force Statistics*

As labor force statistics concern people, it might be presumed that the links with the *SNA* would be tenuous. However, the concept of the labor force, as specified in the international standards published by International Labor Organization, relies directly on the production boundary used in the *SNA*. The main conceptual problem is to determine which members of a household are economically active. This is essentially the same issue as that considered in the previous section, as only those members of a household who are engaged in activity that falls within the production boundary of the *SNA* are treated as being members of the labor force. The size of the labor force, as well as GDP, depends on the conventions adopted with regard to the production boundary. Persons engaged in the production of services for own consumption within the household do not form part of the labor force, whereas those engaged in the own-account production of goods are included. The latter are treated self-employed while the former are not. The concept of employment, which is linked to that of the labor force, clearly refers to persons who receive some kind of remuneration for working in the market sector that generates most of GDP.

6.2 *Balance of Payments Statistics*

Balance of payments statistics record all flows of goods and services, income and capital between an

economy and the rest of the world. International standards for balance of payments statistics are set by the International Monetary Fund, or IMF, in its *Balance of Payments Manual*, the fifth edition of which was produced in 1995. It is intended to be fully consistent with 1993 *SNA*.

The balance of payments is single comprehensive account that records all the transactions between the economic units resident in a country and units resident in the rest of the world. It is essentially the same as the rest of the world account in *SNA*. The balance of payments is divided into four subaccounts. First, a distinction is drawn between current transactions and those relating to assets. This matches the distinction between the current and the accumulation accounts in the *SNA*. The current account of the balance of payments is then subdivided to distinguish trade flows from income flows. The accumulation account is subdivided to distinguish transactions in nonfinancial assets from those in financial assets.

Keeping track of transactions between resident and nonresident units poses special problems. Many of the concepts and definitions in the *SNA* need further precision and refinement in a balance of payments context. Even the concept of residence is becoming increasingly blurred because of increased mobility due to fast air transport and instant international communications through the Internet. The concept of an export or import is also not straightforward when goods are transported through third countries en route to their final destination or are shipped abroad temporarily for processing.

Trade flows may be expected to be among the more reliable items recorded in balance of payments statistics. It is possible to have a check on them because exports from A to B as recorded by A should equal imports by B from A as recorded by B provided they are consistently valued. There are many examples, however, to show that there may be serious inconsistencies between the two sets of data, even for neighboring countries such as the USA and Canada. These difficulties are compounded if there is a lot of smuggling. The difficulties of keeping track of financial transactions can be considerably greater with the emergence of world financial markets on which financial assets are traded continuously and instantaneously as a result of modern communications technology.

As in the *SNA*, the entries in the balance of payments that are usually of the greatest policy interest are balancing items obtained residually, such as the balance on the current account or net lending. In general, however, balancing items are also the most sensitive to error. In principle, the current account balances of all the countries in the world should cancel each other out. One country's deficit is another's surplus. However, when the balance of payments statistics of all the counties in the world are confronted with each other their balances do not cancel and a world current account discrepancy emerges. The IMF has conducted investigations into the factors responsible in order to try to reduce the size of the discrepancy.

Although the concepts and classifications in balance of payments are essentially the same as in the *SNA*, a special manual is needed in the light of the special problems outlined above. It is necessary not only to provide greater precision in certain areas, but also to provide practical operational advice about how best to improve the quality and reliability of the statistics. The IMF has provided technical assistance and training to balance of payments compilers throughout the world over many years.

6.3 Capital Stock and Wealth Data

The aggregate values of a sector's assets and liabilities at the start and end of the accounting period are recorded in the *SNA*'s balance sheets. Total assets less liabilities equal 'net worth,' which is equivalent to 'wealth' in ordinary parlance.

Even though complete information about assets and liabilities is rare at a macroeconomic level, data are often available on holdings of financial assets and on stocks of fixed assets in the form of machinery, equipment, buildings, and other structures. Stocks of fixed assets are often described simply as the 'capital stock.' Estimates of their value are needed for purposes of productivity analysis.

The value of a fixed asset at a point of time is equal to the present, or capitalized, value of the flow of capital services into production that it is capable of providing over the rest of its service life. Capital stock and the flow measures are therefore completely interdependent. Depreciation is measured by the decline, between the start and the end of the accounting period, in the present value of a fixed asset used in production.

Capital stock estimates are usually made using the perpetual inventory method, or PIM, which derives the stock estimate from past data on investment flows, or gross fixed capital formation in *SNA* terminology. If the average life of an asset is estimated to be n years, the investments in the asset over the last n years, revalued at current prices, are cumulated to provide an estimate of the 'gross capital stock.' The cumulative depreciation on each vintage of asset is then deducted to obtain an estimate of the 'net capital stock.' Capital stock estimates derived in this way are available for many countries.

In empirical work on productivity, it is important to note that the input into production is not the capital stock itself but the flow of capital services provided by that stock. Their value is given by depreciation plus the return on the asset. It is possible to estimate volume indexes for flows of capital services for purposes of productivity analysis.

See also: Databases, Core: Demography and Registers; Economic Growth: Measurement; Economic

History, Quantitative: United States; Economic Panel Data; Household Production

Bibliography

Inter-Secretariat Working Group on National Accounts 1993 *System of National Accounts 1993*. Commission of the European Communities, Brussels/Luxembourg

International Monetary Fund 1993 *Balance of Payments Manual*, 5th edn. International Monetary Fund, Washington, DC

Petty W 1662 *A Treatise of Taxes and Contributions*. O. Blagrave, London

Smith A 1950 *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strathan, London

United Nations Statistical Office 1971 *Basic Principles of the System of Balances of the National Economy*. United Nations, New York

T. P. Hill

Macrosociology–Microsociology

Macrosociology generally refers to the study of a host of social phenomena covering wide areas over long periods of time. In contrast, microsociology would rather focus on specific phenomena, involving only limited groups of individuals such as family interactions or face-to-face relations. The theories and concepts of macrosociology operate on a systemic level and use aggregated data, whereas those of microsociology are confined to the individual level. This definition reflects a limited aspect of research practice and epistemoligical discussions of sociologists. Hence, it needs to be completed.

Sociologists thus identify two scales of organization—the microscopic and the macroscopic—which they understand to correspond to two types of phenomena and explanation. Recognizing the existence of levels of reality, however, should not lead us to think that there are ontologically distinct entities that correspond to different types of objects. For the time being, no specific dominant point of view has managed to make the basic structures of these two types entirely homogeneous or unifiable (see Wippler and Lindenberg 1987). Our purpose here is to clarify the debate over the links between macroscopic and microscopic, over whether and how it is possible to move from one to the other, and over the idea that the first is founded on the second.

1. Links Between Micro and Macro and the Problem of Explanation

Although the problems raised by the relations between macro- and microsociology are highly relevant today (to judge from the number of studies done on them in recent years) (see for example Collins 1981, Alexander et al. 1987, Huber 1990), few of them are new. They first arose with the institutionalization of sociology as an autonomous scientific discipline positioned at the center of the debates opposing utilitarianism-inspired theorists like Spencer against sociologists who, in search of a foundation for the specificity of the object and the explanations of social phenomena, rejected the hypotheses of classical economy.

Far from being purely rhetorical, these problems determine researchers' choice of instruments, units of observation, and explanatory procedures. A hypothesis tested at the level of the individual may very well prove false at the systemic level; another, verified at the global or context level may prove invalid for the individual case (see Robinson 1950, Goodman 1953, Duncan and Davis 1953, Goodman 1959). The error of inferring the social from the individual corresponds to that of inferring the individual from the social ('contextual fallacy'): though it is sometimes a mistake to deduce (without any supplementary hypothesis) the voting behavior of individuals from information on how ecological units voted, it is just as wrongheaded to affirm, on the basis of a relation of positive dependence between an individual's degree of education and his or her salary, that when the collective level of education rises, the average salary does too.

Let us suppose, as did the founding fathers, that the aim of social sciences consists of explaining the statics and dynamics of social systems (though this definition is neither necessary nor sufficient). Tocqueville (1856) tried to account for political change, mainly the revolutionary one; Durkheim (1893) for the increasing differentiation of functions and the consequences of such differentiation; Weber (1905) for the birth of a new type of socioeconomic organization; Sombart (1906) for the absence of socialism in the USA. Clearly, all these questions pertain to the systemic level. The size of the system may vary, but whatever its components and their number, it constitutes a unit that must not be broken down or dissolved.

We have three competing procedures for conducting an explanation of such phenomena. The first and the most current one is also the least satisfactory. It consists of analyzing the relations between a given social phenomenon and independent social factors; explaining the social by means of the social is the *leitmotiv* of the Durkheimian sociological program. But in order for that first macrological procedure to be legitimate, one's observations must pertain to the same level. We readily affirm that revolution is the result of improvement in economic, political, and social conditions; that an increase in the volume and density of groups or societies leads to the division of labor; that religious ethics facilitates the birth of capitalism; that a high degree of social mobility makes the development of strong left-wing parties less likely.

It seems, however, that this condition of consistency between the content of a sociological proposition and

the unit of observation and analysis is not often met. In most studies conducted by interview or questionnaire, the unit remains the individual, and a hiatus is created between the objective of the theory, which is to formulate macrosociological propositions, and empirical research, which remains focused on explaining individual behavior. This holds true despite the fact that Lazarsfeld and Menzel (1961) drew sociologists' attention to the necessity of consistency. Without consistency, we are caught in an impasse: many hypotheses deduced from macrosociological theories cannot be adequately tested.

The second procedure consists of collecting observations at an infrasystemic level and developing hypotheses on the behavior of such units (individuals, groups, institutions), with the purpose of explaining systemic relations through an appropriate synthesis of these observations. In this case micrological variables positioned in an intermediary position between dependent and independent macroscopic factors are introduced. One affirms that improvement in conditions (macro) generates a level of frustration (micro) such that it increases the probability of protest at the sociopolitical level (macro); that an increase in volume with growing densification aggravates competition between system units confronting each other on the same market, and that this competition is resolved by division of labor; and that social mobility, whether real or perceived as real by individuals, reduces the search for a collective solution by making people believe in the utility of personal effort, thus making the emergence of strong socialist parties less likely.

An explanation operating at this infrasystemic level is sometimes considered more satisfying, stable, and general than one expressed entirely in macroscopic terms. This is Popper's point of view (1934, 1963), as well as Harsany's (1968), Simon's (1977), and Boudon's (1973, 1977). The argument holds that because the behavior and dynamics of a social system are in fact the result of the actions of its components (individuals, subgroups, institutions), it is these components that should be analyzed if we are to better explain and predict systemic characteristics.

To illustrate this case, let us borrow Tocqueville's (1856) explanation of the stagnation of French agriculture at the end of the Old Regime compared to the English one as Boudon (1998) stylized it according to his cognitivist model. Agriculture stagnation is the effect of administrative centralization in France. Both phenomena are macroscopic but their relation has to be explained by bringing to the fore the reasons of classes of persons. More prestigious civil servants' positions in France than in England cause landlords' absenteeism that in turn provokes a low rate of innovation. By buying positions, French landlords think they can increase their power, prestige, and income. They have good reasons to leave their land, rent it, and serve their King. On the other hand, French farmers cultivate the land in a traditional fashion because they do not have the capacity to innovate. In England, the civil servants' positions are less numerous and less rewarding while local political life offers them good social rewards. What Boudon constructs is a social mechanism to explain the relation between the two macroscopic phenomena, i.e., a set of ultimate causes that have the character of being individual decisions we perceive as understandable.

The third approach of conducting an explanation is to analyze the effects due to the nature of the positions and distributions of certain variables on the behavior of the system's component units *without* formulating hypotheses about individuals (see Blau 1977, Blau and Schwartz 1984). Here we are no longer interested, for example, in the influence of religion (or any other variable) on social relations, but rather in the distribution of religious affiliation and the structural parameters of that distribution, understood to determine the dependent variables. It is important to carefully distinguish this macrostructural research, which works not with variables but positions in the social structure, from studies of the first type like Moore's (1966), Lenski's (1966), Skocpol's (1979), and Turner's (1984), in which collective properties and aggregate variables are subjected to classic treatment in terms of factors, relations, correlations, and influences.

Whereas microsociology seeks to explain social relations in terms of psychological traits, or processes such as exchange, communication, competition, and cooperation that shape relations between individuals, structural macrosociology accounts for the relations between the different components of a given society in terms of differentiation and integration. For this last approach, explaining certain aspects of relations between religious or ethnic groups—intermarriage or friendship, for example—means mechanically deducing them from structural properties of the groups—namely, their respective sizes—without having recourse to supplementary hypotheses about individuals, such as how liberal they are.

By way of illustration, let us consider the following example. Suppose we have to study the relation between religion and intermarriage between religious groups, and that we have established a correlation between the two variables, namely that the percentage of Protestants marrying non-Protestants is greater than the percentage of Catholics marrying non-Catholics. How can this be explained? We could develop hypotheses concerning the behavior of Catholics and Protestants: Protestants are less 'practicing,' less sexually inhibited, more tolerant than Catholics, for example. We would then look for indicators of religious practice, tolerance, and so forth, to account for the observed correlation.

Macrosociology proceeds differently. First, it makes no hypotheses about individual behavior. Rather it analyzes the distribution of individuals among the various positions and the effects of those positions on

behavior. Take two groups of different sizes A and B, with A > B. Individuals in group A can come into contact with individuals in group B and their members can intermarry. The first theorem deduced from these hypotheses is that the rate of intergroup association for group B must be greater than that for group A. If A and B are composed respectively of 10,000 and 100 individuals, and if there are 10 mixed marriages (10 pairs of spouses), the respective intermarriage rates are A: 10/10,000 = 1/1,000 and B: 10/100 = 1/10. If A represents the Catholic group and B the Protestant group, the preceding theorem enables us to affirm that B's intermarriage rate is higher than A's—the correlation has been explained in terms of group size.

But no matter how important the analysis of social systems may be, we cannot say that it is the one and only aim of social sciences. Sociological theories also work to explain individual behavior and may study voting behavior, household consumption, and more generally the attitudes and choices of individuals. With the exception of explanations proposed by psychological theories, sociologists explain these behaviors by analyzing characteristics of individuals in their relation to their environment.

It is not hard to show that such problems run through the entire history of sociology; here, however, I will be proceeding logically rather than chronologically. To this effect, I will be proposing a typology whose categories subsume the main sociological theories and bring out relations of inclusion or exclusion concerning models deduced from those theories. I will examine the objectives, problems, and the solutions of macrosociological and microsociological theories, while asking whether they are intractably antinomic or whether there are potential means of passing from one to the other.

2. *Outline of a Typology of Micro- and Macrotheories*

The first definition of micrology and macrology, the simplest but also the least adequate, as indicated above, is etymological: the terms refer to the size of the units of observation used. Micrology is understood to be limited to small units, whereas macrology analyzes groups, collectivities, communities, organizations, institutions, even entire societies. This first, unidimensional definition, is valid for neither sociology nor economics; the observation units used with certain theories do not fall into either of the two categories. The ultimate observation unit in symbolic interactionism or ethnomethodology, for example, is not the individual but the situation in which individuals interact. Nor is size a more relevant criterion, because the unit of observation for certain macrosociological theories, for example, is a limited number of small groups.

Similar definition problems are encountered in economics. The classical distinction between micro- and macroeconomics also refers to several levels. While microeconomics is the study of the behavior of individual decision-makers (households, firms) or, more exactly, choice as limited by different conditions, macroeconomics seeks to analyze the economy as a whole by structuring its hypotheses around aggregate variables (e.g., national income or consumption). But what is true of sociology is also true of economics: it is not so much the unit of observation that matters as the way in which problems are simplified.

The second definition proposed by Cherkaoui (1997) is multidimensional; it has at least three dimensions: (a) the nature of the patterns to be explained; (b) the type of hypotheses—rationalist or a-rationalist—used in conducting the explanation; (c) the unit of observation and analysis used in empirical testing of propositions. These units may be individuals, situations, or structures; the point is that they must be consistent with the hypothesis demonstrated, as discussed above.

The first dimension pertains to the particular problem under study, which may involve either individual behavior regularities or social and group regularities. Voting for a particular political party, consuming a product, adopting a particular attitude, making an economic or social choice are examples of regularities in individual behavior. On the other hand, accounting for the degree of heterogeneity in a given society (the frequency of intermarriage between ethnic or religious groups, for example); for the existence and intensity of cooperation or conflict within an organization; for integration and differentiation; for macroscopic balances andmbalances, involves studying social patterns.

Let us consider the following problem taken from social stratification studies, one of the major fields of macrosociology. When we conduct empirical research on the effects of the father's occupation or the individual's level of education on the social status attained by that individual, we are positioning ourselves at the microsociological level (though we may not be aware of it)—despite the fact that our project is macrosociological. As long as the propositions demonstrated are micrological, we are not violating logical rules; we can legitimately make assertions of the following type: 'When, for individuals, the level of education increases, so does occupational status.' Such affirmations derive from empirical research conducted in accordance with known, standardized rules: construction of a random sample, use of analytical statistical methods in which the individual is the relevant unit. But in proceeding this way, we are working from several implicit hypotheses, the simplest of which affirms that individuals' decisions are entirely independent of each other. Nothing precludes us from applying such simplifications as long as our hypotheses or empirically tested propositions do not go beyond

the microsociological level. But it is clear that (a) individual choices are in fact interdependent; (b) there is not an infinite number of occupational positions; (c) the set of such positions constitutes a structure whose parameters function as constraints on individual choices. Obviously, when I decide to apply for a post in sociology in a given institution, (a) I am competing with other applicants (interdependence of individuals); (b) I know that the number of positions in my discipline for which applicants are competing is limited; (c) the job openings in sociology that applicants are competing to fill are the direct result of what other jobs are available both in other disciplines of other departments of the same institution (they also depend, though indirectly, on the number of places available in other institutions, such as all institutions of higher learning); and (d) when making its decision, the institution that has activated competition for the jobs will take into account the interest that other institutions of the same type may have in sociology.

The second dimension of the definition of micro- and macrosociology proposed here refers to theoretical simplifications and the objectives of such simplifications. I will summarily distinguish between two types of hypothesis about the behavior of units of observation and analysis: a theory requires either rationalist or nonrationalist hypotheses (this last type may be divided into a-rational and nonrational).The rational individual is not necessarily to be reduced to that of *Homo æconomicus*, whose particularity is to maximize the expected value of his utility function—monetary profit, for example. Following Boudon (1996), we can speak of three different types of rationality: expected, cognitive, and axiological.These theoretical simplifications of course only concern individual actors or units, not structures. If we preclude Hegelian-type visions, which reify universals and endow them with intentionality or rationality, it is obvious that under certain conditions, groups, organizations, and institutions may be assimilated to actors.

The third and last dimension is the level of observation and analysis deemed most relevant to a given theory. That level may be individuals, households, organizations, a context-situation, or population distributions. The individual and the county are the units most commonly used in electoral sociology. Whereas for economic and sociological research, the household is considered the most appropriate unit for studying consumption, in the sociology of organizations and microeconomic theory of the firm, the preferred basic unit is the single firm. For some correlation analysis—that between literacy and the proportion of the state budget allocated for education, for example—rates are required. In the example from social stratification research, we saw why observations made at the individual level were not sufficient to give answers to macrosociological questions; they may even be an inappropriate place to look for such answers as demonstrated by Coleman (1990). The empirical tradition founded at the beginning of the 1950s and popularized by Blau and Duncan's *American Occupational Structure* (1967) and Featherman and Hauser's research (1978), among others, does not allow for a move from the micro to the macro.

		level of analysis (D2) & type of Hypothesis (D3)				
		individual		individual-contextal		structure
		R	NR	R	NR	
problem (D1)	IP	1	3	5	7	9
	SP	2	4	6	8	10

Figure 1
Typology of theories: R, rationalist hypothesis; NR, other types of hypothesis; IP, individual patterns; SP, social patterns; D1, D2, D3, 1st, 2nd, 3rd dimensions

With these few taxonomic principles, we can construct the typology of theories (numbers 1–10 correspond to theory types to be identified) as in Fig. 1. This typology is neither exhaustive nor inalterable, and it can be simplified or complexified, namely by specifying hypotheses here designated 'other than rational.' It does, however, enable us to observe (a) the fundamental distinction between macrological and micrological theories; (b) the paradigmatic principles underlying these theories; (c) the relations of opposition, distance, or inclusion that a given theory has with the others. Finally, it identifies 'at a glance' the historical stages of sociological thought.

To each of the 10 'pure' theory types corresponds one or more theories. Type 1 comprises all theories that share the axioms of the market model of pure and perfect competition, some behaviorist theories, and Homans' (1961) and Blau's (1964) theories of social exchange, among others. For Homans, macrosociological structures are no more than repeated individual behaviors: the interaction between individuals is what generates structures; structures do not determine individual behavior. Type 2 is made up of theories in which the market model is applied to particular sociological fields. Coleman (1990, p. 21 et seq.) presents a much-discussed topical example similar in some points to economic markets: the marriage market or, more exactly, the phenomenon described as 'marriage squeeze.' Psychoanalysis is a type 3 theory, though certain variants of it are closer to type 9. Type 4 comprises stochastic models and some theories of collective behavior: stochastic theories, like the econometrician explanation of the income distribution, do not hypothesize about individuals, while collective behavior theories make use of a-rationalist postulates about agents' behavior, such as those implied in notions of 'suggestion' or 'hypnotic effects' as used by Gustave Le Bon (1895).

Type 5 encompasses the models of adaptive behavior particular to certain behaviorist explanations, together with oligopolistic models like Cournot's,

studied in detail by Simon (1982), while type 6 refers to rational choice theory, including actionist theories that are more complex than type 5. It is no longer simple market models but rather rational behavior models that are called upon in the attempt to build bridges between the micro and the macro. Actionists rediscovered the idea dear to Weber and Popper that all social phenomena (macroscopic level), particularly the structure and function of institutions, should be explained and understood as the result of individual actions, not in terms of universals or collective properties. But in affirming this epistemological principle, actionist theorists did not reduce social phenomena to psychological ones, because in their understanding, action can be influenced by institutions, and institutions are to be explained not only by the actions of individual agents but also by those of other institutions. Here we should cite works by Olson (1966), Schelling (1978), Boudon (1973, 1977, 1990), Hechter (1983), and Coleman (1990).

Type 7 encompasses relativist theories such as ethnomethodology, symbolic interactionism, and phenomenological sociology; primary representatives here are Blumer and Garfinkel. For Blumer, social behavior is negotiated case by case, not imposed by a hidden structure. Macrostructure is unreal, and the empirical foundation of a sociological theory can only be individual behavior in real situations. Whereas Homans left macrosociology intact, interpreting it as a series of individual reinforcements, Blumer denied almost all we claim to know of macrosociological structures. For him, the ultimate basic unit of all research is not the individual, but the situation; that is, interaction between individuals. Psychological principles such as reward and punishment are themselves subject to interpretation in terms of situation (this is why we may categorize this approach as microsociology rather than psychology). Type 8 is made up of role, reference group, and frustration theories. All variants of functionalism and social control theories, from Malinowski through Parsons to Bourdieu, are type 9 theories which explain regularities in individual behavior in terms of variables strictly external to individuals such as values and social rules. Type 9 theories, which stand in opposition to type 1 (classical political economy and Spencerian sociology), dominated sociology until the early 1960s. Finally, macrostructural and network theories make up the 10th and last type (see Fararo 1989).

In fact, only type 2, 4, 6, 8, and 9 theories take on the problem of the connection between micro- and macrolevels. The first four of these ask macrosociological or macroeconomic questions, but are based on the individual as unit of observation. The assumption—that hypotheses about such units enable us to get from the micro to the macro—is operative particularly in normative theories, which try to explain individual behavior in terms of social norms. All the other theories do not go beyond a single level—microsociological for types 1, 3, 5, and 7 and macrosociological for type 10.

The macrostructuralist project is limited to certain aspects of social reality. It cannot, any more than any other theory, offer a solution to the problem of the links between the micro and the macro. While the rational choice theory presents an undeniable advantage over other theories, it cannot serve as a universal solution: presenting it as unconditionally valid makes it vulnerable to the same dangers other theories have encountered. As for functionalism, it's error was to yield to the temptation of hegemony; it claimed the title of general theory of social systems, when some of its principles are only valid for particular, tightly circumscribed domains. This means that there is a right way of using functionalism and normative theories, just as there is a wrong way of using such strong, all-encompassing theories. There can be no single solution to the problem of the links between micro- and macrosociology, any more than there can be a single mode for explaining all phenomena (the first of these problems being only an aspect of the second).

See also: Action, Theories of Social; Bounded Rationality; Coleman, James Samuel (1926–95); Determinism: Social and Economic; Durkheim, Emile (1858–1917); Explanation: Conceptions in the Social Sciences; Functionalism in Sociology; Institutions; Interactionism: Symbolic; Methodological Individualism in Sociology; Rational Choice Theory in Sociology; Reduction, Varieties of; Social Psychology: Sociological; Social Psychology, Theories of; Sociology, Epistemology of; Status and Role: Structural Aspects; Structure: Social; System: Social; Theory: Sociological; Tocqueville, Alexis de (1805–59); Weber, Max (1864–1920)

Bibliography

Alexander J, Giesen B, Münch R, Smelser N J (eds.) 1987 *The Micro–Macro Link*. University of California Press, Berkeley, CA

Blau P 1964 *Exchange and Power in Social Life*. John Wiley, New York

Blau P 1977 *Inequality and Heterogeneity*. The Free Press, New York

Blau P, Duncan O D 1967 *The American Occupational Structure*. John Wiley, New York

Blau P, Schwartz J 1984 *Crosscuting Social Circles. Testing a Macrostructural Theory of Intergroup Relations*. Academic Press, Orlando, FL

Boudon R 1973 *L'inégalité des chances*. Colin, Paris [*Education, Opportunity and Social Inequality*. Wiley, New York]

Boudon R 1977 *Effets pervers et ordre social*. Presses Universitaires de France, Paris [*The Unintended Consequences of Social Action*. MacMillan, London, 1982]

Boudon R 1990 *L'art de se persuader*. Fayard, Paris [*Art of Self-Persuation*. Polity Press, London, 1994]

Boudon R 1996 A cognitivist model—generalized rational-choice. *Rationality and Society* **8**: 123–50
Boudon R 1998 Social mechanisms without black boxes. In: Hedström P, Swedberg R (eds.) *Social Mechanisms. An Analytical Approach to Social Theory*. Cambridge University Press, Cambridge, UK
Cherkaoui M 1997 Le réel et ses niveaux. Peut-on toujours fonder la macrologie sur la micrologie? [Reality and levels of reality: Can macrology always be based on micrology?] *Revue Française de Sociologie* **28**: 497–524
Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University, Cambridge, MA
Collins R 1981 On the microfoundations of macrosociology. *American Journal of Sociology* **86**: 984–1014
Duncan O D, Davis B 1953 An alternative to ecological correlation. *American Sociological Review* **15**: 351–7
Durkheim E 1893 *De la division du travail social*. Alcan, Paris [*The Division of Labor in Society*. Free Press, New York, 1947]
Etzioni A (ed.) 1961 *Complex Organizations*. Holt Rienhart, New York
Fararo T 1989 *The Meaning of General Theoretical Sociology. Tradition and Formalization*. Cambridge University Press,, Cambridge, UK
Featherman D L, Hauser R 1978 *Opportunity and Change*. Free Press, New York
Goodman L A 1953 Ecological regression and behavior of individuals. *American Sociological Review* **18**: 663–4
Goodman L A 1959 Some alternative to ecological correlation. *American Journal of Sociology* **64**: 610–25
Harsany J C 1968 Individualistic and functionalistic explanations in the light of game theory. In: Lakatos I, Musgrave A (eds.) *Problems in the Philosophy of Science*. North Holland, Amsterdam
Hechter M (ed.) 1983 *The Microfoundations of Macrosociology*. Temple University Press, Philadelphia, PA
Hedström P Swedberg R (eds.) 1998 *Social Mechanisms. An Analytical Approach to Social Theory*. Cambridge University Press, Cambridge, UK
Homans G C 1961 *Social Behavior. Its Elementary Forms*. Harcourt and Brace, New York
Huber J 1990 Macro–micro links in gender stratification. *American Sociological Review* **55**: 1–10
Lakatos I Musgrave A (eds.) 1968 *Problems in the Philosophy of Science*. North Holland, Amsterdam
Lazarsfeld P, Menzel H 1961 On the relations between individual and collective properties. In: Etzioni A (ed.) *Complex Organizations*. Holt Rienhart, New York
LeBon G 1895/1975 *Psychologie des Foules*. Retz, Paris
Lenski G 1966 *Power and Privilege. A Theory of Social Stratification*. McGraw-Hill, New York
Moore B 1966 *Social Origins of Dictatorship and Democracy*. Beacon, Boston
Olson M 1966 *Theory of Collective Action*. Harvard University Press, Cambridge, MA
Popper K 1934 *Die Logik der Forschung* [*The Logic of Scientific Research*. Routledge & Kegan Paul, London, 1959]
Popper K 1963 *Conjectures and Refutations*. Routledge & Kegan Paul, London
Robinson W S 1950 Ecological correlations and the behavior of individuals. *American Sociological Review* **15**: 351–7
Schelling T C 1978 *Micromotives and Macrobehavior*. Norton Co., New York
Skocpol T 1979 *States and Social Revolutions*. Cambridge University Press, New York
Simon H 1977 On judging the plausibility of theories. In: Simon (1977) *Models of Discovery*. Reidel Publishing Co., Dordrecht, The Netherlands
Simon H (ed.) 1977 *Models of Discovery*. Reidel Publishing Co., Dordrecht, The Netherlands
Simon H 1982 *Models of Bounded Rationality*. MIT Press, Cambridge, MA
Sombart W 1906 *Warum gibt es in den Vereinigten Staaten keinen Sozialismus*? [*Why is there no Socialism in the United States*. M. E. Sharpe, New York, 1976]
Tocqueville A de 1856 *L'Ancien régime et la Révolution*, in *Œuvres Complètes*. Edition Gallimard, Paris [*The Old Regime and The Revolution*. Doubleday, New York, 1955]
Turner JH 1984 *Societal Stratification. A Theoretical Analysis*. Columbia University Press, New York
Weber M 1905 *Die protestantische Ethik und der 'Geist' des Kapitalismus* [*The Protestant Ethic and the Spirit of Capitalism*. Scribner, New York, 1950]
Wippler R, Lindenberg S 1987 Collective phenomena and rational choice. In Alexander et al. (eds.) *The Micro–Macro Link*. University of California Press, Berkeley, CA

M. Cherkaoui

Macrostructure in Discourse Comprehension, Psychology of

The availability of a macrostructure representation may influence both the reading process and memory for text content. During reading, the identification of macropropositions provides a context for the interpretation of subordinate content. After reading, the macrostructure representation can guide a systematic search of the reader's memory representation.

Discourse comprehension may be thought of as a process of constructing a memory representation. In the case of text comprehension, a reader is credited with good comprehension to the extent that the reader constructs a coherent representation that is consistent with the message intended by the author. It is particularly critical that the reader accurately represent the text's main ideas and their organization, or macrostructure. This article addresses the processes involved in the construction and use of macrostructure representations as a critical component of discourse comprehension.

1. *Two Levels of Discourse Representation*

A text has both a horizontal organization, or microstructure, and a vertical organization, or macrostructure (Kintsch and van Dijk 1978, van Dijk 1980, van Dijk and Kintsch 1983). A full understanding of a text requires that a reader represent both the microstructure and the macrostructure.

A text's horizontal organization derives from the fact that most sentences can be interpreted in the context of the immediately preceding sentence. A microstructure representation captures these local coherence relations. A formal analysis of a text's microstructure requires that the sentences of the text be analyzed into their underlying ideas, or propositions (see, e.g., *Propositional Representations in Psychology*) and their relations. Propositions are related, for example, if they share referents or if they denote events that are causally related (see, e.g., *Text Comprehension: Models in Psychology*; *Narrative Comprehension, Psychology of*). The resulting analysis of a text's microstructure can be summarized in a network representation with propositions represented as nodes and relations represented as connections between the nodes.

A text's vertical organization derives from the fact that expressed and implied propositions are hierarchically related. The macrostructure represents these hierarchical relationships and thus captures the basis of the global coherence of a text. Formally, a macrostructure representation can be derived from a microstructure representation by the application of three macrorules. One rule deletes micropropositions that are not interpretation conditions of other propositions. A second rule creates new macropropositions by generalization from more specific micropropositions in the microstructure. Finally, a third rule constructs new macropropositions by replacing sequences of micropropositions that denote more complex events. The three macrorules are applied recursively to create new macropropositions that dominate increasingly larger sections of the text. Intuitively, it is useful to think of successively higher levels of a macrostructure as corresponding to increasingly concise summaries of the text.

2. *Constructing a Macrostructure Representation While Reading*

2.1 *Some Cognitive Obstacles to Macroprocessing*

A macrostructure represents the 'gist' of a text so an adequate representation of a text's macrostructure is essential to text comprehension. However, even in circumstances where a text is well written and the reader has the requisite vocabulary and background knowledge to understand it, there are several impediments to processing a text's macrostructure. One is that a written text is an incomplete representation of the author's intended communication. Authors omit from the written text information that they assume their audience can compute from the information provided. Sometimes, they will explicitly state a macroproposition; other times, they will leave it to the reader to infer (generalize or construct) the relevant macroproposition from the more specific information that is explicitly provided.

A second obstacle to macroprocessing is that related macropropositions are often widely separated in a text. This is because the written text is linearly organized whereas the underlying macrostructure is hierarchically organized. In many models of discourse comprehension (e.g., Kintsch and van Dijk 1978; see also *Text Comprehension: Models in Psychology*), this fact creates problems for identifying the relationship between text topics because of readers' memory limitations.

Finally, the identification or construction of a macroproposition can involve the reduction of large amounts of information. For example, a page-long description of a sequence of events might ultimately be summarized by a single macroproposition (e.g., John ate at an Italian restaurant). Again, construction of the appropriate macroproposition may place a heavy burden on the memory abilities of the reader.

2.2 *Monitoring the Macrostructure During Reading*

How do readers cope with the cognitive demands of macroprocessing? There are several empirical demonstrations that readers are sensitive to topic changes and pay more attention to sentences that communicate macropropositions. These important junctures in the underlying macrostructure are signaled in the written text in a variety of redundant ways.

An author can use a variety of explicit signals to alert readers to macrorelevant information (Lorch 1989). Headings, overviews, and other devices emphasize the main topics and their organization within a text and have consistent benefits on reader's memory for a text's macrostructure. In addition, function indicators (e.g., 'It is important to note …'; 'To sum up …') can be used to indicate the relevance of specific statements within a text (van Dijk 1980) with the consequence that readers attend more carefully to the cued information. An author can also be very explicit about a change of topics (e.g., 'In contrast to …'; 'Consider another issue').

The underlying macrostructure is also reflected in the distribution of information within a text. A topic is introduced and elaborated, then a new topic is introduced. The identity of the currently relevant discourse topic is usually reflected in the topic-comment structure of the sentences that elaborate the topic (Kieras 1981). Thus, repeated sentence topics provide a basis for identifying the relevant discourse topic. When a change of topic occurs, several writing conventions aid the process of recognizing the change and identifying the new topic. A change of topic often corresponds with a break in local text coherence which, in turn, is often indicated by paragraph structure. The identification of topic transitions may be further facilitated by the use of marked syntactic

constructions that shift the reader's focus to the new topic. Finally, after a transition has been made, the identification of a new topic is facilitated by the convention that the initial sentence of a new paragraph often corresponds to a macroproposition (Kieras 1981).

2.3 Representing Macropropositions and their Relationships

Although authors occasionally denote the macrorelevance of a statement explicitly, macropropositions are frequently left implicit. In that event, readers are expected to use their prior knowledge and their understanding of information presented in the text to infer relevant macropropositions (see, e.g., *Inferences in Discourse, Psychology of*). Even when a macroproposition is explicitly communicated, readers usually must determine its macrorelevance. Syntactic cues, initial position in a paragraph, and other cues may serve as heuristics in identifying the *possible* macrorelevance of a statement. However, the macrorelevance of a proposition must be *verified* by determining that it occupies a superordinate position with respect to other propositions in the text (i.e., it is elaborated by more specific propositions).

When a macroproposition is identified or inferred, its place in the macrostructure must be determined. This involves not only determining its semantic relationships to subordinate micropropositions, but also determining its relationships to previously identified macropropositions. Again, the author sometimes explicitly communicates these relationships; however, readers are often left to infer how a newly identified macroproposition is related. This involves first identifying potentially related information, then evaluating its relationship to the new macroproposition.

There is very little definitive research on how readers infer implicit macropropositions and how they determine unstated relationships between macropropositions. However, such macroprocessing surely entails interactions between (a) cognitive processes that operate automatically in response to information presented in the text and (b) processing strategies that are invoked by readers to meet their particular reading goals. Successful macroprocessing requires that a reader be able to access relevant background knowledge and information in the text representation. There is substantial empirical evidence of powerful, automatic memory retrieval processes that serve this purpose (Kintsch 1988, Myers and O'Brien 1998; see also *Knowledge Activation in Text Comprehension and Problem Solving, Psychology of*). The involvement of strategic processes is demonstrated by evidence that the reading task modulates readers' macroprocessing: readers are sensitive to macrorelevant information even when the reading task renders such information irrelevant, but they pay even greater attention to the macrostructure when it is relevant to their reading goal (e.g., summarization).

3. Using a Macrostructure Representation

3.1 Influences on Comprehension Processes During Reading

The macrostructure representation a reader constructs during reading has a privileged status with respect to processing of subsequent text. Macropropositions are more accessible than micropropositions so they are particularly likely to serve as the context for the interpretation of new information during reading (McKoon and Ratcliff 1992). This is true in two respects. First, because a recently identified macroproposition is likely to be the most available information when a new sentence is read, it will be the first context consulted as the reader attempts to establish the local coherence of the new information (Kintsch and van Dijk 1978, Kintsch 1998). Second, because macropropositions are typically elaborated in a text, macropropositions should generally be relatively accessible from the text representation even when they were not recently processed. Thus, macropropositions should also be relatively available as a context for establishing global coherence relations (Myers and O'Brien 1998). Indeed, there is substantial empirical evidence that macrorelevant information (e.g., titles) can greatly influence how readers interpret statements and organize their text representations.

3.2 Influences on Memory for Text

A macrostructure representation can also play a prominent role in memory for a text after it has been read. For the same reasons that individual macropropositions tend to be relatively accessible during reading, they are also memorable after reading. As elaborated information, there are many retrieval routes to macropropositions in the text representation (see, e.g., *Memory for Text*). Further, in tasks that emphasize reconstructive processes (e.g., free recall), readers often use their macrostructure representations strategically as a retrieval plan to cue their memories for text content.

4. Future Directions

There are several directions in which future research should develop. First, there is a need for new methods of studying macroprocessing as they occur *on-line* (i.e., during reading). Although there are many empirical studies demonstrating that readers are sensitive to macrorelevant information as they read, very few studies address the nature of the cognitive operations involved in verifying macropropositions, inferring

macropropositions, or determining the relationships among macropropositions. Most on-line methods require great precision in specifying the text conditions eliciting a given process (see, e.g., *Priming, Cognitive Psychology of*). Because macroprocessing often involves the integration of several propositions and depends heavily on a reader's background knowledge, such theoretical precision may not be possible. Thus, methods are required that are sensitive to on-line processes, but which do not require pinpointing a particular text event as *the* stimulus eliciting some process.

In addition to the need for new research methods, theoretical advances are needed in several areas. More detailed models are needed of the cognitive processes that underlie the construction of a macrostructure. The models should identify the text conditions that trigger the processes. More detailed models are also needed of how the availability of a macrostructure representation influences subsequent processing. Models are also needed to explain how the construction of a macrostructure interacts with a reader's knowledge of conventional text structures (e.g., narrative; empirical research report) or what van Dijk and Kintsch (1983) call the superstructure of a text. Finally, theoretical advancement on all fronts depends on the development of ways to model the involvement of readers' background knowledge in all aspects of macroprocessing. Current theory ascribes an extensive role to background knowledge, but promising methods to model its role have only recently begun to take shape (Kintsch 1998; see also *Semantic Processing: Statistical Approaches*).

See also: Comprehension, Cognitive Psychology of; Eye Movements in Reading; Learning from Text; Literary Texts: Comprehension and Memory; Memory: Organization and Recall; Memory Retrieval; Reconstructive Memory, Psychology of; Schemas, Frames, and Scripts in Cognitive Psychology; Sentence Comprehension, Psychology of; Situation Model: Psychological.

Bibliography

Kieras D E 1981 Component processes in the comprehension of simple prose. *Journal of Verbal Learning Behavior* **20**: 1–23

Kintsch W 1988 The role of knowledge in discourse processing: A construction–integration model. *Psychological Review* **95**: 163–82

Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, New York

Kintsch W, van Dijk T A 1978 Towards a model of discourse comprehension and production. *Psychological Review* **85**: 363–94

Lorch R F Jr 1989 Text-signaling devices and their effects on reading and memory processes. *Educational Psychological Review* **1**: 209–34

McKoon G, Ratcliff R 1992 Inference during reading. *Psychological Review* **99**: 440–66

Myers J L, O'Brien E J 1998 Accessing the discourse representation during reading. *Discourse Processes* **26**: 131–57

van Dijk T A 1980 *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Lawrence Erlbaum, Hillsdale, NJ

van Dijk T A, Kintsch W 1983 *Strategies of Discourse Comprehension*. Academic Press, New York

R. F. Lorch, Jr.

Mafia

1. The Setting

The term 'mafia' comes originally from Sicily, where it refers to the private use of violence in public domains. Despite its feudal and patrimonial dispositions, mafia is a modern phenomenon. Mafia developed in the slipstream of Italian unification when a modern state structure imposed itself on a predominantly agrarian society still largely feudal in its basic features. In the absence of effective central control over the means of violence, mafia took shape as an *instrumentum regni* of Italian politicians, who chose to rule Sicily through its dominant class of absentee landlords most of whom resided in Palermo, the center of wealth and power of the island.

Mafia presence was particularly strong in and around this city, in the adjacent citrus fruit area of the Conca d'Oro and the vast hinterland of the western interior. This was the area of large cereal-pasture holdings (the so-called *latifondi* or *ex-feudi*), the management of which the absentee owners entrusted to ruthless local, upwardly-mobile leaseholders and overseers who had a reputation for violence. These local bosses and their retainers formed the backbone of the rural mafia, which consisted of loosely-constituted coalitions, called *cosche* or 'families.' Each family or *cosca* controlled its own territory—a village, a town, or an urban neighborhood—and imposed a tribute on all economic transactions. The typical mafioso was a mediator or power-broker, who thrived on the gaps in communication between landlord and peasant, seller and buyer of commodities and services, political candidate and electorate. Ultimately, he held sway over the links between state and local community.

2. Profile

Mafiosi controlled local resources like property, markets, services, and votes. They operated in collusion with members of the elite, most notably urban-based landlords, politicians, government officials, and businessmen. This 'contiguity' involved mutual protection. But the protection which mafiosi offered was some-

times difficult to distinguish from extortion and the boundaries between victim and accomplice were often likewise blurred. In the no-man's-land between public and private domains, mafiosi were left a free hand in the management of local affairs.

Although they exacerbated class tensions between landlords and peasants by their rent-capitalist management of the estates and appropriating land and becoming landowners in their own right, they also controlled these tensions by carving out channels of upward mobility for ambitious and ruthless peasants and shepherds. In Sicily, as probably elsewhere under similar conditions (Brogan 1998, Tilly 1997, 2000, Varese 1994), mafia and politics provided *'carrières ouvertes aux talents'* (Fentress 2000, p. 149). Toward the large mass of landless peasants and shepherds, from whose ranks they usually originated, their attitude involved undisguised disdain and exploitation. When indicted for violent crimes, mafiosi were usually acquitted for lack of evidence because of high-level protection and because no local witness would testify against them. This greatly helped to enhance their power and their reputation as 'men of respect.' Inspired by both fear and admiration, the local population drew up a 'wall of silence' (*omertà*), which ultimately blocked effective prosecution of mafiosi. Until recently, the power of mafiosi, although surrounded and buttressed by silence, was openly displayed. It illustrated the peaceful coexistence between mafia and the State.

Far from being 'a State within a State,' as magistrates and journalists often represented the phenomenon, mafiosi successfully infiltrated public institutions, including political parties, local governments, the judiciary, banks, and legal firms. They did so through their own personal networks of 'friends,' rather than as members of a centralized organization. With the extension of the suffrage, they enlarged their grip on the electorate and controlled more votes. With the extension of postwar government aid to 'develop' the country's southern peripheries, mafiosi swallowed up ever more funds, most notably in the urban construction industry, always capable of placing themselves with cunning and force between State and citizen. Sicilian mafia appears as the violent alternative to civil society. This was the price the Italian State eventually had to pay for the pragmatic accommodation which later became known as 'pollution' of public institutions—the other side of peaceful coexistence.

3. *Representations*

The persistent and popular representation of mafia as 'a State within a State,' as a centralized monolithic organization, is wide of the mark. It makes too much of the organization and too little of it at the same time. Presenting mafia as a single unified structure ignores the organization's fluid character as well as its parasitic relationship to the State. To understand how mafia works is to start the investigation at the local level, because what mafia in Sicily comes down to is local cliques, structured by ties of kinship, marriage, and friendship, that control local resources with violent methods while enjoying a large measure of impunity because of their contiguity with powerful protectors at higher levels of society who need local strongmen as managers of their properties and as canvassers of votes (Blok 2001). The relationships between these 'families' are characterized by conflict and accommodation—strikingly similar to relations between States—rather than being supervised, coordinated, or controlled from the top by a commission of sorts, as current expressions like *Cosa Nostra* ('Our Thing') and 'The Mafia' suggest. In this respect the alliances of mafia families in the United States (largely of Sicilian extraction), from whom the Sicilians adopted this denomination, showed greater stability (Sterling 1990).

The idea that the Sicilian *cosche* operated like sovereignties is neatly illustrated by the failed attempts to coordinate their relations from above. When the so-called 'Commission' of *Cosa Nostra* was put together to contain intramafia violence and to impose a *pax mafiosa*, as happened for the first time in the late 1950s and again on several occasions in the 1970s, it fell apart because representatives of the various factions could not agree about overall policy, or else tried to outmaneuver each other and dominate the Commission. In 1963, what had remained of this board blew up in the notorious Ciaculli affair, which brought an end to the first mafia war and heralded an era of antimafia policy. The second mafia war unfolded around 1980 and ended in the near extermination of one of the warring factions (Sterling 1990, Catanzaro 1998, Stille 1995, Jamieson 2000). These failings alone clearly demonstrate that Sicilian mafia cannot be understood as a single unitary organization. An analysis of these episodes suggests the image of a fluid network that regulates a changing configuration of alliances in and between local 'families.' In all these instances, incipient hierarchy gave way to segmentation.

4. *Transitions*

Since the 1970s, mafia in Sicily gradually lost its role as a pragmatic extension of the State and assumed the character of a hidden power. Mafiosi disappeared as public figures with public identities. Mafia's peaceful coexistence with the State came to an end. Terror replaced accommodation. Like outlaws elsewhere, mafiosi had to hide from the law (if they did not cooperate outright with the judicial authorities), living clandestine lives in modest accommodations. Several factors are responsible for this major transition.

First, beginning in the early 1970s, Sicilian mafiosi moved into the international narcotics trade which produced huge and fast profits but entailed high risks. These ventures, like the building boom in the early 1960s, resulted in internecine struggles between rival factions—between upstarts from the Palermo hinterland, like the Corleonesi led by Luciano Leggio and Salvatore Riina, and the urban establishment in Palermo of which Stefano Bontate was the most striking figure. State interference provoked campaigns of terror against state officials, politicians, magistrates, and policemen. This new development brought mafiosi into open conflict with the Italian State, which could not stand back and accept open warfare on its territory and close its eyes to the massive infiltration of drug money into its economy. Nor could the government ignore the pressure from neighboring States and other countries to take a strong position against the drug traffic and money laundering. The narcotics trade changed the Sicilian and American mafia. Although it became the mafia's most lucrative enterprise, it has also generated more murder, violence, and betrayal than any other criminal activity. As the profits spiraled, so did the need to protect them (Shawcross and Young 1987, p. 307).

The extent to which the tables had been turned on politicians who had either directly or indirectly protected the interests of mafiosi in exchange for electoral support can best be illustrated by the following anecdote. After the first Sicilian politicians and magistrates who tried to withdraw from the mafia or to combat it with determination were assassinated in public places in 1979 and 1980, to be remembered as the first wave of 'excellent cadavers,' Stefano Bontate, the flamboyant mafia leader in Palermo, warned the then prime minister Giulio Andreotti on his visit to the island in February 1980 with the following words: 'We command in Sicily and if you don't want the DC [Christian Democrats] to disappear completely you'll have to do as we say. Otherwise we'll take away not only your votes in Sicily but also votes in Reggio Calabria and the whole of the south of Italy as well. You'll have to make do with the votes in the north where everyone votes Communist' (Jamieson 2000, p. 222).

Yet helped by a new generation of magistrates and special legislation (which included the coinage of the term *associazione mafiosa*), the Italian State, from the early 1980s onwards, started its crackdown on the mafia, which culminated in effective proceedings (the so-called maxi-trials) against hundreds of mafiosi. The prosecution was substantially helped by mafiosi from losing factions who, partly for reasons of revenge and partly in order to survive, repented, turned state's evidence, and helped build a case against their former peers (Stille 1995).

Behind all these changes lurked an even larger transition. This was the demise of communism in Europe, which unintentionally undermined the mafia's *raison d'être*. Since the immediate postwar years, mafiosi had always supported political parties that opposed communist and socialist parties. Faced with the largest communist party in Western Europe, the Italian Christian Democrats could not forgo the support of an informal power that provided huge blocks of votes and thus effectively staved off what was perceived as 'the danger of communism.' With the disintegration of the Soviet Union and the end of the Cold War, the situation radically changed and the shield of protection that kept mafiosi out of prison began to fall apart. The situation resembled the crackdown on the mafia under Fascism. Prefect Mori's large-scale operation in the late 1920s took place after the Sicilian landed classes had obtained more effective and less costly protection of their property from the new government. This turn exposed mafiosi and left them without protection (Duggan 1989).

One cannot help notice another parallel in the different ways the Italian State proceeded against the mafia. In both cases, prosecutors represented the mafia as a corporate organization, as a single, unified, hierarchical organization controlled from the top by one or more super chiefs, who coordinated the relations between local 'families.' As noted above, there is little evidence for this so-called 'monolithic theorem.' Although recurring tendencies toward centralization were not lacking, their very failure shows the strength and structural continuity of local groups tied together by links of kinship, marriage, and friendship. Indeed it has been argued that the absence of a centralized organization at the regional level served the mafia well and helps account for its remarkable longevity in modern society; it reflects the same paucity of communication that made it possible for mafiosi to place themselves at decisive junctions in the relationships between local and national levels (Blok 1974, Catanzaro 1988).

5. *Conclusion*

Today, the term 'mafia' is widely used as a synonym for organized crime. A still larger conflation of the term includes (ironic) references to any coterie wielding illicit power in formal organizations and institutions, including governments, business firms, and universities. To contain the growing inflation of the terminology, one may restrict the use of the term 'mafia' to denote a form of organized crime that includes collusion and contiguity with people representing public institutions. If we agree that mafiosi are agents who operate with violent methods in the border zone of public and private domains to enrich themselves and convert their wealth into honor and social status, there is no need to widen any further the meaning of the term, because the private use of violence in public places has lately expanded on a world scale and has blurred the familiar boundaries

between private and public spheres in Latin America, Eastern Europe, the former Soviet Union, Southeast Asia, and various African countries. All these cases are unique, but one cannot help noticing their family resemblances.

See also: Violence in Anthropology

Bibliography

Blok A 1974 *The Mafia of a Sicilian Village, 1860–1960. A Study of Violent Peasant Entrepreneurs*. Basil Blackwell, Oxford, UK

Blok A 2001 The blood symbolism of mafia. In: Blok A *Honour and Violence*. Polity, Malden, MA

Brogan P 1998 The drug wars (Colombia). In: Brogan P *World Conflicts*. Bloomsbury, Lanham, MD

Catanzaro R 1988 *Il Delitto Come Impresa. Storia Sociale Della Mafia*. Liviana, Padova [1992 *Men of Respect. A Social History of the Sicilian Mafia*. Free Press, New York]

Duggan C 1989 *Fascism and the Mafia*. Yale University Press, New Haven, CT

Fentress J 2000 *Rebels and Mafiosi. Death in a Sicilian Landscape*. Cornell University Press, Ithaca, NY

Hess H 1970 *Mafia. Zentrale Herrschaft und lokale Gegenmacht*. Mohr, Tübingen [English translation—1973 *Mafia and Mafiosi. The Structure of Power*. Lexington Books, Lexington, MA]

Jamieson A 2000 *The Antimafia. Italy's Fight against Organized Crime*. St Martin's Press, New York

Pallotta G 1977 *Dizionario Storico della Mafia*. Newton Compton, Roma, Italy

Shawcross T, Young M 1987 *Men of Honour. The Confessions of Tommaso Buscetta*. Collins, London

Sterling C 1990 *Octopus. The Long Reach of the International Sicilian Mafia*. Norton, New York

Stille A 1995 *Excellent Cadavers. The Mafia and the Death of the First Italian Republic*. Pantheon, New York

Tilly C 1997 War making and state making as organized crime. In: Tilly C. *Roads from Past to Future*. Rowman & Littlefield, Lanham, MD

Tilly C 2000 Preface. In: Blok A *La Mafia di un Villaggio Siciliano, 1860–1960*. Comunità, Torino

Varese F 1994 Mafia in Russia. *Archives européennes de sociologie* **25**: 224–58

A. Blok

Magic, Anthropology of

Magic is a generic term that refers to different kinds of beliefs and practices related to supernatural forces. Among others, it encompasses such areas as witchcraft, sorcery, and shamanism (see *Witchcraft* and *Shamanism*). In anthropology, the meaning of magic and its different manifestations has been the object of contention and debate for many decades. Accordingly, the aim of this article is to trace the main shifts in the way in which anthropologists have interpreted magic over the years and to suggest reasons for these shifts. It will do so by focusing on the work of key anthropological figures.

1. *The Evolutionists: Magic as a Pseudo-science*

The nineteenth century was the time when anthropology became a recognized academic discipline. It was also the time when the numerous achievements of natural science had captured people's imagination, something partly responsible for the claim that Western societies represented the pinnacle of human endeavor. Faith in both natural science and Western cultural superiority were to bear heavily on the way in which magic was understood by nineteenth-century anthropologists.

For Tylor (1874) the paradigmatic figure of Victorian anthropology, magic was 'one of the most pernicious delusions that ever vexed mankind'; it was also a fundamental characteristic of 'the lowest known stages of civilization.' These claims encapsulate two basic assumptions with which Tylor and his contemporaries were working. First is that natives used magic to achieve certain practical results—control of the natural elements, for example, or restoration of a person's health; second, that all societies follow an evolutionary path with fixed stages of development that represent higher degrees of social and cultural complexity.

On the basis of these assumptions, magic emerged in nineteenth-century anthropology as the negative side of science. Indeed, for Tylor and his contemporaries, it was nothing more than a 'pseudo-science,' a system of thought and practice that attempted to effect changes in the empirical world but was unable to do so. Not that magic as a system of thought was the product of minds radically different from those that invented science. Tylor was clearly against such racist charges. The human mind, he argued, operates everywhere and at all times with the same principles of thought—by associating ideas on the basis of analogy and cause and effect. Yet, despite this fundamental sameness—this 'psychic unity of mankind,' as Tylor would have it—there was still ample room for errors. 'Primitive' people, being at the lowest stage of intellectual development, a stage characterized by simplicity and innocence, were particularly prone to making errors, magic being a paradigmatic example.

To use Tylor's well-known expression, 'primitive' people mistake ideal for a real connections, that is, they treat events that are only accidentally related as if they cause one another. For example, there is nothing in reality that connects the crowing rooster with the rising sun, but because the two events occur regularly and sequentially, 'primitive' people assume that the former causes the latter. Magic is the practical ap-

plication of such misconceptions: the rooster is made to crow so that the sun will rise.

The work of Tylor on magic was further elaborated by Sir James Frazer. To begin with, Frazer (1922/1950) made a distinction between 'sympathetic' and 'contagious' magic. The former was the result of associating ideas on the basis of similarity, the latter on the basis of proximity in space or time. Frazer also took Tylor's evolutionism to its logical extreme. As Tambiah (1990) points out, he arranged 'magic, religion and science in an evolutionary lineal scheme, with the unsupportable suggestion that magic preceded religion in time, and with the inescapable inference … that science must inevitably dissolve religion in our time.'

2. *Twentieth-century Paradigms*

2.1 *Magic as a Source of Hope and Optimism*

By his own admission, Branislaw Malinowski was drawn to anthropology after having read Frazer's work. Yet even though inspired by Frazer, Malinowski was already beginning to question the wisdom of Victorian evolutionism. Moreover, having spent the years of World War I on the Trobriand Islands of the south west Pacific, he developed a difference sense of the place of magic in native life. Indeed, one of Malinowski's aims was to show that belief in magic was not equivalent to irrationalism and mysticism. This marks a fundamental shift in Western perceptions of magic and native life in general and constitutes the basis for all subsequent anthropological studies of magico-religious systems. In more general terms, it marks a shift, at this point still quite subtle, toward a less ethnocentric view of the non-Western world.

In his extended essay *Magic, Science and Religion*, Malinowski (1928/1954) sets out to show that much of what has been written about magic was speculative and had little basis in reality. To this effect, he engages both Tylor and the French philosopher-turned-anthropologist Lévy-Bruhl. The former, Malinowski argues, depicts natives as highly contemplative and rational, much like intellectuals; the latter as being 'hopelessly and completely immersed in a mystical frame of mind' (1928/1954). Malinowski found the natives of the Trobriand Islands to be practical people concerned with such matters as fishing, gardening, building canoes, and tribal festivities. This is not to say that they did not also practice magic; but they did so only when their stock of practical knowledge and experience ran out.

In a celebrated example, Malinowski shows that when fishing in the lagoon, natives do not resort to magical rites; they rely instead on their knowledge and skill. In open-sea fishing, however, which is 'full of danger and uncertainty, there is an extensive magical ritual to secure safety and good results' (1928/1954). This and other similar examples, Malinowski argues, show that natives turn to magic only in situations of emotional stress. As such, magic fulfills an important psychological function; it ritualizes optimism and enhances human confidence and hope.

2.2 *Evans-Pritchard: The Transition to Symbolism*

Even though highly suspicious of evolutionism, Malinowski was never able to transcend the other tenet of Victorian anthropology, namely, that the primary aim of magic was the attainment of practical results. Nor by extension was he able to go beyond the notion that, ultimately, magic was a pseudo-science. The shift from this utilitarian perception to one that views magic as a meaningful and meaning-generating phenomenon came with Evans-Pritchard (1937/1976) in his classic study of Zande witchcraft. Evans-Pritchard was working within the broad sociological problematic developed by Emile Durkheim, which emphasized social stability and treated religion as social institution that contributed to it. In his study of Zande witchcraft, Evans-Pritchard extended the argument to include magical beliefs and practices. At the same time, Evans-Pritchard continued the debate with Lévy-Bruhl that Malinowski began. It is in this confrontation with the French philosopher that he lays the ground for the paradigm that was to dominate subsequent anthropological studies of magico-religious systems and is variously known as the symbolic, interpretive, or cultural approach.

Evans-Pritchard, then, analyzes Zande witchcraft along two axes. The first explores how witchcraft contributes to the cohesion of Zande society—its social function. The other examines when and why the Zande have recourse to it—its cultural function. Witchcraft, Evans-Pritchard points out, embraces a system of values that regulate Zande behavior. No one knows for certain who might be a witch, but spiteful, moody, ill-tempered, bad-mannered, or greedy individuals are prime suspects. Belief in witchcraft acts in a way that encourages more positive dispositions and hence curbs anti-social behavior. To begin with, if the Zande suspect that their neighbors are witches, they are careful not to offend them from fear of being bewitched. Those Zande who are spiteful or jealous, on the other hand, will try to curb their spitefulness. Those of whom they are spiteful may themselves be witches and seek to injure them in return for their spitefulness. In this way, Evans-Pritchard argues, friction and conflict and kept in check and the cohesion of Zande society maintained.

Yet social stability is hardly what the Zande have in mind when they turn to witchcraft. Their purpose, according to Evans-Pritchard, is to explain misfortune, to make events such as the accidental injury of a relative, sickness, even death itself socially meaningful and relevant. Explanations of misfortune through witchcraft set in motion checks and balances that

reproduce social cohesion, but this is only the by-product of a system whose rationale lies elsewhere. Witchcraft is first and foremost an idiom (or a symbol, as we would say today), the language that the Zande use to speak about misfortune so that misfortune is brought into the social domain and can be socially dealt with. Evans-Pritchard does not attempt to explain why the Zande feel the need to turn misfortune into a social issue. He is concerned primarily with explaining that witchcraft, which effects this transformation, has nothing to do with mysticism or, at any rate, does not at all mean that 'primitive' people are unable to perceive how the empirical world operates. Such was the accusation leveled by Lévy-Bruhl whom Evans-Pritchard takes to task.

Witchcraft, Evans-Pritchard points out, does not try to explain how things happen, but rather why they happen. For example, it does not try to explain how old granaries collapse, but rather why they collapse at a certain time and injure certain people. The Zande are well aware that termites eat away the foundations of granaries and that wood decays; they know that it is because of these natural causes that old granaries collapse. What they try to explain through witchcraft accusations is the timing of the event. Why should a particular granary collapse 'at the particular moment when these particular people were sitting beneath it? Through the years it might have collapsed, so why should it fall when certain people sought its kindly shelter?' (1937/1976). When the Zande say that it is witchcraft that caused the granary to collapse, then, it is not because they see a witch push it over; they see termites and rotten wood.

Witchcraft offers an explanation over and above natural causation, one that links the collapse of the granary with the fact that certain people were at that point in time sitting underneath it. We, Evans-Pritchard points out, do not have an explanation for the intersection of the two events; we say that they have independent causes and that their intersection is an accident. The Zande provide the missing link; they say that it is witchcraft.

2.3 Magic as a Symbolic Phenomenon

After Evans-Pritchard's ground-breaking book, the notion that magic is a symbolic phenomenon, something that reflects underlying social and cultural realities rather than an irrational, mystical practice becomes the central argument in most anthropological studies. In what follows, I discuss several well-known works, beginning with Douglas's (1966) famous analysis of pollution and taboo (see *Taboo*).

In this book, Douglas deals with the accusation that 'primitive' people do not make a distinction between the holy and the unclean—an accusation whose implication is that 'primitive' societies have a magical orientation toward the physical world; that people in these societies believe that nature is inhabited by unpredictable and often malevolent supernatural forces; and that contact with unclean things, such as menstrual blood or corpses, brings one perilously close to these forces.

Douglas's response to this accusation is that notions of pollution and taboo are symbolic expressions of conceptual and social disorder, the idiom through which people come to terms with unclassifiable things and intractable social contradictions. Dirt, Douglas points out, is matter out of place, the inevitable outcome of every attempt to classify things and establish a conceptual grip on the world. It is the result of a universal human predicament, the need to transform chaos into a cosmos, that is, a meaningful world. Those things that do not fit our conceptual schemes and hence contradict and undermine them are everywhere designated as dirt; they become dangerous and are either suppressed or avoided. The threat is real enough, since unclassifiable things undermine our ability to construct and lead meaningful lives. At the level of culture, this danger is expressed through beliefs in pollution and defilement and concomitant rules of avoidance.

Contradictions in conceptual schemes are often reproduced in contradictions at the social level. Take, for example, the notion of sexual pollution and the related idea that it is dangerous for men to come into contact with menstrual blood. This notion, Douglas points out, is widespread among New Guinea tribes, but it is particularly pronounced among the Mae Enga. In this tribe men believe that sexual intercourse weakens male strength so that even within marriage it is reduced to the minimum necessary for procreation. This belief may appear as little more than an irrational fear, but as Douglas explains, the taboo reflects conflicts embedded in Enga society. The latter consists of exogamous clans that compete fiercely for prestige and power and are hostile to another. The exogamous clan system, however, forces men to marry outside their own clan, that is, women who come from the enemy's camp. In their pollution beliefs and practices of sexual avoidance, then, the Enga are trying to overcome symbolically a fundamental social contradiction—building marriage and family on enmity.

Less than a decade after Douglas's book, Geertz (1973) wrote a seminal essay that was to consolidate the symbolic approach in the study of magico-religious systems. Geertz's essay is entitled *Religion and as a Cultural System* but the reference to 'religion' should be understood to mean any metaphysical system, including magic. Indeed, Geertz's essay is inspired by, and draws on Evans-Pritchard's discussion of Zande witchcraft. It also draws on Max Weber's religious sociology, but unlike the latter who was working within an evolutionary paradigm and treated magic as historically prior to religion, Geertz does not make a distinction between the two.

Geertz's argument is that 'religion' is a cultural system that helps people maintain faith in the ultimate meaningfulness of life and the world. It does so by accounting for all those anomalies in human experience—paradoxes, puzzles, ambiguities—that threaten to undermine the general order of existence. There are three points, according to Geertz, where chaos, 'a *tumult* of events which lack not just interpretations but *interpretability*' threatens to enter into the world: at the limits of people's ability to explain, to endure suffering, and to make sound moral judgements. Coming up against these limits time and again, Geertz (1973) points out, 'sets ordinary human experience in a permanent context of metaphysical concern and raises the dim, back-of-the-mind suspicions that one may be adrift in an absurd world.' It is at such difficult moments that religion intervenes and affirms the ultimate meaningfulness of the world. It does so not by denying that there is ignorance, suffering, and injustice in the world, but rather by denying that such things are intrinsic to reality—an inescapable fact of the world.

The last work to be examined is Taussig's (1980) book *The Devil and Commodity Fetishism in South America*. This is an interesting and influential book. Although situated within the Marxist problematic, which has been historically unreceptive to the significance of magico-religious systems, the book decisively adopts a symbolic approach. It interprets native beliefs in the devil as a means of making sense of, and resisting symbolically an alien way of life imposed on them from the outside. The book has inspired several anthropologists working in other parts of the world who developed arguments along similar lines.

In rural Colombia, Taussig points out, native people working on sugarcane plantation often enter into a secret contract with the devil. In return for their soul, the devil helps them work faster, increase their production, and hence their wages. Such contracts, however, are said to have several negative consequences. To begin with, it is said that there is no point in investing the extra money in capital goods, such as livestock or land. The money earned with the help of the devil is believed to be inherently barren: the animals will die and the land will become sterile. Moreover, those who enter into such contracts are said to die young and in pain. In short, even though people better their conditions in the short run, the long-term effects of dealing with the devil are destructive. How are these beliefs and practices to be interpreted?

The devil, Taussig argues, signifies the way in which local people are trying to come to terms with the capitalist relations of production imposed on them. It also expresses their evaluation of, and resistance to the new way of life. Having been uprooted from their ancestral homes, the local peasants are now forced into wage labor. Their traditional way of life based on reciprocity and cooperation has been displaced by the capitalist mode of production, exploitation by the landowners, and competition among themselves. It may be the case that they are now financially better off than before, but as the devil stories suggest the quality of their lives has drastically decreased. The devil is an apt symbol for expressing the new order of things.

See also: Collective Beliefs: Sociological Explanation; Collective Memory, Anthropology of; Folklore; Mysticism; Myth in Religion; Myths and Symbols: Organizational; Tradition, Anthropology of

Bibliography

Douglas M 1966 *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*. Praeger, New York

Evans-Pritchard E E 1937/1976 *Witchcraft, Oracles, and Magic Among the Azande*. Clarendon, Oxford, UK

Frazer J 1922/1950 *The Golden Bough*. Macmillan, New York

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Malinowski B 1928/1954 *Magic, Science and Religion and Other Essays*. Doubleday Anchor Books, Garden City, NY

Tambiah S J 1990 *Magic, Science, Religion, and the Scope of Rationality*. Cambridge University Press, Cambridge, UK

Taussig M 1980 *The Devil and Commodity Fetishism in South America*. The University of North Carolina Press, Chapel Hill, NC

Tylor E B 1874 *Primitive Culture*. Holt, New York, Vol. I

V. Argyrou

Magnetoencephalography

Magnetoencephalography or MEG refers to the recording of the rapidly changing magnetic field produced by cerebral activity. From these recordings, one can determine where and when the brain activity in question takes place. MEG has provided important information about sensory as well as higher-level information processing in the brain. MEG is also finding its way to the hospital, to aid in diagnosis and to inform the brain surgeon about the precise location of the functional cortical areas, so that causing damage to important cortical areas can be avoided.

With its total noninvasiveness, simple and quick procedures, good accuracy in locating sources, and millisecond-scale time resolution, MEG is a powerful tool for revealing information-processing sequences and their pathologies of the human brain.

1. Background

The first MEG experiment was performed by David Cohen (1968), who recorded the human alpha rhythm using an induction coil with a million turns of copper

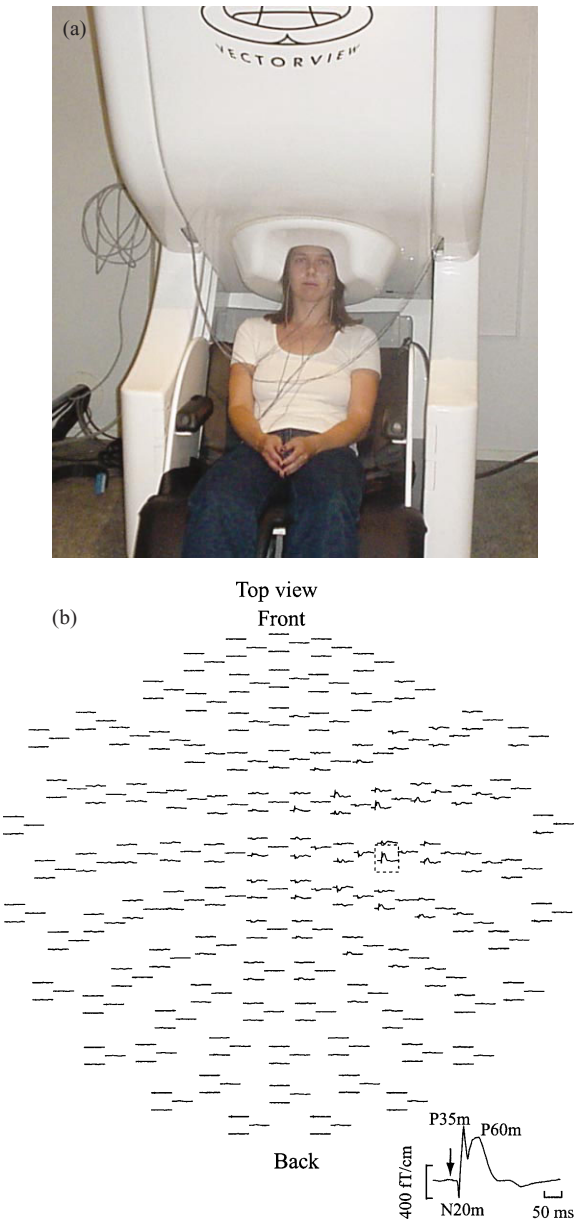

Figure 1
(a) The 306-channel neuromagnetometer at the BioMag Laboratory of the Helsinki University Central Hospital. The superconducting quantum interference device (SQUID) sensors, in a helmet-like array over the head, are immersed in liquid helium at the temperature of 4 K for superconductivity. (b) Cross-section of the sensor array. (c) Signals recorded with the array depicted in (a), when the right median nerve was stimulated at the wrist

wire. Variations in the magnetic flux through the coil induced in it a voltage that could be a detected by computing its correlation with the simultaneous electroencephalography (EEG) alpha rhythm.

Previously, the EEG had been used to record signals close akin those obtained with MEG. However, it was exceedingly difficult to determine which brain structures had given rise to the various EEG signal components. It has been subsequently established that MEG is superior to EEG in determining the location of the underlying brain activity.

The emergence of MEG was made possible by technological advances based on groundbreaking discoveries in physics and technology. Superconductivity, discovered in 1911 (Kamerlingh Onnes, Nobel Prize 1913), was theoretically understood only in the 1950s (Bardeen, Cooper, and Schrieffer, Nobel Prize 1972). This understanding led to the discovery of the Josephson junction (Brian Josephson, Nobel Prize 1973), which made it possible to develop extremely sensitive magnetometers. These magnetometers, based on the superconducting quantum interference device (SQUID), were first developed by Zimmerman and colleagues (Cohen et al. 1970). With the SQUID, the human magnetic alpha rhythm could be detected without correlation or averaging techniques (Cohen 1972); the modulation of this occipitally generated waveform by opening and closing the eyes could be observed in real time.

Initially, building and using SQUIDs was very demanding; only single-channel instruments were available. Sensor arrays were finally developed in the 1980s and early 1990s. At present, the complete magnetic field pattern over the head can be recorded with helmet-shaped magnetometers with up to 306 channels (see Fig. 1).

The 1990s saw an enormous development in various techniques of functional imaging of the human brain. At the same time with the emergence of the large MEG arrays, the so-called functional magnetic resonance imaging (fMRI) was invented. This technique allows the mapping of changes in blood volume and blood oxygenation, indicating areas of enhanced brain activity with millimeter spatial resolution. Being completely harmless to the subject, MEG and fMRI have become very attractive tools for functional imaging of the brain and are increasingly used to complement each other so that both a high spatial and a high temporal resolution can be achieved.

2. *The Genesis of the Neuromagnetic Field*

The magnetic field measured with MEG sensors is produced by electric currents associated with neuronal activity in the brain. When a neuron receives a signal from another neuron at a synapse, an electric current passes through the ion channels that are opened by chemical transmitter molecules. According to the laws of electromagnetism, a magnetic field then encircles the flow of the current. The synaptic currents are

weak, on the order of tens of picoamperes (1 pA = 10^{-12} A). Consequently, the extracerebral magnetic field due to a single postsynaptic potential also is very weak: only on the order of attoteslas (1 aT = 10^{-18} T). Many thousands of these postsynaptic currents have to exist synchronously for the magnetic field to be strong enough to be detected even with the best of the instruments.

It is believed that the magnetic field detectable outside the head is produced by currents initiated at the synapses and guided postsynaptically by the cell structure. Magnetic field lines encircle the flow path of this so-called primary current, extending outside the skull. Because pyramidal cells, which are the largest and most abundant cortical neurons, are predominantly oriented perpendicularly to the cortex, the direction of the primary current also is perpendicular to the cortex. Because MEG detects the tangential component of the primary current only, it is most sensitive to activity in fissural cortex, where the current is oriented parallel to the skull, whereas it does not detect sources that are oriented exactly radially.

Therefore, amplitude comparison between differently oriented sources is possible only if source orientations can be estimated, for example, on the basis of MRI or by simultaneously recorded EEG. It should be noted that the electroencephalogram is produced by precisely the same kinds of mechanisms as the magnetoencephalogram. The main differences between MEG and EEG are (a) that their sensitivity patterns are different, i.e., different cortical locations contribute differently to MEG and EEG signals, and (b) that the sensitivity patterns (the 'lead fields') of EEG electrodes are less accurately known than those of MEG.

The reliable recording of the weak MEG signals requires the use of shielding against external disturbances. For example, variations in the magnetic field of the earth are a million times stronger than the neuromagnetic fields. The recordings are usually performed inside a magnetically shielded room. Noise cancellation is further improved by measuring gradient components of the magnetic field instead of the field itself. In quiet environments, disturbances can be sufficiently dampened by using elaborate compensation methods to permit MEG work even without shielded rooms.

3. *SQUID Magnetometers*

A SQUID magnetometer consists of a superconducting loop of wire (the detector coil) connected inductively to the SQUID loop. The SQUID itself is also a loop made of a superconductor. Its operation is based on allowing a bias current to divide into two alternative paths. When each path includes a weak interruption, the so-called Josephson junction, the voltage over the loop will be a function of the magnetic flux through the loop. The voltage is amplified and from it, the magnetic field strength can be determined.

Magnetoencephalography is limited in spatial resolution in the sense that it is not possible to discern source-current patterns on length scales below 1–2 cm. Nevertheless, one can locate the effective center-of-gravity of the activity with a precision of about 1 mm.

Since the neuromagnetic field outside the head changes slowly as a function of position, near-perfect sampling density can be achieved with a relatively small number of magnetometer channels. Sampling-theoretical calculations (Ahonen et al. 1993) show that most of the field pattern above the head can be obtained with 100–200 channels. All current commercial whole-head devices offer a sufficient number of channels for nearly complete characterization of the neuromagnetic field above the scalp. Therefore, not much will be gained in the future by increasing the number of measurement channels. However, benefits can be expected from efforts to combine MEG with simultaneous EEG recordings and with fMRI and other supplementary information.

4. *Data Analysis*

4.1 *The Forward Problem*

The interpretation of MEG signals requires quantitative understanding of how the electrical currents in the brain produce magnetic fields. This is called the forward problem.

As was pointed out above, neuronal currents are initiated at the synapses and guided postsynaptically by cell structure. The neuronal currents, since ions are moved, change the charge distribution, which, in turn, gives rise to the electric field. The electric field generated in the brain, in its turn, produces passive, so-called volume currents. The volume currents depend on the conductivity structure of the head. The current tends to flow along paths of minimal resistivity. The computation of the complete pattern of electrical currents starting from the active neuronal current (the primary current) requires detailed knowledge of cerebral conductivity and sophisticated finite-element computations. In practice, however, it usually suffices to model the head as a spherically symmetric conductor (consisting of the brain, the skull, and the scalp); then, the extracerebral magnetic field can be easily calculated. In fact, in the spherical model, the complete extracerebral magnetic field can be directly calculated from the primary current; no knowledge about the thickness and conductivity of the skull or the scalp is required.

4.2 *The Inverse Problem*

The benefit from magnetoencephalography is its ability to locate electrical activity in the brain. In the

analysis of neuromagnetic data, it is therefore essential to deal with the inverse problem, i.e., how to determine the internal sources on the basis of measurements performed outside the head. The most common way to tackle this problem is to determine the single source current element (dipole) that most completely explains the MEG pattern.

A current dipole can be thought of as a concentration of primary current to a single point. In MEG applications, a current dipole is used as an equivalent source for unidirectional primary current that may extend over an area of cortex up to several centimeters in diameter. In dipole modeling, the MEG data is interpreted by determining parameters (location, orientation, and amplitude) of a dipole that would best explain the measured signals. This can be done with a computer algorithm that starts from a random dipole position and orientation, and keeps changing these parameters as long as the magnetic field pattern computed from the dipole keeps approaching the experimental one. When no further improvement is obtained, the minimum in the cost function has been reached; a source corresponding to this solution is called the equivalent current dipole (ECD). Dipole fitting is generally done by a computer algorithm that minimizes the sum of squares of the differences of measured signals and those calculated from the dipole. The dipole parameters resulting in minimal discrepancy between measured and calculated signals define the equivalent dipole. The resulting location is near the center of gravity of the underlying activity provided that the active area is small and the primary current, unidirectional.

In cases when a single dipole can not satisfactorily explain the MEG data, one may resort to multidipole modeling. However, the determination of the parameters of several dipoles can be a formidable task as it involves finding a minimum of the sum of squares in a space with the number of dimensions five times the number of dipoles. The algorithm gets easily trapped in a local minimum, which may be far away from the global minimum, corresponding to the actual source locations. In such cases, it may be best to resort to continuous solutions to the inverse problem such as the minimum-norm estimate.

If our prior information of a neural event is limited to knowing that the activity is confined within a specified region of space such as the brain or the cortex, the estimate minimizing the expected error in the inverse-problem solution is the so-called minimum-norm estimate (MNE). This is simply the primary current distribution that has the smallest norm, or overall amplitude, among the current distributions that explain the measurements. MNE was first applied to neuromagnetism by Hämäläinen and Ilmoniemi (1994), who realized that estimates of current distributions could be obtained by suitable linear combinations of magnetometer sensitivity patterns called lead fields. The minimum-norm solution provides a faithful projection of the actual primary current distribution on the lead fields; any improvement in the inverse-problem solution from MNE would require the incorporation of supplementary information about the sources.

When interpreting MEG (or EEG) results, it should be borne in mind that the inverse problem is fundamentally nonunique. Even if the complete electric and magnetic field around the head could be precisely measured, an infinite number of current distributions in the brain could still be constructed that would explain the measured fields. For example, it is always possible that some sources are missed, whatever the measurement set-up. MEG alone is insensitive to radially oriented sources, but even when MEG is complemented with EEG, some sources may remain undetected.

Full utilization of available techniques requires the use of estimation theory to derive optimal solutions to the inverse problem based on all available information, including that provided by MEG and EEG, as well by magnetic resonance imaging (MRI), positron emission tomography (PET), and functional MRI (fMRI). Since MEG is insensitive to the radial component of the primary current, it is essential to perform also EEG measurements in order to get as full a picture of cerebral electrical activity as possible.

If one can assume that the electromagnetically detected electrical activity is limited to areas seen in fMRI or just to the cortex, as revealed by the anatomical MRI, then the inverse solution may be constrained so that no primary current is allowed outside the regions given by these techniques (Ahlfors et al. 1999, Korvenoja et al. 1999). The essential point in fMRI is that it yields the locations of activity while MEG (and/or EEG) gives the temporal sequence and the strength of activation. On the other hand, fMRI can show enhanced activation in regions not visible by MEG or EEG. The time resolution of fMRI, being limited by the dynamics of blood flow and oxygenation, on the order of seconds.

Accordingly, parallel information from these imaging modalities is now being used to constrain the MEG or EEG inverse solutions to limited regions of the cerebrum. This approach provides optimal combined spatial and temporal resolution by exploiting the best aspects of each technology. The combination of the various imaging methods may eliminate the nonuniqueness of the inversion of MEG and EEG data, at least in some applications.

5. *Evoked Magnetic Fields*

MEG signals elicited by sensory stimulation were detected first by Brenner et al. (1975) and Teyler et al. (1975), who were able to record the visually evoked magnetic field. Subsequently, the magnetic field

evoked by auditory (Reite et al. 1978) and other sensory stimuli followed.

Figure 1b shows the kind of evoked-field data that can be obtained with modern MEG systems. In such an arrangement, by presenting stimuli, it is possible to determine the sequence of activation that takes place in the cortex. In addition to somatosensory evoked signals, activity elicited from other modalities can be detected as well. In the visual modality, even face-specific responses, originating from the parietal and temporal lobes, have been reported. Also, MEG has proved to be the best available noninvasive method for revealing the tonotopic organization of the human auditory cortex (Romani et al. 1982). The N1m deflection (peak latency about 100 ms) in the tone-evoked magnetic field, which corresponds to the electrically recorded negative deflection N1 of the event-related potential, was best explained by equivalent current dipoles in tone-frequency-dependent locations of the auditory cortex.

N1m has been reported to show specificity to phonetic features as well. For example, Kuriki and Murase (1989) found that N1m ECDs for speech sounds /ka/ and /a/ differed from one another, with the /ka/ source being located posteriorly to the /a/ source. Interestingly, this difference was obtained in the speech-dominant left hemisphere but not in the right hemisphere, which might indicate that 'the magnetic responses include a component associated with linguistic processing in the auditory cortex.' A further step toward constructing a neuroanatomical phoneme map was taken by Diesch et al. (1996) who reported that neuromagnetic responses to different German vowels had different ECD loci and, further, that the distances between these loci seemed to mirror inter-vowel perceptual distances. This finding supports the idea that the self-organizing neural-network phoneme map of Kohonen may be applicable to model the system of cortical phoneme representation.

MEG has also revealed the functional organization of both primary and secondary somatosensory cortices. The human 'sensory homunculus' of Penfield and Jasper (1954), obtained by stimulating the exposed somatosensory cortex of awake epileptic patients during surgery, was limited to the convexial cortex. MEG, in turn, is at best in studies of fissural cortex. The somatotopic organization of the primary somatosensory cortex was seen with MEG by Okada et al. (1984) and it was further investigated by Hari et al. (1993) and Yang et al. (1993)

Currently there is considerable interest in plastic changes in the functional organization of cortical sensory-receiving areas. Such experience-dependent changes in the somatosensory system were reflected by MEG recordings of Rossini et al. (1994), who first determined ECDs of an early ('N22m') SI response to electrical stimulation of different fingers and thereafter completely anaesthetized four of the fingers. This resulted in the ECD for the unanaesthetized finger shifting in location and becoming stronger. The effect was interpreted by the authors in terms of 'secondary involvement of neighbor neurons within the hand somatic map which usually belong to the anaesthetized fingers,' concluding that 'continuous sensory input from all segments representing a given body district is a prerequisite for normal somatotopic organization' (p. 24).

Kujala et al.'s (2000) MEG recordings from early-blinded subjects showed that their occipital cortex is activated during auditory discrimination. This effect, not found in sighted controls, was elicited by occasional, to-be-detected, changes in the frequency of a repetitive tone.

Menning et al. (2000) demonstrated a training effect in the discrimination of frequency differences in the simple sinusoidal tone. Their standard stimulus was 1000 Hz and deviant stimuli 1005, 1010, and 1050 Hz. The frequency discrimination improved rapidly during the first training week and slowly thereafter. This improvement was associated with enhanced N1m, the magnetically recorded response at about 100 ms from tone onset, as well as in the MMNm response. This enhancement persisted even after training although was somewhat decreased 3 weeks after the training.

MEG is not limited to detecting immediate cortical responses to sensory stimuli—it can also (indirectly) reveal memory traces produced by these stimuli. Previous electrophysiological studies have found the mismatch negativity (MMN), an automatic, attention-independent response to auditory stimulus change, which accurately, albeit indirectly, reflects the properties of stored information, i.e., the neural traces of preceding stimuli (Näätänen 1992). These traces underlying auditory sensory memory can probably be located by determining the sources of the 'mismatch processes,' generating MMN. Several studies using simple tones have indicated that MMNm (the magnetic counterpart of MMN) is generated in supratemporal auditory cortex. This also applies to complex stimuli such as within- and between-category changes in consonant–vowel syllables and to frequency changes in a segment of a complex serial tone pattern.

Using the MMN paradigm, Näätänen et al. (1997) obtained evidence for the existence of phoneme traces in Finnish subjects who were presented with the Finnish vowel /e/ as the standard stimulus and the Finnish vowels /ö/ and /o/ and the Estonian /õ/ as deviants. The stimuli differed from each other only in formant F2 frequency, with F0 and other formants remaining constant. The MMN elicited by the Estonian /õ/ was smaller than that elicited by the Finnish /ö/, even though the acoustical deviation of /õ/ from the standard /e/ was larger than that of /ö/. This phenomenon did not exist in Estonian subjects, who had been exposed to all these vowels in their mother tongue. MEG recordings showed that the

phoneme-related MMN enhancement in the Finnish subjects originated from the left auditory cortex, which was concluded to be the locus of the phoneme traces.

6. Clinical Use of MEG

The noninvasive recording technique of MEG has been used clinically for presurgical mapping, the idea being to locate functional areas, such as the motor cortex or speech areas so that the surgeon may try to avoid harming these (Inoue et al. 1999, Alberstone et al. 2000). Another clinical application is the characterization of epileptic activity (Baumgartner et al. 2000, Minassian et al. 1999), in particular the determination of loci of epileptic activity. This is particularly important when surgery is planned in order to remove the epileptic focus. In children, epileptiform activity has been found to underlie developmental deficits including autism. When characterized with the help of MEG, therapeutic measures can be applied more reliably than without this information.

In addition to clinical applications that benefit the patients themselves, MEG is used in a large variety of studies on different patient groups. For example, Wikström et al. (2000) used MEG to correlate recovery from sensorimotor stroke with the appearance of somatosensory evoked responses in MEG. Pekkonen et al. (1999) studied patients with Alzheimer's disease using MEG and the MMN paradigm, finding impairments in the operation of sensory memory.

7. Conclusion

Magnetoencephalography provides a completely noninvasive tool to probe into the real-time operation of the human brain in experimental conditions that are suitable for studying sensory and cognitive brain functions as well as their disturbances.

In evaluating the future role of MEG in cognitive brain research, it is to be borne in mind that MEG and EEG are the only noninvasive methodologies that provide precise temporal information about the brain activation patterns in its interaction with the environment. Further, of these two methodologies, MEG locates sources considerably more accurately. Of course, this advantage of MEG over EEG only applies to the currents that are visible to MEG, i.e., to those that are not fully radial and not located very deep from the head surface. These limitations, however, should be seen as a strength rather than a weakness of the method, for they could enable one (a) to separately measure a single brain process rather than an amalgamate of multiple temporally overlapping processes which cannot be unequivocally disentangled on the basis of location or orientation, and (b) then, by using this process-specific information, to disambiguate ERP data recorded from the same experimental situation. Therefore, the combined use of the two methodologies is strongly recommended in research aiming at an accurate description of the spatio-temporal activation patterns underlying the brain's cognitive operations. The use of both MEG and EEG, anticipated also in the current developments of the recording systems, will certainly not be made useless or redundant by further developments of EEG signal-analysis techniques which, because of inherent limitations of EEG methodology, cannot challenge the superiority of the combined use. On the contrary, important developments can be made in the field of signal analysis, especially in view of this combined use of the two methodologies, to make their combination especially powerful and an attractive tool of cognitive brain research, which cannot be replaced by any other methodology in the foreseeable future.

See also: Brain Damage: Neuropsychological Rehabilitation; Cortical Activity: Differential Optical Imaging; Functional Brain Imaging; MRI (Magnetic Resonance Imaging) in Psychiatry

Bibliography

Ahlfors S P, Simpson G V, Dale A M, Belliveau J W, Liu A K, Korvenoja A, Virtanen J, Huotilainen M, Tootell R B H, Aronen H J, Ilmoniemi R J 1999 Spatiotemporal activity of a cortical network for processing visual motion revealed by MEG and fMRI. *Journal of Neurophysiology* **82**: 2545–55

Ahonen AI, Hämäläinen M S, Ilmoniemi R J, Kajola M J, Knuutila J E T, Simola J T, Vilkman V T 1993 Sampling theory for neuromagnetic detector arrays. *IEEE Transactions of Biomedical Engineering* **40**: 859–69

Alberstone C D, Skirboll S L, Benzel E C, Sanders J A, Hart B L, Baldwin N G, Tessman C L, Davis J T, Lee R R 2000 Magnetic source imaging and brain surgery: Presurgical and intraoperative planning in 26 patients. *Journal of Neurosurgery* **92**: 79–90

Alho K, Huotilainen M, Tiitinen H, Ilmoniemi R J, Knuutila J, Näätänen R 1993 Memory-related processing of complex sound patterns in human auditory cortex: A MEG study. *Neuroreport* **4**: 391–4

Baumgartner C, Pataraia E, Lindinger G, Deecke L 2000 Neuromagnetic recordings in temporal lobe epilepsy. *Journal of Clinical Neurophysiology* **17**: 177–89

Brenner D, Williamson S J, Kaufman L 1975 Visually evoked magnetic fields of the human brain. *Science* **190**: 480–2

Cohen D 1968 Magnetoencephalography: Evidence of magnetic fields produced by alpha-rhythm currents. *Science* **161**: 784–6

Cohen D, Edelsack E A, Zimmerman J E 1970 Magnetocardiograms taken inside a shielded room with a superconducting point-contact magnetometer. *Applied Physics Letters* **16**: 278–80

Cohen D 1972 Magnetoencephalography: Detection of the brain's electrical activity with a superconducting magnetometer. *Science* **175**: 664–6

Diesch E, Eulitz C, Hampson S, Ross B 1996 The neurotopography of vowels as mirrored by evoked magnetic field measurements. *Brain and Language* **53**: 143–68

Hämäläinen M, Hari R, Ilmoniemi R J, Knuutila J, Lounasmaa O V 1993 Magnetoencephalography—theory, instrument-

ation, and applications to noninvasive studies of the working human brain. *Reviews in Modern Physics* **65**: 413–97
Hämäläinen M S, Ilmoniemi R J 1994 Interpreting magnetic fields of the brain: Minimumnorm estimates. *Medical and Biological Engineering and Computing* **32**: 35–42
Hari R, Karhu J, Hämäläinen M, Knuutila J, Salonen O, Sams M, Vikman V 1993 Functional organization of the human first and second somatosensory cortices: A neuromagnetic study. *European Journal of Neuroscience* **5**: 724–34
Inoue T, Shimizu H, Nakasato N, Kumabe T, Yoshimoto T 1999 Accuracy and limitation of functional magnetic resonance imaging for identification of the central sulcus: Comparison with magnetoencephalography in patients with brain tumors. *Neuroimage* **10**: 738–48
Korvenoja A, Huttunen J, Salli E, Pohjonen H, Martinkauppi S, Palva J M, Lauronen L, Virtanen J, Ilmoniemi R J, Aronen H J 1999 Activation of multiple cortical areas in response to somatosensory stimulation: Combined magnetoencephalographic and functional magnetic resonance imaging. *Human Brain Mapping* **8**: 13–27
Kuriki S, Murase M 1989 Neuromagnetic study of the auditory responses in right and left hemispheres of the human brain evoked by pure tones and speech sounds. *Experimental Brain Research* **77**: 127–34
Menning H, Roberts L E, Pantev C 2000 Plastic changes in the auditory cortex induced by intensive frequency discrimination training. *Neuroreport* **11**: 817–22
Minassian B A, Otsubo H, Weiss S, Elliott I, Rutka J T, Snead O C 3rd 1999 Magnetoencephalographic localization in pediatric epilepsy surgery: Comparison with invasive intracranial electroencephalography. *Annals of Neurology* **46**: 627–33
Näätänen R 1992 *Attention and Brain Function*. Lawrence Erlbaum Associates, Hillsdale, NJ
Näätänen R, Lehtokoski A, Lennes M, Cheour M, Huotilainen M, Iivonen A, Vainio M, Alku P, Ilmoniemi R J, Luuk A, Allik J, Sinkkonen J, Alho K 1997 Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature* **385**: 432–4
Okada Y C, Tanenbaum R, Williamson S J, Kaufman L 1984 Somatotopic organization of the human somatosensory cortex revealed by neuromagnetic measurements. *Experimental Brain Research* **56**: 197–205
Pekkonen E, Jaaskelainen I P, Hietanen M, Huotilainen M, Näätänen R, Ilmoniemi R J, Erkinjuntti T 1999 Impaired preconscious auditory processing and cognitive functions in Alzheimer's disease. *Clinical Neurophysiology* **110**: 1942–7
Penfield W, Jasper H 1954 *Epilepsy and the Functional Anatomy of the Human Brain*. Little, Brown, Boston
Reite M, Edrich J, Zimmerman J T, Zimmerman J E 1978 Human magnetic auditory evoked fields. *Electroencephalography and Clinical Neurophysiology* **45**: 114–17
Romani G L, Williamson S J, Kaufman L 1982 Tonotopic organization of the human auditory cortex. *Science* **216**: 1339–40
Rossini P M, Martino G, Narici L, Pasquarelli A, Peresson M, Pizzella V, Tecchio F, Torrioli G, Romani G L 1994 Short-term brain 'plasticitiy' in humans: transient finger representation changes in sensory cortex somatotopy following ischemic anesthesia. *Brain Research* **642**: 169–77
Teyler T J, Cuffin B N, Cohen D 1975 The visual evoked magnetoencephalogram. *Life Science* **17**: 683–91
Wikström H, Roine R O, Aronen H J, Salonen O, Sinkkonen J, Ilmoniemi R J, Huttunen J 2000 Specific changes in somatosensory evoked magnetic fields during recovery from sensorimotor stroke. *Annals of Neurology* **47**: 353–60
Yang T T, Gallen C C, Schwartz B J, Bloom F E 1993 Noninvasive somatosensory homunculus mapping in humans by using a large-array biomagnetometer. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 3098–102

R. K. Näätänen

Majoritarianism and Majority Rule

This article will use 'majority rule' to denote the voting rule requiring $(n+1)/2$ to carry a decision, and 'majoritarianism' to denote the belief that this rule, and others closely resembling it, are desirable. These ideas may be analyzed by posing four questions:

(a) Is majority rule fair?

(b) Does majority rule efficiently match preferences to outcomes?

(c) Does majority rule comply with basic requirements of formal rationality?

(d) How well do the more complex procedures and electoral systems used in governments match up with majoritarian requirements?

1. *Fairness*

In the long history of human governance, majoritarianism has for the most part been trumped by the rule of minorities. Often, these minorities have been elites based on military prowess, land-ownership, heredity, or priestly knowledge—whether the priests be religious or, for example, socialist. In capitalist and market societies, elites with major ownership stakes and elites constituted by managerial expertise have in many respects held authority denied to popular majorities (Lindblom 1977). In other cases, such as the American Constitution of 1789, *generic* minorities threatened by majorities seeking new uses of governmental power are granted the authority to block action. The US system builds in no bias in favor of a specific minority, and it is indeed prejudicial toward all minorities seeking new uses of government. The favorable bias is specific to defensive minorities seeking to preserve existing policy. The overall historical pattern, we may safely conclude, is that most societies have not been governed in majoritarian fashion.

Majoritarianism arises against this background, in major part, as a demand for political equality. Foremost among the meanings of political equality is the doctrine 'One man [one person], one vote.' This powerful idea links majority rule to the central appeal of popular democracy without specifying exactly the nature of the equality in question. Kenneth May

has conveniently formalized this idea with the criterion of *anonymity*: if a decision process yields the result that X is preferred to Y in a given case, it should again do so for every other case in which the preferences of any two blocks of voters of equal size are transposed (May 1952). Put otherwise, you should always be able to count the votes (and announce a decision) without seeing whose votes they are. Another feature of equality or equity entailed by majority rule concerns the content of the alternatives under consideration. Majority rule is formally neutral in its treatment of alternatives. Whether X or Y is the status quo, for example, makes no difference to the counting of votes. Whether X or Y is pro-labor, respectful of racial minorities, consistent with the tenets of a prevailing ideology, or consistent with the interests of economic elites—none of these or similar considerations about the alternatives in question will matter under majority rule. This additional requirement of 'neutrality' combines with anonymity in 'May's Theorem' to define majority rule as uniquely best among decision rules which respond positively and nonrandomly to voter preferences (May 1952).

The foregoing considerations constitute a strong case for the egalitarian fairness of majoritarianism on an abstractly *a priori* basis. They do not, however, say anything about the fairness of majority rule in specific historical cases where known groups have predictable or announced preferences. Indeed, the neutrality of majority rule toward alternatives implies an embedded incapacity to discriminate in favor of fair, egalitarian, or just policies as against unfair, anti-egalitarian, or unjust ones. Majority rule provides no defense against majorities embracing evil. In racially and religiously divided societies, where strong conflicts predictably divide large groups from small ones, this is a crippling disability which occasions sharp resistance to majority rule. One only needs to think of recent conflict in the Balkans to see this point. The time-honored remedies in cases of this sort are *partition* (to produce more homogeneous, less conflictual, and smaller polities), the protection of *minority rights* against specified forms of majority action within each polity, and, of course, *emigration* opportunities for members of an afflicted minority. The development of civic cultures which surmount and eventually attenuate conflict is probably superior to any of these more mechanical solutions, but requires historical time and good fortune. It is nevertheless worth noting that even some very hierarchical and undemocratic cultures—India, Japan, quite possibly South Africa—have made great strides in this direction (see, in particular, Shapiro 2001).

2. *Preference Efficiency*

Quite apart from considerations of fairness, a case can be made for majority rule based on its efficiency in matching preferences to policy outcomes. Indeed, it can be shown analytically that all other procedures for making collective decisions are inferior to majority rule in their ability to maximize direct fit between preferences and outcomes (See Rae 1969; for a critical evaluation, see Mueller 1989). This applies to other voting rules, rules giving priority to specific types of policy, and all manner of random or arbitrary procedures. Preference-matching efficiency appears most plausible for issues where the determination of basic policy issues and entitlements is concerned. If decisions entail starkly different stakes for different voters ('Shall we destroy the reader's home?') nominal matching of preferences to outcomes is obviously inadequate as a criterion. Where decisions offer gains from trade among the rights so established, market and market-like arrangements appear superior on efficiency criteria. Indeed, as Ronald Coase has shown, trade can turn any well-defined set of initial rights into a Pareto-efficient set of rights on a voluntary basis (Coase 1960). Insofar as such a process of exchange can be embedded in a voting rule, that rule would require unanimity for each change of policy (Wicksell 1896). These market-oriented schemes of course give great leverage to groups and individuals who are privileged by the initial distribution of rights and assets, and are explicitly anti-majoritarian for that reason. It must of course be accepted that neither these market-oriented schemes nor majoritarianism can purport to efficient treatment of preferences which are manipulated, distorted, or subject to intimidation.

3. *Majority Rule and Formal Rationality*

It has been well known since the late eighteenth century that majority rule possesses some quirky features, especially when multiple alternatives are considered in one process. Most famous among these is the 'cycle of voting' identified by the Marquis de Condorcet (1785). In the simplest case, we imagine three voters and three alternatives (Table 1). In sequential application of majority rule, A defeats B on the strength of voters 1 and 2, B beats C via 1 and 3, but then C beats A due to the preferences of voters 2 and 3. This violates the axiom of formal rationality known as transitivity, and opens the possibility that control over agenda sequences is tantamount to control over outcomes. Majority rule is hardly alone in having such quirks. This transitivity requirement, together with others which tap additional aspects of

Table 1
Hypothetical responses of three voters to three alternatives A, B, and C

VOTERS	1	2	3
BEST	A	C	B
	B	A	C
WORST	C	B	A

formal rationality, defines the justly famed Arrow Theorem showing that *all* decision procedures, or 'social welfare functions,' may violate the dictates of formal rationality in one way or another (Arrow 1951).

4. Majoritarianism in Complex Institutions

Far more important than these formal properties for most purposes is the task of realizing majoritarian aims in complex institutions. The American 'electoral college,' for example, contains the very real potential for reversing national majorities—having done so as recently as the fall of 2000. Most single-member electoral systems place an immense premium on spatial distribution of the vote, often giving majority status to parties with well less than majority status in the electorate. Worse yet, where regional parties dominate, as in India, they may fragment outcomes so that nothing close to national majorities are formed. Systems of proportional representation, particularly those based on large districts and exact proportionality, pass the problem of majority formation on to parliaments, and create the opportunity for small parties to exert influence out of all proportion to their size—as with the case of Israel in the 1990s (Taagapera and Shugart 1989). Most, if not all, legislative procedures, entailing many critical veto points, conspicuously violate majority rule. It follows that majoritarianism is less a blueprint than an aspiration.

5. Conclusion

We live in an era when democratic aspirations have very nearly swept the field of direct rivals, particularly among nations that play leading roles in the world economy. Yet in practice, movement toward democratization has been slow, halting, and uncertain in many places. Most of Africa has made little progress, and many key systems from the former Soviet bloc have made minimal progress. Majoritarianism may usefully be seen as one of the key aspirations that should help to shape change and reform in coming decades.

See also: Democracy, History of; Democracy: Normative Theory; Interest Groups; Minimum Winning Coalition, in Politics; Minorities; Party Systems; Political Parties; Political Parties, History of; Rational Choice in Politics

Bibliography

Arrow K J 1951 *Social Choice and Individual Values*. Wiley, New York

Buchanan J M, Tullock G 1962 *The Calculus of Consent*. University of Michigan Press, Ann Arbor, MI

Coase R 1960 The problem of social cost. *Journal of Law & Economics* **3**: 1–44

Condorcet M de 1785 *Essai sur l'Application de l'Analyse à la Probabilité des Decisions Rendues à la Pluraliste des Voix*. De l'Imprim ene royale, Paris

Lindblom C E 1977 *Politics and Markets: The World's Political Economic Systems*. Basic Books, New York

May K 1952 A set of independent, necessary and sufficient conditions for simple majority decision. *Econometrica* **20**: 680–4

Mueller D C 1989 *Public Choice II*. Cambridge University Press, Cambridge, UK

Rae D W 1969 Decision-rules and individual values in constitutional choice. *American Political Science Review* **63**: 40–56

Shapiro I 2001 The state of democratic theory. In: Katznelson I, Milner H H (eds.) *Political Science: The State of a Discipline*. American Political Science Association, Washington DC

Taagapera R, Shugart M S 1989 *Seats and Votes: The Effects and Determinants of Electoral Systems*. Yale University Press, New Haven, CT

Wicksell K 1896 A new principle of just taxation. In: Musgrave R, Peacock A (eds.) 1967 *Classics in the Theory of Public Finance*. St. Martin's Press, New York, pp. 72–118

D. Rae

Majorization and Stochastic Orders

1. Introduction

In a specific context, stochastic orders are orderings of 'largeness' on random variables or vectors: that is, a random variable X is stochastically smaller than another random variable Y or a random vector $\mathbf{X} = (X_1,\ldots, X_d)$ is stochastically smaller than another random vector $\mathbf{Y} = (Y_1,\ldots, Y_d)$ if X tends to take smaller values than Y or the component X_i ($i = 1,\ldots, d$) of $\mathbf{X}$ tend to take smaller values than the corresponding component Y_i of $\mathbf{Y}$. In a broad sense, stochastic orders are orderings that apply to the comparison of random variables or vectors, or probability distributions or measures, and there have been many such stochastic orders. Here, a few stochastic orders that may be of greater interest in the social and behavioral sciences are summarized. These include orderings of dispersion and dependence.

Majorization orderings are orderings of divergence of a vector or function from uniformity. They are useful as a means of proving many inequalities. The vector majorization ordering applies to vectors of the same length which have the same sum of components. It has been generalized to orderings of functions on measure spaces, and in this context, there are overlaps with stochastic orders.

Stochastic and majorization orderings are defined in Sects. 2 and 3, starting with the simplest forms.

Comparisons and contrasts of stochastic and majorization orderings are given in Sect. 4 and illustrative examples of these orderings are given in Sect. 5. The few references which are given provide a guide to the research in majorization and stochastic orderings; they include historical perspectives on the orderings—some often cited early references are Hardy–Littlewood–Pólya, Schur, Lorenz, Dalton for majorization and inequalities, and Lehmann for the simplest stochastic order.

For terminology and notation, increasing (decreasing) is used in place of nondecreasing (nonincreasing), $(z)_+$ denotes $\max\{0, z\}$ over reals, $\stackrel{d}{=}$ denotes random quantities that are equal in distribution, $F^{-1}(v) = \inf\{x: F(x) \geqslant v\}$ is the functional inverse of a univariate cumulative distribution function F, and $X \sim F$ denotes a random variable X having the distribution F.

2. Stochastic Orders

A stochastic order in general means a partial ordering or preordering (the latter need not be reflexive) that compares random variables or vectors, or probability distributions or measures. A stochastic order in a specific sense means the simplest ordering of largeness of random variables; for clarity of terminology, this is referred to as the *stochastic largeness order* (it is called the *usual stochastic order* in Shaked and Shanthikumar 1994). Many stochastic orders generalize this stochastic largeness order in some way. Some main references for this section are Stoyan (1983), Mosler and Scarsini (1993a, 1993b), Shaked and Shanthikumar (1994).

The definitions of the stochastic largeness orders are given next.

Definition 1 (*Stochastic largeness order for random variables.*) *Let X, Y be random variables with respective distribution functions F, G. X is stochastically smaller than Y, written $X \prec^{st} Y$ or $F \prec^{st} G$, if one of the following equivalent conditions hold.*

(a) $\Pr(X > u) \leqslant \Pr(Y > u)$ *for all* $-\infty < u < \infty$, *or* $\Pr(X \leqslant u) \geqslant \Pr(Y \leqslant u)$ *for all* $-\infty < u < \infty$.

(b) $E[\psi(X)] \leqslant E[\psi(Y)]$ *for all increasing functions* ψ *for which the expectations exist.*

(c) *There exist random variables* $\hat{X}$, $\hat{Y}$ *on the same probability space such that* $\hat{X} \stackrel{d}{=} X$, $\hat{Y} \stackrel{d}{=} Y$, *and* $\Pr(\hat{X} \leqslant \hat{Y}) = 1$.

(d) *There exists a monotone kernel* $K: \mathscr{R} \times \mathscr{R} \to [0, 1]$ (*that is*, $K(t, \cdot)$ *is a distribution function for all* $t \in \mathscr{R}$ *and* $K(t, u) = 0$ for $u \leqslant t$) *such that* $G(u) = \int_{-\infty}^{\infty} K(t,u)\,\mathrm{d}F(t)$.

(e) $F^{-1}(v) \leqslant G^{-1}(v)$ *where* F^{-1}, G^{-1} *are the inverse cumulative distribution functions.*

Definition 2 (*Stochastic largeness order in a partially ordered space.*) *Let P_1, P_2 be probability measures on a partially ordered Polish space* $(\mathscr{X}, \mathscr{A})$ [*usually Euclidean space*] *with measure* ν [*usually Lebesgue or counting measure*]. *Let* $\prec$ *be the partial order on* $\mathscr{X}$; *for d-dimensional Euclidean space, the usual partial order is given by* $\mathbf{x} \prec \mathbf{y}$ *if* $x_i \leqslant y_i$, $i = 1,\ldots, d$. P_1 *is stochastically smaller than* P_2, *written* $P_1 \prec^{st} P_2$, *if one of the following equivalent conditions hold.*

(a) $P_1(A) \leqslant P_2(A)$ *for all upper sets* A [*i.e., sets for which* $\mathbf{x} \in A$ *and* $\mathbf{x} \prec \mathbf{y}$ *imply* $\mathbf{y} \in A$].

(b) $\int \psi\,\mathrm{d}P_1 \leqslant \int \psi\,\mathrm{d}P_2$ *for all increasing functions* ψ *for which the expectations exist.*

(c) *There exists random quantities* X_1, X_2 *with values in* $(\mathscr{X}, \mathscr{A})$ *such that* $X_1 \sim P_1$, $X_2 \sim P_2$ *and* $X_1 \prec X_2$ *almost surely.*

(d) *There exists a Markov kernel* $M: \mathscr{X} \times \mathscr{A} \to [0, 1]$ *such that* $M(\mathbf{t}, \cdot)$ *is a measure with support on* $\{\mathbf{z}: \mathbf{t} \prec \mathbf{z}\}$ *and* $P_2(A) = \int M(\mathbf{t}, A) P_1(\mathrm{d}\mathbf{t})$ *for all* $A \in \mathscr{A}$.

Many general stochastic orders take the form of (b) with other classes of functions ψ; for example, convex functions, increasing convex functions, increasing concave functions. Generalizations of the form of (b) are also known as *dilation orderings*. The class of increasing concave functions leads to *second-order dominance*. The class of convex functions leads to a *dispersion ordering* of random variables with the same mean (see Sect. 4), one of many dispersion orderings that have been proposed.

Generalizations for random variables include the hazard rate order, reverse hazard rate order, likelihood ratio order. These are useful for proving results that are intuitively plausible with a stochastic largeness order; sometimes the result will not hold or may be much more difficult to prove with $\prec^{st}$.

Definition 3 (*Some additional univariate stochastic orders.*) *Let X, Y be two random variables with respective cumulative distributions functions F, G and densities f, g.*

(a) *G is larger than F in the likelihood ratio order, denoted by* $F \prec^{LR} G$ or $X \prec^{LR} Y$, *if* $f(x)g(y) \geqslant f(y)g(x)$ *for all* $-\infty < x < y < \infty$.

(b) *G is larger than F in the hazard rate order, denoted by* $F \prec^{HR} G$ *or* $X \prec^{HR} Y$, *if* $f(x)/[1-F(x)] \geqslant g(x)\,[1-G(x)]$ *for all* $-\infty < x < \infty$.

(c) *G is larger than F in the reverse hazard rate order, denoted by* $F \prec^{RHR} G$ *or* $X \prec^{RHR} Y$, *if* $f(x)/F(x) \leqslant g(x)/G(x)$ *for all* $-\infty < x < \infty$.

It is not difficult to check that $F \prec^{LR} G$ implies both $F \prec^{HR} G$, $F \prec^{RHR} G$, and either $F \prec^{HR} G$ or $F \prec^{RHR} G$ implies $F \prec^{st} G$. Intuitively, these concepts provide different ways of comparing the event of X smaller and Y larger vs. the event of Y smaller and X larger. There are multivariate versions of these orderings for random vectors. However, there are also other similar orderings that compare the amount of dependence in multivariate distributions with the set of univariate margins held fixed; these are known as dependence orderings. A basic dependence ordering is given next.

Definition 4 (*Concordance ordering.*) *Let* $\mathscr{F} = \mathscr{F}(F_1,\ldots, F_d)$ *be the class of multivariate distributions with given univariate margins* $F_1,\ldots, F_d$, *where* $d \geqslant 2$. *Let* F, $G \in \mathscr{F}$ *with respective survival functions* $\bar{F}$, $\bar{G}$ (*defined below*). *Suppose* $\mathbf{X} \sim F$ *and* $\mathbf{Y} \sim G$. *Then G is*

more concordant than F, written $F \prec^c G$ *or* $\mathbf{X} \prec^c \mathbf{Y}$, *if* $\forall \mathbf{t} \in \mathscr{R}^d$:

$$F(\mathbf{t}) = \Pr(X_1 \leqslant t_1, \ldots, X_d \leqslant t_d) \leqslant \Pr(Y_1 \leqslant t_1, \ldots, Y_d \leqslant t_d) = G(\mathbf{t})$$

$$\bar{F}(\mathbf{t}) = \Pr(X_1 \geqslant t_1, \ldots, X_d \geqslant t_d) \leqslant \Pr(Y_1 \geqslant t_1, \ldots, Y_d \geqslant t_d) = \bar{G}(\mathbf{t})$$

Note that this *concordance ordering* compares specific upper sets and lower sets rather than all upper sets as in Definition 2. Hence a distribution is larger in the concordance ordering if the random variables are morely likely to simultaneously take large values or simultaneously take small values. Other dependence orderings are given in Scarsini and Shaked (1996) and Joe (1997).

3. *Majorization Orderings*

Vector *majorization* is defined, followed by a general majorization for functions on a measure space. Some main references for this section are Marshall and Olkin (1979), Arnold (1987), Joe (1993).

Definition 5 (*Vector majorization.*) *Let* $\mathbf{x} = (x_1, \ldots, x_n)$ *and* $\mathbf{y} = (y_1, \ldots, y_n)$ *be vectors in* $\mathscr{R}^n$ *such that* $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. $\mathbf{x}$ *is majorized by* $\mathbf{y}$ *or* $\mathbf{y}$ *majorizes* $\mathbf{x}$, *written* $\mathbf{x} \prec^m \mathbf{y}$, *if one of the following equivalent forms holds.*

(a) $\sum_{i=1}^k x_{n-i+1:n} \leqslant \sum_{i=1}^k y_{n-i+1:n}, k = 1, \ldots, n-1$, *or* $\sum_{i=1}^k x_{i:n} \geqslant \sum_{i=1}^k y_{i:n}$, $k = 1, \ldots, n-1$, *where* $(x_{i:n})$, $(y_{i:n})$ *are the elements of* $\mathbf{x}$, $\mathbf{y}$, *respectively, arranged in increasing order.*

(b) $\sum_1^n \psi(x_i) \leqslant \sum_1^n \psi(y_i)$ *for all convex, continuous real-valued functions* ψ.

(c) $\sum_1^n (x_i - t)_+ \leqslant \sum_1^n (y_i - t)_+$ *for all real* t.

(d) *There is a doubly stochastic* $n \times n$ *matrix* P (*non-negative matrix such that row and column sums are* 1) *such that* $\mathbf{y} = \mathbf{x}P$.

(e) $\mathbf{x}$ *is in the convex hull of* $\{(y_{i_1}, \ldots, y_{i_n}) : (i_1, \ldots, i_n)$ *is a permutation of* $(1, \ldots, n)\}$.

Majorization is also called the *Schur order*. There are *weak majorization orders* that do not require the vector sums to be the same. For example, in (b) of Definition 5, for *weak submajorization* and *weak supermajorization*, the class of functions ψ are increasing convex functions or decreasing convex functions, respectively. In economics (e.g., Foster and Shorrocks 1988), in the context of poverty orderings, majorization is called *Lorenz dominance* and the second of the weak majorization orderings is called *general Lorenz dominance*.

Generalizations of Definition 5 have come from generalizations of any one of the equivalent properties. The next definition covers many but not all generalizations of vector majorization. Marshall and Olkin (1979) unified inequalities through majorization, and although generalized majorization leads to inequalities, inequalities have not always been the motivation for extensions.

Let $(\mathscr{X}, \mathscr{A}, \nu)$ be a measure space. For most applications, $\mathscr{X}$ will be a subset of a Euclidean space, and ν will be Lebesgue measure or counting measure. For a non-negative integrable function h on $(\mathscr{X}, \mathscr{A}, \nu)$, let $m_h(t) = \nu(\{x: h(x) > t\})$, $t \geqslant 0$, and let $h^*(u) = m_h^{-1}(u) = \sup\{t: m_h(t) > u\}$, $0 \leqslant u \leqslant \nu(\mathscr{X})$; h^* is the (left-continuous) decreasing rearrangement of h.

Definition 6 *Let* a *and* b *be non-negative integrable functions on* $(\mathscr{X}, \mathscr{A}, \nu)$ *such that* $\int a \, \mathrm{d}\nu = \int b \, \mathrm{d}\nu$. *Then* a *is majorized by* b, *written* $a \prec^m b$, *if one of the following four equivalent conditions hold.*

(a) $\int_0^t a^*(u) \, \mathrm{d}u \leqslant \int_0^t b^*(u) \, \mathrm{d}u$ *for all* $0 \leqslant t < \nu(\mathscr{X})$.

(b) $\int \psi(a) \, \mathrm{d}\nu \leqslant \int \psi(b) \, \mathrm{d}\nu$ *for all convex, continuous real-valued functions* ψ *such that* $\psi(0) = 0$ *and the integrals exist.*

(c) $\int [a-t]_+ \, \mathrm{d}\nu \leqslant \int [b-t]_+ \, \mathrm{d}\nu$ *for all* $t \geqslant 0$.

(d) $\int_t^\infty m_a(s) \, \mathrm{d}s \leqslant \int_t^\infty m_b(s) \, \mathrm{d}s$ *for all* $t \geqslant 0$.

Note that Definition 5 follows from Definition 6 with $\mathscr{X} = \{1, \ldots, n\}$, $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$ and ν being the counting measure. Next a dependence ordering is introduced from majorization to contrast with Definition 4.

Definition 7 (*Dependence ordering.*) *Let* $\mathscr{F} = \mathscr{F}(F_1, \ldots, F_d)$ *be the class of multivariate densities* (*with respect to a measure* ν) *with given univariate margins* $F_1, \ldots, F_d$. *Suppose* F_j *has density* f_j, $j = 1, \ldots, d$. *The* (*constrained*) *majorization ordering* $f \prec^m g$ *of densities* $f, g \in \mathscr{F}$ *is interpretable as an ordering of dependence. The density* $f_I = \Pi_j f_j$ *is among those which are minimal in* $\mathscr{F}$ *with respect to* $\prec^m$.

The interpretation of this ordering is different from the concordance ordering in Sect. 2. Densities in $\mathscr{F}$ are larger in this majorization ordering if they diverge more from $f_I = \Pi_j f_j$. Densities for a random vector that almost satisfy a functional relationship are large in the majorization dependence ordering.

4. *Constrast of Stochastic and Majorization Orderings*

Some relations between stochastic and majorization orderings are summarized in this section. First consider the case of probability vectors over a finite ordered set of size n, indexed as $\{1, \ldots, n\}$. Let $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$ be two probability vectors. If $\mathbf{p}$, $\mathbf{q}$ are increasing, then $\mathbf{p} \prec^{st} \mathbf{q}$ and $\mathbf{p} \prec^m \mathbf{q}$ are the same, and if $\mathbf{p}$, $\mathbf{q}$ are decreasing, then $\mathbf{p} \prec^{st} \mathbf{q}$ and $\mathbf{q} \prec^m \mathbf{p}$ are the same. Next, consider probability measures P_1, P_2 which have masses n^{-1} at $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, respectively. Then Definition 5(b) becomes $\int \psi \, \mathrm{d}P_1 \leqslant \int \psi \, \mathrm{d}P_2$ which is of the form of Definition 2(b).

An example of an overlap of a stochastic ordering for dispersion and the general majorization ordering is

called the *Lorenz ordering* in Arnold (1987), and other names in mathematics and economics. Since majorization orderings are orderings of divergence from uniformity, this connection is not surprising. In Definition 6, let $\mathscr{X} = [0, 1]$, let v be Lebesgue measure, and let $a = F^{-1}$, $b = G^{-1}$, where F, G are the respective distribution functions of non-negative random variables X, Y (with a common finite mean). Since a, b are monotone increasing, the decreasing rearrangements are, respectively, $a^*(t) = F^{-1}(1-t)$ and $b^*(t) = G^{-1}(1-t)$. Condition (b) becomes $\int_0^1 \psi\ (F^{-1}(t))\mathrm{d}t \leqslant \int_0^1 \psi(G^{-1}(t))\mathrm{d}t$ or

$$E[\psi(X)] = \int_0^\infty \psi(s)\ \mathrm{d}F(s) \leqslant \int_0^\infty \psi(s)\ \mathrm{d}G(s) = E[\psi(Y)] \tag{1}$$

for all convex continuous functions ψ. Note that this then implies that $E(X) = E(Y)$, by taking $\psi(s) = s$ and $\psi(s) = -s$.

An interpretation is that with a given mean wealth μ, a distribution of wealth F with a constant F^{-1} is most equitable (this corresponds to a mass of 1 at μ) and a distribution F which is larger in the (Lorenz) ordering is less equitable. In this context, the Lorenz curve of a distribution F is given by

$$L_F(s) = \left[\int_s^1 F^{-1}(1-t)\ \mathrm{d}t\right] \Big/ \left[\int_0^1 F^{-1}(1-t)\mathrm{d}t\right], \quad 0 < s < 1 \tag{2}$$

this is roughly the complement of the term given in Definition 6(a), so a Lorenz curve which is farther below the 45° diagonal is less equitable.

5. *Applications and Illustrative Examples*

Majorization and stochastic orderings are interesting to study for their own sake because of interesting mathematical properties, but they have also been widely applied, for example in operations research, queuing theory, reliability, economics, and many areas of mathematics and statistics. For any ordering, the class of order preserving or isotonic functions is of interest. Member functions of an isotonic class can be used as measures of various types; examples are measures of inequality and measures of dependence.

Example 1 (measures of inequality and dependence). Measures of inequality include Gini indices and extensions, poverty indices in economics, and measures of diversity in biology (Marshall and Olkin 1979, Foster and Shorrocks 1988). For summarizing data consisting of a mix of categorical and continuous variables, measures of association or dependence are useful. Concordance orderings lead to measures monotone association like Kendall tau and Spearman rho (Joe 1990). Constrained majorization orderings lead to measures of divergence from independence like relative entropy measures and uncertainty measures (Joe 1989).

Example 2 (random utility models for choice in mathematical psychology). The setting in random utility choice models is that there is a superset of d items that can be compared in various ways. Suppose the items are assumed to be indexed as $1,\ldots, d$ and $S = \{1,\ldots, d\}$ denotes the index set. Associated with item i is a utility random variable U_i, $i = 1,\ldots, d$; $U_1,\ldots, U_d$ can be independent or dependent. If the U_i are independent, then stochastic ordering results on the U_i lead to preference orderings of various types. The following results are straightforward to prove when the U_i are independent, $U_i \sim F_i$, and there is a zero probability of ties among the U_i.

(a) (i) If $F_k \prec^{st} F_j$, then $p_{jk} = \Pr(U_j > U_k) \geqslant \frac{1}{2}$ or item j is pairwise preferred to item k. (ii) If $F_k \prec^{st} F_j \prec^{st} F_i$, then $p_{ij}, p_{jk} \geqslant \frac{1}{2}$ and $p_{ik} \geqslant \max\{p_{ij}, p_{jk}\}$, or items i, j, k are ordered by strong stochastic transitivity.

(b) If $F_k \prec^{st} F_j$, then $p_{j,S} \geqslant p_{k,S}$ for S such that j, $k \in S$, where $p_{i,S} = \Pr(U_i > U_\ell, \ell \in S\backslash\{i\})$ is a choice probability of the set S, a subset of $\{1,\ldots, d\}$ with cardinality of at least 2. This means the stochastic order implies a preference ordering according to choice probabilities.

(c) Let $r(a_1,\ldots, a_d) = \Pr(U_{a_1} > \cdots > U_{a_d})$, where $(a_1,\ldots, a_d)$ is a permutation of $(1,\ldots, d)$. Let τ_{jk} be a transposition operator so that $\tau_{jk}(a_1,\ldots, a_d)$ is the vector with a_j, a_k transposed. If $F_d \prec^{LR} \cdots \prec^{LR} F_1$, then the items are preference ordered in the strong sense of $r(1,\ldots, d) \geqslant r(\tau_{jk}(1,\ldots, d))$ for all $1 \leqslant j < k \leqslant d$. If the weaker condition $U_k \prec^{HR} U_1$ $(2 \leqslant k \leqslant d)$ is assumed, the strongest conclusion is $r(1,\ldots, d) \geqslant r(\tau_{1k}(1,\ldots, d))$, and if the weaker condition $U_d \prec^{RHR} U_j$ $(1 \leqslant j \leqslant d)$ is assumed, the strongest conclusion is $r(1,\ldots, d) \geqslant r(\tau_{jd}(1,\ldots, d))$.

These results contrast the strength of conclusions that are possible with various stochastic orders of largeness. Generally no preference ordering results follow from stochastic ordering assumptions if the U_i are dependent.

Example 3 (comparison of risks for portfolios). Let X, Y be two continuous random variables representing the returns of two portfolios A, B. B is a better portfolio than A if $E(X) \leqslant E(Y)$ and Y is less dispersed than X in some sense. If the ordering of dispersion is taken from the stochastic/majorization ordering in Eqn. (1), as

$$E(\psi(Y - E[Y])) \leqslant E(\psi(X - E[X])) \tag{3}$$

for all convex continuous functions ψ, then $E[u(X)] \leqslant E[u(Y)]$ for all increasing concave (utility) functions u, and hence this comparison of risk agrees with utility theory with risk averse utilities. See Mosler (1983) for related results.

The examples presented here give an idea of how the orderings have been applied, and the reader can consult the references for many other interesting applications.

See also: Order Statistics; Ordered Relational Structures; Partial Orders; Ranking Models, Mathematics of; Unfolding and Vector Models

Bibliography

Arnold B C 1987 *Majorization and the Lorenz Order*. Springer, New York

Foster J E, Shorrocks A F 1988 Poverty orderings and welfare dominance. *Social Choice Welfare* **5**: 179–98

Joe H 1989 Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association* **84**: 157–64

Joe H 1990 Multivariate concordance. *Journal of Multivariate Analysis* **35**: 12–30

Joe H 1993 Generalized majorization orderings and applications. In: Shaked M, Tong Y (eds.) *Stochastic Inequalities*. IMS Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward, CA, Vol. 22, pp. 145–58

Joe H 1997 *Multivariate Models and Dependence Concepts*. Chapman & Hall, London

Marshall A W, Olkin I 1979 *Inequalities: Theory of Majorization and its Applications*. Academic Press, New York

Mosler K 1983 Increasing multivariate risk: Some definitions. In: Beckmann M, Krelle W (eds.) *Risk and Capital*. Lecture Notes in Economics and Mathematical Systems, 227. Springer-Verlag, Berlin, pp. 88–102

Mosler K, Scarsini M (eds.) 1993a *Stochastic Orders and Decision Under Risk*. IMS Lecture Notes Monograph Series, 19. Institute of Mathematical Statistics, Hayward, CA

Mosler K, Scarsini M 1993b *Stochastic Orders and Applications. A Classified Bibliography*. Lecture Notes in Economics and Mathematical Systems, 401. Springer-Verlag, Berlin

Scarsini M, Shaked M 1996 Positive dependence orders: A survey. *Athens Conference on Applied Probability and Time Series Analysis*. Lecture Notes in Statistics, 114, Springer, New York, Vol. I, pp. 70–91.

Shaked M, Shanthikumar J G (eds.) 1994 *Stochastic Orders and Their Applications*. Academic Press, New York

Stoyan D 1983 *Comparison Methods for Queues and Other Stochastic Models*. Wiley, New York

H. Joe

Male Dominance

The term male dominance evolved in the twentieth century as a conceptual label to characterize the unequal power relations between men as a group and women as a group. This categorical approach to gender relations, which constructs contrastive social classes defined by biological sex in relations of power, is part of a long history of thought regarding the political relations of the sexes beginning with the early Greeks. A correlative of this system of thought has been the relative exclusion of Western women from the public sphere of economic, occupational, and political opportunities compared with their male peers and a tendency to value traits associated with masculinity over those defined as feminine.

The intellectual history of the theory and practice of male dominance is summarized under four major headings, indicating different arguments and time periods. A common thread uniting these arguments is the belief that the biological differences of male and females make male social dominance natural and hence socially desirable. These arguments are reviewed below along with a brief summary of dissenting views.

1. The Misogyny of Early Greeks and the Public/Private Split

Although some scholars look as far back as ancient Mesopotamia for the beginning of the historical process by which male dominance became institutionalized in Western societies (Lerner 1986), the modern arguments for male dominance can be traced to the thought of the early Greeks from Hesiod to Plato and Aristotle. In her treatment of women in Western political thought, Okin (1979) says that from the very beginnings of Greek literature 'strong misogynic strain is obvious.' According to Hesiod, for example, the first woman brought evil and misfortune to the world and the degeneration of the human race.

From Hesiod to Aristotle, the highest words of praise were for males who fulfilled the role of the Homeric aristocratic man. Bravery, skill in warfare, and wealth were manly virtues that no woman could attain. Women were confined to the domestic domain where they were to be quiet, virtuous, beautiful, skilled in weaving and household accomplishments, and above all, faithful to their husbands (Okin 1979).

Plato and Aristotle claimed that 'natural differences' constituted the foundation for segregated male–female social domains, but used this fact to make very different arguments. Plato sought ways to reduce natural differences so that males and females could be equal while Aristotle used natural differences as an argument for the inferiority of women. Plato's utopian view of sexual equality in the *Republic* exposes the central assumption underlying the Western definition of equality: for X to be equal to Y, X must be the same as Y. Plato argues that women in his Utopian city can only be in a position of equality with men if they became like men. To establish equality, he claimed that jobs must be assigned on the basis of ability not sex. Unlike his contemporaries, Plato believed that there was no difference between the sexes apart from their roles in procreation and physical differences in strength. He felt that women were the

wasted sex because their creative and social energies were not sufficiently deployed for the purpose of society and the common good (Okin 1979).

Given the sex role expectations and beliefs of his time, Plato's view was revolutionary with implications for sexual equality that have yet to be realized. Aristotle's approach on the other hand was in keeping with his time. Aristotle accepted the world as it was and developed a philosophical framework to explain it. He also focused on 'natural differences' but departed from Plato in asserting that because women are 'naturally' inferior to men, they are therefore 'naturally' ruled by them (Okin 1979). Aristotle expanded the concept of 'natural' to apply not just to biological sex differences but to sex-linked social traits acquired because of the sexual division of labor due to biological differences. By this schema women's social fate was doubly determined first by her procreative function and second by the assumption that female sex roles follow 'logically from the single fact that women bear children' (Okin 1979).

2. *Arguments over Patriarchal Rule vs. the Social Contract in the Seventeenth and Eighteenth Centuries*

The term 'patriarchy' is very old, first applied to the male leaders of the tribes of Israel whose power was based on kinship not 'contract.' Pateman (1988) says that this changed in response to the controversy that raged in seventeenth-century England about the legitimate source of power in society and how power relations were to be regulated and maintained. Pateman (1988) divides this discussion into two camps: the patriarchalists and the social contract theorists. The patriarchalist approach was represented by the then widely influential book of Sir Robert Filmer, *Patriarcha*. Filmer broke with the biblical tradition associating patriarchy with paternal power by arguing that paternal and political power were 'not merely analogous but *identical*' (Pateman 1988). Filmer argued for absolute monarchy on the patriarchalist grounds that 'kings were fathers and fathers were kings' (ibid). Pateman labels Filmer's view 'classic patriarchalism' as opposed to traditional patriarchy which she defines as paternal rule of the family. Filmer's addition was to make the procreative power of the father in the family the origin of political right in society (Pateman 1988).

Filmer's theory was short lived due to the success of counter arguments proposed by John Locke and Thomas Hobbes. These philosophers argued that all men are 'naturally' free and political right can only be imposed by contract not by patriarchal fiat. Hobbes and Locke separated paternal from political power and claimed that 'contract was the genesis of political right' (Pateman 1988). However, they did not include women in their notions of contract and political right. Women continued to be subordinated to men as fathers and husbands. Hobbes conceived of the family as patriarchal 'wherein the Father or Master is the Sovereign,' and Locke concluded that there is 'a Foundation in Nature' for the legal subjection of women to their husbands (Okin 1979 quoting Hobbes and Locke).

The arguments for the subjugation of women articulated by the social contract theorists leads Pateman to distinguish a third type of patriarchal argument, 'fraternal patriarchy.' According to this argument, although the father no longer figures as the primary power figure women are still not the equal of men. Men now rule the public domain not as fathers but as men. In the family context, women are subordinated first to their fathers and then to their husbands. In the context of public life they are subordinated to men (Pateman 1988). Pateman (1988) calls this 'modern patriarchy' because it is based on the rule of law and 'structures capitalist civil society.'

3. *Darwinian Thought Reflected in Nineteenth-century Evolutionary Models of Cultural Stages*

Another source supporting the argument of fraternal patriarchy or male dominance came from Darwin's evolutionary theory of *The Origin of the Species* published in 1859 and his theory of sexual selection published in *The Descent of Man* in 1871. Darwin's evolutionary theory legitimized nineteenth-century notions of progressive development from the primitive to the complex and encouraged organic analogies leading to reductionist explanations in which 'natural tendencies' were elaborated with terms like 'survival of the fittest' and 'reproductive success.'

In *The Descent of Man* Darwin described the mechanism of sexual selection to explain the evolution of human sex differences. Like natural selection in evolution, Darwin's notion of sexual selection was based on the idea of strife and competition for mates. Building his argument by analogy with animals, Darwin claimed that as there could be no dispute that bulls differ in disposition from cows, human males also differ from females. 'There can be do doubt,' he said 'that the greater size and strength of man, in comparison with women…are due in chief part to inheritance from his half-human male ancestors.' Darwin said that such characteristics 'would have been preserved or even augmented during the long ages of man's savagery, by the success of the strongest and boldest men, both in the general struggle for life, and in their contest for wives.' He measured a man's success by the number of children they left as compared with their 'less favoured brethen' (Darwin 1936). Sanday (1996) argues that the Darwinian concept of reproductive success and sexual competition became

the basis for popular and psychoanalytic views regarding the naturalness of male sexual aggression making it incumbent on women to protect themselves from rape and sexual coercion (e.g., see also *Rape and Sexual Coercion*).

In the second half of the nineteenth century the evolutionary paradigm was applied to notions of cultural progress. Patriarchy was no longer seen as the universal model for human affairs for to do so would be to equate patriarchy with the primitive, 'savage,' stage of human culture. A number of scholars posited cultural stages in which patriarchy was presented as the epitome of social order as opposed to matriarchy which was associated with less advanced social forms.

The most important book sparking interest in ancient matriarchies was the publication in 1861 of Bachofen's *Das Mutterrecht*. In this book Bachofen is content to give women a high position but only at a low level of evolutionary development. Bachofen introduces the term 'mother-right,' which he defines in terms of female rule (gynecocracy) and customs giving mothers control of their children. Thus, in mother-right societies, naming followed the mother's line; the child's status was determined by the mother's; all property rights were passed from mother to daughters and not to sons; and the 'government of the family' fell not to the father but to the mother. 'By a consequent extension of this last principle,' Bachofen said, 'the government of the state was also entrusted to the women (Bachofen 1967). According to Bachofen, mother-right societies constitute the second stage of evolution after 'primitive promiscuity' and before 'the victorious development of the paternal system.' He associates mother-right with 'the pre-Hellenic peoples' and 'patriarchal forms' with Greek culture (1967). His preference for one over the other is seen in his conclusion that paternity represents a 'triumph' because it 'brings with it the liberation of the spirit from the manifestations of nature, a sublimation of human existence over the laws of material life' (Bachofen 1967).

Although Bachofen never used the term matriarchy, preferring instead the terms gynecocracy or mother-right, his definition of mother-right in terms of female rule established the subsequent use of the term matriarchy to mean the mirror image of patriarchy. The anthropological concern with the issue of matriarchy vs. patriarchy, or female vs. male rule, begins with a debate sparked by Henry Maine's 1861 publication of *Ancient Law* and John McLennan's 1865 publication of *Primitive Marriage*. While Maine argued that the patriarchal family and paternal authority was the primordial family form, like Bachofen McLennan argued that matrilineal descent and the maternal family came first, a relic of a primitive state when there was no fixed marriage, so that paternity was not recognized as a social tie and maternity furnished the only relation on which kinship and the family could be founded. A similar argument appears in the better known 1877 book, *Ancient Society*, by Lewis Henry Morgan, one of the founding fathers of anthropology. The subtitle of this book betrays the common framework of the times: *Researches in the Lines of Human Progress from Savagery through Barbarism to Civilization*.

In an 1896 article published in the monthly review, *The Nineteenth Century*, another of the founding fathers of anthropology, E. B. Tylor, comments favorably on McLennan's contribution and introduces the notion of 'the matriarchal family system.' Tylor amasses an impressive array of ethnographic facts to support his contention that there were many communities in which women enjoyed 'greater consideration than in barbaric patriarchal life.' However, half way through the article Tylor rejects the term 'matriarchal' for taking 'too much for granted that women govern the family.' In its place he substitutes the term 'maternal family' on the grounds that the actual power is always in the hands of brothers and uncles on the mother's side. Tylor reaches this conclusion despite the considerable evidence he presents for female familial and social authority. For example, he provides a striking nineteenth-century description of the authoritative role of the senior woman in the matrilineal long house of the Minangkabau of West Sumatra, to this day known to anthropologists as the largest and most stable matrilineal society in the world (see discussion below). Thus, although Tylor comes close to rejecting the male dominance thesis on empirical grounds his conclusion falls into line with the historical tendency to assert its universality.

4. *The Anthropological Debate Regarding Universal Male Dominance in the Twentieth Century*

Tylor's ambivalent position with respect to the issue of male authority marked a great deal of anthropological discussion in the twentieth century. Much of this discussion revolved around the issue of who holds ultimate authority in matrilineal societies. The early twentieth century saw the demise of the term matriarchy, a victim both of the tendency to confine its meaning to exclusive female rule and the exhaustion of the evolutionary paradigm. However, one anthropologist, Rivers (1924), cautioned that abandonment of the assumption of primitive matriarchy should not result in reverting to Maine's doctrine of the priority of father-right, which Rivers noted was still prevalent 'in writings on the history of political institutions.' According to Rivers, Maine's theory of primordial patriarchy was 'even more untenable' than pronouncements concerning the priority of mother-right. Yet, subsequent scholarship did just what Rivers predicted and reverted to an updated version of Maine's argument. In 1961, the topic of matriarchy was revisited by anthropologist David Schneider in

Matrilineal Kinship, the book he edited with Kathleen Gough. This book was published 100 years after the publication of Bachofen's *Das Mutterrecht*. In his theoretical introduction, Schneider (1961) rejects Bachofen's claim of 'generalized authority of women over men,' in matrilineal systems. Schneider (1961) claims that in all unilineal groups, be they patrilineal or matrilineal, the role of women as women is defined in terms of her child-rearing responsibilities and the role of men is defined 'as that of having authority over women and children.' According to Schneider, positions of highest authority within the matrilineal descent group will, therefore, ordinarily be vested in statuses occupied by men.'

In the early 1970s the notion of primordial matriarchies at the dawn of human history was revisited by feminist activist theorists outside of anthropology. By way of response, anthropologists argued for universal male dominance. In their widely influential edited volume, Rosaldo and Lamphere (1974) stated that evolutionary theories positing an earlier stage of human development' in which 'the social world was organized by a principle called matriarchy, in which women had power over men' could be dismissed on both archaeological and ethnographic grounds. Going one step further, they made their famous (but later retracted) statement: 'It seems fair to say then, that all contemporary societies are to some extent male-dominated, and although the degree and expression of female subordination vary greatly, sexual asymmetry is presently a universal fact of human and social life' (Rosaldo and Lamphere 1974; but see also Rosaldo 1980 and Lamphere 1995).

5. *Dissenting Views*

Several anthropologists took exception to the claim of universal male dominance. Annette Weiner's (1976) ethnography of Trobriand mortuary ceremonies demonstrates that women in kinship-based societies do not live in the same world as women in late-capitalist societies, a lesson Eleanor Leacock's (1981) work also teaches. Weiner pointed out that in societies like the Trobriand's the political does not stand alone in an isolated sphere at the apex of social power but functions in conjunction with the cosmological. Even if men dominate the political, their activities are controlled by the larger sociocultural system in which both men and women play defining roles.

In a cross-cultural study of female power and male dominance, Sanday (1981) also takes exception to the conclusion of universal male dominance, demonstrating that there are many societies in which women participate in the economic and political domains of public life. Departing from the standard argument hinging dominance on biological differences and male political activities, Sanday utilizes a large sample of band and tribal societies to investigate the empirical correlates of male and female power. She finds that male dominance is associated with misogyny, male segregation, combativeness and competition leading to significant levels of interpersonal violence, an ideology of male toughness, masculine-defined religious symbolism, and environmental stress. Female power, on the other hand, is associated with respect for the public roles of women, female religious symbolism, reverence rather than exploitation of nature, nurturing paternal roles, an ethic of respect for the individual and social cooperation in decision making.

By the end of the twentieth century, backed by long-term ethnographic research (see Lepowsky 1993 and articles in Sanday and Goodenough 1990) and new theoretical approaches (see Ortner 1990, 1996), the male dominance approach was largely abandoned by cultural anthropologists. Ethnographies of societies like the matrilineal Minangkabau of West Sumatra demonstrated a model of power unlike any posited for the so-called male dominated societies. Summarizing many Minangkabau studies, Nancy Tanner and Lynn Thomas (1985) note that in Minangkabau society there is 'far more centrality and authority for women, specifically as mother and senior women,' than previous models admit such as the one proposed by Schneider (see above).

In her long-term study of the Minangkabau, Sanday (2002) reaches the same conclusion. She finds that men and women share many of the same functions, such as 'childrearing, agricultural labor, and decision making' much in the manner that Plato dreamed of but with important differences. Whether functions assigned to the sexes overlap or differ, they serve the same ends—nurturing and care for lineage members. Such values are ideals expected of all cultivated, cultured Minangkabau individuals, not just women. The public nature of these ideals ensure the well-being of the elderly and sick and the economic security of the mother–child unit independently of the father's support. Among the largest ethnic groups in Indonesia and known there for their commitment to democracy and equality, the Minangkabau proudly refer to themselves as a modern 'matriarchate' and distinguish their system from others which take the patrilineal approach. Their social ideology reflects on the social costs of male dominance and the benefits of incorporating family values and women in the public domain of status of privilege.

In sum, although the ethnographic record demonstrates that male dominance subordinates women in the public arena of political discourse and economic opportunity in some societies, this is not a universal condition. To claim that it is reifies male dominance and displays a teleological approach to human social affairs.

See also: Feminist Movements; Feminist Theory; Gender and Feminist Studies; Gender and Feminist

Studies in Anthropology; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Sociology; Gender, Feminism, and Sexuality in Archaeological Studies; Gender History; Gender Ideology: Cross-cultural Aspects; Masculinities and Femininities; Matrifocality; Rape and Sexual Coercion; Social Movements and Gender

Bibliography

Bachofen J J 1967 *Myth, Religion, & Mother Right: Selected Writings of J J Bachofen*. Princeton University Press, Princeton, NJ

Darwin C 1936 *The Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life and the Descent of Man and Selection in Relation to Sex*. The Modern Library, New York

Lamphere L 1995 Feminist anthropology: The legacy of Elsie Clews Parsons. In: Behar R, Gordon D A (eds.) *Women Writing Culture*. University of California Press, Berkeley, CA

Leacock E B 1981 *Myths of Male Dominance*. Monthly Review Press, New York

Lepowsky M 1993 *Fruit of the Motherland: Gender in an Egalitarian Society*. Columbia University Press, New York

Lerner G 1986 *The Creation of Patriarchy*. Oxford University Press, New York

Okin S M 1979 *Women in Western Political Thought*. Princeton University Press, Princeton, NJ

Ortner S B 1990 Gender hegemonies. *Cultural Critique* **14**: 35–80

Ortner S B 1996 *Making Gender: The Politics and Erotics of Gender*. Beacon Press, Boston

Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA

Rivers W H R 1924 *Social Organization*. Knopf, New York

Rosaldo M Z 1980 The use and abuse of anthropology: Reflections on feminism and cross-cultural understanding. *Signs* **5**(3): 389–417

Rosaldo M Z, Lamphere L 1974 *Woman, Culture and Society*. Stanford University Press, Stanford, CA

Sanday P R 1981 *Female Power and Male Dominance: On the Origins of Sexual Inequality*. Cambridge University Press, Cambridge, UK

Sanday P R 1996 *A Woman Scorned: Acquaintance Rape on Trial*, 1st edn. Doubleday, New York

Sanday P R 2002 *Women at the Center: Journey into the Heart of Matriarchy*. Cornell University Press, Ithaca, NY

Sanday P R, Goodenough R G (eds.) 1990 *Beyond the Second Sex*. University of Pennsylvania Press, Philadelphia, PA

Schneider D M 1961 Preface, introduction. In: Schneder D M, Gough K (eds.) *Matrilineal Kinship*. University of California Press, Berkeley, CA

Tanner N, Thomas L L 1985 Rethinking matriliny: Decision-making and sex roles in Minangkabau. In: Thomas L, von Benda-Beckmann F (eds.) *Change and Continuity in Minangkabau: Local, Regional, and Historical Perspectives on West Sumatra*. Ohio University Press, Athens, OH pp. 45–71

Weiner A B 1976 *Women of Value, Men of Renown: New Perspectives on Trobriand Exchange*. University of Texas Press, Austin, TX

P. R. Sanday

Malinowski, Bronislaw (1884–1942)

A Polish-born anthropologist, Malinowski had a profound influence on the development of British social anthropology between the wars. Under the banner of Functionalism he revolutionized fieldwork methods, contributed to a wide range of academic and public issues, and trained a cosmopolitan cohort of pupils which spread his teachings throughout the world. A passionate, volatile and sometimes abrasive personality, Malinowski provoked controversy throughout his career; but his intellectual integrity was absolute. He promoted the humanitarian uses of anthropology with entrepreneurial flair, attracting Rockefeller funding to his cause. In addition to his ancestral status as a founder of a cohesive academic discipline, Malinowski's historical legacy in Britain and the Commonwealth was an expanded institutional base for its teaching and practice. In Ernest Gellner's ennobling epithet, Malinowski was 'Anthropologist Laureate of the British Empire.'

1. A Polish Anglophile

Bronislaw Malinowski was born on April 7, 1884, of noble (*szlachta*) stock in Cracow, capital of the semi-autonomous province of Galicia in the Austro-Hungarian Empire. As a child he lived for several years among local peasantry and, with his mother, took lengthy trips abroad for the sake of his delicate health. To these early experiences of cultural difference Malinowski attributed his vocation to anthropology. He was also inspired by reading James Frazer's *The Golden Bough*.

During Malinowski's youth, Cracow was a vibrant intellectual and artistic center, home of the Modernist movement called Young Poland. His most important early friendship was with the surrealist artist, novelist and dramatist, Stanisław Ignacy Witkiewicz. Unable to match his friend's artistic genius, Malinowski resolved to excel in the sciences. At the Jagiellonian University he studied mathematics, physics, and philosophy. His brilliant doctoral thesis, 'On the Principle of Economy of Thought' (1906), was a critique of the positivist epistemology of Ernst Mach; it contained the seeds of Malinowski's later thinking about functionalism (Thornton and Skalnik 1993).

In 1908–09, during three semesters at the University of Leipzig, Malinowski turned to the study of *Völkerpsychologie* under Wilhelm Wundt and economic history under Karl Buecher. Possessed by 'a highly developed Anglomania,' in 1910 Malinowski crossed the English Channel to enrol as a graduate student in ethnology at the London School of Economics (LSE) in the University of London. The crossing was his Rubicon, and although not granted British nationality until 1931, he was to spend his most productive years in England.

Under the supervision of Edward Westermarck, Malinowski completed his first English book, *The Family Among the Australian Aborigines* (1913), an evaluation of Aboriginal ethnography and a rigorous critique of evolutionist theories. He presented a signed copy to his Polish compatriot, Joseph Conrad, whom he deeply admired.

2. *The Cult of Fieldwork*

The dominant paradigm of late-nineteenth century anthropology (or ethnology) was evolutionism, based on the premise that mankind had passed through universal stages of development on the journey from savagery to civilization. A division of labor existed in imperial Britain between armchair theorists (like Frazer) and amateur ethnographers (explorers, missionaries, and colonial officers) who reported from outposts of the Empire.

The professionalization of anthropology began following the Cambridge University Expedition to the Torres Strait of 1898–99, led by the zoologist-turned-anthropologist A. C. Haddon. By 1910 anthropology was being taught at Cambridge, Oxford, and London. It was due to Haddon and his disciples W. H. R. Rivers and C. G. Seligman that Malinowski found an almost messianic cult of fieldwork in Britain. In Seligman's words, 'field research in anthropology is what the blood of martyrs is to the Church.' By the time Malinowski embarked on his own fieldwork, at least a dozen students of Haddon and his colleagues had preceded him.

World War One broke out just as Malinowski arrived in Australia. Although cut off from his English funds and under suspicion as an 'enemy alien,' he was permitted entry to Papua (ex-British New Guinea) to conduct his research. His apprentice fieldwork was among the Mailu of the south coast, whence Seligman had directed him to fill a gap left by his own ethnographic survey of Papua a decade earlier. After five busy months, Malinowski returned to Australia where he dashed off *The Natives of Mailu*, a conventional ethnographic report (Young 1988). In June 1915 he returned to eastern Papua with Australian funding for a second spell of fieldwork. Supposedly en route to another field site, Malinowski visited Kiriwina, the largest of the Trobriand Islands, about which Seligman had already reported comprehensively in *The Melanesians of British New Guinea* (1910). For various reasons Malinowski's visit to Kiriwina evolved into a sojourn of 10 consecutive months, followed by a second period of 10 months in 1917–18. Although his movements were restricted by the colonial authorities, Malinowski was not, as his popular legend had it, 'interned' in the Trobriands.

He had learned in Mailu that the richness of his data was a function of the time he had spent living in the village, shunning the company of fellow Europeans and observing everyday life at first hand. It was also imperative to master the local language well enough to have unmediated access to one's subjects. Under the rubric 'the intensive study of limited areas,' such precepts of method had been advocated by Haddon and Rivers during the previous decade, but they had not been put into proper practice until Malinowski's work in Kiriwina.

Thus, the innovative breakthrough his fieldwork effected was based on the simple expedient of living in the same location for many months and probing for information in the vernacular. This method of 'total immersion' encouraged a more intimately interactive style of ethnography, permitting an empathetic understanding which went beyond mere recording. The fieldworker could begin to know another culture from within: 'the final goal' of the ethnographer is 'to grasp the native's point of view… to realize *his* vision of *his* world' (Malinowski 1922, p. 25). 'Participant observation' was to become the catch-cry of his pupils' generation. It still remains the hallmark of anthropological fieldwork. Professionally, it was prolonged and intensive fieldwork, with its loneliness and hardship, that stamped the new breed of anthropologist (Stocking 1995, Chap. 6).

3. *Trobriand Man as Exemplar*

Malinowski was criticized for making his Trobrianders the measure of mankind, but he did so persuasively as Trobriand Man was introduced into the controversies of the day. On close inspection, this exemplar of the Savage was not so very different from Civilized Man. He was demonstrably rational and technologically competent, morally and aesthetically sensitive: an equal, therefore, and not (as imperialists and evolutionists would have it) a benighted inferior. He believed in magic, certainly, but he used it only in situations where technology could not guarantee an outcome. 'Magic helps those who help themselves.' Magic was a tool to which men resorted when the desired end of any enterprise was in doubt. This instrumental, pragmatic view of magic was typical of Malinowski's functionalism.

In 1922 Malinowski published his first Trobriand classic, *Argonauts of the Western Pacific*, which set British anthropology firmly on a new course (Malinowski 1922). It was a richly textured study of *kula*, the circuitous ceremonial exchange of shell valuables which linked the Trobriands to neighboring archipelagoes. Malinowski's combination of ethnographic observation, methodological prescription, and literary romanticism was vivid and compelling. An enthusiastic preface by Sir James Frazer ensured *Argonauts*' successful launch.

With *Crime and Custom in Savage Society* (1926), a little book on 'the seamy side' of Trobriand law, Malinowski pioneered the study of political and legal

anthropology. The following year, *Sex and Repression in Savage Society* set a cat among the Freudian pigeons by contesting the claim that the Oedipal Complex was universal (Malinowski 1927). Malinowski argued that in the matrilineal Trobriands it was the mother's brother, not the father, who was the hated authority figure; likewise, it was the sister, not the mother, who was the main object of a boy's incestuous desire. The debate continues.

The Sexual Life of Savages (Malinowski 1929) was a detailed study of Trobriand courtship, marriage, and domestic life; it was also a covert attack on sexual prudery and hypocrisy, and in the changing mores of the period it had a liberating appeal.

Malinowski's last Trobriand ethnography, *Coral Gardens and their Magic* (1935), dealt exhaustively with horticultural practices and their ritual accompaniments, with land tenure and its politics, and with the language of magic. The two-volume work was a triumph of thick description and a baroque monument to its author's functional method.

Malinowski fancied that his monographs did for anthropology what Joseph Conrad's novels had done for English literature. It is true that through these and many other writings, Malinowski's Trobrianders came to be known as one of the best-documented 'primitive' peoples in the world. Subsequent generations of anthropologists have reinterpreted his material in the light of different theories—the most flattering compliment posterity can pay to any ethnographer.

4. *Functionalism*

The paradigmatic theory that informed Malinowski's Trobriand monographs was functionalism. As an heuristic method which guided fieldwork, functionalism was an appeal for contextualized empirical data purified of historical speculations. Malinowski urged ethnographers to look for the use of a custom (or institution) in the present, so as to determine the part it played in the integral working of a society or culture. This would give the clue to its meaning and the explanation for its existence, one that (in contrast to the speculative explanations of 'antiquarian' anthropologists) was concretely based on empirical observation. Although functionalism was demonstrably deficient when applied to societies undergoing rapid social change, as a 'clinical' method it was immensely influential during the interwar years. The promise it offered 'practical' or applied anthropology in African colonies enabled Malinowski to secure, through the sponsorship of the International African Institute, Rockefeller funding for his students' fieldwork.

His theory of language was based on the premise that function determined form. 'Primitive' language was 'a mode of action' rather than a 'countersign of thought.' Much speech is 'phatic' (agreeable social noises such as 'How are you?') or instrumentally pragmatic, meaningful only in its 'context of situation.' Malinowski foreshadowed Ludwig Wittgenstein's axiom 'do not ask the meaning but the use.'

As grand theory, Malinowski's functional view of culture as an apparatus for the satisfaction of human needs was born out of Machian empiricism, confirmed by Jamesian pragmatism, and blessed by the Durkheimian tenet that culture was *sui generis*. Malinowski's 'scientific theory of culture' (1944a) was built on the premise that man's biological heritage provided the ground plan for all cultures; basic (or primary) needs such as reproduction and derived (or secondary) needs such as education yielded similar cultural solutions, the universal form of which was the family. In a nutshell, he argued (a) that function determines the form of cultural entities; (b) that the essential element of the science of culture is the institution; and (c) that institutions must ultimately be defined in relation to the biological needs they satisfy. As formal theory, then, functionalism explained culture in an instrumental and reductionist fashion. The schematic, utilitarian form in which he presented this theory gained few adherents and it was soon forgotten.

5. *Teacher, Pundit, and Propagandist*

Malinowski's rise to eminence at the LSE had been rapid: from Lecturer in 1923, Reader in 1924, to foundation Professor of Social Anthropology in 1927. The mode of teaching he favored was the seminar. In Socratic fashion, he deployed wit, erudition, and 'intellectual shock tactics' to provoke thoughtful argument. His electrifying seminar became a legend during the 1930s, engaging his brightest pupils and attracting from every continent students of cognate disciplines. The multiethnic roll-call of his pupils would fill a page, but among the most eminent were Raymond Firth, Edward Evans-Pritchard, Ashley Montagu, Isaac Schapera, Audrey Richards, Camilla Wedgwood, Meyer Fortes, Paul Robeson, Siegfried Nadel, Lucy Mair, Reo Fortune, Ralph Bunche, Ian Hogbin, Jomo Kenyatta, Talcott Parsons, Gregory Bateson, Edmund Leach, and Phyllis Kaberry.

Despite his fragile health, Malinowski was a ferociously hard worker who mustered others to his aid. He employed servants at home, dictated most of his writings to stenographers, and employed pupils as long-suffering research assistants. Until she was tragically incapacitated by multiple sclerosis, Elsie, his Australian-born wife, served him as a loving critic and secretary. Together with their three daughters they spent summers at the family villa in the South Tyrol, where pupils often visited to walk and talk (Wayne 1995).

In tandem with his academic activities, Malinowski performed the role of public intellectual, disseminating progressive views on marriage, the family, and sex education. He lectured missionaries and colonial

officers; he gave BBC talks on religion; he actively supported the British Social Hygiene Council and Mass Observation. More importantly, he served on the board of the International African Institute for which he devised a field training program. In 1934 he visited several of his pupils in South and East Africa before conducting some fieldwork of his own in Tanganyika.

Malinowski had lectured widely in the United States in 1926, 1933, and 1936, and he returned for a sabbatical year in October 1938. When war broke out in Europe he was advised to remain in America for the duration. He taught at various universities, anxious to secure a permanent position. As 'war work' he embarked on a lecturing and writing campaign of propaganda (his own term) to convince Americans of the need to fight Hitler. Of all the honors he received he valued most the distinction the Nazis conferred upon him by burning his books. His opposition to totalitarianism was categorical and he wrote passionately against it in *Freedom and Civilization* (Malinowski 1944b).

During the summers of 1940 and 1941 Malinowski conducted fieldwork in the Oaxaca Valley of Mexico with his new wife and a Mexican assistant. This fieldwork did not match the exacting standards he had set in the Trobriands. Just before his sudden death of a heart attack on May 16, 1942 in New Haven, CT, Malinowski was appointed to a permanent professorship at Yale University and elected President of the Polish Institute of Arts and Sciences in New York.

6. *Malinowski's Legacy*

Malinowski bequeathed a controversial legacy. He did more than anyone else of his time to establish a cohesive (even 'clannish') school of anthropology with a unique academic profile. But the year he left for America was the year A. R. Radcliffe-Brown returned to England to create a rival school at Oxford. The latter's more Durkheimian and sociologically austere version of functionalism (known as structural functionalism) attracted many of Malinowski's erstwhile disciples. Radcliffe-Brown dominated British social anthropology during the following decade.

In 1957, a dozen of Malinowski's pupils published a Festschrift which critically evaluated his contribution to different subfields. While conceding that he was an incomparable ethnographer, the majority verdict was that Malinowski was a flawed and inconsistent thinker who had failed to grasp the analytical priority of 'systems' and 'structures.' It thus became a cliche of disciplinary folk history that Malinowski was a brilliant ethnographer but a mediocre theoretician.

This view justified neglect of his work until the late 1960s, when, abetted by Raymond Firth, Malinowski's successor at LSE, the pendulum swung from a theoretical preoccupation with structure to the analysis of process and event. Malinowskian 'methodo logical individualism' became respectable again through decision-making analysis and transactional anthropology. Meanwhile there was a rapprochement between British social and American cultural anthropology as their research agendas converged. American anthropologists had always been receptive to Malinowski's field methods, and eventually the kind of fieldwork they conducted became indistinguishable from that of their British and Commonwealth colleagues.

In 1967, Malinowski's widow published his personal field diaries. They were an embarrassment to many of his pupils and Raymond Firth's preface to the book is an exercise in damage control. Malinowski's agonized musings on the alienation he experienced in the field put the lie to the effortless rapport he was believed to have enjoyed with Trobrianders. His diaries exacerbated, if not precipitated, a moral and epistemological crisis of anthropology during the 1970s, when global decolonization was turning attention to the exploitative aspects of the discipline and its dubious historical role as a handmaid of Imperialism.

The occasion of Malinowski's centenary celebrations in 1984 produced another round of evaluations, the most notable being a collection of essays by Polish scholars. They not only rehabilitated Malinowski's reputation in his homeland, but painstakingly examined his theories in the light of his Polish roots. He was declared a scientific humanist and rationalist, a romantic positivist and political liberal who belonged in the mainstream of modernist European thought (Ellen et al. 1988).

In America the diaries stimulated fresh debates on the assumptions underlying Malinowskian fieldwork and ethnographic representation. Postmodernists appropriated Malinowski for their own rhetorical ends in deconstructing ethnographic authority. To an ironic degree, however, he had pre-empted some of their concerns by the humanistic strain in his work and, except when promoting his 'science' for institutional support on behalf of 'practical anthropology,' he was by no means a naive positivist. Nor was he a 'romantic primitivist' as some critics have charged, despite the Frazerian echoes in his earliest monographs. Fully 50 years before postmodernists deconstructed exoticism, Malinowski had come to the view that anthropology romantically 'over-sensationalized' differences between cultures, and he questioned its fascination with 'freakish' customs such as polyandry, widow burning, and headhunting.

Despite the fragmentation of socio-cultural anthropology at the millennium, Malinowski's fruitful concepts can be detected in almost every subfield, and his legacy persists wherever ethnography is practiced. His name is indelibly linked with Melanesia, but it appears too in anthropological writings from countries as far-flung as Mexico and China, Italy and New Zealand, and India and Ireland. A protean, cosmopolitan

ancestral figure, then, Malinowski continues to fascinate his intellectual descendants.

See also: Economic Anthropology; Ethnography; Exchange in Anthropology; Fieldwork in Social and Cultural Anthropology; Functionalism, History of; Kula Ring, Anthropology of; Participant Observation

Bibliography

Ellen R, Gellner E, Kubica G, Mucha J (eds.) 1988 *Malinowski Between Two Worlds: The Polish Roots of an Anthropological Tradition*. Cambridge University Press, Cambridge, UK

Firth R W (ed.) 1957 *Man and Culture: An Evaluation of the Work of Bronislaw Malinowski*. Routledge and Kegan Paul, London

Malinowski B 1922 *Argonauts of the Western Pacific*. Routledge and Kegan Paul, London

Malinowski B 1926 *Crime and Custom in Savage Society*. Routledge and Kegan Paul, London

Malinowski B 1927 *Sex and Repression in Savage Society*. Routledge and Kegan Paul, London

Malinowski B 1929 *The Sexual Life of Savages*. Routledge and Kegan Paul, London

Malinowski B 1935 *Coral Gardens and their Magic* (2 vols.). Allen & Unwin, London

Malinowski B 1944a *A Scientific Theory of Culture*. Chapel Hill, University of North Carolina Press, NC

Malinowski B 1944b *Freedom and Civilization*. Roy, New York

Malinowski B 1967 *A Diary in the Strict Sense of the Term*. Routledge, London

Stocking G W 1995 *After Tylor*. University of Wisconsin Press, Madison, WI

Thornton R J, Skalnik P (eds.) 1993 *The Early Writings of Bronislaw Malinowski*. Cambridge University Press, Cambridge, UK

Wayne H (ed.) 1995 *The Story of a Marriage: The Letters of Bronislaw Malinowski and Elsie Masson* (2 vols.). Routledge, London

Young M W (ed.) 1988 *Malinowski Among the Magi: 'The Natives of Mailu'*. Routledge, London

M. W. Young

Malthus, Thomas Robert (1766–1834)

The passions aroused even today by the name of Malthus bear witness to the eminent position—almost equal to that of Marx—which he still occupies in the history of ideas.

1. An Unspectacular Life

Thomas Robert Malthus was born on February 13, 1766, in 'The Rookery' in Dorking, near Wooton, Surrey, England. His father, Daniel Malthus, was a prosperous, brilliant, and educated man, a convinced progressive who corresponded with Voltaire and particularly with Rousseau, who visited him on March 9 of the same year.

Very much taken with new ideas, Daniel Malthus brought up his son in accordance with the precepts of *Émile*, and put him in the charge of two dissentient intellectuals: Richard Graves (1715–1804), who had to leave Cambridge as a result of a misalliance, and Gilbert Wakefield (1756–1801), a former pastor who taught classical literature at Warrington Academy and was imprisoned in 1799 for writing that the poor would have nothing to lose from a French invasion. To crown this educational career, Malthus, who had been admitted to Jesus College, Cambridge, in 1784, became a pupil of William Frend (1757–1841), who was dismissed from the university in 1793 for publishing three pamphlets against the Church. It was perhaps Frend who aroused Malthus' interest in economics and demography (Petersen 1979).

In 1788 Malthus took holy orders, in 1793 he was elected a fellow of Jesus College and became rector of Oakwood Chapel (Surrey), 12 or 13 km from the village of Albury, where his father Daniel had settled in 1787. Some years later, he wrote a first essay entitled *The Crisis, a View of the Recent Interesting State of Great Britain, by a Friend of the Constitution*. No publisher could be found for this work, however, and only extracts—published after Malthus' death by his friends William Empson and William Otter—survive (Otter 1836). These reveal a very politically correct Malthus, an advocate of the provision of state assistance at home for those not capable of securing their own livelihood.

Two years later Malthus' polemic pamphlet was published, marking a complete break with the progressive ideas of his entourage: *An Essay on the Principle of Population as it Affects the Future Improvement of Society; With Remarks on the Speculations of Mr. Godwin, M. Condorcet and Other Writers*. The essay was published anonymously, but Malthus was recognized immediately as its author. The pamphlet established Malthus' name, but its central thesis, which was expressed with an apparent scientific detachment, that it is useless, even dangerous, to provide relief for the poor as this encourages them to reproduce, created a scandal.

Malthus, described by all those who met him as a gentle, courteous, sincere, and sensitive man, was thus drawn into a heated and interminable discussion, obliging him to expand and elucidate his ideas on population.

After traveling with his Cambridge friends to Scandinavia (1799), France, and Switzerland (1802), and producing a new essay on price increases (*An Investigation of the Cause of the Present High Price of Provisions* 1800), thus demonstrating his talents as an economist, Malthus published a new book in 1803. This is regarded as a second edition of the 1798 *Essay*;

and indeed entire chapters were taken from the first edition, but the 1803 edition differs fundamentally with respect to its complete title (*An Essay on the Principles of Population, or a View of its Past and Present Effects on Human Happiness with an Inquiry into our Prospects Respecting the Future Removal or Mitigation of the Evils which it Occasions*), its (fourfold) length and, above all, its content, now constituting moral and sociological reflections on population growth. Even during the author's lifetime, the new *Essay* was revised four times with additions and corrections.

In April 1804 Malthus, now 38, married his cousin Harriet Eckersall at Claverton, near Bath. Within four years, she had given birth to one son and two daughters (Petersen 1979).

In 1806 he was appointed professor of history and political economy at the East India Company's college, which opened at Hertford before moving to Haileybury (Hertfordshire) several years later. He held this chair in political economy—the first of its kind—until his death, and publicly defended the college when its existence was jeopardized (Petersen 1979).

For 14 years, Malthus produced little more than a variety of pamphlets—particularly on the importance of grain—but with the five successive editions of his second *Essay* he established such a reputation that he was elected a member of the Royal Society (1819), a corresponding member of the Royal Society of Literature (1824), and later an associate member of the *Académie des Sciences Morales et Politiques* and the Royal Academy of Berlin (1833). In 1821 he was one of the founder members of the Political Economy Club and, in the year of his death, of the Royal Statistical Society.

In 1820 Malthus's second *magnum opus* was published: *Principles of Political Economy Considered with a View to Their Practical Application*. Being of a concrete nature, it was opposed to the more systematic and dogmatic views of David Ricardo (Keynes 1939).

After this, there is little to report apart from the publication of an article entitled 'Population' in the *Encyclopaedia Britannica* (1824) and further trips to the continent and to Scotland. Having gone to Bath to spend Christmas with his family, Malthus died of a heart attack on December 29, 1834, at the age of 68. His grave is to be found in Bath Abbey.

2. *Malthus's Contribution to the Sociology and Economy of Populations*

What the general public, nonspecialists and, regrettably, many historians and economists have taken up of Malthus' theory is a caricatural model according to which population always increases in a geometrical progression, while the means of subsistence only increase in an arithmetical progression, thus making it necessary to limit births by all possible means.

In fact, Malthus' theory went through a complex evolutionary process, developing from one work to the next. His central ideas, however, were already present in the 1798 *Essay*.

Originally, this first *Essay* was merely a philosophical pamphlet intended as a criticism of the utopian optimism of Godwin and Condorcet. The author's main aim was to undermine the theory that it was possible for man and society to progress. In the conclusion to chapter VII, he resumes his arguments as follows:

> Must it not then be acknowledged by an attentive examiner of the histories of mankind, that in every age and in every state in which man has existed, or does now exist,
> —That the increase of population is necessarily limited by the means of subsistence.
> —That the population does invariably increase when the means of subsistence increase. And,
> —That the superior power of population is repressed, and the actual population kept equal to the means of subsistence by misery and vice (Malthus 1798).

In the preface, Malthus states that '(t)he following Essay owes its origin to a conversation with a friend (very probably Daniel Malthus, the author's father, who was to die two years later, taking his generous ideas and fabulous illusions with him to the grave) on the subject of Mr. Godwin's Essay on avarice and profusion in his Enquirer.'

The discussion started the general question of the future improvement of society, and the author at first sat down with an intention of merely stating his thoughts to his friend upon paper in a clearer manner than he thought he could do in conversation. But as the subject opened upon him, some ideas occurred, which he did not recollect to have met with before; and as he conceived that every the least light on a topic so generally interesting might be received with candor, he was determined to put his thoughts in a form for publication (Malthus 1798).

A generation gap? Certainly, but one which only acquired its full significance from the fact that it occurred at a turning point in history. Indeed, in the first chapter, Malthus refers to 'the new and extraordinary lights that have been thrown on political subjects, which dazzle, and astonish the understanding; and particularly that tremendous phenomenon in the political horizon the French Revolution, which, like a blazing comet, seems destined either to inspire with fresh life and vigour, or to scorch up and destroy the shrinking inhabitants of the earth, have all concurred to lead many able men into the opinion that we were touching on a period big with the most important changes, changes that would in some measure be decisive of the future fate of mankind.'

Malthus was well aware of the disillusionment and skepticism engendered by this first historical assault on the prevailing ideology: 'The view which he has given of human life has a melancholy hue; but he feels

conscious that he has drawn these dark tints, from a conviction that they are really in the picture; and not from a jaundiced eye, or an inherent spleen of disposition' (Malthus 1798).

In this first essay, Malthus's argument is backed up by few concrete examples but for a brief examination of the different stages of civilization through which humanity has gone (Chaps. III and IV), case studies of England (Chap. V), the new colonies (Chap. VI), and a note on the speed with which even the old states recovered from the ravages of war, pestilence, famine, and natural catastrophes. The main part of the work (Chaps. VII–XV) is devoted to a rebuttal of the ideas of Wallace, Condorcet, and particularly Godwin, who had just published the third edition of his philosophical treatise *An Enquiry Concerning Political Justice and Its Influence on General Virtue and Happiness* (Godwin 1798). However, two chapters (Chaps. XVI and XVII) initiated the criticism of Adam Smith's economic theory, and two others (Chaps. XVIII and XIX, both omitted from the second *Essay*) seem to provide the key to Malthus' system of values.

In the tradition of the Scottish moralists—particularly Abraham Tucker—whose influence he had been exposed to at Cambridge, Malthus came to believe that the constant pressure that misery exerts on man leads to a conception of life in which hope is directed towards the afterlife (Tucker 1776, in Dupaquier and Fauve-Chamoux 1983). In contrast to these moralists, however, Malthus refused to see life on earth as a trial, but as 'a process necessary, to awaken inert, chaotic matter into spirit; to sublimate the dust of the earth into soul; to elicit an aethereal spark from the clod of clay (Malthus 1799).

Malthus went on to elaborate a theory of stimulation which anticipated that of Toynbee: 'To furnish the most unremitted excitements of this kind, and to urge man to further the gracious designs of Providence by the full cultivation of the earth, it has been ordained that population should increase much faster than food … consider man, as he really is, inert, sluggish and averse from labour, unless compelled by necessity … we may pronounce with certainty that the world would not have been peopled, but for the superiority of the power of population to the means of subsistence … . The principle according to which population increases, prevents the vices of mankind, or the accidents of nature, the partial evils arising from general laws from obstructing the high purposes of the creation. It keeps the inhabitants of the earth always fully up to the means of subsistence; and is constantly acting upon man as a powerful stimulus, urging him to the further cultivation of the earth, and to enable it, consequently, to support a more extended population' (Malthus 1798).

The Malthus of 1798 was a moralist, sociologist, and demographer, rather than an economist. He had begun to reflect on aspects of this domain, however. In his first *Essay*, he criticized Adam Smith's idea that any increase in revenue and capital leads to an increase in the size of the wage fund designated for the upkeep of the labor force, and referred to cases in which increasing wealth had in no way improved the living conditions of poor workers. On the same lines, addressing the problem of how correctly to define the wealth of a nation, he criticized the Physiocrats, and contended that the work of craftsmen and workers was productive for the individuals themselves, although not for the nation as a whole.

Malthus did not return to these thoughts in the first three editions of the second *Essay*, but the gradual development of his theory can be seen in *An Investigation of the Cause of the Present High Price of Provisions* (1800), his teachings at the East India Company's college, his article *Depreciation of Paper Money* in the *Edinburgh Review*, and his publications *Pamphlets on the Bullion Question* in the same journal, *Observations on the Effects of the Corn Laws* (1814), *Principles of Political Economy* (1820), and *Definitions in Political Economy* (1827).

How can Malthus' approach to economics be summed up in a few words? It is not a coherent, systematic and immutable whole: 'Yet we should fall into a serious error,' he wrote in the *Principles*, 'if we were to suppose that any propositions, the practical results of which depend upon the agency of so variable a being as man, and the qualities of so variable a compound as the soil, can ever admit of the same kinds of proof, or lead to the same certain conclusions, as those which relate to figure and number' (Malthus 1820).

Overall, Malthus adhered to the principles of the classical school, but was critical of the idea that human needs and desires are indefinitely and immediately expansible. According to Joseph-J. Spengler, Malthus argued that population growth is conditioned by the 'effective demand' for labor, and not primarily by the productive capacity. 'Such a demand in fact only tends to occur when adequate moral and political conditions coincide, when the social structure is flexible, the landed property correctly divided and commerce active, when a sufficient number of individuals are willing and able to consume more material wealth than they have produced, and when human beings are strong enough to compensate for the inelasticity of the demand for goods and services in terms of effort' (Malthus 1798).

It is primarily where demography, sociology, and morality are concerned that the second *Essay* marks a development in Malthus' thought.

In his first *Essay*, Malthus barely touched upon demographic science or political arithmetics as it was known at the time. He knew his Hume, Wallace, and Price, cited King, Short, and Süssmilch, but, with the exception of the latter, overlooked the work of foreign scholars. He contented himself with rough calculations on 'the unhealthy years' and the relationship between the number of births and the number of burials. In his second *Essay* on the other hand, he devoted one

chapter to the fertility of marriages, another to the effect of epidemics on population movements, and a third to emigration though this cannot be defined as demographic analysis.

Where sociology is concerned, the core of the second *Essay* consists in an analysis of the checks to population in the lowest stage of human society (Book I, Chap. III), among the American-Indians (Chap. IV), in the islands of the South Sea (Chap. V), among the ancient inhabitants of the North of Europe (Chap. VI), among modern pastoral nations (Chap. VII), in different parts of Africa (Chap. VIII), in Siberia (Chap. IX), in the Turkish dominions and Persia (Chap. X), in Indostan and Tibet (Chap. XI), in China and Japan (Chap. XII), among the Greeks (Chap. XIII), among the Romans (Chap. XIV), and in the various states of modern Europe: Norway, Sweden, Russia, middle parts of Europe, Switzerland, France, England, Scotland, and Ireland (Book II, Chaps. I–VIII).

For the most part, the sources of these reflections have been located (cf. the contributions of M. Godelier, N. Broc, and J. Stagl in Fauve-Chamoux *Malthus hier et aujourd'hui* (1984), and Malthus's interpretations—especially those concerning primitive societies—have been discussed, but the central element here is the new way of approaching the issue. Malthus, thus, emerges as one of the pioneers of the sociology of populations. It is on the moral and political level, in particular, that Malthus's theory is completed, confirmed, and substantiated in the second *Essay* and the successive editions.

In the first *Essay*, he asserted that the reproductive power of the human population could only be checked by misery and vice, the word 'misery' being used in the broad sense, and 'vice' in the sense of sexuality being deviated outside of the institution of marriage. While he did not yet use the term 'moral restraint,' he further indicated that the pressure of providing for a family operates to varying degrees in all classes of society, and termed this a 'preventive check' in opposition to the 'repressive check' of misery.

In the second *Essay*, he introduced moral restraint explicitly as another of the checks preventing the population from increasing beyond the limits of the means of subsistence. Two chapters of book IV are devoted to moral restraint, the obligation to practice this virtue, and the effect this would have on society.

All this was embedded in Malthus's reflections on the lot of the poor and the principle of aid. While in 1796 (*The Crisis*) Malthus had defended the principle of state assistance at home, two years later in the *Essay* he asserted not only that the immense sums of money collected in conjunction with the poor laws would not improve the living conditions of paupers, but that 'they have spread the general evil over a much larger surface,' having encouraged the poor to reproduce.

In the *Essay* of 1803, Malthus went even further. In order to illustrate the idea that paupers had no right to aid, he conceived the famous analogy of the banquet, in which the theme was carried to a ridiculous extreme. This offending passage was cut out of following editions, but was unearthed in 1820 by Godwin, who had undertaken, rather late in the day, to refute the Malthusian doctrine. (Godwin 1820) In this, he was followed by most of the later writers, either in good faith or otherwise.

Petersen, in his famous book *Malthus Reconsidered* (1979), shows why Malthus is termed reactionary by ideologists, and points out the injustice of this accusation: 'Malthus was an active member of the Whig Party, and the social reforms he advocated—in addition to the crucial one of universal schooling—included an extension of the suffrage, free medical care for the poor, state assistance to emigrants, and even direct relief to casual labourers or families with more than six children; similarly, he opposed child labour in factories and free trade when it benefited the traders but not the public. Apart from such details he was an honest and beneficent reformer, committed throughout his life to the goal that he shared with every liberal of his day—the betterment of society and of all the people in it.'

3. *The Impact of Malthus and Malthusianism*

Malthus's theory soon became known on the continent, particularly in France and Germany. In France, while the *Essay* of 1798 was not translated until 1980 (by Vilquin), parts of the *Essay* of 1803 were presented in the *Bibliothèque Britannique* as early as 1805 by the Genevian Pierre Prévost, who was also responsible for the translation of the third edition in 1809. The *Principles of Political Economy* appeared in French in the year of their English publication. In Germany, Malthus was translated several times from 1806 onwards. In the rest of Europe (Italy, Spain, Russia), Malthus's works were not translated until the second half of the nineteenth century, but as philosophers and economists read them in English, French, and German, they were much discussed. (Fauve-Chamoux 1984).

In view of the frequency of misquotations, misreadings, second-hand references, silly remarks, and invectives accumulated by these commentators, however, one cannot help but wonder whether the majority of them, even Karl Marx, had really read Malthus's publications.

The most common misconception—particularly among economists—is that Malthus contended that the population really does increase in geometrical progression while the means of subsistence increase in arithmetic progression. In fact, Malthus referred only to tendencies, and refused to allow his theory to be reduced to the simplistic conclusion that population is regulated by subsistence.

An associated misconception casts Malthus as an enemy of demographic growth. In fact, as early as

1805 he made it quite explicit that 'It is an utter misconception of my argument to infer that I am an enemy to population. I am only an enemy to vice and misery, and consequently to that unfavourable proportion between population and food which produces these evils.' This misconception was soon so widespread that the adjective 'Malthusian' was coined to describe not only the practice of restricting births but, in the last half of the nineteenth century, the practice of limiting economic production.

As a rule, socialists have poured scorn on Malthus for his liberal ideas and his skepticism with respect to social intervention policies. In attacking Godwin, he had undermined the very basis of utopian socialism. This resulted in a vigorous, if somewhat delayed, reaction from Godwin himself (Godwin 1820), to which Malthus responded the following year in the *Edinburgh Review*.

This debate came to a head in 1839, when Marcus (1939) accused Malthus of advocating the asphyxiation of 'surplus' newborns, a myth which was taken up and popularized in France by Leroux (1839) with expressions such as 'the somber Protestant of sad England' and 'the selfish defender of the propertied classes.' Karl Marx, champion of invective, is every bit as dismissive, denouncing Malthus as 'superficial,' 'a professional plagiarist,' 'the author of nonsense,' 'the agent of the landed aristocracy,' 'a miserable sinner against science,' 'a paid advocate,' 'the principal enemy of the people,' etc. (cf. Michelle Perrot, *Malthusianism and Socialism* in Dupaquier and Fauve-Chamoux 1983).

However, most socialists apart from Marx agree that there is a connection between overpopulation and misery, but distance themselves from Malthus with respect to the proposed causes and remedies. The most serious discussion of Malthus's theories is to be found in the work of Karl Kautsky (Kautsky 1880). When neo-Malthusianism came to the fore at the end of the nineteenth century, however, socialist criticism once more became more radical. Where his doctrines are concerned, Malthus has had a large number of successors, many of them illegitimate—at least in the author's system of values.

The 'legitimate' successors to Malthus's works include the analysis of the causes, processes, and consequences of the growth of populations and, in political economy, the theory of effective demand, which Keynes revived (1939), affording it great topicality (*The General Theory of Employment, Interest and Money*), as well as the references made to the *Essay on the Principle of Population* by Charles Darwin (1838) and Alfred Russel Wallace (1858).

The illegitimate successors—the neo-Malthusianists—are much more numerous and more visible, having remained in the forefront.

As early as 1822, the Englishman Francis Place, while adopting Malthus's conceptual framework, proposed a much easier and more seductive way of checking population growth than moral restraint: the voluntary limitation of births within marriage (Place 1822). Knowlton also advocated this approach in the United States (Knowlton 1833), and the idea met with the approval of scholars such as Carlisle and Stuart Mill.

In 1860 the journalist Charles Bradlaugh set up the Malthusian League in London. The organization went from strength to strength from 1877 onwards, when Bradlaugh and Annie Besant, one of the pioneers of feminism, were prosecuted for obscenity.

Neo-Malthusianism was taken to France by the anarchist Paul Robin, who founded the League for Human Regeneration in 1896, and to Germany by the social democrat Alfred Bernstein, who in 1913 organized two meetings in Berlin, causing quite a stir. However, the neo-Malthusians came up against not only the hostility of the authorities but the distrust of socialist theorists.

At the time, militant feminism was evolving all over Europe, and birth control was just one element of its main objective: the sexual liberation of women. The decisive turn, however, was taken in the US thanks to Margaret Sanger, who in 1916 had founded a birth control clinic in Brooklyn. In 1925 she organized an international neo-Malthusian probirth control conference in New York, and in 1927 the first international congress on population was held. This was the basis of the International Union for the Scientific Study of Populations, which however almost immediately dissociated itself from the neo-Malthusian network.

See also: Civilization, Concept and History of; Demographic Transition, Second; Demography, History of; Evolution, History of; Fertility Control: Overview; Fertility: Proximate Determinants; Infant and Child Mortality in Industrialized Countries; Political Economy, History of; Population Cycles and Demographic Behavior; Population Dynamics: Classical Applications of Stable Population Theory; Population Dynamics: Momentum of Population Growth; Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems; Population Dynamics: Theory of Nonstable Populations; Population Dynamics: Theory of Stable Populations; Population, Economic Development, and Poverty; Population Forecasts; Population Pressure, Resources, and the Environment: Industrialized World; Ricardo, David (1772–1823); Smith, Adam (1723–90); Social History

Bibliography

Bonar J 1885/1924 *Malthus and His Work*. Macmillan, New York

Condorcet (Caritat M-J-A) 1795/1970 *Esquise d'un tableau historique des progrès de l'esprit humain*. Vrin, Paris

Celass D V (ed.) 1953 *Introduction to Malthus*. Wiley, New York

Charbit Y 1981 *Du Malthusianisme au populationisme, les économistes français et la population (1840–1870)* INED. Travaux et documents Paris
Dupaquier J, Fauve-Chamoux A (eds.) 1983 *Malthus Past and Present*. Academic Press, London
Eversley D E C 1959 *Social Theories of Fertility and the Malthusian Debate*. Clarendon, Oxford, UK
Fauve-Chamoux A (ed.) 1984 *Malthus hier et aujourd'hui*. Société de démographie historique, Paris
Godwin W 1798 *An Enquiry concerning Political Justice and its Influence on Moral and Happiness*. University of Toronto Press, Toronto
Godwin W 1820 *Of Population: An Enquiry Concerning the Power of Increase in the Numbers of Mankind, being an Answer to Mr Malthus's Essay on that Subject*. Longman, London
Gordon L 1976 *A Social History of Birth Control in America*. Viking and Penguin, New York
James P (ed.) 1966 *The Travel Diaries of Thomas Robert Malthus*. Cambridge University Press, Cambridge
James P (ed.) 1979 *Population Malthus, His Life and Times*. Routledge, London
Kautsky K 1880 *Der Einfluß der Volksvermehrung auf den Fortschritt der Gesellschaft*. Block und Hasbach, Vienna
Keynes J M 1939 *Robert Malthus: The First of the Cambridge Economists, Essays in Biography*. New York
Knowlton C 1833 *The Fruits of Philosophy*. Harcourt Brace, New York
Marcus 1939 *The Book of Murder. Vademecum for the Comissioners and Guardians of the New Poor Law. With a Refutation of the Malthusian Doctrine*. London
Malthus T R 1798 *An Essay on the Principle of Population as it Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condocet and other Writers*. London
Malthus T R 1803 *An Essay on the Principle of Population, or a View of his Past and Present Effects on Human Happiness, with an Inquiry into our Prospects respecting the Future Removal or Mitigation of the Evils which it occasions*. J. Johnson, London
Malthus T R 1820 *Principles of Political Economy, Considered with View to their Practical Application*. London
Otter W 1836 Memoir of Robert Malthus. In: *Malthus Principle of Political Economy*. John Murray, London
Petersen W 1979 *Malthus Reconsidered*. Harvard University Press, Cambridge, MA
Place F 1822 *Illustrations and Proofs of the Principle of Population*. Himes, London
Smith K 1951 *The Malthusian Controversy*. Routledge & Paul Kegan, London
Stassart J 1957 *Malthus et la population*. Faculté de droit, Liège

J. Dupâquier

Managed Care

Managed care became a major area of health policy interest in the 1990s. In the early part of the decade it was seen as a potential solution to a wide range of problems in the US health care system. In the latter part of the decade, it became for some the exemplar of what was wrong in the system. However, there is no universally agreed upon definition of managed care. Some apply it to any effort to control health care costs. Others focus on efforts to change the processes by which patients receive care. Some focus on one type of managed care plan, the Health Maintenance Organization (HMO). Finally, some focus not on managed care *per se*, but on its effects on the health care system and the population.

1. Historical Context

The many aspects of managed care—and the debates over it—can only be understood in the context of the US health care system. Although some aspects of managed care may be applicable outside the US, attempts to transfer specific lessons need to be made with care. The United States is an outlier among highly developed countries because of its very high expenditures per capita, substantial proportions of the population without insurance coverage, and less than optimal health status (Anderson and Poullier 1999, National Center for Health Statistics 1999, Vistnes and Zuvekas 1997), reflecting the absence of a national health care delivery system or policy. To examine the development of, and the reaction to, managed care, it is necessary to go back to the period shortly after World War I. At that time, medical care was fee-for-service in its literal sense. Billing was uncommon, and there was no 'third party' between physician and patient. If the patient could not pay, care might not be obtained, unless the physician voluntarily provided care. This fit well the economist's description of medical care as a cottage industry. Physicians tended to practice independently, or with a partner or two.

The development of health insurance plans during the Depression brought a third party into the picture. Indemnity insurance plans reimbursed the patient a fixed amount for services received. Any fees in excess of the indemnification, as well as the paperwork, were the responsibility of the patient. Blue Cross and Blue Shield plans—not-for-profit organizations sponsored by the state hospital and medical associations—provided what is known as a service benefit. They contracted with physicians and hospitals to accept their reimbursement as 'payment in full.' For hospitals, these were literally reimbursements—the Blue plan reimbursed them for their share of total costs. For physicians, the payments were generally based on usual, customary and reasonable fees. Nonetheless, physicians were often uneasy about the involvement of a third party in the doctor–patient relationship.

Enrollment in these health insurance plans was voluntary, and in the classic insurance model, services that could be anticipated, such as maternity care, well-baby checkups, and physical exams, were not covered. Deductibles and coinsurance provided some brake on

the use of services, but there was often an out-of-pocket cap to prevent a patient from being bankrupted by hospitalization or a major illness (Starr 1982).

The rapid expansion of employer-sponsored health insurance after World War II brought the entrance of a fourth party—the employer—into this picture. In most instances, the employer merely paid part or all of the premiums for the health plan it chose to offer, and health plans varied largely in terms of the breadth of coverage and the extent of deductibles and co-insurance. Employers continued to be passive payers of premiums until recently, seeing their 'contributions' to health insurance benefits as just another part of the compensation package.

1.1 Group Practice and Prepayment

These features largely characterized the medical care environment until the early 1970s, but there were a few notable exceptions. As early as the late nineteenth century, some physicians organized themselves into group practices (Trauner 1977). Their goal was usually to share clinical responsibility and coverage, and to have ready access to consultations. While most charged their patients on a fee-for-service basis, and many compensated their physicians on a similar basis, there was often substantial opposition to group practice from independent physicians, probably because the 'in-house' referrals threatened their access to patients.

Prepayment for medical care actually had its origins in the mid-1800s with immigrant societies that hired physicians to take care of their members. By paying a fixed amount per month, members could receive medical care when they needed it without any further out-of-pocket payment (Trauner 1977). This was not just 'insurance for very expensive events,' but true prepayment. This prepayment to a small set of physicians was seen by independent physicians to threaten their livelihood.

During the Depression and shortly after World War II, several new organizations married the concepts of group practice and prepayment. These included the Ross-Loos Clinic in Los Angeles, Kaiser Foundation Health Plan, Group Health Cooperative of Puget Sound, and Group Health Association of Washington DC. These plans used group practice, with its shared medical records and internal physician quality review, in combination with prepayment for all needed services to remove the financial barriers to care.

The early founders of these plans wanted a better way to organize and deliver care. Many were not-for-profit, and some were consumer controlled. They often had a strong public health focus, and offered complete coverage of prenatal care, maternity care, well-baby visits, and immunizations, services excluded by conventional insurance. Some of the plans, notably Group Health Cooperative, Kaiser in Portland, Oregon, and Northern California, the Health Insurance Plan of Greater New York and later, Harvard Community Health plan developed highly-regarded health care research units (Luft and Greenlick 1996).

1.2 Health Maintenance Organizations

The growth of prepaid group practices (PGPs) in the 1950s and 1960s led to increasing evidence that, even though they avoided copayments and deductibles, offered more extensive benefits, had arguably comparable quality of care, (although there were few measures of this), and seemed to keep most of their patients satisfied, their most impressive achievement was lower premiums. By the end of the 1960s, the expansion of insurance coverage by Medicare and Medicaid led to rapid growth in medical care expenditures. The Nixon Administration sought a new, market-based approach that would control costs, but the concept of prepaid group practice was anathema to the American Medical Association.

To address this problem, Lewis Butler and Paul Ellwood coined the term 'Health Maintenance Organization,' or HMO, in 1971. It encompassed both the concept of the prepaid group practice and the San Joaquin-style foundation for medical care (now relabeled the Individual Practice Association, or IPA) in which physicians in their independent practices collectively took responsibility for the costs associated with an enrolled population (Harrington 1971). The IPA paid the physicians on a fee-for-service basis, and there was no need for physicians to work together in group practices. While these two types of HMOs differed significantly in how medical care was organized and delivered, they shared the notion of integrating the insuring function with the responsibility for delivering services to a defined population.

There could be many different organizational and legal structures for these HMOs, often in response to the vagaries of state regulatory requirements. However, the nature of the classic PGP and IPA forced physicians to recognize that the economic implications of the decisions they made would be borne by their organization, rather than some distant insurer or government payer. The PGP, however, typically involved physicians who saw only HMO patients, and on whom the HMO relied to provide nearly all its medical care. This meant that they could develop implicit styles of practice consistent with clinical, professional, and organizational needs. In the IPA, by contrast, the vast majority of patients seen by its physicians were covered by other insurance plans, and because physicians did not practice together, informal approaches to developing consistent practice styles were impossible.

Aside from the usual policy tools of grants and loans to develop new organizations, and federal preemption of state laws prohibiting HMOs, a mech-

anism was developed to facilitate market entry by HMOs. If an employer of 25 or more workers offered a health insurance plan, HMOs could use the mandate authority in the legislation to require that they also be offered to employees at terms no less favorable than the fee-for-service plan.

The mandate option, however, was not unlimited. An employer needed only offer one plan of each type—PGP and IPA. This led to questions about how one should define the two types of plans. The common-sense notion was that the PGP combined prepayment for a set of benefits with the delivery of medical services through a group practice, while the IPA did not rely on the group, or clinic-like practice setting. However, suppose that a large PGP was already well-established in an area. A new PGP could not gain entry to employers wanting to avoid having to deal with multiple plans. If classified as an IPA, however, such a plan could gain entry in the marketplace, and a few IPAs emerged with salaried physicians who practiced in clinic settings.

Even without such subterfuge, the PGP and IPA definitions were inadequate. New HMOs began that contracted primarily with networks of such clinics, much like the Kaiser model, but unlike Kaiser, the clinic physicians continued to see fee-for-service patients. Some called these plans network model HMOs, but they were classified under the IPA rubric for mandating purposes.

An early review of the evidence on HMO performance (Luft 1981), based on data through the 1970s, focused largely on comparisons of HMOs versus fee-for-service insurance (FFS), but recognized the distinction between IPAs and PGPs. In fact, the cost-containing effects of HMOs were most apparent for PGPs. This was not surprising given the economic incentives. The more extensive coverage offered by HMOs increased demand, so some other mechanism, such as salaried payment to physicians or organizationally influenced practice patterns, was needed to constrain supply. Although IPAs may have had a fixed overall budget, they paid their physicians fee-for-service. Predictably, enrollees' assessments of IPAs and FFS were comparable, but differences appeared between PGPs and FFS.

1.3 Managing Costs and Managing Care

The 1980s saw a slow but steady growth, both in the number of HMOs and their enrollment. Much of the growth was in forms other than PGPs. Nevertheless, the policy debate referred to the research on HMOs (generically) even when most of it was on PGPs. Furthermore, much of the research focused on a few highly visible plans, most of which were long-established, not-for-profit, plans with internal research groups. Generalizing from these plans to all PGPs, let alone all HMOs, was clearly unwarranted. But, these studies provided strong evidence that alternatives to FFS could provide high quality care with low cost and reasonable enrollee satisfaction.

Nixon's HMO strategy did not lead to the rapid growth of HMOs. Instead, government price controls were implemented and then private sector approaches to controlling utilization of services, such as 'second opinion programs' in which an insurer would pay for the procedure only if its necessity was confirmed by a second physician. The effectiveness of these programs was questionable, but it set the stage for the refusal by payers to simply reimburse every service rendered by a duly licensed physician (McCarthy and Finkel 1978). Instead, some services would be covered only if approved by the payer. These payment decisions could be quite problematic because the insurer became involved only after the services were rendered and the patient submitted a claim. Not surprisingly, these 'retroactive denials' resulted in substantial patient dissatisfaction.

By the early 1980s, HMOs were becoming increasingly visible outside of the geographic areas in which they had long existed, and employers were asking conventional insurance carriers why they couldn't manage costs as well as did the HMOs. Large employers were self-insuring, thereby capturing the interest earnings on premiums, and avoiding state-mandated insured benefits, such as mental health coverage and chiropractic. Utilization review firms that determined whether benefits should be paid were capturing the cost-control market, and third-party administrators were processing claims. In fact, the very survival of conventional health insurance plans was being threatened because their products were not unique, and others seemed to perform the separate functions better.

California legislation that allowed both its Medicaid program for the poor and private insurers to selectively contract with certain providers changed this situation (Bergthold 1984). Prior to this time, plans controlled the benefits they offered and the financial incentives placed on patients, but patients were free to go to any licensed health care provider. While the Blue plans had contracts with providers, the historical connections between these plans and the local hospital and medical associations meant that all providers interested in participating were accepted.

The intent of the California legislation for its Medicaid program was quite clear. Hospitals were encouraged to bid low in order to increase their market share at the expense of their competitors. Companion legislation allowed the development of what are now called Preferred Provider Organizations (PPOs), which could selectively negotiate contracts with providers for the 'private' market. With fees negotiated, and with a known deductible and co-insurance rate, PPOs were also able to offer their enrollees the advantage of direct billing by the provider. From the patient's perspective, this is markedly

simpler, but the physician may now have to deal with hundreds of different plans, often with differing claims forms and coverage rules.

To further improve the attractiveness of their plans to employers, insurers offering PPOs included the option of various utilization management approaches, such as prior authorization, concurrent review of hospital stays, and the like. Because the PPO now had a contractual relationship with the providers, it could require the provider to agree to these utilization management efforts as part of the contract. Undertaking them in advance would avoid embarrassing retroactive denials.

In most parts of the country there is an oversupply of hospitals and physicians, particularly specialists. Thus, when California's state Medicaid program, or health plans more generally, presented providers with contracts in order to continue to participate, there were two reactions. The first was the hope that as a 'contracting or participating provider,' one would get more patients, even if the fee paid per patient might be somewhat lower. The second was fear that if one did not sign, a large fraction of one's patients would be lost.

PPOs were seen by policy analysts as just a method to constrain payment levels—'fee-for-service in drag.' However, with a unique network of providers, in combination with various utilization management approaches, including using primary care physicians as gatekeepers, a new 'product' was defined. Conventional insurers began using the term 'managed care' in the mid-1980s to distinguish this new approach from the utilization management offered by small stand-alone firms, and from simple PPOs that merely negotiated discounts. Not surprisingly, the large insurers who were attempting to transform themselves in order to survive quickly pointed to the evidence on HMO performance as the rationale for managed care organizations.

1.4 The Managed Care Market Environment

The Clinton health care reform proposals of 1993 added further terminology. In addition to promising close to universal coverage through a variety of funding mechanisms, there was a carefully designed effort to structure 'managed competition' among health care delivery systems. While fee-for-service could be one of these systems, others would be similar to HMOs. Managed competition, however, was supposed to deal with the flow of funds to health plans, to adjust for risk differences among plans, to monitor quality, and address consumer and patient service. Even if all managed care plans were well-designed, well-meaning prepaid group practices, a market organizing structure would still be necessary. Unfortunately, in the wake of the Clinton reform failure there is no such market organizing structure. The decade of the 1990s was one of intense market-based pressures on employers to become more efficient. Corporate restructuring and cost reductions were commonplace in all industries. Whereas wages and salaries were increasing at 1 percent or less per year, health insurance premiums were increasing at over 13 percent per year at the end of the 1980s. Employers began to pressure health plans to reduce that rate of increase. Aggressive cost containment became the watchword for health plan survival and the rate of premium growth fell steadily until by 1995 it was actually negative (Levitt and Lundy 1998).

These external pressures also led to a transformation of the health insurance and health plan industry. New forms of health plans developed and some not-for-profit Blue Cross and Blue Shield plans turned for-profit. Physicians sold their practices to practice management firms and became employees. Pharmaceutical companies began advertising directly to the consumer. These and other changes further eroded the traditional physician–patient relationship.

One core feature of the managed care organization is that it has a network of physicians, regardless of whether they practice together in a group setting such as in the old PGP model, or independently with contractual relationships to the health plan. This means that if a person switches from one plan to another, the physicians available may change. If the switch results from an employer changing health plan options in search of lower premiums, then the anger at having a longstanding relationship broken can be substantial. Reflecting this concern, there has been a shift to broader networks, including nearly all physicians, but this makes it more difficult for plans to concentrate patients and control costs by altering practice patterns.

Cost containment efforts by plans led to constraints on physician fees, both by the federal Medicare plan for the elderly, by state Medicaid plans for the poor, and by health plans. To maintain their incomes, physicians were under pressure to see more patients in less time. Interestingly, there is little evidence that the average patient visit has gotten shorter, or that physician incomes have fallen, but that is a common perception, as is the attribution of the problem to managed care (Luft 1999).

Outpatient drug coverage has become an increasingly common benefit during the 1990s, but the marketing of new, and much more expensive drugs has led to rapidly increasing costs for this benefit. Health plans have limited their formularies, but since coverage may vary across patients in a plan, let alone across plans, patients may find out their new drug is not covered only when arriving at the pharmacy.

The rapid growth of some plans in the mid-1990s, the conversion of some plans to for-profit, and the resulting fortunes made by some industry leaders

attracted enormous attention in the press. Numerous studies pointed to administrative waste and high profits in the industry (Woolhandler and Himmelstein 1997). Alternatively, government data on health expenditures indicate that insurance administration and profit was falling between 1994 and 1997 (Health Care Financing Administration website).

At the same time, the cost pressures of managed care contracts has made it more difficult for hospitals to charge their paying patients more in order to subsidize the uninsured. Premiums for nonmanaged care plans have increased very rapidly, and there has been a general reduction in the prevalence of employer-sponsored insurance, leading to a rising number of uninsured. Thus, while managed care plans may be containing costs for their enrollees, the perception is that this market-focused approach has worsened the situation for many outside the plans, and even for those in the plans, many expectations are not being met.

2. *Evidence on Managed Care Performance*

Has managed care totally supplanted the promise of prepaid group practice? In contrast to earlier reviews, 'current' reviews found that both IPAs and PGPs demonstrated lower costs and utilization of services. As expected, consumers were often less satisfied with the nonfinancial aspects of HMOs, particularly when they felt they did not have a choice of plans. The evidence on quality of care was of particular interest. In contrast to the beliefs of either the advocates or detractors of HMOs, there were as many studies showing better quality in HMOs as worse quality. This balance, however, was not without significant findings—in fact, there were several studies with clinically important and statistically significant differences, but the balance of evidence was maintained. It also means that advocates and detractors can easily point to selected studies to buttress their case (Miller and Luft 1997).

This research has several implications. First, labels matter very little. Some HMOs have very good quality; some very bad. There is not enough data to determine whether 'plan type'—PGP versus IPA—is a distinguishing factor, but the ways in which labels are applied suggest that plan type will not be very informative. While not-for-profits may have certain desirable characteristics, not all for-profit plans are 'bad guys.' Furthermore, it is almost impossible to separate (in a statistical sense) the effects of not-for-profit status from PGP model from 'old established plan.' It may be best to forgo a belief that types and categories truly determine performance, and focus instead on how to evaluate each organization, give it the appropriate incentives and resources to perform well, and then monitor that performance continuously.

An important conceptual limitation is that the standard for comparison matters. Some of the more recent evidence suggests that HMOs may have a smaller cost advantage relative to their competitors than had previously been the case. It also appears that areas with rapid HMO growth have a *slower* rate of growth in costs for FFS-covered enrollees (Chernew 1995, Robinson 1996, Wickizer and Feldstein 1995). These two findings are consistent and suggest that the spread of HMOs creates externalities, or 'spillover effects' outside the plans themselves. This will result in the incorrect observation that HMOs have no impact, when in fact, their impact may be pervasive, and this cannot be determined merely by comparison with FFS. Instead, external measures are needed for what is good care, what utilization of services is appropriate, and what costs are reasonable. Setting such 'gold standard' measures without reference to an existing medical care system may be impossible, and is certainly well beyond the scope of studies focusing only on managed care.

Furthermore, there is important evidence that choice, and the lack thereof, matters in the assessment of plan performance. Americans, especially those with sufficient incomes to see themselves as consumers, rather than as supplicants, value choice highly. The greatest dissatisfaction is expressed by people who have been given no choice but to be in an HMO (Davis et al. 1995). All types of health care systems and coverage imply some tradeoffs to contain costs. HMOs limit the choice of physicians, require prior approval before certain services can be rendered, or encourage physicians to sometimes say 'no' to patient requests. FFS plans use less visible means—restricting the benefit package, having premiums so high only the rich can afford them, or imposing substantial deductibles and copayments. The early PGPs, such as Kaiser, required that a FFS plan be offered as an alternative. This guaranteed that their enrollees were there by choice; if they were dissatisfied, they could, and did, leave. Choice allows people to get what they want, but it causes problems because not all patients are equally costly. Unfortunately, plans have no incentive to provide the best quality care because they will attract the sickest enrollees.

3. *Lessons from Managed Care*

In some ways, it is far too soon to offer lessons because the US is still in the process of developing a health policy, and managed care is but one of many tools being used. However, some speculation may be useful while awaiting more evidence. The early forms of managed care, PGPs, were designed as integrated delivery systems—integrated in the sense that both the financing and delivery components were linked. While some of these organizations owned their hospitals, this

was not universal, and others were able to develop long-term arrangements with their hospital providers. Allowing funds to be flexibly allocated among inpatient and outpatient settings, physicians and other staff and pharmaceuticals seems to be one of the lessons of PGPs that may translate well across boundaries.

It is also important to note that these plans were, and still are, minorities in the context of an oversupplied, overly specialized, and unorganized fee-for-service system. Furthermore, the people who began these PGPs, and to a large extent, those still attracted to working in them, were seen as unusual, if not suspect. In turn, the organizations could design staffing patterns that fit their concept of integrated care, and not have to passively accept the specialty mix preferred by medical training programs. Being able to operate almost as islands in a FFS sea, with voluntary recruitment of staff and enrollees, allowed these plans to develop clinical practice patterns that were mutually acceptable, without having to force change on anyone. The US political system makes massive, uniform change impossible, but the lurches back and forth in other nations suggest that in the absence of knowing precisely what approach is best, experimentation on a smaller scale may be desirable. At the very least, it allows one to attract those most willing to adapt to a new system.

In the late 1990s, many of the problems that people point to in the US health care system reflect the absence of universal coverage, and thus large numbers of uninsured individuals, combined with a surplus of medical resources for those with coverage and a strong desire by the people for the right to choose both health care systems and providers. The presence of a wide variety of plans meets consumer demands for choice, but results in high marketing and administrative costs, along with incentives to not emphasize high quality of care for fear of attracting high-risk enrollees. Efforts to control costs imply changes in the perceived, and possibly actual, control physicians in traditionally FFS practice have over their decisions and thus encounter substantial resistance. The fragmented employer-based system of coverage in the US is not recommended for export, but experimentation with some variations in delivery of care (in contrast to financing coverage) may have some merit.

The changing managed care terminology suggests that new labels will be developed in the future, sometimes to better describe reality, sometimes to expand an accepted concept into new territory. This highlights the importance focusing on the components of health care delivery systems, how (and why) they are organized, the context in which they operate, and the standards by which they are assessed. New types of delivery systems are being created, some will succeed, but most will fail. New technologies such as information systems, will have enormous implications, but exactly how these will develop is unknown. The variability in approaches, however, allows us to examine in more detail what works and what does not, as long as one focuses on fundamental aspects, rather than simply labels.

See also: Reproductive Rights in Affluent Nations

Bibliography

Anderson G F, Poullier J P 1999 Health spending, access, and outcomes: trends in industrialized countries. *Health Affairs* **18**(3): 178–92

Bergthold L 1984 Crabs in a bucket: The politics of health care reform in California. *Journal of Health Politics, Policy and Law* **9**(2): 203

Chernew M 1995 The impact of non-IPA HMOs on the number of hospitals and hospital capacity. *Inquiry Journal of Health Care Organization* **32**(2): 143–54

Davis K, Collins K S, Schoen C, Morris C 1995 Choice matters: Enrollees' view of their health plans. *Health Affairs* **14**(2): 99–112

Harrington D C 1971 San Joaquin Foundation for Medical Care. *Hospitals* **45**(6): 67–8

Health Care Financing Administration website (www.hcfa.-gov/stats/nhe-oact/nhe.htm) 1997 National Health Expenditures, Table 2, National health expenditures aggregate amounts and average annual percent change, by type of expenditure: selected calendar years 1960–97

Hellinger F J 1998 The effect of managed care on quality: a review of recent evidence. *Archives of International Medicine* **158**(8): 833–41

Levitt L, Lundy J 1998 *Trends and Indicators in the Changing Health Care Marketplace*. The Henry J. Kaiser Family Foundation, Menlo Park, CA

Luft H S 1981 *Health Maintenance Organizations: Dimensions of Performance*. Wiley, New York

Luft H S 1999 Why are physicians so upset about managed care? *Journal of Health Politics, Policy and Law* **24**(5): 957–66

Luft H S, Greenlick M R 1996 The contribution of group- and staff-model HMO to American medicine. *Milbank Quarterly* **74**(4): 445–67

McCarthy E G, Finkel M L 1978 Second opinion elective surgery programs: Outcome status over time. *Medical Care* **16**(12): 984–94

Miller R H, Luft H S 1997 Does managed care lead to better or worse quality of care? *Health Affairs* **16**(5): 7–25

National Center for Health Statistics 1999 *Health, United States, 1999*. US Government Printing Office, Washington, DC

Robinson J C 1996 Decline in hospital utilization and cost inflation under managed care in California. *Journal of the American Medical Association* **276**(13): 1060–4

Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York

Trauner J B 1977 *From Benevolence to Negotiation: Prepaid Health Care. in San Francisco*, Ph.D. thesis, University of California, CA

Vistnes J P, Zuvekas S H 1997 Health insurance status of the civilian noninstitutionalized population: 1996. *MEPS Research Findings No. 8*. Agency for Health Care Policy and Research, Rockville, MD

Wickizer T M, Feldstein P J 1995 The impact of HMO competition on private health insurance premiums. *Inquiry Journal of Health Care Organization* **32**(3): 241–51

Woolhandler S, Himmelstein D U 1997 Costs of care and administration at for-profit and other hospitals in the United States. *New England Journal of Medicine* **336**(11): 769–74
Woolhandler S, Himmelstein D U 1997 Erratum. *New England Journal of Medicine* **337**(24): 1783

H. S. Luft

Managed Care and Psychotherapy in the United States

The impact of managed care on the delivery of mental health and substance abuse services has been significant. This entry traces the evolution of managed behavioral care and psychotherapy from the 1970s to the present. Provider criticisms of managed care are addressed. A patient-centric approach to managed behavioral health care is proposed in which providers and managed care work as partners to enhance treatment outcomes and keep health care affordable.

1. The Evolution of Managed Behavioral Health Care

Managed behavioral care is the integration of the financing and delivery of mental health and substance abuse treatment within a system that seeks to manage the accessibility, cost and quality of that care (Marques 1998, p. 411). The managed health care industry is the result of conflicting economic and political pressures to ensure Americans unlimited access to health care services, regardless of their ability to pay, in a market-based system of health care delivery.

The USA is unique among industrialized nations in its reliance on private employers to voluntarily provide health insurance for their employees (Iglehart 1999, p. 6). Approximately 65 percent of Americans receive health insurance through the workplace, 66 percent of them through a managed care plan (*Facts About Managed Care* 1999). In all other industrialized nations, the government provides health care or requires private employers to provide coverage (Studdert et al. 1999). Fears that double-digit health inflation will cause employers to abandon the provision of health care coverage create a government interest in keeping health care premiums low.

1.1 Managing Care in the 1970s

Managed care came into prominence in the 1970s as a result of spiraling health care costs (Mullen 1995, pp. 2–6). Rapid technological advancement in combination with an aging population created exponential increases in health care spending. Health insurance allowed patients to receive treatment without awareness of the costs, permitting practitioners to 'set their price' without consumer flight. Employers turned to managed care as a solution to escalating insurance premiums and fears of the inefficiency of a government-run health care system.

The 1970s approach to managed care focused on utilization review of hospital care. Utilization review is a process by which a managed care organization (MCO) evaluates whether recommended treatment is medically necessary and, as a result, reimbursable under a health insurance policy. This determination is typically conducted through a comparison of proposed treatment plans with medical necessity guidelines. At first, these guidelines were considered proprietary, creating a guessing game as to the definition of medical necessity. It was not until the early 1990s that Green Spring Health Services became the first managed behavioral health care company to make medical necessity criteria public, training providers on their use.

1.2 Managing Behavioral Health Care in the 1980s

As the rapid increase in physical health costs began to slow in the 1980s, mental health costs continued to rise, representing an increasing proportion of the health care dollar (Mihalik and Scherer 1998, p. 1). Mental health benefits typically covered 28–45 days in a psychiatric hospital. This coverage encouraged overutilization of expensive inpatient services and an overbuilding of psychiatric hospitals. Psychiatric facilities competing to fill empty hospital beds advertised all-expense-paid 'vacations' to lush locations, respite from acting-out adolescents, and retreats for alcoholics with health insurance. One study concluded that 40 percent of psychiatric hospitalizations were medically unnecessary (Mechanic et al. 1995, pp. 19, 20).

Also in this decade, the medical cost offset of psychotherapy became widely known. Nick Cummings, through his work with Kaiser Permanente, demonstrated that brief, effective therapy could reduce the number of doctor visits of 'high-utilizers' by as much as 65 percent (Cummings 1985). These 'worried well' sought treatment despite the absence of an organic cause for symptoms. Resolution of physical complaints was successfully achieved through psychological treatment of emotional distress. Meta-analysis of other cost-offset studies found 10–33 percent reductions in hospitalization costs for physical illness following mental health treatment (Mumford et al. 1984, p. 1145). These cost reductions were attributed to the alleviation of the psychological distress associated with a traumatic or chronic illness.

The desire to prevent medically unnecessary psychiatric hospitalization and the potential for medical

cost savings resulted in the creation of mental health 'health maintenance organizations' (HMOs). American Biodyne Centers, California Wellness Plan, and Psychology Management Systems were pioneers in the provision of unlimited outpatient mental health treatment designed to reduce inpatient psychiatric costs and overall medical expenses. As with most staff model HMOs, however, rapid access to routine mental health treatment was problematic. To increase appointment availability managed behavioral care incorporated a network model of service delivery similar to those used in medical managed care systems.

1.3 Managing Behavioral Health Care in the 1990s

Mental health 'carve-out' companies emerged in the 1990s. These carve-outs, called managed behavioral health organizations (MBHOs), specialize in managing mental health and substance abuse treatment. Health plans pay the MBHO a percentage of the health care premium for all members to manage the mental health utilization of the approximate 5–10 percent of members who use the benefit. The MBHO performs medical necessity review of the mental health treatment recommended by a network of independently contracted facilities, programs, and mental health practitioners. In 1999, 177 million Americans were contracted to receive mental health services managed by an MBHO (1999).

In the early 1990s the MBHO managed care through a combination of utilization review, discounted fee arrangements, and benefit management. Of mental health treatment dollars 65–80 percent were paid to inpatient and residential treatment facilities (Hennessy and Green-Hennessy 1997, pp. 340, 341). Because 80 percent of mental health expenditures were associated with treatment for 20 percent of patients, early discharge planning, strategies to improve treatment compliance, and case coordination were managed care interventions for 'high-utilizers' of psychiatric services (Patterson 1994, pp. 51, 54). The effectiveness of these techniques resulted in savings of 25–40 percent, at the same time increasing access to treatment for all members.

The late1990s saw the brief emergence of 'integrated delivery systems.' This approach was based on the premise that practitioners paid to provide all of a member's care would be clinically and financially motivated to treat a patient in the least expensive, but most effective manner possible. Ideally, as the need for treatment intensity diminished, a patient could be 'stepped down' to less restrictive settings within the same facility, maximizing treatment coordination. Instead, costs increased and providers bankrupted. Patients previously treated at one level of care received treatment at all levels of care. Outpatient providers were unable to simultaneously provide and manage care, and still make a profit.

Given the price pressures of the industry, when delivery systems could not integrate, the MBHO industry did. Magellan Health Services acquired four MBHOs covering over 71 million members. FHC purchased Value Health Services and created ValueOptions, an MBHO covering 22 million members. United Healthcare purchased United Behavioral Health covering 20 million lives. This consolidation served as a catalyst for provider political organization and activism.

2. Criticisms of Managed Care

Most mental health providers oppose managed care. Providers contend that managed care organizations face an inherent conflict between quality and cost control. Well-publicized stories of personal tragedy resulting from bad utilization management decisions have created an environment in which managed care has gone from being the centerpiece of health care reform to its target. Provider criticisms of managed care fall into four categories: (a) increased nonreimbursable administrative time; (b) sacrificing of quality for cost savings; (c) intrusion on professional autonomy; and (d) a lack of MBHO accountability.

2.1 Increased Nonreimbursable Administrative Time

Provider dissatisfaction with managed care companies is often reflected in complaints of increased administrative time and decreased income. Many administrative requirements originate in private accreditation bodies' concerns for the quality of care under managed care, but require providers to complete additional paperwork. For example, the National Committee for Quality Assurance (NCQA) requires 'credentialing.' Credentialing begins with the practitioner's completion of a lengthy application and ends following a one- to three-month verification process. The practitioner's education and license are validated. Malpractice history, criminal record, and complaint data banks are reviewed for patterns of misconduct. Credentialing reduces the risk that members will receive treatment from clinicians unqualified to help them or potentially harm them.

Other administrative tasks are associated with ensuring that only medically necessary treatment is reimbursed. Precertification ensures that a member's benefit covers the services to be provided, and the medical necessity of the level and intensity of care proposed. Treatment plan review evaluates patient progress and ongoing necessity for treatment. These reviews also protect the patient from incurring surprise financial liability, and the practitioner from accruing 'bad debt.' Prior to managed care, 15 percent of provider revenue was written off due to an inability to collect (Luft 1999, pp. 957, 964).

2.2 Sacrificing Quality for Cost Savings

Practitioners presume that a cost containment focus precludes concerns about treatment quality. Managed care payment mechanisms such as subcapitation, case rates, or withholds are criticized for encouraging undertreatment. Yet, most studies conclude that more people access mental health treatment under HMO capitation arrangements than under fee-for-service arrangements. Although the treatment intensity is less, outcomes appear to be the same (Mechanic et al. 1995, p. 19).There is some suggestion that the seriously mentally ill do not fare as well in the HMO setting, but research design limitations prevent generalization.

Psychologists point to increased utilization of social workers as evidence of managed care's interest in savings over quality. Yet there are no studies to suggest psychologists' results are superior. In a survey conducted by Consumer Reports (Seligman 1995, p. 965), patients reported equal satisfaction with outcomes whether psychologists, social workers, or psychiatrists provided treatment. Because there are only 70,000 psychologists in the USA, but over 189,000 social workers (Mills 1997) who provide services at an average of $15 less per session, social work referrals improve treatment access.

The debate over managed care's impact on treatment quality is unlikely to be 'won' because treatment outcomes in mental health are difficult to define and measure (Marques et al. 1994, pp. 22–9). Variations in treatment practices, inconsistent use of diagnostic coding, the cyclical nature of mental health symptom exacerbation, and a lack of professional consensus regarding the definition of a positive mental health outcome make it difficult to determine what works and what doesn't. Patient satisfaction with treatment, although measurable, is not correlated with outcome. Patient self-reports of improvement are inconsistent with providers' perceptions of treatment completion. Despite these limitations, research has consistently determined that mental health treatment works and that managed care has not caused deterioration in these results. While patient and provider satisfaction has diminished, the quality of care has not.

2.3 Detracting from Practitioners' Professional Autonomy to Advocate for Patients

Providers resent the imposition on professional autonomy associated with medical necessity review. It is popular to complain of clerical oversight of professional judgment. Although it is possible for unlicensed individuals to authorize or approve recommended treatment according to a professionally developed treatment guideline, the industry standard is that a clinical peer or board-certified psychiatrist renders all nonauthorization decisions. Many states legislate peer review and both NCQA and the Utilization Review Accreditation Commission (URAC) require it.

Fears of 'economic credentialing' pressure clinicians to consider financial factors when recommending treatment, conflicting with professional ethics of patient advocacy. Providers are concerned that only those who treat individuals 'cheaply' will be retained in a managed care network, while providers who favor longer-term treatment, and potentially better outcomes, will be excluded. Yet few providers are removed from managed care networks for patterns of excessive overutilization. The preauthorization requirement often detects and prevents overutilization before it occurs. More common causes for provider termination are noncompliance with credentialing requirements, or provider requests for network removal. Access requirements make narrowing a managed care network solely on utilization unlikely.

2.4 Perceived Lack of Accountability

Managed care reform began to appear on state legislative agendas in 1995 and on the congressional agenda in 1996. Providers organized to fight managed care at the state level. Symbolically targeting 'gag clauses,' physicians began talking to the media, raising consumer anxieties about the quality of health care provided by profit-seeking managed care companies. Despite research to the contrary, the public became convinced through anecdote that managed care was destroying US health care.

Public concern about a perceived lack of managed care accountability has resulted in a proliferation of regulation in the last five years. All 50 states have passed some form of managed care legislation. Federal patient rights legislation has been introduced in Congress every year since 1997. NCQA introduced MBHO quality standards in 1996.

Mental health providers, frustrated with managed care's impact on their practices, are diversifying by providing consultation, education, and forensic services. 'Cash only' practices are emerging, with promises of enhanced privacy due to the absence of managed care oversight. For managed behavioral care to continue to evolve in the new millennium, it must repair its relationship with providers. A patient-centric model of managed behavioral care provides a framework in which practitioners and managed care can partner to achieve the shared goal of improving mental health in the USA.

3. The New Millennium: Patient-centric Care Management

In the patient-centric model of managing behavioral care, an MBHO hires clinician-care managers who:

improve access to needed services;

match patients to providers who have the skills and experience most likely to result in treatment success;

maximize benefit allocation through medical necessity review of recommended services;

enhance treatment outcomes through intervention with ineffective treatment plans;

sustain treatment gains through aftercare monitoring; and

identify or prevent undetected mental illness.

The patient-centric model of managing care is based on the premise that helping patients overcome an emotional disturbance quickly and keeping them well is less expensive than allowing patients to prematurely abandon treatment and subsequently return to intensive levels of care. Promoting early access to effective providers is consistent with this goal.

In a patient-centric model, individual benefits are managed in the context of the well-being of an insured population. Ensuring that care is only reimbursed when medically necessary, that a patient is treated in an effective and less costly setting, and that patients are referred to practitioners with discounted rates maximizes the value of patients' benefits. Anticipation of benefit exhaustion and transition planning prevents individuals who cannot afford needed care from going without it.

Removal of barriers to mental health treatment is the ultimate value managed care can bring to the therapeutic process. Transportation problems, poorly coordinated care, or identification of contraindicated drug combinations are the clinician care manager's province, allowing the practitioner to focus on treatment interventions. By comparing recommended treatment to practice guidelines based on outcome research, managed care can promote the updating of providers' treatment repertoires. Following treatment, periodic monitoring of a patient's condition or follow-through with an aftercare plan can be effective in sustaining treatment gains. Joint pursuit of these patient-centric objectives permits the consumer to take advantage of treatment and care management resources, while reducing the cost of ineffective or unnecessary care.

4. Conclusion

In order to continue to reduce the cost of mental health care while maintaining or improving its quality, managed behavioral health care systems must continue to evolve and adapt. A managed care–provider partnership is unlikely to be easily achieved. Although mental health outcome research improved dramatically over the last 20 years of the twentieth century, there is still much professional disagreement about 'what works.' These debates are often the substance of discussion between providers and care managers. At the heart of the animosity, however, is the providers' perceived loss of professional dignity and autonomy. The next evolution in managed care will not occur unless the providers' sense of professionalism is restored. A patient-centric model of managed behavioral care aligns the goals of managed care employees, mental health providers, and purchaser-employers with the best interests of the patient population.

See also: Health Care Delivery Services; Health Care: Legal Issues; Health Care Organizations: For-profit and Nonprofit; Managed Care

Bibliography

Cummings N A, Dorken H, Pallak M S 1993 The impact of psychological intervention on health care costs and utilization. In: Cummings N A, Pallak M S (eds) *Medicaid, Managed Behavioral Health and Implications for Public Policy*. Foundation for Behavioral Health, San Francisco **2**: 1–23

Hennessy K D, Green-Hennessy S 1997 An economic and clinical rationale for changing utilization review practices for outpatient psychotherapy. *Journal of Mental Health Administration* **24**(3): 340–9

Iglehart J K 1999 Tumult and transformation. *Health Affairs* **18**(2): 6

Luft H S 1999 Why are physicians so upset about managed care? *Journal of Health Politics Policy and Law* **24**(5): 957–66

Marques C C 1998 The evolution of managed behavioral health care: A psychologist's perspective on the role of psychology. In: Bellack A S, Hersen M (eds.) *Comprehensive Clinical Psychology* **411**: 410–18

Marques C, Geraty R, Harbin H, Hoover K, Theis J 1994 Quality and access in the managed behavioral healthcare industry. *Behavioral Healthcare Tomorrow* **3**(5): 22–9

Mechanic D, Schlesinger M, McAlpine D D 1995 Management of mental health and substance abuse services: State of the art and early results. *Milbank Quarterly* **73**(1)

Mihalik G, Scherer M 1998 Fundamental mechanisms of managed behavioral health care. *Journal of Health Care Finance* **24**(3): 1–15

Mills M 1997 US has 114 behavioral health professionals per 100,000 population. Open Minds Newsletter **September** p. 6

Mullen J K 1995 *Intro to Managed Care*

Mumford E, Schlesinger J H, Glass V G, Patrick C, Cuerdon T 1984 A new look at evidence about reduced cost of medical utilization following mental health treatment. *American Journal of Psychiatry* **141**(10): 1145–58

Patterson D Y, Sharfstein S S 1992 The future of mental health care. In: Feldman J, Fitzpatrick R J (eds.) *Managed Mental Health Care: Administrative and Clinical Issues*. American Psychiatric Press, Washington, DC, pp. 335–46

Seligman M E 1995 The effectiveness of psychotherapy: The consumer reports study. *American Psychologist* **50**(12): 965–74

Studdert D M, Sage W M, Gresenz C R, Carole R, Hensler D R 1999 Expanded managed care liability: What impact on employer coverage? *Health Affairs* **8**: 7–27

Satcher, D vs. Surgeon General 1999 Financing and managing mental health care. In: *Mental Health: A Report of the Surgeon General*. Department of Health and Human Services. Washington, DC

L. D. Weaver and C. C. Marques

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Management Accounting

1. Introduction

The Chartered Institute of Management Accountants in the UK defines management accounting as an integral part of management concerned with identifying, generating, presenting, and interpreting information used for

(a) formulating strategy,
(b) planning and controlling activities,
(c) decision taking,
(d) efficient resource usage,
(e) performance improvement and value enhancement,
(f) corporate governance and internal control,
(g) safeguarding tangible and untangible assets.

While the discipline finds its roots in industrial accounting concerns and factory-based cost accounting, it has undergone an important transformation over the past twenty years. The field presently continues to undergo significant changes and remains in a dynamic state of flux.

During the 1980s, many organizations in Western countries perceived a need to alter their management practices. A variety of factors were responsible for this. In part, at the international level, competitive manufacturing methods used by major Japanese corporations and by those of newly industrialized countries were seen as superior to those used in the west (Bhimani and Bromwich 1996). To an extent, this was ascribed to sociocultural characteristics of Japanese workers from factory-floor operators to senior executives. But what was viewed as more significant was the application of flexible manufacturing technologies by the Japanese to produce a greater diversity of high-quality, low-priced products more quickly. Specific modes of work organization, production systems and motivational incentives were seen to differ radically from those used in the West. These forces of change on management approaches have also been viewed as having had important implications for management accounting practices.

The adoption of flexible technologies in manufacturing and service-oriented enterprises, including manufacturing and enterprise resources planning systems, and features of computer-integrated manufacturing, led to conventional accounting methods for cost allocation being questioned. Whereas the manufacturing, environment had become more automated, more specialized, more flexible, and requiring larger capital investments, methods of costing products continued to operate under the conventional notion that company operations were geared toward repetitive production of homogeneous products relying on labor-based activities. Management-accounting techniques were thus seen to be antagonistic to the realities of the new manufacturing environment. Similar critiques were being made of cost-accounting methods in service industries.

Other calls for change in management accounting were also being voiced. North American academics thought that from the post-First World War period, progress in management accounting had virtually ceased, principally as a result of the growing importance of the external reporting priorities, especially in the area of stock valuation over those of internal organizational accounting. Some critics of traditional management accounting asserted that management accounting had, for over seventy years, been subordinated to the requirements of financial-statement preparation for external parties (Johnson and Kaplan 1987). Advocates of this school have supported calls for management-accounting reforms based on the argument that the field must once again be allowed to develop in its own right, rather than remain subordinate to financial accounting.

Many commentators made more substantive arguments, focusing on the propriety of accounting practices, given the roles that management accounting is intended to fulfill. The usefulness of complex methods of organizational costing and pricing, in the face of the belief that the price of a product must ultimately meet the expectations of the marketplace rather than those of its producers, have been questioned. Others have given weight to the argument that management accounting has for too long remained isolated and divorced from other enterprise functions. Conventional accounting practices have put into place channels for information flow and routes for data exchange that have rendered some organizations static, inflexible, and excessively structured, especially in the case of firms operating in dynamic and fast-changing markets. Calls have thereby been made for cost-management tools appealing to wider rationales of enterprise management.

2. Onward with Change

During the 1990s, many organizations pursued a vigorous management change agenda. Until 1994, many organizations shied away from investing in novel enterprise management philosophies because of the malaise experienced in the worldwide economy. Nevertheless, cost-containment programs were widely implemented. Techniques such as activity based costing, throughput accounting, target cost management, life-cycle costing, and the greater use of non-financial performance measures were seen as offering the potential of enhancing value creation and aiding cost-containment (Bromwich and Bhimani 1994, Yoshikawa et al. 1993). From 1995, as the economic engine of major western economies began to generate more steam, organizational searches for wider reaching management technologies grew apace once more. Such pursuits were, however, tempered with a more cautious approach to large-expenditure projects on

altering ways of managing. Moreover, accounting techniques for aiding managerial tasks appeared more cognizant of a need to interlink enterprises' various value-creation activities. Thus, management-accounting technologies stressed processes, as opposed to functions, and emphasized linkages between traditional activities, as well as more flexible organizational structures.

Notions of the impact of brand management on cost accountings were highlighted. Attempts to link organizational learning to customer responsivity and financial tracking were made via establishing novel performance-measure control approaches. Whereas traditional conceptions of cost-accounting techniques for scorekeeping purposes were shunned, achieving more balanced scores came to be deemed important (Kaplan and Norton 1996). Likewise, whereas 'value added' was a significant though short-lived endeavor among proponents of more comprehensive financial reporting in the 1970s, the pursuit of a trademarked notion of economic value added surfaced (Stern et al. 1995). Certainly, the need to be strategically perceptive in all things management accounting was widely perceived during the mid-1990s. The US-based Institute of Management Accountants changed the name of its flagship magazine *Management Accounting* to *Strategic Finance*. Likewise, the Canadian Society of Management Accountants revamped its intellectual capital whereby its published products came to be grouped as a 'strategic management series' package. No doubt many organizations, in both Europe and North America, consider themselves as engaging and supporting strategic management-accounting activities while not appealing to any single definition of the term (Bhimani and Keshtvarz 1999, Guilding et al. 2000).

3. *Strides in Management Accounting*

Perhaps it is well to turn to some significant technologies of management accounting the seeds of which were planted during the 1980s and germinated and flowered in the 1990s. Two of these provide the focus presently.

3.1 *Target Costing*

Target costing indicates for each product (and each cost element of that product) the cost reductions (over time) that must be achieved by the firm and its suppliers in order to ensure that the company's strategy is realized throughout the product's life-cycle, while maintaining a competitive market price for the product offered (Bromwich and Bhimani 1994).

A preliminary step in this process is for the enterprise to come to a view over the whole product life-cycle as to the product strategy in terms of the product attributes to be offered relative to competitors, including further options, quality and after sales services, the technology to be used, the desired profit, and market share. A comparison with expected prices in the market over the life cycle of the product determines the price at which the product with the planned characteristics can be offered in the market and be expected to obtain their target profit and market growth.

Target costing in some Japanese companies is seen as a cost-reduction exercise that seeks to achieve cost reduction in the preproduction stages. It seeks to force designers to reduce costs, though how far target prices should seek to include strategy, and how far it should be seen predominantly as a continuing cost reduction device is a firm-specific issue. Such an approach may be supplemented by levying extra costs on nonstandard parts and excessive tooling, as well as penalizing the use of nonstandard parts or new machines. The procedure contains a number of other lessons. One is that prices are determined in the market and reflect competitors' strategies and customers' demands. Target costing also requires that companies collect and use external information that may not be available or, if available, may not ordinarily be assessed by the Finance department. Finally, it illustrates the possibility of incorporating strategy into accounting and performance measurement more generally.

There are a number of ways of setting target costs, these include:

(a) determine the difference between allowable cost (selling prices–desired profit) and forecast costs determined using current costs,

(b) determine the difference between allowable cost (using planning selling price net of the required return on sales) and current costs,

(c) apply the desired cost reduction to current costs

The above description of target costing should be seen as broadly representative of what ought to be done rather than the organizational reality. There is evidence that Japanese firms often fail to make the target cost, at least prior to manufacture, and either adjust the target costs or decide not to continue with the product in its existing form. Target costing practices in Western enterprise likewise evidence a high degree of organizational specificity.

3.1.1 Tools for target costing. In Japan, value engineering (VE) is an activity which helps to design products which meet customer needs at the lowest cost while assuring the required standard of quality and reliability. A number of tools are generally used in this activity.

3.1.2 Functional analysis. The aim is to determine the functions of the product and indicate how far the design of these functions will satisfy the consumer (as shown by market research) and at what cost relative to that part of the target cost pricing by elements

allocated to that function. This allows the variance between the target cost-per-function and its current cost to be calculated. It also allows the contribution which a function makes to consumer needs to be ascertained relative to its cost, thereby leading to redesign where two items are not well matched (see Yoshikawa et al. 1993).

3.1.3 Cost tables. Cost tables are detailed simulated databases of all the costs of producing a product, either for product functions or product components, based on a large variety of assumptions. They allow 'what if' questions to be posed relating to decisions to change any aspect of the product. A cost table will show the variety of cost performances when using different types of materials for a given component. The data in cost tables will be accepted by all members of the organization, and therefore ensure common thinking throughout. There will be a cost table for each part of the product's lifecycle.

3.1.4 Value engineering collections. These are libraries of case studies of the detailed experience from previous VE activity, the use of which is to improve future target costing and VE. Of course, all the other well-known techniques used in Japan, such as just-in-time methods and quality circles, are utilized, where appropriate, in target-costing exercises.

Enterprise cost structure in Japan is such that, on average, material and components account for some 60 percent of cost, direct labor some 14 percent, and overheads some 24 percent (Bhimani and Bromwich 1996), though these percentages can vary substantially across firms and industries. Comparative US figures are material and component costs (53 percent), direct labor (15 percent) and overheads (32 percent). Possibly, high overhead cost incursions could be a signal that for some firms, much may be gained from the market-derived discipline that is evident in the target-cost-management process.

3.2 Activity Costing

Concerns about overhead allocation have been voiced by many managers and, perhaps more importantly, the need to account in a way that encompasses the activities which cause costs rather than using traditional methods of overhead allocation is recognised by professional accounting bodies.

Activity based costing (ABC) emerged as a cost-management approach during the late 1980s. It can be defined as an approach to costing which focuses on activities as the fundamental cost objects. It uses the cost of these activities as the basis for assigning costs to other cost objects such as products, services or customers (Horngren et al. 2001). During the 1990s, it has increasingly been linked to managerial decisions via time, market, strategy, and quality issues. Examples of ABC calculation can be found in Bromwich and Bhimani (1994) and Yoshikawa et al. (1993).

The published evidence on the payoffs and costs of installing activity-based cost and management information systems is not unequivocal (Bromwich and Bhimani 1994).

Partly, this is because seeking an association between major changes in organizational and management structures and management accounting is a difficult and time-consuming exercise. The practicing management accountant must be satisfied that the benefits likely to accrue from any change will outweigh its costs, though cost-benefit quantification is not a straight forward proposition. Moreover, abstract, theoretical concepts do not tend to succeed in achieving implementation in organizations. To date, there is little evidence that any new accounting method that attempts to facilitate the management decision-making process actually increases the ability to generate greater profits, though providing such evidence is very difficult.

Ascribing increased profitability to the implementation of a costing technique, such as ABC, is a difficult exercise. Problems arise in tracing changes in the 'bottom line' to managerial action associated with altered costing information. The inability to isolate the effects of a single change such as the adoption of ABC *ceteris paribus*, and to use another company as a 'control' for comparative purposes, makes any assertion of a profitability link dubious. This is especially so because changes in organizations often come in bundles, after, for example, the appointment of a new managerial team. Perhaps the ultimate test of establishing whether a switch from a traditional cost system to ABC yields increased benefits is a market test that seeks to ascertain whether such a switch becomes a permanent one for a sufficiently large number of companies adopting ABC. Given the relative novelty of this costing approach, however, this test is difficult to apply. Nevertheless, in various surveys of ABC/ABM implementations in UK firms, Innes et al. (2000) report that, whereas in 1995 13.2 percent of companies rejected the technique, 15.3 percent did so in 1999. They conclude:

> All in all, the survey results are not indicative of any further growth in ABC/M but rather suggest at best a levelling off in the extent of its use in our largest companies.

There are many managers who believe that activity-based accounting methods have changed their approach to decision-making, indirectly yielding increased profits and, more importantly, a much better understanding of their business (see Cooper and Kaplan 1999, for case studies). Some problems still exist with overhead allocation where, for instance, the source of the overhead is difficult to trace or is sunk or is common to diverse activities. Arbitrariness and

judgment will, therefore, continue to significantly affect overhead cost accounting. There is strong support, however, for the argument that dynamic and changing manufacturing (and non-manufacturing) environments render redundant, conventional cost-accounting systems that emphasize information priorities other than those emerging as virtually indispensable in today's competitive marketplace. For example, strategic cost analysis, cost-benefit appraisal for long-term projects, life-cycle consideration, quality, and customer satisfaction data are likely to become increasingly meaningful in service and manufacturing environments (Bromwich and Bhimani 1994).

4. *The Future of Management Accounting*

While many other management-accounting approaches have been, and continue to be, developed, aside from activity accounting and target costing, some concerns have been voiced concerning the future fate of management-accounting practices. King (1997), for instance, believes that many management accountants complain that—while their skills may be valued—they are often excluded from management decision making. Accountants are often not seen as starting players on the management team, in part because of their training, which emphasizes precision and accuracy at the expense of relevance. King (1997) suggests that recognizing that there is more judgment—and less precision—in accounting is a first step in what may have to be an extensive program to ensure that accountants will survive into the twenty-first century.

Foster (1996), likewise, believes that change is required in the way in which management accountants operate. Management accounting is in a continual state of adaptation as the knowledge base in the field increases and the business environment changes. According to Foster 1996), management accounting has often been portrayed as focusing too much on internal reporting and on providing information for statutory financial reporting. This portrayal is seen as inadequate now, and after 2000 it will be even more so. In the future, decisions about which activities to outsource, and how to structure joint ventures with other organizations will be critical management responsibilities. Management will seek financial information to help plan and monitor ongoing relationships with external partners. Foster (1996) notes that, in the future, we will seek greater recognition of root-cause cost drivers as well as cost drivers arising within and across functions.

According to Flegm (1996), skill with numbers is not enough. Accountants of the future must be experts in the basic disciplines of business, yet also generalists who can manage different disciplines, communicate with clients, and motivate employees. The call for more strategic management accountants, likewise, still rings loud. Kaplan and Norton (1996, 2000) considers that advances in the 1990s have made it possible for management accountants to become part of their organization's value-added team. Management accountants must participate in the formulation and implementation of strategy (Kaplan, 1995), and help translate strategic intent and capabilities into operational and managerial measures.

Management accountants have to move away from being scorekeepers of the past and to become the designers of the organization's critical management information systems. Existing performance measurement systems, including those based on activity accounting, focus on improving existing processes. The balanced scorecard, according to Kaplan (1995), by contrast, focuses on what new processes are needed to achieve breakthrough performance objectives for customers and shareholders.

This call has been voiced in parallel to that for altering management accountants' patterns of behavior (Johnson 1995). In the past, management accountants focused on how to disclose and report financial accounting numbers, especially cost numbers, so that managers could draw from them more useful information. According to Johnson (1995), instead of focusing on achieving targets, companies intent on achieving positive long-term results should focus on mastering disciplined and standardized patterns of behavior. Management accountants should work to create channels through which people inquire openly about purpose and method. In this light, Borthick and Roth (1997) note that the management-accounting systems of the future will be a large, real-time database whose data managers anywhere will be able to retrieve, manipulate, and analyze for new problems as they arise. The choice that companies face is whether to be startled by what the future brings or to embrace the future by implementing organization-wide database systems to reap the benefits of better data access.

What seems evident is that each decade brings with it pressures for changing management accounting, as well as different perceptions of the implications of such pressures. The late 1990s heralded in the New Economy where past conceptions of proper organizational management are viewed as being in need of change. A renaissance of liberalization from the past is deemed essential by many old industrial firms facing competition by new giants that did not exist only five years before. The Internet economy has made 'modern' management techniques ancestral. Bricks and mortar cardinal principles are being eschewed and replaced by evolving bricks and mortar truths. Life-cycle costers are now to move at an Internet time pace. Balanced-scorecard adopters must tackle new forms of imbalance. Activity accounting systems must begin to appeal to cyber-based cost drivers. Target cost management must now capture the nuances of ephemeral web markets. While little can be said of where—if

anywhere—the future will take management accounting, pundits will continue to dwell on what is essential for its well being.

See also: Budgeting and Anticipatory Management; Financial Accounting and Auditing; Information and Knowledge: Organizational; Innovation: Organizational; Intelligence: Organizational; Management: General; Organizational Decision Making; Quality Control, Statistical: Methods; Strategy: Organizational; Supply Chain Management; Technology and Organization

Bibliography

Bhimani A, Bromwich M 1996 Management Accounting: Emerging Pathways. In: Drury C (ed.) *Management Accounting Handbook*. Butterworth-Heinemann, Oxford, UK, pp. 6–19

Bhimani A, Keshtvarz M 1999 British Management Accountants: Strategically Inclined? *Journal of Cost Management* **13**: 25–31

Borthick A F, Roth H P 1997 Faster access to more information for better decisions. *Journal of Cost Management* **11**: 25–30

Bromwich M, Bhimani A 1994 *Management Accounting: Pathways to Progress*. CIMA, London

Cooper R, Kaplan R S 1998 *The Design of Cost Management Systems*. Prentice-Hall, Englewood Cliffs, NJ

Flegm E H 1996 The future of management and financial accounting. *Journal of Cost Management* **10**: 44–9

Foster G 1996 Management Accounting in 2000. *Journal of Cost Management* **10**: 36–39

Gulding C, Cravens K S, Tayles M 2000 An international comparison of strategic management accounting practices. *Management Accounting Research* **11**: 113–167

Horngren C, Bhimani A, Foster G, Datar S 1999 *Management and Cost Accounting*. Prentice Hall, Hemel Hempsted, UK

Innes J, Mitchell F, Sinclair D 2000 A tale of two surveys. *Research Update: The Newsletter of CIMA Research* p. 4

Johnson H T 1995 Management accounting in the 21st century. *Journal of Cost Management* **9**: 15–20

Johnson H T, Kaplan R S 1987 *Relevance Lost: the Rise and Fall of Management Accounting*. Harvard Business School Press, Boston

Kaplan K S 1995 New roles for management accountants. *Journal of Cost Management* **9**: 6–14

Kaplan R S, Norton D P 1996 *The Balanced Scorecard*. Harvard Business School Press, Boston

Kaplan R S, Norton D P 2000 *The Strategy-focused Organisation: How Balanced Scorecards Thrive in the New Business Environment*. Harvard Business School Press, Boston

Kato Y 1993 Target costing support systems: Lessons from leading Japanese companies. *Management Accounting Research* **4**: 33–48

King A M 1997 Three significant digits. *Journal of Cost Management* **11**: 31–7

Stern J, Stewart G, Chew D 1995 The EVA® Financial Management System. *Journal of Applied Corporate Finance* 32–46

Yoshikawa T, Innes J, Mitchell F, Tanaka M 1993 *Contemporary Cost Management*. Chapman and Hall, London

A. Bhimani

Management: General

Management describes the complex behavior of those responsible for the decisions that determine the allocation of resources, both human and physical, within an organization. The role of general managers is a multifunctional role: decision maker, strategic leader, and human relations expert, custodian of institutionalized values, mediator between business and society, negotiator and power broker. It is played in a set of institutional contexts and implies specific cultural attitudes and styles of action. According to the simplified model of firm behavior under perfect competition, the managerial function is limited to a set of highly rigid decision rules for determining price and output. In real market situations which deviate from the perfectly competitive, market forces are only one of the variables in the complex context where business management takes place. Other key variables concern the corporate organization, the larger group of organizations in which the firm is embedded, state policies and legal systems, and the stategies of both internal and external stakeholders of the firm. The general management function is therefore a very complex one, that includes a set of managerial roles and implies various forms of managerial behavior and style.

For specific areas of management see articles on *Administration in Organizations*; *Corporate Finance: Financial Control*; *International Marketing*; *Administration in Organizations* and *Organizational Decision Making*. The focus of this article is on general management, on managerial behavior and managerial roles in organizations. More specifically, four main questions will be discussed: (a) the rise of managerial capitalism and the separation of ownership and control; (b) what managers really do: the research on managerial behavior; (c) the role of managers in business organizations; (d) managerial ideologies and managerial styles.

1. The Rise of Managerial Capitalism and the Separation of Ownership and Control

The growing importance of management in the twentieth century is related to the transformation of capitalism and to the emergence of large business organizations, in the form of public companies, that implies the separation of ownership and control. This transformation has been due to a variety of factors, such as technological innovation, generalization of market relations to preindustrial settings, international competititon, industrial concentration, and vertical integration. The change has resulted in increasing organizational complexity, multifunctional structures, greater role differentiation, and more decentralized authority.

The transition from family capitalism to managerial capitalism was facilitated by the legal institution of the

limited company; it was related to the timing and sequence of the industrialization process in different countries and took place earlier in the UK and the USA than in Continental Europe and Japan. Shareholding has often become very dispersed. Boards of directors have been predominantly composed of full-time senior executives who nominate new members and take decisions concerning high management appointments and remuneration. Top management has become effectively responsible not only for operating organizations, but also for strategic decisions concerning investment, finance, internal growth and diversification, profit retentions, and acquisitions. Business firms have displayed a capacity for persistent, although irregular, long-term growth in size, as shown by the growing share of the 100 largest corporations in total industrial value-added, which from less than 10 percent in 1900 grew to 30–50 percent in various industries in the year 2000.

The growth of the large corporation and the increase in capital concentration, however, are not inconsistent with coexistence of other components of the modern economy, where traditional capitalist modes (markets containing large numbers of small owner-managed businesses), and new forms of productive systems (industrial districts with specific competitive-cooperative patterns among producers in the same industry) not only survive, but effectively compete and grow. In addition, in the last two decades of the twentieth century, technological developments, such as personal computers, which increase the comparative efficiency of small organizational units, and new institutional developments, such as corporate divestitures and management buyouts, contribute to checking and even reversing concentrating tendencies.

The large corporation and managerial capitalism are neither a general pattern of business organization, nor necessary steps in economic modernization, but they have marked first the economic history of the UK and the USA and later, that of continental Europe and Japan. The 'complete businessperson'—who was at the same time owner, entrepreneur, and manager—slowly gave place to the separation of the property and control functions and to a plurality of specialized roles within the corporation.

Both Marx and Schumpeter saw the separation of the managerial role from the ownership of capital as a sign of the crisis of capitalism. The limitations of both Marx's and Schumpeter's views in this regard is due largely to their belief that the competitive economy of the individual entrepreneur is the only brand of sustainable capitalism. In reality, capitalism has proved compatible with the existence of large, complex firms, and with state intervention and regulation of the economy. What Marx and Schumpeter saw as a crisis factor, other scholars viewed as a stage in capitalist development and a feature of the fundamental process of the growth of the business firm. Influenced by Weber's analysis of the rise of the modern state, Rathenau (1918) argued that with the depersonalization of property and the rise of '*organisierte Kapitalismus*,' the firm followed the same path as, and became similar to, the modern state.

Berle and Means (1932), in their pathbreaking empirical research on two hundred US firms, argued that the separation of ownership and control was an organizational requirement of the modern corporation and an irreversible trend of economic life. Berle and Means conceived the modern corporation as an organized social group based on the interdependence of different economic interests—those of the owners, the employees, the consumers, and the controlling group. This last group has the 'active property' of the firm—as distinct from the 'passive property' of the shareholders—and performs the entrepreneurial function. Managerial power is highly concentrated in the controlling group, as religious power was in the medieval church and political power in the nation state. Conflicts of interest can arise between managers-entrepreneurs and shareholders; the former are mostly interested in reinvesting most of the profit, in order to enhance the firm's competitive power and long-term development, while the latter want short-term economic gains. A more radical view of this conflict was Burnham's (1941), who foresaw a 'managerial revolution.'

Ideas similar to those of Berle and Means's were developed by Coase (1937), Williamson (1975), and Chandler (1962, 1977), who focused on the firm as an alternative allocative mechanism to the price mechanism, and of the 'visible hand' of management as an alternative to the 'invisible hand' of the market. Chandler, in particular, argued that a fundamental reason for the success of the modern corporation was its ability to obtain higher productivity rates by internalizing activities previously performed by separate autonomous business units, by reducing information and transactions costs, and by better coordinating the flow of goods and factors of production. But the advantages of corporate organization could not be fully exploited until a managerial hierarchy was created, made of managers who are rewarded, selected, and promoted on the basis of their professional skills and performances rather than on the basis of their family ties and the money they have invested in the firm. Once formed, this managerial hierarchy becomes the source of continuity, power, and growth in the large firm, and tends to assume a life of its own, independent from the individual turnover at the top.

According to Chandler, the dividing line between entrepreneur and manager is drawn on the basis of the type of decisions they make: the minority of the general executives and owner-entrepreneurs—who are responsible for strategic decisions—are neatly separated from the majority of sectoral managers—who are in charge of specific processes and perform routine activities. Managerial roles in the modern corporation

are located along a hierarchy of authority and responsibility. The control of finance, personnel, organization, research, and information is a necessary condition, but is not enough to define a full entrepreneurial role; the role implies also the ability to work out long-term general strategies and the power to allocate the resources necessary to achieve the desired goals.

The thesis of the irreversible diffusion of managerial capitalism has been criticized on various grounds. First, it has been pointed out that managerial control applies only to large firms, whereas, as I remarked earlier, the 'other half' of the contemporary economy is made of small firms, most of which are managed by owner-entrepreneurs. Second, even for large firms, family capitalism is much more resilient than the scholars of managerial control would admit. This seems to be the case not only in Continental Europe and Japan but also in the USA, as Burch (1972) tried to show through the survey he made 40 years later of the same firms that Berle and Means studied.

Third, it has been argued that conflictual views on short- and long-term strategies actually vary according to the conception of corporate control developed by high-level executives. The finance conception of control that was pioneered by US conglomerates in the 1980s provides an instance of short-run gain strategy upheld by top managers (Fligstein 1990). There are many instances of top managers and owners basically sharing the same interests and values, in spite of the different roles they play, due to the fact that corporate executives, while not often among the largest shareholders, receive incomes that are highly correlated with stock performance, and that a relevant part of their remuneration is the form of stock, dividend income, and capital gains (Lewellen 1971).

Fourth, scholars like Galbraith (1967) argue that leadership in the large business firm is no longer confined to entrepreneurs, or to top managers; but it is exerted by a technostructure, that is, by all those participants who have specialized knowledge and participate in collective decision making. Technologically advanced and organizationally diversified corporations cannot be governed by single entrepreneurs, or a few top executives; they must be managed by the larger technostructure, or by Meynaud's (1964) technocrats.

Fifth, the idea of the firm as an alternative allocative mechanism for the price mechanism has been criticized for neglecting other important mechanisms in the regulation of economic activity, such as clans, cartels, and trusts, employers, and trade associations (Streeck and Schmitter 1985).

Finally, the very notion of corporate control has been challenged by the studies of interlocking directorates. Mizruchi (1982) and Stockman et al. (1983), among others, argue that public companies, with ownership dispersed among many stockholders, are actually controlled by financial institutions—such as national commercial banks and insurance companies—and through corporate interaction patterns rather than by their top managers. Mintz and Schwartz (1985) found a complex division of labor among financial and nonfinancial institutions. Distinguishing among New York money market commercial banks, regional banks, and national insurance companies, they portray a system where financial institutions simultaneously unite and divide. The densely interlocked groups at the local level—centered around the largest regional banks—are connected to the larger money market banks by a series of bridging links created by the country's largest insurance companies. Control of capital flows enables banks to constrain the actions of nonfinancial corporations and to pursue to some extent centralized decision making.

A related view is that of Useem (1984), who views the formation of an intercorporate management network as the the driving force in the rise of institutional capitalism, that is, the classwide organization of business. The dispersion of ownership and the superior capacity of managerial hierarchies in large firms were the engine behind the transition from family to managerial capitalism. The creation of transcorporate networks of ownership and directorships—as an extension of individual corporate strategies—generated the change for the emergence of classwide organization. Prerequisites for promotion of top executives are not only decision-making and other abilities within the firm, but also demonstrated leadership outside it; prospects of promotion to top positions are generally enhanced if the the aspirant serves effectively on other company boards, plays a relevant role in business associations, and achieves a reputation for having access to higher government posts and for effectively managing the relations with relevant stakeholders of the firm.

Some studies argue that in today's global market, with chief executives operating across national boundaries in transnational corporations and financial networks, managerial capitalism is changing into international capitalism (Vernon 1971, Makler et al. 1982); an international business class, made up of manager-entrepreneurs with interconnected careers and a common cosmopolitan culture, plays a key role in the world economy (Bottomore and Brym 1989); and corporate power stems from the ability to control global production networks and financial flows (Held et al. 1999).

2. *What Managers Really Do: The Research on Managerial Behavior*

After Fayol's pioneering study (1916), research on managerial behavior grew in the 1950s. Early studies, described as the Work Activity School by Mintzberg (1973), recorded and classified the work activities of managers using simple categories. Carlson (1951)

organized a diary analysis of the activities actually performed in four weeks by seven Swedish and two French top managers, supplemented by extensive interviews and by background knowledge. It was a participating research, since it grew out of a long-standing discussion group of Swedish managing directors with Carlson, aimed at improving selection and training procedures for top management posts. The study showed that, with a few exceptions, managers spent most of their time talking and listening (up to more than 90 percent of their time). Most of their activities, apart from formal meetings, were brief, fragmented, and interrupted; they rarely had as much as half an hour a day alone. Managers were exposed to 'administrative pathologies,' such as being overloaded with details or neglecting personal inspection tours.

A second type of empirical research is well exemplified by Mintzberg's (1973) study of the weekly activities of five chief executives in different kinds of organizations. He classifies 10 managerial roles into three main headings: informational, interpersonal, and decisional. Interpersonal roles include figurehead, leader, and liaison, and also negotiator and spokesperson. But the key function is the monitoring and dissemination of information, which makes executive managers the nerve center of the organization and, we can add, provides a major basis for their power. Top managers' foremost professional hazard is the danger of dealing with all issues at a superficial level, because of the variety and brevity of their activities.

A third kind of studies of managerial behavior focuses on differences between managerial jobs. Stewart (1982, 1991) is a good example. She stresses different managerial styles and the constraints, opportunities, and choice models in managerial activity; and she identifies as main criteria of differences the nature and the difficulty of the contacts required in the job, whether the job was a boss-dependent one, and whether the job was an exposed one where the jobholder's mistakes were personally identifiable.

Kotter's (1982) study, based on lengthy interviews and documents of 15 general managers in various companies and their key contacts, introduced agendas and networks as key concepts. He found out that general managers tend to build their agendas—made of loosely connected goals and plans—during the first six months to one year in the job and through the development of a network of contacts. Experienced managers may have the problem of 'tunnel vision,' and therefore require occupational mobility in order to mix with people from various businesses and with different perspectives.

All these studies, and many similar others, contributed to show the complexity of managerial behavior and, at the same time, the difficulty of identifying what is specific about it. Whitley (1989) gives a good summary of what can be considered managerial about managerial activities. He suggests five main features: They are highly interdependent, contextual, and systemic; they are relatively unstandardized; they are changeable and developing; they combine both the maintenance of administrative structures and their changes; they rarely generate visible and separate outputs which can be directly connected to individual inputs.

The limitation of many of these studies is that they underplay the specific organizational contexts where managerial behavior takes place. An exception in this respect is Hannaway's study (1989), based on almost 30,000 random samples of the work activities of 52 managers, divided into three hierarchical levels, over six weeks. She was interested in how managers behave in a hierarchical social system that is characterized by uncertain tasks, ambiguous outcomes, and biased feedback. Among her most interesting findings are the attempts of managers to protect themselves from blame when things go badly by involving others in any potentially risky decision, and to protect themselves from the ambiguity of managerial work signaling their worth to others. Hannaway's study introduces the research topic of the role of managers in the organization.

3. *The Role of Managers in the Organization*

Whereas most psychological studies of management focus on behavior, most sociological studies focus on managerial roles in the organization. A major influence in the study of management in organization was Weber's theory of bureaucratic administration, which fostered a view of managerial authority based on incumbency in a legally defined office and technical competence. Still more influential for the study of management was Barnard's (1938) classical study, which identified in administration, coordination, and control the essential managerial functions. For Barnard, the core of business organizations is the conscious and deliberate cooperation of different individuals for a common goal. Organizations are superior to individuals because of their intrinsic rationality. Accordingly, the function of managers is three-fold: they must take strategic decisions, through the rational evaluation of alternatives; they must assure the consensus and the cooperation of all members of the business organization through shared goals and value premises; and they must achieve a satisfactory mediation between organizational needs and individual expectations. Barnard's view laid the ground for major conceptions of the managerial role, which I will frame in the dominant organizational paradigms.

First, there is the view of the manager as a decision maker and an organization person, as it is derived by the organizational decision-making approach (Simon 1957, March and Simon 1958). This paradigm has been widely accepted in management schools; and

since management schools are important in shaping managerial roles, it is the most widespread among managers and the most influential in shaping their actual behavior. For Simon, organizations help decision makers both by simplifying decisions and by supporting participants in the decisions they need to make. Organizations encourage decision makers to 'satisfice' rather than maximize—that is, to settle for acceptable as opposed to optimal solutions—to attend to problems sequentially rather than simultaneously, and to utilize existing repertoires of action programs rather than to develop novel responses for each situation. Organizations are necessary because of the cognitive limits of individual decision makers; they support rational decision making by setting integrated subgoals and stable expectations and by subdividing responsibilities among managers and other firm members, providing them with the necessary information, resources, and facilities. According to this view, the manager is mainly the organization person, and managerial behavior is an instance of bounded rationality and is largely influenced by organizational structures. Another version of the manager as organizer is Marris's (1979), who sees the ability to organize the complex functioning of modern large-scale firms as the distinctive feature of management.

The second paradigm is the conception of the manager as the informal, human relations-oriented leader, as developed in the human relations model (Roethlisberger and Dickson 1939, Mayo 1945, Likert 1961), and related ones. It stresses the role of managers in exerting informal leadership, and promoting worker morale, increased effort, and cooperation in the service of organizational goals. From this perspective, good managerial leadership aims primarily at influencing the behavior of individual participants, and is conceived as democratic rather than authoritarian, employee-centered rather than production-centered, concerned with interpersonal relations and informal status hierarchies rather than with bureaucratic rules and formal authority. The human relations perspective stimulated relevant research and business policies aimed at increasing productivity through changes in work organization (job enlargement) and/or workers' participation in decision making. But, it was criticized both on ideological grounds—for de-emphasizing the actual conflict of interests between workers and management—and on empirical grounds, since several decades of research have demonstrated no clear relation between worker satisfacton and productivity, between leadership style and worker productivity, and between participation in decision making and worker satisfaction (Strauss 1963, Hollander and Julian 1969).

The third conception, best represented by Selznick (1957), is that of the manager as custodian of institutionalized values and mediator between organization and society. It is couched in the institutional paradigm and seeks to explain changes in the operative goals of the organization in the face of environmental constraints. Intellectually indebted to Michels and Merton, Selznick proposed to study the natural history of organizations by focusing on critical decisions that foster structural change. He discussed how they develop distinctive structures, capacities, and liabilities, through which they become institutionalized, that is, 'infused with value beyond the technical requirements of the task at hand' (Selznick 1957, p. 17). For him, informal structures can never succeed in conquering the nonrational dimensions of organizational behavior, because organizations are concrete structures that must adapt to pressures of their institutional environment, and because they are made of individuals who participate as 'wholes' in the organization and do not act merely in terms of their formal roles within the system. Accordingly, managers issue their directives, although they may be neither understood nor followed by their subordinates. The managerial role is, therefore, defined not in terms of managers' authority, but of their relation to institutionalized values. Since organizations reflect and protect societal values, managers act as catalysts in this process.

The fourth view is that of the manager as a negotiator in a conflict-ridden context and as a power holder, or as I prefer to define it, a power broker. It can be drawn from both Cyert's and March's (1963) behavioral approach, from the students of organizations as negotiated orders, and as political arenas, like Crozier and Friedberg (1977) and Daudi (1989). In this view, the plurality of goals, and the diversity of interests and outlooks among participants, make conflict not a symptom of organizational malfunctioning, but a normal feature of organizational life. According to Cyert and March, managers can adopt various mechanisms for conflict resolution, such as the budget—which stabilizes previous negotiations and expectations—standard procedures and operational rules, informal relations and traditional practices. Real strategies within the firm do not rationally maximize corporate ends but result from compromises within the dominant coalition and between the coalition and other collective actors.

The complex relation between corporate management and control and the insitutional context is also at the core of Fligstein's (1990) study on *The Transformation of Corporate Control*. Rooted in the paradigm which sees organizations as continuously transforming open systems, Fliegstein argues that changes in corporate control take place in three relevant institutional contexts. First of all, the firm itself is the institutional context where the internal strategy and structure of existing firms reflect organized power and interests, and managers' ideas about appropriate corporate behavior in the light of existing conventional wisdom. Second, organizations are embedded in larger groups of organizations which are called organizational fields, that may be defined in terms of product line, industry, or firm size (suppliers,

distributors, financial agencies and, most often, competitors). The organizations that surround the firm provide constant information as to what is occurring in the firm's product markets—which are filtered, interpreted, and greatly affect what actions are possible. Third, the state sets the rules that define what organizations can do and what the limits of legal behavior are. Legislatures and courts define the rights and privileges and set the boundaries of appropriate corporate behavior. Antitrust laws, for instance, define the limits of competitive behavior. Managers' courses of action are not so much constructed in terms of self-interest (whether it be that of the shareholder or the managers themselves), but rather in terms of the legal framework and the self-conscious version of the world that make old and new courses of action possible and desirable. Fligstein reminds us that the worlds of top managers have always been highly structured and their actions shaped by social and political contexts.

With a more radical departure from previous approaches, the perspective of the organization as a negotiated order brings the actor back at the center of organizational analysis, and focuses on concepts like power, strategic game, dominant coalition, and negotiated order. Crozier and Friedberg (1977) do not see organizations as the product of individuals and groups to established roles and procedures, but rather as the result of games in which the freedom of the various actors playing is limited by the formal and informal rules of the organization. Their vision, however, does not amount to diluting organizational action into a sum of individual actions, because the outcomes of interactive strategic games are real organizational properties, and because all participants need a certain amount of cooperation from others and share with them a common interest in organizational survival. By virtue of each actor's constrained freedom, they have power over the other players, since each actor is a source of uncertainty for them and for the organization as a whole. According to this conception, managers can be seen as special players who have greater control over uncertainty areas and are consequently more able to mobilize appropriate resources, like specialized competence, control of information, manipulation of norms, and control of the relations between the organization and the relevant environments.

More explicitly focused on power is Bacharach and Lawler's (1980) model, which introduces the conception of managers as power holders. Indebted to Michel's classical work on political parties as organizational oligarchies, this approach sees managers as power holders, both in the informal structure of influence and in the formal structure of authority. It analyzes the power games, coalitions, alliances, and conflicts organization leaders play both within the dominant coalition and with other antagonistic groups, and the relations between these leaders and their followers.

Related to the view of the manager as a power holder is the study of politics in management. The classical study, Dalton's (1959) *Men who Manage*, analyzes how a group of middle managers distorted the information that they gave to head office to forward their own political goals. Managers have more opportunities than other employees to gain resources and to form alliances in order to forward their career. In the study of politics in management, the image of the manager is often reversed. Managers are not described as acting primarily in the interest of the organization, but in their own interest. From being the heroes of the successful corporation through efficient administration, sound financial control, and profit-maximizing strategies, they become the rulers of bloated bureaucracies that do not produce sufficiently high returns on investment to stockholders, and concentrate on their personal aggrandizement.

4. *Managerial Ideologies and Managerial Styles*

Two major pioneering studies of managerial ideologies were published in the same year: that by Bendix (1956) and that by Sutton et al.(1956). The former is a comparative-historical study of managerial ideologies, i.e., of all those ideas which aim at justifying and explaining managerial power and authority within the firm and, indirectly, the social status of the business class. Managerial ideologies are internalized by entrepreneurs and top managers and are exposed on their behalf. The changes in ideologies of management during the nineteenth and twentieth centuries in Anglo-American and Russian civilizations are both similar and divergent; they are similar insofar as they share the increased managerial concern with the attitudes of workers, which presumably account for their differential productivity; they diverge because in the Anglo-American historical environments the authority relations between employees and workers remained to a large extent autonomous realms of group relations, whereas in pre- and postrevolutionary Russia, the conduct of both employers and workers was regulated by the superordinate authority of the state.

Whereas Bendix's study is a comparative historical analysis of social theories and ideas worked out by thinkers who interpret the ideological needs of business elites to legitimize managerial power, Sutton et al. study the explicit public statements of US managers and entrepreneurs. They conceive ideology not as a specific device for legitimizing power in the business organization, but as a multifunctional mechanism capable of alleviating the strains to which businesspeople are inevitably subject. Three major types of strains are identified: (a) strains stemming from criticism of the dominant position of business in society, that can undermine the businessperson's belief that

corporate values are consistent with US culture; (b) strains originating from the contradictory demands put to the businessperson by the different interlocutors (stakeholders in today's language), i.e., stockholders, employees, competitors, consumers, colleagues; and (c) the strains deriving from the conflicting demands of other social roles that entrepreneurs and managers play outside the firm—in family, community, and various informal groups.

Within the limits of their own cultural traditions, businesspeople adhere to their particular kind of creed to resolve emotional conflicts, overcome the doubts, and alleviate the anxieties engendered by the actions that their roles as entrepreneurs and managers compel them to take. For individuals in business, the function of the ideology is to help them maintain the psychological ability to meet the demands of their job. For instance, the conflict of values between profit and ethics can be eased in two opposite ways: either by affirming that devotion to profit is not inconsistent with social welfare but is, rather, a prerequisite of economic efficiency; or by denying that private profit is, or ought to be, the principal orientation of the business enterprise. Given the importance of business, it follows that ideology has functional importance also for those on whom the actions of business people impinge and for the whole society.

Most research on managerial behavior and roles has been limited to one country. Studies on cultural differences in managerial styles have been more comparative, like Bendix's. Relevant cultural studies of management focus on managerial styles and distinctive national traditions. They usually adopt a comparative perspective and compare Western Europe and North America with Japan, and, until the fall of the USSR, Western and Soviet managerial styles and cultures (Boddewyn 1971). Another group of research does not compare countries and socioeconomic systems, but rather different types of enterprise, along the private–public and the national–multinational dimensions.

The comparison with Japan is framed in the notion of the diversity of capitalisms that compares the US model of 'market-driven capitalism' with the European model of 'market social economy' and with the Japanese model of 'patronage capitalism.' The Anglo-Saxon model seems more competitive and more effective in job creation, whereas the Continental European one appears more effective in combining competitivenes and social cohesion. The Japanese experience is of interest for two main reasons: the successful competitive performance of Japanese firms in the world market, and recent technological changes that have made the Taylorist–Fordist types of work organization obsolete.

Drawing on the Japanese corporations' experience, authors like Dore (1973) and Ouchi (1981) have argued that managerial styles based on hierarchical relations have clear limits to their effectiveness. Overtly bureaucratized organizations, both of the Weberian type and on the Taylorist–Fordist model, lead to internal failures of functioning because of their rigid and participation-stifling nature. Clan forms of authority and the related styles of leadership, like those existing in the work groups and quality circles of Japanese firms, are more efficient, since they rely on the close ties and personal connections among the members of the group. Besides, this style of management develops noneconomic incentives, such as those based on collective identity and a sense of collective goals, which contribute to raise workers' morale and motivation.

Managerial career patterns are consistent with this culture, in the sense that institutional loyalty is a core value for Japanese managers, whereas for US and European ones a professional self-definition seems to prevail, in the sense that moving from one firm to another in order to improve one's career is considered fully legitimate. How far Japanese managerial styles depend on the cultural environment of Japan—which favors an especially compliant, deferential, and hard-working labor force—can be tested by studying the experience of Japanese-run firms recently set up in Western countries. Although the number of studies carried out so far is small, it does seem that Japanese management practices can operate with some success in a more individualistic labor force. Studies of Japanese-run plants in the USA, UK, and other Western European countries (White and Trevor 1983) indicate that bottom-up decision making does work outside Japanese society. Workers seem to respond positively to the greater level of involvement these plants provide, as compared to that of the Western style of management to which they were previously exposed. In recent years, however, the complex impact of globalization and the slower growth of the Japanese economy have put strains on the distinctive Japanese model and redirected it, to some extent, toward the market-led model of capitalism and US styles of management.

Another important instance of of cross-cultural influence in managerial styles has been the imitation, and to some extent the imposition, of US managerial values and practices in France and other Western European countries as a consequence of the USA's victory in World War II (Boltanski 1990).

The comparison of managerial styles and cultures in private and state-controlled enterprises has been another relevant, although declining, area of study. State-controlled firms differ to some extent from private ones because of their patterns of recruitment and socializing agencies and because of the distinctive features of their business environment. Public firms, managers tend to come more often than their private firms colleagues from either government bureaucracy, as in the French case, or politics and the liberal professions, as in the Italian case. They often have a distinct political-ideological imprinting, as did the

managers of postwar British nationalized industries (who had Labour Party affiliations) or the managers of postwar Italy's state-controlled firms (who had been educated at the Catholic University of Milan and had political experience in the Christian Democratic Party). They often showed different attitudes to labor relations and in corporate strategies because their dependence on state policies may have caused diffusion of targets, subordination of corporate goals to broader 'social' ends, and interferences by specific party factions and interest groups.

While some authors (Rees 1984) point out that they can act according to more flexible cost-opportunity criteria of management efficiency, others (Martinelli 1981) argue that they risk experiencing a threefold subordination: to private capital, which uses state-controlled firms to get rid of unprofitable businesses and to socialize losses; to political parties, who use them for quelling social tensions and granting benefits to specific interest groups; and to unions, who find in them a weaker counterpart than private companies, and a place where they can win 'easy' victories at the state budget's expense. However, these differences should not be exaggerated, especially for managers of public firms that compete with private ones in the same market environment.

A growing field is the study on executives' managerial cultures and styles in the extensive literature on multinational corporations (Weinshall 1977) and on globalization (Held et al. 1999). The prevailing view is that managers of multinational corporations do not differ significantly from national firms' managers, except for a greater emphasis on cosmopolitan values and styles of life, a greater sensibility for the scope of their decisions, and a greater alertness to the different societal environments in which they work. An interesting area in the analysis of multinational corporations' leadership styles is industrial relations, where a unified employers' strategy confronts a number of fragmented and diversified trade-union strategies (Makler et al. 1982, Hollingsworth and Boyer 1997).

As for the impact of modern globalization on managerial cultures and styles, contradictory tendencies are at work. On the one hand, distinct national and corporate-specific managerial styles persist, as do distinct institutional arrangements (different functions played by states although declining in their sovereign power, distinct trade-union strategies and organizations, specific organizational devices for increasing efficiency and workers' morale, etc.). On the other, the increasingly interconnected market economy fosters the diffusion of a common managerial culture and of cosmopolitan attitudes and styles among managers. It is still an open question to what extent this cosmopolitan culture is the diffusion of a single US-inspired model or the result of a hybridization of management styles and cultures. This seems the most promising line of future research in managerial roles, cultures, and behavior.

See also: Administration in Organizations; Corporate Culture; Corporate Governance; Entrepreneurship; Leadership in Organizations, Sociology of; Multinational Corporations; Organizational Decision Making; People in Organizations; Stockholders' Ownership and Control; Strategic Human Resources Management; Strategy: Organizational

Bibliography

Bacharach S B, Lawler E J 1980 *Power and Politics in Organizations*. Jossey-Bass, San Francisco

Barnard C 1938 *The Functions of the Executive*. Harvard University Press, Boston, MA

Bendix R 1956 *Work and Authority in Industry*. Wiley, New York

Berle A Jr, Means G C 1932 *The Modern Corporation and Private Property*. Harcourt, New York

Boddewyn J J 1971 *European Industrial Managers, West and East*. International Arts and Sciences Press, White Plains, NY

Boltanski L 1990 Visions of American management in postwar France. In: Zukin S, DiMaggio P (eds.) *Structures of Capital: the Social Organization of Economy*. Cambridge University Press, New York, pp. 343–72

Bottomore T, Brym R (eds.) 1989 *The Capitalist Class. An International Study*. Harvester Wheatsheaf, New York

Burch P H 1972 *The Managerial Revolution Reassessed: Family's Control in America's Large Corporations*. Lexington Books, Lexington, MA

Burnham J 1941 *The Managerial Revolution*. John Day, New York

Carlson S 1951 *Executive Behavior: A Study of the Workload and Working Methods of Managing Directors*. Stromberg, Stockholm

Chandler A D Jr 1962 *Strategy and Structure: Chapters in the History of Industrial Enterprise*. MIT Press, Cambridge, MA

Chandler A D Jr 1977 *The Visible Hand: The Managerial Revolution in American Business*. Harvard University Press, Cambridge, MA

Coase R H 1937 The nature of the firm: Origin, meaning, influence. *Economica* **4**: 386–405

Crozier M, Friedberg E 1977 *L'acteur et le système*. Seuil, Paris

Cyert R M, March J G 1963 *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ

Dalton M 1959 *Men Who Manage*. Wiley, New York

Daudi P 1989 *The Discourse of Power in Managerial Praxis*. Blackwell, Oxford, UK

Dore R 1973 *British Factory, Japanese Factory: The Origins of National Diversity in Industrial Relations*. Allen & Unwin, London

Fayol H 1916 *Administration industrielle et générale*. English translation 1949. Pitman, London

Fligstein N 1990 *The Transformation of Corporate Control*. Harvard University Press, Cambridge, MA

Galbraith J K 1967 *The New Industrial State*. Hamish Hamilton, London

Hannaway J 1989 *Managers Managing*. Oxford University Press, New York

Harbison F, Myers C A 1959 *Management in the Industrial World*. McGraw-Hill, New York

Held D, McGrew A, Goldblatt D, Perraton J 1999 *Global Transformations*. Polity Press, Cambridge, UK
Hollander E P, Julian J W 1969 Contemporary trends in the analysis of leadership process. *Psychological Bulletin* **71**: 387–97
Hollingsworth J R, Boyer R (eds.) 1997 *Contemporary Capitalism. The Embeddedness of Institutions*. Cambridge University Press, New York
Kotter J 1982 *The General Managers*. The Free Press, New York
Lewellen W G 1971 *The Ownership Income of Management*. Princeton University Press, Princeton, NJ
Likert R 1961 *New Patterns of Management*. McGraw-Hill, New York
Makler H, Martinelli A, Smelser N J 1982 *The New International Economy*. Sage, London
March J G, Simon H A 1958 *Organizations*. Wiley, New York
Marris R 1979 *Theory and Future of the Corporate Economy and Society*. North Holland, Amsterdam
Martinelli A 1981 The Italian experience. In: Vernon R, Aharoni Y (eds.) *State-owned Enterprise in the Western Economies*. Croom Helm, London pp. 85–99
Mayo E 1945 *The Social Problems of an Industrial Civilization*. Harvard University Press, Cambridge, MA
Meynaud J 1964 *La technocratie: mythe au realité?* PUF, Paris
Mintz B, Schwartz M 1985 *The Power Structure of American Business*. Chicago University Press, Chicago
Mintzberg H 1973 *The Nature of Managerial Work*. Harper & Row, New York
Mizruchi M S 1982 *The American Corporate Network: 1904–1974*. Sage, Beverly Hills, CA
Ouchi W A 1981 *Theory Z: How American Firms Can Meet the Japanese Challenge*. Addison Wesley, Reading, MA
Rathenau W 1918 *Die neue Wirtschaft*. S Fisher, Berlin
Rees R 1984 *Public Enterprise Economics*. Weidenfeld and Nicholson, London
Roethlisberger F J, Dickson W J 1939 *Management and the Workers*. Harvard University Press, Cambridge, MA
Selznick P H 1957 *Leadership in Administration*. Harper & Row, New York
Simon H 1957 *Administrative Behavior*. Macmillan, New York
Stewart R 1982 *Choices for the Manager*. McGraw-Hill, Maidenhead, UK
Stewart R 1991 Chairmen and chief executives: An exploration of their relationship. *Journal of Management Studies* **28**(5): 511–28
Stockman F N, Ziegler R, Scott J (eds.) 1983 *Intercorporate Structure: Comparative Analysis of Ten Countries*. London
Strauss G 1963 Some notes on power equalization. In: Leavitt H J (ed.) *The Social Science of Organization*. Prentice Hall, Englewood Cliffs, NJ
Streeck W, Schmitter P 1985 Community, market, state and associations. In: (ed.) *Private Interest Government: Beyond Markets and State*. Sage, London
Sutton F X, Harris S E, Kaysen C, Tobin J 1956 *The American Business Creed*. Harvard University Press, Cambridge, MA
Useem M 1984 *The Inner Circle: Large Corporations and the Rise of Business Political Activity in the U.S. and U.K.* Oxford University Press, Oxford, UK
Vernon R 1971 *Sovereignty at Bay*. Basic Books, New York
Weinshall T D (ed.) 1977 *Culture and Management*. Penguin Books, London
White M, Trevor M 1983 *Under Japanese Management: The Experience of British Workers*. Heinemann, London
Whitley R 1989 On the nature of managerial tasks and skills: Their distinguishing characteristics and organization. *Journal of Management Studies* **26**(3): 209–24
Williamson O 1975 *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York

A. Martinelli

Mania

A modern medical dictionary definition of mania refers to the Greek for 'madness' and gives a definition of 'a phase of bipolar disorder characterized by expansiveness, elation, agitation, hyper-excitability, hyperactivity, and increased speed of thought and speech (flight of ideas); called also manic syndrome' (*Dorland's Illustrated Medical Dictionary* 1994). Mania is a concept that has historical roots and modern changes in definition. The term currently refers to a phase of a mental disorder that is clinically defined as bipolar I disorder. A milder form of mania, hypomania, is also a mood state characteristic of a clinical condition, bipolar II disorder. This entry will review the history of mania from the early Greeks through the modern era.

The current psychiatric nomenclature, DSM-IV (American Psychiatric Association 1994), has dropped the term 'manic-depressive illness' in favor of bipolar I and bipolar II mood disorders. However, the term 'manic-depressive illness' is perhaps more descriptive since mania is usually accompanied by some type of depressive or dysphoric state, and manic episodes are most frequently preceded or followed by depressive episodes.

Although there are descriptions of mania as early as the ancient Greeks, the modern concept of manic-depressive illness emanates from the work of Emil Kraepelin in the late nineteenth century. The excellent review of manic-depressive illness by Goodwin and Jamison (1990) relates the Greek sense of mania as being attributed to an excess of yellow bile. Early scholars involved in the description of mania included Hippocrates, Aristotle, and Soranus of Ephesus. The notion that mania followed depression was perhaps first developed by Aretaeus of Cappadocia in the second century AD (Adams 1856). It is described that Aretaeus identified a bipolar disorder, a unipolar manic disorder, and a paranoid psychosis which might now be called schizoaffective mania. Aretaeus' descriptions of mania (as cited by Roccatagliata 1986) describe a period of depression followed by patients being gay, laughing, joking, singing, and dancing all day and all night. Assaultiveness during mania was described for the more serious forms and the notion of suspicion and paranoia was identified as part of mania by Aretaeus. Aretaeus apparently believed that mania

originated in the heart. Galen of Pergamon, writing in the second century, discussed melancholia (depression) as a chronic and recurrent condition.

Other authors who have discussed the history of the evolvement of mania, such as Jackson (1986), relate that throughout the years mania and depression were considered two separate illnesses but having some close connection between them. These notions persisted into the nineteenth century. Falret (1854) described *folie circularie*, a circular form of mania and depression. This description is a forerunner of what we now consider a rapid-cycling form of bipolar disorder in which manic and depressive episodes occur multiple times throughout the year.

In contrast to severe mania, a mild state or what we now call hypomania was first defined by Mendel (1881). Others including Falret and Esquirol had also described mild mania. Kahlbaum (1882) described a mild form of mood shifts which is now considered cyclothymia, or a rather stable characteristic of some individuals to have rapidly alternating high and low mood changes.

Kraepelin (1921) in the late nineteenth century introduced the term 'manic depressive' and separated individuals with mood disorders from those with schizophrenia. The term 'manic depressive' as used by Kraepelin included what we now consider bipolar I disorder (which is mania with and without episodes of depression), bipolar II disorder (which is recurrent depression with preceding or following hypomania), and recurrent major depression (which is now considered unipolar disorder or recurrent major depressive disorder). A major contribution of Kraepelin regarding mania was to give this disorder a name which described the course of illness (manic depressive) as well as to separate this condition from other psychoses, particularly from what is now called schizophrenia.

The Kraepelin view of mania persisted in the nomenclature through about 1980 when there was a movement to change the classification of manic-depressive illness (which included both bipolar and nonbipolar forms) into bipolar and unipolar mood disorders. Prior to 1980 the 'US–UK Study' (Kendell et al. 1971) demonstrated that the difference in prevalence rates for schizophrenia between the USA and UK was less likely due to absolute prevalence rate differences than to differences in definition. In this landmark study patients were diagnosed by psychiatrists trained either in the UK or in the USA. Of 11 patients diagnosed as manic by the UK team only one was so diagnosed by the US psychiatrists. Interestingly, the psychoanalytic approach viewed mania as a defense against depression, and schizophrenia was broadly defined. As psychoanalysis took hold in the USA, there was a reduction in research interest in mania or bipolar disorder (in contrast to schizophrenia) as a specific disorder in comparison to Europe. Considering that US psychiatry had a broad notion of schizophrenia and mania was a 'defense against depression,' it is no wonder that patients who might have been manic were not so diagnosed by the US-trained psychiatrists. The use of lithium carbonate in psychiatry was an important step in the process to change diagnostic terminology in the USA. Lithium carbonate's importance relates to its utility as a treatment for bipolar patients rather than for patients with schizophrenia. Thus it became important for psychiatrists in the USA, who heretofore had not diagnosed mania, to be able to classify those patients who might be lithium responsive. The change in diagnostic nomenclature in 1980 forced US clinicians to diagnose mania based on operational criteria for this condition. These operational criteria reflected the work of Clayton et al. (1965), who defined the symptom complex of patients with mania in such a way that reliability for symptoms could be ascertained. This led to a change in diagnostic style from a subjective to a more objective approach (DSM-II to DSM-III). This transition included proposals for symptom-based criteria for mental disorders (Feighner et al. 1972, Spitzer et al. 1978). This system-based approach has continued through DSM-III-R and DSM-IV (American Psychiatric Association 1994).

The other major change in diagnostic criteria for mania has been the acceptance of hypomania as a condition, resulting in the diagnosis of bipolar II disorder. Modern contributions for this area of research are from Dunner et al. (1976) and Akiskal (1981). Bipolar II disorder was included as a separate diagnostic entity in DSM-IV (Dunner 1993), based on clinical, familial, and treatment data separating such patients from bipolar I and unipolar patients. Additionally, the refinement of the concept of cyclothymic disorder as described by Kahlbaum (1882) has led to its inclusion into the DSM system (Howland and Thase 1993). Finally, the concept of a circular form of bipolar disorder resulted in the inclusion of a rapid-cycling form of bipolar disorder in DSM-IV. This work emanated from the observations of Stancer et al. (1970) and Dunner and Fieve (1974), with subsequent studies identifying rapid-cycling patients as being somewhat treatment-resistant to lithium, more likely to be women than men, and having a diagnostic relationship to bipolar rather than unipolar disorder (Bauer et al. 1994).

1. *The Symptoms of Mania and Hypomania*

The classic symptoms of mania are change in mood state to elation or irritability accompanied by symptoms such as grandiosity, decreased need for sleep, racing thoughts, increase in activity, distractibility, impulsive behavior, and increase in energy (Table 1). The syndrome for bipolar I disorder needs to be psychosocially disruptive and often results in hos-

Table 1
Symptoms during mania

Mood symptoms		Activity and behavior	
*Irritability	80%	*Hyperactivity	87%
*Euphoria	71%	*Decreased sleep	81%
Depression	72%	Violent/assaultive	49%
Lability	69%	*Rapid/pressured speech	98%
Expansiveness	60%	*Hyperverbosity	89%
Cognitive symptoms		Nudity/sexual exposure	29%
*Grandiosity	78%	Hypersexuality	57&
*Flight of ideas/racing thoughts	71%	Extravagance	55%
*Distractability	71%	Religiosity	39%
Poor concentration/confusion	25%	Head decoration	34%
		Regression	28%
		Catatonia	22%
Delusions		Fecal incontinence/ smearing	13%
Any	48%		
Grandiose	47%		
Persecutory/paranoid	28%		
Passivity	15%		
Hallucinations			
Any	15%		
Auditory	18%		
Visual	10%		
Olfactory	17%		

a DSM-IV criteria
*Data are weighted mean percentages of patients
Source: Goodwin and Jamison 1980

pitalization. If the patient is not hospitalized, the symptoms need to last at least one week. Psychotic symptoms may be present.

Hypomania has similar symptoms to mania. The condition needs to last at least four days. However, psychosocial disruption is not part of the hypomanic picture, which is usually a productive time. Psychotic symptoms are absent during hypomania. Patients with manic or hypomanic episodes usually experience depressive episodes just prior to or after the mania/hypomania. Depressive episodes are defined as two weeks or longer periods of a grouping of symptoms: depressed mood, decreased interest in usual activities, sleep changes (insomnia or hypersomnia), psychomotor change (psychomotor retardation or agitation), appetite change (increase or decrease in appetite/weight), loss of energy, difficulty with concentration, psychological symptoms such as guilt, and suicidal ideation.

2. *Epidemiology*

The population frequency of bipolar I disorder is approximately one percent, that is one percent of individuals in Western countries are likely to experience mania at some point in their life (Kessler et al. 1994). This disorder usually begins sometime after puberty and most individuals who manifest bipolar I disorder become ill by the age of 50. The gender ratio for bipolar I disorder reflects equal rates for men and women.

In contrast bipolar II disorder may be more difficult to diagnose and differentiate from unipolar (major) depression since hypomania is a productive state and often not recognized as 'illness' by the patient. The true population prevalence of bipolar II disorder is unknown, although it is generally thought to be somewhat more frequent than bipolar I disorder with perhaps 1–3 percent of the population experiencing lifetime episodes of depression and hypomania. The gender ratio likely reflects more women than men, similar to the ratio found in major depressive disorder, where a 2:1 female to male ratio is found. The age of onset of bipolar II disorder tends to be somewhat later than that for bipolar I disorder with earliest onset in puberty. However later onsets through the fifth and sixth decades are not uncommon.

3. *Bipolar–Unipolar Dichotomy*

Various proposals have been made regarding the separation of bipolar (depression with mania or

hypomania) from unipolar (depression only) patients. The Kraepelin concept of manic depression included both patients with mania and depression, and those with depression only. The bipolar–unipolar concept originated with Leonhard et al. (1962), and was supported by research by Perris (1966) and Angst (1966). These studies showed higher familial rates of mania in relatives of bipolar as compared to relatives of unipolar patients. Further research by Dunner et al. (1976) supported the bipolar–unipolar concept and introduced bipolar II (patients with depression and hypomania) as a distinct subtype. Bipolar II has been included in the most recent diagnostic nomenclature (Dunner 1993).

4. *Mixed Mania*

Mixed mania refers to a manic condition where depressive symptoms and manic symptoms coexist. Patients with an admixture of manic and depressive symptoms are well described in the article by Baastrup and Schou (1967), and generally are lithium non-responsive and may require treatment with mood-stabilizing anticonvulsants. Mixed mania is also included in the correct nomenclature (DSM-IV), although it is less well researched than typical mania (McElroy et al. 1992).

5. *Stage III Mania*

The evolution of psychotic symptoms during mania was described by Carlson and Goodwin (1973). They studied a series of 20 patients and described a progression of symptoms and behavior from a mild prodome stage I through a more typical hypomania with grandiosity (stage II) to mania complicated by frank psychotic symptoms (stage III). The issue of diagnosis of such patients had much to do with the broad concept of schizophrenia in the USA at that time and the notion that anyone who exhibited psychotic symptoms was schizophrenic. Thus the idea that untreated manic patients might progress through phases to become psychotic lent some credence to the psychoanalytic notion of mania being a psychological defense against depression, and yet clearly differentiated this condition from chronic schizophrenia, in which affective symptoms were rare. The differential diagnosis, however, between schizoaffective disorder (which is now defined as chronic psychosis with superimposed manic or depressive symptoms) versus psychotic mania was at that time confusing.

6. *Secondary Mania*

The concept of secondary mania was developed by Krauthammer and Klerman (1978) to describe manic syndromes following or due to exposure to trauma of the nervous system, seizures, medications, drugs, or infectious diseases of the nervous system. This concept has been refined in DSM-IV and classified as 'mania due to a general medical condition.'

7. *Cyclothymic Disorder*

The issue of cyclothymia as a disorder versus a personality or Axis II problem is also one that requires more research (Howland and Thase 1993). Indeed, it is difficult to find research on cyclothymic disorder, as patients with this rather stable condition rarely present for treatment. Patients diagnosed as cyclothymic experienced brief hypomanic periods (lasting fewer than four days) and brief mild depressive periods (lasting fewer than two weeks). Within a day, mood alterations from feeling high to feeling low are common. This mood pattern is persistent for two years or more.

8. *Rapid Cycling*

Rapid cycling is a type of bipolar I or a bipolar II disorder with episodes of mania, hypomania, mixed mania, or depression occurring at least four times a year. This condition was noted to be less responsive to the maintenance treatment effects of lithium carbonate, and thus its distinction as a separate type of bipolar I and bipolar II disorder highlights the difficulty in treatment of this condition for the clinician (Dunner and Fieve 1974). About 15 percent of bipolar I and bipolar II patients experience rapid cycling. The condition may itself be episodic with periods of rapid cycling followed by periods of less frequent mood cycles. Whereas the average frequency of episodes of mania or depression in bipolar I disorder is about four episodes in ten years (Winokur et al. 1969), a tenfold or higher frequency of episodes is found for rapid-cycling patients. The gender ratio for rapid cyclers is such that increased rates of women are found to experience increases in frequency of episodes (Bauer et al. 1994). A greater percentage of patients with rapid cycling experience a history of thyroid difficulties than nonrapid-cycling bipolar patients. Treatment of the rapid-cycling condition may require use of multiple mood stabilizers, avoidance of antidepressant medication which may indeed induce cycling in some patients, and longer treatment trials than one might use for nonrapid-cycling patients.

9. *Differentiation of Rapid Cycling, Mixed Mania, and Cyclothymic Disorder*

Rapid cycling, mixed mania, and cyclothymic disorder are characterized by frequent fluctuations in mood

from high to low. Mixed mania is a manic-like episode wherein there are both manic and depressive symptoms present at the same time. It is severe and disabling. Cyclothymic disorder is a rather stable mild condition with very frequent mood shifts from depression to hypomania. Rapid cycling has formed episodes of illness (hypomanic episodes of at least four days and depressive episodes of at least two weeks). In mixed mania and cyclothymic disorder there is often within-a-day cycling, such that the patient will experience rapid changes of high and low moods. Mixed mania is a briefer and a more severe state than one might experience with a patient who is cyclothymic, where the mood shifts are mild, not especially disruptive, and persistent for two years or more.

10. *Euphoric vs. Irritable Mania*

Clinically, two types of mania may be encountered. In one the patient is jovial, elated, euphoric, and funny, and in the second the patient may be irritable, paranoid, and more threatening. However, in both conditions there are frequently mood alterations from depression to irritability to normal mood to elation. Perhaps the hallmark of the mood disorder in mania is its lability.

11. *Interpersonal Behavior*

Interpersonal effects of mania are described in a classic article by Janowsky et al. (1970), which discusses the effects of manic behavior on others. Often there is a need for discussion among the individuals involved with the treatment of the manic patient so that all have the same approach and are aware of the same information. If not, the manic patient is likely to create a sense of havoc among those trying to give treatment by pointing out discrepancies in the treatment plan.

12. *Treatment*

Treatment of acute mania has indeed been considerably advanced by the introduction of lithium carbonate. Although acute mania is also well treated with antipsychotic drugs these generally lack the long-term mood-stabilizing effects of lithium. More recently, several anticonvulsants have been shown to have long-term mood-stabilizing effects for bipolar I disorder and it is likely that some of the new generation antipsychotic drugs may also be of benefit in long-term stabilization of this illness.

In contrast, bipolar II disorder is usually viewed as recurrent depression and the hypomania as a time of productivity. Thus individuals with this condition are less likely to ask for treatment for the hypomanic phase, which they view as a productive time of their life. However the approach to treating bipolar II disorder is indeed to prevent the recurrence of hypomania, and by stabilizing the hypomania the tendency to recurrent depressions is also stabilized. It should be noted that the suicide rate in bipolar I and bipolar II patients is considerable, particularly during depressed phases, and that the behavior of bipolar I patients when manic may result in their being hospitalized or jailed because of events related to their hyperactivity, manic grandiosity, and at times grandiose delusions.

Sometimes individuals during a manic syndrome will develop the stage III noted by Carlson and Goodwin (1973) and the psychosis may be mood congruent (delusions consistent with the elation or grandiosity evidenced by the patient, such as a belief that one has special powers). On the other hand, the delusions may be mood incongruent and have little to do with the mood state of the patient (for example, a sense that television programs are shown in order to control their thoughts). The presence of psychosis during mania should lead one to combine treatment with mood stabilizers, antianxiety medication, sedatives, and antipsychotic medication (see *Prophylaxis in Psychiatry*).

13. *Summary*

The concept of mania has evolved throughout the centuries from a disorder seen as separate to a condition combined with other mood states (manic-depressive illness) to bipolar I and bipolar II disorders. Further changes in the nomenclature to characterize additional subtypes of bipolar I and bipolar II disorders (such as rapid cycling) have added to our ability to develop treatment patterns for specific patients.

Future research is likely to focus on issues related to patients whose mania is mixed, as they present more difficult treatment issues. Additional research is needed for patients whose mania is more likely to be secondary to a history of polysubstance abuse, medical causes, head trauma, or other brain diseases. Furthermore, since the medications which are currently available are not always effective, there needs to be further development of pharmacotherapy both for the treatment of acute mania as well as for the stabilization of depressive episodes in bipolar patients.

See also: Bipolar Disorder (Including Hypomania and Mania); Comorbidity; Depression; Depression, Hopelessness, Optimism, and Health; Kraepelin, Emil (1856–1926); Mental Illness, Epidemiology of; Prophylaxis in Psychiatry; Psychotherapy and Pharmacotherapy, Combined; Schizophrenia and Bipolar Disorder: Genetic Aspects

Bibliography

Adams F (ed.) 1856 *The Extant Works of Aretaeus, the Cappadocian*. The Sydenham Society, London

Akiskal H S 1981 Subaffective disorders: Dysthymic, cyclothymic and bipolar II disorders in the 'borderline' realm. *Psychiatric Clinics of North America* **4**: 25–46

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. (DSM-IV). American Psychiatric Association, Washington, DC

Angst J 1966 *Zur Ätiologie und Nosologie endogener depressiver Psychosen*. Springer-Verlag, Berlin

Baastrup P C, Schou M 1967 Lithium as a prophylactic agent: Its effect against recurrent depression and manic-depressive psychosis. *Archives of General Psychiatry* **16**: 162–72

Bauer M S, Calabrese J, Dunner D L, Post R, Whybrow P C, Gyulai L, Tay L K et al. 1994 Multisite data reanalysis of the validity of rapid cycling as a course modifier for bipolar disorder in DSM-IV. *American Journal of Psychiatry* **151**: 506–15

Carlson G A, Goodwin F K 1973 Stages of mania: Longitudinal analysis of manic episode. *Archives of General Psychiatry* **28**: 221–8

Clayton P J, Pitts Jr. F N, Winokur G 1965 Affective disorder: IV mania. *Comprehensive Psychiatry* **6**: 313–22

Dorland's Illustrated Medical Dictionary, 28th edn. 1994 W. B. Saunders, Philadelphia, PA

Dunner D L 1993 A review of the diagnostic status of 'Bipolar II' for the DSM-IV work group on mood disorders. *Depression* **1**: 2–10

Dunner D L, Fieve R R 1974 Clinical factors in lithium-carbonate prophylaxis failure. *Archives of General Psychiatry* **30**: 229–33

Dunner D L, Gershon E S, Goodwin F K 1976 Heritable factors in severity of affective illness. *Biological Psychiatry* **11**: 31–42

Falret J 1854 Mémoire sur la folie cuculaire, forme de maladie mentale caractérisée par la reproduction successive et regulaire de l'état maniaque, de l'état melanchlogue, et d'un intervalle lucide plus ou moins prolonge. *Bulletin de l'Académie de Médecine* **19**: 382–415

Feighner J P, Robins E, Guze S B, Woodruff R A, Winokur G, Munoz R 1972 Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry* **26**: 57–63

Goodwin F K, Jamison K R 1990 *Manic-Depressive Illness*. Oxford University Press, New York

Howland R H, Thase N E 1993 A comprehensive review of cyclothymic disorder. *Journal of Nervous and Mental Disease* **181**: 485–93

Jackson S W 1986 *Melancholia and Depression: From Hippocratic Times to Modern Times*. Yale University Press, New Haven, CT

Janowsky D S, Leff M, Epstein R 1970 Playing the manic game: Interpersonal maneuvers of the acutely manic patient. *Archives of General Psychiatry* **22**: 252–61

Kahlbaum K L 1882 Uber cyclisches Irresein. *Der Irrenfreund* **10**: 145–57

Kendell R E, Cooper J E, Gourlay A J, Copeland J R, Sharpe L, Gurland B J 1971 Diagnostic criteria of American and British psychiatrists. *Archives of General Psychiatry* **25**: 123–30

Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshelman S, Wittchen H U, Kendler K S 1994 Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the US: Results from the national comorbidity study. *Archives of General Psychiatry* **51**: 8–19

Kraepelin E 1921 *Manic Depressive Insanity and Paranoia* [trans. Barclay R E, ed. Robertson G M]. Livingstone, Edinburgh, UK

Krauthammer C, Klerman G L 1978 Secondary mania: Manic syndromes associated with antecedent physical illness or drugs. *Archives of General Psychiatry* **35**: 1333–9

Leonhard K, Korf I, Schulz H 1962 Die Temperamente in den Familien der monopolaren und bipolaren phasischen Psychosen. *Psychiatrie, Neurologie und Medizinische Psychologie* **143**: 416–34

McElroy S L, Keck Jr P E, Pope H G, Hudson J I, Faedda G l, Swann A C 1992 Clinical and research implications of the diagnosis of dysphoric or mixed mania or hypomania. *American Journal of Psychiatry* **149**: 1633–44

Mendel E 1881 *Die Manie*. Urban and Schwazeberg, Vienna

Perris C (ed.) 1966 A study of bipolar (manic-depressive) and unipolar recurrent depressive psychoses. *Acta Psychiatrica Scandinavica* **194**, suppl.: 1–152

Roccatagliata G 1986 *A History of Ancient Psychiatry*. Greenwood Press, New York

Spitzer R L, Robins E, Endicott J 1978 *Research Diagnostic Criteria (RDC) for a Selected Group of Functional Disorders*, 3rd edn. New York State Psychiatric Institute of Biometric Research, New York

Stancer H C, Furlong F W, Godse D D 1970 A longitudinal investigation of lithium as a prophylactic agent for recurrent depression. *Canadian Journal of Psychiatry–Revue Canadienne De Psychiatrie* **15**: 29–40

Winokur G W, Clayton P J, Reich T 1969 *Manic-depressive Illness*. C. V. Mosby, St. Louis, MO

D. L. Dunner

Manifestoes, Party

Manifestoes or platforms are the programs usually issued by parties in campaigning for elections. They give an analysis of current problems and future developments endorsed—uniquely—by the party as a whole through its formal procedures. As such they are crucial to the operation of the democratic mandate: the idea that parties are chosen by electors for government on the basis of the competing policies they present. Only in the manifestoes can such competing policies be found, since the manifesto is generally the only policy statement endorsed officially by the party as a whole—in the election or indeed at any other time. The unique standing of the manifesto as a statement of official party policy renders it useful practically, as a guide to the media and (indirectly) electors, about what the party stands for; and also, crucial analytically, as the best indicator of what the party's collective preferences are. It is, however, a coded statement designed to win over votes. Hence its interpretation has to be based on assumptions about how parties compete for votes, which are described below.

1. The Nature of the Manifesto

In the UK, USA, and some continental European countries, manifestoes are published by the party as booklets at the official start of the election campaign. There have been clear physical changes in both appearance and length during the postwar period. They have become markedly longer and noticeably glossier. Starting out as badly printed, unillustrated pamphlets immediately after World War II, they have in some later cases approached the size of books. A more recent tendency is to have a plethora of glossy photographs, colored charts, and figures to reinforce and break up the text, which itself is headed by and interspersed with soundbites.

Elsewhere, practices in regard to the publication of election programs differ. In France, programs have sometimes been published commercially as books, which have attained the best-seller list! In Scandinavia and Canada they have at times appeared as separate pamphlets addressed to women, youth, workers, and other demographic groups. In Japan, they are widely diffused as summaries of interviews with the party secretaries, published in the leading mass circulation newspapers. In Australia and New Zealand they do not even get into print, but are substituted by an hour-long television address from the party leader—the content of which is organised along conventional lines, and so covers much the same range of topics broken into the same kinds of sections as in the USA and UK.

These variations may reflect a growing realization that most electors do not read party programs even if they are pushed through their door. Instead they get an impression of them from media discussion. Media presenters, on the other hand, read programs very carefully, using them to start debates and confront party spokespeople if they try to pull back from concerns expressed in the document. In this sense the manifesto constitutes a policy equilibrium point for the party set at the outset of the campaign (Budge 1994).

Reading the document casually, however, one might be hard pressed to discover much policy content at all. Until recently specific promises have been limited and confined mostly to pledges on peripheral problems where the party knew it could deliver with some certainty if it became (part of) of the government (Rose 1980). Most of the text discusses general trends and developments under such headings as Youth, Unemployment, the Economy, and so forth. Depending on whether the party is in government or opposition it describes the situation in optimistic or pessimistic terms, sometimes going into the historical achievements of the party or stressing the importance of the area.

Rarely, however, is the text more specific than this. The manifestoes of competing parties do nonetheless succeed in presenting quite contrasting diagnoses of the current situation—one painting a glowing picture of free enterprise surging ahead amid general prosperity, another pointing to growing social inequalities and the pressing need for government intervention. Without further information one could not even be sure if they appeared at the same election or at different ones.

2. Electoral Competition and the Manifesto

The reason for structuring election programs in this way is to be found in the view party strategists take of competition for votes. In line with Riker's (1993, p. 105) 'dominance principle,' electors are seen as overwhelmingly endorsing one 'obvious' line of policy in each area: cutting taxes, strengthening law enforcement, extending social provision, and so forth. Particular parties, because of their record and ideology, are associated with each favored line of policy. If you want to cut taxes you will not vote Socialist but you will if you want to extend welfare.

For the parties, competition thus consists in getting 'their' issues—the ones on which they are associated with the popular line of action—on to the agenda. To do this they need to emphasize the importance of such issues compared with those which favor rivals. Their manifesto progresses this process by talking a lot about their party issues, ignoring areas 'belonging' to rival parties. As the task is to convince media and electors of these issues' importance, the manifestoes can ignore fine details of policy in favor of general discussion—a tactic which, conveniently, does not tie parties down to specific actions when in government.

The view of party competition which emerges from analyses of manifestoes (Robertson 1976) is thus one in which relative emphases on different issues substitute for direct confrontation over differing policies on the same issue. This is not to say, however, that parties are not putting forward different programs of government. Parties which talk a lot about taxes mean to cut them and parties discussing welfare mean to extend it. Manifestoes are about priorities between tax, welfare, and other areas, even if the other side to the debate—cutting services or increasing taxes—goes unstated.

3. Using Manifestoes in Political Research

The way manifestoes are written makes them relatively easy to analyze quantitatively. By counting sentences or words into different policy areas and percentaging them out of the total to standardize for length, one can contrast left-leaning and right-leaning parties and even characterize their attitudes within different policy areas (Klingemann et al. 1994). Word counts of electronic texts can be performed easily by computer programs, opening up new possibilities of research once the techniques are better validated. Sentence counts have been carried out manually for all signifi-

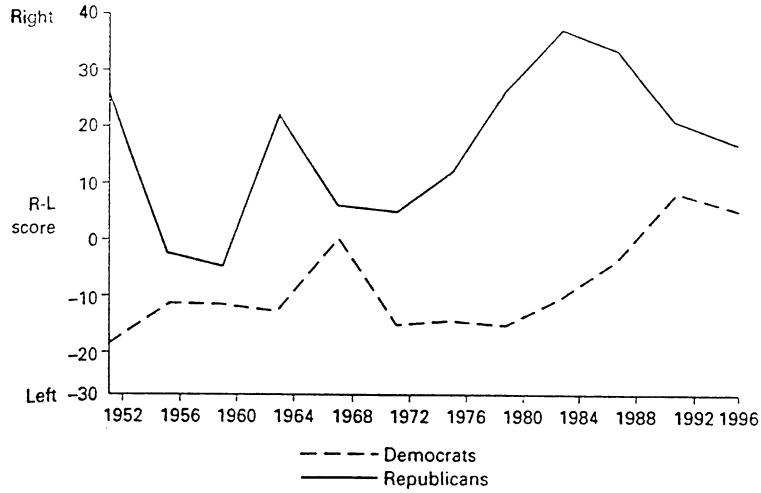

Figure 1
'Leftness' and 'rightness' of US party policy, estimated from platforms 1952–96

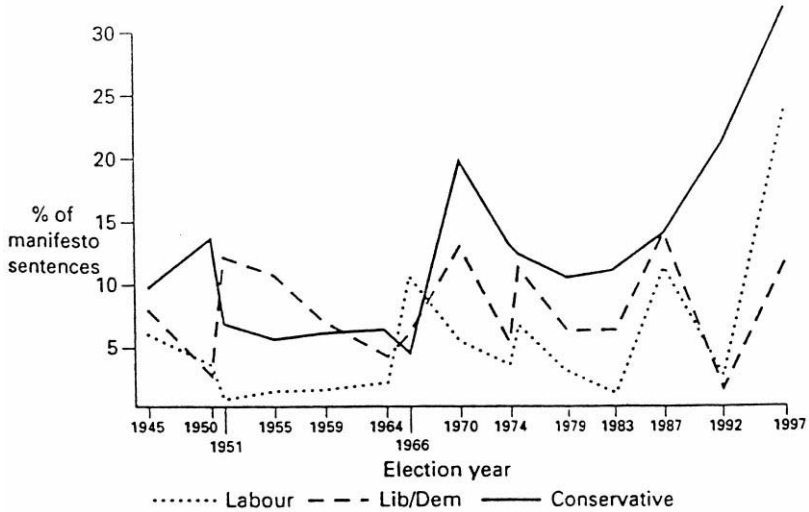

Figure 2
British parties' support for social conservatism, estimated from manifestoes 1945–97

cant parties in 51 countries over the postwar period, from the time of the first democratic election (Budge et al. 1987).

Given the manifestoes' standing as the sole authorized statement of party collective policy preferences, such analyses can give an exact picture of parties' ideological progression and shifting policy stances over the postwar period. This is illustrated in Fig. 1, which traces the US Democrats' and Republicans' left–right movement over the postwar period. (Left–right is measured by adding all percentage mentions of 'rightist' issues and substracting from this the total percentage mention of 'left-wing' issues, Klingemann et al. 1994, p. 40.)

Figure 1 shows first of all that Democrats and Republicans take up the standard left and right positions assumed by Socialist and Conservative parties in other countries. The validity of interpreting US party competition in left–right terms is supported:

(a) by the way in which contrasts between the parties mirror the historical record—for example, far apart at the 'Goldwater' election of 1964, closer in

1992 with Clinton's famous move to the right;

(b) by the high correlation between presidential platform positions and the relative 'liberalism' of subsequent presidential policy (McDonald et al. 1999). (This finding incidentally illustrates the usefulness of comparing manifesto commitments with actual policy as a method of checking out mandate theory, (Klingemann et al. 1994);

Figure 1 also serves as a check on Downsian theories of party competition which predict policy convergence between parties as both seek to position themselves near the median elector (Downs 1957). Clearly this expectation is not fulfilled as the US parties always maintain their ideological distance from each other, even where they converge to some extent, as in 1992. Their patterns of movement are explained better by the 'saliency' theory, where parties are stuck with their traditional issue positions and can act strategically only by varying the emphasis they put on them (Budge et al. 1987).

Such changes can be monitored for the separate issue areas by seeing what percentage of sentences are devoted to them in the manifesto at different elections. Figure 2 traces British parties' concerns with traditional morality and law and order ('social conservatism') over the postwar period. The Conservative Party generally appropriates this area where it presents itself as the defender of 'family values.' However, Labour emphasized it a great deal in the 1997 General Election when it was a major component in the party's rightward shift (Evans and Norris 1999).

4. *Linking Manifestoes to Election Change*

As well as supplementing historical analyses of party change, such 'maps' of party movement can also be related to voting behavior. Instead of estimating party positions indirectly by asking electors about them, we can actually study the effect of a party's strategic adjustments on electors' perceptions and subsequent voting decisions. This extends the possibilities of linking such phenomena as 'revisionism' to electoral reactions, to see if it actually has the effects claimed for it (Evans and Norris 1999, pp. 87–101). Methodologically one can marry textual analysis to survey research to deepen the insight provided by both.

5. *Linking Manifestoes to Government Action*

Manifestoes mediate between electors' preferences —expressed by voting for the closest party—and government actions as mirrored in expenditure or legislation. As well as investigating what impact they have on electors, the priorities in the manifesto can be related to government policy outputs. Research has shown them to be quite closely linked (McDonald et al. 1999), thus upholding 'mandate' ideas that the party or parties in government have been popularly authorized to carry through their program and indeed have a responsibility to carry it through.

This moral commitment is reinforced by two other characteristics of manifestoes, which give them an impact in government.

(a) Manifestoes reaffirm party ideology, applying it to specific policy areas. Thus most party representatives, who belong to a party in the first place because they subscribe to its principles, actually want to carry out manifesto commitments regardless of electoral considerations.

(b) The manifesto is often the only worked out program the party has for government. Even where coalition partners make a policy agreement about what to do, the separate party manifestoes feed into it. For practical reasons of coordination and planning in government therefore, the manifesto is an indispensable document.

6. *Overall Role and Significance of Manifestoes*

Cynicism about the extent to which parties keep their promises, and hence about the significance of the manifesto which contains them, is easy. The election program has often been seen as a 'mere piece of paper' forgotten immediately after the party gets into government. The research cited above, however, shows that the manifesto deserves to be taken more seriously, both by voters and political analysts. Parties are responsible, both in maintaining consistency with previous policy and in carrying it out in government. This would hardly be so if they just shifted policies around in response to electoral expediency. What Figs. 1 and 2 show is considerable stability in the positions taken by the only policy document which parties issue collectively. This renders it an invaluable source for studying party intentions and preferences, and a good guide to what they will do in government.

See also: Conservatism: Theory and Contemporary Political Ideology; Democratic Party; Ideology: Political Aspects; Issue Constraint in Political Science; Party Identification; Party Responsibility; Republican Party

Bibliography

Budge I 1994 A new spatial theory of party competition. *British Journal of Political Science* **24**: 443–67

Budge I, Robertson D, Hearl D (eds.) 1987 *Ideology, Strategy and Party Change*. Cambridge University Press, Cambridge, UK

Downs A 1957 *An Economic Theory of Democracy*. Harper, New York

Evans G, Norris P 1999 *Critical Elections*. Sage, London

Klingemann H-D, Hofferbert R I, Budge I et al. 1994 *Parties, Policies and Democracy*. Westview, Boulder, CO

McDonald M D, Budge I, Hofferbert R I 1999 Party mandate theory and time series analysis. *Electoral Studies* **18**: 587–96
Riker W H 1993 *Agenda Formation*. University of Michigan Press, Ann Arbor, MI
Robertson D 1976 *A Theory of Party Competition*. Wiley, London
Rose R 1980 *The Problem of Party Government*. Macmillan, London

I. Budge

Mannheim, Karl (1893–1947)

As a classic of sociology, Karl Mannheim is the author of one book, with three lives. *Ideologie und Utopie* (Mannheim 1929) was the most widely debated book by a living sociologist in Germany during the climactic years of the Weimar Republic; *Ideology and Utopia* (Mannheim 1936) has been a standard in American-style international academic sociology; and the quite different German and English versions of the book figure in reappraisals of Mannheim initiated by new textual discoveries (Mannheim 1980, 1984) and re-publications (Wolff 1993, Mannheim 1952, 1953, 1956), fostered by a reopening of questions the discipline of sociology had once deemed closed. Mannheim's sociological theorizing has been the subject of numerous book-length studies, evidence of an international interest in his principal themes (e.g., Endreß and Srubar 2000, Gabel 1987, Kettler and Meja 1995, Santambrogio 1990). As a living social thinker, Karl Mannheim was not in fact the author of any work he himself considered a finished book, but rather of some 50 major essays and shorter treatises, many later published in book form.

Born on March 27, 1893 in Budapest, the son of prosperous Jewish parents, Mannheim studied in Budapest, Berlin, Paris, and Heidelberg, held academic posts at Heidelberg, Frankfurt, the London School of Economics and the University of London, and died January 9, 1947 in London. His biography, one of intellectual and geographical migration, falls into three main phases: Hungarian (to 1919), German (1919–33), British (1933–47). Among his most important early intellectual influences are Georg Lukács, Oscar Jászi. Georg Simmel, Edmund Husserl, Heinrich Rickert, and Emil Lask. Mannheim was also strongly influenced by the writings of Karl Marx, Max Weber, Alfred Weber, Max Scheler and Georg Dilthey. Through these and other authors, German historicism, Marxism, phenomenology, sociology and—much later—Anglo-Saxon pragmatism became decisive influences upon his work.

Mannheim spent his first 26 years in Budapest, where he graduated in philosophy and precociously participated in the intellectual life of the 'second reform generation' born a decade earlier. The advanced thinkers of the time were divided between proponents of modernization oriented to French and English social thought and prophets of radical cultural rebirth inspired by Russian and German models. Mannheim did not think that his dedication to the latter group, led by the philosopher Lukács, entailed a blanket rejection of the former, under the sociologist Jászi. Lukács' wartime *Sunday Circle* in Budapest may have devoted its meetings to Dostoevski and Meister Eckhardt, with Mannheim in eager attendance, but Mannheim was also proud of his acceptance in the *Max Weber Circle* when he was in Heidelberg, and during a visiting semester in Berlin in 1914 he attended the lectures of Georg Simmel, the subtle mediator between cultural philosophy and sociology.

Mannheim lived in Germany from 1919, when he fled the counter-revolutionary regime in Hungary, until 1933, when National Socialist decrees forced him out of the university and he left Germany for England. Within a few years of his transfer from the Budapest intellectual scene to German university life, Mannheim began work in Heidelberg on a habilitation thesis in cultural sociology under Alfred Weber. Mannheim's sociological interpretation of the rise and self-differentiation of conservatism was subtitled *A Contribution to the Sociology of Knowledge*, and its submission coincided with Mannheim's publication of an article devoted to his critical encounter with Max Scheler, who recently had brought the concept of the sociology of knowledge into discussion. Mannheim's inaugural address as university instructor at Heidelberg set out the parameters of 'the contemporary state of sociology in German,' dealing with Weber, Troeltsch, and Scheler. Sociology, he believed, provided the frame of reference for twentieth century thinking as a whole.

Mannheim's aloofness, however, from the specialized 'state of sociology' question, as well as his equation of the main currents of contemporary thought with the leading sociological theories, indicate that his move from philosophy to sociology cannot be understood as a simple change of academic specializations. Sociology, in his view, was a more comprehensive undertaking than the academic discipline still taking form. Goaded in 1929 by a charge of 'sociologism' against *Ideologie und Utopie*by the noted literary scholar, Ernst Robert Curtius, Mannheim (in Meja and Stehr 1990) again invoked the heritage of Weber, Troeltsch, and Scheler against the literary scholar's accusation of treason to humanism, presenting these sociologists as modern German classics, as 'a great heritage, a tradition that must be built upon.'

The novelty in Mannheim's approach to the sociology of knowledge is epitomized in three distinct claims. First, and most controversial, is the contention that boundaries between manifestly ideological and ostensibly scientific modes of explaining the cultural as well as the social world are porous, with sociology of knowledge emerging in the border region, as a reflexive

therapy for both domains. Second, is the concomitant conception of ideologies as cognitive structures that are variously flawed, limited, perspectivistically one-sided, subject to correction from other perspectives, and nevertheless productive of knowledge. The third claim, then, is that the sociology of knowledge bears on the answers to substantive questions addressed by ideologies, consequently contributing directly to political orientation. It does so not because knowledge of social genesis can determine judgments of validity, but because systematic pursuit of such knowledge fosters a comprehensive synthesis and renders particularistic ideologies obsolete.

Mannheim's strategy involves two steps. First, the variety of ideas in the modern world is classified according to a scheme of historical ideological types, in keeping with his thesis that the ideological field has moved from atomistic diversity and competition to concentration. Liberalism, conservatism, and socialism are the principal types. Second, each ideology is interpreted as a function of some specific way of being in the social world, as defined by location within the historically changing patterns of class and generational stratification. Liberalism is thus referred to the capitalist bourgeoisie in general, and various stages in its development are linked to generational changes. Similar analyses connect conservatism to social classes harmed by the rise to power of the bourgeoisie, and socialism to the new industrial working class. Each of the ideologies is said to manifest a characteristic 'style' of thinking, a distinctive response to the issues that systematic philosophy has identified as constitutive of human consciousness, such as conceptions of time and space, the structure of reality, human agency, and knowledge itself. The political judgments and recommendations of purely ideological texts must be taken in that larger structural context. The style of thinking is most apparent in the way concepts are formed and in the logic by which they are interlinked. Each style expresses some distinctive design upon the world vitally bound up with the situation of one of the social strata present in the historical setting.

Sociology of knowledge seeks to give an account of the whole ideological field, in its historical interaction and change. To have a method for seeing this means to see in a unified way what ideologically oriented viewers can only see in part. Mannheim draws on Marxism for a conception of politics as a process of dialectical interplay among factors more 'real' than the competing opinions of liberal theory. But neither the proletariat nor any other socio-political force is bearer of a transcendent rationality, historically destined to reintegrate irrationalities in a higher, pacified order. The contesting social forces and their projects in the world are in need of a synthesis that incorporates elements of their diverse social wills and visions. Syntheses in political vision and in the social sciences are interdependent. Sociology of knowledge foreshadows and fosters both.

Mannheim's call to a professorship in Frankfurt followed the remarkable recognition earned by his further work in sociology of knowledge. In his presentation on 'Competition' at the Sixth Conference of German Sociologists in 1928 (in Wolff 1993), Mannheim audaciously used the value-judgment controversy in recent sociology to illustrate his theses about the connectedness to existence (*Seinsverbundenheit*) of social thought and the operations of socially grounded competition to generate syntheses that transcend intellectual conflict. *Ideologie und Utopie*, consisting of three essays on 'politics as a science, 'utopian consciousness,' and an explication of the concepts of '*ideology*' and '*utopia*,' generated great excitement that launched Mannheim in his new setting, recognized in the wider intellectual community as a significant and controversial personality. Questions about varieties of knowing and about the ways in which new knowledge depends on authentic grounding in the contexts of existing knowledge are the major themes of much of Mannheim's work during this period. His seminal paper on *The Problem of Generations* (1927), which continues to have a lasting impact in sociological research and analysis (especially by way of cohort analysis and the sociology of the life course), also belongs here.

The sociology of knowledge dispute about *Ideologie und Utopie* (Meja and Stehr 1990) was mainly philosophical and political, with the focus, first, on Mannheim's hope of overcoming both ideology and political distrust through sociology of knowledge; second, on his conception of the intelligentsia as the social stratum uniquely equipped and even destined for this task; and third, on his activist conception of sociological knowledge, as a mode of mediating between theory and practice. In *Science as a Vocation* (1922), Max Weber had distinguished between words in politics and in science, likening the former to weapons for overpowering opponents and the latter to ploughshares for cultivating knowledge. Mannheim offers the sociology of knowledge as an 'organon for politics as a science,' as a way of bringing about the biblical transformation of swords into pruning hooks prophesied by Isaiah. His proposals were widely canvassed in the leading periodical reviews and subjected to intense criticism, but his reading of the intellectual situation was almost universally applauded. In the cultivated Weimar public and among the participants in the 'Weimar conversation' about social thought after Nietzsche and Marx, *Ideologie und Utopie*figured as the representative book of its time, whether as symptom of cultural crisis or as promise of a way out.

During his five semesters as professor in Frankfurt, however, Mannheim declined the role of public intellectual, He separated the professional aspects of his activities from his public reputation. While he drew close to Paul Tillich and his circle of religious socialists, celebrated and embattled as an 'intellectual,' he

nevertheless increasingly defined himself as a professional sociologist. Never having exposed himself politically, he was caught unaware by the Nazi measure that deprived him of his professorship on grounds of his foreign birth and Jewish ethnicity. *Ideologie und Utopie* had been favorably treated in the Socialist periodical, *Die Gesellschaft*, but the four articles published there (by Hannah Arendt, Herbert Marcuse, Hans Speier, and Paul Tillich, were all more critical than the reception that his work received in *Die Tat*, a periodical of the activist Right. The hard Left treated him as a betrayer of Marxism. Soon Mannheim was a refugee in Amsterdam. Neither his sociology nor his politics had anything to do with his exile from Germany.

In the summer of 1933, Mannheim was appointed to a special lectureship at the London School of Economics. Neither the times nor his situation were conducive to pursuing sociology of knowledge studies, however. Mannheim saw it as his mission to diagnose the general crisis he held responsible for the German disaster and to promote prophylactic and therapeutic measures in Britain. His sense of urgency and his grand theoretical ambitions enthused students, but they estranged the professional sociologists led by Morris Ginsberg, who were engaged in a difficult fight for academic respectability of a discipline still widely dismissed by the English university establishment. Although marginalized at the London School of Economics, the only British institution with a chair in Sociology, Mannheim made a place for himself as a public intellectual, especially after his acceptance by a circle of Christian thinkers that included T. S. Eliot and whose periodic discussions and publications centered on a theme of cultural crisis hospitable to Mannheim's sociological interpretations.

In Britain Mannheim continued to focus on the relationship between knowledge and society, his lifelong topic. Writing in *Man in Society in an Age of Reconstruction*, a work no less influential in the postwar years than *Ideology and Utopia*, he now claims that the sociology of knowledge has lost its strategic centrality with the demise of ideological competition (Mannheim 1940) and calls for 'a new experimental attitude in social affairs,' in view of the 'practical deterioration of the ideals of Liberalism, Communism, and Fascism.' The National Socialist dictatorship, Mannheim argues, exploits a socially unconscious mass response to a worldwide crisis in the institutions of liberal civilization, involving the obsolescence of its regulative social technologies—from markets to parliaments to elitist humanistic education. He pleads for planned social order that strategically utilizes the new social technologies that undermine the spontaneous self-ordering of the previous epoch. A consensual reconstruction could save human qualities and diversities earlier privileged by liberalism, unlike the violent homogenization imposed by communist or national socialist control through command. Timely action guided by awareness of the impending crisis by leading strata, notably the English elite of gentlemanly professionals, can tame the processes that would otherwise destroy liberal civilization and condition mass populations for dictatorial domination. 'Planning for freedom' presupposes a reorientation among traditional elites, their acceptance of a sociological diagnosis of the times, and their willingness to learn prophylactic and therapeutic techniques (Mannheim 1943). Mannheim now claims for sociology the ability to ground and coordinate interdisciplinary approaches in planning. His lectures and writings on 'planning' won him an interested audience, especially during the war and immediate postwar years, and his conception of a post-ideological age was never altogether submerged by the Cold War (Mannheim 1950). As his wartime slogan of 'militant democracy' justified German measures against leftists during the middle decades of the century, his slogan of the Third Way is heard at the beginning of the new millennium in support of political designs he would have found familiar.

Among sociologists, however, Mannheim's standing was defined by the reception of *Ideology and Utopia*. In his preface Louis Wirth casts the work primarily as a contribution to objectivity in social science. The professional consensus is formalized in R. K. Merton's authoritative essay in 1945 on 'The Sociology of Knowledge' (in Merton 1957). Merton includes Mannheim in a group of social theorists from Marx to Sorokin whose diverse approaches to 'the relations between knowledge and other existential factors in the society and culture' he relates to questions and alternatives that provide an agenda for the theoretical clarification and empirical research required to build sociology. Merton's 'paradigm' for the sociology of knowledge sets out five key issues: the existential basis of mental productions, the varieties and aspects of mental productions subject to sociological analysis, the specific relationship(s) between mental productions and existential basis, the functions of existentially conditioned mental productions, and the conditions under which the imputed relations obtain. Crediting him with having sketched the broad contours of the sociology of knowledge with skill and insight, Merton nevertheless finds Mannheim's theory loose, burdened with dubious philosophical claims, and insists that the relations between knowledge and social structure can be clarified only after they are 'shorn of their epistemological impedimenta, with their concepts modified by the lessons of further empirical inquiry' (Merton 1957, p. 508).

The condition for Mannheim's acceptance as deserving sociological pioneer was the discarding of the concept of total ideology and his way of questioning both social science and social knowledge. The editors of three posthumous collections of Mannheim's essays, his intimate friends and noted social scientists, Paul Kecskemeti and Adolph Loewe, reinforced the

consensus view. As Mannheim became acquainted with Anglo-American social science, they argued, empirical social psychology displaced continental philosophies as the theoretical framework for his thinking, and his early writings merit consideration primarily as brilliant anticipations of later developments. In Merton's sociological theory classes during the 1950s, *Ideology and Utopia* often followed Machiavelli's *Prince* in the syllabus, as source material for an exercise in transmuting suggestive ideas into testable propositions. Mannheim's work had joined the list of historical authorities celebrated as legitimating precursors of a settled way of defining and doing sociology.

We can trace the renewal of attention to the historical social thinker, Karl Mannheim, to renewed conflicts about the subject, method, and attitude of sociology. The connection is not simple, since the attack on the disciplinary consensus, where historical models were involved, was more likely to call on Marxist writers or on figures like Theodor W. Adorno, Max Horkheimer, and Herbert Marcuse, who had been among Mannheim's harshest contemporary critics. The effect, nevertheless, was to relegitimate questions about the historicity of social knowledge, the problem of relativism, and the paths of reflexivity open to social thinkers—the issues filtered out of Mannheim's thought in his American reception—and to provide a new point of connection for those sociologists (e.g., Wolff 1993) who had continued puzzling over the issues debated when *Ideologie und Utopie* first appeared. Berger and Luckmann's influential phenomenological reformulation of the sociology of knowledge in *The Social Construction of Reality* (1966) represents a departure from Mannheim's sociology of knowledge. All knowledge, both objective and subjective (and including ideology, propaganda, science and art), is now seen as socially constructed. Berger and Luckmann examine how forms of knowledge are maintained and modified by the institutions and individuals embodying and embracing them.

In the newly fluid state of questions appropriate for sociological theory, Mannheim's famous *Ideology and Utopia* no longer stands alone among his writings, and his insistence on essayistic experimentalism is no longer ignored. There are no propositions to be distilled out of Mannheim's work; there is just the thoughtful encounter with it. Mannheim's prediction about the consequences of uncovering the 'massive fact' of knowledge in society and society in knowledge has been borne out. Yet we are not beneficiaries of any 'progress' in thought, as post-modernists paradoxically often suppose: we may well have to think deeply about old texts, however we class them.

See also: Bourgeoisie/Middle Classes, History of; Class: Social; Conservatism: Theory and Contemporary Political Ideology; Democracy; Democracy, History of; Historicism; Ideology: History of the Concept; Ideology, Sociology of; Knowledge, Sociology of; Liberalism; Marxism in Contemporary Sociology; Marxist Social Thought, History of; National Socialism and Fascism; Objectivity of Research: Ethical Aspects; Objectivity: Philosophical Aspects; Phenomenology: Philosophical Aspects; Planning, Politics of; Pragmatism: Philosophical Aspects; Pragmatist Social Thought, History of; Reflexivity in Anthropology; Relativism: Philosophical Aspects; Social Changes: Models; Social Inequality in History (Stratification and Classes); Socialism; Theory: Sociological; Utopias: Social; Weber, Max (1864–1920); Weberian Social Thought, History Of; Working Classes, History of

Bibliography

Berger P, Luckmann T 1966 *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Doubleday, New York

Endreß M, Srubar I 2000 *Karl Mannheims Analyse der Moderne*. Jahrbuch für Soziologie. Leske und Budrich, Opladen, Germany

Gabel J 1987 *Mannheim et le Marxisme Hongrois*. Méridiens Klincksieck, Paris

Kettler D, Meja V 1995 *Karl Mannheim and the Crisis of Liberalism: The Secret of These New Times*. Transaction Publishers, New Brunswick, NJ

Mannheim K 1929/1953 *Ideologie und Utopie*, 3rd enlarged edn. Schulte-Bulmke, Frankfurt am Main, Germany (1995 8th edn. Klostermann, Frankfurt am Main, Germany)

Mannheim K 1936 *Ideology and Utopia*. Kegan Paul, Trench, Trubner & Co., London

Mannheim K 1940 *Man and Society in an Age of Reconstruction*. Kegan Paul, Trench, Trubner & Co., London

Mannheim K 1943 *Diagnosis of Our Time. Wartime Essays of a Sociologist*. Kegan Paul, Trench, Trubner & Co., London

Mannheim K 1950 *Freedom, Power and Democratic Planning*. Oxford University Press, New York

Mannheim K 1952 *Essays in the Sociology of Knowledge*. Routledge and Kegan Paul, London

Mannheim K 1953 *Essays on Sociology and Social Psychology*. Routledge and Kegan Paul, London

Mannheim K 1956 *Essays on the Sociology of Culture*. Routledge and Kegan Paul, London

Mannheim K 1922–24/1980 *Strukturen des Denkens*. Suhrkamp, Frankfurt am Main, Germany [1982 *Structures of Thinking*. Routledge & Kegan Paul, London]

Mannheim K 1925/1984 *Konservatismus Ein Beitrag zur Soziologie des Wissens*. Suhrkamp, Frankfurt am Main, Germany [trans. 1986 Kettler D, Meja V, Stehr N (eds.) *Conservatism. A Contribution to the Sociology of Knowledge*. Routledge & Kegan Paul, London]

Meja V, Stehr N (eds.) 1990 *Knowledge and Politics: The Sociology of Knowledge Dispute*. Routledge, London [1982 *Der Streit um die Wissenssoziologie*. 2 Vols. Suhrkamp, Frankfurt am Main, Germany]

Merton R K 1957 *Social Theory and Social Structure*. The Free Press, Glencoe, IL
Santambrogio A 1990 *Totalità e Critica del Totalitarismo in Karl Mannheim*. Angeli, Milano, Italy
Weber M 1919/1922 *Wissenschaftslehre*. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany
Wolff K H (ed.) 1993 *From Karl Mannheim*, 2nd extended edn. Transaction Publishers, New Brunswick, NJ

V. Meja

Maoism

Maoism refers primarily to the ideology, politics and writings of Mao Zedong (1893–1976; also romanized Mao Tse-tung). In official Chinese discourse the term 'Mao Zedong Thought' (*sixiang*) is used rather than 'Maoism (*zhuyi*), out of deference to Marxism–Leninism. It can also refer to the ideology of groups who take Mao Zedong as a political model, or to the official role of Mao in the orthodoxy of the People's Republic of China. Mao considered himself to be creatively applying the general theory of Marxism–Leninism to Chinese conditions, and the inextricable interplay of his politics and thought make it difficult to abstract a theoretical essence of Maoism. Because of the radical turn in Mao's politics in his later years, most prominently in the Cultural Revolution, the content and effect of Maoism underwent major changes. Since Mao's political achievements were the prerequisite for his ideological role within China and for his attractiveness as a model for radical political movements elsewhere, Maoism beyond Mao has had a varied and shifting ideological content.

1. *Mao Zedong's Theory and Practice*

Mao never considered himself primarily a theoretical innovator. Indeed, the ideas that he was most proud of, namely, the unity of theory and practice and the particularity of contradiction, both emphasized the primacy of practice and they emerged as general reflections on his political experience. Correct revolutionary leadership—which was Mao's principal concern—depended on a practical dialectic involving an understanding of one's situation derived from grassroots investigation and the experimental application of an ideological plan of attack. Although Mao considered himself to be operating within the framework of Marxism–Leninism, his respect for the problems of practical application went well beyond Marx's philosophical commitment to the unity of theory and practice or Lenin's instrumental notion of tactics.

1.1 *Basic Background*

Mao was an intellectual among peasants and a peasant among intellectuals. In a rural country suffering total chaos, the combination was the key to his success. In a postrevolutionary party–state attempting to modernize, it was the source of catastrophic political interventions.

Born in the hinterland of Hunan, a rural province, Mao rose quickly to leadership among his cohort of scholars in the provincial capital. From the beginning, his leadership combined practical activism with voracious reading and intellectual engagement. His first published article, emphasizing the importance of physical education, appeared in the leading national journal of young intellectuals in early 1917. Mao was a provincial leader in the May Fourth Movement of 1919, the abrupt and radical beginning of mass politics in China, and he became a founding member of the Communist Party of China (CPC) in 1921. The CPC began as an urban and cosmopolitan movement inspired by Bolshevism, and it was almost by accident that Mao rediscovered his rural roots and emphasized the importance of the peasantry to a meaningful revolution in China. *The Hunan Report* (1927), one of his most famous works, captures the enthusiasm of his rediscovery of rural revolutionary potential in the context of optimism about an impending national revolution.

Instead of victory, however, the CPC was destroyed in 1927, and remnants fled to remote rural areas. In the new, bleak context of survival in the countryside, Mao began to work out practices of land reform and guerilla warfare, and his success eventually led to his leadership of the CPC after 1936 and the formulation of his strategy of rural revolution from 1937 to 1945. In conformity with his own experience, Mao's strategy was not a theoretical magic bullet but rather an exhortation to cadres to go down to the basic level and to integrate the Party's leadership with the concrete concerns, limitations, and capacities of each village. From aggregating the primal politicization and militarization of villages came the strength that overwhelmed the government of Jiang Jieshi (Chiang Kaishek) in 1949.

The establishment of the People's Republic of China confirmed Mao's confidence in himself and in Marxism–Leninism, but it also set unfamiliar tasks of consolidation and construction in an unfamiliar environment in which cities, bureaucrats and intellectuals would necessarily play leading roles. The Soviet Union therefore became the model for the initial phase of China's development. By 1956, however, Mao's unease about developments in European communism and his growing confidence in China's own experience led him to radical experimentation in economics and politics. Having completed the transition to socialism, Mao took his goals for continuing the revolution from Marx's brief descriptions of a

communist utopia: critique of 'bourgeois right,' abolition of the differences between town and country and between mental and manual labor, direct mass participation in politics, and so forth. The Great Leap Forward (1958–1961) and the Cultural Revolution (1966–1969) were as much the tragic consequence of Mao's unquestioning commitment to Marxism as they were the result of personal idiosyncrasies. In both cases Mao retreated from the unanticipated consequences of his interventions, but he did not adjust his theoretical framework to account for the failures.

1.2 Essence of Maoism

Mao was not articulate about the basic principles of his thought; from 1921 Marxism–Leninism provided his explicit theoretical framework. Nevertheless, three values are present throughout his career and underpin his commitment to communism: revolutionary populism, practical effectiveness, and dialectics.

Revolutionary populism stems from the conviction that the mobilized masses are ultimately the strongest political power. All exploiters are minorities and therefore vulnerable. The strength of the revolution comes primarily from its closeness to the masses rather than from doctrine or tactics. For Mao, the mass line was not a slogan, but rather a commitment to interactive, mass-regarding behavior by the Party that would maximize its popular support. Mao's successful strategy of protracted rural revolution is the best example of his revolutionary populism; his encouragement of the Red Guards to 'bombard the headquarters' in the Cultural Revolution showed the impossibility of applying revolutionary populism within a regime.

Mao's concern for practical effectiveness is the most prominent characteristic of his political and intellectual style. Mao insisted on personal, concrete investigation and once raised the slogan of 'No investigation, no right to speak.' Clearly Mao's empiricism had dwindled by the time of the Great Leap Forward, when visiting a model commune was considered sufficient evidence for launching a national campaign, but he still adjusted his leftist interventions in the light of unwelcome facts about their effects.

Mao's dialectics had its roots in both Marxism and in traditional Chinese thought. Mao saw a changing world driven by internal contradictions rather than a static world of fixed categories. Dialectics was a flexible method of political analysis that allowed Mao to concentrate his attention on one problem while affirming its interrelationship to other issues. More importantly, the expectation that one situation would transform into another—even its opposite—led Mao to emphasize process in his politics rather than institutions. The ebb and flow of the campaign cycle was better than apparent stability. Likewise, Mao expected unity to emerge from struggle rather than from compromise.

2. Official Maoism

2.1 The Cult of Mao

By the end of 1936 Mao Zedong had become the primary leader of the CPC, the Long March had been concluded, and the war against Japan had begun. Already in his mid-forties, Mao rose to the challenge of distilling his experience into general principles and applying them to the new challenge of war against Japan. The first texts of his major theoretical writings, *Problems of Strategy in China's Revolutionary War*, *On Practice*, and *On Contradiction*, were produced as lectures to the Red Army College in 1936–1937. The Rectification Movement of 1942–1944 was a sustained attempt to inculcate Mao Zedong Thought as the authoritative standard of behavior. The theoretical dimension of Mao's leadership was confirmed at the Seventh Party Congress of 1945. The formulation of the Congress, that Mao Zedong Thought 'creatively applied Marxism–Leninism to Chinese conditions,' has remained the official summary of Mao's accomplishment. While it is deferential with regards to general theory, it made 'Marxism–Leninism Mao Zedong Thought' China's new orthodoxy.

With the founding of the People's Republic of China in 1949, Mao's public persona began a transformation from revolutionary leader to symbol of a new regime. Chinese tradition and Stalinist practice combined to expect a glorified, all-powerful leader, and the image reflected a tightening of Mao's control and a growing aloofness from collegial leadership. The four volumes of Mao's *Selected Works* were published in the 1950s, ironically increasing Mao's intellectual reputation at a time when he was drifting away from writing coherent, persuasive essays. But he was not falling asleep. His major essay from the 1950s, *On the Correct Handling of Contradictions Among the People* (1957), raised the question of how to maintain correct leadership after the revolution. Unfortunately, the rather liberal answer of this essay was replaced by a harsh one based on class struggle even before the essay was published.

The canon of Mao's writings and the cult of personality took a dramatic turn in the leftist era from 1957 to 1976. Mao was supreme leader, and so his public gestures (such as swimming the Yangtze in 1966) and 'latest directives' displaced any need for persuasion. In the chaos of the Cultural Revolution Mao and his thought became an icon of salvation. Statues, buttons, and the *Little Red Book* of quotations became omnipresent symbols of devotion. Meanwhile, various Red Guard groups published various unofficial Mao texts. The nadir of official adulation was his anointed successor Lin Biao's claim that Mao was a genius far above the rest of humanity. After Lin's failed attempt to assassinate Mao in 1971 the claims of omniscience receded, but brief quotations were printed

in red on the masthead of China's newspapers until after his death.

2.2 China after Mao

By the time of Mao's death in September 1976, support for revolutionary leftism had hollowed out to a clique of ideologues known (after their overthrow in October) as the 'Gang of Four.' The regime then faced the difficult challenge of claiming the legitimacy of continuing CPC leadership while abjuring leftism and the Cultural Revolution. In 1978 a speech by Mao was published in which he admitted that the Great Leap Forward was a mistake and that he was responsible for it. Meanwhile, Deng Xiaoping resurrected the Mao quotation 'Seek truth from facts,' and by 1979 had abandoned the remaining policies of the leftist era. In September 1979 Party elder Ye Jianying made the first public analysis of Mao's errors, including the 'appalling catastrophe' of the Cultural Revolution, and his viewpoint was formalized in the Party's 1981 *Resolution on Party History*. The verdict was that Mao was 'seventy percent right and thirty percent wrong,' but in effect he was held to be correct for the first 70 percent of his life and incorrect for the last 30 percent.

The general framework of honoring but criticizing Mao has led to an explosion of available Mao texts, memoirs by those who knew him, and historical research. Since Mao can now be treated as history rather than as sacred authority, texts do not have to be adjusted for current politics as were the texts in the *Selected Works*. However, the cottage industry of academic Mao studies and the popularity of Mao charms and votive pictures should not be mistaken for a continuing influence of Maoism. The Cultural Revolution was too recent and too brutal, and the rural revolution is too distant from current challenges. For the foreseeable future Mao's primary function will be as a teacher by negative example. But Mao Zedong's thought and his example will remain deeply planted in China's heritage.

3. Maoism beyond China

Ever since the visit of Edgar Snow to Yan'an in 1936 produced *Red Star over China*, Mao has been known to the outside world and outsiders have interacted with developments in China. Snow's rendering of Mao's autobiographical interview was translated into Chinese and served for decades as Mao's unofficial biography. On a less flattering note, Stalin quickly detected Mao's originality and maintained a private attitude of derisive distrust towards him. By the same token reversed, in the 1940s many international communists considered Mao and Tito to be the most promising new leaders. The success of China's rural revolution increased Mao's credibility, and countries as distant as Bulgaria were inspired by the Great Leap Forward to attempt more modest leaps themselves.

The introverted radicalization of Chinese politics in the 1960s, highlighted of course by the Cultural Revolution, changed the pattern of external influence. On the one hand, China seemed less attractive and more dangerous to other countries, including other communist countries like Vietnam. On the other hand, those who sought radical political solutions idealized China and Mao. The fearless mobilization, the radical egalitarianism, the thorough rejection of bourgeois values, and the sheer vastness and excitement of the Cultural Revolution induced people to form splinter parties from establishment communist parties. Mao was a major inspiration to the Khmer Rouge in Cambodia and to the Shining Path (*Sendero Luminoso*) in Peru. More broadly, by demonstrating that great things could happen if people massed together Maoism contributed to the world political ferment of the 1960s.

See also: China: Sociocultural Aspects; Chinese Revolutions: Twentieth Century; Communism; Communist Parties; Marx, Karl (1818–89); Marxian Economic Thought; Marxism and Law; Marxism/Leninism; Marxist Social Thought, History of; Revolutions, Theories of

Bibliography

Barme G R 1996 *Shades of Mao: The Posthumous Cult of the Great Leader*. Sharpe, Armonk, NY

http://www.maoism.org/ Text of selected works and links to current Maoist sites

Knight N 1990 *Mao Zedong on Dialectical Materialism: Writings on Philosophy, 1937*. Sharpe, Armonk, NY

MacFarquhar R et al. 1989 *The Secret Speeches of Chairman Mao: From the Hundred Flowers to the Great Leap Forward*. Harvard University Press, Cambridge, MA

Mao T 1965 *Selected Works of Mao Tse-tung* (Originally four volumes, with a fifth added in 1977). Foreign Languages Press, Beijing

Martin H 1978 *Kult und Kanon: Entstehung und Entwicklung des Staatsmaoismus 1935–1978*. Institut für Asienkunde, Hamburg

Mehnert K 1969 *Peking and the New Left: At Home and Abroad*. University of California, Berkeley, CA

Schram S 1969 *The Political Thought of Mao Tse-tung*, rev. and enl. edn. Praeger, New York

Schram S R (ed.) 1992/1999 *Mao's Road to Power: Revolutionary Writings 1912–1949*. Sharpe, Armonk, NY, Vol. 5

Schwartz B I 1951 *Chinese Communism and the Rise of Mao*. Harvard University Press, Cambridge, MA

Short P 2000 *Mao: A Life*, 1st American edn. Henry Holt, New York

Womack B 1982 *Foundations of Mao Zedong's Political Thought, 1917–1935*. University Press of Hawaii, Honolulu

B. Womack

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Mar

Marital Interaction: Effects on Child Development

Socialization research is frequently focused on the effects of parent–child interaction on child development, neglecting the impact of broader family influences, such as marital relations. In fact, associations between marital relations and child adjustment have been reported at least since the 1920s.

1. Marital Discord and Children's Adjustment

The publication of a landmark review by Emery (1982), and influential papers by Rutter and colleagues (e.g., Rutter and Quinton 1984), greatly increased recognition of the significance of relations between marital discord and child adjustment. Accordingly, a particular concern in this area has been the effects of marital conflict on child adjustment.

1.1 Children's Psychological Problems

Relations between marital conflict and children's externalizing problems (e.g., aggression, conduct disorders) have frequently been reported. However, the introduction of more sensitive measures of children's functioning and adjustment has expanded the purview of the possible effects of marital conflict. Thus, evidence for links with internalizing problems (e.g., depression, anxiety) and for relations with children's academic problems and difficulties in peer relations has accumulated in recent years (Grych and Fincham 1990).

1.2 Relations with Other Forms of Family Dysfunction

Marital conflict co-occurs with other forms of family dysfunction (e.g., physical and sexual abuse, Appel and Holden 1998) and is frequently associated with the effects of other family problems on children's adjustment. For example, it has been identified as a significant factor in accounting for the effects of divorce on children's adjustment (Amato and Keith 1991), with high marital conflict linked with adjustment problems in children before, during, and after divorce. Marital conflict is also implicated in the effects of parental depression on child adjustment, and is more closely linked with problematic outcomes in these families than parental depression for some types of adjustment problems (Downey and Coyne 1990).

2. Direct and Indirect Effects of Marital Conflict on Children's Functioning

The research indicates that children are affected by marital conflict (a) owing to exposure to these situations, and (b) because of the effects that marital conflict has on children via changes in parenting practices.

2.1 Effects of Exposure to Marital Conflict

Unresolved interparental discord is distressing for children to observe, even solely as bystanders, from infancy through adolescence. Emotional distress is evident in children's self-reported and behavioral distress reactions (e.g., freezing, crying), physiological reactions (e.g., blood pressure), increased aggressiveness, and their efforts to resolve or otherwise ameliorate interparental conflicts (Cummings and Davies 1994). Children's problems in emotional regulation when coping with anger, including interpersonal anger (Fabes and Eisenberg 1992), have been implicated in their development of adjustment problems (Davies and Cummings 1998).

2.2 Marital Conflict and Parenting

In addition to effects caused by exposure to marital conflict, children may also be affected by changes in family functioning that are the sequelae of marital conflict. In particular, links between marital conflict and a variety of problems in parenting are reported, including parent–child discipline problems, problems in child management, and insecure parent–child attachments (Erel and Burman 1995). Relatively little research has been carried out to demonstrate why parenting is affected, but presumably insensitivity to children or other effects due to the preoccupying demands of conflict with the spouse, and the spillover of interparental hostility to interactions with the children, are involved (Cummings and Davies 1994).

2.3 Age and Gender Differences

Children evidence increased social understanding and disposition towards involvement in marital conflict as they get older. Greater links with internalizing

problems for girls and externalizing problems for boys are frequently reported. However, recent studies suggest that gender differences in the effects of marital conflict change with age, and the pattern of findings for age and gender is quite inconsistent across the body of research in this area. Thus, while effects across age and gender are often reported, one cannot conclude from current work that any one age or gender is more vulnerable to adjustment problems caused by high marital conflict (Cummings and Davies 1994).

3. *The Term 'Marital Conflict'*

However, the relations found are, in part, a function of how marital conflict is defined. Relations with children's adjustment problems are evident when marital conflict is defined as the frequency of verbal or physical hostility between the parents. On the other hand, when marital conflict is defined more broadly as how the parents handle their differences, both positive and negative effects on children's functioning are reported.

4. *It is not Whether but How the Parents Express Disagreements that Matters*

New developments indicate that any assumption, either implicit or explicit, that marital conflict is a homogeneous stimulus oversimplifies matters considerably. Current research suggests that some forms of marital conflict are moderately or even highly associated with children's adjustment problems, whereas other forms of marital conflict may not at all be linked with problems in the children. That is, the effects of marital conflict vary considerably in terms of how it is expressed. Accordingly, research has begun to focus on advancing understanding of the distinction between destructive and constructive forms of marital conflict expression from the perspective of the child.

4.1 *Distinguishing Between Destructive and Constructive Marital Conflict*

The task of differentiating between destructive and constructive marital conflict from the children's perspective is in a relatively early stage of study. A first step is to decide upon criteria for making such distinctions. Among the possible criteria for categorizing conflicts as destructive are whether the behavior is linked with (a) children's adjustment problems, (b) problems in parenting, or (c) distress reactions seen in the children during exposure. Conversely, bases for regarding behaviors as constructive include (a) the absence of links with adjustment or family problems, including the promotion of children's social competency, or (b) positive or neutral emotional reactions by the children during exposure to conflict, including behaviors that ameliorate children's negative emotional reactions (e.g., conflict resolution). By these criteria, aggression between the spouses; aggression towards objects; threatening statements (e.g., 'I'm leaving'); withdrawal or the silent treatment; intense, escalated, proliferating conflict; or blaming the child for marital problems qualify as destructive conflict behaviors. By contrast, conflict resolution; parental explanations to the children that conflicts have been worked out; or conflicts expressed by parents with emotional control and mutual respect merit categorization as constructive conflict behaviors. However, many of the parental behaviors that may occur during marital conflict have not yet been examined in this regard. Moreover, these various criteria for classifying marital conflict may sometimes yield different conclusions regarding the same behaviors, and do not readily apply to complex interparental communications in which multiple behaviors may occur. A well-articulated theoretical model for classifying constructive versus destructive conflict, and empirical evidence to support such a model, are required to support a more advance perspective on this distinction.

4.2 *Explanations for Why Marital Conflict is Linked with Child Development*

A significant challenge is to identify the causal bases for links between marital conflict and child outcomes. Pathways of influence within families are multivariate and complex, and the multiple directions of influence of marital conflict on children are difficult to disentangle (Cummings et al. 2000). Contrary to a common sense expectation that repeated exposure to marital conflict would be linked with habituation to conflict, accumulating evidence indicates that high levels of marital conflict are associated with sensitization to marital conflict and other family stressors. That is, high levels of marital conflict are linked with children's greater distress, anger, and aggression in response to marital conflict, and an increased tendency to intervene in parental disputes. At the level of theory, Davies and Cummings (1994) have proposed that these response are indicative of children's greater insecurity about marital relations in homes characterized by high marital conflict, which contributes to the greater risk for adjustment problems among children from such homes.

5. *Methodologies for Studying the Effects of Marital Conflict on Children*

Examining the effects of marital conflict on children poses significant methodological challenges. For obvious ethical reasons, one cannot test causal propositions by means of a traditional experimental research design. That is, one cannot randomly assign children to groups and induce significant marital conflict in front of the children in the experimental

group in order to observe its effects. Moreover, each of the methodologies that have been developed and are appropriate for the study of this issue (e.g., questionnaire methodologies; laboratory observations of marital interactions; analog presentations of marital conflict; or home observation) have both significant methodological strengths, and weaknesses, in absolute terms and in relation to the other methodologies (Cummings et al. 2000).

6. *Future Directions*

The upshot is that significant further advances towards understanding the effects of different forms of marital conflict on children will probably require the simultaneous and coordinated use of multiple methodologies. The study of causal relations between different forms of marital conflict and child development also requires that effects be examined prospectively over time. A future direction is to test theoretical propositions about why marital conflict affects child development in the context of prospective, longitudinal research designs, which is urgently needed to advance understanding of the nature of the causal processes and pathways of influence on the children. There is also a need for greater study of other positive dimensions of marital relations, which may have directly beneficial effects on children's functioning, or serve to buffer children from the negative impact of marital discord. Little is known about how other familial (e.g., the quality of relations between a child and a sibling) or extra-familial (e.g., the quality of relations with a grandparent) relationships affect the links between marital conflict and children's functioning. An urgent need is to include fathers more directly in research, and to understand more about the child development outcomes linked with the role of fathers in the context of marital relations. Research has focused on affluent, Caucasian, and Western cultures. Greater understanding will be promoted by more cross-cultural research and the inclusion of samples with more diverse ethnic and socioeconomic characteristics.

See also: Child Care and Child Development; Divorce and Children's Social Development; Divorce and Gender; Early Childhood: Socioemotional Risks; Fatherhood; Nontraditional Families and Child Development; Parenting: Attitudes and Beliefs; Queer Theory

Bibliography

Amato P R, Keith B 1991 Parental divorce and the well-being of children: A meta-analysis. *Psychological Bulletin* **110**: 26–46

Appel A E, Holden G W 1998 The co-occurrence of spouse and physical child abuse: A review and appraisal. *Journal of Family Psychology* **17**: 578–99

Cummings E M, Davies P 1994 *Children and Marital Conflict: The Impact of Family Dispute and Resolution*. Guilford Press, London

Cummings E M, Davies P T, Campbell S B 2000 *Developmental Psychopathology and Family Process*. Guilford Press, London

Davies P T, Cummings E M 1994 Marital conflict and child adjustment: An emotional security hypothesis. *Psychological Bulletin* **116**: 387–411

Davies P T, Cummings E M 1998 Exploring children's emotional security as a mediator of the link between marital relations and child adjustment. *Child Development* **69**: 124–139

Downey G, Coyne J C 1990 Children of depressed parents: An integrative review. *Psychological Bulletin* **108**: 50–76

Emery R E 1982 Interparental conflict and the children of discord and divorce. *Psychological Bulletin* **92**: 310–30

Erel O, Burman B 1995 Interrelations of marital relations and parent-child relations: A meta-analytic review. *Psychological Bulletin* **188**: 108–32

Fabes R A, Eisenberg N 1992 Young children's coping with interpersonal anger. *Child Development* **63**: 116–28

Grych J H, Fincham F 1990 Marital conflict and children's adjustment: A cognitive–contextual framework. *Psychological Bulletin* **108**: 267–90

Rutter M, Quinton D 1984 Parental psychiatric disorder: Effects on children. *Psychological Medicine* **14**: 853–80

E. M. Cummings

Market and Nonmarket Allocation

While markets are pervasive forms of social organization, there are many goods and services which we do not distribute through markets. For example, we allow people to buy, sell, and trade cars and shirts, but market exchanges of votes, sex, and kidneys are banned. What reasons can be given in favor of using and refraining from using markets? Many debates in the twentieth century have centered around this question. In this article, the myriad of reasons to use or refrain from using markets—including efficiency, distributive justice, and the effects of markets on democratic institutions, people, and culture—are considered. This article also examines ways of enriching the list of distributive alternatives beyond the two poles of centralized plan and market.

1. *What is a Market?*

Modern economic textbooks treat markets as mechanisms for the production, circulation, and valuation of goods and services. As such, they have three noteworthy features: they are efficient, impersonal, and oriented to individual preference and choice. Let us consider these features in more detail.

Since rational individuals will only exchange goods and services when they have something to gain, markets will be efficient—purging the economy of the

less desirable goods and moving the trading parties to preferred positions. Economists refer to this property of markets in terms of 'Pareto optimality.' A state of affairs is Pareto optimal when no one can be made better off (i.e., their utility increased) without someone else being made worse off (i.e., their utility decreased). An important result in economics, the so-called 'fundamental theorem of welfare economics,' shows that in a world in which everyone could trade everything—including futures, uncertainty, and so on—the allocation of resources would be Pareto optimal. This theorem formalizes the earlier conjectures of Adam Smith and others, that participation in competitive markets would lead rational self-interested individuals to produce, as an unintended consequence of their private actions, collectively good results.

Markets are also impersonal, suitable to regulating the interactions of strangers. They abstract away from particular features of a participant, such as his or her race, religion, or sexual orientation and focus only on his or her money income. The impersonality of market relations means that each individual is indifferent to his or her trading partners at any time. He or she has no precontractual obligations to those he or she trades with, and is free to consider only his or her own interests in pursuing an exchange.

In market exchange, the parties also express their own individual choices, pursuing their own sense of what is good, independent of what the state or others value. Markets are thereby conducive to individual freedom: they accommodate individual preferences without passing judgment upon them. By decentralizing decision making, markets give individuals a significant measure of freedom over their circumstances. Furthermore, one is free to walk away from any exchange—the same deal can be struck elsewhere, as it were. Albert Hirschman (1970) has termed this latter feature the power of 'exit,' to distinguish it from the power of 'voice' (i.e., speaking up).

1.1 A Contrast

The modern, neoclassical conception of economics, developed in the late nineteenth and early twentieth centuries by Edgeworth, Robbins, and Jevons, treats the market as something obvious and simple. In the standard neoclassical general equilibrium model of a market economy formulated by Leon Walras (1926/1954), commodities are identical, the market is concentrated at a single point in space, individuals have complete information about the exchange commodity, and exchange is instantaneous. There are, accordingly, no costs in enforcing agreements and no important distinctions among kinds of markets.

The neoclassical treatment of markets has produced important mathematical results such as the fundamental theorem, and also allowed for the development of powerful models which can characterize the equilibrium properties of an entire economy. But it depends on viewing markets as economic mechanisms whose operation can be understood in abstraction from the particular kinds of goods being exchanged and the relationships between the participants. In the neoclassical conception of markets, it does not matter whether what is being traded is oranges, sexual services, or credit—the nature of market exchange is seen as the same in each case. The participants are considered interchangeable and, given the opportunities each has for exit, no party can exercise significant power over another. In Paul Samuelson's words, 'in a perfectly competitive market, it does not matter who hires whom: so let labor hire "capital."'(Samuelson 1957, p. 894) The result is that the market is viewed as a private realm, whose smooth operation leaves no room for political intervention: intervention is justified only when markets in some goods fail to be established or are inefficient.

By contrast, the classical economists–among them Adam Smith, David Ricardo, and Karl Marx—offered distinct theories of the operation of markets in different domains and focused on the opposing interests of the three main social classes—landlords, capitalists, and workers—who came together to trade and between whom the social wealth was to be distributed. Not only were markets viewed as political, in the sense that they depended on legal property rights, but it was also recognized that their functioning in certain realms inevitably raised questions relevant to the structure of public life. For example, the classical economists noted that the market between buyers and sellers of goods such as oranges differs importantly from the market between buyers and sellers of human labor power. This is because the latter kind of market involves a good which has the unusual status of being embodied in a human being. While we do not have to worry about the noneconomic effects of a market on fruits, we do need to worry about whether a particular market in human labor power produces passivity, alienation, greed, or indifference to others.

The classical economists saw the marketplace as a political and cultural as well as an economic institution, where differing interests wielded power and where some types of exchanges had constitutive effects on their participants. Adam Smith (1776/1937) worried not only about the tendency of the owners of capital to collude and thus block market competition, but also about the effects of a fragmented and monotonous division of labor on the ability of workers to participate in political life.

2. Criteria for Assessing the Use of Markets

The most famous modern debates about the legitimacy of market allocation versus its alternatives have focused on its efficiency. Consider, for example, the

debate between socialist economist Oskar Lange and conservative economist Friedrich Hayek which took place during the 1930s, when the Soviet experiment was underway. Lange and Hayek debated whether a noncapitalist economy with few market structures could experience economic growth. Lange argued that an ideal planned economy could allocate resources in a more optimal manner than a perfectly competitive market economy. In rebuttal, Hayek pointed out that planned economies have a tendency towards bureaucratic sclerosis, a lack of accountability and high information processing costs.

Each side raised important considerations. But from many perspectives their debate about markets can be criticized as narrow and conceptually impoverished. The debate not only tended to operate as if there were just two choices—command centralized planning and the neoclassical rendition of the market as perfectly efficient—but also posed the issue about market or nonmarket choice solely in terms of efficiency.

If we see the market as a complex heterogeneous institution, which brings people into different kinds of relations with each other for different exchanges, then we can also see why the use of markets can raise other concerns besides efficiency. From this broader perspective, not only efficiency but also the effects of the market on the political structure of power and on human development are relevant to determining when and where markets should allocate goods and services and where other institutions should be used.

2.1 *Efficiency*

Through the use of prices, markets indicate what millions of goods and services are worth to sellers and buyers. They thereby function to apportion resources efficiently: they signal to sellers what to produce, to consumers what to buy, and to capitalists where to invest. The continual adjustments of supply and demand, registered in changing prices, allow markets to 'clear' inventory. When inventory is cleared, there is no excess demand or supply: supply equals demand at some price. The efficiency benefits that markets produce constitutes an important *prima facie* case for their inclusion in an economy. Indeed, we know of no mechanism for inducing innovation on an economy-wide basis except market competition.

In certain circumstances, however, markets themselves fail to be efficient. While in the neoclassical textbook model of market exchange transaction costs are taken to be zero, and each participant to have perfect information and act solely on the basis of his or her rational self-interest, it is well recognized that these assumptions do not always hold in reality. Thus, the standard theorems connecting markets and efficiency are coupled with a theory identifying 'market failures,' circumstances in which the utility-enhancing effects of markets are either inoperable or eroded. For example, markets fail when and where certain effects of an exchange—which economists call 'externalities'—are not fully absorbed by the exchanging parties and are passed on to third parties. In some circumstances, these external costs constitute public bads: the defacement of the landscape, pollution, and urban decay are notable examples. Markets also fail where there are natural monopolies, or technologies which give rise to economies of scale, thus making only monopolistic or oligopolistic firms viable.

Recent work within mainstream economics has shown that market failure may be more widespread than previously supposed. This work emphasizes the ubiquity of imperfect and asymmetric information held by the exchanging parties (Akerlof 1970) as well as the fact that in many exchanges the parties cannot costlessly control one another's behavior. (Consider the costs of supervision that employers pay to make sure that their workers are actually working.) Economic models now take into account the incentives agents have to manipulate information and outcomes in order to press their own advantages. Thus, agents may act as if they were willing to pay less for goods and services than they really are, or threaten to withdraw from a trade entirely. It is now recognized that people can act in ways that prevent markets from clearing even in competitive equilibrium—there will be persistent excess supply or excess demand. This is because if one party wants power over the other, he or she can achieve this by giving the other something to lose, so that it now costs something to 'exit.' Thus, because not all prospective borrowers can find loans and not all those who seek jobs can find them, those who do have a strong interest in complying with the banks or owners (Stiglitz and Weiss 1981). The power that banks thus acquire over their creditors and owners acquire over their workers gives these markets a political character. One might say that this recent work emphasizing strategic action, incomplete information, and transaction costs gives mathematical precision to some of the earlier insights of the classical economists.

Various solutions to market failure have been proposed including political intervention in the marketplace, expansion in the number of markets (e.g., allowing externalities themselves to be priced by the market) and the use of nonmarket alternatives such as bargaining and negotiation. The point of such solutions is to restore efficiency.

In thinking about the ways that markets contribute to efficiency, we should note that the actual efficiency outcomes of the market depend on the initial distribution of endowments and assets. This initial distribution is not itself part of the definition of Pareto optimality. A state of the world can be Pareto optimal, even if its original distribution was based on theft and fraud or was otherwise unfair. In order to justify market outcomes, or to use Pareto optimality normatively, the theory of market efficiency needs to be

supplemented with a theory of distributive justice. Some theories of distributive justice focus on the legitimacy of the initial distribution of assets while others focus on the legitimacy of the distributive inequalities that the market produces.

A related point is that efficiency interpreted as Pareto optimality has only modest moral content. This is true for several reasons. First, Pareto optimality deals exclusively with efficiency and pays no attention to distribution. A Pareto optimal state can thus be a state of social misery for some. Simply consider the case in which the utility of the downtrodden cannot be raised without lowering the utility of the millionaires. Second, the Pareto criteria dispenses with interpersonal comparisons, so that we cannot compare the contribution a meal makes to the poor man's utility with that it makes to a millionaire's.

2.2 Freedom

Market allocation is often defended in terms of freedom. Proponents cite a range of important effects which markets have on an individual's ability to develop and exercise the capacity for liberal freedom. Each of these effects, they argue, depends on leaving the market domain largely cordoned off from political interference. Markets (a) present agents with the opportunity to choose between a large array of alternatives, (b) provide incentives for agents to anticipate the results of their choices and thus foster instrumental rationality, (c) decentralize decision making, giving to an agent alone the power to buy and sell things and services without requiring him or her to ask anyone else's permission or take anyone else's values into account, (d) place limits on the viability of coercive social relationships by providing (relatively) unimpeded avenues for exit, (e) decentralize information, (f) (may) enhance an individual's sense of responibility for his or her choices and preferences, (g) allow people to practice and try out various alternatives, (h) create the material wealth which is a precondition for the possibility of having decent alternatives, and (i) allow liberal individuals to cooperate without having to agree on a large range of questions of value (Friedman 1962, Nozick 1974).

Liberal theories which assign substantial weight to individual freedom thus allot a central role for market allocation, pointing to the market realm as a place where the capacity for individual choice—indeed, where the liberal individual—is developed. As liberal theorists are likely to emphasize, respect for markets in goods and services can be an important way of respecting individual (and divergent) conceptions of value. In a market system, there is no preordained pattern of value to which individuals must conform and exchange gives to individuals the freedom to pursue distinct aims. In fact, the market leaves people free not to choose to sell or buy a particular good on the market.

The liberal theory of freedom is essentially (although not entirely) a negative one, emphasizing the space markets provide to individuals to pursue their private ends free from external intrusions by the state or other people. This liberal theory linking freedom and markets has been criticized for not attending to the preconditions for its own realization. These preconditions include the existence of some realms which are protected from the market. For example, most liberals recognize that it is incompatible with liberal freedom to allow a person to sell him- or herself into slavery; free people require rights to consultation, self-judgment and control over the conditions in which they act. But many liberals fail to see that freedom also has implications for other kinds of contracts (e.g., bans on desperate exchanges such as those involved in kidney and baby sales). Some theorists have argued that the securing of liberal freedoms requires the guaranteed provision of a minimal level of basic goods such as healthcare, food, and housing and thus requires the regulation of markets.

2.3 Human Flourishing

Market allocation has often been defended on the grounds that markets, by satisfying people's preferences, contribute directly to human happiness and flourishing. In addition, by stimulating economic growth, markets have the potential to eradicate the poverty and hardship which is everywhere the enemy of a decent quality of life. Certainly, the market makes important contributions to the satisfaction of human wants. But, in considering the effectiveness of the market in maximizing overall human well being, several factors need to be taken into account. First, satisfaction of the subjective preferences of an individual agent may not provide that agent with well being. A lot depends on the content and nature of his or her preferences. Agents may have preferences for things that do not contribute to their overall good, preferences which are mere accommodations to their circumstances, preferences they themselves wish they did not have, preferences based on false information, and preferences which are instilled in them by the market. In some such examples, the case for regulation can be made on paternalistic grounds, excluding exchanges born of desperation, (such as selling oneself into slavery), lack of information, and imprudence.

Second, the market recognizes only those preferences which can be backed up by an ability to pay. The poor man's preferences for a good meal are invisible. Indeed, the notion of 'preference' itself does not distinguish between kinds of preferences, only degrees of strength. Thus, the market cannot distinguish between urgent human needs (the poor man's desire for nourishment) and mere whims (the rich man's desire for a fine burgundy). Third, there are important sources of satisfaction that do not go through the market: public goods that, by their nature, do not go

through the market and interpersonal relations which may be spoiled if they become part of the market's transactional network. So care must be taken to avoid ignoring those preferences that are not expressed on the market.

Some critics have claimed that, even if this care is taken, the use of markets promotes an inferior form of human flourishing. This criticism has a wide interpretation and a narrow interpretation. On the wide interpretation, markets everywhere shape and socialize people in the wrong way. They are taken to do this by promoting selfish behavior, making people more materialistic, and dulling people to important distinctions of value to which they should be responsive.

This worry about markets is sometimes posed in terms of the metaphor of infection—that market norms and relations will spill over and contaminate nonmarket realms such as friendship and love. Thus, it has been alleged that markets erode our appreciation of the true value of other people, since they lead us to think of goods and people as exchangeable items. This wide interpretation of the market's negative effects on human flourishing has only weak social scientific support. There is little evidence that people are more materialistic in market societies than they were in peasant economies, that they devalue love and friendship, or that they are now less likely to engage in moral behavior than in the past (Lane 1991).

A narrower interpretation of this humanistic criticism is that some (but not all) markets have bad feedback effects on particular human behaviors. Studies have shown, for example, that economics and business students (who presumably are more likely to govern their behavior by the axioms of neoclassical economics than are literature or sociology students) are uniquely uncooperative in solving collective action problems (Marwell and Ames 1981).

One important negative feedback effect was identified by Richard Titmuss (1971). Titmuss claimed that allowing blood to be sold and bought as a commodity would have two negative effects: (a) fewer people would give blood for altruistic reasons and (b) the quality of blood would thereby be lower. Titmuss' main thesis concerned the first effect: he argued that over time, a blood market would drive out altruistic blood donation by turning the 'gift of life' into the monetary equivalent of 50 dollars. People would become less willing to give blood freely as market distribution became more prevalent, for their gift loses its benevolent meaning. (Blood quality would be lower since altruistic donors have no incentive to lie about the quality of their blood, while commercial donors clearly do.)

Titmuss aimed to show that markets have broader incentive effects than economists have supposed. His account also challenged the liberal theory of freedom. In his view, allowing a market in blood does not merely add one additional choice to the prospective donor (who can now sell as well as give blood), there is also a feedback mechanism that diminishes the likelihood of donors giving blood for free. If this is so, then sometimes market regulation—including prohibition of certain exchanges—may shore up liberal freedom by allowing us to get the choice that we most want. (Consider the role of minimum wage laws and prohibitions on vote-selling from shrinking people's opportunity sets over time.)

Another possible negative feedback on motivation was the concern of diverse thinkers such as Adam Smith, Karl Marx and John Stuart Mill, who each worried that labor markets which rendered workers as mere appendages to machines would cripple their capacity for independence and political participation. Each thinker noticed that labor market exchanges had a constitutive effect on the parties to the exchange: workers who engaged in servile, menial, and degrading work were likely to be servile citizens.

Where markets have significant feedback effects on human motivation, there is a strong case for social regulation. It is circular to justify markets on the grounds that they maximally satisfy individual preferences in cases where those preferences are themselves shaped by the market. At the very least, recognition of the endogenous nature of some preferences leaves the question of institutional design open since there may be other mechanisms that generate and satisfy more preferences.

2.4 The Nature of the Goods Exchanged

Liberals have traditionally argued that many of the problems that unregulated markets cause with respect to the values of efficiency, justice, freedom, and well-being can be attenuated through state intervention. We can redistribute income and regulate consumption through taxation, publicly provide for social goods such as schools and roads, and establish a minimum level of healthcare for all citizens. Markets can be supported by other social institutions that encourage values such as honesty, reciprocity, and trust.

There is a different criticism of the market that cannot be addressed through market regulation but requires that the use of the market be blocked. This criticism focuses not on the negative consequences of market exchanges but on the ways that markets undermine the intrinsic nature of certain goods. The theorists who make this criticism reject the use of markets in certain domains and for certain goods categorically.

Markets are often taken to be neutral means of exchange. By this proponents mean that the market does not distinguish between my ethical valuation of a good and yours. I may value my Bible infinitely while you assign it a dollar price, but the market is compatible with both of our understandings of this good. The market responds only to effective demand, not to the reasons people have for wanting things.

When the market puts a price on a thing, it leaves the ways that people value it untouched.

In reality, this view of market neutrality is overstated. Some goods are changed or destroyed by being put up for sale. The most obvious examples of this phenomenon are love or friendship. A person who thought that they could buy my friendship would simply not know what it means to be a friend. A proposal to buy love, prizes, honors, friends, or divine grace is conceptually incoherent: it is the nature of these things that they cannot be bought.

Various arguments have been given to show that a wide range of goods are not properly conceived of as market commodities. Some of these arguments depend on the idea that certain goods are degraded when sold, or that some sales are inherently alienating to, or exploitative of, the seller. Items that have been proposed to be withheld from the market in order to preserve their nature include sex, reproductive labor, adoption, military service, healthcare, and political goods such as votes and influence.

Consider sex and reproductive labor. Some feminists have argued that markets are inappropriate for exchanging goods that place people in personal relations with each other. They argue that when sexual relations are governed by the market, such as in prostitution, not only is an inferior good produced since commercial sex is not the same as sex given out of love, but also that the dignity and freedom of the prostitute is compromised. Because practices such as prostitution are alleged to allow women's bodies to be controlled by others, these arguments conclude that women's freedom requires that their sexual and reproductive capacities remain market-inalienable (Radin 1996).

Given the relatively poor economic opportunities open to many women, one may doubt whether paid pregnancy or prostitution is comparatively worse for women's freedom than the alternatives. For example, is prostitution more constraining than low-paid, monotonous assembly-line work? Does commercial surrogacy really place more extensive control over a woman's body than other acceptable labor contracts such as those given to professional athletes? To answer these questions may be more a matter of looking at the empirical consequences of markets in sex and reproduction, and the history of women's roles in the economy and the family, than of examining the intrinsic nature of these goods.

2.5 Democracy

A special case is often made for regulating or banning markets so as to ensure that democratic institutions are supported. This case has several components: (a) a theory of those political goods which must be provided to everyone equally in a democracy—goods such as votes, political rights, and liberties; (b) a theory of the economic preconditions of democracy; and (c) a theory of the kind of institutions which foster the development of people likely to support democratic institutions and function effectively in a democratic environment. It is easy to see that democracy is incompatible with property qualifications on the right to vote, slavery, and vote selling. Current debates center around whether democracy requires greater regulation of the political process including public funding of elections and campaigns, whether democracy requires greater regulation of the economy to allow for 'voice,' and whether expanding the market to education would enhance or inhibit the development of democratic citizens.

3. Regulation vs. Alternatives

Arguments that purport to show that markets in some good do not promote freedom or happiness or well-being or some other value do not directly enable us to conclude that such exchanges should be blocked. Even if markets interfered with or failed to promote certain values, interference with them might be worse overall from the point of view of those same values. Suppose, for example, that the only feasible institutional alternatives to markets all involved significant amounts of coercion. Furthermore, a small amount of economic inefficiency might be better than a whole lot of red tape. Conservatives frequently cite the costs of 'over-regulated' markets such as rent control and limits on parental choice with respect to education. So, a full assessment of market allocation must take into account the alternatives.

3.1 Other Forms of Market Regulation

In addition to regulating the market through redistribution of the income and wealth it produces, other forms of market regulation have been suggested. These include: restrictions on the transferability of income and wealth, as in ownership requirements which stipulate that owners must be active participants in the community in which their property is located, and restrictions as to what people can do to the things they can sell, as in historical building codes.

3.2 Nonmarket Allocation

There are important alternative mechanisms for distribution that do not rely on a market. In the twentieth century, the most important alternative allocation mechanism has been the government. Government decision has allocated such goods to individuals as citizenship, social security, social infrastructure such as roads, highways, and bridges, and education up to early adulthood.

Many allocative decisions are shaped neither by government nor by the market. These include distribution through gift, lottery, merit, the intrafamily regulation of work and distribution, and other principles such as seniority and need. Consider a few cases.

In the USA, the lottery has been used to allocate resources literally bearing on issues of life and death. Military service, especially during wartime, has often been selected for by lottery; the alternative has been to allow people to volunteer.

College admission, election to professional associations, and the award of prizes have been governed neither by market norms nor by state decree but by selective principles of merit and desert. Need has been central in allocating organs for transplantation.

In each of these cases, one can compare the allocation achieved with that which would have been achieved through a market, in terms of the broad list of criteria specified above.

See also: Decision-making Systems: Personal and Collective; Hierarchies and Markets; Market Areas; Market Structure and Performance; Markets and the Law

Bibliography

Akerlof G A 1970 The market for lemons: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* **84**: 488–500

Anderson E 1993 *Value in Ethics and Economics*. Harvard University Press, Cambridge, MA

Bowles S, Gintis H 1993 The revenge of *Homo Economicus*: contested exchange and the revival of political economy. *Journal of Economic Perspectives* **7**(1): 83–102

Downs A 1957 *An Economic Theory of Democracy*. Harper and Row, New York

Friedman M 1962 *Capitalism and Freedom*. Chicago University Press, Chicago

Hirschman A O 1970 *Exit, Voice and Loyalty: Responses to Decline in Firms, Organizations and States*. Harvard University Press, Cambridge, MA

Lane R E 1991 *The Market Experience*. Cambridge University Press, Cambridge, UK

Marwell G, Ames R 1981 'Economists free ride. Does anyone else? Experiments on the provision of public goods. IV. *Journal of Public Economics* **15**(3): 295–310

Marx K 1977 *Capital*. Vintage Books, New York, Vol. 1

Mill J S 1970 *Principles of Political Economy*. Penguin Books, London

Nozick R 1974 *Anarchy, State and Utopia*. Blackwell, Oxford, UK

Ostrom E 1990 *Governing the Commons: the Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK

Polanyi K 1971 *The Great Transformation*. Beacon Press, Boston

Radin M J 1996 *Contested Commodities*. Harvard University Press, Cambridge, MA

Samuelson P 1957 Wages and interests: A dissection of Marxian economics. *American Economic Review* **47**: 884–912

Satz D 1992 Markets in women's reproductive labor. *Philosophy and Public Affairs* **21** (Spring): 107–31

Scitovsky T 1976 *The Joyless Economy*. Oxford University Press, Oxford, UK

Sen A 1987 *On Ethics and Economics*. Blackwell, Oxford, UK

Smith A 1776/1937 *The Wealth of Nations*. Modern Library, New York

Stiglitz J, Weiss A 1981 Credit rationing in markets with imperfect information. *American Economic Review* **71** (June): 393–410

Sunstein C 1991 Politics and preferences. *Philosophy and Public Affairs* **20**: 3–34

Titmuss R 1971 *The Gift Relationship: From Human Blood to Social Policy*. Pantheon Books, New York

Walrus L 1926/1954 *Elements of Pure Ecnonomics* [trans. W. Jaffe]. Richard D Irwin, Homewood, IL

Walzer M 1983 *Spheres of Justice*. Basic Books, New York

D. Satz

Market Areas

1. Market Area as Territorial Expression

A market is the set of actual or potential customers of a given good or service supplier. These consumers may be characterized according to any of a number of sociodemographic dimensions, including their age, their gender, their lifestyle, as well as their location relative to the supplier's own geographic position, to the position of other suppliers, and to the position of other consumers. The territory (or section of the geographic space) over which the entirety or majority of sales of a given good or service supplier takes place constitutes this supplier's market area. The same concept is also known as 'trade area,' 'service area,' or 'catchment area' across various fields of social and behavioral sciences.

In modern economic systems, in which production and consumption have uncoupled, consumption (demand) is rather ubiquitous in the geographic space, while production and distribution functions (supply) are limited to a few locales. The incongruity in the spatial organization of demand and supply is mediated by the mobility of economic agents: market places emerge when consumers are mobile, home delivery requires suppliers to reach out to their customers, while periodic markets take form in sociocultural environments where production and consumption agents are mobile. In all cases, the concept of market area serves to capture the notion of territorial expression of a supplier's clientele. It applies to retailers and service marketers as disparate as grocers (Applebaum 1966), health care providers (Fortney et al. 1999), and library branches (Jue et al. 1999).

Goods and service providers are known to seek common locations in the geographic space (see *Lo-*

cation Theory; *Retail Trade*; *Cities, Internal Organization of*) so as to share into externalities created locally by their agglomeration or clustering. Towns throughout the world, many cities, spontaneous urban retail clusters and business districts, even planned shopping centers, owe their existence to such agglomerative forces. Because it is commonplace for many or all businesses belonging to these geographic clusters to serve customers from the same communities, the narrow definition of market area given in the previous paragraph can be extended to clusters of goods and service providers.

Why firms become located where they do is the central question of location theory. An early school of thought pioneered by Von Thünen approached this question as a least-cost problem (see *Location Theory* for a fuller treatment of this topic). On the other hand, the market-area analysis school emphasizes demand-side considerations by framing the locational decision of the firm as a search for site(s) that command the largest number of customers. It is a straightforward consequence that, for a given spatial distribution of consumers, the size and shape of a firm's market area are indicative of the success of the firm at that site. The market-area analysis school provides the economic underpinnings for a substantial theoretical body on the processes of spatial organization of firms in consumer-oriented industries (Greenhut 1956), but also of market places and settlement locales into hierarchical systems (Christaller 1933, Berry and Parr 1988). The market area approach continues to provide the tenets for applied research in retail location analysis (Ghosh and McLafferty 1987), public facility planning (Thisse and Zoller 1983), and spatial planning (Von Böventer 1965).

2. *Economic Underpinnings*

It was remarked above that a firm's market area is the territorial expression of the firm's clientele. The market area is far more than an ex-post map depiction of a purely aspatial microeconomic process that defines customer-supplier relationships. In his 1880s work, the German economist Launhardt (1882) is credited with unveiling the novel idea that a firm's success in selling its output in a market economy is shaped by considerations of transportation cost and of position of competing firms, one relative to another in the geographic space. In fact, it has only recently been established that the very same ideas were articulated four decades earlier by the German economist Rau (1968) in a little known letter to his colleague Quintelet.

Let us consider initially the simplified case of a single firm supplying a single good to a population of identical consumers scattered over a certain geographic territory. If each consumer is assumed to have an identical demand curve that is elastic to the good's 'real price' (sale price plus transportation cost), individual demand drops with distance from the firms location. The monopolistic firm's market area extends outward to the circular perimeter where demand falls to zero. This distance is known as the range of the good. The market beyond this point remains unserved by the firm; a situation that may raise serious social equity concerns if basic needs of the population remain unmet as a result. For the sake of increasing its sales and reducing its unit production cost, the firm may decide to serve consumers outside the normal market area. It may accomplish this according to several modalities. One approach involves absorbing part of the real price applied to more distant consumers. Any spatial pricing system aimed at accomplishing this goal subsidizes remote consumers and discriminates against nearby buyers. Similar economic circumstances are commonplace when several firms at different locations compete for consumers. Since firms have more control over their nearby customers, they may seek to discriminate against them, whereas consumers at the margin of market areas may enjoy lower prices ensuing inter-firm price competition (see *Location Theory* for a more detailed treatment).

When a second firm whose geographic position is different from the first one also supplies the same good, the situation becomes more complex. Each firm commands sales in the same circular market area if they are located far enough away from each other (i.e., over twice the range of the good). Firms enjoy monopolistic profits that may entice new firms to start business. New competitors can be expected to locate in the interstitial space between monopolistic market areas so as to command the largest possible sales. This leads us to the alternative condition where firms find themselves close enough to each other to create competition at the edge of their monopolistic market areas. The introduction of a competitor establishes another type of boundary line where real prices are equal. The economic significance of this boundary is considerable because it delineates the market areas of competing firms, and consequently determines their sales and profits. Remarkably, as indicated by Isard (1956), this boundary is optimal and socially efficient. Working with Launhardt's (1882) assumption that firms do not charge the same sale price, the boundary line is a hyperbolic curve defined by the equation of the difference of sale prices to the difference of transportation costs to firm locations. The American economist Fetter, who independently discovered the same economic principle, coined the phrase 'Economic Law of Market Areas' (1924).

Figure 1 illustrates the law of market areas under two different conditions. When firm B charges a sale price higher than that of firm A, firm A has a cost advantage over B that translates into a larger market area. The market boundary is hyperbolic and closer to B than to A. In the special case where firms charge identical sale prices, neither competitor has a special advantage. The two market areas are equal and are

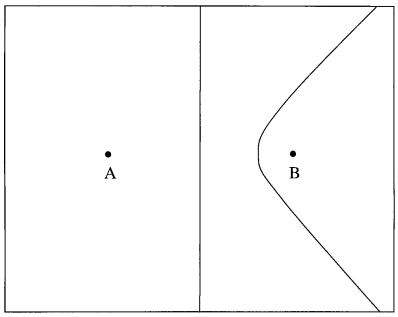

Figure 1
This diagram represents the market areas of two firms, A and B. When A and B charge equal sale price, market areas are separated by the vertical straight line. The hyperbolic curve separates market areas when B charges a price higher than A

separated by a boundary line that is a degenerated hyperbola equidistant from the two firm locations.

As indicated above, free entry of firms affects the size and shape of market areas. Christaller (1933) and Lösch (1954) argued that free entry would compress the circular market areas defined by the range into hexagons forming a lattice covering the entire territory. Their theory of spatial organization was also extended to encompass multiple goods—from the most to the least ubiquitous—to which corresponded multiple overlapping layers of hexagonal market areas. One area of deficiency of this theory was to underrate mutual interdependencies between markets for different goods and their territorial expression, the market areas. Demand-side linkages between markets are particularly meaningful to market-driven industries. One such linkage is created by consumers' effort to reduce the transactional cost of shopping by means of combining the purchase of several goods on the same shopping trip. Multipurpose shopping, as this practice is called, reduces the geographic dispersion of firms where these goods are purchased. Because of enlarged market areas, sites where a larger range of goods is marketed are consequently placed at an advantage over others.

3. *Spatial Analysis*

The previous section introduced the possibility of polygonal, hyperbolic, and circular market areas. The question arises now what shapes are possible and what conditions bring them about. Since market areas constitute a model of the geography of demand, knowledge of geometric and topologic properties of market areas (Hanjoul et al. 1988) may shed some light on various issues of interest to urban and regional scholars such as the life cycle of business districts, people's modality of usage of the urban space, or the quality of geographic accessibility to amenities (e.g., social services, health, and cultural facilities).

It turns out that market areas may take any of a large number of different shapes depending on how the perceived cost of traveling over space is related to the physical distance. For instance, the hyperbolic areas mentioned earlier are related to a linear transportation cost function. When transportation cost increases as the square root of distance (i.e., the cost of the extra mile traveled drops as the trip gets longer), market areas are elliptic. Circles arise when transportation costs follow the log-transform of distance. In contrast to hyperbolae, circular, and elliptic market areas reveal that the firm charging the higher sale price manages to serve consumers behind its own location only if they are located rather close. In these conditions, distant consumers in fact have an economic advantage in bypassing the nearby firm to patronize the more distant and cheaper firm. This economic behavior stands in direct contradiction with common intuition on the matter, yet it can be proved to result from the decreasing marginal cost of transportation. The same fails to hold for a transportation cost that is linear in distance or increases more than proportionately with distance. It follows that firms cannot automatically treat consumers that are distant from competitors as a 'safe' market. How secure a geographic segment of the market is does not only rest on where the segment lies with respect to firms, but also on the structure of transportation costs.

Another interesting property of geographic markets is the inclusion of a firm location in its own market area. Intuition suggests that a firm, whatever the economic and geographic environment it is operating in, ought to be able to competitively serve consumers at its front steps. Mathematical arguments can be used to show that this expectation is sound in most prevailing conditions, namely when transportation costs conform to a concave function of physical distance. In the unlikely event of transportation cost rising at an increasing rate, the firm charging a higher sale price would fail to serve nearby consumers if it is close enough to its main competitor. Quite understandably, this situation translates into greater vulnerability to competitive pressures of the market place.

4. *Behavioral Viewpoint*

The perspective outlined in Sect. 2 of this article may be crucial to foster our conceptual understanding of market areas, but too often it fails the test of operational dependability for the simple reason that it assumes that a locale falls entirely into one or another market areas. Business practitioners and applied social scientists have argued that a firm's ability to command consumers does not stop at the boundary of its market area, but rather that it extends past that boundary and into the areas imputed to other firms. The argument

continues as follows. Because of a host of other factors explaining people's choice of a firm and because of interpersonal variations in preferences, demand is best viewed as the outcome of a probabilistic process. The dominant paradigm is that of Spatial Interaction, according to which one's attraction to a firm (i.e., likelihood to buy from this firm) is directly related to the attractiveness of this firm and inversely related to the travel cost to it. See *Spatial Interaction*, *Spatial Interaction Models*, and *Spatial Choice Models* for a more detailed treatment.

In this view, a firm's market area is recast as the territory where attraction to this firm is greater that to any other firm. Since the 1930s, the approach has been known as Reilly's Law of Retail Gravitation (Reilly 1929, Huff 1962) by analogy with Isaac Newton's Law of Planetary Gravitation. Of course the operational concept of market area is richer than its strict economic counterpart discussed earlier. All the geometric and topological properties of market areas (Sect. 3) are still pertinent under the spatial interaction paradigm. The intensity of demand in and out of the market area adds an important dimension to the characterization of market areas. The principle of spatial interaction generates a pattern of decreasing demand intensity away from the firm location, the rate of which depends on the structure of transportation costs and on interfirm competition.

The spatial interaction approach to spatial market analysis recognizes that firm attractiveness may be a multifaceted notion shaped by price, service, and merchandising properties of the firm. It also recognizes that perceived attractiveness is what influences firm choices and that sociodemographics is a powerful mold of individual perceptions. The capability to segment consumers by their sociodemographic characteristics becomes handy when the study area is highly segregated socioeconomically. It is now commonplace to carry market analysis within the interactive computer environment offered by geographic information systems (see *Geographic Information Systems: Critical Approaches* for a more detailed treatment of this technology). Automated numerical and database manipulations provide an essential toolbox for spatially-enabled market analysis, also know as 'business geographics.'

Consumerism and compelling marketing practices have replaced mass merchandising with ever more finely defined customer segments with their own wants and needs. Where sale price and travel cost used to be the driving forces, markets have become more fluid, uncertain, and varied than ever before. It should not be too much of a surprise therefore that the marriage of spatial interaction and conventional, microeconomic market area analysis has enhanced our grasp of contemporary spatial market systems. The greater propensity of firms to form spatial clusters can be ascribed to the disappearance of exclusive market areas stipulated by the spatial interaction paradigm (de Palma et al. 1985). Concomitantly, a countervailing transformation may manifest itself in the form of greater differentiation of firms from each other (Thill 1992) to better cater to the diversity of consumer segments.

In these times where information technologies are quickly redefining the relationships between demand and supply in post-modern economies, the question naturally come whether electronic commerce (e-commerce) will bring the demise of the notion of market area. While no definite answer can be advanced at this time, partial arguments can be articulated as follows. As a rapidly increasing portion of postmodern economies is siphoned through e-commerce, more commerce takes on a new organizational structure wherein spatial relationships and territoriality may play a lesser role. New commercial relationships contribute to the emergence of a new geography, namely the virtual geography of information and communication flows on the network of computer servers. It has been argued that the notion of market area is not tied to the existence of a physical space and that it has bearing in the space framed by firm characteristics. In the same vein, the notion can be extended to the topological network of the information superhighway. Additionally, even the most optimistic projections on the expansion of e-commerce do not predict that it will entirely replace brick and mortar commerce. Hence, in all likelihood, face-to-face commerce and associated geographic markets will subsist. This is not to say that spatially-defined market will remain unaltered by the digital revolution. If evolution of the past decade gives any indication of future trends, the notion of market area will continue evolving in response to technology-enabled practices of target marketing towards increasingly entwined fields of interaction.

See also: Commodity Chains; Globalization: Political Aspects; International Business; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; International Trade: Geographic Aspects; Market Structure and Performance; Marketing Strategies; Multinational Corporations; World Trade Organization

Bibliography

Applebaum W 1966 Methods for determining store trade areas, market penetration, and potential sales. *Journal of Marketing Research* **3**: 127–41

Berry B L J, Parr J B 1988 *Market Centers and Retail Location. Theory and Applications*. Prentice-Hall, Englewood Cliffs, NJ

Christaller W 1933 *Die Zentralen Orte in Suddeutschland*. Gustav Fischer Verlag, Jena, Germany

de Palma A, Ginsburgh V, Papageorgiou Y Y, Thisse J-F 1985 The principle of minimum differentiation holds under sufficient heterogeneity. *Econometrica* **53**: 767–81

Fetter F A 1924 The economic law of market areas. *Quarterly Journal of Economics* **39**: 520–29

Fortney J C, Lancaster A E, Owen R R, Zhang M 1999 Geographic market areas for psychiatric and medical outpatient treatment. *Journal of Behavioral Health Services and Research* **25**: 108–16
Ghosh A, McLafferty S L 1987 *Location Strategies for Retail and Service Firms*. Lexington Books, Lexington, MA
Greenhut M L 1956 *Plant Location in Theory and Practice: The Economics of Space*. University of North Carolina Press, Chapel Hill, NC
Hanjoul P, Beguin H, Thill J-C 1988 *Theoretical Market Areas under Euclidean Distance*. Institute of Mathematical Geography, Ann Arbor, MI
Huff D L 1962 *Determination of Intra-urban Retail Trade Areas*. School of Business Administration, University of California, Los Angeles
Isard W 1956 *Location and Space-Economy*. MIT Press, Cambridge, MA
Jue D K, Koontz C M, Magpantay J A, Lance K C, Seidl A M 1999 Using public libraries to provide access for individuals in poverty: A nationwide analysis of library market areas using a geographic information system. *Library and Information Science Research* **21**: 299–325
Launhardt W 1882 Kommerzielle trassierung der verkehrswege. *Zeitschrift für Architektur und Ingenieurswesen* **18**: 515–34
Lösch A 1954 *The Economics of Location*. Yale University Press, New Haven, CT
Rau K H 1968 Report to the Bulletin of the Royal Society. In: Baumol W J, Goldfeld S M (eds.) *Precursors in Mathematical Economics: An Anthology*. London School of Economics and Political Science, London, pp. 181–2
Reilly W J 1929 *Methods for the Study of Retail Relationships*. University of Texas Bulletin 2944, Austin, TX
Thill J-C 1992 Competitive strategies for multi-establishment firms. *Economic Geography* **68**(3): 290–309
Thisse J-F, Zoller H G 1983 *Locational Analysis of Public Facilities*. North-Holland, Amsterdam, The Netherlands
Von Böventer E 1965 Spatial organization theory as a basis for regional planning. *Journal of the American Institute of Planners* **30**: 90–100

J.-C. Thill

Market Research

1. Introduction

Market research consists of both academic treatises and practical approaches to the collection, analysis, interpretation, and use of data. When undertaken by academic scholars, the research intends to understand and explain the behavior of market participants. This is accomplished by describing and interpreting real-world behavior, by measuring participants' attitudes and preferences, by experimentally manipulating variables in laboratory and field research, and by developing and testing models of marketplace behavior, laboratory decisions, and survey responses. When undertaken by practitioners, the research aims to predict the behavior of market participants so that managerial decisions can be made more effectively and efficiently based on the criterion that the expected value of information exceeds its cost.

To some extent the academic (theoretical) and practical research activities are related. For example, commercial researchers eagerly adopt new methods that show promise for improved practical applications. These new methods are often developed and tested by academic researchers. Also, the model-building branch of market research tends to insist on pragmatism (Leeflang and Wittink 2000). In that branch, standard questions are: does a new or revised model provide superior predictive validity and will managerial decisions be improved?

2. Theoretical Underpinnings

There is great breadth and richness in market research stemming from its cross-disciplinary nature. The concepts that support inquiries are often based on either economic or psychological theories, and the aim of an academic research study is usually similar to the aim of research in the related discipline.

Economic theory tends to describe and predict normative behavior: the behavior that consumers, managers, or firms should undertake if they want to act optimally. Economic research may show how individuals should behave or it may show support for a theory through econometric modeling of actual consumer or firm behavior. The latter approach is popular in marketing journals as both managers and academic researchers demand empirical support for theoretical predictions. This is not to say that strictly theoretical work does not have a place in the literature. Usually this work is considered a first step in understanding a phenomenon, and other researchers build upon and test the validity and usefulness of new theories. Basu et al. (1985) studied compensation planning. They explain why a firm needs to consider a salesperson's tolerance for risk in designing a motivational compensation plan. Subsequent research shows that real-world compensation schemes broadly support the predictions of the theory.

Apart from purely theoretical research, it is common for academic scholars to propose or adapt a theory and subject it to empirical tests in the same study. Empirical studies in marketing use data from a variety of sources such as laboratory experiments, field experiments, surveys, and marketplace purchases. Household scanner-panel data studies have flourished in the past 20 years due to advances in computer technology that have made purchase data much easier to gather with loyalty cards. The scanner panel derives its name from the supermarket checkout scanner that feeds prices to the cash register. Collecting these data is nearly effortless as households are motivated to use a shopping card or another device each time they make purchases. The card identifies the consumer, and a

computer collects a history of his or her purchases as they are made.

Guadagni and Little (1983) provide the first application of a model of brand choice on household scanner data. Subsequently, other data modelers have tested more elaborate and substantive theories with advanced estimation methods. Gupta (1988), for instance, develops a model to predict not only which brand a consumer will choose, but also when the purchase will be made and how much will be bought.

In contrast to work derived from microeconomics, the research that is based on psychology or behavioral decision theory (BDT) tends to focus on behavior that deviates from strict utility maximization principles. This field of research is partly based on Kahneman and Tversky's (1979) Prospect Theory. Their theory predicts how people make decisions when they are faced with uncertain prospects or outcomes. It differs from traditional economic theory in several respects. For example, people are assumed to judge their well-being against a point of reference. This means that the framing of a tradeoff, how a choice is presented to a consumer, can influence which alternative is chosen. Their theory also predicts that losses have a greater impact on preferences than gains, a concept known as loss aversion. This notion explains why people may engage in risky behavior when faced with choices that involve certain losses, but engage in less risky behavior in choices that involve certain gains.

Marketing studies based on BDT include Simonson and Tversky (1992). They show that a consumer may prefer product A over product B when these are the only products under consideration, but he/she will prefer product B over A when a third product C is included in the choice set. This reversal refutes the regularity condition which states that the addition of a new alternative to a set cannot increase the attractiveness of an original item. In related research, Dhar (1997) shows that choice deferral depends on the characteristics of the alternatives. Again in violation of the regularity condition, he shows that the incidence of deferral increases with the addition of a new alternative when it differs from an original one but is approximately equally attractive. This suggests that if consumers are uncertain about their preferences for attractive options they may prefer to delay the purchase.

Typically, BDT researchers conduct carefully designed laboratory experiments in which they demonstrate that consumer decisions are consistent with a proposed theory. In subsequent experiments they often show that alternative explanations can be ruled out. In this manner, the internal validity of experimental results is very high. Doubt may exist about the external validity in some cases, but real-world results are emerging with increasing frequency. Madrian and Shea (2000) show that at one firm where employees were invited to participate in a retirement plan ('opt in'), it took 10 years for the participation percentage to go from 40 percent at inception to 80 percent. However, when employees were automatically enrolled into a default investment program from which they could 'opt out', the percent participating was a stable 86 percent from the beginning. Uncertainty among employees about the best way to invest may explain the deferral to participate under 'opt in.'

3. *Data Collection and Methods*

Market researchers, like others in the social sciences, use a variety of methods to examine and test consumer behavior. Studies can be broadly classified into two groups, observational and experimental, depending on how the data are gathered. *Observational studies* are based on data tracking the real-world behavior of consumers. For example, a researcher may relate the purchases of consumers to the advertising, promotional, and pricing conditions under which their choices were made. If the study focuses on real-world behavior, the behavior itself is without question. However, explanations of the behavior are subject to potentially severe limitations.

The primary weakness of an observational study is that consumers choose the conditions to which they are exposed (self-selection), making it impossible for the researcher to rule out alternative explanations. Suppose we want to know whether a new advertising campaign increases consumer purchase amounts of a brand. A positive association between the purchases and advertising is consistent with the research hypothesis, but it is also possible that the same consumers who are likely to buy the brand are likely to see the advertising. In addition, other relevant variables may have changed at the same time so that the association is not unique.

Apart from studies of purchase behavior, observational studies include survey responses. Surveys are often used by researchers to obtain an understanding of marketplace behavior. Respondents may describe attitudes towards objects and may provide explanations of past behavior to supplement the information contained in marketplace choices. The field of market research was originally oriented especially toward surveys. A substantial amount of research concerned questionnaire formats, scaling methods, sampling schemes, and response rates. Data analysis often focused on differences in reported purchases or attitudes toward products as a function of demographics. Early research showed, however, that demographics and socioeconomic characteristics provide very modest explanatory power. Survey research now includes customer satisfaction and preference measurement. Respondent heterogeneity in these measures is understood to be a critical property of data that is only partly explained by observable characteristics.

The measurement and analysis of customer satisfaction data occurs both for 'keeping score' and for 'explaining the score.' Explanations of overall satisfaction focus on an identification of the drivers of satisfaction. However, there is not a one-to-one relation between repeat purchase and satisfaction. For the managerial use of satisfaction data to be productive, it is critical that satisfaction is linked with (subsequent) purchase behavior.

While satisfaction measurement occurs after purchases are made, preference measurement precedes purchases. A popular approach to quantify the tradeoffs in (future) purchase behavior is conjoint analysis which has excellent forecast accuracy (Wittink and Bergestuen 2001). Essentially this method asks individuals to evaluate the attractiveness of hypothetical products or services so that implicit tradeoffs between conflicting considerations can be quantified. By including product features not yet available to consumers, it is possible to make predictions of new-product success. In addition, by forcing respondents to consider the benefit of new features or improved quality against each other and against (higher) price levels, researchers learn about consumers' priorities, willingness to pay, and price–quality tradeoffs.

In an *experimental or randomized study*, the researcher assigns individuals to different conditions or treatments at random in order to eliminate the influence of outside factors. Such studies have high internal validity because all factors are held constant except for the one(s) being manipulated. This control by the experimental researcher can be guaranteed in laboratory experiments. However, the behavior individuals exhibit or claim in a laboratory may not generalize to the real world. If it does generalize, it may not be a primary determinant of market behavior. For example, Deighton et al. (1994) find that certain advertising effects, observed earlier in laboratory experiments, do not play a large role in determining the observed brand choices of consumers in selected product categories. We note that this lack of substantiation may also result from product category differences between the studies (i.e., the effect of interest may occur in some but not in all categories).

Commercial market research rarely involves purely experimental methods. Not only do managers insist that the results explain real-world decisions, they are often unwilling to take the risks that are part of experimental research in the real world. Consider the following example. Ackoff, an expert in operations research, was asked by top management at Anheuser Busch to determine the 'correct' amount of spending on advertising. The parties agreed that historical data were insufficient to estimate the relation between sales and advertising. Ackoff suggested that Busch's sales territories in the US be randomly allocated to either receive 50 percent less advertising, the current amount, 50 percent more, or 100 percent more. He argued that such a field experiment would be the only way to determine how sales (and profit) depend on advertising expenditures. Management was concerned, however, that while it did not know the optimal amount of spending it did not want to risk losing business in important territories.

A solution to this dilemma was obtained when management agreed to take risk in some of the smallest territories. When the results, from an experimental design with reduced scope, were very surprising (sales increased in areas with higher advertising but the increase was even higher with lower advertising), management decided that a larger scale experiment was warranted. Unfortunately, the subsequent large-scale results were (still) impossible to interpret (Ackoff and Emshoff 1975). Unlike laboratory experiments, in which all nonmanipulated variables can be controlled, field experiments are subject to interference. Competing manufacturers and distributors were aware of changes in Busch's advertising, and they could modify their own activities. More critically, Anheuser Busch's own sales managers and distributors in territories with reduced advertising support would demand compensation in price or other variables. Also, managers of territories with (randomly assigned) higher advertising levels may have interpreted the increase as an indication of confidence in their actions which would spur them on to greater efforts. As a result it is impossible to attribute changes in sales to changes in advertising, even if advertising is the only variable experimentally manipulated in a field experiment.

Difficulties with the internal validity of field experiments are less likely to occur with new products. Eskin and Barron (1977) report on field experiments in which price and advertising of various new products are manipulated. Across multiple product categories they find that the effect of a given decrease in price on the demand (unit sales) of a new product is stronger for higher amounts of (television) advertising expenditures. This result appears to contradict economic theory which suggests that individual consumers will become less price sensitive with increases in nonprice advertising. An explanation requires that we note the aggregate nature of the study. Higher amounts of advertising attract additional consumers who may differ in price sensitivity from other consumers. For example, the incremental consumers need more advertising before they will consider purchasing the product, and for them a purchase tends to require a lower price. As a consequence, the aggregate demand (sales) curve shows higher average price sensitivity at higher levels of advertising.

The importance of such experimental results for management decisions is illustrated as follows. Management of one firm had expected to choose either 'high price and high advertising' or 'low price and low advertising' for a new product. The argument was that a high price implies a high profit margin which creates resources for advertising, while a low price would not.

The experiment shifts the focus to exploring how demand depends on marketing activities. With the estimated demand function, management can choose the optimal combination of advertising and price (which differs dramatically from either of the margin-based strategies).

Some research combines theories from multiple disciplines to create a broader description of the world or to reconcile conflicting theories. Traditional perspectives on manufacturer advertising as either creating market power or providing useful information implied conflicting results (Kaul and Wittink 1995). To resolve this, Mitra and Lynch (1995) discuss how advertising may both influence the differential strengths of consumer preferences for individual products and increase the consideration set of alternatives. They conduct a laboratory experiment, and use economic measures to show that at the individual consumer level, advertising can decrease price sensitivity (of brand choice) if it strengthens brand preference but increase price sensitivity if it expands the consideration set.

4. *The Future*

Similar to the way it is changing methods of discovery in other fields, computer technology is revolutionizing how market research is conducted. The Internet provides superior means of data collection about consumers' browsing and purchasing behavior, about their preferences for new goods, and about their satisfaction with purchase experiences. Furthermore, it allows these data to be easily shared among analysts and researchers, sometimes making multiple sources of data available for study when before there might have been none. The benefit of this innovation is that firms can more easily learn about the needs of consumers. In the twenty-first century, products and services will better suit consumers' needs.

As with all progress, however, the Internet has created new ethical issues. The issue of privacy, for instance, requires a lot of attention. Firms are collecting data so unobtrusively that people often do not realize that their behavior is being monitored. This is not an entirely new issue, as direct marketing firms have shared or purchased lists of addresses and phone numbers for years, but the scope of the issue is much wider. Inaccurate data can spread quickly and cannot easily be corrected across multiple sources. Consumers are ill-equipped to conceal sensitive data such as their medical histories or to correct inaccurate credit histories that affect loan or credit card applications.

Market researchers need to inform consumers about the benefits they will realize. Only if they provide relevant information will consumers be able to purchase products closer to their tastes. They can decide how much information they are willing to make available based on their self-interests. Safeguards such as firms not selling information without the written consent of consumers and agreeing to destroy data once a certain time period has expired can help protect this exchange.

A more subtle issue that market researchers face is that the results of previous consumer behavior research alter how firms and consumers act in the marketplace. Marketing activities will be customized based on research results for each individual consumer. Future learning then requires that researchers adopt experimental methods. For example, activities can be increased or decreased based on random assignments, from what may have been optimal levels, so as to determine the required adjustments over time. This is similar to the field experiment for Anheuser Busch, except that the randomization occurs over individual consumers.

A variation of this approach is the way direct marketing experiments are conducted. Direct marketing firms often target a small random sample of consumers with multiple offers, see which offers are most effective with which people, and then make targeted offers to a larger population based on the results of the study (Steenburgh et al. 1999). Other practices that need experimental manipulations for continued learning include those used by retailers and e-tailers. Consumers receive coupons in the supermarket based on their current purchases. E-tailers suggest books and music based on customers' previous purchases and other information.

On the Web, it is easy for consumers to compare prices for a given item and to learn about the availability of goods and services that were previously out of reach or invisible. The complexities, however, are vast and for consumers to be effective decision makers it will be important that they have access to decision support systems. Such systems can incorporate a consumer's utility functions for various items so that with access to valid and reliable information on alternatives (infobots), the consumer can maximize his/her utility much more effectively and efficiently than ever before. As a result, markets should become far more efficient.

Managers will offer to serve consumers much more comprehensively than before. Suppliers will provide bundles of products and services that offer convenience. Marketing research will focus on how products and services can be customized and personal relations established so that suppliers effectively solve customers' problems. As a result, marketing research will shift from a product-focused activity to a customer-focused one. The ultimate criterion becomes the profitability of customers.

See also: Advertising Agencies; Advertising: Effects; Advertising: General; Consumer Culture; Consumer Economics; Consumer Psychology; Consumption, Sociology of; Mass Media, Political Economy of; Media Effects; Public Relations in Media

Bibliography

Ackoff R L, Emshoff J R 1975 Advertising research at Anheuser Busch (1963–68). *Sloan Management Review* **16**(Winter): 1–15

Basu A K, Lal R, Srinivasan V, Staelin R 1985 Salesforce compensation plans: an agency theoretic approach. *Marketing Science* **4**(Fall): 267–91

Deighton J, Henderson C M, Neslin S A 1994 The effects of advertising on brand switching and repeat purchasing. *Journal of Marketing Research* **31**(February): 28–43

Dhar R 1997 Consumer preference for a no-choice option. *Journal of Consumer Research* **24**(September): 215–31

Eskin G J, Barron P H 1977 Effects of price and advertising in test-market experiments. *Journal of Marketing Research* **14**(November): 499–508

Guadagni P M, Little J D C 1983 A logit model of brand choice calibrated on scanner data. *Marketing Science* **2**(Summer): 203–38

Gupta S 1988 Impact of sales promotions on when, what and how much to buy. *Journal of Marketing Research* **25**(November): 342–55

Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* **47**(March): 263–91

Kaul A, Wittink D R 1995 Empirical generalizations about the impact of advertising on price and price sensitivity. *Marketing Science* **14**(Summer): G151–60

Leeflang P S H, Wittink D R 2000 Building models for marketing decisions: past, present and future. *International Journal of Research in Marketing* **17**(2–3): 105–26

Madrian B C, Shea D F 2000 *The power of suggestion: Inertia in 401(K) participation and savings behavior*. Working paper 7682, National Bureau of Economic Research, Cambridge, MA

Mitra A, Lynch J G 1995 Toward a reconciliation of market power and information theories of advertising effects on price elasticity. *Journal of Consumer Research* **21**(March): 644–59

Simonson I, Tversky A 1992 Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research* **29**(August): 281–95

Steenburgh T J, Ainslie A, Engebretson H 1999 *Unleashing the information in zipcodes: Bayesian methodologies for massively categorical variables in direct marketing*. Working paper, Cornell University, Ithaca, NY

Wittink D R, Bergestuen T 2001 Conjoint analysis. In: Armstrong J S (ed.) *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, Dordrecht, The Netherlands, pp. 147–67

T. J. Steenburgh and D. R. Wittink

Market Structure and Performance

Why some industries come to be dominated worldwide by a handful of firms, even at the level of the global market, is a question that has attracted continuing interest among economists over the past 50 years—not least because it leads us to some of the most intriguing statistical regularities in the economics literature. Uncovering the forces that drive these regularities provides us with some deep insights into the workings of market mechanisms.

The literature in this area has developed rapidly since 1980. Indeed, a fairly sharp break occurred at the end of the 1970s, with a new generation of models based on game-theoretic methods. These new models offered an alternative approach to the analysis of cross-industry differences in structure and profitability. Before turning to these models, it is useful to begin by looking at the earlier literature.

1. *Preliminaries: Definitions and Measurement*

The structure of an industry is usually described by a simple 'k-firm concentration ratio,' that is the combined share of industry sales revenue enjoyed by the largest k firms in the industry. Official statistics usually report concentration ratios for several values of k, the case of $k = 4$ being the most commonly used. (From a theoretical point of view, the case $k = 1$ is most natural, but is rarely reported for reasons of confidentiality.) A richer description of structure can be provided by reporting both these ratios and the total number of firms in the industry. If ratios are available for many values of k, we can build up a picture of the size distribution of firms, which is usually depicted in the form of a Lorenz curve. Here, firms are ranked in decreasing order of size and the curve shows for each fraction k/n of the n firms in the industry, the combined market share of the largest k firms. (It is more natural in this field to cumulate from the largest unit downwards, rather than from the smallest upwards, as is conventional elsewhere.) Certain summary measures of the size distribution are sometimes used, that of Herfindahl and Hirshman being the most popular (see Hirshman 1964): this is defined as the sum of the squares of firms' market shares, and its value ranges from 0 to 1. While most measures of market structure are based upon firms' sales revenue, other measures of firm size are occasionally used, the most common choice being the level of employment.

2. *The Cross-section Tradition*

The beginnings of the cross-section tradition in the field of industrial organization (IO) are associated with the pioneering work of Joe S. Bain in the 1950s and 1960s (see in particular Bain 1956 and 1966). Bain's work rested on two ideas.

(a) If the structure of the industry is characterized by a high level of concentration, then firms' behaviour will be conducive to a more muted degree of competition, leading to high prices and high profits (or 'performance'). This structure–conduct–performance paradigm posited a direction of causation that ran from concentration to profitability. It therefore raised the question: will the high profits not attract new entry, thereby eroding the high degree of concentration? This leads to Bain's second idea.

(b) Bain attributed the appearance of high levels of concentration to certain 'barriers to entry,' the first of

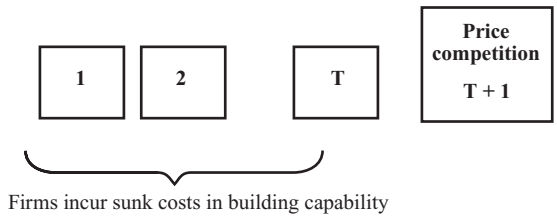

Figure 1
The multistage game framework

which is the presence of scale economies in production (i.e., a falling average-cost curve). In his pioneering book *Barriers to New Competition* (1956), Bain reported measures of the degree of scale economies across a range of US manufacturing industries, and went on to demonstrate a clear correlation between scale economies and concentration. This correlation notwithstanding, it was clear that certain industries that did not exhibit substantial scale economies were nonetheless highly concentrated, and this led Bain and his successors to posit additional barriers to entry which included, among others, the industry's advertising/sales ratio and its R and D/sales ratio. This raises a serious analytical issue however: while the degree of scale economies can be thought of as exogenously given as far as firms are concerned, the levels of advertising and R and D expenditure are the outcomes of choices made by the firms themselves, and so it is natural to think of these levels as being determined jointly with concentration and profitability as part of the same equilibrium outcome (Phillips 1971, Dasgupta and Stiglitz 1980). This remark provides the key link from the Bain tradition to the modern game-theoretic literature.

2.1 The Modern (Game-theoretic) Literature

A new literature which has developed since 1980 takes a rather different approach to that of the Bain paradigm. First, instead of treating the levels of advertising and R and D as 'barriers to entry' facing new firms, these emerge as the outcome of firms individual choices. The models are characterized by 'free entry,' but firms that fail to invest as much as rivals on such 'sunk costs' suffer a penalty in terms of their future profit flows. The second difference from the Bain approach is that the troublesome question of a possible 'feedback' from high profits to subsequent entry is finessed. This is accomplished by formulating the analysis in terms of a simple (game-theoretic) model in which all firms entering the industry anticipate the consequences of their investments on future profit streams.

The multistage game framework is shown in Fig. 1. Over stages 1 to T, firms make various investment decisions (construct a plant, build a brand, spend on R and D). By the end of this process, each firm's 'capability' is determined by the investments it has made. In the final stage subgame (T + 1), firms compete in price, their capabilities being taken as given

The central problem that underlies this kind of analysis is that there are many reasonable ways of formulating the entry and investment process that takes place in stages 1 to T, just as there are many ways of characterizing the nature of price competition that occurs at stage T + 1. Many equally reasonable models could be written down, between which we could not hope to discriminate by reference to the kind of empirical evidence that is normally available. (Such problems are commonplace in the game-theoretic IO literature; see Fisher (1989) and Pelzman (1991) for a critique.)

For this reason, it is useful to begin, not from a single fully specified model, but rather from a 'class of models,' defined by a few simple properties (Sutton 1998). The key innovation here lies in moving away from the idea of identifying a fully specified model within which each firm is assigned a 'set of strategies' (and where our aim is to identify the combination(s) of strategies that form a (perfect Nash) equilibrium). Instead, an equilibrium concept is defined directly on the set of outcomes that emerge from the investment process that occurs over periods 1 to T. An outcome is described by a list of the firms' capabilities, from which we can deduce the pattern of market shares that will emerge in the final stage (price competition) subgame. What we aim to do it to place some restrictions on these outcomes, and so on the form of market structure. The restrictions of interest emerge from certain simple and robust properties that must hold good, independently of the detailed design of the entry and investment process. The only assumption imposed on this process is that each firm ('potential entrant') is assigned some 'date of birth' t between 1 and T, and each firm is permitted to make any (irreversible) investments it chooses either at stage t or at any subsequent stage $t+1, t+2, \ldots, T$. Given the investments made by the firms (which can be represented formally by a set of points in some abstract 'set of products' that the firm is now capable of producing), we define a profit function that summarizes the outcome of the final stage (price competition) subgame. This specifies the ('gross') profit earned by the firm in that final stage, as a function of the set of all firms' capabilities. Finally, it is assumed that there are many potential entrants (in the sense that if all firms made the minimum investment required to enter the industry, then it could not be the case that all firms would earn sufficient profits to recover their outlays).

Within the (complete information) setting just described, the set of equilibrium outcomes (formally, the pure strategy perfect Nash equilibrium outcomes) will satisfy two elementary properties:

(a) 'Viability': The profit earned by each firm in the final stage subgame suffices to cover the investment costs it incurs.

Restriction (a), together with the assumption of 'many potential entrants,' ensures that not all firms will enter the industry. We now focus on any firm that has not entered:

(b) 'Stability': Given the configuration of investments undertaken by all the firms who have entered, then there is no investment available to a potential entrant, at time T, such that it will earn a final stage ('gross') profit that exceeds its cost of investment.

Outcomes that satisfy (a) and (b) are known as 'equilibrium configurations.'

From these properties, two sets of results follow. The first set of results pertains to the special case of industries in which neither advertizing nor R and D play a significant role. In analyzing these industries, the only restriction placed on the final stage ('price competition') subgame is that a rise in firm numbers or a fall in concentration, holding market size constant, reduces the level of prices and the gross profit earned by each firm. (On this assumption, see Sect. 5 below.) The first result relates to the way in which the level of concentration is affected by the size of the market: as we increase the size of the market (by successive replications of the population of consumers) then the minimum level of concentration that can be supported as an equilibrium configuration falls to zero ('convergence'). It is important that increases in market size do not *necessarily* imply convergence to a fragmented market structure: this result specifies only a *lower bound* to concentration. There will in general be many equilibrium configurations in which concentration lies above this bound.

The second result relates to the way in which, for a market of any given size, this lower bound to concentration is affected by the nature of price competition. Here, the key concept is that of the 'toughness of price competition,' which relates to the functional relationship between market structure (number of firms, or level of concentration) and the level of prices or price-cost margins, in the industry. For the sake of illustration, suppose changes in competition law make it more difficult for firms to operate a cartel, or suppose an improvement in transport networks brings into close competition a number of firms that hitherto enjoyed 'local monopolies.' In such cirumstances, we have an increase in the toughness of price competition in the present sense, that is for any given market structure, prices and price-cost margins are now lower. The second result states that an increase in the toughness of price competition leads to an upward shift in the lower bound to concentration (leaving unchanged its asymptotic value of zero; see Fig. 2). This result follows from the viability condition alone: the lower bound is defined by the requirement that the viability condition is *just* satisfied. A rise in the toughness of price competition lowers final stage profit, at a given level of concentration. Restoring viability requires an offsetting rise in concentration.

The available empirical evidence offers clear support for this prediction; see Sutton (1991) and Symeonides (2000).

This simple but basic result offers an interesting shift of perspective on the old idea that creating a more fragmented industry structure might offer a way of generating more intense competition, and so lower prices. What it suggests, rather, is that once we treat entry decisions as endogenous then—at least in the long run—the level of concentration and the equilibrium level of prices are jointly determined by the competition policy regime: introducing tough anti-cartel laws, for example, implies *inter alia* a higher equilibrium level of concentration.

2.2 *The Escalation Effect*

The second set of results relates to the difference between the industries just considered, where advertising and R and D spending are 'low,' and those where advertising and R and D play a significant role. Here, we need to specify how the outlays incurred on advertising and R and D in the earlier stages of the game affect the ('gross') profit earned by the firm in the final stage subgame. The key idea runs as follows: denote by F the fixed and sunk cost incurred by the firm, and by $S\pi$ the profit it earns in the final stage subgame. Here, S denotes market size and π is a function of the vector of products entered and so of the fixed costs incurred by all firms in earlier stages. (So long as firms' marginal cost schedules are flat—an assumption maintained throughout this article—and market size increases by way of a replication of the population of consumers so that the pattern of consumer tastes is unaffected, it follows that final stage profit increases in direct proportion to market size, justifying our writing it in this form.)

The main theorem is as follows (Shaked and Sutton (1987)).

Suppose that for some constants $a > 0$ and $K > 1$, a firm that spends K times as much as any rival on fixed outlays will earn a final stage payoff no less than aS; then there is a lower bound to concentration (as measured by the maximal market share of the largest firm), which is independent of the size of the market.

The idea is this: as market size increases, the incentives to escalate spending on fixed outlays rise. Increases in market size will be associated with a rise in fixed outlays by at least some firms, and this effect will be sufficiently strong to exclude an indefinite fall in the level of concentration.

The lower bound to concentration depends on the degree to which an escalation of fixed outlays results in profits at the final stage and so on the ratio a/K. We choose the pair (a, K) which maximizes this ratio and write the maximal value of the ratio as α. The theorem says that the number α constitutes a lower bound to the (one-firm) sales concentration ratio.

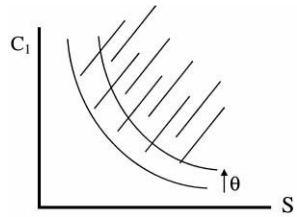

Figure 2
Increasing the toughness of price competition: the lower band shifts upwards but its asymptotic value is still zero

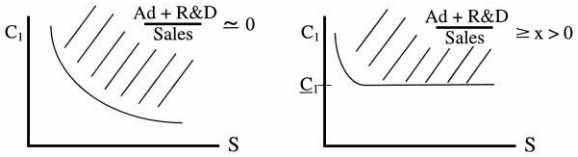

Figure 3
The nonconvergence theorem

In order to proceed to empirical testing, it is necessary to get around the fact that α is not easy to measure directly. So long as we have a well-defined market of the classical kind (a point elaborated in Sect. 4), it is easy to develop an ancillary theorem that allows us to use the level of advertising and/or R and D intensity as a sufficient statistic for α. Here, we obtain a simple 'nonconvergence' prediction, which says that no matter how large the size of the market S, the level of concentration cannot fall below some value C_1. Figure 3 illustrates the basic result for a well-defined market of the classical kind.

One way of testing this prediction is to look at the same set of advertising-intensive industries across a number of countries of varying size. Since firms must spend on advertising separately in each country to develop their brand image locally, this offers a valid 'experiment' relative to the theory. The nonconvergence property illustrated in Fig. 3 has been confirmed by Sutton (1991), Robinson and Chiang (1996), and Lyons et al. (2000).

2.3 *Extensions*

Underlying the above discussion is an important simplifying assumption relating to the definition of the market. It is tacitly assumed here, as in most of the theoretical literature, that we can think of the market as comprising a number of goods that are more or less close substitutes for each other, and that we can regard all other ('outside') goods as being poor substitutes for those in this group. This is the classic approach to the problem of market definition, associated with Robinson (1936): we identify the boundary of the market with a 'break in the chain of substitutes.'

In practice, this is too simple a picture. The linkages between goods may be quite complex. Apart from the demand side, where goods may be more or less close substitutes, there may also be linkages on the supply side: goods which are independent of each other on the demand side may share some technological characteristics, for example, which imply that a firm producing both goods may enjoy some economies of scope in its R and D activities.

In order to justify the analysis of a single market in isolation from other markets ('partial equilibrium'), it is necessary to define the market broadly enough to encompass all such linkages. Once we have widened the set of products to this point, however, we are likely to find that there are certain clusters of products that are much more tightly linked to each other, than to products outside the cluster. In other words, the market encompasses a number of 'submarkets.'

This complication is of particular importance in the context of R and D-intensive industries. It is often the case in these industries that the market comprises a set of submarkets containing products associated with different technologies. The early history of many industries has been marked by competition between rival 'technological trajectories'. Sometimes, a single 'dominant' trajectory emerges (Abernathy and Utterbach 1978). Sometimes, several trajectories meet the preferences of different consumer groups, and all survive. In general, the picture of 'escalation' along a single R and D trajectory, which corresponds to the analysis of Sect. 3 above, needs to be extended in this more complex setting. Depending upon the interplay of technology (the effectiveness of R and D along each trajectory) and tastes (the pattern of consumer preferences across products associated with each trajectory), the industry may evolve in one or other of two ways. The first route involves 'escalation' along a single dominant trajectory, in the manner of Sect. 3 above. The second route involves the 'proliferation' of successive technological trajectories, with their associated product groups ('submarkets'). In this setting, the predictions of the theory need to be recast in a way that is sensitive to the presence of such distinct submarkets; see Sutton (1998), Part I.

3. *Profitability*

A central issue of concern in the traditional literature related to the putative link between the level of concentration of an industry, and the average level of profitability enjoyed by the firms (businesses) in that industry. Here, it is necessary to distinguish two quite different questions. The first relates to the way in which a fall in concentration, due for example to the entry of additional firms to the market, affects the level of prices and so of price-cost margins. Here, matters are uncontroversial; that a fall in concentration will lead to a fall in prices and price-cost margins is well-supported both theoretically and empirically. (This

result was embodied as an assumption in the models set out in Sect. 2 above. While theoretical counter-examples can be constructed, they are of a rather contrived kind.) To test this idea it is appropriate to look at a number of markets for the same product, which differ in size (the number of consumers), so that larger markets support more sellers. It can then be checked whether prices and price-cost margins are, therefore, lower in those larger markets which support more sellers. The key body of evidence is that presented in the collection of papers edited by Weiss (1989). For a comprehensive list of relevant studies, see Schmalensee (1989) p. 987.

In terms of the basic models set out in Sect. 2 above, this first question relates to the description of the final stage (price competition) subgame: we are asking how the level of gross profits per firm earned in the final stage subgame relates to the level of concentration that has results from earlier investment decisions. A second, quite different, question relates to the net profit of firms (gross profit minus the investment costs incurred in earlier stages). In the 'free entry' models described in Sect. 2, entry will occur up to the point where the gross profits of the marginal entrant are just exhausted by its investment outlay. In the special setting where all firms are identical in their cost structure and in their product specifications, the net profit of each firm will be (approximately) zero, whatever the level of concentration. This symmetric setup provides a useful point of reference, while suggesting a number of channels through which some relationship might appear between concentration and net profits (or more conventionally, by the firm's rate of return on its earlier investment—gross profit flow per annum divided by the value of the firm's assets).

There are four channels that are worth noting.

(a) Even if all firms are symmetric (whence we can express net profit per firm π as a function of the number of entrants N) free entry implies zero profit only in the sense that $\pi(\mathrm{N}) \geqslant 0$ and $\pi(\mathrm{N}+1) \leqslant 0$. When N is large, this 'integer effect' is unimportant, but in those markets considered in Sect. 2.2, where the number of firms may be very small even if the market is large, this leaves open the possibility of 'large' net profits for the firms that enter.

(b) If we, realistically, abandon the assumption that all firms are alike in their cost structures, then we would expect the most efficient firms (those with the lowest level of average variable cost) to produce larger volumes of output at equilibrium. If we compare two hypothetical industries that are alike in all respects, except that the firms forming the pool of (potential) entrants in industry A are alike, but those in industry B differ in their efficiency levels, then we would expect industry B to exhibit both a higher level of concentration and a higher level of (average net) profitability (Demsetz 1973).

(c) Another form of asymmetry between firms may arise, not from any inherent differences in efficiency, but from 'strategic asymmetries.' In the models set out in Sects. 2 and 3, the only restriction imposed by the viability condition takes the form of a lower bound to concentration. This lower bound corresponds to a situation in which all earn firms are symmetric *ex post*, and all earn zero net profit. Above this bound there are other, asymmetric, equilibria, where intra-marginal firms enjoy positive net profit. Outcomes of this kind can arise for several reasons, such as the presence of 'first mover advantages.' For example, early entrants to an advertising-intensive industry may spend heavily in advance of rival entry, in building up brand images. Later entrants may then find it optimal to spend less on advertising than these early entrants. Under these circumstances, the first mover may enjoy strictly positive net profits.

(d) If we abandon the maintained assumption of free entry, in favour of the idea—fundamental to Bain's approach—that high concentration levels are associated with some form of 'barriers to entry' that prevent (efficient) firms from entering certain industries, then we would expect these concentrated industries to exhibit high levels of profit, in comparison to a reference group of industries where no such 'barriers' were present.

These points notwithstanding, it is important to note that there is *no* robust theoretical prediction of the kind developed in Sects. 2.1 and 3 above which links concentration to net profits.

4. *Empirical Evidence*

Following Bain's early contributions, a large body of empirical work was devoted to the search for correlations, across different manufacturing industries, between concentration and profitability. While the early literature seemed broadly supportive of such a relationship, the interpretation of the results remained highly controversial, the debate being focused primarily on distinguishing between the Bain interpretation ((d) above) and the Demsetz interpretation ((b) above). A turning point in this literature was marked by the appearance of two new databases that allowed the discussion to move beyond its focus on cross industry studies based on average industry profitability, towards an examination of market shares and profitability of each firm within the industry: the Federal Trade Commission's 'Line of Business' dataset, and the PIMS dataset. (See Scherer (1980) for an overview of the main results.)

The most authoritative review of the evidence is that of Schmalensee (1989). A key finding emerges from attempts to discriminate between the views of Bain and Demsetz by regressing the profitability of a business on both the market share of that business, and the level of concentration in the industry concerned. In regressions of this kind, using datasets that span many industries, profitability is strongly and

positively correlated with market share, while industry concentration tends *not* to be positively related to profitability. One important caveat is needed here, in that this finding might suggest that *within* each industry, firms with larger market shares enjoy higher rates of return: this is not the case. Intra-industries studies of market share versus rates of return suggest that no general relationship of this kind is present. (This, incidentally, is unsurprisingly from a theoretical standpoint: it is easy to find standard examples in which market shares within an industry are either positively or negatively correlated with profitability. In the simple 'linear demand model,' for instance, a firm that enters a larger number of product varieties than its rivals will enjoy a larger market share but a lower level of gross profit per product (Sutton 1998, Appendix 2.1).

What appears to be driving these results, as Schmalensee notes, is the presence of a small number of industries in which there is an unusually strong positive correlation between market share and profitability.

Apart from this kind of cross-sectional investigation, a quite different but complementary perspective is provided by the study of the pattern of firms' profitability over time. Mueller (1986) examined the question of whether, at the firm level, high rates of profitability were transient or persistent over long periods of time. Arguing that observed rates of return showed a substantial degree of persistence over time, Mueller drew attention *inter alia* to one group of industries (highly concentrated, advertising-intensive industries) as exhibiting unusually high and persistent rates of return. (Observation (c) above provides one candidate interpretation of this finding.)

5. *Further Topics: Beyond 'Strategic Interactions'*

The above discussion has focused heavily on the role of 'strategic interactions' between firms as a driver of market outcomes. It was noted in Sect. 2.3, however, that the general run of markets, as conventionally defined, tend to involve a number of (more or less loosely linked) 'submarkets.' The presence of such (approximately) independent submarkets carries some far-reaching implications for the analysis of market structure. In particular, it implies that 'strategic interactions' can not in themselves tell the complete story. Any satisfactory theory of market structure must bring together the role of 'strategic interactions' between firms or products within each submarket, and the presence of 'independence effects' across different submarkets. This becomes particularly important once we begin to look for a richer characterization of market structure by examining the size distribution of firms within each industry. A full discussion of this issue is beyond the scope of this article; for details, the reader is referred to Sutton 1997 and (1998, Part II).

See also: Industrial Geography; International Marketing; Market Areas; Market Research; Marketing Strategies

Bibliography

Abernathy W J, Utterbach J M 1978 Patterns of industrial innovation. *Technology Review* **80**: 40–7

Bain J 1956 *Barriers to New Competition*. Harvard University Press, Cambridge, MA

Bain J 1966 *International Differences in Industrial Structure: Eight Nations in the 1950s*. Yale University Press, New Haven, CT

Dasgupta P, Stiglitz J E 1980 Industrial structure and the nature of innovative activity. *Economic Journal* **90**: 266–93

Demsetz H 1973 Industry structure, market rivalry, and public policy. *Journal of Law and Economics* **16**: 1–9

Fisher F M 1989 Games economists play: A noncooperative view. *Rand Journal of Economics* **20**: 113–24

Hirshman A O 1964 The paternity of an index. *American Economic Review* **54**: 761

Lyons B, Matraves C, Moffat P 2000 Industrial concentration and market integration in the european union. *Economica* **68**: 1–26

Mueller D C 1986 *Profits in the Long Run*. Cambridge University Press, Cambridge, MA

Pelzman S 1991 The handbook of industrial organization: A review article. *Journal Political Economy* **99**: 201–17

Phillips A 1971 *Technology and Market Structure: A study of the Aircraft Industry*. Heath Lexington, Lexington, MA

Reinganum J 1989 The timing of innovation: Research, development and diffusion. In: Schmalensee R, Willig R D (eds.) *The Handbook of Industrial Organization, vol. 1*. Elsevier Science, Amsterdam, pp. 849–908

Robinson J 1936 *Economics of Imperfect Competition*. Macmillan, London

Robinson W T, Chiang J W 1996 Are Sutton's predictions robust? Empirical insights into advertising, R and D and concentration. *Journal of Industrial Economics* **44**: 389–408

Scherer F M 1980 *Industrial Market Structure and Economics Performance*, 2nd edn. Rand McNally College, Chicago, IL

Schmalensee R 1989 Inter-industry studies of structure and performance. In: Schmalensee R, Willig R D (eds.) *The Handbook of Industrial Organisation*, Vol. 2. Elsevier Science, Amsterdam, pp. 951–1009

Shaked A, Sutton J 1987 Product differentiation and industrial structure. *Journal of Industrial Economics* **36**: 131–46

Sutton J 1991 *Sunk Costs and Market Structure*. MIT Press, Cambridge, MA

Sutton J 1997 Gibrat's legacy. *Journal of Economic Literature*. **35**(1): 40–59

Sutton J 1998 *Technology and Market Structure: Theory and History*. MIT Press, Cambridge, MA

Symeonides G 2000 Price competition and market structure: The impact of cartel policy on concentration in the UK. *Journal of Industrial Economics* **48**: 1–26

Tirole J 1988 *The Theory of Industrial Organization*. MIT Press, Cambridge, MA

Weiss L W 1989 *Concentration and Price*. MIT Press, Cambridge, MA

J. Sutton

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Marketing Strategies

1. Introduction

Marketing strategies pursued by firms are aimed at obtaining a sustainable competitive advantage over competitors. In a marketing perspective, firms obtain a competitive advantage when they are able to study, identify, create, and communicate reasons of preference for their customers, with respect to their competitors.

In this article marketing strategies will be reviewed both in an external and internal perspective. Focus will be put on marketing strategies based on market driving capabilities and resources. A model of competitive strategies will then be presented in a circular perspective, considering the shifting nature of movement, imitation, and position wars. Some final points about marketing strategies in the third millennium will be presented, with the proposition of a new firm (proactive, lean, customer and process based, holonic to the limit of virtuality), competing different wars (wars of imagination) with different strategies (resource and capabilities based) and different weapons (trust, knowledge, creativity, integration, and networking abilities).

2. Marketing Strategies in an External and Internal Perspective

Marketing strategies can be viewed in two different ways.

(a) With an external emphasis, typical of the 'Market Power School,' which dominates the 1980s. This school draws on the Harvard tradition (Mason 1939, Bain 1956) in industrial economics studies. However, Porter (1980, 1985) is the most important author in this stream of thought.

(b) With an internal emphasis, typical of the Resource Based View of the Strategy, which dominates the 1990s. Wernerfelt (1984), Rumelt (1984, 1987), Barney (1986), Dierickx and Cool (1989), Itami (1987), Grant (1991), Peteraf (1993), and Prahalad and Hamel (1990, 1994) are the most important authors in this stream of thought.

The 'Market Power School' puts its emphasis on the analysis of the existing environment, following the structure-conduct-performance paradigm of industrial organization. The normative prescriptions of the Market Power School, particularly the three generic strategies suggested by Porter (1980, 1985)—differentiation, cost leadership, and niche strategies of focalization—depend on the external analysis of the environment, conducted mainly with the five-forces model. The five-forces model is an analytical tool developed by Porter (1980, 1985) which helps to individuate the attractiveness of an industry, by analyzing not only the competitors actually in the arena, but also the forces of the potential entrants, the producers of substitutes, and the influences of customers and suppliers.

Therefore, the 'Market Power School' considers strategy primarily as positioning a firm in a given industry context; the focus is on the external analysis of the existing industry and on its attractiveness. The strategic process should determine a 'fit' between the given distinctive strengths of a firm and the given opportunities offered by the external environment.

In synthesis, the Market Power School considers the general strategic process, at the corporate level, as the allocation of the right mix of resources to the various Strategic Business Units, adopting a financial portfolio view, and the competitive and marketing strategies, at the business level, adopting a product-based view of competition.

The Resource based View of the Strategy, rather than on the external environment, puts its emphasis on the unique endowment of resources, assets, competencies, and capabilities of each firm. By creating and continuously incrementing its resource endowment, a firm becomes different from another. So the unique endowment of each firm is at the roots of its competitive advantage, whose sustainability is higher, the higher are the barriers raised to imitability. Marketing strategies are not concerned with the exploitation of the existing resources, but with the continuous management commitment to increase the resource endowment. An example of a resource that should be increased and not only exploited is brand equity (Aaker 1984, 1991).

Resources may be tangible or intangible; it has been widely recognized that marketing strategies based on intangible resources generate hard to imitate competitive advantages.

Typical marketing resources are trust and knowledge (Vicari 1991); both are intangible and strongly linked with customer satisfaction, which is at the beginning and the end of each marketing strategy. At the beginning, customer satisfaction addresses each marketing strategy, because it is the very aim of every successful strategy. At the end, it is the most important result of each marketing strategy, because the long-term survival of a firm depends on the level of satisfaction, loyalty, and trust of its customers (Reicheld 1996). So analysis, measurement, and improvement of customer satisfaction is one of the most important activities in marketing strategies in a resource-based perspective.

3. Marketing Strategies in a Resource-based-view Perspective

In recent years, firms tend to be living in a new competitive milieu (Prahalad and Hamel 1994), in a

new competitive landscape, which is characterized by deregulation, structural changes (mainly connected with the evolution of information technology), a growing number of mergers, acquisitions, and alliances, less protectionism, and changing customer expectations. We can distinguish three forces that are jointly, but radically reshaping the firm's environment: the customer dynamic changes, the enormous potential offered by the I T paradigm, and the consequently inevitable increase of the competitive pressure. Considering these forces together, it is possible to assert that 'during the last ten years competitive space has been dramatically altered' (Prahalad and Hamel 1994).

The increase in competitive pressure is so strong, for many firms and many industries, that some authors (D'Aveni 1994, Valdani 1995) have coined the term 'hypercompetition' to describe this new competitive milieu.

As a matter of fact, in such a competitive and turbulent environment, normative prescriptions based on static industry analysis tend to become unuseful: managers base their decisions less and less on industry analysis, and more and more tend to look inwardly, at their sets of resources and capabilities. As Grant (1991) puts it, when the external environment is in a state of flux, the firm's own resources and capabilities may be a much more stable basis on which to define its identity.

This trend is reinforced when we consider that industry borders are becoming more and more volatile, and in many cases they tend to disappear and to become confused and merged in a process of gradual convergence. Especially, but not only in high-tech environments, a lot of industries tend to converge towards something like a mega-industry, which seems to comprise the previous independent industries of computer, telecommunication, office equipment, entertainment, broadcasting, media publishing, consumer electronics, distribution, financial and insurance services, and so forth. The phenomenon of convergence, like a black hole, gradually is attracting and merging once-different competitive environments: the biggest part of the industrial and services world is falling down in the black hole of convergence, one industry after the other. Only few industries, perhaps the less value-added ones or those that will be marginal in the future, can escape this process of gradual convergence. Many authors have recognized this fact (e.g., Prahalad and Hamel 1994, Hamel 1996, Valdani 1995, and Yoffie 1997).

In this context, the external environment, and particularly industry, tends to lose its importance and potential of explanation in favor of the firm and of its marketing strategies, which are based on its unique stock of resources, competencies, and capabilities.

Whereas in the previous strategic paradigm, competitive and marketing strategies are considered at the product level, in the resource-based perspective marketing strategies recognize the logical and temporal dominance of resources upon products. Products are considered only as the momentaneous expression of the internal resources of the firm. Moreover, resources are not product-specific, but they can lead to the achievement of competitive advantage in a wide range of products and services. For example, Sony has developed the concept of 'pocketability,' in the sense that it has generated a broad set of resources and competencies in order to satisfy, through a lot of products, the customer's need of entertainment linked to pocket dimensions and not bound to specific places. Motorola has developed the concept of 'untethered communication,' through the accumulation of competencies in the field of distribution of voice, images, and data, independently by the specific place where customers are. 3M has defined its company strategy focusing not only on products, but also more generally on its customers' need to use 'subtle supports' and on resources and capabilities in order to serve this need in a broad sense. It is possible to go on, describing the concept of Apple's 'user friendliness' and of Canon's 'C&C (Computer and Communication) strategy' and many others, but it should be clear, on the basis of the examples above, that, in a resource based perspective:

(a) the recognition of customer's needs has a conceptual priority in each marketing strategy;

(b) it requires a high level of marketing capabilities, in order to avoid errors of marketing myopia; and

(c) not only products, but also mainly a firm's resources, competencies, and capabilities are at the basis of the success of a marketing strategy.

Marketing strategies among firms, therefore, are played at three levels that are:

(a) the product-based level of traditional marketing strategies;

(b) the competence- and resource-based level of proactive firms' marketing strategies; and

(c) the 'super-industry-' (which is still resource-based) level of marketing strategies in the era of industry convergence.

Moreover, if we consider the firm as a cognitive system (Weick 1979, Vicari 1991), the industry, and more generally the external environment, tend to become a subjective and not an objective construct. Marketing strategies, therefore, are aimed at enacting the environment, on the basis of the firm's set of resources. In this perspective, competitive and marketing strategies are connected strictly with creativity and knowledge management.

In an attempt at synthesis, it is possible to state that in a resource-based perspective, firms need to adopt a proactive approach (Valdani 1992) to their marketing strategies:

(a) focusing on obtaining, at the same time, variety, quality, speed, and learning;

(b) stretching and leveraging their market driving capabilities, in order to reach a double aim, in a so-called 'dual strategy' (Abell 1994, Valdani 1995).

Firms adopt a dual strategy when they want to manage the present and to prepare the future at the same time. Managing the present means developing capabilities of market and competitive intelligence and re-engineering the processes of the firm in a customer-focused perspective. Preparing the future means developing the managerial and entrepreneurial capabilities of destroying the existing orthodoxies, to imagine, enact, and create a new competitive environment, regenerating strategies and reinventing industries and their borders (Prahalad and Hamel 1994).

4. Marketing Strategies in a Circular Model

It is possible to describe the marketing strategies pursued by firms in the hypercompetition and convergence era by adopting a circular model described in Fig. 1.

The model shows the three possible states of the competitive environment in hypercompetitive conditions. They are: position wars, movement wars, and imitation wars.

4.1 Marketing Strategies in Position Wars

During a position war, the environment is quite stable, the product life cycle is generally in its maturity phase, and customers and competitors are well defined. General marketing strategies in position wars are those of market sharing, aimed at increasing a firm's market share by reducing the competitors' one: pricing strategies are very important in these conditions. Specific marketing strategies for aggressive competitors include the option among (a) a direct and frontal attack to the competitor; (b) lateral and smart moves; (c) guerrilla. Specific marketing strategies for defensive competitors include the options among (a) static defence and inertia; (b) reaction countermoves; (c) strategic retreat.

In the long run, position wars may be dangerous. Firms, in fact, begin to compete on quality and price, increasing quality and decreasing price, in a dangerous and hypercompetitive evolution to the so-called 'ultimate value frontier' (D'Aveni 1994, Valdani 1995). The 'ultimate value frontier' is a situation in which quality is at the highest level, price is at the lowest, and firm's profit at the 'neoclassical' normal level (where

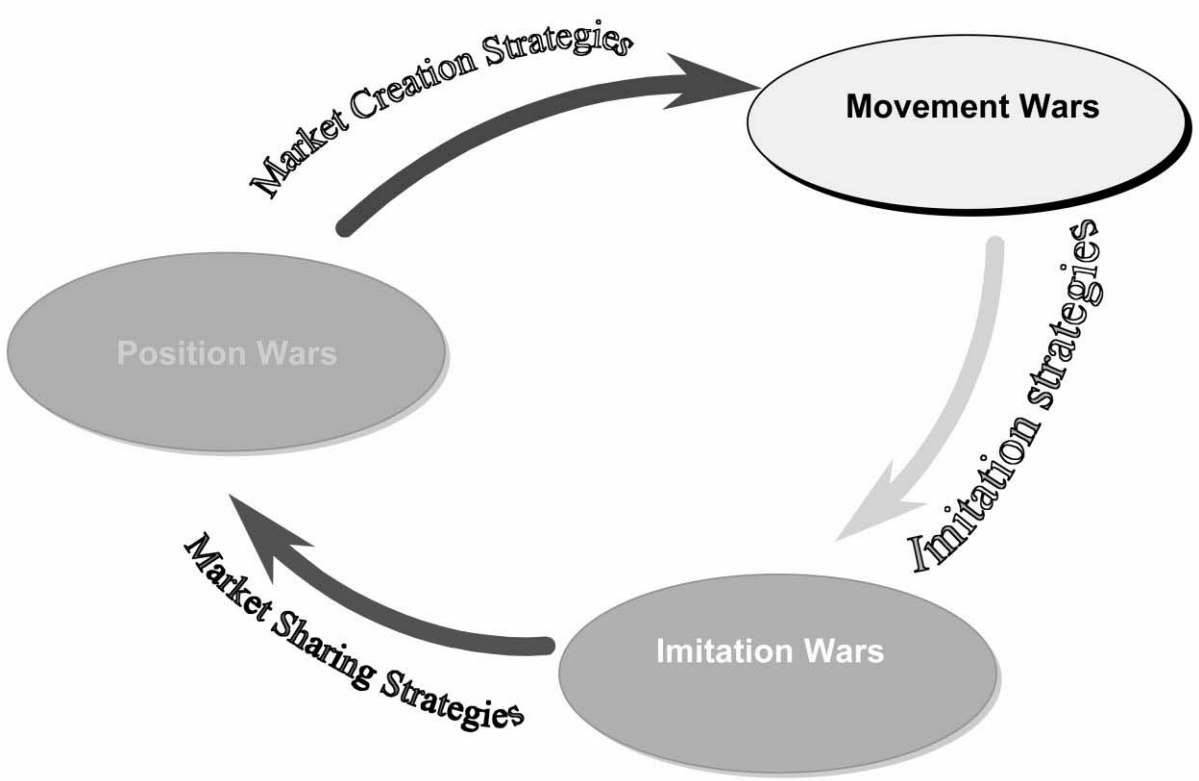

Figure 1
Marketing strategies in a circular model (E Valdani 1995)

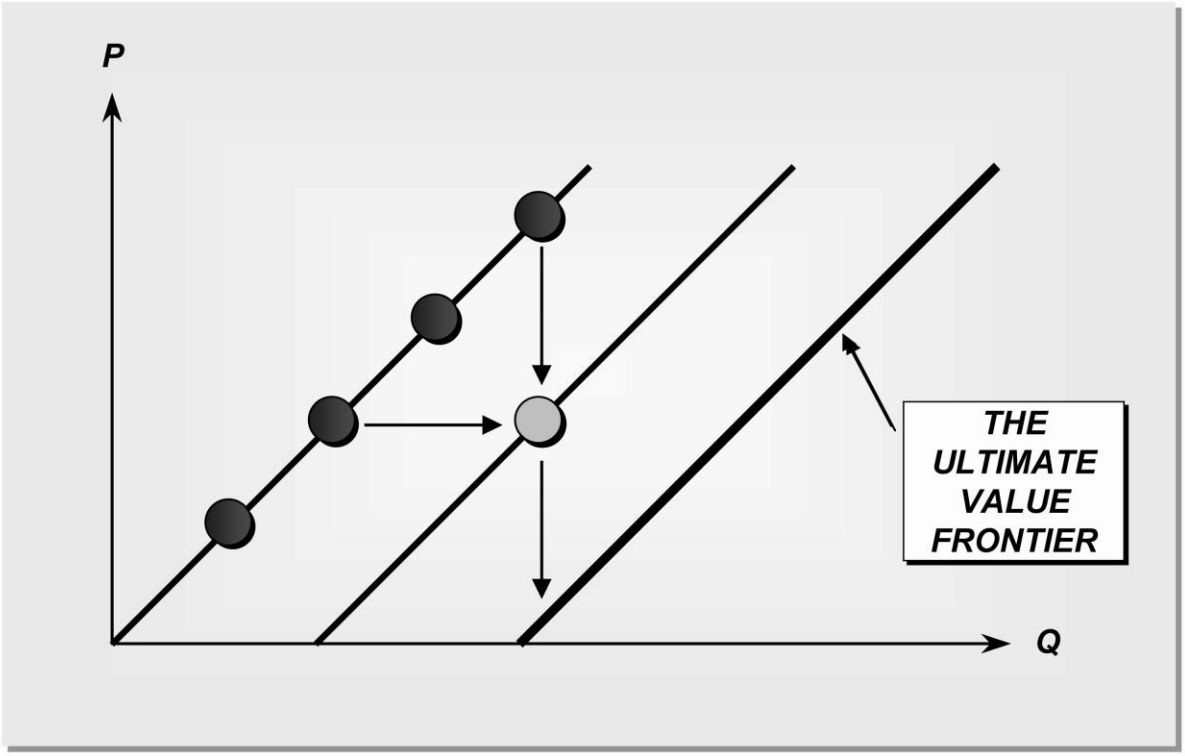

Figure 2
The ultimate value frontier (E Valdani 1995)

normal, in managerial and entrepreneurial terms, means unsatisfactory). The process of evolution to the 'ultimate value frontier' is described in Fig. 2.

4.2 Marketing Strategies in Movement Wars

In order to avoid the evolution to the 'ultimate value frontier,' firms pursue strategies of market creation and enter in movement wars. In a proactive perspective, they increment, leverage, and stretch their market driving capabilities and resources and, by doing so, create new markets or reconfigure radically the existing ones. So, innovation strategies are the best marketing strategies in movement war: they are particularly critical in contexts of high hypercompetition and industry convergence. Innovation and market creation strategies usually require also networking strategies: in fact, a marketing creation strategy is successful only in a co-evolutionary perspective. Last, but not least, strategies of market creation and innovation require, in order to be successful, the ability of a firm to overcome barriers to innovation, both external (i.e., customer inertia) and internal (i.e., specialization).

4.3 Marketing Strategies in Imitation Wars

As it is well known, innovation usually is followed by imitation. Follower firms, pursuing imitation strategies, transform a movement war into an imitation war. Imitation strategies are successful only when firms' endowment of capabilities and resources include rapid learning abilities.

As it should be clear, in the circular model adopted, imitation wars gradually evolve to position wars and these to movement wars, and so on, in a circular perspective.

5. Conclusions: Marketing Strategies in the New Millennium

In the third millennium firms must be able to evolve from an era of hypercompetition to one of imagination. Marketing strategies in an era of imagination are successful only if firms:

(a) adopt a resource based perspective;

(b) focus on customers, pursuing customer sat-i-sfaction and loyalty, and when it is possible, 'customer delight,' that is, exceeding customer expectations;

(c) continuously regenerate their stock of market driving capabilities, in order to avoid the trap of the 'ultimate value frontier' and shift continuously from position to movement and imitation wars;

(d) co-evolve with other firms and institutions, networking with them, with the ultimate aim of reconfiguring existing industries, driving their convergence, and creating new industries in the digital world of the third millennium.

In order to win the wars of the third millennium, firms must develop marketing strategies using weapons completely different from the past. These new weapons are intangible resources and capabilities like trust, knowledge, creativity, integration, and networking abilities. Obviously, new weapons can only be managed by new warriors, that is, new firms, which must be proactive, lean, customer and process based, holonic to the limit of virtuality.

See also: Alliances and Joint Ventures: Organizational; Consumer Economics; Consumer Psychology; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; International Business; International Law and Treaties; International Marketing; International Organization; International Trade: Commercial Policy and Trade Negotiations; Market Areas; Markets and the Law; Monetary Policy; Rational Choice Theory in Sociology; Transaction Costs and Property Rights; World Trade Organization

Bibliography

Aaker D A 1984 *Strategic Marketing Management*. Wiley, New York

Abell D 1993 *Managing Dual Strategy*. Free Press, New York

Bain J 1956 *Barriers to New Competition*. Harvard University Press, Cambridge, MA

Barney J S 1986 Strategic factor markets: Expectations, luck and business strategy. *Management Science* **32**: 1230–41

D'Aveni R 1994 *Hypercompetition*. Free Press, New York

Dierickx I, Cool K 1989 Asset stock acccumulation and sustainability of competitive advantage. *Management Science* **35**: 1504–11

Grant R M 1991 *Contemporary Strategy Analysis. Concepts, Techniques, Applications*. Blackwell, Oxford,UK

Hamel G, Heene A (eds.) 1994 *Competence Based Competition*. Wiley, Chichester, UK

Itami H 1987 *Mobilizing Invisible Assets*. Harvard University Press, Cambridge, MA

Mason E S 1939 Price and production policies of large scale Enterprises. *American Economic Review* **29**: 61–74

Peteraf M A 1993 The cornerstones of competitive advantage: A resource based view. *Strategic Management Journal* **14**: 179–91

Porter M E 1980 *Competitive Strategy: Techniques for Analysing Industries and Competitors*. Free Press, New York

Porter M E 1985 *Competitive Advantage: Creating and Sustaining Superior Performance*. Free Press, New York

Prahalad C K, Hamel G 1994 *Competing for the Future*. Harvard Business School Press, Boston

Reicheld F 1996 *The Loyalty Effect*. Harvard Business School Press, Boston

Rumelt R P 1987 Theory, strategy and entrepreneurship. In: Teece D J (ed.) *The Competitive Challenge*. Ballinger, Cambridge, MA

Valdani E 1995 *Marketing Strategico*. Etas Libri, Milan, Italy

Yoffie D B 1997 *Competing in the Age of Digital Convergence*. Harvard Business School Press, Boston

E. Valdani

Markets and the Law

A 'market' is a forum of exchange, a place where trade is carried out. In a narrow sense the term refers to a physically defined locality where buyers and sellers meet, like a neighborhood farmers' market or a local fair. In its more abstract meaning it refers to any kind of exchange network, in the general sense of a social arrangement in which participants are interconnected through actual and potential exchange transactions. Depending on what criterion is used to define the relevant set of actual or potential trading parties, various kinds of special markets can be distinguished for analytical purposes, even if one acknowledges that in reality such markets cannot be strictly isolated from the more inclusive exchange environment. Financial markets, the US used-car market, the world crude-oil market, or the market for Van Gogh paintings are examples from an indefinitely long list of special markets that might be defined.

1. Markets as Social Arrangements

As social arrangements markets are constituted by bilateral, actual and potential, exchange transactions. By contrast to theft or coercive taking, exchange is a peaceful method of obtaining things that one desires. It is based on mutual agreement between the trading parties. Given the noted alternative methods of personal enrichment, people can be expected to engage in exchange when and where the alternatives appear less attractive. This is normally the case where people meet within a normative-legal-institutional framework that defines and enforces property rights, though, even in the absence of a shared normative order, people may have prudent reasons for pursuing their interests through exchange rather than violent methods. As Max Weber (1978, p. 640) observed, even someone who, prefers to take without pay whatever he can, may choose to resort to peaceful exchange where he is 'confronted with a power equal to his own,' or where he regards it 'as shrewd to do so for the sake of future exchange opportunities which might be endangered otherwise.' In fact, the interest in exploiting potential gains from trade outside of one's own inherited

community can be seen as a principal driving force in the evolution of a normative-legal order that extends beyond traditional community limits. As Weber (1978, p. 637) put it, 'The market is a relationship which transcends the boundaries of neighborhood, kinship group, or tribe. Originally, it is indeed the only peaceful relationship of such kind.'

To say that markets are constituted by *potential* as well as actual exchange transactions points to the role of *competition* as the essential market force, a role that Weber (1978, p. 635) underlines when he defines, 'A market may be said to exist wherever there is competition, even if only unilateral, for opportunities of exchange among a plurality of potential parties.' Competition means that sellers can choose among potential alternative buyers, and that buyers can choose among potential alternative sellers. The terms under which exchanges are actually carried out in a market cannot be adequately understood without considering the potential alternative transactions that the respective parties might have chosen, but did not. As, market transaction 'is always a *social act* insofar as the potential partners are guided in their offers by the potential action of an indeterminately large group of real or imaginary competitors rather than by their own actions alone' (Weber 1978, p. 636).

'*Competition for opportunities of exchange*' (Weber 1978) is constitutive of markets. It is also the source of a fundamental ambiguity in people's attitudes towards markets. While one's own interests are furthered by competition on the other side of the transaction, competition on one's own side is often perceived as a nuisance. As seller one welcomes any increase in the pool of potential buyers, and as buyer one welcomes an increase in the plurality of potential sellers, because this can only improve the terms of trade. Conversely, competition on one's own side of the transaction, be it as buyer or as seller, is much less welcome as it tends to limit the gains that one can hope to realize in the exchange. Despite the benefits that an open market with free entry of potential buyers and sellers has to offer to all parties, there are obvious benefits to be had from the privilege of being in one's own role protected from competition. Interests in securing the benefits of protectionist privileges, on the one side, and interests in realizing the gains that can be had from 'exchange with the highest bidder' (Weber 1978, p. 638), on the other side, are two opposing forces that shape a political community's legal-institutional framework and determine the extent to which it facilitates or inhibits trade within and across its boundaries. Medieval feudal and guild restrictions heavily favored the former; modern market economies are the product of the growing weight of the latter interests.

Markets are a paradigm example of a self-generating or *spontaneous* social order (Hayek 1973, p. 37), i.e., of social arrangements in which the activities of participants are coordinated in a spontaneous manner, through mutual adjustment or adaptation of separate decision-makers, without any deliberate, central direction. In this sense the order of the market can be contrasted 'as a specific type of social structure' (Swedberg 1994, p. 255) to the deliberate, centralized coordination of activities that occurs within *corporate entities* or *organizations*, i.e., within social units such as the 'family, the farm, the plant, the firm, the corporation, and the various associations, and all the public institutions including governments' (Hayek 1973, p. 46). It is one of the central themes in the works of F. A. Hayek that the distinction between 'the two kinds of order' (Hayek 1973, p. 46), *market* and *organization* (Vanberg 1982), is of fundamental importance for an adequate understanding of the nature of societal phenomena in general and of the order of the market in particular. The failure to appreciate adequately the nature of the market as a spontaneous social order is, in Hayek's view, a major source of confusion in discussions on economic theory and, in particular, economic policy, a confusion that he attributes in part to the ambiguity that is implied when the term 'economy' is used to describe the order of the market. As the term is derived from the Greek word *oikonomia*, which means *household-economy*, an 'economy, in the strict sense of the word, is an organization or arrangement in which someone deliberately allocates resources to a unitary order of ends' (Hayek 1978, p. 178). In order to avoid any misleading connotations Hayek suggests speaking of the order of the market not as an economy but as a *catallaxy*—derived from the Greek word *katallatein*, which means 'to exchange' (Hayek 1976, p. 108).

According to Hayek (1976, p. 115), the operation of the market system and the way it coordinates the actions of market-participants can be understood best by thinking of it as a game, 'the game of catallaxy.' The game metaphor is meant to emphasize two essential attributes of the competitive market process. First, it 'proceeds, like all games, according to rules guiding the actions of individual participants' (Hayek 1976, p. 71). And, second, as with all genuine games, the particular outcomes of the 'game of catallaxy' cannot be predetermined but must always remain to a large extent unpredictable, owing to the multitude of contributing factors and to the inventiveness of the participants who are free to choose their strategies within the limits defined by the general rules of the game. Indeed, that particular market outcomes cannot be predetermined is but a consequence of the fact that the 'rules of the game' are the essential device by which the moves of players in the 'game of catallaxy' are coordinated. These rules are typically negative rules that exclude as impermissible certain kinds of strategies, but leave significant scope for choice. By contrast, the essential coordinating device within organizations or corporate arrangements are positive *commands* rather than (negative) general rules of conduct, commands either in the form of specific orders or in the form of generalized commands as they are implied in

'organizational rules' that define the tasks to be performed by persons in particular organizational positions.

By speaking of the market as a 'game of competition' that is played according to certain rules, Hayek underscores the inherent connection between *markets* and the *law*. Since the coordination of actions within markets is based on certain general rules of conduct that impose constraints on the behavior of market participants, it follows that only where suitable rules are in force can a market order be expected to emerge at all, and that the particular nature of the legal-institutional framework within which markets operate will determine their overall working properties. As Hayek (1960, p. 229) puts it: 'If there is to be an efficient adjustment of the different activities in the market, certain minimum requirements must be met; the more important of these are … the prevention of violence and fraud, the protection of property, and the enforcement of contracts, and the recognition of equal rights of all individuals to produce in whatever quantities and sell at whatever prices they choose. Even when these basic conditions have been satisfied, the efficiency of the system will still depend on the particular content of the rules.'

If, as Hayek points out, the order of the market is based on rules, one should expect the 'relation between the character of the legal order and the functioning of the market system' (Hayek 1960 p. 229) to be a central theme of the social science that concerns itself with the study of markets, economics.

2. *Economics as the Science of Markets*

The study of markets as exchange arrangements has traditionally been considered the domain of economics (Weber 1978 p. 635). Indeed, it has even been suggested that one should speak of economics as *catallactics*, in order to underscore the discipline's principal concern with the market as a spontaneous exchange-order or catallaxy (Mises 1949, pp. 233ff, Hayek 1976, p. 108, Buchanan 1979, pp. 19, 27). In light of their universally acknowledged prominence as the discipline's principal subject, it is more than surprising how little attention has been paid in mainstream economics to an explicit discussion of markets as social arrangements and, in particular, to the issue of how their working properties are affected by the legal-institutional framework within which they operate.

The 'peculiar fact that the literature on economics … contains so little discussion of the central institution that underlies neo-classical economics—the market' (North 1977, p. 719) has repeatedly been noted (see, e.g., Hayek 1960, p. 229; Brennan and Buchanan 1985, p. 13, Coase 1994, p. 6, Swedberg 1994, p. 257). In standard economics textbooks, markets are typically defined in terms of the institutionally disembodied mechanics of demand and supply, i.e., 'the forces created by buyers and sellers that establish the prices and quantities exchanged of resources, goods and services,' while, as Coase (1988, p. 7) notes, 'discussion of the market itself has entirely disappeared.' In mainstream economic theory, Coase (1988, p. 5) summarizes, the market is, for the most part, assumed to exist and is not itself the subject of investigation, with the result 'that the crucial role of the law in determining the activities carried out … in the market has been largely ignored.'

The neglect of the institutional dimension of markets in mainstream economics can be directly traced back to the research program that Leon Walras defined in his *Éléments d'Économie Politique Pure* (first published in 1874) for the 'science of pure economics,' a program that gave rise to the modern neoclassical orthodoxy and its theoretical pinnacle, the Arrow–Debreu general equilibrium model. Strongly influenced by the theoretical physics of his time, Walras' ambition was to develop 'the pure theory of economics or the theory of exchange and value in exchange … (as) a physico-mathematical science like mechanics or hydrodynamics' (Walras 1954, p. 71), sciences that, as he saw it, 'abstract ideal-type concepts … and … construct *a priori* the whole framework of their theorems and proofs' (Walras 1954, p. 71). The task he defined for the science of economics was to follow 'this same procedure,' namely to form an ideal-type concept of the market, with 'ideal prices which stand in an exact relation to an ideal demand and supply' (Walras 1954, p. 71). Pure economics, thus conceived, is concerned with 'how prices result from under a hypothetical régime of absolutely free competition' (Walras 1954, p. 71), i.e., a régime with no obstacles to the realization of potential gains from trade and to the attainment of 'the maximum of utility' (Walras 1954, p. 71). Pure economics need not be concerned with whether or not the hypothesized 'absolutely free competition' can be observed 'in the real world' (Walras 1954, p. 255). It supposes 'that the market is perfectly competitive, just as in pure mechanics we suppose, to start with, that machines are perfectly frictionless' (Walras 1954, p. 84).

From the domain of 'pure economics' which studies 'the nature, causes and consequences of free competition' (Walras 1954,, p. 255), Walras explicitly excluded phenomena that are 'to be classified under the heading of institutions' (Walras 1954, p. 63), in particular phenomena concerned with the 'problem of property' or 'the mode of appropriation' of scarce things (Walras 1954, p. 77). In his account, pure economics bears 'clearly the mark of a natural science' (Walras 1954, p. 56). Its subject, value in exchange, 'partakes of the character of a *natural* phenomenon' (Walras 1954, p. 70), because it results 'naturally under given conditions of supply and demand' (Walras 1954, p. 69). By contrast, the 'theory of property' or the 'theory of institutions' is, in Walras' scheme of things, assigned to the domain of 'moral science or ethics'

(Walras 1954, pp. 63, 79), because its subject, the mode of appropriation, is 'not a natural phenomenon' (Walras 1954, p. 76) but 'depends on human decisions' (Walras 1954, p. 77). It is a 'phenomenon which is fundamentally social and which gives rise to questions of justice or of the mutual coordination of human destinies' (Walras 1954, p. 77).

In Walras' research program, pure economics was not meant to be *all* of economics. While the study of institutions and the rules of property was seen to be outside of the domain of 'economics as an exact science' (Walras 1954, p. 47), it was considered to be part of economics as a more broadly conceived enterprise. As Walras put it, 'Appropriation being in essence a moral phenomenon, the theory of property must be in essence a moral science … or, as we shall designate it, *social economics*' (Walras 1954, p. 79). However, the part that his 'social economics' would have had to play in a more broadly conceived economics was never developed in what has come to be known as the Walrasian tradition in economics. The neoclassical mainstream tradition remained occupied with advancing and formalizing in ever more refined ways Walras' program for 'a scientific theory of the determination of prices' (Walras 1954, p. 40), and left unattended the institutional issues that Walras had assigned to 'social economics.'

3. *Institutional Economics: Old and New*

Pioneering as Walras' enterprise of formalizing economic theory along the model of a physico-mathematical science was, it was by no means an entirely unprecedented project. It was very much in line with an established trend among mainstream economists of developing ever more abstract models of the price system, a trend the early beginnings of which are marked by the change in emphasis from the work of Adam Smith with its attention to institutional issues to the writings of David Ricardo from which such attention had largely disappeared (Demsetz 1982, p. 6). Post-Walrasian neoclassical economics must therefore be seen as merely continuing a theoretical course that had already been embarked on before, a course that made it treat markets in a manner about which Demsetz (1982, pp. 6ff) notes, 'Markets became empirically empty conceptualizations of the forums in which exchange costlessly took place. The legal system and the government were relegated to the distant background by the simple device of stating, without clarification, that resources where "privately owned".'

It was in opposition to such narrow focus on 'the market as an abstract price-making mechanism' (Swedberg 1994, p. 255) that unorthodox approaches such as, in particular, the German Historical School and the American institutionalist tradition insisted on the necessity of paying attention to the details of the legal-institutional framework within which market activities are carried out, if one is to arrive at an adequate understanding of real world economic processes.

In his *Grundriss der Allgemeinen Volkswirtschaftslehre*, Gustav Schmoller (1904), head of the 'younger' Historical School, commences an extended discussion on the concept of the 'market' by stressing that in any community of peacefully trading people economic transactions take place under the umbrella of customary, legal, and moral rules, and that knowledge of the historical development of such legal-institutional provisions is prerequisite for an understanding of the development of trade and commerce (Schmoller 1904, p. 15). There never existed, he maintains, anything like 'absolutely free competition,' but economic activities were always embedded in a 'system of norms, of constraints, guidelines, laws, and prohibitions, that regulated the stream of economic life, by ruling that certain kinds of agreements are illegal or not binding, that some contracts are not enforceable while others are void or contestable' (Schmoller 1904, p. 16). After such an introductory reminder, Schmoller provides a detailed historical account of the evolution of market institutions, from their primitive origins in intertribal arrangements for peaceful exchange to their modern incarnations (Schmoller 1904, pp. 17ff).

Among the American institutionalists, John R. Commons is known for his painstaking efforts to portray, in great detail, the legal-institutional framework that conditions the operation of markets as social arrangements (Vanberg 1997). In his *Legal Foundations of Capitalism* (Commons 1957) (originally published in 1924), he traced the gradual process of legal evolution from which the kind of socioeconomic order emerged upon which the analysis of economics as the study of markets concentrates. His central message was that the system of market competition, that is at the core of economic theory is not a 'natural phenomenon' but is rather a societal enterprise in the sense that it presupposes a legal framework which is a product of civilization, not a 'provision of nature.' As he reasoned, the 'simple system of natural liberty' that Adam Smith spoke about, and that his formula of the 'invisible hand' was meant to describe, consisted of 'nothing other than the working rules of an orderly society' (Commons 1957, p. 137), an institutional framework that was by no means a present of nature but 'the fine fruit of evolving centuries of working rules' (Commons 1957, p. 138).

The 'old' institutionalist critique did little to change the analytical focus of mainstream economics, partly because it not only challenged the institution-blindness of economic orthodoxy but also seemed to depart from 'hard-core' elements of the economic tradition, in particular its methodological individualism; partly because, faced with the 'infinite variety of market phenomena' (Schmoller 1904, p. 112), the institutionalists' concern with historical detail tended to result in descriptive studies with little theoretical focus. A more

serious challenge to the self-sufficient Walrasian tradition of pure economics has come, though, from a number of unorthodox approaches that began to emerge in the 1960s and that are, summarily, referred to as the 'New Institutional Economics.' These approaches, including the Economics of Property Rights, Law and Economics, Public Choice, and others, seek to correct for the institutional deficiency of mainstream neoclassical economics, yet—by contrast to the 'old' institutionalism—remain firmly within the classical economic tradition, in fact, they are viewed by their advocates as a revival of essential parts of the Smithian heritage.

One of the most influential contributions to the development of the New Institutional Economics, and one that is of particular interest in the present context, stems from Ronald Coase. In two complementary, path-breaking articles, Coase points to *transaction costs*, i.e., the costs of carrying out economic transactions among persons, as the key to an understanding of the role of market institutions. In his article 'The nature of the firm' (Coase 1988, pp. 33–55), a paper that was originally published in 1937 but made its impact only decades later, he raises the question of why, given the efficiency of market coordination, there exist at all firms, (centrally coordinated organizations) within the spontaneous, noncentralized order of the market. He suggests that firms exist because there are costs to market transactions, such as the costs of finding a suitable exchange partner and of negotiating and enforcing agreements. To the extent that such costs can be reduced by organizing transactions within firms, there is an economic rationale for their existence. In the article 'The problem of social cost' (Coase 1988, pp. 95–156; originally published 1960), Coase argues that the legal framework within which market transactions take place *matters* because it affects the costs at which such transactions can be carried out. Ironically, Coase's argument became mostly known, in its *negative* version as the so-called *Coase theorem*, which says that, as long as property rights are clearly defined, the allocation of resources is not affected by *how* they are defined, *if transaction costs are zero*. The reason is that, no matter what the law says about 'who has the right to do what,' with zero transaction costs, rational economic agents can always trade entitlements until they are in the hands of those who can put them to their most valuable uses (Coase 1988, p. 14).

It is ironic that Coase's argument came to be condensed as the Coase theorem, since the theorem's 'institutions do not matter' message is exactly the opposite of what Coase wanted to emphasize (Coase 1988, pp. 15, 174). The fictitious world of zero transaction costs in which such institutional neutrality would hold is, as he points out, the world 'of standard economic theory' (Coase 1994, p. 10) but not the world that he finds worthwhile to study. If we move, he notes (Coase 1994, p. 11), from the hypothetical regime of zero transaction costs to the real world of positive transaction costs, what becomes immediately clear is '… (that) the legal system will have a profound effect on the working of the economic system and may in certain respects be said to control it.' In Walras's (1954, p. 256) 'hypothetical régime of absolutely free competition,' the model-world of neoclassical orthodoxy, it is sufficient to assume that private property rights are assigned *somehow*. If, as is presumed for such a regime, no obstacles exist to the realization of potential gains from trade, then rational and fully informed economic agents can be predicted to fully exhaust all potential gains so as to attain 'the maximum of utility' (Walras 1954, p. 256). It is in such a context that, for instance, G. Debreu can say of his *Axiomatic Analysis of Economic Equilibrium* that its task is to explain how prices result 'from the interaction of the agents of a private ownership economy' (Debreu 1959, p. vii), without giving any further detail of the kinds of institutions that characterize his 'private ownership economy.'

As strange as the neglect (among mainstream economists) of the market as an institutional arrangement may seem, it is, Coase argues, a systematic consequence of the orthodox model. As he puts it, 'Markets are institutions that exist to facilitate exchange, that is, they exist in order to reduce the cost of carrying out exchange transactions. In an economic theory which assumes that transaction costs are nonexistent, markets have no function to perform' (Coase 1988, p. 7). For an economic theory, that aims at explaining the real world of positive transaction costs, it 'is hardly possible to discuss the functioning of a market without considering the nature of the property rights system, which determines what can be bought and sold and which, by influencing the cost of carrying out various kinds of market transactions, determines what is, in fact, bought and sold, and by whom' (Coase 1994, p. 46).

In other words, the separation that Walras wanted economists to draw between a 'pure economics,' concerned—as an 'exact science'—with the natural mechanics of demand and supply, and a 'social economics,' concerned—as a 'moral science'—with institutional and property rights issues, is difficult to maintain once one acknowledges 'that what are traded on the market are not, as is often supposed by economists, physical entities but the rights to perform certain actions, and the rights which individuals possess are established by the legal system' (Coase 1994, p. 11).

4. Markets and Economic Policy

The difference in their respective outlooks at markets that separates mainstream neoclassical economics from the new institutional economics has its counterpart in differences between their respective approaches to issues of economic policy. Taking the concept of

perfect competition as its starting point, the neoclassical approach tends to judge real world markets in terms of its reference-model, and where it finds reality to fall short of the ideal standard, it diagnoses a need for political correction of 'market failure.' Critics have chastised such reasoning as a 'nirvana approach' (Demsetz), noting that 'we do injustice to the achievement of the market if we judge it … by comparing it with an ideal standard which we have no known way of achieving' (Hayek 1978, p. 185).

By contrast, by looking at the market as a 'social institution which facilitates exchange' (Coase 1988, p. 8), the new institutional economics starts from the recognition that markets are legal-institutional arrangements and that all we can meaningfully compare—and choose among—are alternative, actual or potential legal-institutional frameworks. As Demsetz (1969, p. 1) summarizes the contrast: 'The view that now pervades much public policy economics implicitly presents the relevant choice as between an ideal norm and an existing 'imperfect' institutional arrangement. This *nirvana* approach differs considerably from a *comparative institution* approach in which the relevant choice is between alternative real institutional arrangements. In practice, those who adopt the nirvana approach seek to discover discrepancies between the ideal and the real and, if discrepancies are found, they deduce that the real is inefficient. Users of the comparative institution approach attempt to assess which alternative real institutional arrangement seems best able to cope with the economic problem.'

If markets have to be seen as legally framed social arrangements, there is no other way in which they could be meaningfully described than in terms of their legal-institutional makeup. The way they have used the concept of the 'perfect market' has long served economists as a convenient device to evade the need to be specific about what kind of legal-institutional arrangement they mean to be descriptive of a 'perfect market.' Yet, if that concept is to provide any real guidance to economic policy, it must be specified in legal-institutional terms. And it is indeed, as Coase (1988, p. 9) notes, not without significance, that the kinds of organized markets, like stock exchanges or commodity exchanges, that are 'often used by economists as examples of a perfect market and perfect competition, are markets in which transactions are highly regulated.' It suggests, Coase adds, 'that for anything approaching perfect competition to exist, an intricate system of rules and regulations would normally be needed' (Coase 1988, p. 9).

The comparative institutions approach puts an end to the pretence that anything meaningful is said by reference to 'perfect markets' in economic policy discourse, as it insists that all diagnoses of deficiencies in the operation of existing markets, and all suggestions for corrections, have to be specified in terms of feasible alternative institutional provisions. It views economic policy as 'a choice among alternative social institutions' (Coase 1988, p. 28), a choice that is to be informed by knowledge of how 'the economic system would work with alternative institutional structures' (Coase 1988, pp. 19ff). It recognizes that the task of responsible economic policy cannot be to measure the performance of real world markets in terms of an unattainable theoretical construct, but is 'to devise practical arrangements which will correct defects in one part of the system without causing more serious harm in other parts' (Coase 1988, p. 142).

The comparative institutions approach to economic policy requires, as its theoretical foundation, an economics that provides knowledge of feasible institutional options, of the working properties of alternative institutional arrangements, and of the predictable effects of potential institutional reforms. To provide such knowledge opens up a rich and demanding research program for economists to pursue. To be sure, the adoption of such a program would mean a significant change from the direction that modern economics has come to take, even if, in essence, it would only mean to reactivate an integral part of its original, Smithian research agenda.

See also: Business Law; Hierarchies and Markets; International Trade: Commercial Policy and Trade Negotiations; Law and Economics; Law and Economics: Empirical Dimensions; Lex Mercatoria; Market and Nonmarket Allocation; Market Structure and Performance

Bibliography

Brennan G, Buchanan J M 1985 *The Reason of Rules—Constitutional Political Economy*. Cambridge University Press, Cambridge, UK

Buchanan J M 1979 *What Should Economists Do?* Liberty Press, Indianapolis, IN

Coase R H 1988 *The Firm, the Market, and the Law*. University of Chicago Press, Chicago

Coase R H 1994 *Essays on Economics and Economists*. University of Chicago Press, Chicago

Commons J R 1957 1924 *Legal Foundations of Capitalism*. University of Wisconsin Press, Madison, WI

Debreu G 1959 *Theory of Value, an Axiomatic Analysis of Economic Equilibrium*. Charles Foundation Monograph No. 17, Wiley, New York

Demsetz H 1969 Information and efficiency: Another viewpoint. *Journal of Law and Economics* **12**: 1–22

Demsetz H 1982 *Economic, Legal, and Political Dimensions of Competition*. North-Holland, Amsterdam

Hayek F A 1960 *The Constitution of Liberty*. University of Chicago Press, Chicago

Hayek F A 1973 *Law, Legislation and Liberty*. University of Chicago Press, Chicago, Vol. 1

Hayek F A 1976 *Law, Legislation and Liberty*. University of Chicago Press, Chicago, Vol. 2

Hayek F A 1978 *New Studies in Philosophy, Politics, Economic and the History of Ideas*. University of Chicago Press, Chicago

Mises L von 1949 *Human Action—A Treatise on Economics*. Yale University Press, New Haven, CT

North D C 1977 Markets and other allocation systems in history: The challenge of Karl Polanyi. *Journal of European Economic History* **6**: 703–16
Schmoller G 1904 *Grundriss der Allgemeinen Volkswirtschaftslehre. Teil 2*. Duncker und Humblot, Leipzig
Swedberg R 1994 Markets as social structures. In: Smelser N J, Swedberg R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ, pp. 255–82
Vanberg V J 1982 *Markt und Organisation*. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany
Vanberg V J 1997 Institutional evolution through purposeful selection: The constitutional economics of John R. Commons. *Constitutional Political Economy* **8**: 105–22
Walras L 1954 [1874] *Elements of Pure Economics or the Theory of Social Wealth* [trans. Jaffé W]. Richard D. Irwin, Homewood, IL
Weber M 1978 *Economy and Society: An Outline of Interpretive Sociology* [trans. Fischoff E]. University of California Press, Berkeley, CA, 2 Vols.

V. Vanberg

Markets: Anthropological Aspects

Markets are so routinely regarded as fundamentally economic institutions that long-standing and quite varied anthropological perspectives on them are often overlooked. Anthropological attention focuses on patterns of individual and small-group exchange relationships within specific markets, on institutional structures that organize markets, and on the social, political, and spatial hierarchies through which markets link social classes, ethnic groups, or regional societies into larger systems. Anthropological studies of markets analyze them as nodes of complex social processes and generators of cultural activity as well as realms for economic exchange. Anthropologists' interests in markets, therefore, are partially distinct from—although certainly overlapping with—the concerns of economists.

1. *Markets and Marketplaces*

The term 'market' is inherently ambiguous. Abstractly, 'market' refers to exchange organized around principles such as 'price' or 'supply-and-demand.' 'Market' may also refer to specific social relationships and frameworks through which economic transactions take place. Markets, in the first sense, are networks of economic processes and transactions which may occur without specific locations or spatial boundaries for the transactional universe. In the second sense, markets are social institutions, often located in geographically distinct places, which encompass specific social, legal, and political processes that enable economic transactions, but also extend far beyond them. Plattner (1989) makes a useful distinction:

> 'market' [is] the social institution of exchanges where prices or exchange equivalencies exist. 'Marketplace' refers to these interactions in a customary time and place. … A market can exist without being localized in a marketplace, but it is hard to imagine a marketplace without some sort of institutions governing exchanges. (p. 171)

Marketplaces embody a localized set of social institutions, social actors, property rights, products, transactional relationships, trade practices, and cultural meanings framed by a wide variety of factors including, but not limited to, 'purely economic' or 'market' forces. Of course, anthropological analyses of markets in the first sense are often ethnographically focused on marketplaces in the second sense.

Anthropological approaches to markets sometimes focus on the formal properties of exchange systems as frameworks for organizing behavior, relying on quantitative analyses of exchange relationships. However, anthropologists generally place such analyses within wider ethnographic contexts that see marketplaces as specific locations and social frameworks, characterized not only by economic exchanges in and among them, but also by their equally vital roles as arenas for cultural activity and political expression, nodes in flows of information, landmarks of historical and ritual significance, and centers of civic participation where diverse social, economic, ethnic, and cultural groups combine, collide, cooperate, collude, compete, and clash. Anthropological and sociological analyses emphasize this 'embeddedness' of markets in ongoing patterns of social organization and cultural meaning (Polanyi et al. 1957, Granovetter 1985); that is, economic behavior is not analyzed as an autonomous sphere of human activity, but as inseparably intertwined with a wide variety of social, political, ritual, and other cultural behaviors, institutions, and beliefs.

2. *Marketplaces as Ethnographic Sites*

Marketplaces vary enormously. They differ according to the local, regional, or global scope of production, distribution, or consumption of the goods and services they trade. Retail and wholesale markets are structured quite differently around the distinct activities and social roles of consumers, producers, and traders. Some markets handle physical commodities; others trade intangible financial assets. Many marketplaces are permanent, but in some societies markets are periodic, held at regular or irregular intervals, sometimes as one stop along regional circuits for peddlers who visit specific sites on a fixed cycle. (The tiny handful of ethnographic studies of markets listed in the bibliography give only a glimpse of the wide variety of anthropological analyses of markets, their internal structures, and their wider social and cultural contexts: e.g., Acheson, Bestor, Bohannan, Clark, Cohen, Geertz, Hertz, Meillassoux, Mintz, Polanyi, Plattner, Roseberry, Skinner, Smith, Trager.)

Many small-scale markets are socially embedded in communities, where producers and consumers deal face-to-face over the vegetables, chickens, or bolts of cloth that are the stuff of daily life, whether in a peasant community, an urban bazaar, or a farmer's market in a middle-class suburb. Local markets, as well as much more specialized ones such as urban wholesale markets of professional traders, are often organized around complex, multistranded relationships that intertwine gender, ethnicity, class, and kinship, as well as economic role.

Other very different kinds of markets (not marketplaces) embody diffuse, impersonal (and perhaps anonymous) ties among trade partners, such as in 'spot markets' where economic actors interact only through a one-time transaction, as in many real estate markets, labor markets, and global commodity markets for things such as sugar, coffee, or rubber.

Long-distance trade—both in exotic products and mundane commodities—may pass through highly specialized marketplaces that coordinate a regional or a global industry, such as Dutch flower auctions. Some long-distance markets are organized around small tightly knit communities of professional traders who transact business within networks of trust built on ethnic solidarity, such as New York's diamond exchanges or Ibadan's Hausa traders.

Markets exist along continua between the most informal sectors of society and the most highly regulated. Some markets are organized through informal or quasi-formal institutions (open and above board, but outside the realm of legal or political attention), while others are 'gray,' 'black, ' or entirely illegal. Other specialized markets for professional traders are organized within tightly regulated institutional frameworks that govern access, terms of exchange, reporting requirements, or public accountability; some examples include stock markets, commodity exchanges, and other financial markets. Whether informal or formal, the frameworks of regulation that encompass the smooth functioning of any market usually mix self-regulating mechanisms created by market participants themselves with those imposed by political or legal authorities. The social, institutional construction of trade and markets is evident in the widely varied price mechanisms—bartering, bidding, haggling, setting posted prices, or negotiating contracts, as well as discounts, rebates, or kickbacks—that are established in various markets, reflecting and shaping very different balances of market power among buyers and sellers.

3. *Markets as Institutions*

Anthropologists focus ethnographically on the social structure of markets as institutional systems, the transactional behavior of market participants, and networks among trade partners or among markets (Plattner 1985). They emphasize how economic activity is embedded in social institutions and relationships which structure solutions to economic problems (sometimes conceptualized as 'transaction costs' (Acheson 1994)). These costs, which any market or enterprise inevitably faces, are not only direct overhead expenses on specific exchanges. More generally, the economic and social costs of exchange include those of establishing trust and reliability among trade partners, soliciting or extending credit, guaranteeing stable sources of supply, enforcing compliance with agreements, recruiting labor, distributing profits, monitoring employees, obtaining information on market conditions, creating or enforcing property rights, managing risk, and so forth (Geertz 1978).

Various patterns of social structure that enable markets to form and economic transactions to occur are often conceptualized—by anthropologists influenced by institutional economics and sociology—in terms of 'governance structures,' the institutional structures that organize, constrain, and coordinate economic activities, that sanction some behaviors and provide incentives for others. Different governance structures—different forms of market relationships, different forms of business organization—provide different solutions to the challenges of achieving social and economic integration over the 'transaction costs' that all economic institutions must bear. Governance structures are, therefore, social institutions and systems of norms familiar to anthropologists in many other contexts and subject to similar kinds of social and cultural analyses.

Governance structures range along a theoretical continuum, from 'market governance' to 'governance by hierarchy.' In the former, an economic actor relies on the competitive forces of a spot market to obtain the goods, services, and trustworthiness it requires; in the latter, an economic actor controls goods, services, personnel, and reliability through direct ownership and administrative fiat (as in a vertically integrated industrial corporation). Midway between these extremes is governance by 'relational' or 'obligational' contracting, in which partners in ongoing exchange relationships agree, formally or informally, to do business with one another over time, relying on the strength of personal ties (such as trust) to overcome problems that may arise in the relationship because not all the terms and circumstances of trade are (or can be) specified ahead of time.

Anthropological analyses of markets often focus on the social and cultural patterns sustaining this middle form of governance, such as frameworks of self-regulation, the management of common property, the structural relationships between producers and buyers, the disposition of market power, or the political dynamics of trading communities. Other studies examine the creation of personal ties of trust and reciprocal obligation, and the microlevel transactional behavior among individual traders and other

market participants. Some studies place market relationships within a broader cultural milieu of interpersonal interactions; still others examine negotiating, bargaining, or haggling as a transactional game through which traders form what Plattner calls 'equilibrating relationships.' Implicit and explicit decision-making models used in daily trade have been collected ethnographically to illustrate how economizing strategies are embedded in culturally or socially specific contexts. Information costs and transactions costs have been analyzed within what Geertz (1978) calls 'bazaar economies.'

4. *Markets and Urban Life*

Throughout history, cities and markets have sustained each other, the former providing location, demand, and social context for the latter; the latter providing sustenance, profit, and cultural verve to the former. In many places, towns and marketplaces grew up together, with marketplaces as centers of economic and social life (and eventually political life as well) and as the institutions through which towns and cities were linked to their hinterlands and to other communities. Markets mediate connections and conflicts among very different segments of an economy or a society: across divisions between rural and urban, peasant and capitalist, premodern and modern, colonized and colonizing, or informal and formal sectors. These mediating roles have been examined in urban markets (in the diffuse sense, as in labor markets) and marketplaces (in the more specific sense, of geographically situated hierarchies of trade (Meillassoux 1971)). Within commercialized economies, distribution channels and markets that connect large-scale business to small-scale family firms are another example of market linkages across social and economic sectors.

Market hierarchies have themselves been a major topic of study. Central place theory analyzes the spatial distribution of markets within hierarchies of settlements ('central places'), and within anthropology has been applied to peasant marketing systems and to interrelationships among urban markets. Alignments of trading patterns within market systems have been shown to be important indicators of a wide variety of other social, political, administrative, and ritual aspects of local, regional, and national organization. Also known as 'regional analysis,' this approach was developed in anthropology by Skinner's ethnographic and historical research on China (Skinner 1977), and by extensive studies in Meso-America and elsewhere (Smith 1976).

The cultural environment of trade and marketplaces is also a central aspect of urban life. In the repertoire of crucial social relationships and roles filled by urbanites, Hannerz (1980) includes 'provisioning relationships,' which necessarily involve people in exchange, within and outside of markets. In a spatial sense, marketplaces can be analyzed as a distinctive kind of urban place with economic as well as many political, social, and ecclesiastical functions, but most anthropologists situate marketplaces not simply in spatial terms but within the wider cultural milieu of a society's values, norms, and texture of relationships (Geertz 1979).

Markets do not just organize sources of supply; they also satisfy (or create) demand and desire, as stages upon which consumption is rehearsed and displayed. Many studies of consumption take Bourdieu's perspectives on 'taste,' 'distinction,' and 'cultural capital' as points of departure for examining the cultural force of markets in shaping contemporary urban life. Studies of the logic of market capitalism and how it permeates one's experience of shopping and consumption echo Simmel's perspectives on market mentality as the quintessential condition of urban life. So too, did Redfield and Singer's earlier formulations of 'the cultural role of cities,' which placed the marketplace at the heart of the 'heterogenetic city,' a city that links itself to a wider world and, in the process, transforms the city, the rural hinterlands with which the city is interdependent, and society at large.

5. *Markets and Globalization*

More recently, sweeping transnational economic, political, and social forces have eroded the separability of societies and perhaps have disestablished the primacy of cities as nodes of exchange, but have accentuated the importance of markets. Since the early 1990s, proponents of globalization as a distinct type of social transformation have emphasized markets (in their deterritorialized sense) as central to this change. Globalization, in this view, is the expansion and ultra-integration of markets on an unprecedented scale, creating markets that incorporate societies, social sectors, and commodities that had formerly resisted or been outside market spheres of exchange. This process presumably has weakened nation-states, and has been facilitated by the increasing speed (or fluidity) of communications, transportation, media, and flows of goods, financial assets, and people, all sustained and accelerated by major technological breakthroughs in electronic media and information processing.

Of course, new patterns of global integration formed around markets are themselves nothing new. Anthropologists and sociologists have examined trade and market hierarchies as they establish linkages throughout what Immanuel Wallerstein conceptualized as the 'world system.' Within this expansion of Western European societies to incorporate most of the globe into their spheres of economic, political, and military hegemony, markets have been critical organizing principles for economic, social, political, and cultural phenomena on regional and national as well as global scales. The political economy of the contemporary world system can be seen through complex networks

of 'commodity chains': the links, stages, and hands through which a product passes as it is transformed, fabricated, and distributed between ultimate producers and ultimate consumers (Gereffi and Korzeniewicz 1994). Such chains connect far-flung components of the 'global factory,' the international division of labor among societies whose specialized niches in the world economy may center on resource extraction, low-cost fabrication, transportation efficiencies, or highly developed consumer markets.

Commodity chains are useful for understanding the widely dispersed industrial production characteristic of contemporary transnational trade as well as the fluidity of the circulation of agricultural and other commodities between producer societies on the global periphery and consumer societies of the global core. Sugar, for example, stands as a commodity at the intersection of imperialism, the colonization of the Caribbean as a region of plantation-slave societies, and Western European industrialization (Mintz 1985). The contemporary coffee trade in the Western Hemisphere reflects North American hyperconsumerism and the skillful marketing efforts of mass-market 'boutiques' for 'yuppie coffee,' which form new structural linkages between producer and consumer societies within systems of cultural symbolism and identity based on consumption style (Roseberry 1996).

Against a backdrop of globalizing markets for commodities, people, assets, and images, Appadurai (1996) proposes that contemporary ebbs and flows of transnational culture be conceptualized as 'ethnoscapes,' 'technoscapes,' 'finanscapes,' 'mediascapes,' and 'ideoscapes.' Very roughly, these refer to the complicated tides and undertows of people(s), of technology, of capital, of media representations, and of political ideologies that concurrently link and divide regions of the globe. These 'scapes' resemble diverse and intersecting markets that exchange items and images across the globe, and across domains, creating value and the possibility of exchange through unexpected juxtaposition and disjuncture. Appadurai's vision of global integration (or disintegration) implies a deterritorialized world in which *place* matters little, but the fluidity of exchange is everything. These loosely coupled domains are organized around particular processes (migration, investment, representation, or consumption), and a varied repertoire of influences may travel quickly, in many directions almost simultaneously, across these domains. The center or disseminator of influence on one 'scape' may be simultaneously the periphery or recipient of influence on another. This decentralized, deterritorialized global culture is a world of many markets (broadly conceived) but few marketplaces, few specific central places of interaction. In contrast, Hannerz (1996) locates these processes (or markets) in world cities, nodes that receive, coordinate, and disseminate cultural content and creativity to locally synthesize global elements into diverse, multiple patterns of regional culture.

The trends that created this 'global cultural supermarket' (to use Stuart Hall's phrase) involve markets more than just metaphorically. The commodification of human bodies in global flows of guest workers, sex workers, and refugees involves markets of hope, desire, and misery. Global industry's ability to shift production from place to place has created markets for 'pollution credits' among different jurisdictions. The development of deep-sea diving equipment has reinvigorated debates over ownership of the seabed, and has vastly extended property rights regimes over global oceans. Electronic media have created commercial forms—online auction sites, for example—that raise familiar questions about trade partnerships, rules of trade, and principles of reliability. Global financial markets can now operate 24 hours a day, further transcending spatial and temporal (and hence political) limitations on their operations. The digitization of information into electronic form has created new forms of property rights (patents and copyrights) based on the value of—and the exchange rights inherent in—retrieval systems rather than underlying information. Common cultural heritage has been converted to 'content,' a commodity with which to fill information delivery systems. Biotechnology has created new crop species, and hence questions about the ownership of both new and existing species and the digitized information on human and other genomes as they are decoded. Increasing exploitation of natural resources raises questions about the common property rights of nations and communities over ecosystems, as well as about the ownership and marketability of indigenous knowledge about local biota, itself a source of valuable data for bioprospecting.

The fundamental issues of anthropological interest in markets and exchange (what is property? who can own it? how can it be exchanged?) are not issues simply of trying to understand small-scale, isolated societies in antiquity and the recent past. How markets are constituted, who has access to them, and how they affect the social order as a whole continue to be current issues, affected by the transformations now taking place on a global scale, creating new integrations of local and transnational market systems centered around new forms of property and new media of exchange. Anthropological analyses of these markets address cultural and social issues as fundamental as those raised by analyses of traditional patterns of exchange in peasant marketing systems. Anthropological interest in markets will continue to focus on emerging practices of capitalism as a global cultural and social system.

Bibliography

Acheson J (ed.) 1994 *Anthropology and Institutional Economics*. Monographs in Economic Anthropology, no. 12, Society for

Economic Anthropology. University Press of America, Lanham, MD
Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN
Bestor T C 1999 Wholesale sushi: Culture and commodity in Tokyo's Tsukiji Market. In: Low S M (ed.) *Theorizing the City: The New Urban Anthropology Reader*. Rutgers University Press, New Brunswick, NJ
Bestor T C 2001 *Tokyo's Marketplace: Culture and Trade in the Tsukiji Wholesale Fish Market*. University of California Press, Berkeley, CA
Bohannan P 1955 Some principles of exchange and investment among the Tiv. *American Anthropologist* **57**: 60–70
Clark G 1994 *Onions Are My Husband: Survival and Accumulation by West African Market Women*. University of Chicago Press, Chicago
Cohen A 1969 *Custom and Politics in Urban Africa*. University of California Press, Berkeley, CA
Geertz C 1978 Bazaar economy: Information and search in peasant marketing. *American Economic Review* **68**(2): 28–32
Geertz C 1979 Suq: The bazaar economy of Sefrou. In: Geertz C, Geertz H, Rosen L (eds.) *Meaning and Order in Moroccan Society*. Cambridge University Press, New York, pp. 123–314
Gereffi G, Korzeniewicz M (eds.) 1994 *Commodity Chains and Global Capitalism*. Greenwood Press, Westport, CT
Granovetter M 1985 Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* **91**(3): 481–510
Hannerz U 1980 *Exploring the City: Inquiries Toward an Urban Anthropology*. Columbia University Press, New York
Hannerz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London
Hertz E 1998 *The Trading Crowd: An Ethnography of the Shanghai Stock Market*. Cambridge University Press, New York
Meillassoux C (ed.) 1971 *The Development of Indigenous Trade and Markets in West Africa*. Oxford University Press, London
Mintz S 1961 Pratik: Haitian personal economic relationships. *Proc. Annual Meeting American Ethnological Society*
Mintz S 1985 *Sweetness and Power: The Place of Sugar in Modern History*. Viking, New York
Plattner S (ed.) 1985 *Markets and Marketing*. Monographs in Economic Anthropology, no. 4. Society for Economic Anthropology, University Press of America, Lanham, MD
Plattner S (ed.) 1989 *Economic Anthropology*. Stanford University Press, Stanford, CA
Polanyi K, Arensberg C W, Pearson H W (eds.) 1957 *Trade and Markets in the Early Empires: Economics in History and Theory*. The Free Press, Glencoe, IL
Roseberry W 1996 The rise of yuppie coffees and the reimagination of class in the United States. *American Anthropologist* **98**(4): 762–75
Sahlins M 1972 *Stone Age Economics*. Aldine-Atherton, Chicago
Skinner G W (ed.) 1977 *The City in Late Imperial China*. Stanford University Press, Stanford, CA
Smith C A (ed.) 1976 *Regional Analysis*. Academic Press, New York, Vols. 1 & 2
Trager L 1981 Customers and creditors: Variations in economic personalism in a Nigerian marketing system. *Ethnology* **20**(2): 133–46

T. C. Bestor

Markets: Artistic and Cultural

A market is a forum for exchange between buyers and sellers. The concept has been broadened in its application to the artistic and cultural realms to refer to systems for the production and distribution of the arts. These systems differ from a purely atomistic market in that actors (buyers, sellers, and intermediaries) can be organizations as well as individuals. When the art form is produced by profit-seeking firms, the system is known as a culture industry (see *Culture, Production of*). The term marketization is used to indicate that a type of culture once produced and distributed in nonmarket settings has moved into market (or market-approximating) ones. Broadly, the term also indicates that a business-world way of thinking, referred to in the UK as 'enterprise culture,' has pervaded settings that used to be 'pure.' Key concepts involve the effects on art and culture of uncertainty, the laws of supply and demand, and the responsiveness of markets to buyers' and sellers' desires.

1. *History*

1.1 *Markets for the Fine Arts*

The classic study of market effects on artistic and cultural products, by economists Baumol and Bowen (1966), focused on the performing arts. Their much cited finding is that over time the performing arts become increasingly expensive relative to the cost of living, because the number of personnel needed for performances—the musicians in an orchestra, say, or actors in a Shakespeare play—is difficult to reduce. As a result, the performing arts do not see the productivity gains realizable in manufacturing industries. Performing arts, like other service industries, face increasingly expensive payroll bills. Thus, the performing arts always struggle to find external funding, charge higher ticket prices, or suffer cuts.

At about the same time, White and White (1965/1993) published their sociological study of the transformation of the French art world from an academic system to a dealer-critic system, which occurred at the end of the nineteenth century. Successful for more than two centuries, the former system was rigidly controlled by academics. It failed when the number of aspiring artists in Paris vastly outnumbered the rewards (prizes and academic appointments) available through the academy. The surplus artists wanted an outlet for their work and a means to make a living, so they turned to dealers, people who would sell their works directly to the public. Critics were a key component in this new system, as they worked with dealers and artists to create an aesthetic system that

justified the new forms of painting, and, through publishing, they conveyed these understandings to the public.

White and White demonstrate that the system through which French painting was produced influenced the content of the artworks themselves. The academic system rewarded individual paintings exhibited in salons. In judging work, academics adhered to a strict code. As a result, artists spent several years working on one painting, and although there was some room for creativity, they needed to produce work well within the accepted academic style to win a prize, or even to be invited to show their work. The dealer-critic system, on the other hand, had different requirements. Dealers needed to sell a continuing stream of paintings and, as a result, had an interest in differentiating their artists' products by referring to each artist's unique style. Paintings were purchased by middle-class buyers, as well as more wealthy patrons, and so were more likely to be hung in houses than in museums. This meant that small, easel paintings were preferred to monumental works, and that decorative scenes of landscape and everyday life were preferred to more complicated allegorical or historical works. Each of these requirements, along with a number of others, dovetailed with the new movement in painting, Impressionism. Though clearly there is more to artistic change than just the particular mechanisms that reward artists, White and White demonstrate the influence that these institutional arrangements had on aesthetic possibilities.

The studies by Baumol and Bowen, and White and White were important in setting the stage for serious study of arts markets. Since their studies were published, there has been a convergence of interest in artistic and cultural markets in a variety of disciplines. The main line of agreement among these scholars is that, in contemporary Western societies, the arts are profoundly shaped by their market characteristics.

1.2 Markets for Popular Arts

In an influential article, Hirsch (1972) looked at the distribution of popular arts by profit-seeking organizations in what he termed culture industries. Focusing on popular books and records, he demonstrated that cultural industries are characterized by uncertainty. Record companies and publishers wish to make a profit, but they are not sure which books and albums will succeed. They respond to this situation in a variety of ways, most notably through the strategy of overproduction. They produce many more albums and books than they can possibly sell, with the full knowledge that most will not break even. A small proportion will earn a modest return, and a very few will be phenomenally successful. These last, blockbusters, will generate enough income to cover the losses on the failures and to make a profit for the company. A second strategy taken by firms is to look at what has been recently successful and to overproduce cultural objects in the same genre or format. In this way, change in cultural industries is trendy, following fads and fashion.

A key concept in Hirsch's work is the idea of a gatekeeper, a person or organization who allows only certain objects into (or out of) the system. Gatekeepers are important for the content of artwork, as their guesses on what will succeed, their personal tastes, and even their unquestioned cultural assumptions can affect what they accept and reject (Griswold 1992). Gatekeepers do not directly mold artworks, but in deciding whether or not to produce or distribute an artwork, they determine the content of artwork that *actually reaches the public*.

2. Contemporary Research

Early approaches to artistic and cultural markets made strict distinctions between high and popular culture. High culture, such as opera, theatre, symphony, and the visual arts, was distributed by nonprofit organizations, and was oriented toward the artist. Popular culture, by contrast, was distributed by profit-making organizations and was oriented toward the consumer. Changes in society and in our understanding of culture have blurred these distinctions. Contemporary research may focus on a particular cultural form (high or popular), but is likely to use general tools for understanding organizations and markets. The laws of supply and demand apply regardless of whether the organizations involved are profit seeking. Characteristics of the production system—levels of uncertainty, ambiguity, and risk; market concentration and the level of competition; the ratios of buyers, sellers, and intermediaries—shape not only factors such as the price of an object or performance, but also the content of the art itself.

Becker (1982) suggests that distribution systems matter, especially to artistic *reputation*. What is not distributed cannot gain recognition nor be remembered. He identifies three key mechanisms by which artwork is supported. The first, self-support, where an artist relies on a day job, a spouse, or an inheritance for monetary needs and produces art 'on the side,' does not contain a market element. As Becker notes, self-supported art is often not distributed at all, let alone widely. The other two, patronage and public sale, are of more relevance here. Patronage, whereby a buyer commissions art, either a specific work or a part of the artist's general output—has a clear market component. Patronage sounds genteel, and requires that only one buyer be pleased. It is a form of support that can foster innovation—if the buyer values this—but clearly only within the bounds of the buyer's taste and inclinations.

Becker discusses three forms of public sale of

art. Dealers act as intermediaries between artists and buyers where the artwork is a physical object (a painting, sculpture, or installation). Impresarios mediate between artists and audiences in the performing arts (theatre, symphony, dance, or experimental performance art). Culture industries are the series of organizations that link artists who produce works that exist in multiple copies (books, records, movies, film, television shows—referred to in some literature as 'simulacra') to a large number of consumers.

The potential scope for research in the area of artistic and cultural markets is vast. It encompasses such areas as the international market for visual arts; auctions of contemporary works of art and of works by old masters; the funding dynamics of such arts organizations as museums or symphonies; the Hollywood film production and distribution network; film production elsewhere in the world, such as India's Bollywood; book publishing; and television broadcasting, and the international flow of programming. The list of high and popular art forms could continue. Tourism, visits both to heritage sites and to entertainment centers such as Disneyland, can be considered part of the global cultural market. Further, many consumer products, from soft drinks to jeans, are items of popular culture, and it would be easy to include an array of consumer markets as part of the cultural marketplace. Indeed, some theorists of consumerism would argue that all forms of contemporary purchasing belongs to the cultural realm.

The remainder of this article highlights the characteristics of markets for visual art, and sketches their implications for artists, art lovers, and art institutions. Most of the effects of the market on visual arts have complements in other artistic and cultural markets, although specific details will naturally vary from one art form to another.

3. *Visual Art Markets*

In the visual arts, separate markets exist for consecrated works (old masters and more recent artists with strong reputations), the contemporary avant-garde, art made by anonymous masters from developing countries (so-called primitive art), and contemporary decorative/figurative art. An international market exists for each of these types of art, except the last, which is the most varied and is sold by dealers in national, regional, or local markets, or even in such outlets as craft fairs and street stalls.

The market for consecrated works, unlike most artistic markets, is characterized by undersupply—there are fewer works of art than buyers wanting them. As a result, such works are often sold at auction, where competing buyers can bid up prices to extraordinary heights. But prices can differ substantially from the estimates of the auction houses, because social expectations and the dynamics of the auction floor affect bids along with the laws of supply and demand (Smith 1989). These factors add a degree of uncertainty to the auction market.

Consecrated works are almost always pieces that are being 'resold' by an owner, rather than new works put up by artists. This means that such sales have little direct effect on artists themselves. Clearly, this is the case for masters long dead, and is usually the case as well for artists still living. The European Union, however, has adopted regulations that require a proportion of resale prices to be returned to living artists. The UK is resisting this legislation, arguing that it will put the London auction houses at a disadvantage to those elsewhere, especially in New York, as paying artists makes the resale of art more expensive or less profitable.

Uncertainty is heightened in the avant-garde setting, both for artists and dealers. The prestigious avant-garde markets are concentrated in a few world-class cities, notably New York, London, and Paris. The problem for dealers—the same faced by cultural industry executives—is to choose artists who will succeed. There are huge numbers of artists hoping to be discovered and shown, and it is difficult to predict whom posterity will remember. The solution is to promote a relatively large number of artists as compared with the numbers who will succeed. Of course, aspiring artists may not experience the dealers' strategy as overproducing! Relatively few artists break through even to the level of success indicated by representation by an important dealer. In addition, dealers promote artists who work in styles similar to those that have recently been successful. Unlike cultural industry executives, they do not profess this mimicking strategy, as it goes against central tenets of avant-garde art, namely the sanctity of artistic choice, free from the interference of dealers.

The oversupply of artists leads to suppressed prices for art in general, but clearly dealers have an incentive to limit market entry and to keep prices up to a certain level by, for instance, discouraging their artists from underselling the dealer or trying to convince potential buyers that gallery prices for emerging artists are a fair expression of their future potential. As with other art markets, the avant-garde market is characterized by a 'winner-take-all' reward system where a few superstars receive the lion's share of the returns (see Rosen 1981). This means that a few contemporary artists will be phenomenally successful and become rich, a modest number will just scrape a living, and most will have to supplement their incomes elsewhere.

It is often argued that artists have their fingers on the pulse of society, so if art is ugly, unpleasant, or distressing, it is because it mirrors or foreshadows disturbing things in society. But given the oversupply of artists in the contemporary art market, artists must find a way to gain attention. One could make the case, then, that the sensationalism in some of contemporary art is more a response to the pressure of the art market

than a reflection of the wider cultural scene or the societal zeitgeist: bad-boy (or girl) art has a competitive advantage!

In Europe and the USA, figurative and decorative arts are devalued relative to avant-garde art. In avant-garde markets, the aesthetic system, prestigious dealers, and relative centralization provide entry barriers that do not exist in the decentralized figurative art market. As a result, prices are lower in this market. However, as Martorella (1990) demonstrates, corporations wishing to buy prestigious avant-garde art also prefer noncontroversial, figurative art. Their presence in the art market has helped legitimize some forms of figurative art as avant-garde art. US corporations' interest in collecting regional artists has also reinforced the decentralization in the figurative art market.

The market for 'primitive' art brings up issues such as the exploitation of artists from developing countries by dealers and patrons in the first world (Price 1989). Today, third-world masters are usually named and their work given the honorific title of art, rather than cast anonymously as an expression of tribal visions. Issues such as cultural imperialism are less salient in markets for avant-garde and figurative art (though certainly not nonexistent), but play a key role in markets for other types of culture, notably film and television.

Moulin (1987) argues that the economic value of an artwork and its artistic value are inseparable and synonymous. An economist would argue, then, that the market was efficient in that it properly prices works based on aesthetic value. But it is also possible to argue that the economic value, rather than reflecting aesthetic value, creates it by attracting the attention of art critics and the cultural cognoscenti, and setting in motion the theorizing necessary to underpin an artist's style.

As Menger (1999) demonstrates for a wide variety of artists in different cultural industries, more talent exists than do outlets for it. As a result, artistic careers are risky for artists. Most support themselves economically with a second job. The situation, as with other contingent labor markets (where employers hire workers only when work is available), provides a great deal of flexibility to cultural organizations such as dealers, but shifts costs from the organization to the individual.

4. *The Marketization of Art Museums*

Museums play a number of key roles in the visual art world; notably, they are the venue in which most ordinary viewers confront original art. Much income for US art museums comes in the form of grants for special projects, such as exhibitions and lectures, especially from corporations and government agencies. In addition, US museums increasingly rely on earned income through general admission charges, gate receipts for special shows, and through the sale of merchandise, books, and souvenirs. This funding situation has important implications for the art that is displayed in museums and for museum audiences (Alexander 1996).

Museums must now pay attention to audiences. Not only do audiences bring in money directly in terms of entrance fees and sales, they also influence project funding. Both corporations and government agencies are concerned with the size of the audience their funded projects might attract. This has led museums to operate, in essence, as public sales organizations, in that their exhibitions must appeal to the people who will 'buy' them. Museum curators remain committed to scholarly exhibitions and are less interested in popular ones, but the latter are easier to fund and the public likes them, so curators try hard to mount shows that are both scholarly and popular.

The decline of direct government support of cultural organizations is termed the privatization of culture. This has been a recent phenomenon in Europe as well as in the USA. For instance, the UK has reduced the amount it gives as direct grants to its national museums in an explicit attempt to make museums find a wider source of funding, most of which will be linked, directly or indirectly, to the marketplace.

An interesting change has taken place in US and UK museums over the last few decades: a shift in the basic assumptions of those in charge of museums, from a more scholarly approach to a more managerial one. Collecting, conservation, and connoisseurship remain mainstays of art museums. But a focus on business-world concerns—marketing a product, value for money, and the like—has pervaded museums. Some, particularly those with a business or economics background, see this in a positive light, as evidence that museums have entered the modern world and will finally be managed in a sensible way. Others, especially those vested in the older value system, see this as a corruption of museums, debasing the art and bringing crass considerations like the bottom line to a sacred arena where monetary concerns should be left aside.

5. *Future Directions*

5.1 *Can Culture Survive the Marketplace?*

An important question is how market systems, as opposed to other forms of support, affect culture. Given the pervasive oversupply of artworks and artists, a fundamental characteristic of market systems is that sellers must attract buyers. Art that no-one wants simply cannot be supported by the market. Moreover, markets imply profits. In this case, not only must buyers exist, there must be enough of them for sales to cover production and overhead costs, and to leave a surplus. The larger the audience needed, the

more popular the product must be and, as a corollary, the lower the common denominator. Indeed, as cultural critics for more than a century have lamented the degraded nature of mass culture, contemporary critics worry about the declining state of art. While there might be evidence for this, there is also evidence that even cultural industries can produce innovative material. Factors such as competition, low industry concentration, and decentralization tend to encourage a wider variety of culture than their obverse. In addition, new market techniques and technology, for instance 'narrowcasting' in television, lead to more segmented markets and, in theory at least, allow firms to earn money by catering to more varied tastes.

Not-for-profit enterprises may seem a good alternative. But the need to attract an audience is not limited to for-profit enterprises. Not-for-profit organizations must also reach sufficient audiences while facing an inexorable upward spiral of costs, as Baumol and Bowen showed in 1966. These pressures, along with the privatization of arts funding, lead non-profit enterprises to approximate profit-seeking systems in all but their requirement to return dividends to shareholders.

DiMaggio (1986) argues that large, established arts institutions can survive the marketplace. With proven repertoires of established works well accepted by the public, they can generate significant earned income. They may subsidize an innovative program with known crowd pleasers, for instance *The Nutcracker*, *The Messiah*, or Monet's *Water Lilies*. In addition to their ability to attract the 'middlebrow,' they have the organizational size and expertise to raise funds for such projects.

Since innovative, experimental works may be least likely to survive the marketplace, some have argued that there is a societal interest in subsidizing an oversupply of unpopular artists through government funding (see *Arts Funding*). While generous state sponsorship can have a positive effect on innovation and diversity in the arts, not to mention making the lives of artists easier, it is worth noting that results depend on national politics. For instance, art in Soviet societies was strongly supported by the state—and highly constrained. And political controversies in the USA served to narrow the type of art and artists sponsored by the federal government. Menger (1999) points out the important support given to artists by art colleges and university art departments, an alternative way to nurture talent that will not succeed in the marketplace.

5.2 New Technologies

A key point of uncertainty for the future of artistic and cultural markets is the influence of new technologies, especially the Internet. At the beginning of the twenty-first century, gatekeepers within cultural industry systems are powerful intermediaries—and barriers—between artists and the public. The Internet may change this. Artists have begun to reach out to audiences via the Internet. Some, notably in music, have had some success there, though most bands build a home page in hopes of attracting a recording company to sign them onto a big contract. It is easy to imagine this changing in important ways, but it is not clear exactly how. Perhaps as more people use the Internet (and rely on automated programs to sort information), more people will go to artists directly. Or perhaps through systems of micropayments, where users pay a royalty of a few pennies for use of copyrighted material on the Web, more people will download music, movies, and books from the Internet, eliminating distributors such as music stores, cinemas, and booksellers but not recording companies, film producers, or publishers.

Other cultural institutions, such as museums, are increasingly launching Web pages. It is unclear how these pages will affect their operation. Currently, it is difficult to imagine that being able to view a painting on the Web will have any impact on museums—except to the extent that people who see images on the Net might be intrigued enough to visit the museum in person. But as technology advances, it might become possible to reproduce museum objects in ways that have high fidelity to the original object. Will people then avoid museums, and view art and heritage through their screens? It seems unlikely, although it is worth pointing out that many consumers today prefer listening to orchestral music on compact discs rather than going to hear a live symphony. If the same becomes true for seeing physical objects, museums might suffer. Performing arts organizations might do well in this future scenario, as people might start 'attending' concerts in higher numbers if they could experience virtual 'live' performances in their own living rooms whenever they wanted. Visual artists might change the way they work and sell, producing an object to be digitally scanned, and earning a microroyalty each time someone views it. The implications for arts markets, not to mention artists, artworks, and audiences, are enormous.

See also: Art and Culture, Economics of; Art, Sociology of; Entertainment; Globalization and World Culture; Leisure and Cultural Consumption

Bibliography

Alexander V D 1996 Pictures at an exhibition: Conflicting pressures in museums and the display of art. *American Journal of Sociology* **101**: 797–839

Boorsma P B, Van Hemel A, Van der Wielen N (eds.) 1999 *Privatization and Culture: Experiences in the Arts, Heritage and Cultural Industries in Europe*. Kluwer, Amsterdam

Baumol W J, Bowen W G 1966 *Performing Arts: The Economic Dilemma*. Twentieth Century Fund, New York

Becker H S 1982 *Art Worlds*. University of California Press, Berkeley, CA
Crane D 1992 *The Production of Culture: Media and the Urban Arts*. Sage, Newbury Park, CA
Crane D 1987 *The Transformation of the Avant-garde: The New York Art World, 1940–1985*. University of Chicago Press, Chicago
DiMaggio P J (ed.) 1986 *Nonprofit Enterprise in the Arts: Studies in Mission and Constraint*. Oxford University Press, New York
DiMaggio P J, Stenberg K 1985 Why do some theatres innovate more than others? *Poetics: Journal of Empirical Research on Literature, the Media and Arts* **14**: 107–22
Frey B S, Pommerehne W W 1989 *Muses and Markets: Explorations in the Economics of the Arts*. Blackwell, Cambridge, MA
Greenfeld L 1988 Professional ideologies and patterns of 'gatekeeping': Evaluation and judgment within two art worlds. *Social Forces* **66**: 903–25
Griswold W 1992 The writing on the mud wall: Nigerian novels and the imaginary village. *American Sociological Review* **57**: 709–24
Heilbrun J, Gray C M 1993 *The Economics of Art and Culture: An American Perspective*. Cambridge University Press, New York
Hirsch P M 1972 Processing fads and fashions: An organization-set analysis of cultural industry systems. *American Journal of Sociology* **77**: 639–59
Martorella R 1990 *Corporate Art*. Rutgers University Press, New Brunswick, NJ
Menger P M 1999 Artistic labor markets and careers. *Annual Review of Sociology* **25**: 541–74
Moulin R 1987 *The French Art Market: A Sociological View*, trans. Goldhammer A. Rutgers University Press, New Brunswick, NJ (abridged trans. of 1967 *Le Marché de la Peinture en France*. Les Editions de Minuit, Paris)
Moulin R 1992 *L'Artiste, l'institution, et le marché*. Flammarion, Paris
Price S 1989 *Primitive Art in Civilized Places*. University of Chicago Press, Chicago
Rosen S 1981 The economics of superstars. *American Economic Review* **71**: 845–58
Smith C W 1989 *Auctions: The Social Construction of Value*. University of California Press, Berkeley, CA
White H C, White C A 1965/1993 *Canvasses and Careers: Institutional Change in the French Painting World*. University of Chicago Press, Chicago

V. D. Alexander

Markov Chain Monte Carlo Methods

Markov chain Monte Carlo (MCMC) methods use computer simulation of Markov chains in the parameter space. The Markov chains are defined in such a way that the posterior distribution in the given statistical inference problem is the asymptotic distribution. This allows to use ergodic averages to approximate the desired posterior expectations. Several standard approaches to define such Markov chains exist, including Gibbs sampling, Metropolis–Hastings and reversible jump. Using these algorithms it is possible to implement posterior simulation in essentially any problem which allow pointwise evaluation of the prior distribution and likelihood function.

1. Introduction

In Bayesian statistics the posterior distribution $p(\psi|y)$ contains all relevant information on the unknown parameters ψ given the observed data y. All statistical inference can be deduced from the posterior distribution by reporting appropriate summaries. This typically takes the form of evaluating integrals

$$J = \int f(\psi)\, p(\psi|y)\, d\psi \tag{1}$$

of some function $f(\psi)$ with respect to the posterior distribution. For example, point estimates for unknown parameters are given by the posterior means, that is, $f(\psi) = \psi$; prediction for future data $\tilde{y}$ is based on the posterior predictive distribution $p(\tilde{y}|y) = \int p(\tilde{y}|\psi, y)\, p(\psi|y)\, d\psi$, that is, $f(\psi) = p(\tilde{y}|\psi, y)$, etc. The problem is that these integrals are usually impossible to evaluate analytically. And when the parameter is multidimensional, even numerical methods may fail.

Since the 1980s a barrage of literature has appeared concerned with the evaluation of such integrals by methods collectively known as Markov chain Monte Carlo (MCMC) simulation. The underlying rationale of MCMC is to set up a Markov chain in ψ with ergodic distribution $p(\psi|y)$. Starting with some initial state $\psi^{(0)}$ we simulate M transitions under this Markov chain and record the simulated states $\psi^{(j)}$, $j = 1, \ldots, M$. The ergodic sample average

$$\hat{J} = \frac{1}{M} \sum_{j=1}^{M} f(\psi^{(j)}) \tag{2}$$

converges to the desired integral J (subject to some technical conditions), that is, $\hat{J}$ provides an approximate evaluation. The art of MCMC is to set up a suitable Markov chain with the desired posterior as stationary distribution and to judge when to stop simulation, that is, to diagnose when the chain has practically converged.

In many standard problems it turns out to be surprisingly easy to define a Markov chain with the desired stationary distribution. We will review the most important approaches in this article (for general principle of Monte Carlo simulation, including independent Monte Carlo simulation, see *Monte Carlo Methods and Bayesian Computation: Overview*).

2. The Gibbs Sampler

Consider a variance components model $y_{ij} = \theta_i + e_{ij}$, $i = 1, \ldots, K$ and $j = 1, \ldots, J$, for data y_{ij} from K groups with J observations in each group. Assume independent normal errors $e_{ij} \sim N(0, \sigma_e^2)$ and a normal random effects model $\theta_i \sim N(\mu, \sigma_\theta^2)$ (Gelfand et al. 1990). We assume that $\theta = (\theta_1, \ldots, \theta_k)$, (μ, σ_θ^2), and σ_e^2 are *a priori* independent with $p(\sigma_\theta^2) = IG(a_1, b_1)$, $p(\mu|\sigma_\theta^2) = N(\mu_0, \sigma_\theta^2)$, and $p(\sigma_e^2) = IG(a_2, b_2)$. Here we use $N(m, s^2)$ to indicate a normal distribution with moments m, s, and $IG(a, b)$ to indicate an inverse gamma distribution with parameters a and b. Let $y = (y_{ij}, i = 1, \ldots, K, j = 1, \ldots, J)$ denote the data vector. It can be shown that the conditional posterior distributions $p(\sigma_\theta^2|y, \mu, \theta, \sigma_e^2)$ and $p(\sigma_e^2|y, \mu, \theta, \sigma_\theta^2)$ are inverse gamma distributions, and $p(\mu|y, \theta, \sigma_\theta^2, \sigma_e^2)$, and $p(\theta|y, \mu, \theta, \sigma_\theta^2, \sigma_e^2)$ are normal distributions.

To estimate posterior moments of the type (1) we define a Markov chain in $\psi = (\mu, \theta, \sigma_e^2, \sigma_\theta^2)$. Denote with $\psi^{(t)} = (\mu^{(t)}, \theta^{(t)}, \sigma_e^{2(t)}, \sigma_\theta^{2(t)})$ the state vector of the Markov chain after t transitions. Given the nature of a Markov chain, all we need to define is the transition probability, that is, given a current value for $\psi^{(t)}$, we need to generate a new value $\psi^{(t+1)}$. We do so by sampling from the complete conditional posterior distributions for μ, σ_e^2, σ_θ^2 and θ

(a) $\mu^{(t+1)} \sim p(\mu|y, \theta^{(t)}, \sigma_e^{2(t)}, \sigma_\theta^{2(t)})$

(b) $\theta^{(t+1)} \sim p(\theta|y, \mu^{(t+1)}, \sigma_e^{2(t)}, \sigma_\theta^{2(t)})$

(c) $\sigma_e^{2(t+1)} \sim p(\sigma_e^2|y, \mu^{(t+1)}, \theta^{(t+1)}, \sigma_\theta^{2(t)})$

(d) $\sigma_\theta^{2(t+1)} \sim p(\sigma_\theta^2|y, \mu^{(t+1)}, \theta^{(t+1)}, \sigma_e^{2(t+1)})$

Steps (a) through (d) define a Markov chain $\psi^{(t)}$ which converges to $p(\mu, \theta, \sigma_e^2, \sigma_\theta^2|y)$, as desired. Ergodic averages of the type $\hat{J} = 1/M \sum f(\psi^{(t)})$ provide numerical evaluations of any desired posterior integral J.

The described Markov chain Monte Carlo simulation is a special case of a Gibbs sampler. In general, let $\psi = (\psi_1, \ldots, \psi_p)$ denote the parameter vector. The Gibbs sampler proceeds by iteratively, for $j = 1, \ldots, p$, generating from the conditional posterior distributions

$$\psi_j^{(t+1)} \sim p(\psi_j|\psi_1^{(t+1)}, \ldots, \psi_{j-1}^{(t+1)}, \psi_{j+1}^{(t)}, \ldots, \psi_p^{(t)}, y) \quad (3)$$

If practicable it is advisable to generate from higher dimensional conditionals.

Figure 1 illustrates the Gibbs sampling algorithm. The figure shows simulated parameter values for a hypothetical bivariate posterior distribution $p(\psi_1, \psi_2|y)$.

The seminal paper by Gelfand and Smith (1990) and the companion paper by Gelfand et al. (1990) popularized the Gibbs sampler for posterior simulation in a wide class of important problems. Many earlier papers used essentially the same method in specific problems. For example, a special case of the Gibbs sampler occurs in problems with missing data. In many problems, the actually observed data y can be augmented by missing data z in such a way that simulation from $p(\psi|y, z)$ and $p(z|\psi, y)$ can be implemented in computationally efficient ways, even when simulation from the original posterior distribution $p(\psi|y)$ is difficult. Tanner and Wing (1987) propose what is essentially a Gibbs sampler for the augmented posterior distribution $p(\psi, z|y)$. Geman and Geman (1984) proposed the Gibbs sampler for posterior simulation in a spatial model with a Markov random field prior.

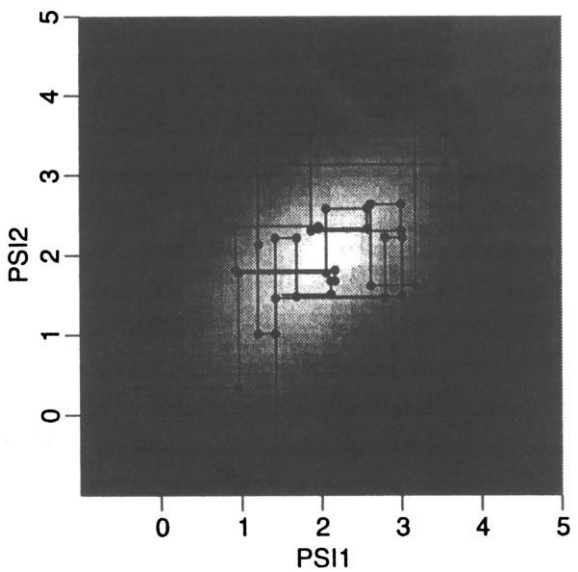

Figure 1
Gibbs sampler. The grey shades show a bivariate posterior distribution $p(\psi_1, \psi_2|y)$. The connected points show the parameter values $\psi^{(t)}$ generated in $M = 40$ transitions of the MCMC simulation. The transition probabilities in the Gibbs sampler are the full conditional posterior distributions (3), leading to the piecewise horizontal and vertical trajectories seen in the figure. Each horizontal line segment corresponds to generating a new value $\psi_1^{(t+1)} \sim p(\psi_1|y, \psi_2^{(t)})$. Each vertical line segment corresponds to generating $\psi_2^{(t+1)} \sim p(\psi_2|y, \psi_1^{(t+1)})$

3. The Metropolis–Hastings Algorithm

The Gibbs sampler owes some of its success and popularity to the fact that in many statistical models the complete conditional posterior distributions $p(\psi_j|\psi_i, i \neq j, y)$ take the form of some well-known distributions, allowing efficient random variate gen-

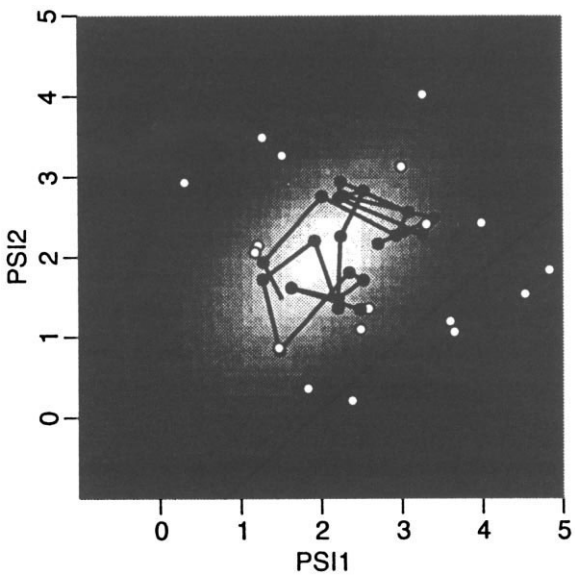

Figure 2
Metropolis sampler. The grey shades show a posterior distribution $p(\psi|y)$. The connected solid points show the parameter values $\psi^{(t)}$ generated in $M = 40$ transitions of a Metropolis chain with bivariate normal proposals $\tilde{\psi} \sim N(\psi, 0.75I)$, where I denotes the 2×2 unit matrix. The empty circles show generated proposals $\tilde{\psi}$ which were rejected using the acceptance probabilities (4). Compare with Fig. 1 which shows a Gibbs sampler with an equal number of transitions

eration. But there remain many important applications where this is not the case, requiring alternative MCMC schemes. Possibly the most generic such scheme is the Metropolis scheme (Metropolis et al. 1953). Consider generating from a posterior distribution $p(\psi|y)$. Denote with ψ the current state of the Markov chain. One transition is defined by the following steps:

(a) Generate a proposal $\tilde{\psi}$ from some proposal generating distribution $q(\tilde{\psi}|\psi)$. The choice of the proposal distribution $q(\cdot)$ is discussed below.

(b) Compute

$$a(\psi, \tilde{\psi}) = \min\left\{1, \frac{p(\tilde{\psi}|y)}{p(\psi|y)} \cdot \frac{q(\psi|\tilde{\psi})}{q(\tilde{\psi}|\psi)}\right\} \tag{4}$$

(c) With probability a replace ψ with the proposal $\tilde{\psi}$. Otherwise, leave ψ unchanged.

Figure 2 illustrates the algorithm. The figure shows the proposals $\tilde{\psi}$ and the (accepted) states $\psi^{(t)}$ for the first 40 iterations of a Metropolis chain simulation for a hyptothetical bivariate posterior distribution. The choice of the proposal distribution $q(\tilde{\psi}|\psi)$ is essentially arbitrary, subject only to some technical constraints. Using a symmetric proposal distribution with $q(\tilde{\psi}|\psi) = q(\psi|\tilde{\psi})$, for example a normal centered at ψ, has the practical advantage that the ratio of proposal distributions $q(\psi|\tilde{\psi})/q(\tilde{\psi}|\psi)$ cancels out of the expression for $a(\cdot)$. Often *Metropolis chain* is used to refer to this special case only. Another practically interesting variation is the use of an independent probing distribution $q(\tilde{\psi})$, that is, the proposal is independent of the current state. Tierney (1994) refers to such algorithms as *independence chains*. Hastings (1970) proposes a larger class of similar algorithms based on a more general expression for the acceptance probability. Chib and Greenberg (1995) give a tutorial introduction to the Metropolis–Hastings algorithm.

4. *Convergence*

The use of integral estimate (2) requires the verification of two conditions related to convergence.

First, the chain has to theoretically, that is, for $M \to \infty$, converge to the desired posterior distribution. Second, even if convergence for $M \to \infty$ is established, we need a convergence diagnostic to decide when we can terminate simulations in a practical implementation.

Tierney (1994, Theorem 1) shows convergence (in total variation norm) under three conditions: irreducibility, aperiodicity, and invariance.

The Markov chains which are used in MCMC schemes generally use a continuous state space, that is, $\psi^{(t)}$ is a real valued vector. For such continuous state spaces the notion of irreducibility is formally defined as π-irreducibility, with respect to some measure π on the state space. For the purpose of the present discussion we only consider $\pi(\psi) = p(\psi|y)$, that is, π denotes the desired stationary distribution. A Markov chain is π-irreducible if for any state ψ and any set B of states with $\pi(B) > 0$ there exists an integer $n \geqslant 1$ such that in n iterations the chain can with positive probability make a transition from ψ to some state in B.

Invariance refers to the property that if we start with a state vector generated from the desired posterior distribution, that is, $\psi^{(t)} \sim \pi$, then a further transition in the Markov chain leaves the marginal sampling distribution of ψ unchanged, that is, $\psi^{(t+1)} \sim \pi$.

The Gibbs sampler and the Metropolis–Hastings scheme define Markov chains which by construction are invariant with respect to the desired posterior distribution. Irreducibility and aperiodicity need to be verified, but are usually not a problem. However, sometimes MCMC implementations suffer from practical violations of irreducibility. There might be some subsets of the parameter space which are such that once the Markov chain simulation enters this set it is very unlikely to leave this subset again within any reasonable number of iterations. Such situations occur, for example, in independence chains if the

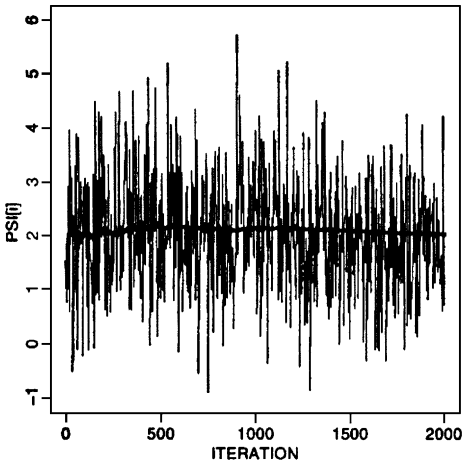

Figure 3
Convergence. The figure plots the first 2000 steps for the Gibbs sampler shown in Fig. 1. The thin black curve plots $\psi_1^{(t)}$ against iteration t. The thick grey curve plots $\hat{J}_t = 1/t \sum_{j=1}^{t} \psi_1^{(t)}$. After about 500 iterations the estimated posterior mean has practically converged. See the text for a discussion of more formal convergence diagnostics

proposal distribution $q(\tilde{\psi})$ has thinner tails than the desired posterior $\pi(\psi)$. The acceptance probabilities, Eqn. (4), include the ratios $\pi(\psi)/q(\psi)$. Assume the chain has generated a parameter value ψ far out in the tail, with very large ratio $\pi(\psi)/q(\psi)$. The chain will then reject any proposed move until a new proposal $\tilde{\psi}$ equally far out in the tail is generated.

Practically more important than establishing theoretical convergence is to recognize practical convergence, that is, to judge when sufficiently many transitions M have been simulated to obtain ergodic averages $\hat{J}$ close to the desired posterior expectations J. The simplest procedure is to plot the trajectories $\psi^{(t)}$ against iteration number t and judge convergence if an informal visual inspection of the plot does not reveal obvious trends. Figure 3 shows a typical trajectory. Several more formal convergence diagnostics have been proposed in the recent literature. Gelman and Rubin (1992) propose to consider several independent parallel runs of the MCMC simulation. Convergence is diagnosed if the differences of $\hat{J}$ across the parallel runs are within a reasonable range. Gelman and Rubin (1992) formalize this with an ANOVA type statistic. Geweke (1992) proposes to compare an ergodic average based on early simulations (say the first 10 percent of the iterations) with an ergodic average based on later iterations (say the last 50 percent). Under convergence the two ergodic averages should be approximately equal. Using an approximate sample standard deviation based on spectral density estimates allows a formal test.

These and other convergence diagnostics are discussed in Best et al. (1995) and implemented in the public domain software BOA described there.

5. *Limitations and Further Reading*

The Gibbs sampler and the Metropolis–Hastings chain implicitely require a fixed dimension parameter space, that is, the dimension of ψ must not change across different values. This excludes, for example, a regression model with an unknown number of co-variates. In other words, the Gibbs sampler or the Metropolis–Hastings algorithm can not be used for model selection.

Several recent monographs provide more complete reviews of MCMC methods. Tanner (1996) provides an introduction including related schemes such as importance sampling. Assuming basic familiarity with the algorithms, Gilks et al. (1996) discuss Markov chain Monte Carlo simulation in the context of important statistical models. Gamerman (1997) and Robert and Casella (1999) review alternative algorithms and related theory.

See also: Markov Models and Social Analysis; Monte Carlo Methods and Bayesian Computation: Overview

Bibliography

Best N G, Cowles M K, Vines S K 1995 *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.3*. MRC Biostatistics Unit, Cambridge, UK

Carlin B P, Chib S 1995 Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society B* **57**: 473–84

Chib S, Greenberg E 1995 Understanding the Metropolis–Hastings algorithm. *The American Statistician* **49**: 327–35

Gamerman D 1997 *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, London

Gelfand A E, Hills S E, Racine–Poon A, Smith A F M 1990 Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**: 972–85

Gelfand A E, Smith A F M 1990 Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409

Gelman A, Rubin D B 1992 Inference from iterative simulation using multiple sequences. *Statistical Science* **7**: 475–73

Geman S, Geman D 1984 Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6): 721–40

Geweke J 1992 Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo J M, Berger J O, Dawid A P, Smith A F M (eds.) *Bayesian Statistics 4*. Oxford University Press, Oxford, UK, pp. 169–94

Gilks W R, Richardson S, Spiegelhalter D J 1996 *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London
Green P 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–32
Hastings W K 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109
Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H, Teller E 1953 Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–91
Robert C P, Casella G 1999 *Monte Carlo Statistical Methods*. Springer-Verlag, New York
Tanner M A 1996 *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd. edn. Springer-Verlag, New York
Tanner M A, Wing H W 1987 The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**: 528–50
Tierney L 1994 Markov chains for exploring posterior distributions. *Annals of Statistics* **22**: 1701–28

P. Müller

Markov Decision Processes

Markov decision processes (MDPs) model decision making in stochastic, sequential environments (see *Sequential Decision Making*). The essence of the model is that a decision maker, or agent (see *Autonomous Agents*), inhabits an environment, which changes state randomly in response to action choices made by the decision maker. The state of the environment affects the immediate reward obtained by the agent, as well as the probabilities of future state transitions. The agent's objective is to select actions to maximize a long-term measure of total reward. This article describes MDPs, an example application, algorithms for finding optimal policies in MDPs, and useful extensions to the basic model.

MDPs are attractive models of decision making and planning. Because of their more general reward and state-transition structure, MDPs extend standard planning models from artificial intelligence. Interesting environments can be expressed as MDPs, relatively efficient algorithms exist for computing optimal decision strategies. Using current computer technology, these algorithms scale to MDPs with on the order of a million states.

1. Animal Behavior Example

For a concrete example of an MDP model, consider this fictitious animal behavior example (see *Animal Cognition*). A flock of sheep grazes in a meadow. At any given moment, the meadow can have abundant food (s_1), adequate food (s_2), scarce food (s_3), or no food (s_4). Each day, the flock can decide to graze (a_1), search locally for another meadow (a_2), or search more globally for another meadow (a_3). The matrices in Tables 1 and 2 define the MDP. Search action a_2 has the advantage of being relatively cheap, but the disadvantage of not changing the state very much.

Table 1
Reward function and state transition function for the graze action in the fictitious sheep grazing example

$R(s, a)$	a_1	a_2	a_3	
s_1	5	−1	−5	
s_2	5	−1	−5	
s_3	2	−1	−5	
s_4	0	−1	−5	
$T(s, a_1, s')$	s_1	s_2	s_3	s_4
s_1	0.6	0.3	0.1	0.0
s_2	0.0	0.4	0.4	0.2
s_3	0.0	0.0	0.4	0.6
s_4	0.0	0.0	0.0	1.0

Table 2
State transition functions for the search actions in the fictitious sheep grazing example

	s_1	s_2	s_3	s_4
$T(s, a_2, s')$				
s_1	0.5	0.4	0.1	0.0
s_2	0.3	0.5	0.2	0.0
s_3	0.0	0.4	0.4	0.2
s_4	0.0	0.0	0.4	0.6
$T(s, a_3, s')$				
s_1	0.1	0.4	0.3	0.2
s_2	0.1	0.4	0.3	0.2
s_3	0.1	0.4	0.3	0.2
s_4	0.1	0.4	0.3	0.2

Roughly speaking, the sheep should graze when there is adequate food, and otherwise search for a better meadow. But, when should they search and what search strategy should they use? Sect. 2 gives a formal definition of MDPs and Sect. 3 describes how to solve them.

2. Definitions

A Markov decision process is a tuple $M = (S, A, T, R, \beta)$, where

S is a finite set of states of the environment;

A is a finite set of actions;

$T : S \times A \to \Pi(S)$ is the state-transition function, giving for each state and agent action, a probability distribution over states ($T(s, a, s')$ is the probability of ending in state s', given that the agent starts in state s and takes action a);

$R : S \times A \to \mathscr{R}$ is the reward function, giving the expected immediate reward gained by the agent for

taking each action in each state ($R(s, a)$ is the expected reward for taking action a in state s; and $0 < \beta < 1$ is a discount factor).

A policy is a description of the behavior of an agent. Because of their simple structure, this article considers only stationary policies, such as $\pi: S \rightarrow A$. These policies specify, for each state, an action to be taken should the agent find itself in that state.

The total expected discounted reward for a policy π from some initial state is a measure of its long-run success. It is defined as the expected value of the discounted sum of immediate rewards. To be more concrete, let $V^{\pi}(s)$ be the total expected discounted reward for starting in state s and executing policy π:

$$V^{\pi}(s) = R(s, \pi(s)) + \beta \sum_{s' \in S} T(s, \pi(s), s') V^{\pi}(s') \quad (1)$$

The value function for policy π is the unique solution of this set of simultaneous linear equations, one for each state s. The system of linear equations can be solved using any of a number of standard techniques from linear algebra.

Equation (1) derives a value function from a policy. Reversing this, the greedy policy with respect to value function V, π_v, is defined as

$$\pi_V(s) = \operatorname*{argmax}_a \left[R(s, a) + \beta \sum_{s' \in S} T(s, a, s') V(s') \right] \quad (2)$$

This policy is obtained by taking the action in each state with the best one-step lookahead value according to V.

Given an initial state s, the goal is to execute the policy π that maximizes $V^{\pi}(s)$. Howard (1960) showed that there exists a single policy π^* that is optimal for every starting state. The value function for this policy, written V^*, is defined by the set of equations

$$V^*(s) = \max_a \left[R(s, a) + \beta \sum_{s' \in S} T(s, a, s') V^*(s') \right] \quad (3)$$

and any greedy policy with respect to this value function is optimal. Thus, the goal of many MDP solution algorithms is finding V^* or a good approximation.

3. *Algorithms for Solving Markov Decision Processes*

Solving an MDP means finding the policy π^* that maximizes the total expected discounted reward. Although many approaches have been studied, the simplest and most general is value iteration.

Value iteration treats the equality in Eqn. (3) as an assignment. It starts off with an arbitrary approximation to V^*, say a value function V that assigns 0 to all states. Over successive iterations, this approximation is improved by using the previous value of V combined with one-step lookahead to get a more refined estimate. The algorithm terminates when the maximum difference between two successive value functions is less than some predetermined $\varepsilon > 0$, at which time the greedy policy with respect to V is returned. The algorithm is guaranteed to terminate, and, if ε is small enough, the returned policy will be optimal.

Table 3
Sequence of value functions generated by value iteration in the fictitious sheep grazing example

Iteration	s_1	s_2	s_3	s_4
0	0.000	0.000	0.000	0.000
1	5.000	5.000	2.000	0.000
2	9.230	7.520	2.720	0.000
3	12.259	8.686	2.979	0.000
4	14.234	9.200	3.200	0.073
…				
86	19.480	12.502	6.502	3.670

While easy to implement, value iteration can be very slow to converge in some MDPs. For additional computational cost per iteration, policy iteration, due to Howard (1960), results in considerably more rapid covergence in practice. Like value iteration, policy iteration computes a sequence of value functions, V. In policy iteration, V is the value function for the greedy policy with respect to the value function V' from the previous iteration: $V = V^{\pi_{V'}}$. This can be done by applying Eqn. (2) and then solving Eqn. (1).

Unlike value iteration, policy iteration finds the optimal value function and policy after a finite number of iterations. This is because there are a finite number of policies, and each iteration finds a policy that is a strict improvement over the previous one.

Returning to the animal behavior example from *Sect.* 1, value iteration produces the sequence of value function given in Table 3. From this, the optimal policy is to graze (a_1) in states s_1 and s_2, local search (a_2) in state s_3, and global search (a_3) in state s_4. Value iteration finds this policy after only four iterations, but convergence to the optimal value function takes considerably longer: For accuracy out to six digits past the decimal point, 170 iterations are required. In contrast, policy iteration steps through only one suboptimal value function (for the policy that always chooses a_1) before obtaining the optimal policy and value function.

As the discount factor β gets close to one, value iteration can require exponentially many iterations to converge. It is possible, although not proven, that policy iteration can also take an exponential number of iterations. The only approach to solving MDPs that has been shown to terminate after a polynomial amount of computation is linear programming (see

Algorithmic Complexity; *Linear and Nonlinear Programming*). Littman et al. (1995) survey computational complexity issues in MDPs.

4. Related Models and Extensions

While the type of MDP described in this article is very general and has been extensively studied, other closely related models are also quite useful.

Markov games. In this model, state transitions and rewards are controlled by multiple agents instead of a single agent. Agents select actions to maximize their personal total expected discounted reward. The zero-sum version of this model (see *Game Theory*), first introduced by Shapley (1953), can be solved via variations of value iteration and policy iteration.

Partially observable Markov decision processes. This model is a hybrid of MDPs with hidden Markov models in which the environment's state is only visible via stochastically generated, and possibly ambiguous, 'observations.' Smallwood and Sondik (1973) provide one of the first algorithmic treatments of this model. Both value iteration and policy iteration algorithms have been studied.

Continuous states, actions, time: Robotics problems, for example, feature continuous state spaces (e.g., arm angle) and action spaces (e.g., joint acceleration) and must run in real time (see *Dynamic Decision Making*). Puterman (1994) surveys some relevant results.

Other objective functions. The problems of maximizing average reward per step, total expected reward (undiscounted), and worst-case reward all share important features with the total expected discounted reward model described here, although each introduces its own complications.

Propositional representations. Many environments can be best described using states or actions in attribute-value form (e.g., health = good, age = young, season = winter, ...). Boutilier et al. (1999) survey prepositional representations of MDPs and recent attempts to take advantage of these representations to speed the solution process.

Value-function approximation. A recent and promising direction is the use of compact parameterized function approximators like neural networks to store value functions. Bertsekas and Tsitsiklis (1996) dubbed this approach 'neuro-dynamic programming' and it has been successfully applied to some very difficult sequential decision problems such as elevator control and move selection in board games.

Reinforcement learning. The MDP model has been embraced by reinforcement-learning researchers as a model of environments faced by agents learning from delayed feedback (Sutton and Barto 1998).

The MDP literature is quite substantial. The early work of Bellman (1957) and Howard (1960) put the field on a firm footing. Later work synthesized the existing material and added new results. Textbooks by Puterman (1994) and Bertsekas (1995) give in-depth summaries. Contributions have been made by researchers in operations research, electrical engineering, dynamic programming, complexity theory, economics, algorithm analysis, artificial intelligence, and the behavioral sciences (see Chap. 1 of Puterman (1994) for a summary). Existing theory and algorithms have been applied successfully to real-world problems in a wide array of environments and renewed attention in recent years has greatly expanded the range of problems to which this formalism can be applied.

See also: Action Planning, Psychology of; Artificial Intelligence: Uncertainty; Decision Research: Behavioral; Decision Theory: Bayesian; Decision Theory: Classical; Dynamic Decision Making; Multi-attribute Decision Making in Urban Studies; Optimal Control Theory; Perception and Action; Utility and Subjective Probability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies

Bibliography

Bellman R 1957 *Dynamic Programming*. Princeton University Press, Princeton, NJ

Bertsekas D P 1995 *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA, Vol. 1 and 2

Bertsekas D P, Tsitsiklis J N 1996 *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA

Boutilier C, Dean T, Hanks S 1999 Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research* **11**: 1–94

Howard R A 1960 *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA

Littman M L, Dean T L, Kaelbling L P 1995 On the complexity of solving Markov decision problems. In: *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence* (*UAI-95*). Morgan Kaufmann Inc., San Francisco, pp. 394–402

Puterman M L 1994 *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. Wiley, New York

Shapley L 1953 Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America* **39**: 1095–100

Smallwood R D, Sondik E J 1973 The optimal control of partially observable Markov processes over a finite horizon. *Operations Research* **21**: 1071–88

Sutton R S, Barto A G 1998 *Reinforcement learning: An Introduction*. MIT Press, Cambridge, MA

M. L. Littman

Markov Models and Social Analysis

There are a plethora of social science processes that are best analyzed through the use of some form of Markov model. In a study of consumer behavior,

Table 1
Brand switching transition probabilities

		Time t		
	p_{ij}	A	B	C
Time $(t-1)$	A	0.72	0.18	0.10
	B	0.09	0.85	0.06
	C	0.21	0.15	0.64

individuals may switch from one brand of a product to another brand. Note that a 'change' here includes staying with the current brand, as well as the possibility of returning to a previously used brand. This switching can be 'random' over time or the individual/customer may in time settle on a particular brand, or may even move away entirely from all brands to a different product. Likewise, voters may change their preferences for candidates from one party to another. Equivalently, workers may change jobs from one sector of the workforce to another; or the focus may be on intergenerational occupational patterns. Households, or individuals, may move across differing income levels, or across various social classes. Credit card holders may move in and out of debt levels.

In a different direction, understanding (and predicting) bed occupancy rates in a hospital can be modeled by Markov processes. Instead of the movement of patients through the stages of illness recovery, the question of interest may be movement up the company's hierarchical promotion ladder(s), or on progress through states (or lack thereof, as in recidivism) in rehabilitation programs (as in health, criminology, etc.). Friendship formation and more generally network analyses and their various manifestations also fall under the Markov model rubric. Bartholomew (1983), among others, provides a compact summary and references for a wide array of applications.

A common thread in all such processes is that at a particular time, t say, the individual under study (i.e., person, system, etc.) is characterized as being in one of a number of clearly identifiable states, s say, $s = 1, \ldots, S$ where S is the total number of states. Another thread is that the state currently occupied by an individual depends on the state occupied at one or more previous times; that is, there is an underlying dependence between the observations. The data record these states for a set of individuals at times $t_0, t_1, \ldots, t_T$ where $T \geqslant 1$ and where t_0 is the initial time. Time itself can be a discrete or continuous variable, leading to Markov chains or Markov processes, respectively (named after the Russian mathematician Andrei Andreevich Markov, 1856–1922, who developed fundamental probabilist results for dependent processes). When time is discrete, the data may typically be thought of as occurring at times $t = 0, 1, \ldots, T$ (with an implied assumption that changes from one time t to the next $(t+1)$ are not dependent on the length of time between such measurements; this assumption is not always valid, see Sect. 3.1 (see also *Longitudinal Data*).

A first-order Markov chain, or simply a Markov chain, is characterized by the probabilities of moving from state i at time $t-1$ to state j at time t, $p_{ij}(t)$, with

$$\begin{aligned} p_{ij}(t) &= P(\text{State } s = j \text{ at } t \,|\, \text{State } s = i \text{ at } t-1) \\ &= P(s = j \text{ at } t \,|\, s = s_0 \text{ at } t = 0, \ldots, s = s_{t-2} \\ &\qquad \text{at } t-2, s = i \text{ at } t-1) \end{aligned} \quad (1)$$

that is, the transition probability depends only on its most recent state and not on the states occupied at other previous times. More generally, an rth order Markov chain is one for which the transition probabilities depend on the r most recent states. Except where specified, this entry assumes a first-order chain as governed by (1). These transition probabilities may be independent of t with $p_{ij}(t) \equiv p_{ij}$ which is the stationary (time-homogeneous) case, or may be a function of t giving rise to nonstationary (time-inhomogeneous) Markov chains. (Note that, for some users, stationarity defines the steady state or equilibrium probabilities obtained after a process has settled down; see Sect. 2.) Homogeneity (or conversely, heterogeneity) refers typically to the fact that the transition probabilities are the same (or not the same) for all individuals in the study.

Continuous time models take two broad forms. The first is a finite state Markov chain based on Eqn. (1). The other form falls under the heading of Markov processes characterized by transition rates $\lambda_{ij}(t)$ of moving from state i at time t, to state j at the infinitesimal time later $t+\Delta t$, also independent of earlier transitions, i.e.,

$$P(s = j \text{ at } t+\Delta t \,|\, s = i \text{ at } t) = \lambda_{ij}(t) + o(\Delta t) \quad (2)$$

where $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$. As for the Markov chain, these rates can be homogeneous (i.e., time-independent with $\lambda_{ij}(t) = \lambda_{ij}$) or nonhomogeneous. In these cases, it is more usual to focus attention on estimating the numbers in each state at time t rather than the transition probabilities of a single individual moving from one state to another; see Sect. 3.2.

Section 1 provides four examples of Markov models from the social science literature. The attention then shifts to the basic models and some of their properties, along with aspects of parameter estimation described in Sects. 2 and 3 for discrete and continuous time models, respectively. Some extensions, such as mover–stayer models, mixed Markov models, semi-Markov models and interactive Markov chains, are also covered briefly in Sect. 4.

1. Some Examples

1.1 Brand Loyalty

Manufacturers, and equally retailers, will always be interested in whether or not the users of their products remain loyal to a particular product and/or switch to another product, and if so, what it is that entices the loyalty/switch that occurs. Is this the effect of advertising, product quality, peer pressure, or what? Draper and Nolin (1964), Lawrence (1966) and Chatfield (1973), among others, first studied this behavior as a Markov chain. Suppose there are three products ($S = 3$): Brand A ($s = 1$), Brand B ($s = 2$) and Brand C ($s = 3$). Suppose the transition probabilities p_{ij}, $i, j = 1, 2, 3$, are as shown in Table 1, where it is assumed implicitly that these probabilities are constant over time. These p_{ij}s represent the probability that an individual who bought one brand ($s = i$) at time ($t-1$) will buy another brand ($s = j$) the next time (t). For example, the probability that a person who bought Brand A and then bought Brand B the next time is $p_{12} = 0.18$. The diagonal elements p_{ss}, $s = 1, 2, 3$ represent product loyalty; e.g., the probability that an individual bought Brand C both last time and this time is $p_{33} = 0.64$. Typically, in practice, these transition probabilities are unknown, and so must be estimated from data. Estimation procedures are discussed in Sect. 2. Once these probabilities are known (or estimated), there are numerous questions of interest (e.g., what is the long-run brand loyalty over many time periods?) that can be answered; some of these will be presented in Sect. 2.

1.2 Occupational Mobility

Occupational mobility, or more broadly social mobility, can be modeled by a Markov chain. The data of Table 2 (adapted from Biblarz and Raftery 1993) display the probabilities associated with the son's (first) occupation and the occupation of the father. There are $S = 5$ possible occupations: self-employed and salaried professionals ($s = 1$); proprietors, clerical workers and retail sellers ($s = 2$); craftsmen in manufacturing, construction, and other industries ($s = 3$); service workers, operatives, and laborers ($s = 4$); and farm managers and workers ($s = 5$). Thus, p_{ii}, $i = 1, \ldots, 5$ represents the probability of a son following his father's footsteps by going into the same occupation; e.g., the probability a son follows his father into farming is $p_{55} = 0.39$. The p_{ij}, $i \neq j$, represent the probabilities that the son goes into a different occupation; e.g., the probability a son goes into a self-employed position while his father worked as a construction worker is $p_{21} = 0.32$. The Biblarz and Raftery (1993) study was actually concerned with whether occupational mobility varied across race and/or across types of family disruption. Thus, instead of Table 2, there is a corresponding table for each of the demographic categories considered. Although it is beyond the scope of this article, logistic regression methods were used subsequently to study what factors most influenced occupational mobility, with the transition probabilities p_{ij} being the starting point.

Table 2
Occupational mobility transition probabilities

Father's occupation	Son's occupation				
	1	2	3	4	5
1	0.48	0.18	0.11	0.22	0.01
2	0.32	0.24	0.11	0.31	0.02
3	0.19	0.16	0.21	0.42	0.02
4	0.13	0.15	0.13	0.55	0.04
5	0.09	0.08	0.09	0.35	0.39

In this example, changes in time t correspond to generational changes. Had the interest focused on occupational mobility of individuals over time (let us say, over decades), then data of the type of Table 2 pertain but where now, e.g., $p_{31} = 0.19$ is the probability that an individual working in construction has moved into a self-employed or salaried professional position one decade later.

Instead of occupational mobility, states may represent social classes so that social mobility questions can be studied. Or, in a study of economic mobility, states would represent levels of economic status, be these states representing financial assets (say), creditworthiness (say), income levels, or whatever economic factor is the focus. In all cases, the probabilities of movement from one state to another take the form of Table 2.

1.3 Voter Patterns

Anderson (1954), Goodman (1962), and Brown and Payne (1986), among numerous other authors, have used Markov chain models to study voting behavioral patterns. In Goodman's study, 455 potential voters were asked in six consecutive months which party's candidate would receive their vote, be that Republican (R, $s = 1$), Democrat (D, $s = 2$) or Don't know (U, $s = 3$). Table 3a gives the numbers of those respondents who indicated in May how they would vote, and again what those responses were in June. Thus, for example, of the 445 subjects interviewed, 11 of those who in May did not know for whom they would vote had by June decided to vote for the Republican candidate. Or, equivalently, from the (estimated) transition probabilities of Table 3b, 83 percent of those who in May planned to vote Democrat were still planning in June to vote for the Democrat candidate.

1.4 Malaria

Let us consider Cohen and Singer's (1979) study of malaria. Each individual was identified at time t as

Table 3
Voting patterns

		(a)					(b)		
		June					June		
	#'s	R	D	U	Totals	p_{ij}	R	D	U
May	R	125	5	16	146	R	0.86	0.03	0.11
	D	7	106	15	128	D	0.05	0.83	0.12
	U	11	18	142	171	U	0.06	0.11	0.83
Totals		143	129	173	445				

Table 4
Malaria disease states transition probabilities

		Time t			
	$\hat{p}_{ij}$	1	2	3	4
Time $(t-1)$	1	0.64	0.03	0.32	0.01
	2	0.66	0.06	0.28	0.00
	3	0.39	0.01	0.53	0.07
	4	0.38	0.05	0.47	0.10

being in one of $S = 4$ states, corresponding to uninfected ($s = 1$), singly infected with *Plasmodium malariae* ($s = 2$), singly infected with *Plasmodium falciparum* ($s = 3$), or jointly infected with both *Plasmodium* species ($s = 4$). Eight surveys were conducted at regular (10-week) intervals, i.e., $T = 7$. The study also separately considered the transitions into and out of the four disease states, for each of seven age groups. Thus, for each age group, there is a 4×4 transition probability matrix for each transition from time $t = t_u$ to $t = t_v > t_u$, of the type shown in Table 4 (extracted from Cohen and Singer 1979), where $t_u, t_v = 0, \ldots, 7$. It is assumed that infection can occur at any time t and not just at those times surveyed, thus making this a continuous time Markov chain (in contrast to the previous examples, all of which were discrete time chains). In a different direction, Cohen and Singer (1979) also studied the malaria epidemic as a Markov process with transitions governed by Eqn. (2).

2. *Discrete Time Models: Markov Chains*

2.1 *Stationary Markov Chains*

Let $\boldsymbol{P}$ be an $S \times S$ matrix of transition probabilities $p_{ij}(t)$ (defined in Eqn. (1)), and let $p_{ij}(t) = p_{ij}$, i.e., the Markov chain is stationary. Note that $\sum_{j=1}^{S} p_{ij} = 1$. Let $p_s^{(0)}$ be the probability that an individual is in the initial state s at $t = 0$, $s = 1, \ldots, S$ and write $\boldsymbol{p}^{(0)}$ as the $S \times 1$ vector whose elements are $p_s^{(0)}$. The matrix $\boldsymbol{P}$ is called a transition probability matrix (tpm) and $\boldsymbol{p}^{(0)}$ is the initial probability vector. The Markov chain is defined completely by $\boldsymbol{P}$ and $\boldsymbol{p}^{(0)}$ (see *Probability: Formal*). Some basic results and definitions include the following.

(a) The n-step-ahead probability matrix $\boldsymbol{P}^{(n)} = (p_{ij}^{(n)})$ $i, j = 1, \ldots, S$, and the n-step-ahead probability vector $\boldsymbol{p}^{(n)} = (p_i^{(n)})$ $i = 1, \ldots, S$, satisfy

$$\boldsymbol{P}^{(n)} = \boldsymbol{P}^n, \quad \boldsymbol{p}^{(n)} = \boldsymbol{p}^{(0)}\boldsymbol{P}^n \tag{3}$$

where $\boldsymbol{P}^n$ implies matrix multiplication of $\boldsymbol{P}$ to the nth power in the usual way; and where the elements of $\boldsymbol{P}^{(n)}$ and $\boldsymbol{p}^{(n)}$, respectively, are $p_{ij}^{(n)}$, which is the probability that an individual in state i at time t is in state j at time $(t+n)$, and $p_i^{(n)}$, which is the probability the individual is in state i after n times units.

To illustrate, Table 5 provides the transition probabilities in $\boldsymbol{P}^{(2)}$ representing the probabilities associated with mobility between a grandfather's and a grandson's occupation, and were obtained from the $\boldsymbol{P}$ of Table 2. Thus, for example, the probability that both the grandfather and grandson are in farming is 0.1704. Note, however, that this says nothing about the father's occupation. Indeed, this term was calculated from

$$p_{55}^2 = (.09)(.01) + (.08)(.02) + (.09)(.02) + (.35)(.04) + (.39)(.39) = .1704$$

that is, the $p_{55}^{(2)}$ takes into account all possible occupations that the father might have had. The probability that all three generations were farmers is $(0.39)^2 = 0.1521$.

(b) A state j is *accessible* from state i if and, only if, from some n, $p_{ij}^{(n)} > 0$. Note that i and j are not necessarily accessible in one step, but are accessible at some time $n > 0$. If no n exists for which $p_{ij}^{(n)} > 0$, then it is never possible to reach j from i. Two states, i and j, *communicate* with each other if each is accessible to the other.

(c) If all states $\{1, \ldots, S\}$ communicate with each other, the chain is said to be *irreducible*.

(d) If there exists an n such that $p_{ij}^{(n)} > 0$ for all i and j, then the chain is *ergodic*.

(e) For an irreducible aperiodic ergodic chain, there exists a stationary distribution $\boldsymbol{\pi}$ which satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$, with elements π_s representing the long-run proportion of times state s is visited and with $\sum \pi_s = 1$.

Table 5
Two-generational transition probabilities

Grandfather's occupation	Grandson's occupation				
	1	2	3	4	5
1	0.3384	0.1810	0.1252	0.3321	0.0233
2	0.2934	0.1809	0.1268	0.3685	0.0304
3	0.2387	0.1708	0.1390	0.4176	0.0339
4	0.2102	0.1659	0.1332	0.4462	0.0445
5	0.1665	0.1335	0.1182	0.4114	0.1704

To illustrate with the brand-switching data of Table 1, it follows that $\boldsymbol{\pi} = (0.299, 0.530, 0.171)$; i.e., in the long run, 29.9 percent of customers purchase Brand A, and likewise 53 percent (17.1 percent) selected Brand B(C).

There are innumerable other properties (see *Probability: Formal*) that govern a Markov chain, providing answers to a wide range of questions of interest, such as whether or not a particular state or group of states is transient (i.e., eventual return is certain), absorbing (i.e., once reached cannot be exited), the expected number of visits to a state, and so forth. Details can be found in a variety of basic texts. All these properties evolve from a knowledge of the tpm $\boldsymbol{P}$ and the initial probability vector $\boldsymbol{p}^{(0)}$. Therefore, attention will now be directed to estimating these entities.

2.2 Statistical Inference—Classical Results

Suppose the data consist of N individuals, for each of whom it is known what state, X_{kt} say, that the kth person was in at each of the $t = 0, 1, \ldots, T$ time points recorded, $k = 1, \ldots, N$. For example, the X_{kt} may represent the brand of a product used by individual k at time t. It is assumed that $N > 1$, giving rise to so-called panel studies. Let $n_s(0)$ be the number of individuals in state s at time $t = 0$. The $n_s(0)$ are assumed to be nonrandom; when they are random, the same results pertain except that the $n_s(0)$ are replaced by their expectations, $s = 1, \ldots, S$. Let $n_{ij}(t)$ be the number of individuals who moved from state i at time $t-1$ to state j at time t; and let $n_{ij} = \sum_{t=1}^{T} n_{ij}(t)$ be the total number of transitions from states i to j.

Then, the initial probabilities $p_s^{(0)}$, $s = 1, \ldots, S$, are estimated by

$$\hat{p}_s^{(0)} = n_s(0) / \sum_{s=1}^{S} n_s(0) \tag{4}$$

The maximum likelihood estimator (mle) of p_{ij}, $i, j = 1, \ldots, S$, is

$$\hat{p}_{ij} = n_{ij} / \sum_{j=1}^{S} n_{ij} = \sum_{t=1}^{T} n_{ij}(t) / \sum_{t=1}^{T} \sum_{j=1}^{S} n_{ij}(t) \tag{5}$$

When the Markov chain is not homogeneous, the mle of the transition probabilities $p_{ij}(t)$ is

$$\hat{p}_{ij}(t) = n_{ij}(t) / n_{i.}(t), \quad t = 1, \ldots, T \tag{6}$$

where $n_i(t) = \sum_{j=1}^{S} n_{ij}(t)$.

The transition probabilities for voter behavior from May to June given in Table 3b were estimated using Eqn. (5). Estimated transition probabilities for each of the $T = 5$ transition times can be found by applying Eqn. (6) to Goodman's (1962) data (not shown).

For properties of these estimators and their original derivation, see Anderson and Goodman (1957). They also derived test statistics for key hypothesis testing and model validation situations. Further, these methods can be treated as special cases of loglinear models (see, e.g., Bishop et al. 1975) (see *Multivariate Analysis: Discrete Variables (Overview)*).

2.3 Empirical Bayes Estimation of $\boldsymbol{P}$

When information from other similar data sets exists, that information can be incorporated into the estimation of the particular tpm $\boldsymbol{P}$ under discussion, by calculating an empirical Bayes estimator of $\boldsymbol{P}$. Not only does this utilize all the available information; the technique is especially useful when sample sizes are small, data are sparse and/or mle's tend to take the extreme value of 0 or 1 for some p_{ij}. Thus, in a brand-switching scenario, data are available for each $l = 1, \ldots, L$ other locations (cities, stores, population sectors, etc.), and the location of particular interest is location $l = L+1$. In effect, the p_{ij}s of the tpm $\boldsymbol{P}_{L+1}$ for location $(L+1)$ are assumed to have an underlying distribution whose unknown parameters are estimated by the data from the $\{l = 1, \ldots, L\}$ other chains; and thence through the use of Bayes' Theorem and the data generated at the location $(L+1)$, the tpm $\boldsymbol{P}_{L+1}$ can be estimated.

In this case, there are data values generated for $n_{ij}(t)$, n_{ij}, $n_s(0)$, $\boldsymbol{P}$, etc., for each $l = 1, \ldots, L, L+1$. Then, the empirical Bayes estimate of the element p_{ij} of $\boldsymbol{P}_{L+1}$ is

$$\hat{p}_{EB;ij} = (n_{ij}^* + \hat{\rho}_{ij}) / (n_{i.}^* + \hat{\rho}_{i+}), \quad i, j = 1, \ldots, S \tag{7}$$

where $n^*_{ij} \equiv n_{ij}$ and $n^*_i \equiv n_i$ for location $L+1$ and where $\hat{\rho}_{i+} = \sum_{j=1}^{S} \hat{\rho}_{ij}$. The $\hat{\rho}_{ij}$ can be obtained by method of moments or maximum likelihood techniques.

The method of moments estimate gives

$$\hat{\rho}_{ij} = \bar{M}_{ij}\bar{X}_{ij}/[\bar{M}_{ij}(1-\bar{M}_{ij})-\bar{X}_{ij}], \quad i, j = 1, \ldots, S \tag{8}$$

with

$$\bar{M}_{ijl} = n_{ijl}/n_{i \cdot l}, \quad X_{ijl} = M_{ijl}(1-M_{ijl}), \quad l = 1, \ldots, L$$

The maximum likelihood estimate of $\hat{\rho}_{ij}$ (and hence $\hat{\rho}_{i+}$) are those values for p_{ij} which maximize the likelihood function

$$L(\rho_{ij} \mid n_{ijl}) \propto \frac{\Gamma(\rho_{i+})}{\Gamma(\rho_{it}+n_{i\cdot\cdot})} \prod_{j=1}^{S} \frac{\Gamma(\rho_{ij}+n_{ij\cdot})}{\Gamma(\rho_{ij})} \tag{9}$$

where $n_{ij\cdot} = \sum_{l=1}^{L} n_{ijl}$, $n_{i\cdot} \sum_{j=1}^{S} \sum_{l=1}^{L} n_{ijl}$ and $\Gamma(\cdot)$ is the usual gamma function. Details of the derivation of the results (7), (8), and (9) can be found in Meshkani and Billard (1992), along with properties of these estimators, and an example. When there is only a single Markov chain at each site (i.e., $N = 1$), the $\hat{\rho}_{ij}$ and $\hat{\rho}_{i+}$ are estimated differently; see Billard and Meshkani (1995).

3. Continuous Time Models: Markov Processes

3.1 Finite State Markov Chains

The Markov chains discussed in Sect. 2 were discussed in the context of discrete time. When there is a natural unit of time for which the data of a Markov chain process are collected, such as week, year, generational, etc., use of the discrete time model is satisfactory. If there is no natural time unit, then use of the discrete time model could be problematical. For example, for a given tpm $\boldsymbol{P}$, does the corresponding tpm for one year later equal $\boldsymbol{P}^{(2)}$, or $\boldsymbol{P}^{(12)}$, for a time unit of six months, or one month, respectively? Clearly, the answers are different though the time period—one year later—is the same. Such difficulties can be circumvented by using a continuous time Markov chain; see Singer and Spilerman (1975), and Shorrocks (1978). When time is continuous the concepts elucidated for the discrete time chains carry over though the notation and some of the mathematics differ.

Let X_t represent the state of the chain at time t, and let

$$p_{ij}(t) = P(X_{t+t_0} = j \mid X_{t_0} = i) \tag{10}$$

be the elements of the $S \times S$ matrix $\boldsymbol{P}(t)$ with $\sum_{j=1}^{S} p_{ij}(t) = 1$ for all $i = 1, \ldots, S$. Then, for continuous times t_1 and $t_2 \geqslant 0$, the $\boldsymbol{P}(t)$ satisfy

$$\boldsymbol{P}(t_1+t_2) = \boldsymbol{P}(t_1)\boldsymbol{P}(t_2) \tag{11}$$

or, equivalently,

$$p_{ij}(t_1+t_2) = \sum_{k=1}^{S} p_{ik}(t_1)p_{kj}(t_2) \tag{12}$$

The Eqns. (11) and (12) are called the Chapman–Kolmogorov equations. These equations are used to show that the matrix $\boldsymbol{P}(t)$ satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}(t) = \boldsymbol{Q}\boldsymbol{P}(t) \tag{13}$$

where $\boldsymbol{Q}$ is an $S \times S$ matrix whose elements q_{ij} satisfy

$$q_{ij} = \lim_{\Delta t \to 0} p_{ij}(\Delta t)/\Delta t, \quad i \neq j,$$

$$q_{ii} = \lim_{\Delta t \to \infty} [1-p_{ii}(\Delta t)]/\Delta t, \quad i = j \tag{14}$$

and are called *intensities*. It follows that

$$q_{ii} = -\sum_{\substack{i=1 \\ i \neq j}}^{S} q_{ij} \equiv -q_i, \quad \text{say} \tag{15}$$

Then, $(-q_{ij}/q_i)$ represents the conditional probability that an individual in state i moves to state j given a transition occurred, and $(-1/q_i)$ is the average time an individual stays in state i.

Equation (13) can be solved to give

$$\boldsymbol{P}(t) = \boldsymbol{e}^{\boldsymbol{Q}t} = \sum_{n=0}^{\infty} \boldsymbol{Q}^n t^n/n! = \boldsymbol{U}\boldsymbol{A}(t)\mathbf{U}^{-1} \tag{16}$$

where $\boldsymbol{A}(t) = (\mathrm{diag}(e^{\lambda_i t}))$, $i = 1, \ldots, S$, is the $S \times S$ diagonal matrix with $\{\lambda_i\}$ being the eigenvalues of $\boldsymbol{Q}$, and where $\boldsymbol{U}$ is the $S \times S$ matrix of the eigenvectors of $\boldsymbol{Q}$.

Conditions (on the $p_{ij}(t)$ values), most notably embeddability and identifiability problems, which allow the solution in Eqn. (16) to exist have received considerable study. Singer and Spilerman's (1976) lengthy article provides a nice summary as well as extensive insights. In a series of papers (see, e.g., Stasny 1986), movement in and out of the labor force entirely embodies these principles (where the $S = 3$ three states represent those who are: employed, unemployed, and not in the labor force). Geweke et al. (1986) expand upon these difficulties within the context of mobility models. Some conditions include the following, where $\hat{\boldsymbol{P}} = (\hat{p}_{ij})$ is the observed estimate of $\boldsymbol{P}$ obtained by the methods of Sect. 2.

(a) If $S = 2$, $\hat{p}_{ij} \geqslant 0$, $\hat{p}_{i1}+\hat{p}_{i2} = 1$, $i = 1, 2$, if and only if $\hat{p}_{11}+\hat{p}_{22} > 1$.

(b) If $\hat{p}_{ij}(t) = 0(+0)$, then $\hat{p}^{(n)}_{ij} = 0(+0)$, for all (respectively, any) integers n.

(c) The determinant $|\hat{\boldsymbol{P}}| > 0$.

(d) Other than the first eigenvalue ($\lambda_1 = 1$ always) no other λ_i equals one and any negative λ_i has even multiplicity.

3.2 Markov Processes

Markov processes are continuous time Markov models based on Eqn. (2), where the focus is on the number of individuals in a given state at time t (rather than the transitions probabilities). The two simplest such models are the so-called Poisson process and the birth and death models. Let $X_s(t)$ be the number of individuals in state s at time t, $s = 1, \ldots, S$.

The possible transitions can be visualized as movements in and out of compartments

$$\boxed{s=0} \longrightarrow \cdots \rightleftarrows \boxed{s=i} \rightleftarrows \cdots \rightleftarrows \boxed{s=j} \rightleftarrows \cdots \longrightarrow \boxed{s=S^*} \quad (17)$$

i and j at transition rates $\lambda_{ij}(t)$; see Eqn. (17). It is assumed that two or more transitions do not occur simultaneously, and that transitions in different time intervals occur independently. Typically, the number of compartments S is finite. In many applications, the total size of the population $\sum_{s=1}^{S} X_s(t) = N$ is bounded at N. Where N is unknown (or bounded) as in a pure birth process (say, which has $S = 1$ state corresponding to the size of the population represented by the number of births), then in the context of (17), the 'source' of new births can be thought of as transitions from a hypothetical state $s = 0$. Similarly, a hypothetical state $s = S^* = S+1$ can be incorporated if necessary. Thus, for a birth and death process, $S = 1$, but births correspond to transitions from an $s = 0$ state and deaths reflect transitions from $s = 1$ to a state $s = S^* = 2$.

Markov processes are governed by a system of differential-difference equations obtained from the basic Chapman–Kolmogorov relations (12). To illustrate, for a birth and death process, these equations become

$$\frac{\mathrm{d}}{\mathrm{d}t} p_x(t) = -(\lambda_x + \mu_x) p_x(t) + \lambda_{x-1} p_{x-1}(t) + \mu_{x+1} p_{x+1}(t), \quad x = 1, 2\ldots, \quad (18)$$

$$\frac{\mathrm{d}}{\mathrm{d}t} p_0(t) = -\mu_0 p_0(t) + \mu_1 p_1(t), \quad x = 0$$

where

$$p_x(t) = P[X(t) = x \mid X(0) = a] \quad (19)$$

and where from (2), λ_x and μ_x are the transition rates of births and deaths, respectively, given the population is of size $X(t) = x$. The initial conditions are $p_a(0) = 1$ and $p_x(0) = 0$ for $x \neq a$.

When the birth and death rates are linearly proportional to the population size, i.e., when $\lambda_x = \lambda x$ and $\mu_x = \mu x$, say, then the solution is, for $a = 1$, and $\lambda \neq \mu$,

$$p_x(t) = \begin{cases} (1-\alpha)(1-\beta)\beta^{x-1}, & x \geqslant 1 \\ \alpha, & x = 0 \end{cases} \quad (20)$$

where

$$\alpha = \mu(\delta-1)/(\lambda\delta-\mu), \quad \beta = \lambda(\delta-1)/(\lambda\delta-\mu),$$
$$\delta = \exp\{(\lambda-\mu)t\}.$$

Thus, when $x \geqslant 1$, the population size is modeled by a geometric distribution.

When $\lambda = \mu$ and $a = 1$,

$$p_x(t) = \begin{cases} (\lambda t)^{x-1}/(1+\lambda t)^{x+1} & x \geqslant 1 \\ (\lambda t)/(1+\lambda t) & x = 0 \end{cases} \quad (21)$$

In the particular case that the birth rate is a constant (i.e., $\lambda_x = \lambda$) and the death rate is zero (i.e., $\mu_x = 0$), the solution of Eqn. (18) gives

$$p_x(t) = P[X(t) = x] = e^{-\lambda t}(\lambda t)^x / x!, \quad x = 0, 1, \ldots \quad (22)$$

where, without loss of generality, $a = 0$. This is the Poisson process. An implication of these results is that the waiting time between occurrence of events is exponentially distributed. Thus, in the Poisson process, the average waiting time between events (whose total number follows a Poisson distribution) is $(1/\lambda)$.

In the birth and death process, the average size of the population at time t is

$$E\{X(t)\} = \begin{cases} a\delta, & \lambda \neq \mu \\ a, & \lambda = \mu \end{cases} \quad (23)$$

with variance

$$\mathrm{Var}\{X(t)\} = \begin{cases} a\delta(\delta_{-1})(\lambda+\mu)/(\lambda-\mu) & \lambda \neq \mu \\ 2a\lambda t & \lambda = \mu \end{cases} \quad (24)$$

The average waiting time for a birth, or a death, to occur given that a transition does occur is $\lambda_x/(\lambda_x+\mu_x)$, or $\mu_x/(\lambda_x+\mu_x)$, respectively. The probability of eventual extinction is one if $\lambda \leqslant \mu$, and $(\mu/\lambda)^a$ if $\lambda > \mu$.

Corresponding results for a pure linear birth (or death) process can be obtained by substituting $\mu = 0$, (or $\lambda = 0$, respectively), into Eqns. (18), (20), (23), and (24).

When there are S states, the system of Eqns. (18) is generalized as follows. Let $\boldsymbol{X}(t) = [(X_1(t), \ldots, X_S(t)]$ be the random variable representing the number of individuals in each state $s = 1, \ldots, S$, at time t. Let $p_x(t) = P\{\boldsymbol{X}(t) = \boldsymbol{x} \mid \boldsymbol{X}(0) = \boldsymbol{a}\}$, where $\boldsymbol{x} = (x_1, \ldots, x_s)$ is a specific value of $\boldsymbol{X}(t)$. Write $x_{ij} = (x_1, \ldots, x_i+1, \ldots, x_j-1, \ldots, x_s)$, i.e., $x_i \equiv x_i+1$ and $x_j \equiv x_j-1$ and all

other $x_k \equiv x_k$; and likewise for the $\lambda_{ij}(t)$. Then, the Markov process is governed by the system of differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t} p_x(t) = -\sum_{i,j=1}^{S} \lambda_{ij}(t) p_x(t) + \sum_{\substack{i,j=1 \\ i \neq j}}^{S} \lambda_{i+1,\, j-1} p_{x_{ij}}(t) \qquad (25)$$

for $\boldsymbol{x}$ in the S-dimensional space of possible values. Typically, there exist some boundary conditions on one or more of the values x_s, $s = 1, \ldots, S$, which necessitate a separate equation for that x_s value. This occurred in the linear birth and death process modeled in Eqn. (18), when $x = 0$. Models for which $S > 1$ are sometimes called interactive Markov models.

In general, except when the transition rates $\lambda_{ij}(t)$ are both independent of time and linear (in one only of i and j, but not in both, such as when, e.g., $\lambda_{ij}(t) = \lambda i$ but not λij), the existence and derivation of solutions to Eqn. (25) are nontrivial tasks. However, when all the transitions across states flow in the same direction, i.e., when in (2), $\lambda_{ij}(t) = 0$ for all $j < i$ to give a so-called right-shift process (or, equivalently, a left-shift process if $\lambda_{ij}(t) = 0$ for all $i < j$), then the recursion theorem of Severo (1969) can be applied successfully to obtain the underlying state probabilities from which all other properties can be obtained. Although written in the context of the progression of HIV to AIDS, Billard and Zhao (1994) give a general solution for time-dependent transitions moving to the next state, $j = i+1$. Nevertheless, the solutions are still computationally intensive.

Unfortunately, since many applications have nonlinear transition rates, these difficulties cannot be dismissed lightly. For example, the spread of epidemics and rumors is modeled invariably with nonlinear rates. For simplicity, suppose $S = 3$ with states $s = 1$, $s = 2$, and $s = 3$ corresponding, respectively, to the number of susceptibles, infectives and removals (either by recovery, death, immunity, etc.). Then, the rate that a susceptible person becomes infected is proportional to both the number of susceptible people and the number of infective people; thus $\lambda_{12}(t) = \beta x_1 x_2$ (where β is the infection rate), and is a nonlinear function of the underlying variables. More generally, there may be several stages of infection and/or removals, leading to an $S > 3$. The spread of rumors and news propagation corresponds to infection spread. Likewise, friendship formations, and indeed network systems and population modeling more generally, typically have nonlinear rates and are modeled as Markov processes.

If it is known that ultimately (as time t extends to ∞) steady-state conditions prevail, then, the derivatives in Eqns. (18) or (25) become $\mathrm{d}p_x(t)/\mathrm{d}t \equiv 0$ with $p_x(t) \equiv p_x$. The resulting equations can then be solved quite easily to obtain expressions for p_x. Notice these do not provide information about the process at specific times t before the steady-state conditions occur.

Some authors have attempted to work with the deterministic versions of the stochastic Eqn. (25). In this case, instead of the probabilities $p_x(t)$, the $\boldsymbol{X}(t)$ are set as fixed rather than random variables, and these $\boldsymbol{X}(t)$ replace the $p_x(t)$ in Eqn. (25). Great care is needed here. While the solution to the deterministic equations equals the expected value of the stochastic variable when the λ_{ij} are linear rates, the equivalence does *not* hold for nonlinear transition rates. However, the deterministic result may provide a reasonable approximation when the population size increases to infinity ($N \to \infty$) for a fixed time t. This does not hold in the reverse case (as $t \to \infty$ for fixed N). Bartholomew (1984) and Lehoczky (1980) provide an excellent discussion of this issue.

4. *Some Other Models*

It can happen for a specific social process and/or a particular data set, that a unique feature exists which precludes the direct application of the basic Markov model as described thus far. This gives rise to models which are still Markovian in nature but which have an added characteristic.

One such model is the mover–stayer model, in which the population consists of two subgroups, the movers and the stayers. The movers are individuals who move (with non-zero probability) between the states according to a Markov chain process, while the stayers are those who always stay in their initial state. Thus, if θ is the proportion of stayers in state s, $s = 1, \ldots, S$, then the transition probability of moving from state i to t to state j at $t+n$ is

$$p^*_{ij} = \theta_i \delta_{ij} + (1-\theta_i) p^{(n)}_{ij} \qquad (26)$$

where $p^{(n)}_{ij}$ is the n-step ahead transition probability (found from Eqn. (3)) and δ_{ij} is the Kronecker delta ($\delta_{ij} = 0$ if $i \neq j$, and 1 if $i = j$). The $p^{(n)}_{ij}$ are estimated by the methods of Sect. 2. The mle estimator of θ_s is

$$\hat{\theta}_s = 1 - \{[1 - n_s \hat{p}^T_{ss}/n_s(0)]/(1 - \hat{p}^T_{ss}) \qquad (27)$$

where $\hat{p}_{ss}$ is the mle estimator of p_{ss}, $n_s(0)$ is the number of individuals in states s at time $t = 0$, and n_s is the number of individuals (stayers) who were in state s for all times $t = 0, 1, \ldots, T$; see Frydman (1984). Sampson (1990) considers a restricted model and illustrates the model with data describing movement of men across seven employment sectors.

While the mover–stayer model accommodates some heterogeneity, more generally there may be K subpopulations. The underlying process consists of a mixture of K Markov chains giving rise to a mixed Markov chain model. Thus, a mover–stayer model is the special case that $K = 2$, and that one of the tpm's is the identity matrix $\boldsymbol{I}$. Vogel (1996) considers a mixture of $K = 2$ populations, those who do and those

who do not enter into extensive plea bargaining in the criminal setting. In a mixed model, individuals can only move within their own subpopulation. When they are able to move out of one subpopulation into another, the model becomes a latent Markov chain model. See Langeheine and Pol (1990) for a detailed discussion of these models.

A semi-Markov model is one that is specified by both the transition probabilities ($\boldsymbol{P}$ as in a Markov chain) and the distribution (specifically the parameters) of the length of time R_{ij} an individual stays in a state i before moving to state j, $i, j = 1, \ldots, S$ (which need not equal the geometric waiting time distribution of the standard Markov chain). Note that transitions and waiting times depend only on the current state and are conditionally independent of previously visited states; see, e.g., Weiss et al. (1982).

5. Conclusion

This article has described briefly the basic principles underlying Markov models. Markov models are particularly useful to describe a wide variety of behavior such as consumer behavior patterns, mobility patterns, friendship formations, networks, voting patterns, environmental management (e.g., patient movement between hospital stations), etc., among innumerable other applications in the social sciences. Four particular examples have been provided herein. The entry also included brief summaries of how data governed by these Markov processes can be assessed and modeled. The texts by Bartholomew (1981, 1982) and Bartholomew and Forbes (1979), along with the review articles by Anderson (1954), Goodman (1962), Singer and Spilerman (1976) and Cohen and Singer (1979), among others, provide more extensive details of these models and their applications.

Bibliography

Anderson T W 1954 Probability models for analyzing time changes in attitudes. In: Lazarsfeld P F (ed.) *Mathematical Thinking in the Social Sciences*. The Free Press, Glencoe, IL, pp. 17–66

Anderson T W, Goodman L A 1957 Statistical inference about Markov chains. *Annals of Mathematical Statistics* **28**: 89–110

Bartholomew D J 1981 *Mathematical Methods in Social Science*, Guidebook 1. Wiley, Chichester, UK

Bartholomew D J 1982 *Stochastic Models for Social Processes*, 3rd edn. Wiley, Chichester, UK

Bartholomew D J 1983 Some recent developments in social statistics. *International Statistical Review* **51**: 1–9

Bartholomew D J 1984 Recent developments in nonlinear stochastic modelling of social processes. *Canadian Journal of Statistics* **12**: 39–52

Bartholomew D J, Forbes A F 1979 *Statistical Techniques for Manpower Planning*. Wiley, Chichester, UK

Biblarz T J, Raftery A E 1993 The effects of family disruption on social mobility. *American Sociological Review* **58**: 97–109

Billard L, Meshkani M R 1995 Estimation of a stationary Markov chain. *Journal of the American Statistical Association* **90**: 307–15

Billard L, Zhao Z 1994 Nonhomogeneous Markov models for the AIDS epidemic. *Journal of the Royal Statistical Society, B* **56**: 673–86

Bishop Y M M, Fienberg S E, Holland P W 1975 *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA

Brown P J, Payne C D 1986 Aggregate data, ecological regression, and voting transitions. *Journal of the American Statistical Association* **81**: 452–60

Chatfield C 1973 Statistical inference regarding Markov chain models. *Applied Statistics* **22**: 7–20

Cohen J E, Singer B 1979 Malaria in Nigeria: Contrained continuous-time Markov models for discrete-time longitudinal data on human mixed-species infections. In: Levin S A (ed.) *Some Mathematical Questions in Biology*. American Mathematical Society, Washington, DC, pp. 69–133

Draper J E, Nolin L H 1964 A Markov chain analysis of brand preferences. *Journal of Advertising Research* **4**: 33–8

Frydman H 1984 Maximum likelihood estimation in the mover–stayer model. *Journal of the American Statistical Association* **79**: 632–38

Goodman L A 1962 Statistical methods for analyzing processes of change. *American Journal of Sociology* **68**: 57–78

Geweke J, Marshall R C, Zarkin G A 1986 Mobility indices in continuous time Markov chains. *Econometrica* **54**: 1407–23

Langeheine R, Van de Pol F 1990 A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research* **18**: 416–41

Lawrence R J 1966 Models of consumer purchasing behaviour. *Applied Statistics* **15**: 216–33

Lehoczky J P 1980 Approximations for interactive Markov chains in discrete and continuous time. *Journal of Mathematical Sociology* **7**: 139–57

Meshkani M R, Billard L 1992 Empirical Bayes estimators for a finite Markov chain. *Biometrika* **79**: 185–93

Sampson M 1990 A Markov chain model for unskilled workers and the highly mobile. *Journal of the American Statistical Association* **85**: 177–80

Severo N C 1969 A recursion theorem on solving differential-difference equations and applications to some stochastic processes. *Journal of Applied Probability* **6**: 673–81

Shorrocks A F 1978 The measurement of mobility. *Econometrica* **46**: 1013–24

Singer B, Spilerman S 1975 Identifying structural parameters of social processes using fragmentary data. *Bulletin International Statistical Institute* **46**: 681–97

Singer B, Spilerman S 1976 The representation of social processes by Markov models. *American Journal of Sociology* **82**: 1–54

Stasny E A 1986 Estimating gross flows using panel data with nonresponse: An example from the Canadian labor force survey. *Journal of the American Statistical Association* **81**: 42–7

Vogel M E 1996 The negotiated guilty plea: Vacancies as an alternative to the caseload pressure explanation. *Journal of Mathematical Sociology* **21**: 241–88

Weiss E N, Cohen M A, Hershey J C 1982 An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital flow. *Operations Research* **30**: 1082–104

L. Billard

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Markov Processes for Knowledge Spaces

This article describes some Markovian procedures designed for assessing the competence of individuals, such as school children, regarding a particular topic. In practice, the assessment is performed by a computer, which asks questions to the individual under examination. The Markovian search takes place in a 'knowledge structure,' a combinatoric object stored in the computer and gathering all the feasible 'knowledge states.' A 'knowledge space' is a particular kind of knowledge structure (see *Knowledge Spaces*). Both of these concepts are defined below. The main reference is Doignon and Falmagne (1999, Chaps. 10 and 11; see also Falmagne and Doignon 1988a, 1988b). For general references on stochastic processes, see, e.g., Parzen (1964), or more specifically, for Markov processes and learning models, Bharucha-Reid (1988), and Norman (1972).

Suppose that the topic is arithmetic, and that a large set Q of questions or problems (or rather, problem types, in the sense that 2-digit addition with carry, without further specification, is a problem type), have been selected to delineate that topic. For concreteness, let us assume that Q contains around 100 such questions. The *knowledge state* of an individual is a particular subset of Q containing all the problems mastered by that individual. The family of all the knowledge states is called the *knowledge structure*, which by convention is assumed to contain at least Q and the empty set. In practice, the number of feasible knowledge states is relatively small. In the case of our arithmetic example, there may be only a few dozen thousand knowledge states, which is a minute fraction of the 2^{100} subsets of Q. The task of the assessment is to uncover, by an efficient questioning, the knowledge state of a particular individual under examination. The situation is similar to that of adaptive testing in psychometrics (Weiss 1983), the critical difference being that the outcome of the assessment here is a knowledge state in the above sense, rather than a numerical estimate of the individual's competence in the field.

Two classes of procedures are outlined in this article. Both of them enter into the general framework depicted by the transition diagram of Fig. 1.

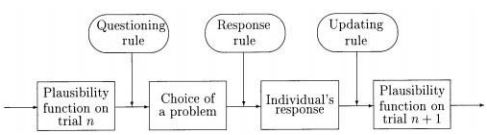

Figure 1
Diagram of the transitions in the two Markov procedures

At the onset of step n of an assessment, all the information accumulated from step 1 to step n–1 is summarized by the value of a 'plausibility function' assigning a value to each of the knowledge states and measuring the plausibility that the individual assessed is in any given state. A 'questioning rule' is then applied to this plausibility function to determine the next question to ask. The individual's response to that question is observed, and it is assumed that this response is generated from the individual's unknown knowledge state via a 'response rule.' In a simple case, it is supposed that the response is correct if the question belongs to the individual's knowledge state, and false otherwise. More sophisticated response rules involve some random factors, such as careless errors or lucky guesses. Finally, the plausibility function is recomputed, through an 'updating rule,' based on its current values, the question asked, and the response given.

For both classes of procedures, we consider a pair $(Q, \mathscr{K})$, where Q is a finite set (of questions) called the domain, and $\mathscr{K}$ is a knowledge structure on Q, that is, a family of subset of Q containing at least Q and the empty set $\varnothing$. A knowledge structure closed under union is called a knowledge space. This closure property plays a generally important role in the theory but is not relevant in the first procedure described below.

1. *A Class of Continuous Markovian Procedures*

We suppose that there is some a priori, positive probability distribution P on $\mathscr{K}$; thus, $0 < P < 1$ and $\sum_{K \in \mathscr{K}} P(K)$. This probability distribution P may be regarded as measuring the uncertainty of the assessor at the beginning of the assessment and may also represent the distribution of the knowledge states in the population of reference. As in Falmagne and Doignon (1988a), we assume that, on every step n of the procedure, the plausibility function is a likelihood function $\mathscr{L}_n$ on the set $\mathscr{K}$ of knowledge states. We set $\mathscr{L}_1 = P$ to specify the likelihood on step 1. We denote by $\boldsymbol{Q}_n \in Q$ and $\boldsymbol{R}_n = 0, 1$ the question asked and the individual's response on step n. Thus, correct and incorrect responses are coded as $\boldsymbol{R}_n = 1$ and $\boldsymbol{R}_n = 0$, respectively. The basic assumptions of this procedure are as follows: on every step $n \geqslant 1$, and given the complete sequence $\boldsymbol{W}_n = ((\mathscr{L}_1, \boldsymbol{Q}_1, \boldsymbol{R}_1), \ldots, (\mathscr{L}_n, \boldsymbol{Q}_n, \boldsymbol{R}_n))$ of likelihood functions, questions and responses,

(a) the value of $\mathscr{L}_{n+1}$ only depends upon the triple $(\mathscr{L}_n, \boldsymbol{Q}_n, \boldsymbol{R}_n)$ via an updating rule

$$u\colon (\mathscr{L}_n, \boldsymbol{Q}_n, \boldsymbol{R}_n) \mapsto u\,(\mathscr{L}_n, \boldsymbol{Q}_n, \boldsymbol{R}_n) = \mathscr{L}_{n+1};$$

(b) the choice of $\boldsymbol{Q}_{n+1}$ only depends upon $\mathscr{L}_{n+1}$ via a questioning rule Ψ;

(c) the individual's response—i.e., the value of $\boldsymbol{R}_{n+1}$—only depends upon $\boldsymbol{Q}_{n+1}$ and the individual's true knowledge state K_0, which is supposed to be constant.

(The last assumption implies that there is no learning during the questioning. In practice, this assumption can be relaxed somewhat without major difficulty.) Once properly formalized, it readily follows from these three assumptions rules that the stochastic process $(\mathscr{L}_n)$ is Markovian with a state space containing all the probability distributions on $\mathscr{K}$. A similar result hold for the stochastic processes $(\mathscr{L}_n, \boldsymbol{Q}_n, \boldsymbol{R}_n)$ and $(\mathscr{L}_n, \boldsymbol{Q}_n)$. In practical applications, one selects an updating rule that raises the probability of all the knowledge states containing a question to which the individual has responded correctly. Similarly, the chosen updating rule must also decrease the probability of all those states containing a questions that has been failed. More precisely, for any knowledge state K in $\mathscr{K}$, we denote by $\iota_K : Q \to \{0, 1\}: q \mapsto \iota_K(q)$ the indicator function of K. Thus, for all $q \in Q$ and $K \in \mathscr{K}$,

$$\iota_K(q) = \begin{cases} 1 & \text{if } q \in K \\ 0 & \text{if } q \in Q \backslash K \end{cases}$$

We also write u_K for the coordinate of u associated with state K; thus

$$\mathscr{L}_{n+1}(K) = u_K(\boldsymbol{R}_n, \boldsymbol{Q}_n, \mathscr{L}_n)$$

A sensible updating rule must satisfy:

$$u_K(\mathscr{L}_n, \boldsymbol{Q}_n, \boldsymbol{R}_n) \begin{cases} > \mathscr{L}_n(K) & \text{if } \iota_K(\boldsymbol{Q}_n) = \boldsymbol{R}_n \\ < \mathscr{L}_n(K) & \text{if } \iota_K(\boldsymbol{Q}_n) \neq \boldsymbol{R}_n \end{cases}$$

Commonsensical consideration also govern the choice of the questioning rule Ψ, which must select the questions to ask so as to maximize the information in the response. We consider below a prominent special case for each of the updating rule u and the questioning rule Ψ.

1.1 A Commutative Updating Rule

Because the data collected at any time during the assessment are as important as the rest, it make sense to favor an updating rule that is 'commutative,' in the sense that its effects do not depend upon the order of the event pairs in the sequence $(\boldsymbol{Q}_1, \boldsymbol{R}_1), \ldots, (\boldsymbol{Q}_n, \boldsymbol{R}_n), \ldots$. The updating rule described here is based on parameters $1 < \zeta_{q,r}$ $(q \in \boldsymbol{Q}, r = 0, 1)$ satisfying the condition: with $\boldsymbol{Q}_n = q$, $\boldsymbol{R}_n = r$,

$$\zeta^K_{q,r} = \begin{cases} 1 & \text{if } \iota_K(q) \neq r \\ \zeta_{q,r} & \text{if } \iota_K(q) = r \end{cases}$$

The updating rule u is defined by

$$u_K(\mathscr{L}_n, q, r) = \frac{\zeta^K_{q,r} \mathscr{L}_n(K)}{\sum_{K' \in \mathscr{K}} \zeta^{K'}_{q,r} \mathscr{L}_n(K')} \tag{1}$$

It is easy to verify that this rule is commutative. (Its operator is *permutable* in the sense of functional equation theory; see Aczél 1966, p. 270.) This updating rule is consistent with a Bayesian updating. (This was pointed out by Mathieu Koppen; see Doignon and Falmagne 1999, pp. 230–32.)

1.2 The Half-split Questioning Rule

For any question q in Q, we write $\mathscr{K}_q$ for the subset of $\mathscr{K}$ containing all the knowledge states containing q, and $\mathscr{K}_{\bar{q}}$ for $\mathscr{K} \backslash \mathscr{K}_{\bar{q}}$. A simple idea for the questioning rule is to always choose, on any step of the procedure, a question q that partitions the set $\mathscr{K}$ into two subsets $\mathscr{K}_{\bar{q}}$ and $\mathscr{K}_{\bar{q}}$ with probability masses as equal as possible. Thus, on step n, we select q in Q such that $\mathscr{L}_n(\mathscr{K}_q)$ is as close as possible to $\mathscr{L}_n(\mathscr{K}_{\bar{q}}) = 1 - \mathscr{L}_n(\mathscr{K}_q)$. Clearly, any likelihood $\mathscr{L}_n$ defines a set $S(\mathscr{L}_n) \subseteq Q$ containing all those questions q minimizing $|2\mathscr{L}_n(\mathscr{K}_q) - 1|$. This questioning rule requires that $\boldsymbol{Q}_n \in S(\mathscr{L}_n)$ with a probability equal to one. The questions in the set $S(\mathscr{L}_n)$ are then chosen with equal probability. We have thus

$$\Psi(q, \mathscr{L}_n) = \frac{\iota_{S(\mathscr{L}_n)}(q)}{|S(\mathscr{L}_n)|}$$

This particular rule for choosing the question is known as the *half-split* questioning rule.

Other examples of updating rules and questioning rules for this type of continuous procedure are considered in Doignon and Falmagne (1999, Chap. 10; see also Falmagne and Doignon, 1988a). Theoretical and simulation results indicate that such procedures are capable of efficiently uncovering an individual's knowledge state.

2. A Class of Discrete Markovian Procedures

These procedures are similar in spirit to those described in the previous section, but different in a key aspect: they are based on finite Markov chains rather than on Markov processes with an uncountable set of Markov states (namely, the set of all likelihood functions on $\mathscr{K}$). As a consequence, these procedures requires less storage and computation and can thus be implemented on a small machine. We only outline the key ideas (for details, see Doignon and Falmagne 1999, Chap. 11, or Falmagne and Doignon 1988b).

As before, a fixed knowledge structure $(Q, \mathscr{K})$ is used by the assessor. It is supposed that, at any time in

the course of the assessment, some of the knowledge states in $\mathscr{K}$ are considered as plausible from the standpoint of the assessor. These 'marked states' are collected in a family which is the value of a random variable $\boldsymbol{M}_n$, where $n = 1, 2, \ldots$ indicates the step number. (Thus, $\boldsymbol{M}_n$ denotes the set of all marked states on step n.) Under certain conditions, $(\boldsymbol{M}_n)$ is a Markov chain. Moreover, during the initial phase of the procedure, this subfamily $\boldsymbol{M}_n \subseteq \mathscr{K}$ decreases in size until a single marked state remains. In the second phase, the single 'marked state' evolves in the structure. This last feature allows the assessor, through a statistical analysis of the observed sequence of problems and answers, to estimate the 'true state.' (A formal definition of 'true state' is given.) Note that, in some cases, a useful estimate can be obtained even if the rule state is not part of the structure.

Both classes of procedures have been applied in practice, especially the continous ones. Results from simulation and from real life applications indicate that these procedures are efficient in uncovering the knowledge state of an individual. Note however that the validity of these procedures heavily depends upon the accuracy of the knowledge structure on which they operate. As is well known, the construction of a knowledge structure faithfully describing the knowledge states in a population is expensive and time consuming.

See also: Conjoint Analysis Applications; Educational Assessment: Major Developments; Educational Evaluation: Overview; Knowledge Representation; Knowledge Spaces; Markov Decision Processes; Mathematical Psychology; Mathematical Psychology, History of; Neural Networks and Related Statistical Latent Variable Models

Bibliography

Aczél J 1966 *Lectures on Functional Equations and Their Applications*. Academic Press, New York

Bharucha-Reid A T 1988 *Elements of the Theory of Markov Processes and Their Applications*. Dover, Minneola, NY

Doignon J-P, Falmagne J-Cl 1999 *Knowledge Spaces*. Springer-Verlag, Berlin–Heidelberg

Falmagne J-C, Doignon J-P 1988a A class of stochastic procedures for the assessment of knowledge. *The British Journal of Mathematical and Statistical Psychology* **41**: 1–23

Falmagne J-C, Doignon J-P 1988b A Markovian procedure for assessing the state of a system. *Journal of Mathematical Psychology* **32**: 232–58

Norman F 1972 *Markov Processes and Learning Models*. Academic Press, New York

Parzen E 1964 *Stochastic Processes*. Holden-Day, San Francisco, CA

Weiss D J 1983 *New Horizons in Testing: Latent Trait Theory and Computerized Testing*. Academic Press, New York

J.-C. Falmagne

Maroons in Anthropology

Maroons—escaped slaves and their descendants—still form semi-independent communities in several parts of the Americas, e.g., Suriname, French Guiana, Jamaica, Colombia, and Brazil. As the most isolated of Afro-Americans, they have since the 1920s been an important focus of anthropological research, contributing to theoretical debates about slave resistance, the heritage of Africa in the Americas, the process of creolization, and the nature of historical knowledge among nonliterate peoples.

1. Maroons and Their Communities

The English word 'maroon' derives from Spanish *cimarrón*—itself based on an Arawakan (Taino) Indian root. *Cimarrón* originally referred to domestic cattle that had taken to the hills in Hispaniola, and soon after to American Indian slaves who had escaped from the Spaniards. By the end of the 1530s, the word was being used primarily to refer to Afro-American runaways.

Communities formed by maroons dotted the fringes of plantation America, from Brazil to Florida, from Peru to Texas. Usually called *palenques* in the Spanish colonies and *mocambos* or *quilombos* in Brazil, they ranged from tiny bands that survived less than a year to powerful states encompassing thousands of members that lasted for generations or even centuries.

During the last decades of the twentieth century, anthropological fieldwork has underlined the strength of historical consciousness among the descendants of these rebel slaves and the dynamism and originality of their cultural institutions. Meanwhile, historical scholarship on maroons has flourished, as new research has done much to dispel the myth of the docile slave. Marronage represented a major form of slave resistance, whether accomplished by lone individuals, by small groups, or in great collective rebellions. Throughout the Americas, maroon communities stood out as an heroic challenge to white authority, as the living proof of the existence of a slave consciousness that refused to be limited by the whites' conception or manipulation of it. It is no accident that throughout the Caribbean today, the historical maroon—often mythologized into a larger-than-life figure—has become a touchstone of identity for the region's writers, artists, and intellectuals, the ultimate symbol of resistance and the fight for freedom.

More generally, Maroons and their communities can be seen to hold a special significance for the study of Afro-American societies. For a while they were, from one perspective, the antithesis of all that slavery stood for; additionally, they were also a widespread and embarrassingly visible part of these systems. Just as the very nature of plantation slavery implied violence and resistance, the wilderness setting of early

New World plantations made marronage and the existence of organized maroon communities a ubiquitous reality.

Planters generally tolerated *petit marronage*—truancy with temporary goals such as visiting a friend or lover on a neighboring plantation. But in most slave-holding colonies, the most brutal punishments—amputation of a leg, castration, suspension from a meat hook through the ribs, slow roasting to death—were reserved for long-term, recidivist maroons, and in many cases these were quickly written into law. Marronage on the grand scale, with individual fugitives banding together to create communities, struck directly at the foundations of the plantation system, presenting military and economic threats that often taxed the colonists to their very limits. Maroon communities, whether hidden near the fringes of the plantations or deep in the forest, periodically raided plantations for firearms, tools, and women, often allowing families that had formed during slavery to be reunited in freedom.

In many cases, the beleaguered colonists were eventually forced to sue their former slaves for peace. In Brazil, Colombia, Cuba, Ecuador, Hispaniola, Jamaica, Mexico, Panama, Peru, Suriname, and Venezuela, for example, the whites reluctantly offered treaties granting maroon communities their freedom, recognizing their territorial integrity, and making some provision for meeting their economic needs, in return for an end to hostilities toward the plantations and an agreement to return future runaways. Of course, many maroon societies were crushed by massive force of arms, and even when treaties were proposed they were sometimes refused or quickly violated. Nevertheless, new maroon communities seemed to appear almost as quickly as the old ones were exterminated, and they remained, from a colonial perspective, the 'chronic plague' and 'gangrene' of many plantation societies right up to final Emancipation.

2. *African Origins, New World Creativity*

The initial maroons in any New World colony hailed from a wide range of societies in West and Central Africa— at the outset, they shared neither language nor other major aspects of culture. Their collective task, once off in the forests or mountains or swamplands, was nothing less than to create new communities and institutions, largely via a process of inter-African cultural syncretism. Those scholars, mainly anthropologists, who have examined contemporary maroon life most closely seem to agree that such societies are often uncannily 'African' in feeling but at the same time largely devoid of directly transplanted systems. However 'African' in character, no maroon social, political, religious, or aesthetic system can be reliably traced to a specific African ethnic provenience—they reveal rather their syncretistic composition, forged in the early meeting of peoples of diverse African, European, and Amerindian origins in the dynamic setting of the New World.

The political system of the great seventeenth-century Brazilian maroon kingdom of Palmares, for example, which R. K. Kent has characterized as an 'African' state, 'did not derive from a *particular* Central African model, but from several' (Price 1996, p. 188). In the development of the kinship system of the Ndyuka Maroons of Suriname, writes André Köbben, 'undoubtedly their West-African heritage played a part ... [and] the influence of the matrilineal Akan tribes is unmistakable, but so is that of patrilineal tribes ... [and there are] significant differences between the Akan and Ndyuka matrilineal systems' (Price 1996, p. 324). Historical and anthropological research has revealed that the magnificent wood-carving of the Suriname Maroons, long considered 'an African art in the Americas' on the basis of formal resemblances, is in fact a fundamentally new, Afro-American art 'for which it would be pointless to seek the origin through direct transmission of any particular African style' (Hurault 1970, p. 84). And detailed investigations—both in museums and in the field—of a range of cultural phenomena among the Saramaka Maroons of Suriname have confirmed the dynamic, creative processes that continue to animate these societies.

Maroon cultures do possess direct and sometimes spectacular continuities from particular African peoples, from military techniques for defense to recipes for warding off sorcery. These are, however, of the same type as those that can be found, if with lesser frequency, in Afro-American communities throughout the hemisphere. In stressing these isolated African 'retentions,' there is a danger of neglecting cultural continuities of a more significant kind. Roger Bastide (1972, pp. 128–51) divided Afro-American religions into those he considered 'preserved' or 'canned'—like Brazilian *candomblé*—and those that he considered 'alive'—like Haitian *vaudou*. The former, he argued, manifest a kind of 'defense mechanism' or 'cultural fossilization,' a fear that any small change may bring on the end, while the latter are more secure of their future and freer to adapt to the changing needs of their adherents. And indeed, tenacious fidelity to 'African' *forms* seems, in many cases, to indicate a culture finally having lost meaningful touch with the vital African past. Certainly, one of the most striking features of West and Central African cultural systems is their internal dynamism, their ability to grow and change. The cultural uniqueness of the more developed maroon societies (e.g., those in Suriname) rests firmly on their fidelity to 'African' cultural principles at these deeper levels—whether aesthetic, political, religious, or domestic—rather than on the frequency of their isolated 'retentions.' With a rare freedom to extrapolate ideas from a variety of African societies and

adapt them to changing circumstance, maroon groups included (and continue to include today) what are in many respects *at once* the most meaningfully African *and* the most truly 'alive' and culturally dynamic of all Afro-American cultures.

3. Four Cases

The most famous maroon societies are Palmares in Brazil, Palenque de San Basilio in Colombia, the Maroons of Jamaica, and the Saramaka and Ndyuka Maroons of Suriname.

Because Palmares was finally crushed by a massive colonial army in 1695, after a century of success and growth, actual knowledge of its internal affairs remains limited, based as it is on soldiers' reports, the testimony of a captive under torture, official documents, modern archaeological work, and the like. But as a modern symbol of black (and anti-colonial) heroism, Palmares continues to evoke strong emotions in Brazil. (For recent scholarship, including archaeology and anthropology, see Reis and dos Santos Gomes 1996.)

Palenque de San Basilio boasts a history stretching back to the seventeenth century. In recent years, historians, anthropologists, and linguists working in collaboration with Palenqueros have uncovered a great deal about continuities and changes in the life of these early Colombian freedom fighters. For an illustrated introduction, see de Friedemann and Cross (1979).

The Jamaica Maroons, who continue to live in two main groups centered in Accompong (in the hills above Montego Bay) and in Moore Town (deep in the Blue Mountains), maintain strong traditions about their days as freedom fighters. Two centuries of scholarship, some written by Maroons themselves, offer diverse windows on the ways these men and women managed to survive and build a vibrant culture within the confines of a relatively small island. (A useful entree is provided in Agorsah (1994).)

The Suriname Maroons now constitute the most fully documented case of how former slaves built new societies and cultures, under conditions of extreme deprivation, in the Americas—and how they developed and maintained semi-independent societies that persist into the beginning of the twenty-first century. From their late seventeenth-century origins and the details of their wars and treaty-making to their current struggles with multinational mining and timber companies, much is now known about these peoples' achievements, in large part because of the extensive recent collaboration by Saramakas and Ndyukas with anthropologists. The relevant bibliography now numbers in the thousands; useful points of entry are Price and Price (1999) and Thoden van Velzen and van Wetering (1988).

4. Anthropological Issues

Since the Herskovitses' fieldwork in Suriname in the 1920s (Herskovits and Herskovits 1934), Maroons have moved to the center of anthropological debates, from the origins of creole languages and the 'accuracy' of oral history to the nature of the African heritage in the Americas and the very definition of Afro-American anthropology. Indeed, David Scott (1991, p. 269) argues that the Saramaka Maroons have by now become 'a sort of anthropological metonym ... providing the exemplary arena in which to argue out certain anthropological claims about a discursive domain called Afro-America.' Much of the most recent anthropological research has focused on Maroon historiography—how Maroons themselves conceptualize and transmit knowledge about the past—and has privileged the voices of individual Maroon historians. Eric Hobsbawm (1990, p. 46), commenting on this work in the more general context of the social sciences, notes that 'Maroon societies raise fundamental questions. How do casual collections of fugitives of widely different origins, possessing nothing in common but the experience of transportation in slave ships and of plantation slavery, come to form structured communities? How, one might say more generally, are societies founded from scratch? What exactly did or could such refugee communities ... derive from the old continent?' Questions such as these are sure to keep students of Maroon societies engaged in active research for many years to come.

See also: Caribbean: Sociocultural Aspects; Ethnogenesis in Anthropology; Latin American Studies: Economics; Race Identity; Racism, History of; Slavery: Comparative Aspects; Slaves/Slavery, History of; South America: Sociocultural Aspects

Bibliography

Agorsah E K (ed.) 1994 *Maroon Heritage: Archaeological, Ethnographic and Historical Perspectives*. Canoe Press, Barbados

Bastide R 1972 *African Civilizations in the New World*. Harper & Row, New York

de Friedemann N S, Cross R 1979 *Ma ngombe: Guerreros y ganaderos en Palenque*, 1st edn. Carlos Valencia Editores, Bogotá

Herskovits M J, Herskovits F S 1934 *Rebel Destiny: Among the Bush Negroes of Dutch Guiana*. McGraw-Hill, New York

Hobsbawm E J 1990 Escaped slaves of the Forest. *New York Review of Books* (Dec 6): 46–8

Hurault J 1970 *Africains de Guyane: la vie matérielle et l'art des Noirs Réfugiés de Guyane*. Mouton, The Hague

Price R 1990 *Alabi's World*. Johns Hopkins University Press, Baltimore, MD

Price R (ed.) 1996 *Maroon Societies: Rebel Slave Communities in the Americas*, 3rd edn. Johns Hopkins University Press, Baltimore, MD

Price S, Price R 1999 *Maroon Arts: Cultural Vitality in the African Diaspora*. Beacon Press, Boston
Reis J J, dos Santos Gomes F (eds.) 1996 *Liberdade por um fio: Historia dos quilombos no Brasil*. Companhia das Letras, São Paulo
Scott D 1991 That event, this memory: Notes on the anthropology of African diasporas in the New World. *Diaspora* **1**(3): 261–84
Thoden van Velzen H U E, van Wetering W 1988 *The Great Father and the Danger: Religious Cults, Material Forces and Collective Fantasies in The World of the Surinamese Maroons*. Foris, Dordrecht, The Netherlands

R. Price

Marr, David (1945–80)

David Courtnay Marr was born on January 19, 1945 in Essex, England. He attended Rugby, the English public school, on a scholarship, and went on to Trinity College, Cambridge. By 1966, he obtained his B.S. and M.S. in mathematics, and proceeded to work on his doctorate in theoretical neuroscience, under the supervision of Giles Brindley. Having studied the literature for a year, Marr commenced writing his dissertation. The results, published in the form of three journal papers between 1969 and 1971, amounted to a theory of mammalian brain function, parts of which remain relevant to the present day, despite vast advances in neurobiology in the past three decades. Marr's theory was formulated in rigorous terms, yet was sufficiently concrete to be examined in view of the then available anatomical and physiological data. Between 1971 and 1972, Marr's attention shifted from general theory of the brain to the study of vision. In 1973, he joined the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology as a visiting scientist, taking on a faculty appointment in the Department of Psychology in 1977, where he was made a tenured full professor in 1980. In the winter of 1978 he was diagnosed with acute leukemia. David Marr died on November 17, 1980, in Cambridge, Massachusetts. His highly influential book, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, which has redefined and revitalized the study of human and machine vision, was published posthumously, in 1982.

1. *A Theory of the Brain*

Marr's initial work in neuroscience combined high-level theoretical speculation with meticulous synthesis of the anatomical data available at the time. The question he chose to address is the *nec plus ultra* of neuroscience: what is it that the brain does? Marr proposed a definite answer to this question for each of three major brain structures: archicortex (the phylogenetically older part of the cerebral cortex), cerebellum, and neocortex. The three answers complement each other, rallying around the idea that the brain's central function is statistical pattern recognition and association, carried out in a very high-dimensional space of 'elemental' features. The basic building block of all three theories is a *codon*, or a subset of features, with which there may be associated a cell, wired so as to fire in the presence of that particular codon.

In the first paper, Marr proposed that the cerebellum's task is to learn the motor skills involved in performing actions and maintaining posture (Marr 1969). The Purkinje cells in the cerebellar cortex, presumably implementing the codon representation, associate (through synaptic modification) a particular action with the context in which it is performed. Subsequently, the context alone causes the Purkinje cell to fire, which in turn precipitates the next elemental movement. Thirty years later, a significant proportion of researchers working on the cerebellum seem to consider this model as 'generally correct'—a striking exception in a field where the *nihil nisi bono* maxim is not known to be observed.

The next paper (Marr 1970) extended the codon theory to encompass a more general kind of statistical concept learning, which he assessed as 'capable of serving many of the aspects of the brain's function' (the vagueness of this aspect of the theory would lead him soon to abandon this approach, which, as he realized all along, was 'once removed from the description of any task the cerebrum might perform'). How can a mere handful of techniques for organizing information (such as the codon representation) support a general theory of the brain function? Marr's views in this matter are profoundly realist, and are based on a postulate of 'the prevalence in the world of a particular kind of statistical redundancy, which is characterized by a 'Fundamental Hypothesis,'' stating that 'Where instances of a particular collection of intrinsic properties (i.e., properties already, diagnosed from sensory information) tend to be grouped such that if some are present, most are, then other useful properties are likely to exist which generalize over such instances. Further, properties often are grouped in this way' (Marr 1970 pp. 150–51). These ideas presaged much of the later work by others on neural network models of brain function, which invoke the intuition of learning as optimization ('mountain climbing') in an underlying probabilistic representation space.

A model at whose core is the tallying of probabilities of events needs an extensive memory of a special kind, allowing retrieval based on the content, rather than the location, of the items. Marr's third theoretical paper considers the hippocampus as a candidate for fulfilling this function (Marr, 1971). In analyzing the memory capacity and the recall characteristics of the hippocampus, Marr integrated abstract mathematical

(combinatorial) constraints on the representational capabilities of codons with concrete data derived from the latest anatomical and electrophysiological studies. The paper postulated the involvement in learning of synaptic connections modifiable by experience—a notion originating with the work of Donald Hebb (see *Hebb, Donald Olding (1904–85)*) in the late 1940s and discussed by Marr's mentor Brindley in a 1969 paper. Marr provided a mathematical proof of efficient partial content-based recall by his model, and offered a functional interpretation of many anatomical structures in the hippocampus, along with concrete testable predictions. Many of these (such as the existence in the hippocampus of experience-modifiable synapses) were subsequently corroborated (see the reviews in Vaina 1990).

2. *The MIT Period*

A consummation of this three-prolonged effort to develop an integrated mathematical-neurobiological understanding of the brain would in any case have earned Marr a prominent place in a gallery, spanning two and a half centuries (from John Locke (see *Locke, John (1632–1704)*) to Kenneth Craik) of British Empiricism, the epistemological stance invariably most popular among neuroscientists. As it were, having abandoned the high-theory road soon after the publication of the hippocampus paper, Marr went on to make his major contribution to the understanding of the brain by essentially inventing a field and a mode of study: *computational neuroscience*. By 1972, the focus of his thinking in theoretical neurobiology shifted away from abstract theories of entire brain systems, following a realization that without an understanding of specific tasks and mechanisms—the issues from which his earlier theories were 'once removed'—any general theory would be glaringly incomplete.

Marr first expressed these views in public at an informal workshop on brain theory, organized in 1972 at the Boston University by Benjamin Kaminer. In his opening remarks, he suggested an 'inverse square law' for theoretical research, according to which the value of a study varies inversely with the square of its generality—an assessment that favors top-down reasoning anchored in functional (computational) understanding, along with bottom-up work grounded in an understanding of the mechanism, but not theories derived from intuition, or models built on second-hand data.

The new methodological stance developed by Marr following the shift in his views is summarized in a remarkably lucid and concise form in a two-page book review in *Science*, titled 'Approaches to Biological Information Processing' (Marr 1975). By that time, Marr came to believe firmly that the field of biological information processing had not yet accrued an empirical basis sufficient for guiding and supporting a principled search for a general theory. Remarking that the brain may turn out to admit 'of no general theories except ones so unspecific as to have only descriptive and not predictive powers'—a concern echoed in one of his last papers (Marr 1981)—he proceeded to mount a formidable critique of the most common of the theories circulated in the early, 1970s, such as catastrophe theory and neural nets (the current popularity of dynamical systems and of connectionism, taken along with the integration of Marr's critical views into the mainstream theoretical neurobiology, should fascinate any student of the history of ideas).

The main grounds for his argument, which was further shaped by an intensive and fruitful interaction with Tomaso Poggio (Marr and Poggio 1977), were provided by an observation that subsequently grew into a central legacy of Marr's career: the understanding of any information processing system is incomplete without insight into the problems it faces, and without a notion of the form that possible solutions to these problems can take. Marr and Poggio termed these two levels of understanding *computational* and *algorithmic*, placing them above the third, *implementational*, level, which, in the study of the brain, refers to the neuroanatomy and neurophysiology of the mechanisms of perception, cognition, and action.

Upon joining the MIT AI Lab, Marr embarked on a vigorous research program seeking computational insights into the working of the visual system, and putting them to the test of implementation as computer models. Marr's thinking in the transitional stage, at which he treated computational results on par with neurobiological findings, is exemplified by the paper on the estimation of lightness in the primate retina (Marr 1974); subsequently, much more weight was given in his work to top-down, computational-theory considerations. This last period in Marr's work is epitomized by the theory of binocular stereopsis, developed in collaboration with Poggio, and presented in a series of ground-breaking papers (Marr and Poggio 1976, Marr and Poggio 1979). At that time, Marr also worked on low-level image representation (Marr 1976, Marr and Hildreth 1980), and on shape and action categorization (Marr and Nishihara 1978, Marr and Vaina 1982). Marr's book, *Vision*, written during the last months of his life, is as much a summary of the views of what came to be known as the MIT school of computational neuroscience as it is a personal credo and a list of achievements of the second part of Marr's scientific endeavor, which lasted from about 1972 to 1980.

3. *Legacy*

The blend of insight, mathematical rigor, and deep knowledge of neurobiology that characterizes Marr's

work is reminiscent of the style of such titans of neuroscience as Warren McCulloch—except that McCulloch's most lasting results were produced in collaboration with a mathematician (Walter Pitts), whereas Marr did his own mathematics. A decade after his quest was cut short, it has been claimed both that Marr is cited more than he is understood (Willshaw and Buckingham 1990), and that his influence permeates theoretical neurobiology more than what one would guess from counting citations (McNaughton 1990). Still, contributors to the mainstream journals in neurobiology now routinely refer to the 'computations' carried out by the brain, and the most exciting developments are those prompted (or at least accompanied) by computational theories.

In computer vision (a branch of artificial intelligence), the influence of Marr's ideas has been complicated by the dominance of the top-down interpretation of his methodology: proceeding from a notion of what needs to be done towards the possible solutions. For some time, Marr's school was identified with the adherents of a particular computational theory of vision, which claims that constructing an internal model of the world is a prerequisite for carrying out any visual task. The accumulation of findings to the contrary in neurobiology and in the behavioral sciences gradually brought to the fore the possibility that vision does not require geometric reconstruction. This encouraged researchers to seek alternative theories, some of which employ concepts and techniques that did not exist in the 1970s, or were not known to the scholars of vision at the time. These new ideas, in turn, are making their way into neuroscience, as envisaged by Marr.

On a more general level, Marr's work provided a solid proof that a good theory in behavior and brain sciences need not have to trade off mathematical rigor for faithfulness to specific findings. More importantly, it emphasized the role of explanation over and above mere curve fitting, making it legitimate to ask *why* a particular brain process is taking place, and not merely what differential equation can describe it.

See also: Cognitive Neuroscience; Computational Neuroscience; Concept Learning and Representation: Models; Feature Representations in Cognitive Psychology; Information Processing Architectures: Fundamental Issues; Mental Representations, Psychology of; Neural Plasticity in Visual Cortex; Neural Representations of Objects; Perception and Action; Visual Perception, Neural Basis of; Visual System in the Brain

Bibliography

Marr D 1969 A theory of cerebellar cortex. *Journal of Physiology London* **202**: 437–70

Marr D 1970 A theory for cerebral neocortex. *Proceedings of the Royal Society of London B* **176**: 161–234

Marr D 1971 Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London B* **262**: 23–81

Marr D 1974 The computation of lightness by the primate retina. *Vision Research* **14**: 1377–88

Marr D 1975 Approaches to biological information processing. *Science* **190**: 875–6

Marr D 1976 Early processing of visual information. *Philosophical Transactions of the Royal Society of London B* **275**: 483–524

Marr D 1981 Artificial intelligence a personal view. In: Haugeland J (ed.) *Mind Design*. MIT Press, Cambridge, MA, Chap. 4, pp. 129–42

Marr D 1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco, CA

Marr D, Hildreth E 1980 Theory of edge detection. *Proceedings of the Royal Society of London B* **207**: 187–217

Marr D, Nishihara H K 1978 Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of the Royal Society of London B* **200**: 269–94

Marr D C, Poggio T 1976 Cooperative computation of stereo disparity. *Science* **194**: 283–7

Marr D C, Poggio T 1977 From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin* **15**: 470–91

Marr D, Poggio T 1979 A computational theory of human stereo vision. *Proceedings of the Royal Society of London B* **204**: 301–28

Marr D, Vaina L M 1982 Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London B* **214**: 501–24

McNaughton B L 1990 Commentary on simple memory: A theory of the archicortex. In: Vaina L M (ed.) *From the Retina to the Neocortex: Selected Papers of David Marr*. Birkhauser, Boston, MA, pp. 121–8

Vaina L M (ed.) 1990 *From the Retina to the Neocortex: Selected Papers of David Marr*. Birkhauser, Boston, MA

Willshaw D J, Buckingham J T 1990 An assessment of Marr's theory of the hippocampus as a temporary memory store. *Philosophical Transactions of the Royal Society of London B* **329**: 205–10

S. Edelman and L. M. Vaina

Marriage

1. *The Definition of Marriage*

Marriage has been a central area of study since the beginnings of anthropology, as a main factor in explaining the variety of kinship systems (Morgan 1870, Rivers 1914). The institution of marriage, however, has not been easy to define as an anthropological concept.

From a descriptive point of view it seems clear that when anthropologists refer to the institution of marriage, they designate a distinct class of phenomena. Yet it is extremely difficult to define marriage as a concept useful in accounting for all the ethnographic cases. Cohabitation, sexual access, affiliation of children, food sharing, and division of labor are very restricted criteria in accounting for the ethnographic spectrum. Let us examine, for example, the definition given in *Notes and Queries* (R. A. I. 1951, p. 110): 'Marriage is a union between a man and a woman such that children born to the woman are recognised legitimate offspring of both parents.' The first element (heterosexual union) does not conform to such ethnographic phenomena as the traditional woman-marriage among the Nuer (Evans-Pritchard 1951) or the homosexual marriages of postmodern societies (Weston 1991). Even if the second element (offspring legitimacy) apparently fits with the marriage customs of the matrilineal Nayar of South India, an anthropological test case for the definition of marriage (Gough 1959), it might be rejected for its vagueness and its limited range of ethnographic cases. As Bell (1997) notes, the statement that marriage is required to produce legitimate children is finalist and tautological. According to the ethnographic data, marriage is neither necessary nor sufficient to define the legitimacy of children and many societies recognize a sharp differentiation between social parenthood and marriage. Leach (1961), recognizing that marriage might be defined as 'a bundle of rights,' identified the following different rights: legal fatherhood, legal motherhood, monopoly of sexual access between married partners, right to domestic services and other forms of labor, right over property accruing to one's spouse, rights to a joint fund of property for the benefit of the children of marriage, and recognized relations of affinity such as that between brothers-in-law. But from this bundle of rights, no single right or set of rights might be defined as central to the universal definition of marriage. Needham (1974), similarly skeptical of a universal definition of marriage, viewed it as a useful word for descriptive purposes, because we know intuitively on which domain the ethnographer will focus, but troublesome when we want to define it, because we risk leaving out of the account the features that are central to the institution in any given society. Needham (1974, p. 44) concludes: 'So "marriage"… is an odd-job word: very handy in all sorts of descriptive sentences, but worse than misleading in comparison and of no real use at all in analysis.' Marriage defined cross-culturally is a concept based upon a serial likeness, with a family resemblance between the elements, rather than a univocal concept defined by a universal structural feature. It is a polythetic concept: a list of uses whose meaning depends upon the ethnographic context. As Needham (1974, p. 43) suggests, 'the comparison of marriage in different societies needs therefore to be contextual, and ultimately "total" in a Maussian sense, if we are to be sure that we understand what we are trying to compare.'

2. *Marriage Prestations*

For anthropologists marriage means the creation of new social relations, not only between husband and wife, but also between kin groups of both sides. As Radcliffe-Brown (1950, p. 43) considered it, 'marriage is essentially a rearrangement of social structure.' In most societies this kind of rearrangement is expressed by a ceremony or a ritual, and is sanctioned by marriage prestations. Two of the most common forms of marriage prestation are bridewealth and dowry. Bridewealth refers to the transfer of valuables from the bridegroom's kin group to that of the bride. In pastoral patrilineal societies such as that of the Nuer, the process of marriage involves handing over cattle from the family and kin of the bridegroom to the family and kin of the bride. By way of this process the children born in the marriage are attached to the lineage of the father who paid the bridewealth. They are 'the children of the cattle' (Evans-Pritchard 1951, p. 98). The marriage payments have not been interpreted as a form of purchase, but as a transfer of rights between kin groups. As Hutchinson (1996) remarks, even if the Nuer nowadays have entered into the sphere of commodification, and money and cattle are interchangeable, they distinguish between 'cattle of girls' which comes from the bridewealth and 'money of cattle' which comes from the market. Cattle remains the dominant metaphor of value; the cultural elaboration of the unique blood links uniting people and cattle allows cattle to enter into the sphere of kinship without the interference of the sphere of money. 'Cattle, like people, have blood,' but 'money has no blood,' say the late twentieth-century Nuer.

Whereas bridewealth always moves in the opposite direction to the bride, dowry moves in the same direction as the bride. It is the transfer of valuables from the bride's family to the bride herself. Due to the fact that marriage prestations have crucial political, economic, and ritual consequences for the society as a whole, they have formed one important class of phenomena to be compared in anthropological studies of marriage. Goody (1976) undertook a grand comparison between bridewealth societies of sub-Saharan Africa and dowry societies of Euro-Asia. The former practice pastoralism and simple agriculture, and the valuables exchanged through the bridewealth form a societal fund of the descent group. The latter have intensive agriculture and developed social stratification. The dowry is part of a conjugal fund, which moves down from parents to daughter, and its amount varies according to the status and wealth of the bride's family. In order to maintain stratification, the dowry system associated with the virginity of the women ensures homogamy and the control of daughters.

Another strategy has been the marriage of near kin in Mediterranean societies, which safeguards the control of property inside the family.

3. Gender and the Meaning of Marriage

Collier (1988) has demonstrated how different gender relations can be correlated with slightly distinct types of marriage prestations in classless societies. In societies with the bride service model, both men and women perceive themselves as autonomous agents. Nevertheless, for men, marriage is the precondition of adult status and enables them to acquire independence. For women, marriage marks a decline in their status and independence. While men consider marriage as an advantageous achievement, women are reluctant to marry. When they become mothers of marriageable daughters, they enjoy access to the labor services of their prospective sons-in-law. In societies with an equal bridewealth model, marriage is the moment when oppositions of gender and age are realized. Both men and women perceive themselves as having opposed and dangerous potentialities. It implies that young men and women depend upon senior kin for marriage arrangements. In societies with an unequal bridewealth model, marriage and rank are mutually defining. Men and women perceive to themselves as bearers of ranks.

According to gender perspectives, marriage involves different situations. Gender asymmetry may also entail the absence of a common name for marriage. As Aristotle stated, in the Greek language 'the union of a man and a woman has no name.' Benveniste demonstrated (1969) that in the Indo-European languages the idea of marriage has no common name for men and women. For a man the terms have a verbal root. He is the agent of an action as he carries a woman to his house. In contrast, for a woman there is no one verb denoting her marriage. She is not the subject of an action; she just changes condition.

4. Marriage Alliance

Lévi-Strauss (1967) has made the major contribution to the study of marriage as the core of the kinship systems. He argues that the rules governing the prohibition of incest lead to exogamy, which in turn produces exchange and reciprocity between kinship groups. 'Marriage alliance' refers to the repetition of intermarriages between exchange units; with cross-cousin marriage as the most elementary form of the system of alliances. One marries into a specific category of the kinship system, which does not distinguish between consanguines and affines. In this category of kinship the potential spouses come from different exchange units, and the system of relationships assures the repetition of alliances. Lévi-Strauss' theory explains the complicated system of marriage classes of the Australian aborigines, the Dravidian system of South India as well as the kinship systems of Asia. The *tour de force* of the theory has been the explanation of so-called Crow–Omaha systems (where there are no positive rules of marriage, but many negative rules applicable to allied descent groups) and the complex systems where there are only negative rules related to some kin degrees. Arab marriage—marriage with the father's brother's daughter—could appear to be out of the scope of the theory, when analyzed as the mirror image of the elementary forms of marriage—marriage with a cross-cousin. When it began to be analyzed as a form of marriage with a near kin degree, however, it could be considered as a case of a complex system of alliance. Héritier (1981) has solved the Crow–Omaha problem and tried to understand the complexities of alliances. Simple exchanges are compatible with limitations of repetition of alliances between groups. The analysis done in some European peasant communities has demonstrated the recurrence of some elementary forms, as well as the social limits of individual marriage choices that have been done between two poles: neither too remote in terms of social status or class nor too close in terms of kin degrees.

5. The Union of Individual Persons and Alliance Systems

In relation to the anthropological perspective of analyzing marriage as an alliance system, marriage in Western societies, as Wolfram (1987) stated, was not traditionally conceptualized as an alliance between groups, but as a union of individual spouses. Marriage creates a relationship between kin on both sides who become connected to each other. Each becomes the 'in-law' relation of the other side, but marriage does not consist of an alliance between both sides. It consists of the union of the spouses, who become 'one flesh.' This view is supposed to have its origin in the Christian concept of marriage. It is probably unique to Western societies and it might be regarded as the centerpoint of the Western kinship system. The Church doctrine of 'one flesh' presumed that the two spouses became 'one person.' Due to the gender asymmetry, the wife was only a part of the husband's person. This relational view of marriage has been replaced by the idea of a couple constituted by two individual persons, and kinship is composed by individual persons who exist prior to any relationship. In non-Western societies kinship constitutes relationships, and individuals are icons of these relationships. Marriage is not the union of individual persons who create new relations, but alliances and flows of prestations between persons who embody diverse relations.

See also: Divorce and Gender; Divorce, Sociology of; Family and Kinship, History of; Family as Institution;

Family Law; Kinship in Anthropology; Property: Legal Aspects

Bibliography

Bell D 1997 Defining marriage and legitimacy. *Current Anthropology* **38**(2): 237–53

Benveniste E 1969 L'expression indo-européenne du 'marriage.' In: Benveniste E (ed.) *Le vocabulaire des institutions Indo-européennes*. Minuit, Paris, Vol. 1

Collier J F 1988 *Marriage and Inequality in Classless Societies*. Stanford University Press, Stanford, CA

Evans-Pritchard E E 1951 *Kinship and Marriage Among the Nuer*. Clarendon Press, Oxford, UK

Goody J 1976 *Production and Reproduction: A Comparative Study of the Domestic Domain*. Cambridge University Press, Cambridge, UK

Gough E K 1959 The Nayars and the definition of marriage. *Journal of the Royal Anthropological Institute* **89**: 23–34

Héritier F 1981 *L'exercise de la parenté*. Hautes Études-Gallimard, Le Seuil, Paris

Hutchinson S E 1996 *Nuer Dilemmas. Coping with Money, War, and the State*. University of California Press, Berkeley, CA

Leach E R 1961 *Rethinking Anthropology*. Athlone Press, London

Lévi-Strauss C 1967 *Les structures eléméntaires de la parenté*, rev. edn. Mouton, Paris

Morgan L H 1870 *Systems of Consanguinity and Affinity of the Human Family*. Smithsonian Institution, Washington, DC

Needham R 1974 *Remarks and Inventions: Skeptical Essays about Kinship*. Tavistock, London

Radcliffe-Brown A R 1950 Introduction. In: Radcliffe-Brown A R, Forde D (eds.) *African Systems of Kinship and Marriage*. Oxford University Press, London

R.A.I. 1951 (Royal Anthropological Institute of Great Britain and Ireland) *Notes and Queries on Anthropology*, 6th edn. Routledge and Kegan Paul, London

Rivers W H R 1914 *Kinship and Social Organisation*. Constable, London

Weston K 1991 *Families We Choose: Lesbians, Gays, Kinship*. Columbia University Press, New York

Wolfram S 1987 *In-Laws and Outlaws. Kinship and Marriage in England*. St. Martin's, New York

J. Bestard-Camps

Marriage and the Dual-career Family: Cultural Concerns

1. Changes in Marriage

The dual-career family begins with marriage, which has changed in important ways over the last century. In the US, men and women are delaying marriage into their mid-to-late twenties, often entering cohabitation first. Divorce rates are high and stable, but rates of remarriage have fallen, so that a larger proportion of adults are unmarried now than in the past. In 1970, unmarried people made up 28 percent of the adult population. In 1996, 40 percent of all adults were unmarried. Seventy-one percent of women born in the early 1950s had married by age 25, compared to 54 percent of those born in the late 1960s (Raley 2000). In fact, the shift away from marriage has been so dramatic for blacks that now a majority of black men and women are not married, compared to about a third of white men and women (Waite 1995). Similar changes in marriage patterns have taken place in most European countries; recent cohorts are marrying at older ages and over a wider range of ages than in the past.

In addition, European countries differ substantially in marriage ages. The Nordic countries of Sweden, Denmark, and Iceland show the highest average ages at marriage for women (around age 29) and the Eastern European countries of Bulgaria, the Czech Republic, Hungary, and Poland the lowest (around age 22). Since societies with relatively high age at marriage also tend to be those in which many people never marry, this diversity suggests that marriage is a more salient component of family in some European countries than others (Kiernan 2000).

Countries in Europe also show a great deal of variation in the proportion of women in marital unions. Marriage is most common in Greece and Portugal, where over 60 percent of women ages 25–29 are married, and least common in the Nordic countries, Italy, and Spain where a third or less are.

In fact, women's increasing commitment to employment and their own career is an often-cited reason for delayed marriage, nonmarriage, and the rise of cohabitation; marriage has declined because women have become economically independent of it. As more women work for pay and as their earnings increase, according to this argument, the economic benefits of marriage fall and divorcing or remaining single becomes both a more feasible and a more attractive option (Becker 1981, McLanahan and Casper 1995). Declines in the earnings of young men, relative to young women, this reasoning goes, also reduce the economic attractiveness of marriage (Oppenheimer and Lew 1995).

An alternative argument sees marriage as beneficial even for women with earnings adequate to support themselves (and for men regardless of women's earnings). Marriage improves the economic well being of its members through the economies of scale, returns to specialization, and the risk sharing it brings. The earnings of wives add to family income, provide insurance against the risk of economic shortfall, and allow families to adjust income to changes in economic needs as children are born and grow to adulthood. These two-earner marriages also provide companionship and collaboration, with spouses sometimes each playing a unique role and sometimes playing similar roles, depending on their abilities and the needs of the family (Goldscheider and Waite 1991, Nock 1998, Oppenheimer 1994).

2. Women's Employment

The rise in women's labor force participation in many countries over the last half-century has been so dramatic as to constitute a revolution in women's lives and roles. In the US, women workers now make up just under half of the labor force. In the early 1960s, only one woman in three was working full time, compared to 86 percent of men. An additional 16 percent of women worked part time. Half of women did not hold a paid job at all. But the revolution had already begun. Women moved from work in the home to work in the office or factory, slowly at first, and then more quickly. Between 1963 and 1975 both women's full time and part time work increased. The shift from work in the home into part time paid employment pretty much stops by the mid 1970s, so that almost all the growth in the 1980s and the 1990s comes from a rapid increase in women's full time work. By 1997, 57 percent of all women were working full time, with another 23 percent working part time. The share of US women who did not work at all for pay shrank to just one in five in 1997—four out of five adult women now hold paying jobs, compared to one in two in 1963.

Many countries have seen similar expansions of women's employment, with very high proportions of women in the labor force in a number of developing countries, such as China, Vietnam, and Bangladesh. Islamic countries generally have low to moderate levels of women's employment (20–50 percent) and Asian countries moderate to high levels (50–80 percent) (United Nations 2000). Within Europe, the Nordic countries show moderately high levels of economic activity among women (60–65 percent), and the Southern European countries moderately low levels (30–40 percent). Married women—especially married mothers—were the shock troops in the revolution in women's employment, dramatically altering their allocation of time to work for pay vs. work in the home. Employment of single women without children changed little over the last 30 or 40 years. These women were about as likely to work full-time in 1997 as in 1963, a little more likely to be working part-time and a little less likely not to be working. Single mothers changed their work behavior more but not dramatically—58 percent worked full time in 1997 compared to half in 1963. One single mother in three did not work for pay in 1963, compared to one in five in 1997.

But changes in women's employment were driven by the choices made by married women, both because many more women are married than not, and because the work choices of married women changed much more than the work choices of unmarried women. Both married mothers and married women without children are much more likely to work for pay and to work full time now than in the early 1960s. In 1963, 41 percent of married women with no children at home worked full time compared to 60 percent in 1997 and the share not working outside the home fell from 43 percent to 19 percent. But married mothers were quite unlikely to work full time in 1963, when fewer than one in four worked this much. By 1997, the proportion working full time had more than doubled to 49 percent. Sixty percent of married mothers did not work outside the home in 1963; by 1997 less than one in four was not employed (Waite and Nielsen 2001). Changes in the employment of married women were similar to the US in Canada and Australia, even more dramatic in the Nordic countries, where over 80 percent of wives are in the labor force, and more modest in Germany, Belgium, and The Netherlands (Spain and Bianchi 1996). In all these countries, the employment of married women creates dual-career couples, while the employment of married mothers forms dual-career families.

3. Working Families

Perhaps as dramatic and far reaching as the alternations in the structure of the family are the changes in the way its members use their time. In the early 1960s in the US, among those in the prime working ages most married couples followed the male breadwinner/female homemaker model; 56 percent had only one earner. The dual-income family was uncommon—both spouses worked full-time in 21 percent of married couples. By 1997, only a quarter of married couples had one earner. In 44 percent of married couples both spouses worked full time, and in another 24 percent one worked full-time and one part-time.

The shift toward the dual-worker family was even more dramatic for couples with children (Waite and Nielsen 2001). Even 30 years ago, most children living with a single parent did not have a parent at home full-time; now most children in married couple families don't either. Among children in married parent families, 60 percent lived in a breadwinner/homemaker family in 1963. By 1997 the situation had reversed so that 57 percent lived in families that had more than one earner, most often both working full-time. Families with a stay-at-home mother became a minority and dual-worker families the most popular choice. Families gained the income the wife earned on the job but lost her time in the home.

4. Work in the Home

Of course, the adults in families have almost always worked, but the shift in the site of women's labor from inside to outside the home changed many other things for families. First, families had more money. Second, families had less of the woman's time in the home. For a while, women in the US seemed to bear the brunt of this change, adding paid employment to their household management and maintenance tasks. But recent evidence suggests that US women—both employed

and not employed—have reduced and men have increased the number of hours that they spend on housework, so that, while not equal, the time contributions of men and women are much closer than they were in the past. For all men and women, the weekly housework gap has fallen from 25 hours in 1965 to 7.5 hours in 1995, and the housework gap between married men and women from 29 to 9 hours per week. The housework gap is smaller in dual-earner than in one-earner couples, because as the wife's hours of paid work increase, her own hours of housework decline whereas her husband's housework hours increase (Bianchi et al. 2000).

Although information on time spent on household work is available for only a small number of countries, the pattern is consistent; women spend at least twice as much time on housework, on average, as men do, even in countries like Latvia and Lithuania where hours of paid employment are high for both men and women. Women's typically lower hours of paid employment and higher hours of work in the household mean that, in many countries, including Japan, Korea, Australia, The Netherlands, and Norway, total hours of paid and unpaid work approximately balance, on average, if not in any particular family. But in other countries, like Spain, Austria, the United Kingdom, Italy, and Lithuania, women work five or more total hours, on average, more than men do, either because their average hours of housework are quite high or because they work long hours at both paid and unpaid work (United Nations 2000, Table 5–6A).

4.1 *Family Hours of Work*

Although Americans feel very short of time, the average employed adult today works about the same number of hours as the average employed adult did 30 years ago. But the time deficit is real; a higher proportion of people are working very long hours (balanced by more people working short hours) and from many more couples jointly contributing 80–100 hours to work per week (Jacobs and Gerson 2001). This latter change has come about as a direct result of the movement of married women into full-time jobs, with the consequent shift from breadwinner/homemaker to dual-earner families. Family hours of paid work are also quite high in the Nordic countries, where a high proportion of married women are employed, and in Latvia, Lithuania, Bulgaria, and Hungary, countries in which women work, on average, quite substantial numbers of hours (United Nations 2000).

5. *The Impact of Marriage on Career*

For most adults, work life and family life are intertwined. The career one chooses defines, in part, the kind of family life one can have. Careers that demand long hours, frequent travel, irregular or rotating schedules, or overtime work with little or no notice constrain time with family and may also affect the quality of that time (Hochschild 1997, Presser 1997). Careers that carry social status, high incomes, health and retirement benefits, and interesting work bring financial and psychological resources to the families of the workers as well as to the workers themselves. Employment and career also affect the chances that people get and remain married. At the same time, marriage affects career choices and success, and family demands may lead men and women to choose different jobs than they otherwise would, to move for a spouse's job, or to work more or less than they would prefer. As might be expected, however, the relationship between the career and family choices is quite different for men and women.

5.1 *Marriage and Men's Careers*

Marriage is good for men's careers. When men marry they tend to earn more, work more, and have better jobs (Nock 1998). The 'marriage premium' in men's earnings seems to be at least five percent at younger ages and considerably more at older ages (Korenman and Neumark 1991). Nor is men's marriage premium a strictly American phenomenon; it appears in the vast majority of developed countries and is generally quite sizeable (Schoeni 1995). Married men are generally more productive than unmarried men, receive higher performance ratings, and are more likely to be promoted (Korenman and Neumark 1991). Scholars link this higher productivity to the more orderly lives of married than single men, to married men's motivation to support their families and the increased work effort this produces, and to investments by the wife in the husband's career (Grossbard-Shechtman 1993).

At the same time, better career prospects and career achievements foster marriage and marital stability for men. Higher-earning men are more likely to get married and marry at younger ages than men with lower earnings (Bergstrom and Schoeni 1996). And high-earning men are less likely to divorce than those with lower earnings, with unemployment especially damaging to marital stability (Grossbard-Shechtman 1993).

5.2 *Marriage and Women's Careers*

While marriage improves men's career attainments, for women it often triggers tradeoffs between work and family that lead to a decline in employment, hours of work, and earnings. Many of these changes appear not at marriage but with the birth of the first child, when some women increase their time at home to

provide care for the baby (Klerman and Leibowitz 1999). Even when mothers work continuously, the demands of childrearing detract from their earning capacity. Comparing US women with similar work histories, Waldfogel (1997) finds that one child still reduces a woman's earnings by almost four percent and two children or more reduces hourly earnings by almost 12 percent. Women do not pay a marriage penalty, but they do pay a substantial motherhood penalty, whether or not they marry.

Marriage and motherhood also seem to decrease the chances that women follow a career trajectory. Even among middle-class, dual-earner couples, most do not pursue two demanding careers but instead typically reduce and scale back their commitment to paid work to mitigate the impact of the job on family life. Often but not always, the wife does the cutting back, generally when children are young but also at other points in the life course (Becker and Moen 1999). As a result, relatively few women manage to achieve both career and family. Women with successful work careers less often marry or remain married than other women (Blair-Loy 1999, Han and Moen 1999). And the substantial majority of women with children, even those with college degrees, do not achieve career success. Among women who graduated from college in the late 1960s through the late 1970s, Goldin (1997) estimates that between 18 percent and 24 percent achieved both a career and a family.

6. *Financial Well-being*

Married couples typically have both higher incomes and greater wealth than unmarried individuals, even after income is adjusted to take into account the number of people in the family (Lupton and Smith forthcoming, Waite and Nielsen 2001). Among married couples, those with two earners have higher incomes, on average, than those with one-earner. In the US, two-earner families have pulled ahead economically as their incomes have increased faster than the incomes of one-earner families. In 1970, the median income of dual-earner families was 1.32 times as high as that of married couple families in which the wife did not work for pay (Spain and Bianchi 1996). In 1998, the ratio was 1.79 (United States Bureau of the Census 2000). But the higher income of two-earner families must be adjusted for the loss of home production. One estimate suggests that dual-earner families need about 35 percent more income to have the same standard of living as families with one spouse—almost always the wife—working full time in the home, to make up for the goods and services produced at home and for clothes, transportation, and other costs of employment (Lazear and Michael 1988).

Women and men are marrying later and spending more time unmarried than in the past. In many countries, levels of education have risen dramatically, especially for women, at the same time that family size has fallen. These changes reduce the need for women's time at home and increase the rewards for their time in paid employment, pushing them and their families toward dual-worker and working parent families.

See also: Career Development, Psychology of; Family as Institution; Family Processes; Family Size Preferences; Family Theory: Economics of Childbearing; Family Theory: Feminist–Economist Critique; Marriage; Motherhood: Economic Aspects

Bibliography

Becker G S 1981 *A Treatise on the Family*. University of Chicago, Chicago, IL

Becker G S, Moen P 1999 Scaling back: Dual-earner couples' work–family stategies. *Journal of Marriage and the Family* **61**: 995–1007

Bergstrom T, Schoeni R F 1996 Income prospects and age at marriage. *Journal of Population Economics* **9**: 115–30

Bianchi S, Milkie S A, Sayer L C, Robinson J P 2000 Is anyone doing the housework? Trends in the gender division of household labor. *Social Forces* **79**: 191–228

Blair-Loy M 1999 Career patterns of executive women in finance: An optimal matching analysis. *American Journal of Sociology* **104**: 1346–97

Jacobs J A, Gerson K 2001 Overworked individuals or overworked families? Explaining trends in work, leisure, and family time. *Work and Occupations*. **28**: 40–63

Goldin C 1997 Career and family: College women look to the past. In: Blau F D, Ehrenberg R G (eds.) *Gender and Family Issues in the Workplace*. Russell Sage, New York, pp. 20–64

Goldscheider F K, Waite L J 1991 New families, no families? In: *The Transformation of the American Home*. University of California Press, CA

Grossbard-Shechtman S 1993 *On the Economics of Marriage: A Theory of Marriage, Labor, and Divorce*. Westview Press, Boulder, CO

Han S-K, Moen P 2001 Coupled careers: pathways through work and marriage in the United States. In: Blossfeld H P, Drobnic S (eds.) *Careers of Couples in Contemporary Societies: A Cross-National Comparison of the Transition from Male Breadwinner to Dual-Earner Families*. Oxford University Press, Oxford, UK

Hochschild A 1997 *The Time to Bind: When Work Becomes Home and Home Becomes Work*. Metropolitan Books, New York

Kiernan K 2000 European perspectives on union formation. In: Waite L, Bachrach C, Hindin M, Thomson E, Thornton A (eds.) *Ties that Bind: Perspectives on Marriage and Cohabitation*. Aldine de Gruyter, New York, pp. 40–58

Klerman J, Leibowitz A 1999 Job continuity among new mothers. *Demography* **36**: 145–55

Korenman S, Neumark D 1991 Does marriage really make men more productive? *Journal of Human Resources* **26**: 282–307

Lazear E, Michael R T *Allocation of Income within the Household*. University of Chicago Press, Chicago, IL

Lupton J, Smith J P in press Marriage, assets, and savings. In: Grossbard-Shechtman S (ed.) *Marriage and the Economy*. Cambridge University Press, Cambridge, UK

McLanahan S, Casper L 1995 Growing diversity and inequality in the American family. In: Farley R (ed.) *State of the Union:*

America in the 1990s. Russell Sage, New York, Vol. 2, pp. 1–46
Nock S L 1998 *Marriage in Men's* Lives. Oxford University Press, New York
Oppenheimer V K, Lew V 1995 American marriage formation in the 1980s: How important was women's economic independence? In: Oppenheim Mason K, Jensen A-M (eds.) *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK, pp. 105–38
Oppenheimer V K 1994 Women's rising employment and the future of the family in industrial societies. *Population and Development Review* **20**: 293–342
Presser H B 1994 Employment schedule among dual-earner spouses and the division of household labor by gender. *American Sociological Review* **59**: 348–64
Raley R K 2000 Recent trends in marriage and cohabitation. In: Waite L, Bachrach C, Hindin M, Thomson E, Thornton A (eds.) *Ties that Bind: Perspectives on Marriage and Cohabitation*. Aldine de Gruyter, New York, pp. 19–39
Schoeni R F 1995 Marital status and earnings in developed countries. *Journal of Population Economics* **8**: 351–9
Spain D, Bianchi S M 1996 *Balancing Act: Motherhood, Marriage and Employment Among American Women*. Russell Sage, New York
United Nations 2000 The World's Women 2000: Trends and Statistics. Table 5–1 United Nations, New York
US Bureau of the Census 2000 www.census.gov/hhes/income/histinc/f013.html
Waite L J 1995 Does marriage matter? *Demography* **32**: 483–508
Waite L J, Haggstrom G W, Kanouse D E 1985 Changes in the employment activities of new parents. *American Sociological Review* **50**: 263–72
Waite L J, Nielsen M 2001 The rise of the dual-worker family: 1963–1997. In: Hertz R, Marshall N (eds.) *Women and Work in the Twentieth Century*. University of California Press, Berkeley, CA
Waldfogel J 1997 The effect of children on women's wages. *American Sociological Review* **62**: 209–17

L. J. Waite

Marriage: Psychological and Experimental Analyses

This article reviews psychological and experimental analysis that provides important insights into the nature of, and influences on, satisfying couple relationships. In addition, there is a description of the application of this knowledge to relationship education for promoting satisfying relationships, and to couple therapy for improving distressed relationships.

1. *Significance of Couple Relationship Satisfaction and Stability*

Most people have an intimate couple relationship at some point in their lives. In western countries, over 90 percent of the population marry by the age of 50 years (McDonald 1995). Even amongst those who choose not to marry, the vast majority engage in marriage-like, committed couple relationships (McDonald 1995). Couples who sustain mutually satisfying relationships experience many benefits. Relative to other people, those in satisfying marriages have lower rates of psychological distress, higher rated life happiness, and greater resistance to the detrimental effects of negative life events (Halford 2000).

Almost all couples report high relationship satisfaction at the time of marriage, but average satisfaction levels deteriorate across the first ten or so years of marriage (Glenn 1998). Around this average trend there is great variability between couples, with some couples sustaining high relationship satisfaction across their life, and others experiencing severe relationship dissatisfaction. Decreased satisfaction is associated with high risk of separation (Gottman 1994). About 42 percent of Australian marriages, 55 percent of American marriages, 42 percent of English marriages and 37 percent of German marriages end in divorce, and about half of these divorces occur in the first seven years of marriage (McDonald 1995).

2. *The Nature of Stable, Satisfying Relationships*

The nature of stable, satisfying relationships varies by culture. For example, the acceptability of particular gender roles is quite different across cultures. Rather than attempting to define how relationships should be, psychologists measure *relationship satisfaction* with standardized questionnaires. While couples vary in what they want from relationships, within western cultures we can define some general characteristics of stable, mutually satisfying relationships.

2.1 *Conflict Management*

No two people can agree on everything, hence conflict must be managed. Both independent observers and spouses report that *effective conflict management* is associated with relationship satisfaction (Weiss and Heyman 1997). When discussing conflict issues, dissatisfied partners often are hostile and critical, they negatively demand change of each other, and do not listen to their partner (Weiss and Heyman 1997). As might be expected, dissatisfied couples find this conflict aversive, and often avoid or withdraw from problem discussions (Gottman 1994).

Couples often develop unhelpful *patterns of conflict management*. Two important examples are the demand-withdraw and mutual avoidance patterns (Weiss and Heyman 1997). In the demand-withdraw pattern, one partner criticizes and demands change, while the spouse attempts to withdraw from the discussion. In mutual avoidance, both partners avoid

discussion of particular conflict topics. Both patterns are associated with a failure to effectively resolve the relationship conflict, which often results in the issue remaining unresolved and being a source of future conflict.

In contrast to dissatisfied couples, satisfied couples are less reactive at an emotional level to their partner's negativity during conflict. Satisfied couples report less anger, and are less likely to reciprocate with hostile, negative behavior in discussions than dissatisfied couples (e.g., Gottman 1994). Relationship distress is associated with high levels of physiological arousal (e.g., elevated heart rate, high stress hormone levels) during interaction (Gottman 1994). This arousal is assumed to be aversive, which may explain the higher rates of withdrawal during problem-focused discussions by distressed partners (Gottman 1994).

Violence is common between partners in intimate relationships. About 30 percent of young couples report at least one incident in which one partner struck the other in the previous year (O'Leary 1999). Violence is most common in couples who are dissatisfied with their relationship; approximately two thirds of distressed couples have been violent in the year prior to their presentation for therapy (O'Leary 1999). While men and women are approximately equally likely to strike each other, when women are the victims of partner violence they are much more likely than men to be injured, to report feeling fear of their partner, and to show psychological disorder (O'Leary 1999).

2.2 Communication and Support

Couples who are satisfied in their relationship report that they communicate well. They talk to each other about positive feelings they have about each other and the relationship (Weiss and Heyman 1997), as well as about things that happen in their day-to-day lives. Over time these conversations build a sense of intimacy. Satisfied couples also support each other. This shows both in emotional support provided through listening carefully to the partner and offering advice when requested, and in practical support, which is provided to assist with life challenges (Bradbury 1998).

2.3 Behavior

Couples in satisfying relationships do more positive things and fewer negative things than distressed couples (Weiss and Heyman 1997). For example, they are more likely to express affection, provide support, state appreciation of the actions of their partner, have enjoyable discussion, and share household tasks. Distressed couples tend to be positive only if their partner recently has been positive. In addition, if one distressed partner behaves negatively, the other often responds negatively immediately (Weiss and Heyman 1997). In contrast, satisfied couples tend to be positive irrespective of their partner's prior actions.

Satisfied couples sustain a balance of shared, enjoyable couple activities and independent interests. Relative to distressed couples, satisfied couples share positive activities more often, and seek out new, shared activities that enhance their sense of intimacy (Baumeister and Bratlavsky 1999). At the same time, satisfied partners also sustain a range of activities and interests independent of their spouse. In contrast, distressed couples typically repeat a narrow range of activities, and often do not have a balance of individual and shared activities, spending either very little or almost all of their free time together (Baumeister and Bratlavsky 1999).

2.4 Perceptual Biases

People are not reliable or objective observers of their relationship. Distressed couples selectively attend to their partner's negative behavior and selectively recall that negative behavior (Weiss and Heyman 1997). In contrast, satisfied partners tend to overlook negative behaviors by their spouse, to have an unrealistically positive view of their partner, and to selectively recall positive aspects of relationship interaction (Weiss and Heyman 1997).

2.5 Beliefs and Thoughts about Relationships

Another characteristic of satisfied couples is holding positive, realistic beliefs about their relationship and their partner. Satisfied couples think about their relationship and its maintenance as a shared challenge, and express a sense of shared goals, affection, commitment and respect within the relationship (Carrere et al. 2000). In contrast, distressed couples focus on their individual needs, and express disappointment, disillusionment and a sense of chaos in their relationship (Carrere et al. 2000). Distressed couples are more likely than satisfied couples to believe that any form of disagreement is destructive, that change by partners is not possible, and that rigid adherence to traditional gender roles is desirable (Baucom and Epstein 1990).

When there are problems in the relationship, distressed couples attribute the causes of those problems to stable, internal, and negative characteristics of their partner (Baucom and Epstein 1990). For example, a partner arriving home late from work may be perceived as 'a generally selfish person who doesn't care about the family' by a distressed partner. The same behavior may be attributed by a maritally satisfied partner as the spouse 'struggling to keep up with a heavy load at work, and being subject to lots of pressure from the boss.' The process of attributing relationship problems to the partner leaves many

people in distressed relationships feeling powerless to improve their relationship.

3. Influences on Relationship Outcomes

Four broad classes of variables seem to impact upon the trajectory of relationship satisfaction and stability: contextual variables, individual characteristics, couple interaction, and life events (Glenn 1998; Karney and Bradbury 1995, Larson and Holman 1994).

3.1 Contextual Variables

Contextual variables are the cultural and social circumstances within which couple relationships exist. High levels of support of the relationship by family and friends, being active in a religion that supports marriage, and living in a country with a low divorce rate and a strong valuing of marriage all predict relationship satisfaction and stability (Larson and Holman 1994). Social disadvantage in the form of low education, poverty, or being a member of an oppressed ethnic group is associated with higher rates of relationship problems and divorce (Larson and Holman 1994). Work, parenting, and friendships all have the capacity to enrich or to cause problems for couple relationships. For example, employment that provides opportunities for utilization of skills and flexibility in meeting family needs is associated with higher relationship satisfaction than employment that is unsatisfying, or associated with work-family role conflicts (Larson and Holman 1994).

3.2 Individual Characteristics

Individual characteristics are relatively stable individual differences that partners bring to the relationship, such as personality traits, psychological disorder, and the effects of personal history. Certain personality traits are associated with increased likelihood of having satisfying relationships. For example, low *neuroticism* and secure *attachment style* are associated with low risk of relationship problems (Karney and Bradbury 1995, Feeney and Noller 1996). Neuroticism is a relatively stable tendency to perceive the world as threatening, and to experience negative emotions such as anxiety and low mood. Attachment style is a relatively stable individual difference in the extent to which people feel anxious about being abandoned by those close to them, or uncomfortable with emotional closeness. Attachment style is the general way a person thinks and responds emotionally in close relationships. Attachment style is believed to be influenced greatly by early experiences within the family of origin (Feeney and Noller 1996).

Psychological disorder in one or both partners increases the risk of relationship problems. Alcohol abuse, depression, and certain anxiety disorders all are commonly associated with relationship problems (Halford et al. 1999). In some cases the psychological disorder seems to cause the relationship problem. For example, alcohol abuse by one partner early in the relationship predicts later relationship violence and deterioration in relationship satisfaction (Halford et al. 1999). In other cases, the relationship problems seem to exacerbate the disorder. For example, marital conflict predicts relapse to problem drinking by men who have previously brought problem drinking under control (Halford et al. 1999).

Personal history is another individual characteristic that predicts relationship satisfaction. For example, experiencing parental divorce or inter-parental violence as a child predicts decreased relationship satisfaction and increased risk of divorce as an adult (Larson and Holman 1994). Also, having been divorced oneself predicts increased risk of problems in the current relationship.

3.3 Couple Interaction

Couple interaction refers to the behavior, thoughts and feelings of partners when they interact. Effective conflict management, good communication, mutual partner support, positive behavior toward each other, shared enjoyable activities and realistic relationship expectations not only are part of current relationship satisfaction, these couple interaction processes also predict future relationship satisfaction and stability (Bradbury 1998, Karney and Bradbury 1995). When trying to promote positive couple relationships, couple interaction is particularly important as it is something that can be changed. For example, more effective communication can be taught to couples.

3.4 Life Events

Life events are the major events and transitions that happen to the couple during the course of their relationship. Life events include both normal changes that most couples experience, such as the birth of a child or a change of job, as well as major, less predictable events such as death of young people in the family, being a victim of crime, or loss of employment. Relationship satisfaction changes most when life events are occurring. For example, when a couple has their first child many couples report either substantial increases or decreases in their relationship satisfaction.

3.5 Influences on Relationships and Risk for Relationship Problems

Context, individual characteristics, couple interaction and life events interact in their influence on the course of couple relationship satisfaction and stability. For

example, low mutual support by partners has the most negative impact upon relationship satisfaction when the couple experience challenging life events (Bradbury 1998). This is particularly so when a contextual factor is operating as well, for example when there are few other social supports available through extended family or friends.

As context, individual characteristics, and couple interaction can be assessed early in a committed relationship, it is possible to assess, for individual couples, indicators of *relationship risk* for developing future problems. Risk level in turn predicts relationship distress and/or separation in the early years of marriage (Bradbury 1998, Gottman 1994). The prediction of future relationship satisfaction is not perfect, both because our understanding of the influences on couple relationships is imprecise, and because we cannot know the future changes in context or life events that will impact upon a couple. However, by identifying *high-risk couples* early in their relationships, it is possible to offer them relationship education to improve their relationship outcomes.

4. Relationship Education

Across many Western countries, relationship education programs are available to marrying and cohabiting couples (van Widenfelt et al. 1997). These programs are intended to assist couples to sustain satisfying relationships and to reduce divorce rates (van Widenfelt et al. 1997). Most relationship education programs are offered by religious and community groups, the content of these programs is often not documented, and the effects of most of these programs have not been evaluated (Halford 2000). Relationship education programs which have been evaluated focus either on inventory-based or skills-based education (Halford 2000).

4.1 Inventory-based Relationship Education

Inventory-based relationship education is a widely used approach in which partners complete standardized self-report forms that assess relationship expectations, beliefs and current patterns of interactions (Halford 2000). The most widely used inventories are PREPARE and Facilitating Open Couples Communication (FOCCUS). In both PREPARE and FOCCUS, trained relationship educators review the partners' individual answers with the couple, helping them to identify areas of agreement and relationship strength, and areas of disagreement and potential difficulty. Sometimes these insights are supplemented with further discussion and exercises to help the couple develop shared, positive, and realistic relationship beliefs and expectations.

The answers partners give to PREPARE and FOCCUS predict the couples' relationship satisfaction and stability over the first few years of marriage (Halford 2000). However, there is no adequate scientific demonstration that participating in PREPARE, FOCCUS or other similar inventory-based relationship education enhances future relationship satisfaction or stability.

4.2 Skills-based Relationship Education

There are several skills-based relationship education programs, and they all place emphasis upon developing couples' conflict management and communication skills. For example, the Guerney Relationship Enhancement Program and the Prevention and Relationship Enhancement Program (PREP) both place emphasis upon training couples in these skills (Halford 2000). At the same time, it is recognized that successful relationships require more than just effective conflict management and good communication. Consequently, skills-training relationship education includes other content such as promoting mutual partner support, expressing affection, and having fun (Markman et al. 1994).

Five to six sessions of skills-training relationship education reliably improves couple conflict management and communication (Markman and Hahlweg 1993). These changes are maintained for months or even years after completing the education program (Markman and Hahlweg 1993). In three published *controlled trials*, PREP enhanced relationship satisfaction up to 5 years after marriage relative to a control group (Markman and Hahlweg 1993), though in one of these studies PREP benefited only high-risk couples (Halford et al. in press).

5. Couple Therapy

Couple therapy is defined by the use of conjoint therapy sessions to alter the relationship between the partners. While there are many forms of couple therapy, only a very small number of the diverse couple therapy approaches have been subjected to scientific evaluation (Baucom et al. 1998). Over three quarters of 40 published controlled trials of couple therapy evaluate cognitive-behavioral couple therapy (CBCT) (Baucom et al. 1998). The only other couple therapy that meets usual scientific criteria for being an evidence-based therapy is Emotion-Focused Therapy (EFT).

5.1 Cognitive-Behavioral Couple Therapy (CBCT)

In CBCT the interaction between partners is viewed as the key to the relationship problem. Careful assessment of the couple's interaction patterns and each partner's thoughts and feelings is undertaken. Education and skills training are used to help the couple

develop more positive interaction and more positive, realistic relationship beliefs (Baucom and Epstein 1990). For example, if a couple showed the demand-withdraw communication pattern that is common in distressed couples, CBCT might seek to alter this pattern by training the couple in better communication and conflict management.

The cognitive aspect of CBCT is based on the assumption that how partners think mediates their emotional and behavioral responses to one another. Unrealistic relationship beliefs such as 'any form of disagreement with my partner is destructive to the relationship' are believed to underlie negative emotions in relationships. Similarly unhelpful, partner-blaming attributions such as 'he/she does these negative things just to annoy me' also mediate negative emotions in relationships. When relationship negativity is excessive, emphasis is placed on changing such negative thoughts (Baucom and Epstein 1990).

In terms of reducing relationship distress, CBCT consistently has been shown to be superior to no treatment, or to less structured supportive counselling with a therapist, (Baucom et al. 1998). CBCT improves couples' communication skills, reduces destructive conflict, increases the frequency of positive day-to-day interactions, increases positive, realistic thinking about the relationship, and enhances relationship satisfaction (Baucom et al. 1998).

Despite the replicated positive effects of CBCT, there are significant limitations to its effects. Approximately 25-30 percent of couples show no measurable improvement with CBCT, and as many as a further 30 percent improve somewhat from therapy, but still report significant relationship distress after treatment (Baucom et al. 1998). Even amongst those couples who initially respond well to CBCT, there is substantial relapse toward relationship distress over the next few years (Baucom et al. 1998).

5.2 *Emotion-Focused Therapy* (*EFT*)

Like CBCT EFT aims to change the couple's interaction. In EFT it is asserted that negative couple interaction often is driven by the partners' internal emotions, which often result from *insecure attachment*. For example, a high fear of abandonment might lead one partner to be very jealous, or to be hyper-sensitive to any criticism by the spouse. The goal of EFT is to access and reprocess the partners' unhelpful emotional responses, to develop more secure attachment styles, and in this way promote more positive couple interaction (Baucom et al. 1998).

EFT has been evaluated in five controlled trials (Baucom et al. 1998), and has been found to significantly improve relationship satisfaction relative to no treatment or unstructured supportive counseling. However, there is little evidence that EFT helps couples via the mechanisms suggested by the proponents. In other words, there is evidence that this form of couple therapy is helpful, but it is not clear why it helps.

6. *Summary*

Mutually satisfying long-term couple relationships confer many advantages to the spouses and their children. Satisfying relationships are characterized by effective conflict management, good communication, high levels of mutual support, positive day-to-day behavior, positive biases in perception of the partner and the relationship, and positive and realistic relationship beliefs. While almost all couples begin with high relationship commitment and satisfaction, only some couples sustain their satisfaction long-term. The influences on whether relationship satisfaction is sustained include the context in which the couple lives, their individual characteristics, their interactions, and the life events that happen to them. Based on an assessment of individual characteristics and couple interaction, it is possible to classify, for an individual couple, their level of risk for developing future relationship problems. Relationship education that teaches key relationship skills, such as conflict management and communication, enhances the maintenance of high relationship satisfaction. CBCT and EFT help at least some couples who are distressed to improve their relationship.

See also: Cultural Variations in Interpersonal Relationships; Divorce and Gender; Divorce, Sociology of; Family Theory: Economics of Marriage and Divorce; Family Therapy, Clinical Psychology of; Love and Intimacy, Psychology of; Marriage; Marriage and the Dual-career Family: Cultural Concerns; Partner Selection across Culture, Psychology of; Partnership Formation and Dissolution in Western Societies; Personality and Marriage

Bibliography

Baucom D H, Epstein N 1990 *Cognitive-Behavioral Marital Therapy*. Brunner Mazel, New York

Baucom D H, Shoham V, Mueser K T, Daiuto A D, Stickle T R 1998 Empirically supported couple and family interventions for marital distress and adult mental health problems. *Journal of Consulting and Clinical Psychology* **66**: 53–88

Baumeister R F, Bratlavsky E 1999 Passion, intimacy, and time: passionate love as a function of change in intimacy. *Personality and Social Psychology Bulletin* **3**: 49–67

Bradbury T N (ed.) 1998 *The Developmental Course of Marital Dysfunction*. Cambridge University Press, New York

Carrere S, Buehlman K T, Gottman J M, Coan J A, Ruckstuhl L 2000 Predicting marital stability and divorce in newlywed couples. *Journal of Family Psychology* **14**: 42–58

Feeney J, Noller P 1996 *Adult Attachment*. Sage, Thousand Oaks, CA

Glenn N D 1998 The course of marital success and failure in five American 10–year cohorts. *Journal of Marriage and the Family* **60**: 569–76

Gottman J M 1994 *What Predicts Divorce? The Relationship Between Marital Processes and Marital Outcomes*. Lawrence Erlbaum, Hillsdale, NJ
Halford W K 2000 *Australian Couples in Millenium Three: A Research and Development Agenda for Marriage and Relationship Education*. Australian Department of Family and Community Services, Canberra, Australia
Halford W K, Bouma R, Kelly A, Young R 1999 Individual psychopathology and marital distress. Analyzing the association and implications for therapy. *Behavior Modification* **23**: 179–216
Karney B R, Bradbury T N 1995 The longitudinal course of marital quality and stability: A review of theory, method and research. *Psychological Bulletin* **118**: 3–34
Larson J H, Holman T B 1994 Premarital predictors of marital quality and stability. *Family Relations* **43**: 228–37
Markman H J, Hahlweg K 1993 Prediction and prevention of marital distress: An international perspective. *Clinical Psychology Reviews* **13**: 29–43
Markman H J, Stanley S, Blumberg S L 1994 *Fighting for Your Marriage* Jossey Bass, San Francisco, CA
McDonald P 1995 *Families in Australia: A Socio-demographic Perspective*. Australian Institute of Family Studies, Melbourne, Australia
O'Leary K D 1999 Developmental and affective issues in assessing and treating partner aggression. *Clinical Psychology Science and Practice* **6**: 400–14
van Widenfelt B, Markman H J, Guerney B, Behrens B C, Hosman C 1997 Prevention of relationship problems. In: Halford W K, Markman H J (eds.) *Clinical Handbook of Marriage and Couples Interventions*. Wiley, Chichester, UK, pp. 651–78
Weiss R L, Heyman R E 1997 A clinical–research overview of couples interactions. In: Halford W K, Markman H J (eds.) *Clinical Handbook of Marriage and Couples Interventions*. Wiley, Chichester, UK, pp. 13–41

W. K. Halford

Marshall, Alfred (1842–1924)

Alfred Marshall, one of the leading economists of his era, was the progenitor of the 'Cambridge School of Economics,' which rose to world prominence in the interwar period. His *Principles of Economics* (1890) was a seminal work in the development of economic thought and introduced tools still prominent in the practice of applied economics.

1. *The Man*

Marshall was born in Bermondsey, a London suburb, on July 26, 1842, second son of William Marshall, a clerk at the Bank of England, and his wife Rebecca, née Oliver. He grew up in London, attending the venerable Merchant Taylors' School where he revealed considerable intellectual power and an aptitude for mathematics. Entering Cambridge University, he pursued the prestigious mathematical tripos, graduating in 1865 in the exalted position of 'second wrangler'—a result guaranteeing election as a fellow of his college, St John's. His interests soon turned from mathematics to questions of human and social existence, and in 1868 he became a lecturer in moral sciences at St. John's. The moral science tripos covered philosophy, logic, and political economy. He lectured on political economy more from necessity than choice. But, perceiving that human and social advance would be possible only if economic constraints became less oppressive he soon settled on the subject as the focus for his life's work and was to be influential in renaming it 'economics.'

He rapidly became Cambridge's dominant teacher of the subject. While publishing little, his early manuscripts (reproduced in Whitaker 1975) demonstrate that by the early 1870s he had developed many of the distinctive theoretical ideas later associated with his name. He also strove to attain a realistic grasp of the economic world both past and present. Reluctant to produce a theoretical monograph devoid of concrete application, he commenced a volume on international trade questions, visiting North America in the summer of 1875 for fieldwork. Although largely completed and accepted for publication, the project was abandoned in the late 1870s. But in 1879 H. Sidgwick privately printed two of its appendices for circulation in Cambridge and a few copies reached economists elsewhere. Known as *The Pure Theory of Foreign Trade* and *The Pure Theory of Domestic Values* these remarkable pieces were the high watermark of Marshall's achievement in pure economic theory but were not widely known or accessible in his lifetime. (Whitaker 1975 provides the fullest version together with surviving portions of the text of the abandoned volume.)

In 1876 Marshall became engaged to Mary Paley, a Newnham lecturer, who had attended his early lectures for women and informally sat the 1874 moral sciences tripos examination. The two were married on August 17, 1877. Marriage required Marshall to resign his fellowship and he found a new livelihood as first principal of University College, Bristol. Here both Marshall and his wife taught economics. Meanwhile, he had taken a hand in the writing of a primer for extension classes, started by Mary Paley before their engagement. It increasingly became his book and appeared under their joint names as *The Economics of Industry* (1879). While ill-judged for its intended audience, it contained much novel material, including a theory of income distribution on marginal productivity lines. Its original features helped establish Marshall's reputation amongst economists, especially abroad.

The fledgling Bristol college lacked financial support, and the tasks of administration and fund raising proved uncongenial to Marshall, anxious as he was to develop and publish his own ideas. Ill health, diag-

nosed as kidney stones, constrained him further and he resigned with relief in July 1881, spending much of the next year traveling in Europe, with an extended stay in Palermo where composition of his *Principles* began.

The following academic year was again spent in Bristol, where the pair taught economics, after which Marshall—who had become a protégé of B. Jowett, the eminent Master of Balliol—moved to Balliol College, Oxford, as lecturer to the probationers for the Indian Civil Service. The move to Oxford had been viewed as permanent with good prospects for elevation, but the sudden death of H. Fawcett—professor of political economy at Cambridge since 1863—offered the possibility for a return to Cambridge in an assured, independent, and well-paid position. Marshall had already come to be recognized, after the 1882 death of W. S. Jevons, as the UK's leading economist and he took up the Cambridge chair in early 1885, retaining it until 1908 when he voluntarily retired at age 65.

Fawcett had carried the professorship lightly, combining it with an active political career. But Cambridge was changing rapidly and expanding its commitment to modern subjects. With the lapsing of celibacy requirements, new-model professors and lecturers devoted themselves to the tasks of teaching and advancing knowledge. Marshall gave much effort over the years to lecturing and advising students, urging the more promising to follow the high calling of the economist. Despite frequent disappointments, he did manage to inspire a number of students who were to be prominent in the development of British economics, among whom A. C. Pigou and J. M. Keynes were conspicuous. From the outset, Marshall strove as professor to expand Cambridge's scope for economic teaching, but with only modest success. He eventually secured in 1903 a new specialized tripos in economics and politics, but few resources were provided for it. The new degree only became firmly established after his retirement, but the seed he planted reached full flowering in the 'Cambridge School' of the interwar period.

The 1880s saw the most copious flow of occasional publications by Marshall as well as his important evidence to the Gold and Silver Commission of 1888–1889. (Most of his occasional writings are collected in Pigou 1925, which includes a virtually complete, bibliography. Materials concerning his participation in government enquiries are collected in Keynes 1926 and Groenewegen 1996.) *Principles*, on which he had been working since 1881, at last appeared to widespread acclaim in 1890. It was a success not only among academic peers at home and abroad but also with an educated public, gratified by its undogmatic and conciliatory tone and its evident concern for human well being. Marshall's intention had been to follow *Principles*—initially subtitled Volume I—with a companion volume devoted to the applied branches of economics, an idea eventually abandoned as unworkable after intermittent efforts over the next decade or more. Meanwhile considerable effort went into revising *Principles*, which was to go through eight editions in his lifetime, the last being in 1920. (Guillebaud 1961 reprints the last edition with variant passages from earlier editions.) The copious rewritings and frequent rearrangements did not mark significant changes in Marshall's views and tended to impair the work's vigor and overall clarity. A condensation of *Principles* appeared as *Elements of Economics of Industry* in 1892, replacing the earlier work written with his wife, which he had come to dislike intensely. (Like the earlier work it included a treatment of trades unions omitted from *Principles*.)

Marshall's next book, *Industry and Trade*, only appeared in 1919, long after his retirement. It was started in 1904 in the aftermath of the UK's intense controversy over the desirability of maintaining free trade and was initially conceived as a short topical book on the issue of the day. But it soon developed into a more ambitious attempt to study the interrelationships between distinctive national conditions of industry and patterns of international trade. A motivating force was Marshall's deep concern for the UK's seemingly parlous economic future. Progress on the book was slow and delayed by war. The volume appearing in 1919 was only a fragment of the complete scheme with a vague promise, never fulfilled, of a continuation. An applied work, rich in contemporary detail but lacking in focus, it had a respectful reception but little impact. It nevertheless remains a remarkable if neglected work, displaying full vigor and wisdom of mind and much acumen about economic arrangements.

Marshall's last book, *Money Credit and Commerce*, was assembled with unprecedented speed, appearing in 1923. It brought together material, often written decades earlier, on money, the business cycle, and international trade and payments. Here the 50-year-old graphical analysis of *The Pure Theory of Foreign Trade* was published at last. Unfortunately, waning powers precluded a new synthesis, but the book has incidental merits.

Marshall died on July 13, 1924 at Balliol Croft, his Cambridge home since 1886. His wife was to outlive him by almost 20 years. She taught economics at Newnham for many years after their return to Cambridge, but any intellectual ambitions of her own were subordinated to the task of relieving her husband of everyday vexations.

Balliol Croft was the center about which Marshall's life had revolved after 1886. Under pleas of weak health he only left reluctantly as professional or social duties required. However, summers were normally spent in lodgings away from Cambridge—usually on the south coast of England or in the Tyrol—where he could write free of everyday distractions. But he was not a recluse. Balliol Croft saw many visitors and there

were frequent 'at homes' for dispensing advice to students. While not relishing academic dispute, he strove to keep in touch with the real world by quizzing all in active life who came in his way: businessmen, trades unionists, social workers, and so on. Factory visits and inspections of working class life were also valued.

Marshall's was a complex and contradictory personality, which it is now difficult to assess. Prone to fuss over details, and sometimes lacking common sense, he was nevertheless inspirational, especially to the young, in his commitment to the advancement of his subject and his intense conviction of its importance for society's well being. He had remarkable conversational powers and an impressive ability to speak extempore. Above all, his subtle and wide-ranging mind saw complexities that eluded others and made it impossible for him to communicate his thought in a cut-and-dried way. His economic writings both benefit and suffer from this characteristic. They benefit from a breadth of vision and a richness of unexpected and thought-provoking insights. They suffer from a lack of full coherence and clear articulation, making it frustratingly difficult to grasp and describe his thought comprehensively. Marshall himself struggled incessantly with the daunting task of transferring his thought to the printed page, a task further complicated by an undue sensitivity to criticism and an incessant urge to rewrite already adequate texts.

2. *The Background to Marshall's Economics*

British economics was at a critical juncture when Marshall came to it around 1870. The ideas of the classical economists, especially J. S. Mill whose *Principles* (1848) was still dominant, seemed out of tune with the needs and spirit of the age. The publication of Jevons's *Theory* (1871) was a notable harbinger of the coming transformation from classical to neoclassical economics, which—while maintaining the deductive tradition of the classicals—was to emphasize the role of demand, the assumption of optimizing behavior, and the use of mathematical formalization. The British deductive tradition itself was coming under increasing attack from historically minded economists, especially in Germany, who sought to rely on induction and case studies as the best means of augmenting economic knowledge. More generally, Darwinian ideas on evolution were challenging traditional certainties, while there was increasing social concern over the growing political power of the working classes and the hard plight of the poorest.

Marshall was strongly influenced by this intellectual milieu. His initial economic studies centered on Mill's rendition of classical theory but his mathematical background, allied with his early discovery of the neglected work of A. A. Cournot (1838), pushed him towards a mathematical—mainly geometrical—mode of argument. In contrast to Jevons he saw no need to overthrow the inherited classical system and only sought to close its gaps while preserving its important insights. He remained firmly in the British deductive tradition but mistrusted extended deductions based on oversimplified formalizations. He sought to guide deduction by maintaining close contact with observation and history and by allowing fully for disturbing causes. Although somewhat sympathetic to the aims of the historical economists, he emphasized that observation without the aid of a coherent analytical framework or 'organon' could yield little useful knowledge. Influenced especially by the views of Herbert Spencer, he placed great emphasis on the way human character was formed by the experiences of home, workplace, and leisure activities. This opened the possibility of gradual change in human character by the intergenerational transfer of acquired traits, a process more Lamarckian than Darwinian. On the other hand, he strongly resisted the idea that human nature could be transformed by drastic social changes. He attached great importance to the gradual improvement of character and abilities, especially in the lower ranks of society. His ideals here were perhaps parochially reflective of the middle class mores of his era but they were heartfelt. Economic progress was to him not an end in itself but a means to the improvement of mankind. Faced by a choice between the two he would perhaps have chosen human over economic gains, but he fortunately saw them as complementing and reinforcing each other. He declined to base economics on an artificial 'economic man' and sought to deal with human nature in the round, giving recognition to the constraints of duty and the striving for recognition among professional peers. While conceding that the economist as such has no special standing as adviser on normative matters, he did not let this restrain him and was indeed something of a preacher.

3. *Theoretical Contributions*

Marshall's mature contributions to the theory of value and distribution build upon his earlier treatments in *The Pure Theory of Domestic Values* (1879) and *The Economics of Industry* (1879 jointly with M. P. Marshall) and are to be found in books 3, 5, and 6 of all editions of *Principles* after the first.

Book 3 of *Principles* deals with consumers' demand. The focus is on the quantity of a particular good demanded by a consumer at a given price and the 'consumer's surplus' consequently obtained. This surplus arises from the fact that a consumer's demand price for an additional unit falls as successive units are acquired and equals purchase price only for the final or marginal unit bought. Marshall's derivation of an

individual's demand curve, contingent on tastes, income, and other prices, and based on the assumption of diminishing marginal utility, provided a simple if limited foundation for a downward sloping market demand curve for a good, relating its price to aggregate quantity demanded. It remains a staple of elementary pedagogy, as does his related concept of demand elasticity. The consumer's surplus notion seems to have been an independent rediscovery of a neglected idea developed by the French engineer J. Dupuit in the 1840s. Marshall was well aware of the limitations of welfare arguments based on this individual-specific monetary measure of benefit. The individual's marginal utility of money must remain constant, and interpersonal aggregation requires strong assumptions. Nevertheless, the concept continues to be widely used in applied economics.

Book 5 of *Principles* brings together market demand and supply for a commodity, taking for granted the broader economic context—an approach known as 'partial equilibrium' analysis which is closely associated with Marshall although hardly original to him. The treatment of supply is restricted to the extreme cases of monopoly and thoroughgoing competition, although the latter is often more akin to modern theory's imperfect competition than to its perfect competition. (*Industry and Trade* (1919) was to probe interestingly into the intermediate case of trusts, cartels, and 'conditional monopolies' always threatened by potential entrants.) The treatment of monopoly does not advance significantly beyond Cournot, apart from a remarkable brief treatment of socially optimal pricing by a publicly owned monopoly. The treatment of supply in the competitive case is more notable. There are two distinct lines of innovation. 'Period analysis' recognizes that quantities of some productive inputs may not be alterable in a limited time so that the industry's supply conditions become contingent upon the length of the period to which the analysis is meant to apply. Moreover, payments to any type of input will be determined by the cost of securing its use to the industry only if there is enough time to vary freely the quantity employed. Otherwise, payment to such inputs has the character of a rent—termed a 'quasi rent'—determined by the product's price rather than helping to determine this price as a necessary element in production cost.

The other notable feature is the explicit consideration of scale economies in production. External economies are due to the growth of the size of the industry as a whole rather than the sizes of its constituent firms. They reflect osmosis of worker skills, increased opportunities for specialization, and so on. Although somewhat elusive, the idea of such economies remains significant in modern economics. Internal economies are due to a growth in firm size that permits more efficient large-scale production methods to be utilized. While often plausible, a persistent ability of larger firms to undercut smaller firms threatens the viability of atomistic competition. Marshall's solution, implied in his 'representative firm' concept, was to combine market and organizational restrictions on rapid firm expansion with a life cycle for the firm due to a waning of entrepreneurial drive on the part of the founder's successors. Such an argument applies more to the Victorian family firm than to the modern joint stock company, which was to take the limelight only in *Industry and Trade* (1919). The modern 'representative agent' concept owes little to Marshall beyond the name.

Scale economies under competition permit the supply price at which any quantity of a good is supplied to the market to fall with quantity, just as the corresponding demand price does. This admits the possibility that there may be more than one market-equilibrium quantity at which supply price equals demand price. The so-called Marshallian adjustment process through which market equilibrium is attained is very much in the classical tradition deriving from A. Smith. Price adjusts to clear the market of whatever quantity is made available, thus settling at the demand price for that quantity. Then the quantity supplied increases (decreases) as the supply price at the initial quantity is below (above) this market price, so that supernormal (infra-normal) profits are being made. An equilibrium quantity is unstable if there is divergence from it.

The possibility that there may be more than one market equilibrium that is stable for small perturbations is sufficient to disprove a claim that a competitive equilibrium must maximize social welfare. Marshall was clear that the most that could be claimed was what came to be called Pareto optimality. But he showed in a remarkable argument that even this might not be true, demonstrating that a social gain may be obtained by taxing an industry whose supply price rises with quantity in order to subsidize an industry having falling supply price. His analysis of this and other welfare issues drew heavily on the consumer-surplus concept and foreshadowed much in the 'new welfare economics' emerging in the 1930s.

Partial equilibrium analysis was extended in book 5 of *Principles* to cover a group of closely interrelated markets, as exemplified by a set of goods jointly demanded or supplied. More general is the interaction between the markets for an output and its several inputs. Here Marshall introduced the idea of derived demand—the demand for inputs being derived from the demand for the output they jointly produce—fruitfully characterizing the determinants of the elasticity of derived demand for any one input. Even if the number of goods involved is large, the analysis of such multi-market cases remains partial equilibrium in nature since an economy-wide setting that is little influenced by what happens in the markets analyzed is still tacitly assumed.

Only in *Principles* book 6 did Marshall turn to a consideration of the entire economy and its complex

mutual interactions between value, production, and income distribution. His approach was macroeconomic, centering on the concept of the 'national dividend' or national income. This was seen as the combined product of the various factors of production, each unit of which claimed a share equal to its marginal product. Increased quantity of a homogenous factor would lower its marginal product but raise the national dividend, while the absolute or relative shares accruing to the entire factor could rise or fall.

The 'law of substitution' in production and consumption (essentially cost minimization and utility maximization), together with the equality of prices to all-inclusive costs required for competitive equilibrium, ultimately settle equilibrium prices for all goods and factors: also the quantity of each good produced, together with the quantities of the various factors used to produce it. None of this is described with any formality. It remains more a general vision than a precise theoretical statement that might rival the general-equilibrium theory of L. Walras. What book 6 effectively does, however, is to integrate a neoclassical marginal-productivity approach to the demand for factors of production with a classical approach to their supply (extensively developed in *Principles* book 4). This results in a real cost explanation of factor price determination in the long period. If factor supplies were permanently fixed, then factor prices would be determined wholly by the derived demand for their services, and the cost of any particular use would only be the opportunity cost of potential earnings in the best alternative use. But if, as Marshall supposes, continual replacement flows are needed just to maintain factor supplies then factor prices must exactly remunerate the real cost of eliciting these replacement supplies while still satisfying the derived-demand or marginal-productivity condition.

For Marshall, capital, labor of all kinds including managerial labor, and even business organization when firms have finite lives, all require replacement flows to maintain their levels. The case of labor is particularly complicated in the long period. Parents often make decisions for their children about occupation and acquisition of human capital, while the expected return to investment in human capital provides inadequate surety for borrowing to undertake it, being both uncertain and incapable of ownership transfer. A further complication is the 'economy of high wages' which results when increased wages improve worker efficiency by boosting physical or mental well-being, a possibility to which Marshall attached considerable importance. Indeed, his treatment of labor markets is one of the most valuable features of *Principles* book 6.

Apart from those embodied in *Principles*, Marshall's theoretical contributions are limited. His foreign trade analysis (1879) with its ingenious offer curves was indeed an important advance on Mill's treatment, but highly restricted circulation limited its impact. On money, he had a masterly grasp of British monetary tradition but added little that was novel. His most significant innovation was to replace the velocity of circulation with a demand for money function, a change which already appeared in an early manuscript and was to remain influential in the Cambridge monetary tradition leading up to Keynes's *General Theory* (1936).

4. *Policy Views*

Marshall was a proponent of economic freedom and free trade but not a doctrinaire one or an advocate of extreme *laissez faire*. He saw government as having important functions in regulating market forces and remedying or alleviating the problems they produced. Redistributive taxation was acceptable, even desirable, if it did not threaten the security of property or incentives to effort. For the well-to-do striving to keep up with their neighbors, or for professionals seeking the acclaim of their peers, Marshall saw absolute incomes as largely symbolic and felt that taxation which preserved income relativities in the middle and upper classes would not have serious disincentive effects on effort, saving, or risk-taking. Town planning and provision of recreational facilities were needed for the well-being of the urban working classes, the young especially, while a government safety net was needed for those in abject poverty, although private charity could be more discriminating in the incentives it offered and in distinguishing the deserving from the undeserving. To 'administrative socialism'—proposals that government take over the active management of most industries—he objected violently, however. He foresaw that it would result in widespread bureaucratic lethargy and an undermining of the freedom to experiment and innovate essential to economic progress and society's future well being. Increasingly he came to feel that the UK's future prosperity must rest on the chivalric competition of those in control of private business, freed from intrusive government regulation and from enervating protection against foreign competition. He took relatively lightly the possibility that cartels, combines and monopolies might come to dominate, viewing them as continually undermined by economic change. His early hopes that involvement in trades unions and cooperative societies would induce among workers an appreciation of and sympathy for the complexities of business management waned and he came to fear organized labor as an ossifying factor second only to the heavy hand of government and itself a distinct threat to the UK's international competitiveness. But, after the sobering experience of war and its expanded government control of the economy, *Industry and Trade* (1919) was to place more reliance on conscious industry-wide

collaboration between management, labor, and government.

Education had a vital role in raising the working classes by improving abilities, skills, and ambitions, as workers, parents, or citizens. Schools and technical colleges, provided or sponsored by the government, were crucial, but Marshall stressed that much practical education occurred in the workplace: also in the home, where the duties of mothers and daughters were urged and extolled.

On monetary arrangements, Marshall's preference was for a managed paper currency operating within an international gold-exchange standard. His proposal in the 1880s for 'symmetalism'—use of a fixed-quantity combination of gold and silver as the monetary base—was only offered as a superior alternative to the clamor of the time for fixed-price-ratio bimetallism. His preferred goal for monetary policy was to maintain fixity of nominal input prices rather than output prices. Productivity increase would thus yield declining output prices. To palliate the effects of unforeseen price-level changes he proposed voluntary indexation of private contracts by means of a price index or 'tabular standard' published by the government.

5. *Place in Economic Thought*

Marshall can claim some subjective, but limited objective, originality as a contributor to the opening round of the so-called neoclassical revolution, initiated in the 1870s by the writings of Jevons, Walras, and C. Menger, although quite fully anticipated in the then largely forgotten writings of Cournot, Dupuit, H. H. Gossen, and others. He has stronger claims to have been an important contributor to the second round, culminating only in the 1890s, which extended marginalism from value to distribution by means of the marginal-productivity concept. Here, he claimed to have been influenced early on by the work of J. H. von Thünen, although the evidence for this is meager. But the lasting impact of Marshall arises less from transforming ideas than from the provision of serviceable tools which still remain in everyday use: partial-equilibrium analysis, consumer's surplus, demand elasticity, derived demand, long-period versus short-period analysis, and so on. These ideas remain prominent in applied work and pedagogy even if they no longer feature prominently at the frontiers of theoretical research. They perhaps serve as mute testimony in defense of Marshall's distrust of extended abstract theorizing and his preference for being roughly relevant rather than precisely and elegantly irrelevant.

But Marshall's impact on the development of economics came as much, if not more, from his influence on his students and his role in the professionalization, or perhaps better academicization, of economics in the UK as from the impact of his ideas. His *Principles* (1890) was prominent for 30 years or more in the training of economists in English speaking countries, while his students took over many of the new economic posts arising in the UK as higher education expanded, or else became prominent in government service. His struggles to establish economics more firmly in Cambridge bore later fruit in the influential 'Cambridge School,' although his intellectual hold on it had greatly weakened by the 1930s. Anxious that economists should present a public façade of unity and cumulative progress, he was led to present neoclassical ideas as more evolutionary than revolutionary and to take a conciliatory tone in methodological disputes, especially in dealing with the claims of the German historical school. While not an active promoter of, or participant in, professional organizations of economists, he placed his ecumenical imprint on the British Economic Association, founded in 1890 and subsequently renamed the Royal Economic Society.

Equilibrium analysis based on fixed conditions and fixed human response-patterns was only a first step for Marshall, the beginning rather than the end of economic knowledge. His evolutionist leanings led him to decry such analysis as merely mechanical and to look for approaches that were more biological in nature and which recognized the mutability and path-dependence of human character and institutions. His interest in such approaches was not to be much developed and remains somewhat enigmatic. Of particular interest in this connection is his analysis of the intertwined emergence of free enterprise and economic thought moved from *Principles* book 1 to *Principles* Appendices A and B after edition 4. Modern mainstream economics, although considerably indebted to Marshall, has hardly adopted his broader agenda, which indeed seems to have had little influence on the work of his students. Today's economic heterodoxy thus can often find both target and support in Marshall's work.

6. *Further Reading*

Keynes's biographical memoir (in Pigou 1925) is not to be missed, and there is now a scholarly full-scale biography (Groenewegen 1995) and a comprehensive edition of correspondence and related documents (Whitaker 1996). There is no satisfactory monograph on Marshall's economic thought but Tullberg (1990) and Whitaker (1990) provide assessments by various authors on the centenary of *Principles*. Wood (1982, 1996) provides a convenient but undiscriminating reprint in eight volumes of pertinent articles from academic journals.

See also: Agriculture, Economics of; Consumer Economics; Cost–Benefit Analysis; Demand for Labor; Diversification and Economies of Scale; Economics, History of; Economics: Overview; Firm Behavior; Income Distribution; Industrial Metabolism; Industrial Policy; Industrialization; Market Structure and Performance; Mill, John Stuart (1806–73); Monetary Policy; Post-Keynesian Thought; Regulation, Economic Theory of; State and Local Taxation; Wage Differentials and Structure

Bibliography

Cournot A A 1838 *Recherches sur les Principes Mathématiques de la Théorie des Richesses*. Hachette, Paris [1897 *Researches into the Mathematical Principles of the Theory of Wealth*. Macmillan, New York]

Groenewegen P 1995 *A Soaring Eagle: Alfred Marshall 1842–1924*. Edward Elgar, Aldershot, UK

Groenewegen P (ed.) 1996 *Official Papers of Alfred Marshall; A Supplement*. Cambridge University Press, Cambridge, UK

Guillebaud C W (ed.) 1961 *Alfred Marshall's Principles of Economics*, 9th (Variorum) edn. Macmillan, London

Jevons W S 1871 *The Theory of Political Economy*. Macmillan, London

Keynes J M (ed.) 1926 *Official Papers of Alfred Marshall*. Macmillan, London

Keynes J M 1936 *The General Theory of Employment Interest and Money*. Macmillan, London

Marshall A 1879 *The Pure Theory of Foreign Trade, The Pure Theory of Domestic Values*. Privately printed [reproduced in Whitaker 1975]

Marshall A 1890 *Principles of Economics*. Macmillan, London, Vol. 1

Marshall A 1892 *Elements of Economics of Industry*. Macmillan, London

Marshall A 1919 *Industry and Trade*. Macmillan, London

Marshall A 1923 *Money Credit and Commerce*. Macmillan, London

Marshall A, Marshall M P 1879 *The Economics of Industry*. Macmillan, London

Mill J S 1848 *Principles of Political Economy*. Parker, London

Pigou A C (ed.) 1925 *Memorials of Alfred Marshall*. Macmillan, London

Tullberg R McW (ed.) 1990 *Alfred Marshall in Retrospect*. Edward Elgar, Aldershot, UK

Whitaker J K (ed.) 1975 *The Early Economic Writings of Alfred Marshall, 1867–1890*. Macmillan, London

Whitaker J K (ed.) 1990 *Centenary Essays on Alfred Marshall*. Cambridge University Press, Cambridge, UK

Whitaker J K (ed.) 1996 *The Correspondence of Alfred Marshall, Economist*. Cambridge University Press, Cambridge, UK

Wood J C (ed.) 1982 *Alfred Marshall: Critical Assessments*. Croom Helm, London

Wood J C 1996 *Alfred Marshall: Critical Assessments*, Second Series. Routledge, London

J. K. Whitaker

Marshall, Thomas Humphrey (1893–1981)

As one the most important British proponents of social and political theory in the twentieth century, T. H. Marshall's work has significantly shaped the literature on citizenship. The never fading relevance of T. H. Marshall's work is evident in the vast literature on citizenship, written in the last two decades of the century, either in response to Marshall's own writings or as a critique of his work. With each re-imagination of citizenship, we are left with no choice but return to Marshall's original formulations.

This recount of T. H. Marshall's accomplishments moves from a brief acknowledgment of him as an academic and citizen to his sociology of citizenship and its students and critics, and ends with a commentary on citizenship today and Marshall's influence on its theory.

1. T. H. Marshall as an Academic and Citizen

Born in Victorian London on December 19, 1893, T. H. Marshall was the son of a successful architect, his youth spent in affluent Bloomsbury and various country retreats. If not disrupted by an imprisonment in a German civilian internment camp during the First World War, he would have followed the familiar path of English upper-middle class career-pattern—from boarding school to Oxbridge, with a likely career in the Foreign Service (Halsey 1984). Instead, equipped with a stronger 'social awareness and commitment,' Marshall opted out for sociology at the London School of Economics (LSE), where he had a distinctive career and provided a 'distinguished role model for aspirant sociologists' (Halsey 1996, Smith 1996). Marshall taught comparative social institutions at the LSE, with a firm interest in sociological explanations of social change and sociology's potential in creating that social change (Lockwood 1974). It is this commitment to social issues and change, both as an academic and as a public intellectual, David Lockwood argued, made Marshall's work a par with those classical texts that paved the way for modern sociology. This commitment is evident in Marshall's combination of academic life with public service. He worked in the Foreign Office Research Department during the Second World War; served as educational adviser to the British High Commissioner in Germany from 1949 to 1950; and took the Directorship of Social Sciences in UNESCO, when he retired from the LSE in 1956. Marshall shared his life with Nadine Hambourg, whom he married in 1934, in homes in London, Cambridge, and the Lakes. Until his death on November 30, 1982 at the age of 88, he remained active in sociological thinking and writing.

To accentuate Marshall's lifelong identification with sociology is no overstatement. Contemplating on his four years of imprisonment in Germany in an auto-

biographical piece, he recalls that in most of his letters to home he wrote evidence of 'a growing sociological curiosity about what was happening' to him and around. Marshall goes on to say that 'if only [he] had been a sociologist then, what an opportunity this would have given [him] to study, by participant observation and any other survey method, the emergence of a structured society from a miscellaneous individuals by chance in a confined space, and its subsequent fortunes' (Marshall 1973, p. 89). It is as if the sociologist in him made intelligible, and bearable, 'the world around' him and 'the remarkable social experiment' he had endured for four years. Marshall left the prison 'deeply affected' by the experience and with a strong resolve to start an academic career.

As a sociologist, Marshall contributed significantly to the study of social stratification and social policy. His seminal piece on citizenship, *Citizenship and Social Class*, established the many themes that have occupied our scholarly and intellectual agendas since then and for years to come, never diminishing in their import and urgency (the essay was first given as a lecture in 1949 at the University of Cambridge and then published in 1950). His collected essays *Sociology at the Crossroads* (first published in 1963; the American edition with a foreword by S. M. Lipset appeared under the title *Class, Citizenship and Social Development* in 1964) and his text *Social Policy* (1965) are still required readings in many countries and many languages. In the 1990s, Marshall's ideas have come to the fore, not only in the excess academic literature on citizenship, but also in politics as A. H. Halsey (1996) asserts, with the transformation of the social democratic welfare state away from its 'socialistic' strands to a position of 'welfare-capitalism.'

His was half-a-century of remarkable achievement in writing on and advocating for just citizenship and policy (Smith 1996). In this lies his resonance in the advent of an age anticipating tumultuous shifts in the order of citizenship as we know it.

2. *Marshall's Sociology of Citizenship*

Marshall's was an evolutionary sociology, grounded in economic history. *Citizenship and Social Class* combines elemental aspects of both of these disciplines very effectively. His interest in equalities and inequalities, and thus, citizenship was shaped through his study of post-feudal England as a Ph.D. student in Cambridge, as well as his day-to-day experiences with 'class' during his internment in Germany and during his brief encounter with politics as a Labor candidate in 1922.

Marshall's work on citizenship is an attempt to 'grand sociology.' It attempts to trace a 'developmental process'—the process of the development of citizenship rights. Marshall did this based on the reading of the British case, but nevertheless with the aim of chartering a more general map for the progressive development of rights. Each citizenship right built upon each other and labored the next one. The development of civil, political, and social rights not only corresponded to the eighteenth, nineteenth, and twentieth centuries in a sequential order, but also each one was the necessary condition for the other. Of this account of advancement of citizenship, Anthony Rees writes: '…[In the 1950s] Marshall almost single-handedly revived the notion of citizenship, and disseminated a particular view of it so successfully that it came to be seen (at least in England) as the only possible account' (1996, p. 3).

Marshall's interest in explaining citizenship rights has a purpose. It stems from his conceptual and political concerns for class and class inequalities. Marshall concedes that citizenship rights, in their earlier stages, were necessary for the flourishing and maintenance of capitalism, a system of inequality in itself. The civil rights of individuals, particularly the right 'to own property and to conclude valid contracts,' were the necessary foundations for a market economy. But it is the inequalities generated by the capitalist market that the subsequent citizenship rights were contrived to alleviate. Here enters social rights as corrective and as a unifying force.

For Marshall, social rights constitute the 'inevitable capstone' of citizenship development. Substantive social entitlements and rights guaranteed by the welfare state should prevent social and economic exclusions that the earlier provisions of civil and political rights could not. This redefinition of citizenship, from a minimum of legal and political rights to a substantive body of social entitlements, is what made Marshall's lasting intellectual contribution (Somers 1993, Rees 1996).

It is this conception of modern citizenship that has been a major inspiration and challenge to those whose intellectual concerns lie in democracy, equality, and social conflict. On both sides of the Atlantic, the difficult relationship between citizenship and social class dominated the scholarly agendas of the 1950s and 1960s (Reinhard Bendix, Ralf Dahrendorf, A. H. Halsey, S. M. Lipset, David Lockwood and Peter Townsend). The question that underlies this early literature, whether the equalizing principles of citizenship can, and do, eliminate the inequalities of social class and incorporate the excluded into the national collectivity as equal members, still occupies a prime place in the sociological imagination, if only asked differently to include other excluded bodies of citizens (women, and diverse contingents of cultural, sexual, racial, and ethnic minorities).

Marshall saw modern citizenship as a product of a larger societal transformation through which '[d]ifferential status, associated with class, function and family was replaced by the single uniform status of citizenship, which provided the foundation of equality on which the structure of inequality could be built'

(1964, pp. 87–8). It is this tension between the expected transforming capacity of citizenship and the continuing inequalities of the modern society that makes Marshall's work still relevant.

3. *Marshall and Citizenship Today*

The concept of citizenship has lived another major intellectual and political renaissance in the aftermath of the political upheavals of the 1990s, mainly in response to the rebirth of new nation-states and the restructuring of the relationship between states and their citizens. These developments have laid bare the differentials in membership and right-bearing, and exposed the limits of citizenship. The theoretical repository set by Marshall's *oeuvre* has been the point of departure for the new burgeoning interest in citizenship. At the current state of citizenship scholarship, it is unavailing to expect a publication that does not have a reference to Marshall's *Citizenship and Social Class*. Whether critically appraised or endorsed, what is striking is how much the recent conceptualizations of citizenship still owe to that of Marshall's.

Three main lines of criticism warrant consideration if only to reveal the influence of Marshall's work on the current undertakings on citizenship. Not so surprisingly, the most scrutinized aspect of Marshall's sociology of citizenship is its evolutionary character. Marshall's periodization of the rise of citizenship has come under attack for not allowing for alternative paths of progression. Moving from different analytical concerns and focusing on different empirical cases, recent research in historical sociology provides comparative contrast to Marshall's sequence (Turner 1986, Mann 1987, Barbalet 1988, Somers 1993). In this body of scholarship, attention to the existing political regimes, the strategies of the ruling elites and the ruled classes, and the varying institutional arrangements and patterns of associational and participatory forms, at both local and national levels, sanctions the redefinition of parameters and trajectory of citizenship. The path to citizenship, as it is evinced by these critical inquiries, is nowhere single and predetermined. Particularly revealing is the movement and importance that social rights displayed in the citizenship schemes of other countries. Bismarckian social reforms, for example, were introduced as substitute for full political citizenship (Mann 1987). In Germany, not only the social rights preceded political rights, but also the imperative relationship between the two sets of rights that Marshall postulated based on the British case was not there.

Even the British case may not fit neatly with Marshall's periodization, it is argued. The empirical re-analysis of the English citizenship practices in the seventeenth and eighteenth centuries reveals that the emergence of national labor markets and development of civil rights happened in the company of extensive 'claims to and practices of political and social citizenship.' These claims and practices were then, under the right conditions, transformed into explicit rights for English working class communities (Somers 1993). Said otherwise, political and social citizienship materialized in varying intensity but significantly before the time frame predicted by Marshall's scheme.

Here a second line of criticism to Marshall surfaces, the lack of agency and struggle in Marshall's explanatory framework. His comes from the dynamic tension between citizenship and social class in the framework of capitalist market development. Marshall's stages of citizenship progress with the formative demands of each emerging social class, but this class agency remains implicit in the main (Somers 1993). In much of the scholarship that follows Marshall's steps class agency, or agency as such, does not get a proper treatment either. In this literature, citizenship is conceptualized as a personal status or legal category, which guarantees equal rights and exact obligations for individuals in national polities. Inasmuch as the scholarship on citizenship moves away from this legalistic/formalistic take on rights and membership, they reveal a much more dynamic understanding of the citizenship development. The active agency is then reconstituted in the analysis of the evolution and elaboration of citizenship as a complex set of rights, statuses, and practices.

An approach which privileges the agency and participatory practices, for instance, effectively unearths women's place in the realm of citizenship, a social group that does not have visibility in the Marshallian framework. In the USA, despite being denied the franchise, women groups were able to mobilize successfully for mothers' pensions and other legislation in the Progressive era (Skocpol 1992). This finding not only reverses Marshall's sequential analysis, but also remedies the accounts of citizenship which omit the formative role of women and their movements in the historic progress of citizenship.

Rights formation is not only a result of contradictions between the egalitarian principles of citizenship and capitalist society, as Marshall's account articulates (Barbalet 1988). Citizenship is actively created by the participation of people in interaction with formal institutions and power structures (Turner 1986, 1990, Giddens 1982). With this revision, that is, the expansion of the Marshallian model to include 'citizenship from below,' the citizen enters the public space as an actor, 'an active bearer of claims,' and partakes in popular struggles to acquire rights. The resulting model thus resolutely incorporates the agency and mobilization of societal groups into the developmental framework cultivated by Marshall and his exponents.

The third line of criticism takes Marshall's framework to test at the end of the millenium. Marshall wrote his account of citizenship when the national character of the state and welfare regimes was taken for granted. Even though Marshall was not so much

attentive to the formation of the 'national' content of citizenship, nevertheless, he was aware that 'the citizenship whose history [he wished to trace was], by definition, national' (1964, p.72). The extension of socioeconomic rights to the previously excluded groups of people, such as the working class, facilitates their incorporation into the 'common culture' of the national collective. Marshall does not elaborate on this collective, neither does he take it to task to define what holds this collective together and what defines the boundaries of this collective.

A necessary corrective to the study of citizenship, brought to the fore more recently, has been to redefine citizenship from simply a legal status to a body of membership forms and practices (Brubaker 1992, Kymlicka 1995). This shift in focus has brought with it a heightened and productive debate, with attention to historical and contemporary processes of boundary making, and the sources of inclusion and exclusion. As revealed by this debate, citizenship, which has evolved to encompass 'national' collectivities, variously following distinct paths and forming peculiar traditions, is no longer singularly located within national boundaries. Neither the distribution of membership rights requires 'national' belonging as a necessary condition of access to rights. Nor the citizenship practices are delimited by the boundaries of the nation-state within which the citizens reside. This *fin de siècle* reconstitution of citizenship implies a multiplicity of membership. In the new terrain of citizenship, different categories of members are accorded with differentiated set of rights—thus breaching the principle of uniform citizenship rights and accentuating disparities between members. Also in the new citizenship, claims to rights and citizenship expand beyond the conventional modalities of political, civil, and social to embrace cultural, sexual, ecological, and global citizenships. The claims are increasingly advanced and legitimated by appeals to group differences and universal human rights, as codified in international treaties and conventions. The collectives and individuals target with their claims not only the nation-states but more and more transnational (such as the European Union) and local institutions and governing bodies.

All these developments project a significantly different topography of citizenship than the one analyzed and conceptualized by Marshall. As citizenship expands and differentiates within and without the nation-state, one may be impatient to register the infirmities and imperfections of Marshall's account. One may promptly find him guilty of not providing all the right answers. However, in the end, an eventual return to Marshall is only inevitable and fitting. His grand narrative of citizenship paves the way for lucid analyses of the new terrains of citizenship that confront us. Through conversations and argumentation with Marshall, the formations of new citizenship become manifestly tangible and substantive. Ultimately it is his preoccupation as a scholar and citizen that motivates our continuing interest in citizenship—that inequalities, independent of their sources and forms, are still part of our scholarly and societal agendas, and still take us, the sociologists, to task.

See also: Citizenship, Historical Development of; Citizenship: Political; Citizenship: Sociological Aspects; Civil Rights; Welfare; Welfare: Philosophical Aspects; Welfare State; Welfare State, History of

Bibliography

Barbalet J M 1988 *Citizenship: Rights, Struggle, and Class Inequality*. Open University Press, Milton Keynes, UK

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Giddens A 1982 Class division, class conflict and citizenship rights. In: Giddens A (ed.) *Profiles and Critiques in Social Theory*. University of California Press, Berkeley, CA

Halsey A H 1996 T H. Marshall and ethical socialism. In: Bulmer M, Rees A M (eds.) *Citizenship Today: The Contemporary Relevance of T. H. Marshall*. UCL Press, London

Halsey A H 1984 Marshall T. H: Past and present, 1893–1981—President of the British Sociological Association 1964–1969. *Sociology* **18**(1): 1–18

Kymlicka W 1995 *Multicultural Citizenship: A Liberal Theory of Minority Rights*. Oxford University Press, Oxford, UK

Lockwood D 1974 For T. H. Marshall. *Sociology* **8**(3): 363–7

Mann M 1987 Ruling class strategies and citizenship. *Sociology* **21**: 339–54

Marshall T H 1950 *Citizenship and Social Class and Other Essays*. Cambridge University Press, Cambridge, UK

Marshall T H 1963 *Sociology at the Crossroads and Other Essays*. Heinemann, London

Marshall T H 1964 *Class, Citizenship and Social Development: Essays*. Doubleday, Garden City, New York [with an introduction by S. M. Lipset]

Marshall T H 1965 *Social Policy*. Hutchinson, London [3rd rev. edn. 1970]

Marshall T H 1973 A British sociological career. *International Social Science Journal* **25**(1/2): 88–100

Rees A 1996 T. H. Marshall and the progress of citizenship. In: Bulmer M, Rees A M (eds.) *Citizenship Today: The Contemporary Relevance of T. H. Marshall*. UCL Press, London

Skocpol T 1992 *Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States*. Harvard University Press, Cambridge, MA

Smith J H 1996 Forward. In: Bulmer M, Rees A M (eds.) *Citizenship Today: The Contemporary Relevance of T. H. Marshall*. UCL Press, London

Somers M 1993 Citizenship and the place of the public sphere: Law, community, and political culture in the transition to democracy. *American Sociological Review* **58**: 587–620

Turner B S 1986 *Citizenship and Capitalism: The Debate over Reformism*. Allen & Unwin, London

Turner B S 1990 Outline of a theory of citizenship. *Sociology* **24**(2): 189–217

Y. Soysal

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Marx, Karl (1818–83)

Karl Marx was hailed at his death as a 'man of science' who had made two great discoveries in but one lifetime: 'the law of development of human history,' and 'the special law of motion governing the present-day capitalist mode of production' (Engels 1980a, p. 429). In this graveside eulogy Marx's longtime friend and sometime co-author Friedrich Engels (1820–95) likened the first of these great discoveries, previously christened 'the materialist interpretation of history' (Engels 1980b, p. 470), to the formulation by Darwin (1809–82) of 'the law of development of organic nature' (Engels 1980a, p. 429). In other overviews of Marx's work, Engels linked Marx's critical work on political economy and capitalist society with a 'dialectical' method in science derived, so he said, from the philosopher Hegel (1770–1831). Engels wrote that 'Marx was … the only one who could undertake the work of extracting from the Hegelian logic the kernel which comprises Hegel's real discoveries' (Engels 1980b, pp. 474–5). Moreover for Engels, science as conceived by Marx was fully compatible with, and indeed an integral part of, a revolutionary political outlook and practice: 'Science was for Marx a historically dynamic, revolutionary force. … His real mission in life was to contribute, in one way or another, to the overthrow of capitalist society' (Engels 1980b, p. 475).

Famously Marx himself denied that he was a 'Marxist' (Engels 1992, p. 356), whatever that may have meant to himself, and to his friends and enemies while he was alive. While the origins of Marxism lie in Engels's accounts of Marx's life and works from 1859 onwards, and in Marx's own published works and manuscripts, Marxism as a science and as a political movement dates from the years after Marx's death in 1883. For over a century it has been an intellectual and political force to be reckoned with, though the initial presumption of methodological and doctrinal unity, implied in the reference to an origin in one man and his works, inevitably began to break down. On both the intellectual and political fronts, the current age is one of post-Marxism.

In another influential overview of Marx's intellectual career, V. I. Lenin (1870–1924) suggested that Marx combined German philosophy, English political economy, and French socialism and revolutionary doctrines in a unique way (Lenin 1964, p. 50). Most intellectual biographies proceed chronologically and thematically through this list. However, in setting out what distinguishes Marx as a social scientist, and what contributions Marxism has made to social science, it is more productive to take the direct approach and to outline the characteristics of the Marxian synthesis. This will necessitate some contextual discussion of what science and philosophy meant in Marx's time and milieu, and what sort of political strategies and actions were possible for a committed socialist when even constitutional politics was radical and subversive.

1. *Science and Philosophy*

Marx was born on May 5, 1818 in Trier in the Rhineland. His family were converted or at least nonpracticing Jews, and he had a classical and liberal education. This took him to the faculties of law and philosophy at Bonn and Berlin universities in the mid-1830s to the early 1840s, preparing him, rather inadvertently, for a career as a progressive journalist and political radical. This was a time of political reaction in the German states when liberals, especially students, resisted the authoritarian doctrines and practices of self-styled autocrats. Liberals argued instead for representative and responsible government that would protect individual rights to free expression and institute a socioeconomic system favorable to individual mobility. While this may sound familiar enough to present-day readers, the political dialogue of Marx's time now seems laden with obscure philosophical and religious issues. This was partly the result of heavy government censorship imposed on overtly political pronouncements, and partly the result of the requirements of the debate: what facts and arguments would sustain a doctrine of popular sovereignty against the reigning doctrines of absolutist rule by divine right?

The Hegelian idealist philosophy within which these coded political debates took place was itself controversial in relation to the Trinitarian truths of the Christian confession, and this had indeed been recently and notoriously explored by the historian D. F. Strauss (1808–74) in his *Life of Jesus* (1835), and by the philosopher Ludwig Feuerbach (1804–72) in his *Essence of Christianity* (1841, 2nd edn., 1843). Marx was swift to draw the radical conclusion of atheism from these writings, and to align himself with a 'new' materialism rooted in human experience, which he defined as materially productive and historically changing. In the 1830s and 1840s, and particularly in the German-speaking world, 'science' as *Wissenschaft* referred to knowledge in the broadest sense systematically presented. Moreover the tradition of natural philosophy was central to this conception of science, which did not readily admit any great distinction in method between natural and human subjects of study. Science as *Wissenschaft* relied on scholastic concepts and commonsensical observation, rather than on mathematical abstraction and testable hypotheses.

Hegelian idealist philosophy was in any case utterly hostile to a materialist ontology of matter-in-motion and an empiricist epistemology through which truth is assigned to concepts in so far as they coincide with the material world. The later nineteenth century saw the rise of materialist accounts of the natural sciences as

empirical modes of study. Subsequently the development in the early twentieth century of the modern social and behavioral sciences was modeled on these scientific presuppositions and practices. In any case English-language philosophy has been hostile to philosophical idealism, and very closely aligned with philosophical empiricism, since the seventeenth and eighteenth centuries. It follows that Marx's early works, and the overarching conception of science extending through his later works, can be quite difficult to understand from a later vantage point in the twenty-first century. Certainly the reader must be very careful in noting any claims in Marx about what 'science' is supposed to be, so that his thought can be understood without anachronism.

In so far as Marx's notions of science have been influential, a conception of science rooted in idealist *Wissenschaft* has thus survived and flourished. This conception may not cover all conceptions of science within Marxism, nor is it necessarily exactly aligned with those that claim to be Hegelian in origin. Rather it is distinguished by these features:

(a) a focus on human ideas as the medium through which science itself exists, rather than a view that knowledge about humanity must in some sense mirror knowledge of the physical world;

(b) a presumption that throughout history the human focus on what is true and meaningful will change, rather than a presumption that truths are fixed by realities outside of human experience;

(c) a view of science as a form of practical knowledge developed for, and incorporated within, a comprehensive range of human activities, rather than as a realm of abstract truth sought for its own sake by objective observers and politically neutral scientists; and

(d) a political presumption that a critical evaluation of present practice, including practices of power and politics, is a necessary condition for productive science, rather than a presumption that proper science is necessarily unconcerned with its own political context, or a view that it simply does not have one worthy of notice.

2. *Natural Science and Social Science*

For Marx science was itself a unity, and natural science was an activity within, rather than a model for, the human social world. The master science, in his view, was political economy, precisely because it dealt with industry (including technology and natural science) and society as a system of productive relationships and regular interactions. However, what he had in mind was a highly reformed version of the works of political economy that he had available at the time, and a version informed by his political project, which was to promote the interests of wage-workers and others in material need.

These ideas and concerns surfaced in Marx's early journalism, published when Prussian censorship was comparatively relaxed in 1842, and in his notebooks and manuscripts from the following two years, during which time his newspaper was closed down, and as he says, he 'retired into his study.' Marx had only limited contact at this time with the social science of political economy, and in common with other German readers, he viewed writers such as Adam Smith (1723–90), Adam Ferguson (1723–1816), and Sir James Steuart (1712–80) through the medium of Hegel's synthetic and critical philosophizing on this subject, particularly in the *Philosophy of Right* (1821). This work was subjected to a lengthy critique in 1843, unpublished by Marx in his lifetime, and by 1844 he had also completed a set of 'Economic and Philosophical Manuscripts,' again unpublished until the 1930s. While there is of course controversy over the relationship between these works and those that Marx wrote later in life, there are demonstrable continuities in terms of text and project that confirm the early 1840s as the period when Marx's critique of the economic categories was truly launched.

Politically the young Marx aimed to produce a critique of political economy informed by his own version of some of the principles of socialism and communism that were then current. These were mostly French in origin, and derived from the writings of Henri Saint-Simon (1760–1825), Charles Fourier (1772–1837), Étienne Cabet (1788–1856). Marx was also influenced by more contemporary writers such as Moses Hess (1812–75), who identified (rather hopefully) the newly emerging industrial poor as prime movers in socialist politics. At this juncture the relationship between socialism as a political doctrine, and social science as the pursuit of knowledge about society, may seem tenuous. However, in the intellectual and political context of the time the very notion that 'society' needed a scientific investigation was itself subversive and 'socialistic.' The established monarchical order was content by contrast to rely on Biblical fundamentalism, commonsensical conservatism, and vigorous repression of any 'free thinking' that might stir up issues concerning social class and political power. This 'free thinking' was exactly what both liberalism and socialism were about, and the two constituted a potent alliance at the time, albeit with many in-built tensions around social class, political participation, and economic systems.

After the revolutionary events of 1848–9 Marx went into exile in England, and during the remainder of his life, his political involvements were necessarily rather problematic in that context. As he himself noted, London was an excellent center for studying both the social science of political economy and the industrial and commercial life that it theorized. While the English tradition in political economy (actually more Scottish than English) was empiricist in its assumptions and methods (rather than idealist, as was German tra-

dition), it was at this stage not rigorously mathematical nor deductive, as it became in the 1870s after the work of William Stanley Jevons (1835–82) and other pioneers of the 'marginalist' revolution. Central to Marx's critique was the work of David Ricardo (1772–1823), whose *On the Principles of Political Economy and Taxation* (1817, 3rd edn., 1821) is classically descriptive and deductive in a mode of natural philosophy not far removed from Aristotelian conceptions. Ricardo defined and explicated the origins of wealth in terms of commercial value, relating this to the human labor involved in manufacture, and assuming the distribution of wages, profit and rent to the 'three great classes' that necessarily constitute modern societies. Marx's critique of this influential view was both political, arguing that class-divided societies are but a transitional phenomenon, and technical, arguing that Ricardo's concept of labor could not account for the accumulation of profit through the exchange of value. This critique was published in stages, most notably *Capital*, Vol. 1 (1867) and has been subsequently enlarged with edited versions of manuscripts for succeeding volumes and studies.

Marx died on March 14, 1883 in London of natural causes, after 35 years of exile from his native Rhenish Prussia. While other ''48 ers' managed to settle their differences with the regime and to regain their civil rights, Marx made only infrequent nonpolitical visits to Germany, and was resolved to maintain his family home in England. He had married Jenny von Westphalen in 1843, and by 1855 they had three surviving daughters, having lost two sons and another daughter (Franziska) to childhood illnesses, certainly exacerbated by harsh financial circumstances. Since 1962 it has been claimed that Marx was the father of Frederick Demuth, the illegitimate son of the family housemaid, but the alleged evidence for this is highly suspect. Marx's wife and daughter (Jenny) both predeceased him by some months, and his two surviving daughters (Laura and Eleanor) both took their own lives some decades later. Against the obvious gloom and despair of family life in the political wilderness, there are memoirs and recollections extant of a happy and cultivated family life, of Marx's evident and loving affection for his wife and daughters, and his intense grief at the death of his two sons, Edgar and Guido.

3. *Marx and Engels*

Any interpretation of Marx is complicated by the fact that from 1844 until the end of his life he worked closely with Friedrich Engels. Though they wrote only three major works together, they engaged in an extensive and well-preserved correspondence (over 7,000 letters between them are currently extant). They also worked in tandem with political groups, particularly those promoting international co-operation among socialists, such as the International Working Men's Association (1864–76). Engels was a very considerable author and publicist before he met Marx, and was at that time the more famous of the two. He was also an autodidact, having availed himself of lectures and student discussions at Berlin University while he was in the army on Prussian national service. Engels was already a participant in the coded politics of post-Hegelian philosophy, and in the German appropriation of 'English' political economy, publishing an outline 'Critique' in 1844 that influenced Marx significantly. Engels also published an analytical and descriptive account of industrial poverty in *The Condition of the Working Class in England* (1845), based on published sources and the testimony of his own eyes in Manchester.

As indicated above Engels was a notable publicist and popularizer in connection with Marx's work, and the biographical, intellectual, and political context developed in his accounts from 1859 onwards set a profound though not unquestioned model for interpreting Marx. By the later 1860s it is evident from Engels's own works that he himself was increasingly influenced by the development of chemistry and physics as rigorous and mathematically precise sciences, closely linked with industrial processes and with the formulation of a philosophy of science. This eventually flowered as positivism, the view that truth of any kind can only follow from the presuppositions and protocols of a singular scientific method, and that the archetypes for this were physics and chemistry as they progressed to increasingly elegant and powerful theorizations and applications.

Engels's grand project in the 1870s was pursued in his own published works until his death in 1895. This was a synthesis of Hegelian 'dialectics,' which he admired for its ability to cope with change and contradiction, with the natural scientific materialism, on a positivist model, that was proving so successful in the laboratory and the factory. Engels formulated three laws of 'dialectics': unity of opposites, negation of the negation, transformation of quantity into quality (Engels 1987, pp. 111–32). His view was that the discoveries of the natural sciences could be decoupled from the logical atomism and reductionism that empiricism usually implied, and linked instead to unifying 'dialectical' laws of nature, history, and logic itself. This then would unify and transcend materialism and idealism, science and philosophy, practical truth and logical judgment.

The extent to which Marx came to endorse this system, or even to lean in this direction at all, is keenly debated, and the evidence—textual and biographical—is deeply ambiguous. One of the chief difficulties here is achieving an understanding of Marx and his work that is not already deeply influenced by the canons of interpretation set by Engels, as only in that way could these questions be resolved. Moreover

Engels was not only Marx's first biographer, but also the biographer of their relationship, as well as legatee and editor of Marx's manuscripts, and authoritative voice on 'Marx.' Engels was the founder of what was becoming, largely through his own efforts at popularizing and developing Marx's ideas, a body of 'dialectical' thought and socialist practice known as 'Marxism.' Chief amongst the terms he used to promote these views was 'scientific socialism,' coined during the late 1870s and widely publicized.

4. *Engels and Marxism*

Science from Engels's perspective was far more influential within socialist practice, and also within the developing social sciences, than any conception of science based on a fresh reading of Marx's works. This is unsurprising, given the biographical and political circumstances, and the philosophies of science and social science current from the 1890s through the 1920s. Engels coined the term 'materialist interpretation of history' (later denominated 'historical materialism') and also the notion of 'materialist dialectics' (later reformulated by one of his intellectual successors, Georgii Plekhanov (1856–1918), as 'dialectical materialism').

In this way Marxism was generally expounded as an outlook that integrated philosophical truth and political practice. This was done by presenting history, that is human society in a developmental sequence, as an instance of scientific knowledge. This scientific knowledge, in turn, was said to be consistent with unifying 'dialectical' laws of nature and logic. Those laws were themselves reflections of the only cosmic reality, that of matter-in-motion, as studied by natural scientists. In that way for Engels, as for Marxists, social science was but a branch of the natural sciences and of philosophical logic, provided that these systems of knowledge were properly understood as unified by a 'great basic process' that was itself dialectical, and so properly captured in Engels's three laws.

In practice Marxist social science was typically organized by the notion that 'history is the history of class struggles', laid out by Marx and Engels in their *Communist Manifesto* of 1848, and illustrated in the abbreviated account of history as successive 'modes of production' offered by Marx in his Preface to *A Contribution to the Critique of Political Economy* (1859), briefly footnoted in *Capital*, Vol. 1. Marxists defined contemporary political struggle as necessarily and profoundly a struggle between the owners of capital (the bourgeoisie or commercial classes) and the proletarians or workers (those who have nothing to sell but their labor), so their methodology was not merely historical but immediately strategic.

Tactically there were of course numerous difficulties in identifying classes and class-fractions and in finding or rejecting coalition partners for working-class interests. Indeed it was also difficult to define the short- and longer-term interests of the any group of workers at all, particularly with respect to religion, nationalism, and numerous other political expressions of identity and purpose. However, in terms of social science there is no doubt that Marxists were amongst the leaders in studying and publicizing industrial poverty, class divisions in society, and social protest and revolution. While pre-Marxist political economy was not entirely devoid of a concern for industrial workers as human beings and as political agents, it was not at all comfortable with the idea of any 'more or less veiled civil war' (Marx and Engels 1976, p. 495). Marxist social science inverted the rather functionalist assumptions of traditional political economy and declared that such warfare was an inevitable stage on the way to human emancipation.

Under communism greater wealth would be produced and distributed in ways far more humane than any money-system of commercial exchange could ever allow. While communism was not itself theorized in any detail, either by Marx or Engels or by subsequent Marxists, the class struggle was itself documented and pursued with single-minded thoroughness. Despite the claims of a dialectical unity of all the social, natural, and logical sciences, this work was persuasive and respected to the extent that it was empirical and evidential. Any claim that Marxism genuinely established a social science different in principles and character from 'bourgeois' science founders on a point of incommensurability (i.e., conventional social scientists did not regard it as science at all because of the 'dialectical' claims, whatever the results) or on a point of disjunction (i.e., the 'dialectical' claims about social science were not the basis of the science that Marxists actually did, whatever their protestations). In practice self-avowed Marxists focused on class division in industrial societies and characteristically did a social science that claimed to be both scientific and political; conversely any social science preoccupied with working-class poverty was often suspected of Marxist political intent that would bias any scientific results. Clearly, different conceptions of science, and of its relationship with politics, have been operative within the social sciences; Marxist social science highlights and largely constitutes this area of debate.

5. *Orthodoxies and Revisionisms*

Marxism became an orthodoxy once it had been challenged, most notably by 'revisionists' in the later 1890s. While this was initially a political challenge concerning the nature and timing of socialist revolution, further doubts were raised concerning the dialectical philosophy of the sciences espoused by Engels and the specific propositions of Marx's critique of political economy. By the 1920s György Lukács

could write that orthodoxy in Marxism refers exclusively to method. This position reflected the success of the marginalist revolution in economics, which isolated Marx's critique of political economy as irrelevant, given that traditional political economy had been effectively superseded. It also reflected the success of more sympathetic attempts to adapt Marx's economic work on capitalism to a form of economics that would at least be intelligible to the mainstream. And most of all it reflected a process of rethinking Marx that was just beginning.

This process was aided by the first attempts to produce Marx's published and unpublished works in accessible form (the Marx-Engels *Gesamtausgabe* of the 1920s and 1930s, edited initially by D. B. Ryanzanov). Engels was of course in the edition as well, and the production of his works alongside Marx's inevitably led to comparisons and questions concerning the consistency of their views, particularly with regard to Engels's claims concerning dialectics and science. These questions of interpretation could be answered either way, but most readers were encouraged to see a coincidence of views between the two and to keep the assumptions of orthodoxy intact. This was unsurprising given the political impetus to unity, and to validating a Marxist tradition, that successive scholarly and popular editions of 'Marx and Engels' reinforced. Engels himself, in his foundational 1859 review of Marx's *A Contribution to the Critique of Political Economy*, stressed the uniqueness and power of Marx's method even at that early date, suggesting that Marx used both a historical and logical method, and that the two were consistent and coincident.

As the working-class uprisings of the 1920s were suppressed, and as the vacuity of Stalinist thought became evident, 'western' Marxism became more theoretical, more concerned with philosophical and methodological issues, more focused on intellectuals than political leaders, and more academically situated in universities and research institutes, most famously the Institute for Social Research at Frankfurt (in the 1930s, and then in a post-Nazi Diaspora). The grander claims of Engelsian synthesis were dropped in favor of multidisciplinary inquiry (taking in psychoanalysis, popular culture, imperialism, ethics, and morality, for example). Explanatory reductionism to class (meaning a person's relationship to the means of production) was replaced by a multifactor approach to causation, and an interpretative turn towards diagnosis and persuasion. This latter took place within a 're-Hegelianizing' of Marx, levering him out of an Engelsian frame of scientific causation, explanation and prediction, and emphasizing his work in conceptualizing long-term phenomena and overarching structures in social development.

This signals an adaptive process whereby Marx has been successively reinterpreted for the social sciences as the intellectual climate, particularly in philosophy of science, has changed. Engels cast Marx in the shadow of Darwin's 'discoveries' and of a positivist account of natural science (doing considerable violence to Hegel in the process). Later Marxists reinterpreted Marx in the light of neo-Kantian epistemologies and ethical doctrines, and in the light of the burgeoning empirical social sciences. After World War II Marxists such as Louis Althusser (1918–90) adapted Marx to structuralism, producing an account of history and social development that reduced human agency to a mere effect of economically driven and only partly predictable large-scale social forces. Althusser proposed to distinguish a nonscientific Marx from a scientist, and Marx was said to have attained this by making an epistemological 'break' and so expunging any trace of Hegelian idealism. Althusser's project was never satisfactorily completed and is noteworthy for its method, that of an intense and large-scale rereading of Marx, always attempting to fit his ideas into a binarized frame, e.g., nonscience versus science, humanist versus materialist.

Perhaps unsurprisingly the contrary reading, that the apparently Hegelian, 'humanist' Marx of the early works is the valuable one, gathered strength at about the same time. Marx's 'Economic and Philosophical Manuscripts' of 1844 were published for the first time in French and English in the late 1950s, and attracted considerable attention from scholars and readers who were intellectually or temperamentally repelled by the aridities of orthodox dialectics. Any social science presupposes various facts or truths about human beings (or 'man,' as he appeared in those days before feminist critique). The 'humanist' Marx was also a reflection of dissatisfaction with behaviorist models that pictured humans as reactive systems, responding to stimuli in predictable ways. The 1844 manuscripts could be read in isolation from any other works by Marx, and frequently were. They appeared to be self-contained meditations on human nature and history, delineating a story of necessary and progressive interactions with nature leading ultimately to the high productivity of modern industry and the discoveries in the natural sciences that made this possible.

Marx's early manuscripts present 'man' as an inherently social being, necessarily engaged in a sensuous 'metabolism' with nature, which is effectively an external but essential part of the body. Labor, in Marx's view, is not merely a vital activity for satisfying needs (as with animals), but rather in the case of humans a 'free conscious' activity that takes human life itself to be an object and fashions things 'according to the laws of beauty.' Humans produce even when free from physical need, and only truly produce when production is freely undertaken. This claim is meant to have descriptive force, distinguishing self-conscious and self-reflexive activity undertaken by humans from the life-activities undertaken by animals, said by Marx to be merely self-identical with what they do. Humans, in his view, produce themselves in different ways, as

different kinds of people (intellectually and physically) in different productive cultures. Communism, in this scheme, is a way of making this self-constructing process more explicit in terms of egalitarian decision making in society.

The inverse of communism that Marx portrays in his manuscripts is the contemporary world of 'alienated' labor, in which the worker is estranged in terms of possession and power from the products of labor, other workers, nature, society at large, and from the quintessential human species-activity, namely creative work. This philosophical analysis sets a context for the consideration of private property as the contemporary system through which production and exchange takes place, and Marx offers his account of 'man' as an alternative to the schemes developed in classic texts of political economy, most notably the highly individualized yet commercially minded hunters and fishermen portrayed in the work of Adam Smith (1723–90). 'Alienation' is used as a general term by Marx for detailing the specifics of class division in commercial societies, and in the early manuscripts he endeavored to spell these out with respect to workers, and also to nonworkers (though he spent little time on this). The thrust of Marx's argument was towards the overthrow of capitalist society as in no one's interest as a human being. This was precisely because material and psychological deprivation contradicted the inherent potential of the human species-being. Notwithstanding the political and metaphysical tinge to Marx's exposition, alienation functioned in the 1960s and 1970s as an empirical catchall concept for sociological and psychological research into and evaluation of modern industrial processes.

Perhaps the most significant development in Marxist social science in the later twentieth century has been the discovery and appropriation of the work of Antonio Gramsci (1891–1937), a communist trade union activist imprisoned by the Fascist government from 1926 to 1935. During his years in prison he composed a lengthy set of political notebooks, which gradually became available after World War II. His views have a distinct affinity with those of the early Lukács and other 'Hegelian' Marxists, at least to the extent that he departed decisively from the supposed certainties and mechanistic models of dialectical materialism. While he did not have access to Marx's early manuscripts, again there is an affinity with the style of thought and political thrust of these works.

Gramsci was in any case facing the problem that a proletarian revolution, declared inevitable by orthodox Marxism, did not seem to be happening in a steady historical development. Far from class position translating easily into political radicalism, his analysis revealed that cultural conservatism was deeply ingrained in peasts and workers, and that inculcating class-conscious political activism would necessitate an ideological counteroffensive to this already existing 'hegemony.' His work inspired a Marxist social science deeply concerned with tracing the flow of power in society into channels that were not reducible, or not directly reducible, to class position and economic circumstances. Ideas, ideologies, and institutions all became much more important in constructing an analysis of contemporary class politics and a set of political strategies and tactics. While this may have seemed to de-emphasize Marx's lifelong project—a critique of the economic categories—it actually mirrored the methodology and content of his writings on contemporary politics, notably *The Class Struggles in France* (1850) and *The Eighteenth Brumaire of Louis Bonaparte* (1852). This Gramscian reading of Marx, then, is another way that he has been reread and reinterpreted, taking a fresh look at which of his texts is most relevant to the times.

For the social sciences Marx and Marxism offer a rich and varied tradition of inquiry and debate, posing difficult questions about objectivity and neutrality in a social science, about what sort of knowledge counts as scientific, and about what sort of intellectual and practical activity science as a whole actually is. These questions may be answered in terms of contextualized readings of Marx, of Engels, of Marxism, and of later interpreters. How directly those contextualizations map onto current circumstances is very much a matter of judgment by those interested in the social and behavioral sciences today, but there is no doubt that many would set a positive value on what a rereading of this tradition can contribute in the present. Indeed it is likely that most educators would count at least a knowledge of this tradition to be essential for any social scientist, whatever their interests. This itself is a tribute to the wide-ranging and as we would say today 'interdisciplinary' value of the legacy left by Marx.

Marx's major works, including the *Economic and Philosophical Manuscripts*, *The German Ideology*, the *Communist Manifesto*, *The Eighteenth Brumaire of Louis Bonaparte*, the *Grundrisse*, the three volumes of *Capital*, and the three volumes of *Theories of Surplus Value*, have all been published in numerous editions and translations and are widely available in English.

The definitive edition of the complete works of Marx and Engels published in their original languages is the renewed *Marx-Engels Gesamtausgabe* (known as MEGA). This was begun in 1975 under communist party sponsorship and published by Dietz Verlag in East Berlin. It is now being continued by an international consortium (IMES) based in Amsterdam and is being published by Akademie Verlag (Berlin). The series includes all published and unpublished works, manuscripts, letters to and from Marx and Engels, and third parties, and considerable bibliographical and contextual background in the notes and apparatus. Current plans are to complete the series in approximately 122 volumes in the next decades.

A vast and scholarly selection from the Marx–Engels materials is published in English in a set of ap-

proximately 50 volumes, begun in 1975 and now near completion. This is the *Collected Works* from Lawrence and Wishart (London).

See also: Bourgeoisie/Middle Classes, History of; Capitalism; Class: Social; Economics, History of; Marxism in Contemporary Sociology; Marxism/Leninism; Marxist Social Thought, History of; Socialism; Socialism: Historical Aspects; Working Classes, History of

Bibliography

Carver T 1989 *Friedrich Engels: His Life and Thought*. Macmillan, Basingstoke, UK

Engels F 1980a Speech at Marx's graveside (1883). In: Marx K, Engels F (eds.) *Collected Works in One Volume*. Lawrence and Wishart, London, pp. 429–30

Engels F 1980b [Review of] Karl Marx, A contribution to the critique of political economy (1859). In: Marx K, Engels F (eds.) *Collected Works*. Lawrence and Wishart, London, Vol. 16, pp. 465–77

Engels F 1987 Anti-Dühring. In: Marx K, Engels F (eds.) *Collected Works*. Lawrence and Wishart, London, Vol. 25, pp. 5–309

Engels F 1992 Letter to Eduard Bernstein, 2–3 November 1882. In: Marx K, Engels F (eds.) *Collected Works*. Lawrence and Wishart, London, Vol. 46, pp. 353–8

Gamble A, Marsh D, Tant T (eds.) 1999 *Marxism and Social Science*. Macmillan, Basingstoke, UK

Lenin V I 1964 Karl Marx: A brief biographical sketch with an exposition of Marxism (1918). In: Lenin V I (ed.) *Collected Works*, 4th edn. Foreign Languages Publishing House, Moscow, Vol. 21, pp. 43–91

McLellan D 1973 *Karl Marx: His Life and Thought*. Macmillan, London

Marx K, Engels F 1976 Communist Manifesto. In: Marx K, Engels F (eds.) *Collected Works*. Lawrence and Wishart, London, Vol. 6, pp. 477–519

Ruben D-H 1979 *Marxism and Materialism*, new & rev. edn., Harvester Press, Brighton, UK

T. Carver

Marxian Economic Thought

Marxian economic thought encompasses the originary doctrines put forward by Karl Marx together with all the developments and controversies concerning them which have evolved since the mid-nineteenth century. The core of the Marxian 'critique of political economy,' and its *differentia specifica* from other currents in economics, may however be encapsulated in a few sentences. The chief and almost exclusive object of analysis is capital understood as a *social relation of production*, where *exploitation* occurs within a monetary commodity-economy. The theoretical instrument employed by Marx to show the link between money and exploitation as well as the endogeneity of crises is the theory that value has its exclusive source in abstract labor as activity—namely, the living labor of the wage workers.

1. Capital as a Social Relation of Production

According to Marx, the capitalist social relation may be defined as the historical situation where the 'objective' conditions of production (i.e., the means of production, including original resources other than labor) are privately owned by one section of society, the capitalist class, to the exclusion of the other, the working class. Separated from the material conditions of labor and hence unable independently to produce their own means of subsistence, workers are compelled to sell to capitalist firms the only thing they own, the 'subjective' condition of production (i.e., their *labor power*), against a money wage to be spent in buying wage goods. Labor power is the *capacity for labor*: it is the mental and physical capabilities set in motion to do useful work, producing use values of any kind, and it is inseparable from the living body of human beings. The labor contract between the capitalists and the wage workers presupposes that the latter are juridicially free subjects (unlike slaves or serfs), and hence that they put their labor power at the disposal of the former only for a limited period of time. The owners of the means of production, the 'industrial capitalists,' need an initial finance from the owners of money, the 'money capitalists,' not only to buy the means of production among themselves (which, from the point of view of the capitalist class as a whole, amounts to a purely 'internal' transaction), but also and primarily to buy workers' labor power (which, from the same point of view, is its only 'external' purchase). The commodity output belongs to the industrial capitalists, who sell it to 'merchant-capitalists' who, in turn, realize it on the market.

Marx assumes that industrial capitalists initially have at their disposal the money they need, and that they sell the output on the market without intermediation (for a classic survey of Marxian economics, see Sweezy 1970; for more recent perspectives, Harvey 1999, Foley 1986). The capitalist process in a given production period may be summarized in the following terms. The first purchase on the so-called labor market is the opening act, and it enables capitalist entrepreneurs to set production going. Firms look forward to selling the commodity product on the output market against money. The receipts must at least cover the initial advance, thereby closing the circuit. Two kinds of monetary circulation are involved here. Wage workers sell commodities, C (which, in this case, cannot but be their own labor power) against money, M, in order to obtain different

commodities, C′ (which, in this case, cannot but be the commodity basket needed to reproduce the workers, arising out from prior production processes and owned by capitalists). Thus, wage earners are trapped in what Marx calls 'simple commodity circulation,' or C–M–C′. On the other hand, capitalist firms buy commodities in order to sell, hence the circulation appears to be an instance of M–C–M′. More precisely: 'money capital' (M) is advanced to purchase commodities (C), which are specified as the means of production (MP) and labor power (LP). MP and LP are the constitutive elements of 'productive capital,' and their joint action in production gives rise to 'commodity capital' (C′) to be sold on the market and transformed back into money (M′). Once expressed in this form, it is clear that capitalist circulation has meaning only in so far as the amount of money at the end is expected to be higher than the money advanced at the beginning of the circuit—that is, if $M' > M$ and the value advanced as money has been able to earn a *surplus value*, consisting in gross money profits (which firms will actually share with financiers, merchant-capitalists, land-owners and rentiers). M–C–M′ is the 'general formula of capital,' because capital is defined by Marx as *self-expanding value*. The class divide between capitalists and wage workers may therefore be reinterpreted as separating those who have access to the advance of money as capital, money 'making' money, from those who have access to money only as income.

The main question addressed by Marx in the first volume of *Capital* is the following: how can the capitalist class get out of economic process more than they put in? What they put in, as a class, is money capital, which represents the means of production and the means of subsistence required for the current process of production. What they get out is the money value of the commodity output sold on the market at the end of the circuit. From a *macroeconomic* point of view, it is clear that the 'valorization' of capital cannot have its origin in the 'internal' exchanges within the capitalist class (i.e., between firms), because any profit one producer gains by buying cheap and selling dear would transfer a loss to other producers. As a consequence, the *source* of surplus value must be traced back to the only exchange which is 'external' to the capitalist class, namely the purchase of labor power.

2. *The Origin of Surplus Value*

To begin, let us assume that capitalist firms produce to meet effective demand, and let us take the *standpoint of the total capital* articulated in different industries. The methods of production (including the intensity and the productive power of labor), employment and the real wage are all known. Marx proceeds by the *method of comparison*. The labor power bought by capital has, like any other commodity, an exchange value and a use value: the former is the monetary form of the 'necessary labor' required to reproduce the means of subsistence, and is *given* before production; the latter is 'living labor,' or labor *in motion* during production. If the living labor extracted from workers were equal to necessary labor (if, that is, the economic system merely allowed for workers' consumption), there would be no surplus value and hence no profits. Though hypothetical and capitalistically impossible, this situation is meaningful and real, since a vital capitalist production process needs to reintegrate the capital advanced to reproduce the working population at the historically given standard of living. In this kind of Marxian analogue of Schumpeter's 'circular flow' relative prices reduce to the ratio between the labor contents embodied in commodities, or 'values,' which are expressed in money as 'simple' or 'direct' prices through the multiplication for the 'monetary expression of labor' (the quantity of money which is produced by one hour of labor).

But the living labor of wage workers is inherently not a constant but a *variable* magnitude, whose actual quantity is *yet to be determined* when the labor contract is bargained, and that will materialize only within production proper. The length of the working day may be extended beyond the limit of necessary labor, so that a surplus labor is created. Indeed, the control and the compulsion by capital of workers' effort guarantee that this potential extension of social working over and above the necessary labor day actually takes place. In this way what may be called 'originary profits' emerge. Marx assumes that the lengthening of the working day is the same for each worker, so that originary profits are proportional to employment. Their sum is total surplus value. So as not to confuse the inquiry into the origin of the capitalist surplus value with that into its distribution among competing capitals, Marx *sticks to the same price rule*, i.e. 'simple prices' proportional to the labor embodied in commodities. He can then subtract from the total quantity of living labor that has really been extorted in capitalist labor processes and objectified in the fresh value added the smaller quantity of labor that the workers really have to perform to produce the equivalent of the wage-goods.

The comparison Marx makes is not between a situation with petty commodity producers, whose wage exhaust income, and a situation where capitalists are present and making profits out of a proportional reduction in wages. It is rather between two actually capitalist situations, where the *determining factor* is the 'continuation' of the social working day (holding constant the given price rule). An implication of the price rule adopted by Marx is that the labor-time *represented* through the value of the money wage bill is the same as the labor-time necessary to *produce* the means of subsistence bought on the market. If the real consumption of the working class determines the

bargaining on the labor market, and firms' expectation about sales are taken to be confirmed on the commodity market, then the process of capital's self-expansion is transparently determined by the *exploitation* of the working class in *production*, and this is simply reflected in circulation as money making profits. Of course, the possibility of surplus labor is there from the start, after the productivity of labor has reached a certain level. However, Marx's key point is that, because the *special* feature of the commodity labor power is that it is inextricably bound to the bodies of the workers, they may *resist* capital's compulsion. In capitalism there is creation of value only in so far as there is the anticipation of the creation of surplus value (i.e. valorization); and the potential valorization expected in the purchase of labor power on the labor market is realized only in so far as the capitalist class wins the class struggle in production and *make workers work* (provided, of course, firms are then able to sell the output). This is the most *basic justification* for labor being the sole source of value. Value is nothing but 'dead,' objectified labor (expressed through money) because surplus value—the real capitalist wealth—*depends causally* on the *objectification* of the *living* labor of the wage-workers in the capitalist labor process as a *contested terrain*: where workers are potentially recalcitrant, and where capital needs to secure labor to get surplus labor.

In capitalism, therefore, the generativity of surplus is an *endogenous* variable influenced by the *social form* taken by production as production for a surplus value to be realized on the market. With given technology and assuming that competition on the labor market establishes a uniform real wage, 'necessary labor' is constant. Surplus value is extracted either by *lengthening* the working day or by speeding up the pace of production with greater *intensity* of labor. Marx calls this method of raising surplus value the production of 'absolute surplus value.' When the length of the working day is legally and/or conflictually limited, capital may enlarge surplus value by the production of 'relative surplus value.' *Technical change*, which increases the *productive power* of labor, lowers the unit-values of commodities. To the extent that the changing organization of production directly or indirectly affects the firms that produce wage-goods, necessary labor, and so the value of labor power, falls. This makes room for a higher surplus labor and thus a higher surplus value. Changes in production techniques leading to relative surplus value is a much more powerful way of controlling worker performance than is the simple personal control needed to obtain absolute surplus value. Moving from 'cooperation' to the 'manufacturing division of labor' to 'the machine and big industry' stage, a *specifically capitalist* mode of production is built up. In this latter, labor is no longer subsumed 'formally' to capital—with surplus value extraction going on within the technological framework historically inherited by capital—but 'really,' through a capitalistically designed system of production. Workers become mere attendants and 'appendages' of the means of production as means of absorption of labor power in motion. They are mere bearers of the value-creating substance. The concrete qualities and skills possessed by laborers spring from a structure of production incessantly revolutionized from within and designed to command living labor within the valorization process. Labor is now purely *abstract*, indifferent to its particular form (which is dictated by capital) in the very moment of *activity*, where it has lost the object of capitalist manipulation in the search for profit. This stripping away from labor of all its qualitative determinateness and its reduction to mere quantity encompasses both the historically dominant tendency to deskilling and the periodically recurring phases of partial reskilling.

3. *Capital and Competition*

The outcome of the total valorization process may be quantitatively summarized with the help of a few definitions. Marx calls the part of the money-capital advanced by firms that is used to buy the means of productions 'constant capital' because, through the mediation of labor as concrete labor, the value of the raw materials and of the instruments of production is transferred to the value of the product. He calls 'variable capital' the remaining portion of the money-capital advanced—namely, the money-form taken by the means of subsistence that buys the workers to incorporate them in the valorization process—because when living labor is pumped out from workers' capacity to labor as abstract labor, it not only replaces the value advanced by the capitalists in purchasing labor power, but also produces value over and above this limit, and, so, surplus value. Constant and variable capital must not be confused with fixed and circulating capital: 'fixed capital' is the capital tied up in plant and equipment lasting more than the chosen time-period; and 'circulating capital' is the capital advanced for wages and raw materials, and it is only partially consumed within the period. The ratio of the surplus value to the variable capital is Marx's 'rate of surplus value.' It accurately expresses the degree of exploitation, this latter being interpreted as the appropriation by capital of surplus labor within the social working day: the higher (lower) the ratio, the higher (lower) the hours the laborers spend working for the capitalist class relative to the hours they spend producing for their own consumption. A similar division between constant capital, variable capital and surplus value, may be detected within the value of the output produced by single capitals as components of total capital. On the other hand, capitalists naturally refer the surplus value to the total capital they advanced. Surplus value as related to the sum of constant and variable capital takes the new name of

'profit,' and this new ratio is thereby known as the 'rate of profit'. Because it connects surplus value not only to variable capital but also to constant capital, the rate of profit obscures the internal necessary relation between surplus value as the effect to living labor as the cause. Profit increasingly comes to be seen as produced by the whole capital as a thing (either as money-capital or as the ensemble of means of production, including workers as things among things) rather than as a social relation between classes. Nevertheless, this fetishistic mystification is not mere illusion; on the contrary, it depends on the fact that, to exploit labor, capital has to be simultaneously advanced as constant capital, and that thereby wage labor is a part of capital on the same footing as the instruments of labor and the raw materials. From this standpoint, the rate of profit accurately express the degree of valorization of all the value advanced as capital.

Before going on, it is necessary to understand the crucial role and the different meaning of competition in Marx. Competition is, for him, an essential feature of capitalist reality. What all capitals have in common—the inner tendency of 'capital in general'—is their systematic ability to make money grow. It is accounted for by the exploitation of the working class by capital as a whole. The nature of capital, however, is realized only through the interrelationship of the many capitals in opposition to each other. This is already clear in the very definition of abstract labor and value (on Marx's labor theory of value the best treatments are: Colletti 1972, Rubin 1973, Napoleoni 1975, Reuten and Williams 1989; on the relationship between Marx and Hegel, see the essay by Arthur in Moseley 1993). The 'socially necessary' amount of abstract labor contained in a commodity comes to be established through the ex post socialization in exchange of dissociated capitalist-commodity producers. Therefore, the determination of 'social values' as regulators of production leading to some 'equilibrium' allocation of social labor—the 'law of value'—affirms itself on individual capitals only through the mediation of the reciprocal interaction on the market.

Marxian competition is of two kinds (Grossmann 1977). The first is intra-branch (or 'dynamic') competition (this side of Marx's legacy was a powerful source of inspiration for Schumpeter). Within a given sector, there is a stratification of conditions of production, and firms may be ranked according to their high, average or low productivity. The social value of a unit of output tends towards the individual value of the firms producing the dominant mass of the commodities sold within the sector. This, of course, implies that a sufficiently strong shift in demand may indirectly affect social value. Those firms whose individual value is lower (higher) than social value earn a surplus value that is higher (lower) than the normal. There is, therefore, a permanent incentive for single capitals to innovate in search of extra-surplus value, whatever the industry involved. This provides the micro-mechanism leading to the systematic production of relative surplus value, independently of the conscious motivations of the individual capitalists. The new, more advanced methods of production increasing the productive power of labor are embodied in more mechanized labor processes. Thus, the 'technical composition of capital'—the number of means of production relative to the number of workers employed—rises. This is represented by a growth in the ratio of constant capital to variable capital, both measured at the values ruling before innovation, what Marx calls the 'organic composition of capital.' But the 'devaluation' (the reduction in unit values) of commodities resulting from innovation permeates also the capital-goods sector and may well result in a fall of the 'value composition of capital,' that is, of the value-index of the composition of capital measured at the values prevailing after the change.

4. *The 'Transformation Problem'*

The struggle to secure, if only temporarily, extra-surplus value expresses a tendency to a *diversification* of the rate of profit within a given sector. On the other hand, the second kind of competition, inter-branch (or 'static') competition, expresses the tendency to an *equalization* of the rate of profit across sectors. Whereas intra-branch competition is enforced by accumulation, which increases the size of capitals, inter-branch competition is enforced by the mobility of capitals of a given size. An apparent contradiction, however, comes to the fore. The rate of profit is the ratio of surplus value to the whole (stock of) capital invested. Assuming, for the sake of simplicity, that all capital is circulating capital, and that the latter is anticipated for the whole period, if both the numerator and the denominator are divided by the variable capital, the rate of profit is a positive function of the rate of surplus value and a negative function of the (value) composition of capital. If, as Marx assumes, competition makes uniform both the length of the working day and the average wage, then the rate of surplus value is the same everywhere. In other words, the labor power bought by variable capital produces a value and a surplus value which is proportional to the labor time expended. But there is no reason to assume a similar uniformity in the compositions of capital. If 'normal' prices were equal to simple prices, then the rate of profit would in general diverge among branches of production: 'prices of production' including a profit equalized across sectors cannot be proportional to values.

Marx offers a solution to the problem in Volume III of *Capital*: 'prices of production' must be interpreted as 'transformed' simple prices that merely *redistribute* surplus value among capitalist commodity-producers.

They are the prices of the capitalist outputs reached through the application to the capital advanced in each industry (still accounted in 'value' terms) of an *average* rate of profit. This latter is theoretically constructed as the ratio of the *total* surplus value to the sum of constant and variable capital invested in the *whole* economy. This 'value' aggregate rate of profit reflects the total abstract labor congealed in the surplus value over the total abstract labor congealed in capital. As such, it acts as the necessary *intermediate bridge* between simple prices proportional to values (highlighting the genesis of surplus value) and prices of production (representing the 'free' operation of inter-branch competition). The total surplus value resulting from the valorization process is now apportioned to individual capitals as a profit proportional to their amount. The profit accruing to a given capital, therefore, may be higher or lower than the surplus value produced by the labor power bought by its own variable capital if the (value) composition of that capital is higher or lower than the average. In Marx's transformation two equalities are respected: the sum of simple prices is equal to the sum of prices of production, and the sum of surplus values is equal to the sum of profits. Moreover, the 'price' and the 'value' rate of profit are identical. Once inter-branch competition is introduced into the theoretical picture, prices of production replace social values as the centres of gravity of effective, market prices.

A long debate developed from the attempt of subsequent authors to correct what seemed to be an 'error' recognized by Marx himself as present in his transformation. There appeared to be a double and inconsistent evaluation of the same commodities when considered as inputs (means of subsistence and elements of workers' subsistence) and as outputs. The former were computed at 'simple prices' and the latter at 'prices of production' (Foley 2000 provides an overview of the discussion; for a denial that there is any error in Marx's transformation, see the chapter by Andrew Kliman in Bellofiore 1998). The tradition starting with Dmitriev, Bortkiewicz and Tugan-Baranovski and reaching maturity with Seton and Steedman's reading of the Sraffa model of price determination abandons Marx's *successivist* method and frames the transformation in the setting of a *simultaneous* equation system. Taking the methods of production and the real wage as the data, it is possible to fix the prices of production, but in general the two equalities cannot be maintained together and the 'price' rate of profit deviates from the 'value' rate of profit. More damagingly, the labor theory of value appears to be *redundant*, since the values expressed by simple prices are known starting from the 'physical' configuration of production and of workers' subsistence, and this is the given from which the prices of production can immediately be determined, so that there is no need for a dual-system accounting. Rather than being a strength of his theory, as Marx thought, the derivation of prices of production from values (through simple prices) seems to end up in the dissolution of the foundation of the whole theoretical construction.

Among the various attempts to counter this negative conclusion, most forcibly put by Samuelson, three positions may be singled out. The first is represented by Duménil, Foley and Lipietz (compare Foley 1986; but see also: Duménil 1980, Lipietz 1982, and the chapter by Desai in Bellofiore 1998). In Foley's version, the key point is a new interpretation of the value of money and of the value of labor power, which are assumed as the constants of the transformation. The 'value of money,' (i.e. the amount of abstract labor time the monetary units represent), is defined as the ratio of the total direct labor time expended in the period to the total money income—which, of course, is the reciprocal of the 'monetary expression of labor'. The 'value of labor power' is no longer defined as the labor embodied in a predetermined commodity bundle consumed by workers, but as the labor represented in the equivalent going to the workers—namely, the given money wage translated into a claim on social labor being multiplied by the value of money. The purchasing power of this money wage, and, so, the labor embodied in the real wage, may change in the transformation from simple prices to prices of production, and workers' consumer choices are allowed to change. Given that the money value of income is postulated to be the measure of the additional value created by workers and that the value of labor power expresses the distribution of this additional value between capital and labor, variable capital is read as the labor 'commanded' in exchange by money wages (the 'paid' portion of living labor) and surplus value as the labor 'commanded' in exchange by gross money profits (the 'unpaid' portion of living labor). From this point of view, the two Marxian equalities are both respected, provided that the equality between the sum of simple prices and the sum of production prices is applied to the total new value added and not to the total value embodied in the commodity product (a point particularly stressed by Duménil and Lipietz); but the 'price' rate of profit may still vary relative to the 'value' rate of profit. A second position is articulated by Fred Moseley in his contribution to Bellofiore (1998). In his view, the givens in the transformation are the value components (constant and variable capital, surplus value) interpreted as money magnitudes, which are taken to be the same whatever the price rule. This is tantamount to saying that constant capital is also to be thought of in terms of the labor-time represented in the equivalent: as the labor 'commanded' by the money advanced by firms to buy the means of production, rather than in terms of the labor embodied in these latter. Through this further rereading, the two Marxian equalities are confirmed in their originary version, with the 'value' and the 'price' average rate of profit being one and the same.

The third position (cf. Bellofiore-Finelli's chapter in Bellofiore 1998) is based on a reconstruction of Marxian theory within a non-commodity money approach, and on an initial differentiation between money capital (identified with the banking system) and industrial capital (identified with the whole firm-sector). It shares with the New Interpretation the view that the core insight of the labor theory of value is the identity between the net product coming out from the living labor of the wage workers evaluated at simple prices and at prices of production. But the third position derives a stronger claim on distribution from the clear separation between firms' monopoly access to money as capital and wage-earners' access only to money as income. Indeed, this distinction means that, through its aggregate investment decision, industrial capital's macro-behavior is able to set the *real* consumption goods which are left available to workers *as a class*, their freedom to choose as individual consumers notwithstanding (a point which was implicitly taken up again by Keynes in some chapters in his *Treatise on Money*, and extended by some post-Keynesian writers and by the Circuit theory of money). As a consequence, the transformation of simple prices into prices of production means a redoubling of the value of labor power, with 'paid labor' (i.e., the labor-time *equivalent* expressed in the money-prices of the wage-goods bought in exchange) departing from 'necessary labor' (i.e., the abstract labor-time actually *performed* to produce those wage-goods). The rate of surplus-value of the first volume of *Capital* even after the transformation accurately depicts the outcome of the struggle over labor time in production proper, and hence the division between the total living labor expended and the share which has been undertaken for the reproduction of the working class. Since, however, prices of production redistribute the new value added among individual capitals in such a way that the producers of wage-goods may obtain a higher or lower amount than actually produced by the labor-power they employed, the gross money profit/money wage rate is a different quantitative measure, a deceptive form of appearance in circulation obscuring the origin of surplus value from labor.

The new approaches see commodity-exchange at simple prices (proportional to values) or at prices of production as *alternative* price rules. The reasons for Marx's moving from the former to the latter may be summarized as follows. Exchange at simple prices makes it transparent that labor is the source of value and surplus value. The redundancy criticism can be rejected once it is realized that the quantity of inputs other than labor and the quantity of output (i.e. the givens in the transformation) are subordinate to the actual *use*—namely, exploitation—of labor power in production. Value as *objectified* labor realized in money, and embodying surplus value and surplus labor, expresses capital's degree of success in constituting the fundamental capital relation, that is, in winning its battle as total capital against the working class and thus becoming a self-valorizing value. Although some partial treatment of intra-branch competition is required to understand why and how much living labor is 'pumped out' in capitalist labor processes, inter-branch competition—and therefore the redistribution of the new value added across sectors and the derivation of prices of production—is a *secondary* logical moment to be abstracted from at the beginning of the inquiry. Marx is here again adopting the method of comparison. The 'simple price' level of abstraction looks at single capitals as aliquot parts of capital as a whole, on the fiction that each of them is getting all the value that is produced in its individual sphere. The 'price of production' level of abstraction allows for capital mobility equalizing the rate of profit across sectors, and each individual capital has to gain profits in proportion to the investments. This second level may be interpreted as a more concrete layer of the theory. The former, 'hypothetical' capitalism, is, however, by no means a first approximation in the sense of giving a preliminary, inexact picture: it rather *truly* reflects the living labor expended in the different branches of production, that is the *hidden essence* of the capitalist process. The more so if the 'value' subdivision of the social working day between the two classes is understood as invariant to the price rule.

See also: Capitalism; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Marx, Karl (1818–89); Marxism in Contemporary Sociology; Marxist Social Thought, History of; Political Economy, History of

Bibliography

Arthur C 1993 Hegel's logic and Marx's capital. In: Moseley (ed.) *Marx's Method in Capital A Reexamination*. Humanities Press, Atlantic Highlands, NJ

Bellofiore R 1998 *Marxian Economics: A Reappraisal*. Macmillan, Basingstoke, 2 vols.

Colletti L 1972 *From Rousseau to Lenin. Studies in Ideology and Society*. New Left Books, London (1st edn. 1968)

Colletti L, Napoleoni C 1970 *Il futuro del capitalismo: crollo o sviluppo?* Laterza, Roma-Bari

Duménil G 1980 *De la Valeur aux Prix de Production*. Economica, Paris

Foley D 1986 *Understanding Capital. Marx's Economic Theory*. Harvard University Press, Cambridge, MA

Foley D 2000 Recent developments in the labor Theory of value. *Review of Radical Political Economics* **32**: 1–39

Grossmann H 1977 Marx, classical political economy and the problem of dynamics. *Capital and Class* **1**(2): 32–55 and **1**(3): 67–99

Harvey D 1999 *The Limits to Capital*. Verso, London, UK

Lipietz A 1982 The so-called 'transformation problem' revisited'. *Journal of Economic Theory* **26**: 59–88

Marx K 1976 *Capital. A Critique of Political Economy*, Vol. I. Penguin, London

Marx K *Capital. A Critique of Political Economy*, Vol. III. Penguin, London
Moseley F 1993 *Marx's Method in Capital. A Reexamination*. Humanities Press, Atlantic Highlands, NJ
Napoleoni C 1975 *Smith Ricardo Marx*. Blackwell, Oxford (1st edn. 1973)
Reuten G, Williams K 1989 *Value-Form and the State. The Tendencies of Accumulation and the Determination of Economic Policy in Capitalist Society*. Routledge, London
Rubin I 1973 *Essays on Marx's Theory of Value*. Black Rose Books, Montreal, Canada (1st edn.1928)
Sweezy P 1970 *The Theory of Capitalist Development. Principles of Marxian Political Economy*. Monthly Review Press, New York (1st edn. 1942)

R. Bellofiore

Marxism and Law

1. Marx and Engels on Law

Marx's and Engels' writings about law are scattered through their works. This article presents key concepts in their theory, then considers debates about law in relation to the base/superstructure metaphor, the state, class, political struggle, and ideology. A post-Marxist direction is suggested in conclusion.

Key concepts in Marx's theory are forces and relations of production, mode of production, and social class. Forces of production are available technologies; relations of production are the social organization of production, distribution, and exchange. Together these constitute a mode of production. When the forces and relations of production are in contradiction, a revolution will give rise to a new mode of production (Marx and Engels, 1845/6, 1847, Marx 1867/1954).

Each mode of production is constituted by a particular class formation, a class being those people who stand in the same relation to the production process. Capitalism is a mode of production in which a class of workers sells their labor power for wages equivalent to their 'use value' (what is necessary to reproduce normal life). The additional value created by that labor in the production process is retained by the owner of the means of production, the capitalist, and realized as 'surplus value' in the process of exchange (Marx 1867/1954). The wage relation defines capitalism.

These formulations are relevant to a theory of law because:

> The sum of these relations of production constitutes the economic structure of society, the real foundation on which rises a legal and political superstructure and to which correspond definite forms of consciousness … (Marx 1857/1973, p. 501)

Subsequent Marxist theorizing has challenged, clarified, and elaborated these concepts. These discussions illuminate Marx's and Engels' own enigmatic statements.

2. Law and the Economic Base

Taking issue with Marx, Plamenatz (1963, pp. 279–81) argued that law is integral to relations of production, not caused by them. Collins (1982) agrees that 'legal rules … constitute the relations of production' (p. 87). Nonetheless, economic determination remains valid because law 'swallows up' pre-existing relations. Both authors deny the priority of the wage relation as the key to capitalist relations of production yet people were paid *de facto* wages while master–servant relationships persisted at law. Second, in these discussions levels of analysis are confused. At a concrete level, law is integral to contracts of employment. All social relationship have political, ideological, and economic dimensions. At an abstract level, however, work may be conceived as logically prior to the law that regulates it.

Renner (1949) argues that, because law appears to be absent from the production process which it actually regulates, it empowers management while fulfilling its social function of preserving the species. This contradictory argument for both the social functionality of law and the class functionality of law's mystifications has limited the development of Renner's approach.

A constant theme in marxist scholarship is the inability of formal legal equality to affect substantive inequality. Engels (1884/1968) is clear that legal equality 'on paper' (p. 500) is undermined by class power. Pashukanis (1924/1980) explores this further. He stresses the inadequacy of explaining each law in term of its content—a tradition running from Stuchka to the conflict theorists—and also of presenting law as 'merely' ideology. Ideologies express a reality in a distorted way, so Pashukanis seeks to identify that material reality of which law is the expression. He distinguishes the individual legal subject, the bearer of rights, as the fundamental unit of law. While found elsewhere, this conceptualization of the legal person is foundational only in capitalist law, because goods produced for exchange (commodities), unlike use values, must be exchanged by abstractly equal subjects in order for the surplus value they contain to be realized. Edelman's work (1979) elaborates this position.

In his analysis of criminal law, Pashukanis noted the conceptual shift from harm to guilt, and to resolution by individualized redemption transactions. Despite criticisms of such analogic extensions, the expansion of human rights law involving abstract individuation may concede the penal terrain. Balbus (1977) has developed this theme, while Kerruish (1991) uses a similar argument about the fetishization of rights and

their relative impotence vis-à-vis the dispossessed. Here the fundamental concept of law, its subject, is economically determined.

3. *The State and Politics*

Both Engels', argument (1884/1970) that when 'irreconcilable antagonisms' arise between classes, the state develops to 'alleviate' the conflict (p. 327) and Marx's similar position in the *Grundrisse* have been regarded as ambiguous (Freeman 1994, p. 8521). Lenin's interpretation (1917, p. 11) that the state 'is an organ of class *domination*' is clear. The *apparent* reconciliation of actually fundamental conflict merely 'legalises and perpetuates the oppression' (p. 11).

Poulantzas (1968/1973) theorizes this process, aided by Althusser and Balibar's (1968/1975) concept of overall determination, which makes possible theorization of particular and complex patterns of interdependence within and between the economic, the political, and the ideological spheres (see also Wright 1985/1997, p. 29–30). The state and law serve the *political* function of unifying the dominant classes and fractions (power bloc) precisely by becoming relatively autonomous from the economy. The state's function of cohesion shapes its diverse and possibly contradictory interventions into the economy—as owner/employer, as law, regulation, inspection, taxation, and so on.

Poulantzas was not alone in his concern to escape 'monolithic' conceptions of the state (see Cain 1977, Hirst 1977). Now Althusser's (1971) influential formulation of ideological and repressive state apparatuses appeared too simplistic and gave place to an understanding of a play of forces within and between state agencies (Hirst 1977) in which law was both object and subject of struggle.

These conceptual advances survived powerful misinterpretation of the concept of overall determination (Thompson 1976, Anderson 1980) but ultimately foundered on their reserve clause of determination by the economic level 'in the last instance' (Althusser and Balibar, 1968/1975). In the end, these approaches to a marxist theory of state and law were, ironically, defeated by a word when all totalizing theory was dubbed 'terrorism' (Lyotard 1984, p. 64).

4. *Marx's Concept of Class*

Specific class concepts based on mid-nineteenth century forms of capitalism no longer serve. Elaboration of the key terms 'forces and relations of production' to include reproduction helps (Althusser and Balibar 1968/1975) but there are many new found relationships which cannot as yet be theoretically spoken.

The separation between the ownership and control of capital was noted by Berle and Means (1932/1948). It has led to the dominance of finance capital over productive capital, with the implications of this shift have been under-theorized as more pension and insurance holders and private individuals become shareholders (but see Carter 1985). Second, nonmanual workers have increased as a proportion of the workforce in the developed world. Are they 'productive' or 'unproductive' workers in terms of value-creation? Third, there are now more service workers than productive workers in the West. Fourth, feminist scholarship reveals that conventional class categories exclude unwaged care work, and make little sense of part-time work or dual incomes. Finally, the urban poor of the less developed world hold an ambiguous place: are they a reserve army without a war, or 'lumpen' workers despite heavy exploitation in informal economies. Marxism has been subjected to 'the merciless laboratory of history' (Elliott 1998).

Poulantzas (1975) addressed the first two questions, arguing for a distinction based on whether a worker fulfills the function of capital (control) or of labor (production of value). This is useful, but little used (but see Wright 1978). Moreover, the continued focus on productive labor exacerbates Marx's original under-theorization of women and workers in the distribution, reproduction, exchange/consumption, and informal sectors.

Moreover, theorization of global and international legal and state forms remains minimal (but see Wallerstein 1984, Holloway and Picciotto 1978, de Sousa Santos 1996). Existing concepts of the state are capable of elaboration to address these questions. However, concepts of contemporary economic classes and fractions which address the global and ethnic segmentation of consumption/production, and which include women and the nonindustrialized poor, remain unavailable prerequizites for understanding global law.

Finally, there are political power groupings whose most significant characteristic is not class membership: people with disabilities, for example, or women concerned about male violence.

5. *The Place of Law Reform in the Struggles of the Oppressed*

Marxist theory develops in part because people need more than political instinct to steer by when they want to improve their situation. But the problems of achieving change in a life by changing a law have been apparent since Marx discussed the workers' struggle for the 10 hour day (1867/1974, chap. X). In an attempt to resolve the conundrum that legal change is needed but law is not positioned to end oppression, I first followed Marx and Engels (1872/1957), seeing law as impotent (Cain 1972); then Hunt and I saw law's power as a staging post of struggle (Cain and Hunt 1979, Hunt 1981a, 1981b). Finally I argued that

'professionalised law' is of defensive but not offensive value for the oppressed (Cain 1985).

Ideologists of the women's movement have done better. Smart (1989, 1995) argues women must challenge masculinist conceptions of the legal subject as a separated, abstracted, detached, decision maker (see also Thornton 1986). Law's discourse calls forth both masculinized and feminized subjects. Simultaneously, however, pragmatic use of law to achieve limited ends makes sense. Lacey (1999) demonstrates the possibility of constructing a *collective* legal subject. Despite similarities to Pashukanis, both arguments are post-Marxist. They envisage transformation of law without transformation of productive relations, evidencing a theoretical revolution.

6. *Law and Ideology*

Marx (1845–6/1976, p. 600) envisaged 'conceptive ideologists' but failed to situate the concept theoretically. Engels, as Phillips (1980) notes, further developed Marx's theory of ideology. He recognized that ideology, including law, 'develops in connection with the given concept material and develops this material further' (Engels 1886/1973, p. 618), yet he also sought to maintain an understanding of law as determined 'in the last resort' by the 'material life conditions' of those developing these apparently independent ideas.

Gramsci (1971) focused directly on this process. New ideas emerge to solve recurrent collective problems. Those who conceive them are usually organically related to a class and therefore think their new thoughts in class-congruent terms. The invention of double-entry bookkeeping is Gramsci's example. Dominant classes invent and use real solutions; more difficult is the task of subaltern classes in promulgating organic solutions to their problems—and free floating intellectual, well-wishers can make matters worse (Gramsci 1971, pp. 1–16, 117; see also Cain 1983). Resistance and struggle about ideas remain ultimately pegged to a class base.

Althusser (1969, p. 232, 1968/1975, and 1971, p. 160) conceives ideology as the material reality which constitutes thinkers as subjects in an imaginary relationship to external (and internal) conditions of existence. Ideology is not true or false consciousness but the materiality of thought itself. This astoundingly postmodern conception relates to law via practice, the class struggle, and control of ideological state apparatuses. But ideology remains connected to forms of economic relations and class struggle through representation (Hirst 1979). However, the fatal flaw in Althusser's theory of ideology was not determinism but the extra-ideological space reserved for science and for scientific marxism.

Foucauldian theory substituted the concept of discourse for that of ideology, clearing a connotation-free space for development. Foucault (1972) theorizes the radical autonomy of discourse/knowledge in two senses. First, discourse is uncaused. It is not to be explained away by reference to authorial intentions, antecedent discourses, or any reading purporting to identify the 'distant presence of the origin' (1972, p. 138). Class, as a possible 'distant presence,' is thus ruled out. Rather discourse must be analyzed in terms of the interplay of its internal and external relations. Second, discourse is *integrally* powerful. It does not depend on a powerful spokesperson. Rather discourse has the power—later conceived as 'positive power' (Foucault 1978, pp. 92–3)—itself to constitute objects, subjects, and the relations between them, authorized spokespeople and sites of enunciation, courses of action and fields of enquiry. In relation to law, Foucault distinguishes the negative 'sovereign' power to repress (itself sustaining an illusion that elsewhere there exists a power-free zone) from the insidious, constitutive powers of the disciplinary discourses evidenced in moves to resocialization rather than punishment.

Here, we are past Marx, past relative autonomy, and into a world where ideology/discourse has its own power, in a play of discursive and extra discursive forces (1972, pp. 4–6), constellations of power which appear and disappear from the pages of historical texts. Foucault tells how the new criminological discourses stigmatized and divided the unpropertied, how the workers' realization and resistance came too late. Here is an often-overlooked intersection of class with power/knowledge—a narrative which makes no claim to truth but which has a wondrous purchase on the (tragic) theoretical imagination (1977, p. 292).

7. *Conclusion*

Theory is fundamental to progressive practice. Such theorization requires elaboration of the new anology implied by Foucault, in which *both* social relationships *and* discourses are self-existent and powerful while having reciprocally constitutive effects. Theories about law in a globalized world require more refined theoretical concepts of race/ethnicity, land, sex/gender, and elaboration of existing concepts of economic relations. Determination, however, needs relegation to an empirical question.

See also: Analytical Marxism; Justice, Access to: Legal Representation of the Poor; Marx, Karl (1818–89); Marxian Economic Thought; Marxism in Contemporary Sociology; Marxism/Leninism; Marxist Social Thought, History of

Bibliography

Althusser L 1971 Ideology and ideological state apparatuses. In: Althusser L (ed.) *Lennin and Philosophy and Other Essays*. New Left Books, London, pp. 123–73

Althusser L, Balibar E 1968/1975 *Reading Capital*. New Left Books, London
Anderson P 1980 *Arguments Within English Socialism*. Verso, London
Balbus I 1977 Commodity form and legal form: An essay on the 'relative autonomy' of the law. *Law and Society Review* **11**: 571–88
Berle A A, Means G 1932/1968 *The Modern Corporation and Private Property*. Harcourt Brace, New York
Cain M 1974 The main themes of Marx' and Engels' sociology of law. *British Journal of Law and Society* **1**: 136–48
Cain M 1977 An ironical departure: The dilemma of contemporary policing. In: *The Year Book of Social Policy in Britain, 1976*. Routledge and Kegan Paul, London
Cain M 1983 Gramsci, the state, and the place of law. In: Sugarman D (ed.) *Legality, Ideology, and the State*. Academic Press, London, pp. 95–117
Cain M 1985 Beyond informal justice. *Contemporary Crises* **9**(4): 335–73
Carter B 1985 *Capitalism, Class Conflict, and the New Middle Class*. Routledge and Kegan Paul, London
Collins H 1982 *Marxism and Law*. Oxford University Press, Oxford, UK
Edelman B 1979 *Ownership of the Image*. Routledge and Kegan Paul, London
Elliot G 1998 *Perry Anderson: The Merciless Laboratory of History*, University of Minnesota Press, Minneapolis, MN
Engels F 1884/1970 Origin of the family, private property, and the state. In: Marx K, Engels F (eds.) *Selected Works*, *Vol. III*. Progress Publishers, Moscow, pp. 191–336
Engels F 1886/1973 Ludwig Feuerbach and the end of classical German philosophy. In: Marx K, Engels F (eds.) *Selected Works*, *Vol. III*. Progress Publishers, Moscow
Foucault M 1972 *The Archaeology of Knowledge*. Tavistock, London
Foucault M 1977 *Discipline and Punish*. Penguin, Harmondsworth, UK
Foucault M 1978 *The History of Sexuality, Vol. I*. Penguin, Harmondsworth, UK
Freeman M 1994 *Lloyd's Introduction to Jurisprudence*, 6th edn. Sweet and Maxwell, London
Gramsci A 1971 *Selections from the Prison Notebooks*. Lawrence and Wishart, London
Hirst P 1977 Economic classes and politics. In: Hunt A (ed.) *Class and Class Structure*. Lawrence and Wishart, London
Hirst P 1979 Althusser and the theory of ideology. In: Hirst P (ed.) *On Law and Ideology*. Macmillan, London, pp. 40–74
Holloway J, Picciotto S 1978 *State and Capital: A Marxist Debate*. Edward Arnold, London
Hunt A 1981a Marxism and the analysis of law. In: Podgorecki A, Whelan J (eds.) *Sociological Approaches to Law*. Croom Helm, London, pp. 91–109
Hunt A 1981b *The Politics of Law and Justice*. *Politics and Power* **10**
Kerruish V 1991 *Jurisprudence as Ideology*. Routledge, London
Lacey N 1998 *Unspeakable Subjects*. Hart, Oxford, UK
Lenin V I 1917 *The State and Revolution* Allen and Unwin, London
Lyotard J-F 1984 *The Postmodern Condition: A Report on Knowledge*. Manchester University Press, Manchester, UK
Marx K 1857/1973 Preface to a contribution to the critique of political economy. In: Marx K, Engels F (eds.) *Selected Works*, *Vol. I*. Progress Publishers, Moscow, pp. 501–6
Marx K 1867/1954 *Capital*, *Volume I*. Lawrence and Wishart, London
Marx K, Engels F 1845–6/1976 The German ideology. In: Marx K, Engels F (eds.) *Collected Works, V*. Lawrence and Wishart, London
Marx K, Engels F 1847/1967 *The Communist Manifesto*. Pelican, Harmondsworth, UK
Marx K, Engels F 1872/1957 *The Holy Family*. Lawrence and Wishart, London
Pashukanis E B 1924/1980 The general theory of law and Marxism. In: Beirne P, Sharlet R (eds.) *Pashukanis: Selected Writings on Marxism and Law*. Academic Press, London
Phillips P 1980 *Marx and Engels on Law and Laws*. Martin Robertson, Oxford, UK
Plamenatz J 1963 *Man and Society*. Longmans Green, London
Poulantzas N 1968/1973 *Political Power and Social Classes*. New Left Books, London
Poulantzas N 1975 *Classes in Contemporary Capitalism*. New Left Books, London
Renner K 1949/1976 *The Institutions of Private Law and their Social Function*. Routledge and Kegan, Paul, London
Santos B de Sousa 1996 *Towards a New Common Sense: Law, Science, and Politics in the Paradigmatic Transition*. Routledge, London
Smart C 1989 *Feminism and the Power of Law*. Routledge, London
Smart C 1995 *Law, Crime and Sexuality, Part III*. Sage, London
Thompson E P 1978 *The Poverty of Theory and Other Essays*. Merlin, London
Thornton M 1986 Feminist jurisprudence: Illusion or reality? *Australian Journal of Law and Society* **3**: 5–29
Wallerstein I 1984 *The Politics of the World Economy: The States, the Movements, and the Civilisations*. Cambridge University Press, Cambridge, UK
Wright E O 1978 *Class, Crisis, and the State*. New Left Books, London
Wright E O 1985/1997 *Classes*. Verso, London

M. Cain

Marxism in Contemporary Sociology

Marxism has a presence in contemporary sociology in three principle forms: assimilating Marxism, using Marxism, and building Marxism. The first of these is closely identified with the way Marxism is incorporated into mainstream sociology, the second with Marxist sociology, and the third with what might be called sociological Marxism.

1. *Assimilating Marxism*

Most sociologists, even those relatively unsympathetic to the theoretical and political agenda of Marxism, recognize that the Marxist tradition has been a source of interesting and suggestive ideas for sociology, one of the important sources of the 'sociological imagination.' Courses in sociological theory typically include respectful discussions of Marx, Weber, and Durkheim

as 'founding fathers' of central theoretical currents in the history of sociology. Durkheim is identified with norms and problems of social integration, Weber with rationalization and the cultural sociology of meaningful action, and Marx with class and conflict. Studies of politics and the state routinely borrow from the Marxist tradition a concern with business influence, economic constraints on state action, and the class bases of political parties and political mobilization (e.g., Lipset 1960). Discussions of work frequently talk about the labor process, the problem of extracting effort from workers and the impact of technology on skills. Discussions of social change talk about contradictions (e.g., Bell 1976,) and, perhaps above all, discussions of social conflict are influenced by the core Marxist idea that conflicts are generated by structurally based social cleavages, not simply subjective identities (e.g., Bourdieu 1984). In these and many other ways, themes integral to the Marxist tradition are imported into the mainstream of sociological scholarship, frequently losing any explicit link to their Marxist pedigree in the process.

2. *Using Marxism*

Marxist sociology represents a more self-conscious, more ambitious use of Marxist ideas in sociology. Here the idea is to take the conceptual framework elaborated within the Marxist tradition—mode of production, exploitation, the labor process, class structure, class struggle, class consciousness, the state, ideology, revolution, and so on—and to use these concepts to understand a wide array of sociological problems. The goal is not for Marxist concepts to lose their identity by being absorbed into the mainstream, but to challenge the mainstream with an alternative set of explanations and predictions. In studies of state policy formation, the claim is not just that there may be business influence and economic constraints on state policies, but that class power sets fundamental limits on the range of possibilities of state action (e.g., Miliband 1969, Offe 1984). In studies of the labor process and technology, the claim is not just that employers face a problem of gaining cooperation from workers and getting them to perform with adequate effort, but that the antagonism of interests between workers and employers is a fundamental property of work organization in capitalist economies (e.g., Braverman 1974). In studies of inequality the claim is not just that class is one of the salient dimensions of social inequality in industrial societies, but that class relations structure fundamental features of the system of inequality (e.g., Wright 1985).

3. *Building Marxism*

The most ambitious articulation of Marxism and sociology goes beyond simply explicitly deploying Marxist categories to tackle a range of sociological problems. Here the goal is to build Marxism itself, to contribute to its development as a coherent theoretical structure by understanding its shortcomings and reconstructing its arguments. (Examples include Althusser 1970, Cohen 1978. An earlier example with considerable influence in contemporary discussions is Gramsci 1971). Building Marxism is important for the other two articulations of Marxism and sociology: It is only in continually building Marxism that it can generate new insights to be appropriated piecemeal into the mainstream, and new arguments to be used by Marxist sociologists in their engagement with sociology.

To understand the tasks involved in building Marxism, it is necessary to briefly lay out the core substantive arguments of the Marxist tradition of social theory. These arguments fall under three theoretical clusters: (a) a theory of the trajectory and destiny of capitalism, (b) a theory of the contradictory reproduction of capitalism, and (c) a normative theory of socialism and communism. The first of these is the core of classical Marxism, the Marxism of Marx and Engels, and is closely identified with what is usually called 'historical materialism.' The second provides the basic ingredients of what might be called sociological Marxism. In the face of empirical and theoretical criticisms of historical materialism from within Marxism, sociological Marxism has become increasingly central to Marxism as a whole. The third comprises the basis for the Marxist critique of capitalism and its emancipatory vision of alternatives.

3.1 *The Theory of the Trajectory and Destiny of Capitalism*

The traditional Marxist theory of the trajectory and destiny of capitalism was grounded in three fundamental theses.

3.1.1 Thesis 1. The long-term nonsustainability of capitalism thesis. In the long run capitalism is an unsustainable social order. Capitalism does not have an indefinite future. Its internal dynamics ('laws of motion') are mainly generated by three interacting forces: the exploitation of workers by capitalists, the competition among capitalists within ever-expanding markets, and the development of the forces of production. These dynamics of capitalism are deeply contradictory and, it is predicted, will eventually destroy the conditions of capitalism's reproducibility. This means that capitalism is not merely characterized by episodes of crisis and decay, but that these episodes have an inherent tendency to intensify over time in ways which make the survival of capitalism increasingly problematic (Marx 1967).

3.1.2 Thesis 2. The intensification of anticapitalist class struggle thesis. As the sustainability of capitalism declines (thesis 1), the class forces arrayed against capitalism increase in numbers and capacity to challenge capitalism. Eventually the social forces arrayed against capitalism will be sufficiently strong and capitalism itself sufficiently weak that capitalism can be overthrown (Marx and Engels 1992, Lukacs 1971).

3.1.3 Thesis 3. The natural transition to socialism thesis. Given the ultimate nonsustainability of capitalism (thesis 1), and the interests and capacities of the social actors arrayed against capitalism, in the aftermath of the destruction of capitalism through intensified class struggle (thesis 2), socialism is its most likely successor (or, in an even stronger version of the thesis, its inevitable successor). Partially this is because capitalism itself creates some of the institutional groundwork for socialism (concentration of ownership, increased productivity, etc.), but mainly socialism emerges in the aftermath of capitalism's demise because the working class would gain tremendously from socialism and it has the power to create it. Given the interests and capacities of the relevant social actors, socialism would be invented through a process of pragmatic, creative, collective experimentalism when it became an 'historical necessity' (Engels 1945).This is an elegant social theory, enormously attractive to people committed to the moral and political agenda of an egalitarian, democratic, socialist future. Since struggles for social change are always arduous affairs, particularly if one aspires to fundamental transformations of social structures, having the confidence that the 'forces of history' are on one's side and that eventually the system against which one is fighting will be unsustainable, provides enormous encouragement. The belief in the truth of this classical theory, arguably, helped to sustain communist struggles in the face of such overwhelming obstacles.

Unfortunately, the empirical evidence for the central theses of the theory of the dynamics and destiny of capitalism are quite weak, and a number of the theoretical foundations for the theses flawed. As a result, most Marxist scholars either explicitly abandon historical materialism or ignore it. Their work, therefore, tends to revolve mainly around the second pillar of the Marxist tradition: the theory of contradictory reproduction.

3.2 The Theory of the Contradictory Reproduction of Capitalism and its Class Relations

The Marxist theory of the contradictory reproduction of capitalism and capitalist class relations is also based on three fundamental theses.

3.2.1 Thesis 1. The social reproduction of class relations thesis. By virtue of their exploitative character, class structures are inherently unstable forms of social relations and require active institutional arrangements for their reproduction. Where class relations exist, therefore, it is predicted that various forms of political and ideological institutions will develop to defend and reproduce them. In capitalism this problem of social reproduction of class relations is further complicated by instabilities generated by capitalist competition and uneven development.

3.2.2 Thesis 2. The contradictions of capitalism thesis. The institutional solutions to the problems of social reproduction of capitalist class relations at any point in time have a systematic tendency to erode and become less functional over time. This is so for two principle reasons: First, capitalist development generates changes in technology, the labor process, class structure, markets, and other aspects of capitalist relations, and these changes continually pose new problems of social reproduction. In general, earlier institutional solutions will cease to be optimal under such changed conditions. Second, class actors adapt their strategies in order to take advantages of weaknesses in existing institutional arrangements. Over time, these adaptive strategies tend to erode the ability of institutions of social reproduction to regulate and contain class struggles effectively.

3.2.3 Thesis 3. Institutional crisis and renovation thesis. Because of the continual need for institutions of social reproduction (thesis 1) and the tendency for the reproductive capacity of given institutional arrangements to erode over time (thesis 2), institutions of social reproduction in capitalist societies will tend to be periodically renovated. The typical circumstance for such renovation will be institutional crisis—a situation in which organized social actors, particularly class actors, come to experience the institutional supports as unsatisfactory, often because they cease to be able to contain class conflicts within tolerable limits. There is no necessary implication here that the new institutional solutions will be optimal or that capitalism will collapse in the face of suboptimal arrangements. What is claimed is that capitalist development will be marked by a sequence of institutional renovation episodes in response to the contradictions in the reproduction of capitalist relations.

Most of the empirical research and theoretical development done by contemporary scholars engaged in building Marxism has in one way or another revolved around these three theses. Thus Marxist theories of advanced capitalism have focused on questions of how the state regulates relations among

capitalists and the relations between capital and labor (e.g., Aglietta 1979). The state also organizes class struggles so that they do not threaten capitalism. It does this through various concessions to the working class such as factory legislation, minimum wages, unemployment compensation, and so forth (e.g., Przeworski 1985). Alternatively, the state disorganizes the working class, for example, through the legal order which constitutes individual citizens (e.g., Poulantzas 1973) or, in some times and places, by promoting racial divisions through discriminatory access to jobs. Sociological Marxism investigates the ways in which consent to capitalism is both organized and potentially challenged within production (Burawoy 1985) as well as within the institutions of civil society—from schools to churches, from trade unions to political parties (e.g., Bowles and Gintis 1976). In these and other ways, the problem of understanding the contradictory reproduction and transformation of capitalist class relations and institutions constitutes the central agenda of sociological Marxism.

3.3 The Normative Theory of Marxism

If one believes the traditional Marxist theory of the dynamics and destiny of capitalism, then there is little need for an elaborate normative theory of the alternatives to capitalism. The problem of socialism can be left to the pragmatic ingenuity of people in the future. It is for this reason that Marxists traditionally have believed an elaborate positive normative theory was unnecessary. The normative dimension of Marxism has thus primarily taken the form of the critique of capitalism as a social order characterized by alienation, exploitation, fetishism, mystification, degradation, immiseration, the anarchy of the market, and so on. The transcendence of capitalism by socialism and, eventually, communism, was then posited as the simple negation of these features, an implicit and undefended theoretical utopia which simply eliminated all the moral deficits of capitalism: a society without alienation, exploitation, fetishism, and the rest.

Once one abandons the optimistic predictions of historical materialism, however, there is no longer a theoretical grounding for bracketing the normative issues. The twentieth century witnessed several historical experiments of trying to build socialism in the aftermath of anticapitalist revolutions without a coherent normative model of socialist institutions. If we have learned anything from the history of revolutionary struggles against capitalism, it is that anticapitalism is an insufficient basis on which to construct a feasible, emancipatory socialist alternative. In addition to a sociological Marxism which explores the contradictory reproduction of class relations in capitalism, therefore, Marxism requires a normative theory that illuminates the egalitarian and communitarian ideals of the emancipatory project, the dilemmas and contradictions in the actual historical attempts at creating socialism, and the design principles of feasible institutions for realizing those emancipatory ideals in the future. The further development of this normative theory is one of the essential tasks for building Marxism in the twenty-first century.

See also: Alienation, Sociology of; Capitalism: Global; Class Consciousness; Class: Social; Socialism; Theory: Sociological

Bibliography

Aglietta M 1979 *A Theory of Capitalist Regulation*. New Left Books, London
Althusser L 1970 *For Marx*. Vintage, New York
Bell D 1976 *The Cultural Contradictions of Capitalism*. Basic Books, New York
Bourdieu P 1984 *Distinction*. Harvard University Press, Cambridge, MA
Bowles S, Gintis H 1976 *Schooling in Capitalist America*. Basic Books, New York
Braverman H 1974 *Labor and Monopoly Capitalism*. Monthly Review Press, New York
Burawoy M 1985 *The Politics of Production*. Verso, London
Cohen G A 1978 *Karl Marx's Theory of History: A Defense*. Clarendon Press, Oxford, UK
Engels F 1945 [1892] *Socialism: Utopian and Scientific*. Scribner, New York
Gramsci A 1971 *Selections from the Prison Notebooks*. International Publishers, New York
Lipset S M 1960 *Political Man*. Doubleday, Garden City, NY
Lukacs G 1971 *History and Class Consciousness*. MIT Press, Cambridge, MA
Marx K 1967 [1867] *Capital*. International Publishers, New York, Vol. I
Marx K, Engels F 1992 [1848] *The Communist Manifesto*. Oxford University Press, Oxford, UK
Miliband R 1969 *The State in Capitalist Society*. Basic Books, New York
Offe C 1984 *The Contradictions of the Welfare State*. MIT Press, Cambridge, MA
Poulantzas N 1973 *Political Power and Social Classes*. New Left Books, London
Przeworski A 1985 *Capitalism and Social Democracy*. Cambridge University Press, Cambridge, UK
Wright E O 1985 *Classes*. Verso, London

M. Burawoy and E. O. Wright

Marxism/Leninism

1. Marxism, Leninism, or Marxism–Leninism?

Scholarly debates about Marxism–Leninism are inextricably bound up with larger intellectual battles

about the relationship—or lack thereof—among the theoretical works of Karl Marx, the theories and political activities of Vladimir I. Lenin, and the tyrannical regime created in the 1930s by Joseph V. Stalin. During the Cold War, scholarship on these questions tended to divide into two highly polarized camps. One view, closely associated with the model of 'totalitarianism' developed by such thinkers as Hannah Arendt (1951), Carl Friedrich and Zbigniew Brzezinski (1956), and Leonard Schapiro (1972), was that the logic of Marxist ideology led inexorably to the more hierarchical, centralized, and violent forms of rule advocated by Marx's Soviet disciples. From this point of view, the generic label 'Marxist–Leninist' could be reasonably applied even to philosophical and ideological works written well before the Bolshevik Revolution of 1917. Thus H. B. Acton's influential *The Illusion of the Epoch: Marxism–Leninism as a Philosophical Creed* (1955) freely interspersed quotations from works by Marx, Lenin, and Stalin in order to illustrate what Acton saw as the philosophical fallacies common to all three theorists. In the hands of lesser scholars, such an approach appeared to hold Marx personally responsible for everything done in the name of his theory, decades after his death, in a country he had never visited in his lifetime; alternative, non-dictatorial tendencies within Marxist socialism in nineteenth and twentieth century Europe were often left unexamined.

By the 1960s and 1970s, efforts on the left to resurrect Marxism—and in some cases Leninism as well—as a type of humanism incompatible with Stalinist tyranny led to a series of revisionist reappraisals of the origins of Soviet ideology (Cohen 1973). Careful study, the revisionists insisted, revealed not continuity but sharp breaks between the thought of Marx and his many of his later self-proclaimed disciples. Such discontinuities—and in particular, the rigidly hierarchical and conspiratorial nature of the Soviet ideological order—were attributed to differences in the cultural milieus of Western and Eastern Europe, to the influence of the Jacobin traditions of the Russian revolutionary intelligentsia on Lenin and Stalin, to the exigencies of state-building in a hostile international environment facing the Bolshevik elite, and even to the peculiar personal psychology of Stalin himself (Tucker 1969, 1971). The term Marxism–Leninism itself, from this perspective, was a sham, useful only (for opposite reasons) to Stalinist and capitalist ideologues (Kautsky 1994).

Since the collapse of the USSR in 1991, however, both totalitarian and revisionist theories about the nature of Marxism–Leninism have come into question. On the one hand, revelations from Soviet and East European archives make it quite clear that Communist Party leaders throughout the bloc really did communicate in the idiom of official Marxism–Leninism, even in their private correspondence, until nearly the end of the Gorbachev era. On the other hand, these documents also show that debates and fissures among party leaders on important ideological issues were a constant feature of Leninist rule, exploding the notion of an entirely uniform and continuous Marxism–Leninism operating from Marx to Stalin and beyond. Thus, the new evidence forces scholars to account simultaneously for continuity *and* change in the ideological discourse linking Marx to Lenin and later Marxist-Leninists. To understand the process by which the complex and multifaceted theories of Marx were transformed into the stultifying orthodox ideology imposed on millions of people by Stalin and his colleagues, then, it is necessary to proceed historically, identifying both the core elements of Marx's theory which united otherwise diverse followers into a coherent intellectual movement, as well as the inconsistencies in Marx's work that generated such heated debates among them. Here the works of intellectual historians like Leszek Kolakowski (1981), Martin Malia (1994), and Andrzej Walicki (1995) have proven especially valuable.

2. *Ideological Continuities and Changes from Marx to Lenin*

The first basic tenet of Soviet Marxism–Leninism—that Marx developed a comprehensive world-view based upon a unique form of dialectical materialism, supposedly reflecting the class interest of the global proletariat—is indeed largely supported by the evidence of Marx's own work. To be sure, some of the later efforts by Soviet ideologists to apply the 'dialectical method' to every conceivable problem of philosophy, science, and art—with damaging and often embarrassing results—had their roots not in Marx, but instead in the writings of his collaborator Engels, whose efforts to synthesize Marx's historical materialism with his own rather peculiar understandings of modern physics and Darwinian evolution were discouraged by Marx himself (Carver 1983). Still, Marx did explicitly claim that his theorizing about revolutionary communism reflected the universal interests of the emerging world proletariat and he expended a great deal of energy attempting to demonstrate that all rival understandings of socialism actually served the interests of nonproletarian, reactionary classes. And while Marx himself never used the exact term 'dialectical materialism,' he consistently claimed that the materialist conception of history that he advocated was thoroughly dialectical in the Hegelian sense—that is, based upon the idea that the logic of history is driven by the struggle between opposed principles, that this struggle generates qualitatively higher stages of development, and that history must lead ultimately to total transcendence and freedom. Seeing human labor rather than Hegel's ethereal 'spirit' as the driving force in history—and as the primary means through which empirical human

beings express and fulfill themselves—Marx recast the Hegelian theory of history in materialist terms, as a struggle between property-owning and oppressed laboring classes, progressing through a series of revolutions in the mode of production from slavery to feudalism to capitalism, and leading ultimately to the transcendence of class society through the final victory of the global proletariat under communism. In this way, Marx theoretically reconciled the seemingly antithetical principles of science, with its insistence on rational analysis of existing material reality, and revolution, with its goal of radical transformation of existing reality through human action. In its basic outlines, this framework was largely preserved in later Soviet presentations of dialectical materialism.

By contrast, the link between Marx's work and the Leninist principle of one party dictatorship is rather less direct. In fact, Marx was quite inconsistent concerning the problem of communist political strategy. Certainly he never explicitly called for the imposition of a hierarchical, centralized political party to run the post-revolutionary proletarian state. Marx did insist more than once that the postrevolutionary government would take the form of a dictatorship of the proletariat; in addition, he saw the communist party in each bourgeois nation as consisting of those workers and intellectuals who had the best theoretical understanding of the long-term strategic goals of the communist movement. However, such passages in Marx's writings coexist with others emphasizing the independent political competence and leadership role of the working class itself. Overall, the theoretical synthesis between science and revolution achieved in Marx's theoretical works falls apart again in his scattered writings on political strategy, where the more scientific Marx tends toward pessimism about the prospects for proletarian collective action before economic conditions are ripe, while the more revolutionary Marx calls for immediate proletarian uprisings in a wide range of developed and less developed societies.

Indeed, precisely this dilemma concerning proper revolutionary timing lay at the heart of debates about communist political strategy during the period of the Second International, the loosely organized coalition of European Marxist parties founded by Engels after Marx's death (Steenson 1978). Revisionist, or right, Marxists, such as Eduard Bernstein, emphasized the scientific aspects of Marx's work over the ideals of revolutionary communism and called for an evolutionary approach to socialism within the context of existing capitalist nation-states. Left Marxists, such as Rosa Luxemburg, took the opposite position, calling for immediate proletarian revolutionary action and downplaying the empirical constraints on revolutionary success emphasized by the revisionists. Finally, orthodox or center Marxists, led by Karl Kautsky, attempted to preserve the synthesis of rational analysis and revolutionary transcendence by calling for fidelity to the entirety of Marx's dialectical outlook. In practice, this theoretical stance led to an emphasis on building the political power of existing Marxist parties, while foregoing premature attempts at communist revolution. It was in this context that Kautsky drew the conclusion that the working class could not create a revolution on its own without prior training by orthodox Marxist intellectuals (Donald 1993).

Indeed, Lenin's 1902 proposal in *What Is To Be Done?* to reorganize the Marxist movement in Russia under the leadership of a strictly disciplined, highly centralized party of professional revolutionaries was originally introduced as an attempt to reinforce Kautsky's orthodox Marxism within the small Russian Social Democratic movement founded by Georgii V. Plekhanov in the 1880s (Harding 1977). While Lenin's conspiratorial conception of the party marked a qualitative break with the procedural democratic norms typical of European socialism at the turn of the century—a factor which played an important role in the 1903 split between Lenin's Bolsheviks and the anti-Leninist Mensheviks—his argument that such a party might itself synthesize revolutionary praxis and modern organizational science was firmly within the orthodox Marxist political tradition.

After World War I had simultaneously destroyed European capitalism, fractured the Second International along national lines, and plunged Lenin's Russian homeland into near anarchy, the popular appeal of Lenin's institutional solution to the dilemmas of political Marxism vastly increased. Stunned by the inaction of Kautsky and the German Social Democratic Party in the early days of the war, Lenin now insisted that, in an age of global imperialism, his model for Marxist politics was universally applicable—and moreover the sole means to combat the opportunism of the 'bourgeois' Second International. With the Bolshevik takeover of central Russia in November 1917 and the subsequent victory of the Red Army in the ensuing Russian Civil War, Lenin's arguments looked prophetic. During this period Lenin's ideas about party organization also became the basis for a new Communist International based in Moscow, spurring the formation of pro-Soviet communist parties throughout Europe and Asia.

3. *Ideological Continuities and Changes from Lenin to Stalin*

The establishment of Lenin's party dictatorship in the Soviet Union, however, did not immediately complete the full institutionalization of Marxism–Leninism. Rather, this hyphenated term emerged only after Lenin's death, during a new struggle among right, left, and orthodox tendencies within the Bolshevik party—this time concerning the proper economic strategy

for the first socialist state. As in the political debates of the Second International, Marx's own works were hardly an adequate guide to action here. The bulk of Marx's economic analysis had been devoted to showing how the class conflict between the proletariat and the bourgeoisie under capitalism would necessarily undermine capitalist profits, increase worker immiseration, and generate a global revolution. Concerning communist economics, the most that could be concluded from Marx's writings was that it would involve some form of central planning, the gradual abolition of the distinction between town and country, and, most importantly, the elimination of bourgeois private property. Beyond this, as in his political writings, Marx's scattered references to post-revolutionary economics tend to oscillate between a scientific insistence upon the need for strict economic efficiency under socialism and a revolutionary depiction of communist society as freed from all previous economic limitations.

Lenin's own writings on economic matters betray a similar ambivalence about the way to combine rational analysis with revolutionary objectives in economic policy. The New Economic Policy (NEP) he instituted in 1921 reestablished markets for agriculture and consumer goods, but left much of heavy industry, communications, and foreign trade under the control of the party-state. After Lenin's death in 1924, a new debate emerged between a right opposition led by Bukharin, advocating a slow, evolutionary, and scientific approach to socialist economics based upon a long-term continuation of the NEP, and a left opposition, led by Trotsky, advocating revolutionary tempos of industrialization at home and communist advance abroad. As in the Second International, however, the right appeared to sacrifice the revolutionary vision of Marxism to support an uninspiring sort of state capitalism, while the left appeared to lack any realistic strategy to build enduring socialist institutions.

The notion of a unified Marxism–Leninism that might serve as a new ideological orthodoxy for Leninist party members in the Soviet 1920s emerged directly out of this struggle. When Lenin suffered the first of a series of strokes in 1922, a triumvirate of Grigorii Zinoviev, Lev Kamenev, and Joseph Stalin took control of the Soviet state in Lenin's's stead. Already by 1923, Zinoviev and Stalin were declaring themselves to be true 'Leninists' in opposition to the 'objectively antirevolutionary' arguments of Trotsky and his Left Opposition. After Lenin's death and Trotsky's defeat in early 1924, Zinoviev and Stalin competed to position themselves as the sole authoritative interpreters of Marxist and Leninist doctrine. However, Zinoviev's thinking on economic matters was comparatively undeveloped, and with the publication of Stalin's slogan 'socialism in one country' in December 1924—an explicit attack on Zinoviev, who then headed the Communist International—control over the definition of Marxism–Leninism shifted decisively toward Stalin's camp.

By 1929, when Stalin turned on his erstwhile supporter Bukharin and eliminated the Right Opposition, he and his intellectual supporters had worked out the third doctrinal pillar of Soviet 'Marxism–Leninism'—namely, a conception of Marxist 'political economy' designed, in effect, to support a Leninist 'professional revolutionary' centralization of all economic activity in the USSR (Hanson, 1997). Five-year plans were intended to synthesize scientific analysis of the Soviet Union's potential production with the revolutionary heroism of workers and managers bent upon overfulfillment of their monthly and yearly plan targets. The abolition of the distinction between town and countryside was to be achieved through the brutal collectivization of agriculture and the liquidation of the independent peasantry as a class. Finally, the elimination of bourgeois private property was to be cemented by perpetual mass purges and arrests of all suspected collaborators with the global bourgeoisie—a policy that ultimately gave rise to the Great Terror of the mid-1930s.

Thus by 1936, when Stalin declared that 'socialism in the main' had been built in the USSR, Marxism–Leninism was established in the form which became the model for other twentieth-century Leninist regimes. The global power of this ideology was vastly strengthened after the Soviet victory in World War II, when the Red Army imposed Marxist–Leninist doctrine, party rule, and economic planning on the occupied countries of Eastern Europe, while new self-described Marxist–Leninist regimes were established in China and North Korea. However, the brutal violence employed by Stalin to enforce total conformity to his own interpretation of Marxism did not suffice to eliminate the inherent tensions between modern organizational rules and Utopian revolutionary aspirations at the core of Marx's original vision of communism. Indeed, even before the dictator's death, new fissures in the international communist movement began to emerge among Leninists emphasizing rapid revolutionary advance, such as Mao, those gravitating toward supposedly more democratic forms of socialism, such as Tito, and various orthodox pro-Soviet communists defending the Stalinist status quo.

During the post-Stalin era, the increasing inability of Soviet Marxism–Leninism to provide a stable basis for long-term economic growth, political loyalty, or cultural commitment led to the gradual decay of the ideology (Evans 1991). Established Marxist–Leninist regimes everywhere were simultaneously undermined by high-level corruption and mass public cynicism. Various 'deviations' from Marxism–Leninism within the Soviet bloc were crushed by Soviet military force. The last-ditch effort by Mikhail Gorbachev in the mid-1980s to reinvigorate Marxism–Leninism by stripping it of its hierarchical and centralizing features only

hastened its institutional demise, first in Eastern Europe and then in the USSR itself. With the collapse of the Soviet Union in 1991, communist parties and regimes worldwide entered a period of profound crisis. By the turn of the twentieth century, Marxism–Leninism officially persisted as a state ideology—at least formally—only in the People's Republic of China, North Korea, Vietnam, Laos, and Cuba. The original Stalinist conception of a single proletarian world-view institutionalized in a genuinely revolutionary party and a uniquely socialist form of economic development, however, was almost certainly dead.

See also: Communism; Communism, History of; Communist Parties; Dictatorship; Russian Revolution, The; Socialist Societies: Anthropological Aspects; Soviet Studies: Society; Totalitarianism

Bibliography

Acton H B 1955 *The Illusion of the Epoch: Marxism–Leninism as a Philosophical Creed*. Cohen & West, London

Arendt H 1951 *The Origins of Totalitarianism*. Harcourt Brace, New York

Avineri S 1968 *The Social and Political Thought of Karl Marx*. Cambridge University Press, Cambridge, UK

Carver T 1983 *Marx and Engels: The Intellectual Partnership*. Wheatsheaf Books, Brighton, UK

Cohen S F 1973 *Bukharin and the Bolshevik Revolution, A Political Biography, 1888–1938*. Knopf, New York

Deutscher I 1954 *The Prophet Armed: Trotsky, 1879–1921*. Oxford University Press, New York

Deutscher I 1959 *The Prophet Unarmed: Trotsky, 1921–1929*. Oxford University Press, London

Donald M 1993 *Marxism and Revolution: Karl Kautsky and the Russian Marxists, 1900–1924*. Yale University Press, New Haven, CT

Evans Jr. A B 1993 *Soviet Marxism–Leninism: The Decline of an Ideology*. Praeger, Westport, CT

Friedrich C J, Brzezinski Z K 1956 *Totalitarian Dictatorship and Autocracy*. Harvard University Press, Cambridge, MA

Hanson S E 1997 *Time and Revolution: Marxism and the Design of Soviet Institutions*. University of North Carolina Press, Chapel Hill, NC

Harding N 1977 *Lenin's Political Thought*, 2 Vols. St Martin's Press, New York

Kautsky J H 1994 *Marxism and Leninism, Not Marxism–Leninism: An Essay in the Sociology of Knowledge*. Greenwood Press, Westport, CT

Kolakowski L 1981 *Main Currents of Marxism: Its Origins, Growth, and Dissolution*. [trans. Falla P S]. Oxford University Press, Oxford, UK, 3 Vols.

Lenin V I 1977 *Selected Works*. Progress Publishers, Moscow, 3 Vols.

Malia M E 1994 *The Soviet Tragedy: A History of Socialism in Russia, 1917–1991*. Free Press, New York

Marx K, Engels F 1975–93 *Karl Marx Frederick Engels Complete Works*, 47 Vols. Lawrence and Wishart, London

Meyer A G 1962 *Leninism*. Praeger, New York

Meyer A G 1970 *Marxism: The Unity of Theory and Practice*. Harvard University Press, Cambridge, MA

Polan A J 1984 *Lenin and the End of Politics*. Methuen, London

Scanlan J P 1985 *Marxism in the USSR: A Critical Survey of Current Soviet Thought*. Cornell University Press, Ithaca, NY

Schapiro L 1972 *Totalitarianism*. Praeger, New York

Stalin J V 1979 *Selected Works*. 8 Nentori, Tirana, Albania

Steenson G P 1978 *Karl Kautsky, 1854–1938: Marxism in the Classical Years*. University of Pittsburgh Press, Pittsburgh, PA

Tucker R C 1969 *The Marxian Revolutionary Idea*. Norton, New York

Tucker R C 1971 *The Soviet Political Mind: Stalinism and Post-Stalin Change*. Norton, New York

Walicki A 1995 *Marxism and the Leap to the Kingdom of Freedom: The Rise and Fall of the Communist Utopia*. Stanford University Press, Stanford, CA

S. E. Hanson

Marxist Archaeology

Strictly speaking, Marxist archaeology is something which cannot exist. Archaeology is an ancillary historical discipline that seeks to exploit a particular category of evidence (the material residues of past human activities) for historical and sociological purposes. In this sense, Marxist archaeology would be a nonsensical category, like Marxist epigraphy or Marxist numismatics. More broadly, then, Marxist archaeology is taken here to be the use of the principles of historical materialism enunciated by Marx and Engels to analyze the dynamics of the past societies whose material remains archaeologists study.

Here one must take into account, of course, Marxism's political importance, which has made its use as a guiding historical theory range from mandatory to proscribed, with various degrees of toleration in between. Archaeology is an expensive and cumbersome activity that requires extensive funding and official permits, so that political controls are relatively easily imposed, with a corresponding effect on how archaeologists present themselves. Completely uncensored archaeologists may decide (in the interests, not of prudence, but of persuasiveness) to omit explicit references to Marx and Engels on matters concerning which Marx and Engels had nothing direct to say (almost all archaeological cases, that is). Indeed, enough of Marxist thought has become assimilated to the mainstream of the social sciences that archaeologists may be unaware of the derivation of their ideas. One must, therefore, take Marxist archaeology to be constituted by what archaeologists who say they are using Marxist principles do.

1. Soviet Marxist Archaeology

The first broad application of a Marxist approach to archaeology occurred in the Soviet Union. This took the form of applying the ethnological stages of Savagery, Barbarism, and Civilization that Engels (1972 [1884]) had adapted from Lewis Henry Morgan's *Ancient Society* (1964 [1877]) to the archaeological evidence from Soviet territories. During the 1920s most archaeologists carried on with traditional descriptive studies much as before, on the assumption that 'because they were studying the history of material culture, their work accorded sufficiently with the materialist perspective of the new social and political order' (Trigger 1989, p. 215). During the 'Third Period' of collectivization and forced industrialization, however, orthodoxy along the lines indicated by Engels and Morgan was imposed from the center: unenthusiastic scholars were dismissed (or worse), local associations promoting the study and preservation of local monuments and antiquities were disbanded for glorifying the past, and so on (Klejn 1993, p. 21).

The result of this imposition, however, was the first systematic, large-scale sociological reading of prehistoric evidence; one that did not fail to render interesting and useful interpretations and to stimulate innovative research. Reading Engels into the preliterate past of the Soviet Union required a broadly functionalist approach that integrated all aspects of the archaeological record. As Trigger (1989) points out, the concrete results achieved in the 1930s in many ways anticipated those of the Anglo-American New Archaeology of the 1960s; a neo-evolutionist processual movement also rooted in Morgan by way of Leslie White (e.g., 1959). At the same time, however, the high political stakes of Marxist theorizing and the intensely hierarchical institutionalization of archaeology under Stalin impeded critical analysis of Engels's propositions in the light of the emerging evidence, a critique all the more necessary given the provisional character of Engels's work.

On the basis of the new, up-to-date ethnology of Morgan, Engels had been willing to revise extensively the initial reflections on primitive society and the development of the State that he and Marx had presented in works such as the *Formen* (Marx 1965, [1857–8]) and the *Anti-Dühring* (Engels 1934 [1877]), but in the Soviet Union the unilineal stage theory formulated by N. I. Marr, the first director of the State Academy of History and Material Culture, was not formally abandoned until Stalin himself denounced it in 1950, 16 years after Marr's death (Klejn 1993, p. 27). The post-Stalin relaxation of political controls led, as might be expected, less to the development of a critical Marxist archaeology than to a concentration on practical work in which such theorizing as did take place concentrated largely on explaining the diversity of the Soviet Union's archaeological record as an outcome. This focus on 'ethnogenesis' had a greater affinity to the nationalist archaeology of Gustav Kossinna than to the theoretical work of Marx and Engels.

2. Childe

Outside the Soviet Union, the first archaeologist to develop Marxist perspectives systematically was V. Gordon Childe (Trigger 1980). As professor of prehistory at Edinburgh in the 1920s, Childe (1925, 1928, 1929) had developed a synthesis of Kossinna's paleoethnological concerns with Oskar Montelius's comparative chronological approach on a scale that encompassed both Europe and the Near East. Childe was a man of the Left (he had been active in the Labour Party of his native Australia) and he took keen professional and political interest in emergent Soviet archaeology. The Soviet experiment showed him how he could fit the prehistory he had elucidated into a larger evolutionary whole. He would recast the transitions from Savagery to Barbarism and from Barbarism to Civilization as the Neolithic and Urban Revolutions respectively (Childe 1936, 1942), stressing the dynamic interplay of production and social organization.

At the same time, he saw no reason to accept the autochthonic unilinealism of the Soviet school: instead, he anticipated recent world-systems approaches by seeing Europe as peripheral to and dependent on the Near East (Childe 1956a). Finally, he grounded his interpretations in an explicitly realist epistemology (Childe 1956b) in which knowledge is socially constructed, but long-lasting to the extent that it is practical. Precisely because he adjusted the details of his historical reconstructions to the best evidence of his day, many of Childe's specific contentions are dated, but his consistent view that historical changes are rooted in the social relations of production continues to inspire research (e.g., Wailes 1996).

3. 'Western' Marxist Archaeology

Childe's eminence (and British tolerance of what could be regarded as his eccentricity) permitted him to be a solitary exponent of Marxist prehistory in the West during the first part of the Cold War. A broader avowal of Marxist approaches to prehistory would wait until a generation of archaeologists trained during the university-based radicalism of the late 1960s and early 1970s attained professional independence. In the United States and Great Britain this renewed, explicit interest in Marxism arose in the context of dissatisfaction with certain aspects of the by-then dominant New Archaeology. This form of neo-evolutionism shared with Marxism many of the same criticisms of the prevailing historicist theory and practice and so was susceptible in turn to a Marxist critique. (For detailed treatment with references the New Archaeology, see *Evolutionary Approaches in Archaeology*;

Theory in Archaeology.) This critique addresses both of the principal strands of the New Archaeology.

First, the New Archaeology was committed to cultural ecology as an explanatory strategy. Cultures were seen as analytically isolable, functionally integrated, homeostatic regulatory mechanisms in which change occurs due to demographic or environmental pressures external to the cultural system itself. Where small-scale, egalitarian societies are concerned, cultural ecology is not incompatible with Marxism (e.g., Marx 1965, pp. 68–9), although attention to the internal tensions (contradictions) of kinship-organized societies may improve upon an innocent systems functionalism (e.g., Gilman 1984, Saitta 1997). In societies that exhibit substantial inequalities, however, notions that their privileged elites act as higher-order regulators of complex adaptive systems are wide open to critique from a Marxist perspective.

Inasmuch as complex societies are riven by internal competition, the winners of which set themselves above, and exploit the mass of the population, they cannot be adequately understood as systems in which an adaptive equilibrium benefits all their members (e.g., Gilman 1981, Zagarell 1986, Brumfiel 1992). Furthermore, complex societies rarely develop in isolation: their class inequalities emerge in the context of external trade and warfare, so that analogues of the world-system approach to capitalism (e.g., Frank 1967) can be applied to ancient cases (e.g., Frankenstein and Rowlands 1978, Kohl 1989, Frank 1993). Finally, against the New Archaeological view that a society's ideational system consists of the information regulating its adaptive response to the environment, archaeologists have appealed to a Marxist view of ideology as the means by which competing classes and factions legitimate their interests so as to control and manipulate others (e.g., Leone 1982, Patterson 1986, De Marrais et al. 1996). These critiques have been so effective that they have been absorbed into the mainstream of Anglo-American archaeological thought.

Second, the New Archaeology insisted upon emulating the methods of the natural sciences by adopting logical-positivist research strategies. This scientism is clearly incompatible with a Marxist view of ideology, according to which archaeology is a body of knowledge developed by social actors in accordance with their interests. Archaeologists have used Marxist ideas to show how the theory and practice of the discipline has been influenced not (or not just) by value-free scientific objectives, but by the social and economic constraints of its setting (Patterson 1995): thus, the pressures of sponsorship (e.g., Gero and Root 1990) would generate archaeological models (e.g., Keene 1983) and popularizations (e.g., Leone 1981) that reinforce the necessity of present-day capital social relations by showing that they existed in the past.

The argument that archaeological claims about knowledge reflect the interests of those who advance those claims is only a short step away from a self-defeating relativism that sees Marxism as an approach to understanding history that can be judged only on its own terms (e.g., Saitta 1983, p. 301). It is a step that some self-styled Marxists have not failed to take. The concrete reconstructions of the past arrived at by archaeologists (overwhelmingly of European descent) often conflict with the accounts of the past developed by indigenous peoples whom the Europeans have conquered. When the author of a work entitled *A Marxist Archaeology* (McGuire 1992, p. 243) writes that:

> Archaeology … needs to be transformed into the writing of specific peoples' histories as a validation of their heritage … This 'righting' of history is part of a global process in which cultural heritage is being reasserted and is replacing the Enlightenment view of human nature as progressively rational

it is clear that his sympathies have caused him to forget then that Marx, if he was anything, was an exponent of the Enlightenment project. If the New Archaeology's naive positivism has lost ground, this is because of its own inherent difficulties (Wylie 2000), not the Western Marxist critique of it.

4. *'Southern' Marxist Archaeology*

In recent years, the most systematic construction of a Marxist archaeology has occurred within the 'social archaeology' current in Latin America (e.g., Lumbreras 1974, Sanoja and Vargas 1978, Bate 1998, cf. Patterson 1994). Here (as in Spain) Marxism has the prestige of being associated with the opposition to a capitalism operating under authoritarian political auspices. It differs critically from the 'Western' Marxist archaeology in that it is not so much a critique of New Archaeological processualism as a substitute for it. In Latin America the dominant archaeological paradigm is the traditional historicism that preceded the New Archaeology in the US and Britain, a historicism oriented toward an ethnopopulist emphasis on the indigenous roots of the modern nations (Vargas 1995) and/or toward the development of archaeological resources for education and tourism.

Archaeologists politically committed to the Left in Latin America have (where their presence has been tolerated) used a Marxist framework to build a processual alternative to the predominant historicism that does not suffer from the limitations of the adaptationist, ahistorical processualism developed in the US. Classic Marxist concepts, such as 'mode of production' and 'social formation' are combined with novel constructs such as *modo de vida* (mode of life) to develop a systematized basis for archaeological research that has none of the post-processual distaste for an evolutionist grand narrative. In Latin America, institutional constraints have prevented this theory from being put into widespread practice, but in Spain

the democratization of the late 1970s was associated with expanded employment and funding in archaeology, and this in turn has led to a remarkably active implementation of historical materialist research strategies (e.g., Ruiz et al. 1986).

5. Conclusions

The societies that most archaeologists are concerned with were not at the center of Marx's attention, and almost no information about them was available in his and Engels's day. To create a Marxist archaeology, one must, then, critically and creatively deploy the central themes of Marxist thinking. Archaeology is a discipline that requires institutional support, and Marxism, as a critical approach to historical inquiry, is not easily or fruitfully institutionalized. At the same time, an archaeological evaluation of critical Marxist variables, such as property relations ('it is always the relationship of the owners … to the direct producers … which reveals the innermost secret, the hidden basis of the entire social structure' [Marx 1967 [1894], p. 791]), is not a straightforward task (Gilman 1997). As a result, Marxist interpretations of the archaeological record are only beginning to be developed.

See also: Ecology, Cultural; Evolutionary Approaches in Archaeology; Marxism in Contemporary Sociology; Marxist Geography; Soviet Studies: Culture

Bibliography

Bate L F 1998 *El Proceso de Investigación en Arqueología*. Crítica, Barcelona, Spain

Brumfiel E M 1992 Breaking and entering the ecosystem: Gender, class, and faction steal the show. *American Anthropologist* **94**: 551–67

Childe V G 1925 *The Dawn of European Civilization*. Kegan Paul, London

Childe V G 1928 *The Most Ancient East*. Kegan Paul, London

Childe V G 1929 *The Danube in Prehistory*. Oxford University Press, Oxford, UK

Childe V G 1936 *Man Makes Himself*. Watts, London

Childe V G 1942 *What Happened in History*. Penguin, Harmondsworth, UK

Childe V G 1956a *The Prehistory of European Society*. Penguin, Harmondsworth, UK

Childe V G 1956b *Society and Knowledge*. George Allen & Unwin, London

De Marrais E, Castillo L J, Earle T 1996 Ideology, materialization, and power strategies. *Current Anthropology* **37**: 15–31

Engels F 1934 [1877] *Herr Eugen Dühring's Revolution in Science*. Charles H. Kerr, Chicago

Engels F 1972 *The Origin of the Family, Private Property and the State*. International Publishers, New York

Frank A G 1967 *Capitalism and Underdevelopment in Latin America: Historical Studies of Chile and Brazil*. Monthly Review Press, New York

Frank A G 1993 Bronze Age world system cycles. *Current Anthropology* **34**: 383–429

Frankenstein S, Rowlands M J 1978 The internal structure and regional context of early Iron Age society in south-western Germany. *Bulletin of the Institute of Archaeology*. University of London **15**: 73–112

Gero J, Root D 1990 Public presentations and private concerns: Archaeology in the pages of National Geographic. In: Gathercole P, Lowenthal D (eds.) *The Politics of the Past*. Unwin Hyman, London, 19–37

Gilman A 1981 The development of social stratification in Bronze Age Europe. *Current Anthropology* **22**: 1–23

Gilman A 1984 Explaining the Upper Palaeolithic revolution. In: Spriggs M (ed.) *Marxist Perspectives in Archaeology*. Cambridge University Press, Cambridge, UK, pp. 115–26

Gilman A 1997 Como valorar los sistemas de propiedad a partir de los datos arqueológicos. *Trabajos de Prehistoria* **54**(2): 81–92

Keene A S 1983 Biology, behavior, and borrowing: A critical examination of optimal foraging theory in archaeology. In: Moore J A, Keene A S (eds.) *Archaeological Hammers and Theories*. Academic Press, New York, pp. 137–55

Klejn L S 1993 *La Arqueología Soviética: Historia y Teoría de una Escuela Desconocida*. Crítica, Barcelona, Spain

Kohl P L 1980 The use and abuse of world systems theory: The case of the pristine West Asian state. In: Lamberg-Karlovsky C C (ed.) *Archaeological Thought in America*. Cambridge University Press, Cambridge, UK, pp. 218–40

Leone M P 1981 The relationship between artifacts and the public in outdoor history museums. *Annals of the New York Academy of Sciences* **376**: 301–13

Leone M P 1982 Some opinions about recovering mind. *American Antiquity* **47**: 742–60

Lumbreras L 1974 *La Arqueología como Ciencia Social*. Histar. Lima, Peru

Marx K 1965 [1857–8] *Pre-Capitalist Economic Formations*. International Publishers, New York

Marx K 1967 [1894] *Capital, Vol. 3: The Process of Capitalist Production as a Whole*. International Publishers, New York

McGuire R H 1992 *A Marxist Archaeology*. Academic Press, San Diego, CA

Morgan L H 1877 *Ancient Society*. Holt, New York

Patterson T C 1986 Ideology, class formation, and resistance in the Inka state. *Critique of Anthropology* **6**(1): 75–85

Patterson T C 1994 Social archaeology in Latin America: An appreciation. *American Antiquity* **59**: 531–7

Patterson T C 1995 *Toward a Social History of Archaeology in the United States*. Harcourt Brace College Publishers, Fort Worth, TX

Ruiz A, Molinos M, Hornos F 1986 *Arqueología en Jaén (Reflexiones desde un Proyecto Arqueológico No Inocente)*. Diputación Provincial de Jaén, Jaén, Spain

Saitta D J 1983 The poverty of philosophy in archaeology. In: Moore J A, Keene A S (eds.) *Archaeological Hammers and Theories*. Academic Press, New York, pp. 299–304

Saitta D J 1997 Power, labor, and the dynamics of change in Chacoan political economy. *American Antiquity* **62**: 7–26

Sanoja M, Vargas I 1978 *Antiguas Formaciones y Modos de Producción Venezolanos*. Monte Avila, Caracas, Venezuela

Trigger B G 1980 *Gordon Childe: Revolutions in Prehistory*. Thames and Hudson, London

Trigger B G 1989 *A History of Archaeological Thought*. Cambridge University Press, Cambridge, UK

Vargas I 1995 The perception of history and archaeology in Latin America: A theoretical approach. In: Schmidt P R, Patterson T C (eds.) *Making Alternative Histories: The Practice of Archaeology and History in Non-Western Settings*.

School of American Research Press, Santa Fe, NM, pp. 47–67
Wailes B (ed.) 1996 *Craft Specialization and Social Evolution: In Memory of Gordon Childe*. University Museum of Archaeology and Anthropology, Philadelphia, PA
White L 1959 *The Evolution of Culture*. McGraw-Hill, New York
Wylie A 2000 Questions of evidence, legitimacy, and the (dis)union of science. *American Antiquity* **65**: 227–37
Zagarell A 1986 Structural discontinuity: A critical factor in the emergence of primary and secondary states. *Dialectical Anthropology* **10**: 155–77

A. Gilman

Marxist Geography

Geographical issues have assumed mounting significance as society's ability to affect nature has increased. Indeed, geographical issues are among the most profound questions facing societies. They bring into question the very survival of modern humanity. In response, many scholars have assumed critical stances with regard to society's influences on the natural world. Marxist geography is one of these critical analyses. Marxist geography agrees with conventional geography that the focus of geographical study should lie on the relations between society and the natural environment, especially the environmental fringe referred to as earth space, where society's multiple effects on nature are concentrated (see *Geography*; *Nature–Society in Geography*). However, Marxist geographers argue that capitalism, as an entire way of life, has inherently destructive effects on nature and space. Marxist geographers argue that crisis is more severe in regions where capitalism is most developed, organizationally, technically, culturally, and economically. For Marxist geographers, ending environmental crisis means ending the basic cause of the problem, and this in turn entails transforming capitalist societies. This radical intention differentiates Marxist geography, analytically and politically, from conventional social and environmental explanations.

1. *History of Marxist Geography*

The radical movement in geography began as part of a culture of political opposition that coalesced in the mid to late 1960s around social issues like inequality, racism, sexism, and the environment, together with geopolitical issues like the Vietnam War. Articles dealing with socially relevant geographical topics began to appear in the discipline's mainstream journals. Until that time, geography had engaged mainly in a conservative intellectual discourse, whether in the form of regional descriptions, or the quantitative analysis of spatial distributions of cities and industries. In 1969, *Antipode: A Radical Journal of Geography* was founded specifically to publish the new kinds of work. Early issues of the journal dealt with urban and regional poverty, discrimination against women and minority groups, unequal spatial access to social services, advocacy planning, Third World development, and similar topics. For several years radical geography used conventional philosophies and analytical techniques for studying this new range of more socially engaged issues. Only gradually did socially relevant radical geography, armed with an eclectic critical politics, turn into a more disciplined Marxist geography, aimed at the achievement of democratic socialism (Peet 1977).

The transformation of radical geography into Marxist geography was informed by a key text, *Social Justice and the City*, written by Harvey (1973). This series of essays records the author's movement from a liberal, critical position, focused on moral issues of urban social justice, to a revolutionary position, based in a philosophy of historical materialism. As scholars and activists began to read Marx, in part stimulated by Harvey's work, in part propelled by the prevailing culture of protest, and in part compelled by theoretical differences with conventional geography, geographical thought was transformed in content and depth, moving from an obsession with the empirical, the descriptive, and the quantitative, to an equally committed obsession with philosophy, social theory, and the politically engaged. Radical geography has had a strong feminist presence since the early 1980s (Rose 1993). The discipline went through poststructural and postmodern phases in the later 1980s and 1990s (Soja 1989). The last 30 years of the twentieth century saw radical and Marxist geography emerging from the backwaters and assuming a place of prominence in contemporary critical social science (Peet 1998).

2. *Historical and Geographical Materialism*

The philosophical position adopted by Marxist geographers, often termed geographical materialism, is derived from the Marxist philosophy of historical materialism. Karl Marx and Friedrich Engels were followers of the German idealist philosopher G. W. F. Hegel. In Hegelian philosophy, events in the real world were determined by prior thoughts in a separate spiritual realm. Marx and Engels eventually turned the spiritual aspect of this belief back on its material base: consciousness, they claimed, was the product of matter, rather than its origin. Marx retained from Hegel's idealism dialectical conceptions like development through contradiction. However, this was rethought in terms of real, materialist dialectics, as with the struggle between classes, or the contradictory relations between society and nature, so that in-

dividuals and societies were forced constantly to transcend their former identities. In a dialectical, geographical materialism, the natural and social worlds are seen not as entities remaining eternally the same, but as interacting processes capable of rapid, and indeed dramatic, change.

Geographical materialism follows historical materialism by beginning its account of history with the natural bases of life, or the environments inhabited by groups of people, and with the modification of nature through human action. As with other animals, the assimilation of natural materials ('metabolism') was 'the everlasting nature-imposed condition of human existence … common to all forms of society in which human beings live' (Marx 1976, p. 290). Marx argued that human survival had to be earned by the application of labor to the natural world. Human consciousness and intentionality were products of this labor process rather than evidence of God's grace. Yet consciousness, as it evolved, enabled the further development of the forces employed in the social production of existence.

The social relations entered into by working people had a contradictory aspect revealed in endemic social struggles. In particular Marx focused on social relations between a ruling class that owned the means of production—the raw materials, machines, buildings, infrastructures, and technologies that physically produced goods and services—and a working class deprived of independent means of livelihood, and forced to work for wages. Class relations between owners and workers, or between capital and labor, were based in the extraction of surplus labor time, in a process Marx called 'exploitation.'

Societies were exploitative when owners took uncompensated labor time from the direct producers. Used in such a way, money became capital (hence capital was formed by a social labor process rather than private initiative) and social relations between capitalist and laborer were exploitative (hence contradictory in the sense of dialectics). Competitive relations between capitalists compelled the reinvestment of capital in more efficient means of production. Exploitation and competition were the social bases of economic growth under capitalism. For Marx, the exploitation process was an arena of intense social struggle, with the owning ruling class using a combination of economic, political, and ideological force to ensure their dominance, and the working class resisting through overt means like social organization and rebellion, and hidden means, like reluctant compliance. Following this analysis, Marxist geographers as dialectical, geographical materialists, see advanced societies, armed with powerful technologies, organized by competitive and exploitative social relations, having disastrously destructive effects on society, culture and nature, for these societies combine rapid economic growth with lack of social control over the economy.

3. *Social Relations with Nature*

In geographical and historical materialism, humans were originally natural beings whose consciousness was directly interwoven with their reproductive activities. But this was a differentiated unity of people with nature. With permanent surpluses, interactions with nature came to be mediated by unequal social relations. For Marxist geographers this was a contradictory movement: social surplus began a process of emancipation from natural constraints; yet emancipation was accompanied by internal social differentiation marked by institutions of domination like the state, patriarchy, and the market. The precise resolution of the contradiction between emancipation and domination forms the specific kind of society.

A further transformation in human nature occurred with development of the productive forces. With division of labor and commodity exchange, the individual was alienated from work, product and nature. When production came to be primarily for exchange, 'pristine' or 'first' nature was separated from the 'socially-produced' or 'second' nature. Natural relations came to be organized primarily through the exchange of commodities, with nature treated as one more commodity.

The production of second nature speeds up emancipation but sharpens social differentiation. This twin process was key to the bourgeois revolution, the formative social process of capitalism, the making of a class society in which economic growth occurs not for the sake of need, but for purposes of profit and capital accumulation. While all societies have socially-mediated relations with nature, capitalism differs in its higher degree of complexity: under capitalism, natural relations are determined by blind forces of competition rather than deliberative human will. Under the abstract dictatorship of competition, capital must expand continuously, stalking the earth in search of raw materials, appending nature to the production process. This argument leads some Marxist geographers (Smith 1984) to the conclusion that nature is now 'socially produced' in the sense that no original relation with nature is left unaltered and no living thing unaffected. Yet the production of nature universalizes contradiction as capitalism creates new barriers—scarcity of resources, nuclear technology, and pollution of the entire earth environment (see *Feminist Political Ecology*). This suggests, for some, the 'natural inevitability' of social transformation spurred by the 'revenge of nature' (e.g., the greenhouse effect).

4. *Mode of Production and Space*

In such a context of competition, exploitation, struggle, and environmental destruction, Marxists theorize that social consciousness has to be diverted into ideological forms. Primarily, ideologies rationalize

and legitimate exploitation, but they also support competition as efficient, and excuse environmental destruction as necessary for economic growth and continued employment. Marxists have extended the notion of ideology deep into culture, claiming that even common sense is a form of ideology (Gramsci 1971). The total ensemble of social relations, forces of production, ideological forms of consciousness, and political institutions that made up any social system was termed, by Marx, the mode of production. For Marx, social transformations essentially involved shifts from one mode of production to another: from gathering and hunting societies, organized by primitive communal relations; to agricultural societies, organized through tribal social structures; to state societies, organized by payments of tribute and taxes; and finally to industrial societies organized by capitalist social relations of competition and exploitation (Marx 1976).

Marxist geography has maintained a serious effort to clarify the relations between modes of production and space. In some of this work, especially in structural Marxist geography following the ideas of Althusser (Althusser and Balibar 1968), space is seen as a 'level' or 'instance' in a mode of production. Each aspect of social life is theorized as having a particular relation with environmental space, each mode of production creates distinct spatial arrangements, and successions of modes of production alter landscapes in any given space. Here the leading spatial theorist was the urban sociologist Castells (1977). For Castells, any social form, like space, can be understood in terms of the historical articulation of several modes of production. By mode of production Castells means the particular combinations of economic, politico-institutional, and ideological systems present in a social formation, with the economy determining the mode of the production of space in the last instance. In terms of this theory one can 'read' the resulting space according to the economic, political, and ideological systems which formed it. An economic reading concentrates on the spatial expressions, or 'realizations,' of production (factories, offices, etc.), consumption (housing, sociocultural institutions), and exchange (means of circulating goods and ideas). The politico–juridical system structures institutional space according to processes of integration, repression, domination, and regulation emanating from the state apparatus. Under the ideological system space is charged with meaning, its built forms and spatial arrangements being articulated one with another into a symbolic landscape. For Castells (1977, p. 127), the social organization of space can be understood in terms of the determination of spatial forms by each of the elements of the three instances (economic, politico–juridical, ideological), by the structural combination of the three (with economy determining which instance is dominant in shaping space), but also (empirically and historically) by the persistence of spatial forms created by earlier social structures, articulated with the new forms, in interaction with the natural environment.

The intention behind this kind of work was to create a science of space. An organized explanation was attempted of the structural order of the forces and relations that impinge on space. The idea was not how economy is reflected in space, but how economy arranges the political, cultural, and social organization of space or, in a more sophisticated statement, how economy is made up from its various spatial instances. Yet this kind of scientific abstraction sometimes gave structuralist work a quality of unreality, as though the contents of space were laid down by driver-less social machines passing over it. Also the return effects of space on society are minimal and mechanical in some structuralist approaches. Basically the approach lacked the active mediation of human agency, people as individuals, organized by social relations into families, classes and institutions, and informed by different kinds of consciousness (Gregory and Urry 1985).

The themes of social practice, consciousness and space were developed further by the urban sociologist Lefebvre (1991). Lefebvre proposed an analysis of the active production of space to bring the various kinds of space and the modalities of their genesis within a single theory. Lefebvre's thesis, simply put, was that social space is a social product: every society, each mode of production, produces its own space. Space consists of natural and social objects (locations) and their relations (networks and pathways). Yet while space contains things, it is often thought of the relations between them. For Lefebvre people are confronted by many, interpenetrated social spaces superimposed one on top of the other, what he calls a 'hypercomplexity,' with each fragment of space reflecting, yet hiding, many social relationships.

Lefebvre proposes an analysis of space, aimed at uncovering the social relations embedded in it, using a triad of spatial concepts:

(a) *Spatial practice*. Production and reproduction secrete a society's space, forming the particular locations and spatial systems, the perceived spaces characteristic of each social formation. Social-spatial practice, for Lefebvre, is practical, employs accumulated knowledge, yet also involves signifying or semiotic processes.

(b) *Representations of space*. These are abstract conceptions of space, employing verbal and graphic signs ('representations') that are riddled with ideologies subordinating people to the logics of social and political systems.

(c) *Representational space*. This is space as directly lived through associated images and symbols by its inhabitants and users, but also by the writers and artists who describe it. Representational space overlays physical space. Redolent with imaginary and allegorical illusions, it makes symbolic use of natural objects and is therefore studied by ethnologists,

anthropologists, and psychoanalysts as well as sociologists and geographers. Representational space has an affective center of everyday life (ego, bedroom, house), embraces the loci of passion, action, lived situations, and is essentially qualitative, fluid, and dynamic.

The three moments of the triad, often called perceived, conceived, and lived spaces, are held together by dialectical relationships. The triad contributes to the production of space in different combinations according to the mode of production and the historical period. Relations between the three moments are never simple or stable, nor are they entirely conscious. For Lefebvre, applying this triad means studying not only the history of space as spatial practice but the history of representations of space, their interrelations, and their relations with practices and ideologies. In particular, ideology achieves consistency by taking on physicality or body in social space, as urban forms, churches or courthouses for instance. Indeed ideology *per se* might be said to consist primarily of a discourse on social space. So for Lefebvre, representations combining ideology and knowledge within sociospatial practice supplant the Marxist concept of ideology alone, and become the most appropriate tool for analyzing space (see Gottdiener 1995 on the semiotics of space).

5. *Spatial Relations*

Marxist theories see the relations between societies—often termed 'spatial relations; by Marxist geographers—in terms of exploitation, conquest, and domination. During the 1970s a series of articles drew out the spatial implications of Marxist theory in a 'capital logic' school of spatial thought. This kind of Marxist geographic theory began with contradictions in the historical dynamic of capitalism and moved to manifestations of these contradictions in space. Spatial relations between environmentally embedded social formations are thought of as modifying, and even transforming, their internal social, economic, and political contents. Thus, the geography of capitalism is composed of unevenly and differently developing social formations. As social contradictions build internally in any social formation, its external sociospatial relations also change, with effects transmitted elsewhere, or antidotal solutions imported. This may slow down or redirect the build up of contradiction in one social formation, but may qualitatively change the development of social formations elsewhere in space where, in interaction with local (class-dominated) processes, new hybrid social formations come into existence. The complex interplay across space between the social formations of a global system is called 'spatial dialectics' (Peet 1981).

In such approaches, interregional social relations are often theorized as spatial components of a more general internal–external dialectic. For Harvey (1975) the Marxist theory of economic growth under capitalism puts capital accumulation at the center of a dynamic and inevitably expansionary mode of production. This dynamism encounters barriers (in labor, means of production, markets) which precipitate crises (chronic unemployment, realization crises, etc.) with the eventual result of a shift in the capital accumulation process to a higher plane (new social wants, new markets, etc.). Newly 'rational' locational patterns and improvements in transportation and communication are an inevitable, necessary part of capital accumulation. But there is a spatial aspect to contradiction: in overcoming spatial barriers and 'annihilating space with time' spatial structures are created which ultimately act as barriers to further accumulation. This is particularly the case when capitalism comes to rely on immobile, fixed capital (i.e., means of production fixed in space) rather than more mobile variable capital (i.e., labor). Then capital has to negotiate a path between preserving the value of past investments in the built environment (the urban landscape, for example) and destroying these investments to make new room for accumulation (urban renewal). The virtue of Marx's location theory, dispersed though it may be in textual fragments, lies in the way space can be integrated into 'fundamental insights into the production of value and the dynamics of accumulation' (Harvey 1975, p. 13).

In terms of geographical theory this dynamic theory of contradictions contrasts with conventional (bourgeois) location theory employing an equilibrium analysis of optimal spatial configurations: as Harvey claimed (1975, p. 13):

> The Marxian theory … commences with the dynamics of accumulation and seeks to derive out of this analysis certain necessities with respect to geographical structures. The landscape which capitalism creates is also seen as the locus of contradiction and tension, rather than as an expression of harmonious equilibrium … The Marxian theory teaches us how to relate theoretically, accumulation and the transformation of spatial structures and ultimately … the reciprocal relationships between geography and history.

Thus, the internal and external dimensions of space are linked to each other and with capital accumulation. Revolutions in the productive forces, the increasing scale of production, and the concentration and centralization of capital are paralleled by urban agglomeration in a widening international capitalist space.

Smith (1984) argued that different conceptions of space were produced by different types of human activities; hence there was a history of theories of space forming part of the history of the human experience. The increasing scale of human intervention in nature, with advance of the forces of production, is seen as the historical-material basis of a theoretical bifurcation of absolute, natural, and physical space (the world of

physical and natural phenomena) from relative and social space (the humanly-constituted field of social events). For Smith, the development of the forces of production transforms absolute (natural) space into relative (social) space in a contradictory movement. Development emancipates society from the constraints of natural space, leading towards 'equalization,' but only by sinking capital into certain spaces, producing 'differentiation' and a relativized space. These contradictory movements determine the specific forms of capitalist space as synthetic outcome: 'Space is neither leveled out of existence nor infinitely differentiated. Rather the pattern that results is one of "uneven development"' (Smith 1984, p. 90). Integrating the geographical notion of 'scales' (urban, national, global) as levels of specificity, with temporal notions of rhythms, cycles and long waves, Smith proposes a 'see-saw' dynamic of uneven development: 'capital attempts to see-saw from a developed to an underdeveloped area, then at a later point back to the first area which is by now underdeveloped' (Smith 1984, p. 149). Smith finds see-sawing to have a higher propensity to occur at the urban level, where mobile capital destroys, then gentrifies, inner city neighborhoods. For Smith the real resolution between equalization and differentiation can be achieved only by global political cooperation among working-class peoples.

The culminating triumph in a series of this series of creative exegeses came with Harvey's *The Limits to Capital* (1982), a reading of Marx's *Capital* organized around three versions of a general theory of capitalist crisis. The first version sees crises of overaccumulation—more capital produced than opportunities for using it—resolved through violent episodes of the destruction of value, times when capitalists turn on each other, yet workers end up paying the social costs (Harvey 1982, pp. 190–203). The second version examines the expression of productive crises in money and finance, speculative booms and busts which require the intervention of the state: in an internationalized economy, crises take the form of interstate competition over shifting the effects of devaluation, with national policies to export the effects of crises, and imperialism, neocolonialism and even war as potential 'solutions' (Harvey 1982, pp. 324–9). The third and last version integrates geographically uneven development: devaluations of capital are systematized into the continuous restructuring of space through interregional competition. Harvey focuses on 'switching crises' in the movement of capital from one place to another and crises in the hierarchy of capitalist institutions, so that problems which originate in the chaos and confusion of local events build upwards into global crises. For Harvey, the basic rhythm of crisis involves regional economies, or more broadly regional class alliances, periodically alleviating internal crises through transformations in external relations in a spatial 'fix' which again has imperialism as its integral component—hence, 'an interpretation of interimperialist wars as constitutive moments in the dynamics of accumulation' (Harvey 1982, p. 443). What better reason, he asks, could there be for declaring that capitalism's time has passed, and that there is a need for a saner mode of production?

6. *Politics of Space and Nature*

Marxist geographers argue, in general, that the social inequalities endemic to capitalist societies result in spatial inequalities, regional and local differences within and between societies. The capitalist world is split between countries that have far more production and consumption than their peoples need, so that a range of problems appear, from overweight children to excessive pollution, and countries that have too little, with their own range of problems, such as starving children and extreme vulnerability to natural disasters like earthquakes and floods. Uneven spatial development also brings entirely different degrees of economic pressure to bear on the natural world, so that environmental crises appear unevenly as they accumulate into global catastrophes. Some Marxist theories of space, based in dependency and world systems theories (Frank 1969, Wallerstein 1974), argue that overdevelopment in the First World is predicated on underdevelopment of the Third World. This interaction between geography's space dimension and its environmental interests provides the theoretical basis for a range of radical politics linking development with nature. Marxist spatial theorists advocate evening out the development process so that basic needs can be met everywhere. Again this requires fundamental changes in social organization, especially in the ways societies relate over space.

Social scientists in the Western tradition of Marxism favor rational and democratic ways of controlling society's environmental and spatial relations. The notion of state ownership of productive resources has always been far weaker in the Western tradition of Marxism than in the Soviet or Eastern tradition. Marxists believing in all-powerful states have long been countered by others believing just as strongly in workers' self-management, cooperative organization, participatory democracy, and democratic planning. Marxists essentially believe in extending democracy from the political sphere to all aspects of social life, especially the productive and reproductive spheres, that is, democratic control over factories, offices, schools, and families. Marxists think that people directly responsible for their fundamental social and economic activities can reach collective decisions that are more rational and far-sighted than those reached by political and economic elites distanced from direct contact with real social practices. In this way democratic socialism is seen by Marxist geographers as the final culmination of the Enlightenment project of rationality, equality, and liberation.

See also: Development: Socioeconomic Aspects; Human–Environment Relationships; Marxian Economic Thought; Marxism in Contemporary Sociology; Marxist Social Thought, History of; Resource Geography

Bibliography

Althusser L, Balibar E 1968 *Reading Capital*. New Left Books, London

Castells M 1977 *The Urban Question: A Marxist Approach* [trans. Sheridan A]. MIT Press, Cambridge, MA

Frank A G 1969 *Capitalism and Underdevelopment in Latin America*. Monthly Review Press, New York

Gottdiener M 1995 *Postmodern Semiotics*. Blackwell, Oxford, UK

Gramsci A 1971 *Selections from the Prison Notebooks*. International Publishers, New York

Gregory D, Urry J 1985 *Social Relations and Spatial Structures*. St. Martins Press, New York

Harvey D 1973 *Social Justice and the City*. The Johns Hopkins University Press, Baltimore, MD

Harvey D 1975 The geography of capitalist accumulation: A reconstruction of the Marxian theory. *Antipode* **7**: 9–21

Harvey D 1982 *The Limits to Capital*. University of Chicago Press, Chicago

Harvey D 1989 *The Condition of Modernity: An Inquiry into the Origins of Cultural Change*. Blackwell, Oxford, UK

Lefebvre H 1991 *The Production of Space* [trans. Nicholson-Smith D]. Blackwell, Oxford, UK

Marx K 1976 *Capital: A Critique of Political Economy*, (trans Fowkes B). Penguin, Harmondsworth, UK, Vol. 1

Peet R 1977 *Radical Geography*. Maaroufa Press, Chicago

Peet R 1981 Spatial dialectics and Marxist geography. *Progress in Human Geography* **5**: 105–10

Peet R 1998 *Modern Geographical Thought*. Blackwell, Oxford, UK

Rose G 1993 *Feminism and Geography*. University of Minnesota Press, Minneapolis, MN

Smith N 1984 *Uneven Development: Nature, Capital and the Production of Space*. Blackwell, Oxford, UK

Soja E 1989 *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. Verso, London

Soja E 1996 *Thirdspace: Journeys to Los Angeles and other Real-and-imagined Places*. Blackwell, Oxford, UK

Wallerstein I 1974 *The Modern World System*. Academic Press, New York, Vol. 1

R. Peet

Marxist Social Thought, History of

1. Critique of Political Economy

In the twentieth century Marxist social thought has taken many different and opposing forms: revisionist, revolutionary, structural, critical, humanist, New Left, social relations, rational choice, etc. Their common denominator is that they appeal to Marx as their original source.

Marx himself began with a critique of modern natural law theory; that is, the way of thinking which dominated the Western tradition of political and economic thought. Modern natural law theory took two main forms, empirical natural law which prevailed within political economy and idealist natural law which prevailed in political philosophy. Marx argued that modern natural law theory made vast strides in advancing social thought and in identifying the domain of the social. However, he thought that the discursive framework within which classical political economy and philosophy moved failed to distinguish adequately between what is social and what is natural, or more precisely failed to distinguish the place of the social in nature as a whole (Fine 1984).

Marx saw classical political economy as having made great advances in understanding the social character and historical origins of the economic forms of the modern age—value, exchange value, price, money, capital, interest, rent, profit, etc. It recognized that human labor is the ground of value and that the economic system as a whole only reached fruition in the modern age—at the end of history rather than at its beginning. Marx argued, however, that classical political economy naturalized labor as the source of value and never questioned in what kind of social organization or under what social circumstances labor would take the form of value. The way Marx put this is that in analytical terms political economy was strong: it perceived that the magnitude of value was determined by the average amount of labor time that goes into the production of a commodity. But dialectically political economy was weak: it treated the fact that things take the form of commodities and are bearers of value as a natural fact of life rather than as the product of determinate social relations, and it treated the historical emergence of the modern commercial, commodity-producing economy as the triumph of reason over the manifold forces which constrained its fruition in the past (see *Political Economy, History of*).

Vis-à-vis empirical natural law theory, Marx saw himself as the first to comprehend adequately the social character of the value form and to release it from the naturalistic framework in which it had been captured by natural law. The claim to be the first was true, although he did not recognize that Hegel's critique of natural law pre-empted and prefigured his own—a fact which has only recently been recognized by Marx–Hegel scholars (Rose 1981). Marx wrote *Capital* as a critique of political economy: not as an economics text but as a study of a society dominated by the dull compulsion of economic forces; not as an economic determinism but as an analysis of a society in which the economic is determinant. The subject matter of *Capital* concerned the inhuman, alienated,

and exploitative social relations that lie hidden behind the fetishized forms of modern economic rationality.

To read Marx as an economic determinist as many of his followers have done, or to accuse him of economism as many of his critics have done, is to miss the mark in as much as *Capital* was a critique of a social world in which:

(a) The exchange of things is a primary condition of intersubjectivity.

(b) Things appear as bearers of exchange value or prices.

(c) Access to things is primarily mediated by the ability to purchase them.

(d) Human activity is subordinated to the movement of things and the fluctuations of the market (for the more traditional reading of Marx's economics, see Dobb 1946, Meek 1973, Sweezy 1969, Mandel 1978).

Capital was a critique of a society which actualizes economism. It treated the very idea of 'the economic' as a *product* of definite social relations of production and conceived a truly human, social world in terms of the overcoming of 'the economic' as such. *Capital* was an attempt to understand a social world in which everything has its price, even the capacities of our body and soul, and humanity is a slave to the products of its own labor (for this social reading of Marx's economics, see Sayer 1979, Clarke 1982).

The key proposition in Marx's 'economic' writings was that economic forms and categories are the visible expression of determinate social relations of production. It is misleading, therefore, to say that Marx called the economy the base on which legal, political, and ideological superstructures rest (Williams 1983). Marx was not consistent in his use of terminology but if we can still speak of a 'base,' it is constituted by social relations and not by their economic forms. The imagery that informs *Capital* is not that of base and superstructure but of form and content—economic form and social content. The approach that Marx adopted was to start by analyzing the forms themselves (commodity, value, use value, price, money, capital, interest, profit, etc.), then uncover the alienated social relations concealed behind these forms, and finally explain why these social relations manifest themselves in the form of material relations between things (Rubin 1972). Marx called this post-enlightenment form of superstition and domination the 'fetishism of the commodity.'

2. *Critique of Political Philosophy*

One potential weakness of this social approach to Marx's 'economic' writings is that it might give the impression that the economic forms of capitalist social relations are its only forms, or at least that they are its essential forms. Other noneconomic forms of modern social life—moral, legal, political, cultural, etc.—might thus be perceived as being in some sense epiphenomenal or inessential. Even when it was recognized that Marx offered a social critique of political economy rather than an economics, there still remained something rather arbitrary in the way in which he treated the economic as a privileged sphere of social life. Within traditional Marxism this led to a number of possible solutions. At one pole, that of revisionist Marxism, legal, political, and cultural forms usually appeared to be entirely dissociated from and independent of capitalist social relations (as in Eduard Bernstein's *Evolutionary Socialism*). At the other pole, that of revolutionary Marxism, they usually appeared to be entirely determined by capitalist social relations (as in Lenin's *State and Revolution*). In the middle they have often appeared in the manner of structuralist Marxism to be 'determined in the last instance' by economic forces or to be 'relatively autonomous' and have their own efficacy (Althusser 1965, Poulantzas 1980). Within most forms of contemporary Marxism, it still seems that there are immediate internal connections between capitalist social relations and economic forms that are not shared between capitalist social relations and the moral, legal, political, and cultural forms of the modern age.

For some Marxists, particularly those confronted by the rise of fascism in the 1930s, there was a need to rescue Marxism from the grip of economists and revive it as a theory of the whole social totality. The key to achieving this was to reconsider the relations between Marx and the tradition of German idealism out of which he emerged, and in particular the relation between Marx and Hegel. This was characteristic of those Hegelian Marxists who belonged to or were influenced by the Frankfurt School of Critical Theory. They argued that, if economic determinism is one aspect of our social world (the aspect which Marx analyzed in detail), the other is that which Hegel called the 'right of subjective freedom' which he analyzed in its various legal, moral, political, and aesthetic dimensions (e.g., Hegel, *Philosophy of Right*). The basic insight of Hegelian Marxism in all its forms was that there is more to capitalism than the circuits of capital: there is Kant as well as Bentham, political philosophy as well as political economy, the fetish of the subject as well as the fetish of the commodity, free will, morality, and personification as well as determination, instrumental rationality, and reification (Lukács 1971, Lukács 1975, Marcuse 1989, Adorno 1973). However, the most characteristic gesture of critical theory was to treat the forms of right, law, and morality either in terms of the logic of illusion (the illusions of liberal individualism) or a logic of anachronism (treating aspects of bourgeois society in the period of its ascendancy as if they were still operative in its period of decline; see *Critical Theory: Frankfurt School*).

The one-sided, economic view of capitalist social relations has also been unconvincing to non-Marxists who have recognized that the modern age conveys

ideas of personality, free will, moral agency, individual rights, legal equality, collective self-regulation, etc. as well as material relations between things (Kolakowski 1978, Lichtheim 1964). In response, Marxists of various stripes have wanted Marxism to acknowledge that the individual is a juridical, moral, political, and cultural subject as well as a 'bearer' of economic forces and that bourgeois society produces 'high' moral, political, and cultural values as well as the bare cash nexus. If it is inadequate to say that the language of individual right and moral agency is merely an illusion, or that it is an anachronism inappropriate to our own age, then the difficulty is how to understand the place of this language within the totality of social life and how to move, as it were, from the circuits of capital to capitalism as a whole.

3. *Humanist Marxism and Stalinism*

In the latter half of the 1950s, a new wind was beginning to blow. In the East it was marked by mounting dissidence partly in reaction to the revelations made at the Twentieth Congress of the CPSU, the 1956 Hungarian uprising and its suppression by Soviet troops. In the West it was marked by the birth of the New Left and of the new social movements (antiwar, antinuclear, shop stewards, etc.), which fought against the moral myopia characteristic of the Cold War and for a perspective capable of looking beyond existing social forms and comprehending new social forces. In the Colonies it was marked by anti-colonial and anti-imperialist revolts whose main aim was national self-determination (Fanon 1965). Within all of these movements there were residues, often strong, of a more traditional Marxism. However there was also growing evidence of a Marxism which put human needs before dogma, and social relations before institutional arrangements, which recognized that choices can and must be made along the way to socialism and that conscious human agency plays a part in the making of history. It was a Marxism which rejected the coupling of socialism and barbarism that was the mark of Stalinist rule and which sought to reconnect Marxism to the idea of humanity (see *Totalitarianism*).

As an alternative to the deformed socialism of the post-Stalinist states, there was posited a humanist Marxism which reaffirmed faith in the revolutionary potential not of the human race or of the dictatorship of the proletariat, but of real men and women. Such dissidents were denounced by the official communist press, and sometimes by the more orthodox Trotskyist opposition, for the sins of idealism, subjectivism, romanticism, clericalism, and humanism. But they represented the rediscovery of the 'social' in socialism that had been squeezed out of existence once the latter was turned into a ruling abstraction (see Dunayavskeya 1964, Draper 1977, James 1992).

Marxist humanism emphasized Marx's own awareness of the limitations of a critique in which 'the connection of political economy with the state, law, morality, civil life, etc. is only dealt with insofar as political economy itself professes to deal with these subjects,' as well as his ambitious life-project: to 'present one after another a critique of law, of morality, politics, etc. … and then finally … to show the connection of the whole.' However, as the Marxist historian Edward Thompson put it, there was also a growing sense that Marx himself was 'trapped within the circuits of capital' and 'only partly sprung that trap in *Capital*' (Thompson 1978). It seemed that Marx was increasingly sucked into the theoretical whirlpool of political economy whose categories were interrogated and re-interrogated but whose main premise, the possibility of isolating the economic from other fields of social study, was left intact. The structure of *Capital* appeared to be dominated by the categories of its antagonist, namely the economic sphere itself, and to remain fixed within this order of things.

It was widely argued that, although Marx could not be blamed for the deformations of traditional Marxism, neither could he be simply exonerated. There was no 'virgin untouchable purity of original Marxism to which Stalinism owes no kinship' (Thompson 1978). If Marx's one-sidedness appeared, like Antigone's commitment to family love, wrong only because it was one-sided and to be valid within its own terrain, the ideology of Stalinism was seen to justify the exercise of totalitarian terror against both its 'own' people and subject nations, and to conceal the social relations (within the workplace, within the party, within the trade unions, within everyday life) which lie beneath absurdly idealized conceptions of state property and workers government.

Affirming the unity of human experience, humanist Marxism treated Marx's failure to explore other subjects systematically as symptomatic of a certain imprisonment within the very economic categories whose social content he dedicated himself to understanding. It argued that law, morality, culture, etc. belonged to a different logic from that of capital and that the influence of the latter should be conceived as one of corruption rather than constitution. This attitude might be exemplified in Thompson's assertion at the end of his study of class law in eighteenth-century England that the rule of law itself is an 'unqualified human good' (Thompson 1977). Most important, Marxist humanism maintained that the philistine character of official Communist ideology lay in its denial of the creative agency of people. When people are considered merely as units in a chain of determined circumstances, what disappears is the fact that we are moral and intellectual beings capable of making our own history, confronting adversity, and surmounting the limitations imposed by circumstances. It was this denial of human agency that Thompson called a 'heresy against man' and which

destroyed from within the social dimension of Marxism. In the Communist world this heresy against people took the shape of a pseudo-Marxist ideology which served only to buttress the ruling bureaucracy.

4. New Left Marxism

The rise of nontraditional forms of Marxism—critical theory in the 1930s, humanist Marxism in the 1950s, and then New Left Marxism in the 1960s—was accompanied by the discovery, translation, and dissemination of the young Marx's early writings. These included his analysis of alienated labor as the key to understanding private property (in the *Economic and Philosophical Manuscripts*), his attack on the latent authoritarianism of the modern state that he (wrongly) attributed to Hegel's philosophy of right (*Critique of Hegel's doctrine of the state*), and his revolt against the anti-Semitism he detected within some sections of the German socialist movement (*On The Jewish Question*). The recovery of Marx's analysis of the modern social division between the self-interested atomism of civil society and the abstract community of the political state—and its manifestation as a split within every individual between the *bourgeois* and *citoyen*—helped rescue the social character of Marx's critique from both the stagnant depths of economic determinism and the giddy heights of political philosophy. In practice this meant the renewal of the critique of the cultural, legal, and political forms of modern society. This included state property as well as private property, collectivity as well as individual personality, bureaucracy and the Party as well as representative government, and the endeavor to overcome the bourgeois and socialist mystifications which surrounded these forms of cultural, legal, and political life (Colletti 1974, McLelland 1969).

The exposure of Marx's early writings to public discussion was accompanied by the publication (in German in 1953 and translated into English in 1973), of Marx's rough draft of *Capital*, known as the *Grundrisse* and written in a rush of revolutionary expectation in 1857–8 (Nicolaus 1973, Rosdolsky 1977). This made Marx's debt to Hegel in his later 'scientific' writings far clearer than it had been. There was interplay between the revelation of previously unknown, untranslated, or obscure writings by Marx and what contemporary Marxists themselves wanted to find in Marx's work. It brought to the fore a thematic present on the margins of *Capital* but more evident in the *Grundrisse*, namely Marx's critique of the alienated social forms of modern 'subjectivity': the forms taken not by the products of human labor but by the producers themselves, the subjects of human labor who produce goods, bring their commodities to the market, interact with one another, express their creativity and take their leisure.

New Left Marxism, with its very strong spirit of rebellion, discovered an implicit hypothesis in Marx's work: that the fetishism of the commodity (the product of human labor) is accompanied by the fetishism of the subject (the producer or laborer) as the split halves of an original unity. Now everything was conceived as a 'social relation'—money, capital, the law, the state, abstract right, etc.—and the times looked promising for Marxist *social* thought. There was a revival or resurgence of Marxist writings on the critique of law (Pashukanis 1983, Fine 1984), the critique of the capitalist and socialist state (Holloway and Picciotto 1978, Clarke 1991), the critique of bourgeois culture (Jameson 2000), the critique of alienation (Meszaros 1970, Ollman 1973), the conditions of radical democracy (Miliband 1988, Habermas 1974), etc.

In post-1968 politics, Marx's early writings were often cited in order to dismiss the whole complex, differentiated edifice of bourgeois authority and democracy as a sham and to put in its place an alternative conception of 'true democracy.' New Left Marxism distinguished between two traditions: that of formal democracy aligned to representation, and that of true democracy aligned to an anti-representational tradition of natural law indebted to Rousseau. The idea was taken from the young Marx, that the so-called democratic state allows the people to appear only as 'fantasy, illusion, representation,' that it offers no more than a 'ceremony' or 'spice' of popular existence, that it expresses 'the lie that the state is in the interest of the people,' and that the existing world has to be 'turned on its head' if the people are to enter the political stage as they should, in person rather than through representation, in flesh and blood rather than in name alone, in actuality rather than in mere form.

This was the thread that Lucio Colletti picked up in his introduction to the Penguin edition of *Marx's Early Writings* (1974). He argued that the revolutionary tradition *From Rousseau to Lenin* (1972) had as its center the critical analysis of parliamentarism and of the modern representative principle itself, and the conviction that sovereignty must no longer be transferred to government by the people but be retained by the people themselves. The idea of revolution was represented not as a transfer of power from one class to another but a transition from one form of power to another—from the alien form of the political state in which the people are only ideally present, to a power that is 'directly into the hands of the people.'

This spirit of radicalism was caught, for example, in Guy Debord's *Society of the Spectacle* (1977). He directed the critique of representation east and west: at a Leninism where the representation of the working class radically opposes itself to the working class, and at a parliamentarism where the people in miniature opposes itself to the people themselves. He depicted the representative principle as 'the quintessence of modern domination … the concrete inversion of life … the heart of the unrealism of the real society.' The

Commune/Council, by contrast, was shown as the form in which all prior divisions between rulers and ruled are finally overcome and where specialization, hierarchy, and separation end. Debord construed the latter as an institution that is fundamentally anti-institutional, a form of representation which only actuates the generalization of communication, a form of organization which recognizes its own dissolution as a separate organization, a power which can no longer combat alienation with alienated forms.

This vision of the Commune/Council was vulnerable to the criticism that it was as abstract as the power against which it protested. It encountered the realist criticism that, viewed as an institutional system and bared of its revolutionary mystique, the Commune/Council manifests its own tendency toward hierarchy and alienation. The rationalist criticism argued that, judged in terms of its capacity for rational decision-making, the face-to-face structures of the Commune/Council fall victim to the vagaries of public opinion or the contingencies of the loudest voice. The modernist criticism argued that the idea of a true democracy echoed an old European tradition of natural law theory, eventually going back to Aristotle, which conceived the political state and civil society as an undifferentiated unity.

5. *Contemporary Marxism*

Marx often defined himself through his opposition to Hegel. He praised Hegel for having discovered the 'correct laws of the dialectic' but indicted him for mystifying the dialectic. He argued that in Hegel the dialectic was 'standing on its head' because it transformed the process of thinking into an independent subject; while he, Marx, recognized that 'the ideal is nothing but the material world reflected in the mind of man and translated into forms of thought' (Postface to *Capital 1*). He argued that Hegel's mystical dialectic served as a philosophy of the state to 'glorify what exists,' while his own dialectic was by contrast 'a scandal and an abomination to the bourgeoisie' because it 'included in its positive understanding of what exists a simultaneous recognition of its negation, its inevitable destruction.' Hegel was Marx's *Doppelganger*: the ghostly double that allowed Marx to be Marx. By making Hegel logical and conceptual, Marx constructed himself as historical and material.

This account was deceptive both in respect of Hegel and Marx himself. In fact, Marx's analysis of the value-form was no less logical and no more historical than Hegel's analysis of the form of abstract right. Their respective methodologies were almost identical (Fine 2001). Within contemporary Marxism the relation between Marx and Hegel continued to confound scholars who echoed Marx's own conviction that Hegel got everything upside down and that Marx was the one who put the world and Hegel back on their feet. There was a tendency within Marxism to treat what Marx said about Hegel as the truth of Hegel without any independent judgement, and to fixate on the logic of inversion which overshadowed Marx's own writings. One result of this was an increasing stress on the historical specificity not only of the objects of Marx's analysis (such as the economic forms of value, money, capital, etc.) but also of the critical consciousness that attempts to grasp them.

For example, in his critique of traditional Marxism and revival of critical theory, Moshe Postone (1996) radicalizes Marx by pushing the idea of historical specificity to its limit. He argues that not only are the economic categories of Marx's theory historically specific but also the categories Marx employs to analyze these economic forms, especially that of labor. Postone's charge against traditional Marxism is that it does not historicise enough, that it treats certain aspect of Marx's theory, say the mode of production in contrast to the mode of distribution, as transhistorical. Postone pushes to the fore the historical specificity not only of the commodity form, but also of industrial production, the working class, the category of labor as a tool of analysis, and even the epistemological opposition between subject and object. He looks to the transformation, abolition, supersession, destruction and/or transcendence not only of the historically specific forms of social life characteristic of our age, but also of the historically specific critical consciousness whose aim is to abolish both these forms.

This type of Marxism takes to the limit Marx's critique of bourgeois ideology: that is, the tendency to universalize the particular and to eternize what are in fact historically specific values, including those of science and critique themselves. Its conception of history has in turn opened up new areas of debate. The idea that a relation is historically specific, however, tells us little about the nature of the relation itself or about whether it ought to be abolished. It reflects rather the culture of a historically self-conscious age that declares that both knowledge and reality are transitory and surpassable and that what seems unquestionably true to one age differs from what seems unquestionably true to another. It is a culture that discloses a variety of conflicting historical 'truths' with no transhistorical criteria for judging between them. It opens the door to a variety of 'post-Marxist' outcomes: a skeptical paralysis which abandons all faith in knowledge and enlightenment; a pragmatic make-believe which hopes that a pretended faith might do the work of an actual faith; a new faith in democracy and democratization which offers a singular political solution to all the alienated forms of social life; or even a new kind of ideological fanaticism which seeks to achieve certainty only by making itself true (see *Ideology, Sociology of*).

Marx argued in *The Communist Manifesto* that the fleeting transitoriness of all values, relations, forms, and types of knowledge was an expression of a

practical nihilism inherent in bourgeois society: 'All fixed, fast frozen relations, with their train of ancient and venerable prejudices and opinions, are swept away, all new formed ones become antiquated before they can ossify. All that is solid melts into air.' Bourgeois society, Marx wrote, drowns religious fervor, chivalrous enthusiasm, and philistine sentimentalism in the icy water of egotistical calculation; it resolves personal worth into exchange value; it strips of its halo every occupation hitherto honored; it reduces the family to a mere money relation; it profanes all that is holy; it substitutes expansion, convulsion, uncertainty, motion, smoke for everything that was permanent, fixed and certain. The bourgeois in this account sees only the transitoriness of things in a world where nothing lasts, nothing has value. Marx emphasized the revolutionary character of the modern bourgeoisie in confronting traditional philosophy and in devaluing all eternal values and timeless ideas. Yet to the extent that Marxism turns the idea of historical specificity into a doctrine of movement, transitoriness, and surpassability, rather than into an investigation of the burden of history bearing down on the present, it begins to mimic the destructive aspect of bourgeois consciousness. It sees the historicity of things only in the sense that every existing social form is equally deprived of value.

In the face of these difficulties within contemporary critical theory, Jürgen Habermas (1996) and the New German School of Critical Theory have reaffirmed the rational aspect of Marxism by appealing to the rational structures of an ideal speech situation. Habermas launched a major critique of the attachment of Marxist social thought to what he reads as the nihilistic ways of thinking characteristic of postmodernism and has endeavored to reconnect Marxism with a republican tradition as indebted to Kant as to Rousseau. In pursuit of a postmetaphysical theory of right, he draws heavily on Kant's *Metaphysics of Justice* in reconceptualizing radical democracy as a differentiated, complex, lawful, and representative system of right, buttressed by an active civil society and open public sphere. Habermas achieves this re-appropriation of Kant by proceduralizing natural right and by thus overcoming what he sees as Hegel's emphatic statisation of right and Marx's equally emphatic socialization of right. Hegel and Marx's radical critiques of natural right theory are both viewed with suspicion.

6. *Conclusion*

The problem of what Marxist social thought is, is not resolved. One wing of contemporary Marxism, sometimes called the Social Relations School, has perhaps protested too much that x and y and z are 'social relations' and has thereby lost sight of the natural world of things and people, use values, and human beings, which inhabit these social relations. Social relations are turned into a subject in their own right and everything else appears by comparison inessential and derivative. The reification of 'the social' as the foundation of all cultural, political, and economic forms has in turn sparked either antifoundationalist critiques of Marxism or attempts to reconstruct Marxism as a form of antifoundationalism (Nancy 1991, Derrida 1994). The other wing of contemporary Marxism is known as analytic or rational choice Marxism. It has been characterized not only by its respect for traditional canons of argumentation to which it subjects Marxist propositions, but also by a methodological and normative individualism which seeks to reinstate the individual in Marxist thought, both in terms of historical explanation of existing society and the formulation of a just society. Rational choice Marxism may be seen as reflecting the rationalization and individualization of contemporary Western society. However, the proposition that it is possible to analyze social life as if its basic units were rational individuals is a statement of social ontology which takes the perceived weakening of social bonds within the modern age as its starting point (Cohen 1978, Elster 1985).

Marxist social thought is now considered 'dead' by many commentators. This perception is due partly to external reasons, such as the decline and fall of official Communism, but also to internal deficiencies within contemporary Marxism itself. Thus the sociology of Erving Goffman and the discourse theory of Michel Foucault seem to have more to say about power relations in the prison, asylum, factory, family, army, school, police, everyday life, etc. than Marxist state-theorists. Yet Marxism remains a crucial resource if the drive to dissociate power from social life or to subordinate social life to the unidimension of power is not to be entrenched.

Marxist social theory is an international phenomenon with strong roots in the West, the East and the Colonies. It has now entered into a new era. It has lost the state support it received from the Soviet Union and its satellites. It has lost the ideological closure with which it was once associated. Its attachment to social movements has also declined. It is now part of a post-Communist age and has to respond to new problems of globalization. In this transition, the history of Marxist social thought is itself being reviewed. Marx's relation both to the natural law tradition and to the other nineteenth century rebels against the tradition, from Hegel to Nietzsche, has been revisited in more open and critical ways. The distinction between Marxism and other forms of twentieth century social thought has been made less sharp and severe. What comes to mind is the thought that Marxist social thought does *not* have its own history and that what we call its history is a manifold of stories each construing itself as the narrative heir to Marx's

writings. The potentiality this opens up is of a new wave of social thought which will draw inspiration from Marx's writings without all the old demarcations and dependencies.

See also: Aristotelian Social Thought; Communism; Critical Theory: Frankfurt School; Existential Social Theory; Gramsci, Antonio (1891–1937); Marx, Karl (1818–89); Marxian Economic Thought; Marxism in Contemporary Sociology; Marxism/Leninism; Marxist Archaeology; Marxist Geography; Nostalgic Social Thought; Pragmatist Social Thought, History of; Socialism; Utilitarian Social Thought, History of; Weberian Social Thought, History Of

Bibliography

Adorno T 1973 *Negative Dialectics* [trans. E. B. Ashton]. Seabury Press, New York; Routledge, London

Althusser L 1965 *For Marx*. Penguin, Harmondsworth, UK

Anderson P 1976 *Considerations on Western Marxism*. NLB, London

Clarke S 1982 *Marx, Marginalism and Modern Sociology: From Adam Smith to Max Weber*. Macmillan, London

Clarke S (ed.) 1991 *The State Debate*. St. Martin's Press, New York

Cohen G A 1978 *Karl Marx's Theory of History: A Defence*. Clarendon Press, Oxford, UK

Colletti L 1972 *From Rousseau to Lenin: Studies in Ideology and Society* [English translation by Merrington J, White J]. New Left Books, London

Colletti L (ed.) 1974 *Karl Marx Ey Writings*. Pelican, Harmondsworth, UK

Debord G 1973 *Society of the Spectacle*. Black and Red, Detroit, MI

Derrida J 1994 *Specters of Marx: The State of Debt, the Work of Mourning and the New International* [English translation by Kanuf P]. Routledge, New York

Dobb M H 1946 *Studies in the Development of Capitalism*. Routledge and Kegan Paul, London

Draper H 1977 *Karl Marx's Theory of Revolution*. Monthly Review, New York

Dunayevskaya R 1964 *Marxism and Freedom: From 1776 Until Today*, 4th edn. Twayne, New York

Elster J 1985 *Making Sense of Marx*. Editions de la Maison des Sciences de l'homme, Paris

Fanon F 1965 *The Wretched of the Earth*. [Translated from the French by C. Farrington] Grove Press, New York [1965 c.1963]

Fine R 1984 *Democracy and the Rule of Law: Liberal Ideals and Marxist Critiques*. Pluto, London

Fine R 2001 *Political Investigations: Hegel, Marx and Arendt*. Routledge, London

Habermas J 1974 *Theory and Practice* [English translation by Viertel J]. Heinemann, London

Habermas J 1996 *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* [trans. W. Rehg]. MIT Press, Cambridge, MA

Holloway J, Picciotto S (eds.) 1978 *State and Capital: A Marxist Debate*. Arnold, London

James C L R 1992 In: Grimshaw A (ed.) *The C L R James Reader*. Blackwell, Oxford, UK

Jameson F 2000 In: Hardt M, Weeks K (eds.) *The Jameson Reader*. Blackwell, Oxford, UK

Kolakowski L 1978 *Main Currents of Marxism* [trans. from the Polish by P. S. Fallen]. Clarendon Press, Oxford

Lichtheim G 1964 *Marxism: an Historical and Critical Study*, 2nd edn. Routledge and Kegan Paul, London

Lukács G 1971 *History and Class Consciousness: Studies in Marxist Dialects [by] Georg Lukás* [Trans. by R. Livingstone]. Merlin, London

Lukács G 1975 *The Young Hegel: Studies in the Relations between Dialectics and Economics* [English translation by Livingstone R]. Merlin Press, London

Mandel E 1978 *Late Capitalism* [English translation by De Bres J]. NLB, London

Marcuse H 1989 *Reason and Revolution: Hegel and the Rise of Social Theory*, 2nd edn. Humanities Press, Atlantic Highlands, NJ

McLelland D 1969 *The Young Hegelians and Karl Marx*. Macmillan, London

Meek L 1973 *Studies in the Labour Theory of Value*. Lawrence & Wishart, London

Meszaros I 1970 *Marx's Theory of Alienation*. Merlin, London

Miliband R 1977 *Marxism and Politics*. Oxford University Press, Oxford, UK

Nancy J-L 1991 *The Inoperative Community* [trans. by P. Connor, L. Garbus, M. Holland and S. Sawhey]. University of Minnesota, Minneapolis, MN

Negri A 1991 *Marx Beyond Marx: Lessons on the Grundrisse* [trans. H. Cleaver, M. Ryan and M. Viano]. Pluto, London

Nicolaus M 1973 'Forward' to Marx, *Grundrisse*. Pelican, Harmondsworth, UK

Ollman B 1973 *Alienation: Marx's Conception of Man in Capitalist Society*. Cambridge University Press, New York

Pashukanis E 1983 *Law and Marxism*. Pluto, London

Postone M 1996 *Time, Labor and Social Domination: A Reinterpretation of Marx's Critical Theory*. Cambridge University Press, Cambridge, New York

Poulantzas N 1980 *State Power Socialism*. Verso, London

Roemer J E 1982 *A General Theory of Exploitation and Class*. Harvard University Press, Cambridge, MA

Rose G 1981 *Hegel Contra Sociology*. Athlone, London

Rosdolsky R 1977 *The Making of Marx's Capital*. Pluto, London

Rubin I 1972 *Essays on Marx's Theory of Value* [originally published in Russian in 1928. English translation by Samardzija M, Perlman F]. Black and Red, Detroit, MI

Sayer D 1979 *Marx's Method: Ideology, Science and Critique in 'Capital.'* Harvester Press, Hassocks, Sussex, UK

Sohn-Rethel A 1978 *Intellectual and Manual Labour: A Critique of Epistemology* [originally published 1970. English translation by Sohn-Rethel M]. Macmillan, London

Sweezy P 1969 *The Theory of Capitalist Development*. Monthly Review, New York

Thompson E P 1977 *Whigs and Hunters*: *The Origin of the Black Act*. Penguin, Harmondsworth, UK

Thompson E P 1978 *The Poverty of Theory and Other Essays*. Merlin, London

Uchida H 1988 Carver T (ed.) *Marx's Grundrisse and Hegel's Logic*. Routledge, London

Williams R 1983 *Keywords: A Vocabulary of Culture and Society*. Fontana, London

R. Fine

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Masculinities and Femininities

Masculinities and femininities refer to the social roles, behaviors, and meanings prescribed for men and women in any given society at any one time. As such, they emphasize gender, not biological sex, and the diversity of identities among different groups of women and men. Although we experience gender to be an internal facet of identity, masculinities and femininities are produced within the institutions of society and through our daily interactions (Kimmel 2000).

1. *'Sex' vs. 'Gender'*

Much popular discourse assumes that biological sex determines one's gender identity, the experience and expression of masculinity and femininity. Instead of focusing on biological universals, social and behavioral scientists are concerned with the different ways in which biological sex comes to mean different things in different contexts. 'Sex' refers to the biological apparatus, the male and the female—our chromosomal, chemical, anatomical, organization. 'Gender' refers to the meanings that are attached to those differences within a culture. 'Sex' is male and female; 'gender' is masculinity and femininity—what it means to be a man or a woman. While biological sex varies very little, gender varies enormously. Sex is biological; gender is socially constructed. Gender takes shape only within specific social and cultural contexts.

2. *Why Use the Plural?*

The use of the plural—masculinities and femininities—recognizes the dramatic variation in how different groups define masculinity and femininity, even in the same society at the same time, as well as individual differences. Although social forces operate to create systematic differences between men and women, on average on some dimensions, even these differences between women and men are not as great as the differences among men or among women.

The meanings of masculinity and femininity vary over four different dimensions; thus four different disciplines are involved in understanding gender.

First, masculinity and femininity vary across cultures. Anthropologists have documented the ways that gender varies cross-culturally. Some cultures encourage men to be stoic and to prove masculinity, especially by sexual conquest. Other cultures prescribe a more relaxed definition of masculinity, based on civic participation, emotional responsiveness, and collective provision for the community's needs. Some cultures encourage women to be decisive and competitive; others insist that women are naturally passive, helpless, and dependent. What it means to be a man in France or among Aboriginal peoples in the Australian outback are so far apart that it belies any notion that gender identity is determined mostly by biological sex differences. The differences between two cultures' version of masculinity or femininity is often greater than the differences between the two genders.

Second, definitions of masculinity and femininity vary considerably in any one country over time. Historians have explored how these definitions have shifted, in response to changes in levels of industrialization and urbanization, position in the larger world geopolitical and economic context, and with the development of new technologies. What it meant to be a woman in seventeenth-century France or in Hellenic Greece is certainly different from what it might mean to be a French or Greek woman today.

Third, definitions of masculinity and femininity change over the course of a person's life. Developmental psychologists have examined how a set of developmental milestones led to difference in our experience and our expression of gender identity. Both chronological age and life-stage require different enactments of gender. In the West, the issues confronting a man about proving himself and feeling successful will change as he ages, as will the social institutions in which he will attempt to enact those experiences. A young single man defines masculinity differently than will a middle-aged father and an elderly grandfather. Similarly, the meanings of femininity are subject to parallel changes, for example, among prepubescent women, women in child-bearing years, and postmenopausal women, as they are different for women entering the labor market and those retiring from it.

Finally, the meanings of masculinity and femininity vary considerably within any given society at any one time. At any given moment, several meanings of masculinity and femininity coexist. Simply put, not all American, Brazilian, or Senegalese men and women are the same. Sociologists have explored the ways in which class, race, ethnicity, age, sexuality, and region all shape gender identity. Each of these axes modifies the others. Imagine, for example, two 'American' men, one, an older, black, gay man in Chicago, the other, a young, white, heterosexual farm boy in Iowa. Wouldn't they have different definitions of masculinity? Or imagine a 22-year old wealthy Asian–American heterosexual woman in San Francisco and a poor white Irish Catholic lesbian in Boston. Wouldn't their ideas about what it means to be a woman be different? Yet each of these people is deeply affected by the gender norms and power arrangements of their society.

If gender varies so significantly—across cultures, over historical time, among men and women within any one culture, and over the life course—we cannot speak of masculinity or femininity as though they were constant, universal essences, common to all women and to all men. Thus, gender must be seen as an ever-

changing fluid assemblage of meanings and behaviors and we must speak of masculinities and femininities. By pluralizing the terms, we acknowledge that masculinity and femininity mean different things to different groups of people at different times.

3. Gender and Power

Recognizing diversity ought not obscure the ways in which gender definitions are constructed in a field of power. Simply put, all masculinities and femininities are not created equal. In every culture, men and women contend with a definition that is held up as the model against which all are expected to measure themselves. This 'hegemonic' definition of masculinity is 'constructed in relation to various subordinated masculinities as well as in relation to women,' writes sociologist Connell (1987). As Goffman (1963, p. 128) once described it:

> In an important sense there is only one complete unblushing male in America: a young, married, white, urban, northern, heterosexual, Protestant, father, of college education, fully employed, of good complexion, weight, and height, and a recent record in sports … Any male who fails to qualify in any one of these ways is likely to view himself—during moments at least—as unworthy, incomplete, and inferior.

Women contend with an equally exaggerated ideal of femininity, which Connell calls 'emphasized femininity.' Emphasized femininity is organized around compliance with gender inequality, and is 'oriented to accommodating the interests and desires of men.' One sees emphasized femininity in 'the display of sociability rather than technical competence, fragility in mating scenes, compliance with men's desire for titillation and ego-stroking in office relationships, acceptance of marriage and childcare as a response to labor-market discrimination against women' (Connell 1987). Emphasized femininity exaggerates gender difference as a strategy of 'adaptation to men's power' stressing empathy and nurturance; 'real' womanhood is described as 'fascinating' and women are advised that they can wrap men around their fingers by knowing and playing by the 'rules.'

4. Gender Identity as Relational

Definitions of masculinity and femininity are not constructed simply in relation to the hegemonic ideals of that gender, but also in constant reference to each other. Gender is not only plural, it also relational. Surveys in Western countries indicate that men construct their ideas of what it means to be men in constant reference to definitions of femininity. What it means to be a man is to be unlike a woman; indeed, social psychologists have emphasized that while different groups of men may disagree about other traits and their significance in gender definitions, the 'anti-femininity' component of masculinity is perhaps the single dominant and universal characteristic.

Gender difference and gender inequality are both produced through our relationships. Chodorow (1979) argued that the structural arrangements by which women are primarily responsible for raising children creates unconscious, internalized desires in both boys and girls that reproduce male dominance and female mothering. For boys, gender identity requires emotional detachment from mother, a process of individuation through separation. The boy comes to define himself as a boy by rejecting whatever he sees as female, by devaluing the feminine in himself (separation) and in others (male superiority). Girls, by contrast, are bound to a pre-Oedipal experience of connection to the same-sex parent; they develop a sense of themselves through their ability to connect, which leads to a desire to become mothers themselves. This cycle of men defining themselves through their distance from, and devaluation, of femininity can end, Chodorow argues, only when parents participate equally in child rearing.

5. Gender as an Institution

Although recognizing gender diversity, we still may conceive masculinities or femininities as attributes of identity only. We think of gendered individuals who bring with them all the attributes and behavioral characteristics of their gendered identity into gender-neutral institutional arenas. But because gender is plural and relational, it is also situational. What it means to be a man or a woman varies in different institutional contexts. Those different institutional contexts demand and produce different forms of masculinity and femininity. 'Boys may be boys,' cleverly comments feminist legal theorist Rhode, 'but they express that identity differently in fraternity parties than in job interviews with a female manager' (Rhode 1997, p. 142).

Gender is, thus, not only a property of individuals, some 'thing' one has, but a specific set of behaviors that are produced in specific social situations. And thus gender changes as the situation changes.

Institutions are themselves gendered. Institutions create gendered normative standards, express a gendered institutional logic, and are major factors in the reproduction of gender inequality. The gendered identity of individuals shapes those gendered institutions, and the gendered institutions express and reproduce the inequalities that compose gender identity. Institutions themselves express a logic—a dynamic—that reproduces gender relations between women and men and the gender order of hierarchy and power.

Not only do gendered individuals negotiate their identities within gendered institutions, but also those institutions produce the very differences we assume are the properties of individuals. Thus, 'the extent to which women and men do different tasks, play widely disparate concrete social roles, strongly influences the extent to which the two sexes develop and/or are expected to manifest widely disparate personal behaviors and characteristics.' Different structured experiences produce the gender differences which we often attribute to people (Chafetz 1980).

For example, take the workplace. In her now-classic work, *Men and Women of the Corporation* Kanter (1977) argued that the differences in men's and women's behaviors in organizations had far less to do with their characteristics as individuals, than it had to do with the structure of the organization and the different jobs men and women held. Organizational positions 'carry characteristic images of the kinds of people that should occupy them,' she argued, and those who do occupy them, whether women or men, exhibited those necessary behaviors. Though the criteria for evaluation of job performance, promotion, and effectiveness seem to be gender neutral, they are, in fact, deeply gendered. 'While organizations were being defined as sex-neutral machines,' she writes, 'masculine principles were dominating their authority structures.' Once again, masculinity—the norm—was invisible (Kanter 1977, 1975). For example, secretaries seemed to stress personal loyalty to their bosses more than did other workers, which led some observers to attribute this to women's greater level of personalism. But Kanter pointed out that the best way for a secretary—of either sex—to get promoted was for the boss to decide to take the secretary with him to the higher job. Thus, the structure of the women's jobs, not the gender of the jobholder, dictated their responses.

Sociologist Joan Acker has expanded on Kanter's early insights, and specified the interplay of structure and gender. It is through our experiences in the workplace, Acker maintains, that the differences between women and men are reproduced and by which the inequality between women and men is legitimated. Institutions are like factories, and one of the things that they produce is gender difference. The overall effect of this is the reproduction of the gender order as a whole (see Acker 1987, 1988, 1989, 1990).

Institutions accomplish the creation of gender difference and the reproduction of the gender order through several gendered processes. Thus, 'advantage and disadvantage, exploitation and control, action and emotion, meaning and identity, are patterned through and in terms of a distinction between male and female, masculine and feminine.' We would err to assume that gendered individuals enter gender-neutral sites, thus maintaining the invisibility of gender-as-hierarchy, and specifically the invisible masculine organizational logic. On the other hand, we would be just as incorrect to assume that genderless 'people' occupy those gender-neutral sites. The problem is that such genderless people are assumed to be able to devote themselves single-mindedly to their jobs, have no children or family responsibilities, and may even have familial supports for such single-minded workplace devotion. Thus, the genderless jobholder turns out to be gendered as a man.

Take, for example, the field of medicine. Many doctors complete college by age 21 or 22, medical school by age 25–27, and then face three more years of internship and residency, during which time they are occasionally on call for long stretches of time, sometimes, even two or three days straight. They thus complete their residencies by their late 20s or early 30s. Such a program is designed for a male doctor—one who is not pressured by the ticking of a biological clock, for whom the birth of children will not disrupt these time demands, and who may even have someone at home taking care of the children while he sleeps at the hospital. No wonder women in medical school—who number nearly one-half of all medical students today—began to complain that they were not able to balance pregnancy and motherhood with their medical training.

In a typical academic career, a scholar completes a Ph.D. about six to seven years after the BA, roughly by age 30, and then begins a career as an Assistant Professor with six more years to earn tenure and promotion. This is usually the most intense academic work period of a scholar's life, and also the most likely childbearing years for professional women. The tenure clock is thus timed to a man's rhythms—not just any man, but one with a wife to relieve him of family obligations as he establishes his credentials. To academics struggling to make tenure, it often feels that publishing requires that family life perish.

Embedded in organizational structures that are gendered, subject to gendered organizational processes, and evaluated by gendered criteria, then, the differences between women and men appear to be the differences solely between gendered individuals. When gender boundaries seem permeable, other dynamics and processes can reproduce the gender order. When women do not meet these criteria (or, perhaps more accurately, when the criteria do not meet women's specific needs), we see a gender segregated workforce and wage, hiring, and promotional disparities as the 'natural' outcomes of already-present differences between women and men. It is in this way that those differences are generated and the inequalities between women and men are legitimated and reproduced.

6. *'Doing Gender'*

There remains one more element in the sociological explanation of masculinities and femininities. Some psychologists and sociologists believe that early child-

hood gender socialization leads to gender identities that become fixed, permanent, and inherent in our personalities. However, many sociologists disagree with this notion today. As they see it, gender is less a component of identity—fixed, static—that we take with us into our interactions, but rather the product of those interactions. In an important article, West and Zimmerman (1987, p. 140) argued that 'a person's gender is not simply an aspect of what one is, but, more fundamentally, it is something that one *does*, and does recurrently, in interaction with others.' We are constantly 'doing' gender, performing the activities and exhibiting the traits that are prescribed for us.

Doing gender is a lifelong process of performances. As we interact with others we are held accountable to display behavior that is consistent with gender norms, at least for that situation. Thus, consistent gender behavior is less a response to deeply internalized norms or personality characteristics, and more a negotiated response to the consistency with which others demand that we act in a recognizable masculine or feminine way. Gender is less an emenation of identity that bubbles up from below in concrete expression; rather, it is an emergent property of interactions, coerced from us by others.

Understanding how we do masculinities and femininities, then, requires that we make visible the performative elements of identity, and also the audience for those performances. It also opens up unimaginable possibilities for social change; as Kessler points out in her study of 'inter-sexed people' (hermaphrodites, those born with anatomical characteristics of both sexes, or with ambiguous genetalia):

> If authenticity for gender rests not in a discoverable nature but in someone else's proclamation, then the power to proclaim something else is available. If physicians recognized that implicit in their management of gender is the notion that finally, and always, people construct gender as well as the social systems that are grounded in gender-based concepts, the possibilities for real societal transformations would be unlimited (Kessler 1990, p. 25).

Kessler's gender Utopianism raises an important issue. In saying that we 'do' gender we are saying that gender is not only something that is done to us. We create and recreate our own gendered identities within the contexts of our interactions with others and within the institutions we inhabit.

See also: Androgyny; Fatherhood; Feminist Movements; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Postmodern; Gay/Lesbian Movements; Gender and Feminist Studies; Gender and Feminist Studies in Sociology; Gender Differences in Personality and Social Behavior; Gender Identity Disorders; Gender-related Development; Lesbians: Historical Perspectives; Male Dominance; Motherhood: Social and Cultural Aspects; Race and Gender Intersections; Rape and Sexual Coercion; Rationality and Feminist Thought; Right-wing Movements in the United States: Women and Gender; Sexual Harassment: Social and Psychological Issues; Sexuality and Gender; Social Movements and Gender; Transsexuality, Transvestism, and Transgender

Bibliography

Acker J 1987 Sex bias in job evaluation: A comparable worth issue. In: Bose C, Spitze G (eds.) *Ingredients for Women's Employment Policy*. SUNY Press, Albany, NY

Acker J 1988 Class, gender and the relations of distribution. *Signs: Journal of Women in Culture and Society* **13**: 473–97

Acker J 1989 *Doing Comparable Worth: Gender, Class and Pay Equity*. Temple University Press, Philadelphia, PA

Acker J 1990 Hierarchies, jobs, bodies: A theory of gendered organizations. *Gender & Society* **4**(2): 139

Acker J, Van Houten D R 1974 Differential Recruitment and Control: The Sex Structuring of Organizations. *Administrative Science Quarterly* **19**(2): 152

Chafetz J 1980 Toward a macro-level theory of sexual stratification. *Current Perspectives in Social Theory* **1**

Connell R W 1987 *Gender and Power*. Stanford University Press, Stanford, CA

Goffman E 1963 *Stigma: Notes on the Management of Spoiled Identity*. Prentice-Hall, Englewood Cliffs, NJ

Kanter R M 1975 Women and the structure of organizations: Explorations in theory and behavior. In: Millman M, Kanter R M (eds.) *Another Voice: Feminist Perspectives on Social Life and Social Science*. Anchor Books, New York

Kanter R M 1977 *Men and Women of the Corporation*. Basic Books, New York

Kessler S J 1990 The medical construction of gender: Case management of intersexed infants. *Signs* **16**(1): 3–26

Kimmel M 2000 *The Gendered Society*. Oxford University Press, New York

Rhode D 1997 *Speaking of Sex*. Harvard University Press, Cambridge, MA

Risman B 1998 *Gender Vertigo*. Yale University Press, New Haven, CT

West C, Zimmerman D 1987 Doing gender. *Gender & Society* **1**(2)

M. Kimmel

Mass Communication: Empirical Research

1. Overview

Evolution of today's industries of mass communication roughly coincided with the development of empirical research methodology and during the middle of the twentieth century the two were closely identified

with US academics. In the late 1930s, scholars began studying mass media institutions and audiences in response to what Wilbur Schramm, the driving force in the creation of mass communication as an academic field, often described as their 'bigness and fewness.' Schramm published readers composed of publications by social scientists who were skeptical of armchair observation (Schramm 1948). He and others created institutes for research in field and laboratory settings to examine popular and critical claims about what newspapers, motion pictures, and radio were doing. Scholars were mostly attracted to media with large and heterogeneous audiences, particularly those that were controlled by a small number of 'gatekeepers.'

Motivated by societal fears more often than by hopefulness, social and behavioral scholars applied systematic methods to analyze media communications, their origins, their popular uses, and their effects on people and society. As mass media reached practically everyone in modern societies, theories about them became commonplace too. Most of these beliefs were expressed quantitatively and many were subjected to statistical testing—some passing these tests, others not. Empiricism, in the broad sense of matching general knowledge claims with general evidence, also spread to more qualitative topics, such as the processes by which mediated communications are produced and consumed (Shoemaker and Reese 1996). The history of this field is documented in Rogers (1994).

By the 1950s, a burgeoning body of multidisciplinary knowledge on mass communication was being taught at several research universities, often in association with education for journalism but mostly as a graduate subject. Empirical scholarship about communication processes and effects was central to most PhD curricula in mass communication through the 1960s. It has remained a significant doctoral option in most schools, alongside more discursive approaches that became popular during and after the Vietnam war. Topics of empirical study have included academic theories about mass communication that cut across media. Predominantly, though, research has centered on the media as social entities and the societal problems each has occasioned.

2. *Characteristic Forms of Empirical Research*

Lasswell (1948) suggested that an act of communication can be analyzed in terms of the answers to the questions, 'Who/Says What/In Which Channel/To Whom/With What Effect?' These five queries became the standard categories of research: studies of communicators, of content, of media, of audiences, and of effects (Berelson and Janowitz 1950). Lasswell's agenda constrained theorizing to a unidirectional transmission model of the human communication process, leaving out ritual uses of media, for example.

2.1 *Audience and Content Surveys*

The most common modes of empirical research on mass communication are descriptive surveys of the content and the audiences of print and broadcast media. Beyond sheer description, audience surveys usually include measures of both media consumption and personal dispositions thought to represent either its causes or its effects. Audience research is mostly conducted by media themselves and includes, for example, newspaper readership studies and broadcast industry monitoring of audience size. Content analyses are mostly produced by academics, often as a way of critiquing media performance (Rosengren 2000). Field observation, institutional investigations, and analyses of documentary records also contribute to the knowledge base.

Most empirical research in this field is aimed at understanding regularities in mass communication processes and effects. Falsifiable research hypotheses are preferred over propositions that cannot be tested, and standards of evidence are assumed to guard against accepting erroneous propositions (Berger and Chaffee 1987). To generalize about mass communication is a formidable task, as each media industry delivers billions of messages to several billion people every day. This constant stream of events commends itself to systematic sampling methods, which are applied in many content analyses as well as in audience surveys. Most media industries are structured along national lines and researchers generally accept 'national sample' data as broad and generalizable. Studies of the internal workings of media are also intranational. The political role of journalism in the USA differs from, for example, that in European systems.

In audience research, representative sampling and measurement are difficult to achieve because interviewing requires cooperation, and objectivity, of the people under study. Self-report data are used, despite their uncertainties, to investigate what other people's experiences are, in contrast with the critic who presumes an ability to know everyone's reactions to a message. Empirical evidence often shows for example that people with contrasting value perspectives draw quite different conclusions from the same message.

2.2 *Experiments*

Experimental study in mass communication is procedurally analogous to the medical sciences, based on random assignment of cases to treatment and control groups or comparison groups. Experimental field testing has overturned some widely held beliefs about media rhetoric. For instance, the stricture that one should not mention arguments opposing an advocated conclusion turned out to be effective with high school dropouts, but it is an ill-advised method of persuading

a college-educated audience (Hovland et al. 1949). Despite time-honored journalistic lore, a television news story that frames an economic recession as seen through the tragedy of a single individual or family directs blame, rather than sympathy, toward these victims (Iyengar 1991).

2.3 Correlational Tests

On topics for which experimental manipulation is impracticable, survey research is often used to test causal propositions. Correlations between exposure to media content and corresponding social behavior are ambiguous in terms of causal directions but the most successful investigators manage to rule out alternative explanations. For instance, the correlation between viewing violent television shows and peer-rated aggressiveness is considered more likely to result from that causal ordering than the reverse, because very little correlation is found between aggressive behavior and a preference for violent programs.

Use of mass media is embedded in so many habitual behaviors that even when causal ordering can be isolated, each direction is supported and relationships are deemed to be reciprocal. For example, newspaper reading is correlated with many forms of political involvement, such as knowledge, opinion-holding, and interpersonal discussion. But field quasi-experiments show that increased newspaper reading leads to higher involvement and that stimulated discussion or political activity in turn stimulates newspaper reading (Chaffee and Frank 1996). Similarly, at a macroscopic level, national mass media infrastructure is one of many systemic factors that are 'both index and agent of development' (Lerner 1957).

3. Normative Theory

Empirical scholars think of themselves as scientists (Berger and Chaffee 1987) but normative theory plays a central role in the field. Many scholars study mass communication because of fears about undesirable influences within society or in the hope that media institutions can be used for beneficial ends. It is common to conclude an empirical study with recommendations for ways to 'improve' the communication phenomenon involved. Communication is mostly a behavioral science that studies mostly social problems.

The divergence between empirical scholars and purely normative theorists arises in the conduct of a study, not in its origins. An empiricist by definition studies those aspects of a phenomenon on which data can be gathered and hypotheses tested and entertains the possibility that evidence will fail to support a theory. For example, the newspaper industry trumpets the 'marketplace of ideas' as the basis for a free and competitive press system. But content analyses in two-newspaper communities find few differences between the news covered in the competing papers. Empirical failure of normative tenets in empirical tests forces a choice between one's beliefs and one's data that divides scholars into separate camps.

4. Media-driven Research

The empirical research community has seized upon each new mass medium as an occasion for investigation. In the face of innovations toward the end of the twentieth century, the scholarly field is questioning the viability of the very term 'mass communication' because no given audience is quite so huge and heterogeneous as was once the case, and control is becoming decentralized. When a new media industry emerges, the earliest studies describe it and speculate on its potential. Soon after come quantitative assessments of its characteristic content and the demography and motivational bases of its use. Eventually investigations turn to questions of effects and of social processes above the level of the individual.

4.1 Newspapers

The newspaper predated social science and in several ways helped spawn it; as a fact-gathering institution, the press is to an extent itself involved in social research and journalism education became the first permanent academic home for mass communication research. The idea that journalism belonged to the social sciences was instituted in the late 1920s at the University of Wisconsin under Willard G. Bleyer when a journalism minor was instituted for PhD students in political science. Bleyer students such as Chilton Bush at Stanford University and Ralph Nafziger at Minnesota and Wisconsin helped to institutionalize mass communication as an academic field of study.

The first newspaper readership studies in *Journalism Quarterly* were published in 1930 by George Gallup and Nafziger (Chaffee 2000). From these beginnings, audience survey research became a routine newspaper marketing tool and public opinion polls a new form of news reporting. Content analysis was promoted by Nafziger and Bush; they and their students developed sampling methods such as the constructed-week description of a newspaper's coverage and built an empirical summary of patterns of public affairs reporting in the American newspaper. Content analysis was also used in developing hypotheses about control of mass media and about media uses and effects (Berelson 1952). Journalists themselves have been subjected to empirical scrutiny, including large-sample surveys (Weaver and Wilhoit 1991) and theories to explain their actions (Shoemaker and Reese 1996)

4.2 Motion Pictures

The first wide-ranging program of research on a specific medium was the Payne Fund project on

motion pictures and youth (Charters 1933). 'The movies' were suspected of broadly undermining education and moral values but the research showed that effects varied widely depending on the individual audience member. This pattern of contradicting conventional fears would be repeated a generation later with television (Schramm et al. 1961). Powdermaker's (1950) analysis of the power structure of Hollywood was perhaps the most insightful anthropological investigation of any mass media institution. But social scientists did not, by and large, find motion pictures amenable to the methods and goals of empirical study after television captured much of the movie-going audience in the 1950s. Films are too disparate, and 'mass' movie attendance too occasional, to meet the assumptions of the empirical research field.

4.3 Radio

During its heyday from the 1920s into the 1950s, radio fit the image of a mass medium more fully than did film, reigning as the channel Americans used most in that period. In many countries of the world today, radio remains the most ubiquitous medium and it is frequently studied in connection with rural development projects (Schramm 1977). But research on radio in the US shrank to a trickle after its social and cultural role was overrun by television in the 1950s. Earlier, though, radio networks headquartered in New York provided the context for the first organized mass communication research institute, Paul Lazarsfeld's Office of Radio Research at Columbia University. Despite its nominal dedication to radio, Lazarsfeld extended his research program to newspapers and interpersonal communication as factors in, for example, presidential election campaigns (Lazarsfeld et al. 1944, Berelson et al. 1954). The research team also probed public reactions to entertainment, such as a dramatization of an invasion from outer space (Cantril 1940). Long after radio lost its main audience to television, Lazarsfeld's radio research group left an important theoretical imprint in the 'limited effects' or 'minimal consequences' model of mass communication (Klapper 1960).

4.4 Magazines

The magazine industry has only occasionally aroused empirical research within the mass communication field. A notable exception within the radio research group was Lowenthal's (1943) detailing of an apparent shift in American cultural values, based on the subjects of biographies in large-circulation magazines; in the 1920s, most such articles concerned 'heroes of production' including industrialists such as Thomas Edison and Henry Ford but, by the late 1930s, magazines mostly featured 'heroes of consumption' such as film stars and radio personalities. This was the forerunner of sociological studies that attempt to draw inferences about a society's values from reading its media content. Competitive business pressures ended the era of mass-circulation consumer magazines by about 1960 and interest among empirical scholars of mass communication waned correspondingly.

4.5 Books

Although the era of mass communication conventionally dates back to the Gutenberg press, books have only occasionally been considered part of the mass media for purposes of media audience, content, or effects analysis. The book has been considered in relation to other media, as in Parker's (1963) finding that the introduction of television into a community reduced per capita public library circulation by approximately one book per year.

4.6 Theory of Mass Communication

Discourse on 'mass media' as a unified entity was well established by World War II (Rogers 1994). Lasswell (1948) provided a very general scheme of the functions a communication system serves in any social system. These included surveillance of the environment, correlation of the various parts of society into a working whole, and transmission of the cultural heritage to new generations of members. These concepts were paralleled in Schramm's (1964) characterization of traditional society in the roles of the town crier, the village council, and the teacher. In empirical research, surveillance is typified by studies of news diffusion, correlation by agenda-setting, and transmission by political socialization. These and other general ideas about communication functions were presumed to apply to all mass media, including those yet to come.

4.7 Television

Mass communication was newly established as a research field when television arrived in most American homes in the 1950s. As television supplanted earlier media in daily life, it also replaced them as the main context for empirical research. Two critical aspects of television became staple research topics—its impact in politics and on children. Although many Americans cite television as their primary source of news, Patterson and McClure (1976) attracted attention by showing that, in an election campaign, network television gave little attention to political issues. Still, empirical scholars have focused on television networks' political coverage and news selection processes (Gans 1979). The first major studies on integration of television into daily life centered on children (Himmelweit et al. 1958, Schramm et al. 1961). Inquiries into television violence led to extensive early studies of effects as well as content (Comstock et al. 1978). Anti-social effects of media violence are

today treated as a truism in the research community. Field testing of this proposition became rare after American television proliferated into many channels in the 1980s but new studies are occasionally conducted when television is introduced into additional countries around the world. Many nations restrict violent television programming based on empirical findings.

4.8 New Media

The era of outer space, the silicon chip, and the computer brought many additional means of communication that are lumped simply as 'new media.' New communication technologies include such rival means as satellite transmission and cable television, tape recordings and compact discs, and computer games and video games. As alternatives to one another, they give consumers more choices, thus eroding the centralized control that typifies mass communication. Many empiricists transferred from traditional mass media to evaluate the social impact of these innovations, examining each new medium in such familiar terms as its audience, its uses, its potential applications, and its social effects for good or ill. After a first wave of descriptive research on a new technology, scholars looked for ways in which it might create 'new communication' and thus social change.

While technological innovation greatly expanded the television industry, it had a limiting impact on the newspaper industry. What empirical scholars consider 'television' grew in the US from three networks in the 1960s to dozens of cable channels in the 1990s (Wilson et al. 1998). With so many rival media claiming audience attention and advertising support, newspapers shrank steadily in number and size throughout the second half of the twentieth century and many were sold into chain ownership. In the more compact nations of Western Europe, newspaper circulation was essentially stable after World War II, although the number of different newspapers declined somewhat. Newspaper reading grew less common in each succeeding US birth cohort throughout the late twentieth century, and aggregate readership dwindled as aging readers died. Concentration of leisure time in electronic media coincided with a reduction in political activity after the turbulent 1960s, leading scholars to propose media-based theories of declining democratic participation (Putnam 2000).

New media are not perceived as having audiences, content, and effects, so much as having users and applications. User control of new media revived the 'active audience' tradition and a flurry of research on 'mass media uses and gratifications' in the 1970s evolved into a similar line of study of the rewards of the new non-mass media in the 1980s and 1990s. New media were seen as potentially very profitable and, as capital was invested in designing more versatile media systems, the techniques of empirical science came into demand. Many of these novel uses of media were seen as pro-social in their goals, attracting empirically trained investigators who might have eschewed putting their skills to work for commercial mass communication firms.

4.9 The Internet

As a worldwide communication system free from centralized control of message production and distribution, the Internet is paradoxically a massive technical infrastructure that promises to 'demassify' human communication. This new medium could obviate studies of such traditional media institutions as local newspapers and national broadcasting. Empiricists have analyzed such new channels as e-mail, chat rooms, and web sites in terms designed for newspapers, radio and television, but the research agenda is itself being modified as users construct their own specialized networks on the Internet.

5. National Priorities for Research

Within policy circles, systematic quantitative behavioral evidence has generally been valued above impressionistic knowledge claims. However, while studies of mass media organizations, content, audiences, and effects often inform policy debates, they have rarely resolved them. Nonetheless, a great deal of the research in this field has been oriented toward societal priorities, such as the role of the media in democratic political processes, and in the socialization of young people. Content analyses are often concerned with social inequities, such as mismatches between the kinds of people and social behaviors portrayed in the media and those that occur in everyday life. The media have been sought out in times of national emergency, and for the promotion of social goals. To a great extent, the academic field, unlike the media organizations themselves, has devoted itself to evaluation of mass communication on behalf of American society. The same is true in other countries. Mass communication scholars in most nations couch their indigenous research as an expression of their culture's unique perspectives. In developing countries, national mass media are evaluated as instruments of domestic development, while foreign inputs (especially US and other western media) are criticized for 'imperialistic' influences that undermine the indigenous culture.

6. Academic Conceptualizations

During the second half of the twentieth century, the social science of mass communication has existed in the form of departments, journals, and academic associations. This field began in the US, spurred

especially in connection with journalism education due to the efforts of Schramm and Bleyer's proteges, and with speech when the National Society for the Study of Communication split off from the Speech Association of America in the 1950s. In some institutions, associations, and journals, communication is studied generically, divided by levels of analysis such as interpersonal, network, community, and macrosocietal (Berger and Chaffee 1987, Rosengren 2000).

This field has been housed primarily at large state universities, especially in the Midwest. It has spread rather unevenly to other countries of the world, being much stronger, for example, in Germany than in France. In Great Britain, and in former colonies that retain the British educational structure, mass communication has been housed in the polytechnic institutions rather than the research universities, so empirical research has not developed strongly; the same could be said of the New England states, where the Ivy League schools were slow to build communication programs, despite early leadership at Yale (e.g., Lasswell, Hovland) and Columbia University (e.g., Lazarsfeld and Berelson).

6.1 Propaganda

Lasswell's (1927) analysis of wartime propaganda techniques marks the beginning of systematic empirical research on mass communication as a general process. For more than a decade, the media were considered important primarily because of their potential for mass persuasion. This alarmist viewpoint foundered both because the first studies of effects on audiences found far less impact than had been feared (Klapper 1960) and because, in World War II, the Allied war effort also included propaganda. The term 'mass communication' was substituted as a less tendentious synonym and soon took on its scholarly meaning. Propaganda was revived during the Cold War as a pejorative applied to the Soviet Union, but was no longer the way mainstream empiricists identified their research.

6.2 Measurement of Meaning

An early research program at the University of Illinois was the attempt to develop an operational definition of the central concept of meaning (Osgood et al. 1957). The semantic differential response scale was developed to track the dimensions along which people orient themselves to external entities, such as people in the news. Three dimensions of meaning were usually found, across a variety of rated objects: evaluation ('good–bad'), activity ('active–passive'), and potency ('strong–weak'). The three-dimensional concept of 'semantic space' was explored for its cross-cultural validity, theoretical implications in cognition, and applications to content analysis. The evaluative rating method persists in the field as a versatile and efficient method for measuring opinion change as a communication effect.

6.3 Agenda Setting

During the 1968 US election campaign, McCombs and Shaw (1972) compared the issues considered important by voters in a North Carolina city and local newspaper coverage of those issues. Finding a strong rank-order correlation between the two and inferring that voter priorities resulted from media priorities, they christened this the 'agenda setting function' of mass communication. This phenomenon, widely studied since, was proposed as a partial answer to the 'limited effects' model of media persuasion; as one writer expressed it, 'the media may not tell people what to think, but they are stunningly successful at telling people whraat to think about' (Cohen 1963).

6.4 Uses and Gratifications

In reaction against the popular emphasis on media effects, empirical scholars began in the 1970s to develop a line of research that is generally called 'media uses and gratifications' (Blumler and Katz 1974). It in effect turned Lasswell's (1948) questions around, asking what people seek when they use mass media. An explosion of research on this kind of question followed a thorough investigation in Israel of what people considered important to them and how well each of five media (newspapers, radio, cinema, books, and television) served those functions (Katz et al. 1973). The main method applied to these issues is audience study, with detailed interviews, usually with a fixed schedule of questions and elaborate statistical analyses exploring, for example, the match between gratifications sought and those received. This approach remains popular as new media have arrived. Weaknesses include reliance on people's self-reported motivations, and circularity of functional explanations.

6.5 Knowledge Gap

A group of rural sociologists studying communication programs in Minnesota in the late 1960s noted a phenomenon they described as a widening 'knowledge gap' across the aggregate population (Tichenor et al. 1970). An information campaign might produce a desired information gain for 'the average person' and yet not necessarily succeed at the social level, if its impact is limited to the upper strata of the population who had already been fairly knowledgeable in the first place. By widening the difference between the 'haves

and have-nots,' the overall pattern could make a bad situation worse, further disadvantaging the lower information strata. This reformulation of information campaign processes argues in effect for a second criterion—reduction of the variance in the system as a whole—alongside that of increasing the average or total store of community knowledge. Many studies found that information campaigns in fact could 'close the knowledge gap,' particularly if their designers worked with this outcome in mind.

6.6 Spiral of Silence

A German theoretical formulation that excited a great deal of empirical research in other western countries and Japan in the 1980s and 1990s was Noelle-Neumann's (1984) scenario of a 'spiral of silence' in public opinion processes. The model assumes that people fear social isolation if they express opinions that are losing favor in the general public and hence tend not to not talk with people of an opposite persuasion; the result is a magnification of people's perception of a shift in popular opinion. This prediction tended to be found in Germany and Japan, but did not hold up well in US replications (Glynn et al. 1997).

6.7 Selective Exposure

A key principle of the limited effects model was that people selectively expose themselves to media messages that are congenial to their social attitudes and political beliefs. This assumption led to the expectation that mass communication's persuasive impact would operate on a heterogeneous audience in two opposing directions, toward the extremes of opinion; the net impact would be negligible ('mere reinforcement'). However, experimental tests showed that media audiences were less self-deceiving than this theory suggested (Sears and Freedman 1967) and the concept gradually disappeared from textbooks.

But it is clear that people seek media messages nonrandomly. The assumption of purposeful exposure to particular experiences via the media, implicit in the uses and gratifications tradition, was revivified in the 1980s by Zillmann and Bryant (1985). Their theory was not one of defensive avoidance but of positive seeking of messages that help an individual manage emotional moods and affective needs.

7. The Need for Empirical Research on Mass Communication

Because huge fortunes and business empires are integral to mass communication, academic claims about the harm these institutions might do, or ways they can serve society better, face skeptics within both industry and policy circles. 'How do you know that?' is a common query faced by critics of media violence, election campaign reformers, and protectors of young consumers alike. Media institutions collectively occupy an ever larger position in society, as do our ways of studying them.

See also: Advertising: General; Agenda-setting; Audience Measurement; Communication and Democracy; Communication: Electronic Networks and Publications; Content Analysis; Film and Video Industry; Internet: Psychological Perspectives; Journalism; Market Research; Mass Communication: Normative Frameworks; Mass Communication: Technology; Mass Media and Cultural Identity; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Media Effects; Media, Uses of; Political Communication; Printing as a Medium; Radio as Medium; Soap Opera/Telenovela; Television: Industry; Violence and Media

Bibliography

Berelson B 1952 *Content Analysis in Communication Research*. Free Press, Glencoe, IL

Berelson B, Janowitz M 1950 *Reader in Public Opinion and Communication*. Free Press, Glencoe, IL

Berelson B B, Lazarsfeld P F, McPhee W 1954 *Voting*. University of Chicago Press, Chicago

Berger C R, Chaffee S H 1987 *Handbook of Communication Science*. Sage, Beverly Hills, CA

Blumler J, Katz E 1974 *The Uses of Mass Communication*. Sage, Beverly Hills, CA

Cantril H 1940 *The Invasion from Mars*. Princeton University Press, Princeton, NJ

Chaffee S 2000 Scholarly milestones. In: Gallup G, Nafziger R (eds.) *Mass Communication and Society*. Vol. 3, pp. 317–327

Chaffee S, Frank S 1996 How Americans get political information. *The Annals of the American Academy of Political and Social Science* **546**: 48–58

Charters W W 1933 *Motion Pictures and Youth*. Macmillan, New York

Cohen B C 1963 *The Press and Foreign Policy*. Princeton University Press, Princeton NJ

Comstock G, Chaffee S, Katzman N, McCombs M, Roberts D 1978 *Television and Human Behavior*. Columbia University Press, New York

Gans H J 1979 *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek and Time*. Pantheon, New York

Glynn C J, Hayes A F, Shanahan J 1997 Perceived support for one's opinions and willingness to speak out—A meta-analysis of survey studies on the 'spiral of silence'. *Public Opinion Quarterly* **61**: 452–63

Himmelweit H, Oppenheim N, Vince P 1958 *Television and the Child*. Oxford University Press, London

Hovland C I, Lumsdaine A A, Sheffield F D 1949 *Experiments on Mass Communication*. Princeton University Press, Princeton, NJ

Iyengar S 1991 *Is Anyone Responsible? How Television Frames Political Issues*. University of Chicago Press, Chicago
Katz E, Gurevitch M, Haas H 1973 On the use of mass media for important things. *American Sociological Review* **38**: 164–81
Klapper J T 1960 *The Effects of Mass Communication*. Free Press, New York
Lasswell H D 1927 *Propaganda Technique in the World War*. Knopf, New York
Lasswell H D 1948 The structure and function of communication in society. In: Bryson L (ed.) *The Communication of Ideas*. Institute for Religious and Social Studies, New York
Lazarsfeld P F, Berelson B, Gaudet H 1944 *The People's Choice*. Duell, Sloan and Pearce, New York
Lerner D 1957 Communication systems and social systems. *Behavioral Science* **2**: 266–275
Lowenthal L 1943 Biographies in popular magazines. In: Lazarsfeld P, Stanton F (eds.) *Radio Research, 1942–43*. Essential Books, Fairlawn, NJ
McCombs M E, Shaw D L 1972 The agenda-setting function of mass media. *Public Opinion Quarterly* **36**: 176–85
Noelle-Neumann E 1984 *The Spiral of Silence: Public Opinion—Our Social Skin*. University of Chicago Press, Chicago
Osgood C, Suci G, Tannenbaum P H 1957 *The Measurement of Meaning*. University of Illinois Press, Urbana, IL
Patterson T E, McClure R D 1976 *The Unseeing Eye: The Myth of Television Power in National Politics*. Putnam, New York
Parker E B 1963 The effects of television on public library circulation. *Public Opinion Quarterly* **27**: 578–89
Powdermaker H 1950 *Hollywood: The Dream Factory*. Little, Brown, Boston
Putnam R D 2000 *Bowling Alone*. Simon & Schuster, New York
Rogers E M 1994 *A History of Communication Study: A Biographical Approach*. Free Press, New York
Rosengren K E 2000 *Communication: An Introduction*. Sage, London
Schramm W 1948 *Communications in Modern Society*. University of Illinois Press, Urbana, IL
Schramm W 1964 *Mass Media and National Development*. Stanford University Press, Stanford, CA
Schramm W 1977 *Big Media, Little Media*. Sage, Beverly Hills, CA
Schramm W, Lyle J, Parker E 1961 *Television in the Lives of Our Children*. Stanford University Press, Stanford, CA
Sears D O, Freedman J L 1967 Selective exposure to information: A critical review. *Public Opinion Quarterly* **31**: 194–213
Shoemaker P J, Reese S D 1996 *Mediating the Message: Theories of Influences on Mass Media Content*, 2nd edn. Longman, White Plains, NY
Tichenor P J, Donohue G A, Olien C N 1970 Mass media flow and differential growth in knowledge. *Public Opinion Quarterly* **34**: 159–70
Weaver D H, Wilhoit G C 1991 *The American Journalist, A Portrait of US News People and their Work*, 2nd edn. Indiana University Press, Bloomington, IN
Wilson B J, Kunkel D, Linz D, Potter W J, Donnestein E, Smith S L, Blumenthal E, Berry M 1998 Violence in television programming overall: University of California, Santa Barbara. In: *National Television Study*. Sage Publications, Newbury Park, CA, Vol. 2, pp. 3–204
Zillmann D, Bryant J 1985 *Selective Exposure to Communication*. Lawrence Erlbaum, Hillsdale, NJ

S. Chaffee

Mass Communication: Normative Frameworks

Mass communication, the public dissemination of symbols that are in principle addressed to everyone, has always aroused controversy, not only about specific policies and practices, but about larger philosophical questions concerning morality, politics and art. Frameworks treating the normative dimensions of mass communication fall into two general groupings: those that distrust it and those that do not. The article examines each in turn.

1. *Distrust of Mass Communication*

Despite its twentieth century ring, mass communication is as old and fiercely debated as civilization itself. Mass communication will always occupy a symbolically charged spot in any social order, whether its medium of communication be pyramids or stained glass, newspapers or television; whether its aim be to keep subjects in awe, believers in the fold, or citizens informed. The lack of discrimination as to audiences and the easy transgression of social boundaries are two key objections to mass communication. Over the twentieth century, such media of mass communication as dime novels, comics and radio, television, videogames and the internet have been blamed for about every ill imaginable in modern life, from the degradation of taste to the corruption of youth and the decline of democracy. Though such attacks are motivated by genuine worries attendant to modern conditions, they have intellectual roots that reach into the Judeo–Christian and Greco–Roman past.

Older normative frameworks are very much alive in academic and popular debates alike about the social meaning of mass communication. They often act today more as moral intuitions about what a social order should expect from its public forms of communication than as systematically developed doctrines. Since such intuitions, at their worst, scapegoat mass media and thus invite simplistic causal attributions for complicated social problems, much social research on the effects of mass communication has explicitly contested older moralistic or normative ideas about mass communication. It would be a mistake, however, to assume that such ideas are of little interest to social science. First, these larger frameworks provide the moral and political resources that drive the enterprise of social science as a quest for truth, humane social organization, and political progress. Second, many age-old topics are very much on the academic and public agenda today, such as worries about the role of the image in politics, the socialization and education of children, the possibility of global understanding, or the place and purpose of the arts. These topics emerge

Table 1
Normative frameworks for mass communication

Doctrine	A characteristic good	A characteristic evil	Representative theorist
Iconoclasm	Representational purity	Idolatry	Moses
Platonism	Reality	Appearance	Plato
Christianity	Compassion	Callousness	Augustine
Stoicism	Cosmopolitanism	Anxiety	Marcus Aurelius
Elitism	Quality	Vulgarity	Arnold, Adorno
Republicanism	Public virtue	Corruption	Cicero, Machiavelli
Liberalism	Free inquiry	Censorship	Locke, Jefferson
Right-wing critics, nineteenth century	Excellence Tradition	Leveling Herd-mentality	Carlyle
Left-wing critics (nineteenth century)	Human fulfillment	Domination	Marx
Liberal critics (nineteenth century)	Expressive variety	Tyranny of majority	Mill, Tocqueville
New deal	Public interest	Irresponsibility	Hutchins commission
Social democracy	Public service	Commercialism	Habermas
Neo-liberalism	Free markets	Regulation	Hayek, Coase
Postmodernism	Free play	Hierarchy	Lyotard, Baudrillard
Cultural Marxism	Empowerment	Dominant ideology	Gramsci, Stuart Hall
Communitarianism	Participation	Apathy Disengagement	Dewey

from distrustful intuitions about mass communication, as is shown below (see Table 1).

1.1 Iconoclasm

The implications of a visually dominated society for politics, cognition and education is a recent concern with a long past. In a deep sense, distrust of the bewitching power of images begins atop Mount Sinai. The second of the ten Mosaic commandments proscribes the making of graven images and of mimetic visual arts in general. Ever since, iconoclasm (the smashing of images) has proven to be an enduring theme in the history of thinking about the public life of signs (Goux 1978). Moses was specifically concerned with the proper representation of deity, but the intuition that images mislead hearts away from the 'true' vision or way of life has long had a more-general purchase. In cultures influenced by Judaism, Christianity and Islam, notions of idolatry and simulation are still used to describe communications and their dangers. Around the world today, TV sets are smashed, authors sentenced to death, or CDs bulldozed in acts of puritanical destruction of media forms considered dangerous. Obviously, such acts are a kind of censorship or repression, as they rest upon the viewpoint that certain things should not be represented. That restraint in depiction can be justifiable, however, is a useful lesson for an age sated with information about Princess Diana's crushed Mercedes or President Clinton's cigar. The principled refusal to look can stem not only from cowardice, but from tact and civility, and offers a rich vein for considering current ethical dilemmas in media representations (Kieran 1998).

1.2 Fabrications and the Corruption of Youth

Fear of the seduction of youth dates at least to Plato's Republic. Plato famously banished poets from his utopia, a gesture archetypal for normative thinking on communication since in at least two respects. First, poets are imitators who multiply appearances rather than realities in a world that already suffers from an overpopulation of copies. Just so, distrust of media veracity is widespread today, whether in the routine suspicion that much in the news is fabricated or the far more sinister inkling that the Holocaust was a carefully orchestrated media conspiracy. Second, poets, claims Plato, have a bad influence on the young, even if their falsehoods can be pedagogically useful. Plato's thought about socialization continues to resonate in debates about mass media, from comics and cinema to popular music and computers. The notion that a just society must patrol public representations appears in the long history of regulating depictions of sexuality and violence in film, and in various content ratings systems for film, television and popular music. Though some twentieth century interpreters of Plato, most notably Sir Karl Popper, read his *Republic* as the fountainhead of authoritarian thinking, few would deny that mass communication aimed at children deserves special care or that the portrayal of 'reality' is a thorny task of public importance.

1.3 Responsibility to Distant Suffering

Both Stoicism and Christianity teach the idea of moral responsibility to the entire world. The Stoics speak of cosmopolitanism, or literally, world citizenship, and the Christians speak of a love that can encompass the entire human family. In the Christian notion of the Gospel, or literally, the good news, there is the ambition, in principle, of dissemination to a worldwide body of communicants. The hope for a great community on a global scale remains a potent political and moral project today, however much it may be stripped of its religious or philosophical content. Here the features of mass communication that usually evoke suspicion are valued. Its lack of discrimination in recipients and ability to cross social barriers make it potentially an inclusive and just form of communication (Peters 1999).

Mass media are prime instigators of responses to distant suffering (Boltanski 1993, Ignatieff 1997). Some suggest that the foreign policy of the USA and NATO is partly dictated by the location of television cameras. Sufferers whose fates can be made graphically visible to influential audiences may, in this argument, be more likely to receive political, humanitarian, or military aid than others who languish beyond the reach of the image. Others suggest that representations of human suffering do not deepen but rather deaden moral sensibilities, as news organizations, fighting for audience share, push the limits on shock and outrage by packaging suffering for popular titillation in ever more gruesome ways. In any case, the ability of the media to alert the public to the rest of the world, to arouse humanitarian interest for those who suffer and to build cross-national bonds of solidarity—or hostility—is clearly one of the most important sites of debate in the politics and ethics of the media today (Moeller 1999). The Christian and Stoic heritage suggests that the representation of one's fellow humans is an ethical responsibility of the first order. Mass communication cannot act in a vacuum: nothing less in this view than global brotherhood and sisterhood is at stake.

1.4 The Degradation of Art

Finally, a tradition with a complex genealogy extending to early modern thinkers such as Montaigne and Pascal or the Roman satirists criticizes mass communication for its cheapness and low level of taste. Often frankly elitist in its condemnation of vulgarity, commercial production or formulaic structure, the high-art tradition wants communication to be an intense and concentrated aesthetic experience. It is perhaps false to say that elitism offers a normative framework for mass communication at all, since mass communication itself—the fact of widely diffused and openly addressed forms of culture—often seems the prime target. Matthew Arnold's dictum that genuine culture consists of 'the best that is known and thought in the world' or the Frankfurt School's fulminations against the 'culture industry' are two instances of the viewpoint that the arts can be morally and socially redemptive when appreciated in the proper, often rarefied, conditions, but catastrophic when they circulate in a bored and restless mass society (Horkheimer and Adorno 1947).

An affirmative case for art in mass communication is found in the republican tradition. For Cicero, who stands at the tradition's head, rhetoric is not simply an art for inflating weak arguments and wooing crowds; it is an authentic technique of political deliberation. Republicans typically argue that theater can be a healthy part of political life, not just a tool of deception or manipulation. Granted, a republican such as Machiavelli has a bad reputation on this point, but he also understands communicative artistry as the very stuff of public life. Whereas many democratic theorists distrust the theater for breeding inauthenticity in citizens and silencing the people, republican theorists insist that rhetoric and drama can have a legitimate place in a just polity (Hariman 1995).

Though concern for aesthetic taste and mass culture was the very heart of twentieth century debates about media, culture and society from the 1920s through the 1980s, the dawn of a new millennium finds confidence thin that a canon of works exists whose quality is beyond dispute. High-brow taste has become increasingly omnivorous and elite art hybrid, jumbling heretofore segregated aesthetic levels. Even so, the intuitions that the products of commercial culture are often appalling and that mass communication could theoretically be a thing of beauty are not likely to disappear from the agenda (Gans 1999).

In sum, several of the most important intellectual sources in the western world—Hebrew, Greek, Roman and Christian—have contributed a number of views about mass communication, all of which converge on the notion that the public dissemination of symbols is of such religious, social, moral and aesthetic weight that it can never be left without some kind of control or restriction. Whether seen as dangerous or valuable, mass communication is held by all these traditions to be powerful indeed.

2. Faith in Openness: Liberalism and its Critics

As opposed to the ancient nervousness about dangerous messages, most modern normative thinking about mass communication developed from or in critical dialog with the Anglo–American tradition, central to which is the proposition that public culture serves society best when left to its own devices. Liberalism, as the tradition has become known owing to its ideal of liberty, emerged from the writings of English Protestants in the seventeenth and eighteenth centuries,

and has spread widely to become the official doctrine not only of communication policy, but of many principles of government generally throughout the modern world. Mass communication, it argues, may threaten extant powers or opinions, but is never truly dangerous.

Distinctive in liberal thought is the special place given to freedom of communication, considered not simply one of many valuable liberties, but as fundamental to all others. The marketplace, whether of goods or ideas, is considered a refuge against the power of the state or church. Liberals conceive of liberty as the absence of constraint. They abhor censorship. Public toleration of ideas held to be repellent or dangerous is, for liberals, the price paid for liberty and an expression of the faith that truth, in the long run, will triumph. Open discussion, like the open market, is held to be self-correcting. The press, said Thomas Jefferson, 'is the best instrument for enlightening the mind of man' and liberal thinkers often give newspapers a privileged mission in civil society, sometimes to the neglect of other institutions. The press, as is often pointed out, is the only nongovernmental institution mentioned in the Constitution of the USA. Liberals dream of the press as a 'fourth estate,' an unofficial branch of government, an all-seeing monitor of society and politics generally, and the public's lifeline of political truth.

The liberal framework, in sum, values openness, criticism, diversity and liberty in mass communication; enjoins an attitude of vigilance, tolerance, public curiosity and rationality on the part of citizens; and believes in a knowable world and in the ultimate fruitfulness of open discussion for societal progress and discovering truth. It owes much to the Enlightenment confidence that scientific progress will banish all shadows and society can be rationally organized. Because of their strong faith in reason and progress, liberals often dismiss, mistakenly, the older frameworks discussed above as censorious and no longer relevant.

2.1 Nineteenth Century Criticisms and Alternatives

The history of modern philosophical norms of mass communication is largely that of the development and criticism of liberalism. In the course of the nineteenth century, liberalism comes under criticism from the right, the left and from within. From the political right comes the charge that liberalism's refusal to identify any positive moral good besides liberty is destructive of tradition and public spirit. Its love of freedom splinters civil society either into private egos pursuing their own interests or herds of bovine citizens haplessly searching for leadership. Such critiques urge that the communication structures shaping public life should not be left to shift for themselves but must be actively cultivated so that the 'best men' and ideas can prevail.

From within, liberal critics of liberalism such as Tocqueville and John Stuart Mill note the potential of its policies to breed conformity and 'soft' forms of despotism. Tocqueville calls for deepened forms of community involvement in voluntary associations, and Mill for a more robust openness in social thought and discussion. Both see the dangers of a tyranny of the majority and the unintended consequences of social atomization. Thenceforth liberals start to consider subtler impediments to liberty besides state and church.

From the left comes an attack on the liberal conception of a self-regulating marketplace, both of goods and ideas. For Marx, most notably, to speak of liberty in the abstract is misguided because it neglects the structures that limit the real experience of free expression. Thinkers in Marxist and socialist traditions have been trenchant critics of the claim that unrestrained discussion will automatically yield truth rather than ideology. Quite like right-wing critics of liberalism, they argue that freedom must mean more than the absence of constraint. It must include meaningful chances for the people to act and speak in public, a point made by feminist and civil rights reformers as well. Freedom of the press is wonderful, the old quip goes, if you happen to own one. For leftist critics, the liberal equation of trade and discussion helps reinforce massive inequalities in the power to propagate ideas. To achieve the ideal of authentically free communication, citizens must have access to the means of communication. More than freedom from censorship, this ideal hinges on people playing meaningful roles in the making of history.

2.2 Twentieth Century Developments

In the twentieth century, fascism emerged from the right-leaning critique of liberalism. European social democracy and the American New Deal developed as left-leaning correctives, with the state in each case assuming an affirmative role in subsidizing social services not delivered by the market, including new mass communication services. In the USA, where the role of the state was weaker than in Europe, broadcasting became a commercial enterprise with only light federal regulation. Even so, broadcasting in theory served the 'public interest, convenience, or necessity'. Though rarely enforced, the public ownership of the airwaves, including the right of listeners to reply and to have programming that met their needs, was found constitutional in several important Supreme Court cases in the USA (Lichtenberg 1990).

In Europe, broadcasting developed as an organ of the state, with a different history in each nation (see *Public Broadcasting*). Whereas media in the USA operated in a capitalist system with occasional governmental oversight, in Europe and elsewhere, government or quasi-government institutions were

developed to provide information and entertainment to the entire nation. The British Broadcasting Corporation, long seen as the international flagship of public-service broadcasting, rested on such principles as universal service, quality programming, a bulwark against creeping commercialism (often read as Americanism) and national integration. Public-service broadcasting was one part of the welfare state's program of providing citizens with the common necessities of life (Calabrese and Burgelman 1999).

Important theoretical treatises on mass communication in the twentieth century reflect a belief that mass communication should serve the public and not simply be free from constraint. In the USA, an important proposal for reform was *A Free and Responsible Press* (1947), better known as the Hutchins Commission. Its title already tells the story: freedom is not enough; the press must actively contribute to the commonweal and stand responsible for its effects. Highly critical of sensationalistic and shallow reporting, which it traced to the economic conditions that require media institutions to grab rather than educate audiences, the report did not, however, call for the state to intervene in the business of public information. Its answer was professional self-regulation within capitalism. Many journalists received the report with hostility, though the project of private ownership combined with public responsibility prevailed in the best postwar American journalism, both print and broadcast (e.g. NBC (National Broadcasting Corporation) and CBS (Columbia Broadcasting Service) news in their heyday). Here the normative vision combines the classic liberal doctrine of freedom from state interference with an affirmative concern for civic duty.

Perhaps the most important theoretical treatise on the normative grounds of mass communication in the second half of the twentieth century was the German philosopher Jürgen Habermas's *Structural Transformation of the Public Sphere* (1962). It articulates the European sense that mass communication is a social good too precious to be left to the market or the state. Notable about this book is its historical and social contextualization of the project of rational public communication. Habermas's notion of the public sphere is a rethinking of the liberal ideal of publicity (now often called 'transparency') for radical democratic purposes. Whether interpersonal or mass, any act of communication discloses, Habermas believes, the potential of persuasion without violence. Habermas' thinking about mass communication turns on the normativity of uncoerced discussion in which the only power is that of the better argument. Rationality and participation are his requirements for just communication; he does not believe that the culture industries (mass media) will do the job. Though his practical alternative is not always clear, his aim is to deepen social democratic ideas and practices of mass communication (Calhoun 1992).

With the globalization of the world economy and the deregulation of the media in the USA, Europe and elsewhere in the last two decades of the twentieth century, classic liberal ideas have come roaring back in the realm of mass communication. Talk of liberty and the free market has been persuasively used to justify massive mergers of newspapers, radio and TV stations, film studios, cable services and internet services, among other outlets of mass communication. Neoliberal arguments have emphasized one side of liberal doctrine (the lack of state regulation of the market of goods and ideas) over the other (the creation of free, open, and diverse public fora). Concentration of ownership in media industries, as well as their conglomeration with non-media industries, has raised fears about conflicts of interest and monopolistic (propagandistic) control. Hence, many recent critics of the deregulation and commercialization of the media have sought to reclaim the affirmative or socially responsible side of liberalism, much like earlier critics (Keane 1991; see *Journalism*).

2.3 Other Criticisms

A long tradition of social research on the production of news shows how the liberal dream can serve as a legitimating ideology for professional journalists and mask the social reality of newswork. The sociology of news production shows that journalists follow formulaic routines and organizational cultures that shape their output more decisively than does a fearless quest for truth. The actual content and practices of journalism seem unaffected by the liberal idea of the press. Exalted rhetoric about the free press coexists with a long history of sensationalism that dates to eighteenth century broadsheets, if not earlier. Walter Lippmann's *Public Opinion* (1922) captures the realist bent in empirical research on journalism well by arguing that liberal dreams should be replaced by a more accurate scientific modes of analysis and reporting.

Postmodernism offers another philosophical attack. The liberal model of communication, postmodernists argue, is simplistic, missing the instability of language. Further, in an age rife with visual rhetorics, censorship should not be the chief worry. Public consciousness can be manipulated through too much information, or information selectively presented or 'spun'. For postmodernists, liberalism's fear of coercion from church or state leaves it unprepared for the simulations and dreamworlds that stem from the market or the psyche itself. Not only suppression but seduction shapes the public agenda.

A similar point is made by cultural Marxists: instead of overt coercion, power largely works through the co-optation of consent or the creation of 'hegemony' (see *Hegemony: Cultural*). The liberal public sphere can serve as a forum of pleasurable domination. In contrast to the liberal emphasis on dissemination,

Marxist-inspired cultural studies often focus on reception: how audiences read messages against the grain of dominant ideologies. Here, the liberal emphasis on active citizens is mirrored, but with a richer, more politicized account of the receiving end of mass communication. As in classical Marxism, neo-Marxists insist on looking at the results, not just the liberties, of communication.

Communitarians, the descendents of nineteenth century rightist and centrist critics of liberal egotism, have a revisionist program for mass communication. They often assail television for its alleged stupefactions and call for new forms of news-gathering and dissemination to invigorate community ties and public involvement. For communitarians, a chief ill of contemporary life is the alienation of ordinary people from public communication. It becomes the task of the press, among other agencies of mass communication, to draw citizens into the common good. The 'public journalism' movement, which reflects communitarian ideas, calls for the participation of all citizens in open and public dialog. Its normative vision of mass communication is not the liberal free dispersion of ideas but a civic conversation of all with all about the important issues of the day (Glasser 1999).

Finally, some fault liberalism for its overly rational emphasis. The American philosopher John Dewey (1927), sometimes a communitarian hero, argued that one fault of the liberal conception of mass communication was its dryness. For him, as for the British Marxist Raymond Williams (1989), drama in its diverse forms is a key resource in modern society. Similarly, recent advocates of public service broadcasting endorse entertainment as more than frivolity; it can play a role in widening sympathies for fellow citizens, allowing a sort of communion among strangers (Curran 1991). These theorists take artistry in mass communication to be a means of expanded social imagination, rather like the republican tradition. They contest the liberal commitment to the left-brain, as it were, of mass communication: the privilege of news in public life. Fiction, they argue, can have as important a civic role as fact (see *Entertainment*).

3. Whither Media Ethics?

Media ethics has been largely concerned with the practices of media professionals, especially the analysis of hard cases that face journalists (deception, disclosure of sources, the blurry borders of objectivity, etc.). The sheer range and diversity of frameworks reflecting on mass communication and its larger social, political, moral, spiritual, pedagogical, economic, cultural and aesthetic implications suggest a richer agenda for media ethics. Questions about the tenor of public life, the experience of childhood, the relief of suffering, or the possibilities of artistic expression, for instance, might take a rightful place in debates about mass communication next to questions about liberty, information and ideology. Though various normative frameworks may be mutually incompatible in their principles, emphases, or conclusions, each one has something to offer.

See also: Art, Sociology of; Broadcasting: Regulation; Censorship and Secrecy: Legal Perspectives; Censorship and Transgressive Art; Communication and Democracy; Communication: Electronic Networks and Publications; Communication: Philosophical Aspects; Freedom of the Press; Mass Media: Introduction and Schools of Thought; Media Imperialism; Media, Uses of; News: General; Norms; Violence and Effects on Children; Violence and Media

Bibliography

Boltanski L 1993 *La Souffrance à Distance: Morale Humanitaire, Médias, et Politique*. Éditions Métailié, Paris

Calabrese A, Burgelman J-C (eds.) 1999 *Communication, Citizenship, and Social Policy*. Rowman and Littlefield, Lanham, MD

Calhoun C (ed.) 1992 *Habermas and the Public Sphere*. MIT Press, Cambridge, MA

Curran J 1991 Rethinking the media as a public sphere. In: Dahlgren P, Sparks C (eds.) *Communication and Citizenship: Journalism and the Public Sphere in the New Media Age*. Routledge, London, pp. 27–57

Dewey J 1927 *The Public and its Problems*. Holt, New York

Gans H 1999 *High Culture and Popular Culture: An Analysis and Evaluation of Taste*. Basic, New York

Glasser T L (ed.) 1999 *The Idea of Public Journalism*. Guilford Press, New York

Goux J-J 1978 *Les Iconoclastes*. Seuil, Paris

Habermas J 1962 *Strukturwandel der Öffentlichkeit*. Luchterhand, Neuwied [trans. 1989 *Structural Transformation of the Public Sphere*. MIT Press, Cambridge, MA]

Hariman R D 1995 *Political Style: The Artistry of Power*. University of Chicago Press, Chicago

Horkheimer M, Adorno T W 1947 *Dialektik der Aufklärung*. Querido, Amsterdam [trans. 1994 *Dialectic of Enlightenment*. Continuum, New York]

Hutchins R 1947 *A Free and Responsible Press. A General Report on Mass Communication: Newspapers, Radio, Motion Pictures, Magazines, and Books*. University of Chicago Press, Chicago

Ignatieff M 1997 *The Warrior's Honor: Ethnic War and the Modern Conscience*. Henry Holt, New York

Keane J 1991 *The Media and Democracy*. Polity Press, Cambridge, UK

Kieran M (ed.) 1998 *Media Ethics*. Routledge, London

Lichtenberg J (ed.) 1990 *Democracy and the Media*. Cambridge University Press, New York

Lippmann W 1922 *Public Opinion*. Harcourt, Brace and Co., New York

Moeller S D 1999 *Compassion Fatigue. How the Media Sell Disease, Famine, War and Death*. Routledge, New York

Peters J D 1999 *Speaking into the Air: A History of the Idea of Communication*. University of Chicago Press, Chicago

Williams R 1989 Drama in a dramatised society. In: O'Connor A (ed.) *Raymond Williams on Television*. Routledge, London, pp. 3–12

J. D. Peters

Mass Communication: Technology

1. Technological Determinism

The received history of the development of communications technologies is extremely straightforward. Gutenberg invented printing by moveable type by 1450; Daguerre photography in 1839; Morse the telegraph in 1843; Bell the telephone in 1878; Edison the phonograph in 1878 and the kinetoscope in 1894 (with Eastman's help in providing the film); the Lumières le cinématographe in 1895; Marconi radio in 1895; Zworykin television in 1934; von Neumann, Mauchly, and Eckert produced the first computer design (EDVAC) in 1944–5; Shockley, Brattain, and Bardeen built the transistor in 1947–8; Arthur C. Clark thought of the communications satellite in 1945 which was launched as Telstar in 1962; Jobs and Wozniak built the personal computer in 1976; and Berners-Lee conceived the World Wide Web in 1990. The history assumes that this sequence of White, Western males had the idea for the device they are credited with inventing, saw it built 'in the metal,' and then witnessed its diffusion with profound effects on society, which was, largely, surprised and transformed by its arrival. This received assumption of technology's social autonomy is termed 'technological determinism' or 'technicism,' a phrase credited to Thorsten Veblen (Elul 1964) and those who propound it are 'technicists' (see, for example, McLuhan 1964). Technicists can either be positive about technological developments ('technophile') or negative ('technophobe,' sometimes termed 'Neo-Luddite'). However, over the last quarter of the twentieth century, technicist accounts of the development of media technologies, especially the often hyperbolized technophile ones, have been increasingly questioned (see, for example, Williams 1981).

Even by its own lights, technicism omits a number of figures whose influence has been profound, for example: Applegarth and Hoe who perfected the rotary press in 1850s, the Abbé Caselli whose telegraph system produced facsimile images in 1862; Pfleumer who created audio tape in the late 1920s; Carlson the patentee of the xerox machine (1938); Kilby who designed the integrated circuit in 1958; Maiman creator of the laser in 1960; Hoff who built the first central processing unit in 1969, etc. Technicism also fails to explain why so many of its roll-call of famous 'inventors' have Dopplegängers who dispute their status—Gutenberg's claim on the press is contested by Waldvogle and Coster; Daguerre's by Fox-Talbot; Morse's by Cooke and Wheatstone; Bell's by Gray; Edison's and the Lumières' by (at least) Friese-Greene and the Skladanowsky brothers; Marconi by Lodge and Popov; Zworykin by Schoenberg, and so on. Not only are there more 'great men' but, mysteriously, they tend to have the same idea at about the same time. Most dramatically, Bell and his rival Gray arrived at the US Patent Office in Washington with designs for telephones on the very same day, February 14, 1876.

More than that, the 'inventors' all have precursors who thought up the same sort of technological application earlier but are forgotten. Francis Ronalds had his elegant static electric telegraph rejected by the British Admiralty in 1816, nearly three decades before telegraphy was diffused. David Hughes, the 'inventor' of the microphone, was told his demonstration of radio was no such thing in 1879 when it was, 16 years before Marconi's 'invention.' A. H. Reeves inaugurated the digital age with a sampling device built in the Paris Laboratory of ITT in 1938 but talk of a 'digital revolution' did not start for another half-century. A succession of small computers, beginning with the very first computer to work on a real problem, the Manchester UK Baby Mark I of 1948, were discarded in favor of large devices until the coming of the Apple II in 1976.

Most grievous of all the problems raised by technicist explanations of the development of communication systems is the central concept that technology engenders social change. The Reformation, for example, is credited to the printing press. As Marshall McLuhan, a major technicist thinker, put it: 'Socially, the typographic extension of man [which is how he described printing from moveable type] brought in nationalism, industrialism, mass markets, and universal literacy' (McLuhan 1964, emphasis added). But, if the press did 'bring in' these things, it took some centuries to accomplish this, making a mock of the norms of causality. The lesser technicist claim is that the press 'caused' the Reformation; but, as Gutenberg biographer, Albert Kapr, noted, 'the stimulus for [Gutenberg's] invention was rooted in the religious controversies of his time and other closely related needs of his age' (Kapr 1996 emphasis added). The same pre-existing pressures, including church reform, drove printing forward. The historical record in fact suggests that it is always factors in the social realm which push communications technologists into assembling their 'inventions.'

2. Social Needs

There are alternative approaches to this history which offer a 'thicker' (in Geertzian terms) account and better answer fundamental questions about the origins of media technologies and the nature of their diffusion.

A great deal of knowledge is accumulated by society (indeed, by different cultures), some formal and much informal. The common insight that leaving washing out in the sun bleaches it, a crucial perception underlying the development of photography, is an example of informal knowledge. Dr. Thomas Young's experiments of 1801 are a good example of formal scientific understanding. He was able to calculate the wavelengths of different colors using light sources shone through narrow slits to interfere with each other. One hundred and fifty years later, the utilization of such light interference patterns informed the creation of holographic imaging systems. Physicist James Maxwell's wave theory, the nineteenth century's best formal explanation of the nature of electromagnetic phenomena, is the foundation of telephony, radio, and television. Alan Turing's 1936 paper solving a cutting edge problem in pure mathematics directly determined computer design a decade later (Turing 1936).

Given the widespread understanding of this 'science,' broadly defined, many can be expected to have ideas about the technical application of such know-how in any given circumstance. For example, 115 years before Turing, Charles Babbage was discovered by the astronomer William Herschel, according to legend, dozing over logarithmic tables in the rooms of the Cambridge (UK) Analytical Society. When asked what he was thinking of, he reportedly said: 'I wish to God these calculations had been executed by steam.' 'It is quite possible,' replied Herschel (Goldstine 1972). Babbage's mechanically driven computer was never built but his ideas for the architecture of a symbol manipulating machine which would alter its operations in the light of its own computations, with a 'mill' (central processing unit) and a 'store' (memory), underpin computing design.

Babbage managed to build only part of his machine which, anyway, required power and tolerances beyond the range of Victorian engineers (Hyman 1984). In other fields, many built prototypes, not all of which failed to work. For example, the coherer was a machine designed to demonstrate the wavelike nature of electricity. A number of physicists constructed variant devices and no physics lab in the late nineteenth century would have been without one. That the coherer could be used for signaling—that it was, in fact, a radio—required the identification of a communications need; and it was the identification of this need which enabled the radio to be, in effect, discovered rather than 'invented.'

Identification occurred because of the parallel development of iron-clad warships. These vessels steamed into battle so far apart that visual contact could not be maintained through the fleet. The first demonstrations of radio by Marconi and Popov took place at sea during the summer maneuvers of the British and Russian Imperial fleets in 1895. In the same way, the railways focused experiments using electricity for signaling, which had been thought of in the mid-eighteenth century and had been producing ignored prototypes since 1816 (Ronalds) at the latest, into the 'invention' of the telegraph in the 1840s. The first telegraph wires ran down the railway tracks between Baltimore and Washington, Windsor and London, St. Germaine and Paris.

In the twentieth century, the need to build nuclear weapons served to transform the development of advanced, massive electrical calculators into an agenda which would produce electronic computers—symbol manipulators which altered their operations in the light of their own calculations—from 1948 on. The threat posed by those same weapons in the 1970s created the need for an impregnable distributed communications system to control essential military computing during nuclear attack. A quarter of a century later, the technology which did this was released to the pubic and diffused globally as the Internet.

The social need which transforms prototypes into 'inventions' need not, however, always be itself a new technology. For example, the legal invention of the modern corporation in the 1860s produced novel office needs. Use was finally found for languishing prototypes such as the typewriter which had been patented in 1714 or the adding machine which now acquired an elegant printing capacity but which had existed in basic form since 1632. The elevator, made safe by Otis in 1857, was perfected and the skyscraper office block was designed to house all these devices and the corporation which used them. It is in this social context that Bell, Grey, and others built telephones. Changes in the structure of post-French Revolution society bring together a range of well-understood chemical and physical phenomena to produce, in photography, a system of image making for use by the dominant middle-class of the nineteenth century. The *camera obscura portabilis* (1550, at the latest) and photokinesic chemistry (explored from the 1720s onwards at the latest) come together to meet this social need. By 1840, photography was 'invented' and the centuries-old Western tradition of painted aristocratic portraits, still-lives, or landscapes was democratized (Eder 1978, Freund 1982).

The general pervasive requirement that the ever-expanding urban mass be housed and entertained governs many developments. The Gutenberg flat-bed press had increased its speed from 8–16 impressions per hour (iph) to 1,200 iph between 1450 and 1800. Driven by the rising literacy required by an industrialized society, between 1800 and 1870 the press, powered by the same steam that had powered the industrial revolution over the previous century and more, reached speeds of up to 40,000 iph.

The seventeenth century lantern slide show, with its cuts, fades, dissolves, and animations, became a major urban diversion. The audience learned to sit in rows in the dark watching images (for example, *La Fantasmagorie*, a hit slide show in Paris during the French

Revolution). The use of photography for slides and then, arising from scientific investigations using photography to stop motion, the idea of a photographic strip to create the illusion of motion dates from 1861. Theater was slowly organized as actor-managers gave way to owners (often with chains of theatres), producers, stars, agents. By the 1890s, 'shows' had become 'show business.' In the US, the owners formed a cartel in 1895 and the artistes, a union. All this, as much as any insight of the Lumière Brothers, 'invented' the cinema (Chanan 1995, Allen 1980).

3. *Social Brakes*

But, as the historian Fernand Braudel has noted, with these 'accelerators' come 'brakes' (Braudel 1979). The 'invented' technology faces social forces suppressing its potential to cause radical disturbance. First when it finally emerges from the shadows it has taken longer to develop than is usually understood. Then, often, complex patent arguments or government regulatory interventions slow its diffusion. Corporations especially tend to be protected by this 'braking' process. The telegraph company Western Union was the world's biggest firm when the telephone was being diffused and although it is now far from being any such thing nevertheless it still exists. New media are meshed with old. The Internet (AOL), the moving image (Warners), and print (Time) join together and all survive. But it takes time for these arrangements and accommodations to be effected.

Thus television was a fully developed system by 1936. The basic idea for the transmission of moving images using mechanical scanning, spinning disks to scan the image had been patented in 1884 and the term 'television' coined in 1906. Electronic scanning was described in 1908. The first signals were displayed on a cathode ray tube at the St. Petersburg Institute of Technology in 1911 by Boris Rozing whose pupils Zworykin and Schoenberg led the research teams at RCA and EMI (UK), respectively. Mechanically scanned and electronic experimental systems were much touted from the 1920s on.

Using what was essentially the RCA solution, the Germans and the British began regular television services in 1936–7. Yet, with the very same technology, the Americans continued to call their transmissions during this period 'experimental.' What was at issue was RCA's dominance. (The Federal government had been struggling with the ATT telephone monopoly for half a century and did not wish to create another monster corporation.) Behind the facade of technical issues a commercial agreement was hammered out to ensure the diffusion of the technology. In 1941, the National Television Standards Committee promulgated signal standards almost the same as those which had been proposed in 1936, but now RCA's radio rival CBS and other manufacturers were able to share the technologies. The next seven years of World War II and its aftermath halted development, although the Germans continued to broadcast throughout the conflict using the system the Americans had put on hold. Then the same suppressive pattern emerged again after the war as the government stopped the award of TV licenses ostensibly to sort out overlapping signals but actually to allow radio and film interests to reach a *modus vivendi*. At the start of this 'freeze' in 1948, television was live from radio headquarters in New York. By 1952, when it was withdrawn, the schedule was dominated by filmed shows made in Hollywood—but radio interests continued to own the television networks and major stations.

The same sort of delays, allowing society fully to absorb the new system, affects all technologies. The growth in patents is often suggested as an indicator of the increased range of contemporary innovation and implies speedier take-up of new developments. However, improvements and modifications can be patented in exactly the same way as breakthroughs and fundamental innovations and the former account for much patent activity. There is no evidence in communications of a greater number of basic innovations. Nor is there evidence of a speed-up in diffusion.

It took up 16 years and an act of the US government (the All Channel Law, 1964) to force manufactures to give up tubes in TVs and use solid-state circuits instead. The audio-CD, based on a computer data storage system, was marketed by Sony and Philips in 1983 to stem the flow of home LP disk recording. Digital sound quality could not be maintained on analogue audiotape. But at the same time digital audiotape (DAT), which could 'clone' CDs, was available. In fact, the first DAT studio recorder had been sold in 1971. Philips, having just built two massive CD manufacturing plants, in 1984 directly frustrated industry discussions about the introduction of DAT which remained a marginal medium for the rest of the century (Gow 1989, Morton 2000).

4. *Transformative Technologies*

The result of the push of social need being constrained by the brake stifling radical disruptive potential is that new communications technologies, however influential on society, do not transform it. Given that communication systems are created, introduced, and diffused in response to pre-existing social needs, it follows that their capacity for radical disruption, pace technicist hyperbole, will be limited.

Of course, a new technology can have unintended consequences—fax machines can facilitate junk mail; the Internet can transform the distribution and profitability of pornography. But technicist rhetoric suggests that transformative end effects cannot be known

and yet, outside of comparatively marginal surprises, this is not the case. The end effect is that transformative radical potential will be contained.

The needs of flexible manufacturing have internationalized the corporation which in turn required the internationalization of communication. To ever-evolving trans-oceanic cable capacity has been added the satellite communications system to facilitate a worldwide market. But a worldwide market is still a market, more efficient and all-embracing but essentially the same. This did not change when Internet communication facilitated data transmission, personal messaging, and commerce.

This is not to say, however, that outside of the West, these technologies cannot have more transformative effects. For example, from 1977–9, the messages of the exiled Ayatollah Khomeini were transmitted from France to Iran by telephone and distributed by audiocassette. The Westernizing Shah was removed (Mohammadi 1995). This illustrates both transformative communications power and, on the other hand, the limits of media effects. As much as Western technology aided the Iranian Islamicist revolution, at the same time it failed the Shah's Westernizing mission.

Centralized governments can also 'resist,' as when China, for example, ordered Rupert Murdoch in 1994 to remove politically offensive channels (BBC World) from AsiaSat 2 and he did so. The technology itself can have limited reach. In 2000, the Internet required literacy, expensive equipment, and easy access to telephones. The International Telecommunications Union estimated that more than two-thirds of humanity lacked even the last. The Internet clearly was not transformative for them.

In the West, although the technologies are pervasive, their revolutionary, as opposed to evolutionary, potential can be disputed as well. This argument depends on the viewpoint. Technicists, whether technophile or technophobe, are amnesiac about technological history and privilege technology as a monocausal social driving force. They erroneously believe that the pace of innovation has quickened over the last two centuries when historical understanding reveals that it has not. They also think, in ignorance of the record, that the processes of innovation have become more structured when, again, human intuition remains as strong a factor as ever. They assume, finally, that more information is in fact significant of itself and ignore its quality and the limited human capacity to absorb it. Henry Thoreau wrote in 1854: 'We are in great haste to construct a magnetic telegraph from Maine to Texas; but Maine and Texas, it may be, have nothing to communicate' (Czitrom 1982). Any Internet search suggests that Thoreau's hesitancy is still merited.

For technicists, each and every change in communications, and in society generally, looms large. A nontechnicist viewpoint, on the other hand, focuses on social continuities. On this basis, the nontechnicist tends to resist even the rhetoric of cumulative technologies converging to effect transformative change. Convergence of signal processing is of significance for communications organizations since it blurs any technological basis for distinguishing their activities. But many already existed across media technologies as conglomerates. Their power and growth has far more to do with national regulation (or its lack) and economic globalization than it does with machines, including digital ones. Communications technology will remain what it has always been, one factor among many, in determining the structures of world communication systems (see Winston 1998).

See also: Broadcasting: Regulation; Communication: Electronic Networks and Publications; Electronic Democracy; Information Society; Information Society, Geography of; Mass Communication: Empirical Research; Media and History: Cultural Concerns; Media Effects; Media Imperialism; Media, Uses of; Technological Determinism; Technology and Organization; Technology and Social Control; Television: History; Television: Industry

Bibliography

Allen R C 1980 *Vaudeville and Film, 1895–1915: A study in Media Interaction*. Arno Press, New York

Braudel F 1979 *Civilisation and Capitalism: 15th–18th Centuries*. Harper Row, New York, Vol. 1

Chanan M 1995 *The Dream That Kicks: The Pre-History and Early Years of Cinema in Britain*. Routledge, London

Czitrom D J 1982 *Media and the American Mind: From Morse to McLuhan*. University of North Carolina Press, Chapel Hill, NC

Eder J M 1978 *History of Photography* [trans. Epstean E]. Dover, New York

Elul J 1964 *The Technological Society*. Knopf, New York

Freund G 1982 *Photography and Society*. David R. Godine, Boston

Goldstine H H 1972 *The Computer from Pascal to von Neumann*. Princeton University Press, Princeton, NJ

Gow J 1989 *The Development and Suppression of the Radical Potential of DAT*. Penn State School of Communications Working Papers, State College, PA

Hyman A 1984 *Charles Babbage: Pioneer of the Computer*. Oxford University Press, Oxford, UK

Kapr A 1996 *Johann Gutenberg: The Man and his Invention* [trans. Martin D]. Scolar, Aldershot, UK

McLuhan M 1964 *Understanding Media: The Extensions of Man*. Methuen, London

Mohammadi A 1995 Cultural imperialism & cultural identity. In: Downing J, Mohammadi A, Sreberny-Mohammadi A (eds.) *Questioning the Media*. Sage, Thousand Oaks, CA

Morton D 2000 *Off the Record: The Technology and Culture of Sound Recording in America*. Rutgers University Press, New Brunswick, NJ

Turing A 1936 On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* **42**: 230–65

Williams R 1981 *Culture*. Fontana, London
Winston B 1998 *Media Technology and Society: A History: From the Telegraph to the Internet*. Routledge, London

B. Winston

Mass Killings and Genocide, Psychology of

To understand and ultimately prevent mass violence, like the Holocaust, the genocide of the Armenians, the genocide in Rwanda, the 'autogenocide' in Cambodia, the mass killing in Argentina and many others, the social conditions, cultural characteristics, psychological processes of groups and individuals that lead to it must be identified. Important questions and issues include: what are the motives of perpetrators, how do these motives evolve, how do inhibitions against killing decline? What are the usual starting points or instigating conditions? What characteristics of cultures and social processes contribute? What is the psychology of perpetrators and bystanders? Intense group violence is usually the outcome of an evolution: how does this take place, how do individuals and groups change along the way? An essential source of groups turning against other groups is the human proclivity to differentiate between 'us' and 'them' and the tendency to devalue 'them.' How can this tendency be mitigated? The passivity of bystanders, ranging from individuals to nations, encourages perpetrators. How can preventive responses be promoted? (The conception of origins that follows is based primarily on Staub 1989; of prevention on Staub 1999. See also Charny 1999.)

1. *Instigators of Collective Violence*

These are conditions in a society or in a group's relationship to another group that have great impact on people. They give rise to psychological reactions in individuals and whole groups of people, and actions and events in a society or social group that lead the group to turn against another group, often a subgroup of the society.

1.1 *Difficult Life Conditions*

One starting point for mass violence is difficult conditions in a society, such as severe economic problems, great political conflicts, rapid social changes, and their combinations. These have intense psychological impacts. They frustrate basic psychological needs for security, for a positive identity, for feelings of effectiveness and control, for positive connections to people, and for a comprehension of reality.

1.2 *Group Conflict*

One type of conflict involves 'vital' interests. Even though these conflicts have objective elements, such as some territory both sides need for living space, the psychological elements (the territory is part of the groups' identity; mutual devaluation, distrust and fear; unfulfilled basic needs) make the conflict especially difficult to resolve. Another type of conflict is between dominant and subordinate groups in a society. Frequently, demands by the subordinate group for greater rights start active violence between the groups that may end in mass killing or genocide (Fein 1993).

Conflicts between groups also tend to frustrate basic human needs. Dominant groups, faced by demands from a subordinate group, often protect not only their rights and privileges, but also their security and identity, as well as their comprehension of reality, which includes their view of the 'right' social arrangements. Difficult life conditions often intensify the impact of group conflict.

1.3 *Self-interest*

When a subgroup of society is greatly devalued (see below), a superior group may engage in mass killing to advance its interests. Mass killing or genocide of indigenous peoples has often been in part due to the desire to gain land or develop resources where these groups have lived (Hitchcock and Twedt 1997).

2. *Turning Against the Other*

Groups could respond to instigating conditions by cooperative efforts to improve conditions or by resolving conflict through negotiation and mutual concessions. Instead, frequently a complex of psychological and social processes arise that turn the group against another and ultimately lead to violence. Individuals turn for identity and security to a group; people elevate their group by devaluing or harming others (Tajfel 1978); they scapegoat another group for life problems or blame the other for the conflict; ideologies are adopted that offer a vision of a better life (nationalism, communism, Nazism, Hutu power in Rwanda and so on), but also identify enemies who must be 'dealt with' in order to fulfill the ideology.

3. *The Evolution of Destructiveness*

The group and its members begin to take actions that harm the other group and its members, which begins an evolution. Individuals and whole groups 'learn by doing.' As they harm others, perpetrators and the whole society they are part of begin to change. Just world thinking, the belief that the world is a just place and those who suffer must have somehow deserved their suffering, leads to greater devaluation of the

victims. In the end, perpetrators, and even bystanders, exclude the victimized group and its members from the moral realm, the realm in which moral values and standards apply. They often replace moral values that protect other people's welfare with values such as obedience to authority or loyalty to the group. As the evolution progresses, individuals change, the norms of group behavior change, new institutions are created to serve violence (for example, paramilitary groups).

4. *Contributing Cultural Characteristics*

Certain characteristics of a culture make it more likely that, in response to instigation, the psychological and social processes that initiate group violence arise.

4.1 *Cultural Devaluation*

One of these is a history of devaluation of another group or subgroup of society. Such devaluation can vary in intensity: the other is lazy, of limited intelligence; the other is manipulative, morally bad, dangerous, an enemy that intends to destroy one's own group. A devalued group that does relatively well—its members have good jobs—is an especially likely victim. Sometimes two groups develop intense, mutual hostility, which has been referred to as an ideology of antagonism (Staub 1989). Seeing the other as their enemy, and themselves as an enemy of the other, becomes part of their identity. This makes group violence more likely.

4.2 *Overly Strong Respect for Authority in a Society*

This makes it especially difficult to deal with instigating conditions. Accustomed to being led, people are more likely to turn to leaders and ideological groups. They are unlikely to oppose it when the group increasingly harms another group. They are also more likely to follow direct orders to engage in violence.

4.3 *Monolithic (and Autocratic) vs. Pluralistic (and Democratic) Societies*

The more varied are the values in a society, the greater the extent that all groups can participate in societal processes, the less likely is the evolution towards mass violence. People will be more likely to oppose harmful, violent policies and practices. Democracies (Rummel 1994), especially mature ones (Staub 1999) that are pluralistic and have a well-developed civic culture, are unlikely to engage in genocide.

4.4 *Unhealed Wounds of a Group Due to Past Victimization or Suffering*

Without healing from past victimization, the group and its members will feel diminished, vulnerable, and see the world as a very dangerous place. At times of difficulty or in the face of conflict, they may engage in what they think of as necessary self-defense. But, instead, this could be the perpetration of violence on others (Staub 1998).

4.5 *Other Cultural Characteristics*

A history of aggression as a means of resolving conflict, as well as certain group self-concepts—a sense of vulnerability, or a feeling of superiority that is frustrated by events, or the combination of the two—make violence also more likely.

5. *The Role of Bystanders*

The passivity of bystanders greatly encourages perpetrators. It helps them believe that what they are doing is right. Unfortunately, bystanders are often passive. Sometimes they support and help perpetrators (Barnett 1999, Charny 1999, Staub 1989, 1999).

Internal bystanders, members of the population, often go along with or participate in discrimination, and ignore violence against victims. As a result, just like perpetrators, they change. Like the perpetrators, bystanders, as members of the same society, have also learned to devalue the victims. They are also affected by instigating conditions. It is difficult for them to oppose their group, especially in difficult times and in an authority-oriented society. To reduce their guilt, and their empathy, which makes them suffer, bystanders often distance themselves from victims. Over time, some become perpetrators.

External bystanders, outside groups, and other nations, also usually remain passive, continue with business as usual, or even support perpetrators. Nations do not see themselves as moral agents. They use national interest, defined as wealth, power, and influence, as their guiding value. When they have ties to another country, they tend to support the leaders, not a persecuted group.

6. *The Role of Leaders*

It is the inclinations of a population, the result of the joining of culture and instigating conditions, that to a substantial degree create the possibility and likelihood of mass murder. To some degree, the people select leaders who respond to their inclinations and fulfill their needs.

Still, leaders and the elite have an important role in shaping and influencing events. They scapegoat and offer destructive ideologies, and use propaganda to intensify negative images and hostility. They create institutions, such as media and paramilitary groups that promote or serve violence. Often such leaders are seen as acting only to gain support or enhance their power. But leaders are also members of their society, impacted by life conditions and group conflict, and, at least in part, act out of the motives and inclinations described above.

7. *Other Influences*

One additional influence is a sudden shift in government, combined with 'state failure,' the new government failing to deal effectively with problems that face the society. An ongoing war, especially a civil war, also adds to the probability of mass killing or genocide. Economic interconnections between a country and other countries make genocide and mass killing less likely (Harff 1996, Melson 1992).

8. *Halting Persecution and Violence*

Once violence against a group has become intense, halting it requires action by nations and the community of nations. Early warning is important, but not enough. Usually, as in the case of the genocide in Rwanda in 1994 (des Forges 1999), when information about impending violence is available, the international community does not respond. For this to change requires changes in values, and actions by citizens to bring them about. It also requires institutions to activate and execute responses by the community of nations. Interconnected institutions within the UN, regional organizations, and national governments are needed.

Appropriate actions include diplomatic efforts: to warn perpetrators, as well as to offer mediation and incentives to stop violence. Such efforts must be accompanied or followed, as needed, by withholding aid, by sanctions and boycotts—ideally designed to affect leaders and elites—and the use of force, if necessary. But early actions, especially preventive actions, are likely to reduce the need for force (Carnegie Commission on the Prevention of Deadly Conflict 1997, Staub 1999).

9. *Preventing Mass Violence*

Preventive actions by bystanders or 'third parties' are important when conditions exist that are likely to lead to mass violence. Helping previously victimized groups heal, making sure that the truth about what happened in a case of prior collective violence is established, helping perpetrators heal—who are wounded, at the very least, by their own violent actions—reduce the likelihood of new or renewed violence. Healing furthers the possibility of reconciliation. Creating positive connections between groups, shared efforts on behalf of joint goals, help people overcome past devaluation and hostility. Coming to understand the other's history and culture is also important. Assumptions of responsibility and expressions of regret by perpetrators (or mutually, when violence was mutual), can further healing, forgiveness, and reconciliation. The punishment of especially responsible perpetrators (but not revenge on a whole group) is important (Staub 1999).

In the long run, the economic development of a country can diminish the likelihood of violence. This is especially so if there is also democratization, which creates pluralism, moderates respect for authority, and lessens differences in power and privilege. The positive socialization of children, the development of inclusive caring, caring that extends beyond the group, is also essential.

10. *Future Directions*

This article has reviewed influences that lead to varied forms of mass violence. Further research ought to consider whether specific forms of violence, such as government persecution, conquest, revolution, civil war, and others, which may ultimately lead to mass killing, also have specific or unique determinants (Totten et al. 1997). Testing our capacity to predict group violence is important. So is the development of techniques to help groups heal and reconcile (Agger and Jensen 1996, Staub 2000). Creating positive bystandership by nations and nongovernmental organizations is essential for prevention. Citizen involvement is required for this. When this happens, the effects of different types of bystander behavior need to be assessed. To create a less violent world, the development of knowledge in this realm and its application has to go hand in hand.

See also: Ethnic Cleansing, History of; Ethnic Conflict, Geography of; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Genocide: Anthropological Aspects; Genocide: Historical Aspects; Holocaust, The; Violence, History of; Violence in Anthropology

Bibliography

Agger I, Jensen S 1996 *Trauma and Recovery Under State Terrorism*. London

Barnett V J 1999 *Bystanders: Conscience and Complicity During the Holocaust*. Greenwood Press, Westport, CT

Carnegie Commission on the Prevention of Deadly Conflict1997 *Preventing Deadly Conflict: Final Report*. Carnegie Corporation of New York, New York

Charny I W (ed.) 1999 *Encyclopedia of Genocide*. ABC-CLIO, Santa Barbara, CA, Vols. 1 and 2

des Forges A 1999 *Leave None To Tell The Story: Genocide in Rwanda*. Human Rights Watch, New York

Fein H 1993 Accounting for genocide after 1945: Theories and some findings. *International Journal of Group Rights*. **1**: 79–106
Harff B 1996 Early warning of potential genocide: The cases of Rwanda, Burundi, Bosnia, and Abkhazia. In: Gurr T R, Harff B (eds.) *Early Warning of Communal Conflicts and Genocide: Linking Empirical Research to International Responses*. United Nations Press, Tokyo
Hitchcock R K, Twedt T M 1997 Physical and cultural genocide of various indigenous peoples. In: Totten S, Parsons W S, Charny I W. (eds.) *Century of Genocide: Eyewitness Accounts and Critical Views*, Garland Publishing, New York
Melson R 1992 *Revolution and Genocide*. University of Chicago Press, Chicago
Rummel R J 1994 *Death by Government*. Transaction, NB, Canada
Staub E 1989 *The Roots of Evil: The Origins of Genocide and Other Group Violence*. Cambridge University Press, New York
Staub E 1998 Breaking the cycle of genocidal violence: Healing and reconciliation. In: Harvey J (ed.) *Perspectives on Loss: A Source Book*, Taylor & Francis, Washington, DC
Staub E 1999 The origins and prevention of genocide and other group violence. *Peace and Conflict: Journal of Peace Psychology*. **5**: 303–36
Staub E 2000 Genocide and mass killing: Origins, prevention, healing and reconciliation. *Political Psychology*. **21**: 367–79
Tajfel H 1978 Social categorization, social identity and social comparison. In: Tajfel H (ed.) *Diferentiation Between Social Groups*. Academic Press, London, pp. 61–76
Totten S, Parsons W S, Charny I W (eds.) 1997 *Century of Genocide: Eyewitness Accounts and Critical Views*. Garland Publishing, New York

E. Staub

Mass Media and Cultural Identity

The term 'mass media' refers to media of communication such as the printed press, cinema, radio, and television which have characteristically addressed large and diverse audiences. The term 'cultural identity' refers to the attribution of a given set of qualities to a particular population. A cultural identity is not static and eternal but rather changes through time. Cultural collectivities commonly think of their identities in terms of how they differ from others. Distinguishing 'us' from 'them' is therefore central to groups' cultural self-identification. In this article, the relationship between mass media and debates about cultural identity is examined.

1. *How Mass Media Relate to Cultural Identity*

Throughout the history of mass communication, there has been a continuing interest in the relationships between mass media and cultural identities. This has been driven by the general, underlying assumption that different media variously form, influence, or shape collective identities as such. How mass media might, at various times, specifically affect the formation and maintenance of national cultures and identities has occupied center stage in this continuing discussion. National identities normally embrace a range of cultural identities based on ethnic, religious, and linguistic diversity. Nationalists, however, often wish to portray national identity as based in a single culture. Nation-states remain, for the most part, the political frameworks in which claims about cultural diversity and expressions of cultural identity within a given society are pursued. Nevertheless, cultural identities may transcend the state system, as for instance in the case of systems of belief that unite members of the world religions. Diasporic communities also share cultural identities despite belonging to different nations and holding diverse citizenships (Schlesinger 1991, Hall and du Gay 1996).

1.1 *Mass Media and National Identities*

From a historical perspective, mass media have been an important part of the nation-building process. Successively, the press, the cinema, radio, and television have been invoked as shapers of collective consciousness, as the bearers of a collectivity's culture, or as objects of policy for shaping collective identities. To the extent that the mass media have contributed to the construction of national cultures, they have also necessarily played a role in the creation of national identities, which are still the most important forms of modern cultural identity in the new millennium.

In the latter half of the nineteenth century, the development of a popular press, nationally distributed magazines, and news agencies making use of the telegraph reinforced the creation of national communication systems in countries such as the USA, the UK, and France. Mass communication developed intensively within the boundaries of national states and through widespread, cross-class consumption, it contributed to the creation of common space for public debate, a shared political agenda, and that sense of collective belonging that characterizes national identities (Carey 1989). Mass communication has therefore played an increasingly central part in the formation of national political cultures. The latter part of the nineteenth century also saw the first major steps in the development of what would come to be a worldwide communications capacity carried through the spread of telegraphy and subsequently telephony in part in line with European imperial expansion. These nineteenth-century beginnings have proved to be the foundation stones of the developing global communications infrastructure which has become steadily more relevant for debate about the devel-

opment of transnational cultural identities at the end of the twentieth century and at the start of the twenty-first.

1.2 Audiovisual Culture

The rapid growth of the cinema before and up to the First World War inaugurated an era in which the moving image ultimately came into its own as the globally dominant cultural form. From its inception, the cinema developed as a national cultural institution. It provided images of the people and, through stardom, of popular icons of identification that played into the elaboration of national identities by representing a culture. At the same time, however, the early cinema was also an international medium that addressed diverse audiences across the boundaries of nation-states. This duality has made the relationship between the moving image and the maintenance of cultural identity within the territorial boundaries of the nation-state an especially potent issue for many countries.

Since World War I, the role of Hollywood in the international cinematic marketplace has been of prime importance. What has accompanied this primacy has been a widespread concern, long articulated in Europe by politicians and cultural elites from the 1920s onwards, about the impact of 'Americanization'—the exporting of US cultural values, tastes, attitudes, and products—on national cultures and identities. The most vociferous critics have seen the transnational flow of the moving image as an instrument of a US government policy intent on global cultural domination. Other detractors have seen 'Americanization' as the unintended outcome of a successful model of production that invites competition as a mode of self-defense.

Given its salience, it is hardly surprising that throughout the twentieth century 'Americanization' has repeatedly come to the center of public debate. Looked at from the vantage point of elites, it has been identified with the reshaping of popular culture and seen as having a negative impact on actions, beliefs, and identities, especially among young people and the working class. For such groups, American cultural forms have been a key source of pleasure: they have embodied modernity and an escape from a dominant version of national culture.

Although Hollywood's output has often been the principal point of attack, there has also been a much more diffuse concern among national elites with the effects of consumerism on the moral fabric of their societies. So, for instance, jazz and rock, fast food and Coca Cola, shopping malls and coffee bars, advertising and blue jeans have each, at times, come to represent the threatening and transformative ubiquity of America to disconcerted moral guardians. The cultural challenge constituted by 'Americanization,' moreover, has also gone deeper into the social fabric to encompass the challenges represented by the economic model of free-market capitalism. This has long been seen as inimical to the social solidarity ideally aspired to by European welfare states. Not surprisingly, then, the history of US market dominance and cultural anxiety in recipient countries has played into intercultural relations across the Atlantic. It also shaped the wider, global debate about cultural and media imperialism.

Alongside the rise of the cinema, the development of nationally based radio broadcasting in the inter-War years and through to the 1950s ensured the dissemination of a range of genres—drama, music, news, comedy, sport—that addressed segments of nationally constituted audiences. Radio domesticated national culture, defined national public spheres, and ensured that cultural identities became deeply mass mediated in many respects.

The audiences for radio and, later, television shared a public world which persisted for most of the twentieth century. The classic vehicle for explicitly pursuing the role of televising the national culture from the 1950s onwards was public service broadcasting, of which the British Broadcasting Corporation (BBC) has been the key model, broadly replicated in many countries (Scannell and Cardiff 1991).

Before its networks and channels began to fragment and specialize with the diversification of distribution systems, general broadcasting offered a container of common mediated experiences to its audiences and articulated these with other forms and sites of cultural consumption. Formally, through the education system, learners would encounter variants of the national folk history and absorb key literary reference points that in combination offered marks of collective identity and thereby constituted collective memories. More informally, outside the sphere of broadcasting, whether as citizens or consumers, people encountered the national cultural gamut in all its variety. This included recorded and live music, popular comedy, theatrical performances, and national and local sporting events. State ceremonies and religious celebrations punctuated the calendar. By both complementing and re-engineering these cultural forms into its own output, broadcasting invited listeners and, later, viewers into a new relationship of social communication. It offered shared ways of speaking and looking that became an integral part of everyday life. It invented schedules that structured temporal patterns and genres that shaped cultural expectations.

Although its transmissions are no respecters of political or cultural boundaries, for the most part radio has been experienced by listeners as a national medium. However, it has always played an important transnational role in wartime propaganda, whether in the hot war of World War II or during the Cold War from the late 1940s until the turn of the 1990s. And music radio has always had strong cross-border

attractions, sustaining international communities of taste.

Besides public service systems, commercial systems of broadcasting such as the US networks—ABC, CBS, NBC—both in their heyday and since, have also had a profound impact on the mediation of national culture by way of providing a wide range of televisual forms. US programs have been pre-eminent in defining the international experience of television for many countries. National television systems, and latterly subscription-based movie channels, moreover, have been important vehicles for the dissemination of cinema, especially of Hollywood's output, which has thereby reached a global audience alongside distribution via the theatrical circuit and through the sale and rental of video recordings.

As the ways in which television was distributed rapidly diversified in the 1980s and 1990s—with the growth of cable and satellite broadcasting, alongside the original terrestrial systems—the relations between cultural identity and the medium became more complex. The national audience's experience of television consumption began to fragment. These changes ushered in the crisis of public service broadcasting in many countries. Fragmentation of the national television experience is likely to be further accentuated as digital signals increasingly replace analogue transmissions, and as the range of channels on offer multiplies rapidly in the most developed parts of the world. It is also going to be affected in increasing measure by the delivery of radio, television, and music over the Internet, as these will impact on the range of available choices and bypass broadcasting schedules.

2. *Cultural Identity, Mass Media and Cultural Imperialism*

The debate about the impact of mass media on cultural identity increasingly rose up the policy and academic agendas in the 1970s. This coincided with the new interest being taken in the cultural impact of films and television programs that were crossing state frontiers. Shaped by the political contours of the Cold War, in which the strategic interests of the capitalist West were both symbolically, and in reality, opposed to those of the communist East, the debate was greatly stimulated by the postcolonial and postimperial interest in the influence exercised over the rest of the world of cultural production deriving from the main capitalist states (Tomlinson 1991).

2.1 *International Debate on Cultural Identity*

Discussion centered on the work of the United Nations Educational, Social and Cultural Organization (UNESCO) and on that of the critical writings of media scholars such as Schiller (1969) and Mattelart (1979). In the early 1970s, the first major studies of international television flows were published and academic debate about the significance of audiovisual trade intensified. In Latin America, which became the key locus for critical thought, arguments centered on the economic and cultural dependence of Third World states on those of the developed First World and the perceived threat posed to indigenous cultures and identities by popular media and consumerism.

UNESCO's interventions stimulated a critique of US cultural and media imperialism. The role of journalism (not least US-, UK-, and French-owned international news agencies) was central to this debate. The news agencies were seen as exercising monopolistic control over news flows and as screening out 'south–south dialogue.' Arguments centered on the worldwide impact of the consumption of Western cinematic and televisual fiction as well as that of news. To counter media imperialism, its critics proposed that a New World Information and Communication Order (NWICO) be instituted. The idea of such an international order implied the existence of a transnational communicative space that should and could be regulated in the interests of more equal dialogue. The defense of national cultures in weak and dependent nations and the pursuit of equity and balance in the flow of reports and images in international cultural trade reached its high-water mark in the UNESCO-supported MacBride report (MacBride 1980). This debate prefigured later attempts to discuss the conditions for creating an international public sphere of open debate (Habermas 1997).

2.2 *Supranationalism and Cultural Identity*

The defense of national cultural identities by nation-states has remained on the international media research agenda, and not just in the Third World. The European Union (EU) has been a key test case for thinking about the relationships between mass media and cultural identity. The EU is the outcome of a half-century-long process dating from World War II, in which previously warring states have sought to develop a framework for political and economic integration. By the end of the twentieth century, 15 European states were bound together by treaty as a trading bloc in the global economy, with a growing line waiting for accession. Official thinking in the EU has assumed a strong causal connection between media consumption and cultural identity. The desired shaping of a new, 'European' cultural identity has been linked to the formation of a supranational political public sphere and a common citizenship.

EU policy-makers' views on imports of US audiovisual products have strikingly echoed arguments aired in the NWICO debate: from a cultural per-

spective, the flow of films and television programs has been represented as a threat to European identity. As noted earlier, ever since World War I, political and cultural elites in Europe have been concerned about the impact of American films on their national cultures and identities (Jarvie 1992). The original national focus has been easily transposed into official, supranational preoccupations about 'European culture.'

The cultural case for domestic film-making has been that home-made moving images represent the nation and invite its members to see themselves as part of a community with a common culture. As television broadcasting became a cultural force across Europe from the 1960s on, official and industrial worries about US dominance of the box office extended to the small screen. There has also always been a major economic interest in limiting imports and safeguarding jobs at home. Hence, the EU successfully negotiated the 'cultural exclusion' of audiovisual services from the final settlement of the General Agreement on Trade and Tariffs (GATT) in 1993. The USA did not accept this principle and the issue remains on the agenda of the successor World Trade Organization.

The EU has found it impossible so far to create a single European cultural marketplace for the moving image. The aspirations have stretched no further, and reasonably so. That is because in the case of radio, television (imported entertainment aside), and the press, cultural preferences have remained predominantly national and regional, following the contours of states and languages. There is no reason to think that this is going to change, even though English is increasingly widely diffused as a second language.

Despite their political convergence, European states still offer a bulwark against widespread cultural homogenization. Publishing, the press, radio, and national education systems continue to sustain diversity and difference in each country. Perhaps it is not so surprising, then, despite some relatively modest investment in financing cross-national film and television coproductions, in supporting producer training and distribution, and in enforcing the provisions of the EU's single-market legislation, that there is little evidence of the growth of a major transnational audience for European audiovisual products. Both national and transnational patterns of cultural consumption co-exist. So, at the start of the new millennium, despite 15 years of policy intervention, Hollywood still provided the common audiovisual diet, with national cinema an also-ran everywhere apart from France, where cultural protectionism had been most vigorously enforced.

Although a supranational polity, the EU is based on nation-states that seek to maintain their distinctive national cultures. The EU's integration process has stimulated a reassertion of regional identities, thereby undercutting the hegemony of nation-states. In some instances, these regions are also 'nations without states' such as Catalonia and the Basque Country in Spain, and Scotland and Wales in the UK. There, regionalism is inescapably connected with the protection of national and cultural identity. Particular importance has been attached to the role of indigenous mass media in sustaining distinct national cultural identities within the wider state.

2.3 *Mass Media, Language, and Cultural Identity*

National media systems have been crucial disseminators of standardized vernaculars. Anderson (1991) has suggested that the convergence of capitalism and printing created the conditions for the dissemination of secular vernaculars in the early modern period. 'Print language' allowed national languages to be disseminated through markets for books and newspapers, creating a common culture. Audio and audiovisual media have performed a similar role in reinforcing the national idiom. For instance, it was not until the post-World War II period that the majority of Italians came to speak standard Italian, and this was substantially due to the role of radio and television broadcasting, reinforcing and complementing the earlier influence of the cinema. Similarly, in Latin America, it was the mass media that first gave audiences images of themselves that they could connect to their everyday life experiences, offering them an entry point into the national culture and the political public sphere. The mass media also provided the idioms of a common culture. Key instances were the Mexican cinema of the 1930s to 1950s, Argentinean radio drama from the 1920s to the 1940s, and popular journalism in Chile from the 1920s to the 1950s (Martín-Barbero 1993).

Although mass media have had a homogenizing impact within national boundaries, in some places they have been used to ensure language diversification. This has depended on the political efficacy of language activists seeking access to the media. Success or failure has depended on the status of the language concerned, the amount of political support for broadcasting or otherwise disseminating a language, whether a language campaign has been effectively led, and the state authorities' willingness to encourage linguistic diversity.

Many European states regulate their broadcasting systems to ensure that minority and 'lesser-used' languages are assured of airtime and thereby encourage and sustain a diversity of cultural expression and identities. The policies of individual states have been supported since 1992 by a European Convention on regional or minority languages and in several cases—notably in the Basque Country and Catalonia in Spain, Wales in the UK, and among the Sami of Scandinavia—minority-language broadcasting appears to have been successful in sustaining their language communities. However, Europe is by no means unique. Movements aimed at securing both public

support for, and state endorsement of, indigenous radio and television production have been active in Australasia and the Americas (Browne 1996).

Canada has been an especially illuminating case of the complex inter-relations between linguistic and cultural identity considerations as the Canadian state has fought its identity wars on two fronts. Externally, the country has had to define itself politically and culturally against the USA. Internally, Canada has had to deal with the continuing debate over the status of Quebec, where claims to sovereignty have long been argued in terms of a distinctive national cultural identity (Collins 1990).

By the middle of the twentieth century, the advent of television extended the existing US media penetration of English-speaking Canada by print and radio. The presence of American media has long shaped Canadian policy-makers' understanding of broadcasting as essential to the maintenance of national identity and cultural sovereignty. As commercial television developed alongside the public sector, governments tried to regulate the system to ensure that Canadian content was carried by Canadian broadcasters, alongside highly popular US material. Consequently, in line with this protectionist stance, when Canada negotiated the North American Free Trade Agreement (NAFTA) with the USA in 1993, it sought successfully to ensure the 'exception' of its cultural industries from the treaty (McAnany and Wilkinson 1996). This paralleled the EU's approach to GATT, for similar reasons.

While the establishment of public service broadcasting in Canada was in part a response to the commercial model prevailing south of the border, it also took the form of reinforcing the country's cultural and linguistic fault-line between Anglophones and Francophones. Although Francophone broadcasting has sustained the national cultural identity of French speakers, this has run counter to the Canadian federal government's attempts to create an overarching political and cultural identity.

Such tensions have also been experienced in the heartland of global media production, the USA. The major migratory impact of Hispanics has transformed the cultures of states such as California and Florida and has provided the market conditions for the development of Spanish-language television channels with a national reach. This assertion of cultural and linguistic difference has provoked debate about the unique position of the English language in the USA and how multiculturalism relates to a common national identity.

2.4 Globalization and New Cultural Identities

In the twenty-first century, the impact of transnational and global changes on what are still largely nation-state-bound systems of communication is likely to increase (Castells 1997). Ideas of communication sovereignty rooted in the heyday of the nation-state are increasingly difficult to sustain in the face of cross-national media mergers, the development of supra-national polities such as the EU, and the rapid, if unequal, diffusion of access to the Internet with its capability of making global communicative connections. Although the worldwide flow of information and cultural products has now permeated the boundaries of states in unprecedented ways, such 'globalization' does not mean that the world is becoming uniform. On the contrary, the reshaping of communicative spaces worldwide means that new cultural identities will emerge continually within, and across, the existing international system of nation-states.

See also: Broadcasting: General; Cultural History; Culture, Sociology of; Entertainment; Film and Video Industry; Identity and Identification: Philosophical Aspects; Identity: Social; International Communication: History; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Mass Media, Representations in; Media and Child Development; Media Imperialism; Public Broadcasting; Public Sphere and the Media

Bibliography

Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Bauman Z 1998 *Globalization: The Human Consequences*. Polity Press, Cambridge, UK

Browne D R 1996 *Electronic Media and Indigenous Peoples: a Voice of Our Own?* Iowa State University Press, Ames, IA

Carey J W 1989 *Communication as Culture: Essays on Media and Society*. Unwin Hyman, Boston

Castells M 1997 *The Power of Identity*. Blackwell, Malden, MA

Collins R 1990 *Culture, Communication and National Identity: the Case of Canadian Television*. University of Toronto Press, Toronto, Canada

Friedman J 1994 *Cultural Identity and Global Process*. Sage, London

Habermas J 1997 *Between Facts and Norms*. Polity Press, Cambridge, UK

Hall S, du Gay P (eds.) 1996 *Questions of Cultural Identity*. Sage, London

Jarvie I 1992 *Hollywood's Overseas Campaign: the North Atlantic Movie Trade, 1920–1950*. Cambridge University Press, New York

MacBride S 1980 *Many Voices: One World*. Kogan Page, London

McAnany E G, Wilkinson K T (eds.) 1996 *Mass Media and Free Trade: NAFTA and the Cultural Industries*. University of Texas Press, Austin, TX

Martín-Barbero M 1993 *Communication, Culture and Hegemony: from the Media to Mediations*. Sage, London

Mattelart A 1979 *Multinational Corporations and the Control of Culture*. Harvester Press, Brighton, UK

Morley D, Robins K 1995 *Spaces of Identity: Global Media, Electronic Landscapes and Cultural Boundaries*. Routledge, London
Scannell P, Cardiff D 1991 *A Social History of British Broadcasting*. Blackwell, Oxford, UK
Schlesinger P 1991 *Media, State and Nation: Political Violence and Collective Identities*. Sage, London
Schiller H 1969 *Mass Communication and American Empire*. A. M. Kelley, New York
Tomlinson J 1991 *Cultural Imperialism*. Pinter, London

P. Schlesinger

Mass Media and Sports

Mass media and sports are formally separate social and cultural institutions that have, over the past century or so, become so inextricably linked that it is now almost impossible to imagine their independent existence. There are two main questions addressed by social and behavioral scientists in relation to these institutions, the first of which is the extent to which the mass media, especially TV, may be said to have 'taken over' or even 'ruined' sports. The second principal area of debate in this area concerns the relationships between media sports 'texts' and audiences, especially their capacity to reproduce or challenge prevailing structures of power. In considering these issues, this article examines first how fundamental social trends have produced the mutually attractive features of mass media and sports that have led them to become so closely entwined.

1. The Historical Emergence of Mass Media and Sports

Both mass media and sports developed rapidly in the late nineteenth and early twentieth centuries in the full flowering of Western modernity, as an unprecedented combination of industrial, capitalist, political, technological and urban change stimulated and confronted deep transformations in work and popular play. The mass media emerged, in both public (state) and private (commercial) manifestations, as the key institutionalized communicator of political discourse, carrier of news, provider of entertainment, and vector of 'promotional' messages of various kinds. As the commercial mass media developed, so did advertising as a major income stream—in the case of free-to-air TV, in particular, the need to capture audiences for advertisers became clearly paramount. At the same time, new media technologies in the print and electronic media (including more-efficient presses and transmission–reception equipment) and the proliferation of media organizations meant a heightened level of 'content hunger'—that is, a need to fill the expanding media space with material of sufficient popularity to create and sustain economically viable enterprises.

While physical play has been a feature of all known societies for millennia, the specific social institution of sports—with its rules, regular competitions, industrial infrastructure and international relations—is a product of late modernity and so is barely a century old (Elias and Dunning 1986). Sports emerged as an important collective manifestation of popular leisure and pleasure, developing (first in the UK, which, not coincidentally, was also the first industrial and capitalist power) from intermittent forms of 'folk' (physical play on feast and holidays) into codified disciplines in which participants were either lovers of the game (amateurs) or paid to perform skillfully in front of paying spectators (professionals). Historically, the British imprint on the formation of contemporary sports is substantial, although at the commencement of the new millennium it has been substantially matched by the American influence on sports television, promotion and marketing.

It is ironic, however, that while the development of sports is closely linked with the processes of internationalization, Anglo–American cultural imperialism and globalization, it is also intimately connected to the consolidation of individual nation-states. Hence, for example, the revival in 1896 of the Olympic Games by the French aristocrat Baron Pierre de Coubertin—some fifteen centuries after the end of the ancient Games (Hill 1992)—was not insignificantly motivated by a felt need to enhance France's sovereignty after a humiliating defeat in the Franco–Prussian War a quarter of a century earlier. In neighboring Belgium, by contrast, soccer was systematically organized in the twentieth century to counter French-speaking influence through a process of vervlaamsing (flemicizing) that used sports to help shape the national cultural landscape (Duke and Crolley 1996).

With the rise in popularity and increasing rationalization of sports since the late nineteenth century, it began to commend itself to the media as an ideal source of content, able to flow across the entire spectrum of news, entertainment and public affairs in a manner that has now made it all-pervasive in contemporary international media (Rowe 1999; Wenner 1998). Despite initial mutual misgivings—concerning the sustained appeal of sports for media audiences on one side and of the deleterious impact of media coverage to paid attendance at sporting events on the other—the possibilities of symbiosis gradually became clear. Sports, while a major staple subject of the press and radio, has formed its closest alliance with TV. This article, therefore, pays particular attention to the TV–sports nexus, demonstrating how their interrelationships have had a profound impact on the economics, politics and social dynamics of contemporary culture.

In general terms, the two principal forces behind the growth of media sports have been nation building and market development, with the balance and rate of development in any given context varying according to social, historical and spatial context. In the UK, for example, a strong impulse existed to nurture an established national culture through the state-fostered pioneering radio and TV broadcasting by the public British Broadcasting Corporation (BBC) of great national sports events like the Grand National and Derby horse races and the Cup Final for soccer. These were prime instances of the mass media communicating the nation to itself, and conveying a sense of a seamless notion of British (especially English) identity (see Whannel 1992). In the USA, however, a much larger, 'newer' nation in which the public organization of broadcasting was subordinated to commercial impulses, the market potential of media sports was in the first instance more speedily realized in a decentralized environment in which the state played a greater role as free-market regulator rather than as promoter of national consciousness and cultural heritage (Goldlust 1987). Such socio-historical differences have similarly marked the development of media sports in other continents and nations, but there is no doubt that media sports is one of the most potent manifestations of international (if not global) culture. One reason for this state of affairs, as noted above, is sports' extraordinary capacity to appear in many forms across a variety of national and international media.

2. *Forms of Mass Media Sports*

The many potential and actual manifestations of mass media sports can be, in part, traced alongside wider developments in media technologies, genres, organizations and audiences. The initial coverage of sports in the mass media was, of course, in newspapers, where the initial prime purpose was to report significant sports events and to record their outcomes in a manner that would attract and sustain substantial readerships. As sports developed as a popular pursuit in the late nineteenth century, sports reports became more elaborate, with more detailed descriptions of play and the often poetic establishment and elaboration of atmosphere and resonance. They also took on a character that was less retrospective and less dependent on prevailing conventions of reportage. As organized sports became more popularly prominent and commercially significant, it was also more routinely discussed in the media, and sports events were anticipated and minutely analyzed—as well as advertised and used as a vehicle for advertising. Sports illustrations, photographs and advertisements (including those that used leading sports and sportspersons to endorse products) helped to develop the visual aesthetics of media sports.

The development of the electronic media provided new possibilities for instantaneous media sports coverage that even accelerated print productions schedules could not match. From the 1920s onwards radio delivered 'real time' coverage and 'ambient' sounds to growing audiences for soccer, boxing, baseball, cricket and other sports. But it was TV that had the unprecedented capacity, above all, to simulate plausibly the experience of physically attending sports contests while the viewer was far distant. As television technology developed, the use of multiple cameras, instant replays, slow (and, later, super slow) motion and other innovations made the remote viewing experience in some ways superior to that of actual attendance at the event. It is for this reason that live TV sport is the most prized and valuable form of mass media sports, despite the development of new media (such as Internet sites).

Many other types of program can be 'spun off' live TV sports, with sports magazine programs; previews; replays and delayed telecasts; retrospectives; quiz and chat shows; and documentaries supplementing the regular sports news bulletins and updates. The print media have adapted to what we might call 'the hegemony of instantaneity' imposed by live TV sports by using the reflective space provided by the written form to expand the sports pages in newspapers (often including daily sports supplements) and to produce many general (the best known of which is *Sports Illustrated*) and specialist sports magazines (such as *Golf World*), all of which are accompanied by sports photography (sometimes in the form of sports 'photo-essays'). At the same time, sports organizations and celebrities are covered routinely in general news, not least by means of proliferating sports celebrity scandals and the more-routine use of the gossip column. When the many novels and films that deal centrally with sports are included, as well as its key role in the promotion, advertising and endorsement of products and services from soft drink and leisurewear to life insurance and fitness programs, the degree to which mass media sports has become both ubiquitous and 'multi-functional' can be fully appreciated. The success of the mass media in insinuating sports into the fabric of everyday social life—even for those who are hostile or indifferent to it—has raised significant questions about its social and ideological ramifications, and has provoked keen debate about what it means to live in societies where sports has passed, in barely a century, from 'rough play' to high finance.

3. *Debating Mass Media Sports: Theory, Research and Method*

Until recently, social and behavioral scientists have on the whole either ignored mass media sports or regarded it, unproblematically, as a symptom of the commodification of popular culture. The argument

that great global and national media sports events, with their reliance on a combination of deep belief and spectacular ritual, constitute a powerful form of secular religion under late modernity, has sometimes been advanced (for example, by Real 1989), but this position has raised rather less controversy than the contention that commerce, especially through the mass media and specifically by means of TV, has colonized (post)modern sports. The economic power that TV wields over sports can be strikingly demonstrated by the modification of sports to fit its requirements. Hence, the introduction of tie-breaks in tennis, one-day cricket, longer time-outs and the 'shot clock' in basketball, the 'designated hitter' in baseball, penalty 'shoot-outs' in soccer, more breaks for commercials in gridiron, minimal scrum rules in rugby league, and the speeding up of 'rucks and mauls' in rugby union, can be attributed wholly or partially to the drive to make sports more 'telegenic' and so commercially valuable. The political economic critique of mass media sports has, therefore, been central to debates about the large-scale growth of the media–sports nexus, not least in relation to the value of broadcast rights for live sports.

3.1 Political Economy

Live TV sport is able to bring global audiences simultaneously to the screen in massive numbers—for example, the cumulative, estimated 37 billion who watched the 1998 soccer World Cup and the guaranteed record audience for each successive opening ceremony of the Olympics. These vast viewing figures also explain why the cost of broadcast rights has escalated over the last three decades. To illustrate this point we may note that the TV broadcast rights in the USA for the summer Olympics paid by the National Broadcasting Company (NBC) network rose from US$72 million in 1980 (before the landmark entrepreneurialist 1984 Los Angeles Games, which are sometimes dubbed 'The Hamburger Olympics') to US$894 million for the 2008 Olympics. The escalating price for the TV rights to the Olympics in the USA can be seen as a response to the somewhat belated recognition by the increasingly hard-nosed rights sellers, the International Olympic Committee (IOC), of the global commercial value of their sporting 'product,' paralleled by the heightened competition between media corporations for the Olympic 'property' that they control.

It is useful at this point to go beyond these rather abstract and almost surreal facts, and to look more closely at the types of organizations and individuals who have the power—and bear the risk and expense—of managing such media sports spectacles on behalf of the viewing world. NBC's huge long-term 'futures' investment package (incorporating all Summer and Winter Games until 2008) was clearly a preemptive strategy to prevent other media corporations from acquiring their prime TV sports assets, especially the Fox network controlled by the world's most powerful contemporary media sports proprietor, Rupert Murdoch. However, there is much more to TV sports than intermittent global sports spectacles like the Olympics. Large local, regional, national and (with the growth of satellite pay TV sports) international audiences can be attracted by regular, 'long-form' sports competitions within single countries. It is for this reason that, for example, in 1998 TV networks in the USA paid over US$15 billion for the eight-year broadcast rights to American football.

Yet free-to-air television sports is, despite its demonstrable popularity, also being supplemented (and in some cases supplanted) by pay television, with media entrepreneurs seeing live sports as (in Murdoch's words) a 'battering ram' which, by achieving 'exclusivity' (that is, by being 'siphoned off' from free-to-air television) lures viewers first to subscribe to pay television and then to purchase its associated interactive services (like home banking and shopping, and telephony). Notably, for example, it was the securing in 1992 by the Rupert Murdoch-controlled BSkyB satellite service of the broadcast rights to Premier League soccer in Britain that turned a loss-making company into one which, by 1996, had profits of £311 million (Goodwin 1998).

This convergence of sports and business interests is most advanced in the ownership by media corporations of sports clubs—these include the European soccer teams A.C. Milan and Paris St Germain by Silvio Berlusconi and Canal Plus respectively, and in the USA, the LA Lakers basketball and Dodgers baseball teams by Murdoch.

However, in 1999 a (perhaps temporary) setback to the literal 'takeover' of sports by the media occurred when the British Monopolies and Mergers Commission blocked a £625 million bid by Murdoch for one of the world's most famous sports 'brands,' Manchester United Football Club, with the Industry and Trade Secretary announcing that the decision was made 'mainly on competition grounds' in the belief that 'the merger would adversely affect competition between broadcasters.' Economic concerns, however, did not entirely hold sway, with a 'public interest' argument advanced that 'the merger would damage the quality of British football … by reinforcing the trend towards growing inequalities between the larger, richer clubs and the smaller, poorer ones.' Such social-democratic resistance to the extension of the power of media and of other commercial corporations over sports has not been restricted to the state—in fact, political reluctance to approve such mergers and acquisitions may be prompted by the grass-roots mobilization of sports fans (in this case the Independent Manchester United Supporters Association).

This instance demonstrates how the complete 'annexation' of sports by the media can be resisted on two

grounds—the prevention of oligopoly in the media industry and also the preservation of sports organizations likely to be destroyed by the unfettered application of commercial logic founded on the concentration of media capital on a small number of elite sports clubs. The latter is demonstrated by current resistance to the mooted replacement of nationally based soccer competitions by an inter- (or, more accurately, trans-) national European soccer 'Super League.' It further demonstrates how, despite the unquestionable contribution of the political economic approach to the social analysis of mass media and sports, it is important to acquire specific knowledge of the popular dynamics of sports power without which there would be no valuable cultural phenomenon for the mass media to covet in the first place.

3.2 Culturalism

In addressing the extent to which the commercial media may have taken over and substantially blighted sports, it is necessary not to be transfixed by imposing numbers—both of spectators and of the costs of gaining access to them through TV. It also means being skeptical of the 'stock' positions that spectating is inferior to participating, and that television spectatorship and other types of media sports use are necessarily inferior to attending sports events in real space and time. Culturalist approaches are, somewhat belatedly, attending to media sports as a significant aspect of the fabric and politics of everyday life in seeking to divine the meanings and uses of media sports, the identities it helps to construct and reinforce, and the social faultlines and points of resistance and incorporation that it lays bare. For example, the somewhat anomalous celebration of African American masculinity represented by global sports superstars like (the now retired) Michael Jordan (Baker and Boyd 1997); the role of media sports in making and remaking diverse images from 'desirable' bodies to 'champion' nations (see various chapters in Martin and Miller 1999); and the media sports (and sports-related) scandals involving such figures as Ben Johnson, O. J. Simpson, Tonya Harding and Nancy Kerrigan, and organizations like the IOC and soccer's governing body FIFA, have all promoted widespread debate about such important social issues as drug taking, racial and gender inequality, violence and institutional corruption (Rowe 1995).

Culturalist analyses of media sports, then, without necessarily downplaying the impact of class relationships or economic forces, are (through interdisciplinary combinations of elements of ethnography, socio-linguistics, semiotics, and so on) more open than orthodox political economy to the complexities, contingencies and uncertainties of 'common cultural' experience. For example, sport is, given the continuing development of global media sports spectacles and infrastructures, a highly suggestive 'test case' for globalization theory. The extent to which global sports may be emerging, or the ways in which local or indigenous sports may survive or adapt to the changes precipitated by commercial media sports, are important concerns. The failure of predominantly American sports like baseball and gridiron to become truly global, despite heavy promotion by media sports corporations, for instance, or the corresponding slowness of the 'world game,' soccer, to consolidate itself commercially at the highest level in the USA despite their hosting the massive media spectacle of the World Cup in 1994, demonstrate how social and cultural groups do not simply 'buy'—and may even actively resist—the media sports made commercially available for their consumption. Important to the dispersion and establishment of sports around the world are the specific histories of the relationships between different countries. For example, international cricket developed in the Indian sub-continent, Australasia and the Caribbean according to the southerly trajectory of the British Empire. Long after independence was gained, cricket has remained, with one sporting 'country' —the West Indies—rather eccentrically consisting of sovereign nations which come together under the same banner solely and explicitly for the purpose of playing cricket.

This point raises significant cultural questions concerning the 'on the ground' relationships between media sports texts and audiences. Instead of seeing sports aficionados as inevitably victimized and exploited, greater allowance is made for how audiences might actively experience and negotiate their encounters with the sports media (see the final section of Wenner 1998). For example, the pleasures that women may take in viewing male sports can be viewed from this perspective as a legitimate leisure choice, rather than as an inevitable sign of patriarchal domination. Similarly, this perspective resists the *a priori* assumption that the media sports gaze of fans is inevitably controlled by capitalist interests for their own acquisitive ends. There is a danger, however, that such 'discoveries' of empowered, active media sports audiences may exaggerate the degrees of freedom of choice of the socially disadvantaged and act as a smokescreen for persistent social inequality. For this reason, it is necessary for continuing social scientific scrutiny both of the density of representations of human subjects in mass media sports (who is most visible) and of their differential social quality (how are they depicted). A striking illustration of this point is the conventional neglect of women's sports in the print and electronic media, such that it is often less than 5 percent of total sports coverage, and is compounded further for 'women of color' (Creedon 1994), while the sexualization of the bodies of sportswomen is often comparable to that of soft pornography (Hargreaves 1994). Hence the role of mass media sports in the stereotyping, marginalization and disempowerment of

subaltern groups can be assessed through a judicious combination of quantitative (including content analysis and audience statistics appraisal) and qualitative method (such as ethnographic studies of media sports audiences and textual analysis of media sports imagery). These forms of research and analysis in the culturalist tradition, when informed by a political economic perspective, present the most promising current lines of inquiry into mass media sports.

4. *Futures for Mass Media Sports and its Research*

There is every indication that sports will continue to develop as a major source of content in the mass media, although, just as it has been suggested that the age of broadcasting is over, it might be argued that the emergence of narrowcasting and interactive media technologies will have a great impact on the representation and use of sports in the media, especially in regard to free-to-air television. There is a requirement, then, to assess industry claims that sports fans can now (following the necessary investment of funds) construct 'home stadia' where they take on the traditional role of TV sports producers with hyperlinked access to a vast, instantly recoverable sports database. This image of the all-powerful mass media sports consumer-turned-producer is quite at variance to that of the TV sports audience as a commodity so valuable that (along with sport itself) it has been first besieged and then conquered by the media. Both polarized images need to be regarded with due skepticism. The processes of industrialization, capital accumulation and 'mediatization' that have transformed sports over the last century, therefore, must be carefully analyzed in the light of continuing contestation over the asserted right to use the sports media without being entirely used by them.

See also: Advertising: General; Entertainment; Leisure and Cultural Consumption; Media Events; Media, Uses of; Ritual; Sport, Sociology of; Sports, Economics of; Television: General

Bibliography

Baker A, Boyd T (eds.) (1997) *Sports, Media, and the Politics of Identity*. Indiana University Press, Bloomington, IN

Barnett S 1990 *Games and Sets: The Changing Face of Sport on Television*. British Film Institute, London

Buscombe E (ed.) 1975 *Football on Television*. British Film Institute, London

Chandler J M 1988 *Television and National Sport: The United States and Britain*. University of Illinois Press, Urbana, IL

Creedon P J (ed.) 1994 *Women, Media and Sport: Challenging Gender Values*. Sage, Thousand Oaks, CA

Duke V, and Crolley L 1996 *Football, Nationality and the State*. Addison Wesley Longman, Harlow, UK

Elias N, Dunning E 1986 *Quest for Excitement: Sport and Leisure in the Civilizing Process*. Blackwell, Oxford, UK

Goldlust J 1987 *Playing for Keeps: Sport, the Media and Society*. Longman Cheshire, Melbourne, Australia

Goodwin P 1998 *Television under the Tories: Broadcasting Policy 1979–1987*. British Film Institute, London

Hargreaves J 1994 *Sporting Females: Critical Issues in the History and Sociology of Women's Sports*. Routledge, London

Hill C R 1992 *Olympic Politics*. Manchester University Press, Manchester, UK

Klatell D A, Marcus N 1988 *Sports for Sale: Television, Money and the Fans*. Oxford University Press, New York

Larson J F, Park H S 1993 *Global Television and the Politics of the Seoul Olympics*. Westview Press, Boulder, CO

Martin R, Miller T (eds.) 1999 *Sportcult*. University of Minnesota Press, Minneapolis, MN

Real M R 1989 *Super Media: A Cultural Studies Approach*. Sage, Newbury Park, CA

Rowe D 1995 *Popular Cultures: Rock Music, Sport and the Politics of Pleasure*. Sage, London

Rowe D 1999 *Sport, Culture and the Media: The Unruly Trinity*. Open University Press, Buckingham, UK

Wenner L A (ed.) 1989 *Media, Sports, and Society*. Sage, Newbury Park, CA

Wenner L A (ed.) 1998 *MediaSport: Cultural Sensibilities and Sport in the Media Age*. Routledge, New York

Whannel G 1992 *Fields in Vision: Television Sport and Cultural Transformation*. Routledge, London

D. Rowe

Mass Media: Introduction and Schools of Thought

The term 'the mass media' refers to institutions that via technologies of largely one-way communication, reach audiences that consist of large numbers of people. The book can be seen as the first mass medium and consequently it could be argued that mass media emerged with the development of printing as a medium at the end of the fifteenth century. However, the concept of the mass media is normally associated with the development of modern, industrial society in the West, beginning in the early nineteenth century. In this period, newspapers joined the book as an important mass medium; later in the century, the telegraph became an important point-to-point media technology that had important consequences for the mass media via its conjunction with the press. In the twentieth century, film, radio, phonograms, and then television and video were added. The growth of the media has of course always been conditioned by prevailing historical circumstances (Winston 1998) and how they are to be best used remains at the core of ongoing media policy issues. The spread of the mass media has proceeded in an accelerating manner and today they constituent a central feature of modern society. From the micro-level of individuals in everyday life to the

macro-level of societal institutions and processes, including the global setting, modernity is inexorably intertwined with the development of the mass media (Thompson 1995).

1. The Field

While research about the mass media can be traced to the early decades of the twentieth century, mass media studies first arose as an academic field during the 1950s, establishing itself institutionally in the following decade, first in the US and then elsewhere. Emerging from several disciplines in the social sciences and humanities, the field launched its own academic departments, journals, associations, and conferences. In the US today, most colleges and universities have departments offering courses and programs about the mass media, a pattern that is increasingly manifested in other countries as well. The field today is broad and eclectic, reflecting both the influences of its earlier intellectual origins, as well as the newer developments in the media's evolution and changing socio-cultural significance.

Mass media studies manifest a heterogeneous collection of theories, methods, and general intellectual profiles (see McQuail 1994 for a useful overview; Boyd-Barrett and Newbold (1995) provide an extensive collection of key texts in the field). Academic departments go under a variety of names, emphasizing different angles on the field, e.g., in the US, many departments have their origins in the older schools of journalism, others developed out of speech communication and rhetoric, others from sociology or literature, and so forth. Programs in journalism studies and training are often coordinated with more general mass media studies, while film history and film studies more generally often tend to maintain a separate institutional identity. Particularly in the UK, the field of mass media at universities is often linked with cultural studies.

1.1 Nomenclature

The labeling of the field varies to some extent. In some academic departments, the dominant thematic is 'communication,' and mass media studies, treated as 'mass communication' may be juxtaposed to such areas as interpersonal communication, organizational communication, and intercultural communication. In the last decade or so, particularly in the light of newer developments in digital media technology that make possible more interactive modes of communication, there is a tendency to drop the adjective 'mass' and simply refer to media studies. This inclination is also prompted by changes in social theory: the term 'mass media' is seen as echoing older theories of 'mass society,' a perspective that today appears out of step with the complexities of contemporary social realities.

1.2 Two Basic Orientations

As a field of teaching and research, mass media studies tends to encompass two basic orientations. On the one hand is academic research and a critical liberal arts perspective on the media and their relationship to society and culture, often aimed at answering scientific questions about their institutional operations, their output, and their audiences. On the other hand one finds an emphasis on applied knowledge. This can be geared to the many operational questions facing media industries, to various organizational actors that make use of the media for their own purposes (corporations, political parties, interest groups, etc.), and to occupational practices within the media. Such occupational practices include production skills within the different media, as well as professional training, in particular for journalism. Other occupational orientations include public relations, advertising, and media management. The lines between the academic and the applied orientations are by no means fixed; the interplay between the two can be both productive as well as problematic. Outside the academy there is considerable media research done from commercial horizons, for example market research, audience measurement, and studies on the effects of advertising. Also, political actors have increasingly been making use of media consultants for strategic purposes.

1.3 Organizing Logics

While mass media studies is a rather sprawling field, there are several organizing logics that give it some structure and coherence. These logics are not mutually exclusive, nor do they have equivalent status. However, they can serve as partial maps of the research terrain. The first such logic derives simply from specifying a particular medium. Many people in the field will identify themselves for example as researchers of television, radio, or the press. Others will specialize still further and focus, for example, on television history or the television industry. Many research studies, however, may encompass more than one medium, for example in a case of the analysis of news coverage.

A second logic is to emphasize the different elements of the communication process: a basic model of communication will often use the elements 'sender,' 'message,' and 'receiver.' Translated into the context of the mass media, this yields, respectively, studies of media institutions, media output, and media audiences. Research on media institutions examines the conditions, both internal and external, that shape the way

they function. This can encompass for instance questions about regulation, political influence, and organizational routines. The study of media output includes the form, content, and modes of representation. Often specific genres, particularly television genres such as soap operas and talk shows, become the object of sustained research attention. Studies of media audiences can include analyses of the impact of specific output, the uses which people make of it, and the interpretations they derive from it.

Third, specific topics or themes provide ongoing structure to the field. These themes can be broad, such as media effects, or (more specifically) media effects on children. Many themes link media studies with particular topics on the wider social research agenda, such as gender and ethnicity; health communication, media and social movements, and media and sports, are also examples of such specialization. Some themes derive instead from particular theoretical and/or methodological developments, for example, agenda setting, media events, or media uses and gratification, and can be applied to a variety of empirical areas.

A fourth organizing logic derives from the still-remaining links between mass media and the 'parent' disciplines; such links serve to focus research (and connect researchers) from particular disciplinary perspectives. For example, political communication is a subfield that interfaces with political science, often incorporating public opinion studies; the psychology of communication examines cognitive processes in relation to mass media. Often, however, within the overall field of mass media studies, intellectual lineage is not made explicit, given that the field has developed a distinct identity over the course of several decades.

Finally, weaving through the above logics are sets of intellectual traditions that very loosely structure the field into schools of thought (see below). It should be emphasized that 'school' is used loosely here; it points to tendencies within the field, not unified factions with clear demarcations. These schools basically reflect differing theoretical and methodological orientations; they point to general intellectual dispositions. The relationships among them have evolved over time, and the development of the field can be understood to a considerable extent by looking at the emergence, interface, and transitions of these schools of thought.

While the exact labeling may be differ in historical accounts of the field, the following constitute the main schools of thought within mass media research: the mainstream perspective, the critical tradition, the culturalist approach, and the policy horizon.

2. *The Mainstream Perspective*

The systematic study of the mass media arose in the early decades of the twentieth century, and the mainstream perspective today can be understood as the intellectual descendants of the pre-World War II pioneers. While the mainstream perspective encompasses a wide array of research orientations and is continually developing new approaches, it is united in its basic adherence to traditional social scientific methodologies and approaches to theory. Its intellectual underpinnings are found chiefly within sociology, psychology, and social psychology, and it constitutes the still-dominant tradition of empirical research in mass communication.

2.1 *Origins*

Its origins can be found in nineteenth century sociological conceptions about the crowd and the mass; early research efforts and attempts to develop theory were geared toward among other things the theme of propaganda, with the experiences of World War I as a point of reference. In the 1920s, the Chicago School of sociology, with its German-trained prominent member Robert E. Park, developed a strongly qualitative, ethnographic micro-sociology of mass media uses and impact, within a generally progressive and reformist social vision. Walter Lippman's (1922) theories about public opinion were an important element in conceptually linking the mass media with the processes of democracy.

The political scientist Harold Lasswell, in the 1930s under President Roosevelt's New Deal, saw new opportunities for studying public opinion and propaganda. His formula of 'who says what to whom through what channel and with what effects' served as a framework for media studies for many years. In psychology, Kurt Lewin's psychology of personality and of group dynamics foreshadowed analyses of the social processes of media influence. After the war, Carl Hovland and his associates (1953) further developed behaviorist perspectives on persuasion.

Interest in the media grew in step with their expanding role. The imperatives of the wars, the importance of the media in political processes, concern about the media's impact on psychology, emotions, and behavior, as well as increasing commercial concerns of the media themselves stimulated research. After World War II, functionalist sociological theory had become the dominant paradigm for fledgling media research, backed up by developments in methodological techniques. Quantitative content analysis, aimed at mapping the manifest dimensions of media messages, became the norm for studying output. Studies of audiences' media preferences, and, most importantly, effects of the media on audiences, began to blossom. Surveys became an established approach; later the strategy was developed to repeat surveys with the same respondents to yield panel studies over time (see Rosengren and Windahl 1989).

A key person in the USA at this time was the Austrian-born Paul Lazarsfeld, who was active in

ushering in a new phase of mass-media research. Lazarsfeld began working for the commercial media in the 1930s, and he viewed the media research tradition he helped to develop as 'administrative' research. This was geared to serve the strategic purposes of media institutions. Gradually, a more society-wide perspective emerged, that helped shape a media-research orientation that was in harmony with the consensus assumptions about society. Together with Elihu Katz (Katz and Lazarsfeld 1955), Lazarsfeld helped introduce the notion of how the impact of the media is filtered through social interaction with others, giving rise to the concept of the two-step flow of communication. Wilbur Schramm (1970), in the 1950s and 1960s, did much to synthesize the diverse elements and promote the identity of mass media research as a distinct field of inquiry.

2.2 The Golden Age: Effects Research and Content Analysis

From about 1950 to the early 1970s, the mainstream perspective emerged, solidified, and remained essentially unchallenged. This was its golden age; since then the heterogeneity of the field has increased. During the 1950s, concern with the effects of the media continued to grow in the USA, especially in the light of television's reach into just about every home. The questions around media effects were not merely scientific: there was popular concern about violence and the media and about portrayals of sexuality, and, later on, about pornography. The media industries landed in a seemingly paradoxical position of trying on the one hand to soothe popular opinion by asserting that such negative effects were minimal, while on the other hand assuring the advertising industry of the effectiveness of media messages. There was also a malaise about 'media manipulation' in the realm of politics, as the question of just how much the media shape attitudes and opinions took on increasing concern. A landmark book in this context was Joseph Klapper's (1960) *The Effects of Mass Communication*, which argued for a 'minimal effects' position.

Within the media effects research tradition, the nature of the 'effects' investigated can vary considerably: on beliefs, attitudes, knowledge, or behavior; intentional or unintentional; long-term or short-term; and so forth. In the USA, the Surgeon General's massive, multivolume report (1972) was perceived to be inconclusive by many, but what it did emphasize was the importance of so-called intervening variables. That is to say, the effects of the mass media must be understood as being qualified by such factors as age, gender, education, social circumstances, audience contexts, and so forth. Generally the drift in media effects research has been away from the earlier psychologism toward increasing socio-cultural complexity. Concern with the effects of mass media continues today, even if the task of identifying and isolating specific media output and linking it to particular effects becomes all the more difficult with the increasing mediatization of society.

The mainstream perspective has been no less interested in the nature of media output. Journalistic output is one of the major areas in which content analysis has been applied, where bias in reporting has been studied by measuring, for example, the visibility of opposing candidates during election campaigns, and coding degrees of positive and negative valence. More encompassing studies have been done where researchers have attempted to elucidate important dimensions of journalistic coverage or even in fictional portrayals that are normal or routine, yet may remain invisible precisely because of their everyday, taken-for-granted character. Measurements have been developed, for example, for tallying acts of violence in television programs. The ambitious Cultural Indicators project of George Gerbner and his colleagues (1973) sought to illuminate the symbolic world of TV fiction in the USA, mapping not only violence, but the more general norms and values manifested in the programs.

Within the tradition of so-called agenda-setting research (see Dearing and Rogers 1996 for an overview), news and current affairs in the mass media are studied and then compared to audiences' understandings of, for instance, what are the major issues of the day. The comparison allows researchers to draw conclusions about the extent to which the media set the political agenda; generally the findings are that the media do have this initiative, but it has been found that audiences, or rather the public, do not always simply mirror the concerns of media coverage.

2.3 Tensions in the Field

The mainstream perspective continues to develop, partly in response to new phenomena in the media, e.g., we witness today considerable attention to the transnational character of the media and a growing research concern with political advertising (Bennett 1996). However, innovation in theory also generates renewal. Thus, in the 1970s the uses and gratifications tradition (Blumler and Katz 1974) arose and posed a set of complementary—if not fully competitive—questions to the effects tradition. Basically, instead of asking what the media do to people, the newer orientation asked what people do with the media and what satisfactions do they thus derive. This shift points to an ongoing key tension within the field, namely the question of the relative power of the media and the degree of autonomy of the audience. The view of minimal effects, in keeping with the intellectual origins of functionalism, liberal democracy, and the needs of commercial media had strong support but did

not go unchallenged. Indeed, critics began claiming that the theories and methods embedded in the mainstream perspective structured a status-quo perspective toward media and society (Gitlin 1978). In 1983 and again in 1993, the leading US journal in the field, *The Journal of Communication*, had themed issues on the tensions within the field (Gerbner 1983, Levy and Gurevitch 1993).

It should also be noted that the prevailing conception of the process of communication at work within the mainstream perspective, operating for the most part implicitly, was that of the transfer of messages or, more generally, of information. This view has its origins in, among other things, Shannon and Weaver's (1949) cybernetic view of communication. Gradually this basic model was augmented by contributions from cognitive psychology, giving rise to more elaborate notions of human information processing. Yet, other voices, with an anchoring in the humanities and cultural theory, were challenging such views, arguing for the centrality of meaning for the process of communication. Both the critical tradition and then the culturalist approach began to coalesce and to offer competing paradigms of mass media research.

3. *The Critical Tradition*

During the late 1960s, critical traditions from various strands of neo-Marxism were beginning to make themselves felt within the social sciences, and by the early 1970s they were manifesting themselves in mass media studies. There are several intertwining intellectual strands here: political economy, the critique of ideology and public sphere theory.

3.1 *Political Economy*

Though the political economy of mass media need not *a priori* be based in the Marxian tradition, within media studies the term does most often connote this lineage rather than, say, neoclassical political economy or conventional economics. The emphasis in Marxian political economy of the media is often on the links between economics and the social, political, and cultural dimensions of modern life; pluralist or consensus models of society are generally rejected as inaccurate. A recurring thematic is the tension between on the one hand, the capitalist logic of media development and operations, and on the other, concerns for the public interest and democracy. The political economy of the media does not anticipate the elimination of commercial imperatives or market forces, but rather seeks to promote an understanding of where and how regulatory initiatives can establish optimal balances between private interest and the public good. The work of the Canadian scholar Dallas Smythe (1977) was an important element in the development of this tradition; among contemporary researchers Nicholas Garnham (1990) has further expended the perspective, which has in turn been systematized in Mosco (1996). Within this perspective, the emphasis is most often on the institutional arrangements of the media: ownership, control and regulation, and their consequences. The political economy of media structures at the global level has increasingly come under research scrutiny (e.g., Herman and McChesney 1996).

A variant of the political economy tradition has focused on media imperialism, in particular how the combination of US corporate and political interests over the years have made use of the media abroad and the consequences that this has had for people in other countries, particularly the developing nations. The work of Herbert Schiller (1976) has been central here. Media flows from the West to 'the Rest' were plotted (Nordenstreng and Varis 1974), including fiction, other entertainment, journalism, and also the role of international advertising (Mattelart et al. 1984). Unesco became a political forum and research promoter for a New International World Information Order (1980). Such research was framed by historical circumstance: the bipolar East–West situation of the Cold War shaped US media strategies abroad in ways that had impact on the North–South global divide. Since the collapse of communism, the North–South socio–economic gap has continued to grow, and though the transnationalization of the mass media is now often framed analytically within paradigms of globalization, critical attention is still addressed to global media disparities.

3.2 *The Critique of Ideology*

Another strand of critical endeavor attends in more detailed ways to media representations, analyzing them from the standpoint of their ideological dimensions. The intellectual roots for the critique of ideology derive not only from Marx, but also from the structural neo-Marxism of Althusser, as well as the more culturally oriented versions of the Frankfurt School's Critical Theory and of Gramsci. Methodologies and epistemologies were at times in conflict, not only between the mainstream perspective and the critical tradition, but also within the critical tradition. Versions both of cultural Marxism and of structural Marxism—the latter making use of Freudian theory adapted via Lacan—became incorporated into the development of cultural studies (see below). A landmark effort in the critique of ideology in media studies was the work of the Glasgow Media Group (1976), which made explicit use of traditional social science methodologies of content analysis to make its points

about the class bias of the British Broadcasting Corporation's journalism.

Ideology as such is a slippery concept, and the earlier assumptions about false consciousness could not hold up to critical scrutiny and increasingly sophisticated epistemology. Gradually, the critique of ideology began to make use of the hermeneutic tradition, treating ideology as an aspect of the interpretation of meaning in the reproduction of relations of social domination. The critique of ideology began to blend with cultural theory, signaling a growing entwinement of the critical and culturalist schools. Efforts to analyze cultural hegemony, in both its domestic and transnational manifestations, strove to relate media representations and the political economy of the media to specific social circumstances. Also, the growing presence of feminism within the critical tradition began to render the exclusive concern with social class as the foundation of ideology untenable: ideology could manifest gendered social relation, not least in the media. Gradually, feminist perspectives began to take hold in media research (van Zoonen 1994). Also, race and ethnicity also became foundations for the critique of ideology in the mass media (Gandy 1998).

3.3 Public Sphere Theory

Discussions of democracy and the media are increasingly framed within the concept of the public sphere, a concept associated with Jürgen Habermas (1989, see also Calhoun 1992). In schematic terms, a public sphere is understood as a constellation of institutional spaces that permit the circulation of information and ideas. These spaces, in which the mass media figure prominently, also serve to foster the development and expression of political views among citizens as well as to facilitate communicative links between citizens and the power holders of society. Beginning in the 1970s, media researchers in Europe and later, the USA, began using the concept to pursue critical analyses of mass media, examining the factors that impede the public sphere. The concept has many parallels with the liberal notion of the 'marketplace of ideas' and similar metaphors, and today it has entered into more mainstream usage, where the problems of journalism are discussed and researched; its Frankfurt School origins often remain in the background.

4. The Culturalist Approach

Within mass media research, contributions from the humanities and from qualitative, interpretive social science had long remained in the shadow of the dominant social scientific methods. Studies based in literature, history, rhetoric, or anthropology were very much the exception. Marshall McLuhan's (1964) work, with a historical view of media technology, was an exception, but it had more impact outside the university than within academic media research.

4.1 Cultural Studies

In the 1960s British Cultural Studies was in its formative years, and in the following decade, under the leadership of Stuart Hall (Hall et al. 1980) had begun to impact on media research (Turner 1996). The eclectic synthesis of intellectual ingredients in cultural studies—including neo-Marxism, structuralism, feminism, psychoanalysis, semiotics, and ethnography—was applied not only to the media but a range of topics having to do with contemporary culture and society. In particular, popular culture has been an ever-present topic of research and debate. Cultural studies has among other things shifted the focus of issues around popular culture (and/versus high culture) from earlier largely aesthetic concerns to questions about social relations and power; taste becomes socially contextualized. There has also been a 'rehabilitation' of popular culture, where 'the popular' is seen as no longer so clearly demarcated from high culture or, alternatively, is defined in a manner to give it political relevance (Fiske 1989). Other research takes the ways in which the media, via advertising and general lifestyle imagery, are central to contemporary consumer culture, not least on the global level (Featherstone 1995).

Cultural studies has grown into a heterogeneous, multidisciplinary field in its own right, with contributions from currents such as postmodernism and postcolonialism, among others. Today studies of the media are only a small part of its vast concerns. Cultural studies can be seen as part of a larger 'cultural turn' within the human sciences in recent decades, but today it is clearly the dominant culturalist tendency, even if there is a lack of consensus about both its core and its boundaries. As cultural studies has expanded and become a global academic phenomenon mingling with many other academic traditions, the critical character of its earlier years, where issues of social and semiotic power were thematized, has not always remained evident. For example, the emphasis on the multiplicity of interpretations of popular culture—at times augmented by postmodern perspectives—has tended to undercut arguments about ideology.

4.2 The Centrality of Meaning

In culturalist mass media studies two themes from the humanities and qualitatively oriented social sciences

have come to assume central positions: sense-making and social constructionism. Research has come to deal with questions of how media representations convey meaning, how media audiences make sense of these representations, and how media-derived meanings further circulate in society. Hall's (1980) model of how meaning is encoded by the media and decoded by audiences became an important organizing concept for much of such work. Studies of reception—how audiences interpret what they encounter in the media—became an important subspecialty within mass media research; the work of Dave Morley (1992) and Ien Ang (1985) are significant contributions in this tradition. Generally, as the media landscape becomes more multidimensional and theorizing about the audience becomes more ambitious, the frontlines of audience research manifest both a convergence of different research approaches and a growing complexity (e.g., Hay et al. 1996)

If semiotics and hermeneutics were the major text-analytic tradition within cultural studies, today they have become augmented and partly replaced by innovations from various other currents such as linguistics and cognitive psychology. These newer efforts have coalesced to generate several versions of what is now termed discourse analysis (van Dijk 1997) that is increasingly used in the elucidation of meaning in media texts.

4.3 Active Subjects

In culturalist approaches to the mass media, the emphasis on the production of meaning and social constructionism tends to privilege versions of the active subject. This perspective is manifest not least in a strand of research that has underscored the audience's sense-making, though debates have emerged (again) as to the degree of semiotic power that should be accorded the audience relative to the media (Curran 1990). Research in some quarters has linked audiences' interpretation of the media as forms of cultural resistance, and tried to delineate interpretations that run counter to hegemonic trends; Radway's (1984) study of women's uses of romance novels for a variety of purposes is a compelling example. More generally, theories of subjectivity have been mobilized to study how individuals construct identities via their active appropriation of media; feminist contributions have been particularly visible here. From the standpoint of collectivities, mass media and cultural identity has become a growing domain of research, not least in transnational contexts (Morley and Robins 1995).

5. The Policy Horizon

It could be argued that the policy horizon is not a school of thought, but rather a thematic organizing logic encompassing diverse perspectives. However, the centrality of policy in the shaping of media development and the distinctiveness of its concerns speak for its status as a school.

5.1 Media Systems and Norms

Situated in the force fields between economics, politics, and technology, the way the media are owned, organized, and financed, and the way that they operate have much to do with the vicissitudes of the policy process. Policy is of course shaped by the specific interests and actors involved, such as the state, commercial media institutions, the advertising industry, media production organizations, citizens groups, and other representatives of the public interest. One line of policy research has addressed the historically anchored traditions in political culture and in philosophies of the public good that buttress prevailing normative frameworks in regard to the media. These frameworks are generally accepted but often given competing interpretations. Freedom of the press, for example, has had a particularly strong status in Western liberal democracies, though how it is to be best applied in a concrete case may be contested. Historically, and today globally, there are competing models for media systems, including authoritarian, communist, and social responsibility versions that build upon differing sets of norms.

5.2 Regulation and Deregulation

Another line of research has been more engaged at the frontlines of ongoing policy issues and battles, and often seeks explicitly to intervene in the policy process. The growing concentration and conglomerization of media ownership has been charted by liberal critics in the USA (Bagdikian 1997), and elsewhere; generally much of the political economy perspective has fed into policy research. Over the past two decades, the increasing commercialization of media industries, coupled with a political climate shaped by neo-liberalism, has contributed to a massive deregulation of the media, giving market forces ever greater free play and repositioning audiences less as citizens or as a public and more as consumers. This has been especially true in the realm of broadcasting regulation, and in Western Europe public broadcasting has undergone a radical realignment as it has lost its monopoly position (Tracey 1998). For over two decades the Euromedia Research Group (1998) has studied these developments. Confronted by commercial channels borne first by cable and satellite systems, and then terrestrial stations, public broadcasters in Europe and elsewhere find that their *raison d'être* has to be redefined as they look to the future. Public broadcasting emerged as a part of the project of the nation-state, and its current circumstances can be seen as part of the dilemmas facing the nation-state in the era of globalization. At

the transnational level, the regulation of international communication becomes all the more important—and complex.

The global media landscape today is dominated by fewer than two dozen conglomerate giants who increasingly control the production of media output and the distribution process, as well as the technical infrastructure. Technical advancements promote a convergence between the technologies of broadcasting, telecommunications, and the computer. Institutionally, major companies in the computer and Internet field have begun to merge with the giants of the traditional mass media. This evokes profound policy issues of diversity, accountability, and democracy (McChesney 1999).

6. *Looking Ahead*

The media are obviously in transition, and research is in the process of adapting itself to the emerging realities. The increasing on-line presence of the traditional media, as well as other forms of functional integration technical convergence, suggest that the field of mass media research will in particular increasingly take on a digital focus. There will be growing attention to such areas as the information society, computers and society, human–computer interface, and electronic democracy. Digitalization has prompted a flurry of research from many disciplines, including social theory (Castells 2000) and thus it can be argued that the field of mass media research will increasingly blend with initiatives from other directions. This may involve some erosion of its status as a separate field, but the enhanced attention being paid to the media will no doubt increase our knowledge and understanding of them.

See also: Agenda-setting; Broadcasting: Regulation; Communication and Democracy; Communication, Two-step Flow of; Cultural Studies: Cultural Concerns; Mass Communication: Empirical Research; Mass Media and Cultural Identity; Mass Media, Political Economy of; Mass Media, Representations in; Media and History: Cultural Concerns; Media Effects; Media Effects on Children; Media Events; Media, Uses of; Printing as a Medium; Radio as Medium; Television: General; Television: History; Violence and Media

Bibliography

Ang I 1985 *Watching 'Dallas'*. Methuen, London

Bagdikian B 1997 *The Media Monopoly*. Beacon Press, Boston

Bennett L 1996 *The Governing Crisis: Media, Money, and Marketing in American Elections*, 2nd edn. St. Martins, New York

Blumler J G, Katz E (eds.) 1974 *The Uses of Mass Communication*. Sage, Beverly Hills, CA

Boyd-Barrett O, Newbold C (eds.) 1995 *Approaches to Media*. Arnold, London

Calhoun C (ed.) 1992 *Habermas and the Public Sphere*. MIT Press, Cambridge, MA

Castells M 2000 *The Rise of the Network Society*, 2nd edn. Blackwell, London

Curran J 1990 The new revisionism in mass communication research: A reappraisal. *European Journal of Communication* **5**: 135–64

Dearing J W, Rogers E 1996 *Agenda-setting*. Sage, London

Euromedia Research Group 1998 *Media Policy: Convergence, Concentration and Commerce*. Sage, London

Featherstone M 1995 *Undoing Culture: Globalization, Postmodernity and Identity*. Sage, London

Fiske J 1989 *Understanding Popular Culture*. Unwin, London

Gandy O 1998 *Communication and Race*. Arnold, London

Garnham N 1990 *Capitalism and Communication*. Sage, London

Gerbner G 1973 Cultural indicators—the third voice. In: Gerbner G, Gross L, Melody W (eds.) *Communications Technology and Social Policy*. Wiley, New York, pp. 553–73

Gerbner G (ed.) 1983 Ferment in the field. *Journal of Communication*. **33**

Gitlin T 1978 Media sociology: the dominant paradigm. *Theory and Society* **6**: 205–53

Glasgow Media Group 1976 *Bad News*. Routledge and Kegan Paul, London

Habermas J 1989 *The Structural Transformation of the Public Sphere*. Polity Press, Cambridge, UK

Hall S 1980 Encoding and decoding in the television discourse. In: Hall et al. (eds.) *Culture, Media Language*. Hutchinson, London, pp. 197–208

Hall S, et al (eds.) 1980 *Culture, Media Language*. Hutchinson, London

Hay J, Grossberg L, Wartella E (eds.) 1996 *The Audience and its Landscape*. Westview Press, Boulder, CO

Herman E, McChesney R 1996 *The Global Media*. Cassell, London

Hovland C, Janis I, Kelly H 1953 *Communication and Persuasion*. Yale University Press, New Haven, CT

Katz E, Lazarsfeld P F 1955 *The People's Choice*. Free Press, New York

Klapper J 1960 *The Effects of Mass Communication*. Free Press, New York

Levy M, Gurevitch M 1983 The future of the field. *Journal of Communication* **43**

Lippman W 1922 *Public Opinion*. Harcourt Brace, New York

MacBride S 1989 *Many Voices, One World*. UNESCO, Paris

Mattelart A, Mattelart M, Delcourt X 1984 *International Image Markets*. Comedia, London

McChesney R 2000 *Rich Media, Poor Democracy: Communication Politics in Dubious Times*. University of Illinois Press, Urbana, IL

McLuhan M 1964 *Understanding Media*. McGraw-Hill, New York

McQuail D 1994 *Mass Communication Theory: An Introduction*. Sage, London

Morley D 1992 *Television Audiences and Cultural Studies*. Routledge, London

Morley D, Robins K 1995 *Spaces of Identity: Global Media, Electronic Landscapes and Cultural Boundaries*. Routledge, London

Mosco V 1996 *The Political Economy of Communication*. Sage, London

Nordenstren K, Varis T 1974 *Television Traffic—A One-way Street?* UNESCO, Paris
Radway J 1984 *Reading the Romance*. University of North Carolina Press, Chapel Hill, NC
Rosengren K E, Windahl S 1989 *Media Matter*. Ablex, Norwood, NJ
Schiller H 1976 *Communication and Cultural Domination*. Sharpe, New York
Schramm W (ed.) 1970 *The Process and Effects of Mass Communication*. University of Illinois Press, Urbana, IL
Shannon C, Weaver W (eds.) 1949 *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL
Smythe D 1977 Communications: Blindspot of western Marxism. *Canadian Journal of Political and Social Theory* **1**: 120–7
Surgeon General's Scientific Advisory Committee 1972 *Television and Growing Up: The Impact of Televised Violence*. GPO, Washington, DC
Thompson J 1995 *The Media and Modernity*. Polity Press, Cambridge, UK
Tracey M 1998 *The Decline and Fall of Public Service Broadcasting*. Oxford University Press, New York
Turner G 1996 *British Cultural Studies*, 2nd edn. Routledge, London
van Dijk T 1997 *Discourse Studies: A Multidisciplinary Study*. Sage, London
van Zoonen L 1994 *Feminist Media Theory*. Sage, London
Winston B 1998 *Media, Technology and Society: A History from the Printing Press to the Superhighway*. Routledge, London

P. Dahlgren

Mass Media, Political Economy of

1. *The Birth of Political Economy*

Political economy first emerged as an intellectual project in the eighteenth century as part of the Enlightenment's concerted effort to replace religious and metaphysical accounts of the world with man-made models of order and change. As a central building block in what were later to become the social sciences it set out to take the new machinery of capitalism apart and to trace its consequences for the organization of social and political life. Following the revolutionary upheavals in America and France these links came to centre more and more on the material and cultural conditions required to secure a democratic order in which people were no longer subjects of a monarch or autocratic ruler but autonomous citizens with the right to participate fully in the life of the nation and to elect their political representatives. Envisioning society as a community of citizens lay at the heart of the Enlightenment's emancipatory project of banishing ignorance, fear, and exploitation and promoting the rational pursuit of liberty, equality, and fraternity.

This ideal continues to provide political economy with its moral touchstone positioning many of its practitioners as permanent critics of capitalism's failure to reconcile possessive individualism with the common good. For many economic commentators however, this is an anachronistic stance, a residue of a time before economic analysis detached itself from its roots in moral philosophy and set out to become a modern analytical science. Consequently, to call oneself a political economist today is as much a declaration as a description. It announces a continuing commitment to teasing out the relations between economic dynamics and the virtuous polity and to the Enlightenment project (Garnham 2000). The political economy of mass media occupies a pivotal position in this endeavour by virtue of the communication system's unique triple role as an essential infrastructural support for economic activity, a significant focus of economic activity in its own right, and the major force organizing the cultural resources through which people come to understand the world and their own possibilities for action.

2. *Capitalism, Democracy, and Communications*

It was immediately clear to many early nineteenth century observers that the emerging system of public communications, centered at that time on the daily press, had a pivotal role to play in securing a democratic social order. First, it provided the most effective way of disseminating the information and knowledge citizens needed to understand the problems facing society and how they might be tackled. Second, it offered a communal space in which alternative interpretations of the situation and competing policies could be tested against the available evidence. Third, by overcoming geographical separation it helped construct the nation-state as an imagined political community as well as an administrative apparatus.

At the same time, by the mid nineteenth century it was equally evident that communications systems also played a central role in organizing the emerging industrial forms of capitalism. First, they provided the means whereby firms could track and coordinate activities that were becoming ever more diverse and dispersed. Second, with the growth of mass production, they provided new outlets for the display advertising that was increasingly central to the task of matching supply to demand. Third, they offered enticing opportunities for capitalist enterprise in their own right. As migrants moved into the cities to work in the new factories and offices, the demand for popular reading matter and later for recorded music and cinema, grew exponentially.

The central question these developments posed for political economy was deceptively simple; 'How did

the increasing integration of public communications into a capitalist economy based on private enterprise affect the former's ability to provide the disinterested information, plural viewpoints, open debate and communal culture required by democratic ideals?

3. *A Free Market in Ideas*

For the first half of the nineteenth century most commentators argued that the best guarantee of open democratic debate was a free market in ideas and arguments, in which a plurality of publishers competed for readers with minimum interference from the government. There were good reasons to identify the state as the principal enemy of press freedom. The American colonists had fought a bitter War of Independence to emancipate themselves from British rule, while in England the government had responded to increasing popular unrest by introducing an array of compulsory levies on the press (including taxes on the paper used and the advertisements carried) in an effort to make radical publications economically unviable. The battle to repeal these 'taxes on knowledge' (as they were known) lasted until 1869, when the last remaining provision was abolished. However, it was at exactly this point, when the nominal freedom to enter the newspaper market was extended, that the introduction of more efficient production technologies (improved presses, mechanized composition and new graphic capabilities) raised the price of market entry and began to concentrate press ownership in the hands of wealthy individuals. This development placed a sizeable question mark against the press's ability to fulfill its promises of providing a forum for debate and acting as a watchdog on abuses of power. In their pamphlet of 1846, *The German Ideology*, Karl Marx and Frederick Engels had famously argued that the capitalist class's control of economic resources allowed them to 'regulate the production and distribution of the ideas of their age' and manipulate public culture in their own interests. Witnessing the rise of the new 'press barons' even commentators who rejected revolutionary politics began to wonder if Marx might not be at least partly right. Whereas most of the old style proprietors had owned a single newspaper, which they often edited themselves, the new owners possessed chains of titles and had significant stakes in key geographical markets. Moreover, the increasing importance of advertising seemed, to many observers, to be extending the influence of capital still further. Drives to maximize readership among the urban working class had led to cuts in cover prices and the spread of the 'penny press.' Shortfalls in income were made up by advertising revenues, altering the previous relations between paid publicity and editorial matter. As the English commentator, Thomas Bowles, noted in 1884: 'Newspapers … profess to sell news and give advertisements to boot. What they really do is sell publicity … and give news to boot'.

The emergence of a press system funded primarily by advertising whose ownership was increasingly concentrated in fewer and fewer hands challenged the ideal of public communications in the service of democracy in several ways. First, it raised the possibility that press proprietors might use the titles they owned to promote their own political enthusiasms or business interests while denying space to opponents and critics. Second, dependence on advertising revenue opened newspapers to editorial pressure from large advertisers wishing to promote their activities or censor unfavorable publicity. Third, the growth of advertising content privileged the speech of commerce and squeezed the space available to other voices. Fourth, because advertisers required access to either mass readerships or wealthy niche markets, publications catering to marginalized minorities or less affluent groups became ever more precarious or unviable. Surveying the situation in the United States in 1910, Edward Ross reluctantly concluded that 'the commercial news-medium does not adequately meet the needs of democratic citizenship.' Many of his contemporaries agreed but saw the entrenched identification of a free press with a free market as an immovable barrier to reform. As Delos Wilcox lamented in 1900; 'The American people must dearly love the freedom of the press, or we should have heard before now much talk of government control or operation of the newspaper (Wilcox 1900).'

The growing tensions between the centrality of the communication system as a resource for capitalist development and as a support for democratic culture prompted two reactions. Some analysts set out to develop a political economy of public communication, measuring corporate performance against the requirements of democratic theory. Others, like Alfred Marshall in his *Principles of Economics* (1890), announced the death of political economy and birth of 'economics.'

4. *Political Economy and Economics*

By dropping the adjective 'political,' economists measured their distance from political economy in three ways. First, whereas political economists set out to trace the general consequences of shifts in economic organization for social and cultural life, economists focus on the working of 'the economy' defined as a separate, bounded, sphere. Second, because economists saw economic behavior as the product of a universalistic individualist psychology rather than of socially constructed practices they generally avoided critical reflection on prevailing institutional arrangements and possible alternatives to them. They saw the economy as a set of interlinked markets, not as a

system of power, focusing on the dynamics of exchange rather than the distribution of property. Third, they defined their role in technical terms and detached themselves from political economy's engagement with moral philosophy. They were concerned with efficiency rather than justice, with individuals as consumers rather than citizens. The field of mass media research is still deeply marked by this division.

Media economists are primarily interested in the peculiarities of the communications industries as industries and tend to see regulated markets as the least worst way of organizing them. In contrast, critical political economists of media have tended to focus on market systems' failures and distortions in the cultural sphere and to champion various forms of public intervention. They begin with the possession of core forms of capital (both material and symbolic) rather than with the dynamics of exchange.

5. *Institutionalizing the Public Interest*

In 1808 Thomas Jefferson's Treasury Secretary, Albert Gallatin, argued strongly for federal control of America's developing road and canal system on the grounds that they would benefit everyone by uniting 'the most remote quarters' of the nation, both physically and imaginatively. Later observers saw the same potential in the telegraph and the telephone, arguing that these new media systems were 'natural monopolies' because a single operator was more likely than competing firms to optimize both economic efficiency and social benefits. This argument was widely accepted, but institutionalized in different ways. In the United States the task was assigned to a private company, American Telephone and Telegraph (AT&T), which was charged with providing a universal service that was both geographically comprehensive and priced at a level that made it accessible to all, regulated (from 1934) by a Federal Communications Commission (FCC). In contrast, in Europe the task of providing universal service was assigned to monopoly Post, Telegraph and Telephone organizations (PTTs) owned and operated by the State. This same division between public enterprise and federally regulated private companies also characterized the institutionalization of broadcasting.

The emergence of broadcasting in the early 1920s was profoundly shaped by two contexts. First, the British experience of managing scarce strategic resources, such as timber, in World War 1, had convinced at least some economists that public corporations were the most appropriate way of addressing the limited space in the radio spectrum available for broadcasting. Second, the social unrest that had followed the end of the War, the victory of the Bolsheviks in Russia, and the extension of the franchise to women, led many commentators to see a public monopoly as the best way of securing responsible and rational participation in mass democracy in uncertain and volatile times. Accordingly, after careful deliberation, in 1926 the British government opted to transform the commercial monopoly initially given to the British Broadcasting Company (operated by a cartel of radio manufacturers) into a public service monopoly, the British Broadcasting Corporation, forbidden to take advertising, funded by a compulsory licence fee on set ownership, and charged with providing informational and cultural resources for citizenship, as defined by political and cultural elites. In the United States, in contrast, radio was widely seen from the outset as a business opportunity. Its promise of widespread and instantaneous access to homes made it an attractive advertising medium in a context where a mass consumer system was expanding rapidly. Despite dissenting voices and a fierce rearguard action by educational and other public interest groups, broadcasting became primarily a privately owned, advertising supported system, oriented to maximizing audiences by providing popular entertainment, with the FCC allocating (and revoking) broadcast licenses and instituting backstop regulation to ensure a minimum degree of access to air-time for minority viewpoints.

These contrasted models of broadcasting as public service and commercial enterprise, communication for citizenship and communication for consumption, provided the basic templates for many other countries, though a number eventually settled for a mixed broadcasting economy, with public and commercial sectors running in tandem.

6. *Consolidation and Critique*

Broadcasting and telecommunications were exceptions, however. The key organizations in the other major mass media sectors—the press, cinema, recorded music, and advertising—remained mostly privately owned and minimally regulated, consolidating the potential control that corporations could exercise over the organization of public knowledge and culture. Critical observers responded by reaffirming and extending the main themes of earlier commentaries. Some, like Upton Sinclair (1920) in his attack on the American commercial press, *The Brass Check*, returned to the potential abuses of owner and advertiser power. Others, notably Danielian in his 1939 study, *The AT&T*, turned their attention to the operation of power in conditions of regulated monopoly. In Britain, Klingender and Legg, observing the monopolistic tendencies in the country's cinema industry and the growing power of the Hollywood studios, produced the first detailed political economy of a major mass media sector, *The Money Behind the Screen* (1937). But the most comprehensive contribution to esta-

blishing a general political economy of media came from two German scholars, Theodore Adorno and Max Horkheimer who were strongly influenced by Marx and whose essay, 'The Culture Industry,' (1944, 1973) sketched out a general model of the possible relations between the increasing commodification and industrialization of cultural production and what they saw as the narrowing possiblities for creativity and critical expression in the public sphere.

Most of these early contributions came from commentators who worked either outside the universities or on the edges of academia. It was not until the expansion of the university system in the 1960s and the rapid development of communications and cultural studies as academic specialisms in their own right, that the political economy of the mass media found a secure base within the academy.

7. *Political Economy and Cultural Analysis*

This new institutionalization led to a significant growth in academic research, but tenured faculty did not enjoy a monopoly of insight and commentators working outside the universities continued to make important interventions in debates. Although this renewed interest in the relations between culture and economy ranged widely, it tended to coalesce around several key strands.

First, the rise of multimedia conglomerates with interests spanning a range of public communications sectors breathed new life into long-standing questions about the possible links between patterns of ownership, structures of corporate control, and the range and diversity of cultural production and produced a significant volume of research charting the new configurations. Debates around the implications of these emerging structures also found an audience outside the academy through books like Ben Bagdikian's *The Media Monopoly* (1983).

Second, analysts attempted to refine the general idea of 'the cultural industries' by investigating the contrasting ways in which production processes were organized within major media sectors and tracing their implications for creative labor and for the range and forms of cultural output. The contributions of French scholars, Patrice Flichy (1980) and Bernard Miege and his colleagues (1986), were particularly significant in establishing this line of research.

Third, scholars returned to political economy's central concern with the constitution of complex democracies by way of Jurgen Habermas's influential idea of 'the public sphere' as a space of open debate situated between the State and capital and relatively independent of both. This led a number of critical analysts in Britain and Europe to reply to commentators calling for an enlarged market sector in television with a vigorous defence of public service broadcasting as the cornerstone of a mediated public sphere. The critique of market-driven broadcasting had developed somewhat earlier in the United States and the ideas first floated by Dallas Smythe, a former FCC economist, and other political economists in the 1950s played an important role in the successful lobby to introduce a Public Broadcasting System a decade later.

Fourth, scholars moved beyond the issues raised by the changing conditions of national media to explore the re-composition of the emerging transnational communications system. Just as the last of the former colonial territories achieved political independence and emerged as autonomous nation states, critics led by Herbert Schiller (1969) argued that they were being incorporated into a new imperial system, based not on the annexation of their territories but the colonization of their cultures by the aggressive consumerism promoted by the major American media and entertainment corporations.

Fifth, encouraged by the rapid growth of the underground press, community radio, and a range of other alternative and oppositional media from the mid 1960s onwards, critical researchers began to explore the political economy of non mainstream and radical media.

These various strands of work attracted widespread criticism from cultural analysts working in other intellectual traditions who accused political economists of reducing cultural life to economic dynamics. This criticism is misplaced. Political economy sets out to show how the underlying dynamics of economic life shape communicative practice by allocating the material and symbolic resources required for action and how the resulting asymmetries in their control and distribution structure both the conditions of creativity in cultural production and the conditions of access and participation in cultural consumption. It focuses on the ways that capitalist dynamics help to organize the playing fields on which everyday social action takes place and write the rules of the game. Epistemologically it is rooted in critical realist rather than interpretive models. It does not deny that in any particular instance the run of play is marked by creativity, improvisation, and not infrequently, opposition and contest but its principal concern is not with events but with the constitution of underlying structures and the opportunities for action they facilitate and constrain. As noted earlier, it is particularly interested in the way that asymmetries in the distribution and control of core resources for cultural action impinge on the vitality of participatory democracy.

8. *Marketizing Media*

In pursuing this argument, critical political economists of media start from the material and cultural resources required for full citizenship. This leads them to evaluate the performance of public communications systems against two main criteria; the promotion of

cultural rights to information, knowledge, representation, and respect—and the provision of a cultural commons hospitable to the negotiation of unities that incorporate diversity. This is in marked contrast to those economists and mainstream political economists who start from the conditions required to maximize market choices and address individuals as consumers of goods rather than as participants in a moral community.

Since 1980, propelled by a revival of neo-liberal economic thinking (unencumbered by Adam Smith's strong moral concerns with benevolence and communality) the market has increasingly become the preferred mechanism for organizing public communicative activity and the touchstone for evaluating the performance of those cultural institutions that remain within the public sector. This general process of marketization has five major dimensions.

(a) *Privatisation*. Media organizations that were formerly part of the public sector and supported by public money have been sold to private investors and converted into private enterprises. Examples include BT, Britain's former PTT and now one of the leading corporations in the national economy, and the former French public service television channel, TFI.

(b) *Liberalization* involves opening markets that have previously been dominated by one or two operators (monopolies and duopolies) to competition from new entrants. Examples include the liberalization of telecommunications in the USA with the break-up of AT&T and the ending of its historic monopoly, and the introduction of commercial cable and satellite television services into European broadcasting markets previously monopolized by State-funded public service channels.

(c) *Re-gearing of the regulatory regimes* by shifting towards 'lighter touch' regulation that allows private companies more freedom of action in key areas such as corporate acquisitions and the amount and type of advertising carried, and removes or weakens established public interest requirements.

(d) *Corporatization* involves reducing the real level of public subsidy, encouraging or compelling public institutions to raise more money from market-based activities, moving them closer to the strategies and rationales of the corporate sector. The BBC's decisions to maximize the commercial returns on its programming, to develop new initiatives with private sector partners, and to launch its own commercial enterprises, provides a particularly good example of this process in action.

(e) *Commodification*. Cinema seats and paperback books and records are commodities. If someone purchases them for their personal use no one else can enjoy them at the same time. In contrast, for most of its history, broadcasting has been a public good. It has been available to anyone with suitable receiving equipment and has allowed any number of people to listen or view at the same time without interfering with each other's enjoyment. This characteristic provided the technical underpinning for the key tenet of universality. Over recent years however, this principle has been challenged by the rapid rise of subscription broadcasting services offered over cable and satellite systems, and more recently by the introduction of pay-per-view events. By converting broadcasting services into commodities these innovations restructure social access linking it ever more closely to ability to pay.

Taken together these intersecting processes of marketization have had three major effects on the ecology of public communications. First, they have consolidated and extended the economic power of the major multimedia corporations by opening up new fields for enterprise. Second, they have eroded the relative countervailing power of public institutions. Third, they have progressively detached the core sectors of public communications from the project of providing everyone with the basic cultural resources required for full citizenship and moved them towards the satisfaction of plural consumer demands. As market segments have proliferated so the cultural commons has contracted.

Nor is marketization confined to the major capitalist economies. With the collapse of the Soviet Union, the increased scope for overseas investment and private enterprise in the world's two most populous countries, China and India, and the almost universal embrace of market disciplines and opportunities in the rest of the world, it has become a defining feature of the present age and the principle dynamic underlying economic globalization.

9. *Digital Convergence and Cosmopolitan Citizenship*

Another pivotal process that is fundamentally transforming contemporary media systems is the new conjunction of previously separate communicative forms and sectors facilitated by the translation of writing, speech, images, data, and sound into a single universal digital code of 0s and 1s. This process of convergence has been widely credited with bringing about a revolution not simply in the communications industries but in the economy as a whole, as commentators see the ability to command and manipulate information and symbolic forms displacing industrial production at the center of capitalist production. Supporters of maximizing the scope for market dynamics celebrate the flexibility and fluidity of the new digital economy as an unequivocal gain for personal liberty and freedom of choice. Critical political economists, on the other hand, argue that because the new economic order is coalescing on ground already prepared by two decades of marketization, the appearance of novelty and change conceals some very familiar problems. They point to three in particular. First, as the merger between Time-Warner (the

world's leading 'old' media company) and AOL (a major Internet enterprise) suggests, the leading communications companies of the future will occupy strategic positions across all the major communications sectors, giving them unprecedented potential control over public culture. Second, they see the enlargement of the market sphere consolidating consumerism's promise of satisfaction and self-realization through the possession of goods as the master ideology of the age. Third, they warn that making access to core informational and cultural resources dependent on the user's ability to pay will widen inequalities of capacity, both within and between countries.

At the same time, critics of the new capitalism point to digital technology's potential to create a new and greatly enlarged public cultural sphere, based on transnational rather than national informational and cultural flows and providing the symbolic resources for a cosmopolitan citizenship rooted in a robust defence of human rights and respect for cultural diversity. They see the problem of translating this ideal into practical policies and institutional forms as the greatest challenge facing a political economy of communications committed to striking a just balance between economic dynamism and universal citizenship. On this reading, far from being overtaken by events, as some commentators have argued, critical political economy's greatest contributions to analysis and debate on the future of media may be yet to come.

See also: Advertising: General; Alternative Media; Communication and Democracy; Communication: Electronic Networks and Publications; Communication: Philosophical Aspects; Enlightenment; Freedom of the Press; Marshall, Alfred (1842–1924); Mass Communication: Normative Frameworks; Mass Media: Introduction and Schools of Thought; Media and History: Cultural Concerns; Media, Uses of; Political Communication; Political Economy, History of; Public Broadcasting; Public Sphere and the Media; Publishing as Medium; Telecommunications and Information Policy; Television: History

Bibliography

Bagdikian B H 2000 *Media Monopoly*, 6th edn. Beacon Press, Boston, MA

Barnouw E et al. 1997 *Conglomerates and the Media*. The New Press, New York

Danielian H R 1939 *The AT&T*. Vanguard, New York

Flichy P 1980 *Les Industries de Imaginaire: Pour une Analyse Economique des Media*. Presses Universitaires de Grenoble, Grenoble, France

Garnham N 2000 *Emancipation, the Media and Modernity: Arguments about the Media and Social Theory*. Oxford University Press, Oxford, UK

Golding P, Murdock G 2000 Culture, communication and political economy. In: Curran J, Gurevitch M (eds.) *Mass Media and Society*, 3rd edn. Arnold, London, pp. 70–92

Horkheimer M, Adorno T W 1973 [1944] The culture industry: Enlightenment as mass deception. In: Horkheimer M, Adorno T W (eds.) *Dialectic of Enlightenment*. Allen Lane, London

Klingender F D, Legg S 1937 *The Money Behind the Screen*. Lawrence and Wishart, London

Marshall A 1890 *Principles of Economics: An Introductory Volume*. Macmillan, London

Marx K, Engels F 1846 [1970] *The German Ideology*. (edited and introduced by Arthur C J). Lawrence and Wishart, London

Mattelart A 1979 *Multinational Corporations and the Control of Culture*. Humanities Press, NJ

McChesney R W 1999 *Rich Media, Poor Democracy: Communication Politics in Dubious Times*. University of Illinois Press, Urbana, IL

Miege B, Pajon P, Salaun J-M 1986 *L'Industrialisation de l'Audiovisuel*. Editions Aubier, Paris

Mosco V 1996 *The Political Economy of Communication*. Sage, London

Schiller H 1969 *Mass Communications and American Empire*. A. M. Kelley, New York

Schiller H 1989 *Culture Inc: The Corporate Takeover of Public Expression*. Oxford University Press, New York

Sinclair U 1920 *The Brass Check: A Study of American Journalism*. Published by the author, Pasadena, CA

Smythe D, Guback T (eds.) 1994 *Counterclockwise: Perspectives on Communication*. Westview Press, Boulder, CO

Wilcox D F 1900 The American newspaper: A study in social psychology. *Annals of the American Academy of Political and Social Science* **July**: 56–92

G. Murdock

Mass Media, Representations in

Representation involves making something perceivable to an audience that is not in its physical presence. Often the item represented—an idea, an ideology, an interest—possesses no tangible physical embodiment so representation means re-presenting it in a new symbolic form or context. Other times a physical antecedent exists, such as a person, a widely recognized social group, an institution, or a place. Even these objects, when represented in the media, typically carry ideological or cultural meanings. There is now an extensive literature on media representations of social and political groups and institutions, and on their implications for the distribution of status, wealth, and power.

No representation is comprehensively accurate. All media representations inevitably omit some aspects of the item represented. Indeed in most cases only a tiny fragment of the actual idea, person, or group will be perceivable to the audience. Audiences may and often do fill in much of the absence themselves, drawing on

their pre-existing mental theories or schemas. Studying representation means probing the combination of absences and presences, the meanings this total package is likely to convey to audiences, and the ramifications of these patterns of representation for power and belief in a society.

It is also vital to understand that there is no one-to-one correspondence between representation in the media text and the representation of the text and its signals as inferences drawn in the minds of audiences. The message that a communicator intentionally encodes in a media text may get decoded in ways that contradict or bear little relation to the communicator's purposes. Because representation involves both encoding and decoding, the conscious intentions of the communicator may have surprisingly little relevance to individuals' reactions and thus to the political and social impacts of media representation. This point is sometimes lost in popular and even academic studies of the media's influence, which frequently take individual media personnel to task for what analysts assume will be the effects of the representations. Vidmar and Rokeach (1974) provided a classic example of the slip between intentions and influence in their study of the popular American television program 'All in the Family.' To illustrate the absurdity of ethnic prejudice, Norman Lear, the show's creator, made its central character, Archie Bunker, a laughably ignorant, outspoken bigot. Yet despite the ridicule heaped upon Archie and his bigotry in most episodes, Vidmar and Rokeach revealed, the show merely reinforced viewers' predispositions. Far from changing their minds or even experiencing discomfort, prejudiced viewers saw Archie's tirades as vindication of their own sentiments.

Communication scholars studying media representation have typically done so to discern likely media influence. But cultural studies analysts and others working in the postmodern mode found surprising allies among those within the positivist and behavioralist traditions in stressing that individuals' interpretations of a given text vary so markedly that studying texts in isolation from their audiences produces flawed accounts of media representation. The major work of representation goes on within each thinking individual's brain rather than in the text, they say. Therefore, much of the scholarly focus in the 1990s shifted towards exploring the audience's interpretations rather than the media text by itself.

This article assumes that the representations in the text influence without determining individuals' sentiments. Beyond this, at the aggregate level, the pattern of representations in a society's mass media illustrate and help define its culture, reinforcing dominant cultural values yet also revealing and stimulating cultural change. These assumptions do not require a one-to-one correspondence between text and mind or text and culture, but do insist that the correlations are often significant. The conflict over just how significant—that is, how powerful media representations are—runs through media scholarship from its earliest days and continues in the most recent work. To cite ideal, polar types, some take the position that individuals make up their own minds, independently reconfiguring anything represented in the media to find meanings that suit their own experiences, knowledge, psychic needs, and goals. Others assert that in 'telling people what to think about,' the media can significantly and sometimes decisively influence an audience's attitudes and preferences. Between these two poles lies a diverse array of theorists who believe an audience's thinking is both influenced by and influences the representations that media transmit, with the degree of media impact depending on many environmental and individual differences.

One of the first major bodies of social scientific research on mass communication explored the representation of violence in movies and its impacts on youth. World War II spawned much research on propaganda—on media texts and their uses of symbols and language to generate potential shifts in mass attitudes. The next two decades saw a dry spell as scholars generalized from this and other early empirical research into media effects, concluding that any consequences of media representation were of minimal significance, limited largely to reinforcing existing predispositions. Advances in conceptualization and in analytical techniques challenged this stance and led to a renaissance in scholarship on media influence.

The more recent social scientific scholarship tends to run in two channels, one focusing on journalism and the news, the other on television and film entertainment (with occasional attention to magazines, radio, popular music, and novels). However, scholars toward the end of the century were increasingly recognizing the blurring of this always-problematic distinction between entertainment and news. Not only was news becoming more 'entertaining,' more shaped by strategies for audience appeal traditionally used by entertainment—in response to rising competition from commercial broadcasters, websites, print media, and other sources—but entertainment offerings on the US broadcast networks were becoming increasingly topical. A resulting hybrid dubbed 'infotainment' occupied increasing portions of the broadcast airwaves and of daily newspapers as well, in the US and elsewhere. As a result of this dissolution of once-distinct lines, an essay on media representation must not limit itself to the traditional news media, even though the bulk of social scientific research has probed the news.

1. *Framing*

Representation is always partial, always requires selection of aspects to stand for the whole (otherwise we would be speaking of presentation, not represen-

tation). Since framing is the process of selecting, highlighting, and suppressing, it is the means through which most media representation takes place. Framing has organized an increasing portion of the literature on media representation. Most of this research rests on the proposition that framing devices are worth studying because they affect (without necessarily determining) the audience's thoughts and perhaps actions. Framing can be defined as selecting aspects of a perceived reality and making them more salient in a communicating text, thereby promoting a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation.

The classic experiments of Kahneman and Tversky show that framing the same basic event or problem differently can have marked effects on an audience's understanding of the item represented. For instance, Kahneman and Tversky (1984) gave subjects in one group a choice between a disease prevention program, A, that would 'save' 200 out of 600 people, and B, which would have a one-third probability that all 600 would be 'saved.' Then, using a matched group, Kahneman and Tversky reversed the frame so that option C meant '400 people will die' and option D offered a one-third probability that 'nobody will die.' Option C was identical to A, as was D to B, excepting only that A and B were framed in terms of 'saving' and C and D in terms of 'dying.' Yet in the first experimental group, 72 percent of subjects chose A, whereas only 22 percent in the second group chose the identical option C—presumably because C was framed to highlight death instead of lives saved.

Kahneman and Tversky's experiments typically involved small changes in wording bringing about significant changes in response (apparently because the participants in the experiments thought about the matter differently, focusing their minds on different aspects of it). Media studies tend to document far more global differences in framing, involving not just word choice but disparities in visual portrayals, aspects of the item emphasized, sources relied upon, people quoted, and more. If differences of a single word can cue markedly different responses, the multidimensional differences in media frames often demonstrated by scholars implicitly suggest considerable media influence.

For example, US media developed starkly different frames for two incidents in which civilian airliners were shot down by military forces, resulting in nearly 300 deaths each. In the first incident, the downing of a Korean Air Lines jet by the Soviet Union in 1983, adjectives and nouns, dramatic graphics, attributions of agency and other aspects of the text emphasized the suffering of the victims and the Soviets' moral guilt. This frame fit with then-president Ronald Reagan's political agenda of heightening Americans' distrust of the Soviet Union and polls showed that the overwhelming majority of Americans accepted Reagan's interpretation of the incident. By contrast, when US forces shot down an Iranian passenger jet in 1988, the media framed the event as a regrettable accident traceable to understandable human error and deficiencies in technology. Language eschewed moral judgments and the victims received scant attention. Here too large majorities of Americans in surveys accepted the Reagan administration's framing—this time exculpatory rather than condemnatory. Public acceptance of the way the media represented the incidents in turn had significant political and policy consequences (Entman 1991).

Nonetheless, framing is far from all-powerful. William Gamson's studies of 'collective action frames' have shown how individuals, interpersonal discussion, and media frames interact. Studying issues including nuclear power and affirmative action, Gamson (1992) found that audience members take cues from news reports, but, depending on their individual thinking, group discussions, and the nature of the issue, can draw from the media-supplied frames new ways of understanding the phenomenon reported. For example, journalists did not construct the dominant news frames and thus the representation of nuclear power to stimulate antinuclear activism. Nonetheless, audience members gathered by Gamson in focus groups were sometimes able through collective deliberation to reframe the issue in a manner that might have moved some to action.

An understanding of the bidirectional influence of media representations and audience thinking immediately cautions against inferring any single meaning or effect from a particular media text. The same text can represent or 'mean' different things to different people in different circumstances—including different historical periods. Thus the words 'racial equality' represented one concept during the 1950s and 1960s—equal treatment and equal rights under the law for black Americans. More recently, the words appear to connote something quite different for many at least in the US, something akin to radical redistribution of income, wealth, and opportunity from the majority Whites to minority groups.

Despite these cautions, the bulk of research on media representations appears to assume that one basic interpretation often—though certainly not always—dominates, at least among the members of any specific culture at a particular time. In fact, one definition of a culture might be sharing cognitive habits and schematic mental linkages that promote similar responses to an attitude object among the individuals who constitute its membership. The (typically implicit) assumption that representations in the media have a 'preferred' or 'privileged' interpretation or meaning in a given culture structures much of the remaining research reviewed here.

To generalize broadly (an inevitability in a literature synthesis covering such a wide area of research), media research explores three aspects of representation corresponding to three kinds of effects (often presumed

rather than empirically demonstrated): the ways representation reflects and affects the general social status of group members such as women and blacks; the ways media representations might affect a mass audience's political preferences in a particular dispute, say in an election or a debate over legislative proposals; and the ways media representations might influence political elites and government officials.

2. Status Reinforcement or Enhancement

Students of media representation frequently examine depictions of groups and how they might reflect, reinforce, or challenge existing status hierarchies. For this purpose, a wide range of human classifications has been invoked, from race and ethnicity to gender, sexual orientation, age, and language groupings.

In this research tradition a major thrust focuses on ethnic groups. For the most part, the research suggests that media images reflect and reinforce dominant status judgments. That is, in framing activities involving members of lower-status groups, the media highlight aspects likely to reinforce perceptions among the dominant group's members of their difference from and superiority to outgroups. Representations tend to carry explicit or implicit judgments of the outgroups on dimensions of morality, intellectual capacity, industriousness, lawfulness, and other valued traits. But because media images are in constant flux and often contain contradictory elements, they also provide cognitive resources for challenges to status markers.

Of particular concern to many researchers has been a pattern whereby the media link crime and dangerous deviance to people of color and especially black males (Entman and Rojecki 2000). Experimental studies repeatedly have suggested that these representations engender cognitive and emotional effects on white audiences that affect the exercise and distribution of power—in this case heightening support for punitive prison sentences or diminishing support for white candidates when political advertisements associate them with black criminality (Gilliam and Iyengar 2000).

Studies of depictions of 'foreigners,' by which is usually meant nations and peoples other than those with white, European ancestry, also fit into this mode. For example, news media coverage of humanitarian crises in Africa, though typically sympathetic to those suffering, manages to endorse racial status hierarchies that place whites at the peak of the pyramid. Thus, Western news media made Bosnian whites appear in a variety of ways more human, more worthy of help, than news of the genocide going on simultaneously in Rwanda (Myers et al. 1996). The same appears true in entertainment. For example, a study of Brazilian TV star Xuxa suggests that her celebrity (and earnings) derived largely from her conformity to North American and North European ideals of blonde sexuality and beauty (Simpson 1998)—even in an ethnically heterogeneous country priding itself on multiracial tolerance. This white cultural preference reflects continuity with a long history in which European and US mass media implicitly derogated black people and valorized whites and the qualities typically associated with whiteness.

In a similar vein, gender representations in the media are found to register cultural change on some dimensions but to endorse traditional patriarchy and gender roles on others. Thus we see women actors portraying business executives in advertisements, and single women living happy lives without obsessing over men. But, although women's sports now receive media attention that gives the lie to assumptions that women are inevitably passive, gentle, timid and the like, closer study suggests the continued impact of gender stereotypes and a preference for males. For instance, research on portrayal of women's sports in Sweden shows that female athletes and their pursuits are implicitly derogated in comparison to males—over 90 percent of the time devoted to sports in a sample of Swedish television programming covered male athletics (Koivula 1999). Beyond this, women still drive the plot in far fewer movies and television programs than men, and still serve as sex objects more often than males. Women's sexuality indeed appears more than ever to be pictured in unrealizable body types that have been linked to an increase in anorexia and other eating disorders. Thus a study of Australian adolescent girls finds a strong influence of media consumption on tendencies to aspire toward unhealthily thin body types, findings replicated in other countries (Wertheim et al. 1997). The studies of female body images typically find that many audience members absorb the preferred meanings and integrate them into their own self-images. These findings challenge any claim that mass audiences always engage in subversive readings of media representations.

On the other hand, Fiske (1992) argues that despite the traditional patriarchal and commercial values inherent in the pop star Madonna's representations of femininity, young women frequently found reinforcement for nontraditional roles and values through their own counter-readings of Madonna's music and videos. Certainly the growth in sheer numerical representation of blacks, women, and other traditionally less-empowered groups in nontraditional roles in Hollywood movies, music videos, and television dramas has both reflected and reinforced a degree of movement toward a more egalitarian status hierarchy.

Looking at status on the dimension of sexual orientation, Joshua Gamson (1999) argues that although the confessional talk shows that became popular in the US during the 1990s, such as 'Jerry Springer' and 'Maury Povich,' seem on the surface to belittle anyone who deviates from the middle class, heterosexual norm, the programs actually offer a more

complex ideological brew. Sometimes the shows construct morality plays that censure gay, lesbian, or transsexual guests, but other times the intolerant snob, ivory tower expert, or sexual bigot assumes the role of villain. Arguably, says Gamson, gays and others living outside the bounds of dominant cultural values find a more welcoming home for self-representation in the 'trash' talk shows than in the rest of the mainstream media. In this sense, tabloid TV, driven as it may be by crass commercial goals, debasing as it may appear to public discourse, may open up new space in the public sphere for persons, ideas, and values previously denied much representation in either entertainment or news.

3. *Media Representation and Political and Policy Disputes*

Moving along a continuum from media representation's impacts on mass publics to impacts on the political linkages between them and elites, we turn now to representation in news coverage of current policy and political contests. Here too the research suggests media most often operate to reinforce structures of power that privilege native born white men.

Thus, for example, female political candidates have been shown to face disadvantages against male candidates. They often receive less news coverage than males, and the news agenda seems to follow male candidates' issue choices most closely. On the other hand, women candidates may benefit from being associated with positive sex-role stereotypes (Kahn 1994). With a black candidate in the race, campaign news tends to emphasize candidates' racial identities and to frame elections as contingent on bloc voting by whites and blacks (Reeves 1997), and this demonstrably handicaps African American candidates.

The representation of nonelectoral dissent against powerful institutions has also received careful scrutiny. Research reveals a tendency for media to accept moderate, polite dissidents who work within the system and make conventional, incremental demands, but to slant decisively against groups that seek more radical objectives, undermining their legitimacy with the public and thus reducing the need for elites to respond positively to their goals (Gitlin 1980).

A subject of much dispute in the US has been representations of parties and candidates in national elections, with some scholars alleging systematic biases favoring one party and others denying these exist. Additional studies explore other kinds of bias than partisanship, suggesting the possible impacts on voting turnout or civic orientations. For example, US journalists increasingly infused their representations of candidates and campaigns with cynicism during the latter decades of the twentieth century, and this appeared to encourage political distrust and withdrawal among audiences.

Alongside such research there also arose theoretical essays that challenged the notion of bias by questioning the legitimacy of its presumed opposite, 'objectivity.' The notion that media can represent people and events in comprehensively accurate and politically neutral ways appears to be a powerful professional ideal for journalists. Nonetheless, scholarship almost universally denies the proposition and asserts that in the course of manufacturing the news, individuals and media institutions must inevitably embed certain kinds of biases (that is, preferred meanings or politically consequential omissions) into their reports. Although US studies, at least, fail to uncover systematic partisan biases, other forms of bias have been analyzed extensively, such as a bias for reporting political conflict over agreement, and political process over policy substance.

4. *Elite Impacts*

The newest and least developed area of research examines the impacts of media representations on elites. In democratic systems, elites are under pressure to respond to 'public opinion,' but it turns out they have few reliable ways to discover what it is on any given issue under debate in the legislature. Surveys are only sporadically useful in trying to gauge the sentiments of a legislator's constituency. Thus, media representations of 'public opinion' to officials become highly consequential. Political elites can only respond to the public thinking that they perceive, and they get much of their data on this from media reports. Moreover, these reports can become self-fulfilling, because they also may influence the public itself. In this view the media become a representative vehicle in the more specific political sense of forging ties (imperfect to be sure) between mass public and elites, by representing 'public opinion' to elites—by making abstractions, here the public's purported preferences and priorities, perceivable so elites can act on that information.

The process becomes even more complicated because elites contending to control the outcome of policy disputes frequently resort to manipulating the media. This may help create impressions of public support or opposition, impressions that augment or derail the momentum of policy proposals. Yet another twist is that media representations have themselves become a contentious political issue—an issue that is itself subject to biased representation. In the US, surveys clearly reveal that increasing majorities over the last few election cycles of the twentieth century came to believe that the media exhibit a liberal bias (Watts et al. 1999). Yet, as noted earlier, empirical studies fail to demonstrate a systematic ideological slant in the news. The public's perception of bias appears rooted in news coverage of news-making itself, and in stories quoting politicians, party elites,

and think tank intellectuals who have made purported media bias a theme of their speeches and writings. Watts et al. (1999, p. 144) find that in campaign coverage, 'conservative candidates, party officials, and supporters have dominated the discourse with allegations of liberally slanted news content.' The claims of left-leaning bias 'overwhelmingly' outnumber those of conservative bias—in the 1996 campaign by 91 percent to 9 percent. Thus there appears to have been a bias in the media's representation of points of view about the media's ideological representations—ironically, conservatives may have succeeded in biasing the US media's representations of media bias. The topic of media representation has itself entered the arena of political contestation and strategy, although the effects on the media, or on power in politics and society, are not yet clear.

5. *Conclusion*

The spread of new media and many new sources of mediated information and images raises questions about the nature of the system of media representation and scholars' ability to generalize about it. The Internet is freely open to almost any representation one might imagine, and its images are readily and simultaneously available to audiences around the globe. The impact of this wider circulation of more diverse images and words could be profound, disrupting the past ability of mainstream media representations to set agendas and shape dominant interpretations. On the one hand, the Internet puts more potential power over individuals' internal representations and thoughts into the hands of each individual, at least those who devote sufficient time to surfing the Web and seeking new views. The seemingly significant media effects discussed here—on group status, on the outcome of elections and other disputes, or on elite behavior—could diminish. On the other hand, this very decline of a homogenizing mass media has some scholars worried that the public sphere will dissipate, only to be replaced by a cacophonous set of specialized subpublics that communicate not with each other but only within their own boundaries. In the future, the very notion of common culture and nationhood forged by public attention to common media of mass communication could come under pressure.

Bibliography

Entman R M 1991 Framing United States coverage of international news: Contrasts in narratives of the KAL and Iran Air incidents. *Journal of Communication* **41**(4): 6–27

Entman R M, Rojecki A 2000 *The Black Image in the White Mind: Media and Race in America*. University of Chicago Press, Chicago

Fiske J 1992 British cultural studies. In: Allen R C (ed.) *Channels of Discourse, Reassembled*. 2nd edn. University of North Carolina Press, Chapel Hill, NC

Gamson J 1999 *Freaks Talk Back: Tabloid Talk Shows and Sexual Nonconformity*. University of Chicago Press, Chicago

Gamson W A 1992 *Talking Politics*. Cambridge University Press, New York

Gilliam F D, Iyengar S 2000 Prime suspects: The influence of local television news on the viewing public. *Am J Pol Sci* **44**: 560–73

Gitlin T 1980 *The Whole World is Watching: Mass Media in the Making and Unmaking of the New Left*. University of California Press, Berkeley, CA

Kahn K F 1994 Does gender make a difference? An experimental examination of sex stereotypes and press patterns in statewide campaigns, coverage. *American Journal of Political Science* **38**: 162–95

Kahneman D, Tversky A 1984 Choices, values, and frames. *American Psychologist* **39**: 341–50

Koivula N 1999 Gender stereotyping in televised media sport coverage. *Sex Roles* **41**: 589–604

Myers G, Klak T, Koehl T 1996 The inscription of difference: news coverage of the conflicts in Rwanda and Bosnia. *Political Geography* **15**: 21–46

Reeves K 1997 *Voting Hopes or Fears?: White Voters, Black Candidates & Racial Politics in America*. Oxford University Press, New York

Simpson A 1998 Representing racial difference: Brazil's Xuxa at the televisual border. *Studies in Latin American Popular Culture* **17**: 197–222

Vidmar N, Rokeach M 1974 Archie Bunker's bigotry: A study in selective perception and exposure. *Journal of Communication* **24**(1): 36–47

Watts M D, Domke D, Shah D V, Fan D P 1999 Elite cues and media bias in presidential campaigns. *Communication Research* **26**: 144–75

Wertheim E H, Paxton S J, Schutz H K, Muir S L 1997 Why do adolescent girls watch their weight? An interview study examining sociocultural pressures to be thin. *Journal of Psychosomatic Research* **42**: 345–55

R. M. Entman

Mass Society: History of the Concept

'Mass society' is a notion central to the assumption that modern, advanced societies possess the following features: a growing internal homogeneity, a combination of elite and bureaucratic control over the majority of the population (the so-called 'masses'), a specific kind of predominant culture ('mass culture,' linked to the 'mass media'), and an illiberal form of politics ('mass politics' and 'mass parties'). They are also said to reflect a new stage in the development of the industrial economy through 'mass production' and

'mass consumption.' 'Mass society' frequently assumes that, in such advanced societies, a certain type of human personality, the so-called 'mass man,' is proliferating and becoming omnipresent. This mass man is thought to embody the mindless uniformity of social conditions and moral attitudes.

1. The Twentieth Century Origin of the Concept

In its contemporary sense, the term 'mass society' (*Massengesellschaft*) was coined by Karl Mannheim in Germany in 1935. However, the first general description of a mass society, as it was to emerge in the social sciences and in the sphere of social ideas in the post-World War II period, was first put forward by the philosopher José Ortega in 1929, in his influential essay *The Revolt of the Masses*. Ortega also launched the expression 'mass man' (*hombre masa* in Spanish), a notion intimately related to that of mass society. These contributions from Mannheim and Ortega were preceded by other notions, such as 'mass production,' but most of the 'mass society' expressions appeared after these had become accepted terms. Of these, 'mass politics,' 'mass culture,' 'massification'—a translation of the German *Vermassung*—and 'mass consumption' are the best known. Still others, which did not always include the qualifying word 'mass,' were often used as synonyms of mass society or equivalents for it: for example, David Riesman's 'lonely crowd' and his concept of 'outer directed personality' which largely overlaps with standard descriptions of 'mass man.'

Despite the twentieth century origin of the expression (as well as of the set of concepts to which it soon became inextricably linked), both the notion of mass society and the arguments (or theories) behind it are much older. Thus the idea of a mass society contains some conceptions (and prejudices) about the nature of human society which are truly ancient, while some of the most elaborate interpretations of the world it claims to describe hark back to classical nineteenth century thinkers, such as Alexis de Tocqueville, to several pre-World War I philosophers and sociologists, and to contributions made by the so-called 'crowd psychologists' of the same period.

2. The Classical Origins of The 'Mass Society' Conception

2.1 The Ancient Roots of the Concept

The set of problems destined to occupy the minds of modern theorists of mass society are clearly related to those which inspired the ideas of the earliest conservative critics of democracy. The latter were concerned that the extension of citizenship and equality to the majority of the people making up the body politic might also entail the access of common, ignorant, vulgar, and politically and legally untrained men to public office. Public office, they thought, ought to be the preserve of legally trained, well educated, and wise citizens. Although Plato in the *Republic* put forward a scheme for the selection of the ablest and best citizens for high office (both women and men) on an egalitarian basis—through equal educational opportunities for all—most conservative observers at the time feared that all actual democracies would easily degenerate into mob rule.

Later, some historians, such as the Greek Polybius who lived under Roman rule, suggested that humankind underwent a cyclical historical process and that periods of democracy quickly and inevitably degenerated into what he called 'ochlocracy,' that is, mob rule and social chaos. Saint Augustine for his part was probably the first to use the word 'mass' for the majority of the people, by referring to those incapable of saving their own souls as the *massa damnata*. He explicitly distinguished an elite—those to be saved, to use his Christian term—and the rest, the undistinguished, united in their failure to free themselves from sin.

Thus the association between crowd behavior, unruly assemblies, emotional and irrational meetings, on the one hand, and the general breakdown of social order, on the other, appeared at the earliest period in the history of Western political thought. Similarly the people's supposedly innate vulgarity and lack of moral fiber was contrasted with the refinement, creativity, courage and distinction, and individuality of the select and responsible few. The dangers of demagoguery—irresponsible leadership of the *plebs* or populace, often coming from isolated members of the elite—were also mentioned by numerous political thinkers during the classical period—from Aristotle to Cicero—in terms not altogether different from those used by some representatives of the modern theory of mass politics.

2.2 Classical Underpinnings from Hobbes to Tocqueville

Neither medieval nor Renaissance thought permitted much further development of the seminal ideas about the majority and its relationship to privileged and powerful minorities, much less the rise of a conception of the social world in which egalitarianism was seen as posing a danger to freedom and a source of human bondage. The social upheaval of the seventeenth century, however, inspired Thomas Hobbes in *Leviathan* and especially in his reflections on the disturbances of the English Civil War, the *Behemoth*, to return to ancient themes about the pernicious results of upsetting the hierarchical order of society by demagoguery and mob rule. Political obedience to the sovereign by the many might entail strains and injustice, but anarchy led to much greater evils, both

for the body politic and for the preservation of civilization and human life. In *Leviathan*, by simplifying the structure of a viable body politic and making it quite homogenous under the power of a single sovereign, Hobbes pointed to some of the features and problems of government to which theorists of mass politics (i.e., of the political dimension of mass societies) would turn their attention in the twentieth century.

In the history of the concept of mass society, it is worth noting that neither the first fully-fledged interpretation of such phenomenon (Ortega 1929) nor the first general presentation of anything approaching the notion of a mass society, Alexis de Tocqueville's *Democracy in America* actually used the expression itself. Although Tocqueville's study (1835 and 1840) did not describe the rise of a mass society, his analysis of the dynamics of a liberal, individualistic and democratic polity, clearly pointed out the trends (both manifest and 'subterranean,' to use his characteristic expression) that would eventually lead to it: thus the pressure of a growing 'equality of conditions' in the modern world would make people more similar to each other and therefore erase the distinctions between would-be unique individuals. The powerful ideology of egalitarianism would also have similar repercussions: concessions to majority wishes made by politicians would produce a more mediocre kind of democracy; conformity, rather than orderly dissent, would undermine the vitality of the new culture of the age; hedonism, rather than striving for higher forms of culture and ideals, would inspire the behavior of a selfish citizenry. In a word, a world of lonely, aimless, vulgar pleasure-seeking and disoriented citizens could eventually—though not necessarily—arise from the general trends to be perceived in the age of democracy.

2.3 Conservatism and 'Fear of the Masses'

Tocqueville's contribution is important because, for the first time, and almost explicitly, he made a distinction between crowd and mob behavior, on the one hand, and the increasingly uniform structure of advanced democratic societies, on the other. Nevertheless, in the decades that followed, observers tended to concentrate on crowd and mass behavior rather more than on the homogenization (or 'massification' as it later was called) of the entire society.

This interest in crowd behavior came, almost entirely, from the conservative quarter. In their eyes, the progress of democracy during the nineteenth century, especially in the great urban settings of the West, would bring many people—i.e., particularly industrial workers—into the streets to confront governments or to defy public order with demonstrations, strikes, and public meetings. An entire series of observers—mostly the so-called crowd psychologists—produced studies and speculated about the irrational, emotional, and manipulable behavior of the masses. Gustave Le Bon's *The Psychology of Crowds* (1896) is the best-known, though many other similar studies were produced, especially in Italy and England. A best-seller for its time, Le Bon's study brought the concept a step closer to the identification of crowd behavior with the general structure of society, already described by some at the time as 'the age of the masses.' A later essay by Ortega bears a title that still refers to the 'masses' rather than to an 'age' or 'society.' Meanwhile Oswald Spengler's *Decline of the West* ends with a somber and catastrophic view of the future of civilization precisely in terms of an 'age of the masses,' the 'barbarians from below.' For his part, Sigmund Freud wrote one of his most significant studies by developing critically some of Le Bon's ideas and intimated, like Le Bon and Spengler, that the future would belong to the rising masses and their dangerous manipulators. The masses were, by definition, incapable of rational and analytical thought and were also undemocratic in every way.

An alternative, democratic view, of the majority had difficulty emerging. At a given moment, anarchist and socialist thinkers began to accept the notion of 'the masses' in order to refer to the people at large (industrial and peasant workers together) in an entirely different vein to that of conservatives. Yet, with the rise of Bolshevism and, later, Stalinism, the 'masses' came to refer abstractly to a given population as controlled by the single party (the 'vanguard of the proletariat') and its 'cadres.' As such, the 'masses' were only assigned a subordinate role of acquiescence and obedience to the monopolistic party, instead of the autonomy characteristic of the people making up a truly free civil society. It cannot be said that, during the period of (Stalinist) Communist predominance in large parts of the world (1917–89), observers within the narrow bounds of the official Marxist–Leninist ideology showed any confidence on the 'masses.' Thus, paradoxically, their views coincided with the pessimistic liberals' view of the same period.

Some Marxists or Neo-Marxists would eventually come to accept the term 'mass society,' which they applied mostly to advanced capitalist and 'imperialist' countries. Nevertheless, Western Marxists who escaped the discipline of their Stalinist counterparts often incorporated the concept of mass society to their analyses, as we shall presently see.

3. The Rise of the Modern Theory of Mass Society

3.1 Ortega and Mannheim

As pointed out above, the 'standard' interpretation of the modern world in terms of a mass society found in Ortega's *The Revolt of the Masses* represented the first general synthesis. His essay also included a number of original ideas (e.g., specialism as a form of 'barbarism,'

or the notion that 'mass man' is not necessarily linked to the lower classes or to the proletariat). The anxieties and preoccupations of liberals (Tocqueville, John Stuart Mill), philosophers of civilization's decline (Spengler, Burckhardt), crowd and mass psychologists (Le Bon, Freud) elegantly and forcefully converged with Ortega's considerations but did not use the expression. Mannheim's notion of 'society of the masses,' more easily translated into 'mass society,' completed Ortega's vision, with his greater insistence on the bureaucratic and industrial structure of such universe. This was developed in his 1935 *Mensch and Gesellschaft im Zeitalter des Umbaus* (*Man and Society in the Age of Reconstruction* 1940). Mannheim, less conservative than Ortega, put forward possible solutions in a reformist, social democratic spirit.

Translated into English by Edward Shils—himself later a critical analyst of the notion—Mannheim's essay opened the way for the widespread use of the expression 'mass society' in the English-speaking world. In Spanish, many writers retained the plural (*sociedad de masas*), but this has not always been the case in other languages, such as French (*société de masse*) or Italian (*società di massa*; also, for mass politics, *regime di massa*; while Ortega's *hombre masa* was translated into Italian as *uomo di massa*).

3.2 Predominance and Expansion of the Mass Society Conception

As Daniel Bell was to point out in his *The End of Ideology* (1960), the mass society conception of advanced societies had become the predominant general vision of the world at the time amongst critics and thoughtful observers, save for the then widespread Marxist views about capitalism and its future. It remained so for a long time, although ramifications, schools of thought, and rival interpretations would also appear. This is not the place for a description nor a classification of such interpretations, however, we should mention that, alongside more conservative (in some cases, overtly pessimistic) views about the rise of a mass society in the framework of Western industrial advanced societies, there emerged some observers who embraced the notion from a left-wing or radical stance. C. W. Mills' attacks against the supposed degeneration of American society and democracy in his two studies *White Collar* (1951) and *The Power Elite* (1956) were typically made from a libertarian position against the development of a mass society on American soil (as a consequence of mass marketing, political manipulation, the undemocratic elite control over the citizenry by a 'military-industrial complex,' etc.)

Likewise, the widely influential neo-Marxist Frankfurt School, on both sides of the Atlantic, embraced the mass society interpretation and had a vast influence upon a wide educated public. Some of its products (see Adorno's *The Authoritarian Personality* (1950), which included a re-working of the mass man notion) were significant contributions in terms of social research; other studies, such as Horkheimer and Adorno's 1944 *Dialectic of Enlightenment* (English translation 1972) included a relatively sophisticated 'hegelianization' of the mass society theory. Yet others from the same school—e.g., Marcuse's *One-dimensional Man*, 1964—popularized the notions of mass society and mass man for a large and quite anxious audience with less philosophical rigor.

More significant perhaps than any division of the notion and theory of mass society into a conservative and a radical version (from the early 1950s to the mid-1970s, by which time it had become truly dominant) was the development of 'specialized' branches of the conception. Two stood out. Firstly, the study of mass politics (further divided into speculations and research on totalitarian mass politics, and the study of pluralistic and democratic mass politics and parties). Secondly, considerable interest was aroused by what Shils had called 'the culture of mass society' (1975a, 1975b). Mass culture and media studies, sociological or otherwise as well as philosophical essays, proliferated and soon became a highly developed field of research and analysis. Most of the debates about the nature and dynamics of mass society took place (or often, raged) within this new field.

4. Accommodation and Persistence of the Concept

The successful emergence of other general concepts for referring to advanced societies in the most general terms (for instance 'post-industrial societies') either relegated or, in some cases, even drove out the use of the term 'mass society' from its frequent use by the mid-1970s. It did not disappear, however. The mass society conception was anything but a fashion, as its credentials in social philosophy, political theory, and the rigorous critique of civilization demonstrate. The predominant tendency, from the 1980s till the first years of the twenty-first century has seen the mass society notion finding accommodation within a number of other general descriptive concepts (and corresponding interpretations). Thus, along with such concepts as 'post-industrial societies,' we find 'post-capitalist' societies, 'corporate' societies, 'consumer' societies, 'information' societies, and several others.

These expressions are not mutually exclusive. Rather, they refer to different facets or dimensions of the same highly complex phenomenon. To give one example, the widespread interest in the development of corporatism, neocorporatism, and the so-called 'corporate societies' in the 1970s and 1980s demonstrated that theories and interpretations produced by students of these phenomena were often highly compatible with the mass society theory. The latter's emphasis on administrative control, the rise of large

transnational corporations, and anonymous social forces seemed to complement theories of corporatism.

The situation at the turn of the twenty-first century seems to indicate a certain decline, although a number of examples can be produced to demonstrate its persistence. Of all its several branches, the concept, research field, and theories of mass culture and the mass media have not lost—the contrary seems to be the case—any ground at all, either in the West or elsewhere.

More recent general interpretations, notoriously the 'globalization' literature, are not only compatible with the classical tenets of the mass society interpretation but are partly, it can be easily claimed, extensions of it. The notion of mass society always included a view of the expansion, interdependence, and general globalization of the supposed mass characteristics of late industrial Western civilization. The idea of world convergence of mass societies was explicitly built into it.

See also: Authoritarian Personality: History of the Concept; Communism, History of; Crowds in History; Hobbes, Thomas (1588–1679); Industrial Society/Post-industrial Society: History of the Concept; Mannheim, Karl (1893–1947); Tocqueville, Alexis de (1805–59); Totalitarianism

Bibliography

Adorno T W 1950 *The Authoritarian Personality*. Harper Row, New York

Arendt H 1958 *The Origins of Totalitarianism*. Meridian Books, New York

Bell D 1960 *The End of Ideology*. Free Press, Glencoe, IL

Chakhotin S 1940 *The Rape of the Masses*. Routledge, London

Giner S 1976 *Mass Society*. Academic Press, New York; Martin Robertson, London

Horkheimer T W, Adorno T 1972 *Dialectic of Enlightenment*. Herder & Herder, New York

Kornhauser W 1959 *The Politics of Mass Society*. Free Press, Glencoe, NY

Le Bon G 1896 *The Psychology of Crowds*. Alcan, Paris

Mannheim K 1940 *Man and Society in the Age of Reconstruction*. Kegan Paul, London

Marcuse H 1964 *One-dimensional Man*. Beacon Press, Boston

Mills C W 1951 *White Collar, The American Middle Classes*. Oxford University Press, New York

Mills C W 1956 *The Power Elite*. Oxford University Press, New York

Olson P (ed.) *America as a Mass Society*. Free Press, Glencoe, NY

Ortega y Gasset J 1929 *La rebelión de las masas*. Revista de Occidente, Madrid

Rosenberg B, White D M (eds.) 1957 *Mass Culture*. Free Press, Glencoe, NY

Shils E 1975a The stratification system of mass society. In: Shils E (ed.) *Center and Periphery: Essays in Macrosociology*. University of Chicago Press, Chicago, pp. 304–16

Shils E 1975b The theory of mass society. In: Shils E (ed.) *Center and Periphery: Essays in Macrosociology*. University of Chicago Press, Chicago, pp. 91–110

Vidich A, Bensman J 1960 *Small Town in Mass Society*. Doubleday, Garden City, NJ

S. Giner

Mastery Learning

1. The Theory and Practice Of Mastery Learning

Since the 1970s, few programs have been implemented as broadly or evaluated as thoroughly as those associated with mastery learning. Programs based on mastery learning operate in the twenty-first century at every level of education in nations throughout the world. When compared to traditionally taught classes, students in mastery-learning classes have been consistently shown to learn better, reach higher levels of achievement, and to develop greater confidence in their ability to learn and in themselves as learners (Guskey and Pigott 1988, Kulik et al. 1990a).

In this section we will describe how mastery learning originated and the essential elements involved in its implementation. We will then discuss the improvements in student learning that typically result from the use of mastery learning and how this strategy provides practical solutions to a variety of pressing instructional problems.

2. The Development of Mastery Learning

Although the basic tenets of mastery learning can be traced to such early educators as Comenius, Pestalozzi, and Herbart, most modern applications stem from the writings of Benjamin S. Bloom of the University of Chicago. In the mid-1960s, Bloom began a series of investigations on how the most powerful aspects of tutoring and individualized instruction might be adapted to improve student learning in group-based classes. He recognized that while students vary widely in their learning rates, virtually all learn well when provided with the necessary time and appropriate learning conditions. If teachers could provide these more appropriate conditions, Bloom (1968) believed that nearly all students could reach a high level of achievement.

To determine how this might be practically achieved, Bloom first considered how teaching and learning take place in typical group-based classroom settings. He observed that most teachers begin by

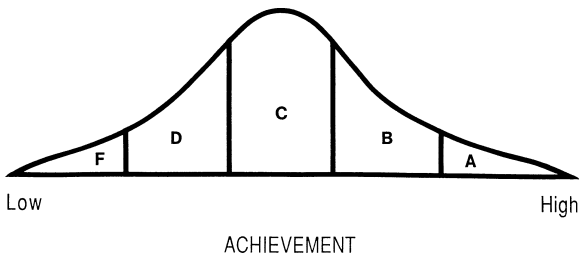

Figure 1
Achievement distribution curve in most traditional classrooms

dividing the material that they want students to learn into smaller learning units. These units are usually sequentially ordered and often correspond to the chapters in the textbook used in teaching. Following instruction on the unit, teachers administer a test to determine how well students have learned the unit material. Based on the test results, students are sorted, ranked, and assigned grades. The test signifies the end of the unit to students and the end of the time they need to spend working on the unit material. It also represents their one and only chance to demonstrate what they have learned.

When teaching and learning proceed in this manner, Bloom found that only a small number of students (about 20 percent) learn the concepts and material from the unit well. Under these conditions, the distribution of achievement at the end of the instructional sequence looks much like a normal bell-shaped curve, as shown in Fig. 1.

Seeking a strategy that would produce better results, Bloom drew upon two sources of information. He considered first the ideal teaching and learning situation in which an excellent tutor is paired with an individual student. In other words, Bloom tried to determine what critical elements in one-to-one tutoring might be transferred to group-based instructional settings. Second, he reviewed descriptions of the learning strategies used by academically successful students. Here Bloom sought to identify the activities of high achieving students in group-based learning environments that distinguish them from their less successful counterparts.

Bloom saw organizing the concepts and material to be learned into small learning units, and checking on students' learning at the end of each unit, as useful instructional techniques. He believed, however, that the unit tests most teachers used did little more than show for whom the initial instruction was or was not appropriate. On the other hand, if these checks on learning were accompanied by a 'feedback and corrective' procedure, they could serve as valuable learning tools. That is, instead of marking the end of the unit, Bloom recommended these tests be used to 'diagnose' individual learning difficulties (feedback) and to 'prescribe' specific remediation procedures (correctives).

This feedback and corrective procedure is precisely what takes place when a student works with an excellent tutor. If the student makes an mistake, the tutor points out the error (feedback), and then follows up with further explanation and clarification (corrective). Similarly, academically successful students typically follow up the mistakes they make on quizzes and tests, seeking further information and greater understanding so that their errors are not repeated.

With this in mind, Bloom outlined an instructional strategy to make use of this feedback and corrective procedure. He labeled the strategy 'Learning For Mastery' (Bloom 1968), and later shortened it to simply 'Mastery Learning.' By this strategy, the important concepts students are to learn are first organized into instructional units, each taking about a week or two of instructional time. Following initial instruction on the unit concepts, a quiz or assessment is administered. Instead of signifying the end of the unit, however, this assessment is primarily used to give students information, or feedback, on their learning. To emphasize this new purpose Bloom suggested calling it a 'formative assessment,' meaning 'to inform or provide information.' A formative assessment precisely identifies for students what they have learned well to that point, and what they need to learn better.

Included with the formative assessment are explicit suggestions to students on what they should do to correct their learning difficulties. These suggested corrective activities are specific to each item or set of prompts within the assessment so that students can work on those particular concepts they have not yet mastered. In other words, the correctives are 'individualized.' They might point out additional sources of information on a particular concept, such as the page numbers in the course textbook or workbook where the concept is discussed. They might identify alternative learning resources such as different textbooks, alternative materials, or computerized instructional lessons. Or they might simply suggest sources of additional practice, such as study guides, independent or guided practice exercises, or collaborative group activities. With the feedback and corrective information gained from a formative assessment, each student has a detailed prescription of what more needs to be done to master the concepts or desired learning outcomes from the unit.

When students complete their corrective activities, usually after a class period or two, they are administered a second, parallel formative assessment. This second assessment serves two important purposes. First, it verifies whether or not the correctives were successful in helping students remedy their individual learning problems. Second and more importantly, it offers students a second chance at success and, hence, serves as a powerful motivation device.

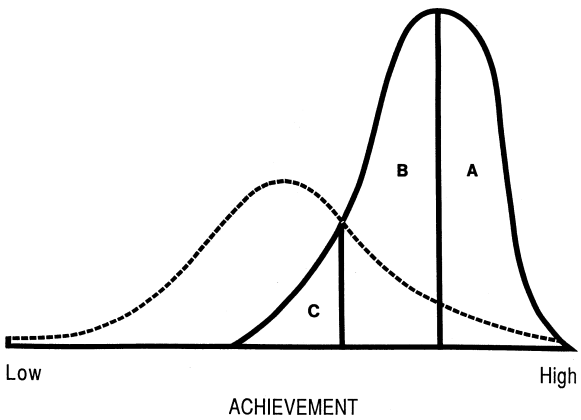

Figure 2
Achievement distribution curve in a mastery learning classroom

Bloom believed that by providing students with these more favorable learning conditions, nearly all could excellently learn and truly master the subject (Bloom 1976). As a result, the distribution of achievement among students would look more like that illustrated in Fig. 2. Note that the grading standards have not changed. Although the same level of achievement is used to assign grades, about 80 percent of the students reach the same high level of achievement under mastery-learning conditions that only about 20 percent do under more traditional approaches to instruction.

3. *The Essential Elements of Mastery Learning*

Since Bloom first outlined his ideas, a great deal has been written about the theory of mastery learning and its accompanying instructional strategies (e.g., Block 1971, 1974, Block and Anderson 1975). Still, programs labeled 'mastery learning' are known to greatly vary from setting to setting. As a result, educators interested in applying mastery learning have found it difficult to get a concise description of the essential elements of the process and the specific changes required for implementation.

In recent years two elements have been described as essential to mastery learning (Guskey 1997a). Although the appearance of these elements may vary, they serve a very specific purpose in a mastery learning classroom and clearly differentiate mastery learning from other instructional strategies. These two essential elements are (a) the feedback, corrective, and enrichment process; and (2) congruence among instructional components or alignment.

4. *Feedback, Correctives, and Enrichment*

To use mastery learning a teacher *must* offer students regular and specific information on their learning progress. Furthermore, that information or 'feedback' must be both diagnostic and prescriptive. That is, it should: (a) precisely reinforce what was important to learn in each unit of instruction; (b) identify what was learned well; and (c) describe what students need to spend more time learning. Effective feedback is also appropriate for students' level of learning.

However, feedback alone will not help students greatly improve their learning. Significant improvement requires that the feedback be paired with specific corrective activities that offer students guidance and direction on how they can remedy their learning problems. It also requires that these activities be qualitatively different from the initial instruction. Simply having students go back and repeat a process that has already proven unsuccessful is unlikely to yield any better results. Therefore, correctives must offer an instructional alternative. They must present the material differently and involve students differently than did the initial teaching. They should incorporate different learning styles or learning modalities. Correctives also should be effective in improving performance. A new or alternative approach that does not help students overcome their learning difficulties is inappropriate as a corrective approach and should be avoided.

In most group-based applications of mastery learning, correctives are accompanied by enrichment or extension activities for students who master the unit concepts from the initial teaching. Enrichment activities provide these students with exciting opportunities to broaden and expand their learning. The best enrichments are both rewarding and challenging. Although they are usually related to the subject area, enrichments need not be tied directly to the content of a particular unit. They offer an excellent means of involving students in challenging, higher level activities like those typically designed for the gifted and talented.

This feedback, corrective, and enrichment process, illustrated in Fig. 3, can be implemented in a variety of ways. Many mastery learning teachers use short, paper-and-pencil quizzes as formative assessments to give students feedback on their learning progress. But a formative assessment can be any device used to gain evidence on students' learning progress. Thus, essays, compositions, projects, reports, performance tasks, skill demonstrations, and oral presentations can all serve as formative assessments.

Following a formative assessment, some teachers divide the class into separate corrective and enrichment groups. While the teacher directs the activities of students involved in correctives, the others work on self-selected, independent enrichment activities that provide opportunities for these students to extend and

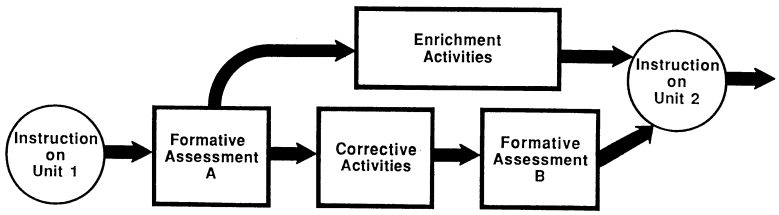

Figure 3
The process of instruction under mastery learning

broaden their learning. Other teachers team with colleagues so that while one teacher oversees corrective activities the other monitors enrichments. Still other teachers use cooperative learning activities in which students work together in teams to ensure all reach the mastery level. If all attain mastery on the second formative assessment, the entire team receives special awards or credit.

Feedback, corrective, and enrichment procedures are crucial to the mastery learning process, for it is through these procedures that mastery learning 'individualizes' instruction. In every unit taught, students who need extended time and opportunity to remedy learning problems are offered these through correctives. Those students who learn quickly and for whom the initial instruction was highly appropriate are provided with opportunities to extend their learning through enrichment. As a result, all students are provided with favorable learning conditions and more appropriate, higher quality instruction.

5. *Congruence Among Instructional Components*

While feedback, correctives, and enrichment are extremely important, they alone do not constitute mastery learning. To be truly effective, they must be combined with the second essential element of mastery learning: congruence among instructional components.

The teaching and learning process is generally perceived to have three major components. To begin we must have some idea about what we want students to learn and be able to do; that is, the learning goals or outcomes. This is followed by instruction that, hopefully, results in competent learners—students who have learned well and whose competence can be assessed through some form of evaluation. Mastery learning adds a feedback and corrective component that allows teachers to determine for whom the initial instruction was appropriate and for whom learning alternatives are required.

Although essentially neutral with regard to what is taught, how it is taught, and how the result is evaluated, mastery learning does demand there be consistency and alignment among these instructional components. For example, if students are expected to learn higher level skills such as those involved in application or analysis, mastery learning stipulates that instructional activities be planned to give students opportunities to actively engage in those skills. It also requires that students be given specific feedback on their learning of those skills, coupled with directions on how to correct any learning errors. Finally, procedures for evaluating students' learning should reflect those skills as well.

Ensuring congruence among instructional components requires teachers to make some crucial decisions. They must decide, for example, what concepts or skills are most important for students to learn and most central to students' understanding of the subject. But in essence, teachers at all levels make these decisions daily. Every time a test is administered, a paper is graded, or any evaluation made, teachers communicate to their students what they consider to be most important. Using mastery learning simply compels teachers to make these decisions more thoughtfully and more intentionally than is typical.

6. *Misinterpretations of Mastery Learning*

Some early attempts to implement mastery learning were based on narrow and inaccurate interpretations of Bloom's ideas. These programs focused on low level cognitive skills, attempted to break learning down into small segments, and insisted students 'master' each segment before being permitted to move on. Teachers were regarded in these programs as little more than managers of materials and record-keepers of student progress. Unfortunately, similar misinterpretations of mastery learning continue in the twenty-first century.

Nowhere in Bloom's writing can the suggestion of such narrowness and rigidity be found. Bloom considered thoughtful and reflective teachers vital to the successful implementation of mastery learning and stressed flexibility in his earliest descriptions of the process:

> There are many alternative strategies for mastery learning. Each strategy must find some way of dealing with individual

> differences in learners through some means of relating the instruction to the needs and characteristics of the learners. The nongraded school is one attempt to provide an organizational structure that permits and encourages mastery learning (Bloom 1968, pp. 7–8).

Bloom also emphasized the need to focus instruction in mastery learning classrooms on higher level learning outcomes, not simply basic skills. He noted:

> I find great emphasis on problem solving, applications of principles, analytical skills, and creativity. Such higher mental processes are emphasized because this type of learning enables the individual to relate his or her learning to the many problems he or she encounters in day-to-day living. These abilities are stressed because they are retained and utilized long after the individual has forgotten the detailed specifics of the subject matter taught in the schools. These abilities are regarded as one set of essential characteristics needed to continue learning and to cope with a rapidly changing world (Bloom 1978, p. 578).

Recent research studies in fact show that mastery learning is highly effective when instruction focuses on high level outcomes such as problem solving, drawing inferences, deductive reasoning, and creative expression (Guskey 1997a).

7. *Research Results and Implications*

Implementing mastery learning does not require drastic alterations in most teachers' instructional procedures. Rather, it builds on the practices teachers have developed and refined over the years. Most excellent teachers are undoubtedly using some form of mastery learning already. Others are likely to find the process blends well with their present teaching strategies. This makes mastery learning particularly attractive to teachers, especially considering the difficulties associated with new approaches that require major changes in teaching.

Despite the relatively modest changes required to implement mastery learning, extensive research evidence shows the use of its essential elements can have very positive effects on student learning (Guskey and Pigott 1988, Kulik et al. 1990a). Providing feedback, correctives, and enrichments; and ensuring congruence among instructional components; takes relatively little time or effort, especially if tasks are shared among teaching colleagues. Still, evidence gathered in the USA, Asia, Australia, Europe, and South America shows the careful and systematic use of these elements can lead to significant improvements in student learning.

Equally important, the positive effects of mastery learning are not only restricted to measures of student achievement. The process has also been shown to yield improvements in students' school attendance rates, their involvement in class lessons, and their attitudes toward learning (Guskey and Pigott 1988). This multidimensional impact has been referred to as the 'multiplier effect' of mastery learning, and makes it one of today's most cost-effective means of educational improvement.

It should be noted that one review of the research on mastery learning, contrary to all previous reviews, indicated that the process had essentially no effect on student achievement (Slavin 1987). This finding not only surprised scholars familiar with the vast research literature on mastery learning showing it to yield very positive results, but also large numbers of practitioners who had experienced its positive impact first hand. A close inspection of this review shows, however, that it was conducted using techniques of questionable validity, employed capricious selection criteria (Kulik et al. 1990b), reported results in a biased manner, and drew conclusions not substantiated by the evidence presented (Guskey 1987). Most importantly, two much more extensive and methodologically sound reviews published since (Guskey and Pigott 1988, Kulik et al. 1990a) have verified mastery learning's consistently positive impact on a broad range of student learning outcomes and, in one case (i.e., Kulik et al. 1990b), clearly showed the distorted nature of this earlier report.

8. *Conclusion*

Researchers in 2001 generally recognize the value of the essential elements of mastery learning and the importance of these elements in effective teaching at any level. As a result, fewer studies are being conducted on the mastery learning process, per se. Instead, researchers are looking for ways to enhance results further, adding to the mastery learning process additional elements that positively contribute to student learning in hopes of attaining even more impressive gains (Bloom 1984). Recent work on the integration of mastery learning with other innovative strategies appears especially promising (Guskey 1997b).

Mastery learning is not an educational panacea and will not solve all the complex problems facing educators today. It also does not reach the limits of what is possible in terms of the potential for teaching and learning. Exciting work is continuing on new ideas designed to attain results far more positive than those typically derived through the use of mastery learning (Bloom 1984, 1988). Careful attention to the essential elements of mastery learning, however, will allow educators at all levels to make great strides toward the goal of all children learning excellently.

Bibliography

Block J H (ed.) 1971 *Mastery Learning: Theory and practice.* Holt, Rinehart and Winston, New York

Block J H (ed.) 1974 *Schools, Society and Mastery Learning.* Holt, Rinehart and Winston, New York

Block J H, Anderson L W 1975 *Mastery Learning in Classroom Instruction.* Macmillan, New York

Bloom B S 1968 Learning for mastery. *Evaluation Comment* **1**(2): 1–12
Bloom B S 1976 *Human Characteristics and School Learning*. McGraw-Hill, New York
Bloom B S 1978 New views of the learner: implications for instruction and curriculum. *Educational Leadership* **35**(7): 563–76
Bloom B S 1984 The two-sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher* **13**(6): 4–16
Bloom B S 1988 Helping all children learn in elementary school and beyond. *Principal* **67**(4): 12–17
Guskey T R 1987 Rethinking mastery learning reconsidered. *Review of Educational Research* **57**: 225–9
Guskey T R 1997a *Implementing Mastery Learning*, 2nd edn. Wadsworth, Belmont, CA
Guskey T R 1997b Putting it all together: integrating educational innovations. In: Caldwell S J (ed.) *Professional Development in Learning-Centered Schools*. National Staff Development Council, Oxford, OH, pp. 130–49
Guskey T R, Pigott T D 1988 Research on group-based mastery learning programs: a meta-analysis. *Journal of Educational Research* **81**: 197–216
Kulik C C, Kulik J A, Bangert-Drowns R L 1990a Effectiveness of mastery learning programs: a meta-analysis. *Review of Educational Research* **60**: 265–99
Kulik J A, Kulik C C, Bangert-Drowns R L 1990b Is there better evidence on mastery learning? A response to Slavin. *Review of Educational Research* **60**: 303–7
Slavin R E 1987 Mastery learning reconsidered. *Review of Educational Research* **57**: 175–213

T. R. Guskey

Mathematical and Logical Abilities, Neural Basis of

1. Introduction

Mathematical and logical abilities are fundamental to our ability to not only calculate but also perform basic skills such as telling time, and dialing a phone number. This article considers both the cognitive processes which provide fundamental numerical skills such as number comprehension and production, number comparison, and arithmetic, and their neural substrates.

2. Fundamental Numerical Abilities

2.1 Numeral Comprehension, Production, and Transcoding

An incredible advance for humanity was the ability to represent numerical quantities symbolically using *numerals*. We are capable of representing numerical quantities in a myriad of ways including arabic numerals (12), written verbal numerals (twelve), spoken numerals ('twelve'), and a variety of concrete representations (e.g., twelve hatch marks on a stick). Each of these symbolic representations must be converted into a meaningful common form (tens: 1, ones: 2) in order to allow comparison of quantities and calculation. The translation rules from numerals to quantity, and quantity to numerals are varied and surprisingly complex. Arabic numerals have important spatial ordering (relative to the rightmost numeral) with some symbols representing quantities (123456789) and others representing a placeholder such as the '0' in '102.' Verbal numerals have different syntactic and quantity related components (one thousand twenty three) which unlike arabic numerals, only provide information about non-zero entries. Critical to success with numerals are: conversion of numerals into their meaning (*numeral comprehension*), production of numerals which correspond to the quantity we intend (*numeral production*), and translation of numerals from one form (e.g., 6) into another (six, *numeral transcoding*).

2.2 Representation of Quantity, Counting and Number Comparison

Our most fundamental numerical abilities include the representation of quantity information, the ability to apprehend quantity by counting, and the ability to compare two numerical quantities to determine (for example) the larger of two quantities.

2.3 Arithmetic

Solving arithmetic problems requires a variety of cognitive processes, including comprehension and production of numerals, retrieval of *arithmetic table facts* (such as $3 \times 7 = 21$), and the execution of procedures specifying the sequence of steps to be carried out (e.g., multiply the digits in the right-most column, write the one digit of the product, and so forth).

3. Components of Numerical Cognition

In this section a purely functional question is considered: What are the different cognitive processes in the brain which allow for our fundamental numerical abilities? A particularly successful line of research in answering this question has come from the study of impairments in numerical abilities after brain injury. This *cognitive neuropsychological* approach seeks to answer the question: How must the cognitive system be constructed such that damage to that system results

in the observed pattern of impaired performance? The existence of multiple numerical cognition components has been based in part on evidence that some numerical components are impaired after brain damage, while other numerical abilities are spared. While the next sections will single out individual patients, there are in each case several documented cases with similar patterns of performance. These remarkable case studies are described below.

3.1 Numeral Comprehension and Production Processes

A distinction between calculation and numeral comprehension and production processes has been drawn in part on the findings from brain-damaged patients such as 'DRC,' who reveal striking impairments in remembering simple arithmetic facts (e.g., $6 \times 4 = 24$), while other abilities such as comprehending and producing written and spoken numerals are unimpaired. This evidence implies separate functional systems for comprehending and producing numerals from those for calculation. Several cases with similar findings, and the opposite pattern of impairment have been reported.

Comprehension and production processes are also clearly composed of several functional components. For example, some patients reveal highly specific impairments only in producing arabic numerals, such as producing '100206' in response to 'one hundred twenty six.' This pattern of performance reveals several separable components of numeral processing, including distinctions between: numeral comprehension and production, verbal and arabic numeral processes, and lexical and syntactic processing within one component (given the correct nonzero numerals in the response, despite impairment in the ability to represent the correct pattern of zeros for the arabic numeral).

An ongoing question in the numerical cognition literature revolves around the question of *asemantic transcoding procedures*. This notion suggests that it may be possible to translate numerals directly into other forms of numerals (e.g., $4 \rightarrow$ four) without having an intermediary step in which the meaning of the numbers is represented. Note that normal transcoding requires the comprehension of the initial problem representation, the representation of that numerical quantity in an abstract, asemantic form, and the subsequent production of the quantity in the format. Are there asemantic transcoding algorithms? The available evidence is mixed, with some studies suggesting there may be asemantic transcoding algorithms, yet there is still some question as to whether or not there is sufficient evidence to support this notion (Dehaene 1997).

3.2 Numerical Magnitude Representations

Response latencies for determining the larger of two numerals (from healthy adults) suggests that we represent numerical quantities in terms of a magnitude representation comparable to that used for light brightness, sound intensity and time duration, and use these magnitude representations to compare two or more numbers. Moyer and Landauer (1967) found that humans are faster at comparing numerals with a large difference (e.g., 1 vs. 9) than they are when they compare two numerals with a small difference (e.g., 4 vs. 5). Further, when the differences between the numerals are equated (e.g., comparison 1: 2 vs. 3, comparison 2: 8 vs. 9) the comparison between numbers which are smaller (e.g., 2 vs. 3) is performed more quickly than the comparison of larger numbers (e.g., 8 vs. 9). These findings led Moyer and Landauer to conclude that the arabic numerals are being translated into a numerical magnitude representation along a mental number line with increasing imprecision the larger the quantity being represented (a psychophysical representation which conforms to Weber's Law). This system appears to be central to the ability to represent numerical quantities, estimate, and compare numerical quantities meaningfully.

These magnitude representations are found in human adults, infants and children, and in animals. They appear to be related to our representation of other magnitudes, such as time duration and distance (Whalen et al. 1999). A crucial challenge for the development of numerical literacy is the formation of mappings between the exact number symbol representations which are learned in school, and the approximate numerical quantity representations which are present very early in life.

3.3 Arithmetic

Evidence from brain-damaged patients, as well as those with normal and impaired development have revealed that there are several functional systems required in order to be able to perform calculations, such as 274×59. Several distinctions can be drawn based on the study of brain-damaged patients and their impairments. First, there is a distinction between the ability to remember simple arithmetic facts (such as $6 \times 4 = 24$), and the ability to perform calculation procedures (such as those required for carrying numbers and placing zeroes in multidigit multiplication). Several brain-damaged patients reveal impairments which are specific to either fact retrieval, or multidigit calculation procedures, indicating that the ability to retrieve arithmetic facts and perform multidigit calculations are represented by different neural substrates.

Within the simple process of remembering arithmetic facts such as $6 \times 4 = 24$, there are multiple

processes involved. Brain-damaged patients have also revealed selective impairment of the ability to recognize the arithmetic operator $(+, -, \times, /)$, despite unimpaired ability to retrieve facts from memory. Others illustrate that it is possible to fully comprehend the process of multiplication or addition, but yet be unable to retrieve multiplication or addition facts (Warrington 1982).

There is currently some debate as to the form in which arithmetic facts are stored. Three major theories are under discussion. One possibility is that arithmetic facts are stored and retrieved in a sound-based representation (consistent with the idea of rote verbal learning, like a nursery rhyme). This notion is popular in part because multilingual individuals often report that they believe they are remembering arithmetic facts in the language in which they acquired arithmetic, even if they have subsequently used a different language in which they are fluent.

However, an alternative view is that arithmetic facts are stored in an abstract, meaning-based form which is not tied to a specific modality (e.g., spoken or written). This proposal was first suggested in part because we perform arithmetic with a variety of numeral forms (primarily arabic and spoken numerals), and so perhaps there is a central store of arithmetic facts which are independent of any numeral form. Several brain-damaged patients have been reported who reveal impairment which is independent of the form in which the problems were presented or answered, consistent with an amodal representation.

Supporters of the sound-based representation of arithmetic facts suggest that representations of numeral magnitude (one likely candidate for a meaning-based representation of arithmetic facts) might not be related to arithmetic fact retrieval. Some brain-damaged patients reveal the inability to determine exact arithmetic responses (e.g., $2+2=3$) but nevertheless can reject a highly implausible and answer such as $2+2=9$, indicating that there is an approximate number representation which provides the meaning of numbers that may be separate from the process of retrieving exact arithmetic facts (Dehaene 1997).

4. *Localization of Number Processes in the Brain*

Neuropsychological evidence has played a crucial role in informing functional theories of numerical cognition and its components. However, most studies have until now been framed within the context of cognitive neuropsychology, which examines the behavior of brain damaged patients at a purely functional level, without concern for brain localization or lesioned site. Accordingly, models of number processing have been framed exclusively in terms of cognitive processes without reference to brain structures. For this reason, much less is known about the relation between these cognitive processes these involved in the processing and their neural substrates.

Interpreting the available evidence is not entirely straightforward, in part because most reports sought to relate brain areas to arithmetic tasks, and not to specific cognitive processes such as arithmetic fact retrieval. For example, studies exploring lesioned-deficit correlations have focused typically on identifying lesion loci associated with impaired performance of some calculation task or tasks. However, an association between lesion site and impairment of a calculation task does not in itself constitute strong evidence for the damaged brain area to be implicated in arithmetic processing, because impaired performance could have resulted from disruption to nonarithmetic processes required by the task (such as attention, working memory, numeral comprehension, and numeral production).

These points also apply to research involving brain recording and cortical stimulation methods. In several recent studies, subjects performed a serial subtractions task (i.e., count backwards by sevens from 1,000) while brain activity was recorded, or cortical simulation was applied. A number of brain areas were found to be associated with performance of the task, including the left parietal lobe, the frontal lobes, and thalamic regions. Although intriguing, these findings do not necessarily imply that these brain areas are implicated in arithmetic processing; the reported data did not exclude the possibility that some or all of the brain areas are involved in nonarithmetic processes required by the serial subtractions task (e.g., comprehending the initial stimulus number, producing responses).

This series of studies left little insight into the localization of specific cognitive processes, and resulted in a report by Kahn and Whitaker (1991), who echo Critchley's statement that 'disorders of calculation may follow lesions in interior or posterior brain, left or right hemisphere, and both cortical and sub cortical structures.' While this may suggest the task of localizing arithmetic processes is daunting, it reveals that more recent data which focus on the localization of separate number processes (e.g., arithmetic fact retrieval), rather than data which seeks to determine which areas are active for a specific task (e.g., serial subtractions) may reveal significant insights into the relation between the mind and brain for numerical cognition.

4.1 *Dehaene's Anatomical and Functional Theory of Number Processing*

Stanislas Dehaene and colleagues were the first researchers to provide a theory of number processing which includes both the different functional components and their localization in the brain. According to the *Triple Code Model* there are three separate number codes in the brain: verbal, arabic, and magnitude (see Fig. 1). Verbal codes are located in the left

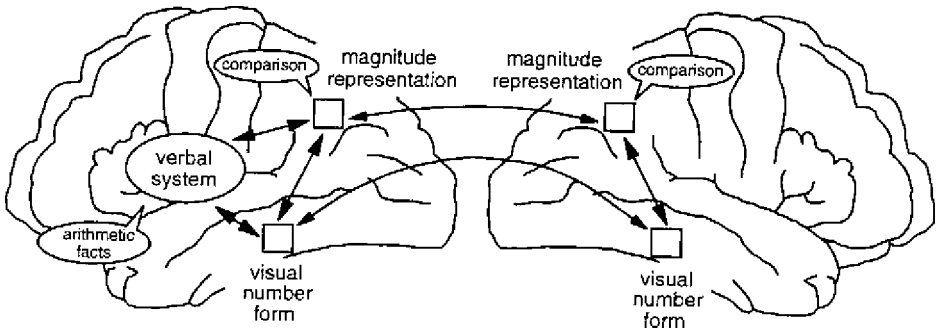

Figure 1
The triple-code model of numerical cognition

hemisphere language areas (e.g., Broca's and Wernicke's areas), and are responsible for holding numbers in memory, arithmetic fact retrieval, and comprehending and producing spoken numerals. Written numerals may also recruit temporal areas involved in visual word recognition. Arabic numerals are thought to be representing in temporal areas which are distinct from the visual word recognition area, and which are thought to be present in both hemispheres. This center is responsible for the recognition and production of arabic numerals. The ability to estimate and compare numbers involves quantity representations found in parietal areas of both hemispheres.

According the Triple Code Model, arithmetic table facts are stored in a sound-based form in language processing centers such as Broca's area. There are four fundamental components involved in calculation, which are: rote verbal memory, semantic elaboration, working memory, and strategy use. Dehaene proposes that retrieval of rote verbal arithmetic facts may be retrieved from a corticostriatal loop through the left basal ganglia, which is thought to store other linguistic material such as rhymes. In some cases, solving simple arithmetic facts may also involve semantic collaboration (such as determining that $9+7=10+6$, and retrieving the answer to $10+6$). If this semantic collaboration is involved, then the Triple Code Model predicts that parietal centers which represent numerical quantity will be involved. The next sections consider the available evidence regarding the localization of different arithmetic processes.

4.2 Numeral Comprehension and Production Processes

There is a large body of evidence suggesting that the comprehension and production of written and spoken verbal numerals is largely subsumed by the lexical and syntactic processing centers used for other linguistic material. For example, brain imaging studies have revealed that left-hemisphere language centers such as Broca's area are active during the production of spoken numerals, while Wernicke's area is active during the comprehension of spoken numerals, consistent with other language-based material (Hochon et al. 1999). Evidence from patients with disconnected brain hemispheres provides converging evidence. The left hemisphere shows a strong advantage over the right hemisphere in both the production and comprehension of verbal numerals. The right hemisphere can eventually produce spoken numerals, though the process is extremely laborious, and there is evidence which suggests that this hemisphere uses the counting string to produce the correct spoken numeral (Dehaene 1997).

There do, however, appear to be differences in the localization of visual numeral processes. Evidence from brain imaging experiments, and from cases in which the two brain hemispheres are disconnected, suggest that both hemispheres represent the ability to recognize arabic numerals. This region, located in the posterior temporal lobe, appears to represent arabic numerals exclusively and not other types of written symbols. In contrast, passive recording of numerical tasks involving written verbal numerals (e.g., twenty-eight) have provided evidence for a single, left hemisphere representation for written verbal numerals, which appears to be a subset of written word representations. In fact, there are striking impairments in which patients are unable to produce any letter symbol, or even their signature, yet are completely able to produce arabic numerals (Butterworth 1999).

4.3 Representations of Numerical Quantity

Perhaps the strongest single finding in the localization of numerical processes is that bilateral parietal regions represent numerical magnitude. Patients with disconnected hemispheres have revealed the ability to perform number comparison in either hemisphere (Dehaene 1997). Studies of the electrical brain signature during number comparison using a technique called *event related potentials* (ERP), has revealed several electrical waveforms during the comparison of two

numerals. Specifically, approximately one-tenth of a second after presenting two numbers, there is bilateral activation of parietal lobes, which varies according to the difficulty of the comparison. Further, this study also looked at the other waveforms and determined that they were involved in either the comprehension of the visual or spoken numerals, or in the action of responding.

As was presented earlier, responses are faster to comparisons involving large differences (e.g., which is more: 1 or 9) relative to comparisons involving smaller differences (e.g., which is more: 4 or 5). The bilateral parietal signal, called N1 (first negative activation) varies in amplitude according to the difficulty of the problem. The larger the difference between the numerals, the larger the N1 brain signal (Dehaene 1996). This suggests that the parietal activation was related to the operation of representing the magnitude of the numbers, and comparing them to decide the larger of the two numbers.

Brain imaging experiments involving functional magnetic resonance imaging (fMRI) have also provided evidence of parietal involvement in representing magnitude. In these experiments, the activation from a number comparison task has been compared with the activation in either a number reading task, or a simple calculation task. Each of these tasks presented a single numeral, which was compared with one in memory and required a spoken response, so that the brain activation for numeral comprehension, numeral production and working memory could be approximately equated, leaving the activation from the cognitive processes of interest to vary. In each case, relative to the control conditions (such as number reading) the inferior parietal lobe of both hemispheres was more active during number comparison than in the control conditions. This suggests that this area is likely to be involved in both the representing of numerical magnitudes, and in the comparison process itself (Pinel et al. 1999).

One question which has not been addressed in the reports of localization is: How do we represent numerical magnitude in the brain? There are a few studies which provide some relevant insights into the brain's representation of magnitude. Studies of parietal cortex in cats have found neurons which are tuned to specific magnitudes. For example, one neuron fired maximally when five items were presented to the cat, and responded in a weaker fashion to related quantities such as 4 and 6). Comparable patterns of performance have also been reported in human subjects (Dehaene 1997).

4.4 Arithmetic

A number of findings from studies of groups of impaired-brain damaged patients suggest that posterior cortical regions, particularly parietal regions, may play a role in arithmetic fact retrieval (Butterworth 1999). Impairment in calculation after brain damage, termed *acalculia*, occurs much more frequently after injury of the parietal lobes than to damage in other centers. Damage to left posterior brain regions impair numerical tasks, including those involving arithmetic, more so than damage elsewhere. However, these studies generally do not distinguish between the multiple components of complex calculation, including arithmetic fact retrieval, calculation procedures, and numeral comprehension and production.

Several single case studies have also implicated left parietal regions as a center for arithmetic fact retrieval, including the previously described patient 'DRC,' and multiple cases studied by Takayama et al. (1994). The application of electrical stimulation to the left parietal lobe (prior to neurosurgery) has also resulted in transient impairment to arithmetic fact retrieval during stimulation when the subject is otherwise completely capable of recalling arithmetic facts. Thus it appears that the left parietal lobe, and perhaps both parietal lobes, play a major role in the retrieval of arithmetic facts from memory.

Parietal cortex may not be the only region involved in retrieving arithmetic facts from memory. Calculation impairments have also been found after damage to frontal lobes and subcortical structures including the basal ganglia (Dehaene 1997). Some of these areas may be involved in processes other than arithmetic fact retrieval. For example, evidence from single case studies suggest that damage to the frontal lobes may produce impairment and an inability to produce multidigit calculation procedures (rather than impairing arithmetic fact retrieval). The hypothesis that frontal lobes play a role in calculation procedures is consistent with the finding from multiple brain imaging studies that complex calculation such as repeated subtractions activate not only parietal centers (thought to be involved in arithmetic fact retrieval) but also other centers such as the frontal area, which maybe involved in holding answers in memory, and performing multidigit calculation procedures.

Patients who have little or no communication between their cerebral hemispheres also provide some evidence as to the localization of arithmetic fact retrieval. When each hemisphere is given a calculation task, only the left hemisphere can retrieve arithmetic facts, and the right hemisphere produces very high error rates (80 percent) (Dehaene 1997). The right hemisphere's inability to perform the arithmetic task cannot be attributed to numeral comprehension or response production impairments. As was discussed earlier, these patients reveal the ability in each hemisphere to perform number comparison, suggesting that both hemispheres can both represent numerical magnitudes and comprehend arabic numerals.

In summary, current evidence indicates that the parietal lobe plays a major role in simple arithmetic

fact retrieval. Other subcortical regions such as the basal ganglia may also be involved. It is currently thought that frontal areas are involved in multidigit calculation procedures, and the working memory demands of complex calculation. These conclusions are somewhat at odds with the assumptions made by the Triple Code Model presented above. Dehaene suggests that the most frequent lesion sites which result in acalculia are in the left inferior parietal region, because this area provides semantic relations between numbers, and can inform fact retrieval. Thus lesions in this area might affect access to arithmetic memory without destroying the rote facts themselves. However, several cases do report specific fact retrieval deficits as a result of parietal lesions.

4.5 Recent Functional Brain Imaging (fMRI) Studies

The advent of fMRI, an excellent tool for measuring the localization of brain activity during cognitive tasks, has reinvigorated the study of the relation between human processing abilities and their corresponding neural substrates. For example, Dehaene and colleagues have performed several comparisons between activation levels during simple numerical tasks such as multiplication and number comparison (deciding the larger of two numbers). In one such fMRI study, Dehaene and colleagues revealed significant activation in bilateral inferior parietal regions in a comparison task relative to the multiplication task. This could be for one of two reasons. First, parietal regions could be active primarily during the comparison and not during arithmetic fact retrieval. However, another possibility is that parietal brain areas are involved in both multiplication and the comparison, but number comparison activates this region to a greater extent than does multiplication.

Additional evidence from comparisons between subtraction, multiplication, number comparison and number naming suggest that parietal brain regions are involved in both arithmetic and number comparison. Each of the arithmetic tasks and number comparison revealed significant parietal activation relative to the number naming control (all of which had comparable number comprehension and production requirements).

5. Summary

Numerical processes in the brain have several subsystems, including those for numeral comprehension and production, representations of number magnitude, remembering specific arithmetic facts, and performing more complex mathematical procedures. Verbal representations of number appear to be represented largely in the left hemisphere, and are considered to be a subset of non-numerical language representations. Arithmetic fact retrieval appears to involve more than one center, but certainly involves parietal and subcortical centers.

Unlike verbal numbers and words, arabic numeral representations have highly bilateral representations, as do representations of number magnitude. Magnitude representations appear to be responsible for number comparison, estimation, and play a crucial role in arithmetic. The representations of number appear to mirror other psychophysical representations such as time duration and light intensity. Finally, complex arithmetic procedures such as those for multidigit calculation appear to recruit frontal brain regions for holding partial facts in memory, and following a stepwise procedure.

Bibliography

Butterworth B 1999 *What Counts: How Every Brain Is Hardwired For Math*. The Free Press, New York

Dehaene S 1996 The organization of brain activations in number comparison: Event-related potentials and the additive-factors method. *Journal of Cognitive Neuroscience* **8**(1): 47–68

Dehaene S 1997 *The Number Sense: How The Mind Creates Mathematics*. Oxford University Press, New York

Hochon F, Cohen L, van de Moortele P, Dehaene S 1999 Differential contributions of the left and right inferior parietal lobules to number processing. *Journal of Cognitive Neuroscience* **11**(6): 617–30

Kahn H, Whitaker H 1991 Acalculia: An historical review of localization. *Brain and Cognition* **17**: 102–15

Moyer R S, Landauer T K 1967 Time required for judgments of numerical inequality. *Nature* **215**: 1519–20

Pinel P, Le Clec'H G, van de Moortese P, Dehaene S 1999 Event-related fMRI analysis of the cerebral circuit for number comparison. *Neuroreport* **10**: 1473–79

Takayama Y, Sugishita M, Akiguchi I, Kimura J 1994 Isolated acalculia due to left parietal lesion. *Archives of Neurology* **51**: 286–91

Warrington E K 1982 The fractionation of arithmetical skills: A single case study. *Quarterly Journal of Experimental Psychology* **34A**: 31–51

Whalen J, Gallistel C R, Gelman R 1999 Non-verbal counting in humans: The psychophysics of number representation. *Psychological Science* **10**(2): 130–7

J. Whalen

Mathematical Education

1. Introduction

Over the past decades mathematical education has been the subject of many controversial discussions relating as much to its goals and content as to the nature of processes of learning and teaching mathematics. For instance, the so-called 'new math reform'

of the 1960s and 1970s, which resulted in substantial changes in the mathematics curriculum in many countries, afterwards often elicited heated debates, and the same holds true for issues relating to teaching and learning mathematics, such as the role of discovery learning, and the use of technological devices in general, and the calculator in particular, in the mathematics classroom. This article does not allow us to give a complete overview of this vast domain of inquiry. Therefore, we will focus selectively on a few major aspects of mathematical education: (a) a dispositional view of the goals of mathematics education; (b) mathematics learning as the construction of knowledge and skills in a sociocultural context; (c) designing powerful teaching–learning environments for mathematics; and (d) constructing innovative forms of assessment instruments tailored to the new goals of mathematics education. For a more complete overview, we refer to the following volumes in which the vast body of studies has been excellently reviewed: *Handbook of Research on Mathematics Teaching and Learning* (Grouws 1992); *International Handbook of Mathematics Education* (Bishop et al. 1996); *Theories of Mathematical Learning* (Steffe et al. 1996). These volumes also show that since the 1970s mathematical education has evolved into an interdisciplinary field of study in which instructional psychologists, mathematics educators, mathematicians, and anthropologists are major participants. Within this community has developed an enriched conception of mathematics learning as involving the construction of meaning, understanding, and problem-solving skills based on the modeling of reality. During the same period, several important shifts have taken place in both conceptual and methodological approaches to mathematical education: from a concentration on the individual to a concern for social and cultural factors; from 'cold' to 'hot' cognition; from the laboratory to the classroom as the primary setting for research; and from a mainly quantitative experimental approach to a more diversified methodological repertoire including qualitative and interpretative techniques.

2. *A Dispositional View of Mathematics Learning*

This section focuses on what students should learn in order to acquire competence in mathematics. In this respect, the more or less implicit view which often still prevails in today's educational practice is that computational and procedural skills are the essential requirements. This contrasts sharply with the view that has emerged from the research referred to above. Indeed, there is nowadays a broad consensus that the major characteristics underlying mathematical cognition and thinking are the following (see De Corte et al. 1996, Schoenfeld 1992):

(a) A well-organized and flexibly accessible domain-specific knowledge base involving facts, symbols, conventions, definitions, formulas, algorithms, concepts, and rules that constitute the contents of mathematics as a subject-matter field.

(b) Heuristic methods, that is search strategies for problem solving that do not guarantee that one will find the solution, but substantially increase the probability of success because they induce a systematic approach to the task (e.g., thinking of an analogous problem; decomposing a problem into subgoals; visualizing the problem using a diagram or a drawing).

(c) Metacognition, involving knowledge and beliefs concerning one's own cognitive functioning on the one hand (e.g., believing that one's mathematical ability is strong), and skills relating to the self-regulation of one's cognitive processes on the other (e.g., planning a mathematical solution process; monitoring an ongoing solution process; evaluating and, if necessary, debugging a solution; reflecting on one's mathematics learning and problem-solving activities).

(d) Affective components involving beliefs about mathematics (e.g., believing that solving a mathematics problem requires effort versus believing that it is a matter of luck), attitudes (e.g., liking versus disliking story problems), and emotions (e.g., satisfaction when one finds the solution to a difficult problem).

It is certainly useful to distinguish these four categories of components, but it is also important to realize that in expert mathematical cognition they are applied integratively and interactively. For example, discovering the applicability of a heuristic to solve a geometry problem is generally based, at least partially, on one's conceptual knowledge about the geometrical figures involved. A negative illustration is that some beliefs observed in students, such as 'solving a math problem should not last more than a few minutes', will inhibit a thorough heuristic and metacognitive approach to a difficult mathematics problem. But, while this integration of the different components is necessary, it is not yet a sufficient condition to overcome the phenomenon of inert knowledge observed in many students: although the relevant knowledge is often available and can even be recalled on request, students do not spontaneously apply it in situations where it is appropriate to solve a new math problem. In other words, acquiring competence in mathematics involves more than the mere sum of the four components listed above. As a further elaboration of this view, the notion of a 'mathematical disposition' introduced in the *Curriculum and Evaluation Standards for School Mathematics* (National Council for Teachers of Mathematics 1989, p. 233) in the USA, points to the integrated availability and application of the different components:

> Learning mathematics extends beyond learning concepts, procedures, and their application. It also includes developing a disposition toward mathematics and seeing mathematics as a powerful way for looking at situations. Disposition refers not simply to attitudes but to a tendency to think and to act in positive ways. Students' mathematical dispositions are

manifested in the way they approach tasks—whether with confidence, willingness to explore alternatives, perseverance, and interest—and in their tendency to reflect on their own thinking.

According to Perkins et al. (1993), the notion of disposition involves, besides ability, inclination and sensitivity; the latter two aspects are essential in view of overcoming the phenomenon of inert knowledge. Inclination is the tendency to engage in a given behavior due to motivation and habits; sensitivity refers to the feeling for, and alertness to, opportunities for implementing the appropriate behavior. Ability, then, combines both the knowledge and the skill—in other words, most of the components mentioned above—to deploy that behavior. The acquisition of a disposition—especially the sensitivity and inclination aspects—requires extensive experience with the different categories of mathematical knowledge and skills in a large variety of situations. As such, a mathematical disposition cannot be directly taught, but has to develop over an extensive period of time.

3. *Mathematics Learning as the Construction of Meaning in Sociocultural Contexts*

The question arises, then, as to what kind of learning processes are conducive to the attainment of the intended mathematical disposition in students. The negative answer seems to be that this disposition cannot be achieved through learning as it occurs predominantly in today's classrooms. Indeed, the international literature bulges with findings indicating that students in our schools are not equipped with the necessary knowledge, skills, beliefs, and motivation to approach new mathematical problems and learning tasks in an efficient and successful way (see, e.g., De Corte 1992). This can largely be accounted for by the prevailing learning activities in today's schools, consisting mainly of listening, watching, and imitating the teacher and the textbook. In other words, the dominating view of learning in the practice of mathematical education is still the information-transmission model: the mathematical knowledge acquired and institutionalized by past generations has to be transmitted as accurately as possible to the next generation (Romberg and Carpenter 1986).

An additional shortcoming of current mathematical education, which is related to the inappropriate view of learning as information absorption, is that knowledge is often acquired independently from the social and physical contexts from which it derives its meaning and usefulness. This has become very obvious through a substantial amount of research carried out since the mid-1980s on the influence of cultural and situational factors on mathematics learning, and commonly classified under the heading 'ethnomathematics and everyday mathematical cognition' (Nunes 1992). For example, a series of investigations on so-called 'street mathematics' has shown that a gap often exists between formal school mathematics and the informal mathematics applied to solve everyday, real-life problems.

The preceding description of the learner as an absorber and consumer of decontextualized mathematical knowledge contrasts sharply with the conception supported by a substantial amount of evidence in the literature showing that learning is an active and constructive process. Learners are not passive recipients of information; rather, they actively construct their mathematical knowledge and skills through interaction in meaningful contexts with their environment, and through reorganization of their prior mental structures.

Although there are conceptual differences along the continuum from radical to realistic constructivism, the idea is broadly shared that learning is also a social process through which students construct mathematical knowledge and skills cooperatively; opportunities for learning mathematics occur during social interaction through collaborative dialog, explanation and justification, and negotiation of meaning (Cobb and Bauersfeld 1995). Research on small-group learning supports this *social constructivist* perspective: cooperative learning can yield positive learning effects in both cognitive and social-emotional respects. However, it has also become obvious that simply putting students in small groups and telling them to work together is not a panacea; it is only under appropriate conditions that small-group learning can be expected to be productive. Moreover, stressing the social dimension of the construction of knowledge does not exclude the possibility that students also develop new knowledge and skills individually. In addition, most scholars share the assumption of the so-called *cultural constructivist* perspective that active and constructive learning can be mediated through appropriate guidance by teachers, peers, and cultural artifacts such as educational media.

4. *Designing Powerful Teaching–Learning Environments*

Taking into account the view of mathematical learning as the construction of meaning and understanding, and the goal of mathematics education as the acquisition of a mathematical disposition involving the mastery of different categories of knowledge and skills, a challenging task has to be addressed. It consists of elaborating a coherent framework of research-based principles for the design of powerful teaching–learning environments, i.e., situations and contexts which can elicit in students the learning activities and processes that are conducive to the intended mathematical disposition.

A variety of projects attempting the theory-based design of powerful mathematics learning environments has already been carried out (see De Corte et al. 1996 for a selective overview), reflecting the methodological shifts toward the application of teaching experiments in real classrooms and toward the use of a diversity of techniques for data collection and analysis, including qualitative and interpretative methods.

For example, Lampert (1986) has designed a learning environment that aims at promoting meaning construction and understanding of multiplication in fourth graders by connecting and integrating principled conceptual knowledge (e.g., the principles of additive and multiplicative composition, associativity, commutativity, and the distributive property of multiplication over addition) with their computational skills. She starts from familiar and realistic problems to allow children to use and explore their informal prior knowledge, and practices collaborative instruction whereby she engages in cooperative work and discussion with the whole class. Students are solicited to propose and invent alternative solutions to problems that are then discussed, including their explanation and justification. It is obvious that this learning environment also embodies a classroom climate and culture that differs fundamentally from what is typical of traditional mathematics lessons.

A second and more comprehensive example is Realistic Mathematics Education (RME) developed in the Netherlands. RME, already initiated by Freudenthal in the 1970s, conceives mathematics learning essentially as doing mathematics starting from the study of phenomena in the real world as topics for mathematical modeling, and resulting in the reinvention of mathematical knowledge. Based on this fundamental conception of doing mathematics the design of 'realistic' learning environments is guided by a set of five inter-related principles: (a) learning mathematics is a constructive activity; (b) progressing toward higher levels of abstraction; (c) encouraging students' free production and reflection; (d) learning through social interaction and cooperation; and (e) interconnecting knowledge components and skills (Treffers 1987).

These and other representative examples (see, for instance, Cobb and Yackel 1998, Cognition and Technology Group at Vanderbilt 1997, Fennema and Romberg 1999, Verschaffel et al. 1999) illustrate efforts within the domain of research on mathematics learning and teaching to implement innovative educational settings embodying to some degree ideas that have emerged from theoretical and empirical studies, such as the constructivist view of learning, the conception of mathematics as human activity, the crucial role of students' prior—informal as well as formal—knowledge, the orientation toward understanding and problem solving, the importance of social interaction and collaboration in doing and learning mathematics, and the need to embed mathematics learning into authentic and meaningful contexts, as well as to create a new mathematics classroom culture. The results of those projects reported so far are promising as they demonstrate that this kind of learning environment can lead to fundamental changes in the sort of mathematics knowledge, skills, and beliefs that children acquire, and to making them more autonomous learners and problem solvers. However, these projects also raise questions for future research. For example, there is a strong need for additional theoretical and empirical work aiming at a better understanding and fine-grained analysis of the acquisition processes that this type of learning environment elicits in students, of the precise nature of the knowledge and beliefs they acquire, and of the critical dimensions that can account for the power of such environments.

5. *Constructing Innovative Forms of Assessment Instruments*

To support the implementation in educational practice of the innovative approach to mathematics learning and teaching, and especially to evaluate the degree of attainment of the new goals of mathematics education, appropriate forms of assessment instruments are required. In this respect, traditional techniques of educational testing, predominantly based on the multiple-choice item format, have been severely criticized as being inappropriate for evaluating students' achievement of the intended objectives of mathematics education. As a consequence, an important line of research has been initiated since the early 1990s aiming at the development of alternative forms of assessment which are tailored to the new conception of the goals and nature of mathematics learning and teaching, and which reflect more complex, real-life or so-called authentic performances (Romberg 1995). At the same time the need for a better integration of assessment with teaching and learning, as well as the importance of assessment instruments to yield information to guide further learning and instruction have been emphasized.

See also: Environments for Learning; Gender and School Learning: Mathematics and Science; Instructional Psychology; Mathematical and Logical Abilities, Neural Basis of; Science Education; Spatial Cognition

Bibliography

Bishop A J, Clements K, Keitel C, Kilpatrick J, Laborde C 1996 *International Handbook of Mathematics Education*. Kluwer, Dordrecht, The Netherlands

Cobb P, Bauersfeld H (eds.) 1995 *The Emergence of Mathematical Meaning*. Erlbaum, Hillsdale, NJ

Cobb P, Yackel E 1998 A constructivist perspective on the

culture of the mathematics classroom. In: Seeger F, Voigt J, Waschescio U (eds.) *The Culture of the Mathematics Classroom*. Cambridge University Press, Cambridge, UK, pp. 158–90
Cognition and Technology Group at Vanderbilt 1997 *The Jasper Project: Lessons in Curriculum, Instruction, Assessment, and Professional Development*. Erlbaum, Mahwah, NJ
De Corte E 1992 On the learning and teaching of problem-solving skills in mathematics and LOGO programming. *Applied Psychology: An International Review* **41**: 317–31
De Corte E, Greer B, Verschaffel L 1996 Mathematics teaching and learning. In: Berliner D C, Calfee R (eds.) *Handbook of Educational Psychology*. Macmillan Library Reference, New York, pp. 491–549
Fennema E, Romberg T A (eds.) 1999 *Mathematics Classrooms that Promote Understanding*. Erlbaum, Mahwah, NJ
Grouws D A (ed.) 1992 *Handbook of Research on Mathematics Teaching and Learning*. Macmillan, New York
Lampert M 1986 Knowing, doing, and teaching multiplication. *Cognition and Instruction* **3**: 305–42
National Council of Teachers of Mathematics 1989 *Curriculum and Evaluation Standards for School Mathematics*. National Council of Teachers of Mathematics, Reston, VA
Nunes T 1992 Ethnomathematics and everyday cognition. In: Grouws D A (ed.) *Handbook of Research on Mathematics Teaching and Learning*. Macmillan, New York, pp. 557–74
Perkins D N, Jay E, Tishman S 1993 Beyond abilities: A dispositional theory of thinking. *The Merrill-Palmer Quarterly* **39**: 1–21
Romberg T A (ed.) 1995 *Reform in School Mathematics and Authentic Assessment*. State University of Albany, Albany, NY
Romberg T A, Carpenter T P 1986 Research on teaching and learning mathematics: Two disciplines of scientific inquiry. In: Wittrock M (ed.) *The Third Handbook of Research on Teaching*. Macmillan, New York, pp. 850–73
Schoenfeld A H 1992 Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In: Grouws D (ed.) *Handbook of Research on Mathematics Learning and Teaching*. Macmillan, New York, pp. 334–70
Steffe L P, Nesher P, Cobb P, Goldin G A, Greer B (eds.) 1996 *Theories of Mathematical Learning*. Erlbaum, Mahwah, NJ
Treffers A 1987 *Three dimensions. A Model of Goal and Theory Description in Mathematics Instruction. The Wiskobas Project*. D. Reidel Publishing, Dordrecht, The Netherlands
Verschaffel L, De Corte E, Lasure S, Van Vaerenbergh G, Bogaerts H, Ratinckx E 1999 Learning to solve mathematical application problems: A design experiment with fifth graders. *Mathematical Thinking and Learning* **1**: 195–229

E. De Corte and L. Verschaffel

Mathematical Learning Theory

'Mathematical learning theory' usually refers to mathematical formalizations of theories for simple 'associative learning' situations such as conditioning, discrimination learning, concept learning, category learning, paired-associate learning, or list learning. By contrast, 'mathematical learning theory' does not typically refer to formal models of language acquisition or machine-learning algorithms and artificial intelligence. Many psychologists think of the pioneering research done in the 1950s and 1960s as representative of mathematical learning theory (see Atkinson et al. 1965, Bower 1994, Bush and Estes 1959, Bush and Mosteller 1955, Coombs et al. 1970, Shepard 1992, but mathematical models of learning have been actively investigated ever since, albeit not always under the same rubric. For a description of the historical context and significance of mathematical learning theory, see *Mathematical Learning Theory, History of*, written by one of the field's preeminent contributors, W. K. Estes.

1. Illustrative Example: Blocking of Associative Learning

One of the most influential models of associative learning was developed by Rescorla and Wagner (1972, cf. Miller et al. 1995, Siegel and Allan 1996), largely in response to the phenomenon called 'blocking' of associative learning, discovered by Kamin (1969).

1.1 Experimental Paradigm and Empirical Effect

Blocking has been found in many species and for many different types of learning. For simplicity of exposition, consider a situation in which a person must learn to diagnose sets of symptoms. In a standard laboratory setting, a learning trial consists of the following sequence of events: the person is shown a list of symptoms; he or she makes a diagnosis; the correct diagnosis is then displayed. After many such trials, the person's accuracy improves. In general, the features to be learned about, such as symptoms, are called 'cues,' and the responses to be made, such as diagnoses, are called 'outcomes.' The strength of the learned associations from cues to outcomes can be measured by the person's subjective rating (e.g., on a scale from 0 to 10, rate how strongly you think this outcome is indicated), or by the objective probability that the person chooses each outcome over repeated trials. Other laboratory procedures use different measures of learning.

When two cues are consistently followed by a certain outcome, each of the two cues acquires moderate associative strength with the outcome. When the well-trained learner is presented with the two cues together, a full-magnitude response is evoked, but when the learner is presented with either cue alone, a somewhat weaker response is evoked. The main point is that both cues do evoke a response.

However, suppose that prior to learning about the pair of cues, the learner is trained with only one of the cues. Subsequently, the second cue is included so that

training with the pair of cues proceeds as previously described. In this situation, it turns out that the added cue gains relatively little associative strength with the outcome. This lack of learning occurs despite the fact that the cue is perfectly indicative of the imminent occurrence of the outcome. Learning about this redundant relevant cue has apparently been prevented, or 'blocked,' by previous learning about the first cue. This blocking effect demonstrates that associative learning does not merely depend on spatiotemporal contiguity of cue and outcome.

1.2 Explanatory Principle

One popular explanation of blocking is that learning only occurs when an outcome is surprising, i.e., unexpected (Kamin 1969). In the second phase of training, the first cue already predicts the outcome because of previous learning, and so there is little surprise when the outcome does in fact occur. Consequently, there is little learning about the redundant relevant cue.

1.3 Formal Expression in a Mathematical Model

The principle of surprise-driven learning was formalized by Rescorla and Wagner (1972). A slight variant of their original model is described here, with an updated notation, but the essentials are preserved. The presence or absence of the ith cue is denoted by c_i, which has value 1 if the cue is present, and which otherwise has value 0. The associative strength between the ith cue and the kth outcome is denoted s_{ki}. This associative strength begins with a value of zero and changes through learning, described later. The predicted magnitude of the kth outcome, denoted p_k, is assumed to be the sum of the associative strengths from whatever cues are present.

$$p_k = \sum_i s_{ki} c_i \tag{1}$$

The actual magnitude of the outcome is denoted by a_k, which has value 1 if the kth outcome is present and otherwise has value 0.

The degree to which the outcome is surprising is measured as the difference between the actual outcome and the predicted outcome. Changes in associative strengths are proportional to the surprisingness:

$$\Delta s_{ki} = \lambda \underbrace{(a_k - p_k)}_{\text{'surprise'}} c_i \tag{2}$$

where λ is a constant of proportionality called the 'learning rate.' Equation (2) implies that an associative strength only changes to the extent that the outcome is unpredicted, and only if the corresponding cue is present.

1.4 Prediction: Simulation or Derivation

Predictions from the model can be ascertained from computer simulation or, in some cases, from mathematical derivation. It will now be shown by explicit computation how the model accounts for blocking. For purposes of this illustration, the learning rate λ in Eqn. (2) is arbitrarily set to 0.5. Suppose the learner first trains on two trials of cue i accompanied by outcome k. The first trial begins with $s_{ki} = 0.0$. According to Eqn. (2), the association from i to k has its strength changed by an amount $\Delta s_{ki} = \lambda(a_k - p_k)c_C = 0.5(1-0)1 = 0.5$. The strength after the first trial is therefore $s_{ki} + \Delta s_{ki} = 0.0 + 0.5 = 0.5$. The second trial of cue i accompanied by outcome k produces another change: $\Delta s_{ki} = \lambda(a_k - p_k)c_A = 0.5(1-0.5) \quad 1 = 0.25$. Adding this change to the strength which began the trial, the resulting associative strength after two trials of learning is $s_{ki} + \Delta s_{ki} = 0.5 + 0.25 = 0.75$.

The model is next trained on two trials of cues i and j, together leading to the same outcome k. The third trial begins with $s_{ki} = 0.75$ and $s_{kj} = 0.0$. According to Eqn. (2), the changes in the associative strengths are $\Delta s_{ki} = \lambda(a_k - p_k)c_i = 0.5(1-0.75)1 = 0.125$ and $\Delta s_{kj} = \lambda(a_k - p_k)c_j = 0.5(1-0.75)1 = 0.125$. Therefore, the third trial ends with $s_{ki} = 0.875$ and $s_{kj} = 0.125$. On the fourth trial, cues i and j are again presented together, and the model perfectly predicts the outcome via Eqn. (1):

$$p_k = \Sigma_i s_{ki} c_i = (0.875)(1) + (0.125)(1) = 1 = a_k$$

Therefore, the associative strengths do not change, because there is no surprise.

If the model had not been initially trained with cue i by itself predicting the outcome, then the associative strengths would have ended up as $s_{ki} = 0.5$ and $s_{kj} = 0.5$. Notice that the associative strength of s_{kj} is stronger in this case than it was after previous training with cue i by itself. The model has therefore successfully predicted that the strength of association from cue j to outcome k is blocked by previous training with cue i.

For some scenarios, the values of the associative strength can be derived mathematically instead of computed numerically trial by trial. For example, if cue i leads to outcome k on proportion P of trials (and if this is the only association being learned), it can be proved that s_{ki} converges to P as training progresses (and as λ grows smaller). If we assume that the strength has reached a stable value so that $\Delta s_{ki} = 0$, then Eqn. (2) implies that $0 = P\lambda\ (1 - s_{ki}) + (1-P)\lambda(0 - s_{ki})$, which reduces to $s_{ki} = P$.

1.5 Fit to Data and Parameter Estimation: Derivation or Search

As presented in Sect. 1.4, the Rescorla–Wagner model has one free parameter, the learning rate λ in Eqn. (2). The exact degree of blocking exhibited by the model depends on the value of this learning rate. The higher the learning rate, the greater the blocking. When attempting to quantitatively fit empirical data, the learning rate is set to best fit the data. There are many methods by which parameter values can be estimated. Typically, a computer program searches many parameter values by trial and error until a nearly optimal value is found. However, in some situations, the best fitting parameter value can be derived by mathematical analysis. In fitting the model to data, additional assumptions are needed to map the model predictions, p_k, to the experimental measurements, such as ratings or choice proportions. These additional mappings can also entail free parameters that must be estimated to best fit the data.

2. Theoretical Issues in Learning

Any mathematical theory of cognitive processes begins with formal representations of (a) the stimuli impinging on the cognizer and (b) the responses emanating from it. In the Rescorla–Wagner model described in Sect. 1, the stimuli are trivially represented by the 1 or 0 values of c_i and a_k and the responses are represented by the predicted magnitudes p_k. Any mathematical theory also posits formal transformations that map the stimulus representation to the response representation. These transformations typically include additional formal constructs, also known as internal representations and processes, that are supposed to be formal descriptions of cognitive processes. In the Rescorla–Wagner model, for example, the associative strengths s_{ki} are internal representations of memory, and Eqn. (2) expresses the process that changes the internal representation. (These representations and processes exist in the model, but it is a separate philosophical issue as to whether homologous representations literally exist in the mind.) Theoretical issues in mathematical learning theory therefore revolve around alternative representations, processes, and their rationales.

2.1 What is Learned (the Representation)

Models of learning must specify a formal representation of the stimuli. In various theories, a stimulus has been represented by a single number (as in the Rescorla–Wagner example in Sect. 2), as a set of elements (e.g., stimulus sampling theory, Estes, 1950), as a vector of random components (used in many models of memory), as a vector of activations across meaningful features (e.g., McClelland and Rumelhart 1986, Rumelhart and McClelland 1986), or as a point in a multi-dimensional psychological space (e.g., Kruschke 1992). The choice of stimulus representations can have consequences for the formal workings of the internal representations and processes of the model, and for the ability of the model to fit empirical learning data.

Models also vary in their assumptions about the internal representations that are learned. Exemplar models assume that the learner can store every unique stimulus. Such models do not necessarily imply that every exemplar can be distinctly recalled or recognized, rather, these models simply assume that internalized exemplar information is used, perhaps imperfectly, for generating responses (e.g., Estes 1991). Non-exemplar models assume that internal representations are summaries of sets of instances. In prototype models, a set of stimuli might be represented by their central tendency on each component. Other models assume that a set of stimuli is represented by an ideal or caricature. Yet another type of abstraction is rule-based representation, which posits necessary and sufficient stimulus conditions for emitting specific responses. Unlike prototypes or exemplars, the conditions for a rule are typically assumed to be strictly satisfied or not satisfied. There is little or no gradation for partially matched stimuli.

2.2 How it is Learned (the Process)

Along with each type of internal representation, there are processes that use the representation to generate responses, and there are processes that adapt the representation as a consequence of new learning. Perhaps the simplest possible learning process is a probabilistic all-or-nothing change from no association to perfect association. Another simple learning process is a fixed-size increment in associative strength between two items whenever they co-occur. However, these simple processes cannot account for the phenomenon of blocking, and it is largely for this reason that the classic models of the 1950s and 1960s have not been more strongly emphasized here. As described earlier, blocking can be accounted for by the process of surprise-driven learning, formalized in Eqn. (2).

In non-rule models, responses are generated by computing the similarity of the stimulus to the various internalized exemplars or prototypes. For a stimulus composed of cues c_j, and for an exemplar or prototype specified by associative strengths s_{kj}, the similarity of the stimulus to the internal representation can be formally expressed as

$$\mathrm{Sim}_k = \exp\left(-\frac{1}{2}\sum_j (s_{kj} - c_j)^2 \right) \tag{3}$$

Equation (3) merely describes a bell-shaped function, centered at the exemplar or prototype. The similarity is maximized (and has value 1.0) when the stimulus components exactly match the internal representation. To the extent that a stimulus is similar to an exemplar or prototype, the response associated with that exemplar or prototype is generated. For an additional discussion of the role of similarity in theories of categorization, see *Categorization and Similarity Models*.

In prototype models, the internal representation of the prototype is gradually learned. One way to make a prototype representation move toward the central tendency of a set of stimuli is to move the prototype a fraction of the distance toward a stimulus when that stimulus occurs. Because outlying stimuli should presumably be represented by other prototypes, the extent to which a prototype moves toward a stimulus should be proportional to the similarity of the prototype with the stimulus. The learning process can then be formalized as follows:

$$\Delta s_{ki} = \lambda \, \mathrm{Sim}_k (c_i - s_{ki}) \tag{4}$$

where λ is, as before, a constant called the learning rate.

Rule-learning models typically assume that the learner hypothesizes one candidate rule at a time, tests the rule for adequacy on successive trials, and retains the rule as long as it works but rejects the rule if it fails. Historically, this kind of learning is called 'hypothesis testing' with a 'win-stay, lose-shift' strategy. Numerous mathematical analyses and empirical tests of these models have been conducted (for summaries, see Levine 1975). More recent rule-based models include mechanisms for gradually tuning the rule conditions, mechanisms for handling exceptions to rules, and mechanisms for combining rules with other representations (e.g., Ashby et al. 1998, Busemeyer and Myung 1992, Erickson and Kruschke 1998, Nosofsky et al. 1994, Vandierendonck 1995).

2.3 Why it is Learned (the Rationale)

The formalisms of the representation and process are supposed to be expressions of explanatory psychological mechanisms. These mechanisms can sometimes be motivated by considerations of how a learner might maximize accuracy, store information efficiently, or learn rapidly. (For an approach to learning and memory from the perspective of optimality different from the following examples (see Anderson 1990).)

The surprise-driven learning process expressed in Eqn. (2) can be motivated by assuming that the learner is attempting to reduce overall error, defined as $E = \frac{1}{2}\Sigma_r (a_r - p_r)^2$, as quickly as possible. Thus, if the association strengths are adjusted proportionally to the (negative of the) gradient of the error, the following formula is implied for changing the strengths:

$$\begin{aligned} \Delta s_{ki} &= -\lambda \frac{\partial E}{\partial s_{ki}} \\ &= -\lambda \frac{1}{2} \sum_r 2(a_r - p_r) \frac{\partial p_r}{\partial s_{ki}} \\ &= -\lambda \, (a_k - p_k) c_i \end{aligned} \tag{5}$$

Notice that Eqn. (5), derived from gradient descent on error, is identical to Eqn. (2), which was a direct formalization of surprise-driven learning. For a generalization of this approach, see *Artificial Neural Networks: Neurocomputation* and *Perceptrons*.

Analogously, the formula for learning of central tendencies (Eqn. (4)) can be derived by assuming that the learner is attempting to maximize the similarity of the prototype to the exemplars. To achieve this goal, the components of the prototype, s_{ki}, should be adjusted proportionally to the gradient of the similarity:

$$\begin{aligned} \Delta s_{ki} &= \lambda \frac{\partial \, \mathrm{Sim}_k}{\partial s_{ki}} \\ &= \lambda \frac{\partial}{\partial s_{ki}} \exp \left(-\frac{1}{2} \sum_j (s_{kj} - c_j)^2 \right) \\ &= \lambda \, \mathrm{Sim}_k \left(-\frac{1}{2} \sum_j 2 (s_{kj} - c_j) \frac{\partial s_{kj}}{\partial s_{ki}} \right) \\ &= \lambda \, \mathrm{Sim}_k \, (c_i - s_{ki}) \end{aligned} \tag{6}$$

Notice that the formula derived in Eqn. (6) is the same as the formula in Eqn. (4). Learning algorithms such as these are sometimes referred to as 'competitive learning' or 'clustering' techniques. For additional information about related approaches to concept learning, see *Connectionist Models of Concept Learning*.

3. Future Trends

3.1 Hybrid Representation Models

Evidence is gradually accumulating that learning by humans, and perhaps by other animals, cannot be fully described with a single representational scheme. Instead, models that incorporate a hybrid of representations are needed. The challenge to theorists is

specifying how the representational options are allocated during learning. This is both an empirical issue and a theoretical one. Experimenters must design experiments that assay the conditions under which learners use each representation, and theorists must invent rationales for the observed representational allocations. Erickson and Kruschke (1998), for example, reported evidence that people use both rules and exemplars, and the researchers described a hybrid model that allocated learning to rules or exemplars depending on which reduced error most quickly. It is likely that future models will also incorporate prototypes or ideals as evidence accumulates for these representational schemes. For additional discussion of types of representation in category, learning, see *Concept Learning and Representation: Models*.

3.2 Rapidly Shifting, Learned Attention

The critical problems confronting a learner are acquisition of many associations from as few examples as possible, without destroying previously learned associations, and doing so in an environment in which there are many potentially relevant cues to be learned about. A natural solution to these problems is to selectively attend to those cues that are most strongly correlated with the desired outcome, and to selectively ignore those cues that are already associated with other outcomes. This selective attention reduces 'overwriting' of previously learned associations, and facilitates the rapid acquisition of new associations. These shifts of attention to particular cues for particular situations is itself a learned response. Numerous perplexing phenomena in human and animal learning can be explained by models of learning that incorporate rapidly shifting, learned selective attention (e.g., Mackintosh 1975, Kruschke 1996, Kruschke and Johansen 1999). In particular, it has been shown that blocking (see Sect. 1) involves learned inattention to the blocked cue, not just surprise-driven associative strength changes (Kruschke and Blair 2000). The allocation of representations to different situations can also be modeled as changes in attention (Erickson and Kruschke 1998).

3.3 Unification with Memory Models

Although mathematical learning theory began in the 1950s and 1960s with similar approaches to research in associative learning and memory, the following decades have seen a divergence of research in these specialized areas. Models of associative learning have emphasized how people and animals learn contingencies between cues and outcomes. Models of memory have emphasized how people encode and retrieve stimuli in particular contexts. Presumably the latter situation is a type of associative learning between stimulus cues and contextual cues. In the near future, mathematical models of memory will need to be unified with mathematical models of associative learning. For more about memory models see *Memory Models: Quantitative*.

3.4 Integration with Neuroscience and Across Species

Mathematical models of learning describe functional algorithms which might be implemented on neural machinery in a variety of different ways. Models that accurately predict behavior can help guide brain research, and a detailed understanding of neural circuitry can help refine the models. With recent advances in brain imagery techniques, this symbiotic relationship between modeling and neuroscience is sure to grow. See *Categorization and Similarity Models: Neuroscience Applications*.

See also: Artificial Neural Networks: Neurocomputation; Categorization and Similarity Models; Categorization and Similarity Models: Neuroscience Applications; Computational Neuroscience; Concept Learning and Representation: Models; Connectionist Models of Concept Learning; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Mathematical Learning Theory, History of; Memory Models: Quantitative; Perceptrons

Bibliography

Anderson J R 1990 *The Adaptive Character of Thought*. Erlbaum, Hillsdale, NJ

Ashby F G, Alfonso-Reese L A, Turken A U, Waldron F M 1998 A neuropsychological theory of multiple systems in category learning. *Psychological Review* **105**: 112–81

Atkinson R C, Bower G H, Crothers E J 1965 *An Introduction to Mathematical Learning Theory*. Wiley, New York

Bower G H 1994 A turning point in mathematical learning theory. *Psychological Review* **101**: 290–300

Busemeyer J R, Myung I J 1992 An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General* **121**: 177–94

Bush R R, Estes W K (eds.) 1959 *Studies in Mathematical Learning Theory*. Stanford University Press, Stanford, CA

Bush R R, Mosteller F 1955 *Stochastic Models for Learning*. Wiley, New York

Coombs C H, Dawes R M, Tversky A 1970 *Mathematical Psychology: An Elementary Introduction*. Prentice-Hall, Englewood Cliffs, NJ, pp. 256–306 (Reprinted in 1981 by: Mathesis Press, Ann Arbor, MI)

Erickson M A, Kruschke J K 1998 Rules and exemplars in category learning. *Journal of Experimental Psychology: General* **127**: 107–40

Estes W K 1950 Toward a statistical theory of learning. *Psychological Review* **57**: 94–107

Estes W K 1991 *Classification and Cognition*. Oxford University Press, New York
Kamin L J 1969 Predictability, surprise, attention, and conditioning. In: Campbell B A, Church R M (eds.) *Punishment*. Appleton-Century-Crofts, New York, pp. 279–96
Kruschke J K 1992 ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* **99**: 22–44
Kruschke J K 1996 Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition* **22**: 3–26
Kruschke J K, Blair N J 2000 Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review* (In press)
Kruschke J K, Johansen M K 1999 A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition* **25**: 1083–119
Levine M 1975 *A Cognitive Theory of Learning: Research on Hypothesis Testing*. Erlbaum, Hillsdale, NJ
Mackintosh N J 1975 A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review* **82**: 276–98
McClelland J L, Rumelhart D E (eds.) 1986 *Parallel Distributed Processing. Vol. 2: Psychological and Biological Models*. MIT Press, Cambridge, MA
Miller R R, Barnet R C, Grahame N J 1995 Assessment of the Rescorla–Wagner model. *Psychological Bulletin* **117**: 363–86
Nosofsky R M, Palmeri T J, McKinley S O 1994 Rule-plus-exception model of classification learning. *Psychological Review* **101**: 53–79
Rescorla R A, Wagner A R 1972 A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In: Black A H, Prokasy W F (eds.) *Classical Conditioning: II. Current Research and Theory*. Appleton-Century-Crofts, New York, pp. 64–99
Rumelhart D E, McClelland J L (eds.) 1986 *Parallel Distributed Processing. Vol. 1: Foundations*. MIT Press, Cambridge, MA
Shepard R N 1992 The advent and continuing influence of mathematical learning theory: Comment on Estes and Burke. *Journal of Experimental Psychology: General* **121**: 419–21
Siegel S, Allan L G 1996 The widespread influence of the Rescorla–Wagner model. *Psychonomic Bulletin & Review* **3**: 314–21
Vandierendonck A 1995 A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin and Review* **2**: 442–59

J. K. Kruschke

Mathematical Learning Theory, History of

Mathematical learning theory in the 1990s is a heterogeneous collection of models having the common theme of expressing basic concepts and assumptions about learning in mathematical form and deriving their empirically testable implications by mathematical reasoning or computer simulations. The history of this body of theory is traced from its beginnings in one of the earliest laboratories of experimental psychology about 1880 through a long period of fixation on the mathematization of learning curves to a radical expansion in the mid 1950s. Mathematical modelling now serves an important role in the direction and interpretation of research across the entire spectrum of learning studies (see also *Mathematical Learning Theory*).

1. *The Earliest Precursor of Mathematical Learning Theory*

The first small step toward mathematical learning theory occurred in the course of a massive experiment on the learning and retention of lists of artificial words ('nonsense syllables') conducted by the founder of the experimental study of memory, H. Ebbinghaus, during the period 1879–80. The learning of each list to a criterion was followed after an interval by a memory test. In plots of amount remembered (denoted b) vs. time between learning and testing (t), b declined over time in a reasonably orderly fashion but with some fluctuations. Ebbinghaus noted that, in view of the fluctuations, the empirical values of b could not reveal a law of retention, but he hypothesized that a smooth curve made to pass through the choppy empirical graph might reveal the underlying trend. In fact, a curve representing the function

$$b = 100k/k(\log t)^c \tag{1}$$

where k and c are constants and $\log t$ is the logarithm of the time between learning and testing, proved to yield an excellent account of the retention data.

This finding must have been satisfying, but Ebbinghaus did not overemphasize what he had accomplished, 'Of course this statement and the formula upon which it rests have no other value than that of a shorthand statement of the above results … Whether they possess a more general significance I cannot at the present time say' (Ebbinghaus 1885/1964, p. 78).

2. *The Mathematical Description of Learning Functions*

2.1 *Fitting the Learning Curve*

Discovering how to demonstrate some general significance for mathematical functions fitted to empirical learning and retention curves was to be the central task of the next half century, during which the pace of relevant work gradually accelerated. A review by H.

Gulliksen (1934) identified two studies in the decade 1901–10, followed by four in 1911–20, and 11 in 1921–30. in J. McGeoch's influential text on the psychology of human learning, published in 1942, an extensive chapter on learning curves was conspicuously placed immediately following the opening chapter on concepts and methods.

The early successors of Ebbinghaus followed his example in choosing equations purely for their ability to describe the forms of empirical curves. More interesting were some instances in which investigators started with analogies between learning and other processes. For example, the first of Ebbinghaus's successors in this strain of research, A. Schukarew adopted an exponential function that had appeared in chemistry as a descriptor of monomolecular chemical reactions. This effort led to no immediate sequel, but the approach was revived many years later with more fruitful results (Audley and Jonckhere 1956, Estes 1950, Thurstone 1930).

2.2 Seeking Rational Foundations for Equations of the Learning Curve

An important step toward providing a rational foundation for mathematical learning functions appeared in a study reported by L. L. Thurstone (1930). Thurstone's advance in methodology was to go beyond defining parameters of learning functions by fiat and, rather, to derive their properties from an underlying process model. He conceived performance to be made up of a population of elementary acts, some of which led to success and some to failure in a task and assumed that acquisition of a skill is the outcome of a random process in which acts leading to error are discarded from the population and successful acts strengthened. Further, Thurstone went beyond the fitting of his model to mean learning curves and showed how it could be used to interpret situations in which the empirical curves exhibited systematic irregularities.

For example, research of W. Kohler and R. M. Yerkes during the 1920s on problem solving in the great apes had found that curves for acquisition of the ability to solve a problem typically started with a sometimes lengthy plateau (period of no improvement) at a chance level of responding followed by an abrupt transition to perfect performance, suggesting that the animal had suddenly acquired insight concerning problem solution. Whether this interpretation was justified was a live issue for some years thereafter. Analyzing the issue in terms of his model, Thurstone concluded that abrupt transitions should occur when a rate parameter in his learning function had a sufficiently high value or a limit parameter (assumed to be related to task complexity) had a sufficiently low value. Thus the occurrence of abrupt transitions in learning curves would be predicted to occur under quite different conditions as a consequence of the same basic learning process and need not be taken as evidence for a qualitatively different process of insight.

Gulliksen (1934) continued the effort to rationalize learning functions with a forward-looking approach that emphasized the importance of treating data from individual learners, rather than only group averages, and the value of going beyond judgments of goodness of fit of equations to data and using estimates of parameters as bases for inferences about underlying processes.

3. The Transition from Learning Curves to Learning Theory

It might have seemed that by the time of the Gulliksen (1934) article, the pieces were all in place for a sharp acceleration of progress toward mathematical learning theory during the second half of the 1930s. In fact, however, the acceleration was delayed for more than 15 years, during which a continuing fixation on the learning curve was accompanied by a dearth of notable new developments.

A major factor contributing to this 'plateau' may have been the scantiness of communication between the psychometricians who were working on mathematization of learning curves and learning theorists who might have made use of their results. During the early 1930s, learning theory had abruptly emerged from obscurity and, under the leadership of C. L. Hull, E. R. Guthrie, B. F. Skinner, and E. C. Tolman and their followers, took center stage in the theoretical psychology of the period. With one exception, none of these theorists took quantitative theory as an objective, and some even argued that psychology was not ready for applications of mathematics. The exception was Hull, whose research group carried out a vigorous program of quantitatively oriented studies that set the stage for a dramatic broadening of mathematical approaches to learning phenomena.

Beyond stewarding this program, Hull almost singlehandedly defined the goals and the general character of a genuine mathematical learning theory. In an epoch-making monograph, Hull (1943) set forth a set of axioms for a general theory of conditioning and elementary learning, together with theorems presenting implications for many aspects of learning, including the way in which learned habits combine with motivation to determine performance, and the quantitative form of generalization from original learning situations to new test situations.

Hull's formal system, the output of a relatively short though intense period of activity in his late years, was programmatic in many aspects and thus was both technically untestable and difficult to work with by anyone who did not share all of his intuitions. Consequently, continuation of Hull's program for quantification of learning theory was left for a new

generation of theorists who came to the field in the 1950s with fresh outlooks and better technical backgrounds for formal theory construction.

3.1 Mathematical Learning Theory in the Mid-twentieth Century

With the stage set for accelerated progress by Hull's heroic efforts, new approaches to mathematical learning theory literally exploded on a number of fronts in the early 1950s. The events of this period have been well portrayed by Bower and Hilgard (1981) and the continuation in the 1960s by Greeno and Bjork (1973), who opened their review with the remark, 'In the decade or so that constitutes the period of our review, progress toward understanding the processes involved in learning can only be described in terms that sound extravagant' (p. 81). Concurrently, the range of empirical targets for modeling expanded drastically and the mode of application shifted from fitting mean performance curves to predicting the almost innumerable statistics describing the fine structure of data (Atkinson et al. 1965). Among the models emerging from this wave of new activity, that of Bush and Mosteller (1955) and that of Estes (1950) and his associates (including R. C. Atkinson and P. Suppes) have had enduring roles in the development of mathematical learning theory. However, the two approaches (fully discussed by Coombs et al. 1970) differ in their basic assumptions about learning.

In Bush and Mosteller's model, an organism's state of learning in a task at any time was assumed to be characterized by a set of response probabilities; and the effect of the events of a learning trial (e.g., reward or punishment) was expressed by a simple difference equation ('linear operator' in their terms) giving the value of response probability at the end of a trial as a function of the value on the preceding trial.

The model of Estes (1950) included difference equations of the same type, but their forms were derived from an underlying theoretical process in which the acquisition of information during learning was represented by changes in the partitioning of an array of abstract 'memory elements.'

The operator model has been attractive for the ease with which it can be applied to predict almost innumerable statistics of simple learning data, but the multilevel approach has become the norm for models of learning, most notably human concept and category learning, in cognitive psychology of the 1980s and 1990s (see *Concept Learning and Representation: Models*).

3.2 Concluding Comment

As the twenty-first century starts, the trends that led to the appearance of a diversified array of 'miniature' models of different aspects of learning beginning in the 1950s have not given way to any consolidation into a single comprehensive model. However, significant lines of cumulative theoretical development are identifiable. A notable example is the embodiment of ideas drawn from Hull's theory in the model of conditioning that has been the pace setter since the early 1970s (Rescorla and Wagner 1972) and recently has appeared as a constituent of new 'adaptive network' models that are flourishing in contemporary cognitive science (see, e.g., *Connectionist Approaches*; *Connectionist Models of Concept Learning*).

See also: Learning Curve, The

Bibliography

Atkinson R C, Bower G H, Crothers E J 1965 *An Introduction to Mathematical Learning Theory*. Wiley, New York

Audley R J, Jonckhere A R 1956 Stochastic processes and learning behaviour. *British Journal of Statistical Psychology* **9**: 87–94

Bower G H, Hilgard E R 1981 *Theories of Learning*, 5th edn. Prentice-Hall, Englewood Cliffs, NJ

Bush R R, Mosteller F 1955 *Stochastic Models for Learning*. Wiley, New York

Coombs C H, Dawes R M, Tversky A 1970 *Mathematical Psychology*. Prentice Hall, Englewood Cliffs, NJ

Ebbinghaus H 1885/1964 *Uber das Gedachnis (About Memory)*. Dover, New York

Estes W K 1950 Toward a statistical theory of learning. *Psychological Review* **57**: 94–107

Greeno J G, Bjork R A 1973 Mathematical learning theory and the new 'mental forestry.' *Annual Review of Psychology* **24**: 81–116

Gulliksen H 1934 A rational equation of the learning curve based on Thorndike's law of effect. *Journal of General Psychology* **11**: 395–434

Hull C L 1943 *Principles of Behavior*. Appleton-Century-Crofts, New York

Rescorla R A, Wagner A R 1972 A theory of Pavlovian conditioning: Variations in the effectiveness of reincorcement and nonreinforcement. In: Black A H, Prokasy W F (eds.) *Classical Conditioning II*. Appleton-Century-Crofts, New York, pp. 64–99

Thurstone L L 1930 The learning function. *Journal of General Psychology* **3**: 469–93

W. K. Estes

Mathematical Models in Geography

1. Introduction

This chapter explores the use of mathematical modeling in human geography. The focus is on providing an overview of the definition and historical

development of modeling in human geography, main research directions, methodological and theoretical issues, and a view of the future prospects of modeling. The alternative definitions of the term model and the evolution of modeling within geography are first briefly reviewed to provide historical context. This is followed by a typology of models in geography as a structured review of the different types of models and their main uses. Attention is then directed at a number of important challenges issues in geographic modeling. The chapter closes with a view of the future for modeling in human geography.

2. Definitions

The term 'model' has taken on multiple definitions within human geography. Moreover, as is discussed in Sect. 2, as the discipline has experienced paradigmatic shifts the status, role and definitions of models have also evolved and varied. As in many social sciences there has been a close connection between the concepts of a theory and a model within geography. Indeed, many of the theories within these disciplines could be more accurately described as models as they are often composed of a set of interrelated hypotheses rather than empirically validated laws (Cadwallader 1985).

A common view of a model within geography is a simplified representation of some aspect of reality. The representation abstracts the central components of interest to allow for an investigation of their role in the process under study. This representation can be articulated in one of three geographic model forms: (a) iconic, where representation is based on a reduction in scale, such as a classroom globe, (b) analogue, relying on both changes in scale and transformation of the properties of the system, e.g., use of a map, and (c) mathematical, in which formal symbols and equations are used to represent the system. The focus here is on this last form of geographical model.

As knowledge is developed through an on-going interaction between theory and observation, mathematical models play a critical role in this process, serving as an empirical counterpart of a theory which is compared against reality. This conventional view of a geographic model might be characterized as the 'model of a process' perspective. There is a second, somewhat more subtle but equally important view of the term model in geography, and this is the 'model as a process' perspective. Models are not simply about the world but are also about our knowledge about the world (Haines-Young 1989). In this way, the process of modeling can be seen to serve three useful purposes. First, it forces the researcher to be explicit concerning theoretical constructs, hypotheses, and other aspects of their understanding of a system. Second, formalization of a problem in an empirical model can be of great utility in revealing complex indirect effects that are not articulated by the theory. Finally, models can provide critical assistance in policy analysis where there is often a gap between existing substantive theory and the real world that policy-makers must confront.

3. Historical Development

In general terms, four broad periods of modeling in modern geography can be identified. A great deal has been written on this evolution and on the role of modeling in geography. Due to space limitations, this cannot be reviewed here. Interested readers are directed to Johnston (1991) for an overview. The first period begins in 1950 and ends in the mid 1960s; it is referred to as the 'quantitative revolution.' Quantitative geographers challenged the traditional approach to human geographic studies, which relied on descriptive analysis and the uniqueness of places. This challenge was led by younger, analytically oriented scholars who argued for a more scientific approach to the study of the spatial distribution of various characteristics of the surface of earth. A central component of that scientific approach was the use of quantitative methods.

The triad of scientific method, quantification, and models took on a peculiarly unique form within geography. As Johnston (1991) suggests, the dependence of models on mathematics coupled with the relatively few geographers who had strong mathematical training meant that little effort was directed towards the development of formal models of geographic reality. Rather, the main thrust of the quantitative revolution focused on multivariate statistical methods and hypothesis testing. Chief among these applications was the use of correlation methods to analyze the associations between different map patterns, as well as factor analytic methods used for a variety of applications, including defining and classification of geographical regions and factorial studies of urban socioeconomic structure. The availability of these and other multivariate methods provided a degree of precision in the analysis of spatial variation that was viewed as a major advancement for the field. While the number of mathematically trained geographers may have been modest, their influence on the practice of geographic modeling was not. Openshaw (1989) argues that these scholars used mathematical modeling to build impressively theoretical and rigorous models of spatial systems.

The second historical stage spans the mid 1960s to the mid 1970s, and was largely a consolidation of the gains made by modelers within the discipline. By the late 1960s, quantitative methods and modeling came to dominate the fields leading academic journals, and the proliferation of modeling and quantitative analysis was in large part responsible for the creation of new specialized journals, such as *Geographical Analysis* and *Environment and Planning A* in 1969.

The heyday of modeling was fairly short-lived, however, as the mid 1970s ushered in an era in human geography where the dominant research paradigm was one characterized by a preoccupation with sabotage and attack (Longley 1993). To a certain extent, these attacks were part of a larger negative reaction to positivism and the uses to which science was put by the political and economic powers of society (Gross and Levitt 1994). Within geography, the main criticisms against modeling and modelers included choosing irrelevant and inconsequential topics, creating highly artificial landscapes and behavioral agents, lacking substantive theoretical foundations, being a-historical, and ignoring the importance of local context.

By the end of this third period, modeling and its practitioners within geography no longer represented a majority position within the discipline. Indeed, the tensions between different modes of inquiry (e.g., humanistic versus scientific) had not been creative ones, and as a result, there was limited cross-fertilization between these approaches. A corollary of this is that much of the substantive theoretical basis for models in human geography has come from other social sciences (e.g., anthropology, economics, psychology, sociology), rather than from within the discipline.

The final historical period of modeling within geography covers the late 1980s to the beginning of the twenty-first century. This last phase is characterized by a revival of interest in geographic modeling, both within geography and other disciplines. Within geography, there is a growing recognition that a continued obsession with ontological and epistemological debates void of an operational methodology to validate the discipline's claims would make human geography less attractive to other social sciences (Sui 2000). This recognition has been reflected in calls for a reconsideration of the merits of integrating quantitative and qualitative approaches towards human geographic research (Lawson 1995) as well as in a growing realization of the important role that modeling and quantitative analysis have played in legitimizing human geography as a social science (King 1993).

Outside of geography, there has been a rediscovery of the importance of the spatial dimension in many social sciences (e.g., Krugman 1998). Part of this rediscovery has been driven by the adoption of recent geospatial technologies, such as geographic information systems (GIS), remote sensing and spatial statistics, in the empirical analyses within these disciplines. Equally important in this renaissance have been efforts that consider the role of geography and space in theoretical frameworks (Fujita et al. 1999).

4. *A Typology of Geographic Models*

Human geographers have employed mathematical models to study processes at widely ranging scales, from the level of individual location decision making to the analysis of broad patterns of trade flows between nation states. Consequently, cataloging these models is clearly beyond the scope of this chapter. Instead, I will categorize human geographic models according to two dimensions: (a) genre; and (b) applications.

4.1 *Spatial Interaction Models*

Spatial interaction modeling has been one of the most important strands of the modern human geographic modeling literature. Spatial interaction models illustrate how different locations are functionally interdependent. This can be seen by considering a simple interaction model:

$$T_{i,j} = \frac{P_i P_j}{d_{i,j}^{\beta}} \tag{1}$$

where the degree of spatial interaction (e.g., migration, trade flows, air travel) $T_{i,j}$ between place i and j is proportionate to the sizes of the population at the origin location P_i and at the destination P_j and inversely related to the distance between these locations $d_{i,j}$. The strength of the decline in interaction with increasing separation is captured by β, the distance decay coefficient. This reflects Tobler's First Law of Geography that 'everything is related to everything else, but near things are more related than distant things.' The process of spatial interaction is therefore a key mechanism for creating the spatial structures underlying human societies. At the same time, spatial interaction can also be influenced by the underlying spatial structures in societies.

4.2 *Urban Geographic Models*

Urban geographic models can be broadly classified into two groups. Models of urban spatial structure focus on the internal spatial organization of individual metropolitan areas, while models of urban systems typically take a national urban system as the unit of analysis and consider the distribution of individual cities within that system.

A rich diversity of models have been developed by urban geographers to capture the internal morphology of cities. These urban land use models attempt to explain empirical regularities in characteristics such as population and building densities, and land values within a given metropolitan area. Early attempts at building formal models of urban spatial structure borrowed heavily from the work of Alonso (1964) who developed the monocentric city model. This model was predicated on an assumed central business district (CBD), the point of maximum access for the surrounding urban area, and increasing transportation costs with movement away from the CBD. From these assumptions together with basic microeconomic

theory, this elegant model can be solved to generate gradients for land values and rents, as well as building and population densities that are negative in distance from the CBD.

From this classic framework, geographers have developed models of multi- or polycentric cities (Odland 1978). In these models the city may have several nodes of concentrated economic activity and the articulation of the role of accessibility in determining land use patterns in such a context results in a highly varied landscape. More recently, geographers have adopted novel fractal geometric approaches to studying urban form and structure (Batty and Longley 1994).

These analytically oriented efforts have led to a number of empirical simulation models of urban spatial structure that have been used for a variety of analyses. These urban models simulate urban development as it evolves dynamically over time and space by considering the interactions between individual choices and actions taken by households, developers, businesses and governments (Wegener 1998). Often these models are integrated with a GIS to provide a spatially explicit depiction of urban evolution.

Models of urban systems and their evolution focus on cities within either a region or nation. A central concern is the size distribution of cities within a system (Carroll 1982). These models consider the population sizes and rankings of a system's cities through the use of so-called rank-size curves which take the following general form:

$$\ln(r_i) = \alpha + \beta \ln(p_i) \tag{2}$$

where $\ln(r_i)$ is the natural logarithm of the rank of city i with respect to size and p_i is city i's population. Many empirical studies have found that a value for β close to 1 well describes many developed national city size distributions. As Fujita et al. (1999) have noted, part of the fascination with the rank size rule is that it comes as close to an empirical law as any relationship in the social sciences.

4.3 Regional Socioeconomic Models

In contrast to urban geographic models which have a finer level of spatial disaggregation, regional models are specified for broader spatial units, such as groupings of states, substate regions, or provinces. Many regional models focus on the economic structure and performance of urban and regional units. Economic base models posit that a city or regional economy can be divided into two components. One consists of those activities that produce for markets outside the region; these activities are thus classified as the 'export base.' The second component, referred to as the local sector, are those activities that produce for internal markets and provide goods and services for local residents and businesses. The central idea in these models is that local service employment is derived from the expansion of the export sector which identifies the importance of the export sector as a basis for economic growth in the region.

Regional input–output models can be viewed as an expansion of the economic base construct to consider a more detailed set of interindustry linkages within an urban or regional economy (Dewhurst et al. 1991). Like economic base models, the earliest regional input–output models focused on a single regional economy; however, input–output modelers have also extended these frameworks to examine the nature of inter-regional inter-industry interactions. These efforts have provided rich insights into the nature of modern space economies.

Regional econometric models represent a third type of macro geographical model used in the study of spatial economic systems. Initially analogues to national macroeconometric forecasting and simulation models, these models contain a set of stochastic behavioral and deterministic equations representing the main components of a region's socioeconomic system. What distinguishes the regional models from their national counterparts is a key concern with inter-regional linkages. At the same time, regional econometric models are typically dynamic frameworks in contrast to regional input–output models, which are detailed but static depictions of regional economies (Rey 2000). As such, regional econometric models lend themselves to the analysis of regional growth processes.

4.4 Integrated Models

The final genre in this classification is found in integrated human geographic models. A model can be classified as integrated in a number of ways. Substantively integrated models encompass a number of different activities within the same analytical framework, recognizing the interdependencies among, for example, regional labor markets, housing markets, demographic processes, transportation networks, and environmental systems (Wegener 1998).

Spatially integrated models provide a comprehensive treatment of one or more geographical dimensions of a process. For example, in multiregional economic models the interactions between different regional economies are treated as endogenous to the system. A second important class of spatially integrated models focuses on the relationships among different levels within a spatial hierarchy, such as global, national, regional and urban (Jin and Wilson 1993).

The final type of model in this genre is the methodologically integrated model. These models

combine different modules, each with a potentially different mathematical structure within the larger comprehensive framework, so as to achieve modularity and flexibility (Rey 2000). All types of integrated models are enjoying increasing popularity as social scientists increasingly recognize that many research problems are best approached from an inter-disciplinary perspective.

5. *Applications of Models in Human Geography*

Model applications can be placed into one of three classes: (a) statistical investigation; (b) optimization; and (c) simulation and forecasting. Statistical investigation is perhaps the most common use of models in human geography. Often this takes the form of using a statistical model to test a hypothesis. For example, human geographers have used specialized regression models to investigate such issues as urban decentralization, regional income inequality, historical voting patterns, and socioeconomic aspects of deforestation. The spatial dimensions are central in these investigations.

Statistical modeling in geography can be either exploratory, where the focus is on the identification of spatial pattern, testing for spatial clustering and departures from random sampling, or confirmatory, where the focus is on spatial autocorrelation and spatial structure in regression models (Anselin 2000). Spatial autocorrelation presents a fundamental problem to the application of traditional (i.e., a-spatial) regression methods, given that their application for inferential purposes is based on an assumption of random sampling, which is rendered invalid in the presence of spatial auto-correlation.

Optimization models identify the optimal solution for a given problem. All optimization models have a quantity that is to maximized (or minimized) as reflected in an objective function, for example in the case of the well-known transportation model:

$$\text{Minimize}\, L = \sum_{i=1}^{n} \sum_{j=1}^{n} T_{i,j} Q_{i,j} \tag{3}$$

subject to:

$$D_i \leqslant \sum_{j}^{n} Q_{i,j}\ \forall i \tag{4}$$

$$S_j \leqslant \sum_{j}^{n} Q_{i,j}\ \forall j \tag{5}$$

$$T_{i,j} Q_{i,j} \geqslant 0\ \forall i, j \tag{6}$$

where the objective is to minimize total transportation costs. The costs are a function of the unit transport cost $T_{i,j}$ and $Q_{i,j}$, the number of units shipped between location i and j. The solution has to respect a number of constraints such that the amount shipped to each region satisfies demand in that region D_j, the amount shipped from a region is no greater than supply in the region S_j and all costs are non-negative (Eqn. (4)).

Several different types of optimization models have been developed within human geography. Location-allocation models (Ghosh and Rushton 1987) typically involve a set of consumers distributed over a given area and a second set of facilities (stores, public services) to serve them. These models are used to allocate the facilities to best serve customers. Optimization methods developed by geographers have also been used to delineate regions for specific objectives, such as school and political zone redistricting.

Simulation and forecasting models are used to study and compare the responses of a system to policy or environmental changes. Examples can be found in the areas of integrated environmental-human analysis (Knight 2000) and the analysis of trade policy impacts on regional economies.

6. *Challenges and Prospects*

6.1 *Challenges*

While modeling in geography is in many ways similar to modeling in other disciplines, a number of important methodological challenges face modelers that deal with geographic dimensions of social processes. These include:

(a) Spatial scale
(b) Space-time modeling
(c) Computational complexity.

6.1.1 Spatial scale issues. A long standing issue in spatial modeling is the so-called modifiable areal unit problem which arises when inferences about the nature of a phenomenon change with the spatial scale used to study it. This has been illustrated in studies of the correlates of voting behavior in which the relationship between age and propensity to vote for a particular party has been demonstrated to vary both in strength and direction as the analysis moved from the individual level, to census tract, to county, to state.

A related challenge facing geographic modelers surrounds the relationships between processes that operate at different scales. For example, there is a growing need to integrate meso level models of urban land use with macro models of a regional economy, so that the interaction between dwindling land supply and aggregate regional economic growth can be

analyzed and better understood. On a broader level, there is also a pressing need to analyze the relationships between regionally focused human induced modification of the physical environment and global climate change.

A final issue related to spatial scale is the proper specification of contextual effects in developing models of socioeconomic systems. Geographers are faced with the twin challenges of building models that capture general systematic properties while capturing the richness of spatially explicit variation in many social phenomena. These two challenges might appear to be mutually exclusive as the search for empirical regularities would seem to rule out local variations and nuances. However, an active branch of geographical modeling has focused on so called 'local spatial analysis' in which methods are being designed to identify 'hot-spots' and deviations from overall general patterns. This opens the possibility of integrating both local and global types of modeling.

6.1.2 Space-time modeling. As geographic modeling becomes more widely utilized in other social sciences, the need to consider temporal as well as spatial dimensions of processes will grow. At present, there is a rich and mature body of methods for dealing with time series, and a modest and growing set of methods for spatial series. Methods for dealing with both dimensions simultaneously within the same model, however, are still in their infancy (Peuquet 1994). Advances in this arena are likely to come from cross-disciplinary work between spatial and time series modelers.

6.1.3 Computational complexity. Technological changes have affected some of the methodological issues facing modelers. In the past, discussions concerning data availability focused on the dearth of good datasets. Today, advances in data collection technologies (i.e., remote sensing, GIS, global positioning systems (GPS)) have shifted the discussion to problems of too much data. Traditional geographic modeling techniques are ill-suited for these enormous datasets and new approaches will be required. Work on spatial data mining and knowledge discovery (Ester et al. 2000) will play a crucial role in meeting this challenge.

The computational complexity associated with human geographic modeling also poses an institutional challenge for practitioners. As modeling becomes ever more sophisticated, the costs of entry into this field also increase. Part of the growth of other paradigms within human geography, such as social theory, has been attributed to their relatively lower start-up costs (Fotheringham 1993). As such, the attraction and training of future generations of geographic modelers requires serious attention.

Finally, while the new geo-spatial technologies have fueled a resurgence of interest in geographical modeling and analysis, these technologies have also redefined the types of questions researchers are addressing. A number of scholars have raised concerns about researchers becoming overly dependent on proprietary spatial modeling software, citing the lack of advanced capabilities and the slowness with which advances in geographical modeling are incorporated. A promising development in this regard is the rise of the open source movement (Raymond 1999) in which software development is shared by members of the research community working in a given area. This offers some exciting possibilities for advancing the state of the art of geographical modeling as reflected by several recent projects (Goodchild et al. 2000). In addition to fostering collaboration among scientists, this approach towards model development has the added advantage of lowering the monetary costs associated with the use of geographical models. This should stimulate their wider dissemination.

6.2 Future Prospects

Recent developments, both internal and external to the field of human geography, bode well for the future of modeling. Within the discipline, the stature of modeling as an important area of research is now well established and its role in legitimizing human geography as a discipline widely accepted. Within the broader social sciences, there is a growing recognition of the importance of space to many social problems. The rediscovery of the geographical dimensions of these problems together with the recognized need for interdisciplinary approaches has given rise to new research agendas within the social sciences. Human geographers active in the development and application of geographic modeling will have an important role to play in strengthening human geography's claim as an integrating discipline within the social sciences.

See also: Geodemographics; Regional Science; Spatial Association, Measures of; Spatial Autocorrelation; Spatial Choice Models; Spatial Data; Spatial Interaction Models; Spatial Optimization Models; Spatial Sampling; Spatial Search Models; Spatial-temporal Modeling; Spectral Analysis

Bibliography

Alonso W 1964 *Location and Land Use*. Harvard University Press, Cambridge, MA

Anselin L 2000 Spatial econometrics. In: Baltagi B (ed.) *Companion to Econometrics*. Blackwell, London

Batty M, Longley P 1994 *Fractal Cities: A Geometry of Form and Function*. Academic Press, New York
Cadwallader M 1985 *Analytical Urban Geography*. Prentice-Hall, Engelwood Cliffs, NJ
Carroll G 1982 National city size distributions: What do we know after 67 years of research? *Progress in Human Geography* **6**: 1–34
Dewhurst J, Hewings G, West R 1991 *Regional Input–Output Modeling: New Developments and Interpretations*. Avebury, Aldershot, UK
Ester M, Frommelt A, Kriegel H-P, Sander J 2000 Spatial data mining: Database primitives, algorithms and efficient dbms support. *Data Mining and Knowledge Discovery* **4**: 193–216
Fotheringham A S 1993 On the future of spatial analysis: The role of {GIS}. *Environment and Planning A* **25**: 30–4
Fujita M, Krugman P, Venables A J 1999 *The Spatial Economy: Cities, Regions, and International Trade*. MIT, Cambridge, MA
Goodchild M F, Anselin L, Applebaum R P, Harthorn B H 2000 Toward spatially integrated social science. *International Regional Science Review* **23**: 139–59
Ghosh A, Rushton G 1987 *Spatial Analysis and Location-allocation Models*. Van Nostrand Reinhold, New York
Gross P R, Levitt N 1994 *Higher Superstition: The Academic Left and Its Quarrels with Science*. Johns Hopkins University Press, Baltimore, MD
Haines-Young R 1989 Modelling geographical knowledge. In: Macmillan B (ed.) *Remodelling Geography*. Blackwell, London, pp. 22–39
Jin Y, Wilson A 1993 Generation of integrated multispatial input–output models of cities {GIMIMoc} 1: Initial step. *Papers in Regional Science* **72**: 351–68
Johnston R 1991 *Geography and Geographers: Anglo-American Human Geography Since 1945*. Edward Arnold, London
King L J 1993 Spatial analysis and the institutionalization of geography as a social science. *Urban Geography* **14**: 538–51
Knight C 2000 Regional assessment, *Encyclopedia of Global Change*. Oxford University Press, New York
Krugman P 1998 Space: The final frontier. *Journal of Economic Perspectives* **12**: 161–74
Lawson V 1995 The politics of difference: Examining the quantitative/qualitative dualism in postructuralist feminist research. *The Professional Geographer* **47**: 440–57
Longley P 1993 Quantitative analysis, technical change, and the dissemination of {GIS} research: Some reflections and prospects. *Environment and Planning A* **25**: 35–7
Odland J 1978 The conditions for multi-center cities. *Economic Geography* **54**: 234–44
Openshaw S 1989 Computer modelling in human geography. In: Macmillan B (ed.) *Remodelling Geography*. Blackwell, London, pp. 70–88
Peuquet D 1994 It's about time: A conceptual framework for the representation of spatiotemporal dynamics in geographic information systems. *Annals of the Association of American Geographers* **84**: 441–61
Raymond E C 1999 *The Cathedral & The Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly, Sebastopol, CA
Rey S J 2000 Integrated regional econometric + input–output modeling: Issues and opportunities. *Papers in Regional Science* **79**: 271–92
Sui D 2000 New directions in ecological inference: An introduction. *Annals of the Association of American Geographers* **90**: 579–82
Theil H 1971 *Principles of Econometrics*. Wiley, New York
Wegener M 1998 Applied models of urban land use, transport and environment: State of the art and future developments. In: Lundqvist L, Mattsson L-G, Kim T (eds.) *Network Infrastructure and the Urban Environment*. Springer Verlag, Berlin, pp. 245–67

S. J. Rey

Mathematical Models in Philosophy of Science

Philosophy of science analyses critically and reflects on the nature of science, its scope, methods, practices, principles, aims, and achievements. The central questions in the philosophy of science are *about* science, addressed in a manner which is usually 'external' to science. Prime examples of questions philosophers ask and try to provide precise answers to tend to be rather broad: 'What is a scientific theory?,' 'How do scientific models refer to reality?,' 'Can some theories be reduced to others?,' 'How does a theory change and evolve?,' etc. Because these sorts of questions are complex and hard, philosophers often disagree about the answers.

Starting in the 1920s, logical positivists (Carnap's Vienna Circle), logical empiricists (Reichenbach's Berlin School) and their numerous successors, impressed by the success and rigor of the sciences, approached the analysis of scientific concepts (including theory, law, evidence, explanation, prediction, confirmation, and reduction) in a distinctive way, by making extensive use of the techniques of formal language and logic. For example, their canonical formulation of a scientific theory is tantamount to a linguistic structure with two disjoint nonlogical vocabularies, consisting of *theoretical* and *nontheoretical* (observation) terms, respectively. The resulting *observation* sublanguage of the latter is presumed to be fully interpreted in a given domain of the phenomenal world of directly observable things, events, and their properties and relations. The accompanying *theoretical* sublanguage (erected only over theoretical terms) is meant to include all theoretical postulates and their deductive consequences. It is linked to the observation sublanguage via *correspondence rules*—a finite string of nontrivial 'mixed sentences' of the joint language structure, with at least one theoretical and one observation term. Because, in general, correspondence rules can provide only a partial interpretation of the theoretical sublanguage, the question of referents of theoretical terms must remain open. One of the most important duties of formalization is to make perfectly clear what is meant by a primitive (observational and theoretical) term, axiom, theorem,

and proof. Logical empiricists believed that casting empirical claims into formal frameworks would aid the clarification of the nature of scientific theories significantly, as well as their relationships to phenomena and data. Other scientific notions, such as prediction and explanation, dropped out as special cases of the concept of deduction. For many years, logical empiricism was the standard, received, syntactic view in philosophy of science, and it exerted considerable power and influence.

Since the 1960s, however, many aspects of the standard view have come under heavy attack. In addition to showing that observational–theoretical and analytic–synthetic distinctions are untenable, a major criticism of the standard view focused on the inadequacy of correspondence rules. Critics of the syntactic view argued that these rules oversimplify unreasonably the actual relationships between theories and phenomena. For example, in the received view, two or more theories cannot be applied conjointly, because correspondence rules work independently within component theories, without any structure for interaction. To remedy the situation, every time several theories are to be applied to the same phenomenon, one would need a simultaneous axiomatization of all pertinent theories in a single theory, together with a brand new set of correspondence rules. What results is an endless proliferation of axiomatic theories. The classic argument against the observational–theoretical dichotomy is that all observation terms are theory-laden and therefore there can be no theory-neutral observation language. By way of illustration, because the seemingly identical notion of velocity in Newtonian mechanics vs. in special relativity is actually calculated differently in these two theories, observational data pertaining to composite velocities will have different empirical meanings, depending on the underlying theory. The same goes for the analytic–synthetic dichotomy.

Another important factor leading to the demise of the standard view and the positivist program was the so-called *globalist* challenge (also known as the *Weltanschauungen* analyses) by Kuhn, Lakatos, and others. Globalist alternatives to the standard view have focused on the problems of scientific change, and the social and temporal dimensions of scientific development. From this perspective, theories are abstract, polysemous, evolving, cultural entities, possessing a deep structure which may be characterized in different linguistic ways.

Around the beginning of the 1970s, several new systematic metatheories of scientific theories emerged, often referred to collectively as the *model-theoretic* or *semantic* conception of scientific theories (as distinguished from the older *syntactic* conception, characterized above). Instead of viewing a theory as a deductively organized body of empirical (true or false) claims, the semantic conception views a theory as a way of specifying a class of models which represents various aspects of observed phenomena. This approach not only corrects the linguistic bias of radical empiricists and handles the globalists' formulation of theory change, it also brings mathematics and philosophy of science closer together.

Differing versions have been developed by a number of philosophers and logicians. In present-day philosophy of science there are three main directions. The first and perhaps most familiar is the so-called *set-theoretical predicate* approach. This view is very close to the practice of scientists, and it provides a natural and convenient way of formulating theories and reasoning about them. For instance, the monograph of Krantz et al. (1971) relies entirely on this mode of presentation. The second approach—popular mainly in philosophy of physics and biology—is the *topological state space* method. And the third, and by far the most formal conception, is implemented by the *structuralist* program. It attempts to address systematically a large number of important problems of philosophy of science, including the structure and dynamics of scientific theories, their equivalence, reduction, approximation, idealization and empirical content. Although recently the semantic conception has been consolidated and extended by a number of philosophers from a variety of perspectives, subtle differences still remain.

1. *The Set-theoretical Predicate Approach*

The set-theoretical predicate view was originally proposed by Suppes (1957, last Chap.). Unlike the logical empiricists, who defined theories to be axiomatic deductive systems (see *Axiomatic Theories*) formalized within a language of (say) first-order logic with equality, Suppes (1988, 1993) argues that a scientific theory should be viewed as being determined by the class of those set-theoretic models in which the empiricists' axioms are true. Accordingly, whenever the standard view specifies an empirical theory by a body of non-logical axioms (e.g., a customary list of ordinary or partial differential equations), Suppes introduces a corresponding set-theoretical predicate whose extension coincides with the collection of those models which satisfy the axioms. The choice of set theory as the best formal tool for representing theories is amply justified by its universal character.

The great classical example of a set-theoretical predicate is given by a simplified theory of classical particle mechanics. In its model-theoretic reconstruction, one begins with a finite *principal* domain (non-empty set) D of (idealized) dimensionless particles or mass points the theory is meant to be about and the *auxiliary* quantity domains of real $\mathbb{R}$ and natural $\mathbb{N}$ numbers. In general, auxiliary domains are viewed as previously defined and known structures, serving as parameters in the definition of a model. Because particles are situated in space–time and the interest is

in explaining their motion caused by various external forces, two additional *ambient* geometric domains are needed: S for physical space points, and T for time instants.

The next step in formulating the pertinent set-theoretical predicate is to enrich the inventory of domains by including a longer list of set-based functions. In intended applications of classical particle mechanics one needs, above all, a *mass* function $m: D \to \mathbb{R}^+$ that assigns to each particle a in D its unique mass $m(a)$, a dimensionless positive number. Axioms of mechanics also require a *position* function $s: D \times T \to S$ that assigns to each particle a at each time instant t its (dimensionless) geometric position $s(a, t)$. Because the relevant space and time structures are carried by the real line, one can use a metric isomorphism $\tau: T \to \mathbb{R}$ to numerically coordinatize the time domain and, likewise, a metric isomorphism $\sigma: S \to \mathbb{R}^3$ can be designated to coordinatize the physical space domain. The job of τ^{-1} is to transport metric and differentiable structure from $\mathbb{R}$ to T. The last item of the conceptual inventory is an *external force* function $f: D \times T \times \mathbb{N} \to \mathbb{R}^3$ such that to any particle a at any time t and to any itemizing natural number n it assigns the nth external force vector $f(a, t, n)$ exerted on a at time t.

The ordered sextuple

$$\langle D, S, T, s, m, f \rangle$$

which simply lists the foregoing set-based conceptual inventory, is the common mathematical shorthand for a possible classical particle model. (To reduce notational clutter, auxiliary domains and isomorphism maps are not included.)

In this simple modeling situation, the pertinent set-theoretical predicate '… is a CPM' (designating the property of being a *classical particle mechanics* model) is defined extensionally as follows:

Definition: $\langle D, S, T, s, m, f \rangle$ is a CPM if and only if the following axioms hold:

Frame Conditions

(a) D, S, and T are nonempty sets, and D is finite;

(b) $m: D \to \mathbb{R}^+$, $s: D \times T \to S$, and f: $D \times T \times \mathbb{N} \to \mathbb{R}^3$ are set-theoretic maps; and

(c) The implicit maps $\tau: T \to \mathbb{R}$ and $\sigma: S \to \mathbb{R}^3$ are (metric) isomorphisms such that the composite map $\sigma \circ s(a, \tau^{-1}(\cdot)): \mathbb{R} \to \mathbb{R}^3$ is twice differentiable for all $a \in D$.

Substantive Laws

The following equation (Newton's second law of motion)

$$m(a) \cdot \frac{d^2}{dx^2} \sigma \circ s(a, \tau^{-1}(x)) = \sum_{n \in \mathbb{N}} f(a, \tau^{-1}(x), n)$$

holds for all particles $a \in D$ and reals $x \in \mathbb{R}$ with a presumed absolutely convergent right-hand side.

A few words about these axioms are in order. According to the set-theoretical predicate view, theories are not treated as families of statements or axioms (i.e., linguistic entities). Rather, they are viewed as certain classes of *set-theoretical structures*, serving as extensions of 'interesting' set-theoretical predicates. It should be noted that this approach does not provide a sharp distinction between theories in mathematics vs. empirical science. Evidently, the set-theoretical predicate program is extremely general; it comprises many models nobody would ever seriously regard as 'useful'. For example, the CPM predicate above applies both to the traditional models of the planetary system and to infinitely many *unintended* numerical and geometric models (see *Mathematical Psychology*).

The foregoing definition and example may lead one to believe that in an attempt to overcome the weaknesses of the standard view with the help of set theory, actually nothing has been accomplished, because the familiar Galois connection between first-order theories and their models automatically justifies a back-and-forth move between syntactic and semantic formulations of theories. Indeed, a first-order formal theory T automatically defines its collection of alternative models $\mathrm{Mod}(T)$, namely, the body of set-theoretic structures in which the theory's axioms are true, and conversely, a given collection of models $\mathscr{M}$ specifies the theory $\mathrm{Th}(\mathscr{M})$, comprised of all sentences satisfied by the models in $\mathscr{M}$. But this is not the case in general. For example, set-theoretic structures of the form $\langle D, S, T, s, m, f \rangle$ can be described in many non-equivalent (usually higher-order) ways, without altering the extension of the predicate '… is a CPM.' The above discussion shows that the semantic view does not require any specific 'laws of nature' and is not committed to any particular linguistic formulation.

In order to be able to draw a demarcation line between classes of models used in pure mathematics and empirical science, various alternatives to the set-theoretical predicate approach have been developed. One idea was to define an empirical theory in terms of a collection of set-theoretic structures of some arbitrary but fixed similarity type (i.e., models in the sense of Tarskian logical model theory), together with a designated subset of models, serving as *intended applications* of which the theory is expected to be definitely true.

Suppes (1962) argued that the deficiency of the syntactic conception in explaining the role of experimental procedures in relating theories to phenomena can be corrected by postulating a *hierarchy of models* which mediate between a designated higher-level theory and an experimental situation. To tie the set-theoretical predicate treatment of theories to more concrete empirical situations, Suppes identified three typical levels of theories, which are sandwiched between a top-level theory and phenomena: theory of experimental design (see *Experimental Design: Over-*

view). Each level in the hierarchy is treated model-theoretically, via a specification of a set-theoretical predicate. This may also resolve the potential drawback of not being able to put a demarcation line between mathematics and empirical science. Also, one can obtain new models out of old ones by means of standard product, quotient and other model-theoretic constructions. The set-theoretical predicate approach has numerous applications in the study of axiomatizability, undecidability, incompleteness, and theory reduction.

In sum, this approach can be thought of as an extension of Bourbaki's '*species of structures*' program in mathematics to empirical science.

2. *The Topological State Space View*

The state space formulation of a theory was initiated by Beth (1961), and further developed by Van Fraassen (1970) and Suppe (1988). On this view, theories of empirical systems are represented by topological state spaces with a dynamic structure. Specifically, the set-based triple $\langle X, G, \Delta \rangle$ is called a *state-space model* of a dynamical system iff X is a topological space, called the *state* (phase) space of the system, G is a topological group, referred to as the *dynamic* group of the system, and $\Delta: G \times X \to X$ is a continuous group action, satisfying $\Delta(1, x) = x$ and $\Delta(g, \Delta(g', x) = \Delta(gg', x)$ for all $g, g' \in G$ and $x \in X$.

The canonical example of a state-space model is the classical phase space of a single particle, moving in a potential field. In this case, the representing state space X of the particle is given by the sixth Cartesian power $\mathbb{R}^6 \cong \mathbb{R}^3 \times \mathbb{R}^3$ of the real line. Its elements are ordered sextuples of the form $\langle x, y, z, p, q, r \rangle$, specified by conjugate pairs of position-momentum vectors, commonly used to provide a complete description of all possible instantaneous states of a dynamical system. Specifically, $\langle x, y, z \rangle$ is the position vector of the unconstrained particle, measured in three independent spatial directions, and likewise $\langle p, q, r \rangle$ codes the particle's momentum vector, viewed in three independent directions. Of course, X is understood to be a topological space, induced by the natural topology of the real line. (For dynamical systems with n unconstrained particles, the representing topological state spaces have $3n + 3n = 6n$ dimensions.)

Generally, physical quantities (observables) are represented by continuous real-valued state functions $f: X \to \mathbb{R}$. In particular, the Hamiltonian (total energy) function h is given by the sum of the particle's potential and kinetic energy

$$h(x, y, z, p, q, r) = V(x, y, z) + \frac{1}{2m}(p^2 + q^2 + r^2)$$

where m stands for the particle's mass. Upon denoting the state of the given particle at time t by $\langle x(t), \ldots, r(t) \rangle$, the coordinate functions $x(t), \ldots, r(t)$ will satisfy the well-known sextuple of first-order, deterministic Hamiltonian differential equations (three for describing how the position coordinates are changing with time, and another three for characterizing how momentum coordinates are changing with time)

$$\frac{dx}{dt} = \frac{\partial h}{\partial p}, \ldots, \frac{dp}{dt} = -\frac{\partial h}{\partial x}, \ldots$$

which in 'practice' tend to have unique solutions for all t. That is, given any time instant t_0 and a state $\langle x_0, \ldots, r_0 \rangle$ in X, there exists a unique differentiable map $\Delta': \mathbb{R} \to \mathbb{X}$ such that $\Delta'(t) = \langle x(t), y(t), z(t), p(t), q(t), r(t) \rangle$ and the coordinate function $x(t), \ldots, r(t)$ satisfy the differential equations with the initial condition $\Delta'(t_0) = \langle x_0, \ldots, r_0 \rangle$. Consequently, somewhat more formally, for any t there is a dynamic map $\Delta'(t): X \to X$ with the interpretation that if $\mathbf{x}$ is the state of the particle at time t', then $\Delta'(t)(\mathbf{x})$ is the particle's state at time $t + t'$. Simply, there is a one-parameter time group $G = \mathbb{R}$, acting on the state space X via $\Delta: \mathbb{R} \times \mathbb{X} \to \mathbb{X}$ where $\Delta(t, x) = \Delta'(t)(\mathbf{x})$.

The Hamiltonian equations determine the phase trajectory of a given state $\mathbf{x}$ in the state space, i.e., a curve (geometric orbit of states) $\{\Delta(t, \mathbf{x}) | \mathbf{t} \in \mathbb{R}\}$ in X, representing the evolution of the dynamical system with initial state $\mathbf{x}$. The set of phase trajectories constitutes a complete phase portrait of the dynamical system under consideration. The phase portrait can be thought of as the 'law' governing the particle's motion. An essential point here is that within the state space framework one can work with the geometric structure of the state space model, even when analytic integration of the Hamiltonian differential equations proves impossible. In sum, the state space view does not require any specific equational laws of nature; qualitative state space geometry can be studied and applied in its own right.

Van Fraassen (1980) interprets the state space model empiricistically. Although in general, state spaces tend to include many states which do not correspond to anything observable, there are special subsets of states that can be specified observationally, by means of designated state functions, representing measurable quantities, such as position and velocity. Understandably, their values can only be measured with finite accuracy and during a limited period of time. Van Fraassen refers to these subsets of 'observable' states together with a string of measurable state functions as *empirical substructures* of a state space model. On the state space view, a state space model is empirically adequate for a dynamical system just in case it possesses an empirical substructure isomorphic to the observable phenomena associated with the dynamical system. Since there may be no objective way to single out the totality of observable phenomena associated with a dynamical system, there is considerable debate

over the seemingly inocuous notions of 'empirical substructure' and 'isomorphism.'

In sum, the state space conception is particularly useful in representing the dynamic aspects of time-dependent phenomena. Specifically, chaotic behavior and stability conditions of nonlinear dynamical systems may be more conveniently handled within the geometric universe of state spaces than by the traditional analytic method of equations. Traditionally, however, state space reconstructions are intended for modeling the behavior of so-called dissipative and conservative systems, and in view of their highly specialized geometric structure, they do not always fit the prevailing views about models and modeling in philosophy of science.

3. *The Structuralist Program*

The structuralist view, advocated by Stegmüller (1976) and his followers (see Balzer et al. 1987, 1996), holds that the most adequate way to understand the essence of scientific theories is by treating them as complexes of set-theoretic structures, made up of simpler structures. On this bottom-up conception, in analogy with the Bourbakian notion of species of structures, a larger variety of classes of models is considered. It is a formal theory about scientific theories, their structure, interaction and dynamics.

The structuralist school provides a systematic formal methodology to reconstruct particular scientific theories and their development. It has great philosophical appeal because it also handles many global issues, including reduction, equivalence, approximation, and specialization of scientific theories.

Very much as in the set-theoretical predicate framework, the simplest structural units of a theory are its models (in the sense of formal semantics), that is, n-tuples of the form

$$\langle D, D', \ldots, R, R', \ldots, f, f', \ldots \rangle$$

understood to be comprised of three types of mathematical objects:

Ontology: D, D',... are the *principal* domains of the model. They contain objects (or rather their idealizations), such as particles, fields, genes, commodities, persons, etc., assumed by the theory. Typical standard *auxiliary* mathematical domains of reals, integers, etc., are also included, but for the sake of brevity, they are not made visible in the model's notation.

Relational Structure: R, R', ... denote the basic comparative relations, defined on (some) of the domains listed above.

Quantitative Structure: f, f', ... designate functions from products of principal and auxiliary domains to real numbers or their vectors.

A scientific theory is defined, in essence, by a collection of models of an arbitrary but fixed similarity type of the foregoing kind. The choice of axioms to be satisfied by the models of such a class is considered to be of secondary importance, so long as alternate collections of axioms continue to individuate the same class of models. Structuralists tend to classify axioms into *frame conditions* and *substantive laws*. Frame conditions specify the domains and codomains of all functions and relations given in the model, and the laws state (in the language of available concepts) suitable structural or technical conditions pertaining to those aspects of the world that are of modeling interest. There are many conditions placed on classes of models in various contexts, and a number of them are summarized below.

In structuralism, the totality of models $\mathscr{M}_p$ satisfying only the frame conditions of a theory is called its collection of *potential models*. The body of models $\mathscr{M}$ satisfying all the frame conditions and the substantive laws of a theory is called its class of *actual* models. Because structuralists lean towards an empiricist outlook on science, they usually assume a dichotomy of theoretical and non-theoretical (observation) relations and functions in their models. The theoretical component of a model (given by its domains, and designated theoretical relations and functions) is regarded as something intrinsic to the ambient theory. That is, the truth values of relations and numerical values of functions can only be determined by assuming the theory. On the other hand, the non-theoretical part of the model can be determined from sources external to the theory. This also means, in particular, that the values of non-theoretical functions and relations in the model can be obtained from measurement (see *Measurement Theory: History and Philosophy*), governed by other (auxiliary) theories.

This distinction is quite similar to the standard dichotomization of statistical concepts into parameters and statistics (see *Statistics: The Field*), and the system engineers' partition of concepts into a pair of disjoint sets of system parameters and input/output variables, respectively. The collection of models $\mathscr{M}_{pp}$ (reducts) obtained from $\mathscr{M}_p$ by simply lopping off all theoretical relations and functions is called the class of *partial potential* models of the theory. (There is a serious debate over the desirability and tenability of assuming the foregoing dichotomy.)

Models of a theory are usually related by various *constraints*. For example, in the theory of classical particle mechanics, it is usually assumed that the mass of a given particle is the same in all models, whose principal domains contain the particle. Formally, constraints $\mathscr{C}$ are families of sets of models. In this example, the class $\mathscr{M}_p$ of possible models consists of sextuples $\langle D, S, T, s, m, f \rangle$, satisfying only the frame conditions of classical particle mechanics, and the class $\mathscr{M}$ of actual models is defined by sextuples $\langle D, S, T, s, m, f \rangle$, satisfying both the frame con-

ditions and the substantive laws. Since external force and mass function are regarded as theoretical, partial potential models have the simple form $\langle D, S, T, s \rangle$.

In a somewhat arcane way, with the notation above, the structuralist school represents a scientific theory in terms of quadruples of classes of models of the following form:

$$\langle \mathscr{M}, \mathscr{M}_p, \mathscr{M}_{pp}, \mathscr{C} \rangle$$

where $\mathscr{M} \subset \mathscr{M}_p$, $\mathscr{C} \subset \mathscr{P}\mathscr{M}_p$, and there is an implicit projection map $\pi : \mathscr{M}_p \to \mathscr{M}_{pp}$.

The realm of observable phenomena to which theory users may want to apply their theory's concepts and laws can also by represented by suitable collections, usually denoted by $\mathscr{I}$. Concretely, structuralists propose to capture bodies of intended applications empiricistically by $\mathscr{I} \subset \mathscr{M}_{pp}$ or realistically by $\mathscr{I} \subset \mathscr{M}_p$. The second alternative is more appropriate when some of the values of theoretical functions and relations are already known. For empiricists, at any given stage of knowledge, the set $\mathscr{I}$ refers to the totality of potential data the theory is supposed to account for. Finally, the structuralists are also able to characterize the *content* of their theories. It is a range of ways the phenomena could be (i.e., possible configurations of data) according to the given theory. Formally, the content $\mathrm{Cont}(\mathscr{M}, \mathscr{M}_p, \mathscr{M}_{pp}, \mathscr{C})$ of a theory consists of all subsets of partial potential models in $\mathscr{M}_{pp}$ whose members are extendable (by adding suitable theoretical components) into a full model that is a member of both $\mathscr{M}$ and $\mathscr{C}$. Now an *empirical claim* of a theory is just a simple condition $\mathscr{I} \in \mathrm{Cont}(\mathscr{M}, \mathscr{M}_p, \mathscr{M}_{pp}, \mathscr{C})$, stating, in an extremely high-flown way, that the entities to which the theory is to be applied actually bear the structure proposed by the theory.

Often scientific theories form a partially ordered network of directly or indirectly related theories (theory nets) which can be aggregated into still larger units, called *theory holons*. Models in theory holons are not necessarily of the same similarity type. These models are usually interrelated by external *links*. For example, links between Newtonian, Hamiltonian, and Lagrangue's mechanics can be understood as part of a physical theory holon.

With this apparatus, Stegmüller and others have (more or less) successfully reconstructed Kuhn's theory of paradigm shifts and Lakatos' research programs. Of course, model-theoretic approaches cannot resolve all problems of philosophy of science, but they can make them clearer and hopefully more tractable.

More recently, some philosophers of science (e.g., Mormann 1996) have considered casting the structural approach to scientific theories in the framework of category theory. In this approach, objects of categories are given by models or model classes, or pairs of such classes (e.g., $\langle \mathscr{M}_p, \mathscr{M} \rangle$), and the arrows are appropriate structure-preserving maps (e.g., sending pairs of model classes to their direct images). It is plausible enough to treat theories category-theoretically, but it seems far less clear how to isolate interesting and powerful categories that include a number of decisive constructions, essential for characterizing the structure and dynamics of actual scientific theories.

Another mathematical direction in twentieth–twenty-first century philosophy of science is the *algorithmic* (computational) approach (see *Algorithms and Algorithmic Complexity*). Along the lines of Chaitin (1992), one can consider syntactically formulated scientific theories of empirical phenomena as analogs of algorithms or computer programs, capable of encapsulating all data available about the phenom ena in question. The reason for doing this is that the interpretation of theories as programs and data as program outputs leads to an important measure of *algorithmic complexity* of scientific theories. The algorithmic complexity of a theory of a particular phenomenon is taken to be the *length* of the shortest program (understood as the counterpart of the theory) that reproduces all pertinent data about the phenomenon. Here, the length can be measured by the number of symbols needed to present the program in a universal programming language. According to the algorithmic approach, the point of a scientific theory is to reduce the arbitrariness in its data. If the program that reproduces the data (and serving as the theory's counterpart) has the same length as the body of data itself, then the theory is obviously redundant. Algorithmically incompressible data possess no regularities whatsoever. They are random in the sense that they can be neither predicted not characterized by laws descriptively shorter than the body of data itself. Chaitin points out that these ideas of complexity can be extended readily to empirical systems. Specifically, algorithmic complexity of an empirical system is the length of the shortest program that simulates (or describes) it.

The dual idea of *organizational complexity* (computational cost) derives from the *running time* for the shortest program that generates the data. To appreciate the nicety of this concept, observe that while the time required for solving finite systems of algebraic equations grows only as a polynomial function of the number of equations, the solution time for the famous Tower of Hanoi toy problem increases exponentially with the number of rings. For example, with only 64 rings the problem requires $2^{64} - 1$ steps, taking over five trillion years with one ring transfer per 10 seconds. In brief, organizational complexity subdivides the class of computable entities into those that are 'practically' computable and those that are not. Algorithmic and computational approaches are at the heart of many modern logical techniques, applied in philosophy of science. These approaches suggest that scientific theories and systems can be compared with respect to both their degrees of algorithmic compressibility and amounts of organizational complexity (intrinsic depth).

4. Conclusion

The most immediate connection between philosophy of science and mathematics is via the structure of scientific theories. Although all three approaches examined here are employed frequently in actual scientific work, they harbor serious difficulties. The popular *set-theoretical predicate* approach does not draw a line between empirical and mathematical theories. A far more specialized *topological state space* method fails to provide sufficient support for the empirical meaning of its topological structure. Finally, the extensive complexity found in the *structuralist* methodology generates pessimism regarding the reconstruction of actual scientific theories. Be that as it may, close analysis of these different approaches will affirm that the semantic method has cast a good deal of light on the nature of scientific theories and therefore will stay for the present. There is, however, a strong possibility that in the foreseeable future the set-theoretic approach will be replaced gradually by the methods of category theory.

See also: Axiomatic Theories; Explanation: Conceptions in the Social Sciences; Functional Explanation: Philosophical Aspects; Kuhn, Thomas S (1922–96); Logical Positivism and Logical Empiricism; Reduction, Varieties of; Structuralism

Bibliography

Balzer W, Moulines C U (eds.) 1996 *Structuralist Theory of Science. Focal Issues, New Results*. Walter de Gruyter, New York

Balzer W, Moulines C U, Sneed J 1987 *An Architectonic for Science: The Structuralist Program*. D Reidel, Dordrecht, The Netherlands

Beth E 1961 Semantics of physical theories. In: Freudenthal H (ed.) *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences*. D Reidel, Dordrecht, The Netherlands

Chaitin G J 1992 *Algorithmic Information Theory*. Cambridge University Press, Cambridge, UK

Krantz D H, Luce R D, Suppes P, Tversky A (eds.) 1971 *Foundations of Measurement, Vol. I*. Academic Press, New York

Mormann T 1996 Categorical structuralism. In: Balzer W, Moulines C U (eds.) *Structuralist Theory of Science. Focal Issues, New Results*. Walter de Gruyter, New York

Stegmüller W 1976 *The Structure and Dynamics of Theories*. Springer, New York

Suppe F 1988 *The Semantic Conception of Theories and Scientific Realism*. University of Illinois Press, Chicago, IL

Suppes P 1957 *Introduction to Logic*. Van Nostrand, Princeton, NJ

Suppes P 1962 Models of data. In: Nagel E, Suppes P, Tarski A (eds.) *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*. Stanford University Press, Stanford, CA

Suppes P 1988 *Scientific Structures and their Representation*, preliminary version. Stanford University, Stanford, CA

Suppes P 1993 *Models and Methods in the Philosophy of Science: Selected Essays*. Kluwer Academic, Dordrecht, The Netherlands

Van Fraassen B C 1970 On the extension of Beth's semantics of physical theories. *Philosophy of Science* **37**: 325–39

Van Fraassen B C 1980 *The Scientific Image*. Oxford University Press, New York

Z. Domotor

Mathematical Psychology

Mathematics has been used in psychology for a long time, and for different purposes. When William James (see Miller 1964) writes

$$\textit{Self-esteem} = \frac{\textit{Success}}{\textit{Pretensions}} \tag{1}$$

he is using mathematical notation metaphorically. Because no method is provided to measure the three variables, the equation cannot be taken literally. James means to convey, by a dramatic formula, the idea that if your pretensions increase without a corresponding increase of your success, your self-esteem will suffer. Such usages of mathematics, which (Miller 1964) calls 'discursive', can be found in psychological discourse at least since Aristotle. They were especially in favor in the last three centuries, often in the guise of some monomial equation resembling those of classical physics, whose success invited emulation. Examples can be found in the works of Francis Hutcheson, Moritz W. Drobisch, Johann F. Herbart, or more recently, Kurt Lewin, Edward C. Tolman, and Clark L. Hull. Despite their historical interest, we shall not review such metaphorical uses of mathematics here (see rather Boring 1950, or Miller 1964).

In this article, we reserve the term 'mathematical psychology' to the elaboration and the *testing* of *mathematical theories* (or models) for behavioral data. In principle, such a theory entails an economical representation of a particular set of data in mathematical terms, where 'economical' means that the number of free parameters of the theory is substantially smaller that the number of degrees of freedom (e.g., independent variables) in the data. In that sense, mathematical psychology plays for behavioral data the role that mathematical physics or mathematical biology play for physics or biology, respectively. In the best cases, the theory is cast in probabilistic terms and is testable by standard statistical methods. The large majority of such mathematical theories for behavioral data have emerged from four partially overlapping traditional fields: psychophysics, learning, choice, and response latencies. Each of these fields is outlined

below from the standpoint of the prominent mathematical models that have been proposed. Other topics to which mathematical psychologists have devoted much work are also mentioned.

1. The Precursors: Fechner and Thurstone

Gustav Theodor Fechner (1801–87; see Boring 1950) was by training an experimental physicist with a strong mathematical background. While his interests were diverse and his contributions many—ranging from experimental physics to philosophy—we only consider him here as the founder of psychophysics. His main purpose was the measurement of 'sensation' in a manner imitating the methods used for the fundamental scales of physics, such as length or mass. Because 'sensation' could not be measured directly, Fechner proposed to evaluate the difference between the 'sensations' evoked by two stimuli by the difficulty of discriminating between them.

To be more specific, we introduce some notation. We write x, y, etc. for some positive real numbers representing physical intensities measured on some ratio scale, e.g., sound pressure level. Let $P(x, y)$ be the probability that stimulus x is judged to be louder than stimulus y. Cast in modern terms (cf. Falmagne 1985), Fechner's idea amounts to finding a real valued function u defined on the set of physical intensities, and a function F (both strictly increasing and continuous) such that

$$P(x, y) = F[u(x) - u(y)] \tag{2}$$

This equation is supposed to hold for all pairs of stimuli (x, y) and (x', y') such that $0 < P(x, y) < 1$ (i.e., subjectively, x is close to y and x' is close to y'). If Eqn. (2) is satisfied, then the function u can be regarded as a candidate scale for the measurement of 'sensation' in the sense of Fechner. *A priori*, it is by no means clear that a scale u satisfying Eqn. (2) necessarily exists. A necessary condition is the so-called *Quadruple Condition*: $P(x, y) \geqslant P(x', y')$ *if and only if* $P(y, y') \leqslant P(y, y')$, where the equivalence is assumed to hold whenever the four probabilities are defined for $0 < P < 1$. In the literature, the problem of constructing such a scale, or of finding sufficient conditions for its existence, has come to be labeled *Fechner's Problem*. An axiomatic discussion of Fechner's Problem can be found in Falmagne (1985).

For various reasons, partly traditional, psychophysicists often prefer to collect their data in terms of discrimination thresholds, which can be obtained as follows from the discrimination probabilities $P(x, y)$. We define a *sensitivity function* $\xi: (y, v) \mapsto \xi_v(y)$ by the equivalence: $\xi_v(y) = x$ *if and only if* $P(x, y) = v$. The Weber function Δ: $(y, v) \mapsto \Delta_v(y)$ (from E. E. Weber 1795–1878, a colleague of Fechner, professor of anatomy and physiology at Leipzig) is then defined by the equation $\Delta_v(y) = \xi_v(y) - \xi_{0.5}(y)$.

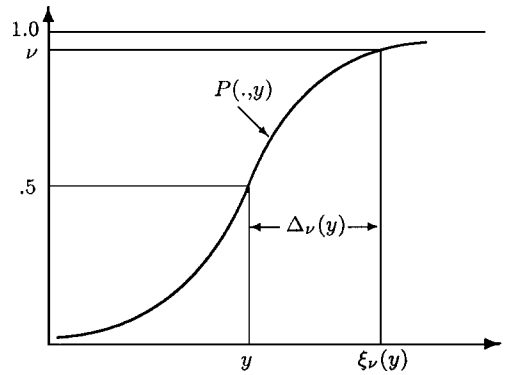

Figure 1
The Weber function Δ in a case where $P(y, y) = .5$; thus, $\xi_{0.5}(y) = y$. The S-shape function $P(., y)$: $x \mapsto P(x, y)$ is called a *psychometric function*

A graphic representation of these fundamental concepts of classic psychophysics is given in Fig. 1, in a special case where $P(y, y) = 0.5$ for a particular stimulus y. For each value of y, the S-shaped function $P(., y)$: $x \mapsto P(x, y)$ is called a *psychometric function*.

Starting with Weber himself, the Weber function Δ has been investigated experimentally for many sensory continua. In practice, $\Delta_v(y)$ is estimated by *stochastic approximation* (e.g. the methods of Robbins and Monro, and Levitt; see Wasan 1969) for one or a few *criterion* values of the discrimination probability v and for many values of the stimulus y. A typical finding is that, at least for stimulus values in the midrange of the sensory continuum and for some criterion values v, $\Delta_v(y)$ grows approximately linearly with y:

$$\Delta_v(y) = yC(v) \tag{3}$$

for some function C depending on the criterion. The label *Weber's Law* is attached to this equation. It is easily shown that Eqn. (3) is equivalent to the homogeneity equation

$$P(\lambda x, \lambda y) = P(x, y) \quad (\lambda > 0)$$

which in turn leads to $P(x, y) = H(\log x - \log y)$, with $H(s) = P(e^s, 1)$. This means that the function u in Eqn. (2) has the form

$$u(y) = A \log y + B \tag{4}$$

where the constants $A > 0$ and B arise from uniqueness considerations for both u and F. Equation (4) has been dubbed *Fechner's Law*. Our discussion indicates that, in the framework of Eqn. (2), Weber's Law and Fechner's Law are equivalent. Much of contemporary psychophysics evolved from Fechner's ideas (see below).

The most durable contribution of Leon Louis Thurstone (1887–1955) to mathematical psychology is

his *Law of Comparative Judgements* (see Bock and Jones 1968), a cornerstone of binary choice theory closely related to Eqn. (2) of Fechner's Problem. Thurstone supposed that a subject confronted with a choice between two alternatives x and y (where x, y are arbitrary labels and do not necessarily represent numerical values) makes a decision by comparing the sampled values of two random variables $\mathbf{U}_x$ and $\mathbf{U}_y$ associated with the alternatives. Suppose that these random variables are independent and normally distributed, with means $\mu(x)$, $\mu(y)$, and variances $\sigma(x)^2$, $\sigma(y)^2$, respectively. Denoting by Φ, as is customary, the distribution function of a standard normal random variable, this leads to

$$P(x, y) = P(\mathbf{U}_x > \mathbf{U}_y) = \Phi\left(\frac{\mu(y)-\mu(y)}{(\sigma(x)^2+\sigma(y)^2)^{\frac{1}{2}}}\right) \tag{5}$$

Assuming that all the random variables have the same variance $\sigma^2 = \alpha^2/2$, we obtain

$$P(x, y) = \Phi[u(x)-u(y)] \tag{6}$$

with $u(x) = \mu(x)/\alpha$ and $F = \Phi$, a special case of Eqn. (2). Equations (6) and (5) are called *Case III* and *Case V* of the Law of Comparative Judgements, respectively (cf. Bock and Jones 1968). Thurstone's model has been widely applied in psychophysics and choice. Other models in the same vein have been proposed by various researchers (e.g. Luce 1959, Luce and Suppes 1965). Thurstone's other important contributions are in learning theory, and especially *Multiple-Factor Analysis*—the title of one of his books (see Miller 1964).

2. *The Beginning: Mathematical Learning Theory*

Two papers mark the beginning of mathematical psychology as a distinguished research field: one by W. K. Estes entitled *Toward a Statistical Theory of Learning* and published in 1950 and the other by R. R. Bush and F. Mosteller which appeared the following year (cf. Atkinson et al. 1965; see also *Mathematical Learning Theory, History of*). These works were the highlights of a movement spear-headed by R. R. Bush and R. D. Luce at the University of Pennsylvania, and R. C. Atkinson, W. K. Estes, and P. Suppes at Stanford, which set out to formalize mathematical learning theories in terms of stochastic processes, and especially, Markov processes (cf. Bharucha-Reid 1960, 1988). There were good reasons for such a development at that time. The previous decade had been plagued by fruitless controversies concerning the basic mechanisms of learning. While a considerable experimental literature on learning was available (cf. Hilgard 1956), the statistical tools in use for the analysis of the data were poor, and the prominent theories ambiguous. When mathematics was used, it was metaphorically (in the sense of the first paragraph of this article). Moreover, the scope of the theories were ambitious, covering a vast class of experimental situations loosely connected to each other conceptually. A typical example of such an endeavor is C. L. Hull's theory (cf. Atkinson et al. 1965, Boring 1950). By contrast, the Markov models developed by Bush, Estes, and their followers were designed for specific experimental situations. Under the influence of Suppes, a philosopher of science from Stanford who played a major role in the development of mathematical psychology, these models were often stated axiomatically. As a consequence, the predictions of the models could, in many cases, be derived by straightforward mathematical arguments. Two classes of models were investigated.

2.1 *Finite State Markov Chains*

The basic idea here is that the subject's responses in a learning experiment are the manisfestations of some internal states, which coincide with the states of a finite Markov chain. The transitions of the Markov chain are governed by the triple of events (stimulus, response, reinforcement) occurring on each trial. In many situations, the number of states of the Markov chain is small. For the sake of illustration, we sketch a simple case of the so-called *one-element model*, in which the chain has only two states which we denote by N and K. Intuitively, N stands for the 'naive' state, and K for the 'cognizant' state. On each trial of the experiment, the subject is presented with a stimulus and has to provide a response, which is either labeled as C (correct) or as F (false). For concreteness, suppose that the subject has to identify a rule used to classify some cards into two piles. Say, a drawing on each card has either one or two lines, which can be either (both) straight or (both) curvy, and either (both) vertical or (both) horizontal. The subject must discover that all the cards with curvy lines and only those, go on the left pile. The subject is told on each trial whether the chosen pile is the correct one.

In words, the axioms of the model are as follows. [M1] The subject always begins the experiment in the naive state N; [M2] the probability of a transition from the naive state N to the cognizant state K is equal to a parameter $0 < \theta \leqslant 1$, constant over trials, regardless of the subject's response; [M3] in state N, the probability of a correct placement is equal to a parameter $0 < \alpha \leqslant 1$, constant over trials; [M4] in state K, the response is always correct. The derivation of the model can either be based on a two-state Markov chain with state space $\{N, K\}$, or on a three-state Markov chain with state space $\{NC, NF, KC\}$ (where N, K denote the subject's cognitive states and C, F the responses). In either case, the derivations are straightforward. Notice that, from axioms [M1] and [M2], the trial number of

the occurrence of the first K state has a geometric distribution with parameter θ. Writing $\mathbf{S}_n$ and $\mathbf{R}_n$ for the cognitive state and the response provided on trial $n = 0, 1, \ldots$ (thus, $\mathbf{S}_n = N, K$ and $\mathbf{R}_n = C, F$), we get easily for $n = 0, 1\ldots$,

$$P(\mathbf{S}_n = N) = (1-\theta)^n$$
$$P(\mathbf{R}_n = C|\mathbf{S}_n = N) = \alpha, \quad P(\mathbf{R}_n = C|\mathbf{S}_n = K) = 1$$
$$P(\mathbf{R}_n = C) = 1-(1-\theta)^n(1-\alpha)$$

which implies, with $p_n = P(\mathbf{R}_n = C)$

$$p_{n+1} = (1-\theta)p_n + \theta \tag{7}$$

A strong prediction of this model is worth pointing out: the number of correct responses recorded before an error occurring on trial number $n+1$ should be binomially distributed with parameters α and n. (Indeed, all the subjects' responses have been generated by a naive state.) As demonstrated in the work of P. Suppes and R. Ginsberg (see Atkinson et al. 1965), this prediction is surprisingly difficult to reject experimentally, at least in some situations.

This model is only one of a large class of finite state Markov models that were developed for diverse experimental situations ranging from concept learning to psychophysics and reaction time, mostly between the early 1950s and the late 1960s. Another closely related but essentially different class of models, intended for the same kind of learning experiments, was also investigated.

2.2 Linear Operator Models

To facilitate the comparison, we take the same concept identification experiment as above, involving the classification of cards into two piles, and we consider a simple representative of this class of models. As before, we denote by $\mathbf{R}_n$ the response on trial n. We take as the sample space the set of all sequences ($\mathbf{R}_0$, $\mathbf{R}_1,\ldots, \mathbf{R}_n,\ldots$), with $\mathbf{R}_n = C, F$ for $n = 0, 1,\ldots$. We also define $p_{\omega,0} = P(\mathbf{R}_0 = C)$, $p_{\omega,n} = P(\mathbf{R}_n = C|\mathbf{R}_{n-1},\ldots, \mathbf{R}_0)$, $n = 1, 2,\ldots$. Let $0 < \theta \leqslant 1$ be a parameter. The two axioms of the model are as follows: for $n = 0, 1,\ldots$

$$\textbf{[L1]}\ p_{\omega,n+1} = (1-\theta)p_{\omega,n} + \theta, \quad \text{if } \mathbf{R}_{n-1} = F$$
$$\textbf{[L2]}\ p_{\omega,n+1} = p_{\omega,n}, \quad \text{if } \mathbf{R}_{n-1} = C$$

Thus, the model has two parameters $p_{\omega,0}$ and θ, and learning occurs only when false responses are provided. As in the case of the finite Markov chain models, many predictions could be computed from such models, which could then be tested on the type of learning data traditionally collected by the experimenters. For both classes of models, the results of such enterprises were often quite successful. Nevertheless, the interest for such models waned during the 1960s, at least for learning situations, because the researchers gradually realized that their simple mechanisms were far too primitive to capture all the intricacies revealed by more sophisticated analyses of learning data (see especially Yellott 1969). General presentations of this topic can be found in Atkinson et al. (1965) for the finite state Markov learning models, and in Estes and Suppes (1959) for the linear operator models. A mathematical discussion of Markov processes for learning models is contained in Norman (1972).

Despite the partial failure of these models to provide a satisfactory detailed explanation of traditional learning data, their role was nevertheless essential in the introduction of modern probability theory (in particular stochastic processes) and axiomatic methods in theoretical psychology, and in promoting the emergence of mathematical psychology as a field of research. In recent years, a renewed interest in learning theory has appeared from the part of some economists (see *Mathematical Learning Theory*).

3. Psychophysics

The main research topics in psychophysics can be traced back to the ideas of Fechner and Thurstone outlined in Sect. 1. Fechner's method of measuring sensation is indirect, and based on the difficulty of discriminating between two stimuli. Under the impetus of S. S. Stevens a psychologist from Harvard, different experimental methods for 'scaling' sensation became popular.

3.1 Direct Scaling Methods

In the case of the magnitude estimation method, Stevens (1957) asked his subjects to make direct numerical judgements of the intensities of stimuli. For example, a subject may be presented with a pure tone of some intensity x presented binaurally, and would be required to estimate the magnitude of the tone on a scale from 1 to 100. Typically, each subject would only be asked to provide one or a couple of each estimations, and the data of many subjects would be combined into an average or median result which we denote by $\phi(x)$. In many cases, these results would be fitted reasonably well by the so-called *Power Law* $\phi(x) = \alpha x^\beta$ or such variants as $\phi(x) = \alpha x^\beta + \gamma$ and $\phi(x) = \alpha(x+\gamma)^\beta$. In the *cross-modality matching* method, the subject is presented with some stimulus from one sensory continuum (e.g., loudness), and is required to select a stimulus from another sensory continuum (e.g., brightness) so as to match the two subjective intensities. Power laws were often also obtained in these situations. In the context of a discussion concerning the measurement of sensation, the difference of forms between Eqn. (4) and the Power law was deemed important. While not much mathematical theorizing was involved in any particular application of these ideas, a real challenge was offered by the need to construct a comprehensive theory linking all important aspects of psychophysical methods and data.

		Responses	
		YES	NO
Stimulus	YES	4(Hit)	-2(Miss)
	NO	-1(False Alarm)	3(Correct Rejection)

Figure 2
An example of payoff matrix. The subject collects 4 monetary units in the case of a correct detection (or hit)

The most ambitious effort in this direction is due to Krantz (1972). For a slightly different approach to direct scaling, see Anderson (1981).

3.2 Functional Equation Methods

Because the data of psychophysical experiments are typically noisy, the theoretician may be reluctant to make specific assumptions regarding the form of some functions entering in the equations of a model. An example of such a model is Fechner Eqn. (2), in which the functions u and F are not specified *a priori*. In such cases, the equations themselves may sometimes specify the functions implicitly. For instance, if we assume that both Weber's Law and Eqn. (2) hold, then Fechner's Law must also hold, that is, the function u must be logarithmic, as in Eqn. (4). Many more difficult cases have been analyzed (cf. Falmagne 1985) (see *Functional Equations in Behavioral and Social Sciences*).

3.3 Signal Detection Theory

Response strategies are often available to a subject in a psychophysical experiment. Consider a situation in which the subject must detect a low intensity stimulus presented over a background noise. On some trials, just the background noise is presented. The subject may have a bias to respond 'YES' on some trials even though no clear detection occurred. This phenomenon prevents a straightforward analysis of the data because some successful 'YES' responses may be due to lucky guesses. A number of 'signal detection' theories have been designed for parsing out the subject's response strategy from the data. The key idea is to manipulate the subject's strategy by systematically varying the *payoff matrix*, that is, the system of rewards and penalties given to the subject for his or her responses. These fall into four categories: correct detection or '*hit*'; correct rejection; incorrect detection or '*false alarm*'; and incorrect rejection or '*miss*.' An example of a payoff matrix is displayed in Fig. 2. (Thus, the subject collects four monetary units in the case of a correct detection, and looses one such unit in the case of a false alarm.)

For any payoff matrix θ, we denote by $p_s(\theta)$ and $p_n(\theta)$ the probabilities of a correct detection and of false alarm, respectively. Varying the payoff matrix θ over conditions yields estimates of points $(p_s(\theta), p_s(\theta))$ in the unit square. It is assumed that (except for experimental errors) these points lie on a ROC (*Receiver Operator Characteristic*) curve representing the graph of some ROC function ρ: $p_n(\theta) \mapsto p_s(\theta)$. The function ρ is typically assumed to be continuous and increasing. The basic notion is that the subject's strategy varies along the ROC curves, while the discriminating ability varies across these curves. The following basic random variable model illustrates this interpretation. Suppose that to each stimulus s is attached a random variable $\mathbf{U}_s$ representing the effect of the stimulus on the subject sensory system. Similarly, let $\mathbf{U}_n$ be a random variable representing the effect of the noise on that system. The random variables $\mathbf{U}_s$ and $\mathbf{U}_n$ are assumed to be independent. We also suppose that the subject responds 'YES' whenever some threshold λ_θ (depending on the payoff matrix θ) is exceeded. We obtain the two equations

$$p_s(\theta) = P(\mathbf{U}_s > \lambda_\theta), \quad p_n(\theta) = P(\mathbf{U}_n > \lambda_\theta) \tag{8}$$

The combined effects of detection ability and strategy on the subject's performance can be disentangled in this model, however. Under some general continuity and monotonicity conditions and because $\mathbf{U}_s$ and $\mathbf{U}_n$ are independent, we get

$$P(\mathbf{U}_s > \mathbf{U}_n) = \int_{-\infty}^{\infty} P(\mathbf{U}_s > \lambda)\, \mathrm{d}P(\mathbf{U}_n \leqslant \lambda) = \int_0^1 \rho(p)\, \mathrm{d}p \tag{9}$$

with ρ the ROC function and after changing the variable from λ to $p_n(\lambda) = p$. Thus, for a fixed pair (s, n), the area under the ROC curve, which does not depend on the subjects' strategy, is a measure of the probability that $\mathbf{U}_s$ exceeds $\mathbf{U}_n$. Note that Eqn. (9) remains true under any arbitrary continuous strictly increasing transformation of the random variables. For practical reasons, specific hypotheses are often made on the distributions of these random variables, which are (in most cases) assumed to be Gaussian, with expectations $\mu_s = E(\mathbf{U}_s)$ and $\mu_n = E(\mathbf{U}_n)$, and a common variance equal to 1. Replotting the ROC curves in (standard) normal–normal coordinates, we see that each replotted ROC curve is a straight line with a slope equal to 1 and an intercept equal to $\mu_s - \mu_n$.

Obviously, this model is closely related to Thurstone's Law of Comparative Judgements. Using derivations similar to those leading to Eqns. (5) and (6) and defining $d'(s, n) = \mu_s - \mu_n$, we obtain $P(\mathbf{U}_s > \mathbf{U}_n) = \Phi(d'(s, n)/\sqrt{2})$, an equation linking the basic signal detectability index d' and the area under the ROC

curve. The index d' has become a standard tool not only in sensory psychology, but also in other fields where the paradigm is suitable and the subject's guessing strategy is of concern. Multidimensional versions of the Gaussian signal detection model have been developed. Various other models have also been considered for such data, involving either different assumptions on the distributions of the random variables $\mathbf{U}_s$ and $\mathbf{U}_n$, or even completely different models (such as 'threshold' models). Presentations of this topic can be found in Green and Swets (1974), still a useful reference, and MacMillan and Creelman (1991) (see *Signal Detection Theory*).

Mathematical models for 'multidimensional' psychophysics were also developed in the guise of *Geometric representations of perceptual phenomena*, which is the title of a recent volume on the topic (Luce et al. 1995; see in particular Indow's chapter).

4. Measurement and Choice

Because they were proccupied with the scientific bases of their discipline, a number of mathematical psychologists have devoted considerable efforts to the elucidation of the foundation of measurement theory, that is, the set of principles governing the use of numbers in the statement and discussion of scientific facts and theories. The reader is referred to the articles of this encyclopedia dealing with this rather specialized topic of philosophy of science, which concerns not only psychology, but science in general. An up-to-date account of the results can be found in the three volumes of *Foundation of Measurement* by D. H. Krantz, R. D. Luce, P. Suppes and A. Tversky in varying orders depending on the volume (see *Measurement Theory: History and Philosophy*).

The literature on Choice Theory is extensive. Contributors come from diverse fields including mathematical psychology, but also microeconomics, political science, and business—the latter two being concerned with the study of voters or consumers' choices. Early on, the literature was dominated by Thurstone's Law of Comparative Judgements, which still remains influential. Many other situations and models have been analyzed however (see *Risk: Theories of Decision and Choice*). We only give a few pointers here. A generalization of the Thurstone model is obtained by dropping the assumption of normality of the random variables $\mathbf{U}_x$ and $\mathbf{U}_y$ in Eqn. (5) (i.e., the last equation in (5) does not necessarily hold). Despite many attempts, the problem of characterizing this model in terms of conditions on the binary choice probabilities, posed by Block and Marschak (in 1960; see Suppes et al. 1989), is still unsolved. In other words, we do not know which set of necessary and sufficient conditions on the choice probabilities $P(x, y)$ guarantee the existence of the random variables $\mathbf{U}_x$, $\mathbf{U}_y$ satisfying the first equation in (5) for all x, y in the choice set. A number of partial results have been obtained by various authors, however.

In the *multiple choice paradigm*, the subject is presented with a subset Y of a basic finite set $\mathscr{X}$ of objects and is required to select one of the objects in Y. We denote by $P(x; Y)$ the probability of selecting x in Y. By abuse of notation, we also write $P(X; Y) = \sum_{x \in X} P(x; Y)$. Suppose that $P > 0$. The *Choice Axiom*, proposed by Luce (1959), states that, for all $Z \subseteq Y \subseteq W \subseteq \mathscr{X}$, we have

$$P(Z; Y)P(Y; W) = P(Z; W) \tag{10}$$

Defining the function $v: x \mapsto P(x; X)$, Eqn. (10) yields immediately (with $Z = \{x\}$ and $W = X$)

$$P(x; Y) = \frac{v(x)}{\sum_{y \in Y} v(y)}$$

for all $Y \subseteq \mathscr{X}$ and $x \in Y$. This model plays an important role in the literature. In the binary case, it has an interpretation in terms of random variables as in the Thurstone model, but these random variables, rather than being Gaussian, have a negative exponential distribution (see Falmagne 1985).

In the general case of such a random variable model for the multiple choice paradigm, we simply suppose that to each x in $\mathscr{X}$ is attached a random variable $\mathbf{U}_x$ such that, for all subsets X of $\mathscr{X}$ and all x in X, we have

$$P(x; X) = P(\mathbf{U}_x = \max\{\mathbf{U}_y \mid y \in X\}) \tag{11}$$

where 'max' stands for 'maximum' (of the random variables in the set X). The characterization problem for this model has been solved by Falmagne 1978). As in the binary case, specific assumptions can be made on the distributions of these random variables.

Models based on different principles have also been proposed for the multiple choice paradigm. For example, in the *elimination by aspects* model, due to A. Tversky (see Suppes et al. 1989), a subject's choice of some object x in a set X is regarded as resulting from an implicit Markovian-type process gradually narrowing down the acceptable possibilites.

Among other important lines of research on choice theory germane of mathematical psychology, we mention *subjective utility theory*—especially as formulated in the work of L. J. Savage—and its recent generalization *rank-dependent utility* (see Luce 2000). For reviews of probabilistic choice models, the reader may consult Luce and Suppes (1965) or Suppes et al. (1989). A sample of some recent works can be found in Marley (1997) (see *Risk: Theories of Decision and Choice*).

5. Response Latency Mechanisms

The latency of a response has been used as a behavioral index of the sensory or mental processes involved in the task since the inception of experimental psychology in the nineteenth century. Many mathematical models are based on F. C. Donders' idea (originally published in 1868–9) that the observed latency of a response is a sum of a number of unobservable components including at least a sensory, a decision, and a motor response part (see, especially, the work of S. Sternberg; cf. Luce 1986). These models make various assumptions on the distributions of the component latencies, which are often taken to be independent. w. G. McGill, for instance, assumes that the component latencies are all distributed exponentially and independently, with possibly different parameters, so that their sum is distributed as a general gamma random variable (see Luce 1986). Another category of models is grounded on the assumption that the observed response results from an unobservable random walk with absorbing barriers as in the work of Donald Lanning or Steve Link. The basic reference in this field is Luce (1986) (see also Townsend and Ashby 1983).

6. Other Topics

From these four traditional areas, research in mathematical psychology has grown to include a wide variety of subjects. Current research includes many aspects of perception, cognition, memory, and more generally *information processing* (cf. Dosher and Sperling 1998). In some cases, the models can be seen as more or less direct descendants of those proposed by earlier researchers in the field. The *multinomial process tree models* (Batchelder and Riefer 2000), for instance, while not dealing in principle with learning, is in the spirit of the Markovian models of the learning theorists. The same remark applies to the *stochastic assessment* models used in *knowledge spaces research* (Doignon and Falmagne 1999; see *Knowledge Spaces*). However, the advent of powerful computers also gave rise to different types of models for which the predictions could be obtained by simulation, rather than by mathematical derivation. Representative of this trend are the *parallel distributed processing* models (Rumelhart and McClelland 1986), the complex memory models such as those reviewed by Clark and Grondlund (1996), the *neural networks* models and the *multidimensional scaling* techniques (see *Multidimensional Scaling in Psychology*).

The research in *psychometrics* concerns the elaboration of statistical models and techniques for the analysis of test results. As suggested by the term, the main objective is the assignment of one or more numbers to a subject for the purpose of measuring some mental or physical traits. In principle, such a topic could be regarded as part of our subject. For historical reasons, however, this line of work has remained separate from mathematical psychology.

7. The Journals, the Researchers, the Society

The research results in mathematical psychology are mostly published in specialized journals such as the *Journal of Mathematical Psychology*, *Mathematical Social Sciences*, *Psychometrika*, *Econometrica* and *Mathématiques, Informatique et Sciences Humaines*. Some of the work also appears in mainstream publications, e.g., *Psychological Review* or *Perception and Psychophysics*. Early on, the research was typically produced by psychologists, and the work often had a strong experimental component. Over the last couple of decades other researchers became interested in the field, coming especially from economics and applied mathematics.

The Society for Mathematical Psychology was founded in 1979. The society manages the *Journal of Mathematical Psychology* via its Board of Editors, and organizes a yearly meeting gathering about 100 participants coming from all over the world. The European Mathematical Psychology Group (EMPG) is an informal association of about 100 scientists meeting every summer at some European university. The first meeting was in 1971. For details regarding the history of the field, see *Mathematical Psychology, History of*.

See also: Measurement Theory: History and Philosophy; Mathematical Learning Theory, History of; Mathematical Psychology, History of; Mathematical Learning Theory; Psychophysics

Bibliography

Anderson N M 1981 *Foundation of Information Integration Theory*. Academic Press, New York

Atkinson R C, Bower G, Crothers E 1965 *An Introduction to Mathematical Learning Theory*. John Wiley & Sons, New York

Batchelder W H, Riefer D M 1999 Theoretical and empirical review of multinomial processing tree modelling. *Psychonomic Bulletin & Review* **6**: 57–86

Bharucha-Reid A T 1988 *Elements of the Theory of Markov Processes and their Applications*. Dover, Minneola, NY (Originally published by McGraw-Hill Book Company, New York, 1960)

Bock R D, Jones L D 1968 *The Measurement and Prediction of Judgement and Choice*. Holden-Day, San Francisco, CA

Boring E G 1950 *A history of experimental psychology*. In: Elliot M (ed.) *The Century Psychology Series*. Appleton-Century-Crofts, New York

Clark S E, Grondlund S D 1996 Global matching models of recognition memory: how the models match the data. *Psychonomic Bulletin & Review* **3**: 37–60

Doignon J-P, Falmagne J-C 1999 *Knowledge Spaces*. Springer, Berlin

Dosher B, Sperling G 1998 A century of information processing theory: vision, attention and memory. In: Hochbery J (ed.) *Handbook of Perception and Cognition at Century's End: History, Philosophy Theory*. Academic Press, San Diego
Estes W K, Suppes P 1959 Foundations of linear models. In: Bush R R, Estes W K (eds.) *Studies in Mathematical Learning Theory*. Stanford University Press, Stanford, CA
Falmagne J-C 1978 A representation theorem for finite random scale systems. *Journal of Mathematical Psychology* **18**: 52–72
Falmagne J-C 1985 *Elements of Psychophysical Theory*. Oxford University Press, New York
Green D M, Swets J A 1974 *Signal Detection Theory and Psychophysics*. E. Krieger Publishing Co., Huntingdon, NY
Hilgard E R 1956 *Theories of Learning*, 2nd edn. Appleton Century Crofts, New York
Indow T 1995 Psychophysical scaling: scientific and practical applications. In: Luce R D, D'Zmura M, Hoffman D, Iverson G J, Romney A K (eds.) *Geometric Representation of Perceptual Phenomena (Papers in Honor of Tarrow Indow for his 70th Birthday)*. Lawrence Erlbaum Associates, Mahwah, NJ
Krantz D H 1972 A theory of magnitude estimation and cross-modality matching. *Journal of Mathematical Psychology* **9**: 168–99
Luce R D 1959 *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York
Luce R D 1986 *Response times—Their Role in Inferring Elementary Mental Organization*. Oxford University Press, New York
Luce R D 2000 *Utility of Certain and Uncertain Alternatives: Measurement-Theoretical and Experimental Approaches*. Lawrence Erlbaum Associates, Mahwah, NJ
Luce R D, Suppes P 1965 Preference, utility and subjective probability. In: Luce R D, Bush R R, Galanter G (eds.) *Handbook of Mathematical Psychology*. Wiley, New York, Vol. III
Luce R D, D'Zmura M, Hoffman D, Iverson G J, Romney A K (eds.) 1995 *Geometric Representation of Perceptual Phenomena (Papers in Honor of Tarrow Indow for his 70th Birthday)*. Lawrence Erlbaum, Mahwah, NJ
MacMillan N A, Creelman C D 1991 *Detection Theory: A User's Guide*. Cambridge University Press, New York
Marley A A J (ed.) 1997 *Choice, Decision and Measurement (Essays in honor of R. Duncan Luce)*. L. Erlbaum, Mahwah, NJ
Miller G A 1964 *Mathematics and Psychology*. Wiley, New York
Norman M F 1972 *Markov Processes and Learning Models*. Academic Press, New York
Rumelhart D E, McClelland J L (eds.) 1986 *Parallel Distributed Processing. Exploration in the Microstructure of Cognition*. MIT Press, Cambridge, MA, Vol. 1
Stevens S S 1957 On the psychophysical law. *Psychological Review* **64**: 153–81
Suppes P, Krantz D H, Luce R D, Tversky A 1989 *Foundations of Measurement*. Academic Press, San Diego, CA, Vol. II
Townsend J T, Ashby F G 1983 *The Stochastic Modelling of Elementary Psychological Processes*. Cambridge University Press, New York
Wasan M T 1969 *Stochastic Approximation. Cambridge Tracts in Mathematics and Mathematical Physics, 58*. Cambridge University Press, Cambridge, MA
Yellott J I Jr 1969 Probability learning with non contingent success. *Journal of Mathematical Psychology* **6**: 541–75

J.-C. Falmagne

Mathematical Psychology, History of

'Mathematical psychology' constitutes the uses of mathematics in extracting information from psychological data and in constructing formal models of behavioral and cognitive phenomena. Its evolution is traced from faltering beginnings in the early nineteenth century to a flowering in the mid-1900s.

1. *What Do Mathematical Psychologists Do?*

During the history of psychology, the role of mathematics in research and theory construction has assumed many different forms. Even before psychology had gained recognition as a science independent of philosophy, efforts were underway to formulate quantitative laws describing unitary relations between indicators of mental processes and temporal or other physical variables. This activity, conducted in the hope of emulating the achievements of physical scientists like Kepler and Galileo has yielded laws describing, for example, the growth of sensation as a function of intensity of a light or sound, the fading of memories over time, and the rate at which people or animals work as a function of the rate at which the environment makes rewards available.

As psychological research expanded in scope and complexity during the early decades of the twentieth century, the search for simple laws largely gave way to the formulation of more elaborate models in which relations between behavioral measures and multiple causal variables are mediated by connections to hypothesized underlying processes. By mid-century, this approach had matured to the point of contributing substantially to the interpretation of many aspects of human cognition and behavior, including the encoding of stimulus representations during perception, searching of memory during recognition and recall, and making decisions in situations involving uncertainty and risk.

The focus of activity of the new specialty broadened as it gained a secure footing in the fabric of psychological sciences. An early phase in which the role of mathematical psychologists was limited to adapting familiar mathematical tools to existing problems of data analysis and theory construction gradually gave way to a new phase in which the goal was to anticipate new problems. Thus, to an increasing extent during the later 1900s, mathematical psychologists have devoted intensive efforts to the investigation of mathematical structures and systems that might offer potentialities for future application in psychological theory construction. Examples include studies of non-Euclidean geometries, algebraic structures, functional equations, graphs, and networks. Results of these investigations, originally conducted at an abstract level, have already found important applications to aspects of visual perception, memory, complex learn-

Table 1
Professional milestones in the history of mathematical psychology

Date	Event
1963	Publication of *Handbook of Mathematical Psychology*
1964	Founding of *Journal of Mathematical Psychology*
1965	Founding of *British Journal of Mathematical and Statistical Psychology*
1965	Publication of Atkinson, Bower, and Crothers' textbook of mathematical learning theory
1968	Founding of Society for Mathematical Psychology
1970	Publication of Restle and Greeno's textbook of mathematical psychology
1971	Formation of European Psychology Group
1973	Publication of Laming's textbook of mathematical psychology
1976	Publication of Coombs, Dawes, and Tversky's textbook of mathematical psychology

ing, psychological measurement, and decision making. (For reviews of these developments, see Dowling et al. 1998, Geissler et al. 1992, Suppes et al. 1994.)

2. *The Route to Mathematical Psychology*

Although precursors can be identified as far back as the early 1700s, mathematical psychology in the sense of a member of the array of psychological sciences took form with remarkable abruptness shortly after the middle of the twentieth century. In the next two sections, the course of development is traced in terms of notable events bearing on both the professional and the scientific aspects of the discipline.

2.1 *Professional Milestones*

A scientific discipline is defined not only by the common research focus of its continually changing membership but by the means it has developed to cope with the central problem of communication. Before 1960, investigators with common interests in the conjunction of mathematics and psychology had no publications devoted to communication and preservation of research results, no conventions to provide for face-to-face exchange of ideas, no textbook listings for mathematical psychology courses in university departments. The suddenness with which the situation changed can be seen in Table 1, which traces the emergence of an academic specialty of mathematical psychology in terms of some professional, or organizational, milestones.

A quantitatively oriented journal, *Psychometrika*, founded in 1936 by the Psychometric Society, has continued publication to the present, but both the journal and the society have specialized in the limited domain of theory and methods pertaining to psychological tests. On the appearance in 1963 of the three-volume *Handbook of Mathematical Psychology*, both psychometricians and psychologists at large must have been surprised at the volume of work extending across general experimental and applied psychology being produced by a heterogeneous but as yet little known post-World War II cohort of mathematical psychologists. Some of the more senior investigators in this growing specialty, among them R. C. Atkinson, R. R. Bush, C. Coombs, W. K. Estes, R. D. Luce, and P. Suppes, had even before publication of the *Handbook* begun informal discussions of ways of meeting the need for a new publication outlet, and these led to the founding of the *Journal of Mathematical Psychology* in 1964. Clearly the train of events was paralleled in countries other than the US, for the *British Journal of Mathematical and Statistical Psychology* (replacing the *British Journal of Statistical Psychology*) began publication the following year. By the middle of the 1970s, the discipline of mathematical psychology, now somewhat more visible, was being well served by its own societies, journals, and textbooks.

Table 2
Scientific milestones in the history of mathematical psychology

Date	Contribution
1738	Measurement of risk
1816	Model for processes of thought
1834	Weber's law
1860	Fechner's law
1904	Factor analysis
1927	Theory of comparative judgment
1944	Game theory

2.2 *Scientific Milestones*

In terms of theoretical and methodological advances, progress along the path from the earliest precursors to the dramatic expansion of mathematical psychology in the 1950s is marked by a long sequence of scientific milestones, from which a small sample is listed in Table 2 and reviewed in this section.

2.2.1 Measurement of risk. Choice of a starting point for the sequence is somewhat arbitrary, but Miller (1964) suggests for this role a monumental achievement of the probabilist D. Bernoulli. In the early eighteenth century as in the late twentieth, coping with risk was a central concern of many people, both rich and poor. Thus there was, and continues to be, need for principled guidance for making wise choices, not only in games of chance, but in everyday economic activities like buying insurance. Bernoulli's contribution was to show how one could, beginning with a few simple but compelling assumptions about choice situations, derive by pure mathematical reasoning the likelihood of success or failure for various alternative tactics or strategies even for situations that had seemed intractably complex.

However, addressing the highly pertinent question of why people generally fail to use the results of Bernoulli's mathematical analyses to make only wise choices had to wait until a later era when mathematical reasoning would be coupled with experimental research.

2.2.2 Modeling processes of thought. The next milestone in Table 2 marks the effort by one of the founders of scientific pedagogy, J. Herbart (also discussed by Miller 1964), to build a mathematical model for human thought in terms of a competitive interplay of ideas. Mathematically, Herbart's model resembles one formulated a century and a half later by the 'behavior theorist' C. L. Hull for the oscillation of excitatory and inhibitory tendencies during generation of a reaction to a stimulus. However, Herbart's preconception that mathematics and experiment are antithetical kept his theory on the sidelines of early psychological research and cost him his chance at greatness.

2.2.3 Weber's law. No such preconception hindered the work of physicists and physiologists who, contemporaneously with Herbart, were attacking the ancient philosophical problem of the connection between mental and physical worlds by means of experiments on relations between stimuli and sensations. Like many others, no doubt, the German physiologist E. Weber noted that the smallest change in a stimulus one can detect depends on its magnitude—a bright light or a loud sound, for example, requiring a larger change than a dim light or a weak sound. Weber went beyond this observation and, by systematic testing of observers in the laboratory with several sensory modalities, determined that the just noticeable change in a stimulus required for detection is a constant fraction of its magnitude, that is,

$$\Delta M/M = K \tag{1}$$

where ΔM is the just noticeble change in stimulus magnitude M, and K is a constant. Equation (1) which became famous as Weber's law, only holds strictly for a limited range of magnitudes in any sensory modality, but it applies so widely to a useful degree of approximation that it appears ubiquitously in human engineering applications.

2.2.4 Fechner's law. The Weber function has entered into many theoretical developments in the realm of sensation and perception. Among these is the work of the physicist G. Fechner, famous for his formulation of a quantitative relation between a sensation and the magnitude of the stimulus giving rise to it. Fechner showed that for all stimulus dimensions to which Weber's law applies, the relation is logarithmic, in its simplest form expressible as

$$S = K\log(R) \tag{2}$$

where S denotes the measured magnitude of sensation produced by a stimulus of physical magnitude R, K is a constant, and log is the logarithmic function. Investigators in later years questioned whether Fechner's formula is the best choice for scaling the relation between stimulus magnitude and sensation, but none have questioned the importance of the work of Weber and Fechner in producing the first glimmerings of the power of mathematics and experimental observation in combination for analyzing psychological phenomena (see *Psychophysical Theory and Laws, History of*).

2.2.5 Factor analysis. The next contribution to merit the term 'milestone,' the founding of factor analysis by C. Spearman about 1904, requires a digression from the chain of research developments listed in Table 2. The data addressed by Spearman constituted not measures of response to stimulating situations but coefficients of correlation (a measure of degree of association) among assessments of academic achievement or talent and scores on psychological tests obtained for a group of school children. An orderly pattern characterizing a matrix of these correlations suggested that all of the abilities tapped depended on a general mental ability factor coupled with some minor factors specific to individual tests. Spearman identified the general factor, termed g, with general intelligence, an interpretation that has been vigorously debated both in the technical literature and the popular press up to the present day. Generalization of Spearman's approach gave rise to a family of methods of 'factor analysis' that for some sets of test data yielded other structures, the most popular being a small set of factors with generality

intermediate between Spearman's *g* and his specific factors. Factor analysis continues to be a major focus of research in psychometrics, but it has remained outside the mainstream of mathematical psychology (see *Psychometrics*).

2.2.6 Theory of comparative judgment. Returning to the main line of development represented in Table 2, the direct consequence of the pioneering accomplishments of Weber and Fechner was the shaping of psychophysics, a quantitative approach to elementary phenomena of sensation and perception that continues a vigorous existence in experimental and engineering psychology. To psychologists at large, psychophysics may seem a narrow specialty, but Fechner, its principal founder, was by no means a narrow specialist. Though he pursued his research in simple experimental situations chosen for their tractability, he foresaw broad applicability for the conception of psychological measurement that emerged from it. In the vein of a proposition put forward by the great French mathematician LaPlace much earlier, Fechner noted that physical goods or commodities, including money, are significant to people only because they generate within us representations on an internalized scale of 'psychic value.' Weber's and Fechner's laws should be applicable to the relation between objectively measurable commodities and psychic values just as between physical stimuli and sensation. Developing formal machinery for operating with the internal scale of value was no trivial matter, however, and had to wait until another intellectual giant, L. L. Thurstone came on the scene nearly three-quarters of a century later.

Thurstone began with the observation that when people make judgments concerning almost any kind of stimulus quality, whether as simple as shades of gray or as complex as attractiveness of pictures or efficiency of workers' performance, the judgments for any given stimulus vary somewhat over repeated tests but that, on average, the judgments exhibit an orderly pattern, as though based on a single attribute that might correspond to an internalized scale of measurement. Thurstone took a measure of the variability of repeated judgments (the 'standard deviation' in statistical parlance) as the unit of measurement on the hypothesized internal ('subjective,' or 'psychological') scale and derived characterizations of types of scales that arise under various conditions. Although the Thurstonian scales were not definable by reference to a physical device like a ruler or a thermometer, they had some of the useful properties of familiar physical scales, as, for example, that different pairs of pictures separated by the same amount on a scale of attractiveness would be judged equally similar on this attribute. The body of scaling theory flowing from Thurstone's work has had almost innumerable applications in ability testing and assessment of performance (see *Multidimensional Scaling in Psychology*).

2.2.7 Game theory. The title of Von Neumann and Morgenstern's *Theory of Games and Economic Behavior* (1944) scarcely hints at its epoch-making role in setting the framework within which theories of human decision making have developed. The scope of the theory includes not only social interactions that involve conflicts of interest but all situations in which individuals or groups make choices among actions or strategies whose outcomes involve uncertainty and risk. von Neumann and Morgenstern converted the intuitive notion of a behavioral strategy, prevalent under the rubic *hypothesis*in early learning theories, to a formal characterization of a decision space, and, with their treatment of psychological value, or utility, forged a close and enduring association between behavioral decision theory and psychological measurement (see *Decision Research: Behavioral*; *Decision and Choice: Economic Psychology*; *Risk: Theories of Decision and Choice*; *Game Theory*).

3. *Mathematical psychology in the 1950s*

The pace with which important developments in mathematical psychology appeared during the preceding two centuries could not have prepared one to foresee the burst of activity in the 1950s that provided most of the material for a multivolume handbook by the end of the decade. Major research initiatives that sprang up and flourished during these years included the first mathematical theories of learning and memory, graph models of group structure, models of signal detection and risky choice, computer simulation of logical thinking and reasoning, and modeling of the human operator as a component of a control system (see *Mathematical Learning Theory, History of*; *Mathematical Psychology*).

What could account for this quantum leap in the evolution of mathematical psychology? A major factor appears to be a convergence of influences from new developments in disciplines outside of psychology:

(a) The advent of stored-program digital computers and, with a slight lag, the branching off from computer science of the specialty of artificial intelligence.

(b) Analyses of the foundations of physical measurement by P. W. Bridgman and N. R. Campbell and explication by G. G. Stevens (1946) of their implications for psychological measurement (see e.g., *Measurement Theory: History and Philosophy*).

(c) The appearance of N. Wiener's *Cybernetics* (1948) and treatments of the mathematics of control systems (see e.g., *Motor Control Models: Learning and Performance*; *Optimal Control Theory*).

(d) The flowering of communication theory (Shannon and Weaver 1949) and exploration by Attneave (1959) of the uses of communication and information theory in psychology (see *Information Theory*).

(e) N. Chomsky's first formal models of grammar (see *Computational Psycholinguistics*).

(f) The work of engineers on detection of signals occurring against noisy backgrounds in communication systems, leading to the theory of ideal observers (see *Signal Detection Theory*).

These events contributed directly to the expansion of mathematical psychology by supplying concepts and methods that could enter directly into the formulation of models for psychological phenomena. But perhaps at least as important, they created a milieu in which there was heightened motivation for students to acquire the knowledge of mathematics that would allow them to participate in the exciting lines of research opened up for the fledgling discipline of mathematical psychology.

See also: Categorization and Similarity Models; Categorization and Similarity Models: Neuroscience Applications; Fechnerian Psychophysics; Information Processing Architectures: Fundamental Issues; Mathematical Learning Theory; Mathematical Learning Theory, History of; Mathematical Psychology; Memory Models: Quantitative; Psychophysical Theory and Laws, History of; Psychophysics; Stochastic Models

Bibliography

Atkinson R C, Bower G H, Crothers E J 1965 *An Introduction to Mathematical Learning Theory*. Wiley, New York

Attneave F 1959 *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results*. Holt, New York

Coombs C H, Dawes R M, Tversky A 1976 *Mathematical Psychology*. Prentice-Hall, Englewood Cliffs, NJ

Dowling C E, Roberts F S, Theuns P 1998 *Recent Progress in Mathematical Psychology*. L. Erlbaum, Mahwah, NJ

Geissler H-G, Link S W, Townsend J T (eds.) 1992 *Cognition, Information Processing, and Psychophysics: Basic Issues*. L. Erlbaum, Hillsdale, NJ

Luce R D, Bush R R, Galanter E 1963 *Handbook of Mathematical Psychology*. Wiley, New York

Miller G A 1964 *Mathematics and Psychology*. Wiley, New York

Restle F, Greeno J G 1970 *Introduction to Mathematical Psychology*. Addison-Wesley, Reading, MA

Shannon C E, Weaver W 1949 *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL

Spearman C 1904 General intelligence objectively determined and measured. *American Journal of Psychology* **15**: 201–93

Stevens S S 1951 Mathematics, measurement and psychophysics. In: Stevens S S (ed.) *Handbook of Experimental Psychology*. Wiley, New York, pp. 1–49

Suppes P, Pavel M, Falmagne J-Cl 1994 Representations and models in psychology. *Annual Review of Psychology* **45**: 517–44

Von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

Wiener N 1948 *Cybernetics*. Wiley, New York

W. K. Estes

Matrifocality

1. Introduction: Definition and Derivation of the Term

Although there is no generally agreed upon definition of matrifocality, and indeed the term has been applied so variously that its meaning is quite vague, for the purposes of this article it is defined as follows: 'matrifocality is a property of kinship systems where the complex of affective ties among mother and children assumes a structural prominence because of the diminution (but not disappearance) of male authority in domestic relations.'

The term 'matrifocal family' was introduced to solve a problem of empirical description as well as to suggest an alternative theoretical view of the domestic organization of rural African-Americans in the then British Guiana (Smith 1956). After considering (and rejecting) such terms as 'matricentral' and 'matriarchal' to refer to the internal relations among members of households, many of which were female-headed (but certainly not all), the term 'matrifocal' was used as follows:

> The household group tends to be matri-focal in the sense that a woman in the status of 'mother' is usually the *de facto* leader of the group, and conversely the husband-father, although *de jure* head of the household group (if present), is usually marginal to the complex of internal relationships of the group. By 'marginal' we mean that he associates relatively infrequently with the other members of the group, and is on the fringe of the effective ties which bind the group together (Smith 1956).

In the last sentence the words "effective ties" is an undetected printer's error; in the manuscript it reads "affective ties" which was, and is, the author's intention.

The term was quickly taken up and widely used in Caribbean studies though not consistently or without debate; it frequently was confused with 'female-headed.' Since the 1950s both the use and the meaning of matrifocality has been extended beyond the Caribbean as it tracked the worldwide revolution in gender relations, in marriage patterns, and in domestic relations. Its appeal as a less precise, but more evocative,

expression has value in pointing to the changes in kinship practices and gender relations in the context of broader socioeconomic and political changes, but for the analysis of those changes something closer to its original meaning is necessary.

2. *The Term 'Matrifocal' and the Context of its Introduction*

Beginning in the 1930s an upsurge of labor unrest and nationalism in the Caribbean colonies led to policies of 'development and welfare.' Family life became a central issue in those policies, on the assumption that anything differing from middle-class European family ideals was unsuited to the modern world. Sociology had developed quite complex theories purporting to show that the isolated nuclear family of man, wife, and children was not only the most highly evolved, but also the most adapted to modern industrial society with its complex division of labor and high degree of social mobility based on male occupations. These ideas converged with the assumption—the false assumption—that a stable nuclear family based upon legal marriage is natural, being found in all human societies (see *Family, Anthropology of*). Given these assumptions, households containing women with children by several fathers, unstable conjugal relations, conjugal partners living together without being legally married, were all considered to be deviant, abnormal, or pathological.

More technical anthropological analysis had shown that counting descent through females did not alter male dominance; however, matrilineal descent, that is, the inheritance or transmission of power and economic resources from a man to his sister's sons (rather than to his own sons), could lead to unstable marriage and complicated domestic organization when the contrary pulls of marriage and parenthood on the one hand, and loyalty to the lineage on the other, precipitated an array of varying domestic arrangements (see Fortes 1949).

The analysis of matrilineal systems was important for countering the ethnocentric assumptions of nuclear family universality and paternal authority patterns; unfortunately, it had little influence on the emerging discussions of matrifocality. A widely cited publication was a paper entitled *A Survey of the Consanguine or Matrifocal Family* in which the author defined the matrifocal family as 'a co-residential kinship group which includes no regularly present male in the role of husband-father' (Kunstadter 1963). Given this definition, it is not surprising that the author classifies extreme matrilineal cases where women and children live with, and under the jurisdiction of, the woman's brother, as 'matrifocal,' but it also excludes cases where a husband father is actually present, and abandons the criterion specified by Smith (1956) who later stated that, the term 'matrifocal'—specifically intended to convey that it is women in their role as mothers who come to be the focus of relationships, rather than head of the household as such—the nuclear family is both ideally normal, and a real stage in the development of practically all domestic groups (Smith 1973). Kunstadter also ignored the developmental cycle of domestic relations, pioneered by Fortes in his study of Ashanti. It was central to the original concept of matrifocality because it analyzed the changing focus of domestic authority and cohesion as children are born and mature. Kunstadter's paper, because of its place of publication, its title, its simplicity, and apparent authority, became the most widely cited source on matrifocality, drastically modifying its meaning by equating it with a static definition of female-headed households.

3. *Family Policy and the Search for the Causes of Poverty*

In European and North American societies it had long been assumed that poverty and disorganized families were at once the cause and the result of each other. These assumptions were given vigorous new life, and a decided racial slant, by the declaration in 1965 of a 'War on Poverty' in the US, when it became evident that the material well-being of the African–American population actually had deteriorated in spite of revolutionary civil rights legislation. A report by the United States Department of Labor (1965) purported to find the cause of continuing social deprivation in family structure, and more specifically in the 'matriarchal' pattern of Negro families in which the roles of husbands and wives are reversed. The term matrifocal had been introduced precisely to avoid this kind of false reading of the data, but both critics and defenders of this report managed to negate that original meaning, moving towards Kunstadter's view and equating it with matriarchal or female headed.

It is ironic that most of the suggested cures for poverty require a return to traditional gender roles (on the assumption that a stable nuclear family based upon marriage is both natural and necessary for the adequate socialization of children), while at the same time gender roles themselves were undergoing a radical transformation as women of all classes increasingly entered the labor force, bore children out of wedlock, and frequently raised them in households without a male head.

4. *Matrifocality and Feminism*

It was inevitable that the term would be taken up in the new wave of feminism that has done so much to transform societies in the latter quarter of the twen-

tieth century. The idea of a stage of human evolution in which women were dominant is not new. The Swiss jurist Bachofen (1861), took Herodotus's descriptions of the matriarchal system of the Lycians as its point of departure, but contributed to nineteenth-century speculations about the development of patriarchy out of primitive promiscuity through a stage of matriarchy (see *Kinship in Anthropology*). The same kind of research into myth and religion has been employed by recent theorists of cognitive or symbolic archaeology, most notably Marija Gimbutas and her followers. Admirers and opponents alike of Gimbutas's theory of a woman-centered stage of European social development have made free use of matrifocality in their commentaries, referring to a matrifocal stage in the development of European civilization (Marler 1997). This kind of generalized use has spread far outside anthropological circles, sometimes being used to refer to almost any female focussed activity, whether motherhood is involved or not. The term has also found its way into discussions of animal behavior.

5. *The Analytic Dimension of Matrifocality*

Matrifocality has been identified by anthropologists in societies too numerous to list here, and including such varied cases as Java, Portugal, Thailand, Italy, South Africa, and Brazil, as well as the *locus classicus* of the Caribbean and urban North America. Many have used the broader definition of female-headed or female-dominated households and pointed to the economic difficulties faced by husband-fathers as a presumed cause of matrifocality, thus, implicitly (if not explicitly) characterizing such families as abnormal forms of the nuclear family. However, all raise the theoretical issue of the relations among class, status, gender, and the family. Matrifocality is never an isolated condition or a simple cultural trait; it must always be considered in the context of the entire social system of which it is a part.

One of the first studies to address that issue was the study of marriage in the context of race and class in nineteenth century Cuba (Martinez-Alier 1974). She showed the effect upon domestic organization of the marginalization of colored men and women in a society where racial purity and legal marriage defined social status and social honor. In later work Smith (1987) argued that the matrifocal family in the Anglophone Caribbean is an integral part of the complex status system that emerged out of the *slave* regime, where men of higher status were able to marry equals and enter nonlegal unions (coresidential or visiting) with women of lower status, thus institutionalizing a dual marriage system that operated at all levels of the social system. It is not poverty that produces matrifocal families but the combination of societal norms, status, and property considerations stressing nuclear families based on legal marriage, coexisting with nonlegal unions, illegitimacy, and matrifocal families at all levels of the hierarchy for those whose unions do not serve to conserve either status or significant property. Another example is O'Neill's demonstration that in rural Portugal, illegitimacy and matrifocal families are closely related to the status and land-owning structure of local communities (O'Neill 1987).

See also: Family and Gender; Family, Anthropology of; Family as Institution; Family Theory: Feminist–Economist Critique; Feminist Theory; Kinship Terminology; Male Dominance; Motherhood: Social and Cultural Aspects; Psychoanalysis in Sociology

Bibliography

Bachofen J J 1861 *Das Mutterrecht*. Krais and Hoffmann, Stuttgart (1967 *Myth, Religion, and Mother Right: Selected Writings of J J Bachofen*. Princeton University Press Princeton, NJ)

Fortes M 1949 Time and social structure: An Ashanti case study. In: Fortes M (ed.) *Social Structure: Essays Presented to Radcliffe-Brown*. Clarendon Press, Oxford, UK

Kunstadter P 1963 A survey of the consanguine or matrifocal family. *American Anthropologist* **65**: 56–66

Marler J (ed.) 1997 *From the Realm of the Ancestors: An Anthology in Honor of Marija Gimbutas*. Knowledge, Ideas & Trends, Manchester, CT

Martinez-Alier V 1974 *Marriage, Class, and Colour in Nineteenth-century Cuba: A Study of Racial Attitudes and Sexual Values in a Slave Society*. Cambridge University Press, London and New York (reprinted as Stolcke V 1989 *Marriage, Class, and Colour in Nineteenth-century Cuba: A Study of Racial Attitudes and Sexual Values in a Slave Society*, 2nd edn. University of Michigan Press, Ann Arbor, MI)

O'Neill B J 1987 *Social Inequality in a Portuguese Hamlet: Land, Late Marriage, and Bastardy, 1870–1978*. Cambridge University Press, New York

Smith R T 1956 *The Negro Family in British Guiana: Family Structure and Social Status in the Villages*. Routledge and Kegan Paul, London, (the full text of this book is now available, with the kind permission of Routledge, on-line at *http://home.uchicago.edu/~rts1/first.htm/)*

Smith R T 1973 The matrifocal family. In: Goody J (ed.) *The Character of Kinship*. Cambridge University Press, Cambridge, UK

Smith R T 1987 Hierarchy and the dual marriage system in West Indian society. In: Collier J F, Yanagisako S J (eds.) *Gender and Kinship: Essays Toward a Unified Analysis*. Stanford University Press, Stanford, CA

United States Department of Labor 1965 *The Negro Family: The Case for National Action*. Office of Policy Planning and Research, US Department of Labor, Washington, DC

R. T. Smith

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Mauss, Marcel (1872–1950)

The father figure of contemporary social and cultural anthropology in France, Mauss was among the most influential social scientists of the early twentieth century. His scholarly activities and achievements were closely connected to the birth, organization, and coming to power in the French academe of the 'Sociological School' headed by Emile Durkheim and gathering among other luminaries Célestin Bouglé, Maurice Halbwachs, Robert Hertz, Henri Hubert, or François Simiand, the intellectual heritage of which is still being largely claimed, discussed, and studied worldwide (see *Durkheimian Studies*, Oxford) as one of the major foundation acts of modern social sciences proper.

Mauss was born on May 10, 1872 in Epinal (Vosges county) in a middle class Jewish family long established in Eastern France and having fully adopted the pattern of French Liberal patriotism. His personal career followed in many respects the rise of the first generation of famous French–Jewish intellectuals represented by Bergson, Durkheim, or Lucien Lévy-Bruhl. He shared with them their Republican commitment, notably as a young activist engaged in the battle in defense of Captain Dreyfus, as one of the earliest contributors (reporting on the cooperative movement) to the socialist journal *L'Humanité* founded by Jean Jaurès and, in the interwar years, as a dedicated expert of the (anti-Bolshevik) French socialist party (notably in matters financial and those related to the stock exchange). Mauss died on February 10, 1950 in Paris.

With formal university training in philosophy (a quasi must for would-be social scientists in his time), Mauss skipped the then almost inescapable stint of secondary school teaching, but also the stages of a normal academic career in the network of the Faculties of Letters. Several study trips and stays in Britain, Holland, and Germany enabled him to commit himself from the outstart to research and the training of scholars. He was first appointed as a lecturer on archaic religions (*The Religion of Peoples Without Civilization*) at the 5th Section (Sciences of Religion) of the *Ecole Pratique des Hautes Etudes* (1901), a unique agency housed in, but independent from, the Sorbonne and aiming at the instruction of specialists in various fields of erudition. He combined later (1926) these courses with seminars on 'descriptive ethnology' at the newly founded *Institut d'Ethnologie* under the aegis of the Sorbonne, only to occupy the chair of sociology at the *Collège de France* (1931), an even more prestigious institution of higher learning. While lecturing there he also shared in a Socratic manner with a number of (often mature) students the insights of a uniquely ingenious armchair anthropologist, until he was forced to early retirement under the German occupation in 1941. His positions in these both central, but academically marginal Parisian institutions helped him to achieve a rather atypical career of a scholar exclusively dedicated to research on exotic societies without visiting any of them and producing neither a thesis, nor even books proper, but, instead, a number of substantial studies which—though highly specialized by their thematic focus—refer equally to various tribal civilizations of Australia, Amerindia, Africa, and Oceania as well as to a vast variety of relevant 'social facts' of the Hindu, Hebrew or Muslim world.

His education, intellectual development, and academic options were directed and supervised by Durkheim, his uncle and life long tutor. Their works—though intimately linked—appear to be considerably divergent. With the benefit of historical hindsight Mauss is credited nowadays as a less theory-bound, more inspiring and heuristically rewarding ancestor for contemporary successors of the 'French School of Sociology,' his personal impact extending upon many significant anthropologists, sociologists, folklorists, political scientists, social psychologists, or even human geographers or social historians starting a career in France after the First World War, especially after 1945. His association with Durkheim took a decisive form by his participation in the empirical elaboration of databanks for Durkheim's *Suicide* (1897) and continued, more importantly, with the editorial work of the *Année sociologique* (12 volumes between 1898 and 1913), the central scholarly organ of the 'Sociological School.' This quasi yearly survey of relevant publications embodied the first major attempt since August Comte (see *Comte, Auguste (1798–1857)*) to organize empirically the burgeoning social sciences into an inter-related system of disciplines. In charge of a large sector of the *Année* covering 'religious sociology,' Mauss became the author of one-quarter, approximately, of the 10,000 odd pages of review articles published in the *Année* under Durkheim, and took over the direction of the *Année*—transformed in the interwar years into the irregularly published *Annales sociologiques*—following the uncle's untimely death in 1917.

In the division of labor of what became the 'French School of Sociology' Mauss was, from the outset, dealt out the sector or archaic civilizations, especially that of 'primitive religion,' strategic field for the experimentation of Durkheim's basic methodological precepts. In this view indeed the main 'social functions' appear to be apprehensible in a straightforward form in archaic societies, so much so that they may inform the interpretation of more complex societal realities as well. Thus, tribal societies—not without ambiguity—can be equated to the simplest, the most essential, and the historically earliest patterns of social organization. To boot, following the 'religious turn' occurring around 1896 in Durkheim's scholarly evolution, religion was considered as central among 'social representations' or objectivations of 'collective consciousness,' instrumental in the integration of archaic and,

by implication, of all social formations. Hence the epistemological priority granted to the study of 'elementary religion' and the allocation of such pursuits to sociology proper as a master discipline, whatever the degree of development, power, extension, or complexity of societies concerned could be. Thus most of Mauss' work centered on comparative problems related to belief systems, rituals, and mental habits, and other collective practices serving as a foundation for 'social cohesion' in extra-European and preindustrial civilizations, could qualify for being sociological, without much reference to anthropology, ethnology, ethnography, or folklore; such terms being practically absent from the topical listing of 'branches of social sciences' in the *Année*.

Mauss' major early studies are the outcome of his brotherly collaboration with Henri Hubert, the ancient historian and archeologist of the 'Sociological School.' They deal with the social functions of '*Sacrifice*' (1899) (biblographic details of Mauss' studies are listed by years of publication at the end of the third voume of his *Oeuvres* (pp. 642–94) and propose 'A general theory of magic' (1904), to which Mauss added a number of personal essays, especially those concerning 'The sources of magic power in Australian societies' (1904) and the first part of his unpublished (and unpresented!) doctoral dissertation on 'Prayer' (1909). Some of these essays will be published in their common book (*Mélanges d'histoire des religions*, 1909), where the authors offer in their introduction the outline of a sociological theory of religion. Denominational systems of faith are social facts and should hence be put on the agenda of sociology. They divide the life world of societies into a sacred and a secular (profane) sphere. The former is invested with essential collective values, so much so that most if not all moral customs, legal, educational, artistic, technical, and scientific practices display religious foundations. Mauss' first ever review article discussed already a book on 'Religion and the origin of penal law' (1896). In the light of his and his companions' comparative studies concerning 'religious representations' (including totemism, 'positive and negative rites' and their mythical rationalizations, the Melanesian notion of *mana* as the paradigm of socio-religious force, the religious origins of values—like that of money, etc.), religion emerges as a fundamental 'social function,' even if secularized modern societies tend to replace it by functional equivalents (like patriotic rituals) to bring about a desired degree of integration.

Mauss will be less systematic (some would say, less dogmatic) than his uncle in the generalization and theoretical exploitation of observations gained from the study of archaic religions. But he shares with Durkheim the conviction that the social prevails over the individual among norms commanding human behavior and, more specifically, that sociohistorical models are the sources of mental categories epitomized in the 'reason' or 'logos' of Western philosophy, permitting the intelligence of time, space, causality, genres, force, etc., though this approach was often condemned as 'sociocentrism' or 'sociological imperialism' by contemporary critics. In concrete terms Mauss' main interest lay, on the one hand, in the exploration of interconnections between collective practices and ideas, in the manner how mental elaborations respond to the organizational patterns of societies contributing thus to their cohesion and, on the other hand, in the exemplification of the arbitrary nature of cultural facts which become meaningful in particular social configurations only. The first focus would lead to decisive methodological insights for the sociology (or the social history) of knowledge. The latter would inspire developments in structuralist social anthropology.

Major statements about the covariation and the fundamental integrity of various fields of collective conduct are made in the essay of comparative anthropology published together with Durkheim on *Primitive social classifications* (1903) and in a *Study of social morphology*, produced with the collaboration of Mauss' disciple Henri Beuchat on *Seasonal Changes of Eskimo Societies* (1906).

In the much cited (and not less criticized) first study a vast array of evidence, drawn from a number of extremely different Australian and American societies, is presented and analyzed in order to illustrate the functional unity of societies (as totalities) in which mental habits, religious ideas, and even basic categories guiding the perception of environment and other—residential, economic, technical—practices are demonstrably interconnected. Thus, 'the classification of things reproduces that of people,' as exemplified in many tribal civilizations where membership in phratries, matrimonial classes, clans, gender, or age groups serves as a principle for the organization of religious and other social activities as well as for the interpretation of reality in 'logical classes.' The spatial situation of the tribe is a basic model for categories of orientation, like cardinal points. Hence the intimation, that categories of reasoning, instead of following universal patterns, are sociohistorical products and even modern science owes some of its principles to primitive categorizations. Such results will later encourage Mauss to affirm the pretensions of 'social anthropology to be once in the position of substituting itself to philosophy by drafting the history of the human mind.'

The study of Eskimo life offers a more empirical evidence for a case of covariation in time of the material substratum—morphology—of societies ('the mass, density, shape, and composition of human groups') and the moral, religious, and legal aspects of collective existence. The approach is, exceptionally, not comparative—though references are made to other exotic and even European societies too—only to be more closely focused on a privileged example where

even the pattern of settlement of families concerned (gathering in winter, dispersion in summer) differs with the seasons, together with the intensity of social interactions, religious activities, and kinship relations. Winter represents the high point of social togetherness accompanied by festivities, the unity of residence of partner families, and various forms of economic, sexual and emotional exchange among their members, generating in them a strong sense of political and moral integrity.

In the second part of his career after World War I Mauss' work appears to be thematically more diversified without losing its initial foci centered upon the ways and means by which society determines individual behavior and the conditions of collective 'solidarity' or 'cohesion,' the latter—while realized thanks to contingent or 'arbitrary' assets—forming an operational whole called 'culture' or 'civilization.' This helps to restate the historical and ethnographically localized nature of mental categories. These problems and ideas inform some major studies published by Mauss in his elderly years like 'Relations between psychology and sociology' (1924), 'The physical impact of the idea of death suggested by society' (1926), 'The problem of civilisations' (1929), 'Bodily techniques' (1935), 'A category of the mind, the concept of person, that of the self' (1938). But the masterpiece emerging from that period remains the famous comparative essay *On Gift*, *Form and Reason of Exchange in Archaic Societies* (1925).

'The gift' is a vast, though fragmentary accomplishment of a program to study primitive economic systems where much of the exchange of goods is carried out by apparent donations accompanied by other—mostly symbolic or ritual—services, which make it a 'total social phenomenon.' The main focus of the study is what some North American Indians name *potlatch*, a system of 'total exchange' of 'agonistic' nature, whereby rivalry between groups, relationships of prestige and force, forms of social inequality and alliance are expressed, confirmed and enforced. The exchange in question may equally imply the circulation of goods, honors, courtesy, rituals, women, etc., and its essential principle is the obligation to offer and accept as well as return gifts, even if the things given hold purely symbolic value only, since many acts of donation consist of the ritualized destruction of a maximum of goods—which is the very meaning of the *potlatch*. Destruction of valuables is in such systems an essential source of power and prestige by the demonstration of one's possessions and the capacity to dispense with them on an apparently voluntary, but in fact strictly compulsory basis. Survivals of such arrangements, proper to several archaic societies, can be traced in the code of behavior of industrial societies as well. Modern contractual trading relations are still often completed or accompanied by moral customs of inescapable reciprocities grounded in gifts, sacrifice, symbolic grants, acts of generosity, the generalization and institutionalization of which (possibly in the regime of social welfare) appears in Mauss' view as a desirable development. He will extend the scope of the study of 'total social facts' in several essays, especially in one on 'Joking relationships' (1926), socially organized delivery of jokes and insults among kins, destined to strengthen the cohesion of the clan by the enforcement of verbal reciprocities.

Mauss' scholarly achievement has left a considerable mark on the social sciences in France and elsewhere. In contemporary French sociological tradition he is considered—alongside may be with Maurice Halbwachs—as the most 'legitimate' ancestor from the 'Sociological School.' The historic importance of his work has been selectively and critically received but warmly appraised by men like Pierre Bourdieu (who is holding a chair of sociology illustrated by Mauss at the *Collège de France*), Roger Caillois, Georges Dumézil, Louis Dumont, Georges Gurvitch, Michel Leiris, Claude Lévi-Strauss (who succeeded Mauss at the *Collège de France*).

See also: Anthropology, History of; Belief, Anthropology of; Civilizational Analysis, History of; Classification: Conceptions in the Social Sciences; Collective Beliefs: Sociological Explanation; Communes, Anthropology of; Community/Society: History of the Concept; Durkheim, Emile (1858–1917); Exchange in Anthropology; Exchange: Social; Folk Religion; Folklore; Functionalism in Anthropology; Habit: History of the Concept; Individual/Society: History of the Concept; Magic, Anthropology of; Potlatch in Anthropology; Religion: Evolution and Development; Religion: Family and Kinship; Religion, History of; Religion: Morality and Social Control; Religion, Phenomenology of; Religion, Sociology of; Ritual; Sacrifice; Social History; Sociology, History of; Solidarity: History of the Concept; Solidarity, Sociology of; Totemism; Tribe

Bibliography

Fournier M 1994 *Marcel Mauss*. Fayard, Paris

Karsenti B 1997 *L'homme total. Sociologie, anthropologie et philosophie chez Marcel Mauss*. Presses Universitaires de France, Paris

James W, Allen N J (eds.) 1998 *Marcel Mauss, A Centenary Tribute*. Berghahn Books, New York

Mauss M 1950 *Sociology et anthropologie*. Presses Universitaires de France, Paris

Mauss M 1968–1969 *Oeuvres*. Éditions de Minuit, Paris, 3 Vols.

Mauss M 19997 *Écrits politiques*. Fayard, Paris

Mauss M 1998 *Lettres écrites à Marcel Mauss*. Presses Universitaires de France, Paris

V. Karaday

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

McCarthyism

The term McCarthyism is an eponym now widely used as an epithet in American politics. It bears the name of a Republican Senator, Joseph R. McCarthy, who triggered widespread controversy in the United States by alleging in the 1950s that a number of individuals holding important posts in the national government were Communists. While McCarthy made this charge when a Democrat, Harry S. Truman, was president, he continued to press his claim even after a Republican, Dwight D. Eisenhower, took over the White House in 1952.

It was during Eisenhower's early years as president that McCarthy's charges of Communists in government resonated most strongly across the country. Allegations of this sort were by no means a novel phenomenon in American politics. Beginning in the 1930s a variety of legislative committees, especially the House Un-American Affairs Committee (HUAC), conducted investigations aimed at exposing the presence of Communists in the ranks of public employees at all levels of government in the United States, as well as in other sectors of American life such as the entertainment industry.

Why, then, should McCarthy's assaults on Communists in government have attracted so much public attention in the 1950s? The answer lies partly in the fact that McCarthy focused his anti-Communist assault upon citadels of power in the American political system, especially the White House and the State and Defense Department. As a result, his charges were a source of immediate fascination for the media, for whom stories about the president and other high executives enjoy a very high level of priority.

So, it is not surprising that McCarthy's pronouncements quickly became front-page material in American newspapers. Equally valuable to McCarthy was the fact that his arrival on the political scene coincided with the emergence of television as the primary medium through which the nation's political dialog was to be conducted. Each step in the unfolding drama of conflict between McCarthy and high officials of the Truman and Eisenhower administrations was thus taken in full view of a national political audience, as a growing number of Americans began to rely on television as the principal source from which they received their information on public affairs.

But if the ascent of McCarthy's career owed a substantial debt to television, this new medium of communication also contributed a great deal to his eventual undoing. In 1954 the Senate decided to hold nationally television hearings on the Senator's allegation that Communists had infiltrated the Defense Department. When it did, McCarthy proved unable to provide convincing evidence that this had in fact happened. From that point on his influence began to wane. Not long afterwards he was censured by his colleagues in the Senate, and he died three years later at the age of 57.

But at the height of his career the power that McCarthy exerted in American politics was indeed extraordinary, dividing the country and attracting both devout support and fierce hostility from widely different sectors of the public. During his early days in office President Eisenhower went to great lengths to avoid giving the Senator any offense, and at the peak of his career McCarthy received a degree of deference from his Congressional colleagues that few legislators in modern times have been accorded. No article of faith was stronger in the legislature at that time than the belief that it would be risking political suicide to oppose McCarthy in public.

As McCarthy's power eventually receded, his decline triggered a variety of efforts to explain how or why a single Senator had been able to exert such enormous influence on the American body politic during a comparatively brief career on the national political scene. Much of this attention was centered on the political setting in which McCarthy's charges erupted, and on features of that period which gave his accusations a credibility they would never otherwise have enjoyed.

1. *The Contemporary Political Setting*

The decade of the 1950s was a period of widespread frustration and uncertainty in American politics. The country had emerged triumphant just a few short years earlier from World War II, the most destructive war of the century. Scarcely had the United States finished celebrating this victory, however, when it found itself entangled in a wide-ranging conflict with the Soviet Union, the nation that had been its most powerful ally in the struggle just concluded. Moreover, ultimate success in the war had only been achieved through the use of atomic weapons against Japan, one of the country's major adversaries. While successful in forcing Japan to surrender, this first use of atomic weaponry opened up the possibility of a major war in the future that would have horrifying effects upon the entire world.

Such fears of calamities looming on the horizon were quickly overshadowed by a development of more immediate concern. In 1950 the Communist regime installed by the Soviets in North Korea launched an invasion aimed at toppling the pro-Western regime in South Korea, a country the United Nations immediately acted to defend with a coalition of military organizations in which the United States played a leading role. Moreover, just when the United Nations forces were on the verge of victory in this conflict, Chinese forces entered the fray on the side of the North Koreans, and the two contending forces soon found themselves locked into a military stalemate.

At the same time a similar impasse was emerging within American domestic politics. Neither of the two major political parties was able to gain a commanding political position as the decade unfolded. In the most remarkable upset of modern presidential politics, the Democrats had won their fifth straight presidential election in 1948, when President Harry S. Truman was returned to office. Except in foreign affairs, however, this victory did not prove to be an empowering event for Truman. A coalition of Republicans and conservative Southern Democrats remained the governing force in Congress on domestic policy issues, as had been true since 1938. It was in this political setting of weak White House leadership in the domestic arena, bitter partisanship in Congress, and the prospect of unending crisis abroad that Senator McCarthy was able to vault into a commanding position in American politics.

2. *McCarthyism in Retrospect*

A half-century since his departure from the American political scene it is not easy to believe that Senator McCarthy was once the formidable force that he appeared to be in his own day. The rest of the twentieth century following his fall from power showed little in the way of lasting effects from the climate of repression that McCarthyism seemed to symbolize or the pressure toward political conformity that his activities appeared to generate. On the contrary, American politics in the decades following the 1950s was for the most part far removed in spirit from the conservative ethos of the McCarthy era. At the height of McCarthy's power in the 1950s, it would be hard to imagine either the Senator or his critics anticipating the broad-ranging set of movements for change that would erupt in the remainder of the twentieth century or the radical steps these new forces would take to achieve their goals.

Witness, for example, the civil rights revolution launched in the United States under the leadership of Martin Luther King in the 1960s, which led to far-reaching changes in the status of Afro-Americans in the United States, a movement that was to be the precursor of protest activities in the pursuit of other political causes in the years ahead. Opponents of the Vietnam War in the 1960s and 1970s used tactics of civil disobedience that would have been unthinkable in the McCarthy era, creating a political climate in the United States in which resistance to official authority became commonplace. These tactics soon became part of the political arsenal of groups working on behalf of a variety of other causes, including women's rights and environmental protection.

What the political and social history of the United States since the 1950s does indeed suggest is that, destructive as McCarthy's activities were for the lives and careers of many individuals whose loyalty to the United States was called into question, he did not inflict any permanent damage on the willingness of Americans to defy authority. Fears that this might happen were voiced widely when the Senator was at the height of his power. Actually, the reverse may actually have occurred. Far from foreshadowing the shape of things to come, McCarthyism may have provoked such a negative reaction to its own excesses that it had the effect of stimulating voices of dissent in American society.

One thing that McCarthyism did produce in its immediate aftermath was a body of searching critiques by scholars who sought to explain why the United States had experienced so much domestic travail over the possibility of domestic subversion at a time when it appeared to have far less to fear from such a development than any of the other modern democracies, most of which had Communist parties competing within their domestic political system to worry about. One historian, Hofstadter (1964) saw McCarthy as a contemporary manifestation of a continuing feature of American political life, the periodic convulsion of fear that the country faced a dire threat from a domestic group with foreign connections, such as Masons or Catholics. Hofstadter described this as 'the paranoid style in American politics.'

Others, however, questioned whether McCarthyism ever had the strength or the staying power sufficient for it to become a significant and continuing force in the American political system, posing a serious threat to democratic institutions in the United States. At the peak of McCarthy's power in the mid-1950s, the task of rooting out alleged subversives from the executive branch had already been bureaucratized, with the establishment of an executive security apparatus increasingly jealous of its own prerogatives in carrying out this task. McCarthy himself seemed to have little ambition or talent for converting his impromptu crusade into anything resembling a more permanent operation. His choice of Roy Cohn and David Schine as his chief aides revealed this incapacity. These aides managed to convert his crusade into a comic event, as they toured Europe seeking to expose the presence of subversive books or librarians lurking in American agencies overseas.

Perhaps the simplest and most persuasive explanation of McCarthy's sudden rise and precipitous fall in American politics has been offered by Polsby (1960, 1974), with support from other leading political scientists, including Seymour Lipset (Lipset and Raab (1970) and Michael Rogin (1967). Polsby argues that the sudden ascent of McCarthy in the 1950s in American political life rested not so much on the public support his charges received, but on the willingness of his Congressional colleagues, largely though not entirely Republicans, to either support or refrain from criticizing his activities. The event that sealed his doom was the outcome of the 1952 election, when, for the first time since 1928, the Republicans

won the presidency. In the wake of this development, McCarthy's continuing attacks on the White House and the executive branch soon became dysfunctional for his partisan supporters, and as his utility for these longstanding allies eroded, so too did their support.

See also: Cold War, The; Communism; Communist Parties; Parties/Movements: Extreme Right; Political Protest and Civil Disobedience; Social Movements, Sociology of

Bibliography

Bell D 1964 *The Radical Right*. Doubleday Anchor Books, Garden City, NY

Hofstadter R 1964 *The Paranoid Style in American Politics and Other Essays*. Wiley, New York

Lipset S M, Raab E 1970 *The Politics of Unreason*. Harper and Row, New York

Polsby N W 1960 Towards an explanation of McCarthyism. *Political Studies* **8**: 250–71

Polsby N W 1974 *Political Promises*. Oxford University Press, New York

Reeves T C 1982 *The Life and Times of Joe McCarthy*. Stein and Day, New York

Rogin M P 1967 *The Intellectuals and McCarthy: The Radical Specter*. MIT Press, Cambridge, MA

Rovere R 1959 *Senator Joe McCarthy*. University of California Press, London

Schrecker E 1998 *Many are the Crimes: McCarthyism in America*. Little, Brown, Boston

F. E. Rourke

Mead, George Herbert (1863–1931)

Together with Charles Peirce, William James, and John Dewey, George Herbert Mead is considered one of the classic representatives of American pragmatism. He is most famous for his ideas about the specificities of human communication and sociality and about the genesis of the 'self' in infantile development. By developing these ideas, Mead became one of the founders of social psychology and—mostly via his influence on the school of symbolic interactionism—one of the most respected figures in contemporary sociology. Compared with that enormous status, other parts of his work like his general approach to action and his ethics are relatively neglected.

1. Life and Context

Mead was born the son of a Protestant clergyman in Massachusetts (South Hadley, February 27, 1863). He spent the majority of his childhood and youth at Oberlin College, Ohio, where his father was appointed professor and where he himself studied. After four years of bread-and-butter employment and intense intellectual struggle with the Darwinian revolution and Kant's moral philosophy, Mead entered graduate study at Harvard and continued at the universities of Leipzig and Berlin, Germany, specializing in questions of psychology and philosophy. After teaching at the University of Michigan (1891–94), Mead was brought by John Dewey to the newly founded University of Chicago where he taught until his death on April 26, 1931. Publishing very little, but increasingly influential through his teaching and his life as an activist citizen during the Progressive Era, Mead's reputation has grown since his death. All his major works were published posthumously, based partly on student notes, partly on unfinished manuscripts from his remaining papers.

2. The Work

In his foundations of social psychology, Mead starts not from the behavior of the individual organism but from a cooperating group of distinctively human organisms, from what he called the 'social act.'

Groups of humans are subject to conditions that differ fundamentally from those of prehuman stages. For human societies, the problem is how individual behavior not fixed by nature can be differentiated yet also, via mutual expectations, be integrated into group activity. Mead's anthropological theory of the origins of specifically human communication seeks to uncover the mechanism that makes such differentiation and reintegration possible.

Charles Darwin's analysis of expressive animal behavior and Wilhelm V. Wundt's concept of gestures were crucial stimuli for Mead's own thinking on this matter. He shares with them the idea that a 'gesture' is a 'syncopated act,' the incipient phase of an action that may be employed for the regulation of social relationships. Such regulation is possible when an animal reacts to another animal's action during this incipient phase as it would react to the action as a whole. If such a connection is working properly, the early phase of the action can become the 'sign' for the whole action and serve to replace it.

For a gesture to have the same meaning for both sides, its originator must be able to trigger in him or herself the reaction that he or she will excite in the partner to communication, so that the other's reaction is already represented inside him or herself. In other words, it must be possible for the gesture to be perceived by its actual originator. Among humans, this is the case particularly with a type of gestures that can also be most widely varied according to the precise situation: namely, vocal gestures. For Mead, they are a necessary condition for the emergence of self-consciousness in the history of the species, but not a sufficient condition (otherwise the path of self-consciousness would, for example, have been open to birds as well).

Mead also regarded as crucial, the typically human uncertainty of response, and the hesitancy facilitated by the structure of the nervous system. These entail that the originator's virtual reaction to their own gesture does not just take place simultaneously with the reaction of their partner, but actually precedes that reaction. Their own virtual reaction is also registered in its incipient phase and can be checked by other reactions, even before it finds expression in behavior. Thus, anticipatory representation of the other's behavior is possible. Perception of one's own gestures leads not to the emergence of signs as substitute stimuli, but to the bursting of the whole stimulus-response schema of behavior and to the constitution of 'significant symbols.' It thus becomes possible to gear one's own behavior to the potential reactions of others, and intentionally to associate different actions with one another. Action is here oriented to expectations of behavior. And since, in principle, one's communicative partners have the same capacity, a binding pattern of reciprocal behavioral expectations becomes the premise of collective action.

This anthropological analysis, which Mead extends into a comparison between human and animal sociality, provides the key concepts of his social psychology. The concept of 'role' designates precisely a pattern of behavioral expectation; 'taking the role of the other' means to anticipate the other's behavior, but not to assume the other's place in an organized social context. This inner representation of the other's behavior entails that different instances take shape within the individual. The individual makes their own behavior (like their partner's behavior) the object of their perception. Alongside the dimension of instinctive impulses, there appears an evaluative authority made up of expectations on how the other will react to an expression of those impulses.

Mead speaks of an 'I' and a 'me.' The 'I' refers in the traditional philosophical sense to the principle of creativity and spontaneity, but in Mead it also refers biologically to the human instinctual make-up. This duality in Mead's use of the term is often experienced as contradictory, since 'instinct,' 'impulse,' or 'drive' are associated with a dull natural compulsion. Mead, however, considers that humans are endowed with a constitutional surplus of impulses, which—beyond any question of satisfaction—creates space for itself in fantasy and can be only channeled by normativization. The 'me' refers to my idea of how the other sees me or, at a more primal level, to my internalization of what the other expects me to do or be. The 'me,' qua precipitation within myself of a reference person, is an evaluative authority for my structuring of spontaneous impulses and a basic element of my developing self-image. If I encounter several persons who are significant references for me, I thus acquire several different 'me's,' which must be synthesized into a unitary self-image for consistent behavior to be possible. If this synthesization is successful, the 'self' comes into being: that is, a unitary self-evaluation and action-orientation which allows interaction with more and more communicative partners; and at the same time, a stable personality structure develops which is certain of its needs. Mead's model, more than Freud's, is oriented to dialogue between instinctual impulses and social expectations.

Mead's theory of personality passes into a developmental logic of the formation of the self that is applicable to both species and individual. Central here are the two forms of children's conduct designated by the terms 'play' and 'game.' 'Play' is interaction with an imagined partner in which the child uses behavioral anticipation to act out both sides; the other's conduct is directly represented and complemented by the child's own conduct. The child reaches this stage when it becomes capable of interacting with different individual reference-persons and adopting the other's perspective—that is, when the reference-person at who the child's instinctual impulses are mainly directed is no longer the only one who counts. The child then also develops a capacity for group 'game,' where anticipation of an individual partner's behavior is no longer enough and action must be guided by the conduct of all other participants. These others are by no means disjointed parts, but occupy functions within groups. The individual actor must orientate himself or herself to a common goal—which Mead calls the 'generalized other.' The behavioral expectations of this generalized other are, for instance, the rules of a game, or, more generally, the norms and values of a group. Orientation to a particular 'generalized other' reproduces at a new stage the orientation to a particular concrete other. The problem of orienting to ever broader generalized others becomes the guiding thought in Mead's ethical theory.

If Mead's introductory lectures on social psychology published as *Mind, Self, and Society* (Mead 1934), and the great series of essays that developed his basic ideas for the first time between 1908 and 1912, are taken as his answer to how cooperation and individuation are possible, then the much less well-known collection of Mead's remaining papers—*The Philosophy of the Act* (Mead 1938)—represents an even more fundamental starting point. The problem that Mead addresses here is how instrumental action itself is possible. In particular, he considers the essential prerequisite for any purposive handling of things: that is, the constitution of permanent objects. His analysis of the ability for role taking as an important precondition for the constitution of the 'physical thing' is a major attempt to combine the development of communicative and instrumental capabilities within a theory of socialization.

In Mead's model, action is made up of four stages: impulse, perception, manipulation, and (need-satisfying) consummation. The most distinctively human of these is the third, the stage of manipulation, whose interposition and independence express the reduced

importance of the instincts in humans and provide the link for the emergence of thought. Hand *and* speech are for Mead the two roots of the development from ape to human. If impressions of distance initially trigger a response only in movements of the body, the retardation of response due to distance and the autonomy of the sphere of contact experience then make possible a reciprocal relationship between eye and hand: the two cooperate and control each other. Intelligent perception and the constitution of objects take place, in Mead's view, when distance experience is consciously related to contact experience. But this becomes possible, he further argues, only when the role-taking capability develops to the point where it can be transferred to nonsocial objects.

A thing is perceived as a thing only when we attribute to it an 'inside' that exerts pressure on us as soon as we touch it. Mead calls this, 'taking the role of the thing.' If I also succeed in doing this by anticipation, I will be able to deal with things in a controlled manner and accumulate experiences of manipulative action. Combined with the cooperation of eye and hand, this means that the body's distance senses can and actually do trigger the experience of resistance proper to manipulation. The distant object is then perceived as an anticipated 'contact value'; the thing *looks* heavy, hard or hot. Only interactive experience allows what stands before me to appear as active (as 'pressing'). If this is correct, social experience is the premise upon which sense perception can be synthesized into 'things.' Mead thereby also explains why at first—that is, in the consciousness of the infant or of primitive cultures—all things are perceived as living partners in a schema of interaction, and why it is only later that social objects are differentiated from physical objects. The constitution of permanent objects is, in turn, the precondition for the separation of the organism from other objects and its self-reflective development as a unitary body. Self-identity is thus formed in the same process whereby 'things' take shape for actors. Mead is thus trying to grasp the social constitution of things without falling prey to a linguistically restricted concept of meaning. Mead develops a slightly different formulation of the same ideas in those of his works that connect up with philosophical discussions of relativity theory and which make central use of the concept of 'perspective.'

Mead's ethics and moral psychology are as much grounded upon his theory of action and his social psychology as they set an axiological framework for these scientific parts of his work. Mead's approach to ethics develops from a critique of both the utilitarian and Kantian positions. He does not regard as satisfactory an orientation simply to the results of action or simply to the intentions of the actor; he wants to overcome both the utilitarian lack of interest in motives and the Kantian failure to deal adequately with the goals and objective results of action. He criticizes the psychological basis common to both ethical theories. Mead argues that the separation between motive and object of the will is a consequence of the empiricist concept of experience, and that beneath the surface this also characterizes Kant's concept of inclination. For Mead, the value of an object is associated with the consummatory stage of the action, so that value is experienced as obligation or desire. According to him the relation expressed in the concept of value cannot be limited either to subjective evaluation or to an objective quality of value; it results from a relationship between subject and object which should not, however, be understood as a relationship of knowledge. The value relation is thus an objectively existing relation between subject and object, but it differs structurally from the perception of primary or secondary qualities. This difference is not due to a higher degree of subjective arbitrariness, but to the reference of values to the phase of need satisfaction rather than the phase of manipulation or perception. The claim to objectivity on the part of scientific knowledge bound up with perception or manipulation is, therefore, a matter of course also as far as moral action is concerned. This does not mean that Mead reduces ethics to one more science among others. For science, in his analysis, investigates the relations of ends and means, whereas ethics investigates the relationship among ends themselves.

Epigrammatically, one might say that for Mead the moral situation is a personality crisis. It confronts the personality with a conflict between various of its own values, or between its own values and those of direct partners or the generalized other, or between its own values and impulses. This crisis can be overcome only by one's own creative, and hence ever risky, actions. Mead's ethics, then, seeks not to prescribe rules of conduct but to elucidate the situation in which 'moral discoveries' are necessary. Expectations and impulses must be restructured, so that it becomes possible to rebuild an integral identity and to outline a moral strategy appropriate to the situation. If this is done successfully, the self is raised to a higher stage, since regard for further interests has now been incorporated into conduct.

Mead attempts to describe stages of self-formation as stages of moral development and, at the same time, as stages in the development of society toward freedom from domination. Orientation to a concrete other is followed by orientation to organized others within a group. Beyond this stage and beyond conflicts between different generalized others, there is an orientation to ever more comprehensive social units, and finally to a universalist perspective with an ideal of full development of the human species. We attain this universalist perspective by attempting to understand all values that appear before us—not relativistically in a nonjudgmental juxtaposition, but by assessing them in the light of a universalist community based upon communication and cooperation. Comprehensive communication with partners in the moral situation,

and rational conduct oriented to achievement of the ideal community, are thus two rules to be applied in solving the crisis. This perspective lifts us outside any concrete community or society and leads to ruthless questioning of the legitimacy of all prevailing standards. In each moral decision is a reference to a better society.

The moral value of a given society is shown in the degree to which it involves rational procedures for the reaching of agreement and an openness of all institutions to communicative change. Mead uses the term 'democracy' for such a society; democracy is for him 'institutionalized revolution.' Individuals do not acquire their identity within it through identification with the group or society as such, in its struggle against internal or external enemies. Mead investigated the power-stabilizing and socially integrative functions of punitive justice, and looked at patriotism as an ethical and psychological problem. He recognized that both are functionally necessary in a society, which, because not everyone can publicly express their needs, requires an artificial unity. For Mead, the generation of a universalist perspective is by no means just a moral demand; he is aware that it is achievable only when all humans share a real context in which to act—something that can come about by means of the world market.

3. *Mead's Influence in Social Theory*

During Mead's lifetime, his influence was almost entirely limited to his students and a few colleagues in Chicago, and to his friend, the leading pragmatist philosopher John Dewey. The paths of influence there joining pragmatist philosophy, functionalist psychology, institutionalist economics, empirical sociology, and progressive social reformism can hardly be disentangled from one another. In the history of philosophy, Mead's main service is to have developed a pragmatist analysis of social interaction and individual self-reflection. This same achievement enabled him, in the age of classical sociological theory, to clear a way for it to escape fruitless oppositions such as that between individualism and collectivism. Mead's grasp of the unity of individuation and socialization defines his place in the history of sociology.

After Mead's death, the school of 'symbolic interactionism' played a decisive role in assuring his influence in sociology. Herbert Blumer, a former student of Mead's, became the founder and key organizer in the USA of a rich sociological research tradition which turned against the dominance of behaviorist psychology, quantitative methods of empirical social research, and social theories that abstracted from the action of members of society. This school, by contrast, emphasized the openness of social structures, the creativity of social actors, and the need for interpretation of the data of social science. Mead came to be seen as the school's progenitor and classical reference, although his work was consulted only fragmentarily. In the dominant postwar theory of Talcott Parsons it remained marginal; Mead's ideas were mentioned, alongside the works of Durkheim, Freud, and Cooley, as important for the understanding of the internalization of norms.

An important strand of the reception of his work can be found in Germany. Jürgen Habermas, in his *Theory of Communicative Action*, identified Mead as one of the crucial inspirers of the paradigm shift 'from purposive to communicative action.' By this time at the latest, Mead was not just considered the originator of one sociological approach among many but as a classical theorist of the whole discipline. The pragmatist renaissance that is working itself out in philosophy and public life has focused attention more on Dewey than on Mead. One can also try to sound the potential of Mead's work and American pragmatism in general for a revision of sociological action theory, the theory of norms and values, and macrosociological theory. The innovative potential of Mead's pragmatic social theory is evident far beyond the narrow field of qualitative microsociological research, for which symbolic interactionism has primarily laid claim to his legacy.

See also: Groups, Sociology of; Interactionism: Symbolic; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Self-development in Childhood; Self: History of the Concept; Social Psychology; Socialization in Infancy and Childhood; Socialization, Sociology of; Sociology, History of

Bibliography

Cook G A 1993 *George Herbert Mead: The Making of a Social Pragmatist*. University of Illinois Press, Urbana, IL

Habermas J 1984/1987 *Theory of Communicative Action*. Beacon Press, Boston, 2 Vols

Hamilton P (ed.) 1992 *G H Mead. Critical Assessments*. Routledge London/New York, 4 Vols

Joas H 1985 *G H Mead. A Contemporary Re-examination of His Thought*, (translated by Raymond Meyer), 2nd edn. 1997. MIT Press, Cambridge, MA

Mead G H 1932 In: Murphy A E (ed.) *The Philosophy of the Present*. Open Court, Chicago, London

Mead G H 1934 In: Morris C W (ed.) *Mind, Self, and Society*. University of Chicago Press, Chicago, IL

Mead G H 1936 In: Moore M H (ed.) *Movements of Thought in the Nineteenth Century*. University of Chicago Press, Chicago, IL

Mead G H 1938 In: Morris C W, Brewster J M, Dunham A M, Miller D L (eds.) *The Philosophy of the Act*. University of Chicago Press, Chicago, IL

Mead G H 1964 In: Reck A J (ed.) *Selected Writings*. Bobbs-Merrill, Indianapolis, IN
Miller D L 1973 *G H Mead: Self, Language, and the World*. University of Texas Press, Austin, TX

H. Joas

Mead, Margaret (1901–78)

Margaret Mead, noted twentieth century American cultural anthropologist, was born December 16, 1901, in Philadelphia, Pennsylvania, to Emily Fogg Mead, sociologist and social activist, and Edward Sherwood Mead, economist and professor at the Wharton School of Finance and Commerce at the University of Pennsylvania. She grew up in a household open to new ideas and social change. Her mother was an ardent proponent of woman's suffrage; her grandmother, Martha Ramsay Mead, who lived with them, was, with her mother, interested in experimenting with the latest ideas in education. Her father, while somewhat progressive, also acted as devil's advocate in lively debates over issues at mealtimes.

Her education was a combination of home schooling and local schools. She attended Buckingham Friends School and graduated from Doylestown High School, both in Bucks County, outside of Philadelphia. She spent her first year of college at DePauw University in Indiana, but found herself an outsider there. She transferred to Barnard, a highly regarded women's college in New York City, where she made life-long friends and from which she graduated in the spring of 1923.

In the fall of 1923 she married Luther Cressman, a young man she had met in Bucks County and been engaged to since leaving for DePauw. He was a college graduate and a graduate of General Theological Seminary in New York, which he attended while she was at Barnard. For both, the ideal marriage was an equal partnership.

They attended graduate school at Columbia University in New York City together, he in sociology, she receiving her Master's degree in psychology, then her Ph.D. in anthropology. At Columbia, she was a student of Franz Boas, the preeminent anthropologist in the US, and part of a distinguished anthropological community that included Melville Herskovits, Ruth Bunzel, Esther Goldfrank, and Ruth Benedict, who began as a mentor, and through the graduate school years became a colleague, close friend, and lover. A bisexual, throughout her life Mead would develop relationships with both men and women. During this time she also met anthropologists from outside New York, of whom Edward Sapir would be the most important.

She defended her dissertation, a piece of library research titled, *An Inquiry into the Question of Cultural Stability in Polynesia*, in the spring of 1925 (published in 1928). After accepting a position as an assistant curator at the American Museum of Natural History in New York upon her return, that fall she set off for her first field trip, to American Samoa in the Pacific. She was gone altogether 9 months, while Luther studied in Europe. The result of that field trip was her book, *Coming of Age in Samoa* (1928), which showed Americans there could be a different system of morality in which premarital sex was accepted, and adolescence itself was not necessarily a time of 'storm and stress.' This book became a best-seller and began the process which, by the time of her death, would make her the best-known anthropologist in the US.

On the trip home, she met New Zealander Reo Fortune, on his way to Britain to study psychology. They fell in love and over the course of 2 years of internal struggle she and Luther decided to divorce. In 1928 she married Reo, who had changed his field of study to anthropology and had already taken his first field trip to the Dobu people in the Pacific, and they embarked on their first field trip together, to the Manus people of the Admiralty Islands in the Pacific. This resulted in her book, *Growing Up in New Guinea* (1930), a study of the socialization and education of Manus children.

Two other field trips with Reo, a summer trip to the American West and a return to the Pacific, this time the mainland of New Guinea, resulted in her books, *The Changing Culture of an Indian Tribe* (1932), which studied assimilation among the Omaha Indians, and *Sex and Temperament in Three Primitive Societies* (1935), a comparative study of three New Guinea peoples, the Arapesh, Mundugumor, and Tchambuli, which gave evidence for the first time of the constructed nature of gender systems. Prior to this, gender had been viewed as biologically innate. Problems in the marriage came to a head on New Guinea, resulting in Mead's and Reo's divorce on their return from the field. She married British anthropologist Gregory Bateson, whom she had met in New Guinea, in 1936, and they set out for their first field trip to Bali in the Pacific, which resulted in their co-authored book, *Balinese Character: A Photographic Analysis* (1942). After this field trip their daughter and Mead's only child, Mary Catherine Bateson, was born in 1939.

During World War II Mead served as Executive Secretary of the National Research Council's Committee on Food Habits which did serious research on nutrition planning and cultural food problems, such as how to help people used to rice adopt wheat flour. She wrote a pioneering popular book, *And Keep Your Powder Dry* (1942), the first by an anthropologist to attempt to use anthropological insights and models to deal with a nonnative culture, that of the US as a whole.

After the war ended, she became involved with Ruth Benedict in a project studying nonnative cultures around the world titled Research in Contemporary

Cultures, popularly known as National Character studies. When Benedict died in 1948, Mead took over as director of the project and brought it to a conclusion in 1952. In 1949, using information from the eight cultures she had done field work in, she wrote *Male and Female: A Study of the Sexes in a Changing World*. In 1950, she and Gregory Bateson divorced, the war having drawn them apart.

In the 1950s Mead decided to revisit the Manus in the Pacific, to see what a difference 25 years had made for them. The result was *New Lives for Old: Cultural Transformation—Manus, 1928–1953*. This began a pattern of revisiting old field-work sites that continued almost until her death and resulted in several other shorter publications.

In 1964, she wrote *Continuities in Cultural Evolution*, which she considered one of her best works. Earlier, in 1961, she had agreed to write columns for *Redbook Magazine*, a popular woman's magazine in the US. This, coupled with her many appearances on television talk shows and radio, along with her lectures throughout the world, made her a celebrity. She retired as a curator at the American Museum of Natural History in 1969.

In 1970, she published *Culture and Commitment: A Study of the Generation Gap*, and in 1971 published a conversation between herself and African–American author James Baldwin, titled *A Rap on Race*. Her autobiography, *Blackberry Winter*, appeared in 1972. She continued teaching classes on a visiting or adjunct status, as she had done throughout her career, as well as lecturing and writing. She developed pancreatic cancer and died in New York City, on November 15, 1978, at the age of 76.

During her lifetime, Mead, beyond her fieldwork, was most noted in the profession of anthropology for her pioneering use of audio-visual techniques in the field, for her continuing efforts to develop field training and techniques for use in the field, and predominantly as a founder and leader of the American 'Culture and Personality' school of anthropology, which, in the 1930s and 1940s tried to meld interdisciplinary, particularly psychological, ideas, and methodologies with those of anthropology. Mead pioneered the use of psychological testing among native populations, worked with psychologists and psychoanalysts to develop or test theories, and fostered a wide variety of interdisciplinary connections throughout her life. After World War II she also became known as an advocate of anthropological studies of nonnative peoples and was a founder of the Society for Applied Anthropology.

In American culture and the world at large, Mead was known for popularizing the ideas of anthropology and applying them to specific areas of life. She was noted as an expert on child-rearing, the family, education, ecology, community building, and women's issues. She coined the phrase, 'postmenopausal zest,' for the renewal and refocusing of women's energies after menopause. She was not afraid to be controversial, advocating trial marriage, the decriminalization of marijuana, and women's rights in the 1960s.

But she had her critics, both during her lifetime and after. During her long career, Mead was criticized as well as lauded by other anthropologists for her popularizing efforts and for her interdisciplinary initiatives. Feminist Betty Friedan, in her 1963 book, *The Feminine Mystique*, accused Mead of falling into biological determinism in her book *Male and Female*, and retarding the cause of women's rights through the 1950s. By the 1970s, educated New Guineans were beginning to look at her fieldwork critically as a product of colonial thinking. This was part of a general movement among former colonial peoples concerning anthropologists who had done field work in their countries.

Her anthropological fieldwork was not questioned seriously until after her death, and the controversy remains ongoing. In his book New Zealand anthropologist Freeman (1983) attacked Mead's fieldwork in *Coming of Age in Samoa*, accusing her of total misunderstanding of Samoan culture, and of describing Samoan culture and customs inaccurately. He depicted her as a young, naive girl deceived by Samoans who lied to her, or made up stories. He suggested that, as a young woman eager to succeed in the eyes of her mentor, Franz Boas, she skewed her work to prove his theory of cultural determinism. While some defended him (e.g., Appell 1984), Freeman's accusations were vigorously critiqued by other anthropologists. Holmes, who had restudied Ta'u, a village where Mead had worked in American Samoa, while he had his own critiques of her work, saw her study overall as generally sound for its time and place (Holmes 1987). Freeman's own use of scientific method was questioned (e.g., Ember 1985, Patience and Smith 1986). Several researchers saw Samoan culture as containing elements of both Freeman's and Mead's studies (e.g., Shore 1983, Scheper-Hughes 1984, Feinberg 1988). Others felt that both had overgeneralized (e.g., Côté 1994, Orans 1996). Some noted the differences between Western Samoa, where Freeman had done his fieldwork in the 1940s and 1960s, and American Samoa, where Mead did hers in the 1920s, and the role of the ethnographer's gender, age, and personality (e.g., Ember 1985, Holmes 1987, Côté 1994). Several also noted Freeman's selective quoting from other anthropologists, other writers on Samoa, and Mead's work, which misrepresented these people's positions (e.g., Weiner 1983, Holmes 1987, Leacock 1988, Shankman 1996).

Freeman's 1999 book focused on arguing that Mead was hoaxed in Samoa by two Samoan girls, who told her exaggerated and untrue stories of Samoan sexual practices (Freeman 1999). But information in a letter from Mead to a friend, Eleanor Steele, May 5, 1926, in

the Margaret Mead Papers at the Library of Congress, Washington, DC, shows that Mead did not rely on these girls for her knowledge of sexual practices.

Arguments continue because Freeman's accusations tapped into deeper questions anthropologists continue to wrestle with, and writing about Mead's work has become a way to focus on them. One of these was a re-evaluation of the roles of biology and culture going on in the 1980s and 1990s, as Freeman was seen as championing biology and attacking Boas and Mead as cultural determinists (e.g., Weiner 1983, Leacock 1988). Another was the question whether anthropology was indeed a science or belonged instead in the humanities. Freeman's definitions of science and objectivity as he questioned Mead's work gave others an opportunity to reflect on the nature of science, of objectivity, of ethnographic authority, and the nature of anthropology (e.g., Ember 1985, Marshall 1993). Freeman saw no room for a relative view of fieldwork. He believed there could only be one truth, not many, and his was the correct vision, which led other writers to reflect on the nature of truth. Freeman tied Mead's supposed inadequacies in the field to her gender, which provoked a strong response (Nardi 1984).

The overall consensus at the turn of the twenty-first century is that both Mead's and Freeman's work contain truths, as both are products of different time periods in Samoan history, the different statuses allowed to the two fieldworkers within the culture of those times due to age and gender, and the different geographical areas studied. Neither contains the whole truth because they are products of one person's viewpoint and one person cannot see everything.

In the 1990s, as the study of colonialism and imperialism gained emphasis, Mead's work, like that of other anthropologists, was seriously critiqued by both native scholars and Western scholars for its pro-Western stance and representations of native people (e.g., Iamo 1992, Newman 1999).

During her lifetime, Mead played a leadership role in anthropology and science. She was elected to membership in the National Academy of Sciences in 1975, and that same year became president of the American Association for the Advancement of Science. She had been president of the Society for Applied Anthropology (1940), the World Federation of Mental Health (1956–57), and the American Anthropological Association (1960). She received over 40 awards during her lifetime, including the Viking Medal in general anthropology and, after her death, the Presidential Medal of Freedom. But her work has become a matter of history to most contemporary anthropologists.

Within American culture as a whole, Mead's lasting legacy will be both the model of her life as scholar and activist and her promotion of broad ideas: the idea of the relativity of sexual mores that differ from culture to culture; the relative shape of the family, ranging from the nuclear, to widely extended by blood and adoption; the idea of the cultural construction of gender, of masculinity and femininity, as categories of human identity; and the connection between child-rearing practices and the shape of a culture.

See also: Age: Anthropological Aspects; Anthropology; Anthropology, History of; Benedict, Ruth (1887–1948); Boas, Franz (1858–1942); Cultural Relativism, Anthropology of; Culture: Contemporary Views; Culture, Production of; Determinism: Social and Economic; Ethnography; Fieldwork in Social and Cultural Anthropology; Gender and Feminist Studies; Gender and Feminist Studies in Anthropology; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender Ideology: Cross-cultural Aspects; Identity in Anthropology; Interpretation in Anthropology; Marriage; Relativism: Philosophical Aspects; Sapir, Edward (1884–1939)

Bibliography

Appell G N 1984 Freeman's refutation of Mead's coming of age in Samoa: The implications for anthropological inquiry. *Eastern Anthropologist* **37**: 183–214

Bateson M C 1984 *With a Daughter's Eye: A Memoir of Margaret Mead and Gregory Bateson*. Morrow, New York

Cassidy R 1982 *Margaret Mead: A Voice for the Century*. Universe Books, New York

Côté J E 1994 *Adolescent Storm and Stress: An Evaluation of the Mead–Freeman Controversy*. Erlbaum, Hillsdale, NJ

Ember M 1985 Evidence and science in ethnography: Reflections on the Freeman–Mead controversy. *American Anthropologist* **87**: 906–10

Feinberg R 1988 Margaret Mead and Samoa: Coming of age in fact and fiction. *American Anthropologist* **90**: 656–63

Freeman D 1983 *Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth*. Harvard University Press, Cambridge, MA

Freeman D 1999 *The Fateful Hoaxing of Margaret Mead: A Historical Analysis of her Samoan Research*. Westview Press, Boulder, CO

Foerstel L, Gilliam A (eds.) 1992 *Confronting the Margaret Mead Legacy: Scholarship, Empire, and the South Pacific*. Temple University Press, Philadelphia, PA

Gordan J (ed.) 1976 *Margaret Mead: The Complete Bibliography, 1925–75*. Mouton, The Hague

Grinager P 1999 *Uncommon Lives: My Lifelong Friendship with Margaret Mead*. Rowman and Littlefield, Lanham, MD

Holmes L D 1987 *Quest for the Real Samoa: The Mead/Freeman Controversy and Beyond*. Bergin and Garvey, South Hadley, MA

Howard J 1984 *Margaret Mead: A Life*. Simon and Schuster, New York

Iamo W 1992 The stigma of New Guinea: Reflections on anthropology and anthropologists. In: Foerstal L, Gilliam A (eds.) *Confronting the Margaret Mead Legacy: Scholarship, Empire, and the South Pacific*. Temple University Press, Philadelphia, PA

Lapsley H 1999 *Margaret Mead and Ruth Benedict: The Kinship of Women*. University of Massachusetts Press, Amherst, MA
Leacock E 1988 Anthropologists in search of a culture: Margaret Mead, Derek Freeman, and all the rest of us. *Central Issues in Anthropology* **8**: 3–23
Marshall M 1993 Wizard from Oz meets the wicked witch of the East: Freeman, Mead, and ethnographic authority. *American Ethnologist* **20**: 604–17
Nardi B 1984 The height of her powers: Margaret Mead's Samoa. *Feminist Studies* **10**: 323–37
Newman L M 1999 *White Women's Rights: The Racial Origins of Feminism in the United States*. Oxford University Press, New York
Olmsted D L 1980 In memoriam Margaret Mead. *American Anthropologist* **82**(2): 262–373
Orans M 1996 *Not Even Wrong: Margaret Mead, Derek Freeman, and the Samoans*. Chandler and Sharp, Novato, CA
Patience A, Smith J W 1986 Derek Freeman and Samoa: The making and unmaking of a behavioral myth. *American Anthropologist* **88**: 157–62
Saler B 1984 Ethnographies and refutations. *Eastern Anthropologist* **37**: 215–25
Scheper-Hughes N 1984 The Margaret Mead controversy: Culture, biology and anthropological inquiry. *Human Organization: Journal of the Society for Applied Anthropology* **43**: 85–93
Shankman P 1996 The history of Samoan sexual conduct and the Mead–Freeman controversy. *American Anthropologist* **98**: 555–67
Shore B 1983 Paradox regained: Freeman's Margaret Mead and Samoa. *American Anthropologist* **85**: 935–44
Weiner A B 1983 Ethnographic determinism: Samoa and the Margaret Mead controversy. *American Anthropologist* **85**: 909–19

M. M. Caffrey

Meaning and Rule-following: Philosophical Aspects

The rule-following considerations consist of a cluster of arguments that purport to show that the ordinary notion of following a rule is illusory; this in turn has profound consequences for the concept of meaning. In response, some have tried to show that it is necessary to revise the ordinary concept of meaning; others have tried to find flaws in the arguments.

1. *Introduction*

It is a common thought that an individual's words mean what they do because the individual is following a rule for the use of that word: a rule that determines what is a correct, and what is an incorrect, use of the term. This idea has been challenged in a cluster of arguments given by Wittgenstein, which in turn have received great impetus in the interpretation given by Saul Kripke. According to Kripke, the central problem is this: there does not seem to be any set of facts that determines that a speaker is following one rule rather than another. In a little more detail: suppose that, in using a certain word, an individual seems to be following a certain rule, R. Then we can always come up with a different rule, R′, which fits all the facts about the individual's use of the word just as well as R. Why is the individual following R rather than R′?

According to Kripke there is no straightforward answer to this question. The ordinary conception of what it is to follow a rule is simply mistaken; so, in turn, is the ordinary conception of what it is for a word to be meaningful. The best that is available is a surrogate for the ordinary conception; what Kripke calls a 'skeptical solution.' This takes the form of a social criterion of rule-following: an individual follows a rule in so far as their behavior conforms to that of the other members of their community. A mistake simply consists in deviation from what the others do. Thus an individual's words mean whatever the other members of their community use them to mean. This, in turn, can be seen as providing the basis for Wittgenstein's famous denial of the possibility of a private language. A private language is impossible since there would be no community-wide use to provide the correct use of its words, and hence the words would have no meaning.

Kripke's presentation of the rule-following worries has engendered a large literature. His skeptical solution has not been widely accepted. Responses have typically fallen into one of two classes. Either it has been argued that there are further facts about an individual, overlooked by Kripke that determine which rule they are following. Or it has been argued that Kripke's whole presentation is vitiated by an over reductionism approach. These responses will be examined in what follows; first it is necessary to examine the nature of the rule-following problem itself.

2. *The Rule-following Problem*

The passages of Wittgenstein that provide the inspiration for the argument appear, in typically compressed and gnomic form, in Wittgenstein (1953, pp. 138–242) and Wittgenstein (1978, Chap. 6). Kripke's account was published in its final form in Kripke (1980), having been widely presented and discussed before; a similar interpretation was given independently in Fogelin 1987 (first edition 1976). There has been some controversy as to how accurate Kripke's presentation of Wittgenstein's views are; a number of authors have embraced the term 'Kripkenstein' to refer to the imaginary proponent of the views discussed by Kripke. The exegetical issue will not be of concern here. Kripke will only with the argument present the concern. (For discussion of the exegetical issues see Boghossian 1991.)

The central idea is easily put. Imagine an individual who makes statements using the sign '+.' For in-

stance, they say '14 + 7 = 21,' '3 + 23 = 26.' It might be thought that they are following the rule that '+' denotes the plus function. But consider the sums using '+' that they have never performed before (there must be infinitely many of these, since they can only have performed a finite number of sums). Suppose that '68 + 57' is one such sum. Now consider the quus function, which is stipulated to be just like the plus function, except that 68 quus 57 is 5. What is it about the individual that makes it true that they have been using '=' to denote the plus function rather than the quus function? By hypothesis it cannot be that they have returned the answer '125' to the question 'what is 68 + 57?,' since they have never performed that sum before.

The immediate response is that the individual meant plus in virtue of having mastered some further rule: for instance, the rule that, to obtain the answer to the question 'what is 68 + 57?' one counts out a heap of 68 marbles, counts out another of 57, combines the two heaps, and then counts the result. But now reapply the worry. How can it be known that by 'count' the individual did not mean 'quount,' where, of course, this means the same as 'count' except when applied to a heap constructed from two piles, one containing 68 objects, the other containing 57, in which case one correctly quounts the pile if one simply returns the answer 5? One might try to fix the meaning of 'count' by some further rule; but this will just invite further worries about what is meant by the words occurring in that rule. Clearly there is a regress. Any rule that is offered to fix the interpretation of a rule will always be open to further interpretations itself.

At this point it might be suggested that it is a fact about the individual's dispositions that determines that they meant plus: the relevant fact is that they would have answered '125' if one had asked them the question 'what is 68 + 57?' But that response falls foul of the normativity constraint on rules. Rules are things that tell us what we ought to do; dispositions are simply facts about what we would do. To think that facts about the rules we follow can be derived from facts about our dispositions is to try, illegitimately, to derive an 'ought' from an 'is.' Thus even if, due to some cognitive blip, the individual would in fact have answered '5' to the question 'what is 68 + 57?,' one would still want to say that by '+' they meant plus, but that they simply made a mistake. If one identifies what their rule requires them to do with what they would in fact do, no possibility is left that they might go wrong. Yet it is part of the very idea of a rule that one can make sense of the idea of breaking it. So one cannot identify the rule that an individual would follow with what they are disposed to do. No answer has been found to the question of what the individual means by their words.

It is important to see just how wide the repercussions of this are. First, note that although Kripke presents the problem as an epistemic one (how can one *know* that the individual meant plus rather than quus), the deeper problem is constitutive: there seems to be nothing for meaning to consist in. Second, note that the problems do not arise just because Kripke restricts himself to behavioral evidence. The argument is that there is nothing in the individual's mind, not just their behavior that can determine what they mean. (Here there is an important difference between Kripke's arguments and the influential arguments for the indeterminacy of meaning given by Quine (1960); Quine does restrict himself to behavioral evidence.) Third, note that it is not just a matter of the meaning of words in a public language. The same considerations apply to the very contents of an individual's thoughts themselves. How can an individual know that their thought that 2 + 2 = 4 is a thought about plus, and not a thought about quus? This makes the worry go much deeper; and it also makes Kripke's skeptical solution much more shocking. For it is one thing to say that the meaning of words in a public language are determined by the behavior of the community; there is a sense in which, given the conventional nature of language, something like that is bound to be true. It is another thing to say that the very contents of an individual's thoughts are somehow determined by the community around them. It is now time to examine the skeptical solution a little more deeply.

3. *The Skeptical Solution*

Recall that the basic problem that arose in identifying following a rule with having a disposition: it gives us no purchase on the idea of making a mistake. But once one identifies following a rule with conforming to the dispositions of the community, the idea of a mistake can come back: an individual makes a mistake in so far as they do not conform. What this approach can make no sense of is the idea of the whole community going wrong. In proposing this solution, Kripke is not suggesting that he has recaptured everything in the ordinary notion of meaning. After all, it is normally thought that the reason that the whole community gives the same answer to 'what is 68 + 57?' is because they have grasped the same concept. But as in Kripke (1982, p. 97), the skeptical solution does not allow us to say that; grasping the same concept isn't some independent state that explains convergence in judgement. Rather, the skeptical solution is meant as a surrogate for our ordinary conception of meaning: it is the best that there is.

However, it is questionable whether it is good enough. It might be accepted that the ordinary philosophical picture of the nature of meaning is a false one, one that the skeptical solution corrects. It would be much worse if the skeptical solution were at odds with the ordinary practice of making judgements. Yet it seems that it is. As Paul Boghossian has noted,

there are plenty of cases in which, for ordinary predictably reasons, the whole community tends to go wrong (Boghossian, 1990, pp. 535–6). Thus consider a good magician. With sleight of hand they can get us all to believe that they have put the egg into the hat; only they know that it is under the table. So the community agrees that the predicate 'is in the hat' applies to the egg; the magician is the odd one out. Yet it would be ridiculous to say, as it seems the skeptical solution must, that the community is right and the magician is wrong. The magician is the one who knows the truth. The rest of the community has simply been tricked.

A natural response to this is to say that not every disposition of the community should be respected. It is important to pick and choose. In this case, the disposition of the magician to say that the egg isn't in the hat should count for far more than that of the rest of the community to say that it is. However, once that move is made, the primary motivation for the skeptical solution has been lost. The whole point was that using the actual dispositions of an individual could not give rise to the possibility of error; so it was necessary to look to the actual dispositions of the community instead. But once one starts to pick and choose amongst the dispositions of the community, saying that some must be given greater weight than others, the possibility of error is reintroduced into the individual case. Here too, if some dispositions are favored over others, then the individual can be thought of as falling into error: they will be in error whenever they go with a less favored over a more favored disposition. To say this is not to deny the role of the community altogether; for instance, it could be that in sorting out the preferred dispositions it will be necessary to have recourse to the dispositions of the experts, and this is a position conferred by the community. But the essential role of the community will have been lost, and with it, what is distinctive about the skeptical solution. Let us now turn to this alternative proposal, and examine in more detail the grounds on which some dispositions might be preferred to others.

4. *Laundering the Dispositions*

The central thought is that the correct application of a term cannot be identified with that which an individual is disposed to apply it to, nor with what the community is disposed to apply it to; but it might be possible to identify it with what the individual or the community is disposed to apply it to under certain circumstances. This is the device that enables some dispositions to be preferred to others. The proposal here parallels certain versions of the disposition theory of color, and useful lessons can be gained from considering that.

According to a crude dispositional theory of color, an object is, say, red, if and only if observers have the disposition to judge it as red. But that crude form of the theory is hopeless. Some observers are color-blind; some circumstances, for instance colored lighting, will distort the color judgements even of someone who is not. So, a more plausible dispositional account of red will say that an object is red if, and only if, competent observers in favorable circumstances have the disposition to judge it as red. This makes the account more plausible; and it reintroduces the possibility of error. Yet it raises a difficulty, that of specifying what competent observers and favorable circumstances are. For if all that can be said about them is that the are, respectively, the people who are good at identifying red things, and the circumstances that are good for such identifications, then the account will be circular.

Parallel issues arise for the rule-following concerns. Suppose it were said that an individual means plus rather than quus by '+' just in case, under favorable circumstances, they have the disposition to use it to denote the addition function i.e., to answer '125' to 'what is 68 + 57?'; '2' to 'what is 1 + 1?' and so on. Now the possibility of error has been reintroduced, since sometimes the circumstances will not be favorable. But how should the favorable circumstances be specified? They cannot be specified as those circumstances in which the individual uses '+' to denote the addition function; for that makes the proposal trivial. Worse, the proposal can provide no support for thinking that the individual does in fact use '+' to mean plus, since it is equally true that they have a disposition to use '+' to denote quus in those circumstances in which they use it to denote the quaddition function.

Moreover, it seems that it won't do to say that favorable circumstances are those in which the individual is thinking hard, and has no distractions, for even here people make mistakes. Everyone does. Of course, in such cases the individual will not be thinking hard enough; but if that is shorthand for 'not thinking hard enough to get it right,' then the proposal has lapsed back to triviality.

What other ways might there be to launder dispositions without triviality? One response, stemming from those who seek to give a biological foundation to meaning, is to seek to spell out competent speakers and favorable circumstances in terms of proper functioning, where this, in turn, is spelled out in terms of the function that has been selected for. This is typically presented as part of a broader project of understanding semantics notions in terms of evolutionary ones. Whatever the merits of the broader project, it faces great difficulties in providing a response to the rule-following worries. One kind of problem arises when we try to give an account of highly abstract terms, terms that do not bring obvious evolutionary advantage to those who grasp them: the case of addition, discussed so far, provides a case in point, let alone the concepts of, say, algebraic topology. But suppose one were to consider more common concepts, like that of *cow*. Could it be argued that the reason a given individual means *cow* by 'cow' rather than, say, *cow unless today is January 25, 2150 in which case horse* is

that they have been selected to do so? It appears not; for, since we have not yet reached January 25, 2150 it will be true of every cow that the individual has in fact met that they are: a cow unless today is January 25, in which case they are a horse. (This example makes clear the similarities that the rule-following problem bears to Goodman's '*New Riddle of Induction*' and the famous example of grue; see (Goodman 1973), and Kripke's discussion 1982, pp. 20, 58–9.) Similar examples could be given using disjunctive concepts, for instance that of being *a cow or a marshmallow*: it is true of every cow that it is either a cow or a marshmallow. For discussion, and a suggested way of avoiding this latter problem, see Fodor (1990, Chaps. 3 and 4). It might be argued that the property of being a cow unless today is January 25, 2150 (in which case a horse), is not a very natural one. That is surely correct. But once that property is ruled out as an eligible referent for the term 'cow' on the grounds of its unnaturalness it is unclear that the talk of proper function is actually doing any work at all. (For an example of a teleological response to the rule-following problem, see Millikan (1990); for discussion of the problems with the approach in general see Boghossian (1989, pp. 537ff).) For a discussion of the advantages to be gained by treating naturalness as a primitive feature of properties, including an application to rule-following, see Lewis (1983, pp. 375–7).

A more promising approach to laundering dispositions is suggested by Philip Pettit (1999). Why not think of the ideal conditions as involving an idealization of our actual practices of correcting ourselves and others in the light of discrepancy and divergence? This reintroduces a social element: individuals mean plus rather than quus by '+' just in case the process of correcting themselves and others in which they actually engage will, if taken as far as it can be taken, result in them using it to denote the addition function. However, there is in this proposal no unwanted guarantee that the community, as it now stands, is right in its use of terms; for the process of correction might not yet have been carried out. Provided that the idealization is not understood in a trivializing way (i.e., as the process that will bring individuals to apply their terms correctly according to what they mean by them) the proposal will not be trivial. Nonetheless a worry remains. It seems that there is no *a priori* guarantee that a community's actual practices, even if idealized, will lead them to use their terms correctly. For couldn't it be that their practices are prey to a deep and systematic error, such that however much they are idealized, the error will remain? Perhaps, to extend the example given before, nature itself is the magician, and the community will never see through its tricks. In such a case, following Pettit's method of resolving differences and contradictions would perhaps lead them to a consistent theory, and one that they would be quite happy with; but it would be a theory involving misapplications of their own concepts—misapplications which, given their cast of mind, they would never be in a position to shake off. All this seems possible. Yet if what is meant by a word is what its users would be disposed to apply it to after they had resolved all discrepancies and divergences, no such errors would be possible. It might seem very implausible that communities are in fact prey to errors of this sort; but it is not clear that there should be an *a priori* guarantee that they are not.

5. *Antireductionist Approaches*

The rule-following worry consists in the thought that we cannot give a characterization of what following one linguistic rule rather than another amounts to. But what kind of characterization is required? It might appear that what Kripke has shown to be unobtainable is a characterization of facts about rule-following in terms of other sorts of facts. What it shows then is that a certain kind of reductionism is unavailable. So another response to the rule-following worries is just to give up on reductionism: conclude that meaning facts are simply not reducible to other sorts of facts, in particular, conclude that they are not reducible to facts about actual or possible behavior. This is the approach taken by McDowell (1984, 1993); see also Wright (1990).

The flight from reductionism has been a feature of much recent analytic philosophy; and social scientists schooled in the *verstehen* approach will find the response natural. But what exactly does a rejection of reductionism with respect to meaning consist in? One response would be to say that each individual meaning fact (that 'plus' means plus, for example) is *sui generis* and primitive; one simply grasps it or one doesn't. Kripke (1982, p. 51) briefly discusses and rejects such an approach and he is surely right that it is most implausible. No matter how skeptical one might be about the existence of analytic equivalences, it cannot be denied that there are systematic connections between different meaning facts; how else could we explain the meanings of new words in terms of old ones? A more plausible claim is that, while there are connections between meaning facts, no meaning fact can be reduced to a nonmeaning fact. This is how proponents of anti reductionism typically conceive of it. The contention is that the rule-following worries only arise if one tries to give an account of meaning facts in terms of facts that make no reference to meaning. (Note: if this really is to be anti reductionism, the position had better be that there is no way whatsoever of deriving meaning facts from nonmeaning facts, not just that there are ways of formulating meaning facts that do not admit of reduction.) Thus, for instance, Paul Horwich denies that there is any reductive substitution for 'F-ness' in the sentence '"Plus" means F-ness'; but he still claims that meaning facts are reducible to facts about use. As he

correctly says, this is a rejection of the nonreductionist position. (See Horwich 1998).

An obvious difficulty with this antireductionist approach concerns how to fit it into a naturalistic world picture. Naturalism has been understood by many to require that every fact is ultimately reducible to a fact of natural science. This general issue of naturalistic reducibility is an enormous one, which cannot be pursued here. Clearly the success of a anti reductive response to the rule-following worries will depend on the success of anti reductive approaches more generally. However, the worry can be somewhat lessened by pointing out that whilst anti reductionists reject the idea that meaning facts are reducible to nonmeaning facts, it is not obvious that they need reject the idea that they *supervene* on natural facts. Reduction requires that, for every meaning fact, a corresponding nonmeaning fact can be given. Supervenience requires only the weaker theory that the class of meaning facts is somehow made true by the class of nonmeaning facts. That is, someone who embraces supervenience without reduction will accept that nonmeaning facts (in particular, facts about usage, belief, and the like) make true the meaning facts, in the sense that there could not be a change in the meaning facts without a change in the nonmeaning facts. Whether such a picture can really be made good, however, remains to be seen. (There is a gesture towards such a picture in McDowell 1984, p. 348; for a fuller discussion of supervenience of meaning on use, without reduction, see Williamson 1994, pp. 205–9.)

See also: Reduction, Varieties of; Verstehen und Erklren, Philosophy of; Wittgenstein, Ludwig (1889–1951); Word Meaning: Psychological Aspects

Bibliography

Blackburn S 1984 The individual strikes back. *Synthese* **58**: 281–301

Boghossian P 1989 The rule-following considerations 5. *Mind* **98**: 507–49

Boghossian P 1991 The problem of meaning in Wittgenstein. In: Puhl K (ed.) *Meaning Skepticism*. De Gruyter, Berlin, Germany

Chomsky N 1986 *Knowledge of Language: Its Nature, Origin and Use*. Praeger, New York, Chap. 4

Fodor J A 1990 *A Theory of Content and Other Essays*. MIT Press, Cambridge, MA

Fogelin R 1987 *Wittgenstein*, 2nd edn. Routledge, London

Goodman N 1973 *Fact, Fiction, and Forecast*, 3rd edn. Bobbs-Merrill, Indianapolis, IN

Horwich P 1998 *Meaning*. Clarendon Press, Oxford, UK

Kripke S A 1982 *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Harvard University Press, Cambridge, MA

Lewis D 1983 New work for a theory of universals. *Australasian Journal of Philosophy* **61**: 343–77

McDowell J 1984 Wittgenstein on following a rule. *Synthese* **58**: 325–63

McDowell J 1993 Meaning and intentionality in Wittgenstein's later philosophy. In: French P, Uehling T, Wettstein H (eds.) *Midwest Studies in Philosophy: The Wittgenstein Legacy 17*. University of Notre Dame Press, Notre Dame, IN

McGinn C 1984 *Wittgenstein on Meaning: An Interpretation and Evaluation*. Blackwell, Oxford, UK

Millikan R 1990 Truth rules, hoverflies, and the Kripke–Wittgenstein paradox. *Philosophical Review* **99**: 323–53

Pettit P 1990 The reality of rule following. *Mind* **99**: 1–21

Pettit P 1999 A theory of normal and ideal conditions. *Philosophical Studies* **96**: 21–44

Quine W V 1960 *Word and Object*. MIT Press, Cambridge, MA

Williamson T 1994 *Vagueness*. Routledge, London

Wittgenstein L 1958 *Philosophical Investigations*, 2nd edn. Blackwell, Oxford, UK

Wittgenstein L 1978 *Remarks on the Foundations of Mathematics*, 3rd edn. rev. Blackwell, Oxford, UK

Wright C 1990a Kripke's account of the argument against private language. *Journal of Philosophy* **81**: 759–78

Wright C 1990b Wittgenstein's rule-following considerations and the central project of theoretical linguistics. In: George A (ed.) *Reflections on Chomsky*. Blackwell, Oxford, UK

R. Holton

Measurement Error Models

The term *measurement error model* is used to denote a regression model, either linear or nonlinear, where at least one of the covariates or predictors is observed with error. If x_j denotes the value of the covariate for the *i*th sample unit, then x_i is unobserved and we observe instead $X_i = f(x_i, u_i)$, where u_i is known as the *measurement error*. The observed (or *indicator*) variable X_i is assumed to be associated to the unobserved (or *latent*) variable x_i via the function f; the form of this association defines the different types of measurement error models described in the literature. For a comprehensive bibliography on this literature, see Fuller (1987) and Carroll (1995).

In the simple additive measurement error model, measurement error has the effect of obscuring the relationship between the response variable and the covariate that is measured with error. This effect is sometimes known as *attenuation*, as in the linear regression case, the ordinary least squares estimator of the regression coefficient associated to the variable that is measured with error is biased towards zero (e.g., Fuller 1987).

The next section describes the effect of additive measurement error in the simple regression case, and notes the relationship between this topic and those in other entries. Subsequent sections describe extensions

to the multivariate and nonlinear regression cases. The article ends with a brief historical overview of the development of measurement error models.

1. Simple Linear Regression and Measurement Error

The simple linear regression model (see *Linear Hypothesis: Regression (Basics)*) is given by

$$Y_i = \beta_0 + \beta_1 x_i + e_i \tag{1}$$

where Y_i is the response observed for the ith sample item, x_i is the explanatory variable (or covariate) measured for the ith sample item, β_0 and β_1 are unknown regression coefficients, and e_i is random error typically assumed to be distributed as a normal random variable with mean 0 and variance σ_e^2. Here we talk about *functional models* when the explanatory variable x is assumed to be a fixed value, and about *structural models* when instead x is assumed to be a random draw from some distribution. Typically, we choose $x \sim N(\mu_x, \sigma_x^2)$ (e.g., Fuller 1987). [Carroll (1995), however, makes a different distinction between functional and structural models. They refer to *functional modeling* when the variable x is either fixed or random but minimal assumptions are made about its distribution, and to *structural modeling* when a parametric form is chosen for the distribution of x. Carroll (1995) argues that when a consistent parameter estimate under the structural model formulation can be found, it tends to be robust to misspecification of the distribution of x.]

It is well known that the ordinary least squares estimator (see *Linear Hypothesis: Regression (Basics)*) of β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2}$$

is unbiased and has smallest variance among linear unbiased estimators. If $x \sim N(\mu_x, \sigma_x^2)$ and $Cov(x, e) = 0$, then the estimator $\hat{\beta}_1$ in (2) is the maximum likelihood estimator of β_1 (see *Estimation: Point and Interval*).

1.1 Models for Measurement Error

Suppose now that we cannot observe the explanatory variable x_i directly. Instead, we observe the *manifest* or *indicator* variable X_i which is a noisy measurement of x_i. We distinguish two different types of measurement error models depending on the relationship between X_i and x_i:

(a) Error models, discussed in this entry, where we model the distribution of X given x.

(b) Regression calibration models, where the distribution of x given X is of interest.

Within the class of error models, we distinguish two variants: *classical additive error models*, and *error calibration models*. The *classical additive error model* establishes that the observed variable X is an unbiased measure of x. That is

$$X = x + u \tag{3}$$

where $u \sim (0, \sigma_u^2)$. As an example, consider the problem of estimating the relationship between serum cholesterol levels and usual intake of saturated fat for a sample of individuals. Suppose further that intake of cholesterol for an individual is measured by observing the individual's intake over a randomly chosen 24-hour period. Because daily nutrient intake varies from day to day for an individual, the intake measured on any one day is a noisy measurement of the individual's usual or habitual intake of the nutrient.

We talk about *error calibration models* when the observed variable is a *biased* measurement of the variable of interest. In this case, we use a regression model to associate the two variables

$$X = \alpha_0 + \alpha_1 x + u \tag{4}$$

where as before, $E(u) = 0$, but now $E(X) = \alpha_0 + \alpha_1 x$. As an example, consider the problem of assessing the habitual daily alcohol consumption for a sample of individuals. Typically, individuals in the sample are asked to record their alcohol consumption over a short period, for example over a week. The mean of the observed alcohol intakes over the seven days cannot be assumed to be an unbiased estimator of the usual daily alcohol consumption, as there is evidence (see Freedman et al. 1991; Carroll 1995) that individuals tend to under-report the amount of alcohol they drink. Suppose in addition that the under-reporting is not correlated with alcohol consumption. In this case, the measurement X is biased for x and must be corrected. If information about relationship (4) is available, the measurement X can be *calibrated* by using $\alpha_1^{-1}(X - \alpha_0)$.

We briefly discuss *regression calibration models* later in this article.

1.2 The Effect of Measurement Error

Consider the simple example of a linear regression with additive measurement error in the covariate, as in expressions (1) and (3). Further, assume that the measurement error is independent of the true measurements x and of the errors e. If $X = x + u$, $x \sim (\mu_x, \sigma_x^2)$ and $u \sim (0, \sigma_u^2)$ then the regression coefficient estimated using the noisy measurements is biased towards zero. If $\hat{\gamma}_1$ is the estimated regression coefficient obtained from using X in the regression equation, then

$$E(\hat{\gamma}_1) = \lambda \beta_1 \tag{5}$$

where β_1 is the regression coefficient associated to x and λ is called a *reliability ratio* (e.g., Fuller 1987,

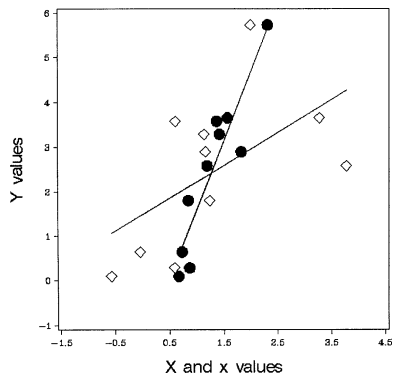

Figure 1
Simulated example. Black circles represent observations without measurement error. Diamonds represent the observations subject to measurement error in the covariate. The steeper line is the fitted regression line in the absence of measurement error, while the flatter line shows the fitted regression line in the presence of measurement error.

Sect. 1.1; Carroll 1995, Sect. 2.2.1), computed as

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \tag{6}$$

Clearly, $\lambda < 1$, and approaches 1 as the variance of the measurement error goes to zero. Therefore, ordinary least squares estimation in the model with X instead of x produces an estimator of the regression coefficient that is biased towards zero, or is *attenuated*.

Measurement error has in addition an effect on the variability of points around the estimated regression line. Note that under model (1) with measurement error as in (3),

$$var(Y \mid X) = \sigma_e^2 + \frac{\beta_1^2 \sigma_u^2 \sigma_x^2}{\sigma_x^2 + \sigma_u^2} \tag{7}$$

Thus measurement error has the effect of increasing the noise about the regression line (Carroll 1995).

1.3 A simple simulated example

To illustrate the effect of measurement error on regression estimation, we constructed 10 responses $Y = \beta_0 + \beta_1 x + e$, where $\beta_0 = 0, \beta_1 = 2, e \sim N(0, 1)$, and $x \sim N(1, 1)$. We then constructed a noisy measurement of x by adding measurement error u where $u \sim N(0, 3)$. We fitted two simple linear regression models, using x and using X as the covariate in the model. The data, together with the two fitted regression lines are shown in Fig. 1. In the figure, the dark dots represent the true measurements, while the diamonds represent the noisy measurements. Note that the regression line fitted to the noisy data is attenuated towards zero. In fact, in this simulated example, the true attenuation coefficient or reliability ratio is only $\frac{1}{4}$.

Attenuation of the slope is greatly dependent on the simple additive error model. In other cases, the ordinary least squares estimator of the regression coefficient may *overestimate* the true slope. Consider, for example, the case where X is a *biased* measurement of x, and in addition, the measurement error is correlated with the equation error. That is, X is as in Eqn. (4) and $\mathrm{Corr}(e, u) = \rho_{eu}$. It can be shown that

$$E\{\hat{\gamma}_1\} = \frac{\beta_1^2 \alpha_1 \sigma_x^2 + \rho_{eu} \sigma_e \sigma_u}{\alpha_1^2 \sigma_x^2 + \sigma_u^2} \tag{8}$$

where, as in Eqn. (5), $\hat{\gamma}_1$ denotes the estimated regression coefficient obtained by using X in the model, and β_1 is the true slope. Notice that, depending on the values of α_1 and of ρ_{eu}, $\hat{\gamma}_1$ may actually overestimate β_1.

2. Multiple Linear Regression and Measurement Error

Extension of results presented in Sect. 1 to the case of multiple predictors is not straightforward, even if only one of the predictors is subject to measurement error and the measurement error model is as in Eqn. (3).

Let

$$Y = \beta_0 + \beta_1 x + \beta_w' \mathbf{w} + e \tag{9}$$

where now $\mathbf{w}$ is a $p \times 1$ vector of predictors that are measured without error, β_w is the vector of regression coefficients associated to $\mathbf{w}$, and x is measured with simple additive measurement error as in Eqn. (3). Furthermore, let $\sigma_{x|w}^2$ denote the variance of the residuals in the regression of x on $\mathbf{w}$. If model (9) is fitted using X in place of x, it can be shown (e.g., Gleser et al. 1987) that, as in Eqn. (5), the estimated regression coefficient for X is $\hat{\gamma}_1 = \lambda \beta_1$, where

$$\lambda = \frac{\sigma_{x|w}^2}{\sigma_{x|w}^2 + \sigma_u^2}. \tag{10}$$

Notice that when x and $\mathbf{w}$ are uncorrelated, λ in Eqn. (10) reduces to Eqn. (6). Gleser et al. (1987) show that when one of the covariates is measured with error, the regression coefficients associated to $\mathbf{w}$ that are measured without error are also biased, unless x and $\mathbf{w}$ are uncorrelated.

If now $\mathbf{x}$ is a $q \times 1$ vector of covariates measured with error, and $\mathbf{w}$ is as above, then the ordinary least

squares estimates have expectation

$$E\begin{bmatrix}\hat{\gamma}_x\\ \hat{\gamma}_w\end{bmatrix} = \begin{bmatrix}\sum_{xx}+\sum_{uu} & \sum_{xw}\\ \sum_{wx} & \sum_{ww}\end{bmatrix}^{-1} \times \left\{\begin{bmatrix}\sum_{xx} & \sum_{xw}\\ \sum_{wx} & \sum_{ww}\end{bmatrix}\begin{bmatrix}\beta_x\\ \beta_w\end{bmatrix}+\begin{bmatrix}\sum_{ue}\\ \mathbf{0}\end{bmatrix}\right\}. \quad (11)$$

The ordinary least squares estimate of the regression coefficients will be unbiased only when $\sum_{uu} = 0$ and $\sum_{ue} = 0$.

3. *Estimation in Measurement Error Models*

Consider first the simple model (1) with measurement error as in Eqn. (3). Assume that $(x, e, u)' \sim N[(\mu_x, 0, 0)', \text{diag}\,(\sigma_x^2, \sigma_e^2, \sigma_u^2)]$, and suppose that σ_u^2 is known. Under this model, the vector, $(\mathbf{Y}, \mathbf{X})$ is distributed as a bivariate normal vector, and therefore, the sample first- and second-order moments form a set of sufficient statistics for the parameters. Fuller (1987, Sect. 1.2.1) and Carroll (1995, Sect. 2.3.1) derive *method of moments* (see *Estimation: Point and Interval*) parameter estimates that can be obtained from a one-to-one mapping from the minimal sufficient statistics to the vector $(\hat{\mu}_x, \hat{\sigma}_x^2, \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_e^2)$. For example,

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y}-\hat{\beta}_1\bar{X}\\ \hat{\beta}_1 &= (s_X^2-\sigma_u^2)^{-1}s_{XY}\end{aligned} \quad (12)$$

where $\bar{Y}$ and $\bar{X}$ are the sample means of Y and X, and s_X^2 and s_{XY} are the sample variance of X and the sample covariance between X and Y, respectively. Fuller (1987) shows that when σ_u^2 is known and as $n \to \infty$,

$$n^{1/2}[(\hat{\beta}_0-\beta_0), (\hat{\beta}_1-\beta_1)]' \xrightarrow{\mathscr{L}} N\{(0, 0)', \mathbf{T}\} \quad (13)$$

where $(\hat{\beta}_0, \hat{\beta}_1)$ are as in (12) and $\mathbf{T}$ is a positive definite matrix (see Fuller 1987, Sect. 1.2.2).

Suppose now that rather than assuming that σ_u^2 is known, we consider the case where the *relative* sizes of the variances $\sigma_u^{-2}\,\sigma_e^2$ are known. We use δ to denote the ratio of variances. Under the assumption of normality of $(x, e, u)'$ as above, method of moments estimators for $(\beta_0, \beta_1, \sigma_u^2)$ are

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y}-\hat{\beta}_1\bar{X}\\ \hat{\beta}_1 &= \frac{s_Y^2-\delta s_X^2+[(s_Y^2-\delta s_x^2)^2+4\delta s_{XY}^2]^{1/2}}{2s_{XY}}\\ \hat{\sigma}_u^2 &= \frac{s_Y^2+\delta s_X^2-[(s_Y^2-\delta s_X^2)^2+4\delta s_{XY}^2]^{1/2}}{2\delta}\end{aligned} \quad (14)$$

where for $\hat{\sigma}_u^2$ we choose the smallest root of the quadratic expression above, so that $\hat{\sigma}_u^2 < s_X^2$. Fuller (1987, Theorem 1.3.1) shows that, for $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_u^2)'$, and $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1, \sigma_u^2)'$ (as given in Eqn. (14)), as $n \to \infty$,

$$n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \xrightarrow{\mathscr{L}} \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}) \quad (15)$$

for $\boldsymbol{\Gamma}$ a positive definite covariance matrix. Asymptotic normality of $(\hat{\beta}_0, \hat{\beta}_1)$ holds even if the distribution of x is not normal.

The estimators above can be extended to the case where a vector $\mathbf{x}$ is measured with error. While in the simple linear regression case results apply in the functional and in the structural model framework, in the multivariate case we must distinguish between models in which $\mathbf{x}$ is assumed to be random and those in which $\mathbf{x}$ is assumed to be fixed. We consider first the *structural model* set-up, where $\mathbf{x}$ is taken to be random. In this case, the maximum likelihood estimators of model parameters derived under the normal model can be shown to be consistent and asymptotically normal under a wide range of assumptions.

Let

$$\begin{aligned}Y &= \beta_0+\mathbf{x}'\beta_1+e\\ \mathbf{X} &= \mathbf{x}+\mathbf{u}\end{aligned} \quad (16)$$

where

$$\begin{bmatrix}\mathbf{x}\\ e\\ \mathbf{u}\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ 0\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sum_{xx} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \sigma_e^2 & \sum_{eu}\\ \mathbf{0} & \sum_{ue} & \sum_{uu}\end{bmatrix}\right) \quad (17)$$

Here, $\sum_{uu}$ and $\sum_{eu}$ are known. The maximum likelihood estimators of (β_0, β_1) are

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y}-\bar{\mathbf{X}}\hat{\beta}_1\\ \hat{\beta}_1 &= \left(\mathrm{S}_{XX}-\sum_{uu}\right)^{-1}(\mathrm{S}_{XY}-\sum_{ue})\end{aligned} \quad (18)$$

where $\mathbf{S}_{XX}$ is the sample covariance matrix of $\mathbf{X}$ and $\mathbf{S}_{XY}$ is the covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$. A modified maximum likelihood estimator of (β_0, β_1) for the case where $\sum_{uu}$ and $\sum_{eu}$ are unknown and must be estimated from the data is given by Fuller (1987, Sect. 2.2).

If the vector $\mathbf{x}$ is taken to be fixed, rather than random, the method of maximum likelihood fails to produce consistent estimators of the parameters in the model. This is because the vector of unknown model parameters $(\beta_0, \beta_1, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \sigma_e^2)$ is $n+3$-dimensional, but the sample size is only n. In this case, then, the number of parameters increases with sample size.

Estimation of model parameters in these cases has been discussed by, for example, Morton (1981).

Consider now model (16) with assumptions as in (17), and suppose that we do not know $\sum_{uu}$ and $\sum_{eu}$ as above. Instead, let $\boldsymbol{\varepsilon} = (e, \mathbf{u})' \sim N(\mathbf{0}, \sum_{\varepsilon}\sum_{\varepsilon})$, where $\sum_{\varepsilon\varepsilon} = \boldsymbol{\Omega}_{\varepsilon\varepsilon}\,\sigma^2$, and $\boldsymbol{\Omega}_{\varepsilon\varepsilon}$ is known. In this case, the maximum likelihood estimators of β_1 and σ^2 are

$$\begin{aligned}\hat{\beta}_1 &= (\bar{\mathbf{X}}'\bar{\mathbf{X}} - \hat{\lambda}\Omega_{uu})^{-1}(\bar{\mathbf{X}}'\bar{\mathbf{Y}} - \hat{\lambda}\Omega_{ue})\\ \hat{\sigma}^2 &= (q+1)^{-1}\hat{\lambda}\end{aligned} \tag{19}$$

where $\bar{\mathbf{X}}'\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}'\bar{\mathbf{Y}}$ are averages taken over the n sample observations, and $\hat{\lambda}$ is the smallest root of $|\bar{\mathbf{Z}}'\bar{\mathbf{Z}} - \lambda\,\Omega_{\varepsilon\varepsilon}|$ with $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$. The estimator of β_1 is the same whether $\mathbf{x}$ is fixed or random. Under relatively mild assumptions, the estimator $\hat{\beta}_1$ is distributed, in the limit, as a normal random vector (see Fuller 1987, Theorem 2.3.2).

4. Related Methods

4.1 Regression with Instrumental Variables

In the preceding sections we have assumed implicitly that when the variance of the measurement error is not known, it can be estimated from data at hand. Sometimes, however, it is not possible to obtain an estimate of the measurement error variance. Instead, we assume that we have available another variable, which we denote Z, and which is known to be correlated with x. If Z is independent of e and u, and Z is correlated with x, we say that Z is an *instrumental variable* (see *Instrumental Variables in Statistics and Econometrics*). Fuller (1987) presents an in-depth discussion of instrumental variable estimation in linear models (Sects. 1.4 and 2.4).

He shows that method of moments estimators for the regression coefficient can be obtained by noticing that the first and second sample moments of the vector (Y, X, Z) are the set of minimal sufficient statistics for the model parameters. The resulting estimators are

$$\begin{aligned}\hat{\beta}_1 &= s_{XY}^{-1}s_{YZ}\\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1\bar{X}\end{aligned} \tag{20}$$

where s_{XZ} and s_{YZ} are the sample covariances of the covariate and the response variables with the instrumental variable, respectively. Because the sample moments are consistent estimators of the population moments, the estimators in (20) are consistent estimators of β_0 and β_1. Notice, however, that if the assumption $\sigma_{xZ} \neq 0$ is violated, the denominator in the expression for $\hat{\beta}_1$ above estimates zero and the estimator is therefore not consistent.

Amemiya (1985, 1990a, 1990b), Carroll and Stefanski (1994) and Buzas and Stefanski (1996) discuss instrumental variable estimation in nonlinear models.

4.2 Factor Analysis

Factor analysis is heavily used in psychology, sociology, business, and economics (see *Factor Analysis and Latent Structure: Overview*). We consider a model similar to the model used for estimating an instrumental variable, with a few additional assumptions. The simplest version of a factor model is

$$\begin{aligned}Y_1 &= \beta_{01} + \beta_{11}x + e_1\\ Y_2 &= \beta_{02} + \beta_{12}x + e_2\\ X &= x + u\end{aligned} \tag{21}$$

Here, (Y_1, Y_2, X) are observed. We assume that the coefficients associated to x are nonzero, and that $(x, e_1, e_2, u)' \sim N[(\mu_x, 0, 0, 0)', \text{diag}\,(\sigma_x^2, \sigma_{e1}^2, \sigma_{e2}^2, \sigma_u^2)]$.

In factor analysis, the unobservable variable x is denoted *common factor* or *latent factor*. For example, x might represent intelligence, and the observed variables (Y_1, Y_2, X) might represent the individual's scores in three appropriate tests. The ratio of the variance of x to the variance of the response variable is called the *communality* of the response variable. For example, the communality of Y_1 above is calculated as

$$\lambda_1^2 = \frac{\beta_1^2\sigma_x^2}{\beta_1^2\sigma_x^2 + \sigma_{e1}^2} \tag{22}$$

Note that λ_1^2 is equal to the reliability ratio that was presented in Sect. 1.

Estimators for β_{11} and for β_{12} can be obtained using the same methods that were used for estimating instrumental variables in Sect. 4.1. For example, $\hat{\beta}_{11}$ is the instrumental variable estimator of β_{11} under the model $Y_1 = \beta_{01} + \beta_{11}x + e$ using Y_2 as an instrumental variable. Similarly, an estimator for β_{12} is obtained using Y_1 as an instrumental variable.

For an extensive discussion of factor analysis, refer to Jöreskog (1981) and Fuller (1987).

4.3 Regression Calibration

Suppose that for each sample item we observe Y, X, and $\mathbf{w}$, where $X = x + u$, and $\mathbf{w}$ is a vector of covariates measured without error. The idea behind regression

calibration is simple. Using information from X and $\mathbf{w}$, estimate the regression of x on $(X, \mathbf{w})$. Then fit the regression model for Y using $E\{x \mid X, \mathbf{w}\}$ using standard methods. The standard errors of the parameters in the regression model for Y must be adjusted to account for the fact that x is estimated rather than observed.

Estimation of the regression of x on $(X, \mathbf{w})$ is straightforward when replicate observations are available. Suppose that there are m replicate measurements of x for each individual, and that $\bar{X}$ is their mean. In that case, the *best linear estimator* of x is $E\{x \mid \bar{X}, \mathbf{w}\}$, where

$$E\{x \mid \bar{X}, \mathbf{w}\} = \mu_x + \begin{pmatrix} \sigma_x^2 \\ \sum_{\mathbf{w}x} \end{pmatrix} \times \begin{bmatrix} \sigma_x^2 + \sigma_u^2/m & \sum_{x\mathbf{w}} \\ \sum_{x\mathbf{w}}' & \sum_{\mathbf{ww}} \end{bmatrix}^{-1} \begin{pmatrix} \bar{X} - \mu_X \\ \mathbf{w} - \boldsymbol{\mu}_w \end{pmatrix} \quad (23)$$

Moment estimators for all unknown parameters in the expression above can be obtained as long as replicate observations are available and an estimator for the unknown measurement error variance can be computed. If replicate observations are not available, it is possible to use an external estimate of the measurement error variance.

Regression calibration was proposed by Gleser (1990). Carroll (1995, Chap. 3) discusses regression calibration in nonlinear models.

An alternative approach called SIMEX (for simulation extrapolation) that is well suited for the additive measurement error case was proposed by Cook and Stefanski (1994).

5. *Nonlinear Models and Measurement Error*

Carroll (1995, Chap. 6) presents an in-depth discussion of the problem of estimating the parameters in a nonlinear model when one or more of the covariates in the model are measured with error. Most of their discussion assumes that the measurement error is additive and normal, but some results are presented for more general cases. Other references include Amemiya (1990), Gleser (1990), and Roeder et al. (1996).

Carroll (1995) presents two approaches for estimating the parameters in the nonlinear model subject to measurement error, denoted *conditional score* and *corrected score* methods, that are based on an *estimating equations* approach. In either case, the resulting estimators do not depend on the distribution of x.

The *conditional score* method is derived for the case of the canonical *generalized linear model* (McCullagh and Nelder 1989) (see *Linear Hypothesis*). Without loss of generality we discuss the case where both x and w are scalar valued. The conditional density of Y given (x, w) is given by

$$f(y \mid x, w, \Theta) = \exp\left\{ \frac{y\eta - D(\eta)}{\phi} + c(y, \phi) \right\} \quad (24)$$

(expression 6.4 in Carroll 1995), where $\eta = \beta_0 + \beta_1 x + \beta_2 w$ is called a *natural parameter*, $\Theta = (\beta_0, \beta_1, \beta_2, \phi)$ is the unknown parameter to be estimated, and where the first and second moments of the distribution of Y are proportional to the first and second derivatives of $D(\eta)$ with respect to η. The form of η, $D(\eta)$ ϕ and $c(y, \phi)$ depends on the model choice.

If x were observed, then Θ is estimated by solving the following equations

$$\sum_{i=1}^{n} \{Y_i - D^{(1)}(\eta_i)\} \begin{pmatrix} 1 \\ w_i \\ x_i \end{pmatrix} = 0$$

$$\sum_{i=1}^{n} \left[\left(\frac{n-1}{n} \right) \phi - \frac{(Y_i - D^{(1)}(\eta_i))^2}{D^{(2)}(\eta_i)} \right] = 0 \quad (25)$$

where $D^{(k)}(\eta)$ is the kth order partial derivative of $D(\eta)$ with respect to η.

We now assume that x is measured with normal additive measurement error with variance σ_u^2. A sufficient statistic Δ can be found for x (Stefanski and Carroll 1987) so that the distribution of Y conditional on (Δ, w) has the same form as the density given in (24). The form of the functions $D(\eta)$, η, ϕ and $c(y, \phi)$ must be modified accordingly (see, for example, Carroll et al. 1995, p. 127). Therefore, this suggests that an estimator for Θ when x is observed with measurement error can be obtained by solving a set of equations similar to Eqn. (25).

The approach described above can be used only in the case of canonical generalized linear models. An approach applicable to the broader class of generalized linear regression models is the *corrected score* method, also derived under an estimating equations framework. Suppose, first, that x is observable, and that an estimator of Θ is obtained by solving

$$\sum_{i=1}^{n} \xi(Y_i, x_i, w_i, \Theta) = 0 \quad (26)$$

where ξ is a likelihood score from the model when x is observed without measurement error. Suppose now that we can find a function $\xi^*(Y, X, w, \Theta, \sigma_u^2)$, such that

$$E\{\xi^*(Y, X, w, \boldsymbol{\Theta} \mid Y, x, w\} = \xi(Y, x, w, \boldsymbol{\Theta}) \quad (27)$$

for all (Y, x, w). An estimate for the unknown parameter Θ can then be obtained by solving Eqn. (26) where the corrected score ξ^* is used in place of ξ. Corrected score functions that satisfy Eqn. (27) do not always exist, and when they do exist they can be very difficult to find. Stefanski (1989) derived the corrected score functions for several models that are commonly used in practice.

6. A Brief Historical Overview

The problem of fitting a simple linear regression model in the presence of measurement error was first considered by Adcock (1877). Adcock proposed an estimator for the slope in the regression line that accounted for the measurement error in the covariate for the special case where $\sigma_a^2 = \sigma_e^2$. The extension to the p-variate case was presented by Pearson (1901).

It was not until the mid-1930s that the terms *errors in variables models* and *measurement error models* were coined, and a systematic study of these types of model was undertaken. The pace of research in this area picked up in the late 1940s, with papers by Lindley (1947), Neyman and Scott (1951), and Kiefel and Wolfowitz (1956). These authors addressed issues of identifiability and consistency in measurement error models, and proposed various approaches for parameter estimation in those models. Lindley (1947) presented maximum likelihood estimators for the slope in a simple linear regression model with additive measurement error for the functional model. He concluded that, when the two variance components are unknown, the method of maximum likelihood cannot be used for estimating β_1, as the resulting estimator must satisfy the relation $\hat{\beta}_1$, $\hat{\sigma}_u^2 = \hat{\sigma}_e^2$. Solari (1969) later showed that this unexpected result could be explained by noticing that the maximum likelihood solution was a saddle-point and not a maximum of the likelihood function. Lindley (1947) also showed that the generalized least squares estimator of the slope is the maximum likelihood estimator when the ratio of variances is assumed known.

Extensions of the simple linear regression model to the multiple regression model case, with one or more covariates measured with error were presented in the late 1950s, 1960s and 1970s. Excellent early in-depth reviews of the literature in measurement error models are given by Madansky (1959), and Morton (1981).

In the early 1990s, Schmidt and Rosner (1993) proposed Bayesian approaches for estimating the parameters in the regression model in the presence of measurement error. More recently, Robins and Rodnitzky (1995) proposed a semi-parametric method that is more robust than the maximum likelihood estimator of the slope in structural models when the distributional assumptions are not correct. Carroll et al. (1997), however, showed that the Robins and Rodnitzky estimator can be very inefficient relative to the parametric estimation when the distributional assumptions are correct. Roeder et al. (1996) and Carroll et al. (1999) have argued that the sensitivity of the maximum likelihood estimator to model specification can be greatly reduced by using flexible parametric models for the measurement error. In particular, the authors suggest that a normal mixture model be used as the flexible model. In the 1999 paper, Carroll et al. (1999) proceed from a Bayesian viewpoint, and use Markov chain Monte Carlo methods to obtain approximations to the marginal posterior distributions of interest.

The literature on measurement error models is vast, and we have cited a very small proportion of the published books or manuscripts. In addition to the review articles listed above, the text books by Fuller (1987) and Carroll (1995) are widely considered to be the best references for linear and nonlinear measurement error models, respectively. The recent manuscript by Stefanski (2000) cites several papers that have appeared in the literature since 1995.

Bibliography

Adcock R J 1877 Note on the method of least squares. *The Analyst* **4**: 183–4

Amemiya Y 1990 Two stage instrumental variable estimators for the nonlinear errors in variables model. *Journal of Econometrics* **44**: 311–32

Buzas J S, Stefanski L A 1996 Instrumental variable estimation in generalized linear measurement error models. *Journal of the American Statistical Association* **91**: 999–1006

Carroll R J, Stefanski L A 1994 Measurement error, instrumental variables, and corrections for attenuation with applications to metaanalyses. *Statistics in Medicine* **13**: 1265–82

Carroll R J, Freedman L S, Pee D 1997 Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error. *Biometrics* **53**: 1440–57

Carroll R J, Roeder K, Wasserman L 1999 Flexible parametric measurement error models. *Biometrics* **55**: 44–54

Carroll R J 1995 *Measurement Error in Nonlinear Models*. Chapman and Hall, London

Cook J, Stefanski L A 1994 Simulation extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* **89**: 1314–28

Freedman L S, Carroll R J, Wax Y 1991 Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology* **134**: 510–20

Fuller W A 1987 *Measurement Error Models*. Wiley, New York

Gleser L J 1985 A note on Dolby G. R. unreplicated ultrastructural model. *Biometrika* **72**: 117–24

Gleser L J 1990 Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In: Brown P J, Fuller W A (eds.) *Statistical Analysis of Measurement Error Models and Applications*. Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference. American Mathematical Society, Rhode Island, RI, pp. 99–114

Gleser L J, Carroll R J, Gallo P P 1987 The limiting distribution of least squares in an errors-in-variable regression model. *Annals of Statistics* **15**: 220–33

Jöreskog K G 1981 Structural analysis of covariance structures. *Scandinavian Journal of Statistics* **8**: 65–92

Kiefer J, Wolfowitz J 1956 Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27**: 887–906
Lindley D V 1947 Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society*, Supplement **9**: 218–44
Madansky A 1959 The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association* **54**: 173–205
McCullagh P, Nelder J A 1989 *Generalized Linear Models*, 2nd edn. Chapman and Hall, London
Morton R 1981 Efficiency of estimating equations and the use of pivots. *Biometrika* **68**: 227–33
Neyman J, Scott E L 1951 On certain methods of estimating the linear structural relations. *Annals of Mathematical Statistics* **22**: 352–61
Pearson K 1901 On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**: 559–72
Robins J M, Rodnitzky A 1995 Semi-parametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**: 122–9
Roeder K, Carroll R J, Lindsay B G 1996 A semi-parametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* **91**: 722–32
Rosner B, Spiegelman D, Willett W C 1990 Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology* **132**: 734–45
Schmidt C H, Rosner B 1993 A Bayesian approach to logistic regression models having measurement error following a mixture distribution. *Statistics in Medicine* **12**: 1141–53
Solari M E 1969 The 'maximum likelihood solution' to the problem of estimating a linear functional relationship. *Journal of the Royal Statistical Society B* **31**: 372–5
Stefanski L A 1989 Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics, Series A* **18**: 4335–58
Stefanski L A 2000 Measurement error models. *Journal of the American Statistical Association* **95**: 1353–7
Stefanski L A, Carroll R J 1987 Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika* **74**: 703–16

A. Carriquiry

Measurement, Representational Theory of

The representational theory of measurement—RTM for short—is a sort of 'interface' between a qualitative, empirical, or material domain of research and the use of numbers to quantify the empirical observations and to draw inferences from the data. Roughly, the main tenets of RTM consist of the following steps (for definition of the main concepts see *Ordered Relational Structures*):

(a) Describe the situation under scrutiny by a qualitative relational structure, consisting of a base set A and a finite number of relations. Among these relations is usually an order relation and sometimes special elements such as units. These entities are the primitives of the structure. They must have an empirical identification. The properties of this qualitative relational structure are stated as conditions or axioms on its primitives, which are true statements about the structure, at least in the intended empirical identification.

(b) Verify by empirical procedures (experiments, observations) the truth of those conditions that are testable (which should be almost all of them) and find arguments (plausibility, consequences of failure) for those that are not testable.

(c) Find a structure with a numerical base set (a subset of $\mathbb{R}$ or $\mathbb{R}^n$) and find homomorphism between both structures. Finding a homomorphism may consist of providing a representation theorem which, essentially, states the existence of at least one homomorphism.

(d) Find all homomorphisms between the qualitative and the numerical structure. This step amounts to proving a uniqueness theorem, which in most cases yields a procedure to calculate any homomorphism from a given one.

(e) Clarify on the basis of the uniqueness properties of the set of homomorphisms which numerical statements about the empirical domain are meaningful and which are not. Meaningful statements can be used for further analysis, such as statistical tests.

This theory is thoroughly described in the three volumes *Foundations of Measurement* (see Krantz et al. 1971, Suppes et al. 1989, Luce et al. 1990. Other comprehensive expositions, differing slightly in emphasis and philosophical argumentation, are Pfanzagl (1971), Roberts (1979), and Narens (1985)

1. *History and Contemporary Formulation of Measurement Within RTM*

The above kind of formalization of measurement goes back to Helmholtz (1887). Hölder (1901) was the first to prove a representation theorem in the above sense. One version or other of Hölder's theorem lies at the heart of almost all representation theorems proved in the following years up to the present time. It did not, however, stimulate much research until the late 1950s. The 1930s and 1940s saw an ongoing debate about the measurability of sensory variables, which seemed to some a prerequisite to laying a firm foundation for psychology, not unlike that of most of the natural sciences. In this debate, the introduction of scale types by Stevens (1946) was a cornerstone. The search for a precise formulation led to the model-theoretic formulation of Scott and Suppes (1958). This framework was extended in the three volumes *Foundations of*

Measurement (Krantz et al. 1971, Suppes et al. 1989, Luce et al. 1990) and in the other books cited above.

In its present form, RTM consists of two relational structures $\mathscr{A}$ and $\mathscr{N}$. The base set of $\mathscr{N}$ is some subset of the reals or a real vector space. The objects which are empirically identified with the elements of A, the base set of $\mathscr{A}$, can be measured if it is possible to give or construct a homomorphism φ between the two structures. In most cases the structure $\mathscr{A}$ contains among its relations a weak order $\lesssim$. Some authors distinguish functions and constants such as a unit element from the proper relations. In our notation, these variants are all subsumed in $(R_i, i \in I)$. This sequence of relations can even be infinite. Sometimes it is necessary to endow the structures with several base sets and relations among them. This case can also be formulated within the present scheme, although it looks a bit clumsy.

The following are three of the major issues that confront RTM:

Representation. Does there exist a homomorphism between two structures, the qualitative and the numerical one, i.e., is there a mapping which preserves structure? This question is concerned with the measurability in principle of the variables under scrutiny.

Uniqueness. How can one describe the set of all representations of a given structure? This topic includes questions about the precise definition of scale type.

Meaningfulness. Which assertions about a measurement structure make sense, which do not? What is the meaning of numerical operations performed on the scale values? These questions are relevant to the applicability of statistical tests and to the content of descriptive indices.

The representation question is answered in RTM by formulating representation theorems; the next section gives two prominent examples. The uniqueness problem was initially dealt with by describing the set of all homomorphisms, this kind of result is usually part of the respective representation theorems (see Sect. 2.1). However, the modern theory differs in several aspects from this approach (see Sects. 3 and 4). The discussion of meaningfulness seems to have gone way beyond RTM. Several notions of invariance and meaningfulness have been proposed and analyzed, partly under the headline of dimensional analysis. We give but a short overview in Sect. 5.

2. Representation Theorems

The procedure outlined in the introductory paragraph and the formal definition is exemplified in this section by presenting a few representation and uniqueness theorems and by discussing examples. Basically, there are a few general properties that can be exploited to produce relational structures which can be successfully mapped into a numerical structure. Primarily these are ordered structures. The empirical identification of the elements of the base set and the order relation is in most cases rather straightforward. The other relations of the structures are operations, or are derived from the comparison of intervals, and the property of conjointness (see *Measurement Theory: Conjoint*). In the present article we focus on extensive structures and difference structures.

2.1 Extensive Measurement

Our first structure was also historically the first to be considered. It goes back to Helmholtz (1887). It postulates a qualitative ordering and a concatenation operation, i.e., a binary operation $\circ$: $A \times A \to A$ which resembles the operation of addition of real numbers.

Definition 1. Let A be a nonempty set, $\lesssim$ a binary relation on A, and o a closed operation on A, i.e., for all elements $a, b \in A$ there exists an element $c \in A$ such that $a \circ b = c$. The triple $(A, \lesssim, \circ)$ is a closed extensive structure if the following conditions (E1), (E2), (E3), and (E4) are satisfied for all $a, b, c, d \in A$; if in addition (E5) is satisfied it is a positive structure.

(E1) $\lesssim$ is a weak order on A, i.e., it satisfies $a \lesssim b$ and $b \lesssim c$ implies $a \lesssim c$, and $a \lesssim b$ or $b \lesssim a$.

(E2) $a \circ (b \circ c) \sim (a \circ b) \circ c$.

(E3) $a \lesssim b$ iff $a \circ c \lesssim b \circ c$ iff $c \circ a \lesssim c \circ b$.

(E4) If $a \prec b$, then for any c, d there exists a positive integer n such that $na \circ c \lesssim nb \circ d$, where na is defined inductively by $1a = a$ and $(n+1)a = na \circ a$.

(E5) $a \prec a \circ b$.

The example of weights in Sect. 2 of the entry *Ordered Relational Structures* can be regarded as such a structure if one is willing to accept certain idealizations. It formalizes the measurement of mass by comparing objects on a pan balance. Another obvious example is the measurement of length, also modulo several idealizations.

The following theorem not only guarantees the existence of a numerical representation (a homomorphism into the reals) but also states that this existence implies the conditions of Definition 1. It reads as follows:

Theorem 1. Let $(A, \lesssim, \circ)$ be a system with $\lesssim$ a binary relation and o a binary operation on A. Then $(A, \lesssim, \circ)$ is a closed extensive structure if and only if there exists a function φ: $A \to \mathbb{R}$ such that for all a, $b \in A$

(i) $a \lesssim b$ iff $\varphi(a) \leqslant \varphi(b)$;

(ii) $\varphi(a \circ b) = \varphi(a) + \varphi(b)$.

Another function φ': $A \to \mathbb{R}$ satisfies these two properties iff there is a constant $r \in \mathbb{R}$ such that $\varphi' = r \cdot \varphi$. The structure is positive iff $\varphi(a) > 0$.

The first part is the existence theorem, the second part the uniqueness theorem. The latter can be rephrased as: A closed concatenation structure allows the measurement of the objects on a ratio scale.

This result has been extended in several ways. One point of improvement consists of dropping the assumption of closedness of the operation. A closed operation forces the possibility to concatenate arbitrary large and arbitrary small elements, an assumption that might be too strong in practice. Krantz et al. (1971) give sufficient conditions when the operation is not necessarily closed. Other modifications are concerned with minimal and maximal elements in the structure, such as the velocity of light in relativistic velocity concatenation. Recently, this problem has been generalized by regarding 'singular points' in the structure, i.e., elements that behave differently with respect to o than do the rest of the elements.

2.2 *Difference Measurement*

While extensive operations abound in physics, they are rarely encountered in social sciences. However, the comparison of intervals is often used in psychological experiments. For instance, one can ask subjects whether the political system of two nations a, b is more similar to each other than that of the nations c, d. In economics the probability of choosing one alternative over another is used as an index of the size of a utility interval. In psychophysics, the method of bisection requires an observer to choose a stimulus whose value on some psychological scale such as loudness or brightness is between two given stimuli, such that the ensuing sensory differences are equal. This procedure assumes that people are able to compare intervals. Further examples are described in Chap. 4 of Krantz et al. (1971).

The following definition assumes a weak order $\lesssim$ on a subset A^* of pairs. For simplicity, the reader might assume that exactly one of the pairs ab and ba is in A^* for all $a, b \in A$. The definition and theorem are more general, but this is the most important case.

Definition 2. Let A be a a nonempty set, A^* a nonempty subset of the Cartesian product $A \times A$, and $\lesssim$ a binary relation on A^*. The triple $(A, A^*, \lesssim)$ is a positive difference structure if for all $a, b, c, d, a', b', c', a_1, a_2, \ldots \in A$.

(D1) $\lesssim$ is a weak order on A^*.

(D2) If $ab, bc \in A^*$, then $ac \in A^*$.

(D3) If $ab, bc \in A^*$, then $ab, bc \prec ac$.

(D4) If ab, bc, $a'b'$, $b'c' \in A^*$, $ab \lesssim a'b'$, and $bc \lesssim b'c'$, then $ac \lesssim a'c'$.

(D5) If $ab, cd \in A^*$ and $ab \succ cd$ then there exist d', $d'' \in A$ such that ad', $d'b$, ad'', $d''b \in A^*$ and $ad' \sim cd \sim d''b$.

(D6) If a_1, $a_2, \ldots$ is a strictly bounded standard sequence, i.e., $a_i a_{i+1} \in A^*$ and $a_i a_{i+1} \sim a_1 a_2$ for $i = 1, 2, \ldots$ and for some $d'd'' \in A^*$ holds $a_1 a_i \prec d'd''$, then it is finite.

Theorem 2. Let $(A, A^*, \lesssim)$ be a positive difference structure. Then there exists a function ψ: $A^* \to \mathbb{R}^+$ such that for all $a, b, c, d \in A$

(a) if $ab, cd \in A^*$, then $ab \lesssim cd$ iff $\psi(ab) \leqslant \psi(cd)$;

(b) if $ab, bc \in A^*$, then $\psi(ac) = \psi(ab) + \psi(bc)$

If ψ' is another function with these properties, then there is some $r > 0$ such that $\psi' = r \cdot \psi$.

If for all $a \neq b$, $a, b \in A$ either ab or ba in A, then there exists φ: $A \to \mathbb{R}$ such that for all $ab \in A^*$, $\psi(ab) = \varphi(b) - \varphi(a)$. If φ' has the same property, then for some $s \in \mathbb{R}$, $\varphi' = \varphi + s$.

The uniqueness assertion of this theorem roughly says that the intervals, i.e., the elements of A^*, are measured on a ratio scale, while the elements of A are measured on an interval scale.

Other difference structures, called algebraic-difference structures, have been developed, allowing both for positive and negative intervals (Krantz et al. 1971). Even the case of two simultaneous orderings of intervals compatible in a well-defined sense has been considered. The representation theorems guarantee the measurement of one ordering by differences of the representing function, while the other is represented by their ratios. A modification of this theorem was used by Suck (1998) for a measurement theoretic characterization of the exponential distribution.

2.3 *Comments on the Axioms*

Some remarks are in order regarding the nature of the different axioms. Some describe particular properties of the structures, such as the order axioms, or associativity (E2) in Definition 1, or (D4) in Definition 2. They are called *structural* and are all testable for given elements of A. Another class of axioms are the solvability conditions such as (D5). They postulate the existence of certain elements, given a few others, such that a given 'equation' is solved. Often a consequence of this kind of axiom is that the structure becomes infinite. Characteristically, a solvability axiom is often a part of conditions that are sufficient for numerical representations; however, it is not a necessary condition. Testing it is a delicate matter if one does not find elements d', d'' in (D5), then their non-existence is in general not proved.

A third class of conditions are Archimedean axioms, with Definition 1 (E4) and Definition 2 (D6) being examples. They are necessary for a representation with real numbers, because $\mathbb{R}$ satisfies a similar property (for all $x > y > 0$ there is an integer n such that $ny > x$). In ordered relational structures, the corresponding formulation employs the notion of standard sequences. These are increasing or decreasing sequences of elements of A, which are equally spaced in some sense specific for the structure. Examples are the sequences in (E4) and (D6).

Standard sequences are also very important for the construction of the homomorphisms in the represen-

tation theorems. This fact becomes clear if one studies the proofs of results such as Theorems 1 and 2. The key idea is as follows: a first standard sequence a_1, a_2,... is chosen. An element $a \in A$ is enclosed in $a_k \lesssim a \prec a_{k+1}$. The value k is taken as a first approximation to $\varphi(a)$. Then a second standard sequence b_1, b_2,... is constructed with each $b_{2i} = a_i$, i.e., a new element is introduced between two consecutive elements of the first sequence. If now $b_i \lesssim a \prec b_{i+1}$ then $i/2$ is the second approximation to the desired $\varphi(a)$. Proceeding in this way, a sequence of approximations is constructed. If this sequence can be shown to converge, then one has a value $\varphi(a)$. This φ can in many cases be shown to be one homomorphism. Other homomorphisms drop out by beginning with a different standard sequence. The uniqueness statements are derived from observing how these sequences are interrelated. This kind of proof mimics in some sense the actual procedures used in physics to measure objects. It is constructive and, depending on how many approximating steps are employed, controls the accuracy of the measurement.

3. *Scale Type*

Stevens (1946) introduced the concept of 'scale type' in response to a vivid debate on the possibility of evaluating subjective sensory events quantitatively. Much of the development of the present form of RTM in the 1950s and 1960s was influenced by the endeavor to give a sound foundation to Stevens' classical scale types, namely nominal, ordinal, interval, and ratio scales.

Roughly, the idea is to characterize scale type by describing the set of all homomorphisms of a representation in a uniqueness theorem. Such a description can be performed by giving the set of admissible transformations of a given homomorphism. Thus, if any strictly increasing transformation of a representation is admissible, then we have an ordinal scale if transformations of the form $x \to rx + s$ with $r > 0$ are admissible, we have an interval scale, and with $x \to rx$ we have a ratio scale. If necessary, one can define a nominal scale as one with all injective mappings (see *Ordered Relational Structures*) as admissible transformations.

Over time this simple theory needed several modifications. Log-interval scales and difference scales were introduced, mainly because the choice of the numerical structure was recognized to be highly arbitrary. Furthermore, not all scales could be given a clear place in this scheme. Roberts and Franke (1976) found that the scale type of semiorders (see *Partial Orders*) was 'irregular' compared with the previously considered types. This discussion was soon linked with invariance considerations and the meaningfulness problem. The scale type question was finally resolved by the work of Louis Narens (e.g., Narens 1981a, b), which led to the consideration of the automorphism group and the degree of homogeneity and uniqueness of an ordered structure (see *Ordered Relational Structures*). The next section discusses this theory.

4. *Modern Theory: Automorphisms of Measurement Structures*

The article on *Ordered Relational Structures* presents the homogeneity-uniqueness definition of scale type. Using automorphisms instead of homomorphisms and isomorphisms in this context frees the scale type and meaningfulness discussion from the arbitrariness of the choice of the representing structure. Furthermore, instead of applying an admissible transformation to a given homomorphism, one can apply an automorphism to the structure and then use the homomorphism. This way the possible representations drop out more naturally.

The first major success of this approach, which has been championed primarily by Narens, was the successful characterization of scale types (M, N) on the reals in the Alper/Narens Theorem. The type (1, 1) corresponds to a ratio scale, (2, 2) is an interval scale, (1, 2) is something in between, and (∞, ∞) is an ordinal scale. Other scale types do not exist if the homomorphism is onto a real interval. This deep result explains why science does not employ scales of a type intermediate between the interval and ordinal type. However, if the range of the homomorphism is the rational numbers, Cameron (1989) and Macpherson (1996) showed that all types (M, N) with $1 \leqslant M \leqslant N$ are possible.

The main problem is now to classify structures by scale type. In the investigation of this question a subset of the automorphism group, namely the translations, plays a crucial role. An automorphism $\tau \in \mathrm{Aut}\mathscr{A}$ is a *translation* if $\tau(a) \neq a$ for all $a \in A$ or if it is the identity map id_A. Luce and Narens (1994) formulate three major questions about translations which arose from the analysis of the proof of the Alper/Narens theorem:

(a) Do the translations form a group?

(b) Can the translations be ordered such that they satisfy an Archimedean property?

(c) Are the translations 1-homogeneous?

If all these questions are answered affirmatively, a numerical representation can be constructed (see Luce and Narens 1994).

Thus, a major problem is to find structural conditions that imply these three properties of the translations. But this seems to be a difficult problem, and not very much is known about it (see Luce in press). However, this shift of focus in the scale type discussion has already proved fruitful and has considerably increased the understanding of the philosophical problems of measurement. Possibly, final answers require a simultaneous solution of the meaningfulness problem which we discuss next.

5. Meaningfulness

In this section a short overview of the various aspects of the meaningfulness problem is given.

The 'meaning' of words or statements used to be a much debated issue in analytical philosophy in the first half of the twentieth century, as witnessed, for example, in the work of Carnap and others. In the same vein, and probably influenced by the philosophical discussion, the meaningfulness problem has been of interest from the very beginning of RTM. It is just the meaning of statements involving numerical values which is being debated. To give an example, consider the two statements:

(a) The ratio between yesterday's maximum temperature and today's is 2.

(b) The ratio between the difference of yesterday's maximum and minimum temperature and today's difference of the same values is 2.

Clearly, Statement (a) depends on the employed temperature scale. If it is correct using a Fahrenheit scale it is wrong in a Celsius scale, and vice versa (except for a few particular values). The second statement remains true or false in each temperature scale which is used nowadays. As a result, (a) seems meaningless, while (b) has a chance to be meaningful.

In its first formulation, the meaningfulness problem was closely linked to the uniqueness of scales and to scale type. Roughly, one seemed to settle with the formulation: a statement involving numerical scale values is meaningful if its truth value (i.e., true or false) does not change when an admissible transformation is applied to any of the scale values occurring in its formulation.

This approach falls short for several reasons, among which is the detection of irregular scales by Roberts and Franke (1976). More advanced theories of meaningfulness distinguish several concepts of invariance, such as reference and structure invariance. By the former it is understood that a relation is invariant under permissible transformations, while by the latter it is invariant under endomorphisms, i.e., homomorphisms of the structure into itself. These invariance properties were shown to provide necessary conditions on numerical relations for empirical meaningfulness. By contrast, the definition of a relation using the primitives of the structure may influence its meaningfulness. This is an intricate question connecting the meaningfulness problem to questions of definability.

Another aspect is dimensional analysis, which is a powerful tool in physics. The idea of this theory is to consider an equation involving the values of several different measurement homomorphisms, such as $f(x_1,\ldots, x_k) = 0$. If admissible transformations are applied to the scale values, the postulation of invariance narrows down the possibilities for f. in most cases, this argument leads to a functional equation which, when solved, gives the form of the physical law underlying the original equation. Luce (1959) applied this principle to psychophysical laws. Roberts and Rosenbaum (1986) undertook a systematic investigation of this technique, and Narens and Mausfeld (1992) presented a theory on the relationship of the psychological and the physical in psychophysics resulting in what they call an 'equivalence principal.' This method turns out to be a tool to sort out quantitative formulae and statements that have psychological significance. The role of dimensional analysis to find or discuss 'lawful' expressions is elaborated at a more advanced level in Luce (1978) and Chap. 22.7 of Luce et al. (1990).

One of the issues of the meaningfulness discussion was the question of meaningful statistics. In it an answer is sought to the question of which statistical procedures are applicable to which kind of data. For example, one can look at the test statistic of a t-test. It consists of a ratio of a difference of mean values and a standard deviation (or a 'mixed' standard deviation calculated from two samples). It is easy to see that this value is invariant under affine transformations of the data, but may vary considerably if the data are transformed monotonically. Does this fact imply that the t-test is applicable to interval scales and not to ordinal scales? Opinions on this question seem to differ widely. Luce et al. (1990) give an overview of what RTM can contribute to this question. Their point is that it is in general not the employed statistical value (e.g., the mean value) itself which might be inappropriate, but rather the numerical relations involving this value. Therefore, a difference has to be made between the meaningfulness of hypotheses on populations and the meaningfulness of statistics used for the description of samples.

A completely new way of looking at the meaningfulness issue was recently suggested by Narens (in press). He seems to turn the tables, in the sense that RTM is justified in terms of a new formulation of meaningfulness. In this formulation one still captures the domain of interest in a qualitative relational structure, but instead of deriving meaningfulness from invariance properties of this structure, one assumes a theory of meaningfulness in terms of which the concept of a meaningful scale can be formulated. In his paper it is shown that for each (in this sense) meaningful scale there exists a mathematical representing structure such that the scale consists of homomorphisms between these structures. Thus, the meaningful problem is solved by giving it a more fundamental place in the scientific process. Future discussion and elaboration of this point of view will decide on its viability.

6. Problems of RTM

Philosophical debate regarding the foundations of RTM continues—see, for instance, critiques by Michell (1986), Niederée (1987, 1992, 1994), and

Dzhafarov (1995). The main points of criticism are analyzed and discussed in Luce and Narens (1994).

A major driving force of RTM has been the idea that applied scientists would test the axioms of representation theorems, and if the fit was satisfactory, they would construct a scale accordingly, using the uniqueness part for a justification of the operations performed in the course of the research on the topic. This program, if it was a program, seems not yet to have been implemented. Others would say it has failed. There are very few attempts to test axioms for this purpose. Luce's (2000) book on utility theory is an exception and may point a way of renewing this program.

A major obstacle to an application of RTM in the way described is the lack of a satisfactory error theory of measurement within RTM. To what extent must an axiom be violated to be sure it does not hold? To give an example: even for physical length measurement, one can observe many instances of violations of the testable axioms of Theorem 1 (see Sect. 2.1). This will necessarily happen if the investigated objects differ only very little in length, less than the accuracy of the measuring device.

There have been numerous attempts to extend the classical 'deterministic' RTM to 'nondeterministic' versions by adding a probabilistic structure (see *Measurement Theory: Probabilistic*). However, none has provided a practical device for a theory of measurement error within RTM.

Another critical issue is the almost ubiquitous use of infinite structures. More than that the most elegant and useful results, such as the Alper/Narens Theorem, are obtained for continuum structures, i.e., ordered structures which satisfy Dedekind completeness (all infima and suprema exist, see *Ordered Relational Structures*). It can be surmised that a powerful error theory calls for mappings on a continuum.

The idea of embedding structures in Dedekind complete ones using extension procedures such as Dedekind cuts works in some cases but is intricate in others (see *Ordered Relational Structures* and *Partial Orders*). In what sense finite data sets are approximating such structures is a problem of theoretical importance for which not very much is known as yet. Sommer and Suppes (1997) proposes a critical approach to the use of the continuum in this context.

There are more results needed on classification by scale type. In particular, structural equivalents of homogeneity and uniqueness of automorphisms are not yet known. By structural equivalents, we mean conditions formulated within the structure and not within the automorphism group. Likewise, it is unclear how properties of the translations can be formulated or derived within $\mathscr{A}$.

Finally, the meaningfulness question is one of the outstanding open problems of RTM. The recent approach of Narens (submitted) mentioned in Sect. 5 appears completely new from the perspective of RTM, and it might possess the quality of a Copernicanean revolution.

7. *Applications of RTM*

It has already been mentioned that axiom testing is only rarely performed. But there are such investigations, mostly in connection with conjoint measurement (see *Conjoint Analysis Applications*). Much of the research on color vision, spatial vision, and other psychophysical questions uses representations of measurement structures. The benefit of RTM in such investigations is often theoretical: it provides a firm basis for formulating hypotheses and performing tests. As an example we cite Heller (1997) on binocular space perception and the elaborate mathematical development ensuing from this paper (see Aczél et al. 1999).

An interesting application of RTM can be found in dimensional analysis and the derivation of numerical laws in physics and psychophysics. Chapter 10 of Krantz et al. (1971) is the paradigm for this kind of application. Since then many results have been published. We cite one important article: Narens and Mausfeld (1992).

Finally, we mention the application of measurement theory in decision making. Expected utility theory, its subjective counterparts, e.g., Prospect theory, rank dependent utility (Luce 2000), etc, all rely to a great extent on results from RTM. This tendency is kept to the present day and is likely to remain so for some time. Luce (2000) is an outstanding recent example. The papers Luce and Narens (1994) and Luce (1995) give an in-depth overview of the general discussion.

See also: Binocular Space Perception Models; Conjoint Analysis Applications; Mathematical Psychology; Measurement Theory: Conjoint; Order Statistics; Ordered Relational Structures; Statistical Analysis, Special Problems of: Transformations of Data; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Aczél J, Baros Z, Heller J, Ng C T 1999 Functional equations in binocular space perception. *Journal of Mathematical Psychology* **43**: 71–101

Cameron P J 1989 Groups of order-automorphisms of the rationals with prescribed scale type. *Journal of Mathematical Psychology* **33**: 163–71

Dzhafarov E N 1995 Empirical meaningfulness, measurement-dependent constants, and dimensional analysis. In: Luce R D, D'Zmura M, Hoffman D, Iverson G J, Romney A K (eds.) *Geometric Representations of Perceptual Phenomena: Papers in Honor of Tarow Indow on his 70th Birthday*. Erlbaum, Mahwah, NJ

Heller J 1997 On the psychophysics of binocular space perception. *Journal of Mathematical Psychology* **41**: 29–43
Helmholtz H 1887 *Zählen und Messen erkenntnistheoretisch betrachtet, Philosophische Aufsätze Edward Zeller gewidmet*. Leipzig. [Reprinted in *Gesammelte Abhandlungen*. 1895, Vol. 3, pp. 356–91. Englisch translation by Bryan C L 1930 *Counting and Measuring*. Van Nostrand, Princeton, NJ]
Hölder O 1901 Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* **53**: 1–64
Krantz D H, Luce R D, Suppes P, Tversky A 1971 *Foundations of Measurement*. Academic Press, New York, Vol. 1
Luce R D 1959 On the possible psychophysical laws. *Psychological Review* **66**: 81–95
Luce R D 1978 Dimensionally invariant numerical laws correspond to meaningful qualitative relations. *Philosophy of Science* **45**: 1–16
Luce R D 1995 Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology* **46**: 1–26
Luce R D 2000 *Utility of Gains and Losses: Measurement-theoretical and Experimental Approaches*. Erlbaum, Mahwah, NJ
Luce R D in press Conditions equivalent to unit representations of ordered relational structures. *Journal of Mathematical Psychology*
Luce R D, Narens L 1994 Fifteen problems concerning the representational theory of measurement. In: Humphreys P. (ed.) *Patrick Suppes: Scientific Philosopher*. Kluwer Dordrecht, The Netherlands, Vol. 2, pp. 219–49
Luce R D, Krantz D H, Suppes P, Tversky A 1990 *Foundations of Measurement*. Academic Press, New York, Vol. 3
Macpherson D 1996 Sharply multiply homogeneous permutation groups, and national scale types. *Forum Mathematicum* **8**: 501–7
Michell J 1986 Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin* **100**: 398–407
Narens L 1981a A general theory of ratio scalability with remarks about the measurement—theoretic concept of meaningfulness. *Theory and Decision* **13**: 1–70
Narens L 1981b On the scales of measurement. *Journal of Mathematical Psychology* **24**: 249–75
Narens L 1985 *Abstract Measurement Theory*. MIT Press, Cambridge, MA
Narens L in press Meaningfulness. *Journal of Mathematical Psychology*
Narens L, Mausfeld R 1992 On the relationship of the psychological and the physical in psychophysics. *Psychological Review* **99**: 467–79
Niederée R 1987 On the reference to real numbers in fundamental measurement: A model-theoretic approach. In: Roskam E E, Suck R (eds.) *Progress in Mathematical Psychology*. North-Holland, Amsterdam, Vol. 1, pp. 3–23
Niederée R 1992 What do numbers measure? A new approach to fundamental measurement. *Mathematical Social Sciences* **24**: 237–76
Niederée R 1994 There is more to measurement than just measurement: Measurement theory, symmetry, and substantive theorizing. *Journal of Mathematical Psychology* **38**: 527–94
Pfanzagl J 1971 *Theory of Measurement*. Physika, Würzburg, Germany
Roberts F S 1979 *Measurement Theory with Applications to Decision making, Utility, and the Social Sciences*. Addison-Wesley, Reading, MA
Roberts F S, Franke C H 1976 On the theory of uniqueness in measurement. *Journal of Mathematical Psychology* **14**: 211–8
Roberts F S, Rosenbaum Z 1986 Scale type, meaningfulness, and the possible psychophysical laws. *Mathematical Social Sciences* **12**: 77–95
Scott D, Suppes P 1958 Foundational aspects of theories of measurement. *Journal of Symbolic Logic* **23**: 113–28
Sommer R, Suppes P 1997 Dispensing with the continuum. *Journal of Mathematical Psychology* **41**: 3–10
Stevens S S 1946 On the theory of scales of measurement. *Science* **103**: 677–80
Suck R 1998 A qualitative characterization of the exponential distribution. *Journal of Mathematical Psychology* **42**: 418–31
Suppes P, Krantz D H, Luce R D, Tversky A 1989 *Foundations of Measurement: Geometrical, Threshold, and Probabilistic Representations*. Academic Press, New York, Vol. 2

R. Suck

Measurement Theory: Conjoint

The theory of conjoint measurement is a branch of measurement theory in which objects described by several attributes are assigned numerical values that preserve an ordering of the objects by a qualitative relation such as preference or likelihood. Numerical values of objects usually are presumed to equal the sum of values assigned separately to each attribute. A conjoint measurement theory consists of the structure assumed for objects and their attributes, assumptions about the qualitative relation among objects, a numerical representation that connects the qualitative ordering of objects to numerical values, and a description of all sets of numerical values that satisfy the representation.

1. Objects and Attributes

The objects are members x, y, … of a subset X of the Cartesian product $X_1 \times X_2 \times \ldots \times X_n$ of $n \geqslant 2$ other sets referred to as attributes, factors, or criteria. When x_i is the member of attribute X_i associated with object x, $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, … Any set of entities that share a common set of attributes could serve as X, including consumption bundles (each attribute is a good, x_i is the amount of attribute i in bundle x), investment portfolios (stocks, bonds), houses (location, style, size, cost), candidates for a position (education, experience, skills, quality of references), and a host of other possibilities.

A particular theory specifies the structure of the X_i and X. Structural aspects are the number of attributes, whether X equals or is a proper subset of $X_1 \times \ldots \times X_n$, and the nature of each X_i (finite, infinite but discrete, a continuum). One theory assumes that n is any integer greater than 1, X is an arbitrary subset of

$X_1 \times \ldots \times X_n$, and every X_i is a finite set. Another assumes that $n = 2$, $X = X_1 \times X_2$, and each X_i is a continuum. Additional restrictions on structure may be imposed by preference axioms.

2. Preferences and Axioms

Qualitative comparisons between objects in X are described by a binary relation $\succsim$ on X and induced relations $\succ$ and $\sim$ defined by $x \succ y$ if $x \succsim y$ and not $(y \succsim x)$, and $x \sim y$ if $x \succsim y$ and $y \succsim x$. When preference is primitive, $x \succsim y$ signifies that x is at least as preferable as y, $x \succ y$ means that x is strictly preferred to y, and $x \sim y$ denotes equally preferred objects. When objects are uncertain events, $x \succsim y$ if x is judged to be at least as likely to occur as y, $x \succ y$ if x is judged to be more probable than y, and $x \sim y$ if x and y are considered equally likely.

Conditions imposed on the behavior of $\succsim$ on X are referred to as axioms. Although some axioms describe intuitively rational or reasonable behaviors, they are empirically refutable assumptions (see *Utility and Subjective Probability: Empirical Studies*).

2.1 Ordering Axioms

It is often assumed that $\succsim$ is a weak order, that is, that $\succsim$ is complete (for all $x, y \in X$: $x \succsim y$ or $y \succsim x$) and transitive (for all $x, y, z \in X$: if $x \succsim y$ and $y \succsim z$ then $x \succsim z$). When $\succsim$ is a weak order, $\sim$ is an equivalence relation (reflexive: $x \sim x$; symmetric: $x \sim y \Rightarrow y \sim x$; transitive) that partitions X into equivalence classes, with $x \sim y$ if and only if x and y are in the same class. A less common but more realistic axiom presumes only that $\succ$ is asymmetric ($x \succ y \Rightarrow$ not $(y \succ x)$) and transitive, in which case $\sim$ is not necessarily transitive, as when $x \sim y$, $y \sim z$, and $x \succ z$. Still less common is the axiom which assumes that $\succsim$ is complete but not necessarily transitive.

2.2 Independence and Cancellation Conditions

Independence and cancellation conditions are needed for the additive utility representation in the next section. The most basic independence condition assumes that, for each i, if $x_i = z_i$ and $y_i = w_i$, and $x_j = y_j$ and $z_j = w_j$ for all $j \neq i$, then $x \succsim y \Leftrightarrow z \succsim w$. When $\succsim$ is a weak order on $X = X_1 \times \ldots \times X_n$, this implies that the relation $\succsim_i$ on X_i defined by

$$x_i \succsim_i y_i \text{ if } x \succsim y \text{ whenever } x_j = y_j \text{ for all } j \neq i \qquad (1)$$

is a weak order. Moreover, if $x_i \succsim_i y_i$ for all i, then transitivity implies $x \succsim y$, with $x \succ y$ if $x_i \succ_i y_i$ for at least one i.

A more general first-order cancellation axiom asserts that if $\{x_i, z_i\} = \{y_i, w_i\}$ for all i, then $x \succsim y \Leftrightarrow w \succsim z$. This extends to the general cancellation condition which says that if $m \geqslant 2$, if $x^1, \ldots, x^m$, $y^1, \ldots, y^m \in X$, and if $y_i^1, \ldots, y_i^m$ is a permutation of $x_i^1, \ldots, x_i^m$ for every i, then it is not true that $x^j \succsim y^j$ for $j = 1, \ldots, m$ with $x^j \succ y^j$ for some j. The second-order cancellation axiom is the special case of general cancellation for $m = 3$. An instance for $n = 2$ is : it is false that $(x_1, x_2) \succsim (y_1, z_2)$, $(y_1, y_2) \succsim (z_1, x_2)$, and $(z_1, z_2) \succ (x_1, y_2)$. Its $\sim$ version, whose conclusion is $\{x^1 \sim y^1,\ x^2 \sim y^2\} \Rightarrow x^3 \sim y^3$, illustrates a notion of higher-order tradeoffs and is sometimes referred to as the Thomsen condition.

2.3 Solvability and Archimean Axioms

Solvability axioms posit the existence of attribute values that allow compensating tradeoffs and thereby impose further restrictions on the X_i and X. An example for $n = 2$ is : for every $(x_1, x_2) \in X$ and every $y_1 \in X_1$, there is a $y_2 \in X_2$ for which $(y_1, y_2) \sim (x_1, x_2)$.

Archimedean axioms, sometimes referred to as continuity conditions, are technical conditions for infinite structures that ensure the existence of real-valued utilities in representations of $\succsim$. When each X_i is a nondegenerate real interval and $\succ$ is the lexicographic order defined by $(x_1, x_2) \succ (y_1, y_2)$ if $x_1 > y_1$ or $(x_1 = y_1, x_2 > y_2)$, there is no real-valued function u on X that satisfies $x \succ y \Leftrightarrow u(x_1, x_2) > u(y_1, y_2)$. Archimedean axioms forbid this type of noncompensatory dominance structure.

3. Additive Representations

The basic additive conjoint representation of $\succsim$ on X consists of a real-valued function u_i on X_i for $i = 1, \ldots, n$ such that, for all $x, y \in X$

$$x \succsim y \Leftrightarrow \sum_{i=1}^{n} u_i(x_i) \geqslant \sum_{i=1}^{n} u_i(y_i) \qquad (2)$$

An alternative is the multiplicative representation $x \succsim y \Leftrightarrow f_1(x_1) \ldots f_n(x_n) \geqslant f_1(y_1) \ldots f_n(y_n)$, where $f_i = e^{u_i}$. Other algebraic combinations of attribute utilities are discussed in Krantz et al. (1971) under polynomial measurement. Additional restrictions may be imposed on the u_i of Eqn. (2) for representations of qualitative probability.

3.1 Uniqueness

Let $v_i = au_i + b_i$ with $a > 0$ and $b_1, \ldots, b_n$ any real numbers. When Eqn. (2) holds, it holds also for $v_1, \ldots, v_n$ in place of $u_1, \ldots, u_n$. When no other transformations

of the u_i satisfy Eqn. (2), the attribute utilities for the additive representation are unique up to similar positive affine transformations. The common $a > 0$ for the transformations preserves scale commensurability of the attribute utility functions. When X is finite, more general transformations are possible unless enough $\sim$ comparisons hold to force uniqueness up to similar positive affine transformations (Fishburn and Roberts 1989).

3.2 Specific Theories

When X is finite, Eqn. (2) reduces to a finite number of linear inequalities with coefficients in $\{1, 0, -1\}$ which have a solution if and only if $\succsim$ satisfies weak order and general cancellation (Kraft et al. 1959, Fishburn 1970, Krantz et al. 1971).

When the X_i are infinite, $X = X_1 \times \ldots \times X_n$, and $n \geqslant 3$, weak order, first-order cancellation, and an Archimedean axiom lead to u_i for Eqn. (2) that are unique up to similar positive affine transformations. A topological approach (Debreu 1960, Fishburn 1970, Wakker 1989) also endows X with special topological structure, and an algebraic approach (Fishburn 1970, Krantz et al. 1971) uses a solvability condition. Similar approaches apply for $n = 2$, where a second-order cancellation condition is also required.

Another weak-order theory considers a denumerable number of attributes that refer to successive time periods and involve notions of discounting and impatience (Fishburn 1970). The relaxation of weak order where $\succ$ is asymmetric and transitive leads to the one-way representation

$$x \succ y \Rightarrow \sum_{i=1}^{n} u_i(x_i) > \sum_{i=1}^{n} u_i(y_i) \tag{3}$$

for partially ordered preferences (Fishburn 1970). Transitivity can be omitted without affecting additivity as shown by the nontransitive additive conjoint representation

$$x \succsim y \Leftrightarrow \sum_{i=1}^{n} \phi_i(x_i, y_i) \geqslant 0 \tag{4}$$

in which ϕ_i is a skew-symmetric ($\phi_i(a, b) + \phi_i(b, a) = 0$) real-valued function on $X_i \times X_i$ (Fishburn 1991).

4. Conjoint Measurement and Social Choice

Ideas from conjoint measurement have been applied to social choice theory (Fishburn 1973) under the following interpretations. Each i indexes one of n voters, and $\succsim_i$ denotes the preference relation of voter i. Unlike Eqn. (1), the $\succsim_i$ are primitive relations not derived from $\succsim$, and $\succsim$ takes the role of a social preference relation which is a function of the voter preference profile ($\succsim_1, \ldots, \succsim_n$). Conditions of order, fairness, positive responsiveness, and so forth, can then be proposed to relate $\succsim$ to realizations of ($\succsim_1, \ldots, \succsim_n$). Some sets of conditions lead to additive representations for simple majority, weighted majority, and more general rules, whereas others are collectively incompatible (Arrow 1963) and result in impossibility theorems.

5. Decision Under Risk and Uncertainty

Conjoint measurement is used extensively for multiattribute outcomes in decision under risk and uncertainty (Keeney and Raiffa 1976). Consider the expected-utility representation

$$p \succsim q \Leftrightarrow \sum_{x \in X} p(x)\, u(x) \geqslant \sum_{x \in X} q(x)\, u(x) \tag{5}$$

where p and q are simple probability distributions on $X = X_1 \times \ldots \times X_n$ and u on X is unique up to a positive affine transformation $au + b$, $a > 0$. Let p_i be the marginal distribution on X_i of p. The independence condition which says that $p \sim q$ whenever $p_i = q_i$ for all i implies an additive decomposition $u(x_1, \ldots, x_n) = u_1(x_1) + \ldots + u_n(x_n)$ for outcome utilities (Fishburn 1970, p. 149). A less-restrictive condition says that, for every i, the preference order over marginal distributions on X_i at fixed values of the other attributes is independent of those fixed values. This implies either the additive decomposition or a multilinear form which under somewhat stronger conditions can be expressed as

$$ku(x_1, \ldots, x_n) + 1 = [ku_1(x_1) + 1] \ldots [ku_n(x_n) + 1] \tag{6}$$

where k is a nonzero constant. A number of other decompositions of u are also described in Keeney and Raiffa (1976).

See also: Conjoint Analysis Applications; Measurement Theory: History and Philosophy; Utility and Subjective Probability: Contemporary Theories

Bibliography

Arrow K J 1963 *Social Choice and Individual Values*, 2nd edn. Wiley, New York

Debreu G 1960 Topological methods in cardinal utility theory. In: Arrow K J, Karlin S, Suppes P (eds.) *Mathematical Methods in the Social Sciences, 1959*. Stanford University Press, Stanford, CA

Fishburn P C 1970 *Utility Theory for Decision Making*. Wiley, New York

Fishburn P C 1973 *The Theory of Social Choice*. Princeton University Press, Princeton, NJ

Fishburn P C 1991 Nontransitive additive conjoint measurement. *Journal of Mathematical Psychology* **35**: 1–40

Fishburn P C, Roberts F S 1989 Uniqueness in finite measurement. In: Roberts F S (ed.) *Applications of Combinatorics and Graph Theory in the Biological and Social Sciences*. Springer-Verlag, Berlin

Keeney R L, Raiffa H 1976 *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York
Kraft C H, Pratt J W, Seidenberg A 1959 Intuitive probability on finite sets. *Annals of Mathematical Statistics* **30**: 408–19
Krantz D H, Luce R D, Suppes P, Tversky A 1971 *Foundations of Measurement*. Academic Press, New York, Vol. 1
Wakker P P 1989 *Additive Representations of Preferences*. Kluwer, Dordrecht, Holland

P. Fishburn

Measurement Theory: History and Philosophy

Measurement theorists attempt to understand quantification and its place in science. The success of quantification in the physical sciences suggested that measurement of psychological attributes might also be possible. Attempting this proved difficult and putative psychological measurement procedures remain controversial. In considering such attempts and the issue of whether they amount to measurement, it is useful to understand the way in which the concept of measurement has unfolded, adjusting to problems both internal and external to quantitative science. This is the history and philosophy of measurement theory.

1. *Measurement Theory and Problems Internal to Quantitative Science*

1.1 *Measurement of Extensive Magnitudes*

The first treatise touching on the theory of measurement was Book V of Euclid's *Elements* (Heath 1908), written about 300 BC and attributed to Eudoxus. It defined the central concepts of *magnitude* and *ratio* and explained the place of numbers in measurement. This explanation was accepted as standard for more than 2,000 years. The early Greeks divided quantity into multitude (or discrete quantity, e.g., the size of a football crowd) and magnitude (or continuous quantity, e.g., the length of a stadium). The measure of a quantity was given relative to numbers of appropriately defined units. Number, then, was what today is called 'whole number.' This posed a problem because it was known that magnitudes could be mutually incommensurable (two magnitudes are mutually incommensurable if and only if the ratio of one to the other cannot be expressed as a ratio of whole numbers, e.g., as is the case with the lengths of the side and diagonal of a square). The problem for Euclid was to explain how his discrete concept of measure applied to continuous magnitudes.

It was solved by liberalizing the concept of measure to that of ratio and by requiring that for each specific magnitude (say, each specific length) there is the series of multiples of that magnitude (e.g., for any specific length, L, there exists mL, for each whole number, m). Considering just a single pair of lengths, say L_1 and L_2, and letting n and m stand for any two whole numbers, the ratio between L_1 and L_2 is given by the following three classes of numerical ratios:

(a) The class of all n/m such that $mL_1 < nL_2$;
(b) The class of all n/m such that $mL_1 = nL_2$; and
(c) The class of all n/m such that $mL_1 > nL_2$.

Defined this way, a ratio exists between each magnitude and any arbitrary unit, whether incommensurable or not, and, in principle, this ratio may be estimated. This is the classical theory of measurement. Of course, unless there is some physical operation for obtaining multiples of magnitudes it is not possible to estimate these ratios. Hence, this elegant and powerful solution was not taken to apply beyond the range of extensive magnitudes (i.e., those for which some operation of addition can be defined). At the time this included only the geometric magnitudes, weight and time.

1.2 *Measurement of Intensive Magnitudes in Physics*

The early Greeks recognized that other attributes, for example, temperature, admitted of degrees and Aristotle even discussed these hypothetically in terms of ratios (*On Generation and Corruption* 334 b, 5–20). This raised the possibility of nonextensive magnitudes (i.e., magnitudes for which an operation of addition was not known), a possibility explored during the Middle Ages via the concept of intensive magnitude. For example, Oresme thought intensive magnitudes similar enough to line lengths to assert that 'whatever ratio is found to exist between intensity and intensity, … a similar ratio is found to exist between line and line and vice versa' (Clagget 1968, p. 167), thereby attributing to them an internal structure sufficient to sustain ratios.

The medieval philosophers, however, did not attempt to show that their intensive magnitudes really were, like length, additive in structure. While they contemplated intensive magnitudes as theoretical possibilities, they were not interested in actually measuring them. Attempts to expand the range of measurable attributes were more successful from the Scientific Revolution of the seventeenth century onwards, when science became more closely identified with quantitative methods. This success raised another problem: it was one thing to distinguish intensive from extensive magnitudes in theory; it was quite another to show that a particular procedure actually measured any of the former. Gradually, the issue dawned of what was to count as evidence that hitherto unmeasured attributes are measurable.

Most of the physical magnitudes found measurable after the Scientific Revolution were not extensive and

were measured indirectly. For example, velocity was measured via distance and time, density, via mass and volume. In each case, the magnitude measured in this derived way was thought to be some quantitative attribute of the object or system involved, but one for which a physical operation directly reflecting the supposed underlying additive structure of the attribute could not be (or, possibly, had not yet been) identified. Because quantitative physics, as a body of science, was held in high regard, other disciplines aped its quantitative character. For example, Hutcheson (1725) proposed a quantitative moral psychology. He was criticized by Reid (1748) for 'applying measures to things that properly have not quantity' (p. 717), a charge that was to become increasingly familiar as further attempts were made to integrate quantitative thinking into psychology. The problem is how to tell quantitative from nonquantitative attributes. Solving it requires specifying quantitative structure.

Hölder (1901) achieved this, specifying conditions defining the concept of an unbounded, continuous quantitative attribute. Call the attribute Q and let its different levels (the specific magnitudes of Q) be designated by $a, b, c, \ldots$. For any three levels, a, b, and c, of Q, let $a+b=c$ if and only if c is entirely composed of discrete parts a and b. Then according to Hölder, Q being quantitative means that the following seven conditions hold.

(a) Given any two magnitudes, a and b, of Q, one and only one of the following is true:

(i) a is identical to b (i.e., $a=b$ and $b=a$);

(ii) a is greater than b and b is less than a (i.e., $a>b$ and $b<a$); or

(iii) b is greater than a and a is less than b (i.e., $b>a$ and $a<b$).

(b) For every magnitude, a, of Q, there exists a b in Q such that $b<a$.

(c) For every pair of magnitudes, a and b, in Q, there exists a magnitude, c, in Q such that $a+b=c$.

(d) For every pair of magnitudes, a and b, in Q, $a+b>a$ and $a+b>b$.

(e) For every pair of magnitudes, a and b, in Q, if $a<b$, then there exists magnitudes, c and d, in Q such that $a+c=b$ and $d+a=b$.

(f) For every triple of magnitudes, a, b, and c, in Q, $(a+b)+c=a+(b+c)$.

(g) For every pair of classes, ϕ and ψ, of magnitudes of Q, such that

(i) each magnitude of Q belongs to one and only one of ϕ and ψ;

(ii) neither ϕ nor ψ is empty; and

(iii) every magnitude in ϕ is less than each magnitude in ψ, there exists a magnitude x in Q such that for every other magnitude, x', in Q, if $x'<x$, then $x'\in\phi$ and if $x'>x$, then $x'\in\psi$ (depending on the particular case, x may belong to either class).

Hölder had one eye on Euclid's concept of ratio and, so, what he was able to prove from these conditions was that if an attribute has this sort of structure, then the system of ratios of its magnitudes is isomorphic to the system of positive real numbers. Among other things, this entails that each level of the attribute may be measured by any other level taken as the unit.

Using Hölder's conditions, the issue of the sort of evidence needed to confirm the hypothesis that some attribute is quantitative may be considered. If, relative to the attribute in question, a concatenation operation upon objects can be identified which directly reflects quantitative additivity, then one has evidence that the attribute is quantitative. For example, rigid straight rods of various lengths may be joined lengthwise so that the resulting length equals the sum of the lengths of the rods concatenated. This issue was considered in a systematic way by Helmholtz (1887) and Campbell (1920) and Campbell's term, 'fundamental measurement,' has become the standard one to describe this way of establishing measurement.

Campbell (1920) also identified the category he called 'derived measurement.' He held that 'the constant in a numerical law is always the measure of a magnitude' (p. 346). For example, the ratio of mass to volume is a constant for each different kind of substance, say, one constant for gold, another for silver, and so on. That such system-dependent constants identify magnitudes of quantitative attributes is perfectly sound scientific thinking. In the case of density, the fact that this constant is always observed must have a cause within the structure of the system or object involved. Since the effect is quantitative, likewise must be the cause. Hence, the magnitude identified by the constant, or as we call it, the density of the substance, is quantitative and measured by this constant.

A similar line of thought suffices for all cases of indirect quantification in physics, and because Campbell (1920) thought that physics was 'the science of measurement' (p. 267), he concluded that the issue of the sort of evidence needed to show that an attribute is quantitative is exhausted by the categories of fundamental and derived measurement. Campbell did not know that Hölder (1901) had already refuted this conclusion using a simple model. If all that could be known of any two points on a line is their order and of any two intervals between such points, whether they are equal, then Hölder showed this knowledge sufficient to prove distance quantitative. His 10 axioms applying just to order between points and equality between intervals exploit the fact that these nonadditive relations may indirectly capture the underlying additivity between distances (viz., if $A<B<C$ (where A, B, and C are points), then the distance from A to C equals that from A to B plus that from B to C). Applied to other attributes, this result entails that there are ways other than those standard in physics for showing attributes as quantitative. Hence, the possibility of measurement in disciplines lacking fundamental measurement cannot be ruled out.

1.3 Measurement of Psychological Magnitudes

Attempts at measurement in psychology presented entirely new challenges for measurement theory. These attempts are hampered by the fact that there are no fundamentally measured, extensive psychological magnitudes upon which to base the measurement of theoretical, psychological attributes. Hence, in attempting to gain evidence that psychological attributes of interest are quantitative, Campbell's categories of fundamental and derived measurement cannot be utilized. To cope with this, a new approach to measurement had to be devised.

This challenge was met by the theory of conjoint measurement (see *Measurement Theory: Conjoint*). Unlike fundamental and derived measurement, conjoint measurement does not depend upon finding extensive magnitudes. This theory (Krantz *et al.* 1971) applies in circumstances where a dependent attribute is a noninteractive function of two independent attributes and where distinct levels of the independent attributes can at least be classified (as same or different) and the dependent attribute weakly ordered. Since, in psychology, such classifications and orderings are sometimes possible, this theory is applicable. If the attributes involved are quantitative, then this is detectable via trade-offs between the contributions of the conjoining independent attributes to the order upon levels of the dependent attribute. The theory has been applied to the measurement of psychological attributes (see *Conjoint Analysis Applications*) and has been extended to situations involving more than two independent attributes.

2. Measurement Theory and Problems External to Quantitative Science

2.1 The Philosophy of Mathematics

The most potent external factor affecting measurement theory has been the philosophy of mathematics, especially as it relates to the concept of number. Hölder's (1901) proof of an isomorphism between ratios of magnitudes of a continuous quantitative attribute and the positive real numbers appeared to secure the view, popular at least since the Scientific Revolution (and, arguably, implicit in Euclid) that numbers are such ratios (see, e.g., Frege 1903). However, at least since Russell (1897), some philosophers sought to divorce the concept of number from that of quantity. The view that came to dominate twentieth century philosophy was the thesis that numbers are 'abstract entities' and able to be conceptualized as elements of formal systems. Accordingly, they were thought not to be empirically located, as are quantities. As a result, measurement ceased to be defined in the classical manner, i.e., as the numerical estimation of ratios of magnitudes. Russell (1903) proposed that measurement be defined instead as the representation of magnitudes by numbers. This representational approach has dominated thinking about measurement this century (e.g., Campbell 1920, Nagel 1931, Ellis 1966), being developed most comprehensively by Suppes, Luce, Krantz, and Tversky (e.g., Krantz et al. 1971, Suppes et al. 1989, Luce et al. 1990) and their associates (see *Measurement, Representational Theory of*).

The application of this approach to the measurement of any attribute reduces to four steps. First, an empirical system is specified as an empirically identifiable set of some kind (e.g., of objects or attributes) together with a finite number of empirical relations between its elements. Second, a set of axioms is stated for this empirical system. As far as possible these should be empirically testable. Third, a numerical structure is identified such that a set of many-to-one mappings (homomorphisms) between the empirical system and this numerical structure can be proved from the axioms (the 'representation theorem'). Fourth, inter-relations between the elements of this set of homomorphisms are specified, generally by identifying the class of mathematical functions which includes all transformations of any one element of this set into the other elements (the 'uniqueness theorem'). This allows the distinctions between types of scales, popularized by Stevens (1951), to be precisely defined. This approach to measurement theory has produced, among other important developments, the theory of conjoint measurement.

The numerical representation of empirical systems depends on identifying structures within those systems which possess numerical correlates. Insofar as mathematics may be thought of as the study of the general forms of structures (empirical or otherwise), proofs of representation theorems, therefore, rely upon identifying mathematical structure within empirical systems. Insofar as quantitative theories in science require something like Hölder's concept of continuous quantity as the empirical system to be numerically represented, it follows that the mathematical structure of the positive real numbers might be thought of as already present (as ratios of magnitudes) in that empirical system. It is along these lines that a reconciliation between classical and representational theory may be possible.

2.2 The Philosophy of Science

Developments in the philosophy of science have also affected measurement theory. For psychology, the most significant factor was operationism (Bridgman 1927). Operationism was developed into a theory of measurement by Stevens (1951). If, as Bridgman held, the meaning of a concept is the set of operations used

to specify it, then measurement is the 'assignment of numerals to objects or events according to rules' (Stevens 1951, p. 1) and the attribute measured via any such assignment is defined by the rules used. This approach proved useful in the social and behavioral sciences before it was known how to test the hypothesis that such attributes are quantitative.

Michell (1999) contains further material on the history and philosophy of measurement in psychology.

See also: Conjoint Analysis Applications; Measurement, Representational Theory of; Measurement Theory: Conjoint; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Aristotle 1941 On generation and corruption. In: McKeon R (ed.) *The Basic Works of Aristotle*. Random House, New York

Bridgman P W 1927 *The Logic of Modern Physics*. Macmillan, New York

Campbell N R 1920 *Physics, the Elements*. Cambridge University Press, Cambridge, UK

Clagget M (ed.) 1968 *Nicole Oresme and the Medieval Geometry of Qualities and Motions*. University of Wisconsin Press, Madison, WI

Ellis B B 1966 *Basic Concepts of Measurement*. Cambridge University Press, Cambridge, UK

Frege G 1903 *Grundgesetze der Arithmetik*. Hildesheim, Germany, Georg Olms

Heath T L 1908 *The Thirteen Books of Euclid's Elements*. Cambridge University Press, Cambridge, UK, Vol. 2

Helmholtz H von 1887 Zählen and Messen erkenntnistheortisch betrachtet. *Philosophische Aufsätze Eduard Zeller zu seinum fünfzigjährigen Doktorjubiläum gewindmet*. Fues' Verlag, Leipzig [1971 An Epistemological Analysis of Counting and Measurement. In: Kahl R (ed.) *Selected Writings of Hermann von Helmholtz*, 1st edn. Wesleyan University Press, Middletown, CT]

Hölder O 1901 Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* **53**: 1–46 [1996 The axioms of quantity and the theory of measurement, Part I. *Journal of Mathematical Psychology* **40**: 235–52; 1997, Part II. *Journal of Mathematical Psychology* **41**: 345–56]

Hutcheson F 1725 *An Inquiry into the Original of our Ideas of Beauty and Virtue*. Darby, London

Krantz D H, Luce R D, Suppes P, Tversky A 1971 *Foundations of Measurement*. Academic Press, New York, Vol. 1

Luce R D, Krantz D H, Suppes P, Tversky A 1990 *Foundations of Measurement*. Academic Press, San Diego, CA, Vol. 3

Michell J 1999 *Measurement in Psychology: A Critical History of a Methodological Concept*. Cambridge University Press, New York

Nagel E 1931 Measurement. *Erkenntnis* **2**: 313–33

Reid T 1748 An essay on quantity. [1849 In: Hamilton W (ed.) *The Works of Thomas Reid*. Maclachlan, Stewart & Co, Edinburgh, UK]

Russell B 1897 On the relations of number and quantity. *Mind* **6**: 326–41

Russell B 1903 *Principles of Mathematics*. Cambridge University Press, New York

Stevens S S 1951 Mathematics, measurement and psychophysics. In: Stevens S S (ed.) *Handbook of Experimental Psychology*. Wiley, New York

Suppes P, Krantz D H, Luce R D, Tversky A 1989 *Foundations of Measurement*. Academic Press, New York, Vol. 2

J. Michell

Measurement Theory: Probabilistic

In the theory of probabilistic measurement concepts of measurement are developed in a probabilistic framework to provide a theoretically sound treatment of variability of measurements. Concepts of probabilistic measurement evolved in the context of the behavioral and social sciences, for instance in psychophysics and utility theory. They differ from the standard Gaussian theory of error as applied in physics, where at a theoretical level the underlying scales considered (e.g., length, mass) are conceived in a deterministic fashion. To deal with the variability of measurements ('error') that arise in practical applications, statistical concepts are then introduced in an *adhoc* manner on top of these deterministic concepts. In the Social and Behavioral Sciences, in contrast, variability (e.g., of psychophysical or preference judgments made by the subjects) usually is considered a feature of the domain considered. Therefore, concepts of probabilistic measurement aim at the theoretical (possibly axiomatic) foundation of the scales themselves in a probabilistic theoretical framework.

Roughly, two traditions can be distinguished: the psychometric and the measurement-theoretic approach. The following account focuses mainly on the measurement-theoretic perspective which strives for suitable probabilistic counterparts to concepts of the representational theory of measurement (see *Measurement Theory: History and Philosophy*; *Measurement, Representational Theory of*).

Several general strategies of probabilization will be discussed. For brevity's sake they will be explained using the examples of additive conjoint and algebraic difference measurement (see *Measurement Theory: Conjoint*; *Conjoint Analysis Applications*; Krantz et al. 1971).

1. Fechnerian and Thurstonian Scaling

This section briefly outlines two basic approaches to the scaling of subjective magnitudes which played an important part in the development of probabilistic measurement.

Consider an experimental paradigm where a subject is asked to compare two stimuli out of a stimulus set A according to some criterion (e.g., loudness of tones, brightness of flashes, subjective utilities of commodities). If in repeated comparisons of the same pair of stimuli the subject made the same judgments throughout, this would define a relation $\gtrsim$ on A, with $a \gtrsim b$ meaning that a is chosen over b. If $\gtrsim$ fulfilled the (testable) axioms of a weak order (see *Ordered Relational Structures*) this would give rise to the introduction of ordinal subjective scales $u: A \to \mathrm{Re}$ satisfying $u(a) \geqslant u(b)$ if and only if $a \gtrsim b$.

However, ordinarily the subject's judgments will vary between experimental trials, particularly for stimuli 'close' to each other. This variability can be accounted for by considering the corresponding system of choice probabilities $p(a, b)$, where $p(a, b)$ denotes the probability that a is chosen over b.

The weak-utility model assumes that the relation $\gtrsim$ defined by

$$a \gtrsim b \quad \text{if and only if } p(a, b) \geqslant 1/2$$

is a weak order giving rise to an ordinal scale u as described before. A necessary testable condition for this model is weak stochastic transitivity

$$\text{If} \quad p(a, b) \geqslant 1/2 \quad \text{and} \quad p(b, c) \geqslant 1/2$$
$$\text{then} \quad p(a, c) \geqslant 1/2.$$

The weak-utility model is implied by the strong-utility model, which postulates that $p(a, b)$ is a measure of the strength-of-preference or discriminability in the sense that for some scale u

$$p(a, b) \geqslant p(c, d)$$
$$\text{if and only if } u(a) - u(b) \geqslant u(c) - u(d).$$

The strong-utility model is essentially equivalent to the Fechnerian model which is based on the assumption that there is a strictly increasing function F: $\mathrm{Re} \to \mathrm{Re}$ and a scale u satisfying

$$p(a, b) = F[u(a) - u(b)].$$

(for relations to Fechnerian psychophysics see Falmagne 1986).

A different path was taken in the twenties by L. L. Thurstone. On his account, subjective magnitudes are not represented by single numbers but by real random variables. The basic assumption is that there is a family of jointly distributed random variables, U_a ($a \in A$) on a sampling space Ω that corresponds to different experimental situations $\omega \in \Omega$, such that in situation ω the stimulus a is chosen over b if and only if $U_a(\omega) \geqslant U_b(\omega)$. It is further assumed that for a suitable probability measure P

$$p(a, b) = P(\{\omega \in \Omega \mid U_a(\omega) \geqslant U_b(\omega)\}) =: \Pr[U_a \geqslant U_b]$$

Models of that kind are called random utility models. An example of a testable necessary condition for this model is the so-called triangle inequality,

$$p(a, b) + p(b, c) - p(a, c) \leqslant 1.$$

Under certain conditions this model implies the Fechnerian model. For instance, Thurstone assumed that the random variables U_a are normally distributed; under the additional assumption that these random variables are pairwise independent and have equal variances σ^2 (Case V of Thurstone's law of comparative judgment) it follows that a particular case of the Fechnerian model holds (with $u(a)$ the expected value of U_a and F the cumulative distribution function of $N(0, 2\sigma^2)$).

Note that the strict utility model (see *Luce's Choice Axiom*) is equivalent both to a special case of the Fechnerian and of the random utility model (compare Suppes et al. 1989, Chap. 17).

While these and other instantiations of the Fechnerian and the random utility model involve specific distributional assumptions, attention will be focused in the following on distribution-free approaches. The concepts that are introduced are probabilistic versions of standard concepts of the representational theory of measurement.

2. *Probabilistic Difference and Additive Conjoint Measurement*

2.1 *Fechnerian Scaling as Probabilistic Difference Measurement*

Recall that a structure $\langle A, \gtrsim \rangle$, where $\gtrsim$ is a weak ordering on $A \times A$, is called an algebraic difference structure if there is representation $u: A \to \mathrm{Re}$ such that

$$(a, b) \gtrsim (c, d)$$
$$\text{if and only if } u(a) - u(b) \geqslant u(c) - u(d).$$

If one defines a quaternary relation $\gtrsim$ by

$$(a, b) \gtrsim (c, d) \text{ if and only if } p(a, b) \geqslant p(c, d),$$

then the strong-utility model holds if and only if the structure $\langle A, \gtrsim \rangle$ is an algebraic difference structure. Hence, Fechnerian scaling can be conceived as probabilistic difference measurement.

In consequence, axiom systems (Krantz et al. 1971, Chap. 4) for algebraic difference structures directly yield testable conditions that have to be satisfied by the choice probabilities for this model to apply.

2.2 Probabilistic Additive Conjoint Measurement

By way of illustration, consider an experimental paradigm where subjects are asked to compare the binaurally induced subjective loudness of successively presented pairs (a, x), (b, y) of tones, with a and b being presented to the left ear and x and y to the right ear. Here two sets of stimuli A and X are referred to (tones presented to the left/right ear) and subjects have to compare *pairs* $(a, x) \in A \times X$ of stimuli. A possible deterministic approach could be to assume that those judgments induce a weak ordering $\gtrsim$ on $A \times X$ and that the structure $\langle A \times X, \gtrsim \rangle$ is an additive conjoint structure, meaning that there are scales $u: A \to \mathrm{Re}$ and $v: X \to \mathrm{Re}$ such that

$$(a, x) \gtrsim (b, y) \quad \text{if and only if} \quad u(a) + v(x) \geqslant u(b) + v(y).$$

Again, in practice a probabilistic variant of this approach is preferable. This means that choice probabilities $p(a, x; b, y)$ have to be considered. A probabilistic variant of the above approach can be obtained by assuming that the ordering $\gtrsim$ defined by

$$(a, x) \gtrsim (b, y) \text{ if and only if } p(a, x; b, y) \geqslant 1/2$$

yields an additive conjoint structure. As before, testable conditions can be directly derived from the standard axiomatizations of additive conjoint measurement. For example, the axiom of double cancellation translates into the condition

$$\text{if } p(a, x; b, y) \geqslant 1/2 \quad \text{and} \quad p(c, y; a, z) \geqslant 1/2, \quad \text{then } p(c, x; b, z) \geqslant 1/2.$$

This model together with the condition $p(a, x; b, y) \leqslant p(d, z; b, y) \Leftrightarrow p(a, x; d, z) \leqslant 1/2$ is equivalent to the existence of a representation of the form

$$p(a, x; b, y) = F[u(a) + v(x), u(b) + v(y)]$$

where F is a function, strictly increasing in the first and strictly decreasing in the second argument. Falmagne (1979) discusses a number of more specific cases of this model and their interrelations. Some of these combine concepts of algebraic-difference measurement (Fechnerian scaling) and additive conjoint measurement, a different such combination being the additive-difference model of preference (see Suppes et al. 1989, Chap. 17). An analogous probabilization of extensive measurement was developed in Falmagne (1980).

3. Mixture Models and Generalized Random Utility Models

A different perspective is taken in some recent developments in which generalized distribution-free random utility models are considered and proven to be equivalent to models based on probabilistic mixtures of standard deterministic measurement structures (Niederée and Heyer 1997; Regenwetter and Marley in press). Being a generalization of Block and Marschak's (1960) classical account of binary choice systems induced by rankings, it lends itself to a probabilization of a wide class of concepts of the representational theory of measurement. The underlying general principles are explained by considering the case of additive conjoint measurement. For simplicity, only two-factorial additive conjoint structures with finite domains $A = X$ will be considered for which there is a representation $u: A \to \mathrm{Re}$ satisfying $(a, b) \gtrsim (c, d)$ if and only if $u(a) + u(b) \geqslant u(c) + u(d)$.

3.1 Mixture Models

Assume choice probabilities $p(a, b; c, d)$ to be given as considered above. The basic assumption underlying an additive conjoint mixture model is that in each experimental trial the subject is in a specific state corresponding to an additive conjoint structure. These structures may vary, each of them occurring with a certain probability. The states themselves will usually be unobservable, but they are assumed to determine the observable individual response (choice) in each situation.

Formally, this can be stated as follows. Let $\mathscr{M}$ be the set of all conjoint structures on $A \times A$, that is, structures $\langle A \times A, \gtrsim \rangle$ with different $\gtrsim$, and let P a probability measure on $\mathscr{M}$ describing the probability with which states, that is, conjoint structures, occur. The measure P is then said to *explain* the choice probabilities $p(a, b; c, d)$ if and only if

$$p(a, b; c, d) = P(\{\langle A \times A, \gtrsim \rangle \in \mathscr{M} | (a, b) \gtrsim (c, d)\})$$

for all $a, b, c, d \in A$.

This concept is related closely to the following distribution-free random utility model.

3.2 Generalized Random Utility Models and Their Relation to Mixture Models

The underlying basic assumption is that a subject's states correspond to sample points ω in the sampling space Ω and that for each $a \in A$ there is a real random variable U_a such that in state ω the subject chooses (a, b) over (c, d) if and only if $U_a(\omega) + U_b(\omega) \geqslant U_c(\omega) + U_d(\omega)$. Again, it is assumed that states occur with certain probabilities.

Formally, let P be a probability measure on Ω. The measure P and a family U_a of random variables are

then said to *explain* the choice probabilities $p(a, b; c, d)$ if and only if

$$p(a, b; c, d) = P(\{\omega \in \Omega | U_a(\omega) + U_b(\omega) \geqslant U_c(\omega) + U_d(\omega)\}).$$

A probabilistic representation theorem can be proven which states that for each system of choice probabilities $p(a, b; c, d)$ there is a probability measure on $\mathscr{M}$ explaining these probabilities if and only if there is a probability measure on Ω and a family U_a of real random variables explaining these probabilities. Hence, the above additive conjoint mixture and random utility model explain the same systems of choice probabilities.

3.3 The Characterization Problem

The question arises whether a set of necessary and jointly sufficient testable conditions can be specified which a system $p(a, b; c, d)$ has to fulfill for it to be explainable by an additive conjoint mixture model. For each finite domain A there is indeed a finite set of linear inequalities of the form

$$d_1 \cdot p(a_1, b_1; c_1, d_1) + \cdots + d_k \cdot p(a_k, b_k; c_k, d_k) \leqslant d$$

($d_1, \ldots, d_k, d \in \mathrm{Re}$) with this property. (The proof refers to a polytope associated with $\mathscr{M}$, the inequalities describing half-spaces generating the polytope.)

From a computational viewpoint, finding such systems of inequalities (associated with the corresponding polytope) is not a trivial task. However, there is a handy general method that allows one to generate *necessary* inequalities corresponding to axioms for additive conjoint measurement. These include

$$\sum_{i=0}^{k-1} p(a_i, b_i; a_{\pi(i)} b_{\sigma(i)}) - p(a_{\pi(k)}, b_{\sigma(k)}; a_k, b_k) \leqslant k-1$$

for all permutations π, σ of $\{1, \ldots, k\}$.

3.4 The General Case

The concepts and results outlined in the previous subsections can be extended to a wide class of relational structures, including linear orders, weak orders, semiorders, partial orders, equivalence relations, algebraic difference structures, extensive structures, and other classes of structures studied in the representational theory of measurement (see *Measurement, Representational Theory of*; *Ordered Relational Structures*). In Niederée and Heyer (1997) a corresponding unifying general conceptual framework and general representation and characterization theorems are presented, which include the above results as special cases (cf. also Regenwetter and Marley (in press); see *Characterization Theorems in Random Utility Theory* for related developments).

4. Falmagne's Concept of Random Additive Conjoint Measurement

A different strategy of probabilizing the concept of additive conjoint measurement underlies Falmagne's notion of random conjoint measurement. By way of illustration, consider again the above example of binaural loudness comparison. For simplicity, stimuli will be identified with suitable numbers (e.g., denoting the stimulus energy). The experimental paradigm is now modified in such a way that for stimuli a, x, y chosen by the experimenter, subjects are asked to determine a stimulus b with the property that (a, x) appears equally loud as (b, y). In a deterministic conjoint measurement framework, a plausible assumption would be that this is the case if and only if $u(a) + v(x) = u(b) + v(y)$. Assume instead that for each triple a, x, y the settings b can be conceived as a random variable $U_{x,y}(a)$ with uniquely defined median $m_{xy}(a)$. Falmagne's model then postulates that there are scales u, v such that

$$u[U_{xy}(a)] = v(x) - v(y) + u(a) + \varepsilon_{xy}(a)$$

where $\varepsilon_{xy}(a)$ is a random variable (denoting error) which has a unique median equal to zero. For such a representation to exist, certain testable conditions must be satisfied. For instance, it must hold that $m_{xy}[m_{yz}(a)] = m_{xz}(a)$.

5. Concluding Remarks

All of the above probabilistic approaches are derived from deterministic measurement concepts. While in Sects. 1 and 2 algebraic measurement structures are considered whose ordering is defined in terms of (choice) probabilities, the mixture models of Sect. 3 refer to an 'urn' of algebraic measurement structures, each of which corresponds to a latent state. The concept of random conjoint measurement of Sect. 4 refers to a single algebraic structure where the experimental outcome is assumed to be pertubated by 'error.' Despite these fundamental differences in perspective, relationships can be established for certain special cases. This was illustrated in Section 1 by Thurstone's Case V. (For the relation between probabilistic and random conjoint measurement see Falmagne 1986, Sect. 9.7.) Further research is needed on the relation among these approaches and the role of independence and distributional assumptions. In this connection it could be profitable to combine these approaches with qualitative characterizations of distributions as found, for example, in Suck (1998).

As in the axiomatic tradition of the representational theory of measurement, emphasis is placed in all of these developments on testable conditions. In the present context, however, empirical tests of these conditions require suitable statistical procedures, many of which still need to be developed (for some results see Iverson and Falmagne 1985).

Which of the above concepts of probabilistic measurement, if any, is 'the correct' one for a given situation will certainly depend on the substantive area under consideration. This suggests that future developments of more specific concepts of probabilistic measurement should be tied to the corresponding substantive theories.

In the above description of problems and results of probabilistic measurement details had to be omitted. For a more detailed and comprehensive account of the material in Sects. 1, 2, and 4 see, for example, Falmagne (1986, Roberts (1979), and Suppes et al. (1989, Chaps. 16–7).

See also: Characterization Theorems in Random Utility Theory; Conjoint Analysis Applications; Elicitation of Probabilities and Probability Distributions; Measurement, Representational Theory of; Measurement Theory: Conjoint; Measurement Theory: History and Philosophy; Probability: Formal; Probability: Interpretations; Psychophysical Theory and Laws, History of; Psychophysics; Test Theory: Applied Probabilistic Measurement Structures; Utility and Subjective Probability: Contemporary Theories; Utility and Subjective Probability: Empirical Studies

Bibliography

Block H D, Marschak J 1960 Random orderings and stochastic theories of responses. In: Olkin I, Ghurye S, Hoeffding W, Meadow W, Mann H (eds.) *Contributions to Probability and Statistics*. Stanford University Press, Stanford, CA

Falmagne J C 1979 On a class of probabilistic conjoint measurement models: some diagnostic properties. *Journal of Mathematical Psychology* **19**: 73–88

Falmagne J C 1980 A probabilistic theory of extensive measurement. *Philosophy of Science* **47**: 277–96

Falmagne J C 1986 Psychophysical measurement and theory. In: Boff K R, Kaufman L, Thomas J P (eds.) *Handbook of Perception and Human Performance, Vol. I: Sensory Processes and Perception*. Wiley, New York

Iverson G J, Falmagne J C 1985 Statistical issues in measurement. *Mathematical Social Sciences* **10**: 131–53

Krantz D H, Luce R D, Suppes P, Tversky A 1971 *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*. Academic Press, San Diego, CA

Niederée R, Heyer D 1997 Generalized random utility models and the representational theory of measurement: a conceptual link. In: Marley A A J (ed.) *Choice, Decision, and Measurement. Essays in Honor of R. Duncan Luce*. Erlbaum, Mahwah, NJ

Regenwetter M, Marley A A J in press Random relations, random utilities, and random functions. *Journal of Mathematical Psychology*

Roberts F S 1979 *Measurement Theory with Applications to Decision Making. Utility, and the Social Sciences*. Addison-Wesley, Reading, MA

Suck R 1998 A qualitative characterization of the exponential distribution. *Journal of Mathematical Psychology* **42**: 418–31

Suppes P, Krantz D H, Luce R D, Tversky A 1989 *Foundations of Measurement, Vol. II: Geometrical, Threshold, and Probabilistic Representations*. Academic Press, San Diego, CA

D. Heyer and R. Niederée

Media and Child Development

Using the Internet, watching television, and listening to music are the main leisure-time activities for most children and adults around the globe. Therefore, media and communication technology is perceived as a major influence on human development, interacting with genetic dispositions and the social environment.

People are attracted to the media: to seek information, to be entertained, or to look for role models. The effect of media and communication technology on human development depends on several factors: (a) the psychological and social modes, which steer the motives of the media user; (b) personal predispositions; (c) the situational and cultural circumstances of the user; (d) kind, form, and content of the media and, (e) the time devoted to media use.

1. Media violence

Children in the twenty-first century are brought up in a media environment where the idea of communication convergence has become reality. There is a permanent cross-over between TV- and Internet-content; between computer-games and telecommunication; between editorial media contributions, and merchandising. Whereas most of this may serve socially constructive structures, i.e., increase information, facilitate communication, etc., it can also be applied in negative ways. Violence has always been a particularly successful media-market factor. It attracts high attention among male adolescents, its language is universal, and with the, often simple, dramaturgy, it can be more easily produced than the complex dialogue-based stories. Television was the dominating medium in the life of children during the second half of the twentieth century. It was often blamed for having negative effects on the young, but it undoubtedly also created numerous prosocial consequences, from programs such as 'Sesame Street,' 'Mr. Rogers' Neighborhood' in the USA, 'Die Sendung mit der Maus' in Germany, or 'Villa Klokhuis' in The Netherlands, to name but a few.

Television may still be popularly utilized, but in most Western children's lives, it is no longer the case that one dominating medium is the 'single source' for information and passive entertainment. In the twenty-first century, people are growing up in a 'digital environment.' With Internet-TV, and technologies, such as 'replay,' and with mobile phones connected to the World Wide Web, any content is potentially accessible at any given moment in any given situation.

However, the media and media-effects debate in the twentieth century has predominantly been a Western, and more specifically an American/Anglo-Saxon issue. However, relatively little research has been conducted concerning a far-reaching global approach to media violence. In the 1980s, Eron and Huesmann (1986) presented a cross-cultural study involving seven countries, including Australia, Israel, Poland, and the USA. In the 1990s, a group of researchers started an international analysis on the media environment of European children in the tradition of Hilde Himmelweit's landmark 1958 study on television (Livingstone 1998). Groebel (2000) conducted a global study for UNESCO with 5,000 children from 23 countries participating in all regions of the world. The purpose was to analyze:

(a) the impact of different cultural norms on possible media effects;

(b) the interaction between media violence and *real* violence in the immediate environment of children, and

(c) the differences between world regions with a highly developed media landscape and those with only a few basic media available.

Children and adolescents have always been interested in arousing, and often even violent, stories and fairy tales (see, Singer and Singer 1990). With the arrival of mass media, film and television, however, the quantity of aggressive content daily consumed by these age groups increased dramatically. As real violence, especially among youths, is still growing, it seems plausible to correlate the two: media violence = aggressive behavior. With more media developments such as video-recorders, computer games and the Internet, one can see a further increase in violent images, which attract great attention. Videos can present realistic torture scenes and even real murder. Computer games enable the user to actively simulate the mutilation of 'enemies.' The Internet has, apart from its prosocial possibilities, become a platform for child pornography, violent cults, and terrorist guidelines. Even with these phenomena, however, it is crucial to realize that the primary causes for aggressive behavior will still likely be found in the family environment, the peer groups, and in particular the social and economic conditions, in which children are raised (Groebel and Hinde 1991).

Yet media play a major role in the development of cultural orientations, world views and beliefs, as well as the global distribution of values and (often stereotyped) images. They are not only mirrors of cultural trends but can also channel them, and are themselves major constituents of society. Sometimes they can even be direct means of intergroup violence and war propaganda. It is important to identify their contribution to the propagation of violence, when one considers the possibilities of prevention.

With the technical means of automatization and, more recently, of digitization, any media content has become potentially global. Not only does individual news reach nearly any part of the world, but mass entertainment has also become an international enterprise. For example, American or Indian movies can be watched in most world regions. Much of what is presented contains violence. In literature, art, as well as in popular culture, violence has always been a major topic of human communication. Whether it is the Gilgamesh, a Shakespearean drama, the Shuihu Zhuan of Luo Guanzhong, Kurosawa's Ran, stories of Wole Soyinka, or ordinary detective series, humankind seems always to be fascinated by aggression. This fascination does not necessarily mean that destructive behavior is innate, however, it draws attention, as it is one of the phenomena of human life which cannot be immediately explained and yet demands consideration of how to cope with it, if it occurs. Nearly all studies around the world show that men are much more attracted to violence than women. One can assume that, in a mixture of biological predispositions and gender role socializations, men often experience aggression as rewarding. It fits their role in society, but may once also have served as a motivation to seek adventure when exploring new territory or protecting the family and the group. Without an internal (physiological thrill seeking) and an external (status and mating) reward mechanism, men may rather have fled leaving their dependants unprotected. But apart from 'functional' aggression, humankind has developed 'destructive' aggression, mass-murder, hedonistic torture, and humiliation, which cannot be explained in terms of survival. It is often these which are distributed widely in the media.

The type of media differ in impact. Audio-visual media are more graphic in their depiction of violence than are books or newspapers. They leave less freedom in the individual images, which the viewers associate with the stories. As the media become ever more perfect with the introduction of three dimensions (virtual reality) and interactivity (computer games and multimedia), and as they are always accessible and universal (video and Internet) the representation of violence 'merges' increasingly with reality.

Another crucial distinction is that between 'context-rich' and 'context-free' depictions of violence. Novels or sophisticated movies usually offer a story around the occurrence of violence: its background, its consequences. Violence as a pure entertainment product, however, often lacks any embedding in a context that is more than a cliched image of good and bad.

Table 1
Potentially problematic and nonproblematic forms of media content

Modes	Investigative	Message oriented	Entertainment
Problematic	Voyeurism	Censorship; propaganda	Rewarded violence
Nonproblematic	Classical journalism	Antiviolence campaigns	Stories; thrills

The final difference between the individual media forms concerns their distribution. A theatre play or a novel, are nearly always singular events. The modern mass media, however, create a time- and space-omnipresence. Even here, a distinction between problematic and nonproblematic forms of media violence has to be made. A news program or a television documentary, which presents the cruelty of war and the suffering of its victims in a nonvoyeuristic way, is part of an objective investigation, or may even serve conflict-reduction purposes. Hate campaigns, on the other hand, or the 'glorification of violence,' stress the 'reward' characteristics of extreme aggression. In general, one can distinguish between three different modes of media content (see Table 1):

(a) purely investigative (typically news);
(b) message-oriented (campaigns, advertisement), and
(c) entertainment (movies, shows).

Although, these criteria often may not be easy to determine, there are clear examples of each of the different forms. Reality television, or paparazzi activities may have to do with the truth but they also, in the extreme, influence this very truth through their own behavior (e.g., the discussion surrounding Princess Diana's death). Through the informal communication patterns on the Internet, rumors also have become part of 'serious' news reporting, as the discussion around Bill Clinton and Monica Lewinsky has shown. Whether true or not, deviant groups and cults can influence the global information streams more efficiency than ever before. The cases of Serbia and Rwanda on the other hand have demonstrated the role which 'traditional' mass propaganda can still play in genocide.

Finally, many incidents worldwide indicate that children often lack the capacity to distinguish between reality and fiction and take for granted what they see in entertainment films, thus stimulating their own aggression (Singer and Singer 1990). If they are exposed permanently to messages which promote the idea that violence is fun or is adequate to solve problems and gain status, then the risk that they learn respective attitudes and behavior patterns is very high.

Many scientific theories and studies have dealt with the problem of media violence since the beginning of the twentieth century. Most of them originate in North America, Australia/New Zealand or Western Europe. But increasingly, Asia, Latin-America and Africa are contributing to the scientific debate. The studies cover a broad range of different paradigms: cultural studies, content analyses of media programs, and behavioral research. However, the terms aggression and violence are defined exclusively here in terms of behavior, which leads to the harm of another person. For phenomena, where activity and creativity have positive consequences for those involved, other terms are used.

Recently, scientists have overcome their traditional dissent and have come to some common conclusions. They assume media effect risk, which depends on the message content, the characteristics of the media users, and their families, as well as their social and cultural environments. In essence, children are more at risk of being immediately influenced than adults. But certain effects, like habituation, do affect older age groups. While short-term effects may be described in terms of simple causal relationships, the long-term impact is more adequately described as an interactive process, which involves many different factors and conditions. Yet as the commercial and the political world strongly rely on the influence of images and messages (as seen in the billion-dollar turnover of the advertising industry or the important role of media in politics), it seems naive to exclude media violence from any effects probability.

The most influential theory on this matter is probably the Social Learning Approach by Albert Bandura (1977) and his colleagues. As most learning comes from observation of the immediate environment, it can be concluded that similar processes work through the media. Many studies have demonstrated that children either directly imitate what they see on the screen, or they integrate the observed behavior patterns into their own repertoire. An extension of this theory considers the role of cognitions. 'If I see that certain behavior, e.g., aggression, is successful, I believe that the same is true in my own life.' Groebel and Gleich (1993) and Donnerstein (1997) both show in European and US studies that nearly 75 percent of the aggressive acts depicted on the screen remain without any negative consequences for the aggressor in the movie, or are even rewarded. The Script Theory, among others, propagated by Huesmann and Eron (1986), assumes the development of complex world views ('scripts') through media influence. 'If I overestimate the probability of violence in real life (e.g., through its frequency on the TV-screen), I develop a belief system where violence is a normal and adequate part of modern society.' The role of the personal state of the viewer is stressed in the Frustration–Aggression Hypothesis (see Berkowitz 1962). Viewers who have been frustrated in their actual environment, e.g.,

through having been punished, insulted, or physically deprived, 'read' the media violence as a signal to channel their frustration into aggression. This theory would explain why in particular, children in social problem areas are open to media-aggression effects.

The contrary tendency has been assumed in the Catharsis Theory, and later the Inhibition Theory by Seyrnour Feshbach. As in the Greek tragedy, aggressive moods would be reduced through the observation of similar tastes with others (substitute coping). Inhibition would occur when the stimulation of own aggressive tendencies would lead to learned fear of punishment and thus contribute to its reduction. While both approaches may still be valid under certain circumstances, they have not been confirmed in the majority of studies, and their original author, Feshbach, now also assumes a negative effects risk.

Much of the fascination of media violence has to do with physiological arousal. The action scenes, which are usually part of media violence, grab the viewer's attention and create at least a 'kick,' most usually among males. At the same time, people tend to react more aggressively in a state of arousal. This would again explain why arousing television scenes lead to higher aggression among frustrated/angered viewers, as Zillmann (1971) explains in his Excitation Transfer Theory. In this context, it is not the content but the formal features, sound, and visual effects that would be responsible for the result. Among others, Donnerstein, Malamuth, and Linz have investigated the effect of long-term exposure to extremely violent images. Men in particular get used to frequent bloody scenes and their empathy towards aggression victims is reduced.

The impact of media violence on anxiety has also been analyzed. Gerbner (1993) and Groebel (Groebel and Hinde 1991) have both demonstrated in longitudinal studies, that the frequent depiction of the world as threatening and dangerous leads to more fearsome and cautious attitudes towards the actual environment. As soon as people are already afraid or lack contrary experiences, they develop an Anxious World View, and have difficulties in distinguishing between reality and fiction. Cultural studies have discussed the role of the cultural construction of meaning. The decoding and interpretation of an image depends on traditions and conventions. This could explain why an aggressive picture may be 'read' differently, e.g., in Singapore than in Switzerland, or even within a national culture by different groups. These cultural differences have definitely to be taken into account. Yet, the question is, whether certain images can also immediately create emotional reactions on a fundamental (not culture-bound) level and to what extent the international mass media have developed a more homogeneous (culture-overspanning) visual language. Increasingly, theories from a non-Anglo-Saxon background have offered important contributions to the discussion.

Groebel has formulated the Compass Theory; depending on already existing experiences, social control, and the cultural environment, media content offers an orientation, a frame of reference which determines the direction of one's own behavior. Viewers do not necessarily adapt simultaneously to what they have observed; but they measure their own behavior in terms of distance to the perceived media models. If extreme cruelty is 'common,' just 'kicking' the other seems to be innocent by comparison, if the cultural environment has not established a working alternative frame of reference (e.g., social control, values).

In general, the impact of media violence depends on several conditions: media content—roughly 10 acts of violence per hour in average programming (see the recent *US National TV Violence Study*, by Donnerstein 1997); media frequency; culture and actual situation; and the characteristics of the viewer and his family surrounding. Yet, as the media now are a mass phenomenon, the probability of a problematic combination of these conditions is high. This is demonstrated in many studies. Based on scientific evidence, one can conclude that the risk of media violence prevails.

2. *Media use*

2.1 *Global Statistics*

Of the school areas in the UNESCO sample, 97 percent can be reached by at least one television broadcast channel. For most areas the average is four to nine channels (34 percent); 8.5 percent receive one; 3 percent two channels; 9 percent three channels; 10 percent 10 to 20 channels, and 18 percent more than 20 channels.

In the global sample, 91 percent of the children have access to a TV set, primarily at home. Thus, the screen has become a universal medium around the world. Whether it is the 'favelas,' a South Pacific island, or a skyscraper in Asia, television is omnipresent, even when regions where television is not available, are considered. This result justifies the assumption that it is still the most powerful source of information and entertainment outside face-to-face communication. Even radio and books do not have the same distribution (91 and 92 percent, respectively).

The remaining media follow some way behind: newspapers, 85 percent; tapes (e.g., cassette), 75 percent; comics, 66 percent; videos, 47 percent; videogames, 40 percent; personal computers, 23 percent; Internet, 9 percent.

Children spend an average of 3 hours daily in front of the screen. That is at least 50 percent more time spent with this medium than with any other activity: homework (2 hours), helping the family (1.6 hours);

playing outside (1.5 hours); being with friends (1.4 hours); reading (1.1 hours); listening to the radio (1.1 hours); listening to tapes/CDs (0.9 hours), and using the computer (0.4 hours).

Thus, television still dominates the life of the children around the globe.

There is a clear correlation between the presence of television and reporting action heroes as favorites. The favorites in Africa are pop stars/musicians (24 percent), with Asia the lowest (12 percent). Africa has also high rankings for religious leaders (18 percent), as compared to Europe/Canada (2 percent); Latin-America (6 percent), and Asia (6 percent). Military leaders score highest in Asia (9.6 percent), and lowest in Europe/Canada (2.6 percent). Journalists score well in Europe/Canada (10 percent), and low in Latin-America (2 percent). Politicians rank lowest in Europe (1 percent), and highest in Africa (7 percent). Again, there may be a correlation with the distribution of mass media: the more TV, the higher the rank of mass-media personalities, and the lower the traditional ones (politicians, religious leaders). There is a strong correlation between the accessibility of modern media and the predominant values and orientations.

At this stage, we can summarize the role of the media for the perception and application of aggression as follows:

(a) Media violence is universal. It is presented primarily in a rewarding context. Depending on the personality characteristics of the children, and depending on their everyday life experiences, media violence satisfies different needs.

(b) It 'compensates' own frustrations and deficits in problem-areas.

(c) It offers 'thrills' for children in a less problematic environment.

(d) For boys, it creates a frame of reference for 'attractive role-models.'

There are many cultural differences, and yet, the basic patterns of the media violence implications are similar around the world. Individual movies are not a problem. However, the extent and omnipresence of media violence contributes to the development of a global aggressive culture. The 'reward-characteristics' of aggression are promoted more systematically than nonaggressive ways of coping with life. Therefore, the risk of media violence prevails.

The results demonstrate the omnipresence of television in all areas of the world. Most children around the globe seem to spend most of their free time with the medium. What they get is a high portion of violent content.

Combined with the real violence, which many children experience, there is a high probability that aggressive orientations are promoted, rather than peaceful ones. But also in lower-aggression areas, violent media content is presented in a rewarding context. Although children cope differently with this content in different cultures, the transcultural communality of the problem is the fact that aggression is interpreted as a good problem-solver for a variety of situations.

Children desire a functioning social and family environment. As they often seem to lack these, they seek role models, which offer compensation through power and aggression. This explains the universal success of movie characters like *The Terminator*. Individual preferences for films of this nature are not the problem. However, when violent content becomes a common phenomenon up to the occurrence of an aggressive media environment, the probability that children will develop a new frame of reference, and that problematic predispositions are channeled into destructive attitudes and behavior, increases immensely.

See also: Alternative Media; Educational Media; Lifelong Learning and its Support with New Media: Cultural Concerns; Mass Media and Cultural Identity; Mass Media: Introduction and Schools of Thought; Media Effects on Children; Media Ethics; Media Literacy; Science and the Media; Violence and Media

Bibliography

Bandura A 1977 *Social Learning Theory*. Prentice-Hall, Englewood Cliffs,

Berkowitz L 1962 Violence in the mass media. *Paris-Stanford Studies in Communication*. Stanford University, Stanford, CA, pp. 107–37

Donnerstein E 1997 *National Violence Study*. University of California, Santa Barbara, CA

Feshbach S 1985 Media and delinquency. In: Groebel J (ed.) *Proceedings of the 7th United Nations Congress Symposium on the Prevention of Crime; Scientific Aspects of Crime Prevention*. Milan

Gerbner G 1993 *Violence in cable-originated television programs: A report to the national cable television association*. NCTA, Washington, DC

Gerbner G, Gross L 1976 Living with television: Violence profile. *Journal of Communication* **26**: 173–99

Groebel J (ed.) 1997 New media developments, *Trends in Communication, 1*. Boom, Amsterdam

Groebel J, Gleich U 1993 *Gewaltprofil des deutschen Fernsehens* [Violence Profile of German Television]. Landesanstalt fur Rundfunk Nordrhein-Westfalen, Leske and Budrich, Leverkusen, Germany

Groebel J, Hinde R A (eds.) 1991 *Aggression and War. Their Biological and Social Bases*. Cambridge University Press, Cambridge, UK

Groebel J, Smit L 1997 *Gewalt im Intemet* [Violence on the Internet]. Report for the German Parliament. Deutscher Bundestag, Bonn, Germany

Huesmann L R (ed.) 1994 *Aggressive Behavior: Current Perspectives*. Plenum, New York

Huesmann L R, Eron L D 1986 *Television and the Aggressive Child: A Cross-national Comparison*, Lawrence Erlbaum, Hillsdale, NJ

Livingstone S 1998 *A European Study on Children's Media Environment*. World Summit on Children's Television, London
Malamuth N 1988 *Do Sexually Violent Media Indirectly Contribute to Antisocial Behavior*. Academic Press, New York, Vol. 2
Singer J L, Singer D G 1990 *The House of Make-Believe*. Harvard University Press, Cambridge, MA
Zillmann D 1971 Excitation transfer in communication mediated aggressive behavior. *Journal of Personal and Social Psychology* **7**: 419–34

J. Groebel

Media and History: Cultural Concerns

1. Text and Image: Relation Reversed

This heading has stopped surprising anyone since the convergence of the two terms has occurred. Nevertheless, right after World War II, studying film as a document and proceeding from there to an analytic rebuttal of the written sources seemed incongruous, even sacrilegious. At that time, when radio and cinema already coexisted with the print media, the legitimacy of pictures remained contested. Only their blue-bloods—painting, museums, private collections—crossed the thresholds of cultivated society, entered the chambers of power and scaled the ivory towers of the historian. Undoubtedly, during the 1930s, a few 'degenerate' countries, such as the Soviet republic, recognized the seventh art—thanks to the newsreels of Dziga Vertov and the analyses of Sergei Eisenstein, or of *Chapayev*, as a source and agent of history. But in their own countries Charlie Chaplin, Jean Renoir, Roberto Rosselini, Fritz Lang were never truly acknowledged as artists, much less as shapers of thought. In fact, it was not until the 1960s that New Wave cinematographers such as Jean-Luc Goddard and François Truffaut succeeded in employing pens and cameras to shape an art which would rival the six others and consequently gain a leading role in historiography. No doubt cinema had commented on history for a long time, but full recognition and sufficient legitimization were not really achieved until this point.

Eisenstein, Luchino Visconti, Rainer Werner Fassbinder, and Steven Spielberg can be counted among the greatest historians of our times, each for his own reason. Eisenstein, because he knew how to translate socialist doctrine into intelligible images in *Strike* by analysing its strategic aspects in a capitalist country at a specific historical juncture: the rise of discontent, the strike, the time of waiting, the provocation, repression. Visconti, for unraveling the relationship of seduction and murder between an aristocratic family and the plebeian power of the Nazis in *La caduta degli dei* (The Damned). Fassbinder for researching and decoding the prohibitions and taboos of contemporary society. All of these were works of historians who go by way of fiction to find paths of analysis, while Spielberg, in *Schindler's List*, plays the instruments of cinema to elicit emotions. The film is more a product of a false memory than a portrayal of history, like Nikita Mikhalkov's. Nonetheless it is a monumental work.

Today we are observing a reversal, the triumph of the image, and an inversion: the image has fallen under suspicion of becoming a prevaricator. One of the most poignant war photographs, *Grief*, by Dmitri Baltermants (1942), has been exposed as a montage. The filmed impaling of Chinese children in Shanghai (1938) has turned out to be a fake, as were many films and photos depicting Lenin, more and more so as the Stalinist era progressed, as first the backdrops of the 2nd Congress of the 3rd International, then Radek, followed by Zinoviev, disappeared. But moving pictures and today images on television, the dictator of morals and opinions, are purported to be a more faithful message—'pictures don't lie'—than political statements. However, this is another illusion. If you change the focus and angle of the view, you alter the message of a picture, as Kuleshov demonstrated in the 1920s.

Elites and leaders tried for a long time to ignore television, just as the preceding generation treated motion pictures with contempt. However, the world of politics recognized its impact in the 1960s and has not stopped trying to exploit this medium. The Nixon–Kennedy debates were the opening round of this struggle. De Gaulle understood the stakes involved and knew how to use the platform to his advantage.

2. Images as Agents of History

The relationship of film to the history of societies assumes several patterns. It is both a reflection and a source of history and may even constitute a historical event in its own right. In the first place, film—initially on the big screen, then on television—acts as an agent of history. Naturally its social and political impact is all the greater if the agencies or institutions that control production and distribution see themselves as propagators of an ideology. The extreme case is the propaganda film. The Nazis went furthest and were the most thorough in the development and production of this genre. They controlled and coordinated the script, the shooting, the selection of actors and the music, and assured distribution and screening by issuing 70,000 16mm projectors in schools and universities from 1936; staging multiple premieres and offering free showings (for example *Jud Süß*), etc.

There is no doubt that the Bolsheviks—above all Lunacharsky and Trotsky—were the first to divine that cinema would become an art for the masses and the art of the future, as well as an educational tool. They were

able to contribute to the grandeur and the glory of Soviet motion pictures, but, unlike the Nazi leaders, they did not control all phases of production because, being members of the intelligentsia, as lawyers, professors, and physicians, they were unfamiliar with, filmmaking and were initially condescending toward the art. They were content to control the explicit denotation, the written screenplay, in the manner of an auditor. As a result, the big names in Soviet film (S. M. Eisenstein, Lev Kuleshov, Vsevolod Pudovkin, etc.), while sympathizers of the new regime, were capable of making films which, although criticizing the Tsarist past, did not really conform to the ruling ideology. This changed during the Stalin era, when a film like *Chapayev* expressed it very consciously. This work was a certain turning point, to the satisfaction of the rulers (1934).

This desire to use film for propaganda purposes should not be considered the exclusive domain of extremists. Democracies have also produced propaganda films and have sometimes even shaped and pursued a policy of using film to this end. This was particularly so in wartime, and most conspicuously in the United States, where from 1941 to 1945 the Roosevelt administration launched an ambitious programme of films designed to justify American intervention in World War II and an alliance with the USSR, etc. Even with no government prompting, the studios had long been marketing movies which served to glorify American social and political systems. It was not by chance that in 1941 one of the most zealous advocates of the American way of life, Frank Capra, was entrusted with making *Why We Fight*. Nonetheless, the US did not lack snipers, soon to become scapegoats of the McCarthy era, such as Dalton Trumbo, Herbert J. Biberman, and a little later Elia Kazan, victims of the hostility of those who did not want their clear consciences troubled.

As an agent of history, film appears not only in its best-known forms: feature, documentary and newsreel. Commercials and promotional films are other vehicles which have more limited targets but are no less effective. Today television performs some of these functions, such as when reducing the role of theatrical film production, or, alternatively, lifting it to heretofore unknown heights when it shows films originally produced for other media. Nevertheless, the functions of cinema and television have been able to remain distinct. An example is the USSR during the Brezhnev era, when the Agitprop torch was passed to television. Cinema managed to escape to a greater extent from the constraints of official ideology, as films like Andrei Tarkovsky's *Andrei Rublev* and Gleb Panfilov's *I Want the Floor* demonstrate. Yet it was television which, during the reign of Andropov and Chernenko, opened a political breach in the system thanks to in-depth reporting on the lamentable conditions in factories and hospitals, absenteeism and incompetence of administrators, summed up as a 'bureaucratic *katastroika*.' Until then, only journals (not newspapers) had had a dissident influence on public opinion, albeit only on a small fraction of the population. Suddenly, television exposed the degenerate trends of the society to a wide audience.

3. *Images as Documents*

Film has won respect as a document, but more in anthropology than in history, and more in Anglo-Saxon countries than in France, Italy, Germany, and Russia. In the late 1990s we recognized that film, after all, constitutes an archive; a document which, directly or indirectly, furnishes testimony on contemporary society, whether it aspires to that function (newsreels, documentaries) or has other objectives (feature films). One of the particular characteristics of the latter was to take a page from novelists like Zola, Camus, Gorky, etc., and present various facts as demonstrations of social and political behaviour. Fritz Lang's *M*, Jean Renoir's *The Crime of Monsieur Lange* and Victorio De Sica's *The Bicycle Thief* (which was actually based on a novel) are a few precedent-setting examples.

Their approach differs radically from that of Joris Ivens and Michelangelo Antonioni's *Chung Kuo Cina*. They take their political assessment as the starting point in filming the 'real' and the captivating. The phenomenon in the early twenty-first century is to employ video for documentary purposes, i.e., using it to write the history of our times. Film investigations which solicit reminiscences and oral accounts of witnesses are legion: Claude Lanzmann's *Shoah* and Jacques Kébadian's *Mémoire arménienne* are legendary examples. They are a continuation of a series of films made by conquered peoples, including the Lakotas in Osawa's *The Black Hills Are Not for Sale*, or of another genre, successors of investigations and documentaries which have appeared since the first motion pictures with John Grierson's English school, Walter Ruttmann in Germany, and Ivens in the Netherlands.

4. *Media Events*

Besides documenting an event, a film can create one. In 1962 Erwin Leiser's *Mein Kampf* triggered an outbreak of children's accusations against their parents in Germany. In 1972 *Le chagrin et la pitié* played the same explosive role in provoking the French national conscience. It far outstripped the effect of works like that of the American historian Robert Paxton on the voluntary collaboration of the Vichy State. A dozen years earlier similar charges by Henri Michel and Eberhard Jaeckel had left the French cold. Another aspect is the media use of some so-called 'events.' An example is the discovery of the so-called unpublished works of Goebbels. They show that the

hype jeopardizes factualness because it creates a hierarchy in information which corresponds to the status of the author rather than to the significance of the information contained in the work.

Another illusion, which was clearly demonstrated during the Gulf War, is that of 'live' history, which claims to be the event and claims to be simultaneous and omnipresent thanks to satellite communications. Reporters at different locations go all out to escalate the drama, filling in the gaps in available footage, and the TV screen becomes a screenplay instead of a mirror of reality. The event ceases to be what is supposed to be the object of the coverage and turns out to be the reporters' picture of the story.

Moreover, Gulf War television coverage illustrated that live history is a myth. A historical event cannot be subjected to a time schedule like a sporting event and it does not abide by any other ground rules unless it degenerates into a show. That was the case at the end of World War II when the assaults of the US Marines were planned to coincide with the positioning of cameras. Nor is a historical event restricted to what is shown. It includes its antecedents and aspects which are not broadcast.

5. *Media Multiplication and Spread of Knowledge*

The multiplication of sources of information, of media, today creates new obstacles to the intelligibility of historical issues because each transmits different bits of knowledge, which are rarely tied together into a more comprehensive picture. We are a long way from the time when knowledge consisted of the contents of a textbook plus a few supplementary readings which made minor or major corrections or additions. However, this development presents several obvious disadvantages.

During the first stages of our education, in secondary school, then at a university, we start by studying separate fields of knowledge—history, foreign languages, literature, economics, etc.,—which usually pay no heed to each other. Those who take Spanish are not acquainted with Dostoyevsky. Even worse, in France, for example, one learns in literature that Jean Jacques Rousseau (see *Rousseau, Jean-Jacques (1772–78)*) drafted a constitution for Corsica around 1760, but the required history of France does not indicate that Corsica had a republican regime before the French Revolution. Later, in higher education, each subject tends to be subdivided into several new disciplines—from archaeological anthropology to culinary epistemology—and most of these try to swallow the others and dominate them by means of a kind of 'transdisciplinary imperialism.' For example, in the social sciences and history, aggressive encroachment passes from economics to linguistics, demography and anthropology. Each field, in turn, attempts to impose its own account of most academic issues.

Later in life, reading the press, we observe that newspapers organize information differently. We see this if we compare five dailies at the recent turn of the century—*El Pais*, *Frankfurter Allgemeine Zeitung*, the *New York Times*, *Le Monde*, *Corriera della Sera*. Naturally, the first page is occupied by the celebration of the new millennium, the Indian Air Lines hostage tragedy, etc. Constancy is maintained by the division into sections. It may be patterned on the organization of government, of ministries, or on the activities of society: foreign affairs, internal affairs, health, etc. Then there are variously placed pages on the courts, culture and sports and a special financial section. The distribution is about the same in the German newspaper, which has a few pages devoted to art, social problems and careers. These sections do not interact naturally unless an occasional editorial or lead story refers readers to another article.

Moreover, the print media often publish an article on history, for example on the anniversary of an event. But this historical piece lives in a returned-to past and is not related to developments which have surfaced in the meantime, as if the past were dead and gone forever. In fact, this attitude is not limited to the print media; a biography of Bismarck does not mention the influence of his ideas after he died. We see the same obliviousness in most biographies, written or unwritten.

Let us now turn to the structure of television programming. It reflects the bureaucratic organization of the channel and its professional and technical divisions: news, features, movies, cartoons, documentaries, etc. Each department is unaware of what its neighbor is doing, unless it is to dispute time slots, since exposure is a matter of vying for the limited amount of viewing windows. Comparing the items devoted to the situation in Russia at the very beginning of 2001; the BBC News reported that Putin spent New Year's Eve with soldiers in Chechnya, a special mentioned his past in East Germany and his encounter with Sobchak in Saint Petersburg, etc. The French TV Journal presented news bulletins but had no in-depth coverage, while features provided analyses but no news items. There were certainly some parallels between French TV news and the contents of daily newspapers of the same day. The former derives its contents from the latter, while the inverse applies to a lesser extent, first, because of the professional rivalry between print and broadcast journalists, the former looking down on the latter, and also because a daily furnishes, to all appearances, much more information and opinion than television, in the light of the amount of air time available and television's goal of entertaining as well as informing.

Apart from that, it is evident that the situation of each medium is different in different countries and that it does not carry the same weight in the political habits

of their citizens. Compared to the influence of other sources of knowledge, the status of television is much weaker in Germany than in the US or France. It is very strong in North Africa and mistrusted in Russia because it is under government control. It did not become the main platform of debate there until the days of *perestroika*. Instead journals were the main opinion makers—not newspapers, not television, not cinema—and the journals transmitted the culture found in school books and the teaching profession. Subsequently everything changed. At the beginning of the twenty-first century there was no longer a dominant pole.

Finally, as far as movies are concerned, going to see a film is a free choice, even more so than turning on the television set (you can switch channels whenever you want) or read a newspaper, usually the same one each day out of habit or preference, or take required university courses. Surfing the Internet is also a choice. Either you master its navigation or you do not. It may be like browsing in an encyclopedia. (The Internet will not be mentioned again in this article.) The public makes choices which indicate its receptivity or lack of interest in subjects which are on offer. The French, for example, wholeheartedly embraced Italian neorealism, Fassbinder's films and those of Woody Allen, which are far from being the most popular in their counties of origin. Is this because they cruelly or humorously describe the shortcomings of their own societies? It has been said that a taboo concerning the war in Algeria prevails in France. As a matter of fact, the conflict engendered more than 50 films, but the French are in no hurry to see them.

The desire not to know disconcerting facts is frequently stronger than the desire to learn the truth. This phenomenon is well known, both in Germany and France. But the same thing goes for Russia, Japan, and Algeria. For whatever the reason, the problem presented here is the non-connectivity of bodies of knowledge; (a) within each community, and (b) from one community to the other.

The problem is therefore to verify the existence or nonexistence of connections between knowledge provided by textbooks, historical scholarship and news media, between professional journalism and the products of the film industry, with the understanding that history is not only knowledge of the past but also the relationship of the past to our times, the analysis of continuities and breaks.

6. *Production of History*

Before continuing, it is a good idea to remind ourselves that there are several approaches to historical questions. First, the philosophical or political approach which makes 'sense' out of history: history is 'progress', history is 'decadence', etc. This vision has prevailed for ages and still exists as a Marxist or Christian view, for example. To express the duality the Russians have two terms while the French have only one. '*Sens*' translates as '*znachenie*' (meaning) and '*napravlenie*' (direction, orientation). The second approach is erudite or scholarly. It involves recording all pieces of information in a chronological calendar of sources and documents of all kinds and attempts to reconstruct the past. This approach is dominant in textbooks. The third is exemplified by the *Annales* school, which regards itself as experimental. It opposes the first, takes advantage of the second and distrusts an academic free-for-all. Risking an analogy, I would say that experimental history is modeled on experimental medicine in the sense that conventional medicine and traditional history address universal problems (life and death, the meaning of history), whereas experimental medicine and new history seek to classify illnesses or phenomena—prices and wages, taxes and strikes, types of war—and then attempt to solve problems. New teaching systems incorporate this aspect but go too far in excluding the second approach. The fourth comprises written fiction and docudrama.

Having said this, how does one explain that a work of fiction, Visconti's *The Damned* (*La caduta degli die*), for example, facilitates comprehension—better than any well-documented text—of the fact that part of the German elite succumbed to Nazism? And how does one explain that two works of fiction, a Solzhenitsyn novel and Abuladze's *Repentance*, enabled readers to understand how the Stalinist regime worked better than any analysis by a political scientist or historian? We could proceed further with the analyses of nineteenth and twentieth century French society by Balzac, Zola and Chabrol, which were more convincingly presented than the essays of many sociologists and humanities scholars. This poses the question of the way works of this type convey a message and suggests the desirability of comparing it to the way journalists, historians, filmmakers and teachers perform their functions.

The collapse of ideologies and the formidable force of history in the making, released partly by the renaissance of the former colonies, has shaken the Eurocentric model of universal history. Now every nation and social group knows how to come to terms with its own history and is starting to compile elements from its memory to recreate its own past.

Today several forms of history survive and coexist alongside factual reporting and fiction: general, traditional history, which is by no means restricted to the Western world and which is the most commonly used model for school textbooks in most countries; memorial history, which is one of the processes of history in the making; and experimental history, exemplified by the journal *Annales*.

Clearly, a historical work usually combines all these forms or models in varying proportions. Only fiction claims to be different. Yet, like the other kinds of

history, it takes pride in scholarship and is not necessarily devoid of significance, at times even becoming a vehicle for ideological thought. This schematic comparison of the role of different approaches and how they work is merely an attempt to answer the question of how news reporting, history, education, and fiction overlap.

The choice of information, or more precisely, the selection criteria, depends on the approach. The first case, which, for the sake of convenience, we will term traditional history, that of textbooks and encyclopedias, certainly involves a sifting process. This follows a hierarchical principle. Some authors and sources rank higher than others. Information originating from people in positions of authority receives more attention than private records. A letter handwritten by Churchill is treated with more respect than accounts of an anonymous witness. Consequently this form of history tends to be a reproduction of statements made by leaders, and often of verbal exchanges between opponents; in short, between those who steer its course. We could call it official history, or more accurately, institutional history, for it is the history of the establishment and opposition institutions. In this model mundane details count for little, because they reveal the flaws in a society and its institutions and fail to highlight changes.

Accumulation is the governing principle in memorial history. Each discovery is another piece in the puzzle recreating a picture of the past. I can rebuild the history of my village, of my community, by collecting and compiling snippets of information.

Experimental history is defined by how it justifies its choice of information. Simply specifying sources is not enough. Experimental historians spell out the selection principle they use to answer questions such as: What is the divorce trend in France? How does the Kurds' status vary in the different States governing their territory? 'These are the documents I will use, and this is how I propose to proceed.'

Film fiction focuses on the facts which seem the most relevant when the movie is being made. This is not the past which is at our fingertips, as in memorial history. It concentrates on what will interest viewers today.

Television and often print media exploit shocking images and sensational headlines to make a lasting impression on people. Vivid illustrations and catchy phrases stimulate the imagination and clamor for attention.

These forms are also based on different organizational principles. The first two stay within a chronological framework, for dating is one of the main criteria for historical documentation, where authenticity is judged by record-keeping accuracy. Experimental history, in contrast, adheres to the dictates of logic. The quality of the text is determined by the stringency of the argument. Historical fiction stresses dramatic and aesthetic aspects. Here history is a matter of beautiful designs, twists and turns, and suspense. But true history, as people experience it or events unfold, does not conform to aesthetic rules. Neither does it abide by the laws of melodrama or tragedy. Just as with news broadcasting, imagining that you are seeing history happen live is something of an illusion. Admittedly, as Ignacio Ramonet recognized, in most cases the images appear in an ordered pattern, like a football match, with a universally fixed structure. However, the format of a game is governed by a rule book. The players know the rules and abide by them. This is not the case with war or the history of a nation.

The specific roles of works of art depend on their nature. And they do not always give those who produce them the same advantages. The task of general history is to instill patriotism and to establish the legitimacy of the powers that be. While the USSR and Islamic countries have reined it in history too tightly, Western regimes have allowed it more leeway. But those who are most successful in achieving these objectives rarely refuse the honors and privileges attached to their vocation—which reminds us of clergy who glorify their church.

Memorial history is an identifier. A group's history fosters coherence. It recounts its experiences, guided by concern for its self-dignity and recognition. This type of history is often anonymous and collective and usually affects one particular occupation or ethnic group. The common heritage is frequently expressed in celebrations, rituals or some kind of ceremony. Experimental history, accompanied by demography, anthropology, sociology, economy, etc, has the latent function of establishing the sovereignty of science. Its practitioners vie for intellectual power. Finally, filmmakers follow the pleasure principle. Their admittedly selfish quest for prestige determines the nature of their creations, which may be aimed at a large audience or target the avant-garde. News coverage serves to enforce the public's right to know 'what is going on.' Freedom of information is one but not the only prerequisite. High visibility and the power of images on screen are the features which influence awareness.

Where is the creativity in these very different types of work? In general history, it is found in the classification and organization of the facts, or at least those which are chosen. In contrast, creativity is not important to memorial history, for this form is based on devotion and reverence. The individual who uncovers and logs the tracks fades into the background and becomes anonymous. In experimental history, creativity is expressed in the choice of problems and not of imaginary situations conceived by the filmmaker or the novelist. Finally, in news reporting, the perspective reflects the creative input of the journalist. As far as television is concerned, late-breaking news can upset the applecart at any moment. Incidentally, news programmes are the part of television broadcasting in which the relationship between the medium

and its audience has changed the least in the 20 years since 1980. It remains a one-way transmission of knowledge from those in the know to those they wish to inform.

Interaction plays only a minor role (in talk shows) while some 'informative' radio programs encourage listener participation. Viewers only intervene to the extent that they are taken into consideration by those who choose particular news items or images. What has changed the most in broadcasting is the degree of autonomy of channels which are more politically independent but more dependent on outside sources, such as the US networks, to obtain news footage. This poses the question whether the demand for images may create a bias, an obstacle to a reasoned analysis of the situation and to the definition of newsworthiness. No pictures—no news. Yet in the beginning of the twentieth century image transmission was unknown. What a turnaround! What a paradox!

Ordering facts is therefore becoming more and more complex and taking a multitude of forms. School books have lost their privileged status. In France, for example, the policy designed to democratize education by not forcing families to pay for textbooks means that they often do not even belong to the pupils any more. The schools issue them to students on loan. The same pupils own cassettes and video games and surf the Internet. The devaluation of textbooks is a serious issue. After all, they provided a framework for knowledge, which we could always subsequently analyze or correct.

See also: Advertising: Effects; Audiences; Film: Genres and Genre Theory; Film: History; Human–Computer Interface; Media Effects on Children; Radio as Medium; Telegraphworks

Bibliography

Ceplair L, Englund S C 1983 *The Inquisition in Hollywood: Politics in the Film Community, 1930–1960*. University of California Press, Berkeley, CA

Ferro M 1993 [1976] *Cinema et Histoire*, completely rev. edn. Gallimard, coll. Folio; 1988 *Cinema and History*. Wayne State University Press, Detroit. Translations into Spanish, Turkish, Italian, Korean (Éditions Kachi), Chinese

Hockings P (ed.) 1975 *Principles of Visual Anthropology*. Mouton, The Hague, The Netherlands

Sorlin P 1980 *The Film in History. Restaging the Past*. Barnes and Noble, Totawa, NJ

Taylor R 1979 *Film Propaganda, Soviet Russia and Nazi Germany*. Croom Helm, London

Wolton D 1991 *War game, l'information et la guerre*. Presses Universitaires de France, Paris

M. Ferro

Media and Social Movements

Contemporary social movements engage in a complicated and uneasy dance with the mass media. Historically, movements have typically had their own media, controlled by movement participants and directed at the people they hoped to rally to their cause. Where mass media were once peripheral to understanding movements, in the last half-century, with the rise of television, they have become central. It is no longer possible to understand the actions of contemporary movements without understanding the nature of the movement–mass media transaction.

Social movements are fields of collective actors maintaining a sustained challenge to institutional practices, policies, and power arrangements and/or to cultural codes by mobilizing people for collective action, at least some of which is contentious. Many of the above terms require some brief elaboration. Collective actors (generally called social movement organizations) are typically formal with designated officers but may also include more informal advocacy networks, often grouped around some publication or think-tank. The field of actors has some means of sustaining the challenge over time; a one-time march or rally does not make a social movement. Challenges may have a strong cultural component as well as representing efforts to change institutions and policies. In democracies, there are institutionalized and prescribed ways of citizens exercising influence on policy using the legal system, electoral system, and the peaceful petitioning of public officials ('lobbying'). Unless at least one actor in the relevant field is engaged in other, more contentious forms of collective action—be it rallies, demonstrations, publicity campaigns aimed at embarrassing opponents, boycotts, strikes, violence, or various other types of unruly behavior—it does not qualify as a social movement.

1. Role of the Mass Media

Social movement organizations typically make use of direct media in the form of newsletters to members and, increasingly, e-mail. In addition, there are often alternative media read by many of their constituents that report sympathetically on their activities (see *Alternative Media*). However, they cannot afford to ignore general audience mass media for several reasons. First, the mass media provide a master forum in the sense that the players in every other forum also use the mass media, either as participants or as part of the gallery. Among the various forums of public discourse, the mass media provide the most generally available and shared set of cultural tools. Social movement organizations must assume that their own constituents are part of the mass media gallery and the messages their would-be supporters hear cannot be ignored, no

matter how extensive the movement's own alternative media may be.

Second, mass media forums are *the* major site of contest politically in part because all of the potential or actual sponsors of meaning—be they authorities, members, or challengers—*assume* pervasive influence (whether justified or not). The mass media often become the critical gallery for discourse carried on in other forums, with success measured by whether a speech in a legislative body or other venue is featured prominently in elite newspapers or in television news coverage.

Finally, the mass media forum is not simply a site of contest and indicator of broader cultural changes in the civil society. The meanings constructed at this site influence changes in language use and political consciousness in the workplace and other settings in which people go about the public part of their daily lives. When a cultural code is being challenged successfully, changes in the media forum both signal and spread the change. To have one's preferred understanding of a policy issue increase significantly in mass media forums is both an important outcome in itself and carries a strong promise of a ripple effect.

It is useful to think of mass media output in terms of the relative prominence of competing *frames*. A frame is a central organizing idea for making sense of relevant events, suggesting what is at issue. 'Media frames,' Gitlin (1980, p. 7) writes, 'largely unspoken and unacknowledged, organize the world both for journalists who report it and, in some important degree, for us who rely on their reports.' Movement organizations compete with others as sponsors of frames, attempting to promote the career of their favorite through a variety of tangible activities—speech-making, pamphlet writing, demonstrations, advertising, interviews with journalists, and the like.

2. *Media Opportunity Structure*

Social movement organizations must compete with other sponsors of frames—public officials, political parties, corporations, vested interest groups, and journalists themselves. Mass media norms and practices can sometimes facilitate movement success in this contest but on balance tend to present a series of obstacles that put them at a relative disadvantage. This media opportunity structure is one subpart of the larger political opportunity structure in which movements must operate in attempting to shape an effective symbolic strategy.

2.1 *Movement Opportunities*

Social movements often make good copy for the media. They provide drama, conflict, and action, colorful copy, and photo opportunities. Because visual material puts a higher premium on spectacle, television is more likely than print media to emphasize it. Spectacle means drama and confrontation, emotional events involving people who have fire in the belly, who are extravagant and unpredictable. This puts a high premium on novelty, on costume, and on confrontation.

Violent action has most of these media-valued elements. Fire in the belly is fine but fire on the ground photographs better. Burning buildings and tires make better television than peaceful vigils and orderly marches. Visual spectacle is high in entertainment value. When news is a vehicle for procuring an audience to sell to advertisers, then one needs to worry about people tuning out. The media opportunity structure provides an incentive for action strategies that provide strong visual material.

2.2 *Movement Constraints*

Social movement organizations typically need to overcome four major obstacles in media norms and practices. The focus here is on the American news media but to the extent that the news media in other countries are similar, the same obstacles will be present. First, journalists have a strong tendency to give official frames the benefit of the doubt. In many cases, the assumptions of these officials are taken for granted; but even when they are challenged by social movement organizations, it is these challengers who carry the burden of proof. A weaker form of this argument is that journalists make official frames the starting point for discussing an issue.

Various observers have noted how subtly and unconsciously this process operates. Halberstam (1979, p. 414) describes how former television news anchor Walter Cronkite's concern with avoiding controversy led to his acceptance of the assumptions underlying official packages. 'To him, editorializing was going against the government. He had little awareness, nor did his employers want him to, of the editorializing which he did automatically by unconsciously going along with the government's position.'

A second disadvantage for social movement organizations stems from the daily news routines of journalists. These news routines bring reporters into regular—sometimes even daily—contact with sources representing public agencies, political parties, corporations, and other well-established groups. They rarely include regular contact with groups who are challenging the dominant interpretations of events. Social movement organizations, then, lack the routine access to journalists that are characteristic of their rivals.

Third, other media norms and practices in the United States—particularly the *balance norm*—favor certain sponsors of alternative frames. In news accounts, interpretation is generally provided through

quotations and balance is provided by quoting spokespersons with competing views. The balance norm is vague and the practices it gives rise to favor certain spokespersons over others. Organized opposition is a necessary condition for activating the norm. Once invoked, there is a strong tendency to reduce controversy to two competing positions—an official one and (if there is one) the alternative sponsored by the most vested member of the polity. In many cases, the critics may share the same unstated common frame as officials, differing only on secondary issues of the most effective means of achieving particular outcomes.

The balance norm, however, is rarely interpreted to include challenger frames, even when no other alternative is available. Tuchman (1978, p. 112) argues that balance in television news in the United States 'means in practice that Republicans may rebut Democrats and vice-versa' but that 'supposedly illegitimate challengers' are rarely offered the opportunity to criticize government statements. Instead, she suggests, 'reporters search for an "establishment critic" or for a "responsible spokesman" whom they have themselves created or promoted to a position of prominence.'

Finally, there is a culture clash between the more pragmatic and cynical subculture of American journalists and the more idealistic and righteous subculture typical of movement organizations (see Eliasoph 1998). Movements seem to demand unreasonable and unrealistic things and often have a righteousness that is unappealing to those who are living with the inevitable compromises of daily life. Movements hector people and call them to account. This means that internal movement conflicts and peccadilloes will have a special fascination for journalists, giving them an opportunity to even the score from their standpoint. As Gamson and Wolfsfeld (1993, p. 120) put it, 'The fall of the righteous is a favored media story wherever it can be found, and movements offer a happy hunting ground.'

3. *Strategic Dilemmas of Movements*

These characteristics of the media opportunity structure create a series of dilemmas for the symbolic strategies of social movement organizations. Movement-media communication is, to quote Gamson and Wolfsfeld (1993, p. 119), 'like a conversation between a monolingual and a bilingual speaker.' The media speak mainstreamese, and movements are tempted to adopt this language since journalists are prone to misunderstand or never hear the alternate language and its underlying ideas. Movement activists often feel that something has been lost in translation. One can accept the dominant cultural codes and not challenge what is taken for granted but for many movements this would surrender fundamental aspects of their *raison d'être*.

There is, then, a fundamental ambivalence and, for some, an estrangement between movements and media. Movement activists tend to view mainstream media not as autonomous and neutral actors but as agents and handmaidens of dominant groups. The media carry the cultural codes being challenged, maintaining and reproducing them. In this sense, they are a target as much as a medium of communication. But they are also the latter since one tries to speak *through* the media rather than to them. This dual media role is the central problematic from the movement standpoint and gives rise to the following dilemmas.

3.1 *Publicity as Validation*

When demonstrators chant, 'The whole world is watching,' it means that they matter, that they are making history. The media spotlight validates the fact that the movement is an important player. Conversely, a demonstration with no media coverage at all is a nonevent, unlikely to have any positive influence either on mobilizing followers or changing the target. No news is bad news.

Given the lack of routine access, getting attention is a necessary step in promoting a preferred frame. But the very tactics that are employed to garner this attention can detract from the preferred frame one is sponsoring. Members of the club enter the media forum through the front door when they choose, are treated with respect, and given the benefit of the doubt. Movement organizations must find some gimmick or act of disorder to force their way in. But entering in this manner, they risk being framed as crazies or fanatics and the preferred frame they are promoting may be obscured in the process. 'Those who dress up in costume to be admitted to the media's party will not be allowed to change before being photographed,' write Gamson and Wolfsfeld (1993, p. 122).

3.2 *Weak Control Over Followers*

Movement organizations face the additional dilemma that, even if they engage in strategic planning, their weak control over the actions of their followers may make it impossible to implement their plan. The media influence internal movement leadership by certifying some people or groups and ignoring others. Some media-designated leaders may have had few followers before their annointment but with their media-generated celebrity, they soon find followers. As Gitlin (1980) argues, it is precisely those leaders who are attached to followers only through their media image and are unaccountable to the rank and file who are likely to advocate the extravagant and dramatic

actions that generate good media copy. This symbiotic relationship can undercut the carefully planned media strategies of more accountable leaders.

4. Measuring Success

A movement strategy can be judged successful if it solves two problems: gaining *media standing* and increasing the *relative prominence of its preferred frame* in mass media discourse.

4.1 Media Standing

In legal discourse, standing refers to the right of a person or group to challenge in a judicial forum the conduct of another, especially with regard to governmental conduct. The rules for according legal standing have been anything but fixed and clear; rather than a matter of clear definition, legal standing is a battle ground.

By analogy, media standing is also contested terrain. In news accounts, it refers to gaining the status of a regular media source whose interpretations are directly quoted. Note that standing is not the same as being covered or mentioned in the news; a group may be in the news in the sense that it is described or criticized but has no opportunity to provide interpretation and meaning to the events in which it is involved. Standing refers to a group being treated as an agent, not merely as an object being discussed by others.

From the standpoint of most journalists who are attempting to be 'objective,' the granting of standing is anything but arbitrary. Sources are selected, in this view, because they speak as or for serious players in any given policy domain: individuals or groups who have enough political power to make a potential difference in what happens. Standing is a measure of achieved cultural power.

Journalists operating in a news forum try to reflect their perceptions of who the key players are, but in practice, they are influenced by various other factors in choosing sources and quotes. In cultural contests, sources are often chosen because they are seen as representing a particular perspective. Rather than being seen as representative in the sense of typical, they are chosen as prototypes who represent a particular cultural tendency in a compelling and dramatic way. In this sense, standing still reflects a journalistic political judgment about which cultural movements make a difference or are serious players.

4.2 Frame Prominence

If one charts mass media coverage of some issue domain over time, frames and their associated idea elements and symbols will ebb and flow in prominence. Success in gaining new advantages in cultural terms is measured by changes in the relative prominence of the movement's preferred frames compared to antagonistic or rival frames. This measure of success can be extended to other non-news forums as well. Frames can be extracted from cartoons, films, advertising, and entertainment as readily as from news accounts. The prominence of preferred movement frames can be assessed over time in such forums in the same way as in news forums.

Using these two measures, four media outcomes can be defined for movements. *Full response* means that the movement organization receives both media standing and a significant increase in the prominence of its preferred frame. *Collapse* means it receives neither standing nor increased prominence for its preferred frame. *Co-optation* means that the group receives media standing but no significant increase in its preferred frame. Finally, *pre-emption* means that the challenger's preferred frame has significantly increased in media prominence in spite of the absence of media standing for its sponsor.

See also: Adolescents: Leisure-time Activities; Cultural Expression and Action; Information Society; Internet: Psychological Perspectives; Mass Communication: Normative Frameworks; Mass Media and Cultural Identity; Media and History: Cultural Concerns; Media Ethics; Media Imperialism; Media, Uses of; Political Communication; Political Discourse; Popular Culture; Social Movements, History of: General; Television: Genres; Television: Industry

Bibliography

Eliasoph N 1998 *Avoiding Politics*. Cambridge University Press, Cambridge, UK

Epstein E J 1973 *News from Nowhere*. 1st edn. Random House, New York

Gamson W A, Wolfsfeld G 1993 Movements and media as interacting systems. *The Annals of the American Academy of Political and Social Science* **528**: 114–25

Gans H J 1979 *Deciding What's News*. Pantheon Books, New York

Gitlin T 1980 *The Whole World Is Watching*. University of California Press, Berkeley, CA

Halberstam D 1979 *The Powers That Be*. Knopf, New York

Hallin D C, Mancini P 1984 Political structure and representational form in US and Italian television news. *Theory and Society* **13**: 829–50

Iyengar S, Kinder D R 1987 *News that Matters*. University of Chicago Press, Chicago, IL

Morley D 1980 *The 'Nationwide' Audience*. British Film Institute, London

Paletz D L, Entman R M 1981 *Media, Power, Politics*. The Free Press, New York

Ryan C 1991 *Prime Time Activism*. 1st edn. South End Press, Boston
Sigal L V 1973 *Reporters and Officials*. D. C. Heath, Lexington, MA
Tuchman G 1978 *Making News*. Free Press, New York

W. A. Gamson

Media Effects

1. Introduction

Communications research, or media studies, is about effect. It might have been otherwise—consider the study of art, for example—but it is not. However, the field is subdivided—audience measurement, content analysis, production process, reception studies—the underlying aim, not always acknowledged, is to account for the power of the media.

From Plato's admonition that the written word might corrupt unintended audiences to Postman's admonition that television would corrupt rational discourse, there has been continuous speculation—both scholarly and popular—about the effects of media. The questions, however, are much better than the answers. How effective was the use of radio for propaganda in World War I? What role did radio play in Hitler's rise to power? Did Roosevelt's 'fireside chats' persuade Americans to join the allies in World War II? Did the televised pictures from Vietnam and Bosnia hasten American withdrawal in the one case and engagement in the other? Is television responsible for the personalization of politics? Do television debates affect the outcome of presidential elections? Are the media simply salesmen of the *status quo*? Does cinematic glorification of violence induce real-world violence? Why is it taking so long for the media to get people to quit smoking? How does representation of minorities on television affect intergroup relations? Will global media homogenize or exacerbate cultural differences?

There are some answers. Scholarship has shown, for example, that the printing press played a part in the Protestant Reformation by undercutting the Church's monopoly on the word of God; that the newspaper fostered integration of nation–states; that the radio broadcast of an 'invasion from Mars,' resulted in a near-panic; that the live telecast of Anwar Sadat's visit to Jerusalem changed Israeli opinion towards the prospect of Middle East peace; that the televised debates of 1960 helped John Kennedy win the presidency; that a heavy diet of media violence in childhood may influence later behavior. Even this handful of questions and answers is enough to illustrate the futility of generalizing about media effects without a conceptual scheme to link types of 'effects' to particular attributes of 'media.' Even if an agreed scheme were available, the complexity of operationalizing it makes evident why the empirical study of media effects should have veered so strongly towards the ostensibly simple study of media campaigns, that is, towards the power of persuasive messages to change the opinions, attitudes and actions of individuals in the very short-run.

One attempt to map the field was made by Lazarsfeld (1948). Coming from the man whose name is so closely associated with the study of campaigns and their 'limited effects,' it is reassuring to note that he was quite clear about what he was and was not doing. Lazarsfeld's scheme is based on the cross-tabulation of response-time (immediate, short-run, long-run, institutional) to media stimuli ranging from single units (a particular broadcast, for example), to genres (soap opera, for example), to type of ownership (public vs. private broadcasting, for example), to types of media technology (print vs. broadcast, for example). Thus, addressing the long run, he speculated that Cervantes's Don Quixote (a single unit) influenced our image of chivalry, that foreign language radio broadcasts (genre) contributed to the assimilation of immigrants, that commercial advertising (ownership) cultivated cynicism, and that radio (technology) demeaned the experience of listening to classical music. Note how this typology equates effect with change, ignoring the possibility that the media may also be effective in *slowing* change in the service of the status quo, cognitive or societal. Note, too, that Lazarsfeld's typology is focused largely on individuals rather than on societal change.

Amending Lazarsfeld's scheme the discussion that follows distinguishes among (a) the nature of effect: change vs. reinforcement; (b) the object of effect: opinions, social organization; (c) the unit affected: individual, group, institution, nation; (d) the time frame of the response: immediate, short-run, long-run; and (e) the active ingredient, or attribute, on the media side: technology, ownership, content, and the context associated with reception. This still-crude schema will help differentiate among several major traditions of research on media effects, to parse how each defines media and effect and to identify the theory that is invoked to connect them.

2. Persuasion

From about 1940 to 1960, research on media effects focused on the study of persuasion, that is, on whether the media can tell us 'what to think' or what to do. More formally, the question is whether the media are able to induce (a) change (b) in the opinions, attitudes, and actions (c) of individuals (d) in the short-run (e) by

means of campaigns of persuasive appeals. This work centered on Carl Hovland at Yale and Paul Lazarsfeld at Columbia, but the paradigm also applies to studies of advertising's wartime propaganda and, somewhat circuitously, to studies of the effect on children of media portrayals of violence.

Hovland's group were experimentalists (McGuire 1996) who varied the rhetoric of persuasive messages–fear appeals, repetition, one-sided vs. two-sided arguments—in order to assess their impact on attitude change both in the laboratory and in the real-life orientation of American soldiers during World War II. Hovland (1959) gave much thought to the problem of why change is more likely in the laboratory than in the field.

Lazarsfeld and his co-workers worked in the field, applying survey research to the study of mass persuasion, particularly to how people make up their minds in election campaigns. The well-known results of this work, as summarized by Klapper (1960), is that the media have only 'limited effects,' that is, that the response of attempts to influence individuals via mass media is more likely to reinforce prior attitudes than to change them. Explanation of these results, in turn, led to an interest in the obstacles that persuasive message encounter enroute to their targets as well as the conditions under which they are most likely to succeed. Two of the identified obstacles were the tendency to protect prior cognitions and prejudices through selective attention, perception, and retention, and the tendency to anchor opinions and attitudes in small groups of valued others, from whose norms and network, members are reluctant to depart unilaterally. Academically, these studies reflected the then-current interest in the social psychology of influence, in the validity of the theories of mass society in which individuals were thought to be disconnected from each other, in the power of radio which had just come of age, and, not least, in the methodology of survey research which nurtured the twin fields of opinion and communication.

Persuasive messages are more likely of success, it was thought, if they could be harnessed to existing cognitions (canalization), and endorsed by existing social networks (supplementation). Monopolization of the message, as in totalitarian societies, was also thought to enhance effect (Lazarsfeld and Merton 1948). It is ironic, but probably true, that these assertions comforted the media establishment by absolving them of the sin of brainwashing (even at the expense of deflating their egos). Researchers were comforted, too, that the mass was less vulnerable than had been feared (even at the expense of failing to find 'powerful effects').

Belief in the persuasive potential of the media—certainly among politicians and advertisers, but also among researchers in fields such as public health, for example—has fueled continuing efforts to identify the conditions under which communications campaigns are more likely to succeed (McGuire 1986). The ability of the new media to custom-tailor their messages is a new condition of research interest (Turow 1997). Most researchers do not deny 'limited effects.' Some, however, dissent (e.g., Zaller 1996, Iyengar and Simon 2000), arguing that research on short-run effects is poorly designed in that it fails to take account both of the variance in the reach of competing messages—say in an election campaign—and variance in the attention of message receivers to information and influence.

3. *Diffusion*

Having identified interpersonal networks and selectivity as keys to persuasion, researchers invoked these processes to explore two additional paradigms; one is called 'diffusion' and the other 'uses and gratifications.'

Diffusion research is an effort to trace the adoption of an idea or innovation as it spreads, over time, among a community of potential adopters exposed both to the media and to each other. It reaches out to the many disciplines that are interested in how things get from here to there, such as the spread of religion, stylistic change in fashion and the arts, change in language and naming, or the epidemiology of disease. It builds on the finding of campaign research that interpersonal relations are a major key both to influence and adoption.

Diffusion studies, ideally, combine the sample survey with sociometry, as in Coleman et al.'s (1966) study of the adoption of a new antibiotic by communities of doctors and Ryan and Gross' (1943) study of the adoption of a new farm practices. Rogers (1983) summarizes this work, and applies it also to studies of developmental communication and certain aspects of nation building (Lerner 1958). These several branches of diffusion research posit that we are told 'when to think' (or act). They focus on (a) change, (b) in the behavior (c) of networks, (d) in the middle to short-run, (e) in response to mass and interpersonal messages. Diffusion differs from persuasion in the perception that influence takes time, especially as it has to wend its way through the norms and networks of a community.

If persuasion research had to lower its expectations of the media, so did diffusion research, which had no illusions. But diffusion research was beset by the difficulty of generalizing across its many case studies; even the self-same innovation, while diffusing, may take on new meanings. And the diffusion paradigm did not take adequate account of power, it was said, in its over simple assumption that decision-making was in the hands of autonomous individual adopters. A further argument against diffusion research is that its findings have a tautological ring: people adopt (things that are closest to) what they already have. For these reasons, diffusion research has waned.

4. Gratifications

The other lead from persuasion studies—the idea of selectivity—points in a different direction. Selectivity, as has been noted already, suggests that audiences have to 'resist' media influence or bend it to their own wishes. This possibility turned attention to what audiences do with the media rather than vice versa, and to hypotheses about the ways in which persons with particular interests or particular roles adapt media messages—even the media themselves—to their own purposes. Early studies looked at the advice-giving functions of the soap opera or the self-testing implicit in quiz programs, or the uses of classical music as anticipatory socialization for the upwardly mobile (for reviews see Blumler and Katz 1974, Rosengren et al. 1985). The pioneering studies tended to result in a list of functions as reported by audiences, while later studies proceeded more rigorously to identify media uses associated with different 'interests' or 'needs' or roles. Thus, there were studies comparing use of the media by children who were well-integrated and poorly-integrated in their peer groups; adolescents whose parents had and had not coached them in sexual knowledge; towns whose newspapers were and were not on strike; voters who had and had not made up their minds.

Even more oriented to limited effects, gratification studies lean towards a paradigm of nonchange, or (a) reinforcement (b) of role or interest (c) of individuals (d) in the relatively short-run (e) via a media smorgasbord (or tool-chest). They imply that the media tell us 'what to think with.' While the selectivity of gratification studies coincides with one of the by-products of persuasion research, it is somewhat exaggerated to claim that gratification research is a direct heir to persuasion studies; in fact, the two paradigms co-existed at Columbia. Gratification draws more on functional theory than persuasion research, and can be applied to groups and whole societies, not only, to individuals. Ball-Rokeach (1985) has addressed these issues in her theory of 'media dependency.'

Gratifications studies are also precursors of the more recent 'reception theory,' proposed by an alliance of literary and communications theorists. Reading, they say, is a collaborative effort between text and audience—and, more radically, that readers are the authors of their texts.

Like its partners in the tradition of limited media effects, gratification research has waned somewhat, victimized by the disenchantment with functional theory. Specifically, critics ask gratificationists whether the needs and roles which they attribute to persons and groups were not themselves created by the media so as to gratify them. They also ask—as they do of reader reception theory—whether theory has not gone too far in abolishing the power of the text in favor of the reader's power to read upside down or inside out.

5. Agenda Setting

Media effects are limited—as we have seen so far—because they are intercepted by prior attitudes and group norms, that is, by the selectivity imposed by prior commitments to cognitive and social structures. It follows that if there are powerful media effects, they will have to circumnavigate these obstacles. This inference underlies several major movements to break loose from limited effects, and redeem communication research for bigger things. If so, we may define theories that claim powerful effects by whether and how they neutralize or side step the barriers to persuasion.

Beniger (1987) suggests that one way to discover powerful effects is to focus on the operation of information, rather than influence. Information, says Beniger, is neutral—unlike influence-attempts, it does not threaten prior cognitive structures or structures of social relations. A prime example is the agenda-setting paradigm, which does not presume to tell us 'what to think' but, rather, 'what to think about.' Agenda-setting research, hand in hand with priming and framing (Price) is a flourishing enterprise since the pioneering papers of McCombs and Shaw (McCombs 1981). It fits well with the idea that media confer status on persons not just on issues (Lazarsfeld and Merton 1948). It maps on our scheme as (a) change (b) in the attention (c) of individuals or society (d) in the short-run (e) in response to information. This resonates with some of the classic writings on media effects, and is not as innocent as it sounds. It is, evidently, a powerful effect in itself, but also can operate indirectly. If the media can prime citizens to think 'domestic policy' rather than 'foreign policy,' it may be that the President's performance will be evaluated differently.

6. Knowledge Gap

Cognitive theory also fits nicely with another paradigm of media effects called information-gap or knowledge-gap. Its thesis is that an information (not influence) campaign will be more readily absorbed by those who are information-rich on the subject of the campaign than by the information-poor (Tichenor et al. 1970). Even though everybody learns something—about ways of investing money, or contraceptive techniques–the net result will be that the gap between the classes will further widen, thus paving the road to hell with another good intention. The explanation, it appears, lies not with the valence of a cognitive structure—since the message is not challenging an existing attitude—but the complexity of the structure, inasmuch as well-developed structures can more easily make room for yet more. Thus, knowledge gap studies may be mapped on our scheme as (a) reinforcement (b) of the stratification (c) of society (d) in the short-run (e) as a result of an information campaign. The requisite theory, of course, proceeds from the level of

individual response to the level of social structure. It can be sloganized as the media telling us 'who should think.'

7. Critical Theory

Reinforcement, that is, nonchange is the aim of the media according to its most famous proponent, the Frankfurt School. Unlike the reinforcement of gratifications or of knowledge-gap studies, which are based on 'audience power,' the critical Frankfurt School sees the media as agents of hegemonic power, telling their audiences 'what *not* to think.'

Ironically, critical theory and classic persuasion theory shared similar views of a powerless mass audience and of media omnipotence. But the Frankfurt School was interested in long-run impositions on society not in the short-run effects of campaigns on individuals. They conjure up a conspiratorial culture industry, presided over by captains of industry, determined to impose a soothing false-consciousness. This is the best of possible worlds, say the media, and, with luck, you'll be discovered (just as Hollywood discovered the girl next door). Lowenthal's (1944) study of biographies in popular magazines captures this dynamic.

Ownership of the media is what concerns the Frankfurt School, and the mass-production and distribution of entertainment is their strategy for neutralizing the power of the audience to defend itself. The seemingly egalitarian mix of 'Benny Goodman and the Budapest String Quartet' (Horkheimer and Adorno 1972) is a way of packaging the message of classlessness, and thereby effacing a working-class culture that might keep class consciousness alive. Conformity and consent is the end-product of 'what not to think.' This is a neo-Marxist theory which continues to enlist a distinguished following, albeit with a less conspiratorial bent (e.g., Gitlin 1978, Gerbner and Gross 1976).

There is critical theory on the right as well. Noelle-Neumann (1984), for example, indicts journalists (not owners) for falsifying reality. The result, as in the Frankfurt School, is to promulgate a fabricated consensus which serves to silence opposition—even of a real majority who come to believe, erroneously, that they are part of an embarrassed minority.

Classical critical theory (no less than functional theory) has trouble explaining change and, *a fortiori*, representation of change in the media, for the obvious reason that the theory expects the media to reinforce the status quo. In response to the strong movements for change in the second half of the twentieth century—civil rights, feminism, etc.—critical theorists have, ironically, become interested in variations in audience reception. This openness to the possibility that members of the audience may resist media dominance, and that different readers may read differently, has been championed by the Birmingham branch of critical theory, under the leadership of Hall (1980), founded on the neo-Marxism of Raymond Williams and Richard Hoggart. This group has pioneered the establishment of cultural studies that conceptualizes the media as a site for negotiation. Critical and functional theorists now find themselves ringing the same doorbells to observe viewers' negotiations with the once unchallenged text.

The critical approach to media effects would map on our scheme as (a) reinforcement (b) of the stratification (c) of society (d) in the long-run (e) powered by the hegemonic consortium of media managers and owners. The Birmingham branch of critical theory would allow for much more variation in audience reception of the media.

8. Technological Theory

The powerful effects proposed by technological theory circumvent audience defenses by dismissing the influence of content altogether. McLuhan (1964) is the most radical advocate of this approach, arguing that the media do not tell us what to think, but 'how to think.' This controversial guru of media effects believed that the media technologies of each epoch discipline our brains. Thus, cognitive processing of the linearity of print creates personalities and societies that think in lines (e.g., before-and-after causuality, hewing to lines, delaying gratifications until the end of the line, etc.) and act in lines (assembly lines, railroad lines, chains of command). The dotted screen of television, on the other hand, is much more diffuse, and invites viewer participation to make sense of the ambiguous signal. McLuhan's colleague and mentor, Innis (1950) emphasized the socio-organizational implications of different media technologies—how 'space-oriented' media such as papyrus, for example, enabled the Egyptian empire, while 'time-bound' media, such as pyramids, make for religious continuity across generations. These are the themes of the Toronto School.

More in the spirit of Innis than McLuhan, technological theory has enlisted many first-rank researchers. Goody and Watt (1963) have spelled out the effects of transition from orality to literacy; Eisenstein (1979) has explored the effect of the printing press on religion, science, and scholarship; Tarde (1989) credits the newspaper with the rise of the public; Carey (1989) has shown how the telegraph affected the economic integration of the United States; Gouldner (1976) suggests that the proliferation of paper created a need for ideology; Meyrowitz (1985) argues that television's accessibility has lowered the boundaries that separate generations, genders, classes, etc. Rather than wild speculation, each of these theories is specific about the particular attribute of media technology (the fixedness of print, the accessibility of television, the simultaneity of the telegraph) that is responsible for the hypothesized effect. As a group, technological theories can be

mapped in two related ways. McLuhan proposes (a) change (b) in the mental processing (c) of individuals (d) in the long-run (e) as a result of unique technological attributes of the different media. This combines with the more characteristic emphasis on (a) change (b) in social organization (c) of societies and institutions (d) in the long-run (e) in response to media technologies.

Technological determinism is much criticized, of course, not just in communications research. It is easy to show that the same technology—radio, for example—has been put to different uses at different times, to different effect. It can also be argued that the same technology has different effects under different social and cultural circumstances—as is the case with European-style public television and American-style commercial television.

9. *Sociological Theory*

For such reasons, the sociology of media effects proposes a 'soft determinism' that harnesses media technologies to social management. It is interested in the effect of managed media technologies on social organization emphasizing social maintenance and reinforcement rather more than social change. Its best-known proponents were at the University of Chicago in the first decades of the twentieth century, and were concerned with the assimilation of immigrants, the coherence of cities, the integration of nation–states, the fostering of participatory democracy, and the role of the media—among other factors—in these processes (Carey 1996). This loosely-knit group, which included John Dewey, Charles Cooley, Robert Park, Louis Wirth, Herbert Blumer, W. I. Thomas and, later, David Riesman, Morris Janowitz, Kurt and Gladys Lang, were much influenced by the social psychology of Tarde and Toennies. Some of the earliest studies on the possible relation between media and deviance were conducted in Chicago.

Chicago is where modern communication research began, and it is an irony of intellectual history that its immediate successor, the Columbia School, hardly took notice. While Chicago and Frankfurt share the idea that the media reinforce social structure, Chicago places emphasis on social organization and disorganization, whereas Frankfurt (and knowledge-gap studies) emphasizes stratification and domination. It shares an interest in technology with Toronto. But it shares almost nothing with Columbia. Where Columbia is interested in (a) change (b) of opinion (c) of individuals (d) in the short-run, (e) due to media content, Chicago may be said to be interested in (a) reinforcement (b) of the integration (c) of society, city, nation, ethnicities (d) in the long-run due to (e) the shared experience made possible by content and technology. A group at MIT added a cybernetic paradigm to this perspective in their studies of the media in nation building (e.g., Deutsch 1966). Work on the role of media and 'media events' in the integration of nation–states continues in this tradition (Dayan and Katz 1992, Cardiff and Scannell 1987) and approaches Carey's (1989) call for study of the ritual aspects of mass communication.

10. *Situational Theory*

The social and physical locus of media reception cannot be considered as intrinsic as can technology or even content, and yet it is so much a part of the media that it may be said to have its own effect. Stereotypically, the newspaper is read alone, television is watched at home with family, and movies are viewed at a downtown theater with friends. The fact that these settings are in constant flux makes it possible to ask, for example, what difference it makes to read the news on paper or on the Internet, or to watch a film at a theater or on TV.

It is comparisons of this kind that are implicit in the tradition of psychoanalytic study of the cinema (Flitterman-Lewis 1987). Surprisingly, perhaps, cinema theorists wed technology and psychology in proposing that a darkened theater and projection from behind-the-head (abetted by the seamless Hollywood narrative) regress cinema-goers to an infantile stage in which a movie becomes their dream or their voyeuristic experience, with the reenactments and renegotiations that that implies. This contrasts with the television experience which invites a more dialogic stance based on the vis-a-visness of the set and the intimacy of the setting (Horton and Wohl 1956). Houston (1984) has developed this comparison.

This approach to 'positioning' is also central to reception studies. In its most elementary form readers know whether a text is intended for them—which is what makes books different from TV, according to Meyrowitz (1985)—and its more subtle forms gives the reader an identity or a role to play. Along the same line, television events—presidential debates, for example, or the World Cup—position viewers as citizens, or as sport fans, in roles very different from the consumer (and other) roles ascribed by everyday television. Cyberspace, perhaps, proposes altogether different identities.

That cinema and books are central to this tradition of work reveals how differently researchers themselves approach the several media. The concept of 'identification,' for example, so central to thinking about the effects of film does not figure at all centrally in the empirical study of television effects. Some studies of the effects of TV violence on children have used the concept (Bandura 1963) as have some studies of soap opera and television drama.

The performing arts share this interest. Adorno (1942) insists, for example, that listening to classical music at home is a different experience than listening in a concert hall. But, it might be added, many of the

compositions one hears reverently in the concert hall were once background music in some feudal court.

It is relevant, too, to recall the concern with which the Royal Family confronted the BBC's proposal to broadcast the coronation of Elizabeth II in 1953, lest the viewers at home fail to conform to the decorum expected of them. Cinema audiences can be controlled, went the argument, so can the audience on the route of march, but how can one make certain (Scannell 1996) that the television audiences will take the proper 'position'?

Perhaps the earliest paper in this tradition is Freidson (1953) and its flowering is in psychoanalytic studies of cinema and reception theory. It maps on our scheme as (a) change or reinforcement (b) of the identity or role (c) of individuals (d) in the relatively short-run (e) in response to media 'positioning.' Perhaps this approach can be sloganized as the media telling me 'who I am (now),' and about my selves.

11. Recapitulation

This has been an effort to present multiple approaches to the study of media effects without attempting to present the findings of these approaches, or to assess their validity. It is an effort to bring some order to the conceptualizations of media effects, and to put the overblown 'persuasion' paradigm into perspective.

Each approach is characterized by linking an aspect of the media (technology, ownership, content, situation of contact) to a particular kind of effect. Effects are classified as change vs. reinforcement of some aspect (e.g., opinion, role, organization) of individuals or social systems within a given time frame. It takes theory to spell out the processes that link 'cause' and 'effect, and relevant theories have been specified wherever possible. A catchphrase ('what to think,' 'what to think about,' 'what not to think,' etc.) was proposed to caricature each approach. And, whenever appropriate, a School name (Chicago, Columbia, Toronto, etc.) was assigned to each (see Table 1).

This summary highlights the unease in the field between the findings of so-called 'limited effects,' and the both popular and theoretical intuition that the media are 'powerful.' It does so, first of all, by showing that different kinds of effects may be limited or powerful, and may well co-exist. Second, it spells out the operational meaning of 'limited' in terms of audience resistance, whether cognitive or interpersonal, finding resistance particularly strong when media content make frontal assaults on embedded attitudes and behavior. It follows, therefore, that proposals of powerful effects must identify processes that circumvent or overcome such resistance, and, indeed, this is the case for theories labeled here as cognitive, critical, technological, psychoanalytic.

The history of study of media effects is full of discontinuities, suffering less from competition among

Table 1
Map of Media Effects

Theory	Approach	Slogan	"School"	Effect				
				(1) Change?	(2) What?	(3) Unit	(4) Time	(5) Active media attribute
Social psych	Persuasion	What to think	Yale Columbia	Change	Opinion	Individual	Immediate	Content (appeals)
Network	Diffusion	When to think	Columbia Land Grant Universities	Change	Behavior	Networks	Short-run	Content (information)
Functional	Gratifications	With what to think	Columbia	Reinforce	Role	Individual	Short-run	Content (tool kit)
Cognitive	Agenda-Setting	What to think *about*		Change	Attention	Individual Society	Short-run	Content (information)
Cognitive	Knowledge-Gap	Who should think		Reinforce	Stratification	Society	Short-run	Content (information)
Neo-Marx	Critical	What *not* to think	Frankfurt	Reinforce	Stratification	Society	Long-run	Ownership
Neo-Marx	Cultural Studies	What to negotiate	Birmingham	Reinforce	Resistance	Individual Groups	Short-run	Content
Cognitive; Social-History	Technological-Determinist	How to think How to organize	Toronto	Change	Organization	Individual Society	Long-run	Technology
Sociology	Sociological	With whom to think	Chicago	Reinforce	Organization	Society	Long-run	Technology
Anthropology; Poli Sci; Psycho-analytic	Situational	Who am I		Reinforce change	Identity role	Individual	Short-run	Situation

paradigms than from paradigm exhaustion. Two kinds of requiems have been recited, repeatedly: (a) that media effects are exaggerated, and/or that there is nothing much left to learn (Berelson 1959) and (b) that effects research is exaggerated—either too positivistic or too wild—and that study of the media will benefit from abandoning the obsessive search for provable effects (Carey 1989). Both suffer from an impoverished conceptualization of 'effect.'

See also: Advertising: Effects; Advertising: General; Agenda-setting; Audiences; Broadcasting: General; Consumer Psychology; Diffusion, Sociology of; Market Research; Mass Media and Cultural Identity; Mass Media: Introduction and Schools of Thought; Mass Media, Representations in; Media Effects on Children; Media Ethics; Media Imperialism; Media, Uses of; News: General; Printing as a Medium; Radio as Medium; Television: History; Television: Industry

Bibliography

Adorno T W 1942 The radio symphony. In: Lazarsfeld P F, Stanton F (eds.) *Radio Research 1941*. Duell, Sloane and Pearce, New York

Ball-Rokeach S J 1985 The origins of individual media system dependency. *Communication Research* **12**: 485–510

Bandura A 1963 Influence of model's reinforcement contingencies on imitative responses. *Journal of Personality and Social Psychology* **1**: 589–95

Beniger J R 1987 Toward an old new paradigm: The half-century flirtation with mass society. *Public Opinion Quarterly* **51**: S46–S66

Berelson B 1959 The state of communication research. *Public Opinion Quarterly* **23**: 1–5

Blumler J G, Katz E 1974 *The Uses of Mass Communication*. Sage Publications, Beverly Hills, CA

Cardiff D, Scannell P 1987 Broadcasting and national unity. In: Curran J, Smith A, Wingate P (eds.) *Impacts and Influences*. Methuen, London, pp. 157–77

Carey J W 1989 *Communication as Culture: Essays on Media and Society*. Unwin Hyman, Boston

Carey J W 1996 The Chicago school and mass communication research. In: Dennis E E, Wartella E (eds.) *American Communication Research: The Remembered History*. Erlbaum, New York

Coleman J S, Katz E, Menzel H 1966 *Medical Innovation: A Diffusion Study*. Bobbs-Merrill, Indianapolis

Dayan D, Katz E 1992 *Media Events: The Live Broadcasting of History*. Harvard University Press, Cambridge, MA

Deutsch K W 1953/1966 *Nationalism and Social Communication* Second Edition. Cambridge:MIT Press

Eisenstein E 1979 *The Printing Press as an Agent of Change: Communications and Cultural Transformation in Early-Modern Europe, Vol 1*. Cambridge University Press, Cambridge, UK

Flitterman-Lewis S 1987 Psychoanalysis, film and television. In: Allen R C (ed.) *Channels of Discourse: Television and Contemporary Criticism*. The University of North Carolina Press, Chapel Hill, NC, pp. 172–210

Freidson E 1953 The relation of the social situation of contact to the media in mass communications. *Public Opinion Quarterly* **17**: 230–38

Gerbner G, Gross L 1976 Living with television: The violence profile. *Journal of Communication* **26**(2): 173–99

Gitlin T 1978 Media sociology: The dominant paradigm. *Theory and Society* **6**: 205–53

Goody J, Watt I 1963 Consequences of literacy. *Comparative Studies in Society and History* **5**: 304–45

Gouldner A M 1976 *The Dialectics of Ideology and Technology*. Macmillan, London

Hall S 1980 Encoding/decoding. In: Hall S, Hobson D, Low A, Wills P (eds.) *Culture, Media, Language: Working Papers in Cultural Studies (1972–1979)*. Hutchingson, London

Horkheimer M, Adorno T 1972 *The Dialectics of Enlightenment*. Allen Lane, London, pp. 126–67

Horton D, Wohl R R 1956 Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry* **19**(3): 215–29

Houston B 1984 Viewing television: The metapsychology of endless consumption. *Quarterly Review of Film Studies* **9**: 183–95

Horland C I 1959 Reconciling conflicting results from experimental and survey studies of attitude change. *American Psychologist* **14**: 8–17

Innis H A 1950 *Empire and Communication*. University of Toronto Press, Toronto

Iyengar S, Simon A F 2000 New perspective and evidence on political communication and campaign effects. *Annual Review of Psychology*

Klapper J T 1960 *The Effects of Mass Communication*. Free Press, Glencoe, IL

Lasswell H 1927 *Propaganda Techniques in the World War*. Knopf, New York

Lazarsfeld P F 1948 Communication research and the social psychologist. In: Dennis W (ed.) *Current Trends in Social Psychology*. University of Pittsburgh Press, Pittsburgh

Lazarsfeld P F, Merton R K 1948 Mass communication, popular taste and organized social action. In: Bryson L (ed.) *The Communication of Ideas*. Harper & Row, New York

Lerner D 1958 *The Passing of Traditional Society: Modernizing the Middle East*. Free Press, Glencoe, IL

Lowenthal L 1944/1979 Biographers in popular magazines. In: Lazardsfeld P, Stanton F N (eds.) *Radio Research: 1942–1943*. Duell, Sloan and Pearce and Arno Press, New York, pp. 507–48

McCombs M E 1981 The agenda-setting approach. In: Nimoro D D, Sanders K R (eds.) *Handbook of Political Communication*. Sage, Beverly Hills, CA, pp. 121–40

McGuire W J 1986 The myth of massive media impact. In: Comstock G (ed.) *Public Communication and Behavior*. Academic Press, Orlando, FL

McGuireW J 1996 Persuasion and communications at Yale. In: Dennis E E, Wartella E (eds.) *American Communication Research: The Remembered History*. Erlbaum, Mahwah, NJ

McLuhan M 1964 *Understanding Media: The Extensions of Man*. McGraw Hill, New York

Meyrowitz J 1985 *No Sense of Place: The Impact of Electronic Media on Social Behavior*. Oxford University Press, New York

Noelle-Neumann E 1984 *The Spiral of Silence*. The University of Chicago Press, Chicago

Price V, Tewksbury D 1997 News, values and public opinion: A theoretical account of media priming and framing. *Progress in the Communication Sciences* **13**: 173–212

Rogers E M 1983 *Diffusion of Innovation*, 3rd edn. Free Press, New York

Rosengren K L, Wenner P, Palmgreen P 1985 *Media Gratifications Research: Current Perspectives*. Sage, Beverly Hills, CA
Ryan B, Gross N C 1943 The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology* **8**: 15–24
Scannell P 1996 *Radio, Television and Modern Life: A Phenomenological Approach*. Blackwell, Cambridge, UK
Tarde G 1989/1901 *L'opinion et la Foule*. Presses Universitaires de France, Paris
Tichenor P J, Donahue G A, Olien C N 1970 Mass media, flow and differential growth in knowledge. *Public Opinion Quarterly* **34**: 159–70
Turow J 1997 *Breaking Up America: Advertisers and the New Media World*. University of Chicago Press, Chicago
Zaller J 1996 The myth of massive media impact revisited: New support for a discredited idea. In: Mutz D, Sniderman P, Brody R (eds.) *Political Persuasion and Attitude Change*. University of Michigan Press, Ann Arbor, MI

E. Katz

Media Effects on Children

The twentieth century brought with it the dawn of electronically based media that became integrated into children's daily lives. Children's media environments include information delivered in print such as comic books, by mass broadcast systems such as television, film, and radio, by audio and video cassette recorders, and by the more recent digital interactive computer-based technologies such as video games, desktop computers, CD-ROM, the Internet, and virtual reality. Although a shift took place from technologies that delivered information to a mass audience to those that were increasingly responsive to the individual, broadcast television was the dominant and most influential medium used by children in the twentieth century. As the major creator and exporter of television programs, the USA played a key role in the kinds of programs available to children throughout the world. Recurrent questions, debates, and policy initiatives included the impact of media on children's (a) moral and social development, (b) cognitive development, (c) gender and ethnic stereotypes, and (d) use of time. These same issues will guide children and media research in the digital media environments of the twenty-first century.

1. *Impact of Media on Children's Moral and Social Development*

1.1 *Media Violence*

Issues about how media impact children's moral and social development typically focus on the area of violence. Wartella and Reeves (1985) traced the history of recurring media issues, including violence, in American society. The role that films play in shaping children's morality was a research focus of the Payne Fund in the 1920s and 1930s. Antisocial radio and comic book content also emerged as a social concern. Television, however, became the medium in which the controversy about media violence became a matter of serious and sustained public and scientific inquiry. Over the last five decades of the twentieth century, the United States Congress periodically held hearings, animated in part by public pressure to protect children and society from antisocial, aggressive behavior.

Several key theoretical models have been used to examine the impact of violent content on children's behavior. These models included (a) *cultivation theory*, which examined the degree of violent content in television programs over time and how they might shape dominant and prevailing views of how much violence there was in a culture, (b) *social learning* or *social cognitive theory*, which predicted *imitation* of violent acts as well as *disinhibition* of acts that children typically inhibit, but which are acted upon when they observe media models act aggressively with impunity, (c) *arousal theory*, which hypothesized that heavy viewers of violent content would become *habituated* and *desensitized* to violent content, (d) *script theory*, which used observational and enactive learning to explain the durability of aggressive behavior, and (e) *psychoanalytic theory*, which predicted decreases in children's aggressive conduct after viewing violent media content because innate aggressive impulses were harmlessly drained off via fantasy experiences, a process known as *catharsis*.

Thousands of research studies have been conducted on the impact of violent media content on children, including the Surgeon General's Report on *Television and Social Behavior*, published in 1972. Taken together, the results of this body of research reveal that (a) the amount of violent television content has remained steady over time, with children's cartoons portraying the most frequent number of violent acts per episode, (b) children imitate and become disinhibited in aggressive behaviors after exposure to violent television programs, (c) children who view programs with heavy concentrations of violence become habituated and desensitized to the content, (d) aggressive effects of media are durable over time, and (e) there is little evidence of catharsis. Similar outcomes are found for the newer technologies, notably violent video and virtual reality games.

A key methodological criticism was that most studies showing a causal link between exposure to media violence and childhood aggression were conducted in laboratory, rather than in naturalistic, environments. However, field experiments using naturalistic methods of inquiry also demonstrated deleterious effects of violent content on children and so did longitudinal correlational studies. Similarly,

meta-analyses, in which large numbers of studies were examined simultaneously, supported the thesis that television violence is one cause of childhood aggression.

1.2 Prosocial Media

Just as media can cultivate aggressive antisocial conduct, so too can they cultivate prosocial behaviors that can contribute to the moral development of children. Prosocial behavior includes altruistic acts such as helping, sharing, and understanding the feelings of others. Stein and Friedrich (1972) first demonstrated the potential of television to cultivate prosocial behaviors by comparing violent cartoons with prosocial content. Preschoolers who saw the violent content became more aggressive in their subsequent play while those who saw the prosocial programs increased in task persistence, delay of gratification, and rule obedience. Children from lower socioeconomic backgrounds also increased in cooperation, nurturance, and verbalization of feelings. Later studies of both commercial and public prosocial television programs demonstrated similar beneficial effects for children of all social classes. Effects were particularly pronounced when rehearsal techniques, such as verbally labeling and role playing (i.e., enacting) the prosocial messages, were utilized.

1.3 Commercial Media

Media are often commercial enterprises with the intent to sell products to users and to export programs world wide. A commercial financial base, which characterizes American television but not that of all nations, impacts the kind of content that children have available to them. Arguments have been advanced that the increased commercialism in children's entertainment media has undermined social values by increasing materialistic attitudes.

In the beginnings of television, industry created many educational and prosocial programs for children, in part so that families would buy a television set. Once there was a critical mass of television owners, networks began to concentrate their efforts on selling audiences to advertisers, their source of revenue. Because children are a smaller and less lucrative audience than adults, children's programs became much less available than adult programs.

Young children are also limited by the cognitive skills that they bring to bear on media content. They believe, for instance, that commercials are there to assist them in their buying decisions. By contrast, older children increasingly understand the persuasive intent of commercials, and many become cynical about advertised products by the end of middle childhood, although they still want many of the products. Older children have successively been taught the deceptive techniques used by commercial advertisers, but younger children do not have the cognitive capacities to understand commercial persuasive intent.

Action for Children's Television (ACT) became a major public advocacy group arguing for the regulation of advertisements in American children's television programs. Public pressure on the American television industry during the 1970s led broadcasters to (a) limit the amount of advertising on children's programs, (b) use commercial separators to segment the commercial from the program content, and (c) ban program hosts from selling products to children within their own program or in segments surrounding their program.

By the 1980s, however, the Federal Communications Commission (FCC), the American government agency that regulates broadcast media, changed directions and instituted a deregulation policy. Children were no longer considered a special audience in need of protection, and the market place was left to regulate itself. In that deregulatory environment, program-length commercials (or product-based programs) emerged. Program-length commercials were programs created around a product that advertisers wanted to sell to children. Most American children growing up in the 1980s viewed programs that were intimately linked to a specific product with a consumer focus rather than one that was meant to foster their development. Moreover, the amount of time that broadcasters spent in traditional kinds of advertising to children increased.

By the 1990s, the Internet emerged as a new medium for advertising to consumers, including children. With little initial governmental regulation of Internet commercial practices, advertisers created sites where they asked children for personal information about who they were and about what kinds of buying practices they had. Favorite media characters targeted individual buyers by writing personal online messages to specific children. Environments were seamless so that it was difficult for children to tell the difference between a commercial and other kinds of content.

Public advocacy groups continued to pressure the government to regulate both the Internet and commercial television practices. The United States Congress passed the Children's Television Act of 1990, placing time limits on the amount of commercial advertising that could take place on children's programs, and the Children's Online Privacy Protection Act, forcing American advertisers to stop soliciting on-line, personally identifying information from children. Nonetheless, it remains legal for advertisers to use cookies, an Internet tracking device that keeps detailed records of which users visit specific Internet sites.

2. Cognitive Media Effects

With the advent of each new technological innovation, predictions are made about the vast potential for educational practice to be revolutionized. These predictions remain largely unfulfilled because of the commercial nature of most media and early limitations inherent in media delivery. Nonetheless, educational innovations and successes take place with the development of new media.

Cognitive developmental approaches, such as those advanced by Bruner, and information processing theory are typically used to examine children's learning from educational media. In cognitive developmental theory, children are thought to represent and to think about information in different ways, depending upon their stage of development. The three levels of thought are enactive representations of information with the body, iconic visual representations, and symbolic verbal representations.

The ways that children think map onto the symbol systems used to convey content by various representational media. These symbol systems, known as formal features, structure, mark, and represent content through visual and auditory production techniques such as action, sound effects, and dialogue. Radio, for instance, relies on auditory features for delivering messages whereas television relies on visual and auditory features. Because very young children tend to think in pictures as well as words, television may provide a developmentally appropriate code for them to represent and remember content in iconic and symbolic modes.

Information processing theory examines the flow of information through cognitive processes such as perception, attention, and memory of events. The perceptual attention-getting characteristics of media, such as the use of sound effects, can improve children's attention to, and memory for, plot-relevant television content. Children's memories of the explicitly presented content that is central to a program and children's memories of abstract inferential content, such as character feelings, motivations, and temporal links of causal program events, are enhanced by the use of character action. While young children can remember important program events that are explicitly presented, older children have metacognitive skills that allow them to understand the abstract, inferential material as well.

Sesame Street stands out as a landmark achievement in creating effective educational television programs that children choose to view. This carefully constructed program uses production techniques, such as singing and sound effects, to carry important verbal information. Evaluations of the program suggest that advantaged and disadvantaged children learn cognitive lessons, are better prepared to enter school, and are better readers early in life. Longitudinal follow-ups reveal that adolescents who had been heavy viewers of Sesame Street as children had better grades in high school. Empirical examinations of Sesame Street also led to the development of the comprehensibility model, an approach emphasizing the active, strategic nature of children's attention as they search for meaning in television programs (Anderson et al. 1981).

Realizing the unique appeal and promise of broadcast television as an educator of children, the United States Congress passed the Children's Television Act of 1990. As a condition for license renewal, commercial broadcasters had to provide educational and informational programming for children, defined as content that could improve the cognitive intellectual and/or social and emotional development of children aged 16 years and under. After several years, in which the educational merit of broadcaster offerings were of dubious quality, the FCC implemented the three hour rule, requiring each broadcaster to provide three hours of educational and informational programming each week.

Because of the broad definition used by the FCC in defining educational programming, most broadcasters presented prosocial rather than academic content for their educational offerings. Although children often take away social and emotional lessons from these programs, the potential of television to educate children in academic content areas as it entertains them still remains vastly underutilized in American society. Public broadcasting remains the main source of academically oriented television programs for American children.

Although other educational television programs can also enhance children's attention and learning, most have been unable to sustain an audience or the test of time. Inherent limitations exist in educational television presentations that have been difficult to overcome. These problems include a lack of user control, the different knowledge bases of different learners, and the suppression of imaginative and creative skills. The video cassette recorder greatly expanded children's rehearsal options. As multimedia environments emerged, different media could also activate different cognitive skills and improve learning by fostering interactive, individualized exchanges between children and educational content.

3. Displacement Effects

What have children given up to listen, to view, or to interact with media? Uses and gratification theory provides the framework for understanding these shifts in leisure time activities. In this theory, children and adults have certain needs, such as entertainment or education. Children use media to fulfil or to gratify those needs.

Film and television have been heavily criticized for replacing valuable educational activities, such as reading books. Early research suggests that television

displaced activities that filled similar needs for children in the USA as well as in other nations (Murray and Kippax 1978, Schramm et al. 1961). For instance, television replaced comic book reading, not book reading. Although later cross-cultural research did implicate television viewing with a decrease in children's reading skills, watching educational television is related to reading books. Thus, displacement effects depend on children's specific uses of media.

Media also displace one another. Just as television displaced going to films and listening to the radio in the middle of the twentieth century, video games and Internet computer use began to take time away from television viewing by the end of the century. Nonetheless, television remained children's dominant and most used technology of the twentieth century with American children spending two–three hours per day in the presence of an operating television set. Access to cable and satellite television content increased children's exposure. The digital age will add increased clarity to images and multiple options to use media in interactive ways. Historical media usage patterns suggest that entertainment needs have and will continue to drive American children's use of media though the children of certain other cultures, such as Israel, seem to invest more effort in extracting educational messages.

4. *Impact of Gender and Ethnic Media Stereotypes*

Children look to media to learn information about themselves as well as others. Because children bring their own unique qualities to media experiences, schematic information processing models, in which learned expectations guide perception, memory, and inferences about others, have been key theoretical approaches for understanding how stereotyped media content impacts them.

In American television programs, stereotypes about men and women have been prevalent throughout the twentieth century in terms of the number of major roles portrayed and the kinds of personality characteristics, occupational roles, and behaviors displayed. Children understand and remember the stereotyped presentations that they see. Counterstereotyped portrayals can also be learned by children, but children sometimes distort the content of these presentations to fit traditional gender stereotypes. For instance, children tend to remember television portrayals of male nurses and female doctors as male doctors and female nurses.

Video games are overwhelmingly created around male interests, such as aggression, fostering boys' early interest in the world of computers. Although many video games contain violent content, children also develop visual spatial skills by playing them. Computer content also often reflects male interests, such as mathematics. Boys spend more time interacting with computers than girls do, particularly as children get older and as programming skills are required. Girls, however, do participate with computers, particularly when they perceive the interaction as gender appropriate.

Internet explorations by adolescents often take the direction of identity experimentation. In a virtual world where one can create his or her own identity, many young men and women engage in gender-swapping, in which males present themselves as females and females present themselves as males (Turkle 1995). These presentations are particularly prevalent on MUDs, multiuser domains or dungeons, in which adolescents enact various roles and create imaginative places in cyberspace.

Ethnic stereotypes are also prevalent in television portrayals. The Civil Rights Movement led to changes in the representation of African Americans. Roles became less stereotyped, and the number of African Americans males on television approximated their real-life numbers in the American population. At the end of the twentieth century, Latinos were the most under-represented group of American citizens in comparison with their prevalence in society. Women of all ethnic minority groups remained virtually invisible and under-represented. There is relatively little literature about how these portrayals impact children. What is known is that children of ethnic minority groups prefer seeing characters that are of their own ethnic background, and these characters sometimes serve as models for them.

Computer purchases reveal that Asian Americans buy the most computers, followed by Caucasian Americans, Latino Americans, and African Americans. These buying patterns suggest that some ethnic groups will have earlier access to the information highway and to computer technologies than will others, perhaps creating a digital divide that accentuates disparities in future educational and employment opportunities among these groups.

5. *Conclusion*

The twentieth century witnessed the emergence and displacement of various media, yet each maintained its own unique place in children's lives. Across media, themes about morality, stereotypes, education, and children's use of their free time recurred as public interest groups attempted to protect and improve children's media environments. These same themes will continue to guide research and policy initiatives in the twenty-first century as multimedia systems converge and merge.

See also: Advertising: Effects; Advertising, Psychology of; Childhood Depression; Childhood Health; Media

and Child Development; Media Effects; Media, Uses of; Television: General; Violence and Effects on Children; Violence and Media; Visual Images in the Media; War, Political Violence and their Psychological Effects on Children: Cultural Concerns

Bibliography

Anderson D, Lorch E P, Field D E, Sanders J 1981 The effects of TV program comprehensibility on preschool children's visual attention to television. *Child Development* **52**: 151–7

Calvert S L 1999 *Children's Journeys Through the Information Age*, 1st edn. McGraw-Hill, Boston

Calvert S L, Tan S 1994 Impact of virtual reality on young adults' physiological arousal and aggressive thoughts: Interaction versus observation. *Journal of Applied Developmental Psychology* **15**: 125–39

Children's Television Act of 1990 1990 Publ. L. No. 101–437, 104 Stat. 996–1000, codified at 47 USC Sections 303a, 303b, 394

Corteen R, Williams T 1986 Television and reading skills. In: Williams T M (ed.) *The Impact of Television: A Natural Experiment in Three Communities*. Academic Press, Orlando, FL, pp. 39–84

Gerbner G, Gross L, Morgan M, Signorielli N 1980 The mainstreaming of America: Violence profile No. 9. *Journal of Communication* **32**: 100–27

Greenberg B, Dominick J 1970 Television behavior among disadvantaged children. In: Greenberg B, Dervin B (eds.) *Uses of the Mass Media by the Urban Poor*. Praeger, New York, pp. 51–72

Greenfield P 1996 Video games as cultural artifacts. In: Greenfield P M, Cocking R R (eds.) *Interacting with Video*. Ablex, Norwood, NJ, pp. 85–94

Federal Communications Commission 1991 Report and order: In the matter of policies and rules concerning children's television programming. *Federal Communications Commission Reports*, Report No. 2111

Federal Communications Commission 1996 August 8 FCC adopts new children's TV rules (MM Docket 93–48). *Federal Communications Commission News*, Report No. DC 96–81

Huston A, Wright J 1998 Mass media and children's development. In: Damon W, Sigel I, Renninger K (eds.) *Handbook of Child Psychology, Vol. 4: Child Psychology in Practice*, 5th edn., Wiley, New York, pp. 999–1058

Jordan A, Woodard E 1998 Growing pains: Children's television in the new regulatory environment. In: Jordan A, Jamieson K (Special Volume eds.) *Children and Television: The Annals of the American Academy of Political and Social Science*. Sage, Thousand Oaks, CA, pp. 83–95

Montgomery K, Pasnik S 1996 *Web of Deception: Threats to Children from Online Marketing*. Center for Media Education, Washington, DC

Murray J P, Kippax S 1978 Children's social behavior in three towns with differing television experience. *Journal of Communication* **28**: 19–29

Schramm W, Lyle J, Parker E B 1961 *Television in the Lives of Our Children*. Stanford University Press, Stanford, CA

Stein A, Friedrich L 1972 Television content and young children's behavior. In: Murray J P, Rubenstein E A, Comstock G A (eds.) *Television and Social Behavior. Vol. II: Television and Social Learning*. US Government Printing Office, Washington, DC, pp. 202–317

Turkle S 1995 *Life on the Screen*. Simon and Schuster, New York

Wartella E 1988 The public context of debates about television and children. In: Oskamp S (ed.) *Applied Social Psychology Annual*. Sage, Newbury Park, CA, Vol. 8, pp. 59–69

Wartella E, Reeves B 1985 Historical trends in research on children and the media: 1900–1960. *Journal of Communication* **35**: 118–33

S. L. Calvert

Media Ethics

Media ethics is a branch of ethics that addresses moral issues arising in connection with the obtaining, preparation, storage, presentation, and dissemination of information through the means of mass media. Mass media include print media (newspapers, magazines, and books), recordings, motion pictures, and electronic media (radio, television, and the computer). Media ethics seeks to help the media practitioners resolve various moral problems arising in all the areas of media communications: journalism, advertising, public relations, and entertainment.

The media exercise a strong and complex influence upon the perception and understanding of the world by the public and, consequently, upon shaping the personality of each individual and the interactions of individuals with one another. News and reportage, commercials and advertisements, soap operas and films—all exert in the long run more or less subtle influence on people's views, choices, and behavior. Because of the ubiquity of the media and their growing presence, the ethical problems that the media practitioners face become increasingly important.

There are a number of issues discussed in media-ethics, and the importance of these issues depends on the function of the media and on the nature of the media. For example, the problem of objectivity is far more important in journalism than in entertainment or advertisement; moreover, the problem of objectivity is much more affected by the visual media than by print media.

1. Truth in the Media

Traditionally, discussion concerning media ethics concentrated around journalistic ethics for which the problem of truth is of critical importance. There are two reasons for this importance. First, the media are the primary source of information in a democracy and reliable information is indispensable for a genuine democratic political system. The government must be accountable to the people it represents, and one means of ensuring this is access to information related to governmental policies and informed discussion about government. For this reason, the press is often viewed

as the fourth estate that monitors the actions of the three governmental estates: the legislature, executive, and judiciary (Thomas B. Macaulay's concept). Second, the use of inaccurate information contributes to a decline in the autonomy of the individual. Deception is an exercise of self-interest; it makes making choices harder and undermines human freedom.

One facet of truth in journalism is objectivity, that is, the commitment to impartiality, to reflecting the world as it is, without any bias or distortion. The journalist should be a neutral, detached observer who reports facts free from personal opinion. There are aspects of each event that are independent of the convictions of reporters. News reports attempt to capture the essence of an event and its underlying causes, and locate the event in a social, historical, and political context. However, perfect objectivity is impossible to achieve. First, there is a problem of selection. News is not just reporting what recently happened; otherwise the most trivial events would qualify as news. It is the selection of the news based on its importance for the community as reinforced by showing these events in a larger context. Also, what is news depends very much on where the reporters and cameras are. In the television era, reporters are influenced by the existence of good pictures and this, in turn, determines how much time is assigned to a story.

Second, journalism striving for objectivity relies on an unrealistic view of the separation of facts and values. For this reason, fairness usually becomes the goal as the best approximation of objectivity. Fairness, or balance, is evenhandedness in the treatment of a presented issue; it is the shedding of light on an issue from different angles. Its opposite, unfairness or bias, often means the misleading slanting of news; because it conflicts with the central journalistic duty to inform the public, bias should be avoided. Bias can be the result of faulty thinking (giving too much weight to eyewitness testimonies which are never purely objective; looking for confirmation and not falsification of a view; using unrepresentative samples; or committing the *post hoc ergo propter hoc* error), the way of presenting the material (presenting news out of context; using emotionally charged language; focusing on the negative; or using close-ups rather than medium-long shots), or the nature of the organization (insufficient time allotted for news presentation).

Although perfect impartiality is unattainable, making explicit journalists' own prejudices and assumptions can bring them closer to the truth. If journalism did not make seeking the truth untainted by subjectivity its goal, it would severely undermine its credibility and usefulness. By assuming, after the postmodernist thinkers, that truth is just a social construct, the borderline between fact and fiction is blurred; whereby no news could be trusted.

It is often stressed that in the interest of objectivity, journalists should avoid, if possible, conflicts of interest and commitment. Not infrequently, story sources want to receive positive coverage, and attempt to influence journalists through gifts, free tickets to events to be reviewed, free travel, etc. Some codes of ethics explicitly prohibit receiving such favors, but for financially struggling organizations such opportunities can be hard to pass over. Also, there may be a conflict of interest when journalists are members of, or sit on the boards of, organizations they cover. Sometimes even the appearance of such a conflict can be as damaging to the credibility of the media outlets as the conflict itself. However, if a conflict cannot be avoided, it should be acknowledged to the public.

The position of truth in the context of advertising is different to that in journalism. The role of advertising is to persuade people to do something: in commercial advertising, to buy a product, or, in political advertising, to accept a person or views. This requires the stressing of the strong points of the product and omitting the weaknesses. Because the only goal of advertisement is to promote a product or idea, it is one sided and should not be expected to be balanced and complete. The public must be aware of this and treat the advertising accordingly: *caveat emptor*. Also, advertising is a persuasive, not rational, communication. As such, it wants to catch potential client's attention, and is not primarily concerned with the product's quality. But this does not mean that truth is irrelevant for advertising. Advertising should not be deceptive by suggesting nonexisting qualities of a product; it should avoid deceptive means, such as using staged testimonials (see *Deceptive Methods: Ethical Aspects*).

This element is particularly important in political advertising. Politicians must win public support to win an election, so they must have legitimacy obtained by favorable media visibility which stresses the positive and downplays the negative aspects of the politician's views and activity. This often means diverting the public's attention from important political and economic issues to what plays well in the media, particularly on television: polished demeanor, good looks, witty ripostes. It is claimed that the voters are more interested in trustworthiness of politicians than in their particular promises. Realization of this problem leads the politicians and their public relations advisors to project the best image of these politicians through the media, to shape in public opinion the image of their great personality and character. For this reason, political public relations advisors and consultants become increasingly important on the political scene.

2. *Freedom and Responsibility*

The freedom to publish opinions is defended with the argument that it is necessary for democracy, and that the discovery of the truth requires a free marketplace of ideas. However, others say, the powerful dominate

the media, and consequently the competition of ideas is not genuine because the access to the media is unequal. Also, an argument is used that free speech fosters individual autonomy and development, but, on the other hand, it may encourage hate speech and obscenity.

Unqualified free speech, even in democratic society, is unattainable and it is legally restrained, e.g., in the interest of national security and public safety. There remains a problem of defining the situation when national security is endangered. It is sometimes maintained that the United States lost the war in Vietnam because the media coverage largely undermined the image of the military. Hence, media coverage of the British military operations in the Falklands (1982) and the United States operation in Iraq (1991) was restricted much more severely than necessary.

Most media practitioners realize that—notwithstanding legal issues—free speech must be qualified with responsibility. Journalists in a democracy have a responsibility to foster pluralism and diversity of political debate, and to refrain from perpetuating myths about different social groups. Also, the media must take into account the consequences of their actions. For instance, reporters can sometimes anticipate that their approach to news gathering may induce antisocial behavior: the camera can turn a demonstration into a riot, and a live coverage of a hijacking, which gives the hijackers free publicity, makes hard the release of the captives. When their stories bring some harm, journalists defend themselves by saying that they just report a story. But news organizations are very willing to take credit for positive consequences. News is created by journalists, so they need to be sensitive to possible influences their news may have. They need to take into account who may be harmed by their actions. This is particularly important in the context of citizens' privacy (see also *Privacy of Individuals in Social Research: Confidentiality*).

2.1 Privacy

Journalists justify breaching someone's privacy by saying that the public has the right to know. This is an important argument, because it implies that the media are obligated to publicize some reports, whereas freedom of the press merely permits such publicizing. However, the argument leaves unspecified what it is that the public has the right to know. Although the media, journalism in particular, ought to act in the public interest, it does not mean that they are obligated to fully satisfy the public's interest. Public interest is not tantamount to mere curiosity. The critics indicate that the media much too often publicize information to satisfy the public's curiosity, but refrain from publishing revelations about the media themselves; journalists routinely avoid stories that put other journalists in a negative light (Hausman 1992).

The problem of privacy violation is particularly acute in the case of public figures. It is often said that public figures relinquish their right to privacy by the very fact that they become public figures; the loss of privacy is a price for being famous and thereby for high social status, wealth, power, etc. For this reason, all details of private lives of public figures can be published. It is claimed that in the case of politicians such details are important, because their personal qualities can interfere with their public duties and affect their decision making, especially under pressure. Also, the character of the people associated with a politician influence the politician's priorities and decisions. People elect a whole person, not only a public side; consequently, they should know about the private side as well. Hence, notwithstanding the motives of the media, they are acting on the public's behalf as a watchdog of the political arena, and, therefore, they have the right and responsibility to know the facts concerning people who shape political and economic policies of the country. In particular, the media should expose the hypocrisy of politicians who publicize goals and policies that are at variance with their own life. However, others say, in line with La Rochefoucauld's maxim, that hypocrisy is the homage which vice renders to virtue, and, hence, if positive policies are enacted in spite of violation of them by the enacters, so much the better for the policies. For this reason, a hasty exposure of politicians' vices can be counterproductive because it may lead to the demise of the politicians and, consequently, to the demise of their positive and desirable policies. Besides, an invasion of privacy discourages many people from entering an election race, which is detrimental to the democratic process. It is also said that to maintain their mental heath, public persons should have time off from the sight of the public.

Privacy is violated by undercover investigation and the use of long-range cameras, electronic eavesdropping devices, or surreptitious taping. The use of such means is often justified as indispensable in the investigation of crime and corruption, public health dangers, and safety negligence because it may be the only available means to unearth the story, and it may contribute to greater objectivity of the story. The critics maintain that the use of such means may lead to a general erosion of trust, and discussion of the news obtained by questionable methods often focuses on these methods rather than on the news itself.

2.2 Self-regulatory Mechanisms

The media recognize that some self-regulation is necessary to avoid an erosion of confidence and public demand for governmental regulation. There are several mechanisms used to develop and maintain responsibility of the media. Codes of ethics mark the stage of maturity of a profession, provide a framework for self-regulation, and can help to define and direct

the work of journalists; however, they are usually vague and often leave open the problem of implementing the enforcement mechanism. Another mechanism is internal criticism by a neutral arbitrator within the organization to investigate and to respond to complaints; an ombudsman can increase an organization's credibility, because the existence of an in-house critic signals to the public its sensitivity to public opinion and its willingness to modify the organization's practices. Also, news councils are formed, with variable success, to debate complaints against the press (e.g., British Press Council since 1953, National News Council in the USA in 1973–1984). Finally, education offered in classes on journalism ethics, conferences, meetings, and seminars explore issues of media ethics.

3. *Quality of Media Content*

The media frequently argue that they show and publish what the audience wants, which is determined democratically by conducting polls, and that criticism of the media for low quality programs is elitist. This argument is rebutted by the claim that the media respond only to the immediate interests of the public, and what goes beyond these interests is inevitably disregarded. Also, this way of determining interests has a tendency to confine them to the lowest tastes. The pursuit of higher ratings leads to trivialization and sensationalism of journalism and entertainment. The proliferation of tabloid journalism is the result of a conviction that the media only reflect the public's tastes; interests in sensational and bizarre aspects of our culture is natural, and no harm results from satisfying them. However, it is claimed that the constant exposure to tabloid journalism diverts public attention from important social and political problems, and gives credibility to sensational aspects of human existence. Tabloid journalism has become mainstream because sensationalism is one of the most important elements shaping the news, particularly in television, which, by definition, gives rise to distortion. Visual media call attention to conflict, magnitude, the unusual, the sensational, and the spectacular. Television relies on brevity and swift transitions. Visual events (fires, crashes) gain prominent coverage on television; an explanation of a tax reform receives only a short sound bite. The public will eventually come to think that the sensational aspects are the norm.

The reliance on sensationalism is apparent in entertainment. For example, talk shows live on the voyeurism of the public and exhibitionism of the guests. The primary goal of exposing the unusual behavior or appearance of talk show guests is to entertain the public, not to help the guests. The shows break taboos, exhibit deviant behavior, and make it appear ordinary.

Stressing the sensational manifests itself in the saturation of entertainment with violence and erotic content. Many media critics express a concern that watching violence stimulates violent behavior generally, and copying the watched behavior in particular, people who watch violence regularly become desensitized to its horror and may even develop a positive attitude to it. However, it is possible that pictures of violence increase the level of revulsion concerning its consequence, that it inhibits aggressive behavior, that is, has a cathartic effect—Aristotle claimed that a tragedy with its depiction of 'serious actions' has a cathartic effect (*Poetics* vi).

The effect on children of watching television is probably the most researched subject today in media influence on people. Results of many research studies indicate that television, especially depiction of violence, has a profound effect on children's behavior. This is also true for computer games which have become increasingly saturated with violence. Also, action movies engage viewers' emotions and reduce them to primal instincts, leaving out intellect (Sloterdijk 1993). However, some say, even granted that it may be difficult to establish empirically the existence of a causal link between viewing and future behavior, many actions of the media are predicated on the existence of such a link. For example, if there were no influence of the media on the public, no organization would invest any money for advertising. Therefore, if noncriminal behavior can be influenced, it is difficult to assume that the media have no impact on criminal behavior. Moreover, constant repetition is both a tool of education for acquiring and holding information, and of propaganda for putting across a political message. It would be remarkable if constant viewing of violence in movies or games did not have a similar effect. The media often use rating systems to indicate the content of films and shows, but, some contend, this system is often used as an allurement rather than a warning, and the movie producers are tempted to include gratuitous violence, pornography, and profanity in their products to make them more appealing.

The media create popular culture, and in this culture they gravitate toward what is most exciting. Therefore, popularity seems to be a crucial factor in deciding the content of the media. But if popularity is a deciding factor, then there is no room for creativity and acceptance of new ideas. There is little room for real choices because the media become the subject of a form of Grasham's law which says that poor quality entertainment drives out high quality entertainment, news, and educational programming (Fink 1988).

See also: Advertising, Psychology of; Freedom of the Press; Mass Media: Introduction and Schools of Thought; Media and Child Development; Media and History: Cultural Concerns; Media Effects; Media Effects on Children; Media Imperialism; Political Advertising; Science and the Media; Violence and Media

Bibliography

Alix F X 1997 *Une Éthique pour l'Information: de Gutenberg à Internet*. L'Harmattan, Paris
Christians C G, Fackler M, Rotzoll K B, McKee K B (eds.) 2001 *Media Ethics and Moral Reasoning*, 6th edn. Longman, White Plains, NY
Cohen E D, Elliott D 1997 *Journalism Ethics: A Reference Handbook*. ABC-Clio, Santa Barbara, CA
Day L A 1999 *Ethics in Media Communications: Cases and Controversies*, 3rd edn. Wadsworth, Belmont, CA
Fink C C 1988 *Media Ethics: In the Newsroom and Beyond*. McGraw-Hill, New York
Hausman C 1992 *Crisis of Conscience: Perspectives on Journalism Ethics*. HarperCollins, New York
Kieran M (ed.) 1998 *Media Ethics*. Routledge, London
Limburg V E 1994 *Electronic Media Ethics*. Focal, Boston, MA
Pigeat H 1997 *Médias et Déontologie: Règles du Jeu ou Jeu sans Règles*. PUF, Paris
Sloterdijk P 1993 *Medien-Zeit: Drei gegenwartsdiagnostische Versuche*. Cantz, Stuttgart, Germany
Wiegerling K 1998 *Medienethik*. Metzler, Stuttgart, Germany

A. Drozdek

Media Events

Media events are large, important events that interrupt the normal media schedule and audience activities, inviting special coverage and attention. State funerals, royal weddings, diplomatic visits, Senate hearings, and the Olympic games are representative examples. In these cases an agency independent of the media sponsors an event with a claim of historical importance. The media validate that claim by interrupting their normal schedules for live coverage; the audience delivers the final validation by granting the media event their special attention. When successful such events have an integrating effect on societies, while their status as an apparently new form of dispersed ceremony raises serious questions.

The ideal type of media event can be defined according to form, substance, and outcome (adapted from Dayan and Katz 1992). Formally, media events are interruptions of normal broadcasting routines for the live presentation of preplanned events, organized by agencies other than the media. Substantively, media events are presented with reverence and ceremony and are declared historic; they usually aim at reconciliation and emphasize the voluntary actions of heroic individuals (see Carey 1998 on conflict media events). The ideal type outcomes of media events are that they enforce a norm of viewing, exciting large audiences, who celebrate and renew loyalties, integrating their societies. Media events, then, fall within the category of secular and political ritual in modern life (Rothenbuhler 1998) (see *Ritual*).

Outside the scholarly literature there are two related uses of the term media event. Among the professionals who coordinate events for media coverage, a media event is anything like a press conference, public appearance, or photo opportunity arranged for the purposes of media coverage. The corresponding pejorative use of the term casts a media event as a 'pseudo-event' (Boorstin 1961), as something without stature independent of media coverage. There are important grounds for criticism of the media in these regards that will be addressed below. These criticisms should not distract, however, from the evident social functions of genuinely special media events.

The live television broadcast of the funeral of President John F. Kennedy and the surrounding events attracted an extraordinary audience. For extended periods of time it appears a majority of all Americans were watching and hearing the same events at the same time on television. These were ceremonial events that followed a social crisis and preceded a return to normality. Research conducted at the time showed that huge numbers of people were attending to the television coverage, reporting that they thought it important, and rearranging their schedules to watch it, often in groups of family and friends. Many people also reported symptoms of stress, anxiety, and uncertainty in the days following the assassination, which receded to normal levels in the days and weeks following the televised funeral (Greenberg and Parker 1965).

Every four years the broadcasting media interrupt their normal schedules for special coverage of the Olympic games, a large, regularly reoccurring event, organized by agencies officially independent of the media. Much of the coverage is live and both broadcasters and audiences betray an ethical preference, in criticizing edited and tape-delayed coverage and packaged programming segments. Media agents emphasize the importance of the games within and beyond sports; they make frequent historical comparisons and provide coverage of preceding and concurrent social, political, and cultural events. The games attract an audience that is unusually large, heterogeneous, and attentive. They are more likely than normal television audience members to be in groups, to be eating and drinking, to have planned their viewing ahead of time, to have rearranged their schedule to accommodate the viewing, and to have visitors from outside the home (e.g., Rothenbuhler 1988).

Though the televised Kennedy funeral and the Olympic games might be thought to fall in different categories of seriousness, they have many indicators of seriousness in common. As in such diverse examples as Anwar Sadat's first trip to Jerusalem, the funeral of Lord Mountbatten, the early trips of Pope John Paul II, or selected Senate hearings, the television coverage itself takes on a ceremonial style and the audience is recruited into a sense of participation at a distance. The whole of the event, the coverage, and the audience

activity constitutes a new event, a media event that is constructed as and functions as a ritual in mediated communication.

If the media event constitutes a ritual, and it is not a false one, that has major implications for social theory more broadly. The face-to-face context is conventionally presumed to be necessary for ritual and ceremony; such communicative products as television or newspaper coverage, commemorative art, or memoirs are considered byproducts. The audience on the scene usually is considered primary, and may participate in the ceremony as congregation or witness. Media audiences usually would be considered secondary spectators. In the media event, though, the happenings in the face-to-face context are only a part of the event, no more important than the coverage and the activities of the at-home audience. The audience on the scene of a state funeral media event, for example, plays a role on the stage of the event, while the audience at home becomes the official audience, or even a kind of dispersed congregation.

The media event is comparable to the broadcast distribution of religious ritual, both in its apparent functions for audiences and in the questions it raises for analysts. Broadcast coverage of a ritual would not appear to be the same thing as the ritual, anymore than a painting of a scene could be mistaken for the scene. Yet, observation consistently shows that for the faithful the broadcast coverage can function as participation in the ritual and have many of the same effects that physical participation would have. Who is to tell the faithful that television viewing does not count, if their belief and behavior confirms that it is special in their experience, and if religious leaders encourage such behavior and confirm their feelings? While the Pope himself has on occasion blessed the television audience, research shows that both commercial entertainment dramas devoted to religious subjects and apparently exploitative, money-raising television preachers also inspire genuine faith and devotional activity in some of their audience. Much of the television audience of Lady Diana's funeral also, like the audience for the Kennedy funeral 35 years earlier, acted as if and felt themselves to be participating in the mourning. Of course the audience at home does not get wet if it rains, but that may not matter for meanings and effects that can work through mediated communication. The implication is that the media event is a new form for the ritual and ceremony that Durkheim (1912/1965) argued provides the quasireligious foundations of society.

Media events can invite critical, if not cynical, responses as opportunities for duplicity and bad faith abound. The intense audience attention they inspire is highly desirable to the media, their commercial or governmental sponsors, and the organizers of events. The ceremonial attitude is vulnerable to manipulation for political or commercial gain. The quasi-religious aspects can appear as pre-modern residuals, as failures of rationality. Any given media event may deserve its critics, but when criticism is an intellectual habit not informed by the specifics of the case, then it is an instance of the larger category of distrust of mass communication. That twentieth century attitude can be traced back through a history of distrust of the image, all the way to Plato. But there are substantive grounds for concern as well.

Inevitably the work routines, technical requirements, and institutional values of media institutions come to bear on would-be media events. The most attractive, dynamic, or compelling parts of a ceremony or event may be scheduled for the greater convenience of the media or the easier access of large audiences. Scheduling of concurrent or succeeding portions of the event can be arranged to make it easy for television crews to move from one to the other, and to minimize the chances of either long periods of waiting or being caught by surprise. The scene itself can be accommodated to television cameras, with greater emphasis on a fixed central stage and activities that are highly visual, colorful, featuring dramatic movement, and adaptable to television scheduling. Television coverage will bring some of its own habits to bear as well, including close-up shots of emotional reactions and the reduction of complex events to simple narratives of personal tragedy, redemption, and triumph. As the work routines and values of the media come to bear on events, they alter those events and shift the balance of power among the sponsoring agencies, media, and audiences—and some of what is unique and valuable about media events depends on that balance of power.

The ceremonial aspect of media events is quite distinct from what is normally expected of news coverage. There is reason for concern that that aspect could be misused, and that the attitudes and behaviors appropriate to ritual and ceremony might spread to other domains where they would not be appropriate. Becker (1995), Chaney (1983, 1986), Liebes (1998), and others have pointed to dangers of ritualizing media coverage. Politicians will prefer ceremonial treatment as it implies their actions are that much more important, while giving journalists and the public reason to regard them reverentially rather than suspiciously. From the media point of view, interrupting normal programming for special coverage can give the audience special reason to watch, it can enhance the status of the journalists and their channel, and the excitement can be its own reward. There are pressures, then, to ritualize even small events such as the opening of a press conference. Simultaneously, we can expect media organizations to be on their guard for any event large enough to warrant special coverage and the claim to being a participant in historic events. Such tendencies can short-circuit rational consideration and democratic process, and diffuse the resource of audience attention.

In sum, media events represent an important, new social form of dispersed ceremony. It is a form that

raises important questions. On the one hand are questions about the power of communication and the necessity of the face-to-face context. On the other hand are the host of questions arising from the incongruity of serious, quasireligious social functions being served by media normally devoted to entertainment and profit. It is a form that can be abused or inappropriately imitated by politicians, government, the media, and other institutional interests.

See also: Entertainment; Media Effects; News: General; Ritual; Rumors and Urban Legends; Television: History

Bibliography

Becker K 1995 Media and the ritual process. *Media, Culture, and Society* **17**: 629–46

Boorstin D-J 1961 *The Image: A Guide to Pseudo-events in America*. Harper & Row, New York

Carey J W 1998 Political rituals on television: Episodes in the history of shame, degradation and excommunication. In: Liebes T, Curran J (eds.) *Media, Ritual, and Identity*. Routledge, New York

Chaney D 1983 A symbolic mirror of ourselves: Civic ritual in mass society. *Media, Culture, and Society* **5**: 119–35

Chaney D 1986 The symbolic form of ritual in mass communication. In: Golding P, Murdock G, Schlesinger P (eds.) *Communicating Politics: Mass Communications and the Political Process*. Holmes & Meier, New York

Dayan D, Katz E 1992 *Media Events: The Live Broadcasting of History*. Harvard University Press, Cambridge, MA

Durkheim É 1912 *The Elementary Forms of the Religious Life* [Trans. Swain J W 1965]. Free Press, New York

Greenberg B S, Parker E B (eds.) 1965 *The Kennedy Assassination and the American Public: Social Communication in Crisis*. Stanford University Press, Stanford, CA

Liebes T 1998 Television's disaster marathons: A danger for democratic processes? In: Liebes T, Curran J (eds.) *Media, Ritual, and Identity*. Routledge, New York

Rothenbuhler E W 1988 The living room celebration of the Olympic Games. *Journal of Communication* **38**(3): 61–81

Rothenbuhler E W 1998 *Ritual Communication: From Everyday Conversation to Mediated Ceremony*. Sage, Thousand Oaks, CA

E. W. Rothenbuhler

Media Imperialism

'Media imperialism' has been one of the most influential models in international communication. It has given rise to a significant body of empirical research, considerable theoretical debate, as well as support for international policy interventions about global communications imbalance. However it is increasingly subject to critique, particularly as a reflection of a particular moment of international media history that is now passing.

1. The Broader Paradigm of Cultural Imperialism

Media imperialism needs to be seen as a subset of the earlier and broader paradigm of 'cultural imperialism' that is most closely associated with Herbert Schiller (1969, 1976). This critical Marxisant paradigm itself developed out of broader critiques of the triumphalist paradigm of 'modernization' propounded in the late 1950s and the 1960s predominantly by US theorists. This 'dominant' model proposed a single global process of modernization through a unilinear diffusion of Western technologies, social institutions, modes of living, and value systems to the eponymous 'Third World' (Sreberny-Mohammadi 1996). Critical views, especially those advanced by dependency theorists such as Gunder-Frank, saw the process instead as the spread of inegalitarian capitalism from a few core advanced industrial countries to the periphery, the South. In this process, the West was supplied with raw materials and cheap labor as well as with markets for their manufactured goods. The spread of Western influence was seen as establishing an ideological bulwark against the much-feared spread of communism.

Cultural imperialism, Schiller argued, 'best describes the sum of the processes by which a society is brought into the modern world system and how its dominating stratum is attracted, pressured, forced, and sometimes bribed into shaping social institutions to correspond to, or even promote, the values and structures of the dominating center of the sytem' (1976, p. 9). For Schiller, the carriers of Western cultural influence included communication technologies themselves, which were not value-neutral instruments but imbued with capitalist values; language (evident as English takes over as the global linguistic currency from the varieties of colonial linguistic heritage); business practices; the genres as well as content of soap opera, blockbuster film, popular music, etc. Schiller's view was US-centric, and he saw media as central elements in the global expansion of capitalism centered on the US, fueled by advertising and consumerism. In the context of the Cold War, US policy makers' promotion of the 'free flow of information'—with no trade or legal barriers to the production or movement of mediated cultural products—undoubtedly helped the spread of US hegemony in ways that were consonant with, if not totally driven by, US foreign policy.

2. Media Imperialism

Media imperialism, a more focused subset of this model, was defined as 'the process whereby the ownership, structure, distribution or content of the media in any one country are singly or together subject to substantial external pressures from the media interests of any other country or countries without proportionate reciprocation of influence by the

country so affected' (Boyd-Barrett 1977, p. 117). It differentiated between the shape of the communications vehicle, a set of industrial arrangements, a body of values about best practice, and media content, although it is the latter claim that received most research attention.

It highlighted both the issues around 'cultural invasion' as well as a more general imbalance of power resources. In essence, the model argued that the world patterns of communication flow, both in density and direction, mirrored the system of domination in the international economic and political order, the control by the West over the rest.

3. *Empirical Research*

Part of the strength of the paradigm of media/cultural imperialism was its compelling clarity, which gripped many a critical academic imagination, and which suggested many fertile avenues of empirical research. It precipitated numerous studies that sought to examine the unidirectional flow of mediated product from the West to the South. The 'one-way street' of television traffic was mapped by Nordenstreng and Varis (1974); film production and distribution was studied by Guback and Varis (1982); international news flow by the IAMCR/UNESCO study (Sreberny-Mohammadi et al. 1985). Scholars investigated the reach and dominant role of Western news agencies (Boyd-Barrett 1980), supplemented latterly by study of the televisual news agencies (Paterson 1998). Others showed that Third World media organizations were modeled on those of the earlier mother empires (Katz and Wedell 1977). More recent studies detail the ongoing transnational processes of conglomeratization and the increasing vertical and horizontal linkages of media corporations (Herman and McChesney 1997).

Cultural imperialism and media imperialism were among the first models to take the global dynamics of media production and diffusion seriously. These paradigms posed important questions about cultural homogenization, about the diffusion of values such as consumerism and individualism, and about the possible impacts of Western media product on indigenous cultures in the South (Hamelink 1983).

4. *Political/Policy Struggles*

These models also provided the conceptual and empirical framework for the NWICO (New World Information and Communication Order) debates which preoccupied UNESCO and the nonaligned movement through the 1970s until the USA, the UK, and Singapore walked out of UNESCO in 1984–5 (see Vincent and Galtung 1992). The MacBride Commission gathered input from over 100 international experts, and in its final report, *Many Voices, One World* (UNESCO 1980) challenged as inadequate the prevailing idea of 'free flow of information' since it functioned to privilege Western hegemony and cultural diffusion and to turn the South into mere consumers of Western news values, entertainment, and advertising images. Instead the report suggested the important qualification of a 'free and balanced' flow, which was adopted as UNESCO policy, although it defaulted from precise suggestions as to how, for example, international news coverage could be appropriately 'balanced.'

It is significant that UNESCO funded many of the early landmarks in critical international communication research yet from the mid-1980s, its focus and rhetoric has shifted toward a more neo-liberal position and a greater concern with issues around freedom and democratization than around balance/flow.

The debates about global media inequality and concern about the logics of conglomeratization and the threats to local cultures continue even in 2001 at the forums of the MacBride Round Tables; the demand for a Peoples Communication Charter and in the activities of the Cultural Environment Movement (see Voices 21 1999). However, while the debates in the 1970s involved Third World policy makers, the critical voices are now mainly Western academics, while Southern politicians have become more involved in nitty-gritty decisions around deregulation, digitalization, and media convergence.

5. *Critiques of Media Imperialism*

5.1 *Arguments From and About Historical Process and Effects*

'Media imperialism' was a problematic argument both theoretically and empirically from the beginning, and by year 2001 it appears increasingly ossified. An argument that might have had some validity for some countries at a certain period of time became the dominant critical paradigm, operating as a form of political correctness by which critics were seen as apologists for the USA and its demand for a 'free flow international regime for trade in cultural products' (Sinclair et al. 1995, p. 6). By the 1980s when the global television and other media industries really changed shape, the 'critical' paradigm had become the 'orthodoxy' (Sinclair et al. 1995, p. 5).

Historically it is doubtless the case in most, especially technologically based, endeavors that the prime mover sets the model: in cars, in high-rise architecture, in media. Broadcasting did not develop with world domination in mind, even if some of its spread has been consonant with Western foreign policy interests. However, the model of state-controlled broadcasting which was adopted in much of

the South hardly supported the 'free flow of information' of Western interests. It is also the case that subsequent media players 'catch up,' begin to produce their own content, and rework genres to fit their cultural habits. Already in the 1970s, Tunstall (1977) was arguing that Latin America's high level of television imports in the 1960s was a transitional phase while nationally based television systems were being established.

Empirically, it is mainly from the 1990s with the expansion of global satellite systems and development of infrastructure that much of the population of the rural South has actually been introduced to television. Thus, oddly enough, media could not have had the strong effects imputed by the model (nor by the modernization theorists like Lerner 1958) because the televisual media actually didn't reach most people in the South until the satellite expansion of the late 1980s.

The crude adoption of 'media imperialism' arguments promoted a truncated historical analysis and tended to disregard earlier historical processes. It is probably the case that the cultural legacies of imperialism—the spread of Christianity, training in and use of European languages, establishment of systems of schooling and higher education, spread of administrative cultures—have had more enduring and far-reaching effect as carriers of Eurocentric modernization to the South than have the subsequent arrival of electronic media (Sreberny-Mohammadi 1997, Bitterlli 1989).

Additionally, many other industries, connected to but not reducible to the media industries, produce and market symbolic goods (fashion, foodstuffs, architecture and interior design, consumer durables) while the global tourism and travel industries literally transport millions of people a year. A focus on only one element of the contemporary production of culture cannot be read for the whole. Capitalism is also not organized and instituted through media, albeit that media content diffuses images of the goods, lifestyles, attitudes that are increasingly commodified and available for consumption in the capitalist market. In that sense, the media can support processes of economic transformation and the profound shift from culture as 'practice' to culture as 'commodity,' and here reconnect to the bigger issues around modernization and development that tend to become submerged in the focus on media structures alone.

5.2 *Arguments About Production and Flow*

'Media imperialism' was a model developed at a time of comparative media scarcity and newly established broadcasting structures in the South. By the year 2001 there are many significant cultural industries in the South: Globo in Brazil, and Televisa in Mexico produce telenovelas; a huge multimedia complex near Cairo supports the production of Islamic soap operas, which Turkey also produces. And if the focus shifts away from television alone to include other cultural products, the diversity increases: Bollywood, the familiar label for the Indian film industry, is the Eastern challenge to Hollywood in the sheer number of film titles produced yearly, with the Asian diaspora constituting a sizeable audience. The Iranian and Chinese film industries are gaining global recognition and audiences. The marketing of 'world music' has helped the diffusion of Algerian, Senegalese, Cuban, and Brazilian contemporary music. The Indian ZeeTV is a powerful independent newscaster while Qatar's Al-Jazeera is revolutionizing factual programming in the Arab World.

Thus the model's metaphors of the 'one-way flow' down the 'one-way street' from the West to the rest (Nordenstreng and Varis 1974) are challenged by more diverse production, and more complex images of the 'patchwork quilt' (Tracey 1988). Southern media exports into the West have been dubbed 'reverse imperialism' while many of the new media mogul empires are not based in the West and Asian corporations and entrepreneurs (SONY, Masushita) have bought up segments of Western media.

5.3 *Arguments About Audiences and Effects*

'Media imperialism' also implied direct unmediated effects of television programming on audiences, turning them into Americanized consumers. It tended to ignore the particularities of local culture and their meaning systems as well as other processes of modernization (shifts in material production, migration and tourism, political institutionalization and democratization) that often operate alongside media development.

New approaches to the 'active audience' within media studies have forced a rethinking of international effects also. As with domestic viewing practices, analyses of the negotiation of meanings around US-produced programs such as Dallas by 'foreign' audiences show different audience viewing strategies including the imposition of local interpretive frames on 'foreign' media product (Liebes and Katz 1990). There is also strong evidence that audiences prefer locally made mediated culture where and when that becomes available (Sepstrup and Goonasekura 1994). The original logic of argument also depended on a rather ossified notion of 'national' culture, whereas even US product changes, so that the *Cosby Show* has become a popular, if rather inaccurate, symbol of changing race relations in the USA.

All this means that a single global product doesn't always work across cultural divides. While a film such as *Titanic* may have global reach, falling into that magical, much-desired and heavily marketed category of the 'global popular,' most media production is now tailored for local, that is, linguistically bounded, sometimes still nationally-bounded, cultural markets. The mantra is to 'think global, produce local.'

Regional linguistic-cultural markets are growing in economic significance as the 'cultural discount' (Hoskins and McFadyen 1991) of local cultural preference kicks in. While this can produce regional exchanges of news and television programming, it can also produce localized versions of the old paradigm, as in the concern about India's cultural dominance over the subcontinent.

5.4 Arguments About Genres

Discussion about the nature and definition of media genres in relation to the South has perhaps not been fully engaged. The advent of an externally originated cultural form into a society has often caused controversy in, for example, fear about the impact of free verse on traditions of formal poetry or about the impact of the novel on non-Western writing cultures. Yet now Latin American magic realism and post-colonial family sagas crossing generations and national boundaries have transformed the very form of the novel. A debate once raged about whether Latin American telenovelas were a new form of television series, inflected with socially developmental themes, or whether they were simply soap operas with commercial promotional messages. Anthropologists have analyzed the rise of the Islamic soap opera (Abu-Lughod 1993) and it seems clear that Southern media texts are not 'simply' copies of Western texts, but neither are they utterly novel genres either. Textual give and take, multiple sources of inspiration and derivation in media content, needs to be taken seriously, and attention paid to how forms evolve over time, both within developed television industries like the USA and the UK as well as within 'younger' media systems. For example, Zee-TV is not the same as MTV, although it may also show music videos.

5.5 Arguments About Media Policy and Regulation

As the global media environment changes with growing convergence between broadcasting and telecommunications, growing digitalization and the spread of the Internet, so there has also been a move from notions of imperialism to ideas of cultural hegemony. Such constructs are less overtly spatial and usefully deconstruct a singular and unproblematic 'West.' There have been fierce intra-Western disputes, most obviously the French disagreements about the deregulated flow of US film within GATT talks, or the differences and competition between US and UK modes of news construction. Parts of South East Asia may now be more concerned about the dominance of Indian cultural product than US; and national law in one Western country encounters a different set of rules in another, as in the US-French dispute about the sale of Nazi ephemera through the Yahoo gateway on the Net.

5.6 Arguments About the State and Democratization

The concern of much of the debate with transnational media domination by the West and the dominance of the market often precluded serious analysis of national processes, particularly the complex class formations in the South and the role of states. Thus, for example in Latin America, repressive military juntas have used media as key institutions of authoritarian government (Waisbord 2000) while in rapidly modernizing South East Asian economies the ruling elites have aligned with global corporate interests, often against the needs of their own population. As popular movements for democratization and equality have grown inside such states, this has prompted greater analytic focus on the relations between transnational processes and national politics and media control, and how states and markets interact within different political and economic formations.

It is amongst the autocratic, sometimes religious, regimes of the Middle East and South East Asia that the argument about 'media imperialism' is still heard. The 'protection of indigenous culture' can be used as a weapon to prevent internal change and demands for democratization, and the 'outside' can be constructed solely as a predator involved in acts of rape against the feminized, vulnerable nation. Yet, in numerous Southern countries—Indonesia, Nigeria, Iran, Brazil—secular intellectuals, ethnic groups, and women are struggling for greater political openness and basic human rights as well as a media environment which facilitates the construction of an inclusive national public sphere. Sometimes the transnational space can act as one of solidarity and support for such struggles.

6. The Rise of 'Globalization'

Many of the early debates around cultural imperialism and media imperialism actually prefigure key strands of argument that were taken up by the rhetoric of globalization in the 1990s: the recognition of the collapse of space and time through communications technologies; the commodification of social life; the significance of cultural flows. But within the model of globalization is also a sense of collapse of the older, single-centered hegemon into a more disorganised post-Fordist set of global production processes (Lash and Urry 1994); a more decentered and unstable political and cultural environment; and a variety of resistances to Western modernity with the rise of primordial religious and ethnic loyalities. The interpenetration of the global and local actually produces multiple formations of modernity.

'Cultural imperialism' tended to see cultures as singular, somewhat ossified, nationally bounded entities: 'American' television impacting on 'Iran,' for example. There is increased awareness of the loosening

of the hyphen in the nation-state and thus the 'national culture' is better seen as a site of contest. These are not just issues for countries in the South, but for most countries around the world. In the UK, this is enacted politically with the devolution of Scotland and Wales, and in debates around immigration policy; and culturally by debates about the religious content of the millenial exhibition in the Dome, and about the representation of ethnic minorities in the media. National identity may be practiced daily through a variety of symbolic forms and organizational routines, but its actual content may be more fluid and disputed than ever.

Globalization highlights the manner in which pressures to modernize are being met with particular local orientations and practices. It can still be argued that Western values lie hidden behind a seemingly neutral 'universal' and abstract set of processes, but the precise way these are encountered, understood, and dealt with varies from country to country, even within the same region. Latin America presents considerable national variation as does the Arab world; hence, even an increased awareness of regionalism is insufficient. Among media academics, more particularism, more case studies, more nuance is required.

The real collapse of the three worlds of development has left a conceptual lacuna. It is hard to propose such totalizing models any longer. Rather, we need regional frames that take cultural specificities, local structures, political values, seriously. Globalization has exacerbated economic disparities among—but also within—countries and has not ended political repression. Contemporary 'flows' include massive movements of capital and of people, so others are no longer 'there' but 'here' and building transnational communities through various kinds of diasporic media.

Work on the media has also tended to be focused on a single medium, initially mainly television, now heavily focused on the Net; yet rather bizarre situations exist whereby entire newspapers are banned, for example in Tunisia, Egypt, and Iran, while Internet-based information circulates with impunity. Looking at only one element of a complex media environment can also lead to wrong conclusions.

A recent attempt to rework the 'media imperialism' model suggested the need to 'encompass neo-colonialisms of inter-ethnic, inter-cultural, inter-gender, inter-generational and inter-class relations' (Boyd-Barrett 1998, p. 167) but instead helps us to recognize that in the end the model is simply about power imbalances, which do not require the 'imperial' stamp to be of importance to examine. Indeed the 'imperialism' obscures the differentiated processes of economic inequality and complex power dynamics. Access to cultural expression often involves complex internal struggles for power as well as pushes from the outside. As Tomlinson (1991) has argued, cultural imperialism always consisted of many different discourses; the ongoing attempt to rewrap them into one through the trope of 'media imperialism' is an increasingly forlorn task. The world has changed and so must our language and our theoretical frames.

7. *Conclusion*

Perhaps the enduring cultural legacy of the USA lies in the development of a commercial model of television as a medium: using entertainment content to attract audiences who could be sold to advertisers (Sinclair et al. 1995). While this is still not a universal model and many states still try to control broadcasting, commercial pressures and availability of multiple forms of media distribution/reception such as satellite and cable mean the globalization of capitalism and the manufacture of mediated culture are in the ascendant. In that sense, the focus on US-centric media diffusion obscures a far wider concern around the generic mass production of culture that is now flourishing globally.

See also: Advertising: Effects; British Cultural Studies; Globalization and World Culture; Media Effects; Media Effects on Children; Media Ethics; Media, Uses of; Soap Opera/Telenovela; Television: History

Bibliography

Abu-Lughod L 1993 Finding a place for Islam: Egyptian television serials and the national interest. *Public Culture* **5**(3): 494–512

Bitterlli U 1989 *Cultures in Conflict: Encounters between European and Non-European Cultures 1492–1800*. Polity Press, Cambridge, UK

Boyd-Barrett O 1977 Media imperialism: Towards an international framework for the analysis of media systems. In: Curran J, Gurevitch M, Woolacott J (eds.) *Mass Communication and Society*. Edward Arnold, London, pp. 116–35

Boyd-Barrett O 1980 *The International News Agencies*. Constable, London

Boyd-Barrett O 1998 Media imperialism reformulated. In: Thussu D (ed.) *Electronic Empires*. Arnold, London, pp. 157–17

Boyd-Barrett O, Rantanen T 1998 *The Globalization of News*. Sage, London

Galtung J, Vincent R 1992 *Global Glasnost: Toward a New World Information and Communication Order?* Hampton Press, Cresskill, NJ

Guback T, Varis T 1982 Transnational communication and cultural industries. *Reports and Papers on Mass Communication*. UNESCO, Paris, No. 92

Hamelink C 1983 *Cultural Autonomy in Global Communications*. Longman, New York

Herman E, McChesney R 1997 *The Global Media*. Cassell, London

Hoskins C, McFadyen S 1991 The US competitive advantage in the global television market: Is it sustainable in the new broadcasting environment? *Canadian Journal of Communication* **16**(2): 207–24

Katz E, Wedell G 1977 *Broadcasting in the Third World*. Harvard University Press, Cambridge, MA

Lash S, Urry J 1994 *Economies of Signs and Space*. Sage, London
Leibes T, Katz E 1990 *The Export of Meaning*. Oxford University Press, Oxford, UK
Lerner D 1958 *The Passing of Traditional Society*. Free Press, New York
Nordenstreng K, Varis T 1974 Television traffic—a one-way street? *Reports and Papers on Mass Communication*, No. 70 UNESCO, Paris
Paterson C 1998 In: Boyd-Barrett O, Rantanen T (eds.) *The Globalization of News*. Sage, London
Peoples Communication Charter 1999 http://www.tbsjournal.com/Archives/Spring99/Documents/Congress/Charter_intro/Charter/charter.html
Schiller H 1969 *Mass Communications and American Empire*. A. M. Kelley, New York
Schiller H 1976 *Communication and Cultural Domination*. International Arts and Sciences Press, White Plains, NY
Sepstrup P, Goonasekura A 1994 TV transnationalization: Europe and Asia. *Reports and Papers on Mass Communication*. UNESCO, Paris
Sinclair J, Jacka E, Cunningham S 1995 *New Patterns in Global Television: Peripheral Vision*. Oxford University Press, New York
Sreberny-Mohammadi A 1996 The global and the local in international communication. In: Curran J, Gurevitch M (eds.) *Mass Media and Society*. Arnold, London, pp. 177–203
Sreberny-Mohammadi A 1997 The many cultural faces of imperialism. In: Golding P, Harris P (eds.) *Beyond Cultural Imperialism*. Sage, Thousand Oaks, CA, pp. 49–68
Sreberny-Mohammadi A, Nordenstreng K, Stevenson R, Ugboajah F 1985 Foreign news in the media: International reporting in 29 countries. *Reports and Papers in Mass Communication*. UNESCO, Paris, No. 90
Tomlinson J 1991 *Cultural Imperialism*, Johns Hopkins University Press, Baltimore, MD
Tracey M 1988 Popular culture and the economics of global television. *Intermedia* **16**: 2
Tunstall J 1977 *The Media are American*. Columbia University Press, New York
Voices 21 1999 http://www.tbsjournal.com/Archives/Spring99/Documents/Congress/congress.html
Waisbord S 2000 Media in South America: Between the rock of the state and the hard place of the market. In: Curran J, Park M (eds.) *De-Westernizing Media Studies*. Routledge, London
UNESCO 1980 *Many Voices, One World*. International Commission for the Study of Communication Problems. UNESCO, Paris

A. Sreberny

Media Literacy

The core concept of 'media literacy' is intelligent engagement, as both a user and a creator, with media and technology. In Canada, the Ontario Ministry of Education in 1989 determined that media literacy is defined as that which is

> concerned with helping students develop an informed and critical understanding of the nature of the mass media, the techniques used by them, and the impact of these techniques. More specifically, it is education that aims to increase students' understanding and enjoyment of how media work, how they produce meaning, how they are organized, and how they construct reality. Media literacy also aims to provide students with the ability to create media products. (cited in Tyner 1998, p. 119).

In the USA, where media literacy has generally not been part of the school curriculum, the Aspen Institute Leadership Conference on Media Literacy determined that

> Media literacy, the movement to expand notions of literacy to include the powerful post-print media that dominate our informational landscape, helps people understand, produce, and negotiate meanings in a culture made up of powerful images, words, and sounds. A media-literate person—everyone should have the opportunity to become one—can decode, evaluate, analyze, and produce both print and electronic media. (Aufderheide and Firestone 1993, p. 1).

Two other constructs are closely related to media literacy, sometimes used interchangeably with it, and prevalent in countries such as the UK, Australia, and Canada. 'Media education' is teaching *about* media, as distinguished from teaching *with* media. Ordinarily, media education emphasizes the acquisition both of cognitive knowledge about how media are produced and distributed and of analytic skills for interpreting and valuing media content. In contrast, 'media studies' ordinarily emphasizes hands-on experiences with media production. Both media education and media studies proponents intend to achieve media literacy goals through learning activities with young people.

Appropriation of the term 'literacy' for media literacy has been criticized. Some believe it inaccurately glorifies this newer literacy by connecting it with the older, culturally valued reading and writing of text (or print literacy, alphabetic literacy, or simply literacy). Others argue that being literate connotes achieving a state in which one is a passive user rather than a lifelong learner and enthusiastic media participant. Media literacy proponents contend that the concept is and should be related to print literacy and depicts an active, not passive user: The media-literate person is a capable recipient and creator of content, understanding sociopolitical context, and using codes and representational systems effectively to live responsibly in society and the world at large.

1. *Need for Media Literacy*

Media literacy, or media education, has been part of the scholarly literature since at least the 1950s. Then, as now, work has been stimulated by the realities of modern life. Media and technology are frequently used outside school and work. Much content is produced and used for entertainment more than information. Much is 'low-brow' popular culture. Much is visual or audiovisual rather than (or in addition to) text—the main representational system

taught in school. All is created and distributed within particular sociopolitical and economic systems; often, profit-making is a primary goal. People like much of what media and technology offer and much of what can be done with them. Very often people accept content uncritically, apparently without taking account of the contextual factors in production, distribution, and interpretation and, for younger people, without understanding how production processes can create seemingly realistic or accurate content. The confluence of these realities has motivated scholars, educators, and public intellectuals to conceptualize and advocate for media literacy.

Today, in most developed countries, homes are full of media and technology. For example, a recent survey (Roberts et al. 1999) of a large nationally representative sample of US homes with children 2–18 years of age found the following:

(a) 99 percent of the homes had at least one television, 97 percent had at least one VCR, 74 percent had cable or satellite, and 44 percent had premium cable;

(b) 97 percent had at least one radio, 94 percent had at least one tape player, and 90 percent had at least one CD player;

(c) 69 percent had at least one computer, 59 percent had at least one CD-ROM drive, and 45 percent had access to the Internet; and

(d) 70 percent had at least one videogame player.

On average, children in these same homes were spending somewhat more than six hours a day in their own personal nonschool uses of media and technology. Usage time was quite unevenly distributed:

(a) 42 percent with television and 13 percent more with other television-like media,

(b) 22 percent with audio media,

(c) 5 percent with computer technology,

(d) 5 percent with videogames, and

(e) 12 percent with print media.

Children's usage patterns changed somewhat with age, with respect to total time invested in, and distribution of usage time among, different media and technology. There were also some variations according to other demographic characteristics. Overall, however, these differences were small enough to be considered variations on a theme rather than different themes. The USA may be among the most media-saturated of cultures, but the general picture is not that different in other developed countries. Clearly, there is ample reason to argue that media literacy is needed in the USA and elsewhere.

Theoretically, media literacy could be a relevant concept and goal for every culture that has any representational system whatsoever. In fact, media literacy work ordinarily focuses on postprint media and technology and the content they convey. Except in the USA, media literacy is prominent in almost all developed countries and typically part of formal schooling. It is probably more common in countries in which media and technology purvey much popular culture and in those in which there is a marked presence of content originally produced elsewhere (most often, the USA).

2. *Elements of Media Literacy*

As an area of scholarship, advocacy, and action, media literacy is to a notable degree contested terrain. However, there are a few agreed-upon fundamental facts (paraphrased from Aufderheide and Firestone 1993):

(a) the content of media and technology is constructed;

(b) the content is produced within economic, social, political, historical, and esthetic contexts that have influenced the production;

(c) the interpretive meaning-making processes involved in content reception consist of an interaction among the user, the content, and the culture;

(d) media and technology have unique 'languages,' characteristics which typify various forms, genres, and symbol systems of communication; and

(e) media and technology representations play a role in people's understanding of social reality.

Drawing on these facts, in something of a caricature, four 'pure' orientations can be found under the media literacy umbrella. As real people, media literacy leaders are a mix of these types, while typically favoring one orientation over the others or self-consciously combining orientations.

One orientation celebrates the richness that media and technology can bring to humankind. The production techniques, genres, and forms of media become the focus, along with storytelling and artistry. Film studies, arts-based curricula, and hands-on production predominate in media education designed from this orientation; production is typically an important part of the curriculum. It is an orientation that has been in the field for some time and is likely to continue.

Another orientation reflects fear and dislike of media and technology. Content is believed to have powerful potential to affect the user. Most content (or most content that users select) is judged negatively. The very act of using the medium or technology may be addictive. In this orientation, media literacy goals become those of protecting or 'inoculating' users. Much media literacy work in the 1970s reflected this orientation, at least to some extent. Not so today.

Yet another orientation takes a more agnostic position about media and technology. Both good and bad uses and content exist; what is good and bad may vary among individuals and circumstances. Media literacy becomes self-awareness, good taste, good choices, interpretive tools and skills, an evaluative stance toward all content, and self-determination as both a user and creator of media and technology.

A final orientation reflects the recognition that cultural forces are at work with media and technology, privileging the views and welfare of some groups over other groups. Media literacy is participation in expanding the voices and visions available in media and technology and active resistance to any hegemonic influences. The dominant theoretical paradigm in this orientation is critical theory or cultural studies.

A critical component of media literacy is interpreting and creating messages conveyed in various 'languages.' Of these, the visual elements found in print, film, television, computers, and other media and technology have received the most attention. Visual literacy is needed to interpret, create, and control the effects of images—images alone and images combined with words; images one receives and images one creates. Particularly when images are realistic, as in much photography, film, and television, the naïve user may not realize how much these images too are constructed and must be interpreted. Visual literacy has a long history of its own, coming out of photography and film.

3. Relationship to Information Literacy

Although media literacy is broadly defined, work has tended to emphasize popular audiovisual media such as film and television. Recently, as use of computer-based technologies, the Internet, and the World Wide Web has increased, a related literacy, information literacy, has gained prominence. Whereas proponents of media literacy (including visual literacy and media education) come primarily from the fields of communications, media studies, the arts, education, and cultural studies, the most vigorous proponents of information literacy have been librarians and other information professionals. The American Library Association Presidential Committee on Information Literacy wrote: 'to be information-literate, a person must be able to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information' (American Library Association 1989, p.1)

At first glance, this short definition seems quite similar to a short definition for media literacy. Closer examination reveals notable distinctions. Information literacy assumes that one has an identified information need; media literacy does not. Media literacy issues apply when one is engaged with media and technology idly, for relaxation and for fun, as well as intentionally to learn something. Moreover, in contrast to information literacy, media literacy makes much of the fact that the content one encounters or creates is problematic and merits intelligent, critical engagement. Furthermore, that content utilizes a variety of representational systems that one needs to learn to interpret and utilize. Despite these differences, there are as well many similarities in the two modern-day literacies and several efforts to connect together.

4. Issues and Problems

The incomplete integration of various orientations into a cohesive conceptualization of media literacy is one issue or problem in the field. Beyond the basic facts presented earlier, conceptual differences abound. Differences can be productive, but many believe the field would benefit from a more coherent conceptualization, which would then support more cohesive advocacy and education efforts. Another problem arises from the fact that much media literacy work is action-oriented, developing media literacy goals and curricula, providing media education. With the exception of work in cultural studies and critical theory, there is too little scholarly and research-based understanding of what media literacy is and how it functions. Furthermore, too few media education programs have been seriously evaluated so that successful approaches can be replicated and less successful approaches improved. Finally, the field has done little to move from the user side of the equation to the medium side, to examine how media and technology themselves can promote media literacy via their system and/or content characteristics.

See also: Advertising: Effects; Audiences; Electronic Democracy; Entertainment; Literacy and Illiteracy, History of; Literacy Education; Media Effects; Media Effects on Children; Media Ethics; Media Imperialism; Media, Uses of; Printing as a Medium; Television: History

Bibliography

American Library Association 1989 *Presidential Committee on Information Literacy: Final Report*. American Library Association, Chicago

Aufderheide P, Firestone C 1993 *Media Literacy: A Report of the National Leadership Conference on Media Literacy*. The Aspen Institute, Queenstown, MD

Bazalgette C, Bevort E, Saviano J (eds.) 1992 *New Directions: Media Education Worldwide*. British Film Institute, London

Brunner C, Tally W 1999 *The New Media Literacy Handbook*. Anchor Books, Doubleday, New York

Buckingham D, Sefton-Green J 1994 *Cultural Studies Goes to School: Reading and Teaching Popular Media*. Taylor & Francis, London

Considine D M, Haley G E 1999 *Visual Messages: Integrating Imagery into Instruction. A Media Literacy Resource for Teachers*, 2nd edn. Libraries Unlimited, Englewood, CO

Duncan B, D'Ippolito J, McPherson C, Wilson C 1996 *Mass Media and Popular Culture*, version 2. Harcourt Brace, Toronto, Canada

Giroux H A, Simon R L 1989 *Popular Culture, Schooling and Everyday Life*. Ontario Institute for Studies in Education Press, Toronto, Canada

Kellner D 1995 *Media Culture*. Routledge, London

Kubey R (ed.) 1997 *Media Literacy in the Information Age*. Transaction, New Brunswick, NJ

Masterman L 1985 *Teaching the Media*. Comedia, London

Masterman L, Mariet F 1994 *Media Education in 1990s Europe: A Teachers' Guide*. Manhattan Publishing, Croton, NY

McLaren P, Hammer R, Sholle D, Reilly S 1995 *Rethinking Media Literacy: A Critical Pedagogy of Representation*. Peter Lang, New York
Messaris P 1994 *Visual 'Literacy': Image, Mind and Reality*. Westview Press, Boulder, CO
Moore D M, Dwyer F M (eds.) 1994 *Visual Literacy: A Spectrum of Visual Learning*. Educational Technology Publications, Englewood Cliffs, NJ
Quin R, Aparici R (eds.) 1996 Media education. *Continuum: The Australian Journal of Media & Culture* **9**(2) (whole issue)
Roberts D F, Foehr U G, Rideout V J, Brodie M 1999 *Kids & Media in The New Millenium*. The Henry J. Kaiser Family Foundation, Menlo Park, CA
Symposium: Media literacy 1998 *Journal of Communication* **48**(1): 5–120
Tyner K 1998 *Literacy in a Digital World: Teaching and Learning in the Age of Information*. Erlbaum, Mahwah, NJ

A. Dorr

Media, Uses of

Mass media (mainly TV, newspapers, magazines, radio, and now Internet) have been found to perform a variety of 'functions' for the audiences who use them (Wright 1974). Among the major controversies in the literature on media use (which is limited by most research reviewed below having been conducted in the USA) is the question of how much each medium's content informs rather than simply entertains its audience. Television is often assumed to serve primarily as an entertainment medium while print media are looked to for information.

The information vs. entertainment distinction may overlook more long-term uses and subtler or latent impacts, such as when media content is used as a stimulus for subsequent conversation or for consumer purchases. Such latter impacts also can affect longer-range cultural values and beliefs, such as fostering materialism or the belief that commercial products can solve personal problems.

The media may also perform interrelated information or entertainment 'functions,' as when viewers learn headline information about a news story on television, but then seek in-depth information from newspapers or magazines. Much literature has been devoted to studying the diffusion of stories of landmark events (Budd et al. 1966), such as the Kennedy assassination or the death of Princess Diana—or the audience response to events created for the media themselves (e.g., Roots, Kennedy–Nixon and other presidential debates).

There is also the more general question of how the various media complement, compete with, or displace each other to perform these functions, particularly in terms of user or audience time. As each new medium is introduced, it performs functions similar to those performed by existing media. This is referred to as 'functional equivalence' (Weiss 1969). Unless individuals devote more overall time to media use, time spent on older media must be displaced. Some researchers argue that overall media time tends to remain constant so that displacement is likely if new media become popular (Wober 1989). To the extent that a new medium like television performs the entertainment function in a way that is more compelling and attractive (than existing media like radio, magazines, movies, or print fiction), according to this 'functional equivalence' argument those earlier media will suffer losses of audience time and attention. Time displacement is becoming a central concern in research on Internet use, as researchers seek to gauge its potential to displace use of TV, radio, or newspapers.

1. Methods

Media use issues raise various questions about study methodology and measurement, since research results can vary depending on the methods used. For example, people's media use can be measured by electronic meters attached to their TVs (or computers), by 'beepers' attached to the person's clothing, by in-person or video monitoring, or by simply asking these people questions about their media use. Survey questions generally have the value of being more flexible, more global, and more economical than monitoring equipment, but there are problems of respondent reliability and validity. Long-range usage questions (e.g., 'How many hours of TV did you watch in a typical week?') may reveal more about a respondent's typical patterns, but they are limited by that person's memory or recall. Shorter-range questions (e.g., 'How many hours of TV did you watch yesterday?') appear to overcome these problems, but the day may not be typical of long-range usage. There is evidence that in the shorter range, complete time diaries of all the activities the person did on the prior day may be needed to capture the full extent of media use engaged in; an example of the power of the diary approach to reflect the full extent of media impact is given in Fig. 1.

In much the same way, different results are obtained when different types of questions are asked about the ways and purposes for which the media are used and how media usage has affected audiences. Thus different results are obtained when one asks respondents which news medium has informed them most compared to a more 'microbehavioral' approach, in which respondents are asked actual information questions and these answers are then correlated with the extent of each medium used (Davis and Robinson 1989)—or when respondents are asked to keep an information log of each new piece of information obtained and where/how it was obtained.

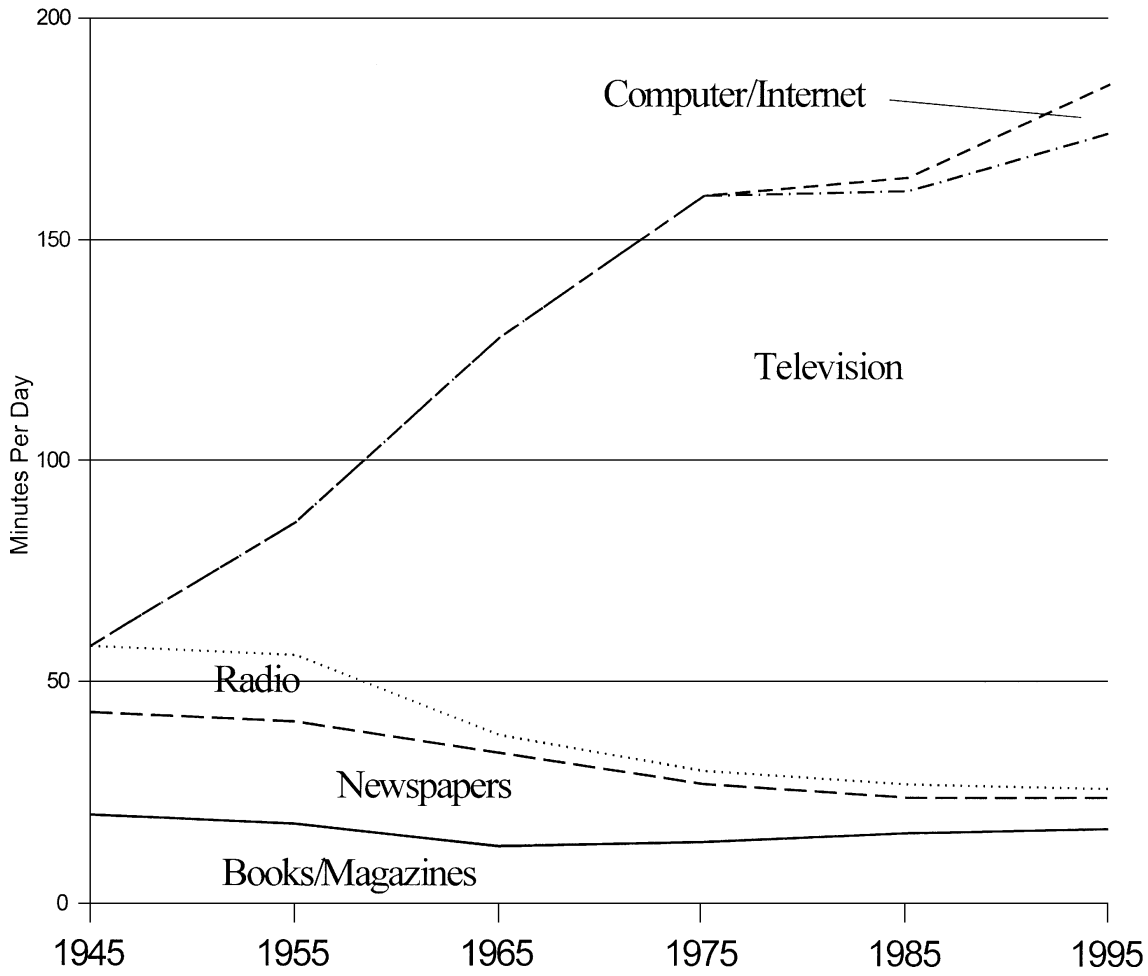

Figure 1
Minutes per day using mass media as a primary activity in US time-diary studies. Source: Americans' Use of Time Project

Similarly, different results and perspectives obtain when respondents are asked about the gratifications they receive from particular types of media content (such as news or soap operas) rather than the 'channels approach' (see below), which focuses more on respondents' preference for, and habitual use of, specific media to serve certain functions.

2. *Perspectives on Media Use*

The four topic areas covered below mainly concern quantative studies of media audiences. Qualitative studies of media audiences since the 1980s have become increasingly popular in cultural studies and communication studies (see *Audiences*).

2.1 *Uses and Gratifications*

Research on media use has been dominated by the uses and gratifications perspective since the 1970s (Rubin 1994). This perspective asserts that individuals have certain communication needs and that they actively seek to gratify those needs by using media. This active use of media is contrasted with older audience theories in which individuals were assumed to be directly affected by passively consumed media content.

The most commonly studied needs are information, entertainment, and 'social utility' (discussing media content with others or accessing media content to be used in subsequent conversation). Individuals who experience gratification when they use media are likely to develop a media use habit that becomes routine. Without such gratifications habits will not be de-

veloped and existing habits could erode. Across time, gratifications become strongly associated with high levels of media use.

Various forms of gratification are measured by different sets of questionnaire items. Individuals have been found to vary widely in their communication needs and gratifications, with the strongest and most universal communication need satisfied by media repeatedly found for entertainment, then information, and then social utility. Women tend to report stronger need for social utility than men and their use of media is more likely to reflect this need. The central assertions of uses and gratifications theory have been confirmed in a very large number of studies that have been done, focusing on different types of media content ranging from TV news, political broadcasts, and sports programs, to soap operas and situation comedies. Frequency of media use is almost always found to be strongly correlated with reports of gratification. Persons who report more social utility gratification tend to be more socially active. Persons who report more information gratification tend to be better educated and have higher social status.

Uses and gratifications research has been unable to establish strong or consistent links between gratification and subsequent effects (Rubin 1994). For example, persons who report strong entertainment gratification when viewing soap operas should be more influenced by such programs, and persons who report strong information gratification when viewing TV news should learn more.

However, it has proved very difficult to locate such effects consistently; and when effects are found, they tend to be weak, especially in relation to demographic factors. One reason for this may be that most research has been done using one-time, isolated surveys. Another is that individuals may have a limited consciousness of the extent to which media content gratifies them. To increase the accuracy of gratification self-reports, researchers have sought to identify contingent conditions or intervening variables that would identify gratification effects. Other researchers have argued that the type of media use has to be considered, such as routine or ritualized use of media vs. conscious or instrumental use. These efforts have had modest success. Some researchers reject the idea of looking for effects, arguing that the focus of uses and gratifications research should be on the media use *process*, without concern for the *outcome* of this process.

2.2 Channels

Recently, a channels approach to media research has emerged (Reagan 1996) that assumes that people develop 'repertoires' of media and of the channels into which such media are divided. These 'repertoires' consist of sets of channels that individuals prefer to use to serve certain purposes. Channels are rank ordered within repetoires. If all channels are equally accessible, then more preferred channels will be used. For example, some individuals may have a channel repertoire for national news which prioritizes network TV news channels first (e.g., ABC first, CBS second, NBC third), ranks radio news channels second, newspaper channels third, and so on. If they miss a network news broadcast, they will turn next to radio channels and then to newspapers. Reagan (1996) reports that these repertoires are larger (include more channels) when individuals have strong interests in specific forms of content. When interest is low, people report using few channels.

The channels approach was developed in an effort to make sense of the way that individuals deal with 'information-rich' media environments in which many media with many channels compete to provide the same services. When only a handful of channels is available, there is no need for people to develop repertoires as a way of coping with abundance. In a highly competitive media environment with many competing channels, the success of a new channel (or a new medium containing many channels) can be gauged first by its inclusion in channel repertoires and later by ranking accorded to these channels. For example, research shows that the Internet is increasingly being included in channel repertoires for many different purposes ranging from national news and health information to sports scores and travel information. In most cases, Internet channels are not highly ranked but their inclusion in channel repertoires is taken as evidence that these channels are gaining strength in relation to other channels.

2.3 Information Diffusion

2.3.1 Diffusion of news about routine events. Perhaps the most widely believed and influential conclusion about the news media is that television is the public's dominant source for news. Indeed, most survey respondents in the USA and other countries around the world clearly do perceive that television is their 'main source' for news about the world, and they have increasingly believed so since the 1960s. As noted above, that belief has been brought into question by studies that examine the diffusion of actual news more concretely. Indeed, Robinson and Levy (1986) found that not one of 15 studies using more direct approaches to the question found TV to be the most effective news medium; indeed, there was some evidence that TV news viewing was insignificantly associated with gains in information about various news stories. While the choice, popularity, and variety of news media have changed notably since the early 1990s, much the same conclusion—that newspapers and other media were more informative than television—still held in the mid-1990s.

That does not mean that news viewers do not accrue important news or information from the programs they view. Davis and Robinson (1989) found that viewers typically comprehend the gist of almost 40 percent of the stories covered in a daily national newscast, despite the failure of TV news journalists to take advantage of the most powerful factors related to viewer comprehension (especially redundancy). However, these information effects seem short-lived, as the succeeding days' tide of new news events washes away memories of earlier events. It appears that print media —with their ability for readers to provide their own redundancy (by rereading stories or deriving cues from the story's position by page or space)—have more sticking power. Moreover, the authors found that more prominent and longer TV news stories often conveyed less information because these stories lacked redundancy or tended to wander from their main points.

These results were obtained with national newscasts and need to be replicated with local newscasts, which are designed and monitored more by 'news consultants' who keep newscasters in more intimate touch with local audience interests and abilities than at the national network level. The declining audience for network flagship newscasts since the mid-1980s further suggests that viewers' perceptions of the 'news' for which TV is their major source may be news at the local level.

2.3.2 Diffusion of news about critical events. Another approach to news diffusion research has focused on atypical news events that attract widespread public attention and interest. Kraus et al. (1975) argued that this 'critical events' approach developed because researchers found that the news diffusion process and its effects are radically different when critical events are reported. Early critical events research focused on news diffusion after national crises such as the assassination of President Kennedy or President Eisenhower's heart attack. One early finding was that interpersonal communication played a much larger role in diffusion of news about critical events, often so important that it produced an exponential increase in diffusion over routine events. Broadcast media were found to be more important than print media in providing initial information about critical events.

In Lang and Lang's (1983) critical events analysis of news diffusion and public opinion during Nixon's Watergate crisis, both news diffusion and public opinion underwent radical changes for many months. The Langs argued that this crisis had two very different stages, first during the 1972 election campaign and second after the campaign ended. Watergate was largely ignored during the election campaign and its importance was marginalized, news diffusion about Watergate resembling diffusion of routine campaign information. Knowledge about Watergate was never widespread and this knowledge actually declined from June 1972 until November 1972. But the 1973 Senate hearings drew public attention back to it with an unfolding sequence of events that ultimately led to erosion of public opinion support for Nixon.

Other researchers have argued that the media increasingly create artificial crises in order to attract audiences. News coverage is similar to what would be used to report critical events. This means that the importance of minor events is exaggerated. For example, Watergate has been followed by an unending series of other 'gates' culminating in Whitewatergate, Travelgate, and Monicagate. Some researchers argue that coverage of the O. J. Simpson trial was exaggerated to provoke an artificial media crisis. Research on these media crises indicates that they do induce higher levels of news diffusion and that frequently the role of interpersonal communication is stimulated beyond what it would be for routine event coverage. Knowledge of basic facts contained in news coverage becomes widespread without the usual differences for education or social class. One interesting side-effect is that public attitudes toward media coverage of such events has been found to be mixed, with less than half the public approving of the way such events are covered. There is growing evidence that the public blames news media for this coverage and that this is one of the reasons why public esteem of the media has been declining.

2.4 Time Displacement

2.4.1 Print. Earliest societies depended on oral traditions in which information and entertainment, and culture generally, were transmitted by word of mouth. While their large gatherings and ceremonies could reach hundreds and in some cases thousands of individuals simultaneously, most communication occurred in face-to-face conversations or small groups. The dawn of 'mass' media is usually associated with the invention of the printing press. Print media posed a severe challenge to oral channels of communication. As print media evolved from the publication of pamphlets, posters, and books through newspapers and journals/magazines, we have little empirical basis for knowing how each subsequent form of print media affected existing communication channels. We do know that print channels were held in high regard and that they conferred status on persons or ideas (Lazarsfeld and Merton 1948). Print media are thought to have played a critical role in the spread of the Protestant Reformation and in promoting egalitarian and libertarian values. Government censorship and regulation of print media

became widespread in an effort to curtail their influence.

Based on current research findings of the 'more … more' principle described below, however, we suspect—for simple reasons of the literacy skills involved—that the popularity of older forms of print media such as pamphlets and books made it possible for later media to find audiences. Existing book readers would have been more likely to adapt and spend time reading journals and newspapers than would book nonreaders. Some researchers argue that Protestantism was responsible for the rise of print channels generally because Protestants believed that all individuals should read the Bible on their own. This led them to focus on teaching literacy. Print media were more widespread and influential in Protestant Northern Europe than predominantly Catholic Southern Europe during the early part of the seventeenth century.

2.4.2 Cinema and radio. Early in the twentieth century, movies quickly became a popular way of spending time; and anecdotal reports suggest that weekly moviegoing had become a ritual for most of the US population by the 1920s. The impact of movies was widely discussed, but there are no contemporary reports about whether the movies were frequented more or less by readers of books, magazines, or newspapers—although one might suspect so for the simple economic reason that the more affluent could afford both old and new media. Movies were especially attractive to immigrant populations in large cities because many immigrants were not literate in English.

Much the same could be expected for the early owners of radio equipment in the 1920s. Whether radio listeners chose to go to the movies, or read, less than nonowners again appears largely lost in history. By the 1930s and 1940s, however, radio appears to have been used for at least an hour a day; but it is unclear how much of that listening was performing the 'secondary activity' function that it has become today—rather than the radio serving as the primary focus of attention, as suggested in period photographs of families huddled around the main set in the living room.

2.4.3 Television. Greater insight into the time displacements brought about by TV was provided by several network and non-network studies, which documented notable declines in radio listening, moviegoing, and fiction reading among new TV owners. These were explained as being media activities that were 'functional equivalents' of the content conveyed by these earlier media (Weiss 1969).

However carefully many of these studies were conducted (some employing panel and other longitudinal designs, as in the 'Videotown' study), they were limited by their primary focus on media activities. Using the full-time diary approach and examining a sample of almost 25,000 respondents across 10 societies at varying levels of TV diffusion, Robinson and Godbey (1999) reported how several nonmedia activities were affected by TV as well, with TV owners consistently reporting less time in social interaction with friends and relatives, in hobbies, and in leisure travel. Moreover, nonleisure activities were also affected, such as personal care, garden/pet care and almost 1.5 hours less weekly sleep—activities that do not easily fit into the functional equivalence model. The diaries of TV owners also showed them spending almost four hours more time per week in their own homes and three hours more time with other nuclear family members.

Some of TV's longer-term impacts did not show up in these earliest stages of TV. For example, it took another 10 to 15 years for TV to reduce newspaper reading time, which it has continued to do since the 1960s—primarily it seems because of sophisticated local TV news advisors. Another explanation is that TV's minimal reliance on literacy skills tends to erode over time, and that it took 15 years for literacy skills to erode to the point where more people found it less convenient to read newspapers. In contrast, since the initial impact of TV in the 1950s, time spent on radio listening as a secondary/background activity has increased by almost 50 percent and books, magazines, and movies have also recaptured some of their original lost audiences. There has been a dramatic proliferation of specialty magazines to replace the general interest magazines apparently made obsolete by television. Some researchers argue that these magazines disappeared because advertisers deserted them in favor of TV. In terms of nonmedia activities, Putnam (2000) has argued that TV has been responsible for a loss in time for several 'social capital' activities since the 1960s.

Figure 1 shows the relative declines in primary activity newspaper reading and radio listening times since the 1960s, both standing out as free-time activities that declined during a period in which the free time to use them expanded by about five hours a week. TV time as a secondary activity has also increased by about two hours a week since the 1960s.

A somewhat different perspective on TV's long-term effect on media use is afforded by time series data on media use for the particular purpose of following elections. As shown in Fig. 2, TV rapidly had become the dominant medium for following electoral politics by 1964. Since then each of the other media has suffered declines in political use as TV use has maintained its nearly 90 percent usage level; more detailed data would probably show greater TV use per election follower across time as well. However, the three other media have hardly disappeared from people's media environment and continue to be used

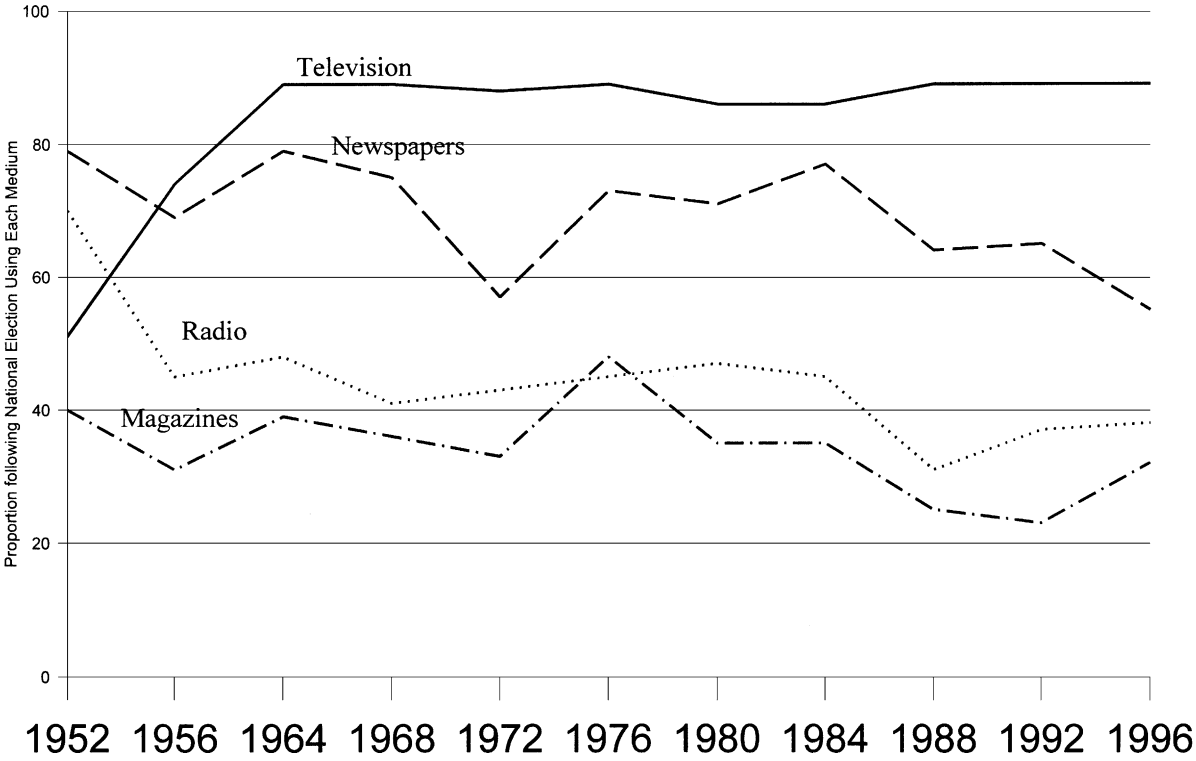

Figure 2
Proportions following each medium. Source: American National Election Studies

as important sources by large segments of the electorate in Fig. 2, such as radio talk shows in the 1990s.

What these electoral data also show, however, is that the composition of the audience for these media has changed as well. Newspapers and magazines no longer appeal as much to better educated individuals as they once did; much the same is true for TV political usage. Age differences are also changing, with newspaper and TV political followers becoming increasingly older, while radio political audiences are getting proportionately younger.

At the same time, there are clear tendencies for media political audiences to become more similar to one another. Greater uses of all media for political purposes are reported by more educated respondents, particularly newspapers and magazines. This finding appears to be consistent with the channel repertoire approach since educated respondents likely have more interest in politics and thus develop a larger channel repertoire. Older respondents report slightly more political media use than younger people, and men more than women, and people who use one political medium are more likely to use others as well.

These are patterns for political content, however; and notably different patterns would be found if the content were fashion or rock/country music and not politics. Yet there is still a general tendency for users of newspapers to use other media more (particularly for broad information purposes), bringing us back again to the overriding 'more … more' pattern of media use.

3. *Conclusions*

A recurrent and almost universal theme in media use research is the tendency of the media/information rich to become richer following the 'more … more' model. Users of one medium are more likely to use others, interested persons develop larger channel repertoires, and the college educated become more informed if news content is involved according to the 'increasing knowledge gap hypothesis' (Gaziano 1997). Robinson and Godbey (1999) have found more general evidence of this 'Newtonian model' of behavior in which more active people stay in motion while those at rest stay at rest. The increasing gap between the 'entertainment rich and poor' may likely be found as well, as when users of TV for entertainment can watch more entertainment programs when more channels or cable connections become available. Persons who have

strong communication needs and experience more gratification from media tend to make more use of media.

A prime outcome of media use then may be to create wider differences in society than were there prior to the media's presence. As media channels proliferate, people respond by developing repertoires that enable them to use these channels efficiently to serve personal purposes. These repertoires vary widely from person to person and strongly reflect the interests of each individual.

The Internet and home computer promise to bring literally millions of channels into peoples' homes. How will people deal with this flood of channels? Existing research on Internet and computer use indicate that unlike early TV users, Internet/computer users are not abandoning older channels in favor of Internet-based channels. Instead, they are merely adding Internet channels to existing repertoires and increasing their overall use of media (Robinson and Godbey 1999). These new media users also report the same amount of TV viewing, perhaps even increasing their viewing as a secondary activity while they are online. It is probably too early to conclude that these elevated levels of total media use will persist even after the Internet has been widely used for several years. The ease with which Internet use can coincide with use of other media will make it difficult to arrive at an accurate assessment of its use in relation to other media. In fact, some futurists argue that TV will survive and evolve as a medium by incorporating the Internet to create an enhanced TV viewing experience.

Advocates of the potentially democratizing influence of the Internet will probably be disappointed to find they are swimming upstream as far as the human limits of media information/entertainment flow is concerned. The more interesting question may be whether 'functionally equivalent' or nonequivalent activities will be those that will be replaced if the Internet continues its present growth trends.

See also: Advertising: Effects; Advertising: General; Audiences; Broadcasting: General; Consumer Psychology; Diffusion, Sociology of; Film and Video Industry; Film: Genres and Genre Theory; Film: History; Internet: Psychological Perspectives; Mass Media, Political Economy of; Mass Media, Representations in; Media Effects; Media Effects on Children; Media Ethics; Media Imperialism; News: General; Printing as a Medium; Radio as Medium; Television: Genres; Television: History; Television: Industry

Bibliography

Blumler J G, Katz E (eds.) 1974 *The Uses of Mass Communications: Current Perspectives on Gratifications Research*. Sage, Beverly Hills, CA

Budd R W, MacLean M S, Barnes A M 1966 Regularities in the diffusion of two news events. *Journalism Quarterly* **43**: 221–30

Chaffee S H 1975 The diffusion of political information. In: Chaffee S H (ed.) *Political Communication*. Sage, Beverly Hills, CA, pp. 85–128

Davis D, Robinson J 1989 News flow and democratic society in an age of electronic media. In: Comstock G (ed.) *Public Communication and Behavior*. Academic Press, New York, Vol. 2, pp. 60–102

Gaziano C 1997 Forecast 2000: Widening knowledge gaps. *Journalism and Mass Communication Quarterly* **74**(2): 237–64

Kraus S, Davis D, Lang G E, Lang K 1975 Critical events analysis. In: Chaffee S H (ed.) *Political Communication*. Sage, Beverly Hills, CA, pp. 195–216

Lang K, Lang G E 1983 *The Battle For Public Opinion: The President, the Press, and the Polls During Watergate*. Columbia University Press, New York

Lazersfeld P F, Merton R K 1948 Mass communication, popular taste and organized action. In: Bryson L (ed.) *The Communication of Ideas*. Harper and Bros, New York, pp. 95–118

Putnam R D 2000 *Bowling Alone*. Simon and Schuster, New York

Reagan J 1996 The 'repertoire' of information sources. *Journal of Broadcasting & Electronic Media* **40**: 112–21

Robinson J, Godbey G 1999 *Time for Life*. Penn State Press, State College, PA

Robinson J, Levy M 1986 *The Main Source: Learning from Television News*. Sage, Beverly Hills, CA

Rubin A M 1994 Media uses and effects: a uses-and-gratifications perspective. In: Bryant J, Zillmann D (eds.) *Media Effects: Advances in Theory and Research*. Erlbaum, Hillsdale, NJ, pp. 417–36

Weiss W 1969 Effects of the mass media on communications. In: Lindzey G, Aronson E (eds.) 1968–69 *The Handbook of Social Psychology*, Vol. 5, 2nd edn. Addison-Wesley, Reading, MA, pp. 77–195

Wober J M 1989 The U.K.: The constancy of audience behavior. In: Becker L B, Schoenbach K (eds.) *Audience Responses to Media Diversification*. Erlbaum, Hillsdale, NJ, pp. 91–108

Wright C R 1974 Functional analysis and mass communication revisited. In: Blumler J G, Katz E (eds.) *The Uses of Mass Communications: Current Perspectives on Gratifications Research*. Sage, Beverly Hills, CA, pp. 197–212

J. Robinson and D. Davis

Mediating Variable

A mediating variable explains and identifies the causal process underlying the relationship between two other variables. A mediating variable (M) is intermediate in the causal sequence relating an independent variable (X) to a dependent variable (Y) such that the independent variable causes the mediating variable which in turn causes the dependent variable. Mediating variables are known as intervening variables or

intermediate variables because they 'come between' the independent and the dependent variable. Important aspects of mediating variables are their close link with theory and the potential that the mediating variables identified in one context may operate in a wide variety of contexts. This article starts with examples of mediating variables, then outlines tests for mediating variables and describes limitations and extensions of mediating variable models.

1. The Mediated Effect

When quantified, the mediating variable effect is known as the mediated or indirect effect. It is called the indirect effect because it represents an effect of X on Y that is transmitted indirectly through the mediating variable. If some, but not all, of the effect of X on Y is transmitted through M, the effect is partially mediated because a direct effect of X on Y exists even after adjustment for the mediator.

The mediating variable effect differs from other 'third-variable' effects used to understand the relationship between two variables. For a moderator or interaction variable, the effect of X on Y differs for different values of the moderator variable, but a moderator does not transmit the effect of X on Y like the mediating variable. A confounding variable, when included in the investigation of the association between X and Y, changes the association between X and Y. A confounder changes the association between X and Y because it is related to both X and Y but not because it is in a causal sequence relating X to Y. Although the conceptual distinction between a confounder and a mediator is clear, it can be difficult to differentiate between them with actual data. The underlying causal sequence is the important aspect of the mediating variable.

2. Examples

Because the mediating variable elaborates the relationship between two variables by explaining why or how an effect occurs, it is of considerable importance in many fields. One of the first applications of the mediating variable was in psychological theories that learning processes, such as habit strength, explained the association of stimulus and response (Hull 1943). Later, a distinction was made between hypothetical mediating constructs representing processes or entities not directly observed and intervening variables which were observed measures of these hypothetical constructs (MacCorquodale and Meehl 1948). Duncan (1966) established the use of mediating variables in sociology when he applied path analysis techniques developed earlier by Sewall Wright (1934) to examine causal mediating relations among variables. One of the first sociological examples was that father's socioeconomic status causes the mediating variable, child's educational achievement, which causes child's socioeconomic status. The effect of father's socioeconomic status on child's socioeconomic status was not entirely mediated by child's educational achievement, so a direct effect between father's and child's socioeconomic status was also included in the model. Decomposition of effects into direct and indirect effects is a common focus of sociological research, whereas identifying mediating variables that entirely explain an effect is more common in psychological research.

A new application of the mediating variable is in chronic disease research (Schatzkin et al. 1990). Prospective studies examining the effects of etiological factors or prevention strategies on chronic disease require lengthy follow-up measurements because of the low frequency of occurrence and the slow development of the disease. As a result, researchers attempt to identify mediating variables on the causal pathway relating risk factors to disease. The mediating variables are called intermediate or surrogate endpoints. Using colon cancer as an example, the proliferation of epithelial cells in the large bowel occurs before colon cancer and is causally related to colon cancer. A study targeting cell proliferation rather than colon cancer requires less time and fewer subjects.

Mediating variables are critical in the development and application of prevention programs. Here the prevention program is designed to change mediating variables hypothesized to be causally related to the dependent variable. It is assumed that the prevention program causes change in the mediating variable which in turn causes change in the dependent variable. Programs to prevent coronary heart disease target behaviors such as diet and smoking and biological factors such as cholesterol level and blood pressure. Social influences-based drug prevention programs are designed to increase skills to resist drug offers and engender norms less tolerant of drug use. Treatment programs for substance abuse target mediators such as communication skills and social support to prevent a relapse to drug abuse.

Theories of health behavior and observed empirical relationships guide the selection of mediating variables for prevention programs. A prevention program based on established theory regarding mediating variables may be more likely to change the outcome measure and the results provide a test of the theoretical basis of the prevention program. Competing theories of the onset of drug abuse, for example, may suggest alternative mediators that can be tested in an experimental design. Prevention programs will also cost less and will have greater benefits if effective and ineffective mediating processes are identified.

Mediating variables in prevention serve a different purpose than other applications of mediating variable

methodology. In the prevention case, mediating variables are selected before the study and are fundamental to the prevention program because of their causal relationship to the dependent variable. In most other applications, the purpose of the mediating variable is to identify the processes that generated an effect, after the effect has been found.

3. *Tests of the Mediated Effect*

The single mediator model with symbols to represent relationships between variables is shown in Fig. 1. The parameter estimates and standard errors from the following three equations provide the information for tests of the mediated effect (MacKinnon and Dwyer 1993):

$$Y = \tau X + \varepsilon_1 \quad (1)$$

$$M = \alpha X + \varepsilon_2 \quad (2)$$

$$Y = \tau' X + \beta M + \varepsilon_3 \quad (3)$$

where Y is the dependent variable, X is the independent variable, M is the mediating variable, τ codes the relationship between the independent variable and the dependent variable, τ' is the coefficient relating the independent variable to the dependent variable adjusted for the effects of the mediating variable, α is the coefficient relating the independent variable to the mediating variable, β is the coefficient relating the mediating variable to the dependent variable adjusted for the independent variable, and ε_1, ε_2, and ε_3 code unexplained variability. The intercepts are not included to simplify the presentation. There are two estimators of the mediated effect, $\alpha\beta$ and τ–τ', which are algebraically equivalent in ordinary regression but not in other analyses such as multilevel and logistic regression.

There are three major types of tests of the mediated effect that use the information in the above regression models: (a) causal step tests, (b) difference in coefficients tests, and (c) product of coefficients tests. Methods to assess mediation based on causal steps entail tests of the different logical relationships among the three variables involved that must be true for a variable to be a mediator. The following sequence of causal steps described in Baron and Kenny (1986) is the most widely used method to assess mediation. (a) The independent variable (X) must affect the dependent variable (Y), τ in Equation (1). (b) The independent variable (X) must affect the mediator (M), α in Equation (2). (c) The mediator must affect the dependent variable (Y) when the independent variable (X) is controlled, β in Equation (3). The conceptual links between each necessary causal relationship and the statistical tests are clear in the causal step method. However, the causal step method has no direct estimate of the mediated effect and standard error to construct confidence limits. The first requirement, a significant relationship between the independent and dependent variable, excludes models where mediation exists but the relationship between the independent variable and the dependent variable is not significant.

The second method to test for mediation compares the relationship between the independent variable and the dependent variable before and after adjustment for the mediator. The method tests whether a third variable, here a mediator, significantly changes the relationship between two variables. The difference in the regression coefficients (τ–τ') described above is an example of this approach. Formulas for the standard error of τ–τ' can be applied to construct confidence limits for the mediated effect (Clogg et al. 1992). A drawback of the change in coefficient method is that it is conceptually more similar to a confounding variable than a mediating variable.

The third method to test the significance of the mediated effect is based on the product of coefficients which is more consistent with the causal sequence in mediation. The estimator of the mediated effect is $\alpha\beta$, the product of regression coefficients α and β. The most commonly used standard error is the first-order Taylor series solution for the product of two random variables derived by Sobel (1982) using the multivariate delta method, where σ_α and σ_β are the standard errors of a and b, respectively:

$$\sigma_{\alpha\beta} = \sqrt{(\alpha^2\sigma_\beta^2 + \beta^2\sigma_\alpha^2)} \quad (4)$$

The matrix formulas for the computation of this standard error are included in most covariance structure analysis software programs. Confidence limits for the mediated effect can be calculated using the standard error in Eqn. (4), and the result is then compared to a standard normal distribution to test for significance. The mediated effect divided by its standard error, $\alpha\beta/\sigma_{\alpha\beta}$, does not always follow a normal distribution, however.

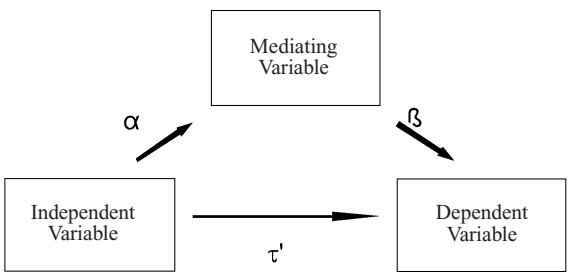

Figure 1
The mediating variable model

The product of coefficients methods provide an estimate of the mediated effect and the standard error of the mediated effect. In addition, the underlying model is a mediation model where the mediated effect is the product of coefficients hypothesized to measure causal relationships. For an independent variable coding experimental assignment, the α parameter tests whether the manipulation successfully changed the mediating variable it was designed to change and the β parameter tests whether the mediating variable is related to the dependent variable, as suggested by theory.

3.1 Causal Analysis of Mediating Variables

Methods based on the regression approach described above have been criticized based on causal analysis of the relationships among variables. For example, if X, M, and Y are measured simultaneously, there are other models (e.g., X is the mediator of the M to Y relationship or M and Y both cause X) that would explain the data equally well and it is not possible to distinguish these alternatives without more information (Spirtes et al. 1993).

The case where X represents random assignment to conditions improves causal interpretation of mediating variables (Holland 1988, Robins and Greenland 1992). Holland applied Rubin's (1974) causal model to a design where students are randomized to one of two groups, either to a group receiving motivation to study or to a control group that did not receive motivation. The mediating process is that assignment to the motivation group affects the number of hours studied which affects test performance. Under some assumptions, the typical regression coefficient for the group effect on test score, τ, and the group effect on number of hours studied, α, are valid estimators of the true causal effect, primarily because of the randomization of units to treatment. The relationship between the mediating variable of the number of hours studied and test score is more problematic (e.g., Y may cause M) and the regression coefficient β is not an accurate estimator of the causal effect because this relationship is correlational, not the result of random assignment. The estimator τ' is also not an accurate causal estimator of the direct effect. The missing information for the causal effects is whether the relationship between the number of hours studied and test score would have been different for subjects in the treatment group if they had instead participated in the control group. Recent applications of this causal approach investigate exposure to treatment as the mediating variable. For example Angrist et al. (1996) investigated the effect of Vietnam war service on health with random selection in the draft as the independent variable, serving in Vietnam as the mediating variable, and health as the dependent variable.

4. More Complicated Mediation Models

The single mediator model described above is easily expanded to include a chain of mediating variables. In fact, most mediating variables are actually part of a longer theoretical mediational chain (Cook and Campbell 1979). For example, it is possible to measure each of the four constructs in a theoretical chain from exposure to a prevention program, to comprehension of the program, to short-term attitude change, to change in social norms, to change in the dependent variable. Typically, researchers measure an overall social norms mediator rather than all mediators in the chain, even though a more detailed chain is theorized.

The single mediator methods can be extended for multiple mediators and multiple outcomes with correspondingly more mediated effects (Bollen 1987). Multiple mediator models are justified because most independent variables have effects through multiple mediating processes. The true causal relationships are difficult to disentangle in this model because of the number of alternative relationships among variables. One solution to the problems inherent in the causal interpretation of multiple as well as single mediator models is to view the identification of mediating variables as a sustained research effort requiring a variety of experimental and nonexperimental approaches to identify mediating variables. The analysis of multiple mediators in one study informs the design of randomized experiments to contrast alternative mediating variables leading to refined understanding of mediating processes (West and Aiken 1997). Meta-analytical studies provide information about the consistency of mediating variable effects across many situations (Cook et al. 1992). Furthermore, the identification of mediating variables requires examination of additional sources including ethnographic, historical, and clinical information (Cronbach 1982).

5. Future Directions

Mediating variables will continue to play a major role in the social and behavioral sciences because of the need to understand how and why variables are related. In particular, the practical and theoretical benefits of mediating variables in prevention research should guide the development of effective prevention programs. Investigators will endeavor to find the information needed for the application of Rubin's causal model and related approaches to mediating variable models. Accurate point and interval estimators of mediated effects will continue to be developed for various statistical methods including categorical and longitudinal models. In addition to statistical approaches, sustained study of mediating variables will include information from a variety of sources including history, journalism, and clinical experience.

Such comprehensive efforts are necessary to determine if a variable is truly intermediate in the causal sequence between two other variables.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Control Variable in Research; Experimenter and Subject Artifacts: Methodology; Instrumental Variables in Statistics and Econometrics; Latent Structure and Casual Variables; Moderator Variable: Methodology; Systems Modeling

Bibliography

Angrist J D, Imbens G W, Rubin D B 1996 Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**: 444–55

Baron R M, Kenny D A 1986 The moderator–mediator distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51**: 1173–82

Bollen K A 1987 Total direct and indirect effects in structural equation models. In: Clogg C C (ed.) *Sociological Methodology*. American Sociological Association, Washington, DC, pp. 37–69

Clogg C C, Petkova E, Shihadeh E S 1992 Statistical methods for analyzing collapsibility in regression models. *Journal of Educational Statistics* **17**: 51–74

Cook T D, Campbell D T 1979 *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Rand McNally College Pub. Co., Chicago

Cook T D, Cooper H, Cordray D S, Hartmann H, Hedges L V, Light R J, Louis T A, Mosteller F 1992 *Meta-Analysis for Explanation: A Casebook*. Russell Sage, New York

Cronbach L J 1982 *Designing Evaluations of Educational and Social Programs*, 1st edn. Jossey-Bass, San Francisco

Duncan O D 1966 Path analysis: sociological examples. *American Journal of Sociology* **72**: 1–16

Holland P W 1988 Causal inference, path analysis, and recursive structural equations models. In: Clogg C C (ed.) *Sociological Methodology*. American Sociological Association, Washington, DC, pp. 449–93

Hull C L 1943 *Principles of Behavior*. D. Appleton-Century, New York

MacCorquodale K, Meehl P E 1948 Operational validity of intervening constructs. *Psychological Review* **55**: 95–107

MacKinnon D P, Dwyer J H 1993 Estimating mediated effects in prevention studies. *Evaluation Review* **17**: 144–58

Rubin D B 1974 Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**: 688–701

Robins J M, Greenland S 1992 Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**: 143–55

Schatzkin A, Freedman L S, Schiffman M H, Dawsey S M 1990 Validation of intermediate endpoints in cancer research. *Journal of the National Cancer Institute* **82**: 1746–52

Sobel M E 1982 Asymptotic confidence intervals for indirect effects in structural equation models. In: Leinhardt S (ed.) *Sociological Methodology*. American Sociological Association, Washington, DC, pp. 290–312

Spirtes C, Glymour P, Scheines R 1993 *Causation, Prediction, and Search*. Springer-Verlag, New York

West S G, Aiken L S 1997 Toward understanding individual effects in multicomponent prevention programs: design and analysis strategies. In: Bryant K J, Windle M, West S G (eds.) *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*. American Psychological Association, Washington, DC

Wright S 1934 The method of path coefficients. *Annals of Mathematical Statistics* **5**: 161–215

D. P. MacKinnon

Mediation, Arbitration, and Alternative Dispute Resolution (ADR)

Mediation, arbitration and ADR ('alternative' dispute resolution) are processes used to resolve disputes, either within or outside of the formal legal system, without formal adjudication and decision by an officer of the state. The term 'appropriate' dispute resolution is used to express the idea that different kinds of disputes may require different kinds of processes—there is no one legal or dispute resolution process that serves for all kinds of human disputing. Mediation is a process in which a third party (usually neutral and unbiased) facilitates a negotiated consensual agreement among parties, without rendering a formal decision. In arbitration, which is the most like formal adjudication, a third party or panel of arbitrators, most often chosen by the parties themselves, renders a decision, in terms less formal than a court, often without a written or reasoned opinion, and without formal rules of evidence being applied. As noted below, the full panoply of processes denominated under the rubric of ADR now includes a variety of primary and hybrid processes, with elements of dyadic negotiation, facilitative, advisory and decisional action by a wide variety of third party neutrals, sometimes combined with each other to create new formats of dispute processing (see *Negotiation and Bargaining: Role of Lawyers*; *International Arbitration*; *Litigation*; *Courts and Adjudication*; *Disputes, Social Construction and Transformation of*; *Legal Systems: Private*; *Lex Mercatoria*; *Legal Pluralism*; *Lawyers*; *Judges*; *Paralawyers: Other Legal Occupations*).

1. Definitions and Types of Processes

In an era characterized by a wide variety of processes for resolving disputes among individuals, organizations, and nations, process pluralism has become the norm in both formal disputing systems, like legal systems and courts, and in more informal, private settings, as in private contracts and transactions, family disputes, and internal organizational grievance systems. There are a number of factors that delimit the kinds of processes which parties may choose or may be ordered to use under rules of law, court, or contract.

The 'primary' processes consist of individual action (self-help, avoidance), dyadic bargaining (negotia-

tion), and third party facilitated approaches (mediation), or third party decisional formats (arbitration and adjudication). 'Hybrid' or 'secondary' processes combine elements of these processes and include med-arb (facilitated negotiation followed by decision), minitrials (shortened evidentiary proceedings followed by negotiation), summary jury/judge trials (use of mock jurors or judges to hear evidence and issue 'advisory' verdicts to assist in negotiation, often conducted within the formal court system), and early neutral evaluation (third parties, usually lawyers or other experts, who hear arguments and evidence, and 'advise' about the issues or values of the dispute, for purposes of facilitating a settlement or structuring the dispute process). Increasing judicial involvement in dispute settlement suggests that judicial, and often mandatory, settlement conferences are another form of hybrid dispute mechanism. Retired judges provide a hybrid form of arbitration or adjudication in private 'rent-a-judge' schemes that are sometimes authorized by the state.

Dispute processes are also characterized by the extent to which they are voluntary and consensual (whether in predispute contract agreements, ADR *ex ante*, or voluntarily undertaken after the dispute ripens, ADR *ex post*), or whether they are mandated (by a predispute contract commitment) or by court rule or referral. The ideology that contributed to the founding of modern mediation urges that mediation should be entered into voluntarily and all agreements should be arrived at consensually (Menkel-Meadow 1995a). Nevertheless, as courts have sought increasingly to 'manage' or reduce their caseloads, and have looked to ADR processes as a means of diverting cases to other fora, even mediation may be 'mandated,' although it is usually participation in, not substantive agreement, that is required.

The taxonomy of different dispute processes also differentiates between binding and non-binding processes. Arbitration, for example, can be structured either way. Under some contractual and statutory schemes (such as the American Federal Arbitration Act), decisions by private arbitrators are final and binding on the parties, and subject to very limited court review, including only such claims as fraud, corruption of the arbitrator, or, in a few jurisdictions, serious errors of law or extreme 'miscarriages of justice.' Nonbinding processes, including nonbinding decisions in some arbitrations, allow appeals or follow-through to other processes, such as mediation or full trial. Many court annexed arbitration programs, for example, allow a de novo trial following an arbitration if one party seeks it, often having to post a bond or deposit for costs. The process of mediation itself is non-binding, in that, as it is a consensual process, a party may exit at any time; on the other hand, once an agreement in mediation is reached, a binding contract may be signed, which will be enforceable in a court of law.

Finally, dispute processes are often subject to different requirements depending on whether they are used in private settings (by contract, in employment or other organizational settings) or in public arenas such as courts. Court related or 'court-annexed' ADR programs, now encompassing the full panoply of dispute processes, may be subject to greater legal regulation, including selection, training, and credentialing of the arbitrators or mediators, ethics, confidentiality, and conflicts of interest rules, as well as providing for greater immunity from legal liability.

ADR processes are often differentiated from each other also by the degree of control the third party neutral has over both the process (the rules of proceedings) and the substance (decision, advice, or facilitation) and the formality of the proceeding (whether held in private or public setttings, with or without formal rules of evidence, informal separate meetings, or 'caucuses' with the parties, and with or without participation of more than the principal disputants). ADR processes are being applied increasingly to diverse kinds of conflicts, disputes, and transactions, some requiring expertise in the subject matter (such as scientific and policy disputes) and spawning new hybrid processes such as 'consensus building' which engage multiple parties in complex, multi-issue problem solving, drawing on negotiation, mediation and other nonadjudicative processes (Susskind et al. 1999).

Although there have been efforts to develop taxonomies or predictive factors for assignment of particular case types to particular processes (and some courts which assign or prohibit certain case types in some categories of dispute resolution), for the most part these efforts 'to fit the forum to the fuss' (Sander and Goldberg 1994) have been unsuccessful. Amenability of different cases to different processes just as often depends on the personalities of the disputants, parties, lawyers, and third party neutrals as on any particular case type characteristic.

2. *Theory and History of ADR*

The modern growth of arbitration, mediation, and other ADR processes can be attributed to at least two different animating concerns. On the one hand, scholars, practitioners, consumers, and advocates for justice in the 1960s and 1970s noted the lack of responsiveness of the formal judicial system and sought better 'quality' processes and outcomes for members of society seeking to resolve disputes with each other, with the government, or with private organizations. This strand of concern with the quality of dispute resolution processes sought deprofessionalization of judicial processes (a reduction of the lawyer monopoly over dispute representation), with greater access to more locally based institutions, such as neighborhood justice centers, which utilized com-

munity members, as well as those with expertise in particular problems, with the hope of generating greater party participation in dispute resolution processes (Merry and Milner 1993). Others sought better outcomes than those commonly provided by the formal justice system, which tend toward the binary, polarized results of litigation in which one party is declared a loser, while the other is, at least nominally, a winner. More flexible and party controlled processes were believed to deliver the possibility of more creative, Pareto-optimal solutions which were geared to joint outcomes, reduction of harm or waste to as many parties as possible, improvement of long term relationships, and greater responsiveness to the underlying needs and interests of the parties, rather than to the stylized arguments and 'limited remedial imaginations' of courts and the formal justice system (Menkel-Meadow 1984, Fisher et al. 1991). Some legal and ADR processes (like arbitration) are rule based, but other forms of ADR (negotiation and mediation) are thought to provide individualized solutions to problems, rather than generalized notions of 'justice.'

A second strand of argument contributing to the development of ADR was, however, more quantitatively or efficiency based. Judicial officers, including those at the top of the American and English justice systems, argued that the excessive cost and delay in the litigation system required devices that would divert cases from court and reduce case backlog, as well as provide other and more efficient ways of providing access to justice (Burger 1976, Woolf 1996). This efficiency based impetus behind ADR encouraged both court-mandated programs like court-annexed arbitration for cases with lower economic stakes, and encouraged contractual requirements to arbitrate any and all disputes arising from services and products provided in banking, health case, consumer, securities, educational, and communication based industries.

Modern ADR structures are related only loosely to their historical antecedents. In many countries, arbitration had its origins in private commercial arbitrations, outside of the formal court structure, and used principally by merchants when disputing with each other (Dezalay and Garth 1996). In the United States, labor arbitration developed to secure 'labor peace,' as well as to develop a specialized substantive 'law of the shop floor' (Fuller 1963).

Early use of mediation or conciliation occurred in some courts and communities seeking both to reduce caseloads and to provide more consensual agreements in ethnically or religiously homogeneous areas (Auerbach 1983). Indeed, mediation and other consensually based processes are thought to work best in regimes where there are shared values, whether based on common ethnicity, or communitarian or political values (Shapiro 1981). In Asian and other nations with more communitarian and harmony based cultures (as contrasted to more litigative or individualistic cultures), mediation is often the preferred form of dispute resolution, but it too has been used for system or political regime purposes beyond resolving the disputes of the parties (Lubman 1967). Thus, most political regimes have had to deal with both public and private forms of dispute resolution that often supplement, but sometimes challenge or compete with, each other.

The introduction or 'revival' of multiple forms of dispute resolution (including mediation, arbitration, ombuds, and conciliation) within the legal system probably dates to the 1976 conference on the 'Causes of Popular Dissatisfaction with the Administration of Justice' at which the idea of a 'multidoor courthouse' was introduced in order to meet both the caseload needs of the judicial system and the 'quality of justice' needs of consumers in a rapidly growing arena of legally and culturally cognizable claims (Sander 1976). More deeply contextualized study of the social transformation of conflicts into legally cognizable claims by a community of sociolegal scholars (Felstiner et al. 1980–81), drawing on anthropological, sociological, political, and psychological insights, also contributed to the theoretical, as well as practical, significance of pluralism in disputing.

3. *Applications*

Each of the ADR processes have their own logic, purposes, and jurisprudential justifications. Mediation and conciliation are often used to improve communications between parties, especially those with preexisting relationships, to 'reorient the parties to each other' (Fuller 1971) and to develop future oriented solutions to broadly defined conflicts. Arbitration, on the other hand, being more like adjudication (Fuller 1963, 1978) is used more often to resolve definitively a concrete dispute about an event which has transpired and requires fact finding, interpretation of contractual terms, or application of legal principles.

These basic forms have been adapted to a number of subject areas and dispute sites. As regular use of these formats of dispute resolution becomes more common, mediation seems to be overtaking arbitration as a preferred method of dispute resolution (because of the ideology of party self-determination and the flexibility of agreements). Arbitration, still most commonly used in labor disputes, is now the method of choice in form contracts signed by consumers, as well as merchants. Arbitration has, thus far, been the mode of choice for resolving international commercial, investment, and trade disputes, such as in the World Trade Organization (WTO) and the General Agreement on Tariffs and Trade (GATT). Arbitration has also been deployed in new forms of disputes developing under both domestic and international intellectual property regimes. Various forms of mediation and arbitration are also being used increasingly to resolve transnational disputes of various kinds (political, economic, natural resource allocation, and ethnic violence) and

are employed by international organizations such as the United Nations and the Organization of American States, as well as multinational trade and treaty groups (NAFTA, the European Union, and Mercosur) and nongovernmental organizations in human rights and other issue related disputes (Greenberg et al. 2000).

Beginning in the United States, but now in use internationally, mass injury (class action) cases, both involving personal and property damages, have been allocated to ADR claims facilities, utilizing both arbitral and mediative forms of individual case processing. In legal regimes all over the world, family disputes are assigned increasingly to mediative processes, both for child custody, and support and maintenance issues. In many nations, this growth in family mediation has spurred the development of a new profession of mediators, drawn from social work or psychology, who sometimes compete with lawyers both in private practice and as court officers (Palmer and Roberts 1998).

In many jurisdictions some form of referral to ADR is now required before a case may be tried. Increasingly, however, parties to particularly complex disputes, such as environmental, mass torts, or governmental budgeting, may convene their own ADR processes, with a third party neutral facilitating a new form of public participatory process which combines negotiation, fact-finding, mediation, and joint problem solving. Such 'consensus building' processes have also been applied to the administrative tribunal processes of both rule-making and administrative adjudication in a new process called 'reg-neg' (negotiated rule-making or regulation).

Although ADR has been considered, until quite recently, principally an American alternative to courts, the use of ADR is spreading slowly around the world, being used to relieve court congestion, provide expertise in various subject matter disputes (e.g., construction, labor matters, family law), build transnational dispute systems for economic, human rights, and political issues, and to offer alternative justice systems where there is distrust of existing judicial institutions. The use of ADR across borders and cultures, raises complex questions about intercultural negotiations (Salacuse 1998) and multijurisdictional sources of law or other principles for dispute resolution.

4. *Controversies*

The use of mediation, arbitration, and ADR processes, in lieu of more traditional adjudication, has not been without its controversies, reviewed briefly in this section.

4.1 *Privatization of Jurisprudence*

With the increased use of negotiated settlements, mediation, and private arbitration, there has been concern that fewer and fewer cases will be available in the public arena for the making of precedent (Fiss 1984), and debate about and creation of rules and political values for the larger community (Luban 1995). As settlements are conducted in private and often have confidentiality or secrecy clauses attached to them, others will not learn about wrongs committed by defendants, and information which might otherwise be discoverable will be shielded from public view. Settlements may be based on non-legal criteria, threatening compliance with and enforcement of law. Claims are more likely to be individualized than collectivized.

Whether there is more privatization or secrecy in the settlement of legal disputes than at some previous time remains itself a subject of controversy as empirical studies document relatively stable rates of non-judicial case terminations (at over 90 percent in many jurisdictions and across all types of disputes) (Kritzer 1991).

Related concerns about the privatization of the judicial system include increased indirect state intervention in the affairs of the citizenry through more disputing institutions, at the same time that the exit of wealthier litigants gives them less stake in the quality and financing of public justice systems (Abel 1982). The debate centers on whether dispute resolution systems can serve simultaneously the private interests of disputants before them and the polity's need for the articulation of publicly enforced norms and values (Menkel-Meadow 1995b).

4.2 *Inequalities of Bargaining Power*

A number of critics have suggested that less powerful members of society, particularly those subordinated by race, ethnicity, class, or gender, will be disadvantaged disproportionately in ADR processes where there are no judges, formal rules or, in some cases, legal representatives to protect the parties and advise them of their legal entitlements (Delgado et al. 1985, Grillo 1990–91). Responses from ADR theorists suggest that there is little empirical evidence that less advantaged individuals or groups necessarily fare better in the formal justice system, and that sophisticated mediators and arbitrators are indeed sensitive to power imbalances and can be trained to 'correct' for them without endangering their 'neutrality' in the ADR process. Many private ADR organizations have begun developing standards for good practices and Due Process protocols to protect the parties and ensure the integrity of the process.

4.3 *Evaluation and Empirical Verification of Effectiveness*

There are few robust research findings with respect to the effectiveness of ADR in meeting its claimed advantages. Recent findings from studies of ADR in the American federal courts have been contradictory about whether or not arbitration, mediation, and

some forms of early neutral evaluation do decrease case processing time or costs, either for the parties or the system. Preliminary studies from England demonstrate low usage of mediation schemes (Genn 1999). Yet studies continue to demonstrate high satisfaction rates among users of arbitration and mediation programs (MacCoun et al. 1992), and higher compliance rates with mediated outcomes than traditional adjudication (McEwen and Maiman 1986). In light of the variation in ADR programs, it is too early for there to be sufficient data bases for accurate comparisons between processes.

4.4 Distortions and Deformations of ADR Processes

Within the nascent ADR profession there is concern that the early animating ideologies of ADR are being distorted by their assimilation into the conventional justice system. Within a movement that sought to deprofessionalize conflict resolution there are now competing professional claims for control of standards, ethics, credentialing, and quality control between lawyers and nonlawyers. Processes like mediation that were conceived as voluntary and consensual are now being mandated by court rules and contracts. Processes that were supposed to be creative, flexible and facilitative are becoming more rigid, rule and law based, and judicialized as more common law is created by courts about ADR, and more laws are passed by legislatures. The overall concern is that a set of processes developed to be 'alternative' to the traditional judicial system are themselves being co-opted within the traditional judicial process with its overwhelming adversary culture. Policy makers and practitioners in the field are concerned about whether a private market in ADR is good for 'disciplining' and competing with the public justice system or whether, on the other hand, there will be insufficient accountability within a private market of dispute resolution.

5. The Future of ADR

There is no question that the use of a variety of different processes to resolve individual, organizational, and international problems is continuing to expand. New hybrid forms of ADR (as in mediation on the Internet) are developing to help resolve new problems, with greater participation by more parties. Large organizations are creating their own internal dispute resolution systems. There are clear trends in favor of mediation and arbitration in the international arena, where globalization of enterprises and governmental interests require creative and simple processes that are not overly attached to any one jurisdiction's substantive law, to promote goals of efficiency, fairness, clarity, and legitimacy, particularly in regimes with underdeveloped formal legal systems. It is also clear that there is competition over who will control such processes, and which processes will dominate in which spheres of human disputing and deal-making. The likely result is that the creative pluralism and flexibility of ADR will be subject increasingly to its own forms of formality and regulation in an effort to keep its promises of efficiency, participation, better quality outcomes, and justice.

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict: Anthropological Aspects; Conflict Sociology; Dispute Resolution in Economics; International Arbitration; Lex Mercatoria; Parties: Litigants and Claimants

Bibliography

Abel R 1982 The contradictions of informal justice. In: Abel R (ed.) *The Politics of Informal Justice: The American Experience*. Academic Press, New York

Auerbach J 1983 *Justice without Law? Resolving Disputes Without Lawyers*. Oxford University Press, New York

Burger W 1976 Agenda for 2000 AD—need for systematic anticipation. *Federal Rules Decisions* **70**: 92–4

Delgado R et al. 1985 Fairness and formality: Minimizing the risk of prejudice in alternative dispute resolution. *Wisconsin Law Review* **1985**: 1359–404

Dezalay Y, Garth B 1996 *Dealing in Virtue: International Commercial Arbitration and the Construction of a Transnational Legal Order*. University of Chicago Press, Chicago

Felstiner W, Abel R, Sarat A 1980–81 The emergence and transformation of disputes: Naming, blaming and claiming. *Law & Society Review* **15**: 631–54

Fisher R, Ury W, Patton B 1991 *Getting to Yes: Negotiating Agreement Without Giving In*, 2nd edn. Viking Penguin, New York

Fiss O 1984 Against settlement. *Yale Law Journal* **93**: 1073–90

Fuller L 1963 Collective bargaining and the arbitrator. *Wisconsin Law Review* **18**: 3–47

Fuller L 1971 Mediation: Its form and functions. *Southern California Law Review* **44**: 305–39

Fuller L 1978 The forms and limits of adjudication. *Harvard Law Review* **92**: 353–409

Genn H 1999 *The Central London County Court Pilot Mediation Scheme: Final Report*. Lord Chancellor's Department, London

Greenberg M C, Barton J, McGuinness M E 2000 *Words Over War: Mediation and Arbitration to Prevent Deadly Conflict*. Rowman & Littlefield, Lanham, MD

Grillo T 1990–91 The mediation alternative: Process dangers for women. *Yale Law Journal* **100**: 1545–610

Kritzer H 1991 *Let's Make A Deal: Understanding the Negotiation Process in Ordinary Litigation*. University of Wisconsin Press, Madison, WI

Luban D 1995 Settlements and the erosion of the public realm. *Georgetown Law Journal* **83**: 2619–62

Lubman S 1967 Mao and mediation: Politics and dispute resolution in communist China. *California Law Review* **55**: 1284–359

MacCoun R, Lind E A, Tyler T 1992 Alternative dispute resolution in trial and appellate courts. In: Kagehiro D K, Laufer W S (eds.) *Handbook of Psychology and Law*. Springer-Verlag, New York

McEwen C A, Maiman R J 1986 The relative significance of disputing forum and dispute characteristics for outcome and compliance. *Law & Society Review* **20**: 439–47

Menkel-Meadow C 1984 Toward another view of legal negotiation: The structure of problem solving. *UCLA Law Review* **31**: 754–842
Menkel-Meadow C 1995a The many ways of mediation: The transformation of traditions, ideologies, paradigms and practices. *Negotiation Journal* **11**(3): 217–42
Menkel-Meadow C 1995b Whose dispute is it anyway? A philosophical and democratic defense of settlement (in some cases). *Georgetown Law Journal* **83**: 2663–96
Menkel-Meadow C 1997 When dispute resolution begets disputes of its own: Conflicts among dispute professionals. *UCLA Law Review* **44**: 1871–933
Merry S, Milner N 1993 *The Possibility of Popular Justice: A Case Study of American Community Justice*. University of Michigan Press, Ann Arbor, MI
Palmer M, Roberts S 1998 *Dispute Processes: ADR and the Primary Forms of Decision Making*. Butterworth, London
Salacuse J 1998 Ten ways that culture affects negotiating style: Some survey results. *Negotiation Journal* **14**(3): 221–40
Sander F 1976 Varieties of dispute processing. *Federal Rules Decisions* **70**: 111–34
Sander F, Goldberg S 1994 Fitting the forum to the fuss: A user friendly guide to selecting an ADR procedure. *Negotiation Journal* **10**: 49–68
Shapiro M 1981 *Courts: A Comparative and Political Analysis*. University of Chicago Press, Chicago
Susskind L, McKearnan S, Thomas-Larmer J 1999 *The Consensus Building Handbook: A Comprehensive Guide to Reaching Agreement*. Sage, Thousand Oaks, CA
Woolf Lord 1996 *Access to Justice: Final Report to the Lord Chancellor on the Civil Justice System*. HMSO, London

C. Menkel-Meadow

Medical Experiments: Ethical Aspects

Experimental methods are of great importance in social and natural science. This article describes the nature of experiment, the ethical nature of the experimenter–subject relationship, the rights and interests approach to subject protection, the social impact of experimental methods, and the social control of experiment.

1. *The Nature of Experiment*

While a precise definition of an experiment is hard to give, for present purposes we can identify the key elements as (a) a definite state of affairs for investigation (the experimental object); (b) a second definite state of affairs, similar in all relevant respects to the experimental object (the control object); (c) deliberate and controlled modification of the experimental object; (d) observation of the experimental object, the process of modification, the modified object and the control; (e) systematic comparison of the modified object and the control. 'Object' or 'state of affairs' is deliberately vague: experiments can be performed on physical systems (for instance, atoms), biological systems (organisms, populations or ecosystems), or social individuals or systems (people, social groups, economies). The experimental and control objects are normally constructed through selection or statistical sampling to represent a specified population or natural or social kind.

1.1 *The Ethical Orientation of Experimental Methods*

The ethical issues in experiments arise out of the various operations performed to constitute, observe, modify and compare the experimental object and the control. These issues fall into three main groups: the interests of the experimental object and control; the character, motivation and behavior of the experimenters; and the impact the experiment has on current and future social interests. In addition, consideration must be given to the ethical issues of scientific research generally: including sound methodology, accurate and open publication, and fair dealing with the public.

1.2 *Experimental Objects, Subjects and Participants*

Confusingly, in the literature what is called here the 'experimental object and control' are usually referred to as the subjects of the experiment. There is a whole history of philosophy leading up to this terminology. Here it has been avoided, in order not to prejudice the question of what kinds of thing are the objects of the experiment. For instance, plants would not normally be thought of as 'subjects,' but people would be. The term 'object' is preferred here, to designate the 'thing' on which the experimenter acts. Most experiments presume or create a docile object which is malleable to the experimenter's will. The language of 'subjects' is illuminating too. The subject in an experiment is both 'subject to' the experimenter in the political sense, albeit normally with the subject's consent (again, there is a political analogy). But in philosophy or linguistics, 'subject' normally means the agent (the subject of a sentence, the knowing subject), as distinct from the object, which is acted upon.

These terminological issues are far from academic. Practically, there is a debate in medical research with patients about whether it is better to call the 'subjects' participants, in part because human-subject research usually requires the cooperation of its subjects and their action in compliance with the requests of the researcher, and in part because the 'subject' designation is felt to be demeaning and oppressive.

Theoretically, what we call the experimental object tells us something about what the relationship is between the researcher and the object. In medical research into new drugs, early phase research will involve experimentation on the bodies of animals and, later, healthy humans, to carry out pharmacological investigations into the chemistry of the drugs in complex biological systems. Here the human or animal body is the research object, not the person of the individual who 'owns' that body and who is the 'subject' governing that body. The researcher requires the cooperation of the individual subject, however this cooperation is achieved, in order that the body of the subject be docile and available for objective, pharmacological study.

Thus it is clear that the relationship between researcher and object in an experiment is philosophically complex, and essentially ethical. This ethical relationship may take on a range of values from the wholly good to the unequivocally evil, however.

2. *The Rights and Interests of the Experimental Object*

The recent history of the ethics of research has tended to concentrate on the rights and interests of the experimental object, and control. Humans certainly have interests, although these may be very difficult to identify in some circumstances. Adult humans (historically, adult male, white, middle-class humans) form the paradigmatic instance of the subjects with interests. At the very least, consideration of interest involves avoiding unnecessary harm and exploitation of the experimental object and control. For instance, placebo controlled trials in medicine can only be justified if the treatment of the control group is no worse than it would be if they were not taking part in the experiment.

2.1 *Vulnerable Subjects*

Children, the mentally ill, and the unconscious are often characterized as 'vulnerable' subjects. This relates to their physical vulnerability (exposure to risk of assault or degradation), and to their inability (or diminished ability) to consent or refuse to participate in the experiment. Other kinds of subject (prisoners or students and employees of the researcher) are defined as vulnerable because they can consent, but are vulnerable to exploitation or oppression. More controversially, some kinds of animals are regarded by some as possessing interests, by virtue of their degree of consciousness or their ability to suffer pain. Even more difficult is the relationship between rights and potential interests, as in the case of embryo research, for example. In these instances, we move away from a consideration of the rights and interests of the experimental object, towards a focus on the duties and moral character of the experimenter.

2.2 *Limits of the Rights and Interests Approach*

While the greater part of the academic and practical literature stresses the rights and interests of individuals as the foundation of research ethics, there are important tensions. First, the rights/interests focus tends to be difficult to assess directly. Most argue, moreover, that it is indeed paternalistic to try to assess what other adults ought to want or is in their interest. Hence, great stress is placed on informed consent. There is a large research literature on the best ways of obtaining informed consent, including a literature on the readability of consent forms. This leads to a number of philosophical and legal problems. Is signing a consent form the same as consenting 'authentically'? Does consent indicate a true, reflective judgment of understanding and a wish to participate? Can people consent to something that is not in their interest? Is consent a sufficient protection? Consent could be governed by a 'freedom of contract' doctrine, which allows that where two informed consenting adults make an agreement, no one else has a right to interfere, unless some third party is harmed by the arrangement. But this would disallow consideration of exploitative relationships, in which one of the parties is not harmed, but on the other hand is not fairly benefited. More generally, consent focuses on individuals, where the role of the group might be more significant.

The difficulties around consent do not invalidate its importance, but they do remind us that other factors are significant too. In the case of subjects unable to consent, we are forced to consider their interests objectively, so far as we can, but also the ethical status of the experimenter. In other words, the duties and character of the researcher are quite as significant as the rights and interests of the subjects.

2.3 *Respect and Dignity*

In the European literature on research ethics, considerable attention is paid to the 'dignity' of the human subject. This is often understood to be a peculiar property of human beings, from embryo to cadaver, although arguably there is such a thing as feline dignity or the dignity of other species too. It would be unusual to refer to the dignity of a landscape, but many would argue that while speaking of the rights of a landscape is absurd, assigning it 'dignity' or something similar, such as 'grandeur' or 'intrinsic value' would be quite sensible. Dignity arguments are hard to

apply, and it has been said that 'dignity' is something assigned only to 'objects' which are not autonomous. Thus it is best understood as a concept which orientates the researcher's attitude: ethical researchers take up a certain attitude of respect towards their subjects.

Most of the ethics of experiments is common morality—avoiding unnecessary harm, seeking permission, treating people (or animals) with respect and fairness. What gives experiments their special moral character is the experimental relationship, which transforms natural subjects (people, animals, societies, environments) into objects of deliberate manipulation. A basic principle of ethics, formulated by Kant, is that one should treat other people as ends in themselves, rather than merely as means to reach one's own ends. Here the important words are 'merely' and 'people.' Kant's principle has been criticized because it apparently excludes societies, animals and environments from moral consideration. His position can be rescued in two ways: extending the concept of 'person' to include any creature which has interests, and arguing that mistreatment of nonhuman objects involves indirect mistreatment of the humans whose interests are affected indirectly. More imaginative explorations of the ethical relationship between researcher and the object have been proposed, notably by Heidegger, who proposes that the 'technological' attitude itself is what requires moral assessment. Heidegger strongly criticizes the attitude to nature which is involved in turning people and things into mere 'resources'. Marxian concepts of alienation and exploitation, and feminist discussions of the construction of the docile (female) body run along similar lines. While these concerns are alien to most medical ethics and most analytic philosophy, these themes do arise in many bioethics contexts—for instance, in discussions of cloning, genetic engineering, environmental ethics, and 'healthy volunteer' research in pharmacology.

3. *The Social Impact of Experiments*

An issue now widely discussed is who should have a say in the moral evaluation and social control of experiments. Most scientific experiments on physical systems have from the nineteenth century until the end of the twentieth century been regarded as morally neutral, progressive and acquiring moral significance only when 'applied'. Medical experiments have had a much more complex public ideology. Some sorts of intervention—for instance, in economics—have not been regarded as experiments at all until recently (because some aspect of controlled and observed manipulation has been lacking).

Most experiments have a social impact, internally (the way humans must behave in order for the experiment to succeed) and externally (the socio-economic impact of the research). Experiments in economics or medicine, for instance, can have important consequences for the participants, which may persist long after the completion of the study. Appropriate mechanisms for rewarding and compensating subjects are needed. Participants may be required to adhere to the experimental protocol longer than they would otherwise choose, in order to preserve the integrity of the experiment. On the other hand, it should also be noted that nonexperimental interventions can be even more consequential than controlled and observed experimental ones, as the move toward 'evidence-based medicine' attests.

In many large scale social experiments, participation needs to be protected by democratic assent to the study, which can stand in tension with the need for the subjects to be 'disciplined' to behave in compliance with the protocol. This leads to a need for the social control of experiments.

3.1 *The Social Control of Experiments*

In most democratic societies it is now believed that public scrutiny of research is a good thing, through participation in the governance and ethical oversight of research. One important form of participation in this process is the 'ethics committee,' which has been adopted widely as the chief form of ethical review. Other forms of public participation, including helping to set research agendas and standards of performance, have begun to be tried, particularly in areas of widespread public concern (genetic engineering or nuclear power, for instance). A second form of ethical or social assessment is now extensively practiced, the 'technology' or 'impact' assessment. This involves social and economic analysis of interventions, usually on the basis of more experimentation, and often involving sociological research.

See also: Bioethics: Philosophical Aspects; Deceptive Methods: Ethical Aspects; Ethical Dilemmas: Research and Treatment Priorities; Ethical Issues in the 'New' Genetics; Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives; Ethics and Psychiatry; Ethics for Biomedical Research Involving Humans: International Codes; Experimentation in Psychology, History of; Objectivity of Research: Ethical Aspects; Psychiatry: Informed Consent and Care; Research Conduct: Ethical Codes; Research Ethics: Research; Research Subjects, Informed and Implied Consent of

Bibliography

Ashcroft R E, Chadwick D W, Clark S R L, Edwards R H T, Frith L J, Hutton J L 1997 Implications of socio-cultural contexts for ethics of clinical trials. *Health Technology Assessment* **1**(9): 1–65

Bernard C 1957 *An Introduction to the Study of Experimental Medicine*. Dover Books, New York
Edwards S J L, Lilford R J, Braunholtz D A, Jackson J C, Hewison J, Thornton J 1998 Ethical issues in the design and conduct of randomised controlled trials. *Health Technology Assessment* **2**(15): 1–128
Foucault M 1975 *The Birth of the Clinic: An Archaeology of Medical Perception*. Routledge, London
Habermas J 1978 *Knowledge and Human Interests*. Heinemann International, London
Heidegger M 1977 *The Question Concerning Technology and Other Essays*. Harper & Row, New York
Moreno J 2000 *Undue Risk: Secret State Experiments on Humans*. Routledge, New York
Reverby S (ed.) 2000 *Tuskegee's Truths: Rethinking the Tuskegee Syphilis Study*. University of North Carolina Press, Chapel Hill, NC
Singer P 1990 *Animal Liberation*. Random House, New York
McNeill P 1993 *The Ethics and Politics of Human Experimentation*. Cambridge University Press, Cambridge, UK

R. E. Ashcroft

Medical Expertise, Cognitive Psychology of

Investigations of medical expertise endeavor to characterize the ways experienced clinical practitioners and medical scientists perform on a range of experimental and 'real-world' tasks. Expertise research employs a cross-sectional approach that contrasts subjects at various levels of training and experience in view to understand the mediators of skilled performance. Medicine is a complex, knowledge-rich, and ill-structured domain. Ill-structured indicates that the initial states, the definite goal state, and the necessary constraints are unknown at the beginning of the problem-solving process.

1. *What is Medical Expertise?*

The cognitive characteristics of expertise are a highly organized and differentiated knowledge base, refined technical skills, and perceptual capabilities that allow experts to see the world in distinctly different ways from nonexperts. These are some of the cognitive mediators of high-level performance as exemplified by experts. Experts are also well versed in disciplinary discourse and cultural practices, and can smoothly coordinate action in team settings. As discussed in other entries in this volume, the development of expertise is a function of extensive domain-related practice and training, typically exceeding 10 years (Ericsson and Smith 1991).

Although exceptional performance is one of the distinguishing characteristics of expert, it is problematic to define expertise in terms of performance. In cognitive research, expertise is sometimes defined relative to other cohorts. For example, writers who have 5 years of experience are more expert than newcomers to the discipline. More commonly, expertise is defined according to socially-sanctioned criteria. For example, Grand Master chess players are awarded such status when they achieve a certain Elo rating on the basis of tournament play. The medical expert is typically a board-certified medical specialist in subdomains such as cardiology and endocrinology with some years of experience. Like other professional domains, medical practice is highly specialized and medical expertise is narrowly constituted. Research has documented gradations in performance as experts work in areas increasingly distal to their domain. For example, a typical cardiologist is a genuine expert in cardiology, has substantial aptitude in respiratory medicine, and would be rather challenged to treat dermatological patients. Medical knowledge has proliferated in recent years and by necessity, this has led to greater specialization and narrower bandwidths of expertise.

2. *Medical Cognition*

The term medical cognition refers to studies of cognitive processes, such as problem-solving, reasoning, decision making, comprehension, memory retrieval, and perception in medical practice or in experimental tasks representative of medical practice. These studies often focus on contrasting subjects of varying levels of expertise on sets of tasks. The subjects in these studies include medical students, physicians at varying levels of training and experience, clinical-specialists, and biomedical scientists. There are two principle experimental traditions in the study of medical cognition: a decision-making and judgment approach in which a subjects' decisions are contrasted with a normative model, indicating optimal choices under conditions of uncertainty and a problem-solving/protocol analytic tradition in which the focus is on characterizing cognitive performance on reasoning and explanation tasks. This article is principally concerned with the latter approach. For a more detailed discussion, see Patel et al. (1994).

3. *Diagnostic Reasoning*

There are three classes of interdependent cognitive tasks that comprise clinical medicine: diagnoses, therapeutics, and patient monitoring and management. Diagnostic reasoning has been the focal point of much research in medical cognition. Diagnoses can be defined as the process of identifying and classifying malfunctions in a system. In the process of diagnosis, a physician makes a series of inferences derived from

observations including patient's history, physical examination findings, laboratory tests and responses to therapeutic interventions. The foundational studies of Elstein et al. (1978) on diagnostic reasoning employed an information-processing analysis of performance influenced by Newell and Simon's (1972) seminal work on problem solving. They characterized the process of diagnostic reasoning as involving a hypothetico-deductive process in which sets of diagnostic hypotheses (typically 4 or 5) are weighed against pieces of available evidence. These investigations, as well as others that employed a similar approach, found no differences between students and clinicians or between physicians of varying levels of competency in their use of diagnostic strategies. The characterization of hypothetico-deductive reasoning as an expert strategy seemed anomalous as it was widely regarded as a weak method of problem-solving used in problems where little knowledge was available.

As the field of expertise research increasingly focussed on knowledge-rich domains such as medicine and physics, the focus of analysis shifted away from domain-general strategies to the organization of expert knowledge and its effect on reasoning. Subsequent investigations documented systematic differences in reasoning strategies between expert and nonexpert physicians. They found that when solving familiar problems, experts' employed a form of reasoning characterized by the generation of inferences from the available patient data to the diagnostic hypotheses. This forward-directed reasoning strategy was in contrast to backward-directed reasoning, where the direction of inference is from a diagnosis to the explanation of given patient data. Highly skilled physicians, who are not experts in the problem domain, are more likely to use a mixture of forward and backward-directed reasoning. Backward reasoning, which is characterized by the generating and testing of hypotheses, is akin to the hypothetico-deductive method. Pure forward reasoning is only successful when working on familiar problems. Backward reasoning is considerably less efficient and makes heavy demands on working memory because one has to keep track of goals and hypotheses. This strategy is more likely to be used when domain knowledge is insufficient. Forward reasoning is disrupted under conditions of problem complexity and uncertainty. The differential use of strategies and skills is a function of better organized knowledge structures that emerge as a consequence of medical training and clinical experience.

4. *Conceptual Understanding in Biomedicine*

This section addresses studies that have investigated the role of basic science concepts in clinical explanations and research that has focused on subjects' understanding of physiological systems. The first area of research has examined the ways in which biomedical knowledge is used by subjects at different levels of expertise on problems of varying complexity. Patel and coworkers have demonstrated that basic science concepts are used sparingly by experts on routine problems in their own domain of expertise. Intermediate subjects as well as novices tend to introduce more basic science concepts into their explanations of clinical problems. These concepts appear to be integral to the development of explanatory coherence. However, at certain stages in the acquisition of expertise, the use of biomedical concepts can actually impede diagnostic reasoning, resulting in incoherent explanations and misleading diagnostic hypotheses. In more complex clinical problems, mastery of biomedical knowledge is correlated with discriminating between relevant and less relevant problem features and selecting among competing diagnostic hypotheses.

The biomedical sciences are prototypical of domains of advanced knowledge acquisition. These domains are characterized by complex subject matter necessitating substantial prior knowledge as well as standards of coherence that exceed earlier forms of knowledge acquisition. Several studies have documented that medical students exhibit significant misconceptions in areas as diverse as cellular respiration, genetics, and cardiovascular physiology. Misconceptions emerge as a function of both formal and informal learning. These misunderstandings are not the mere result of a single piece of wrong knowledge. They reflect networks of knowledge that consist of elements that may be correct, partially correct, or substantially flawed. The study of misconceptions is oriented towards the detailed analysis of knowledge structures in order to identify the multiple sources of knowledge that comprise them. Instructional interventions have been developed to target specific misconceptions with the goal of developing more generative and robust understanding of biomedical systems and concepts (Feltovich et al. 1989).

There are competing theories on how best to characterize the acquisition of biomedical expertise. The knowledge encapsulation theory suggests that as a function of clinical experience, biomedical concepts become increasingly subsumed under clinical concepts at higher levels of abstraction with the same explanatory power (Boshuizen and Schmidt 1992). For example, explaining symptoms of chest pain and shortness of breath on exertion can invoke a more detailed explanation of circulatory function or can be more succinctly expressed in terms of symptoms leading to a particular pathophysiological condition. With increasing levels of expertise, physicians generate explanations at higher levels of generality using fewer biomedical concepts. An alternative theory expresses growth of understanding in terms of progressions of mental models. A mental model is a dynamic knowledge structure that is composed to make sense of experience and to reason across space and time. Reasoning with mental models involves a

process of mental simulation and can be used to generate predictions about future states or derive causal explanations to account for the development of a particular problem (Patel et al. 1994).

5. Dynamic Decision-making in Medical Settings

Early medical decision-making research typically contrasted physician performance with normative statistical models. These studies documented several weaknesses in physicians' decisions, most notably insensitivity to probabilities and the lack of a rational approach for weighing evidence. Subsequent research examined factors such as how problems are framed, and how uncertainty and risk affect decision outcomes. This research was shaped by the highly influential work of Tversky and Kahneman (1974). Human decision makers are constrained by limitations of information processing including selective perceptual capabilities, limited attentional resources, and errors associated with memory retrieval. These limitations guide the decision maker to construct simplified decision models and to use heuristics. Heuristics are rules of thumb that direct decision-making and sometimes result in biases that cause particular patterns of errors. Research in medical decision making has similarly found that physicians exhibit a range of biases that lead to suboptimal decisions. In general, this research portrays the expert decision maker as a fallible reasoner with marked deficiencies in decision analysis. This is at variance with other perspectives on expertise. Conventional decision-making research has been critiqued for the kinds of artificial tasks employed and the limited perspective on characteristics of domain competence such as domain knowledge and cognitive skills involved in making difficult decisions.

An emerging area of research concerns investigations of cognition in dynamic real-world environments. Naturalistic decision-making research differs from conventional decision research that typically focuses on a single decision event among a fixed set of alternatives in a stable environment. In realistic settings, decisions are embedded in a broader social context involving multiple participants and are part of an extended complex decision action process. Stress, time pressure, and communication patterns among different individuals critically affect decision processes. Naturalistic medical decision-making research within this tradition has been carried out in areas such as anesthesiology, intensive care medicine, critical care nursing, and emergency telephone consultation (e.g., Gaba 1992, Patel et al. 1996). These complex social settings all involve situations of varying levels of urgency and the smooth coordination of decisions among individuals with different kinds of expertise (e.g., cardiologists, nurses, and pharmacists). For example, three principle objectives in caring for patients are first to stabilize the patient, then to identify and treat the underlying disorder, and finally to plan a longerterm course of action. The information gathering process as well as knowledge is distributed among several team members. There are some commonalties across these investigations including how level of urgency affects decision-making strategies. Under conditions of high urgency, decisions and actions are taken with minimal justification and deliberation and follow a pattern of satisficing. In addition, senior team members such as the attending physician are essentially responsible for the course of action taken. Under less urgent conditions, multiple options may be considered and deliberated, and the decision process is somewhat more democratic in nature. Decisions leading to actions, for example to administer a certain drug, have immediate consequences (positive and negative) and iteratively inform future decisions. Recent research has also begun to characterize the factors that contribute to successful collaboration and performance among medical teams. This research suggests the need for a view of medical expertise that situates the individual practitioner within a particular social context and recognizes that although knowledge and skill are critical determinants of expertise, the ability to employ distributed cognitive resources (e.g., the knowledge and memory of other individuals), coordinate decision-action cycles, monitor continuously changing situations, and when necessary, offload information to others are equally valued attributes.

See also: Ethics for Biomedical Research Involving Humans: International Codes; Expert Systems in Cognitive Science; Expert Systems in Medicine; Medical Profession, The; Science and Technology Studies: Experts and Expertise

Bibliography

Boshuizen H P A, Schmidt H G 1992 On the role of biomedical knowledge in clinical reasoning by experts, intermediates, and novices. *Cognitive Science* **16**: 153–84

Elstein A S, Shulman L S, Sprafka S A 1978 *Medical Problem Solving: An Analysis of Clinical Reasoning*. Harvard University Press, Cambridge, MA

Ericsson K A, Smith J 1991 *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press, New York

Feltovich P J, Spiro R, Coulson R L 1989 The nature of conceptual understanding in biomedicine: The deep structure of complex ideas and the development of misconceptions. In: Evans D A, Patel V L (eds.) *Cognitive Science in Medicine: Biomedical Modeling*. MIT Press, Cambridge, MA, pp. 113–72

Gaba D 1992 Dynamic decision-making in anesthesiology: Cognitive models and training approaches. In: Evans D A, Patel V L (eds.) *Advanced Models of Cognition for Medical Training and Practice*. Springer-Verlag, Berlin

Newell A, Simon M A 1972 *Human Problem Solving*. Prentice Hall, Englewood Cliffs, NJ

Patel V L, Arocha J F, Kaufman D R 1994 Diagnostic reasoning and medical expertise. *Psychology of Learning and Motivation* **31**: 187–252

Patel V L, Kaufman D R, Arocha J F 2000 Conceptual change in the biomedical and health sciences domain. In: Glaser R (ed.) *Advances in Instructional Psychology*. Erlbaum, Mahwah, NJ, pp. 329–92

Patel V L, Kaufman D R, Magder S A 1996 The acquisition of medical expertise in complex dynamic environments. In: Ericsson K A (ed.) *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports and Games*. Erlbaum, Mahwah, NJ

Tversky A, Kahneman D 1974 Judgement under uncertainty: Heuristics and biases. *Science* **185**: 1124–31

V. L. Patel and D. R. Kaufman

Medical Geography

Medical geography deals with the application of major concepts and theories derived from human and physical geography to issues of health and disease. As such, it is not only a rapidly growing subfield of geography, but should also be considered to be a field within public health. Most of the rapid growth of medical geography has occurred since the end of World War II, and has been based on developments in geography, the social sciences generally, and the health sciences. Recently, in a reaction to what has been seen as an overemphasis in medical geography on the 'medical model' of disease, some have suggested that the field be renamed 'the geography of health' or 'the geography of health and disease' (Kearns 1993). The development of medical geography reflects the acknowledgment that issues of health and disease are not only scientifically interesting, but that health is crucial to human welfare.

1. *Traditions of Medical Geography*

There are four major foci of medical geography. These are: (a) the analysis of relationships between culture, behavior, population, environment, and disease (usually termed 'disease ecology'); (b) the identification of spatial patterns of disease, and the explanation of those patterns based upon the social, environmental, and cultural processes that generate those patterns; (c) the analysis of geographical considerations in the planning and administration of health systems; and (d) the consideration of health and disease in the broader contexts of society, political economy, social structure, and patterns of power.

For much of the twentieth century, there was a division between research in the geography of disease and the geography of healthcare (Mayer 1982). The former was concerned with the analysis and understanding of spatial patterns of disease; disease as the result of maladaptive relations between people and the environment; and disease as the result of interrelations between population, behavior, and environment (Meade 1977). Much of this was based on the work of Jacques May (May 1958), a surgeon turned medical geographer at the American Geographical Society, who wrote extensively about the ecology of disease, and the geography of nutrition. The ecology of disease considers the complex relations between the physical, sociocultural, and biological environments on the one hand, and the resulting patterns of disease on the other. May argued that for disease to occur, there must be a coincidence in time and space of agent and host. This statement reflects the implicit emphasis in early medical geography on the geography of infectious diseases. Agents, in this tradition, are the organisms that cause infectious diseases. Most of these are microorganisms such as viruses, bacteria, protozoa, and helminths (worms), although many helminths are also large enough to be grossly visible. 'Hosts' are those organisms that harbor the disease. Most medical geography has been concerned with human hosts, though there is no reason that diseases among nonhuman organisms cannot also be analyzed geographically. Some diseases, such as African Trypanosomiasis (African Sleeping Sickness) and yellow fever involve animal reservoirs, and disease ecologic analyses must therefore consider animal behavioral and spatial patterns.

The other major tradition of medical geographic research has dealt with geographical aspects of healthcare provision. Themes such as geographical aspects of access to healthcare; optimal location of facilities, ambulances, and services; and the regional planning and distribution of health systems have figured prominently in this tradition.

Until recently, the two major traditions of thought in medical geography—the geography of disease and the geography of healthcare—have been considered separately, the two are frequently inseparable and usually have mutual implications for one another. For example, the geographical distribution of disease is a major determinant of demand and need for healthcare, and this should be an important influence on how and where healthcare is provided to populations. Similarly, the geographical distribution of care can exert a strong influence on the changing distribution of disease.

Though the geography of disease has major implications for healthcare provision, the analysis of disease patterns and the underlying explanation of those patterns have scientific importance in itself. Biological and epidemiological characteristics of the population, environmental exposure to both infectious agents and environmental pollutants and toxins, social and economic characteristics of the population, and the genetic predisposition of the population to certain diseases influence the distribution of disease. The role of genetic influences has been increasingly apparent in medicine and medical science, but the health-related social sciences, including medical geography, have tended to minimize the importance of genetic influences. Genetic

influences operate at two levels: one is the genetic predisposition to certain diseases, and the other is the genetic determination of disease. To illustrate the former, family history is a major risk factor for diabetes mellitus, coronary artery disease, and many cancers. Genetic influences in this context operate in the context of individual experience—certainly, habits, customs, and behaviors all contribute to these diseases. The genetic determination of some diseases is less complex on a social level. The cause of diseases such as Tay-Sachs Disease, sickle cell anemia, Down's Syndrome, and others are completely genetic.

2. *The Evolution of Medical Geography*

Most medical geographers identify Hippocrates as the first known medical geographer. In addition to being a prominent Greek physician, he was also a major thinker in the relationships between the physical environment and health. In his essay, *Of Airs, Waters, and Places* (Dubos 1965), Hippocrates noted the importance of climate, water, and other physical characteristics as influences on the health conditions of areas. This was a major advance at the time, in terms of an implicit recognition that disease was not due to divine or mystical forces, but rather to various aspects of the physical environment. Malaria, for example, is a 'vectored disease'—anopheline mosquitoes transmit the parasites responsible for malaria (plasmodia) from person to person. The only way in which malaria can be contracted without the anophelines that are implicated in its transmission is through blood transfusions. Anophelines can only live in certain climatic, hydrologic, and geomorphologic conditions, depending upon the species of anopheline. Thus, certain environments are not conducive to the existence of anophelines. When this is the case, malaria is absent. This physical relationship was a major theme of Hippocratic thought, and attests to the importance of the human–environment relationship as an influence on disease.

Few advances were apparent in medical geography until the latter part of the eighteenth century and the nineteenth century, with the advent of the 'public health movement' in Europe and the Americas. It may be incorrect to think of this as a monolithic movement, but the major idea behind it is that a number of thinkers and social activists saw the local environment as a determinant of public health. Much of the attention was somewhat naively devoted to cleanliness and the minimization of filth for aesthetic purposes, but a major portion of this movement sought to improve the health conditions of the population by improving cleanliness, particularly in urban areas. Rodent control, water quality, and air quality were all major issues in the public health movement, and many diseases are associated with an abundance of rodents, poor water quality, and air pollution. The bubonic plague—the Black Death—which exerted a major influence on European history and has been important in other areas of the world, cannot occur in the absence of rats that carry the fleas that transmit the plague bacillus. Many diseases are waterborne, and these waterborne diseases such as cholera and other diarrheal diseases are major sources of illness in developing countries.

Much of the attention given to environmental influences on disease was countered by the discovery that specific microorganisms caused many diseases. This was caused by the development of microscopic techniques, and focussed attention on the microscopic level. The 'germ theory of disease' exerted a profound influence on science and medicine that countered the major emphasis of geographic studies on the macro-level. It was not until the second half of the twentieth century, with the development and emphasis on the field of environmental health in public health, that there was a major refocusing on the larger scale.

Other roots of medical geography in the nineteenth century came from efforts at disease mapping at the global and national scales. Some of this was the result of the need for imperial powers to know what diseases they could expect to encounter in foreign lands. Another impetus was the scientific curiosity of individuals such as August Hirsch, in his *Handbook of Historic and Geographical Pathology* (Hirsch 1883). This began the tradition of disease mapping which has always been important in medical geography.

The practical utility of disease mapping and medical geography was further realized during World Wars I and II, where again it became important to know what diseases would be encountered by troops who went abroad from their homelands. Disease mapping thus played a role in the wars.

Following World War II, the major emphasis on medical geography was initially on the development of disease ecology, first based largely on the work of May, and subsequently on that of Learmonth (1988) and others who also sought to understand the interplay of biological, cultural, and environmental factors in disease. The mobility of populations also became a theme for investigation. One of the major themes in geographical thought has been the investigation of migration and spatial interaction. Such has also been the case in medical geography. Prothero (1965), for example, investigated the role of migration in malaria transmission. This research was influential in the formation of policy with the World Health Organization (WHO).

The 'quantitative revolution' in geography of the 1960s provided an impetus to new techniques of studying disease and healthcare. Some of the earliest research using more sophisticated quantitative techniques were by researchers such as Gerald Pyle (Pyle 1971) who investigated the spatial patterns of cancers, heart disease, and stroke in Chicago. This research was also significant in that it represented some of the

earliest research by geographers to relate the spatial patterns of disease at a metropolitan level to urban structure, and to emphasize the geographical patterns of noninfectious diseases.

The quantitative revolution was also significant in that geographers, who were educated in this tradition, and particularly in urban and transportation geography, began to study the geography of healthcare. They also assessed the spatial behavior of those seeking healthcare and geographical access to facilities, and patterns of inequity and inequality based upon statistical studies and large databases. Some of the most significant work came out of the Chicago Regional Hospital Study, co-directed by the geographer Richard Morrill. This large project resulted in numerous publications and a heightened understanding of the role of numerous spatial factors in the structure of urban healthcare systems (Morrill and Earickson 1968). Some of the major findings of this project were that individuals did not always seek the nearest sources of care; that hospitals were not located optimally for the population that they served; and that factors other than geographical accessibility were important in where people got their care. It also resulted in thought about the differences between equality and equity in the geography of healthcare.

Another subject of many studies following the beginning of the quantitative revolution was spatial studies of disease diffusion or spread. Most of this work concerned and continues to concern contagious infectious diseases. One of the earliest was Pyle's study of the diffusion of cholera in the USA in the nineteenth century (Pyle 1969). This research demonstrated that the spread of cholera, in three separate epidemics in the nineteenth century, reflected the change in regional structure and connectivity. As transportation systems developed and places became better connected from the early to the late nineteenth century, the pattern of spread of cholera changed from a distance-based contagious diffusion pattern to one based more on the settlement hierarchy, and therefore usually termed a 'hierarchical' pattern by geographers.

3. *Diffusion Studies and Geographical Information Systems* (*GIS*)

3.1 *Diffusion Studies*

Disease diffusion has been a dominant theme in scientific medical geography. Subsequent to Pyle's analysis of cholera, the major advances in understanding diffusion have come mainly from the mathematical modeling of contagious disease diffusion. Among the most influential studies have been those by Cliff and others (Cliff et al. 1981) which combined mathematical models with historical analyses of measles in Iceland. Iceland was chosen because it has been, historically, a relatively isolated island community, and the disease dynamics may be studied with less concern from constant introduction of new cases from outside than in virtually any other region of its population and size. Thus, the internal spatial distribution of disease is more easily studied in such an environment. One of the major findings of these models of measles diffusion is that it is possible to predict the location of an epidemic but not its severity, or its severity but not its specific location. One major task for the future is to develop a truly predictive medical geography that can be precise in predicting the time, location, and severity of epidemics. However, this level of prediction remains elusive.

Reflecting the worldwide importance of HIV/AIDS, there have been many studies of the geographical patterns and diffusion of this pandemic. Major works have been written on the spread of HIV/AIDS, the dynamics of the epidemic, and its regional context at the national and international levels (e.g., Gould 1993, Shannon et al. 1991).

3.2 *Geographical Information Systems*

As in other areas of geography and public health, the development of geographical information systems (GIS) has proven to be a major innovation in medical geography. It allows the interactive display of many layers of spatial data, so that, for example, knowing the environmental requirements for the survival and propagation of a vector can lead to a precise display of areas where that vector might survive. An overlay with a 'buffer' depicting the range of flight for the vector can illustrate graphically the predicted areas to which the vector can fly. This can then be combined with a view of population distribution. Such a depiction may facilitate the specification of where a disease may remain endemic. This does not necessarily represent a fundamental conceptual innovation, but certainly is a technical advance. It gives more precision to the concepts of disease ecology that also deal with the range and habitats of vectors. Predictions of the future distribution of the disease are then possible based upon the predicted changes in population distribution and environmental modification. These techniques have been used to address the predicted changes in the distribution of vectored diseases that might transpire with conditions of global warming and sea level rise. The use of GIS has not been restricted to research by medical geographers *per se*, but has been responsible in itself for the growing awareness of the importance of geographical patterns of disease generally in public health and medical disciplines.

4. *Health and Social Geography*

Healthcare provision is one element of social policy, and this observation has been responsible for some writers placing medical geography within the purview

of social geography (Kearns). Health is not merely a medical phenomenon, but is partly a product of social and economic conditions. Just as there has been the development of social epidemiology as a field within public health, so too have writers analyzed the relationship between socioeconomic patterns and patterns of healthcare provision. Some of this emphasis came out of the earlier 'welfare geography,' whose proponents sought to develop effective indicators of social well being, and then to appreciate the geographical variation and explanation of those. This also parallels the WHO's definition of health as a positive state, and not just the absence of disease. This reconsideration has itself represented the major impetus between the 'new public health,' which seeks to understand the conditions that promote health, and then to implement policies at a variety of scales that will promote positive health.

Some have suggested further that the use of the new 'social theory' in geography should extend to medical geography as well. It is difficult to describe the contours of social theory in geography in 2000, but it is clearly something other than classical sociological theory. Rather, it is an amalgam of Marxist, structuralist, and postmodern concepts. There are some major overlaps in this somewhat broad school of thought with the arguments that medical geography is a subset of social geography, but the major concepts that social theory has brought to medical geography go somewhat further than that. Health is a social condition, created by social and economic structures. Moreover, health, in the tradition of social theory, is a culturally and socially defined concept. Thus, the very categories that are used in defining and discussing health and disease are products of society, and what the discussion reflects, more than anything else, are social concepts. The subject of study thus becomes transformed from the geography of health and disease to the social meaning of health and disease. This is why there is resistance among the proponents of this approach to referring to the geography of health and disease as 'medical geography'—because 'medical' implies the 'medical model' of disease, diagnosis, and treatment. It is seen as reflecting the dominance and power structure of contemporary medicine, and the biologically oriented concepts of medicine and medically related sciences, rather than the social concepts described above. A balanced discussion of the geography of health and disease must include this approach, yet it is difficult to evaluate the contribution of these concepts beyond the truisms that they reflect. A valid question is has this set of approaches led to a greater understanding of the geography of health and disease? They certainly provide context. Whether they lead to greater understanding, though, of how and why health and disease are distributed as they are remains to be seen.

Studying the social meanings of health and disease can bear a strong relationship with literary analysis, treating discussions of health and disease as 'discourses' or texts to be analyzed in the same way as literary texts. This is done in the hope that this will lead to greater understanding of the 'meaning' of the language and concepts that are used to discuss health. In the social sciences generally, such an approach is frequently termed 'discourse analysis.'

An outgrowth of this tradition, and the emphasis on the sense of place and the 'lived experience' of place in geography is the recent interest in medical geography on the importance of place for health, disease, and healing. Places with which people can identify have been termed 'therapeutic landscapes' (Kearns and Gesler 1998). The human dimensions of place are taken to be important in the process of healing. People are ill, recover, and remain well in specific places that they experience in personal ways. Explorations of the sense of place in medical geography are an important new direction for the field.

5. *Political Ecology of Disease*

Human modifications of the environment have led to changing patterns of health and disease. The human impact on the landscape is usually the result of broader social and political forces. For example, Fonaroff (1968) demonstrated how changes in Trinidad's wage economy removed people from cacao cultivation as their major occupations and resulted in increased rates of malaria. This was because people continued to cultivate cacao and other crops in fields at their homes, and did this at dawn and dusk—precisely when anopheline mosquitoes are most likely to take a blood meal. Thus, changes in the overall economy and in patterns of employment essentially caused increases in malaria incidence and prevalence. The same was true in Meade's study of land use changes in Malaysia (Meade 1976) where she demonstrated that the intensive development of rubber plantations exposed laborers to malaria because of land use changes that resulted in the development of environments favorable to anopheline reproduction at precisely those locations where laborers congregated on rubber plantations. Taking a longer historical view, the same is true of Lyme disease in the northeastern USA (Mayer 2000). Suburbanization has led to the exposure of people to 'edge environments' on the urban periphery adjacent to second growth forest. These are areas that deer preferentially settle, and deer are implicated in the transmission of ticks that serve as the vectors for Lyme disease. Thus, people, agents, and insect vectors are brought into contact with one another, as a result of broader population pressures, land use changes, housing developments, and other social factors. Understanding how government power and human action has resulted in changing patterns of disease has suggested that the concepts of political ecology may be applied to understanding disease patterns—they can

be the unintended consequences of decisions and plans in a variety of arenas that result in environmental modification (Mayer 1976, Mayer 2000).

6. Geography and Health Systems Planning

One of the applications of medical geography to health services policy is particularly important in centrally planned and financed health systems such as in the UK. Geographical concepts are crucial to the planning, budget allocation, and needs assessment processes in the National Health Service (NHS). For example, the NHS is divided into regions and subregions (districts). Funds are allocated to districts and regions, and plans developed based upon regional needs. These needs are assessed on the basis of health indicators, regional demographics, and past trends in specific areas. Facility construction and location are anticipated well into the future, and hospitals cannot be modified in their scope of services, size, or facilities without formal approval at the national level. Referral patterns from primary care practitioners (GPs—or 'general practitioners') are based largely on proximity to hospitals and specialist consultants. Patients are not free to travel long distances to facilities of their own choice under the NHS—rather, regions are allocated to specific facilities that serve those regions. This is very different to the system that has evolved in the USA, in which there is very little regional planning and co-ordination of facilities and services. Though there were mechanisms for regional planning in the past, these programs were virtually eliminated in the early 1980s as the US healthcare system moved to more of a market orientation.

This is not to say that geographical considerations are not used in the delivery of care in the USA. Trauma systems—formal arrangements of hospitals and prehospital care providers for the treatment of serious injuries—are highly regionalized, with formal patient transfer protocols to higher order centers for the most serious injuries. The same is true for burn care, organ transplantation, and certain other services in selected states and metropolitan areas. For the most part, however, hospitals are able to offer whatever services their planners and administrators deem appropriate. The assumption is that the institution (and presumably the public) will benefit if the market can support those services, and if the market cannot support those services, then they will eventually be eliminated. Needs and demand assessment based upon the population geography of hospital service areas is crucial in individual institutional planning. Thus, needs are assessed in the USA by private facilities in much the same way as any business will assess regional demand for goods. A government orientation that is evident in centrally planned and financed systems such as the NHS is thus not evident in the USA, though individual institutions still focus on regional need. It is more marketing, though, than public service planning. The US system is anomalous in a global context, where most countries have some form of nationalized system. Along with these nationalized systems comes the regional planning of facilities and institutional interrelationships.

7. Medical Geography and Indigenous Health Systems

Many societies, and particularly developing countries, have at least two different healthcare systems that are usually not well integrated. The first is Western-style medicine, with its coterie of diagnostic and therapeutic technologies, medications, and vaccinations. The second is a less formal but important system of traditional or 'indigenous' medicine and healing (Good 1987). As the name implies, this system is deeply ingrained in the local culture, and is frequently better accepted by the population than is Western medicine, which is frequently imposed on the society from the outside, or from the national healthcare bureaucracy. Traditional healing may involve herbs and herbal medicine, spiritual healing, and even shamanism. One of the major challenges is integrating the two very different systems together. They have different assumptions of the nature of illness and disease, and obviously different practices of healing and diagnosis. Since indigenous medicine is so frequently better accepted by the local population than is Western-style medicine, people may choose preferentially to use traditional healers rather than Western physicians. This makes traditional medicine more accessible to the population because of fewer cultural barriers, and also because there is usually a greater number of traditional healers than physicians in developing countries. This suggests that the geographical accessibility of traditional healers may be greater than that of Western-style medical personnel. However, for many diagnosable conditions, such as infections, surgical conditions, trauma, and other problems, Western medicine has more effective treatments than does traditional medicine. For other conditions, such as stress-related disorders, traditional healers are often effective. Anthropologists have studied the cultural integration and barriers, and geographers have studied the relationship between these two systems. The greater accessibility to the population of traditional healers suggests that a major opportunity to serve the population resides in integrating the two systems, such that people treated with conditions that are most appropriately treated by Western medical personnel can be seen by physicians and auxiliary personnel, whereas those with conditions that are treatable by traditional healers can be seen by those individuals.

This whole approach may be included in the cultural study of healthcare, in which 'alternative' or 'complementary' medicine become major subjects of study.

Data from developed countries also indicate that the majority of people frequently consult personnel such as chiropractors, massage therapists, and homeopaths who are usually not considered part of the formal or 'legitimate' system of healthcare (Gesler 1991).

8. Future Research

Future research in medical geography is difficult to predict. The following areas are likely to be major foci in the next two decades: (a) the development of a predictive medical geography, in which future geographical patterns of disease may be foreseen with reliability; (b) the analysis of health and disease in the contexts of society and social values; (c) the relationships between global change, including global environmental change, such as global warming, and patterns of disease; (d) the political economy of health systems, and the political ecology of disease; (e) the development of new methodologies and the integration of existing methods, such as mathematical medical geography and geographical information systems.

See also: Healing; Health: Anthropological Aspects; Health Policy; Medical Sociology; Public Health; Public Health as a Social Science

Bibliograpy

Cliff A D, Haggett P, Ord J K, Versey G R 1981 *Spatial Diffusion: An Historical Geography of Epidemics in an Island Community*. Cambridge University Press, Cambridge, New York

Dubos R J 1965 *Man Adapting*. Yale University Press, New Haven, CT

Fonaroff L S 1968 Man and malaria in Trinidad: Ecological perspectives of a changing health hazard. *Annals of the Association of American Geographers* **58**: 526–56

Gesler W M 1991 *The Cultural Geography of Health Care*. University of Pittsburgh Press, Pittsburgh, PA

Good C M 1987 *Ethnomedical Systems in Africa: Patterns of Traditional Medicine in Rural and Urban Kenya*. Guilford Press, New York

Gould P 1993 *The Slow Plague: A Geography of the AIDS Pandemic*. Blackwell, Oxford, UK

Hirsch A 1883 *A Handbook of Geographical and Historical Pathology*. New Sydenham Society, London

Kearns R A 1993 Place and health: Toward a reformed medical geography. *The Professional Geographer* **45**: 139–47

Kearns R A, Gesler W M (eds.) 1998 *Putting Health into Place: Landscape, Identity, and Well-being*. Syracuse University Press, Syracuse, NY

Learmonth A A 1988 *Disease Ecology: An Introduction*. Blackwell, Oxford, UK

May J M 1958 *The Ecology of Human Disease*. M.D. Publications, New York

Mayer J D 1982 Relations between two traditions of medical geography: Health systems planning and geographical epidemiology. *Progress in Human Geography* **6**: 216–30

Mayer J D 1996 The political ecology of disease as one new focus for medical geography. *Progress in Human Geography* **20**: 441–56

Mayer J D 2000 Geography, ecology and emerging infectious diseases. *Social Science and Medicine* **50**: 937–52

Meade M S 1976 Land development and human health in west Malaysia. *Annals of the Association of American Geographers* **66**: 428–39

Meade M S 1977 Medical geography as human ecology: Dimension of population movement. *Geographical Review* **67**: 379–93

Morrill R L, Earickson R J 1968 Hospital variation and patient travel distances. *Inquiry* **5**: 1–9

Prothero R M 1965 *Migrants and Malaria*. Longmans, London

Pyle G F 1969 The diffusion of cholera in the United States in the nineteenth century. *Geographical Analysis* **1**: 59–75

Pyle G F 1971 *Heart Disease, Cancer, and Stroke in Chicago: A Geographical Analysis with Facilities Plans for 1980*. University of Chicago, Chicago, Department of Geography Research Monographs No. 134

Shannon G W, Pyle G F, Bashshur R L 1991 *The Geography of AIDS: Origins and Course of an Epidemic*. Guilford Press, New York

J. D. Mayer

Medical Profession, The

Social studies of the medical profession involve the analysis of medicine as an occupation. The emphasis lies not on the technical discoveries of medicine or on the biographies of great physicians but on the ways in which the production and application of knowledge in helping prevent, cure, or care for illness is a social activity. Social science studies of medicine are historical and structural and not only cross-sectional and psychological. To understand medicine we need to explain its historical development. This article focuses on medicine in the English-speaking world.

1. Introduction

1.1 Medicine as a Profession

The traditional professions have a history rooted in the guilds of the Middle Ages (Krause 1996). In the nineteenth century the professions promised a mode of work which provided an alternative to the deadening factory systems of the day. Today bureaucracy and profession are contrasting work principles, the former

characterized by hierarchy and routine, the latter by autonomy and self-direction.

In the 1930 to 1960s and in comparison to 'ordinary' occupations, the professions were viewed as having unique attributes including an esoteric body of knowledge, a code of ethics, and an altruistic or community orientation. Yet studies of specific professions, including medicine, indicated that they were not as rigorously self-regulating, ethical, or 'community oriented' as their doctrines asserted. By the latter part of the twentieth century the professions were seen as exploiting their power to reduce competition, to secure higher incomes, to entrench their task monopolies, and to protect as much as correct or discipline, their own miscreants. The core of professional identity came to be defined as control of the content of work and professionalization as a strategy for attaining and maintaining control over a set of tasks (Freidson 1970, Larson 1977). Social studies of medicine reflect this increasing skepticism about professional altruism (Saks 1995).

1.2 A Heterogeneous Profession

To speak of 'the medical profession' is to ignore the profession's internal and international variation. Within particular countries there are dozens of medical boards, associations, and regulatory agencies. Public health and curative medicine have divergent aims and somewhat competing philosophies. Physicians work in a variety of organizations and settings which differentially shape their work and orientations.

Internationally, medicine ranges from doctors using basic medical equipment to highly skilled experts employing complex computer-assisted procedures. It encompasses practitioners who routinely accept bribes to high status doctors catering to elite clients. While some countries or regions have a physician oversupply, others need more physicians. Internationally, the ratio of physicians relative to population ranges from two (in Chadd, Eritrea, Gambia, and Malawi) to over 400 per 100,000 population (Cuba, Georgia, Israel, Spain, Ukraine). Globalization has, however, been accompanied by an almost universal reverence for highly technical medicine to which elites around the world want access. Doctors are generally of high social standing.

2. Historical Development

In the second half of the nineteenth century medicine was fragmented and fractious, composed almost entirely of males and faced with a variety of competitors from midwifery to homeopathy within pluralistic healing systems. Many laypersons thought they knew as much about health and illness as physicians. While anesthetics and aseptic techniques had helped to improve surgical procedures, medical therapies generally were not far removed from the bleeding, blistering, and purging prevalent earlier.

By the middle of the twentieth century, in less than 100 years, medicine had risen from its previously lowly position to one of social and cultural authority (Starr 1982). Medical power was considered to be legitimate and justified. Physicians controlled key health care institutions and were the acknowledged experts in what constituted disease and what should be done about it. In mid-century, unlike in earlier times patients generally benefited rather than suffered from visiting a doctor. Medical education was lengthy, rigorous, and science and technology based. Entrance into medicine was highly prized and competitive. Physicians had high status, unprecedented autonomy and authority, and high incomes. Medicine was dominant within health care (Freidson 1970, Larkin 1983, Starr 1982, Willis 1983/89).

At the end of the twentieth century, ever more complex instruments permit a new gaze into and through the human body. Innovative therapies and technologies are announced daily. Treatments promise to make even cancer more like a chronic illness than an acute fatal disease. The amelioration of symptoms is important.

The major trend for medicine has thus been that of an increasingly scientifically sophisticated, powerful and efficacious profession. There are, however, dissenting views. Medicine has possible iatrogenic effects on the individual, the institutional, and the social level (Illich 1975). Powerful technologies and pharmacologies often have negative side effects. Medicine still is more effective at treating injuries and trauma and acute disease than it is at curing the rising tide of more chronic diseases in aging populations. Medicine individualizes the social causes of illness and may rob the public of the inclination to view themselves as responsible or competent regarding their own health. The proliferating use of antibiotics creates more deadly drug-resistant organisms. Ever-more aspects of life are viewed as posing a threat to health rather than accepted as a normal part of life. Analyses indicate that more doctors and more health care are not the major factors leading to the improved health of populations (Evans et al. 1994, McKeown 1979). Medicine is criticized both for being too scientific and not 'humane' and for not being scientific enough in testing its own procedures. The profession is accused of self-interest in its attempts to hinder, prevent, or shape the establishment of health care systems designed to ensure more equal public access to health care.

2.1 Medicine, Central and Powerful yet Challenged

Medicine is always part of, rather than 'above' society. In the nineteenth and early twentieth centuries medi-

cine was overtly, sexist reflecting the patriarchal social structures of the time. Women patients were viewed as fragile victims of their own uteruses (Ehrenreich and English 1973). Before and during World War II, the profession in Germany actively displayed anti-Semitism and physicians took part in the Holocaust. Under state command economies, medicine frequently, as in psychiatry, served the purposes of the state in oppressing or subduing citizens. In Europe and North America, medicine early on was eugenicist and, in many jurisdictions, physicians sterilized the mentally ill or retarded. Later, tranquilizers used by doctors to treat depression in women also served to 'adjust' women to what are now perceived as oppressive social roles. Given the tendency to reflect the interests, values, and regulatory regimes of the era in which it is embedded, medicine can be seen as part of the movement to control and regulate populations (Lupton 2000).

Medicine is part of the movement towards medicalization, the tendency to view human phenomena, not in terms of right or wrong, but in terms of health and illness. Yet the medical profession is not the only source of such medicalization. Health care is the source of profits within a rapidly expanding medical–industrial complex. Nonprescription drugs and food supplements are as popular as prescription drugs. The drug industry is one of the most powerful drivers of the health field and its research directions and applications.

We are living in a 'risk society' (Beck et al. 1994). Almost any activity now carries with it an assessment of the possible threat to health and safety that it contains. In a secular and materialist society, human life is not viewed as preparation for an afterlife but as an end in itself. There is a desperate desire to delay death, avoid or control risks, and to appear and be, young and vigorous, or at least not old and decrepit. In all of these areas medicine has something to say.

Nevertheless, seemingly at the height of its powers, and with new genetic discoveries promising much more in the coming decades, medicine faces challenges from all those whom it had previously controlled or influenced, from patients to governments (Hafferty and McKinlay 1993). In what follows, the historical rise of, and contemporary challenges to, medicine are traced through a brief analysis of medical knowledge, education, practice, and discussion of 'the rise and fall' of medicine.

3. *Medical Knowledge*

Medical knowledge and medical power are intimately related. Control over the production of knowledge as well as credentialism are methods whereby medicine has attained and maintains control over health care. The profession administers the production of knowledge in medical schools or health science complexes. Physicians are the guardians of access to patients and to prescription drugs, crucial aspects of the research endeavor.

Yet, doctors in the eighteenth and nineteenth centuries claimed authority, not on the basis of science, but because of their gentlemanly status. This authority was underwritten by attendance at elite universities and knowledge of classical learning. Medicine was a male middle-class enterprise. Ever-increasing educational requirements served to discourage, or weed out, working class and minority groups.

In the nineteenth and twentieth centuries medicine asserted authority because of its association with the rapidly rising prestige of science. Medicine was one of many groups claiming to be using science to improve human well-being. But this was a particular kind of medicine. A medicine based on individual biology and the germ theory gained ascendance over a public health medicine focused on the social roots of illness in poverty and poor living conditions.

Recent developments have undermined the authority of medical claims to expertise. New forms of knowledge challenge these professional underpinnings. In the late twentieth century nonphysician experts question the efficacy of particular medical procedures. New corps of clinical epidemiologists and others, often at the behest of governments, are busy formulating what works and what does not. Even clinical work can be routinized through medical guidelines or protocols.

In the application of medical knowledge to practical problems, the essence of what it means to be a profession, medicine develops interests not necessarily pregiven by medical knowledge and not necessarily congruent with the needs or interests of patients, the public, or payers (Freidson 1970). To serve its own interests medicine in many countries opposed 'state intrusion' into health care or attempts to make access to care more equitable (Hafferty and McKinlay 1993).

The current 'rationalization' of health care has weakened the links between medical knowledge and medical authority both regarding medical control over health care systems and over clinical work (Gabe et al. 1994, Hafferty and McKinlay 1993). Economists, planners, and managers claim more expertise in health care policy than physicians. State or corporate involvement in health care systems, new modes of payment, financing of equipment, insurance company scrutiny, for-profit health corporations, state protocols, and clinical guidelines shape what doctors do in their day-to-day work.

The relationship between medicine and the social sciences has itself changed. Once medicine used the social sciences to accomplish its own tasks. Typically, social scientists were asked to help understand why some patients did not 'comply with' medical regimes. Now, social science has begun to undermine medical

authority through challenging the biomedical paradigm, through its focus on medical power and self-interests, and for its support of the view that health is determined more by social factors than it is by health care. Social science analyses were once 'in' medicine but are increasingly now 'of' medicine.

The foundations of medical knowledge are being challenged by new views of science in which knowledge is viewed as being socially constructed or determined rather than science directly reflecting 'nature.' There are now claims that medicine is 'constructing' the body rather than simply 'discovering' it (Lupton 2000). Biomedicine is not viewed as scientifically neutral but as a truth claim in contests over power. From this perspective the profession is also conceived as part of state attempts to control populations through implicit methods of 'surveillance' and of inculcating norms of behavior through various medically approved ways of thinking and acting (Petersen and Bunton 1997, Turner 1995).

4. *Medical Education*

In the late nineteenth and early twentieth centuries university education became the official certification of professional standing. Medical training changed from a rather haphazard process, given in some countries in profit-making schools, to a scientific education, most often accomplished in public universities (though in some countries the most prestigious schools were in private colleges and universities). The Flexner Report in the United States in 1910 reflected the triumph of the new biomedicine and was important in making university education more scientific and laboratory based, on the German model. Dozens of schools were closed and exclusionary tactics and credentialism raised the status of the profession, controlled competition within, and helped reduce competition without (Larson 1977).

By the mid-twentieth century medical students generally took four years of education followed by internship and perhaps another three or four years of specialist training. Usually the first two years were science based with clinical training being introduced over the subsequent years. Throughout the century the power of the universities and medical schools or the state over medical education increased and the direct power of the profession over curriculum and even the number of students being trained decreased.

Medical students are now more diverse in background. Jews are no longer overtly excluded or ostracized and, beginning in the 1960s, women formed an increasing proportion of the medical school population. Other minority groups are still not equitably represented among medical students and practitioners but now as much for class reasons as because of race or ethnicity.

Early social science perspectives had held to the view that what physicians did in their practices was a reflection of the training they had undergone. By contrast, in the 1970s it became accepted wisdom that the behavior of physicians was more a reflection of the work situation in which they practiced than it was of their educational experiences (Freidson 1970). Medical students were being socialized into the role of the medical student rather than into the role of the physician. They learned how to study, how to be emotionally detached from patients, how to tolerate clinical ambiguity and, in their confrontations with their clinical teachers, to become adept at presenting themselves as knowledgeable and competent. Medical students are now collectively less passive. In some countries interns and residents have organized into associations and unions to improve their conditions of study and work and their pay.

5. *Medical Practice*

From the earliest times doctors practiced by themselves within markets. In the eighteenth and nineteenth centuries doctors faced competition both from other physicians and from 'other' healers ranging from midwives to bonesetters. The use of physicians was a last resort and hospitals were largely a place for the poor to die. In any event few could afford physician fees. With improved perceived efficacy and the rise of a market for health care, physicians emerged as the pre-eminent free-enterprise professionals and adopted a corresponding individualist ideology. In the 1930s medicine was still largely based on solo practice. But within half a century there was a major transformation of medicine from physicians as petty bourgeois entrepreneurs catering to patients in the domestic sphere to physicians as part of a mass health care market treating patients in offices or in the new 'health factories.' In most Western countries spending on health care now ranges around a mean of about 9 percent of national GNPs. Worldwide, however, health care expenditures varies widely, from 2 percent of GNP (in low GNP countries) to over 14 percent (the United States).

As part of the movement towards the welfare state, the provision of medical care became a public concern. Internationally many forms of payment mechanism and system developed, from health care systems as in Britain to national health insurance (e.g., Canada) to more private systems (e.g., the United States) or to mixed forms (e.g., Australia, Germany). These systems influenced what physicians did, and where, for how much, although in somewhat different ways (Hafferty and McKinlay 1993, Johnson et al. 1995). Today increasing proportions of physicians are paid by salary or capitation rather than the traditional fee-for-service, although in some places such as in Eastern Europe, this trend is reversed. Though medicine

resisted the intrusion of state or corporate power or tried to shape it, everywhere there is more external control than there was previously over the nature, conditions, and quality of practice.

Medical autonomy is being undermined in both private and publicly financed or organized systems. In the most predominantly private system, that in the United States, privately owned provider organizations control the day-to-day work of doctors in the name of profit or saving costs. In publicly financed systems, once some form of control over physician costs had been instituted, the state tends to leave purely 'clinical' matters in the hands of the profession. Yet general health care policies, such as the almost universal controls over the use of technologies or over payment mechanisms, do have a profound impact on what individual physicians do at the clinical level. Nevertheless, medicine in more private or more state run systems is subject to different constraints.

The heart of professional standing is self-regulation, yet, within the market and outside of it, by competition, by corporate fiat and/or by state regulation, medical practice is subject to more surveillance and regulation. There is a deepening division in many jurisdictions between the function of representing the profession and that of licensing and regulation. Organizations with the latter aims, even if still largely controlled by physicians, are more constrained by public or governmental input, direction, or representation than they were previously. Professional self-regulation, the core measure of professional standing and autonomy, is never complete and is always qualified.

Physicians are a cohesive and closed group, dependent on colleagues for patient referrals, for recommendations, and for mentorship though now competing for patients. There is competition in some big city markets for patients, and, in various state-administered systems, there is physician unemployment. Perhaps as a result of competition, for the first time in the United States in 1994 the average income of physicians dropped. Internal fragmentation made more difficult the efforts of medical elites to mobilize the profession to face the political, economic, and social challenges of the late twentieth and early twenty-first centuries.

The last third of the twentieth century witnessed the rise of neo-liberalism and globalization accompanied by attacks on the role of the state and the decline or transformation of the welfare state. Health care systems became more controlled, rationalized, and privatized. The interests of medicine are no longer necessarily congruent with the forces which led to welfare state decline. Business and states are in an implicit or explicit coalition to control health care costs hence the work of physicians. Medical practice has thus historically moved from a situation of patronage, in which clients controlled physicians, to one of collegiate or colleague control in a medically dominated system, to one in which doctor–patient relationships are mediated by (public or private) third parties (Johnson 1972).

5.1 *Understanding Changes in Medicine*

What do developments in the last century and a half tell us about the medical profession? There is more agreement regarding the general trends than there is about how to understand these. Medicine rose to power in the nineteenth and early twentieth centuries. By the mid-twentieth century, medicine was dominant in health and health care, controlling the content of care, patients, other health occupations, and the context within which medicine is practiced (Freidson 1970). Other health occupations were absorbed, subordinated, limited, or excluded from the official health care system (Willis 1983/89). Powerful within health care systems, medicine had important effects on the societies in which it was embedded.

Today, however, medicine faces challenges from all those with whom it interacts. Patients use unorthodox forms of healing and herbal medicines, demand particular forms of care, and complain about or sue physicians. Health care occupations, many of them 'female' occupations, from nursing to physiotherapy chafe under medical control, seek autonomy, and chip away at the borders of the medical monopoly through credentialism, claims to a unique knowledge, and through political bargaining. A proliferating variety of forms of unorthodox healing, from shiatsu to chiropractic, claim legitimacy (Cant and Sharma 1999). States and corporations come to shape medical work. The impact of a technologically sophisticated medicine on human well-being or the health of nations is not as clear as it once appeared. Moreover, just when medicine is facing external challenge it is weakened by internal fragmentation by specialty and gender and by a developing hierarchy between an academic and research elite and practitioners. The major issue for social science analysts has thus become to understand a historical trajectory of rise, consolidation, and challenge or possible decline of medical power.

A common-sense explanation for medical power is that medicine became authoritative as it gained in scientific knowledge and became more effective in curing disease. However, while improving effectiveness, or at least the perception of such efficacy, was important (doctors have to be sought out by patients), historians have noted that medicine became cohesive and dominant before it became more efficacious. Efficacy based theories of the rise of medicine also have difficulty explaining recent weakening medical power (though changing disease patterns are important).

Many social science analyses of medicine have been descriptive narratives loosely based on a simple interest group perspective and have focused on inter-

professional relationships. Freidson's emphasis on medical dominance in the 1970s stimulated broader debates about the reasons for the rise of medical power. Pluralist interest group theories, or forms of closure theory (Murphy 1988, Witz 1992) which stress medicine's own struggles as part of a 'system' of professions, contrast with structurally oriented theories which emphasize the rules within which interest group struggles occur (Abbott 1988, Murphy 1988, Navarro 1986, Witz 1992). Closure theory describes the strategies of demarcation of occupational boundaries and of exclusion used by medicine in attaining and maintaining its dominance over other health occupations. By contrast, structural theories claim that medical power is fundamentally anchored in the congruence of its interests and ideologies with those of dominant elites, classes, and the state. Biomedicine became predominant partly because its individualist oriented explanations drew attention away from the social causes of disease. In general a focus on medical 'agency' views medicine as imposing biomedicine on society, while structural analyses focus more on the 'selection' of a particular type of medicine by prevailing structures of power.

5.2 Is Medical Power Declining?

Has the social and cultural authority of medicine declined since the medical dominance thesis was first proposed? Most analysts acknowledge that medicine is being challenged, particularly since governments, as part of the rise of welfare states, are viewed as 'intruding' on medical power within health care. A few theorists argue that state involvement in health care actually crystallized medical power in place (Larkin 1983) or that medicine as a corporate entity retains its power externally although individual practitioners may have lost some of their autonomy to medical elites. But the consensus is that medical power is not what it once was.

The terms proletarianization or corporatization (McKinlay and Stoeckle 1988) depict trends influencing medicine as similar to those that affected workers in the nineteenth century as skilled workers became employees and their work came to be managerially controlled, if not routinized and fragmented. Medicine, previously an autonomous and dominant profession, has become subject to state, corporate, and bureaucratic imperatives. Although the proletarianization thesis applies more closely to the United States with the major instance the corporate micro-management of physician work, still, state or corporate rationalization of health care everywhere involves more control over what physicians do and how they do it. Deprofessionalization refers to the decreasing distance, in knowledge or education, between doctors and patients or the rise of consumerism generally (Haug 1975).

It may be that all of these different perspectives reveal something about medical power, although referring to somewhat different levels of analysis and/or to different aspects of medical power. The proletarianization thesis most clearly applies to medical work; deprofessionalization to medical knowledge; and social closure to inter-professional struggles. Proletarianization is the most general concept since employment status and managerial control as part of a general process of the corporatization and bureaucratization of care can affect most aspects of medical dominance. For example, current reforms of health care almost inevitably involve pressure for a reorganization of interprofessional relationships within the health care division of labor in the name of efficiency, potentially leading to a weakening of the medical monopoly.

The notion of professionalization as a strategy of control over a particular work domain, which includes as a key component professional autonomy and self-regulation, implies continual struggles over such control. State and private interests now coincide in viewing medical power as a barrier to a reorganized and less costly health care. In this sense, both the rise of medicine to a dominant position and its current possible decline can be viewed in terms of the congruence or incongruence of its interests and ideologies with class and state forces external to health care. Medical power, as both Freidson and Navarro in their different ways have pointed out, is always contingent on its fit within the power structure of the societies of which it is a part.

Analyses of medicine have been given recent impetus both by world events and by new social science perspectives. Economic globalization and the rise of an emphasis on markets in health care have brought a new urgency to attempts to understand the role of medicine. For example, might a return of health care markets bring a revival of medical power? It seems unlikely to do so because market oriented governments regard the professions as unneeded market monopolies and health markets are themselves constraining. If some physicians in some countries benefit from neo-conservative or even 'third way' policies, their gains are somewhat accidental and contingent. Medicine no longer has the power to define the health agenda.

Explanations for the changing nature of medicine are also caught in the debates between agency and structural determination which pre-occupy the social sciences. The notion of differing levels of analysis, individual and group actions within particular changing social structural conditions might provide a way out of this dilemma. New perspectives in the social sciences may help bridge or transcend these dichotomies (Scambler and Higgs 1998). For example, notions from Foucault of power as enabling and not simply repressive may change our views of doctor–patient power conflicts (Petersen and Bunton 1997). Social

constructionist views that biomedicine is simply one among a number of different and presumably equally valid ways of viewing the world and the body may fundamentally alter traditional views of science and knowledge. Theories about the effects of globalization on state autonomy, class structure, and professional powers may prove important (Coburn 1999) as might attempts from critical realism to view agency as operating within structures. Certainly, any perspective on contemporary developments has to come to terms with the broader economic and political changes, many of these towards more market oriented economies and societies, which are sweeping the world. Such analyses, however, are more promises for the future than current accomplishments.

6. *Conclusions*

Medicine is like other forms of work and subject to many of the same general pressures. Physicians are both active in shaping and reproducing particular social structures yet the profession cannot escape being subject to such structures. Most recent social studies of the medical profession are oriented to its power, to changes in that power and evince an increasing skepticism to professional claims of altruism. The focus is on the self-interests of the profession and its changing accommodation with external powers. This view contrasts with perspectives which emphasize the many individual (Norman Bethune in China) and collective (Medicine Sans Frontieres) acts of heroism and self-sacrifice of physicians or groups of physicians. The social science emphasis on self-interests and power is a little one-sided, medicine is not only self-interested. Moreover, a health care controlled by private corporations, managers, or the state may presage a worse future for patients than one in which medicine, or at least individual physicians, retain a relative autonomy from external powers.

Medicine rose during a particular period of industrial capitalism and flourished, perhaps had its 'golden age', during the period of the development of the welfare state. In that milieu medicine could partially reconcile both service to others and its own self-interests. We are today in a new era, one which promises a greater emphasis on individualism and on markets. Thus, medicine, as the societies of which it is a part, is currently in the midst of profound changes, the ramifications of which are as yet unclear. At the center of societal and individual concern and interest yet challenged by others as to its pre-eminence. Embedded within a huge industry much of it driven by the desire for financial gain, and enmeshed with state bureaucracies with their own imperatives. Medicine is revered yet envied, sought after but viewed with skepticism, scientific yet saturated with self-interests and motives not pregiven by its science or service, with many potentially altruistic practitioners yet apparently mercenary organizations. The challenge for medicine is to fulfil its service to others in an era of profit seeking. The social sciences focused on medicine face the task of providing an understanding of medicine more adequate for contemporary transformations and contradictions.

See also: Health Professionals, Allied; Medical Sociology; Professions, Sociology of

Bibliography

Abbott A 1988 *The System of Professions: An Essay on the Division of Expert Labour*. University of Chicago Press, Chicago

Beck U, Giddens A, Lash S 1994 *Reflexive Modernization: Politics, Tradition, and Aesthetics in the Modern Social Order*. Polity Press, Cambridge, UK

Cant S, Sharma U 1999 *A New Medical Pluralism: Alternative Medicine, Doctors, Patients and the State*. UCL Press, London

Coburn D 1999 Phases of capitalism, welfare states, medical dominance and health care in Ontario. *International Journal of Health Services* **29**(4): 833–51

Ehrenreich B, English D 1973 *Complaints and Disorders: The Sexual Politics of Sickness*. Feminist Press, Old Westbury, NY

Evans R G, Barer M L, Marmar T R (eds.) 1994 *Why are Some People Healthy and Others Not?* A de Gruyter, New York

Freidson E 1970 *Profession of Medicine: A Study of the Sociology of Applied Knowledge*. Dodd, Mead and Co., New York

Gabe J, Kelleher D, Williams G (eds.) 1994 *Challenging Medicine*. Routledge, London

Hafferty F W, McKinlay J B (eds.) 1993 *The Changing Medical Profession: An International Perspective*. Oxford University Press, New York

Haug M R 1975 The erosion of professional authority: A cross-cultural inquiry in the case of the physician. *Milbank Memorial Fund Quarterly* **54**: 83–106

Illich I 1975 *Medical Nemesis: The Expropriation of Health*. McClelland and Stewart and Marion Boyars, London

Johnson T 1972 *Professions and Power*. Macmillan, London

Johnson T, Larkin G, Saks M (eds.) 1995 *Health Professions and the State in Europe*. Routledge, London and New York

Krause E A 1996 *Death of the Guilds: Professions, States and the Advance of Capitalism, 1930 to the Present*. Yale University Press, New Haven, CT

Larkin G 1983 *Occupational Monopoly and Modern Medicine*. Tavistock, London

Larson M S 1977 *The Rise of Professionalism: A Sociological Analysis*. University of California Press, Berkeley, CA

Lupton D 2000 Social construction of medicine and the body. In: Albrecht G L, Fitzpatrick R, Scrimshaw S C (eds.) *Handbook of Social Studies in Health and Medicine*. Sage, London, Thousand Oaks, CA

McKinlay J B, Stoeckle J D 1988 Corporatization and the social transformation of doctoring. *International Journal of Health Services* **18**: 191–205

McKeown T 1979 *The Role of Medicine: Dream, Mirage or Nemesis?* Basil Blackwell, Oxford, UK

Murphy R 1988 *Social Closure: The Theory of Monopolization and Exclusion*. Clarendon Press, Oxford, UK

Navarro V 1986 *Crisis, Health and Medicine: A Social Critique*. Tavistock, New York
Petersen A, Bunton R (eds.) 1997 *Foucault, Health and Medicine*. Routledge, London and New York
Navarro V 1989 Professional dominance or proletarianization?: neither. *The Milbank Quarterly* **66**(suppl. 2): 57–75
Saks M 1995 *Professions and the Public Interest: Medical Power, Altruism and Alternative Medicine*. Routledge, London
Scambler G, Higgs P (eds.) 1998 *Modernity, Medicine and Health: Medical Sociology Towards 2000*. Routledge, New York
Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York
Turner B S 1995 *Medical Power and Social Knowledge*, 2nd edn. Sage, London
Willis E 1989 *Medical Dominance: The Division of Labour in Australian Health Care*, 2nd edn. George Allen and Unwin, Sydney (1st edn. 1983)
Witz A 1992 *Professions and Patriarchy*. Routledge, London

D. Coburn

Medical Sociology

Medical Sociology is a subdiscipline that draws on the methodologies and middle range theories of substantive sociological specialities to elucidate important health, health services organization, and health care utilization issues. The fields drawn on most commonly include social stratification, organizational analysis, occupations and professions, social psychology, gender, and political sociology. Medical sociology also shares concepts and methods with related fields such as public health, health services research, medical economics, medical anthropology, social epidemiology, demography, and ecology. Sociologists working in health may use health behavior and institutions as areas of study to advance theory and methods in sociology generally, or may be motivated primarily to solve applied problems relating to improvement of health care, organizational arrangements, and processes of care (Freeman and Levine 1989).

1. *Social Stratification and Health*

The single most dominant theme in medical sociological research addresses the way in which stratification by socio-economic status (SES), gender, race, ethnicity, and age affects patterns of health and illness behavior, illness risk, morbidity, disability, mortality, access to health care services, utilization of care, functional health outcomes, subjective well-being, and health-related quality of life (Mechanic 1978). The link between SES and health has been repeatedly documented over the twentieth century and throughout the world (Amick et al. 1995). There is renewed interest in studying economic inequalities in health in many countries and in examining whether inequality itself, independent of income, contributes to poorer health outcomes (Wilkinson 1996). A variety of hypotheses have been suggested, linking inequality to health outcomes at both the community level (social capital and social cohesion) and at the individual level (sense of personal control, perceptions of equity). Ascertaining the role of inequality requires careful specification of types of income inequality and data sources that allow the controls needed to examine causal patterns convincingly.

1.1 *Socioeconomic Status and Health*

Much effort is given to specify and test pathways through which social stratification indicators affect health, focusing on a wide range of health-related behaviors, environmental risks, and personal traits and dispositions. Link and Phelan (1995), in contrast, building on Durkheim's (1951) classic analysis of suicide, have presented data in support of the argument that socioeconomic differences are fundamental causes of poor health outcomes because those who are more advantaged in any situation or historical period will have greater knowledge of, or be better able to manipulate, the causal influences in their favor. Thus, while the pathways operative in any situation or historical period are specific to the circumstances, the knowledge, power, and resources characteristic of higher SES that provide advantage are more general and fundamental determinants of health across time and social circumstances.

1.2 *Gender and Health*

Gender studies in health have also advanced, consistent with their growing importance in sociology more generally. Medical sociologists have a longstanding interest in gender differences in illness, death, and health care utilization. The key challenges have been to explain the female advantage in length of life; the contrasting prevalence of specific disorders among men and women—with particular focus on the high prevalence of depression among women; and different patterns of illness behavior, informal uses of social support, and help-seeking (Mechanic 1978). Studies of mortality generally show that while women may have a biological advantage in longevity, much of the variation in mortality rates can be explained by differential exposure of men and women to the major causes of mortality. These causes are substantially affected by culture, health habits, and exposure to risk.

While the medical sciences define depression as a discrete entity, sociologists have been intrigued by the fact that the female excess of depression tends to be balanced by the male excess of substance abuse, antisocial behavior, and violence. This observation is consistent with a nonmedical conception that views disorders as alternative modes of expressing distress and tension and not simply discrete entities with independent causes. Research on variations in illness behavior tests a range of hypotheses concerning symptom perception and identification, willingness to acknowledge distress and request assistance, and differential barriers and facilitative factors in seeking care for men and women. Patterns of response tend to be specific to disorders and types of services studied, and studies of utilization of care among men and women need to be especially sensitive to the large differences in the use of services related to reproductive health.

1.3 Age and Health

Age, like gender, is a crucial biological and social determinant of many important health outcomes. Aging is associated with increased prevalence of diseases and disability but varying birth cohorts experience different health outcomes at any given age, reflecting changes in social and cultural factors, social developmental processes, lifestyles and behavior, and the availability and effectiveness of medical technology. Debate about the consequences of increased longevity for health is a substantial focus. One view is that the mortality curve increasingly follows a rectangular pattern whereby successive cohorts experience more disability-free years prior to death. A contrasting view focuses on the increasing prevalence of chronic illness and disability with extension of life and increased disability life years among the elderly. The literature provides some support for both perspectives but no clear resolution of the issue. Age is clearly associated with increased prevalence of chronic disease and disability but researchers now more clearly differentiate between the consequences of the aging process itself and the consequences of disease and are less likely to attribute disease-associated impairments to aging.

1.4 Race, Culture, and Ethnicity and Health

There is much interest in differential health patterns by race, culture, and ethnicity. Such research, in contrast to research based on other stratification indicators, is more descriptive of particular groups in time and place and less based on theory. This is inevitable given the diversity of cultural groups and rapid acculturation. Indeed, the categories used to describe race and ethnic groups in most studies and databases (nonwhite, Hispanic origin, etc.) are themselves heterogeneous and their usefulness is in dispute. The literature consistently describes the poorer health outcomes of disadvantaged race/ethnic groupings. In the US literature there is much interest in the persistent lower birthweight rates and higher infant mortality among nonwhites (mostly African Americans), as compared with whites, and the extent to which these differences are attributable to socioeconomic differences versus other factors. There is also much interest in why some disadvantaged subgroups achieve relatively good health outcomes despite poverty and many adversities.

2. The Study of Health Care Organization and Provision

It is well established that national systems of health care develop from their unique historical trajectories, national cultures, political systems, and economic status. But medical care, and the systems in which they are embedded, are part of worldwide networks, sharing knowledge, technology, and organizational innovations. In addition, common trends among developed nations in demography, urbanization, epidemiology of disease, medical demand and rising public expectations, and the growing gap between medical possibilities and the ability or will to fund them, pose many common problems for health systems (Mechanic and Rochefort 1996). These problems include means of promoting health, organizing levels of care, expanding health care access, linking health and social services, and cost containment and health care rationing. Most organizational studies on health care systems bear on aspects of each of these problems as they relate to the organization of particular health care systems or subsystems. There have also been efforts to study health care systems comparatively across nations.

2.1 Changing Role of the Physician

Sociological work on the organization of care has focused on the health professions—and particularly the changing role of the physician, the politics of health care dominance and authority, intraorganizational networks and relationships, and integrating levels and sectors of care. There has also been much interest in how incentives and other organizational factors affect processes and quality of care. Much attention has been given to structures and processes of care in particular types of organizations such as hospitals, intensive care units, perinatal units, nursing homes, and health maintenance organizations, and to the role of changing technologies.

Much excellent sociological work has been focused on the development of the medical profession and the

means used to achieve dominance over medical work and other health occupations (Freidson 1970). Early work that sought to define the special characteristics of the medical profession has been succeeded by more dynamic analyses of how professions and occupations compete to gain control over varying spheres of work (Abbott 1988). As the authority and control over organizational decisions have devolved from physicians to managers and administrators with the corporatization of medicine (Starr 1982), sociologists have been concerned with whether, and in what ways, the changes in autonomy and control contribute to the proletarianization of physicians.

2.2 Managed Care

In the USA since about 1990 most sociological research on health organization has been focused on the growth of managed care and its variety of new organizational forms. Although organizational forms differ widely, from staff and group model prepaid practices to a variety of network models, they share in common certain strategies that are of great research interest. These strategies include the use of capitation and transfer of risk through organizational levels, the use of utilization review to limit the provision of unneeded care, and new systems of constraint on physicians, including practice guidelines, disease management programs, and physician profiling. With the growth of health care organized by private corporations, there is renewed interest in the consequences of nonprofit versus for- profit forms of organization.

Much of the research on managed care relates to the effects of such management on plan and physician choice, access to care, use of inpatient and outpatient services, satisfaction and trust, and health outcomes. Studies are diverse, ranging from how physicians and utilization reviewers communicate and negotiate over care to how physician choice and continuity of care affect patient satisfaction. Methods of study range widely, from conversational analysis and participant observation to experimental studies and surveys. The survey is the most common approach to studying managed care practices but there is increasing use of large administrative data sets to better understand patterns of care and to address specific hypotheses about how insurance coverage, social characteristics, and patterns of illness affect utilization and costs of care.

The most important changes in US health care include the growing corporatization and privatization of health care, the shift from fee-for-service to capitated health insurance plans, and the imposition of rationing devices such as gatekeepers to specialized services, utilization management, and payment incentives for physicians to meet utilization targets. Other central issues involve initiatives to achieve organizational and clinical integration of services, to develop structures more appropriate to treat chronic disease and provide longterm care, and to better coordinate health and other related services for persons with severe and persistent disabilities. Deinstitutionalization trends in developed countries keep clients with mental illness and developmental disabilities, impaired elderly, and other persons with disabilities in their own homes or communities. These trends create profound challenges to the reproduction of, in decentralized community settings, the types and range of services available in total institutions (Mechanic 1999). Although many promising programs and innovations have been developed, most studies report significant barriers to achieving cooperation among service sectors with varying eligibility criteria, bureaucratic cultures, priorities, and reward systems. Seemingly simple in theory, community integration of health services continues as a significant challenge in reality.

3. The Social Psychology of Health and Health Care

The study of stress, coping and social support has been central in medical sociological research and constitutes perhaps the best integrated body of knowledge tied together by common theoretical interests and methodological concerns. This area also interlinks work in medical sociology with efforts in psychosocial epidemiology and health psychology.

3.1 The Stress and Coping Process

The basic paradigm is simple and posits that stressful life events cause negative health outcomes when individuals lack the counterharm resources needed to deal with the challenges embedded in the stress event (Lazarus 1966). Counterharm resources include the ability to deal with the instrumental challenge and to manage the emotions evoked by the situation. Social support is seen as having either a buffering effect on the influence of stress events on outcomes, a more direct positive effect, or both. A measure of sense of personal control (or a related proxy such as low helplessness or mastery) is also often included in these research models and is found to contribute to more positive outcomes. There are a variety of models elucidating the pathways of the stress process and the relative influence of different intervening variables.

The basic ideas informing the model have been in the literature since the 1960s but much effort has been devoted to refining the model and related hypotheses, developing improved measures, and perfecting the quality of data and analytic approaches. Initial measures of life events were crude and ambiguous. Although the underlying theory was that life events called for psychological and somatic readjustments that strained the organism, and thus positive life changes were significant as well as negative events, the

scales combined and confused these items and did not allow a clear test. Moreover, studies treated meaning in a confused way or completely ignored it, did not carefully specify dependent measures of illness, and commonly made incorrect causal inferences from cross-sectional data.

Many important questions persist in the stress-coping literature despite vast improvement in measurement, data quality, and analysis. It is difficult empirically to capture the fact that many stressful life events are not single occurrences but part of a sequence of related stressors. Some researchers have focused on the role of more minor but persistent events such as commonplace daily hassles. Although there has been broad recognition that the meaning of events is central, most investigators have not measured meaning in a manner independent of the phenomena they wish to predict. Brown and Harris (1978) have developed a novel approach to contextual analysis in which detailed descriptions of the event are obtained and trained raters attribute meaning to the event, on the basis of their assessment of how the average person would respond to an event in that context, following a carefully specified protocol.

Focus on the availability of social networks and social support has become increasingly important, with growing sophistication about the complexity of the processes involved. Researchers are now much more aware that persons who attract support are quite different from those who do not, and that selective processes are already evident early in life. Thus, the relative role of social support, as compared with the personal attributes that attract support, is more difficult to assess. Similarly, researchers in the area are beginning to take note of the obligations inherent in support networks and possible deleterious aspects.

The most difficult component of stress and coping research has been the measurement of coping. Most such measures, such as those that focus on task-oriented coping versus denial, are quite limited. However, there is a rich tradition in medical sociology of intensive qualitative study of coping processes in a wide range of illness and disability conditions, including impending death (Fox 1988). Many of these studies proceed on the methodology of grounded theory and seek to understand the stages of illness and how illness trajectories vary. Such studies focus on how people become aware of their illness, the attributions they make, the uses of informal and formal sources of care, how coping changes at different stages of illness, and how persons affected interact with family, friends, and health care providers. Such studies often lack the generality and rigor of large quantitative population studies but they provide a richness of understanding that is not captured in the measurement systems used in the larger studies. Coping is a complex, iterative process, with causal sequences occurring in close temporal proximity, which is almost impossible to capture in the usual survey.

3.2 Disability

The study of disability has become increasingly important as the disability rights movement has advanced in many countries. The ideology of the movement has been built around a sociological conception of disability which maintains that impairments may or may not result in disability, depending on social organization and the physical environment. In essence, whether an impairment becomes a disability depends on physical access to community facilities, social attitudes, and discrimination with respect to employment and other types of social participation, and the types of accommodations the community is willing to make. These, in turn, affect clients' attitudes, motivation, participation, and productivity. At a macro level, sociologists and economists study the relationship between the economy and use of the formal disability insurance system. Disability status varies as economic conditions and other community characteristics change.

4. Other Areas of Inquiry

Because medicine can be both a source of data to illuminate sociological hypotheses and a basis for addressing applied problems, research in the area ranges very widely across sociological and medical areas. From a broad sociological perspective, illness is a form of deviance and medicine an institution of social control (Parsons 1951). Illness is often manipulated by individuals to seek release from usual obligations and other secondary gains, to excuse failure, and to gain social support. Community agents also seek to control illness definitions to insure societal productivity and protect the public purse from demands that would be too burdensome. In this sense, definitions of illness are differentially constructed by agents who have varying needs, ranging from the individual client seeking certain advantages to government agencies attempting to constrain entitlement payments.

From the standpoint of sociology, medicine has been a basis for studying occupational socialization, professional organization and control, intraorganizational processes, power and influence, demographic processes, political conflict, collective behavior, and much more. From a medical perspective, sociological investigation has contributed to the study of every major disease category, examining etiological and risk factors and other influences on the course of disease and social responses and adaptations. Medicine has also been a context for methodological innovation in experimental approaches, use of surveys, measurement approaches, and analytic methods.

See also: Disability: Sociological Aspects; Doctor–Patient Interaction: Psychosocial Aspects; Health

Behavior: Psychosocial Theories; Health Behaviors; Health Care Delivery Services; Health Care Organizations: For-profit and Nonprofit; Health Care, Rationing of; Health Care Technology; Health Economics; Health Policy; Medical Geography; Medical Profession, The; Medicalization: Cultural Concerns; Professions, Sociology of; Public Health; Public Health as a Social Science; Socioeconomic Status and Health; Socioeconomic Status and Health Care; Stress and Health Research

Bibliography

Abbott A 1988 *The System of Professions: An Essay on the Division of Expert Labor*. University of Chicago Press, Chicago

Amick III B C, Levine S, Tarlov A R, Walsh D C W (eds.) 1995 *Society and Health*. Oxford University Press, New York

Brown G W, Harris T 1978 *Social Origins of Depression: A Study of Psychiatric Disorder in Women*. Free Press, New York

Durkheim E [1930] 1951 *Suicide*. Spaulding J A, Simpson G (trans). Free Press, New York

Fox R C 1988 *Essays in Medical Sociology: Journeys into the Field*. Transaction Books, New Brunswick, NJ

Freeman H E, Levine S (eds.) 1989 *Handbook of Medical Sociology*, 4th edn. Prentice Hall, Englewood Cliffs, NJ

Freidson E 1970 *Profession of Medicine: A Study of the Sociology of Applied Knowledge*. Dodd, Mead and Company, New York

Lazarus R S 1966 *Psychological Stress and the Coping Process*. McGraw Hill, New York

Link B G, Phelan, J 1995 Social conditions as fundamental causes of disease. *Journal of Health and Social Behavior*. Special issue: 80–94

Mechanic D 1978 *Medical Sociology*, 2nd edn. Free Press, New York

Mechanic D 1999 *Mental Health and Social Policy: The Emergence of Managed Care*, 4th edn. Allyn and Bacon, Boston

Mechanic D, Rochefort D A 1996 Comparative medical systems. *Annual Review of Sociology* **22**: 239–70

Parsons T 1951 *The Social System*. Free Press, New York

Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York

Wilkinson R G 1996 *Unhealthy Societies: The Afflictions of Inequality*. Routledge, London

D. Mechanic

Medicalization: Cultural Concerns

One of the abiding interests of social scientists concerned with modernization and its effects, particularly those who took up the legacy of Emile Durkheim and Talcott Parsons, has been to show how social order is produced and sustained in contemporary society. In this vein the sociologist Irving Zola argued in the early 1970s that medicine was becoming a major institution of social control, replacing the more 'traditional' institutions of religion and law, resulting in the 'medicalizing' of much of daily life in the name of health. Zola's publication, by no means totally opposed to the process he highlights, gave birth to a genre of research in which the cumbersome word medicalization— 'to make medical'— was adopted as a key concept (Zola 1972).

It can be argued that medicalization commenced many hundreds of years before Zola's observation, from the time that healing specialists were first recognized among human groups. With the consolidation between approximately 250 BC and 600 AD of the several literate medical traditions of Europe and Asia that remain powerful today, healers were sought out to purge the stresses of everyday life in addition to dealing with physical malfunction. Common to these literate medical traditions is an emphasis on self-discipline, vigilance, and virtue as essential for sustained good health, and the function of healers is to restore order to the bodies of those patients no longer in balance with the social and environmental milieu.

1. Modernization and Medicalization

With modernity several fundamental shifts can be detected in the form taken by medicalization, most notably the involvement of the state in the systematic monitoring of the health of populations. Beginning in the seventeenth century, European and North America modernization fostered an 'engineering mentality' one manifestation of which was a concerted effort to establish increased control over the vagaries of the natural world through the application of science. As a result, by the eighteenth century, health came to be understood by numerous physicians and by the emerging middle classes alike as a commodity, and the physical body as something that could be improved upon. At the same time, legitimized through state support, the consolidation of medicine as a profession was taking place, together with the formation of medical specialties and the systematic accumulation, compilation, and distribution of new medical knowledge. Systematization of the medical domain, in turn, was part of a more general process of modernization to which industrial capitalism and technological production was central, both intimately associated with the bureaucratization and rationalization of everyday life.

Medicalization expanded in several directions during the eighteenth and nineteenth centuries. First, there was an increased involvement on the part of medical professionals in the management not only of individual pathology, but of life-cycle events. Birth had been entirely the provenance of women, but from

the early eighteenth century, in both Europe and North America, male midwives trained and worked at the lying-in hospitals located in major urban centers to deliver the babies of well-off women. These *accoucheurs* later consolidated themselves as the profession of obstetrics. By the mid-nineteenth century other life-cycle transitions, including adolescence, menopause, aging, and death had been medicalized, followed by infancy in the first years of the twentieth century. In practice, however, large segments of the population remained unaffected by these changes until the mid-twentieth century.

Another aspect of medicalization can be glossed in the idiom of 'governmentality' proposed by Michel Foucault. With the pervasive moves throughout the nineteenth century by the state, the law, and professional associations to increase standardization through the rational application of science to everyday life, medicine was integrated into an extensive network whose function was to regulate the health and moral behavior of entire populations. These disciplines of surveillance, the 'bio-politics of the population,' as Foucault described it, function in two ways. First, everyday behaviors are normalized so that, for example, emotions and sexuality become targets of medical technology, with the result that reproduction of populations and even of the species are medicalized. Other activities, including breastfeeding, hygiene, exercise, deportment, and numerous other aspects of daily life, are medicalized—largely by means of public health initiatives and with the assistance of the popular media.

Medical and public health management of everyday life is evident not only in Europe and North America, but also in nineteenth-century Japan and to a lesser extent in China. In India, Africa, South East Asia, and parts of the Americas, medicalization is intimately associated with colonization. Activities of military doctors and medical missionaries, the development of tropical medicine and of public health initiatives, designed more to protect the colonizers and to 'civilize' the colonized than to ameliorate their health, are integral to colonizing regimes. As with medicalization elsewhere, large segments of the population remained untouched by these activities (Arnold 1993).

From the late eighteenth century, another arm of medicalization became evident. Those populations labeled as mentally ill, individuals designated as morally unsound, together with targeted individuals living in poverty were for the first time incarcerated in asylums and penitentiaries where they were subjected to what Foucault termed 'panopticism.' Inspired by Jeremy Bentham's plans for the perfect prison in which prisoners are in constant view of the guards, the Panopticon was, for Foucault, a mechanism of power reduced to its ideal form—an institution devoted to surveillance.

These changes could not have taken place without several innovations in medical knowledge, technologies, and practice, among which four are prominent. First, the consolidation of the anatomical and pathological sciences whereby the older humoral pathology is all but eclipsed so that belief in individualized pathologies is essentially abandoned in favor of a universal representation of the normal body from which sick bodies deviate. Second, the introduction of the autopsy enabling systematization of pathological science. Third, routinization of the physical examination and of the collection of case studies. Fourth, the application of the concept of 'population' as a means to monitor and control the health of society, central to which is the idea of a norm about which variation, which can be measured statistically, is distributed. The belief that disease can be understood both as individual pathology and as a statistical deviation from a norm of health becomes engrained in medical thinking as a result of these changes. Treatment of pathology remains, as was formerly the case, the core activity of clinical medicine, but the new epistemology of disease causation based on numeration gradually gained ground. Public health and preventive medicine, always closely allied with the state, made the health of populations its domain.

Other related characteristics of medicalization, well established by the late nineteenth century, and still evident today, can be summarized following Nikolas Rose (1994), into 'dividing practices,' whereby sickness is distinguished from health, illness from crime, madness from sanity, and so on. Following this type of reasoning certain persons and populations are made into objects of medical attention and distinguished from others who are subjected to different authorities including the law, religion, or education. At the same time various 'assemblages' are deployed—a combination of spaces, persons, and techniques, that constitute the domain of medicine. These assemblages include hospitals, dispensaries, and clinics, in addition to which are the home, schools, the army, communities, and so on. Recognized medical experts function in these spaces making use of instruments and technologies to assess and measure the condition of both body and mind. The stethoscope, invented in the early nineteenth century, was one such major innovation, the first of many technologies that permit experts to assess the condition of the interior of the body directly, rendering the patient's subjective account of malaise secondary to the 'truth' of science.

Several noted historians and social scientists argue that from the mid-nineteenth century, with the placement in hospitals for the first time not only of wealthy individuals but of citizens of all classes, the medical profession was able to exert power over passive patients in a way never before possible. This transition, aided by the production of new technologies, has been described as medical 'imperialism.' Certain researchers limit use of the term medicalization to these particular changes whereas other scholars insist that the development of hospitalized patient populations is just one

aspect of a more pervasive process of medicalization, to which both major institutional and conceptual changes contribute. Included are fundamentally transformed ideas about the body, health, and illness, not only among experts but also among the population at large.

2. The Medicalization Critique

In writing a review article of medicalization Conrad (1992) states that over the previous 20 years the term had been used most often as a critique of inappropriate medicalization rather than simply to convey the idea that something had been made medical. Sociological literature of this period argued uniformly that health professionals had become agents of social control. This position was influenced by the publications of Thomas Szasz and R. D. Laing in the 1960s in connection with psychiatry where they insisted that the social determinants of irrational behavior were being neglected in favor of an approach dominated by a biologically deterministic medical model. Zola, Conrad, and others argued in turn that alcoholism, homosexuality, hyperactivity, and other behaviors were increasingly being biologized and labeled as diseases. While in theory this move from 'badness to sickness' ensured that patients could no longer be thought of as morally culpable, it permitted medical professionals to make judgments that inevitably had profound moral repercussions.

Once it became clear that life-cycle transitions and everyday behaviors were increasingly being represented as diseases or disease-like, a reaction set in during the 1970s against medicalization, particularly among sociologists and feminists. It was Ivan Illich's stinging critique of scientific medicine in *Medical Nemesis* that had the greatest effect on health care professionals and the public at large. Illich argued that, through overmedication, biomedicine itself inadvertently produces iatrogenic side effects (an argument that no one denies) and, further, that the autonomy of ordinary people in dealing with health and illnesses is compromised by medicalization.

In the 1970s a majority of feminist critics characterized medicine as a patriarchal institution because, in their estimation, the female body was increasingly being made into a site for technological intervention in connection with childbirth and the reproductive life cycle in general. On the other hand it has also been argued within the feminist movement that insufficient attention is paid to women's health, and that medical research and the development of medications have been directed primarily at the diseases most common to men. The white male body has been taken as the standard for all. This countercurrent in the feminist movement has, in the long run, proved to be the more robust and has contributed to the line of argument most common in current publications on medicalization in which the question of individual agency is central and a rapprochement with medicine is sought out.

In contemporary writing it is common to assert that members of the public are not usually made into victims of medical ascendancy and, in any case, that to cast women in a passive role is to perpetuate the very kinds of assumptions that feminists have been trying to challenge. Although active resistance to medicalization has contributed to the rise of the home-birth movement and to widespread use of alternative therapies and remedies of numerous kinds, empirical research makes it clear that the responses of individuals to the availability of biomedical interventions are pragmatic, and based upon what are perceived to be in the best interests, not only of women themselves, but often of their families and at times their communities (Lock and Kaufert 1998). This is particularly evident in connection with reproduction when, for example, women who believe themselves to be infertile make extensive use of new reproductive technologies, despite the high failure rate and the expense and emotional upheaval that is involved. In Europe and North America resort to these technologies may be primarily to fulfill individual desires, but in other locations more may be at stake. Women can be subjected to ostracism, divorced, or thrown into the street, if they do not produce a healthy child, in many situations preferably a male child. Even when these drastic measures are not applied, it is clear that the majority of women internalize the norm that their prime task in life is to reproduce a family of the ideal size and composition, and that failure to do so diminishes them in the eyes of others. Under these circumstances it is not surprising that a pragmatic approach to medical technology and medical services is much more common than is outright resistance to these products of modernization. At times, as in the case of the breast cancer movement, AIDS, or where toxic environments are at issue, people unite to fight for more effective medical surveillance. Under these circumstances, the knowledge and interests of users results in an expansion of medicalization.

3. Medicalized Identities

Social science critiques of medicalization, whether associated more closely with labeling theory and the social control of deviance, or with Foucaldian theory and the relationship of power to knowledge, have documented the way in which identities and subjectivity are shaped through this process. When individuals are publicly labeled as schizophrenic, anorexic, menopausal, a heart transplant, a trauma victim, and so on, transformations in their subjectivity are readily apparent. At times medicalization may function to exculpate individuals from responsibility for being sick and thus unable to function effectively in society.

Medicalization is not limited to sickness and 'deviance' however. Wellness—the avoidance of disease and illness, and the 'improvement' of health—is today a widespread 'virtue,' notably among the middle classes.

As part of modernity, practices designed to support individual health have been actively promoted for over a century, and are now widely followed among the population at large. The sight of jogging wives of fishermen in Newfoundland's most isolated communities is testimony to this. The individual body, separated from mind and society, is 'managed' according to criteria elaborated in the biomedical sciences, and this activity becomes one form of self-expression. Body aesthetics are clearly the prime goal of some individuals, but a worry about the 'risk' of becoming sick is at work for the majority. By taking personal responsibility for health, individuals display a desire for autonomy and in so doing they actively cooperate in the creation of 'normal,' healthy, citizens, thus validating the dominant moral order (Crawford 1984). Health is commoditized.

As evidence is amassed to demonstrate conclusively how social inequity and discrimination of various kinds contribute massively to disease, ranging from infections to cancer, the idea of health as virtue appears increasingly out of place. Due to poverty large segments of the population in most countries of the world have shorter life expectancies and a greater burden of ill-health than do their compatriots. The pervasive value of individual responsibility for health enables governments to narrow their interests to economic development, while ignoring redistribution of wealth and the associated cost for those individuals who, no matter how virtuous they may be, are unable to thrive.

4. *Risk as Self-governance*

Activities designed to assist with the avoidance of misfortune and danger are ubiquitous in the history of humankind, but the idea of being at 'risk' in its technical, epidemiological meaning is a construct of modernity. In theory morally neutral, risk provides a means whereby experts can distance themselves from direct intervention into people's lives while employed the agency of subjects in their own self-regulation through 'risk-management.' Among the numerous examples of this process, the transformation of aging, in particular female aging, into a discourse of risk is illustrative. Given that women live longer than men it seems odd that female aging has been targeted for medicalization, but this move is in part driven by a fear of the enormous expense to health care budgets that very old infirm people, the majority of them women, are believed to incur.

Medicalization of female middle age, and in particular the end of menstruation, commenced early in the nineteenth century but it was not until the 1930s, after the discovery of the endocrine system, that menopause was represented in North America and Europe as a disease-like state characterized by a deficiency of estrogen. In order to sustain this argument, the bodies of young premenopausal women must be set up as the standard by which all female bodies will be measured. Postmenopausal, postreproductive life can then be understood as deviant. This 'expert knowledge' is buttressed through comparisons between human populations and those of other mammals, where postmenopausal life is very unusual. The arguments of biological anthropologists that older women are essential to the survival of highly dependent human infants and their mothers in early hominid life are ignored. Moreover it is argued erroneously that women have lived past the age of 50 only since the turn of the twentieth century, and that postreproductive life is due entirely to improved medical care and living conditions.

Today older women are warned repeatedly about heart disease, osteoporosis, memory loss and Alzheimer's disease, and numerous other conditions for which they are said to be at increased risk due to their estrogen-deficient condition. Daily medication on a permanent basis with medical monitoring has been recommended by gynecological organizations in many countries for virtually all older women, although some reversals of these blanket suggestions are now taking place. Few commentators deny that drug company interests have contributed to this situation. Misleading interpretations of often poorly executed epidemiological research create confusion about estimates of risk for major disease in women who are past menopause.

Furthermore, cross-cultural research indicates that the common bodily experiences of populations of middle-class North Americans at menopause is significantly different from that of women in other parts of the world, a situation to which local biologies and culture both contribute, making medicalization of this part of the life cycle exceedingly problematic (Lock 1993). Nevertheless, many thousands of women willingly medicate themselves daily in the firm belief that aging is pathology. Among them a few, but we cannot predict who exactly, may avert or postpone the onset of debilitating and life-threatening diseases, while others may hasten their death.

The combination of the emerging technologies of the new molecular genetics with those of population genetics is currently opening the door to an exponential growth in medicalization in the form of what has been termed 'geneticization.' As genetic testing and screening of fetuses, newborns, and adults becomes increasingly institutionalized, the possibilities for surveillance are boundless, particularly so because proponents of genetic determinism promote the idea that genes are destiny. Our genes are increasingly thought of as 'quasi-pathogens' that place us at

increased risk for a spectrum of diseases. We are all deviants from the normative standards set by the mapping of the human genome, itself an abstraction that corresponds to the individual genome of no one.

This biomedicalization of life itself comes with extravagant promises about our ability in the near future to harness nature as we wish, the enhancement of human potential, and access to the knowledge that makes it possible to know what it is that makes us uniquely human. Based on the results of genetic testing, a *laissez-faire* eugenics, grounded in individual choice and the inalienable right to 'health' for our offspring, is already taking place. Of course, suffering may be reduced if, for example, some families choose to abort fetuses with Tay Sach's disease. On the other hand, how can discrimination of the basis of genetics, already evident in some workplaces, insurance companies, and the police force be controlled? Should entrepreneurial enterprises have a monopoly over the development and testing for multifactorial diseases such as breast and prostate cancer as is currently the case? How should the findings of these tests be interpreted by clinicians and the public when all that can be surmised if the results are positive is the possibility of elevated risk for disease in the future? Negative results do not indicate that individuals will not contract the disease. Who will store and have access to genetic information, and under what conditions? And who, if anyone, may patent and commodify genetic materials and for what purpose? The new genetics requires urgent answers to such questions.

It has been proposed that we are currently witnessing a new '*bio*medicalization' associated with late modernity or postmodernity (Clarke et al. in press). A technoscientific revolution is taking place involving increased molecularization, geneticization, digitization, computerization, and globalization, and these changes are in turn associated with a complete transformation of the organization (including the privatization of a great deal of research), expertise, and practices associated with the medical enterprise. In this milieu the potential exists to make the body increasingly a site of control and modification carried out in the name of individual rights or desire.

Such modifications are often highly profitable to involved companies and frequently correspond to the utilitarian interests of society, one prime objective is to save on health care expenditure. Risk estimates permit the escalation of anxiety among targeted populations as when, for example, the public is repeatedly informed that one in eight women will contract breast cancer, and one in 25 will die of this disease. The one in eight estimate is incorrect because it is based on the cumulative probability of sustaining the disease over an entire lifetime, that is, somewhere between birth and 110 years of age; even for elderly women the probability is never as high as one in eight. Nevertheless this figure induces women to submit to breast cancer testing that then raises a new round of anxieties and uncertainties.

A danger exists of overestimating the consequences of technological innovation and associated biomedicalization. A large proportion of the world's population effectively remains outside the reach of biomedicine. That medication for HIV/AIDS is not universally available is a gross injustice, as is the increasing incidence of antibiotic-resistant tuberculosis that could be controlled efficiently if biomedicine was competently practiced. At the other extreme, despite an enormous promotion of hormone-replacement therapy, research shows that less than a quarter of the targeted female population takes this medication, and even then not consistently. It is also clear that many women do not make use of selective abortion should a fetal 'flaw' be found through genetic testing.

Medicalization understood as enforced surveillance is misleading. So too is an argument that emphasizes the social construction of disease at the expense of recognizing the very real, debilitating condition of individuals who seek out medical help. Rather, an investigation of the values embedded in biomedical discourse and practice and in popular knowledge about the body, health, and illness that situate or resist various states and conditions as residing within the purview of medicine better indicates the complexity at work.

See also: Biomedical Sciences and Technology: History and Sociology; Depression, Clinical Psychology of; Health: Anthropological Aspects; Health Policy; Medical Profession, The; Medicine, History of; Psychiatry, History of; Public Health as a Social Science

Bibliography

Arnold R 1993 *Colonizing the Body: State Medicine and Epidemic Disease in Nineteenth-century America*. University of California Press, Berkeley, CA

Clark A E, Mamo L, Shim J K, Fishman J R, Fosket J R 2000 Technoscience and the new biomedicalization: Western roots, global rhizomes. *Sciences Sociales et Santé* **18**(2): 11–42

Conrad P 1992 Medicalization and social control. *Annual Review of Sociology* **18**: 209–32

Crawford R 1984 A cultural account of 'health': Control, release and the social body. In: McKinlay J B (ed.) *Issues in the Political Economy of Health*. Methuen-Tavistock, New York pp. 60–106

Illich I D 1976 *Medical Nemesis*. Pantheon, New York

Lock M 1993 *Encounters with Aging: Mythologies of Menopause in Japan and North America*. University of California Press, Berkeley, CA

Lock M, Kaufert P 1998 *Pragmatic Women and Body Politics*. Cambridge University Press, Cambridge, UK

Rose N 1994 Medicine, history, and the present. In: Jones C, Porter R (eds.) *Reassessing Foucault: Power, Medicine and the Body*. Routledge, London
Zola I 1972 Medicine as an institution of social control. *Sociological Review* **20**: 487–504

M. Lock

Medicine, History of

At all times and in all cultures attempts have been made to cure diseases and alleviate suffering. Healing methods nowadays range from experience-based medical treatment within the family and the lay world to the theories and methods of experts. Medicine in a narrow sense refers to 'scientific' knowledge. Such special knowledge is distinct from other forms of knowledge following culturally characteristic criteria. The physician's appropriate action is derived from medico–scientific knowledge: we may only speak of 'medicine' and 'physician' when diseases are not primarily seen in magical, mythical, religious, or empirical terms. Medicine developed as soon as diseases came to be explained 'according to nature' and treated with scientific rationality. Historically, this took place in the fourth century BC in ancient Greece. Scientific medicine developed from these ancient beginnings. This concept of modern Western medicine has come to be the leading model of medicine worldwide. At the same time, there have always been alternative understandings and methods of medicine, which will also be mentioned briefly in this article. Concepts of medicine comprise a continuous causal interrelation between a particular physiology, a particular pathology and its deducible therapy. The inner perspective of the scientific and conceptual history of Western medicine will be outlined in Sect. 1. This view will then be a perspective guided by socio–historical, historico–sociological, and anthropological approaches in Sect. 2. Sect. 3 provides a conclusion and perspective on future developments.

1. *The Scientific and Conceptual History of Modern Western Medicine*

The magical–mythical art of healing of indigenous cultures is part of a world in which human relationships with nature and with societal structures are not understood as objects in their own right. In hunter–gatherer societies and in the early agriculturally determined cultures, diseases were seen as the results of violated taboos, offended demons or neglected ancestors. Therapies aimed at restoring the old order: exorcism was practiced, demons and ancestors were pacified. With this, the spiritual, physical, and social order were restored. This animistic–demonic medicine presented a closed rational system of thinking and acting.

Magical conceptions of the world were not confined to prescientific times. Rather, iatro-magic ('iatro' from ancient Greek iatros, 'the physician') influenced medical theory and even practical medical action far into the seventeenth century. Even today, simile magic (e.g., the use of ginseng roots) or singularity magic (e.g., the use of charms, symbols, spells) are found in popular medicine, nature healing, and so-called complementary medicine.

The highly developed medicine of the ancient cultures of Mesopotamia, Egypt, and Asia Minor was characterized by a religious world order. This order was underpinned by (national) deities and the (divine) ruler. Diseases were seen as conscious or unconscious violations of the world order with which the healers were familiar. These healers possessed practical as well as prophetic skills. Thus physicians and priests were hardly distinguishable. There was an abundance of empirical practices both in magical–mythical medicine and in the archaic art of healing. The one criterion that clearly distinguishes the latter from medicine in the narrow sense was that the entire art of healing remained metaphysically oriented. Even those treatments which worked in the modern sense as well were in their time assigned to a magical–mythical, i.e., religious, domain. This also applied to rules of hygiene such as sanitary and dietary laws.

It was Hippocrates of Kos (ca. 460–375 BC) who took the decisive step towards scientific medicine of classical antiquity. Hippocrates and his school emancipated medicine from religious concepts. Diseases came to be explained 'according to nature.' Natural philosophical elements (fire, water, air, earth) and their qualities (warm, humid, cold, dry) were combined with the teachings of the body fluids (blood, phlegm, yellow and black bile) to produce a theory of humoral pathology. Apart from this concept, ancient medicine also developed a wide spectrum of theories and methods associated with various philosophical schools. The spectrum ranged from the pathology of the body fluids (humoral pathology, humoralism) to the pathology of solid body parts (solid pathology). This included, for example, functional anatomy (including vivisections), surgery (e.g., vascular ligatures), and pharmacy (as practiced by Dioscorides Pedanius, first century BC). The Alexandrine medical schools of Herophilus of Chalcedon (ca. 330–250 BC) and Erasistratus of Kos (ca. 320–245 BC) followed empirical, partly experimental concepts in the modern sense. With that, almost all possible patterns of modern medical thinking had already been put into practice, or at least anticipated, in classical antiquity.

Galen of Pergamum (AD 130–200) subsumed the theory and practice of ancient medicine under a hum-

oral pathological view. Authoritative Hippocratic–Galenian medicine entered Persian and Arabian medicine via Byzantium (e.g., Onbasius of Pergamum, ca. 325–400, and Alexander of Tralles (ca. 525–600). As a result of Islamization, Persian–Arabian medicine spread eastward to South-east Asia and westward to the Iberian Peninsula. In the heyday of Arabian–Islamic culture between the seventh and twelfth centuries, ancient medicine was transformed into practice-oriented compendia in a process of independent scientific work. Far beyond the high Middle Ages, the *Canon Medicinae* of Ibn Sina (980–1037) (known in the West as Avicenna) was the handbook even of Western European physicians.

For approximately 1,500 years humoralism was the dominating medical concept from India to Europe. The original theory of the four body fluids, qualities, and elements was constantly expanded: by the four temperaments (sanguine, phlegmatic, choleric, melancholic), by the four seasons, the 12 signs of the zodiac, then later by the four evangelists, certain musical scales, etc. This concept allowed the physician to put the patient with his/her respective symptoms at the center of a comprehensive model: the powers of the macrocosm and the microcosm influenced one another within the individual on a continuous basis. By means of thorough anamnesis, prognoses, and dietary instructions ('contraria contrariis'; allopathy), physicians supported the healing powers of nature ('vis medicatrix naturae'; 'medicus curat, natura sanat'), a model which—according to the principle of 'similia similibus'—is still valid today in homeopathy.

With the political decline of the Western Roman Empire during the fifth and sixth centuries, a Christian art of healing came to the fore in Western Europe. In this, the traditions of ancient practice (Cassiodor, 487–583) mingled with a magic-influenced popular art of healing and Christian cosmology to create a world of 'Christus Medicus.' Such a theological–soteriological concept of medicine was able to take up the ancient cult of Asclepius. In early Christian and medieval iatro-theology diseases were interpreted as the result of sin or as a divine act of providence. This still continues to have an effect on contemporary religious practices, such as the use of votive tablets, intercessions, and pilgrimages. In her medical writings (e.g., *Physica*; *Causae et Curae*), Hildegard of Bingen (1098–1179) left abundant evidence of the medieval art of healing.

The so-called monastic medicine, or pre-Salernitarian medicine (sixth to twelfth century) was superseded in southern Italy and Spain by the adoption of ancient traditions from Arabian medicine. Beginning in Salerno in the eleventh century with Constantinus Africanus (1018–87) and Toledo in the twelfth century with Gerhard of Cremona (1114–87) and his followers, the era of scientifically oriented medicine taught at universities began. Northern Italy, Montpellier, and Paris became major centers of medical study at this time. Hippocratic–Galenian medicine became the all-embracing and generally binding concept.

Neoplatonism and astrology (of which Marsilius Ficino, 1433–99 was a prominent practitioner), contributed to shaping Hippocratic–Galenian medicine into an all-encompassing model during the Renaissance. Modern medicine has been gradually developing ever since. This development was furthered by the discovery of numerous ancient texts, e.g., the *De re medicina* of Celsus, discovered in 1455. Andreas Vesalius (1514–64), in his secular anatomical work *De humani corporis fabrica* (1543), wanted to purify the ancient authorities of errors that had arisen through mistranslation. More and more, medicine based itself on its own experience, on its own theories, and on its own 'experiments.' Theophrast of Hohenheim known by the name of Paracelsus (1493/94–1541) introduced a chemistry-oriented way of thinking into medicine, while René Descartes (1596–1650) contributed a mathematical–mechanical thought pattern; both were complemented by Francis Bacon's (1561–1626) model of empirical thinking. The human body was released from religious cosmology and gradually became the outstanding subject of scientific examination. This initially took the form of anatomical explorations.

A secular step was taken at the end of the eighteenth century and early in the nineteenth century in English and French hospitals: the focus of medicine and of the physicians shifted from the patient to his/her disease. The sick man disappeared from medical cosmology. Seventeenth-century attempts at scientific systematization led to extensive nosologies in medicine, e.g., that of Thomas Sydenham (1624–89). Hippocratic–Galenian symptomatology, which was directed at the individual patient, was replaced by described syndromes. These syndromes became detached from the individual sick person, who turned into a mere 'case.' With this, a scientific–experimental field of separate research into disease entities opened up for medicine.

The man-machine model of iatro-physics and the man-reagent model of iatro-chemistry finally introduced modern scientific thinking into medicine. As medical theory was progressively freed from any religious or metaphysical reasoning, the question as to the source of life was raised. It was answered by dynamic concepts of disease, by psychodynamism and biodynamism. These concepts finally led, via Brownianism, to modern psychopathology and psychotherapy. In contrast to that, iatro-morphology built on the visible and observable. Georgio Baglivi (1668–1708) paved the way from humoral to solid pathology. Pathology improved from Giovanni Battista Morgagni's (1682–1771) organ pathology via François-Xavier Bichat's (1771–1802) tissue pathology to the functional cellular pathology of Rudolf Virchow (1821–1902). Cellular pathology established the physiological and pathological fundamentals of scientific medicine.

The actual step towards modern scientific medicine was taken in the late eighteenth and early nineteenth centuries. Chemical, physical, and then biological thinking has been determining the theories of medicine ever since. Deliberately reproducible (research) results led to a 'triumphal victory' for modern medicine in the second half of the nineteenth century with developments such as anesthesia (nitric oxide 1844–6; ether 1846; chloroform 1847), antisepsis (Joseph Lister, 1827–1912), asepsis (Ernst von Bergmann, 1836–1907; Curt Schimmelbusch 1860–95), immunology, serum therapy, and the diagnostic (1895–6) and therapeutic (1896) possibilities of x-rays. With acetylsalicylic acid (asprin), the first chemically representable fever-reducing drug was introduced into medicine in 1873.

Modern scientific medicine follows the iatro-technical concept pathophysiology, a causal analysis of distinguishable disease entities, objectifying and gauging methods, and causal therapy characterize such medicine. The driving force of this scientifically and technically oriented medicine is its endeavor to put physiology, pathology, and therapy on a scientific footing; first chemistry, then physics became central guiding sciences, and laboratories and experiments were introduced even into clinical practice. With the animal model, bacteriology launched a biological experimental situation. From the start, the iatro-technical concept created specific tensions in medicine: on the one hand, the scientific fundamentals which aim at general insights get into the maelstrom of the physician's need for action (consequence: the 'autistic-undisciplined thinking' and acting, as denounced by Eugen Bleuler, 1857–1915). On the other hand, medical action directed at the sick individual patient as a subject gets caught in the maelstrom of generalized technical applications and scientifically justifiable therapies (consequence: the 'therapeutic nihilism' of the Second Viennese School, i.e., proscription of any insecure therapeutical measures; Josef Skoda, 1805–81). As a result, since the 1840s both patients and medical practitioners have been turning to and from homeopathy and nature healing over and over again.

2. *The Institutional and Organizational History of Western Medicine*

Medicine is a science built neither on pure knowledge nor on mere empiricism. It is rather an action-based science directed at the sick individual. As already formulated in Hippocrates' first aphorism, the patient and his/her environment are granted equal parts in medicine. Therefore, a scientific and conceptional internal perspective can only partly explain the significance of medicine. From a historical perspective, diachronic questions are necessary which are directed at medicine from outside, i.e., from society. Consequently, a comprehensive history of medicine also demands socio–historical, historico–sociological, and anthropological perspectives.

The interaction between scientific knowledge and practical experience has been part of medicine since ancient times. In the course of the scientification of medicine during the nineteenth century, the historical argument simply disappeared from scientific reasoning. As a well-aimed counter-reaction against the purely scientific–technical concept, the history of medicine was reintroduced into the medical curriculum in the early twentieth century. Far into the 1990s, the history of medicine was considered in medical faculties and schools to be the substantial representative of the humane aspects of medicine. However, just as the objects and methods of the historical sciences began to expand in the late twentieth century, medicine also became an object of general history, i.e., especially one of social and cultural history. Yet, although methodological professionalization raised the academic standards of the history of medicine, the position of medical history within medicine has been impeded by the emphasis on professional historiography, including its technical jargon. Since the 1990s, humanist issues of medicine have come increasingly under the category of ethics in medicine. As a result, the existence of history as an academic subject within medical faculties and schools has been called into question. If history in medicine wants to survive, it has to meet the expectations of its humanist–pragmatic task which the rapid progress in medicine dictates. Due to the different tasks—which partly refer to the current problems of legitimization and action, partly to the debate about contents and methods in historiography—it is necessary to make a distinction between a 'history of medicine' and a 'history in medicine.'

Within the doctor–patient encounter, the patients receive special attention due to their special need for help, as do doctors due to their specific need for action. From antiquity to the beginnings of scientific–technical medicine, the conceptual world of medicine and the patients' lay interpretations overlapped in large areas. Humoral pathological medicine lives on in popular medicine, in nature healing, and in alternative medicine. With the advent of the scientific concept, the world of medicine and that of the patient drifted apart. From the eighteenth century, the physician as an expert began to face the patient as a layperson. At the same time, medicine came to hold the monopoly of definition and action over the human body. In the sociogenesis of health as a social good and value, medicine and the physicians were endowed with an exclusive task; and they were thus given a special position in society. The historical and sociological discussion of the 1970s called this process 'the professionalization of medicine.'

The medical marketplace has always been characterized by enormous competition. Because of that, the healing professions have almost always depended

on their clients, the patients. In medieval and early modern times, the patronage system made the academically trained physician subject to his socially superior patron-client. This was only inverted with the gradual scientification of medicine in the eighteenth century. With the compulsory linkage to scientific medicine, orthodox medicine has developed since the 1840s. Orthodox medicine was confronted with traditional or modern alternatives as kinds of 'outsider medicine'; healers who applied these methods were both internally and externally discriminated against as quack doctors. Industrial societies with their various social security systems and their general claim of social inclusion opened up a new clientele of the working class and the lower strata for medicine.

A significant medical institution both in the Western and in the Islamic tradition is the hospital. Hospitals were established partly for pilgrims, partly for anybody in need of help, in order to fulfill the Christian ideal of 'caritas.' Having migrated eastwards along with ancient medicine, the hospital became the place of training for physicians in Islam around the turn of the first millennium. Within the European tradition, the charitably oriented hospital changed only in the late eighteenth century: the modern hospital came to be a place to which only sick people were admitted, with the purpose of discharging them after a well-directed treatment within a calculable period of time. Such was initially the responsibility of proficient nursing. Around the same time selected hospitals had become places of practical training for future physicians. Scientific–medical research was also conducted in these university hospitals from the 1840s onwards; only as a result of this, antisepsis medicine entered the General Hospitals in the 1860s. This was the beginning of the history of modern nursing, of scientifically oriented medical training, of the many medical disciplines, of the modern Medical Departments and Schools, and of the hospitals which only then became medically oriented.

At first sight, diseases embody the biological reality in the encounter between physician and patient. As a study of leprosy or plagues shows, diseases have had a number of different names over the course of time. Therefore it is hardly possible to determine with certainty which diseases were referred to in reports handed down to us. However culturally interpreted, diseases induce the encounter between physician and patient. It is therefore the sick human being who accounts for the existence of medicine. Over and above the conceptual and institutional interpretations, disease has always been experienced as a crisis in human existence. In different cultures and eras sick people and their environment interpreted diseases with reference to their metaphysical reasoning. Diseases are given a cultural meaning as a divine trial or as a retaliation for violated principles of life and nature. The scientific–analytical notion of disease in the iatro-technical concept is not capable of taking the patient's quest for meaning into account. The chronically ill, the terminally ill, and the dying therefore present a constant challenge for scientific–technical medicine—including its necessarily inherent thoughts of progress. The result is a continuous push towards alternative and complementary medicine as well as towards lay and self-help.

From a historical and sociological point of view, even the interpretations of health take place in a repetitive process. In antiquity, health resulted from the philosophical ideal of the beautiful and good human being (ancient Greek: 'kalokagathia'. The Islamic world and the Christian Middle Ages assigned the health of the body to religious categories. In the sciences of early modern times, the nature-bound human body came to be understood as an objectified body. In early rationalism, Gottfried Wilhelm Leibniz (1646–1716) elevated medicine to a means of ordering private and public life according to scientific rules: the order of the soul, the order of the body and the order of society corresponded with piety, health, and justice; they had to be guaranteed by the church, by medicine, and the judiciary. In critical rationalism, reason, physics, and morality were declared a unity. In the early nineteenth century, the interpretation of the concept of health became a scientific activity. The question of the meaning of health was excluded from medicine. Within the iatro-technical concept, health as a chemical, physical, or statistical norm measured by percussion, auscultation, the thermometer, etc., became a negative definition of scientific notions of disease. Those values and ideas of order which were inextricably bound up with the notion of health seem to have disappeared in the course of scientific–rational progress.

With the scientific notion of health and disease, modern medicine was finally granted the monopoly of interpretation and action for a rational design of the biological human resources in modern societies. Disease is the cause for medical action, health is its aim. Via the societal value of health medical knowledge and medical action are integrated into their societal surroundings. The definition of health mediates in its valid interpretation of a particular civilization between the individual, the social, and the 'natural' nature of mankind. Thus, health integrates the 'natural' biological basis, the social basis and the individuality—the 'I' and 'me'—of human existence. In terms of philosophical anthropology, the valid notions of health mediate between the outer world/nature, the social world/society, and the inner world/individuality with regard to the body/living corpse. The place and the scope of medical action in society are defined via the interpretation and the effect of the valid definitions of health.

To the same extent to which the human relationship towards nature becomes more scientific and leads to a scientific–technological civilization, more and more aspects become scientific within the combined in-

terplay of the body as a part of oneself and the body as an object. 'Homo hygienicus,' i.e., mankind, who defines himself and is defined by others in terms of medical standards, are typical examples of the paradox of modernity: the autonomously progressing discoveries of scientifically oriented medicine liberated the person, communities, and society from the task of having to account for the values and norms of their bodily existence. From this position, the individual and public 'rationalization of the body'—as stated by Max Weber (1864–1920), amongst others, at the beginning of the twentieth century—could begin their widely welcomed and universally demanded 'victory march of medicine.' With the end of the belief in progress, the disadvantages of this process have been perceived since the 1970s. So in historical, sociological, and anthropological debates medicine has turned into an 'institution of social control.' Medicalization—as Michel Foucault (1926–84), amongst many others, pointed out from the early 1970s onwards—became the focus of a fundamental critique on the penetrating social impact of modern medicine.

The doctor–patient relationship is no anthropological constant. Rather, this relationship is an integral part of the economic and social organization of a community. This is especially true for public medicine and the health sciences. Early forms of the modern public health service developed from the end of the thirteenth century in northern Italian cities. The driving forces were the endemic plagues which kept returning at regular intervals after the Great Plague of 1347–51. In early modern times those public health measures evolved which then became regular institutions in the great commercial cities: general regulations which also had health effects (such as food inspection and market regulations, etc.); municipal supervision of those practicing medicine; city hospitals for the infirm, and others for special isolation (leprosaria, plague hospitals, etc.); the beginnings of a municipal medical and surgical service.

When the territorial states were thoroughly consolidated on the legal and administrative level, medical and sanitary supervision unfolded in the late seventeenth and early eighteenth century. A public medicine arose which was linked to administrative and political goals. Within the framework of mercantilism one intention in calculating power was to increase the population. The paternalistic welfare state towards the end of the absolutist period developed public health as part of its 'populating policy.' This is where the special, still heavily disputed relationship between state, medicine, and women, begins: to the extent to which women were made the 'human capital' of a well-directed population policy their role came to be substantially determined by medicine (medicalized midwifery, medicalization of birth and childhood, medicalization of housekeeping, childrearing, etc.).

Through the bourgeois revolutions the nation as sovereign made itself the object of public health. Besides the idea of a qualitative evaluation of the population the idea arose of calculating the value of a human being in terms of money. In a combined act of statistics, epidemiology, physics, geography, metereology, etc., and driven by a paternalistic idea of welfare the first modern health sciences came into being. The nineteenth-century cholera pandemics accelerated this development. Only with the beginning of industrialization was comprehensive healthcare promoted, due to the 'public value' of health. The modern health sciences finally defined in a biological chain of causes and effects a closed circle of human hygiene: the environment as conditional hygiene of Michel Levy (1809–72), Jules Guérin (1801–86), Edmund Parkes (1819–76) or Max von Pettenkofer (1818–1901); the microbiology or bacteriology as infectious hygiene of Louis Pasteur (1822–95) or Robert Koch (1843–1910); the dynamic relation of exposition and disposition as constitutional hygiene; the forthcoming human life as racial hygiene and eugenics; and finally health and society as social hygiene—all these combined approaches took hold of basically all areas of human existence in the late nineteenth and early twentieth centuries. This is especially true for unborn life. At this point the historical reflection on public healthcare, medical statistics, and epidemiology, health economics, health laws, and health systems, etc., begins.

The idea of obligatory collective health was developed after the turn of the twentieth century. In the biologistic ideology of National Socialism it was molded into a highly rational model of future-oriented health. The goal of National Socialist healthcare was a 'genetically healthy' and 'racially pure' population of 'Aryan–German blood.' This genetically and racially cleansed 'Volk' should be fit to survive the secular 'war of races.' The goal of obligatory national—which really means racial—health was located beyond all individual rights. This model of a totalitarian program of healthcare was based categorically on exclusion. It was carried out in decisive areas: at first sterilization (at least 350,000 victims, all of them Germans), then 'euthanasia' (over 100,000 victims, including so-called 'wild euthanasia,' probably more than 250,000 victims, all of them Germans), and finally the extermination of 'Völkerparasiten' during the Holocaust (approximately 6 million victims, of whom 180,000 were Germans).

Medicine in National Socialism requires special mention here, as it is historically unique in the world. Basically an excluding function is always inherent in medicine. In the daily routine of medicine this becomes apparent wherever decisions have to be made on the allocation of scarce goods and services in short supply (e.g., transplantation medicine, assessment and social medicine, allocation of medical services). As to their forcefully carried out eugenic actions the National Socialists themselves referred to the examples of other countries, especially the USA. Indeed there had been, prior to 1933 and after 1945, an 'internationale' of

eugenicists. The National Socialist model of health can serve worldwide as an historical example of the conflicting nature of the rationalization process on which modern times are based—a process which is irreversible. National Socialism is therefore to be seen as the Janus-faced flipside of the modern era, which is as obsessed with progress as it is blinded by it. This means that National Socialism and medicine in the National Socialist era are not only particular phenomena of German history. They are rather a problem of our times and thereby a problem of all societies which have built on modernism.

3. *Conclusion and Outlook: Western Medicine and the Molecular Health of Tomorrow*

In the narrow scientific–practical sense medicine is determined by the dialectics of medical knowledge and medical action. Such dialectics have accompanied medicine from its first historical steps onwards to the scientific–rational establishment in Greek antiquity. The controversy between 'dogmatists' and 'empiricists,' between 'medics' and 'doctors,' has raged throughout the entire history of medicine. It is the patient who gives this dialectic a direction: the patient's need for help constitutes medicine. Because of the patient and the special situation in which the patient and the doctor meet, medicine possesses a domain distinct from that of other sciences. In the encounter between doctor and patient medicine faces the necessity to take action. With this, the doctor–patient encounter is placed within a historical context. This is the starting point of historical reflection as an indispensable aspect both of a history in medicine and a history of medicine.

The position of medicine within society is changing fundamentally. The scientific medicine of the late nineteenth and early twentieth centuries was the reproductive rearguard of industrialization. By contrast, the medicine of the late twentieth and early twenty-first centuries is a forerunner of scientific, economic and social change. Since the 1980s, molecular biology has become the referential discipline of medicine. With that, it has become possible to diagnose diseases in their genetic informational principles and their molecular mechanisms. The molecular transition means a secular change in the concept of medicine. The molecular biological thought pattern transfers the process of health and disease into the informational bases of the production and function of proteins. With that, genetics and information processing in the cellular and extracellular realm step into the limelight.

Historical analysis has shown to what extent medical knowledge and medical action affect society. Thus, experimental hygiene and bacteriology have cleansed and pasteurized conditions and behavior since the middle of the nineteenth century. Since the 1960s the pharmacological possibilities of contraception have altered the self-definition and the role of women. Molecular genetics and clinical reproductive medicine have already started to show their effects on the generative behavior of mankind (*in vitro* and *in vivo* fertilization, preimplantation diagnostics, etc.). The transition to molecular medicine will also change the living world and the social world: molecular medicine leads to a new image of humans. The medicine of the last decades followed the model of 'exposition,': in which diseases had an external origin (e.g., bacteria, viruses, stress factors, etc.). Molecular medicine shifts the etiology from exposition (i.e., the outside world of all human beings) to disposition and therefore to the (genetic) inner world of the individual. Predispositions for certain diseases are internalized as an individual (biological) fate. The result is the 'genetization' of individual life and of its subsequent societal acts (e.g., life and health insurance). The interpretation of health, including the notion of normality of the body, will in substance be linked with the knowledge of genetically determined dispositions. Such a genetic living world will be as 'real' for human beings of the twenty-first century as the bacteriological living world was 'real' for the twentieth century. See the following for further information sources and contemporary research: *Bibliography of the History of Medicine*; *Bulletin of the History of Medicine*; *Current Work in the History of Medicine* (an international bibliography); *Isis* (an international review devoted to the history of science and its cultural influences); *Journal of the History of Medicine and Allied Sciences*; *Medizinhistorisches Journal*; *Social History of Medicine*.

See also: Body, History of; Health: Anthropological Aspects; History of Science; Medical Geography; Medical Sociology; Medicalization: Cultural Concerns; Plagues and Diseases in History; Scientific Disciplines, History of

Bibliography

Bynum W F, Porter R (eds.) 1993 *Companion Encyclopedia of the History of Medicine*. Routledge London/New York

Garrison F H 1929 *An Introduction to the History of Medicine, with Medical Chronology, Suggestions for Study and Bibliographic Data*, 4th edn. Saunders, Philadelphia/London

Kiple K F (ed.) 1993 *The Cambridge World History of Human Disease*. Cambridge University Press, Cambridge, UK

Loudon I (ed.) 1997 *Western Medicine. An Illustrated History*. Oxford University Press, Oxford, UK

Norman J M (ed.) 5th edn. 1991 *Morton's Medical Bibliography. An Annotated Check-list of Texts Illustrating the History of Medicine*, 5th edn. Scolar, Aldershot, UK

Porter R (ed.) 1996 *The Cambridge Illustrated History of Medicine*. Cambridge University Press, Cambridge, UK

Rothschuh K E 1978 *Konzepte der Medizin in Vergangenheit und Gegenwart*. Hippokrates, Stuttgart, Germany

Schott H 1993 *Die Chronik der Medizin*. Chronik, Dortmund, Germany

Shryock R H 1936 *The Development of Modern Medicine. An interpretation of the Social and Scientific Factors Involved.* University of Pennsylvania Press, Philadelphia, PA
Sigerist H E 1951–61 *A History of Medicine, Vol. 1: Primitive and Archaic Medicine; Vol. 2: Early Greek, Hindu and Persian Medicine*. Oxford University Press, New York

A. Labisch

Meiji Revolution, The

1. *The Meiji Revolution* (*1853–77*)

Meiji ('Enlightened Government') is the era name adopted in 1868 by the new government that overthrew Japan's previous military government of the Tokugawa *shoguns* (*bakufu*) and re-adopted the emperor as the country's sovereign. Since 1853, when Matthew C. Perry, Commander of the American Squadron to the China Seas and Japan, first put military pressure on Japan to open up its borders to foreign intercourse, the country had been quickly incorporated into the world's capitalist system. For the small feudal state that Japan had been up to that moment, the problem was to find a way to survive as an independent country and to decide on the type of state that would be able to deal with world capitalism and the international relations of East Asia. The great political and social changes that were the result of the political struggle, which developed around this problem, we shall call here the Meiji Revolution. The year 1877, when the last rebellion of the samurai class was suppressed, we shall consider the year of its completion. The political struggle can be divided into three stages.

2. *The First Stage: The Overthrow of the Bakufu* (*1853–68*)

The *bakufu's* authority was greatly damaged by the fact that it had been forced to open up the country to intercourse with foreign powers. In order to stabilize its political foundation, the *bakufu* asked the *daimyo*, or feudal lords of Japan, for their opinions and groped towards achieving some kind of consensus among the *daimyo* class as a whole. On the other hand, it wanted to clarify the legitimacy of its own national government by strengthening its connection with the emperor in Kyoto, who nominally still was Japan's monarch. Moreover, the *bakufu* began to employ military drills in the Western style and made a start to build its own navy tutored by naval officers from Holland. However, when the news that a joint English and French squadron which had just defeated China in the Second Opium War was about to visit Japan in 1858, the *bakufu* buckled under this military pressure to sign treaties allowing free trade with the United States, Holland, Russia, England, and France. What is more, these treaties were unequal treaties in that they included clauses on extraterritoriality and fixed tariffs that denied Japan its legal and tax autonomy. The *bakufu* sought ratification of the treaties from Emperor Komei, but the emperor refused on the ground that Japan would cease to exist as an independent state if this attitude of acquiescence to the demands of the foreigners were to persist. As a result, the discussion about Japan's existence as a nation-state became divided between those supporting the *bakufu* and those backing the position of the emperor.

As a result of the opening, in 1859, of the harbors of Yokohama, Nagasaki, and Hakodate to free trade, Japan was quickly absorbed into the world's capitalist system. Prices rose precipitously, and the markets of the country were thrown into confusion. Voices denouncing the signing of the treaties and the *bakufu*'s policy of submission to the foreigners arose not only among the members of the ruling class, but also among well-to-do farmers and merchants, and so became a protest of the Japanese people as a whole. As a result, the *bakufu* lost confidence in its ability to rule the country. By contrast, the idea of a government that would take a hard line towards the foreigners and make the *bakufu* cooperate with that effort was gaining in popularity among those supporting Emperor Komei, who continued to refuse to ratify the treaties. Powerful *daimyo*, who, in their turn, were thinking of taking up the right to govern the country themselves, were also found supporting this idea.

Warriors of the faction that proposed radical military reforms wanted to throw out the foreigners even more badly. They were of the opinion that there was no military advantage to be gained just by superficially adopting Western-style drilling methods, while keeping the old feudal system linking military ranks with family background. They stressed that Japan should aim to modernize its army while engaged in actual battle, and so they wanted to use war to accomplish internal reform. Thus, in June of 1863, radicals from the domain of Choshu first attacked foreign shipping in the Straits of Shimonoseki in order to force the shogun to take military action against the foreigners, for the country's leadership in such matters was considered to be his responsibility. However, this type of behavior by the radical warriors, which showed no regard for the feudal order, frightened Emperor Komei and the highest officials of his court. They proceeded, in September of 1863, to chase the Choshu warriors from Kyoto with the help of the warriors from the Aizu domain in north-eastern Japan, which traditionally supported the *bakufu*.

On the other hand, the Great Powers of Europe and the United States, which had concluded treaties with Japan, would not allow a native movement that aimed to break away from world capitalism by military force. Therefore, they decided to complete Japan's incorporation into the world's capitalist system by dealing its radical faction some harsh blows. In August of

1863, an English naval squadron attacked the home base of the Satsuma at Kagoshima. Next, in September of 1864, a joint naval force of English, French, American, and Dutch warships attacked the strategic port of Shimonoseki in Choshu. This fief had been declared an 'Enemy of the Imperial Court' by Emperor Komei, because its soldiers had entered the capital against his will. Moreover, in November of 1865, the joint naval force of these four nations penetrated Osaka Bay, and obtained from Emperor Komei, through military pressure, the ratification of the treaties. Thus, Satsuma and Choshu both painfully experienced Western military power at first hand. They were, therefore, also the first to understand that, in order to resist foreign pressure, it would be necessary to bundle the power of the entire Japanese people through the establishment of a unified nation-state. They became convinced that, as a first step to accomplish this, they should take immediate and drastic measures to reform their military forces.

In Satsuma, the old system based on lineage was abandoned and a new homogeneous army of samurai warriors established. In Choshu, lower ranking samurai, farmers, and townsmen joined to form an armed militia or *shotai*. The *bakufu*, on the other hand, used the opportunity offered by the fact that Choshu had been named an 'Enemy of the Imperial Court' and the domain's defeat at Shimonoseki to try and re-establish its authority. It succeeded in co-opting the Court and in destroying the lateral alliance of the powerful *daimyo*. In 1866, with the sanction of the Emperor, the *bakufu* organized an expedition to punish Choshu, but the domain had just concluded a military alliance with Satsuma, and with its newly organized militia it was able to defeat the army of the *bakufu*.

3. *The Second Stage: The Restoration of Imperial Rule* (*1868–1871*)

In a *coup d'etat* of January 1868, samurai from Satsuma, convinced that it would be impossible to establish a strong nation-state under the leadership of the Tokugawa, together with Tosa warriors and members of the Imperial Court, who were concerned with the loss of prestige the Imperial House had suffered because of its recent connection with the *bakufu*, took over power and established a government based on the restoration of imperial rule (*osei fukko*). Warriors from Satsuma, Choshu (which had obtained a cancellation of the order declaring it to be an 'Enemy of the Imperial Court'), and Tosa defeated the *bakufu*'s forces near Kyoto in the same month. By means of a civil war, which continued until June of 1869, the restoration government was able, with the cooperation from the domains, to eradicate the power of the old *bakufu* and to seize the lands which had been under its jurisdiction. The Tokugawa family was reduced to nothing more than being one of many *daimyo*. The restoration government made Emperor Meiji, who had been born in 1852 and had succeeded to the throne after the death of his father Komei in 1866, the country's new sovereign. Instead of the slogan *Sonno joi* ('Revere the Emperor and Expel the Barbarian') through which it had come to power, the new government now stressed Japan's national power with a new slogan *Bankoku Taiji* ('Holding Out Against All Nations of the World'). Also, in order to obtain support for the new government, the new leaders criticized the autocratic nature of the old *bakufu*, professed its respect for public opinion, and established a deliberative assembly the members of which were elected by the domains.

During the one-and-a-half-year long civil war, which claimed more than 10,000 lives, the idea that birth should determine social standing was abandoned in the domains of the various *daimyo*, which also westernized their military forces and accepted the participation of rich farmers and merchants in their local governments. For these men, it was natural to expect that the army of the new Japanese nation-state would consist of the combined military forces of the domains. But to the restoration government, it seemed that trying to take a hard line against the foreign powers in this manner might spiral the power of the local samurai out of control. Conversely, the leaders of the domains attacked the weakness of the government's foreign policy. The effort to stabilize a hard-line government based on a consensus and with the cooperation of the domains was abandoned. In August of 1871, after several thousand elite warriors from the three domains of Satsuma, Choshu, and Tosa, had gathered in Tokyo, the government took the drastic measure of abolishing the domains and establishing prefectures, saying that 'The existence of the domains hurts the realization of the *Bankoku Taiji* policy'.

4. *The Third Stage: The Centralized Nation-state (1871–77)*

Therefore, the problem of establishing this was felt to be an urgent matter to be accomplished through the cooperation of the three domains mentioned above. These, together with the domain of Hizen, formed the backbone of the new government. In December of the same year, Iwakura Tomomi (the Court), Okubo Toshimichi (Satsuma), and Kido Koin (Choshu) and others were sent as the representatives of the new government on an embassy to the United States and Europe in order to revise the unequal treaties. Saigo Takamori (Satsuma) was the most important man of the caretaker government remaining behind in Japan and he forcefully pursued the legal and institutional reforms that would make the Great Powers agree to a

revision of the treaties. In order to accomplish this, the armament factories and the ships built and owned for the sake of modernization by the *bakufu* and the different domains were brought under government control, students were sent overseas, and many able foreigners were employed. Ten percent of the original assessed income of their former domain was given as personal property to the *daimyo*, and the courtiers (even those of the lowest ranks) were assured an income at the same level as the former *daimyo*. The government, however, aimed at building a nation-state on the model of those in Europe and America, which required an enormous outlay of funds. For this reason it did not relieve the tax burden required of the farmers. Also, there was no agreement between the four domains on the question of how to deal with the warrior class. Even though their stipends had been guaranteed, from 1873 an army draft had been started from among all layers of the population in order to be able to suppress internal rebellions. Therefore, it became necessary to face the problem of the position of the samurai in the new society.

When the caretaker government in Japan realized that, is the short term, a revision of the treaties was out of the question, it strengthened its intention to display the energy and prestige of the new Japanese nation-state within East Asia. The Japanese government wanted to assure its prestige in an East Asian world that seemed dominated by continuing concessions to the Great Powers of Europe and the United States since the defeat of the Qing in 1842. It hoped in this way to heighten the prestige of the new nation-state at home and decided it could use the samurai class to project its military power. Already in 1871, during negotiations with the Qing the Japanese had started to build new international relationships in East Asia. When aborigines from Taiwan harmed and killed some inhabitants of the Ryukyu Islands, which traditionally paid tribute to both China and Japan, this was used as an excuse to plan a military expedition to Taiwan. Moreover, the government considered the attitude of Korea, which had refused, since 1868, to negotiate with Japan because its diplomatic form was no longer that used by the Tokugawa *bakufu*, an affront to its national prestige. For this reason, in August of 1873, the caretaker government decided to send an embassy with the message that it was ready to go to war over this issue. However, the chief ambassador of the delegation that had been sent to Europe and the United States, Iwakura, came back to Japan in September of that year and forcefully opposed this idea. He was supported by that part of the samurai class that was in the process of being transformed into the bureaucracy of the new nation-state.

Because Emperor Meiji cancelled the decision to send the embassy, Saigo and Eto Shinpei (Hizen), who had been the main members of the caretaker government, resigned from the government. The Satsuma and Tosa warriors, who were living in Tokyo as part of the Imperial Household guards, seceded from this force, and in February of 1874 the samurai of Hizen rebelled under the leadership of Eto. The Meiji government, now mainly led by Iwakura and Okubo, used the warships it had collected and the newly installed nationwide telegraph network to the maximum extent possible and subdued the insurrection within a short time, thus preventing it from spreading. However, in order to win over the armed forces and its officer corps, which both still mainly consisted of samurai, the government proceeded with the invasion of Taiwan in April 1874. Ignoring China's protest that this was an infringement of its national sovereignty and showing its readiness to go to war with the Qing, the government showed that it was able to take a hard line. When the Qing recognized, at the very last moment in October 1874, that the Taiwan expedition had been a 'heroic act,' the Japanese leadership learned it need not respect the power of the Qing. On the other hand, from the attitude China had taken on the Taiwan issue, it was possible to forecast that, in case of a Japanese dispute with Korea, the Qing, even though it was officially Korea's suzerain, would not come to its rescue. Moreover, it was becoming clear that if Japan engaged in military ventures in East Asia the Western powers were likely to remain neutral.

On the national level, this success earned the government the loyalty of its officer corps, and, across the board, the pace towards a system of general conscription was accelerated in order to prepare for war. For the first time, it seemed possible to replace the samurai class and establish bureaucratic control over the army. In February 1876, Japan imposed, through military force, an unequal treaty on Korea, which had kept itself isolated in a manner similar to Japan before Perry's arrival. Also, with the treaty concluded with Russia in May 1875 by which Sachalin was exchanged for the Kuril islands, the government succeeded in solving a border dispute, which had existed since the last years of the eighteenth century.

So between 1874 and 1876, the Meiji government gave permanent shape to Japan's borders and reorganized its international relations with East Asia. In this way, it put an end to the general feeling among the population of Japan, existing since the 1840s, that the state was in a crisis which might lead to its dissolution. The only problem left was the revision of the unequal treaties with the Great Powers of Europe and the United States.

With the national prestige the government harvested from these developments, it decided to embark on some radical internal reforms. These reforms were meant to create the conditions necessary to establish capitalism in Japan, which, in turn, would enable the country to respond adequately to world capitalism. One of these conditions was to revise the land tax and to establish the right to private ownership of land

without diminishing the national tax income as a whole, and the government meant to accomplish this within 1876. Another was to convert the stipends of the former *daimyo*, the courtiers, and the samurai into government bonds (August 1876). This latter reform was no rejection of feudalism, but rather a decision to change it into a tool for the adoption of capitalism.

The land tax reform called forth a movement that opposed the fixation of the taxes at the high levels of feudalism. This opposition was led by rich farmers, and peasant rebellions exploded, one after another, in virtually all regions of Japan. At the same time, from October 1876, a revolt by samurai broke out which reached its peak with the Satsuma rebellion of February 1877. In January of that year, therefore, the government lowered the land tax from 3 percent to 2.5 percent of the land value and also diminished the local taxes. This drove a wedge between the peasant movement and the samurai rebellion and allowed the government, after a violent struggle lasting eight months, to defeat, with its conscription army of 60,000 men, the 40,000 samurai of Satsuma. As a result, the movement to incorporate the samurai class into the structure of the nation-state was extinguished and the power of the new state was no longer challenged.

However, after its defeat, the samurai class held the government to its original promise to respect public opinion and it became the foremost power in the struggle for the opening of the National Diet to all classes of Japanese society. In this struggle, the samurai, allied with the rich farmers and merchants who stressed that there could be no taxation without representation, opposed the increasingly authoritarian Meiji government. And so the struggle, which had centered on the manner in which national power was to be wielded, now transformed itself into one for democratic rights for everyone.

5. *Historiography of the Meiji Revolution*

These events were, until the second decade of the twentieth century, considered to be the epochal restoration of direct imperial rule, accomplished by the imperial house with the reverent support of the Japanese people. When Japanese capitalism reached a dead end during the 1920s, with its repressive system at home and its increasingly aggressive policies abroad, Japanese social scientists, and Marxists in particular, started to link the unholy trinity of monopoly capital dominating the country's political system, its half-feudal land ownership, and its imperially indoctrinated army and bureaucratic leadership with these events that were now known as the Meiji Restoration. This train of thought divided itself into two schools, one maintaining that the Meiji Restoration was an incomplete bourgeois revolution, and the other that it was nothing but the establishment of absolutism. After World War II, and especially during the 1960s, the pre-war interpretations were considered to be flawed because they neglected the concrete facts of the international situation around Japan at the time of the revolution and had been narrowly focused on the events as one stage in the development of the Japanese nation by itself. According to this thinking, foreign pressure in the middle of the nineteenth century forced the realization that if Japan wanted to keep its independence, it needed to cooperate in achieving a thorough overhaul of the state. Because a bourgeois revolution was not necessarily an indispensable ingredient for the success of the capitalist system transplanted to Japan in the second half of the nineteenth century, the revolution compromised and merged with the feudal elements of Japanese society.

However, today the positions are divided as follows. One group of historians stresses the fact that the transition to modern society was comparatively easy because of the high levels of social and economic development in Japanese society during the second half of the Edo period (1600–1867). Another group of historians considers the Meiji Restoration a popular or a nationalist revolution, ignoring such problems as feudalism, class, and the status of power holders. While recognizing a certain level of development in pre-modern Japan, a third group maintains that the Meiji Restoration was the formation of a centralized imperial state, which aimed to establish capitalism on a basis of support for the old nobility and the *daimyo* class of Tokugawa society. During this process of introducing world capitalism into Japan, the state itself went through several transformations.

6. *Issues of the Debate on the Meiji Revolution*

The Meiji Revolution presents, by its very nature, a complicated problem. On the one hand, it should be called a 'restoration of imperial rule' because of the efforts that were made to revive the structure of the imperial state that had centralized power in Japan more than 1,200 years earlier. Until 1945, the position of the emperor as the supreme wielder of power in the modern Japanese state seemed to be fixed, and this in itself continued to confront and suppress all political thought stressing the sovereignty of the Japanese people. On the other hand, in the sense that the feudal state made up of *shogun* and *daimyo*, which had existed for more than 260 years, was destroyed, it can be called a sort of revolution. This Asian-style nation-state, which had existed since the beginning of the seventeenth century, could not have gradually reorganized and modernized itself from above without the overwhelming military and economic pressures from world capitalism of the second half of the nineteenth century.

In Japan, the men responsible for these revolutionary changes did not come from the bourgeoisie as in the case of Western-style revolutions. During Japan's

closed-country policy, a new class of rich farmers and merchants had, to a certain degree, emerged with the formation of a national market that involved the whole country, but not the outside world. Their nationalism formed the basis of and fueled the changes that took place. However, the leaders of the Meiji period were a group of men, who had their roots in the warrior class and had changed themselves into a literary and military bureaucracy while treating the state as their private property. By 1877, they had gained complete control of the Japanese state with the emperor at its head. However, without the support of the rich farmers, merchants, and ex-samurai, political and social stability still eluded them. This had to wait until the 1890s when a constitution had been put in place and a parliament was established so that other groups might have some say in the running of the government. In the very broadest sense, Japan's Meiji Revolution ranks with Germany, Italy, and Russia as one of the latecomers in the drive towards modernization. Further generalizing seems to be meaningless, and this author would like to stress that the revolution creating a nation-state in response to foreign pressure was probably the result of the specific circumstances existing in the second half of the nineteenth century in East Asia.

See also: East Asian Studies: Politics; East Asian Studies: Society; Ethnic Cleansing, History of; Feudalism; Imperialism, History of; Japan: Sociocultural Aspects; Modernization and Modernity in History; Nationalism, Historical Aspects of: East Asia; Nations and Nation-states in History; Parliaments, History of; Racism, History of; Revolutions, History of; Southeast Asian Studies: Politics

Bibliography

Beasley W G 1972 *The Meiji Restoration*
Jansen M B 1961 *Sakamoto Ryoma and the Meiji Restoration*
Norman E H 1940 *Japan's Emergence as a Modern State*
Walthall H 1999 *The Weak Body of a Useless Woman: Matsuo Taseko and the Meiji Restoration*
Waters N L 1983 *Japan's Local Pragmatists: The Transition from Bakumatsu to Meiji in the Kawasaki Region*

M. Miyachi

Melanesia: Sociocultural Aspects

The convention of dividing Oceania into three distinct geographic regions or culture areas—Polynesia, Micronesia, and Melanesia—originated in 1832 with the French navigator J-C-S. Dumont d'Urvilles. By this convention, Melanesia today encompasses the independent nation-states of Papua New Guinea, Solomon Islands, Vanuatu, and Fiji; the French overseas territory of New Caledonia (Kanaky); the Indonesian province of Irian Jaya or Papua (West Papua); and the Torres Straits Islands, part of the Australian state of Queensland. These islands comprise about 960,000 square kilometers with a total population of more than eight million people. The large island of New Guinea and its smaller offshore islands account for almost 90 percent of the land area and 80 percent of the population of Melanesia (see Fig. 1).

1. Melanesian Diversities

The diversity of contemporary Melanesian political arrangements is exceeded by a linguistic, cultural, ecological, and biological diversity that impressed early European explorers and challenged later social and natural scientists. Although mostly pig-raising root-crop farmers, Melanesians occupy physical environments ranging from montane rainforests to lowland river plains to small coral islands; they live in longhouses, nucleated villages, dispersed hamlets, and urban settlements. While racially stereotyped as black-skinned and woolly-haired people, they display a wide range of phenotypic features. Similarly, Melanesian social organizations illustrate numerous anthropologically recognized possibilities with respect to rules of descent, residence, marriage, and kin terminology. Melanesians speak more than 1,100 native languages, some with tens of thousands of speakers and others with less than a few hundred. English, French, and Bahasa Indonesia are recognized as official languages; several pidgin-creoles are widely spoken and also recognized as official languages in Papua New Guinea and Vanuatu. Melanesia is thus culturally heterogeneous to the point that its very status as a useful ethnological category is dubious; many anthropologists would assign Fiji to Polynesia and classify Torres Straits Islanders with Australian Aborigines. Ethnographic generalizations about the region often show a marked bias toward cases from some areas (e.g., the New Guinea Highlands) rather than others (e.g., south-coast New Guinea), and in any case remain rare.

Linguists, archeologists, and geneticists offer persuasive narratives of prehistoric migrations that account for Melanesian heterogeneity. These narratives, although not universally accepted, reveal how Dumont D'Urvilles' convention obscures historical and cultural connections between Melanesian and other Pacific populations as well as fundamental differences within Melanesian populations. At least 40,000 years ago, human populations of hunter–gatherers settled the super-continent of Sahul (Australasia), reaching what is today the island of

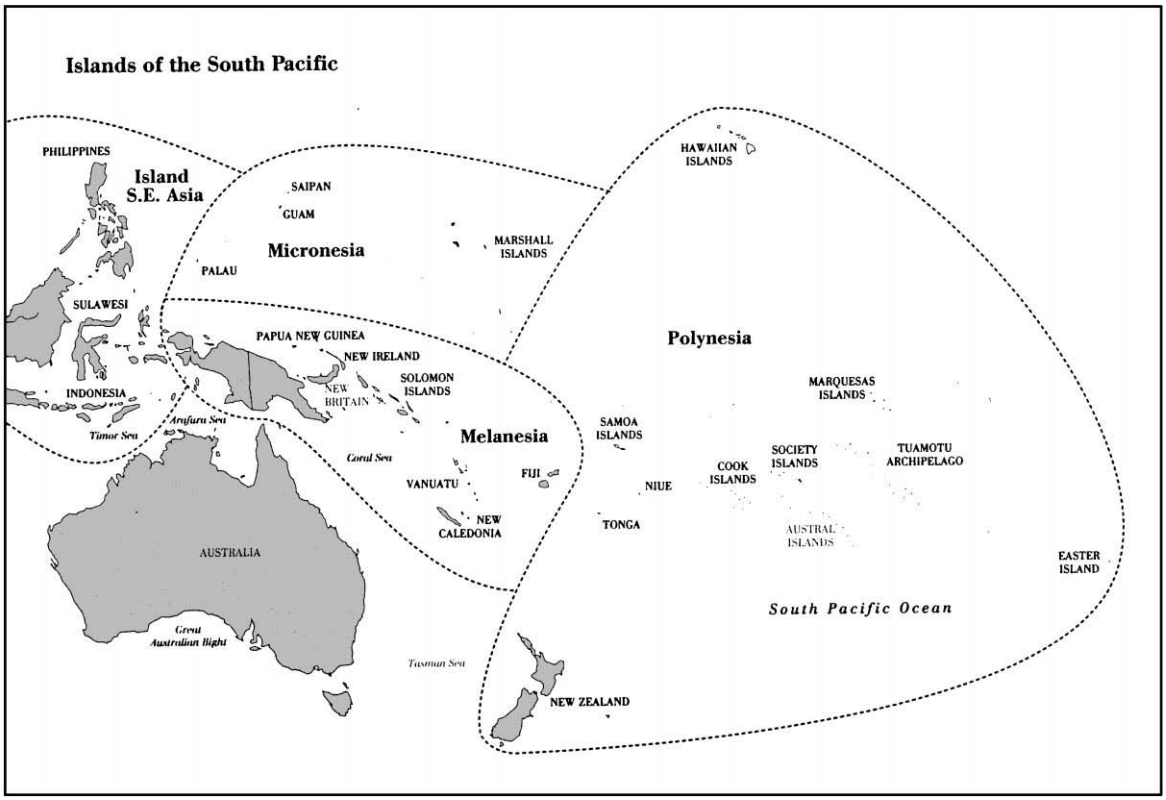

Figure 1
Islands of the South Pacific

New Guinea, the Bismarck Archipelago and the Solomon chain by at least 30,000 years ago. Linguistic evidence suggests that these settlers spoke Non-Austronesian (NAN or Papuan) languages, which today number around 750 and occur mainly in the interior of New Guinea, with isolated instances through the islands to the east as far as the Santa Cruz Archipelago. By 3,500 years ago, a second migration out of Southeast Asia had brought another population into sustained contact with the original settlers of the Bismarcks. This migration is associated by archeologists with the distinctive Lapita pottery complex, distribution of which suggests that 'Lapita peoples' had by 3,000 years ago moved as far south and east as hitherto uninhabited Fiji and western Polynesia. Lapita peoples spoke an early language of the Austronesian (AN or Malayo-Polynesian) family labeled Proto-Oceanic. Austronesian languages are today distributed along coastal eastern New Guinea and through island Melanesia; all Polynesian languages and most Micronesian ones are also descended from Proto-Oceanic. Hence the extraordinary linguistic, cultural, and biological diversity of Melanesia by comparison with Polynesia and Micronesia derives from at least two prehistoric sources: (a) processes of divergence and differentiation among relatively separated populations over the very long period of human occupation in the region, especially on the island of New Guinea, and (b) processes of mixing or creolization between the original NAN-speaking settlers and the later AN-speaking arrivals in island Melanesia.

2. *Social and Cultural Anthropology in/of Melanesia*

Dumont D'Urvilles' classificatory legacy has been provocatively criticized on different grounds, namely that 'Melanesia' often designates one term of an invidious comparison with 'Polynesia.' Such comparisons, manifestly racist in many nineteenth-century versions, portray Melanesians as economically, politically, and culturally less developed than Polynesians and, by evolutionary extension, people of 'the West.' To Western political and economic interests, moreover, Melanesia has historically appeared marginal; Melanesianist anthropologists, for example, never found much support within the post World War II

institutional apparatus of area studies. This marginality has rendered Melanesian studies as the exemplar of anthropology's authoritative discourse on 'primitives,' and contributed to the popular perception of Melanesian studies as the anthropological subfield most engaged in the contemporary study of radical (non-Western) Otherness. Such a perception not only turns Melanesian people into exotic vehicles for Western fantasy, but also suppresses the continuing history of colonialism in the region and misrepresents much Melanesianist anthropology, past and especially present.

Melanesian ethnography has played a central role in the history of the larger discipline's methods and theories. The work of W. H. R. Rivers on the 1898 Cambridge Expedition to the Torres Straits and later trips to Melanesia, and of Bronislaw Malinowski during his World War I era research in the Trobriand Islands, both mark important moments in the development of genealogical inquiry and extended participant observation as characteristic methods of social anthropology. Similarly, the Melanesian ethnography of M. Mead, G. Bateson, R. Fortune, and I. Hogbin, among others, guided anthropological approaches to questions of socialization and emotions, culture and personality, gender relations and sorcery, ritual process and social structure. From 1950 onwards, anthropology drew heavily on Melanesian ethnography—including the first waves of intensive fieldwork in the New Guinea Highlands—for diverse theoretical advances in the areas of cultural ecology, gender and sexuality, and in the study of metaphor, meaning, and knowledge transmission.

3. *Gift/Commodity*

Most prominent among the gate-keeping concepts that make Melanesia a distinctive anthropological place is exchange. Malinowski's classic description of *kula*, and Marcel Mauss's treatment of Malinowski's ethnography in *The Gift*, provided the grounds for understanding Melanesian sociality as fluid and contingent, performed through pervasive acts of give and take. In the 1960s, the critique of models of segmentary lineage structure in the New Guinea Highlands further enabled recognition of reciprocity and exchange as the Melanesian equivalents of the principle of descent in Africa. Exchange brings into being the social categories that it perforce defines and mediates: affines and kin, men and women, clans and lineages, friends and trade partners, living descendants and ancestral spirits. The rhythms of Melanesian sociality thus develop out of infrequent large-scale prestations associated with collective identities and male prestige; occasional gifts of net bags from a sister to her brother or of magic from a maternal uncle to his nephew; and everyday exchanges of cooked food between households and betel nut between companions.

Before contact with Westerners, Melanesia was laced by extensive trade networks that connected coastal and inland areas, and linked numerous islands and archipelagoes. These regional networks crossed different ecological zones and organized the specialized production of food, pigs, crafts, and rare objects of wealth such as pearl shells and bird plumes. Wealth objects were in turn often displayed and circulated in sequential chained exchanges, of which the most complex and geographically far-reaching include *kula*, and the Melpa *moka* and Enga *tee* systems of the central New Guinea Highlands. These exchange systems often define a sphere of male-dominated activity that depends on the labor of women, but instances of women circulating women's wealth in public exchanges do occur throughout Melanesia. During the colonial period, many trade networks atrophied, while other forms of competitive exchange effloresced, incorporating new items of wealth. State-issued money now commonly circulates in exchange alongside indigenous currencies and wealth, especially as part of marriage payments and homicide compensations.

Melanesians sometimes self-consciously contrast their own reciprocal gift-giving with the market transactions of modern Western society; for example, the distinction between *kastam* ('custom') and *bisnis* ('business') occurs in the English-based pidgin-creoles of the region. In related fashion, anthropologists have used an ideal–typical contrast between a gift economy and a commodity economy to describe the social relations of reproduction characteristic of indigenous Melanesia. For Gregory (1982), 'gift reproduction' involves the creation of personal qualitative relationships between subjects transacting inalienable objects, including women in marriage; and 'commodity reproduction' involves the creation of objective quantitative relationships between the alienable objects that subjects transact. This dichotomy provides a way of talking about situations now common in Melanesia in which the status of an object as gift or commodity changes from one social context to another, and varies from one person's perspective to another. Such situations occur, for example, when kin-based gift relations become intertwined with the patronage system associated with elected political office or when marriage payments become inflated by cash contributions derived from wage labor and the sale of coffee and copra.

While Gregory's formulation has been criticized as obscuring the extent to which commodity-like transactions—barter, trade, localized markets—were always a feature of Melanesian social life (and gifts a feature of highly commoditized Western societies), it nonetheless highlights how Melanesians produce crucial social relations out of the movement of persons and things. Anthropologists have thus used the concept of personhood as a complement to the concept of the gift in understanding Melanesian moral relationships; for it is in the context of gift exchange that the

agency of persons and the meaning of things appear most clearly as the effects of coordinated intentions. Marilyn Strathern contrasts a Melanesian relational person or dividual with a Western individual. Whereas Western individuals are discrete, distinct from and prior to the social relations that unite them, Melanesian dividuals are composite, the site of the social relations that define them. Similarly, whereas the behavior of individuals is interpreted as the external expression of internal states (e.g., anger, greed, etc.), the behavior of dividuals is interpreted in terms of other people's responses and reactions. The notion of dividuality makes sense of many Melanesian life-cycle rites that recognize the contributions of various kin to a person's bodily growth, as well as of mythological preoccupations with the management of ever-labile gender differences.

4. *Equality/Hierarchy*

Melanesian polities have often been characterized as egalitarian relative to the chiefly hierarchies of Polynesia; instances of ascribed leadership and hereditary office, numerous throughout insular Melanesia, have accordingly been regarded as somewhat anomalous, while the categorical distinction between 'chiefs' and 'people of the land' has served to mark Fiji's problematic inclusion within Melanesia. Traditional Melanesian political organization has been defined negatively, as not only stateless but also lacking any permanent form capable of defining large-scale territorial units. Precolonial warriors, such as those of south-coast New Guinea, distinguished themselves through vigorous practices of raiding, head-hunting, and cannibalism that ensured social and cosmological fertility. Local autonomy, armed conflict, and ongoing competition among men for leadership have thus been taken as the hallmarks of traditional Melanesian politics. This definition of precolonial Melanesian polities invokes a sharp contrast with the centralized, bureaucratic, violence-monopolizing states of the West. Yet, at the same time, Melanesian politics, epitomized in the figure of the self-made 'big man,' have sometimes been thought to recall values of individualism and entrepreneurship familiar from the West's own self-image.

Big men typically validate their personally achieved leadership through generous distributions of wealth. They manipulate networks of reciprocal exchange relationships with other men, thus successfully managing the paradox of generating hierarchy—even if only temporarily—out of equality. This paradox is variously expressed within Melanesia. In classic big-man exchange systems, known best from areas of high population density and intensive pig husbandry and horticulture in the New Guinea Highlands (Melpa, Enga, Simbu, Dani), a constant and imbalanced flow of shells, pigs, and money underwrites individual prestige. Such systems are not found everywhere in Melanesia, however, prompting Godelier (1986) to invent a new personification of power, the 'great man.' Great men emerge where public life turns on ritual initiations, where marriage involves the direct exchange of women ('sister exchange'), and where warfare once similarly prescribed the balanced exchange of homicides. The prestige and status of great men often depends upon their success in keeping especially valued 'inalienable possessions' out of general circulation. While this typological innovation has been challenged, political life in the Sepik area and in eastern Melanesia frequently involves transmitting ritual resources (heirlooms, myths, names, and so forth) within a delimited kin grouping such as a clan. In the Mountain Ok area, ancestral relics (bones and skulls) provide the material focus of a secret men's cult that links contiguous communities and complements clan affairs; in insular Melanesia, forms of voluntary male political association detached from kin groupings are found, such as the public graded societies of northern Vanuatu. Nevertheless, the scope for establishing political relations in precolonial Melanesia was restricted by the instability of big men's exchange relations, or by the demands placed on high-ranking men to sponsor youthful aspirants, or by the limited extent of the communities or kin groups beyond which ritual resources have no currency.

Whereas Melanesian political relations have been deemed egalitarian by comparison with the West, gender relations have usually been characterized as manifestly inegalitarian. Ethnography of highlands New Guinea in particular suggests that women's productive labor in gardening and pig husbandry is eclipsed if not wholly appropriated by men who act exclusively as transactors in public exchanges. In areas with once secret and lengthy rites of male initiation, senior men socialized boys into ideologies of dangerous female power and pollution while enforcing strict physical separation of the sexes. Ethnography of insular Melanesia (AN speakers), by contrast, suggests that gender inequality is less severe, tempered by women's social and genealogical roles as sisters rather than wives, and as both producers and transactors of cosmologically valued wealth items such as finely woven mats. But Margaret Mead's pioneering work on sex and temperament in the Sepik area, and Knauft's (1993) review of south-coast New Guinea ethnography suggest the extreme variation possible in gender relations among even geographically proximate societies.

New forms of social inequality and hierarchy have taken shape since the mid-nineteenth century through intensified interactions with foreign traders, missionaries, and government officials. During the colonial period, 'chiefs' were often appointed in areas where they had not existed, while indigenously recognized leaders were often able to use their status to monopolize access to the expanding cash economy.

Male labor migration put extra burdens on women who remained at home, their contacts with new ideas and values inhibited. In rural areas today, 'rich peasants' have emerged by manipulating customary land tenure—and sometimes using 'development' funds of the postcolonial state—to accumulate smallholdings of cash crops; conflicts have arisen over claims to cash crops that women raise on land owned by their husbands. In urban areas, political and economic elites attempt to reproduce their dominant class position through private education and overseas investments as well as cosmopolitan consumption practices.

5. *Melanesian Modernities*

The legacies of different colonial states (Great Britain, France, Germany, The Netherlands, Australia, and Indonesia) variously condition Melanesian political economies, especially with regard to control of land. In Papua New Guinea, limited land alienation by the crown, an indirect subsidy of contract labor, has preserved a vital resource in the defense against encroaching commodification; but French settlers in New Caledonia and, more recently, Javanese transmigrants in Irian Jaya violently displaced indigenous populations. In Fiji, complicated arrangements for leasing agricultural land, a consequence of British colonial policy, generate tension between Fijian landowners and their largely Indo-Fijian tenants, descendants of indentured plantation workers brought from India in the late nineteenth and early twentieth centuries. In addition, uneven capitalist markets and proliferating Christian denominations have added new dimensions to the diversity of Melanesia. Anthropologists readily acknowledge that even in those parts of Melanesia where the presence of foreign agents and institutions has been relatively weak (once a tacit justification for rendering colonial entanglements ethnographically invisible), received cultural orientations have become articulated with globalized elements of Western modernity. In other words, Melanesia is emerging as a key anthropological site for the comparative study of plural or local modernities.

Current Melanesianist research, historical as well as ethnographic, extends earlier studies of north-coast New Guinea cargo cults that sought to demonstrate how indigenous cosmologies shaped local interpretations of European material wealth and religious culture. This research highlights the syncretic products of intercultural encounters, noting how Melanesians have appropriated and reinterpreted such notable ideological imports as Christianity, millenarian and otherwise, and democracy. Even the impact of large-scale mining and timber projects has been shown to be mediated by highly local understandings of land tenure, kinship, and moral economy. In West Papua and Papua New Guinea, these projects define sites of struggle among transnational corporations, national states, and local stakeholders, often generating destructive environmental consequences (the Ok Tedi and Grasberg gold mines) and massive social dislocation (the civil war on Bougainville).

By the same token, current research emphasizes how the historical encounter with Western agents and institutions often transforms indigenous social practices, including language socialization, thus revaluing the meaning of tradition. For example, the flourishing gift economies of rural villagers are often a corollary of the remittances sent back by wage workers in town, while the ritual initiation of young boys might provide a spectacle for paying tourists from Europe and the USA. Since the 1970s, the emergence of independent nation-states has stimulated attempts to identify national traditions with habits such as betel-nut chewing or artifacts such as net bags, both of which have become much more geographically widespread since colonization. In many of these states, 'chieftainship' furnishes the newly traditional terms in which elected political leaders compete for power and resources. In New Caledonia, the assertion of a unifying aboriginal (Kanak) identity and culture has been a feature of collective anticolonial protest.

Contemporary Melanesian societies and cultures are thus defined by a tense double process. On the one hand, they are increasingly drawn into a world system through the spread of commercial mass media; the force of economic reforms mandated by interstate regulatory agencies; and the sometimes tragic exigencies of creating political communities on the global model of the nation-state. On the other hand, they continue to reproduce themselves in distinctive and local ways, organizing processes of historical change in terms of home-grown categories and interests.

See also: Big Man, Anthropology of; Colonialism, Anthropology of; Exchange in Anthropology; Kula Ring, Anthropology of; Malinowski, Bronislaw (1884–1942); Mauss, Marcel (1872–1950); Polynesia and Micronesia: Sociocultural Aspects

Bibliography

Douglas B 1998 *Across the Great Divide: Journeys in History and Anthropology*. Harwood Academic Publishers, Amsterdam

Foster R J (ed.) 1995 *Nation Making: Emergent Identities in Postcolonial Melanesia*. University of Michigan Press, Ann Arbor, MI

Gewertz D B, Errington F K 1999 *Emerging Class in Papua New Guinea: The Telling of Difference*. Cambridge University Press, Cambridge, UK

Godelier M 1986 *The Making of Great Men: Male Domination and Power Among the New Guinea Baruya* [trans. Swyer R]. Cambridge University Press, Cambridge, UK

Gregory C A 1982 *Gifts and Commodities*. Academic Press, New York

Jolly M 1994 *Women of the Place: Kastom, Colonialism and Gender in Vanuatu*. Harwood Academic Publishers, Philadelphia

Kirch P V 1997 *The Lapita Peoples: Ancestors of the Oceanic World*. Blackwell, Cambridge, MA
Knauft B M 1993 *South Coast New Guinea Cultures: History, Comparison, Dialectic*. Cambridge University Press, Cambridge, UK
Knauft B M 1999 *From Primitive to Postcolonial in Melanesia and Anthropology*. University of Michigan Press, Ann Arbor, MI
Kulick D 1992 *Language Shift and Cultural Reproduction: Socialization, Self and Syncretism in a Papua New Guinea Village*. Cambridge University Press, Cambridge, UK
Lederman R 1998 Globalization and the future of culture areas: Melanesianist anthropology in transition. *Annual Review of Anthropology* **27**: 427–49
Pawley A 1981 Melanesian diversity and Polynesian homogeneity: A unified explanation for language. In: J Hollyman, A Pawley (eds.) *Studies in Pacific Languages & Cultures in Honour of Bruce Biggs*. Linguistic Society of New Zealand, Auckland, NZ, pp. 269–309
Spriggs M 1997 *The Island Melanesians*. Blackwell, Oxford, UK
Strathern M 1988 *The Gender of the Gift: Problems with Women and Problems with Society in Melanesia*. University of California Press, Berkeley, CA
Weiner A B 1992 *Inalienable Possessions: The Paradox of Keeping-While-Giving*. University of California Press, Berkeley, CA

R. J. Foster

Memes and Cultural Viruses

1. Definition

The meme is a hypothetical unit of cultural transmission, sometimes restricted specifically to forms of culture which are transmitted by imitation. Memes are replicators in that they are copied (with imperfect heredity) when one individual imitates another. Given that memes have some properties (their outward effects on human behavior) which affect the rate at which they replicate, their copy fidelity, and their longevity, they are thus subject to natural selection and are expected to accumulate adaptations.

The key aspect of the meme hypothesis is that the adaptations generated after many generations of meme transmission (cultural evolution) are not necessarily expected to increase the reproductive success of the persons that carry the memes (as traditional sociobiological theory explains cultural traits), nor are they expected to aid any higher levels of organization such as familial groups (as functionalist strains of sociology and anthropology often explain the existence of cultural characteristics). Memes are expected to garner adaptations which aid the memes themselves as they spread through human social networks. In this way, memes are described as being selfish replicators.

Memes are sometimes referred to as 'cultural viruses,' in reference to their putatively maladaptive (or at least nonadaptive) nature. In this case, humans are referred to as hosts to parasitic memes. Host behavior (and in some cases, host phenotype in general) is not always under the control of the genotypes which built the host. In some cases, the host may be considered an extended phenotype, which acts in the interest of parasite genes. Thus the aggressive nature of rabid mammals is understood as a manipulation by the rabies parasite, which is spread by the saliva into bite wounds caused by the rabid animal. Memes are proposed to affect their spread in a similar manner, through the manipulation of a human host's behavior. From this perspective, much human behavior is the result of insidious cultural parasites which manipulate the instincts and motivations of human hosts in order to 'trick' host bodies into performing maladaptive behaviors which increase the reproductive success of the cultural traits (memes) themselves.

2. History

The term 'meme' was coined by the British biologist, Richard Dawkins, in *The Selfish Gene*, and its original definition is found in the following oft-copied excerpt:

> Examples of memes are tunes, catch-phrases, clothes fashions, ways of making pots or of building arches. Just as genes propagate themselves in the gene pool by leaping from body to body via sperm or eggs, so memes propagate themselves in the meme pool by leaping from brain to brain via a process which, in the broad sense, can be called imitation … If the idea catches on, it can be said to propagate itself, spreading from brain to brain (Dawkins 1976 p. 206).

Dawkins' text, however, was not a work on culture, but on biology. He had added the section on memes in order to demonstrate that the principals of natural selection and adaptation were likely applicable outside of the realm of genetic evolution. Dawkins claimed that 'I am an enthusiastic Darwinian, but I think Darwinism is too narrow a theory to be confined to the narrow context of the gene.' Dawkins wished to use the idea of meme as a way to enrich his thinking about evolution, to support a concept of universal Darwinism, and to shed some suspicion upon the nascent discipline of sociobiology, which was emerging out of evolutionary ecology, the academic milieu in which Dawkins worked. Thus from the onset, the meme hypothesis was proposed as an adjunct aspect of a theory of evolution, not as a serious proposal as a theory of culture. However, many of the aspects of meme theory have appealed to those who are on the fringe of social theory, especially to communities who have backgrounds in evolution, population genetics, and artificial life.

While Dawkins is credited with the origin of the word 'meme,' R. A. Fisher, a pioneer in the discipline of population genetics, had claimed as early

as 1912 that natural selection should apply to such diverse phenomena as 'languages, religions, habits, customs … and everything else to which the terms stable and unstable can be applied.' While the desire to apply ideas from evolutionary biology to social phenomena was nothing new, the way in which Fisher thought of the process was distinct from the inter-society competition which other early twentieth-century biologists thought of social evolution. He outlined in a paper as an undergraduate the first meme-like theory of cultural evolution, with his claim that playing cards had spread widely due to games like bridge, as the eager bridge player must find several other naive persons and teach them the rules of bridge to get a game. This way of thinking of cultural evolution at the level of the cultural trait, and not at the level of the society at large, differentiates meme theory from other forms of social theory derived from evolutionary biology.

While other terms have been introduced into the literature to describe putative units of culture (e.g., Lumsden and Wilson's culturegen, 1981), and some investigators have proposed explicit theories of cultural transmission without coining new terms to refer to their cultural units (Feldman and Cavalli-Sforza 1976) the term meme has achieved a comparatively wide distribution and has been added to dictionaries, further reinforcing its legitimacy. While the use of the word meme has spread in the popular literature, especially on the internet (see the *Journal of Memetics* at http://www.cpm.mmu.ac.uk/jom-emit/), its use within academic circles is limited. The term 'meme,' however, when compared to other terms which attempt to describe culturally transmitted units, has been particularly successful. A number of popular works on 'memes' have sprung up in recent years (Lynch 1996, Brodie 1996, Blackmoore 2000). These books are only tangentially related to work on cultural transmission theory and evolutionary biology. While they have produced substantial media splashes in the popular press, none of them has received a substantial positive response within academic circles. Their uncritical (positively zealous) position on memes has helped in associating the term meme with popular pseudoscientific literature. Time will tell whether the rapid spread and acceptance of the term meme in the absence of reasonable experiments or formal theory is used as evidence for, or against, the meme hypothesis.

3. *Context*

In the discussion within evolutionary biology from which the term meme arose, the scale at which natural selection operates (species, group, organism, or gene) was of great importance (Williams 1966). It is now recognized generally that selection at larger scales, while possible, is likely a weak force in comparison with selection at the smaller scales. That the meme is an important agent in cultural evolution (as opposed to the individual, the social class, or the society as a whole) is an outgrowth of this mode of thinking.

Scale is a key concept in the description of memes, but while there is agreement on the importance of scale, there is not yet a consensus on what scale is appropriate. Cultural phenomena on scales as small as the phoneme, and as large as entire religious systems, have been considered memes. Some have even claimed that all cultural phenomena within that wide range of scales can be viewed as memes (Durham 1991). However, most definitions of meme focus on relatively smaller scales.

Dennett (1995) claims that memes are 'the smallest elements that replicate themselves with reliability and fecundity,' while Pocklington and Best (1997) argue that 'appropriate units of selection will be the largest units of socially transmitted information that reliably and repeatedly withstand transmission.' From both of these perspectives, replicative integrity and repeated transmission are important. However, Pocklington and Best argue that the very smallest elements within culture are likely to be unable to accumulate adaptations, as they do not have subcomponents which can vary, providing grist for the mill of selection and preventing adaptation at the level of the replicator. They argue that the largest clearly coherent cultural material will be most likely to respond to selection pressure and generate self-serving adaptations. Thus, while there may be no single, uniform answer to the question, 'How large is a meme?,' a focus on scales of cultural phenomena makes evolutionary theory more useful in describing the observed phenomena.

Questions of scale in evolutionary theory are prominent in discussions of the topic of group selection and the repercussions of individual based vs. group based models as possible explanations for human altruism. Those arguing for true human altruism often suggest group selection as a mechanism through which the differential survival of human populations has caused group-level adaptations. The more cynical gene and individual selectionists usually describe all altruism as being the reflection of underlying selfishness and suggest that true altruism would be unlikely (if not impossible) to evolve. Thus, within this debate, important moral questions are claimed to be resolved by a better understanding of the scale at which natural selection works on human populations. The addition of the idea of selection at the level of the meme, as opposed to individual level selection (or group selection), is a simple continuation of this general argument.

While the term meme is sometimes used simply to refer to a 'cultural particle' without reference to concepts of selfish cultural adaptation, the majority of work making use of the term meme makes explicit analogies to the exploitation of host animals by parasites, and suggests that culture itself creates culture, with humans and their societies as inter-

mediate 'vehicles.' A substantial exception to this is the work on bird song making use of the term meme (Lynch and Baker 1993). In this literature, the term meme is used simply to refer to bird song types, with no reference to selfish cultural adaptation. Similarly, there is substantial work on cultural transmission using quantitative models (mostly from population genetics) which does not make use of the term meme.

In summary, the central concept that differentiates the idea of a meme from the idea of culture in general is that memes, having replicated (through a process of imitation) are expected to evolve adaptations that do not help their host, but instead will have adaptations that serve their own replications, much as parasites do. Neither human reproductive success nor social function at any higher level of scale needs be involved. Thus memetics is inherently a functionalist and adaptationist viewpoint, but with an important shift of emphasis away from the individual (as sociobiological theories emphasize) and the social group (as traditional anthropological functionalism emphasizes) to the cultural trait itself. This conclusion has important consequences for many areas of social thought which have yet to be explored seriously. At this point, the meme is purely hypothetical, merely the lineal descendant of one of Darwin's more dangerous ideas.

See also: Adaptation, Fitness, and Evolution; Cognitive Psychology: Overview; Collective Memory, Anthropology of; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Evolution: Diffusion of Innovations; Genes and Culture, Coevolution of; Natural Selection; Sociobiology: Overview; Sociology: Overview

Bibliography

Blackmoore S B 2000 *The Meme Machine*. Oxford University Press, Oxford, UK

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Brodie R 1996 *Virus of the Mind: The New Science of the Meme*. Integral Press

Bull J 1994 Virulence. *Evolution* **48**: 1423–37

Cavalli-Sforza L L, Feldman M W 1981 *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press, Princeton, NJ

Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford, UK

Dawkins R 1982 *The Extended Phenotype: The Gene as the Unit of Selection*. W H Freeman, Oxford, UK

Dennett D C 1995 *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon and Schuster, New York

Durham W H 1991 *Coevolution: Genes, Culture and Human Diversity*. Stanford University Press, Stanford, CA

Feldman M W, Cavalli-Sforza L L 1976 Cultural and biological evolutionary processes. *Theoretical Population Biology* **9**: 238–59

Lumsden C J, Wilson E O 1981 *Genes, Mind and Culture*. Harvard University Press, Cambridge MA

Lynch A 1996 *Thought Contagion: How Belief Spreads Through Society*. Basic Books, New York

Lynch A, Baker A J 1993 *A* population memetics approach to cultural-evolution in chaffinch song: Meme diversity within populations. *American Naturalist* **141**: 597–620

Pocklington R, Best M L 1997 Cultural evolution and units of selection in replicating text. *Journal of Theoretical Biology* **188**: 79–87

Williams G C 1966 *Adaptation and Natural Selection*. Princeton University Press, Princeton, NJ

Wilson E O 1975 *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA

R. Pocklington

Memory and Aging, Cognitive Psychology of

'As I get older, can I maintain or even improve my memory for names?' Questions such as these motivate a strong practical interest in the relation of memory and aging. The diversity of practical issues related to memory problems is easily matched by the conceptual sophistication of memory-related taxonomies emerging from cognitive research (e.g., short-term or working memory vs. long-term memory, episodic vs. semantic memory, explicit vs. implicit memory). Indeed, the observation that age differences vary by the type of memory under consideration is invariably part of the empirical justification for these distinctions. Furthermore, in the field of cognitive aging a number of rival accounts have been put forward with the claim that age-related memory problems are foremost a consequence of a general weakening of cognitive processing efficiency which is traced conceptually to a decline in a hypothetical mental processing resource, a decline in the speed of processing, or a decline in inhibition of irrelevant information—to name but three prominent examples. Accordingly, this review of research on memory and aging will be structured under three superordinate questions: Do memory systems age differently? Are memory-related age differences an epiphenomenon of a more general deficit? Can one prevent or even reverse age-related memory declines with training programs?

1. Differential Aging of Memory Systems

Research on the varieties of memory have always been at the core of cognitive research and invariably found their way into aging research (for reviews see Light 1991, Zacks et al. 2000; for a less technical review see

Schacter 1997). The most traditional distinction is the one between short-term memory and long-term memory. Within each of these categories, research has spawned more refined subcategories.

1.1 Short-term Memory, Working Memory, and Executive Control Processes

The traditional measure of *short-term memory* is the memory span task which measures the number of items (e.g., digits, words) presented at a fast pace that can be reproduced in correct serial order with a 50 percent success rate (forward span). Older adults exhibit a minor but reliable decline in this measure; age effects are much more pronounced when items must be reported in the reverse order of presentation (backward span). This pattern of results maps well onto the more recent distinction between phonological loop and central executive in the *working memory* framework. Forward span seems to rely on the phonological loop. In contrast, backward span requires the reorganization of information before production, presumably in the central executive component. Some of the largest (i.e., disproportionate) effects of age have been reported for tasks requiring such simultaneous processing and storage of information. This specific age deficit is mirrored by age-related physiological differences in the frontal cortex, specifically the frontostriatal dopaminergic system (for a review see Prull et al. 2000). Recent proposals question the notion of a single central executive and argue that processing costs and associated age effects are tied to the implementation of specific *executive control processes* which are linked strongly to specific task domains. For example, important executive control processes required for reasoning about novel figural material show very large decline with age, whereas those involved in sentence comprehension do not.

1.2 Semantic, Episodic, and Source Memory

One of the most striking dissociations of age and long-term memory relates to the distinction between *semantic memory* (factual knowledge) and *episodic memory* (knowledge tied to specific coordinates of place and time). Interestingly, age invariance in semantic memory appears to hold not only in terms of amount of knowledge (as is long known from the adult-age stability of crystallized intelligence) but also in terms of speed of accessing or retrieving information (see also Sect. 2.2 below). Such results obviously limit proposals of general age-related slowing. Moreover, age invariance in accuracy of comprehension is found for many other syntactic, semantic, and pragmatic aspects of language use once the influence of contributions of executive control processes is taken into account (for a review see Kemper and Kliegl 1999).

In contrast to semantic memory, older adults exhibit reliable deficits in episodic memory (i.e., in the ability to learn new material; for a review see Kausler 1994 and Smith 1996). Despite a long research tradition it still remains unclear whether the problem relates to encoding or retrieval components of the task, but most agree that storage of information appears to be least affected. There is quite solid evidence pointing to a specific age-related increase in susceptibility to proactive interference (i.e., the harmful consequences of old age for the acquisition of new knowledge). In general, however, older adults respond like younger ones to experimental manipulations of material to be learned (e.g., degree of organization, familiarity, concreteness of words).

A promising line of current research focuses on the defining component of episodic memory. Older adults have problems in generating or retrieving contextual details or the source of newly acquired information (for a review see Johnson et al. 1993). For example, they are worse than young adults in recalling the color in which an item was presented and they recall less well who told them about a specific event. Problems with *source memory* impede the ability to update or revise one's memory (e.g., the current parking slot of the car, keeping track of developments in a scientific field). They obviously might underlie difficulties of older adults in a very wide spectrum of everyday activities.

1.3 Procedural and Implicit Memory

The research reviewed in Sects. 1.1 and 1.2 focused largely on deliberate or explicit memory tasks. One prominent recent line of research examines *procedural or implicit memory*, that is, the influence of past experiences (i.e., memories) on current performance unbeknownst to the person or without their deliberate attempt. Motor skills are the prime example of procedural memory; their availability even after years of disuse is well known and there is good evidence that this remains so into old age. At a somewhat subtler level it has been shown that reading a word will make this word more perceptible later in a word identification task. Most importantly, such implicit memories appear to exert their effects almost equally strongly in younger and older adults and across many days. Conceivably, this type of memory may turn out not to be affected by normal aging after removal of task contaminations due to executive control processes.

2. Age Effects on Memory as an Indicator of General Processing Efficiency

Distinctions between different memory systems originating in cognitive research map onto a varied pattern of aging with a strong gradation. As age is an

individual difference, not an experimental variable (i.e., we cannot randomly assign people to be young or old), age differences in memory tasks are typically correlated with other cognitive abilities (such as general intelligence). Consequently, age differences reported in cognitive memory research might reflect the participants' general cognitive processing efficiency. This section reviews the most prominent accounts in this line of research.

2.1 Decline of Processing Resources

Age differences in memory tasks could result from age differences in a limited processing resource such as, for example, processing speed or working memory capacity. The amount of resources needed for successful performance depends also on characteristics of the memory task, such as the amount of cues provided to support retrieval. Thus, age differences in memory are quite sensitive to the particular constellation of internal (i.e., decline of resources) and external factors (i.e., environmental support) associated with a task (Craik 1983). Moreover, limits of processing speed could be the cause or the consequence of limits of working memory capacity or its associated subsystems. For example, with faster cognitive processing one can use more information from working memory before it decays. Alternatively, the number of processing steps for solving a given task, and hence the total processing time required, may depend inversely on the size of working memory capacity. In correlational research, regression analyses have been used to determine the chain of causation (e.g., Salthouse 1996). Measures of processing speed (e.g., WAIS digit-symbol substitution) mediate age differences in measures of working memory capacity (e.g., computational span) much better than vice versa. Measures of processing speed also account for a large share of age-related individual differences in episodic memory tasks. Note, however, that psychometric measures of speed and working memory are complex cognitive tasks with an unclear contribution of the processes which they attempt to explain.

An alternative approach to link external and internal resources is to determine the amount of presentation time (i.e., an external resource) needed by younger and older adults for the same level of accuracy across a wide variety of tasks or experimental conditions as a proxy of the inverse of the internally available processing resource (Kliegl et al. 1994). The age ratio of such time demands can be determined at several accuracy levels and varies distinctly across processing domains. For example, as mentioned above, older and younger adults do not differ at all in semantic memory access times, whereas older adults need about 1.5 to twice the time of younger adults for responses to relatively novel environmental stimuli as long as stimulus–response associations can be characterized in terms of one-step if-then rules ('If red, press left'). And older adults need about four times the amount of presentation time of younger adults for complex working memory tasks requiring the coordination of two or more such rules and for a wide variety of list memory tasks (e.g., different list lengths, accuracy levels, word material). Finally, there are task domains (e.g., complex memory updating operations) with age differences in the asymptotic accuracy reached at sufficiently long presentation times. Slowing of processing could cause such effects if intermediate products of processing needed for later processes are lost due to time-based decay or interference (Salthouse 1996). The mechanisms generating different domain specific slowing functions are not yet clear. One expectation is that they will emerge with a better understanding of executive control processes (see Sect. 1.1 above).

2.2 Inhibitory Control

Age-related memory problems have also been linked to a decline in attentional *inhibitory control* over the contents of working memory (for reviews see Hasher et al. 1999, Zacks et al. 2000). The basic idea is that it is difficult for older adults to keep task-irrelevant thoughts from interfering with ongoing mental activities and, in the case of episodic memory tasks, that this leads to less efficient encoding and retrieval operations. Hasher et al. (1999) distinguish between *access*, *deletion*, and *restraint* functions of inhibitory control which characterize efficient cognitive processing. The role of the access function is to prevent goal-irrelevant material (e.g., autobiographical associations in the context of writing a grant proposal) from access to consciousness. The deletion function serves to suppress material already active but no longer needed (e.g., outdated instructions). The restraint function concerns primarily the inhibition of situationally inappropriate responses and links to the larger domain of action regulation (e.g., off-target speech in conversational settings).

This approach shares considerable overlap with attempts to reconceptualize working memory in terms of executive control processes (see Sect. 1.1) and, therefore, holds considerable potential for convergence of two important current lines of research. There are, however, questions about its adequacy with respect to language functions. For example, off-target speech, which is a prime example of inhibitory control problems, generates systematically higher ratings of interestingness irrespective of age and may reflect subtle positive age differences in sensitivity towards conversational situations rather than an age deficit (Burke 1999). The approach has also been applied to account for age-differential effects of arousal and

circadian rhythms on memory performance and thus provides a theory-guided perspective on the context sensitivity of memory functions.

3. *Memory Training*

Given the ubiquity of age-related memory complaints, the question of prevention or reversal of decline has received considerable attention in research, mostly in laboratory settings (for a review see Camp 1998). Most commonly, participants learned to apply mnemonic techniques (e.g., forming vivid mental images between word pairs) to improve their recall of word lists. Statistical analyses of the effect sizes of 49 such memory interventions based on pretest-posttest designs showed that training and use of mnemonics improved performance in comparison to a same-age control group (Verhaeghen et al. 1992). As in many other domains of skill acquisition, transfer was quite limited, that is, training gain was restricted to the task in the focus of the training. With respect to age-differential training gains, younger adults benefit more than older adults from memory training but there is also good evidence that older adults can clearly surpass untrained younger adults with a mnemonic strategy tailored to the task (Baltes and Kliegl 1992). Studies about the stability of the intervention found that memory techniques apparently are remembered quite well over at least three years but also that people do not use them for lack of opportunity or because more efficient means to circumvent the problem are available (e.g., external aids). There are, however, also successful interventions demonstrating, for example, improved medication adherence after instruction in mnemonic devices.

Current research focuses on factors that lead to an efficient deployment of memory strategies in everyday life (Camp 1998). One problem of earlier intervention research was that the gap between laboratory memory tasks and real-life demands on memory was simply too large. With affordable and accessible computer technology it is now possible to implement realistic simulations of everyday memory tasks and of training programs tailored to the individual's level of performance. Moreover, some memory tasks such as associating faces and names likely require a skill acquisition course similar to one envisioned for learning to play a new musical instrument or learning to speak a new language, including systematic coaching and deliberate practice (Kliegl et al. 2000). Such expertise programs aim at circumventing past performance limitations with qualitatively different strategies and task-specific knowledge. They have already been shown to enable, for example, the acquisition of a digit span of 120 by a 70-year-old adult (Kliegl and Baltes 1987). However, taking into account the necessary investment of time, not many of those persons complaining about a poor face–name memory are likely to be motivated enough to carry out a strenuous training program and rather prefer to sustain the embarrassment of occasionally not remembering a name. Information about such cost–benefit ratios will become highly relevant for assessing the quality of cognitive interventions in general; the area of memory and aging appears to be particularly well suited for the development of prototypes in this respect.

4. *Conclusion*

Research on memory and aging has been a fertile testing ground for mainstream cognitive theories of memory as well as for the development of theories of cognitive aging. Distinctions between different memory systems receive considerable support from interactions with age. The search for the level at which a unifying theoretical framework can be formulated productively remains an important goal. Recent advances in monitoring of brain activities may yield new perspectives and constraints for the next steps of theoretical development. Two limitations must be mentioned. Research in this field is typically based on cross-sectional contrasts of students (18–25 years of age) and healthy, well-educated older adults (65–75 years). Therefore, results are not representative of the population at large but reflective of an upper-bound estimate given fortunate life circumstances; research on pathological older populations, especially dementia, was not covered (for a review of interventions see Camp 1998). Age effects may also be contaminated with effects due to birth cohort (i.e., older adults were born and raised in different times and cultural settings). In general, however, the core results presented here appear to hold up in longitudinal research (for a very extensive report on longitudinal memory changes see Hultsch et al. 1998). Over 90 percent of older adults express concern about their memory when asked about the downsides of age. Obviously, memory and aging is not only a multifaceted theoretical challenge but also a domain with many opportunities for cognitive psychology to prove its value in real-life settings.

Bibliography

Baltes P B, Kliegl R 1992 Further testing of limits of cognitive plasticity: Negative age differences in a mnemonic skill are robust. *Developmental Psychology* **28**: 121–5

Burke D M 1999 Language production and aging. In: Kemper S, Kliegl R (eds.) *Constraints on Language: Aging, Grammar, and Memory*. Kluwer Academic Publishers, Boston, pp. 3–28

Camp C J 1998 Memory interventions for normal and pathological older adults. *Annual Review of Gerontology and Geriatrics* **18**: 155–89

Craik F I M 1983 On the transfer of information from temporary to permanent memory. *Philosophical Transactions of the Royal Society of London* **B302**: 341–59
Hasher L, Zacks R T, May C P 1999 Inhibitory control, circadian arousal, and age. In: Gopher D, Koriat A (eds.) *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application*. MIT Press, Cambridge, MA, pp. 653–675
Hultsch D F, Hertzog C, Dixon R A, Small B J 1998 *Memory Change in the Aged*. Cambridge University Press, Cambridge, UK
Johnson M K, Hashtroudi S, Lindsay D S 1993 Source monitoring. *Psychological Bulletin* **114**: 3–28
Kausler D H 1994 *Learning and Memory in Normal Aging*. Academic Press, San Diego, CA
Kemper S, Kliegl R (eds.) 1999 *Constraints on Language: Aging, Grammar, and Memory*. Kluwer Academic Publishers, Boston
Kliegl R, Baltes P B 1987 Theory-guided analysis of mechanisms of development and aging through testing-the-limits and research on expertise. In: Schooler C, Schaie K W (eds.) *Cognitive Functioning and Social Structure Over The Life Course*. Ablex, Norwood, NJ, pp. 95–119
Kliegl R, Mayr U, Krampe R T 1994 Time-accuracy functions for determining process and person differences, an application to cognitive aging. *Cognitive Psychology* **26**: 134–64
Kliegl R, Philipp D, Luckner M, Krampe R 2000 Face memory skill acquisition. In: Charness N, Park D C, Sabel B (eds.) *Aging and Communication*. Springer, New York, pp. 169–186
Light L L 1991 Memory and aging: Four hypothesis in search of data. *Annual Review of Psychology* **42**: 333–76
Prull M W, Gabrieli J D E, Bunge S A 2000 Age-related changes in memory: A cognitive neuroscience perspective. In: Craik F I M, Salthouse T A (eds.) *Handbook of Aging and Cognition*. L. Erlbaum Associates, Hillsdale, NJ, Vol. 2, pp. 91–153
Salthouse T A 1996 The processing-speed theory of adult age differences in cognition. *Psychological Review* **103**: 403–28
Schacter D L 1997 *In Search of Memory: The Brain, the Mind, and the Past*. Basic Books, New York
Smith A D 1996 Memory. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 236–250
Verhaeghen P, Marcoen A, Goossens L 1992 Improving memory performance in aged through mnemonic training: A meta-analytic study. *Psychology and Aging* **7**: 242–51
Zacks R T, Hasher L, Li K Z H 2000 Human memory. In: Craik F I M, Salthouse T A (eds.) *Handbook of Aging and Cognition*. L. Erlbaum Associates, Hillsdale, NJ, Vol. 2, pp. 293–357

R. Kliegl

Memory and Aging, Neural Basis of

A predominant view of the neural basis of memory and aging in the twentieth century was that tens of thousands of neurons were lost on a daily basis in older adults resulting in impaired cognition and memory in old age. The early twenty-first century view of this phenomenon has changed rather dramatically. In the 1990s there was growing recognition of the fact that neuronal numbers in many brain structures remain relatively constant. Nevertheless, on a functional level, learning and memory are impacted by processes of aging. Research on the neural basis of memory and aging has identified neurochemical and electrophysiological correlates of age-related memory impairment in some older animals, including some older humans. Other older individuals show little impairment. The inclusion of older adults in early stages of neurodegenerative diseases in studies of normal aging has also contributed to a perspective of functional decline in the whole population. Both the magnitude of neuronal loss and the degree of functional memory impairment were likely overestimated in previous research.

1. Neuronal Counts and Aging

It has become evident with advances in research techniques and technology that traditional cell counting methods were flawed, often under-representing the number of cells in older tissue. Stereological cell counting techniques that provide unbiased estimates of cell numbers have replaced cell density measures. The application of stereological techniques in studies of aging in the nervous system has resulted in a revision of the view that large numbers of neurons die during normal aging, at least for some structures. Age-related loss of larger numbers of neurons in memory-related circuits does not appear to be a concomitant of normal aging but rather a consequence of dementing diseases such as Alzheimer's disease (AD; Morrison and Hof 1997). Medial-temporal lobe structures including the hippocampus and the neocortex are essential for the form of memory used for conscious recollection of facts and events. Although there is age-related decline in this form of memory, neuronal loss in the medial-temporal lobes does not appear to be the cause of functional impairment.

2. Brain Memory Systems, Normal Aging, and Neuropathological Aging

Human long-term memory systems have been classified into two major forms: declarative (or explicit) and nondeclarative (implicit or procedural). Declarative learning and memory, which can be conceptualized as learning with awareness, refers to the acquisition and retention of information about events and facts. It is assessed by accuracy on recall and recognition tests (e.g., word-list learning). Four decades of research support the conclusion that medial-temporal/diencephalic memory circuits are critical for establishing long-term memory for events and facts (Squire

1992). In contrast, there is considerable evidence that medial-temporal/diencephalic memory circuits are not critical for several kinds of nondeclarative memory.

Brain substrates of declarative and nondeclarative memory systems can function in parallel and independently from one another. Within the nondeclarative category, brain substrates of various forms of learning and memory are often physically remote from one another and composed of qualitatively different types of neurons, neurotransmitters, and projections. Most importantly for an understanding of the neural basis of memory and aging, the various brain memory systems are probably affected differentially by processes of aging.

2.1 Declarative Learning and Memory

Among brain memory systems, it is the declarative form that draws the greatest research attention in gerontology. The subjective awareness of 'trying to learn' is typically present in tasks that assess declarative learning and memory. Lesions in the medial-temporal lobe region including the hippocampus are associated with deficits in declarative learning and memory. It has become increasingly evident that lesions in these regions are not the consequence of normal aging but rather a part of the neuropathology of AD. Early in the course of AD, impairments in memory and cognition are hardly detectable. It is difficult to distinguish these patients with early AD from normal older adults. Yet, in the earliest stages of the disease, as many as half of the neurons in some critical layers of the medial-temporal lobes essential in the declarative memory circuit are lost. Without a doubt, many studies of normal aging have included in the sample individuals with early AD.

For example, there was evidence of hippocampal atrophy from radiological brain scans in about one-third of 154 community-residing older adults ranging in age from 55 to 88 years. These 'normal' older adults participated in a longitudinal study of aging and memory (Golomb et al. 1993). It was primarily the participants with hippocampal atrophy who performed poorly on tests of immediate and delayed recall of prose paragraphs and paired associate words. Participants with hippocampal atrophy were probably in the early stages of AD. Stereological cell counting techniques of post mortem brain tissue demonstrated that almost complete loss of pyramidal cells in the CA1 field of the hippocampus was a characteristic feature of AD that did not occur in normal aging (West et al. 1994).

Magnetic resonance imaging (MRI) assessment of hippocampal atrophy has become useful in the diagnosis of AD early in the course of the disease (Jack et al. 1997). Older adults with mild cognitive impairment who are at risk of AD have significant volume reduction in the hippocampus. Furthermore, individuals at risk of developing AD but presymptomatic at the initial testing develop memory impairment that is paralleled with hippocampal loss when followed up a year or two after initial MRI testing. There are observations from many laboratories including our own of an association between hippocampal atrophy in the early states of degenerative disease and declarative memory impairment (see Fig. 1).

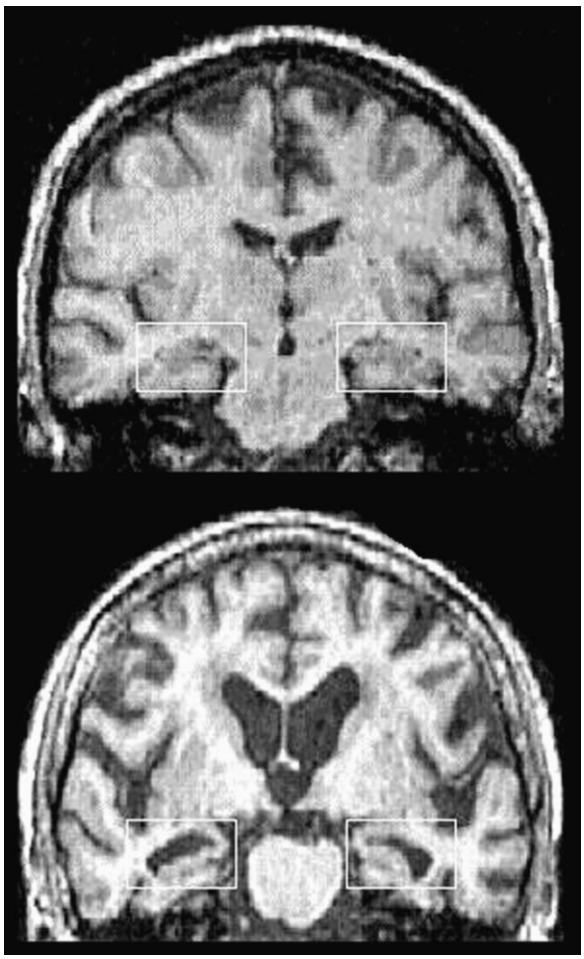

Figure 1
MRI of a coronal view of the cerebral cortex and medial temporal lobes, including the hippocampus (indicated by a box) in two older women. The 80-year-old woman in the top plate has little hippocampal atrophy and intact declarative memory function. The 95-year-old woman in the bottom plate has extensive hippocampal atrophy (as indicated by the absence of tissue in the hippocampus) and moderate impairment on memory tests

Whereas there is only limited neuronal loss in medial-temporal lobe regions in normal aging, declarative memory shows functional age-related deficits

in a majority of older adults, with recall affected more than recognition. Age differences in memory processes such as the encoding of new information that engages the prefrontal cortex as well as medial-temporal lobes have also been documented. If neuronal loss cannot account for the cognitive impairment, what neural mechanisms might cause the behavioral deficits?

Age-related changes in neurotransmitter levels and in electrophysiological properties have been associated with impaired memory. However, these neurobiological changes vary among individuals. Individual differences in memory abilities are larger in old age in a number of species, with some older animals performing as well as the young, while other older animals are seriously impaired. At a neurotransmitter level, there is evidence that the wide variation in acetylcholine input to the hippocampus is associated with the greater variability in memory ability in older rats. A marker for acetylcholine, choline acetyltransferase (ChAT), was significantly reduced in the medial septal region only in older rats with memory impairments (Gallagher and Rapp 1997). Nevertheless, the ChAT levels and memory performance in some older rats is equal to that of young rats. There are many electrophysiological properties of neurons in aging rodents that remain intact, but one learning and memory-related phenomenon that may become impaired is long-term potentiation (LTP; Barnes 1998). Again, the pattern is that some older organisms that exhibit severely impaired performance on memory tasks have lower levels of LTP, but some older organisms have LTP and memory performance levels similar to levels in young animals.

2.2 Nondeclarative Learning and Memory

Nondeclarative learning and memory may be conceptualized as 'learning without awareness' that is measured by changes in performance. Nondeclarative learning and memory consist of multiple, dissociable processes, including (a) repetition priming effects, (b) the acquisition and retention of motor, perceptual, or problem solving skills, and (c) simple forms of classical conditioning. Repetition priming is assessed by a decrease in reaction time or a bias in response to particular words or patterns as a consequence of prior exposure (during the experiment) to those words or patterns. Skill learning is assessed by improvement in speed and accuracy across trials on repetitive sensorimotor tasks. Classical conditioning is assessed with the pairing of a neutral and a reflex-eliciting stimulus and is observed when the neutral stimulus elicits the reflexive response. The various neural substrates of nondeclarative learning and memory are physically remote from one another and have little overlap. Brain imaging studies in humans have demonstrated that circuitry in the cerebellum is activated in eyeblink classical conditioning. Motor skill learning engages the motor cortex. Repetition priming involves the occipital cortex with the frontal cortex engaged in priming tasks requiring production.

The initial brain memory systems perspective of aging was that nondeclarative forms of learning and memory remained relatively intact whereas declarative forms were impaired. In the case of repetition priming, additional investigation has revealed a more complex perspective. An overview of the literature indicates a mild but consistent reduction in priming in normal aging, but the underlying substrate for this result is unknown. Mild age-related impairment in repetition priming may reflect reduced priming in a subgroup of older adults who are aging less successfully or have as yet undetected AD, or it may represent an impairment that all older adults experience as a consequence of normal aging (Fleischman and Gabrieli 1998).

The studies that have examined motor skill learning in normal aging are not consistent in their results, but the perspective again is one of minimal age-related effects. Various studies of skill learning and aging have reported three different kinds of results: superior performance by older adults, equal performance by young and old, and superior performance by young adults. Comparing age-matched older adults to patients with a diagnosis of probable AD, Eslinger and Damasio (1986) found no differences in acquisition of a rotary pursuit task. However, the AD patients were seriously impaired on tests of declarative memory. Again, the perspective for motor skill learning is that there is a mild effect of normal aging that is less dramatic than observed on declarative memory tasks.

A different perspective emerges from studies of simple eyeblink classical conditioning. Since the first studies were carried out in the 1950s comparing young and older adults on this task, striking age differences were apparent. When adults over the age range of 20 to 90 are tested, age-related effects appear in the decade of the 40s. Direct comparisons of a declarative memory measure (California Test of Verbal Learning) and a nondeclarative measure (400 ms delay eyeblink classical conditioning) in the same young and older adults revealed a larger age effect on the nondeclarative than the declarative task (Woodruff-Pak and Finkbiner 1995).

Like all investigations of normal aging, when participants have subclinical pathology (as in the early stages of AD) performance is affected on eyeblink classical conditioning. Indeed, several laboratories have demonstrated that patients diagnosed with probable AD are severely impaired on this task. However, in elderly adults with normal neuropsychological profiles, it is the status of the cerebellum that is related to performance on eyeblink conditioning. The correlation between eyeblink conditioning performance and cerebellar volume corrected for total brain volume was .81 ($p < .02$) in a sample ranging in age from 77 to 95 years (Woodruff-Pak et al. 2000).

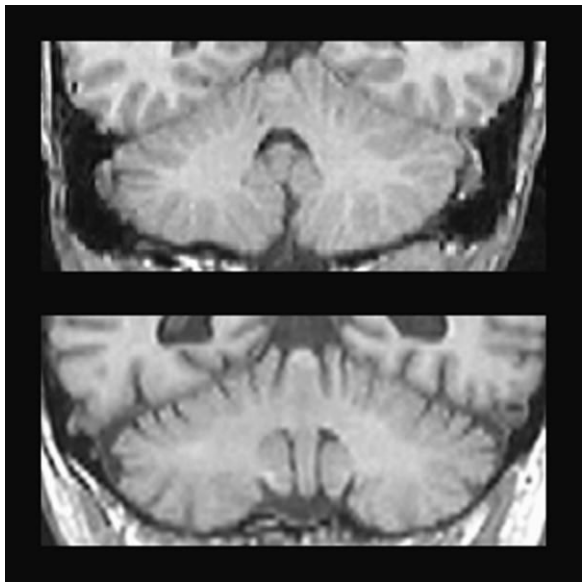

Figure 2
MRI of a coronal view of the cerebellum of a young man (20 years old; Top) and of an older man (79 years old; Bottom). Shrinkage in the older cerebellum is evident as deep sulcal grooves in cerebellar cortex

Until quite recently, the role of the cerebellum in cognition received little research attention whereas the role of medial-temporal lobe structures including the hippocampus in memory has been a central focus of research in cognitive neuroscience. Studies of cognition in normal aging have overlooked the potential role of the cerebellum. There are striking differences in cerebellar volume between young and older adults (see Fig. 2), and tasks that are cerebellar-dependent show age-related deficits earlier than hippocampally-dependent tasks. Further exploration of the role of the cerebellum in cognitive and aging processes is warranted.

See also: Aging Mind: Facets and Levels of Analysis; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Cognitive Aging; Memory and Aging, Cognitive Psychology of; Social Cognition and Aging; Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

Bibliography

Barnes C A 1998 Spatial cognition and functional alterations of aged rat hippocampus. In: Wang E, Snyder D S (eds.) *Handbook of the Aging Brain*. Academic Press, San Diego, CA

Eslinger P J, Damasio A R 1986 Preserved motor learning in Alzheimer's disease: Implications for anatomy and behavior. *Journal of Neuroscience* **6**: 3006–9

Fleischman D A, Gabrieli J D E 1998 Repetition priming in normal aging and Alzheimer's disease: A review of findings and theories. *Psychology and Aging* **13**: 88–119

Gallagher M, Rapp P R 1997 The use of animal models to study the effects of aging on cognition. *Annual Review of Psychology* **48**: 339–70

Golomb J, de Leon M J, Kluger A, George A E, Tarshsh C, Ferris S H 1993 Hippocampal atrophy in normal aging: An association with recent memory impairment. *Archives of Neurology* **50**: 967–73

Jack C R Jr., Petersen R C, Xu Y C, Waring S C, O'Brien P C, Tangalos E G, Smith G E, Ivnik R J, Kokman E 1997 Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* **49**: 786–94

Morrison J H, Hof P R 1997 Life and death of neurons in the aging brain. *Science* **278**: 412–19

Squire L R 1992 Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* **99**: 195–231

West M J, Coleman P D, Flood D G, Troncoso J C 1994 Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. *The Lancet* **344**: 769–72

Woodruff-Pak D S, Finkbiner R G 1995 Larger nondeclarative than declarative deficits in learning and memory in human aging. *Psychology and Aging* **10**: 416–26

Woodruff-Pak D S, Goldenberg G, Downey-Lamb M M, Boyko O B, Lemieux S K 2000 Cerebellar volume in humans related to magnitude of classical conditioning. *NeuroReport* **11**: 609–15

D. S. Woodruff-Pak and S. K. Lemieux

Memory: Autobiographical

The term *autobiographical memory* refers to memory for the events of our lives and also to memory for more abstract personal knowledge such as schools we attended, people we had relationships with, places we have lived, places we have worked, and so on. Autobiographical memory is, then, an intricate part of the self and one of the more complex forms of cognition in which knowledge, emotion, identity, and culture, all intersect during the course of remembering: It is the complex nature of this intersection that will be considered here.

1. Autobiographical Knowledge

In general a distinction can be drawn between memory for highly specific details of events and memory for more conceptual or abstract aspects of experience. For example, a person remembering a holiday from some years ago might recall a hot day on a particular beach, the heat of the sand on bare feet, the rhythmic sound of the waves, the coolness of the water, other people talking, etc., and these details may come to mind in the

form of images, feelings, and sensations. This *event-specific knowledge*, or ESK, is near to the sensory experiences from which they originate, they preserve some of the phenomenal nature of moment-by-moment conscious experience (Conway and Pleydell-Pearce 2000), and when we recall them they trigger *recollective experience* resulting in a sense of the self in the past (Gardiner and Richardson-Klavehn 2000, Tulving 1985, Wheeler et al. 1997).

Probably the best description of this experience of remembering is provided by the great French writer Marcel Proust, who relates how the taste of a madeleine cake dipped in warm tea led to a sudden onrush of childhood autobiographical memories. In Proust's account, however, the cue (the taste of the cake) occasions not just the recall of ESK but a whole flood of sensations, recollections of knowledge, and feelings for his previously 'forgotten' childhood (see Proust [1913] 1981). Indeed, a hallmark of autobiographical memories is that they always contain several different types of knowledge, some which is specific and some general. Thus, the person recalling the day on the beach might also recall that this occurred during a holiday in Italy, which in turn took place when the children were little. Conway and Pleydell-Pearce (2000) refer to these types of more abstract or conceptual autobiographical knowledge as *general events* and *lifetime periods* respectively.

2. Constructing Memories from Knowledge

2.1 Generating Autobiographical Memories

When Proust's onrush of memories occurs he enters a state termed *retrieval mode* (Tulving 1983). In retrieval mode an intentional conscious attempt is made to recall memories, and a subset of attentional processes become directed toward long-term memory and to lifetime periods, general events, and ESK. Recalling memories is, however, a complicated process, and the evidence indicates that memories are constructed as a pattern of activation is set up across knowledge structures in long-term memory. Turning again to the example of the holiday memory, one can imagine that the memory was recalled while the rememberer was in conversation with a family member during which they were jointly recalling valued moments from their shared past. Thus, the cue is shared holidays, and this can be used to search lifetime periods and general events.

These knowledge structures contain information at different levels of specificity about activities, locations, others, motives, feelings, common to the period or general event and quite possibly several lifetime periods will be identified that contain general events of shared holidays. The lifetime period 'when the children were little' (a common lifetime period listed by older adults, Conway 1990) will contain knowledge that can be used to access the general event 'holiday in Italy.' Knowledge represented at this point can then be used to access related ESK, and once this occurs a whole pattern of activation is established across the different layers of autobiographical memory knowledge and a memory is formed. This process of generative retrieval is controlled by the *working self* and access to autobiographical knowledge is channeled through working self goal structures that determine the configuration of patterns of activation in long-term memory—which may or may not become full autobiographical memories.

2.2 Controlling Memory Construction

The influence of the working self on the process of memory generation is extensive and powerful and goals characteristic of certain personality and/or attachment styles selectively increase accessibility to goal-relevant autobiographical knowledge (e.g., Bakermans-Kranenburg and IJzendoorn 1993, McAdams 1993, McAdams et al. 1997, Mikulincer and Orbach 1995, Mikulincer 1998, Strauman 1996, Woike 1995). Thus, individuals with a personality type centered around notions of power predominantly recall and value memories of events in which they controlled others, achieved high status, performed acts of leadership, demonstrated independence, and so forth. In contrast, individuals with a personality type centered around notions of intimacy and dependence have preferential access to memories of interactions with important others, social events, moments of dependency and interdependency, etc. (see Conway and Pleydell-Pearce 2000). The working self can also attenuate access to goal-incongruent or self-threatening autobiographical knowledge. A frequently observed feature of clinical depression is an inability to form memories that contain ESK (Williams 1996). Instead the generation process terminates at the level of general events and, presumably, this disruption of retrieval has the protective effect of preventing the sufferer from recalling negative experiences that might otherwise serve to increase feelings of worthlessness and despair.

2.3 'Spontaneous' Recall of Memories

The construction of autobiographical memories is, then, extended in time and is a product of a complex interaction between working self goals and long-term memory. But this process of generative retrieval, so frequently observed in experimental studies of autobiographical memory, can be bypassed by a sufficiently specific cue. A cue which corresponds to a representation of ESK will automatically activate that representation, and activation spreading from the ESK will activate a general event which in turn will activate a lifetime period. The result is that a pattern of

activation is suddenly and spontaneously formed that, if linked to working memory goals, would immediately become a specific memory. Such spontaneous recall appears to be what Proust experienced and what many other writes have also described (see Salaman 1970 for a review). Indeed, Berntsen (1996) in a recent survey found that on average people report involuntary recalling two to three memories per day. The spontaneous recall of autobiographical memory most often occurs when we are in retrieval mode and actively attempting to recall a memory: It is then that other memories come to mind, often surprisingly, and occasionally of 'forgotten' experiences.

All these instances of spontaneous retrieval arise because a cue has activated representations of ESK. However, it is also the case that cues constantly activate autobiographical knowledge which is highly sensitive to both externally presented and internally generated cues (Conway and Pleydell-Pearce 2000). Such activations do not usually cause spontaneous remembering, because activation is restricted mainly to general events and lifetime periods and so does not reach a sufficient level of intensity to activate ESK. Nor, typically, do these activations engage the working self and because of this they occur outside conscious awareness. Nevertheless, the fact that direct retrieval does occur, and with some frequency, is good evidence for the nonconscious construction of autobiographical memories.

3. *Vivid Memories*

Memories can be effortfully generated or come to mind relatively effortlessly by a process of direct retrieval but, however a memory is constructed, at some point the activated knowledge becomes joined to the goal structures of the working self and this can facilitate or inhibit recall depending upon goal compatability. Some experiences are so self-engaging (positively or negatively) that they may be encoded in ways that make them especially available to construction and more resistant to forgetting than other memories. A good example of this type of memory are those termed *flashbulb memories* by Brown and Kulik (1977). In a formative study of memories for learning of various major news events that occurred in the 1960s, e.g., the assassinations of President John F. Kennedy (JFK), Martin Luther King, etc., a decade prior to their study, Brown and Kulik sampled memories from groups of white and black North Americans. They found that both groups had vivid and detailed memories of where they were and what they were doing when first learning the news of JFK's murder. In contrast, less than half the white group had detailed memories for learning of the killing of Martin Luther King compared to all of the black group.

The selective formation of flashbulb memories by groups to whom the news event is most self-engaging has now been reported for a range of different events, e.g., the attempted assassination of President Ronald Reagan (Pillemer 1984), the resignation of the British Prime Minister Margaret Thatcher (Conway et al. 1994), the 1989 Californian earthquake (Neisser et al. 1996), the death of the Belgium king (Finkenauer et al. 1997), and others. As Neisser (1982) points out, flashbulb memories are the points in our lives where we line up our personal history with public history and in effect say: 'I was there.'

4. *Memories and Motives*

There are, of course, many other ways in which memories and the self intersect with our social world. Pillemer (1998) documents several of these and shows how often they are related to important moments of change when new life goals were adopted, revised, or abandoned. One rich area for these types of memories is in education and it seems that quite a few people have vivid flashbulb-type memories for moments when a career path quite suddenly became apparent to them. To take an example from the many listed in Pillemer (1998), a student recalled how in an undergraduate class a professor, when challenged, was able to give the exact reference (act, scene, and line) for Iago's line 'Put up your bright swords for the dew will only rust them' from Shakespeare's *Othello*. This virtuoso act led the student into a sudden realization that she too would like to acquire such an exact and detailed body of knowledge and this in turn led her to apply to graduate school and to a career in research. These vivid self-defining autobiographical memories (Singer and Salovey 1993) often date to a period when people were 15 to 25 years of age, a period known as the *reminiscence bump*, because when memories are recalled across the lifespan there is an increase in memories retrieved from this period (Rubin 1982, Rubin et al. 1998). One explanation here is that this is a period in which a stable and durable self-system finally forms, when processes such as generation identity formation take place, and the last phase of separation from the family occurs. Experiences during this period maintain an enduring connection to fundamental life goals and hence their prolonged high accessibility in long-term memory.

5. *Memories and the Brain*

An important development since the 1990s has been the study of the neural substrate of autobiographical remembering. This is of importance not only from a theoretical point of view but also because autobiographical memory is often disrupted following brain injury, and is one of the main cognitive abilities that becomes impaired, severely and irreversibly, in dementing illnesses of old age. Indeed, one very general form of impairment following varying types of brain injury is loss of access to ESK, with some sparing of

more general autobiographical knowledge. Studies of people with brain injuries and current neuroimaging studies all indicate that autobiographical remembering is distributed over several different brain regions (see Conway and Fthenaki 2000, Markowitsch 1998). It appears that the constructive retrieval processes are mediated by neural networks in the frontal lobes, bilaterally, but more prominent in the left cortical hemisphere than in the right. As construction proceeds, and a memory is formed, activation shifts to the right cortical hemisphere to networks at the frontal-temporal junction (temporal poles) and temporal-occipital junction. Once a specific and detailed memory is 'in mind' activation can be detected in the right temporal and occipital lobes (and to a lesser extent in the left). The latter site of activation is important as it may reflect the activation of sensory-perceptual details of the recalled experience—the ESK that is such a hallmark of autobiographical remembering, access to which may be lost following neurological injury. Neurologically, then, the construction of autobiographical memories features activation of brain processing regions at the front of the brain in the neocortex and unfolds over time as a memory is constructed, to areas in the middle and toward the posterior of the brain.

6. *Summary*

Autobiographical remembering is a dynamic cognitive process leading to the transitory formation of specific memories. These memories are constructed from several different types of knowledge and have an intricate relation to self. Indeed, autobiographical memories are one of the key sources of identity and they provide a crucial psychological link from personal history of the self to selves embedded in society.

See also: Elaboration in Memory; Episodic and Autobiographical Memory: Psychological and Neural Aspects; Memory and Aging, Cognitive Psychology of; Memory Retrieval; Mood-dependent Memory; Personal Identity: Philosophical Aspects; Reconstructive Memory, Psychology of; Self-knowledge: Philosophical Aspects

Bibliography

Bakermans-Kranenburg M J, IJzendoorn M H 1993 A psychometric study of the adult attachment interview: Reliability and discriminant validity. *Development Psychology* **29**(5): 870–9

Berntsen D 1996 Involuntary autobiographical memories. *Applied Cognitive Psychology* **10**(5): 435–54

Brown R, Kulik J 1977 Flashbulb memories. *Cognition* **5**: 73–99

Conway M A 1990 *Autobiographical Memory: An Introduction*. Open University Press, Buckingham, UK

Conway M A, Anderson S J, Larsen S F, Donnelly C M, McDaniel M A, McClelland A G R, Rawles R E, Logie R H 1994 The formation of flashbulb memories. *Memory and Cognition* **22**: 326–43

Conway M A, Fthenaki A 2000 Disruption and loss of autobiographical memory. In: Cermak L (ed.) *Handbook of Neuropsychology: Memory and Its Disorders*, 2nd edn. Elsevier, Amsterdam, pp. 257–88

Conway M A, Pleydell-Pearce C W 2000 The construction of autobiographical memories in the self memory system. *Psychological Review* **107**: 261–88

Finkenauer C, Luminet O, Gisle L, El-Ahmadi A, van der Linden M, Philippot P 1998 Flashbulb memories and the underlying mechanisms of their formation: Towards an emotional-integrative model. *Memory and Cognition* **26**(3): 516–31

Gardiner J M, Richardson-Klavehn A 2000 Remembering and knowing. In: Tulving E, Craik F I M (eds.) *Handbook of Memory*. Oxford University Press, Oxford, UK, pp. 229–44

Markowitsch H J 1998 Cognitive neuroscience of memory. *Neurocase* **4**: 429–35

McAdams D P 1993 *Stories We Live By: Personal Myths and the Making of the Self*. Morrow, New York

McAdams D P, Diamond A, de Aubin E, Mansfield E 1997 Stories of commitment: The psychosocial construction of generative lives. *Journal of Personality and Social Psychology* **72**(3): 678–94

Mikulincer M 1998 Adult attachment style and individual differences in functional versus dysfunctional experiences of anger. *Journal of Personality and Social Psychology* **74**(2): 513–24

Mikulincer M, Orbach I 1995 Attachment styles and repressive defensiveness: The accessibility and architecture of affective memories. *Journal of Personality and Social Psychology* **68**: 917–25

Neisser U 1982 Snapshots or benchmarks? In: Neisser U (ed.) *Memory Observed: Remembering in Natural Contexts*. W.H. Freeman, San Francisco, CA, pp. 43–8

Neisser U, Winograd E, Bergman E T, Schreiber C A, Palmer S E, Weldon M S 1996 Remembering the earthquake: Direct experience vs. hearing the news. *Memory* **4**: 337–57

Pillemer D B 1984 Flashbulb memories of the assassination attempt on President Reagan. *Cognition* **16**: 63–80

Pillemer D B 1998 *Momentous Event, Vivid Memories*. Harvard University Press, Cambridge, MA

Proust M [1913] 1981 *Rememberance of Things Past*. Random House, New York

Rubin D C 1982 On the retention function for autobiographical memory. *Journal of Verbal Learning and Verbal Behavior* **21**: 21–38

Rubin D C, Rahhal T A, Poon L W 1998 Things learned in early adulthood are remembered best. *Memory and Cognition* **26**: 3–19

Salaman E 1970 *A Collection of Moments: A Study of Involuntary Memories*. Longman, Harlow, UK

Singer J A, Salovey P 1993 *The Remembered Self*. Free Press, New York

Strauman T J 1996 Stability within the self: A longitudinal study of the structural implications of self-discrepancy theory. *Journal of Personality and Social Psychology* **71**(6): 1142–53

Tulving E 1983 *Elements of Episodic Memory*. Clarendon Press, Oxford, UK

Tulving E 1985 Memory and consciousness. *Canadian Psychologist* **26**: 1–12

Wheeler M A, Stuss D T, Tulving E 1997 Towards a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin* **121**: 331–54

Williams J M G 1996 Depression and the specificity of autobiographical memory. In: Rubin D C (ed.) *Remembering Our Past: Studies in Autobiographical Memory*. Cambridge University Press, Cambridge, UK, pp. 244–67

Woike B A 1995 Most-memorable experiences: Evidence for a link between implicit and explicit motives and social cognitive processes in everyday life. *Journal of Personality and Social Psychology* **68**(6): 1081–91

M. A. Conway

Memory, Consolidation of

1. Historical Bases

At the start of the twentieth century, Müller and Pilzecker (1900) proposed that memory formation required a period of stabilization. They suggested that the neurobiological formation of memory proceeded in two steps: initially memories perseverated in a labile form and then gradually consolidated into a stable and permanent form. The idea that memories were initially labile was used to explain retroactive interference, the forgetting of recent information induced by later learning of other information.

These ideas were the precursors of the more formal neurobiological view of memory formation proposed by Donald Hebb (1949). Hebb proposed a dual-trace hypothesis of memory consolidation. According to this hypothesis, a learning experience initiates reverberating activity in neural circuits which serves both as a holding mechanism for short-term memory and as a process integral to changing synaptic connections to establish new circuits for long-term memory. This hypothesis therefore used two memory traces, working in a serial manner, to explain how a memory evident immediately after an experience is transitional to a permanent memory trace. The reverberating circuits, dependent on sustained neurophysiological activity, were labile, resulting in memory that was susceptible to perturbation until relatively permanent and stable circuit modifications were established.

The dual-trace hypothesis was supported by clinical findings that head trauma produced retrograde amnesia, the loss of recent memories prior to injury, and anterograde amnesia, the loss of the ability to form new memories. According to a dual-trace hypothesis, retrograde amnesia is a time-dependent phenomenon resulting from the disruption of the continuing formation of memories not yet consolidated into long-term memory, while anterograde amnesia may be either transient or permanent resulting from disruption of brain mechanisms or systems responsible for initiating the processes that would lead to long-term memory.

The dual-trace model also drew support from clinical findings emerging during the preceding decade from the use of electroconvulsive therapy (ECT) to treat psychiatric disorders. Reports from clinical studies indicating that patients experienced retrograde amnesia for recent events prior to treatment were followed by extensive studies in laboratory animals testing the effects on memory of electroconvulsive shock and other treatments that alter brain functions. Many treatments, generally administered soon after training, induce retrograde amnesia. Retrograde amnesia has been noted in a wide range of species, including mammals and other vertebrates, as well as insects and molluscs, suggesting that memory consolidation is a process conserved in evolution.

2. Temporal Gradients of Retrograde Amnesia

The defining characteristic of retrograde amnesia is that the efficacy of a given treatment decreases with time after learning. For many years, considerable effort attempted to identify the temporal gradient of retrograde amnesia. The importance of this effort is that the temporal properties might reveal the time course of memory consolidation, a time constant of great importance for identifying likely biological candidates for the mechanisms of memory formation. However, under some conditions the temporal gradient was very short, one second or less, and under other conditions the gradient was very long. Thus, the time-course proved to be quite variable and the gradient was eventually found to depend on many factors. Some, such as task, species, motivation, time of day, might be considered to be intrinsic to memory formation itself. Others such as severity of head trauma in humans, or intensity or dose of a treatment in laboratory animals, are not properties of memory *per se*. Therefore, it became evident that it is unlikely that a single temporal gradient reveals the time required for memory formation. Rather, the findings indicate that memory gradually becomes less susceptible to modification with time after an experience (Gold and McGaugh 1975).

The time-course for loss of susceptibility spanning several hours, and the multiple temporal gradients observed with different treatments, suggests that memory consolidation may reflect the accrual of biological consequences of an experience, including gene expression and protein synthesis. The role of protein synthesis in consolidation of memory is further supported by extensive findings showing that drugs that interfere with protein synthesis produce retrograde amnesia. However, interpretation of this evidence is complicated by findings that many classes of drugs, including drugs acting to augment or to impair

neurotransmitter functions, block the effects of antibiotics on memory without reversing the inhibition of protein synthesis.

Recently, the term-memory consolidation has been used to define a phenomenon with somewhat different properties. Experiments indicate that ECT treatments administered to humans can induce retrograde amnesia for events occurring years prior to the treatments. Also, under some circumstances, surgical removal of the hippocampus produces retrograde amnesia for events that preceded surgery by weeks in rodents and months in nonhuman primates. In one sense, these experiments are analogous to those of other studies of memory consolidation, demonstrating time-dependent retrograde amnesia. However, there are also significant differences. First, the time-course of the amnesia extends well beyond the neurobiological processes likely to underlie the formation of memory. Second, most experiments involve treatments, for example, removal of the hippocampus, that result in a permanently altered brain. Therefore, these long temporal gradients seem more likely to reflect continuing reorganization of the neural substrates of memory storage, perhaps modifying the continual integration of old and new memories.

3. *Posttraining Design*

Beyond the specific information offered about memory formation, research on memory consolidation has also provided an experimental design that has been extremely important in examining the effects of drugs and other treatments on memory. When drugs are administered chronically or soon before training, any effects observed on learning or on later memory might be attributable to actions on sensory functions, motor functions, or motivational factors at the time of learning. Distinguishing effects such as these from primary effects on memory is very difficult. In contrast, studies of memory consolidation, characterized by retroactive effects of treatments on memory, avoid all of these potential confounds. In these experiments, the treatment is administered after training. Also, memory is typically assessed one day or more later, at a time when the acute effects of the treatment have likely dissipated. Therefore, the experiment subjects are unaffected by the treatment at the times of training and testing. In addition, demonstration of retrograde amnesia gradients offers further support for interpretations based on effects memory. As the delay between training and treatment increases, the efficacy of the treatment diminishes. Such findings strongly support the view that any effects observed at the time of testing reflects actions on memory formation and not an extended proactive effect of the treatment at the time of testing. Because of the clarity of interpretation afforded by posttraining treatment experiments, studies of memory consolidation are used extensively to distinguish effects of treatments on memory from potential influences on performance.

4. *Retrograde Enhancement of Memory*

In addition to studies demonstrating that memory consolidation can be impaired by many treatments, there is also extensive evidence that memory can be enhanced by treatments administered after training. With findings spanning many species of research animals and many learning tasks, memory for recent information can be enhanced by stimulant drugs and by low-level stimulation of specific brain regions such as the mesencephalic reticular formation and the amygdala. Importantly, the enhancement of memory seen with these treatments, like those of amnestic treatments, is greatest when the treatment is administered soon after training and loses efficacy with time after training. Thus, memory consolidation can be both enhanced or impaired by posttraining treatments.

5. *Endogenous Modulation of Memory Consolidation*

Depending on the drug dose used and other experimental conditions, many treatments can either enhance and impair memory. The dose-response curve for most treatments follows an inverted-U function, in which low doses are ineffective, intermediate doses enhance memory, and high doses impair memory. Moreover, the peak of the dose-response curve interacts with such factors as arousal or stress induced by the training conditions. Generally, higher training-related arousal results in lower doses to enhance or to impair memory. Thus, physiological responses to training apparently summate with physiological responses to the treatments in modifying memory formation.

5.1 *Hormonal Regulation of Memory Formation*

These findings support the view that endogenous physiological responses to training, such as neuroendocrine events, may modulate memory formation. Considerable evidence indicates that some hormones, when administered soon after training, modulate later memory. These hormones include epinephrine, norepinephrine, vasopressin, adrenocorticotropin, and glucocorticoids.

The susceptibility of memory consolidation to modulation by hormones appears to reflect an ability of hormonal consequences of an experience to regulate the strength of memory formed for that experience. According to this view, memory consolidation enables

the significance of events, reflected in hormonal responses, to control how well the events will be remembered. In this context, the conservation of the phenomenon of memory consolidation across evolution may represent the establishment and maintenance of a mechanism for selecting important memories as those that should be retained (Gold and McGaugh 1975).

5.2 Epinephrine Regulation of Memory Formation

The adrenal medullary hormone epinephrine is one of the best-studied examples of hormonal modulation of learning and memory. Posttraining injections of epinephrine enhance memory for a wide range of tasks in experiments with rats and mice. The enhancement follows an inverted-U dose-response curve, in which high doses can impair memory. Moreover, the effective doses, and interactions with training-related arousal, match well against measurements of circulating epinephrine levels after training. In humans, propranolol, a β-adrenergic antagonist that blocks epinephrine receptors, blocks memory for story elements containing high arousal. Propranolol also blocks enhancement of memory for word lists by experimentally induced arousal.

Because epinephrine does not enter the central nervous system in appreciable extent, it is likely that the hormone enhances memory consolidation by peripheral actions. Two major candidates, which are not mutually exclusive, for peripheral mechanisms mediating epinephrine effects on memory are actions on adrenergic receptors on the ascending vagus nerve, with terminations in the brain in the region of the nucleus of the solitary tract, and classic physiological actions on hepatic receptors to increase blood levels of glucose, a substance that has ready access to the brain. Electrical stimulation of vagal afferents to the central nervous system enhances memory in both rats and humans. Also, inactivation of the nucleus of the solitary tract blocks memory-enhancement induced by posttraining administration of epinephrine.

5.3 Glucose Regulation of Memory Formation

A second mechanism by which epinephrine may act to enhance memory is by liberating glucose from hepatic stores, thereby raising blood glucose levels. Because glucose enters the brain readily via a facilitated transport mechanism, glucose then may be an intermediate step between peripherally acting epinephrine and actions on the brain to modulate memory (Gold 2001).

Administration of glucose enhances memory in rats, mice, and humans. The effects of glucose on memory in these species are characterized by an inverted-U dose-response curve. In humans, glucose enhances memory consolidation in a posttraining design, one rarely used in studies of pharmacology of memory in humans. Also, glucose enhances memory in a wide range of humans, including healthy young and elderly adults, as well as individuals with Down syndrome and Alzheimer's disease. The parallels in humans and laboratory animals between the effects on memory of glucose, as well as of epinephrine and other treatments noted here, are striking and suggest that transitions from studies of memory consolidation in nonhuman animals may contribute significantly to the development of pharmacological treatments to ameliorate cognitive deficits in a range of human conditions (Korol and Gold 1998).

In rats, glucose enhances memory for some tasks when infused in small volume into specific brain areas, such as the medial septum, hippocampus, and amygdala. Although traditional thought was that the brain had a surplus supply of glucose, except under conditions of severe food deprivation, recent evidence indicates that cognitive activity can deplete extracellular glucose levels in the hippocampus during spatial learning. Administration of glucose at doses that enhance memory blocks that depletion. Thus, enhancement of memory may reflect actions to reverse a natural limitation on the glucose reserves of some brain areas. While the proximal mechanism by which glucose enhances memory is not clear, in many instances the effects of glucose on memory appear to be associated with augmented release of acetylcholine in the hippocampus, an action that may be indirect.

5.4 Amygdala Integration of Modulatory Influences on Memory

Other neurotransmitters also appear to have important roles in modulating memory processes. Considerable evidence indicates that norepinephrine, particularly in the basolateral nucleus of the amygdala (BLA), is important in mediating the effects of many treatments on memory. For example, infusions of β-adrenergic agonists into the BLA after training enhance memory. Conversely, lesions of the BLA or infusion of β-adrenergic antagonists into the BLA block both enhancement and impairment of memory consolidation by hormones and other treatments. Also, in contrast to findings with healthy human subjects, emotional arousal does not enhance memory in individuals with lesions of the amygdala. Together, these findings support the view that the BLA may play a key role in memory formation by integrating the effects on memory of a broad range of treatments (McGaugh 2000, Cahill and McGaugh 1996).

6. Neural Systems and Memory Consolidation

There is considerable evidence that different neural systems mediate processing of different forms of memory. When administered systemically, many

drugs enhance or impair memory for many kinds of tasks. However, when drugs are administered directly into specific brain regions, they may modulate memory for a more restricted set of tasks. For example, amphetamine infused into the hippocampus enhances memory for spatial learning in a swim task but not for learning to find a platform identified with a salient proximal cue. Amphetamine infused into the striatum enhances the cued but not the spatial version of the task. Amphetamine enhanced memory for both tasks when infused into the amygdala after training, consistent with the view that the amygdala integrates modulation of many types of memory.

Many other tasks involve contributions to learning of more than one neural system at a time. Interactions between different memory systems can determine the manner in which a rat solves a maze, perhaps learning on the basis of extra maze cues (e.g., turn toward the light) or perhaps learning on the basis of making a particular response (e.g., turn right). Recent studies have begun to integrate aspects of memory consolidation with neural system approaches to learning and memory. When viewed in this context, modulation of memory may have an important role in biasing what is learned by boosting or blunting memory processing in different neural systems.

7. *Cell and Molecular Biology Bases of Memory Formation*

Studies of memory consolidation provide a firm basis for examination of memory storage processing that continues for a considerable time after an experience (McGaugh 2000). With this information, several approaches have been taken to identify the nature of the brain systems and mechanisms involved in these processes. Studies of functional brain activation in humans show changes over a period of several hours in the regions activated by learning, perhaps indicating that memory consolidation involves reorganization of brain representations of memory during the time after an experience. The cellular and molecular events that are initiated by an experience, continue for considerable time after that experience, and are important to memory formation for that experience, are currently a topic of much investigation. By evident analogy with short- and long-term memory, cellular and molecular biology contributions are often sorted into those that affect early- or late (protein synthesis-dependent)-phase memory. The molecular bases are derived also from an analogy, or mechanism depending on one's view, between long-term potentiation and memory. The early stage of memory consolidation appears to involve CaMKII (calcium-calmodulin-dependent protein kinase II). Inhibitors of CaMKII injected into either the hippocampus or amygdala shortly after training produce retrograde amnesia. When injected into the hippocampus, inhibitors of PKC (protein kinase C) or PKA (protein kinase A) produce retrograde amnesia even if administered hours after training. Moreover, PKA activity and CREB (cAMP response element-binding protein) immunoreactivity increase in the hippocampus up to hours after training, suggesting that the late phases of memory consolidation may involve cAMP-mediated activation, by PKA phosphorylation, of the CREB transcription factor.

There is evidence that early- and late-phase memory consolidation can be impaired independently by some treatments. However, there are also conditions in which early- and late-phases are both impaired by a single treatment. These issues return to some raised by Hebb's initial dual process hypothesis of memory formation. At the levels of both neural systems and cellular and molecular neurobiology, it now seems quite likely that research findings will eventually reveal some neural systems and biochemical processes that act in series and others that act in parallel.

See also: Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Memory in the Bee; Memory Retrieval; Memory: Synaptic Mechanisms

Bibliography

Cahill L, McGaugh J L 1996 Modulation of memory storage. *Current Opinion in Neurobiology* **6**: 237–42

Gold P E 2001 Drug enhancement of memory in aged rodents and humans. In: Carroll M E, Overmeir J B (eds.) *Animal Research and Human Health: Advancing Human Welfare Through Behavioural Science*. American Psychological Association, Washington, DC, pp. 293–304

Gold P E, McGaugh J L 1975 A single trace, two process view of memory storage processes. In: Deutsch D, Deutsch J A (eds.) *Short Term Memory*. Academic Press, New York, pp. 355–90

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Korol D L, Gold P E 1998 Glucose, memory and aging. *The American Journal of Clinical Nutrition* **67**: 764S–71S

McGaugh J L 2000 Memory—a century of consolidation. *Science* **287**: 248–51

Müller G E, Pilzecker A 1900 Experimentelle Beiträge zur Lehre vom Gedächtnis. *Zeitschrift fur Psychologie* **1**: 1–288

P. E. Gold and J. L. McGaugh

Memory: Collaborative

Although 'collaborative memory' may be briefly defined as remembering activities involving multiple persons, a more precise delineation of the phenomenon is the major goal of this article. Accordingly,

several key premises of the phenomenon of collaborative memory should be noted at the outset. These include the following: (a) collaborative memory refers to a form or characteristic of remembering activity rather than to a theoretical system or function of memory; (b) scientific analysis of collaborative memory performance is based in part on principles derived from, and complementary to, theories pertaining to individual-level memory phenomena; (c) as collaborative memory performance is often influenced by both cognitive and interactive processes, research methods and interpretations from neighboring disciplines may be consulted or incorporated; (d) contemporary collaborative memory research and theory is neutral (i.e., empirically open) to questions regarding the benefits and costs of multiperson remembering, and may exist independent of specific effects; and (e) that numerous everyday remembering activities occur in the context of other cooperating individuals provides a necessary but not sufficient rationale for investigating collaborative memory.

1. *Phases and Forms of Collaborative Memory*

Collaboration in remembering may be differentiated by (a) whether it occurs during the encoding or retrieval phase of individual-level memory, or (b) the form of memory (i.e., the system or task) in which it occurs (or to which it is addressed). Although no definitive frequency data are available, collaborative memory activities occur perhaps most commonly during a retrieval phase. Thus, n 1 individuals attempt to recall information to which they have been previously exposed and for which no single individual possesses a perfect representation. Collaborative memory during this phase is often directed at remembering a commonly experienced personal event (i.e., collaborative episodic memory) or an item of knowledge or information to which the individuals would likely have been exposed (i.e., collaborative semantic memory). In addition, collaborating in recalling an intention or to-be-performed action (i.e., collaborative prospective memory) may be an example. Family or friendship groups attempting collectively to reconstruct stories from their shared past provide numerous fascinating (and entertaining) examples of collaborative episodic remembering. Laboratory illustrations include collaborative remembering of verbal or nonverbal information. Partners attempting to cue one another in recalling historical knowledge, such as the major Kings of Sweden or the combatants and outcome of the Peloponnesian War, are collaboratively performing a semantic memory task.

In addition to retrieval, remembering activities in the context of other individuals may occasionally concentrate on the encoding phase, wherein individuals collectively work to record incoming information in 'multisite' temporary storage (i.e., collaborative working memory). To be sure, in an inescapable sense, the 'memory' that is being encoded 'resides' in one brain, in personal or individual-level storage. Thus, collaborative encoding involves incidental or cooperative activities that may enhance the probability that a to-be-remembered (TBR) item is recorded, and thereby accessible (i.e., retrievable) at a later date by one or more individuals. For example, partners may develop encoding mnemonics or divide the task in order to enhance the storage of rapidly incoming information.

Collaboration may occur if the to-be-remembered information is not immediately or comprehensively available to any single participating individual. That is, a deficit is present—i.e., the answer is known neither immediately nor definitively by one individual—there is no opportunity (or no need) for collaboration. In collaborative episodic or semantic memory, an assumption is that the individuals may offer both common and unique memories. In addition, in the process of collaboration some mutual cuing may occur, such that novel items are produced. In collaborative prospective memory, the deficit may be an anticipated one. An individual may be concerned about forgetting to perform a future action (e.g., birthday, appointment, message) and enlist a spouse to help them remember at an appropriate time. Although both retrieval and encoding phases may occur in the (presumably assistive) context of participating individuals, both remain fundamentally individual-level functions. In this sense, collaboration qualifies the process and products of remembering, but does not necessarily carry emergent properties.

2. *Scope and Selected Conceptual Issue*

For several decades, researchers in a surprising variety of fields have addressed aspects of everyday memory activity that appear to operate in the influential context of other individuals. Although some cross-field differences in terminology and assumptions exist, several disputed conceptual issues are common to them all. One still-unresolved issue concerns the extent to which collaboration is optimally or acceptably effective.

2.1 *Scope of the Phenomenon*

The phenomenon has been also called collective (e.g., Middleton and Edwards 1990), situated (e.g., Greeno 1998), group (e.g., Clark and Stephenson 1989), socially shared (e.g., Resnick et al. 1991), interactive (e.g., Baltes and Staudinger 1996), transactive (Wegner 1987), or collaborative (e.g., Dixon 1996) memory. The fields in which this phenomenon has historically been of interest include educational psychology, cognitive science, social-cognitive psychology, industrial-organizational psychology, child

developmental psychology, and adult developmental psychology (recent reviews include Baltes and Staudinger 1996, Engeström and Middleton 1998, Kirshner and Whitson 1997, Lave and Wenger 1991).

2.2 Continuing Theoretical and Research Issues

Common to these literatures is the basic fact that two or more individuals attend to the same set of learning or memory tasks and are working cooperatively (although not necessarily effectively) to achieve a recall-related goal. Notably, the members of the collaborating group can be variously passive listeners, conversational interactants, productive collaborators, seasoned tutors, counterproductive or even disruptive influences, or optimally effective partners. Therefore, according to the neutral definition of collaborative memory espoused in this article, no *a priori* assumptions are made about the effectiveness or logical priority of the memory-related interaction. It has long been clear that group processes can vary in their effectiveness and thus group products can vary in their accuracy and completeness (e.g., Steiner 1972).

The issue of the extent to which collaborative memory is effective has been evaluated from numerous perspectives for several decades. Indeed, much research has focused on this contentious issue (e.g., Dixon 1999, Hill 1982), with several key factors appearing to play a role in the observations and inferences. These factors include: (a) whether the participants are collaborative-interactive experts (e.g., friends or couples), (b) the type of outcome measure observed (i.e, a simple product such as total items recalled or a variety of recall-related products such as elaborations and inferences), (c) the extent to which the actual processes (e.g., strategic negotiations) and byproducts (e.g., affect and sharing) of the collaborative communication are investigated, and (d) the comparison or baseline by which the effectiveness of collaborative performance is evaluated. In general, little extra benefit is observed under conditions in which researchers reduce the dimensionality of the tasks, the familiarity of the interactants, the variety of the memory-related products measured, and the richness of the collaborative communication (e.g., Meudell et al. 1992). In contrast, evidence for notable collaborative benefit may be observed when researchers attend to collaborative expertise, multidimensional outcomes, measurement of actual collaborative processes, and comparisons accommodated to memory-impaired or vulnerable groups (e.g., Dixon and Gould 1998).

See also: Episodic and Autobiographical Memory: Psychological and Neural Aspects; Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Productivity, Social Psychology of; Prospective Memory, Psychology of; Semantic Knowledge: Neural Basis of

Bibliography

Baltes P B, Staudinger U M (eds.) 1996 *Interactive Minds: Lifespan Perspectives on the Social Foundation of Cognition*. Cambridge University Press, Cambridge, UK

Clark N K, Stephenson G M 1989 Group remembering. In: Paulus P B (ed.) *Psychology of Group Influence*, 2nd edn. L. Erlbaum, Hillsdale, NJ

Dixon R A 1996 Collaborative memory and aging. In: Herrman D, McEvoy C, Hertzog C, Hertel P, Johnson M K (eds.) *Basic and Applied Memory Research*. Lawrence Erlbaum Associates, Mahwah, NJ

Dixon R A 1999 Exploring cognition in interactive situations: The aging of N + 1 minds. In: Hess T M, Blanchard-Fields F (eds.) *Social Cognition and Aging*. Academic Press, San Diego, CA

Dixon R A, Gould O N 1998 Younger and older adults collaborating on retelling everyday stories. *Applied Developmental Science* **2**: 160–71

Engeström Y 1992 Interactive expertise: Studies in distributed intelligence. *Research Bulletin* **83**. Department of Education, University of Helsinki, Finland

Engeström Y, Middleton D (eds.) 1998 *Cognition and Communication at Work*. Cambridge University Press, Cambridge, UK

Greeno J G 1998 The situativity of knowing, learning, and research. *American Psychologist* **53**: 5–26

Hill G W 1982 Group versus individual performance: Are N + 1 heads better than 1? *Psychological Bulletin* **91**: 517–39

Kirshner D, Whitson J A (eds.) 1997 *Situated Cognition*. L. Erlbaum, Mahwah, NJ

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK

Meudell P R, Hitch G J, Kirby P 1992 Are two heads better than one? Experimental investigations of the social facilitation of memory. *Applied Cognitive Psychology* **6**: 525–43

Middleton D, Edwards D (eds.) 1990 *Collective Remembering*. Sage, Newbury Park, CA

Resnick L B, Levine J M, Teasley (eds.) 1991 *Perspectives on Socially Shared Cognition*, 1st edn. American Psychological Association, Washington, DC

Steiner I D 1972 *Group Process and Productivity*. Academic Press, New York

Wegner D M 1987 Transactive memory: A contemporary analysis of the group mind. In: Mullin B, Goethals G R (eds.) *Theories of Group Behavior*. Springer-Verlag, New York

R. A. Dixon

Memory Development in Children

Although scientific research on memory development has about the same long tradition as the scientific study of psychology (i.e., about 1880), the majority of studies have been conducted since the 1970s, stimulated by a shift away from behaviorist theories toward information-processing theories (Schneider and Pressley 1997). Given that hundreds and thousands of empirical studies have investigated the development of children's ability to store and retrieve information from memory, only a rough overview of the most

important research trends can be presented in this article (for recent reviews see Cowan 1997, Schneider and Bjorklund 1998, Schneider and Pressley 1989).

1. Memory Development in Very Young Children

From birth on, infants can remember things (e.g., faces, pictures, objects) for rather long periods of time. Basic memory activities such as the association of a stimulus with a response and the distinction between old and new stimuli (i.e., recognition processes) are especially dominant early in life. Moreover, young infants can also remember activities that they had performed at an earlier point in time. Research focusing on older infants (between 10 and 20 months of age) has used deferred (delayed) imitation techniques to measure memory boundaries. In several of these experiments, infants watched as an experimenter demonstrated some novel, unusual behavior with an unfamiliar toy. Infants later imitated such strange behaviors, indicating that these events had been stored in long-term memory. Although levels of imitations are typically greater for older than for younger infants, the findings indicate that the neurological systems underlying long-term recall are present at the beginning of the second year of life (Meltzoff 1995).

Interestingly, age does not seem to be the primary determinant of whether or for how long memory will be remembered once the capacity is in place. Rather, the literature shows that the organization of events and the availability of cues or reminders determine young children's long-term memory. Young children tend to organize events in terms of 'scripts' that are a form of schematic organization with real-world events structured in terms of their causal and temporal characteristics. Scripts develop most routinely for common, repeated events. Children learn what 'usually happens' in a situation, for instance, a birthday party or a visit to a restaurant. Memory for routine events makes it possible for infants and toddlers to anticipate events and to take part in, and possibly take control of, these events (Nelson 1996). Repeated experience clearly facilitates long-term recall in preverbal and early-verbal children.

Moreover, cues or reminders also result in better recall for very young children. Longitudinal research on young children's long-term memory for events experienced either 6 or 18 months before showed that younger children (who were about 3 years old when the event happened) needed more prompts (retrieval cues) to reconstruct their memories than children who were about a year older at the time of the event. Although most children showed low levels of free recall, they could remember much more when specific cues were presented (for a review of this research, see Fivush 1997).

What are important mechanisms of young children's (verbal) memory development? Recent research indicates that parents play an important role. Interchanges between parents and their young children seem highly relevant for children's recall proficiency. Children learn to remember by interacting with their parents, jointly carrying out activities that are later performed by the child alone. Through these conversations, children learn to notice the important details of their experiences and to store the to-be-recalled information in an organized way.

2. (Verbal) Memory Development between 5 and 15

The vast majority of studies on memory development has been carried out with older children, mainly dealing with explicit, that is, conscious remembering of facts and events. This skill was also labeled declarative memory and distinguished from procedural memory, which refers to unconscious memory for skills. It was found repeatedly that particularly clear improvements in declarative memory can be observed for the age range between 6 and 12 years, which roughly corresponds to the elementary school period in most countries (Schneider and Pressley 1997). In order to explain these rapid increases over time, different sources of memory development have been identified. According to most researchers, changes in basic capacities, memory strategies, metacognitive knowledge, and domain knowledge all contribute to developmental changes in memory performance. There is also broad agreement that some of these sources of development contribute more than others, and that some play an important role in certain periods of childhood but not in others.

2.1 The Role of Basic Capacities

One of the most controversial issues about children's information processing is whether the amount of information they can actively process at one time changes with age. The concept of memory capacity usually refers to the amount of information that can be held in the short-term store (STS) and has been typically assessed via memory span tasks or measures of working memory. Whereas the former tasks require children to immediately recall a given number of items in the correct order, the latter are more complex in that they not only require the exact reproduction of items but are embedded in an additional task in which children must transform information held in the STS. The maximum number of items a person can correctly recall in those tasks define their memory span. In general, children's performance on working-memory tasks shows the same age-related increase as their performance on memory-span tasks, although the absolute performance level is somewhat reduced in working-memory tasks. Numerous studies have

shown that this development is related to significant increases in information processing speed which are most obvious in early ages, with the rate of changes slowing thereafter (Kail 1991).

Although there is little doubt that performance on memory span tasks improves with age, the implications for whether working-memory capacity changes with age are not so obvious. It still remains unknown whether the total capacity store factually increases with age or whether changes in information processing speed, strategies and knowledge allow more material to be stored within the same overall capacity.

2.2 *Effects of Memory Strategies*

Memory strategies have been defined as mental or behavioral activities that achieve cognitive purposes and are effort-consuming, potentially conscious and controllable (Flavell et al. 1993). Since the early 1970s numerous studies have investigated the role of strategies in memory development. Strategies can be executed either at the time of learning (encoding) or later on when information is accessed in long-term memory (retrieval). The encoding strategies explored in the majority of studies include rehearsal, which involves the repetition of target information, organization, which involves the combination of different items in categories, and elaboration, which involves the association of two or more items through the generation of relations connecting these items. Retrieval strategies refer to strategic efforts at the time of testing, when the task is to access stored information and bring it back into consciousness.

Typically, these strategies are not observed in children younger than 5 or 6. The lack of strategic behaviors in very young children was labeled 'mediational deficiency,' indicating that children of a particular (preschool) age do not benefit from strategies, even after having been instructed how to use them. The term 'production deficiency' refers to the fact that slightly older children do not spontaneously use memory strategies but can benefit substantially from strategies when told how to use them. More recently, the construct of a 'utilization deficiency' has been proposed to account for the opposite phenomenon, that is, the finding that strategies initially often fail to improve young children's memory performance (Flavell et al. 1993, Schneider and Bjorklund 1998). The explanation for this discrepancy favored by most researchers is that executing new strategies may consume too much of young children's memory capacity.

Although strategies develop most rapidly over the elementary school years, recent research has shown that the ages of strategy acquisition are relative, and variable within and between strategies. Even preschoolers and kindergarten children are able to use intentional strategies, both in ecologically valid settings such as hide-and-seek tasks, and in the context of a laboratory task. Although the majority of (cross-sectional) studies suggest that strategy development is continuous over the school years, recent longitudinal research has shown that children typically acquire memory strategies very rapidly (Schneider and Bjorklund 1998). Moreover, there is increasing evidence of substantial inter- and intrasubject variability in strategy use, with children using different strategies and combinations of strategies on any given memory problem (Siegler 1996).

Taken together, age-related improvements in the frequency of use and quality of children's strategies play a large role in memory development between the preschool years and adolescence. However, there is now an increasing realization that the use of encoding and retrieval strategies depends largely on children's strategic as well as nonstrategic knowledge. There is broad consensus that the narrow focus on developmental changes in strategy use should be replaced by an approach that takes into account the effects of various forms of knowledge on strategy execution.

3. *The Role of Metacognitive Knowledge (Metamemory)*

One knowledge component that has been explored systematically since the early 1970s concerns children's knowledge about memory. The term metamemory was introduced to refer to a person's potentially verbalizable knowledge about memory storage and retrieval (Flavell et al. 1993). Two broad categories of metacognitive knowledge have been distinguished in the literature. Declarative metacognitive knowledge refers to what children factually know about their memory. This type of knowledge is explicit and verbalizable and includes knowledge about the importance of person variables (e.g., age or IQ), task characteristics such as task difficulty, or strategies for resulting memory performances. In contrast, procedural metacognitive knowledge is mostly implicit (subconscious) and relates to children's self-monitoring and self-regulation activities while dealing with a memory problem.

Empirical research exploring the development of declarative metamemory revealed that children's knowledge of facts about memory increases considerably over the primary-grade years, but is still incomplete by the end of childhood. Recent studies also showed that increases in knowledge about strategies are paralleled by the acquisition of strategies, and that metamemory-memory behavior relationships tend to be moderately strong (Schneider and Pressley 1997). Thus, what children know about their memory obviously influences how they try to remember. Nonetheless, although late-grade-school children know

much about strategies, there is increasing evidence that many adolescents (including college students) have little or no knowledge of some important and powerful memory strategies.

The situation regarding developmental trends in procedural metamemory is not entirely clear. Several studies explored how children use their knowledge to monitor their own memory status and regulate their memory activities. There is evidence that older children are better able to predict future performance on memory tasks than younger children, and that there are similar age trends when the task is to judge performance accuracy after the fact. Also, older children seem better able to judge whether the name of an object that they currently cannot recall would be recognized later if the experimenter provided it (feeling-of-knowing judgments). However, although monitoring skills seem to improve continuously across childhood and adolescence, it is important to note that even young children can be fairly accurate in such metamemory tasks, and that developmental trends in self-monitoring are less pronounced than those observed for declarative metamemory. It appears that the major developmental improvements in procedural metamemory observable in elementary school children are due mainly to an increasingly better interplay between monitoring and self-regulatory activities. That is, even though young children may be similarly capable of identifiying memory problems than older ones, in most cases only the older children will effectively regulate their behavior in order to overcome these problems.

4. *The Impact of Domain Knowledge*

Striking effects of domain knowledge on performance in memory tasks has been provided in numerous developmental studies. In most domains, older children know more than younger ones, and differences in knowledge are linked closely to performance differences. How can we explain this phenomenon? First, one effect that rich domain knowledge has on memory is to increase the speed of processing for domain-specific information. Second, rich domain knowledge enables more competent strategy use. Finally, rich domain knowledge can have nonstrategic effects, that is, diminish the need for strategy activation.

Evidence for the latter phenomenon comes from studies using the expert-novice paradigm. These studies compared experts and novices in a given domain (e.g., baseball, chess, or soccer) on a memory task related to that domain. It could be demonstrated that rich domain knowledge enabled a child expert to perform much like an adult expert and better than an adult novice—thus showing a disappearance and sometimes reversal of usual developmental trends. Experts and novices not only differed with regard to quantity of knowledge but also regarding the quality of knowledge, that is, in the way their knowledge is represented in the mind. Moreover, several studies also confirmed the assumption that rich domain knowledge can compensate for low overall aptitude on domain-related memory tasks, as no differences were found between high- and low-aptitude experts on various recall and comprehension measures (Bjorklund and Schneider 1996).

Taken together, these findings indicate that domain knowledge increases greatly with age, and is clearly related to how much and what children remember. Domain knowledge also contributes to the development of other competencies that have been proposed as sources of memory development, namely basic capacities, memory strategies, and metacognitive knowledge. Undoubtedly, changes in domain knowledge play a large role in memory development, probably larger than that of the other sources of memory improvement described above. However, although the various components of memory development have been described separately so far, it seems important to note that all of these components interact in producing memory changes, and that it is difficult at times to disentangle the effects of specific sources from that of other influences.

5. *Current Research Trends*

During the 1990s, research interests in memory development shifted from the more basic issues discussed above to certain aspects of event memory or autobiographical memory. In particular, children's eyewitness memories have attracted substantial research attention. Not surprisingly, much of the recent interest has been stimulated by children's increasing participation in the legal system, either as victims of or as witnesses to reported crimes (Ceci and Bruck 1998). Major research interests concerned age trends regarding the reliability of children's memory of witnessed events, age-related forgetting processes, and age trends regarding children's susceptibility to suggestion (Brainerd and Reyna 1998).

The main findings of this research can be summarized as follows: (a) Children's free recall of witnessed events is generally accurate and increases with age. Despite low levels of recall, what preschoolers do recall is usually accurate and central to the witnessed event. (b) Preschoolers are especially vulnerable to the effects of misleading questions and stereotypes. Although young children's erroneous answers to misleading questions do not necessarily reflect an actual change in memory representations, such changes may occur, with young children being more likely to make such changes than older children (Ceci and Bruck 1998). (c) To obtain the most accurate recall, questions should be asked in a neutral fashion, and should not be repeated more often than necessary. (d) Autobiographical memories are never perfectly reliable. Although there is great development during the preschool years and into the elementary school

years with respect to the accuracy and completeness of event memories, no major differences between the event memories of young school children and adults can be found. This does not mean that primary school children's event memory is already close to perfect but simply illustrates the fact that even adults' memories of witnessed events are fallible at times, particularly when delays between the to-be-remembered events and the time of testing are long.

As noted above, the focus of this overview was on the development of explicit, verbal memory. Although there has been less research on other memory systems such as implicit memory (i.e., memory for some information without being consciously aware that one is remembering) or visuo-spatial memory, the available evidence suggests that age differences found for these kinds of memory—if noticed at all—are typically small and far less pronounced than those observed for verbal memory, indicating that findings from one area cannot be transferred to other domains.

See also: Cognitive Development: Child Education; Cognitive Development in Childhood and Adolescence; Cognitive Development in Infancy: Neural Mechanisms; Cognitive Development: Learning and Instruction; Infant Development: Physical and Social Cognition; Lifespan Theories of Cognitive Development; Memory Models: Quantitative; Prefrontal Cortex Development and Development of Cognitive Function; Schooling: Impact on Cognitive and Motivational Development

Bibliography

Bjorklund D F, Schneider W 1996 The interaction of knowledge, aptitudes, and strategies in children's memory performance. In: Reese H W (ed.) *Advances in Child Development and Behavior*. Academic Press, New York

Brainerd C J, Reyna V F 1998 Fuzzy-trace theory and children's false memories. *Journal of Experimental Child Psychology* **71**: 81–129

Ceci S J, Bruck M 1998 Children's testimony: Applied and basic issues. In: Damon W, Sigel I, Renninger K A (eds.) *Handbook of Child Psychology, Vol 4: Child Psychology in Practice*. Wiley, New York

Cowan N (ed.) 1997 *The Development of Memory in Childhood*. Psychology Press, Hove, UK

Fivush R 1997 Event memory in early childhood. In: Cowan N (ed.) *The Development of Memory in Childhood*. Psychology Press, Hove, UK

Flavell J H, Miller P H, Miller S A 1993 *Cognitive Development*, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ

Kail R 1991 Development of processing speed in childhood and adolescence. In: Reese H W (ed.) *Advances in Child Development and Behavior*, Vol. 23. Academic Press, New York

Meltzoff A N 1995 What infants tell us about infantile amnesia: Long-term recall and deferred imitation. *Journal of Experimental Child Psychology* **59**: 497–515

Nelson K 1996 *Language in Cognitive Development: Emergence of the Mediated Mind*. Cambridge University Press, New York

Schneider W, Bjorklund D F 1998 Memory. In: Damon W, Kuhn D, Siegler R (eds.) *Handbook of Child Psychology, Vol. 2: Cognitive, language, and perceptual development*. Wiley, New York

Schneider W, Pressley M 1989 *Memory Development Between 2 and 20*. Springer Verlag, New York

Siegler R S 1996 *Emerging Minds. The Process of Change in Children's Thinking*. Oxford University Press, New York

W. Schneider

Memory for Meaning and Surface Memory

Linguists, psychologists, and philosophers draw a distinction between the meaning of a sentence and the exact wording, or surface form, which conveys that meaning. Consider the following:

(a) The policeman chased the suspect.
(b) The officer chased the suspect.
(c) The suspect was chased by the officer.

Each of these sentences uses a different surface form to convey exactly the same meaning. Psycholinguistic research has shown that people recall the meaning of a sentence much better than its surface form, but that memory for form, under some conditions, is surprisingly robust.

1. Early Research on Memory for Meaning and Surface Form

Sachs (1967) carried out one of the first comparisons between memory for meaning and memory for surface form. She presented listeners with recorded passages that were interrupted zero, 80, or 160 syllables after a critical sentence. During the interruption, listeners were asked to judge whether a test sentence had occurred, in exactly the same form, earlier in the passage. Sometimes the test sentence was identical to the critical sentence (e.g., 'He sent a letter about it to Galileo, the great Italian scientist.'), sometimes it was a paraphrase (e.g., 'He sent Galileo, the great Italian scientist, a letter about it.'), and sometimes it altered the meaning of the critical sentence (e.g., 'Galileo, the great Italian scientist, sent him a letter about it.'). On the immediate test, listeners correctly rejected both the paraphrases and the sentences with altered meanings. But after just 80 syllables of intervening text, performance on the paraphrases showed a precipitous decline. These results were taken as evidence that the

surface form of a sentence is held in working memory long enough for the meaning to be extracted, but only the meaning is stored in long-term memory.

In a subsequent study, Jarvella (1971) interrupted recorded narratives and asked listeners to write down as much of the preceding narrative as they could recall 'exactly, word-for-word.' The two sentences preceding each interruption took one of two forms. The final two clauses were identical in both forms, but in one the next-to-last clause was part of the final sentence while in the other it was part of the preceding sentence. Consider the following examples:

(d) The confidence of Kofach was not unfounded. To stack the meeting for McDonald the union had even brought in outsiders.

(e) Kofach had been persuaded by the international to stack the meeting for McDonald. The union had even brought in outsiders.

About 80 percent of the words in the next-to-last clause (e.g., 'to stack the meeting for McDonald') were recalled verbatim when it was part of the final sentence as in (d), but fewer than 50 percent were recalled correctly when it was part of the penultimate sentence as in (e). These results suggest that sentence boundaries mark the point where the meaning of a sentence is stored in long-term memory and its surface form is irretrievably lost.

2. *Surface Memory*

In the late 1970s and early 1980s several experiments were published demonstrating reliable surface memory for sentences in natural settings. Kintsch and Bates (1977) tested memory for statements made during classroom lectures. They found that college students could reliably discriminate between statements they had actually heard ('Galton was the brilliant younger cousin of Darwin.') and meaning preserving paraphrases of those statements ('Darwin was the older cousin of the extremely intelligent Galton.') as much as five days later. Similar results were reported by Keenan et al. (1977) who investigated memory for statements made during a faculty lunchtime conversation, and by Bates et al. (1980) who tested memory for statements from a television soap opera.

Two factors appear to explain why these researchers found evidence for surface memory while Sachs (1967) and Jarvella (1971) did not. First, the experiments were conducted in natural settings where the surface form of a sentence *can* be socially relevant. This claim is supported by findings that surface memory is most robust for jokes (Kintsch and Bates 1977), mock insults (Keenan et al. 1977), and other socially significant utterances. Second, each of these studies included one or more control groups whose members completed the memory test without prior exposure to the lecture, conversation, or soap opera. The performance of these groups provides a baseline against which surface memory is more easily detected.

Additional evidence of long-term surface memory comes from studies that employ 'indirect' measures of memory. In one such study, Tardif and Craik (1989) created two versions of the same passage. Each sentence in one passage was paraphrased in the other. One week after reading the passage, participants were asked to read it again. Half saw the same version on both readings while half saw different versions. Rereading times were faster when the surface form remained constant.

3. *Memory for Meaning*

Van Dijk and Kintsch (1983) have proposed that the meaning of a discourse is represented in memory by two distinct, but interrelated, knowledge structures that they call the propositional textbase and the situation model.

3.1 *The Propositional Textbase*

Bransford and Franks (1971) argued that the meaning of a discourse is stored in memory as a network of ideas that transcend sentence boundaries. Consider the following:

(f) The ants ate the sweet jelly that was on the table.

(g) The ants in the kitchen ate the jelly.

(h) The ants in the kitchen ate the sweet jelly, which was on the table.

(i) The jelly was sweet.

These sentences convey four basic ideas: the ants ate the jelly, the ants were in the kitchen, the jelly was sweet, and the jelly was on the table. Bransford and Franks presented listeners with sentences like these, followed by a recognition memory test. The primary determinant of recognition performance was the number of ideas conveyed by each test sentence. Listeners were most likely to 'remember' sentence (h), which includes all four ideas, even if they never actually heard it. These results suggest that as each sentence is heard, the ideas are extracted and stored in long-term memory together with related ideas. Test sentences are then recognized by comparing them to the complete, integrated set of ideas in memory.

To account for results such as these, psychologists borrowed the idea of a proposition from linguistics and philosophy. Propositions are sometimes defined as the smallest units of meaning to which we can assign a truth value. It makes sense to ask whether 'the ants ate the jelly' (a complete proposition) is true, but not to ask whether 'ants' by itself is true. Propositions are also defined by their structure. Each proposition consists of a single predicate term (a verb, adverb,

adjective or preposition) and one or more arguments (nouns or other propositions). By either definition, each of Bransford and Franks' 'ideas' corresponds to a single proposition that we can represent as follows:

(j) EAT (ANTS, JELLY)
(k) IN (ANTS, KITCHEN)
(l) SWEET (JELLY)
(m) ON (JELLY, TABLE)

Most psychologists now agree that the meaning of a discourse is represented in memory as a propositional textbase: a network of propositions connected by shared arguments. Thus, EAT (ANTS, JELLY) is connected to SWEET (JELLY) by JELLY and to IN (ANTS, KITCHEN) by ANTS. This claim is supported by over 25 years of research on the comprehension and recall of discourse (for a review, see van Dijk and Kintsch 1983, Chap. 2). Among the most compelling is a study by McKoon and Ratcliff (1980) who presented readers with brief stories followed by a speeded recognition task. Readers were shown a series of probe words and asked to judge, as quickly as possible, whether each word had occurred in one of the stories. Probe words were recognized more quickly when the preceding test word was from the same story. This facilitation effect became more pronounced as the distance between the words in the propositional textbase decreased.

3.2 The Situation Model

During comprehension, listeners and readers use their knowledge of the world to go beyond the propositional textbase and create a representation of the situation described by a discourse, a representation that is as much as possible like the representation that would result from direct experience. Van Dijk and Kintsch (1983) refer to this as a situation model. A situation model can take many forms and often includes sensory-motor information that is not normally associated with strictly linguistic representations. Among the most studied types of situation models are those associated with spatial descriptions and simple narratives.

Mani and Johnson-Laird (1982) demonstrated that a propositional textbase does not adequately describe how spatial descriptions are represented in memory. They presented readers with two types of descriptions:

(n) The bookshelf is to the right of the chair. The chair is in front of the table. The bed is behind the table. (determinate)

(o) The bookshelf is to the right of the chair. The chair is in front of the table. The bed is behind the chair. (indeterminate)

These descriptions differ by a single argument to a single proposition: BEHIND (BED, *TABLE*) versus BEHIND (BED, *CHAIR*). This suggests that there should be little or no difference in how these descriptions are understood and remembered. But the situation associated with the indeterminate description (o) is ambiguous. The bed could be either in front of or behind the table. Because of this, Mani and Johnson-Laird hypothesized that readers would have greater difficulty understanding and remembering indeterminate descriptions. Their experiments confirm this prediction.

Trabasso et al. (1984) were among the first to propose that causal connections play a central role in the mental representation of narratives. They claim that each event in a narrative is understood by determining its causes and consequences, and that the narrative as a whole is understood by finding a chain of causally related events that connect its opening to its eventual outcome. In support of this claim they showed that events with many causal connections to the rest of a narrative are remembered better and rated more important than otherwise similar events with fewer causal connections. They also showed that events on the causal chain that connects the opening of a narrative to its outcome are recalled better and rated more important than causal 'dead ends.' These results suggest that narratives are represented in memory by a complex situation model that includes a representation of each event and the causal connections that tie the events together.

4. Current and Future Directions

A major goal of psycholinguistic research is to create computer models that simulate the comprehension and recall of discourse (e.g., Kintsch 1998). Achieving this goal will depend on understanding, in detail, how the surface forms and meanings of sentences are represented in memory. Another significant research goal is to identify the specific brain regions involved in understanding and remembering discourse. Up to the end of the twentieth century, most research of this type focused on isolated words and out-of-context sentences. Research on how, and where, the brain creates a coherent propositional textbase or situation model is still in its infancy (e.g., Beeman and Chiarello 1998).

See also: Comprehension, Cognitive Psychology of; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Memory Models: Quantitative; Psycholinguistics: Overview; Sentence Comprehension, Psychology of; Text Comprehension: Models in Psychology

Bibliography

Bates E, Kintsch W, Fletcher C R, Giuliani V 1980 The role of pronominalization and ellipsis in texts: Some memory experiments. *Journal of Experimental Psychology: Human Learning and Memory* **6**: 676–91

Beeman M, Chiarello C (eds.) 1998 *Right Hemisphere Language Comprehension: Perspectives from Cognitive Neuroscience*. Erlbaum, Mahwah, NJ
Bransford J D, Franks J J 1971 The abstraction of linguistic ideas. *Cognitive Psychology* **2**: 331–50
Jarvella R J 1971 Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior* **10**: 409–16
Keenan J M, MacWhinney B, Mayhew D 1977 Pragmatics in memory: A study of natural conversation. *Journal of Verbal Learning and Verbal Behavior* **16**: 549–60
Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK
Kintsch W, Bates E 1977 Recognition memory for statements from a classroom lecture. *Journal of Experimental Psychology: Human Learning and Memory* **3**: 150–9
Mani K, Johnson-Laird P N 1982 The mental representations of spatial descriptions. *Memory & Cognition* **10**: 181–7
McKoon G, Ratcliff R 1980 Priming in item recognition: The organization of propositions in memory for text. *Journal of Verbal Learning and Verbal Behavior* **19**: 369–86
Sachs J 1967 Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics* **2**: 437–42
Tardif T, Craik F I M 1989 Reading a week later: Perceptual and conceptual factors. *Journal of Memory and Language* **28**: 107–25
Trabasso T, Secco T, van den Broek P 1984 Causal cohesion and story coherence. In: Mandl H, Stein N L, Trabasso T (eds.) *Learning and Comprehension of Text*. Erlbaum, Hillsdale, NJ
van Dijk T, Kintsch W 1983 *Strategies of Discourse Comprehension*. Academic Press, New York

C. R. Fletcher

Memory for Text

A central issue in theorizing about what readers remember and learn from text is how activation from knowledge in working memory is regulated and stored in memory (Kintsch 1998, Myers and O'Brien 1998). Processes including making inferences and integrations, and constructing and updating situation models determine the quality of the representation that is stored in memory. What is remembered or learned from text is dependent upon this representation. In this article, factors are examined that are relevant to the process of construction of this episodic memory representation.

1. *Relation Between Processing and Memory for Text*

Making inferences and integrating information play important roles in the process of constructing a mental representation during comprehension. In making inferences, readers activate knowledge from their semantic memory that is relevant to, but left implicit in, the text. Integration consists of connecting concepts or propositions to earlier propositions by searching episodic memory, and may be followed by making an inference. The memory representation depends upon these processes. Little, however, is known about the extent to which readers actually perform inference and integration activities during reading. In particular, little attention has been paid to the way in which activation of knowledge in a reader's memory is regulated. A similar problem arises with regard to the construction and updating of situation models (representations of the situation the text refers to or is about) in episodic memory. To what extent do readers construct and update precise situation models, and what factors influence these processes? This article seeks to contribute to a better understanding of the regulation processes involved in making inferences and integrations, in constructing and updating of situation models during comprehension, and, subsequently, influencing memory for text.

2. *Completeness of the Representation*

One block of factors that are relevant to the regulation processes involved in making inferences and integrations are person-related factors. These factors, such as reading goal and habitual reading style of an individual reader, determine the manner in which inferencing and integrating occur. Readers with a high or strict comprehension criterion infer concepts related to a schematic structure (or script) underlying a story and integrate concepts to a greater extent than do readers with a low or careless comprehension criterion. It seems that if a reader is interested in minimal comprehension the standards for coherence are met relatively easily and little activation of background knowledge takes place. They read, for instance, sentences with script arguments (for example, booking office in the context of a train journey story) in a script-based story faster than sentences with nonscript arguments (e.g., bookstall), regardless of whether the arguments have been mentioned before in that story or not. In contrast, if a reader is interested in attaining a thorough understanding of a text, the standards for coherence are very demanding, reading is slow and involves extensive recruiting of background knowledge or of information from the mental representation that has been constructed so far. These slow or careful readers can be characterized as readers who completely instantiate scripts and also integrate encountered concepts in the episodic memory representation. The outcomes of a recognition task presented afterwards, consisting of judging the implicit script arguments as 'new' or 'old,' as Van Oostendorp (1991) showed, support this interpretation. Slow or careful readers infer the scriptal arguments in the implicit condition—when the script argument was not

mentioned—while fast or careless readers do not. They are, therefore, slower in judging the script argument as new because they have similar or related information represented in episodic memory, making the judgment 'new' difficult and slow. These findings indicate how processing influences the memory representation that is being stored.

Regulation also occurs on the basis of textual characteristics. On a sentence level, for instance, it appears that subjects process concepts in semantically less-related sentences more extensively than concepts in semantically high-related sentences (Van Oostendorp 1994). Semantic relatedness was here assessed—among other measures—by means of a rating task in which subjects judge the meaning overlap within each pair of content words in a sentence. Subjects read context sentences such as 'The cat caught a mouse in the kitchen' (which contains highly related concepts) as opposed to 'The cat seized a mole in the field' (less related). These sentences were also embedded in stories. Immediately after reading such a context sentence, a verification question was presented which referred to a relevant attribute of one concept (has claws for cat). These attributes, particularly low-typical attributes, are verified faster after reading semantically less-related context sentences than after reading highly related sentences. Also the cued recall performance of readers is better for sentences that are semantically less related than for high-related sentences. Semantic high-relatedness may, thus, lead to less activation of knowledge, result in the failure to make inferences and, consequently, to a less elaborate episodic memory trace. In a study by Cairns et al. (1981), subjects read sentences with a predictable or with an unpredictable target word in relation to a preceding context. An example of a predictable word is 'gum' in 'Because she was chewing so loudly in class, Sarah was asked to get rid of her gum promptly.' It is unpredictable in 'Because it was annoying the others, Sarah was asked to get rid of her gum promptly.' The reading time for the second sentence is longer than for the first type of sentence. The recognition and reproduction of unpredictable word sentences is also better than of predictable word sentences. Cairns et al. (1981) assume that more knowledge is activated while processing the former sentence, which leads to prolonged processing and to a more elaborate propositional representation. Semantic relatedness may even induce failures to notice errors in sentences. Van Oostendorp and De Mul (1990) presented subjects with sentences such as 'Moses took two animals of each kind on the Ark. True or false?' A majority of subjects answer erroneously 'true' although the subjects know the correct name (Noah), as was shown by a later test. Furthermore, sentences with high-related inaccurate names (Moses) lead to more semantic illusions than sentences with low-related inaccurate names (e.g., Adam).

One may conclude that readers continuously monitor the semantic cohesion of a mental representation under construction and regulate further processing on the basis of a comparison of the perceived cohesion to some internal comprehension standard. Myers and O'Brien (1998) take a similar standpoint on the control of processing. During initial processing of a sentence, the perceived cohesion of a propositional representation is primarily dependent on the semantic relatedness between involved concepts (Van Oostendorp 1994, see also Kintsch 1974, p. 214 for the same idea). Often, readers have the competence to making inferences and integrations, but frequently they don't completely employ this capacity because the initial perceived cohesion is above the standard readers have set.

3. *Updating Mental Representations*

The same problem with regard to the completeness of the propositional representation can be raised concerning the construction and updating of situation models. Do readers construct and represent in memory detailed situation models under naturalistic conditions, that is, with a more naturalistic text and with a more naturalistic reading task than those often used in laboratory studies? And, also, do they accurately update their model when reading new (correcting) information? Readers often do not form integrated spatial-situation models during comprehension of a naturalistic story, even when they have the opportunity to reread the text, nor do they update their model accurately in episodic memory (Zwaan and Van Oostendorp 1994). The text used in these studies was a part of a detective story. Only when readers are specifically instructed to construct a mental map of the situation described in the story do they form and update spatial situation models. Other studies basically confirm these findings (e.g., Hakala 1999).

Under certain circumstances, a text with too much coherence can also be detrimental to constructing an adequate situation model. A text that is fully explicit and coherent at both the local and global level may result in impaired comprehension at a situation-model level—measured by problem-solving questions—at least for readers with high, prior knowledge (McNamara et al. 1996). Apparently, a highly coherent text may hinder deeper understanding of high knowledge readers because it reduces their amount of active processing during reading and, as a result, they fail to construct an adequate situation model.

Also when newspaper articles are used updating in episodic memory of situation models is not always effective. In one study, for instance, subjects were presented with a text about the situation in Somalia at the time of the US operation, 'Restore Hope,' followed by a second, related text (Van Oostendorp 1996). The first text reported that 'operation Restore Hope started

under American control,' and in the second text it was said that the command structure of the operation had been changed, such that 'The United Nations took over the command in Somalia.' If updating is correct, the old information is replaced by the new information. After reading the second text, readers received an inference test with test items such as 'The USA troops operate under the UN flag (True/False?).' Results show that the updating performance is, in general, very low. Furthermore, readers who have available an originally appropriate situation model perform a higher degree of updating. That is, the more accurate the original information is represented in memory, the better and faster they judge inferences concerning transformations in the second text. A second, more remarkable, finding is that transformations that are more important to the situation described are less updated than less important transformations. This result has been observed in two experiments with different materials, test items, etc. (Van Oostendorp 1996). It seems that the central part of a situation model may be less easily updated than the peripheral parts. The central part of a mental representation is also less updated when this information is in focus, at least for readers with small initial conceptual networks (Van Oostendorp and Van der Puil 2000). Focus is here manipulated by letting readers compare one text with another along some dimension. The way of examining whether the representation has been updated is based on the cued-association task designed by Ferstl and Kintsch (1999). With this task subjects are presented with a word and are asked to provide an association to it. Based on these data, a proximity matrix is calculated for each subject. Subsequently, the similarity of these matrices or conceptual networks, before, as well as after, reading a second text containing new, correcting information, is calculated for the focus group and for a control, nonfocus group. For subjects with a small initial conceptual network there is less updating in the focus group compared to the nonfocus group. These results correspond to what was mentioned previously: changes can be updated less easily with important information, i.e., information in focus, than with less important information.

One important reason why updating may fail is that it is often difficult for readers to discredit old information completely and to exchange that for new information. For example, in reading a story on a fire in a warehouse (Van Oostendorp and Bonebakker 1999), readers in the experimental condition read a sentence such as 'inflammable materials were carelessly stored in a side room.' Later, they read that 'the side room happened to be empty.' Instead of the sentence 'inflammable materials were carelessly stored in a side room,' readers in a control condition received, a neutral sentence, irrelevant to the cause of the fire. The influence of old, obsolete information in the experimental condition could not be fully neutralized by the new, discrediting information. Answers on inference questions, such as what was the cause of the explosion or, for what reason could an insurance company here refuse a claim, were frequently based on the old information, even by subjects who were aware of the fact that information was discredited. Readers in the experimental condition more often gave answers such as 'because of careless behavior of the owner' than in the control condition. Recall and direct questions showed that almost all readers had the corrections available but still did not use it during processing of the text. In these experiments, even the explicit instruction that information might be corrected does not lead to a better updating. Readers continue to use the misinformation represented in episodic memory, and keep making inferences based on the corrected information (Johnson and Seifert 1999, Wilkes and Reynolds 1999). It is interesting to know the limits within which readers hold on to old information in memory and don't take into account new, correcting information. This issue has been explored by Van Oostendorp et al. (in press) using expository text with scientific content. For example, a text that was used, explained a method aimed at increasing the strength of ceramic materials. The old information stated that 'the treatment (of adding silicon atoms) takes some days,' while a number of sentences later it was mentioned that 'by recent advancements the treatment takes as little as a few minutes.' Inference questions were presented that could tap the interpretation of readers about intermediate events, in order to examine whether the old or the new information source influences this interpretation. For instance, an intermediate sentence contained 'Reaction speed can be measured by an external device.' The inference questions were presented after reading the text. The subjects were asked, for example, about the unit of time this device should use to control the process of hardening ceramic materials. Answers of hours or days would mean that readers mainly base their answer on the old information source, as opposed to answers of minutes or seconds, which would mean that they use primarily the new information source. Strengthening of old information in the memory representation—by repeating it in paraphrased form or referring to it indirectly—appears to lead to less updating, that is, to less use of the new information. And, alternatively, readers who read a text in which the new information is reinforced are more likely to use the new information as the basis of inferences. A regulation mechanism seems to be present that is based on weighting evidence favoring either old or new information. According to the outcomes of this evaluation process of the sources, readers choose one or the other point of view, and use that for making inferences, even backwards ones. Thus, in terms of the recent Construction–Integration model of Kintsch (1998), the balance between the strength of sources—the activation values of old and new information in

the mental representation—may influence the degree of updating. Subtle reinforcements of old and new information can activate qualitatively different updating strategies, such as holding on to the old information and rejecting the new information, or, on the contrary, switching to a new perspective (as we saw in the studies briefly discussed here).

4. Conclusions

Memory for text depends upon the construction of the mental representation during understanding, more specifically upon the completeness of inferencing and integration, and the extent of updating the mental representation. A number of factors, textual, individual and contextual, are involved with the construction process.

Regarding textual conditions, it appears that texts with semantically highly related concepts led to superficial processing and not noticing errors (as in, e.g., the Moses-illusion experiments and the cat-caught-a-mouse experiments) and, consequently, to a shallow memory representation. Furthermore, the type of text seems to be important to updating. Expository texts were used in some studies, and readers were able to update their situation models (as in the strengthening-of-ceramic-materials experiments). In contrast, this updating seems to be difficult when stories are used about everyday events (as in the fire-in-warehouse experiments). In addition, the exact character of the correction itself is important. Logical inconsistencies are probably easy to detect but difficult to repair, while a correction or reported change in the world (as in the Somalia experiments) may be more difficult to detect but easy to understand (and to repair). The explicitness, relevance or saliency of old and new information, respectively also influence updating.

Individual characteristics of readers constitute the second block of variables. Relevant here are reading style (Van Oostendorp 1991), prior knowledge, working memory capacity, and beliefs, such as the epistemological belief that integration of ideas implied by a text is important to understanding.

Finally, completeness of processing and updating memory also depends on contextual conditions, such as setting, instruction, and reading goals (Van Oostendorp 1991).

In summary, the processing of readers can often be incomplete in several ways, and these factors have to be taken into account in order to achieve a valid theory of memory for text that can explain the imperfect but also the adaptive memory performance of readers.

See also: Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Literary Texts: Comprehension and Memory; Semantic Knowledge: Neural Basis of; Semantic Processing: Statistical Approaches; Text Comprehension: Models in Psychology

Bibliography

Cairns H S, Cowart W, Jablon A D 1981 Effects of prior context upon integration of lexical information during sentence processing. *Journal of Verbal Learning and Verbal Behavior* **20**: 445–53

Ferstl E C, Kintsch W 1999 Learning from text: Structural knowledge assessment in the study of discourse comprehension. In: Van Oostendorp H, Goldman S R (eds.) *The Construction of Mental Representations During Reading*. Erlbaum Associates, Mahwah, NJ, pp. 247–78

Hakala C M 1999 Accessibility of spatial information in a situation model. *Discourse Processes* **27**: 261–79

Johnson H M, Seifert C M 1999 Modifying mental representations: Comprehending corrections. In: Van Oostendorp H, Goldman S R (eds.) *The Construction of Mental Representations During Reading*. Erlbaum Associates, Mahwah, NJ, pp. 303–18

Kintsch W 1974 *The Representation of Meaning in Memory*. Erlbaum Associates, Hillsdale, NJ

Kintsch W 1998 *Comprehension. A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK

McNamara D S, Kintsch E, Songer N B, Kintsch W 1996 Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction* **14**: 1–43

Myers J L, O'Brien E J 1998 Accessing the discourse representation during reading. *Discourse Processes* **26**: 131–57

Van Oostendorp H 1991 Inferences and integrations made by readers of script-based texts. *Journal of Research in Reading* **14**: 3–21

Van Oostendorp H 1994 Text processing in terms of semantic cohesion monitoring. In: Van Oostendorp H, Zwaan R A (eds.) *Naturalistic Text Comprehension*, Ablex Publishing Corporation, Norwood, NJ, pp. 97–115

Van Oostendorp H 1996 Updating situation models derived from newspaper articles. *Medienpsychologie. Zeitschrift fur Individual- und Massenkommunikation* **8**: 21–33

Van Oostendorp H, Bonebakker H 1999 Difficulties in updating mental representations during reading news reports. In: Van Oostendorp H, Goldman S R (eds.) *The Construction of Mental Representations During Reading*, Erlbaum Associates, Mahwah, NJ, pp. 319–39

Van Oostendorp H, De Mul S 1990 Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta Psychologica* **74**: 35–46

Van Oostendrop H, Otero J, Campanario J in press Conditions of updating during reading. In: Louwerse M, Van Peer W (eds.) *Thematics: Interdisciplinary Studies*. John Benjamins, Amsterdam

Van Oostendorp H, Van der Puil C 2000 *The Influence of Focus on Updating a Mental Representation*. Proceedings of the 10th Annual Conference of the Society for Text and Discourse, University of Lyon, France

Wilkes A L, Reynolds D J 1999 On certain limitations accompanying readers' interpretations of corrections in episodic text. *The Quarterly Journal of Experimental Psychology* **52A**: 165–83

Zwaan R A, Van Oostendorp H 1994 Spatial information and naturalistic story comprehension. In: Van Oostendorp H, Zwaan R A (eds.) *Naturalistic Text Comprehension*, Ablex Publishing Corporation, Norwood, NJ, pp. 97–115

H. Van Oostendorp

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Memory: Genetic Approaches

There is a long history documenting the usefulness of genetic approaches in studies of brain function, including learning and memory. These approaches fall into two general categories: forward and reverse genetics. Forward genetics is concerned with the identification of genes involved in biological processes such as learning and memory. The starting point of these studies is usually the identification of mutant organisms with interesting phenotypic changes, and their goal is to identify the mutations underlying these changes. In reverse genetic studies, the gene is already at hand, and the goal is to define its role in biological processes of interest. This normally involves the derivation and study of organisms with defined genetic changes. Although the principal purpose of genetic approaches is to study how genetic information determines biological function, recently animals with genetically engineered mutations have been used to develop and test multidisciplinary theories of learning and memory that go well beyond gene function.

1. *The Role of Genetics in Biology*

To explore the role of genetics, it is important to place it in the large context of biological investigations. The ultimate goal of biological research is to develop and test explanations of complex phenomena such as learning and memory. At the heart of this process is the establishment of causal connections between phenomena of interest, such as changes in synaptic function and learning. There are four complementary general strategies that science uses to make causal connections between phenomena of interest. One of these strategies is the *Lesion* strategy. Thus, pharmacological and genetic lesions of calmodulin-induced kinase II (CaMKII) are known to result in deficient long-term potentiation (LTP) of synaptic function, suggesting a connection between the activation of this synaptic kinase and LTP (Silva et al. 1997). It is noteworthy that genetics and pharmacology are the only two approaches to interfere with molecular function in biology. The second strategy that science uses to make causal connections between phenomena of interest is the direct *observation* of these phenomena in their natural context. For example, the induction of LTP is accompanied by observable increases in CaMKII activity. The third strategy involves the *induction* of one phenomenon by the other. For example, injection of activated CaMKII into pyramidal neurons in hippocampal slices induces an LTP-like phenomenon. Finally, *modeling* plays a critical role in making causal connections between phenomena of interest. To assert that two natural phenomena are connected, it is essential to understand something about the mechanism that links them. Thus, CaMKII activation is thought to trigger LTP by phosphorylating and thus enhancing the function of synaptic glutamate receptors. Each of the strategies mentioned above is insufficient to connect two phenomena of interest. Instead, convergent evidence from all four strategies is needed. Therefore, since genetics is one of only two general molecular-lesion approaches available, it is easy to see why it has played a key role in biology. Besides its key role in testing hypotheses (i.e., CaMKII is required for LTP induction), it can be argued that the principal role that genetics has played in biology has been to suggest possible hypotheses or explanations of natural phenomena. Indeed, forward genetic screens have allowed biologists to make major discoveries, even in the absence of a well-delineated hypothesis.

2. *Forward Genetics*

Long before we had the ability to directly manipulate genes in animals such as flies and mice, geneticists were busy using chemical mutagens to alter genetic information in living systems (forward genetics). The goal of classical or forward genetics, which continues to be used extensively to this day, is to identify the genes critical for biological processes of interest. The idea is that study of those genes is often a critical first hint for unraveling underlying biological processes. In forward genetic screens, animals are first exposed to a mutagen, for example, the DNA-altering compound ethyl-nitroso-urea, mated, and the progeny are screened for phenotypic changes of interest. The phenotype of a mutant is the sum total of observed biological changes caused by a genetic manipulation. Recent application of this approach in the study of mammalian circadian rhythms resulted in the identification of *clock*, a crucial link in the cascade of transcriptional events that marks molecular time in organisms as diverse as Drosophila and mice (Wilsbacher and Takahashi 1998).

Other molecular components of this pathway, such as *per*, were isolated in mutagenesis screens in Drosophila. By identifying novel and unexpected molecular components of biological processes of interest, forward genetics has often reshaped entire fields of research. At times, science can go in circles, obsessively chasing its own tail of half-truths, incapable of escaping the gravitational pull of its worn out paradigms. Forward genetics, in the hands of masters such as Edward Lewis (developmental mutants) and Seymor Benzer (learning mutants), has the ability to turn paradigms upside down, and initiate new lines of scientific inquiry. The key to the success of forward genetics is the design of biological screens with which the randomly mutagenized animals are tested. If the screens are too stringent, or if the fundamental biological insights underlying the screen's design are off mark, one runs the risk of ending up with empty hands, or even worse, with a number of misleading

mutants. In contrast, nonstringent designs lead to overwhelming numbers of nonspecific mutants that are essentially useless.

3. The First Screens for Learning and Memory Mutants

Seymor Benzer and colleagues working with Drosophila at the California Institute of Technology designed the first successful screen for learning and memory mutants in the 1970s (Dudai 1988). Benzer and colleagues developed a behavioral procedure with operant and Pavlovian components. During training the flies were allowed to enter two chambers, each with a different odorant, but they only got shocked in one of the chambers. During testing approximately two-thirds of the trained flies avoided the chamber with the odorant that previously had been paired with shock. With this procedure, Benzer and colleagues tested a number of *Drosophila* lines derived from flies treated with ethylmethane sulfonate (EMS). The first mutant line isolated from this screen was *dunce* (Dudai 1988).

Remarkably, three out of the four learning and memory mutations, first discovered in genetic screens in *Drosophila*, code for members of the cAMP-signaling pathway. For example, *dunce* lacks a phosphodiesterase that degrades cAMP. Importantly, these findings have recently been extended into vertebrates, where electrophysiological and behavioral studies have confirmed the critical importance of cAMP signaling to learning and memory (Silva et al. 1998). Remarkably, in the early 1970s Eric Kandel and his colleagues at Colombia University also found evidence for the importance of cAMP signaling in learning and memory with a completely different approach. They used a reduced cellular preparation to study sensitization, a nonassociative form of learning, in the sea snail *Aplysia* (Byrne and Kandel 1996). They also found that sensitization depends on cAMP signaling. This is a fascinating example of convergent evidence in science, but it also serves to illustrate that genetics, like any other tool in science, is most successful when used in parallel with other approaches. The persuasive power of convergent evidence cannot be overemphasized. Besides identifying new genes, genetics can also be used to test hypotheses about the function of cloned genes (reverse genetics).

4. Reverse Genetics

In classical genetics an interesting phenotype is usually the driving force behind the molecular experiments required to identify the underlying mutant gene(s). In contrast, in reverse genetics, the interesting molecular properties of a gene usually drive the generation and study of the mutant animal (hence, the word reverse). It is now possible to delete and add genes to many species, ranging from bacteria to mice. For example, mice can be derived with the deletion (knockouts) or overexpression (transgenics) of almost any cloned gene. These manipulations can involve whole genes or they can target specific domains or even single base pairs.

To generate knockout mice, the desired mutation is engineered within the cloned gene, and this mutant DNA is introduced into embryonic stem (ES) cells. Since ES cells are pluripotent, they can be used to derive mice with the genetically engineered lesion. For that purpose, they are injected into blastocysts (early embryos), and the blastocysts are implanted in host mothers. The resulting chimeric (having mutant and normal cells) offspring are then mated to obtain mutant mice. In contrast, transgenic mice are derived by injecting the pronuclei of fertilized eggs with a DNA construct carrying a gene of interest under the regulation of an appropriate promoter. The injected eggs are transplanted into pregnant females and some of the resulting progeny will have the transgenic construct inserted randomly in one of its chromosomes.

With classical knockout and transgenic techniques it is not possible to regulate the time and the regions affected by the mutation/transgene. However, recent techniques promise to circumvent these limitations with a variety of techniques. For example, the expression of the gene of interest can be regulated by gene promoters that can be controlled by exogenously provided substances, such as tetracycline derivatives (Mayford et al. 1997). Alternatively, it is also possible to regulate the function of a protein of interest by fusing it with another protein that can be regulated by synthetic ligands such as tamoxifen (Picard 1993). For example, our laboratory has recently showed that a transcriptional repressor called CREB can be activated at will when fused with a ligand-binding domain (LBD^m) of a modified estrogen receptor. Addition of tamoxifen (the ligand of the modified receptor) activates the $CREB^r/LBD^m$ fusion protein. It is important to note that irrespective of the exact method used, the general idea of reverse genetic studies is that the function of a gene can be deduced from the phenotype of the mutant animal.

5. Knockouts, Transgenics, and Learning

The first knockout/transgenic studies of learning and memory analyzed mice with a targeted mutation of the α isoform of CaMKII (Grant and Silva 1994). Pharmacological studies had previously shown that this family of calcium calmodulin induced kinases present in synapses were required for LTP, a stable enhancement in synaptic efficacy thought to contribute to learning and memory. Remarkably, deleting αCaMKII resulted in profound deficits in hippocampal LTP and in hippocampal-dependent learning and memory. Additional studies showed that either the overexpression of a constitutively expressed form

of the kinase or a mutation that prevented its autophosphorylation also disrupted LTP and learning. Importantly, studies of hippocampal circuits that fire in a place-specific manner (place fields) showed that these CaMKII genetic manipulations disrupted the stability of these place representations (but not their induction) in the hippocampus. Altogether, these studies suggested the provocative hypothesis that this kinase is important for the induction of stable synaptic changes, that the stability of synaptic changes is crucial for the stability of hippocampal circuits coding for place, and that these circuits are essential for spatial learning (Elgersma and Silva 1999). Even this very abbreviated summary demonstrates that these studies went well beyond gene function. Instead, they used mutations to test hypotheses that connected molecular, cellular, circuit, and behavioral phenomena. Although it is reasonable to claim that the αCaMKII has a direct role in the regulation of synaptic function (for example, by phosphorylating glutamate receptors), it is more problematic to argue that the kinase is regulating spatial learning. There are many more phenomenological steps between kinase function and the animal's ability to find a hidden platform in a water maze than between the biochemical properties of this kinase and its role in synaptic plasticity. By comparison, it is easier to see how the stability of place fields in the hippocampus could be an important component of hippocampal-dependent spatial learning.

6. *Common Concerns with the Interpretation of Transgenic/Knockout Studies*

Despite the power of genetics there are a number of concerns that must be kept in mind when using genetic approaches. One of the most commonly discussed is the possibility that developmental effects or any other change caused by the mutation preceding the study could confound its interpretation. Another pertains to the possible effects of genetic compensation. Since proteins do not work alone, but instead function in highly dynamic networks, it is often observed that specific genetic changes lead to alterations/compensations in the function of other related proteins. A related concern pertains to genetic background. Extensive studies have shown that the genetic background of a mutation has a profound effect on its phenotype. The concerns listed above are not limitations of genetics, but simply reflect the properties of the biological systems that genetics manipulates. At the heart of many of the concerns described above are two misconceptions concerning the nature and organization of biological systems.

First, genetics is essentially a lesion tool. Like other lesion tools, it cannot be used in isolation. To establish causal connections between any two phenomena (A and B) in science, it is never enough to lesion A and document the alteration of B. As described above, it is also critical to fulfil three other criteria: first, A must be observed to precede B; second, triggering A should result in B; finally, it is essential to have a clear hypothesis of how A triggers B. Fulfilling only one or two of those four criteria is simply not enough to establish a causal connection between A and B. Therefore, although studying the effects of a deleted protein is an important component in determining its function, it is by no means sufficient.

Second, biological systems are dynamic and adaptive, and, therefore, the lesion of any one component is always followed by changes in several other components. Although it is often a helpful simplification to think of biological components as independent functional units, it is important to remember that they are not. Thus, it is hardly surprising that the effect of a mutation is dependent on biological variables such as genetic background.

7. *The Future of Genetic Manipulations*

In the near future it will be possible to delete or modify any gene, anywhere in most organisms of interest, and at any time of choice. Additionally, more powerful forward genetic strategies will allow the isolation of entire pathways of genes involved in any neurobiological phenomenon of interest, including learning, attention, emotion, addition, etc. In parallel with expected advances in genetics, there will also be advances in the methods used to analyze mutants. These advances are just as critical to genetic studies as advances in genetic methodology. For example, imaging the brain of mutant mice may yield insights into how molecular lesions affect the function of brain systems. Most genetic studies of learning and memory in mice have focused on the relationship between cellular phenomena (i.e., LTP) and behavior. Advances in small animal magnetic resonance imaging (MRI) may enable the kind of functional system analysis in the brains of mutant mice that have so far only been possible in large primates. Similarly, multiple-single unit recording techniques are starting to yield system-wide snapshots of circuit activity in the brains of mutant mice. At a molecular level, small-size positron emission tomography (PET) devices will allow the imaging of molecular function in living animals such as mice. For example, it may be possible to image the activation of a receptor such as the dopamine receptor during learning or memory. Microarray techniques and other molecular cloning approaches will allow the identification of gene profiles in mutant mice. These molecular profiles will be critical to delineating the molecular changes behind the expression of a mutant phenotype. It is important to note that genetics allows us to reprogram the biology of organisms. The finer and more sophisticated the phenotypic and genotypic tools that we have at our

disposal, the deeper we may be able to probe the magical natural programs embedded in our genes.

See also: Learning and Memory, Neural Basis of; Memory in the Bee; Memory in the Fly, Genetics of

Bibliography

Byrne J H, Kandel E R 1996 Presynaptic facilitation revisited: State and time dependence. *Journal of Neuroscience* **16**(2): 425–35

Dudai Y 1988 Neurogenetic dissection of learning and short-term memory in Drosophila. *Annual Review of Neuroscience* **11**: 537–63

Elgersma Y, Silva A J 1999 Molecular mechanisms of synaptic plasticity and memory. *Current Opinion in Neurobiology* **9**(2): 209–13

Grant S G, Silva A J 1994 Targeting learning. *Trends in Neurosciences* **17**(2): 71–5

Mayford M, Mansuy I M et al. 1997 Memory and behavior: A second generation of genetically modified mice. *Current Biology* **7**(9): R580–9

Picard D 1993 Steroid-binding domains for regulating functions of heterologous proteins in cis. *Trends in Cell Biology* **3**: 278–80

Silva A J, Kogan J H et al. 1998 CREB and memory. *Annual Review of Neuroscience* **21**: 127–48

Silva A J, Smith A M et al. 1997 Gene targeting and the biology of learning and memory. *Annual Review of Genetics* **31**(5352): 527–46

Wilsbacher L D, Takahashi J S 1998 Circadian rhythms: Molecular basis of the clock. *Current Opinion in Genetics and Development* **8**(5): 595–602

A. J. Silva

Memory in the Bee

Memory relates to changes in the brain initiated by learning. These changes must represent the acquired information in forms that can be used to guide adapted perception and behavior. Since the memory content's code is unknown, memory is accessible only via retrieval from a store, and thus memory always has two intimately entangled aspects: storage and retrieval. Mechanistic approaches to memory have focused on the storage processes, and retrieval has been mostly neglected because of the lack of experimental tools. Since Ebbinghaus (1885) it has been known that learning leads to memory, but memories are not made instantaneously by learning, rather, they develop and change over time. When Hermann Ebbinghaus began the scientific study of memory in humans, he discovered that memory formation is a time-consuming process and depends on the interval between sequential learning trials. Later, James (1890) developed the concept of primary and secondary memory, referring to the limited timespan and information capacity of primary memory, and the seemingly unlimited duration and content of secondary memory. Around the turn of the century, psychologists had established a framework of thinking about sequential memory stages, which was captured by the perserveration-consolidation hypothesis of Müller and Pilzecker (1900). Neural processes underlying newly formed memories initially perseverate in a labile form and then, over time, become consolidated into lasting neural traces.

In this sense, memory dynamics is not restricted to humans and mammals, but is a general property in animals. Using invertebrate model systems one can ask basic questions of memory formation. Why should memory take hours, days, or weeks for final adjustment of the circuit? Is the neural machinery so slow? The analysis of the memory trace's molecular and neural properties has gained from studies of invertebrate species, because of their small number of large neurons, as in the marine slug *Aplysia*; their well-worked-out classical and molecular genetics, as in the fruit fly *Drosophila*; or their potential to record memory stage correlates in alert and behaving animals even at the level of single neurons and circuits, as in an insect, the honeybee *Apis mellifera*. These studies prove that even in such rather simple systems memory is a highly dynamic process of multiple memory traces at multiple neural sites.

The combined temporal and spatial properties of the memory trace can be studied very well in the bee, because a bee will associate an odor with reward quickly, even under conditions when the brain is exposed to neural recordings (Menzel 1999). Natural learning behavior in bees is well defined, because appetitive memory formation occurs during foraging on flowers, which are rather unpredictable and widely distributed food sources. Perseveration and consolidation of memories appear to be adapted to the demands and constraints of the specific requirements to which the bee is exposed in its natural environment. Thus a bee, like other animals, behaves at any certain time with reference to information gathered over long periods of time, and at any particular moment information is evaluated according to genetically controlled internal conditions (thus the phylogenetic history of the species) and parameters of the experience gathering process such as reliability, context-dependence, and what new information means to the animal. These aspects are a function of time. New memories must be incorporated into existing ones based on their relevance.

The concepts evolving from these arguments can be studied in bees using a classical conditioning paradigm in which a harnessed bee learns to associate an odor with reward and expresses its retention by extending its tongue in response to the odor alone. The neural circuit (Fig. 1(A)) underlying this behavior is well

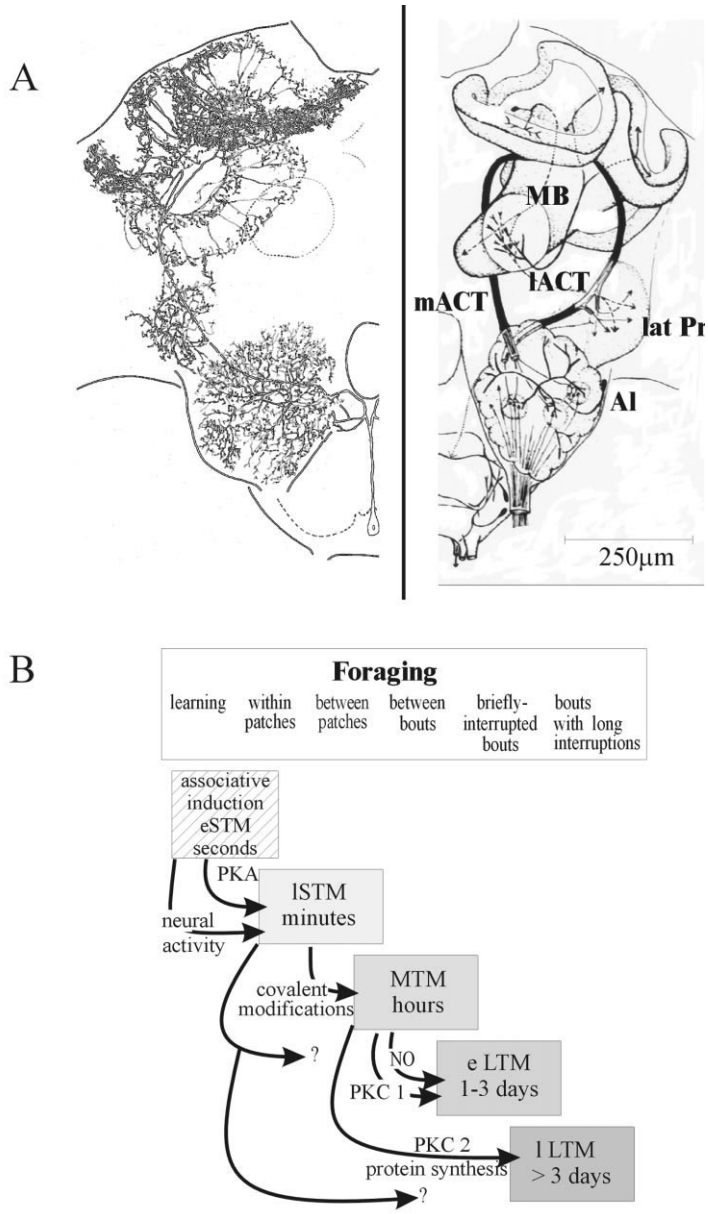

Figure 1
(A) Schematic representation of the olfactory (right) and the reward (left) pathways in the bee brain. These pathways are symmetrical on both sides of the brain; here one side of each is shown. The reward pathway is represented by a single identified neuron, the VUMmx1 (Hammer 1993). The antennal lobe (Al) is the first sensory integration center. The projection neurons (mACT, lACT) connect the Al with the mushroom bodies (MB), a second-order multisensory integration center, and the lateral protocerebrum (lat Pr), a premotor area. (B) Model of the memory stages and their respective cellular substrates (see text). Upper panel gives the sequence of behavioral events during foraging on flowers (Menzel 1999). eSTm: early short-term memory; lSTM: late short-term memory; MTM: mid-term memory; eLTM: early long-term memory; lLTM: late long-term memory; PKA: protein kinase A; NO: nitrogen monoxide; PKC: protein kinase C; PKC1: protease-dependent PKC activity; PKC2: protein synthesis-dependent PKC activity. The arrows indicate transitions between the memory phases as related to the cellular events indicated.

described (Hammer and Menzel 1995, Hammer 1997), and learning-related changes can be monitored by intracellular electrodes or by imaging fluorescence patterns of Ca^{2+} activity (Faber et al. 1999, Menzel and Müller 1996). Furthermore, molecular correlates can be measured at high temporal resolution by quickly blocking enzymatic activity and specific detection of second messenger-dependent reaction cascades in small compartments of the bee brain (Müller 1996).

The dynamics and distribution of the memory trace can be conceptualized as a stepwise, time- and event-dependent process (Fig. 1(B)) reflecting the temporal-spatial characteristics of multiple memory traces. Such an analysis helps correlate the underlying mechanisms of behavioral data.

1. *Associative Induction and Early Short-term Memory (eSTM)*

An associative learning trial leads to associative induction and an early form of short-term memory (eSTM) in the range of seconds. This memory is initially localized in the primary sensory neuropil, the antennal lobe, and is highly dominated by appetitive arousal and sensitization induced by the unconditioned stimulus (US), sucrose. Thus eSTM is restricted to the first synaptic connections in the sensory pathway. It is rather unspecific and imprecise. The associative component, which connects specific stimuli with motor actions, already exists at a low level, and develops later over a period of several minutes (consolidation). eSTM covers the time window during which bees can expect to be exposed to the same stimuli, since flowers most frequently occur in patches. No specific choices need to be performed at this time, and general arousal (depending on the strength of the US) will suffice to control whether the animal stays in the patch or postpones choices for a later time.

At the cellular level, stimulus association is reflected in the convergence of excitation of the pathways for the conditioned stimulus (CS) and the unconditioned, rewarding stimulus (US). There are three neuroanatomical convergence sites of these pathways (Fig. 1(A): antennal lobe, lip region of the mushroom bodies, lateral protocerebrum); two of these sites can independently form an associative memory trace for an odor stimulus. The dynamics of these two traces are different: the antennal lobe (first-order neuropil) establishes the trace quickly and gradually, the mushroom body stepwise by a consolidation process in the range of minutes during the transition to a late form of STM (lSTM). The molecular substrates for the formation of the associative links are unknown. Neither a glutamate receptor, as shown for the hippocampus, nor adenylylcyclase, as proposed for *Aplysia*, appear to play a role in the bee. However, cAMP upregulation of PKA during STM is necessary for memory transition to mid-term memory (MTM) and long-term memory (LTM), and an increase in PKA activity is specifically connected with associative trials, indicating PKA's role in consolidation to lasting memories established during consolidation (Fig. 1(B)).

2. *Late Short-term Memory (lSTM)*

The transition to the selective associative memory trace during lSTM is a rather slow process after a single learning trial lasting up to several minutes, and a quick one after multiple learning trials. Thus, consolidation is both time- and event-dependent, where events must be associative experiences and not just CS or US repetitions. The mushroom bodies, a high-order, multiple sensory integration neuropil, appear to be selectively involved.

The behavioral relevance of these findings for foraging behavior under natural conditions may be related to the temporal separation between intra- and interpatch visits. First, memory needs to be specific after leaving a patch, because distinctions need to be made between similar and different flowers. Second, such a specific memory trace should also be established after a single learning trial, because in some rare cases a single flower may offer a very high amount of reward. Third, discovering a rewarding flower in a different patch means that the local cues just learned are now presented in a different context. lSTM is, therefore, a component of a working memory phase (together with eSTM), during which context dependencies are learned. Fourth, consolidation might be a highly beneficial property for a foraging bee which must optimize its foraging efforts with respect to reward distribution. In the flower market bees can extract the profitability of a flower type only on the basis of its frequency within the foraging range and its probability of offering food. Time- and event-dependence of memory consolidation in lSTM could be a simple means of extracting this information.

Consolidation into lasting memories depends not only on PKA activity but also on NO synthase activity, indicating that both cAMP and cGMP are important second messengers for consolidation. It is likely that the target of both second messengers is PKA, supporting the interpretation that high PKA activity is essential for the transition to long-lasting memories during consolidation (Fig. 1(B)).

3. *Mid-term Memory (MTM)*

At the beginning of MTM, behavior is controlled by consolidated, highly specific memory. At this stage memory is more resistant to extinction, conflicting

information, and elapsing time, and some information about the context dependencies may have already been stored. Under natural conditions bees have usually returned to the hive and departed on a new foraging bout within the time window of MTM. Upon arrival at the feeding area, memory for flower cues no longer resides in working memory (eSTM), but needs to be retrieved from a more permanent store. Therefore, MTM is a memory stage clearly disconnected from a continuous stream of working memory.

MTM is physiologically characterized by a wave of protease-dependent PKC activity in the antennal lobe (Fig. 1(B)). It might, be therefore, that the primary sensory neuropil is a substrate for MTM. Since the mushroom bodies appear to be the critical substrate for consolidation during lSTM, it has been speculated that the MTM trace in the antennal lobe is established under the guidance of the mushroom bodies. The mushroom bodies provide the information which relates the memory traces in the primary sensory neuropils to context stimuli across modalities. Output neurons of the mushroom bodies feed back to the antennal lobe. A particular output neuron, the Pe 1, projecting to a premotor area, shows associative plasticity only during lSTM (Mauelshagen 1993). It might thus be that the mushroom bodies are only involved during working memory (eSTM and lSTM), and provide the information necessary to couple the long-term stores in the sensory (and possibly also the motor) neuropils.

4. *Long-term Memory (LTM)*

LTM is divided into two forms, an early LTM (eLTM, 1–3 days) characterized by protein synthesis-dependent PKC activity, but not by protein synthesis-dependent retention, and late LTM (lLTM, > 3 days) protein synthesis-dependent retention and no enhanced PKC activity. The transition from lSTM to both forms of LTM appears to be independent of MTM, because inhibiting the characteristic substrate of MTM (protease-dependent enhancement of PKC activity) does not prevent eLTM and lLTM being formed (Fig. 1(B)). LTM requires multiple learning trials, indicating that specific information which can be extracted only from multiple experiences (signal reliability, context dependence) controls transfer to LTM. This transfer can be related to a change in the proportion of the activator and repressor forms of the PKA-responsive transcription factor CREB in *Drosophila* (Yin et al. 1995a, 1995b), but the role of CREB in the bee is still unknown. The picture emerging from findings on *Drosophila* is that LTM formation can be actively suppressed, rather than it being automatically produced with associative events accumulating and time elapsing. The balance between the activator and repressor form of CREB should therefore depend on the information content gained by multiple learning trials, rather than their mere accumulation, a proposal which needs to be tested.

Structural changes in the connectivity between neurons have been proposed as the substrates for LTM in vertebrates and invertebrates (Bailey et al. 1996) and are believed to be at least one target of the interference effects of protein synthesis inhibitors and memory blockers (for a review see Milner et al. 1998). In bees, direct evidence for LTM-related structural changes is lacking, but measurements of mushroom body subcompartment volume indicated that more experienced bees have bigger volumes and more elaborate dendrite branches (Durst et al. 1994, Fahrbach et al. 1995).

The biological circumstances of two forms of LTM may be related to the distinction between those forms of learning which usually lead to lifelong memories (e.g., visual and olfactory cues characterizing the home colony) and those which are stable but need updating on a regular basis (e.g., visual and olfactory cues of feeding places).

The bee, a small animal with a brain of merely $1\,mm^3$ and a total of 950,000 neurons, establishes multiple and distributed memory traces not very different from mammalian memory in the general temporal dynamics, characteristics of contents, and cellular substrates. It may thus serve as a suitable model for the study of memory structure and formation.

See also: Genes and Behavior: Animal Models; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Memory: Genetic Approaches; Memory in the Fly, Genetics of; Memory: Synaptic Mechanisms

Bibliograhy

Bailey C H, Bartsch D, Kandel E R 1996 Toward a molecular definition of long-term memory storage. *Proceedings of the National Academy of Sciences USA* **93**: 13445–52

Durst C, Eichmüller S, Menzel R 1994 Development and experience lead to increased volume of subcompartments of the honeybee mushroom body. *Behavioral and Neural Biology* **62**: 259–63

Ebbinghaus H 1885 *Über das Gedächtnis*. Duncker and Humblot, Leipzig, Germany

Faber T, Joerges J, Menzel R 1999 Associative learning modifies neural representations of odors in the insect brain. *Nature Neuroscience* **2**: 74–8

Fahrbach S E, Giray T, Robinson G E 1995 Volume changes in the mushroom bodies of adult honey bee queens. *Neurobiology of Learning and Memory* **63**: 181–91

Hammer M 1993 An identified neuron mediates the unconditioned stimulus in associative olfactory learning in honeybees. *Nature* **366**: 59–63

Hammer M 1997 The neural basis of associative reward learning in honeybees. *Trends in Neurosciences* **20**: 245–52

Hammer M, Menzel R 1995 Learning and memory in the honeybee. *Journal of Neuroscience* **15**: 1617–30
James W 1890 *The Principles of Psychology*. L. H. Holt, New York
Mauelshagen J 1993 Neural correlates of olfactory learning: Paradigms in an identified neuron in the honey bee brain. *Journal of Neurophysiology* **69**: 609–25
Menzel R 1999 Memory dynamics in the honeybee. *Journal of Comparative Physiology A* **185**: 323–40
Menzel R, Müller U 1996 Learning and memory in honeybees: From behavior to neural substrates. *Annual Review of Neuroscience* **19**: 379–404
Milner B, Squire L R, Kandel E R 1998 Cognitive neuroscience and the study of memory. *Neuron* **20**: 445–68
Müller G E, Pilzecker A 1900 Experimentelle Beiträge zur Lehre vom Gedächtnis. *Zeitschrift für Psychologie* **1**: 1–288
Müller U 1996 Inhibition of nitric oxide synthase impairs a distinct form of long-term memory in the honeybee, *Apis mellifera*. *Neuron* **16**: 541–9
Yin J C P, Del Vecchio M, Zhou H, Tully T 1995a CREB as a memory modulator: Induced expression of a *dCREB2* activator isoform enhances long-term memory in *Drosophila*. *Cell* **81**: 107–15
Yin J C P, Wallach J S, Del Vecchio M, Wilder E L, Zhou H, Quinn W G, Tully T 1995b Induction of a dominant negative CREB transgene specifically blocks long-term memory in Drosophila. *Cell* **79**: 49–58

R. Menzel

Memory in the Fly, Genetics of

1. The Genetic Approach to Learning and Memory

The genetic approach to understand the molecular processes mediating learning and memory was founded upon the principles of genetics and molecular biology that were discovered in the first half of the twentieth century. Studies of mutant organisms with physical defects, along with knowledge of the structure of DNA, the hereditary material, led to the realization that a mutation in a single nucleotide of DNA offered the ultimate way of performing a biological dissection. In other words, a mutation that inactivates a single gene of an animal offers the biologist a way of studying the biological consequences of removing but a single building block from the animal. This genetic approach has been used to study many different questions in biology. Its application to the study of learning and memory and other behaviors can be traced to the laboratory of Seymour Benzer, beginning around 1970 (Benzer 1971).

This approach for dissecting behavior is reductionist in the sense that it reduces the problem to the molecular level. It begins with the idea that a mutation in an animal that produces a learning or memory deficiency identifies a gene and its protein product required for normal learning in nonmutant animals. With the identification of a sufficient number of genes and proteins that are involved, some mechanistic understanding of the molecular dynamics underlying the process can be gained. While this is true, it is now accepted that a genetic connection between gene and behavior is insufficient to gain the necessary depth of understanding. This is because the behavior of an animal emerges from complex interactions at levels other than the molecular. Molecules mediate learning through their biological functions and interactions but this occurs within the context of certain neurons. These neurons, in turn, can be part of complex neural networks that convey information or are involved in behavioral output. Thus, research in the late twentieth century revealed that a multilevel analysis of learning and memory is essential for a complete understanding of the process. In other words, it is necessary to understand learning and memory at the genetic level, the cellular level, the neuroanatomical level, and the behavioral level, since behavior emerges from biological processes at all of these functional levels. Thus, scientists now use the genetic method along with other approaches to gain a greater appreciation of the fundamental mechanisms underlying learning and memory.

2. Learning in the Fly

Adult *Drosophila* are able to learn many different types of information (Davis 1996). After walking into a chamber in which an odor cue is paired with an aversive stimulus of mild electric shock, the animals will tend to avoid the odor for many hours, indicating that they have learned this association (Quinn et al. 1974). They can also learn associations between the color of light and an aversive stimulus, or odors presented with a food reward. They can learn to avoid walking into one side of a small chamber after being punished by heat upon entering that side (Wustmann and Heisenberg 1997), and to avoid flying in the direction of a particular visual cue if that flight direction is punished by mild heat.

The characteristics of *Drosophila* learning are not unique, but reflect the learning principles established with other animals including mammals. For example, the memory of odors can be disrupted by anesthesia shortly after learning, like the amnestic effects produced by anesthesia, electroconvulsive shock therapy, or protein synthesis inhibitors when administered to other species shortly after learning (Squire 1987). Learning an association between an odor and an aversive stimulus requires that they be presented at the same time, reflecting the rule for simultaneous presentation of cues, or pairing, found with many forms of learning in vertebrates. Furthermore, giving *Droso-*

phila repeated learning trials that are separated in time (spaced training) is more effective than presenting the trials with no spacing (Tully et al. 1994). This greater effect of spaced training is also a principle of learning for other animals.

3. *Genes Involved in* Drosophila *Learning*

Although all of the aforementioned types of *Drosophila* learning could, in principle, be the focus of genetic studies, most research has been concentrated on odor learning. Many different mutants that disrupt odor learning have been isolated and the responsible genes have been cloned to identify the protein product of the gene. The best studied of these are listed in Table 1. Most of these mutants have pointed to the fact that intracellular signaling through the small molecule, cyclic AMP, is critical for *Drosophila* odor learning. The normal function of the *dunce* gene is required for normal odor learning and its product, as revealed by molecular cloning and expression, is the enzyme cyclic AMP phosphodiesterase (Chen et al. 1986). This enzyme removes excess cyclic AMP from the cell and in its absence, cyclic AMP levels become significantly elevated. It is thought that the high level of cyclic AMP found in *dunce* mutants prohibits the dynamic changes in cyclic AMP levels that occur during normal learning, or cause the desensitization of other signaling components (Fig. 1), producing a learning deficiency. The product of the *rutabaga* gene performs the opposite role. Its product, adenylyl cyclase, is responsible for synthesizing cyclic AMP in cells (Levin et al. 1992). Another protein that works to help adenylyl cyclase perform its function is neurofibromin (Guo et al. 2000). Cyclic AMP has its effects by activating a protein kinase known as protein kinase A. Mutations in the gene that codes for protein kinase A also disrupt odor learning (Drain et al. 1991, Skoulakis et al. 1993). Protein kinase A, in turn, is known to phosphorylate and activate numerous proteins. Some of these phosphorylations result in the activation of proteins involved in short-term memory. One protein, however, known as CREB, is activated by protein kinase A and is required specifically for long-term memory. Inactivation of the CREB gene in *Drosophila* blocks the formation of long-term, but not short-term, odor memories (Yin et al. 1994). The role of CREB is that of a transcription factor, that is, to turn on or off the expression of other genes. Thus, CREB functions in long-term memory by regulating the expression of other genes.

Although cyclic AMP signaling has emerged as a dominant theme for *Drosophila* odor learning, other genes that may be part of other signaling pathways and cell adhesion proteins have also been found to be essential (Table 1). The *amnesiac* (*amn*) gene codes for a neuropeptide with similarities to the vertebrate neuropeptide, pituitary adenylyl cyclase activating peptide (PACAP) (Feany and Quinn 1995). It is also likely involved in the cyclic AMP signaling system (Fig. 1) but its relationship to other components of the pathway is not yet established. The *leonardo* (*leo*) gene encodes a protein known as 14-3-3, which can function in the activation of other types of protein kinase, including the *raf* protein kinase and protein kinase C (Skoulakis and Davis 1996). Cell adhesion molecules of the integrin family have been implicated in short-term odor memory through the discovery of the *Volado* (*Vol*) gene, a gene required for flies to establish short-term odor memory and one that codes for an α-integrin (Grotewiel et al. 1998). These cell adhesion molecules function in a dynamic way, in that they are modulated by intracellular signaling pathways to rapidly form and break contacts with neighboring cells. It is currently thought that integrins work at the synapse to alter the adhesion of components at the synapse and to alter intercellular signaling.

Table 1
Genes mediating odor learning in *Drosophila*

Gene name	Gene product	Expression in brain	References
dunce (*dnc*)	cyclic AMP phosphodiesterase	mushroom bodies	Chen et al. 1986 Nighorn et al. 1991
rutabaga (*rut*)	adenylyl cyclase	mushroom bodies	Levin et al. 1992
DCO	protein kinase A	mushroom bodies	Drain et al. 1992 Skoulakis et al. 1993
CREB	CREB		Yin et al. 1994
NF1	neurofibromin		Guo et al. 2000
leonardo (*leo*)	14-3-3	mushroom bodies	Skoulakis et al. 1996
amnesiac (*amn*)	PACAP-like neuropeptide		Feany and Quinn 1995
Volado (*Vol*)	α-integrin	mushroom bodies	Grotewiel et al. 1998

Notes: The best-characterized genes mediating odor learning in *Drosophila* along with selected references. The gene name, its symbol, and protein product are listed in the first two columns. The genes listed, except for *CREB*, *NF1*, and *amnesiac*, are expressed at relatively high levels in mushroom body neurons. See also Fig. 1.

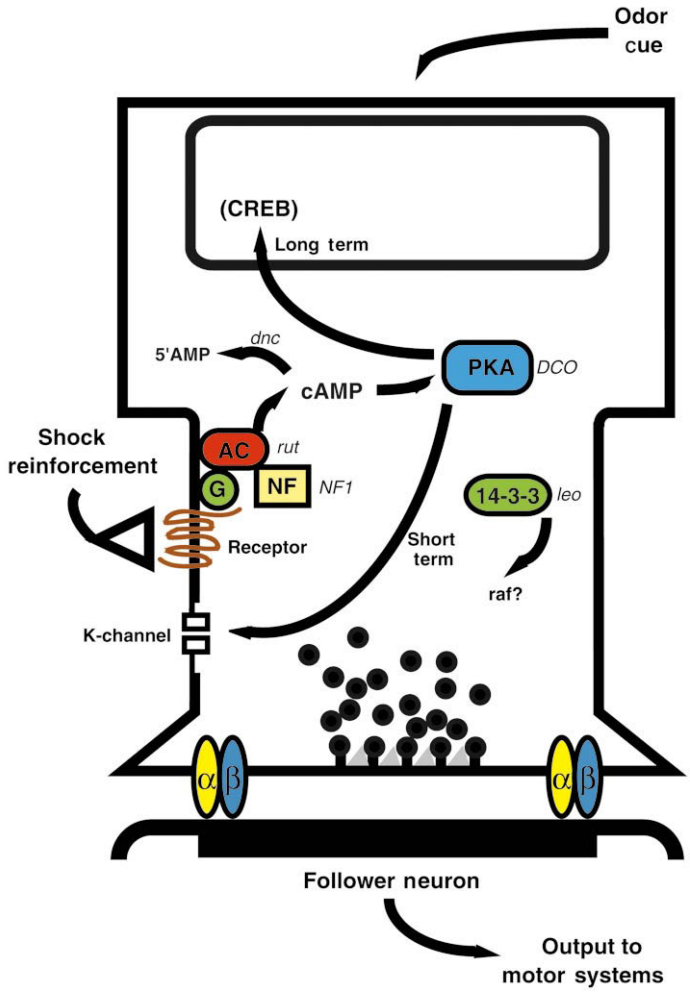

Figure 1
The figure depicts the current cellular and molecular model for olfactory classical conditioning in *Drosophila*. Mushroom body neurons receive information about the environment through neural circuits, including the type and concentration of odors experienced, along with information about whether the animal experiences an aversive stimulus (shock reinforcement). This environmental information is integrated by the mushroom bodies through the actions of genes and molecules identified to participate in learning. The odor cue simply activates the mushroom body neuron and the shock reinforcement acts as a modulator of this activation. The modulation occurs by the activation of adenylyl cyclase which elevates internal cyclic AMP levels. Several different proteins are involved in elevating cyclic AMP, including the enzyme adenylyl cyclase which is the product of the *rutabaga* (*rut*) gene, a G-protein (G) that couples the adenylyl cyclase to a neurotransmitter receptor (Receptor), and neurofibromin (NF, encoded by the *NF1* gene), which functions to help in the activation of the AC. The product of the *dunce* (*dnc*) locus, cyclic AMP phosphodiesterase, controls the dynamic changes in cyclic AMP levels. Cyclic AMP activates protein kinase A (from the *DCO* gene) which phosphorylates numerous proteins including potassium channels (K-channel) and CREB, a transcription factor in the nucleus required for long-term memory formation. Other proteins known to be involved include a 14-3-3 protein (from the *leonardo* gene) which may be involved in the activation of other types of protein kinases (*raf*, for example) and integrin proteins. The *Volado* (*Vol*) gene encodes the α-subunit of an integrin heterodimer, and is depicted at the synapse since integrins are cell surface proteins involved in the dynamic adhesion between cells.

4. Brain Neurons Mediating Insect Learning

A major refocus of the *Drosophila* learning field began in 1991 with the discovery of the neurons clearly required for odor learning (Nighorn et al. 1991). There are about 5,000 mushroom body neurons in *Drosophila* and these neurons are similar to neurons in the primary olfactory cortex of the human brain. The primary olfactory cortex is known to be important for odor learning in vertebrate species. The fact that mushroom body neurons are largely responsible for odor learning by insects is supported by many different types of research, among which is the discovery that many of the genes required for learning are highly expressed in mushroom body neurons relative to other neurons (Davis 1993). The product of *dunce* is highly expressed in mushroom body neurons as are protein kinase A and the products of *rutabaga*, *leonardo*, and *Volado* (Table 1). It is not yet known whether CREB, neurofibromin, or *amnesic* is highly expressed in these neurons. These observations have led to a cellular model for *Drosophila* odor learning (Nighorn et al. 1991, Fig. 1). This model envisions mushroom body neurons as integrators of the information presented during training, that being specific odors and electric shock. This integration changes the physiology of mushroom body neurons using the cyclic AMP signaling system such that they activate neural circuits for avoidance behavior after learning.

Drosophila is a powerful biological system for the discovery of genes involved in learning and memory, and for elucidating principles for the molecular events underlying learning. It is of additional interest that many of the genes identified as participating in *Drosophila* learning have been implicated in learning in other species (Davis 2000). And the field blossomed in the 1990s with the discovery of the neurons that mediate odor learning. Nevertheless, the physiological changes that occur within mushroom body neurons during odor learning remain speculative and the subject of models (Fig. 1). To use genetics to make that final link between changes in cellular physiology and learning remains a challenge for the future.

See also: Learning and Memory, Neural Basis of; Memory: Genetic Approaches; Memory in the Bee; Memory: Synaptic Mechanisms; Protein Synthesis and Memory

Bibliography

Benzer S 1971 From the gene to behavior. *Journal of the American Medical Association* **281**: 24–37

Chen C-N, Denome S, Davis R L 1986 Molecular analysis of cDNA clones and the corresponding genomic coding sequences of the *Drosophila dunce*$^+$ gene, the structural gene for cAMP phosphodiesterase. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 9313–17

Davis R L 1993 Mushroom bodies and *Drosophila* learning. *Neuron* **11**: 1–4

Davis R L 1996 Biochemistry and physiology of *Drosophila* learning mutants. *Physiological Reviews* **76**: 299–317

Davis R L 2000 Neurofibromin progress in the fly. *Nature News & Views* **403**: 846–47

Drain P, Folkers E, Quinn W G 1991 cAMP-dependent protein kinase and the disruption of learning in transgenic flies. *Neuron* **6**: 71–72

Feany M, Quinn W 1995 A neuropeptide gene defined by the *Drosophila* memory mutant *amnesiac*. *Science* **268**: 825–26

Grotewiel M S, Beck C D O, Wu K-H, Zhu X-R, Davis R L 1998 Integrin-mediated short-term memory in *Drosophila*. *Nature* **391**: 455–60

Guo H-F, Tong J, Hannan F, Luo L, Zhong Y 2000 A neurofibromastosis-1-regulated pathway is required for learning in *Drosophila*. *Nature* **403**: 895–98

Levin L, Han P-L, Hwang P M, Feinstein P G, Davis R L, Reed R R 1992 The *Drosophila* learning and memory gene *rutabaga* encodes a Ca^{2+}/calmodulin-responsive adenylyl cyclase. *Cell* **68**: 479–89

Nighorn A, Healy M, Davis R L 1991 The cAMP phosphodiesterase encoded by the *Drosophila dunce* gene is concentrated in mushroom body neuropil. *Neuron* **6**: 455–67

Quinn W G, Harris W A, Benzer S 1974 Conditioned behavior in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America* **71**: 708–12

Skoulakis E M C, Davis R L 1996 Olfactory learning deficits in mutants for *Leonardo*, a *Drosophila* gene encoding a 14-3-3 protein. *Neuron* **17**: 931–44

Skoulakis E M C, Kalderon D, Davis R L 1993 Preferential expression of the catalytic subunit of PKA in the mushroom bodies and its role in learning and memory. *Neuron* **11**: 197–208

Squire L R 1987 *Memory and Brain*. Oxford University Press, New York

Tully T, Preat T, Boynton S C, Del Vecchio M 1994 Genetic dissection of consolidated memory in *Drosophila*. *Cell* **79**: 35–47

Wustmann G, Heisenberg M 1997 Behavioral manipulation of retrieval in a spatial memory task for *Drosophila melanogaster*. *Learning & Memory* **4**: 328–36

Yin J C P, Wallach J S, Del Vecchio M, Wilder E L, Zhou H, Quinn W G, Tully T 1994 Induction of a dominant negative CREB transgene specifically blocks long-term memory in *Drosophila*. *Cell* **79**: 49–58

R. L. Davis

Memory: Levels of Processing

In the 1960s, theories of human memory were dominated by the notion of memory stores and the transfer of encoded information from one store to another. The associated experimental work was designed to elucidate various features of the stores; for example, their coding characteristics, their capacities, and their forgetting functions. Craik and Lockhart (1972) critic-

ized the stores concept and suggested instead that human memory could be understood in terms of the qualitative type of processing carried out on the material to be learned and later remembered. However, proponents of the memory stores view (e.g., Atkinson and Shiffrin 1971) argued for sensory stores, a short-term buffer, and a long-term store. Craik and Lockhart proposed that the qualitative differences between remembered events, and their different retention characteristics could be described in terms of different mental processes, as opposed to a variety of structures. The term 'levels of processing' was coined to capture the idea that incoming stimuli were processed first in terms of their sensory features and then progressively in terms of their meanings and implications. The further suggestion was that 'deeply' encoded stimuli (that is, those that were fully analyzed for meaning) were also the ones that would be remembered best in a later test. The purpose of the present article is to lay out these ideas and arguments more fully along with a review of the evidence that supports them. Critical points of view will also be described and discussed.

1. *Basic Ideas*

Craik and Lockhart's (1972) main objection to the memory stores perspective was that the defining characteristics of the various stores did not appear to be constant from one situation to another. For example, both the capacity and the rate of forgetting associated with the short-term store varied as a function of the meaningfulness of the material held in the store. As an alternative formulation, Craik and Lockhart proposed that the primary functions of the cognitive system are the perception and understanding of incoming material, and that the formation of the memory trace is an incidental by-product of these primary processing operations. In this formulation there is no special 'faculty' of memory, and no memory stores as such; memory is simply a function of the processing carried out on perceived stimuli—for whatever reason.

The further suggestion was that deeper levels of processing were associated with longer lasting memory traces. This notion has its roots in the idea that the input side of the cognitive system is organized hierarchically, with early sensory analyses gradually developing into analyses of meaning, association, and implication. Specifically, Craik and Lockhart based their levels of processing (LOP) view of memory on Anne Treisman's (1964) levels of analysis formulation of the processes of selective attention. Treisman proposed that incoming stimuli are subjected to a series of 'tests,' organized hierarchically, with early tests being concerned with analysis of sensory features and later tests dealing with identification of words and objects. All incoming stimuli pass the early sensory tests, implying that sensory information is analyzed and therefore available for later memory, but only a progressively smaller proportion of stimuli penetrate through the successive tests to a full analysis of identification and meaning (the mechanism of selective attention). Craik and Lockhart capitalized on this general set of ideas and added the further notion that deeper (i.e., more meaningful) levels of analysis were associated with semantically richer and more durable memory traces. One of Craik and Lockhart's main points was therefore that the processes of memory and attention are intimately interlinked. Indeed, in their formulation the processes of perceiving, attending, understanding, and remembering are all aspects of the overall cognitive system. Memory encoding processes are simply those processes carried out essentially for the purposes of perception and comprehension; memory retrieval processes may be thought of as a reinstatement or recapitulation of some substantial proportion of the processes that occurred during encoding.

Two further points discussed by Craik and Lockhart (1972) should be mentioned. The first is the distinction between two types of rehearsal: one type functions by continuing to process encoded material at the same level of analysis, whereas the second involves operations that enrich and elaborate the material by carrying processing to deeper levels. In the original paper Craik and Lockhart referred to these two functions by the somewhat uninspired names of Type I and Type II rehearsal, but preferable terms are 'maintenance' and 'elaborative rehearsal.' If later memory is simply a function of the deepest level of analysis obtained, then memory performance should increase as a function of greater amounts of elaborative processing, but should be independent of the amount of maintenance processing. This prediction was borne out in the case of subsequent recall (Craik and Watkins 1973, Woodward et al. 1973) but, interestingly, not for recognition. In this latter case, greater amounts of maintenance rehearsal are associated with increased levels of recognition memory (Woodward et al. 1973). Apparently recognition, but not recall, is sensitive to some strengthening aspect of maintenance processing.

The second point from the original paper is that the distinction between 'short-term' and 'long-term' memory was maintained, but not in the form of separate memory stores. For Craik and Lockhart, short-term or 'primary memory' was synonymous with active processing of some qualitative aspect of a stimulus or small set of stimuli. Material 'held in primary memory' was thus held to be equivalent to attention paid to that material. Given that attentional resources can be deployed flexibly to a large variety of different processes, this formulation solves the problem of why material held 'in the short-term store' can be of many different qualitative types, including phonemic, lexical, semantic, and even imaginal information. Primary memory was thus seen as a large set of possible

processing activities, rather than as the single structural store envisaged in the Atkinson and Shiffrin (1971) model.

2. Empirical Evidence

If memory is a function of the deepest level of processing obtained, it should not matter how that level of analysis is produced; for example, intention to learn or memorize the material should be irrelevant. This thought led to a series of experiments by Craik and Tulving (1975) in which words were processed to various depths by preceding each word by a question. These types of question (or 'orienting tasks') were designed so that the following word need only be processed to a shallow level (e.g., 'is the word printed in upper case letters?'), to an intermediate level (e.g., 'does the word rhyme with train?'), or to a relatively deep semantic level (e.g., 'is the word a type of animal?'). In a typical experiment, 60 concrete nouns were each preceded by one question of this type, 20 concerning case, 20 rhyme, and 20 semantic, and with half of the questions associated with a 'yes' answer and half with a 'no' answer (e.g., the word TABLE preceded by 'does the word rhyme with fable?' or by 'does the word rhyme with stopper?'). The encoding phase was then followed either by a recall test for the 60 words or by a recognition test in which the 60 target words were mixed randomly with 120 new words of a similar type.

Table 1 shows the results from two experiments in the Craik and Tulving series. For both recall and recognition memory performance was a function of depth of processing, but unexpectedly 'yes' answers in the initial encoding phase gave rise to higher levels of performance than did 'no' answers, for rhyme and semantic questions at least. Craik and Tulving suggested that this latter result reflected the greater degrees of elaboration associated with 'yes' answers. For example, the word TIGER would be elaborated more following the question 'is the word a jungle animal?' than following 'is the word a type of furniture?' That is, the compatible question serves to specify and perhaps enrich the encoding of the target word to a greater degree.

Table 1
Proportions of words recalled and recognized as a function of depth of processing

Response type	Case	Rhyme	Semantic
Words recalled (Experiment 3)			
Yes	0.14	0.30	0.68
No	0.05	0.15	0.40
Words recognized (Experiment 9)			
Yes	0.23	0.59	0.81
No	0.28	0.33	0.62

Source: Craik and Tulving 1975

Further experiments in the Craik and Tulving (1975) paper showed that processing time had little effect on subsequent memory performance; type (or depth) of processing was much more important. Studying the words under intentional or incidental learning conditions also made little difference to the pattern of results. Finally, motivation differences appeared to be unimportant in such experiments, as varying the reward associated with specific words had no differential effect on performance.

3. Criticisms and Rebuttals

The LOP notions attracted a lot of attention in the 1970s, probably because the process-oriented perspective had been in the back of many researchers' minds at that time. The ideas also drew criticisms, however, and excellent critical reviews were published by Nelson (1977) and Baddeley (1978); replies to these critical points were made by Lockhart and Craik (1990).

The major criticism was that the LOP ideas were extremely vague, and difficult to disprove; the ideas did not constitute a predictive theory. A linked criticism concerned the circularity of the concept of 'depth.' Given that no independent index of depth of processing had been proposed, it seemed all too easy to claim that any event that was well remembered must have been processed deeply. One answer to this point is that the LOP ideas were always intended to provide a framework for memory research, rather than a tight predictive theory. Thinking of memory in terms of mental processes that vary in terms of the qualitative types of information they represent, suggests different concepts and different experiments than those suggested by a structural viewpoint. The absence of an independent index of depth is certainly a drawback, although judges show good agreement when asked to rank the relative depths of a set of orienting tasks. One possibility is that neuroimaging or neurophysiological techniques may provide such an index (e.g., Kapur et al. 1994, Vincent et al. 1996).

Baddeley (1978) cited evidence to show that pictures appeared to be relatively insensitive to LOP manipulations, and suggested that the LOP ideas may be restricted to verbal materials. However, another way to interpret this result is that pictures are simply very compatible with our cognitive analyzing processes, and are therefore processed deeply and meaningfully regardless of the ostensible orienting task.

Some criticisms seemed well founded, and the LOP framework was modified accordingly (Lockhart and Craik 1990). For example, Craik and Lockhart's original formulation had implied that incoming stimuli were analyzed in a constant linear fashion from

shallow to deep processing. This seems unnecessarily restrictive, and a more realistic account would allow for interactive processing throughout the cognitive system, but with the resulting memory trace reflecting those processing operations that were carried out, regardless of the order in which they were achieved. A second point concerns the durability of shallow, sensory processing operations. The 1972 paper asserted that the results of such analyses were quite transient, in line with observations that 'sensory memory traces' decayed within a matter of seconds. However, subsequent work using paradigms of implicit or procedural memory showed that shallow processing operations can have extremely long-lasting effects when tested in sensitive ways (e.g., Kolers 1979). Apparently the transient effects of sensory analyses are associated with explicit memory and conscious recollection.

A third critical point requiring acknowledgment stems from observations of amnesic patients. Such people can certainly process information deeply in the sense that they can comprehend incoming material and make appropriate meaningful responses, yet they have little or no recollection of the material at a later time. It therefore seems that deep processing is a necessary but not sufficient correlate of good memory performance. Some further set of operations (associated perhaps with neurophysiological processes taking place in the hippocampus and medial temporal lobes of the cerebral cortex) are also necessary, and it is these latter operations that are impaired in amnesic patients.

A final criticism modifies the point that deep semantic processing is inevitably best for subsequent memory. Several theorists have made the point that there is no such thing as a universally 'good' encoding condition; rather, a specific encoding condition is good to the extent that it is compatible with the cues available in the later retrieval environment. This idea is captured in the notions of transfer-appropriate processing, repetition of operations, and the encoding specificity principle. As one illustration, Morris et al. (1977) showed that when the retrieval test involves rhyming cues, rhyme encoding operations were associated with higher levels of recognition than were semantic encoding operations. It should also be noted, however, that the combination of semantic encoding and semantic retrieval was associated with a substantially higher level of recognition than the combination of rhyme encoding and rhyme retrieval (0.68 vs. 0.40 averaged over Experiments 1 and 2). It therefore seems that any complete account of the psychology of memory must involve principles that capture both notions of depth of processing and transfer-appropriate processing.

This suggestion is illustrated by an experiment by Fisher and Craik (1977). They had subjects encode words either by rhyme or semantically, and then provided retrieval cues that were either identical, similar, or different from the way the word was originally encoded. For example, if the word DOG was encoded by rhyme (rhymes with log—DOG), the respective retrieval cues would be 'rhymes with log—,' 'rhymes with frog—,' 'associated with cat—.' Table 2 shows that both LOP and transfer-appropriate processing affect the pattern of results. That is, semantic encoding is generally superior to rhyme encoding and the more similar the encoding and retrieval cues, the better is performance. Also, the two manipulations interact in the sense that the benefits of deeper semantic processing are greater with compatible cues (or alternatively the effects of encoding-retrieval compatibility are greater at deeper levels of processing).

Table 2
Proportions of words recalled as a function of encoding context and similarity between encoding context and retrieval cue

Encoding/retrieval similarity	Encoding context		
	Rhyme	Semantic	Mean
Identical	0.24	0.54	0.39
Similar	0.18	0.36	0.27
Different	0.16	0.22	0.19
Mean	0.19	0.37	

Source: Fisher and Craik 1977

4. *Further Developments*

Since the 1970s the LOP ideas have been used in a wide variety of theoretical and applied contexts, from social cognition and personality theory to cognitive neuroscience (Lockhart and Craik 1990). It has been shown, for example, that deep semantic processing is reliably associated with activation of the left prefrontal cortex (Kapur et al. 1994). It has also been shown that associating words with the self results in particularly high levels of recollection, arguably because a person's 'self-schema' is richly detailed and meaningful.

The central observation that deep processing results in good memory, is undeniable. The theoretical reason for this finding is less certain, but a reasonable proposal is that deeper levels of processing result in a record that is distinctive from other encoded events, and therefore more discriminable at the time of retrieval. A further reason is that the interconnected schematic structure of deeper representations facilitates the processes of reconstruction at the time of retrieval (Lockhart and Craik 1990). In any event it appears that the concept of depth of processing, or some similar concept, is a necessary one for any complete account of human memory processes.

See also: Attention: Models; Attention, Neural Basis of; Elaboration in Memory; Encoding Specificity in Memory; Implicit Memory, Cognitive Psychology of;

Memory: Organization and Recall; Memory Retrieval; Short-term Memory, Cognitive Psychology of; Working Memory, Psychology of

Bibliography

Atkinson R C, Shiffrin R M 1971 The control of short-term memory. *Scientific American* **225**: 82–90

Baddeley A D 1978 The trouble with levels: A re-examination of Craik and Lockhart's framework for memory research. *Psychological Review* **85**: 139–52

Cermak L S, Craik F I M 1979 *Levels of Processing in Human Memory*. Erlbaum, Hillsdale, NJ

Challis B H, Velichkovsky B M 1999 *Stratification in Cognition and Consciousness*. John Benjamins, Amsterdam

Craik F I M, Lockhart R S 1972 Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* **11**: 671–84

Craik F I M, Tulving E 1975 Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General* **104**: 268–94

Craik F I M, Watkins M J 1973 The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior* **12**: 599–607

Fisher R P, Craik F I M 1977 Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory* **3**: 701–11

Kapur S, Craik F I M, Tulving E, Wilson A A, Houle S, Brown G 1994 Neuroanatomical correlates of encoding in episodic memory: Levels of processing effect. *Proceedings of the National Academy of Sciences of the USA* **91**: 2008–11

Kolers P A 1979 A pattern-analyzing basis of recognition. In: Cermak L S, Craik F I M (eds.) *Levels of Processing in Human Memory*. Erlbaum, Hillsdale, NJ, pp. 363–84

Lockhart R S, Craik F I M 1990 Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology* **44**: 87–112

Morris C D, Bransford J D, Franks J J 1977 Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior* **16**: 519–33

Nelson T O 1977 Repetition and levels of processing. *Journal of Verbal Learning and Verbal Behavior* **16**: 151–77

Treisman A 1964 Selective attention in man. *British Medical Bulletin* **20**: 12–16

Vincent A, Craik F I M, Furedy J J 1996 Relations among memory performance, mental workload and cardiovascular responses. *International Journal of Psychophysiology* **23**: 181–98

Woodward A E, Bjork R A, Jongeward R M 1973 Recall and recognition as a function of primary rehearsal. *Journal of Verbal Learning and Verbal Behavior* **12**: 608–17

F. I. M. Craik

Memory Models: Quantitative

Quantitative models of human memory describe the storage and retrieval of three types of information: item information, associative information, and serial-order information. Item information deals with familiarity: how we can recognize objects, events, and words as old. Associative information deals with the unification or binding of two items: names and faces, sights and sounds, labels and referents, words and their meaning. Serial-order information deals with the temporal binding of sequential events: the days of the week, the letters of the alphabet, the months of the year, or how to spell words. Storage subsumes the initial encoding and its persistence over time (though these are clearly separate issues). Retrieval deals with performance at the time of test: recall, recognition, and sometimes reconstruction or rearrangement. Recall requires producing or generating the target item, while recognition only requires responding (e.g., 'familiar' or 'unfamiliar') to the presented item.

There are many quantitative models of memory (perhaps several dozen); rather than describing them one-by-one I shall describe a number of ways in which they differ, and then describe a selected few in more detail.

1. Dimensions

1.1 Item Representation

Items are usually represented as vectors of attributes or features. These features can be discrete (generally binary or ternary) or continuous. If binary, they are usually 0, 1 but sometimes -1, 1; if ternary -1, 0, 1, where 0 represents uncertainty or loss. If continuous, they are usually considered as N random samples from a normal distribution, where N is the dimensionally of the vector. The feature distribution is generally assumed to have mean zero and, perhaps, variance $1/N$. If the variance is $1/N$, then we start with unit vectors; they are statistically normalized to 1.0.

While one can map between discrete and continuous vectors, the choice has implications for the measurement of similarity. For discrete vectors, similarity is generally measured by Hamming distance, the number of mismatching features. For continuous vectors, one can compute the similarity of two item vectors by the dot product or the cosine of the angle between them, and one can also set up a prototype and generate a set of exemplars (instances or examples) with any desired similarity. One can also normalize these exemplars so we still work with unit vectors.

1.2 Associations

There are three basic ways to represent associations. One is by partitioned vectors, another is by the outer product, and the third is by convolution. With partitioned vectors, if we have items A and B to associate, the A item is stored in the left-hand side of the vector and the B item is stored in the right-hand

side of the vector. With the outer product, every feature or element of A is multiplied by every element of B and the result is stored in an $A \otimes B$ outer-product matrix. With convolution, the two items are convolved the way any two discrete probability density functions would be convolved (i.e., outcomes sum but probabilities multiply), and the association is the resulting $A * B$ convolution.

1.3 Serial Order

Serial order can be represented by item-to-item associations (chaining), item-to-position associations (time tags or bins), or multiple convolutions. In a chaining model, each item is associated both with its predecessor and its successor, much like the links in a chain, and recalling a string is like traversing a chain link by link. In a bin model, items are associated to positions (temporal or spatial), and to retrieve a string the position of each item is retrieved and then the item is recalled in that position. With multiple convolutions, each item is convolved (associated) with all prior items and recall proceeds from first to last.

1.4 Storage

There are basically two options here: localized storage or superposition. With localized storage, each item or association is stored in a separate bin or array, much like the storage registers of a computer. With superposition, the items or associations are summed and stored in a single memory array; this is an essential characteristic of distributed memory models. There are other logical possibilities for superposition (e.g., items and associations could be stored separately) but, to my knowledge, this possibility has not been explored.

The storage choice has implications for both binding and retrieval. With superposition, some non-linear transform is necessary to effect binding. That is, with pairs A–B and C–D there must be some way of knowing whether A goes with B or D, and likewise for C. For retrieval, localized storage implies a search process (as in a computer), whereas superposition implies direct access or a content-addressable memory. Many years ago, von Neumann (1958) gave compelling arguments why the brain was not like a computer, but his arguments have often been ignored.

1.5 Retrieval

Both recognition and recall are generally assumed to be two-stage processes. For recognition, the two stages are memory and decision; for recall, they are generation and deblurring or cleanup. For recognition, the probe item operates on the stored information in memory and the result is fed into a decision system. The comparison process is the dot (or inner) product or a counting process which tallies the number of matching features. The decision system is generally assumed to be a threshold function where continuous probability–density functions are mapped into a binary (yes/no) function (above or below a criterion).

For recall, the result of the generation process generally will be an imperfect replica of the target item because of deterioration of the stored memory trace over time or interaction with other memory traces. Consequently, a clean-up process is necessary to map the retrieved approximation into an exact copy of the target item. The most popular deblurring process is a Hopfield net (Hopfield 1982); however, many models do not specify the process but merely assess the results of some hypothetical process by a winner-take-all principle. That is, the recalled item is assumed to be the best match of the retrieved approximation to all list items.

Some models do not specify the generation process but assume it is a sampling process. Here strengths are mapped into probabilities by the Luce choice axiom (Luce 1959, see *Luce's Choice Axiom*) which says that the probability of sampling a given item is the ratio of its strength to the sum of the strengths of all the items. Such models then may assume a recovery process that implements a sampling-without-replacement algorithm; this is to avoid recalling an item which has already been recalled.

Another approach to the retrieval process is to assume a random walk or diffusion process. The comparison process involves accumulating information (e.g., computing a dot-product sequentially) and there are decision boundaries which terminate the process. With multiple random walks occurring in parallel, a number of items can each have their own random walk. For the upper boundary, the result is the first to arrive, and otherwise there is either a time limitation or all items must reach a lower boundary if negative 'strengths' are envisioned. One advantage of random-walk models is that they can predict both response accuracy and latency, and latency is often used as a dependent variable to supplement accuracy in recognition studies (see *Diffusion and Random Walk Processes*).

1.6 Context

Context is important in memory processes; how we encode the word 'bank' depends upon whether it is accompanied by money and loan, or river and stream. Also, the items used in studies of recognition memory are seldom novel, so these studies are assessing knowledge of list membership (i.e., items presented in the context of the list). Studies of source memory reverse this; given the item(s), what was the context?

Quantitative models of memory generally represent context in the same way as they represent items; namely, as random vectors. Such context vectors may be concatenated or associated with the item vectors. For distributed memory models, there must be a binding operation (e.g., auto-association) without which the context would be free floating. Context is generally assumed to drift over time, from item to item or list to list. Recognition or recall can occur in the presence or absence of context; accuracy (and latency) should be better when context is present (or available) at both study and test.

1.7 Forgetting

The classical issue has been whether forgetting occurs because of decay (atrophy through disuse, or mere passage of time) or interference (degradation of memory traces because of subsequent events). The experimental evidence favors interference over decay although, in some situations, small amounts of decay may occur. In terms of the models, interference may occur because of context change or because of trace degradation. The degradation may be loss of features (e.g., $+1/-1$ features may be set to 0), diminution of trace strength, or (less often) a gradual diffusion of each feature away from some original value.

Loss of trace strength can be modeled by simple first-order linear difference equations. The interest is not so much in solving the difference equations but in treating them as evolution equations so we can trace (and make predictions about) the time course of the process. There are analytic models which can yield some closed-form solutions (often in terms of signal/noise ratios where one derives explicit expressions for means and variances of the comparison or generation process as a function of theoretical parameters) or simulation models where one can get numerical results of the same type. In general, the complexity of both types of model increases with the number of parameters, and simulation can also be a useful check on the accuracy of the derivations for analytic models.

1.8 Repetition

Learning can be considered as the rate of change of memory with respect to the number of presentations, so repetition is an important issue for memory models. Some memory models do not deal with repetition effects, but those that do need a learning rule or principle to specify the effect of repetition. Repetition can either lay down a new trace (multiplex models) or strengthen existing traces. The latter can be open loop or closed loop (independent or dependent on the state of the memory trace at the time of study); either way, various possibilities exist. The new item vector or selective features can be added to the old item vector. Connectionist memory models generally use either back-propagation or a 'Boltzmann machine' modeled after thermodynamics, but these are outside the scope of this article.

2. Early Models

The first models of the sort discussed here were the matched-filter and the linear-associator model of Anderson (1970). The former was a model for the storage and retrieval of item information, and the latter for the storage and retrieval of associative information. In the matched filter model, items were represented as random vectors with elements selected from a normal distribution with mean zero and variance $1/N$ and stored by superposition in a common memory vector. If we denote the ith item in a list of items as $\mathbf{f}_i$ and have a memory vector $\mathbf{M}$ with a forgetting parameter α, where $0 \leqslant \alpha \leqslant 1$, then the storage equation was

$$\mathbf{M}_j = \alpha \mathbf{M}_{j-1} + \mathbf{f}_j, \quad \mathbf{M}_0 = 0. \tag{1}$$

This is the simple linear first-order difference equation mentioned above, and it gives the contents of memory after a list of items has been presented once.

For retrieval, the probe item—call it $\mathbf{f}_k$—is compared with the memory vector by the dot product. If $\mathbf{f}_k$ was an item that had been in the list with $L-1$ other items, then

$$\mathbf{f}_k \cdot \mathbf{M} = \mathbf{f}_k \cdot \sum_i^L \mathbf{f}_i = \mathbf{f}_k \cdot \mathbf{f}_k + \mathbf{f}_k \cdot \sum_{j \neq k} \mathbf{f}_j \cong \alpha^{L-k} \tag{2}$$

because, given the way the item vectors were constructed, if E is the expected value then $E[\mathbf{f}_i \cdot \mathbf{f}_i] \cong 1$ but $E[\mathbf{f}_i \cdot \mathbf{f}_j] \cong 0$, $i \neq j$. Thus, one can 'recognize' an old item, the familiarity is greater the later its presentation in the list, and one can distinguish it from a new item probe which has an expected value of zero. The recognition accuracy and, more particularly, the slope of the retention curve, will be a function of the forgetting parameter α.

It is relatively easy to derive an expression for the old- and new-item variances (they are almost the same; the old-item variance for a single item is about twice that for a new item but weighted by $1/L$) so, assuming that the strength (i.e., dot product) must exceed some criterion, the recognition performance is characterized by the probability density functions shown in Fig. 1. Thus, the matched-filter model provides a well-motivated explanation for the popular application of signal-detection theory to human memory (Murdock 1974—also see *Signal Detection Theory*). However, the

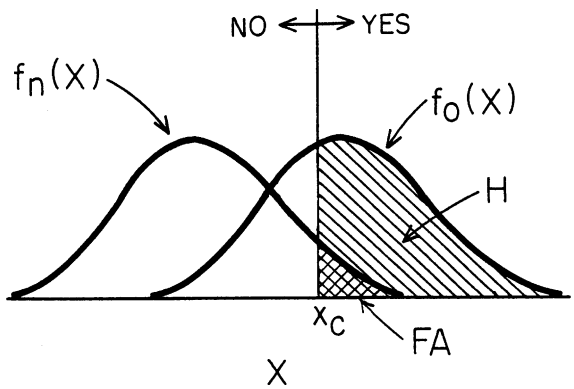

Figure 1
Old $f_0(x)$ and new $f_n(x)$ item distributions with a criterion placed at x_c where x is the decision axis. The shaded area under $f_0(x)$ is the probability of a hit (H: correct response to an old item) while the cross-hatched area under $f_n(x)$ is the probability of a false alarm (FA: calling a new item old). (Fig. 2.8 in Murdock 1974.) (Reproduced by permission of Lawrence Erlbaum Publishers)

matched-filter model is quite limited; it cannot explain associations, there is no representation of context, and there is no learning rule.

The linear-associator model goes a step further and assumes that associations can be represented by the outer product of two item vectors. Although originally intended for autoassociations, thus giving a content-addressable memory, it could certainly be applied to the A–B associative paradigm. If **f** and **g** are two (possibly correlated) item vectors then a comparable storage equation would be

$$\mathbf{M}_j = \alpha \mathbf{M}_{j-1} + \mathbf{f}_j * \mathbf{g}_j \tag{3}$$

where $\mathbf{f}_j$ and $\mathbf{g}_j$ are the jth items, and one could retrieve the associate to any item by vector-matrix multiplication. (Since this is not a commutative operation, one would need pre- or postmultiplication depending upon whether the probe was the B or the A item.)

We still have no way of implementing serial-order memory. Using a slightly different approach (but many of the same ideas) Liepa (1977) used convolution rather than the outer product for the associative operation and extended it to multiple convolutions for serial order. To illustrate with $L = 3$, if we had item vectors **f**, **g**, and **h** then the list might be encoded as

$$\mathbf{M} = \mathbf{f} + \mathbf{f} * \mathbf{g} + \mathbf{f} * \mathbf{g} * \mathbf{h}. \tag{4}$$

If we use δ to denote the Dirac delta vector and # to denote correlation (roughly, $*$ and # are inverse operations) then we can retrieve the items in order because

$$\delta \# \mathbf{M} \cong \mathbf{f},\ \mathbf{f} \# \mathbf{M} \cong \mathbf{g}, \quad \text{and } (\mathbf{f} * \mathbf{g}) \# \mathbf{M} \cong \mathbf{h}. \tag{5}$$

Each retrieval operation generates a fuzzy replica which must be deblurred, but the clean-up operation is not shown here. Liepa's approach was based on the convolution/correlation formalism of Borsellino and Poggio (1973), who showed that this was a viable storage/retrieval algorithm.

These simple schemes show how information can be stored and retrieved in the three cases of interest, but they do not begin to cope with the complexities of the experimental data. We now consider some of the more recent models which attempt to explain these data in more detail.

3. *Current Models*

3.1 *Resonance-retrieval Theory*

This theory (Ratcliff et al. 1999) deals only with recognition, and it has no representation assumptions. Items are characterized by their resonance where resonance values are assumed to be normally distributed with positive means for old items and negative means for new items. When a probe item is presented, each item in the search set (generally the most recently-presented list) resonates and this determines the mean drift rate for the diffusion processes. Each drift rate has a variance as well, and this is an important feature of the model.

There is a variable starting point with upper and lower boundaries; the diffusion processes stop when one item (not necessarily the correct one) hits the upper boundary, or when all items have hit the lower boundary without any hitting the upper boundary. Decision factors affect the starting value and the placement of the boundaries. This model is able to give a very detailed account of accuracy and latency data for both correct and error responses.

3.2 *SAM Model*

This 'search of associative memory' model (Shiffrin and Raaijmakers 1992) is an updated version of the earlier and very influential Atkinson and Shiffrin buffer model (see Izawa 1999). Both models focus on the importance of a short-term store or rehearsal buffer. This has a very limited capacity (generally four items or two pairs) which serves as the antechamber to a much more capacious long-term memory. Although, again, there are no representation

assumptions in SAM, the strength of each item is multidimensional; there is a context strength, a self-strength, and an associative strength (with all the other items currently in the buffer), and background noise.

For recognition, the 'familiarity' of a test item is the sum of the products of the context and self-strengths of all the images in the retrieval matrix, so (like the resonance-retrieval theory) it is a global matching model, or GMM. In GMMs, all the items in the list (or search set) enter into the memory comparison process, and this characteristic seems necessary to account for experimental data. The SAM model handles recall by a search (sampling) and recovery process, although it is free recall (no order requirement), not serial recall. Although there are a number of parameters in the model (a dozen or more) it is probably able to account for a wider range of data than any other single model and, as a consequence, has been very influential.

3.3 MINERVA2

This simulation model (Hintzman 1988) represents items as concatenated vectors of binary $(+1, -1)$ features, and partitioned vectors are used to store the associations between two items in a memory array. Each item or pair of items has flanking context features, and it too is a GMM. For recognition, the probe item or pair is compared with all the items in the memory array, and the resulting Hamming distances are cubed, which produces binding. Thus, even though (like SAM and the resonance-retrieval theory) it is a localist model, again like these other two models all the information in memory enters into the comparison process.

MINERVA2 can also carry out recall. If we assume that pairs had been stored, the probe item is compared to all the items in the memory matrix and a fuzzy reconstruction of the target item is generated. There is also an iterative deblurring algorithm, so this fuzzy reconstruction can often be mapped onto an extra replica of one of the presented items. There is no forgetting rule; some features are randomly set to zero to reduce the accuracy level appropriately. While the model has not yet been applied to serial-order information, not only can it handle much recognition data but also judgments of frequency and judgments of recency. These judgments are usually not considered in other models.

3.4 Dual-trace Array-perturbation Model

The perturbation model (Estes 1997) is a limited model for short-term ordered recall. It is a bin model which assumes there could be place changes between adjacent items, and that this perturbation occurred during the study of new items or the recall of old items. It provided a good explanation for distance functions (position gradients for recalled items). The array model is a more general model of memory derived from instance-based (i.e., localized) models of categorization. In many ways it is like MINERVA2, but there is no cubing. Recognition probability is computed as the ratio of summed similarity to the background level. The model can also be applied to multidimensional stimuli by assuming perturbation occurs independently and in parallel on all dimensions.

The dual-trace framework is an attempt to explain recall–recognition differences. It assumes two distinct traces—stimulus traces (a record of items encoded in terms of their features) and response traces (a record of the person's response to these traces). The stimulus trace is the primary basis for recognition but it is insufficient for recall. Traces of verbal responses (names, descriptive terms) provide a sufficient basis for recall. The model can explain memory loss, recovery, and distortions, and only has a few parameters, but it has not yet been applied to a wide variety of human memory data.

3.5 OSCAR

This is an OSCillator-based Associative Recall model (Brown et al. 2000) with a learning-context signal based on a temporal hierarchy of oscillators. The outer-product of item vectors and a dynamic context vector are stored in memory, paralleling the linear associator model of Anderson (1970), except that here each item is associated with the context rather than two items with each other. There is assumed to be an array of N oscillators; each element of the context vector is made up of the combined output of several oscillators of different frequencies. Each oscillator has a value that varies sinusoidally over time, and slow-moving oscillators carry more weight.

For serial recall, it is assumed that some or all of the oscillators (context) can be reset to their initial values and then recycled to regenerate (recall); thus it is a position-based and not an inter-item-based model for serial recall. The most similar item is recalled (no deblurring mechanism is specified), and after recall there is a temporary inhibition to implement a sampling-without-replacement rule. The model can explain a wide variety of serial-order effects, but little is said about item recognition, or associative recognition or recall. Presumably item recognition would be very easy to implement, but associative recall and recognition is another matter. Perhaps a three-way output product as in the matrix model of Humphreys et al. (1989) could be used but, to my knowledge, there has as yet been no commitment on this point.

3.6 TODAM2

This model (a theory of distributed associative memory) (Murdock 1997) is an extension of the Liepa model which also uses the convolution/correlation formalism, but is designed to provide an integrated

account of the storage and retrieval of item, associative, and serial-order information. Like Liepa, it uses convolution for storage and correlation for retrieval, and a useful feature is the recall–recognition identity. It can be shown (Weber 1988) that, using primes for approximations,

$$\mathbf{g} \cdot \mathbf{g}' = \mathbf{g} \cdot \{\mathbf{f} \# (\mathbf{f} * \mathbf{g})\} = (\mathbf{f} * \mathbf{g}) \cdot (\mathbf{f} * \mathbf{g}).$$

That is, if the retrieved item is obtained by correlating the probe $\mathbf{f}$ with the association $\mathbf{f} * \mathbf{g}$, the similarity of the retrieved approximation $\mathbf{g}'$ to the target item $\mathbf{g}$ is identical to the dot product of the associated pair with itself.

TODAM2 uses chunks for serial-order effects where a chunk is the sum of n-grams and an n-gram is the n-way association of the sum of n items. Then, like Liepa, multiple convolutions are used for recall. Like MINERVA2, context vectors flank the items, and each item (or, for pairs, the sum of the two items) is autoassociated to bind context to items or associations. Mediators can be included in the association, where a mediator is an intervening item which facilitates recall of unrelated items (e.g., ring—RUNG—ladder). So far TODAM2 is the only model designed explicitly to apply to all three types of information (item, associative, and serial-order), but it has not been applied to data as extensively as SAM or (in the serial-order area) OSCAR.

4. Discussion

There is diversity rather than consensus in this area of quantitative models of memory, at least partly attributable to the immaturity of our field. While the plethora of models may seem disturbing (and I have not even been able to include connectionist memory models such as McClelland and Chappell 1998, or Sikstrom 1996, it is hoped that some agreement will emerge in the not-too-distant future.

See also: Diffusion and Random Walk Processes; Knowledge Representation; Learning and Memory: Computational Models; Luce's Choice Axiom; Mathematical Learning Theory; Mathematical Learning Theory, History of; Mathematical Psychology; Mathematical Psychology, History of; Memory Retrieval; Recognition Memory, Psychology of; Signal Detection Theory

Bibliography

Anderson J A 1970 Two models for memory organization using interacting traces. *Mathematical Biosciences* **8**: 137–60

Borsellino A, Poggio T 1973 Convolution and correlation algebras. *Kybernetik* **122**: 113–22

Brown G D A, Preece T, Hulme C 2000 Oscillator-based memory for serial order. *Psychological Review* **107**: 127–81

Estes W K 1997 Processes of memory loss, recovery, and distortion. *Psychological Review* **104**: 148–69

Hintzman D L 1988 Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review* **95**: 528–51

Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA* **79**: 2554–8

Humphreys M S, Bain J D, Pike R 1989 Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review* **96**: 208–33

Izawa C (ed.) 1999 *On Human Memory: Evolution, Progress, and Reflections on the 30th Anniversary of the Atkinson-Shiffrin Buffer Model*. Lawrence Erlbaum Associates, Mahwah, NJ

Liepa P 1977 Models of content addressable distributed associative memory (CADAM). Unpublished manuscript, University of Toronto, Canada

Luce R D 1959 *Individual Choice Behavior, a Theoretical Analysis*. Wiley, New York

McClelland J L, Chappell M 1998 Familiarity breeds differentiation: A subjective–likelihood approach to the effects of experience in recognition memory. *Psychological Review* **105**: 724–60

Metcalfe J E 1982 A composite holographic associative recall model. *Psychological Review* **89**: 627–61

Murdock B B 1974 *Human Memory: Theory and Data*. Lawrence Erlbaum Associates, Potomac, MD

Murdock B B 1997 Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review* **104**: 839–62

Ratcliff R, Van Zandt T, McKoon G 1999 Connectionist and diffusion models of reaction time. *Psychological Review* **106**: 261–300

Shiffrin R M, Raaijmakers J 1992 The SAM retrieval model: A retrospective and prospective. In: Healy A F, Kosslyn S M, Shiffrin R M (eds.) *From Learning Processes to Cognitive Processes: Essay in Honor of William K. Estes*. Lawrence Erlbaum Associates, Hillsdale, NJ, Vol. 2

Sikstrom P S 1996 The TECO connectionist theory of recognition memory. *European Journal of Cognitive Psychology* **8**: 341–80

von Neumann J 1958 *The Computer and the Brain*. Yale University Press, New Haven, CT

Weber E U 1988 Expectation and variance of item resemblance distributions in a convolution-correlation model of distributed memory. *Journal of Mathematical Psychology* **32**: 1–43

B. Murdock

Memory: Organization and Recall

The concept of organization played a major role in memory research during the 1960s and 1970s. It was believed that the study of organizational processes in recall would be a useful vehicle for understanding how human mind processes information. Much research on organization and recall was generated during this

period. Since then, the interest in this topic has declined. The concept does not play the key role in memory as it used to do.

1. Definition

The term organization has been used in a number of ways in the context of recall and memory. One basic distinction is that between primary and secondary organization (Tulving 1968). *Primary organization* refers to the effects that are attributable to task characteristics. *Secondary organization* refers to organization that is imposed on the to-be-remembered information by the individual. It is this latter type of organization that has been of primary interest in trying to understand information processing and memory. On the basis of this Voss (1972) formulated one definition of organization that seems to capture the essence of what many memory researchers mean by organization in this context. 'Organization is a process that intervenes between input and output in such a way that there is not a 1:1 input–output relation' (Voss 1972, p. 176). This definition focuses on a particular output order that reflects a higher-order grouping of the to-be-remembered information and it emphasizes that the subject adds a structure to the information presented, which makes input and output different. A similar flavor is contained in the definition proposed by Tulving (1968): 'organization occurs when the output order of items is governed by semantic or phonetic relations among items or by the subject's prior extra-experimental or intra-experimental acquaintance with the items constituting a list' (Tulving 1968, p. 16).

2. Historical Roots

Organization theory of memory had its heydays during the 1960s and 1970s. However, the importance of organization in relation to learning and memory had been emphasized much earlier in *Gestalt psychology*. Although the main focus of the Gestalt psychologists was that of perceptual groupings (e.g., similarity, proximity), the notion was that this grouping had implications also for memory and retention. As noted by Postman (1972, p. 4), Wolfgang Köhler stated that those factors that are active in organization of primary experience should also affect recall. Postman (1972) noted four basic principles in the Gestalt conceptualization of organization that are of relevance for how organization came to be used by organization theorists during the 1960s and 1970s. First, he claimed that the Gestalt view was that organization is largely established by the initial perception of the events to be remembered. Second, the form of organization is determined by the relations among the component units, such as proximity and similarity, indicating that there is a natural or optimal organization. Third, the availability of prior experiences for recall over time depends on the temporal stability of the memory traces laid down by these experiences. Fourth, the accessibility of available traces for recall is a function of the similarity between these traces and current stimuli.

Quite a different influence of how organization came to play a major role in memory research dates even further back than that of the Gestalt psychologists. This is the way organization was used in the context of text recall. The role of organization in text processing was first specified early in the twentieth century and was later refined by Bartlett (1932). The term used by Bartlett was '*schema*,' which he defined as an active organization of past experience. The concept of schema and later the concept of *script* was taken up by many researchers in the area of text recall and discourse processing (e.g., Anderson et al. 1977, Kintsch and Vipond 1979, Schank and Abelson 1977).

3. Research Issues

Many research problems were invented in studies of organization and recall during the 1960s and 1970s. It was extensively explored whether there was an optimal number of categories to organize for a maximal recall to occur, whether the number of items per category was critical for *clustering*, and whether blocked vs. random presentation of category items made a difference.

In particular one topic was of major interest in this research field, namely that of developing measures or indices of organization. Several such measures were developed during the 1960s and 1970s. One commonly used way to classify these measures is whether organization is imposed by the experimenter or by the subject. For measures based on *experimenter-based organization* only one study trial is usually required, whereas for the *subject-based organization*, a multitrial design is required.

For experimenter-based measures, the most frequently used paradigm is that the items to be remembered belong to two or more different semantic categories. The words from the different conceptual categories usually appear in a random order in the list presented, but can, as mentioned, also appear in a blocked fashion in the study list. Even if no specific instruction to the subjects is given about organizing the words into semantic clusters, the output is typically organized such that the words recalled appear in clusters. This form of organization is referred to as categorical organization or categorical clustering. Word lists used in studies of organization and recall can also be composed of an associative rather than a categorical relationship.

Subject-based measures are used in experiments where the words comprising a list are unrelated, conceptually or associatively. That is, subjects are free to organize the words at output in whatever other

order they prefer. Organization in this case is based on idiosyncrasies of the subjects and it is determined by the extent to which they recall the items of a list in the same order at two successive trials.

Although the basis for organization of the study items differ between experimenter- and subject-based measures, it is generally assumed that both types of measures are indicating the same underlying psychological process of organization. Several different indices of each form of organization have been developed.

Bousfield (1953) developed several such indices for categorical clustering. The basic rationale for these different indices was to relate the number of successive repetitions from a given category in a response protocol to the total number of words recalled. A repetition in this case is the occurrence of two items from the same category directly after one another. One problem with some of the first indices developed was that they did not take into account the transition from one category to another. As such transitions are crucial for perfect clustering of a list of words with more than one category, a proper measure was not found until the item-clustering index (ICI) was developed. ICI took this aspect of clustering into account, but failed in other ways. It did not take into consideration the possibility that clustering can be perfect although not all items of all categories are recalled. The adjusted-ratio-of-clustering (ARC) score proposed by Roenker et al. (1971) took this into account. The formula developed to obtain a quantitative measure for organization by means of ARC is:

$$ARC = R - E(R)/maxR - E(R) \quad (1)$$

where R is the total number of observed category repetitions, E(R) is the expected number of category repetitions, and max$R$ is the maximum possible number of category repetitions. $E(R)$ is computed as the ratio between the sum of the squared number of items recalled from each category and the total number of items recalled, minus unity; max$R$ is the difference between the total number of items recalled and the number of categories represented in the response protocol. A comprehensive review of clustering measures was presented by Murphy (1979), who concluded that the ARC score is one of the best measures available for assessing clustering.

Measures of categorical clustering, e.g., ARC, were frequently used in memory research during the 1960s and 1970s. Today, these measures are less common in articles on memory, but are sometimes used as a dependent variable in its own right. For example, in studies of aging and memory, organization, as measured by means of the ARC score, has been used as a dependent variable to estimate cognitive deficits as a function of age in much the same way as memory deficits as a function of age is estimated by recall scores (e.g., Bäckman and Wahlin 1996). Bäckman and Wahlin (1996) concluded that organizability of the study list in semantic categories could serve as a cognitive support for recall and that the ability to utilize this support remained relatively constant across the age range that these authors studied, 75–96 years.

There are two main measures of subject-based organization. For one of these, proposed by Bousfield and Bousfield (1966), the critical index is the number of intertrial repetitions (ITR). This index is determined by constructing a matrix for each subject, with the items of the word list represented along both rows and columns of the matrix. Rows represent the nth word recalled and columns represent the $(n+1)$th word recalled. This matrix is then used for making a tabulation of the frequency with which given pairs occur adjacent to one another on two successive trials (Shuell 1969). This frequency value is the observed ITR value. The measure of subjective organization according to Bousfield and Bousfield (1966) is this observed ITR subtracted by an expected ITR

$$E(ITR) = c(c-1)/hk \quad (2)$$

where c is the number of items common to two recalls, h is the number of words recalled on trial n, and k is the number of words recalled on trial $n+1$.

The other measure of subject-based organization, proposed by Tulving (1962), is called *subjective organization*. Computing subjective organization this way, the data are organized in 2 x 2 contingency tables, with recall and nonrecall on trial n on one axis and recall and nonrecall on trial $n+1$ on the other axis. Tulving argued that the component for recall on two successive trials increased logarithmically over trials, while the component for nonrecall and recall on the two trials remained essentially constant with practice. Subjective organization is expressed quantitatively as:

$$SO = \Sigma n_{ij}\, logn_{ij}/\Sigma n_i\, logn_i \quad (3)$$

where n_{ij} represents the numerical value of the cell in the ith row and the jth column, and ni represents the marginal total of the ith row (see Shuell 1969).

The measures of subjective organization by Bousfield and Bousfield (1966) and Tulving (1962), respectively, are still used when subjective organization constitutes a dependent variable in recall experiments. The measures are of comparable power, but the ITR measure by Bousfield and Bousfield (1966) may be a bit more commonly used, simply because it is easier to compute.

One theoretical issue of great concern during the heydays of research on organization and recall was about the locus of organization. To this end it should be stated immediately that clustering is an output phenomenon from which some sort of organizational process is inferred. However, organization of the to-be-remembered information can equally well take place at the time of encoding. In a review article Shuell (1969) concluded that this was still an open question. He argued that a prerequisite for solving this issue was

that a measure of organization was developed that was independent of recall. Such a measure has not yet been developed and the basic issue of the locus of organization at encoding or retrieval is still an open question. Even in later reviews, for example those collected in Puff (1979), empirical data and theoretical statements appear to be rather mixed on the topic of the locus of organization.

Another research question examined in the area of organization and recall was the relationship between the phenomenon of organization and the notion of memory as an entity of limited capacity. Although the concept of organization was of great interest in itself, its popularity or central role in memory theory was also due to the fact that it provided, and still provides, a nice illustration of the notion of a capacity-limited memory. The basic finding is that recall performance is improved whenever an organizational structure is imposed on the to-be-remembered materials compared to cases when no organizational principle can be applied. This has been demonstrated in experiments comparing recall performance in organizable and nonorganizable lists of words. Thus, the use of organization in this way can be said to be an extension of the principles of chunking proposed by Miller (1956). However, organization has also been used in a broader sense, namely as a means for compensating for a limited capacity of memory for larger sets of to-be-remembered information. This type of organization need not be limited to clustering of words in categorized lists, but can be based on other principles. For example, the type of imagery structure of the materials to be remembered in mnemonic techniques, like the loci method, is one form of organization used in this broader sense.

Still another research question discussed was whether the organization processes underlying category clustering and subjective organization were the same or different, and whether the organization processes underlying clustering according to semantic category were same or different as organization processes behind grouping on the basis of other nonconceptual or nonassociative principles. In a vast majority of the studies carried out on organization and recall, the underlying organization principles has been based on semantic properties of the to-be-remembered materials. However, in a series of studies during the 1970s it was demonstrated that organization of the to-be-remembered materials can also be organized according to modality of presentation. Mixed lists of auditorily and visually presented words were presented and output in an immediate free recall test tended to be in clusters of auditorily presented words and visually presented words (Nilsson 1973). In a later study by Nilsson (1974), the list presented contained words from different semantic categories. Half of the words in each category was presented auditorily and the other half visually. The results showed that organization by modality was dominating in an immediate free recall test, whereas in a delayed final free recall test, the words in the recall output were ordered according to semantic category.

4. *Current Role of Organization in Memory Research*

As stated initially, the concept of organization played a major role in memory theory during the 1960s and 1970s. Many research questions were invented and many empirical variables were explored. However, progress in solving these questions and mapping the variables into existing organization theory was not overwhelming at the time (Shuell 1969). To some extent, this state of affairs might have been due to the fact that organization theory, as it was referred to at the time, was not really a theory in a formal sense. It was more of a framework or a point of view emphasizing higher-order cognitive processes beyond simple associationistic principles of learning and memory that had dominated the scene for many years.

Looking back now at the issues explored in the 1960s and 1970s, it can be concluded that not much has happened in answering the specific questions that were asked then. The question of whether there is some optimal number of categories in a list for maximal organization to occur has not been given a final answer. The question about the locus of organization at encoding or retrieval has not yet been answered in a way that has reached consensus in the scientific community. The interest in solving these and the other questions in focus of attention at the time seems to have declined although single papers on this topic have appeared in recent years (e.g., Kahana and Wingfield 2000).

Focus and interest changed from organization during the 1960s and 1970s to interactions between storage and retrieval. Experiments on recall and recognition of individual, unique word events became fashion during the 1970s. Perhaps the strongest expression of this is the great impact that the levels-of-processing framework (Craik and Lockhart 1972) had in reorienting the research community in memory to different research questions than before. Another approach to memory research that contributed to a shift in orientation was the new interest in structural and functional properties of episodic and semantic memory (Tulving 1972) and subsequently in the issue of memory systems in general. This reorientation meant a considerable broadening of the research issues of primary interest and focus. For one thing, this wider approach to studies of memory meant that many new disciplines got involved in a collaborative enterprise to understand how mind and brain process information. Technological inventions for brain imaging (positron emission tomography and functional magnetic resonance imaging) currently play a major role in studying memory. Future will tell whether the concept

of organization will be brought back into research focus again. Perhaps by using these brain imaging techniques as a means to obtain independent measures of organization.

See also: Elaboration in Memory; Encoding Specificity in Memory; Memory, Consolidation of; Memory: Levels of Processing; Reconstructive Memory, Psychology of

Bibliography

Anderson R C, Reynolds R E, Schallert D L, Goetz E T 1977 Frameworks for comprehending discourse. *American Educational Research Journal* **14**: 367–81

Bäckman L, Wahlin Å 1996 Influences of item organizability and semantic retrieval cues on word recall in very old age. *Aging and Cognition* **2**: 312–25

Bartlett F C 1932 *Remembering*. Cambridge University Press, London

Bousfield A K, Bousfield W A 1966 Measurement of clustering and of sequential constancies in repeated free recall. *Psychological Reports* **19**: 935–42

Bousfield W A 1953 The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology* **49**: 229–40

Craik F I M, Lockhart R S 1972 Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* **11**: 671–84

Kahana M J, Wingfield A 2000 A functional relation between learning and organization in free recall. *Psychonomic Bulletin & Review* **7**: 516–21

Kintsch W, Vipond D 1979 Reading comprehension and readability in educational practice and psychological theory. In: Nilsson L-G (ed.) *Perspectives on Memory Research: Essays in Honor of Uppsala University's 500th Anniversary*. Erlbaum, Hillsdale, NJ, pp. 329–65

Miller G A 1956 *Language and Communication*. McGraw-Hill, New York

Murphy M D 1979 Measurement of category clustering in free recall. In: Puff R C (ed.) *Memory Organization and Structure*. Academic Press, New York, pp. 51–83

Nilsson L-G 1973 Organization by modality in short-term memory. *Journal of Experimental Psychology* **100**: 246–53

Nilsson L-G 1974 Further evidence for organization by modality in immediate free recall. *Journal of Experimental Psychology* **103**: 948–57

Postman L 1972 A pragmatic view of organization theory. In: Tulving E, Donaldson W (eds.) *Organization of Memory*. Academic Press, New York, pp. 3–48

Puff R C 1979 (ed.) *Memory Organization and Structure*. Academic Press, New York

Roenker D L, Brown S C, Thompson C P 1971 Comparisons of measures for the estimation of clustering in free recall. *Psychological Bulletin* **76**: 45–8

Schank R, Abelson R 1977 *Scripts, Plans, Goals, and Understanding*. Erlbaum, Hillsdale, NJ

Shuell T J 1969 Clustering and organization in free recall. *Psychological Bulletin* **72**: 353–74

Tulving E 1962 Subjective organization in free recall of 'unrelated words.' *Psychological Review* **69**: 344–54

Tulving E 1968 Theoretical issues in free recall. In: Dixon T R, Horton D L (eds.) *Verbal Behavior and General Behavior Theory*. Prentice-Hall, Englewood Cliffs, NJ, pp. 2–36

Tulving E 1972 Episodic and semantic memory. In: Tulving E, Donalson W (eds.) *Organization of Memory*. Academic Press, New York, pp. 381–403

Voss J F 1972 On the relationship of associative and organizational processes. In: Tulving E, Donalson W (eds.) *Organization of Memory*. Academic Press, New York, pp. 167–94

L.-G. Nilsson

Memory Problems and Rehabilitation

Models of memory emphasize the distinction between declarative and procedural types of knowledge (Squire 1987). *Declarative* memory can be further broken down into *episodic* (i.e., memory for experiential information such as what one had for breakfast), and *semantic* (i.e., knowledge of general information such as dates in history, scientific formulas, state capitals, etc.) processes. *Procedural* memory refers to skills, habits, and classically conditioned behaviors (e.g., playing the piano and driving a car). These different memory processes are mediated by different brain structures; hence, they break down differentially in the context of neurological injury. In this article, remediation for problems in the declarative memory domain will be discussed. First, factors to consider when selecting a rehabilitation strategy will be addressed. Then, a wide variety of strategies will be described, including internal strategies, external strategies, use of augmentative technology, and pharmacological intervention.

1. Selection of Strategy

In the clinical context disruption in various stages of memory can result in memory impairment. In order to form a new memory one must first be able to encode or analyze the information or event. The second process in memory formation is consolidation or storage. Finally, one must be able to retrieve or access the information as needed. Remediation strategies vary according to the particular stage of memory that is disrupted. Disruptions in encoding and retrieval can often be remediated with internal aids. On the other hand, disruptions in consolidation (or storage) of information are often not improved with internal strategies and therefore must rely on external aids for compensation.

Different types of memory depend on different brain structures, and therefore are differentially affected in relation to various disease processes. Injury to the hippocampus and adjacent structures may disrupt consolidation. Examples of neurological conditions that affect the hippocampus include Alzheimer's disease, encephalitis, and anoxia (loss of oxygen to the brain resulting in death of brain cells). Other neuro-

logical conditions affect subcortical or frontal brain areas (e.g., multiple sclerosis, Parkinson's disease, frontal lobe dementia, head injury, and stroke). In these conditions, attentionally mediated memory abilities are disrupted, rendering patients vulnerable to problems with encoding and retrieval.

In order to determine the most appropriate remediation strategy, the type of memory problem must first be characterized through a comprehensive neuropsychological assessment. In addition to memory, other cognitive domains, such as attention, language, reasoning, visuospatial abilities, and motor skills, should be assessed in order to determine which memory remediation strategies are optimal for a particular patient. It is important to note that insight into one's medical condition and memory status is vital to the remediation process (Prigatano et al. 1986). Finally, an important aspect of a complete neuropsychological assessment includes assessment of emotional functioning, as significant depression or anxiety can adversely affect memory ability, regardless of other disease processes. Only after emotional difficulties are assessed and treated can the memory problem be accurately assessed.

2. *Internal Strategies*

Internal memory strategies include techniques to improve encoding and retrieval. Like other skills, these strategies must be practiced over time. Furthermore, to execute such skills, one must be able to focus attention on the task at hand.

Many internal strategies target the encoding process. Through various methods, associations between new information and old (i.e., semantic) knowledge are formed in ways that make sense to that individual. These new associations provide cues for later retrieval. Because these strategies depend upon previously stored semantic knowledge, in cases where semantic memory is impaired, internal strategies might not be effective. Three common internal strategies are imagery, verbal elaboration, and organization (West 1995). As noted earlier, internal strategies are often not successful for people with significant consolidation problems, as they will not remember to employ them when necessary (Wilson and Moffat 1984, Glisky 1995).

Imagery involves the association of interactive visual images. This method has proven especially effective in the learning of names and faces (Yesavage et al. 1983). In this instance, one would study a new person to identify a characteristic that could be linked to that person's name. For example, when meeting a person named 'Karen Singer,' one could conjure up an image of the person singing. This technique can also be used to link words together, such as items from a grocery list, errands that must be done, or daily chores. For example, if one needs to pay the bills and do the vacuuming, one could create a visual image of vacuuming away the bills. A single image would therefore provide the cues for remembering both chores.

Verbal elaboration involves the creation of new words or sentences, rather than images, to link information together. These techniques are dependent upon semantic and verbal abilities. If one wanted to remember a new name, one could combine the first and last name into a new sentence. For example, the name 'Darren Wernick' could be remembered with the sentence 'Darren wore a neck-tie,' as the words 'wore' and 'neck' would provide the cue for the name. Mentally linking this sentence to a visual image, such as Darren putting on a tie, would further strengthen the elaboration and enhance later retrieval. Making up rhymes is a common method for recalling information (e.g., Thirty days hath September ...). Another method is first-letter elaboration, in which one uses the first letters in words to create a new word or pseudo-word. For example, to remember emergency procedure for a hospital fire, employees are taught to remember the word 'RACE.' This not only reminds workers of the steps to take in case of a fire, but also the order of steps ('Rescue' patients, sound the 'Alarm,' 'Confine' the fire, 'Evacuate' the premises).

It is important to note in both imagery and verbal elaboration that the imaginary links created between objects or words are most effectively remembered when they make personal sense to that individual, even if they seem nonsensical to others. Often the first and most creative idea is the most salient cue for later recall.

Finally, *organization* of incoming information is beneficial to emphasize natural links between stimuli in order to enhance later retrieval. Organizational techniques can be particularly useful in situations where the individual has to learn information that exceeds attention span limitations (e.g., a lengthy list). One effective method of organization is 'chunking,' or grouping items into categories (Cermak 1975). When trying to remember items for the grocery store one could organize the list so that the five dairy items and the five vegetables are grouped together. Therefore, simply remembering the two general categories of dairy items and vegetables would enhance memory retrieval. This technique is also used to remember sequences of numbers. For example to remember a phone number, one could chunk ten individual numbers into three larger numbers.

3. *External Strategies*

Internal strategies enhance encoding and retrieval abilities of patients with attentionally based memory problems. However, those with impaired consolidation capacity will often not benefit from such

techniques. Often consolidation problems occur as the result of damage to the hippocampus and related cortices. Because they rely less on cognitive capacity and effort, external memory strategies are often more effective than internal strategies. Wilson (1991) found that patients taught both internal and external strategies tended to utilize external aids more often and for a longer amount of time following intervention. External strategies include use of both environmental and portable cues.

Environmental memory techniques involve structuring one's environment so that memory cues are naturally provided by the surroundings. Kapur (1995) describes three main environmental strategies that include personal, proximal, and distal cues.

Personal environmental cues involve making changes on one's person that serve as reminders. For example a person might place a ring on an unusual finger as a prompt to remember to make a phone call. The ring would be returned to the proper finger once the task has been accomplished. The drawback to this method is that it serves as a cue to remember an event, but does not provide specific information regarding that event.

Proximal environmental cues involve manipulating the layout within a space or a room so that cues are permanently provided concerning the location of objects and procedures that should be attended to. Examples include having a desk drawer dedicated to letter-writing and bill-paying supplies, having a basket within eyesight to place unpaid bills, and having a shelf near the door for outgoing mail. In this instance, the room itself provides cues to remember an important chore (i.e., paying the bills), where to find supplies (i.e., desk drawer), and a cue to complete the last step of the task (i.e., taking bills to the mailbox). Clear labels describing contents of storage areas are also often helpful for people with severe memory problems, as they might not remember where to find things on a day-to-day basis.

Distal environmental cues provide directions to places and instructions on how to safely navigate one's surroundings. These can include navigating a home, a building, a neighborhood, or transportation networks. Cues can include maps, visual signs pointing out how to get to certain places, labels reminding people where they are going, and warning signs about possible hazards that they may encounter (e.g., steep stairs, heavy doors, etc.). A professional assessment of the home environment is helpful in determining which proximal and distal cues are necessary best to create a supportive and safe setting.

In contrast to environmental aids, *portable* external memory aids are devices that people can take with them into other environments. Such devices are intended to provide cues and reminders about important information. Examples are notebooks, calendar books, and day planners. Often people with memory problems must be taught to use these devices until they become habitual and no longer require novel memory effort.

Memory books are special notebooks created by rehabilitation therapists that contain specific information to meet the needs of the person with memory problems. Such books are personally tailored for the individual, and may contain such sections as daily schedules, orientation calendars, addresses/phone numbers, maps for getting around, pictures and names of people that the individual will interact with, pictures of buildings the individual should know, and detailed instructions for accomplishing certain tasks. The book must be easy for the individual to use (i.e., clearly labeled, easy to carry, self-explanatory). Often therapists spend a good deal of time teaching patients how to use memory books. Correct usage often requires practice and rehearsal until the procedure becomes automatic. Without proper instruction and cueing, individuals with significant memory problems often cannot use external aids successfully.

4. *Augmentative Technology*

In recent years there has been a proliferation of electronic organizers and computer programs that serve as external memory aids for normal healthy individuals as well as in the rehabilitation setting. In order for patients with memory problems to use such technology, they must retain sufficient cognitive capacity to learn complex new skills with new equipment. Factors that affect ability to utilize such technology include age, educational level, and prior familiarity with electronics or computers (Kapur 1995). Success with this approach requires careful analysis and selection of specific memory aids to compensate for the specific memory problems. Attempts should be made to avoid depending on other areas of cognition that may be compromised. Furthermore, patients often must have support in their home environment, usually by family members, as they learn to use and rely upon technology on a daily basis.

An older type of electronic aid is the dictating machine. These are still useful tools for recording verbal information for later use, especially if the information is abundant and given quickly, as in a classroom or lecture situation. Furthermore, dictation machines can be used by people with visual or writing impairments, and are fairly simple to use even for patients with significant memory impairment. However, they do not provide a way to organize incoming information.

Much new computer software that can facilitate the organizational skills of memory-impaired individuals is available. Computer programs can be specifically designed or altered to compensate for a particular disability. Research indicates that successful computer usage increases self-esteem in brain-injured patients (Glisky and Schacter 1987). However, the cost-

effectiveness of individualized computers and computer programs as compensatory tools is questionable because these techniques are narrow in application and require a great allocation of resources (Glisky 1995).

A very promising development in electronic memory aids is the personal digital assistant (electronic organizer), such as the Palm Pilot. These portable machines can be individually tailored to meet a variety of needs, serving as a notebook, daily schedule, calendar, address book, reminder list, and even auditory alarm. However, the user must be able to learn three basic operations: entering new information, reviewing stored information, and deleting obsolete information.

In several case studies, patients with memory problems were introduced to electronic organizers. Findings indicated that patients with very poor memory, lack of initiative, lack of insight into memory problems, difficulties with visual attention, poor motor control, and limited problem-solving skills had difficulty learning to utilize these aids successfully. However, patients with mild impairment in everyday memory functioning may find the electronic organizer extremely helpful in compensating for memory lapses (Kapur 1995).

5. *Pharmacology*

Some types of memory problems occur due to poor encoding or retrieval secondary to attentional limitations. Medications that facilitate attention, such as Ritalin and other stimulants, have the potential to increase functional memory abilities in such patients. However, many memory problems are due to consolidation problems secondary to hippocampal damage. Alzheimer's disease (AD) is the most common type of disorder that impairs consolidation. Because AD affects a large segment of the population, considerable resources have been devoted to developing effective pharmacological intervention. Hence, most of our knowledge regarding psychopharmacological remediation of memory disorders is based on research of patients with AD. Although no medications exist for repairing brain damage once it has occurred, several medications are being developed to prevent or slow down brain degeneration associated with progressive dementia, such as AD. Other medications focus on enhancing specific neurotransmitters that may be depleted in this disorder.

Two drugs have been approved by the FDA for use in memory remediation in patients with AD: tacrine and donepezil. Both are cholinesterase inhibitors, as they prevent the breakdown of acetylcholine, a neurotransmitter needed for memory processing and thought to be diminished as a result of AD (Frautschy 1999). Both drugs have demonstrated some ability to slow down memory decline in some, but not all, AD patients in the early stages of disease (Doody 1999). However, tacrine was shown to have serious side effects in some patients and therefore is no longer commonly prescribed. As of 2001, several new cholinesterase inhibitors, such as metrifonate and eptastigmine, are undergoing clinical trials. These may prove to be promising agents in AD treatment.

Currently there are many other drugs in development, applying different strategies and theories to the slowing of AD. Many are focused upon eliminating or slowing formation of abnormalities characteristically found in Alzheimer's brains, such as plaques and tangles. Although many pieces of the puzzle continue to fall into place with ongoing study, the cascade of events that creates these abnormalities is still not well characterized.

Other research approaches have used antioxidants, such as Vitamin E, as these compounds may help to prevent AD-induced neurotoxicity in the brain. Because AD patients' brains often show evidence of plaque-associated inflammatory response, anti-inflammatory agents may also prove to be beneficial. Multiple ongoing research trials are currently investigating beneficial effects of steroids, such as corticosteroids, estrogens, and androgens. However, extensive study of human subjects will be required before benefits of these approaches are clearly understood (Frautschy 1999).

6. *Conclusions*

Memory impairment can occur for many reasons, and memory can break down in different ways. Professional consultation by neuropsychological and rehabilitation specialists may help to characterize the type and extent of memory deficits, associated cognitive impairment, and emotional distress. Specific cognitive strengths and weaknesses and details regarding the particular type of memory deficit will dictate the compensatory strategies that are selected and taught to the individual. These can include both internal and external strategies. Furthermore, some patients with specific neuropsychological profiles may benefit from pharmacological intervention. In particular, patients with attentionally based memory problems may benefit from use of stimulants. Increased use of strategies, aids, and techniques has improved day-to-day memory performance in memory-impaired patients, even at long-term follow up (Wilson 1991). These findings underscore the potential effectiveness of a strong and precise program of rehabilitation.

See also: Brain Damage: Neuropsychological Rehabilitation; External Memory, Psychology of; Memory and Aging, Cognitive Psychology of; Memory: Organization and Recall; Memory Retrieval

Bibliography

Cermak L S 1975 *Improving Your Memory*. McGraw-Hill, New York
Doody R S 1999 Clinical profile of Donepezil in the treatment of Alzheimer's Disease. *Gerontology* **45**: 23–32
Frautschy S 1999 Alzheimer's disease: Current research and future therapy. *Primary Psychiatry* **6**: 46–68
Glisky E L 1995 Computers in memory rehabilitation. In: Baddeley A D, Wilson B A, Fraser N W (eds.) *Handbook of Memory Disorders*. Wiley, New York
Glisky E L, Schacter D L 1987 Acquisition of domain-specific knowledge in organic amnesia: Training for computer-related work. *Neuropsychologia* **25**: 893–906
Kapur N 1995 Memory aids in the rehabilitation of memory disordered patients. In: Baddeley A D, Wilson B A, Fraser N W (eds.) *Handbook of Memory Disorders*. Wiley, New York
O'Connor M, Cermak L S 1987 Rehabilitation of organic memory disorders. In: Meier M J, Benton A L, Diller L (eds.) *Neuropsychological Rehabilitation*. Guilford, New York
Prigatano G P, Fordyce D J, Zeiner H K, Roueche J R, Pepping M, Wood B C 1986 *Neuropsychological Rehabilitation after Brain Injury*. Johns Hopkins University Press, Baltimore
Squire L R 1987 *Memory and Brain*. Oxford University Press, New York
West R L 1995 Compensatory strategies for age-associated memory impairment. In: Baddeley A D, Wilson B A, Fraser N W (eds.) *Handbook of Memory Disorders*. Wiley, New York
Wilson B A 1991 Long-term prognosis of patients with severe memory disorders. *Neuropsychological Rehabilitation* **1**: 117–34
Wilson B A 1995 Management and remediation of memory problems in brain-injured adults. In: Baddeley A D, Wilson B A, Fraser N W (eds.) *Handbook of Memory Disorders*. Wiley, New York
Wilson B, Moffat N 1984 Rehabilitation of memory for everyday life. In: Harris J E, Morris P E (eds.) *Everyday Memory: Actions and Absentmindedness*. Academic Press, London
Yesavage J A, Rose T L, Bower G H 1983 Interactive imagery and affective judgements improve face-name learning in the elderly. *Journal of Gerontology* **38**: 197–203

A. M. Sherman and M. O'Connor

Memory Psychophysics

Memory psychophysics, or mnemophysics, is the branch of psychophysics that treats the functional relations between physical stimuli and their remembered sensory responses or representations. The name implicates the parent disciplines of memory and psychophysics. Memory researchers have long explored the reliability of representing the past, pursuing the relationship between the objective properties of stimuli that are no longer present(ed) and their current, remembered representations. Classical psychophysics, however, deals with the relationship between the physical properties of the momentary stimuli impinging on the sensory surface and the instantaneous sensations or perceptual representations. With the affinity of the respective pursuits recognized, students of memory psychophysics have sought to deploy psychophysical theory and methods to elucidate the quantitative properties of memory-based judgments and representations. Memory psychophysics is a young field of inquiry, yet several emerging trends are already discernible. These developments are examined in chronological order. Considered first is their historical context, and the article concludes with remarks concerning the prospects of memory psychophysics.

1. Memory Psychophysics: Antecedents, Conjectures, and Modern Beginnings

The idea that memory is based on images, on seeing internal pictures, can be traced back to at least Aristotle, if not Plato (see Algom 1992a,b, for a historical perspective). Plato's notion of *eikon*, prevalent in his discussions of memory, refers to a copy or an image that holds considerable similarity to the original perception. In *De Memoria*, Aristotle suggests that people rely on small-scale internal models of the perceptual referent when they remember. Aristotle proposes a spatial-analog representation underlying the memory of scenes and events. The spatial images undergo continuous transformations that enable the person to date the original perception in time. Implying a shrinkage over time in the size of the memory image, Aristotle comes remarkably close to a re-perception theory of memory, anticipating the current mnemophysical hypothesis bearing that name.

In the nineteenth century, several pioneers of psychophysics and psychology considered the idea of a psychophysics applied to memory. Fechner's little-read article from 1882 is titled, 'Some thoughts on the psychophysical representations of memories,' and Titchner's 1906 classic, *An Outline of Psychology*, explicitly acknowledge the discipline of memory psychophysics along with the strong assertion that memory obeys Weber's law. Wundt's psychology of consciousness similarly invites the perceptual study of memory, as does the work of Ebbinghaus on human memory, immensely influenced by Fechner's psychophysics.

Modern mnemophysics was (re)born in the Laboratory of Psychophysics at the University of Stockholm, Sweden. Bjorkman, Lundeberg, and Tarnblom (in 1960) were probably the first modern researchers to study, via rigorous psychophysical methods, the subjective magnitudes of perceptual (physically presented) and remembered (symbolically represented) stimuli in tandem. In the same laboratory, Dornic first applied Steven's power law of sensation to judgments of remembered stimuli. In a similar vein, Ekman and Bratfish and other researchers have asked people to assess inter- and intra-city distances, estimates that

must be considered to be forms of 'cognitive psychophysics' (cf. Wiest and Bell 1985, who should be consulted for a review) inasmuch as these stimuli can not be presented for view. Shepard and Chipman (1970) had people compare states of the continental USA on geometric similarity once perceptually (on the basis of outline shapes) and once from memory (on the basis of mere names). The respective sets of data were similar to the extent that Shepard and Chipman concluded that internal representations are second-order isomorphic to the referent physical stimuli. In another landmark study, Moyer (1973) had people select, while timed, the larger of two animals based on presentation of their names. Response time decreased as the difference in size between the referent animals increased ('symbolic distance effect'), mirroring the perceptual relation by which reaction time is inversely related to stimulus difference. Moyer invoked the notion of 'internal psychophysical judgment' to account for the data. Finally, Baird, another veteran of the Stockholm laboratory, explicitly proposed a cognitive theory of psychophysics in the early 1970s.

2. *Unidimensional Mnemophysics: Psychophysical Functions for Remembered Stimuli*

Ultimately, it was 1978 that ushered in the discipline of memory psychophysics. Two independent studies were published in prestigious journals, deriving psychophysical functions for remembered stimuli. In one, Moyer et al. (1978) had separate groups of people estimate the sizes of perceived and of remembered stimuli. Participants in the memory conditions were told to imagine each stimulus as its name (learned earlier) was called out, then assign a number in accord with standard magnitude estimation instructions. Separate psychophysical functions were then derived for the common set of physical stimuli, one for perceptual judgments, the other for judgments made from memory. The same method has been used in subsequent studies probing the gamut of sensory modalities from visual length, distance, area, volume, and brightness, to auditory loudness and volume, to dimensions of touch, taste, odor, and pain (see Algom 1992b, for a comprehensive review).

The following fundamental questions were pursued. How do remembered sensory magnitudes depend on referent physical magnitudes? Do memory scale values map onto their physical referents by means of the same functional relation (e.g., Steven's power transform) as do perceptual scale values? And if so, do the same parameters (exponents) govern perceptual and memory functions? Again, standard psychophysical methods and analyses can and have been used with both perceptual and memory types of judgment, the only procedural difference being that stimuli are in one case physically presented, in the other, symbolically represented.

Two general findings have emerged from these unidimensional studies of memory psychophysics. First, judgments from memory relate to the corresponding physical stimuli via power transforms in much the same way as perceptual judgments. Second, systematically smaller exponents govern the memory functions compared with the respective perceptual functions. For instance, in the aforementioned study by Moyer et al. (1978), perceptual area related to physical area by a power function with an exponent of 0.64, whereas remembered are related to physical area by a power function with an exponent of 0.46.

Two rival formulations have been suggested to account for properties of memory-based magnitude judgments. According to the reperception hypothesis, perception and memory perform identical (power) transformations on the input data. Because the input to memory processing has already been transformed perceptually (by exponent b), the memory exponent should reflect two transformations, and thus equal b^2. The alternative uncertainty hypothesis posits that greater uncertainty causes people to constrict the range of judgments or to expand the range of the underlying stimulus dimension, thereby producing an attenuated memory exponent.

The cumulative results provide qualified support for the reperception and the uncertainty hypotheses—in that order. Both theories predict smaller memory exponents for compressive perceptual continua (characterized by smaller-than-unity power function exponents), a prediction amply borne out by the extant data. However, clear resolution must await examination of expansive continua (characterized by greater-than-unity power function exponents) for which the reperception hypothesis predicts a steeper memory function, whereas the uncertainty hypothesis still predicts an attenuated memory function. Mnemophysical investigation of pain (Algom and Lubel 1994), an expansive sensory dimension, has provided qualified support for the reperception hypothesis.

The good fits to the memory functions obtained throughout the various studies affirm that, likely, more than just ordinal information is conserved. Nevertheless, the same studies are vulnerable on counts of validity, beset by indeterminacy of the functions relating sensations and memories to the corresponding overt responses. The issue cannot be solved within the confines of the univariate designs used.

3. *Multidimensional Mnemophysics: Rules of Stimulus Integration in Perception and in Memory*

Steven's contention that the psychophysical law is a power function depends on the strong assumption that the numbers given by participants (their overt responses) are proportional to sensation or memory magnitudes (their inner perceptions). However,

Stevens provided no justification for this assumption. As a result, the validity of the power law and the associated findings and interpretations are suspect. The need for an adequate metric structure to support validity is as vital for memory psychophysics as indeed it is for the entire edifice of psychophysics (see Narens 1996, for a firm formal foundation for Stevens's approach).

Examination of multidimensional stimuli such as rectangles, mixtures of odor and of taste, or painful compounds, does provide the underlying metric structure needed for authenticating the overt response. The problem becomes tractable because the rules by which the components integrate—uncovered and tested by methods such as conjoint measurement or functional measurement—provide the necessary constraints to validate the psychophysical functions. Indeed, specification of the appropriate model of integration entails the scale—the psychophysical function—as its natural derivative. Most important in the context of memory psychophysics, deployment of multivariate methods enables the examination of a wholly novel class of questions. They concern the respective rules of integration in perception and memory pertaining to the same set of physical stimuli. Integration models have been specified for numerous stimuli using a perceptual response. Multivariate studies of memory psychophysics have established the complementary integration models for remembered stimuli. At issue is the constancy in form of stimulus integration in perception and memory.

For area of rectangles presented for view (Algom et al. 1985), the veridical height X width rule has been shown to underlie the perceptual judgments. The same multiplicative rule reappeared when the rectangles were not presented for view, but were represented instead by previously learned names ('remembered composites'). In another condition ('mental composites'), participants were trained to name a set of horizontal and vertical line stimuli varying in length. Subsequently, the participants were instructed to form imaginary rectangles whose sides were made of previously shown pairs of lines represented by a pair of appropriate names. Notably, the same height X width rule applied again, despite the fact that the stimuli were wholly imaginary (no physical rectangles had been presented). A third condition ('semimental composites') resembled 'mental composites' in that it entailed no presentation of the physical stimulus; judgments were based on the (separate) presentation of a physical line and a name (standing for another, perpendicular, line) to be considered by the participant as the respective sides of an imaginary rectangle. For semimental composites, too, the same height X width rule applied.

In displaying an invariance in the rules of multidimensional integration, visual area joins other continua, including additional visual attributes, smell, taste, and pain. For olfaction, to cite a single example, the same rule of synthesis has been shown to hold for physical, remembered, mental, and semimental mixtures of given odorants. The transrepresentational invariance also holds developmentally. When, in the course of development, perceptual integration changes (e.g., from addition to multiplication), so too does the corresponding memorial integration. The results may tap a general-purpose compositional strategy: integration rules are invariant across various processing stages of a given stimulus. The validated functions derived within the framework of multifactorial mnemophysics are commensurate with the earlier findings: memory functions are markedly more compressive than the respective perceptual functions (except for young children). Therefore, a given invariant rule of integration acts on different sets of scale values in perception and memory.

4. *Symbolic Comparisons*

A common mental activity of everyday life consists of comparing pairs of stimuli in their absence and deciding which contains more of the attribute of interest. Deciding which numeral in the pair, 8–2, is larger numerically, or deciding which name in the pair, cat–dog, refers to the larger animal, are common examples. Such comparisons entail symbols (typically, though not exclusively names or numbers) standing for the referent stimuli, and thus are necessarily based on information retrieved from memory. The symbolic distance effect, mentioned earlier, documents one characteristic of symbolic comparisons. The semantic congruity effect by which large pairs are compared faster under 'choose the larger stimulus' instructions and small pairs are compared faster under 'choose the smaller stimulus' instructions, documents another. A third phenomenon, the end effect, pertains to the fact that pairs of stimuli containing the smallest or the largest stimulus as a member are compared faster.

Note that the main dependent variable used in this research is response time (augmented, at times, by a measure of error), not magnitude estimation or other (nonspeeded) scaling estimates. This reflects a preoccupation with the substantive processes of learning, comparison, representation, and decision. Earlier developments in memory psychophysics, by contrast, mainly concerned scaling and organization. Indeed, research on symbolic comparisons and research on scaling and composition of single remembered stimuli has not been satisfactorily integrated to date. With the application to comparison data of advanced psychophysical techniques, this domain comes squarely under the purview of memory psychophysics.

A discrete, proposition-based, semantic coding account (Banks 1977) has been influential against rival 'analog' theories of the mental comparison process. More recently, evidence accrual decision models (e.g., Petrusic's 1992, Slow- and Fast-Guessing Theory),

characterized at times as a 'random walk' (e.g., Link's 'Wave Discrimination Theory'), and connectionist networks (e.g., Leth-Steensen and Marley 2000), have been developed. A notable feature of these theories is their full alliance with mainstream psychophysics and cognitive psychology. By the same token, however, many recent models lack a treatment of the difference between comparisons made with physical and remembered stimuli. This difference, one should recall, is the *raison d'être* for the establishment of an independent discipline of memory psychophysics.

5. Concluding Remarks

The primary function of the senses is to guide ongoing behavior, yet they also exert considerable influences on memory associated with the original sensations. Sensations endure and inform cognition well beyond the physical presence of the triggering stimulus. It is those cognitions and memories that are captured and elucidated within memory psychophysics. As this article shows, there is a conceptual shift from scaling to cognizing in the pertinent research. This allows for the examination of a richer class of phenomena including decision, context, learning, and representation. The framework of single-variable scaling has been inhospitable to explicating process-based phenomena. Nevertheless, integrating scaling and comparison data remains a challenge to be met in future memory psychophysics.

See also: Learning Curve, The; Psychophysical Theory and Laws, History of; Psychophysics; Sensation and Perception: Direct Scaling; Visual Space, Geometry of

Bibliography

Algom D 1992a Introduction. In: Algom D (ed.) *Psychophysical Approaches to Cognition*, Elsevier, Amsterdam, The Netherlands

Algom D 1992b Memory psychophysics: An examination of its perceptual and cognitive prospects. In: Algom D (ed.) *Psychophysical Approaches to Cognition*, Elsevier, Amsterdam, The Netherlands

Algom D, Wolf Y, Bergman B 1985 Integration of stimulus dimensions in perception and memory: Composition rules and psychophysical relations. *Journal of Experimental Psychology: General* **114**: 451–71

Banks W P 1977 Encoding and processing of symbolic information in comparison tasks. In: Bower G H (ed.) *The Psychology of Learning and Motivation*, Academic Press, San Diego, CA, Vol. 11

Baranski J V, Petrusic W M 1992 The discriminability of remembered magnitudes. *Memory & Cognition* **20**: 254–70

Leth-Steensen C, Marley A A J 2000 A model of response time effects in symbolic comparison. *Psychological Review* **107**: 62–100

Moyer R S 1973 Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics* **13**: 180–4

Moyer R S, Bradley D R, Sorensen M H, Whiting J C, Mansfield D P 1978 Psychophysical functions for perceived and remembered size. *Science* **200**: 330–2

Narens L 1996 A theory of ratio magnitude estimation. *Journal of Mathematical Psychology* **40**: 109–29

Petrusic W M 1992 Semantic congruity effects and theories of the comparison process. *Journal of Experimental Psychology: Human Perception and Performance* **18**: 962–86

Shepard R N, Chipman S 1970 Second order isomorphism of internal representations: The shapes of states. *Cognitive Psychology* **1**: 1–17

Wiest W M, Bell B 1985 Stevens's exponent for psychophysical scaling of perceived, remembered, and inferred distance. *Psychological Bulletin* **98**: 457–70

D. Algom

Memory Retrieval

Memory retrieval is the recovery of stored information. It is a feature of virtually all cognitive activity, whether it be conducting an everyday conversation, planning a course of action, or making simple decisions. Yet, as with many cognitive processes, memory retrieval is a skill largely taken for granted. Only the experience of retrieval failure—a friend's name refusing to come to mind, or an interaction with someone suffering a severe memory disorder such as Alzheimer's disease—leads to an appreciation of memory's pervasiveness and to questions about the processes that underlie it. How does someone extract from the vast amount of stored information the one piece of information that the occasion demands? This article offers a brief account of what is currently known about these retrieval processes.

1. Two Views of Memory Retrieval

Two simple views provide a necessary foundation for the presentation of contemporary accounts of memory retrieval. One view claims that retrieval is determined by the state of the memory trace. The other claims that it depends on the presence of an effective retrieval cue.

1.1 The Trace-dependent View of Retrieval

The simplest account of memory retrieval assumes that it depends only on the strength of the memory trace relative to other traces. Retrieval failure occurs because the memory trace has become too weak, or because competing traces have become stronger than the target trace. Such a trace-dependent view of

retrieval held sway in much of experimental psychology until well into the 1960s.

According to the trace-dependent view the strength of the memory trace (and thus the likelihood of retrieval) is a function of its initial strength and the length of time between this original experience and the attempted retrieval. The body of evidence supporting both these claims is overwhelming. The second assumption is supported by countless experiments that show a gradual decrease in retrieval success as the retention interval increases. Consider now the first assumption. The initial strength of a trace is determined by a number of factors, most notably by the way in which the experience is encoded, and it is not at all difficult to demonstrate that such encoding processes influence retrieval (see *Memory: Levels of Processing*). For example, in experiments requiring the recall of names of simple objects, recall levels can be greatly enhanced if, during presentation of the names, participants are asked to form a visual image of each object.

The trace-dependent view can account for many of the basic phenomena of memory retrieval. For example, the fact that retrieval becomes more difficult with passing time is a consequence of a weakening of trace strength. The fact that an unrecallable item can nevertheless often be recognized is explained by the claim that recall requires a stronger trace than does recognition. By appealing to fluctuations in trace strength over time it can even account for the common experience of being unable to recall something at one moment but being able to do so at a later time.

In fact, the trace-dependent view of retrieval is not so much false as seriously incomplete. It is incomplete in at least three ways. First, strength is not the only feature of the memory trace that is important to retrieval; so too is its content. A unidimensional concept such as strength fails to capture the qualitative (or multidimensional, semantic) aspects of the encoding process. As will be seen, this is a feature of the memory trace that is of great importance in understanding retrieval. Second, the state of the memory trace, no matter how it is characterized, is not the only factor that influences retrieval. These further influences will be described in the remainder of this article. Third, the fact that encoding processes strongly influence retrieval does not constitute an account of the retrieval process itself, any more than the fact that practice increases skill at a game such as chess is an account of the processes involved in skilled chess playing.

1.2 The Cue-dependent View of Retrieval

An alternative, or rather a complement, to the trace-dependent view of memory retrieval is that of cue dependency. According to this view, retrieval failure is often a consequence of ineffective cueing rather than of a faded memory trace. From this perspective, the problem is one of trace *access*. The trace may be quite intact but without an effective retrieval cue it remains inaccessible, rather like a mistakenly shelved library book. The book is quite intact, but the retrieval code is no longer effective. Similarly, information stored in computer memory is of no use without an effective means of accessing it, and a major research area in computer science is the development of efficient systems of information retrieval. Although such systems have had only limited success as models of human memory retrieval, they do embody one important principle. Memory retrieval depends on the relationship between how a memory is stored and the type of information available at the time of retrieval. Successful retrieval therefore depends not only on the state of the memory trace, but also on the conditions at time of retrieval.

It is not difficult to show that material that resists all efforts at retrieval, and thus gives every appearance of having been forgotten, can subsequently be remembered in response to an appropriate cue. Such findings demonstrate that in many cases inability to retrieve information may be a failure to access information that is potentially available, rather than of forgetting in any permanent sense. This fact accounts for the common experience described as the 'tip of the tongue' state in which people feel they are on the verge of recall. Such states tend to occur in association with attempts to retrieve simple factual information such as the name of a person or place. Formal studies of the phenomenon confirm its validity. When they report being in a tip-of-the-tongue state, participants in experiments can often accurately describe features of the word they cannot quite recall—its first letter, or the number of syllables. A closely related phenomenon that has also been studied experimentally is called 'feeling of knowing.' With better than chance accuracy, participants can predict whether or not they will be able subsequently to recognize an item of information they cannot presently recall. For a more detailed description of these phenomena see Tulving and Craik (2000, Chap. 13).

The success at recovering apparently forgotten memories in response to appropriate cueing has led some to entertain an extreme form of the cue-dependent view of forgetting. The claim is that nothing is ever forgotten but, rather, memories simply become inaccessible. This extreme view has dubious scientific status in that it cannot be falsified by empirical methods. Even the failure of the most exhaustive efforts to cue the lost memory will remain unconvincing to the firm believer: it can always be claimed that such failure simply reflects an inability to find an effective cue. For this reason, and because of its general implausibility, this view has never found much support from within experimental psychology, although it is widely held in other circles (for a review of this point see Loftus and Loftus 1980).

2. The Context-sensitive View of Memory Retrieval

This section develops the view that retrieval is a more complicated process than that assumed by either the trace-dependent or the cue-dependent accounts. The claim is that successful retrieval is a constructive process based on an interactive relationship between features of the trace and those of the retrieval cue. For convenience of exposition a distinction is drawn between retrieval of general knowledge (commonly referred to as semantic memory) and retrieval of personally experienced past events (episodic memory).

2.1 Retrieval from Semantic Memory

What makes a retrieval cue effective? In the case of memory for factual material (semantic memory) the answer is fairly straightforward, at least at a general level. Obviously the cue, if it is to be effective, must be knowledge associated with the material that is to be retrieved, or knowledge from which features of the target material can be inferred. The challenge is to embody this general principle into a more precise description and to this end there have been a number of attempts to describe in detail how such retrieval-by-association might work. The strategy has been to develop a formal description of a person's knowledge and then explain how this structure can be used to retrieve information. The description usually takes the form of a network, typically embodied in a computer program. The nodes of the network may be words or concepts, and the interconnections between nodes represent various forms of relationships among the nodes. Thus, a node denoting 'bird' may be linked to a higher-order node (animal), to an attribute list (has wings, feathers, etc.), and to a list of exemplars (robin, sparrow, etc.). Such networks can retrieve information either directly from the network or by inference. Thus, the fact that a bird has wings may be retrieved directly, but answering the question 'does a bird breathe?' may require an inference because the attribute 'needs to breathe' is connected, not directly to the node 'bird,' but to the higher-order node 'animal.' A powerful feature of such networks is that they can explain how intelligent responses can be made to novel questions (Did Isaac Newton have a telephone?) the answers to which cannot have been learned directly.

Connectionism, or parallel-distributed processing (PDP), is a more recent development (see *Cognition, Distributed*). In a connectionist network, words and concepts are not represented by nodes, but are themselves networks, hence the term *distributed*. A typical network consists of a set of input and output units, between which there may be one or more layers of hidden units. Knowledge (say the concept dog) is represented by the pattern of interconnectedness among these units, and the weight of the interconnections. Stimulation of input units sends excitation through the network. Retrieval consists of the resulting firing of appropriate output units. Although PDP models have scored a number of successes, the extent to which such networks accurately model human acquisition of knowledge and its retrieval remains a controversial matter.

2.2 Retrieval from Episodic Memory

What of the retrieval of personally experienced past events, a form of remembering referred to as episodic memory? How can people retrieve the details of the traffic accident they witnessed yesterday, or how they spent New Year's Eve, 1999? By definition, such experiences represent unique events tied to a specific point in time. General knowledge on the other hand is likely to be used in a variety of contexts that strips its original acquisition of any temporal specificity. Most people know that Paris is the capital of France, for example, without reference to the events through which that knowledge was originally acquired. The major consequence of the difference between these two types of memory is that the retrieval of elements from single events is more strongly influenced by the context within which such elements occurred. To illustrate this point, consider the following example based on an experiment reported by Thomson and Tulving (see Tulving 1983 for a more detailed account). The experiment will be reported in some detail because it embodies a very fundamental principle of memory retrieval.

Suppose the word FLOWER is a sample member of a list of 24 words presented to participants who will be asked subsequently to recall as many of these words as possible. During the initial presentation, some participants see the word FLOWER paired with another word, fruit. The two words of each pair are weakly associated, that is, for example, the word 'FLOWER' is occasionally (about 1 percent of the time) given as a response to the word 'fruit' in normative free association studies. Participants in a second group see FLOWER without any pairing. For the recall test, each of these groups is divided into two; one subgroup receives the word fruit as a cue, the other subgroup receives no cue. Thus, the experiment has four conditions. In one condition the words are presented and tested in isolation. Consider this the baseline condition. In a second condition words are presented with a paired word and this paired word is re-presented as a cue at the time of recall. For the other two conditions the paired word appears at presentation but not at recall, or vice versa. How does recall for the three conditions involving at least one presentation of the weakly associated word compare with the baseline condition? Not surprisingly, the condition in which the associate appears at both study and test produces substantially better recall than the baseline condition.

Even more interesting is that the other two conditions yield equivalent recall levels that are lower than the baseline condition.

Suppose a further condition is added to the experiment. In this new condition the target word is again paired with a weak associate at presentation (fruit—FLOWER), but in the recall phase, this cue is replaced with a strong associate of flower (bloom) not previously presented. A reasonable prediction for this condition is that the presence of the strong associate bloom when recall is being attempted should increase the likelihood of the recall of FLOWER, perhaps to a level higher than for the condition in which the weak associate fruit was present, even although fruit but not bloom was present at the study phase. In fact, exactly the opposite is true. Not only is bloom a much poorer cue than fruit, its presence reduces recall to well below the baseline level. On the other hand, if FLOWER is presented in isolation then bloom is a better cue than fruit.

What can be concluded from this experiment? The experiment demonstrates a principle that is of wide significance and of fundamental importance to an understanding of memory retrieval. In its most general form, this principle states that successful retrieval depends not only on the state of the memory trace, nor only on the nature of the retrieval cue but rather on the interactive relationship between the two. Retrieval will, of course, be influenced both by the form of the initial processing, and by the nature of the available retrieval cues, but neither of these factors considered separately will be sufficient to explain retrieval.

Consider the Thomson and Tulving results in these terms. The event to be recalled is the presentation of the word FLOWER. If one asks if this event will be better remembered if it occurs in relative isolation or in the context of the word bloom, there is no simple answer. It will depend on what cues are available at the time. Similarly if one asks which cue is better, bloom or flower, for retrieving this event, again there is no simple answer. All that can be said is that cue effectiveness depends on the state of the memory trace, which in turn depends on the context within which the event occurred—in isolation, or in the presence of bloom. Memory retrieval is context-sensitive (see *Encoding Specificity in Memory*).

2.3 Some Examples of Context-sensitive Retrieval

This context-sensitive view of memory retrieval has important implications for the understanding of common memory phenomena. In everyday remembering retrieval contexts will differ from one occasion to another, providing a complex and varying set of retrieval cues over which there can be little control. Such contexts may vary greatly in their cue effectiveness depending on their relationship to the conditions under which the original event was experienced. This strong effect of contextual cueing on retrieval offers a more convincing explanation (than the fluctuating trace strength view) as to why a memory that remains stubbornly inaccessible on one occasion readily springs to mind on another.

Even changes in the physical environment are sufficient to influence retrieval. An experiment reported by Godden and Baddeley (see Baddeley 1993) provides an interesting example. These researchers were involved with the training of deep-sea divers. They had divers learn material either while under water or while sitting on land, and then tested them either on land or under water. Their results show a strong context effect in that memory is better if the retrieval and learning environment are the same. Similar results have been obtained by having learning and retrieval occur in different rooms. The practical implications of such result are straightforward and exemplify some general principles. First, retrieval in an environment radically different from that of the original experience may be poorer compared with recall in the same environment. Baddeley describes previous anecdotal reports of the difficulty divers had remembering underwater experiences when being debriefed on the surface. Second, when there is some control over the study environment, that environment should be as similar as possible to the environment in which retrieval will be required.

What is true of changes in the external physical environment is also true of the internal or mental environment, giving rise to a phenomenon known as state dependency. For example, attempting to retrieve material while in a state of marijuana or alcohol intoxication is more likely to succeed if the original event was also experienced in that state, rather than a state of nonintoxication. (It should be noted, however, that memory retrieval is always better when both encoding and retrieval occur in the non-intoxicated state.) Emotionally arousing material shows a similar effect. Thus, people in a depressed mood may find it easier to retrieve unpleasant experiences than pleasant ones, a bias that can only serve to sustain the depressed mood (see *Mood-dependent Memory*).

Such results can be used to develop techniques that might improve a person's ability to retrieve past events. The dependency of retrieval on physical and mental context suggests several possibilities. If someone is attempting retrieval in a physical environment different from that of the original experience (which is probably the usual situation), then retrieval can be improved by having the person mentally place themselves in the original environment (Smith 1979). This reinstatement of context is an important aspect of an interview technique developed by Fisher and Geiselman (1992) to facilitate the recall of eyewitnesses. As an aside, it is interesting to note that this interview technique achieves results that are at least as good as those obtained through the use of hypnosis. Contrary to popular opinion, there is nothing magical

about hypnosis as a means of eliciting memories; insofar as it succeeds, hypnosis exploits the same principles of context reinstatement.

2.4 Failures of Memory

It may seem that human memory retrieval as it has been thus far described is not radically different from information retrieval in computer systems. In both cases success depends on a compatible relation between a system of storage and a method of retrieval. However, there are important differences. Computer retrieval is typically an all-or-none affair; when successful, the information is retrieved with perfect fidelity. In the absence of physical damage or explicit programming instructions, the stored information remains unchanged and is uninfluenced by other input. The downside of this fidelity is the relative inflexibility of what will function as a successful retrieval cue. Human memory retrieval is different. The highly interrelated structure of human memory allows for great flexibility, enabling it to answer all kinds of unanticipated questions. Rather than locating a complete and intact memory trace, a great deal of human memory retrieval (especially episodic memory) is better thought of as a form of pattern completion, rather like the paleontologist's reconstruction of an entire animal from a few fossil bones.

The cost of this flexibility and constructive processing is reduced accuracy. With computer retrieval, the function of the retrieval cue (file name, key word, etc.) is to locate the stored information; neither the retrieval process nor the cue itself modifies this information. With human memory the situation is different. There is extensive evidence that, in the case of human memory, the constructive, pattern-completion process, along with the retrieval cue itself, play a role in forming the content of the reported memory.

Errors attributable to the constructive process usually reflect the misleading influence of existing knowledge. Schacter (1996, p.103) describes the following simple demonstration based on experimental work by Roediger and McDermott (1995). Participants studied a list of 15 words all associated with the word *sweet* (candy, sour, sugar, etc.) although the word sweet itself did not appear. A short time later, they were asked whether the word sweet was in the list. Schacter reports that in demonstrations with very large audiences, 80–90 percent erroneously claim that sweet was part of the list. The compelling aspect of this demonstration is not merely the false recognition of sweet, but that people claim to have a quite vivid memory of the word and are highly confident that it was on the list.

The content of retrieval cues, in the form of questions and hints, can also help determine the content of retrieved memories. In a classic study (Loftus and Palmer 1974), participants saw a film of two cars colliding. Some participants were asked to estimate the speed of the cars when they hit each other. For other participants the word hit was replaced with other words such as smashed, or contacted. This variation produced large differences in estimates, ranging from 40.8 mph for smashed, 34 mph for hit, and 31.8 mph for contacted. The differences in cues produced other effects. For example, participants whose question had used smashed were more likely to report that there had been broken glass although, in fact, there had been none.

The potential distorting effects associated with the constructive nature of retrieval have obvious implications for a wide range of activities ranging from legal eyewitness testimony to psychotherapies that place a strong emphasis on the recollection of past experiences. These implications are often lost to interviewers or therapists who regard memories as akin to books in a library or files in computer memory—a passive entity that once located and activated will emerge fully intact. The reality of memory retrieval is different. On the one hand the flexible constructive nature of memory retrieval does make it possible to recall experiences that at first appear to have been totally forgotten. On the other hand, however, this same property makes such recollections vulnerable to the influences of the particular wording of questions, and the general context (both physical and mental) within which retrieval is attempted. For a more detailed account of these matters see *Eyewitness Memory: Psychological Aspects* and *Reconstructive Memory, Psychology of*.

3. Conclusion

Our understanding of retrieval has advanced considerably beyond the notion that the only determining factor is the strength of the memory trace. It has also become clear that a simple cue-dependency account is inadequate. The memory trace is not a static entity waiting for an appropriate cue to pluck it from some mental shelf. Rather, retrieved memories are the result of a complex interplay of stored information and retrieval cue. The principles governing this interaction are likely to be the major focus of future research. Such research will provide greater understanding as to why human memory, usually so dependable, can sometimes be inaccurate or even totally false. A particular challenge is to understand, not simply why mistaken memories occur, but why such memories are often indistinguishable from those that are genuine, and are believed to be veridical with a high degree of confidence. In this matter behavioral research is being supplemented by brain imaging techniques that promise to identify the brain mechanisms that distinguish false from real memories.

See also: Declarative Memory, Neural Basis of; Encoding Specificity in Memory; Episodic and Autobiographical Memory: Psychological and Neural Aspects; Implicit Learning and Memory: Psychological and Neural Aspects; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Memory, Consolidation of; Memory: Levels of Processing; Memory Models: Quantitative; Memory: Organization and Recall; Memory Problems and Rehabilitation; Memory: Synaptic Mechanisms; Short-term Memory: Psychological and Neural Aspects; Working Memory, Neural Basis of

Bibliography

Baddeley A D 1993 *Your Memory: A User's Guide*. Macmillan, New York

Fisher R P, Geiselman R E 1992 *Memory-enhancing Techniques for Investigative Interviewing: The Cognitive Interview*. Thomas, Springfield, IL

Loftus E F, Loftus G R 1980 On the permanence of stored information in the human brain. *American Psychologist* **35**: 409–20

Loftus E F, Palmer J C 1974 Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior* **13**: 585–89

Roediger H L, McDermott K B 1995 Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 803–14

Schacter D L 1996 *Searching for Memory*. Basic Books, New York

Smith S M 1979 Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory* **5**: 460–71

Tulving E 1983 *Elements of Episodic Memory*. Oxford University Press, New York

Tulving E, Craik F I M (eds.) 2000 *The Oxford Handbook of Memory*. Oxford University Press, New York

R. S. Lockhart

Memory: Synaptic Mechanisms

Memory has fascinated Western thinkers since the era of the pre-Socratic philosophers. With time, however, the focus of interest in memory has shifted first from philosophy to psychology and, more recently, from psychology to neural science. The reason that memory has so occupied scholarship is that memory is a bridge between the humanities, concerned with the nature of human experience, and the natural sciences, concerned with the mechanistic basis of that experience. Memory is the glue that binds mental processes together across time and that gives continuity and coherence to our lives. Many of the characteristics that make each of us unique arise from the enduring effects of our individual experience.

The shift from philosophical to empirical investigation into memory occurred in the nineteenth century and involved three fundamental distinctions. At the end of the nineteenth century, as academics gained increasing faith in the scope and power of empirical scientific research, they became emboldened to apply experimental techniques to mental processes such as learning and memory. The first steps were taken in 1885 by the German philosopher Hermann Ebbinghaus, who transformed philosophical speculation about memory into an experimental science. Ebbinghaus was much influenced by the work of Weber, Fechner, and Wundt in the psychophysical study of sensation. These scientists showed that one could apply objective experimental techniques to the study of a behavioral process. Although the measurement was the subject's subjective response, these responses proved to be quite reliable when the probe—the stimulus used to elicit the response—was objective and quantifiable.

In trying to develop a probe for memory, Ebbinghaus hit upon the use of three-letter nonsense words that had no relation to any language, thereby preventing previous associations or experience from affecting the process of learning and recall. By memorizing syllable lists of varying length and by testing his recall at different points in time, Ebbinghaus was able to deduce two important principles about memory storage. First, he found that memory is graded—that practice makes perfect. There was a linear relationship between the number of training repetitions and the extent of recall on the following day. Second, Ebbinghaus anticipated the distinction between short and long-term memory that has dominated modern thinking by noting that whereas a list of six or seven items could be learned and retained in only one presentation, longer lists required repeated presentation. The same distinction between short and long-term memory was apparent when Ebbinghaus plotted a 'forgetting curve,' which he found to be biphasic: he rapidly forgot information during the hour after training, but forgetting was much more gradual after that, out to periods as long as a month.

At the beginning of the twentieth century a second important distinction was introduced into thinking about memory by the theorist Richard Semon. Semon divided memory into three components: (a) encoding or acquisition of new information, (b) storage of that information over time (in the form of what he termed the engram), and (c) retrieval or decoding of the information in a behaviorally conducive context (Semon 1909–1923). Of these three components, the engram, or the storage mechanism, has proven most amenable to a biological line of inquiry.

A third distinction—not between temporal phases or components of a given memory but between

different types of memory and the brain systems that mediate them—emerged only in the latter half of the twentieth century. Studies by Brenda Milner and others of patients with lesions of the brain in the hippocampal formation and medial temporal lobe have made it clear that memory is not a unitary faculty of mind but can be divided into at least two functionally distinct forms: explicit (or declarative) memory which, at least in humans, involves conscious recall and can be put into words; and implicit (nondeclarative or procedural) memory, which is revealed by a lasting, unconscious change in behavior (Milner et al. 1998).

Despite the existence of these two fundamentally different forms of memory, both Ebbinghaus' distinction between short and long-term memory and Semon's distinction between encoding, storage, and retrieval are quite general and apply to both. The differences between implicit and explicit memory arise from the way the molecular mechanisms for short and long-term memory storage are embedded in the brain systems that mediate the different forms of memory.

Here we will focus primarily on the mechanistic aspects of memory, and specifically on the mechanisms of storage. The different functional forms of memory are addressed elsewhere in this *Encyclopedia*.

1. *How Memory is Stored: The Cellular Mechanisms of the Engram*

In the early part of the twentieth century, Santiago Ramon y Cajal, the great Spanish neuroanatomist, introduced a theory about memory storage that is now almost universally accepted. Cajal proposed that synapses, the connections between neurons, are plastic—that is, they can change over time and are sensitive to their history. As a result, information can be stored in the brain by changing the strength of pre-existing synaptic connections between neurons (Ramon y Cajal 1893). Cajal further suggested that the change in the strength of connections between neurons results from anatomical changes in the shape of individual dendritic spines, the protuberances from the dendrites of excitatory neurons on which many synapses form.

In the subsequent decades there have been two major elaborations of Cajal's ideas, the first by Donald Hebb in 1949 and the second by Eric Kandel and Ladislau Tauc in 1965. In his oft-cited book, Hebb (1949) proposed a homosynaptic rule for strengthening synaptic connections, according to which the events that trigger synaptic strengthening occur at the strengthened synapse itself. Hebb hypothesized that 'when an action of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of

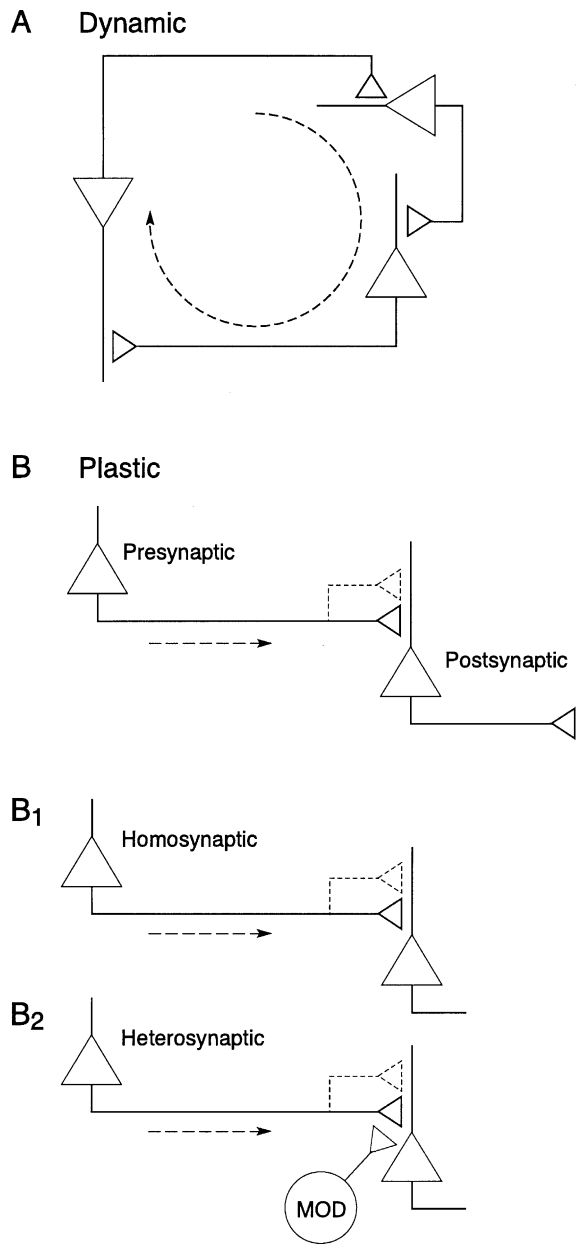

Figure 1
Models of information storage in the brain. (**A**) Dynamic storage in a reverberating circuit. (**B**) Synaptic plasticity, proposed by Ramon y Cajal. Synaptic plasticity can be homosynaptic. (**B1**), in which patterned or repetitive synaptic activity alters the strength of that synapse; or heterosynaptic (**B2**), in which a modulatory interneuron alters synaptic strength. See text for details

the cells firing B, is increased.' Because the strength of the connection between a pre- and postsynaptic neuron is increased when the firing of the postsynaptic neuron is correlated or associated with the firing of the presynaptic neuron, this sort of synaptic strengthening has been termed associative.

After such an event, when the first of the two neurons is activated (perhaps by a stimulus that resembles the one to which it fired earlier) it has an increased chance of leading to the firing of the second. In addition to being homosynaptic and associative, Hebb proposed that the synaptic strengthening be input-specific: when two neurons fire coincidentally the synapse between them should be strengthened, but other synapses on either neuron should remain unchanged.

Kandel and Tauc proposed in 1965 an additional, heterosynaptic rule for strengthening synaptic connections. They further proposed that this heterosynaptic strengthening could be of two sorts, nonassociative or associative. In nonassociative heterosynaptic facilitation, a synapse could be strengthened without any change in the firing of either the presynaptic or the postsynaptic neuron. This occurs as a result of the firing of yet a third neuron, a modulatory interneuron, that acts on the synapse to increase its strength. In associative heterosynaptic facilitation, the strengthening effect of the modulatory neuron is further enhanced when the firing of the modulatory input is associated in time with the firing of the presynaptic neuron. Kandel and Tauc demonstrated homosynaptic and both types of heterosynaptic plasticity in the mollusc *Aplysia* by manipulating one and two inputs to a single cell (Kandel and Tauc (1965).

The plastic change hypothesis of Cajal and its elaboration by Hebb and by Kandel and Tauc, however, represent only one of two competing theories of the nature of the engram. An alternative to plastic change at the synapse as the principle for information storage was the idea of dynamic change, advanced by Alexander Forbes and Lorente de Nó. According to their idea, information could be stored in neural circuits without altering the synapses as a result of the firing of reverberant networks of neurons or pairs of mutually excitatory cells (Delisle Burns 1958).

Tests of these competing ideas in the 1960s showed that it was difficult to establish reverberating activity in the brain because of the abundance of inhibitory interneurons that prevent the continuous cycling of activity in practically every neural circuit. By contrast, exploration of synapses showed that most chemical synapses in the vertebrate brain as well as in invertebrates have remarkable plastic properties. Beginning in 1965 with the developmental studies of Hubel and Wiesel and the studies of Kandel and Tauc, and the work of Bliss and Lømo (to which we will return below) several years later, a variety of forms of synaptic plasticity were demonstrated in different neuronal systems. It soon became clear that built into the molecular architecture of many chemical synapses is a remarkable capacity for modification.

2. *Homo- and Heterosynaptic Plasticity are Recruited by Learning and Serve for Memory Storage*

Showing that chemical synapses are plastic was one thing; showing that such plastic changes are induced in a behavioral context by learning is another. The first rigorous experiment designed to explore whether or not plastic mechanisms are induced by learning was carried out in 1970 by Vincent Castellucci, Irving Kupferman and their colleagues in *Aplysia* (Castellucci et al. 1970). They identified a neural circuit mediating a simple reflex—the gill withdrawal reflex to stimulation of the siphon—and showed that it can be modified by two simple forms of learning, habituation and sensitization. In habituation, repeated presentation of a novel stimulus leads to a gradual decrease in the animal's reflex withdrawal response as it learns that the stimulus is innocuous. Habituation of this gill-withdrawal reflex was accompanied by a homosynaptic decrease in the strength of the connection between the siphon sensory neuron and the gill motor neuron in the circuit. Sensitization is a form of learned fear in which the animal recognizes a stimulus, such as a shock to the tail, as being aversive and learns to enhance its gill withdrawal response to a previously neutral stimulus, such as a weak tactile stimulus to the siphon. Sensitization of the gill withdrawal reflex was accompanied by a nonassociative, heterosynaptic increase in the strength of the same synaptic connection between sensory and motor cells, brought about by the activity in modulatory interneurons induced by the noxious tail stimulus. Later studies at the same synapse by Thomas Carew, Robert Hawkins, Tom Abrams and their colleagues demonstrated that classical conditioning gives rise to associative heterosynaptic facilitation. These several studies show that multiple mechanisms of plasticity can coexist at a single synapse and can be recruited by different forms of learning.

In 1973, Bliss and Lømo described a long-lasting Hebbian form of homosynaptic plasticity in the mammalian brain with potential relevance for explicit forms of memory (Bliss and Lømo 1973). They found that when the perforant path, a fiber track in the hippocampal formation in the temporal lobe, was repetitively stimulated in an anaesthetized animal, subsequent responses in the dentate gyrus to single stimuli were enhanced: the perforant path synapse had been strengthened. This result was important for several reasons. First, it showed that an enduring form of plasticity exists in the adult mammalian brain. Second, the hippocampus had been clearly implicated in work by Penfield and Perot (1963) and by Scoville

and Milner (1957) to be involved in human explicit memory, so finding plasticity here was particularly intriguing. Third, the plasticity was homosynaptic. Further investigation of this form of plasticity (generically termed Long-Term Potentiation, or LTP) showed that in most cases it is dependent on coincident pre- and postsynaptic firing and is input-specific. It thus has all the characteristics of a Hebb synapse.

More recent studies have revealed that the distinction between heterosynaptic and homosynaptic mechanisms of facilitation is not absolute and that both can commonly occur at a plastic synapse. For example, detailed studies of classical conditioning in *Aplysia* by Glanzman, Hawkins and their colleagues revealed both associative heterosynaptic and homosynaptic mechanisms; interfering with either can disrupt synaptic plasticity (Bao et al. 1997). In the hippocampus, LTP can be induced in a purely homosynaptic way, without heterosynaptic participation. But for the potentiation to persist for more than a few hours in duration, heterosynaptic modulatory processes seem to be required (Bailey et al. 2000). Such observations suggest that memory storage often requires both forms of plasticity. Whereas homosynaptic plasticity can serve as a learning mechanism and as a mechanism of short-term memory, heterosynaptic mechanisms are often required for persistence of long-term memory. This combinatorial utilization of homo and heterosnaptic plasticity is emerging as a major point in the study of synaptic plasticity, to which we will return below.

3. *Molecular Mechanisms of Plasticity in Specific Model Systems*

The experimental elimination of dynamic changes as a major contributor to memory storage focused attention on the chemical synapse as a site of information storage in the brain, a focus that allowed a concerted attack on its molecular mechanisms. As a result of the classic early work of Bernard Katz and Paul Fatt and a long list of others more recently, we now know a great deal about the cell and molecular mechanisms of synaptic transmission. This large body of knowledge is now being brought to bear on the synaptic mechanisms of memory storage and represents the starting point of all subsequent studies of the mechanisms of plastic change.

The molecular mechanisms of plasticity and the relationship of plasticity to simple forms of memory were first studied in *Aplysia* (Kandel 1976). Studies of the gill withdrawal reflex in the intact animal and studies of components of the neural circuit in cell culture have revealed several mechanistic themes that seem to be of general importance. First, just as memory has at least two temporal phases, many forms of synaptic plasticity have an early and a late phase. These phases are pharmacologically dissociable: the early phase does not require new protein synthesis, whereas the late phase is blocked by inhibitors of either protein or mRNA synthesis. This implies that the induction of synaptic plasticity recruits two distinct signaling pathways, one of which operates rapidly at the synapse and the other of which operates more slowly and transmits a signal to the nucleus, where it induces the activation of new genes required for long-lasting plastic change. These two temporal phases of synaptic plasticity seem to correspond to the two temporal phases of memory described by Ebbinghaus.

The early phase of heterosynaptic plasticity in *Aplysia* is initiated by tail stimuli that activate modulatory interneurons, some of which release the neurotransmitter serotonin. This serotonin activates serotonergic receptors in the sensory neurons of the reflex, which in turn activate an enzyme, adenylyl cyclase, that increases the amount of the small intracellular signaling molecule cAMP in the cell. cAMP leads to early-phase enhancement of synaptic transmission both by alteration of the electrical properties of the sensory cell and by changes in the synaptic machinery. Repeated activation of the modulatory interneurons initiates the same category of events but also leads to late-phase plasticity, which can last days and is characterized by demonstrable structural alterations at the synapse. Stimuli that induce late-phase plasticity lead to the translocation of several different signaling molecules to the nucleus of the presynaptic cell (Bailey et al. 1996).

Despite the involvement of events in the nucleus, late-phase plasticity in *Aplysia* and in other systems retains its synapse specificity. This implies that the products of the new genes that are activated during the induction of enduring plasticity are somehow specifically targeted to the relevant synapse(s). The leading hypothesis for this phenomenon at the beginning of the twenty-first century involves synaptic tagging: the synapses where plasticity is induced are somehow marked such that the products of the relevant induced genes are effective only at those synapses. A consequence of this hypothesis is the phenomenon of synaptic capture, in which weak stimulation of one synapse can lead to lasting potentiation if the stimulation occurs during a window after strong stimulation of a separate synapse on the same cell. Synaptic capture has been observed in *Aplysia* by Kelsey Martin and her colleagues (Martin et al. 1997) and in mice by Uwe Frey and Richard Morris (Frey and Morris 1997). Since synaptic capture in either homosynaptic or heterosynaptic plasticity represents a specific breakdown of synapse specificity, synaptic capture may have significant consequences for the encoding and storage of information, but the role of this phenomenon remains unclear.

Parallel work has been carried out in the fruit fly *Drosophila* by Ron Davis, Chip Quinn, Tim Tully, and their colleagues. Genetic screens have revealed that

many of the molecules implicated in synaptic plasticity in *Aplysia* are also involved in learning in *Drosophila*. For example, modulatory neurotransmitter receptors and components of the cAMP second messenger system have been associated with learning in this organism. Such results suggest that not only the general principles but also some of the details of the mechanisms that emerge from study of an organism such as *Aplysia* may be very generally applicable (Davis 1996). Synaptic plasticity in mammals is best characterized in the hippocampus. The most work has been done at the Schaeffer collateral synapse, distinct from the perforant path synapse first studied by Bliss and Lømo (1973). At the Schaeffer collateral synapse, LTP can be induced by various patterns of repetitive stimulation of the axons coming into the synapse or by simultaneous brief stimulation of the axons and depolarization of the postsynaptic cell; both instances conform to Hebb's requirements. Hippocampal LTP has an early and a late phase; the intracellular messenger cAMP seems to be critical for the late phase but less so for the early, which is induced by calcium influx. Synapse specificity and synaptic capture have been demonstrated. There remains considerable controversy as to the precise molecular changes that accompany plasticity at this synapse; it is likely that changes in both the presynaptic and the postsynaptic cell contribute, with the balance of contributors shifting depending on the specifics of the inducing stimulus.

While this form of homosynaptic plasticity has been well studied, there is considerable evidence that heterosynaptic plasticity also has an important role at the same synapses. Intracellular cAMP can be produced either by appropriate patterns of calcium buildup or by modulatory neurotransmitters in a heterosynaptic fashion, and there is good evidence that both contribute. As we note above, modulatory neurotransmitters can enhance the capacity for plasticity or even induce it in the absence of neuronal stimulation, and blockade of modulatory neurotransmitters blocks the late phase of LTP at most synapses in the hippocampus and amygdala. Blockade of modulatory neurotransmitters in intact animals and in people can interfere with the formation of memories under certain circumstances (Cahill et al. 1994). Such observations once again suggest that homosynaptic and heterosynaptic plasticity interact importantly at a variety of synapses in the hippocampus and elsewhere in the formation of memories.

Plasticity has been described in many other places in the mammalian brain, most of which are described in detail elsewhere in this *Encyclopedia*. In particular, LTP has been described at the other synapses in the hippocampus, in the amygdala, the neocortex, and the striatum. A conjugate phenomenon, long-term depression (LTD), has also been described in many of these structures as well as in the cerebellum. In many of these cases the characteristics emphasized above are retained: early and late phases of plasticity, a critical role for calcium influx, and the participation of second messengers like cAMP. It seems likely that a set of common molecular themes exists for most or all forms of synaptic change.

4. *Correlating Synaptic Change with Behavior*

The postulate that synaptic change underlies various forms of learning in invertebrates and in mammals is difficult to demonstrate in more than a correlative way. In many cases, manipulations that disrupt synaptic plasticity also disrupt related forms of learning. This is particularly so in the correlation between LTP at the Schaeffer collateral synapse in the hippocampus and forms of complex, hippocampus-dependent spatial learning in rodents (Martin et al. 2000). While such correlations become impressive as their number increases, there are some documented exceptions. Furthermore, the manipulations involved are rarely definitive. Whether by lesion, drug application, or genetic change, any manipulation that affects synaptic plasticity is likely to also have subtle effects on normal synaptic function and on the properties of neurons away from the synapse specifically being studied.

Such correlational studies suggest that synaptic plasticity is necessary for at least some forms of learning. Two other modes of investigation of the connection between synaptic strengthening and learning are required to substantiate the hypothesis. First, specific changes in the strength of a defined synapse or set of synapses should be able to mimic learning and thus to have a defined effect on the animal's behavior—that is, induction of plasticity should be sufficient for learning. Second, learning in a normal animal should in principle lead to changes in the strength of some specific set of synapses. These two types of experiment have been more challenging than simple correlation of synaptic plasticity with capacity for learning.

The first approach, changing behavior in a way that resembles learning by altering the strength of a synapse, has best been attempted in the cerebellum. As described elsewhere in this *Encyclopedia*, the cerebellum is involved in learning of coordinated motor patterns and in the acquisition of certain classically conditioned responses to arbitrary stimuli. The core of the cerebellum's neural architecture is the Purkinje cell, which receives strong input from a single climbing fiber(or a few of them) and weak input from as many as 100,000 parallel fibers. The Purkinje cell is thus ideally suited to learn associations between a single, specific stimulus (represented by the climbing fiber) and any of a large number of arbitrary stimuli (represented by the parallel fibers). Learning in this circuit is thought to be mediated in large part by alteration in the strength of the synapses between the

parallel fibers and the Purkinje cell, an alteration which is controlled by the relative timing of climbing fiber and mossy fiber activation.

In an early, pioneering study, Brogden and Gantt showed that (a) when they stimulated the cerebellar white matter they could sometimes produce an identifiable motor output, and (b) when they paired this stimulation with a tone, that tone came to induce the motor output independent of neural stimulation (Krupa et al. 1993). This means that the direct stimulation of the cerebellar white matter can substitute for the unconditioned stimulus in a classical conditioning paradigm. In an elegant series of studies using electrophysiological recording, lesions, stimulation, and reversible inactivation, Thompson and his colleagues have identified the climbing fiber of the cerebellum as the critical white matter component that can encode the unconditioned stimulus and they have developed very strong evidence that the essential memory trace for certain forms of classical conditioning is indeed formed and stored in the cerebellum (Krupa et al. 1993). While such studies do not specifically demonstrate a change in synaptic strength, they do demonstrate that stimulation conditions known to induce changes in synaptic strength can mimic one or both of the stimuli in a learning paradigm.

The most difficult aspect of the connection between synaptic plasticity and learning is the demonstration of changes in synaptic strength after demonstrable learning has occurred in a normal animal. This has been demonstrated in *Aplysia* and in the rat amygdala. In *Aplysia*, sensitization, habituation, and classical conditioning, three simple learning processes by which the reflex response to a light touch is modulated, have been shown to result in changes in the functional strength of a specific synapse in the circuit that underlies the reflex (Kandel 1976, Murphy and Glanzman 1997). Such results greatly strengthen the idea that synaptic change underlies learning in this organism.

The rat amygdala is required for fear conditioning to a tone—that is, for learning to associate a previously neutral audible tone with a foot shock. Part of the circuit underlying this learning phenomenon consists of the projections from the auditory portion of the thalamus, which relays auditory information about the tone, to the lateral subnucleus of the amygdala. There is some controversy as to whether the amygdala then stores this information or simply plays a critical role in orchestrating the storage of the tone-shock pairing in other structures (Cahill et al. 1999). If the former is the case, the synaptic plasticity concept would predict a measurable change in synaptic strength in the amygdala after fear conditioning. Joseph LeDoux and Michael Rogan have demonstrated LTP *in vivo* at the synapse between the thalamic projections and the lateral amygdala, and, importantly, they have demonstrated change at this synapse after a specific tone-shock pairing (Rogan et al. 1997). This result lends critical support to the notion that learning leads to plasticity of specific synapses.

5. *Top-down and Bottom-up Approaches to Learning and Plasticity*

It is not coincidental that the clearest demonstrations of a connection between learning and plasticity come from *Aplysia* and from the amygdala, rather than from the intensively studied hippocampus. The *Aplysia* reflex circuit and the amygdala fear conditioning circuit represent simpler neural systems in which to study this connection, because the circuit that mediates learning can (at least in principle) be clearly identified. This allows one to start with a defined learning behavior, delineate the circuit that mediates it, and thus identify a relatively small number of synapses at which the relevant plasticity may be occurring. By examining only these synapses, one can maximize the chance of finding meaningful synaptic change. In brief, in simple systems, it is possible to identify where to look. This sort of approach, where the behavior comes first and the synapses are examined later, can be called a top-down approach to the study of learning.

The case is quite different in the hippocampus. In this case, plasticity was described first (albeit in a structure known to be required for certain forms of learning), and behavioral tasks were tested later to look for correlates to disruptions of that plasticity: the approach is bottom-up. Furthermore, the type of learning task in which the hippocampus is involved (explicit or spatial learning) is vastly more complicated than the simple associations studied in *Aplysia* and in the amygdala, and the representation of information in the hippocampus is likely to be more abstract. While the Schaeffer collateral synapse is often referred to as a single site, it actually represents approximately 1,000,000,000 individual synapses from cells in the CA3 field of the hippocampus on to cells in the CA1 field. Even if the idea that plasticity at this synapse is important for hippocampus-dependent learning is correct, only a small subset of these synapses is likely to be meaningfully altered in a given learning task: it is not remotely clear where to look for the relevant changes. For this reason, compelling demonstration of the role hippocampal synaptic plasticity in learning has proven difficult to achieve.

Plasticity in the cerebellum is perhaps an intermediate case. The cerebellum's role in learning was described before plasticity, and we have a fairly clear model of how information is encoded: in a classical conditioning task, the climbing fiber encodes the unconditioned stimulus, the parallel fiber encodes the conditioned stimulus, and the Purkinje cell output modulates a motor response. This gives one an idea of where to look for meaningful synaptic change. How-

ever, there are millions of Purkinje cells, and each receives as many as 200,000 synapses from individual parallel fibers. While there is strong evidence that change at some parallel fiber-Purkinje cell synapses is important, identifying the specific site (or sites) where plasticity occurs in a given learning task is a daunting task indeed and may be no easier than in the hippocampus.

6. A Broader View

We have emphasized the two central models of how synaptic plasticity may store information: homosynaptic plasticity and heterosynaptic plasticity. We have also described evidence that the two may operate both independently and in tandem at certain important synapses in the mammalian brain. A critical current question in the biology of synaptic plasticity is the relation between the two. While homosynaptic plasticity seems to have a greater capacity for information storage because of its greater specificity, heterosynaptic plasticity, at least in most systems, leads to more long-lasting change. One clue to their relationship comes from studies of the role of modulatory neurotransmitters in the hippocampus. In general, modulatory neurotransmitters have an effect on the late phase of plasticity but not on the early phase: enhancing modulatory transmitters can increase the capacity for long-lasting change, and blocking them can truncate potentiation to just the early phase. One possibility is that whereas homosynaptic plasticity, with all its potential for specificity of connections, contributes primarily to learning and short-term memory, heterosynaptic plasticity determines what information is destined to enter into long-term storage: it is the mechanism of memory.

We began this article with the suggestion that memory has become a particularly attractive field for the biologically minded neuroscientist because the means of information storage have proven amenable to analysis even without an understanding of the very complicated processes by which behavioral information is encoded into memory. The concept that information can be stored in the changing strength of individual synapses is an entrée into this biological mechanism. However, as we have seen, the complexity of information encoding cannot be escaped for long: it is only in *Aplysia* and in maximally simplified mammalian systems like the amygdala that it is at all clear where to look for the plasticity that accompanies learning. Correlative studies can increase our confidence that forms of synaptic plasticity in the hippocampus and elsewhere do matter for learning, but a definitive demonstration will have to wait for a fuller understanding of the more abstract encoding of memory in these circuits.

See also: Amygdala (Amygdaloid Complex); Cerebellum: Associative Learning; Classical Conditioning, Neural Basis of; Fear Conditioning; Long-term Depression (Cerebellum); Long-term Potentiation (Hippocampus); Memory, Consolidation of

Bibliography

Andersen P 1977 Long-lasting facilitation of synaptic transmission. *Ciba Foundation Symposia* **58**: 87–108

Bailey C H, Bartsch D, Kandel E R 1996 Toward a molecular definition of long-term memory storage. *Proceedings of the National Academy of Science USA* **93**: 13445–13452

Bailey C H, Giustetto M, Huang Y Y, Hawkins R D, Kandel E R 2000 Is heterosynaptic modulation essential for stabilizing Hebbian plasticity and memory? *Nature Reviews Neuroscience* **1**: 11–20

Bao J-X, Kandel E R, Hawkins R D 1997 Involvement of pre- and postsynaptic mechanisms in posttetanic potentiation at *Aplysia* synapses. *Science* **275**: 969–973

Bliss T V, Lømo T 1973 Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* **232**: 331–356

Cahill L, Prins B, Weber M, McGaugh J L 1994 Beta-adrenergic activation and memory for emotional events. *Nature* **371**: 702–704

Cahill L, Weinberger N M, Roozendaal B, McGaugh J L 1999 Is the amygdala a locus of 'conditioned fear'? Some questions and caveats. *Neuron* **23**: 227–228

Castellucci V, Pinsker H, Kupfermann I, Kandel E R 1970 Neuronal mechanisms of habituation and dishabituation of the gill-withdrawal reflex in *Aplysia*. *Science* **167**: 1745–1748

Davis R L 1996 Physiology and biochemistry of Drosophila learning mutants. *Physiological Reviews* **76**: 299–317

Delisle Burns B 1958 *The Mammalian Cerebral Cortex*. Arnold, London

Frey U, Morris R G 1997 Synaptic tagging and long-term potentiation. *Nature* **385**: 533–6

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Kandel E R 1976 *The Synaptic Basis of Behavior*. W H Freeman & Co, San Francisco, CA

Kandel E R, Tauc L 1965 Heterosynaptic facilitation in neurones of the abdominal ganglion of *Aplysia depilans*. *Journal of Physiology* **181**: 1–27

Krupa D J, Thompson J K, Thompson R F 1993 Localization of a memory trace in the mammalian brain. *Science* **260**: 989–991

Martin K C, Casadio A, Zhu H E Y, Rose J C, Chen M, Bailey C H, Kandel E R 1997 Synapse-specific, long-term facilitation of *Aplysia* sensory to motor synapses: A function for local protein synthesis in memory storage. *Cell* **91**: 927–38

Martin S J, Grimwood P D, Morris R G 2000 Synaptic plasticity and memory: An evaluation of the hypothesis. *Annual Review of Neuroscience* **23**: 649–711

Milner B, Squire L R, Kandel E R 1998 Cognitive neuroscience and the study of memory. *Neuron* **20**: 445–68

Murphy G G, Glanzman D L 1997 Mediation of classical conditioning in Aplysia californica by long-term potentiation of sensorimotor synapses. *Science* **268**: 467–471

Penfield W, Perot P 1963 The brain's record of auditory and visual experience. *Brain* **86**: 595–696

Ramon Y, Cajal S 1893 *Neue Darstellung vom Histologischen Bau des Centralnervensystem. Archiv für Anatomie und Entwicklungsgeschichte* 319–428

Rogan M T, Staubli U V, LeDoux J E 1997 Fear conditioning induces associative long-term potentiation in the amygdala. *Nature* **390**: 604–607

Scoville W B, Milner B 1957 Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry* **20**: 11–21

Semon R 1909–1923 *Mnemic Psychology*. George Allen and Unwin, London

C. Pittenger and E. Kandel

Men's Health

The term 'men's health' is used here to refer both to the physical and mental health problems that are of concern for men and to health differentials among men. Moreover, when one speaks of 'men's health' one also draws attention to differences in the health and health care needs of boys and men as compared to girls and women. These differences extend far beyond the obvious differences in the reproductive systems. Men in most developed countries suffer more severe chronic conditions than women. They also have higher death rates for most leading causes of death and die about six years younger than women on average. Biological and social factors contribute to gender differences in health. From a biological perspective, these gender differences can be attributed to anatomical and physiological differences between men and women. Health behaviors are important factors influencing health and longevity, and men are more likely than women to engage in behaviors that increase the risk of disease and death. A social constructivist approach argues that the practices that undermine men's health are often signifiers of masculinity and instruments men use in the negotiation of social power and status. Because most of the research on men's health has been done in developed countries, this review is strongly biased in that direction.

1. *Mortality*

Life expectancy is a synthetic measure of current mortality conditions in a particular year, and it is widely used as a general indicator of health and mortality (see *Life Expectancy and Adult Mortality in Industrialized Countries*). International comparisons (World Health Organization 2000) indicate that in 1999 men's life expectancy at birth was highest in Japan with 77.6 years, followed by Sweden (77.1 years), and Australia (76.8 years). Life expectancy was lowest in Sierra Leone (33.2 years), Niger (37.2 years), and Zambia (38.0 years). Men's life expectancy at birth rose dramatically during the twentieth century in developed countries. For example, from 1900 to 1995, men's life expectancy increased by 30 years in France, 25 years in Sweden, and 26 years in the USA. Progress in the area of mortality was evident also at older ages. Men's remaining life expectancy at age 80 almost doubled—from about four years in 1900 to about seven years in 1995. Improvements in survival at older ages have been more pronounced in recent decades than they were in earlier ones.

Although these secular improvements are impressive, they lag behind the improvements observed for women. As a consequence the gender gap in mortality, which favors women, widened during the twentieth century. In developed countries today men die about six years earlier than women on average. Gender disparities in mortality vary considerably in magnitude across countries, with gender differences in life expectancy as high as 11.3 years in the Russian Federation and 11.1 years in Kazakhstan (World Health Organization 2000). Also in most developing countries men have a lower life expectancy than women. However, the gender differences in these countries are somewhat smaller in magnitude. Gender differences in life expectancy are very small or even reversed in countries that exhibit pronounced discrimination against women, such as India.

The mortality disadvantage of men is present at all ages. There is some evidence that this male disadvantage starts even before birth. The sex ratio at conception is unknown. More than two-thirds of prenatal mortality occurs before pregnancies are clinically recognized, and the sex ratios for those very early deaths are also unknown. Evidence reviewed by Waldron (1998) suggests that between the second and fifth months of pregnancy male embryos have higher mortality risk than female embryos. Data for developed countries show that males had higher rates of late fetal mortality than females in the early and mid-century but that those sex differences decreased late in the twentieth century. No significant sex differences in late fetal mortality risk have been observed in recent data for a number of developed countries.

At the start of the twenty-first century the sex ratio at birth (boys/girls) in developed countries varies between 1.04 and 1.07. This ratio is elevated in countries with a strong son preference such as China (with sex ratios of 1.11 to 1.13 in the late 1980s). Infant and childhood mortality is higher for boys than for girls and these higher death rates for males continue throughout the entire life span. Higher male death rates at all ages translate into a population sex ratio that is increasingly unbalanced with age. Among octogenarians in developed countries there are more than twice as many women as men, and among centenarians women outnumber men by a factor of 4 to 6.

Consistently higher male mortality across ages and countries has led some authors to conclude that men are inherently more fragile than women for biological reasons. Sex hormones have been identified as making

a major contribution to the gender gap in mortality. These hormones modulate the cholesterol-carrying lipoprotein patterns, and women have a healthier lipoprotein pattern than men (Hazzard 1986). It has also been suggested that there is an advantage to having two X-chromosomes (Christensen et al. 2000). Sexual selection theory, a part of evolutionary theory, has been used to explain the origins of sex differences in mortality. In humans and some other species, females typically make a greater parental investment in their offspring than do males. For a woman, this investment includes a nine-month gestation period, which is followed by lactation and much subsequent nurture. Evolutionary psychologists argue that sex differences in parental investment favor different reproductive strategies for men and women and, consequently, that different traits were selected for in men and women. The greater female parental investment becomes a resource for which males compete and which thus limits their fitness. Wang and Hertwig (1999) argued that reproductive success and personal survival tend to be antagonistic goals in human males, because males' design for personal survival is compromised by the requirements for achieving success in intrasexual competition. In contrast, reproductive success and personal survival tend to be interdependent goals in human females, because the mother's presence is critical for the survival of the child. This argument suggests that evolution has favored traits that increase reproduction in males and traits that increase survival in females. Sexual selection theory is consistent with the female survival advantage that is evident across ages and countries. However, sexual selection theory alone cannot explain why gender differences in mortality increased so noticeably during the twentieth century. In addition to biological factors, social and behavioral influences are clearly important determinants of men's health and of gender differences in death rates.

2. *Morbidity and Disease*

The leading cause of male deaths in developed countries is heart disease, followed by cancer, accidents, and cerebrovascular disease. The leading causes of death differ by age. In the USA in 1998, accidents were the leading cause of death for boys and men aged 1 to 44 (Murphy 2000). Cancer was the leading cause of death for those aged 45 to 64, while heart disease was the leading cause for those aged 65 and older. These age-related patterns were similar for women and men. For each of the 10 leading causes of death, however, age-adjusted death rates were higher for men than for women. The greatest gender difference in age-adjusted death rates was for suicides, with a mortality sex ratio of 4.3 male deaths to 1 female death. The smallest gender differences were for stroke and hypertension, each with a ratio of 1.1 male deaths to 1 female death.

It has long been believed in social epidemiology that men die earlier than women, but that women have poorer health than men. This paradigm of higher morbidity for women has recently been challenged. MacIntyre et al. (1996) examined two large British data sets and concluded that the magnitude and direction of gender differences in morbidity vary according to the condition in question and to the phase of the life span. Excess female morbidity was found consistently across the life span only for psychological distress, and it was far less apparent, if not reversed, for a number of physical symptoms and conditions. In a similar vein, Verbrugge (1989) pointed out that women's excess morbidity tends to be limited to symptoms and less serious conditions, while men have higher prevalence rates for heart disease, atherosclerosis, emphysema, and other fatal conditions.

Men's lower rates of anxiety, depression, and general emotional malaise seem to be a consistent, international phenomenon. Courtenay (2000b) suggested that denial of depression is one of the means men use to demonstrate their manhood and to avoid relegation to a lower status position relative to women and other men. A consequence of such denial may be a form of unhappiness that finds expression in higher drug use and alcohol consumption (Schofield et al. 2000).

3. *Health Disparities Among Men*

There are pronounced differences in men's health and life expectancy across countries. Boys born in affluent societies, such as Japan and the USA, can expect to live more than 30 years longer than boys in the poor countries of sub-Saharan Africa. There are also considerable health disparities within developed regions. Of particular concern is the recent mortality crisis affecting men in the former Soviet republics. From 1992 to 1994, the life expectancy of Russian men dropped by 6.1 years, and the gender difference in life expectancy increased to an astounding 13.6 years (Shkolnikov et al. 1998). Efforts to detect the underlying causes of the crisis suggest that the mortality upsurge cannot be explained by the collapse of the health care system or environmental pollution. Instead, psychological stress caused by the shock of an abrupt and severe economic transition is likely to have played a major role, perhaps mediated by the adverse health effects of excessive alcohol consumption. Similar patterns are also evident in the former Soviet republics other than Russia, although they are less pronounced there. These patterns provide a telling example of the profound effects that psychological and behavioral factors can have on men's health.

Particular groups of men and boys within a society have elevated health risks. These groups include African-American men (in the USA), gay men, homeless men, men in prison, men with disabilities, and

unemployed men. Social hierarchies within societies are linked with health differentials. Men of lower socioeconomic status and those who are less socially integrated exhibit poorer health outcomes when measured in terms of mortality, disability, chronic illness, or injury rates (see *Life Expectancy and Adult Mortality in Industrialized Countries*; *Mortality Differentials: Selection and Causation*).

4. Health Behavior

Many health scientists believe that health behaviors and lifestyles are among the most important factors influencing health and longevity and that both men and women can substantially decrease their health risks by adopting appropriate preventive practices. Men are more likely to adopt beliefs and behaviors that increase their risks and less likely to engage in behaviors that are linked with health and longevity. Gender differences in alcohol abuse are particularly pronounced. National survey data from the USA indicate that about 13 percent of men are classified as heavy drinkers, as compared to only 3 percent of women (Waldron 1995).

Cigarette smoking is probably the single most important behavioral risk factor for disease and premature mortality. In many developed countries, the prevalence of cigarette smoking is declining and gender differences in smoking behavior have been decreasing for decades, but even today cigarette smoking continues to be more prevalent among men.

There is evidence that regular physical activity reduces the risk of major chronic diseases. Many authors have noted that in general, men are slightly more physically active than women. However, the type and intensity of physical activity seem to differ by gender. It appears that men are more likely to engage in infrequent but strenuous activities that may increase the risk of injury (such as weight-lifting or team sports). In contrast, women seem to be more likely to engage in regular, light to moderate exercise (e.g., walking, gardening, housework) that confers optimal health benefits (Courtenay 2000a).

Regular medical examinations are critical for the early detection of many diseases and may result in a better prognosis. It is true that women visit physicians more often than men, but this difference is observed primarily for conditions that are not major causes of death. Gender differences in health care utilization generally begin to disappear when the health problem is more serious. For heart disease and most types of cancer, women delay seeking medical care as long as or longer than men, and thus do not have a better prognosis in the case of illness (Waldron 1995).

Courtenay (2000a) reviewed the evidence on gender differences in more than 30 health practices, including self-examinations, dietary practices, safety belt use, violence, and various risk-taking behaviors. He concluded that men are significantly more likely than women to engage in practices that increase the risk of disease, injury, and death. These behaviors tend to co-occur in healthy or unhealthy clusters, and the interaction of unhealthy practices (e.g., cigarette smoking combined with alcohol abuse) may compound men's health risks.

5. Masculinity and Men's Health

Some recent sociocultural studies on men's health and illness have been influenced by critical feminist theories. These studies focus on gender as a key factor for understanding the patterns of men's health risks and men's psychological adjustment to illness. The term gender encompasses expectations and behaviors that individuals learn about masculinity and femininity (see *Masculinities and Femininities*). From a social constructivist perspective, both men and women construct gender by adopting from their culture concepts of masculinity and femininity. In this view a person's gender is not something one is, but rather something one does in social interactions.

There is very high level of agreement within societies about what are considered to be typically masculine and typically feminine characteristics. For example, typically masculine gender stereotypes in contemporary Western societies include aggression, competitiveness, dominance, independence, and invulnerability (Moynihan 1998). These stereotypes provide collective and dichotomous meanings of gender, and they often become widely shared beliefs about who men and women innately are. People are encouraged to conform to these stereotypic beliefs and behaviors. In most instances men and women do conform, and they adopt these dominant norms of masculinity and femininity.

Courtenay (2000b) argued that health-related beliefs and behaviors are prominent means for demonstrating masculinity and femininity. Health behaviors can be understood as ways of constructing or demonstrating gender, and the 'doing of health' is a form of 'doing gender.' Men use health beliefs and behaviors to demonstrate masculine ideals that clearly establish them as men. Among these health-related beliefs and behaviors are the denial of weakness and vulnerability, emotional and physical control, the appearance of being strong and robust, dismissal of any need for help, a ceaseless interest in sex, and the display of aggressive behavior and physical dominance.

> In exhibiting masculine ideals with health behavior, men reinforce cultural beliefs that men are more powerful and less vulnerable than women; that men's bodies are structurally more efficient than and superior to women's bodies; that asking for help and caring for one's health are feminine; and that the most powerful men among men are those

for whom health and safety are irrelevant (Courtenay 2000b, p. 1389).

In these ways masculinity is constructed and defined at the expense of men's health. For instance, fear of being 'soft' may deter men from applying sunscreen to prevent skin cancer, and the need to display toughness to win peer approval may result in violence or risky driving. Similarly, men are demonstrating dominant norms of manhood when they refuse to take sick leave from work, when they boast that drinking does not impair their driving, or when they brag that they have not been to a doctor for a very long time.

6. Conclusion

Although health research has frequently used males as study subjects, it has typically neglected to examine men and the health risks associated with men's gender. However, men's greater susceptibility to disease and premature death is being increasingly noted, and the health of men is becoming a public health concern. Men's health advocates have pointed out that it is important to contest stereotypes of masculinity that are widely disseminated and deeply entrenched in the healthcare system and other institutions. It has also been suggested that behavioral interventions designed to improve health behaviors should prove effective for prevention, because these behaviors are important determinants of health and longevity. Considering that many health practices are differentially linked to notions of masculinity and femininity, the design of gender-specific interventions may be required to yield effective outcomes (Weidner 2000).

See also: Gender and Cardiovascular Health; Gender and Health Care; Gender and Physical Health; Gender Role Stress and Health; Health Behaviors; Physical Activity and Health; Reproductive Rights in Developing Nations; Smoking and Health; Women's Health; Y-chromosomes and Evolution

Bibliography

Christensen K, Kristiansen M, Hagen-Larsen H, Skytthe A, Bathum L, Jeune B, Andersen-Ranberg K, Vaupel J W, Ørstavik K H 2000 X-linked genetic factors regulate hematopoietic stem-cell kinetics in females. *Blood* **95**: 2449–51

Courtenay W H 2000a Behavioral factors associated with disease, injury and death among men: Evidence and implications for prevention. *Journal of Men's Studies* **9**: 81–142

Courtenay W H 2000b Constructions of masculinity and their influence on men's well-being: A theory of gender and health. *Social Science and Medicine* **50**: 1385–1401

Hazzard W R 1986 Biological basis of the sex differential in longevity. *Journal of the American Geriatrics Society* **34**: 455–71

MacIntyre S, Hunt K, Sweeting H 1996 Gender differences in health: Are things really as simple as they seem? *Social Science and Medicine* **42**: 617–24

Moynihan C 1998 Theories in health care and research: Theories of masculinity. *British Medical Journal* **317**: 1072–5

Murphy S L 2000 *Deaths: Final Data for 1998*. National vital statistics reports, Vol. 48, No. 11. National Center for Health Statistics, Hyattsville, MD

Schofield T, Connell R W, Walker L, Wood J F, Butland D L 2000 Understanding men's health and illness: A gender-relations approach to policy, research, and practice. *Journal of American College Health* **48**: 247–56

Shkolnikov V M, Cornia G A, Leon D A, Mesle F 1998 Causes of the Russian mortality crisis: Evidence and interpretations. *World Development* **26**: 1995–2011

Verbrugge L M 1989 The twain meet: Empirical explanations of sex differences in health and mortality. *Journal of Health and Social Behavior* **30**: 282–304

Waldron I 1995 Contributions of changing gender differences in behavior and social roles to changing gender differences in mortality. In: Sabo D, Gordon D F (eds.) *Men's Health and Illness: Gender, Power, and the Body*. Sage, Thousand Oaks, CA, 22–45

Waldron I 1998 Factors determining the sex ratio at birth. In United Nations (ed.) *Too Young to Die: Genes or Gender?* United Nations, New York, 53–63

Wang X T, Hertwig R 1999 How is maternal survival related to reproductive success? *Behavioral and Brain Sciences* **22**: 236–7

Weidner G 2000 Why do men get more heart disease than women? An international perspective. *Journal of American College Health* **48**: 291–4

World Health Organization 2000 *The World Health Report: Health Systems: Improving Performance*. World Health Organization, Geneva, Switzerland

E. Brähler and H. Maier

Mental and Behavioral Disorders, Diagnosis and Classification of

1. Introduction: Principles and Aims of Classification Systems for Mental Disorders

All scientific disciplines use conventions to specify how the relevant phenomena of their respective fields should be defined and labeled (the nomenclature) and according to which criteria and aspects they should be organized and classified in order to simplify complex data and phenomena on the basis of similarities and differences in order to facilitate communication. In medicine, such nomenclatures and classification systems are key prerequisites for the diagnostic identification of patients and their disorders within the clinician's diagnostic process. This is the process by which a diagnostician assesses and evaluates specific phenomena of an individual's laboratory findings, complaints, and behaviors in order to assign the

patient subsequently to one or multiple diagnostic classes in the respective system.

Key requirements of diagnostic classifications systems in medicine are:

(a) reliability in terms of the consistency with which diagnostic classificatory decisions are made;

(b) validity in terms of agreement between basic research and clinical utility for intervention, for example;

(c) comprehensiveness in terms of coverage of all relevant disease phenomena across the age span;

(d) relevance for prognosis in terms of further course, remission, and relapse;

(e) utility for research in terms of the consistent and reliable collection of data to permit the generation and testing of hypotheses.

Since Hippocrates' time, there have been numerous attempts to classify psychiatric diseases according to a few major classes. The systematic and scientifically based search for such classification systems is, however, a relative recent phenomenon. Kraepelin's influential experimental-based distinctions and various nosological modeling exercises that separated mood disorders from dementia praecox (later known as schizophrenia), along with the increasing availability of research methods and competing concepts, gave rise to the development of numerous phenomenological classification attempts in the first half of the twentieth century. Most of these systems of thought are, however, based on dubious or speculative etiological assumptions and use a wide range of variably defined cognitive, affective, and behavioral signs and symptoms along with aspects of course, outcome, and intensity as the key defining characteristics of mental disorders. Despite the fact that almost all of these systems focused on only a few forms of mental disorders, there was little agreement among these different classification systems in terms of principles and common concepts. Consequently, until 1950, there were hundreds of various psychiatric classification systems worldwide, which did not comply with the key requirements of classification systems in terms of comprehensiveness, reliability, validity, and utility for research and intervention (Blashfield 1984).

This heterogeneity reflected, on the one hand, the multifaceted presentations of mental disorders with regard to biological, social, and psychological manifestations and, on the other, the poor state of scientific research and knowledge concerning the nature and the causes of mental disorders until the mid-twentieth century. A further obstacle is that, for modern medicine, the ideal and ultimate goal of diagnostic classification systems is a 'nosological' classification (see *Nosology in Psychiatry*). The term nosology refers to an integrative and comprehensive derivation of models of diseases within the framework of a logical order, according to specified 'natural criteria' that are based not only on shared similarities and differences of all objectively given logical phenomena, but also on their respective etiological and pathogenic factors (e.g., genetic, neurobiological, psychological) proven to be of relevance in experimental and empirical research. For mental disorders, the derivation of such logical, natural, and comprehensive systems of illnesses seems premature. Therefore, even now, current classification systems for mental and behavioral disorders (previously called psychiatric classification systems) remain mainly descriptive and phenomenological. Thus, strictly speaking, they have to be considered as typologies.

Although recent research progress in neurobiology and psychology have led to increasing agreement about more adequate classification principles, it seems unlikely that there will ever be absolute and infallibly perfect criteria for mental disorders as well as a true classification of diseases: first, because of the complexity of what could be summarized under mental and behavioral disorders; second, because of the dependence of progress in scientific knowledge about the causes, nature, course, and outcome of mental disorders. Furthermore, the term mental disorders implies a distinction between 'mental' disorders and 'physical' disorders, which is, according to our current knowledge, a reductionist anachronism of mind-body dualism. There is compelling scientific evidence that there is a lot that is physical in mental disorders, as well as much that is mental in physical disorders. It seems fair to emphasize that the problem raised by the term mental disorders has been much clearer than any solution; thus the term unfortunately persists in the titles of all classificatory systems because no appropriate substitute has yet been found.

In light of this continuing dilemma, it was agreed in the 1980s and 1990s to use descriptive, phenomenological, psychopathological classification systems of mental and behavioral disorders, instead of nosological 'psychiatric' systems. Examples of up-to-date nomenclatures and respective classifications for mental and behavioral disorders that will be highlighted here are chapter F (Mental and Behavioural Disorders) of the tenth revision of the International Classification of Diseases (*ICD-10*, WHO 1991), as well as the *Diagnostic and Statistical Manual of Mental Disorders*, 3rd and 4th revisions (*DSM-III* and *DSM-IV*, American Psychiatric Association 1980, 1994). As *DSM-IV* is the major diagnostic classification system used in scientific research around the world, this overview focuses primarily on *DSM-IV* because of its greater stringency and scientific utility.

2. *Towards a Systematic and Comprehensive International Classification of Mental Disorders*

The 8th revision of the *WHO International Classification of Diseases*, injuries and causes of death (*ICD-8*, WHO 1971) in the 1960s signaled the beginning of

systematic efforts at an international level to make a serious attempt to develop a unified system of classification of mental disorders. However, despite improvements brought about by this attempt and its 9th revision (*ICD-9*, 1978) with regard to agreement about the core spectrum of classes of relevant disorders and increased comprehensiveness, the diagnostic guidelines remained limited to narrative descriptions that provided neither a well-defined nomenclature and glossary nor specific diagnostic criteria. Although *ICD-8/9* presented for the first time a more comprehensive and systematic tabulation of mental disorders, they relied almost exclusively on clinical judgment and its application by psychiatrists. Therefore, they were appropriately labeled *Psychiatric Classification Systems*. Due to *ICD-8/9*'s narrative nature, which did not include explicit criteria and rules and which was strongly bound to so-called 'clinical judgments by psychiatrists,' diagnosing of mental disorders was more an art than a science, plagued by unreliability and lack of validity and research and clinical utility.

With the increasing interest in systematic basic and applied research on mental disorders and the need for cross-national multicenter studies, it became quickly evident that even broad *ICD* diagnoses such as anxiety neurosis, depressive neurosis, and schizophrenia could not be compared between countries or even between various psychiatric schools within any given country. Another shortcoming of these systems was that they were not uniformly accepted by other core scientific disciplines such as psychology, neuropharmacology, and other allied disciplines that had gained greater influence in mental disorders research and intervention since the 1960s.

The increasing degree of dissatisfaction in psychiatry, clinical psychology, and the allied disciplines stimulated greater opposition and a search for alternative approaches: (a) antipsychiatric movements and sociological alternatives that defied the usefulness of any psychiatric diagnoses; (b) unitarian positions that classified mental disorders by degree of disturbance of key functions; and (c) a 'diagnostic nihilism' that was especially pronounced in both clinical psychology and psychoanalytical schools of thought. The latter, in particular, emphasized the need for alternative classification systems built, however, on yet partly unproven etiological and therapeutic models that seemed to promise at least greater utility for therapeutic decisions.

At the same time psychology, and clinical psychology in particular, started to provide new, promising methods and behavioral, psychophysiological, and cognitive models based on experimental research. Increasingly psychometric and psychopathometric studies attracted attention, especially in the context of new behavioral and pharmacological treatments that became available for many mental disorders. Empirical psychological methods, such as the use of psychometric instruments (i.e., self-report and clinician-rated) based on test theory and key psychometric properties of objectivity, reliability, and validity, were used in more detailed and reliable assessments and descriptions of observable manifestations of acute and chronic psychopathological phenomena. These developments offered various rational ways of improving the diagnostic classification process in terms of reliability, as a prerequisite for improved future validity (Matarazzo 1983), and were also instrumental for subsequent progress in neurobiological, genetic, and psychopharmacological research in mental disorders.

3. *Explicit Criteria and Operationalized Diagnoses within a Comprehensive Multi-axial Classification System*

The advent of *DSM-III* (American Psychiatric Association 1980) and their fourth revision *DSM-IV* (American Psychiatric Association 1994) marked the beginning of the contemporary era in the diagnosis and classification of mental disorders. More consistent with the preceding psychometric studies and linked validation work, every diagnostic category—from infancy to old age—was now given an operational definition, strictly descriptive and as 'neutral' as possible. Necessary symptoms and phenomena were specified, along with rules on how to calculate a diagnostic decision based on signs, symptoms, and their duration; syndromes and diagnostic exclusions were also included (Table 1).

This 'atheoretical' (in the sense of deleting 'unproven' nosologies) and descriptive approach substantially increased the diagnostic reliability of the assessment of psychopathology and the process of making diagnostic decisions of mental disorders (Wittchen and Lachner 1996). Furthermore, due to the empirical nature of this approach, subsequent revisions and improvements of the reliability, validity, and clinical and research utility of this classification system could begin to be systematically related to empirical evidence derived from the application of these strict criteria. Although one needs to acknowledge that not all of these criteria have yet been fully validated by data about important correlates, clinical course and outcome, family history, and treatment response, they are at least partially based on empirical studies. These *DSM* criteria could be regarded as helpful intermediate steps which we can expect will lead to further refinements following further studies.

3.1 *Operational Diagnostic Criteria*

The *DSM-III/IV* approach is superior to previous classification systems in terms of observability (emphasis on explicit behavioral symptoms), reliability (in

Table 1
DSM-IV—Multiaxial classification and groups of disorders covered

Axis I	**Clinical Disorders**
	Other Conditions That May Be a Focus of Clinical Attention
Axis II	**Personality Disorders**
	Mental Retardation
Axis III	**General Medical Conditions**
Axis IV	**Psychosocial and Environmental Problems**
Axis V	**Global Assessment of Functioning**

Axis I disorders

Disorder Usually First Diagnosed in Infancy, Childhood, or Adolescence (*excluding Mental Retardation, which is diagnosed on Axis II*)
Delirium, Dementia, and Amnesic and Other Cognitive Disorders
Substance-Related Disorders
Schizophrenia and Other Psychotic Disorders
Mood Disorders
Anxiety Disorders
Somatoform Disorders
Factitious Disorders
Dissociative Disorders
Sexual and Gender Identity Disorders
Eating Disorders
Sleep Disorders
Impulse-Control Disorders Not Elsewhere Classified
Adjustment Disorders
Other Conditions That May Be a Focus of Clinical Attention

Axis II disorders

Paranoid Personality Disorder
Dependent Personality Disorder
Schizoid Personality Disorder
Obsessive-Compulsive Personality
Schizotypal Personality Disorder
Antisocial Personality Disorder
Borderline Personality Disorder
Histrionic Personality Disorder
Narcissistic Personality Disorder
Avoidant Personality Disorder
Personality Disorder Not Otherwise Specified

Mental Retardation

Axis IV Psychosocial and Environmental Problems

Problems with primary support group
Problems related to the social environment
Educational problems
Occupational problems
Housing problems
Economic problems
Problems with access to health care services
Problems related to interaction with the legal system/crime
Other psychosocial and environmental problems

terms of agreement among clinicians), validity (in terms of agreement with other criteria), feasibility (ease of administration), coverage (covering all disorders of clinical significance in the mental health system), and age sensitivity (from infancy to old age). It also acknowledges the continuous need to increase further the observability, reliability, validity, feasibility, coverage, and age sensitivity of the system, as well as to modify the specific diagnostic categories in light of empirical advances in mental health research.

3.2 Definition of Mental Disorder

Although there is still no perfect definition or precise boundaries for what constitutes a mental disorder (similar to the concepts of somatic disorder and physical health), *DSM-III* has at least attempted to specify what constitutes a mental disorder:

> Each mental disorder if conceptualized as a clinically significant behavioral or psychological syndrome or pattern that occurs in a person and thus is associated with present distress (a painful symptom) or disability (impairment in one or more important areas of functioning) or with a significantly increased risk of suffering death, pain, disability or an important loss of freedom. In addition, this syndrome or pattern must not be merely an expectable response to a particular event, e.g., the death of a loved one. Whatever its original cause, it must currently be considered a manifestation of a behavioral, psychological or biological dysfunction in the person. Neither deviant behavior (political, religious or sexual) nor conflicts that are primarily between the individual and the society are mental disorders unless the deviance or conflict is a symptom of dysfunction in the person as described above. There is no assumption that each mental disorder is a discrete entity with sharp boundaries (discontinuity between it) and other mental disorders or between it or no mental disorder. (American Psychiatric Association 1994)

3.3 Multi-axial Evaluation

There is increasing evidence that the complexity of mental and behavioral disorders cannot be described adequately by focusing merely on psychiatric symptoms. Therefore, in accordance with previous studies, *DSM* suggested classifying patients with regard to phenomena on five axes (Table 1). In its entirety, the multiaxial system provides a biopsychosocial approach to mental disorders and their assessment.

3.4 Descriptive Approach

The etiology or pathophysiological processes are known in detail for only a few mental disorders, such as for so called 'organic mental disorders' where organic (biological) factors have been proven to be necessary for the development and maintenance of the respective disorder. For most of the other disorders, however, the etiology is unknown, beyond the finding that, for all disorders, a complex vulnerability-stress model seems to be the most adequate. Although many theories have been advanced to explain core-psychopathological processes for some disorders, the explanations at this point remain unsatisfactory and thus do not provide a firm basis for classification. Therefore, current classification systems for mental disorders are largely atheoretical with regard to the specific etiology or pathophysiological processes. The current system is descriptive, referring to the fact that the definitions of disorders are generally limited to descriptions of the clinical features of the disorders, consisting of easily identifiable behavior signs or symptoms, such as pathological anxiety reactions (i.e., panic attack), mood disturbances, and certain psychotic symptoms that require a minimal amount of inference on the part of the observer (Table 2).

The descriptive approach is also used to group mental disorders into diagnostic classes by the use of operationalized definitions which specify in a prototypical way all the mandatory symptoms and additional criteria used for the diagnosis of a disorder. All of the disorders with a known etiology of pathophysiology are grouped into classes on the basis of shared clinical features. Others are classified according to a key syndrome, such as the presence of anxiety reactions and avoidance behavior in anxiety disorders.

3.5 Systematic Description

Beyond the operationalization of diagnosis, the descriptions of disorders are more comprehensively described in terms of essential features, associated features, age-of-onset, course impairment complications, predisposing factors, prevalence, familial patterns, and differential diagnosis. This provides a better understanding of the disorder itself and has been shown to reduce the degree of misinterpretation of the explicit diagnostic criteria for symptoms, syndromes, and diagnosis by users in research and clinical settings.

3.6 Comorbidity

DSM's goal of a broader descriptive and atheoretical psychopathological classification system has necessarily led to a considerable increase in the number of diagnoses covered and in a substantially higher rate of patients with multiple diagnoses. The occurrence of more than one specific mental disorder in a person has been labeled 'comorbidity' (see *Comorbidity*). Unlike previous systems, which used questionable hierarchical rules to reduce the substantial degree of comorbidity in an attempt to arrive at a core or main diagnosis, *DSM-IV* encourages the assignment of multiple diagnoses, both cross-sectionally as well as

Table 2
Examples for explicit criteria and operationalized diagnoses in *DSM-IV*

A. Syndromal Level—Criteria for Panic Attack
A discrete period of intense fear or discomfort, in which four (or more) of the following symptoms developed abruptly and reached a peak within 10 minutes:

(1) palpitations, pounding heart, or accelerated heart rate
(2) sweating
(3) trembling or shaking
(4) sensations of shortness of breath or smothering
(5) feeling of choking
(6) chest pain or discomfort
(7) nausea or abdominal distress
(8) feeling dizzy, unsteady, lightheaded, or faint
(9) derealization (feelings of unreality) or depersonalization (being detached from oneself)
(10) fear of losing control or going crazy
(11) fear of dying
(12) paresthesia (numbness or tingling sensations)
(13) chills or hot flushes

B. Diagnostic criteria for Panic Disorder without Agoraphobia
Both (1) and (2)
recurrent unexpected panic attacks
at least one of the attacks has been followed by at least 1 month (or more) of one (or more) of the following:
—persistent concern about having additional attacks
—worry about the implications of the attack or its consequences (losing control, heart attack)
—significant change in behaviour related to the attacks

I. Absence of Agoraphobia.
I. The panic attacks are not due to the direct physiological effects of a substance (e.g., drug of abuse, a medication) or a general medical condition (e.g., hyperthyroidism).
I. The panic attacks are not better accounted for by another mental disorder, such as social phobia, specific phobia, obsessive-compulsive disorder, post-traumatic stress disorder or separation anxiety disorder.

longitudinally (over the person's life) for various reasons:

(a) to increase the reliability of diagnostic assessments,

(b) to ensure that empirical psychopathological data are collected to better understand the boundaries of mental disorders and their inter-relationships,

(c) ultimately to arrive at a crisper and scientifically sound classification of mental disorders based on core psychopathological processes, and

(d) to enhance the system's utility in terms of clinical and basic research purposes (Wittchen and Lachner 1996).

Studies such as the US National Comorbidity Survey (Kessler et al. 1994) have demonstrated that comorbidity is a basic characteristic of all mental disorders, caused by neither a methodological artifact nor a help-seeking bias, and that comorbidity might have diagnosis-specific implications for severity, impairment, course and outcome (Wittchen 1996). There is also some evidence of the existence of broad, higher-order factors of phenotypic psychopathology that, in the future, might actually allow the organization of common psychopathological variables in terms of common genetic, neurobiological, and psychological factors.

4. The Relationship of the ICD-10 and the DSM-IV Classification of Mental Disorders

Although the 1980s and 1990s have seen an increasing degree of agreement about the major classes of mental and behavioral disorders along with the common language and nomenclature, it is important to recognize that there are still two major international classification systems in use, the *ICD-10* and *DSM-IV*. The wide acceptance with increased clinical utility of the *DSM* approach prompted the World Health Organization in the early 1990s to develop, jointly

with the *DSM* task forces, a congruent and quite similar system of more explicit diagnostic classification organized in the same way as *DSM*, but for worldwide use. Nevertheless, due to the fact that the *ICD* has been adopted by all countries and health care systems, some minor differences with regard to the organization and the scope of diagnoses covered between *ICD-10* and *DSM-IV* need to be highlighted.

(a) *ICD-10* has been produced by the World Health Organization in different user versions. In addition to the clinical diagnostic guidelines for clinical use, WHO has published separate versions for research and for the purposes of administration, primary care, and disability.

(b) *DSM-IV* operationalizes mental disorders within its diagnostic manual with emphasis on associated psychosocial dysfunctions. *ICD-10*, on the other hand, tends to shift the assessment of psychosocial impairments and disabilities towards different axes for classification (*ICD-10* Classification of Impairments, Disabilities, and Handicaps) as part of the *ICD-10* multi-axial system.

(c) *DSM-IV* is characterized by clearer and explicit structuring in defining mental disorders, whereas *ICD-10* tends to be more general and unspecific, thus retaining, at least partially, the narrative guidelines format.

(d) Furthermore, there are some conceptual differences in the definition of some mental disorders such as schizophrenia and other psychotic disorders as well as anxiety disorder that might lead to some diagnostic differences, despite the same presentation of the patient when coding the more subtle distinctions of diagnostic classes (e.g., panic and agoraphobia, personality disorder).

Yet, it needs to be acknowledged that the process of coordination between the two partly competing systems has been taken seriously in that all *DSM* diagnoses covered in the 4th revision also match comparable *ICD-10* F-codes, resulting in a high degree of convergence.

5. *Conceptual Changes*

Making our current classification systems *ICD-10* and *DSM-IV* comprehensive, atheoretical, and descriptive implies a number of conceptual changes, in comparison to their predecessors. These changes do not only refer to the increased number of diagnoses along with the emphasis on diagnosing comorbid conditions whenever appropriate, they also influence our understanding of specific types of disorders.

5.1 *Schizophrenia and Psychotic Disorders*

To match new scientific evidence, demonstrating significantly different patterns of course and outcome and risk factors, as well as to increase reliability, both systems, *ICD-10* and *DSM-IV*, have differentiated considerably the classification of schizophrenia and other psychotic disorders. By using stricter time and symptom criteria, the concept of schizophrenia is narrowed in favor of a better delineation from other psychotic disorders, as well as from psychotic symptoms occurring in the course of affective disorders. These changes built on various studies that have highlighted differences in prognosis and treatment response of psychotic disorders and also aim at the reduction of a possible stigmatizing and negative effect of false positive diagnosis of schizophrenia.

5.2 *The Departure from the Concept of Neurosis*

In previous systems, the so-called 'neurotic disorders' comprised a large heterogeneous group of manifestations of anxiety (i.e., anxiety neurosis, phobic neuroses), depressive (i.e., depressive neurosis), somatoform conditions (hysterical neurosis), and personality disorders (characterological neurosis). However, the common etiological denominator 'neurosis' has never been proven, nor could the disorders be reliably diagnosed. Therefore, the broad concept of neurosis was discarded in favor of a disorder-specific classification of numerous types of specific conditions, for each of which firmer scientific evidence was available. This change resulted in a considerable increase of diagnoses; for example, panic disorder and generalized anxiety disorder largely replaced former anxiety neurosis; depressive neurosis was replaced by the descriptive term major depression or dysthymia; phobic neurosis was broken down into agoraphobia, social, and various forms of specific phobias. Furthermore, this departure from neurosis allowed for a more clear-cut differentiation of affective (mood) disorders from anxiety disorders and a departure from the controversial and unproven differentiation of so-called endogenous (caused by biological factors) and neurotic (reactive and neurotic-developmental etiology) depressions.

5.3 *Addictive Disorder*

Whereas in the past *ICD-8* and *-9* used a fairly generalized and poorly defined concept of addiction, *ICD-10* and *DSM-IV* built on a substance-specific syndromatic classification of 'abuse and dependence,' supplemented by numerous substance-related specific problems (i.e., intoxication syndrome), frequently seen and underdiagnosed in health care settings. The differentiation of substance-specific criteria accounts for research findings that have highlighted the

substance-specific risk factors and correlates and treatment implications on the one hand, and the need for a reliable assessment of dependence syndromes on the other.

These examples highlight that the modern classification systems for mental and behavioral disorders have brought the field much closer to the paradigms and procedures of basic and clinical research—for example, neurobiology, neuropharmacology, and clinical psychology. Each of these core disciplines can now build in research on the same uniform and reliable criteria as they are used in clinical research and practice. This achievement has, for the first time, opened the door to subsequent empirically based revisions and changes in future systems.

6. *Improving the Diagnostic Process and Diagnostic Instruments*

Assigning diagnoses of mental disorders according to the current classification systems can be described as a complex process in which a clinician makes inferences about the patient's 'true' health state, using various measures ranging from structured assessments about an individual's past and current physical, mental, and social history to laboratory tests. Within this framework, diagnostic decision theories (Fig. 1) continue to assist in identifying the key aspects involved in making inferences about the 'objective' state of the patient, by using various sources of information of varying degrees of reliability and validity, and taking into account the clinical evaluation process needed to interpret these findings. Such investigations are essential in improving classification systems of mental disorders, because they largely depend on the verbal interaction between a diagnostician and the patient. Building on the current state of knowledge about the nature and course of mental disorders, the key problematic issues of diagnostic classification in mental disorders have become apparent:

(a) For mental disorders, reliable and objective laboratory or neuropsychological markers are not currently available; therefore, improved classification systems should focus on clearer guidelines of assessment for key psychopathological symptoms in terms of their subjective (patient self-report) and clinician -observed (clinical rating) aspects.

(b) Mental disorders vary in terms of onset, risk factors, course, duration, severity, and associated disability. Therefore, it is necessary to establish clear criteria regarding symptoms and phenomena for the diagnosis of specific disorders, along with clear criteria regarding duration, onset, severity, and other associated behavioral aspects.

(c) Many psychopathological terms used for classification purposes in the past stem from various theoretical, yet unproven and untested, hypotheses referring to various theoretical schools of thought for mental disorders. Examples of such ambiguous terms that are neither shared by all users nor supported by experimental studies are diagnostic terms, such as 'endogenous depression' and 'neurosis,' and psychopathological symptoms, such as 'frustration tolerance,' all of which are only vaguely described and not consistently assessable. Therefore, there is the need for well-defined psychopathological terms that can be evaluated with high reliability by all mental health professionals, irrespective of their theoretical orientation.

Referring to the model of the diagnostic process that structured the diagnostic process for mental disorders in a systematic way (Fig. 2), numerous empirical studies have resulted in specific suggestions regarding how to improve its consistency and reliability (Wittchen and Lachner 1986). Ideally, this diagnostic process starts with the reliable assessment of all relevant cognitive, affective, behavioral as well as biological, psychosocial, and social-interactive phenomena by assessing systematically all complaints, signs and symptoms of a patient. In agreement with the medical decision process, the core features of these manifestations can be labeled symptoms. According to certain rules, symptoms can be further translated into syndromes, characterized by a systematic association of frequent symptom configurations. Then, according to further clinical-nosological rules and hierarchical decisions, symptoms can be translated to diagnoses of mental disorders.

The extensive psychometric and methodological work in this domain resulted in subsequent and still ongoing attempts to improve not only the reliability of the diagnostic process (by specifying core symptoms and syndromes in *DSM-III*, *IV* and the *ICD-10*) and their systematic validation, in terms of agreement with other basic and clinical markers (Robins and Barrett 1989), but also their translation into diagnostic assessment instruments for research and clinical use. Research in diagnostic assessment has high priority for the developers of *ICD-10* and *DSM-IV*. In fact, the development of structured diagnostic interviews and their more radical next step, namely, the development of fully diagnostic standardized diagnostic interviews (see American Psychiatric Association 2000), has been made possible only by the existence of explicit descriptive diagnostic criteria and the specification of diagnostic algorithms for hundreds of specified diagnoses of mental disorders, and the intimate interaction between both the developers and users of the instruments, as well as developers of the classification system. It is now routine in basic and clinical research to require the use of comprehensive and systematic interviews that call for a standardized review of all explicit criteria, explicit wording to initiate more or less strictly defined probing of specific symptoms, and consensus standards to define thresholds for dichot-

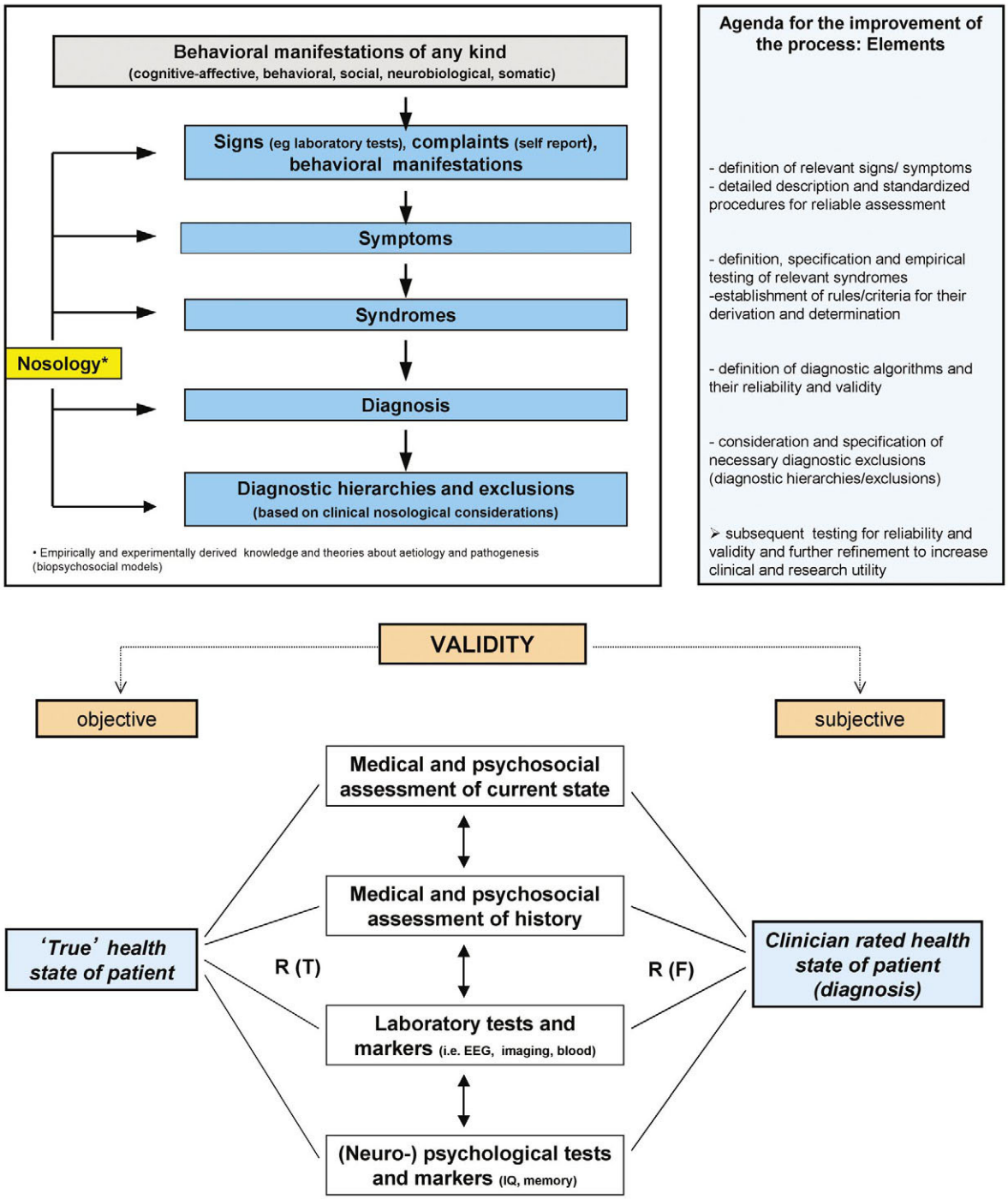

Figure 1
Diagnostic decision process (Wittchen and Lachner 1996)

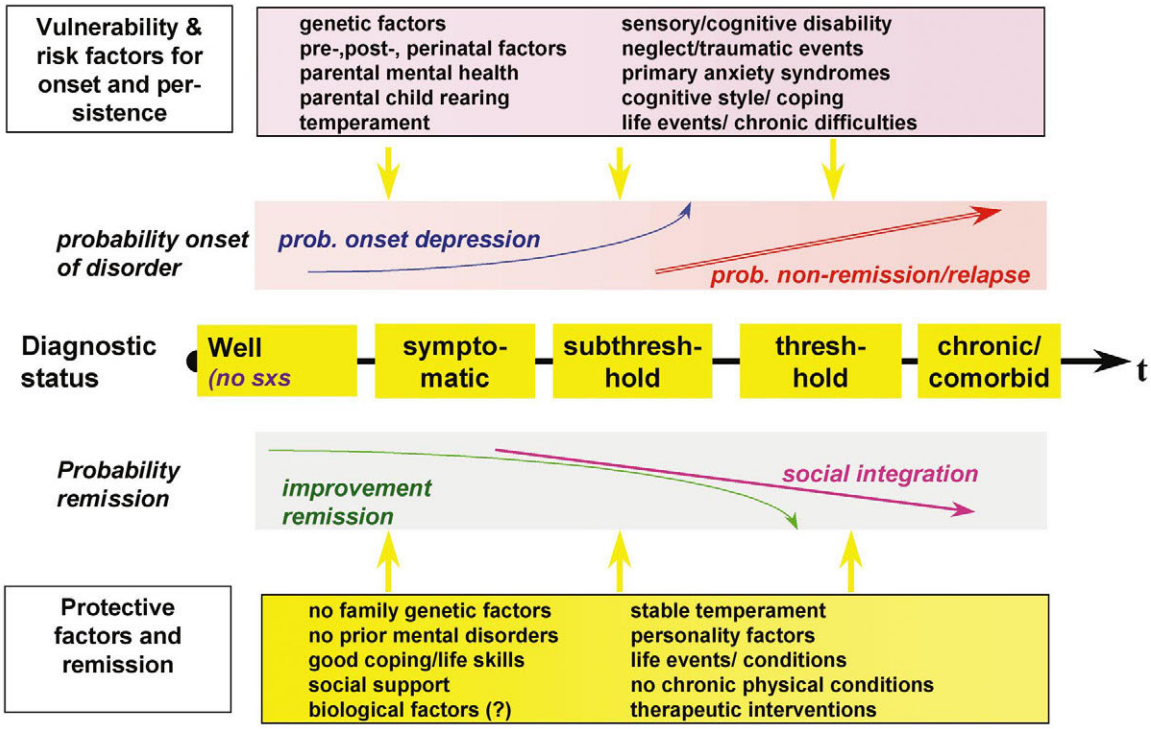

Figure 2
Diagnostic process

omous symptom and disorders classifications of open-ended responses (Wittchen 1994). With careful training and close monitoring of diagnostic ratings, it is now possible to obtain good inter-rater agreement when using diagnostic interviews.

There are several reasons for standardizing the diagnostic process and the use of instruments: (a) Learning the new classifications systems: Unlike the past when clinicians and researchers had to deal only with a handful of diagnoses, they now have to diagnose hundreds of diagnoses for mental disorders, each with its unique set of criteria. It is hard to see how they are able to use such complex classification systems appropriately and reliably without the use of diagnostic tools that guide them through the system. (b) Quality assurance: diagnosing mental disorders requires a comprehensive cross-sectional and longitudinal assessment of the patient's symptomatology along with the need to make subtle distinctions with regard to key phenomena, time, and severity criteria. Because for the vast majority of mental disorders, this process relies almost entirely on subjective-verbal interactions between the patient and the diagnostican, it is unlikely that 'clinical judgment' within an unstructured clinical assessment alone will result in adequate decision-making. As the 1980s and 1990s have witnessed a tremendous increase in effective behavioral and medication treatments for specific disorders assessed with such instruments, reliable and valid diagnostic interviews should be made a standard requirement for routine care, as well as to ensure proper diagnostic standards and appropriate treatment allocation.

7. *Future Perspectives*

7.1 *Diagnostic Assessment and the 'Science of Self-report'*

Beyond the continued need for refinement of the diagnostic process and the improvement of diagnostic instruments, classification systems for mental disorders have also started to profit from an increasingly closer collaboration of developers of diagnostic interviews with cognitive psychologists and survey methodologists (Stone et al. 2000). Diagnostic interviewing for classificatory purposes is a highly complex process between a patient and a diagnostician. Asking the right question and making sure that the patient understands the question is a challenge for the diagnostician. Challenges for the patient include: understanding the task (thoughtful and honest responding), being willing to carry out the task (motivation), and being able to carry out the task (dealing with questions

that might be ambiguous or exceed the limits of memory or cognition). Cognitive psychologists have developed a number of strategies to deal with such problems, few of which are currently systematically applied in clinical interviews. 'Think aloud' experiments and cognitive debriefing interviews are two examples of effective methods of detecting and correcting the problem of misunderstanding and other limitations of diagnostic interviews. It can be expected that a more systematic application of such clinical validation strategies will considerably lessen diagnostic assessment difficulties and ultimately result in improved classification rules and definitions.

7.2 Biological Tests and Correlates

Currently no biological findings are available that can be used as reliable diagnostic markers for mental disorders, nor are discoveries in sight that might provide an immediately available, firmer, and more objective basis for classifying mental disorders. However, referring to the tremendous progress in innovative neurobiological methods, there are some potential areas from which such markers, at least for some diagnoses, might be derived.

Among neuroendocrine tests the hypothalamic-pituitary-adrenocortical axis has received the most attention. The Dexamethasone Suppression test (DST) was the first and most studied marker in research of depressive disorders (Carroll et al. 1986). Based on the observation that depressed patients fail to suppress plasma cortisol, it was believed that a positive DST might be a specific laboratory test for severe forms of depression. Although subsequent studies confirmed the high specificity versus normal and non-psychiatric controls, there was no specificity in comparison to other mental disorders or healthy controls who had experienced a recent stressful life event. Similar specificity problems were also found for other neuroendocrine tests, such as the adrenocorticotrophic hormone (ACTH) response to corticotropin releasing factor (CRF) and growth hormone releasing hormone (GHRH), as well as sleep and neuroimaging studies (Nemeroff 1996, Holsboer 1993). It is not entirely clear at this point to what degree the partly disappointing findings of such methods for diagnostic purposes are due to the fact that our currently phenomenological classifications may be inadequate targets for such experimental laboratory marker technologies. Yet, such findings as those emerging from molecular biology, functional brain imaging and genetic studies have and will considerably advance our basic knowledge about the complex functional pathways of mental functioning and mental disorders and remain key targets for improved future classification systems.

8. Conclusion

The past two decades have witnessed tremendous advances in the classification of mental and behavioral disorders with regard to increased reliability, validity, clinical and research utility, comprehensiveness, and improved communication. The current available systems, however, are far from being satisfactory. More research is clearly needed not only to improve further the reliability of explicit criteria and to clarify the boundaries of disorders, but, in particular, to explain more effectively the complex etiology and pathogenesis of mental disorders on neurobiological, psychological, and social-behavioral levels. Such research will result, hopefully, in the identification of common and disorder-specific core psychopathological processes that might provide more valid and comprehensive criteria for a sharper and more satisfactory classification system.

See also: Classification: Conceptions in the Social Sciences; Clinical Psychology: Validity of Judgment; Clinical versus Actuarial Prediction; Decision Support Systems; Differential Diagnosis in Psychiatry; Medical Expertise, Cognitive Psychology of; Mental Health and Normality; Risk Screening, Testing, and Diagnosis: Ethical Aspects; Syndromal Diagnosis versus Dimensional Assessment, Clinical Psychology of

Bibliography

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental DSM-III*, 3rd edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington, DC

American Psychiatric Association 2000 *Handbook of Psychiatric Measures*. American Psychiatric Association, Washington, DC

Blashfield R K 1984 *The Classification of Psychopathology—Neo-Kraepelinian and Quantitative Approaches*. Plenum Press, New York

Carroll B J, Martin F, Davis B 1968 Pituitary-adrenal function in depression. *Lancet* **556**: 1373–74

Holsboer F 1993 Hormones and brain functions. In: Mendlewitz J, Brunello N, Langer S Z, Racagny G (eds.) *New Pharmacological Approaches to the Therapy of Depressive Disorders*. Kargar, Basel, Switzerland

Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshleman S, Wittchen H-U, Kendler K S 1994 Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**: 8–19

Matarazzo J D 1983 The reliability of psychiatric and psychological diagnosis. *Clinical Psychology Review* **3**: 3–145

Nemeroff C B 1996 The corticotropin-releasing factor (CRF) hypothesis of depression: New findings and new directions. *Molecular Psychiatry* **1**: 336–42

Robins E, Guze S B 1970 Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry* **126**: 983–7
Robins L N, Barrett J E (eds.) 1989 *The Validity of Psychiatric Diagnoses*. Raven Press, New York
Stone A A, Turrkan C A, Bachran C A, Jobe H S, Kurtzman H S, Cain V S (eds.) 2000 *The Science of Self-report: Implications for Research and Practice*. Erlbaum, Mahwah, NJ
Sudman S, Bradburn N M, Schwarz N 1996 *Thinking about Answers: The Application of Cognitive Processes in Survey Methodology*. Jossey-Bass, San Francisco
Wittchen H-U 1994 Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): A critical review. *Journal of Psychiatric Research* **28**(1): 57–84
Wittchen H-U 1996 Critical issues in the evaluation of comorbidity. *British Journal of Psychiatry* **168**(Suppl. 30): 9–16
Wittchen H-U, Höfler M, Merikangas K R 1999 Towards the identification of core psychopathological processes? Commentary. *Archives of General Psychiatry* **56**(10): 929–31
Wittchen, H-U, Lachner G 1996 Klassifikation. In: Ehlers A, Hahlweg K (eds.) *Enzyklopädie der Psychologie. Themenbereich D Praxisgebiete, Serie 2 Klinische Psychologie, Band 1, Grundlagen der Psychologie*. Hogrefe, Göttingen, Germany
World Health Organization 1971 *International Classification of Diseases* (8th revision). World Health Organization, Geneva, Switzerland
World Health Organization 1991 Mental and Behavioral Disorders (including disorders of psychological development). Clinical descriptions and diagnostic guidelines. In: *Tenth Revision of the International Classification of Diseases*. World Health Organisation Division of Mental Health, Geneva, Switzerland, Chap. 5(f)

H.-U. Wittchen

Mental Health and Normality

What constitutes mental health? This has been referred to as '[o]ne of the most perplexing but seldom asked questions in psychiatry' (Stoudemire 1998, p. 2). Indeed, while recent years have seen substantial steps forward with regard to defining mental illness—the fourth edition of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* (1994) is an example of such progress—psychiatrists remain at odds over the definition of mental health.

1. Definitions of Normality

Within the field of psychiatry, there are multiple definitions of what is normal. Offer and Sabshin (1991) have created a framework for these definitions, which consists of four different perspectives on normality: normality as health; normality as utopia; normality as average; and normality as transactional systems.

1.1 Normality as Health

This model is still predominant throughout medicine and states that normality is equivalent to the absence of disease. Illnesses (or diseases) are categorized and defined; normality (or health) is not defined *per se* other than stating that all behavior not meeting the criteria for illness can be considered to be normal. When a patient is treated for his or her illness, the goal of the treatment is to remove the illness. When that is accomplished, for all practical purposes the person is considered to be normal (or healthy). In the mental health field, normal behavior is defined as the absence of mental illness. The emphasis is upon the absence of disease rather than upon optimal functioning.

1.2 Normality as Utopia

In this model the criteria for normality (or health) are much higher than 'the absence of pathology.' The emphasis is upon optimal functioning, and this type of normality is like a Platonic ideal, to be strived for, but rarely achieved. Most psychoanalysts define the goal of analytic treatment as helping their patients to achieve their full potential of psychological functioning, whatever that might be. Sabshin (1988) has pointed out that psychoanalysis lacks a literal or detailed language and methodology to describe healthy functioning. According to this model, however, and in contrast to the normality as health model, relatively few people are healthy.

1.3 Normality as Average

This definition utilizes statistical data and essentially states that normal behavior falls within two standard deviations of the mean of whatever behavioral variable is being studied. In this context the boundary between health and illness is arbitrary. One special problem for this definition involves those occasions when the population being studied is dysfunctional as a whole.

1.4 Normality as Transactional Systems

In this approach variables are studied over time so that normality encompasses interacting variables measured more than once. This model differs from the three previous approaches, which utilize cross-sectional measurements. In addition, many systems and their interactions are studied in a longitudinal fashion.

While examples of every perspective can be found in current psychiatric theory and practice, the first model—normality as health—has consistently dominated both research and clinical work. The concept of a 'normal' control group, for instance, is based on this view of mental health. Epidemiological studies, by

definition, assume that those who are not identified as ill are in fact healthy or normal. Finally, the *DSM-IV*—the gold standard of psychodiagnostics—is based on the supposition that mental health equals the absence of pathology.

1.5 Normatology

The scientific study of normality over the lifecycle is referred to as normatology. The roots of normatology can be found in Offer and Sabshin's 1963 declaration of a 'theoretical interest in an operational definition of normality and health' (1963, p. 428). Since Offer and Sabshin began their work on normatology in the 1960s, they have repeatedly called upon mental health professionals to seriously consider the concept of normality. As Brodie and Banner (1997, p. 13) write in their review of normatology, [Offer and Sabshin's] genius lies in their insistence that rather than responding casually or thoughtlessly with received wisdom, we seek instead to answer seriously, methodically, empirically, and longitudinally the vast, seemingly unanswerable question that our patients, our colleagues, and our society persists in asking: 'Doc, is this normal?'

The way in which psychiatrists answer the question 'Doc, is this normal?' has implications that extend beyond mere theoretical interest. Indeed, definitions of normality are central to numerous policy debates. One arena in which the question of what constitutes a psychiatric disorder has received much attention is the health insurance debate. Broad definitions of psychiatric illness are of concern to insurance companies, who traditionally view mental health expenditures as a bottomless pit. Fearing endless expenditures for the treatment of psychopathology, insurance carriers have fought particularly hard against it. Claiming that much of what passes as psychiatric treatment is in fact unnecessary, insurers have argued that too many psychiatrists spend their time treating the 'worried well.' Recent trends in managed care illustrate the extent to which insurance companies have succeeded in narrowly defining what constitutes a disorder and which disorders warrant treatment. Managed care companies continue to lobby for an ever-expanding universe of normality, in which psychopathology is minimized to reduce costs associated with its treatment.

Definitions of normality are also critical to social policy debates over such issues as homosexuality. Even as homosexuality gains increased acceptance within American society (Vermont's ruling on domestic partnerships is an example of this trend), there are those who still argue that homosexuality should be considered abnormal, and as such should not be officially sanctioned. Deciding which mental disorders should qualify for Social Security Disability benefits is another example of the ways in which definitions of normality carry serious consequences. When the decision was made to stop providing disability benefits to persons suffering primarily from substance abuse disorders, many people lost their benefits.

In light of the considerable debate over what constitutes normality and what should be considered a psychiatric disorder, Offer and Sabshin have urged researchers to direct their efforts toward answering this question empirically. In *The Diversity of Normal Behavior* (1991), their most recent book on the subject, the authors note that interest in the field has developed significantly. They refer to an emerging empirical literature, and point out the impact of normatology on the *DSM-IV*. The authors attribute this change to a growing movement in psychiatry toward objectifying its nosological base, and to a desire among mental health policymakers to better define mental disorders. In spite of these developments, however, Offer and Sabshin (1991) state that psychiatrists have yet to focus adequately on the normal aspects of their patients' behavior.

In this article we will review current trends in normatology, and will assess the extent to which the field of psychiatry has made efforts to better define what is normal. We will begin by looking at the ways in which researchers have attempted to develop and refine empirically based measurements of psychological normality. We will discuss normatological developments in such disciplines as child therapy, family therapy, and sex therapy. We will also discuss current challenges to the ways in which presumptions about normality influence psychiatric diagnoses and treatment. In order to assess whether normatological concepts are being integrated into the training of new mental health professionals, we will present the results of a review of current psychiatry and psychology textbooks, as well as of a survey of clinical psychology doctoral students. Finally, we will discuss where the field of normatology appears to be headed as we move into the twenty-first century.

2. Current Trends in Research on Normality

2.1 Empirically Based Approaches to Normality

One area that has seen some progress since the 1990s is the development of empirically based approaches to distinguish between the normal and abnormal. While such approaches are not new (the original MMPI is an example of an early attempt), researchers have continued to refine existing instruments and create new ones. Achenbach (1997, p. 94) has been a particularly strong proponent of such an approach. He suggests that in order to distinguish normality from abnormality, both should be seen as existing on a continuum. Placing an individual at a specific point on the continuum, according to Achenbach, requires an 'empirically based bootstrapping strategy.' Such a strategy involves 'lift[ing] ourselves by our own boot-

straps' through an iterative process of progressively testing and refining various fallible criteria for judging the normal versus abnormal.'

Achenbach's approach places a strong emphasis on psychometric testing. As he points out, however, such tests must be based on standardized data from subjects who are independently judged to be normal as opposed to demographically similar subjects who are judged to be abnormal. Failure to base normal/abnormal distinctions on empirically validated measures can lead to misdiagnosis and confusion. As an example, Achenbach points out that when the 'always on the go' criterion for Attention Deficit Disorder with Hyperactivity from the *DSM-III* (1980) was tested empirically, healthy children scored higher than those diagnosed with the disorder.

The most thoroughly researched and widely applied psychometric instrument for differentiating between normal and abnormal personalities is the Minnesota Multiphasic Personality Instrument (MMPI) (Hathaway and McKinley 1943). In 1989, the MMPI was restandardized and updated to the MMPI-2 (Butcher et al. 1989). As Ben-Porath writes in his (1994) review, the MMPI has evolved from a strict typological instrument to a test that places individuals along continua of various personality trait dimensions. In line with the methodology proposed by Achenbach (1997) outlined above, the MMPI-2 is used to quantitatively (rather than qualitatively) classify individuals into distinct groups.

Millon's model of personality (1981) as well as the psychometric instruments that have grown out of it (e.g., Strack 1991, Millon 1994a, 1994b) also incorporate a quantitative approach to the distinction between normal and abnormal. The underlying assumption of Millon's work is that abnormal personality traits are distortions or exaggerations of normal ones. Pathology, therefore, is seen as lying on one end of the normal–abnormal continuum. The Millon Clinical Multiaxial Inventory (MCMI) (Millon 1977)—a well-researched and widely used personality instrument—recently underwent its third revision (Millon 1994a). The MCMI is used primarily with clinical populations to assess individual personality characteristics along various continuum points.

The Personality Adjective CheckList (PACL) (Strack 1991), a recent outgrowth of Millon's work, was designed for use with nonpsychiatric patients. Based on Millon's principles of personality assessment, the PACL may be used to assess normal versions of the character types frequently encountered in clinical settings. While pathological versions of Millon's personality types have been well studied, relatively little work has gone into understanding their normal correlates. The PACL was developed in large part to provide researchers with a means of elucidating the normal end of the normal–abnormal continuum. While the development of the instrument is a step toward a better definition of psychological normality, investigators have yet to make use of the PACL in applied research settings.

One relatively new psychometric instrument used to differentiate between normal and abnormal personalities is the Personality Assessment Inventory (PAI) (Morey 1991). Like the PACL and MMPI-2, the PAI assumes that clinical constructs occur in both mild and more severe forms. The PAI was developed in order to provide researchers and clinicians with another tool for distinguishing between individuals at various points of the normal–abnormal continuum. In summarizing the current state of psychodiagnostics with regard to the normal/abnormal distinction, Morey and Glutting (1994, p. 411) write that 'the *American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders*, despite attempts to objectify many critical distinctions, is still unclear on distinctions between normal personality, abnormal personality, and clinical syndromes.' The authors point out, however, that while conceptual difficulties remain in differentiating the normal from abnormal, clear differences exist between psychometric instruments that were designed to measure normal or abnormal constructs. The PAI was designed to primarily measure abnormal constructs in such a way as to capture meaningful variation even within the normal range.

As previously mentioned, a common assumption underlying such psychometric instruments as the MMPI-2, PACL, and PAI is that normal and abnormal characteristics can be seen as lying on opposite ends of the same continua. Eley (1997) calls this assumption into question with her study of childhood depressive symptoms. Looking at 589 same-sex twins eight to 16 years old, Eley found that genetic factors significantly contribute to the etiology of both individual differences and extreme group membership for self-reported depressive symptoms. In other words, Eley's results suggest that depressed children differ qualitatively from children who are not depressed. These results support a categorical, rather than continual approach to the normal/abnormal distinction.

2.1.1 Child and family therapy. Child and family therapy is an area in which clinicians are frequently asked to make distinctions between normal and abnormal behavior. As a recent study by Cederborg (1995) demonstrates, these distinctions are negotiated through a complicated process involving a number of participants, namely therapist, parent(s), and child. In order to understand better how these participants define and redefine a child's behavior, Cederborg analyzed the content of 28 family therapy sessions with seven different families. She concludes that rather than classifying children according to psychiatric diagnoses, family therapists (at least those in Sweden) tend to define abnormality in terms of 'interactional problems.' What might initially be defined as psychi-

atric problems, writes Cederborg, are transformed through therapy into social phenomena in the child's environment.

Nevertheless, as Gardner (1988) points out, for all children referred to psychiatric treatment, a level of classification has already taken place. This classification occurs at the level of *difference*. For instance, a parent might contrast the behavior of their child with that of the child's peers. One criterion that is often used to classify childhood behavior, writes Gardner, is whether a child's behavior is gender-appropriate. In order to determine how gender norms affect distinctions between normal and abnormal children, Gardner studied 15 boys and 15 girls, six to nine years old who were referred to a child guidance clinic in the UK (southwest England). She found that for over half of the sample, the problem behavior that resulted in the original referral was considered by the children's mothers to be gender-inappropriate. Gardner concludes that mental health professionals working with children should 'consciously increase [their] gender awareness and knowledge about the individual variability of behaviors and so encourage others to be more flexible and less judgmental about deviations from the norm' (Gardner 1988, p. 80).

Walsh (1993, pp. 3–4), in her review of normal family processes, comes to a conclusion in keeping with the findings of Cederborg (1995) and Gardner (1988): namely, that 'all views of normality are socially constructed.' As a result, Walsh writes that it is 'imperative to examine the explicit and implicit assumptions about family normality that are embedded in our cultural and clinical belief systems.' A study conducted by Kazak et al. (1989, pp. 289–90) supports Walsh's assertion. In Kazak et al.'s study, four samples (20 families with young children, 172 college undergraduates, 24 grandmothers, and 21 therapists) completed a battery of standardized family assessment instruments. All participants were asked to indicate on the instruments how a 'normal' family would respond. The investigators found significant and substantial differences in perceptions of normality according to developmental variables, ethnic background, and gender. Of particular interest in this study is the gap between the therapists' perceptions of normality and those of the other subjects. The authors suggest that these differences point to 'inherent tensions between the values and expectations that therapists have, and those of families.'

As McGoldrick et al. (1993, p. 405) point out, one result of the cultural basis for judgements of normality is that as times change, so do perceptions of what is normal. '… our dramatically changing family patterns' write the authors, 'which in this day and age can assume many varied configurations over the life span, are forcing us to take a broader view of both development and normalcy. It is becoming increasingly difficult to determine what family life cycle patterns are "normal" …'

2.1.2 Sex and marital therapy. Another area in which culturally driven judgments of normality affect psychiatric treatment is sex and marital therapy. Kernberg (1997) writes that when normality is equated with predominant or average behavior (re: normality as average), treatment may become a matter of 'adjustment,' and normality may lose its usefulness as a standard of health. However, if normality represents an ideal pattern of behavior (re: normality as an ideal fiction), then treatment may become ideologically motivated. Warning of the risks of ideologically and culturally biased treatment, Kernberg reminds us that 'only a hundred years ago psychoanalysis was at one with a scientific community that regarded masturbation as a dangerous form of pathology …' and 'that our literature lumped homosexuality and sexual perversions together for many years without a focus on their significantly differentiating features' (p. 20).

2.2 Review of Current Psychiatry and Psychology Textbooks

Graduate education provides future mental health professionals with an opportunity to become acquainted with normatological concepts. This process may occur formally through coursework and lectures, or informally through clinical experience and observations. Graduate training is a time when professional conceptualizations of normality are first formed; how this happens is of critical importance to an understanding of current and future trends in normatology. Indeed, what graduate students learn about normality today is perhaps the best barometer of how the field will view normality tomorrow.

In order to assess how normatological concepts are being formally introduced to current graduate students, we reviewed the most up-to-date psychiatry and psychology textbooks available at a private medical school in a major mid-Western city in the USA. In all, we examined 11 textbooks (many of which are used at medical schools throughout the USA): six general psychiatry textbooks (Sadock and Sadock 2000, Stoudemire 1998, Goldman 1995, Waldinger 1997, Nicholi 1999, Hales et al. 1999); two general clinical psychology textbooks (Heiden and Hersen 1995, Hersen et al. 1991); two abnormal psychology textbooks (Turner and Hersen 1997, Peterson 1996); and one developmental psychopathology textbook (Luthar et al. 1997).

We searched the indexes of these textbooks for the following entries: 'normality'; 'mental health'; 'mental illness'; and 'mental disorder.' Of the 11 textbooks, only three included entries for 'normality' (27 percent); one included 'mental health' (9 percent); two included 'mental illness' (18 percent); and four included 'mental disorder' (36 percent). In all, it appears that a student searching the indexes of these textbooks for terms relating to normality would find disappointingly little.

We then examined the textbooks in order to determine whether they explicitly addressed the concept of psychological normality. Of the 11 textbooks, only four (36 percent) included sections that dealt explicitly with normatological issues.

In order to determine what messages these textbooks are conveying—explicitly or implicitly—about normality, we divided the textbooks into groups based on the models of normality outlined in Sect. 2 (normality as health; normality as utopia; normality as average; and normality as transactional systems). Textbooks that stated either directly or indirectly that normality equals the absence of pathology were placed in the first group, and so on. The results of our analysis are as follows: five textbooks (45 percent) supported the 'normality as health' model; none (0 percent) supported 'normality as utopia'; four (36 percent) supported 'normality as average'; and one (9 percent) supported 'normality as transactional systems.' Additionally, one text (9 percent) listed Offer and Sabshin's five models of normality without favoring any particular viewpoint.

It appears, therefore, that the majority of the introductory psychiatry and psychology textbooks currently used by students of at least one major medical school teaches students that normality equals either the absence of pathology, or that which lies in the middle of the bell-shaped curve. It also appears that explicit information relating to normatological concepts is in short supply, at least as far as graduate textbooks are concerned. Needless to say, the implications of our review are limited by the small number of textbooks examined, as well as by the fact that we focused on only one school. Nevertheless, our results raise serious concerns about the extent to which normality is overlooked by the authors of introductory psychiatry and clinical psychology textbooks.

2.3 *Survey of Clinical Psychology Doctoral Students*

In order to learn more about how current mental health professionals-in-training view normality, we conducted a survey of clinical psychology doctoral students at two mid-Western universities. The brief nine-question survey was sent out by e-mail to all of those students who had at least six months of clinical training. A total of 53 students received the survey. The students' response rate was 64 percent.

Respondents had an average of 30 months of clinical experience. Seventy-nine percent of the respondents were involved in clinical work when they answered the survey. During the course of their clinical work, 91 percent of the students said that a patient had asked them 'Am I normal?' Only 52 percent of those asked directly answered their patient's question.

Ninety-four percent of the respondents said that they believed that an understanding of normality was necessary for their future clinical work. Only 62 percent of the respondents stated that they had taken a course in graduate school that dealt with issues of normality. Of those students who had taken such a class, 91 percent reported that they found the course helpful.

With regard to the models of normality outlined in Sect. 2, the majority of the respondents (56 percent) aligned themselves with the 'normality as transactional systems' approach. Twenty-seven percent of the students agreed with the 'normality as average' perspective. Twelve percent viewed normality as a utopian ideal, and the rest (6 percent) agreed with the 'normality as health' model.

Seventy-seven percent of the students believed that the 'normality as health' perspective was the dominant view within the field of mental health. The rest of the students were nearly evenly divided between the 'normality as utopia,' 'normality as average,' and 'normality as transactional systems' models with regard to what view they felt dominated (6, 9, and 9 percent, respectively).

These findings suggest that the vast majority of current clinical psychologists-in-training who responded to our survey are directly confronted with issues of normality in their clinical work. Just over half of those confronted, however, deal with the issue explicitly. Moreover, while nearly all of those students asked believe that an understanding of normality is necessary to their clinical work, less than two-thirds of them had actually taken a class that dealt with the subject. In keeping with the results of the previous section, it would seem as though graduate training in psychology is not meeting the needs of students where normatological issues are concerned.

While over half of those students who responded to our survey believed that the 'normality as transactional systems' approach is the correct one, over three-quarters of them saw the 'normality as health' model as dominating the field. One wonders if this discrepancy is indicative of a forthcoming shift in how mental health professionals approach issues of normality.

As with our review of textbooks in the previous section, a cautionary note is in order with regard to generalizing beyond those students who responded to our survey. At the very least, however, the results of this survey raise important questions about the way in which clinical psychology graduate training addresses questions of normality.

3. *Summary*

Defining mental health has always been a more difficult task for mental health professionals than defining mental illness. Within the field of psychiatry, there are multiple, competing perspectives on nor-

mality. As outlined by Offer and Sabshin (1991) these perspectives are: normality as health; normality as utopia; normality as average; and normality as transactional systems. While examples of every perspective may be found in current psychiatric theory and practice, 'normality as health' continues to dominate the field.

Normatology is the scientific study of normality over the lifecycle. Recent developments in normatology have occurred in the areas of psychometrics, child and family therapy, and sex and marital therapy, among others. A number of writers have been increasingly critical of current nosological assumptions about normality, especially with regard to the *DSM-IV*.

A review of introductory psychiatry and psychology textbooks found few references to normality. Most of the textbooks reviewed explicitly or implicitly supported the 'normality as health' and 'normality as average' models.

A survey of clinical psychology doctoral students found a general interest in normatological issues, despite a lack of formal training on the subject. The majority of those students who responded to the survey aligned themselves with the 'normality as transactional systems' model, perhaps signaling a future shift away from the 'normality as health' perspective.

See also: Culture as a Determinant of Mental Health; Health and Illness: Mental Representations in Different Cultures; Health: Anthropological Aspects; Health Psychology; Mental and Behavioral Disorders, Diagnosis and Classification of; Personality Theory and Psychopathology

Bibliography

Achenbach T M 1997 What is normal? What is abnormal? Developmental perspectives on behavioral and emotional problems. In: Luthar S S, Bevack A, Ciccheppi D, Weisz J (eds.) *Developmental Psychopathology: Perspectives on Adjustment, Risk, and Disorder*. Cambridge University Press, Cambridge, UK, pp. 93–114

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. American Psychiatric Press, Washington, DC

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Press, Washington, DC

Ben-Porath Y S 1994 The MMPI and MMPI-2: Fifty years of differentiating normal and abnormal personality. In: Strack S, Lorr M (eds.) *Differentiating Normal and Abnormal Personality*. Springer, New York, pp. 361–401

Brodie H K H, Banner L 1997 Normatology: A review and commentary with reference to abortion and physician-assisted suicide. *American Journal of Psychiatry* **154**(6): suppl. 13–20

Butcher J N, Dahlstrom W G, Graham J R, Tellegen A, Kaemmer B 1989 *The Minnesota Multiphasic Personality Inventory—2 (MMPI-2): Manual for Administration and Scoring*. University of Minnesota Press, Minneapolis, MN

Cederborg A 1995 The negotiation of normality in therapeutic discourse about young children. *Family Therapy* **22**(3): 193–211

Eley T C 1997 Depressive symptoms in children and adolescents: Etiological links between normality and abnormality: A research note. *Journal of Child Psychology and Psychiatry* **38**(7): 861–65

Gardner F 1988 The parameters of normality in child mental health—A gender perspective. *Journal of Independent Social Work* **3**(1): 71–82

Goldman H H (ed.) 1995 *Review of General Psychiatry*, 4th edn. McGraw-Hill, New York

Hales R E, Yudofsky S C, Talbott J A (eds.) 1999 *The American Psychiatric Press Textbook of Psychiatry*. American Psychiatric Press, Washington, DC

Hathaway S R, McKinley J C 1943 *The Minnesota Multiphasic Personality Inventory*. University of Minnesota Press, Minneapolis, MN

Heiden L A, Hersen M (eds.) 1995 *Introduction to Clinical Psychology*. Plenum Press, New York

Hersen M, Kazdin A E, Bellack A S (eds.) 1991 *The Clinical Psychology Handbook*, 2nd edn. Pergamon Press, New York

Kazak A E, McCannell K, Adkins E, Himmelberg P, Grace J 1989 Perceptions of normality in familes: Four samples. *Journal of Family Psychology* **2**(3): 277–91

Kernberg O 1997 Perversions, perversity, and normality: Diagnostic and therapeutic considerations. *Psychoanalysis and Psychotherapy* **14**(1): 19–40

Luthar S S, Burack J A, Cicchetti D, Weisz J R (eds.) 1997 *Developmental Psychopathology: Perspectives on Adjustment, Risk, and Disorder*. Cambridge University Press, New York

McGoldrick M, Heiman M, Carter B 1993 The changing family life cycle: A perspective on normalcy. In: Walsh F (ed.) *Normal Family Processes*, 2nd edn. Guilford Press, New York, pp. 405–43

Millon T 1977 *Millon Clinical Multiaxial Inventory Manual*. Computer Systems, Minneapolis, MN

Millon T 1981 *Disorders of Personality*. Wiley, New York

Millon T 1994a *Millon Clinical Multiaxial Inventory—III Manual*. Computer Systems, Minneapolis, MN

Millon T 1994b *Millon Index of Personality Styles Manual*. Psychological Corporation, San Antonio, TX

Morey L C 1991 *The Personality Assessment Inventory Professional Manual*. Psychological Assessment Resources, Odessa, FL

Morey L C, Glutting J H 1994 The Personality Assessment Inventory and the measurement of normal and abnormal personality constructs. In: Strack S, Lorr M (eds.) *Differentiating Normal and Abnormal Personality*. Springer, New York, pp. 402–20

Nicholi A M (ed.) 1999 *The Harvard Guide to Psychiatry*, 3rd edn. Belknap Press of Harvard University Press, Cambridge, MA

Offer D, Sabshin M 1963 The psychiatrist and the normal adolescent. *Archives of General Psychiatry* **9**: 427–32

Offer D, Sabshin M (eds.) 1991 *The Diversity of Normal Behavior: Further Contributions to Normatology*. Basic Books, New York

Peterson C 1996 *The Psychology of Abnormality*. H. B College, Fort Worth, TX

Sabshin M 1988 Normality and the boundaries of psychopathology. Paper presented at the *First International Congress on the Disorders of Personality*, Copenhagen, Denmark

Sadock B J, Sadock V A (eds.) 2000 *Kaplan & Sadock's Comprehensive Textbook of Psychiatry*, 7th edn. Lippincott, Williams & Williams, Philadelphia
Strack S 1991 *Manual for the Personality Adjective Checklist (PACL) (rev.)*. 21st Century Assessment, South Pasadena, CA
Stoudemire A (ed.) 1998 *Clinical Psychiatry for Medical Students*, 3rd edn. Lippincott-Raven, Philadelphia
Turner S M, Hersen M (eds.) 1997 *Adult Psychopathology and Diagnosis*, 3rd edn. Wiley, New York
Waldinger R J 1997 *Psychiatry for Medical Students*, 3rd edn. American Psychiatric Press, Washington, DC
Walsh F 1993 Conceptualization of normal family processes. In: Walsh F (ed.) *Normal Family Processes*, 2nd edn. Guilford Press, New York, pp. 3–69

D. Offer, M. Sabshin, and D. Albert

Mental Health: Community Interventions

Community intervention in mental health involves two primary areas where mental health professionals have attempted to improve psychological well-being through interventions in the social and cultural environments of individuals and groups: (a) efforts to affect the communities themselves in terms of changing local norms and social networks, creating health-promoting processes, developing new social settings, and affecting policies which negatively impact on the well-being of citizens; and (b) interventions located in such community settings as schools and community organizations intended to provide both more easily accessible services and a wide range of interventions, most often preventive in nature.

In the United States, the impetus to move toward community interventions in delivering mental health services was linked to both political issues and conceptual developments in mental health and public health during the last 40 years of the twentieth century. Political support was provided by the 1963 Mental Retardation and Community Mental Health Centers Act (Public Law 88-164) signed by John Kennedy. Fueled in part by the report of the federally-sponsored Joint Commission on Mental Health and Mental Illness, this act was designed to localize mental health services in communities and expand the range of professional activities to include consultation and education as well as direct service.

Conceptually, the theoretical writings of Gerald Caplan (1964) provided an approach to the provision of services which was both community-based and focused on prevention as well as remediation. Caplan's vision included a variety of assumptions about human behavior and the role of the mental health professional which would be increasingly realized over time. First, the community of concern was the entire community at risk from psychological problems, not only those already identified as having difficulty. Service integration based on a knowledge of the community was highlighted, with services tailored to the specific populations of concern. Second, individual behavior was to be assessed in terms of its adaptive significance in its local context, arguing against the universalist assumption about signs and symptoms of disorder. Third, the role between professional and citizen was redefined from one of expert–client to one of collaboration.

Finally, Caplan over time (1989) elaborated on four specific components of his population-oriented approach which have become significant areas of community intervention in mental health: (a) Prevention (discussed below); (b) Crisis Theory, in which intervention is provided during times of acute stress due to either natural disasters such as floods or incidents such as schoolyard shootings; (c) Mental Health Consultation, where scarce professional resources are used in community settings such as schools to increase indigenous capabilities of high impact individuals such as teachers; and (d) Support Systems, where social support to deal with difficult life circumstance is either provided through professional intervention or through working with the ongoing support systems of individuals.

Caplan's overarching goal of promoting mental health and reducing the rates of disorder in a community has been articulated in public health as well. Altman (1995) suggests that a fundamental rationale for the importance of community-level intervention involves the understanding that shared determinant risks such as social class and social isolation themselves affect a broad variety of health outcomes. He outlines a five phase framework for conducting community-level interventions designed to benefit the public health, including mental health: (a) research, (b) transfer, (c) transition, (d) regeneration, and (e) empowerment. The focus of these phases is to move from a databased understanding of the problem to a collaborative community-level intervention which creates community empowerment over time. Feedback of data to local citizens is critical to this process.

1. *Population-oriented Community Interventions*

Most population-oriented community interventions have focused public health concerns. Thus, the Stanford Five-City Project involved a comprehensive, multilevel intervention to decrease the risk of cardiovascular disease in two intervention and three control cities. Multimedia campaigns were complemented by educational interventions in public schools and workplaces and health-related messages in restaurants and food stores. Results showed not only a greater decrease in cardiovascular risk scores in the experimental communities but increased reductions in overall smoking level, cholesterol level, and blood pressure (Farquar et al. 1990). While the specific

implications of the intervention with respect to mental health outcomes were not assessed, the project serves as a model for approaching mental health using a community-wide, multilevel, collaborative intervention paradigm.

Similar projects focusing on community mobilization and multilevel intervention around health issues have been reported in many other countries. Higgenbotham, Heading, McElduff, Dobson, and Heller (1999) evaluated the effects of a 10-year community intervention in the Australian coalfields to reduce coronary heart disease. In this project lack of overall community mobilization resulted in a focus on specifically affected subgroups in the community such as families of heart disease patients, school children, retired persons, and women responsible for family nutrition. The process aspect of conducting community-level intervention is outlined by Pena, Thorn, and Aragon (1994) in their description of a community intervention project in the El Limon, an isolated gold-mining community in Nicaragua. Focus was on the health and living conditions for the population of the town in general and the working environment of miners in particular. Local citizens—in particular miners—served as data collectors. Prevalence rates of various physical and mental health conditions, such as acute respiratory disease and nervous disorders, were calculated and now serve as the basis for developing community intervention programs.

Community-level interventions have also been targeted at social norms which support risk behaviors. Framing much of this work is diffusion of innovation theory which historically focused on how agricultural innovations were disseminated over time. Winett, Anderson, Desiderato, Solomon, Perry, Kelly, Sikkema, Roffman, Norman, Lombard, and Lombard (1995) outline this perspective and report on a series of papers by Kelly which employ diffusion of innovation theory in an effort to alter norms related to AIDS risk behavior. These studies focus on gay men living in small cities who frequent gay bars. With the help of bar owners, Kelly identified popular and influential persons who attended the bars and trained them to seek out others and deliver up-to-date information on AIDS transmission and prevention. Greater reductions in risk behavior among gay men frequenting gay bars were reported in the participating cities than in control cities.

Additional reports of efforts to alter community attitudes and responses have come from countries other than the United States. Fawcett, Heise, Isita-Espeje, and Pick (1999) report on a research and demonstration project in Izlacalco, Mexico to modify the response of friends, family, and the community at large to women suffering abuse from their husbands. This project began with focus groups and in-depth interviews which identified community attitudes and beliefs related to violence against women. A two-part intervention consisting of a 12-session consciousness-raising and skill-development workshop and a community-wide media campaign designed to reduce victim-blaming and increase community ownership of violence against women as a community-wide issue.

2. *Preventive Community Interventions in Community Settings*

Most community interventions geared toward changing aspects of the community have focused on public health issues social problems or issues such as AIDS, substance abuse, and violence. During the latter years of the twentieth century, however, an increasing number of community interventions in mental health focused on the prevention of varying disorders and the promotion of well being. These efforts were originally grounded in Caplan's designation of three levels of prevention: primary, secondary, and tertiary. Primary focuses on intervention before the appearance of disorder, while secondary involves intervention based on early identification of behaviors which are risk factors for later disorder. Tertiary prevention, in contrast, seeks to prevent the worsening of an already existing condition. Each of these levels of prevention has been implemented in community settings, and in each the concept of risk-reduction is central to the rationale for the intervention.

The prevention area has made considerable conceptual and empirical progress, and summaries of this progress are found in Felner, DuBois, and Adan (1991), Mrazek and Haggerty (1994), and Durlak and Wells (1997). The following community interventions only begin to outline the range of topics, populations involved, and process issues which are particularly salient in community interventions in mental health.

With respect to the area of tertiary prevention, perhaps the most well-known and discussed is the deinstitutionalization movement which began in the 1960s. As 'deinstitutionalization' implies, this movement was intended to mitigate the potentially negative effects of being in total institutions removed from family and friends by returning state hospital clients to the community. The intent was to provide sufficient supportive services in the local community to allow previously hospitalized individuals to rejoin their social networks and families and reduce the incidence of relapse. However, the development and integration of such community-based services was never adequately accomplished. Some deinstitutionalized individuals significantly improved as a result of this policy. However, in the absence of adequate community-based services, others were not able to handle community living. For them, the conditions necessary for effective tertiary prevention were not provided.

The potential of tertiary prevention for this group, however, is well documented in the work of Fairweather (Fairweather and Davidson 1987). In these projects, individuals released from state facilities

were placed in a 'lodge' located in the community. Here, as a group, they developed job skills and interpersonal skills honed by living in the same environment. Evaluations of this program showed positive effects in terms of gainful employment and lower return rates to state hospitals than control groups. This model has been built on in more recent times through the development of mutual help organizations for chronically mentally ill individuals which provide an ongoing setting, leadership roles, and social support.

Efforts at secondary prevention focus on proactively seeking out potential risk factors or precursors of later mental health-related problems. Many secondary prevention projects have been carried out in such community settings as hospitals and schools, including screening programs for toddlers, parent education, and school involvement programs for immigrant adults, and programs for parents of premature infants. In each instance an at-risk group is identified and educational or consultative services developed.

One of the best documented and successful secondary prevention programs is the Primary Mental Health Project (Cowen et al. 1996). This project has historically focused on children in the elementary grades who are screened for level of behavior problems in the early school years. Careful longitudinal research was done to develop screening procedures which included input from teachers and parents, and to ascertain whether or not those identified as having behavior problems manifested in the classroom were indeed more likely to develop academic and mental-health related problems later on. Interventions were then designed and carried out primarily by non-professionals in the direct child-aide role. Extensive evaluation data collected both at the end of the intervention program showed that intervention group children were rated as having better mental health outcomes such as less depression, anxiety, and learning difficulty than control group children. Further, follow-up data several years later showed that these gains were maintained. In addition, the program has been widely disseminated in school districts in the United States.

The area of primary prevention has taken hold in many different settings and across the life cycle. The blueprint from much of this work is found in the five step preventive intervention research cycle (Mrazek and Haggerty 1994). First a problem is identified and information relevant to its extent is gathered. A review of risk and protective factors related to the problem is then undertaken. Next, pilot studies based on a theoretical framework generated from a review of information of risk and protective factors are undertaken to gain preliminary information on the efficacy of the intervention and issues which need further investigation. Only after these steps have been undertaken is a full-fledged large-scale trial of the preventive intervention carried out. The final step is to develop materials and community relationships which promote the implementation and evaluation of the preventive intervention in the community. Information from community implementation becomes part of a feedback loop which informs community interventionists about how to improve the intervention or issues in transporting it to other communities.

The Mrazek and Haggerty (1994) report outlines some of the many areas where community interventions have been aimed at the prevention of disorder across the life cycle. With respect to programs designed to enhance social competence, for example, they describe multicomponent projects such as the Montreal Longitudinal-Experimental Study for disruptive seven year-old boys (Trembley et al. 1991). This project provided both home-based training for parents and in-school social skills training for their children rated by teachers as highly disruptive. Three years after the intervention, children who were rated as being less aggressive were placed in special classes, schools, or institutions less often, and initiated fewer acts of delinquency by age 12.

Programs aimed at adolescents, adults, and the elderly are also abundant. With respect to adolescents, programs designed to provide social influence, resistance training, and promoting norms which counter drug use are described, as are family and school-based programs to prevent conduct disorder. Successful programs have also been reported which redesign the structure and social organization of schools to promote both academic achievement and reduce dropping out. Here the promotion of parent involvement, school-based mental health services, and mechanisms to provide consistent and supportive adult–adolescent relationships, have been found to serve preventive functions.

With respect to adults, programs aimed at altering the marital relationship have been found to affect the level of adult depression and divorce rates. Occupational stress and job loss have also been identified as risk factors for depression. Here, interventions with unemployed workers designed to promote self-efficacy and job-seeking skills have shown success in terms of both mental health and economic outcomes. Household burden and limited English skills have been found to be risk factors for depression in low income, migrant Mexican American women, and interventions using indigenous helpers (*Servidoras*) in the local community.

Community interventions with the elderly have focused on the developmentally related risks of relationship loss, chronic illness, social isolation, and caregiver burden. The provision of social support is often central here, as is the involvement of non-professionals as intervention agents. For example, programs where widows provide a support and coping strategy-oriented intervention with other recently widowed women has been shown to affect well-being.

3. Concluding Comments

Overall, community interventions in mental health have been reported across a range of contexts and populations. Most are person-based, focusing on developing skills and competencies in individuals which will serve as protective factors in dealing with life circumstances. These circumstances may either involve ongoing issues such as promoting positive peer relations or may focus on dealing with particularly stressful situations such as parental divorce or the transition into a new school. However, other interventions target the social conditions and contexts within which individuals function. Together, the intent is to reduce risk factors and promote the development of positive coping of individuals in community contexts of importance to them.

Community intervention in mental health have also raised a new set of issues relating to methods, the research relationship between scholars and citizens, how to assess outcomes at the community level, and ethics. In addition, while many positive findings have been reported, questions remain about the conditions under which community interventions in mental health can fulfill the hopes on which they are built. The future of community interventions will involve an increasing focus on the contributions of factors at differing levels of the ecological environment to individual well-being; the ways in which multiple methods may complement each other in understanding the development and evaluation of community interventions, and models of collaboration in dealing with mental health issues in community contexts.

See also: Community Environmental Psychology; Community Health; Health Interventions: Community-based; Mental Health Programs: Children and Adolescents; Public Psychiatry and Public Mental Health Systems

E. J. Trickett

Mental Health Programs: Children and Adolescents

History demonstrates that the meaning of mental health programs for children and adolescents changes dramatically with changing views of children, adolescents, and mental health. Here, then, the subject of mental health programs for children and adolescents will be discussed from a historical perspective, one that explains current approaches in the context of the following major historical shifts:

(a) A shift away from competing perspectives, disciplines, and models working on a small scale and toward complementary perspectives, disciplines, and models working to create comprehensive systems of care.

(b) A shift away from segregating and toward integrating troubled children into 'normal' environments, with 'normal' defined in terms of family, school, community, and culture.

(c) A shift away from focusing on treatment and guidance only and toward focusing on prevention as well, especially on prevention for children deemed at risk.

This discussion will explain some of the main reasons for these shifts as well as explain why, despite the progress made in thinking about mental health programing, so many troubled children and adolescents still receive inadequate care or no care at all.

1. A Brief Historical Overview

Because of space limitations, this discussion focuses on developments in Europe and the United States Premodern Developments.

Just about every historical account of mental health programs begins with the fact that for a long time there were no such programs specifically for children and adolescents (Parry-Jones 1994). Rather, there were institutions for the mentally ill where young and old were housed together, sometimes cared for, but rarely treated. This fusion of programs for children, adolescents and adults no doubt reflected the relative lack of differentiation between childhood and adulthood—as evidenced by the fact that even the very young were forced to work for food and shelter. Prior to the modern era of mental health programing, the main developments in programing for children and adolescents with mental health problems had to do with making mental institutions more habitable.

This premodern focus on institutions without treatment is often mentioned to indicate the progress made in the way we now care for children and adolescents with mental health problems. However, it is possible that the appearance of progress masks a different reality, the reality that premodern times took care of most troubled children not within institutions but within fairly homogenous and defined communities—making the quality of life for those children actually superior to that of troubled children today. An absence of programs and systems of care does not mean, then, an absence of care. We will return to this theme of community and care at the very end.

1.1 The Nineteenth Century

The nineteenth century brought several important developments in the way children were understood and treated, developments that paved the way for the

next century's boom in both theory and programing. Two developments in particular bear special mention. The first is the development of public education and child labor laws that applied to all children. The second is the development of a science of psychopathology.

Public education and child labor laws did much to create a separation of childhood from adulthood and to usher in the modern era with its assumptions about children having 'rights.' However, initially, the rights of children were quite limited. Whether in home or school, most children experienced an authoritarian style of parenting and teaching that meant their mental health was informally measured according to how well they conformed to the rules and requirements of those in authority.

The nascent science of psychopathology did much to pave the way for the proliferation of twentieth century theories and diagnostic systems which have so shaped the development of today's mental health programs. But that science was faulty and limited with respect to its applicability to children. For example, Emil Kraepelin's system for classifying mental disorders was used in the same way that medical diagnoses defined discrete biologically based diseases. Furthermore, when referring to children with serious mental health problems, Kraepelin's system was simply extended 'downward' (Cantwell and Rutter 1994). There was, then, no separate field of study and no separate mental health system focusing on psychopathology in childhood.

1.2 The First Half of the Twentieth Century

The first half of the twentieth century is significant for establishing separate fields of child study, mental health agencies for children, and child welfare systems all of which developed into the current fields, agencies, and systems for addressing the needs of troubled children and adolescents. With the establishment of juvenile court systems and child welfare laws, attention was drawn to the underlying causes of young peoples' offenses and to developing child welfare agencies to deal with parental abuse and neglect. With Alfred Binet's work in intelligence testing, attention shifted to developing special education for children with mental health problems. And with the establishment of a science of childhood and child psychopathology, mental health programs for children and adolescents became increasingly theory-driven.

Three theoretical perspectives that were developed during this time deserve special mention, not only because of their popular appeal but also because of their influence on mental health programing. Sigmund Freud and psychoanalytic theory laid the groundwork for the development of pyschotherapies for children and for dramatic changes in residential treatment. John Watson, B. F. Skinner, and behavioral theory did the same for the development of the behavioral treatments so common today. Adolf Meyer and psychobiological theory forecast today's comprehensive programs for preventing mental health problems and for providing systems of care.

In contrast to Freud's psychoanalytic theory and Watson's behavioral theory, Meyer's psychobiological theory had an immediate impact on mental health programing. This theory (or perhaps framework is a better word—since Meyer and his followers placed so much emphasis on using common sense) changed the focus in discussions of mental illness from psychosis and disease to the prevention of psychosis and delinquency through attending to everyday problems. Meyer gave the term mental hygiene to a movement begun by Clifford Beers, a movement dedicated to the tasks of prevention and providing guidance. The most obvious and practical outcome of this movement was the establishment of child guidance clinics and the development of a new kind of interdisciplinary team made up of a psychiatrist, psychologist, and social worker. Child guidance clinics provided needed services, but they never came close to bringing prevention and treatment to an adequate scale.

The evolution of child psychiatry also belongs to this era. With the publication in 1935 of Leo Kanner's text *Child Psychiatry* (1962), this new field took on an identity as a go-between. The child psychiatrist was to be rooted in science and pediatrics while also being rooted in education and the nonscientific traditions of advising parents. The new discipline of child psychiatry was to take from the emerging disciplines of child study and psychoanalysis but follow Meyer's lead in valuing common sense. Kanner's vision, then, was of the child psychiatrist as the wise physician eschewiing extremes, dogmatic positions on theory and practice, and narrow identities. This new discipline, then, lent itself to being open to what is unique about particular child patients—which may have been why, in 1943, Kanner became the first to distinguish infantile autism (his term) from mental retardation and psychosis.

The story of child mental health and mental health programing for children and adolescents in the first half of the twentieth century would not be complete without mentioning two other strands of theory. The first, cognitive developmental theory, had as its major proponents the Swiss 'genetic epistemologist,' Jean Piaget, and the German born psychologist, Heinz Werner. The second, Gestalt psychology (not to be confused with Gestalt therapy) had as its major proponents Wolfgang Kohler, Kurt Lewin, and several others working first in Germany and later in the United States.

During the first half of the twentieth century, neither of these theoretical traditions had much influence on child mental health or on programing for children, but over time, their influence has been enormous, though often hidden. However, the field of child mental health

has yet to plumb all the riches within these two traditions. For example, Lewin's work provides a powerful framework for promoting family-centered systems of care—yet few today have been trained to represent and analyze child and family problems using a Lewinian framework. In a similar vein, the constructivist child psychology of Piaget and the organismic-developmental psychology of Heinz Werner offer powerful means to assess children with problems and to provide special education that matches up with children's developmental level and engages their interests and strengths. However, only a minority of special educators and clinicians have been trained in constructivist and organismic-developmental approaches to education and mental health.

There are, however, a few positive exceptions—for example, Edward Zigler. Zigler's training within both the Piagetian and Wernerian traditions prepared him well to make significant contributions to how mentally retarded children should be assessed and treated as well as contributions to programing for poor children 'at risk' for developing problems, including mental health problems. More will be said about Zigler when we come to discussing comprehensive programing and programing designed to prevent mental health problems.

1.3 The Second Half of the Twentieth Century

With respect to programing for children and adolescents with mental health problems, the second half of the twentieth century can be usefully divided into three periods as follows: (a) from 1950 to the mid-1960s and America's civil rights movement, (b) from 1965 to 1982 and Jane Knitzer's 'call to arms' in her book *Unclaimed Children* (1982), and (c) from 1982 to the present and the era of 'systems of care.'

During the first period, the development of psychoanalytic theory and its offshoots culminated in a number of innovative treatments and approaches to programing. This period also witnessed the cognitive revolution and the beginnings of an integration of developmental psychology with clinical child psychology. Also during this period, behavioral theory began to shift its principal home from the laboratories for academics to the clinics, hospitals, and special programs for children and adolescents with mental health problems (Lovaas 1977, Meichenbaum 1974, Patterson 1979).

During the second period, public policy instituted profound changes in mental health programing and special education—changes prompted less by theory and more by a newfound commitment to social justice. During this period, too, the rise of community mental health replaced residential treatment with alternative, community-based means of care. Also during this period, there was a renewed commitment to combining clinical research, assessment, and treatment with good science—as evidenced by the tremendous increase in money expended on mental health research, by breakthroughs in research on the biological determinants of several psychiatric disorders, by the commitment to revising systems for classifying disorders of childhood to make them more reliable and valid, and by scrutinizing therapies and mental health programs using scientific methods for evaluating program effects.

During the third and last period the new multidisciplinary field of developmental psychopathology has emerged as has the educational practice of 'inclusion.' But the hallmark of this period may well be the advent of 'systems of care' and its paradigm shift in mental health programing. Let us now backtrack a bit to discuss further the developments in each of these late twentieth century periods.

1.3.1 1950–65. When the second half of the twentieth century began, psychoanalytic perspectives dominated the mental health field. Many of these perspectives were constructive developments correcting old Freudian theory or extending psychoanalytic theory into areas uncharted by Freud (cf. Winnicott 1958). For example, Freud's daughter, Anna, did much to develop the method of treating children through analyzing their play. Erik Erikson extended psychoanalytic theory into the study of culture—and simultaneously gave us today's umbrella term, identity, for explaining adolescence. Margaret Mahler and Donald Winnicott corrected Freud's overemphasis of the Oedipal complex by demonstrating the centrality of object relations in the process of an infant and young child's 'individuating.' And Bruno Bettelheim, Fritz Redl and others extended psychoanalysis into the design of residential treatment. In a discussion of mental health programs, this last development requires some explaining.

Bettleheim, Redl, and others called their work 'milieu' therapy. By this they meant the shaping of virtually everything that went on in residential treatment to support a child's 'ego'—from picking out furniture that would withstand the not-so-occasional abuse of the troubled child to using a child's tantrums to promote insight through 'life space' interviewing. The milieu of these residential treatment centers worked, then, o help children and adolescents develop their own 'inner controls.' The writings of Redl and Wineman (1965) in particular provide fresh insights into what it takes to support troubled children and their development. Sadly, much of this extraordinary history in mental health programing is forgotten today—another example of how younger practitioners with new 'medicines' may fare no better than older practitioners who knew how to use older 'medicines' in extraordinary ways.

Psychoanalytic theory in the 1950s and early 60s also helped to spawn a number of offshoots that defined themselves as reactions to features in psychoanalytic theory and practice. Humanistic psychology and Virginia Axline's play therapy provide one example. Attachment theory and John Bowlby's work on the 'secure base' phenomenon provide another. During this period, Axline's play therapy had a widespread influence on how psychotherapy for children was conducted, but it was Bowlby's (1973) work on the dangers of separating child from caregiver that influenced changes in mental health programing. Bowlby's work and Renee Spitz's (1945) previous work on institutionalized children provided conceptual fuel for the later trend toward family preservation and keeping even pathogenic families together.

However, during this first period, psychoanalytic perspectives and their offshoots were not alone. This was the period when Piaget, Vygotsky, and, to a lesser extent, Werner became widely read—ushering in the so-called cognitive revolution. As mentioned earlier, these cognitive perspectives on children did not at first have much impact on mental health programing, but, in subsequent periods, their influence has been increasing.

As mentioned before, behavioral theory developed into a clinical tool—to be used everywhere that clinicians could define dysfunction in terms of 'target behaviors.' From autism to anorexia, from infancy to adolescence, behaviorists worked to demonstrate that 'All behavior is one.' Most of these new behavioral therapies derived directly from the work on operant conditioning pioneered by B. F. Skinner. They differed from today's behavioral treatments and programs mainly in the limited and sporadic nature of their interventions.

Finally, with respect to relevant developments during this period, family systems theory and family therapy developed rapidly to become a major alternative to the traditional therapies which focused on pathology within the child (Barnes 1994). Family systems theory demonstrated that children's and adolescents' dysfunctional behavior often serves important functions within a larger system, usually that of the family. The leading figures, such as Jay Haley and Salvadore Minuchin, gave the movement an almost swashbuckling style as they poked fun at traditional perspectives and challenged family members with provocative prescriptions and ways of labeling their family roles. But today, family therapists often do their work as part of a mental health team—as happens in many hospital based crisis centers where children and adolescents come to be stabilized, assessed, and referred.

1.3.2 1965–82. As we have already seen, not every major development in mental health programing results from developments in clinical theory or innovations in clinical practice. A good many developments result from changes occurring in the larger society. This was certainly true during America's 'civil rights era,' in the late 1960s and on into the 1970s.

The civil rights era began as a protest against the social injustices caused by racism, but it extended to the social injustices caused by sexism and discrimination against those with disabilities. The common themes throughout were those of integration and equal opportunity.

With respect to programing for children and adolescents with mental health problems, the civil rights movement's main influence was on the education of children with disabilities. Prior to this movement, segregating these children had been the rule. That changed with the passage of Public Law 94–142 mandating education of all children with disabilities and in the least restrictive environment.

This law effectively established a new and separate system of special education—one that was intended to address the educational needs of children and adolescents with mental health problems. For this reason, the new special education can be considered a mental health program. Furthermore, special education's being an entitlement program gives it added significance since poor and even middle-class families often could not afford adequate treatment. However, this new special education failed to deliver on its promise. Not only did children with problems continue to be unnecessarily segregated from the mainstream, but the education they received came to be a 'curriculum of control' (Knitzer et al. 1990).

The second major and relevant offshoot of the civil rights movement was the 'Head Start' program—a comprehensive program for poor children and their families and one implemented on a very large scale (Zigler and Muenchow 1992). Led by its first director, Edward Zigler, Head Start was and still is one of the most ambitious antipoverty programs for children. Furthermore, it embodies an approach to preventing mental health problems—one emphasizing improving the overall quality of living for poor families and their children. Within Head Start, parents have found jobs and job training, and children have found a variety of services to improve their health and education. Finally, from its inception, Head Start has been a social science laboratory—with built-in research funding for ongoing program evaluation and program development. Despite chronic problems associated with underfunding, Head Start remains an important program for thousands of children at risk for developing mental health problems. Furthermore, it represents a growing interest in and focus on multidisciplinary work directed at understanding the conditions underlying the phenomenon of resilience—as evidenced by the emergence of the new field of developmental psychopathology (Cicchetti 1990).

Within the field of mental health, this period also witnessed a major shift away from residential treat-

ment for adults and toward community mental health as psychological and sociological studies such as Irving Goffman's classic study (1961) emphasized the pathogenic influences of institutionalization. However, most funding for child and adolescent services continued to be applied toward inpatient and residential treatment. In fact, this period saw an increase in inpatient facilities for children and adolescents—as assessment units were established within for-profit hospitals to provide therapy as well as crisis management and assessment—with hospital stays lasting for several months.

The last major contribution of this period is not one that can be described as a movement. Hard science has long been the aim of those working to understand and treat mental illness. However, during this period, the enormous increases in funding for research on mental illness provided the needed support to make hard science in this area a reality. As a result, there were numerous breakthroughs in the field of psychopharmacology as well as scientific confirmation of there being biological causes for such serious disorders as schizophrenia.

As important as these developments were in their own right, they were, perhaps, not as important as the establishment of a pervasive scientific paradigm or frame for thinking about 'disorders' of childhood and adolescence. Nowhere is this more evident than in the push during this period to revise classification systems—to make them not simply more reliable and valid but also atheoretical so that a variety of researchers and clinicians holding different theoretical perspectives could pool data and communicate with one another. For perhaps the first time in its history, the field of child mental health was joined together by its scientific frame of thinking.

One of the casualties of this frame was psychoanalysis and psychoanalytic theory. Deemed largely untestable, psychoanalytic theory lost its previous commanding influence to the newer, highly testable drug therapies. Behavioral and cognitive–behavioral therapies have flourished under the scientific frame—with their emphasis on precise measurement, controlled comparisons, and the use of data to drive therapies (cf. Lovaas 1980, Meichenbaum 1974, Patterson 1979).

However, despite its obvious strengths, the scientific frame has its limitations, especially when it comes to mental health programming for children and adolescents. The complexities involved in predicting, preventing, and treating mental illness in childhood and adolescence are often too large to fit within the scientific frame. The result has been a need to rely on alternative frames—such as the one operating during the third period under the label, 'systems of care.'

1.3.3 1982–present. In 1982, Jane Knitzer's monograph, *Unclaimed Children*, documented what practitioners such as Fritz Redl had been arguing all along. She found that most children with serious emotional and behavioral disorders were not getting the services they needed. Neither the new special education nor the old mental health services were reaching the great majority of these children and adolescents. Furthermore, even when these children and adolescents were provided special services, they still were getting neither a proper education nor proper care.

The result of Knitzer's 'call to arms' was the development of a new paradigm and approach to mental health programing, an approach called 'Systems of Care.' Several features define its identity (Stroul and Friedman 1996). First, the systems of care approach calls for extraordinary hierarchic organization between federal, state, and local agencies. Second, the systems of care approach pushes for 'wraparound services,' that is, for sets of services designed to meet the special needs of individual families. Third, the systems of care approach provides support for families lost in a confusing array of services—mostly through the efforts of case managers who help families set goals, make plans, and advocate for needed services. Fourth, the systems of care approach requires professionals and agencies to be culturally competent—to value diversity and to develop skills for establishing positive relationships with those from different cultures. To understand the significance of this new paradigm and approach, we need mention only a few comparisons with old ways of programing.

Children with serious emotional and behavioral problems often come from families with serious problems as well. In the past, this fact led to the practice of separating children from their families, on the assumption that the children would have a better chance of improving. However, this practice ignored the fact that problematic families often remain committed to their children long after the commitments by professionals have ended. This practice also ignored the fact that dysfunctional families have strengths to enlist in helping their children. Most important, this practice ignored the fact that with additional supports, dysfunctional families can become better able to meet the needs of their children. The systems of care approach has taken these facts to heart by making family support central.

Older approaches to mental health programing also ignored cultural differences or treated cultures as being disadvantaged,' even dysfunctional. As a result, minority groups were subjected to subtle and not-so-subtle prejudice when being 'helped' by mental health professionals. For example, single mothers from cultures which value extended family members (grandparents, aunts, etc.), which feel that young children at night should be in their parents' bed, which feel children need occasional physical reminders (i.e., spanks) to behave properly—these mothers often found themselves being judged harshly or unfairly as

professionals referred to their 'broken homes,' their fostering overly dependent relationships between themselves and their children, and their authoritarian parenting styles. With the systems of care approach, the trend has been away from this kind of cultural insensitivity and toward providing services in culturally sensitive ways (Issaacs-Shockley et al. 1996).

Older approaches to programing also presented few options for treatment: outpatient office-based therapy, inpatient hospital care, and residential treatment. Furthermore, educational, welfare and mental health programs often functioned separately and sometimes in conflict with one another. With the systems of care approach, the effort has been to coordinate work done by different agencies and different professionals. This effort at coordination shows in a variety of ways. It shows in the increase in interagency teams. Most especially, it shows in professional–family relationships as case managers and home visitors work to get families and children the services they need. In the systems of care approach, then, old barriers between mental health and other kinds of child-related programs are broken down. For example, child welfare programs designed to protect abused and neglected children—often by removing children from their homes—now work to have mental health programs train parents and foster parents so that children temporarily removed from their homes can return safely to their families.

The systems of care approach and the advent of managed health care has dramatically changed the nature and functions of inpatient and residential treatment for children and adolescents (Woolston 1996). Now, the average length of stay in inpatient facilities has been reduced considerably—from several months to two weeks. The result has been a change in how inpatient facilities function. Where before, therapy was central, now inpatient facilities focus solely on crisis management (stabilization), assessment, and doing what they can to help establish the community-based system of care that children and adolescents will need when they leave the hospital. Residential treatment centers, too, are being hard-pressed to reduce length of stay and to turn outward toward the family and community. Financial constraints, public policy, and the difficulty of demonstrating empirically that inpatient and residential treatment are more effective, all have conspired to change the mental health programing landscape and to turn professionals toward developing community-based systems of care.

2. *Concluding Remarks*

The new era of systems of care may well have brought improvements in mental health programing. However, it has not brought a reduction in the percentage of children and adolescents with serious mental health problems. How can this be so? How can conceptual progress and progress in mental health programing fail to make a noticeable difference in the prevalence of problems?

There seem to be several reasons, not one. First, as long as schools exclude troubled children from the mainstream and educate them in classrooms with a curriculum of control, systems of care can do only so much. Second, as long as managed health care and inconsistent political support make it difficult for agencies to recruit trained professionals and to get done all that needs to get done to provide quality care, systems of care will be limited in their usefulness. Third, and perhaps most important, as long as there are mental health problems associated with the erosion of community life, systems of care will fail to make a significant difference. We may call our services 'community based' and our overall system 'community mental health.' However, surrounding families with systems of care will never substitute for surrounding families with communities that care. Perhaps community building will define the next era.

See also: Adolescent Health and Health Behaviors; Adolescent Vulnerability and Psychological Intervention; Health Promotion in Schools; Infant and Child Development; Theories of; Infant and Child Mortality in Industrialized Countries; Family Health

Bibliography

Barnes G G 1994 Family therapy. In: Rutter E, Taylor L, Hersov L (eds.) *Child and Adolescent Psychiatry: Modern Approaches*, 3rd edn. Blackwell Science, London

Bowlby J 1973 *Separation: Anxiety and Anger. Psychology of Attachment And Loss Series*. Basic Books, New York, Vol. 2

Cantwell D P, Rutter M 1994 Classification: conceptual issues and substantive findings. In: Rutter M, Taylor E, Hersov L (eds.) *Child and Adolescent Psychiatry: Modern Approaches*, 3rd edn. Blackwell Science, USA

Cicchetti D 1990 A historical perspective on the discipline of developmental psychopathology. In: Rolf J, Masten A S, Cicchetti D, Nuechterlein K H, Weintraub S (eds.) *Risk and Protective Factors in the Development of Psychopathology*. Cambridge University Press, New York

Goffman I 1961 *Asylums: Essays in the Social Situation of Mental Patients and Other Inmates*. Anchor Books, New York

Issaacs-Shockley M T, Cross B J, Bazrom K, Dennis I, Benjamin M 1996 Frameworks for a culturally competent system of care. In: Stroul B A, Friedman R M (eds.) *Children's Mental Health: Creating Systems of Care in a Changing Society*. Paul Brookes, Baltimore, MD

Kanner L 1962 *Child Psychiatry*, 3rd edn. Blackwell, Oxford, UK

Knitzer J 1982 *Unclaimed Children: The Failure of Public Responsibility to Children and Adolescents in Need of Mental Health Services*. Children's Defense Fund, Washington, DC
Knitzer J, Steninberg Z, Fleisch B 1990 *At the Schoolhouse Door*. Bank Street College of Education, New York
Lovaas O I 1980 *The Autistic Child: Language Development Through Behavior Modification*. Holsted Press, New York
Meichenbaum D 1974 *Cognitive Behavior Modification: An Integrative Approach*. Plenum, New York
Parry-Jones W 1994 History of child and adolescent psychiatry. In: Rutter M, Taylor E, Hersov L (eds.) *Child and Adolescent Psychiatry: Modern Approaches*, 3rd edn. Blackwell, London
Patterson G R 1979 Treatment for children with conduct problems: a review of outcome studies. In: Feshback S, Fraczek A (eds.) *Aggression and Behavior Change: Biological and Social Processes*. Praeger, New York
Redl F, Wineman D 1957 *Controls from Within*. Free Press, Illinois
Spitz R A 1945 Hospitalization: an inquiry into the genesis of psychiatric conditions of early childhood. In: Eissler R S, Freud A, Hartman H, Kris E (eds.) *The Psychoanalytic Study of the Child*. Yare University Press, New Haven, CT, Vol. 1
Stroul B A, Friedman R M 1996 The system of care concept and philosophy. In: Stroul B A (ed.) *Children's Mental Health: Creating Systems of Care in a Changing Society*. Paul Brookes, Baltimore, MD
Winnicott D W 1958 *Collected Papers: Through Pediatrics to Psychoanalysis*. Tavistock Publications, London
Woolston J L 1996 Psychiatric inpatient services. In: Lewis M (ed.) *Child and Adolescent Psychiatry*, 2nd edn. Williams & Williams, Baltimore, MD
Zigler E, Muenchow S 1992 Head Start: *The Inside Story of America's Most Successful Educational Experiment*. Basic Books, New York

W. G. Scarlett

Mental Illness, Epidemiology of

Epidemiology in the field of mental disorders is the study of their distribution in populations and of the risk factors associated with their onset and course. Epidemiological methods provide tools to conduct research in etiology and genetics, and serve as the basis of outcome studies, clinical trials, and public health research. Key requirements of epidemiological studies are (a) the definition of the target population (e.g., total population or representative samples of a region or a country, or representative fractions thereof); (b) explicit, reliable and valid criteria for disorders or what constitutes a case (e.g., symptoms or syndromes); (c) explicit, reliable and valid criteria for variables and factors that might be associated with a disease (i.e., gender, social class, genetic factors, infectious agents), and use of epidemiological methods for measuring outcome occurrence (e.g., *prevalence*, defined as the proportion of individuals affected by a particular disorder at a specified time period or point in time, and *incidence*, defined as the number of new cases among persons in a population, who were initially free from the disorder and developed the disorder over a given period of time, such as a lifetime, 30 days, or 1 year), and (d) for measuring associations (risk and protective factors), and impact (i.e., course of illness, associated impairments, and disability). Epidemiology can be divided further into two interrelated orientations and methodologies, namely *descriptive epidemiology*, aiming at measuring the extent of mental disorders in the community, and *analytic epidemiology* (e.g., longitudinal cohort designs, case-control studies), that focuses on understanding the etiology of mental disorders. Examples are laboratory or genetic markers to test etiologic hypotheses. Epidemiology offers some unique and promising research strategies for describing and clarifying the nature, etiology, course, and outcome of mental disorders, because patients in treatment settings usually represent a small and highly selective segment of the full spectrum of mental disorders. Thus, findings for risk factors, prognosis and etiology might be biased by selection biases as well as the severity of the studied condition (Friis and Sellers 1999).

Originally, epidemiology and epidemiological research designs were developed to study chronic and infectious diseases, and have only been adapted slowly for use in mental and behavioral disorders since the 1950s, largely due to the initially controversial status of past diagnostic classification systems for mental disorders. The complex manifestations, poorly understood etiologies, and variability of the course of mental disorders, in addition to their previously low diagnostic reliability, were often difficult to capture in basic epidemiologic designs. This difficulty was particularly true for those involving one or two points in time (e.g., cohort studies). In addition, risk factors for mental disorders can be as difficult to conceptualize and assess as variables of relevance for mental disorders. Despite early difficulties, these problems have been partly overcome in recent years.

Strongly related to the introduction of more reliable and valid classification systems for mental disorders, based on explicit criteria and operationalized diagnoses in 1980 (APA 1980, 1994) and the increasing availability of structured and standardized diagnostic assessment instruments, the last two decades of the twentieth century witnessed an unprecedented progress in epidemiological research on mental disorders. The field has advanced in terms of number of studies, their scope, degree of methodological sophistication, and their linkages to allied disciplines such as psychology, neurobiology, and sociology (see Table 1). What started in the 1950s as a scientifically problematic and quite restricted area of 'psychiatric epidemiology' (Robins 1992) has now reached firmer ground and opened up new perspectives.

Table 1
Epidemiology of mental disorders in the 1990s

Progress	Deficits
Availability of large-scale general population studies (prevalence, sociodemographic correlates, comorbidity, age of onset)	Incomplete coverage of diagnoses (deficits: somatoform, sleep disorders, comorbidity with somatic conditions, personality disorders)
Increasingly sophisticated sampling, design and statistical procedures	Childhood disorders and the developmental continuity of adult mental disorders
International collaboration (joint re-analyses on prevalence, risk factors, comorbidity)	Explorations of thresholds for and boundaries between disorders
Documentation of associated impairments and disabilities	Assessment of diagnosis-specific disability and impairment
Crude documentation of poor recognition and intervention	Service utilization and need assessment
Improvements in diagnostic instruments and contributions to psychopathology and diagnostic classification	Lack of prospective–longitudinal studies to identify vulnerabilities and causal risk factors and natural course over lifespan
Slowly increasing number of prospective–longitudinal and family-genetic studies: contributions to nosological research	Linkage to genetic research and neurobiology

1. Descriptive Epidemiological Findings

Starting with the landmark Epidemiological Catchment Area (ECA) program in the early 1980s (Robins and Regier 1991), increasingly sophisticated large-scale studies in the general population have made it evident that mental disorders are very frequent disorders of the brain, affecting almost every person over their life courses. These studies have also highlighted the correlates and the variability in the manifestations of disorders of emotion, cognition, and behavior, demonstrating that mental disorders are not as uniform as previously believed in terms of their risk factors, courses, outcome, associated disabilities, and impairments.

1.1 Prevalence and Correlates of Mental Disorders in the Community

Community surveys in the 1990s, such as the National Comorbidity Survey (NCS) (Kessler et al. 1994), the Health Examination and Interview Survey—Mental Health Supplement (Wittchen et al. 1999) and similar studies in other countries (WHO 2000) have estimated the lifetime rates for mood, anxiety, and substance disorders alone to range between 36.3 percent and 48.6 percent. Across studies point-(30-day) prevalence estimates vary between 10–17 percent, in spite of differences in coverage of diagnoses, design, and cultural setting. Relatively stable estimates were also found for specific disorders, such as panic disorder (lifetime estimates: 3–4 percent, point prevalence: 1–2 percent) as well as psychotic disorders (1–3 percent).

Similar convergent evidence has also become available from these studies with regard to sociodemographic correlates and impairment. For example: (a) The majority of people affected by a mental disorder usually report the onset of their condition in adolescence, largely due to a predominant early onset of anxiety and substance-use disorders (alcohol, drug abuse, and dependence), whereas depressive disorders occur at higher ages and over the whole lifespan; (b) Females are 2–3 times more frequently affected by various depressive and anxiety disorders; (c) Males more frequently develop substance-use and antisocial personality disorders; (d) Low social class has not been confirmed consistently as being associated with the onset of mental disorder in all countries; and (e) The presence of mental disorders has been found to be associated consistently with increased rates of disability and impairment days, which vary by type of disorder and comorbidity.

Data also reveal (WHO 2000) that only about 22 percent of cases in Canada and the USA, and only slightly more in the Netherlands (31.7 percent) and Germany (29.2 percent) receive any type of treatment. The vast majority in the three countries were cared for exclusively in the primary health care sector, with only few receiving treatment by mental health specialists, even though treatment was merely defined as 'any treatment contact,' irrespective of appropriateness in terms of type, dose, and duration of treatment. A further disturbing finding is the fact that the majority of patients delay many years after the first onset of their disorder before getting treatment (Olfson et al. 1998). It is not entirely clear to what degree these low treatment rates are due to patients' poor helpseeking behavior, structural barriers in the health care system, or health service providers' lack of recognition and diagnostic skills. However, the fact that low treatment rates are not confined to countries that do not cover

health care and treatment cost by insurance plans suggests that there are many reasons for this problem.

Another important key finding of studies is that, even in community samples, a high proportion of subjects suffer from more than one mental disorder. In the NCS (Kessler et al. 1994), the vast majority of people with one mental disorder had other comorbid conditions as well; 54 percent of the population with a lifetime mental disorder had three or more lifetime comorbid conditions, and among those with 12-month diagnoses 59 percent had three or more additional disorders. Most frequent patterns of comorbidity were among anxiety, affective, and substance-use disorders. The consistency of comorbidity findings across studies has made it clear that comorbidity is a fundamental characteristic of most mental disorders. Comorbidity has been shown to be neither simply an artifact of random association of frequent disorders, nor influenced by helpseeking behavior or methodological aspects of the studies (Wittchen 1996). Comorbidity has specific effects on the degree of suffering as well as the likelihood of seeking professional help. Further comorbidity has been demonstrated to have important etiological, pathogenetic, clinical, and psychopathological implications that have become a major research topic in mental health research (Wittchen et al. 1999a).

Beyond the mere demonstration of the size of the problem of mental disorders, there is now increased national and international collaboration on the basis of more stringent methodologies and designs, which allow for powerful coordinated cross-national reanalyzes with regard to risk factors as well as worldwide studies such as the ongoing World Mental Health 2000 study (Kessler 1999). The significance of such endeavors can be highlighted by findings of The Cross-National Collaborative Group (1992) that demonstrated in powerful analyses that the rates of depressive disorders are increasing in each successively younger birth cohort in industrialized countries. In addition, age of first onset of depressive disorders had declined into early adolescence. This finding has prompted a series of studies that explore the reasons for a continuing increase in depression rates, and further led to a reconsideration of projections with regard to the burden of depressive disorders in the future.

1.2 Prevalence of Mental Disorders in Primary Care

Although large-scale comprehensive studies about the prevalence of mental disorders in primary care are lacking, there is considerable evidence from international collaborative primary care studies on selected anxiety and depressive disorders (Üstün and Sartorius 1995) that these disorders are highly prevalent in primary care settings around the world: point prevalence estimates for depressive disorder range between 8–11 percent, and those for anxiety disorders range between 4–10 percent. Such studies have also revealed—consistent with more recent studies (Wittchen 2001)—that mental disorders in primary care are usually poorly recognized and rarely treated. Only every second case is recognized, and of those recognized, only one-third receives some type of state of the art treatment.

1.3 The Burden of Mental Disorders

Along with the increased emphasis of epidemiological research on the role of general and diagnostic-specific disabilities and impairment, community studies have also been able to demonstrate the substantial burden that mental disorders have on the subjects' lives, social functioning, and interpersonal environment. Based on available epidemiological data, the Global Burden of Disease study (Murray and Lopez 1996) showed that the burden of mental disorders has been underestimated greatly. Of the ten leading causes of disability worldwide in 1990, measured in years lived with a disability, five were psychiatric conditions: Unipolar depression, alcohol use, bipolar affective disorders, schizophrenia, and obsessive-compulsive disorder.

2. Unresolved Issues in Descriptive Epidemiology

Despite considerable progress, descriptive epidemiological knowledge is still limited, given in particular that only a few of all mental disorders have been investigated. Other areas with significant deficits are (a) mental disorders in children, adolescents, and the elderly; (b) the dimensional and subthreshold characterization of mental disorders; (c) a more comprehensive identification of risk factors; (d) the range of associated disabilities; and (e) the use of health services with emphasis on national and cultural variations.

2.1 Coverage of Disorders

Epidemiological research is still far away from having examined the full range of clinically significant psychopathological conditions. Noteworthy deficits exist, for example, with regard to somatoform disorders, sleep disorders, substance-related disorders beyond abuse and dependence, personality disorders, and some forms of childhood disorders. Each of these conditions merits considerable epidemiological attention from a public health, and even more from a pathogenetic, perspective, because they have different onset and course characteristics, and are frequently comorbid. For example, somatoform disorders (i.e., pain disorders, hypochondriasis, somatization), which were shown to be highly prevalent conditions that start early in life and constitute a major burden on the health care system, have rarely been included in past community surveys. Similarly, sleep disorders, which rank high as a principal reason for primary care consultations, have not been studied systematically in

community surveys. The unclear nosological status of these conditions may be responsible for this neglect. Obstacles to research on personality disorders are of a methodological nature because there are not sufficiently reliable and time-efficient assessment tools available.

For childhood conditions, the challenge lies in consensus regarding appropriate choice of age/developmental-stage-specific diagnostic assessments. Researchers must also concur on the degree to which multiple sources of information (i.e., parents and teachers) can be combined into one coherent strategy that mirrors the continuity from childhood to adolescent and adult mental disorders. Despite growing collaboration, there is still a remarkable division among investigators regarding epidemiological designs and methods used in childhood, adolescent, and adult mental disorder research. Inherent in this division is the issue of the developmental continuity of psychopathological features, which also touches partly on the ongoing controversy of dimensional versus categorical measures in epidemiological studies of mental disorders. Intensified research in this area is needed, especially because of evidence that most adult mental disorders begin in adolescence.

2.2 Subthreshold Conditions and Dimensional Approaches

With few exceptions, descriptive epidemiological evidence is based on a restricted range of threshold diagnoses, assessed with diagnostic instruments, such as the WHO Composite International Diagnostic Interview (CIDI) without sufficiently detailed consideration of available duration, persistence, and severity information. The exclusive reliance on categorical threshold diagnoses carries substantial risks of artifactual explanations (such as in comorbidity analyses) and fails to acknowledge the dimensional nature of most expressions of psychopathology. The DSM-IV (APA 1994), as the most frequently used classification system in research into mental disorders, makes only a few attempts to derive discrete categories that are mutually exclusive and lead to a single classification of an individual. In fact, the system was intended to stimulate further development of research on the thresholds for and boundaries among disorders (APA 1994). Available data with their primary reliance on categorical diagnostic decisions are not an optimal source for modifying diagnostic systems (Wittchen et al. 1999c).

2.3 Assessment of Disability and Need of Treatment

Over recent years, increasing recognition has emerged that diagnoses of mental disorders themselves cannot appropriately answer questions about need for care, service utilization, and treatment match. Such domains are thought to be determined largely by the individual's functioning status and disability. Epidemiological studies from the last two decades of the twentieth century have not provided coherent and comprehensive information about these service-related issues, and do not allow for reliable characterizations of diagnosis-specific degrees of disability with regard to various social roles. Also, they do not provide sufficiently detailed data about helpseeking behavior, service utilization, and services needs (Regier et al. 1998). Recently, the World Health Organization has started collaborative, systematic, conceptual, and psychometric developmental work to design generic and diagnosis-specific assessment instruments modules to assess disability. These measures might also provide a better basis for need assessment in the area of treatment.

2.4 Assessment Instruments and Diagnostic Classification

Population studies and methods-related epidemiological work have been instrumental in the improvement of diagnostic classification systems for mental disorders. Reliable symptom and diagnostic assessment instruments of mental disorders have been created for use in epidemiology and clinical research. This work has not only significantly influenced the content and structure of clinical instruments: Structured Clinical Interview for DSM-IV (SCID) (First et al. 1997); Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Brugha et al. 1999b) and non-clinical tools: Composite International Diagnostic Instrument (CIDI) (WHO 1990), but also played an important role in the revision processes of diagnostic classification systems (DSM-IV and ICD-10).

Yet these conceptual models of mental disorders are not, and have never been, a paragon of elegance, nor have they resulted in sufficiently neat and crisp classification systems that match basic research findings, and clinical management and decision-making. The introduction of these operationalized and descriptive manuals have resulted in greater diagnostic reliability and consistency in the use of diagnostic terms around the world. In particular, they have been a key prerequisite for epidemiological progress. However, major problems (i.e., thresholds, overlap, and comorbidity), which remain a source of significant dissatisfaction and controversy, will require extensive future work.

At the center of this agenda is the need for convincing clinical and nosological validation in terms of prognostic value and stability, family and genetic findings, and laboratory findings for almost all mental disorders, allowing a sharper genotypical and pheno-

typical classification. Current diagnostic classification manuals (DSM-IV and ICD-10) deliberately do not contain mutually exclusive diagnostic categories in order to simulate research inquiries into diagnostic boundaries and thresholds—a valuable target for epidemiological research. Consensus is lacking on how to tailor appropriate psychopathological assessment instruments that are able to address such threshold issues appropriately. Further, despite the substantial scientific exploration and examination that went into instruments like the CIDI (WHO 1990) and the SCAN (Brugha et al. 1999a), some basic problems of reliability and validity inherent in the assessment of some mental disorder are yet unresolved.

At the center of discussion is no longer the traditional question of whether to go for categorical or dimensional, but rather to what degree and for which psychological conditions 'clinical judgment and probing' should be regarded as a mandatory core element. Empirical evidence needs to be gathered to determine in which diagnostic domains and clinical instruments are superior to fully standardized instruments, such as the CIDI, which try to identify explicitly the latent variables behind the vagueness of clinical judgment. Progress in the resolution of this issue will also offer ultimately more appropriate strategies in resolving the 'gold standard' question of the optimal strategy for validating epidemiological instruments (Brugha et al. 1999a; Wittchen et al. 1999c) (see *Nosology in Psychiatry*; *Differential Diagnosis in Psychiatry*)

3. *Longitudinal Studies and Causal–Analytic Epidemiology*

Despite a slowly growing number of costly large-scale prospective–longitudinal studies (Grasbeck et al. 1998, Wittchen and Nelson 1998) that have become available, knowledge about natural course, longitudinal stability of symptoms, and comorbid associations, as well as vulnerability and risk factors for the onset and persistence of mental disorders is still quite meager. This deficiency is particularly true for children, adolescents, and young adults; Thus it remains difficult to characterize mental disorders reliably by patterns of course and incidence across the life span. Further the 'causal risk factors' status (Kraemer et al. 1997) has not yet been established for most putative risk factors of mental disorders. At this point it remains unclear what might be cause, consequence, or a mere correlate.

Prospective longitudinal studies such as the Dunedin and the Early Developmental Stages of Psychopathology study (Wittchen and Nelson 1998) can advance our etiological knowledge through the collection of information on early signs and risk factors, which have been gleaned from high-risk studies, and from studies on protective factors and processes related to the onset of mental disorders. From such studies a substantial body of evidence has already become available with regard to partly diagnosis-specific, partly general risk factors for many disorders. Well established risk factors include: A family history of mental disorders, the effect of threatening or stressful disruptions in the individuals' immediate social environment, childhood loss, abuse, and trauma. There is also further evidence for more complex interactions: Familiar genetic factors may enhance the effects of loss events; patterns of symptom progression from preceding early anxiety disorders relate to the risk of developing secondary depression; and certain childhood temperamental traits increase the risk of subsequent mental disorders. What remains, however, is the large challenge of completing the complex vulnerability–stress interaction, and understanding how vulnerability and risk factors from formerly competing paradigms interact with each other in specific disorders or groups of mental disorders.

3.1 *Linking Epidemiology Closer to Clinical and Basic Neurobiological Research*

The limitations of convenience samples from clinical and other settings for etiological and pathogenic research are becoming recognized increasingly. Limitations include the risks of artifactual findings, the over- and underestimation of effects, confoundation by comorbidity, and the impossibility of establishing causal risk factors for first onset of a disorder. Prospective–longitudinal studies and causal–analytic designs in representative population samples will be of key importance to overcome these limitations. A requirement for such designs is a comprehensive evaluation of the epidemiological triad of host (i.e., genetic variables or temperamental predispositions), agent (life stage transitions or life stress), and environment (social processes or environmental agents). A considerable challenge for such studies is the identification of how and when certain environmental factors potentiate or protect against genetic and biological vulnerability factors. Neuroscience, and in particular, genetic research, is likely to be essential for our future understanding of the etiology and pathogenesis of mental disorders, and ultimately a better genotypic classification. The complexity of longitudinal designs, and the time and costs involved, are substantial; however, findings promise a better understanding of the relevant developmental pathways of specific mental disorders and their interrelationships among each other (comorbidity) (see Table 2).

Family and other genetic studies have demonstrated convincingly that an individual's genetic makeup is an important factor in determining vulnerability for almost all mental disorders (Merikangas 1999). Methods of genetic epidemiology for studying risk factors and etiology of familiar diseases appear to be the most promising ways to unravel the complex

Table 2
Relevance of epidemiology of contemporary psychiatry

Examination of the generalizability of clinical samples
Identification of potential etiologic factors and correlates of psychiatric disorders:
• Sex differences in psychiatric disorders
• Age-specific patterns of onset and offset
• Cohort effects
• Environmental agents (e.g., viral exposure, toxins, diet, and agents)
• Geographic patterns (e.g., cultural pattern of expression, risk factors, migration)
Genetic and family-genetic studies in community samples
Estimation of attributable risk of neurobiological and genetic factors
Identifying core psychopathological processes over the lifespan
Identifying common and unique vulnerabilities and risk factors for disorders
Establishing of norms for biological and genetic markers
Testing the generalizability of biological markers in community samples

Source: Adapted from Merikangas 1999.

mechanisms through which genes exert their influence. The integration of population genetics and epidemiology is critical for determining the attributable risk of particular DNA markers for disease as well as how environmental conditions increase or reduce expressions of genetic vulnerability. Such studies might be directed at states or traits conferring susceptibility. In the near future it can be expected that such types of genetic epidemiologic studies will identify some of those genetic mechanisms that place individuals at increased risks for disorders such as substance dependence, anxiety, and depression (see *Mental Illness, Genetics of*).

4. Need Evaluation and Its Implications on Interventions in the General Population

Since the 1980s, quite comprehensive, interdisciplinary mental health systems of providers have emerged in most industrialized countries. However, given the size of the problem of mental disorders it is evident that no system could afford comprehensive treatments for all affected persons. Epidemiological data are a key prerequisite for identifying deficits and problems in health care systems, and offering guidance in service planning and resource allocation. At a time of increasing numbers of effective pharmacological and psychological treatments, competing provider models for mental disorders, and tighter health care budgets, epidemiology can be expected to gain further importance, especially in highlighting the efficacy of interventions. But certainly the available studies and data do not yet provide us with appropriate level of detail for this important task (Regier et al. 1998; Wittchen 2000) (see Table 3).

Core elements of need assessment are a reliable definition of a disorder, clearly defined associated disabilities, and existing effective interventions (taking into account both their limitations and modes of delivery). Nevertheless, despite the existence of many effective psychological and drug treatments shown to prevent disability, relapse, chronicity, and suffering for many mental disorders, few epidemiological studies are available that answer with sufficient detail core questions such as: How many anxiety or depressive disorders were treated by psychiatrists, psychotherapists or other types of providers; how many were treated by medication or some form of psychological treatment, or even treated at in- or outpatient institutions; how many in need are treated appropriately; and how many remain untreated. Instead, mostly crude data regarding rates of service utilization and unmet service needs are available that emphasize predominantly the role of primary care physicians. Even though the quite limited role and poor recognition abilities of primary care physicians for mental disorders has been noted repeatedly throughout the world (Üstün and Sartorius 1995), it seems that traditional psychiatric epidemiology still seems to favor this model strongly. Neglected are more comprehensive service utilization and need assessment strategies for specialist treatment and other interventions. Another unfortunate deficit in this field is its strong emphasis on so-called 'major mental disorders,' such as psychotic disorders, as opposed to 'minor and neurotic disorders.' Although this oversimplified dichotomy has clearly outlived its usefulness and scientific justification since the 1980s, many health utilization

Table 3
Types of questions related to comprehensive need assessment

Stepwise questions	Epidemiological activity
What is a case?	Prevalence, incidence
What do patients need?	Subjective/objective
Do patients get what they need?	Inappropriate use and service delivery
How can structure and process of services/treatment be improved?	Evaluation of changes, stepwise planning of improvements

and needs surveys still overemphasize major morbidity, neglecting what are actually the most prevalent and persisting 'minor morbidity conditions,' that cause by far the greatest financial burden of all disorders (Rice and Miller 1998). Current attempts that merely link diagnosis with measures of disability in order to improve need assessment (see above) might not solve this critical problem. Rather, they might result once again in an inappropriately strong emphasis on the most severely ill, neglecting those in earlier stages of their illness process who might profit most from modern treatment methods.

The challenge for epidemiology on a cross-national and international basis lies in a systematic comparison of such competing models. Development of appropriate assessment instruments for use in epidemiological studies to identify advantages and disadvantages of each of these perspectives, in terms of legal, cost, comprehensiveness, and effectiveness issues has been identified as a necessary first step (Jagger et al. 1998). Current perspectives on this issue seem to overemphasize two search strategies: the development of reliable and valid measures of disability (Regier et al. 1998), and the search for other 'marker' variables of those in greatest need or the 'most severe.' This perspective might fall too short. In search of improved approaches for comprehensive need assessment and evaluation, future epidemiological research should additionally emphasize: (a) a more comprehensive assessment of helpseeking behaviors that covers the full spectrum of professional providers in the respective country or region; (b) a wider coverage of types of interventions received, contingent on the availability of treatments in that country for that diagnosis; and (c) a detailed inquiry into perceived barriers to recognition and treatment.

5. Conclusion

In light of the ongoing rapid developments in neuroscience, clinical psychology, and psychiatry, as well as public health research, epidemiology can be expected to play an increasingly important role in basic, clinical, and public health research of mental disorders. The key challenge will be to understand how multiple risk and vulnerability factors interact over time and over the lifespan in producing a single or multiple mental disorder over brief or longer periods of time.

See also: Dementia: Overview; Depression; Depression, Clinical Psychology of; Mental and Behavioral Disorders, Diagnosis and Classification of; Mental Health and Normality; Mental Health: Community Interventions; Public Psychiatry and Public Mental Health Systems; Schizophrenia

Bibliography

APA (American Psychiatric Association) 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd end. APA, Washington, DC

APA (American Psychiatric Association) 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. APA, Washington, DC

Brugha T S, Bebbington P E, Jenkins R 1999a A difference that matters: Comparisons of structured and semistructured psychiatric diagnostic interviews in the general population (Editorial). *Psychological Medicine* **29**: 1013–20

Brugha T S, Nienhuis F J, Bagchi D, Smith J, Meltzer H 1999b The survey form of SCAN: The feasibility of using experienced lay survey interviewers to administer a semi-structured systematic clinical assessment of psychotic and nonpsychotic disorders. *Psychological Medicine* **29**: 703–11

First M B, Spitzer R L, Gibbon M, Williams J B W 1997 *User's Guide for the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I)*. New York American Psychiatric Press, New York

Friis R, Sellers T 1999 *Epidemiology for Public Health Practice*. Aspen, Gaithersburg, CO

Grasbeck A, Hansson F, Rorsmann B, Sigfrid I, Hagnell O 1998 First incidence anxiety in the Lundby Study: Course and predictors of outcome. *Acta Psychiatrica Scandinavica* **98**: 14–22

Hansson L, Muus S, Vinding H R, Gostas G, Saarento O, Sandlund M, Lonnerberg O, Oiesvold T 1998 The Nordic Comparative Study on Sectorized Psychiatry: Contact rates and use of services for patients with a functional psychosis. *Acta Psychiatrica Scandinavica* **97**: 315–20

Jagger C, Ritchie K, Bronnum-Hansen H, Deeg D, Gisbert R, Grimley Evans J, Hibbett M, Lawlor B 1998 The Medical Research Council Cognitive Function and Ageing Study Group Peerenboom R, Polge, C van Oyen H. Mental health expectancy—the European perspective: A synopsis of results presented at the Conference of the European Network for the Calculation of Health Expectancies (Euro-REVES). *Acta Psychiatrica Scandinavica* **98**: 85–91

Kessler R C 1999 The World Health Organization Consortium in Psychiatric Epidemiology (ICPE): Initial work and future direction—the NAPE Lecture (1998). *Acta Psychiatrica Scandinavica* **99**: 2–9

Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshleman S, Wittchen H-U, Kendler K S 1994 Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**: 8–19

Kraemer H C, Kazdin A E, Offord D R, Kessler R C, Jensen P S, Kupfer D J 1997 Coming to terms with the terms of risk. *Archives of General Psychiatry* **54**(4): 337–43

Krueger R F, Caspi A, Moffit T E, Silva P A 1998 The structure and stability of common mental disorders (DSM-III-R): A longitudinal–epidemiological study. *Journal of Abnormal Psychology* **107**(2): 216–27

Lieb R, Wittchen H-U, Höfler M, Fuetsch M, Stein M, Merikangas K R 2000 Parental psychopathology, parenting styles, and the risk of social phobia in offspring. A prospective-longitudinal community study. *Archives of General Psychiatry* **57**: 859–66

Merikangas K R 1999 Editorial: The next decade of psychiatric epidemiology. *International Journal of Methods in Psychiatric Research* **8**(1): 1–5

Murray C J L, Lopez A D (eds.) 1996 *The Global Burden of Disease*. World Health Organization, Geneva, Switzerland
Oldehinkel A, Wittchen H-U, Schuster P 1999 Prevalence, 20-month incidence and outcome of unipolar depressive disorders in a community sample of adolescents. *Psychological Medicine* **29**: 655–68
Olfson M, Kessler R C, Berglund P A, Lin E 1998 Psychiatric disorder onset and first treatment contact in the United States and Ontario. *American Journal of Psychiatry* **115**: 1415–22
Regier D A, Kaelber C T, Rae D S, Farmer M E, Knauper B, Kessler R C, Norquist G S 1998 Limitations of diagnostic criteria and assessment instruments for mental disorders. *Archives of General Psychiatry* **55**: 109–15
Rice D P, Miller L S 1998 Health economics and cost implications of anxiety and other mental disorders in the United States. *British Journal of Psychiatry* **173**(34): 4–9
Robins L 1992 The future of psychiatric epidemiology. *International Journal of Methods in Psychiatric Research* **2**: 1–3
Robins L, Regier D A 1991 *Psychiatric Disorders in America*. The Free Press, New York
The Cross National Collaborative Group 1992 The changing rate of major depression: Cross national comparisons. *The Journal of the American Medical Association* **268**(21): 3096–105
Üstün T B, Sartorius N 1995 Editorial: Measuring functioning and disability; a common framework. *International Journal of Methods on Psychiatric Research* **7**(2): 79–83
Üstün T B, Sartorius N 1995 *Mental Illness in General Health Care: An International Study*. John Wiley & Son, Chichester, NY
Wittchen H-U 1996 Critical issues in the evaluation of comorbidity. *British Journal of Psychiatry* **168**(Suppl. 30): 9–16
Wittchen H-U 2000 Met and unmet need for intervention in community cases with anxiety disorders. In: Andrews G, Henderson S (eds.) *Unmet Need in Psychiatry. Problems, Resources, Responses*. Cambridge University Press, Cambridge, UK, p. 256
Wittchen H-U 2001 Generalisierte Angst und Depression in der primavarzthilieen Versorjung. *Fortschriffe der Medicin* **119** (Supplement 1)
Wittchen H-U, Nelson C B (eds.) 1998 Early developmental stages of substance abuse. *European Addiction Research* **4**(1–2)
Wittchen H-U, Höfler M, Merikangas K R 1999 Towards the identification of core psychopathological processes? Commentary. *Archives of General Psychiatry* **56**(10): 929–31
Wittchen H-U, Üstün B, Kessler R C 1999 Diagnosing mental disorders in the community. A difference that matters? *Psychological Medicine* **29**: 1021–27
Wittchen H-U, Miller N, Pfister H, Winter S, Schmidtkunz B 1999 Affelitive, somatoforme und angststorungen in Deutschland. *Das Gesundheitswesen* **61**(Supplement 2): 216–20
WHO (World Health Organization) 1990 *Composite International Diagnostic Interview* (CIDI): (a) CIDI-interview (version 1.0), (b) CIDI-user manual, (c) CIDI-training manual, (d) CIDI-computer programs. World Health Organization, Geneva
WHO (World Health Organization) 2000 Cross-national comparisons of the prevalences and correlates of mental disorders. WHO International Consortium in Psychiatric Epidemiology. *Bulletin of the World Health Organization* **78**(4): 413–26

H.-U. Wittchen

Mental Illness, Etiology of

Mental illness, like most types of illness, is multifactorial in origin. It can be influenced by constitutional and genetic factors, by both early and recent environmental factors and by cultural and social elements. These are best described in chronological format, with those more distant described first and the most recent described last.

1. *Early Causes*

1.1 *Genetic Influences*

The advances made by genetics in the last 25 years of the twentieth century have been enormous, and are discussed elsewhere (see, e.g., *Genetic Studies of Behavior: Methodology*; *Genetics of Complex Traits Through the Life Cycle*; *Mental Illness, Genetics of*). It may appear that few of these advances have yet had much impact on the causation of mental illness, because there have been few dramatic advances comparable with that of the identification of the gene for Huntington's chorea in 1987 (Gilliam et al. 1987). However, there has been a steady growth in knowledge, best summarized as incremental mapping, which has helped to place the genetic contribution to a wide range of mental illness, in context.

The standard methods of studying the genetic component of causation of illness are the same in psychiatry as in other branches of medicine. The relative genetic risk of having, or developing, a particular mental disorder is determined by three research approaches; family, twin and adoption studies. Family studies identify people with the disorder through questions (probands); the rates of disorder are identified in other family members and relatives, and then compared with the rates in the general population. From these figures it is possible to compute the numbers of people likely to develop the condition at some time in the future (the expectancy rate or morbid risk). A higher rate in immediate relatives compared with others, and greater rates than the general population, indicate a genetic contribution to the disorder.

However, environmental factors may also contribute to risk in family studies, and twin studies are better at distinguishing the genetic component. Monozygotic and dizygotic twin probands are compared with their co-twins for concordance rates (i.e., the rate of co-occurrence of the disorder in the other twin). If the ratio of the concordance rates for monozygotic (MZ) twins (who are similar to clones as they have identical genes) and dizygotic (DZ) twins is significantly greater than one, the disorder has a genetic component.

Because there is often a bias in reporting disorder, it is preferable to derive probands and their co-twins from twin registers wherever possible (Puri and Tyrer 1998).

Even with twin studies, the effect of environment cannot be completely separated since there is a tendency, fortunately exhibited less now than formerly, for monozygotic twins to be treated as though they were the same person and to identify psychologically with each other to a greater extent than dizygotic twins. Adoption studies overcome this effect, and when they are carried out with monozygotic twins, are seen to their best advantage. Other studies include adoptee approaches in which one or both of the biological parents has the disorder but the adoptive parents have not, so the rates of disorder in the adoptees can be compared with a control population, and adoptee family studies in which the adoptee is the proband and the rate of disorder determined in the biological and adoptive parents.

The results of these studies in general, show that there is a strong genetic contribution to bipolar mood disorder and schizophrenia with MZ/DZ ratios of greater than 3 (Hauge et al. 1968, Gottesman and Shields 1976) but this falls far short of that shown by a disorder such as Huntington's chorea. A more common finding is that illustrated by personality disorder, in which twin studies suggest that most personality traits are equally contributed to by genetic and environmental factors (Livesley et al. 1998) and several models have been postulated for the interaction between these (Kendler and Eaves 1986). Now that the human genome is close to being identified, the techniques of molecular genetics are being used increasingly to identify specific genes associated with psychiatric disorder, but to date, only one clear disorder has been identified, fragile X syndrome, which affects males mainly and which has clear locus on the X chromosome (Sutherland 1979). However, a range of other disorders has been shown to be associated with chromosomal abnormalities, and approximately 40 percent of the causes of severe mental handicap (Down's syndrome, phenylketonuria, tuberose sclerosis, Hurler's syndrome, Lesch-Nyhan's syndrome and Tay Sach's disease) are known to be caused by identifiable chromosome abnormalities that are potentially preventable by genetic counselling (Weatherall 1991; see *Mental Illness, Genetics of*).

1.2 Environmental Factors Before and During Birth

The time immediately before, during and shortly after birth is a very vulnerable one for the fetus or new-born infant. It is at these times that anoxia can lead to brain damage. The most obvious manifestation of this is loss of brain cells leading to learning disability in adult life, and better obstetric care and supervision of babies with low birth weight reduces this risk (Illsley and Mitchell 1984). There have also been a number of studies suggesting that the development of other mental illness, particularly schizophrenia and related disorders, is linked to obstetric difficulties (Geddes and Lawrie 1995), but there is some doubt about the methodology of some of these studies. More careful matching of cases and controls in a larger sample, however, has failed to replicate these findings, apart from prolonged labor and emergency Caesarian section being more common in schizophrenic probands (Kendell et al. 2000).

There is also evidence that infectious diseases, particularly when contracted before birth, may be associated with mental disorder. Schizophrenia is more common in those born in the winter months (Hare et al. 1973) and the case has been made, not altogether satisfactorily, that this could be a consequence of maternal influenza in the winter months (Adams et al. 1993). Another infectious disease linked to mental illness is rubella, which is associated strongly with learning disability (due to microcephaly), if infection is contracted in the first trimester of pregnancy.

1.3 Early Years of Development

The hypothesis that mental illness is a consequence of problems in early development has been active for many years and is one of the fundamental components of psychoanalysis. This argues that psychological conflicts encountered in early development that remain unresolved are likely to surface later in life in distorted form that psychoanalysis can succeed in decoding. This is perhaps best encapsulated in the phrase 'give us a child to the age of five and let who will have him thereafter.' Another approach initiated by the work of John Bowlby is the possibility that the earliest stages of the infant's relationship with its mother can also lead to problems that arise again in adult life. This has led to the growth of attachment theory (Bowlby 1988) and has also stimulated psychoanalytical approaches to treatment.

These difficulties are often not associated with overt psychological trauma, or, if they are, it is usually within the bounds of everyday experience. More severe trauma is associated with a different range of pathology. There are now many conditions which have been shown to be associated with severe psychological trauma in the early years of life, including borderline personality disorder, post-traumatic stress disorder, recurrent depressive episodes and a range of problems of sexual dysfunction. However, research into the subject has come in for some criticism, as data are almost entirely retrospective and subject to significant bias. This has been accentuated by the high profile of the multiple personality disorder, and its relationship with past childhood abuse, principally sexual abuse. Often, such abuse is unknown to the sufferer at first,

but is evoked by assessment and treatment under hypnosis. There have been many claims that this form of assessment is suspect and creates what is commonly known as the 'false memory syndrome' (Brandon et al. 1998).

Depressive illness in adult life is also more likely to occur in those who have suffered the loss of a parent through separation, divorce or death (Harris et al. 1986). This is postulated to be due to increased vulnerability to adversity after adverse early experiences. The long gap between initial loss and the onset of depression has been studied by life events researchers, and the concept of 'brought forward time' introduced to compare it with the immediate response to a major life event.

2. *Late Causes*

2.1 *Later Environmental Causes*

There are a host of environmental causes of mental illness from late adolescence onwards that, unlike the previous ones described earlier, are determined to a much greater extent by the individual concerned. The group of conditions showing this characteristic most prominently, are the substance misuse disorders, as absence of indulgence can never lead to disorder. Although there are likely to be genetic factors that influence the extent that people become addicted to a substance (Kendler et al. 1995) by far the most important factor is repeated exposure and this is primarily an environmental issue. Alcohol is the most important of the substance misuse disorders, accounting for one third of all mental disorders and having harmful effects on around 10 percent of the population per year (Kessler et al. 1994), a staggering figure however it is interpreted.

Other important environmental causes are infections. These include viral, protozoal and bacterial organisms, and most recently, prion infections, such as the human form of bovine spongiform encephalopathy (BSE), and mental disorder is created when they affect the brain. The acquired immune deficiency syndrome (AIDS), neurosyphilis and viral encephalitis are the most important. Trauma to the brain is also an important cause of personality change and functional handicap, including dementia. Whenever the oxygen supply to the brain is significantly impaired, there is the danger of neuronal death and consequent organic mental illness, primarily dementia. Similar problems can be created by metabolic disturbances but these are more likely to be reversible. Thus a range of medical illnesses; hypertension, atherosclerosis, renal failure, hepatic failure, thyrotoxicosis, myxodema, sarcoidoisis, autoimmune diseases and Cushing's disease (or any medical illness treated with steroids) may all be associated with mental disorder, of which anxiety, depression and dementia are the most common.

A more common environmental factor influencing the onset and nature of many common mental disorders is social deprivation. Epidemiological studies have consistently shown that almost all mental illness is more common in those of lower than higher social classes ((Hollingshead and Redlich 1958, Kessler et al. 1994). The reasons for this are much more difficult to unravel, but include specific aspects of deprivation (poorer material circumstances—greater crime, traffic pollution, sub-standard housing, unemployment), less control over direction in life and a higher rate of adverse life events (Boulton 1998).

2.2 *Recent Stress as a Precipitant of Mental Disorder*

The notion that stress is the cause of much mental (and physical) disorder is a very old one. It has been researched heavily in recent years by those interested in the effect of life events on mental illness. Although some of the early causes of mental illness might be perceived as stressors that create mental illness, the general notion of stress as a recent precipitant is rather than a distant one. Early experiences may create a vulnerability to mental illness that is then activated by a stressful experience. These are often combined in the stress–diathesis model of mental illness, which places mental illness and stress on separate scales: low stress being associated with no mental illness; greater levels of stress creating illness in those who have a vulnerable diathesis, and very severe stress which may create symptoms of mental illness in almost anybody.

There are three specific mental illnesses which are defined as stress related: acute stress disorders, adjustment reactions and post-traumatic stress disorder. All of these present primarily with symptoms within the neurotic spectrum (e.g., anxiety, panic, depression, irritability, social withdrawal), and satisfy the common definition that they would not have occurred in the absence of a psychosocial stressor. Acute stress disorders can occur in almost anybody when under intolerable pressures (e.g., car accident) but resolve rapidly when the natural processes of restitution take place (usually within 24 hours). Adjustment disorders are longer-lasting, either because the stress is persistent (e.g., stress in a relationship), or because it is more intense. Post-traumatic stress disorder is associated with stress that is beyond the range of normal experience (e.g., earthquake, observing homicide) and which is associated with specific symptoms, including flash-backs of the traumatic event, nightmares, extreme avoidance of situations that evoke memories of the event, as well as the range of symptoms shown in adjustment disorders with generally greater intensity (particularly depressive symptoms).

However, although these conditions are defined by the nature of the stressful experiences preceding them, there are many others in which stress is a provoking

factor. These are specific codes for stressful events in both the *DSM* and *ICD* classifications (American Psychiatric Association 2000, World Health Organization 1992) and in a full diagnostic summary the nature of any associated stresses should be noted. In research work, stresses are often quantified into life events, whose intensity depends both on their nature and the extent of their (contextual) threat (Brown and Harris 1978) and there have been numerous studies that show that the onset of almost all mental illnesses is accompanied by a greater rate of life events than those without such illness. It is difficult to know to what extent, however, these could be regarded as causes of the illness in the true etiological sense, and it is more common to regard them as triggers or provoking factors. Thus, for example, it has been shown that exposure to high levels of critical expressed emotion in patients with schizophrenia is more likely to provoke relapse than if more normal expression of emotion is shown (Vaughn and Leff 1976) and interventions to reduce such emotion in families may have a beneficial effect (Pharoah et al. 2000). Even with this evidence, it is not suggested that critical expressed emotion is the cause of schizophrenia; it is the specific combination of a specific type of stress and its relationship to a particular illness that makes its effects shown. This 'Achilles heel' phenomenon, the activation of an illness by a particular combination of circumstances, is very common in mental illness, and gives support to the stress–diathesis model for many of the most common conditions encountered, particularly those in primary care.

The multifactorial nature of causation of mental illness is illustrated in the following case report, based on a real patient but with some important elements distorted to prevent identification.

> Mrs A came as a refugee from a country in Sub-Saharan Africa where three members of her family were killed in front of her by rebels involved in fighting a civil war. She had become withdrawn and depressed after this, but was persuaded to flee the country by relatives and came to another country (the UK), where at the time she was being seen clinically, her application to stay in the country was being considered. Her relatives were also concerned that at times she had become more seriously disturbed and believed that her mind had been taken over by alien beings that had entered her body after a native witch doctor in her country had cursed her.
>
> On assessment, she was shy but cooperative and keen to please. Her family history revealed that her mother had been unwell after one of her children was born and believed her body had been possessed by spirits that were preventing her from looking after her baby. A niece had also killed herself by walking into a swamp when unwell and allegedly possessed. During her childhood she did well at school but had times when she lost confidence in her own abilities and had always been more anxious than her brothers and sisters. In general, she had always had low self-esteem. Mental state examination revealed that she was persistently anxious with episodes of panic, particularly at night and when her husband was away from home (he worked variable hours). She also had vivid nightmares of her parents being killed in front of her by men wielding machetes, and was unable to have any sharp knives in her kitchen because they aroused such high levels of anxiety. She also had periods of quite deep depression when she felt that life was such a struggle that it would be easier to take her own life than continue to fight. Direct questioning about her alien experiences revealed she still believed that she had been cursed, as indeed had her whole family, and that twitching movements of her hands and arms indicated that she was still under external control. She was also very anxious that her application to remain in the country as a refugee might not be granted and kept asking the services to intervene on her behalf so she could stay.

A diagnosis of post-traumatic stress disorder, comorbid generalized anxiety disorder and recurrent depressive episode was made, and an unspecified diagnosis in the schizophrenia group, probably persistent delusional disorder, was also made.

In considering the causes of this complex problem, which is representative of many others who present to psychiatric services, the following etiological factors were considered:

(a) the family history of possession and being cursed, which could have a cultural explanation, but which might be better explained by a family history of schizophrenia, which Mrs A would be then more likely to have, because of its genetic component;

(b) the extreme stress occasioned by seeing members of her family killed in front of her (one of the characteristic features of post-traumatic stress disorder);

(c) her lifelong tendency of low self-esteem and tendency to be anxious; and

(d) her current anxiety about the real possibility of being asked to leave the country if she did not achieve refugee status.

The etiology of these problems can be described in the form of a tree, in which the main trunk indicates her fundamental diatheses, her possible tendency to develop schizophrenia and the likelihood that she has some of the features of personality disturbance in the anxious/fearful group (cluster C) (a mixture of genetic and constitutional factors). The branches include the relatively recent experiences of seeing her family killed, and the most recent, more characteristic of an adjustment disorder, indicate the stress of waiting to hear whether she would be allowed to stay in the country.

In constructing this tree, we need to be aware that there is a natural tendency for people to attempt to find causal explanations for everything in life, and sometimes these may be quite wrong. However, most of them have face validity, because in the search to make sense of a chaotic world, success in finding explanations make it more bearable and also easier to recall—a phenomenon nicely described as 'effort after meaning' by one of the earlier researchers on the subject (Bartlett 1932). Sometimes it is more accurate to say that the cause of a particular mental illness is

unknown (idiopathic) but it makes us more authoritative to pick on a trunk, branch or leaf of the etiology tree and give a suitable explanation.

See also: Differential Diagnosis in Psychiatry; Post-traumatic Stress Disorder

Bibliography

Adams W, Kendell R E, Hare E H, Munk-Jorgensen P 1993 Epidemiological evidence that maternal influenza contributes to the aetiology of schizophrenia: An analysis of Scottish, English and Danish data. *British Journal of Psychiatry* **163**: 522–34

American Psychiatric Association (APA) 2000 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Text revision DSM-IV-TR. American Psychiatric Association, Washington, DC

Bartlett F C 1932 *Remembering*. Cambridge University Press, Cambridge, UK

Birtchnell J 1972 The inter-relationship between social class, early parental death and mental illness. *Psychological Medicine* **2**: 166–75

Boulton M 1998 Sociology. In: Puri B K, Tyres P J (eds.) *Sciences Basic to Psychiatry*, 2nd edn. Churchill Livingstone, Edinburgh, UK, pp. 327–35

Bowlby J 1988 *A Secure Base: Parent–child Attachment and Healthy Human Development*. Basic Books, New York

Brandon S, Boakes J, Glaser D, Green R 1998 Recovered memories of childhood sexual abuse: Implications for clinical practice. *British Journal of Psychiatry* **171**: 296–307

Brown G W, Harris T O 1978 *Social Origins of Depression: A Study of Psychiatric Disorders in Women*. Tavistock, London

Geddes J R, Lawrie S M 1995 Obstetric complications and schizophrenia: A meta-analysis. *British Journal of Psychiatry* **167**: 786–93

Gilliam T C, Bucan M, MacDonald M E, Zimmer M, Haines J L, Cheng S V, Pohl T M, Meyers R H, Whaley W L, Allitto B A 1987 A DNA segment encoding two genes very tightly linked to Huntington's disease. *Science* **238**: 950–2

Gottesman I I, Shields J 1976 Genetics of schizophrenia—Critical review of recent adoption, twin and family studies—behavioral genetics perspectives. *Schizophrenia Bulletin* **2**: 360–401

Harris T, Brown C W, Bifulco A 1986 Loss of parent in childhood and adult psychiatric disorder: The role of lack of adequate parental care. *Psychological Medicine*. **16**: 641–59

Hauge M, Harvard B, Fischer M, Gotlieb–Jewen K, Juel–Nielsen N, Raebild I, Shapiro R, Videbeck T 1968 The Danish twin register. *Acta Geneticae Medicae et Gemellologiae* **17**: 315–32

Hare E H, Price J S, Slater E T O 1973 Mental disorder and season of birth. *Nature* **241**: 480

Hollingshead A B, Redlich F C 1958 *Social Class and Mental Illness: A Community Study*. Wiley, New York

Illsley R, Mitchell R J 1984 *Low Birth Weight: A Medical, Psychological, and Social Study*. Wiley, Chichester, UK

Kendell R E, McInnenny K, Juszczak E, Bain M 2000 Obstetric complications and schizophrenia: Two case-control studies based on structured obstetric records. *British Journal of Psychiatry* **176**: 516–22

Kendler K S, Eaves L J 1986 Models for the joint effects of genotype and environment on liability to psychiatric illness. *American Journal of Psychiatry* **143**: 279–89

Kendler K S, Walters E E, Neale M C, Kessler R C, Heath A C, Eaves L J 1995 The structure of the genetic and environmental risk factors for six major psychiatric disorders in women. Phobia, generalized anxiety disorder, panic disorder, bulimia, major depression, and alcoholism. *Archives of General Psychiatry* **52**: 374–83

Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshleman S, Wittchen H U, Kendler K S 1994 Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**: 8–19

Livesley W J, Jang K L, Vernon P A 1998 Phenotypic and genetic structure of traits delineating personality disorder. *Archives of General Psychiatry* **55**: 941–48

Pharoah F M, Mari J J, Streiner D 2000 Family intervention for schizophrenia (Cochrane Review). *Cochrane Database of Systematic Reviews*, 2. Update Software. Cochrane Library Issue 2, Oxford, UK

Puri B, Tyrer P 1998 *Sciences Basic to Psychiatry*, 2nd edn. Churchill Livingstone, Edinburgh, UK

Sutherland G R 1979 Heritable fragile sites on human chromosomes II. Distribution, phenotypic effects and cytogenetics. *Human genetics* **53**: 136–48

Vaughn C, Leff J 1976 The measurement of expressed emotion in the families of psychiatric patients. *British Journal of Social and Clinical Psychology* **15**: 157–65

Weatherall D J 1991 *The New Genetics and Clinical Practice*. Oxford University Press, Oxford, UK

World Health Organisation (WHO) 1992 *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. World Health Organisation, Geneva, Switzerland

P. Tyrer

Mental Illness: Family and Patient Organizations

1. Mental Health Policies and the Family

In November 1999 Dr. Gro Harlem Brundtland, director general of the World Health Organisation (WHO) launched the WHO's new global strategies for mental health. These strategies aim to ease the 'burden' of mental disorders and neurological diseases currently affecting about 400 million people, by improving the quality of care throughout the world, particularly in the developing countries.

WHO aims to achieve these objectives through a number of measures in both developing and developed countries. The organization plans to raise awareness of the relative importance of mental disorders as a major contributor to the global burden of disease in different groups, including health professionals and public health decision makers, and also among the general public. WHO will fight the social stigma, misconceptions, and discrimination associated with mental illness, as well as promoting the human rights of people with a mental illness: 'Very often and in

many countries, individuals who are affected by neuropsychiatric disorders endure double suffering, namely from the conditions themselves and from the social stigma and discrimination attached to them.' (WHO/67 Press Release, Nov. 12, 1999—Raising Awareness, Fighting Stigma, Improving Care.) In December 2001, the 10th anniversary of the Principle for the Protection of Persons with Mental Illness and the Improvement of Mental Health Care adopted by the UN General Assembly in 1991, Dr. Brundtland proposed launching measures to foster the implementation of these principles with, for example, an International Convention on the Rights of Persons with Mental Disorders.

Mental health promotion should go beyond simply achieving the absence of a mental disorder, but aim to improve mental well-being, a state in which individuals can realize their abilities, cope with the stresses of life, work productively, and make a positive contribution. WHO has identified poverty as a great obstacle to such mental well-being. In both mental health promotion and in treatment and care Dr. Brundtland saw the family, often extended by the closest community network, as having a central role to play: 'Much of the burden of caring for the mentally ill, or for prevention for those who are in danger of becoming ill, is left to the family …' (Office of Director General, WHO, Oct. 13, 1999.) The family needed more support and better information, in recognition of their work. She commended the 1995 'Barcelona Manifesto,' put forward by the European Union Federation of Family Associations of Mentally Ill People, which outlines the needs and perceived rights of the family of a person with mental illness, and would consider its principles in WHO's work.

At a regional level, within the European Union, the Council of Ministers of Health has also focused on mental health, in particular the promotion of mental health within the context of improving social inclusion. It has invited the European Commission to analyze the impact of Community activities on mental health, for example, in the fields of education, youth policy, social affairs, and employment, and to consider the need to draw up—after consulting with member states—a proposal for Council recommendations on the promotion of mental health in the European Union.

2. *User and Family Organizations Related to Mental Illness*

There are organizations focusing on users and families' concerns operating at a number of levels: global, regional, national, and within individual countries. Some organizations are run by users for users, while others involve mental health professionals. At a global level, the World Association for Psychiatric Rehabilitation and the World Federation for Mental Health each has an interest in the perspective of users and carers.

An example of a user-run organization at a regional level is the European Network of (ex-)Users and Survivors of Psychiatry, which describes itself as 'a regional initiative to give (ex-)users and survivors of psychiatric services a means to communicate, to exchange opinions, views and experiences in order to support each other in the personal, political and social struggle against expulsion, injustice and stigma in our respective countries.'

The history of the Network goes back to 1990 when an initiative was taken in the Netherlands to form a network of associations of (former) psychiatric patients from various European countries. Since then the Network has organized four European conferences. At the last conference in Luxembourg in 1999 more than 90 delegates, all of them (ex-)users/survivors from 26 European countries, met and created an action plan for the coming years.

The aims and objectives of the European Network are:

(a) The European Network is against any unilateral approach to, and stigmatization of, mental and emotional distress, madness, human suffering, and unconventional behavior.

(b) The European Network should support (ex-) users/survivors' autonomy and responsibility in making their own decisions (self-determination).

Priorities for the network include the following:

(a) Act against any kind of discrimination in society (both inside and outside the mental health care system) of people who have been subject to the psychiatric system;

(b) Support the development of (ex-)user/survivor groups throughout Europe (with a particular emphasis on those countries where there are no existing organizations);

(c) Create and support new alternatives to the psychiatric system and collect and share information on the existing ones;

(d) Influence and try to change present treatment in psychiatry.

The European Network attempts to influence policy at a European level and maintains contacts with other international organizations active in the mental health field. The Network collaborates with the WHO Regional Office for Europe, the European Union, the European Disability Forum, the International Labour Organisation (ILO), Mental Health Europe/Santé Mental Europe (the former European Regional Council of the World Federation for Mental Health) and the Geneva Initiative on Psychiatry.

The Network is a federal structure of national and local associations of (ex-)users and survivors and of mixed organizations with a significant (ex-)user/survivor membership. For countries where there are no such associations, exceptionally individual (ex-)user/survivor members may become members. The aim is

for the Network to be a grassroots, democratic and fully (ex-)user/survivor controlled organization. Through the membership of its member organizations the network represents several ten of thousands of (ex-)users/survivors from across Europe.

In the USA during the last 20 years the National Alliance for the Mentally Ill (NAMI) has grown, with a membership now of over 200,000 family members and service users. It operates at a national, state, and local level to 'provide the nation's voice on mental illness.' It represents family members and service users who seek more equitable services for people with a severe mental illness.

In Europe, voluntary organizations from a number of countries, which represent the relatives of people with a mental illness came together in DeHaan, Belgium in 1990. A manifesto signed by family associations from 10 European countries led to the establishment of the European Federation of Associations of Families of Mentally Ill People (EUFAMI) in 1992. At their third congress held in Sweden in 1999, 22 member associations from 15 countries—directly representing 55,000 families—set out their aim to work toward the reduction of stigma and discrimination against people with a mental illness and to seek adoption by health professionals of the highest standards of good practice.

3. *Function of User and Family Organizations Related to Mental Illness*

There are some descriptions of individual groups and programs available. One group focuses on issues of loss and how to overcome it (Baxter and Diehl 1998). The BRIDGES program and the Journey of Hope are peer-taught programs that offer education and support to people with a severe mental illness and their families. The aim is to validate the participants' sense of loss as normal, and provide a structure for a new sense of self, so that individuals and their families can move from isolation and loneliness to empowerment and reconnection with ordinary life.

Families of people with dementia may have particular needs given the progressive nature of the illness and its increasing impact on carers as a person's mental and also physical health declines. Support groups for carers allow members to trade information about the disease and daily care requirements (Wormstall et al. 1996).

4. *The Relationship Between Family Organizations and Statutory Services*

Members of family organizations may be dissatisfied with the care provided by statutory services. Members of the Dutch organization Ypsilon, which supports the relatives of people with a long-term psychotic illness, took part in a questionnaire study. They reported a lack of access to treatment professionals, a lack of information about their relative's illness, and a lack of family involvement in treatment planning (Schene and van Wijngaarden 1995).

In some family organizations, health professionals may be more involved. In Massachusetts, Alliance for the Mentally Ill (AMI) provide eight family support groups, with a total average monthly attendance of 86 relatives and 25 professionals (Bouricius et al. 1994). Family members lead the groups, but they encourage health professionals to attend. Results of a questionnaire given to the relatives showed that a large majority found the attendance of professionals at the meetings to be helpful in accessing needed services. The researchers also reported that the support groups tend to improve services for people with a mental illness in the area.

Across the USA there is a wide variation in the contact between doctors training in psychiatry and psychoeducational programs involving families of people with a mental illness. In less than half of the psychiatry training schemes is NAMI formally involved. Barbee et al. (1991) give some examples of successful collaboration. In one program, an exercise in which trainee psychiatrists and family members change roles is recommended for its ability to improve communication between trainees and families.

Such programs may have some effect in reducing the delay between a person being diagnosed as having a mental health problem, and family members getting in touch with a support organization. A survey in Quebec of AMI members found that half of the members experienced a delay of over two years and only 10 percent were referred to the organization by a psychiatrist (Looper et al. 1998). The majority of respondents would have preferred earlier contact, and some approved of the more proactive methods developed by AMI-Quebec such as a telephone call after a relative's first admission to hospital.

5. *Benefits of Involvement with a Family Organization*

There is some evidence of family members gaining benefit from joining a support group. For example, in a study of 225 families with a relative with a mental illness, there were significant differences between those who participated in a support group and those who did not (Mannion et al. 1996). The support group participants were more likely to be a parent with a higher functioning relative who had been ill for a longer time. They reported less subjective burden, and a greater use of adaptive coping mechanisms than did members of families who did not participate in a support group.

It is possible to demonstrate that there are some ingredients of a group that are particularly helpful. A US study compared two types of group, an interactive psychoeducational model and a support model offering nonstructured discussions (Kane et al. 1990). Each was held over four sessions for the families of people with a relative with schizophrenia or schizoaffective disorder. In the psychoeducational group, over three-quarters of the participants rated the quality of the information received as excellent, and 94 percent said they had received the kind of information they wanted. This compares with a quarter of the support group rating the information as excellent and none feeling that they had received the help they wanted. There was a lower level of depression in members of the psychoeducational group.

A study in Germany reported benefits for family members who joined a self-help group (Schulze-Mönking 1994). In this case the families of men with a severe mental illness were most likely to join the group. Over the two years of the study, there was a tendency for the relatives of the families which participated to have a better outcome. Family members in the group developed more social contacts and reported fewer physical complaints.

As well as providing support and information about a relative's illness, these organizations can fulfill a number of other functions. For example, in the Circle Model, family members caring for a relative who is cognitively impaired are enabled to get some respite (Jansson et al. 1998). Equal numbers of family members and volunteers train together in a study circle, and once trained, the volunteers can replace the carer on a regular basis in the person's home. The relatives gained support from meeting others in the same position as themselves, and they also expressed feelings of security and relaxation in relation to their respite from the home situation. The volunteer caregivers also expressed satisfaction and appreciation of the knowledge they gained from the caregivers.

In many caregiving situations an elderly parent is looking after a grown-up son or daughter at home or providing substantial amounts of support out of the home. One of the biggest worries for the carer may be what will happen after his or her death. The Planned Lifetime Assistance Network (PLAN) is available in some parts of the USA through NAMI. It aims to provide lifetime assistance to individuals with a mental health problem whose parents or other family carers are no longer alive, or are no longer able to provide care (Lefley and Hatfield 1999).

Another area of growth outside the statutory sector is in self-help groups in the voluntary sector. Their focus is diverse and represents an interest in promoting self-help and support for people with a range of mental health problems including anxiety disorders, mood disorders, schizophrenia, and dementia. These groups are created and operate at a local, regional, and national level. Some have developed groupings at an international level whose main focus is to act as a lobby group and advocate for intergovernmental initiatives in policies around mental health.

6. *Self-help Groups*

In the USA, Borkman (1997) has linked the development of contemporary self-help groups with the founding of Alcoholics Anonymous (AA) in 1935. AA is a model for other self-help groups, which are also nonhierarchical direct democracies that avoid advocacy but focus on providing support. The major growth of all these groups took place during and after the 1970s alongside the development of the civil rights and women's movements, both of which challenged bureaucracies and traditional authority.

To understand more about the people involved in self-help groups, Chamberlin et al. (1996), using a participatory action research paradigm, with an advisory committee of individuals who had used such a program, carried out a survey of participants in six representative programs. Respondents spent an average 15 hours per week in their program and had been attending for almost five years. Members both received and gave help. Overall they reported that attending a self-help program had a salutory effect on their quality of life

Other organizations controlled by consumers or survivors may have different aims, carrying out different activities. Trainor et al. (1997) define consumer/survivor organizations as being operated for and controlled and staffed by people who have used the mental health system. In Ontario, Canada, the Consumer/Survivor Development Initiative funds 36 such organizations. They carry out a range of activities, which include offering mutual support, cultural activities, advocacy skills training, and education for the public and professionals. Involvement in a self-help group leads to a drop in contact with services, with less time spent in in-patient care, and less contact with crisis services, suggesting that contact with such groups may help individuals handle difficulties in a different way.

In the UK information from members of different user groups suggested that these groups could have a role in ensuring that individual users' rights were respected (Barnes and Shardlow 1997). They also enabled people with a mental health problem to improve the accountability of services, and supported their wider participation as citizens.

See also: Health Interventions: Community-based; Mental Health: Community Interventions; Mentally Ill: Public Attitudes; Public Psychiatry and Public Mental Health Systems; Social Support and Health; Support and Self-help Groups and Health

Bibliography

Barbee J G, Kasten A M, Rosenson M K 1991 Towards a new alliance: Psychiatric residents and family support groups. *Academic Psychiatry* **15**: 40–9

Barnes M, Shardlow P 1997 From passive recipient to active citizen: Participation in mental health user groups. *Journal of Mental Health UK* **6**: 289–300

Baxter E A, Diehl S 1998 Emotional stages: Consumers and family members recovering from the trauma of mental illness. *Psychiatric Rehabilitation Journal* **21**: 349–55

Borkman T 1997 A selective look at self-help groups in the United States. *Health and Social Care in the Community* **5**: 357–64

Bouricius J K, Kersten E, Nagy M, McCartney P L, Stein R 1994 Family support groups. AMI of western mass style. *Innovations and Research* **3**: 33–40

Chamberlin J, Rogers E S, Ellison M L 1996 Self-help programmes: A description of their characteristics and their members. *Psychiatric Rehabilitation Journal* **19**(3): 33–42

Jansson W, Almberg B, Grafström M, Winblad B 1998 The circle model: Support for relatives of people with dementia. *International Journal of Geriatric Psychiatry* **13**: 674–81

Kane C F, DiMartino E, Jimenez M 1990 A comparison of short-term psychoeducational and support groups for relatives coping with chronic schizophrenia. *Archives of Psychiatric Nursing* **4**: 343–53

Lefley H P, Hatfield A B 1999 Helping parental caregivers and mental health consumers cope with parental aging and loss. *Psychiatric Services* **50**: 369–75

Looper K, Fielding A, Latimer E, Amir E 1998 Improving access to family support organizations: A member survey of the AMI-Quebec alliance for the mentally ill. *Psychiatric Services* **49**: 1491–2

Mannion E, Meisel M, Solomon P, Draine J 1996 A comparative analysis of families with mentally ill adult relatives: Support group members versus non-members. *Psychiatric Rehabilitation Journal* **20**(1): 43–50

Schene A H, van Wijngaarden B 1995 A survey of an organization for families of patients with serious mental illness in the Netherlands. *Psychiatric Services* **46**: 807–13

Schulze Mönking H 1994 Self-help groups for families of schizophrenic patients: Formation, development and therapeutic impact. *Social Psychiatry and Psychiatric Epidemiology* **29**: 149–54

Trainor J, Shepherd M, Boydell K M, Leff A, Crawford E 1997 Beyond the service paradigm: The impact and implications of consumer/survivor initiatives. *Psychiatric Rehabilitation Journal* **21**: 132–40

Wormstall H, Gunthner A, Morawetz C, Schmidt W 1996 Groups for family caregivers of patients with Alzheimer's disease in Germany. *Nervenarzt* **67**: 751–6

R. Ramsay and J. H. Henderson

Mental Illness, Genetics of

Psychotic and affective disturbances were first described in the philosophic literature of ancient Greece under the concepts of mania and melancholia, which were considered as extremes of character profiles. In this epoch the familial inheritance of these disturbances was already noticed. This wisdom continued till the beginning of the nineteenth century when these disturbances were first considered as medical diseases. During this century the new concept of degeneration stressed the accumulating effects of familial inheritance across generations and was applied to psychotic and emotional disorders. At the end of the nineteenth century alcoholism was additionally recognized as a major source of degeneration. Although the classical concepts of inheritance and degeneration did not specify the mechanisms of transgenerational transmission, both concepts prepared the basis for the eugenic movement starting in Anglo-Saxon countries. This ideology proposed methods of reproductive planning for increasing the health and welfare of humankind and for reducing the impact of diseases, criminality, and bad habits.

1. Twentieth Century Developments

Mendel was the first to postulate specific mechanisms of genetic transmission in plants. He was lucky in that he was working on traits which were due to a monogenic variation (one mutation, one changing trait). The relevance of this work was not immediately recognized. The redetection of the Mendelian law and the invention of the concept of genes in 1900 induced rapidly growing interest in the genetic hypothesis and stimulated its application to humans. Remarkably, the original concept of genes coding for physical and possibly also for psychological properties was hypothetical and was developed without any knowledge of the physical existence or structure of genes (which became apparent only several decades later through the work of Watson and Crick). The early concept of genes already included the possibility that a gene could have multiple variants (mutations) with differential functional consequences (polymorphism). Thus, the distinction between genetic and nongenetic influences became feasible. The familial transmission of patterns and disturbances of human behavior was also explored in a genetic perspective. The advantage of this development was the development of systematic family and twin studies and their application to mental diseases, mainly in the 'Munich School' headed by E. Rüdin at the beginning of the twentieth century. Familial aggregation and genetic determination of most mental disorders were concluded.

These results were misinterpreted in the predominant one disease/one gene perspective (monogenetic concept) which was valid mainly for rare genetic diseases but not so for common mental diseases. The eugenic movement in many countries was influenced by this misinterpretation, and stimulated programs to prohibit people affected from reproducing. Nazi Germany in particular established forced sterilization programs among patients with schizophrenia and affective disorders. This development is particularly surprising as it was noticed that many inherited

behavioral traits did not stick to a Mendelian mode of familial transmission, and as the application of Mendelian genetic transmission to all inherited traits was called into question by the biometric school of geneticists, particularly in the UK. A polygenetic etiology to explain familial aggregation without Mendelian patterns of transmission was early proposed in a paper by the British statistician Sir Ronald Fisher in 1918. It is difficult to understand why leading psychiatric geneticists (e.g., E. Rüdin in Germany) appeared to be ignorant of these insights and continued to claim the monogenic Mendelian origin of schizophrenia and manic-depressive disease. In a polygenetic perspective the rationale of the eugenic movement becomes fully invalid as those programs cannot change the frequency of most targeted mental disorders which are not transmitted in a Mendelian manner.

Despite excellent empirical work (e.g., Kallmann) the hypothesis of the genetic origin of mental disorders did not receive a lot of attention between the 1940s and 1970s. Several reasons can be cited for the lack of interest: first, the genetic hypothesis was seriously discredited by the morally disreputable eugenic movement focusing on these diseases; second, psychoanalytic thinking predominated in these decades in many countries (especially in the USA until about 1970); and third, the research tools for studying genetically influenced disorders without a Mendelian mode of transmission were very limited before the molecular-genetics era; finally, tools for understanding complex interactions of genetic and environmental influences were underdeveloped in these times. Given this lack of genetic research potential the environmental causes of mental diseases were overstressed.

Motivated by the emerging biological psychiatry during the 1970s again family-genetic approaches in searching for causes of mental diseases received growing attention. Advanced methods of family, twin, and adoption studies using the recent developments of epidemiological and biostatistical techniques were performed exploring the etiology of, and the relationship between, mental disorders. Particularly the landmark Danish adoption studies by Kety and Goodwin changed the predominating assumption of the psychosocial origin of psychoses and alcoholism and motivated a nature–nurture discussion which dominated the field for a decade. A refined arsenal of genetic-epidemiological methods was used to demonstrate the interactions of environmental and genetic forces as the etiological basis of common mental diseases. Gene–environmental interactions turned out to be the rule but not an exception for the common mental diseases. The results strongly argued against the oversimplification of the nature vs. nurture debate.

Based on the detection of the biochemical structure of chromosomes as DNA strings, and the process of expression of genes in the functionally relevant proteins, the polymorphic nature of genes and the functional consequences became apparent in the middle of the twentieth century. The genetic variability might occur as different alleles of a gene (allelic variants) which might also result in different gene products (proteins), or it might occur as different variants in promoter regions of genes which influence the degree of gene expression resulting in a variation of the quantity of the gene product. Thus, genetic causes of diseases could be identified on a DNA level as well as on a protein level. Genetics thus became a basic tool for unraveling the pathophysiology of diseases and for the understanding and development of effective treatments.

The rapidly developing field of molecular genetics shifted the focus from the nature–nurture debate to the search for the genetic origin of diseases in the variation on the DNA level (from about 1990). Stimulated by the progress in unraveling the causes of monogenic diseases, extensive efforts are now undertaken to identify genetic variants of etiological relevance. The resulting neglect of environmental influences in psychiatric research is compensated by the hope that once impacting genes are identified specific environmental contributions to the etiology will more easily be detected.

In the meantime the genetic causes of most monogenic diseases have been detected. As expected from this success story, the first successes in detection of gene mutations for mental diseases came for rare genetic diseases characterized by mental retardation and early onset dementia. Causal mutations of specific genes have now been identified for rare variants of mental retardation such as fragile X syndrome and Rett's syndrome (Amir et al. 1999). Causal genes for rare variants of dementia of Alzheimer type characterized by early onset were also identified. Genes accounting for more common diseases have less deleterious effects. Thus, in spite of intensive research work, only very few genes influencing the common mental diseases have been identified. The major difficulty is the polygenetic origin of these disorders. The progress of the Human Genome Project will accelerate progress in finding disease genes in the future.

Among the common mental diseases schizophrenia, affective disorders, and addiction, and also anxiety disorders received most intensive study. The genetics of rare mental diseases, particularly early-onset Alzheimer's disease and specific subtypes of mental retardation, have been extensively and very successfully studied; given the limitations of space these rare disorders are not included in this overview.

2. *Specific Disorders*

2.1 *Schizophrenia*

Among mental disorders schizophrenia is most intensively studied in a genetic perspective. Schizo-

phrenia is a disabling lifelong disorder with the first signs in early childhood, with an often insidious onset in early childhood (lifetime prevalence 1 percent). The symptoms are heterogeneous and vary across the lifetime. An extensive body of evidence proposes that the symptoms of schizophrenia emerge from a maladaptive brain development. The familial-genetic basis had already been established at the beginning of the twentieth century. Recurrent risks among siblings is about 5–7 percent, resulting in a relative risk of about 10 (lifetime risk in the group of interest divided by lifetime risk in the general population). The risk among parents of schizophrenics is reduced (3–5 percent) because of a reduced fertility of schizophrenics (i.e., affected subjects have fewer children than random subjects in the general population). Currently, having a first-degree relative with schizophrenia shows the highest predictive power among all known risk factors for schizophrenia (see *Schizophrenia*).

2.1.1 Genetic vs. nongenetic sources. Twin studies explore the genetic impact by systematically varying the genetic similarity between twins (i.e., by comparing mono- and dizygotic twins). A higher concordance rate for schizophrenia was consistently found among monozygotic compared to dizygotic twins (50 percent to 10 percent). On the one hand, this difference proves the operation of genes. On the other hand, the concordance rate of monozygotic twins is far from 100 percent, arguing for the impact of nongenetic environmental forces. The application of variance-analytic methods to twin data combined with prevalence rates in the population make it possible to distinguish three sources of etiological variance under the assumption of specific modes of familial transmission (genetic, nongenetic familial environment, and nongenetic individual-specific environment). Model-dependent variance analyses propose about 50 percent of the etiological variance to be due to genetic factors, whereas the remainder is mainly allocated to individual-specific environment. Adoption studies systematically vary environment by teasing apart the biological background (i.e., being a biological child of affected parents) and foster environment (i.e., being adopted away into another familial environment). Following this strategy a strong genetic influence interacting with environmental forces was concluded. The familial environment cannot be excluded on the basis of adoption studies. Up to now there have been different conclusions from twin compared to adoption studies with regard to the relevance of familial environment for the emergence of schizophrenia; whereas this problem is currently unresolved, the relevance of strong genetic influences is unquestionable (Kety et al. 1994, Tienari et al. 1994).

A Mendelian pattern of familial transmission cannot be observed and the precise nature of the transmission mechanism remains obscure. Thus, schizophrenia is a complex disease like insulin-dependent diabetes, coronary heart disease or autoimmune disorders. Schizophrenia shares a series of features with other complex diseases:

(a) Environmental factors are operating in concert with genetic factors (evidenced by less than 100 percent concordance among monozygotic twins).

(b) Penetrance of the phenotype is incomplete: the offspring of an unaffected twin of a schizophrenic case (discordant monozygotic twins) have the same risk (10 percent) of transmitting the disorder to their children as monozygotic twins with both being affected.

(c) The boundaries of the transmitted familial phenotype of schizophrenia are not distinct. Also related syndromes (other psychoses) and isolated symptoms are aggregating. Among the relatives without any psychiatric disorder during lifetime neurobiological characteristics of schizophrenia occur more often than expected by chance (e.g., attention deficit, deviant patterns of evoked potentials, memory problems, slow pursuit eye movement disturbances) (Tsuang 2000).

2.1.2 Molecular approaches. The strong support for a genetic influence on schizophrenia encouraged the search for predisposing genetic variants. Despite several promising leads there is not yet a definite association of a genetic variant with the illness. Although many genes coding for proteins which are considered as relevant for the pathophysiology or as treatment targets are polymorphic, an association of these variants with the disease was not consistently found. However, meta-analyses demonstrated very mild effects of genetic variants of the genes for serotonin receptor 2a and for dopamine receptor D3 (relative risks about 1.2).

The genetic linkage strategy working on families with multiple affected individuals was enormously successful in monogenic diseases and was applied also to complex diseases. The strategy makes use of the transmission of genetic information via

(a) chromosomes where genes are placed in a fixed order, and

(b) crossing over between chromosomal pairs of maternal and paternal origin.

The cosegregation of the disease and the variation at a genetic locus is explored by this strategy. Cosegregating markers point to the location of the impacting gene in close spatial neighborhood. A major advantage of this strategy is that linkage can be explored genome-wide by applying a limited number of equally spaced markers (at least 400).

In the first stage the application of this strategy identified several candidate regions on the genome

hosting susceptibility genes which have yet to be identified. The first candidate region which was confirmed later on was reported by Straub et al. (1995). In the meantime several genome-wide linkage scans were completed, and multiple candidate regions were reported. Some of the candidate regions were confirmed in different linkage studies. Currently candidate regions on 1q, 5q, 6p, 6q, 8p, 10p, 13q, 18p and 22q are most well confirmed across the various linkage studies. Taken together, evidence for a major gene locus was not found in the vast majority of the studies. The identified candidate regions for schizophrenia and other complex diseases are broad, including hundreds of genes. It will take a long time to identify the susceptibility genes in a systematic manner. The progress in mapping genes and in characterizing their function (e.g., due to the Human Genome Project which is being conducted in North America and Europe) will accelerate this search for genes.

It can be concluded from the multiplicity of candidate regions that there is no single causal or major gene that explains most of the genetic variance but that multiple susceptibility genes influence the risk of schizophrenia. The demonstrated polygenetic basis of schizophrenia explains:

(a) the complex pattern of familial aggregation not fitting to a Mendelian mode of transmission, and

(b) that the prevalence of schizophrenia remains stable over time although this genetically influenced disease with an onset in adolescence and early childhood is associated with a significant reduction of fertility (Gershon 2000, Maier et al. 2000).

2.2 *Affective Disorders*

Affective disorders run in families and are genetically influenced. The familial pattern of aggregation of specific affective syndromes proposed a genetic-epidemiological split between bipolar disorder (manic episodes in combination with depressive episodes) and unipolar disorder (recurrent depressive episodes only). More than three decades ago Angst and Winokur observed that bipolar disorders were more common than in the general population, 1 percent among parents, children, and siblings of patients with bipolar disorder, $\sim$ 7 percent with the same disorder, but not so among relatives with unipolar depression (1–2 percent); in contrast, unipolar depression aggregates in families with both syndromes 20 percent compared to about 10 percent in the general population) (Winokur et al. 1995).

Twin and adoption studies strongly suggest a genetic influence which is stronger for the bipolar than for the unipolar variant. Mean monozygotic concordance rates are 30–70 percent for unipolar depression and 50–80 percent for bipolar disorder. Dizygotic concordance rates are 10–30 percent for unipolar depression and 10–20 percent for bipolar disorder. The calculated heritability rates are 30–50 percent for unipolar depression and 50–80 percent for bipolar disorder, leaving space for environmental risk factors (Sullivan et al. 2000).

Given the relatively high lifetime prevalence of unipolar depression in the general population, very informative longitudinal studies exploring the relationship between different risk factors from various domains are feasible. A population-based prospective twin study (Kendler et al. 1990) proposed that genetic and environmental risk factors (such as early parental loss, perceived parental warmth in childhood and critical live events and social support later on) interact. The risk factors occurring already in childhood also influence personality features (e.g., neuroticism) and coping strategies which also operate as risk factors for depression and mediate the early environmental and genetic sources for the final disease status.

Mendelian patterns of transmission were only observed in selected extended families; most families with more cases show a more complex pattern. Also the recently feasible systematic search for genes using genome-wide linkage approaches primarily focused on bipolar disorder. In 1994, the first candidate region for bipolar disorder obtained by molecular-genetic tools was found on chromosome 18p (Berrettini et al. 1994); this result was confirmed by other studies. Other confirmed candidate regions are on 4p, 13q, 18q, and 21q.

A lot of genetic association studies comparing allele frequencies between cases and controls were performed mainly for unipolar depression focusing on polymorphic genes which are believed to be involved in the pathophysiology or pharmacology. A promoter variant of the 5-HT transporter received particular attention. Classical linkage studies using monogenic phenotypes as markers—e.g., red–green color blindness, which is localized on the X chromosome—in bipolar disorder produced ambiguous results. As with schizophrenia, multiple regions on the genome were proposed by linkage studies to host genes contributing to the risk of bipolar disorder. Some of these regions were confirmed by independent groups. Specific susceptibility genes have not yet been identified (Craddock and Jones 1999) (see *Depression*).

2.3 *Anxiety Disorders*

Some behavioral disorders can phenomenologically be considered as extremes of behavioral variants with a broad variation in the general population. For example, anxiety is a complex behavioral reaction physiologically revealed in dangerous situations; another example is eating disorders (anorexia, bulimia) which might be considered as variants of dieting. The physiological reactions are expressed in an interindividually, quantitatively, and qualitatively variable manner under the same situational context. The interindividual variation of these behavioral traits is partly under genetic control, as evidenced by twin

studies. The degree of genetic impact is variable across traits with strong effects on anxiety proneness and smaller effects on dieting. Thus, it did not come as a surprise that the phenomenologically related disorders demonstrate familial similarity and genetic influences. However, the magnitude of genetic influence may vary along the behavioral continuum, and a qualitatively additional effect might operate on the extremes (i.e., on the disorders). An additional genetic effect was indeed observed for anorexia whereas a qualitatively additional effect on anxiety disorders seems to be less likely.

Anxiety disorders display a phenomenologically heterogeneous and variable symptomatology overlapping with nearly all other psychological disorders. Subtyping of anxiety disorders is widely accepted on the basis of distinct phenomenological features. The various clinical variants (generalized anxiety disorders, panic disorders, phobias) reveal specific familial aggregation in family studies, although substantial intrafamilial cosegregation between various specific anxiety disorders and also with depression (especially generalized anxiety disorder but panic disorder substantially less so) and addictive behavior (panic disorder, phobias) was observed. The absolute prevalence rates for specific disorders vary considerably between studies due to methods of case identification and sampling. The reported relative risks vary between 2 and 10. Twin studies including multiple anxiety/affective disorders demonstrated that:

(a) Generalized anxiety disorder, panic disorder and specific phobic disorders are under genetic influence (with heritability rates between 30 percent and 45 percent); obsessive-compulsive disorder seems to have the lowest level of genetic influence.

(b) The genetic contribution to each anxiety syndrome is neither highly specific nor highly unspecific (partly syndrome-specific and partly shared by other anxiety syndromes and by unipolar depression).

(c) Different anxiety disorders are genetically heterogeneous with at least two genetically distinct groups: panic disorder, phobias and bulimia defining a group of disorders with broad overlap of influencing genetic factors, and generalized anxiety and depression defining a separate genetically overlapping group (Kendler et al. 1995).

The phenotype transmitted in families is not only restricted to specific clinical syndromes. Increased anxiety proneness, behavioral disinhibition, increased sensitivity to hyperventilation or elevated autonomic reaction were also observed more commonly than expected by chance among healthy relatives (Merikangas et al. 1999).

The search for specific genes influencing anxiety disorders has been unsuccessful up to now. Some genome-wide linkage studies were performed for panic disorder without providing conclusive results on the localization of susceptibility genes. In any case, a major gene effect was not found. Thus, it is very likely that multiple genes influence the risk for panic disorder, each with an effect too small to be detected by linkage analysis. On the other side, associations with variants of candidate genes which are known to be involved in the pathophysiology of anxiety could also not be detected up to now (Van den Heuvel et al. 2000) (see *Anxiety and Anxiety Disorders*).

2.4 Alcoholism

Alcoholism is a brain disease. In order to emerge, drinking alcohol is a prerequisite. Drinking is common in the general population and does not necessarily induce alcoholism as a disease. Only in a subgroup does drinking proceed to alcoholism. Twin studies in the general population have demonstrated that drinking of alcohol itself is genetically determined in a complex manner. Three behavioral dimensions are widely independently influenced by different factors:

(a) time pattern of abstinence,
(b) frequency of drinking,
(c) quantity of drinking.

Particularly the frequency and the quantity of drinking seem to be under genetic control in the general population but counter to expectancy both traits are not influenced by the same genes.

Alcoholism is a complex behavioral condition characterized by compulsive drinking; crucial signs are abuse (consumption in spite of anticipated adversity) and/or dependence/addiction (e.g., loss of control on drinking, tolerance to alcohol, unsuccessful attempts to quit drinking, repeated withdrawal syndromes, inability to abstain and continuous consumption even in the morning). Both conditions require the availability and the consumption of alcohol. Given the variability of these conditions across countries and milieu conditions, the prevalence rates vary due to sociocultural sources. Consistently, alcoholism is less common among females. Although these nonbiological factors explain a limited degree of familial aggregation of alcoholism which is reported in a series of family studies (relative risks varied between 2 and 10 with a very broad range of absolute lifetime prevalences), genetic factors are an even stronger contributor (at least among males). Five twin studies report heritability rates between 30 and 60 percent with mainly stronger effects in males.

Some adoption studies also point in the same direction. Adoption studies have also revealed the genetic heterogeneity of alcoholism (Cloninger 1987, Cadoret et al. 1995). It was proposed that inability to abstain (combined with early onset of the disease) on the one hand, and lack of control on the other hand were genetically independent. It was also suggested that a common subtype of alcoholism characterized by a combination with antisocial personality disorder is genetically distinct from alcoholism beyond a familial loading with antisocial personality disorders.

These observations motivated subtype classification of alcoholism with an early onset variant, inability to abstain, and antisocial and criminal behavior as a subtype with the strongest genetic basis which is qualitatively and quantitatively different from the genetic forces determining drinking in the general population; the other subtypes are also genetically influenced but less strongly so (Heath et al. 1991, 1997, Cadoret et al. 1995).

The metabolism of alcohol is under the control of enzymes which present as various isoforms each with various genetic variants with differential activity (aldehyde dehydrogenase—ALDH, alcohol dehydrogenase—ADH). Carriers of two copies of the less active variant of one isoform of ALDH react to alcohol in an aversive manner (flushes in the face, nausea) creating a barrier for excessive or long-term use of alcohol with a reduced risk for alcoholism as a consequence. Similarly, carriers of genetic ADH variants with reduced activity are associated with lower risks for alcoholism. These influential allelic variants for alcoholism are protective against alcoholism. Thus, apart from late-onset Alzheimer's disease, alcoholism is the only common mental disorder with well confirmed susceptibility genes. However, the frequency of allelic variants associated with reduced metabolic activity vary across populations with relatively high frequencies in Asian populations and with neglectible prevalences in Caucasian (European) populations.

Other risk factors for alcoholism and other substance abuse are personality features such as antisocial behaviour (disorder) or novelty seeking. These personality patterns are partly under genetic control and may influence the use, abuse and addiction of alcoholism on a genetic basis (Cloninger 1987, Cadoret et al. 1995). The genetic impact of personality on alcoholism, however, is via a gene–environmental interaction as the availability of alcohol or other substances is a prerequisite.

Drug addiction is currently the only common psychiatric disorder with available valid animal disease models. Genetic manipulations of addictive behavior became feasible using transgenic techniques. Thus, major progress in unraveling the genetic basis of drug addiction in mice has been made and will elucidate the molecular genetics of alcoholism in the near future (Nestler 2000) (see *Alcohol-related Disorders*).

3. *Conclusion*

The genetics of mental disorders has been the topic of a long and controversial debate. Only various rare variants (prevalences substantially lower than 1 percent) of mental retardation and early onset Alzheimer's disease have been demonstrated to be classical genetic diseases with mutations at one gene locus causing the clinical syndrome (monogenic diseases). Most mental diseases are highly prevalent in the general population, such as schizophrenia, affective disorders, anxiety disorders or addiction (at least 1 percent lifetime prevalence), also called common diseases. Common mental disorders share a series of features with common diseases (e.g., hypertension, cardiac heart disease, diabetes mellitus):

(a) All common mental diseases are familial (i.e., relatives of patients are more likely to be affected with the same disease than random subjects in the general population).

(b) Twin studies were performed for all common mental diseases pointing to genetic causes.

(c) There is evidence that influences on the manifestation of the disease at least partly derives from the genetically influenced underlying behavioral traits (e.g., personality dimensions as neuroticism) ranging between mental health and illness (Bouchard 1994).

(d) The concordance rate among monozygotic twins is far from 100 percent; thus, the etiology is multifactorial with environmental as well as genetic factors contributing.

(e) The phenotype as transmitted in families is variable between relatives; unlike monogenic diseases a clear distinction between affected and healthy status is impossible.

(f) The familial aggregation of one disease often goes together with the co-aggregation of another disease (e.g., generalized anxiety disorder and depression); common genetic factors are mainly responsible for this coaggregation in families and partly explain the excess comorbidity among patients.

(g) The genetics of each common mental disease is complex and associated with genetic heterogeneity; the genetics does not follow a clear Mendelian mode of transmission, but genome-wide linkage studies argue for the contribution of multiple genes to each disorder. Consequently, the contribution of a single genetic variation is not causal but only probabilistic (susceptibility genes influencing the risk of the disease).

(h) Multiple susceptibility genes for a disease may either emerge from multiple monogenic, clinically unrecognized subtypes or from the simultaneous and cumulative contribution of several genes. Whereas the first possibility cannot be excluded definitively, the second possibility is substantially more plausible given the lack of Mendelian transmission. Whereas the contributing gene variants are likely to be rare under the first condition, the contrary should be the case under the second condition. Each of these common gene variants influencing the risk is likely to be of ancient origin dating back about 100,000 years. The functional consequence of each variant is presumably modest, facilitating overcoming the process of selection. In contrast, rare genetic variants causing monogenic diseases are more recent in origin.

(i) Strong efforts using very similar techniques are currently being undertaken for each of these diseases

to identify the contributing genes. Recently, major progress has come from linkage studies providing knowledge of the localization of susceptibility genes. Currently only very few susceptibility genes are identified. Due to the progress of the Human Genome Project and the development of high throughput techniques the detection of susceptibility genes for most common mental disorders can be expected in the near future.

See also: Behavioral Genetics: Psychological Perspectives; Familial Studies: Genetic Inferences; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Factors in Cognition/Intelligence; Genetic Studies of Personality; Genetic Testing and Counseling: Educational and Professional Aspects; Genetics of Complex Traits Through the Life Cycle; Mental and Behavioral Disorders, Diagnosis and Classification of; Mental Illness, Epidemiology of; Schizophrenia and Bipolar Disorder: Genetic Aspects

Bibliography

Amir R E, Van den Veyver I B, Wan M, Tran C Q, Francke U, Zoghbi H Y 1999 Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics* **23**: 185–8

Berrettini W H, Ferraro T N, Goldin L R, Weeks D E, Detera-Wadleigh S, Nurnberger J I, Gershon E S 1994 Chromosome 18 DNA markers and manic-depressive illness: Evidence for a susceptibility gene. *Proceedings of the National Academy of Science USA* **91**: 5918–21

Bouchard T 1994 Genes, environment, and personality. *Science* **61**: 1700–1

Cadoret R J, Yates W R, Troughton E, Woodworth G, Stewart M A 1995 Adoption study demonstrating two genetic pathways to drug abuse. *Archives of General Psychiatry* **52**: 42–52

Cloninger C R 1987 Neurogenetic adaptive mechanism in alcoholism. *Science* **236**: 412–20

Craddock N, Jones I 1999 Genetics of bipolar disorder. *Journal of Medical Genetics* **36**: 585–94

Gershon E S 2000 Bipolar illness and schizophrenia as oligogenic diseases: Implications for the future. *Biological Psychiatry* **47**: 240–4

Heath A C, Bucholz K K, Madden P A F, Dinwiddie S H, Slutske W S, Bierut L J, Statham D J, Dunne M P, Whitfield J B, Martin N G 1997 Genetic and environmental contributions to alcohol dependence risk in a national twin sample: Consistency of findings in women and men. *Psychological Medicine* **27**: 1381–96

Heath A C, Meyer J, Jardine R, Martin N G 1991 The inheritance of alcohol consumption patterns in a general population twin sample: II. Determinants of consumption frequency and quantity consumed. *Journal of Studies on Alcohol* **52**: 425–33

Kendler K S, Kessler R C, Neale M C, Heath A C, Eaves L J 1990 The prediction of major depression in women: Toward an integrated etiologic model. *American Journal of Psychiatry* **150**: 1139–48

Kendler K S, Walters E E, Neale M C, Kessler R C, Heath A C, Eaves L J 1995 The structure of the genetic and environmental risk factors for six major psychiatric disorders in women. Phobia, generalized anxiety disorder, panic disorder, bulimia, major depression, and alcoholism. *Archives of General Psychiatry* **52**: 374–83

Kety S S, Wender P H, Jacobsen B, Ingraham L J, Jansson L, Faber B, Kinney D K 1994 Mental illness in the biological and adoptive relatives of schizophrenic adoptees. *Archives of General Psychiatry* **51**: 442–55

Maier W, Schwab S, Rietschel M 2000 The genetics of schizophrenia. *Current Opinion in Psychiatry* **13**: 3–9

Merikangas K R, Avenevoli S, Dierker L, Grillon C 1999 Vulnerability factors among children at risk for anxiety disorders. *Biological Psychiatry* **46**: 1523–35

Nestler E J 2000 Genes and addiction. *Nature Genetics* **26**: 277–81

Straub R E, MacLean C J, O'Neill F O, Burke J, Murphy B, Duke F, Shinkwin R, Webb B T, Zhang J, Walsh D et al. 1995 A potential vulnerability locus for schizophrenia on chromosome 6p24-22: Evidence for genetic heterogeneity. *Nature Genetics* **11**: 287–93

Sullivan P, Neale M C, Kendler K S 2000 Genetic epidemiology of major depression: Review and meta-analysis. *American Journal of Psychiatry* **157**: 1552–62

Tienari P, Wynne L C, Moring J, Lahti I, Naarala M, Sorr A, Wahlberg K-E, Saarento O, Seitamaa M, Kaleva M, Läksy K 1994 The Finnish adoptive family study of schizophrenia. *British Journal of Psychiatry, Supplement* **23**: 20–6

Tsuang M 2000 Schizophrenia: genes and environment. *Biological Psychiatry* **47**: 210–20

Van den Heuvel O A, Van de Wetering J M, Veltman D J, Pauls D L 2000 Genetic studies on panic disorder: A review. *Journal of Clinical Psychiatry* **61**: 756–66

Winokur G, Coryell W, Keller M, Endicott J, Leon A 1995 A family study of manic-depressive (bipolar I) disease. *Archives of General Psychiatry* **52**: 367–73

W. Maier

Mental Imagery, Psychology of

The capacity of the human mind to store traces of past sensory events undoubtedly has considerable adaptive value, by enabling human beings to retrieve and consult information about absent objects or remote events. Obviously, the destiny of large portions of the human's daily experience is to be forgotten, but the ability to preserve and reactivate sensory traces of objects or events in the form of conscious internal events is a feature of great significance for a living organism.

1. Definitions

A 'mental image' is a cognitive event that encodes figural information about currently nonperceived objects and, in metaphorical terms, renders absent objects present to the mind. 'Mental imagery' refers to the mechanisms involved when a person builds the internal representations that encode the figural content of objects or events, stores them in a memory store, and later reinstates the original information by means

of some form of reactivation. Reactivation can ultimately result in a material output, such as a graphic production intended to reflect the appearance of a memorized object. Reactivation can also remain an internal, private event. Some of the information can be externalized through verbal discourse, when the person describes an imaged object, whether this is in response to an external request or because the person spontaneously intends to express knowledge about the object.

Even if it were to be limited to the evocation of past experience, mental imagery would still be a very valuable capacity. Actually, once an organism is endowed with the capacity of creating images, this enables it to process remote or absent objects, which extends the range of its potential actions in the world considerably. Moreover, beyond simply reproducing past perceptions, mental imagery can be used for creative purposes, to create mentally structures or patterns that have never been experienced perceptually, and in fact could well be impossible for the person to experience. Imagination, as the faculty of constructing new patterns, extends the range of human action. Creative or anticipatory imagery is essential in many forms of human activity, such as engineering, architecture, art, and design.

Imaginal experiences can be generated in relation to all sensory modalities. So far it is the visual modality that seems to have elicited most of the theoretical and empirical research in psychology, but auditory imagery and, more recently, olfactory and kinesthetic imagery, are now attracting the interest of numerous researchers. However, beyond identifying the range of imagery experiences, the main challenge for research is, first, to confirm the existence of psychological events that can be characterized as 'images,' and then go on to describe their content and structure, examine their role in cognitive functioning, and, last but not least, provide evidence that mental images have specific features that make them a distinct category of mental representation. The major thrust of research in recent years has been to consider images as specific forms of representation, and to account for their unique contribution to human cognition. As a result, imagery tends to be envisaged in the context of multimodal theories of the human mind.

2. *Physiological and Brain Correlates of Mental Imagery*

One of the greatest difficulties for psychology is to assess the occurrence and content of private events. Psychologists have attempted to go beyond collecting verbal and graphic expressions of these events, and collect indirect, but it is hoped reliable, indicators of mental imagery. The underlying assumption is that if cognitive events can only be 'viewed' by the minds that generate them, and not by external observers, then measurable correlates are needed that can be related to the properties of the reported images. For instance, electrodermal response and heart rate have been shown to be affected by the emotional charge of imagined objects. However, it is still unclear whether it is the images themselves that cause the observed physiological responses, or whether some more abstract cognitive entities are responsible for both the subjective experience of imagery and the concomitant physiological responses.

Several forms of ocular activity have been envisaged as potential reflections of imaginal experience in the visual modality. If imagery consists of reinstating at least part of the patterns of activity that took place during initial input processing, it may be useful to find out whether ocular motor functions reflect anything of this processing. Jean Piaget initially promoted this approach in the context of his theory of imitative imagery. However, the data have never confirmed any clear relationship between the patterns of eye activity during perception and in the imagery of specific objects. The only effect that has been established clearly is the pattern of ocular activity that accompanies the imagination of objects in motion (such as a pendulum). Another measurement that has elicited a great deal of interest is pupil dilation. The time course of pupil diameter change exhibits quite different patterns depending on how easily images can be generated in response to concrete or abstract nouns (so that the generation of readily available images of concrete objects contrasts with that of associative images attached to abstract concepts). However, this is more likely to reflect the cognitive load associated with the generation of images for abstract concepts, rather than the processes underlying image construction *per se*.

Not surprisingly, research has tended to question the real value of information collected from neurovegetative responses, and concentrate more and more on evidence obtained from measurements of brain activity. Electroencephalographic recordings of the brain activity that accompanies the formation of visual images have established reliably that the alpha rhythm decreases in the occipital areas when a person is generating and 'looking at' visual images. Similarly, studies based on evoked potentials have shown maximal positivity in the occipital and posterior temporal regions of the brains of individuals who form visual images of familiar objects in response to object nouns. Thus, the regions of the brain involved in visual processing also appear to be implicated in the generation and manipulation of mental images. These empirical findings have been interpreted as suggesting that perception and imagery may not only share specific sites of the neural architecture, but may involve similar mechanisms. In addition, a large amount of empirical evidence of intimate functional interactions between images and percepts (both facilitation and interference) suggests that a common neuronal sub-

strate underlies both activities. When people are invited to detect a visually presented shape while imagining simultaneously the same or another distinct shape, the amplitude of the evoked potentials is greater when the two shapes match.

Neuroimaging techniques have provided numerous data that corroborate the indications based on electrophysiological measurements, and provide still more accurate information about the regions involved in generating and manipulating visual images. Single photon emission computerized tomography (SPECT) initially indicated that the occipital cortex is implicated in the production of visual images. In tasks where the participants had to verify sentences, and presumably relied on visual representations of the situations described, the verification of sentences that were more likely to involve imagery (such as '*The green of fir trees is darker than that of grass*') was accompanied by more occipital activity than the verification of sentences which did not call upon imagery ('*The intensity of electrical current is measured in amperes*'). Subsequently, positron emission tomography (PET) has provided converging evidence for a variety of imagery tasks. However, evidence for the role of some specific occipital regions is not consistent or unambiguous. Some authors have reported activity of the primary visual cortex in visual imagery, but others failed to find any such activity in this region, but only in the associative regions (mainly the temporo-occipital and parieto-occipital regions). Furthermore, the studies that reported activity in the primary visual cortex also suggested the topographical organization of the cortical regions involved in visual imagery, whereas the other studies seemed to indicate that the cortical areas serving mental imagery were only a subset of the areas involved in visual perception.

The recent development of more sophisticated techniques, such as functional magnetic resonance imaging (fMRI), has provided strong evidence in support of the view that early visual areas may be responsible for both visual perception and imagery. The discrepancies among the various neuroimaging studies can in fact mainly be attributed to individual differences. They may also reflect the differing degrees of image resolution required by different imagery tasks. Different imagery tasks may not require similar 'grain' or resolution to be achieved, and may therefore involve different regions of the brain. The concept developed by Stephen Kosslyn distinguishes between three sorts of imagery. The first is 'spatial imagery,' or imagery dedicated to the mental representation of spatial relationships, in which the visual character of the image currently evoked is not crucial to performing the task. Occipito-parietal regions of the brain seem mainly to be responsible for this type of imagery. The second form is 'figural imagery,' which occurs only when a low-resolution topographic image is generated from the activation of stored representations of shapes or objects. The inferior temporal cortex seems to be primarily involved in this form of imagery, which does not call for high-resolution components. The third form is 'depictive imagery,' which is thought to rely on high-resolution representation in the primary visual cortex. It is involved in tasks that require a fine-grained representation, such as when shapes have to be interpreted or compared. Shared neural processes are thought to underlie both perception and depictive imagery.

Converging evidence about the role of the occipital regions in mental imagery is also available from neuropsychology, in particular from cases involving a documented syndrome of 'imagery loss.' This deficit can occur without any impairment of perceptual recognition. Neuropsychological investigations attest that the brains of patients with an impaired capacity to generate visual images also display cortical lesions in the occipital regions. Furthermore, patients with temporo-occipital lesions are unable to gain any advantage from imagery instructions in verbal learning tasks that normally benefit from imagery. It is worth mentioning that patients who are suffering from unilateral perceptual neglect may also show similar unilateral deficit when they are invited to report the content of scenes from their imagination. These data suggest that the central mechanisms used for imaginal processing mirror the mechanisms involved in perception. However, cases of pure representational neglect also occur, without any neglect of objects present in the patient's visual environment.

3. *Behavioral Attestations of Mental Imagery*

So far it looks as though several objective measurements can be linked systematically to the occurrence of mental imagery. Some of them are presumably no more than physiological events that accompany imagery activity, whereas measurements of brain activity are thought to be more intimately related to the actual process of image construction. However, strictly speaking, neither of these measurements gives psychologists any information about the cognitive events that interest them. This shortcoming has led to another line of research, which has also been quite popular among imagery researchers. The argument is that if the intimate mechanisms of images cannot be defined, even by a thorough analysis of the changes of their physiological concomitants, it could be more helpful to look at the effects of imagery on the behavior of people invited to form mental images while performing a cognitive task.

The paradigm used in this approach is very simple. A person is invited to carry out a cognitive task that will be expressed in the form of a measurable performance. Under control conditions, only the instructions to carry out the task are given. Under the test conditions, the same task has to be performed, but in addition the participant is instructed to make use of

mental imagery while carrying it out. In conventional memory tests, for instance, participants presented with a list of concrete nouns may be invited to memorize these nouns (control condition), or they may receive the additional instruction to generate visual images of the objects designated by these nouns (experimental condition). When recall is measured at the end of the experiment, the comparison of the two scores can be expected to show whether imagery instructions have had any impact, and if so, whether this has been beneficial or detrimental. Many empirical studies carried out using this very simple approach have provided evidence that imagery does have a positive impact on the memory of both verbal and pictorial materials. Provided the participants are allowed enough time, the effect is not very different from that produced by presenting pictures depicting the objects in addition to the nouns. However, the effect cannot be considered to be just a nonspecific result of the fact that extra processing has been imposed by the instructions, since other powerful strategies, as those based on linguistic elaboration, do not produce an effect of the same magnitude.

The facilitating effects of imagery on verbal memory have been assessed for more complex materials than lists of words. They have been shown to occur for tasks ranging from paired-associate learning through sentence memory, to memory of paragraphs or texts. In addition, in a variant of the paradigm in which no imagery instructions are given, but where the investigator compares the performance of individuals identified as 'high' and 'low' imagers respectively (on the basis of specific psychometric assessment), the former have higher memory scores, suggesting that they spontaneously make use of their capacity to convert verbal inputs into mental images. The impact of individual differences, however, is only found for materials that can be readily imaged. For instance, high imagers do better than low imagers in remembering narratives involving characters, scenery, and events that are easy to image, but the groups do not differ with regard to their memory of abstract texts. In comprehension tasks, high imagers process descriptions of spatial configurations more quickly than low imagers, and they recall their content more accurately. This is especially true when the poor sequential structure of these descriptions creates special demands on the readers' cognitive resources. High imagers need less time than their counterparts to process spatial descriptions in which configurations are described according to unexpected or incoherent sequences.

If one considers other domains of cognitive processing, such as reasoning or problem solving, there is ample demonstration that strategies based on the visualization of the data and their combination into integrated figures facilitates the successful performance of the task. This is true, for instance, for the resolution of spatial problems, as well as the resolution of three-term syllogisms. However, in the case of syllogisms, only visuo-spatial relationships (such as those expressed by '*above–below*') and relationships that can be visualized metaphorically ('*better–worse*') benefit from imagery instructions, but this is not the case for relationships with no spatial content ('*darker–brighter*'). It is worth noting that imagery is being used here as an alternative to other powerful methods, such as reasoning based on the rules of logic. Imagery achieves the visual picturing of displays from which solutions can be 'read out' without any recourse to formal reasoning.

The same is true of visuo-spatial problems such as the following: '*Think of a cube, all six surfaces of which are painted red. Divide the cube into 27 equal cubes by making two horizontal cuts and two sets of two vertical cuts each. How many of the resulting cubes will have three faces painted red, how many two, how many one, and how many none?*' This problem can be solved by a reasoning procedure totally devoid of any imagery. For instance, to decide how many cubes have three faces painted red, people can access available information in their knowledge base, indicating that a cube has eight corners, and that since a corner is defined by the intersection of three faces, there will be eight cubes with three red faces. However, it is remarkable that the vast majority of people, even those who have mastered sophisticated reasoning methods, tend to rely on visual imagery when they have to solve this sort of problem.

In most of the cases mentioned above, researchers are forced to conclude that imagery does indeed have positive effects, but that alternative strategies that do not call upon mental imagery can also be quite efficient. Imagery, then, is not to be seen as a 'cognitive panacea,' but rather as an especially efficient cognitive procedure among a variety of strategies. Other situations of interest are those where the question to which the person must respond concerns an aspect of an object or a situation that he or she has never processed before, but which is nevertheless accessible from memory. In other words, the person is invited to consider some aspect of the situation for the first time, and asked a question, the answer to which is unlikely ever to have been stored in the memory in a linguistic or propositional form. For instance, on a map of Europe, do Paris, Berlin, and Moscow lie on a straight line? Or in which hand does the Statue of Liberty hold her torch? If people are able to answer such questions, this is probably because information has been retained from previous exposure to the objects in a form that preserves their visual and spatial characteristics. In such cases, imagery seems to be the only way a person could possibly access the relevant information.

Imagery does not only involve reinstating the visual patterns of previously seen objects. In most problem-solving situations, imagery involves carrying out a series of transformations of imagined figures. By combining images of individual objects or parts of objects, the person can produce imaged patterns and

subject them to novel interpretations. This process is presumably an important component of discovery processes. It works in contexts where the transformations are carried out in response to explicit verbal instructions from an investigator, but also in less constrained conditions, such as creative visual synthesis. Creative processes in art and design obviously rely on these capabilities of the human mind.

4. *The Structural Properties of Mental Images*

When brought into play in the context of cognitive functioning, imagery is generally shown to have beneficial effects on performance. Although these effects have been confirmed by hundreds of experiments, in themselves they do not tell us anything about the properties of mental images, or why they are so powerful in cognitive processing. In some cases, the effects of imagery can be explained as providing an extra opportunity to process information. Thus, simply because it is more advantageous to encode any item of information under two forms rather than just one, the addition of imagery to cognitive processing should enhance performance. The concept of dual coding, as illustrated and advocated by Allan Paivio, is mainly based on this type of explanation, although the dual code theory also introduces the further argument that image codes have intrinsic properties that render them more powerful than other, mainly verbal codes.

This situation has led researchers to explore the possibility that mental images may have structural properties that distinguish them from other forms of representation, and that these properties could account for their functional properties. Research has thus shifted toward using empirical methods intended to assess the intimate characteristics of mental images. The basic tenet of this approach is the distinction between long-term memory representations that encode figural information (in a form that is left unspecified), and images as transient cognitive events resulting from the activation of these long-term representations. The subjective counterpart of the activation of these patterns of figural information is the occurrence of a conscious image. The theory of mental imagery developed by Stephen Kosslyn delineates the properties of the mental 'medium' (or 'visual buffer') on which images are thought to be activated.

The visual buffer is conceived of as a matrix endowed with the functional properties of a coordinate space. It has limited resolution and is subject to processing constraints similar to those that affect visual perception. In particular, images, like percepts, are constrained with regard to the apparent size of imagined objects. Just as it is more difficult to distinguish details of an object on a tiny photograph than on a larger one, so a detail of an imagined object is more rapidly 'detected' in images of objects that are imagined at a larger size. Moreover, when an object is imagined at a given size, the parts that occupy a larger space are more rapidly 'detected' than the less extensive parts. The processing device in which visual images are constructed has nonextensible limits which constrain the amount of information present simultaneously in these images.

A set of hypotheses on the processes underlying mental imagery has been developed in the light of this theoretical framework. These hypotheses concern the mental medium on which images are displayed, and they promote a 'modular' concept of imagery that distinguishes between the processes of generation, maintenance, exploration, and transformation of images. The view entertained here rejects the concept of imagery as a single undifferentiated function, but instead assumes that imagery corresponds to a set of distinct subcapacities that are largely independent of each other. This approach implies that people may differ with regard to one or more of these capacities, and the concept of a 'high imager' should be defined in terms of the specific processes actually contributing to this greater capacity. Consequently, there may be several ways of being a 'high imager,' depending on the imagery modules involved. Furthermore, people may be said to be 'high imagers' because they possess especially well-developed aptitudes, but also because they are inclined naturally to use imagery in preference to other strategies. Individual orientation toward imagery may also be determined by metacognitive awareness of the efficiency of images in memory and thinking.

Most attempts to identify the sources of efficacy of images converge on the assumption that images draw their efficacy from features that make them the best cognitive substitutes for actual perceptual events. The comparison of human behavior in perceptual and imaginal situations often reveals similarities of response patterns. Of course, people usually discriminate between their perceptions and their imaginal experiences, but there are many similarities in the way they access these two types of information. For instance, people's verbal descriptions of an object available to their current visual inspection or from memory are very similar. In other words, a percept and an image seem to yield patterns of information that impose comparable modes of processing. Furthermore, the parts of a configuration that are best remembered, because they are more salient or more remarkable, are also those best remembered from images of these configurations.

Chronometric measurements have proved a valuable way of identifying the critical features of the image structure. It has been shown that when people scan mentally between two points on the visual image of an imagined configuration, this takes a time that is proportional to the distance between the two points on the original visual configuration. This finding seems to

imply that visual images have a structure that analogically reflects the metric structure of the objects evoked. There is no suggestion that the pattern of activation corresponding to imaginal experience occurs on spatially defined portions of the brain, but simply that the spatial properties of objects, including their metric properties, are in some way reflected in their imaginal counterparts. The spatial characteristics of objects previously perceived are represented in visual images and stored in the memory. A further feature of interest is that spatial information can be included in the visual images of configurations that have been constructed from verbal descriptions and that the person has never actually seen. Even if the distances between objects are not explicitly expressed in the verbal description of a visuospatial configuration, the very process of constructing a mental image requires the visualization of the distances between the items that compose the configuration. This is due to an essential characteristic of analog representations, where the fact of positing objects at specific locations at the same time inevitably displays the spatial relationships among these objects.

Research on mental rotation has also yielded data suggesting the analogous character of visual images. The basic finding from Roger Shepard's paradigm is that the time it takes to rotate an object mentally increases as a function of the size of the angle of rotation. This finding complements discoveries from mental scanning, and shows that images not only possess a structure that reflects the object's structure in an analogous fashion, but also that images are transformed in a manner analogous to the way actual objects are perceived or manipulated. A remarkable fact is that people who exhibit such chronometric patterns in mental scanning or mental rotation experiments are not at all aware of these relationships. They do seem to realize how useful images can be in daily tasks, but they have no intuitive perception of the intimate mechanisms of mental imagery.

5. *Conclusions*

Like any psychological event accessible to introspection, images can be described by people in terms of their content, vividness, clarity, and degree of detail. Researchers obviously favor objective assessments of internal events, through the use of indicators expected to correlate with described images. Data that provide information about the neural structures that are involved when mental images are generated are even more valuable. A particular advantage of this approach in recent years is that it has allowed researchers to uncover the many similarities between imagery and perception. In this respect, behavioral and neuroimaging studies have progressed hand-in-hand in an especially productive manner.

An important objective of research on imagery is to account for the relationships between the structural properties of images and how they function when they are brought into play in cognitive activities. The assumption is that images draw their functional effectiveness from the properties that they share uniquely with perceptual events. Unlike other, more abstract, forms of representation, images contain information structured analogously to perceptual information, and this gives them particular adaptive value. Imagery provides representations that allow individuals to retrieve information in the absence of the objects that they evoke, and so to process objects that are temporarily or definitely out of sight. The fact that the processes that are applied to images exhibit similar patterns to those of the perceptual processes gives them an obvious cognitive advantage.

To summarize, mental imagery is not unrelated to the other cognitive functions. In particular, it is intimately interconnected with perception, from which it derives its content, and for which it is a valuable functional substitute in many types of cognitive activity.

See also: Imagery versus Propositional Reasoning; Mental Models, Psychology of; Mental Representations, Psychology of; Visual Imagery, Neural Basis of

Bibliography

Denis M 1979 *Les Images Mentales*. Presses Universitaires de France, Paris

Denis M 1991 *Image and Cognition*. Harvester Wheatsheaf, New York

Denis M, Logie R H, Cornoldi C, de Vega M, Engelkamp J (eds.) 2001 *Imagery, Language, and Visuo-spatial Thinking*. Psychology Press, Hove, UK

Finke R A 1989 *Principles of Mental Imagery*. MIT Press, Cambridge, MA

Kosslyn S M 1980 *Image and Mind*. Harvard University Press, Cambridge, MA

Kosslyn S M 1983 *Ghosts in the Mind's Machine: Creating and Using Images in the Brain*. W. W. Norton, New York

Kosslyn S M 1994 *Image and Brain: The Resolution of the Imagery Debate*. MIT Press, Cambridge, MA

Miller A I 1984 *Imagery in Scientific Thought: Creating 20th-century Physics*. Birkhäuser, Boston

Morris P E, Hampson P J 1983 *Imagery and Consciousness*. Academic Press, London

Paivio A 1971 *Imagery and Verbal Processes*. Holt Rinehart & Winston, New York

Paivio A 1986 *Mental Representations: A Dual Coding Approach*. Oxford University Press, New York

Piaget J, Inhelder B 1966 *L'image Mentale Chez l'Enfant*. Presses Universitaires de France, Paris

Richardson J T E 1980 *Mental Imagery and Human Memory*. Macmillan, London

Richardson J T E 1999 *Imagery*. Psychology Press, Hove, UK
Shepard R N, Cooper L A 1982 *Mental Images and their Transformations*. MIT Press, Cambridge, MA

M. Denis

Mental Maps, Psychology of

Consider the following thought experiment. You are blindfolded in a large room. Someone guides you walking along an L-shaped path. That is, they lead you some distance straight ahead, then guide you in a turn, and then lead you another distance in the new direction. After this you are instructed, still blindfolded, to return directly on your own to the place where your guided walk started. Most people can accomplish this task reasonably accurately. Logically, to accomplish this task you would seem to have to keep in mind the location where you started and how its distance and direction changes as you walk. Interestingly, and perhaps surprisingly, blind persons do not perform this as accurately as blindfolded sighted persons. In some sense, this task seems to involve a form of mental representation, a mental map, of this very simple space, one that includes the starting location and this elementary path.

The idea of mental maps is more interestingly applied to more complex spaces and in cases where vision is not completely eliminated by blindfold or blindness. Edward Tolman (1948) may have been the first to raise this concept for psychology in the course of investigating maze learning in rats. He noted that his animal subjects moved in a generally appropriate direction when a well-learned path to the goal was blocked. He considered this as evidence for their general spatial orientation to the situation and some general knowledge of the spatial layout. He used the term 'cognitive map' to refer to this knowledge. Of particular interest was that the rats seemed to be going beyond the information with which they had been specifically trained. They had learned a series of right and left turns to get to the goal but then were able to respond in a spatially appropriate way when that sequence of turns could not be executed. Since Tolman's original observations, the presence of cognitive maps in his sense have been observed in a wide variety of species. See Gallistel (1990) for a review.

1. *The Organization of Spatial Knowledge*

In humans, for example, even young children demonstrate behavior in which detours or short cuts are taken by which the child traverses parts of the spatial layout with which he or she has had no previous experience. As with Tolman's rats, children can be taught a zigzag route through a complex of rooms and then be tested on their ability to make detours or take alternative routes with which they have had no direct experience. By the age of six or seven years they can do this quite readily (Hazen et al. 1978). In fact, much of the research on spatial cognition in humans has focused on how well people acquire configurational knowledge of an overall spatial layout with only experience with specific routes through the space. Since routes often are defined partially, if not totally, in terms of making left and right turns, walking straight ahead, etc., route knowledge is associated with egocentric representation of spatial information. That is, locations are defined or described in terms of their relation to the observer. Configurational knowledge is usually understood as knowledge of the locations within a space in relation to each other. In this respect configurational knowledge is often associated with allocentric representation of spatial information. That is, locations are defined or described in relation to features of the space outside of the observer. Tolman's term of cognitive map has usually been applied to configurational knowledge as opposed to route knowledge. However, the more general question is how spatial knowledge is organized, and perhaps it is better to use the less loaded term of 'mental map.'

1.1 *The Role of Experience in Determining the Nature of the Mental Map*

With this in mind it is natural to ask under what conditions does one's mental map reflect a route organization and under what condition does it reflect configurational organization of spatial information. Returning to the blind and blindfolded participants mentioned above, it seems that perhaps having had visual experience might predispose one toward configurational organization to a greater degree. This possibility is congruent with observations of professional blind mobility instructors who often observe that their clients have difficulty in making detours in areas in which they have been trained on particular routes. There is considerable classical research that also suggests that blind people, who know their way from place to place in an area, often have more difficulty than sighted persons in making detours when their usual path is blocked. That may be due to blind persons having a greater tendency than sighted to maintain a routelike organization of their spatial knowledge. Sighted people are more likely to shift to a configurational organization, at least after some familiarity with an area. Why might this be? As sighted people move around they can see how the directions and distances to objects and locations changes. With this knowledge at every point along a path they can see

where they are in relation to other locations. This is the basis of configurational knowledge.

In support of this possibility there is evidence that sighted people, even when blindfolded, update where they are in relation to other locations in the environment; there is much less tendency for blind people to do this (Rieser et al. 1988). What is there about visual experience that would cause such a difference in updating? One hypothesis is that the optical flow stimulation, available visually whenever a sighted person moves, provides information about the changing distance and direction of the visible locations in the environment. This information, which is not available to blind people, calibrates sighted persons' stepping behavior and that calibration can be used when walking without vision. The bottom line is that having visual experience facilitates the generation of mental maps with a configural organization of spatial knowledge.

2. *The Nature of Configural Mental Maps*

What is the nature of these mental maps reflecting configural organization? A perusal of the research literature suggests that in the minds of many researchers the mental maps are like cartographic maps with a two-dimensional layout of a space viewed like a bird from above. But if operational criteria of configural mental maps include simultaneous awareness of the spatial relations of every location from every other location and the changing directions and distances to locations as one moves, there are other possibilities. One of the most interesting examples is that of the navigation of the seafarers of the Caroline Islands. They have a system that appears very strange to us but enables them to travel effectively across hundreds of miles of the South Pacific in outrigger canoes.

Their navigation system has been carefully studied by a number of anthropologists (e.g., Lewis 1978, Gladwin 1970). One of the most compelling analyses is that of Hutchins (Hutchins 1995, Hutchins and Hinton 1984). The navigators essentially represent the directions of islands to which they want to sail in terms of a celestial compass. That is, they know which direction to sail in terms of where along the horizon particular stars rise and set. And these directions are known for travel from every island to every other island in their sailing environment. (Thus to become a navigator requires a huge amount of memorization.) They index the progress of a trip in terms of the changing direction of an out-of-sight reference island. That direction is again specified in relation to the location of rising and setting stars. At the beginning of a voyage the reference island may be in the direction of a star which rises somewhat in the direction of the heading of the boat, say at a relative bearing of 45 degrees. As the journey proceeds the direction of the reference island moves toward the side and then maybe to the rear. The reference island is out of sight during the entire trip and, in fact, may not even exist in reality. (A hypothetical reference island is just as useful for keeping track of the progress of the trip as a real out-of-sight island.) From our perspective a most interesting and strange aspect of the Caroline Island conceptualization is that their canoe is stationary and the islands are moving with respect to the boat. Since, in travel, movement is relative, logically it makes no difference whether the traveler is moving or the environment is moving with respect to a stationary traveler.

3. *The Origin of Spatial Knowledge in Mental Maps*

How does the medium from which we acquire spatial information affect our mental map? This is another aspect of the experiential question. Much of our spatial knowledge comes from actual experience within a space. However, we also acquire spatial knowledge in a variety of other ways such as from verbal descriptions, from maps, from exposure to virtual reality, etc.

Take the case of maps, for example. Thorndyke and Hayes-Roth (1982) compared observers' mental maps after studying a map of a large building and after experience in actually navigating around the building. Map experience led to more accurate estimation of the straight line or Euclidean distances between locations characteristic of configural knowledge, whereas actual navigation led to more accurate route distance estimation and to more accurate judgments of actual direction of locations from station points within the building. Uttal (2000) has suggested that experience with maps has a more general effect on the kinds of mental maps we form. He suggests that the overhead view that prototypical maps provide, and the fact that maps by their scale transformation make available spatial relations which cannot be easily grasped by ordinary exploration, have a general effect on how children come to think about space and spatial relations.

There has been considerable recent interest in the relation between language and space (see, e.g., Bloom et al. 1996). In the case of language, it would seem possible to vary a text description to facilitate different kinds of mental maps, for example, biasing toward configural organization or biasing toward route organization. Tversky and her colleagues have found that the narrative provided for a listener can determine the perspective that the listener takes in thinking about the space (Tversky 1996). This is reflected in their reaction times to name the various objects in a space. They also found that observers in describing spaces from maps used either a route description or a

configural description or a mixture of the two, and rarely any other kind.

These questions are not just an issue of academic interest. We have all often had frustrating experience trying to understand verbal directions about how to get somewhere or trying to grasp the layout of an area by means of a map. Virtual reality is currently being proposed as having great potential for training people, e.g., soldiers, for tasks in new environments. However, with the present state-of-the-art it is difficult to build a good sense of the layout (a good mental map) of a virtual world one is moving through. How to use these media most effectively to enable the most desirable mental map is a goal for future research.

Bibliography

Bloom P, Peterson M A, Nadel L, Garrett M F (eds.) 1996 *Language and Space*. MIT Press, Cambridge, MA

Gallistel C R 1990 *The Organization of Learning*. MIT Press, Cambridge, MA

Gladwin T 1970 *East is a Big Bird*. Harvard University Press, Cambridge, MA

Hazen N L, Lockman J J, Pick H L Jr 1978 The development of children's representations of large-scale environments. *Child Development* **49**: 623–36

Hutchins E 1995 *Cognition in the Wild*. MIT Press, Cambridge, MA

Hutchins E, Hinton G E 1984 Why the islands move. *Perception* **13**: 629–32

Lewis D 1978 *The Voyaging Stars: Secrets of the Pacific Island Navigators*. Norton, New York

Loomis J M, Klatzky R L, Golledge R G, Cicinelli J G, Pellegrino J W, Fry P A 1993 Nonvisual navigation by blind and sighted: Assessment of path integration ability. *Journal of Experimental Psychology: General* **122**: 73–91

Rieser J J, Guth D A, Hill E W 1988 Sensitivity to perspective structure while walking without vision. *Perception* **15**: 173–88

Thorndyke P W, Hayes-Roth B 1982 Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology* **14**: 560–89

Tversky B 1996 Spatial perspective in descriptions. In: Bloom P, Peterson M A, Nadel L, Garrett M F (eds.) *Language and Space*. MIT Press, Cambridge, MA pp. 463–91

Tolman E C 1948 Cognitive maps in rats and men. *Psychological Review* **56**: 144–55

Uttal D H 2001 Seeing the big picture: Map use and the development of spatial cognition. *Developmental Science* **3**: 247–86

H. L. Pick Jr.

Mental Models, Psychology of

A mental model is a representation of some domain or situation that supports understanding, reasoning, and prediction. There are two main approaches to the study of mental models. One approach seeks to characterize the knowledge and processes that support understanding and reasoning in knowledge-rich domains. The other approach focuses on mental models as working-memory constructs that support logical reasoning (see *Reasoning with Mental Models*). This article focuses chiefly on the knowledge-based approach.

Mental models are used in everyday reasoning. For example, if a glass of water is spilled on the table, people can rapidly mentally simulate the ensuing events, tracing the water through its course of falling downward and spreading across the table, and inferring with reasonable accuracy whether the water will go over the table's edge onto the floor. People's ability to infer and predict events goes well beyond their direct experience. For example, if asked 'Which can you throw further, a potato or a potato chip?' most people can give an answer immediately (the potato) even if they have never actually tossed either item.

However, mental models are not always accurate. Mental models researchers aim to capture human knowledge, including incorrect beliefs. The study of incorrect models is important for two reasons. First, the errors that a learner makes can help reveal what the learning processes must be. Second, if typical incorrect models are understood, then instructors and designers can create materials that minimize the changes of triggering errors.

A striking example of an incorrect mental model is the *curvilinear momentum* error (Clement 1983, McCloskey 1983). When college students are asked: 'If a ball on a string is spun in a circle and then let go, what path will it take?', many of them correctly say that the ball will travel at a tangent to the circle. However, a fair proportion states that the ball will move in a curved path, retaining some of the curvilinear momentum gained from being spun in a circle. The usual intuition is that the ball will gradually lose this 'curvilinear momentum', so that the path will straighten out over time. This erroneous intuition is fairly general; for example, the same error turns up when people are asked about the path of a ball blown through a circular tube. Further, the error does not yield immediately to training; it is found even in students with a few years of physics. However, it does diminish with increasing expertise.

Another striking error is seen when people are asked what trajectory a ball will follow if it rolls off the edge of a table (McCloskey 1983). Instead of the correct answer, that the ball will fall in a parabolic path (Fig. 1a), many people believe the ball will continue traveling straight, and begin falling (either straight down or in a curved path) only when its forward momentum begins to flag (Fig. 1c and 1b). People seem to believe that sufficient forward momentum will overcome the tendency to fall. This error is sometimes called 'Roadrunner physics' because it

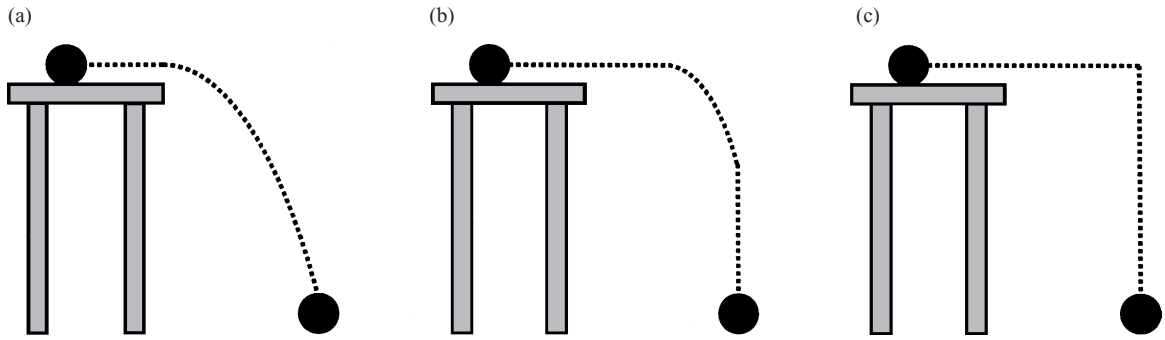

Figure 1
Responses to the question 'What path will the ball take after it rolls off the table?' (adapted from McCloskey 1983)

resembles the event in which a cartoon character runs off a cliff but does not fall until some distance over the edge. However, McCloskey noted that the same error occurs in the writings of Jean Buridan and other fourteenth-century Aristotelian philosophers. It appears that cartoon events were created to match a mental model that arises naturally from experience, possibly by overgeneralizing from experiences with linear momentum.

Mental models can facilitate learning, particularly when the structure of the new learning is consistent with the model. For example, Kieras and Bovair (1984) showed that subjects could operate a simulated device more accurately and could diagnose malfunctions better when they had a causal mental model of its functioning, rather than a merely procedural grasp of how to operate it. Similarly, Gentner and Schumacher (1986) showed that subjects were better able to transfer an operating procedure from one device to another when they had a causal mental model of the operation of the first device, rather than just a set of procedures. The degree of facilitation depended greatly on the match between the original model and the new material.

Mental models are used to explain human reasoning about physical systems: devices and mechanisms (de Kleer and Brown 1983, Hegarty and Just 1993, Kieras and Bovair 1984, Williams et al. 1983); electricity (Gentner and Gentner 1983); the interactions of people with computers and other devices (Norman 1988), and knowledge of home heating systems (Kempton 1986). They have also been applied to spatial representation and navigation (Forbus 1995, Hutchins 1983, Tversky 1991); ecology (Kempton et al. 1995), human population growth (Gentner and Whitley 1997), and the development of astronomical knowledge (Vosniadou and Brewer 1992).

Mental models are related to several other kinds of representational structures (see Markman 1999 for a comprehensive discussion). *Schemas* (or *schemata*) are general belief structures. *Scripts* are schemas summarizing event sequences, characterized by a chiefly linear temporal order, with limited inferential flexibility. *Naïve theories* or *folk theories* are global systems of belief, typically encompassing larger domains such as biology. The terms *mental models* and *naïve* or *folk theories* overlap in their application, though mental models are typically more specific than theories.

1. *Characteristics of Mental Models*

Mental models reasoning relies on *qualitative relations*, rather than on quantitative relations. People can reason well about the fact that one quantity is less than another without invoking the precise values of the quantities. This principle forms the basis for qualitative process theory, discussed below (Forbus 1984).

Mental models often permit *mental simulation*: the sense of being able to run a mental model internally, so that one can observe how it will behave and what the outcome of the process will be. The processes that underlie mental simulation are still under study. However, there is good evidence that people are able, within limits, to mentally simulate the behavior of a device, even if they are simply shown a static display (Hegarty and Just 1993). There is an apparent tradeoff between online simulation and retrieval of stored outcomes (Schwartz and Black 1996). As people become familiar with a system, they no longer carry out full simulations of behavior in all cases, but instead simply access their stored knowledge of the outcome.

Another finding of mental models research is that people are capable of holding two or more inconsistent models within the same domain, a pattern referred to as *pastiche models* (Collins and Gentner 1987) or *knowledge in pieces* (diSessa 1982). For example, Collins and Gentner (1987) found that many novice subjects had 'pastiche' models of evaporation. A novice learner may give one explanation of what causes a towel to dry in the sun and a completely different explanation of what causes a puddle of water

to evaporate, failing to see any connection between the two phenomena. Novices often use locally coherent but globally inconsistent accounts, often quite closely tied to the details of the particular example. This pattern emphasizes the tendency of novices to learn conservatively, with knowledge cached in highly specific, context-bound categories. So long as each model is narrowly accessed in contexts specific to it, the inconsistencies may never come to the learner's attention.

2. Mental Models in Everyday Life

Kempton et al. (1995) note that mental models 'give an underlying structure to environmental beliefs and a critical underpinning to environmental values.' For example, Kempton (1986) proposed on the basis of interviews that people used two distinct models of home heating systems. In the (incorrect) valve model, the thermostat is thought to regulate the rate at which the furnace produces heat; setting higher makes the furnace work harder. In the threshold model, the thermostat is viewed as setting the goal temperature, but not as controlling the rate of heating; the furnace runs at a constant rate. (This is the correct model for most current household systems.)

Having derived these two models from interviews, Kempton asked whether these models could explain people's real behavior in running their household furnaces. He examined thermostat records collected by Socolow (1978) from real households and found that the patterns of thermostat settings fitted nicely with the two models he had found. In particular, some families simply set their thermostat twice a day—low at night, higher by day, consistent with the threshold model—while others constantly adjusted their thermostats and used a range from extremely high to much lower temperatures. This is an extremely expensive strategy, in terms of fuel consumption, but it follows from the valve model. In this model, the thermostat setting controls how hard the furnace works, so the higher the setting, the faster the house will warm up. This reasoning can be seen in the analogies produced by Kempton's interviewees. Those with the valve model often compared the furnace to other valve devices, such as a gas pedal or a faucet and suggested that you need to 'turn 'er up high' to make the house warm up quickly. Thus, there is evidence that mental models can influence real-life environmental decision making.

Three significant generalizations can be made so far. First, people use mental models to reason with; they are not merely a convenient way of talking. Second, mental models can facilitate problem solving and reasoning in a domain. Third, mental models can yield incorrect results as well as correct ones. The next issues are where mental models come from and how they are used in learning and instruction.

3. Analogies and Mental Models

Mental models are often based on implicit or explicit analogies with other knowledge. The incorrect valve models used by Kempton's informants, discussed above, were apparently drawn from experiential analogies. However, analogical models can also be a useful way to extend knowledge from well-understood domains to less familiar domains. For example, Gentner and Gentner (1983) identified two common mental models of electricity, the *flowing water* model and the *moving crowd* model. In the flowing water model, current flows through a wire the way water flows through a pipe, and a resistor is a narrow pipe. In the moving crowd model, current is viewed as the rate of movement of a crowd through a hall, and a resistor as a gate through to the next hall. Although both analogies can account for many simple facts about d.c. circuits, they each have drawbacks. Voltage is easy to map in the flowing water model (the number of batteries corresponds to the number of pumps pushing the water forward), but it is awkward to map in the moving crowd model (unless perhaps to a loud noise impelling the crowd forward). In contrast, the behavior of resistors is easier to predict if they are seen as gates (as in the moving crowd model) than if they are seen as constrictions (as in the flowing water model). Thus, if these analogical models are really used in reasoning, people with the water model should reason more accurately about combinations of batteries than people with the crowd model, and the reverse for resistors. Indeed, that was what was found. When people filled out a questionnaire about their mental model of electricity, and then made simple predictions about combination circuits, people who held the flowing water model were more accurate about combinations of batteries, and those with the moving crowd model were more accurate about combinations of resistors.

4. Methods of Studying Mental Models

The initial elicitation of mental models is often done by the direct method of interviews or questionnaires that explicitly ask people about their beliefs (for example, Collins and Gentner 1987, Kempton 1986) or by analyzing think-aloud protocols collected during reasoning (Ericksson and Simon 1984) (see *Protocol Analysis in Psychology*). However, directly asking people about their mental models is not enough, for people are often unable to fully articulate their knowledge. Therefore, many researchers follow this direct interview with other methods of validating the proposed mental models. Once the mental models in a domain are roughly known or guessed, materials can be designed to bear down on the details. For example, problems are designed such that subjects' mental models can be inferred from patterns of correct and

incorrect answers, response times, eye movements, or particular errors made (Gentner and Gentner 1983, Hegarty and Just 1993, Schwartz and Black 1996) or patterns of retention for new materials in the domain (Bostrom et al. 1994).

5. Representing Mental Models

Mental models research often includes an explicit representation of the knowledge. For example, in Patrick Hayes' (1985) classic paper on the naïve physics of liquids, roughly 80 axioms are used to represent the knowledge involved in understanding the possible states a liquid can take and the possible transitions that can occur between states. These axioms capture knowledge about when a liquid will flow, stand still, or spread into a thin sheet on a surface.

A useful formalism for representing mental models is *qualitative process* (*QP*) *theory* (Forbus 1984). This theory, originating in artificial intelligence, aims to capture the representations and reasoning that underlie human reasoning about physical processes in a manner sufficiently precise to permit computer simulation. A central intuition is that human reasoning relies on qualitative relations, such as whether one quantity is greater or less than another, rather than on quantitative relations. For example, in QP theory, a mental model is represented in terms of (a) the entities in the domain—e.g., water in a pan; (b) qualitative relations between quantities in the domain—e.g., that the temperature of water is above freezing and below boiling; (c) the processes that create change—e.g., heat flow or liquid flow; and (d) the preconditions that must hold for processes to operate. An important feature of QP theory is that it uses *ordinal relationships* between quantities, such as that one quantity is greater than another, rather than representing quantities as numerical values. The idea is to match human patterns of reliance on qualitative relations rather than on exact values. A second important feature is that instead of using exact equations, QP theory uses a *qualitative mathematics* to provide a causal language that expresses partial knowledge about relationships between quantities. For instance, *qualitative proportionalities* express simple causal relations between two quantities. The idea is that people may know, for example, that greater force leads to greater acceleration, without knowing the exact numerical nature of the function (linear, exponential, etc.). An interesting aspect of QP theory is that, in addition to representing novice models, it can also capture an important aspect of expert knowledge: namely, that experts typically parse a situation into qualitatively distinct subsystems before applying more exact equations. QP theory allows researchers to describe people's knowledge about what is happening in a situation at a particular time, how the system is changing, and what further changes will occur.

6. Implications for Instruction and Design

Mental models developed from experience can be resistant to instruction. In the case of curvilinear momentum cited above, even students who had learned Newton's laws in physics classes often maintained their belief in curvilinear momentum. One technique that has been used to induce model revision is that of *bridging analogies* (Clement 1991). Learners are given a series of analogs. The first analog is a close match to the learner's existing model (and therefore easy to map). The final step exemplifies the desired new model. The progression of analogs in small steps helps the learner to move gradually to another way of conceptualizing the domain.

Mental models have been used in intelligent learning environments (see *Intelligent Tutoring Systems*). For example, White and Frederiksen's (1990) system for teaching physical reasoning begins with a simple mental model and gradually builds up a more complex causal model. Early in learning, they suggest, learners may have only rudimentary knowledge, such as whether a particular quantity is present or absent at a particular location. By adding knowledge of how changes in one quantity affect others, and then progressing to more complex relationships among quantities, learners can acquire a robust model.

Another implication of mental models research is that the pervasiveness and persistence of mental models needs to be taken into account in designing systems for human use. Norman (1988) argues that designers' ignorance of human mental models leads to design errors that plague their intended users. Sometimes these are merely annoying—e.g., a door that looks as though it should be pulled, but that needs to be pushed instead. However, failure to take mental models into account can lead to serious costs.

An example of such a failure of mental models occurred in the Three-mile Island nuclear disaster. Early in the events that led to the melt-down, operators noted that the reactor's coolant water was registering at a high pressure level. They interpreted this to mean that there was too much coolant and accordingly they pumped off large amounts of coolant. In fact, the level was dangerously low, so much so that the coolant was turning into steam—which, of course, led to a sharp increase in pressure. Had this alternate model been at hand, the operators might have taken different action.

7. Mental Models as Temporary Aids to Logical Reasoning

Another approach to mental models is taken by Johnson-Laird (1983) and his colleagues (see *Reasoning with Mental Models*). This approach differs from the research cited in the remainder of this article in that it views mental models as temporary working-memory sketches set up for the purposes of immediate

reasoning tasks such as propositional inference (Johnson-Laird 1983). The focus on immediate working-memory tasks in this approach has led to a relative lack of emphasis on long-term knowledge and causal relations. However, there may be value in bringing together the working-memory approach with the knowledge-intensive approach. There is evidence that long-term causal mental models can influence the working-memory representations that are set up in speeded tasks (Hegarty and Just 1993, Schwartz and Black 1996).

See also: Informal Reasoning, Psychology of; Mental Imagery, Psychology of; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning; Reasoning with Mental Models; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Bostrom A, Atman C J, Fischhoff B, Morgan M G 1994 Evaluating risk communications: Completing and correcting mental models of hazardous processes. Part II. *Risk Analysis* **14**(5): 789–98

Clement J 1983 A conceptual model discussed by Galileo and used intuitively by physics students. In: Gentner D, Stevens A L (eds.) *Mental Models*. Erlbaum, Hillsdale, NJ, pp. 325–40

Clement J 1991 Nonformal reasoning in experts and in science students: The use of analogies, extreme cases, and physical intuition. In: Voss J, Perkins D, Siegal J (eds.) *Informal Reasoning and Education*. Erlbaum, Hillsdale, NJ, pp. 345–62

Collins A, Gentner D 1987 How people construct mental models. In: Holland D, Quinn N (eds.) *Cultural Models in Language and Thought*. Cambridge University Press, Cambridge, UK, pp. 243–65

de Kleer J, Brown J S 1983 Assumptions and ambiguities in mechanistic mental models. In: Gentner D, Stevens A L (eds.) *Mental Models*. Erlbaum, Hillsdale, NJ, pp. 155–90

diSessa A A 1982 Unlearning Aristotelian physics: A study of knowledge-based learning. *Cognitive Science* **6**: 37–75

Ericksson K A, Simon H A 1984 *Protocol Analysis*. MIT Press, Cambridge, MA

Forbus K D 1984 Qualitative process theory. *Journal of Artificial Intelligence* **24**: 85–168

Forbus K 1995 Qualitative spatial reasoning: Framework and frontiers. In: Glasgow J, Narayanan N, Chandrasekaran B (eds.) *Diagrammatic Reasoning: Cognitive and Computational Perspectives*. MIT Press, Cambridge, MA, pp. 183–202

Gentner D, Gentner D R 1983 Flowing waters or teeming crowds: Mental models of electricity. In: Gentner D, Stevens A L (eds.) *Mental Models*. Erlbaum, Hillsdale, NJ, pp. 99–129

Gentner D, Schumacher R M 1986 Use of structure-mapping theory for complex systems. In: *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 252–8

Gentner D, Stevens A L (eds.) 1983 *Mental Models*. Erlbaum, Hillsdale, NJ

Gentner D, Whitley E W 1997 Mental models of population growth: A preliminary investigation. In: Bazerman M, Messick D M, Tenbrunsel A E, Wade-Benzoni K (eds.) *Environment, Ethics, and Behavior: The Psychology of Environmental Valuation and Degradation*. New Lexington Press, San Francisco, CA, pp. 209–33

Hayes P J 1985 Naive physics I: Ontology for liquids. In: Hobbs J R, Moore R C (eds.) *Formal Theories of the Commonsense World*. Ablex Publishing Corporation, Norwood, NJ

Hegarty M, Just M A 1993 Constructing mental models of machines from text and diagrams. *Journal of Memory and Language* **32**: 717–42

Hutchins E 1983 Understanding micronesian navigation. In: Gentner D, Stevens A L (eds.) *Mental Models*. Erlbaum, Hillsdale, NJ, pp. 191–225

Johnson-Laird P N 1983 *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge, MA

Kempton W 1986 Two theories of home heat control. *Cognitive Science* **10**: 75–90

Kempton W, Boster J S, Hartley J 1995 *Environmental Values in American Culture*. MIT Press, Cambridge, MA

Kieras D E, Bovair S 1984 The role of a mental model in learning to operate a device. *Cognitive Science* **8**: 255–73

Markman A B 1999 *Knowledge Representation*. Erlbaum, Mahwah, NJ

McCloskey M 1983 Intuitive physics. *Scientific American* **248**(4): 122–30

Norman D A 1988 *The Psychology of Everyday Things*. Basic Books, New York

Schwartz D L, Black J B 1996 Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology* **30**: 154–219

Socolow R H (ed.) 1978 *Saving Energy in the Home: Princeton's Experiments at Twin Rivers*. Ballinger, Cambridge, MA

Stevens A, Collins A 1980 Multiple conceptual models of a complex system. In: Snow R, Federico P, Montague W (eds.) *Aptitude, Learning and Instruction: Cognitive Process Analysis*. Erlbaum, Hillsdale, NJ, Vol. 2, pp. 177–97

Tversky B 1991 Distortions in memory for visual displays. In: Ellis S R, Kaiser M, Grunewald A (eds.) *Spatial Instruments and Spatial Displays*. Erlbaum, Hillsdale, NJ, pp. 61–75

Vosniadou S, Brewer W F 1992 Mental models of the Earth: A study of conceptual change in childhood. *Cognitive Psychology* **24**(4): 535–85

Williams M D, Hollan J D, Stevens A L 1983 Human reasoning about a simple physical system. In: Gentner D, Stevens A L (eds.) *Mental Models*. Erlbaum, Hillsdale, NJ, pp. 131–53

White B Y, Fredricksen J R 1990 Causal model progressions as a foundation for intelligent learning environments. *Artificial Intelligence* **42**(1): 99–157

D. Gentner

Mental Representation of Persons, Psychology of

Within social psychology, a major research focus is the processes by which perceivers form impressions of other persons, and the nature of the mental representations that they construct as a result. Mental representations or impressions of persons are organized configurations including many types of information, such as physical appearance, personality characteristics, and group memberships, as well as the perceiver's

reactions to the person (e.g., like or dislike). These representations influence strongly the perceiver's actions toward the person, such as choosing to interact with or to avoid the person, or helping or aggressing against the person (see *Social Psychology*).

1. Historical Development

In the period since the 1940s that representations of persons have been studied within social psychology, shifts of emphasis afford a rough chronological division into three periods.

1.1 The 1940s and 1950s

Research on mental representations of persons first flourished in the late 1940s, when Solomon Asch (1946) investigated how people construct impressions from limited, often conflicting information. Researchers presented lists of personality traits, such as 'honest, organized, and critical,' to research participants as descriptions of real persons. The participants then made judgments and evaluative ratings of the target persons. Researchers assessed the relative weights of different traits in the overall impression. For example, negative traits were often found to have more impact than positive ones and certain specific traits (such as 'warm' vs. 'cold') were found to have a disproportionate impact on an overall impression.

Another important question concerned how perceivers combined multiple items of information, such as traits. Some held that traits were combined algebraically, with the evaluative implications of different traits being combined by averaging (or other similar processes) into an overall evaluative judgment. In contrast, psychologists influenced by Gestalt theories, including Asch, held that traits were combined 'configurally,' with specific combinations taking on new and emergent meanings. Someone described as both 'cold' and 'intelligent,' for example, might be seen as using his intelligence in a calculating way and be judged as thoroughly negative, rather than as neutral (the result of averaging one negative and one positive trait).

A third major question concerned the accuracy of people's impressions of others. Researchers had participants rate other people (based on personal acquaintance or from limited information, such as an interview) on trait scales. These ratings could be compared to criteria such as ratings provided by the target or his or her close friends, to determine the accuracy of each rater. This line of research seemed important for both practical and theoretical reasons. What types of people could perceive others accurately, and therefore might make good counselors or personnel officers? But the line of work was dealt a near-fatal blow by Cronbach's (1955) critique dealing with problems in selection of the accuracy criterion as well as statistical and methodological issues. The question of accuracy received little attention for a generation after that date.

Through this period, person impressions were implicitly assumed to be relatively unorganized lists of traits. This assumption was generally consistent with mainstream memory research at the time, which used paradigms such as learning lists of unrelated words.

1.2 The 1960s and 1970s

The transition to the next era of research is clearly dated by the publication of Fritz Heider's *Psychology of Interpersonal Relations* (Heider 1958). Heider's work aid the groundwork for a focus on the process (rather than content or organization) of person perception, which characterized the next two decades. Attributional inferences—inferences about a person's inner qualities (such as traits) that cause observed behaviors—were a major focus of Heider's thinking and remained a central theme throughout this period (see *Attributional Processes: Psychological*). The issue is clearly central in person perception: we usually learn about other people by observing their behaviors (rather than by being presented lists of traits), so if impressions are ultimately composed of traits, we must understand how behaviors are mentally translated into traits.

Theorists such as Edward Jones and Harold Kelley advanced models of attributional inference that were tested in numerous experiments. Characteristics of the actor (e.g., his or her past behaviors), the behavior itself (e.g., its desirability or undesirability), and the social context (e.g., whether other actors behave in the same or a different manner) were investigated as influences on attributional judgments and therefore person impressions. Heider, Jones, and Kelley all began with logical, rational analyses of what a reasonable perceiver would do. But researchers quickly uncovered a host of biases, reproducible ways in which actual judgments about persons departed from those predicted by rational models. Biases included actor–observer differences in attribution (people make systematically different attributions for their own behaviors than for those of others); false consensus bias (people assume that others act the same way that they themselves do); and negativity bias (people make more extreme attributions for negative than for positive behaviors). Most important was the correspondence bias: people tend to attribute traits to an actor that correspond to an observed behavior even when obvious and effective situational causes induce the behavior.

Research in this period maintained the assumption implicitly that traits constitute the core of person impressions, by investigating the processes by which perceivers infer traits based on behavioral evidence. As in the previous period, there was little attention to the organization of impressions, which were treated as unorganized lists of traits.

1.3 Late 1970s Through the 1990s

Cognitive psychology emerged in the 1960s and 1970s with the fall of behaviorism and the rise of the information-processing perspective (see *Cognitive Psychology: Overview*). As it became respectable throughout psychology to consider mental representations and processes as causes of behavior, the impact on social psychology was dramatic. Pioneering 'social cognition' researchers borrowed from cognitive psychology both theoretical and methodological tools to apply to social psychological phenomena. Central among the borrowings were associative models of memory (embodying the assumption that mental representations were formed by linking atomic elements into larger structures), schema theories (describing how organized memory representations guide attention and interpretation of new information), and response time measurement methods (which shed light on mental representations and processes by timing people's responses to specific types of questions). Research driven by these trends flourished, with the appearance of new journals and conferences devoted to social cognition. In a continuation of the attributional work of the earlier 1970s, a central focus of the early social cognition researchers was on person perception. This newer work had several major themes.

1.3.1 Changing conceptions of bias. Though a concern with 'biases' in judgment continued, bias was no longer defined purely in operational terms as a deviation from the predictions of some rational model of judgment. Instead the focus was on identifying the underlying processes that resulted in those judgmental patterns. For example, researchers examined effects of accessibility (the ease with which mental representations can be activated and used, based on the recency and frequency of prior use) on judgments and behaviors (Higgins et al. 1977). They also investigated the types of 'heuristic' processing (using simple rules of thumb) that occur when people are unable or unwilling to exert much effort (Chaiken 1980) (see *Heuristics in Social Cognition*). Processes like these turned out to account for many phenotypically distinct 'biases' that had previously been conceptualized and investigated in isolation from one another.

1.3.2 Focus on trait–behavior relations. Workers began to examine the organization of impressions, particularly the ways behaviors and related traits were linked together in memory representations. Drawing on associative models of memory, researchers such as Reid Hastie, Thomas Srull, and David Hamilton formulated models of person memory. Hastie (1980) and Srull independently developed similar models of the associative structures created when perceivers encounter inconsistent information about a target person—say, information that an individual who performs mostly honest acts also performs a few dishonest behaviors. These models successfully accounted for several types of empirical observations, including the fact that people tend to recall a higher proportion of the unexpected behaviors than of the expected ones.

1.3.3 Role of general knowledge structures. Researchers in this period examined the role of more general knowledge structures, particularly stereotypes (representations of social groups, such as gender, racial, or occupational groups), in the construction of person impressions (see *Stereotypes, Social Psychology of*). In some circumstances, a discrete impression of an individual group member is not formed at all, the person being mentally represented as simply an interchangeable member of the group. Even when an individual impression is formed, it is often greatly influenced by the perceiver's stereotype of the group as a whole. Research by Susan Fiske, Marilynn Brewer (1988), and others investigated the informational and motivational conditions under which stereotypes have this effect. Stereotypes were found to be activated automatically, and to affect impressions of individual group members as a 'default' condition unless perceivers are particularly moti vated (e.g., by a conscious desire to avoid using stereotypes) and cognitively able (e.g., through freedom from time pressure or distraction) to overcome these effects.

1.3.4 Diverse information included in impressions. Research on group stereotypes and on the role of behaviors in impressions made it clear that representations of persons typically include more than just traits. Group membership, specific behaviors, and the perceiver's emotional reactions to the target person are often part of person impressions (Carlston 1994), and researchers are beginning to examine how these multiple forms of information are interrelated.

1.3.5 Renewed attention to accuracy. New methodological developments in the 1980s allowed person perception researchers to revisit the questions of accuracy that had been dormant since Cronbach's critique in the 1950s. Work by David Kenny (1994) and others showed how to separate out effects on accuracy due to theoretically irrelevant factors (such as perceivers' tendencies to use different parts of rating

scales) and focus on more meaningful questions. Research using these techniques showed, for example, that observers can be remarkably accurate in judging certain personality attributes (such as extroversion) based on extremely limited information, such as a video clip of the target person lasting just a few seconds.

2. Current Themes and Future Directions

2.1 Close Connections with Research on Memory

Through the whole half-century covered in this article, assumptions regarding mental representations of persons have loosely tracked assumptions regarding memory representations in general. When memory researchers studied learning lists of words, person impressions were considered to involve lists of traits. When memory researchers postulated associatively linked structures, person impressions were conceptualized as linked complexes of traits and behaviors. When memory researchers studied large-scale, organized knowledge structures such as schemas, person perception researchers invoked organized stereotypic knowledge about social groups.

Recent trends in the study of memory include (a) an emphasis on specific (episodic or exemplar) knowledge, displacing the earlier focus on general schematic knowledge, and (b) investigations of connectionist (or parallel distributed) representations as alternatives to the traditional symbolic representations. Both of these trends have been applied to representations of persons (Smith and Zarate 1992, Read and Miller 1999). Studies testing such newer models of representation have often looked very like studies from a cognitive laboratory, with stripped-down tasks, priming manipulations, and response time measurements. These methods have sparked concerns about external validity, as observers have wondered what implications tiny response-time differences, for example, might have for real-world person perception. Yet the methods have offered powerful tools for investigating underlying representations and processes, even those (often termed 'implicit' knowledge) that the perceivers cannot consciously access (Greenwald et al. 1998).

2.2 Likely Directions for Future Research

Ongoing and future research on person representations seems likely to place a greater emphasis on the social and interpersonal context of person perception.

2.2.1 The self. Since the work of Daryl Bem in the 1960s, social psychologists have assumed that self-perception and perception of other people are closely related. Recent advances in models of person representations have correspondingly been applied to the self. For example, there have been studies of implicit self-esteem and of the way people organize positively and negatively valenced self-knowledge into discrete self-aspects, and their implications for the individual's personal and social functioning.

2.2.2 Linkages of other persons and groups to the self. Henri Tajfel demonstrated that representations of social groups to which an individual belongs become part of the person's self-representation. Similarly, Arthur Aron argued that in a close relationship, mental representations of the partner and the self become linked. Thus, representations of other persons and social groups can in effect become part of the self, with direct implications for self-regulatory processes (including emotional responses) and behavior including social influence, cooperation, and intergroup relations.

2.2.3 Impression change. Remarkably little study has been given to impression change. Yet in everyday life our impressions of others do change from time to time, and Bernadette Park and other researchers are now addressing this issue rather than continuing to focus on the initial formation of impressions of strangers.

2.2.4 Beyond verbal stimuli. In the first two decades, person impression research relied largely on lists of trait words, and since then, on written descriptions of behaviors. Researchers are now beginning to understand the limitations of verbal stimulus materials, which introduce extraneous issues such as the communicator's intention, the specific choice of words, etc. Some studies today use photographs, audio recordings, or video materials as stimuli, though verbal materials are still the overwhelming majority due to their ease of use and the possibility of close experimental control.

The methodological and conceptual shifts over the half-century of work on mental representations of persons should not obscure an underlying continuity. The same idea that motivated Asch's research still stands: perceivers construct representations of other people based on behaviors they observe, group memberships, and other types of information. The perceivers then draw on those representations as they make judgments about others, decide to form relationships with them or to avoid them, accept or resist social influence from them, and treat them with justice and altruism, or with prejudice and discrimination.

Representations of persons thereby play a crucial mediating role in virtually all forms of social behavior.

See also: Attitude Formation: Function and Structure; Attitudes and Behavior; Concept Learning and Representation: Models; Feature Representations in Cognitive Psychology; Social Categorization, Psychology of; Stereotypes, Social Psychology of

Bibliography

Asch S E 1946 Forming impressions of personality. *Journal of Abnormal Social Psychology* **41**: 258–90

Brewer M B 1988 A dual process model of impression formation. In: Srull T K, Wyer R S (eds.) *Advances in Social Cognition*. Lawrence Erlbaum Associates, Hillsdale, NJ, Vol. 1, pp. 1–36

Carlston, D E 1994 Associated Systems Theory: A systematic approach to cognitive representations of persons. In: Wyer R S (ed.) *Advances in Social Cognition*. Lawrence Erlbaum Associates, Hillsdale, NJ, Vol. 7, pp. 1–78

Chaiken S 1980 Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personal and Social Psychology* **39**: 752–66

Cronbach L J 1955 Processes affecting scores on 'understanding of others' and 'assumed similarity'. *Psychology Bulletin* **52**: 177–93

Greenwald A G, McGhee D E, Schwartz J L K 1998 Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personal and Social Psychology* **74**: 1464–80

Hastie R 1980 Memory for information which confirms or contradicts a general impression. In: Hastie R, Ostrom T M, Ebbesen E B, Wyer R S, Hamilton D L, Carlston D E (eds.) *Person Memory*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 155–77

Heider F 1958 *The Psychology of Interpersonal Relations*. Wiley, New York

Higgins E T, Rholes W S, Jones C R 1977 Category accessibility and impression formation. *Journal of Experimental and Social Psychology* **13**: 141–54

Kenny D A 1994 *Interpersonal Perception*. Guilford Press, New York

Read S J, Miller L C 1999 *Connectionist Models of Social Reasoning and Social Behavior*. Lawrence Erlbaum Associates, Mahwah NJ

Smith E R, Zarate M A 1992 Exemplar-based model of social judgment. *Psychological Review* **99**: 3–21

E. R. Smith

Mental Representations, Psychology of

Mental representations were banned from scientific psychology by the behaviorists. They came back into psychology during the so-called cognitive revolution, when information processing models came to dominate psychological theorizing.

They were banned by the behaviorists for two reasons. First, they are not directly observable; they must be inferred from their observable behavioral consequences. Radical behaviorists believed that inferred entities had no valid role to play in a scientific psychology (Skinner 1938, 1950, 1990). Second, mental representations are not neurobiologically transparent: it has been and remains difficult to say how the entities and processes central to many kinds of hypothesized mental representations might be realized by currently understood neurobiological processes and structures. Not surprisingly, efforts to eliminate mental representations from psychological theorizing have often been driven by a desire to anchor psychological theorizing in neurobiology. (See, for example, Edelman and Tononi 2000, Hull, 1930, 1952, Rumelhart and McClelland 1986.)

Another difficulty is that it has not always been clear what cognitive psychologists understood by the term representation. This lack of clarity is due to the inherent complexity and abstraction of the concept. Although mental representations are central to prescientific folk psychology, folk psychology does not provide a rigorous definition of representation, any more than folk physics provides a rigorous definition of mass and energy. Representation, rigorously defined, is a mathematical and computational concept.

The cognitive revolution was closely tied to the emergence of computer science because computer science created indubitably physical machines that unequivocally computed. This dispelled the widespread belief that computing was an inherently mental activity in the dualistic sense—mental and therefore not physical. More importantly, computer science led to a deeper understanding of what it meant—from a physical and mathematical perspective—to say that something computed (Turing 1936). Computation became an object of mathematical thought rather than merely a tool of such thought.

A representation, mental or otherwise, is a system of symbols. The system of symbols is isomorphic to another system (the represented system) so that conclusions drawn through the processing of the symbols in the representing system constitute valid inferences about the represented system. Isomorphic means 'having the same form.' The form in question is mathematical form, the forms of the equations specifying the relations among the symbols and among the things that the symbols represent. For example, Ohm's law—$I = V/R$—which is the equation for the relation between current (I), voltage (V), and resistance (R) in an electrical circuit, has the same form as the equation for the relation between speed (S), force (F), and viscous resistance (R) in a mechanical system like a shock absorber—$S = F/R$. The identical form of these two equations is suggestive of the much broader isomorphism (mathematical equivalence) between electrical, mechanical, hydraulic, and thermal systems that gives rise to linear systems theory in engineering.

The symbols in the above two equations differ, but that is only to remind us that the physical variables they refer to differ. The important thing is that the equations that describe the two systems are the same. Because the forms of the relations the variables enter into are the same, we can represent a mechanical system with an electrical system (and vice versa). And we can represent either of them with a paper and pencil system that we endow with a suitable mathematical form. What matters in representations is form, not substance.

The symbols in an information processing system (a symbol system) have two fundamental properties: they refer to things outside the system and they enter into symbol processing operations. The symbol processing operations in the above examples are the operations of arithmetic (V divided by R) and the rewrite rules (rules of algebra) dictated by the principles that define the system of arithmetic operations.

We believe that we understand simple electrical circuits because the inferences we draw from manipulating the symbols on paper correctly predict what we observe when we make the corresponding manipulations of the electrical circuit itself. Thus, for example, simple algebra allows us to deduce from $I = V/R$ that $IR = V$. When we measure I and R and compute the numerical product of the two measurements, the number we get turns out to be the same number that we get when we measure V. Our paper and pencil representation of the electrical circuit, which includes both the symbols themselves and the rewrite rules that we observe in deriving $IR = V$ from $I = V/R$, correctly predicts the results of the measurements that we make on the circuit itself.

The above example of a symbolic system contains three distinct contrivances—symbols, rules that govern the manipulation of those symbols, and measuring processes. The measuring processes relate the numerical values of the symbols to the voltages, resistances, and currents to which they refer. Because these are obviously human contrivances, it might seem that representations are artifacts of a purely human manner of interacting with the world, requiring perhaps some form of consciousness. However, the same three contrivances are present in a process control computer. Such a computer is also a human contrivance, but it interacts with the world without human intervention. It measures physical variables using digitizing transducers, symbolizes those variables by means of bit patterns in its memory banks, manipulates those symbols in accord with the applicable rules of algebra and physics, and uses the results to control observable actions—all without human intervention.

The cognitive revolution was predicated on the possibility that brains—both human and animal—are process control computers contrived by evolution through natural selection. They assess their environment through sensory or perceptual processes; they symbolize the results of these assessments by values stored in memory; they manipulate those values by means of the relevant mental operations (the operations of perception and thought); and they use the results (percepts, inferences, and deductions) to control behavior. If so, then mental representations are the very stuff of psychology. A psychology without mental representations is no more possible than a physics without masses and energies.

Thus, from the standpoint of a cognitivist, psychology is the science of mental representations. The essential questions in psychology are: What representations does the mind compute? From what data does it compute them? How does it compute them? How does a given representation get translated into observable behavior?

1. *Information*

An important development in the mathematical treatment of computation and representation was the rigorous definition and quantification of the information carried by signals (Shannon 1948). Signals are symbols that convey information from one location in space and time to another—like, for example, the nerve signals that carry information about the environment from sensors in the periphery to the brain. The amount of information conveyed by a signal is a function of the amount of information about the world already present at the site where the signal is received and processed. When a digitizing thermometer sends a bit pattern to a computer specifying the temperature of a fluid, the signal, that is, the transmitted bit pattern, conveys information about the environment to the computer. The less information the computer already has about the temperature, and the more precisely the bit pattern received specifies what the temperature is, the more information is conveyed by the signal. (See Rieke et al. 1997 for the rigorous development of these ideas in the analysis of neural signaling.)

This idea that the amount of information conveyed by a signal is measured by the amount by which the signal reduces the receiver's uncertainty about the state of the world is highly intuitive: If we already know that the temperature is 70°, then a signal indicating that it is 70° tells us nothing. This simple idea has, however, nonintuitive mathematical consequences. It implies, for example, that signaling presupposes prior knowledge on the part of the receiver regarding the range of possibilities. If the mind of the newborn baby is truly a blank slate, with no beliefs about what temperature its environment might have, then any signal that tells it what the temperature is—even if imprecisely—conveys an infinite amount of information. In information theory, no signal can convey an infinite amount of information in a finite amount of time. Thus, in order for us to acquire information about the world from our experience of it,

we must have built into our information processing structures implicit representations of the range of environments that could be encountered. We must know in advance something about the world we are to experience.

This implication of information theory, together with the consideration that the machinery of computation itself seems unlikely to arise from the impact of experience on a system not endowed with some initial computational capacity, gives a nativist cast to information processing theories of mind. If the brain is fundamentally an organ of computation devoted to the computation of the mental representations that enter into the decisions leading to actions, then it does not seem that it could get up and running without a non-trivial amount of genetically specified structure, much of which contains implicit knowledge about the world to be represented. That is why extreme empiricists tend to be anti-representational: they tend to reject the cognitivist assumption that the mental representations are the stuff of psychology.

2. *Decision Processes*

Symbols are translated into observable behavior by means of control variables and decision processes. A control variable specifies a parameter of an action, for example, the angle that is to be maintained with respect to a directional reference stimulus like the sun (see illustrative example below). A decision variable is a computed symbolic value representing some aspect of the current environment that merits a response just in case it exceeds some criterion, called the decision criterion. The analysis of decision processes in modern psychology has been heavily influenced by statistical decision theory, which treats the structural and formal features of decisions made in the face of ambiguous information (Green and Swets 1966). The information about the world carried by symbols is ambiguous for two reasons. First, the processes that generate the symbolic values are inherently and inescapably noisy. The temperature of the fluid cannot be exactly known and hence it cannot be known with certainty whether an environmental variable actually does exceed some criterion; it can only be known with varying degrees of probability. Thus, decisions are inherently statistical in nature. Optimal decision processes must take account of the statistical uncertainty about the true value of critical variables. Second, one and the same sensory input may be generated by more than one state of the world. For example, radically different arrangements of surfaces in the three-dimensional environment can produce identical images when projected on to the two-dimensional retina. In computing a representation of the three-dimensional environment that generated these inputs, the brain must be sensitive to the relative likelihoods of various three-dimensional configurations, given the two-dimensional projection that has been sensed. The analysis of optimal decision-making under these conditions brings in another aspect of statistical decision theory, Bayesian inference (Knill and Richards 1996).

3. *Illustrative Examples*

The development of mathematical analyses of information processing and decision making inspired a psychology focused on mental representations. What has sustained it are the many examples of human and animal behavior that imply an underlying representation. Some of the simplest and most illuminating examples are found in learned behavior in nonhuman animals that depend on underlying representations of abstract but basic properties of the world like distance, direction, duration, and time of day (phase of the day–night cycle).

A honeybee forager, when it returns to the hive after discovering or revisiting a source of rich nectar, does a dance that symbolizes the solar bearing (direction relative to the sun) and distance of the source from the hive. Foragers that have followed it while it danced later leave the hive and fly in the indicated direction for the indicated distance before they begin to look for the source. Because the dance directly symbolizes direction and distance and because the witnesses to the dance base their own flight directions and distances on what they have observed, it seems inescapable that the direction and distance of the source must be represented in the system that controls bee behavior, the bee brain. In this case, the representational nature of mental processes is manifest in a behavior that is itself representational. (See Gallistel 1998 for a recent review of insect navigation and bee dancing, emphasizing the information processing implications.)

The dance is in the form of a figure eight. It is performed on the vertical surface of the interior of the hive out of sight of the sun. When running the middle bar of the eight (the part common to the two circles), the dancing bee waggles rapidly from side to side. The angle of this waggle run with respect to the vertical symbolizes the direction of the source relative to the sun, while the number of waggles symbolizes the distance.

The angle of the waggle run with respect to the vertical changes during the day so as to take into account the changing direction of the sun, even under conditions where the dancers have not seen the sun or anything else in the sky that indicates where the sun is for hours or even days. Both the dancer and its audience are able to represent compass direction (direction relative to the earth's surface) by reference to the sun's compass direction, first, because they have learned the solar ephemeris, the compass direction of the sun as a function of the time of day, and, second, because they possess a circadian clock. The circadian

clock symbolizes the time of day. It is a cyclical molecular process within nerve cells (Gekakis et al. 1998, Sehgal 1995), with approximately the same period as the day–night cycle that is synchronized to the sun's cycle every dawn and dusk by signals coming from photoreceptors. Because this biochemical cycle within cells is synchronized with the day–night cycle, phases within this biochemical cycle—the momentary concentrations of the different molecules whose concentration varies cyclically—indicate the phase of the earth's rotational cycle, that is, the time of day.

The solar bearing symbolized by the direction of the waggle run is computed from the representation of two different aspects of the bee's previous experience. One set of experiences are those from which it learns the solar ephemeris (Dyer and Dickinson 1996). The other is the foraging experience from which it learns the compass direction of the source from the hive. The solar bearing is the angular difference between the compass direction of the source from the hive and the current direction of sun (as given by the solar ephemeris).

There does not appear to be a way to account for the bee's behavior without endowing its brain with the capacity to symbolize the time of day, compass direction, and distance. It has also to have the capacity to learn functions like the solar ephemeris. A function is a set of paired symbols, an input symbol and an output symbol. The input symbol in the solar ephemeris represents the time of day, while the output symbol represents the compass direction of the sun. A function may be realized by means of a look-up table, which stores the possible pairs of input and output symbols, but this can make large demands on memory. Alternatively, a function may be generated by a neuronal process that transforms an input signal into an output signal. In that case, the relation between the input and the output of this process must have the same mathematical form as the solar ephemeris itself.

Finally, the bee brain must be able to compute an angular difference, a symbol representing the difference between the compass direction given by its solar ephemeris function and the compass direction of the source. This latter symbol is retrieved when needed from the memory generated at the time the bee found the source. The result of this computation, the symbol representing the angular difference between the directions represented by two other symbols, represents the solar bearing of the source. It is this angle that we observe when a dancer makes its waggle run. A psychology focused on mental representations rests on the claim that there is no way to explain this robust and reliable fact about bee behavior except by an appeal to the kind of information processing just described.

A second example of the fundamental role that information processing plays in the control of behavior comes from the extensive studies of conditioned behavior in the common laboratory animals—the rat, the pigeon, and the rabbit. In Pavlovian conditioning, the experimenter repeatedly presents temporally paired elementary stimuli. For example, using rabbits as subjects, the experimenter may repeatedly present a tone followed at a short latency by an annoying puff of air directed at the sclera of the eye or an annoying shock to the skin around the eye. The tone is called a conditioned stimulus (CS), because it elicits observable behavior only after conditioning, while the puff or shock is called an unconditioned stimulus (US), because it elicits observable behavior in the absence of any conditioning. When a US has reliably followed a CS, the subject responds to the CS in anticipation of the US. In the present example, the rabbit blinks when it hears the tone. This blink is called the conditioned response. It is so timed that the moment of peak closure more or less coincides with the moment when the US is expected. If the US sometimes comes at a latency of 0.4 seconds and sometimes at a latency of 0.9 seconds, the rabbit learns to blink twice, with the first blink peaking at about 0.4 seconds and the second at about 0.9 seconds (Kehoe et al. 1989).

Evidently, the rabbit measures and remembers the durations of the intervals between the onsets of the tone and the onsets of the US. How else can we explain the fact that it matches the latency of its response to the latency of the US? The rabbit must possess a memory like the memory that Alan Turing (1936) placed at the heart of his mathematical abstraction of a computing device, the so-called Turing machine. This notional machine has a memory to which it writes and from which it reads symbols. If the rabbit did not have a memory in which it could store a symbol representing the CS–US latency and from which it could subsequently retrieve that symbol, its ability to match its conditioned response to that latency would be inexplicable.

It is a general property of conditioned behavior that the latency of the conditioned response is proportional to the CS–US latency (Gallistel and Gibbon 2000). Moreover, from the nature of the variability in conditioned response latencies, it appears that the decision about when to make a conditioned response following the onset of a CS must be based on the ratio between the remembered CS–US interval and the interval elapsed since the onset of the current CS (Gibbon et al. 1984). Thus, when the tone sounds, the rabbit retrieves from memory a symbolic value representing the CS–US interval, measures the time elapsed since the onset of the tone, to generate a constantly growing signal whose momentary magnitude represents the duration of the currently elapsed interval, computes the ratio of the two values, and responds when the ratio exceeds a critical value.

Just as bees can compute an angular difference from directions (compass angles) stored in memory, so rats can compute a temporal difference from durations stored in memory, as shown by experiments using what is called backward conditioning. In a backward

conditioning experiment, the US precedes the CS. For example, a tone CS comes on 1 second after a shock US ends. Under these conditions, subjects respond weakly or not at all to the tone, because it no longer gives advanced warning of the US. Although they do not respond to the tone, they learn the (negative) interval between it and the shock. This is shown by also teaching them a forward temporal relation between a light and the tone. When they have also been taught that the onset of the light predicts the onset of the tone after a latency of 5 seconds, then they respond strongly to the light (Barnet et al. 1997). From their representation of the tone–shock interval ($= -1$ second) and their representation of the light–tone interval ($= +5$ seconds), they appear to have computed the expected light–shock interval (4 seconds). Consequently, they react fearfully to the light, even though it has never been followed by shock. Its only connection to shock is by way of the tone, but they do not react fearfully to the tone itself, because it has always followed the shock. The predictive relation of the light to the shock has been inferred by computations performed with the symbols that represent the two durations.

As these illustrative examples show, animals are able to function effectively in a complex world because their brains construct mental representations of behaviorally important aspects of that world—spatial relations, temporal relations, numerical relations, social relations, and so on. Wherever there is regularity and form in the world, animals represent that regularity and that form in order to exploit it for their own ends. The most basic mechanism of life itself—the genetic mechanism—is a mechanism for copying, transmitting, and processing information. A significant fraction of that information specifies the immensely complex structure of the brain, an organ dedicated to the processing of information about the animal's environment.

See also: Concept Learning and Representation: Models; Feature Representations in Cognitive Psychology; Knowledge Representation; Neural Representations of Direction (Head Direction Cells); Neural Representations of Objects; Propositional Representations in Psychology; Reasoning with Mental Models; Reference and Representation: Philosophical Aspects

Bibliography

Barnet R C, Cole R P, Miller R R 1997 Temporal integration in second-order conditioning and sensory preconditioning. *Animal Learning and Behavior* **25**: 2 221–33

Dyer F, Dickinson J A 1996 Sun-compass learning in insects: representation in a simple mind. *Current Direction in Psychological Science* **53**: 67–72

Edelman G, Tononi G 2000 *A Universe of Consciousness: How Matter Becomes Imagination*. Basic Books/Allen Lane, New York

Gallistel C R 1998 Brains as symbol processors: the case of insect navigation. In: Sternberg S, Scarborough D (eds.) *Conceptual and Methodological Foundations*. Vol. 4 of *An Invitation to Cognitive Science*, 2nd edn. MIT Press, Cambridge, MA, pp. 1–51

Gallistel C R, Gibbon J 2000 Time, rate and conditioning. *Psychological Review* **107**: 289–344

Gekakis N, Staknis D, Nguyen H B, Davis F C, Wilsbacher L D, King D P, Takahashi J S, Weitz C J 1998 Role of the CLOCK protein in the mammalian circadian mechanism. *Science* **280**: 1564–70

Gibbon J, Church R M, Meck W H 1984 Scalar timing in memory. In: Gibbon J, Allan L (eds.) *Timing and Time Perception*. New York Academy of Sciences, New York, Vol. 423, pp. 52–77

Green D M, Swets J A 1966 *Signal Detection Theory and Psychophysics*. Wiley and Sons, New York

Hull C L 1930 Knowledge and purpose as habit mechanisms. *Psychological Review* **37**: 511–25

Hull C L 1952 *A Behavior System*. Yale University Press, New Haven, CT

Kehoe E J, Graham-Clarke P, Schreurs B G 1989 Temporal patterns of the rabbit's nictitating membrane response to compound and component stimuli under mixed CS-US intervals. *Behavioral Neuroscience* **103**: 283–95

Knill D, Richards W (eds.) 1996 *Perception as Bayesian Inference*. Cambridge University Press, New York

Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W 1997 *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA

Rumelhart D E, McClelland J L (eds.) 1986 *Parallel Distributed Processing*. MIT Press, Cambridge, MA

Sehgal A 1995 Molecular genetic analysis of circadian rhythms in vertebrates and invertebrates. *Current Opinion in Neurobiology* **56**: 824–31

Shannon C E 1948 A mathematical theory of communication. *Bell Systems Technical Journal* **27**: 379–423, 623–56

Skinner B F 1938 *The Behavior of Organisms*. Appleton-Century-Crofts, New York

Skinner B F 1950 Are theories of learning necessary? *Psychological Review* **57**: 193–216

Skinner B F 1990 Can psychology be a science of mind? *American Psychologist* **45**: 1206–10

Turing A M 1936 On computable numbers, with an application to the *Entscheidungs* problem. *Proceedings of the London Mathematical Society* 2nd series **42**: 230–65

C. R. Gallistel

Mental Retardation: Clinical Aspects

Recognized throughout history, *mental retardation* (also referred to by other terms) is a condition of substantial limitations in various aspects of present intellectual functioning that affects one's performance in daily living. Mental retardation begins in childhood

(birth to age 18) and is characterized by co-existing limitations in intelligence and adaptive skills. The definition and its clinical diagnosis, system of classification, and etiological assessment have changed repeatedly in the past century, influencing prevalence estimates, services and social policy, and intervention approaches. This article describes (a) the interrelationships between the supports people with mental retardation require to function well and (b) the role clinical services play to enable these individuals to create personally satisfying lives for themselves.

1. Concept of Support

One common theme in current conceptualizations of mental retardation has been the need individuals with this disability have for support. *Supports* which are resources and strategies that promote the interests and causes of individuals with or without disabilities, (Luckasson et al. 1992) require individualization for several reasons. First, a person's capacity is influenced by their intelligence, adaptive skills, physical and mental health, and the disability's etiology. Second, the extent and types of limitations and capabilities exhibited by individuals with mental retardation are influenced contextually by chronological age, community setting, and culture. Mental retardation refers to 'a specific pattern of intellectual limitation, not a state of global incompetence' (Luckasson et al. 1992). Third, a person's support needs are balanced with their capabilities, with the demands made upon the individual by the environments used, and with the individual's resultant functioning (Fig. 1). *Capabilities* include intellectually dependent competencies such as social, practical, and academic skills. *Environments* encompass the home and school but extend to the community and work setting with advancing age. *Functioning* refers to how one copes with the ordinary challenges of routine life; the significance of one's limitations in functioning is influenced by the demands made and the supports available in an environment. The intensities and types of support necessary for individuals with mental retardation relate to the interaction between an individual's capacity and the conditions or requirements of the environments used, and the acceptability of functioning that results. This support triangle conceptualization represents a shift in thinking from a clinical-psychological model of disability to a contextual or social-cultural model.

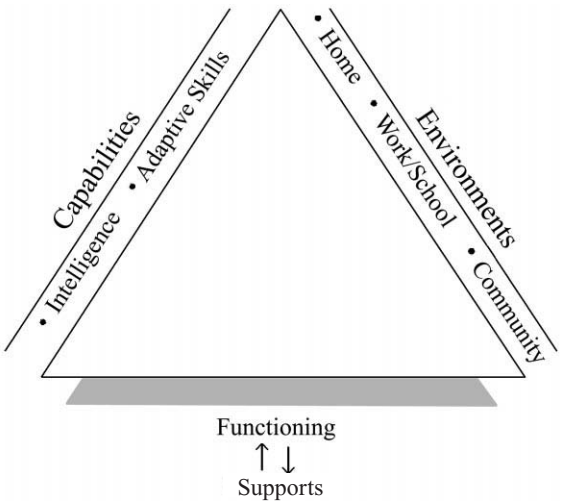

Figure 1
General structure of the definition of mental retardation (from Luckasson et al. 1992)

Four assumptions are central to the application of this definition of mental retardation (Luckasson et al. 1992):

(a) valid assessment considers cultural and linguistic diversity as well as differences in communication and behavioral factors;

(b) the existence of limitations in adaptive skills occurs within the context of community environments typical of the individual's age peers and is indexed to the person's individualized needs for supports;

(c) specific adaptive limitations often coexist with strengths in other adaptive skills or other personal capabilities; and

(d) with appropriate supports over a sustained period, the life functioning of the person with mental retardation will generally improve.

2. Diagnosis

Given the framework of these assumptions, several evaluation phases are required to diagnose mental retardation. First, the person's IQ is tested. Second, their adaptive skills are measured, identifying the strengths and limitations in each applicable area of adaptive skill. Third, the age at which the suspected disability first occurred is determined. Finally, an individual's performance on these assessments and the age of disability onset are evaluated against the three criteria required for a diagnosis of mental retardation:

(a) intellectual functioning is significantly below average: an IQ approximately 70 to 75 or less;

(b) consequential limitations exist in several or many crucial areas of adaptive skill: two or more of 10 applicable areas including communication, self-care, home living, social skills, community use, self-direction, health and safety, functional academics, leisure, and work; and

(c) the onset of the suspected disability occurs during the developmental period: before age 18.

In most instances, a diagnosis of mental retardation is preceded by observable, early delays in cognitive, social, and communication development. While some etiologies can be associated with a wide range of

capability, others are more universally accompanied by mental retardation (e.g., chromosomal disorders like Down's syndrome, Angelman syndrome, and trisomy 13 syndrome).

3. *Procedure for Profiling Supports*

When a social-cultural model is used to conceptualize mental retardation, clinicians aim not only to make a diagnosis or a determination of whether mental retardation exists, but also to describe the resources and services that person will require to function satisfactorily during a given period of life. This second aim, referred to as an *individualized profile of supports*, is a complex task requiring input from a team of professionals, family members, and whenever possible, the focus individual. Because a person's support needs change with time, support profiles must be reexamined regularly and whenever significant changes are predicted or occur.

Like those without mental retardation, the lives of individuals with mental retardation have multiple *dimensions*: (a) capacity or ability (i.e., intellectual functioning and adaptive skills), (b) psychological or emotional well-being, (c) physical characteristics including health and the etiology of the disability, and (d) environmental influences of home, school, employment, and daily routines. For those who meet the diagnostic criteria for mental retardation, their team members proceed through several steps to construct a profile of supports across these life dimensions. Including diagnosis, this process consists of six steps:

(a) the individual's strengths and weaknesses in intellectual and adaptive skills are identified;

(b) if significant limitations exist, the individual is assessed against the criteria for intellectual functioning and age of onset and a diagnosis for mental retardation is made or not made;

(c) the individual's strengths and weaknesses in psychological and emotional wellbeing are identified;

(d) the individual's strengths and weaknesses in physical characteristics, health, and etiological considerations of the disability are determined;

(e) assessment is made of the environmental influences on the individual to determine strengths and weaknesses; and

(f) the types and intensities of support are identified that will onset any limitations across the four dimensions.

4. *Classification*

After defining mental retardation and determining who has the disability and who does not, additional systems of classification may have value. The modern view is to classify according to a system of intensities of needed supports. In the 1992 American Association

Table 1
Supports intensity decision grid

	Intermittent	Limited	Extensive	Pervasive
Time duration	As needed	Time limited, occasionally ongoing	Usually ongoing	Possibly lifelong
Time frequency	Infrequent, low occurrence	Regular, anticipated, could be high frequency		High rate, continuous, constant
Settings: living, work, recreation, leisure, health, community, etc	Few settings, typically one or two settings	Across several settings, typically not all settings		All or nearly all settings
Resources: professional/ technological assistance	Occasional consultation or discussion, ordinary appointment schedule, occasional monitoring	Occasional contact, or time limited but frequent regular contact	Regular, ongoing contact or monitoring by professionals typically at least weekly	Constant contact and monitoring by professionals
Intrusiveness	Predominantly all natural supports, high degree of choice and autonomy	Mixture of service-based supports, lesser degree of choice and autonomy		Predominantly service-based supports, controlled by others

Source: Luckasson et al. (1996).

for Mental Retardation (AAMR) system, for example, the individual's support needs in each applicable adaptive skill area of the capability dimension (and in the other dimensions as well) are classified as being Intermittent, Limited, Extensive, or Pervasive. This is sometimes referred to as the ILEP classification system.

Determining the intensities of an individual's support needs constitutes a team activity but also requires clinical judgment. There are no bright lines between each of the four intensities. However, consideration of five factors contributes to decisions about support intensity: time-duration (how long a support is needed); time-frequency (how often a support is required); settings in which the supports are needed; resources required for the supports (cost, personnel, expertise, etc.); and the degree of intrusiveness in one's life. A decision grid of the five factors and the four intensities may contribute to team planning with the individual (Luckasson et al. 1996) (see Table 1).

An older classification system categorized individuals with mental retardation by IQ ranges rather than classifying their needs for supports. This system was last described in the 1983 AAMR manual (Grossman 1983) (not a part of the current manual) and was retained as an option in *DSM-IV* (American Psychiatric Association 1994) (see *Mental and Behavioral Disorders, Diagnosis and Classification of*). The system used four to five IQ ranges: mild (IQ 50–55 to 70–75), moderate (IQ 35–40 to 50–55), severe (IQ 20–25 to 35–40), profound (IQ below 20–25), and sometimes borderline, a range between about 70–75 and 85 (technically a score outside the category of mental retardation but nevertheless a score indicating cognitive impairments). This system of classifying according to IQ ranges may have seemed more useful in earlier times when the IQ scores of individuals who did not receive education or who resided in large, isolated institutions were often predictive of functioning or placements. Today in many countries, however, with universal access to education, increasing access to supports, expectations that people will be part of their ordinary communities, and an understanding of the interaction between people and their environments, IQ scores alone are less predictive of functioning. Thus, a classification system that merely transforms the initial eligibility IQ score into a classification of mild, moderate, severe, or profound is less useful and even misleading. For example, to label those scoring in the relatively higher IQ range as 'mild' within a group already categorized as having significantly impaired cognitive functioning mischaracterizes the disability.

5. *The Lives of People with Mental Retardation*

All individuals with mental retardation, regardless of their age or IQ, should expect lives with personal meaning and opportunities to contribute to their families and communities. People with mental retardation will, of course, have different skills, motivations, temperaments, and opportunities. Mental retardation may affect how people find meaning in their lives but should not affect *whether* they find meaning.

Several principles should guide public policies, services, and supports for people with mental retardation:

(a) normalization: making available the normal opportunities of a society;

(b) inclusion: assuring that the person has the opportunity and support to be part of society; and

(c) choice: supporting the individual's personal decision-making.

In this section, we consider personal meaning and the guiding principles in the lives of people with mental retardation at four stages: early childhood, school years, adulthood, and older years.

For infants and young children, the primary concern of society should be supporting their families' ability to raise the child in a loving home and assuring that early intervention services are provided. Depending on the infant's or child's needs, services may include medical care, assistive technology, health supports, family supports, physical, occupational, and communication therapy, and early childhood education.

For school-aged children, the primary concern should be providing meaningful access to education for *all* students, including students with mental retardation. The goal is to provide students with individually functional skills that will contribute to a satisfying adulthood. In the US, the Individuals with Disabilities Education Act (IDEA) guarantees that every child with a disability, ages 3 through 22, has a legal right to a free appropriate public education in the least restrictive setting with any related services necessary to benefit from education. Transition planning and services, beginning at age 14, are required to facilitate a student's entry to an adult role in society. Being included in schools and classrooms with peers without disabilities throughout their education appears to enhance both the abilities of students with mental retardation and nondisabled peers to learn from each other and establish lifelong relationships.

For adults with mental retardation, the primary concerns will be finding and keeping jobs, and establishing adult personal lives, including homes, families, and friendships. Financial assistance such as social security is critical in laying a financial base for individuals who otherwise will likely face extreme poverty. Job training and supports will also be critical to avoid the problems of unemployment and isolation. Similarly, antidiscrimination legislation such as the Americans with Disabilities Act (ADA) and Section 504 of the Rehabilitation Act of 1973 are important in fighting discriminatory denial of services and benefits.

Many people with mental retardation now live long lives, due to improved medical care and general health care improvements. For people in their older years, the

primary support concerns will be maintaining family and community ties and preserving health. Issues of how to spend retirement years in satisfying activities and relationships will build on earlier experiences with work, family, and personal choices.

6. *Continuing Controversies*

Several areas of continuing controversy effect the concept and assessment of mental retardation. First, there is disagreement on the soundness of adaptive behavior (AB) measures and a lack of consensus on the construct of adaptive competence. AB test results appear to overlap with factors measured by IQ tests. Second, concerns exist about the use of IQ measures with people who have mental retardation. While IQ measures generally yield stable results, they are criticized as being racially discriminating, non-accommodating to physical, sensory, or behavioral limitations, and impractical guides for interventionists. Some argue that intelligence should be conceptualized broadly as personal competence, and thus involve assessment of physical, affective, adaptive, and everyday competence. Beyond the assessment of IQ in people with disabilities, there is still widespread controversy over the relationships between IQ, culture, and genetics. Third, many US consumer and advocacy groups have rejected the term 'mental retardation' as stigmatizing and negative. This stigmatization may be at the root of the otherwise unexplainable reductions in many states of school-aged individuals with mental retardation labels and the corresponding increases in the number of individuals labeled with learning disabilities.

A final area of dispute involves the primacy of professionals over individuals with mental retardation in defining the label and its purpose and in determining needed supports. In recent years, many people with mental retardation have learned to advocate for themselves and have been influential in sensitizing professionals to (a) the value of person-centered planning, (b) the negative implications of the label, (c) the benefits of self-advocacy, and (d) the lowered expectations that result from a deficit orientation. Despite these trends, many people with mental retardation continue to be only minimally involved in decision-making about their own lives.

7. *Future Directions*

The future of mental retardation will likely be affected by several trends. First, the developed world is just now experiencing the first generation of young people who had universal access to special education and basic supports. These young people and their families have increased expectations of inclusion in their societies, and see themselves and their struggles in a more positive and political light than earlier generations who were neglected, deprived, and isolated because of their disabilities. Second, significant new neurological research on the brain functioning draws many old presumptions into question. It now appears that brain plasticity may extend over many years, that fetal surgery may prevent some brain conditions previously thought inevitable, and that many troubling behaviors have neurological bases and perhaps cures. These and other advances significantly change the prospects for many people with mental retardation.

Third, health-care providers face challenging questions daily about the rationing of health care, care for infants born with significant disabilities, and the value of life when a person has a disability. Societal doubts continue despite the increased demands of people with disabilities and their families for health care systems untainted by discrimination because of disability. It remains to be seen whether societies will genuinely commit to value all people with mental retardation, and to promote their sharing of all the benefits of the society or whether their tenuous place is protected only as long as they are not too disabled. Fourth, the construct of mental retardation is changing. It is changing from a purely scientific construct to incorporate a social construct, from an IQ-based definition to a functional, supports-based definition. These changes have implications both for assessing disability and for providing supports.

Finally, there is a growing acknowledgment of the forgotten generation, people with mild cognitive limitations who do not technically come within the traditional definition of mental retardation but whose daily functioning is limited by lowered intelligence. These individuals, because of their cognitive limitations, live at the margins of society, attempting to cope with poverty and unable to achieve adequate access to healthcare, social services, the justice system, or other benefits of society. They have received none of the specialized education, health care, or justice considerations because they failed the diagnostic criteria for mental retardation; but they have compelling needs. Perhaps societies will draw on lessons learned from addressing the needs of people with disabilities to assist the forgotten generation.

See also: Disability: Psychological and Social Aspects; Disability: Sociological Aspects; Intellectual Functioning, Assessment of; Intelligence: Central Conceptions and Psychometric Models; Mental Retardation: Cognitive Aspects; Special Education in the United States: Legal History

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*), 4th edn. APA, Washington, DC

Grossman H J (ed.) 1983 *Classification in Mental Retardation*. American Association on Mental Retardation, Washington, DC

Luckasson R, Coulter D, Polloway E A, Reiss S, Schalock RL, Snell M E, Spitalnik D M, Stark J A 1992 *Mental Retardation: Definition, Classification and Systems of Support*, 9th edn. American Association on Mental Retardation, Washington, DC

Luckasson R, Schalock R L, Snell M E, Spitalnik D 1996 The 1992 AAMR definition and preschool children: A response from the Committee on Terminology and Classification. *Mental Retardation* **34**: 247–53

Schalock R L, Stark J A, Snell M E, Coulter D L, Polloway E A, Luckasson R, Reiss S, Spitalnik D M 1994 The changing conception of mental retardation: Implications for the field. *Mental Retardation* **32**: 181–93

M. E. Snell and R. Luckasson

Mental Retardation: Cognitive Aspects

Standard definitions of mental retardation include three criteria: (a) significantly subaverage intelligence, (b) significant limitations in adaptive skills, (c) onset during the developmental period. Given criteria (a) and (b), it is clear that cognitive limitations must play a major role in mental retardation. Both the extent of overall cognitive difficulty and the particular pattern of cognitive problems vary as a function of etiology (e.g., Down syndrome, Williams syndrome, lead poisoning).

1. *Definitions of Mental Retardation*

The most commonly used definition of mental retardation is that provided by the fourth edition of the *Diagnostic and Statistical Manual* (*DSM-IV*; American Psychiatric Association 1994). According to this definition, individuals are considered to have mental retardation if:

(a) Their current IQ, based on an individually administered test, is at least 2 standard deviations below the mean (below about 70).

(b) They have significant limitations, relative to those expected for their chronological age (CA) and sociocultural background, in at least two of the following domains: communication, social/interpersonal skills, self-care, home living, self-direction, leisure, functional academic skills, use of community resources, work, health, and safety.

(c) These difficulties were first evidenced prior to age 18 years.

Each of the other major organizations involved in the treatment of individuals with mental retardation—the American Association on Mental Retardation (AAMR), the American Psychological Association (APA), and the World Health Organization (WHO)—accepts the same criteria but implements them in slightly different ways (see Luckasson et al. 1992, American Psychological Association 1996, World Health Organization 1996).

In all of these definitions, mental retardation is subdivided into four levels. The *DSM-IV*, APA, and WHO subdivisions are based on IQ, as follows: mild (IQ between 50–5 and about 70), moderate (IQ between 35–40 and 50–5), severe (IQ between 20–5 and 35–40), and profound (IQ below 20–5). The AAMR subdivisions are based on the intensity of support required to enhance independence, productivity, and community integration.

2. *Historical Approaches to the Cognitive Aspects of Mental Retardation*

2.1 *Difference Approach*

Mental retardation may be due to genetic causes (e.g., Down syndrome, fragile X syndrome, a bad roll of the genetic dice for genes that impact on intelligence in the general population; see *Genetic Factors in Cognition/Intelligence*) and/or environmental ones (e.g., consumption of alcohol by the mother during pregnancy, lead poisoning, closed head injury, extreme poverty). Historically, psychologists and educators have focused on level of mental retardation rather than cause, both for educational intervention and for research. Extensive characterizations, independent of etiology, are provided in the *Manual of Diagnosis and Professional Practice in Mental Retardation* (American Psychological Association 1996) and are summarized in Table 1. Both educational programs and prognosis have been based on level of mental retardation, without taking into account etiology. A standard research design for studies of cognitive aspects of mental retardation included two groups of participants: individuals with mental retardation and individuals of normal intelligence. In determining the individuals to be assigned to the mentally retarded group, etiology was characteristically ignored; individuals with cultural-familial mental retardation (bad roll of the genetic dice combined with poverty), individuals with mental retardation due to environmental causes such as lead poisoning, and individuals with various syndromes were all likely to be included. Researchers and practitioners commonly assumed that individuals with mental retardation differed from individuals of normal intelligence in specific and fundamental ways, beyond simply intellectual slowness. Proposed bases for these differences included behavioral rigidity, deficits in verbal mediation ability, deficits in short-term memory, and deficits in attention (see, e.g., Burack et al. 1998, Zigler and Balla 1982). This type of position is currently referred to as the 'difference approach.'

Table 1
Characteristics associated with individuals of differing levels of mental retardation

Level of mental retardation	Mental age (MA)	Language ability	Academic skills	Living skills and employment
Mild	8–12 years	Fluent by adolescence	Reading, arithmetic between 1st and 6th grade levels	Independent
Moderate	6–8 years	Functional by adolescence	Functional reading and arithmetic abilities not attained	Some supervision required
Severe	4–6 years	Limited	None	Extensive supervision required
Profound	0–4 years	At most, single words	None	Pervasive supervision required

2.2 Developmental Approach

In the late 1960s, Edward Zigler offered the first formal proposal of the 'developmental approach' (see Burack et al. 1998). Zigler argued that individuals whose mental retardation was due to cultural-familial causes should be considered to be the lowest part of the normal distribution of intelligence. As such, these individuals should follow the same developmental path as individuals of normal intelligence. In particular, individuals with cultural-familial mental retardation should acquire cognitive abilities in the same sequence as individuals of normal intelligence ('similar sequence hypothesis'; see also *Piaget's Theory of Child Development*) although each step might take longer and later steps might not be attained. Furthermore, when matched to individuals of normal intelligence of the same mental age (MA), individuals with cultural-familial mental retardation should show no particular areas of cognitive strength or weakness ('similar structure hypothesis'). Thus, in contrast to the difference theorists, proponents of the developmental approach argued that there was no specific cognitive deficit associated with cultural-familial mental retardation; individuals with this type of mental retardation were simply cognitively slow.

Beginning in the early 1980s, a few psychologists began to argue that the developmental approach should be extended to individuals with mental retardation due to organic causes. In one of the first papers to take this position, Dante Cicchetti and Petra Pogge-Hesse (1982) argued that despite the well-known organic basis for Down syndrome, the development of young children with this syndrome fit the similar structure and similar sequence hypotheses. Cicchetti and Pogge-Hesse further argued that the developmental approach was likely to be appropriate for individuals with mental retardation due to other organic causes. Later studies confirmed the applicability of the developmental approach, especially the similar sequence hypothesis, for individuals with mental retardation due to a wide range of organic causes (Hodapp and Burack 1990).

By the mid 1990s, however, more than 500 genetic disorders associated with mental retardation had been identified, and it was clear that many more would be found. Similarly, large numbers of teratogenic causes had been identified, and it was apparent that more would be identified. These realizations renewed concern among some researchers that etiology should be taken into account for both intervention purposes and research. Many of the same researchers who had demonstrated the applicability of the similar sequence hypothesis for individuals with various genetic syndromes began also to examine potential differences among syndromes with regard to cognitive strengths and weaknesses. These researchers were especially concerned that the similar structure hypothesis might not appropriately characterize many genetic syndromes. Although the majority of mental retardation researchers continued to group participants by level of retardation rather than etiology, an increasing number focused their research on specific etiologies of mental retardation. The findings of many of these studies strongly suggested the importance of taking etiology into account, whether one was concerned with intervention or with theory.

3. Etiological Approach to Cognitive Aspects of Mental Retardation

The etiological approach to mental retardation was founded on the assumption that the various genetic and environmental causes of mental retardation were likely to have differing effects on brain structure and function. The areas and functions most affected were expected to vary due to differences in which genes were involved (deleted, duplicated, or mutated) and the role these particular genes, in transaction with the environment, play in development, or which aspects of

the brain were growing most rapidly at the time of exposure to a particular teratogen. Because of these differences, it is likely that some aspects of cognition will be more severely impacted than others, and that the domains of most serious impact will vary across syndromes. If so, a person's overall MA or IQ (used to assign level of mental retardation, as indicated in Table 1) may not accurately reflect that person's abilities in specific domains. In some domains the person may be performing at a higher level than expected, whereas in others he or she may be performing below the expected level. Furthermore, this pattern would characterize the cognitive abilities of most individuals with mental retardation of the same etiology. Cognitive profiles associated with particular syndromes have begun to be identified (see the papers in Tager-Flusberg 1999 and Denckla 2000). As illustration, the profiles associated with two syndromes that have often been contrasted are described below.

3.1 Cognitive Profile for Williams Syndrome

Williams syndrome is perhaps the syndrome most identified with a specific pattern of cognitive strengths and weaknesses. This very rare syndrome (1 in 20,000 live births) is caused by a hemizygous deletion of about 1.5 megabases on the long arm of chromosome 7 (7q11.23), encompassing at least 18 genes. Early reports noted that individuals with Williams syndrome, although mentally retarded, were extremely gregarious and highly verbal. Popular interpretation of these results has led to claims that despite 'severe' mental retardation, the language of individuals with Williams syndrome is 'spared,' and may even be above the level expected for a normally developing child of the same chronological age (CA). As such, Williams syndrome is often taken to provide a strong case for the independence of language from cognition. This popular interpretation provides a caricature of individuals with this syndrome and is seriously inaccurate in some regards: individuals with Williams syndrome typically have mild mental retardation and only rarely severe mental retardation, and language abilities typically are below, rather than above, CA expectations. Furthermore, language abilities are not independent of nonlinguistic abilities. Performance on language tasks is highly correlated with performance on a variety of other types of cognitive but nonlinguistic tasks (Mervis et al. 1999; see also *Language Acquisition*). However, the popular interpretation does capture the fact, established in empirical studies of the cognitive characteristics of individuals with this syndrome (see below), that some aspects of the language abilities of individuals with Williams syndrome are more advanced than would be expected for IQ or MA.

To determine if a particular syndrome is associated with a specific pattern of cognitive strengths and weaknesses (*cognitive profile*), researchers must establish that most individuals with the syndrome show the same pattern of cognitive strengths and weaknesses. This is best accomplished by comparing the performance of individuals on a single well-standardized test that includes measures of many different types of cognitive abilities. To determine if this pattern is unusual among individuals with mental retardation, researchers may include a contrast group of individuals with other forms of mental retardation in their study (see Mervis and Robinson 1999). Research following this strategy has indicated that Williams syndrome is characterized by a specific cognitive profile: relative strengths in auditory (verbal) short-term memory and language (vocabulary) and extreme weakness in visuospatial construction (e.g., pattern construction, drawing). Most individuals with Williams syndrome, but relatively few individuals with other forms of mental retardation, fit this cognitive profile (Mervis and Klein-Tasman 2000).

The adaptive behavior profile associated with Williams syndrome is consistent with the cognitive profile: communication skills of individuals with Williams syndrome (which rely primarily on verbal ability) are more advanced than their daily living skills (which rely heavily on visuomotor integration or visuospatial construction). Results of magnetic resonance imaging (MRI) studies are also consistent with the Williams syndrome cognitive profile: although brain volume is significantly reduced relative to sex and CA-matched individuals of normal intelligence, the volume of the superior temporal region (important for auditory and language processing) is preserved (Reiss et al. 2000). As the authors note, the direction of the relation between relative strength in auditory memory and language and preserved superior temporal volume is not clear. Although it may be tempting to assume that the relative strength in auditory memory and language shown by individuals with Williams syndrome is due to the preserved volume of the superior temporal region, the reverse is also possible; the preserved volume could be secondary to greater use of auditory memory and language skills over time, resulting in larger cortical representation.

3.2 Cognitive Profile for Down's Syndrome

The most common genetic cause of mental retardation (1 in 800 live births) is Down syndrome (trisomy 21). Most individuals with Down syndrome have mild to moderate mental retardation. It often has been argued that individuals with Down syndrome have a flat cognitive profile. In fact, results of studies finding no significant differences between individuals with Down syndrome and individuals with mental retardation of cultural-familial origin were used as support for the difference approach in contrast to the early developmental approach. More recent studies have shown, however, that Down syndrome is associated with a specific cognitive profile: relative

strength in visuospatial skills (both memory and construction) and relative weaknesses in verbal short-term memory and expressive language (Chapman and Hesketh 2000). The extent to which this cognitive profile is limited to Down syndrome has not been determined. As with Williams syndrome, the adaptive behavior profile for Down syndrome is consistent with the cognitive profile: individuals with Down syndrome have stronger daily living skills than communication skills.

3.3 Within-syndrome Variability

Research has established that there often are major differences across syndromes both in characteristic cognitive profiles and in characteristic level of mental retardation. At the same time, it is important to acknowledge that there are individual differences among people with a particular syndrome. For example, 12 percent of individuals with Williams syndrome do not fit the typical cognitive profile for this syndrome; many of these individuals evidence cognitive strengths and weaknesses similar to those typically associated with Down syndrome. Similarly, a minority of individuals with Down syndrome do not fit the typical profile for that syndrome; these individuals often fit the Williams syndrome cognitive profile.

Another area in which there is wide within-syndrome variability is overall intelligence, or IQ. For example, on the Kaufman Brief Intelligence Test (K-BIT, composed of a verbal subtest and a nonverbal reasoning subtest), the average IQ for a large sample of individuals with Williams syndrome was 66 (in the range of mild mental retardation), with a standard deviation of almost 15 (the same as in the general population) and a range from 40 (the lowest possible IQ on the K-BIT) to 106. Thus, the variability in intelligence among individuals with Williams syndrome is as great as for individuals in the general population; the difference is that the mean IQ for Williams syndrome is about 2 standard deviations below the mean for the general population. Intelligence levels for individuals with Williams syndrome range from severe mental retardation to average intelligence. This range is due to the combination of variability in the genes that impact on general intelligence (which are outside the Williams syndrome deletion region) and the transaction of these genes with varying environmental factors.

Both etiology and level of intelligence are important in understanding the cognitive aspects of mental retardation. As with the general population, large differences among individuals with mental retardation (regardless of etiology) are typically associated with large differences in both academic skills and adaptive functioning. For example, some individuals with Williams syndrome read at the high school level and hold jobs in competitive employment; at the other extreme, some cannot read at all and work in sheltered workshops. At the same time, knowledge of etiology allows for reliable prediction of patterns of cognitive strengths and weaknesses, independent of overall level of intelligence. The existence of these patterns, and the correlations among level of performance in domains of strength and weakness, is important from a theoretical perspective to our understanding of the structure of cognition. These patterns are important from an applied perspective as well: they can be used as a basis for developing intervention strategies geared to individuals with a particular etiology of mental retardation.

See also: Genetic Factors in Cognition/Intelligence; Information Processing and Psychopathology; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Mental and Behavioral Disorders, Diagnosis and Classification of; Mental Retardation: Clinical Aspects; Special Education in the United States: Legal History; Syndromal Diagnosis versus Dimensional Assessment, Clinical Psychology of

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association, Washington, DC

American Psychological Association 1996 Definition of mental retardation. In: Jacobson J W, Mulick J D (eds.) *Manual of Diagnosis and Professional Practice in Mental Retardation*. American Psychological Association, Editorial Board of Division 33, Washington, DC, pp. 13–53

Burack J A, Hodapp R M, Zigler E (eds.) 1998 *Handbook of Mental Retardation and Development*. Cambridge University Press, Cambridge, UK

Chapman R S, Hesketh L J 2000 Behavioral phenotype of individuals with Down syndrome. *Mental Retardation and Developmental Disabilities Research Reviews* **6**: 84–95

Cicchetti D, Pogge-Hesse P 1982 Possible contributions of the study of organically retarded persons to developmental theory. In: Zigler E, Balla D (eds.) *Mental Retardation: The Developmental-Difference Controversy*. Erlbaum, Hillsdale, NJ, pp. 277–318

Denckla M B (ed.) 2000 Specific behavioral/cognitive phenotypes of genetic disorders. Special issue of *Mental Retardation and Developmental Disabilities Research Reviews* **6** (2)

Hodapp R M, Burack J A 1990 What mental retardation tells us about typical development: the examples of sequences, rates, and cross-domain relations. *Development and Psychopathology* **2**: 213–25

Luckasson R, Coulter D L, Pollaway E A, Reiss S, Schalock R L, Snell M E, Spitalnik D M, Stark J A 1992 *Mental Retardation: Definition, Classification, and System of Supports*, 9th edn. American Association on Mental Retardation, Washington, DC

Mervis C B, Robinson B F 1999 Methodological issues in cross-syndrome comparisons: Matching procedures, sensitivity (*Se*), and specificity (*Sp*). *Monographs of the Society for Research in Child Development* **64**: 115–30 (Serial No. 256)

Mervis C B, Klein-Tasman B P 2000 Williams syndrome: Cognition, personality, and adaptive behavior. *Mental Retardation and Developmental Disabilities Research Reviews* **6**: 148–58

Mervis C B, Morris C A, Bertrand J, Robinson B F 1999 Williams syndrome: Findings from an integrated program of research. In: Tager-Flusberg H (ed.) *Neurodevelopmental Disorders*. MIT Press, Cambridge, MA, pp. 65–110

Reiss A L, Eliez S, Schmitt J E, Straus E, Lai Z, Jones W, Bellugi U 2000 Neuroanatomy of Williams syndrome: A high-resolution MRI study. *Journal of Cognitive Neuroscience* **12**: Suppl. 65–73

Tager-Flusberg H (ed.) 1999 *Neurodevelopmental Disorders*. MIT Press, Cambridge, MA

World Health Organization 1996 *Multiaxial Classification of Child and Adolescent Psychiatric Disorders: The ICD-10 Classification of Mental and Behavioral Disorders in Children and Adolescents*. Cambridge University Press, Cambridge, UK

Zigler E, Balla D (eds.) 1982 *Mental Retardation: The Developmental-Difference Controversy*. Erlbaum, Hillsdale, NJ

C. B. Mervis

Mentally Ill: Public Attitudes

Public attitudes to mental illness and mentally ill individuals vary according to a number of internal and external factors. In addition, these attitudes are related to how the illness is perceived in itself, and its causation. The portrayal of mentally ill individuals and psychiatrists in the written and visual media are often negative. Such portrayals not only influence attitudes of the general public but are also influenced in turn by the public's attitudes. Although some cultures may attribute special powers to the mentally ill, by and large societies see mentally ill individuals as violent, frightening, and dangerous, and mental illness as an incurable condition without distinguishing between different types of mental illness. There is considerable research evidence to suggest that common mental disorders such as depression and anxiety may well affect up to one third of the adult population, but most of the negative attitudes are linked with severe major mental illness such as schizophrenia or bipolar affective illness. It can be argued that it is not the illnesses *per se* which provoke negative attitudes, but what they represent (i.e., 'the outsiderness') that is more frightening. These attitudes may or may not always translate into behavior, but will significantly influence the allocation of resources into the delivery of mental health services, recruitment into various branches of mental health professions, and general acceptance of mentally ill individuals by the larger community. In addition, attitudes toward medication can prove to be negative.

1. Definitions

Mental illness is a broad generic term which includes major mental illnesses such as schizophrenia, bipolar affective disorders, common mental disorders such as anxiety and depression, personality disorders, and various other conditions sometimes named after etiology (for example, post-traumatic stress disorder), or after abnormal behavior (for example, episodic dyscontrol syndrome). Therefore, when discussing attitudes towards mental illness, it is important that individual categories are accepted rather than broad generic terms.

1.1 Attitudes

Attitudes form a key point of an individual's thinking, especially toward things that matter to the individual. Attitudes have several components—some a function of the dimensions of personality traits, whereas others are a function of access. By assessing attitudes, one can measure beliefs, but it is not always likely that these beliefs can be turned into behavior. Thus, both cognitive and affective components of these attitudes must be assessed, separately and together. These attitudes do not always remain static and are influenced markedly by a number of factors such as media, education, and societal conditions.

1.2 Stigma

Stigma is the expectation of stereotypical and discrediting judgment of oneself by others in a particular context. The term originates from Goffman's (1963) definition of stigma, which states that stigma originates when an individual, because of some attributes, is disqualified from full social acceptance. Stigma, like the attributes, is deeply embedded in the sociocultural milieu in which it originates and is thus negotiable.

The Greeks in ancient times used the word stigma to refer to bodily signs that exposed something unusual and negative about the signifier. Further conceptualization of stigma has included a mark that sets an individual apart (by virtue of mental illness in this case), that links that individual to some undesirable characteristic (perceived or real in cases of mental illness using violence, fear, and aggression) and rejection, isolation, or discrimination against the individual (e.g., at the workplace by virtue of the sick role and mental illness). Stigma thus includes cognitive and behavioral components and can be applied to age, race, sexual orientation, criminal behavior, and substance misuse, as well as to mental illness. Stigmatizing reactions are always negative and related to negative stereotypes. Stigma allows the society to extend the label of being outside the society to all those who have negative stereotypes. Such an exclusion may be seen as

necessary for bifurcation of society, and will take the same role as scapegoating. The definitions of sickness and sick roles will change with the society, and stigma will also influence how distress is expressed, what idioms are used, and how help is sought.

Stigma and negative attitudes to mental illness are by and large universal. These are briefly discussed below to highlight similarities prior to discussing various factors responsible for these attitudes.

2. *Attitudes in Europe*

The attitudes of public and professionals alike across different countries of Europe have been shown to vary enormously. In this section some of the key studies and the findings are reported.

2.1 *Historical*

In ancient Greece, stigma and shame were interlinked and these can be studied using theories of causation and the balance within the culture between exclusion of mentally ill individuals and the ritual means of cleansing and re-inclusion. The interwoven theories of shame, stigma, pollution, and the erratic unpredictability of the behavior of severely mentally ill individuals suggest that, in spite of scientific and religious explanations, attitudes were generally negative. Simon (1992) suggests that the ambivalence of the ancient Greeks can explain their attitudes toward mental illness. The diseases of the mind were seen in parallel with those of the body, and even the doctors were part and parcel of the same cultural ambience of competition, shame, and fear of failure. During the Medieval and Renaissance periods, mental illness was perceived as a result of imbalance of various humors of the body, and related concepts of shame contributed to this. As Christianity encouraged a God-fearing attitude to the world, mental illness was seen as a sign that God was displeased and punishing the individual, thereby making negative attitudes more likely and more acceptable.

In the seventeenth and eighteenth centuries attitudes shifted, but overall remained negative and in turn provided entertainment to the visitors to Bedlam asylum where individual visitors paid a penny to see the inmates.

2.2 *Current/Recent*

In the nineteenth century, with the establishment of various psychiatric asylums in the UK and the USA, there was generally a growing perception among the lay public and physicians alike that the mind is a function of the brain, thereby influencing a shift toward more humane treatments—although these were still called moral therapy, thereby giving it a moral/religious tinge. The Athenian thinking on psyche has influenced these attitudes. Porter (1987) proposed that the growing importance of science and technology (among other things) was influential in channeling the power of right thinking people in imposing the social norms. The men of the church (also men of power) influenced public opinion, and informed attitudes and behavior toward marginal social elements that would then become disturbed and alien. Central state or market economy influenced the expectations which then divided those who set and met the norms from those who did not.

In a survey, the British Broadcasting Corporation (BBC 1957) reported that the public's tolerance of mentally ill individuals depended upon the circumstances in which contact was made. More than three quarters of those surveyed were willing to mix with mentally ill individuals in areas where low personal involvement occurred; only half were willing to work with people with mental illness, whereas only a quarter were agreeable to such an individual being in authority over them. Attitudes toward mental illness and mentally ill individuals is influenced by a number of external factors such as legal conditions, expectations of treatment, etc.

In Turkey, for example, Eker (1989) reported that paranoid schizophrenia was identified easily among various groups. These attitudes influenced family perceptions, helpseeking, and caring for the individual who is mentally ill, and are influenced by educational status, gender, age, social class, and knowledge. Brändli (1999) observed that in Switzerland stigma was more common in older nonurban males with low education, and poorly informed. There were differences in public knowledge about Alzheimer's disease, depression, and other psychiatric conditions. The findings also suggested that the general public was more accepting of seeking help from their primary care physician rather than a psychiatrist. In Greece, a study revealed that the younger, more educated, and higher social class individuals saw mental illness as a psychosocial problem, whereas older people saw it as a disorder of the nervous system (Lyketsos et al. 1985). Follow-up studies showed an improvement in people's attitudes. Religion has been shown to play a role in forming these attitudes, but a sample from Israel (Shurka 1983) demonstrated that attitudes were more likely to be mixed and inconsistent. From the UK, the role of gender (Bhugra and Scott 1989) and ethnicity (Wolf et al. 1999) have been shown to influence attitudes to mentally ill individuals. The attitudes of other medical and nonmedical professions are also important, but are beyond the scope of this chapter.

3. *Attitudes in America*

Most of the studies from North America have been carried out by social scientists and have illustrated a

range of attitudes to a number of conditions. Here only some of the important studies will be discussed.

3.1 Historical

In America from the seventeenth century onward, the traditional belief prevailed among Christians that madness is often a punishment visited by God on the sinner, and this influenced subsequent views and attitudes. Although both physical and mental illnesses were seen as punishments, it was only in the latter that moral components were seen, especially in the context of teachings from the Bible. A lack of faith and persistence of sinful and lewd thoughts, were seen as key causation factors in the development of insanity.

In the early 1940s, Cumming and Cumming (1957), in their classic study of lay public perceptions of the mentally ill, demonstrated that, when asked to agree with the proposition that anyone in the community could become mentally ill, as normal and abnormal occur on a continuum, the whole educational package was rejected. An explanation was that expression of shame and inferiority was seen as important.

In the 1940s, Ramsey and Seipp (1948) demonstrated that subjects who had higher socioeconomic and educational levels were less likely to view mental illness as a punishment for sins and/or poor living conditions. Yet they were less pessimistic about recovery. In the 1960s, studies from the USA were seen as either optimistic or pessimistic depending upon whether the researcher saw changes in attitudes toward the mentally ill as positive/accepting, or negative/rejecting. It is often difficult to ascertain whether stigma or deviant behavior causes the rejection.

3.2 Recent Studies

Star vignettes have been used in determining public knowledge of mental illness, and public attitudes to mental illness and mentally ill people (Rabkin 1974). Crocetti and Lemkau (1963) used vignettes to demonstrate that virtually all the subjects in their sample agreed with the question, 'Do you think that people who are mentally ill require a doctor's care as people who have any other sort of illness do?' In this sample, low social class and low educational levels did not predict negative attitudes. In another study using case vignettes, it was shown that the largest increase in rejection rates occurred when a person had been admitted to a mental hospital and this was seen as due not to the fact that they were unable to help themselves, but that a psychiatrist or mental health professional had confirmed their status as mentally ill (Phillips 1963). The rejection appeared to be based on how visibly the behavior deviates from the customary role expectations. Continuance of symptoms, and public visibility of these symptoms, and behavior, influences attitudes and makes them more negative.

Social distance is the distance between mentally ill individuals and the lay public on a number of parameters, especially in social interactions. It may also show how individuals will employ mentally ill persons at different levels of responsibility, thereby denoting acceptance or rejection. People may feel reasonably comfortable with mentally ill individuals if they do not have to mix with them on a regular basis. Meyer (1964) reported that three quarters of the subjects in his sample were willing to work with mentally ill people, but only 44 percent would imagine falling in love with such a person. There is no doubt that some illnesses are more stigmatizing than others, as is the severity of the illness, presence of a diagnostic label, and availability of alternative roles. In addition, attitudes can be influenced by the type of treatment and the patient's response to it. Dovidio et al. (1985) reported that people are ambivalent in their attitude toward persons with psychological problems. In an undergraduate class of 94 males and 81 females, the students were asked to give their first impressions of mentally ill applicants in the process of college applications. The results showed that individuals with mental illness were seen favorably in terms of character and competence, but negatively in the context of security and sociability.

Link et al. (1992) suggest that patients' perceptions of how others see them are extremely important. A mental illness label gives personal relevance to an individual's beliefs about how most people respond to mental patients. The degree to which the individual with mental illness expects to be rejected is associated with demoralization, income loss, and unemployment in individuals labeled mentally ill, but not in those without mental illness. Thus it appears that labeling activates beliefs that lead to negative consequences. Patients may use such strategies to protect themselves or their networks.

Such stigmatized attitudes are also reflected in relationships with other outsider groups, such as the homeless, and influence portrayals of negative spoken and visual media (see Fink and Tasman 1992).

4. Attitudes Elsewhere in the World

Attitudes to the mentally ill from other societies and cultural settings are also influenced by a number of factors. Studies on the topic vary tremendously in their methods and data access. The beliefs about mentally ill people and toward mental illness are likely to be related to types of illness prevalent in that group.

4.1 Historical

In ancient Ayurvedic texts, the description of mental illness also suggested physical forms of treatment. In

addition, diet and seasons were reported as playing key roles in the genesis of such conditions. This approach allowed the locus of control to be shifted to some degree away from the individual, thereby making that individual less responsible and less likely to be stigmatized.

4.2 Current

Ilechukwu (1988), in a survey of 50 male and 50 female psychiatric outpatients in Lagos, was able to demonstrate that some patients did believe in supernatural causes of neuroses and disorders, but psychosocial causes were cited most commonly. The attitudes and beliefs of patients are important, but beyond the scope of this article.

Wig et al. (1980), in a three-site study from India, Sudan, and the Philippines, found that, when asked to comment on case vignettes, community leaders were able to identify mental retardation (in all sites), alcohol and drug-related problems (in the Sudanese and Philippines areas), and acute psychosis (in India). Thus it appears that there are cultural differences in the identification of different clinical conditions. They also reported that attitudes toward mentally ill people were more negative in India compared with the other two sites. Thus the studies can be used to establish the needs of the general public and will allow planners to develop appropriate services.

Verghese and Beig (1974), in a survey from South India, reported that over 40 percent of Muslim and Christian respondents reported that marriage can help mental illness, although only 20 percent of Hindus shared this belief. Virtually no respondents reported believing in evil spirits. The most commonly recognized causes mentioned were emotional factors, including excessive thinking. In respondents over the age of 40, one fifth saw mental illness as God's punishment. Christians were again more likely than Hindus to fear mentally ill people (although this fear disappeared with education), and yet the Christians believed strongly in the possibility of a complete cure. Nearly three-quarters of Hindus believed that the moon influences mental illness. These are interesting findings in that they address religious differences in public attitudes.

5. Reasons for Negative Attitudes

Reactions to any taboo or outsider group depend upon a number of factors. These include the frequency of the actual or anticipated behavioral events, intensity and visibility of such behavior, and circumstances and location of such behavior on the one hand, and personal factors on the other. In this section we focus on the latter.

5.1 Age

Several studies have shown that older people tend to have more negative attitudes toward mentally ill people. The reasons for this are many. Older people, in spite of their life experiences, are generally more conservative and equally rejecting of behavior which is seen as odd and alien. The role of age is likely to be mediated by other factors such as education, social and economic class, etc.

5.2 Types of Illness

Attitudes towards people with schizophrenia are likely to differ from those reported toward people with depression or personality disorder. This may reflect the stereotypic images of the condition or fear related to the condition. It may also be due to previous knowledge about the illness. The perceived causative/etiological factors also play a role in attitudes.

5.3 Gender

Males tend to have more negative attitudes and, as noted above, are also more likely to be rejected when they suffer mental illness. Females may be more sympathetic, for a number of reasons. Other studies of attitudes toward other alien groups also demonstrate that females are more positive. They are also more likely to be carers, and may be the first to contact psychiatric services on behalf of individuals. For women, change in role after mental illness is likely to produce more stigma. Yet they are more likely to be admitted with more 'masculine' illnesses such as personality disorder and drug abuse. The gender roles in the context of illness may well play a role in generating negative attitudes.

5.4 Religious Beliefs

Some studies have demonstrated that Christians tend to have more negative attitudes toward mental illness, but again this is not a consistent finding. The individual subject's level of religiosity and depth of religious values must be studied rather than simple religious ascription.

5.5 Educational Status

The effects of education on attitudes are mixed (Bhugra 1989). Some studies have related negative attitudes clearly with low educational status, whereas others have failed to show such association, or showed that highly educated subjects held more negative attitudes, including those studying professionally.

5.6 Professions

Medical students may hold more negative and stereotypical attitudes, and other branches of medicine too have been shown to have negative attitudes to mentally ill people (e.g., schizophrenics), those with odd behavior (e.g., deliberate self-harm), and psychiatry as a profession. The psychiatrist is often lampooned and seen as a 'head shrinker' or a modern-day witch-doctor. There are individuals who are well educated and belong to a high socioeconomic class and yet hold negative attitudes.

5.7 Others

Some ethnic groups such as African-Caribbeans in the UK, and Hispanics, Asian-Americans, and Mexican-Americans in the USA have been found to have more negative attitudes toward mentally ill individuals or be far more restrictive in their description of etiology of mental illness compared to the white majority population. Nonacculturated individuals appeared to have more old-fashioned attitudes (i.e., negative stereotypes).

These are some of the complex set of factors that influence attitudes and stereotypes.

6. Educational Interventions

Educational interventions in order to reduce stigma toward mentally ill individuals and foster positive attitudes towards mental illness and those who are mentally ill are based on several levels of education. One-off educational programs and fact sheets on any illness are not likely to produce long-term changes. There is considerable evidence in the literature to suggest that education, if given at an appropriate and stable level, and repeated as required, will produce changes that can be sustained. In addition, educational interventions must target specific populations, bearing in mind their age, gender, ethnic composition, primary language, educational attainments, social class, etc.

Wolff et al. (1999) reported on the results of follow-up of neighbors, where two group homes for people with mental illness had been set up and one set of neighbors had received extensive education. They found that although the public education intervention may have had at best only a modest effect on knowledge, behavior toward the mentally ill residents changed. There was a decrease in fear and exclusion, and increased levels of social contact in the experimental area. They observed that educational intervention *per se* did not in itself lead directly to a reduction in fearful attitudes, whereas contact with patients did. Thus, any campaign which encourages subjects to increase contact with mentally ill individuals may prove to be more successful. The educational programs must be paced slowly and in a sustainable manner. The patients and their carers must be involved in planning these interventions without turning the whole exercise into a circus.

Any educational intervention must target the intervention either at the group most at risk of negative attitudes or at those already having negative attitudes. Repeated packages and adequate time for consultation and discussion will influence attitudes. Using small groups with experiential teaching is more likely to be successful compared with seminars or lectures with large groups. Any educational campaign must be sustained and momentum maintained. The effect on attitudes following several interventions is more sustained, and greater than the sum of the individual effects. Using a number of strategies, involving participation on the part of the public, carers, and patients, can influence attitudes.

The educational packages and interventions aimed at improving attitudes will, in the long run, influence resource allocation into mental health, improve recruitment, and move toward acceptance of community care. Such interventions can reduce fear, make expectations more realistic, and prevent attitudes from hardening. However, these interventions must be clear, focused, and appropriate, based upon the needs of the group that is being educated rather than on perceptions of need by the professionals. These packages must deal with alienation experienced by patients at different levels and in different settings, such as housing, employment, and social settings.

7. Conclusions

The attitudes of the public toward mentally ill individuals reflect prevalent social norms and mores. Expectations from psychiatric services, response to treatment, type of treatment, type of mental illness, and inaccessibility to treatment are external factors that will influence public attitudes. Age, gender, personality traits, and social and economic class are some of the personal factors that will influence these attitudes. These attitudes will not always translate into behavior, but one will be influenced by the other. Labeling of mental illness and the perception of mentally ill individuals as dangerous along with associated fear will all influence attitude formation. Any educational intervention must include some or all of these factors. Negative attitudes will influence helpseeking as well as compliance. These attitudes may well be linked to illness, but also to stereotypes of the illness, and to associated or perceived impairment as a result of that illness. These negative attitudes will also influence rejection of any preventative strategies the psychiatric profession may wish to advocate.

See also: Attitude Formation: Function and Structure; Discrimination; Foucault, Michel (1926–84); Health and Illness: Mental Representations in Different Cul-

tures; Mental Health: Community Interventions; Mental Health Programs: Children and Adolescents; Mental Illness: Family and Patient Organizations; Prejudice in Society; Social Constructivism

Bibliography

BBC 1957 *The Hurt Mind: An Audience Research Report*. BBC, London

Bhugra D 1989 Attitudes towards mental illness. *Acta Psychiatrica Scandinavica* **80**: 1–12

Bhugra D, Scott J 1989 Public image of psychiatry: A pilot study. *Psychiatric Bulletin of the Royal College of Psychiatry* **13** : 330–3

Brändli H 1999 The image of mental illness in Switzerland. In: Guimon M J, Fischer W, Sartorius N (eds.) *The Image of Madness*. Karger, Basle, Switzerland

Crocetti G M, Lemkau P V 1963 Public opinion of psychiatric home care in an urban area. *American Journal of Public Health* **53**: 409–17

Cumming E, Cumming J 1957 *Closed Ranks: An Experiment in Mental Health Education*. Harvard University Press, Cambridge, MA

Dovidio J, Fishbane R, Sibicky M 1985 Perceptions of people with psychological problems. *Psychological Reports* **57**: 1263–70

Eker D 1985 Attitudes of Turkish and American clinicians and Turkish psychology students toward mental patients. *International Journal of Psychiatry* **31**: 223–229

Fink P, Tasman A 1992 *Stigma and Mental Illness*. APA Press, Washington, DC

Goffman E 1963 *Stigma*. Prentice Hall, Englewood Cliffs, NJ

Guimón J, Fischer W, Sartorius N 1999 *The Image of Madness*. Karger, Basle, Switzerland

Ilechukwu S 1988 Interrelationship of beliefs about mental illness psychiatric diagnosis and mental health care delivery among Africans. *International Journal of Social Psychiatry* **34**: 200–6

Link B, Cullen F, Mirotznik J, Struening E 1992 The consequences of stigma for persons with mental illness. In: Fink P J, Tasman A (eds.) *Stigma and Mental Illness* 1st edn. APA Press, Washington, DC

Lyketsos G, Mouyas A, Malliori M, Lyketsos C 1985 Opinion of public and patients about mental illness and psychiatric care in Greece. *British Journal of Clinical and Social Psychiatry* **3**: 59–66

Meyer J K 1964 Attitudes towards mental illness in a Maryland community. *Public Health Reports* **79**: 769–72

Phillips D 1964 Rejection of the mentally ill: The influence of behavior and sex. *American Sociological Review* **29**: 679–89

Porter R 1987 *Mind Forg'd Manacles*. Athlone, London

Rabkin J 1974 Public attitudes towards mental illness. *Schizophrenia Bulletin* **10**: 9–23

Ramsey G, Seipp M 1948 Attitudes and opinions concerning mental illness. *Psychiatric Quarterly* **22**: 428–44

Shurka E 1983 Attitudes of Israeli Arabs towards the mentally ill. *International Journal of Social Psychiatry* **29**: 95–100

Simon B 1992 Shame, stigma and mental illness in Ancient Greece. In: Fink P J, Tasman A (eds.) *Stigma and Mental Illness*, 1st edn. APA Press, Washington, DC

Verghese A, Beig A 1974 Public attitudes towards mental illness: The Vellore study. *Indian Journal of Psychiatry* **16**: 8–18

Wig N N, Suleiman M A, Routledge R et al. 1980 Community reactions to mental disorders. *Acta Psychiatrica Scandinavica* **61**: 111–26

Wolff G, Pathare S, Craig T, Left J 1999 Public education for community care. In: Guimon J, Fischer W, Sartorious N (eds.) *The Image of Madness*. Karger, Basle, Switzerland

D. Bhugra and W. Cutter

Mentoring: Supervision of Young Researchers

1. Context

The Division of Science Resources Studies of the US National Science Foundation reports in statistical tables compiled in 1999 (Hill 1999) that US universities awarded close to 43,000 doctoral degrees (including a small number of professional degrees) in 1998. In the USA, these numbers had increased from 33,500 in 1988, a decade ago, with most of the increase occurring by 1994. Standing at the peak of a large educational enterprise which represents a large social investment, these numbers indicate the changed context in which postbaccalaureate education operates. According to statistics compiled in many regions (excluding primarily African countries, because of lack of access to statistical data), in 1997 160,433 Ph.D.s were awarded, of which 91,372 Ph.D. theses were written in the fields of science and engineering, including the social and behavioral sciences (National Science Board 2000).

Given the scale of these investments and the importance for society of results from research activities using graduate and postdoctoral assistance, it is not surprising that questions arise concerning the roles and responsibilities of graduate students, postdoctoral associates, faculty, and their institutions. The system is showing the strains of increasing numbers as well as mounting calls for accountability from graduate students themselves and the governmental funders and sponsors of graduate education. Graduate students have begun to organize and gain collective bargaining rights on campuses and postdoctoral associates are pushing for national reforms to their status, including written contracts, uniform job title and benefits, support for postdoctoral associations, representation on institutional policy-making committees, and a national postdoctoral organization.

Adequate response to these stresses requires the attention of academic institutions and professional associations. Most important, perhaps, are the responses at the level of departments and programs actually producing new Ph.D.s. The beginnings of change can be seen in the developments described

below; but better understanding of changes in mentoring practice and supervision requires systematic research attention, of a kind they have not previously received, in order to identify and assess effects on individuals, departments, programs, and institutions, and to make improvements in the future.

Increased social and governmental attention to questions of research misconduct has given particular impetus for change in this arena. These concerns resulted in new regulations in the USA requiring institutions to establish procedures in response to allegations and many other nations have implemented such procedures. Governmental and academic institutions, as well as scientific and professional associations, regard this as a broader mandate for education in research integrity. In the USA, the National Institutes of Health has established requirements for ethics training for all research staff, which have implications for mentoring and supervision.

Professional associations have been custodians of good practice for their members, as well as for practitioners and researchers who may not be members but draw on the body of knowledge that these associations incorporate, and represent themselves as having the relevant expertise. Today's world requires that experts work with others with very different specialties more often than in the past; the need for these group or team efforts raises additional issues for good mentoring and supervision and for the effective review of research results.

2. *Supervision*

The US Council of Graduate Schools issued guidance for supervisory practice in 1990 (Council of Graduate Schools 1990). Recognizing that there are no recipes for the creative impetuses that underlie the most successful interactions between faculty members and students, the brochure provides useful recommendations that can assist departments, programs, and supervisors to enable their students to make good progress. These recommendations are appropriate and can be adapted easily for use with postdoctoral fellows. The guidance indicates that departments should have a clear written framework outlining the stages students are expected to complete in graduate training. Supervisors should help their students plan their time carefully, meet with them regularly, and be ready, at an early stage in the program, to make an assessment, which will be perceived to be fair and appropriate, about whether they should continue towards the Ph.D. Supervisors should engage the students early in research activities, in order to promote knowledge about the field and facilitate the selection of a dissertation topic. In the research endeavors for graduate students, students should discouraged from achieving perfect results at the expense of good research and should not be allowed to become distracted by interesting, but tangential, issues.

In readying students to undertake research, the department or program should clarify the standards it expects students to meet. Will a literature survey be necessary? Will experimental research or fieldwork be required? Are there certain techniques or tools that the student must master in the training program or dissertation? What writing standards must reports and theses meet? In the middle stages of training, graduate students should expect to complete most of the original work that will form the dissertation. Since research may lead in unexpected directions, supervisors and students need to be alert to considerations that might require modifying the original goals, particularly to allow completion of the project in a timely fashion. Careful records need to be kept, and here supervisory assistance and monitoring is most important and lays the foundations for good research practice throughout a career.

Team research may require additional considerations. There needs to be a careful definition of students' contribution to research work and the students should be given opportunities to demonstrate a broad understanding of project purposes and methods. Of course, students in individual projects also need opportunities for presentation about their status and findings. Writing should begin early in the project. Milestones here are useful. An introduction or overview as well as a list of references can be drafted and revised as work progresses. Questions about interim findings and the relation of the research to the work of others need to be posed so that students can stay abreast of developments in their fields. Experience demonstrates that students should be allowed to complete their dissertations before being asked to continue as a postdoctoral research associate.

2.1 *Checklist for Supervisors and Departments*

Does the department have a document describing its views on good supervisory practice? What steps exist to make good matches between supervisors and prospective students or postdoctoral associates? Should students and postdoctoral associates present reports for assessment by nonsupervisors early in their tenure? Do they make public presentations satisfactorily? Are adequate meetings scheduled between supervisors and students or postdoctoral associates? Are there regular assessments of progress and background knowledge? Do supervisees and supervisors see the assessment procedures as satisfactory? For students, how are research topics refined early in training? For postdoctoral associates, how are research projects assessed to provide adequate stimulation and chance for advancement? Have appropriate

times been scheduled for developing long-term programs of research and for identifying milestones? Do supervisors check research records for accuracy, quality, and completeness?

2.2 Checklist for Students

Have you developed a systematic work-plan and identified major difficulties and relevant references? Are your records in good order, so that you can find the relevant material about a problem you worked on six months ago? Do you have draft portions of any completed aspects of the research? Do others find your written work difficult or easy to understand? Have interim projects, such as figures, tables, or other matter, been identified for preparation during the project?

3. Mentoring

In 1997, the National Academy of Sciences published a handbook for faculty who want to develop good mentoring skills. It points out that mentoring is a personal, as well as a professional, relationship. Good mentors want to help their students get the most from their education, assist their socialization into the field, and aid them in finding employment. Students and postdoctoral associates benefit from having many mentors with different skills and priorities.

Successful mentors today must deal with issues involving ethnicity, culture, sex, and disability. Currently, women account for 42 percent of doctorates earned in the USA, while the number of doctorates earned by US minority group members increased at a faster pace than the total, rising from 9 to 15 percent. Non-US citizens make up about a quarter of the new doctorates. Students may need access to a number of different faculty or graduate student peers for assistance. This may be particularly important when they come from different cultural or minority backgrounds. As faculty advisors, mentors must build mutual respect with their mentees. Besides recognizing how issues of diversity may affect students' priorities and perceptions, faculty need to recognize that their students are likely to pursue careers outside academe and are legitimately concerned about time to degree and funding for their graduate education.

Junior faculty are also often in need of mentoring. They have the stresses created by considerably increased responsibilities for teaching and research and for professional service. Departments can establish numerous, albeit simple, ways to ease this transition, through appropriate workshops or by encouraging senior faculty to pair with junior colleagues. Departments can provide clear written guidance about tenure requirements and teaching policies and facilitate the acquisition of necessary resources for research, as well as nonresearch resources, such as childcare, which may be very important for entering faculty. They should arrange for frequent feedback and evaluation. Minority and women faculty in particular may need protection from demands on their time created by the lack of others of their race or gender in their departments.

Students and postdoctoral associates often need assistance in mapping out career plans and in being introduced to contacts that may be able to help them. Students should be reminded that prospective employers also expect candidates to have other skills in addition to their competence in a chosen research field. Sometimes, mentors might need to provide frank and honest advice that identifies weaknesses, but also indicates areas of strength which mentees could develop. Over the course of the training, areas of weakness can also be corrected. Practice at teaching or making presentations, for instance, or critiques of draft grant proposals and resumes as well as research reports offer opportunities for improvement. Areas of strength might provide the basis for a good letter of recommendation from the mentor to the appropriate potential employer.

There are many ways that departments and institutions can improve the quality of mentoring, all requiring systematic and on-going attention. Departments and institutions can track the progress of former students for information about careers. They can collect other relevant data or ask advanced students to assess how well their mentors or others have contributed to their progress. Institutions can schedule special activities and provide instructions on mentoring for new faculty and advisors. They can monitor faculty abuses and take a more active role in selecting faculty for advisory roles. They can provide workshops and discussions for students. They can encourage electives and other classes to broaden skills of students and postdoctoral associates.

All discussions of good mentoring and supervisory practice indicate that attention to these matters is imperative if trainees are to emerge with an intellectual and ethical framework that promotes research integrity and with some degree of satisfaction with their training experiences. Supervisors and mentors transmit to their students many of the attitudes and information needed to carry out significant research and good teaching and to practice responsible research. Faculty must recognize the need for human relationships to be built on respect and trust, as well as the special issues that the vulnerability of trainees raise.

In selecting mentors, students need to consider their publication records and national recognition, history of finding support for trainees, availability during the period of the student's training, prior training record and positions of recent graduates, practices of recognition for student accomplishments, and laboratory or other research arrangements. Students and post-

doctoral associates should also recognize that mentors have different personalities. Some are take an active interest and become involved in student projects; some are disengaged; some may develop friendships with their mentees. Each can pose difficulties for both parties. Open communication needs to be encouraged so that issues and potential misunderstandings can be clarified and any changes that might be necessary can be initiated in timely fashion.

To produce good work, mentors and trainees also need to be educated in the research standards of their fields. They need to be familiar with relevant codes of ethics. In experimental work, standards of laboratory practice or laboratory safety need to be discussed. This also holds true for standards of fieldwork and survey practice. Regulatory requirements concerning animal and human subjects need to be well understood. Standards for ethical practice of human subjects research include matters of privacy and confidentiality, consent, and community protection. With regard to research in the social and behavioral sciences, complex questions concerning deception or stigmatization of individuals or communities may need particular care, especially in the context of research on vulnerable populations in other cultures. All should be familiar with definitions of research misconduct and the policies of their program and institution. Students must be trained in responsible use of statistical methods, where appropriate, and in the importance of stewardship of records and data, as well as in respect for ownership of ideas and intellectual property. Standards for responsible authorship and peer review also need to be understood and discrepancies that may exist between different fields need to be examined. These are areas where departments and institutions as well as professional associations and scientific societies have much to offer.

4. Conclusion

For persons in academic research environments to achieve good results, they require and deserve assurance that others will do their share and that they will be protected if difficult problems arise. To foster this ethic, graduate departments or groups should develop adequate written statements of the terms of graduate study, including policies about conflict of interest, and should assign necessary duties to individual faculty members. These statements should be distributed to trainees. Individual faculty often have responsibilities for the admission of students, which allows them to assume similar responsibilities for advising students and postdoctoral associates and for placing graduate students in postdoctoral positions and jobs. Tasks that can help graduate students and postdoctoral associates succeed—such activities as writing proposals or research reports—can also be shared among faculty, as can assignments for making students aware of ethical aspects of research and professional activities.

Departments should create systems that specify advisors' duties and methods for evaluating their performance. Recent publications and guidance from national bodies and professional associations, as well as books and articles by individual scholars, can be helpful in developing materials and activities responsive to the needs of new faculty, postdoctoral researchers, and graduate students. Training for advisors will be needed. Increased attention to structural reforms and ethical research practice might also lead to the diminution of research misconduct.

See also: Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives; Higher Education; Research Conduct: Ethical Codes; Research Ethics: Research; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects; Science, Sociology of; Universities, in the History of the Social Sciences

Bibliography

Buchanan A 1996 Perfecting imperfect duties: Collective action to create moral obligations. *Business Ethics Quarterly* **6**: 27–42

Council of Graduate Schools 1990 *Research Student and Supervisor: An Approach to Good Supervisory Practice*. Council of Graduate Schools, Washington DC

Hill S T 1999 *Science and Engineering Doctorate Awards: 1998, Detailed Statistical Tables*. Division of Science Resources Studies, National Science Foundation, Arlington, VA

Hollander R D 1999 Guaranteeing good scientific practice in the US. In: Wolfrum R (ed.) *Ethos of Research*. Max-Planck-Gesellschaft, Munich, Germany, pp. 199–212

LaPidus J B 1997 Doctoral education: preparing for the future. *CGS Communicator* **XXX**: 10, Special Edition, November, pp. 1–10

Leatherman C 1998 Graduate students gather to learn 'organizing 101'. *The Chronicle of Higher Education* **August 14**: A10–11

Macrina F L 1995 Mentoring. *Scientific Integrity: An Introductory Text with Cases*. ASM Press, Washington, DC, pp. 15–39

Mervis J et al. 1999 Postdocs working for respect. *Science* **285**(5433): 1513–33

National Academy of Science, National Academy of Engineering, National Academy of Medicine, Committee of Science, Engineering, and Public Policy 1997 *Advisor, Teacher, Role Model, Friend: On Being A Mentor to Students in Science and Engineering*. National Academy Press, Washington, DC

National Science Board 2000 *Science and Engineering Indicators 2000*. National Science Foundation, Arlington, VA

North Carolina State University Resource Guide in Research Ethics, http://www.fis.ncsu.edu/grad/ethics/resource.htm

Sieber J E 1992 *Planning Ethically Responsible Research: A Guide for Students and Internal Review Boards*. Sage, Newbury Park, CA

Upham S 1995 University of Oregon Guidelines for Good Practice in Graduate Education. *CGS Communicator* **XXVIII**:10, October,10–2

Weil V 1999 Mentoring: Some ethical considerations. Unpublished paper, 17 pp, available from Professor Vivian Weil, Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago

Zigmond M J 1999 Promoting responsible conduct: Striving for change rather than consensus. *Science and Engineering Ethics* **5**: 219–28

R. Hollander

Meta-analysis in Practice

If meta-analysis describes the quantitative approach to combining information from a number of different studies, then the practice of meta-analysis is concerned with the details of implementation of that approach. In effect, the practice of meta-analysis is concerned with ensuring the validity and robustness of the entire enterprise of research synthesis, including, problem formulation, design and data collection, data quality and data-base management, and analysis and interpretation. The practice of meta-analysis is really about the practice of data analysis.

The scope of this article is, by necessity, modest. However, a comprehensive guide to the science and art of research synthesis is available in *The Handbook of Research Synthesis* (Cooper and Hedges 1994a) and is recommended to the interested reader. Here, selected topics are presented that represent current intellectual themes in the practice of meta-analysis, such as: (a) the role of decisions and judgments, particularly judgments about similarity of studies; (b) the importance of sensitivity analysis to investigate the robustness of those decisions; and (c) the role research synthesis plays in the process of scientific discovery, with brief illustrations of how meta-analysis contributes to that process.

1. Selected Issues in Practice

1.1 Similarity Considerations and Combining Information

Implicit in the decision to combine information across studies is the assumption that the studies are similar or exchangeable. The dimensions on which studies might be similar are quite broad and include, for instance, design features, subject pool, quality, measurements, and even context. This assessment about the similarity of studies is central in the practice of meta-analysis because it is at the heart of deciding what information to combine and how to combine that information (Gaver et al. 1992). Underlying the choice of statistical models for combining information in a meta-analysis, for example, are explicit assumptions about the similarity of the studies. Thus, specifying a fixed-effects model for a meta-analysis is essentially equivalent to the assumption that the studies are homogeneous, measuring the same underlying parameter, and any observed differences among studies is due to sampling variability. Alternatively, a random-effects model formally incorporates study heterogeneity through the specification of a between-study component of variance.

It will be assumed here that each study in a meta-analysis provides summary information that is common across studies, including measures of an outcome of interest and standard errors of those measures. Often there is little control over the choice of the summary measure, and different summary measures may be reported across the studies. In practice, the meta-analyst will need to decide upon a single meaningful measure, frequently a standardized measure such as an effect-size or risk ratio, and subsequently transform the study-specific summaries to the chosen measure. For example, in a study of the effect of teacher expectancy on pupil performance in an IQ test, the summary measure used for each study was the effect-size, a standardized mean difference calculated as the mean IQ score of the experimental group minus the mean of the control group divided by a pooled standard deviation (see *Meta-analysis: Overview*). Effect sizes are most useful when study outcomes are assessed with different instruments and are measured on a continuous rather than dichotomous scale. In the case when all the studies to be combined report the same summary measure, as in the IQ study, an important practical decision for the analyst is whether different studies are more usefully compared by their effect sizes or by the simple unscaled differences in means, a quantity often of more direct scientific relevance and interpretability.

Assessments about what studies to combine requires subjective judgments on the part of the investigator, and will depend on the motivation and context of the research synthesis. In addition, inherent in the specification of a statistical model to combine information in a meta-analysis are explicit formal assumptions about the similarity of the studies and the data to be combined (Draper et al. 1993). Assume that there are k independent studies available for a meta-analysis, and let T_i denote the study-specific effect from study i ($i = 1,\ldots, k$). If all the studies are measuring the same thing, for example, the same 'true' change in IQ scores, the k studies are assumed to be similar in that each study is assumed to provide an estimate of the true effect θ ($= \theta_1 = \cdots = \theta_k$). This assumption of similarity implies the following fixed-effects model for combining the study-specific effects,

$$T_i = \theta + \varepsilon_i \tag{1}$$

where ε_i is typically assumed normal with mean zero and variance V_i, and the ε_i are regarded as inde-

pendent. (For simplicity, assume the V_i are known.) An optimal estimator of θ is

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \tag{2}$$

a weighted average of the T_i in which the weights w_i are proportional to $1/V_i$ (see also *Meta-analysis: Overview* and *Meta-analysis: Tools*).

Alternatively, the meta-analyst might conclude that the studies are heterogeneous, implying that differences in the T_i come both from experimental error and from actual differences between the studies. If explicit information on the source of heterogeneity is available, it may be possible to account for study-level differences in the modeling, but in the absence of such information a useful and simple way to express study-level heterogeneity is with a random-effects or hierarchical model:

$$\begin{aligned} &\text{Stage I:} \quad T_i = \theta_i + \varepsilon_i \\ &\text{Stage II:} \quad \theta_i = \theta + \zeta_i \end{aligned} \tag{3}$$

Here the ε_i are, for example, taken to be normal with mean zero, and variance V_i (the within-study component of variation), and the ζ_i are regarded as normal with mean zero and variance τ^2 (the between-study component of variation). All the ε_i and ζ_i are considered independent. When a probability distribution is specified for the second stage parameters, θ and τ, this class of models is known as a Bayesian hierarchical model (see *Bayesian Statistics* and *Hierarchical Models: Random and Fixed Effects*).

A key implication of the second stage model in Eqn. (3) is that there is information about a particular θ_i from all of the other study-specific effects, $\{\theta_1, \ldots, \theta_k\}$. Specifically, an estimate of θ_i is

$$\hat{\theta}_i(\tau) = (1 - B_i(\tau))T_i + B_i(\tau)\hat{\theta}(\tau) \tag{4}$$

a weighted average of the observed ith study-specific effect, T_i, and the estimate of the population mean effect, $\hat{\theta}(\tau)$, where the weights are $B_i(\tau) = V_i/(V_i + \tau^2)$. Although most meta-analyses are interested in learning about θ and τ^2, it is informative to examine the behavior of Eqn. (4) as a function of τ. The weighting-factor $B(\tau)$ controls how much the estimate of $\hat{\theta}_i$ 'shrinks' towards the population mean effect. When τ is taken to be zero, $B(\tau)$ is 1 and the random-effects model is equivalent to the fixed-effects model. Small values of τ describe situations in which studies can 'borrow strength' from each other and information is meaningfully shared across studies. Very large values of τ, on the other hand, imply that $B(\tau)$ is close to zero, so that not much is gained in combining the studies. The Bayesian approach to meta-analysis particularly through the specification of a probability distribution for τ provides a formalism for investigating how similar the studies are, and for sharing information among the studies when they are not identical. Many practical issues in the implementation of Bayesian meta-analysis are discussed and illustrated in DuMouchel and Normand (2000).

1.2 Sensitivity Analysis

At every step in a research synthesis decisions and judgments are made that have an impact upon the conclusions and inferences drawn from that meta-analysis. Sometimes a decision may be easy to defend, for example, the omission of a very poor quality study from the meta-analysis, or the use of a random-effects model instead of a fixed-effects model. At other times, a decision may be much less convincing: for example, the use of only the published literature, the omission of a study that appears to yield an unusual effect size estimate, or the use of one regression model instead of another. When the basis for a decision is tenuous, it is important to check whether plausible departures from that basis appreciably affect the conclusions; in other words, to check the sensitivity of the conclusions to small changes in the analysis.

The objective of sensitivity analysis is to evaluate the impact of the various decisions and judgments made at each step of a research synthesis. The following examples are illustrative of the types of issues, posed as questions, that can arise at different steps in a meta-analysis.

(a) How reliable is the information obtained by a single individual who identifies, reviews and abstracts the relevant literature? Should there be at least two individuals who conduct independent searches of the literature and assess the relevant results from the studies? Should the abstracters of the literature be blind to the authorship of the studies?

(b) How do the conclusions change if we assume different statistical models for combining study-specific effects? Do the conclusions change if we use a weighted versus an unweighted summary measure of effect-size? How sensitive are the inferences to distributional assumptions, including the assumption of independence of studies? How sensitive are the estimates of θ_i and θ to τ^2?

(c) Which studies are outliers? How sensitive is the combined estimate of effect size to any one particular study? In other words, which studies are influential in the sense that the conclusions would change if that study was not included? What features of a study, for example, randomized versus not randomized, blind versus not blind, characterize influential studies?

(d) If only statistically significant studies are published in the literature, then the studies that have appeared in the published literature are not a representative sample of all studies addressing the specific

research question, a problem known as publication bias. What are the effects of publication bias on the conclusions of a research synthesis? Can the magnitude and the impact of publication bias be assessed in a research synthesis? What is the effect on the results of a research synthesis if the retrieval of the literature is not exhaustive?

(e) How robust are statistically significant associations found in a meta-analysis? Meta-analysis is an observational study, and as such, is subject to the influence of uncontrolled variables. How sensitive are the conclusions of a meta-analysis to the effects of confounding variables, either measured or unmeasured? Could alternative hypotheses explain the observed associations?

Little has been written about methods for sensitivity analysis in research synthesis (see Greenhouse and Iyengar 1994). The decisions made over the course of a meta-analysis have either a substantive or statistical character. By and large, the latter are easier to deal with than the former, because computer packages have made it possible to reanalyze data fairly easily, while some substantive decisions are irrevocable. For instance, early decisions regarding the focus of the primary and secondary research questions determine the nature of the literature search, the kind of information that is extracted from the retrieved literature, and the way that information is processed for later synthesis. In this context, changing the goals in midstream calls for an entirely new meta-analysis. Despite that, it may still be useful to ponder what might have happened if, for example, the search for primary studies were more complete.

2. *Meta-analysis for Explanation*

Research synthesis plays a central role in the process of scientific discovery with meta-analysis providing a formal methodology for the systematic accumulation and evaluation of scientific evidence. As Cooper and Hedges (1994b) note, 'Research syntheses attempt to integrate empirical research for the purpose of creating generalizations. Implicit in this definition is the notion that seeking generalizations also involves seeking the limits and modifiers of generalizations.' Perhaps the single most important strength of research synthesis over individual studies is the ability to investigate the robustness of a relationship and the conditions moderating its magnitude and direction. For example, in 1955, Beecher used a meta-analysis of 15 studies to demonstrate the consistency of the placebo-response and the effectiveness of placebos in the relief of pain across a number of different conditions, including severe postoperative pain, angina, and headache among others (Mosteller and Chalmers 1992). By comparing results across studies it is possible to investigate systematic variation in effect size due to personal characteristics, design features, outcomes, and time. The contemporary practice of meta-analysis goes beyond the simple description and estimation of effects to address the problem of explaining why, or under what conditions, a given effect can be observed.

The achievements of meta-analysis have been considerable and its use in many disciplines, especially in the social, behavioral, and medical sciences, continues to grow. This growth is facilitated, in part, by the existence of electronic bibliographic databases such as ERIC, PsychINFO and MEDLINE, and initiatives such as the Cochrane Collaboration, an international organization that prepares, maintains and ensures the accessibility of systematic reviews of the effects of health care interventions (Chalmers et al. 1997). Meta-analysis has become a standard methodology in the researcher's and policy maker's tool box.

2.1 *A New Conceptual Framework for Meta-analysis*

Rubin (1992) has proposed a perspective on meta-analysis which is useful to consider and contrast with the goals of a traditional quantitative literature review. The idea is that the objective of combining information across studies is to build and extrapolate response surfaces in an attempt to estimate 'true effects' from ideal studies rather than estimate effects from some population of 'flawed' studies. In this approach, each study can be classified by two types of characteristics: S, which are variables of scientific interest (e.g., dose of treatment, subject characteristics, etc.) and D, which are design variables (e.g., sample size, whether the study was randomized or not, investigator characteristics, etc.). If T represents the observed outcome for each study, then in principle a response surface model can be built for T given (S, D) using the observed studies. However, the region where the data can fall is not necessarily of primary scientific interest, because it reflects the various choices that were made by necessity about values of D. Interest would be in the extrapolated response surface, $E(T \mid S, D = D_0)$ with D fixed at specified ideal values, say D_0. The details of implementation of such an approach remain to be developed. Nevertheless, Rubin suggests a number of intriguing practical implications that follow from this perspective, for example, there is no need to be selective about which studies to include in a meta-analysis since all studies contribute to the estimate of the response surface and modeling will automatically down-weight poorer studies. Also, using principles of experimental design, this approach provides guidance in the choice of new studies to add to the database based on choosing studies, i.e., values of S and D, that increase the precision of estimation of the response surface, $E(T \mid S, D = D_0)$. The basis for this conceptualization is grounded in ideas underlying response surface methodology, hierarchical statistical models, and statistical methods for causal inference

and for missing data (see *Causal Inference and Statistical Fallacies* and *Statistical Data, Missing*).

2.2 Applications of Meta-analysis in Practice: Vignettes

The following vignettes are meant to provide a flavor of some of the more contemporary applications of meta-analysis and the central role that meta-analysis plays in explanation, program evaluation, and in informing policy decisions. Additional illustrations of this kind can be found in Cook et al. (1992) which presents four detailed case studies of research syntheses along with expert commentary highlighting the value of meta-analysis for scientific explanation and the challenges of implementation.

(a) Meta-analyses that include moderator variables provide an opportunity to investigate relations that may never have been considered in any of the primary studies. Consider, for example, a meta-analysis based on 120 studies of whether or not homework worked. Cooper (1989) found that students given homework assignments did better than those not given assignments, and furthermore, the more time spent on homework the better. An unexpected finding that had not been anticipated in the primary studies, however, was that although homework was very effective in high school it was almost totally ineffective in elementary school. This discovery has had important practical implications, and has stimulated the consideration of hypotheses as to why this might be, illustrating the impact of what Cooper has called 'review-generated evidence.'

(b) 'Cross-design synthesis' is a relatively new research strategy for combining the results from diverse, complementary studies, such as randomized controlled studies and administrative databases, to improve knowledge about interventions and to inform policy decisions. The goal is to capture the diverse strengths of the different designs while minimizing weaknesses. The approach is anchored in meta-analytic principles and methods, and in addition relies on methods for assessing the generalizability of data sources, and methods for statistical adjustment for potential biases in intervention comparisons (Droitcour et al. 1993).

(c) The goal of the Agency for Health Care Policy and Research (AHCPR) was to enhance the quality, appropriateness, and effectiveness of health services and to improve access to health care. The Federal legislation authorizing the creation of the agency (P.L. 101–239, December 1989) directed the agency to issue guidelines on clinical practice and required the guidelines to be based, in part, on a systematic synthesis of research evidence. Large-scale, AHCPR multidisciplinary studies called PORTS (Patient Outcomes Research Teams) as part of this mandate carried out formal meta-analyses to determine what was known, what was not known and what needed to be learned about variations in clinical practice and outcomes for a particular disease or condition (see, for example, Lehman et al. 1995).

(d) When individual subject-level data is available from a number of similar studies, meta-analytic techniques can be used to combine information across the studies. Not only can study specific effects be estimated but individual subject characteristics can be used as covariates. An example of the use of such a meta-analysis is the study of nearly 600 outpatients treated for acute depression, who ranged in age from 18 to 80 years, from six studies conducted during a 10-year period. The goal of this study was to investigate the differential efficacy of psychotherapy alone, in comparison with the combination of psychotherapy and pharmacotherapy (Thase et al. 1997).

(e) Given the body of research available and the quality of that research with respect to a particular question, it may be determined that doing a meta-analysis is premature for that field. Nevertheless, a research synthesis can be extremely useful in identifying obstacles in the literature that need to be addressed and successes that can be built on to guide future studies. For example, the Committee on the Assessment of Family Violence Interventions (1998) concluded that 'The quality of the existing research base of evaluations of family violence interventions is therefore insufficient to provide confident inferences to guide policy and practice, except in a few areas that we identify in this report. Nevertheless, this pool of evaluation studies and additional review articles represents a foundation of research knowledge that will guide the next generation of family violence evaluation efforts and allows broad lessons to be derived from the research literature.'

3. Suggestions for Further Reading

The articles *Meta-analysis: Overview* and *Meta-analysis: Tools* provide further details on analytical aspects of doing meta-analysis. For more information on the practice of meta-analysis, see Cook et al. (1992), Cooper and Hedges (1994a), Gaver et al. (1992), Hunt (1997), Rosenthal (1984), and Stangl and Berry (2000).

Bibliography

Chalmers I, Sackett D, Silagy C 1997 The Cochrane collaboration. In: Maynard A, Chalmers I (eds.) *Non-random Reflections on Health Services Research: On the 25th Anniversary of Archie Cochran's Effectiveness and Efficiency*. BMJ Books, London

Committee on the Assessment of Family Violence Interventions 1998 *Violence in Families: Assessing Prevention and Treatment Programs. A Report of the National Research Council/ Institute of Medicine*. National Academy Press, Washington, DC

Cook T D, Cooper H, Cordray D S, Hartmann H, Hedges L V, Light R J, Louis T A, Mosteller F 1992 *Meta-analysis for*

Explanation: A Casebook. Russell Sage Foundation, New York
Cooper H 1989 *Homework*. Longman, New York
Cooper H, Hedges L V (eds.) 1994a *The Handbook of Research Synthesis*. Russell Sage Foundation, New York
Cooper H, Hedges L V 1994b Research synthesis as a scientific enterprise. In: Cooper H M, Hedges L (eds.) *The Handbook of Research Synthesis*. Russell Sage Foundation, New York
Draper D, Hodges J S, Mallows C L, Pregibon D L 1993 Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A* **156**: 9–37
Droitcour J, Silberman G, Chemlimsky E 1993 Cross-design synthesis: A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *International Journal of Technology Assessment in Health Care* **9**: 440–9
DuMouchel W, Normand S-L 2000 Computer-modeling and graphical strategies for meta-analysis. In: Stangl D K, Berry D A (eds.) *Meta-analysis in Medicine and Health Policy*. Dekker, New York
Gaver D P, Draper D, Goel P K, Greenhouse J B, Hedges L V, Morris C N, Waternaux C 1992 *Combining Information: Statistical Issues and Opportunities for Research*. National Academy Press, Washington, DC
Greenhouse J B, Iyengar S 1994 Sensitivity analysis and diagnostics: Issues and methods. In: Cooper H M, Hedges L (eds.) *The Handbook of Research Synthesis*. Russell Sage Foundation, New York
Hunt M 1997 *How Science Takes Stock: The Story of Meta-analysis*. Russell Sage Foundation, New York
Lehman A F, Thompson J W, Dixon L B, Scott J E (eds.) 1995 *Schizophrenia Bulletin. Special Issue Theme: Schizophrenia: Treatment Outcomes Research* **21**: 561–676
Mosteller F, Chalmers T C 1992 Some progress and problems in meta-analysis of clinical trials. *Statistical Science* **7**: 227–36
Rosenthal R 1984 *Meta-analytic Procedures for Social Research*. Sage, Beverly Hills, CA
Rubin D B 1992 Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics* **17**: 363–74
Stangl D K, Berry D A (eds.) 2000 *Meta-analysis in Medicine and Health Policy*. Dekker, New York
Thase M E, Greenhouse J B, Frank E, Reynolds C F, Pilkonis P A, Hurley K, Grochocinski V, Kupfer D J 1997 Treatment of major depression with psychotherapy or psychotherapy–pharmacotherapy combinations. *Archives of General Psychiatry* **54**: 1009–15

J. B. Greenhouse

Meta-analysis: Overview

1. What is Meta-analysis?

1.1 The Origins of Meta-analysis

Meta-analysis (sometimes called 'quantitative synthesis' or 'overview analysis') is the term used to describe *quantitative* methods for combining information across different studies. The term 'meta-analysis' seems to have been coined by Glass (1976) to describe this idea of utilizing information in many studies of the same effect, although the concept itself is very much older (dating back at least to the 1930s, when it was studied by Fisher and Pearson).

Glass (1976) also introduced the idea of combining different summary statistics from different studies in a scale-free form (known as 'effect sizes'). Most commonly, in the sociological literature, these forms include standardized mean differences and correlation coefficients. These techniques extend the applicability of the concept of meta-analysis, since one then does not need identical measures of the effect in each of the studies considered.

The ideas have proved to be very powerful, and since their introduction there has been a veritable explosion in the use of such techniques, with the prime growth probably occurring in analysis of sociological effects and medical or epidemiological results.

Of course, it has long been common to see reviews of an area, with an expert bringing together the different information and synthesizing a conclusion from many disparate sources. Such overviews often contain qualitative or subjective impressions of the totality of information available. In this sense, the idea of combining study information is an old and appealing one, especially when considering subtle effects that might be hard to assess conclusively in one study. The key contributions of meta-analysis lie in various attempts to formalize this approach, and the term is usually reserved for the situation where one is combining numerical effect sizes from a collection of studies, rather than giving a more general nonquantitative overview.

1.2 The Goals of Meta-analysis

meta-analysis has become particularly popular in situations where the studies individually do not show that an effect is statistically significant. In this context, it is often the case that a combination of studies is more powerful in evaluating the effect. Methods can be divided into several groups:

(a) those which enable the overall significance of an effect to be evaluated, based on the multiple studies available;

(b) those which attempt to estimate an overall effect size θ, by combining the individual estimates in multiple studies in appropriate ways;

(c) those which evaluate the heterogeneity in a group of studies, often as a prelude to carrying out (a) and (b); and

(d) those which evaluate possible systematic biases in meta-analytic methods.

This article provides a brief overview of ideas in each of these groups, and illustrates them with an analysis of one specific example.

It must be stressed that underlying the general popularity of meta-analysis is the assumption that for

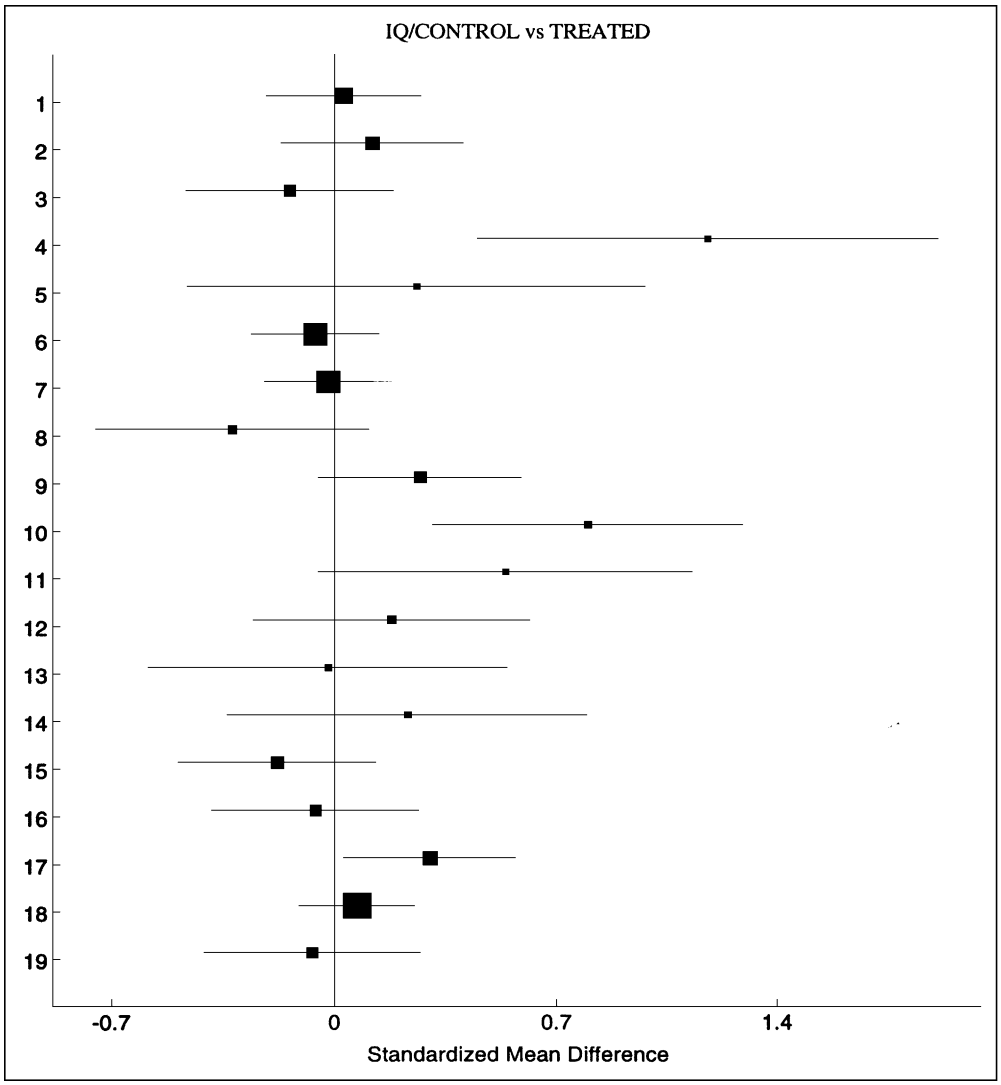

Figure 1
Ladder plot of 19 studies of student IQ modified by teacher expectations. Size of squares is proportional to accuracy of study

most effects of interest, there is some grand overall value of θ that can actually be found more accurately by summarizing the information in many studies. If this concept is correct, then the methods are of considerable power and value.

However, in reality the value of the estimated effect size in a particular study is conditional on many specific and nongeneralizable constraints pertaining in that study. To generalize, or even to combine, the values from such studies is not necessarily valid unless these constraints are reasonably satisfied across the range to which generalization is sought.

One option to deal with this is to develop linear or hierarchical models which incorporate covariate information (Hedges and Olkin 1985; DuMouchel 1990), or to use a response-surface approach as advocated in Rubin (1990). These methods often have the practical drawback that the information available on covariates in the individual studies is often sparse or nonexistent, and that many studies do no more than announce that the estimates available are 'adjusted' for some named or even un-named covariates.

Because of this partial knowledge of individual studies, it seems inevitable that many users will opt for

the simple methods of meta-analysis described below, and will then take the overall values as applying much more widely than is often justified. At the very least, such naive approaches should be tempered by a careful evaluation of the heterogeneity involved.

1.3 IQ Assessment: A Typical Meta-analysis

To carry out an ideal meta-analysis, there are a number of steps. We first identify an effect we wish to study. We then collect all of the studies on the subject, and combine their results using a method of meta-analysis. This then gives us both an overall measure of the relationship, and a statistical assessment of the significance of the relationship taking into account all the studies.

As an illustrative example we will consider a set of 19 randomized studies of the effects of teacher expectancy on later pupil performance on an IQ test, taken from Raudenbusch and Bryk (1985) and analyzed also in detail in various chapters of Cooper and Hedges (1994). In each study the 'treated' group consisted of students identified to their teachers as 'likely to experience substantial intellectual growth,' and the 'control' group was not so identified. In this case, the effect size for each study represents the mean IQ score of the treated group minus the mean of the control group, divided by a pooled standard deviation: that is, each effect size in this example is a standardized mean difference.

The data are illustrated in a 'ladder plot' in Fig. 1. This plot is typical of that used in meta-analyses. Each of these studies is represented with its mean effect size together with the associated 95 percent confidence interval (CI). The size of the means in this illustration is proportional to the precision, indicating which studies are likely to carry more weight.

As Fig. 1 shows, the various effect sizes do not give a conclusive picture. In some studies the effect sizes are positive, in others negative. Only three are statistically significant on the face of it. The goal of meta-analysis is to try to combine these in some effective way.

2. A Formal Overview of Meta-analysis

2.1 Testing Significance of an Effect

In the typical formalism for meta-analysis, we assume that study i provides a value T_i (the effect size), all of which are assumed to measure the same overall effect θ; and we also assume we know the standard error v_i of the ith effect size. One of the nontrivial aspects of a meta-analysis is the collection of these T_i, and care must be taken to extract appropriate information. It is easy to bias the outcome of a meta-analysis by careless selection of the T_i, and methods such as blinding of the extractors of the effect sizes, or multiple extraction by independent reviewers, are often used.

The first goal of meta-analysis is to try and decide whether or not an overall effect is significant. There are two common and simple approaches to this. The first is just vote counting: how many studies have positive and how many have negative effect sizes? If there is no overall effect this should be a binomial variable. In the IQ example of Fig. 1, such vote counting yields 11 out of 19 in favor of a positive effect, but clearly this is not significant ($p = 0.33$).

This simple-minded approach clearly fails to take into account any precision attached to each study. A somewhat more sophisticated idea, of 'combining p-values,' was introduced by Fisher (1932). Here the individual p-value is taken to encapsulate the information in the individual study, and all other aspects of the study are ignored in combining this information. If p_i is the p-value of the ith study, since $\chi = -2\sum_1^n \log p_i$ has a χ^2_{2n} distribution under the null hypothesis, χ can be used to assess whether the totality of the p-values leads to rejection or not. In the IQ example, this value is around 70 on 38 df, indicating now that the null hypothesis is rejected at the 99 percent level. It is worth noting that rounding in reporting of the effect sizes and variances can lead to surprising inaccuracies in the evaluation of χ; and that the significance here is very largely due to just two or three significant studies, with the negative and neutral studies having limited ability to overcome these strong positive results.

2.2 Combining Effect Sizes: Fixed Effects

While the combined p-value approach is valid under the appropriate assumptions of independence between studies, it has the clear drawback that it does not permit any estimate of an overall effect. Other approaches to meta-analysis use somewhat more information from the individual studies, and attempt to combine them in a sharper way. In the simplest such method (the so-called 'fixed effects' model), the effect size is assumed to be normally distributed, so that formally, for the ith study we assume

$$T_i = \theta + e_i \tag{1}$$

It is assumed that e_i are independent $N(0, v_i^2)$ random variables. To use this type of approach some steps may be needed to render the assumption of normality at least approximately correct: for example, if the effect size is a correlation, a transform to normality is required. In other cases effect sizes may be (possibly standardized) differences between means, as in the IQ example; in yet other cases, they may be the estimates of slopes in a regression model, and so on.

The goal now is to estimate the overall value θ, assuming the homogeneity inherent in (1). Typically it is assumed that the standard errors v_i are known accurately, in which case standard theory indicates

that θ is best estimated by

$$\bar{T}_\bullet = \left[\sum_i T_i/v_i^2 \right] \Big/ \left[\sum_i 1/v_i^2 \right]$$

which has variance

$$\sigma_\bullet^2 = \left[\sum_i 1/v_i^2 \right]^{-1}$$

In the IQ data, this fixed effects model leads to an overall estimate of the difference between means of $\bar{T}_\bullet = 0.06$, with a 95 percent CI of $(-0.01, 0.13)$. Although most of the individual studies are not significant, we have thus been able to use meta-analysis to establish that, at least on the face of it, the overall effect is significant at the 10 percent level, though not at the 5 percent level: not a strong result but one indicating that further study might well be warranted.

2.3 Combining Effect Sizes: Random Effects and Nonhomogeneity

The fixed effects model does not allow for heterogeneity between studies. When there is an indication that the studies are not homogeneous, it is common to combine estimates via a 'random effects' model (Draper et al. 1993), which attempts to allow for inter-study variation. In the random effects model we consider the formalization

$$\begin{aligned} T_i &= \theta_i + \varepsilon_i \\ \theta_i &= \mu_\bullet + \xi_i \end{aligned} \tag{2}$$

Here T_i is the observed effect size for each study, θ_i is the corresponding true i^{th} effect size, and it is assumed that ε_i are independent $N(0, v_i^2)$ random variables, that the ξ_i are independent $N(0, \tau^2)$ random variables, and that the $\xi\iota_i$ and ε_i are mutually independent. The fixed effects model takes $\tau^2 = 0$; by allowing $\tau^2 > 0$, the random effects model enables us to capture some of the inhomogeneity since it assumes different studies have mean values θ_i which may differ from $\mu_\bullet$.

In this case the meta-analysis estimator of $\mu_\bullet$ is given by

$$\bar{T}_\bullet^* = \left[\sum_i Ti/[v_i^2 + \hat{\tau}^2] \right] \Big/ \left[\sum_i 1/[v_i^2 + \hat{\tau}^2] \right]$$

which has variance

$$\sigma_\bullet^{2*} = \left[\sum_i 1/[v_i^2 + \hat{\tau}^2] \right]^{-1}.$$

There are various methods to give the estimator $\hat{\tau}^2$, the most common of which is the estimator of DerSimonian-Laird (1986), whose variance is given in Biggerstaff and Tweedie (1997).

The random effects model can also be analyzed in a Bayesian context, and extends logically to hierarchical models (DuMouchel 1990, Draper et al. 1993) by the addition of priors on θ and τ^2.

In the IQ data, this random effects model leads to an overall estimate of the difference between means of overall estimate of $\bar{T}_\bullet^* = 0.089$, with 95 percent CI $(-0.020, 0.199)$. The DerSimonian–Laird estimator of $\tau^2 = 0.026$, with a 95 percent CI (0.004, 0.095), indicates significant lack of homogeneity; and now we see that by allowing for this heterogeneity, the significance of the overall $\bar{T}_\bullet^*$ is at 11 percent.

This of course indicates a different conclusion from, say, the method of combining p-values. The difference is explained by the rather different rejection regions underlying the different methods, and in general the results from combining effect sizes will be preferred, as they use considerably more detailed information.

2.4 Combining Effect Sizes: Using Covariates

One further extension of Eqn. (2) is to incorporate covariates, in the form

$$\theta_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3}$$

where X_i is a vector of covariates in study i and β_1 is a vector of parameters.

This is attractive when the individual studies contain sufficient information to enable the model to be fitted, since it helps explain the variability in the random effects model.

In our IQ example, there exist data on the length of time (in weeks) that the teachers are exposed to the children being tested. When this is factored into the model, it is found that the estimate of $\beta_0 = 0.424$ and the estimate of $\beta_1 = -0.168$, and both are highly significant (see Cooper and Hedges 1994, Chap. 20). In this case the covariate appears to explain much of what is going on: in all but one of the negative results, the teacher had known the children for more than two weeks, but in only one of the positive studies was this the case. Thus without direct knowledge of children's abilities, there seems to be a real effect of the treatment; but direct knowledge mitigates this almost entirely.

3. Problems with Meta-analyses

3.1 Possible Difficulties

There are several provisos that need to be taken into account before accepting a formal summation of the studies as in the section above, and with the huge increase in the use of meta-analysis, there has come a

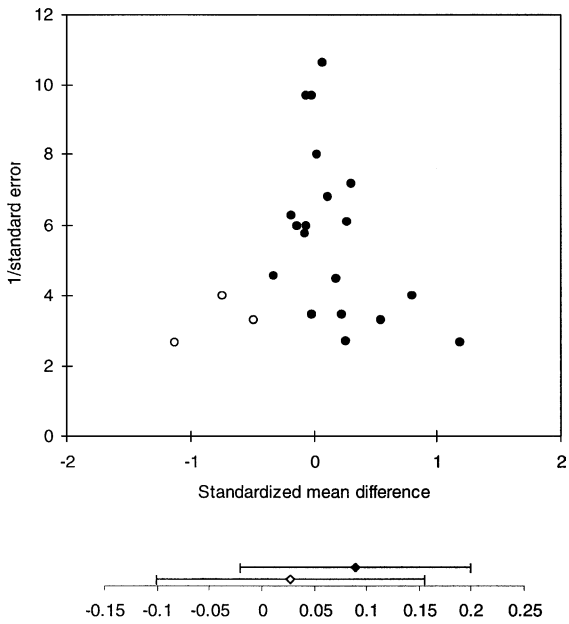

Figure 2
Possible publication bias in studies of teacher expectancy of IQ. Top panel is a funnel plot of standardized mean differences: the solid circles are original data, the open circles are three imputed 'missing studies.' Bottom panel shows overall mean and 95 percent CI before and after allowing for the missing studies

large number of books and discussion papers which assess the benefits, drawbacks, and problems of these techniques (Glass et al. 1981, Hedges and Olkin 1985, Draper et al. 1993, Cooper and Hedges 1994, Mengersen et al. 1995). Three of the key concerns that meta-analysis raises, and which differ from those in general statistical methodology, are:

(a) the problem of comparability of data and study design, since for the meta-analysis to be meaningfully interpreted, we must not combine 'apples and oranges';

(b) the effect of 'publication bias,' recognizing that failure to obtain all relevant studies, both published and unpublished, may result in a quite distorted meta-analysis; and

(c) the effect of different quality in different studies, so that one should not rely totally evenly on the studies used.

3.2 Are We Comparing Apples with Oranges?

Meta-analysis is designed to enable combination of results from studies which are comparable. The interpretation of comparability is a subjective and often difficult one. In order to paint an honest picture of the aims and applicability of any meta-analysis, we must first carefully define the relevant effect with which we are concerned, and ensure that all studies collected do address this same effect. This can be quite nontrivial.

In the IQ example, for instance, we would need to be sure that the tests for IQ were measuring similar attributes. Some comparability (at least of the scaling of the test) is provided by the standardization of the mean differences. We also need to be convinced that the concept of 'teacher expectancy' that we are evaluating is appropriately similar across the different studies, and from the written papers this is not always easy to decide.

There are three different methods one might suggest for handling such heterogeneity.

The first is by using models that specifically allow for such an effect, such as the random effects models discussed above. More subtly, to allow for the types of inhomogeneity with which we are concerned, Bayesian methods might well be used. In this context the priors on the various parameters can perhaps be thought of not as describing 'prior information' in any strong sense, but rather as describing in more detail the way in which the studies might be heterogeneous.

A second method of handling variability is by building more complex models where covariates are introduced. This is clearly preferable when it can explain the variability otherwise swept into τ^2 in the random effects model. There are some who advocate that the random effects model should never be used, but that one should instead search out appropriate covariates to explain heterogeneity between studies, and, as we have seen in the IQ example, this can be very fruitful. The drawback to this is that, since meta-analysis seeks to use the published results of studies without recourse to raw data which is often lost or unavailable, the user is often unable to use covariates since these are not published in sufficient detail.

A third (very simple) method used to account for heterogeneity is to give results separately for different subsets of the data which are thought to be heterogeneous, rather than to attempt to develop a parametric model for the effects of this stratification. This also is only possible if there are sufficient studies to allow reasonable estimation in each stratum.

3.3 Publication Bias

One of the most interesting phenomena in meta-analysis is 'publication bias.'

It is obviously important in principle in meta-analysis to attempt to collect all published and unpublished studies relevant to the relationship in question. The problem here is that unpublished studies, by their nature, are likely to differ from published studies. They are likely to be less significant, since journals differentially accept significant studies.

They are likely to show less 'interesting' effects, since such studies are often not submitted; or in the case of non-English speaking authors, are submitted only to local journals that are missed in scanning. Hence their omission from a meta-analysis may well bias the combined result away from the null value.

Missing studies due to publication bias are not easy to allow for. Unlike traditional missing data problems, there is an unknown number of them. Their effect could be huge, or it could be minute, and developing a sensitivity analysis that accounts for them is not trivial. Publication bias seems to be a new form of bias that needs new methods to account for it.

There are several ways used to evaluate the existence and effect of such bias. The first is the 'funnel plot,' introduced in Light and Pillemer (1984), which gives a good graphical indication of the possible existence of some forms of publication bias. If one plots the effect size against some measure of size of study, then under the normal assumptions of the fixed and random effects models, there should be symmetry around the true θ; and since (for practical reasons) there are generally more small studies than large ones, one should typically see a funnel or tree shape for the pattern of data. If the plot does not exhibit such symmetry then one might deduce that there are missing studies. This is illustrated on the IQ data in Fig. 2.

Such graphical indications are the most frequently used diagnostic for publication bias, but give little information of what difference the 'missing studies' might make. There are a number of rather complex approaches to this problem (Iyengar and Greenhouse 1988, Berlin et al. 1989, Dear and Begg 1992, Hedges 1992, Givens et al. 1997). In Duval and Tweedie (2000) a simpler method for handling such studies is developed which seems to give results consistent with more complex methods and quantifies the subjective impression given by using funnel plots.

For the IQ data, the methods in Duval and Tweedie (2000) estimate that the number of missing studies is around two to three, with positions as indicated in Fig. 2. Allowing for three such missing studies leads to a random effects estimate $\bar{T}^*_\bullet$ of 0.027 with 95 percent CI of $(-0.10, 0.16)$: that is, much of the observed estimate of θ might well be due to studies not included in the collection. Such a sensitivity analysis can aid in assuring that we do not become overconfident in assuming we have a full and correct estimate of the final answer.

3.4 Quality of Studies

Clearly different studies are of different quality, and there is considerable debate about whether to exclude studies that are unreliable.

A policy of deliberate exclusion of poor quality studies also helps in many cases to mitigate the problems of publication bias. If the studies that are not published are poor quality, which is quite conceivable, then there may be reasons for excluding them even if they exist on the fringes of research publication.

Some quality aspects are readily agreed on. For example, there is general concensus that studies which are randomized in some way are better than purely observational studies. In the medical literature, the Cochrane Collaboration, which is attempting to develop a full set of information on various diseases and treatments, will only accept into its base of studies for inclusion those studies which are randomized clinical trials. However, while there may be a rationale for only using (or conducting) randomized trials, in many sociological areas there is little possibility of using other than observational trials, and so this objective criterion for inclusion is not always of use.

There has been some work done on methods of allowing for quality (Cooper and Hedges 1994). Most of these methods involve weighting schemes, where the weighted averages in (1) are modified to depend, not just on the variances of the studies, but on other attributes of the studies. One such approach consists of drawing up lists of quality 'attributes' and then, based on a formal scoring of papers, to weight studies according to the quality.

The problem with most schemes for assessing and accounting for quality differences is their subjectivity. In general, it seems best to ensure that studies are included rather than having some excluded or down-weighted on grounds which are not clear and open. If there are real concerns about the quality of particular studies, then a viable alternative is to construct the analysis with and without these studies: as with many areas of meta-analysis, such sensitivity considerations can rapidly settle the role of the poor quality studies in the overall outcome.

4. Implementing Meta-analyses

4.1 Collecting Data

There is no formal way of ensuring that all sources of data have been covered in developing the values for a meta-analysis. However, most meta-analyses at least attempt to carry out searches of all relevant databases, and then work from this list to a wider search. In the IQ example there are various sources of literature (relevant to many other sociological and educational meta-analyses) that might be formally searched: for example, the ERIC (Educational Resources Informational Center) database, *PsycINFO* and *Psychological Abstracts*, or *Sociological Abstracts*. The list will vary from field to field.

As well as such formal and usually computerized searches, it is also valuable to use other more informal methods. Following up on references in articles already found (especially review articles), consideration of citation indexes, and general conversations and

communications with others in the field will also assist in locating studies. In particular the last form, informal followup, is perhaps the best method for finding the otherwise missing studies, unpublished theses, or non-mainstream articles whose omission may lead to publication bias.

In all cases, it is imperative that the meta-analyst gives a full and detailed description of the search process used and the actual data derived. This is of particular importance in situations where the basic article gives more than one summary statistic that might be used.

4.2 Software for Calculations

In order to apply the ideas above it would be ideal to point the potential user to appropriate software that could carry out the full range of meta-analysis. The ideal software is not yet available, although there are many homegrown versions in the statistical literature, with a variety of features, not all of them intuitively easy and—rather more problematically—not all of them giving correct results.

Methods to use SAS or BUGS to carry out both frequentist and Bayesian meta-analyses are described in Normand (1999) and DuMouchel and Normand (2000), while a range of recent SAS macros is described in Wang and Bushman (1999). The Cochrane Collaboration, which aims to become a full registry of studies in clinical trials areas, also has developed some analytic software, although this has to date only been available for studies in their collection. Various commercial software packages are currently under development which have many of the desirable features required by the non-expert.

Nonetheless, there is still a long way to go before meta-analysis can be carried out totally routinely.

4.3 Conclusions

The IQ example with which this overview is illustrated indicates many of the advantages and some of the pitfalls of implementing a meta-analysis.

The advantages are threefold. We have been able to establish that, despite the existence of positive and negative studies, the overall effect is positive. We have found that, when lack of homogeneity is taken into account, the positive effect is not yet known to be statistically significant. And we have seen that the influence of covariates in these datasets may well be crucial, so that when they are taken into account, a much more clearcut picture appears.

The pitfalls are several. We have seen that the simplistic use of voting procedures, combined p-values or fixed effects models may give conflicting answers, and much thought needs to go into deciding how to use random effects or possibly Bayesian models in these circumstances. We have indicated that on the face of it, there may well be publication bias in this dataset, and that this might account for much of the observed overall effect.

The implementation of this series of meta-analyses used a number of one-off pieces of software, for analysis and for graphical presentation. As this example shows, however, even when the mathematical methodology becomes routine to implement, there will still be a need for the practitioner to take every precaution to ensure that the results really do reflect a coherent picture of the overall effect being evaluated.

5. Further Reading

Detailed reviews of almost all aspects of meta-analysis are given in the books of Glass et al. (1981), Hedges and Olkin (1985), Rosenthal (1991), Cooper and Hedges (1994), and Sutton et al. (2000). A very readable account of the general problems of combining information is in Light and Pillemer (1984).

See also: Explanation: Conceptions in the Social Sciences; Meta-analysis in Practice; Meta-analysis: Tools; Quantification in the History of the Social Sciences

Bibliography

Berlin J A, Begg C B, Louis T A 1989 An assessment of publication bias using a sample of published clinical trials. *Journal of the American Statistical Association* **84**: 381–92

Biggerstaff B J, Tweedie R L 1997 Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* **16**: 753–68

Cooper H, Hedges L V (eds.) 1994 *The Handbook of Research Synthesis*. Russell Sage Foundation, New York

Dear K B G, Begg C B 1992 An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science* **7**: 237–45

DerSimonian R, Laird N 1986 Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**: 177–88

Draper D, Gaver D P, Goel P K, Greenhouse J B, Hedges L V, Morris C N, Tucker J R, Waterman C 1993 *Combining Information: Statistical Issues and Opportunities for Research*. American Statistical Association, Washington DC

DuMouchel W H, Normand S-L T 2000 Computer modeling and graphical strategies for meta-analysis. In: Stangl D K, Berry D A (eds.) *Meta-analysis in Medicine and Health Policy*. Marcel Dekker, New York, pp. 127–78

DuMouchel W H 1990 Bayesian meta-analysis. In: Berry D (ed.) *Statistical Methods for Pharmacology*. Marcel Dekker, New York, pp. 509–29

Duval S J, Tweedie R L 2000 A non-parametric 'trim and fill' method of assessing publication bias in meta-analysis. *Journal of the American Statistical Association* **95**: 89–98

Fisher R A 1932 *Statistical Methods for Research Workers*, 4th edn. Oliver and Boyd

Givens G H, Smith D D, Tweedie R L 1997 Publication bias in meta-analysis: a Bayesian data-augmentation approach to

account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science* **12**: 221–50
Glass G V 1976 Primary, secondary, and meta-analysis of research. *Educational Researcher* **5**: 3–8
Glass G V, MgGaw B, Smith M L 1981 *Meta-analysis of Social Research*. Sage Publications, Newbury Park, CA
Hedges L V, 1992 Modeling publication selection effects in Meta Analysis. *Statistical Science* **7**: 246–55
Hedges L V, Olkin I 1985 *Statistical Methods for meta-analysis*. Academic Press, Orlando, FL
Iyengar S, Greenhouse J B 1988 Selection models and the file drawer problem. *Statistical Science* **3**: 133–35
Light R J, Pillemer D B 1984 *Summing Up: the Science of Reviewing Research*. Harvard University Press, Cambridge, MA
Mengersen K L, Tweedie R L, Biggerstaff B J 1995 The impact of method choice in meta-analysis. *Australian Journal of Statistics* **37**: 19–44
Normand S-L T 1999 Meta-analysis: Formulating, evaluating, combining and reporting. *Statistics in Medicine* **18**: 321–59
Raudenbush S W, Bryk A S 1985 Empirical Bayes Meta Analysis. *Journal of Educational Statistics* **106**: 75–98
Rosenthal R 1991 *Meta-analysis Procedures for Social Research*. Sage Publications, Newbury Park
Rubin D 1990 A new perspective. In: Wachter K W, Straf M L (eds.) *The Future of meta-analysis*. Russell Sage Foundation, New York, pp. 155–66
Wang M C, Bushman B J 1999 *Integrating Results through Meta Analytic Review Using SAS Software*. SAS Institute, Cary, NC

R. L. Tweedie

Meta-analysis: Tools

The replication of results is a hallmark of the scientific process. Some form of research synthesis is essential to determine whether results from different studies are consistent, and to summarize the evidence. Meta-analysis is the use of statistical procedures in that endeavor; it differs from secondary analysis in that the information comes from statistical summaries of the data in the original studies and not directly from the raw data (Glass 1976). Meta-analysis is a fundamental scientific activity that is practiced in all of the empirical sciences (Draper et al. 1993). In meta-analysis, the results of individual studies are usually represented by a measure of effect size, and an average or typical effect size is estimated. The analysis may also assess variation in results across studies. This article provides an introduction to the analytic tools most widely used in meta-analysis.

1. *Effect Sizes*

An effect-size measure is used to represent the outcome of each study in a meta-analysis. Effect sizes make the results of different studies comparable, even though there may be differences in the ways the studies collect or analyze data. The appropriate effect size depends on the design of the studies and the nature of the outcome variable. In experimental studies that contrast treatments and employ continuous outcome variables, the standardized mean difference (the difference between the treatment and control group means divided by the within-group standard deviation) is often used. In experimental studies with dichotomous outcomes, the effect sizes are typically odds ratios or risk ratios. When both the independent and dependent variables are continuous, correlation coefficients or standardized regression coefficients are appropriate effect sizes. Each study in a meta-analysis provides both an estimate of effect size and a standard error that indicates the sampling uncertainty of the estimate. In most cases, effect-size estimates are approximately normally distributed; transformations such as the log of odds ratios are often used to improve the normal approximation (see *Statistical Analysis, Special Problems of: Transformations of Data*).

2. *Combining Effect-size Estimates Across Studies to Estimate an Average Effect*

Estimation of the average effect size across studies involves combining the individual effect-size estimates from different studies, usually by averaging. Because some studies (i.e., those with larger sample sizes) produce estimates with lower uncertainty than others, it is common practice to weight the average proportionally to the inverse of the variance (squared standard error) of each effect-size estimate. Weighting is used to increase precision of the combined estimate; it has no effect on bias.

2.1 *Inference Models*

There are two sorts of questions that meta-analysis can address. The analyst may inquire about the parameters underlying the particular set of studies observed in the meta-analysis, or there may be more general interest in a putative population of studies that *could* exist. These questions require different inference models, designated *conditional* and *unconditional* (Hedges and Vevea 1998).

The *conditional* inference model involves generalizations to parameters contained in the set of studies that are actually observed in the meta-analysis (or in studies that are, *a priori*, equivalent to them). For example, conditional inference about the mean effect size in a set of studies addresses the mean of the effect size parameters in these particular studies; the effect sizes of any other studies are conceptually irrelevant. The only uncertainty in the conditional inference model arises because the sample effect-size estimate in each study is not identical to the effect-size parameter that it estimates; rather, there is random variation associated with the sampling of persons (or other units) in the studies. Conditional inference may be appropriate, for

example, when a particular set of studies has been commissioned to answer a particular question (e.g., a regulatory question), and these studies provide the entire evidentiary base on which the question is to be determined.

In contrast, the *unconditional* inference model involves inferences about the parameters of a universe or hyperpopulation of studies from which the observed studies are a sample. The hypothetical other studies in this universe are conceptually relevant. The observed studies are regarded as a sample from that universe, which implies that there is additional uncertainty arising from the process of sampling studies. Thus, the effect-size estimate of a study involves a component of uncertainty associated with the sampling of units into that study, and an additional component associated with the sampling of the study itself. The unconditional inference model can be seen as a Bayesian model where the hyperpopulation of studies is conceived as generating a prior distribution of study effect size parameters (see *Bayesian Statistics*). The unconditional model may be appropriate when combining evidence about a treatment that is implemented somewhat differently in every study, giving rise to a distribution of treatment effects corresponding to variation in implementation.

Inference models are distinguished from the statistical procedures described below because the type of inference intended is logically prior and should determine the statistical procedures that are used. Fixed-effects statistical procedures are appropriate for conditional inferences and random-effects or Bayesian procedures are appropriate for unconditional inferences.

2.2 Fixed-effects Statistical Procedures

Fixed-effects statistical procedures are designed to make conditional inferences. Let $\theta_1,\ldots,\theta_k$ be the effect-size parameters from k studies; let $T_1,\ldots,T_k$ be the corresponding estimates observed in the studies, and let $v_1,\ldots,v_k$ be the variances (squared standard errors) of those estimates. These v_i are called the *conditional* error variances to emphasize that they are the variances of T_i *conditional* on θ_i.

Assume that the T_i are independent and that each T_i is approximately normally distributed given θ_i. Note that the normality assumption may not be realistic in all situations. Where it is not, other procedures are needed that are adapted to the particular situation. If the effect-size parameters are equal or nearly so (that is, if $\theta_1 = \ldots = \theta_k = \theta$), then the fixed-effects estimate of θ is the weighted mean $\bar{T}_\bullet$, given by

$$\bar{T}_\bullet = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \tag{1}$$

where the weight $w_i = 1/v_i$. The variance of $\bar{T}_\bullet$ is the reciprocal of the sum of the weights

$$\sigma_\bullet^2 = \frac{1}{\sum_{i=1}^{k} w_i} \tag{2}$$

and the standard error $SE(\bar{T}_\bullet)$ of $\bar{T}_\bullet$ is the square root of $\sigma_\bullet^2$; that is, $SE(\bar{T}_\bullet) = \sigma_\bullet$. $T_1, \ldots, T_k$ are normally distributed; it follows that the linear combination $\bar{T}_\bullet$ is also normally distributed.

If $T_1, \ldots, T_k$ all estimate the same underlying effect size, that is if $\theta_1 = \ldots = \theta_k = \theta$, then $\bar{T}_\bullet$ estimates θ, and a $100(1-\alpha)$ percent confidence interval for θ is given by

$$\bar{T}_\bullet - C_{\alpha/2}\sigma_\bullet \leqslant \theta \leqslant \bar{T}_\bullet + C_{\alpha/2}\sigma_\bullet \tag{3}$$

where $C_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution (e.g., $C_{\alpha/2} = 1.96$ for $\alpha = .05$) and $\sigma_\bullet^2$ is the sampling variance of $\bar{T}_\bullet$ (Eqn. (2)). If the confidence interval does not include zero, we reject the hypothesis that $\theta = 0$ at the 100α percent significance level (see *Hypothesis Testing in Statistics*; *Estimation: Point and Interval*).

To assess the heterogeneity of effect parameters in fixed-effects models, a test of the hypothesis that $\theta_1 = \ldots = \theta_k$ is often employed. The test uses the statistic

$$Q = \sum_{i=1}^{k} w_i (T_i - \bar{T}_\bullet)^2 \tag{4}$$

which has a chi-square distribution with $k-1$ degrees of freedom when the effect size parameters are the same in all studies. Large values of Q lead to rejection of the hypothesis that the effect sizes are homogeneous.

2.3 Random-effects Statistical Procedures

Random-effects statistical procedures are designed to make unconditional inferences. That is, they allow inferences about effect-size parameters in a universe of studies from which the observed studies are a sample. Use the same notation for effect-size parameters, their estimates, and the sampling error variances of the estimates. Let $\mu_\bullet$ be the mean of the population from which the effect-size parameters $\theta_1, \ldots, \theta_k$ were sampled; let τ^2 be the variance of that mean. The model for the effect size estimates is

$$T_i = \theta_i + \varepsilon_i = \mu_\bullet + \xi_i + \varepsilon_i \tag{5}$$

where $\xi_i = \theta_i - \mu_\bullet$ is a study-specific random effect and $\varepsilon_i = T_i - \theta_i$ is the sampling error as in the fixed-effects model. Thus τ^2 is the variance of ξ_i and represents a component of variation due to between-study differences in effect parameters.

Although statistical procedures can be more complex in the case of random-effects models, the most frequently used procedure estimates $\mu_\bullet$ by a weighted

mean of the effect-size estimates using slightly different weights from those employed in the fixed-effects procedure. Here, the weights include an additional component of variation, τ^2, which describes between-study variation in effect size parameters. This variance component is estimated by

$$\hat{\tau}^2 = \frac{Q-(k-1)}{c} \tag{6}$$

where c is given by

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} (w_i)^2}{\sum_{i=1}^{k} w_i} \tag{7}$$

and Q is given in Eqn. (4). When Eqn. (6) produces a negative result, the estimate is set to zero. The unconditional variance v_i^* of T_i is

$$v_i^* = v_i + \hat{\tau}^2 \tag{8}$$

where the asterisk denotes that this is the unconditional or random-effects variance. The weights used to compute the weighted mean are $w_i^* = 1/v_i^*$. The random-effects weighted mean is given by

$$\bar{T}_\bullet^* = \frac{\sum_{i=1}^{k} w_i^* T_i}{\sum_{i=1}^{k} w_i^*} \tag{9}$$

Note that $\bar{T}_\bullet^*$ depends on $\hat{\tau}^2$ because w_i^* does. The variance of $\bar{T}_\bullet^*$ is the reciprocal of the sum of the random-effects weights, that is,

$$\sigma_\bullet^{*2} = \frac{1}{\sum_{i=1}^{k} w_i^*} \tag{10}$$

The standard error $SE(\bar{T}_\bullet^*)$ of the mean effect estimate $\bar{T}_\bullet^*$ is the square root of its sampling variance: $SE(\bar{T}_\bullet^*) = \sigma_\bullet^*$. If both the effect-size estimates and the effect-size parameters are normally distributed, then $\bar{T}_\bullet^*$ is also normally distributed.

An approximate $100(1-\alpha)$ percent confidence interval for the mean effect $\mu_\bullet$ is given by

$$\bar{T}_\bullet^* - C_{\alpha/2}\sigma_\bullet^* \leqslant \mu_\bullet \leqslant \bar{T}_\bullet^* + C_{\alpha/2\sigma_\bullet^*} \tag{11}$$

where $C_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution and $\sigma_\bullet^{*2}$ is the variance of $\bar{T}_\bullet^*$ given in Eqn. (10). If the confidence interval does not include zero, one rejects the hypothesis that $\mu_\bullet = 0$ at the 100α percent significance level.

The between-studies variance component τ^2 is the natural characterization of between-study heterogeneity in random-effects models. The test for heterogeneity using the Q statistic described in connection with fixed-effects models is also a test of the hypothesis that $\tau^2 = 0$.

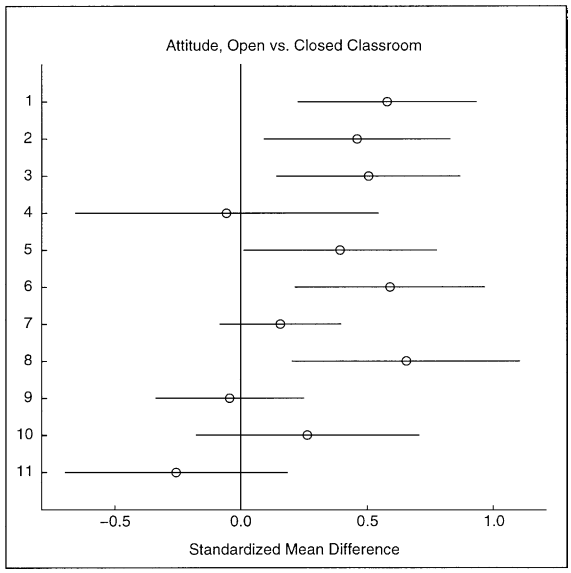

Figure 1
Student attitude effect size estimates

2.4 Bayesian Statistical Procedures

The random-effects procedure above gives an estimate of the average effect size that depends on $\hat{\tau}^2$, which is estimated from the data. When the number of studies is small, the uncertainty of this estimate can be rather large, and a variety of values of τ^2 are almost equally consistent with the data. Bayesian procedures for estimating the average effect size essentially compute the random-effects estimate of the mean conditional on τ^2, and average these conditional means, giving weight to each that is proportional to the posterior probability of τ^2 given the data. Bayesian procedures can also provide improved estimates of individual study effect sizes using information from the entire collection of effect sizes, and have some logical advantages for unconditional inference (for more information see DuMouchel 1990 or Hedges 1998).

2.5 Fixed and Random Effects: An Example

Figure 1 depicts effect sizes with 95 percent confidence intervals from a typical small meta-analysis. The data are taken from a study of the effects of open education on student attitudes toward school (Hedges et al. 1981). Positive effect sizes correspond to improve-

ments in attitude associated with open education. The studies are equivocal with respect to the effect of open education on attitude, with five of the 11 confidence intervals including zero.

In this data set, the fixed-effects estimate, $\bar{T}_{\bullet}$, of the population effect (from Eqn. (1)) is 0.288. The standard error of that estimate (Eqn. (2)) is 0.056, so that a 95 percent confidence interval for θ (Eqn. (3)) is 0.178–0.398. However, the Q statistic (Eqn. (4)) for this data set is 23.459, which is distributed as a chi-square on 10 degrees of freedom if the effects are homogeneous. The test rejects homogeneity, $p < 0.01$. That is, the effects appear to be heterogeneous.

The estimated variance component $\hat{\tau}^2$ (Eqn. (6)) is 0.047. When that additional component of variation is incorporated in the weights, the resulting random-effects estimate of $\mu_{\bullet}$ (Eqn. (9)) is 0.303. The corresponding standard error (Eqn. (10)) is 0.088, and a 95 percent confidence interval for the mean of the distribution of true effects (Eqn. (11)) is 0.131–0.475. Note that the precision of estimation is lower in the random-effects analysis.

3. *Modeling Between-study Variation in Effect Sizes*

It is often desirable to examine the relation between effect sizes and study-level covariates that may represent methodologically or substantively important characteristics of the studies (such as treatment intensity or dosage, treatment type, study design, subject population). Such modeling must be considered correlational, since there is no random assignment of studies to have different characteristics. Care should be employed in the interpretation of patterns of relations between study-level covariates and effect size. It is not uncommon for several covariates to explain essentially the same pattern of variation, making interpretations ambiguous.

Just as in the case of estimating the mean effect across studies, there are two inference models: conditional (which allows inferences about the parameters of the particular studies observed) and unconditional (which allows inferences about the parameters of studies from which the studies observed are considered a sample). In either inference model, the inferences are usually considered conditional on the values of the study level covariates observed (see *Linear Hypothesis: Regression (Basics)*; *Linear Hypothesis*).

3.1 *Fixed-effects Models*

In describing the modeling of between-study variation, it is helpful to use matrix notation. Denote the k-dimensional vectors of population and sample effect sizes by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ and $\boldsymbol{T} = (T_1, \ldots, T_k)'$, respectively. Suppose that the effect-size parameter θ_i for the ith study depends on a vector of $p \leqslant k$ fixed study-level covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ such as treatment dosage or length of follow-up. Specifically, assume that

$$\theta_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \quad i = 1, \ldots, \quad k \tag{12}$$

where $\beta_1, \ldots, \beta_p$ are unknown regression coefficients.

The matrix

$$\mathbf{X} = \begin{matrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \cdots & \mathbf{x}_{1p} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \cdots & \mathbf{x}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{k1} & \mathbf{x}_{k2} & \cdots & \mathbf{x}_{kp} \end{matrix} \tag{13}$$

is called the design matrix in regression analysis, and is assumed to have no linearly dependent columns; that is, $\mathbf{X}$ has rank p. The set of equations can be written succinctly in matrix notation as

$$\boldsymbol{\theta} = \mathrm{X}\boldsymbol{\beta} \tag{14}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the p-dimensional vector of regression coefficients.

The model for $\mathbf{T}$ can be written as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{15}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k)' = \mathbf{T} - \boldsymbol{\theta}$ is a k-dimensional vector of residuals. The linear model $\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ for the effect sizes is analogous to the model that is the basis for ordinary least squares regression analysis. Because $\boldsymbol{\varepsilon} = \mathbf{T} - \boldsymbol{\theta}$, the distribution of ε is approximately k-variate normal with means zero and diagonal covariance matrix $\mathbf{V}$ given by

$$\mathrm{V} = \mathrm{Diag}(v_1, v_2, \ldots, v_k) \tag{16}$$

Because the elements of ε are independent but not identically distributed, we use the method of weighed least squares to obtain an estimate of β.

The generalized least-squares estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{T} \tag{17}$$

which has a normal distribution with mean β and covariance matrix $\boldsymbol{\Sigma}$ given by

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{18}$$

The distribution of $\hat{\boldsymbol{\beta}}$ can be used to obtain tests of significance or confidence intervals for components of $\boldsymbol{\beta}$. If σ_{jj} is the jth diagonal element of $\boldsymbol{\Sigma}$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ then a $100(1-\alpha)$ percent confidence interval for β_j, $1 \leqslant j \leqslant p$, is given by

$$\hat{\beta}_j - C_{\alpha/2}\sqrt{\sigma_{jj}} \leqslant \beta_j \leqslant \hat{\beta}_j + C_{\alpha/2}\sqrt{\sigma_{jj}} \tag{19}$$

where $C_{\alpha/2}$ is the 100α percent two-tailed critical value of the standard normal distribution.

The error sum of squares statistic

$$Q_E = \mathbf{T}'[\mathbf{V}^{-1} - \mathbf{X}\mathbf{V}^{-1}(\mathbf{X}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{T} \quad (20)$$

tests the goodness of fit of the regression model. When the model is correctly specified (that is, when $\theta = \mathbf{X}\boldsymbol{\beta}$), Q_E has a chi-square distribution with $k-p$ degrees of freedom, so that large values of Q_E indicate that fit is poor.

3.2 Mixed-effects Models

As in the fixed-effects model, we assume that each T_i is normally distributed about θ_i with (conditional) variance v_i. However, in the case of the mixed model, the predictor variables are not assumed to explain all of the variation in θ_i. Rather, there is an additional component of variance τ^2 that is not explained by the p study-level covariates $\mathbf{X}_1, \ldots, \mathbf{X}_p$. The linear model is

$$\theta_i = \beta_1 x_{i1} + \cdots + \beta_{ip} + \xi_i \quad (21)$$

where $x_{i1}, \ldots, x_{ip}$ are the values of the predictor variables $X_1, \ldots, X_p$ for the ith study, $\beta_1, \ldots, \beta_p$ are unknown regression coefficients, and ξ_i is a random effect with variance τ^2.

The $k \times p$ design matrix $\mathbf{X}$, and the k-dimensional vectors of population and sample effect sizes $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ and $\mathbf{T} = (T_1, \ldots, T_k)'$, are the same as in the fixed-effects model. The set of equations can be written in matrix notation as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \quad (22)$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$ is the p-dimensional vector of regression coefficients, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)'$ is the k-dimensional vector of study-specific random effects. The model for the observations $\mathbf{T}$ can be written as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} \quad (23)$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k)' = \mathbf{T} - \boldsymbol{\theta}$ is a k-dimensional vector of residuals of $\mathbf{T}$ about θ, and $\eta = \xi + \varepsilon$ is a k-dimensional vector of residuals of $\mathbf{T}$ about $\mathbf{X}\boldsymbol{\beta}$.

The usual estimator of the residual variance component is given by

$$\hat{\tau}^2 = (Q_E - k + p)/c \quad (24)$$

where Q_E is the residual sum of squares from the *fixed-effects* weighted regression (Eqn. (20)). The constant c is given by

$$c = \mathrm{tr}(\mathbf{V}^{-1}) - \mathrm{tr}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-2}\mathbf{X}] \quad (25)$$

where $\mathbf{V} = \mathrm{Diag}(v_1, \ldots, v_k)$ is a $k \times k$ diagonal matrix of conditional variances and $\mathrm{tr}(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$.

The distribution of $\boldsymbol{\eta} = \boldsymbol{\xi} + \boldsymbol{\varepsilon}$ has mean zero and a diagonal covariance matrix given by $\mathrm{Diag}(v_1 + \tau^2, v_2 + \tau^2, \ldots, v_k + \tau^2)$. Using the estimate $\hat{\tau}^2$ in place of τ^2, one may compute the $k \times k$ diagonal matrix $\mathbf{V}^*$, an estimate of the unconditional covariance matrix of $\mathbf{T}$, by

$$\mathbf{V}^* = \mathrm{Diag}(v_1 + \hat{\tau}^2, v_2 + \hat{\tau}^2, \ldots, v_k + \hat{\tau}^2) \quad (26)$$

The generalized least squares estimator of $\boldsymbol{\beta}$ under the model using the estimated covariance matrix $\mathbf{V}^*$ is

$$\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{*-1}\mathbf{T} \quad (27)$$

which is approximately normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix Σ^* given by

$$\boldsymbol{\Sigma}^* = (\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1} \quad (28)$$

The distribution of $\boldsymbol{\beta}^*$ can be used to obtain tests of significance or confidence intervals for components of $\boldsymbol{\beta}$. If σ^*_{jj} is the jth diagonal element of Σ^*, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, then an approximate $100(1-\alpha)$ percent confidence interval for β_j, $1 \leqslant j \leqslant p$, is given by

$$\beta^*_j - C_{\alpha/2}\sqrt{\sigma^*_{jj}} \leqslant \beta_j \leqslant \beta^*_j + C_{\alpha/2}\sqrt{\sigma^*_{jj}} \quad (29)$$

where $C_{\alpha/2}$ is the 100α percent two-tailed critical value of the standard normal distribution.

3.2.1 Relation to hierarchical linear models. The mixed models considered here are related to the hierarchical linear model, a special case of the general mixed linear model, which finds wide application in the social sciences (see *Hierarchical Models: Random and Fixed Effects*; also Goldstein 1987, Bryk and Raudenbush 1992). There has been considerable progress in developing software to estimate and test the statistical significance of parameters in these models. In these models, the data are regarded as hierarchically structured, and a model is defined for each level of the hierarchy. In meta-analysis, level I is that of the study and the model for level I is

$$T_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, k \quad (30)$$

where ε_i is a sampling error of T_i as an estimate of θ_i. Level II of the model describes between-study variation in the study-specific effect-size parameters (the θ_i). The linear model above implies a level II model like

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \xi_i \quad (31)$$

where ξ_i is a study-specific random effect. Most of the attention in estimating these models has focused on the case where both the ε_i and the ξ_i are normally distributed with zero mean and unknown variances, that is $\varepsilon_i \sim N(0, v)$ and $\xi_i \sim N(0, \tau^2)$, $i = 1, \ldots, k$.

There are two important differences between the hierarchical linear models usually studied and the model used in meta-analysis. The first is that, in meta-analytic models, the variances of the sampling errors $\varepsilon_1, \ldots, \varepsilon_k$ are *not* identical across studies. The sampling error variances depend on various aspects of study

design (particularly sample size) that are rarely constant across studies. The second is that the sampling error variances in meta-analysis are assumed to be known. Therefore the model used in meta-analysis is a special case of the general hierarchical linear model where the level I variances are unequal, but known. Software for the analysis of hierarchical linear models can be used for mixed model meta-analysis if it permits the specification of first-level variances that are unequal but known, as do to the programs HLM (Raudenbush et al. 2000) and SAS PROC MIXED (SAS 6.12 1996).

Methods for Bayesian analyses of hierarchical mixed linear models are also available (see Seltzer et al. 1996 for a general treatment or DuMouchel 1990 or Hedges 1998 for applications to meta-analysis).

3.3 Fixed- and Mixed-effects Explanatory Models: An Example

In the data set depicted in Fig. 1 (and employed in Sect. 2.5 to illustrate fixed- and random-effects analyses), the first three effects are from studies of kindergarten through third-grade classrooms, and the remaining effects are from grades four through six. The tendency for the K-3 effects to be positive warrants further investigation. If we construct a linear model that incorporates an indicator of grade level as a predictor, the fixed-effects estimates of the intercept (K-3 effect) and slope (which represents a decrement associated with the higher grade levels) from Eqn. (17) are 0.517 and -0.316. The standard error of the slope (from Eqn. (18)) is 0.122, so that a 95 percent confidence interval for the difference between the K-3 effect and the 4–6 effect (Eqn. (19)) is -0.555 to -0.077. Since that interval fails to include zero, we can reject at the 0.05 level the null hypothesis that there is no difference between grade levels. However, Q_E, the test of model fit (Eqn. (20)), is 16.997 on nine degrees of freedom, $p < 0.05$. Hence, we conclude that, even though the explanatory variable is significant, the effects are still heterogeneous.

The estimated variance component (Eqn. (24)) is 0.028. Note that this is substantially lower than the estimate of 0.047 in the simple random-effects analysis (Sect. 2.5). That is because a part of the variability that was absorbed by the variance component has now been attributed to the grade-level indicator. The random-effects estimates of the slope and intercept in this mixed model are 0.517 and -0.302, and the standard error of the slope is 0.124. A 95 percent confidence interval for the decrement associated with being in a higher grade is now -0.635 to 0.031. Hence, if our interests demand unconditional inference, we would not reject the null hypothesis that the two age groups exhibit the same average attitude change.

4. Conclusions and Suggestions for Further Reading

There are two basic inference models in meta-analysis (conditional and unconditional) and two classes of statistical procedures used in combining effect size estimates (fixed- and mixed-effects procedures). Bayesian procedures are closely related to mixed models. In either case, statistical methods that weight estimates by the inverse of the conditional or unconditional variance are used for analysis. Typical analyses involve computing an overall mean effect size, assessing heterogeneity, and modeling variation in effect sizes across studies.

The articles *Meta-analysis: Overview* and *Meta-analysis in Practice* provide further introduction to meta-analysis. For more information on the general process of integrating research and related statistical tools, the reader may find useful: Cooper (1989), Cooper and Hedges (1994), Hedges and Olkin (2001), Rosenthal (1991), and Strangl and Berry (2000).

Bibliography

Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models*. Sage Publications, Newbury Park, CA

Cooper H M 1989 *Integrating Research: A Guide to Literature Reviews*. Sage Publications, Newbury Park, CA

Cooper H M, Hedges L V 1994 *The Handbook of Research Synthesis*. The Russell Sage Foundation, New York

Draper D, Gaver D P, Goel P K, Greenhouse J B, Hedges L V, Morris C N, Tucker J R, Waternaux C 1993 *Combining Information: Statistical Issues and Opportunities for Research*. American Statistical Association, Washington, DC

DuMouchel W H 1990 Bayesian meta-analysis. In: Berry D (ed.) *Statistical Methods for Pharmacology*. Marcel Dekker, New York, pp. 509–29

Glass G V 1976 Primary, secondary, and meta-analysis. *Educational Researcher* **5**: 3–8

Goldstein H 1987 *Multi-Level Models in Educational and Social Research*. Oxford University Press, London

Hedges L V 1998 Bayesian approaches to meta-analysis. In: Everitt B, Dunn G (eds.) *Recent Advances in the Statistical Analysis of Medical Data*. Edward Arnold, London, pp. 251–75

Hedges L V, Giaconia R M, Gage N L 1981 *Meta-analysis of the Effects of Open and Traditional Instruction*. Stanford University Program on Teaching, Effectiveness Meta-Analysis Project, Final Report, Vol. 2, Stanford, CA

Hedges L V, Olkin I 2001 *Statistical Methods for Meta-Analysis in the Medical and Social Sciences*. Academic Press, New York

Hedges L V, Vevea J L 1998 Fixed- and random-effects models in meta-analysis. *Psychological Methods* **3**: 486–504

Raudenbush S W, Bryk A S, Cheong Y F, Congdon R T Jr 2000 *HLM 5*. Computer software. Scientific Software International, Chicago

Rosenthal R 1991 *Meta-Analytic Procedures for Social Research*. Sage Publications, Newbury Park, CA

SAS 6.12 Computer software 1996 The SAS Institute, Cary, NC

Seltzer M H, Wong W H, Bryk A S 1996 Bayesian analysis in

applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics* **21**: 131–67
Strangl D K, Berry D A 2000 *Meta-Analysis in Medicine and Health Policy*. Marcel Dekker, New York

L. V. Hedges and J. L. Vevea

Metacognitive Development: Educational Implications

Metacognition is broadly defined as cognition about one's own cognitions. John Flavell's (1979) seminal article on developmental aspects of metacognition and metacognitive monitoring stimulated two decades of empirical and conceptual work in developmental as well as educational and mainstream cognitive psychology. The basic proposal outlined by Flavell was that developmental changes in metacognitive knowledge mediate developmental changes in cognitive performance. *Metamemory*, one subtype of metacognition that refers to knowledge about memory, has attracted particularly strong research interest in developmental psychology. Given that research devoted to aspects of metamemory and its importance for education has created by far the most findings in the field, the emphasis of the present overview will be on memory knowledge and its implications. However, most of what will be said about metamemory can be easily generalized to metacognitive knowledge related to a variety of problem-solving activities.

1. *Conceptualizations and Models of Metamemory*

The concept of metamemory was introduced by John Flavell and colleagues to refer to knowledge about memory processes and contents. In their taxonomy of metamemory, Flavell and Wellman (1977) distinguished between two main categories, 'sensitivity' and 'variables.' The 'sensitivity' category referred to mostly implicit, unconscious behavioral knowledge of when memory is necessary, and thus was very close to subsequent conceptualizations of *procedural metacognitive knowledge*. The 'variables' category referred to explicit, conscious, and factual knowledge about the importance of person, task, and strategy variables for memory performance. This is also known as *declarative metacognitive knowledge*. Metamemory about person variables includes knowledge about how, when, and why one remembers or forgets. Metamemory about task variables comprises knowledge about task influences on memory performance, for instance, knowledge that shorter item lists are easier to remember than long lists. Finally, metamemory about strategies refers to knowledge about advantages and possible problems of memory strategies.

This taxonomy of metamemory was not intended to be exhaustive. Since the late 1970s, a number of other theorists have contributed to the development of metamemory theory (for recent reviews, see Holland Joyner and Kurtz-Costes 1997, Schneider 1999, Schneider and Pressley 1997). For instance, Paris and colleagues (e.g., Paris and Oka 1986) introduced a component called 'conditional metacognitive knowledge' that focused on children's ability to justify or explain their decision concerning memory activities. Whereas the declarative metamemory component introduced by Flavell and Wellman focused on 'knowing that', conditional metamemory referred to 'knowing why.' The procedural metamemory component, that is, children's ability to monitor and regulate their memory behavior ('knowing how') was thoroughly analyzed by Ann Brown and colleagues (e.g., Brown et al. 1983). Here, the frame of reference was the competent information processor, one possessing an efficient 'executive' that regulated cognitive behaviors. Brown and colleagues could demonstrate that memory monitoring and regulation processes play a large role in complex cognitive tasks such as comprehending and memorizing text materials.

Although more recent conceptualizations of metacognition have expanded the scope of the theoretical construct, they also make use of the basic distinction between declarative and procedural knowledge. For instance, Wellman (1990) has linked the declarative metacognitive component to the broader concept of children's 'theory of mind', which focuses on classes of knowledge about the inner mental world and cognitive processes that develop during the preschool years. Pressley, Borkowski, and colleagues have systematically considered declarative and procedural components of metacognition in developing a theoretical model that emphasizes the dynamic interrelations among strategies, monitoring abilities, and motivation (e.g., Pressley et al. 1989). In their 'good information-processing model', metamemory is conceptualized in terms of a number of interactive, mutually dependent components such as domain knowledge, metacognitive knowledge, and monitoring efficiency.

It should be noted that conceptualizations of metacognition and metamemory in the fields of general cognitive psychology, social psychology, and the psychology of aging differ from this taxonomy. For instance, popular conceptualizations of metacognition in the field of cognitive psychology exclusively elaborate on the procedural component, focusing on the interplay between monitoring and self-control (see Nelson 1996). On the other hand, when issues of declarative metamemory are analyzed in the fields of social psychology and gerontology, the focus is on a person's belief about memory phenomena and not on veridical knowledge.

In a recent article, O'Sullivan and Howe (1998) propose a similar view for developmental research, arguing that metamemory should be conceptualized as personalized, constructed knowledge consisting of both accurate and naive beliefs. Although there are certain advantages of such a view, in particular with regard to research devoted to developmental changes in young children, O'Sullivan and Howe's position that the true–false belief distinction is not of pre-eminent importance in metamemory development remains controversial in the field. Anyway, the fact that different conceptualizations of metamemory exist in different areas of psychology illustrates the fuzziness of the concept and the need to carefully define the term in order to avoid misunderstandings. The concept of metamemory used in the rest of this article refers to both declarative and procedural knowledge components.

2. *Assessment of Metamemory*

There are a variety of measures that have been used to capture what children know about memory. Most measures of declarative, factual knowledge have utilized interviews or questionnaires that focus on knowledge about person variables, task demands, and strategies. Whereas earlier instruments suffered from methodological problems, more recent interviews and questionnaires showed better psychometric properties such as sufficient reliability and validity (Schneider and Pressley 1997). Moreover, it was shown that nonverbal techniques helped in assessing young children's declarative knowledge. For instance, one task required young children to distinguish effective from poor strategies while watching a model executing the strategies on video. In another successful procedure ('peer-tutoring task') older children were asked to tutor a younger child about how to do a certain memory task in order to maximize learning. Peer tutoring is likely more motivating to young children than interviews, and they tend to be more explicit when answering a question of an older child as compared with that of an adult (who already seems to know everything).

Procedural metamemory has been assessed through concurrent measurement of memory and metamemory. In this case, children are asked to perform a memory task and to simultaneously (or immediately afterwards) report their knowledge about how they performed the task and about factors that may have influenced their performance. Whereas some tasks focus on monitoring activities, others assess the impact of self-control and regulation. Examples of the former category include performance predictions which are made prior to study and involve estimation of how much will be remembered, as well as feeling-of-knowing (FOK) judgments. FOK tasks require children to estimate whether an item that currently cannot be recalled will be recognized if the experimenter provided it. These FOK ratings are then related to subsequent performance on a recognition test that includes nonrecalled items.

A number of concerns have been raised about measures of procedural metamemory. However, although there is not a perfect index of metamemory, many of the measurement problems that metamemory researchers confront are similar to measurement problems in other areas of psychology.

3. *Development of Metamemory*

A lot of metamemory data have been produced since the early 1980s, much of which was highly informative about children's knowledge about memory. We now understand better both (a) children's long-term, factual (declarative) knowledge about memory, and (b) their abilities to monitor memory.

3.1 *Children's Factual Knowledge About Memory*

Using sensitive methods that minimize demands on the child, it is possible to demonstrate some rudimentary knowledge of metamemorial facts in preschoolers. For instance, they understand mental words, distinguish between important vs. less important elements in picture stories, and know about the relevance of retrieval strategies when performing on a hide-and-seek task. Knowledge of facts about memory is more impressive in the elementary-grade years, and much more complete by the end of childhood. Nonetheless, knowledge of memory, in particular, knowledge about the interactions of metamemory variables or understanding of the relative importance of text elements continues to develop.

One of the most important findings produced by metamemory researchers is that there is increasing strategy knowledge with increasing age. However, there is also increasing evidence that many adolescents (including college students) have little knowledge of some powerful and important memory strategies (for details, see Pressley and McCormick 1995).

3.2 *Development of Procedural Metamemory*

Compared with the age trends demonstrated for declarative metamemory, the situation regarding procedural metamemory is less clear. Research focusing on monitoring (e.g., FOK tasks) has shown that even young children seem to possess the skills relevant for the solution of FOK and performance prediction problems when task difficulty is adequate. Thus the ability to monitor one's own memory activities can be already high in young children and seems to improve

continuously during the early elementary school years. However, the evidence regarding developmental trends is not consistent, with some studies showing better performance in younger than in older children (for a review, see Schneider 1999).

On the other hand, the available evidence on the development of self-regulation skills shows that there are clear increases from middle childhood to adolescence. Spontaneous and effective use of self-regulation skills occurs only in highly constrained situations during the grade-school years and continues to develop well into adolescence. Comparisons of younger and older children in 'study-time apportionment' tasks indicate that it is the interplay between monitoring and self-regulatory activities that develops with age. That is, when the task is to learn item pairs of varying difficulty, both younger and older children show adequate monitoring skills in that they are well able to distinguish difficult from easy item pairs. However, only the older children allocate study time differentially, spending more time on the difficult than on the easier items. In comparison, younger children typically spend about the same amount of time on easy pairs as they spend on hard pairs.

3.3 Metamemory–Memory Relations

From a developmental and educational perspective, the metamemory concept seems well-suited to explain children's production deficiencies on a broad variety of memory tasks. Empirical research was stimulated by the belief that young children do not spontaneously use memory strategies because they are not familiar with those tasks and thus do not know anything about the advantages of strategy use. This should change soon after children enter school and are confronted with a large number of memory tasks. Experience with such tasks should improve strategy knowledge (metamemory), which in turn should have an impact on memory behavior (i.e., strategy use). Thus, the major motivation behind studying metamemory and its development has been the assumption that although links between metamemory and memory may be weak in early childhood, they should become much stronger with increasing age.

Overall, the empirical findings do not indicate such a strong relationship. Narrative and statistical meta-analyses have shown that there are moderate, non-trivial quantitative associations between metamemory and memory. For instance, Schneider and Pressley (1997) reported an overall correlation of about 0.40 based on about 60 publications and more than 7000 children and adolescents. Qualitative reviews of the various metamemory–memory relations make obvious, however, that no single statistic could capture the diversity and richness of the findings. In general, the evidence points to a bidirectional, reciprocal relationship between metamemory and memory behavior. Metamemory can influence memory behavior, which in turn leads to enhanced metamemory.

Although there are several reasons for the fact that the link between metacognition and cognitive performance is not always as strong as it could be, undoubtedly one of the major mediators is motivation. Empirical research has shown that metamemory–memory performance relationships are particularly impressive when participants are highly motivated to achieve a certain goal. Under these circumstances, all available ressources will be activated to cope with task demands, including declarative and procedural metacognitive knowledge (cf. Borkowski and Mutukrishna 1995).

4. Metacognition and Education

Most of memory development is not so much a product of age but of education. One way in which parents and teachers facilitate cognititive development is by nurturing the development of children's metacognition. For instance, a cross-cultural study by Carr and colleagues (Carr et al. 1989) examined differences in the amount of instruction that US and German parents provided to their children at home. Carr and colleagues found that German parents, in contrast to US parents, reported more instruction of strategies, checked homework more often, and also used games that promoted thinking. These differences in home instruction were accompanied by superior strategy use by German children on the memory tasks provided in school. Carr et al. (1989) concluded from this that parental instruction is an important facilitator of children's further metacognitive development.

One rather new area of metamemory research involves applications of metacognitive theory to educational settings. It was assumed that children's experiences at school must play an important role in shaping their use of and knowledge about how to learn and remember. However, observations of normal classroom situations did not always prove that teachers foster children's metacognitive development. For instance, Moely and colleagues observed in classrooms to find out how elementary teachers instructed strategy use and memory knowledge as they presented lessons to children in Grades K to 6 (see the overview by Moely et al. 1995). Teachers varied widely in the extent to which they focused on how children might adjust or regulate their cognitive activities in order to master a task. Efforts also varied widely depending on the subject matter under consideration. When Moely and colleagues first looked at a broad range of instructional effort, they found low levels of strategy instruction. However, teachers were found to be more likely to offer suggestions for strategy use when teaching math problem solving.

Although rich strategy instruction is not common in schools, it can be successfully implemented. Several

comprehensive research projects focused on reading instruction and comprehension monitoring (cf. Borkowski and Mutukrishna 1995, Paris and Oka 1986). A particularly important instructional procedure in this context is *reciprocal teaching* (Palincsar and Brown 1984). Reciprocal teaching takes place in a collaborative learning context and involves guided practice in the flexible use of the following four comprehension monitoring strategies: questioning, summarizing, clarifying, and predicting. The novice's role is facilitated by the provision of scaffolding by the expert (teacher). Skills and strategies are practiced in the context of reciprocal teaching dialogs. The teacher and the students take turns leading discussions regarding the contents of a text they are jointly attempting to understand. Overall, this instructional approach has proven to be extraordinarily successful both with normal and learning-disabled students.

Further, very ambitious programs have been undertaken by Pressley and colleagues in order to evaluate effective instructional programs in US public school systems (see Pressley and McCormick 1995, Schneider and Pressley 1997). Strategy instruction was not conducted in isolation, but was viewed as an integral part of the curriculum, and thus was taught as part of language arts, mathematics, science, and social studies. The goal was to simultaneously enhance children's repertoires of strategies, knowledge, metacognition, and motivation. Pressley and colleagues found that effective teachers regularly incorporated strategy instruction and metacognitive information about flexible strategy use and modification as a part of daily instruction. This research and the applied studies outlined above have enhanced greatly our understanding of how to establish long-term strategy instruction in educational contexts that not only is rich in metamemory and motivational enhancement, but also helps most students to accomplish their academic goals.

See also: Cognitive Development: Learning and Instruction; Cognitive Styles and Learning Styles; Competencies and Key Competencies: Educational Perspective; Explanation-based Learning, Cognitive Psychology of; Instructional Technology: Cognitive Science Perspectives; Learning to Learn; Memory Development in Children; Metamemory, Psychology of; Self-regulated Learning; Tacit Knowledge, Psychology of

Bibliography

Borkowski J H, Muthukrishna N 1995 Learning environments and skill generalization: How contexts facilitate regulatory processes and efficacy beliefs. In: Weinert F E, Schneider W (eds.) *Memory Performance and Competencies—Issues in Growth and Development*. Erlbaum, Mahwah, NJ, pp. 283–300

Brown A L, Bransford J D, Ferrara R A, Campione J C 1983 Learning, remembering, and understanding. In: Flavell J H, Markman E M (eds.) *Handbook of Child Psychology*, Vol 3: *Cognitive Development*. Wiley, New York, pp. 77–166

Carr M, Kurtz B E, Schneider W, Turner L A, Borkowski J G 1989 Strategy instruction and transfer among American and German children: Environmental influences on metacognitive development. *Developmental Psychology* **25**: 765–71

Flavell J H 1979 Metacognition and cognitive monitoring—A new area of cognitive-developmental inquiry. *American Psychologist* **34**: 906–11

Flavell J H, Wellman H M 1977 Metamemory. In: Kail R V, Hagen W (eds.) *Perspectives on the Development of Memory and Cognition*. Erlbaum, Hillsdale, NJ, pp. 3–33

Holland Joyner M, Kurtz-Costes B 1997 Metamemory development. In: Cowan N (ed.) *The Development of Memory in Childhood*. Psychology Press, Hove, UK, pp. 275–300

Moely B, Santulli K, Obach M 1995 Strategy instruction, metacognition, and motivation in the elementary school classroom. In: Weinert F E, Schneider W (eds.) *Memory Performance and Competencies—Issues in Growth and Development*. Erlbaum, Mahwah, NJ, pp. 301–21

Nelson T O 1996 Consciousness and metacognition. *American Psychologist* **51**: 102–16

O'Sullivan J T, Howe M L 1998 A different view of metamemory with illustrations from children's beliefs about long-term retention. *European Journal of Psychology of Education* **13**: 9–28

Palincsar A S, Brown A L 1984 Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction* **1**: 117–75

Paris S G, Oka E R 1986 Children's reading strategies, metacognition, and motivation. *Developmental Review* **6**: 25–56

Pressley M, Borkowski J G, Schneider W 1989 Good information processing: What it is and what education can do to promote it. *International Journal of Educational Research* **13**: 857–67

Pressley M, McCormick C B 1995 *Advanced Educational Psychology for Educators, Researchers, and Policymakers*. Harper Collins, New York

Schneider W 1999 The development of metamemory in children. In: Gopher D, Koriat A (eds.) *Attention and Performance XVII—Cognitive Regulation of Performance: Interaction of Theory and Application*. MIT Press, Cambridge, MA

Schneider W, Pressley M 1997 *Memory Development Between Two and Twenty*, 2nd edn. Erlbaum, Mahwah, NJ

Wellman H M 1990 *The Child's Theory of Mind*. MIT Press, Cambridge, MA

W. Schneider

Metamemory, Psychology of

This article is an overview of psychological research on metamemory, which is a subset of metacognition. Metacognition is the scientific investigation of an individual's cognitions about his or her own cognitions. In particular, metamemory is the subset of metacognition that emphasizes the monitoring and

control of one's own memory processing, both during the acquisition of new information into memory and during the retrieval of previously acquired memories.

What makes the investigation of metacognition scientific is that the theories of metacognition attempt to account for empirical data about metacognition. Related to that, one of the oldest topics in psychology is the topic of consciousness, and early twentieth-century textbooks about psychology frequently defined psychology as the scientific investigation of consciousness. Theories of an individual's cognitions about his or her own cognitions may appear to be similar to what some people would regard as consciousness, especially self-consciousness. This is not surprising, because the development of theories of consciousness can be affected by the empirical findings about metacognition, in at least two ways (Nelson 1996): first, the empirical findings pose a challenge to theories of consciousness insofar as such theories should be able to account for the empirical findings about how people monitor their own cognitions, and hence such theories can sometimes be disconfirmed by particular empirical findings; second, the empirical findings may provide clues that can inspire new theories of consciousness (e.g., see Flanagan 1992). Thus the interplay between metacognition and consciousness can be expected to be symbiotic, with each of those affecting the other (e.g., special issue of *Consciousness and Cognition*, 2000).

Two major subdivisions of metacognition are (a) metacognitive knowledge (i.e., what people know about their own cognitions as based on their life history) and (b) on-line metacognitions comprised of metacognitive monitoring and metacognitive control of one's own cognitions.

The first of those subdivisions includes autobiographical facts such as 'I remember things better when I see them than when I hear them' or 'I usually remember the gist of the text better than the exact words that were in the text.' Metacognitive knowledge, especially the development of metacognitive knowledge, has been studied extensively in children (e.g., Kreutzer et al. 1975).

The second of those subdivisions involves questions about the way in which people monitor their on-going cognitions and also the way in which people control their on-going cognitions. The key notion is that the distinction between the meta level versus the object level is relational rather than absolute. Put another way, no particular aspect of cognition is always at the meta-level in any absolute sense. Instead, if one aspect of cognition is monitoring or controlling another aspect of cognition, then we regard the former aspect as metacognitive in relation to the latter aspect. An example may help to clarify this and may also help illustrate the kinds of metacognition that currently are being researched. Imagine that you are asked what the capital of Australia is. You might recall the name of an Australian city, and after you say it, you might be asked how confident you are that the answer you recalled is correct. Thus your confidence judgment (e.g., '50 percent') is at the meta-level, relative to your recall response (e.g., 'Sydney'). However, if you are then asked to tell how accurate that particular confidence judgment is, you might put an interval around the confidence judgment (e.g., 'somewhere between 40 and 60 percent' or 'somewhere between 30 and 70 percent'); then the original confidence judgment of '50 percent' is at the object level, relative to the confidence interval. Thus a given cognition can be either an object-level cognition (if it is the object of some other cognition that is monitoring or controlling it) or a meta-level cognition (if it is monitoring or controlling some other cognition).

In the heavily researched area of metacognition referred to as 'metamemory,' the monitoring and control are of one's own memory during the acquisition of new information and during the retrieval of previously acquired information. Before about 1950, many researchers conceptualized people as blank slates, and the way that acquisition was believed to occur was that the individual was assumed to be passive, having little or no control over his or her own acquisition. However, since the 1950s, researchers began to conceptualize the individual as having substantial control over acquisition and as being active rather than passive, both during the acquisition of new information and during the retrieval of previously learned information.

Consider this concrete example. Suppose that a student is studying for an examination that will soon occur on French–English vocabulary such as *chateau* = castle. We suppose that several monitoring and control processes will be activated while the student is learning the new vocabulary and while the student is attempting to retrieve the answers during the subsequent examination. Some of those monitoring and control processes are discussed in the next section (a theoretical framework that integrates these processes into an overall system can be found in an article by Nelson and Narens 1990).

1. *Metacognitive Monitoring*

The various metacognitive monitoring processes are differentiated in terms of when they occur during acquisition and retrieval, and also in terms of whether they pertain to the person's future performance (prospective monitoring) or the person's past performance (retrospective monitoring). Consider each of those in turn.

1.1 *Prospective Monitoring*

1.1.1 Ease-of-learning judgments. Even prior to initiating the intentional acquisition of to-be-learned

items, some metacognitive monitoring occurs. An ease-of-learning judgment is the person's judgment of how easy or difficult the items will be to acquire. For instance, the person might believe that *cheval* = horse will be more difficult to learn than *chateau* = castle. Underwood (1966) showed that learners are somewhat accurate—not perfectly, but well above chance—at predicting which items will be easiest to learn. The learners' predictions of how easy it would be to learn each item, made in advance of the presentation of those items for study, were positively correlated with subsequent recall after a constant amount of study time on every item. That is, the items people predicted would be easiest to learn had a greater subsequent likelihood of being recalled than items predicted to be hardest to learn.

1.1.2 Judgments of learning. The next kind of monitoring occurs during or soon after acquisition. The learner's judgment of learning is his or her prediction of the likelihood that a given item will be remembered correctly on a future test. Arbuckle and Cuddy (1969) showed that the predictive accuracy of people's judgments of learning is above chance but far from perfect, similar to the situation for ease-of-learning judgments. Research by Leonesio and Nelson (1990) showed that judgments of learning are more accurate than ease-of-learning judgments for predicting eventual recall, perhaps because people's judgments of learning can be based on what learners notice about how well they are mastering the items during acquisition. Mazzoni and Nelson (1995) showed that judgments of learning are more accurate when the learning is intentional rather than incidental, even when the amount recalled is the same for intentional versus incidental learning.

In regard to intentional leaning, two examples illustrate how widely the accuracy of judgments of learning can vary. First, in situations such as the acquisition of foreign-language vocabulary as discussed above, Nelson and Dunlosky (1991) found that people's judgments of learning can be almost perfectly accurate if the judgment of learning is made not immediately after studying a given item but rather after a short delay; this finding has been replicated in many laboratories and is called the 'delayed-JOL effect' (where 'JOL' stands for judgment of learning). The delayed-JOL effect is exciting because it shows that under the proper conditions, people can monitor their learning extremely accurately. However, there currently is controversy over the theoretical mechanisms that give rise to the high accuracy of delayed JOLs (e.g., Dunlosky and Nelson 1992, Kelemen and Weaver 1997, Spellman and Bjork 1992), and this seems a fruitful topic for future research.

Second, however, in situations such as the acquisition of text, JOLs can be extremely inaccurate. Glenberg and his colleagues had people read passages of text and make a JOL after each passage, followed by a test consisting of one true–false inference derived from each passage. The recurring finding was that the JOLs were not above chance accuracy for predicting that test performance; Glenberg et al. (1982) referred to this as an 'illusion of knowing.' However, subsequent researchers discovered that those findings had a highly limited domain and that people's JOL accuracy for assessing their comprehension of text could be well above chance if small changes were made in the tests that assessed the accuracy of the JOLs. In particular, Weaver (1990) found that the accuracy of JOLs increased as the number of true–false inference questions increased, and Maki et al. (1990) found that the accuracy of JOLs increased when the test question was a multiple-choice item rather than a true–false item. Thus the way in which the accuracy of JOLs is assessed can affect conclusions about the degree of metacognitive accuracy.

1.1.3 Feeling-of-knowing judgments. Another kind of metacognitive monitoring judgment is people's prediction of whether they will eventually remember an answer that they currently do not recall. This was the first metamemory judgment examined in the laboratory to assess people's accuracy at predicting their subsequent memory performance. Hart (1965) found that feeling-of-knowing judgments were somewhat accurate at predicting subsequent memory performance. The likelihood of correctly recognizing a nonrecalled answer was higher for nonrecalled items that people said they knew than for nonrecalled items people said they didn't know. However, people frequently did not recognize answers that they had claimed that they would recognize, and people sometimes did recognize answers that they had claimed they wouldn't recognize (although in part, these correct recognitions could sometimes be due to guessing factors in the multiple-choice recognition test). Subsequently, the accuracy of predicting other kinds of memory performance such as relearning was investigated by Nelson et al. (1984), who also offered several theoretical explanations for how people might make their feeling-of-knowing judgments. Currently, the most widely accepted explanation (e.g., Koriat 1997, Metcalfe et al. 1993, Reder and Ritter 1992) is that rather than monitoring directly the nonrecalled information in memory (almost as if by magic—see Nelson and Narens 1990), what people do when they make feeling-of-knowing judgments is to assess both their familiarity with the stimulus cue (aka 'stimulus recognition') and the partial components that they can recall from the requested response (e.g., tip-of-the-tongue components such as the first letter or the number of syllables), and then draw an inference based on that assessment. For instance, when people recognize the stimulus *cheval* as having been studied previously and/or recall that 'h' is the first letter of

the requested response, then they might infer that they will recognize the requested response if they saw it in a multiple-choice test item.

1.2 Retrospective Confidence Judgments

In contrast to the aforementioned monitoring judgments in which people attempt to predict their future memory performance, retrospective confidence judgments occur after someone recalls or recognizes an answer. They are judgments of how confident the person is that his or her answer was correct. For instance, if someone were asked for the English translation equivalent of *chateau*, the person might recall 'castle' (the correct answer) or might recall 'hat' (the incorrect answer, probably occurring because the person confused *chateau* with *chapeau*) and then, without feedback from the experimenter, would make a confidence judgment about the likelihood that the recalled answer was correct. Fischhoff et al. (1977) demonstrated that these retrospective confidence judgments have substantial accuracy, but there is a strong tendency for people to be overconfident, especially when the test is one of recognition. For instance, for the items that people had given a confidence judgment of '90 percent likely to be correct,' the actual percentage of correct recognition was substantially below that. Subsequent research by Koriat et al. (1980) found that people's accuracy could be increased—and people's tendency to be overconfident decreased—if at the time of making each retrospective confidence judgment, the people were asked to give a reason that their response (in either recall or recognition) might have been wrong. However, no change in accuracy occurred when people were asked to give a reason that the response might have been correct, so the conclusion of the researchers was that in the usual situation of making retrospective confidence judgments, people have a 'confirmation bias' to think of reasons why they were correct and fail to think of reasons why they might have been wrong.

2. Metacognitive Control

Although it is interesting that people can monitor their progress during acquisition and retrieval, this would be little more than a curiosity if it had no other role in learning and memory. However, people can control aspects of their acquisition and retrieval. First, consider what people can control during self-paced acquisition; second, consider what they can control during retrieval.

2.1 Control During Self-paced Acquisition

2.1.1 Allocation of self-paced study time during acquisition. A student who is learning foreign-language vocabulary can allocate various amounts of study time to the various to-be-learned items, for example, allocating extra study time to the most difficult items. Bisanz et al. (1978) found that the allocation of study time may be related to people's JOLs. Learners in the early years of primary school make accurate JOLs but do not utilize the JOLs when allocating study time across the items, whereas slightly older children do utilize their JOLs when allocating study time. The older children allocated extra study time to items that they judged not yet to have been learned and did not allocate extra study time to items that they judged to have been learned.

2.1.2 Strategies during self-paced study. People can control not only *how much* study time they allocate to various items, but also *which strategy* they use during that study time. Often there are strategies that are more effective than rote repetition, but do people know about them? People's utilization of a mnemonic strategy for the acquisition of foreign-language vocabulary was investigated by Pressley et al. (1984). After people learned some foreign-language vocabulary by rote and learned other foreign-language vocabulary by the mnemonic strategy, they chose whichever strategy they preferred for a final trial of learning new foreign-language vocabulary. Only 12 percent of the adults chose the mnemonic strategy if they had not received any test trials during the earlier phase. However, 87 percent chose the mnemonic strategy if they had received test trials during the earlier acquisition phase. Thus, test trials help people to see the effectiveness of different strategies. When the subjects were children instead of adults, they not only needed test trials but also needed experimenter-provided feedback after those test trials so as to know how well they had performed on the rote-learned items versus the mnemonic-learned items. Without both the test trials and the feedback, the children were unlikely to adopt the advantageous mnemonic strategy.

2.2 Control During Retrieval

2.2.1 Control of initiating attempts at retrieval. Immediately after someone is asked a question, and before attempting to search memory for the answer, a metacognitive decision occurs about whether the answer is likely to be found in memory. If you are asked what the telephone number is for the President of the United States, you probably would decide immediately that the answer is not in your memory. Notice that you do not need to search through all the telephone numbers that you know, nor do you need to search through all the information you have stored in your memory about the President. Consider how different that situation is from one in which you are asked the telephone number of one of your friends.

This rapid feeling-of-knowing judgment that precedes an attempt to retrieve an answer was investigated by Reder (1987). She found that people are faster at making a feeling-of-knowing decision about whether or not they know the answer to a general information question (e.g., 'What is the capital of Australia?') than they are at answering that question (e.g., saying 'Canberra'). Thus a metacognitive decision can be made prior to (as well as after) retrieving the answer. Only if people feel that they know the answer will they continue their attempts to retrieve the answer. When they feel they do not know the answer, they don't even attempt to search memory (as in the aforementioned example of your response to a query for the President's telephone number).

2.2.2 Control of the termination of retrieval. People may initially believe that they know an answer, but after extended attempts at retrieval without producing the answer, they eventually terminate searching for the answer. The metacognitive decision to terminate such an extended search of memory was investigated by Nelson et al. (1984). They found that the amount of time elapsing before someone gives up searching memory for a nonretrieved answer is greater when the person's on-going feeling of knowing for the answer is high rather than low. As an example, someone might spend a long time during an examination attempting to retrieve the English equivalent of *chateau* (which the person studied the night before) but little or no time attempting to retrieve the English equivalent of *boîte* (which the person did not study previously). The metacognitive decision to continue versus terminate attempts at retrieving an answer from memory can also be affected by other factors, such as the total amount of time available during the examination. A recent theory of the metacognitive components involved in retrieval has been offered by Barnes et al. (1999).

3. Neuropsychological Aspects of Metacognition

Neuropsychological patients have been investigated to determine if any of them have particular deficits of metacognition. For instance, Korsakoff patients, who have frontal-lobe damage as well as other brain damage (Shimamura et al. 1988), have deficits in the accuracy of their JOLs (Bauer et al. 1984). Also, Korsakoff patients have extremely low feeling-of-knowing accuracy (Shimamura and Squire 1986) but normal retrospective-confidence-judgment accuracy (Shimamura and Squire 1988). Patients with primarily frontal-lobe deficits sometimes show normal recall but reduced feeling-of-knowing accuracy (Janowsky et al. 1989).

Many of the experiments cited in this article have been reprinted in a book edited by Nelson (1992). Other books containing more recent findings about metacognition have been edited by Metcalfe and Shimamura (1994), Reder (1996), and Mazzoni and Nelson (1998).

See also: Cognitive Neuropsychology, Methodology of; Cognitive Neuroscience; Cognitive Psychology: History; Cognitive Psychology: Overview; Incidental versus Intentional Memory; Memory Retrieval; Prefrontal Cortex; Prospective Memory, Psychology of; Reconstructive Memory, Psychology of; Self-monitoring, Psychology of

Bibliography

Arbuckle T Y, Cuddy L L 1969 Discrimination of item strength at time of presentation. *Journal of Experimental Psychology* **81**: 126–31

Barnes A E, Nelson T O, Dunlosky J, Mazzoni G, Narens L 1999 An integrative system of metamemory components involved in retrieval. *Attention and Performance* **17**: 287–313

Bauer R H, Kyaw D, Kilbey M M 1984 Metamemory of alcoholic Korsakoff patients. *Society for Neurosciences Abstracts* **10**: 318

Bisanz G L, Vesonder G L, Voss J F 1978 Knowledge of one's own responding and the relation of such knowledge to learning: A developmental study. *Journal of Experimental Child Psychology* **25**: 116–28

Dunlosky J, Nelson T O 1992 Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition* **20**: 374–80

Fischhoff B, Slovic P, Lichtenstein S 1977 Knowing with certainty: the appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance* **3**: 552–64

Flanagan O 1992 *Consciousness Reconsidered*. MIT Press, Cambridge, MA

Glenberg A, Wilkinson A C, Epstein W 1982 The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition* **10**: 597–602

Hart J T 1965 Memory and the feeling-of-knowing experience. *Journal of Educational Psychology* **56**: 208–16

Janowsky J S, Shimamura A P, Squire L R 1989 Memory and metamemory: Comparisons between patients with frontal lobe lesions and amnesic patients. *Psychobiology* **17**: 3–11

Johnson M K, Raye C L 1981 Reality monitoring. *Psychological Review* **88**: 67–85

Kelemen W L, Weaver C A 1997 Enhanced metamemory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, & Cognition* **23**: 1394–409

Koriat A 1997 Monitoring one's own knowledge during study: A cue-utilization approach of judgments of learning. *Journal of Experimental Psychology: General* **126**: 349–70

Koriat A, Lichtenstein S, Fischhoff B 1980 Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory* **6**: 107–18

Kreutzer M A, Leonard C, Flavell J H 1975 An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development* **40**: 1–60

Leonesio R J, Nelson T O 1990 Do different metamemory judgments tap the same underlying aspects of memory?

Journal of Experimental Psychology: Learning, Memory, & Cognition **16**: 464–70
Maki R H, Foley J M, Kajer W K, Thompson R C, Willert M G 1990 *Journal of Experimental Psychology: Learning, Memory, & Cognition* **16**: 609–16
Mazzoni G, Nelson T O 1995 Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 1263–74
Mazzoni G, Nelson T O 1998 *Metacognition and Cognitive Neuropsychology: Monitoring and Control Processes*. Erlbaum, Mahwah, NJ
Metcalfe J, Schwartz B L, Joaquim S G 1993 The cue familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**: 851–61
Metcalfe J, Shimamura A P 1994 *Metacognition: Knowing about Knowing*. MIT Press, Cambridge, MA
Nelson T O 1992 *Metacognition: Core Readings*. Allyn and Bacon, Boston
Nelson T O 1996 Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item comment. *Applied Cognitive Psychology* **10**: 257–60
Nelson T O, Dunlosky J 1991 When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The 'delayed-JOL effect.' *Psychological Science* **2**: 267–70
Nelson T O, Gerler D, Narens L 1984 Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General* **113**: 282–300
Nelson T O, Narens L 1990 Metamemory: A theoretical framework and new findings. In: Bower G H (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory*. Academic Press, San Diego, CA, Vol. 26, pp. 125–73
Pressley M, Levin J R, Ghatala E 1984 Memory strategy monitoring in adults and children. *Journal of Verbal Learning and Verbal Behavior* **23**: 270–88
Reder L M 1987 Strategy selection in question answering. *Cognitive Psychology* **19**: 90–138
Reder L M 1996 *Implicit Memory and Metacognition*. Erlbaum, Mahwah, NJ
Reder L M, Ritter F E 1992 What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**: 435–51
Shimamura A P, Jernigan T L, Squire L R 1988 Korsakoff's syndrome: Radiological (CT) findings and neuropsychological correlates. *Journal of Neuroscience* **8**: 4400–10
Shimamura A P, Squire L R 1986 Memory and metamemory: A study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **12**: 452–60
Shimamura A P, Squire L R 1988 Long-term memory in amnesia: Cued recall, recognition memory, and confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14**: 763–70
Spellman B A, Bjork R A 1992 When predictions create reality-judgments of learning may alter what they are intended to assess. *Psychological Science* **3**: 315–16
Underwood B J 1966 Individual and group predictions of item difficulty for free learning. *Journal of Experimental Psychology* **71**: 673–79
Weaver C A 1990 Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**: 214–22

T. O. Nelson

Metaphor and its Role in Social Thought: History of the Concept

Metaphor has resisted any wide agreement as concept, yet the last few decades have witnessed a burgeoning of work and interest in metaphor and its related tropes. The many attempts to theorize metaphor have included, *inter alia*, characterizations of metaphor as 'comparison,' as 'without meaning,' as 'anomaly,' 'speech act,' 'loose talk,' 'interaction,' and 'intentional category mistakes.' These attempts have had the form of trying to assimilate metaphor under a previously existing understanding of language—a move that is itself metaphorical. This resistance to conceptualizing suggests that it might be more helpful to consider metaphor in terms of its uses.

1. *Aristotle*

For most of its life metaphor has had its home in rhetoric, and it has been to Aristotle that writers have usually turned for the earliest considerations. He understood it as 'giving a thing a name that belongs to something else,' and thought that 'the greatest thing by far is to have a command of metaphor.' Not all writers have heeded the contexts within which Aristotle wrote. His teacher, Plato, had sufficient doubts about the value of metaphor that he barred poets from his Republic. Aristotle, by contrast, found uses for metaphor, not only in politics where rhetoric enabled a man to be heard effectively in public, but also in law, where juries were suspicious of evidence that could be faked, and witnesses who could be bribed. They were, rather, influenced by arguments turning around a balance of probabilities, providing an important context for rhetoric. Aristotle cannot easily be read as limiting metaphor to the realm of ornament, or efficiency of utterance: he sees clear cognitive uses:

> … strange words simply puzzle us; ordinary words convey only we what we know already; it is from metaphor that we can best get hold of new ideas (1984 *Rhetoric*, lll,1410b).

We may suppose that Aristotle's description of metaphor as giving something a name which belongs to something else, detailed as: 'the transference being either from genus to species, or from species to genus, or from species to species, or on grounds of analogy'

(1984 *Poetics*, 1457b), also owed something to his practice as taxonomist of the natural world. Although Aristotle's comment on Plato's forms that '(to say) they are patterns and the other things share in them is to use empty words and poetical metaphors' (1984 *Metaphysics* 991a21), may be taken as evidence of his ambivalence to metaphor, but should rather be understood as evidence of a variety of rhetorical uses. His reference to analogy and his remark that 'to make good metaphors implies an eye for resemblances' has led his theory to be dubbed 'the comparative theory.' Aristotle may be understood, in short, as praising metaphor in poetry and drama, in law, in politics, and in what we might now refer to as natural science (see *Aristotelian Social Thought*).

The subsequent history of metaphor until the end of the Renaissance may best be understood as part of the history of rhetoric.

2. *Early Modernity and Early Resistance to Metaphor*

Part of the development of rhetoric in seventeenth-century Europe was a 'simplification' of the tropes to just four: metaphor; metonymy in which the name of an attribute or other adjunct is substituted for the intended object as with 'scepter' for 'authority'; synecdoche in which a part signifies the whole; and irony where a term is used which is opposite to the one intended to be read.

The first major philosophy written in English, credited to Thomas Hobbes (1588–1679) and his *Leviathan* (1651), published a year after Descartes' death, is severe on metaphor and claims that, of the seven causes of absurdity 'The sixth, (is) to the use of metaphors, tropes, and other rhetorical figures, instead of words proper.' Yet the opening paragraph of the English version of the text elaborates what Black (1962) would later call the 'associated commonplaces' of the title image, Leviathan, Hobbes's metaphor for the State. Hobbes was writing in complicated, changing, and dangerous times, with differing purposes, addressing different audiences and with changing beliefs, about which Quentin Skinner has written elucidating the use of rhetoric in a context of humanism, of the growth of science, and the growing tensions with the church (1996). John Locke (1632–1704), following Hobbes, thought that 'all the artificial and figurative application of words … (are) in all discourses that pretend to inform or instruct, wholly to be avoided; and where truth and knowledge are concerned, cannot but be thought a great fault.' Here began a tradition later called the 'purity thesis,' still sustained in the British Isles at the beginning of the twenty-first century.

Meanwhile, in continental Europe, other voices were heard in praise of metaphor. Giambatista Vico (1668–1744) used metaphor as a tool within philology to produce a theory of 'poetic logic.' This included the notion that names were given to features of the natural world that were already well known, drawing for example upon the body—rivers and bottles were given 'mouths,' etc.

> Whence we derive the following (principle of) criticism for the times in which the metaphors were born in (various) languages. All metaphors which, by means of likenesses taken from bodies, come to signify the labors of abstract minds, must belong to times in which philosophies had begun to become more refined. This is shown by the following: that in all languages the words necessary to cultivated arts and recondite sciences have rural origins (Vico 1968, para. 404).

Vico's view led him to reverse the standard relationship of the poetic to the literal.

> From all this it follows that all the tropes (and they are all reducible to the four types discussed) which have hitherto been considered ingenious inventions of writers, were necessary modes of expression of all the first poetic nations, and had originally their full native propriety. But these expressions of the first nations later became figurative when, with the further development of the human mind, words were invented which signified abstract forms or genera comprising their species or relating parts with their wholes. And here begins the overthrow of two common errors of the grammarians: that prose speech is proper speech, and poetic speech improper; and that prose speech came first and afterward speech in verse (Vico 1968, para. 409).

Vico's theory links the actions of metaphor, metonymy, and synecdoche. Metaphor identifies a new domain of human experience making it possible for metonymy to identify its parts, synecdoche to elaborate the relations between the attributes of parts to the whole, and irony to inspect the earlier practices for their adequacy, by a test of opposition. When irony is correctly detected the new object may be regarded as appropriately established in the consciousness of the person or group. Vico's views were not fully effective until the nineteenth century but views in some respects similar to his were held by Herder (1744–1803), born in the year Vico died. Herder's application was not the diachronic one of Vico, but rather the synchronic, through which he came to celebrate difference in cultures and provided an intellectual basis for the beginnings of anthropology and ethnology (see *Romanticism: Impact on Social Thought*).

Some dissident voices were also heard in England. Coleridge (1772–1834), when asked why he attended Davy's chemistry lectures, replied that he wanted to renew his stock of metaphors. Shelley (1792–1822) provided a connection between metaphor, language, and culture:

> Language is vitally metaphorical; that is, it marks the before unapprehended relations of things and perpetuates their apprehension, until words, which represent them, become, through time, signs for portions or classes of thought instead of pictures of integral thoughts: and then, if no new poets

should arise to create afresh the associations which have thus been disorganized, language will be dead to all the nobler purposes of human intercourse.

3. Nietzsche and Freud

Nietzsche (1844–1900) marks a significant development. Moving beyond both Herder and Vico and their celebration of cultural plurality, extending Shelly's view, and in a Darwinian context, he proposed that metaphor serves a creative and critical plurality. With earlier writers we may be tempted to observe that with a change of view of metaphor there seems to be a linked change of the theory of the human being. With Nietzsche, however, the linkage is explicit:

> The drive toward the formation of metaphors is the fundamental human drive, which one cannot for an instant dispense with in thought, for one would thereby dispense with man himself. This drive is not truly vanquished and scarcely subdued by the fact that a regular and rigid new world is constructed as its prison from its own ephemeral products, the concepts (Nietzsche 1979, p. 88).

Turning towards the domain of philosophy, he remarks in an early writing:

> What then is truth? A movable host of metaphors, metonymies, and anthropomorphisms: in short, a sum of human relations which have been poetically and rhetorically intensified, transferred, and embellished, and which, after long usage, seem to a people to be fixed, canonical, and binding. Truths are illusions that we have forgotten are illusions; they are metaphors that have become worn out and have been drained of sensuous force, coins that have lost embossing and are now considered as metal and no longer as coins (Nietzsche 1979, p. 84).

The metaphor for truth, of coin which loses its face, has attracted many writers, few of who have commented on the phrase 'a sum of human relations' as applied to the tropes, a perception yet remaining to be developed.

Freud (1856–1939), some 12 years younger than Nietzsche, with mutual intimate acquaintances, made the claim that Nietzsche's 'guesses and intuitions often agree in the most astonishing way with the laborious findings of psychoanalysis.' Whatever that relationship might yet turn out to have been, Freud's work had the effect of applying Nietzsche's perceptions to the 'unconscious,' a new metaphor of depth, drawing upon the classical myths for his lexicon—which, in turn, made available rich seams of linked metaphors. I. A. Richards (1893–1979) who provided terms for the study of metaphor: 'tenor,' 'vehicle,' and 'tension.' These terms, though by no means unambiguous, and failing to encompass the full range of metaphors, nevertheless helped to revive talk of metaphor. Richards linked psychoanalysis to metaphor.

> The psychoanalysts have shown us with their discussions of 'transference'—another name for metaphor—how constantly modes of regarding, of loving, of acting, that have developed with one set of things or people, are shifted to another. They have shown us chiefly the pathology of these transferences, cases where the vehicle—the borrowed attitude, the parental fixation, say—tyrannizes over the new situation, the tenor, and behavior is inappropriate. The victim is unable to see the new person except in terms of the old passion and its accidents. The victim reads the situation only in terms of the figure, the archetypal image, the vehicle. But in healthy growth, tenor and vehicle—the new human relationship and the family constellation—cooperate freely; and the resultant behavior derives in due measure from both. Thus in happy living the same patterns are exemplified and the same risks of error are avoided as in tactful and discerning reading. The general form of the interpretative process is the same, with a small-scale instance—the right understanding of a figure of speech—or with a large-scale instance—the conduct of a friendship (Richards 1936, 135ff).

More recently, Freud's writing has begun to be redescribed in ways that reveal its tropical basis (e.g., Lacan *The Four Fundamental Concepts of Psychoanalysis*). (See also *Psychoanalysis, History of*.)

4. The Tide Turns, Natural Science Exposed as Metaphor-dependent

Quine (1908–2000) writing as an empiricist nevertheless thought it: 'a mistake, then, to think of language usage as literalistic in its main body and metaphorical in its trimming. Metaphor, or something like it, governs both the growth of language and our acquisition of it' (1979, p. 160). Mary Hesse has extensively described the work of metaphor in science as compatible with this view. Max Black (1962) offered an 'interactive' theory of metaphor as a development of the 'comparison' view, and further linked metaphor and model, thereby linking the worlds (languages?) of literature and science. In the same year, Thomas Kuhn (1962), offered a sociological description of natural science, one recognized more by sociologists and historians than by philosophers, which had as one of its central notions that of a 'paradigm' which, in some uses, may be taken as a synonym for metaphor. This provided a stimulus for developing a social critique of natural science (see Latour and Woolgar, *Laboratory Life: The Construction of Scientific Facts*). In 1970, Rom Harré distinguished models in science into two kinds, homeomorphs which are modeled upon its subject, as is a doll, and paramorphs where the model is only distantly related, Harré's examples included Harvey's use of hydraulics to model, and discover, the circulation of blood. Paramorphs, close cousins to metaphor, are creative: homeomorphs have heuristic uses only. It is perhaps ironic that the stimulus for much work on metaphor may have come from this identification of the place of metaphor in science.

5. *The Tide Turns, Within the Philosophy of Language*

Ferdinand de Saussure (1857–1913) within his structural view of language, made a distinction between paradigmatic and syntagmatic axes that can be seen to correspond to a distinction made by Roman Jakobson (1896–1982) between two types of aphasia. In one, the victims have difficulty with the selection of a term and are dependent upon the context, which is on the contiguity of the discourse to continue to be part of it. By contrast, other aphasics, were seen to substitute equivalent terms e.g., table for lamp, smoke for pipe. Jakobson saw these as examples of metonymy and metaphor on the grounds that these were the most condensed expression of the contiguity or similarity distinctions. This changed perception of the relation between metaphor and metonymy led to developments in structuralism for example, by Levi-Strauss (1908–) in anthropology, and by Jacques Lacan (1901–81) in psychoanalysis. Accepting Jakobson's suggestions, David Lodge offered a typology of modern writing based upon an opposition between metaphor and metonymy (1977).

Heidegger's writing is rich with metaphor yet he denied it. He construed the sensory/nonsensory distinction as belonging to metaphysics and drew the conclusion that 'the metaphorical exists only within the boundaries of the metaphysical.' By expressing a preference for the 'poetical' over the 'mathematical,' Heidegger has been influential in offering a reading of Nietzsche recovering a role for philosophy by means of what we might identify as metaphor of the kind which, as a sentence, is obviously false: 'Philosophy is essentially untimely because it is one of those few things that can never find an echo in the present.'

Derrida (1930–) radicalized Saussure's structural description of language, both in his conceptualizing and also in the style of his writing—in so doing, he matched the congruence in Nietzsche who linked his valuation of metaphor to his style of writing. Derrida's view of language relies upon a play of differences, and understands metaphor as operating within that play of differences—he was unable to use the otherwise conventional distinction between the metaphorical and the literal. His identification of phallogocentrism undermined the metaphor of depth for more perspicuous knowledge, paving the way for a metaphor of breadth. Hans-Georg Gadamer (1900–) understood as a major contributor to hermeneutics also wrote of metaphor.

> It is the genius of linguistic consciousness to be able to give expression to these similarities. This is its fundamental metaphorical nature, and it is important to see that it is the prejudice of a theory of logic that is alien to language if the metaphorical use of a word is regarded as not its real sense (Gadamer 1989, p. 429).

This view strengthens the possibility of substituting a metaphor of breadth for that of depth to describe knowing and understanding, begun by Derrida. Paul Ricoeur (1913–) in a conservative review of the principal writers, sees metaphor as the continuing source for the redescription of the world, but only in a perspective within which philosophy would be concerned to manage the legitimacy of its uses, limiting its influence and restricting any epistemological damage.

While it may be interesting to speculate what part the earlier publication of Ludwig Wittgenstein's (1889–1951) *Tractatus* (1922) might have played in making the *Philosophical Investigations* (1953) as influential as it has proved, the former, though built upon metaphor, has no place for it within its closed system, whereas the latter uses metaphor throughout, but without attempting an explicit theory of it. The style of the *Philosophical Investigations* is as if we hear just one side of a conversation—simply a limitation imposed if the writing is not to be that of drama or a novel? Perhaps the *Investigations* will come to be judged as the writing which, more than any other, licensed the growth of academic interest in metaphor (see *Hermeneutics, Including Critical Theory*; *Linguistic Turn*).

6. *The Tide Turns, The Broader Academic Response*

Kenneth Burke (1897–), following Vico, selected 'metonymy, irony, metaphor and synecdoche' and labeled them the 'Four Master Tropes' and set them to new (as he thought) uses:

> I refer to metaphor, metonymy, synecdoche, and irony. And my primary concern with them here will be not with their purely figurative usage, but with their rôle in the discovery and description of 'the truth.' It is an evanescent moment that we shall deal with—for not only does the dividing line between the figurative and literal uses shift, but also the four tropes shade into one another. Give a man but one of them, tell him to exploit its possibilities, and if he is thorough in doing so, he will come upon the other three' (1969, p. 503).

Four influential interdisciplinary conferences were held in the academic year beginning September 1977, following the publication of a bibliography, *Metaphor* (Shibbles 1971). Later, some 3,500 publications were identified in the period 1985–90 (van Noppen and Hols 1990).

In one, came a significant moment with a sharp difference between Max Black and Donald Davidson. Following Richards and Black, it had been normal to speak of 'literal meaning' and 'metaphorical meaning.' Davidson relied upon a distinction between what words mean and what they do, and took metaphor to belong exclusively to the domain of use. The terms used in metaphor should be held to mean exactly what they mean in their literal use.

Andrew Ortony collected an influential number of papers representing a variety of perspectives on

metaphor (1979). The rate of change merited a second edition as early as 1993 (1993). Within philosophy, Cooper provided a comprehensive and well-received discussion (1986). Seen by many as a significant event, the publication in 1980 by Lakoff and Johnson (1980) of *Metaphors We Live By*, copiously illustrated the endemic nature of metaphor in everyday understanding and extended metaphor's claim to cognitive uses. More recently, this has been developed and broadened by Raymond Gibbs (1994).

By the end of the twentieth century there was scarcely a field of academic enquiry within the behavioral sciences which had not begun to explore its areas of interest with the benefit of an analysis using metaphor and its related tropes as tools e.g., Brown (1977) in sociology, Leary (1990) in psychology, McCloskey (1986) in economics, and much is owed to Hayden White who developed Vico's insights and applied them to the writing of history providing a most virtuous circle (1973, 1985).

7. *Metaphor and 'The Social'*

We have, almost in passing, encountered pointers towards a closer association between the metaphorical and the social. Darwin's *Origin of Species* provided the opportunity to replace a view of language as correspondence with reality, with that of a series of tools to support survival. Metaphor in this view is a tool-making tool. Nietzsche referred to the tropes as 'a sum of human relations.' Cohen (1979) drew attention to similarities between metaphors and jokes, and to the experience which he calls the achievement of intimacy:

> There is a unique way in which the maker and the appreciator of a metaphor are drawn closer to one another. Three aspects are involved: the speaker issues a kind of concealed invitation; the hearer expends a special effort to accept the invitation; and this transaction constitutes the acknowledgment of a community.

This description may easily be extended into a research program. Davidson has recently maintained that thought itself absolutely depends on a three-way relationship between at least two people and a series of events that are shared in the world.

Critical both of Saussure's 'objectivist' linguistics and of 'subjectivist' perspectives, Bakhtin (1895–1975) developed a view of language as sustained and developed in particular social relationships, in turn embedded in wider political and economic conditions and, paradigmatically, with consciousness being thought of as inhering in the person-within-the-group. This view allows for full social interaction, including the non-verbal, to be understood as modifying the associations which attached to the terms deployed in conversation. Wittgenstein's style in the *Philosophical Investigations* is reminiscent of Bakhtin's view.

Although computer technology has made possible the assembly of large corpora of searchable language texts drawn from a variety of publications, it seems not yet to be practicable to incorporate significant contextual detail into these databases, other than of an immediate textual kind, providing a limitation on the development of a social understanding of metaphor, by this route. For the time being, it would seem that some kind of ethnographic research will be most suited to explore the human-relational sources of metaphor use, with Bakhtin's perspective suggesting the possibility of developments in ethnomethodology. Some suggestions, focusing on 'conversational realities,' have been made by John Shotter (1993) and a comprehensive, utopian, and pragmatist redescription of modern culture and politics drawing upon the changing status of metaphor, has been offered by Richard Rorty (1989). Metaphor is the meeting place of the human with the natural.

See also: Anthropology; Aristotle (384–322 BC); Conceptual Blending; Freud, Sigmund (1856–1939); Hobbes, Thomas (1588–1679); Language and Poetic Structure; Language and Thought: The Modern Whorfian Hypothesis; Models, Metaphors, Narrative, and Rhetoric: Philosophical Aspects; Modernity; Nietzsche, Friedrich (1844–1900); Organizations, Metaphors and Paradigms in; Psychoanalysis: Overview; Saussure, Ferdinand de (1857–1913); Wittgenstein, Ludwig (1889–1951)

Bibliography

Aristotle 1984 *The Complete Works of Aristotle: The Revised Oxford Translation*. 2 vols. Princeton University Press, Princeton, NJ

Bakhtin M 1979 *Estetika Slovesnogo Tvorchestra* (Moscow) [English translation 1986 *Speech Genres & Other Late Essays*. University of Texas, Austin, TX]

Black M 1962 *Models and Metaphor*. Cornell University Press, Ithaca, NY

Blumenberg H 1979 *Schiffbruch mit Zuschauer, Paradigma einer Daseins metaphor*. Suhkamp Verlag, Frankfurt am Main, Germany [English translation 1997 *Shipwreck with Spectator Paradigm of a Metaphor for Existence*. MIT Press, Cambridge, MA]

Brown R H 1977 *A Poetic for Sociology*. Cambridge University Press, Cambridge, UK

Burke K 1969 The four master tropes. In: *A Grammar of Motives*. University of California, Berkeley, CA

Cohen T 1979 Metaphor and the cultivation of intimacy. In: Sachs S (ed.) *On Metaphor*. University of Chicago Press, Chicago

Cooper D 1986 *Metaphor*. Blackwell, Oxford, UK

Derrida J 1972 La mythologie blanche. In: *Marges de la Philosophie*. Les Editions de Minuit, Paris [English translation 1982 White mythology. In: *Margins of Philosophy*. Harvester, Hemel Hempstead, UK]

Derrida J 1987 Le retrait de metaphor. *Enclitic* **2**(2): 5–33 [English translation 1998 The retreat of metaphor. In: Wolfreys J (ed.) *The Derrida Reader: Writing Performances*, Edinburgh University Press, UK]

Fiumara G C 1995 *The Metaphoric Process*. Routledge, London

Gadamer H-G 1960 *Wahrheit und Methode*. Mohr (Paul Siebeck), Tübingen, Germany [English translation 1989 *Truth and Method*, 2nd edn. Sheed & Ward, London]
Gibbs R 1994 *The Poetics of Mind*. Cambridge University Press, Cambridge, UK
Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
Leary D E 1990 *Metaphors in the History of Psychology*. Cambridge University Press, Cambridge, UK
Lakoff G, Johnson M 1980 *Metaphors We Live By*. University of Chicago Press, Chicago
Lodge D 1977 *The Modes of Modern Writing*. Arnold, London
McCloskey D N 1986 *The Rhetoric of Economics*. Wheatsheaf, Brighton, UK
Nietzsche F 1979 *Philosophy and Truth*. Humanities Press, Atlantic Highlands, NJ
Ortony A 1993 *Metaphor and Thought*, 2nd edn. Cambridge University Press, Cambridge, UK [1st edn. 1979]
Quine W V 1979 A postscript on metaphor. In: Sachs S (ed.) *On Metaphor*. University of Chicago Press, Chicago
Richards I A 1936 *A Philosophy of Rhetoric*. Oxford University Press, Oxford, UK
Ricoeur P 1975 *La Metaphore Vive*. Éditions du Seuil, Paris [English translation 1978 *The Rule of Metaphor*. Routledge and Kegan Paul, London]
Rorty R 1989 *Contingency, Irony and Solidarity*. Cambridge University Press, Cambridge, UK
Sacks S 1979 *On Metaphor*. University of Chicago Press, Chicago
Shibbles W A 1971 *Metaphor: An Annotated Bibliography and History*. Language Press, Whitewater, WI
Shotter J 1993 *Conversational Realities*. Sage, London
Skinner Q 1996 *Reason and Rhetoric in the Philosophy of Hobbes*. Cambridge University Press, Cambridge, UK
Van Noppen J-P, Hols E 1990 *Metaphor II: A Classified Bibliography of Publications 1985 to 1990*. John Benjamins, Amsterdam
Vico G B 1744 *Scienza Nuova*, 3rd edn. Mosca, Naples, Italy [English translation 1968 *New Science*. Cornell University Press, Ithaca, NY]
White H 1973 *Metahistory: The Historical Imagination in Nineteenth Century Europe*. Johns Hopkins University Press, Baltimore, MD
White H 1985 *Tropics of Discourse*. Johns Hopkins University Press, Baltimore, MD
Wittgenstein L 1953 *Philosophical Investigations*. Blackwell, Oxford, UK

D. Lambourn

Method of Moments

1. Introduction

In many empirical investigations of dynamic economic systems, statistical analysis of a fully specified stochastic process model of the time series evolution is too ambitious. Instead it is fruitful to focus on specific features of the time series without being compelled to provide a complete description. This leads to the investigation of partially specified dynamic models. For instance, the linkages between consumption and asset returns can be investigated without a full description of production and capital accumulation. The behavior of monetary policy can be explored without a complete specification of the macroeconomic economy. Models of inventory behavior can be estimated and tested without requiring a complete characterization of input and output prices. All of these economic models imply moment conditions of the form:

$$Ef(x_t, \beta_o) = 0 \tag{1}$$

where f is known a priori, x_t is an observed time series vector, and β_o is an unknown parameter vector. These moment relations may fall short of providing a complete depiction of the dynamic economic system. Generalized Method of Moments (GMM) estimation, presented in Hansen (1982), aims to estimate the unknown parameter vector β_o and test these moment relations in a computationally tractable way. This entry first compares the form of a GMM estimator to closely related estimators from the statistics literature. It then reviews applications of these estimators to partially specified models of economic time series. Finally, it considers GMM-related moment-matching problems in fully specified models economic dynamics.

2. Minimum Chi-square Estimation

To help place GMM estimation in a statistical context, I explore a closely related minimum chi-square estimation method. Statisticians developed minimum chi-square estimators to handle restricted models of multinomial data and a variety of generalizations. Neyman (1949) and Burankin and Gurland (1951), among others, aimed to produce statistically efficient and computationally tractable alternatives to maximum likelihood estimators. In the restricted multinomial model, estimators are constructed by forming empirical frequencies and minimizing Pearson's chi-square criterion or some modification of it.

The method has direct extensions to any moment-matching problem. Suppose that $\{x_t\}$ is a vector process, which temporarily is treated as being *iid*. Use a function ψ with n coordinates to define target moments associated with the vector x_t. A model takes the form:

$$E[\psi(x_t)] = \phi(\beta_o)$$

where β_o is an unknown parameter. The moment-matching problem is to estimate β_o by making the empirical average of $\{\psi(x_t)\}$ close to its population counterpart $\phi(\beta_o)$:

$$\min_{\beta} \frac{1}{T} \sum_{t=1}^{T} [\psi(x_t) - \phi(\beta)]' \, V \sum_{t=1}^{T} [\psi(x_t) - \phi(\beta)] \tag{2}$$

where V is the distance or weighting matrix. The distance matrix sometimes depends on the data and/or

the candidate parameter vector β. The use of V to denote a weighting matrix that may actually depend on parameters or data is an abuse of notation. This simple notation is used because it is the probability limit of the weighting matrix evaluated at the parameter estimator that dictates the first-order asymptotic properties.

The limiting distribution of the criterion in Eqn. (2) is chi-square distributed with n degrees of freedom if the parameter vector β_o is known, and if V is computed by forming the inverse of either the population or sample covariance matrix of $\psi(x)$. When β is unknown, estimates may be extracted by minimizing this chi-square criterion; hence the name. To preserve the chi-square property of the minimum (with an appropriate reduction in the degrees of freedom), we again form the inverse sample covariance matrix of $\psi(x)$, or form the inverse population covariance matrix for each value of β. The minimized chi-square property of the criterion may be exploited to build tests of overidentification and to construct confidence sets for parameter values. Results like these require extra regularity conditions, and this rigor is supplied in some of the cited papers.

While the aim of this research was to form computationally tractable alternatives to maximum likelihood estimation, critical to statistical efficiency is the construction of a function ψ of the data that is a sufficient statistic for the parameter vector (see Burankin and Gurland 1951). Berkson (1944) and Taylor (1953) generalize the minimum chi-square approach by taking a smooth one-to-one function h and building a quadratic form of $h[\frac{1}{T}\sum_t \phi(x_t)]-h[\psi(\beta)]$. Many distributions fail to have a finite number of sufficient statistics; but the minimum chi-square method continues to produce consistent, asymptotically normal estimators provided that identification can be established.

GMM estimators can have a structure very similar to the minimum chi-square estimators. Notice that the core ingredient to the moment-matching problem can be depicted as in Eqn. (1) with a separable function f:

$$f(x, \beta) = \psi(x) - \phi(\beta)$$

used in the chi-square criterion function. Target moments are one of many ways for economists to construct inputs into chi-square criteria, and it is important to relax this separability. Moreover, in *GMM* estimation, the emphasis on statistical efficiency is weakened in order to accommodate partially specified models. Finally, an explicit time series structure is added, when appropriate.

3. GMM Estimation

Our treatment of *GMM* estimation follows Hansen (1982), but it builds from Sargan's (1958, 1959) analyses of linear and nonlinear instrumental variables (see *Instrumental Variables in Statistics and Econometrics*). See Ogaki (1993) for a valuable discussion of the practical implementation of *GMM* estimation methods. *GMM* estimators are constructed in terms of a function f that satisfies Eqn. (1) where f has more coordinates, say n, than there are components to the parameter vector β_o. Another related estimation method is M-estimation. M-estimation is a generalization of maximum likelihood and least squares estimation. M-estimators are typically designed to be less sensitive to specific distributional assumptions (see *Robustness in Statistics*). These estimators may be depicted as solving a sample counterpart to Eqn. (1) with a function f that is nonseparable, but with the same number of moment conditions as parameter estimators.

In Sects. 4–6 we will give examples of the construction of the f function, including ones that are not separable in x and β and ones for which there have more coordinates than parameter vectors. A minimum chi-square type criterion is often employed in *GMM* estimation. For instance, it is common to define the *GMM* estimator as the solution to:

$$b_T = \arg\min_{\beta} \; T g_T(\beta)' V g_T(\beta)$$

where

$$g_T(\beta) = \frac{1}{T}\sum_{t=1}^{T} f(x_t, \beta)$$

and V is a positive definite weighting matrix. This quadratic form has the chi-square property provided that V is an estimator of the inverse of an appropriately chosen covariance matrix, one that accounts for temporal dependence.

The sections that follow survey some applications of *GMM* estimators to economic time series. A feature of many of these examples is that the parameter β by itself may not admit a full depiction of the stochastic process that generates data. *GMM* estimators are constructed to achieve 'partial' identification of the stochastic evolution and to be robust to the remaining unmodeled components.

3.1 Time Series Central Limit Theory

Time series estimation problems must make appropriate adjustments for the serial correlation for the stochastic process $\{f(x_t, \beta_o)\}$. A key input into the large sample properties of *GMM* estimators is a central limit approximation:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f(x_t, \beta_o) \Rightarrow \text{Normal}\,(0, \Sigma_o)$$

for an appropriately chosen covariance matrix Σ_o. An early example of such a result was supplied by Gordin (1969) who was used martingale approximations for

partial sums of stationary, ergodic processes. See Hall and Heyde (1980) for an extensive discussion of this and related results. The matrix Σ_o must include adjustments for temporal dependence:

$$\Sigma_o = \sum_{j=-\infty}^{+\infty} Ef(x_t, \beta_o) f(x_{t-j}, \beta_o)'$$

which is the long-run notion of a covariance matrix that emerges from spectral analysis of time series. In many *GMM* applications, martingale arguments show that the formula for Σ_o simplifies to include only a small number of nonzero terms. It is the adjustment to the covariance matrix that makes the time series implementation differ from the *iid* implementation (Hansen 1982).

Adapting the minimum chi-square apparatus to this environment requires that we estimate the covariance matrix Σ_o. Since f is not typically separable in x and β, an estimator of Σ_o requires an estimator of β_o. This problem is familiar to statisticians and econometricians through construction of feasible generalized least squares estimators. One approach is to form an initial estimator of β_o with an arbitrary nonsingular weighting matrix V and to use the initial estimator of β_o to construct an estimator of Σ_o. Hansen (1982), Newey and West (1987) and many others provide consistency results for the estimators of Σ_o. Another approach is to iterate back and forth between parameter estimation and weighting matrix estimation until a fixed point is reached, if it exists. A third approach is to construct an estimator of $\Sigma(\beta)$ and to replace V in the chi-square criterion by an estimator of $\Sigma(\beta)^{-1}$ constructed for each β. Given the partial specification of the model, it is not possible to construct $\Sigma(\beta)$ without the use of the time series data. Long run covariance estimates, however, can be formed for the process $\{f(x_t, \beta)\}$ for each choice of β. Hansen et al. (1996) refer to this method as *GMM* estimation with a continuously updated weighting matrix. Hansen et al. (1996), Newey and Smith (2000), and Stock and Wright (2000) describe advantages to using continuous-updating, and Sargan (1958) shows that in some special circumstances this method reproduces a quasi-maximum likelihood estimator.

3.2 Efficiency

Since the parameter vector β that enters moment condition (1) may not fully characterize the data evolution, direct efficiency comparisons of *GMM* estimators to parametric maximum likelihood are either not possible, or not interesting. However, efficiency statements can be made for narrower classes of estimators.

For the study of *GMM* efficiency, instead of beginning with a distance formulation, index a family of *GMM* estimators by the moment conditions used in estimation. Specifically study

$$a g_T(b_T) = 0$$

where a is a k by n selection matrix. The selection matrix isolates which (linear combination of) moment conditions will be used in estimation and indexes alternative *GMM* estimators. Estimators with the same selection matrix have the same asymptotic efficiency. Without further normalizations, multiple indices imply the same estimator. Premultiplication of the selection matrix a by a nonsingular matrix e results in the same system of nonlinear equations. In practice, the selection matrix can depend on data or even the parameter estimator provided that the selection matrix has a probability limit. As with weighting matrices for minimum chi-square criteria, we suppress the possible dependence of the selection matrix a on data or parameters for notational simplicity. The resulting *GMM* estimators are asymptotically equivalent (to possibly infeasible) estimators in which the selection matrix is replaced by its probability limit. As has been emphasized by Sargan (1958, 1959) in his studies of instrumental variables estimators, estimation accuracy can be studied conveniently as a choice a^* of an efficient selection matrix. The link between a weighting matrix V and selection matrix a is seen in the first-order conditions:

$$a = V \frac{1}{T} \sum_{t=1}^{T} \frac{\partial f(x_t, b_T)}{\partial \beta}$$

or their population counterpart:

$$a = Vd$$

where

$$d = E\left[\frac{\partial f(x_t, \beta_o)}{\partial \beta}\right]$$

Other distance measures including analogs to the ones studied by Berkson (1944) and Taylor (1953) can also be depicted as a selection matrix applied to the sample moment conditions $g_T(\beta)$. Moreover, *GMM* estimators that do not solve a single minimization problem may still be depicted conveniently in terms of selection matrices. For example, see Heckman (1976) and Hansen (1982) for a discussion of recursive estimation problems, which require solving two minimization problems in sequence.

Among the class of estimators indexed by a, the ones with the smallest asymptotic covariance matrix satisfy:

$$a = ed' \Sigma_o^{-1}$$

where e is any nonsingular matrix, and the best covariance matrix is $d'(\Sigma_o)^{-1}d$. This may be shown by imitating the proof of the famed Gauss-Markov Theorem, which establishes that the ordinary least squares estimator is the best, linear unbiased estimator.

3.3 Semiparametric Efficiency

Many *GMM* applications, including ones that we describe subsequently, imply an extensive (infinite) collection of moment conditions. To apply the previous analysis requires an *ad hoc* choice of focal relations to use in estimation. Hansen (1985) extends this indexation approach to time-series problems with an infinite number of moment conditions. The efficiency bound for an infinite family of *GMM* estimators can be related to the efficiency of other estimators using a semiparametric notion of efficiency. A semiparametric notion is appropriate because an infinite-dimensional nuisance parameter is needed to specify fully in the underlying model. It has been employed by Chamberlain (1987) and Hansen (1993) in their study of *GMM* estimators constructed from conditional moment restrictions.

4. Linear Models

Researchers in econometrics and statistics have long struggled with the idea of how to identify an unknown coefficient vector β_o in a linear model of the form:

$$\beta_o \cdot y_t = u_t$$

where y_t is a k-dimensional vector of variables observed by an econometrician. Least squares solves this problem by calling one of the variables, y_{1t}, the dependent variable and requiring the remaining variables, y_{2t}, to be orthogonal to the disturbance term:

$$E(u_t y_{2t}) = 0$$

Alternatively, as suggested by Karl Pearson and others, when there is no natural choice of a left-hand side variable, we may identify β_o as the first principal component, the linear combination of y_t with maximal variance subject to the constraint $|\beta| = 1$.

A third identification scheme exploits the time series structure and has an explicit economic motivation. It is a time-series analog to the instrumental variables and two-stage least squares estimators familiar to economists. Suppose that a linear combination of y_t cannot be predicted given data sufficiently far back into the past. That is,

$$E(\beta_o \cdot y_t \,|\, \mathscr{F}_{t-m}) = 0 \tag{3}$$

where $\mathscr{F}_t$ is a conditioning information set that contains at least current and past values of y_t. This conditional moment restriction can be used to identify the parameter vector β_o, up to scale. In this setup there may be no single variable to be designated as endogenous with the remainder being exogenous or even predetermined. Neither least squares nor principal components are appropriate for identifying β_o.

The model or the precise context of the application dictates the choice of lag m. For instance, restriction (3) for a specified value of m follows from martingale pricing relations for multiperiod securities, from Euler equations from the investment problems faced by decision-makers, or from the preference horizons of policy-makers.

To apply *GMM* to this problem, use Eqn. (3) to deduce the matrix equation:

$$E(z_{t-m} y_t')\, \beta_o = 0$$

where z_t is an n-dimensional vector of variables in the conditioning information set $\mathscr{F}_t$ and $n \geqslant k-1$. By taking unconditional expectations we see that β_o is in the null space of the n by k matrix $E(z_{t-m} y_t')$. The model is over-identified if $n \geqslant k$. In this case the matrix $E(z_{t-m} y_t')$ must be of reduced rank and in this sense is special. This moment condition may be depicted as in Eqn. (1) with:

$$f(x_t, \beta) = z_{t-m} y_t' \beta_o$$

where x_t is a vector containing the entries of z_{t-m} and y_t. The *GMM* test of over-identification, based on say a minimum chi-square objective, aims to detect this reduced rank. The *GMM* estimator of β_o seeks to exploit this reduced rank by approximating the direction of the null space.

For the *GMM* applications in this and Sect. 5, there is a related but independent statistics literature. Godambe and Heyde (1987) and others study efficiency criteria of martingale estimation equations for a fully specified probability model. By contrast, our models are partially specified and have estimation equations, e.g. $\sum_{t=1}^{T} [z_{t-m} y_t'] \beta_o$, that is a martingale only when $m = 1$. When the estimation equation is not a martingale, Hansen (1982) and Hansen (1985) construct an alternative martingale approximation to analyze statistical efficiency.

Given that the null space of $Ez_{t-m} y_t'$ is not degenerate, it might have more dimensions than one. While the moment conditions are satisfied (the null space of $E(z_{t-m} y_t')$ is nondegenerate), the parameter vector itself is under-identified. Recent literature on weak instruments is aimed at blurring the notion of being underidentified. See Stock and Wright (2000) for an analysis of weak instruments in the context of *GMM* estimation.

As posed here the vector β_o is not identified but is at best identified up to scale. This problem is analogous to that of principal component analysis. At best we might hope to identify a one-dimensional subspace. Perhaps a sensible estimation method should therefore

locate a subspace rather than a parameter vector. Normalization should be inessential to the identification beyond mechanically selecting the parameter vector within this subspace. See Hansen et al. (1996) for a discussion of normalization invariant *GMM* estimators. This argument is harder to defend than it might seem at first blush. Prior information about the normalized parameter vector of interest may eliminate interest in an estimation method that is invariant to normalization.

There are some immediate extensions of the previous analysis. The underlying economic model may impose further, possibly nonlinear, restrictions on the linear subspace. Alternatively, the economic model might imply multiple conditional moment relations. Both situations are straightforward to handle. In the case of multiple equations, prior restrictions are needed to distinguish one equation from another, but this identification problem is well studied in the econometrics literature.

4.1 Applications

A variety of economic applications produce conditional moment restrictions of the form in Eqn. (3). For instance, Hansen and Hodrick (1980) studied the relation between forward exchange rates and future spot rates where m is the contract horizon. Conditional moment restrictions also appear extensively in the study of a linear-quadratic model of production smoothing and inventories (see West 1995 for a survey of this empirical literature). The first-order or Euler conditions for an optimizing firm may be written as Eqn. (3) with $m = 2$. Hall (1988) and Hansen and Singleton (1996) also implement linear conditional-moment restrictions in the lognormal models of consumption and asset returns. These moment conditions are again derived from first-order conditions, but in this case for a utility-maximizing consumer/investor with access to security markets. Serial correlation due to time aggregation may cause m to be two instead of one. Finally, the linear conditional moment restrictions occur in the literature on monetary policy response functions when the monetary authority is forward-looking. Clarida et al. (2000) estimate such models in which m is dictated by preference horizon of the Federal Reserve Bank in targeting nominal interest rates.

4.2 Efficiency

The choice of z_{t-m} in applications is typically *ad hoc*. The one-dimensional conditional moment restriction of Eqn. (3) actually gives rise to an infinite number of unconditional moment restrictions through the choice of the vector z_{t-m} in the conditioning information set $\mathscr{F}_{t-m}$. Any notion of efficiency based on a preliminary choice z_{t-m} runs the risk of missing some potentially important information about the unknown parameter vector β_o. It is arguably more interesting to examine asymptotic efficiency across an infinite-dimensional family of *GMM* estimators indexed by feasible choices of z_{t-m}. This is the approach adopted by Hansen (1985), Hansen et al. (1988) and West (2000). The efficient *GMM* estimator that emerges from this analysis is infeasible because it depends on details of the time series evolution. West and Wilcox (1996) and Hansen and Singleton (1996) construct feasible counterpart estimators based on approximating the time series evolution needed to construct the efficient z_{t-m}. In particular, West and Wilcox (1996) show important improvements in the efficiency and finite sample performance of the resulting estimators.

Efficient *GMM* estimators based on an infinite number of moment conditions often place an extra burden on the model-builder by requiring that a full dynamic model be specified. The cost of misspecification, however, is not so severe. While a mistaken approximation of the dynamics may cause an efficiency loss, by design it will not undermine the statistical consistency of the *GMM* estimator.

5. Models of Financial Markets

Models of well-functioning financial markets often take the form:

$$E\,(d_t z_t \,|\, \mathscr{F}_{t-m}) = q_{t-m} \tag{4}$$

where z_t is a vector of asset payoffs at date t, q_{t-m} is a vector of the market prices of those assets at date t, and $\mathscr{F}_t$ is an information set available to economic investors at date t. By assumption, the price vector q_t is included in the information set $\mathscr{F}_t$. The random variable d_t is referred to as a 'stochastic discount factor' between dates $t-m$ and date t. This discount factor varies with states of the world that are realized at date t and encodes risk adjustments in the security market prices. These risk adjustments are present because some states of the world are discounted more than others, which is then reflected in the prices. The existence of such a depiction of asset prices is well known since the work of Harrison and Kreps (1979), and its conditional moment form used here and elsewhere in the empirical asset pricing literature is justified in Hansen and Richard (1987). In asset pricing formula (4), m is the length of the financial contract, the number of time periods between purchase date and payoff date.

An economic model of financial markets is conveniently posed as a specification of the stochastic discount factor d_t. It is frequently modeled parametrically:

$$d_t = g(y_t, \beta_o) \tag{5}$$

where the function g is given *a priori* but the parameter β_o is unknown, and a target of estimation. Models of investor preferences may be written in this manner as can observable factor models in which d_t is a function (often linear) or vector y_t of 'factors' observed by an econometrician. See Cochrane (2001) for examples and a discussion of how this approach is connected to other empirical methods in financial economics.

To apply *GMM* estimation, inference, and testing to this problem we do two things. First we replace Eqn. (4) by its unconditional counterpart:

$$E(d_t z_t - q_{t-m}) = 0$$

which can be justified by applying the Law of Iterated Expectations. Second we substitute Eqn. (5) into this unconditional moment implication:

$$E[g(y_t, \beta_o) z_t - q_{t-m}] = 0$$

This may be depicted as (1) by writing:

$$f(x_t, \beta) = g(y_t, \beta) z_t - q_{t-m}$$

where x_t contains the entries of the factors y_t, the asset payoffs z_t, and the corresponding prices q_{t-m}. As is evident from Hansen and Singleton (1982) and Kocherlakota (1996), *GMM*-based statistical tests of this model give rise to one characterization of what is commonly termed as the 'equity-premium puzzle.' This puzzle is a formal statement of how the observed risk–return trade-off from financial market data is anomalous when viewed through the guises of many standard dynamic equilibrium models from the macroeconomics and finance literatures.

Studying unconditional rather than conditional moment relations may entail a loss of valuable information for estimation and testing. As emphasized by Hansen and Richard (1987), however, information in the set $\mathscr{F}_{t-m}$ may be used by an econometrician to form synthetic payoffs and prices. This information is available at the initiation date of the financial contract. While use of conditioning information to construct synthetic portfolios reduces the distinction between conditional and unconditional moment restrictions, it introduces additional challenges for estimation and inference that are being confronted in ongoing econometric research.

6. *From Densities to Diffusions*

A century ago Karl Pearson proposed a family of density functions and a moments-based approach to estimating a parameterized family of densities. The densities within this family have logarithmic derivatives that are the ratio of a first to a second-order polynomial. More recently, Wong (1964) provided scalar diffusion models with stationary distributions in the Pearson family. Diffusion models are commonly used in economic dynamics and finance.

A scalar diffusion is a solution to a stochastic differential equation:

$$\mathrm{d}x_t = \mu(x_t)\,\mathrm{d}t + \sigma(x_t)\,\mathrm{d}B_t$$

where μ is the local mean or drift for the diffusion, σ^2 is the local variance or diffusion coefficient, and B_t is standard Brownian motion. I now revisit Pearson's estimation problem and method, but in the context of a data generated by a diffusion.

The stationary density q of a diffusion satisfies the integral equation:

$$\int_{\underline{r}}^{\bar{r}} \left(\mu\phi + \frac{\sigma^2}{2}\frac{\mathrm{d}\phi}{\mathrm{d}x} \right) q\,\mathrm{d}x = 0 \tag{6}$$

for a rich class of ϕ's, referred to as test functions. This gives an extensive family of moment conditions for any candidate (μ, σ^2). In particular, moment conditions of the form (1) may be built by parameterizing μ and σ^2 and by using a vector of test functions ϕ. Pearson's moment recursions are of this form with test functions that are low-order polynomials.

Pearson's method of estimation was criticized by R. A. Fischer because it failed to be efficient for many members of the Pearson family of densities. Pearson and Fischer both presumed that the data generation is *iid*. To attain asymptotic efficiency with *iid* data, linear combinations of $f(x, \beta_o)$ should reproduce the score vector for an implied likelihood function. Low-order polynomials fail to accomplish this for many members of the Pearson family of densities, hence the loss in statistical efficiency.

When the data are generated by a diffusion, the analysis of efficiency is altered in a fundamental way. Parameterize the time series model via $(\mu_\beta, \sigma_\beta^2)$. There is no need restrict μ and σ^2 to be low-order polynomials. Let Φ be a vector of test functions with at least as many component functions as parameters to estimate. Associated with each Φ is a *GMM* estimator by forming:

$$f(x, \beta) = \mu_\beta(x)\Phi(x) + \frac{1}{2}\sigma_\beta^2(x)\frac{\mathrm{d}\Phi}{\mathrm{d}x}(x)$$

The moment conditions (1) follow from Eqn. (6).

Conley et al. (1997) calculate the efficiency bounds for this estimation problem using the method described in Hansen (1985). They perform this calculation under the simplifying fiction that a continuous-data record is available and show that an efficient choice of Φ is:

$$\Phi_\beta(x) = \frac{\partial}{\partial\beta}\left[\frac{2\mu_\beta(x) + \mathrm{d}\sigma_\beta^2(x)/\mathrm{d}x}{\sigma_\beta^2(x)} \right] \tag{7}$$

The choice of test function turns out to be the 'derivative' of the score vector with respect to the Markov state x. This efficient choice of Φ depends on

the true parameter vector β_o and hence is infeasible to implement. Conley et al. (1997) show that the same efficiency can be attained by a feasible estimator in which Φ is allowed to depend on β as in Eqn. (7). The score function derivative is easier to compute than the score function itself because the constant of integration for implied density does not have to be evaluated for each β. While tractable, this efficient test function solution to the time series problem will typically not lead one to use low-order polynomials as test functions even for Pearson's parameterizations.

Pearson used moment recursions to construct tractable density estimates without numerical integration. Computational concerns have subsided, and the diffusion model restricts more than just the stationary density. Transition densities can also be inferred numerically for a given (μ, σ^2) pair. To motivate fitting only the stationary density, a model-builder must suppose potential model misspecification in the transition dynamics. One example of this form of misspecification is a subordinated diffusion model used to model financial time series in which calendar time and a more relevant information-based notion of economic time are distinct.

Integral equation (6) is 'localized' by using smooth test functions that concentrate their mass in the vicinity of given points in the state space. Banon (1978) exploited this insight to produce a nonparametric drift estimator using a locally constant parameterization of μ. Conley et al. (1997) justify a local linear version of the *GMM* test function estimator described in the previous subsection.

7. Related Approaches

Since around 1990 statisticians have explored empirical likelihood and other related methods. These methods are aimed at fitting empirical data distributions subject to *a priori* restrictions, including moment restrictions that depend on an unknown parameter. In particular, Qin and Lawless (1994) have shown how to use empirical likelihood methods to estimate parameters from moment restrictions like those given in (1) for *iid* data. While Qin and Lawless (1994) use the statistics literature on estimation equations for motivation, they were apparently unaware of closely-related econometrics literature on *GMM* estimation. Baggerly (1998) describes a generalization of the method of empirical likelihood based on the Cressie-Read divergence criterion. Imbens et al. (1998) and Bonnal and Renault (2001) use this generalization to unify the results of Qin and Lawless (1994), Imbens (1997), Kitamura and Stutzer (1997), and others and *GMM* estimation with a continuously-updated weighting matrix. (See Newey and Smith (2000) for a related discussion.) Within the *iid* framework, Bonnal and Renault (2001) show that the continuously updated *GMM* estimator is a counterpart to empirical likelihood except that it uses Pearson's χ^2 criterion. The relative entropy or information-based estimation of Imbens (1997) and Kitamura and Stutzer (1997) is of the same type but based on an information criterion of fit for the empirical distribution.

Kitamura and Stutzer (1997) and others use clever blocking approaches for weakly dependent data to adapt these methods to time series estimation problems. Many *GMM* applications imply 'conditional' moment restrictions where the conditioning information is lagged m time periods. Extensions of these empirical distribution methods to accommodate time series 'conditional' moment implications is an important area of research.

8. Moment-matching Reconsidered

Statistical methods for partially specified models allow researchers to focus an empirical investigation and to understand sources of empirical anomalies. Nevertheless, the construction of fully specified models is required to address many questions of interest to economists such as the effect of hypothetical interventions or policy changes.

Models of economic dynamical systems remain highly stylized, however; and they are not rich enough empirically to confront a full array of empirical inputs. Producing interesting comparative dynamic results for even highly stylized dynamic systems is often difficult, if not impossible, without limiting substantially the range of parameter values that are considered. Since analytical solutions are typically not feasible, computational methods are required. As in other disciplines, this has led researchers to seek ways to calibrate models based on at least some empirical inputs. See Hansen and Heckman (1996) for a discussion of this literature. The analysis of dynamical economic systems brings to the forefront both computational and conceptual problems.

The computational problems are reminiscent to those confronted by the inventors of minimum chi-square methods. Two clever and valuable estimation methods closely related to moment matching have recently emerged from the econometrics literature. One method is called 'indirect inference' and has been advanced by Smith (1993) and Gourieroux et al. (1993). The idea is to fit a conveniently chosen, but misspecified approximating model that is easy to estimate. The empirical estimates from the approximating model are used as targets for the estimation of the dynamic economic model. These targets are 'matched' using a minimum chi-square criteria.

This method is sometimes difficult to implement because the implied approximating models associated with the dynamic economic model of interest may be hard to compute. Gallant and Tauchen (1996) circumvent this difficulty by using the score function for the approximating model (evaluated at the empirical estimates) as targets of estimation. The computational

burden is reduced to evaluating an empirical score expectation as a function of the underlying parameter of interest. This allows researchers to use an expanded set of approximating statistical models.

In both of these methods there are two models in play, an approximating statistical model and an underlying economic model. These methods address some of the computational hurdles in constructing parameter estimators through their use of convenient approximating statistical models. On the other hand, it is harder to defend their use when underlying dynamic economic model is itself misspecified. Gallant and Tauchen (1996), for instance, use statistical efficiency as their guide in choosing approximating statistical models. Better approximating models results in more accurate estimators of the parameters of interest. This justification, however, neglects the role of misspecification in the underlying dynamic economic model. For 'indirect inference' it is often not evident how to construct approximating models that leave the estimates immune to the stylized nature of the underlying economic model. Thus it remains an important challenge for econometricians to devise methods for infusing empirical credibility into 'highly stylized' models of dynamical economic systems. Dismissing this problem through advocating only the analysis of more complicated 'empirically realistic' models will likely leave econometrics and statistics on the periphery of important applied research. Numerical characterizations of highly stylized models will continue with or without the aid of statistics and econometrics.

Bibliography

Baggerly K A 1998 Empirical likelihood as a goodness-of-fit measure. *Biometrika* **85**: 535–47

Banon G 1978 Nonparametric identification for diffusion processes. *SIAM Journal of Control and Optimization* **16**: 380–95

Berkson J 1944 Application of the logistic function to bioassay. *Journal of the American Statistical Association* **39**: 357–65

Bonnal H, Renault E 2001 *Minimal Chi-square Estimation with Conditional Moment Restrictions*. CIRANO, Montreal, PQ

Burankin E, Gurland J 1951 On asymptotically normal estimators: I. *University of California Publications in Statistics* **1**: 86–130

Chamberlain G 1987 Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34**: 305–34

Clarida R, Gali J, Gertler M 2000 Monetary policy rules and macroeconomic stability: Evidence and some theory. *Quarterly Journal of Economics* **115**: 147–80

Cochrane J 2001 *Asset Pricing*. Princeton University Press, Princeton, NJ

Conley T G, Hansen L P, Luttmer E G J, Scheinkman J A 1997 Short-term interest rates as subordinated diffusions. *Review of Financial Studies* **10**: 525–77

Gallant A R, Tauchen G 1996 Which moments to match. *Econometric Theory* **12**: 657–81

Godambe V P, Heyde C C 1987 Quasi-likelihood and optimal estimation. *International Statistical Review* **55**: 231–44

Gordin M I 1969 Central limit theorem for stationary processes. *Soviet Mathematics. Doklady* **10**: 1174–6

Gourieroux C, Monfort A, Renault E 1993 Indirect inference. *Journal of Applied Econometrics* **8**: S85–S118

Hall P, Heyde C C 1980 *Martingale Limit Theory and Its Application*. Academic Press, Boston

Hall R E 1988 Intertemporal substitution and consumption. *Journal of Political Economy* **96**: 339–57

Hansen L P 1982 Large sample properties of generalized method of moments estimators. *Econometrica* **50**: 1029–54

Hansen L P 1985 A method for calculating bounds on asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics* **30**: 203–38

Hansen L P 1993 Semiparametric efficiency bounds for linear time-series models. In: Phillips P C B (ed.) *Models, Methods and Applications of Econometrics: Essays in Honor of A. R. Bergstrom*. Blackwell, Cambridge, MA Chap. 17, pp. 253–71

Hansen L P, Heaton J C, Ogaki M 1988 Efficiency bounds implied by multi-period conditional moment restrictions. *Journal of the American Statistical Society* **88**: 863–71

Hansen L P, Heaton J C, Yaron A 1996 Finite sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* **14**: 262–80

Hansen L P, Heckman J J 1996 The empirical foundations of calibration. *Journal of Economic Perspectives* **10**: 87–104

Hansen L P, Hodrick R J 1980 Forward exchange rates as optimal predictors of future spot rates. *Journal of Political Economy* **88**: 829–53

Hansen L P, Richard S F 1987 The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. *Econometrica* **55**: 587–613

Hansen L P, Singleton K J 1982 Generalized instrumental variables of nonlinear rational expectations models. *Econometrica* **50**: 1269–86

Hansen L P, Singleton K J 1996 Efficient estimation of linear asset pricing models with moving-average errors. *Journal of Business and Economic Statistics* **14**: 53–68

Harrison J M, Kreps D M 1979 Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* **20**: 381–408

Heckman J J 1976 The common structure of statistical methods of truncation, sample selection, and limited dependent variable and a simple estimator for such models. *Annals of Economic and Social Measurement* **5**: 475–92

Imbens G W 1997 One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies* **64**: 359–83

Imbens G W, Spady R H, Johnson P 1998 Information theoretic approaches to inference in moment condition estimation. *Econometrica* **66**: 333–57

Kitamura Y, Stutzer M 1997 An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65**: 861–74

Kocherlakota N 1996 The equity premium: It's still a puzzle. *Journal of Economic Literature* **34**: 42–71

Newey W K, West K D 1987 A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**: 703–8

Newey W K, Smith R J 2000 Asymptotic bias and equivalence of gmm and gel. Manuscript

Neyman J 1949 Contributions to the theory of the χ^2 test. *Proceedings of the Berkeley Symposium on Mathematical*

Statistics and Probability. University of California Press, CA
Ogaki M 1993 Generalized method of moments: Econometric applications. In: Maddala G S, Rao C R, Vinod H D (eds.) *Handbook of Statistics*. Elsevier Science, Amsterdam, Vol. 11, Chap. 17, pp. 455–86
Qin J, Lawless J 1994 Empirical likelihood and general estimating equations. *Annals of Statistics* **22**: 300–25
Sargan J D 1958 The estimation of economic relationships using instrumental variables. *Econometrica* **26**: 393–415
Sargan J D 1959 The estimation of relationships with autocorrelated residuals by the use of instrumental variables. *Journal of the Royal Statistical Society* **21**: 91–105
Smith A 1993 Estimating nonlinear time series models using simulated vector autoregressions. *Journal of Applied Econometrics* **8**: 63–84
Stock J H, Wright J H 2000 GMM with weak identification. *Econometrica* **68**: 1055–96
Taylor W 1953 Distance functions and regular best asymptotically normal estimates. *Annals of Mathematical Statistics* **24**: 85–92
West K D 1995 Inventory models. In: *Handbook in Applied Econometrics (Macroeconometrics)*. Basil Blackwell, Oxford, UK
West K D 2000 On optimal instrumental variables estimation of stationary time series models. *International Economic Review* forthcoming.
West K D, Wilcox D W 1996 A comparison of alternative instrumental variables estimators of a dynamic linear model. *Journal of Business and Economic Statistics* **14**: 281–93
Wong E 1964 The construction of a class of stationary markoff processes. In: Bellman R (ed.) *Sixteenth Symposium in Applied Mathematics—Stochastic Processes in Mathematical Physics and Engineering*. American Mathematical Society, Providence, RI, pp. 264–76

L. P. Hansen

Methodological Individualism in Sociology

Methodological individualism in sociology refers to the explanatory and modeling strategies in which human individuals (with their motivations) and human actions (with their causes or reasons) are given a prominent role in explanations and models. Social phenomena are viewed as the aggregate results of individual actions. Explanation thus proceeds from the parts to the whole: individual action has an explanatory primacy in relation to social facts, society's properties, and observed macroregularities. Among the problems associated with this methodology, the following three are of special interest: (a) how should individual characteristics be selected and connected? (b) does the relevance of individualism as a methodology depend on the way models and theories are used?, and (c) how should individualistic social science take individual cognition into account?

1. Individualism and the Form of Explanations

1.1 Giving Primacy to Individuals

While the constitution of sociology as an autonomous discipline has involved the recognition of a separate layer of 'social facts,' methodological individualism, which presupposes the existence of social facts as an *explanandum* for social science, is not alien to the sociological tradition, as exemplified by the work of Weber, Pareto, and others. A general feature of individualistic explanations is that individual motivations, preferences, reasons, propensities, or individual characteristics generally speaking figure explicitly in the proposed models and explanations, together with the description of relevant technological and natural circumstances.

Methodological individualism has gained influence in the twentieth century through the work of such authors as Mises, Popper, and Hayek, but its emergence is traceable to the debates in nineteenth century Germany and Austria about the nature of explanation in history, the status of economic theory, and the respective scientific roles of nomological explanation and particular understanding. It is associated classically with the requirement of a real understanding of the motivations of the social actors themselves (as illustrated by Simmel 1900, Weber 1922).

This methodology can be contrasted with several types of nonindividualistic methods (Boyer 1992). It is violated by those theories which rely on the operation of impersonal forces (such as nature, mind, history, progress, or destiny), and by 'holistic' explanations in which the properties of collective entities (such as nations, classes, social groups, or society as a whole) or unconscious forces are given an independent explanatory role.

Methodological individualism is compatible with ontological holism about collective entities, in the following sense: individualistic social scientists may recognize the existence of social entities (such as 'cultures' or 'traditions') which are irreducible to individual component parts. But they postulate that only individuals have goals and interests, and that these have explanatory value with respect to human conduct. They reject or reinterpret the notion of collective belief. Finally, they recognize that the social set-up can be transformed through the action of individuals. This makes methodological individualism hardly compatible with the more deterministic versions of historical materialism, although it is congenial to some Marxian themes, in particular Marx's criticism of the belief in the historical effectiveness of abstract notions of man, society, and consciousness.

1.2 Individual Motivation

Some individual characteristics should figure in the *explanans* in an explicit manner. But which are the

relevant ones? The selection should allow the theorist to reconstruct some basic facts about the choices of reasoning agents, and formulate those relationships which are considered useful for explanation or prediction. Both cognitive factors (such as beliefs or probability judgements) and volitional ones (desires or preferences concerning states of affairs) play a decisive role. There is some uncertainty, however, in the determination of the appropriate concepts of beliefs and preferences. In particular, should they be revealed by introspection, or in actual choice?

Behavioral social scientists insist that they should be inferred from actual choices, but most social scientists hypothesize in a bolder way. In particular, the Popperian version of methodological individualism emphasizes the crucial importance of a reconstruction of the knowledge available to the agents themselves. Elucidation of this situation-relative knowledge gives the social scientist a clue to the appropriate behavior of the actors (Watkins 1970, Popper 1972). If we turn to the goals of the actors: should their preferences be confined to (narrowly conceived) self-interest, for example a combination of pleasures across time? This would be empirically unconvincing. It is then tempting to distinguish several types of action: self-interested action and action out of respect for norms or values, for example. But such a contrast is rather uncertain in many empirical situations. Correlatively, it appears difficult to disentangle the economic dimension from other aspects of human action (Demeulenaere 1996).

1.3 Rationality Assumption and Equilibrium

Should the social scientist consider any kind of beliefs, desires, and choices? It would seem awkward to leave unused our knowledge about rational action. Several criteria of rational choice can be used. One basic requirement is the existence of a complete and transitive ranking of social states of affairs which, combined with beliefs about uncertain events, yields a ranking of actions. More specialized criteria involve specific relationships between beliefs and evaluative judgements, as exemplified by the classical expected utility formula for choice under risk or uncertainty.

It can hardly be assumed that flesh-and-blood individuals are perfect optimizing entities in all circumstances: given the drastic limitations on our information-processing capabilities, the bounded rationality assumption is more realistic (Simon 1982). In many models, agents seem to be confined to single-exit situations: individual goals and the characteristics of the situation are the important explanatory inputs (Latsis 1972). In most real-world cases, however, individual attitudes in the gathering and use of incomplete knowledge are crucial explanatory factors. In Simon's words, a shift is needed from substantive to procedural rationality. This shift in emphasis has been decisive since the 1970s.

This psychological orientation could be (and has been) resisted for the following reason: sociological explanations of aggregate social facts are not supposed to capture all the idiosynchrasies of the actions of existing individuals, but only the central tendencies in individual action, among which we should undoubtedly count the tendency to act rationally through the selection of appropriate means in a purposeful manner (Goldthorpe 1998). It can thus be argued that sociology, as part of empirical science, should neither restrict itself to the elucidation of the logical consequences of the interaction of ideally rational agents nor try to replicate the psychological functionings of particular agents.

In practice, sociologists can hardly focus on social situations exclusively or psychological processes exclusively. Stylized models of both are required. In some situations, however, a more complex treatment of psychological processes is required. The two imperatives of psychological relevance and situational analysis are complementary in a deep sense, since the rational agent's adaptive cognitive strategy in the face of incomplete knowledge must depend on the characteristics of her incomplete-knowledge situation.

In most social contexts, the reasons for action of distinct individuals are not independent from one another, and the hypothesized relationships between individual preferences, beliefs, and actions assume the form of an equilibrium. This was brought out neatly in Cournot's analysis of imperfect economic competition. In the definition of equilibrium concepts, the interrelationships between the predicted social facts and individual reasoning are made clear. In some cases, however, the predictive value of equilibrium concepts (such as 'Nash equilibrium' in which each agent's action is the best response to the actions of others) appears to presuppose a fairly sophisticated coordination mechanism for individual expectations, or even sociologically or culturally determined common focal points. This suggests that some preexisting irreducibly 'social' facts might have independent explanatory value after all (Janssen 1998).

2. Methodological Individualism and the Contrasting Uses of Social Theory

Models and theories may have several uses. How does this plurality affect the significance of methodological individualism? The relationship between positive and normative uses of theories should be considered. Of equal importance is the level of description at which theory-based inferences and predictions are formulated.

2.1 The Individualistic Point of View and Normative Thinking

Scientific commitment to methodological individualism for explanatory purposes implies no particular

normative criterion. In particular, as Popper noted, an individualistic understanding of social life need not lead to underrate the importance or value of altruistic feelings or other-oriented interests generally speaking: 'What really matters are human individuals, but I do not take this to mean that it is I who matter very much' (Popper 1945).

Nevertheless, most normative criteria presuppose an individualistic description of social life. For example, the usual criteria of freedom and responsibility have no clear significance unless social life is understood with a view to the causal powers of persons, portrayed as autonomous units of evaluation and action. It thus appears that common normative views implicitly rely on an individualist explanation of social phenomena. It is also the case, at a more fundamental level, that some kind of individualistic description of social life is involved in the formulation of many evaluative criteria. Most equality or efficiency criteria can be applied only if we are able to ascertain that definite states of affairs are experienced by distinct individuals. Individualism, here, stands for an understanding of social life according to which the separate lives and experiences of the individuals (rather than social groups) are the *loci* at which things of value are to be found.

2.2 *Individualism and the Levels of Description of Complex Social Facts*

Several explanations might appear relevant and informative at different levels of description. The projected inferential and predictive uses of social theory would thus appear decisive for the relevance of individual-based descriptions and explanations of social life. As a matter of fact, the very notion of sociology as a separate discipline is sometimes understood in a holistic manner because sociology is supposed to describe and explain complex 'social facts' which cannot be described and explained adequately at the microlevel of individual action.

In Durkheim's classic study of suicide, for example, individual conduct is understood on the basis of the social characteristics which account for the scientifically relevant individual characteristics. Thus individual propensities to suicide can only be explained by referring to the state of society as a whole, in particular its degree of 'integration' or cohesiveness. 'Egoist suicide' is characterized by the attitude of individuals who give a particular importance to their own personality or destiny (their individual self) in the face of social rules or collective identity (the collective self), and the empirical frequency of such an attitude is explained by the degree of social integration.

The irreducibly holistic dimension in Durkheimian sociology is the belief in the independent explanatory value of 'macro' notions with respect to observed individual conduct. If we turn to rules and institutions, this means that individual agreements cannot be treated as independent explanatory factors. Individualistic theorists, however, try to explain rules and institutions by focusing on the circumstances of individual rational choice (as in Olson 1965, Coleman 1990).

Durkheimian explanation can be reinterpreted in individualistic or interactionist terms to some extent, by focusing on the psychological dimension of complex microsociological patterns of interdependencies that aggregate into macroproperties such as 'integration.' Given the parallelism between the two levels of description, the choice between an individualistic perspective and a holistic one should depend upon the projected use of theory: do we intend to describe macro phenomena, or do we try to understand micro interactions and their aggregate effects? The relevance of methodological individualism is clearer in the second case.

3. *Individual Cognition and Social Explanation*

Individuals are reasoning entities: they develop elaborate cognitive structures—even social theories—about their own action and interaction. How should individualistic social science take this into account? Arguably, any kind of rational action involves cognitive operations to some extent. This is true of instrumentally rational actions, after the pattern of 'logical actions' in Pareto's sense, since the adequate association of ends and means is only possible on the basis of some representation of the world, and some reasoning about it. Sometimes the goal and its relationship with the available means are obvious, so that the cognitive process will not figure prominently in the proposed explanations. In other cases, the interesting part of the explanation is the cognitive process itself.

The explicit treatment of this cognitive dimension appears to be the best response to the critique according to which individualistic models rely on an arbitrarily simplified account of human action as instrumentally rational action. It is useless to deny the importance of instrumental, means-end reasoning in social interaction. But much is to be gained by paying attention to the cognitive processes through which agents frame their choices (Fillieule 1996). Cognitive sociologists also explore the inner logic of argumentative strategies (Boudon 1990, Bouvier 1999). A challenging dimension of the theorist's work is to reconstruct the intrinsically convincing reasons which are intermingled with mistaken beliefs, in a manner which results in seemingly irrational behavior. Of particular interest are the processes through which individual agents evaluate and adjust their own goals, and select the norms of behavior they are willing to accept as binding. This could bring new significance to

the classical Weberian notion of axiological rationality, now understood as a special dimension of a general theory of cognitive rationality (Boudon 1995, 1998, Mesure 1998, Saint-Sernin et al. 1998).

This implies that rationality requirements are applied to the actor's goals, beyond the instrumental adequacy of the means to the goals. These are strong requirements, but they are motivated by an effort to gain a better fit between theoretical predictions and observed social facts. For instance, a long-standing puzzle in individualistic political studies is the so-called 'paradox of not voting': since individual voting in general elections has small disadvantages and no clear causal effect, why do people vote? A particular line of individualistic inquiry starts from the idea that most people are able to understand the basic justifications of certain norms (such as 'citizens should vote').

Explicit treatment of individual cognition is a valuable goal, but it raises a number of questions concerning the precise drawing of the dividing lines between the building blocks of individualistic sociological explanation: between individual characteristics and natural facts on one hand, and between individual characteristics and social relations on the other hand.

At some level of description, individual action is a series of natural (neural and external) states of affairs. It thus appears that individual 'good reasons' are screened off by the natural causes of behavior, which are a proper object of inquiry for natural science, and methodological individualism in its classical versions might appear misleading because it treats individual good reasons as ultimate explanations of action.

Psychological processes are implemented in the human brain by neurons, but this does not imply that the 'ultimate' explanation of social phenomena is to be found at the neurophysiological level, or at the molecular level. Science is constructed in layers, in such a manner that it includes both higher level theories which work at the macrolevel appropriate to the macrophenomena to be explained (such as sociological theories) and other theories which show how the higher level objects and relations can be accounted for at the level below (Simon 1997). The hierarchical relationship between two classes of phenomena does not automatically imply that explanation should proceed from the lower level to the higher level.

At a certain level of description, individual action is the sequel of an intention which can only be depicted in a certain language, using definite mental categories. Arguably, language, categories, and meanings are social objects. Hence it might appear appropriate to consider human actions as understandable only on the basis of pre-existing rules, or 'institutions' in Mauss's sense—acts or ideas which people find before them as pre-existing facts. This radical holism draws our attention to the constitutive role of social norms with respect to the meaning of actions—undoubtedly an important feature of many social situations. This feature can find a place in individualistic explanations, once we allow for psychological motives which depend on complex, holistic representations of social interaction, along the lines of institutional individualism (Agassi 1975). As a matter of fact, most social rules are not just given as 'facts': they are agreed upon, or rejected, by individuals.

Individualistic explanatory and predictive strategies rely on a sharp distinction between the choice problems of individual actors, on the one hand, and social institutions, regularities, and norms on the other hand. On this account, the emergence and stability of social regularities, social norms and institutions should be explained on the basis of the underlying individual reasons. This creates an opportunity for the application of both classical rational-choice models and elaborate theories of individual cognition.

See also: Action Theory: Psychological; Action, Collective; Action, Theories of Social; Bounded Rationality; Collective Behavior, Sociology of; Collective Beliefs: Sociological Explanation; Durkheim, Emile (1858–1917); Methodological Individualism: Philosophical Aspects; Pareto, Vilfredo (1848–1923); Rational Choice Theory in Sociology; Rationality in Society; Simmel, Georg (1858–1918); Sociology, Epistemology of; Sociology: Overview; Traditions in Sociology; Verstehen und Erklären, Philosophy of; Weber, Max (1864–1920)

Bibliography

Agassi J 1975 Institutional individualism. *British Journal of Sociology* **26**: 144–55

Boudon R 1990 *L'art de se persuader*. Fayard, Paris

Boudon R 1995 *Le juste et le vrai*. Fayard, Paris

Boudon R 1998 Social mechanisms without black boxes. In: Hedström P, Swedberg R (eds.) *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge University Press, Cambridge, UK

Bouvier A 1999 *Philosophie des Sciences Sociales*. Presses Universitaires de France, Paris

Boyer A 1992 *L'explication en Histoire*. Presses Universitaires de Lille, France

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Demeulenaere P 1996 *Homo oeconomicus*. Presses Universitaires de France, Paris

Fillieule R 1996 Frames, inferences, and rationality: Some light on the controversies about rationality. *Rationality and Society* **8(2)**: 151–65

Goldthorpe J H 1998 Rational action theory for sociology. *British Journal of Sociology* **49**(2): 167–92

Janssen M 1998 Individualism and equilibrium coordination in games. In: Backhouse R E, Hausman D, Mäki U (eds.) *Economics and Methodology*: *Crossing Boundaries*. St. Martin's Press, New York

Latsis S J 1972 Situational determinism in economics. *British Journal of Philosophy* **23**: 207–45

Mesure S (ed.) 1998 *La nationalité des valeurs*. Presses Universitaires de France, Paris
Olson Jr M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
Popper K R 1945 *The Open Society and its Enemies*, 5th edn. rev. 1966. Routledge and Kegan Paul, London
Popper K R 1972 *Objective Knowledge*. Clarendon Press, Oxford, UK
Saint-Sernin B, Picavet E, Fillieule R, Demeulenaere P (eds.) 1998 *Les modèles de l'action*. Presses Universitaires de France, Paris
Simmel G 1900 *Philosophie des geldes*. Duncker & Humblot, Leipzig, Germany
Simon H A 1982 *Models of Bounded Rationality*. MIT Press, Cambridge, MA
Simon H A 1997 *An Empirically Based Microeconomics*. Cambridge University Press, Cambridge, UK
Watkins J W N 1970 Imperfect rationality. In: Borger R, Cioffi F (eds.) *Explanation in the Behavioral Sciences*. Cambridge University Press, Cambridge, UK
Weber M 1922 *Wirtschaft und Gesellschaft*. Mohr, Tübingen, Germany

E. Picavet

Methodological Individualism: Philosophical Aspects

Methodological individualism (MI) can be stated roughly as the thesis that explanations of social phenomena should appeal only to facts about individual people. This is a directive or rule, rather than a statement of purported fact—hence methodological individualism. But whether a rule is useful depends, in part, on the facts; the fact of gravity makes the rule 'exit by the window' poor advice. We need to ask whether society, individuals, and their relations are such as to make MI a sensible rule to follow, where explanation is our goal. Also, MI is a quite general rule. Few would deny that an individualistic explanation is sometimes in order. The question is whether it always is.

1. Historical Background

The expression 'methodological individualism' seems first to have been used in 1908 by Schumpeter, but the doctrine was expounded by Menger in the 1880s; it was a cornerstone of what came to be known as 'Austrian method' which stood in opposition to the German historical school (Hayek 1968, Nozick 1977). Emerging from within economic thought, MI has always had its strongest adherents—as well, it is claimed, as its greatest successes—in that domain. Success for holistic styles of explanation—explanation by reference to social 'wholes', their structure and their function—has been claimed most often in anthropology and sociology, the latter much influenced by Durkheim's anti-individualism. Weber declared himself an individualist, but did not always follow its rule, and the best example of individualistic sociology in recent times may be the work of Homans (1967). History has been very much disputed territory, with both practicing historians and theorists of history divided as to whether MI is a rule they can or should adopt. Examples of 'pure' methodological individualism in the practice of any social science are hard to find, and one theoretical issue is whether this is symptomatic of some deeper incoherence in the position.

Karl Popper's writings, and those of his student John Watkins, sparked a debate between individualists and collectivists or holists that lasted through the 1960s (O'Neill 1973). But a second wave of interest in this dispute emerged in the 1980s as a number of authors turned to the issue with claims informed by recent developments in philosophy of mind, language, and biology, as well as by some rethinking of Marxist views. By and large, the terms of debate in the first wave were set by individualists who saw themselves as opposed to political movements of the extreme left and right, and opposition to MI was portrayed as depending on a dangerous belief in inexorable laws of history and society. These considerations largely have been put aside in the second wave, and a number of broadly holistic positions have emerged which are not committed to the existence of laws of history, and which display no affiliation to the Hegelean tradition.

In the view of some, the debate over MI can be settled only by assessing how successful individualistic social science is by comparison with its holistic rival. On that reading, the question is an empirical one. This article is written from a somewhat different perspective, which aims at getting clear about the basic positions and argumentative moves of both sides. Without such clarity it will be easy to misinterpret the empirical results.

2. Collective Intention

This section begins by considering an objection that tends to be raised when individualists insist that it is the beliefs, motivations, and decisions of individuals that matter, though it will be put to one side quickly. The objection is that there is a sense in which groups of individuals can, say, form intentions to do things. Some accounts of group intentions (or what are sometimes called 'we-intentions') need present no problem for the individualist. On one account a group intention to do something consists of all the members of the group intending to play their part in doing it, and believing that other members of the group similarly intend; here there is no group intention over and above the intentions and beliefs of the members of the group (see, e.g., Toumela 1989). Searle has argued that such accounts are incorrect, and that we need to

acknowledge the irreducibility of we-intentions to intentions with a singular subject. But Searle also argues that this is consistent with MI: when you and I intend that we play tennis, what happens is that I have an intention that we should play, and you have an intention with the same content; there is no supra-personal 'we' having an intention (Searle 1995, pp. 24–6). However, Velleman has argued that a collective may literally have an intention and that this is not a case simply of a conjunction of intentions, whatever their form, possessed by individuals (Velleman 1997). Velleman's shared intentions are not supposed to be lodged in a suprapersonal mind, because they are not supposed to be mental entities at all. Rather, they are linguistic. Still, on Velleman's account, it can happen that the unit of intending is the collective and not the individual. So an explanation of events couched in terms of intentions need not be an explanation in terms of the mental states of individuals, and a restriction of our attention to the intentions of individuals might obscure a relevant explanatory factor. But at the very least the theory of linguistic intentions requires elaboration before it can be considered to pose a serious threat to MI. (See also the work of Gilbert 1989 who treats social groups as the subject of 'we.')

3. *Reduction*

In *The Open Society and its Enemies* Popper describes MI as the view that collectives must be reduced to individuals and their actions (Popper 1966, p. 91). What does reduction involve? An influential model for this sort of approach is microphysical reduction in the sciences; we reduce water to H_2O by showing that the properties of a body of water are just the properties of a collection of suitably related H_2O molecules (relations are important because the molecules have to have the right kinds of bonds, for example). A comparable reduction of social entities to individuals aims to show that social entities like collectives are nothing over and above collections of suitably related individuals.

A number of arguments have been offered for thinking that this condition is not fulfilled. Reductive MI is rejected by some because it is supposed to depend on an implausible semantic thesis: that we can define all our social terms—like 'class,' 'class conflict,' 'nation'—in individualistic terms (see, e.g., Mandelbaum 1955, reprinted in O'Neill 1973). Whether that semantic thesis can be defended is a difficult issue, but we need not resolve it here. One can believe that social entities like nations and classes are nothing over and above collections of suitably related individuals without believing that the concept of a nation or of a class can be defined in terms of individualistic concepts. Reductionism is a claim about what exists, and not a claim about definability. Similar claims about reduction without definition have been fruitfully pursued in the philosophy of mind (Davidson 1980) and in the philosophy of time (Mellor 1998).

But could a social entity, France, say, be reduced to individuals? Ruben, a contributor to the second wave, argues that France cannot be so reduced, because no class or group of individuals can have exactly the properties France has. For example, France has the property that it could have had a different population, but the class of French people (past, present, and future) could not have had any person in it other than the persons it does have in it, since classes are defined by their members. And if two entities do not share all their properties, they cannot be identical. On Ruben's view, France is a holistic entity, not even constructed out of individuals, which bear to it only contingent relations (Ruben 1985, chap. 1). It should be said that Ruben is otherwise quite sympathetic to MI.

This argument is open to dispute if we think of terms like 'France' as referring to whatever thing occupies a certain role, just as 'The Prime Minister' refers to the person, whoever it is, who occupies a certain party-political role. There is no doubt that in 1999 John Howard was the PM of Australia and that counting John Howard and the PM would be double counting, despite the fact that the PM could have been someone other than Howard, while Howard could not have been someone other than himself. And when we speculate on what would have happened had the opposition leader, Kim Beazley, been PM, we are imagining a situation in which Beazley occupies that very same role, or something very like it. The qualification 'or something very like it' is necessary because the PM's role might be altered to some degree by, say, constitutional amendment. But there are limits to what is possible for the role by way of alteration; some alterations would be so substantial that we would regard them as resulting in the replacement of that role with something else—a role of President, for example.

What role would something occupy in order to be France? France is a geo-political entity founded at a certain time in a certain way, with a subsequent history of relations to other geo-political entities like The Holy Roman Empire, Germany, and the EEC. Call that role F. If we are going to say that, by definition, France is the thing that occupies precisely the role F then it will be impossible to say that the history of France might have been different. On the other hand it makes dubious sense to say that France might, for example, have had exactly the history of Italy (though of course Italy might have been called 'France'). The solution, as with the role of Prime Minister, is to say that France is, by definition, the thing that occupies a role sufficiently like role F; that leaves the matter vague, but since questions about what might have happened in history really are vague, this is no criticism of the present approach.

Now it may be that the thing that occupies the France-role at any given time is a collection of individuals standing in certain relations—along with other things like certain purely physical resources, perhaps. And different collections of individuals will fill this role at different times, just as different individuals are Prime Minister at different times. Of course, the France-role and the Prime Minister-role are different in that we think of the France-role as occupied by a single entity over the whole period that the role is filled at all, while we think of the Prime Minister role as occupied by many different things—there is one France but there are many Prime Ministers. But treating France as one continuous entity is not ruled out on the present approach; a single living organism is constituted by the cells that make it up, but no cell need last the whole life of the organism. We may think of France as similarly constituted by an overlapping sequence of individual people, suitably related.

However, the idea of the individuals that constitute a social whole being 'suitably related' raises another objection to the individualistic reading of social entities. The objection is that the relations in which individuals have to stand in order for collections of them to fill macrosocial roles are in fact social relations. So all we have is a reduction of the social to the social. Thus, it is arguable that a collection of persons constitutes a nation only if there are relations of subservience, or equality, or other obviously social relations, between those persons.

The first thing to be said about this argument is that, even if it is correct, it allows us a reduction of the macrosocial to the microsocial. It would allow us to say that, while there are limits to how far individualism can be pressed, we have shown at least that the only social relations constitutive of collectives are relations between individuals. That would be progress from the individualist's point of view. The second thing to be said is that the argument may not be correct. We might counter the argument by making the same move with respect to the social relations between individuals that we made in the case of supposed social wholes. We said that deciding whether there is such a thing as France at a given time is a matter of deciding whether something occupies the France-role at that time. Similarly, deciding whether two individuals stand in a social relation of equality at a time may be a matter of deciding whether there is something that occupies what we might call 'the equal-citizenship role.' It might turn out that the thing that occupies the equal-citizenship role is, say, a presumably very complex psychological relation. How might we spell out this role? The following is on the right lines, though greatly simplified: the equal-citizenship relation is that relation, whatever it is, which in normal circumstances ensures that the law will be applied in the same way in similar cases and regardless of class, race, sex, and various other factors. And we may then find that the thing which does ensure this is a complex psychological relation between individuals, involving beliefs, desires, values, and behavioral dispositions.

These seem the most promising sorts of moves open to the individualist, though whether they will ultimately succeed is something we cannot settle here. Rather, it will be assumed simply from now on that it is successful—thereby conceding apparently vital ground to the individualist—and ask whether we would then be justified in agreeing with MI that explanations of social phenomena ought to appeal only to facts about individuals.

4. *Reduction and Explanation*

Before we try to answer this question it is worth making some points about the relations between the claim that the social is reducible to the individual and the claim that explanations of the social should appeal only to facts about individuals. First, the claim that social entities are reducible to individuals is a factual claim and not a methodological rule. But recall our decision to see whether MI is an appropriate rule. It would certainly seem that if reducibility is true, the conditions for the application of MI could scarcely be more favorable. We shall see that even if reducibility is true, the victory of MI is not assured. Second, MI might be formulated in a way that does not depend on a claim about reduction. Popper has said that '… we must try to understand all collective phenomena as due to the actions, interactions, aims, hopes, and thoughts of individual men (Popper 1960, p.157), which suggests the view that while social entities are real and (perhaps) not reducible to individuals, their existence, persistence, character, and alteration ought to be accounted for in terms of individuals alone. But it is worth remarking that even those who deny reducibility of the kind described above would probably assent to a weaker thesis of the dependence of the social on the individual. A way to make this view more precise is by appeal to the idea of supervenience. Supervenience theses tell us what features of the world need to be independently specified, and what features we get for free, so to speak, by specifying other features. And it is certainly a plausible thought that by specifying all the facts about individual thought and action, we thereby fix the facts about the social. Imagine a world that is absolutely indistinguishable from our world from the point of individual thought and action; would there be any social differences between the worlds? Assuming, rather plausibly, that the answer is no, we can conclude that the social supervenes on the individual (Currie 1984). But notice that this, sometimes called global supervenience, is a very weak claim, and one that tells us little about the prospects for individualistic explanation of social phenomena. In particular, it does not give us any right to think that we shall be able to demarcate some manageable subclass of the individual facts in terms

of which to explain any given social phenomenon. For all supervenience tells us, the most we could say is that the totality of individual facts explains the totality of social phenomena. A workable theory of individualistic explanation requires us to find local connections between the individual and social.

Third, an issue of the kind that delights philosophers arises when we look more closely at the claim that facts about social wholes like countries and classes can be reduced to facts about individuals and individual psychology. Perhaps the most complex group of problems that philosophy of mind will take into the new millennium is that raised by the doctrine of 'externalism.' At least since Descartes there has been a tendency in philosophy to assume that nothing follows about the world from what goes on in the mind; one can have an inner life of meaningful thoughts and still not be sure that anything external corresponds to those thoughts—hence Descartes' worry about skepticism. But this mind–world dualism (not at all the same as mind–body dualism) has been under attack, and many philosophers now hold that the contents of some thoughts depend essentially on the existence of the things they refer to. This is externalism (see Putnam 1975 and the essays in Pettit and McDowell 1986). If externalism is true, then, so the argument goes, supposed reductions or explanations of the social in terms of the individual may turn out to be reductions or explanations of the social in terms of, among other things, the social. While the true implications of externalism are notoriously difficult to see, a preliminary response to this argument would be to find some way of ensuring that the mental states that we appeal to in explaining social phenomena are never ones with the kinds of contents that raise these externalist problems. While this and other potential solutions to the difficulty take us beyond the limits of the present survey, we do have here a striking illustration of the interconnection of social philosophy and the philosophy of mind.

Fourth, a claim about the reducibility of social entities should not be confused with a claim to the effect that social entities do not exist. The reduction of water to H_2O is not grounds for doubting that water exists; in fact successful reductions (as opposed to what philosophers sometimes call 'eliminations') show that the reduced entities do exist. But this creates a problem for an individualist. If social entities are reducible, then why should an individualist not refer to them in her explanations? And what, in that case, would be left of the idea of a distinctively individualistic explanation? Perhaps the individualist should be concerned, not about whether, but about how we refer to social entities. Their objection would then be to references to social entities which fail to make manifest their reducibility; we should refer to France, not as, say, 'France' but as 'such and such a group of individuals bearing such and such relations to one another.' This is an issue we shall come back to, but on the face of it the claim is implausible. Who would insist, comparably, that we never satisfactorily explain an accident by saying that there was water on the street without also making it manifest how the body of water in question was made up of H_2O molecules? Perhaps the individualist thinks that we should at least not refer to social entities unless we are confident that we could give an individualistic reduction of them. But who would claim that all explanations that appealed to water were illegitimate prior to the chemical revolution in the late eighteenth century?

What is left to the individualist? They may say that while all sorts of nonindividualistic explanations may be given and accepted, because they are the best we can get in the circumstances, epistemic considerations always favor the move to an individualistic explanation where one is available. For such a move will always be from an explanation which refers to social entities in a way which does not make their individualistic constitution manifest, to an explanation which refers to social entities in a way which does. And such a move will always represent an increase in relevant explanatory information. This is a plausible and influential thought that deserves some attention.

5. *Explanation and Information*

One person who holds this view is Elster (see Elster 1985). Elster advocates MI from a Marxist position—an indication of the shift in the terms of the debate that has taken place between the first and second waves, though some other writers influenced by Marx continue to oppose MI (see Cohen 1978, Wright et al. 1992). Elster argues that we all have reason to prefer individualistic explanations to explanations in terms of social aggregates. First of all, microlevel explanations reduce the time-lag between cause and (putative) effect, making us less vulnerable to the confusion of causation with mere correlation. Second, understanding is increased when we 'open up the black box' and look at the 'desires and beliefs that generate the aggregate outcomes' (Elster 1985, p. 5). This second remark is somewhat misleading in that Elster does not argue that we ought to construct social explanations by examining the actual beliefs and desires of the real people involved in the phenomenon to be explained. Rather, his commitment to MI is a commitment to 'rational actor explanations' wherein we explain what happened as a rational response on the part of individuals to their situation, without looking at the actual details of their mental states (without, in other words, opening the black box at all).

But individualists who do not follow Elster's enthusiasm for rational actor explanations and who think that our business is with the empirical study of actual mental states will still agree with him in this:

That the great strength of individualistic explanation is that they give us more explanatorily relevant information than holistic explanations do. Thus, Watkins says that 'There may be unfinished or half-way explanations of large-scale social phenomena …; but we shall not have arrived at rock-bottom explanations … until we have statements about the dispositions, beliefs, resources, and interrelations of individuals' (Watkins 1957, p. 168). Individualists may agree that in many situations we have to make do with aggregate-level explanations because the relevant individualistic detail is not available to us. But they will say that it is one thing to tolerate aggregate-level explanations when nothing better is available, and quite another to prefer them.

6. *Levels of Explanation*

An obvious response to this line of thought is to point out that it can be extended in unwanted ways. Note that the individualist's case, outlined in Sect. 3, for a reduction of the social to the individual has a parallel in an argument for the reduction of the individual to microphysics. Those who adopt a functionalist view of the mind think that the idea of a mental state is the idea of something that fits a certain causal role, and that the things that do fit those mental states are purely physical states of the brain. On this view, pain is (roughly) whatever is caused by damage to the body and which in its turn causes us to avoid the cause of the damage. Most scientists and philosophers think that what has that causal role is some physical state of the brain. So pain will turn out to be that brain state. So why, if we reject unanalyzed references to social wholes, ought we to be content with similarly unanalyzed references to persons, minds, and mental states? Why not insist that these be spelt out in terms of the physical states that occupy the relevant roles? Popper and Watkins would deny the claim that mental state-roles are occupied by purely physical states because they believe that there is a nonphysical mind which makes an independent contribution to causation (see, e.g., Popper et al. 1984), so they would be able to claim that descent below the level of the individual and her mental states to the level of physical causation is a transition that involves loss of explanatory information. But belief in this kind of dualism is not widely shared. Another response would be to say that our methodological obligation is not always to descend to the most basic level, but to the most basic level within the relevant domain. If our interest is in the domain of the social sciences, then the most basic level is that of the individual; anything below that ceases to be social science. But suppose I declare my interest to be in holistic social science; what is an individualist to say when I refuse to countenance explanations that go below the level of the most basic social wholes, whatever they are?

It is time to question the assumption, shared by individualists like Elster and Watkins, that descending to lower levels always makes for increases in explanatory information. Jackson and Pettit have argued that higher level explanations may in fact convey information, including causally relevant information, not even implicitly conveyed by lower level ones (Jackson and Pettit 1992a, 1992b). Jackson and Pettit make a distinction between causally efficacious and causally relevant facts, the latter being a more inclusive class than the former. They contend that it is sometimes the case that, by appeal to a higher order explanation, we can locate causally relevant facts that we could not locate by appeal to an explanation of lower order which tells us only about what is causally efficacious. Consider an explanation of the decline of religious practice in terms of increased urbanization, an explanation that accounts for changes in one social institution by appeal to changes in another. What would we lose by replacing this explanation with a much more detailed one in terms of the changing values, employment, and location of individuals, the decisions of individuals to close certain parishes, the inexperience of ministers in new parishes, etc.? By deleting urbanization from our explanation we lose sight of this, that the effect brought about by these individual decisions, namely reduced religious observance, would have been brought about by any other combination of individual decisions as long as it was a combination that manifested itself, at a higher level, as increased urbanization. We would go wrong to suppose that what was crucial to the decline of religion was that Smith moved from village A to town B, that he went from working on this farm to working in that factory, along with all the other specific, causally efficacious facts about Smith and about other individual people. What was crucial was that something took place at the level of individuals that constituted urbanization, and there are many combinations of individual events other than the combination that actually occurred that would have done that. So the shift from the holistic to the individualistic explanation involves loss, as well as gain, of causally relevant information. The gain is in the specificity of detail: we now know which combination of 'microfacts' determined the macrofacts about the decline of religion. The loss is in breadth: the macro-explanation identified for us a class of combinations of microfacts, any one of which would equally have explained the decline of religion—namely the class of combinations which determine that urbanization occurred.

7. *Conclusion*

The conclusion, then, is that even on epistemic grounds alone, and ignoring pragmatic factors, an individualistic explanation is not automatically to be preferred to a nonindividualistic one. But it does

not follow that appeals to the beliefs and motives of individuals—real or 'typical'—can simply be jettisoned from the project of social explanation. In one vital respect, the relation between the individual and the social is not like that between the microphysical and the social. While in both cases we have a supervenience relation, in the former case but not in the latter we have something more. Facts about individual belief and motivation contribute to the intelligibility of social phenomena, while facts about the microstructure of someone's brain do not contribute to the intelligibility of her beliefs and motives (Rudder Baker 1995). Consequently, whatever higher-level social facts we may invoke to explain a revolution, the decay of rural life, or the victory of republicanism, we cannot say in advance that we shall simply ignore individuals. Further, as we have acknowledged, the social exists at the level of the individual; facts about individuals are often themselves social facts. If this is right, the appropriate judgment on MI would seem to be this: that we cannot accept the claim that individualistic explanations are always to be preferred to explanations which invoke social entities without displaying how those entities can be reduced to individuals, that indeed such unreduced references to social entities sometimes carry information valuable for explanatory purposes. But we owe a debt, nonetheless, to those who have elaborated modes of individualistic explanation, since social inquiry cannot do without such explanations. The remaining methodological questions are about how and in what circumstances individualistic and holistic factors can best be combined, and about whether there are any styles of holistic explanation that we can say on philosophical or empirical grounds are always to be avoided.

See also: Action, Collective; Atomism and Holism: Philosophical Aspects; Causation: Physical, Mental, and Social; Evolutionary Selection, Levels of: Group versus Individual; Explanation: Conceptions in the Social Sciences; Functional Explanation: Philosophical Aspects; Individual/Society: History of the Concept; Individualism versus Collectivism: Philosophical Aspects; Popper, Karl Raimund (1902–94); Reduction, Varieties of

Bibliography

Cohen G A 1978 *Marx's Theory of History: A Defence*. Clarendon Press, Oxford, UK

Currie G 1984 Individualism and global supervenience. *British Journal for the Philosophy of Science* **35**: 345–58

Davidson D 1980 *Essays on Actions and Events*. Clarendon Press, Oxford, UK

Elster J 1985 *Making Sense of Marx*. Cambridge University Press, Cambridge, UK

Gilbert M 1989 *On Social Facts*. Routledge, London

Hayek F A von 1968 Economic thought: The Austrian school. In: Sills D (ed.) *Encyclopedia of the Social Sciences*. Macmillan, New York, p. 4

Homans G C 1967 *The Nature of Social Science*, 1st edn. Harcourt, Brace & World, New York

Jackson F, Pettit P 1992a In defense of explanatory ecumenism. *Economics and Philosophy* **8**: 1–21

Jackson F, Pettit P 1992b Structural explanation and social theory. In: Charles D, Lennon K (eds.) *Reductionism and Anti-reductionism*. Oxford University Press, Oxford, UK

Mandelbaum M 1955 Societal facts. *British Journal of Sociology* **6**: 305–17

Mellor D H 1998 *Real Time II*. Routledge, London

Nozick R 1977 Austrian methodology. *Synthese* **36**: 353–92

O'Neill J (ed.) 1973 *Modes of Individualism and Collectivism*. Heinemann Educational, London

Pettit P, McDowell J (eds.) 1986 *Subject, Thought and Context*. Clarendon Press, Oxford, UK

Popper K R 1960 *The Poverty of Historicism*. Routledge & Kegan Paul, London

Popper K R 1966 *The Open Society and its Enemies. Volume II: The High Tide of Prophecy: Hegel, Marx, and the Aftermath*. Routledge & Kegan Paul, London

Popper K R, Eccles J 1984 *The Self and its Brain*. Routledge & Kegan Paul, Spring Intern, New York

Putnam H 1975 The meaning of 'meaning'. In: Putnam H (ed.) *Mind, Language and Reality. Philosophical Papers, Vol. 2*. Cambridge University Press, Cambridge, UK

Ruben D-H 1985 *The Metaphysics of the Social World*. Routledge & Kegan Paul, London

Rudder Baker L 1995 *Explaining Attitudes*. Cambridge University Press, Cambridge, UK

Searle J R 1995 *The Construction of Social Reality*. Allen Lane, London

Toumela R 1989 Actions by collectives. *Philosophical Perspectives* **3**: 471–96

Velleman J D 1997 How to share an intention. *Philosophy and Phenomenological Research* **57**: 29–50

Watkins J 1957 Historical explanation in the social sciences. *British Journal for the Philosophy of Science* **8**: 104–17

Wright E O, Levine A et al. 1992 *Reconstructing Marxism: Essays on Explanation and the Theory of History*. Verso, London.

G. Currie

Metropolitan Growth and Change: International Perspectives

The twentieth century was widely considered to be 'the urban century.' This was the period when most countries, even those with less advanced economies, were transformed from rural to urban-based societies, and then to a situation of metropolitan dominance. At the beginning of the new century, however, it is unclear what the next stage in the urbanization process will be. Is it to be a post-metropolis era characterized by

dispersed internet-linked settlements, or the beginning of a new metropolitan era, but one with a differing organization dominated by a small number of very large global metropolises or 'world' cities? Debates on these trends, however, continue to suffer from three problems: the increasing complexity and diversity of urban processes around the world (United Nations Centre for Human Settlements [Habitat] 1996), the lack of data on the linkages between urban areas, and the continued ambiguity of concepts and measurement criteria.

One concept that embraces this complexity and offers a framework in which to describe and understand both historical and contemporary trends in the urbanization process, at different spatial scales, is the concept of the 'urban system' or system of cities. This essay introduces the concept and then uses it to outline recent shifts in metropolitan growth.

1. *The Concept of an Urban System*

Over the last few decades, it has become widely accepted that individual cities and metropolitan regions do not exist in isolation. Cities, in fact, only exist because of their hinterlands and their interconnectedness to other places. They are, simultaneously, part of several larger sets or networks of urban areas—within a region, a nation and internationally—that interact closely with each other. In other words, they function as nodes in an integrated series of urban systems. Within those systems, individual places assume particular attributes, trade certain goods, and play particular functional roles, because of their relative position in one or other of those systems, and change as the structures of those systems evolve over time.

A system of metropolitan areas is a subset of the urban system that focuses on those larger areas that dominate the national economies and societies and their territorial organization. They are large enough to meet the standard measurement criteria employed by statistical agencies (e.g., in the census of any given country) for metropolitan status. In the US Census, for example, metropolitan statistical areas (MSAs) have a minimum population size of 100,000, with at least 50,000 in their core area municipality. The resulting 280 MSAs, in total, house over 77 percent of the US population and over 50 percent of that population now lives in those MSAs with over one million population. In Europe and Asia similar designations of 'functional urban regions,' also intended to replace out-dated political definitions of what is urban, are now widely employed, and they exhibit similar levels of metropolitan dominance.

The logic underlying the idea of studying the urbanization process, both past and present, through a system of urban or metropolitan regions is relatively straight-forward. The first place to look in examining the growth experience of any metropolitan area is to look outside, that is at its relative situation and its external connections or 'external relations.' Where does that area fit within the larger systems of which it is a part? The economic fortunes of urban areas, in this framework, are determined by the sum of their networks of linkages, flows and interactions with other places. These linkages, in turn, define and then reenforce the emergence of a hierarchy of urban places, based on systematic differences in size and the range of functions performed, as well as the relative level of diversification of the economy of each urban area, and the mix of populations, occupations and facilities within those places. This interconnectedness, of course, is not new. Historically, the interdependence of urban places has always been an essential feature of the urbanization process. Think, for example, of Carthage as a node in the far-flung colonial urban system of the Roman era, or the extensive networks of relations maintained by cities in the Hanseatic League. What is different now is that the interconnections are more intense, varied and universal.

2. *A Brief History of the Idea*

The concept of the urban system has been around for some time, but in its current form dates from the 1960s (see Bourne 1974, Berry 1976, Pred 1977, Geyer and Kontuly 1996). It developed primarily out of the need to find a conceptual framework that would overcome the limitations of existing concepts of the urban process that focused narrowly on studies of a few cities or attributes, or one theory. Typically, these included detailed descriptions and classifications of the characteristics of urban areas, but provided little sense of the full range of linkages underlying those attributes. Two principal bodies of urban theories were on offer at the time. The first addressed the location of manufacturing cities, and the structure of their networks of suppliers and markets. The spatial organization and hierarchical properties of the entire urban system, therefore, depended on the specific requirements of the production process alone.

The second dominant paradigm was based on classical central place theory. This theory purported to explain, simultaneously, the size, functions and spatial distribution of urban settlements that had developed primarily to provide different bundles of goods and services to a local, and largely rural, population living in their immediate hinterlands. The assumption that there is a distinct hierarchy of goods and services, varying from low- to high-order, each requiring a different threshold size market area to survive, resulted in a settlement system that was also explicitly hierarchical and geographically regular.

Although both theories provided conceptually elegant and analytically rigorous frameworks, each

suffered as the economy, settlement structure and the nature of trade and other interdependencies, changed. For example, manufacturing has represented a declining proportion of national output and employment in most western countries, and the location of individual manufacturing firms has been increasingly driven by factors other than minimizing the distance between sources of raw materials, suppliers and markets. At the same time, and despite numerous adaptations, classical central place theory has also become increasingly inadequate as an explanation for urban location and growth in a period of rapid urbanization that is overwhelmingly dominated by large metropolitan areas. These areas provide a wide range of diverse functions for both local and distant populations.

The concept of an urban system then offered an approach to, and a way of thinking about the study of urbanization at a macro-geographical scale, rather than a new technique or theory *per se*. It became common currency in the field precisely because it was sufficiently robust to accommodate very different levels and patterns of urbanization, and because it overcame the weaknesses and limited utility of models based on only one sector of activity (e.g., manufacturing), or one location factor (e.g., the minimization of transportation costs), or one form of interaction (e.g., the movement of material goods). The central role in the urban systems concept assigned to interactions of all forms also situates the concept in a strategic position to capture the growing importance for urban development of new communications technologies and the information age more generally.

Applying the urban/metropolitan system concept does not imply that we should ignore rural areas or small towns or marginalize their role. Instead, the concept argues that in contemporary and highly-urbanized societies, in which (on average) less than 25 percent of populations live outside of metropolitan areas and only 3 percent on farms, the entire economy is effectively 'organized' by and through its urban nodes and increasingly by its largest metropolitan regions. In such societies, all parts of national territories can then be considered as 'urban,' as integral components of metropolitan regions and the zones of influence surrounding those regions. Rather than being external to the metropolitan economy and social system, rural areas become fundamental building blocks of the regional urban systems associated with each metropolitan area. Even in developing countries with much lower levels of urbanization, the larger metropolitan areas serve as the gateways linking rural areas to the world economy.

In this context, individual metropolitan areas then grow by competing with each other, in order to expand their zones of influence or hinterland. This is the spatial equivalent of attempts to increase (or decrease) market share. The outcome of this competition is a urban system that is spatially differentiated, as well as temporally dynamic. Chicago, for example, garnered tributary territory for itself in the mid-west during the late nineteenth century that was formerly dependent on St. Louis for services and supplies. Denver, Dallas and Minneapolis, in the post-war period, expanded their hinterlands westward, truncating former service areas of Chicago, while Minneapolis' service area has been reduced in the north-west by the growth of Portland and Seattle. Similarly, in Canada, Toronto has assumed national dominance, in both culture and economy, displacing Montreal from its historical position at the top of the Canadian urban hierarchy, while Vancouver, Calgary and Edmonton have replaced Winnipeg as the gateways to and service capitals for the west.

Perhaps the best recent examples of the increasing importance of inter-metropolitan competition, and the evolution from national to trans-national urban systems, are in western Europe. The creation of the single integrated European market—the European Union—has not only accelerated economic integration across national borders, but is leading to a substantial reorganization of long-established systems of metropolitan centers and hinterlands within each country. All major centers are now competing, intensely and directly, with each other for a larger share of the entire community market, not just for those markets formerly protected within national boundaries and trade barriers.

3. *Postwar Phases in the Evolution of Metropolitan Systems*

Most scholars, when faced with the daunting task of summarizing the evolution of urbanization, not to mention the complex shifts of functions and influence among metropolitan areas, over the postwar period, tend to rely on constructing periodizations of that history. My own approach is to classify the last half of the twentieth century, admittedly in an overly simplistic fashion, into four periods or phases. Each is characterized by different rates, determinants and geographies of urbanization and by varying trajectories of metropolitan growth. Phase one, encompassing the 1950s and 1960s, was a period of rapid economic and population growth and with it increasing urban and metropolitan dominance. Almost every economic sector, region and urban area grew, floating as it were on the baby-boom population, the expansion of both manufacturing and services, the growth of the public sector, and increasing levels of prosperity and domestic consumption. Some places, of course, grew faster than others, and levels of income 'polarization' between regions increased. Most rural areas, and many small towns, in contrast, showed little or no growth; some went into sharp decline. Because of the continuing benefits associated with proximity (or spatial agglomeration), industrial production,

wealth and power became even more concentrated in the larger metropolitan areas.

The second phase, beginning in the 1970s, witnessed a marked reversal, or 'turn-around,' in the macro-geography of urban growth. Manufacturing activity declined in relative terms and the location of new economic activity shifted away from the older and large industrial metropolises as the benefits of agglomeration declined and the costs rose. In the US, this sectoral shift resulted in a parallel locational shift, as population and employment growth migrated from the old Northeast manufacturing belt (the so-called 'rust' belt) to newer cities in the south and west (the 'sun' belt). This period is frequently labeled as one of deconcentration and decentralization within national urban systems. Terms such as 'de-urbanization,' 'counter-urbanization' (Berry 1976) and 'polarization reversal' (Geyer and Kontuly 1996), became popular catch-phrases in the literature on urban North America and Western Europe.

During this phase, metropolitan areas in most industrialized countries grew slowly, if at all, and certainly more slowly than middle-size cities and smaller urban centers. Some metropolitan areas registered absolute declines. Population growth reappeared in a few selected rural areas. Net migration flows shifted markedly toward nonmetropolitan regions and, for the first time in the statistical record, more people in North America and in much of Western Europe left metropolitan areas than moved in. To some observers, these trends represented a 'clean-break' with the past. In effect, it was argued, they spelled the end of the traditional hierarchical ordering of urban places, if not the end of the large industrial metropolis and the reversal of metropolitan dominance. It was also seen as the beginning of a rural and small town revival.

The 1980s, however, produced yet another distinctive pattern of urban growth, and another turn-around in the pattern and hierarchical organization of urban systems. Many, but not all, of the older and larger metropolitan areas showed a rebound in terms of relative growth rates as manufacturing activity underwent a modest renaissance, but especially because service and financial sectors expanded rapidly. Almost all services in the expanding financial and producer services sectors are concentrated in the larger metropolitan areas. The industrial metropolis has, it appears, been supplanted by the service-based or 'money' metropolis, and metropolitan dominance has re-asserted itself. The renaissance of small towns and rural areas, on the other hand, seemed to be short-lived. Except for those small towns located in close proximity (i.e., within commuting distance) to metropolitan regions, and a few selected retirement and recreational outposts in locations that are rich in environmental amenities (e.g., warm climate, attractive scenery, cultural and heritage sites), the rural revival was apparently over.

The final phase in this periodization followed the recession of the early 1990s, and continues through to the present. Since this period is still unfolding, it is difficult, indeed premature, to provide a formal label. Nevertheless, as described above, manufacturing activity continues to decentralize, either to outer suburbs or new industrial districts located outside of metropolitan regions, or to countries in the developing world. At the same time, fractions of the 'new economy,' in addition to financial and producer services, such as the media and knowledge industries, computer technology, e-commerce and entertainment, have continued to concentrate in the upper levels of the urban hierarchy (e.g., New York, Los Angeles, London, Tokyo). Other functions have tended to locate in smaller centers, but ones that are destined to be the new metropolises of the twenty-first century (e.g., Seattle, Austin, Frankfurt, Barcelona). Thus, even when the hierarchical ordering of metropolitan areas remains more or less intact, the functions performed by those centers and the linkages among them—that is, their network and hierarchical organization and their behavior as a system—have shifted with surprising rapidity.

4. *Issues of Current Debate and Research*

Current debates focus around a set of questions on how systems of metropolitan areas are being reorganized in response to economic restructuring, social and demographic change, and the increasing global integration of financial markets, production facilities and culture. Both research and policy concerns have shifted from a focus on deterministic economic models and those based on a single sector, to more holistic frameworks that incorporate multiple sectors and a wider range of explanatory variables—including the crucial roles of social capital, institutional capacity and political factors—in accounting for variations in metropolitan growth. Cost effectiveness, rather than cost minimization, and 'quality of life' considerations rather than traditional agglomeration factors, have become the critical factors in understanding the location decisions of investors, firms and individuals.

There has also been a shift in research from the effects of economic specialization and agglomeration economies to more flexible paradigms. One example is the re-interpretation of traditional ideas of initial and comparative advantage and the introduction of models of 'competitiveness' and innovation (Nijkamp 1990, Mills and MacDonald 1992, Brotchie et al. 1995). Competition between cities is hardly a new idea, as any reading of urban history will confirm, but its nature, intensity and scale are now different. Moreover, the players in this new competitive landscape are not simply individual metropolitan areas (such as San Francisco), but much larger multi-national con-

glomerates operating out of clusters of metropolitan areas and nearby communities that combined constitute spatially extensive metropolitan regions (e.g., the greater New York region with 20 million people, Los Angeles–San Diego with 16 million, San Francisco–Oakland–San Jose with 10 million).

The shift to a focus on innovation and competitiveness in both the research and policy literature has also heightened awareness of the central importance of two other factors—the capacity of local institutions to adapt to change and the depth of human capital available, the latter embodied in skills, knowledge, local culture and innovative talent—in determining which metropolitan areas grow and which do not. While no one has as yet come up with a simple formula for what determines urban success (although Peter Hall 1998 tries hard to do so), those that grow tend to be those that can stimulate, animate, coordinate, and otherwise take advantage of these skills and knowledge. That is, those places that exhibit the properties of and behave as 'learning regions' within the new economy paradigm. Of course, as the urban system concept suggests, they also have to be in a position within the system—i.e., have the external linkages—necessary to take advantage of such attributes.

These debates are perhaps most intense in the recent literature on globalization (King 1990, Sassen 1994, Lo and Yeung 1998). Although this literature is prone to exaggeration on the scale, timing and effects of globalization, often due to the lack of historical perspective, there is little doubt that the process is important. As everyone knows, the global economy has become increasingly integrated and interdependent, especially over the last three to four decades. Trade, transnational travel, tourism, and immigration have increased dramatically. Finance capital now moves around the globe in an instant; manufacturing supply systems are now often global in their reach, with parts designed in one country, produced in another, assembled in still another, and marketed everywhere. Information, including the media and popular culture, have also become 'internationalized,' carried by communications systems, managed by multinational media firms, and facilitated by the internet. Interestingly, most of this expanding activity, as in the initial stages of the previous industrial era, is organized by and through the largest metropolitan areas in those countries.

In effect, a new and more tightly interconnected global urban system is developing. This system, ironically, is dominated by financial and cultural world cities such as New York, London, Tokyo and Paris, that have served as 'command' centers of the global economy for some time, but now with renewed vigor (Sassen 1991, Knox and Taylor 1995). Below that upper tier are emerging second-order global centers, such as Los Angeles, Hong Kong, Frankfurt, Geneva, Sydney, San Francisco, Singapore, Bombay, Toronto, that jockey for relative position and prominence. Here the competition for status and market share is especially intense, and the rankings of places most fluid. There are, however, a number of problems with the world city hypothesis. One is that it emphasizes attributes rather than networks of external relations. Second, it tends to ignore the inter-connections between all urban places, large and small, within and between nations. In other words, it looks only at the properties and few of the linkages among the top tier cities rather than those linkages among and between all tiers, and at different spatial scales from the regional to the global level. In effect, the world city paradigm lacks the broader and more flexible perspective provided by the urban/metropolitan system framework.

All of the larger metropolitan areas are now less dependent on their respective local hinterlands and more dependent on the largest centers and the global market-place. They have, as a result, become more and more detached from their respective 'national' urban systems. The global economy, it is widely argued, is now driven more by direct competition among such global urban regions rather than among countries. Nation states, and their governments, are said to be in relative decline, their importance diminished by transnational trade, global corporations and international exchange agreements, while the global urban system is in its ascendency. Western Europe, as an economic space, can now be seen as a 'continental system of metropolitan regions' competing amongst themselves for a rich, expanding and (almost) borderless market. Borders still matter, of course, but largely for functions other than defining market share. National governments also still matter, but should be seen as part of a wider 'rescaling' or redistribution of economic and political power. Metropolitan areas are now critical agents in this rescaling.

For some observers, the information revolution, new network technologies, and global integration, are combining to rewrite the basis of spatial competition and thus the logic of metropolitan growth. In the extreme scenario, these new parameters are expected to reduce, if not annihilate, the costs (and thus the role) of the 'friction' of distance, and to eliminate many of the benefits formerly associated with the concentration of economic activity. A widespread dispersal of economic growth is then feasible, if not more efficient. This, in theory, reduces the need for large urban agglomerations. Does this mean the end of geography, the end of the large metropolis? Do space and place no longer matter? Evidence to date, however, suggests that these claims are at best exaggerated, and weakly documented, and therefore are likely incorrect. The net redistributive effects of information technologies and the rise of global cultural networks, on conditions of metropolitan dominance and polarization are problematic; the same factors can in fact lead to further concentration rather than dispersion, and to a renewed 'sense of place' rather than its demise.

While physical distance has declined as a factor in the locational calculus of most economic activities and as a cost in maintaining networks of communication, location itself remains vitally important. In other words, place still matters, and metropolitan places still matter most. It is simply that the elements which define the attributes of space and the value of location today, such as the quality of life and the need for social (and face-to-face) interaction, are weighted differently than in the past. There is no empirical evidence to support the argument that we are doing away with the large metropolis, although some of the former heavily industrial cities (e.g., Detroit, Cleveland, Manchester, Birmingham) that have failed to adapt to the new realities, have certainly declined. Instead, we are expanding many of the traditionally dominant metropolises while rapidly building new ones. In so doing we are reorganizing the metropolitan systems of which they are a part. This is perhaps especially evident in the emergence of global financial centers in Europe and Asia, such as Frankfurt, Singapore, Shanghai and Bombay. Moreover, there is substantial evidence that these metropolitan regions, operating within larger competitive urban systems, rather than nation states, are becoming more central to any explanation of global economic growth, and the diffusion of social and cultural norms.

5. *Conclusions*

The concept of the urban (now increasingly the metropolitan) system offers a way of approaching, a way of thinking about the macrogeography of the urban process. It stresses the importance of networks of connections among all urban places—that is the summation of their external relations, at regional, national and international scales—and the fact that these places behave as members of several urban systems simultaneously. The complex shifts in the fortunes of individual metropolitan areas documented above suggest the need for such a robust conceptual framework and analytical paradigm. First, the dominant trend identified here is a continuation of metropolitan dominance, although in different forms, and with a somewhat different set of winners and losers. Second, intense competition—both national and global—for a larger share of the economic pie and for cultural dominance, is increasingly seen as competition among a set (a system) of metropolitan regions rather than among national or subnational states. The metropolitan system, writ large, is now the dominant global change phenomenon.

See also: Globalization: Geographical Aspects; Planning, Politics of; Policy Knowledge: Universities; Population, Economic Development, and Poverty; Urban Geography; Urban Places, Planning for the Use of: Design Guide; Urban Planning: Growth Management; Urban Sprawl; Urban System in Geography

Bibliography

Berry B J L (ed.) 1976 *Urbanization and Counterurbanization*. Sage Publications, Beverly Hills, CA

Bourne L S 1974 *Urban Systems: Strategies for Regulation*. Clarendan, Oxford, UK

Bourne L S, Sinclair R, Dziewonski K (eds.) 1984 *Urbanization and Settlement Systems: International Perspectives*. Oxford University Press, Oxford

Brotchie J, Newton P, Hall P, Nijkamp P (eds.) 1995 *Cities in Competition: Productive and Sustainable Cities for the 21st Century*. Longman, Melbourne, Victoria

Dunford M, Kafkalas G (eds.) 1992 *Cities and Regions in the New Europe*. Bellhaven, London

Geyer M, Kontuly T M 1996 *Differential Urbanization: Integrating Spatial Models*. Arnold, London and New York

Habitat, United Nations Centre for Human Settlements 1996 *An Urbanizing World: Global Report on Human Settlements*. Oxford University Press, Oxford, UK

Hall P 1998 *Cities in Civilization: Culture, Innovation and Urban Order*. Weidenfeld and Nicolson, London

King A 1990 *Global Cities*. Routledge, London

Knox P, Taylor P (eds.) 1995 *World Cities in a World System*. Cambridge University Press, Cambridge, UK

Lo F-C, Yeung Y-M (eds.) 1998 *Globalization and the World of Large Cities*. United Nations University Press, Tokyo

Mills E S, McDonald J F (eds.) 1992 *Sources of Metropolitan Growth*. Rutgers, Center for Urban Policy Research, New Brunswick, NJ

Nijkamp P (ed.) 1990 *Sustainability of Urban Systems: A Cross-National Evolutionary Analysis of Urban Innovation*. Gower, Aldershot, UK

Pred A 1977 *City Systems in Advanced Economies*. John Wiley, Chichester, UK

Sassen S 1991 *The Global City: London, New York, Tokyo*. Princeton University Press, Princeton, NJ

Sassen S 1994 *Cities in the World Economy*. Pine Forge Press, Thousand Oaks, CA

L. S. Bourne

Mexican Revolution, The

Although its causes stretch back into the later nineteenth century—some would say further—the Mexican Revolution began in November 1910 with an armed insurrection against the regime of President Porfirio Díaz (1876–1911). It is harder to say when (if ever) it ended, hence to define the exact boundaries of the phenomenon. The 1910 revolt inaugurated a decade of intense fighting and political mobilization, which peaked in 1914–15, when Mexico became a mosaic of regional revolutionary forces, loosely allied

in shifting coalitions. The coalition led by Venustiano Carranza triumphed on the battlefield in 1915 and, for five years, Carranza's government struggled to survive; its overthrow in 1920—the last successful violent overthrow of an incumbent government—ended the armed revolutionary cycle. For a further 20 years (1920–40) the new regime, led by veterans of the armed revolution, consolidated itself, rebuilding the economy, achieving political stability under dominant party rule, and undertaking a series of social—particularly labor, agrarian, and educational—reforms which culminated in the radical administration of Lázaro Cárdenas (1934–40). Mobilizing workers and peasants, the regime resisted the challenge of dissident army leaders, conservative opponents, foreign companies, and the United States. Since the 1940s the dominant party, the PRI, continued to monopolize political power, but the thrust of its policies became more moderate—or downright conservative. Capital accumulation and industrialization now took priority over social reform and redistribution. Old enemies—the Church, the United States, foreign investors—were conciliated. And a new post-revolutionary generation of civilian politicans, technocrats, and businessmen came to dominate the booming, urbanizing Mexico of the 1950s. In generational terms, the revolution had clearly ended; and the revolutionary project—reformist, populist, and nationalist—had also faded. However, the political system set in place during 1920–40 remained, *mutatis mutandis*: and revolutionary symbols and policies—for example, the figure of Emiliano Zapata, and the agrarian reform cause which he had championed—survived, not only in official rhetoric, but also in local protests, popular movements, and even outright rebellions, such as the Zapatista uprising in Chiapas in 1994. By now, the revolution figured not only as a national myth, but also as a source of radical opposition to a regime whose radical claims had worn thin.

As a historical phenomenon, therefore, the Revolution is best seen as two consecutive phases: the armed revolution of 1910–ca.1920, and the institutional-reformist revolution of ca. 1920–40. Analytically, it can be usefully analyzed under three heads: causes, process, and outcome.

1. *Causes of the Revolution*

While some historians would seek the causes of the Revolution in the distant legacy of Spanish colonialism (large landed estates, a powerful Church, a tradition of coercive labour and ethnic discrimination), such a view fails to explain either why Mexico alone in Spanish America experienced a major revolution prior to World War II (Bolivia 1952, and Cuba 1959 bear rough comparison), or why Mexico's revolution occurred a century after independence (1821). A more persuasive interpretation focuses on the regime of Díaz, a liberal general who, ending 50 years of political instability, civil and international war, and economic stagnation, seized power and constructed a stable, authoritarian regime, which provided a generation of peace and ostensible prosperity. Behind the facade of the liberal 1857 Constitution, and invoking a positivist rationale, Díaz ran a personal regime, reelecting himself and his cronies to high office; he imposed central control, balanced the budget, wooed foreign investors, and, thanks to the new railway system, promoted a successful project of export-led economic growth. Political stability—by the 1900s, political ossification—contrasted with rapid socioeconomic change; cities grew, and with them a literate, urban middle class; a new industrial proletariat sprang up alongside the more numerous artisanate; peasant communities, especially in the dense heartland of central Mexico, lost their lands to expansionist, commercial *haciendas* (estates); throughout the country, rural communities, which comprised three-quarters of the population, came under tighter political controls and economic constraints.

Initially popular, Díaz was, by the 1900s, an aging tyrant, incapable of appreciating, still less controlling, the forces of change he had encouraged. As the memory of nineteenth-century instability faded, and the example of US and European liberalism exerted a growing appeal, protest mounted, perhaps exacerbated by the economic recession of 1907. Two principal forms of protest emerged: first, a loosely urban middle-class liberalism, which sought to make a reality of the 1857 Constitution, making the government—national, state, and local—responsible to the electorate (hence the slogan: 'effective suffrage, no reelection'); second, a broader, popular, largely peasant, social protest, directed against expansionist landlords and the abusive agents of the state—especially the *iefes politicos* (local political bosses). Two secondary sources of opposition also developed: elite families who had been excluded from power by Díaz's narrow coterie of cronies; and an urban working class which combined an 'economist' strategy of incipient unionization with support for liberal democracy. (Although some more radical working-class groups flirted with anarchism, these were a distinct minority). Liberalism—the demand for civil rights and clean elections—offered a broad platform on which these disparate forces could come together; social, nationalist and anticlerical demands were, as yet, marginal. In 1908–10 the urban middle class mounted an electoral opposition to the regime; in 1910–11, when a rigged election again returned Díaz to power, the more radical elements of the opposition resorted to armed insurrection, at which point the rural communities of northern and central Mexico took the lead.

2. *The Process of the Revolution*

The ensuing process of revolution (1910–20) varied by region and involved a politico-military narrative of bewildering complexity. It can be schematically summarized in terms of four collective actors: (a) the old regime, represented by the Porfirian political elite, the military, the landlord, and business class, and many of the Church hierarchy; (b) the urban middle class, eager for political representation but social stability; (c) the working class, similarly drawn to liberalism, but keen for economic betterment; and (d) the peasantry, concerned to conserve (or recover) lost land, while asserting local autonomy in face of an authoritarian state.

These actors were involved in a drama which comprised five acts (three of them episodes of civil war).

(a) In 1910–11 the initial broad coalition of middle-class reformists and peasant rebels forced Díaz to resign (May 1911), but failed to dismantle the old regime.

(b) The nominal leader of the opposition, the well-meaning northern landlord Francisco Madero was elected president and sought—with sincerity and some success—to promote political democracy. But free elections and free expression alarmed the old regime, while failing to satisfy Madero's insurgent peasant allies, notably Emiliano Zapata. Caught between these two fires, Madero succumbed to a military coup and was murdered in February 1913.

(c) During 1913–14 a more extensive and savage civil war ensued, as Madero's erstwhile supporters—middle class, peasants, workers—opposed the military government of Victoriano Huerta, who sought to crush the revolution 'cost what it may' and to restore the old regime. Phase (c) was, therefore, a more prolonged and decisive re-run of phase (a), Huerta, a naive autocrat, failed. Popular leaders like Zapata and Villa combined with canny northern politicians like Carranza in a powerful coalition which, winning the support of President Woodrow Wilson, defeated Huerta's army and destroyed much of the old regime. In August 1914 Huerta had fled, his army was disbanded, and the old elite definitively relinquished national power.

(d) The Revolution was, however, too divided to rule. During phase (d), the two major northern leaders—Carranza and Villa—fought for supremacy. Now, in a new twist, rival revolutionary armies clashed in a series of bloody battles, which Carranza won. Some historians see this outcome as the victory of the revolutionary bourgeoisie over the worker-peasant forces of Villa and Zapata. This is unconvincing. Carranza's broad coalition embraced peasants and workers (indeed, Carranza's leading general, Obregón, recruited working-class radicals into his 'Red Batallions'); and Villa's coalition contained 'bourgeois' reformists, provincial elites, and timeserving opportunists (it also enjoyed greater sympathy in the US). Villa's defeat did not, therefore, save Mexico from socialism. It did, however, bring to power a new political elite, nominally led by Carranza, including a clutch of hardheaded, 'self-made' men from the north—notably Obregón and Calles—who would become the architects of the postrevolutionary order.

(e) During 1915–20 Carranza clung tenuously to power, governing a country wearied by war, disease, hunger, hyperinflation, and endemic lawlessness. The new Constitution of 1917, which combined democratic political provisions with more radical commitments to land and labor reform, anticlericalism, and economic nationalism, was more a rhetorical statement of intent than a practical blueprint for government. Seeking to rig his own succession—and thus to defy the revolutionary army—Carranza was ousted and killed in 1920: the last case of a successful armed insurrection in Mexican history.

3. *Outcome of the Revolution*

While the era of major civil war had now ended, the outcome of the revolution remained uncertain. If the old regime had been destroyed, the form of the new revolutionary state was still sketchy and—despite the bold rhetoric of the Constitution—there had as yet been no major structural socioeconomic reform. Given the greater continuity of the post-1920 period, the Revolution's outcome can be schematically/thematically outlined. However, a basic chronological distinction must be made. From 1920 to 1934 Mexico was under the sway of the 'Sonoran dynasty,' northwestern leaders who combined military prowess with acute political acumen. During the presidencies of Obregón (1920–24) and Calles (1924–28), and the period of Calles' informal dominance known as the Maximato (1926–34), the regime consolidated politically, while delivering social and nationalist reform in moderate doses. Despite talk of 'socialism,' Mexico remained within the capitalist camp and, indeed, forged closer economic ties with the United States. However, the Depression coincided with a political crisis provoked by the assassination of president-elect Obregón in 1928.

Politico-economic crisis pushed the regime to the left, and under Cárdenas (1934–40) the social demands of the Revolution were belatedly met: Cardenas enacted a sweeping agrarian reform: encouraged labour organization; extended state education (while introducing a 'socialist' curriculum); and expropriated the foreign oil companies. The 'outcome' of the Revolution thus emerged incrementally over two decades, sometimes in response to critical conjunctures Obregón's assasination, the Depression). Not everything that happened in Mexican public life post-1920 was the product of 'the Revolution' (if by that we

mean the armed struggle of 1910–20); however, public life *had* been altered decisively by that struggle, often in informal ways (by virtue of migration, inflation, the erosion of old hierarchies, and the rise to power of new elites); hence the subsequent two decades of Sonoran state-building and Cardenista social reform can be seen as the continuation of the revolution by other—more peaceful, political and institutional—means.

This process can be summarized under six heads, which are diagnostic of the Mexican revolution; state-building; labour and agrarian reform; education; anticlericalism; and nationalism. Of course, these interact: agrarian reform, education, and anticlericalism all served the goals of state- and nation-building. This does not mean, however, that the story was, as some revisionist historians argue, essentially one of centralized state-building; for these initiatives also welled up 'from below,' from a restless civil society. The outcome, therefore, reflected a dialectic involving both the embryonic state and a civil society which, though weary of war, was ready to struggle for popular, often local, goals; land, schools, jobs, and a say in politics.

The Revolution spawned a new state, run by political parvenus (many from the ranks of the revolutionary armies), committed to a 'populist' policy of incorporating mass publics into new institutions, notably the *sindicatos* (trade unions), *ejidos* (agrarian reform communities), and political parties. During the 1920s the latter proliferated, and the infant state struggled against US opposition, praetorian rebellion, and popular Catholic insurrection (the Cristero War, 1926–29). But following Obregón's death. President Calles established the dominant official party, the PNR (1929), which loosely united most revolutionary groups and gradually established a monopoly of political power at all levels. Mexico did not become a one-party state (opposition, right and left, was guardedly tolerated), but, down to the 1990s, it was a dominant-party state, invulnerable to military coup, popular revolt, or party competition, hence unique in Latin America. Meanwhile, the powers of the state increased, notably in the late 1930s, when Cárdenas implemented social reforms and expanded the state's role in the economy.

Cárdenas' reforms consummated the revolutionary project of labor and agrarian reform. Trade unions benefited from legal protection, but in return accepted the partial tutelage of the state; the dominant labor confederation (in the 1920s the CROM, in the 1930s the CTM) became a close ally of the state. More radically, the government responded to peasant demands by distributing *hacienda* land to rural communities—gradually in the 1920s, rapidly in the 1930s. The agrarian reform communities (*ejidos*) benefited about half Mexico's peasantry, providing access to (but not individual ownership of) land, while promoting peasant politicization and education. By the 1930s, an extensive network of rural schools served not only to impart literacy, but also to inculcate nationalism, indigenismo (a revalorization of Mexico's indigenous culture) and, briefly, 'socialism.' Education also served the state's campaign against the Catholic Church, which has depicted—literally in the murals of Diego Rivers—as a reactionary enemy of progress, reform, and national sovereignty. The state weathered the bloody Cristero uprising (1926–29), but by the later 1930s began to mute its anticlericalism, aware that the Catholic Church was a formidable opponent and perhaps, an ill-chosen target.

Hostility to the Church was paralleled by fear of the United States. Both were threats to Mexico's fragile sovereignty and fledgling revolution. Desperate for recognition, the regime conciliated the US (the Bucareli agreement, 1923); but, as Calles consolidated power, he pursued policies which offended the US: attacking the Church, supporting Sandino and the Nicaraguan liberals, and asserting Mexico's control of its petroleum deposits. Diplomatic tension (1927) gave way to detente; but, a decade later, the chronic labor problems of the oil industry, aggravated by the oil companies' quasi-colonial attitudes induced President Cárdenas to nationalize the industry (1938). This unprecedented demonstration of Third World economic nationalism, grudgingly tolerated by a US government alarmed by Axis power and unsympathetic to the oil lobby, crowned Cárdenas's radical administration and, by provoking widespread nationalist support in Mexico, illustrated the populist and mobilizing capacity of the revolutionary state.

After 1938, however, the momentum of the revolution ebbed. A new political generation—civilian, technocratic and business-friendly—came to power: detente with both the Church and the US accelerated; land and labour reform stalled. The institutions of the regime—the dominant party, the CTM, the *ejido*—survived, underpinning a generation of political stability and economic growth (c. 1950–80), but they now pursued more conservative goals and increasingly became instruments of top-down control rather than bottom-up representation.

4. *Interpretations and Conclusions*

Like any major revolution, that of Mexico has provoked rival interpretations some of which respond to political ends. The regime itself projected a foundational myth of a popular, progressive, patriotic revolution which toppled the tyrant Díaz, confronted clerical, landlord, and foreign enemies, and brought peace and social reform to a grateful people. Though 'socialism' was trumpeted—especially in the 1930s—this was a *suj generis* socialism, independent of Moscow, rooted in Mexican soil. It offered social reform and national redemption but did not promise the abolition of the market or the creation of a planned economy. Critics on the far left (neither numerous nor

powerful) veered between outright opposition to and constructive engagement with this regime (e.g., during the years of Popular Frontism, 1935–9). Liberal critics—some veterans of the Madero revolution of 1910—castigated the regime for its authoritarianism; but their demands for political pluralism, though recurrent, went unheeded, at least until the 1980s. More powerful opposition stemmed from the right, especially the Catholic right, which attempted insurrection in the 1920s and mass mobilization (*Sinarquismo*) in the 1930s. Each of these currents—the 'official,' leftist, liberal, and Catholic—produced distinct interpretations of the Revolution.

From the 1950s, as the historical events receded and scholarly historians (especially in the United States) deived into the Revolution, interpretations again diverged. Mainstream opinion—which recognized the revolution as a genuine social movement, responding to popular grievances—yielded to revisionist critiques. Historians rightly rejected the notion of a monolithic national revolution, positing instead a multifaceted revolution, which assumed different forms in different times and places: 'many Mexicos' produced 'many revolutions.' Some historians, influenced by revisionist critiques of other revolutions, especially the French, questioned the very notion of 'revolution,' arguing (wrongly) that Mexico experienced no more than a 'great rebellion,' a chaotic blend of social protest, aimless violence, and Machiavellian careerism. Similarly, the antistatist reaction of the 1980s, evident on both the neoliberal-right and the loosely postmodern left, fueled criticism of the revolutionary state.

The right, excoriating the revolutionary state's corruption, demagogy and rent-seeking, sought to rehabilitate Díaz and justify the neoliberal state-shrinking project of the 1980s: while the new left, flirting with Foucault, denounced the state's pervasive power and applauded its historic opponents—for example, the Cristero rebels whom the state had dismissed as clerical reactionaries. The revisionist tide, therefore, contained several contradictory currents (some of them quite old) and, while it usefully swept away the complacent myth of a benign, progressive, homogenous, revolution, it carried the risk—as revisionist tides do—of simply inverting the old Manichaean certainties. If the Revolution was bad, its victims and opponents must have been good; revolutionary claims were mere demagogic cant; essentially, the Revolution conned the people and constructed an authoritarian Leviathan state.

Since these interpretations embody politico-philosophical assumptions, they cannot be easily adjudicated by recourse to empirical evidence; especially since, with the expansion of archives, research, and publications, evidence can be found pointing in all directions. It is the *balance* of this vast evidential universe which counts. My initial overview of the Revolution suggests (*contra* extreme revisionism) that the Revolution *was* a genuine revolution, embodying popular grievances and mobilization: and that the regime which ensued, despite its undoubted authoritarian and demagogic tendencies, did not—and could not—ride roughshod over a victimized or inert people. Social reform responded to state-building goals and favored the centralization of power; but it also responded to demands and initiatives from below, without which unions would not have been formed, land would not have been distributed, and schools would not have been built. State and civil society interacted in many—and contrasting—forms, throughout a large and heterogeneous country. Mexico's 'many revolutions,' therefore, varied, as regional historians have expertly shown. In aggregate, this complex dialectical process, unique within Latin America at the time, did not destroy capitalism and install socialism; indeed, it served more to remove ('feudal'?) brakes on market activity (*latifundios*, coerced labour, peasant indigence), thus contributing to a more dynamic capitalism. At the same time, it enhanced national integration and, for a generation (1910–40), made possible the grassroots popular empowerment which is the hallmark of any genuine revolutionary episode.

See also: Latin American Studies: Politics; Nation-states, Nationalism, and Gender; Revolutions, History of; Revolutions, Sociology of; Revolutions, Theories of; Violence, History of; Warfare in History

Bibliography

Benjamin T, Wasserman M 1990 *Provinces of the Revolution. Essays on Regional Mexican History 1910–29*, 1st edn. University of New Mexico Press, Albuquerque

Bethell L (ed.) 1991 *Mexico Since Independence*. Cambridge University Press, Cambridge, UK

Brading D A (ed.) 1980 *Caudillo and Peasant in the Mexican Revolution*. Cambridge University Press, Cambridge, UK

Friedrich P 1977 *Agrarian Revolt in a Mexican Village*. University of Chicago Press, Chicago

Gruening E H 1928 *Mexico and its Heritage*. The Century, London

Hamilton N 1982 *The Limits of State Autonomy: Post-Revolutionary Mexico*. Princeton University Press, Princeton, NJ

Joseph G M, Nugent D (eds.) 1994 *Everyday Forms of State Formation. Revolution and the Negotiation of Rule in Modern Mexico*. Duke University Press, Durham, NC

Katz F 1998 *The Life and Times of Pancho Villa*. Stanford University Press, Stanford

Knight A 1986 *The Mexican Revolution*, 2 vols. Cambridge University Press, Cambridge, UK

Knight A 1994 Cardenismo: Juggernaut or jalopy? *Journal of Latin American Studies* **26**: 73–107

Meyer J A 1976 *The Cristero Rebellion*. Cambridge University Press, Cambridge, UK

Rodriguez J E 1990 *The Revolutionary Process in Mexico Essays on Political and Social Change, 1880–1940*. University of California Press, Los Angeles

Ruiz R E 1980 *The Great Rebellion Mexico 1905–1924*. Norton, New York
Smith P H 1979 *Labyrinths of Power: Political Recruitment in Twentieth-Century Mexico*. Princeton University Press, Princeton, NJ
Smith R F 1972 *The United States and Revolutionary Nationalism in Mexico, 1916–32*. Chicago University Press, Chicago
Womack J 1968 *Zapata and the Mexican Revolution*, 1st edn. Knopf, New York

A. Knight

Microdatabases: Economic

Additions to scientific knowledge generally require interactions between theory and measurement. For example, the Hubbel Telescope provides astronomers with information about the universe, biologists have the Human Genome Project to provide them with a mapping of the genetic characteristics of the human species, and physicists have large scale particle accelerators to allow experiments designed to identify types of matter that could not have been observed in the past because the instrumentation was not powerful enough. It is a common observation among scientists that the rate of scientific progress is a function of the degree to which new measurement techniques enable us to observe events that could not have been observed in the past.

What are the social science counterparts of the Hubbel Telescope, the Human Genome Project, and the particle accelerator? It turns out that, in the social sciences generally and in economics in particular, the evolution of the sample survey provides a close counterpart to the natural sciences' powerful measurement devices.

In this article, the evolution of sample survey data in economics, often called microdata, is examined; databases that represent households as well as those that represent business establishments are discussed; the distinction between cross-sectional and longitudinal microdatabases are examined; and the principal features of a selection of the most important databases both in the USA and in other countries are summarized. The article concludes with a brief assessment of future developments.

1. *History of Microeconomic Databases*

The earliest use of quantitative empirical information in the development of economics as a science was probably in the living standards surveys conducted in the UK during the nineteenth and early twentieth centuries. These early survey studies were descriptions of income, expenditures, and nutrition in cities or towns, and were designed to help understand poverty. Perhaps the best known studies are those by Charles Booth (1891–7), and by Seebohm Rountree (1902), which examined York. Earlier work, more in the mode of intensive case studies than surveys, was carried out by Frederic LePlay (1855). Another set of very early studies was the establishment of firm surveys focused on measuring production. Systematic work along those lines was done in the USA during the early part of the twentieth century by Simon Kuznets and his colleagues at the National Bureau of Economic Research (Kuznets 1941). The Kuznets studies were basically macro descriptions of change over time in the level of production in the US economy, along with a variety of industry sub-aggregates. The information in these early studies was typically collected from sample surveys (or censuses) of business establishments, conducted by the US Census Bureau (see *Statistical Systems: Censuses of Population*).

A distinguishing feature of these early establishment measurements is that the basic microdata were not available to outside analysts. These studies were solely used to produce aggregate estimates for the economic system as a whole, and were not thought of as sources of information about micro-level economic behavior.

Over the twentieth century, there has been an explosion of economic measurements based on sample surveys. In the early part of the century the focus was mainly on business establishments and on the measurement of aggregate economic activity, as previously noted. Later in the twentieth century, starting with the Great Depression of the 1930s, microeconomic databases, particularly those relating to households, began to appear with increasing frequencies. In 1935 and 1936, for example, a massive study of household expenditures was undertaken by the US Bureau of Labor Statistics, in an attempt to get a better understanding of the impact of the depression on living standards. The Current Population Survey (CPS), which provides the basic measure of unemployment in the US, began in the early 1940s. The BLS conducted consumer expenditure surveys designed to produce weights for the Consumer Price Index as early as 1890, in 1935–36 as just noted, and in their current form in 1950, the early 1960s, and the early 1970s.

All of the sets of microdata just noted were designed and collected by governmental units. During the latter part of the twentieth century, these governmental efforts began to be supplemented with microdata collections funded by governments but designed and conducted in the private sector, typically involving academic survey research units, such as the National Opinion Research Center (NORC) at the University of Chicago and the Survey Research Center (SRC) at the University of Michigan. These private sector microdata sets were almost entirely household data, and included the series of studies of consumer net worth titled the Survey of Consumer Finances (begun in the late 1940s), the National Longitudinal Surveys

of labor market activity, the Panel Study of Income Dynamics, and the Retirement History Survey (all begun in the late 1960s). During the 1970s, 1980s, and 1990s there was a dramatic increase in both public and private household microdata collections (see *Databases, Core: Sociology*). Finally, the last decades of the twentieth century saw the beginnings of publicly available business establishment microdata in the USA.

2. *Structure of Microeconomic Databases*

Microeconomic databases belong to one of four possible categories, depending on whether they sample households or firms, and whether they are cross-sectional or longitudinal. Whether surveys are cross-sectional or longitudinal depends on whether or not the same household or establishment is followed through time, or whether a new sample is drawn for each study so that the same household or establishment is never in consecutive studies except by chance. Studies can also be focused on households or individuals, often called demographic surveys, or they can examine business firms or establishments, often-called economic surveys.

The great bulk of the microdatabases available to the scientific and policy communities for analysis consist of household or individual microdatabases, not establishment or firm databases. These household microdatabases tended to be cross-sectional earlier in the twentieth century, and increasingly longitudinal later in the century.

In part, this is due to the fact that the early databases were likely to be focused on a particular policy variable, where measurement of that variable was the principle focus of the study. Thus, for example, the Current Population Survey (CPS) has a very large monthly sample of cases because its principle objective is to estimate the unemployment rate with great precision, not only for the country as a whole, but also for states and other geographic areas. The CPS has limited information with which to explain employment rates, but the very large sample provides a rich description of the level, change, and regional variation in unemployment rates. In contrast, microdata like the National Longitudinal Surveys or the Panel Study of Income Dynamics were designed as longitudinal microdatabases able to track and eventually model the dynamics of change over time.

These tendencies (for household microdata to be used more than firm or establishment data and for analysis to be increasingly longitudinal) are clearly visible in the published literature (Manser 1998, Stafford 1986). Over the period 1984–1993, for example, there were over 550 articles in the major US economic journals dealing with household microdata, compared to a little over 100 using firm or establishment microdata. Moreover, the great bulk of the latter were not based on nationally representative establishment data collected by the Census or by the BLS, but were typically local studies with the data collected by the researcher.

The very limited use of establishment data in the published literature compared to household data is probably due to three factors. First, until very recently it was difficult, if not impossible, for researchers to access establishment microdata in the USA, due to the (accurate) perception that privacy/confidentiality considerations were a much more serious problem for establishments than for households; after all, it is impossible to hide GM or AT&T in a microdata set. Second, the theory of household or individual behavior is much better developed than the theory of business behavior, probably due in part to the greater availability of rich household microdata. Third, microdata sets for establishments are likely to be produced in the public sector and to be designed to track change over time in policy relevant aggregate variables. Thus they will be relatively strong on well-measured policy variables, and relatively weak on a rich set of explanatory variables—a combination that makes analysis of the dataset less likely to show up in scientific journals.

The tendency for household studies to be increasingly longitudinal is probably due to two factors. First, it took the economics profession some time to discover the advantages of longitudinal over cross-sectional analysis, and to develop the appropriate statistical tools. Second, the earlier sets of public microdata were strongest at providing careful measurements of the policy relevant variables that justified the survey, which requires only cross-sectional data; they were typically less strong on including the explanatory variables essential to successful longitudinal modeling.

The principal features of a number of microdatabases are summarized, including databases where the major focus is on measurement of a dependent variable (where the design is apt to be cross-sectional), and others where the focus is on providing a rich explanation of behavior (where the focus is likely to be on careful longitudinal measurement of a very large set of explanatory variables). The series that we cover in this discussion are mainly household data series available in the USA, but some information on microdatabases available from Western Europe, Asia, and the developing world is provided.

The series covered includes the Surveys of Consumer Finances (SCF); the Survey of Consumer Attitudes (SCA); the National Longitudinal Surveys (NLS); the Panel Study of Income Dynamics (PSID); the Survey of Income and Program Participation (SIPP); the Consumer Expenditure Survey (CES); the Time Use survey (TU); the Current Population Survey (CPS); the Health and Retirement Study (HRS); the Luxembourg Income Study (LIS); the Living Standards Measurement Surveys (LSMS); and the British,

German and European versions of the Panel Study of Income Dynamics (The British Household Panel Survey, BHPS; the German Socio-Economic Panel, GSOEP; and the European Household Panel Survey, EHPS). A brief discussion of establishment databases is also included.

2.1 *Survey of Consumer Finances* (*SCF*)

The Survey of Consumer Finances, sponsored by the Federal Reserve Board, was initiated just after World War II to provide some insight into the likely behavior of consumers who had accumulated large amounts of assets during the war, when many types of consumer goods and services were unavailable and thus a large fraction of income was saved. The early surveys contain detailed descriptions of asset ownership and amounts of assets held across a rich spectrum of financial and tangible assets—financial assets like savings accounts, checking accounts, CDs, stocks and bonds, in later years money market accounts, IRAs and Keoghs, etc., along with tangible assets like investment real estate, businesses and farms, and vehicles.

Growing dissatisfaction with certain features of the SCF data, in particular with the fact that total financial asset holdings as measured by the survey were much lower than total financial asset holdings as measured in the Federal Reserve Board Flow of Funds accounts, substantially reduced Federal Reserve Board support after 1960. The series was continued with less financial detail and broader sponsorship during the 1960s, and basically came to an end in its original form in 1969. The series was revived in 1983 with a sampling feature aimed at remedying the substantial underestimate of asset holdings shown by all of the previous surveys. This design feature was the addition of a high income/high wealth sample of households selected from statistical records derived from tax files. SCFs of this sort were conducted in 1983, 1989, 1992, 1995, and 1998 and are scheduled on an every-third-year basis.

In addition to the detailed asset data, the SCF contains a comprehensive income section, has occasionally obtained substantial data on the pension holdings of respondents from the companies providing those pension plans, has a relatively standard set of household demographics and typically contains other variables of analytic interest—savings behavior, subjective health status, inheritances received, intended bequests, etc. (For more information about the SCF, see their website: http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html.)

2.2 *Survey of Consumer Attitudes* (*SCA*)

Surveys of Consumer Attitudes began in the USA in the late 1940s, originally as a 'soft' introduction to the questions about detailed asset holdings and income sources obtained in the Surveys of Consumer Finances. These attitude surveys moved to an intermittent quarterly basis in the early 1950s, to a regular quarterly basis in the early 1960s, and to a regular monthly basis in 1978.

The basic content of these surveys includes a set of measures designed to reflect consumer assessments of their current financial situation compared to the past, their expectations about future financial conditions, their expectations about business conditions during the next year and the next five years, and their assessments of whether the present is a good or bad time to buy durable goods. Three of these core consumer attitude questions are part of the US statistical system of leading economic indicators. The SCA contains substantially more data than the indicator series, including information about expected price change, perceptions of the effectiveness of current economic policy, assessments of buying conditions for houses and cars, expectations about changes in unemployment rates, etc.

Data on consumer attitudes based on US experience are now routinely collected in a number of other countries, including Austria, Australia, Belgium, Canada, China, Czech Republic, Denmark, Finland, France, Germany, Great Britain, Greece, Hungary, Ireland, Italy, Japan, Luxembourg, Norway, Poland, Russia, Spain, South Africa, Sweden, Switzerland, and Taiwan. (Information about the SCA can be obtained from their website: http://athena.sca.isr.umich.edu/scripts/contents.asp.)

3. *National Longitudinal Surveys*

The Bureau of Labor Statistics at the US Department of Labor sponsors the National Longitudinal Surveys (NLS). The surveys began in 1966 with studies of the age cohorts of men aged 45–59 and of women in the age cohorts of 30–44. Young men and women aged 14–24 were added in the late 1960s, and young men and women in the age cohorts of 14–22 (called NLSY) were added in 1979. The children of the 1979 cohort were added in 1986. Of the six cohorts, data collection on four are continuing, while data collection on the remaining two (the original older male cohort aged 45–59, and the cohort of young men aged 14–24 in 1966) have been terminated.

The NLS surveys combine respondent interviews with a series of separately fielded administrative data collections, especially for the NLSY and the children of the NLSY cohorts. The information collected directly from NLS sample members includes information about work, family, educational experience, home ownership status, pensions, delinquency, and basic demographic information. The administrative data collection includes school characteristics obtained for the young women and young men cohorts

and for the NLSY group, as well as for the cohort of children of the NLSY. In addition, school transcripts including coursework and attendance records were collected for the NLSY respondents in the early 1980s, and similar information is being obtained for children of the NLSY. Finally, aptitude and achievement scores from standardized tests were transcribed from school records in the late 1960s for the young men and young women cohorts, in the early 1980s for NLSY respondents, and during 1995 for the children of the NLSY cohort. Summary pension plan descriptions are collected for the mature women respondents and/or their spouses who report pension plan coverage, and death certificate information was collected in the early 1990s for most of the deceased members of the older men cohort. (More information about the NLS can be obtained from their website: http://stats.bls.gov/nlshome.htm.)

3.1 *Panel Study of Income Dynamics* (*PSID*)

The PSID, started in 1968, was based on combining a sample from the Survey of Economic Opportunity, originally conducted by the Department of Health and Human Services in 1966 and 1967, with a probability sample of individuals 18 years of age and older. The sample was heavily overweighed from the beginning with relatively poor households, since the SEO sample was designed to study poverty and had a disproportionate number of poor households.

The PSID has unique design features that have contributed to its status as being (probably) the most widely used microdataset in the world. A critical design feature is the way in which the PSID continues to be cross-sectional representative of the full US population while maintaining the longitudinal characteristic that enables analysts to trace the dynamic of change over time. Representativeness for the US population (except for new immigrants) is maintained by following PSID family members who leave their original sample households and begin their own households. This feature ensures a continued representation of newly formed households in the sample, accurately reflecting how the population changes with the birth of newly formed households. Thus PSID members are born, live their lives, and die just as individuals in the population do, and PSID traces the original sample individuals through their entire lifespan accompanied by whichever family they happen to be attached to—the original family in which they were a child, a new family when they formed their own household, etc.

The PSID content is directed mainly at labor force experience and economic status, with a standard set of demographics. The study is especially rich in terms of data on jobs, job transitions, and detailed sources of income from work.

In recent years, the PSID has added significant modules on new topics. Health status, health insurance coverage, and functional health began to be collected in the early 1990s; wealth measures were obtained at five-year intervals starting in 1984; and detailed pension and savings data were first collected in 1989. The wealth measures have been particularly successful. Finally, the PSID has had a major influence on microdata developments in other countries. Counterparts to the PSID can be found in the British Household Panel Survey (BHPS), the German Socio-Economic Panel (GSOEP), and the European Household Panel Survey (EHPS). (For more information about the PSID, see their website: http://www.isr.umich.edu/src/psid/index.html.)

3.2 *Survey of Income and Program Participation* (*SIPP*)

The SIPP was begun in the early 1980s, with the basic objective of providing much richer income information for households at the lower end of the income distribution who might be eligible for various types of government public assistance programs. Thus the survey contains considerable detail on work experience, participation in a variety of public welfare programs, and on sources of income, along with family composition, a few health measures, and a standard set of demographics.

The SIPP basically has a cross-sectional design, although a number of observations are obtained for each respondent household in order to obtain very detailed information about income flows and work experience. Thus the SIPP is conducted three times a year, with respondents providing information about the most recent four-month period. Respondents stay in the sample for approximately six quarters, although in recent years the duration of stay has been extended so that a more genuinely longitudinal feature is available. In addition to income, work experience, program participation and some health items, the SIPP also has a wealth module that is administered several times to each participating household. Selected outgoing rotation groups from the SIPP are currently being followed longitudinally to observe the impact of the 1990s welfare reform on income, labor force participation, and poverty. (More information about the SIPP can be obtained from their website: http://www.sipp.census.gov/sipp.)

3.3 *Consumer Expenditure Survey* (*CES*)

The Consumer Expenditure Survey, conducted by the Bureau of Labor Statistics, has the principal objective of providing weights for the calculation of the Con-

sumer Price Index. These studies have a long history: the first survey was conducted in 1888–1891, the next in 1901 to get a better measure of price inflation in food, the third in 1917–1919 to provide weights for a cost-of-living index, the next two during the depression of the 1930s to look at poverty issues, and the sixth, seventh and eighth in 1950, the early 1960s, and early 1970s, respectively. The current version began in the early 1980s, when the survey was modified to be a continuous study with new samples appearing approximately every other year.

The survey goes into enormous detail on household expenditures across a long list of major product classifications (e.g., clothing, food, recreation, travel, insurance, cars, household durables, etc.) and provides estimates of consumer expenditures over past time periods for these categories. In addition, respondents are asked to keep a detailed diary, where they report each item purchased over a two-week period. The diary information and the expenditure data for the product classifications (e.g. clothing for the respondent, for the spouse, for any children, etc.) are then integrated to produce an estimate of total consumer spending. In addition to consumption data, the CES includes detailed information about income, and also has some asset data along with the standard demographics. (More information about the CES can be obtained from their website: http://stats.bls.gov/csxhome.htm.)

3.4 *The Time Use Survey (TU)*

Time Use studies have a relatively lengthy history, and actually go back into the very early part of the twentieth century when studies of local communities in the US, and of cities in the then USSR, were conducted. As a major survey effort designed to provide national data on both market and non-market activities, the Time Use studies effectively started in 1965, when a cross-national study of urban time use was organized by a set of researchers under the direction of the Hungarian sociologist Alexander Szalai (1972). The countries included Belgium, Bulgaria, Czechoslovakia, France, Federal Republic of Germany, German Democratic Republic, Hungary, Peru, Poland, USA, USSR, and Yugoslavia. Subsequently, time use studies in the US with roughly comparable methodologies were conducted during 1975–1976, and during 1981–1982. Other time use studies in the USA were conducted in the late 1980s and early 1990s, although some design differences in these later studies make comparability with the earlier 1965, 1975–1976 and 1981–1982 studies difficult. Most European (and some Asian) countries conduct time use studies periodically, typically every 5 or 10 years, with a common survey design.

One feature of the time use studies that differentiates it from most other microdatabase efforts is the difference in design between the US and other countries. All studies that focus explicitly on time use collect data for a (usually retrospective) 24-hour period—a time diary, as contrasted to a series of questions about the frequency of a set of activities. For the most part, these studies have been used as parts of national accounting systems, in which the major focus is on attempting to estimate the volume of non-market work, travel, leisure activities, etc. as a supplement to the National Income and Product Accounts. Thus the major focus, particularly in Europe, is on the macroeconomic uses of time diary data.

In contrast, time diary data in the US have been designed for use in a microeconomic as well as a macroeconomic framework. For that purpose, data for a single day consist mainly of statistical noise, and multiple time diaries for different days of the week and seasons of the year need to be collected for each sample member. Thus the US design specified that data were to be collected for two weekdays, one Saturday and one Sunday, weighted to produce an estimate of time use during a typical week. In contrast, time diary studies in Europe typically (but not always) collect data for a single day for each respondent; in some recent studies multiple time diaries are collected for each respondent, often getting information for both a weekday and a weekend day in the same survey.

Data archives on time use studies are maintained at Essex University in the UK. (For more information, see their website: http://www.iser.essex.ac.uk/mtus.)

4. *Current Population Survey (CPS)*

The Current Population Survey, probably the most widely reported survey in the US if not in the world, is basically designed to produce an estimate of the unemployment rate for the nation as a whole, as well as for geographic subdivisions like states or SMSAs. The CPS, started in the early 1940s, is designed to produce estimates of work status during the prior week, including time spent looking for jobs as well as time spent working, for a very large (about 50,000 cases) monthly sample. The survey is a very simple and very brief instrument that collects virtually nothing else besides the policy relevant variables—unemployment and work status for the adult population in the USA—from which the unemployment rate for the US as a whole, or for particular states or SMSAs, can be calculated. The CPS does have additional survey content, since there are a series of annual supplements that include such variables as a detailed assessment of family income, assessments of health status for family members, etc. But the core CPS is a very brief interview focused almost entirely on employment and job search experience. (For more

information about the CPS, see their website: http://stats.bls.gov/cpshome.htm.)

4.1 Health and Retirement Study (*HRS*)

The Health and Retirement Study, sponsored by the National Institute on Aging, was begun in 1992 with the recognition that the US statistical system had virtually no current data focused on the work situation, health status, family responsibilities, and pension status of the very large cohort of Baby Boom individuals (those born between 1946 and 1964) who would be starting to retire in the early twenty-first century. The behavior of this cohort was thought to be critically important for assessing the societal stresses associated with the prospective dramatic change in the proportion of the working population to the older dependent population.

The study, which collects information every other year, was originally focused on the US population between the ages of 51 and 61 (those in the birth cohorts of 1931 to 1941). This design would provide information with which to model the labor force participation decisions of individuals who were (mainly) not yet retired but who would retire shortly. In subsequent years, the survey was modified from one designed to follow an original cohort through its retirement experience, to a continuing study of the US population of age 50 and older. It was designed to follow the retirement experience of successive cohorts of individuals starting with those born between 1931 and 1941 and continuing with those born up through the Baby Boom years. Thus the current study includes the original cohort of those born between 1931 and 1941 (and their spouses if married), and includes other cohorts added in subsequent years (those born before 1923, those born between 1924 and 1930, and most recently those born between 1942 and 1947).

The content areas of the HRS were subject to much more than the usual amount of developmental work. Survey content is decided by a set of researchers organized into working groups and headed by a Steering or Oversight Committee. The criteria that were uniformly adopted were that variables that played well-defined roles in modeling retirement decisions (or other critical decisions, such as saving rates) were eligible for inclusion, but other variables were not. Thus, the working groups pulled together sets of variables measuring job characteristics and work history; family structure and intrafamily transfers; health status and health insurance; economic status, including both income and wealth; several batteries of cognitive tests; a set of probability questions relating to continued work activity; future health status; longevity, bequests, etc.; and detailed information about pensions. The magnitude, duration, and cost of this planning effort is unique and gives every indication of having produced a very high quality dataset; it is certainly possible that the HRS planning model represents the future of microeconomic database design activities. (For more information about HRS, see their website: http://www.umich.edu/~hrswww.)

4.2 Luxembourg Income Study (*LIS*)

The Luxembourg Income Study represents one of the few attempts to collect comparable data on economic status across a large sample of countries. The basic idea is to collect national data on all income sources from as wide a range of countries as feasible, process the data so as to make it as comparable as possible across countries, and then make these data available to the research and policy communities for analysis of differences in economic status and income distribution. The core of the LIS activities is to process the data collected by national statistical agencies, to have a staff of trained analysts interact with their counterparts in these national agencies, and to develop measurements of income and income components that are as comparable as possible across countries.

The most important goal for LIS is harmonization—reshaping and reclassifying the components of income or definitions of household structure into comparable categories. Such harmonization allows the researcher to address important social issues without having to invest countless hours in getting every variable that will be analyzed into a comparable format.

Since its beginning, the LIS project has grown into a cooperative research project with a membership that includes countries in Europe, North America, the Far East and Australia. The LIS database now contains information for more than 25 countries for one or more years, covering over 90 datasets over the period 1968 to 1997. The countries currently in the LIS database include: Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Hungary, Ireland, Israel, Italy, Luxembourg, The Netherlands, Norway, Poland, Portugal, ROC Taiwan, Russia, Slovak Republic, Spain, Sweden, Switzerland, UK, and the USA. Negotiations are currently under way to add data from Korea, Japan, New Zealand, and other countries. (For more information about LIS, see their website: http://www.lis.ceps.lu/access.htm.)

4.3 Living Standards Measurement Study (*LSMS*)

The World Bank established the Living Standards Measurement Study in 1980 to explore ways of improving the type and quality of household data collected by government statistical offices in developing countries. The objectives of the LSMS were to develop new methods for monitoring progress in raising levels of living, to identify the consequences for

households of current and proposed government policies, and to improve communications between survey statisticians, analysts, and policy makers. To accomplish these objectives, LSMS activities have encompassed a range of tasks concerned with the design, implementation and analysis of household surveys in developing countries. The main objective of LSMS surveys is to collect household data that can be used to assess household welfare, to understand household behavior, and to evaluate the effects of various policies on the living conditions of the population. Accordingly, LSMS surveys collect data on many dimensions of household well-being, including consumption, income, savings, employment, health, education, fertility, nutrition, housing, and migration. Three different kinds of questionnaires are normally used: the household questionnaire, which collects detailed information about household members; the community questionnaire, in which key community leaders and groups are asked about community infrastructure; and the price questionnaire, in which market vendors are asked about prices.

Because welfare is measured by consumption in most LSMS research on poverty, the measurement of consumption is strongly emphasized in the questionnaires. There are detailed questions on cash expenditures, on the value of food items grown at home or received as gifts, and on the ownership of housing and durable goods. A wide range of income information is also collected, including detailed questions about wages, bonuses, and various types of in-kind compensation.

LSMS-type surveys have been conducted in a long list of countries, including Peru, Ivory Coast, Ghana, Mauritania, Jamaica, Bolivia, Morocco, Pakistan, Venezuela, Russia, Guyana, Nicaragua, South Africa, Vietnam, Tanzania, Kyrgyzstan, Ecuador, Romania, and Bulgaria. For more information on the Living Standards Measurement Surveys, see Grosh and Glewwe 1995. (See also the LSMS website at: http://worldbank.org/html/prdph/lsms/lsmshome.html.)

5. *Establishment Microdata*

Central Statistical Bureaus produce virtually all nationally representative establishment datasets. In the US, which has a unique decentralized statistical system, establishment databases are typically collected by the Census Bureau and designed (and commissioned) either by Census or by the Bureau of Labor Statistics. As noted in the introductory remarks for this essay, the major use of establishment databases tends to be macroeconomic rather than microeconomic, in that they are used to produce national estimates and industry distributions of production, wage costs, employment, fringe benefits, work hours, etc.

The history of establishment databases sponsored by the BLS goes back to the 1800s, starting with occupational pay surveys. The current employment statistics series (CES or 790) began at BLS in 1915, the unemployment insurance series (ES 202) was started in 1935, the Occupational Employment Statistics (OES) survey in 1971, the Employment Cost Index (ECI) series in 1976, the Employee Benefits Survey (EBS) in 1980, and the hours at work survey in 1981. All of these surveys are designed to produce aggregate statistics from which monthly, quarterly, or annual estimates of change can be calculated and used by policy makers and business forecasters. The microdata underlying these aggregates are typically unavailable to the research community, although there are occasional exceptions under tightly controlled circumstances.

Establishment databases that are the responsibility of the Census Bureau basically include a census of establishments in all industries conducted every five years (available starting in 1963), and an annual survey of manufacturing establishments (available starting in 1972). The census data for all industries are designed to provide, among other statistics, benchmark data for total production, price deflators, and product detail. The annual surveys of manufacturing provide data on shipments, wage costs, capital expenditures, materials consumption, and energy use.

These census data are, as with the BLS data described earlier, designed to produce aggregate industry statistics relevant to policy makers and forecasters. However, in recent years Census has tried to increase the use of the microdata in these surveys by creating data enclaves—Longitudinal Research Data Centers. These enclaves exist not only at the Census Bureau, but have also been installed at a number of other locations (Boston, Pittsburgh, Los Angeles, and Berkeley). The intent is to encourage researchers throughout the country to use the microdata, with the expectation that such research will not only turn up interesting scientific findings but will also suggest improvements in the basic data.

There is somewhat greater use of establishment microdata in Western European countries than in the USA, largely because there appears to be a bit less concern with privacy (confidentiality) considerations in Europe. In some countries, establishment microdata can be combined with individual microdata, permitting, for example, analyses of the way in which worker wage trajectories vary with industry.

6. *The Future of Economic Microdata*

Developments during the last decades of the twentieth century suggest that the twenty-first century is likely to see a continued expansion in the use of longitudinal household microdata sets, along with a vigorous growth in the use of establishment databases for micro-level analysis. These developments will be the result of three compelling forces: first, the increased

ability of economists to understand behavior based on the combination of richer theoretical insights and the greater availability of relevant data; second, the increased involvement of academic economists in the design of microeconomic datasets; and third, the development of more effective strategies for safeguarding privacy and confidentiality while still permitting researchers access to microdata files.

See also: Censuses: History and Methods; Data Archives: International; Databases, Core: Demography and Registers; Databases, Core: Sociology; Economic Panel Data; Quantification in History; Survey Research: National Centers

Bibliography

Booth C 1891–1897 *Life and Labours of the People in London*. Macmillan, London and New York

Davis S J, Haltiwanger J, Schuh S 1996 *Job Creation and Destruction*. The MIT Press, Cambridge, MA

Duncan G J, Hofferth S, Stafford F P (Forthcoming) Evolution and change in family income, wealth and health: The panel study of income dynamics, 1968–2000 and beyond. In: House J S, Kahn R L, Juster F T, Schuman H, Singer E (eds.) *A Telescope on Society: Survey Research and Social Science in the 20th and 21st Centuries*. University of Michigan Press, Ann Arbor, Chap. 6

Grosh M E, Glewwe P 1995 *A Guide to Living Standards Measurement Study Surveys and Their Datasets*. World Bank, Washington, DC

Kuznets S S 1941 *National Income and its Composition, 1919–1938*. National Bureau of Economic Research, New York

LePlay F 1855 *Les Ouvners Europeans*, 6 Vols

Manser M E 1998 Existing labor market data: Current and potential research uses. In: Haltiwanger J, Manser M E, Topel R (eds.) *Labor Statistics Measurement Issues*. University of Chicago Press, Chicago

Rountree B, Seebohm 1902 *Poverty: A Study of Town Life*, 2nd edn. Macmillan, London

Smeeding T M 1999 Problems of international availability of microdata: The experience of the Luxembourg income study (LIS) infrastructure project. In: Chlumsky J B, Schimpl-Neimanns B, Wagner G G (eds.) *Kooperation zwischen Wissenschaft und amtlicher Statistik—Praxis und Perspektiven*. Metzler Poeschel Publishers, Wiesbaden, Germany, pp. 186–98

Stafford F P 1986 Forestalling the demise of empirical economics: The role of microdata in labor Economics Research. In: Ashenfelter O C, Layard R (eds.) *Handbook of Labor Economics*. Elsevier, New York, Vol. 1, pp. 387–423

Szalai A 1972 *The Use of Time*. Morton, The Hague

Woodbury R (ed.). 2000 The health and retirement study, part I—history and overview. *Research Highlights*, **7** (May 2000).

F. T. Juster

Microevolution

1. *Concept*

The word 'evolution' derives from the Latin term *evolutio*, which means 'an unrolling.' It can be used in this sense, or in others which involve the idea of change. But not all changes are evolutionary. The ocean's surface is always changing, but this is not an evolutionary process. Implicit in the concept of evolution are those of (a) continued change; (b) divergence; (c) restriction of opportunities; and (d) in a large number of situations, irreversibility. It is questionable whether there is a general direction (hence, 'progress') in the organic evolutionary process.

The primary factors which determine evolution are mutation (genetic change) and selection (in Charles Darwin's words 'The preservation of favorable individual differences and variations, and the destruction of those which are injurious' (Darwin 1859)). The two, however, can be influenced by population structure. Among the variables that should be considered in this latter context, mention should be made of population size, mobility and life histories of its members, sex and the genetic system, and assortative mating.

The dazzling array of organic diversity present on earth has always fascinated mankind. An important property of this variability is that it is discontinuous. Organic matter is organized in discrete individuals and groups of individuals. Moreover, these groups may be isolated reproductively from other similar ensembles. When this occurs it is said that they are different species.

Microevolutionary processes are those that occur within a species. *Homo sapiens* is a particularly favorable subject for the investigation of these processes because, of course, humans know much more about themselves than is true for any other organism. On the other hand, humans are unique among other forms of life because they developed culture. This species-specific trait varies according to its own laws, and largely independently of biological factors. The reciprocal, however, is not true; the organic evolution in humans may be strongly influenced by cultural processes.

2. *Mutation*

The information needed to assure the transmission of the biological inheritance is given by a molecule, deoxyribonucleic acid (DNA). The particular sequence of its component units (nucleotides) is important because it codes (through another molecule, ribonucleic acid, or RNA) for the sequences of amino acids, the protein units that combined in arrays make up physical bodies. The unit of heredity is the gene

(present in a locus), composed of coding (exons) and noncoding (introns, that probably have regulatory functions, i.e., they may modulate gene expression). The transmission takes place in DNA-protein arrays called chromosomes. The genetic material can occur either in the nucleus of the cells or in the cytoplasm, in an organelle called the mitochondrion.

Detrimental agents or problems during the replication process may lead to mutations, changes in the previously existent DNA sequence. Advances in the molecular study of these structures revealed that many different types of change may occur (alteration in just one nucleotide, deletions or insertions of many of them, triplet (the coding unit is made of trios) expansions, variations in nucleotide position, changes in the process of DNA → RNA → protein information transfer, unequal recombination or gene conversion between different DNA strands. Moreover, these variations do not occur at random. There are changes that are allowed and others that are forbidden, depending on the importance of the particular DNA region for cell physiology. Their fate will also be much different if they occur in noncoding or coding regions. A special type of change is that which occurs due to horizontal DNA transfer, that is, pieces of DNA (transposons) move from one species to the other, or between chromosome regions of the same species, helped by infectious agents.

3. *Natural Selection*

Natural selection is also a name that involves many different processes. For instance, selection may be directional, when it changes the adaptive norm of a population; stabilizing, when this norm is protected; or balancing, which includes all those factors that maintain genetic heterogeneity or polymorphisms (common variants). Selection occurs at all levels of the biological hierarchy (molecular, cellular, tissue, organ, individual, population, species, community of different taxonomic entities), and its agents accordingly diverse (Salzano 1975, Williams 1992).

Selection can act through differences in mortality or fertility, and a dialectic dilemma that faces any organism is how much energy it should invest in individual survival, as opposed to reproductive efforts. Throughout human history significant changes have occurred in the emphasis given to these two processes, and they have been connected with subsistence patterns greatly influenced by culture. Among hunters and gatherers, mobility was very important, and the nomadic way of life would disfavor large numbers of offspring. Several measures were then developed to assure a small number of children (herbal contraceptives, mating taboos, abortion induction by mechanical means, infanticide). Variance in number of children, a key evolutionary parameter, was thus restricted. Since population numbers were small, infectious agents would not be very important, and mortality would be mainly determined by other environmental factors or by violent deaths determined by rival groups.

The situation changed with the domestication of plants and animals. The premium now was to have a large number of children, to help in the cultivation and harvest of plants, or to take care of the animals. The possibility of storing large quantities of food made life more independent of the environment, and there was a concomitant increase in population size. Higher mortality levels then developed, due to epidemics and also due to the fact that it is easier to take care of two as compared to, say, 10 children.

Modern technology made possible a better control of number of children and of infectious agents. The action of natural selection, therefore, became more restricted.

This does not mean that the human species is no longer evolving. Our interaction with infectious agents and other parasites can be compared to an arms race, in which development of a given trait in the host may be followed almost immediately by a corresponding change in the attacking organism. Fascinating cases of coevolution can then emerge; some of these were reviewed in Levin et al. (1999). As an example of the possibilities of analytical studies in this area, mention can be made of Andrade et al.'s (1999) work. Using a specific molecular technique, Andrade and colleagues were able to detect *Trypanosoma cruzi* (the protozoan responsible for Chagas' disease) directly on tissues of affected organisms, and confirmed experimentally the clear differential tissue distribution of diverse clones. This, of course, is important information for the understanding of the disease's pathogenesis, with concomitant implications for therapy.

Indirect evidence for the action of natural selection can be obtained using homozygote frequencies at the allele level, and the ratio of synonymous to nonsynonymous DNA changes (that is, modifications that lead to different amino acids as contrasted to those that result in the same amino acid). Salamon et al. (1999) considered these points in relation to three loci of the human leukocyte antigen system, which regulate human immune response against pathogenic agents. For all of them they found strong indication of balancing selection at the amino acid level.

4. *Population Size*

Natural selection can only operate on the material that is available to it. There are several factors, therefore, that can also influence the destiny of a population or of its gene pool. Size is one of the most important. In a small population, random factors can determine the fate of a given variant independently of its adaptive value. Wright (1978) discussed in detail the several misunderstandings that developed in the adoption and

discussion of the concept of random drift. It is important to specify, for instance, whether the phenomenon arose from accidents of sampling, or from fluctuations in the systematic (mutation, selection) pressures. These stochastic events, in combination with the deterministic factors indicated above and migration rates, may lead in a species with multiple partially isolated local populations to what this author calls the shifting balance process of evolution.

Also important, in human populations, are unique events such as the adoption of a given invention or spread of a major technological advance. Thus, the population expansion of *Homo sapiens* that seems to have occurred about 50,000 years ago is generally associated with the 'creative explosion' of the upper-Paleolithic-type technology.

Given the importance of population size for evolution, it may be asked what numbers have prevailed for the most part of human history. Of course, only inferences can be made for prehistoric populations, but a series of statistical or mathematical methods have been developed which relate present genetic diversity or the overall branch length of a genealogical tree to a parameter called effective size. The latter is the breeding size of an abstract population in which the effects of population size and subdivision are taken into consideration. It has been suggested that the effective size of human populations is about one-half of their census size.

Harpending et al. (1998) contrasted two hypotheses about past population sizes. The 'hourglass' hypothesis proposed that there was a contraction (bottleneck) in the number of human ancestors at some time before the last interglacial in the Pleistocene, but that the previous population was large and distributed over a large part of the Old World. The 'long-neck' hypothesis, on the other hand, postulated that the human ancestral population was small during most of the Pleistocene. Data they assembled on the variability of the mitochondrial DNA and *Alu* insertions (transposition-type elements) suggested that the second hypothesis was the correct one, and that the human population in that period would have had an effective population size of 10,000. Jin et al. (2000), using two sets of dinucleotide repeats (28 and 64, respectively), obtained even lower numbers (2,301–3,275).

5. *Migration*

Besides population size, the amount of migration among groups is critical in the interpretation of genetic variability. At the tribal level, what happens can be described broadly by the fission–fusion model proposed by Neel and Salzano (1967). Hunter-gatherer groups experience cyclic events of fissions (mainly due to social tensions, the migrating units being composed of lineal relatives) and fusions, as convenience dictates. This demographic pattern changes dramatically as socioeconomic conditions lead to larger agricultural and urban populations. Large-scale migrations also conditioned the mixing of diverse continental groups. The genetic consequences of this population amalgamation were considered by Chakraborty et al. (1988). They concluded that the treatment of these agglomerates as panmictic (random mating) populations can lead to erroneous estimates of mutation rates, selective pressures, and effective population sizes.

A special aspect of the consequences of migration is the 'founder effect' (Mayr 1942). This term designates the establishment of a new population by few founders, who carry only a small fraction of the total genetic variability of the parental population. Distinct alleles or gene arrangements can become more prevalent in different regions due to this phenomenon. For instance, in African-derived Americans, the beta-S haplotypes (distinct allele arrangements present in the hemoglobin gene) Benin and Bantu show clearly different prevalences in Brazil and Mexico, as compared to other nations of the continent. Bantu is the most frequent in these two countries, while Benin is the most prevalent haplotype in North America and the Caribbean area (review in Bortolini and Salzano 1999). This difference probably arose due to the diverse source of African slaves that came to the New World during the sixteenth to nineteenth centuries.

6. *Assortative Mating*

Mating choice is a complex behavior characteristic, which involves psychological, cultural, socioeconomic, and biological variables. Its main evolutionary influence is on the distribution of genotype frequencies. In multiethnic communities there is a clear preference for homogamic matings. Recently, with the use of exclusive matrilineal (mitochondrial DNA) or patrilineal (Y chromosome) genetic markers, it is possible to evaluate the influence of sex when heterogamic unions occurs, in historical perspective. Thus, in Latin Americans of mixed ancestry, the European component was mainly contributed by males, while the Amerindian fraction is mostly derived from females. This is a reflection of unions that occurred in the Colonial period and that could be detected independently of the demographic, cultural, and biological changes that occurred afterwards (Salzano and Bortolini 2001).

7. *Intra- and Interpopulation Variability*

Studies at the blood group, protein, and DNA levels, all indicated that in humans the intrapopulation variation is far higher than the interpopulation variation. An overall evaluation of this question was made by Barbujani et al. (1997), who indicated that the

intrapopulation variability, independently of the markers used, is of the order of 85 percent, that which occurs among populations within a continent is of 5 percent, and that among continents of 10 percent. Intercontinental differences, therefore, are small and do not support the concept of continental 'races.' This does not mean that, using a convenient array of genetic markers, we cannot establish with complete confidence the continental origin of a given population or of the ancestors of a given individual (Salzano 1997). Most of the interpopulation variability, also, consists of gradients of allele frequencies (see, for instance, Rothhammer et al. 1997), and not of abrupt discontinuities. The biological data, therefore, is in complete agreement with the ethical concept of the brotherhood of humankind.

See also: Brain, Evolution of; Cultural Evolution: Theory and Models; Darwin, Charles Robert (1809–82); Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionary Epistemology; Evolutionary Selection, Levels of: Group versus Individual; Evolutionary Social Psychology; Genes and Culture, Coevolution of; Human Evolutionary Genetics; Intelligence, Evolution of; Lifespan Development: Evolutionary Perspectives; Natural Selection; Social Evolution, Sociology of; Sociality, Evolution of

Bibliography

Andrade L O, Machado C R S, Chiari E, Pena S D J, Macedo A M 1999 Differential tissue distribution of diverse clones of *Trypanosoma cruzi* in infected mice. *Molecular and Biochemical Parasitology* **100**: 163–72

Barbujani G, Magagni A, Minch E, Cavalli-Sforza L L 1997 An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences USA* **94** 4516–9

Bortolini M C, Salzano F M 1999 β^S haplotype diversity in Afro-Americans, Africans, and Euro-Asiatics—An attempt at a synthesis. *Ciência e Cultura* **51**: 175–80

Chakraborty R, Smouse P E, Neel J V 1988 Population amalgamation and genetic variation: Observations on artificially agglomerated tribal populations of Central and South America. *American Journal of Human Genetics* **43**: 709–25

Darwin C 1859 *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London

Harpending H C, Batzer M A, Gurven M, Jorde L B, Rogers A R, Sherry S T 1998 Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences USA* **95**: 1961–7

Jin L, Baskett M L, Cavalli-Sforza L L, Zhivotovsky L A, Feldman M W, Rosenberg N A 2000 Microsatellite evolution in modern humans: A comparison of two data sets from the same populations. *Annals of Human Genetics* **64**: 117–34

Levin B R, Lipsitch M, Bonhoeffer S 1999 Evolution and disease—population biology, evolution, and infectious disease: Convergence and synthesis. *Science* **283**: 806–9

Mayr E 1942 *Systematics and the Origin of Species*. Columbia University Press, New York

Neel J V, Salzano F M 1967 Further studies on the Xavante Indians. X. Some hypotheses-generalizations resulting from these studies. *American Journal of Human Genetics* **19**: 554–74

Rothhammer F, Silva C, Callegari-Jacques S M, Llop E, Salzano F M 1997 Gradients of HLA diversity in South American Indians. *Annals of Human Biology* **24**: 197–208

Salamon H, Klitz W, Easteal S, Gao X, Erlich H A, Fernandez-Viña M, Trachtenberg E A, McWeeney S K, Nelson M P, Thomson G 1999 Evolution of HLA Class II molecules: Allelic and amino acid site variability across populations. *Genetics* **152**: 393–400

Salzano F M (ed.) 1975 *The Role of Natural Selection in Human Evolution*. North-Holland, Amsterdam

Salzano F M 1997 Human races: Myth, invention, or reality? *Interciencia* **22**: 221–7

Salzano F M, Bortolini M C 2001 *The Genetics and Evolution of Latin American Populations*. Cambridge University Press, Cambridge, UK

Williams G C 1992 *Natural Selection: Domains, Levels, and Challenges*. Oxford University Press, New York

Wright S 1978 *Evolution and Genetics of Population. Vol. 3. Experimental Results and Evolutionary Deductions*. University of Chicago Press, Chicago

F. M. Salzano

Microsimulation in Demographic Research

Microsimulation is a computer-dependent technique for simulating a set of data according to predetermined probabilistic rules. It was originally applied primarily to problems in the physical sciences, such as gamma-ray scattering and neutron diffusion. Three characteristics of the method, highlighted in an early description (McCracken 1955), are that the problems to which it is applied depend in some important way on probability; that experimentation is impracticable; and that the creation of an exact formula is impossible. Each of these characteristics makes the method particularly useful in a variety of demographic applications.

1. The Monte Carlo Method and Microsimulation

The Monte Carlo method takes its name from the games of chance popularly associated with the resort of the same name. At its simplest, the method is akin to coin-tossing. There is a known probability—one-half, if the coin is fair—of the toss of a coin resulting in a head or, conversely, in a tail. If we toss a fair coin a number of times we expect to get heads half the time and tails half the time.

A demographic application is the calculation of the average waiting time to conception, that is, the average number of months a woman will take to conceive. In this example, the monthly probability of conception (fecundability) might be set at 0.2 (rather than 0.5, as

in the coin-tossing example). One can think of this as requiring a coin that comes up heads with probability 0.2, and tails with probability 0.8: a very unfair coin indeed. The weighted coin is tossed until it comes up heads, and the number of tosses until this happens, which represents (or simulates) the number of months it will take for an individual to conceive, is recorded. The trial is repeated until the desired sample size of women has been achieved. Then, the average waiting time to conception is calculated by summing the number of months each individual 'woman' took to conceive, and dividing this sum by the total number of women.

Such a physical experiment—'a simple game of chance with children's marbles'—was carried out by de Bethune (1963, p. 1632) in his examination of the expected spacing between births. Taking two green marbles, representing the number of days in a menstrual cycle during which conception is possible, and 26 red marbles, representing the remaining days, de Bethune drew a marble at random and if it was red, returned it to the pot and drew again. His objective was to count the number of draws necessary to produce a green marble, which is equivalent to the number of 28-day cycles until conception occurred. He repeated the experiment 200 times, and tabulated the results. These two problems are so simple that they can be solved, with little effort, algebraically. But with some elaboration of either input distributions or process, or both, such problems quickly become algebraically intractable.

Mechanical solutions like de Bethune's, while logically possible, are time-consuming and cumbersome. More elegantly, the overriding principle of random selection inherent in the toss of a coin or the selection of a colored marble can be preserved through use of a table of random numbers or, in probabilistic applications, a table of random probabilities. An experimental outcome is simulated by determining where a random probability falls on a known cumulative probability distribution. For example, given a probability of conception in a particular month of 0.18, the selection of a random probability up to and including 0.18 indicates that conception occurs in that month, while selection of a larger random probability indicates that it does not.

What made microsimulation a practical tool was the development of computer technology. The investigator uses a computer program that follows the paths laid out in a flow chart representing the process under investigation. At each point at which a decision must be made (such as whether a simulated individual conceives in a particular month), the program selects a random probability whose value, in comparison with a set of input probabilities, determines what will happen next. (Such computer-generated random numbers are more accurately termed pseudorandom since they are derived algorithmically, but this distinction has no practical significance here.) Information about each simulated individual is saved, and the resulting data set analyzed as though it had been derived in a more conventional way.

2. *A Brief History of Microsimulation in Demography*

Concerns about the possibility of massive population increase in the nonindustrialized countries after the end of World War II stimulated a great deal of demographic research into the determinants and components of high fertility. This research, united by an interest in how—and sometimes why—high fertility was brought about, took various directions: that of American demographers such as Notestein attempted to codify the forces that had produced the demo graphic transition of Western countries; that of French demographers such as Henry examined the patterns of childbearing of historical populations; that of Americans such as Sheps investigated the childbearing of contemporary high-fertility populations such as the Hutterites. Statistical models drawing on renewal theory appeared to hold out considerable promise, since it is a straightforward matter to envisage childbearing as a Markov renewal process (see for example Sheps and Perrin 1964). The problem was, however, that only oversimplified models were amenable to solution. As models gained in realism, they became intractable.

In the mid-1960s demographers whose investigations based on renewal theory had come to an impasse began to turn, with considerable enthusiasm, to microsimulation (see for example Hyrenius 1965). Notable among these were Sheps and Perrin in the United States, Hyrenius in Sweden, Barrett in Britain, Jacquard in France, and Venkatacharya in India. Their results were interesting and stimulating, and held out considerable promise of continuing gains in the future (see for example Barrett 1971). Undoubtedly, some of the enthusiasm for microsimulation at that time was related to enthusiasm for computer technology itself. Although it is now difficult to imagine performing empirical population analysis without the aid of a computer, this is a recent development in the history of the analysis of population data. It was only in 1964, for example, that the Population Commission of the Economic and Social Council of the United Nations 'recommended that a study be made of the possibilities of use of electronic computers for expediting and enlarging the scope of demographic analysis' (United Nations 1965, p. 1). The fact that the scientist commissioned to undertake this study was Hannes Hyrenius, who had a particular interest in microsimulation, may have given modeling in general and microsimulation in particular a degree of prominence that they might otherwise not have achieved.

Somehow, though, the promise held out in the 1960s that microsimulation might become a major weapon in the demographic armory has not been realized. What happened, instead, was that the demographic enterprise turned increasingly to data collection. Under the aegis of the World Fertility Survey (WFS), fertility surveys were conducted between 1973 and 1983 in more than 40 countries some of which had never before been subject even to a census. The expansion of the global demographic database was profound, and a great deal of energy was then directed toward data analysis. Paradoxically, the continued development of computer technology, which in the 1960s had made microsimulation a viable research tool, from the 1970s made data analysis feasible on a scale previously unimagined. This was further enabled by the development and marketing of specialized computer software. Today, new surveys continue to be conducted by the successor to the WFS, the Demographic and Health Surveys (DHS). Extraordinary gains in understanding have resulted from this burgeoning of information, and from the continued expansion of computing capability; but the construction and analysis of simulated data have, as a consequence, assumed a secondary role in the demographic enterprise.

This being said, microsimulation remains an important demographic tool. The technique continues to provide insights in various demographic applications that would otherwise not be forthcoming, as is described below.

3. *Applications of Microsimulation in Demography*

3.1 *The Simulation of Fertility, Family, and Household*

The demographic problems to which microsimulation was applied originally in the 1960s and 1970s were dominated by the quest for a better understanding of the relative importance of different components of the birth process, especially in high-fertility populations, and the estimation of levels of fertility implied by different combinations of these input components. Such applications probably remain the most common.

Demographers use the term 'fertility' to refer to reproductive performance (the term 'fecundity' being reserved for the innate biological capacity to conceive), and the estimation of fertility involves the calculation of such statistics as age-specific fertility rates, or the distributions of the number of children ever borne according to maternal age, or average numbers of children ever borne. All these measures can be obtained from a collection of individual fertility histories consisting of the dates of birth of each of a simulated individual's children. In contrast to real survey-derived fertility histories in which maternal age at each birth is determined by subtracting the mother's date of birth from the child's, simulated dates of birth are produced directly in terms of the mother's age (generally in months) at the time of confinement.

The starting place for a microsimulation model of fertility is a flow chart of the reproductive process and sets of real or hypothetical probability distributions which can take either empirical or functional forms. An individual woman enters the reproductive process at the age when she is first exposed to the risk of conception, such as when she marries, and progresses through 'life' one month, or menstrual cycle, at a time. It is generally assumed that she has passed menarche at this point. It is then convenient to assess the age at which she will become sterile, for example, through menopause, although some small proportion of women will become infecund before menopause, or even be infecund before marriage. The simulation then moves forward one month, or menstrual cycle, at a time until the individual conceives, conception being determined in the Monte Carlo manner by comparing a random probability with the individual's probability of conception (known as fecundability) in that month. If the individual does not conceive in a particular month, then, so long as she remains fecund, she is given a chance to conceive in the next one. Once she conceives, there is a possibility that she will abort spontaneously (miscarry), undergo induced abortion, experience a stillbirth, or produce a live child.

Each of these outcomes has associated with it a duration of gestation (pregnancy), and a duration of postpartum infecundity: pregnancies are short in the case of spontaneous abortions, longer in the case of stillbirths, and longest for live births; postpartum infecundity lasts for at most a few months in the case of a nonlive outcome, but may extend for up to a maximum of about 18 months in the presence of prolonged and intense breastfeeding. Subsequently, if the individual is not deemed to have become infecund in the meantime, she is once again at risk of conception. Ultimately, she does become infecund, and her reproductive history is then brought to an end.

There are many variations and elaborations on this basic structure. One such elaboration concerns exposure to the risk of pregnancy. In the simple example presented above, each individual's exposure is assumed to continue from some initiating point at least until the individual is no longer fecund, which is equivalent to assuming universal marriage, no separation or widowhood while women are still fecund, and no female mortality. Alternatively, one might posit nonuniversal marriage that can be interrupted by the death of a husband or by separation, and might posit in addition a possibility of remarriage. One might also expose individuals themselves to the risk of death. This may be a useful strategy in certain applications since the simulated histories will then

reflect the experience of all women and not just that of survivors as is perforce the case with histories that have been collected by retrospective interviewing.

Another complex of possible elaborations concerns child survival. Although the goal of such models is to estimate fertility, there are a number of reasons for taking account of child death. One is that the death of an unweaned child interrupts breastfeeding and thus leads to the mother's becoming fecund sooner than if the child had survived. Another is that if the simulated population uses contraception the mother may suspend precautions in response to the death of a child in order to conceive and bear another. This introduces the possibility of another elaboration, related to contraceptive use, which can be incorporated into a simulation model as a method-specific proportional reduction in the monthly probability of conception.

One notable absence from all of these models, however simple or elaborate, is that of men. Microsimulation focuses on an individual at a time which means, in a simulation of fertility, that the focus is on women. However, many parameters that appear to pertain to women only actually pertain to a couple: fecundability, for example, assumes the presence of a man.

The model developed by Hyrenius and Adolfsson (1964) follows in broad terms the flow chart described earlier, and employs rather simple input data: fecundability, for example, is treated as a constant. Even so, the model produced interesting output, and by attracting considerable attention is probably the best candidate for the title of ancestor of all subsequent microsimulation models of fertility. Many of these were directed, like their forebear, at illustrating the implications of certain constellations of input data (Barrett 1971). Santow (1978), for example, calibrated a fertility model by comparison with data on the Hutterites, a religious group resident in farming communities in north America who traditionally demonstrated high and natural fertility. The model was then modified to apply to the Yoruba of Western Nigeria, notably by incorporating long periods of postpartum sexual abstinence. Information on abstinence had been collected in a survey, but not detailed information on fertility other than the number of children borne to women of particular ages. The simulation model filled this gap by providing schedules of age-specific fertility rates.

In principle, variants of the basic model permit the modeling of household structure and kin relations. Once a woman's childbearing history has been simulated, the simulation of kin relations and household structure is a matter only of extending the model to incorporate factors such as survival or widowhood of the mother, and survival, leaving home, and marriage of her offspring, along with assumptions about the formation of separate households when offspring marry (for examples see Ruggles 1987, Wachter et al. 1978).

3.2 Evaluation of Methods of Analysis

Microsimulation has proved useful for evaluating the validity of a method of analysis, for demonstrating that a method is invalid, for illustrating those conditions under which a method breaks down, and for demonstrating the extent to which a method is biased. It is often possible to show by mathematics whether or not a method 'works,' but a demonstration by means of microsimulation is both simpler to present and more compelling to an audience of nonmathematicians. In addition, in the event that a method does not work, it may be difficult to quantify by mathematical means alone the extent to which its output is misleading.

In the absence of direct information on births and deaths of individual children, levels of infant and child mortality can be estimated indirectly from tabulations of the proportions of children who have died according to their mother's age. This indirect method for the estimation of child mortality can be tested by constructing a microsimulation model of childbearing over the reproductive span, with allowance made for given levels of infant and child mortality. The output from the model, consisting of individual histories of childbearing and child survival, can be manipulated to show the total number of children ever borne and children surviving by age of mother; implied mortality probabilities can be calculated according to the indirect method, and the results compared with the known input probabilities of mortality. One such exercise (Santow 1978, pp. 144–9) demonstrated that the indirect method worked well so long as the pace of childbearing was independent of the number of children already borne, but that once measures were adopted to limit family size to a particular number the method broke down. One might have anticipated that this would be the case because the derivation of the indirect method incorporated an empirical function representing the age pattern of natural fertility, which differs from that of controlled fertility, but the simulations permitted quantification of the extent of the bias.

Microsimulation has also been used to evaluate the operation of two indirect methods of detecting fertility control (Okun 1994). The methods are 'indirect' in the sense that the data to which they are conventionally applied—those compiled from historical or Third-World sources—do not include direct measures of contraceptive use. Nevertheless, just as childbearing is affected by fertility control, so are histories of childbearing and the aggregate statistics derived from them; and the methods were devised with the intention of detecting the telltale traces of such control. As with the example of indirect estimation of mortality, the evaluation followed the form of a controlled experiment. Fertility histories were simulated under known conditions of contraceptive use, and the extent of fertility control was calculated by means of each of the

indirect methods being evaluated. The inferred levels of fertility control were then compared with the known levels, which had been used as input in the microsimulation models. The exercise was valuable because it cast doubt on both the indirect methods examined. In this application, moreover, it is difficult to know how the indirect methods could have been evaluated comprehensively except by means of microsimulation.

For further examples of the use of microsimulation to evaluate methods of analysis see Bracher and Santow (1981) and Reinis (1992).

3.3 Inference of Probable Input when Outcomes are Known

The examples described thus far can be viewed as forward projections, since they deal with the implications for demographic outcomes of particular known sets of input conditions. It is possible, however, also to 'go backwards'—to use microsimulation to assess what sort of input conditions are consistent with particular known outcomes. This is perhaps a more delicate exercise than the more usual one since it is conceivable that the same outcomes may derive from different combinations of input factors. To take a simple example, a low overall birth rate can result from high birth rates within marriage, very low ones outside marriage, and low proportions married; or from low birth rates within marriage and higher proportions married.

A recent example of such work sought to infer the levels of contraceptive use and the proportions of pregnant teenagers seeking abortion that were consistent with observed shifts in teenage birth rates and teenage abortion rates in Sweden (Santow and Bracher 1999). Microsimulation was indicated because abortion rates expressed in terms of women, which is the form in which official data are tabulated, can fall either because fewer conceive, or because fewer pregnant women seek abortion. The application was justified because the models demonstrated that particular combinations of the proportion using contraception and the proportion who would seek abortion if pregnant created unique pairs of birth and abortion rates.

4. Independence and Levels of Aggregation

In formulating a microsimulation exercise the analyst is always confronted with a critical question concerning the level of detail, or conversely of aggregation, that is necessary in order to represent sufficiently faithfully the process whose implications are being investigated. Hyrenius and Adolfsson (1964), for example, created useful results from simulations that incorporated a monthly probability of conception (fecundability) that was invariant not just according to age, but also over the entire population. In later applications it was found useful to vary fecundability by age (for example, Santow 1978) or between women (for example, Bracher 1992). The latter application, the aim of which was to assess the effect on the timing of the subsequent conception of various combinations of no breastfeeding, full breastfeeding, and contraception of various efficiencies adopted at six weeks postpartum or at the appearance of the first menstrual period, also incorporated a prospectively obtained distribution of times to the first menstruation among breastfeeding women, menstrual cycles of varying lengths, and varying probabilities of ovulation's preceding the first postpartum menstruation.

We might contrast this very fine disaggregation of a segment of the childbearing process with the approach typically taken by microsimulators of kinship relations in historical populations. The former problem, concerning the timing of a subsequent conception under varying patterns of breastfeeding and the adoption of postpartum contraception, does not seek to go so far as to simulate complete birth histories of women whence are derivable conventional fertility statistics such as age-specific fertility rates and distributions of children ever born (although it would be possible to elaborate the model to produce such output). Problems of the latter type, however, typically take such conventional fertility measures, as derived from an initial, historical population, as their starting point. Since the aim of such exercises is to examine the implications of particular demographic scenarios for subsequent generations, the microsimulators quite naturally want to start with a known scenario which, if it is to be constructed from the basic flow chart described earlier, would require a great deal of trawling through the available scanty data, and a great deal of trial and error. They thus employ rather aggregated demographic parameters, such as the distribution of children ever born, so that their simulation output will mimic from the outset the properties of a known population.

The problem, as Ruggles (1993) has pointed out, is that demographic behavior is correlated between the generations. Fertility is correlated within population subgroups defined by family background, and also within families; and mortality probabilities are correlated within members of the same kin group. Failure to take such correlations into account leads to an underestimate of the heterogeneity in simulated populations of kin. Ruggles terms the assumption that the characteristics of members of a kin group are uncorrelated the Whopper Assumption, and shows that the resulting error can be large.

The heart of the problem is an assumption of independence where true independence does not exist. Yet one aspect of the power of microsimulation lies in its ability to incorporate heterogeneity in underlying characteristics (such as fecundability and the risk of child death) into the simulated population. Another aspect of the power of microsimulation is its ability to

introduce dependence between events—to take account of the fact that different events lead to different subsequent pathways. (Indeed, it was precisely the failure to take account of the lack of independence of such demographic behaviors as using contraception, being infecund after a birth, and being married, that invalidated the analytical model discredited, by means of microsimulation, by Reinis [1992].) Selecting the level of disaggregation of a process, and hence of population heterogeneity, that is appropriate for the issue under examination is a critical element of a useful microsimulation model.

See also: Families and Households, Formal Demography of

Bibliography

Barrett J C 1971 A Monte Carlo simulation of reproduction. In: Brass W (ed.) *Biological Aspects of Demography*. Taylor and Francis, London

de Bethune A 1963 Child spacing: The mathematical probabilities. *Science* **142**: 1629–34

Bracher M 1992 Breastfeeding, lactational infecundity, contraception and the spacing of births: Implications of the Bellagio Consensus Statement. *Health Transition Review* **2**: 19–47

Bracher M, Santow G 1981 Some methodological considerations in the analysis of current status data. *Population Studies* **35**: 425–37

Hyrenius H 1965 Demographic simulation models with the aid of electronic computers. World Population Conference. IUSSP, Liege, Vol. 3, pp. 224–6

Hyrenius H, Adolfsson I 1964 A fertility simulation model. Reports 2. Almquist & Wiksell, Stockholm, Sweden

McCracken D D 1955 The Monte Carlo method. *Scientific American* **192**(5): 90–6

Okun B S 1994 Evaluating methods for detecting fertility control: Coale and Trussell's model and cohort parity analysis. *Population Studies* **48**: 193–222

Reinis K I 1992 The impact of the proximate determinants of fertility: Evaluating Bongaarts's and Hobcraft and Little's methods of estimation. *Population Studies* **46**: 309–26

Ruggles S 1987 *Prolonged Connections: The Rise of the Extended Family in Nineteenth-Century England and America*. University of Wisconsin Press, Madison, WI

Ruggles S 1993 Confessions of a microsimulator: Problems in modelling the demography of kinship. *Historical Methods* **26**: 161–9

Santow G 1978 *A Simulation Approach to the Study of Human Fertility*. Martinus Nijhoff, Leiden, The Netherlands

Santow G, Bracher M 1999 Explaining trends in teenage childbearing in Sweden. *Studies in Family Planning* **30**: 169–82

Sheps M C, Perrin E B 1964 The distribution of birth intervals under a class of stochastic fertility models. *Biometrics* **20**: 395

United Nations 1965 Study on the use of electronic computers in demography, with special reference to the work of the United Nations. Economic and Social Council E/CN.9/195

Wachter K W, Hammel E A, Laslett P 1978 *Statistical Studies of Historical Social Structure*. Academic Press, New York

G. Santow

Middle Ages, The

The Middle Ages are good to think about. Situated by definition between Antiquity and Modernity, they constitute an intermediate period between the grand beginnings of European civilization and its contemporary achievements, endeavoring to attain these standards and even to surpass them. An intermediate status is always ambivalent. If one thinks in strictly dualistic terms, it is no longer the old and not yet the new, or, conversely, it allows the partial survival of the old and prepares the birth of the new. But it can also become an equally dignified participant in a triad, an age with special characteristics of its own. All this offers a paradise for conflicting interpretations, deprecations, and apologies; for an eternal to and fro of disdain, glorification, exoticism, and rediscovery.

1. The Notion of the Middle Ages

The most enduring characterization of the Middle Ages appeared before the term itself was coined. It was inherited from Italian Humanists who had been accustomed since Petrarch to lament the 'dark' and 'barbarous' period between the glories of Greek and Latin Antiquity and their rebirth in their own age. This negativity was reinforced by the Enlightenment, which bluntly labeled it an age of ignorance and superstition (Voltaire). On the other hand, no sooner had the Renaissance begun to evolve into the new (post-Columbian, Gutenbergian, Copernican) world of Modernity; no sooner had the Middle Ages come to be regarded as definitely over; than condemnation was sidelined by waves of nostalgia: chivalric dreams resurrected by Ariosto, Cervantes, and Tasso; ideals of sainthood and a unified *Christianitas* cherished by the Catholic Reformation, the Bollandists, Chateaubriand, and the Romantics. Medieval achievements such as Gothic cathedrals inspired both freemasonry and waves of occultism. The brutish energy of castles and armor haunted the imagination with Gothic Romance, then they became crystallization points for heroic-chivalric romantics and Wagnerian 'new barbarians,' to degenerate finally into the comic world of Monty Python. The organic beauty of medieval cities and crafts was rediscovered by the Pre-Raphaelites, Burckhardt, and Huizinga, art nouveau and art deco.

To these ambivalent judgments a third dimension must be added, that of different identities (local, institutional, national, and European). Medievalists inherited monastic annals, urban or royal chronicles, order histories, and the history of western Christendom presented as world history. Much reasoning on the Middle Ages was motivated, after Machiavelli's *Discorsi* (1525), Guicciardini's *Storia d'Italia* (1540),

or Shakespeare's king-dramas by an interest in 'national' roots. The antiquarian curiosity of Muratori (1672–1750) resulted in his path-breaking Rerum Italicarum Scriptores; the mighty and systematic collection of sources on the German past, the Monumenta Germaniae Historica, was started by Pertz (1826), the Histoire de France by Michelet (1831), the Récits des temps mérovingiens by Thierry (1840), and the studies on English common law by Maitland (1895). In the twentieth century, historical attention turned towards the universalistic frameworks—Christianity, Papacy, Empire—which could be analyzed as the 'progenitors' of Europe. The Middle Ages meant the formation of modern Europe, which, on the basis of those earlier achievements, was able to prepare itself for global expansion.

After the twentieth century's world wars and revolutions there was another resurgence of interest in the Middle Ages. Interdisciplinary approaches found a treasure-trove in the 'total' history of medieval civilization, and an unexpected series of bestsellers appeared: not only 'new history' monographs in the footsteps of *Montaillou* by Emmanuel Le Roy Ladurie (1975), but also new-style historical romances such as *A Distant Mirror* by Barbara Tuchman (1978) and *The Name of the Rose* by Umberto Eco (1980). From the 1980s onwards, châteaux, monasteries, and city halls, but also modern university campuses became involved in the ongoing reinvention of the Middle Ages. More recently, two other elements were added to this picture. The collapse of communism in Eastern Europe reoriented historical consciousness in the direction of medieval origins, sometimes in a nineteenth-century style, as a claim for historical–territorial legitimacy, sometimes as a foundation of the 'common roots' of a new European identity. The other novelty was provided by chronology: the year 2000 brought about a new fascination with the apocalypse, with some divergence in Eastern Europe, where, instead of chiliastic calculations, pompous millennial celebrations were held to commemorate the glorious foundation of their Christian states. Which Middle Ages should we address in the new Millennium? Whose concept should we adopt? We shall approach these issues from three angles: (a) periodization, (b) structures, and (c) a brief survey of narratives, personalities, and artifacts.

2. *Extension and Limits of the Middle Ages*

As already indicated, the concept of the Middle Ages originated in the Renaissance. The term *media tempestas* made its appearance in 1469, in a eulogy of Cusanus written by Giovanni Andrea dei Bussi, a pontifical librarian. Analogous expressions (*media aetas*, *media antiquitas*) appeared frequently and became allied with the already established concept of the 'dark ages' familiar from the writings of Petrarch, Leonardo Bruni, Lorenzo Valla, and other Italian humanists.

However, the notion of a transitory period between antiquity and its rebirth was older and this affects the chronological limits of the Middle Ages. Laments upon the lost greatness of ancient Rome and hopes of its imminent revival were voiced by Gregory the Great. From Justinian I and Charlemagne to Otto III and Frederick Barbarossa, the slogan '*renovatio imperii Romani*' repeatedly allowed the conceptualization of intermediate decadent periods. For the 'renaissance of the twelfth century,' Arabic learning could perhaps be regarded as the mediator between the wisdom of the *antiqui* and the Christian learning of the *moderni*.

The negative image of the medieval decadence of classical civilization, however, was duly counterbalanced by the positive image of the advent of Christianity, the triumph of which over Roman 'paganism' was praised as the demise of 'superstition,' 'tyranny,' 'idolatry,' and 'magic.' Judgments upon the period following the fall of the Roman Empire were thus rather positive during the Middle Ages, or at worst ambivalent. This is also reflected in the first elaborate system imposing a ternary division upon world history, that of Joachim of Fiore (d. 1202). Projecting the model of the Holy Trinity upon the history of mankind, Joachim distinguishes between the age of the Father (Old Testament), the age of the Son (New Testament)—which lasts from Jesus to the writer's own age (which corresponds to the Middle Ages)—and the imminent new period of the Holy Spirit, which will have an 'Eternal Evangile,' and in which everybody will follow Christian life-precepts to the letter.

Early-modern historiography inherited from the Renaissance the ambivalent image of the Middle Ages and extended them to embrace universal history. Georg Horn, a Leiden historian, put the *medium aevum* between 300 and 1500 in his *Arca Noe* (1666). Christoph Cellarius, a historian from Halle, was the first to realize his world history on the basis of the ternary division into Antiquity (*Historia antiqua*, 1685), Modernity (*Historia nova*, 1686), and the Middle Ages between them (*Historia medii aevi*, 1688). The start of the *medium aevum* was for him the reign of Constantine and the acceptance of Christianity (313), and the end was marked by the fall of Constantinople (1453).

These limits were challenged by later suggestions. The 'start' was increasingly identified with the forced abdication, in 476, of Romulus Augustulus, the Western Roman Emperor, whose vacant throne was never taken. The 'end' came to be associated with the discovery of America (1492) or the start of the Reformation (1517). National histories have tended to adjust the dates in accordance with their own temporal divisions: the English tend to count modernity from the advent of the Tudors (1485), while the Hungarians take the fatal defeat at the hands of the Ottomans at

Mohács in 1526 as the end of medieval Hungarian history.

With an eye to Polish discourse on the 'feudal economy' in early modern times Jacques Le Goff made the provocative proposal of a 'long Middle Ages' lasting until the French Revolution. Extension can go in both directions, however: in the wake of Peter Brown's work on 'late antiquity,' a five-century-long transitional period has come to replace the divide constituted by the 'fall' of the Roman Empire, a period embracing the history of the Mediterranean basin from Diocletian to Charlemagne. The extension of the Middle Ages also has a geographical dimension. With the Roman Empire and Christianity as basic points of reference, this notion, born in Italy and nurtured in Western Europe, has always embraced more than *Europa Occidens*.

The Mediterranean basin as a lively unit of civilization throughout the Middle Ages has always been a truism for historians of commerce, crusades, or navigation, a tradition resumed by Pirenne in his *Mohammed and Charlemagne* (1937) and by Braudel in his *Méditerranée* (1949). This rediscovery obliged historians of Latin Christianity to become acquainted with the large and autonomous tradition of Byzantine studies (Ostrogorski, Obolenski, Patlagean) and to engage in systematic comparisons. In parallel with this renewed interest in an integrated panorama of the South, there was also a resurgence of nineteenth-century visions of the 'barbaric,' individualistic vitality and communal freedoms of the Germanic and Slavic North (*Markgenossenschaft*, *zadruga*). The medieval Roman and Germanic synthesis also became the object of a new type of structural history for Pirenne (1933) and Bloch (1939). The medieval origins of Eastern, Central, East-Central, and South-Eastern Europe—these 'borderlands of Western civilization'—have also been examined in impressive syntheses by the Romanian Nicolae Iorga (1908–12), the Czech Francis Dvornik (1949), the Polish Oscar Halecki (1952), and the Hungarian Jenö Szücs (1981).

3. *Structures in the Middle Ages*

Defining a period within history requires the characterization of the structures that provide its identity and coherence. These might be 'objective,' 'deep' structures; the product of a momentary combination of various factors; or simply an order observed by those seeking genealogies *a posteriori*.

The most general structural account imposed upon the medieval period is that of 'feudalism' (see *Feudalism*), a notion which embraces social, economic, and cultural relations. Based on medieval legal concepts, '*féodalité*' served as the most general characterization of the Ancien Régime during the Enlightenment (Montesquieu, Voltaire, Vico); was further developed by economic and sociological theory (Smith, Marx, Weber); and finally, with Bloch, Duby, Boutruche, and Le Goff, provided a comprehensive account of 'feudal society' and its economic structures, social stratification, habits, *mentalités*, rituals, and cultural and religious achievements. Historical analysis found an attractive subject in these personalized power-structures which supplemented the early medieval weakness or lack of sovereign states. Sealed with the ritual of vassality and with the conditional donation of a *feudum* (fief), 'feudalism' constituted a complex relationship of military, economic, and social (and, within that, parental–familial) dependence and solidarity, one of the most original features of the Middle Ages and seemingly the foundation of Europe's subsequent achievements.

Departing from the chaotic self-organization of early-medieval times and evolving into the flowering of the 'second feudal age' (Bloch) social/relations were analyzed in accordance with their ecclesiastical, chivalric, courtly, and urban ramifications. The concept of feudalism also made possible the comparison of medieval Europe with other world civilizations—Marxist historiography relied upon this when coining its universal evolutionary pattern of world history. In recent decades, new approaches have discussed the 'feudal economy,' 'feudal mutation' around the year 1000 (Poly and Bournazel 1991), and regional European variants of feudal social structures. On the other hand, the exaggerated extension of this notion also led to an outright rejection of the concept (Brown 1974, Reynolds 1994).

Besides this synthetic vision of an all-encompassing pyramid of 'feudalist' dependencies, other original medieval types of social structure should be mentioned. 'Orders,' appearing around the year 1000 as a horizontal classificatory category, referred to alliances of monasteries such as the Cluny *congregatio* or later the Cistercians; they were also used for a general functional tripartition within society (*oratores–bellatores–laboratores*). The latter 'ideology' (Duby 1980) prepared the late medieval appearance of 'estates' as social entities meriting political representation. Chivalric orders, religious confraternities, urban guilds, and medieval universities gave this 'corporative' principle a crucial role within medieval society.

Social classes, stratification, and social types are frequently discussed as a combination of contemporary distinctions and posterior classification. Within the principal division of ecclesiastical society and laity may be distinguished the nobility, with its higher and lower strata; courtly and urban society, with the respective microcosms of rank-related, social, institutional, religious, and ethnic distinctions among its permanent and temporary inhabitants; the peasantry, with various phases of unification or diversification of status; and the broad space of roads, forests, and suburbs left for a marginal existence, where the poor, outlaws, pilgrims, merchants, hermits, and wandering

knights mingled. Within these social categories one can also observe elementary structures: the family (with its gender and age divisions, and the changing status of women, children, and the old); 'parenthood' (understood as extended kinship group); the 'individual'; 'community'; '*gens*'; and '*natio*.'

Disruptions and transformations caused by invasions and internal wars should also be considered: attacks by and, eventually, the settlement of migrating tribal warrior societies (Goths, Franks, Lombards, Huns, Turks, Avars, Vikings, Hungarians, Cumans), and conflicts with Arabic peoples from the fall of the Visigothic state to the successful *Reconquista*. External threats not only disturbed but also unified, however: the consciousness of being '*europeenses*' emerged among Carolingians confronting the Arabic threat. Medieval Europe entered the phase of expansion with the Crusades and the evolution of long-distance trade, and the concept of European identity reappeared with the *timor Tartarorum* in the thirteenth century and the progress of the Ottoman occupation in the fourteenth and fifteenth centuries.

Throughout the Middle Ages, Europe possessed a general structure of universal bonds which both mingled with a variety of local and regional particularisms and existed in continuous tension and conflict with them. Politically speaking, this universality was embodied in the Empire, or rather in the two inheritors of the Roman Empire, the Carolingian (subsequently the Holy Roman) Empire in the West and Byzantium in the East. These universal powers were confronted by a jungle of local, regional, institutional, communal, and individual counterpowers. Resisting tribal retinues, territorial lordships supported by fortified castles, ecclesiastical immunities, urban liberties, noble privileges, constitutional rights, courtly intrigues, popular revolts, and bloody wars paved the way for the emergence of 'national' kingdoms. Finally, the principle of *rex imperator in regno suo* subjugated both universal structures and local autonomies; the birth of the modern Europe of sovereign nation-states was among the most significant outcomes of the Middle Ages.

From the religious and cultural standpoints, the universal bond within medieval Europe was *Christianitas*, based upon the ecclesiastical structures (bishoprics, dioceses, parishes) of the Roman Papacy and the Eastern Churches with their three sacred languages, Latin, Greek, and Church Slavonic. The history of medieval Christendom brought a continuous expansion of internal and external conversions, defeating the resistance of the '*barbaricum*,' integrating local beliefs and cultural diversity, and relying upon the omnipresent cult of saints increasingly controlled and standardized by the Papacy, the universal networks of monastic and mendicant orders, and an emerging intellectual elite issued from medieval universities, imposing a unified pattern of literacy, Roman and Canon Law, scholastic theology, a kind of *philosophia perennis*. Within such universal frameworks, historical evolution unfolded along the conflict-ridden lines of movements of religious renewal (often repressed as heresies) and a series of renaissances (the Carolingian, Ottonian, and twelfth-century renaissances and 'the' Renaissance) which made possible an increasing absorption of the Antique cultural heritage. Ongoing conflicts between 'learned' and 'popular,' Old- and New-Testament-based, 'mystic,' 'liturgical,' 'rational,' and 'secularized' variants of Christianity should also be mentioned. Towards the end of the Middle Ages, the universal frameworks could no longer contain the internal tensions, and they succumbed, as in the political sphere, to the ascending nationalization of churches, legal systems, and vernacular literatures, and to the plurality of confessions within modern Europe, giving way to an era of religious wars, popular revolts, and witch-hunts. This enumeration of structures could continue with the 'hard' realities of geographic, climatic, ecological, demographic, and economic 'deep structures' (Braudel 1949); with the 'soft' but equally enduring structures of '*mentalité*' and '*imaginaire*' (Le Goff); with the 'thick' descriptions furnished by the case studies of historical anthropology (Le Roy Ladurie 1975, Ginzburg 1976, Davis 1975); with eternal dualities such as 'oral' and 'written' (Goody 1987, Stock 1983, Clanchy 1993), 'body' and 'mind' (Bynum 1995); with basic media such as 'text' and 'image' (Petrucci 1992, Schmitt , Belting 1990, Camille 1989); or with the intricacies and variabilities of 'old' and 'new' philology (Cerquiglini 1989). In lieu of this, let me conclude with a short and necessarily arbitrary overview of some of the concrete forms in which the European Middle Ages continue to live.

4. *Narratives, Personalities, and Artifacts in a Chronological Order*

Let us begin with the grand destroyers of the 'old' (Roman, pagan Antiquity) and the grand constructors of the 'new' (medieval Christian Europe). Constantine the Great (312–337) had both qualities: renouncing the divinity of the Roman Emperor, he preserved its sacredness in a Christian form; terminating the persecution of Christianity, he found a form of survival for classical culture; and by shifting the center of his empire to the East, he contributed both to the thousand-year-long Byzantine continuation, and to the creation of the fertile dualism of the Papacy and (an ever-renewing) Empire in the Latin West. Among the 'grand destroyers', Attila, King of the Huns (434–453) merits pride of place: his archetypal figure and his exploits became a model for all oriental invaders until Genghis Khan. We should perhaps pay a little more attention to the 'grand constructors.' A

model saint, renouncing the military life for that of catechumen, hermit, monk, and bishop: St. Martin of Tours (d. 397); two church fathers: St. John Chrysostome (437–407), Patriarch of Constantinople, and St. Augustine (354–430), Bishop of Hippo, pathbreakers of personalized religiosity; an abbot, St. Benedict of Nursia, founding in 529 the model of Western monastic communities in Montecassino and writing the most influential Rule; an emperor, Justinian I (527–565), with his last attempt to restore and the first to renovate the Roman Empire, his most enduring success being the codification of Roman Law (*Corpus Juris Civilis*); and finally a pope, Gregory the Great (590–604), organizer of the medieval papacy, church hierarchy, liturgy, and conversion.

In the dim light of early medieval legends, annals, and chronicles, one may perceive influential converters, such as St. Patrick, the apostle of the Irish (390–461); impressive rulers, such as the Frank Clovis (481–511), the Goth Theoderic (493–526), or the Northumbrian St. Oswald (634–642); and powerful Merovingian queens, such as the wicked Fredegund, or the pious St. Radegund (520/5–587). Collective memory also preserves the narratives of the '*origines gentium*,' barbarian migration, conversion, and dynastic myths of emerging ruling houses (claiming divine origin). Following the *Germania* of Tacitus, we hear of the Goths, the Franks (with the Merovingian dynasty), the Anglo-Saxons (with their 'Woden-sprung kings'), and the Lombards (or later legends and stories related to the Scandinavian Yngling, the Slavic Přemysl, Piast, and Rurik houses, and the migrating Hungarians with their Árpádian dynasty).

A new chapter started with the Carolingians: the emergence of an influential family from the status of *majordomus* to that of royal and imperial dignity. Charles Martel stopped the Arab invasion at Poitiers (732), and Pepin the Short (741–768) and Charlemagne (768–814) allied with the Papacy and the latter received the imperial crown from Pope Leo III (800)—a symbolic moment in medieval history (which found its counterpoint in 1077 in Henry IV's penance at Canossa). The Carolingian renewal of state—church structures and social networks of dependence, the take-off in demographic, agrarian, and urban growth, the 'renaissance' of high culture within closed courtly and monastic circles, the successful resistance against invasions (Avars, Arabs, Normans, Hungarians): all this laid the foundations of a new, triumphant evolution which began to unfold in the West around the year 1000. There was a similar new start in the East with the Macedonian dynasty (867–1056), reorganizing the Byzantine Empire; successfully confronting Bulgarians, Hungarians, and Russians; and systematizing the army, the administration, and ceremonial under Leo VI and Constantine VII.

As for artifacts, while the splendor of Byzantium was incontestable, and radiated even in the remote mosaics of Ravenna and Venice, in the West the early Middle Ages, though unspectacular, were of vital importance, involving the widespread diffusion of the horseshoe, the iron plough, the water-mill, and the evolving naval technology of the Vikings and Italian merchants. In the cultural sphere, new technologies of writing, book-illumination, Gregorian chant, goldsmithery, evolving monastic and urban architecture, and fortified castles point to a new stage of evolution.

Around the year 1000, one ruler stands out: the 'marvel of the world,' the Saxon Otto III, third emperor of the Holy Roman Empire (983–1002), allied to his French teacher Gerbert d'Aurillac, Pope Sylvester II (999–1003), and the Czech St. Adalbert (d. 997). They created a powerful new political vision for *Europa Occidens*, relating the integration of the Germanic North and the Mediterranean South to the extension of the influence of Latin Christianity towards the Slavic and Hungarian East. In the wake of the earlier conversion of the Croats (Tomislav, 924) and the Czech lands of St. Wenceslas (d. 929), the Poland of Bolesław the Brave (992–1025) and the Hungary of St. Stephen (997–1038) became part of this 'new Europe,' which was also extended towards the North. The Danes of Harald Blueteeth (940–985) and Sweyn Forkbeard (985–1014), the Norwegians of St. Olaf (1015–1030), and later the Swedes were converted under Anglo-Saxon and Germanic influence. This Christianization was conjugated with parallel activities on the part of Byzantium: along the routes established by ninth-century Cyrillo-Methodian missions in Bulgaria and Moravia, in 996 St. Vladimir's (980–1015) Kievan Rus' was converted.

Medieval Europe was profoundly renewed after the turn of the Millennium: new dynasties (Capetians, Salians, Normans in England and Sicily) reshaped religious life under the influence of Cluny, the 'reform papacy' of Gregory VII (1073–85), and the Crusades proclaimed by Pope Urban II in 1095. In tandem with agricultural growth, the population nearly doubled in two centuries, medieval towns multiplied and acquired important immunities and liberties, and a new dynamism began to emerge.

The twelfth century brought the explosion of 'modernity,' fertile tensions and conflicts, and a wide array of important figures. Abelard (1079–1142), the daring and unfortunate 'knight' of dialectics, self-confident in his reliance upon reason, archetype of the medieval intellectual; his opponent, St. Bernard of Clairvaux (1090–1153), the charismatic leading figure of the Cistercians, founding father of the Templars and organizer of the Second Crusade (1147–49), abbot, politician, healer, and mystic; Manuel I Komnenos (1143–80), Byzantine Emperor who tried to restore the ancient splendor of the East by starting a chivalrous rivalry with the crusading West; Arnold of Brescia (d. 1155), student of Abelard, apostolic preacher, and tribune of a republican Rome that he hoped to resurrect—but hoping in vain for the help of Frederick Barbarossa (1152–90), the great emperor of the

Hohenstaufen dynasty, who was inclined towards the *renovatio imperii*—and ultimately executed for his heresies; the martyr–bishop St. Thomas Becket, who stood up to his strong-willed, youthful friend, Henry II Plantagenet (1154–89) and paid for it with his life in 1170. This was also a century of women: Heloise (c. 1100–63), the seduced student, lover, and wife of Abelard, who persisted in their intellectual and emotional partnership after Abelard was castrated by her angry relatives; Eleanor of Aquitaine (1122–1204), grand-daughter of the first *troubadour*, Duke William IX of Aquitaine (1087–1127), herself a patron of poetry, wife of Louis VII, King of France (1137–80), whom she divorced in 1152 to marry Henry II Plantagenet, taking one-third of France as a dowry; and Hildegard of Bingen (1098–1179), the visionary nun of Rupertsberg, the most prolific female writer of the age.

The twelfth century was also notable for urban handicrafts, flowering fairs, schools, and the beginning of the construction of cathedrals. It was also the century of crusades and knighthood under the banner of St. George the 'dragon-slayer,' Roland, and King Arthur and the knights of the Round Table; it was the century of orthodox or heretical itinerant preachers, from Peter the Hermit, who called for crusade, and Robert d'Arbrissel, who intervened for the emancipation of women, to Waldes, the rich Lyon merchant who distributed his wealth to the poor in 1173 and decided to live in mendicancy. When confronted with interdiction, with his disciples, the Poor of Lyon, he accepted persecution rather than submission. Another heretical movement, the Cathars, were remote followers of the antique Manichean dualists, more immediately related to the South-East European Bogomils.

At the beginning of the thirteenth century Christianity seemed fraught with conflict. There were ongoing struggles between the Empire—within the framework of which Frederick II (1212–50), the last great Hohenstauf, was preparing for his Ghibelline exploits—and the Papacy, which became a terrible rival backed by the Guelph-dominated cities. The cities influenced international politics, both in the Mediterranean and in the north: the fourth Crusade was diverted by Venice to conquer Byzantium in 1204, and the Hansa alliance, the future master of the Scandinavian–Baltic region, was growing. New problems arose within the cities: confraternities, 'people's parties,' lay preachers, rebellious students at Paris, Oxford, and Bologna universities, and bank crashes in Florence and Siena. The fight against the heretics took the form of the crusade against the 'Albigensians' (1209–29), which subjugated southern France to the north. France was also at war with England and the Empire, in the course of which Philip II Augustus (1180–1223) secured an important victory at Bouvines (1214). Nobilities confronted their rulers to gain constitutional rights (Magna Carta, 1215; Hungarian Golden Bull, 1222). This burgeoning situation was mastered with the help of two religious reformers, the founders of the two mendicant orders, St. Dominic (1170–1221) and St. Francis (1181–1226), and two popes, Innocent III (1198–1216) and Gregory IX (1227–41), who were able to restructure old resources and to integrate new forces.

With the assistance of the Dominicans and the Franciscans, the challenge of heresy could be faced and the position of the Church was re-established in the cities, where the pride of the emerging cathedrals and accumulating wealth was balanced by religious control of fashion, luxury, and usury, and by charitable institutions of burghers moved by their '*alter Christus*,' the Poverello of Assisi. The Dominicans helped to establish the feared Inquisition (1231) against the heretics and popular 'superstition'; to strike a balance between reason and faith, Aristotle and the Bible (St. Thomas Aquinas: *Summa Theologiae*, ca 1267–73); and in princely courts to Christianize the rulers. The practical morality of their 'mirrors of princes' found an echo in the ears of St. Louis IX of France (1226–70), the most pious of rulers, his ambitious brother, Charles of Anjou, King of Naples and Sicily (1266–85), and many others. The religious message of the mendicant orders was also amplified by pious princesses: St. Hedwig of Silesia (1174/8–1243), St. Elizabeth of Thuringia (1207–31), Blanche of Castile (1188–1252), St. Agnes of Bohemia (1205–82), and St. Margaret of Hungary (1242–70), not to mention the daughters of Italian merchants, bankers, and burghers: St. Claire of Assisi (1194–1253), Umiliana dei Cerchi (1219–46), Margaret of Cortona (1247–97), and Angela of Foligno (1248–1309).

By the end of the thirteenth century, medieval Europe was prospering: cloth and textile manufactures (partly dependent upon machinery), mechanical clocks, silk, paper, and a developing steel industry indicate the progress, gold coins came into circulation again. Although the devastating assault by the Mongols (1240–41) in the East (subjugating the Russian principalities) and the rise of the Ottomans in the Near East represented a serious menace, this only served to strengthen the cohesion of an *Occidens* which now included Poles, Czechs, and Hungarians. There was also a new situation in South-Eastern Europe: the second Bulgarian empire of the Asens, the reborn Byzantium of the Palaeologues (1361–1453), and the emerging Serbian state of the Nemanjic dynasty. New expansion was imminent: ambassadors and travelers reached the Far East (Marco Polo: 1254–95), Aragon was casting its gaze towards North Africa, and Norway was incorporating Iceland and Greenland.

The fourteenth century is traditionally considered a period of crisis: famines, economic decline, the Black Death (1347–52), flagellant movements, pogroms, the extinction of many royal dynasties (Árpádians, 1301;

Přemysls, 1309; Capetians, 1328; Piasts, 1370), the Hundred Years' War (1337–1453), and ravaging popular revolts (the Jacquerie, 1358; Wat Tyler, 1381). The notion that this constituted some kind of 'waning' is misguided, however. It is only partly valid for the West, and is counterbalanced by the prosperity in Polish, Czech, Hungarian, and Serbian Central Europe. The Prague of Charles IV (1346–78) now became the center of Europe. Drawing upon the richness of the Low Countries, the splendid court of Burgundy lived out its 'epics of temerity and pride.' Despite the ravages of the plague, Italian cities continued to prosper. Following Dante (1265–1321), Giotto (1266–1337), and Petrarch (1304–74)—who saw himself as already somehow beyond the 'dark' Middle Ages—Boccaccio's *Decameron* (1350) and Chaucer's *The Canterbury Tales* (1390s) offer a rich synthesis of medieval narrative tradition. With all its crises and schisms, with Meister Eckhardt (1260–1327), St. Bridget of Sweden (1303–73), St. Catherine of Siena (1347–80), John Wyclif (1330–84), and John Hus (c. 1370–1415), Christianity was preparing for the renewal of personal and secular religiosity.

The formidable new tensions and the grand achievements of the fifteenth century represent the concluding period of this 'long Middle Ages.' The power of the emerging prophecies and religious anxieties is demonstrated by the explosion of the Hussite rebellion; Joan of Arc, who in 1429 brought about a decisive turn in the Hundred Years' War; and the beginnings of witch-hunts in Switzerland, Austria, Italy, and France. The political map of Europe was being reshaped by the emergence of powerful new states. The Ottoman conquest was only temporarily slowed down by John of Hunyad and the crusade organized by St. John Capistran at Belgrade (1456). The formation of a multinational Central European empire, attempted by the Anjous, the Luxemburgs, the Jagiellonians, and Mathias Corvinus (1458–90), was finally realized by the Habsburgs under Maximilian I (1493–1519). A large Eastern Europe re-emerged with the Polish-Lithuanian Confederation and the new Russia of Ivan III (1462–1505). In the West, an impressive new candidate for statehood, Burgundy, finally succumbed, and, besides the France of Louis XI (1461–83) and Tudor England after the Wars of the Roses, a unified Spain under Isabel of Castile (1474–1504) and Ferdinand of Aragon (1479–1516), and a miraculously ascending Portugal were the states which led the worldwide expansion of Europe in early modern times. Germany and Italy, on the other hand, contributed the largest proportion of the inventions, wealth, and culture which made this expansion possible: the invention of printing by Gutenberg (c. 1450), the banks of the Medici and the Fugger, the map of Toscanelli (1474), and the bold expedition of the Genoese sailor Columbus (1492). The narrative must end here as a new story begins in an enlarged new world.

See also: Bloch, Mark Léopold Benjamin (1886–1944); Catholicism; Christianity Origins: Primitive and 'Western' History; Feudalism; Historiography and Historical Thought: Christian Tradition; International Trade: Geographic Aspects; Islam and Gender; Marx, Karl (1818–89); Pilgrimage; Plagues and Diseases in History; Protestantism and Gender; Time, Chronology, and Periodization in History; Weber, Max (1864–1920)

Bibliography

Arnaldi G, Cavallo G (eds.) 1997 *Europa medievale e mondo Bizantino. Contatti effettivi e possibilità di studi comparati: Tavola rotonda del XVIII Congresso del CISH, Montreal, 29 agosta*. Istituto Storico Italiano per il Medioevo, Rome

Baschet J, Schmitt J-C (eds.) 1996 *L'image. Fonctions et usages des images dans l'Occident médiéval actes du 6e International Workshop on Medieval Societies, Centro Ettore Majorana*. Le Léopard d'or, Paris

Belting H 1994 (1990) *Likeness and Presence: A History of the Image Before the Era of Art*. University of Chicago Press, Chicago

Benson R L, Constable G (ed.) 1991 *Renaissance and Renewal in the Twelfth Century*. Toronto University Press, Toronto, ON

Bloch M 1989 (1939) *Feudal Society*. Routledge, London

Boutruche R 1970 *Seigneurie et féodalité au Moyen Age*. Presses universitaires de France, Paris

Branca V (ed.) 1973 *Concetto, storia, miti e immagini del Medio Evo*. Sansoni, Firenze, Italy

Braudel F 1972–73 (1949) *The Mediterranean and the Mediterranean World in the Age of Philip II*. Harper and Row, New York

Brown E 1974 The tyranny of a construct: feudalism and historians of Medieval Europe. *American Historical Review* **79**: 1063–88; also in: Little L K, Rosenwein B H (eds.) *Debating the Middle Ages: Issues and Readings*. Blackwell, Oxford, UK, pp. 148–69

Brown P 1978 *The Making of Late Antiquity*. Harvard University Press, Cambridge, MA

Brown P 1995 *The Rise of Western Christendom: Triumph and Diversity. 200–1000 AD*. Blackwell, Oxford, UK

Burckhardt J 1954 (1878) *The Civilization of the Renaissance in Italy*. [Holburn H (ed.)] Random House, New York

Bynum C W 1995 *The Resurrection of the Body in Western Christianity, 200–1336*. Columbia University Press, New York

Camille M 1989 *The Gothic Idol: Ideology and Image-making in Medieval Art*. Cambridge University Press, Cambridge, UK

Cerquiglini B 1989 *L'Eloge de la variante: Histoire critique de la philologie*. Seuil, Paris

Clanchy M T 1993 *From Memory to Written Record, England 1066–1307*. Blackwell, Oxford, UK

Davis N Z 1975 *Society and Culture in Early Modern France: Eight Essays*. Stanford University Press, Stanford,CA

Duby G 1980 (1978) *The Three Orders: Feudal Society Imagined*. University of Chicago Press, Chicago

Duby G 1981 (1966–67) *The Age of the Cathedrals: Art and Society, 980–1420*. University of Chicago Press, Chicago

Duby G, Perrot M, Pantel-Schmitt P (eds.) 1992–94 *A History of Women in the West*. Belknap Press of Harvard University Press, Cambridge, MA
Dvornik F 1949 *The Making of Central and Eastern Europe*. The Polish Research Centre Ltd., London
Eco U 1986 The return of the Middle Ages. In: Eco U (ed.) *Travels in Hyperreality: Essays*. Picador, London, pp. 59–85
Folz R 1969 *The Concept of Empire in Western Europe from the Fifth to the Fourteenth Century*. Edward Arnold, London
Folz R 1984 *Les saints rois du Moyen Age en Occident, VIe-XIIIe siècles*. Société des Bollandistes, Bruxelles
Geary P J 1988 *Before France and Germany: The Creation and Transformation of the Merovingian World*. Oxford University Press, Oxford, UK
Geremek B 1994 (1987) *Poverty: A History*. Blackwell, Oxford, UK
Geremek B 1996 (1991) *The Common Roots of Europe*. Polity Press, Cambridge, UK
Gieysztor A 1997 *L'Europe nouvelle autour de l'An Mil. La Papauté, l'Empire et les 'nouveaux venus'*. Accademia dei Lincei, Rome
Ginzburg C 1980 (1976) *The Cheese and the Worms: The Cosmos of a Sixteenth-century Miller*. Johns Hopkins University Press, Baltimore, MD
Goody J 1987 *The Interface Between the Written and the Oral*. Cambridge University Press, Cambridge, UK
Gurevich A I 1995 *The Origins of European Individualism*. Blackwell, Oxford, UK
Halecki O 1952 *The Borderlands of Western Civilization: A History of East Central Europe*. Ronald Press, New York
Huizinga J 1996 (1919) *The Autumn of the Middle Ages*. University of Chicago Press, Chicago
Klaniczay G 1990 *The Uses of Supernatural Power: The Transformation of Popular Religion in Medieval and Early-modern Europe*. Polity Press, Cambridge, UK
Le Goff J 1988 (1985) *The Medieval Imagination*. University of Chicago Press, Chicago
Le Goff J (ed.) 1990 (1989) *Medieval Callings*. University of Chicago Press, Chicago
Le Goff J 1994 (1972) *Medieval Civilization, 400–1500*. Blackwell, Oxford, UK
Le Goff J 1999 *Un autre Moyen Age*. Gallimard, Paris
Le Goff J, Schmitt J-C (eds.) 1999 *Dictionnaire raisonné de l'Occident médiéval*. Fayard, Paris
Le Roy Ladurie E 1979 (1975) *Montaillou. The Promised Land of Error*. Vintage Books, New York
Obolensky D 1982 *The Byzantine Commonwealth. Eastern Europe, 500–1453*. St. Vladimir's Seminary Press, Crestwood, NY
Ostrogorski G 1969 *History of the Byzantine State*. Rutgers University Press, New Brunswick, NJ
Patlagean E 1988 *Europe, seigneurie, féodalité: Marc Bloch et les limites orientales d'un espace de comparison*. Centro Italiano di Studi sull'alto Medioevo, Spoleto, Italy
Petrucci A 1992 *Medioevo da leggere: guida allo studio delle testimonianze scritte del Medioevo Italiano*. Einaudi, Torino, Italy
Pirenne H 1937 (1933) *Economic and Social History of Medieval Europe*. Harcourt Brace, New York
Pirenne H 1965 (1937) *Mohammed and Charlemagne*. The World Publishing Company, Cleveland, OH, New York
Poly J-P, Bournazel E 1991 *The Feudal Transformation: 900–1200*. Holmes and Meier, New York
Reynolds S 1994 *Fiefs and Vassals: The Medieval Evidence Reinterpreted*. Oxford University Press, Oxford, UK
Sergi G 1998 *L'idée de Moyen Age. Entre sens commun et pratique historique*. Flammarion, Paris
Southern R W 1990 (1970) *Western Society and the Church in the Middle Ages*. Penguin, London
Stock B 1983 *The Implications of Literacy. Written Language and Models of Interpretation in the Eleventh and Twelfth Centuries*. Princeton University Press, Princeton, NJ
Szücs J 1990 (1981) *Die drei historischen Regionen Europas*. Verlag Neue Kritik, Frankfurt am Main
Vauchez A 1997 (1981) *Sainthood in the Later Middle Ages*. Cambridge University Press, Cambridge, UK

G. Klaniczay

Middle East and North Africa: Sociocultural Aspects

1. *Representations and Misrepresentations*

The term, 'The Middle East and North Africa,' refers to a territory characterized by considerable diversity, extending from Morocco in the West to Iran in the East, spread over Africa and Asia and extending into Europe. This is a region that has given rise to the earliest forms of urban life and state organization in human civilization, produced three major monotheistic religions (Judaism, Christianity, and Islam), includes three major language groups (Arabic, Persian, and Turkish) as well as countless other languages and dialects, and exhibits sharp contrasts in physical and human geography. Since the eighth century AD, successive Islamic empires have ruled and to some extent united this geographical expanse, the most recent of which was the Ottoman. In other words, this is an area of great complexity, with its long historical record, literate traditions, and mixture of social and cultural groups.

In the parlance of early anthropology, this was regarded as a 'culture area,' and from an area studies perspective it was seen as a region broadly sharing linguistic, religious, and cultural characteristics as well as a common geopolitical position in the world order. Yet, because of its complexity and diversity, the term and the territory associated with it are coupled in a shifting, unstable relationship. Afghanistan is sometimes included, sometimes not. The Sudan falls in or out depending on whether the emphasis is on the African heterogeneous south or the Arab Muslim north. Turkey is usually included, although many of its intellectuals and politicians prefer to find continuities with Europe and the Balkans, and/or with Central Asia.

The first part of the term, the Middle East, originates from nineteenth-century European, specifically British, strategic military divisions of the world. Other terms also compete for the same territories or carve it up in different ways: older terms such as the Near East or the Levant still have occasional currency, the latter having been recently proposed as a substitute term that emphasizes linkages, rather than boundaries, between the societies and cultures of the Greeks, Turks, Arabs, and Jews (Alcalay 1993). Similar to this are arguments for the Mediterranean as a cultural unit. The term 'the Arab world' is often seen, especially indigenously, as more accurate in depicting an analytically coherent unit. In Arabic, the historical distinction between the Maghreb (the West, or North African countries) and the Mashreq (the East, or the Eastern Mediterranean countries) is widely used and this is reflected in French scholarship on the region, as exemplified by the journal *Maghreb-Machrek*. In German, the term 'the Orient' still largely refers to Islamic and Arab societies and the disciplines and learning related to them.

New regionalisms also compete with these designations in different ways. The renewal of historical and religious connections with Central Asia and the Caucasus, after a rupture of more than a century of Russian and Soviet rule, gives rise to new/old redrawing of the boundaries of the region. The freeing of scholarly imaginations by theories of globalization leads to the investigation of historical and contemporary links across the Indian Ocean with South Asia and Southeast Asia. Finally, studies of diaspora communities (e.g., Turks in Germany, the Lebanese in South America and Australia, the North Africans in Europe, Arab Americans, and so on) are beginning to look at translocal and transnational linkages that are also part of the making and unmaking of the region.

What these examples of shifting boundaries and their referents illustrate is the contested reasoning behind the proposed unity of the region. At the heart of this contest lies the role of Islam perceived as a unifying and even homogenizing force. Despite the demographic reality that the largest numbers of Muslims do not live in the Middle East, the region seems inextricably linked with Islam as its historical fount and heartland. The ways in which some scholars now subsume Central Asia within the Middle East, on the basis of the 'Islamic nature' of its societies, shows how Islam is considered as a unifying force that brings together societies and histories with widely disparate historical experiences. Yet, the question of whether Islam, or any religion, can be seen as providing the sociocultural cohesion of any geographical region remains a problematic issue, as does the question of where the cohesive quality of Islam is to be located. Islam as a belief system? Or an institutional framework? Or historical experience? The different answers lead towards different routes of comparisons and distinctions (Asad 1986).

This identification with a unified Islamic essence also led to an enduring interpretation of the region through dichotomous notions of East and West. Edward Said's seminal work, *Orientalism* (1978), is the most prominent discussion of the relations of knowledge and power that accompanies this dichotomy and their implications. Other works have also carefully examined when the boundary between East and West was discursively set and how it shifted geographically throughout history, sometimes including the Balkans and Greece and sometimes not (Todorova 1997). The question of where the Near East begins inexplicably continues to bedevil some contemporary popular writers (e.g., Kaplan 2000), who see this boundary as explanatory of a range of phenomena from wars to political formations to economic structures to lifestyles and fashions.

How the region was imagined and hence 'made' brings together a number of different actors besides the Great Powers and their strategic interests. The 'sand-mad' English men and women, like Sir Richard Burton, Charles Doughty, and Gertrude Bell, who traveled the region as explorers and spies, wrote accounts that, alongside the textual and philological studies of the Orientalists, became part of the discursive universe called the Orient and fired European and American imaginations. Some of these nineteenth-century accounts were highly ethnographic, as in the case of Edward Lane's classic *An Account of the Manners and Customs of the Modern Egyptians* (1836) and W. Robertson Smith's *Kinship and Marriage in Early Arabia* (1885). The French colonies in North Africa were a particular destination for colonial ethnographers. Missionaries, among whom Americans figured largely, were another source of information that shaped the contours of this world for the Western imagination. Finally, novelists from Flaubert to Mark Twain took part in filling in the details of the imagined Orient as the region opened up to Western infiltration and then domination with the gradual demise of the Ottoman Empire.

However neutral the use of the term 'the Middle East' in contemporary scholarship, it is important not to deny the power of the images of the East, the Orient, and the world of Islam in which this scholarship is rooted.

2. *Meta-narratives and Their Critiques*

Within the framework of area studies, scholarship on the region was marked, as in other parts of the world, by modernization theory and 'developmentalism.' Following World War I and the collapse of the Ottoman Empire, various states were carved out by European colonial powers and came under their direct rule, with a few notable exceptions such as Turkey and

Iran. Most of the colonial states became independent after World War II, again with a few notable exceptions such as Algeria, which gained independence only in 1962. The region held a prominent place in illustrating the debates of the time on emergent states and nation building through such studies as Daniel Lerner's *The Passing of Traditional Society* (1958). This makes the current eclipse of the region in theoretical and comparative works, particularly in political science, and its perceived 'exceptionalism' from global currents of liberalization and democratization, all the more noticeable.

Another founding text, especially for anthropologists, was Carleton Coon's *Caravan: The Story of the Middle East* (1951). This powerful text, which has been critiqued at many levels, provided a long-enduring framework, explanation, and diagnosis for the region through the concept of the Middle East as a 'mosaic society.' It was instrumental in creating a vision of the Middle East as a region united by Islam (as belief, law, and ritual), made interdependent by ecology and an ethnic division of labor, and with its inherent disorder kept at bay through successive authoritarian empires and states. Ethnicity (loosely deployed to refer to all sorts of groups whether linguistic, religious, sectarian, or occupational) was seen as a primary divisive factor, offset by a traditional market exchange economy and a strong state that lends society a rigid stability.

However flawed and ahistorical its narrative, *Caravan* was the perfect textbook, presenting a clear story-line (as the subtitle indicates) but permitting digressions and discrepancies. Many of the textbooks, reviews, and encyclopedia articles that followed the publication of this work employed its handy tripartite division of the region into ecological zones of pastoral/rural/urban, each with its economic, social, and cultural traits. Islam—as a 'high' tradition—was largely left to the purview of scholars trained in Orientalist studies with their knowledge of language and written texts; so too, to a great extent, was the study of the city and of urbanism as the crucibles of this high culture. However, ethnographic documentation of 'traditional society' did include the study of local forms of worship and ritual, focusing mainly on rural and pastoral groups. The study of social change focused largely on the transformation of rural economies, as in the influential compendium by Antoun and Harik (1972). Anthropologists generally paid little attention to national politics and modes of political participation, although perhaps noting the 'accommodation' of traditional social structures, and more importantly cultural values, to modernization.

Alongside this, however, was a number of works and circles of scholars who were gradually transforming the notions governing the understanding of Middle Eastern societies, and also more generally of many of the basic tenets of anthropology and social science. One strand of literature developed the critical approach to colonialism, neoimperialism and knowledge within which Said's *Orientalism* can be situated. This included works by Abdullah Laroui, Anouar Abdul-Malek, and Talal Asad, as well as the contributors to journals such as the *Review of Middle East Studies* and *Khamsin* ('the desert wind that scorches'), which were published in England.

A second strand, sometimes featuring the same authors, focused on political economy and modes of production, well demonstrated by the US journal *MERIP* (*Middle East Research and Information Project*) in its early years. Works by authors like Samir Amin brought the region into central debates on dependency and unequal development. In this vein were studies of the oil economy, which was gaining importance in the 1970s, focusing on the structure of oil-based rentier states as well as the regional and international migration of labor. This focus was well placed. Labor markets within the Middle East, and migration more generally, worked to integrate the region in ways that states could not. Integration, however, did not mean the easy accessibility of different societies to one another but, rather, painful competition over economic and social goods, discoveries of cultural disjuncture, the rise of racist stereotyping and the creation of entirely new hierarchies of status and power. In addition, the impact of migration and of remittances was wide ranging, which changed household structures, local economic patterns, authority relations, and perceptions of identity.

A third strand of scholarship, building upon the first two, set about revising the received wisdom on the Ottoman and colonial periods, and focused on the incorporation of the region into the world capitalist system as well as on nationalism and ideological currents in the area. Turkish scholars were particularly productive in this literature, which also helped encourage the turn to social history, local history, and the study of class and family.

In anthropology, scholars working on the region now included such distinguished names as Pierre Bourdieu, Ernest Gellner, and Clifford Geertz. In their very different ways, these theorists laid the foundations for new departures in scholarship through an emphasis on social relations and cultural symbols, a focus on field research, and an appreciation of the diversity of historical and contemporary sociopolitical formations. Work by Geertz' students particularly challenged formal notions of social structure, kinship, and neighborhood by focusing instead on the fluid, transactional, and negotiated nature of the ties that form the basis of social solidarity and networks. Morocco particularly became the site for early explorations in reflexivity, intersubjectivity and the questioning of long-held notions of ethnographic objectivity, as seen in works by Vincent Crapanzano and Paul Rabinow.

It is worth noting that these anthropological works are all on North Africa and, with the exception of

Bourdieu's work on Algeria, focus on Morocco, and indeed on particular towns and locations. This highlights the fact that generalizations about the region tend to be based on the accumulation of knowledge from very specific sites. Countries that act as especial poles of attraction for anthropologists are Morocco, Yemen, and Egypt as well as Iran before the Islamic revolution. Another privileged site of research is Israel, which has attracted a fair number of scholars in addition to having its own well-developed research community. However, scholars working on Israel tend not to be in dialogue with scholars of the rest of the Middle East. This is not merely a reflection of actual tensions in the region but is also due to the fact that these scholars often adopt frames of analysis for Israel that are very different from those developed for interpreting the 'Islamic' Middle East. For example, only in the literature on Israel do we find the serious study of ethnicity. Yet much of this work is conducted within US-inspired frameworks with the focus being on the immigrant nation, the melting pot, and the frontier society. Furthermore, this work remains largely focused on relations between various Jewish groups with hardly a mention of Arabs, whether Muslim, Christian, or Druze, within Israeli society.

Excluding these sites, the literature on all other countries of the region, including even Turkey, was sporadic until recently, spread out in time and topical focus. Thus it did not contribute to the creation of schools of thought or to central concerns in theory and methodology. It is also interesting to note, that even in the 1980s, overview courses of the Middle East showed a significant time lag from actual scholarly production and did not abandon their tripartite schema of nomads/villagers/urban dwellers, except for the occasional addition of a unit on women, sometimes entitled 'women and the moral order,' apparently thus concerned with putting women in their place. Yet this rider onto the course outline is a harbinger of important changes that, over the past two decades, may be finally putting to rest the master narrative, the 'story' of the Middle East.

3. *Directions, Trends, and the Resurgence of Islam*

With the availability of new frames of interpretation, researchers on the region have begun increasingly to work within cultural, feminist, and postcolonial perspectives. Current research shows a healthy diversity, with interests varying from oral history to identity, to health practices, to urban politics, to everyday life. However, there appears to be a too sanguine assumption that the ghosts of Orientalism have been laid to rest. Works that have explicitly met the challenge of examining and historicizing the politics of representation are rare, the most prominent being Timothy Mitchell's *Colonizing Egypt* (1988).

It is important to draw attention to scholarship in the languages of the region itself, which are often marginalized in Western scholarship. These works have followed global theoretical trends but also exhibit particular interest in issues of national identity, political and economic underdevelopment, intellectual heritage, and the rewriting of national histories, as exemplified in works by Mohamed Arkoun, Abdullah al-Jabiri, and George Tarabishi, to name but a few. A powerful trend of 'indigenization' represents an important experimental moment in this scholarship. In the 1980s, the slew of writings on the 'Arabization' of the social sciences, the possibility of alternative epistemological frameworks, and the challenge to universalist notions of science presented a formidable task for many social scientists in the region. It is important to note that these discursive trends were empowered through funding, institutional support, and channels of dissemination, such as that provided by the Center for Arab Unity Studies in Beirut, Lebanon. Interestingly, a similar energy, but bolstered by much larger resources and global rather than regional networks, now calls for the Islamization of the social sciences. This literature can be traced from the works in the 1970s of the Iranian scholar, Ali Shariati (see Algar 1979), to current Egyptian and South Asian thinkers who are linked in ways previously not possible in the East/West division of knowledge and power.

Islam is once again a central concern of scholars working on the region and now stands firmly within the purview of anthropology as well. The Iranian revolution and the rise of Islamist movements both across the region and globally present particular challenges to the understanding of sociocultural change. Often this literature seems driven by a sense of political urgency or, as in the case of European scholars of immigration, by a sense of impending doom. While what constitutes the object of study in an anthropology of Islam has continued to be a problematic issue, as discussed by Asad (1986) and Gilsenan (1982), various works usefully look at the practice, significance, and power of Islam in different settings, such as education, village politics, expressive culture, or pilgrimage.

The study of gender is one area where tensions with past representations have worked to help transcend old debates and open new horizons. Whereas exotic notions of femininity were at the heart of Orientalist constructions of the Orient, critical approaches to gender have worked to analyze, subvert, and transcend these representations. Gender and gender inequality, have been studied in a multitude of ways: as colonial experience, as underdevelopment, as identity, as representation, as memory, and as embodied practice. This literature, which is growing at a great pace and now includes research on masculinity and sexuality, is among the topics most in dialogue with theoretical

writings in anthropology specifically and the social sciences generally. It provides routes and models that other research topics, similarly fraught with long and contested genealogies, could usefully follow (Abu-Lughod 1989, 1993).

Clear trends within the region include such global processes as structural adjustment, the opening up to world markets in unprecedented ways, and the intervention of international organizations into the microcosms of the family, the neighborhood, and local authorities. Migration and displacement continue to frame the lives of a majority of the population of the region, directly and indirectly. Refugee studies is now a growing field, given that the region has produced as well as hosted some of the largest refugee populations in the world. Large numbers of peoples in the region, most famously the Palestinians, conceive of their identity through collective memories centering on dreams of a lost land (Said 1986).

The Palestinian case also shows clearly the intimate intersections of a multitude of local, national, and global levels as refugee camps, dispersed refugees, militias, governments, international bodies, and *ad hoc* alliances all function to govern and determine the fates of several million people. Less studied are diaspora populations—as one important link between nationalism and transnationalism, between the global, the national, and the local. The Iranian case is one exception where the emergence of diasporic public spheres have been studied, as well as the creation of ethnic virtual realities embedded in concrete social and economic exchanges. Los Angeles, or Irangeles, forms a crucial node in the relations that link Iranian communities in many European cities, in India, and in Japan (Naficy 1993).

Finally, one promising trend is the increasing number of works attempting to locate the appropriate intersection of history and ethnography that can interpret the layers of complexity in Middle Eastern societies. Past and present worlds mirror one another through connections and disconnections (Ghosh 1992), while the objects of memory range from village homes to artwork to oral narratives (Slymovics 1998). Furthermore, the growing focus on the construction of local social worlds, of everyday life, of popular culture, and of interpersonal relations opens up the scope of inquiry to bring out the multiple sites of modernity in the Middle East.

See also: Gellner, Ernest (1925–95); Islam: Middle East; Nationalism, Historical Aspects of: Arab World; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Economics; Near Middle East/North African Studies: Gender; Near Middle East/North African Studies: Geography; Near Middle East/North African Studies: Politics; Near Middle East/North African Studies: Religion; Near Middle East/North African Studies: Society and History; Orientalism

Bibliography

Abu-Lughod L 1989 Zones of theory in the anthropology of the Arab world. *Annual Review of Anthropology* **18**: 267–306

Abu-Lughod L 1993 *Writing Women's Worlds: Bedouin Stories*. University of California Press, Berkeley, CA

Alcalay A 1993 *After Jews and Arabs: Remaking Levantine Culture*. University of Minnesota Press, Minneapolis, MN

Shari'ati A 1979 *On the Sociology of Islam: Lectures by Ali Shari'ati*. Mizan Press, Berkeley, CA

Antoun R, Harik I (eds.) 1972 *Rural Politics and Social Change in the Middle East*. Indiana University Press, Bloomington, IN

Asad T 1973 *Anthropology & the Colonial Encounter*. Ithaca Press, New York

Asad T 1986 *The Idea of an Anthropology of Islam*. Occasional Paper Series, Center for Contemporary Arab Studies. Georgetown University Press, Washington, DC

Coon C 1951 *Caravan: The Story of the Middle East*. Holt, Rinehart and Winston, New York

Ghosh A 1992 *In an Antique Land*. Ravi Dayal Pub, New Delhi

Gilsenan M 1982 *Recognizing Islam: Religion and Society in the Modern Arab World*. Pantheon Books, New York

Kaplan R 2000 *Eastward to Tartary: Travels in the Balkans, the Middle East and the Caucasus*. Random House, New York

Lerner D 1958 *The Passing of Traditional Society: Modernizing the Middle East*. Free Press, Glencoe, Ill

Middle East Report 1997 *Middle East Studies Networks: The Politics of a Field* **27**(4), no. 205

Mitchell T 1988 *Colonising Egypt*. Cambridge University Press, Cambridge, UK

Naficy H 1993 *Making of Exile Cultures: Iranian Television in Los Angeles*. University of Minnesota Press, Minneapolis, MN

Said E W 1978 *Orientalism*. Pantheon Books, New York

Said E W 1986 *After the Last Sky: Palestinian Lives*. Faber and Faber, London

Slyomovics S 1998 *The Object of Memory: Arab and Jew Narrate the Palestinian Village*. University of Pennsylvania Press, Philadelphia, PA

Todorova M 1997 *Imagining the Balkans*. Oxford University Press, New York

S. Shami

Midlife Psychological Development

The timing of the midlife period has been the subject of a great deal of controversy. The modal responses to survey questions suggest that midlife begins at 40 and ends at 60 years of age (Lachman et al. 1994). However, there is great variability in the responses, with the age range for midlife typically from 30 to 75. Moreover, the ages of entry and exit are positively correlated with age, such that the older one's present age the later the expected timing of midlife. For example, on average, those in their twenties report that midlife begins at 30, whereas those in their seventies often include themselves within the period of

midlife (Lachman et al. 1994). This is particularly salient because research indicates that one's subjective conception of age is a better indicator of well-being and functioning than chronological age.

The middle years are an important and central part of the lifespan, yet there has been little research focused directly on this period (Brim 1992). Those in midlife are often in leadership positions; thus, the well-being of middle-aged adults has an impact on the welfare of those who are both younger and older. Middle-aged adults may have responsibility for mentoring younger workers and taking care of multiple generations of family members. While raising their own children, middle-aged adults may also need to provide care for their own parents, if they are widowed, disabled, or frail (Moen and Wethington 1999).

1. *Perspectives on Midlife*

It is useful to examine the midlife period within the context of the entire life span (Baltes et al. 1997). The nature of psychological functioning in the middle years is likely to be more meaningful when considered relative to what has come before and what lies ahead. On most dimensions, middle-aged adults may fall somewhere in between younger and older adults. However, in some domains, middle-aged adults may function more like those who are younger, and in other domains they may be more similar to older adults.

A life-span perspective suggests that the ratio of gains and losses changes during adulthood (Baltes et al. 1997). In young adulthood the gains outnumber the losses. In midlife, the tendency is toward a balance of gains and losses. This balance is tipped in later life, when the number of losses may exceed the potential gains. As this transition occurs, the process of adapting to losses, some of which are uncontrollable, is important to well-being. As adults age they begin to use more compensatory strategies, such as reducing the number of demands or responsibilities (Brim 1992). Brandstädter and Renner (1990) found that during middle age and beyond, adults were more inclined to use accommodative strategies, that is changing or even giving up one's goals to be more consistent with one's abilities or revised priorities. In contrast, those in young adulthood more frequently use assimilative strategies, taking on more goals and persisting in the face of obstacles. These adaptive processes can be applied to the physical, psychological, and social changes that occur during the middle years.

2. *Theories of Development in Midlife*

There are a number of stage theories that include the midlife period (Erikson 1963, Levinson et al. 1978). According to Erikson (1963), the key focus of midlife involves the life task of generativity vs. stagnation. Successful negotiation of this stage entails focusing one's efforts on the younger generation. In midlife one is able to guide the next generation by sharing knowledge and wisdom with younger co-workers and family members. The middle-aged adult serves important roles in the family and the workplace. They are often guides for the younger and older generation, with an ability to draw on their experience in multiple domains (McAdams 2001). According to Erikson, if adults are not successful in achieving generativity, they may be considered stagnant or nonproductive because a key task for midlife is to transmit knowledge and values to the next generation.

3. *Physical and Psychological Changes*

One of the hallmarks of midlife is that it is a time of peak performance in a variety of psychological domains, while at the same time there are declines in physical and some aspects of cognitive functioning. The midlife period is accompanied by physical changes such as changes in sensory functioning and hormonal changes (Avis 1999). Although some aspects of physical functioning (e.g., lung function, reaction time) or cognitive performance (e.g., working memory) typically show declines by midlife, many cognitive functions are at their peak (e.g., reasoning) or still improving (e.g., knowledge) (Miller and Lachman 2000, Willis and Schaie 1999). Some functions may have begun to decline, such as speed of processing or working memory (Dixon et al. 2001), while other aspects are increasing, such as verbal ability and wisdom (Miller and Lachman 2001, Staudinger et al. 1992). Other abilities peak at midlife, including problem solving and reasoning abilities (Willis and Schaie 1999). Some psychosocial variables such as positive mood (Mroczek and Kolarz 1998), self-confidence, and sense of mastery also appear to peak in midlife (Lachman et al. 1994, Neugarten 1968).

4. *Multiple Roles and Transitions*

A number of role transitions typically occur during the middle years. Middle-aged parents are said to experience the empty nest syndrome when their adult children leave home. Although some have assumed this would be a traumatic time, especially for women, research shows that this is often a welcome period when middle-aged adults pursue new interests, enjoy the freedom from raising children, and often strengthen the intimacy of the marital bond (Paul 1997).

The menopause for women marks a major shift in biological functioning, with changes in hormone levels

and reproductive ability. Recent medical developments, however, may make it more common for postmenopausal women to bear children. There are wide individual differences in the experiences of menopause. Physiological symptoms such as hot flushes or sleeplessness are not universal. (Avis 1999). There are a number of myths associated with the menopause, including that it is accompanied by depression, mood swings, and fatigue. In fact, research shows that the menopause is generally not a negative experience, but is often associated with a sense of relief that menstruation is over and a welcomed freedom from concern about pregnancy (Avis 1999).

5. *Personality and the Self in Midlife*

Personality traits are relatively stable throughout adulthood (Costa and McCrae 1980). These enduring characteristics play a major role in shaping the course of development. Some aspects of the self are more malleable and influenced by experience, making the person resilient or vulnerable to stress. Thus, the self-concept can serve as a resource or a risk factor. Self-esteem, self-efficacy, and the sense of control affect choice and selection of goals as well as the amount of effort expended to accomplish them in midlife (Bandura 1997). There are some areas in which the sense of control increases in later life, such as over work, marriage, and finances (Lachman and Weaver 1998). In other domains, including memory, sex life, and relationship with children, the sense of control declines with age. The sense of control also contributes to health (Lachman and Weaver 1998). Those who have a greater sense of mastery and lower perceived constraints on their life have better health and functional status. Those who believe they have more control over health are more likely to engage in health-promoting behaviors including exercising and eating a healthy diet.

Sex role characteristics also shift during the middle years. In early adulthood men are more agentic and women are more communal. However, there is evidence that during the middle years both genders become less sex-stereotyped and adopt more of the opposite-sex characteristics, but they do not necessarily lose their sex-typed characteristics. Thus, both genders appear to develop more integrated and androgenous characteristics (Parker and Aldwin 1997, James and Lewkowicz 1997).

6. *The Midlife Crisis*

Perhaps the most ubiquitous association with midlife is the midlife crisis. The media and popular literature have portrayed midlife as a period involving inner turmoil, regret about growing older, disappointment about one's marriage or career, and a sense of loss when children leave home (Brim 1992). However, there is little evidence in support of a universal midlife crisis of this nature. Indeed there are some who experience distress, but they tend to be those who have had tumultuous periods throughout their lives. Those who have more neurotic personality styles are the ones who are more likely to experience a midlife crisis (Costa and McCrae 1980, Whitbourne and Connolly 1999). Even among those who do have a crisis, the nature of the midlife crisis also varies considerably. For some it is a reappraisal of life, for others it is a fear of getting older, or it may be a realization that one has not fulfilled one's goals (Rosenberg et al. 1999). Rather than being an inevitable part of middle age (Levinson et al. 1978), empirical evidence supports the view that a crisis in midlife is more a function of stable personality characteristics or coping styles than of a particular age period (Costa and McCrae 1980, Whitbourne and Connolly 1999). People who have crises during midlife are likely to have them at other ages or transitions as well.

7. *Future Directions*

More research is needed to understand the nature of the midlife period. It is important to understand how behaviors and patterns laid down in the early and middle years of adulthood impact the nature and course of aging. Clearly, if midlife serves as a window on aging, there may be early warning signs apparent in midlife that may be important to identify for prevention efforts. Mental health and physical well-being in adulthood is determined by multiple factors, many of which are under one's control. Lifestyle factors involving social relationships, diet, exercise, mental stimulation, smoking, and alcohol intake all have a bearing on one's quality of life (Rowe and Kahn 1997). Well-being in midlife is an important area of study because the outcome impacts multiple generations who are touched by middle-aged adults in the family and the workplace. The nature of midlife functioning and behavior may be a key indicator of the length and quality of life in the later years.

See also: Adult Cognitive Development: Post-Piagetian Perspectives; Adult Development, Psychology of; Adulthood: Developmental Tasks and Critical Life Events; Coping across the Lifespan; Life Course in History; Life Course: Sociological Aspects; Lifespan Development, Theory of; Personality Development in Childhood; Plasticity in Human Behavior across the Lifespan; Self-regulation in Adulthood

Bibliography

Avis N E 1999 Women's health at midlife. In: Willis S L, Reid J D (eds.) *Life in the Middle: Psychological and Social*

Development in Middle Age. Academic Press, San Diego, CA, pp. 105–46
Baltes P B, Lindenberger U, Staudinger U M 1997 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology: Vol. 1. Theoretical Models of Human Development*, 5th edn. Wiley, New York, pp. 1029–43
Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman, New York
Brandstädter J, Renner G 1990 Tenancious goal pursuit and flexible goal adjustment: Explication and age-related analysis of assimilative and accommodative strategies of coping. *Psychology and Aging* **5**: 58–67
Brim G 1992 *Ambition: How We Manage Success and Failure Throughout our Lives*. Basic Books, New York
Costa P T, McCrae R R 1980 Still stable after all these years: Personality as a key to some issues in adulthod and old age. In:Baltes P B, Brim O G, Jr. (eds) *Life-span Development and Behavior*. Academic Press, New York, Vol. 3, pp. 65–102
Dixon R, de Frias C M, Maitland S B 2001 Memory in midlife. In: Lachman M E (ed.) *Handbook of Midlife Development*. Wiley, New York, pp. 248–78
Erikson E H 1963 *Childhood and Society*, 2nd edn. W W Norton, New York
James J B, Lewkowicz C J 1997 Themes of power and affiliation across time. In: Lachman M E, James J B (eds.) *Multiple Paths of Midlife Development*. University of Chicago Press, Chicago, pp. 109–43
Lachman M E, James J B 1997 Charting the course of midlife development: An overview. In: Lachman M E, James J B (eds.) *Multiple Paths of Midlife Development*. University of Chicago Press, Chicago, pp. 1–17
Lachman M E, Lewkowicz C, Marcus A, Peng Y 1994 Images of midlife development among young, middle-aged, and elderly adults. *Journal of Adult Development* **1**: 203–11
Lachman M E, Weaver S L 1998 Sociodemographic variations in the sense of control by domain: Findings from the MacArthur studies of midlife. *Psychology and Aging* **13**: 553–62
Levinson D J, Darrow C N, Klein E B, Levinson M H, McKee B 1978 *The Seasons of a Man's Life*. Knopf, New York
McAdams D P 2001 Generativity in midlife. In: Lachman M E (ed.) *Handbook of Midlife Development*. Wiley, New York, pp. 395–443
Miller L S, Lachman M E 2000 Cognitive performance in midlife and the role of control beliefs. *Neuropsychology, Cognition, and Aging* **7**: 69–85
Moen P, Wethington E 1999 Midlife development in life course context. In: Willis S L, Reid J D (eds.) *Life in the Middle: Psychological and Social Development in Middle Age*. Academic Press, San Diego, CA, pp. 3–24
Mroczek D K, Kolarz C M 1998 The effect of age on positive and negative affect: A developmental perspective on happiness. *Journal of Personality and Social Psychology* **7**: 1333–49
Neugarten B L 1968 The awareness of middle age. In: Neugarten B L (ed.) *Middle Age and Aging*. University of Chicago Press, Chicago, pp. 93–8
Parker R, Aldwin C 1997 Do aspects of gender identity change from early to middle adulthood? Disentangling age, cohort, and period effects. In: Lachman M E, James J B (eds.) *Multiple Paths of Midlife Development*. University of Chicago Press, Chicago, pp. 67–107
Paul E 1997 A longitudinal analysis of midlife interpersonal relationships and well-being. In: Lachman M E, James J B (eds.) *Multiple Paths of Midlife Development*. University of Chicago Press, Chicago, pp. 171–206
Rosenberg S D, Rosenberg H, Farrell M P 1999 The midlife crisis revisited. In: Willis S L, Reid J D (eds.) *Life in the Middle: Psychological and Social Development in Middle Age*. Academic Press, San Diego, CA, pp. 47–73
Rowe J W, Kahn R L 1997 Successful aging. *The Gerontologist* **37**: 433–40
Staudinger U M, Smith J, Baltes P B 1992 Wisdom-related knowledge in a life-review task: Age differences and the role of professional specialization. *Psychology and Aging* **7**: 271–81
Whitbourne S K, Connolly L A 1999 The developing self in midlife. In: Willis S L, Reid J D (eds.) *Life in the Middle: Psychological and Social Development in Middle Age*. Academic Press, San Diego, CA, pp. 25–45
Willis S L, Schaie K W 1999 Intellectual functioning in midlife. In: Willis S L, Reid J D (eds.) *Life in the Middle: Psychological and Social Development in Middle Age*. Academic Press, San Diego, CA, pp. 233–47

M. E. Lachman

Migration and Health

Migration can impact health in several ways, having consequences for places of origin, and places of destination, as well as for the migrants themselves. Although most migration is internal in that it takes place within countries (e.g., rural to urban migration), researchers have shown greater interest in international migration and specifically how it impacts the health of immigrants. It is recognized that migration can have both negative and positive consequences on health yet most research has focused primarily on the health problems of immigrants.

1. *Historical Development of the Field*

Demographers and social scientists have always been interested in migration, one of the major demographic processes which, along with fertility and mortality, determines population size, composition, and distribution (Pol and Thomas 1992) However, until relatively recently, there has been little interest in how migration impacts health, except perhaps in the area of mental health and psychological well-being. Early research, for example, had suggested that immigration to North America had negative consequences on the mental health of immigrants (see Friis et al. 1998).

Interest in the impact of migration on health has increased over the latter part of the twentieth century.

Although most migration in the world today takes place within countries, there has been little interest in its impact on health primarily because it is not typically as disruptive as international migration. Even in the case of international migration, there has traditionally been little interest in the health of immigrants. Most health research has been conducted in Western societies where the focus has been on the health status and healthcare needs of native populations. Immigrants were either excluded or were too few, both in number and as a proportion of study samples, to yield reliable estimates of their health status and health needs.

Because the late twentieth century has seen increasing immigration to Western societies from poorer nations, scholars and policy makers have made a concerted effort to understand the health status and health care needs of rising numbers of immigrants and how they impact the host countries' health and social service systems (Eschbach et al. 1999, Friis et al. 1998). There has also been some interest in the negative impact of dislocation and forced migration as well as in the impact of smuggling and trafficking of migrants on their health (Gushulak and MacPhearson 2000).

2. *Methodological Issues*

Friis et al. (1998) outlined a number of methodological issues in the study of how migration influences health. They argue that because migration results in changes in numerous environmental variables that cannot be adequately controlled, studies often reach mistaken conclusions about how migration influences health. They cite Kasl and Berkman's (1983) formulation of the ideal study design that would involve migrants-to-be, non-migrants at the point of origin, and native residents at the point of destination. The health status and psychosocial characteristics of the three groups would be assessed initially and at several follow-up contacts. In addition, studies would need large sample sizes that would enable sufficient power to adjust for numerous confounders so that the impact of migration on health can be appropriately estimated.

Needless to say such studies are scarce with selection factors, type of migration, and differences in historical time and context often leading to biased conclusions.

3. *Healthy Migrants*

Despite the focus of recent research on negative aspects of migration, a number of studies have suggested that immigrants to Western societies appear to be as healthy as if not healthier than native populations. Markides and Coreil (1986) suggested such a 'healthy migrant' or 'migration selection' effect to explain the relatively low mortality rates and good health of Mexican Americans in the Southwestern United States. The term 'epidemiologic paradox' was applied because Mexican Americans shared similar socioeconomic characteristics and conditions with African Americans, yet their mortality and health conditions were similar to the more advantaged non-Hispanic Whites. More recently, Hummer et al. (1999) undertook a comprehensive analysis of how race-/ethnicity and nativity are associated with mortality in the US population. They found consistently lower mortality rates among foreign-born persons in all major ethnic groups (Blacks, Hispanics, and persons of Asian origin) than among native-born persons. Interestingly, foreign-born Blacks along with persons of Asian origin had the lowest odds of mortality, with native-born Blacks having the highest odds of mortality.

Similar results were found with respect to other health status indicators by Stephen et al. (1994). Again, foreign-born persons were generally healthier than native-born Americans at all ages, in both genders and from all major ethnic origins. Such findings are not restricted to the United States. For example, Chen et al. (1996) found that immigrants to Canada, especially recent immigrants, have lower rates of chronic conditions and disabilities than native-born Canadians. Similar patterns have also been found in Australia (Donavan et al. 1992).

The above and other similar studies conclude that there is a 'healthy migrant' effect at work. Healthy people are more prone to immigrate than less healthy people. In addition, Western countries require medical screenings by prospective immigrants to ensure relatively good health, and most people immigrate for employment or occupational reasons which require a relatively good level of health. Finally, people who immigrate tend to have a positive outlook on their lives and futures, which is consistent with good health.

Despite these recent findings, few studies have focused on positive aspects of the immigrant experience, and most focus on the health problems of immigrants and how they impact the host societies' health and social service systems. Although many immigrants do have special health problems that demand attention, the almost exclusive focus on their health *problems* tends to perpetuate stereotypes that often foster anti-immigrant feelings (Junghans 1998).

4. *A Model Proposed by Friis et al.* (*1998*)

Friis et al. (1998) have noted that the impact of migration on health may be approached using the well-known 'stress-illness' model, where migration is considered a major life event and can be conceptualized as a source of stress (see *Stress and Health Research*). Along with migration, acculturation into a

new society is also a potentially stressful experience that can impact physical and mental health. As immigrants become more acculturated and integrated into the larger society, the level of stress they experience often is reduced.

As in the stress-illness model, Friis et al. (1998) hypothesize a number of factors that mediate and/or modify the impact of migration and acculturation on health. These include social support and lifestyle factors. Social support may take numerous forms. These include the nature and type of social networks, support from family and friends, as well as from the community at large. Immigrants are often socially isolated because of linguistic and cultural barriers. This is often the case with older immigrants whose children become assimilated and acculturated into the host society which leads to intergenerational strains and further isolation of the elderly (Markides and Black 1995).

Friis et al. (1998) suggest that migration sometimes leads to lifestyle changes, such as poor diet, alcohol and smoking, reduced physical activity, and other personal behaviors that can have adverse health outcomes. For example, studies have shown decreasing intake of fruits and vegetables among immigrant children in Western countries and high rates of obesity among children and adults because of poor and overnutrition, factors which may be associated with increased incidence of certain chronic conditions (e.g., cardiovascular diseases and diabetes later in life).

Another health outcome of interest is how immigration and level of acculturation influence health care utilization. As Friis et al. (1998) point out there has been little systematic information in both Europe and North America in this area. Clearly this is an area where more research is needed.

5. *Special Health Concerns*

As indicated previously, much of the literature has focused on specific health concerns and problems of immigrants. Some of these are addressed briefly in this section.

5.1 *Communicable Conditions*

Special health concerns often involving immigrants are communicable conditions, such as HIV/AIDS and tuberculosis. HIV/AIDS has been a worldwide epidemic for two decades that has attracted the attention of scholars and policy makers. Although it has probably spread through migration, it is not clear what the impact of migration on its prevalence has been in specific countries. Carballo et al. (1998) have noted that the prevalence of AIDS is lower among migrants than among natives in Italy and Belgium, while the opposite is the case in Germany, where most cases are traced to Africa, Asia, and the Americas. However, AIDS prevalence is lower in Germany among immigrants from Turkey and Eastern Europe than among native-born Germans.

Carballo et al. (1998) also note rates of HIV and other sexually transmitted diseases (STDs) are higher among immigrants in Sweden, especially those from Africa. They also note that there appears to be a higher risk of STDs in countries like Belgium, where migration has been primarily among males who are more likely than non-migrants to use sex workers.

Another major communicable disease of great concern in recent years has been the rise of tuberculosis as a major public health problem worldwide, especially in poor countries. It has also become a major concern in Western countries, where its prevalence is higher among immigrant groups, especially those from poor countries (Carballo et al. 1998). Although the overall impact on the host countries does not appear to be major, it has a major impact among the immigrant communities, especially where immigrants live in crowded and unsanitary environments, conditions that may promote the spread of tuberculosis.

5.2 *Chronic Conditions*

With chronic conditions accounting for the vast majority of deaths in Western societies, there has always been interest in how migration and 'Westernization' influence the prevalence of major chronic conditions. This has especially been the case for cardiovascular diseases (CVDs). Early research with Japanese and Japanese Americans showed an increasing gradient of coronary heart disease (CHD) mortality from Japan to Hawaii to California which was explained by changes in diet and lifestyle in general. Low CHD mortality was related to retention of traditional Japanese lifestyle (see discussion by Friis et al. 1998, p. 175).

Research in Europe has also examined CVD prevalence among immigrant groups. For example, high rates have been observed among immigrants from India in the United Kingdom. Stroke mortality is also high among immigrants from India and the Caribbean who have high rates of diabetes, a contributing factor to stroke (see discussion in Carballo et al. 1998, p. 938). High rates of diabetes have also been found among certain immigrant groups in the United States, especially among persons of Mexican origin. These high rates have been attributed to excess obesity and Native American admixture (Stern and Haffner 1990).

5.3 *Mental Health and Psychological Well-being*

As mentioned previously, early research in North America (e.g., Malzberg 1967) found higher rates of

mental illness, as measured by high rates of first admission to mental hospitals, among immigrants than among non-immigrants. Nevertheless, based on the recent data indicating better physical health among immigrants to the United States, Canada, and Australia reviewed earlier, one would expect good overall mental health among immigrants today, especially recent ones. However, wide-ranging studies of mental health among immigrants are scarce.

There is also some evidence both from North America and Europe that migration may impact mental health and psychological well-being differently in different age groups. While younger immigrants often assimilate into the main stream through education and employment, their parents become increasingly isolated because of linguistic and cultural barriers. This situation often produces frictions and conflicts between the generations and high rates of depression and anxiety in the older generation (see Carballo et al. 1998 and Markides and Black 1995).

Marital problems have been noted in some immigrant groups, especially when migration involves separation of spouses. When reunification eventually occurs, separation and divorce are common. Separation and divorce impact negatively, especially on children and women, whose status in the local immigrant community is usually tied to marriage and family. Evidence from several European countries reviewed by Carballo et al. (1998) suggests that when immigrant couples break up they often have difficulty finding culturally sensitive support systems with social isolation, depression and other psychological problems being quite common.

Other European evidence also points to psychosomatic problems during the first years of migration, such as ulcers, headaches, anxiety attacks, sleeping disorders, gastrointestinal complaints, and alcohol and drug abuse (Carballo et al. 1998). These and other psychological problems appear to be worse among political refugees than among voluntary migrants. Yet very few studies of the mental health (or physical health) of refugees have been conducted. In addition, the mental health needs of refugees are usually neglected despite the fact that most are highly traumatized and stressed (Friis et al. 1998). With the number of refugees worldwide on the rise, there is a great need to better understand their unique experiences and mental health problems as well as to provide them with better mental and physical health care.

5.4 Special Problems Associated with Smuggling and Trafficking of Immigrants

Gushulak and McPhearson (2000) have provided a very valuable overview of health issues associated with trafficking and smuggling of immigrants. With rising globalization, the numbers of people wanting to immigrate to more-developed countries is on the rise. In addition, there is a growing demand for unskilled workers in developed countries and despite more stringent entry requirements a rapidly growing number of people (around 4 million each year) become victims of international trafficking.

There are a number of issues impacting on the health of illegally smuggled immigrants. For example, receiving nations have put in place certain health and other screening barriers which select relatively healthy immigrants. Persons bypassing legal channels are thus not health selected. As Gushulak and MacPhearson (2000) point out, the very existence of screening barriers may encourage less healthy persons to seek illegal means of migrating.

The type of transportation may also have health consequences leading to illness, injury, or even death. A tragic example was the case of 58 Chinese-origin immigrants who died while being smuggled to England in June of 2000. They apparently suffocated in a compartment of a truck carrying tomatoes. Eshbach et al. (1999) have documented hundreds of deaths per year at the United States–Mexico border among undocumented migrants, many of whom were being smuggled.

Illegally smuggled immigrants are sometimes victims of environmental conditions, such as extreme heat or cold. Lacking social and legal protection, they often become victims of violence. And, they are more likely to suffer a variety of psychosocial problems such as sexual abuse, isolation, and psychosocial illness (Gushulak and McPhearson 2000). As with refugees, there is a need to better understand and to better treat the mental and physical health problems of illegally smuggled immigrants.

6. Conclusions

The issue of how migration impacts health is complex and multi-faced. The type of migration (voluntary vs. involuntary, legal vs. illegal) is of great importance. There are issues associated with the point of origin, point of destination, the journey itself, as well as with the health of the migrants themselves.

Most research has taken place in Western societies and has focused almost exclusively on voluntary international migrants. This type of research can be approached using a version of the well-known stress-illness model. Studies have focused almost exclusively on the health problems of immigrants despite considerable evidence that many immigrants are often healthier than non-immigrant native persons. This practice can lead to stereotyping of immigrants and to anti-immigrant sentiments.

With international migration of all types on the rise, clearly there is a great need to better understand the

health and health care needs of migrants and to develop better ways to address their physical and mental health needs.

See also: AIDS, Geography of; Community Health; Infectious Diseases: Psychosocial Aspects; Migration, Economics of; Migration: Sociological Aspects; Migration, Theory of; Mortality and the HIV/AIDS Epidemic; Social Integration, Social Networks, and Health; Social Support and Health; Unemployment and Mental Health

Bibliography

Carballo M, Divino J J, Zeric D 1998 Migration and health in the European union. *Tropical Medicine and International Health* **3**: 936–44

Chen J, Ng E, Wilkins R 1996 The health of Canada's immigrants. *Health Reports* **7**: 33–45, 47–50

Donovan J W, d'Espaignet E, Merton C, van Ommeren M (eds.) 1992 *Immigrants in Australia: A Health Profile*. Australian Institute of Health and Welfare, Ethnic Health Series, No. 1. AGPS, Canberra, Australia

Eschbach K, Hagan J, Rodriguez N, Hernández-León, Bailey S 1999 Death at the border. *International Migration Review* **34**: 430–54

Friis R, Yngue A, Perssan V 1998 Review of social epidemiologic research on migrants' health: findings, methodological cautions, and theoretical perspectives 1998. *Scandinavian Journal of Social Medicine* **26**: 173–80

Gushulak B D, MacPherson D W 2000 Health issues associated with the smuggling and trafficking of migrants. *Journal of Immigrant Health* **2**: 67–78

Hummer R A, Rogers R G, Nam C B, LeClere F B 1999 Race/ethnicity, nativity, and US adult mortality. *Social Science Quarterly* **80**: 136–53

Junghans T 1998 How unhealthy is migrating? *Tropical Medicine and International Health* **3**: 933–4

Kasl S V, Berkman L F 1983 Health consequences of the experience of migration. *Annual Review of Public Health* **4**: 69–90

Malzberg B 1967 Internal migration and mental disease among the white population of New York State, 1960–1961. *International Journal of Social Psychiatry* **13**: 184–91

Markides K S, Black S A 1995 *Race, Ethnicity and Aging: the Impact of Inequality*. *Handbook of Aging and the Social Sciences*, (4th edn.). Academic Press, San Diego, CA

Markides K S, Coreil J 1986 The health of Hispanics in the Southwestern United States: an epidemiologic paradox. *Public Health Reports* **101**: 253–65

Pol L G, Thomas R K 1992 *The Demography of Health and Health Care*. Plenum Press, New York

Stephen E H, Foote K, Hendershot G E, Schoenborn C A 1994 *Health of the Foreign-Born Population: United States, 1987–90. Advance Data from Vital and Health Statistics,* **241**: 1–10. National Center for Health Statistics, Hyattsville, MD

Stern M P, Haffner S M 1990 Type II diabetes and its complications in Mexican Americans. *Diabetes/Metabolism Review* **6**: 29–45

K. S. Markides

Migration, Economics of

Migration is the move from one geographic area to another. Residential migration occurs when the household (or person) changes its place of residence by moving from one neighborhood to another within the same local area. Internal migration occurs when the household moves across larger geographically distinct units—such as counties, metropolitan areas, states, or provinces—but remains within the same country. International migration occurs when the household moves across national boundaries.

1. Migration and Economic Efficiency

The study of migration lies at the core of labor economics because the analysis of labor flows—whether within or across countries—is a central ingredient in any discussion of labor market equilibrium. Workers respond to regional differences in economic outcomes by voting with their feet. These labor flows improve labor market efficiency.

Suppose there are two regional labor markets in a particular country, the North and the South, and that these two markets employ workers of similar skills. Suppose further that the current wage in the North exceeds the wage in the South.

Under some conditions, the wage differential between the two regions will not persist once the economy attains a competitive national equilibrium. After all, the wage differential encourages some Southern workers to pack up and move North, where they can earn higher wages and presumably attain a higher level of utility. The flow of Southern workers into the North would raise the Southern wage and depress the Northern wage. If there were free entry and exit of workers in and out of labor markets, the national economy would eventually be characterized by a single wage.

The single wage property of competitive equilibrium has important implications for economic efficiency. The theory of labor demand shows that the wage equals the value of marginal product of labor in a competitive market. As workers move to the region that provides the best opportunities, they eliminate regional wage differentials. Therefore, workers of given skills have the same value of marginal product of labor in all markets. The allocation of workers to firms that equates the value of marginal product across markets is an efficient allocation because it maximizes national income.

To see why a competitive equilibrium is efficient, suppose that a benevolent dictator takes over the economy and that this dictator has the power to dispatch workers across regions. In making allocation decisions, this dictator has one over-riding objective:

to maximize the country's income. When the dictator first takes over, he sees that the wage in the North exceeds the wage in the South. This wage gap implies that the value of marginal product of labor is greater in the North than in the South.

The dictator picks a worker at random. Where should this worker be allocated? Because the dictator wants to maximize national income, the worker will be sent to the region where he is most productive, the North. In fact, the dictator will keep allocating workers to the Northern region as long as the value of marginal product of labor is greater in the North than in the South. The law of diminishing returns implies that as the dictator forces more and more people to work in the North, the value of marginal product of Northern workers declines. In the end, the dictator will maximize national income only when the value of marginal product of workers is the same in all labor markets.

In sum, migration and economic efficiency are closely linked in a competitive economy. Through an 'invisible hand,' workers who search selfishly for better opportunities accomplish a goal that no one in the economy had in mind: an efficient allocation of resources.

2. An Economic Model of Migration

In 1932, Sir John Hicks argued that 'differences in net economic advantages, chiefly differences in wages, are the main causes of migration.' Practically all modern studies of migration use this hypothesis as the point of departure and view the migration of workers as a type of human capital investment (Sjaastad 1962). Workers calculate the value of the opportunities available in each of the alternative labor markets, net out the cost of making the move, and choose whichever option maximizes the net present value of lifetime income.

Suppose there are two labor markets where a particular worker can be employed. The worker is currently employed in region i and is considering a move to region j. The worker, who is t years old, earns w_{it} dollars. If he were to move, he would earn w_{jt} dollars. It costs M dollars to move from i to j. These migration costs include the actual expenditures incurred in transporting the worker and his family, as well as the dollar value of the 'psychic cost'—the pain and suffering that inevitably occurs when one moves away from family, neighbors, and social networks.

Like all other human capital investments, migration decisions are guided by the comparison of the present value of lifetime earnings in the alternative opportunities. The net gain to migration is given by:

$$\text{Net Gain} = \sum_{k=t}^{T} \frac{w_{jk} - w_{ik}}{(1+r)^{k-t}} - M \qquad (1)$$

Where r is the discount rate and T is the age of retirement. The worker moves if the net gain is positive.

A number of empirically testable propositions follow immediately from this framework:

(a) An improvement in the economic opportunities available in the destination increases the net gains to migration, and raises the likelihood that the worker moves.

(b) An improvement in the economic opportunities at the current location decreases the net gains to migration, and lowers the probability that the worker moves.

(c) An increase in migration costs lowers the net gains to migration, and reduces the likelihood of a move.

In sum, migration occurs when there is a good chance that the worker will recoup his human capital investment. As a result, migrants will tend to gravitate from low-income to high-income regions and the larger the income differential between the regions or the cheaper it is to move the greater the number of migrants.

This framework has been extended to address a slightly different question: *Which* persons in the source region are most likely to move? The answer to this question is particularly important for evaluating the impact of migration flows. Suppose that 10 percent of region i's population chooses to move to region j. The economic impact of this migration will depend on whether the migrants are randomly chosen from the source region's population or are chosen from a particular tail of the skill distribution.

Differences in migration costs across workers help sort out the migrants from the source region's population. Other things being equal, those workers who find it easier to move are the ones who will, in fact, move.

The shape of the income distributions in the two regions will also help sort out the migrants—even if all workers faced the same migration costs (Borjas 1987). Suppose that the skills of workers are portable across the two regions—so that employers value the same types of workers in each of the labor markets. Higher skills typically lead to higher earnings, but the rate of return to skills typically varies across labor markets. Two types of selection may characterize the migrant flow:

(a) Positive Selection occurs when the migrants have above-average skills. The migrant flow from i to j is positively selected when the destination offers a higher rate of return to skills. The migrants are then drawn from the upper tail of the skill distribution because region i, in a sense, 'taxes' high-skill workers and 'insures' less-skilled workers against poor labor market outcomes.

(b) Negative Selection occurs when the migrants have below-average skills. The migrant flow is negatively selected when the source region offers a larger

payoff to skills. Few skilled workers will then want to move from region *i*.

In short, as long as regional income differences (net of migration costs) are large enough to induce migration, highly skilled workers will naturally gravitate to those regions where the rate of return to skills is high. In the optimal sorting of workers to regions, highly skilled workers live in regions that offer high rates of return to skills and less-skilled workers live in regions where the rate of return to skills is relatively low.

The simple model in Eqn. (1) stresses how alternative income streams determine the migration decision for a particular worker. However, it is the alternative streams of utility that determine the worker's decision. After all, the worker cares not only about relative incomes, but also about the amenities and disamenities offered by the various regions.

The fact that migration maximizes utility introduces a number of interesting twists into the study of migration decisions. For instance, Eqn. (1) ignores why there are regional wage differences in the first place, implicitly assuming that the national labor market is in disequilibrium (in the sense that different regions offer different opportunities to the same worker). However, regional wage differences may partly reflect compensating wage differentials that reward workers for the varying set of amenities that different regions offer (Roback 1982). The wage would then be relatively lower in more pleasant localities. Even though a particular worker might face different wages in different labor markets, the worker's utility would be constant across labor markets. The wage differentials that are the focus of the human capital approach—and that determine the migration decision in Eqn. (1)—are the ones that persist after the analysis has controlled for regional differences in the value of amenities and disamenities (see *Wage Differentials and Structure*).

3. *Internal Migration*

Despite problems of comparability, the available data indicate that there is a great deal of internalmigration in developed economies. In Canada and the USA, the five-year migration rate across 'local areas' (localities in Canada, counties in the USA) is around 20 percent. In Ireland and Japan, the one-year migration rate across local areas (counties in Ireland, prefectures in Japan) is between 2 and 3 percent, compared to a one-year migration rate of 6.2 percent in the USA.

3.1 Determinants of Internal Migration

A large empirical literature attempts to determine which variables best explain the size and direction of internal migration flows (see the survey in Greenwood 1997). A few variables have been found to be good predictors in many countries

3.1.1 Age. Migration is most common among younger workers. The human capital model provides a simple explanation for this pattern. Older workers have a shorter period over which they can collect the returns to the migration investment. The shorter payoff period decreases the net gains to migration, and hence lowers the probability of migration.

3.1.2 Education. Migration is more common among more educated workers. This correlation could arise because highly educated workers may be more efficient at learning about employment opportunities in alternative labor markets, thus reducing migration costs. It is also possible that the geographic region that makes up the relevant labor market is larger for more educated workers. Consider, for example, the labor market faced by college professors. There are few firms in any given city and professors' skills are highly portable across colleges and universities. In effect, college professors sell their skills in a national—and perhaps even an international—labor market.

3.1.3 Distance. Greater distances deter migration because greater distances imply larger migration costs. The relationship between distance and migration was an integral part of the 'gravity models' that dominated the literature before the introduction of the human capital framework. The gravity model presumed a (mechanical) direct relationship between migration and the size of the destination and origin regions, as well as an inverse relationship between migration and distance.

3.1.4 Other variables. The evidence linking migration and other variables is more ambiguous. Some studies suggest that persons who are unemployed are more likely to migrate than persons who are not, and that the elasticity of migration with respect to the source region's unemployment rate seems to be stronger for workers who are unemployed (at least in the USA). These studies, however, typically ignore that both employment and migration propensities are endogenously determined. Other studies investigate if the wage differential between the destination and source regions has a positive impact on migration. This direct test of the human capital model requires information on the earnings stream that the

worker would face in both the source and destination regions. As a result, the evidence is typically very sensitive to how the selection bias problem is handled. Finally, a series of studies have begun to examine the link between internal migration and differences in social welfare benefits provided by different jurisdictions in the USA. The evidence is mixed, however, on whether 'welfare magnets' attract particular types of person to generous localities (Brueckner 2000).

3.2 Return and Repeat Migration

Workers who have just migrated are very likely to move back to their original location (generating return migration flows), and are also very likely to move onward to still another location (generating repeat migration flows). In the USA, the probability of an interstate migrant returning to the state of origin within a year is about 13 percent, while the probability of a migrant moving on to yet another state is 15 percent (DaVanzo 1983).

Unless economic conditions in the various states change drastically soon after the migration takes place, the high propensity of migrants to move again is not consistent with the simple human capital model summarized in Eqn. (1). Prior to the initial migration, the worker's cost-benefit calculation indicated that a move from region i to region j maximized the present value of lifetime earnings (net of migration costs). How can a similar calculation made just a few weeks after the move indicate that returning to i or perhaps moving on to region k maximizes the worker's income?

Two distinct factors can generate return and repeat migration flows. The worker might learn that the initial decision was a mistake. After all, a worker contemplating a move faces a great deal of uncertainty about economic conditions in the region of destination. Once he migrates, he might discover that the available economic opportunities are far worse than what he expected to find. Return and repeat migration flows arise as workers attempt to correct their errors.

Return or repeat migration might also be the career path that maximizes the present value of lifetime earnings in some occupations, even in the absence of any uncertainty about job opportunities in different locations. Some workers might find that a brief experience with a particular employer in a different city provides a 'stepping-stone' to better job opportunities in the future. In other words, the temporary stay in region j is but one rung in the career ladder that maximizes lifetime earnings.

3.3 Family Migration

The discussion has focused on the behavior of a single worker as he or she compares employment opportunities across regions and chooses the location that maximizes the present value of lifetime earnings. Most migration decisions, however, are not made by single workers, but by families. The migration decision, therefore, should not be based on whether a particular member of the household is better off at the destination than at the origin, but on whether the family as a whole is better off (Mincer 1978).

Suppose that the household is composed of two persons, a husband and a wife. Let ΔPV_H be the change in the present value of the husband's earnings stream if he were to move geographically from region i to region j—net of migration costs. And let ΔPV_W be the same change for the wife. If the husband were single, he would migrate if the 'private gains' ΔPV_H were positive. If the wife were single, she would migrate if ΔPV_W were positive.

The family unit (that is, the husband and the wife) will move if the net gains to the family are positive, or if $\Delta PV_H + \Delta PV_W > 0$. The optimal decision for the family unit is not necessarily the same as what is optimal for a single person. Suppose, for example, that the wife would move on her own if she were single, for she gains from the move (that is, $\Delta PV_W > 0$), but that the husband's loss exceeds her gain (so that $\Delta PV_H + \Delta PV_W < 0$). Hence it is not optimal for the family to move. The wife is, in effect, a 'tied stayer.' She will sacrifice the better employment opportunities available elsewhere because her husband is much better off in their current region of residence.

Similarly, consider the situation where the husband experiences an income loss if he moves on his own (so $\Delta PV_H < 0$). Nevertheless, when he moves as part of a family unit, the wife's gain exceeds the husband's loss (or $\Delta PV_H + \Delta PV_W > 0$), and it is optimal for the family to move. The husband is a 'tied mover.' He follows the wife even though his employment outlook is better at their current residence.

The analysis suggests that marriage deters migration because a person's private gains may be outweighed by the spouse's losses. The analysis also suggests that migration need not 'pay' for all workers in the family. A comparison of the pre- and postmigration earnings of tied movers would indicate that migration reduced their earnings. The family, however, is better off.

4. International Migration

There was a resurgence of immigration in the USA and in many other countries at the end of the twentieth century. By the year 2000, about 140 million persons—or roughly 2 percent of the world's population—resided in a country where they were not born. Nearly 6 percent of the population in Austria, 17 percent in Canada, 11 percent in France, 17 percent in Switzerland, and 10 percent in the USA was foreign-born.

Different sets of concerns have motivated economic research on internal and international migration. For the most part, the internal migration literature has examined the determinants of migration—who moves and where. In contrast, the international migration literature has been concerned with the issues that propel the debate over immigration policy. In particular, what is the economic impact of international migration on the host country?

4.1 Assimilation and Cohort Effects

The impact of immigration on the host country depends on how the skill distribution of immigrants compares to the skill distribution of the native-born population. A great deal of research attempts to document trends in the skill endowment of the immigrant population, and how that skill endowment adapts to economic and social conditions in the host country through the process of assimilation.

Early studies in this literature used cross-section data sets to trace out the age-earnings profiles of immigrants and natives (Chiswick 1978). A cross-section survey allows us to compare the current earnings of newly arrived immigrants (measured as of the time of the survey) with the current earnings of immigrants who migrated years ago. It was typically found that the earnings of newly arrived immigrant men were substantially lower than the earnings of native-born men. In contrast, immigrants who had been in the host country for two or three decades earned more than native workers did.

These findings were typically interpreted as follows. When immigrants first arrive in their destination, they lack many of the skills valued by host-country employers, including language, educational credentials, and information on what the best-paying jobs are and where they are located. As immigrants learn about the host country, their human capital grows relative to that of natives, and economic assimilation occurs in the sense that immigrant earnings begin to converge to the earnings of natives. Immigrant earnings could eventually surpass native earnings if immigrants are positively selected from the population of the source countries. The cross-section evidence thus seemed to indicate that upward mobility was an important aspect of the immigrant experience in many host countries.

This interpretation draws inferences about how the earnings of immigrant workers evolve over time from a single snapshot of the population. Suppose, however, that today's newly arrived immigrants are inherently less skilled than those who arrived twenty years ago. The poor economic performance of recent arrivals may then indicate that the newest immigrants have few skills and poor economic potential and will always have low earnings, while the economic success enjoyed by the earlier arrivals may indicate that the earlier waves were always better skilled and had greater economic potential. Because of these intrinsic differences in skills across immigrant cohorts, one cannot use the current labor market experiences of those who arrived twenty or thirty years ago to forecast the future earnings of newly arrived immigrants (Borjas 1985).

In short, a cross-section survey yields an incorrect picture of the assimilation process if there are skill differentials among immigrant cohorts at the time they entered the host country. These 'cohort effects' can be generated by shifts in immigration policy, by changing economic conditions in the host and source countries, and by selective outmigration of immigrants in the host country.

Many studies, using either longitudinal data or repeated cross-sections, have calculated the rate of economic assimilation and measured the importance of cohort effects in immigrant-receiving countries (see the survey in Borjas 1994). The findings typically differ across countries. In the USA, the immigrant waves that entered the country in the 1980s and 1990s were relatively less skilled than the waves that entered in the 1960s and 1970s. Immigrants in the USA also experience some economic assimilation.

Many of the international differences in the skill endowment of the immigrant population and in the rate of economic assimilation are probably due to differences in immigration or adaptation policies. The existing research, however, does not examine this relationship.

4.2 Labor Market Impact of Immigration

The entry of immigrants into a particular labor market should lower the wage of competing workers (workers who have the same types of skills as immigrants), and increase the wage of complementary workers (workers whose skills become more valuable because of immigration). For example, an influx of foreign-born laborers into a particular locality reduces the economic opportunities for laborers who already live in the locality—all laborers now face stiffer competition in the labor market. At the same time, highly skilled natives may gain substantially. They pay less for the services that laborers provide, such as painting the house and mowing the lawn and natives who hire these laborers can now specialize in producing the goods and services that better suit their skills.

In many host countries, immigrants cluster in a limited number of geographic areas. Many empirical studies exploit this fact to identify the labor market impact of immigration by comparing labor market conditions in 'immigrant cities' with conditions in markets untouched by immigration (Card 1990, Altonji and Card 1991). These studies typically correlate some measure of economic outcomes for native

workers with a measure of immigrant penetration in the locality and often report a weak 'spatial' correlation. The evidence is often interpreted as indicating that immigration has little impact on the labor market opportunities of the native-born.

The 'short-run' perspective that frames this type of research can be very misleading. Over time, natives who live in the immigrant cities—as well as natives who live in other cities—will likely respond to the entry of immigrants. It is not in the best interest of native-owned firms or native workers to sit still and watch immigrants change economic opportunities. All natives now have incentives to change their behavior to take advantage of the altered economic landscape.

For instance, native-owned firms see that cities flooded by less skilled immigrants tend to pay lower wages to laborers. Employers who hire laborers will want to relocate to those cities, and entrepreneurs thinking about starting up new firms will find it more profitable to open them in immigrant areas. The flow of jobs to the immigrant-hit areas helps cushion the adverse effect of immigration on the wage of competing workers in these localities.

In addition, laborers living in areas not directly affected by immigration might have been thinking about moving to the cities penetrated by immigrants *before* the immigrants entered the country. They will now choose to move elsewhere. And some native-born laborers living in the immigrant cities will seek better opportunities elsewhere.

The internal migration of native workers and firms within the host country, in effect, can accomplish what the immigrant flow, with its tendency to cluster in a small number of gateway localities, did not—a 'spreading out' of the additional workers over the entire nation, rather than in just a limited number of localities. A spatial comparison of the employment opportunities of native workers in different localities might show little or no difference because, in the end, immigration affected every city, not just the ones that actually received immigrants.

Because local labor market conditions may not provide valuable information about the economic impact of immigration, some studies have attempted to measure the impact at the level of the national market (Borjas et al. 1997). The 'factor proportions approach' compares the host country's actual supplies of workers in particular skill groups to those it would have had in the absence of immigration, and then uses outside information on labor demand elasticities to calculate the wage consequences of immigration. Suppose, for instance, that in the absence of immigration there would be one unskilled worker per skilled worker. Immigration may change this factor proportion so that there are now two unskilled workers per skilled worker. Such a change in factor proportions should widen the wage gap between skilled and unskilled workers. During the 1980s and 1990s, the immigrant flow in the USA was relatively less skilled. As a result, the factor proportions approach implies that immigration had a sizable adverse impact on the relative wage of native-born workers at the bottom of the skill distribution.

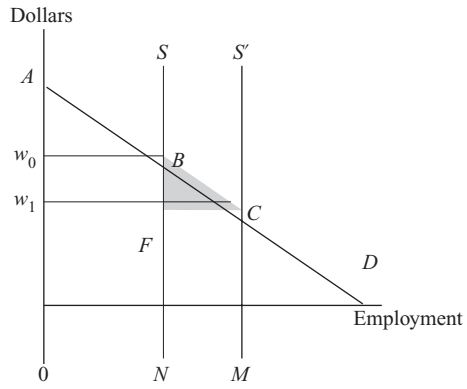

Figure 1
The benefits from immigration to the host country

The factor proportions approach is unsatisfactory in one important sense. It does not estimate the impact of immigration on the labor market by directly observing how this shock affects some workers and not others. Instead, the approach simulates the impact of immigration at the national level. For given labor demand elasticities, the factor proportions approach mechanically predicts the relative wage consequences of shifts in supply. The results, therefore, are sensitive to the assumptions made about the value of the elasticity that links changes in relative wages to relative supplies.

4.3 The Gains to the Host Country

It is simple to describe how the host country benefits from immigration in the context of the 'textbook model' of a competitive labor market. Consider the standard supply-demand analysis presented in Fig. 1. The supply curve of labor is given by S and the demand curve for labor is given by D. For simplicity, suppose that the labor supply curve is inelastic, so that there are N native-born workers. A competitive market equilibrium implies that the N native workers are initially employed at a wage of w_0.

Each point on the labor demand curve gives the value of marginal product of the last worker hired (see *Demand for Labor*). As a result, the area under the demand curve gives the total product of all workers hired. Hence the area in the trapezoid $ABN0$ measures the value of national income prior to immigration.

Suppose that immigrants enter the country and that, in the short run, the capital stock remains fixed so

that the demand curve does not shift. If we assume that immigrants and natives are perfect substitutes in production, the supply curve shifts to S' and the market wage falls to w_1. National income is now given by the area in the trapezoid $ACM0$. The figure shows that the total wage bill paid to immigrants is given by the area in the rectangle $FCMN$, so that the increase in national income accruing to natives is given by the area in the triangle BCF. This triangle is the *immigration surplus*, and measures the increase in national income that occurs as a result of immigration and that accrues to natives.

Why does the host country benefit? Because the market wage equals the productivity of the last immigrant hired. As a result, immigrants increase national income by more than what it costs to employ them. Put differently, all the immigrants hired except for the last one contribute more to the economy than they get paid.

The simulation of this model for the USA suggests that the net benefit from immigration is quite small (Borjas 1995). If a 10 percent increase in supply lowers the wage by 3 percent, the immigration surplus is on the order of 0.1 percent of GDP. This small net gain, however, disguises the fact that there may have substantial wealth transfers from native-born workers to the capitalists who employ the services that immigrants provide.

Immigration alters labor market conditions in both the host and source countries. Moreover, the immigrant population in the host country often makes large financial transfers, or remittances, to family members left in the source country. These remittances can have a substantial impact on economic activity, particularly in developing source countries. The impact of international migration flows on the source countries has not been studied systematically.

Bibliography

Altonji J G, Card D 1991 The effects of immigration on the labor market outcomes of less-skilled natives. In: Abowd J M, Freeman R B (eds.) *Immigration, Trade, and the Labor Market*. University of Chicago Press, Chicago, pp. 201–34

Borjas G J 1985 Assimilation, changes in cohort quality, and the earnings of immigrants. *Journal of Labor Economics* **3**: 463–89

Borjas G J 1987 Self-selection and the earnings of immigrants. *American Economic Review* **77**: 531–53

Borjas G J 1994 The economics of immigration. *Journal of Economic Literature* **32**: 1667–717

Borjas G J 1995 The economic benefits from immigration. *Journal of Economic Perspectives* **9**: 3–22

Borjas C J, Freeman R B, Katz L F 1997 How much do immigration and trade affect labor market outcomes? *Brookings Papers on Economic Activity* **1**: 1–67

Brueckner J K 2000 Welfare reform and the race to the bottom: Theory and evidence. *Southern Economic Journal* **66**: 505–25

Card D 1990 The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review* **43**: 245–57

Chiswick B R 1978 The effect of Americanization on the earnings of foreign-born men. *Journal of Political Economy* **86**: 897–57

DaVanzo J 1983 Repeat migration in the United States: Who moves back and who moves on? *Review of Economics and Statistics* **65**: 552–9

Greenwood M J 1997 Internal migration in developed countries. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, Vol. 1B, pp. 647–720

Mincer J 1978 Family migration decisions. *Journal of Political Economy* **86**: 749–73

Roback J 1982 Wages, rents, and the quality of life. *Journal of Political Economy* **90**: 1257–78

Sjaastad L A 1962 The costs and returns of human migration. *Journal of Political Economy* **70**: 80–93

G. J. Borjas

Migration History

The term 'migration history' has historiographical and historical dimensions. This article is thus divided into two sections. The first part outlines concepts and methods, focusing on tasks, fields, and problems of interdisciplinary historical migration research. The second section gives some insight into the history of migration itself, based primarily on modern European history. Non-European migration is included where it was linked in some way to European migration.

1. Historical Migration Research: Issues and Concepts

1.1 Motivations, Patterns, Typologies

Migration has always been a constitutive element of the *conditio humana* as *homo sapiens* spread over the world as *homo migrans*. The history of migration is part of general history and can only be understood in context. As social processes, after all, migration movements are responses to complex economic, ecological, social and cultural, religious, ethnic, and political conditions and challenges. Migration thereby gradually penetrates all spheres of life. Historical migration research branches spaces, cultures, and historical times. Therefore it is necessary for researchers to adopt an interdisciplinary approach. Varying in degree according to the problem at hand, interdisciplinary research strategies cover almost all

human sciences in both contemporary empirical and historical migration research.

There was and still is a broad variety of reasons for migration. We can, for example, posit economic or social motivations. Within this field, we could then distinguish between subsistence, betterment, or career migration (Tilly 1978). These types of movements can in turn be distinguished from migrations motivated by religious and political or ethno-nationalistic and racial reasons, which also cause flight and forced migrations. These included the expulsions and forced repatriations in the twentieth century, where the movement of borders across people often caused movements of people across borders.

Of crucial importance for any critical analysis of historical migration trends—and for an insight into the fate of many migrants, often less a matter of choice than of circumstance—is an awareness of the fact that definitions and attributes such as 'emigrant' and 'immigrant,' 'labor migrant,' or 'refugee' and 'asylum seeker' are ascribed classifications. So far, they have often been assigned for administrative or tax purposes, or for epistemological scientific reasons—which also depend on specific classification criteria—but in no way describe the generally multiple identities of migrants (Castles and Miller 1998).

The matter is even more complicated because when migration was controlled and restricted, migrants had to assume the official classification of their status in order to get past immigration officials, often leaving a 'false trace' in official records and statistics. This is one more reason why we have to distinguish between the way migrants classified themselves and the way they were classified, for example by the state or contemporary researchers.

1.2 Spatial Dimensions and Research Concepts

In examining spatial mobility, we have to distinguish between movements in geographical spaces and those in social spaces. Geographically, the scope of historical migration research ranges from the macrocosmos of international and intercontinental mass movements to the microcosmos of interregional or interlocal migrations, and thus from large scale studies on a highly abstract level to small scale case studies with a larger socio-historical focus. Levels and methods of historical migration research thus range from microhistorical to meso- and macrohistorical approaches, including even multilevel theories of migration research, and from individual or group-specific dimensions to quantitative analyses of highly aggregated mass data serving to determine collective behavior during mass movements. On the temporal axis, the field of historical migration research stretches from long term studies of single migration movements to cross-sectional analyses of the entire, simultaneous migration in one region or beyond its borders (J. Lucassen and L. Lucassen 1997).

The varying approaches by individual disciplines and the various emphases in interdisciplinary approaches lead to different interpretations of migration history. As a social and historical phenomenon, for instance, migration should be seen as a complex process. According to 'classical' historical migration research, this process was often triggered by an increasing propensity to migrate, e.g., in the case of European transatlantic mass emigration in the nineteenth century, and by the consequently more or less gradual mental segregation from the social context of the home region. In this process, transatlantic migrant networks played an important role. A next phase would be the transformation—often provoked by some external cause—of this propensity into an actual decision to migrate, followed by the act itself. In the case of dense transnational networks resulting from chain migrations, the departure often took place more or less abruptly. The last phase—provided the migration process had not been aborted or been reversed by remigration—was described as assimilation into the social and cultural context of the immigration region. In the case of large discrepancies in material culture, socio-cultural norms, or collective mentalities, assimilation could become a long term social and cultural process, sometimes even reaching intergenerational dimensions ('second generation immigrant').

The 'classical' approaches of historical migration research focused on movement in geographical spaces. In the 1990s, however, new approaches emerged which were to focus upon the movement and positioning of migrants in social spaces. This applied especially to mesolevel network theories as well as to theories and typologies of transnational social spaces and migrant identities. These new approaches were derived mainly from the social and political sciences (Faist 1998, 2000, Pries 1999), dealing with the accelerated development of transnational structures in the economy, in society, and politics against the background of rapidly forced globalization. These new approaches in migration research sometimes lead to unproductive claims of exclusivity and needlessly hostile confrontation with the 'classical' approaches of migration research, even though both approaches can be usefully incorporated into complex research concepts (Gerber 2000).

1.3 Tasks of Historical Migration Research

Historical migration research has three main tasks. The first is to investigate migration movements in terms of volume, courses, and structures. The second task is to study the behavioral patterns of migrants with respect to, for example, region, class, group, and gender. The third task is to embed migration move-

ments and behavioral patterns of migrants into the framework of population, economy, society, and culture of both emigration and immigration areas. This includes the economic, social, and cultural tensions between both sides encouraging migrations as well as the effects of migration on both sides.

Despite such comprehensive tasks and the fact that the movement of peoples rates as one of the most 'moving' moments in history, historical migration research is not an independent discipline. It is, in fact, an interdisciplinary field of research, to which humanities as well as social and behavioral sciences contribute.

Historical migration research as an interdisciplinary branch is relatively young. Its disciplinary main foci are the historical sciences, including *inter alia* the history of population, economic, social, cultural, and gender history as well as ethno-history and historical anthropology. They also include the branches of legal and political history analyzing the structure of migration procedures and their repercussions on migration processes. And there are links to mainly empirical disciplines or approaches that can be also applied to historical issues and historiographical approaches (e.g., sociology, social geography, or social psychology).

2. *Migration Movements: Global and European Perspectives*

Migration was and still is by no means a uniquely western phenomenon. People in Africa, Asia, and the New World moved for similar reasons as did Europeans. As in Europe, most people moved over short distances, usually remaining within their own region or continent. Nonetheless, there have been several non-western Diasporas. Between 600 and 1500, the Arabs and the Turks expanded into parts of southern Europe and controlled the Balkans between 1500 and 1914. The Japanese and Russians colonized Korea and Siberia respectively, and the Chinese created trading communities along the shores of the Indian Ocean (Hoerder 2001).

We can classify the history of global migrations according to regions, periods, or migration patterns. We shall focus here, by way of example, on the modern history of European migrations, looking first at the non-European references linked to the external migration of Europeans. We know that the most voluminous migration movements outside of Europe were triggered by the process of European expansion and contraction. Many Europeans and non-Europeans met for the first time during the colonial migrations that followed the age of European 'discoveries,' and, for the last time, during post-colonial migrations from the former European colonies back to the 'mother countries.'

2.1 *Global Dimensions*

Some migrations outside of Europe were directly carried out or instigated by Europeans, e.g., the settler migrations from Europe to North and South America, South Africa, Australia, and New Zealand (totalling about 70 million people); the transatlantic African slave trade (12 million); and the migration of colonial soldiers and indentured laborers from China, India, Java, and Japan to the colonial labor markets in Asia, Africa, and South America (2.5 million). In total, the volume of intercontinental migrations directly caused by the European expansion can be estimated at about 100 million migrants, both voluntary and involuntary.

In addition to migratory movements controlled by Europeans, an unknown number of Africans, Asians, and Amerindians moved as an indirect consequence of European expansion. In the New World, the indigenous Amerindian population migrated or was driven to those regions left to them by the intruding Europeans, a fate shared by the indigenous populations in parts of southern Africa, Australia, and New Zealand. In Asia and Africa, European expansion led to new jobs in textile mills, mining, cash-crop agriculture, railway construction, and in the colonial armies, causing large scale internal labor migrations. From this point on, we shall only mention migration movements directly controlled by the Europeans.

The migrations resulting from the European expansion can be divided into three types of movements and three periods: (a) the migration to the colonial plantations, mines and railway construction sites between 1500 and 1900; (b) the transatlantic mass exodus of Europeans from the mid-nineteenth century until 1914, and their overseas emigration between 1945 and 1960; (c) the colonial and overseas return migration as well as the non-European labor migration to Western Europe after 1945. We shall briefly outline the first two movements, and include the third in the review of developments in Europe (Sect. 2.2).

The first overseas migration movement started around 1500, just after the Iberians had discovered the New World. The Amerindian population declined rapidly because of European epidemic diseases and economic exploitation, while the number of Europeans willing to settle in the American colonies was too small to satisfy the demand for labor in the mines and plantations. The Iberians had been faced with a similar labor shortage in the south of their own peninsula and had imported slave labor from sub-Saharan Africa. Some of these slaves were brought to the New World, and soon the Iberians were also transporting slaves directly from West Africa to the New World. The British, French, and Dutch followed suit, and between 1500 and 1850 roughly 5 million slaves were brought to Brazil, 5 million to the Caribbean, and 1 million each to Spanish and North America. In three and a half centuries, less than 4 million Europeans migrated to the New World,

making it a demographic extension of Africa, not of Europe (Emmer 1991).

After 1850, humanitarian protests totally suppressed the Atlantic slave trade and won emancipation for all slaves in the New World. However, the demand for tropical labor remained. The recruitment of free African migrants failed, and although the number of European migrants rose dramatically after the middle of the nineteenth century, they chose to settle in the more temperate overseas zones. Employers in the tropical world then turned to Asia, and between 1850 and 1914 about 2.5 million migrant workers from India, China, Japan, Java, and Polynesia went to work in the coffee fields around Sao Paulo and on islands off the Peruvian coast to dig guano. They also worked on sugar plantations in the Caribbean, Natal, Fiji, Hawaii, and Queensland, in the tobacco fields of Sumatra, on rubber plantations in Indochina, and even helped to build the railways of East Africa (Indians) and California (Chinese) (Engerman 1986, Northrup 1995).

Most of these Asian workers had signed a contract of indenture that granted them free passage in exchange for overseas employment for a fixed period, and entitled them to return home once their contracts expired. The majority did not return, creating ex-Indian communities in South Africa, Fiji, Hawaii, Trinidad, Guyana, and Suriname, ex-Japanese communities in Brazil and Hawaii, and ex-Chinese communities in the Caribbean and Indonesia. Humanitarian protests in the West and nationalist protests in China and India brought some of these migration movements to an end. After World War Two, millions of Asian migrants began to cast their sights on the Middle East.

One special chapter in the encounter between the European and non-European world through migration was marked by the European transatlantic mass exodus from the mid-nineteenth to early twentieth century (Nugent 1992, Baines 1995). This exodus was far stronger than the European colonial migrations. Up to the 1830s, the continental migration from central Europe to eastern and southeastern Europe—in contrast to the emigration from England, Ireland, Scotland, and Wales—was much more powerful than transatlantic migration, which had become a mass movement only by the middle of the century.

As a secular social mass movement, transatlantic migration accompanied the shift in Europe from agriculture to industry. It was the transportation revolution that ultimately facilitated the mass exodus. Of the same importance, however, were the transatlantic networks established by migrations of people and exports of goods and capital even prior to the Industrial Age. The mass migration was preceded by the transatlantic movement of colonial labor and settlement migrations of destitute Europeans—men and women—who worked off their passage to the New World in 'indentured servitude' and were eventually rewarded with a small amount of start-up capital or a piece of land (Wokeck 1999).

Essential factors for the development of transatlantic emigration in the nineteenth century were the liberty to emigrate from the sending countries, the need for immigrants, and their full acceptance in the receiving countries. The advent of the steamship and the expansion of railroad systems on both sides of the Atlantic made cheap passages possible. Chain migrations established transatlantic networks and a dense transatlantic communication. Once underway, the transatlantic mass movement developed a growing internal dynamic up to World War One, after which migration controls and restrictions began to curb the trend.

Up to the late 1880s, the 'classical' sending regions for mass emigration in the nineteenth century—excluding France, which was hardly affected—were the relatively well-developed industrial countries of western, central, and northern Europe. Once employment increased due to industrialization and the economic tensions towards the US decreased, transatlantic migration from these regions slowed down in the 1890s, except for Great Britain (Ireland). From the 1880s on, however, emigration from southern, southeastern, and eastern European regions increased all the more, roughly corresponding to the north-to-south and west-to-east rate of industrialization. In the USA, this transatlantic movement soon came to be known as the 'new immigration.'

A gross total of 63 million (including returnees) and a net total of 50–55 million Europeans emigrated overseas from 1820 to 1915. The main destinations up to the late nineteenth century were in North America, with the US far ahead of Canada. New Zealand and Australia began to catch up in the 1860s, as did, from the 1870s on, South American countries which had had mixed populations since the colonial era. Beginning in the 1880s and 1890s, Argentina and Brazil attracted large numbers of emigrants in the growing wave of migration from southern Europe, especially from Spain and Portugal, causing a fall in the percentage entering the US—from about 80 percent prior to 1850, to about 75 percent from 1851 to 1890, to roughly a half from then on.

European overseas emigration reached a final climax between 1945 and 1960. Up to the mid-1960s, it still included more people than immigration to Europe from Turkey, Asia, Africa, and the Caribbean (Münz 1997, p. 225).

2.2 European Dimensions

Europe was really on the move in the nineteenth and twentieth centuries, but also in the Middle Ages and the early modern period. If we could look at spatial mobility from the Middle Ages up to the end of the

twentieth century and at the same time could eliminate reduced traveling times and longer distances due to innovations like steamship, railway, and airplane, we probably would not see a major rise in mobility towards industrial and post-industrial Europe. In the Middle Ages, after all, the majority of the European population, at least for certain parts of their lifetimes, had to be mobile to survive; only a minority of them stayed at home living off subsistence farming or local employment throughout their lives (Schubert 1995).

In the early modern period, various groups of European migrants temporarily or permanently moved over great distances via land or water. Among other groups there were migrating artists and artisans, architects and technical experts, seasonal or itinerant laborers and migrating tradesmen of fixed abode, travellers, laborers, mercenaries, sailors, and laborers for the maritime and colonial labor markets. There were settlement migrations recruited or invited by state authorities, e.g., the 'Peuplierung' of Prussia, the 'impopulation' of the Donau monarchy, and the colonial settlements in Russia under Katharine II. And there were the movements of religious refugees and expellees, who were often welcomed with open arms for economic reasons by the host authorities. All these long distance movements were paralleled by an even greater number of short distance moves between villages and small cities.

The three most relevant large scale types of migration in modern Europe were: (a) flight and forced migrations in Europe; (b) economically motivated migrations in Europe; and (c) flight, minority, and economically motivated migrations to Europe (Page Moch 1992, Bade 2000).

(a) Among flight and forced migrations, movements for religious but also for political reasons predominated during the European cultural crisis of the early modern era. In the late eighteenth and especially in the nineteenth centuries, however, the most common reason for flight and exile was political persecution. Perceived as a threat to the European stability pact negotiated at the Congress of Vienna in 1815, all constitutional reform, national and social movements were radically suppressed. Revolts and revolutions were crushed, causing a dramatic rise in the numbers of political refugees. The period from 1830 to 1848/49 qualified the nineteenth century as the era of political exile. Still, the number of political refugees was low in comparison to that of new types of refugees, casualties of the epoch of the nation-state. The founding of the nation-states created minorities within national borders, thereby laying the foundations for the flight and forced migrations of the twentieth century, which would go down in history as the 'century of refugees.'

The series of benchmarks of the tragedy of flight and forced migrations in the twentieth century began with World War One, during which millions fled war zones or were deported. During the inter-war period, about 1.5 million people fled from revolutionary and Soviet Russia, while around 5 million were resettled as a result of the great 'population swap' carved out by the new nation-states that had emerged from the ashes of the three multi-ethnic empires. During World War Two all this, however, was surpassed by the flight and deportations of 50–60 million people out of which 6 million Jews were killed in mass executions or industrialized mass murder in German concentration and extermination camps. Immediately after the War came the expulsion of about 14 million Germans from the former eastern territories of the German Reich and from the settlement districts of ethnic Germans in the East. After the era of flight from eastern to central and western Europe during the Cold War, the 'ethnic cleansing' of the inter-war period and of World War Two was re-enacted in the early 1990s in the expulsions and flight caused by the wars in the former multi-ethnic republic of Yugoslavia.

(b) Besides migrant trade—very important up to the expansion of commodity markets in the nineteenth century—labor migrations were the most important ones among economically motivated movements. They evolved into predominantly agricultural migration systems with fixed circular movements (Lucassen 1987). These systems were held together by long seasonal migration traditions between regions with virtually inverse requirements: the poor rural sending regions did not have enough work or just work at very low wages and had seasonally available surplus labor. The target areas, usually involved in intensive monoculture, offered seasonal work and much higher wages than the sending regions. Apart from agriculture, seasonal construction work in developing cities and their surrounding areas was also attractive.

Out of about 20 European labor migration systems operating at the turn from the eighteenth to the nineteenth century, J. Lucassen reconstructed seven larger systems from a far earlier period. In these systems, more than 300,000 labor migrants, men and women, traveled up to 250–300 km at the turn of the century, within and across state borders. The most important of these migrating systems was the 'North Sea system' from the start of the seventeenth to the middle of the nineteenth century. Starting in the Netherlands, it spanned the whole north European coast. Besides supplying seasonal farm labor, the coastal harbors also provided access to maritime and colonial labor markets of the North Sea system.

During the industrialization process, the magnetic field of migration in north central Europe underwent a dramatic change in its powers of attraction. The North Sea system was overtaken by the new industrial coal and steel mining centers, especially in the Ruhr area and in Lorraine. In north central Europe, the fall-off in seasonal agricultural migrations to the Netherlands was followed—up to World War One—by an increase in east-to-west migrations to the German northeastern territories, with women comprising about 50 percent

of the labor force on the huge east-Elbian estates of Prussia. Poles and Italians formed the main contingents of foreign labor in Germany and France, in the *longue durée* hinting at the large scale inner-European south-to-north and east-to-west migrations which eventually became the hallmark of the second half of the twentieth century.

The two World Wars suspended transnational mobility in the new migration systems, changed it with the restructuring of the German northeastern territories after World War One, and brought it to a complete standstill with the loss of the German eastern territories after World War Two. During World War Two, labor migrations in Nazi Germany and in German-occupied Europe were largely replaced by the deportation and forced employment of disenfranchised slave laborers, who accounted for most of the 11 million 'displaced persons' (DPs) after the War.

Only ten years after a war that Germany had started, Germany and Italy signed the Labour Recruitment Treaty of 1955, paving the way for the future pan-European system of 'guest worker migrations.' Leaving Turkey aside, all the sending regions lay in the northern Mediterranean. Receiving areas were the highly industrialized and economically thriving countries of central, northern, and western Europe up to the first half of the 1970s, when, in the face of projected economic crises, recruitment bans and immigration restrictions were introduced.

Despite the high level of remigration, many labor migrants at that time settled in the host countries and sent for their families to join them. Short term migrations became long term ones, ultimately evolving into genuine immigration processes. Since the late 1970s, they have gradually transformed the host countries of 'guest worker migrations' into immigration countries.

In the former colonial nations, particularly in England, France, and the Netherlands, the role of 'guest workers' was first occupied by postcolonial immigrations, which were generally permanent from the outset. During the whole decolonization process, about 7 million people, including remigrating Europeans, migrated from the former European colonies to Europe after World War Two. Profound changes took place, when, in the 1980s, the now thriving former south European sending regions of 'guest worker migrations' received increasing inter-continental south-to-north migrations. This ultimately transformed Europe as a whole from an emigration into an immigration continent.

(c) The 'new immigrations' to Europe included both inter-continental south-to-north migrations and the new east-to-west migrations since the fall of the Iron Curtain. As in postcolonial migrations, privileged migrations legitimized by historical or cultural links to the host regions were predominant here. This mainly included minorities from the former Soviet Empire and its successor states, such as Armenians, ethnic Germans, and Jews (Bade and Oltmer 1999, Fassmann and Münz 2000). Apart from and often coinciding with non-European labor migrations to Europe, there were growing global migrations of asylum-seeking refugees in the south-to-north direction since the early 1980s, and, since the fall of the Iron Curtain, increasingly also in the east-to-west direction.

Worldwide migration movements have increased in this age of globalization, media, and information networks. Yet, for the most part migrants remained in the surrounding areas of the sending regions and the percentage reaching Europe was still only about 5 percent at the end of the twentieth century. Nevertheless, horror visions of global mass migrations towards the continent have captured the European imagination, equating migration policy with security and defense policy.

Experts differ in their assessment of the 'migration pressure' from southern and eastern regions. The key questions are whether it is even directed at Europe, whether it will gradually and inexorably grow, and whether it can be curbed by coordinated—i.e., global, not just European—intervention ('global governance') to control the causes of migration (Nuscheler 1995, Opitz 1997). From the entire range of conceivable strategies, Europe up to 2001 has done least to tackle the causes of involuntary migration in the sending areas, and most to combat flight migrations to Europe.

With the freedom of movement within the European Union, internal 'vulnerability' grew—to use defense policy jargon—due to immigrations from outside the Community. The flipside of opening the borders has therefore been the increased closing of a 'fortress Europe.' Apart from private visits, tourism, and other short term stays, the European defense system against non-European immigration only admits people welcome for economic, cultural, or other grounds, e.g., highly qualified specialists, scientists, artists, and people who are accepted as members of privileged postcolonial or ethnic minorities, or have to be tolerated to some extent because of universalist or human-rights principles (family reunions, refugees, asylum seekers) (Santel 1995).

In current migration debates and migration policies the tension has increased between self-descriptions and official ascriptions, i.e., between the way migrants perceive themselves and the identities assigned to them by immigration authorities. Migrants must do their best to fit these assigned identities in order to have a chance of acceptance. Ascriptions, e.g., of 'refugee characteristics' are the codes of administrative systems controlling and managing migrants' destinies. The decision on who is a 'real' refugee, depends on the fulfillment of these one-sided criteria. What matters most to asylum-seeking refugees is, therefore, often not what has happened to them, but whether their story fits into the catalogue of available ascriptions laid down by the host country. Hence the approaches of migration policy and migration research to such

conceptual problems of migration may seem quite similar to some extent, despite the fundamental conflict of interest concerning the ascriptions used on both sides (Bade 2000).

See also: Colonization and Colonialism, History of; Global Population Trends; Globalization: Geographical Aspects; Migration, Economics of; Migration into the Americas and between American Countries; Migration out of Europe; Migration: Sociological Aspects; Migration, Theory of; Migrations, Colonizations, and Diasporas in Archaeology; Refugees in Anthropology

Bibliography

Bade K J 2000 *Europa in Bewegung. Migration vom späten 18. Jahrhundert bis zur Gegenwart*. CH Beck Munich, Germany (Engl., French, Ital., Span. editions forthcoming 2001/02)

Bade K J 2001 Historische Migrations forschung. In: Oltmer J (ed.) *Migrationsforschung und Interkulturelle Studien*. IO Jare IMIS, Osnabrueck

Bade K J, Oltmer J (eds.) 1999 *Aussiedler—deutsche Einwanderer aus Osteuropa*. Rasch, Osnabrück, Germany

Baines D 1995 *Emigration from Europe, 1815–1930*. Cambridge University Press, Cambridge, MA

Castles S, Miller M J 1998 *The Age of Migration. International Population Movements in the Modern World*, 2nd edn. Macmillan, London

Emmer P C 1991 European expansion and migration: The European colonial past and intercontinental migration, an overview. In: Emmer P C, Mörner M (eds.) *European Expansion and Migration. Essays on the Intercontinental Migration from Africa, Asia, and Europe*. Benz Publishers, New York, pp. 1–12

Engerman S L 1986 Servants to slaves to servants: Contract labour and European expansion. In: Emmer P C (ed.) *Colonialism and Migration. Indentured Labour before and after Slavery*. Martinus Nijhoff Publishers, Boston, pp. 263–94

Faist T 1998 *International Migration and Transnational Social Spaces*. Institut für Interkulturelle und Internationale Studien, Bremen, Germany

Faist T 2000 *The Volume and Dynamics of International Migration and Transnational Social Spaces*. Clarendon Press, Oxford, UK

Fassmann H, Münz R (eds.) 2000 *Ost-West-Wanderung in Europa*. Böhlau, Vienna, Austria

Gerber D A 2000 Theories and lives: Transnationalism and the conceptualization of international migrations to the United States. In: Bommes M (ed.) *Transnationalismus und Kulturvergleich* (IMIS-Beiträge, vol. 15). Rasch, Osnabrück, Germany, pp. 31–53

Hoerder D (forthcoming) 2001 *Cultures in Contact. European and World Migrations, 11th Century to 1990s*. Duke University Press, Durham, NC

Lucassen J 1987 *Migrant Labour in Europe, 1600–1900. The Drift to the North Sea*. Croom Helm, London

Lucassen J, Lucassen L (eds.) 1997 *Migration, Migration History, History. Old Paradigms and New Perspectives*. Peter Lang, Bern, Switzerland

Münz R 1997 Woher—wohin? Massenmigration im Europa des 20. Jahrhunderts. In: Pries L (ed.) *Transnationale Migration (Soziale Welt, Sonderband 12)*. Nomos, Baden-Baden, Germany, pp. 221–43

Northrup D 1995 *Indentured Labour in the Age of Imperialism, 1834–1933*. Cambridge University Press, Cambridge, UK

Nugent W 1992 *Crossings. The Great Transatlantic Migrations, 1870–1914*. Bloomington, IN

Nuscheler F 1995 *Internationale Migration, Flucht und Asyl*. Opladen

Opitz P J (ed.) 1997 *Der globale Marsch. Flucht und Migration als Weltproblem*. C.H.Beck, Munich, Germany

Page Moch L 1992 *Moving Europeans. Migration in Western Europe since 1650*. Indiana University Press, Bloomington, IN

Pries L (ed.) 1999 *Migration and Transnational Social Spaces*. Ashgate, Aldershot, UK

Santel B 1995 *Migration in und nach Europa*. Opladen

Schubert E 1995 *Fahrendes Volk im Mittelalter*. Bielefeld, Germany

Tilly C 1978 Migration in modern European history. In: McNeill W H, Adams R S (eds.) *Human Migration. Patterns and Policies*. Indiana University Press, Bloomington, IN, pp. 48–72

Wokeck M 1999 *Trade in Strangers. The Beginning of Mass Migration to North America*. Pennsylvania State University Press, University Park, PA

K. J. Bade

Migration into the Americas and between American Countries

The history of the Americas is deeply marked by migration in its various forms. Immigration from outside the region has had a demographic impact in the 500 years that elapsed since colonization began: first, by the occupation of territories by the settlers, then by the enforced migration of African peoples, and finally by the immigration of Europeans and Asians that took place in the nineteenth century and the first decades of the twentieth century.

During the second half of the twentieth century the migration trends of Anglo-Saxon North America and that of the Latin American region took opposite directions. While the USA and Canada consolidated their position as great immigration receivers, Latin America turned from an immigration area into an emigration one, with displacements within the region or towards the developed world, especially the USA.

This article contains a summary of the voluntary migrations that have taken place towards and within

the Americas. No attention is given to enforced migration from Africa, because of its different nature.

1. Colonization

Settlement of the New World entailed the large-scale transfer of population from the colonial metropolises in order to dominate the natives and consolidate the colonizing process. Through the expansion of trade and the predominance of the European continent in the fifteenth and sixteenth centuries, what Wallenstein has called 'the European world economy' (Wallerstein 1974, Cohen 1995) came about. In a systematic way, population movements became part of the general development of human societies under the aegis of the great empires.

The English, French, Dutch, Spanish and Portuguese had the greatest influence on the colonizing process. In turn, each colonial empire made its own mark on the new societies that emerged.

European colonizing emigration generally occurred on a voluntary basis, although governments took an active part in promoting it. In the British Isles emigration towards colonizing areas was systematically planned: the idea was to export people in order to consolidate imperial hegemony and also to solve social problems caused by overpopulation in the metropolitan territory.

The Spanish Empire went even further in conducting and controlling the process. The Laws of the Indies reflected a planning drive that went from the design of new cities to management of the emigrating groups, seeking to ensure that they included only Spanish subjects capable of proving their 'blood purity' (which meant that they were not descended from gypsies, moors, or converted Jews) (Mörner 1985). Among the consequences of this zeal is a register, kept at the Archive of the Indies, that has made it possible to estimate the total flow of emigrants in a reasonably reliable way. Mörner (1985) figures that there were approximately 450,000 emigrants from Spain between 1504 and 1650. This was a significant part of the total population of Spain (estimated at about 8 million around 1590), but its quantitative impact on the population of the Americas was of lesser importance. Throughout the colonial period, immigrants coming from Spain constituted a minority as compared with the indigenous population (Sanchez Albornoz 1994).

Spanish immigration was essentially made up by single men. During the first century of the Conquest women made up only 5 percent of the total that crossed the ocean; a century later they constituted 35 percent of the total. This settlement by unattached men quite naturally favored crossbreeding between the colonizers, American natives and Africans, so that the growth of a large *mestizo* and mulatto population is one of the features of Spanish and Portuguese America.

Estimates of emigration to the North American colonies are more nebulous, since immigrants had much more varied origins. Those from the British Isles were predominant, although many also came from Germany, Switzerland, France and the Netherlands (see *Migration out of Europe*).

In the eighteenth century British emigration came mainly from two population groups, differing both in geographical origin and demographic profile: the 'metropolitan' migration from London and the central region of England, and the stream from England's northern counties and from Scotland. Unattached men were a majority in the first group, mainly craftsmen and tradesmen in search of labor. Immigrants from Scotland and the northern counties of England were usually married couples with many children, mainly working on the land (Bailyn 1986).

The differences in colonizing style in the North and in the South gave rise to different types of societies. This would determine the subsequent development of the Americas.

2. Immigration from Outside the Americas

One of the predominant ideas among the elites that led the independence of the States of the Americas was to attract settlers to the new nations. Besides the need to have people to work and colonize empty territories, and to promote agriculture, they encouraged free trade and the forging of links between industrial centers and the new nations that produced raw materials.

Until the end of the nineteenth century the USA maintained an open-door policy, with only limited restrictions on immigration. Up to the mid-1900s the Federal Government took no measures to encourage the inflow of persons: instead manpower scarcity made the various states compete with each other to capture immigrants by financing their travel and offering them favorable conditions for the purchase of land. The promise of economic success and freedom radiated by the new republic led to an important inflow of immigrants, estimated at 250,000 from 1776 to 1820 (US INS, 1990), but only after 1830 did European immigration become a mass phenomenon.

In Central and South America demographic density varied greatly among subregions, but in general there was a shortage of manpower and attracting European immigrants was regarded as among the high priorities of the new Latin American Republics. From an ideological point of view, immigration schemes were based on the doctrinaire assumptions then in vogue in Europe; namely, that population volumes were equated with economic progress and military power. Additionally, human settlement helped to demarcate

the still vague frontiers of the new countries. The example of the USA and the success of immigration in that great republic of the North had a decisive influence on the leaders of the new republics of the South.

In Latin America, the intent to incorporate European migrants (preferably from Northern Europe) whom, besides their families and skills were expected to bring along a spirit of order and hard work was added to populationist arguments. More or less explicitly, the purpose was to 'upgrade the race,' an attempt by the ruling elites to fortify their ascendancy over the mulatto and *mestizo* masses, whose participation in the independence and civil wars had given them considerable autonomy and self-confidence.

Meanwhile the population of Europe was undergoing a transformation unthinkable in previous periods. The effects of the agricultural and industrial revolutions that swept the continent created great mobility towards the towns and cities and a consequent break-up of the links that tied peasants to the land and to their lifelong habits. Internal migrations went hand in hand with international emigration in search of new environments less restrictive of the realization of personal goals. The weakening of feudal bonds also meant the disintegration of traditional patterns of life and family models, and the appearance of new forms of labor relationships and social organization. Migration to the cities and emigration to America represented the extremes of this adventure which implied breaking away from the past and facing the future in unknown worlds.

European emigration was directed mainly to North American countries and, to a lesser extent, to some countries of the south, like Argentina, Uruguay and southern Brazil.

According to figures provided by Chesnais (1986 164) some 56 million people were involved in intercontinental emigration from 1821 to 1932. Sixty percent of them left for the USA, 22 percent for Latin America, 9 percent for Canada, and 9 percent for Australia, New Zealand and South Africa. Half of the 12 million whose destination was Latin America went to Argentina, 36 percent to Brazil, 6 percent to Uruguay and 7 percent to Cuba. Immigrants arrived in considerable numbers in the USA towards the end of the eighteenth century and the beginning of the nineteenth century. During most of the nineteenth century emigration to the USA originated mainly in the British Isles and the countries of northern and western Europe. From 1880 onward, immigration from southern and eastern Europe grew in volume. The presence of Asians on the Pacific coast increased gradually. These were Chinese at first and Japanese later. Resistance to this immigration gave rise to the first round of restrictions of immigration into the USA. The Immigration Act of 1882 set limits to the entry of the Chinese into US territory and barred the possibility of their obtaining US citizenship.

Throughout the nineteenth and twentieth centuries the USA was the largest receiver of immigrants. While European immigrants had predominated from colonial times to the mid 1900s, Chinese migration became increasingly important on the Pacific Coast. The first two decades of the twentieth century set a record in the number of immigrants. From 1900 to 1920, 14.5 million people were admitted, and in several years of that period arrivals exceeded one million persons (US INS 1990).

In Latin America, the struggle for independence and civil wars prevented immigration from developing until the second half of the nineteenth century. It became massive only in the last decade of that century and the initial decades of the twentieth century. The 1929 crisis halted European immigration, and it recovered for only a brief period after World War II.

Italians were predominant in the emigration movement towards Latin America until 1905, when the Spanish became the most numerous contingent. The Portuguese emigrated at first almost exclusively to Brazil and subsequently to the USA and Venezuela. People from the Middle East and Eastern Europe joined the migratory currents to North and South America since the late 1800s, and they grew in numbers in the 1920s. Colonies of German immigrants settled in several Latin American countries and wielded considerable influence in the south of Brazil and in Chile.

The scarcity of population in Argentina and Uruguay meant that immigrants were a significant factor of the population. Around 1860, some 30 percent of the population of those countries had been born abroad. In Brazil the highest proportion of inhabitants born outside the national territory (7.2 percent) was reached in 1900, although in the southern states of the country the scale of immigration was similar to that of Argentina and Uruguay.

In the late 1950s and more definitely in the 1960s European immigration towards the Americas ceased almost completely, both to the USA and to South America. This ended a trend that had persisted with great intensity for more than a century.

3. *Regional and Frontier Migrations and Mobility*

Intra-regional migration has been very important in the Americas. The causes for such movements have varied. Sometimes they occurred in areas where political borders resulting from the wars of independence divided communities with a common identity and a shared history. In other cases there were regions with different levels of demographic density, land, or manpower availability. In all cases those

frontiers were easily crossed and there were no major physical obstacles for population displacements.

The USA has received such migrants throughout its whole history, across both its northern and southern frontiers. Throughout the nineteenth century as well as in several periods of the twentieth century, there was continual immigration from Canada, consisting not only of Canadian natives but also of immigrants from other countries who preferred to continue their quest for more promising lands. Mexican migration to the USA was originally similar to the frontier movements in Latin American regions. Migration through frontier territories took place all the time, but during the Mexican Revolution (1910–1917) population movements towards the north became more important. The participation of the USA in World War I generated a demand for more workers and the first Bracero Program was implemented in 1917 and 1921 (Durand 1996). In 1942 a new program of this kind was carried out, intended to recruit workers for agriculture, the railroads and mining, and this became an important precedent to current Mexican immigration.

In the Caribbean, after the Slaves' Emancipation Acts (1838), mobility and emigration were a means to achieve freedom from the plantations and to escape the limitations they imposed (Hope 1996). During the nineteenth century, migration occurred mainly among the islands of the region that had sugar plantations (Cuba, Puerto Rico, Dominican Republic) and towards the banana plantations of Costa Rica. The construction of extensive infrastructural systems (the Central American Railroad and the Panama Canal) towards the end of the nineteenth century and the beginning of the twentieth century attracted immigrants from the Caribbean islands, as did oil exploitation in Venezuela and the Dutch Antilles (Aruba and Curaçao) some decades later (Hope 1996).

As early as in the 1930s and in a more marked way since the 1950s two things occurred in Latin America that bore directly upon the increase of migratory movements, both inside countries and across their borders.

First high rates of population growth were experienced (on average they reached a peak from 1955 to 1965) owing to decreased mortality and the prolonged persistence of high fertility rates.

Second, largely as a result of the crisis of the 1930s in the central countries, some Latin American nations shifted from an economic model based on the export of agricultural commodities to an 'inward growth' project of industrial development, initially intended to cover domestic needs. This scheme gathered new strength during World War II, but it developed unevenly. In some countries (Argentina, Brazil, Costa Rica, Chile, Uruguay, and Mexico) economies tended to become diversified through the increasingly important role of industry. In others development was slower or came later, but on the whole the region attained high economic growth rates, over and above those of industrialized countries, between 1950 and 1975.

As regards migration, the phenomenon of greatest quantitative significance and economic impact during those years was the increase in urbanization. People moved from rural areas into towns in great numbers with the result that urban populations swelled remarkably, particularly in metropolitan cities.

Until the 1960s, intra-regional migration was confined largely to movements between neighboring countries, and it could be taken as a prolongation of internal migration beyond state borders. Originally, such movements were mostly rural and usually temporary, such as seasonal transfers for harvesting or implementing other concrete tasks. As towns developed and agricultural transformation entailed rural–urban displacements, frontier migration also changed, replacing native populations in rural areas and joining the current towards the cities in the receiving countries.

One case in point is Argentina. Besides receiving European immigrants this country was also a center of attraction for migrants from neighboring countries. Those movements were confined to frontier regions in the beginning, but from the 1950s they shifted their target to urban areas, mainly to the zone of influence of the city of Buenos Aires, where the development of industry and services was concentrated.

In the period after World War II Venezuela also became a destination for both European and Latin American migrants. The Venezuelan government implemented policies to attract professionals and skilled workers from abroad for large investment projects. The rise in oil prices led to a remarkable increase of fiscal revenues and investments from 1974 to 1981. Latin American immigration came not only from neighboring countries but also from more distant areas. Although the economic situation of Venezuela changed in the 1980s and 1990s, Colombian immigration continued during the 1990s, due to the links forged in previous years and the violence prevailing in Colombia.

In the last few decades, the settlement of Brazilian peasants and rural workers (the 'Braziguayans'), in the Upper Paraná River, at the frontier between Paraguay and Brazil has been one of the latest examples in Latin America of rural population expansion through frontiers. This is a movement stemming from infrastructural construction and the increase in economic and political interchanges between Paraguay and Brazil.

Mexico and Costa Rica have been traditional recipients of frontier migrants (Colombians, Nicaraguans and Guatemalans). Since the mid-1970s the lack of stability and violence have been the cause of population movements: internally displaced persons, international migrants, refugees, people seeking a place within the region and people trying to find their

Table 1
Emigrants from Latin America and Caribbean countries in other countries in the Americas

	1960	1970	1980	1990	Annual growth rates (o/oo)		
					1960–1970	1970–80	1980–90
Total Latin American and Caribbean migrants in the Americas	1582489	3091632	6538914	11030846	69.3	77.8	53.7
To the USA	908309	1803970	4372487	8407837	71.0	92.6	67.6
Mexican to USA	575902	759711	2199221	4298014	28.1	112.1	69.3
Non-Mexican to USA	332407	1044259	2173266	4109823	121.3	76.0	65.8
To Canada[1]		82685	323415	523880		146.1	49.4
To other countries in Latin America & the Caribbean[2]	674180	1204977	1843012	2099129	59.8	43.4	13.1

1 For Canada, 1986 1996 Census, no data for 1960.
2 For 1960 8 countries; for 1970 20 countries; for 1980 19 countries; for 1990 18 countries.
Source: Pellegrino, A. Report on International Migration in Latin America (1999). Based on IMILA-CELADE National Census Data Base.

way to the USA have made up the migratory picture of the region. According to information gathered by the United Nations High Commissioner for Refugees, the displaced population in countries of the region reached 1.163 million persons at the beginning of 1990. Mexico, Costa Rica, Guatemala and Honduras, in that order, harbored the greatest numbers of refugees (Staton Russel 1995).

4. *Latin America, a Region of Emigration*

In the second half of the twentieth century, and particularly after the 1960s, migratory destinations for Latin Americans became more diversified. On the one hand, the streams towards the USA grew in volume, as did emigration from the Caribbean into Canada (Table 1). The oil crisis of the 1970s had shattered the economic development of Latin American countries. Those that were oil producers enjoyed a period of affluence that enabled them to increase investments and make their economies more dynamic. Others went into a permanent crisis that foreshadowed what would happen in the 1980s.

The decade of the 1980s, which the UN Economic Commission for Latin America and the Caribbean (ECLAC) has called 'the lost decade for development,' had its effects on international migration. Countries that had been traditional recipients of manpower migration (Argentina and Venezuela) saw immigration from neighboring nations slow down, and there was a new increase in the flow of Latin Americans towards the USA, and to a lesser extent also to Canada. In addition Latin Americans emigrated to Europe and Australia, though in significantly smaller numbers.

There has also been a process of return to the motherland of the children and grandchildren of European immigrants, mainly to Spain and Italy, whose governments implemented policies to promote the retrieval of nationals scattered over the world. By obtaining passports and the possibility of enjoying the rights of European citizenship, many Latin Americans have emigrated and recovered the nationality of their ancestors. In Brazil and Peru, descendants of the previously large generations of Japanese immigrants have also returned to Japan.

The Latin American population recorded in the USA grew from nearly 1 million in 1960 to almost 8.5 million in 1990. In Canada, the number of Latin Americans grew from 80,000 in 1970 to more than half a million in 1990. Changes in legislation in the receiving countries as well as the increasing emigration propensity in the countries of origin produced a change in geographical sources of immigration to North America. Latin America and Asia became the main source of immigrants during this period while European immigration decreased and

represented a much smaller proportion of the total.

Emigration from Latin America and the Caribbean to the USA reached its highest figures from 1960 to 1970, namely from Cuba, the Dominican Republic, Haiti, and the English-speaking Caribbean. Although Mexican migration has always been the strongest, its greatest thrust took place from 1970 to 1980. During the 1980s, the highest growth rates were from Central American countries.

Reasons for emigration to the USA vary, and, although in most cases they are the result of economic crises in the countries of origin, political violence has also played a decisive role. Political violence gave rise to emigration from Cuba, Haiti, and the Dominican Republic in the 1960s, as in the Southern Cone in the 1970s and in Central America in the 1980s. Moves originally due to political violence and persecution triggered off subsequent currents where economic and labor motives predominated.

Emigration to Canada has been considerably less than to the USA throughout, although it has increased remarkably in the last three decades. The English-speaking Caribbean, in particular Jamaica, Trinidad and Tobago, and Guyana, have been the main providers of immigrants. Canada had concluded special agreements with those countries to contract short-term workers. Haitians, who were next in numbers of migrants, went mainly to French Canada.

Latin American migration towards the USA is only part of a number of consequences of the hegemony of the USA in the Americas. Apart from the strengthening of economic ties that has taken place in the last decades of the twentieth century, the globalization of the mass media has meant not only greater access to information but also a homogenization of aspirations and values. Common expectations regarding ways and patterns of life have accentuated migratory potentials.

The existence of numerous and consolidated local communities has a feedback effect on migratory flows, even when initial causes lose their strength. Networks are built up through the ties that immigrants maintain with their families or friends in the colonies and countries of origin, and generate solidarity mechanisms that reduce the costs and risks of emigration.

Latin Americans have entered country-specific sectors of the labor market. For instance in the USA Mexican immigration has historically been oriented towards the agricultural sector in the southern states. In recent decades it has diversified into urban activities and services. As from the 1970s, changes in the US labor market have led to transformations in the occupational role of Latin American immigrants. Schematically they can be divided into two groups nowadays: namely (a) those who get highly qualified jobs in science and technology or in managerial circles; and (b) a more numerous group of workers in the area of social and personal services, where Latin American immigrants have traditionally been numerous.

See also: Colonization and Colonialism, History of; Emigration: Consequences for Social Structure in the Sending Population; Immigration; Immigration: Consequences for Fiscal Developments in the Receiving Population; Latin American Studies: Economics; Latin American Studies: Society; Migration, Economics of; Migration: Sociological Aspects; Migration, Theory of

Bibliography

Bailyn B 1988 *The Peopling of British North America. An Introduction*. The Curti Lectures. University of Wisconsin, WI

Body G S 1991 *Les Etats Unis et leurs immigrants*. Les etudes de la Documentation Française, Paris

Castles S, Miller M J 1993 *The Age of Migration. International Population Movements in the Modern World*. The Guilford Press, New York

Chesnais J C 1986 *La Transition Démographique*. Institut National d'Etudes Démographiques-Presses Universitaires de France, Cahier N°. 113, Paris

Cohen R (ed.) 1995 *The Cambridge Survey of World Migration*, Cambridge University Press

Durand J 1996 *Migrations mexicaines aux Etats Unis*. CNRS Editions, Paris

Ferenczi I, Willcox W F (eds.) 1929 *International Migrations*. 2 Vols. National Bureau of Economic Research, New York

Hope T E 1996 Emigration dynamics in the Anglophone Caribbean. *Policy Workshop on Emigration Dynamics in México, Central America and the Caribbean*. San José, Costa Rica, June 17th–18th

Kritz M, Lim L L, Zoltnik H (eds.) 1992 *International Migration Systems: A Global Approach*. Clarendon Press, Oxford, UK

Maingot A 1999 Emigration dynamics in the Caribbean: The cases of Haiti and Dominican Republic. In: Appleyard R (ed.) *Emigration Dynamic in Developing Countries: Mexico Central America and the Caribbean* Vol. 3. UNFPA-OIM, USA

Massey D 1988 International migration and economic development in comparative perspective. *Population and Development Review* **14**: 383–414

Massey D, Arango J, Hugo G, Kouaouci A, Pellegrino A, Taylor J E 1998 *Worlds in Motion. Undestanding International Migration at the End of the Millennium*. Clarendon Press, Oxford, UK

Mörner M 1985 *Adventurers and Proletarians. The Story of Migrants in Latin America*. University of Pittsburgh Press UNESCO, París

Oddone J A 1966 *La emigración europea al Río de la Plata*. Ediciones de la Banda Oriental, Montevideo, Uruguay

Pellegrino A 1989a *Historia de la Inmigración en Venezuela. Siglos XIX y XX*. Academia Nacional de Ciencias Económicas, Caracas, Venezuela

Pellegrino A 1989b *Migración Internacional de Latinoamericanos en las Américas*. Centro latinoamericano de Demografía de las Naciones Unidas (CELADE), Universidad Católica Andrés Bello, Agencia canadiense para el desarrollo internacional, Caracas

Pellegrino A 1999 *Report on International Migration in Latin America*. Based on IMILA-CELADE National Census Data Base

Portes A, Rumbaut R G 1990 *Immigrant America. A portrait*. University of California Press, CA

Sanchez Albornoz N 1994 *La población de América latina. Desde los tiempos precolombnos al año 2025*. Alianza Universidad, Madrid, Spain

Staton Russel S 1995 Migration patterns of U.S. Foreign Policy Interest. In: Teitelbaun M, Weiner M (eds.) *Threatened Peoples, Threatened Borders. World Migration and U.S. Policy*. An American Assembly Book, Norton, New York, pp. 39–87

U.S. Department of Justice Immigration and Naturalization Service (INS) 1991 Trends in immigration. In: *1990 Statistical Year Book of the Immigration and Naturalization Service*

Wallerstein I 1974 *The Modern World System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

A. Pellegrino

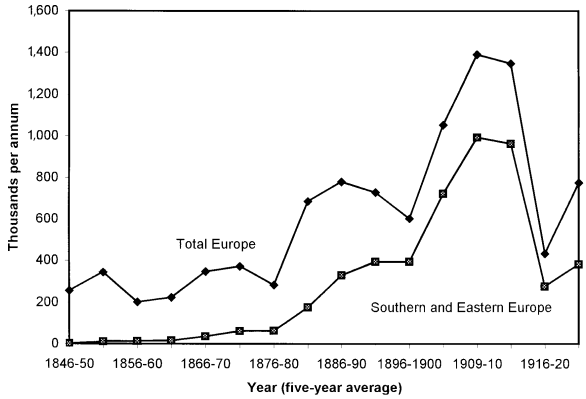

Figure 1
Emigration from Europe, 1846–1924 (five-year averages)

Migration out of Europe

International migration has been a key element in the development of the Western world since the industrial revolution. Its most profound influence has been on the peopling of the continents of North and South America and Australasia with emigrants of European stock and their descendants. This radical shift in population has been the focus of a wide range of studies by social scientists and historians. Much of the analysis addresses the questions of who migrated, when, and, above all, what motivated the migration. Much of the discussion has focused on what might be called 'the age of mass migration,' from the middle of the nineteenth century to World War I. One reason among many for concentrating on this period is that it is possible to observe migration behavior relatively free of the legal restrictions later introduced. Accordingly, the discussion here pays only brief reference to later periods.

1. Patterns of Migration

Early settlements of Europeans were progressively established in different parts of the New World from the sixteenth to the eighteenth century. Between 1820 and 1920 about 55,000,000 Europeans set sail for these resource-abundant and labor-scarce destinations, and about three-fifths of them went to the USA. Earlier migration had been a trickle by comparison. Some of the early migrants were pioneer settlers seeking to establish new communities free from the religious or political persecution they faced in Europe. There were also streams of convicts and indentured servants sent to work on frontier farms and plantations. These were rapidly overtaken by the mounting numbers of free migrants, but it was not until the middle of the nineteenth century that mass migration can really be said to have taken hold. Annual emigration rose from a steady flow of about 300,000 in the 1850s to 1870s, climbing steeply to exceed 1,000,000 at the turn of the century (see Fig. 1).

The emigrants in 1900 were certainly different from those in 1800. Early nineteenth-century migrant streams were often led by farmers and artisans from rural areas travelling in family groups, intending to acquire land and settle permanently at the New World's frontier. In the late nineteenth century, while many still had rural roots, the emigrants from any given country were increasingly drawn from urban areas and from nonagricultural occupations. About two-thirds were men and the overwhelming majority travelled as single individuals rather than in family groups. Of migrants to the USA between 1868 and 1910, 76 percent were between the ages of 15 and 40. While the young and single might be regarded as inherently more adventurous and more mobile, these characteristics also reflect a deeper economic calculus. By emigrating as young adults they were able to reap the gains over most of their working lives. By moving as individuals they were able to minimize the costs of the move, including earnings foregone during passage and job search. And since the emigrants were typically unskilled they also had little country- or technology-specific human capital invested and hence stood to lose few of the rents from such acquired skills (except language). Finally, these young adults had the smallest commitment to family and assets at home.

As mass migration mounted so the composition by country of origin changed. In the middle decades of the nineteenth century the emigrants came from North-western Europe, chiefly from Britain, Ireland, Germany and the Scandinavian countries. But as Fig. 1 shows, the great surge in migration after 1880 was contributed largely by the rise in emigration from the countries of Southern and Eastern Europe. Notable among these so-called 'new emigrant' countries were Italy, Spain, Austria-Hungary, and Russia. Statistics

such as these hide the enormous variations in *rates* of emigration (per thousand of the source population). The highest was from Ireland, with an average emigration rate of 13 per thousand per annum between 1850 and 1913. Countries such as Sweden and Norway had rates approaching five per thousand from 1870–1913, while the rates for Germany and Belgium were under two per thousand, and those for France were very small. Furthermore, the long term trends in emigration differed widely: from the 1880s rates of emigration declined sharply in Ireland, Germany, and Norway while they underwent a dramatic increase in Southern and Eastern Europe.

2. *The Determinants of Emigration*

Various hypotheses have been offered to explain variations in emigration across time and place, drawing on perspectives from economics, sociology, demography, and geography (Lowell 1987, Chap. 2). One important fact that such theories must explain is that, during the course of modern economic growth in Europe, national emigration rates often rose gradually at first from very low levels, rising more steeply to a peak and then gradually falling. Clearly, economic incentives were (and are) a key determinant of migration: in virtually every case where there is a net flow from one country to another it is from the relatively poor to the relatively rich country. New internationally-comparable real wage data for the late nineteenth century illustrate this. Average unskilled wage rates in 12 (Western) European countries were barely more than half those in the New World (weighted by proportion of emigrants). Recent analysis shows that wage ratios were an important determinant of emigration rates by decade and across countries. On average an increase in the New World real wage by 10 percent relative to that in Europe generated a rise in the emigration rate of 1.3 per thousand (Hatton and Williamson 1998, p. 43). But real wage ratios alone cannot explain why poor countries often had low emigration rates and why emigration often increased as development took place. Other influences must also be taken into account.

One factor is the poverty trap. As countries and regions industrialized, and as incomes rose, a larger share of potential migrants could afford to save up sufficient resources to finance the move and hence the number of emigrants increased. Thus, in its early stages, increases in emigration could be positively associated with an increase in wage rates at home (even though it narrowed the foreign/home wage gap). A second factor is rapid population growth, producing large cohorts of young adults whose opportunities to inherit smallholdings or to enter into family businesses were limited. Recent analysis of cross-country trends in emigration across Europe in the late nineteenth century finds strong evidence of these demographic effects and only limited evidence of the poverty trap (Hatton and Williamson 1998, p. 43). There is also some evidence that the more urban the country the higher the emigration—supporting the notion of greater mobility among urban populations.

3. *Persistence and Volatility in Migration*

Once established, migration streams perpetuated themselves as previous migrants provided new migrants with prepaid tickets for passage, food and shelter on arrival, and established immigrant networks to help gain access to job opportunities. This 'friends and relatives effect' proved to be a very powerful force in the late nineteenth century. Evidence from US immigration records suggests that as many as 90 percent of arrivals were traveling to meet a friend or relative who had previously emigrated. Recent estimates suggest that for each thousand migrants from a particular country, the friends and relatives effect 'pulled' between 20 and 100 further migrants each year, even when other factors such as relative wages are taken into account. Thus the larger the stock of previous emigrants, the greater would be the flow, and that flow would in turn lead to further flows.

Much of the literature has focused on this as the key element in determining mass migration. While social networks may have been the mechanism through which many individuals migrated, they also have an economic dimension. Not only did the friends and relatives effect reduce the costs and uncertainties of migration to the prospective migrant, it also eased the poverty trap. This also helps to explain different emigration patterns from countries at similar levels of development. In Ireland, the Great Famine effectively ejected a million migrants who formed a substantial migrant stock, particularly in the USA. Thus even the poorest Irish migrant would have benefited from the release of the poverty constraint, and Irish emigration declined as conditions in Ireland gradually improved. By contrast, in Italy emigration increased as the economy developed and the stock of previous emigrants grew.

Sharp year-to-year fluctuations in migration flows were often superimposed on these long-term trends. Some studies of annual movements in emigration suggest that business cycle conditions, especially in the destination country, were the key influence and that other influences (such as wages) were relatively unimportant. This has been interpreted to mean that migrants were driven by the 'pull' effects of opportunities in the New World, rather than the 'push' effects of conditions at home. Recent work has suggested that a specification along the lines of Todaro (1969), where both the wage and the probability of employment matter, can explain time series move-

ments in emigration reasonably well. The proportionately larger effect of unemployment, particularly in the destination, reflects risk aversion on the part of potential migrants. The year-to-year volatility of migration can also be interpreted as reflecting the option value of waiting. Thus, even though it may be worthwhile to emigrate now, it may be better still to wait until conditions in the receiving country have improved (Hatton 1995).

4. *Destination Choice and Return Migration*

In the past, emigration streams from a given country were often dominated by one destination, for example, Scandinavian emigrants went almost exclusively to the USA. Consequently, choice among alternative destinations is less well understood. Choice of destination *within* a receiving country is associated with measures of regional income as might have been expected, and especially with the stock of previous migrants to that state/region. But choice among countries involves additional factors such as cultural and linguistic affinity with the country of origin. Thus emigrants from Italy, Spain, and Portugal revealed much stronger preferences for South American countries such as Argentina and Brazil than did other European emigrants. Given these affinities and the pulling power of previous migrants, these streams persisted in spite of the income advantage of going instead to the USA. However, when new immigrant streams arose, such as that from southern Italy at the end of the nineteenth century, economic advantage carried more weight. Thus migrants from the north of Italy continued to favour South America over North America (despite their urban backgrounds) while those Italians from the rural south migrated in increasing numbers to the urban USA.

Although most migrants were permanent settlers, there were mounting flows of return migrants. By the end of the nineteenth century about a third of European migrants to the USA were returning, usually after a few years. Increasing destination wages relative to transport costs and falling voyage times contributed to the trend. But the upward trend in return migration owes most to the changing country composition of emigrants—particularly the growing share from Southern Europe. Many of these emigrants intended to return home and use their accumulated savings to establish families, and often to start farms or businesses. In such cases the outward flows were more male-dominated than where permanent settlement was the goal. While return migration strategies are not well understood one thing is clear: differences in rates of return migration are associated more with the country of origin than with the country of destination. Thus the high rates of return migration among southern Europeans at the turn of the century applied equally to those emigrating initially to North America and to South America.

5. *Immigrant Assimilation*

It has often been argued that immigrants faced discrimination and prejudice—factors which placed them under social and economic disadvantages and which may have made migration less attractive or encouraged return migration. The US Immigration Commission, which reported in 1911, argued that immigrants themselves, particularly the 'new immigrants' from Southern and Eastern Europe, were unable or unwilling to integrate into American society and that they lacked the skills and motivation to be successful in the labor market. Instead they crowded into ghettos and into unskilled occupations with little hope of upward mobility. Revisionist writers have argued instead that immigrant communities were not backward-looking and isolationist; rather, they provided a means though which immigrants could maintain their ethnic identity and benefit from social support networks at the same time as gaining access to the means of economic advancement (Bodnar 1985).

A key issue is whether immigrants did suffer economic disadvantage and whether they caught up with the native-born as they acquired knowledge, skills, and experience in the American labor market. The literature on immigrant earnings in the postwar period (following Chiswick 1978) suggests that, on arrival, immigrants suffered a wage disadvantage relative to the native-born but that their wages grew faster and caught up with the native-born after 10 to 15 years. Recent analyses for immigrants at the turn of the century suggest that (contrary to an earlier view) a similar pattern can be found among immigrants who arrived as adults (Hatton and Williamson 1998, Chap. 7). Those who arrived as children suffered little or no disadvantage and second-generation immigrants often had higher earnings than the native-born. While the 'new immigrants' faced some economic disadvantages, much of this can be associated with limited education and English language proficiency—disadvantages which diminished over time.

6. *The End of the Age of Mass Migration*

Migration decisions were also influenced by policy and/or prejudice towards immigrants in the receiving country. Sometimes this worked to the advantage of the immigrant, for example, subsidized passages were offered to British emigrants to Australia and to (northern) Italian emigrants to the state of Sao Paolo (Brazil). Sometimes migrants were discouraged by legal or administrative obstacles (such as non-British

migrants to Australia before 1945). But immigration restrictions grew gradually in the early twentieth century as immigration itself increased and became associated with growing inequality. The US Immigration Acts of 1921 and 1924, which introduced immigration quotas, are often seen as putting an abrupt end to the age of mass migration. However, it is likely that immigration from some European countries (those on the downswing of the emigration cycle) would have diminished anyway. The world depression of the 1930s, which was particularly severe in the New World, discouraged immigrants from most European countries except those fleeing totalitarian regimes.

In the period following World War II, emigration revived and grew as prosperity returned to the world economy. But it did not match the heights of the decade before World War I. One reason is that migration chains, which had been such an important factor in the earlier period, had been broken by 30 years of war and economic upheaval. Emigration from Europe grew over the following decades but not as fast as it had in the 50 years before 1914. This was partly due to continuing restrictions on immigration, but it also reflects the rapid growth of living standards in Europe and their convergence on those of destination countries such as the USA. As a result, Europeans declined as a proportion of immigrants into the USA, from 66 percent in the 1950s to 10 percent in the 1980s. This period also marked the transition of Europe from a region of emigration to one of immigration. At the same time countries in Asia, Africa, and the Caribbean entered a phase comparable with that of Europe in the age of mass migration.

See also: Colonialism, Anthropology of; Colonialism: Political Aspects; Colonization and Colonialism, History of; Historical Demography; Immigration; Immigration and Migration: Cultural Concerns; Migration and Health; Migration, Economics of; Migration into the Americas and between American Countries; Migration: Sociological Aspects; Migration, Theory of; Migrations, Colonizations, and Diasporas in Archaeology; Population Pressure, Resources, and the Environment: Industrialized World

Bibliography

Bodnar J 1985 *The Transplanted: A History of Immigrants in Urban America*. University of Indiana Press, Bloomington, IN

Chiswick B R 1978 The effect of Americanisation on the earnings of foreign-born men. *Journal of Political Economy* **86**: 897–921

Hatton T J 1995 A model of UK emigration, 1870–1913. *Review of Economics and Statistics* **7**: 407–15

Hatton T J, Williamson J G 1998 *The Age of Mass Migration*. Oxford University Press, New York

Lowell B L 1987 *Scandinavian Exodus: Demography and Social Development of 19th Century Rural Communities*. Westview Press, Colorado, CO

Todaro M P 1969 A model of labor migration and urban unemployment in less developed countries. *American Economic Review* **59**: 138–48

T. J. Hatton

Migration: Sociological Aspects

Few people today spend their whole lives in their native village or neighborhood: most experience mobility from country to town, or between regions in one country, while a minority migrates across national borders. Even those who do not migrate are affected by movements of people in or out of their communities, and by the resulting changes. Migration is an important factor in the erosion of traditional boundaries between languages, cultures, ethnic groups, and nation-states. Migration is not a single act of crossing a border, but rather a lifelong process that affects all aspects of the lives of those involved. The outcome may be temporary residence abroad followed by return, or permanent settlement in the new country. The latter may lead to complete integration into the receiving population, formation of ethnic minorities which remain distinct from the majority population, or emergence of transnational communities (or diasporas) that have close links with members of the same ethnic groups in the country of origin as well as in other migrant-receiving countries. In view of its all-embracing character, migration studies is an interdisciplinary field of study, to which all the social sciences make important contributions. However, this article will concentrate on sociological aspects of migration.

1. Definitions and Categories

Migration means crossing the boundary of a political or administrative unit for a certain minimum period (Boyle et al. 1998, Chap. 2). Internal migration refers to a move from one area (a province, district, or municipality) to another within one country. International migration means crossing the frontiers that separate one of the worlds' approximately 200 states from another. Some scholars argue that internal and international migration are part of the same process, and should be analyzed together (Skeldon 1997, pp. 9–10). However, this article focuses specifically on international migration, because of its links to globalization and its significance in creating multi-ethnic societies.

Table 1
Migrant population by region, 1965 and 1990[a]

	Estimated foreign-born population					
	Millions		As percentage of total population of region		As percentage of migrant stock world total	
Region	1965	1990	1965	1990	1965	1990
World total	75.2	119.8	2.3	2.3	100.0	100.0
Developed countries	30.4	54.2	3.1	4.5	40.4	45.3
Developing countries	44.8	65.5	1.9	1.6	59.6	54.7
Africa	7.9	15.6	2.5	2.5	10.6	13.1
Asia	31.4	43.0	1.7	1.4	41.8	35.9
Latin America and the Caribbean	5.9	7.5	2.4	1.7	7.9	6.2
Northern America	12.7	23.9	6.0	8.6	16.9	20.0
Europe and the former USSR	14.7	25.1	2.2	3.2	19.6	20.9
Oceania	2.5	4.7	14.4	17.8	3.3	3.9

a Adapted from Zlotnik H 1999 Trends of international migration since 1965: what existing data reveal. *International Migration* **37**: 21–62, Tables 1a and 1b.

The great majority of border crossings do not imply migration: most travelers are tourists or business visitors who have no intention of staying for long. Migration means taking up residence for a certain minimum period. Most countries have a number of official migration categories. For instance, Australia distinguishes between permanent immigrants, long-term temporary immigrants who stay at least 12 months, and short-term temporary visitors. However, Australia is a 'classical immigration country' with a tradition of nation-building through immigration, so public debate concentrates on permanent immigrants, who are expected to settle and become citizens. Other countries prefer to see immigration as essentially temporary. When Germany recruited so-called 'guest-workers' from the late 1950s to the early 1970s, they only received one-year residence permits. In time, it became difficult to limit residence and migrants were granted permits for two years, then five years, and finally for unlimited residence. Similarly, the policy-makers of labor-recruiting countries in Asia or the Middle East today do not want foreign workers to stay permanently, but may find this difficult to prevent in the long run.

Such variations show that there is nothing objective about definitions of migration: they are the result of state policies, introduced in response to political and economic goals and public attitudes. International migration arises in a world divided into nation-states, in which remaining in the country of birth is still seen as a norm and moving to another country as a deviation. One way in which states seek to improve control is by dividing migrants into categories.

(a) Temporary labor migrants, men and women who migrate for a limited period in order to take up employment and send home money (remittances).

(b) Highly skilled and business migrants, who have qualifications as managers, executives, professionals, technicians, or similar.

(c) Irregular migrants (also known as undocumented or illegal migrants), who enter a country, usually in search of employment, without the necessary documents.

(d) Refugees, defined by the 1951 UN Convention as persons residing outside their country of nationality, who are unable or unwilling to return because of a 'well-founded fear of persecution on account of race, religion, nationality, membership in a particular social group, or political opinion.'

(e) Asylum-seekers, people who move across borders in search of protection, but who may not fulfill the strict criteria laid down by the 1951 UN Convention.

(f) Family reunion, migration to join relatives who have already entered an immigration country under one of the above categories.

(g) Return migrants, people who return to their countries of origin after a period in another country.

2. *The Volume of Contemporary Migration*

Since the Second World War, international migration has grown considerably. Two main phases can be distinguished. The first lasted from 1945 to 1973: the long boom stimulated large-scale labor migration to Western Europe, North America, and Oceania from less-developed areas. This phase ended around 1973, with the 'Oil Crisis,' which triggered a major recession. In a second phase from the mid-1970s, capital investment shifted away from the old centers, and

transnational forms of production and distribution reshaped the world economy. The older industrial countries experienced new types of inflows, while new immigration countries emerged in Southern Europe, the Gulf oil countries, Latin America, Africa, and Asia. The late 1980s and early 1990s were a period of unprecedented migration (Castles and Miller 1998).

According to UN figures (Table 1), the global migrant stock (the number of people resident in a place outside their country of birth) grew from 75 million in 1965 to 120 million in 1990. International migration appears to have grown more rapidly in the 1990s, reaching an estimated 135–140 million people, including some 13 million United Nations High Commissioner for Refugees-recognized refugees by 1997. Nonetheless, international migrants make up only about 2 percent of the world's population (Zlotnik 1999).

However, migration is concentrated in certain countries and regions. The UN study shows that 90 percent of the world's migrants were living in just 55 countries. In absolute numbers, most migration is between less-developed countries, but in relative terms, the developed world has been more affected by immigration. The 1990 immigrant share in total population was highest in Oceania (17.8 percent) followed by North America (8.6 percent) and Western Europe (6.1 percent). The immigrant share in population was far lower in Asia (1.4 percent), Latin America and the Caribbean (1.7 percent), and Africa (2.5 percent) (Zlotnik 1999). In the 1980s and 1990s, flows from less-developed to developed countries have been grown rapidly, despite attempts by receiving countries to restrict such movements. In addition, there have been large flows of labor migrants from the least developed countries of the South to the newly industrializing countries, especially in East Asia. Although women have always formed a large proportion of migrants, their share has gradually increased: by 1995 about 48 percent of all international migrants were women, and they outnumbered male migrants in about a quarter of receiving countries (Zlotnik 1999). There was a shift in the character of female migration, with a trend away from movement as family members of male workers or refugees and an increase in the number of women who moved independently or as heads of households (Lutz et al. 1995).

3. *Causes of Migration*

The causes of migration are highly complex, and various social sciences explain them in their own ways (Boyle et al. 1998, Massey et al. 1993). Economists focus on disparities in levels of income and employment between sending and receiving areas, while demographers examine differences in fertility, mortality, age-structure, and labor-force growth. However, the mere existence of disparities does not always lead to mobility: the very poor rarely migrate, and there are many areas with huge reserves of underemployed people who remain where they are. Sociologists therefore focus on two sets of causes, which may be seen as macro- and microfactors. The former refers to the role of powerful institutions—states, corporations, markets, and international organizations—in initiating and regulating migration, or in putting up barriers against it. For instance, temporary labor migration to Germany in the 1960s and the Gulf oil states in the 1980s was organized by states and employers for economic reasons. Attempts to restrict labor migration and asylum-seeker movements to Western Europe in the 1990s were also state actions involving a fair degree of intergovernmental collaboration.

Microfactors refer to the social networks developed by migrants and their communities (Boyd 1989). Networks based on family or on common place-of-origin help provide shelter, work, assistance with bureaucratic procedures, and support in personal difficulties. Access to migration networks can be seen as a form of 'social capital,' a resource that makes it possible for migrants and their families to face the challenges of displacement and sometimes-hostile environments. Some people (both migrants and non-migrants) become facilitators of migration. A 'migration industry' emerges, consisting of recruitment organizations, lawyers, agents, smugglers, and other middle-people. Such people can be both helpers and exploiters of migrants. The strong interest of the migration industry in the continuation of migration has often confounded government efforts to control movements.

A useful approach is 'migration systems theory' which analyses the linkages between macro- and microfactors (Kritz et al. 1992). Typically, migratory chains are started by an external factor, such as recruitment or military service, or by an initial movement of young (usually male) pioneers. The 'cultural capital' needed to embark on migration can also be provided by access to education or international media, which raise awareness of opportunities elsewhere. Once a movement is established, the migrants follow established routes and are helped by relatives and friends already in the area of immigration. The connections between migrant community and area of origin may persist over generations. Remittances fall off and visits home may decline in frequency, but familial and cultural links remain. People stay in touch with their area of origin, and may seek marriage partners there. Migration continues along the established chains—and may increase dramatically at a time of crisis.

4. *Migration and Development*

The most important question for countries of origin is whether migration assists or hinders development.

Migration may hinder development by siphoning of qualified personnel (the 'brain drain'), removing dynamic young workers and reducing pressures for social change. Migration often involves a transfer of the most valuable economic resource—human capital—from a poor country to a rich one. It is only worthwhile for the emigration country if the gain in human capital (enhanced skills and productivity) through working abroad can be productively utilized upon return and the transfer of income from immigration to emigration country outweighs the costs of upbringing of the migrant.

Labor-exporting countries often pursue short-term aims, concerned with generating jobs for an underutilized workforce and with getting the maximum possible inflow of worker remittances (Abella 1995). Global migrant remittances increased from US$2 billion in 1970 to US$70 billion in 1995 (Taylor 1999). Many countries therefore encourage emigration for employment. This may mean government involvement in recruitment and deployment of workers, regulation of nongovernmental recruitment agencies, or simply *laissez-faire* with regard to spontaneous movements. Most emigration-country governments have policies to prevent abuse or exploitation of their citizens while abroad, and to provide assistance in case of illness, accident, death, trouble with the law, disputes with employers or other emergencies. However, regulation of emigration from less-developed countries is often ineffective, as the large number of irregular migrants demonstrates. This allows exploitative employment and abuses like the trafficking of women and children for prostitution.

There is a lack of coordinated strategies to assisting returning migrants with re-integration. Most migrants are left to their own devices and frequently face difficulty in finding employment commensurate with the skills they have acquired abroad. They may end up running small unproductive businesses which often fail. Savings may be spent on consumption and dowries rather than investment. Research indicates that adequate counseling and information both before and after return, as well as help in obtaining investment credits are factors conducive to successful reinsertion and maximization of positive effects on development. Maintenance of social networks in the home country is crucial for a successful return.

5. *Settlement and Ethnic Diversity*

For receiving countries, the key question is whether immigration will lead to settlement, formation of ethnic communities, and new forms of ethnic and cultural diversity. For instance Gulf oil countries do not allow family reunion and settlement, yet their economies are structurally dependent on foreign labor. This is leading to increased length of stay and family formation, despite the rules. Similarly, there is evidence of settlement and emergence of ethnic neighborhoods in Japan and other Asian labor-importing countries (Mori 1997). Migration generally leads to settlement of a certain proportion of the migrants due to the social networks mentioned above. Another factor is the increasing strength of human rights safeguards in many countries, which make it difficult for governments to deport migrants or to deny them the right to live with their families.

Immigrants often differ from the receiving populations: they may come from different types of societies (for example, agrarian–rural rather than urban–industrial) with different traditions, religions and political institutions. They often speak a different language and follow different cultural practices. They may be visibly different, through physical appearance (skin color, features, and hair type) or style of dress. Some migrant groups become concentrated in certain types of work (generally of low social status) and live in low-income residential areas. The position of immigrants is often marked by a specific legal status: that of the foreigner or noncitizen. In many cases, immigration complicates existing ethnic or racial divisions in societies with long-standing minorities.

Culturally distinct settler groups almost always maintain their languages and some elements of their homeland cultures, at least for a few generations. Where governments have accepted permanent settlement, there has been a tendency to move from expectations of individual assimilation to recognition of cultural difference. The result has been the policies of pluralism or multiculturalism introduced in various forms in North America, Oceania and parts of Western Europe since the 1970s. Where governments refuse to recognize right to community formation and cultural difference, immigrants tend to turn into marginalized ethnic minorities.

At a time of economic restructuring and far-reaching social change, some groups in the receiving populations may see immigrants as a danger to living standards, life styles and social cohesion. In Europe, extreme-right parties have grown and flourished through anti-immigrant campaigns. Similarly, one reaction to the Asian crisis of 1997–9 was to blame immigrants for unemployment and other social ills, and to introduce policies for large-scale deportations. The overall experience since the 1950s is that immigration almost always leads to cultural changes, which may be perceived as threatening. The result is often a politicization of issues connected with migration and the emergence of conflicts that may take many years to resolve.

6. *Migration, National Identity, and Citizenship*

If, as Albrow (1996) argues, the age of modernity is being replaced by a 'global age,' it seems that in-

ternational migration is even more crucial in the new epoch than in the preceding one. This is not surprising: if the central mechanisms of globalization are cross-border flows and transnational networks (Castells 1996, Held et al. 1999), then flows of people are clearly as important as flows of finance, commodities and ideas. However, while states generally welcome these other types due to their economic benefits, they are often suspicious of migrants, whom they see as a threat to national culture and identity, and hence as a major factor challenging the nation-state.

The nation-state, as it has developed since the eighteenth century, is often based on myths of ethnic and cultural homogeneity. Immigration and ethnic diversity threaten such ideas, and the emergence of multicultural societies creates major challenges to national identities. Fundamental institutions, such as citizenship itself, are likely to change in response to diverse values and needs. The trend to development of transnational communities is a further challenge to the nation-state: modern forms of transport and communication make it possible for immigrants and their descendants to maintain long-term links with the ancestral homeland or with diaspora groups elsewhere (Basch et al. 1994, Cohen 1997).

The classical countries of immigration have been able to cope with this situation most easily, since absorption of immigrants has been part of their myth of nation building. But countries that place a common culture at the heart of their nation-building process have found it very difficult to resolve the contradiction. This applies to many European countries, but also to many postcolonial nation-states in other continents. Asian states have tended to adopt quite restrictive rules on naturalization and citizenship, and find it very difficult to accept the possibility of integrating new immigrant populations (Castles and Davidson 2000, Davidson and Weekley 1999). However, migration continues to grow for economic and cultural reasons, and is likely to remain a potent force for social transformation in the future.

See also: Assimilation of Immigrants; Crime and Ethnicity (Including Race); Cultural Assimilation; Development and the State; Development: Socioeconomic Aspects; Globalization and World Culture; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Immigrants: Economic Performance; Immigration and Migration: Cultural Concerns; Immigration: Public Policy; Migration and Health; Migration, Economics of; Migration History; Migration, Theory of; Multiculturalism: Sociological Aspects

Bibliography

Abella M I 1995 Policies and Institutions for the orderly movement of labour abroad. In: Abella M I, Lönnroth K J (eds.) *Orderly International Migration of Workers and Incentives to Stay: Options for Emigration Countries*. International Labour Office, Geneva, Switzerland

Albrow M 1996 *The Global Age*. Polity Press, Cambridge, UK

Basch L, Glick-Schiller N, Blanc C S 1994 *Nations Unbound: Transnational Projects, Post-Colonial Predicaments and Deterritorialized Nation-States*. Gordon and Breach, New York

Boyd M 1989 Family and personal networks in migration. *International Migration Review* **23**: 638–70

Boyle P, Halfacree K, Robinson V 1998 *Exploring Contemporary Migration*. Longman, Harlow, UK

Castells M (ed.) 1996 *The Rise of the Network Society*. Blackwells, Oxford, UK

Castles S, Davidson A 2000 *Citizenship and Migration: Globalisation and the Politics of Belonging*. 2nd edn. Macmillan, London

Castles S, Miller M J 1998 *The Age of Migration: International Population Movements in the Modern World*. Macmillan, London

Cohen R 1997 *Global Diasporas: An Introduction*. UCL Press, London

Davidson A, Weekley K (eds.) 1999 *Globalization and Citizenship in the Asia-Pacific*. Macmillan, London

Held D, McGrew A, Goldblatt D, Perraton J 1999 *Global Transformations: Politics, Economics and Culture*. Polity Press, Cambridge

Kritz M M, Lin L L, Zlotnik H (eds.) 1992 *International Migration Systems: A Global Approach*. Clarendon Press, Oxford, UK

Lutz H, Phoenix A, Yuval-Davis N 1995 Introduction: nationalism, racism and gender. In: Lutz H, Phoenix A, Yuval-Davis N (eds.) *Crossfires: Nationalism Racism and Gender in Europe*. Pluto Press, London, pp. 1–25

Massey D S, Arango J, Hugo G, Kouaouci A, Taylor J E, Pellegrino A 1993 Theories of international migration: a review and appraisal. *Population and Development Review* **19**: 431–66

Mori H 1997 *Immigration Policy and Foreign Workers in Japan*. Macmillan, London

Skeldon R 1997 *Migration and Development: A Global Perspective*. Addison Wesley Longman, London

Taylor J E 1999 The new economics of labour migration and the role of remittances in the migration process. *International Migration* **37**: 63–88

Zlotnik H 1999 Trends of international migration since 1965: what existing data reveal. *International Migration* **37**: 21–62

S. Castles

Migration, Theory of

Social scientists have theorized about four basic aspects of human migration in their efforts to explain it: the structural forces that promote 'out-migration' from sending regions; the structural forces that attract 'in-migrants' to receiving societies; the motivations, goals, and aspirations of people who respond to these

structural forces by becoming migrants; and the social and economic structures that arise to connect areas of out- and in-migration. The ensuing sections review the theoretical perspectives that address these issues, and then a theoretical synthesis is developed that includes empirically-supported propositions from each one.

1. Neoclassical Economics

The oldest and best-known theoretical model, neoclassical economics, argues that migration is caused by geographic differences in the supply of and demand for labor (Lewis 1954, Ranis and Fei 1961). A region with a large endowment of labor relative to capital will have a low equilibrium wage, whereas an area with a limited endowment of labor relative to capital will be characterized by a high market wage. The resulting wage gap causes workers from the low-wage or labor-surplus area to move to the high-wage or labor-scarce region. As a result of this movement, the supply of labor decreases and wages eventually rise in the capital-poor area, while the supply of labor increases and wages ultimately fall in the capital-rich area, leading, at equilibrium, to a geographic wage differential that exactly reflects the costs of inter-regional movement, pecuniary and psychic.

Associated with this macroeconomic theory is an accompanying microeconomic model of individual choice (Todaro 1976). In this scheme, rational actors decide to migrate because a cost-benefit calculation leads them to expect a positive net return, usually monetary, from movement. Migration is conceptualized as a form of investment in human capital (Sjaastad 1962). People choose to move where they can be most productive, given their skills; but before they can reap the higher wages associated with greater labor productivity they must invest in the material costs of travel, the costs of self-support while moving and looking for work, the effort involved in learning a new environment and possibly a different culture, the difficulty experienced in adapting to a new labor market, and the psychological costs of cutting old ties and forging new ones.

Potential migrants estimate the costs and benefits of moving to alternative locations and migrate to where the expected discounted net returns are greatest over their projected working lives. Net future returns are estimated by taking earnings observed in the destination area and multiplying them by the probability of obtaining a job there to derive 'expected destination earnings,' which are then subtracted from earnings estimated for the place of origin (observed earnings times the probability of employment). The difference is then summed over future years and discounted by a factor that reflects the greater utility of money earned in the present than in the future. From this integrated difference the estimated costs are subtracted to yield the expected net returns to migration. In theory, a potential migrant goes to wherever the expected net returns are greatest.

2. The New Economics of Migration

In recent years, a 'new economics of labor migration' has arisen to challenge the assumptions and conclusions of neoclassical theory (Stark 1991). A key insight of this approach is that migration decisions are not made by isolated individuals, but within larger units of interrelated people—typically families or households but sometimes entire communities—and that people act collectively not only to maximize expected income, but also maximize status within an embedded hierarchy, to overcome barriers to capital and credit, and to minimize risk.

In most developed countries, risks to household income are managed through institutional mechanisms. Crop insurance programs and futures markets give farmers a way of protecting themselves from natural disasters and price fluctuations, whereas unemployment insurance and government transfer programs protect workers against recessions and job loss through structural transformation. Private or government-sponsored retirement programs, meanwhile, offer citizens a means of insuring against the risk of poverty in old age.

In the absence of such programs, households are in a better position than individuals to control risks to economic well-being. They can easily diversify sources of income by allocating different family workers to different labor markets. As long as economic conditions across labor markets are negatively or weakly correlated, households minimize risk by geographically diversifying their labor portfolios. In the event that economic conditions at home deteriorate and productive activities there fail to generate sufficient income, the household can rely on migrant remittances for support.

Markets for credit and capital are incomplete or inaccessible in many settings, and in the absence of an efficient banking system, migration becomes attractive as a strategy for accumulating funds that can be used in lieu of borrowing. Families send one or more workers to a higher wage area to accumulate savings or send them back in the form of remittances. Although most migrant savings and remittances go toward consumption, some of the funds may also be channeled into productive investment.

A key proposition of the new economic model is that income is not a homogeneous good. The source of the income really matters, and households have significant incentives to invest scarce family resources in activities that provide access to new income sources, even if these activities do not increase total income. The new economics of migration also questions the

assumption that income has a constant effect on utility—i.e., that a $100 real increase in income means the same thing to a person regardless of community conditions and irrespective of his or her position in the local income distribution. It holds that households send workers abroad not only to improve absolute incomes, but also to increase them *relative* to others, and, hence, to reduce *relative* deprivation compared with some reference group (Stark 1991). A household's sense of relative deprivation depends on the total income earned by households above it in the income distribution. If utility is negatively affected by relative deprivation, then even if a household's income and expected gains from migration remain unchanged, it may acquire an incentive to migrate if there is a change in *other* households' incomes.

3. Segmented Labor-market Theory

Although neoclassical theory and the new economics of labor migration offer divergent explanations for the origins of migration, both are micro-level decision models. What differs are the units assumed to make the decision (the individual or the household), the entity being maximized or minimized (income vs. risk), the assumptions about the economic context of decision making (complete and well-functioning markets versus missing or imperfect markets), and the extent to which the migration decision is socially contextual (whether income is evaluated in absolute terms or relative to some reference group).

Standing distinctly apart from these models is segmented labor-market theory, which argues that migration stems from the intrinsic characteristics built into modern industrial society. According to Piore (1979), migration is not caused by push factors in sending regions (low wages or high unemployment), but by pull factors in receiving areas (a chronic and unavoidable need for migrant workers). The built-in demand for inexpensive and flexible workers stems from three basic features of advanced industrial economies.

The first is *structural inflation*. Wages not only reflect conditions of supply and demand; they also confer status and prestige, social qualities that inhere in jobs. In general, people believe that wages should reflect social status, and they have clear notions about the correlation between occupational status and pay. As a result, employers are not entirely free to respond to changes in the supply of workers. A variety of informal social expectations and formal institutional mechanisms (union contracts, civil service rules, bureaucratic regulations, company job classifications) ensure that wages correspond to the hierarchies of prestige and status that people perceive and expect.

The demand for cheap, flexible labor is also augmented by *social constraints on motivation* within occupational hierarchies. Most people work not only to generate income, but also to accumulate social status. Acute motivational problems arise at the bottom of any job hierarchy because there is no status to be maintained and there are few avenues for upward mobility. The problem is structural because the bottom cannot be eliminated from the labor market. Mechanization to eliminate the lowest and least desirable class of jobs will simply create a new bottom tier composed of jobs that used to be just above the bottom rung. Since there must always be a bottom of any hierarchy, motivational problems are inescapable.

Thus, employers need workers who view bottom-level jobs simply as a means to the end of earning money, and for whom employment is reduced solely to income, with no implications for status or prestige. For a variety of reasons, migrants satisfy this need. They begin as target earners, seeking to earn money for a specific goal that will improve their status or well-being at home—building a house, paying for school, buying land, acquiring consumer goods. Moreover, the disjuncture in living standards between origin and destination communities often means that low urban or foreign wages appear to be generous by the standards of the sending society.

The demand for migrant labor also stems from *economic dualism*. Capital is a fixed factor of production that can be idled by lower demand but not laid off; owners of capital must bear the costs of its unemployment. Labor is a variable factor of production that can be released when demand falls, so that workers bear the costs of their own unemployment. Whenever possible, therefore, capitalists seek out the stable, permanent portion of demand and reserve it for the employment of equipment, whereas the variable portion of demand is met by adding labor. Thus, capital-intensive methods are used to meet basic demand, and labor-intensive methods are reserved for the seasonal, fluctuating component. This dualism creates distinctions among workers, leading to a bifurcation of the labor force.

Workers in the capital-intensive primary sector get stable, skilled jobs working with the best equipment and tools. Employers are forced to invest in these workers by providing specialized training and education. Their jobs are complicated and require considerable knowledge and experience to perform well, leading to the accumulation of firm-specific human capital. Primary-sector workers thus tend to be unionized or professionalized, with contracts that require employers to bear a substantial share of the costs of their idling (severance pay and unemployment benefits). Because of these costs and continuing obligations, workers in the primary sector become expensive to let go; they become more like capital.

In the labor-intensive secondary sector, however, workers hold unstable, unskilled jobs; they may be laid off at any time with little or no cost to the employer. Indeed, the employer will generally lose money by

retaining workers during slack periods. During down cycles the first thing secondary-sector employers do is to cut payroll. As a result, employers force workers in this sector to bear the costs of their unemployment. They remain a variable factor of production and are, hence, expendable.

The inherent dualism between labor and capital extends to the labor force in the form of a segmented market structure. Low wages, unstable conditions, and the lack of reasonable prospects for mobility in the secondary sector make it difficult to attract local workers, who are instead drawn into the primary, capital-intensive sector, where wages are higher, jobs are more secure, and there is a possibility of occupational improvement. To fill the shortfall in demand within the secondary sector, employers turn to migrants.

In their analysis of the process by which Cuban immigrants were incorporated into the United States, Portes and Bach (1985) uncovered evidence of a third possible sector that blends features of primary and secondary labor markets. Like the secondary sector, ethnic enclaves contain low-status jobs characterized by low pay, chronic instability, and unpleasant working conditions, jobs that are routinely shunned by natives. Unlike the secondary sector, however, they provide immigrants with significant economic returns to education and experience, as well as the very real prospect of upward socioeconomic mobility, thus replicating features of the primary sector.

The existence of a large, concentrated ethnic population creates a demand for specialized cultural products and services that immigrant entrepreneurs are uniquely qualified to fill. In addition, privileged access to the growing pool of migrant labor gives them an advantage when competing with mainstream firms. Migrants working in the enclave trade low wages and the acceptance of strict discipline upon arrival for a greater chance of advancement and independence later on. In order to function effectively over time, therefore, an ethnic enclave requires a steady stream of new arrivals willing to trade low initial wages for the possibility of later mobility, yielding an independent segmented source of labor demand for migrant workers, complementing that emanating from the secondary sector.

4. *World-systems Theory*

Growing out of the historical–structural tradition in social science, world-system theory argues that migration stems from the penetration of capitalist economic relations into non-capitalist or pre-capitalist areas to create a mobile population (Fligstein 1979 Sassen 1988). Driven by a desire for higher profits and greater wealth, owners and managers of capitalist firms in core regions enter poor peripheral areas in search of land, raw materials, labor, and consumer markets. Migration emerges in response to the disruptions and dislocations that inevitably occur in the process of capitalist development.

In order to achieve the greatest profit from existing agrarian resources, and to compete within global commodity markets, capitalist farmers seek to consolidate landholding, mechanize production, introduce cash crops, and apply industrially produced inputs such as fertilizer, insecticides, and high-yield seeds. Land consolidation destroys traditional systems of land tenure based on inheritance and common ownership. Mechanization decreases the need for manual labor and makes many agrarian workers redundant. The substitution of cash crops for staples undermines traditional social and economic relations based on subsistence; and the use of modern inputs produces high crop yields at low unit prices. All of these forces drive peasant farmers out of local markets and create a mobile labor force of people displaced from the land.

The extraction of raw materials for sale on global markets requires industrial methods that rely on paid labor. The offer of wages to former peasants undermines traditional forms of social and economic organization based on norms of reciprocity and fixed role relations, and creates incipient labor markets based on new conceptions of individualism, private gain, and social change. At the same time, capitalist firms enter developing regions to establish assembly plants, often within special export-processing zones created by sympathetic governments. The demand for factory workers strengthens local labor markets and weakens traditional productive relations. The insertion of foreign-owned factories into peripheral regions undermines regional economies by producing goods that compete with those made locally; by feminizing the workforce without providing factory-based employment opportunities for men; and by socializing women for industrial work and modern consumption without providing a lifetime income capable of meeting these needs. The result once again is the creation of a population that is socially and economically uprooted and prone to migration.

The same capitalist economic processes that create migrants in peripheral regions simultaneously attract them into certain highly developed urban areas. The investment that drives economic globalization is managed from a small number of *global cities*, whose structural characteristics create a strong demand for migrant labor. In order to ship goods, deliver machinery, extract and export raw materials, coordinate business operations, and manage expatriate assembly plants, capitalists in global cities build and expand transportation and communication links to the peripheral regions where they have invested. These links not only facilitate the movement of goods, products information, and capital, they also promote

the movement of people by reducing the costs of movement along certain pathways.

The creation and perpetuation of a global trading regime also requires an underlying system of international security. Core capitalist nations have both an economic interest in, and the military means of preserving, geopolitical order, and leading powers thus maintain relatively large armed forces to deploy as needed to preserve the integrity of the global capitalist system. Threats to the stability of that system are met by military force projected from one or more of the core nations. Each military base and armed intervention, however, creates a range of social and political connections that promote the subsequent movement of migrants.

Finally, processes of economic globalization also create ideological or cultural links between core capitalist regions and their peripheries. In many cases, these cultural links are longstanding, reflecting a colonial past in which core countries established administrative and educational systems that mirrored their own in order to govern and exploit a peripheral region. Ideological connections are presently reinforced by mass communications and advertising campaigns directed from the core urban centers. The diffusion of core cultural patterns and the spread of modern consumption interact with the emergence of a transportation-communication infrastructure to channel migrants disproporationately to global cities.

5. *Social Capital Theory*

Although it was Loury (1977) who first introduced the concept of social capital, it was Bourdieu (1986) who pointed out its broader relevance to human society. According to Bourdieu and Wacquant (1992, p. 119), 'social capital is the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition.' The key characteristic of social capital is its convertibility—it may be translated into other forms of capital, notably financial capital. People gain access to social capital through membership in networks and social institutions and then convert it into other forms of capital to improve or maintain their position in society (Coleman 1988).

Massey et al. (1987, p. 170) were the first to identify migrant networks as a form of social capital. Following Coleman's (1990, p. 304) dictum that 'social capital … is created when the relations among persons change in ways that facilitate action,' they identified migration itself as the catalyst for change in the nature of social relations. Everyday ties of friendship and kinship provide few advantages, in and of themselves, to people seeking to migrate. Once someone in a personal network has migrated, however, the ties are transformed into a resource that can be used to gain access to employment and high wages at points of destination. Each act of migration creates social capital among people to whom the new migrant is related, thereby raising the odds of their migration.

Goss and Lindquist (1995) point to migrant institutions as a structural complement to migrant networks, arguing that interpersonal ties are not the only means by which social capital is created and that 'migration is best examined not as a result of individual motivations and structural determinations, although these must play a part in any explanation, but as the articulation of agents with particular interests and playing specific roles within an institutional environment … ' (p. 345). For-profit organizations and private entrepreneurs provide services to migrants in exchange for fees set on the underground market: surreptitious smuggling across borders; clandestine transport to internal destinations; labor contracting between employers and migrants; counterfeit documents and visas; arranged marriages between migrants and legal residents or citizens of the destination country; and lodging, credit, and other assistance in countries of destination. Over time, individuals, firms, and organizations become institutionally stable and offer another source of social capital to would-be migrants.

6. *Cumulative Causation*

The theory of cumulative causation basically argues that human migration changes individual motivations and social structures in ways that make additional movement progressively likely, a process first identified by Myrdal (1957) and reintroduced to the field by Massey (1990). Causation is cumulative in the sense that each act of migration alters the context within which subsequent migration decisions are made to make additional trips of longer duration more likely. Social scientists have discussed eight ways that migration has cumulatively caused: the expansion of networks, the distribution of income, the distribution of land, the organization of agriculture, culture, the regional distribution of human capital, the social meaning of work, and the structure of production. Feedbacks through other variables are also possible, but have not been systematically treated.

7. *A Theoretical Synthesis*

Massey et al. (1998) recently completed a systematic review of theory and research throughout the world to derive a synthetic model human migration. Contemporary migration appears to originate in the social, economic, political, and cultural transformations that

accompany the penetration of capitalist markets into non-market or pre-market societies (world systems theory). In the context of a globalizing economy, the entry of markets into peripheral regions disrupts existing social and economic arrangements and brings about the displacement of people from customary livelihoods, creating a mobile population of workers who actively search for new ways of earning income, managing risk, and acquiring capital. Migration does not stem from a lack of economic development, but from development itself.

One means by which people displaced from traditional livelihoods seek to assure their well-being is by selling their labor on emerging markets (neoclassical economics). Because wages are generally higher in urban than in rural areas, much of this process of labor commodification is expressed in the form of rural–urban migration; but wages are even higher, in developed countries overseas, and the larger size of these transnational wage differentials inevitably prompts some adventurous people to sell their labor on international markets by moving abroad for work.

International wage differentials are not the only factor motivating people to migrate, however. Evidence also suggests that people displaced in the course of economic transformation move not simply in order to reap higher lifetime earnings by relocating permanently to a foreign setting; rather, households struggling to cope with the jarring transformations of early economic development use labor migration as a means of managing risk and overcoming barriers to capital and credit (the new economics of labor migration).

In many regions, markets or government programs for insurance, futures, capital, credit, and retirement are poorly developed or nonexistent, and households turn to migration in order to compensate for these market failures. By sending members away to work, households diversify their labor portfolios to control risks stemming from unemployment, crop failures, or price fluctuations. Migrant labor also permits households to accumulate cash for large consumer purchases or productive investments, or to build up savings for retirement. Whereas the rational actor posited by neoclassical economics takes advantage of a geographic disequilibrium in labor markets to move away *permanently* to achieve higher lifetime earnings, the rational actor assumed by the new economics of labor migration seeks to cope with market failures by moving *temporarily* to repatriate earnings in the form of remittances.

While the early phases of economic development promote emigration, transformations in post-industrial cities yield a bifurcation of labor markets. Jobs in the primary labor market provide steady work and high pay for local workers, but those in the secondary labor market offer low pay, little stability, and few opportunities, thus repelling locals and generating a structural demand for migrant workers (segmented labor-market theory). This process of labor market bifurcation is most acute in global cities, where a concentration of managerial, administrative, and technical expertise leads to a concentration of wealth and a strong ancillary demand for low-wage services (world-systems theory). Unable to attract native workers, employers turn to immigrants and initiate immigrant flows directly through formal recruitment (segmented labor-market theory).

Although instrumental in initiating migration flows, recruitment becomes less important over time because the same processes of economic globalization that create mobile populations in developing regions, and which generate a demand for their services in certain cities, also create links of transportation, communication, politics, and culture to make the movement of people increasingly cheap and easy (world-systems theory). International migration is also promoted by the foreign policies of, and the military actions taken by, core nations to maintain international security, protect investments, and guarantee access to raw materials, entanglements that create links and obligations that often generate ancillary flows of refugees, asylees, and military dependents.

However a migration stream begins, it displays a strong tendency to continue because of the growth and elaboration of networks (social-capital theory). The concentration of migrants in certain destination areas creates a 'family and friends' effect that channels later cohorts to the same places and facilitates their arrival and incorporation. If enough migrants arrive under the right conditions, an enclave economy may form, which further augments the specialized demand for migrant workers (segmented labor-market theory).

The spread of migratory behavior within sending communities sets off ancillary structural changes, shifting distributions of income and land and modifying local cultures in ways that promote additional out-migration. Over time, the process of network expansion tends to become self-perpetuating because each act of migration causes social and economic changes that promote additional movement (the theory of cumulative causation). As receiving areas implement restrictive policies to counter the rising tides of migrants, they create a lucrative niche into which enterprising agents, contractors, and other middlemen move to create migrant-supporting institutions, providing migrants with yet another infrastructure capable of supporting and sustaining movement (social-capital theory).

During the initial phases of out-migration, the effects of capital penetration, market failure, social-network expansion, and cumulative causation dominate to cause a rapid upsurge of the flow, but as the level of emigration reaches high levels, and the costs and risks of movement drop, movement is increasingly determined by wage differentials (neoclassical economics) and labor demand (segmented labor-market theory). As economic growth in sending areas occurs,

wage gaps gradually diminish and well-functioning markets for capital, credit, insurance, and futures come into existence, progressively lowering the incentives for emigration. If these trends continue, the country or region ultimately becomes integrated into the global or national economy, whereupon it undergoes a transformation: massive out-migration progressively trails off and shifts to a pattern of circulatory movement whose net balance fluctuates in response to changing economic conditions at origin and destination (Hatton and Williamson, 1998), yielding a characteristic curve in the course of the mobility transition that Martin and Taylor (1996) have called the 'migration hump.' Crossing this so-called hump quickly and painlessly is one of the fundamental challenges of economic development.

See also: Assimilation of Immigrants; Immigrants: Economic Performance; Immigration; Immigration and Migration: Cultural Concerns; Immigration: Consequences for Fiscal Developments in the Receiving Population; Immigration: Public Policy; Migration and Health; Migration, Economics of; Migration out of Europe; Migration: Sociological Aspects; Migrations, Colonizations, and Diasporas in Archaeology

Bibliography

Bourdieu P 1986 The forms of capital. In: Richardson J G (ed.) *Handbook of Theory and Research for the Sociology of Education*. Greenwood, New York

Bourdieu P, Wacquant L J D 1992 *An Invitation to Reflexive Sociology*. University of Chicago Press, Chicago

Coleman J S 1988 Social capital in the creation of human capital. *American Journal of Sociology* **94S**: 95–120

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Fligstein N 1981 *Going North, Migration of Blacks and Whites from the South, 1900–1950*. Academic Press, New York

Goss J, Lindquist B 1995 Conceptualizing international labor migration: a structuration perspective. *International Migration Review* **29**: 317–51

Hatton T J, Williamson J G 1998 *The Age of Mass Migration: Causes and Economic Impact*. Oxford University Press, Oxford

Lewis W A 1954 Economic development with unlimited supplies of labour. *Manchester School of Economic and Social Studies* **22**: 139–91

Loury G C 1977 A dynamic theory of racial income differences. In: Wallace P A, LaMond A M (eds.) *Women, Minorities, and Employment Discrimination*. Lexington Books, Lexington, MA

Martin P L, Taylor J E 1996 The anatomy of a migration hump. In: Taylor J E (ed.) *Development Strategy, Employment, and Migration: Insights from Models*, Organisation for Economic Co-operation and Development, Development Centre, Paris

Massey D S 1990 Social structure household strategies, and the cumulative causation of migration. *Population Index* **56**: 3–26

Massey D S, Alarcón F, Durand J, González H 1987 *Return to Aztlan: The Social Process of International Migration from Western Mexico*. University of California Press, Berkeley CA

Massey D S, Arango J, Hugo G, Kouaouci A, Pellegrino A, Taylor J E 1998 *Worlds in Motion: Understanding International Migration at the End of the Millennium*. Oxford University Press, Oxford

Myrdal G 1957 *Rich Lands and Poor: The Road to World Prosperity*. Harper and Row, New York

Piore M J 1979 *Birds of Passage: Migrant Labor and Industrial Societies*. Cambridge University Press, New York

Portes A, Bach R L 1985 *Latin Journey: Cuban and Mexican Immigrants in the United States*. University of California Press, Berkeley

Ranis G, Fei J C H 1961 A theory of economic development. *American Economic Review* **51**: 533–65

Sassen S 1988 *The Mobility of Labor and Capital: A Study in International Investment and Labor Flow*. Cambridge University Press, Cambridge, UK

Sjaastad L A 1962 The costs and returns of human migration. *Journal of Political Economy* **70S**: 80–93

Stark O 1991 *The Migration of Labor*. Blackwell, Oxford, UK

Todaro M P 1976 *Internal Migration in Developing Countries: A Review of Theory, Evidence, Methodology, and Research Priorities*. International Labour Office, Geneva

D. S. Massey

Migration to the United States: Gender Aspects

Despite the overwhelming presence of women in migration flows, and statistical data gathering on the migration of women as well as men, until recently the gender aspects of migration had been totally neglected and the pervasive assumption was that the international migrant is a young, economically motivated male. However, legal immigration to the United States—still very much the largest of all international flows—was dominated by women for most of the twentieth century. For the United States a crossover in sex differentials in migration occurred in 1930, after which women have annually outnumbered men (Houstoun et al. 1984) with the sole exception of a couple of years after the passage of the Immigration Reform and Control Act (IRCA) in 1986 that granted amnesty to illegals. As demographers, Houstoun et al. highlighted this glaring neglect for the discipline of sociology, just as Seller (1975) had highlighted it for the field of history. After these various calls, attention began to be paid to women and migration until the topic has now mushroomed and we are beginning to develop what Stacey and Thorne (1985, pp. 305–6) called a gendered understanding of all aspects of human culture—one that traces 'the significance of gender organization and relations in all institutions

and in shaping men's as well as women's lives' that leads to a fuller understanding of the causes, processes, and consequences of migration.

Such a gendered understanding should elucidate those aspects of the process of migration which were neglected by the exclusive focus on men. As Tilly (1989) underscored, bringing women into the humanities and the social sciences takes place in stages: first, by filling in the gaps in knowledge resulting from their absence; second, by transforming the conceptual and theoretical frameworks of their disciplines.

The study of immigration is by its very nature interdisciplinary. A natural division of labor has arisen whereby sociologists attend most to contemporary immigration flows (the Latin American and Asian), historians are concerned with past flows (the Southern and Eastern European), and anthropologists observe the impact of emigration and return on sending communities in underdeveloped nations. Within sociology, much research on women has filled in gaps and yielded new insights and directions, but until recently the field itself had undergone little transformation. Traditionally, sociology was neither totally male-defined, as history of literature, nor basically gender-sensitive, as anthropology. Stacey and Thorne (1985) judged that the feminist contributions to sociology were contained by the delimiting capacity of functionalism to explain male–female differences; of empiricism to treat gender as a variable, rather than as a central theoretical concept; of Marxist sociology to ghettoize it; and by the underdevelopment of feminist theory itself.

Striving to contribute to a gendered understanding of the social process of migration, this article has been organized according to these major issues. How is gender related to the decision to migrate? What are the patterns of labor market incorporation of women immigrants? What is the relationship of the public and the private?

1. The Decision To Migrate

1.1 Micro and Macro Linkages

The underlying assumption in studies of migration has been that of the male pauper—a single or married man who seeks to amass capital with which to return to his native country. The corollary assumption has been that it is men who typically make the decision to migrate and women follow to create or reunite a family, generating secondary movements (cf. Lee 1966, Houstoun et al. 1984).

In sociology, the traditional, individual micro approach was best developed by Lee's (1966) theory that focused on the individual migrant's decision to migrate—the 'push' and 'pull' factors that hold or attract people, as well as the intervening obstacles (distance, physical barriers, immigration laws, cost), the influence of personal traits (stage in the life cycle, contact with earlier migrants), and the effect of transitions (marriage or retirement).

More recently, the structural, macro approach to the study of migration developed as the link between migration and world patterns of unequal development increasingly became evident as North America continued to attract the world's poor and in Western Europe the periphery countries of Spain, Italy, Greece, and Turkey became suppliers of labor to the industrialized core countries of France, Germany, and Switzerland. The structural perspective argued that a system of economic migration had developed from the flow of labor between developed and underdeveloped nations due to the functions that this system of labor migration performed (cf. Castells 1975, Burawoy 1976, Portes 1978, Pedraza-Bailey 1985), providing host countries, such as the United States or France, with a dependable source of cheap labor, while also providing the countries of emigration, such as Mexico or Turkey, with a 'safety valve' for the discontent of their poor and lower-middle classes. Though this perspective largely ignored the gendered aspects of migration, some analysts, such as Fernandez-Kelly's (1983) in-depth study of the *maquilladora* industries in the US-Mexican border, emphasized its gender-specific nature. Most undocumented aliens working the fields of the US continued to be men, while most of those working in the export-manufacturing plants along the Mexican border were women.

The danger of the structural emphasis, however, lies in its tendency to lose sight of the individual migrants who do make decisions. The theoretical and empirical challenge now facing immigration research inheres in its capacity to capture both individuals as agents, and social structure as delimiting and enabling. Such a link between micro and macro levels of analysis has developed through studies of gender and family as well as on social networks. Massey et al. (1987), in *Return to Aztlan*, their analysis of Mexican migration to the US, showed that while international migration originates historically in transformations of social and economic structures in sending and receiving societies, once begun, migration comes to fuel itself. As migrants' social networks grow and develop, families make migration part of their survival strategies and individual motivations, household strategies, and community structures are altered by migration, making further migration more likely. In *Between Two Islands* (1991), Grasmuck and Pessar's analysis of contemporary migration from the Dominican Republic to New York City (the two islands), also emphasized that the household is the social unit which makes decisions as to whether migration will take place, who in the family will migrate, what resources will be allocated to the migration, what remittances will be sent back, whether or not family members will

return back home, and whether the migration will be temporary or permanent. All of these decisions are guided by normatively prescribed kinship and gender roles as well as by the hierarchy of power within the household, decision-making that often betrays enormous interpersonal conflict (cf. Hondagneu-Sotelo 1994), but which also constitutes a family strategy to meet the challenges that accompany economic and political transformation in the Third World.

2. *The Demographic Composition of the Migration Flow*

Emigration is a process experienced differently by women and men; hence, it can be sex-selective and also forms sex imbalances, both in the old country and the new (Parr 1987). Examining international migration patterns shows that women have predominated among immigrants to Argentina, Israel, and the US, and constitute an increasing share of migrants in areas such as West Africa and the Persian Gulf states (Tyree and Donato 1986). Examining internal migration patterns shows that whereas in Africa and South Asia men predominate in migration to the cities and women remain in rural areas to farm the land, in Latin America, the Caribbean, and Philippines, most migrants to cities are women (e.g., Khoo et al. 1984, Lee 1989, Hojman 1989, Gugler 1989), contrasts that depend on the land tenure and agricultural production arrangements. Moreover, government policy can create imbalanced migration flows by legally restricting the migration of men or women (e.g., Wilkinson 1983).

Flows of migration that are dominated by men require that we consider 'the woman's side' when the women themselves are left behind in the communities. For example, Brettell's (1988) analysis of the long-standing male emigration from Portugal to Brazil and Spain in the late ninteenth and twentieth centuries showed the impact of the emigration on the matricentric characteristics of the way of life in this area. Households adjusted to the centuries of absence of men by becoming female-headed, often extended three-generation households, and, contrary to the Mediterranean pattern, patriuxorilocal, since grooms often moved in with the wife's family. Moreover, the migration of men also affected the lives of women by promoting a delayed age at marriage and high rates of spinsterhood as well as illegitimacy, all of which bound the women more firmly to their families of origin.

Ultimately, the demographic composition of migration flows is important not only because its causes are various but also because it has consequences. In his comparative analysis of Italian and Jewish mobility in New York at the turn of the century, Kessner (1977) underscored that their patterns of social mobility depended on the varying composition of the migration flows. Newcomers that arrive as temporary migrants—as 'birds of passage,' in Piore's (1979) phrase—work with the goal to return home, tolerating the most abysmal working conditions to accumulate capital for their investments back home. By contrast, permanent immigrants must make their future in the new land and cannot tolerate abysmal working conditions as temporary. Thus, they seek to attain social mobility in the new society, taking greater risks and making more long-term investments, such as setting up family businesses. The two types of migration are reflected in the demographic composition of the flows. Flows of temporary migrants, such as the Italian or the Mexican, are by and large non-family movements of men in the productive years who intend to make money and return home. By contrast, flows of permanent immigrants, such as the Jewish or Cuban, are characterized by the migration of families who intend to remake their lives and homes. It is quite common for flows of refugees—who leave their country in fear, seeking safety—to be initially dominated by women and children, as in the early years of both the Cuban and Indochinese exodus to the US, sex imbalance that, in the Cuban case, was thereafter reversed (Pedraza 1996). As refugees, women are particularly prone to victimization due to violence and indifference to their plight.

Diner (1983) studied a female-dominated flow of migration, that of Irish immigrant women in the nineteenth century. The Irish migration was 'pushed' by conditions that prevailed throughout much of Europe then—poverty, landlessness, and the social and economic dislocations that accompanied the transition from feudalism to capitalism (cf. Bodnar 1985), exacerbated by the Famine at mid-century. Coupled with the Irish system of single inheritance and single dowry, Ireland increasingly became the home of the unmarried and the late married. In Ireland, women no longer had realistic chances for marriage or employment; to attain either they had to turn their backs on the land of their birth. As the century wore on, the migration became basically a female mass movement. Hence, not just famine and poverty but what Jackson (1984, pp. 1007–8) called 'the interlocking relationship of land-family-marriage' caused the preponderance of women in the migration. As a consequence of land scarcity, both arranged marriages and the practice of dowries spread, and celibacy and late marriages rose. One escape from family and spinsterhood was for women to join a religious order; another was emigration. The usual kin chain migration became a female migratory chain.

3. *The Incorporation of Women*

3.1 *Labor Force Participation*

That immigration has a decided impact on the labor force participation of women is a central fact of

immigration research. For example, in contrast to the very low rates of labor force participation of women in Cuba prior to the revolution, and Mexican and Puerto Rican women in the US, Cuban women immigrants in the United States have a very high rate of labor force participation (Perez 1988). Achieving the upward mobility of the Cuban family in the United States made women's work necessary and broke with the traditional Cuban notion that a woman's place is in the home, justifying the massive entrance of women into the labor force (Prieto 1987), employment that was not necessarily accompanied by a change in traditional values.

Numerous research studies have examined the labor market outcomes of immigrant women—the occupations and income they attained and the disadvantages reflected in different 'payoffs' to their characteristics. Most of these studies, however, suffered from the problem Stacey and Thorne identified as treating gender as a variable, rather than as a central theoretical principle.

3.2 Occupational Concentration

Like men, immigrant women became occupationally concentrated in particular types of occupations, by and large clustering in just a few of them. Typically, women work as domestic servants, sew for the garment industry, serve in small family businesses, or, most recently, work in highly skilled service occupations, such as nursing.

The major consequence of the predominantly female and single nature of the Irish migration in the nineteenth century was that the women overwhelmingly entered domestic service, an occupation in which there was a labor vacuum because other poor women did not want it since the expectation that they 'live in' interfered with their own family life. But they were able to amass impressive savings with which to bring other relatives over; to send remittances back home; to support their church; to secure an American marriage; and to finance the foundation for a small business or an education. Thus, Irish women experienced higher rates of social mobility than Irish men. Glenn's (1986) study of three generations of Japanese Women in domestic service, *Issei*, *Nisei*, *War Bride*, depicted domestic service as one of the few occupations open to women of color in American history.

Yesterday's immigrants, as well as today's, became concentrated in the garment industry since it relied on a traditional skill that throughout much of the world defined womanhood; moreover, it relied on homework and subcontracting, allowing women to stay at home to care for their children. This advantage always led women to accept low wages and exploitative conditions (e.g., Howe 1976, Safa 1984, Sanchez-Korrol 1983, Lamphere 1987). In his study of the garment industry, *Through the Eye of a Needle*, Waldinger (1986) showed how New York became the leading center of the garment industry, its growth spurred by the arrival of an immigrant labor force—poor, industrious, and lacking in other skills—of Russian Jews and Italians. Since many of the Russian Jews had previously developed skills in the needle trades, garments quickly became the Jewish trade, though specialization was gendered as men worked on tailoring (coats and suits) while women worked on the lighter trade (dresses, undergarments, and children's clothes). That same immigrant labor force is today Latin American and Asian.

Ethnic enterprise—the concentration of certain immigrant groups in small business—describes the immigrant experience of yesterday (Jews, Chinese, Italians, Greeks) as well as today (Koreans, Arabs, Chaldeans, Cubans) in the US. The unpaid family labor donated by women—wives, daughters, mothers, grandmothers—is a large part of what allows immigrants to amass profits and turn them into savings that are reinvested in the development and growth of family businesses. Thus, women's contribution is the key to the success of these enterprises and to the achievement of the class position best described by the phrase *la petite bourgeoisie*.

Women immigrants also figure importantly as technicians, teachers, doctors, and nurses. Shin and Chang's (1988) study of how Korean immigrant physicians are incorporated into the American medical profession found that women physicians were much more likely to immigrate to the US than men. Moreover, while all immigrant physicians were more likely to enter the peripheral specialties of American medicine, gender contributed significantly to that peripheralization.

4. The Public and the Private

Research on immigrant women has sought to chronicle both the private world of immigrant women and their families, as well as the contribution immigrant women made to the private sphere of other women's families. Weinberg's (1988) study of *The World of Our Mothers* (so deliberately a corrective to Howe's (1976) *The World of Our Fathers*) on the Jewish immigrant experience in New York city at the turn of the century showed how the lives of Jewish immigrant women, centered on the domestic sphere, differed from the lives of men, defined by work and the synagogue. Moreover, immigrant women played a mediating role between the old world and the new. Immigration exposed daughters to the ways of a modern, secular world they were eager to accept. Mothers themselves clung to traditional, Orthodox ways. But within the family these women, who lived for and through others,

played the role of mediators between fathers and daughters.

The labor that immigrant women donated as servants also contributed to the changing role of the housewife in America. Matthews (1987) pointed out that the cult of domesticity arose in the early to mid-nineteenth century among middle class and upper middle class women because the availability of domestic servants allowed time for the development of the arts of baking and needlework. While historically women had relied on other women as 'help,' and worked side by side with them on domestic chores, in the nineteenth century 'domestic servants' that required supervision replaced the 'help'—a change that was facilitated by the increasing number of poor immigrant women coming to America. While in mid twentieth-century, American technology (dishwashers, microwaves, and the like) replaced the servants, at the century's end, poor, immigrant women again began to serve as domestic servants as the dual career marriage now rendered it a necessity (cf. Repak 1994).

The deeply felt needs of immigrant women also found expression in their popular, religious tradition. In his analysis of the devotion Italian immigrants in New York poured onto *The Madonna of 115th Street*, Orsi (1985) underscored that while the Madonna came from Italy with the immigrants, and as such was a symbol to all Italians of nation, history, and tradition, above all she was a woman's devotion as these women turned to the Madonna with petitions for help with the hardship and powerlessness of their lives. That private relation became public at the annual *festa* when both men and women participated as a community that served to regenerate their national culture and to console them for the physical and spiritual trials of immigration.

5. Conclusion

In sum, gender plays a central role in the decision to migrate and the composition of the migration flows, composition which holds consequences for the subsequent form of immigrant incorporation. The experience of immigration also profoundly impacts the public and private lives of women—their labor force participation, their occupational concentration, their religiosity, their marital roles and satisfaction, and their autonomy and self-esteem. Hence, difficult as the experience of immigration was, it was often far more positive for women than for men, as it allowed women to break with traditional roles and patterns of dependence, join the labor force, and assert a newfound (if meager) freedom (cf. Foner 1978). Pessar (1984) studied Dominican women immigrants that had previously not worked in the Dominican Republic but went to work outside the home for the first time in the US. This change transformed the patriarchal roles in the household, heightened the women's self-esteem, increased their income, as well as their capacity to participate as equals in household decision-making. However, employment did not provide women with a new status as working women that challenged or subordinated their primary identities as wives and mothers. Rather, it often reinforced these very identities, allowing women to redefine them in a more satisfying manner than prior to the migration.

See also: Economic Development and Women; Family and Kinship, History of; Feminist Economics; Gender and Place; Gender, Economics of; Migration, Economics of; Migration: Sociological Aspects; Migration, Theory of; Wage Differentials and Structure; Work, Sociology of

Bibliography

Bodnar J 1985 *The Transplanted: A History of Immigrants in Urban America*. Indiana University Press, Bloomington, IN, p. 294

Brettell C B 1988 Emigration and household structure in a Portuguese parish, 1850–1920. *Journal of Family History* **13**: 33–57

Burawoy M 1976 The functions and reproduction of migrant labor: Comparative material from Southern Africa and the United States. *American Journal of Sociology* **81**: 1050–87

Castells M 1975 Immigrant workers and class struggles in advanced capitalism: The Western European experience. *Politics and Society* **5**: 33–66

Diner H R 1983 *Erin's Daughters in America: Irish Immigrant Women in the Nineteenth Century*. John Hopkins University Press, Baltimore, MD, p. 192

Fernandez-Kelly M P 1983 Mexican border industrialization, female labor force participation and migration. In: Nash J, Fernandez-Kelly M P (eds.) *Women, Men, and the International Division of Labor*. SUNY Press, Albany, NY

Foner N 1978 *Jamaica Farewell: Jamaican Migrants in London*. University of California Press, Berkeley, CA, p. 262

Glenn E N 1986 *Issei, Nisei, Warbride: Three Generations of Japanese American Women in Domestic Service*. Temple University Press, Philadelphia, p. 290

Grasmuck S, Pessar P R 1991 *Between Two Islands: Dominican International Migration*. University of California Press, Berkeley, CA, p. 24

Gugler J 1989 Women stay on the farm no more: Changing patterns of rival–urban migration in sub-Saharan Africa. *Journal of Modern African Studies* **27**: 347–52

Hojman D 1989 Land reform, female migration and the market for domestic service in Chile. *Journal of Latin American Studies* **21**: 105–32

Hondagneu-Sotelo P 1994 Regulating the unregulated?: Domestic workers' social networks. *Social Problems* **41**: 50–64

Houstoun M F, Kramer R G, Barrett J M 1984 Female predominance of immigration to the United States since 1930: A first look. *International Migration Review* **18**: 908–63

Howe I 1976 *World of Our Fathers*. Simon and Schuster, New York, p. 714
Jackson P 1984 Women in 19th Century Irish emigration. *International Migration Review* **18**: 1004–20
Kessner T 1977 *The Golden Door: Italian and Jewish Immigrant Mobility in New York City, 1880–1915*. Oxford University Press, New York, p. 224
Khoo S E, Smith P C, Fawcett J T 1984 Migration of women to cities: The Asian situation in comparative perspective. *International Migration Review* **18**: 1247–63
Lamphere L 1987 *From Working Daughters to Working Mothers: Immigrant Women in a New England Industrial Community*. Cornell University Press, Ithaca, NY, p. 390
Lee E S 1966 A theory of migration. *Demography* **3**: 47–57
Lee S M 1989 Female immigrants and labor in Colonial Malaya: 1860–1974. *International Migration Review* **23**: 309–31
Massey D S, Alarcon R, Durand J, Gonzalez H 1987 *Return to Aztlan: The Social Process of International Migration from Western Mexico*. University of California Press, Berkeley, CA, p. 335
Matthews G 1987 *Just a Housewife: The Rise and Fall of Domesticity in America*. Oxford University Press, New York, p. 281
Orsi R A 1985 *The Madonna of 115th Street: Faith and Community in Italian Harlem, 1880–1950*. Yale University Press, New Haven, CT, p. 287
Parr J 1987 The skilled emigrant and her kin: Gender, culture, and labour recruitment. *Canadian History Review* **68**: 529–51
Pedraza S 1996 Cuba's refugees: Manifold migrations. In: Pedraza S, Rumbaut R (eds.) *Origins and Destinies: Immigration, Race, and Ethnicity in America*. Wadsworth, Belmont, CA
Pedraza-Bailey S 1985 *Political and Economic Migrants in America: Cubans and Mexicans*. University of Texas Press, Austin, TX, p. 242
Pedraza-Bailey S 1990 Immigration research: A conceptual map. *Social Science History* **14**: 43–67
Perez L 1988 Cuban women in the U.S. labor force: A comment. *Cuban Studies* **18**: 159–64
Pessar P R 1984 The linkage between the household and workplace of Dominican women in the United States. *International Migration Review* **18**: 1188–211
Piore M 1979 *Birds of Passage*. Cambridge University Press, New York, p. 229
Portes A 1978 Migration and underdevelopment. *Politics and Society* **8**: 1–48
Prieto Y 1987 Cuban women in the U.S. labor force: Perspectives on the nature of the change. *Cuban Studies* **17**: 73–94
Repak T A 1994 Labor market incorporation of central American immigrants in Washington, DC. *Social Problems* **41** (February): 114–28
Safa H I 1984 Female employment and the social reproduction of the Puerto Rican working class. *International Migration Review* **18**: 1168–87
Sanchez-Korrol V E 1983 *From Colonia to Community: The History of Puerto Ricans in New York City 1917–1948*. Greenwood Press, Westport, CT, p. 242
Seller M S 1975 Beyond the stereotype: A new look at the immigrant woman. *Journal of Ethnic Studies* **3**: 59–68
Shin E H, Chang K P 1988 Peripherization of immigrant professionals: Korean physicians in the United States. *International Migration Review* **22**: 609–26
Stacey J, Thorne B 1985 The missing feminist revolution in sociology. *Social Problems* **32**: 301–16
Tilly L A 1989 Gender, women's history, and social history. *Social Science History* **13**: 439–62
Tyree A, Donato K 1986 A demographic overview of the international migration of women. In: Simon R J, Brettell C B (eds.) *International Migration: The Female Experience*. Rowman and Allanheld, Totowa, NJ, p. 310
Waldinger R 1986 *Through the Eye of the Needle: Immigrants and Enterprise in New York's Garment Trades*. New York University Press, New York, p. 231
Weinberg S S 1988 *The World of Our Mothers: The Lives of Jewish Immigrant Women*. Schocken Books, New York, p. 325
Wilkinson R C 1983 Migration in Lesotho: Some comparative aspects, with particular reference to the role of women. *Geography* **68**: 208–24

S. Pedraza

Migrations, Colonizations, and Diasporas in Archaeology

Migrations are the basis for widespread human presence in the world today. The specificity of human migrations is related to their adaptive mode, and is entirely different from animal dispersions. The flexibility of cultural systems permits, in principle, their adaptation to any natural environment encountered. Thus, this form of expansion can be broader and more intense than modifications of a biological order. Demographic developments occurring in human populations due to technological adaptations are often evoked to justify such human migrations. However, causes of a metaphysical order seem to act more powerfully to drive populations toward dispersal.

Regardless of the cause(s), migrations are a constant observable in human behaviour up to the present. In reality, a sedentary way of life seems rather to constitute an exception rather than the rule, considering the immense history during which humanity was exclusively nomadic. Once techno-economical capacities were adequate, they were immediately put to use for displacements which required new forms of adaptation of the social group, including the readjustment of values.

In terms of methodology, we often distinguish seasonal migration from permanent displacement. In practice, the distinction is reduced rather to a notion of scale, applied to time, rates and proportions of migrations. Humanity being a mobile species, its territorial extension was a result of perpetual movement and only large-scale phenomena can be considered permanent where diversity is constant and processes unending.

From the beginning, the distinctive feature of humanity was its ability to leave the natural environment favorable to the way of life followed by

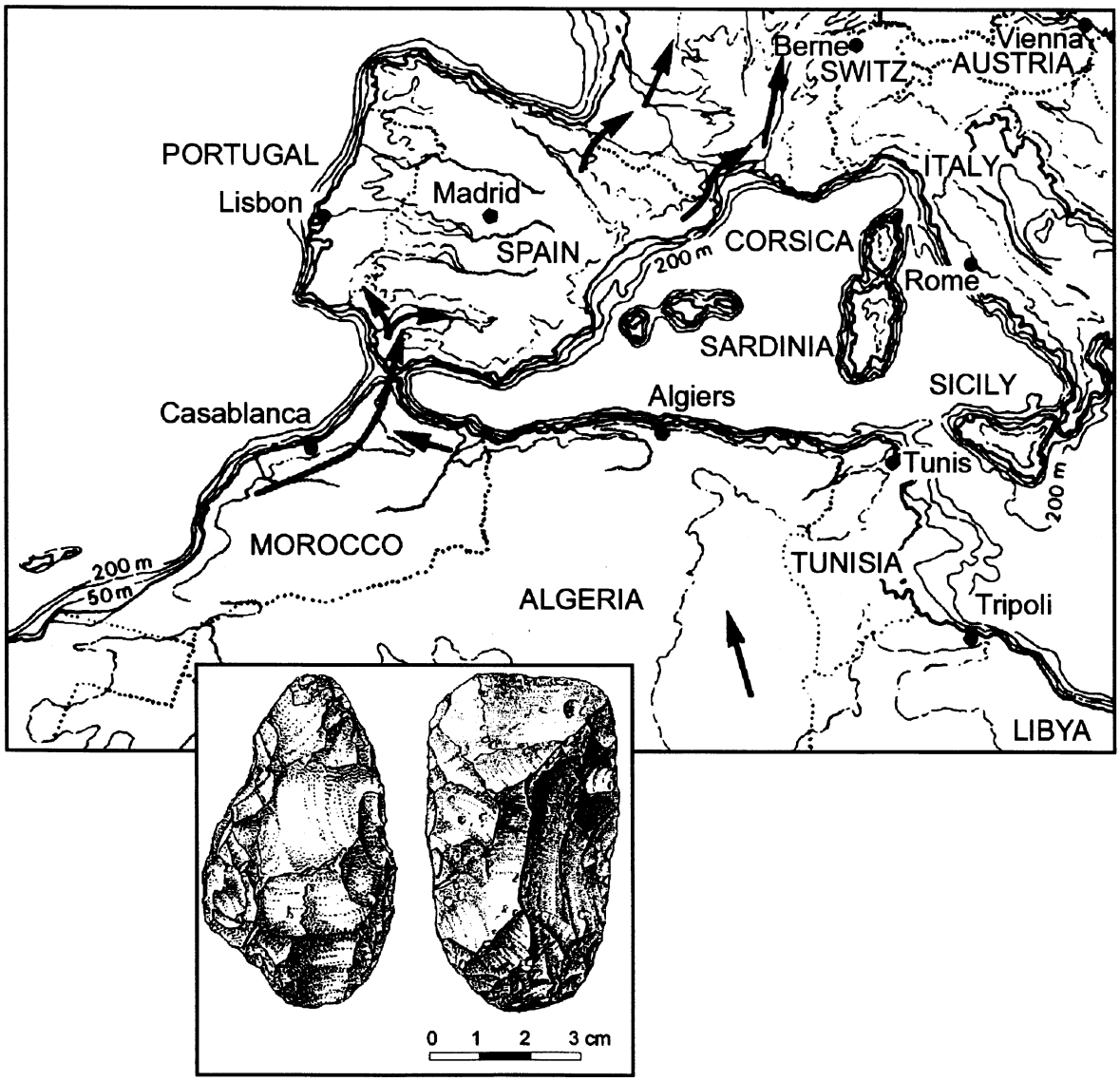

Figure 1
The first Acheulean migration in Europe passed from Africa to Spain via Gibraltar, after an earlier population from Asia arrived in Europe with a flake-based industry (after Roe 1981)

other primates. Via bipedalism and the manipulation of objects, man was able to leave the protective and nourishing environment of the tropical forests. The bipedal position, adapted to open environments, provoked these displacements, while a meat-based diet compensated for the loss of calories from vegetal foods, which were greatly reduced in the new savannah environments. Migration and the cultural adaptation that bipedalism and such a diet necessitated, were therefore the driving forces behind the origins of humanity. The freedom acquired in relation to environmental determinism became the inevitable destiny of humanity, under the ultimate threat of its complete disappearance.

The natural environments successively encountered in the process of hominid expansion toward northern latitudes, required the development and adaptation of more complex and suitable physical forms (Wolpoff 1998, Klein 1989). These early migratory movements extended to the north and east of the Old World, each

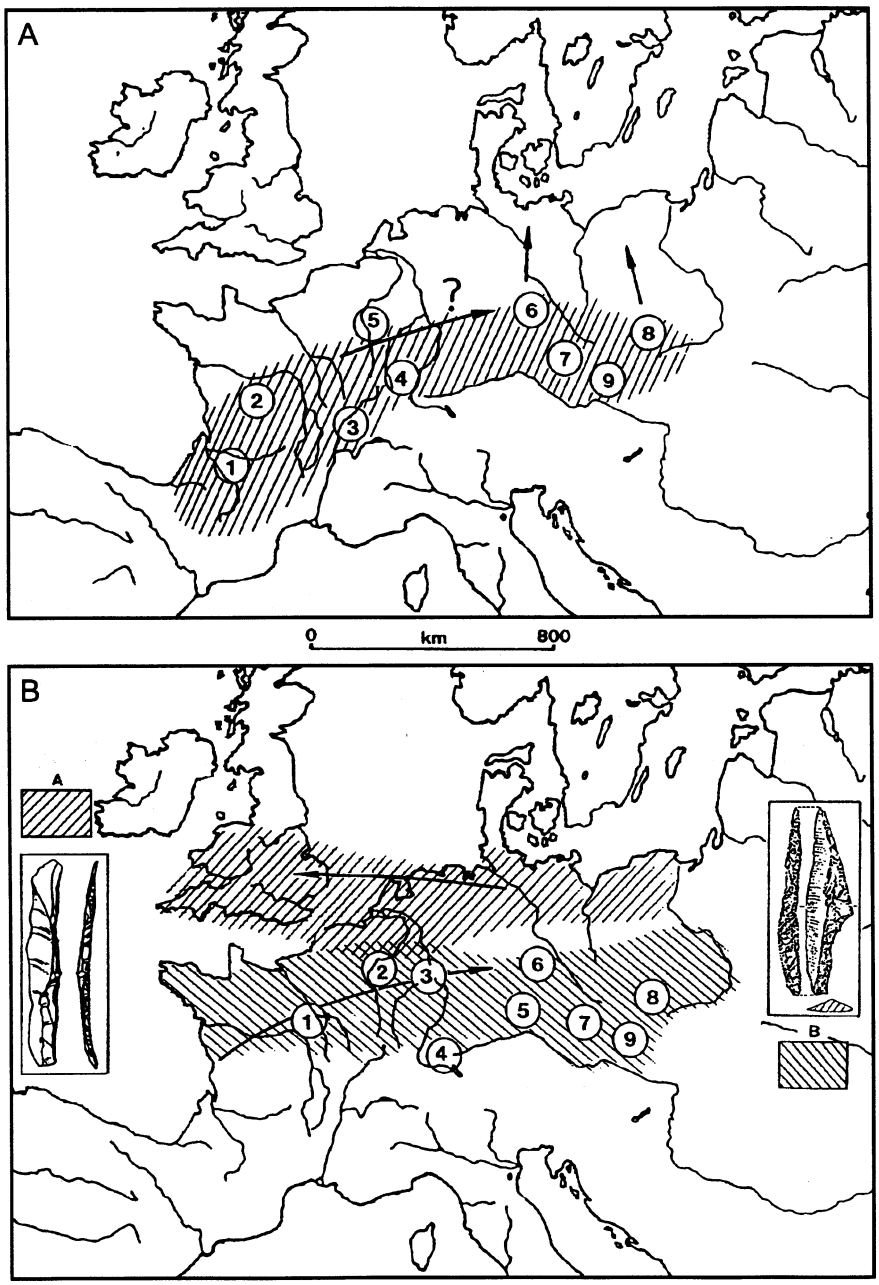

Figure 2
A: Magdalenian movement in hilly regions during the Dryas I. B: Dual movement during the Late-glacial, toward the northern plains and west across the then-dry North Sea (after Otte 1996)

time reaching new areas and demanding new dietary resources. The two keys explaining the range and the success of these migrations are the production of tools and the use of fire. Both were invented through imagination and audacity, forces that alone hold the explanation for such pioneering enterprises (Fig. 1). Humanity eventually progressed throughout history according to this continuous thread of making ever-more extended territorial conquests—towards the islands, the Americas, the moon. No force other than

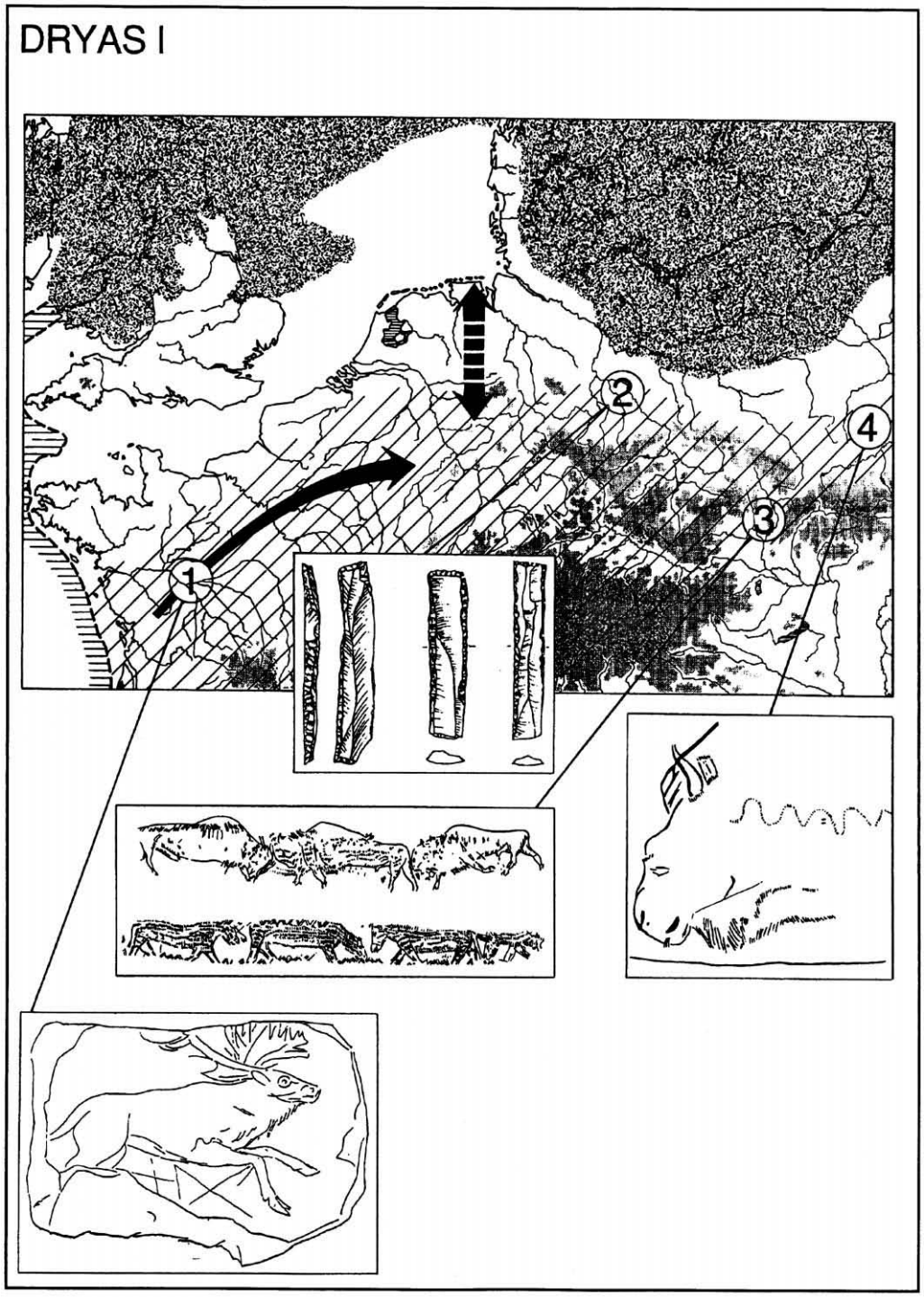

Figure 3
Two forms of migrations are observed in the northern plains during the Late-glacial: conquest in the eastern hilly regions and seasonal movements from north to south, following fluvial axes (after Otte 1997b)

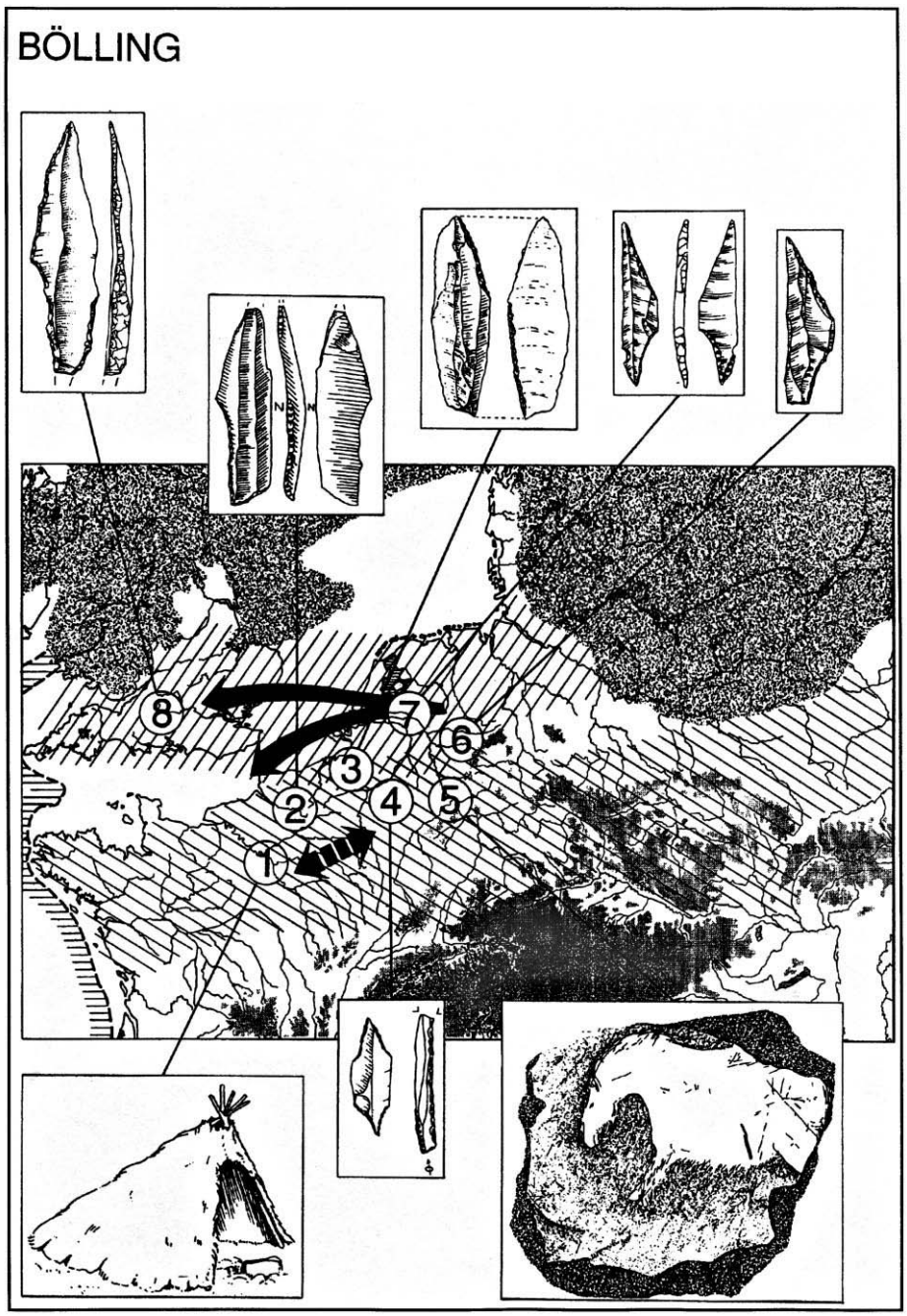

Figure 4
While Magdalenian migrations followed the axis of the plateaus, the northern plains were crossed by Hambourgian and Creswellian groups (after Otte 1997b)

imagination pushed humanity, yet—retrospectively—we can retrace the high points, the pauses, the adaptations and changes in rhythm. Within the confines of already-occupied areas, complex phenomena resulting from contacts and exchanges proper to humanity, were produced. These phenomena can be

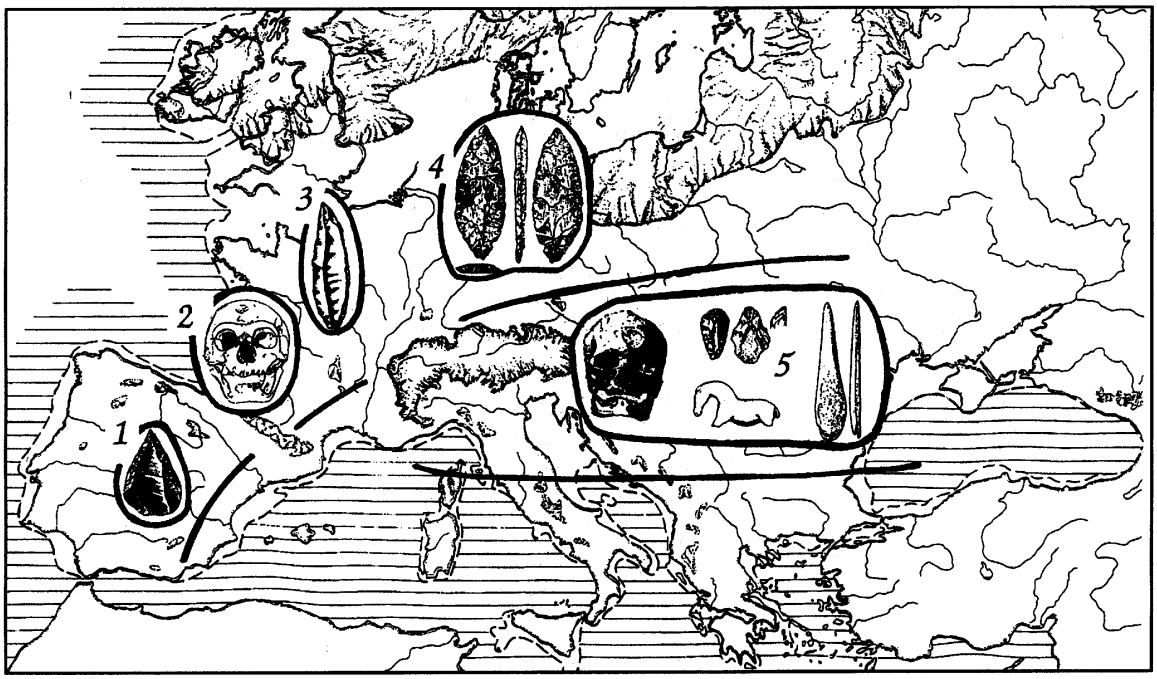

Figure 5
The first migration by modern humans in Europe met the established local Neanderthals. The Aurignacian remained homogeneous with respect to people and traditions. At the margins of this movement, contact phenomena arose: the acculturation of the Neanderthals by the Aurignacian culture (after Otte 1995)

classified in categories such as acculturation, colonization, and diaspora.

In order to appreciate the movements of migration toward unoccupied areas, the examination of conquests on the margins can be instructive. For diffusion to the high seas of Oceania, migration is clearly linked to the mastery of navigation. An analogous situation is found on the northern margins of Europe during the Late-glacial: Hambourgian and Magdalenian reindeer hunters overcame climatic constraints as temperatures rose (Figs. 2–4).

In order that massive displacements towards unoccupied regions may take place, technological equipment, in the subtlety of their elaboration, permitted adaptation to new environments and not simply to climatic modifications. This 'call to conquest' and the temptation of mobility seem to act as the only agents leading to displacement. Within occupied areas, more limited migratory cycles can be observed, e.g., seasonal rounds (following the displacement of herds) or for ceremonial reasons (trade/exchange, marriage). These are such that potential mobility remains constant and is effectively begun from the moment that a threshold of ability is attained in relation to new constraints in a foreign environment. Technology, hunting and group solidarity therefore had to become specialised before a migration could take place outside the original area of development.

The conquest continues east to Easter Island and north to Scandinavia. The case of the Americas remains perhaps the most revealing: migrations first followed the Siberian coast, then the islands of Beringia, and finally the northwest Pacific Coast. Migratory waves then switched from the high latitudes to the southern opening formed by the Californian coast. The invasions, extremely intense in North America, passed through a bottleneck between glaciers where modes of adaptation could reverse abruptly. In other terms, it seems quite probable that the first migration took place via coasts, directly toward the south. This would explain the early dates in South America (e.g., Brazil). Moreover, the northern plains were colonised later: either after the separation of the glaciers to the north or even by a return, coming from the south after having crossed the continent only in Central America. As means of communications were developed, displacements were organized, first by coastal or marine environments, later by land. The modes of occupation of the prehistoric Americas still reflect migratory waves following the first conquest.

The contacts with the cultural milieus already in place during migrations appear more complicated

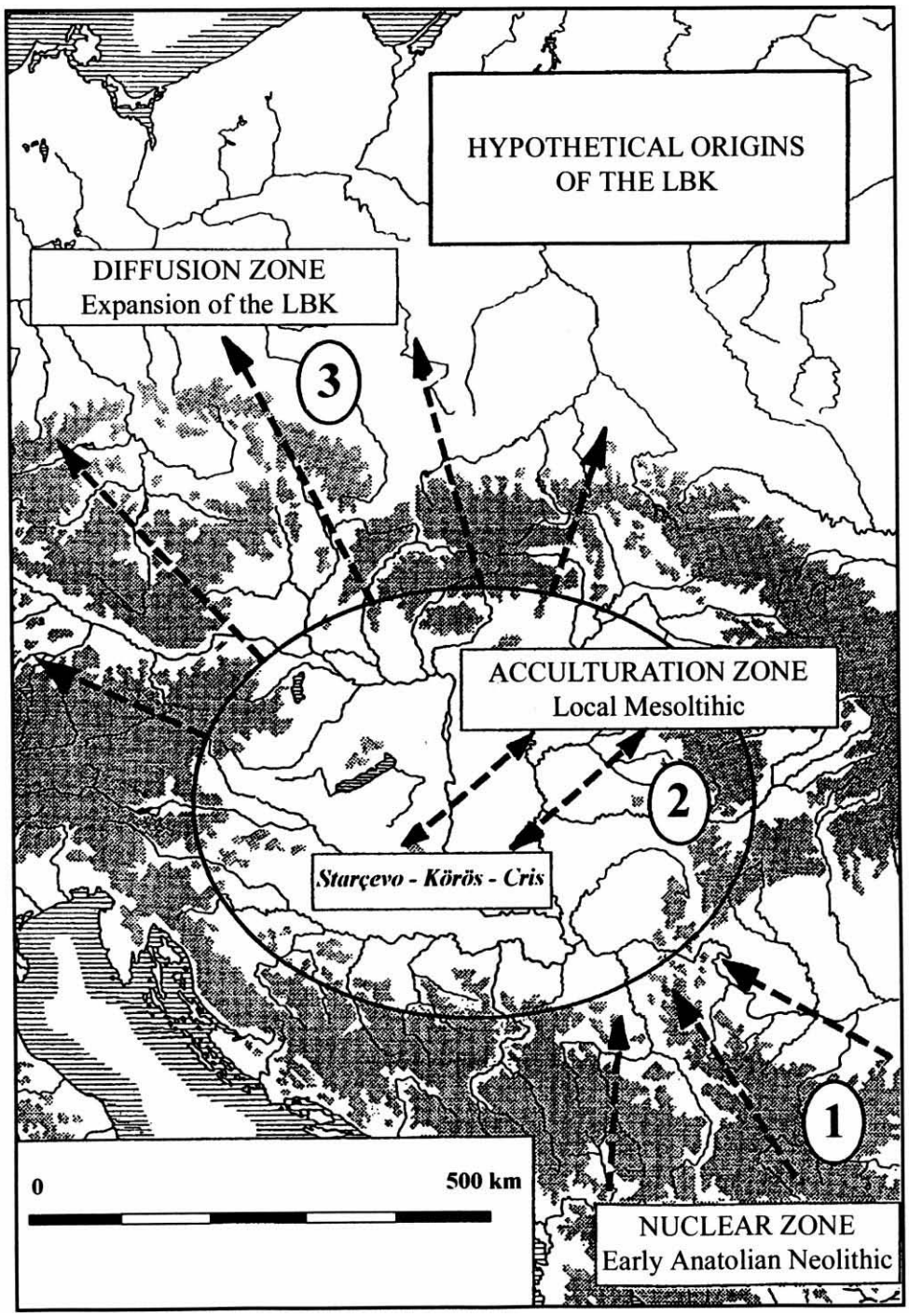

Figure 6
The LBK appears as a movement of colonization from a cultural center in modern Hungary. The settlers kept contact with nuclear areas and remained stable (after Otte in press)

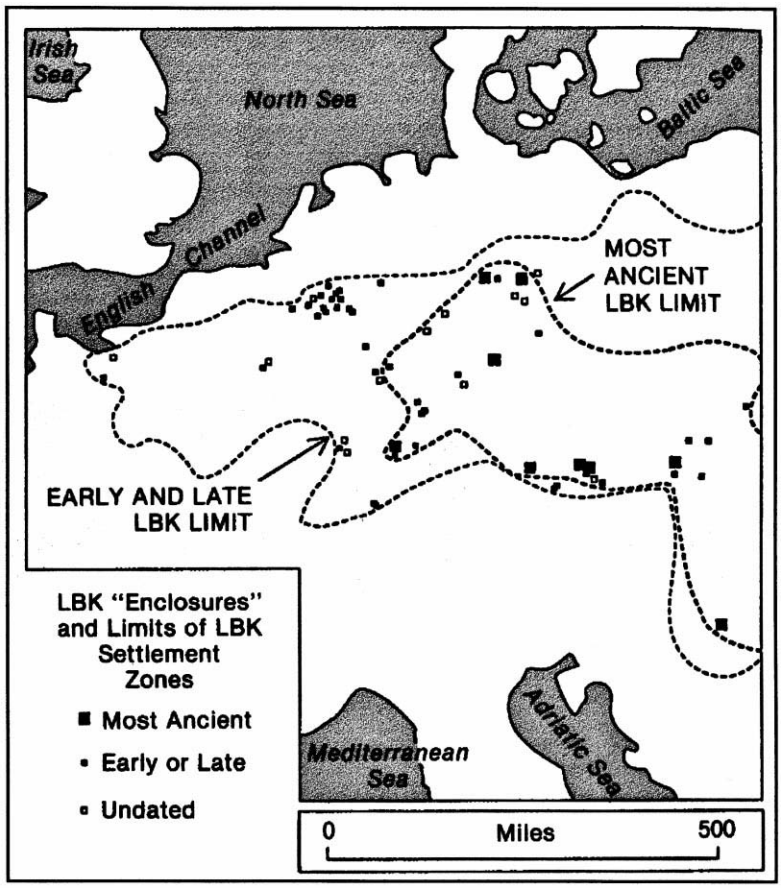

Figure 7
The LBK expansion represents the clearest model of prehistoric colonization: artifacts remained unchanged over a vast territory and are radically different from those found in preceding periods in the regions crossed (after Keeley 1996)

than simple adaptation to new environments. Some classic cases are known. One example is the arrival of modern humans in Europe, evidence of which is given both by a new anatomical form and by a completely novel system of symbolic and technological behavior. We can be certain at least that a migratory wave from external populations can be revealed at the level of their customs, and attested by archaeology. A point of contact is then established, as it was with the Celts or the Germans, whose migrations were—elsewhere—attested by texts. A double control confirms the validity of archaeological data in the study of migration: anthropological on one hand, historical on the other. The Aurignacian trail crossed a Europe already regularly occupied by Neanderthals of the Mousterian culture.

The two milieus are clearly opposed, according to archaeological data, but on the margins, acculturation phenomena were produced locally, evidenced by 'mixed' assemblages: the Chatelperronian industry, foliate point cultures (Fig. 5). In this way, tertiary phenomena (acculturation) emerge, from which later cultures develop. The case of the Aurignacian, at first clearly imposing itself, then completely disappearing in favour of a local Gravettian, forms a good example of migrations melting into a new and finally dominant population. These types of migrations to already occupied regions, occur throughout the Palaeolithic with more or less clarity.

Another instructive case is illustrated by the southern Solutrean industry which includes one element that seems to originate in North Africa and another from northern Gravettian migrations. When these two movements from north and south met, a new region of acculturation made its appearance, provoking the creation of a culture (the Solutrean) lasting

Figure 8
Small groups of Bell Beaker people had an enormous territorial expansion. They are characterized by graves, a particular ceramic form and copper objects. Their distribution, superposed on to local traditions, evokes the diaspora observed in historical times (after Briard, in Mohen 1996)

several thousand years in Southwest Europe. Only general artistic trends reflect continuity with the preceding period, tracing their development under a harmonious form. Other elements of a stylistic character are profoundly modified at the end of the period of contact between these two opposing migratory waves.

A third migratory model also appears within the archaeological domain: the establishment of colonies, far from their points of origin but maintaining contact via exchanges with the intervening regions. This mode of colonization is well known from recent historical periods: Greek and Roman colonies, and later European colonial expansion. Colonial migrations first appear during the LBK (Linearbandkeramik) period of the Neolithic, six thousand years ago, when early Neolithic groups expanded across the European continent from the east and south-east.

The settlers traveled in a distinctive manner, under a permanent form but without modifying their own way of life, in this way maintaining contact with their original culture. The LBK people had an architecture and an urbanism in balance with a stable economy (agriculture and animal husbandry) and an effective and appropriate technology (production of ceramics, grindstones, polished adzes). A coherent way of life was harmonized with a new, totally mastered, environment. Contacts with indigenous populations modified the internal equilibrium acquired by the new arrivals. Archaeological data clearly illustrate this migratory model, with a great constancy presented by

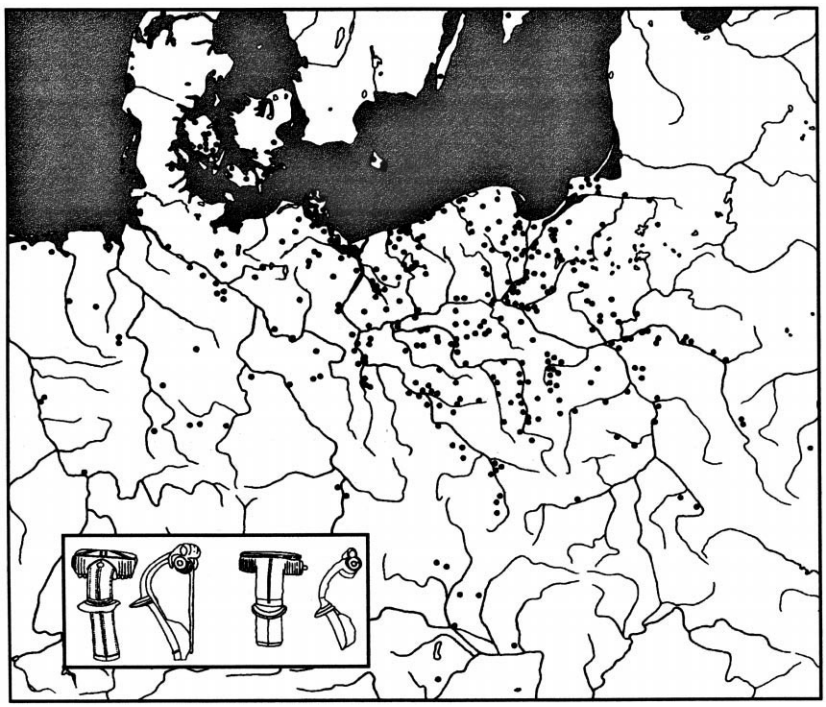

Figure 9
Methodological example of distribution of a particular type of decorative elements along the Baltic coast. These Germanic elements identify an ethnic group and permit us to follow their migrations until the time of Christ (after Böhm 1996)

habitats and ceramic styles (Figs. 6–7). Contacts are revealed by the long-distance exchange of raw materials from previously occupied regions. The impression of a strong cultural unit thus dominates the entire area of expansion of the archaeological culture.

Finally, the diaspora shows a people whose migrations are integrated into other milieus within disturbing their existence and nature. Modern cases are eloquent and well known, such as the Armenians, the gypsies and the Jews. Archaeological cases of widely dispersed cultures superimposed on to local traditions yet remaining stable, are also known. For example, the Bell Beaker people during the Neolithic period, identified by the presence of distinctive, globular goblets, were spread across western Europe, particularly along seacoasts and rivers. Their simple burials, individual and with highly standardized grave goods, suggests the existence of a network of traveling merchants, perhaps linked to the exchange of the first metals (Fig. 8). Cultural links remain obvious between these different populations, because they participated in an extended exchange network over a vast territory. However, not a single center seems to exist as a unifying and referential model for the culture.

Migrations in archaeological contexts are thus frequent and refer to all humanity. We are able to recognize clear traces throughout virgin territories such as the Americas or northern Europe: here it is the artifacts themselves which directly retrace the paths of human expansion. Stylistic criteria permit us to recognize different cultures and to follow their movement across space and through time.

Mitigating cases where such migrations are superimposed on indigenous populations demand a different and more elaborate methodology. The styles of equipment must be compared between the two groups in order not only to distinguish them but also to identify possible symbioses or influences. Maps of dispersion aid, stage by stage, in understanding archaeological evidence across space (Fig. 9). In order to distinguish the influences of possible convergences, and similarities produced in different environments but at similar developmental stages, critical analysis must be applied to the archaeological record. The classic case, for example, is the Late Mousterian in Europe, where the same 'inventions' are known across all cultural environments in different regions but in similar circumstances. Migration can only appear in the global distribution and cohesion of a phenomenon, paired with territorial expansion. The more stylistic

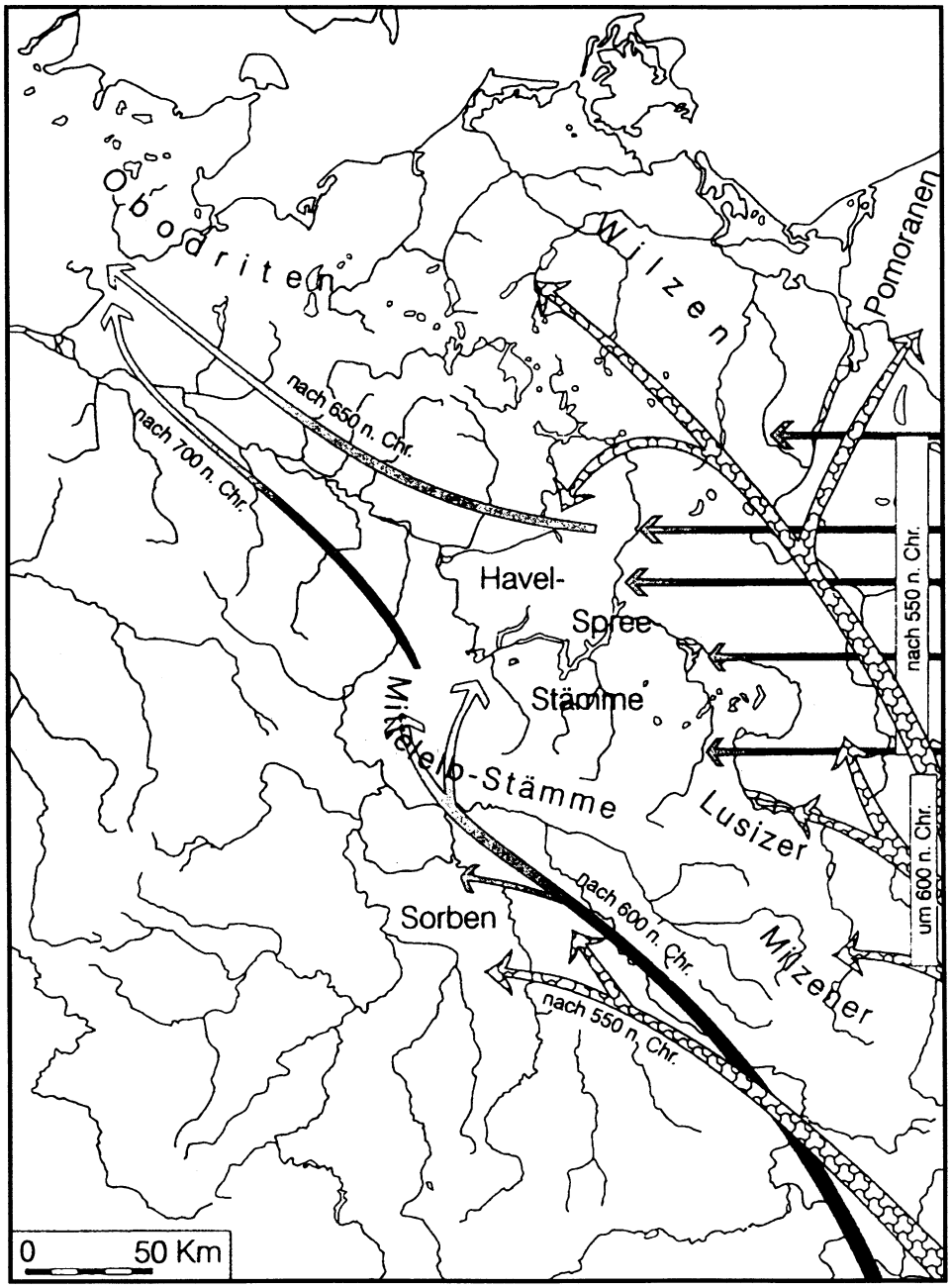

Figure 10
Historical migrations (here, the Slavs) are reconstructed by the distribution of languages, written sources and archaeological information (after Kóčka-Krenz 1996)

criteria are elaborated, the less random they are, and the more therefore they correspond to the system of values conveyed by a homogeneous population. It is this general key which archaeologists follow in their reconstructions of displacement. Although this law is not infallible and demands constant vigilance, comparisons with recent culturally well-defined peoples, permit us to ensure its effectiveness (Fig. 10).

See also: Demographic Techniques: Data Adjustment and Correction

Bibliography

Böhme H W 1996 Kontinuität und Traditionen bei Wanderungsbewegungen im frühmittelalterlichen Europa vom 1–6 Jahrhundert. *Archäologische Informationen* **19**: 89–103

Dixon J 1999 Late Pleistocene maritime adaptations and colonisation in the Americas. *Pre-prints of the World Archaeological Congress* **4**: 10–14

Garanger J 1987 Le peuplement de l'Océanie insulaire. *L'Anthropologie* **91**(3): 803–16

Keeley L H 1996 *War Before Civilization*. Oxford University Press, Oxford

Klein R 1989 *The Human Career*. The University of Chicago Press, Chicago

Kóčka-Krenz H 1996 Die Westwanderung der Slawen. *Archäologische Informationen* **19**: 125–34

Mohen J-P (ed.) 1996 *La vie préhistorique*. Faton, Dijon, France

Otte M 1995 Traditions bifaces. In: *Les industries à pointes foliacées d'Europe centrale*. *Paléo* supplement no. 1 195–200

Otte M 1996 Aires culturelles au Paléolithique supérieur d'Europe. In: Mohen J-P (ed.) *La vie préhistorique*. Faton, Dijon, France, pp. 286–9

Otte M 1997a Contacts trans-méditerranéens au Paléolithique. In: Fullola J M, Soler N (eds.) *El món mediterrani després del Pleniglacial (18.000–12.000BP)*. Museu d'Arqueologia de Catalunya, Girona, pp. 29–39

Otte M 1997b Paléolithique final du nord-ouest, migrations et saisons. In: Fagnart J-P, Thévenin A (eds.) *Le Tardiglaciaire du Nord-Ouest de l'Europe*. CTHS, Paris, pp. 353–66

Otte M in press. Le Mésolithique du Bassin Pannonien et la formation du Rubané. In: *Proceedings of the Conference 'From the Mesolithic to the Neolithic'*, Szolnok, 996

Roe D A 1981 *The Lower and Middle Palaeolithic in Britain*. Routledge, London

Wolpoff M 1998 *Paleoanthopology*. McGraw-Hill, Maidenhead, UK

M. Otte

Military and Disaster Psychiatry

1. Introduction

Whether by force of humans or nature, massive destruction creates an atmosphere of chaos and compels individuals to face the terror of unexpected injury, loss and death. In times of disaster or war, psychological injury may occur as a consequence of exposure to physical injury, disruption of the environment, or the terror or helplessness produced by these events. To address such injury in a timely manner, mental health care must be provided in environments near chaos and destruction, as well as in the hospital. The disciplines of military and disaster psychiatry address care demands in nontraditional environments and in mass casualty situations, where resources are overwhelmed. Care in these environments relies on contributions not only from psychiatrists, but also from other physicians, social scientists, epidemiologists, psychologists, nurses and emergency responders such as police and firemen.

This overview of military and disaster psychiatry begins with an examination of the consequences of disasters and wars for communities, and the evolution of medical responses to these traumatic experiences. A discussion of the phenomenology of trauma-related psychiatric morbidity and principles of prevention, mitigation of consequences, and management follows. Finally, transnational economic, ethical and legal trends are presented as issues requiring further study.

2. Practice Environments in Military Operations and Disasters

'Disaster' has numerous definitions. The word is derived from the Latin *dis* ('against') and *astrum* ('stars')—'the stars are evil'. A disaster such as an earthquake or a flood overwhelms a community's capacity to respond. The distinction between 'natural' disasters (e.g., earthquakes) and human-made or technological ones such as explosions, or train derailments is increasingly difficult to make. For example, much of the death and destruction from an earthquake may be due to poorly constructed housing—thus, there is a human-made element to the consequences of even 'natural' disaster. From a psychological standpoint, a more critical distinction concerns whether the disaster was inflicted intentionally, as is the case with acts of war or terrorism.

War may be defined as a political act (generally involving violence) to achieve national objectives or protect national interests. During the last 30 years, militaries around the world have increasingly become involved in peacekeeping and humanitarian relief missions. The use of military forces in these endeavors also maintains a country's influence and minimizes political instability in the affected nation.

The potential stressors in all disaster environments include exposure to the dead and grotesque, threat to life, loss of loved ones, loss of property, and physical injury. Although the military brings supplies and a portable living environment to protect soldiers, civilians (frequently exposed to combat environments in modern times) may be subject to large-scale devastation, become refugees, and experience shortages that threaten life. Frequently, such victims do not receive treatment for psychiatric symptoms that emerge from bombings, battle, rape, torture and unrestrained murder. Although an earthquake may be concluded in seconds, the consequent traumatic experience may

continue for weeks, months and possibly years. For both soldiers and civilians in combat environments, exposure over time may include anticipated or entirely unexpected life-threatening experiences followed by daily life in an austere and disrupted environment.

The emotional and behavioral responses following a disaster occur in four phases. The first immediately following a disaster generally consists of strong emotions including feelings of disbelief, numbness, fear and confusion—normal emotional responses to an abnormal event. The second phase usually lasts from a week to several months and is accompanied by the appearance of assistance from outside agencies and communities. Adaptation to the austere environment as well as intrusive symptoms (unbidden thoughts and feelings accompanied by hyper-arousal) occur during this phase. Somatic symptoms such as fatigue, dizziness, headaches and nausea may develop. Anger, irritability, apathy, and social withdrawal are often present. The third phase is marked by feelings of disappointment and resentment when hopes for aid and restoration are not met. Here, often, the sense of community is weakened as individuals focus on their personal needs. The final phase, reconstruction, may last for years. During this period, survivors rebuild their lives, make homes and find work using available social supports. Individuals may progress through these phases at various rates. Many persons may be unable to reconstruct their lives fully and instead develop persistent symptoms.

The causes of disaster and war have been historically attributed to sources ranging from the gods, to the wind of a passing cannonball, and various natural, unnatural or supernatural sources of contagion. Emotional consequences of disaster are described in the *Iliad*, and references to the terror induced by the attack of this hero are diverse. Ancient Greeks attributed epidemic illness to Apollo's wrath after the desecration of his temple. The French military surgeon Larrey commented clearly on the ill effects of war upon the health of Napoleon's soldiers. Others commented on combat-related pathological behaviors during the US Civil War, and recent studies have noted the descriptions of veterans of that war hospitalized for symptoms very similar to those of today's Post Traumatic Stress Disorder (PTSD).

The science of neurology entered military medicine with Weir Mitchell's work during and after the Civil War. Over the remainder of the nineteenth and twentieth centuries studies increasingly distinguished between diseases of the nervous system for which traumatic lesions could be demonstrated and those for which no such lesion could be identified. The concepts of neurasthenia, dissociation, hysteria and psychological suggestion were developed to define psycho–neurological states without demonstrable anatomic abnormality.

Military physicians in the Russo–Japanese War made similar diagnostic distinctions. Recognition of the 'nontraumatic' injuries that followed railway accidents and other technological disasters occurred at the same time. Military psychiatric experience in World Wars I and II led to the development of specific treatment principles. During World War I, physicians from various armies addressed the problem of soldiers with emotional or behavioral disturbances with a variety of diagnostic labels such as shell shock, gas neurosis and conversion paralysis. Treatment ranged from prolonged psychiatric hospitalization, to punishing electric shock and various talk therapies. Gradually, US, Canadian and British forces incorporated into their treatments the expectation that these soldiers return to battle after brief evaluations. German military scientists recognized the importance of unit cohesion in mitigating psychological injury. Elsewhere, efforts were made to screen out soldiers felt to be at risk for psychological disturbances on the assumption that these soldiers were genetically weak. Although the terms proximity (treatment near the combat zone), immediacy (early identification of stress-related disorders), simplicity (treatment with rest, food and brief support) and expectancy (expectation of prompt recovery and return to duty) were defined in later conflicts, these practices evolved to varying degrees during World War II. These principles, along with the development of psychotropic medications, the failures of screening programs, and the recognition of the problems of drug abuse in operational environments greatly influenced the management practices of subsequent military and disaster responders.

Civilian physicians have also long recognized the trauma of war as a cause of human suffering. In 1859, Jean Henri Dunant arranged for civilian medical services for the injured after observing soldiers die from lack of medical attention during the Battle of Solferino. His efforts led to the establishment of the International Red Cross, and to international guidelines for humane care to the sick and wounded in times of war. During the later part of the twentieth century the Red Cross, and other international medical and relief agencies such as Doctors Without Borders increasingly provided mental health-related consultation, education and direct care in the aftermath of war, natural and human-made disasters. The World Health Organization and the Pacific–Asian Health Organization have also supported international disaster relief efforts.

3. *Phenomenology*

3.1 *Symptoms versus Functioning*

Military and disaster psychiatry must address the clinical concerns of identified patients, but must also strive to prevent potentially incapacitating morbidity

in entire populations. Distress-related symptoms are universal during disasters and combat. Initial psychiatric response in the aftermath of war and disaster must focus on mobilizing effective functioning. Symptoms occurring in persons who are not impaired are a secondary concern. Such symptoms can become 'medicalized' if clinicians cause impaired functioning by unjustifiably reinforcing a view that symptoms are due to a disease. While the ultimate label given to clusters of symptoms has political, economic and research-related significance, the self-perception that one is ill can become a powerful determinant of impaired functioning both during and after combat and disasters.

3.2 Military Operations, Disasters and Psychiatric Syndromes

Much of military and disaster psychiatry focuses on the myriad behavioral reactions to stressful events—'stressors'. Well-defined psychiatric syndromes describe many of these responses. The precipitating stressor for PTSD involves a threat to the physical integrity of self or others, so immediate that the exposed individual suffers a potent sense of helplessness, horror or fear. A characteristic distress response may follow such trauma. This response consists of symptoms that involve: (a) 'reliving' the original event (e.g., nightmares, distressing vivid recollections or fear when exposed to events resembling the original trauma); (b) numbing of responsiveness or behavioral avoidance of events or situations that somehow resemble or symbolize the original trauma; and (c) symptoms of increased vigilance, such as exaggerated startle, outbursts of anger or other evidence of hyper-arousal. If these symptoms of severe distress persist for over a month, then a diagnosis of PTSD is appropriate. Symptoms may first occur months or even years after the triggering event, but this is not the norm. If symptoms occur within the first month after the trauma and have not lasted longer than a month, then Acute Stress Disorder (ASD) is diagnosed. Controversy persists regarding the diagnostic validity PTSD, probably because it was defined in the aftermath of the Vietnam War in the wake of political and antiwar pressures. Nonetheless, PTSD and ASD are conceptualized as modal distress responses to severe or catastrophic stressors, and have been as carefully defined and delineated as other psychiatric disorders.

Disabling distress reactions occur in response to less significant trauma and present in patterns not described by PTSD or ASD. Adjustment Disorder, for example, is a maladaptive behavioral and/or emotional response to a diverse array of stressors. Conversion Disorder may be diagnosed when one develops unexplained symptoms or deficits affecting voluntary motor or sensory function (e.g., sudden paralysis of the 'trigger' finger) without demonstrable neuro–anatomical injury. Bereavement, a normal grief reaction after the death of someone who is valued or loved, may also occur in response to losses incurred during war or disaster. Other distress responses to war or disaster include anxiety and depressive syndromes, and antisocial behavior (involving acts of violence, criminal behavior, military misconduct, or war-related atrocities). Alterations in health-related behaviors (e.g., misuse of tobacco, drugs or alcohol, poor eating habits) may also develop after exposure to disaster or war.

3.3 Battle Fatigue

The term 'battle fatigue' provides a framework to encompass the variety of responses to operational stress, but does not define a specific constellation of symptoms, as in Major Depressive Disorder or PTSD. A wide range of physical and emotional symptoms and signs can occur among individuals with battle fatigue including gastrointestinal distress, tremulousness, anxiety, perceptual disturbance, a sense of unreality, and a dazed look (i.e., 'thousand-yard stare'). The diversity and non-specific nature of presentation distinguish this entity from ASD. Battle fatigue occurs in combatants who have exhausted physiological and psychosocial coping mechanisms with the intense combat experience. Minor injury, parasitic infection, starvation, heat exhaustion, and cold injury may decrease the coping resources of a combatant.

3.4 Medically Unexplained Physical Symptoms

War historians have observed that unexplained physical symptom syndromes are common sequelae of combat since at least the US Civil War. Syndromes such as 'soldier's heart' and illnesses characterized by physical symptoms attributed (by sufferers) to war-related exposure to Agent Orange are examples. Contentious debates between scientists, clinicians, veterans and their advocates, and journalists persist around putative etiology. Some argue that the consistent appearance of these syndromes after war speaks to the likelihood that psychosocial factors contribute to their etiology.

3.5 Other Psychiatric Illnesses

Depression, anxiety disorders and personality changes have all been associated with exposure to the trauma of disaster and war. These psychiatric disorders may be accompanied by somatic complaints. Such illnesses have been described in large numbers of persons exposed neither to war nor other disasters. Therefore, biological, genetic and environmental risk factors are all likely involved in the development of these illnesses.

4. Etiology and Epidemiology

4.1 Predisposing Factors

PTSD, other anxiety and depressive disorders, and physical symptom syndromes are more frequently diagnosed among women than men in association with any given stressor. Explanations for this involve neurobiological and psychosocial factors including the greater rate at which women seek treatment for stress-related symptoms and that duration of illness (e.g., PTSD) may be longer for women and therefore more likely reach clinical attention. Men are at higher risk for post-war problems with alcohol and substance use, and antisocial and violent behavior. Gender-specific neurophysiological factors as well as cultural factors are again implicated in these differences.

Level of functioning after combat and disasters also relates to pre-trauma functioning. Individuals who function marginally in various roles (e.g., occupational and social) prior to disaster or combat exposure are at increased risk for poor functioning after trauma compared with individuals who were previously high functioning. Individuals who have successfully negotiated past traumatic experiences may be resilient ('hardened') in similar future situations. However, if past traumatic events resulted in PTSD or psychiatric distress syndromes, subsequent traumatic exposures may make future episodes of these disorders more likely.

4.2 Protective Factors

Protective factors may be present to varying degrees in groups such as military units, police, or firefighters exposed to trauma. Strong leadership can create powerful loyalty and interpersonal cohesion with populations. Potent leaders can create a unit dynamic wherein leaders are so valued and trusted by members of the unit as to enable voluntary participation in extremely high-risk combat or rescue–recovery situations. A common symptom of poor leadership is the occurrence of destructive inter-group conflicts and organizational splits.

An axiom of professional soldiers is 'we will fight as we have trained, therefore we must train as we expect to fight'. If the level of training is high, individuals in the unit (military or civilian) more frequently trust ingrained basic principles aimed at supporting one another in a quest for mission success. Recently, nonmilitary disaster responders (firemen, police, physicians and civic leaders) in developed countries have assembled to train for response to terrorist attack or natural disaster. Government emergency preparedness agencies such as the US Federal Emergency Management Agency are increasingly coordinating such training. The quality and extent of fit between persons, equipment (e.g., comfort and mobility of chemical protective suits, familiarity with operation of remotely controlled bomb or mine detectors or personnel recovery devices), and the environment may modify stress responses of individuals or communities. The extent to which the living or working environment may modify response is evident in studies of those forced to exist in close quarters for extended periods of time with only limited contact with the outside world, such as those aboard ships or submarines. The 'fit' between pilot and aircraft as well as between aircrew members may be improved through specific training. Finally, the effectiveness (or perceived effectiveness) of leadership response to crisis is a factor that may modify community response.

4.3 Precipitating Factors

Precipitating factors are the proximate circumstances that initiate the various sequelae of trauma. For disaster responders and military populations, deployments and peacekeeping missions disrupt families and are often 'poorly timed' with regard to other life events. High intensity and duration of disaster or combat exposure relate directly to the likelihood of psychiatric casualties. Specific experiences, such as physical injury, witnessing grotesque deaths, torture or other atrocities place individuals at increased risk for adverse mental health consequences. Victimization in the form of rape, harassment, or assault can precipitate distress reactions in those victimized. Sexual assault is a potent precipitant of adverse neurobehavioral changes.

4.4 Mitigating and Perpetuating Factors

Ongoing factors, including the security and safety of recovery environments, extent of secondary traumatization and—in military populations—rotation schedules, extent of recognition or compensation for efforts and belief in the mission effect the rate and severity of distress symptoms. Symptoms in civilian victims of war or in the aftermath of disaster may be mitigated or exacerbated by perceptions of community leadership's preparedness for disaster, response to crisis, recognition of 'heroes', and provision of medical, financial or emotional assistance both immediately after crisis, and over time.

Nonmilitary, nongovernmental organizations such as the American Red Cross and the Salvation Army help to minimize the stress following a disaster. By attending to basic human needs such as food, clothing and shelter, they reduce both the psychological and the physiological effects of the event. In recent years the Red Cross has developed training for volunteer health care workers to recognize, minimize and treat stress responses in disaster workers and victims of disaster.

5. Management and Care Delivery

5.1 General Principles

Often disasters or military conflicts shatter the expectation of a just and safe world within populations where notions of basic justice and safety are cultural norms. In such populations, establishing the sense of safety and expectation of justice is an important aspect of recovery. Other interventions vary with the stage of the disaster. Initially, establishing a safe environment, and managing life-threatening injury and disease possibilities, such as those resulting from infection or absence of potable water, can be the most important psychiatric interventions. Subsequently, identifying high-risk populations such as disaster workers, firefighters, police, persons at impact zones and children can focus intervention strategies. Outreach programs are critical, since disaster victims rarely seek mental health care. Those who are physically injured are also at great risk for psychiatric disturbance. Educating medical and community groups about normal responses to abnormal events as well as when mental health referral is indicated is an important part of outreach programs. Advising community leaders on expected behavioral problems and needs is required to ensure availability of resources to care for victims. This work must involve planning for expected natural or human-made disasters, and allocating funds for the care of anticipated victims before disasters actually occur.

Responsibility for preventive measures, and recognition and treatment of the psychological consequences of such wars and disaster cannot be limited to the few (if any) available psychiatrists. General physicians, psychologists and other social scientists must use their diverse skills to care for disaster and war victims. They must diagnose and treat disorders associated with trauma, (e.g., PTSD, depression and anxiety disorders), provide consultation to medical and surgical colleagues and other first responders, and educate community leaders about predictable responses to abnormal events.

5.2 Military Mental Health Care

The US military has attempted to decrease the incidence and severity of combat and operationally induced psychiatric disorders. Mental health teams are now routinely assigned to US forces in combat and deployed operations other than war. Each branch of the US military service has specialized rapid intervention teams to provide consultation and acute treatment to units that have experienced traumatic events. These teams instruct commanders on likely behavioral responses to stress and recommend leadership actions that may reduce negative responses to stressful situations. Post-incident debriefings assess the effect on the unit, and attempt to reduce long-term consequences of traumatic events. Open discussion of an incident is believed to foster unit cohesion, facilitate accurate individual and group understanding, and reduce the development of psychiatric disorders. However, in the few groups actually studied, there is no convincing evidence that acute incident debriefing has any effect on the later development of psychiatric illness. Debriefings may be useful in identifying individuals who require further mental health attention and decreasing individual isolation and stigma.

Despite the absence of consensus data supporting their effectiveness, there is increasing interest in expanding the use of rapid intervention teams. The US military currently proposes to establish a unified, multi-service policy on the composition and use of these teams. This effort follows the widely publicized 'Gulf War Illness' complaints of veterans from that campaign. Some believe that since these symptoms are largely tied to psychological problems, increased attention to stress during military operations could have reduced their incidence or severity.

Different missions, patterns of deployment, and medical support systems among US military services pose major problems to the development of a unified approach to managing operational stress. Armies typically deploy large units for extensive periods of time and allocate large amounts of medical assets to support these units. This medical support includes specialty services. The US Navy and Marine Corps deploy smaller units both at sea and ashore. General medical officers and nonphysician providers furnish medical support, and specialty care is not routinely available in the operational theater. The US Air Force has both short- and long-range missions. Operational stress management doctrine must consider these differences. Military physicians also provide medical and psychiatric assistance to civilian populations in times of natural and human-made disasters. In addition to direct patient care, military psychiatrists consult with community leaders and with civilian physicians not accustomed to responding to large-scale physical and emotional traumas.

In the USA, definitive treatment of psychiatric illness is often provided in the military's system of hospitals. Medical care is provided to active duty personnel and to their families. Other mental health specialists, nurses, social workers and psychologists augment this care. Military members who develop psychiatric disorders while on active duty are eligible for medical retirement disability pay, and continued treatment through a system of Veterans Administration hospitals. Individuals may be separated from service due to personality problems without disability payment or ongoing medical care from the military.

Other nations with recent wartime experience, such as Israel and Croatia, have developed programs to evaluate and treat soldiers and civilians exposed to combat. Their experiences are somewhat different

from the US, since they rely much more heavily on reserve forces. These nations have a more inclusive social medical infrastructure therefore treatment programs are less reliant on the military medical system. Other nations are increasingly confronted with management of operational stress in peacekeeping and humanitarian missions. Asian nations that have recently experienced natural disasters and terrorist events are also studying approaches to evaluating and treating individuals exposed to trauma.

5.3 Medical Education

Several nations provide medical education specifically for members of their armed forces. The US Congress in 1975 authorized The Uniformed Services University of the Health Sciences (USUHS) to provide medical education and produce physicians for military service. USUHS provides a four-year medical degree program and a number of graduate degree programs in the basic and clinical sciences. The USUHS Center for the Study of Traumatic Stress conducts research, and consults to communities, and federal and international agencies on matters surrounding individual and community responses to trauma, disaster and war. Japan, the UK and Russia are among nations with institutions that teach military specific curriculum to military medical care providers. As in other nations, these countries also call to national duty physicians not specifically trained in military institutions during times of war or crisis.

6. Future Challenges and Evolving Issues

As political, social, scientific, and technological factors evolve, societies will change their responses to the consequences of disasters and wars. Psychiatric practice associated with wars and disaster has changed with the evolution of scientific understanding of illness. In the future, the resources to deal with the consequences of disaster or war and the relative importance assigned to dealing with the resultant injuries and disabilities are likely to be influenced by political and socio–cultural values.

6.1 Ethical Challenges

The hyper-suggestibility of recently traumatized individuals has provided an occasion for exercising political influence and manipulating loyalties. Providing care in the mass casualty situation raises ethical questions about the equitable distribution of resources and the moral values to consider in determining their apportionment. Governments in trouble have withheld treatment to minority racial or political groups—clearly an ethical breach. Since governmental terrorism is a common form of terrorism, care providers and leaders must be sensitive to the possibility that disasters will afford tyrants an opportunity to manipulate citizens for their own purposes.

To facilitate command assessment of troop health status, militaries have denied members confidentiality in medical communication. Mental healthcare providers must strike a balance between a promise of privacy that encourages persons to seek care, and responsible reporting to higher command regarding situations that pose danger to larger groups. Thus dual allegiance to both individuals and to the larger community presents an ethical challenge that must be negotiated by the military care provider. Persons in extreme circumstances may behave in ways that they later view as shameful. Shame may contribute to posttraumatic symptoms and disturb one's capacity to use social supports. Disaster triage is frequently carried out in large open areas that allow everyone present to hear what patients say to caregivers. Given the social stigma assigned to the manifestations of psychiatric illness it is easy to understand both patients' reluctance to communicate and doctors' reluctance to inquire. Perhaps re-educating the population can reduce ethical and therapeutic problems associated with stigma. However, altering deeply ingrained cultural expectations is just as challenging as providing privacy in chaotic triage environments.

6.2 Technological Advances

New technologies in combat will modify the means of sorting and treating persons with medical and psychiatric injury. Future militaries in technologically advanced nations are likely to become much smaller, move rapidly across the battlefield, use advanced sensors, and direct intense fire across a considerable distance. These capabilities, coupled with the possible use of weapons of mass destruction, will likely make the battlefield more chaotic and inhospitable to human life. Emergency care and evacuation of those with disease and injury may become increasingly difficult. The inability to maintain contact with rapidly moving units may preclude returning individuals to their original units. Future military casualties may increasingly rely on care by unit buddies, medics and frontline leaders rather than specialized medical units or specialists at hospitals in the rear.

Underdeveloped nations may have limited access to advanced technologies, so more traditional ways of organizing medical and psychiatric practice may continue to be relevant.

The evolution of highly mobile units on widely disbursed battlefields will decrease the opportunity for exchanging rested troops from the rear area for those exhausted by frontline combat. Provision of brief respite for exhausted troops—a hallmark of management of battle fatigue—may become impossible

as each individual may be performing a critical specialized task. Small medical units operating within the area of combat are likely to be eliminated from this technology-intensive battlefield. While treatment may by necessity move to the battlefront, medical specialists at the rear may render triage decisions and diagnoses through the use of telemedicine communication technology. Experience has shown that front-line mental health providers take a pragmatic view of acute psychiatric symptoms, and tend not to make hasty formal diagnoses on overstressed troops. Rear-echelon providers, by contrast, tend to assign formal psychiatric labels that may be inaccurate and may stigmatize troops without contributing to treatment. Rear-echelon mental health specialists in future battles must address the challenge of providing useful therapeutic advice from afar while avoiding meaningless diagnostic stigmatization.

Advanced technology will have similar implications for those responding to human-made disasters such as terrorist attacks especially as terrorists gain increased access to so called weapons of mass destruction (chemical and biological agents). Clarifying the roles of military and civilian responders in terms of triage, treatment, consultation and education in any joint response to crisis is another challenge for military and disaster psychiatrists.

6.3 Cultural Issues

Social scientists note that responses to trauma may be considered either normal or pathological, depending on the interested party. Many have expressed fear that mental health practitioners, motivated by profit, will try to convince individuals experiencing normal uncomfortable responses that they need treatment. Overcoming this fear or the belief that accepting assistance signals weakness is a challenge in circumstances where necessary and available external assistance is rejected by a nation in crisis.

Most individuals exposed to traumatic events deserve to be reassured that with return to work, community and family, they will recover. However, some individuals (and perhaps some cultures) will experience greater psychopathologic responses and more prolonged symptoms following trauma. The nature of the behaviors and symptoms associated with trauma response across cultures is still uncertain. The extent to which social supports, biological–genetic predisposition, concurrent illnesses, and other political economic parameters contribute to variability across cultures is not known. While it appears clear that the severity of the trauma is important, reliable measures of severity remain to be determined. Trauma may, under some conditions, create the opportunity for personal growth as well, and further understanding of this potential must also be exploited to reduce morbidity.

7. Conclusion

Natural and human-made disasters result in traumatic disruption of societal function. Wars and acts of terrorism with their attendant large-scale death, injury and destruction affect populations in much the same way as massive natural disasters. Whatever the cause, disasters and military operations leave in their wake populations experiencing psychological disturbances that have been described by social scientists, civil leaders, physicians and other care providers throughout the ages.

The consequences of exposure to disaster and war may take the form of psychological disorders such as PTSD or may manifest as various (and sometimes more subtle) forms of behavioral change, anxiety or depression. Symptoms may present at different times during and after traumatic exposure. Many factors complicate the evaluation and treatment of neuropsychological syndromes in the aftermath of war or disaster. Resources are overwhelmed, life-threatening illnesses require immediate treatment, and psychological casualties are often reluctant to seek assistance.

Progress has been made in identifying the nature of trauma-related psychological responses. Predisposing, exacerbating and mitigating factors have been identified. The value of multidisciplinary preparation and training for disaster management and the need for outreach programs have also been demonstrated. Further study will focus and clarify the roles of psychotropic medications and various forms of psychosocial support and psychotherapy in the treatment of war and disaster-related morbidity. With technological advances and global economic shifts, the nature of war and other human-made disasters will change. Military and disaster mental health care delivery must anticipate such changes to develop improved methods of prevention, evaluation, and care for individuals and groups devastated by war or disaster.

See also: Disasters, Coping with; Disasters, Sociology of; Military Psychology: United States; Post-traumatic Stress Disorder; Reconstruction/Disaster Planning: Germany; Reconstruction/Disaster Planning: Japan; Reconstruction/Disaster Planning: United States

Bibliography

Geiger H J, Cooke-Deegan R M 1993 The role of physicians in conflicts and humanitarian crises: Case studies from the field missions of physicians for human rights, 1988 to 1993. *Journal of the American Medical Association* **270**: 616–20

Glass A J (ed.) *Neuropsychiatry in World War II: II. Overseas*. US Government Printing Office, Washington, DC

Jones F D, Sparacino L R, Wilcox V L, Rothberg J M (eds.) 1995 War psychiatry. In: *Textbook of Military Medicine. Part I, Warfare, Weaponry, and the Casualty*. Office of the Surgeon General, Washington, DC

Jones J C, Barlow D H 1990 The etiology of posttraumatic stress disorder. *Clinical Psychology Review* **10**: 299–328
Holloway H C, Benedek D M 1999 The changing face of terrorism and military psychiatry. *Psychiatric Annals* **29**: 363–75
Iacopino V, Waldman R J 1999 War and health: From Solferino to Kosovo—the evolving role of physicians. *Journal of the American Medical Association* **282**: 479–81
Mollica R F, McInnes K, Sarajlić N, Lavelle J, Sarajlić I, Massagli M P 1999 Disability associated with psychiatric comorbidity and health status in Bosnian refugees living in Croatia. *Journal of the American Medical Association* **282**: 433–39
Shalev A Y, Solomon Z S 1996 The threat and fear of missile attack on Israelis in the Gulf War. In: Ursano R J, Norwood A E (eds.) *Emotional Aftermath of the Persian Gulf War: Veterans, Families, Communities and Nations*, 1st edn. American Psychiatric Press, Washington DC, pp. 143–62
Ursano R J, Holloway H C 1985 Military psychiatry. In: Kaplan H I, Sadock B J (eds.) *Comprehensive Textbook of Psychiatry*, 4th edn. William and Wilkins, Baltimore, pp. 1900–9
Ursano R J, McCaughey B G, Fullerton C S (eds.) 1994 *Individual and Community Responses to Trauma and Disaster: The Structure of Human Chaos*. Cambridge University Press, New York
Weisaeth L 1994 Psychological and psychiatric aspects of technological disasters. In: Ursano R J, McCaughey B G, Fullerton C S (eds.) *Individual and Community Responses to Trauma and Disaster*. Cambridge University Press, New York, pp. 72–102

D. M. Benedek and R. J. Ursano

Military and Politics

Virtually all nations have some form of military force for protection against external foes, for international prestige, and often to maintain internal order. The relationship between a nation's political life and its military is a fundamental and enduring problem which may be understood as a matter of managing the boundary between them. Civil authorities desire to control the military; but, militaries are more effective when they are professionalized, which requires substantial autonomy and minimal civilian penetration into their internal operations (Wilensky 1964).

1. *Civilian Control of the Military*

The scale of the problem of relations between military and politics differs between modern democracies and less well developed and differentiated societies (see *Modernization, Political: Development of the Concept*). In stable democratic regimes, widely accepted political norms and formal institutional mechanisms serve to maintain the boundary (see *Political Culture*). Historically, standing militaries have been viewed as contributing to tyranny because of the expense of their maintenance. In modern times they gain political influence through symbiotic relationships with the private enterprises that produce their weapons systems—the 'military–industrial complex' that former United States President Eisenhower warned against in 1961.

Within democracies, institutions of civilian control include constitutional, legal, and administrative mechanisms such as military budgets and establishments, civilian confirmation of officer commissions, appointment of top military officials by civilian authorities, and prohibitions on military employment for domestic problems. Even powerful and popular military officers who exceed existing boundaries may be removed from their positions by their civilian superiors, as when, in April 1951, US President Harry Truman summarily relieved General Douglas MacArthur.

Professional, full-time military, consists of members who devote all of their time to their duties, minimizing conflicts of interest. In some regimes, civilian authorities worry about militaries with a capacity to compete with their authority. In both communist and fascist regimes, specialized political officers have been employed within military units with lines of authority parallel to military commanders as a means of ensuring the latter's compliance with regime dictates.

However, such institutions are not effective absent an underlying foundation of well-developed and widely accepted norms in the broader political culture, which may take centuries to develop (Landau 1971). Political norms include general acceptance of military subordination to civilian authorities and specific prohibitions on serving officers engaging in political activities such as legislative lobbying or standing for elected or appointed office. These norms constitute essentially a social contract about the roles and functions of civil and military authorities, respectively. The exact character of this contract tends to be renegotiated over time.

In democratic regimes, such as Britain, a pattern of norms developed over centuries in which both civilian and military bureaucracies were subordinated to the control of Parliament. In the United States, representative political institutions were constitutionally established before any other, with the result that control of the military by civilian authority has never been at issue, nor has the legitimacy of representative institutions relative to the military (see *Public Bureaucracies*). In developing nations, representative political institutions may still have to compete with the military for legitimacy (Stepan 1971).

Development of separate and effective institutions for domestic law enforcement and state militias, combined with firmly established political norms allowing employment of militias internally only in extreme circumstances of natural disaster or civil unrest, have reduced pressures to use military forces internally.

2. Military Participation in Politics

In developing nations, the military, because it tends to be socially conservative, is typically the best organized of institutions, and controls lethal force, has been both motivated and able to act as an independent political factor. It has superseded civil authority by *coup d'etat*, influenced public policy by threat of intervention, been used as an instrument of terror. It has made successful claims for special privileges, such as special stores and facilities, achieved large budget allocations, or achieved direct military control of economic enterprises as in former communist regimes.

When a nation's military is dominated by particular ethnic or racial groups, or by geographic regions, the probability increases that it will be used in internal political conflicts, or for repression of certain ethnic groups, as events during 1999 in Indonesia concerning East Timor have shown. Only occasionally has the military acted to facilitate the establishment or restoration of democracy in such nations (see *Democratic Transitions*).

Following World War II, because the perceived role of the German and Japanese militaries in causing the war, and of their role in supporting authoritarian regimes, the Allies deemed it of utmost importance to demilitarize both nations, including constitutional proscriptions against the use of offensive military force. Efforts by western democratic states to professionalize the militaries of developing nations in Latin America, Africa, and Asia have been only partially successful given the military's role in internal political control (O'Donnell and Schmitter 1986). It remains to be seen whether analogous western efforts to assist former Soviet-bloc militaries to adjust to existence in democratic states will prove more fruitful (Linz 1996).

3. Autonomy for the Military

The military also benefits from well-defined boundaries between it and civilian politics. When militaries are relatively insulated from civilian intrusion into promotion and assignment to duty they have proven more effective (Chisholm 2000). To the extent that militaries are perceived by their civil societies to be professional organizations whose members are engaged in service to their nations and are not principally mechanisms for patronage, nepotism, and informal welfare, militaries are accorded a relatively higher status (Janowitz 1960).

The use of force is usually considered at three levels of analysis: strategic, operational, and tactical. Civilian leadership typically takes responsibility for the strategic level of decision, advised by the military. The mere presence of military capability may indirectly influence decisions by civilian leaders about strategy. In 1999, for example, NATO found it politically feasible to intervene militarily in Kosovo because the sophistication of its air forces promised military effectiveness with low risk of casualties. Disputes between civilian leaders and military officers most often occur at the operational level of decision. The military is usually primarily responsible for the tactical level. In recent years, with dramatic improvements in communications capacities, civilian leaders have become increasingly involved in operational level decisions, and even in tactical decisions, such as the ill-fated US effort in 1976 to rescue the crew of the merchant ship *Mayaguez* from their Cambodian captors.

Where civilian leaders have heeded their professional military officers on questions concerning the use of force, ill-advised adventures and misuse of the military have been less likely. As the technology of warfare has become more complex, the need for expert military advice to civilian decision makers has become more acute. Civilian leaders have not always shown themselves capable of seeking advice, or listening to it, or understanding it. Ironically, senior military officials are frequently less prone to use force than their civilian counterparts. Changes in warfare have also effectively doomed the hastily thrown together citizen army as an effective means of national defense and created pressure for standing militaries.

4. Interpenetration of Military and Politics

Established boundaries do not mean impermeable barriers, however. Historically, there has been concern that a military set completely apart from civil society might prove dangerous to the latter. In the late nineteenth century the United States Navy relied predominately on citizens from other countries to staff its ships, and sailors were considered social pariahs to be excluded from polite society, while officers were drawn disproportionately from higher social strata. Both factors contributed to separation of that service from American civil life (Karsten 1972).

Universal conscription has lessened such separation, as have reserve officer training programs in civilian universities, both of which create a regular flow of individuals in and out of the military. This at once increases civilian influence on the professional military, enhances civilian understanding of the military, and provides mechanisms by which militaries can expand and contract in response to external threats. Ironically, it also increases the potential political cost of military actions, as the United States found with Vietnam, and Russia discovered in its Chechen endeavors during the 1990s, in which parents of conscripts pressured the government to end the action.

Development of smaller, all-volunteer, career militaries runs counter to this historical trend, and to the extent that in the generations following World War II fewer top civilian leaders in western nations have performed military service, trust between civilian and

military leaders appears to have diminished (Ricks 1997). Moreover, it may mean reduced understanding by civilian leaders of the appropriate uses of and limitations on military force as an instrument of national policy, especially as nations expand the use of their militaries for limited armed conflicts and for operations other than war. Employment in the 1980s of the United States Marines in Lebanon at the behest of the State Department and against the opposition of senior military officials exemplifies the misuse of military forces at the operational level. Pressured to act by domestic constituencies during various international crises, civilian leaders appear to turn to the use of military force, even if they cannot achieve the ends sought.

As a relatively closed institution and with socially conservative inclinations, the military has often had difficulty adjusting to changes in civil society. However, changes of values in the broader society do penetrate the military, evidenced by improved treatment of enlisted personnel. Desegregation and the integration of women and gays into the military have proceeded at a faster pace than in civilian institutions, perhaps because of its hierarchical, command organization. Changes have been instituted by the command of civilian authorities, an approach impossible in most sectors of civil society.

Militaries also have acted as interest groups with their own political agendas. In stable democracies, they must compete with other institutions for legitimacy and for scarce resources. The influence of militaries waxes and wanes with the centrality of external threats to their societies, the perceived ability of the military to contend with threats, and in some cases, the degree of internal political unrest. A military's failure, as the Argentine military found in its war with Britain over the Falkland Islands in 1982, may have profound consequences for its domestic standing, even where it has previously functioned to suppress political dissent.

Militaries have historically also served as a mechanism for social mobility for less advantaged groups, particularly during economic recession and following wars when veterans have received pensions, tax relief, health care, and education and real estate subsidies. For example, the US GI Bill following World War II provided college education for a broad sector of society that was previously excluded, contributing to the sustained growth of the postwar American economy. Military training in technical specialities such as aviation or electronics also provides skilled personnel at subsidy to private economies. The importance of this has increased as technical skills in demand by military and civilian sectors have converged. Convergence makes it more difficult for militaries to retain skilled personnel, especially during periods of strong economic growth.

Political dictators have been educated and trained in the military, but in democratic regimes this has not been a problem because of their inherent flexibility and because agents of change come from a range of institutions.

See also: Democratic Transitions; Military Psychology: United States; Military Sociology; Modernization, Political: Development of the Concept; Political Culture; Public Bureaucracies; War: Anthropological Aspects; War: Causes and Patterns

Bibliography

Chisholm D 2000 *Waiting for Dead Men's Shoes: Origins and Development of the U.S. Navy's Officer Personnel System. 1793–1944*. Stanford University Press, Stanford, CA

Janowitz M 1960 *The Professional Soldier*. Free Press, Glencoe, IL

Karsten P 1972 *The Naval Aristocracy: The Golden Age of Annapolis and the Emergence of Modern American Navalism*. Free Press, New York

Landau M 1971 Linkage, coding, and intermediacy. *Journal of Comparative Administration* **2**: 401–29

Linz J J 1996 *Problems of Democratic Transition and Consolidation: Southern Europe, South America, and Post-Communist Europe*. Johns Hopkins University Press, Baltimore, MD

O'Donnell G, Schmitter P C (eds.) 1986 *Transitions from Authoritarian Rule: Tentative Conclusions About Uncertain Democracies*. Johns Hopkins University Press, Baltimore, MD

Ricks T E 1997 *Making the Corps*. Scribner, New York

Stepan A 1971 *The Military in Politics: Changing Patterns in Brazil*. Princeton University Press, Princeton, NJ

Wilensky H 1964 The professionalization of everyone? *American Journal of Sociology* **87**: 548–77

D. Chisholm

Military Geography

Military geography is the subfield of geography that deals with the impact of geography on military affairs. Its purpose is to provide understanding and appreciation of the significance of geographic concepts and realities to military plans and operations at the tactical and operational levels of war, and to military concerns at the strategic level. Thus in using geographic methods and information, military geography 'concentrates on the influence of physical and cultural environments over political-military policies, plans, programs, and combat/support operations of all types in local, regional, and global contexts' (Collins 1998).

While military geography is accepted by recognized scholars in the discipline as a viable subdivision of geography, its growth and contribution to scholarship has been marked by periods of intense activity and

corresponding periods of benign neglect. Quite logically, the most vigorous activity has been during wartime or shortly thereafter, although some of the best studies have been only recently published. Recognizing the historic distrust between nations and peoples, and with it the possibility or reality of military action or intervention, the continued geographic study of military issues and conflict resolution would seem to have a place in scholarly research (see J. V. Coniglio, *Military geography: Legacy of the past and new directions* in Garver 1981). This article presents the nature and scope of military geography, reviews its historical development and considerable literature, comments on global strategic views, and suggests some directions for future research and study in the field.

1. Nature and Scope

1.1 Military Operating Environment

At a meeting of the Royal Geographical Society in London in 1948, Field Marshall Lord B. Montgomery concluded his remarks about strategy and the making of war in saying: 'I feel the making of war resolves itself into very simple issues and the simplest in my view is what is possible and what is not possible? …What is possible will depend firstly on geography, secondly on transportation in its widest sense, and thirdly on administration. Really very simple issues, but geography I think comes first.' The problems of war are rooted in geography. War is fought to gain control of areas and peoples of the world. The conduct of war and battles is conditioned by the character of the area of operations—the military operating environment. Some military operating environments contain critical objectives and some contain essential lines of communication; some favor the attacker while others favor the defender; some pose severe environmental constraints while still others allow for extensive maneuver and the employment of large, mechanized formations. The outcome of battles and wars rests in a large part on how well military leaders at all levels seize upon tactical and strategic opportunities for success provided by the military operating environment. On a world regional level, the different military operating environments are classified as follows: (a) temperate forest and grassland; (b) humid tropical; (c) desert; (d) mountain; (e) cold weather or winter; and (f) urban areas. Clearly force structure and size and weapons/equipment requirements will vary depending on the special environment they are to be deployed in as will the optimum tactical doctrine vary under differing environmental conditions.

1.2 Environmental Matrix

Understanding the significant elements of the environmental matrix is critical to military commanders and staff officers in planning and executing operations at all levels. Analysis of the environmental matrix, defined as 'the sum of all of the factors and forces which operate at a place and which can have an effect upon the performance of any function there' (Peltier and Pearcy 1966), shows that its nature results from the coexistence and interrelation of a host of different elements. Some of these elements are physical, others are cultural. In addition to location, or 'place,' the physical elements include landforms, hydrology, weather and climate, surface materials, vegetation, and minerals. In their manifold combinations, these features comprise the varying physical regions, which occur over the earth in differing regional military operating environments. The cultural or human elements of the landscape consist of population characteristics, settlement and land use, economies, transportation networks, and cultural groups, institutions, and capabilities. In summary, it is the integration of the interacting physical and human resources of the environmental matrix upon which the economic, political, and military power base of a country, or region of any size, is derived. Every area is constantly in flux. Natural processes modify it, while humans change both the environment and their own use of it. Analysis of the implications of changes in the environmental matrix of a country or region is also a continuing process.

1.3 Levels

As with all geography, the value of military geography lies in its unique spatial perspective and methodology. The mere listing of information and data concerning an area of operations does not in itself constitute military geography. The contribution of military geography is in the analysis and identification of significant elements of the environmental matrix in the military area of potential or actual operations, which assists the commander and staff in preparing estimates, and in formulating and executing military plans. Military geography provides a coherent and selective mission-oriented assessment of the environmental matrix in military operating environments at the tactical, operational, and strategic levels of war.

The tactical level of war deals with the geography of the battlefield and involves the geographical factors of the battle itself. It includes small unit plans and operations and the concerns of direct combat at the division level and below. Primary interest is in enemy resources and maneuverability in a relatively *small area* of operations. The identification and use of 'COCOA' in area analysis is essential: Critical Terrain, Obstacles, Cover and Concealment, Observation and Fields of Fire, and Avenues of Approach. Thus tactical planning is 'military planning' in the narrowest sense of the word; it deals with how to fight battles, that is, *the weather and terrain, the movement of troops*, employment of weapons and other resources on the battlefield.

The operational level of war relates to large unit (corps and army level) plans and operations and is concerned with logistics, administration, and support of smaller battle units as well as combat over a *large area* of operations. Knowledge of general enemy resources and military capabilities and consideration of the climate and physical nature of the area of operations is critical. Operational planning is 'military planning' in a broader sense of the word, but it is still military and is involved with the movement of troops and resources to and within the larger area (theater) of operations and with arrangements to allow battles and campaigns to be fought on favorable terms. At the operational level, geography narrows down then to the geography of campaign plans and operational plans.

The strategic level of war (or geostrategic) deals with plans and operations on the national and global levels and with concern for political, economic, social, and psychological factors as well as environmental and military matters. It is interested in total enemy resources and culture, the entire earth as the potential area of operations. Geostrategy is also called national strategy, grand strategy, global strategy, or international strategy. It has historically involved itself in geopolitics and the presentation of global strategic views of the world. Strategic planning is the most general of military planning and includes factors that affect national policy and strategy and thus has a strong affiliation with political geography. The military factor is one of many elements that enter into the formulation of strategic estimates and plans at this level. Strategic geography is involved with the planned use of military force to achieve national objectives and with the questions of when, where, and how wars are to be fought.

2. *History*

Extant writings on the relationship of geography to military planning and operations go back some 2500 years to about 500 BC, when S. Tzu wrote his treatise on 'the Art of War.' He placed terrain and weather among the top five matters to be deliberated before going to war, noting that 'A general ignorant of the conditions of mountains and forests, hazardous defiles, marshes and swamps, cannot conduct the march of an army' (Sun Tzu 1963). Xenophon's account of the march of the 10000 across Asia Minor, from 401–399 BC, contains many examples of a lack of understanding of military geography. Varus, a Roman General, in AD 9, cost Emperor Augustus all the territory between the Rhine and the Elbe. Having conducted earlier campaigns in Mediterranean dry regions, Varus had no experience in the cold, wet, swampy, mixed-deciduous forest of middle Europe. While his cavalry and wagons were immobilized in mud and water, the Germans destroyed his army at the Battle of the Teutoberg Forest (Jackman 1971). Eight centuries later, in 1747, Frederick the Great's instructions for his Generals reflected his understanding of the importance of military geography when he wrote: 'Knowledge of the country is to a general what a rifle is to an infantryman and what rules of arithmetic are to a geometrician. If he does not know the country, he will do nothing but make gross mistakes. Without this knowledge, his projects, be they otherwise admirable, become ridiculous and often impracticable.' Had Napoleon's cavalry been aware of the sunken road at Waterloo in 1815, perhaps a different battle outcome may have resulted (O'Sullivan and Miller 1983).

The development of military geography as a separate field of study is clearly demonstrated by reviewing the *Bibliography of Military Geography*, volumes 1–4 (1,059 pp.), published by the United States Military Academy. Entries range from the highly speculative consideration of global strategic theories at the geostrategic level to combat operations at the platoon tactical level in varying terrain and weather situations. Military geography's contemporary beginnings in mid-nineteenth century Europe were characterized by regional studies dealing with the natural and man-made features of several theaters of war (see, for example, T. Lavallee's *Géographie Physique, Historique et Militaire* 1836 and A. von Roon's *Militärische Länderbeschreibung von Europa* 1837). The Lavallee and Roon books mark the first attempts to consciously relate geography and the art of war. As the concept of total war for national purposes matured in the last half of the nineteenth century, many more military geographies were published. Military geographies of Germany, Switzerland, and the Rhine and Danube are but a few examples (Ruhierre 1875, Bollinger 1884 and Maguire 1891). At the tactical level, D. W. Johnson's monograph on the military geography/geomorphology of the battlefields of World War I presented an informative study of the importance of terrain in the tactical advances and retreats in trench warfare (Johnson 1921). Johnson's *Battlefields of the World War* has been called 'the one outstanding monograph by an American scholar dealing with military geography.' Little was written on military geography in the period between World Wars I and II.

3. *Global Strategic Views*

The early nineteenth century German military strategist C. von Clausewitz maintained in his writings '*On War*' that land-power was the ultimate political force. However, he put war in its proper place, not as an end in itself, but as a means to an end—the advancement of national power. His definitions of strategy and tactics enlarged the scope of military geography to include the concepts of geostrategy and geopolitics which in turn supported the global power determinism

of European nations in the nineteenth and early twentieth centuries (see *Jomini, Clausewitz, and Schlieffen*, Department of Military Art and Engineering, United States Military Academy, West Point, NY 1951). Global strategic views were formulated in which the existence of political power and influence among nations was explained as a function of geographical configurations, especially the global layout of continents, oceans, and connecting seas, and the nations that controlled these areas. Thus US Navy Admiral A. T. Mahan's thesis in 1900 that seapower, not landpower, was the key to world domination gained the support of several major world nations (Mahan 1890). This was soon followed by the British Oxford Geographer, Sir H. Mackinder's heartland theory on the global dominance of Eurasia as the resource base for world landpower (Mackinder 1904). Professor N. Spykman of Yale University countered with the proposal that the 'Rimland,' with its marginal seas, contained greater world power potential than the 'Heartland.' (Spykman 1944). The post-World War II containment policy of the United States and NATO nations was conceptually anchored on Spykman's Rimland thesis (see Cohen 1973, Norris and Haring 1980). Discussion of these global strategic views in the era spanning both world wars brought military geography scholars into the study of political geography and geopolitics (see *Political Geography* and *Geopolitics*).

4. Modern Literature

World War II found most military geographers involved in preparing area intelligence reports bearing *on military, economic, or transportation* problems in the various theaters of war as basic geographic knowledge of these regions was seriously deficient. The finest examples of wartime area reports were the Joint Army and Navy Intelligence Studies (JANIS). While these volumes were prepared by specialists from a number of disciplines, the director of each area research team was a military geographer whose academic training had been in the field of geography. Thus the standard area intelligence report was organized topically in the framework of regional geography. As a result of the increased needs for geographic knowledge in World War II, military geography soon came into its own as a subfield of geography shortly after the war when the professional geographers returned to college and university departments of geography. It was first identified as such in an essay titled 'Military Geography' in the book, *American Geography: Inventory and Prospect* prepared and published by distinguished geographers under the sponsorship of the Association of American Geographers (Russell 1954).

Following World War II, interest and research in military geography languished while military concerns moved to new strategies to confront the issues of nuclear deterrence and Cold War policies to prevent military conflict between superpower nations. Nevertheless, military journals published numerous articles concerning the effect of terrain and weather on military operations in different military operating environments. The history, philosophy, and theory of modern military geography were presented in academic papers by military officers attending civilian universities (Thompson 1962, Brinkerhoff 1963). In the early 1960s, the US Army's equivalent institution for professional graduate training, the Command and General Staff College, introduced a required course titled 'Military Geography' consisting of a series of readings and area studies. A book titled 'Military Geography' was commercially published in 1966 and immediately became the primer for both scholars and military officers interested in the subject (Peltier and Pearcy 1966).

A bibliographic survey will demonstrate that while strategic studies came into their own as an interdisciplinary endeavor, work accomplished in the pure name of military geography began to decline sharply in quantity after the mid-1960s (see *Bibliography of Military Geography*). The reason for this decline is a lack of interest by academic geographers in conducting research in military geography. Nevertheless, revitalization occurred, beginning in the late 1970s. A course titled 'Military Geography' was introduced in 1978 at the United States Military Academy and was supported by a new publication, *Readings in Military Geography* (Garver 1981). In 1983 P. O'Sullivan and J. Miller published a monograph, *The Geography of Warfare*, which reviewed the interaction between geography and the history of selected military campaigns involving do's and don'ts of tactics and strategy. This was followed by *Terrain and Tactics* (O'Sullivan 1991), resulting in a more focused academic treatment of the interplay between differing terrain settings and tactical options. In 1996, with the primary sponsorship of the Geography Department faculty at the United States Military Academy, the Association of American Geographers approved the establishment of the Military Geography Specialty Group within the Association's listing of topical specialty groups.

A recent major contribution to the field is J. Collin's exceptional book, *Military Geography: For Professionals and the Public*, published in 1998. Clearly the most comprehensive treatment of military geography in print, the subject is presented in a traditional geographic format with the following major sequential topics: (a) physical geography; (b) cultural geography (c) political-military geography; and (d) area analyses. Each of the 19 chapters terminates with a list of key points for easy reference. Stated aims of the author are to offer a college-level textbook, provide a handbook for political-military professionals, and to enhance public appreciation for the impact of geography on military affairs.

A second new publication, which merits special attention, is H. A. Winters et al., *Battling the Elements: Weather and Terrain in the Conduct of War* 1998. The authors examine the connections between major battles in world history and their geographic components, revealing what role weather, climate, terrain, soil, and vegetation have played in combat. Each of the 12 chapters offers a detailed and informative explanation of a specific environmental factor and then looks at several battles that highlight its effects on military operations. Among the many battles examined are the American Revolution's Bunker Hill, the Civil War's Gettysburg and Wilderness campaigns, World War I's Verdun and Flander's Fields, World War II's beaches at Normandy and Iwo Jima, and the Rhine crossing at Remagen, Vietnam's battles of Dien Bien Phu and the Ia Drang Valley, and Napoleon and Hitler in Russia. As this thoughtful analysis makes clear, those leaders who know more about the physical nature of battlefield conditions will have a significant advantage over opposing leaders who do not.

5. *Future Directions for Research and Study*

With the recent and continuing great advances in communications, surveillance and intelligence-gathering technology, computer programming capabilities, and weapons sophistication, applications of remote sensing, geographical information systems (GIS), battlefield simulation, and war gaming techniques can be made by military geographers toward better understanding of the complex relationship between geography and military matters. Perhaps the single most important lesson to be gained from this essay on military geography is the danger of neglecting or misunderstanding geographic concepts and realities when planning and executing military operations at any level. Clearly, General Eisenhower recognized the value of knowledge of military geography in the conduct of war when on April 22, 1959 he wrote in his frontispiece to Volume I of the *West Point Atlas of American Wars* that 'The "Principles of War" are not, in the final analysis, limited to any one type of warfare or even limited exclusively to war itself… but principles as such can rarely be studied in a vacuum; military operations are drastically affected by many considerations, one of the most important of which is the geography of the region.'

See also: Military History

Bibliography

Bibliography of Military Geography, Vols. 1–4 n.d. Department of Geography and Computer Science, United States Military Academy, West Point, NY

Bollinger H 1884 *Militär-Geographie der Schweiz*, 2nd edn. Orell Fussli and Company, Zurich, Switzerland.

Brinkerhoff J R 1963 M.A. thesis, Columbia University

Cohen S B 1973 *Geography and Politics in a World Divided*, 2nd edn. Oxford University Press, NY

Collins J M 1998 *Military Geography for Professionals and the Public*. National Defense University Press, Washington, DC

Garver J B (ed.) 1981 *Readings in Military Geography*. Department of Geography and Computer Science, United States Military Academy, West Point, NY

Jackman A 1971 Military geography. In: Leestma R A (ed.) *Research Institute Lectures on Geography*. U.S. Army Engineer Topographic Laboratories Special Report ETL-SR-71-1, Fort Belvoir, VA

Johnson D W 1921 *Battlefields of the World War, Western and Southern Fronts, A Study in Military Geography*. Oxford University Press, NY

Mackinder H J 1904 The Geographical Pivot of History. *Geographical Journal* **23**: 421–37

Maguire T M 1891 *Strategic Geography: The Theatres of War of the Rhine and Danube*. Edward Stanford, London.

Mahan A T 1890 *The Influence of Seapower on History, 1660–1783*. Little, Brown, Boston

Norris R E, Haring L L 1980 *Political Geography*. Charles E Merrill, Columbus, OH

O'Sullivan P, Miller J W Jr 1983 *The Geography of Warfare*. St. Martin's Press, NY

O'Sullivan P 1991 *Terrain and Tactics*. Greenwood Press, NY

Peltier L C, Pearcy G E 1966 *Military Geography*. D Van Nostrand, Princeton, NJ

Ruhierre H 1875 *Géographie Militaire de L' Empire D'Allemagne*, Sandoz et Fishbacher, Paris.

Russell J A 1954 Military geography. In: James P E, Jones C F (eds.) *American Geography: Inventory & Prospect*. Syracuse University Press, Syracuse, NY

Spykman N J 1944 *The Geography of the Peace*. Harcourt, Brace and World, New York

Sun Tzu 1963 *Art of War*. Translated with introduction by Griffith S H, Clarendon Press, Oxford, UK

Thompson E R 1962 Ph.D. thesis, Syracuse University, NY

Winters H A, Galloway G E Jr, Reynolds W J, Rhyne D W 1998 *Battling the Elements: Weather and Terrain in the Conduct of War*. Johns Hopkins University Press, Baltimore, MD

J. B. Garver

Military History

The easy definition of military history is that it is the history of wars. And yet this is too imprecise. Wars have social, economic, and political dimensions which have been analyzed more by the historians of those subdisciplines than by military historians. That is not to say that there are not important links to be made between military history and other historical subdisciplines, nor is it to deny that the good military historian endeavors to make those connections. But in terms of their subject matter military historians have been concerned primarily with the histories of armed forces, not only in war but also in peace. Military history has therefore been more comfortable with wars fought by armies and navies than with wars fought between warrior societies, or before soldiering became a distinct profession.

1. The Emergence of Military History as a Separate Subdiscipline

Thucydides wrote the history of a war, and—like some other ancient historians—himself saw service. But it would not be reasonable to call him a military historian. It required the growth of professional armies and the concomitant influence of the Enlightenment for the ancient historians who wrote about conflict to be treated as military historians rather than as historians *tout court*.

1.1 The Influence of Professional Armies

In the 1590s, Maurice of Nassau developed systems of infantry drill and of military organization which standardized tactics and which were emulated throughout Europe. This was the basis for what in 1956 Michael Roberts called the 'military revolution'—a clutch of changes that occurred in warfare between 1560 and 1660. The Thirty Years' War in Central Europe, the Fronde in France, and the British civil wars served to disseminate the principles of the 'military revolution,' which found their seventeenth-century apogee in the Swedish army of Gustavus Adolphus. Subsequent historians have elongated the chronology of the 'military revolution.' Geoffrey Parker sees its origins as before 1560, in the growth of a new style of artillery-resistant fortification, the *trace italienne*. Others have highlighted developments after 1660, calling the growth in army size under Louis XIV and the maintenance of armies in peace as well as in war a 'second military revolution' (Rogers 1995). The essential points are that over the course of the period 1500 to 1800, armies as we now recognize them evolved, and that those armies then became the basis of imitation within Europe and the foundation of empire outside it.

During the course of the eighteenth century the aristocracies of Europe, now subordinated to the crown and state, made soldiering their vocation. Military academies were established, not only for the training of the scientific arms—the artillery and engineers—but also for the cavalry and infantry. The combination of institutional continuity and professionalization promoted the study of war. The writings of the ancients were supplemented by the effusions of eighteenth-century practitioners—including Maurice de Saxe, Henry Lloyd, and G. F. von Tempelhoff—who combined theoretical musings with a distillation of their own experiences. Military history had didactic rather than scholarly origins.

1.2 The Enlightenment

The study of past wars as a means of improving present practice was, at one level, entirely consistent with the precepts of the Enlightenment. War, like any other human activity, could be seen as a science, based on unchanging principles, themselves derived from actual experience. The wars of Frederick the Great, a child of the Enlightenment and a major writer on war, served to promote these connections. Over 70 works of military theory were published in the seventeenth century, but more than twice that in the eighteenth, and over 100 in the years 1756 to 1789 (Gat 1989).

Napoleon was both a product of this tradition and its most distinguished advocate. William Napier, an unstinting admirer of the emperor, established British military history with his account of the Peninsular War. Although Napier eschewed theory, his history was influenced by the most important military theorist of the nineteenth century, A. H. Jomini. Jomini's *Traité des grandes opérations militaires* (first two volumes 1804) became a five-volume history of the wars of Frederick the Great and Napoleon, with a resumé of the general principles of war tacked on as a conclusion. His *Précis de l'art de la guerre* (1838), which shaped the syllabuses of military academies into the twentieth century, put the theory first but still relied on history to make its points (Shy 1986, Alger 1982).

Carl von Clausewitz was critical of Jomini, but like him wrote far more history than theory. Moreover, his principal work, the unfinished and posthumously published *Vom Kriege* (1832), relies on history for its evidential base. Clausewitz stressed his anxiety to break with the nostrums of earlier military writers, but he is no different from them in his readiness to cull military history for normative purposes.

At one level, therefore, the influence of the Enlightenment on the development of military history operated internally—it helped define how and by whom war was fought. But its consequences were also external to the subject. The phenomenon of war itself appalled the *philosophes*—not least Frederick's friend, Voltaire. Their efforts to curb and moderate its effects, stoked by the memory of the Thirty Years' War and its terrors for the civilian population, took shape through international law. War became an activity clearly distinguished from peace, undertaken by specialists separated from civilian society but who, crucially, acted not on their own account but on behalf of the state. For the *philosophes*, war was not necessarily endemic in society. Those who built on the legacy of the Enlightenment, the liberals of Victorian Britain in particular, could see it as an activity that was not honorable but reprehensible, maintained by the aristocracy with the aim of sustaining their own hold on society, and replaceable by alternative forms of interstate competition, particularly trade.

2. Academic Neglect of Military History

The legacy of liberalism was a belief that military history was not a proper study for universities. As history gained a foothold in the curriculums of

European higher education in the last third of the nineteenth century, war ought to have been central to its preoccupations. Indeed, at one level it was. Historians of the ancient or medieval worlds could not neglect war entirely, but they still preferred to focus on more humane developments, on 'progress' in law, religion, or the machinery of state. Equally, no nineteenth-century scholar could neglect the impact on Europe of the Napoleonic Wars or the wars of German Unification—as Archibald Allison showed for the former, and Heinrich Friedjung for the latter. But they neglected the conduct of war itself: war was an aberration, inimical to the 'Whig' view of history. The advent of nuclear weapons in 1945, with their threat of a war so awful and so complete that war and peace became absolute terms rather than points on a scale of relative values, completed this marginalization. Diplomatic historians looked at why wars occurred, at how they were ended, and at efforts to avoid them thereafter, but the history of war itself was left to its practitioners.

This picture of academic neglect, still propagated but with much less reason in the twenty-first century, is overdrawn. There were major exceptions, even if it remains true that those exceptions prove the rule.

The founding father of academic military history, Hans Delbrück, served in the Franco-Prussian War and believed that universities should recognize military history. He planned to do his *Habilitation* on a military historical topic, but both Heinrich von Treitschke and Theodor von Mommsen, the leading historians in Berlin, opposed the idea. Delbrück persisted with his enthusiasm and in 1883 announced his intention to write a general history of war in its political context. Leopold von Ranke told him that the study of war was not appropriate to a university, and when Delbrück was appointed to a chair in 1895 (ironically as Treitschke's successor) it was in 'universal and world history.' In 1900, Delbrück published the first of the four volumes of his history of war. Mommsen told the author that he would not have time to read it (Bucholz 1985, Deist 1998).

Delbrück also alienated those to whom his work ought to have appealed. He argued that Frederick the Great had not sought decisive battle but had preferred to exhaust his enemies by maneuver. The German army's general staff was incensed. Its own studies of Frederick's campaigns were extensive, but their purposes were more didactic than scholarly. Institutionally, this loyalty to Frederick (rather than to Napoleon) as the founding father of modern war predisposed them to see him as the advocate of a strategy of 'annihilation' rather than of 'attrition' (Bucholz 1985, Lange 1995).

At one level, therefore, Delbrück was an isolated figure—virtually unique in the academic world of Wilhelmine Germany, and vilified by the one institution that thought deeply about the history of war. But he was not uninfluential. He had been tutor to the Kaiser, he was a Reichstag deputy and he edited the *Preussische Jahrbücher*. He emerged from World War I a leading figure in Germany's public and intellectual life. As a role model his legacy spread in two directions. First, his interpretation of military history was essentially Clausewitzian; he studied war as a state activity, and as an agent for the implementation of strategy. Second, he treated war as a discrete phenomenon, possessing an integrity from the ancient world to the modern. Historical knowledge so defined was the basis for pronouncements on current strategic issues.

Delbrück's institutional legacy proved more short-lived. Military history was established at the Friedrich Wilhelm University in Berlin after World War I. But the subject was usurped by the Nazis. For a long time after World War II there was no established chair of military history at a German university, with the exception of the post created for the Clausewitz scholar, Werner Hahlweg, at Münster in 1969. Only in the 1990s did the subject reassert itself in research terms, and a professorship was established at Potsdam.

In Britain, too, individual careers punctured the image of academic neglect. The Regius Professor of Modern History at Oxford between 1904 and 1925, Sir Charles Firth, was a historian of the English Civil War and of Cromwell's army. The same university's Chichele Professor of Modern History from 1905, Sir Charles Oman, wrote a massive and definitive history of the Peninsular War. In the same year Oxford created a lectureship in military history, strategy and tactics, and in 1909 established the Chichele chair in the history of war. Thus in Britain, unlike Germany, the subject acquired an institutional focus. Two years later, Cambridge followed suit with the Vere Harmsworth chair of naval history. King's College London, whose professor of modern history, J. K. Laughton, had been involved in the formation of the Naval Records Society in 1893, formed a department of naval history in 1913 but failed to make an appointment (N. A. M. Rodger in Hattendorf 1994).

In Germany, the effect of World War I was to institutionalize personal initiative; in Britain, its effect was to undermine the progress already made. Spenser Wilkinson, who had been appointed to the Oxford chair, had been a journalist, and—like Delbrück—had linked the study of military history to current policy issues. But, unlike Delbrück, Wilkinson did not join the public debate on the strategy of World War I. He turned to scholarly studies of the eighteenth century and, his eyesight failing, became a marginal figure. Until the appointment of Sir Michael Howard to the chair in 1977, his successors lacked academic punch (Ernest Swinton), were tarred with the brush of journalism (Cyril Falls), or were unproductive in publication terms (Norman Gibbs). The burden of actual teaching throughout this period fell on C. T. Alkinson, whose output was considerable, but who tended to antiquarianism as well as personal eccentricity. In Cambridge, the Harmsworth chair married

naval and imperial history in 1932, to the eventual detriment of the former.

3. General Staffs and Official Histories

The failure of the universities in Britain to master military history was displayed by the allocation of responsibility for the official histories of World War I. The accounts of naval and military operations were both put in the hands of civilians—Sir Julian Corbett and Sir John Fortescue, respectively. Neither man held an academic post. Corbett was a lawyer by training, but his interest in contemporary maritime doctrine, as well as in naval history, qualified him admirably for the task. Fortescue was writing a *History of the British Army* in what proved ultimately to be 13 volumes, but he derived his principal income from his appointment as Royal Librarian. His enthusiasm for colorful accounts of Napoleonic battles produced a prose style and a conceptual approach unequal to the nature of modern war. He was replaced by a team of soldiers, headed by Sir James Edmonds.

Thus the bulk of Britain's official history program ended up, by default, being run on lines not dissimilar to those of the other belligerents. In the wake of Prussia's victory in 1871, the major land powers had established or re-established their general staffs, charged with planning in peacetime and operational direction in wartime. The staffs' study of history flowed from their responsibility for doctrine and training. During the period 1871 to 1914, massive military histories of continuing scholarly importance were published, either by the historical sections of the general staffs (the Germans devoted 19 volumes to the campaigns of Frederick the Great) or by individual officers with staff training. The conceptual basis of current understanding of Napoleonic warfare was established in this period, by French officers such as Jean Colin and Hubert Camon.

These were the bodies and individuals made responsible for the official histories of World War I. The Austro-Hungarian was in the hands of the Bundesministerium für Heereswesen and the Kriegsarchiv, the French in those of the Ministry of War, and the German in those of the Marinearchiv (for naval operations) and the Reichsarchiv (for land operations). Although the latter were nominally independent of the armed forces, in reality the personalities and the agenda showed the thread of continuity.

Thus the biggest projects in military history ever undertaken were given to soldiers rather than scholars. Their justifications became didactic, and their focus tactical and operational rather than political or economic. Full accounts of campaigns were not matched by comparable analyses of economic mobilization or strategic direction.

After World War II, academics regained some of the ground they had lost, but the effects were partial. The British official history was conceived in much broader terms, and a division of labor between dons and soldiers established, the former being tasked with both the 'civil' and the 'strategic' series. In the USA the history of naval operations was entrusted to S. E. Morison, a Harvard professor. Elsewhere, however, military history narrowly defined remained largely in the hands of the successors to the general staffs' historical sections. France has *services historiques* for each of its armed forces, and there are few historians of warfare in university posts. With distinguished exceptions, such as the medievalist Philippe Contamine and the modernist Guy Pedroncini, the study of war in French academic life has been dominated by its relationship to social history—either the social composition of armies (André Corvisier has been a pioneer here) or the impact of war on society itself (where Jean-Jacques Becker has been a leader). In Italy, although in the 1960s and 1970s Giorgio Rochat, Massimo Mazzetti, and others began to challenge the domination of the official histories, the armed services' grip on the subject has remained powerful. In Germany, the Militärgeschichtliche Forschungsamt was created in 1957 under the auspices of the Bundeswehr. It has, however, recruited civilian scholars as well as uniformed members, and its publications have maintained rigorous standards that defy the self-serving purposes of the didactic tradition to which it is nominally the heir. Its monumental history of Germany in World War II shows a willingness to confront its armed forces' recent past, which contrasts strongly with the pattern in Japan. Here, the rejection of a nation's militarist inheritance has also resulted in the atrophy of military history.

4. The Didactic Tradition

Governments and armed forces sponsored the writing of official history because they hoped to derive from it 'lessons' for future application. But their labors produced accounts so voluminous and so long in production that the lessons were in danger of being lost or forgotten. Into this gap stepped others, preeminently but not only J. F. C. Fuller and Basil Liddell Hart in Britain, who unashamedly used military history as a prescriptive basis for the future. Like Delbrück, they ranged from the ancient world to the present, but unlike him, their focus was tactical rather than strategic. Even Fuller's and Liddell Hart's ostensibly historical works, whose popularity endures, retain a theoretical purpose.

5. The 'New' Military History

In the short term, therefore, the effect of the two world wars was the reappropriation of military history by the armed forces for purposes that were instructional

rather than academic. But in the long term the all-embracing nature of both wars made the marginalization of military history within university circles unsustainable. Once the first half of the twentieth century entered the curriculum, war could no longer be left out.

The idea of total war was used to underpin the idea of total history. War became central, rather than peripheral, not only to the period 1914–45 but to every other period as well. Ancient and medieval history was particularly susceptible to this approach. But its success can best be measured by consideration of the eighteenth century. The 'didactic' tradition had seen this as the era of limited warfare, when armies maneuvered rather than sought battle, and wars were indecisive affairs conducted by professional armies cut off from civil society. Current scholarship sees eighteenth-century warfare as involving a much larger cross-section of society, and is far readier to apply the vocabulary of modern war to conflicts like the Seven Years' War and the American War of Independence. Their length and geographical extent are seen by some as prefigurings of world war.

The 'new military history,' a term coined in the USA in the 1960s, indicated that the history of war was not simply about 'great captains' and their conduct of operations (Paret 1992). Military history embraced the study of armies in peacetime, and war's cultural, economic, and social effects; it had a responsibility to explore its links with other subdisciplines in history. The main directions in which the subject moved were toward 'war and society,' both narrowly and broadly defined. The narrow definition meant the analysis of an army's social composition, and also its effect on civil–military relations. The broader definition considered war's effects—economic and cultural—on civil society.

As a campaigning tool the 'new military history' proved remarkably successful, not least in the USA. In 1954, only 37 out of 493 colleges and universities in the USA offered courses in military history. By 1989 164 out of 554 institutions taught military history, and more than 2 per cent of all historians in America were military historians. This calculation excluded the officer training courses (the ROTC), which accounted for about 400 more military historians. In the mid-1970s in the USA about 100 doctoral theses a year embraced military history; 20 years later the figure had tripled (Don Higginbotham in Charters et al. 1992, Kennedy 1989).

Military historians established their subject by redefining themselves in interdisciplinary terms: they explained why they were not military historians, or at least not in the traditional sense, but 'total' historians. Although its advocates continue to promulgate the 'new military history,' its point has been made—it is now not so 'new.' So well absorbed have its points become that academic military history is increasingly returning to a narrower definition of its roots. Warfare itself is now a subject fit for self-contained study, and academics working in the field publish in journals devoted exclusively to military history rather than to general history. The argument that war is part of total history no longer has to be made.

6. *The Rediscovery of Clausewitz*

One of the reasons why Delbrück was at odds with the general staff was that he and they were not comparing like with like. Delbrück's focus was on war as a political act, as the means to fulfill strategic objectives. The general staff saw war as a self-contained phenomenon. Since 1945, Gordon Craig, Gerhard Ritter and, above all, Michael Howard and Peter Paret have operated within the Delbrückian interpretation. It is hard to underestimate the intellectual consequences in the anglophone world of Howard's and Paret's translation of Clausewitz's *Vom Kriege*, which appeared in 1976. It relieved Clausewitz of the incubus of obscurity and ambiguity, and it put the focus firmly on books 1 and 8 of *Vom Kriege*—those that deal with the relationship between war and politics: this was a primer for the vocabulary of nuclear deterrence as much as it was an insight into Napoleonic warfare.

Although Paret approached the task as a historian of ideas, Howard did so as a student of strategic studies. Through his revitalization of the war studies department of King's College London in the 1960s, Howard reforged the links between military history and current policy (Howard 1983). His own subsequent distinction in both fields, at Oxford and then at Yale, ensured that, in the UK at least, the health of military history was tied in to the study of international relations. In the USA as well as in continental Europe the relationship between military history and strategic studies has been more antagonistic. The cynic would say that, deprived of the armed services' commitment to the study of the subject, evident especially in Germany and France, and lacking the financial resources of the history departments of American universities, British military history has had no option but to embrace a marriage of convenience.

There are therefore distinct national approaches to military history, reflections in large part of the institutional diversity of its roots. Comparative military history has barely begun. Moreover, the strengths and weaknesses of each national school can reflect the fact that this form of history above all is tied in to the nation state, its formation and its self-identity. German military history since 1945 has predictably been concerned with World War II to the detriment of earlier wars; the Soviet Union largely ignored Russian military history before 1917. But both these patterns are being broken down. And Britain, which according to this sort of logic ought to be leading the world in naval history, actually let the subject crumble in the 1970s and 1980s.

7. 'Popular' Military History

The popularity of military history outside the confines of universities and staff colleges, at least within the UK and the USA, makes military history an obvious route by which to introduce the nonspecialist public to history more generally. Some of the writing, such as that of Henri Lachouque or Georges Blond in France, can come close to romantic nostalgia; the work of others, such as Paul Carell on the Wehrmacht in World War II, can seem striking for what it leaves out. But the best work in this genre—both German Werth and Alistair Horne on Verdun, for example—deserves, and has received, serious attention from scholars. In Britain, John Keegan and Antony Beevor have thrived on the back of military history's flowering in academic circles. Much of what they do is traditional. Keegan, like Liddell Hart, ranges across time and space; like Liddell Hart he too writes with the fluency and ease of a journalist; and, as with Liddell Hart, there is a didactic thrust. But the appeal is also in the narrative, as the staggering success of Beevor's *Stalingrad* (1998) displays. Keegan (1976) criticized historians of war who focused on its operational and tactical level but failed to explain the experience of combat itself. Beevor's book puts such precepts into practice. The academic world has also responded to Keegan's call. Research on war 'from below' not only has popular appeal but also interdisciplinary potential.

Military historians continue to complain that they are marginal figures in the academic world. In mainland Europe, that complaint has some substance, but becoming less year on year. And in the USA, where it is stated most vociferously, it is without serious foundation. However, its current strength derives in large part from its past battles to establish itself. Lacking a linear pedigree, military history is a hybrid whose resilience derives from the multiplicity of its approaches.

See also: Arms Control; Chemical Sciences: History and Sociology; Enlightenment; Military and Politics; Military Geography; Military Sociology; Science, Technology, and the Military; War: Causes and Patterns; War, Sociology of; Warfare in History

Bibliography

Alger J I 1982 *The Quest for Victory: The History of the Principles of War*. Greenwood, Westport, CT

Beevor A 1998 *Stalingrad*. Viking, London

Bucholz A 1985 *Hans Delbrück and the German Military Establishment*. University of Iowa Press, Iowa City, IA

Bucholz A 1997 *Delbrück's Modern Military History*. University of Nebraska Press, Lincoln, NE

Charters D A, Milner M, Wilson J B 1992 *Military History and the Military Profession*. Greenwood, Westport, CT

Clausewitz C von 1976 *On War* [trans. Howard M, Paret P]. Princeton University Press, Princeton, NJ

Deist W 1998 Hans Delbrück. Militärhistoriker und Publizist. *Militärgeschichtliche Mitteilungen* **57**: 371–83

Delbrück H 1975–85 History of the Art of War within the Framework of Political History. Greenwood, Westport, CT

Gat A 1989 *The Origins of Military Thought from the Enlightenment to Clausewitz*. Oxford University Press, Oxford

Hattendorf J B (ed.) 1994 *Ubi Sumus? The State of Naval and Maritime History*. Naval War College Press, Newport, RI

Howard M 1983 The Use and Abuse of Military History. In: Paret P (ed.) *The Causes of Wars*. Temple Smith, London

Jomini A H de 1992 *The Art of War*. Greenhill, London

Keegan J 1976 *The Face of Battle*. Jonathan Cape, London

Kennedy P 1989 The fall and rise of military history. *The Yale Journal of World Affairs* **1**(2): 12–19

Lange S 1995 *Hans Delbrück und der 'Strategiestreit'. Kreigsführung und Kreigsgeschichte in der Kontoverse 1879–1914*. Rombach, Freiburg im Breisgau, Germany

Paret P 1992 The history of war and the new military history. In: Paret P (ed.) *Understanding War: Essays on Clausewitz and the History of Military Power*. Princeton University Press, Princeton, NJ

Rogers C 1995 *The Military Revolution Debate: Readings on the Military Transformation of Early Modern Europe*. Westview, Boulder, CO

Shy J 1986 Jomini. In: Paret P (ed.) *Makers of Modern Strategy from Machiavelli to the Nuclear Age*. Oxford University Press, Oxford, UK

H. F. A. Strachan

Military Psychology: United States

Military psychology, a special discipline of work psychology, has as its primary focus the application of psychological principles and methods to the many facets of specialized work environments in innumerable military settings. Military psychologists work in government and university settings where they conduct both laboratory and field research. They also work in schools of medicine, or at military installation outpatient mental health or family counseling clinics. Psychologists provide clinical treatment to military populations, either by improving the lives of armed services personnel and their families away from home, or by providing support for those who are separated from loved ones while deployed to other countries with unfamiliar cultures and surroundings. Uniformed psychologists may work in troop units on field assignments where occasionally they deploy on dangerous military missions. Military psychologists supply guidance to military leaders and decision makers on behavioral issues of individual combatant or team performance, and on procedural matters to prevent or reduce physical and psychological casualties that accompany battlefield exigencies of war. Some psychologists serve as advisors at staff headquarters or for defense contract consultant groups. Occasionally,

they serve on governmental legislative committees with oversight of a broad range of national personnel policies impacting millions of military personnel. In new military venues, psychologists analyze humanitarian and peacekeeping missions to determine procedures for saving military and civilian lives.

Military forces helped psychology develop both as an applied profession and as a scientific discipline. From 1940 to 2000, military forces were the largest single employers and trainers of research psychologists. Following diminished threats of Soviet communist world domination (ca. 1989–90), countries downsized their military forces, and consequently decreased sponsorship, financial support, and personnel positions for military psychologists. Even so, in 2000, the three US armed services employ 300–400 clinical psychologists in uniformed service. About an equal number of research psychologists do military research. These include uniformed psychology officers and full time government civil servants in military research laboratories, and a sizable number of defense contractors, do military research.

1. *Military Psychology's Roots: World Wars I and II*

Military personnel have always been interested in the psychology and behavior of leaders and warriors in combat. Writings about such Captains of War as Alexander the Great, Caesar, and Napoleon, and by military philosophers such as Carl von Clausewitz, readily strike an appreciation of the psychological implications of leading men in combat. Military strategists assess troop readiness by gathering intelligence information regarding an adversary's vulnerabilities to gain tactical advantage in combat. However, throughout the nineteenth century there still was no 'organized body of knowledge' concerning the principles or practice of military psychology.

In the last quarter of the nineteenth century, university-based intellectual study and laboratory research on predicting human behavior formally established the scientific discipline and profession of psychology, albeit predominately an academic one. World War I wedded applications of the relatively young discipline of psychology to the military. In helping to resolve national conscript and military personnel issues, psychologists' close working relationships to the military catapulted psychology as an applied profession in the workplace. Many behavioral scientists in Europe, and hundreds of psychologists in the American Psychological Association (APA; founded in 1892), dedicated themselves to determining solutions to numerous wartime specialized military work-related issues. In 1917, Robert M. Yerkes, President of the APA and an experimental psychologist, formally organized APA psychologists into committees to apply scientific principles of psychology in aiding the US government in the war effort. Their pioneering work expanded upon the psychology known at the time, adapted it, and applied it to military support roles. Such work became identifiable as military psychology.

1.1 *Selection and Placement*

The influx of millions of conscripted men into the US Army required an economical and efficient method of classifying new soldiers and identifying potential officer candidates. Psychologists in the US Army Surgeon General's Division of Psychology developed the Army Alpha and Beta tests, as expansions of the research of the French psychometrician Alfred Binet. These first large-scale group-administered tests of intellectual ability rapidly screened and identified the intellectual level of 1.7 million conscript recruits, and classified those young men for placement into jobs, for training, and for preparation as combatants for war. The Division's tests selected 42,000 of the recruits for admission to officer training. World War (WW) I psychological tests and measurement work constituted the first formalized psychological research in military settings.

The Division of Psychology developed a system to grade individuals and grouped them according to abilities to learn; they provided lectures on training methods, and advised training officers. They measured troop morale and assimilation into the military; developed special trade tests to assess skills or combat leadership abilities, and contributed to development of methods and procedures to improve combat effectiveness and morale. In adopting mental measurement, and using psychometric screening tests for personnel selection and classification programs, the US Armed Forces made them a principal instrument of manpower management, and thereby gave credence to applied psychology within the academically based APA (Johnson 1991).

1.2 *Clinical Psychology*

Early in WW I, psychologists played an educational role in military medical settings by training hospital staff and surveying patients. In 1918, the US Army Surgeon General authorized the first duty assignments of psychologists to assist in evaluation of neuropsychiatric patients at the Walter Reed Army General Hospital in Washington, DC. This boosted the clinical practice role of psychologists within the military.

1.3 *Military Psychologists in World War II*

As the world's military forces demobilized after WW I, most psychologists returned to academia to advance

the science of psychology. As a consequence, there was a paucity of military psychology efforts until the onset of WW II. During the 1930s and early 1940s, military forces resumed interest in psychological applications for selection, classification, and assignment of military personnel. In several European countries, military establishments created and maintained behavioral science activities and research groups. In the USA, over 2,000 civilian and uniformed psychologists addressed WW II military problems, firmly establishing the role of psychology in the military. Their well-documented work pervaded published articles in APA journals in the mid- to late 1940s.

To replace the Army Alpha test, US Army psychologists developed the new Army General Classification Test (AGCT) in the early 1940s. It was administered to 12 million men during WW II. Instead of striving to eliminate bad risks, the newer psychometric screening tools sought to identify individuals who could effectively acquire certain military skills or perform specific tasks. These tests evolved to become the widely used Armed Services Vocational Aptitude Battery (ASVAB). In Western countries, WW II psychologists used psychomotor tests of coordination and physical ability for the selection of pilot candidates, and employed specialized tests for navigators and other military specialties. Psychological assessment centers were formed to develop performance-oriented tests and to select and train military operators for the British Special Operations Executive (SOE) and the US Office of Strategic Services (OSS).

1.4 Leadership

During WW II, leadership was established as a topic of military research. The build-up of military forces required identifying leaders at all levels of new command structures. Psychologists initially developed selection tests to identify individuals possessing innate characteristics and abilities desirable in leaders. However, some psychologists downplayed the innate approach, and instead insisted leadership could be trained and developed as an acquired skill. Military studies of the leadership performance of officers, and instructional innovations, gave credence to both viewpoints. These continuously have been in the forefront of military psychology since.

1.5 Human Factors and Engineering Psychology

Radar, sonar, command and control centers, high-performance aircraft, submarines, large naval surface vessels, and other new military hardware challenged the cognitive capabilities of military personnel to operate complex equipment systems effectively. In the early 1940s through the late 1950s, hundreds of experimental psychologists teamed with military system design engineers in conducting laboratory and simulation research to assess demands for increased human performance. Numerous experimental psychology studies were done assessing sensory and perceptual demands, aviator visual capabilities, visual search techniques, psychomotor skills required of equipment operators, cognitive skills of sonar and radar operators, design and location of controls and displays in aircraft and other vehicles, other man-machine interfaces, and work–rest schedules for command and control personnel. Important military research topics included studies of effects of extreme heat, cold, high altitude, and other environmental factors on military performance.

Military engineering psychologists aided weapon-system designers to apply an understanding of human capabilities and limitations in the design of equipment, materials, and jobs so as to optimize the integration of human operators into a 'total system' design within a military operational concept. Together with engineers, psychologists used the 'systems approach' to analyze complex human–machine environments in *system* terms. They used techniques such as functional, task, and time-line analyses of proposed operational procedures, information flow and decision making, and simulation in experimental trial testing (Parsons 1972). In Europe, engineering psychology was embedded in the field of ergonomics with a particular emphasis on biomechanics and physiology (Zinchenko and Munipov 1989), whereas in the USA it was variously called engineering psychology, human factors psychology, or human engineering, with more focus on cognitive processing. System engineering practices integrated principles of engineering psychology, and became trends in military equipment design centers in industrialized countries. (See *Engineering Psychology*.)

1.6 Social Psychology

Military social psychologists conducted hundreds of attitude surveys and experiments concerning soldier and sailor morale and motivation to support formulation of US military personnel policies for WW II. Social psychologists developed small group performance assessment techniques, expanded psychological warfare techniques, added new psychological perspectives to enemy intelligence analyses, initiated studies of prisoners of war, and developed small group performance assessment techniques.

Social psychological studies in several allied countries provided useful information to WW II military policy makers and established use of the social survey as a military personnel management tool. Their applied research solidified generalizable social psychological findings. Most impact-making findings, described at length in *The American Soldier* (Stouffer et

al. 1949), concerned the importance of (a) cultural and personality influences in understanding and predicting behavior, (b) the role of attitudes in predicting and controlling behavior, and (c) the primary group in determining the morale and motivation of soldiers (Johnson 1991).

2. *Uniqueness of Military Psychology*

The unique breadth, scope, and diversity of subject matter pertinent to specialized work in military psychology sets it apart from other domains of psychology. Examples of military psychology topics not readily found in other realms of psychology are presented below.

2.1 *Stress in Extreme Environments*

Since the mid-1950s, military psychologists have studied extensively psychological and performance effects of highly stressful military environments. Diverse stressors not commonly found in civilian life include conditions of fear, sensory overload, sensory deprivation, social isolation, sleep deprivation, sustained operations, high mountain altitudes, climatic temperature extremes of deserts and tropics, severe winters, and living under the sea, in outer space, and on remote stark land masses. Military personnel are exposed to extreme heat in combat vehicles, high rates of vehicle acceleration, vibration, high acoustical noise, high levels of toxic gases and air pollutants in the work station, and even unusual dietary and nutritional mixes (Krueger 1991, 1998). Military research psychologists develop programs involving equipment, operational procedures, preventive medicine guidance, and training to alleviate stress or to increase an individual's ability to cope with multiple stressors to preserve health and performance. On the clinical side, other military psychology programs help uniformed service members and their families adjust to the general stresses of military life.

2.2 *Nuclear, Biological, and Chemical (NBC) Warfare*

Military forces face severe psychological stresses and performance-related problems anticipating sustaining military operations on a NBC battlefield. These range from coping with uncertainty and fear, to compromised communication capabilities, to general lack of confidence in equipment and procedures. Expected performance degradations attributable to wearing chemical protective uniforms are disconcerting to the untrained (Krueger and Banderet 1997). Military psychologists assist military forces, but also pass along lessons to national civil defense and first responder emergency personnel preparing responses to terrorist threats, or to national disasters.

2.3 *Cross-cultural Operations*

Cross-cultural interactions have been studied under both simulated and field conditions. Much of this work was directed toward small numbers of highly specialized Special Operations Forces who possess language skills and cultural training for geographical areas in which they are expected to work. Under United Nations and NATO auspices, military forces conduct peacekeeping and nation-building activities, preventing civil outbreaks while war-torn countries are restructured. These situations present unique psychological challenges to military personnel who are otherwise equipped and trained for combat. Military training must prepare thousands of combatants to perform peacekeeping activities, and to deal one-on-one with paramilitary rebels, civil insurrectionists, and civilian populace refugees in a culture foreign to their own. What is learned about improving the ability of a military person to deal with his or her foreign counterpart has wide application beyond defense activities, and is relevant to other international efforts of government and nongovernmental agencies. Relatively new fields of political psychology and peace psychology benefit from this work.

2.4 *Research Laboratory Continuity*

Military research laboratories in the 50-year NATO alliance have endured as integral parts of one or another military service. Such continuity promotes longitudinal research and development programs based upon frequent sharing of research instrumentation, expertise, and research data, government technical reports and published articles, and frequent exchanges and customer dialogue with operational military agencies—the consumer–user of research and consulting. Collaboration crosses international boundaries. Military research laboratories supply military field commanders with analyzed data, suggested solutions, and recommendations for solving tricky human behavioral problems on the battlefield. Military research psychologists, benefiting from such laboratory continuity, help military services enact numerous improvements in operations and management. For example, prior to the 1991 Persian Gulf War in Iraq, Kuwait, and Saudi Arabia, US forces had not fought in desert climatic extremes since WW II. Military psychologists spearheaded the rapid provision of

preventive medicine guidance for preserving the health and performance of combatants in the desert for US military forces who were expected to rehearse tactical scenarios while wearing bulky chemical protective clothing as they acclimatized to extreme heat in Middle Eastern deserts. Thirty years of biomedical and psychological research data were culled and transmitted in useable handbook format to the fighting forces of several alliance nations (Glenn et al. 1991). The preventive medicine guidance was highly successful in preventing numerous casualties.

2.5 Tackling Large Military Concerns

Military forces contend with, and resolve, major social-psychological problems. Kenneth E. Clark pointed out that military services are concerned with big operations, and the behavioral and social science research that supports those services must tackle big problems—a kind of work he referred to as 'macro psychology' (Crawford 1970). Because of substantial government backing and financial sponsorship, military psychology often mounts a concerted attack on major social problems. Large research projects combine extensive logistical support and well-orchestrated efforts of teams of multidisciplinary scientists. On large practical problems, hundreds of military personnel may serve as research participants for field experiments, and tens of thousands may participate in survey work. Ample research instrumentation is arranged, millions of data points are collected and analyzed, and results are often presented to ultimate policy decision makers at the highest levels of the military or civilian government prompting such study. Changes decided upon as a result of psychological research can be implemented because of the relatively closed-loop military organizational structure, and impact results are subsequently fed back to the source where they can affect even more change in service-wide procedures or policies affecting millions of people.

3. Current and Future Issues in Military Psychology

For eight decades, military psychology has continued to contribute meaningfully to national defense in countries that maintain military forces. Military psychologists bring psychological principles to bear in tackling 'real working world' problems and issues that confront military service personnel. Military psychologists are pacesetters in numerous topical matters of critical importance to military organizations, but which also have far-reaching implications for the civilian populace. In that process, military psychologists have also made significant contributions to psychology as a whole.

3.1 Hot Topics in Military Psychology Research

Military psychologists tackle practical operational military personnel performance-related issues. Those relevant to the twenty-first century include the following: (a) how do armed forces cope with a large influx of military women into jobs traditionally held by men?; (b) how do volunteer military services obtain quality recruits and maintain retention to meet goals of a proficient, ready combat force in fluctuating global economic and employment markets?; (c) how will future soldiers, sailors, airmen, and marines handle wearable personal computers, and how will they cope with possible inundation by computerized data streams on a digitized battlefield?; (d) what role will information warfare play, and once reliance on computerized battlefields has been solidified, how does one deal with data disruptions?; (e) will computerization demote or leave behind those soldiers who are not computer literate?; (f) what role and treatment regimen should be considered for prescribed ingestion of chemical substances proposed as enhancements to human performance?; (g) how will changing paradigms of leadership affect future military operations and what will psychological research have to say about them?; and (h) how must military leadership change to accommodate intercultural alliances?

3.2 The Military as a Social Psychological Experimental Proving Ground

When social, politically correct trends grab the attention of democratically free societies, the governmentally controlled military system often takes on a role as society's social-psychological experimental laboratory. The military is a closed-loop system, in which uniformed personnel literally belong to, and work for, their military bosses 24 hours per day, 365 days per year. In many countries, military personnel are provided with housing, food supply, pay, and medical care systems. Compared with performing social studies in society at large, conducting military 'social experiments,' collecting performance data, and obtaining feedback on how well 'treatments' work in the military is almost assured.

During their military careers, most US military service personnel have typically participated in one or more 'social experiments.' Examples of social studies targeted toward military personnel include the following: (a) integrating the work force through influx of members of all religions, racial and ethnic minorities, women, and, more recently, homosexuals; (b) instituting sexual harassment awareness and other sensitivity training in the workplace; (c) implementing tobacco-smoking cessation, control of recreational drugs and alcohol use, family advocacy programs, personal weight control, physical fitness, and uniform dress regulations; (d) adoption of the British Army's

regimental unit replacement personnel transfer policies whereby a whole military unit's personnel, and dependent families, relocate together as a group from one military assignment to another; and (e) making it mandatory for military personnel to subject themselves to inoculations, experimental drugs and therapeutics, or, owing to insufficient supplies, withholding drug treatments for some personnel.

Military psychologists, therefore, have the opportunity to participate in the enactment of social and organizational change in the military, and their work can have far-reaching implications for society at large.

See also: Engineering Psychology; Military and Disaster Psychiatry

Bibliography

Crawford M P 1970 Military psychology and general psychology. *American Psychologist* **25**: 328–36

Cronin C (ed.) 1998 *Military Psychology: an Introduction*. Simon and Schuster, Needham Heights, MA

Gal R, Mangelsdorff A D (eds.) 1991 *Handbook of Military Psychology*. Wiley, Chichester, UK

Glenn J F, Burr R E, Hubbard R W, Mays M Z, Moore R J, Jones B H, Krueger G P (eds.) 1991 *Sustaining Health and Performance in the Desert: Environmental Medicine Guidance for Operations in Southwest Asia*. USARIEM Technical Note Nos. 91-1 and 91-2, pocket version. DTIC Nos. AD: A229-643 and AD: A229-846. US Army Research Institute of Environmental Medicine, Natick, MA

Johnson E 1991 Foreword In: Gal R, Mangelsdorff A D (eds.) *Handbook of Military Psychology*. Wiley, Chichester, UK, pp. xxi–xxiv

Krueger G P 1991 Introduction of section 3: Environmental factors and military perspectives In: Gal R, Mangelsdorff A D (eds.) *Handbook of Military Psychology*. Wiley, Chichester, UK, pp. 211–13

Krueger G P 1998 Military performance under adverse conditions In: Cronin C (ed.) *Military Psychology: An Introduction*. Simon and Schuster, Needham Heights, MA, pp. 88–111

Krueger G P 2000 Military culture. In: Kazdin A E (ed.) *Encyclopedia of Psychology*. American Psychological Association, Washington, DC and Oxford University Press, New York, Vol. 5, pp. 252–59

Krueger G P, Banderet L E 1997 Effects of chemical protective clothing on military performance: a review of the issues. *Military Psychology* **9**: 255–86

Mangelsdorff A D 2000 Military psychology: history of the field. In: Kazdin A E (ed.) *Encyclopedia of Psychology*. American Psychological Association, Washington, DC and Oxford University Press, New York, Vol. 5, pp. 259–63

Parsons H M 1972 *Man–Machine System Experiments*. Johns Hopkins Press, Baltimore, MD

Stouffer S A, Lumsdaine A A, Lumsdaine M H, Williams R M, Smith M B, Janis I L, Star S A, Cottrell L S 1949 *The American Soldier: Combat and Its Aftermath*. Princeton University Press, Princeton, NJ, Vol. II

Taylor H L Alluisi E A 1994 Military psychology. In: Ramachandran V S (ed.) *Encyclopedia of Human Behavior*. Academic Press, New York, Vol. 3, pp. 191–201

Uhlaner J E 1968 *The Research Psychologist in the Army—1917 to 1967*. US Army Behavioral Science Research Laboratory Technical Research Report No. 1155, US Army Behavioral Science Research Laboratory, Arlington, VA

Wiskoff M F (ed.) 1988–99 *Military Psychology: The Official Journal of the Division of Military Psychology, American Psychological Association*. Erlbaum, Mahwah, NJ

Wiskoff M F 1997 Defense of the nation: military psychologists. In: Sternberg R J (ed.) *Career Paths in Psychology: Where Your Degree Can Take You*. American Psychological Association, Washington, DC, Chap. 13, pp. 245–68

Zinchenko V, Munipov V 1989 *Fundamentals of Ergonomics*. Progress, Moscow

G. P. Krueger

Military Sociology

The study of armed forces is somewhat of an anomaly in the sociological discipline. Although possessing an extensive and cumulative literature, the sociology of the military is rarely included in the university curriculum. Moreover, discipline boundaries for students of the armed forces have been exceptionally permeable. Sociologists of the armed forces have long relied on the work of other students of military in such allied disciplines as political science, psychology, and history. In recent years, there has been an increasing overlap with peace studies and national security studies. Beyond academia there is a larger group—variously, present and past members of the military, defenders and critics of military organization, and journalists—who both give insights and serve as a corrective for professional sociologists of the military. Indeed, few substantive areas in sociology have such a diffuse and broad constituency as does the study of armed forces and society.

One readily observed trend in the sociological study of military phenomena is its widening purview. Where earlier accounts saw the military as a self-contained organizational entity, contemporary accounts regard the military and civilian spheres as interactive. The sense of the broadened scope is captured in the contemporary preference for the term 'armed forces and society' with its more inclusive connotations, as opposed to the more delimited 'military sociology.' Precisely because the study of armed forces and society has become so overarching, it is convenient to present the extant literature by discrete topical constructs: (a) the professional soldier; (b) the combat soldier; (c) the common soldier; (d) the citizen soldier; and (e) organizational change.

1. The Professional Soldier

The basic referents for discussion of military professionalism are to be found in two landmark studies that first appeared in the interwar years between Korea and Vietnam. Samuel P. Huntington, *The*

Soldier and the State (1957), and Morris Janowitz, *The Professional Soldier* (1960), shared a common perspective in that they eschewed negative stereotypes of the military officer. This was in contrast to the contemporaneous thesis of C. Wright Mills (1956) characterizing military leaders as 'warlords' wielding enormous influence in the 'power elite.'

Huntington and Janowitz also agreed that the complexities of modern warfare and international polices required new formulation of military officership. They differed, however, in their conceptual and programmatic portrayal of modern military professionalism. For Huntington, military efficiency and political neutrality require a form of insulation from the values of the larger and more liberal society. Janowitz, on the other hand, proposes that military professionalism should be responsive to, but not overwhelmed by, external conditions such as managerial skills, civilian educational influences, and emergent social forces. Subsequent studies of the professional officer have been strongly influenced by these contrasting ideal types.

A hardy perennial in the professional soldier literature has been the examination of the social origins of career officers and socialization at military academies. Research on this subject has been as notable in European military sociology as in the USA. The general conclusion is that professional self-definitions are much more shaped by anticipatory and concurrent socialization than by social background variables.

In the USA, media attention in 1999 was focused on studies that presented evidence of a 'civil-military gap.' The overall finding was one of a growing social conservatism within the officer corps that was increasingly alienated from the social values of the larger society (Feaver and Kohn 1999). At the same time, however, public opinion surveys reported that the armed forces were accorded the highest evaluation among US institutions.

If research on military professionalism in the USA, Western Europe, and other advanced democracies was becoming more notable in the contemporary period, studies of military officers in other areas followed a different pattern. During the 1970s the literature on the military in Third World countries was quite extensive, but has since declined. The literature on military officers in underdeveloped areas was marked by two quite opposing schools, one seeing the armed forces as 'moderinizers,' the other as 'praetorians.'

2. *The Combat Soldier*

Any discussion of the combat soldier must use as a benchmark the surveys of World War II reported in the volumes of *The American Soldier* by Samuel Stouffer and his associates (1949 Vol. II). These studies reveal a profoundly nonideological soldier. The key explanation of combat motivation was seen as a function of the soldier's solidarity and social cohesion with fellow soldiers at small group levels. Shils and Janowitz (1948) reported similar findings based on interviews with German prisoners of war. The overriding salience of the primary group became an accepted tenet of military sociology.

Moskos' (1970) observations of US combat soldiers in Vietnam, however, indicated that the concept of primary groups had limitations. The combat soldier in Vietnam had a more privatized view of the war fostered by the one-year rotation system in contrast to his World War II counterpart who was in the war for the duration. Moskos' Vietnam research, moreover, found that although the US soldiers had a general aversion to overt patriotic appeals, this should not obscure underlying beliefs as to the war's legitimacy, or 'latent ideology,' as a factor affecting combat performance and commitment.

The increasing use of armed forces in peacekeeping missions starting in the 1990s has focused attention on the contrast between soldiers as 'warriors' or 'humanitarians.' One the one hand, the conventional wisdom is that 'operations other than war' undermine combat effectiveness. Field research, however, indicates that many soldiers themselves view peacekeeping as conducive to overall military effectiveness (Miller 1997). In any event, the peacekeeping literature has become another genre in military sociology, replacing to a major extent the earlier interest on the combat soldier. Much of this was anticipated by Janowitz's (1960) earlier formulation of the emerging 'constabulary' role of the military.

3. *The Common or Enlisted Soldier*

The benchmark referent for any discussion of the common or enlisted soldier ('other ranks' in British terminology) is again the volumes of *The American Soldier* (Stouffer et al. 1949, Vol. I). Never before or since have so many aspects of military life been so systematically studied. These materials largely revolved around the enlisted culture and race relations as well as combat motivation. These issues continue to interest military sociologists, with the more recent topical additions of gender and sexual orientation. A lacuna in the military sociology of enlisted personnel has been the near absence of studies of sailors, airmen, or marines.

The overriding finding of *The American Soldier* (Stouffer et al. 1949, Vol II) was the pervasive enlisted resentment toward the privileged status of officers. The centrality of the enlisted–officer cleavage was further corroborated by other sociologists, who described the military from the vantage of active-duty participation in World War II. Starting in the Cold War period, another distinction in the military structure appeared. The college-educated draftee is described as far more alienated from his enlisted peers of

lower socioeconomic background than he is from officers with whom he shares a similar class background. In the Vietnam War, the most significant cleavage was between single-term servicemen and career servicemen, cutting across ranks (Moskos 1970). In the post-Cold War era, yet another cleavage has appeared, that between soldiers serving in combat units and those in support units.

One of the most celebrated findings of *The American Soldier* was the discovery that the more contact white soldiers had with black troops, the more favorable was their reaction toward racial integration (Stouffer et al. 1949, Vol. II). Such social science findings were used to buttress the arguments that led to the abolishment of racial segregation in the armed forces. By the early 1950s this integration was an accomplished fact, resulting in a far-reaching transformation of a major US institution. Following ups and downs in race relations during the 1960s and 1970s, the armed forces by the 1990s were viewed as model for black leadership in a racially integrated institution. One key finding, however, was that blacks consistently take a more negative view of race relations than do whites.

If race relations, relatively speaking, were positive in the armed forces, the interactions between men and women were viewed as more problematic. By the 1990s, the role of women had greatly expanded in the US armed forces to the point where women were in nearly all positions excepting direct ground combat. Much public and media attention was focused on recurrent scandals involving sexual harassment and adultery in the military. Indeed, between 1995 and 2000 more books were written on gender than on any other topic in the armed forces. One key finding is that enlisted women and women officers were not in accord on the role of females in the armed forces, the former favoring a more limited role than the latter (Miller 1998).

4. *The Citizen Soldier*

A running theme in American military life has been the juxtaposition of the professional soldier and the citizen soldier. The notion of the citizen soldier raises the twin issues of the extent to which military life affects civilian sensibilities of noncareer soldiers and civilian input affects the military system. Although topics such as reserve forces and officer training programs on college campuses are directly related to the concept of the citizen soldier, these topics have not been objects of major research by military sociologists.

The controversies over conscription during the Vietnam War did relate conceptual issues and empirical findings to the sociology of the citizen soldier. Even with the end of the draft in 1973, sociological interest in the citizen soldier remained strong (Segal 1989). The policy debate on the all-volunteer force and military recruitment has largely become one between sociologists and economists.

5. *Organizational Change*

A major paradigm for understanding change in the military organization is the institutional–occupation thesis (Moskos and Wood 1988). Where an institution is legitimated in terms of values and norms, an occupation is based on the marketplace economy. In an institution, role commitments tend to be diffuse, reference groups are 'vertical' (i.e., within the organization), and compensation is based on rank and seniority. In an occupation, role commitments tend to be specific, reference groups are 'horizontal' (i.e., with like workers external to the organization), and compensation is based on skill level and labor market considerations. An ideal type formulation, the 'I/O' thesis has served as a basis for much subsequent research in Western military systems outside the USA. The overarching thesis is that contemporary military organizations are moving away from an institutional format to one more resembling that of an occupational one.

In the wake of the end of the Cold War, even more momentous changes are occurring within armed forces of Western societies. The modern military that emerged in the nineteenth century was associated with the rise of the nation-state. It was a conscripted mass army, war-oriented in mission, masculine in makeup and ethos, and sharply differentiated in structure and culture from civilian society. The 'postmodern' military, by contrast, loosens the ties with the nation-state, becomes multipurpose in mission, and moves toward a smaller volunteer force. It is increasingly androgynous in makeup and ethos and has a greater permeability with civilian society (Moskos et al. 2000).

At the turn of the new century, military sociology has yet to find a significant niche within the academic community. Yet military sociologists are increasingly being noted by the media and policy makers.

See also: Cold War, The; Military and Disaster Psychiatry; Military and Politics; Military History; Military Psychology: United States; Police, Sociology of; Professionalization/Professions in History; Professions, Sociology of; Racial Relations; Violence: Public; War: Anthropological Aspects; War, Sociology of

Bibliography

Feaver P D, Kohn R H 1999 *Project on the Gap Between the Military and Civilian Society*. Triangle Institute for Security Studies, Durham, NC

Huntington S P 1957 *The Soldier and the State*. Belknap Press of Harvard University Press, Cambridge, MA

Janowitz M 1960 *The Professional Soldier*. Free Press, Glencoe, IL

Miller L L 1997 Do soldiers hate peacekeeping? *Armed Forces and Society* **23**: 415–50

Miller L L 1998 Feminism and the exclusion of army women from combat. *Gender Issues* **16**: 333–64

Mills C W 1956 *The Power Elite*. Oxford University Press, New York
Moskos C C Jr 1970 *The American Enlisted Man*. Russell Sage Foundation, New York
Moskos C C, Wood F R 1988 *The Military More Than Just a Job?* 1st edn. Pergamon–Brassey's International Defense Publishers, Washington, DC
Moskos C C Williams J A Segal D R (eds.) 2000 *The Postmodern Military*. Oxford University Press, New York
Segal D R 1989 *Recruiting for Uncle Sam*. University Press of Kansas, Lawrence, IN
Shils E A, Janowitz M 1948 Cohesion and disintegration in the wehrmacht in World War II. *Public Opinion Quarterly* **12**: 280–315
Stouffer S A et al. 1949 *The American Soldier*. Vol. I: *Adjustment During Army Life*. Princeton University Press, Princeton, NJ
Stouffer S A et al. 1949 *The American Soldier*. Vol. II: *Combat and its Aftermath*. Princeton University Press, Princeton, NJ

C. Moskos

Mill, John Stuart (1806–73)

J. S. Mill's main contributions to Social Science lay in three areas, political economy, political philosophy, and the philosophy of social science; the major works identified with these three fields are: *Principles of Political Economy*, *On Utilitarianism*, *On Liberty*, *Considerations on Representative Government*, and his *Logic*. Although educated in the intellectual environment of classical Utilitarianism associated with Jeremy Bentham and Bentham's collaborator, Mill's father James Mill, Mill is famous for his amendment to utilitarianism strictly considered, particularly in his essay *On Liberty*, a classic statement of political liberalism. In many works he evinced a recognition of the importance of historical process that was absent from the aspiration to a deductive social science which characterised the previous generation of utilitarians, and the thinking of liberal political economists who would claim his legacy. Mill's iconic status as a liberal has made his intellectual legacy a site of fierce ideological contestation.

Born in London, on May 20, 1806, his precocious education—famously described in the celebrated *Autobiography* (1873)—deliberately prepared him for a career as a social and political thinker and reformer. Although commonly held to have inculcated him with utilitarian principles, his early education was grounded in the classics—his father started him with Greek at age three—and was much wider than this, wider, indeed, than that of most modern social scientists. He spent a year in France in 1820–2 before beginning a career in the Civil Service, following his father's footsteps at India House.

In his *Autobiography* he describes an emotional crisis at age 20 which had important intellectual dimensions and was finally resolved only following the death of his father in 1836. In the intervening period Mill remained an active member of the Philosophical Radicals—a largely extra-parliamentary grouping on the more radical wing of those seeking reform of parliament culminating in the 1832 Reform Act—but his underlying philosophical position was undergoing a radical reappraisal. This was first outlined in a series of articles on 'The Spirit of the Age' in the periodical *The Examiner*. Establishing some distance from his father and his upbringing, he claimed, liberated his thinking from Benthamite utilitarianism, and led him to promote the ideal of self-cultivation (which his own *Autobiography* sought to encourage). He was inspired in the development of the imagination and emotions by his relationship with Harriet Taylor, a married woman whom he met and fell in love with in 1830. To her he also credited considerable influence in the intellectual development of his arguments. They were married eventually in 1851, after her husband's death. Despite this awakening of feeling, Mill rejected intuitionism as a basis for philosophy and was committed to extending the experiential methods of the natural sciences to the social. The new dimensions in his thinking did, however, stress the limitations of the associationist psychology which had informed his own upbringing, and introduced ideal-regarding criteria—standards which look to the fulfilment of certain ideal principles—into the predominantly consequentialist character of his inherited utilitarianism. Furthermore Mill increasingly recognised the thick historical texture required for an understanding and appraisal of social and political institutions. Through his father's Scots education and his teacher Dugald Stewart, Mill had access to the thinkers of the Scottish Enlightenment—David Hume, Adam Smith, William Robertson, and Adam Ferguson—who had developed sophisticated accounts of what would today be called historical sociology, then termed philosophical history. Mill's interest in this can be seen as early as his essay *Civilisation* (1836).

There were new influences which he also acknowledged. These included that of S. T. Coleridge to whom he devoted an essay with his concern to institutionalize historically acquired learning and cultivation in a national clerisy—a kind of secularized church establishment; Alexis de Tocqueville, whose two volumes of *Democracy in America* Mill reviewed for the *Westminster Review* in 1835 and 1840 and who had a profound effect on his thinking about the need to manage the political effects of the historical movement to more democratic societies; and Auguste Comte, whose sense of history as a process of rational amelioration survived (frailly) in Mill, even after his rejection of the more elitist policy implications which Comte had drawn from it.

Impressed as Mill was by Tocqueville's account of local democracy in American townships, and seeing democratic culture as the inevitable future for Euro-

pean commercial polities, Tocqueville's worries about egalitarian mediocrity reinforced Mill's concerns, already voiced in *Civilisation*, about the stifling of individuality, energy, and dissent in the emergent mass societies. Although America had abolished an aristocracy of taste or intellect, this had not eradicated deference, which Mill now saw being abjectly paid to 'public opinion (which) becomes in such countries a species of religion and the majority is its prophet.' A continuing concern for Mill, henceforth, was, as he put it in his review of Tocqueville's work 'that there should exist somewhere a great social support for opinions different from the mass.'

Mill's commitment to radical politics itself faltered at times following disillusionment with the Radicals' Whig allies in the Reform Parliament of 1832. The major focus of his political activity at this time became a new radical periodical, the *London and Westminster Review* (1835). It was from the early 1830s too that he began work on his *Logic*.

Mill recognized that, since the application of the principle of utility required an assessment of consequences, a knowledge of social and historical context and process was needed to make such assessments. He also recognized that traditional utilitarianism—considered as a foundation of social analysis, rather than as a regulative principle—had failed to provide that rich contextualization or knowledge of historical processes. Macaulay famously had lambasted Mill senior's *a-prioristic*, deductivist *Essay on Government*. The influences on Mill's early maturity, following his 'breakdown'—Saint-Simon, Comte, Coleridge, and Carlyle, and, in due course de Toqueville—all reflect a preoccupation with remedying these defects. Mill, in fact, had access through his father and his father's tutor Dugald Stewart, to an earlier tradition of historical thinking in the Scottish enlightenment of the eighteenth century. Such thinking had operated through theorizing the effects of large-scale social changes, for example, economic and social mobilization, on human reasoning processes and ultimately on social conventions and forms of political association. By contrast, Bentham and Mill's father, James, had attempted to model, understand, and explain political phenomena deductively, drawing out the implications of utilitarian psychology.

Mill rejects this approach. His *Logic* is an analysis of the methods appropriate to the different branches of science, based on the assumption that the only grounds for deductivism is the assured existence of the uniform operation of physical causes. In Book VI *On the Logic of the Moral Sciences* Mill discussed the extent to which the human sciences diverged from the natural. His discussion has become a suitable point of departure for much teaching and theoretical reflection in the social sciences ever since.

In the course of outlining the possibility of a social science Mill subjected to criticism two alternative methods which he called the geometric and the chemical methods, by analogy with sciences already treated earlier in the *Logic*.

The geometric method was that pursued by his father in his ill-fated *Essay on Government*—an attempt to deduce the optimum form of government from certain posited axioms about the selfish motivations of human beings. This cannot work, because deductive operations presume a uniformity that cannot cope with the quintessential sociopolitical circumstance of multiple, simultaneous, and conflicting causes impinging on a given circumstance. Geometrical deduction is not even yet mechanics, which, crude though it is, can at least cope with this multiplicity of forces.

The chemical method is the attempt to infer causal laws from the interaction of complex entities at (at least) one remove from their underlying physical causes. Mill's thought here seems to be that if real causes take place at the level of physics, 'chemical' phenomena represent epiphenomena which are related *in some systematic way* to those underlying real physical causes, and that by a complex of observation and deduction the chemical properties of physical elements might be derivable at this level. The limitations of this method are that it can only apply where the observed phenomena are *invariantly* related to underlying causal properties—as in the case of chemical compounds whose properties relate systematically to those of their constituent elements. Since the underlying causes of social phenomena are many and various and not elemental or invariant in their operation, it is not possible to treat political facts in the manner of chemical facts, nor to induce from experiment or observation, any reliable causal regularities. Mill instances the attempt to answer the question whether regulating imports is conducive to economic prosperity or not. Even supposing two apparently similar polities pursuing different policies in this respect would not constitute a test case, argues Mill. The reason is that different trade policies are not ultimate 'properties of Kinds' but presuppose a whole raft of differences in opinions, habits, and so on, which might themselves influence the success of the economy, with or without free trade.

For Mill, 'true' causation operated only at the level of local physical events. A human science that sought to base itself on causal generalizations, therefore, must ultimately be grounded in the discovery of physiological causes. However this, he recognized, was too far below the level of 'the general circumstances of society' which he sought to understand, for the latter, at least in the present state of knowledge, to be unpacked in terms of the former. In the meantime he thought, empirical induction about the effects of circumstance on human character formation (which he called 'ethology'), although it does not yield full causal knowledge, could produce 'empirical laws' of greater or lesser range and accuracy, and which might certainly be useful presuppositions against which to consider the likely consequences of policy initiatives.

Such empirical laws yielded what he called (following Comte) both 'uniformities of co-existence' (social statics) and 'uniformities of succession' (social dynamics). Although neither the deductive nor the inductive method was sufficient in itself, such confidence as we can have in deductive conclusions can be derived, Mill thought, from 'collating the conclusions of the ratiocination either with the concrete phenomena themselves, or when this is unobtainable [as in the case of social science], with their empirical laws.' This combination of deductive and inductive reasoning brought to bear on even very complex phenomena, whilst it could not achieve the level of predictive accuracy available in the natural sciences, could prove extremely useful to the 'wise conduct of the affairs of society' (*Logic* VI, ix, §§1,2). It might even, in time, be linked with or 'resolved into' the real causal laws at the level of physiology, so effecting a complete union of the social with the physical sciences. Yet even if this were not fully achieved, there was, Mill insisted, the possibility of constructing a science of society or what he called 'that convenient barbarism'—sociology.

Following his *Logic*, which was a considerable publishing success, Mill returned to political economy. His father had taught him economics as a child, and he published in the subject at age 16. His *Essays on Some Unsettled Questions in Political Economy*, written in 1830–1, now found a publisher (1844) and he settled to the composition of the *Principles of Political Economy, with Some of their Applications to Social Philosophy* (1848). Not only (as the subtitle indicates) did he regard economics as inseparable from the rest of social philosophy, he also regarded the relationship between economics and social theory more widely considered as being the most important part of it. Whilst composing it, Mill was painfully aware of the Irish potato famine, and the impact of economic orthodoxies on its course. Seeking to avoid specific polemical issues Mill was nevertheless concerned to stress that political economy was 'far from being a set of maxims and rules, to be applied without regard to times, places, and circumstances' (*Hansard* v.190, 1525). Thus, although in pure economics he regarded himself as a follower of Ricardo, he distanced himself from those who took Ricardo to have established a purely deductive school. Historians of economic thought disagree as to precisely how to characterize this relationship, and who these economists might be (see Hollander 1985, Vol. II, pp. 914ff). It is clear however, that Mill regarded good economic method to combine both deductive and inductive operations, and in policy stressed the importance of knowledge of local conditions—such as those in Ireland or India—in applying that method. Economic science, he thought, operated under those limitations applying to the social sciences generally, outlined in the last volume of the *Logic*—deductions were only reliable if the generalizations which generated them were the result of physical causes, and even then only if other causal effects did not intervene. Neither of these conditions, in his view, obtained in economics any more than in the rest of social science. This also informed his view of the limited role which mathematical or formal modeling might be expected to play in economics—for without certainty or quantification such precision was likely to be misleading.

Mill's insistence on subordinating economic analysis to policy is clear in his famous chapters on socialism in the *Principles*, which undergo considerable development in the first three editions (1848, 1849, 1852), under the influence of the European revolutions of 1848 which had reopened debate on the subject, and of Harriet Taylor whom he finally married in 1851. Originally sceptical about co-operation, Mill now discussed the ideas of Fourier as a supposedly superior synthesis of the ideas of Owen and Saint-Simon, which overcome earlier worries about incentives. He came to see the emergence of co-operative forms as both an evolutionary possibility within capitalist economies and, once population could be voluntarily controlled and inequalities overcome through control of inheritance, as a desirable social form in which individuality and moral progress, far from being under threat, could flourish better than in conventional economies.

Such themes integrate with the political writing to which Mill returned, at the end of the 1850s. He produced a trio of important essays, long in the making: *Utilitarianism*, *On Liberty*, and *On Representative Government*. These worked out his concern to articulate a conception of progress consistent with utility, as a part of which he sought to support diversity of opinion and to protect social and political institutions from the effects of mediocre uniformity—-whether deriving from the middle classes or the dominance of a socialist-inspired working-class movement.

Utilitarianism articulated a defence of the principle of utility into which Mill introduced significant modifications to Benthamite utility. The first was the claim that apart from the various dimensions of utility analysed by Bentham, and contrary to the latter's claim (made notorious by Mill himself?) that 'pushpin was as good as poetry,' it was possible to distinguish higher from lower qualities of pleasure, and that quality as well as quantity was to be given weight in seeking to apply the utilitarian criterion that actions were 'right in proportion as they tend to promote happiness.' Amongst the higher pleasures were those of altruism, the cultivation of the higher feelings and of the intellect. The former constituted a rejection of the Benthamite reliance on psychological egoism, the latter introduced what many have seen as an alien, ideal-regarding element into utilitarianism. Mill's second innovation was his claim that the principle of utility should be deployed as a criterion of moral or social principles, practices or institutions, and *not* applied directly to regulate individual actions. The

distinction has since been refined and elaborated as that between rule utilitarianism and act utilitarianism. Although Mill does not make the distinction in these terms, he makes it clear that utility is not meant to substitute for conventional morality, but rather is used to appraise and correct its rules. He is also clear that although justice is grounded in utility, appeals to utility cannot be used to overthrow (any but?) the hardest cases.

Classical utilitarianism was premised on the integrity of each individual's experience of utility. Each was to count for one and no more than one, and each was the best judge of their own utility. Mill's two modifications allow privileged judgements to be made, capable of overturning the subjective claims of the individual but they also allowed judgements of progress to be made which were not purely quantitative.

In *On Liberty* Mill formulated what he disarmingly called a 'very simple principle' introduced, in all but name, in his *Principles of Political Economy*. There he remarked that 'there is a circle around every human being, which no government, be it that of one, the few or the many, ought to be permitted to overstep' (*PPE*, Vxi, §2). Mill's principle was that: 'the sole end for which mankind are warranted, individually or collectively, in interfering with the liberty of action of any of their number, is self-protection, that the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant' (Acton 1967, pp. 72–3).

Mill's argument has become an icon for liberals and occasioned a huge literature both academic and polemical. The academic controversy has centered around two issues—whether Mill's defense of the principle of liberty was consistent with his claimed loyalty to the principle of utility, and whether his defence of the integrity of the sphere of individual liberty is, in the end, coherent.

Mill denied that his argument involved grounding liberty in abstract right distinct from utility, which was still the 'ultimate appeal on all ethical questions.' It was, though, 'utility in the largest sense grounded on the permanent interests of man as a progressive being' to which one had to appeal, and not the maximization of existing preferences. This invokes the modifications made to the doctrine of utility in terms of the possibilities of increasing not merely the quantities, but the qualities and refinements of pleasure available to a population. Briefly, moral progress presupposes innovation, and innovation presupposes (though it does not guarantee) liberty. Liberty of thought and speech could generate and promulgate new ideas and test old ones, avoiding the static mediocrity that he feared so much. Liberty in the sphere of private action enabled experiments in living amongst which might emerge new and more beneficial social forms, new forms of domestic partnership and co-operative production in the sphere of the economy, being two such examples. On this view liberty is justified in instrumental terms as a means to a social progress defined, or definable, with reference to utility ('in the largest sense … '). However, Mill's enthusiasm for liberty—redoubled in the closing passages of *On the Subjugation of Women*—his celebration of what he disarmingly calls the 'intrinsic worth of individuality' and the cultivation of the personality as a 'noble and beautiful object of contemplation'—has a kind of autonomous aesthetic which have seemed to many commentators to go well beyond the instrumental role claimed for it in his more programmatic statements.

Mill also had great difficulty sustaining the boundaries of the 'sphere of private action' which he attempted to mark by distinguishing 'self regarding actions'—which were not to be interfered with, and 'other regarding actions' which were, in principle, candidates for legislative restraint. The viability of this in turn involved distinguishing between actions which merely 'affect' and those that 'affect the interests of' others. But the notion of interests, whilst clearly different from wants or preferences, is a notoriously difficult one to deploy—particularly for liberals—without establishing paternalistic claims over individuals' preferences.

Mill's discussion of the application of this principle is an object lesson of the sensitivity required in the application of philosophical principles to areas of policy, but to many it has revealed a residual paternalism.

Considerations on Representative Government, Mill claimed, embodied the principles to which he 'had been working up during the greater part of my life.' It was certainly, amongst other things, an attempt to show how his commitment to liberty as a part of a revised conception of utility could be embodied in political institutions. It revealed an acute historical sensitivity to the broad circumstances that made representative government possible, to the generation of dispositions in the population and to those processes—the race, as he put it, between democratization and education—that might imperil it. Representative government is fixed firmly within the presuppositions of a progressive, national, metropolitan state, and for all Mill's celebration of diversity he evinces little enthusiasm for sustaining relict regional or national cultures such as the Celtic-speaking nations.

Once representative government is possible—and it is not possible for all stages of human development—Mill considers there are two criteria by which government should be judged—efficiency and education. The one seeks the optimum use of the existing good qualities; the second seeks to augment the virtuous qualities in the population at large. These can be seen as a version of two principles widely seen as central to many nineteenth-century thinkers—stability and progress. Much of the work is a consideration of various political institutions in light of the trade-off between

these two principles. Two examples may suffice—those of voting and local government.

Like all utilitarians, Mill rejected abstract or natural rights, and did not derive the vote, any more than liberty, from such considerations. The right to vote was grounded in the important necessity of ensuring that each—even, as he later publicly argued, women—had an institutional expression through which they could safeguard their interests. But this did not entail the privacy, equality, or unlimited domain of the franchise. All of these were to be assessed in consequentialist terms, through their contribution to Utility. Consequentialist judgments are sensitive to the empirical circumstances of that which is being judged. Mill acknowledged that privacy of the ballot box might once have been necessary to avoid undue pressure from employees, landlords, or social superiors. The danger had now, he thought, come to be from individuals regarding the vote as an expression of their selfish wishes. Public voting would, he thought, make it necessary for voters to justify their decisions to their fellows and this might require them to decide on grounds of public reason rather than private rationality. Consequentialist considerations also informed his thoughts about the electorate. Though all adults must have access to the vote for protection, equal voting is not conducive to the best outcome—part of which is to make the best use of the talents available. Mill advocates an educational test (at least partly to encourage self-improvement), disfranchizing those receiving welfare benefit, and proportional representation and plural voting for university graduates and professional classes. This ensured that minority and educated opinion would at least be heard in parliament, and mitigated the potential for majority tyranny, exercizing benign influence on less-informed representatives. For similar reasons he opposed the mandating of representatives. Finally Mill regarded representative bodies as incompetent to draft or engage in detailed amendment of legislation, which should be left to commissions of experts. Mill thought such considerations enabled a distinction to be drawn between 'True and False Democracy.' In True Democracy the role of the People is 'strong enough to make reason prevail but not strong enough to prevail against reason.'

Mill's discussion of local government stresses the educative role it can play in giving the inexperienced scope for political participation in a limited sphere where it cannot damage great national interests. Devolution must not, however, be so great as to exclude the interest of at least some of the educated from whom lessons might be learnt, nor, he thought, should it ever escape the superintendence of the national government.

Mill sat as Liberal MP for Westminster—a notoriously radical seat since the 1780s—in the 1865–8 Parliament, his candidacy provoking huge demand for cheap popular editions of his *Political Economy*, *Liberty*, and *On Representative Government*. True to the spirit of his observations in *On Representative Government* he refused to pledge—or indeed campaign—prior to the election. The most famous cause he espoused during that period was to seek—albeit unsuccessfully—to amend the 1867 Reform Bill so as to grant votes to women. This became a major cause for him during the rest of his life and his *The Subjection of Women* (1869) has become one of the canonical texts in the study of women's political emancipation.

Harriet Taylor had died in Avignon in 1858, and from that time on until his death Mill spent half of each year there. Mill himself died in Avignon on May 7, 1873. Harriet's daughter, Mill's stepdaughter, Helen, acted as Mill's secretary and arranged the publication of a number of his papers after his death, including the *Autobiography* and the *Chapters on Socialism*. In the latter Mill pursued themes explored in the later chapters of his *Principles of Political Economy*, although with even greater open-mindedness as to the possibility of one form or another of socialism proving itself superior. Although he thought revolutionary socialism and a centralised economy both unjust and 'chimerical,' and that even temperate, though extensive, forms of socialism were 'not available as a present resource,' he recognized the high moral values implicit in some of these and was firmly of the opinion that various socialist schemes 'have a case for a trial.' For him though, the criteria to be applied in such a trial were essentially those of the development of human character. The irresistibility of the advance of democracy was a given for Mill, and he recognized that the implications of this were still working their way through and that when they did so they would penetrate to 'the very first principles of existing society'—in which he included private property. Yet, whilst accepting this, his concern to balance values which the mass of the people might not yet share is one of the central tensions in his work, informing what has come to be seen as his central preoccupation—the defense of liberty.

Mill's *Principles of Political Economy* enjoyed huge—almost canonical—status in the middle of the nineteenth century, down to the marginalist revolution associated with Jevons and Marshall. It is now no longer of technical interest to economists but remains an important chapter in the history of economic, and social and political, thought. That part of his *Logic* devoted to the philosophy of the social sciences remains an important statement of a particular approach in sociology. Mill's political writings have a more ambiguous legacy. Political Science now takes forms—and has preoccupations—Mill could not have foreseen; as a result his *Representative Government* now has a less practical character than he might have hoped for it. *On Liberty*, however, is still often read in the spirit in which it was written—as a statement of moral and political philosophy to be considered as a success or failure in terms of the arguments advanced there.

Perhaps what would have been most satisfying of all to Mill, his own character as a public moralist, and the ideal of character formation, have re-emerged recently as an acknowledgedly central preoccupation of his work.

See also: Comparative Studies: Method and Design; Consequentialism Including Utilitarianism; Democracy; Emergent Properties; Enlightenment; Freedom/Liberty: Impact on the Social Sciences; Freedom: Political; Individual/Society: History of the Concept; Liberalism; Liberalism: Historical Aspects; Natural Law; Political Economy, History of; Utilitarianism: Contemporary Applications

Bibliography

Acton H B 1967 *A System of Logic*. Everyman, London (new impression of the 8th edn. *Utilitarianism, On Liberty and Considerations on Representative Government*. London, 1910 and often reprinted)

Berger F R 1984 *Happiness, Justice and Freedom: The Moral and Political Philosophy of John Stuart Mill*. University of California Press, Berkeley, CA

Burns J H 1957 J. S. Mill and democracy 1829–1861. *Political Studies* **V**

Collini S (ed.) 1989 *On Liberty with the Subjection of Women and Chapters on Socialism*. Cambridge University Press, Cambridge, UK

Collini S, Winch D, Burrow J 1983 *That Noble Science of Politics*. Cambridge University Press, Cambridge, UK

Cowling M 1963 *Mill and Liberalism*. Cambridge University Press, Cambridge, UK

Eisenach E J 1998 *Mill and the Moral Character of Liberalism*. University Park,

Grey J 1983 *Mill on Liberty: A Defence*. Routledge and Kegan Paul, London

Hamburger J 1965 *Intellectuals in Politics: John Stuart Mill and the Philosophical Radicals*. Yale University Press, New Haven, CT

Hollander S 1985 *The Economics of John Stuart Mill*, 2 Vols. Blackwell, Oxford, UK

Laine M (ed.) 1991 *A Cultured Mind: Essays on Mill Presented to J. M. Robson*. Toronto University Press, Toronto

Leavis F R (ed.) 1967 *Mill on Bentham and Coleridge*. Chatto and Windus, London

Packe J M St. 1954 *The Life of John Stuart Mill*. Secker and Warburg, London

Rees J 1985 *John Stuart Mill's On Liberty*. Oxford University Press, Oxford, UK

Robson J M (ed.) 1963/1991 *Collected Works of John Stuart Mill*. Blackwell, Toronto

Robson J M 1968 *The Improvement of Mankind*. Blackwell, Toronto

Ryan A 1975 *John Stuart Mill*. Routledge and Kegal Paul, London

Schneewind J B (ed.) 1968 *Mill: A Collection of Critical Essays*. Macmillan, London

Schwarz P 1968 *The New Political Economy of John Stuart Mill*. Weidenfeld and Nicolson, London

Skorupski J (ed.) 1998 *Cambridge Companion to John Stuart Mill*. Cambridge University Press, Cambridge, UK

Stillinger J (ed.) 1969 *Autobiography*. Oxford University Press, Oxford, UK

Ten C L 1980 *Mill on Liberty*. Oxford University Press, Oxford, UK

Thompson D 1976 *John Stuart Mill and Representative Government*. Princeton University Press, Princeton, NJ

Winch D (ed.) 1985 *Principles of Political Economy (Books IV and V)*. Penguin, Harmondsworth, UK

I. Hampsher-Monk

Millennialism

Millennialism (derived from the Latin term, millennium, which means literally 1,000 years) is the belief that a prolonged period of bliss is a future prospect, usually expected to commence after catastrophe which will end the present dispensation and perhaps the world itself. Salvation and survival into the new era will be confined to those who believe this prophecy and conform to certain stipulated demands of faith and morals. Millennialism has inspired numerous movements throughout the world and given rise to outbursts of intense social unrest, but has also created new patterns of social consciousness.

1. The Term Millennialism

The terms millennialism, millenarism, and millenarianism are more or less synonymous, although some writers (Harrison 1979) have sought to distinguish millennialism as scholarly theological concern with prophetic exegesis, and millenarianism as applicable to social movements preaching the imminence of the millennium. This distinction, however, is not observed generally. The term chiliasm (Greek *chilioi*—a thousand) is also employed in reference to social manifestations of belief in the early outworking of prophecy.

2. Historical Provenance

Millennial ideas arose in Zoroastrianism which exerted an influence on Judaic eschatology in which it became a strong theme. Jewish apocalyptic literature linked the millennium to the expectation of a returning messiah, a king to save his people, and this association informed the expectation among early Christians that Christ would soon return to earth to rule over his kingdom for 1,000 years. The concept gradually has acquired wider application to contexts other than the

Judeo–Christian, and is used loosely to refer to a long period of time rather than specifically to 1,000 years, but it is in Christianity that millennialism is most fully, albeit controversially, articulated.

2.1 Development in Early and Medieval Christianity

In Christian history, an assortment of New Testament texts, especially Revelation 20, have been cited alongside the Old Testament books of Isaiah, Daniel, and Ezekiel to provide both a sequence of events and allegorical reference points as a basis for a chronology for the prophesied millennium. Some early Church fathers and various heretical movements expected an imminent literal outworking of prophecy, and looked forward to indulgences and pleasures which they had hitherto denied themselves. Augustine changed Christian thinking by regarding the contemporary epoch of the Church to be the millennium, at the end of which Christ would return and establish the conditions of eternity. Millennial ideas did not disappear, however, and found expression in the theories of Joachim of Fiore in the twelfth century and in various heretical movements, and they received a new impulse from sects that arose in Reformation times.

3. Pre- and Postmillennialism

For later Christians, Daniel Whitby (1638–1726) formulated the position which became widely accepted, that the advent would occur at the end of the millennium rather than before, thus establishing a distinction between postmillennialists (who believed that the advent would follow the millennium) and premillennialists. Postmillennialists held optimistic beliefs about social progress and the diffusion of the Christian message through missionary activity, thus paving the way for the return of Christ. The designation 'millennialist' was henceforth applied to premillennialists, those expecting the imminent second advent of Christ. Among themselves, premillennialists differed about other matters, among which were: whether the tribulation alluded to in I and II Thessalonians would occur before or after the 'rapture' of the saints (i.e., their being caught up in the clouds to join Christ); and whether the millennial kingdom would be established on earth (as believed by Jehovah's Witnesses) or in heaven (as maintained by Seventh-day Adventists). Most millennialists have been 'mortalists,' denying that man possessed an immortal soul.

3.1 The Development of Premillennialism

Millennialism has given rise to numerous episodes of fanaticism. After the Reformation, the 'open Bible' policy of Protestants gave millennial expectations common currency. Many movements inspired by expectations of the early fulfillment of prophecy have been of short duration, particularly those with adherents who took it upon themselves to take up arms to help bring the advent and the millennium about. Such were the Anabaptists who occupied the city of Münster in Westphalia in 1534, and the Fifth Monarchy Men who challenged the secular authorities in the 1650s in England.

3.2 Dating in Prophetic Exegesis

Millennialists have not depended on the secular calendar in affirming when the millennium would begin, but have often set dates. Only the most proximate dates have inspired active movements. Expected dates have been calculated by the choice of specific past events as the departure point for the outworking of chronology and the application of numerological principles. Those calculations concerning the fulfillment of prophecy generally have not coincided with the completion of the first or second millennia of the western calendar. Not all movements have committed themselves to predicting specific dates, and others have done so guardedly in recollection of the repeated scriptural warnings that Christ would come 'like a thief in the night.' Jehovah's Witnesses have been perhaps the most persistent in setting dates for the fulfillment of prophecy (having, formally or informally, at different times and successively harbored keen expectations of 1874, 1914, 1925, and 1975). The Seventh-day Adventists, after the disappointed hopes of William Miller's predictions for 1843 and 1844, have abandoned date-setting, while the Christadelphians, although much exercised on the subject and disposed, especially in their early years (1840s and 1850s), to look for signs in contemporary political events to match biblical prophecies, have for a long time disavowed the possibility of acquiring accurate foreknowledge of the time of Christ's return.

3.3 The Size of Modern Millennial Movements

Whilst postmillennial orientations stimulated missionary and revivalist activity in the eighteenth century, particularly in North America, the pessimism implicit in premillennialism previously had inspired movements which operated in considerable tension with the state and the wider society, and these hostile movements usually were small and localized. Growth of literacy and improved communications in the nineteenth century allowed sectarian premillennialists to organize much larger, enduring, and widespread movements, some of which, by the end of the twentieth century, had followings counted in millions (e.g., the Seventh-day Adventists, Jehovah's Witnesses). As the need for better accommodation with modern society

and the agencies of the state grew, so these sects became less aggressive and their insistence on the imminent end of the existing dispensation less strident.

3.4 The Social Constituency of Millennialism

It has been a general assumption that those attracted to millennial sects are not only pessimistic about general social conditions, but are likely to be drawn from disinherited social groups, and the relative deprivation thesis has been frequently invoked to explain their appeal. As a generalization this may be warranted, but by no means all of those attracted to millennialism have formed their views in socially or culturally disadvantaged circumstances. Prominent Protestant thinkers—Joseph Mede, the theologian, and the scientist, Isaac Newton, among them—were devoted to numerological exegeses of biblical prophecy. The initial leaders of the strongly adventist (premillennial) Catholic Apostolic Church included several Anglican priests and Church of Scotland ministers, and all but one of that church's 12 apostles were either lesser aristocracy or professional men. Several lay supporters were, or became, MPs or higher civil servants. The originators of the contemporaneous movement devoted to premillennialist prophetic exegesis (later known as the Plymouth Brethren), were clerics and ordinands, and local leaders were often men of some substance. Nineteenth century millennialism had an appeal which extended beyond the working class.

4. When Prophecy Fails

Millennial movements all face the prospect of the failure of prophecy, and examination of their history reveals diverse responses to this problem. Gager, following Festinger's study of a contemporary adventist cult (*When Prophecy Fails* 1956) has contended that the expansion of early Christianity was attributable, at least in part, to the failure of millennial expectations (Gager 1975). Believers sought to overcome the cognitive dissonance of failed prophecy by persuading others that they were not wrong: the next best thing to being proved right was to persuade others to join you and to accept your rationalizations for the error. The evidence on early Christianity is perhaps too sketchy to substantiate this contention, while Festinger may have studied too small and ephemeral a movement (in which he and his team constituted too influential a group) to warrant the comparison.

4.1 Two Responses to Failure

Even so, some millennialist movements have triumphed over repeated failure of prophecy, and have done so by various different strategies. Perhaps the most successful of contemporary millennial bodies, the Seventh-day Adventist Church, has flourished by cultivating an encompassing program of educational, dietetic, and therapeutic welfare in which to engage members' energies. This policy originated early in the movement's history, and has expanded despite the apparent inconsistency of devoting resources to welfare in a world due soon to be overturned by the cataclysmic events supposedly foretold in the Scriptures. In contrast to this many-sided strategy to occupy the faithful, Jehovah's Witnesses, whose theology teaches that God's Kingdom is already established but not yet visible, seek to engage their following single-mindedly in the work of recruitment. This is a response to failed prophecy more or less on the Festinger model. By deploying their 'publishers' (the movement's designation of its members) in extensive house-to-house calls, the movement has developed a systematic method not only of winning new adherents but, perhaps more importantly, of involving members in the regular reaffirmation of their own commitment. The primacy of this latter function becomes apparent when it is acknowledged that at least 200 h of canvassing are on average expended by publishers for each additional new convert.

4.2 The Introversionist Response

The Plymouth (Exclusive) Brethren response to hope deferred has been a classic withdrawal from involvement in the wider society into a more introversionist position. The movement has ceased to proselytize: although still undertaking token public proclamation of 'the truth,' they are far from exhorting listeners 'to come and join us'—joining is not an option that is made easy. The Brethren, who regard their assemblies as already a foretaste of heaven, expect the Holy Spirit to guide all sincere Christians to them and their way of life. Maintaining communal purity has become a primary concern, hence the insistence on separation from the inherently evil world, and the diffidence respecting would-be converts. A less extreme example of the same shift of orientation from keen millennialist expectations to a more introverted position is also found among the Christadelphians.

4.3 The Demise of the Catholic Apostolics

Perhaps the most radical resolution of the experience of failed prophecy was that of the Catholic Apostolic Church. The Church confidently expected the advent to occur in the lifetime of its original 12 apostles, who were designated (by members who claimed the gift of prophecy) in the early days of the Church's formation in c. 1830–5. An apostle died in 1855, and a shocked church acknowledged that there was no brief to

replace him. The last of the 12 died in 1901. Thereafter, the Church, according to its own rubrics, lacked legitimacy to ordain ministers. Eventually, in the 1970s, the last minister, ordained by the last apostle, also died. The self-prescribed denial of the competence to make new appointments virtually pre-ordained the demise of the Church if prophecy failed. A schismatic branch in Germany survives, the Neue Apostolische Gemeinde, in which designated prophets took a different view and nominated successors to apostles as they died.

5. *Millennialism in Other Religious Traditions*

The millennial theme represents a final outworking of history, in which the wicked are punished, and the righteous are saved in a process which is envisaged generally as involving severe tribulations. The term is applied loosely, however, and the constituent elements of millennialism, as manifested in Christian eschatology, are often absent in so-called millennialist movements in other religious traditions. In Judaism, the messianic element is strong, and has stimulated the emergence of messianic claimants. The most celebrated of these was Shabbatei Tzevi (1626–76), a self-styled messiah whose following, and that of his various successors, was widespread throughout the Middle East and Europe, and which persisted even after he recanted and converted to Islam. In Islam, the Shi'ite tradition of a hidden imam, due to return to lead the faithful in some future time, incorporates both messianic and millennial themes, but the rising of the Mahdi, Mohammed Ahmed ibn Seyyid, in nineteenth-century Sudan appears to have been a more explicit admixture of revolutionary political aims and religious fervor. The visions of future salvation, following a period of tribulation, found in the claims made in the Hindu tradition for Kalki, an avatar of Vishnu, and the prospect of the forthcoming troubled age of Mappo, in Mahayana Buddhism, have lacked the power to inspire significant social movements and scarcely qualify to be designated millennialist. The other-worldliness of Indian religion appears to render it inhospitable to millennialist themes. A more explicitly millennialist movement occurred in late nineteenth-century China, in the T'ai'ping rebellion, but this movement was influenced directly by Christian ideas.

6. *Third World Instances*

A wide assortment of religious uprisings among tribal peoples are often termed millennialist. Most of these movements are ephemeral. Some are openly militant, proclaiming either the early return of the ancestors and engaging in practices designation as nativistic, or claiming to be the true originators of all the innovative accoutrements of western civilization. Sometimes a messiah is promised as the savior who will usher in the time of prosperity and bliss, and the prophet harbinger of the new era may himself be cast in the role of messiah. The cargo cults of Melanesia exemplify the collective messiahship of the ancestors for whose return votaries practiced ecstatic rituals or sought to dance the new dispensation into being. The Ghost Dance of 1870 among the tribes of California and its more celebrated successor of 1890 proclaimed collective salvation once the ancestors were induced to return to practice the old ways.

6.1 *Christian Influences in Third World Cases*

Many of these movements were influenced powerfully by Christian millennialism. Christian missionaries were often clergy of fundamentalist persuasion, or were drawn from fundamentalist denominations: millennialism was a stock part of their message. The Kitawala movement in what is today Zambia, Zimbabwe, and Malawi was a radical and semipoliticized millennial cult which owed its inspiration to Watchtower preachers (later designated as Jehovah's Witnesses) of whom Joseph Booth, a sometime Seventh-day Adventist, was a conspicuously influential figure. Suggestions that the Tupinamba and Tupi–Guaraní tribes in South America who, as early as the mid-sixteenth century, were discovered to embrace a cult which taught that they might dance themselves to weightlessness and be carried off to a promised land of deathlessness in the East, illustrated the possibility of autochthonous millennialism, uninfluenced by Christianity, may be discounted, since those who first reported this phenomenon were Jesuits.

6.2 *Mass Magical Movements*

Various Third World movements were designated readily as millennial even though all that they promised were miraculous therapies or new means for detecting witches. Such were the various Zionist sects of South Africa; the pocomania cults of Jamaica, which claimed to counteract obiah, so-called black magic; and the Lumpa Church of Alice Lenshina which fell victim to military intervention ordered by the Zambian government in 1964. These cults have operated essentially at the individual level, seeking to play very little part in inducing changes in social consciousness.

6.3 *Revolutionist Movements and Social Consciousness*

The more avowedly revolutionist movements have often enjoyed greater influence. Such was the case with

the Hau-Hau movement among the Maoris which played a role in the Maori Wars; the Maji Maji Rebellion in East Africa; the afore-mentioned Kitawala movement in Central Africa in provoking the Chilembwe uprising; and the collective resistance movement among erstwhile mutually hostile North American tribes under Tenskwatawa, the Shawnee prophet. Millennialism has often served as an unwitting agency in raising the consciousness of tribal or ethnic identity, or of the need to transcend it in a wider cause. This appears to have occurred among Rastafarians in the West Indies, and more effectively among the Mahdists of the Sudan in the formation of a new state. Prospects of salvation as a chosen people in a new millennium caused Paul to declare that Christians were 'neither Jew nor Greek, bond nor free,' but a new people with a new collective identity.

See also: Cargo Cults; Charisma and Charismatic; Christianity Origins: Primitive and 'Western' History; Fundamentalism (New Christian Right); Judaism; Mission: Proselytization and Conversion; Prophetism; Religion: Nationalism and Identity; Secular Religions

Bibliography

Benz E (ed.) 1965 *Messianische Kirchen, Sekten und Bewegungen im heutigen Afrika*. Brill, Leiden, The Netherlands

Cohn N 1961 *The Pursuit of the Millennium*, 2nd edn. Harper, New York

Festinger L, Riecken H W, Schachter S 1956 *When Prophecy Fails*. University of Minnesota Press, Minneapolis, MN

Flegg C G 1992 *Gathered Under Apostles: A Study of the Catholic Apostolic Church*. Clarendon Press, Oxford, UK

Gager J G 1975 *Kingdom and Community: The Social World of Early Christianity*. Prentice-Hall, Englewood Cliffs, NJ

Harrison J F C 1979 *The Second Coming: Popular Millenarianism 1780–1850*. Rutgers University Press, New Brunswick, NJ

Holt P M 1958 *The Mahdist State in the Sudan 1881–1898*. Clarendon Press, Oxford, UK

Mühlmann W E 1961 *Chiliasmus und Nativismus: Studien zur Psychologie, Soziologie und historischen Kasuistik der Umsturzbewegungen*. Reimer, Berlin

O'Leary S D 1994 *Arguing the Apocalypse: A Theory of Millennial Rhetoric*. Oxford University Press, New York

Pearson M 1990 *Millennial Dreams and Moral Dilemmas: Seventh-day Adventists and Contemporary Ethics*. Cambridge University Press, Cambridge, UK

Penton M J 1985 *Apocalypse Delayed: The Story of Jehovah's Witnesses*. University of Toronto Press, Toronto, ON

Sandeen E R 1970 *The Roots of Fundamentalism: British and American Millenarianism, 1800–1930*. University of Chicago Press, Chicago

Sharot S 1982 *Messianism, Mysticism, and Magic: A Sociological Analysis of Jewish Religious Movements*. University of North Carolina Press, Chapel Hill, NJ

Wilson B R 1973 *Magic and the Millennium: A Sociological Study of Religious Movements of Protest among Tribal and Third-World Peoples*. Heinemann, London

Worsley P 1957 *The Trumpet Shall Sound: A Study of Cargo Cults in Melanesia*. McGibbon & Kee, London

B. R. Wilson

Mind–Body Dualism

1. Introduction

According to one version of the dualist account, a person is a union of two radically different types of entity, one a material body and the other an immaterial mind. Some, like Descartes (1641), have held that mental phenomena are attributes of a mental substance. While our bodies are all fashioned out of the same stuff, each individual mind is a unique 'thinking thing,' an ego identified with the soul. Others, like Locke (1690) and Hume (1739), have held that mental phenomena are immaterial entities, that, taken together, constitute the mind. The mind is a 'bundle' of perceptions and feelings. Since we are aware not only of the messages of the senses but also of at least some of our feelings, desires, and ideas, each person must not only be a center of consciousness but also of self-consciousness. There is another version of dualism based on a distinction between the kinds of predicates that can be used to describe the material and the mental aspects of human beings. There is only one kind of substance but it has two very distinct kinds of properties.

Mind/body dualism, in either form, has raised two major philosophical questions:

(a) 'How is it possible for immaterial mind and material body to interact, as it seems they obviously do, be they substances or properties?'

(b) 'Where does the core of personal identity lie, in the material body or the immaterial mind, in the material or the mental aspects of a person?'

The monistic accounts offered by materialists and idealists purport to be solutions to some of the problems inherent in either version of the dualist account.

2. The Ontological Version of Mind/Body Dualism

The origin of the thesis that mind and body are different *substances* is usually credited to Descartes. He came to this view in the course of reflecting on whether there was any belief that could not be doubted. The program of analysis that led him to his ontological version of mind/body dualism is set out in his *Sixth Meditation*.

> To begin with, I will go back over all the things which I previously took to be perceived by the senses, and reckoned to

> be true; and I will go over my reasons for thinking this. Next, I will set out my reasons for subsequently calling these things into doubt. And finally I will consider what I should now believe about them (Descartes 1641 p. 51).

He convinced himself that his mental states were better known to him than his material surroundings, including his own body, so the indubitable truth must be mental. The first step in recovering knowledge of himself and the world must be to find an undubitable premise or premises from which the main characteristics of the world and of myself as a center of consciousness, that is as a thinking thing, can be deduced. Descartes' method in philosophy and in the sciences was based on his faith in the power of logic to ensure that, step by step, the mind passes from truths to truths. '… there is nothing so far removed from us as to be beyond our reach, or so hidden that we cannot discover it, provided only we abstain from accepting the false for the true, and always preserve in our thoughts the order necessary for the deduction of one truth from another' (Descartes 1644 p. 16). But what could serve as a premise for the deduction of secure knowledge about myself and from that to secure knowledge about the natural world? It must be both thinkable and indubitable. Borrowing from earlier sources Descartes gives us the most famous philosophical argument in history, '*cogito ergo sum*.' It is supposed to undermine the possibility of universal doubt. If someone is doubting that he or she is thinking, then, since doubting is a species of thinking, that person *is* thinking. So the existence of thinking is indubitable.

However, Descartes draws a much stronger conclusion. His argument, so often quoted and parodied, is supposed to prove that a thinking *thing* exists, namely a mind, soul, or ego, that is truly Descartes (Descartes 1641 p. 19). The argument begins with 'I am thinking therefore I exist' (Descartes 1644 p. 194). Though meant to highlight the indubitability of the existence of the thinker as thinking thing, it led straight to the thesis that the mind and the body of a person are two radically different substances. The attributes of a thinking substance, a mind, are immaterial. It is indivisible and, being indivisible, must be immortal. A mind has no properties in common with a material, divisible, spatially, and temporally locatable body. But the body supplies the mind with sensations that are not at the whim of the person, while the mind makes the body move in ways that the person intends. So mind and body must interact. How could such interactions occur when Descartes' analysis leaves each with properties that could have no causal relations with one another? The person or self, being an entity, can be the target of acts of reference, just as material entities like Mt. Blanc are denoted by proper names. While the name 'Descartes' refers to the whole ensemble, body and mind, the pronoun 'I' seems to refer to the ego, the really real Descartes, the immaterial entity that is the mind. It also serves as the source of personal identity.

There is one obvious objection to the move from the indubitability of the existence of acts of thinking to the conclusion that there must be immaterial thinking things. At most we know thinkings must exist, not that there must a substance, an entity that is doing the thinking. The circularity of the argument is clearer in French than in Latin.

> *Je* pense
> donc
> *Je* suis.

Far from this argument proving the existence of a Cartesian ego, since 'je,' appears in the conclusion of the argument, and in the premise, the argument is circular. The only valid argument that the doubting paradox supports is this:

> Doubting exists
> therefore
> Thinking exists.

One might well reject the idea that personal identity is rooted in an immaterial entity while still holding that thinking, feeling, and willing are immaterial processes. The mind/body problem appears on the one hand as the problem of how two quite different kinds of entities could interact, and on the other as the problem of assigning priority between bodily continuity and individuality as the basis of personal identity, or on continuity of mind, as given in consciousness through the power of memory.

3. *The Discursive Version of Mind/Body Dualism*

An alternative way of making sense of the radical differences that seem to mark off mental states of persons from material states of their bodies has been to propose that there are two irreducible sets of predicates, one used for describing the body and its states and the other for describing the mental life of conscious beings. Some body predicates can be used to describe any material thing, while others can be used only of other organisms. Mental predicates, for the most part, can be used only of human beings, though some can also be used of animals, such as 'alert' of a dog. But there is just one entity to which these descriptions are to be applied. For some philosophers this entity is the human body as organism (Searle 1992). For others it is the person, the basic particular of human life (Strawson 1956).

We recollect that when Descartes realized that the mental and material properties of human beings were very different he was tempted into thinking that, therefore, a person must be constituted of two different substances each with its proper set of attributes. By an

ingenious analysis of the material conditions that are involved in the basic criteria of personal identity, Strawson was able to show that the concept of a person is logically prior to 'having a body' and 'having a mind.' The Cartesian distinctions among predicates do not lead inexorably to the theory of two substances, the material body and the immaterial mind, the alleged thinking thing. The basic particulars of the human world are persons. A person must be embodied in order for it to be possible for that person to be reidentified as one and the same person who one had met before. This involves M or material predicates. But a person is also the subject of ascriptions of P or person predicates, which involve consciousness. According to Strawson a person can ascribe consciousness to him or herself only if that person can ascribe such states to others. There must be logically adequate criteria for ascriptions of P predicates. So we are left with two sets of predicates for describing people, both of which are necessary and neither of which is sufficient to encompass the whole of human life.

In Searle's version of this approach, the dualism is founded on the intentionality of mental states, a feature not shared by material states of the organism. Intentionality is the property that a significant sign has of pointing beyond itself, of meaning something, whether or not there is anything in the world to which it corresponds. Mental states are just those states of the organism that do point beyond themselves. But thought is not anything other than an aspect of the biology of human organisms, just as is digestion. This view is far from new. Cabanis remarked that thought is a secretion of the brain, just as bile is of the liver.

4. *Supervenience*

A third suggestion (Kim 1998) is based on the idea that there could be no thought without a material system, such as a brain and nervous system, while there are many material systems that are not conscious or cognitively active. It has been said that the mental supervenes on the material without being reducible to it. This proposal does not seem to touch the deep question of how the states of the material system and the mental states that supervene upon it are related.

5. *Criticisms*

Ontological dualism has been under attack from its beginnings in the seventeenth century. There have been two lines of criticism, materialist and idealist. If either of these monisms could be defended it would put paid to the discursive dualism based on two radically distinct sets of predicates. Both criticisms are based on arguments designed to show that there really is only one kind of property that human beings possess. It follows that there would be only one kind of predicate which could be ascribed to a human being.

5.1 *How Materialists get Rid of the Mind*

Traditionally materialists have tried to solve the interactional aspect of the mind/body problem by eliminating the mental side of the incommensurable pair. If it could be shown that the phenomenon that people had taken to be mental and so radically different from the material attributes of a human being were really material after all, then there would be no ontological problem about the possibility of a causal relation between the states of the sensory systems and sensations and perceptions. One material state can cause another. If it could be shown that all mental concepts could be replaced by material concepts, without loss of content, the mind/body problem would evaporate in a more fundamental way. The logical subject to which the reformed vocabulary would be predicated would be denoted by an expression for a material thing, the body.

How could one show that a vocabulary was so radically defective that it ought to be replaced by a different set of words? In the 1950s several philosophers of science, notably Hanson (1958), who coined the phrase 'the theory-ladeness of observations,' reviving a thesis popular at the beginning of the nineteenth century, pointed out that there is no sharp division between words used in theories and words used to describe phenomena, since the latter incorporate theoretical concepts. For example when the word 'heat' is used in the description of a calorific phenomenon, people literate in science understand it as having a complex meaning, partly phenomenological, partly in terms of the kinetic theory, of molecules in motion.

In an attempt to discredit the psychological vocabulary of one vernacular, namely, English, Churchland (1984) argued that it must be faulty since it was theory-laden with 'folk psychology,' a theory based on mistaken ideas about the nature of beliefs and memories, which are taken to be mental entities. There are no such mental entities, argued Churchland. Folk psychology is false psychological theory. In favor of what terminology should we eliminate the lexicon of everyday life? There is a developing and true theory of cognition, he argued, which is based on a materialist ontology, that of neurophysiology. Thus we should eliminate expressions like 'pain' in favor of 'firing in the c-fibers' and 'the sound of the flute' by an expression for the Fourier analysis of the waveform of the propagated sound wave. There is no mind/body problem, because there is no scientifically acceptable vocabulary for referring to and describing the 'mind' side of the problematic pair of terms. This is not, be it

noted, a proposal to translate the mental vocabulary, term by term, into a physiological vocabulary, but to replace one vocabulary by another, across the board.

5.2 How Idealists Get Rid of Matter

If the assumption of the existence of matter, as the real substrate of the world, is a mistake, then what is left, the mental realm, must be all that there is. There is, therefore, no interaction problem to be solved. This was the route taken by Berkeley (1729). In order to follow his arguments we need to look at a double distinction made famous by Locke (1690), but one which was by no means exclusive to him. Berkeley argued that the distinction between material and mental qualities is bogus. So there is no need for the hypothesis of the existence of matter to be the substance which carries the primary or material properties. According to Locke (1690), ideas are what a person is conscious of, and qualities are the attributes of material things that cause these ideas in a conscious subject. Perception is the having of ideas of qualities. The ideas of primary qualities, such as the bulk, figure, motion, and texture of material things, resemble the material qualities which cause the ideas, while the ideas of secondary qualities, such as experiences of color, taste, and warmth, do not resemble their causes. Though we cannot say just what the secondary qualities of color and so on are in material things, we can say that there must be powers in material things to produce the ideas of them. What grounds these powers? Natural science offers hypotheses about the corpuscular structures on which these powers depend. Newtonian science elegantly fills out these hypotheses with primary qualities, the 'bulk, figure, motion, and texture of the insensible parts.' So felt warmth is the effect on a person of the motion of molecules in the stuff that is felt to be warm. It is only too clear that this account is both close to the way physical science seems to work, and highly contentious. How could the movement of molecules cause a feeling that has nothing of motion in it? The mind/body problem surfaces in even more intractable form in Locke's famous theory of perception.

Berkeley's idealism springs straight from his criticism of the Lockean distinction between primary and secondary qualities. According to Berkeley (1729) the *distinctio*n must be rejected, and with it the distinction between qualities and ideas. If these linked dichotomies are abandoned, then the hypothesis of a material substrate to underpin experience is gratuitous. Furthermore, since the properties of this alleged matter are clearly passive, and have nothing of power and activity in them, by dropping the whole scheme the problem of how matter could *act on* mind disappears. As Berkeley remarks, 'only spirits are active' and the most active of all is God.

There is no mind/body problem because there is no body in the sense of a material substance, unobservable but foundational to the existence of each human being, a something that lies behind the qualities that we perceive. To be is to be able to be perceived. To perceive some thing is to have an idea of it. Only that which is perceived or perceivable exists, so the world is a world of ideas. And this, Berkeley insists, is the ordinary world we all know well. In a way we should perhaps say that the ordinary world is neither mental nor material, since Berkeley's argument attacks the roots of this distinction. There is just the world as we perceive it.

6. Expression and Description

Recently the discursive account has taken a deeper and more subtle turn, in Wittgenstein's querying of the assumption that mental predicates are used to describe states of mind known only to the person who enjoys them. By looking very carefully at the sorts of occasions we use the mental predicates, Wittgenstein (1953) saw that it would be more correct to say that a statement like 'I have an itch' expressed one's feeling rather than described it. The point is this: when we describe something the description and what is described are independent of one another in the sense that the description could be wrong. But an expression is related to the feeling expressed internally, that is to have an itch is not only to have a certain feeling, but to be disposed to do such characteristic things as scratching and saying 'I itch!' Only if the material and immaterial predicates were used in descriptions could the relation between the states they severally described be causal. The feeling and the tendency to express it are both necessary components of what is to 'have an itch,' 'be in pain,' and so on. So in using the first person I do not refer to something within the person, the self. Rather in using first person pronouns, for instance, I express my identity. To be able to use such pronouns is an integral aspect of what it is to have a sense of personal identity.

See also: Biology's Influence on Sociology: Human Sociobiology; Body, History of; Brain, Evolution of; Consciousness and Sensation: Philosophical Aspects; Consciousness, Cognitive Psychology of; Discourse and Identity; Hume, David (1711–76); Identity and Identification: Philosophical Aspects; Identity in Anthropology; Locke, John (1632–1704); Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects; Personal Identity: Philosophical Aspects; Topographic Maps in the Brain; Wittgenstein, Ludwig (1889–1951)

Bibliography

Berkeley G 1729 [1985] *Principles of Human Knowledge*. Fontana, London

Churchland P M 1984 *Matter and Consciousness*. MIT Press, Cambridge, MA
Descartes R 1641/1985 Meditations on first philosophy. In: Cottingham J, Stoothoff R, Murdoch D (eds.) *The Philosophical Writings of Descartes*. Cambridge University Press, Cambridge, UK, Vol. II
Descartes R 1644 Principles of philosophy. In: Cottingham J, Stoothoff R, Murdoch D (eds.) *The Philosophical Writings of Descartes*. Cambridge University Press, Cambridge, UK, Vol. I
Hanson N R 1958 *Patterns of Discovery*. Cambridge University Press, Cambridge, UK
Hume D 1739 [1965] *A Treatise of Human Nature*. Clarendon Press, Oxford, UK
Kim J 1998 *Mind in the Physical World*. MIT Press, Cambridge, MA
Locke J 1690 [1974] *An Essay Concerning Human Understanding*. J. M. Dent, London
Searle J R 1992 *The Rediscovery of the Mind*. MIT Press, Cambridge, MA
Strawson P F 1956 *Individuals*. Methuen, London
Wittgenstein L 1953 *Philosophical Investigations* (trans. Anscombe G E M). Macmillan, New York

R. Harré

Minimum Winning Coalition, in Politics

1. Introduction

Minimal winning coalitions are alignments of parties or politicians just large enough to defeat rivals and no larger. In such coalitions, defection by a single member is sufficient to render the coalition no longer large enough to 'win.' Riker (1962) introduced the idea of minimal winning coalitions in the study of electoral and legislative politics as an alternative to the view expressed in Downs (1957). Downs argued that politicians are primarily office seekers rather than policy makers or allocators of resources. As such, they maximize electoral support and, therefore, forge coalitions as large as possible. Riker's decision makers make authoritative allocation decisions and so seek to minimize the number of claimants on the distribution of resources. A vast literature on coalition formation and government stability has grown out of the debate between Riker and Downs.

The Downsian model indicates that on unidimensional issues and in winner-take-all elections, politicians adopt (usually centrist) policy positions in order to maximize their vote share. Downs's politicians care only about winning office. They do not concern themselves with the policy or private goods concessions they must make to others in order to win.

Riker, in contrast, argued that maximizing votes is costly. Voters are attracted to a candidate by promises about personal benefits. Candidates have preferences of their own about the distribution of scarce resources in the form of private goods to their backers and leftover resources for their own use. To attract the maximum number of votes, politicians must pay a cost by sacrificing some personal interests or granting private side-payments to prospective supporters in an effort to avoid alienating potential voters. Riker argued that rational politicians, motivated primarily by a desire to control resources, seek to attract just enough votes to win and no more, subject to variation above minimal winning size only because of uncertainty about the preferences of voters or their loyalty. By forming minimal winning coalitions, politicians make as few concessions as possible while still controlling sufficient support to maintain governmental authority and pass legislation. Larger than minimal winning coalitions reduce the vulnerability of leaders to defections by individual legislators or small, highly issue-focused parties, but at a price.

2. Policy Preferences and Coalition Formation

In neither Downs's nor Riker's theories are decision makers motivated primarily by policy preferences. Subsequent developments in the coalitions literature refine and challenge elements in the Riker–Downs debate, in part by introducing policy preference as another factor, besides victory or resource maximization, that motivates coalition formation. Axelrod (1970) and De Swaan (1973) argue that policy preferences restrict electoral coalitions to political parties with similar policy agendas. Both maintain that ideologically diverse minimal winning coalitions are less likely to form or survive than are ideologically compact coalitions. Axelrod calls such coalitions minimal-connected coalitions and contends that these arrangements are privileged when prospective coalition partners negotiate with one another. The contention that ideological affinities play a part in selecting minimal winning coalitions is generally supported by the experiences of political parties in, for example, Israel, but is contradicted by the experiences with coalition formation in India.

3. Issue Trading and Multidimensional Issues

The idea that parties or politicians are election oriented, resource oriented, and policy oriented has been developed further. Several scholars extend these ideas by examining coalition formation in multidimensional spaces, that is, in circumstances when policy preferences cannot be portrayed on a single line, but rather reflect greater complexity or linkages across issues. Stokman and Van den Bos (1992) constructed an exchange model that identifies likely vote-trading partners and the trades that can bring them into coalitional alignment. This model and variations on it have been subjected to extensive empirical testing, with impressive results. Others, notably Laver and Schofield (1990) and Laver and Shepsle (1996) have used game theory and spatial models to investigate the theoretical and empirical

basis for arguing that coalition formation is compact in terms of policy differences in parliamentary settings with multidimensional issue spaces. This body of research, also with significant empirical testing, moves the study of minimal winning coalitions beyond the single-issue or unidimensionality envisioned in the original Downs–Riker debate and places the issues in the context of issue trading and logrolling. Furthermore, it moves beyond the cooperative game theory framework proposed by Riker, making use of a mix of insights from noncooperative game theory in which promises are not binding, but are enforced only by self-interest, and from advances in social choice theory. This expanded literature examines how political systems can use institutions or agenda control to circumvent the problems identified in Arrow's Impossibility Theorem. That theorem shows that under specific, generally reasonable conditions of democratic choice, no rule for aggregating preferences exists that can guarantee a positive direct translation of individual preferences into policy outcomes, with the exceptions of unanimity or dictatorship. In analyses of multiple, connected issues, relatively closely shared policy preferences among contending coalition partners is a necessary condition for coalition formation. Especially in Laver and Shepsle, multidimensional issue preferences are used to determine coalition membership and membership is rewarded through policy influence. Policy influence is disseminated, according to them, by using cabinet portfolios as the principal payoff to parties or individuals in parliamentary coalitions.

4. *Coalition Payoffs*

The idea of linking coalition membership to cabinet posts goes back to the earliest formal and empirical examinations of coalition formation. The size of coalitions carries implications regarding the distribution of payoffs or benefits to the members, especially regarding cabinet posts. Research on minimal winning coalitions, however, generally assumes that the benefits distributed to members do not influence their subsequent legislative or electoral competitiveness, that is, the payoffs from joining a successful coalition—including cabinet posts, greater influence over the legislative agenda, control over patronage, etc.—frequently are treated as if they do not translate into resources that subsequently affect the competitiveness of political parties. Browne and Franklin (1973), for instance, argue that such payoffs are distributed in proportion to the party's size in the coalition, and give no consideration to whether payoff allocations influence the future size of parties. Bueno de Mesquita (1978), in contrast, offers evidence that the most valuable cabinet portfolios influence electoral prospects and, therefore, alter the future size of political parties. His analysis shows that party size is endogenous to decisions about coalition payoffs. This suggests that valuable payoffs are highly contested and so are not routinely distributed according to a simple proportionality formula.

5. *Challenges to the Idea of Minimal Winning Coalitions*

Groseclose and Snyder (1996) challenge the foundations of the debate over minimal winning coalitions. They note that virtually all formal models of coalition formation, as well as the related literatures on logrolling and vote-trading, either predict or assume minimal winning coalitions. They, however, suggest a model in which super-majority coalitions are preferred to minimal winning coalitions. In doing so, they endorse Riker's idea that politicians seek to form coalitions that are as cheap as possible. They show, however, that if coalition builders—they call them vote buyers—move sequentially rather than simultaneously, and if the losing vote buyer is always given a last chance to seek defectors from the winning vote buyer's coalition, then minimal winning coalitions will not generally be cheapest for the buyer. Rather, larger than minimal winning coalitions will generally be expected to form in equilibrium. Bueno de Mesquita et al. (2000) build on this work to show the conditions under which super-majority or minimal winning coalitions will form when politicians are motivated by reselection and, secondarily, by a desire to maximize their personal welfare. They show that political systems governed by small coalitions emphasize rent-seeking at the expense of efficient production of public goods, while leaders of systems dependent on larger coalitions shift resources into the provision of public goods and away from private goods allocations.

Strom (1990) also presents a challenge to the minimal winning coalition perspective. He examines minority coalitions, primarily in Europe, and shows that, in seeming contradiction to the equilibrium expectations formed in the literature, they are a common feature of the political landscape and in many cases they endure as the basis of on-going governance. Usually they consist of a single, minority party government that seeks legislative support on an *ad hoc* basis as it moves from one issue to another. Strom's argument relies on political institutions to induce equilibria, in this case equilibria that include minority coalitions. He further supports minority coalition formation by showing a strategic dependence between party behavior in government and subsequent electoral accountability.

The size of winning coalitions proves to be highly variable empirically. Though there is a central tendency to favor nearly minimal winning coalitions, variations in the motivations of leaders seem to tilt them toward or away from minimal winning coalitions. Those who are more concerned with power

and control, rather than policy, tend to pursue larger coalitions, as suggested by Downs. Those with a stronger concern for promoting a policy agenda tend to form coalitions just large enough to win and no larger. Because of this difference, the theory of minimal winning coalitions not only makes predictions about the size of legislative or electoral groupings, but also provides clues for discerning the motives of leaders from their strategies for gaining support to win and hold office.

See also: Agenda-setting; Electoral Systems; Game Theory; Game Theory: Noncooperative Games; Rational Choice in Politics; Voting: Tactical

Bibliography

Axelrod R M 1970 *Conflict of Interest; A Theory of Divergent Goals with Applications to Politics*. Markham, Chicago

Browne E, Franklin M 1973 Aspects of coalition payoffs in parliamentary democracies. *American Political Science Review* **67**: 453–69

Bueno de Mesquita B 1978 Coalition payoffs and electoral performance in European democracies. *Comparative Political Studies* **11**: 61–81

Bueno de Mesquita B, Morrow J D, Siverson R, Smith A 2000 Political institutions, political survival, and policy success. In: Bueno de Mesquita B, Root H L (eds.) *Governing for Prosperity*. Yale University Press, New Haven, CT, pp. 59–84

De Swaan A 1973 *Coalition Theories and Cabinet Formations*. Elsevier, Amsterdam

Downs A 1957 *An Economic Theory of Democracy*. Harper, New York

Groseclose T, Snyder J M Jr 1996 Buying supermajorities. *American Journal of Political Science* **90**: 303–15

Laver M, Schofield N 1990 *Multiparty Government: The Politics of Coalition in Europe*. Oxford University Press, Oxford, UK

Laver M, Shepsle K A 1996 *Making and Breaking Governments: Cabinets and Legislatures in Parliamentary Democracies*. Cambridge University Press, New York

Riker W H 1962 *The Theory of Political Coalitions*. Yale University Press, New Haven, CT

Stokman F N, Van den Bos J 1992 A two-stage model of policy making with an empirical test in the US energy policy domain. In: Moore G, Whitt J A (eds.) *The Political Consequences of Social Networks. Research in Politics and Society*, Vol. 4, JAI Press, Greenwich, CT

Strom K 1990 *Minority Government and Majority Rule*. Cambridge University Press, New York

B. Bueno de Mesquita

Minnesota Multiphasic Personality Inventory (MMPI)

The MMPI, the Minnesota Multiphasic Personality Inventory, was developed in the 1940s as a means of evaluating mental health problems in psychiatric and medical settings. The test authors, Starke Hathaway and J. C. McKinley, thought that it was important in evaluating patients' problems to ask them about what they felt and thought. Their instrument was a self-report inventory that included a very broad range of problems and could be answered with a sixth-grade reading level. The MMPI was developed according to rigorous empirical research methods and rapidly became the standard personality instrument in clinical settings (Hathaway and McKinley 1940). The popularity of the true-false personality inventory was in large part due to its easy-to-use format and to the fact that the scales have well-established validity in assessing clinical symptoms and syndromes (Butcher 1999). The MMPI underwent a major revision in the 1980s resulting in two forms of the test—an adult version, the MMPI-2 (Butcher et al. 1989) and an adolescent form, MMPI-A, (Butcher et al. 1992). The MMPI-2 is the most widely researched instrument and is used for the evaluation of clinical problems in a broad range of settings including mental health, health psychology, correctional settings, and personnel screening, and in many forensic applications such as child custody and personal injury (Lees-Haley et al. 1996, Piotrowski and Keller 1992).

The MMPI-2 contains 567 true–false questions addressing mental health symptoms, beliefs, and attitudes. The items on the MMPI-2 are grouped into scales (clusters of items) that address specific clinical problems such as depression or anxiety. After the inventory is completed, the items are scored or grouped according to the scales that have been developed. An MMPI scale allows the clinician to compare the responses of the client with those of thousands of other people. Initially, the scores are compared to the normative sample, a large representative sample of people from across the USA, in order to determine if the person's responses are different from people who do not have mental health problems. If the person obtains scores in the extreme ranges, for example on the depression scale, compared with the normative sample then they are likely to be experiencing problems comparable to the clinical samples of depressed clients that have been studied.

1. Evaluating Cooperative Responding

In some situations clients might be motivated to present personality characteristics and problems in ways that are different than they actually are. For example, if people are being tested to determine whether or not they are 'sane' enough to stand trial in a criminal court case they might attempt to exaggerate symptoms or problems in order to avoid responsibility. Alternatively, people being evaluated in pre-employment psychological evaluation might be inclined to present themselves in an extremely positive way to cover up problems. It is important, in MMPI-2 profile interpretation, to evaluate the way in which

people approach the task of self-revelation. Are they sufficiently cooperative with the testing to produce a valid result?

There are several indices on the MMPI-2 to address honesty and cooperativeness in responding to the items (Baer et al. 1995):

1.1 Cannot Say Score

This index is simply the total number of unanswered items. If the client leaves out many items on the test (e.g., 30 items), the profile may be invalidated. A high score on this index suggests that clients have not cooperated sufficiently to provide a clear picture of their personality functioning.

1.2 The L Scale

The L or Lie scale is a measure of the client's willingness to acknowledge personal faults or problems. Individuals who score high on this scale are presenting an overly virtuous picture of themselves. They claim virtue and positive characteristics that are not typically found among people in general. The L scale is particularly valuable in situations where the individual has something to gain by not appearing to have problems, such as in personnel screening or in child-custody disputes. In these cases, people might try to put their best foot forward and present themselves as 'better' adjusted than they actually are.

1.3 The K Scale

The K scale was developed as a means of evaluating the tendency that some people have to minimize problems. This scale is also used as an index for correcting some clinical scales that have been shown to be affected by defensiveness; that is, it serves as a correction factor to compensate for the tendency of some people to deny problems. Empirically derived percentages of the K scale score are added to five of the clinical scales in order to improve test discrimination.

1.4 The Infrequency Scales

The MMPI-2 contains three measures to assess the tendency of some people to exaggerate their problems or 'fake' the test by over-responding to extreme items (the F or Infrequency scale, the Back F or F(B) scale, and the P(p) scale). Each scale performs a somewhat different function in interpretation. The items on these scales are very rare or bizarre symptoms. Individuals who endorse a lot of these items tend to exaggerate symptoms on the MMPI-2, perhaps as a way of trying to convince professionals that they need to have psychological services. Some people who have a need to claim problems in order to influence court decisions will tend to elevate the infrequency scales. The infrequency scales can be elevated for several possible reasons: the profile could be invalid because the client became confused or disoriented or responded in a random manner. High F and F(B) scores are commonly found among clients who are malingering or producing exaggerated responses in order falsely to claim mental illness (Graham et al. 1991).

1.5 TRIN AND VRIN scales

Two scales to assess inconsistent responding have been developed for the MMPI-2. These scales were derived by examination of the individual's responses to pairs of items that have similar content. A particular response to one of the items in the pair requires that the other item be endorsed in a given way in order to be consistent: for example, a pair of items that contains content that cannot logically be answered in the same direction if the subject is responding consistently to the content.

2. Profile Interpretation

The assessment of clinical problems is approached by examining the client's responses to the items using several strategies. We will examine three types of scales that comprise the MMPI-2's problem measures: the traditional clinical scales and profile codes, the MMPI-2 content scales, and the supplemental or specific-problems scales.

2.1 The MMPI-2 Clinical Scales

Hathaway and McKinley developed the original MMPI clinical scales to classify empirically or group patients into problem types. For example, they developed scales to assess hypochondriasis (the Hs scale), depression (the D scale), hysteria (the Hy scale), psychopathic deviation (the Pd scale), paranoid thinking (the Pa scale), psychasthenia (the Pt scale), schizophrenia (the Sc scale), and mania (the Ma scale). In addition, two other scales were included on the basic profile to address problems of sex-role identification (the Mf scale) and social introversion and extraversion (the Si scale). These scales have clearly defined correlates or empirical relationships. If a scale is elevated in the significant range, the individual is considered to possess the attributes measured by the scale.

Some clients have only one of the clinical scales elevated in their profile. In these cases, the profile interpretation is straightforward—the clinician simply examines the published correlates for that scale and

incorporates these into the psychological report. However, some clients might have several of the clinical scales elevated in their profile. The full pattern or profile configuration needs to be interpreted. Fortunately, extensive research has been undertaken to evaluate complex profiles or code types. These profile types result from having two or more of the clinical scales elevated in the interpretive range. In these cases, the practitioner refers to published 'code books' which provide empirical descriptions for the various patterns that have been studied.

2.2 Content-based Scales

The MMPI-2 contains a number of scales that assesses content themes that the individual endorsed in the item pool. The content scales are homogeneous item clusters that summarize themes and represent direct communications about problems to the practitioner. There are 15 content scales measuring different symptom areas and problems: for example, Antisocial Practices (ASP), Bizarre Mentation (BIZ), and Low Self Esteem (LSE).

2.3 Special-problem or Supplemental Scales

Several supplemental scales have been developed to assess specific problems such as the potential to develop problems of addiction (the MacAndrew Addiction scale or MAC-R and the Addiction Potential scale or APS) and whether or not the individual acknowledges having problems with drugs or alcohol. Another scale, the Marital Distress scale, addressed a client's attitudes toward their marital relationship. The MMPI-2 supplemental scales provide the practitioner with the means of evaluating specific problems (such as post-traumatic stress disorder) that are not addressed in the clinical or content scales.

3. Applications of the MMPI-2

Contemporary uses of the MMPI-2 include such activities as evaluating clients following their admission to an inpatient psychiatric facility; providing personality information about client's problems and attitudes in the early stages of psychological treatment to assess treatment amenability; providing personality information for therapists to employ in giving the client feedback in psychotherapy; and providing information that would aid the clinician in arriving at an appropriate clinical diagnosis. The MMPI-2 is also widely used in medical settings to provide psychological adjustment information on patients—for example, in rehabilitation programs as a measure of behavior problems and symptoms following a stroke as part of the neuropsychological evaluation; or in the evaluation of drug treatment effects.

A great deal of contemporary personality assessment research is also aimed at further exploring the MMPI-2 test variables. The two most active research areas involve further exploration into the external validity of the MMPI-2 scales, and the translation and adaptation of the instrument for other languages and cultures. For example, research carried out in the mid-1990s has further explored the validity of the test in diverse settings such as personnel screening (Butcher 1995); forensic assessment (Megargee 1997); and test correlates with outpatients (Graham et al. 2000). Several studies have explored the use of MMPI-2-based personality descriptions in providing test feedback to psychotherapy clients. This research has shown that providing psychological test feedback in treatment increases the effectiveness of treatment (Finn 1996, Finn and Tonsager 1992, Newman and Greenway 1997).

The MMPI-2 is also widely used as a cross-cultural research technique to study psychopathology across cultures. The MMPI-2 items and scales have shown remarkable resilience when used in other languages and cultures (Butcher 1996), and computer-based interpretation of the MMPI-2 has been shown to be highly effective in describing clinical patients in other cultures (Butcher et al. 1998).

4. Conclusion

The MMPI-2 is the most widely used and most versatile clinical personality instrument that clinical psychologists have available. It is a self-report personality questionnaire that provides the test user with scores on a number of scales that address important clinical problem areas. In addition the MMPI-2 provides mental health professionals with information that allows them to appraise the client's willingness and frankness in the appraisal—key information about the client's response attitudes that allows for a determination of the credibility of the client's self-report. Once cooperation in the assessment is assured, the MMPI-2 clinical measures can be effective in delineating the mental health symptoms, personality traits, and special problems that the client is likely to be experiencing.

See also: Personality Assessment; Personality Theories

Bibliography

Baer R A, Wetter M W, Nichols D S, Greene R, Berry D T R 1995 Sensitivity of MMPI-2 validity scales to underreporting of symptoms. *Psychological Assessment* **7**: 419–23

Butcher J N 1995 *User's Guide for The Minnesota Report: Revised Personnel Report*. National Computer Systems, Minneapolis, MN

Butcher J N 1996 *International Adaptations of the MMPI-2*. University of Minnesota Press, Minneapolis, MN

Butcher J N 1999 *The MMPI-2: A Beginners Guide*. American Psychological Association, Washington, DC
Butcher J N, Berah E, Ellertsen B, Miach P, Lim J, Nezami E, Pancheri P, Derksen J, Almagor M 1998 Objective personality assessment: Computer-based MMPI-2 interpretation in international clinical settings. In: Belar C (eds.) *Comprehensive Clinical Psychology: Sociocultural and Individual Differences*. National Computer Systems, Minneapolis, MN, pp. 277–312
Butcher J N, Graham J R, Dahlstrom W G, Tellegen A M, Kaemmer B 1989 *Minnesota Multiphasic Personality Inventory-2 (MMPT-2): Manual for Administration and Scoring*. University of Minnesota Press, Minneapolis, MN
Butcher J N, Williams C L, Graham J R, Tellegen A, Ben-Porath Y S, Archer R P, Kaemmer B 1992 *Manual for Administration, Scoring, and Interpretation of the Minnesota Multiphasic Personality Inventory for Adolescents: MMPI-A*. University of Minnesota Press, Minneapolis, MN
Finn S E 1966 Assessing feedback integrating MMPI-2 and Rorschach findings. *Journal of Personality Assessment* **67**: 543–57
Finn S E, Tonsager M 1992 Therapeutic effects of providing MMPI-2 test feedback to college students awaiting therapy. *Psychological Assessment* **4**: 278–87
Graham J R, Ben-Porath Y S, McNulty J 2000 *Using the MMPI-2 in Outpatient Mental Health Settings*. University of Minnesota Press, Minneapolis, MN
Graham J R, Watts D, Timbrook R 1991 Detecting fake-good and fake-bad MMPI-2 profiles. *Journal of Personality Assessment* **57**: 264–77
Hathaway S R, McKinley J C 1940 A multiphasic personality schedule (Minnesota): 1. Construction of the schedule. *Journal of Psychology* **10**: 249–54
Lees-Haley P R, Smith H H, Williams C W, Dunn J T 1996 Forensic neuropsychological test usage: An empirical survey. *Archives of Clinical Neuropsychology* **11**: 45–51
Megargee E I 1977 Using the Megargee MMPI-based classification system with the MMPI-2s of female prison inmates. *Psychological Assessment* **9**: 75–82
Newman M L, Greenway P 1997 Therapeutic effects of providing MMPI-2 test feedback to clients at a university counseling service: A collaborative approach. *Psychological Assessment* **9**: 122–31
Piotrowski C, Keller J W 1992 Psychological testing in applied settings: A literature review from 1982–1992. *Journal of Training and Practice in Professional Psychology* **6**: 74–82

J. N. Butcher

Minorities

A minority group is usually defined as a group of people with common interests or characteristics which distinguish them from the more numerous majority of the population of which they form a part or with whom they live in close proximity within a common political jurisdiction. Minorities may be composed of people with common interests in a range of activities such as sports, environmental issues, professional occupations, and leisure pursuits. The term 'minority interest' is often used to suggest that a particular activity attracts the interest of less than a majority of the relevant population.

1. *Social Science Definition*

In social science and more widely in public discourse, the term has come to mean groups that have a minority status that is relatively long-lasting or permanent and is based on ethnic or cultural attributes. 'Minority' is thus a term used to define people with distinct ethnic, national, religious, or linguistic affiliations. People belonging to minorities are thus often recognizable by their dress, language, place of residence, special cultural practices, or ethnicity. Minority status may be imposed on a group of people by majority prejudice and discrimination. Thus gypsies are defined as minorities in many European countries by majorities who disapprove of their lifestyle and modes of behavior.

2. *Minorities and Power*

An important characteristic of minorities is that they are assumed to have less power than the majority group. The term minority thus not only implies a numerical minority but political powerlessness as well. Minorities are assumed to be weak and inferior, often discriminated against by the majority, and excluded from sharing political power. The description of a group as a minority thus reflects their lack of power and inferior status as much or more than their numerical status. A majority group, by contrast, is assumed to be politically, economically, and ideologically dominant, and references to it as a majority reflect its political ascendancy. The majority may use its numerical and political superiority to try to eliminate the minority by forcing its members to adopt the language, religion, and customs of the majority. However, such pressures usually have the reverse effect, in strengthening the identity of the minority, its internal cohesion, and its pride in its cultural traditions. Oppressed minorities are capable of fierce and prolonged resistance to majority pressure as can be seen by the long civil war in the Sudan and the nationalist campaign in Northern Ireland.

3. *Minorities and Democracy*

In democracies, minorities are often considered to deserve special protection against the arbitrary actions of the majority. This is partly a recognition that minorities should be allowed to contribute to the policy-making process if they are to be subject to its

decisions and accept them as legitimate. However, it is also a recognition that the oppression of minorities is against liberal democratic traditions and a violation of human rights. It is also a matter of realpolitik, as minorities that are pushed too far can resist violently and impose huge costs on the majority for their oppression. Thus in many countries minorities are given special status and are accorded special institutional arrangements and legislation to combat their social, economic, and political exclusion. These arrangements may involve federalism, rights entrenched in a constitution, local autonomy, or anti-discrimination legislation.

4. *Minorities in Power*

This assumption of minority powerlessness may not always be correct. History is full of examples of minorities ruling majorities: European colonial governments ruled overseas territories in Africa, Asia, and the Americas; minority white governments exercised power in Rhodesia, as did the apartheid regime in South Africa; and the Tutsi minority rules the majority Hutus in contemporary Rwanda and Burundi. However, even when minorities are in power, this is considered in our democratic age to be illegitimate, fragile, and vulnerable to overthrow by the majority.

5. *Reseatment Towards Successful Minorities*

Political powerlessness may not mean exclusion in all areas of social activity. Sometimes minorities occupy important niches: for example, the Chinese in Indonesia and Malaysia have powerful roles in the economy. Successful minorities who occupy important and prominent positions are often the focus of majority resentment and frustrations and may become the targets for political attacks and violence. The Chinese in Indonesia have periodically been the victims of savage anti-Chinese riots, while the appalling example of the Holocaust is a constant reminder of how a successful and apparently well-integrated minority can become the victim of extreme racist scapegoating and stereotyping leading to genocide.

6. *Women and Minority Status*

The position of women provides an interesting example of the confusion between the numerical and political dimensions of the term, and shows how a majority group can be redefined as a minority. In most societies women form a majority of the population but they are systematically excluded from the most power ful roles in society. Women form a minority, often a small minority, of politicians, judges, industrialists, bankers, administrators, church leaders, and officers in the armed forces. This exclusion of women from the powerful positions in society means that they appear to be a minority even though they form an overall numerical majority. The exclusion of women from positions of power is fundamental to their construction as a minority.

7. *Types of Minorities*

In the contemporary world there are two types of minorities that are critically important in the government of modern states. These are indigenous, territorially based minorities, often known as 'national minorities' and 'transnational minorities,' created by the processes of international migration.

7.1 *Indigenous Minorities*

Indigenous, territorially based minorities can also be of two types. The first are often described as 'first nations.' These are the indigenous, aboriginal, peoples who were often overwhelmed by later migrations. Thus in North and South America and in Oceania, European migration in the eighteenth and nineteenth centuries caused the original indigenous peoples to be defeated, displaced, and almost exterminated. In recent decades, first nations have had considerable success in asserting their claims to be regarded as nations and gaining recompense for broken treaties signed with their ancestors. Canada in particular tried to redefine its relationship with its native peoples in the *Royal Commission Report on Aboriginal Peoples in Canada* (1996). The report advocated a new partnership between native peoples and Canada, with Canada recognizing their right to 'nation' status and autonomy. It recommended that native peoples be treated as partners at federal, provincial, and local levels. Treaty agreements have been signed at provincial level as well as federal level, for example by the government of British Columbia granting native peoples land rights and political autonomy. In 1999 the Nunavut area became a territory and the first part of Canada to be governed by one of Canada's 'first nations,' in this case the Inuit.

In New Zealand the Maori have raised issues regarding their community rights under the Treaty of Waitangi and these demands have been recognized by the New Zealand government. This has resulted in the significant transfer of resources such as land, control over fisheries, and the promotion of their language. There are five Maori seats out of a total of 120 in the New Zealand House of Representatives; these are elected from a separate Maori roll. Maori can choose to register on either the Maori or general electoral rolls. The change from a single-member, simple majority electoral system to a form of proportional representation in 1996 has significantly increased

minority representation in the New Zealand parliament.

In Australia and Brazil, the campaigns on behalf of aboriginal people are finding it much more difficult to get government recognition and support and much less progress has been made in recognizing their rights and protecting their interests.

7.2 National Minorities

In Europe, the term 'national minorities' is used to refer to territorially based minorities such as the Basques and Catalans in Spain, the Scots or Welsh in the UK, or the Bretons and Corsicans in France. These minorities were created during the processes of state-building in the eighteenth and nineteenth centuries as core populations such as the Castilians, English, and French expanded their political sovereignty and incorporated neighboring territories inhabited by different ethnic groups, or by previously autonomous national groups sometimes enjoying their own constitutions and laws as in the cases of the Catalans and Scots.

7.2.1 Managing national minorities. Territorial minorities can present major problems of statecraft for the countries into which they are incorporated. They may have close affiliations to people in neighboring states and therefore can be a source of claims for secession or international conflict. Alleged ill-treatment of ethnic Germans in Poland, whose rights were protected by bilateral treaties, was the pretext for Hitler's invasion of Poland in 1939 and the start of World War II. Similar accusations of Czech mistreatment of the Sudeten Germans the previous year had been used to justify German intervention in Czechoslovakia in 1938. There are numerous examples of territorial minorities which overlap borders in Europe, such as the Hungarian minority in Romania, the German minorities in Denmark and the Netherlands, and the German-speaking Austrian minority in the Alto-Adige region of Italy. Africa provides as many examples of such territorial minorities as Europe, owing to the arbitrary nature of the boundaries drawn by the colonial powers, which split tribal and linguistic groups with impunity. Thus there is a Somali minority in Kenya and Fulani minorities in Nigeria, Cameroon, and Niger. Some national minorities are well integrated and identify strongly with the nation-states in which they reside, for example, the French- and Italian-speaking Swiss or the Swedish-speaking Finns. Other national minorities resist their incorporation and campaign strongly for independence or union with another state; examples of these are the Basques in Spain, or the nationalists in Northern Ireland who wish to leave the UK and join the Republic of Ireland.

7.2.2 Claims for independence. The problem for multinational states is that if territorially based minorities feel oppressed and alienated, then they can easily challenge the legitimacy of their continued incorporation, generate nationalist claims, and demand independence. This is less true, of course, for first nations who tend to be very small minorities in the countries where they reside and whose objectives tend to be limited to demands for cultural recognition, limited autonomy, and control over traditional lands and other resources. National minorities represent the most serious challenge. There are many states that have been created through secession from another state, for example, Ireland, Norway, and Finland. The collapse of multinational empires such as the Austro-Hungarian, Ottoman, and Soviet empires left many nation-states in their wake. The collapse of the former Republic of Yugoslavia in 1991 led to the creation of five new states. Multinational states would thus seem to have a high potential for fragmentation into their constituent parts. This explains why territorial minorities are often granted special rights of autonomy, language and national representation in order to encourage their acceptance of the legitimacy of existing political arrangements. This process of acceding to minority demands represents a shift in the politics of multinational states, as such rights were usually lost during the earlier processes of nation-building. At the beginning of the twenty-first century, the unique position of Quebec in Canada is a good example of what is sometimes called asymmetrical federalism, where one province has special rights, for example, over language and immigration, that are not accorded to the other provinces.

7.2.3 National minority nights. Bauböck (1994) argues that there are three kinds of reasons for respecting national minority rights. First, there are those grounded in historical agreements that promised autonomy as a condition for voluntary integration. These apply to many first nations, and countries such as Canada, New Zealand, and the USA are renegotiating treaties with first nations and compensating them for treaties that were broken. Secondly, minorities may have strong claims for rights owing to persistent discrimination and disadvantage resulting from the dominance of a hegemonic majority. The problem here is that suppression of rights is often fiercest when the majority group fears territorial secession by the minority. The systematic suppression of the Irish nationalists in Northern Ireland, of the Kurds in Iraq and Turkey, and of groups such as the Chechens in the Russian Federation, are based on fears that these minority groups will create their own states or join neighboring states, thus reducing the national territory. States tend to regard their territorial integrity as an inviolable prin-

ciple and will use extreme measures to maintain this. Finally, Bauböck argues, liberal democracies should recognize the general value of secure membership in cultural and political communities for individual autonomy and well-being. This might mean granting local autonomy, protecting the local language and religion, and encouraging investment in minority areas. These policies still might not reduce the local demands for complete independence but might well make them less intense.

8. *International Migration*

The second major process that has led to the creation of ethnic and cultural minorities has been international migration. In the nineteenth century, migration was largely a process of European settlement of colonial or former colonial territories in America, Africa, Asia, and the Pacific. European settlers established themselves in particular in North and South America, Southern Africa, Australia, and New Zealand. In many of these places, Europeans settled as majorities, but in others as minorities. There was also some migration of subject peoples at the behest of colonial powers so that indentured Indians, for example, were recruited by the British authorities to work on the sugar plantations in Natal, Fiji, and in the Caribbean. This recruitment led to the establishment of thriving Indian communities in Guyana, Trinidad, South Africa, and Fiji. In the twentieth century, Indians were recruited to build the railways in East Africa and this led to the development of Asian communities in Kenya, Uganda, Tanzania, Malawi, and Zambia. In Fiji, the establishment of a large and successful Indian community has led to considerable tensions between native Fijians who control the land and politics and the economically successful Indians who, in 1946, became a majority of the population. In 1987, after a native Fijian coup, thousands of Indo-Fijians left the country, leaving native Fijians in the majority once more. The new constitution entrenched the rights of native Fijians, reducing Indo-Fijians to second class status.

8.1 *Economic Migration*

In Western Europe, immigration has been of considerable importance since World War II. Some 30 million people from all over the world have moved to Western Europe to find work and settle in these secure and prosperous economies. A great deal of migration has been internal to Europe, with people moving from Eastern and Southern Europe to Germany, France, and the UK, but large numbers of Turks, Africans, Asians, and Caribbeans have also been in involved in this migration. Germany has attracted the largest numbers of immigrants and in 1999 had a foreign population of nearly seven and a half million, over 9 percent of the population of reunited Germany. France has a foreign population of 3.6 million, many of whom are Portuguese, Algerian, and Moroccan, but also Africans from francophone West Africa. Britain's immigrants came initially from Ireland and then from the Caribbean and the Indian subcontinent. In 1999 Britain had an ethnic minority population of 3.7 million, some 6.3 percent of the total UK population.

8.2 *Including Minorities*

Immigrant communities established through international migration pose very different issues of political management to those posed by indigenous territorial minorities. Immigrants make few political demands on their new countries until they have established substantial communities. Their demands are likely to be for equal treatment, recognition of cultural values and activities, and access to citizenship. They rarely occupy specific territories. They therefore have no claims to be nations. They tend to settle in cities, to have diverse backgrounds, and usually have limited political rights at first, though extensive social and economic rights. They may be welcomed for their skills, labor, or investment, but initially are regarded as newcomers who must develop close ties with their new country before they can be admitted as citizens. Only if they have ethnic or colonial connections with their host country, as ethnic Germans coming to Germany from Eastern Europe or Commonwealth citizens coming to the UK, do they have citizenship rights from the time of settlement.

8.3 *The Challenge of Migration*

Postwar migration has contributed to the creation of a greater diversity of peoples in Western Europe, but also to concerns about discrimination, disadvantage, and racial violence. The settlement of large numbers of non-European immigrants was unexpected, as few believed that the huge level of economic growth and prosperity in Europe that developed after World War II would be so sustained, creating such a massive demand for both skilled and unskilled labor. Some countries such as France, the Netherlands, and Sweden welcomed immigrants and encouraged them to settle and to naturalise. Other countries, such as the UK and Germany, were more cautious. The popular response to immigration was neutral during periods of rapid economic growth, but more hostile in periods of recession. In the 1970s there was strong support for immigration controls in the UK, led by the populist politician Enoch Powell. In the 1980s the anti-immigrant *Front National* gained significant popular support in France, which was sustained for two decades. In Austria in the 1990s, the Austrian Freedom Party, another anti-immigrant party, gained 27 percent of the vote and a place in a coalition government.

8.4 Integrating Minorities

Even countries which were founded by immigrants and consider themselves as countries of immigration, such as the USA and Canada, are concerned about the integration of immigrants and their willingness to become loyal citizens of their new country. Traditional assumptions that immigrants would assimilate into the US melting pot and disappear into the general population have proved outdated as minorities continue to exist and thrive across the generations. Schlesinger (1992) has even warned that the huge levels of immigration and the creation of new minorities could lead to the 'disuniting of America' (by which he means the USA). Similar concerns are often raised in Europe. Will new minorities create their own identities, loyalties, and institutions separate from those of the majority?

The evidence is mixed. It is clear that immigrants place a high priority on learning the language of their new country of residence and encouraging their children to learn the language and integrate into the education and labor markets. Naturally immigrants wish to preserve their customs, religion, and traditions, but a high degree of integration quickly takes place if it is allowed to do so. Attempts at forcible integration of new minorities usually result in resistance and hostility by the minority towards the majority. Schlesinger's concerns are thus exaggerated as there is plenty of evidence of the rapid integration of new immigrant minorities and high rates of their naturalization. Long-settled minorities such as African Americans can face much greater problems of acceptance and integration due to their historical legacy of slavery and oppression. Sustained discrimination and racism against a minority will increase its feelings of solidarity, alienation, and frustration. This may result in violent explosions as happened in Watts in Los Angeles in 1968 and in Brixton in London in 1981. A positive welcome to minorities and a recognition of their economic, cultural, and political contribution to their new society is the best strategy for fostering equality, justice, and unity.

See also: Assimilation of Immigrants; Diaspora; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Race Identity; Racial Relations; Racism, History of; Racism, Sociology of; Refugees: Geographical Aspects; Refugees in Anthropology

Bibliography

Bauböck R 1994 *Transnational Citizenship: Membership and Rights in International Migration*. Edward Elgar, Aldershot, UK

Bauböck R 2000 Political community beyond the sovereign state: Supranational federalism and transnational minorities. Paper presented at the Conceptualizing Cosmopolitanism Conference, University of Warwick, April 2000

Bosniak L 2000 Citizenship denationalised. *Indiana Journal of Global Legal Studies* **7** (Spring): 447–508

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Guibernau M 1999 *Nations Without States: Political Communities in a Global Age*. Polity Press, Oxford, UK

Kondo A (ed.) 2001 *Citizenship in a Global World*. Palgrave, Basingstoke, UK

Kymlicka W 1995 *Multicultural Citizenship*. Clarendon Press, Oxford, UK

Royal Commission Report on Aboriginal Peoples in Canada Minister of Supply and Services (1996) Ottawa, Canada. Communications Group Publishing, Canada

Rose A, Rose C B (eds.) 1965 *Minority Problems*. Harper, New York

Schlesinger A D 1992 *The Disuniting of America: Reflections on a Multicultural Society*. Norton, New York

Spoonley P 1993 *Racism and Ethnicity*. Oxford University Press, Auckland, New Zealand

Z. Layton-Henry

Minority Access to Health Services: United States

Ethnic minority populations comprise large segments of the US population and are growing especially rapidly. Their use of healthcare services is relatively low. Certain factors accounting for racial and ethnic disparities are well understood but others remain to be determined. Identifying and overcoming barriers to healthcare is important to insure an equitable distribution of healthcare related resources in order to promote equal well-being in all sectors of the population. Problems in minority access to healthcare are evident in other developed countries as well, though not as pervasively as in the United States.

1. Ethnic Minority Populations: Size, Growth, and Risk

Black Americans make up at least 12.5 percent of the United States population, not allowing for a probable census undercount (O'Hare et al. 1991). The black population is relatively young and represented only about 8 percent of the 65 and over population in 1995, because of a persistently higher mortality rate that reverses only in very old age ('crossover effect').

Persons of Hispanic heritage made up about 11 percent of the United States population in 1996 (del Pinal and Singer 1997). The Hispanic population consists of a number of diverse groups, including Cuban Americans, Puerto Ricans, persons from Central and South American countries, and, by far the

largest group, comprising almost two-thirds of the Hispanic population, Mexican Americans. The Hispanic population is growing especially rapidly due to immigration and differential birth rates and a high proportion of children and youth. According to current projections (del Pinal and Singer 1997), Hispanics by 2050 will comprise about 24 percent of the United States population.

In 1997 about 4 percent of the US population was Asian American (Lee 1998), up from 3 percent in 1990 and 2 percent in 1980. Asian Americans make up about 2 percent of persons 65 years and older, and are projected to make up 7 percent in 2050 (Treas 1995). The increase reflects a shift in immigration policy in 1965 toward the elimination of national quotas, and the arrival of successive waves of refugees from the Vietnam war (Lee 1998). American Indians and Alaska Natives are made up of more than 550 tribes recognized by the federal government and Alaska Native Villages, the largest of which are the Navajo and Cherokee. In 1990 there were almost 2 million American Indians and Alaska Natives, more than twice as many as were counted in the mid-1970s (Norton in press). They comprise about 0.4 percent of persons 65 years of age and older and are expected to increase to a proportion of 0.6 percent by 2050 (Treas 1995).

2. *Socioeconomic Distress*

Among the most important differences between racial and ethnic groups are those in education, occupational status, and income—that is, in socioeconomic standing (Krieger et al. 1997). Poverty is especially influential. Over a number of years, African and Native American families have been roughly three times as likely as white families to have incomes placing them below the federally established poverty line. The disparity is even greater when considering extreme poverty—family incomes at a level less than half of the poverty threshold—and is also large when considering children and the elderly. The effects of poverty are compounded by differences in total value of accumulated assets, or total wealth (O'Hare et al. 1991).

Many minority poor live among other people who are poor in areas in which such individuals and families are found in high concentrations. The neighborhoods tend to be distressed, marked by high rates of unemployment, homelessness, crime, and substance abuse, with few local resources to help offset these problems (Wilson 1987).

3. *Access to Healthcare and Health Services*

Minority access to and patterns of receiving healthcare are problematic. Levels and occasions of use have been best documented among African Americans. Despite relatively poor health status and higher rates of illness and mortality, African Americans make fewer visits for ambulatory care than whites. Moreover, their visits are considerably more likely than those of whites to be made to hospital outpatient and emergency departments than to physicians' offices (National Center for Health Statistics 1996).

Excess emergency treatment of minorities appears to be more selective in other developed countries than in the United States. Investigators studying the West Midlands in the United Kingdom (Gilthorpe et al. 1998), for example, found that Black children and older Indian and Bangladeshi men and women were more likely than Whites to be admitted to hospital on an emergency basis with asthma-related suffering. The difference occurred despite a lack of evidence indicating underlying racial and ethnic differences in the prevalence of asthma. Disproportionate emergency care is not pervasive, however.

Latinos and Asian Americans enjoy better health status when considering indicators of morbidity and mortality, although on overall ratings of health status, Latinos were more likely to indicate 'fair/poor' health and less likely to indicate 'excellent' health. African Americans in the study closely resembled Latinos in pattern of perceived health status (Weigers and Drilea 1996). Despite considering themselves less healthy than Whites, however, Latinos and African Americans have proven less likely to make use of ambulatory medical care (National Center for Health Statistics 1996).

Most groups receive healthcare not only at relatively low levels but also in less than optimal ways. Problems are especially evident in receiving routine and preventive care. When pregnant, women from almost all minority groups are less likely than whites to receive prenatal care during the first trimester (National Center for Health Statistics 1996).

Racial and ethnic disparities such as these are less common in other developed countries than in the United States. A study conducted in Amsterdam (Reijneveld 1998) found that in keeping with relatively poor health, immigrants from Turkey and Morocco made relatively heavy use of primary care and other health services. Thus, the pattern of minority–white differences is opposite that found in the United States. In another study, researchers investigating birth records from three hospitals in East London (Collins et al. 1997) compared the pregnancies of African, West Indian, and European White women. West Indian and African women were more likely than European White women to have delivered low birthweight infants, but they had not received less prenatal care.

Related to these problems is the lack for many minorities in the United States of a continuing, single point of access—a usual source of healthcare. In 1996, 76.3 percent of whites had an office-based usual source of care, whereas only 63.6 percent of African

Americans and 57.9 percent of Latinos had a comparable usual source. The disparities are greater still when considering children under age 18.

4. Factors Affecting Minority Access

To a notable extent, problems in access can be attributed to a lack of health insurance coverage. African Americans and Latinos are notably less likely than others to have health insurance. Disparities are found both among males and females and are greatest among children and young adults (National Center for Health Statistics 1996).

Because in the United States health insurance often is provided as a benefit of employment, differences between racial and ethnic groups in insurance coverage are tied to differences in patterns of labor force participation. Minority persons are more likely to be unemployed and, if employed, to work in jobs that fail to provide health insurance. The problem is particularly acute among Hispanic males, only 49.7 percent of whom had job-related healthcare coverage in 1996.

Differences in rates of insurance coverage would be greater still but for widespread availability and use of publicly-supported insurance programs, especially Medicaid. Medicaid is a federal-state program supporting healthcare to the poor. In keeping with their higher rates of poverty, minorities are more likely than whites to be covered by Medicaid; blacks and Hispanics are more than twice as likely as whites to be covered (Banthin and Cohen 1999). Medicaid benefits vary from state to state and in some areas they provide only minimal coverage.

Also, many private office-based healthcare providers do not accept Medicaid as a source of payment and are unavailable to minority and other low-income patients. Moreover, minorities are over-represented in many inner city and rural areas in which office-based medical providers are found less often than in other areas perceived to be more hospitable (Perloff et al. 1997). In this manner, geography and financing help to dictate minority under-representation in office-based practice as a usual source of care.

Minority beliefs and attitudes toward illness and toward seeking healthcare have been proposed as another explanation for minority–white differences in healthcare. Conceivably, some minority communities believe that healthcare is less desirable or necessary than whites do, or exhibit a greater preference for religious or holistic alternatives to scientific medicine. Put another way, competing 'explanatory models' (Kleinman 1978) of health and illness might lead minority group members away from mainstream healthcare providers and toward alternative sources of help. However, Snowden et al. (1997) found that blacks and whites differed on few attitudes toward the importance of insurance coverage and use of healthcare, and that the attitude differences failed to explain substantial differences in receiving outpatient medical care, although the study did not touch on possible underlying differences in how symptoms of illness are recognized and interpreted in the first place.

The attitudes of healthcare providers and their sensitivity to ethnic and cultural differences have been invoked to explain differences in healthcare. Critics have stressed the existence of mistrust and a lack of comprehension on the part of some healthcare providers to beliefs, behavior patterns, and linguistic patterns prevalent in many minority communities and the mutual alienation of patients and providers which can result (e.g., Kavanagh and Kennedy 1992). They have called for a 'culturally competent' healthcare practice which would seek to overcome social distance as a barrier.

Unlike members of other minority groups, many Native Americans are eligible for a government-sponsored system of healthcare earmarked for their use. The Indian Health Service (HIS) and Urban Indian Health Programs offer various forms of healthcare, although the greater part of the effort is directed toward primary care. In recent years the budgetary limitations have caused the HIS to curtail its offerings (Manson 1998). In moving toward self-determination, many tribes have assumed direct control of local programs, including those providing healthcare. We know little about levels and types of care provided under any of these arrangements (Manson 1998).

5. Future Directions

Although the role of financing as a barrier to equal provision of healthcare is well established, the situation demands a nuanced and comprehensive understanding. Puzzles remain. When asked whether they had experienced any difficulty or delay in obtaining healthcare, more Hispanics than whites indicated the existence of such a problem, but fewer African Americans so indicated (Weinick et al. 1997). Among persons experiencing a problem, many more Hispanics than whites reported an inability to afford care, but only slightly more African Americans so indicated. Other concerns, including location and receptivity of providers, and knowledge and willingness to seek care of patients, must be taken into account.

There is also a need for greater understanding of exactly which members of ethnic minority populations are unlikely to seek healthcare. Greater reluctance to use the health care system might be linked not only to socioeconomic standing, but also to factors such as country of birth and language preference—many minorities are immigrants and refugees—as well as adherence to traditional cultural beliefs and practices.

More effort should be made to learn from international comparisons. The experience of minorities in the United States appears to differ from that of minorities in other developed countries. Explanations should be sought by pinpointing key differences in the structure of national and local healthcare delivery systems. All such considerations and approaches should be employed to provide a comprehensive explanation.

See also: Gender and Health Care; Health Care Delivery Services; Health Care, Rationing of; Health Care Systems: Comparative; National Health Care and Insurance Systems; Socioeconomic Status and Health Care

Bibliography

Banthin J S, Cohen J W 1999 *Changes in the Medicaid Community Population: 1987–96* MEPS Research Findings No. 9. AHCPR Pub. No. 99-0042. Agency for Health Care Policy and Research, Rockville, MD

Collins J W, Derrick M, Hilder L, Kempley S 1997 Relation of maternal ethnicity to infant birthweight in east London, England. *Ethnicity and Disease* **7**(Winter): 1–4

del Pinal J, Singer A 1997 Generations of diversity: Latinos in the United States. *Population Bulletin* **52**: 1–44

Gilthorpe M S, Lay-Yee R, Wilson R C, Walters S, Griffiths R K, Bedi R 1998 Variations in hospitalization rates for asthma among black and minority communities. *Respiratory Medicine* **92**: 642–8

Kavanagh K H, Kennedy P H 1992 *Promoting Cultural Diversity: Strategies for Health Care Professionals*. Sage, London

Kleinman A 1978 Clinical relevance of anthropological and cross-cultural research: Concepts and strategies. *American Journal of Psychiatry* **135**: 427–31

Krieger N, Williams D R, Moss N E 1997 Measuring social class in US public health research: Concepts, methodologies, and guidelines. *Annual Review of Public Health* **18**: 341–78

Lee S M 1998 Asian Americans: Diverse and growing. *Population Bulletin* **53**: 1–39

Manson S M 1998 Mental health services for American Indians: Need, use, and barriers to effective care. Manuscript under review, University of Colorado Health Sciences Center

National Center for Health Statistics 1996 *Health, United States, 1995*. Public Health Service, Hyattsville, MD

Norton I M in press American Indians and mental health: Issues in psychiatric assessment and diagnosis. In: Herrera J M, Lawson W B, Sramek J J (eds.) *Cross Cultural Psychiatry*. Wiley, Chichester, UK

O'Hare W P, Pollard K M, Mann T L, Kent M M 1991 African Americans in the 1990s. *Population Bulletin* **46**: 1–40

Perloff J D, Kletke P R, Fossett J W, Banks S 1997 Medicaid participation among urban primary care physicians. *Medical Care* **35**(2): 142–57

Reijneveld S A 1998 Reported heatlh, lifestyles, and use of health care of first generation immigrants in The Netherlands: Do socioeconomic factors explain their adverse position? *Journal of Epidemiology and Community Health* **52**: 298–304

Snowden L R, Libby A, Thomas K 1997 Health-care-related attitudes and utilization among African American women. *Women's Health: Research on Gender, Behavior, and Policy* **3**: 301–14

Treas J 1995 Older Americans in the 1990s and beyond. *Population Bulletin* **50**: 1–46

Weigers M E, Drilea S K 1996 *Health Status and Limitations: A Comparison of Hispanics, Blacks, and Whites* MEPS Research Findings No. 10. AHCPR Pub. No. 00-0001. Agency for Health Care Policy and Research, Rockville, MD

Weinick R M, Zuvekas S H, Drilea S K 1997 *Access to Health Care—Sources and Barriers*. MEPS Research Findings No. 3. AHCPR Pub. No. 98-0001. Agency for Health Care Policy and Research, Rockville, MD

Wilson W J 1987 *The Truly Disadvantaged: The Inner City, The Underclass, and Public Policy*. University of Chicago Press, Chicago

L. R. Snowden

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Mis

Mission: Proselytization and Conversion

The concept of a universal mission to the world is embedded in different ways in the monotheistic religious traditions of Judaism, Christianity, and Islam. Judaism, though its scriptures contain strong elements of universalism, has in modern times ceased to be a missionary religion. Islam, by contrast, continues to win converts in large numbers, in part because of the appeal to peoples threatened by modernization of the Muslim concept of the immutability of the Shari'a or divine law. In this article 'mission' refers primarily to the activity of the churches in commending the doctrinal beliefs and moral values of Christianity as worthy of universal acceptance. 'Proselytization' (or proselytism) derives from the Greek *proselytos*—the term used by the Septuagint (the Greek version of the Old Testament) to describe a non-Jew who adopted the Israelite faith in Yahweh—and has come to mean the recruitment of converts from other religious systems and communities. Proselytization traditionally has been regarded as central to Christian mission, but does not necessarily constitute its totality. Medical and relief work, Christian education, the promotion of agricultural development, and agitation for a just international economic order, are now seen by Christians as integral parts of the mission to bring the world into conformity with Christian values. This article surveys the history of the term 'mission' and the missionary activities of the churches since the eighteenth century, reviews recent scholarship on the missionary movement and conversion, and notes attempts by many theologians to dissociate the concept of mission from proselytization.

1. The History of 'Mission' and 'Missions'

1.1 The Terms 'Mission' and 'Missions'

The use of the Latin term *missio* to describe the sending of Christians to propagate the faith among unbelievers was pioneered by Ignatius Loyola in the sixteenth century. Senior members of the Jesuit order vowed to go wherever the Pope might send them, to 'missions' among either the Catholic faithful or infidels, to promote the glory of God and the good of souls (Ross 1994). The widespread application of the singular vernacular term 'mission' to the overall activity of the church in the world is, however, more recent, dating only from the 1950s, when both political challenges and theological questioning compelled missionary theorists to re-evaluate precisely what it was that God had supposedly sent the church into the world to do. Until then, the plural term 'missions' was much more common, denoting the range of agencies deployed by both Catholic and Protestant churches to pursue the agreed aim of converting non-Christians. That usage remains widespread among American Christians and also among historians and social scientists concerned to investigate Christian missions simply as a human phenomenon.

1.2 The Modern Missionary Movement

Christian missions up to the mid-eighteenth century were conducted chiefly by the Roman Catholic and Russian Orthodox churches (the latter mainly in Siberia). The Roman Catholic missionary endeavor lost impetus following the suppression of the Jesuit order by the papacy in 1773, and the initiative passed to Protestants influenced by German pietism and the ensuing movements of evangelical revival which spread from central Europe to North America and Britain. The Protestant awakening culminated in the formation of numerous voluntary societies, beginning with the Baptist Missionary Society (1792) in Britain, the American Board of Commissioners for Foreign Missions (1810) in America, and the Basel Mission (1815) in continental Europe. For much of the nineteenth century these societies experienced only limited success in terms of both domestic recruitment and conversions overseas. From the late 1870s this picture changed as missionary enthusiasm converged with forms of popular imperialism, and as indigenous peoples became increasingly attracted by the literacy and access to literature (both vernacular and English) which the missions offered. By this time, European Roman Catholic missions had also revived, stimulated by the recovery of the papacy and the formation of new religious orders, such as the White Fathers, founded by Cardinal Lavigerie in 1868. Competition between Protestant and Catholic missions was frequently intense, and often coincided with imperial rivalries. North America was regarded by the Catholic Church as technically a mission field until 1908, and the first American Catholic missionary order, the Maryknoll Fathers and Brothers, was not founded until 1911.

By 1910 Protestant foreign missionaries working in Africa, Asia, and the Pacific numbered some 19,000, of whom 71 percent were male or female lay persons. Catholic missionary numbers were almost certainly larger. In both Protestant and Catholic missions, however, the bulk of the work of evangelization was performed by indigenous (and primarily lay) Christians. After the First World War, the Americans supplanted the British as the predominant force in Protestant missions, though the older denominational boards progressively were eclipsed by newer interdenominational missions of evangelical or fundamentalist inclination (Hutchison 1987). Since the 1960s, many western mission agencies have dwindled in size, but their place has been taken by new missions from the nonwestern world, notably from Korea. By 2000, the total Christian foreign missionary force had risen to an estimated 420,000.

2. *Recent Scholarship on Christian Missions*

2.1 *The Historiography of Missions*

Until the late 1960s most writing on Christian missions tended to be confessional, often hagiographical, rather than scholarly. This was despite the fact that missionaries had made substantial contributions to the development of secular disciplines such as linguistics, Sanskritic and oriental studies, and anthropology (Walls 1996). Since the 1960s historical and anthropological scholarship on the missionary movement has proliferated, prompted by heightened awareness of the importance of religion in colonial encounters and the shaping of indigenous nationalist movements. Some of the pioneering African historiography of the 1960s identified a growth in racist and colonialist attitudes among late Victorian missionaries as crucial in the origins of West African nationalism (e.g., Ajayi 1965). The anticolonial reaction of the 1960s and 1970s encouraged writing by historians, anthropologists, and theologians that interpreted the missionary movement as the ideological expression of western colonialism. Such interpretations were able to point to the close connections between missions and western imperial expansion in contexts such as China, and to the consistent use by missions of terminology which branded nonwestern peoples as benighted heathens who needed 'civilizing' by the supposedly Christian west.

Since the early 1980s a strengthening reaction to, or at least modification of this tradition has been evident. Scholars from a variety of disciplines (including, notably, literary studies) have become increasingly dissatisfied with models which erect a simple dichotomy between 'colonialist' missions and the 'victims' of colonialism (see *Religion: Culture Contact*). As the focus of imperial historiography has shifted from the materialist preoccupations of an earlier generation of Marxist-influenced historiography to a characteristically postmodern concern with the power of narrative and the construction of mentality and identity, scholarship has become less interested in the formal connections of missions to the apparatus of colonial rule and more concerned to elucidate the complex patterns of religious and cultural exchange created by missionary encounters (Peterson and Allman 1999). Rather than regarding indigenous converts simply as passive dupes of colonial pressure, scholars now are inclined to delineate a range of indigenous responses varying from outright rejection to transformative appropriations of the Christian message and the biblical text. This emphasis accords with the primacy of indigenous agents in transmitting the message, and with the fact that the rate of conversion to Christianity in the nonwestern world has accelerated, not slackened, since decolonization. Particular attention has been devoted to the cultural significance of vernacular translation, which committed missionaries, whether consciously or not, to a 'radical pluralism' wherein 'all languages and cultures are, in principle, equal in expressing the word of God' (Sanneh 1989, Walls 1996). The development of a vernacular literature, and especially a popular religious literature, to which mission Christianity contributed so fundamentally in nonwestern contexts, has been identified as the essential precondition for the transmutation of an ethnicity into a nation (Hastings 1997). The current academic trend does not, however, necessarily imply a more favorable estimation of the missionary impact. Perhaps its most influential exponents, the American anthropologists Jean and John Comaroff, despite an emphasis on the complexity and two-sided nature of missionary encounters in nineteenth-century southern Africa, continue to depict missionaries as 'colonial evangelists,' for whom language was not a medium of cultural liberation but rather a field of discourse whose terms they dominated in order to achieve 'the colonization of consciousness' (Comaroff and Comaroff 1991, 1997).

2.2 *Regional Studies*

The shift in focus towards the dynamics of conversion has been most marked in African studies, perhaps because of the remarkable growth of Christianity in sub-Saharan Africa (from some 60 million adherents in the 1960s to as many as 300 million by 2000). There are signs, however, of a similar trend in South Asian studies. The literary scholar, Gauri Viswanathan, though an adherent of Edward Said's school of 'Orientalism,' has argued that in British India, conversion to mission Christianity was a profoundly destabilizing influence, disrupting the stable categories of religion and nationhood on which the Raj depended, and compelling a professedly secular colonial administration to grapple with fundamental issues of civil rights (Viswanathan 1998). Studies of conversion by tribal and untouchable (or Dalit) peoples in India

have followed precedents set by Africanist scholarship in interpreting conversion to Christianity as a process, usually collective, which frequently offered those involved a more satisfactory framework of cosmological meaning and a more advantageous communal identity than the obvious alternative of adherence to Brahmanical versions of Hinduism. Frykenberg (in Hastings 1999) has observed that 'no mass movement of conversions in India was ever led by a European.'

Scholarship on conversion to Christianity in China, where the material linkages between missions, converts, and the protection of the imperial powers were more blatant than anywhere else, has been relatively slow to follow the trend set by the Africanists and South Asianists. Interpretations of Christianity in China during the missionary era have tended until recently to concentrate either on mission ideologies and strategies or on hostile antiforeign reactions to missions, rather than on the motivations of those Chinese who did convert (but for a recent collection which pioneers a more indigenous perspective on the Christian history of China see Bays 1996). The substantial growth of popular Christianity in Communist China since the era of the Cultural Revolution (1966–79) necessarily remains largely uncharted territory for scholarship.

Scholarly analysis of the impact of Christian missions on Latin America for long concentrated on the close connections between the Catholic Church and Iberian expansion which followed Columbus's 'discovery' of America in 1492 and the papal bull of 1493 which in effect divided ecclesiastical control of the continent between the Spanish and Portuguese crowns. More recently, attention has been turned to the phenomenal growth of popular Protestantism (especially Pentecostalism) in Latin America since the 1950s (see *Religion: Culture Contact*; *Christianity: Evangelical, Revivalist, and Pentecostal*). The primary question currently being debated is parallel to that posed by writing on African Christianity during the colonial period: to what extent can the current success of Protestant proselytism in this most traditionally Catholic of continents be interpreted satisfactorily as a function of North American cultural and political influence, or should greater emphasis in fact be placed on the ability of Pentecostalism to meet the felt needs of the Latin American poor? (Hastings 1999, Martin 1990).

2.3 Sociological and Gender Studies

Within the social sciences, interest in missions has been for obvious reasons more pronounced among anthropologists than among sociologists. Very few sociologists have devoted attention to the unusual nature of mission organizations as typically voluntary bodies which have needed to maintain control at a formidable distance over highly motivated and frequently eccentric individuals, and at the same time satisfy the expectations of their supporting domestic constituencies. One such study of how religious zeal was both exploited and contained through institutional mechanisms examines the Basel Mission on the Gold Coast (now Ghana) in the nineteenth century (Miller 1994).

Much of the recent writing on missions within the social sciences examines the significance of the missionary movement from a feminist perspective. For most of the nineteenth century, Protestant missions were overwhelmingly male institutions. Most societies did not accept single women candidates, and missionary wives were not counted as missionaries in their own right, despite making in many cases major contributions to the work of the mission. The first Protestant missionary society specifically for women (the Female Education Society) dates from 1834. The first such societies were often inspired by the needs of the high-caste women of the Indian zenanas, who, being secluded from public contact with men, could be reached and educated only by women. From the late 1860s, female or zenana missionary societies became common on both sides of the Atlantic, either as autonomous institutions, or as female auxiliaries to existing denominational societies. These bodies enabled women to fulfil public roles both in their homelands and overseas which had few parallels in other contexts. From the 1870s, the denominational societies themselves began to accept single women candidates. They did so to varying extents, but the trend was accentuated by the readiness of many of the new nondenominational 'faith missions' to recruit single women (Robert 1996). By 1910 the missionary movement had become a predominantly female enterprise still controlled by men. Whilst still influenced by male assumptions about the proper restriction of women to the domestic sphere, it modeled a different reality. In that year, single women made up nearly 28 percent of the entire Protestant missionary force, and married women a further 28 percent. It seems likely that a similar gender distribution obtained for most of the twentieth century. The proliferation of missionary sisterhoods, first in France in the late nineteenth century, and then more widely, produced an even greater predominance of women in the Catholic missionary movement. Missionary women have thus become a major concern of scholars investigating what has been termed 'imperial feminism,' which is held to have combined 'colonial' views of racial difference with stances that were in practice subversive of traditional concepts of gender difference (Chaudhuri and Strobel 1992).

3. Mission and Proselytism in Theological Discourse

Whilst secular scholarship on missions has tended to move away from approaches which interpreted the

proselytizing imperative as inescapably connected with the colonial system of material domination towards indigenous perspectives which emphasize the volitional agency of the convert or convert community, much Christian theology since the 1960s ironically has continued to regard coercion as intrinsic to the concept of proselytization, and has therefore been anxious to divorce 'mission' entirely from proselytism. One animating impulse of this concern has been the experience of the Roman Catholic, Orthodox, and other ancient eastern churches in finding their faithful to be the objects of often aggressive and well-funded evangelism by Protestant groups, particularly from the USA. A report jointly produced by the Roman Catholic Church and the World Council of Churches in 1970 sought to distinguish Christian witness from proselytism, defining the latter as 'improper' attitudes or behavior which employed coercion and thus violated the rights of the human person. Such a distinction has proved an unstable one, since almost any act of religious witness can be interpreted by the recipient as being coercive or invasive. Since the collapse of Communist regimes in eastern Europe in 1989 and the dissolution of the Soviet Union, opposition to proselytism from the west has motivated legislation designed to curb the missionary work of those regarded by the Orthodox churches as 'sects.' More broadly, from the perspective of a Muslim, Hindu, Buddhist, or Jew, all Christian missionary activity is proselytizing in intention and effect, leading inevitably to the severance of the convert from community and even nation. Some Christian theologians accordingly have taken their opposition to proselytism towards its logical conclusion by denying that the conversion of adherents of other religions is a valid missionary objective for Christians. It then becomes difficult to resist the application of the same principle to the conversion of humanists or adherents of any belief system whatever, and the universalism intrinsic to the Christian theological tradition becomes unsustainable. Religious pluralism is no new phenomenon, but all religious believers in the western world now face unprecedented social and cultural pressures to accept the permanence and ideological absoluteness of pluralism. In the younger churches of the non-western world, whose role in world Christianity is becoming increasingly dominant, insistence on the inescapable connection between mission and proselytization remains strong. It may therefore be predicted that the twenty-first century is likely to see an even more heated debate within the churches about the validity and meaning of mission, proselytization, and conversion.

See also: African Studies: Religion; Christianity in Asia; Christianity in Central and South America; Colonialism, Anthropology of; Colonization and Colonialism, History of; Latin American Studies: Religion; Religious Fundamentalism: Cultural Concerns; Southeast Asian Studies: Religion

Bibliography

Ajayi J F A 1965 *Christian Missions in Nigeria 1841–1891: The Making of a New Elite*. Northwestern University Press, Evanston, IL

Bays D H (ed.) 1996 *Christianity in China: From the Eighteenth Century to the Present*. Stanford University Press, Stanford, CA

Chaudhuri N, Strobel M (eds.) 1992 *Western Women and Imperialism: Complicity and Resistance*. Indiana University Press, Bloomington, IN

Comaroff J, Comaroff J 1991 *Of Revelation and Revolution*. University of Chicago Press, Chicago

Comaroff J, Comaroff J 1997 *Of Revelation and Revolution*. University of Chicago Press, Chicago

Hastings A 1997 *The Construction of Nationhood: Ethnicity, Religion and Nationalism*. Cambridge University Press, Cambridge, UK

Hastings A 1999 (ed.) *A World History of Christianity*. Eerdmans, Grand Rapids, MI

Hutchison W R 1987 *Errand to the World: American Protestant Thought and Foreign Missions*. University of Chicago Press, Chicago

Martin D 1990 *Tongues of Fire: The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK

Miller J 1994 *The Social Control of Religious Zeal: A Study of Organizational Contradictions*. Rutgers University Press, New Brunswick, NJ

Peterson D, Allman J 1999 Africans meeting missionaries. *Journal of Religious History* **23**: 1

Robert D 1996 *American Women in Mission: A Social History of Their Thought and Practice*. Mercer University Press, Macon, GA

Ross A C 1994 *A Vision Betrayed: The Jesuits in Japan and China, 1542–1742*. Edinburgh University Press, Edinburgh, UK

Sanneh L 1989 *Translating the Message: The Missionary Impact on Culture*. Orbis Books, Maryknoll, NY

Viswanathan G 1998 *Outside the Fold: Conversion, Modernity, and Belief*. Princeton University Press, Princeton, NJ

Walls A F 1996 *The Missionary Movement in Christian History: Studies in the Transmission of Faith*. Orbis Books, Maryknoll, NY

B. Stanley

Mitochondria and Evolution

Mitochondria in eukaryotic cells are self-reproducing, semi-autonomous organelles that are the site of ATP synthesis, the cell's source of chemical energy. Many gene products necessary for the production of ATP are encoded by the mitochondrion's own DNA, often organized as a circular, double-stranded molecule within the inner membrane of the organelle. This ancient and distinct genome is designated MtDNA, to emphasize its separate evolutionary history and underscore the fact that it is replicated and inherited

differently from the ordinary DNA on chromosomes in the cell's nucleus (Wolstenholme 1992). In contrast to nuclear genes, most animal mitochondrial genes are linked on a single small loop less than 20 kilobases in size, and commonly show uniparental inheritance, usually through the female parent. A single female germ cell can have thousands of copies of the mitochondrial genome, and to a first approximation, each copy is identical.

Sperm also contain mitochondria, but a chemical signal, tagging of sperm mitochondria with the protein ubiquitin, triggers its destruction. This MtDNA is eventually lost in the developing embryo, usually before the third embryonic cleavage (Sutovsky et al. 1999). The maturing embryo thus contains a mixture of both parent's nuclear genes, but generally inherits only the mitochondrial genes of its mother.

1. Heteroplasmy Complicates Simple Uniparental Inheritance of MtDNA

Paternal transmission of MtDNA in most vertebrates is thus a rare event, mostly confined to interspecific crosses, and can be one source of the condition known as genetic heteroplasmy. A heteroplasmic individual is a genetic chimera and contains multiple MtDNA sequences. As a consequence, in species with high heteroplasmy, molecular taxonomies based on only MtDNA markers become unreliable. Instead of following a simple maternal gene genealogy, the population genetics of MtDNA transmission in heteroplasmic species can become difficult to follow.

The probability of transmission of any particular genotype in heteroplasmic taxa requires knowing the frequency of different genotypes in the sex cells of parents involved in any particular mating. Offspring from the same parents can even differ in the proportions of mixed genotypes they inherit under these circumstances, and cases illustrating this problem are known in domestic cattle herds, bovine interspecies crosses, murine hybrids, humans with complex metabolic disorders, or those babies produced as a result of fertilization assisted by spermatid or sperm microinjection (Kaneda et al. 1995, Sutovsky et al. 1999).

A second source of individual heteroplasmy is the rapid fixation of MtDNA mutations, which can result in both sequence and length differences in the mitochondrial genome within the same individual. Unlike the first case, heteroplasmy of this nature will have fewer long-term evolutionary consequences, if germline cells are unaffected. As an individual ages, mutations often accumulate in the MtDNA of somatic cells. Since differentiated cells may divide infrequently, especially if damaged; a given organ can eventually become a mixture of different MtDNA genotypes. Recombinational repair of mutations, demonstrated in vitro using rat liver mitochondrial protein extracts may counteract this (Thyagarajan et al. 1996), but the lack of nucleotide excision repair activity may limit the replication of heavily damaged MtDNA.

Mitochondria are known to be the accumulation site of highly reactive molecules and free oxygen radicals associated with ATP production during metabolic activity. A lively debate surrounds the question of whether or not MtDNA damage might be generally repaired by genetic recombination mechanisms. Evidence for mitochondrial genetic recombination is circumstantial in vertebrates and hinges on a more complete understanding of mutations at hypervariable regions within the organelle's genome.

Recombination is known to occur in the MtDNA of numerous fungi, and MtDNA Pseudo-genes (nonfunctional MtDNA sequences) have been found transferred into the nuclear genome of most organisms studied. While mammalian mitochondria possess a recombinational DNA repair pathway, the same enzymes might not be associated with genetic recombination like in yeast. The mechanism(s) mediating the transfer of mitochondrial gene fragments into the nuclear genome are unknown at present.

1.1 Elevated MtDNA Mutation Rates Offer Opportunities

Maintenance of function often places such extreme constraints on biochemically essential pathways that mutations are incompatible with life. MtDNA contains sequences absolutely essential for producing a limited number of gene products involved in ATP synthesis (cytochrome oxidases, cytochromes, NADH dehydrogenases, and ATPases, along with specialized ribosomal and transfer rRNA elements, but they exist in high copy numbers and also contain regions within genes that are not under such tight functional constraints. MtDNA basically shows a 10-fold elevated mutation rate compared to nuclear DNAs of the same functional class (including ribosomal and transfer RNAs, and the protein coding regions).

Mitochondrial genomes as a group are specialized and streamlined, with fewer noncoding regions, compared to the numerous introns within genes in the nucleus. The noncoding regions retained by mitochondrial genomes are of interest to many evolutionary biologists, because they allow fairly personalized identification of maternal lineage groups in a nested, time-dependent fashion as new mutations rapidly accumulate on the old genetic backgrounds or haplotypes. The elevated mutation rate in noncoding MtDNA almost guarantees finding adequate polymorphism within populations such that individuals will possess unique DNA sequences. Only close maternal relatives (siblings with the same mother, maternal cousins, and maternal grandmothers) will likely carry the same MtDNA sequence.

The major noncoding region in MtDNA is commonly called the D-loop, a term borrowed from

bacterial genetics and referring to a region that contains sequences for the initiation of DNA replication. Population studies within mammalian species often focus on the approximately 1,000 base-pair hyper-variable region, where substitution rates may be 10–100 times more than the average rate for the rest of the mitochondrial genome (considered about 2 percent sequence change per 1 million years of separation from the last common maternal ancestor). Mutations segregating in the noncoding regions can be used to examine sex-biased dispersal and sexual selection, the stability of specific demes contributing to biodiversity in refugia, demographic bottlenecks and founder events, as well as hybridization processes.

Local populations are often dominated by two or three MTDNA lineages that have diverged from a common ancestor less than 100,000 years ago. Mitochondrial and nuclear genetic systems such as Microsatellites, Y-chromosome markers, short tandem repeats (STRs), or variable number tandem repeats (VNTRs) may lead to different conclusions about the history of a given group. High levels of MtDNA lineage resolution may uncover unanticipated patterns of dispersal related to altered behavior within hybrid zones, previous instances of demographic catastrophes linked to environmental perturbations or disease epidemics, and also hitchhiking of nearly neutral hyper-variable MtDNA markers in linkage to co-evolving haplotype sequences that convey some strong adaptive advantage to the population.

1.2 Other General Applications of Hyper-variable MtDNA Markers

There are tens of thousands of individual mitochondrial sequences from different parts of the genome now reported in data banks, further evidence of the high mutation rate associated with these genes. Complete mitochondrial genomes are known for a variety of vertebrates and invertebrates, including organisms with poor taxonomic histories that are difficult to collect, breed, or maintain in a laboratory setting. Knowledge of some MtDNA sequences has stimulated the design strategy of degenerate Polymerase Chain Reaction (PCR) primers (Simon et al. 1994) that can be employed to study cryptic species, model the past population history of now-extinct organisms from museum specimens, and survey for biodiversity in barely studied ecosystems. Researchers also argue for the conservation of genetically distinct lineages with poorly differentiated morphologies but showing complex mating, foraging, or predator avoidance behaviors, tied to as yet unknown quantitative behavioral genes.

MtDNA mutations have also become important genetic markers in humans for clinical research on cancer, heart disease, neurodegenerative disorders, and other conditions. Affected populations of cells may take on distinct sets of mitochondrial substitutions at particular nucleotides after exposure to mutagens, other toxins in the environment, chemotherapy, radiation, or simple aging. Since MtDNA makes up 1–2 percent of all the DNA present in an animal's cell, it is well suited to forensic and ancient population studies where only a few cell fragments survive.

2. Evolutionary Insights gained using MtDNA

Original enthusiasm expressed by some that the problem of speciation, Darwin's 'mystery of mysteries,' would be eventually understood with molecular phylogenies based on mitochondrial markers has been tempered by the realization that even with a well-supported MtDNA tree, species concepts, reproductive isolation, and morphological discontinuities will require broad-based research on behavioral and ecological fronts. No case perhaps illustrates this point more clearly than the continued debate over inclusion of Neanderthals in the direct line leading to modern humans. Two independent groups showed that the tree connecting the maternal ancestor of MtDNA hypervariable sequences from two different Neandertal humans to modern humans far exceeds the degree of variation found in over 5,000 modern human sequences (Ovchinnikov et al. 2000).

Mitochondrial markers have been particularly important for demonstrating the stages of genetic differentiation that many animal species show as a population evolves through paraphyly to reciprocal monophyly and complete differentiation. Indeed, discussions about molecular data and modern species origins often center on the most appropriate interpretation of these stages in speciation (Morrow et al. 2000). Patterns of allelic lineage distributions in spatially isolated subgroups have been useful for generating hypotheses concerning Pleistocene refugia, dispersal corridors, behavioral reinforcement, and incomplete barriers to gene flow in plants as well as animals. Controversy over legal protection of specific populations that show MtDNA lineage patterns, coincident with habitat disturbances disrupting ecological boundaries, often transcends dry systematic debates on the proper designation of hybrid taxa.

2.1 Intra and Intergenomic Conflicts of Interest

Hybrid taxa are of particular interest for mitochondrial genetics because hybrid dysgenesis may result when mitochondria from one taxa find themselves functioning in the company of a nuclear genome from another distinct lineage, as in the case of *Mus musculus* and *Mus domesticus* (Sage et al. 1986). Two or more genes that must interact within a single organism may display a spectrum of evolutionary responses grading from mutualism to antagonism. Speculation on why mitochondria retain some genes

while shunting others into the nucleus often revolves around the idea of genomic conflicts of interest.

Mitochondrial genes normally coevolve in linkage as a uniparental chromosomal unit, and this was recognized by Hurst and Hamilton (1992) as one of two ways to prevent cytoplasmic gene conflict. In their view, sperm evolved to prevent the mixing of incompatible cytoplasmic genes. For many common species, the male sex has resigned attempts to contribute MtDNA to the next generation. Since the topic of hybrid male sterility has been so important to research on the genetics of speciation, molecular exploration of the loci controlling the ubiquitin-tagging pathway is likely to increase our sophistication of understanding how genomic conflicts have been solved by various species.

The suggestion that sex is an integral part of intragenomic conflict helps explain why nuclear genetic mechanisms distinct from sex but linked with incompatibility have arisen in a number of organisms, from acellular slime molds to yeast and certain ciliates. These taxa, as predicted, show unusual numbers of mating types and variations in the uniparental inheritance pattern of MtDNA. Highly inbred organisms also tend to have biparental inheritance of organelle DNA, since conflict-reducing mutations can easily spread among identical genotypic backgrounds.

2.2 *Origin of Complex Cells*

The prevailing view of mitochondrial genomes is that they are circular, but a large number of organisms possess linear MtDNA, including some ciliates, parasitic apicomplexan protozoa such as *Plasmodium* the other animal parasites *Babesia* and *Theileria*, some algae, slime molds, oomycetous fungi, many yeast, some Cnidaria species, and several Hydra species (Nosek et al. 1998). Baker's yeast (*Saccharomyces cerevisiae*) MtDNA is normally portrayed as circular, with linear molecules commonly found *in vivo*, but they had been interpreted as degraded circles, as found in many plant MtDNAs. This view has now changed based on pulsed-field gel electrophoresis employed to study very large molecules, and a more sophisticated understanding of the telomeres of linear mitochondrial genomes has resulted (they are of intense interest to molecular biologists as potential sites for new drug development in diseases caused by organisms with these specialized linear structures). The question of linear vs. circular genomes is also important in order to evaluate the endosymbiont hypothesis for the origin of mitochondria, since circular genomes are widespread among hypothetical prokaryote ancestors, the purple bacteria (Wilson et al. 1987).

Do organisms with linear MtDNAs represent a separate prokaryotic lineage, compared with those with circular ones? Gene sequences as well as gene order have been examined. Phylogenetic trees of yeasts possessing both types of genomes make it clear that based on ribosomal sequences, there is not a radical difference between circles or lines. Analysis of the mechanisms that have evolved to solve the problem of replication of 5′ends of DNA are likely to reveal how linear and circular organelle genomes are related.

2.3 *Adaptation, Biogeography, and Phylogeography*

An extensive literature of original research dealing with the use of MtDNA markers in evolutionary biology was summarized by Avise (1994). Aside from providing phylogenetic trees for species with no fossil record, the mitochondrial marker approach is particularly useful in the mapping of organismal traits onto testable hypotheses of phylogeny based on parsimony, distance, and likelihood methods. Characters such as powered flight and the loss of it, tetrapod limb structure, brood parasitism, eusociality, melanism, endothermy, carnivory, vocal signal complexity, androdioecy, and even multicellularity have been placed on MtDNA phylogenies.

Aside from higher systematics and the phylogenetic distribution of specific adaptations, mitochondrial haplotypes can, like other single-locus markers, be analyzed as single genes with frequencies and geographic distributions that help portray population history, even in species with limited genetic variation. The relationships of the alleles in a population are informative for longer-term processes of phylogeography and population subdivision (Burke 1998), and can sometimes untangle current from historical events and reveal demographic trends.

2.4 *Nonequilibrium Population Genetics*

MtDNA characteristically is more sensitive to aspects of population biology and history owing to its uniparental transmission that reflects a lower effective population size (Ne). Populations not in Hardy-Weinberg equilibrium for genetic markers (Slatkin 1993) include many taxa of interest to conservation biologists as well as weedy or invasive species showing recent expansions of geographic range. When measured for genotype diversity, groups undergoing selection, nonrandom mating, enhanced mutation, and characterized by tiny population sizes often show deviations from expected Hardy-Weinberg allele frequencies.

MtDNA genotypes in these groups sometimes evidence 'waves' in the frequency distributions of particular haplotypes compared on a pairwise basis with count mismatches (Rogers and Harpending 1992). Thus, mitochondrial genotypes are particularly useful in identification of migrants and hybrid individuals, asymmetrical mating preferences, and historical effects

stemming from stochastic changes in populations where an ancestral source population was polymorphic for a particular set of markers (Waser and Stroebeck 1998).

Reductions in effective population size may be the result of either population bottlenecks or selective sweeps and so mismatch distributions alone cannot be used to infer population history. However, the general shape of MtDNA trees, where a star-like pattern of alleles indicates few mutational differences between the major variants in a population, is consistent with the hypothesis of recent deme expansion. Deep branches in gene trees, in contrast, usually reflect long periods of population subdivision. Highly polymorphic nuclear genetic markers and developments of phylogenetic estimation theory will be combined with mitochondrial haplotype information to yield detailed scenarios of population history during the final stages of the Pleistocene era for some highly vagile species (Bowcock et al. 1994 Cogdon et al. 2000).

See also: Adaptation, Fitness, and Evolution; Behavioral Genetics: Psychological Perspectives; Brain, Evolution of; Evolutionary Selection, Levels of: Group versus Individual; Natural Selection

Bibliography

Avise J 1994 *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York

Bowcock A M, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd J R, Cavalli-Sforza L L 1994 High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **386**: 455–7

Burke T (ed.) 1998 Special Issue: Phylogeography. *Molecular Ecology* **7**: 367–545

Cogdon B C, Piatt J F, Martin K, Friesen V L 2000 Mechanisms of population differentiation in marbled murrelets: Historical versus contemporary processes. *Evolution* **54**: 974–86

Hurst L D, Hamilton W 1992 Cytoplasmic fusion and the nature of sexes. *Proceedings of the Royal Society, London B* **247**: 189–94

Kaneda H, Hayashi J-I, Takahama S, Taya C, Fischer-Lindahl K, Yonekawa H 1995 Elimination of paternal mitochondrial DNA in intrasepcific crosses during early mouse embryogenesis. *Proceedings of the National Academy of Sciences USA* **92**: 4542–6

Morrow J, Scott L, Cogdon B, Yeates D, Frommer M, Sved J 2000 Close genetic similarity between two symatric species of Tephritid fruit fly reproductively isolated by mating time. *Evolution* **54**: 899–910

Nosek J, Tomaska L, Fukuhara H, Suyama Y, Kovac L 1998 Linear mitochondrial genomes: 30 years down the line. *Trends in Genetics* **14**: 184–8

Ovchinnikov I V, Gotherstrom A, Romanova G P, Kharitonov V M, Liden K, Goodwin W 2000 Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404**: 490–3

Rogers A R, Harpending H C 1992 Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* **9**: 552–69

Sage R D, Heyneman D, Lim K-C, Wilson A C 1986 Wormy mice in a hybrid zone. *Nature* **324**: 60–3

Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P 1994 Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America* **87**: 651–701

Slatkin M 1993 Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**: 264–79

Sutovsky P, Moreno R D, Ramalho-Santos J, Dominko T, Simerly C, Schatten G 1999 Ubiquitin tag for sperm mitochondria. *Nature* **402**: 371–2

Thyagarajan B, Padua R A, Campbell C 1996 Mammalian mitochondria possess homologous recombination activity. *The Journal of Biological Chemistry* **271**: 27536–43

Waser P M, Stroebeck C 1998 Genetic signatures of interpopulational dispersal. *Trends in Ecology and Evolution* **13**: 33–4

Wilson A C, Ochman H, Prager E M 1987 Molecular time scale for evolution. *Trends in Genetics* **3**: 241–7

Wolstenholme D R 1992 Animal mitochondrial DNA: Structure and evolution. *International Review of Cytology* **141**: 173–216

R. L. Cann

Mixture Models in Statistics

Mixture models are typically used where a heterogeneous population of sampling units consists of several subpopulations within each of which a relatively simpler model applies. The source of heterogeneity could be gender, age, geographical origin, cohort status, etc. The use of *Mixture models* or, in particular, of *finite mixture distributions* for modeling phenomena goes back to the early years of statistics (see McLachlan and Peel (2000) for an account of the history of mixture models). We note in particular Newcomb (1886), who used a normal mixture to model outliers, and Pearson (1894), who employed the method of moments to decompose normal mixtures. Since then numerous papers have been published in which the theory and applications of mixture distributions were discussed.

Mixture models can be interpreted in two different ways which can be referred to as direct and indirect interpretations (Titterington et al. 1985). By a direct interpretation, it is assumed that the underlying population is comprised of k subpopulations and that the experimental unit on which the observation X is made belongs to one of these subpopulations, and that there is no knowledge of the identity of this subpopulation. In other words, the mixture model is intended as a direct representation of the underlying phenomenon. By an indirect interpretation, the mixture model is simply used as an approximation to an unknown distribution, where the 'subpopulations' have no direct physical interpretation. As an example of an indirect application, we adapt an example and

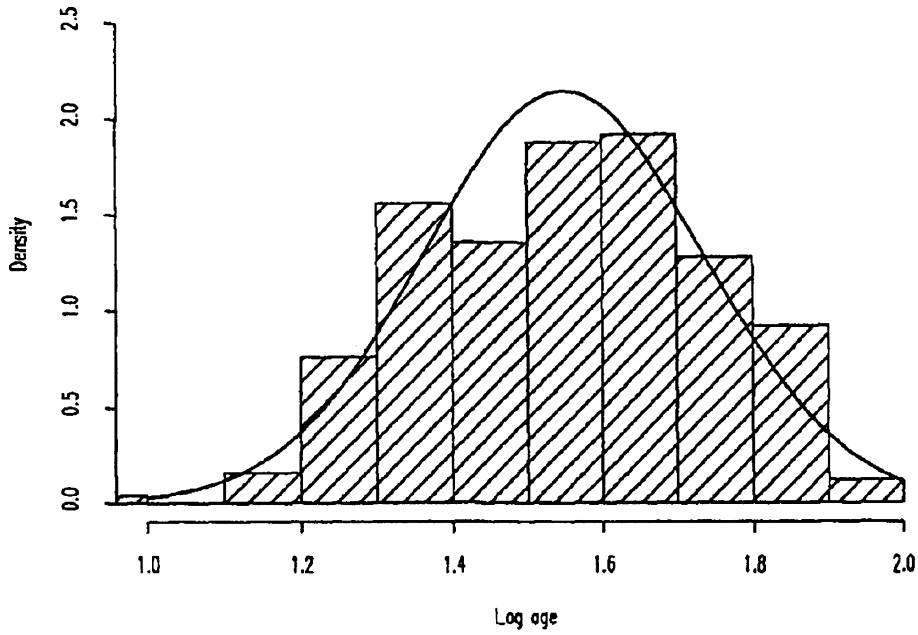

Figure 1
Age of onset data and fitted single normal density for male schizophrenics

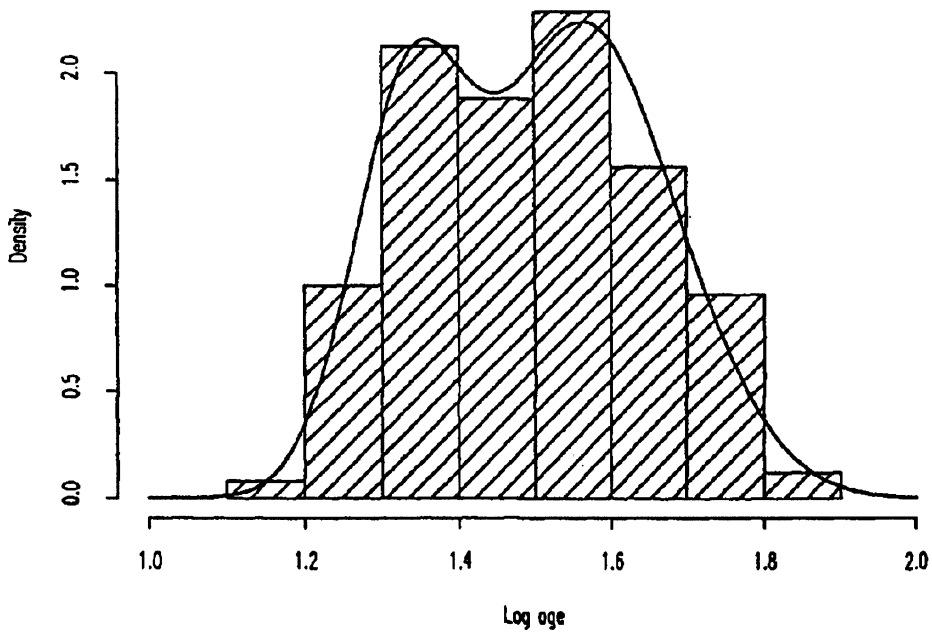

Figure 2
Age of onset data and fitted two component normal mixture for female schizophrenics

figures reported in Everitt (1996). The distribution of the (log) age of onset of schizophrenia was studied separately for men and women. It transpired (see Figs. 1 and 2) that while the data for men could adequately be described by a single normal density peak at the age of about 35, for the data for women, a two-component mixture was needed with an additional peak at about 22.

In order to define a mixture model mathematically, suppose the distribution of a random variable (or vector) X can be represented by a probability density function (or probability mass function, a term used here for both continuous and discrete random variables) of the form

$$p(x \mid \psi) = \pi_1 f_1(x \mid \theta_1) + \ldots + \pi_k f_k(x \mid \theta_k) \qquad (1)$$

where $\psi = (\pi, \theta)$, $\pi = (\pi_1, \ldots, \pi_k)$, $\theta = (\theta_1, \ldots, \theta_k)$, $\pi_j \geqslant 0$, $j = 1, \ldots, k$, $\pi_1 + \ldots + \pi_k = 1$, $f_j(x \mid \theta_j) \geqslant 0$, and $\int f_j(x \mid \theta_j) \mathrm{d}x = 1$.

The presentation in Eqn. (1) is usually referred to as a *finite mixture model*. Here θ_j is a parameter vector characterizing the component density function $f_j()$ and the π_j's are mixing weights. Model (1) allows each component density to belong to a different parametric family, allowing in principle for mixed continuous and categorical variables. In most applications, a common parametric family is assumed and thus the mixture model takes the following form

$$p(x \mid \psi) = \pi_1 f(x \mid \theta_1) + \ldots + \pi_k f(x \mid \theta_k) \qquad (2)$$

which we assume to hold in the sequel.

The mixing weight π can be regarded as a discrete probability function over θ, describing the variation in the choice of θ across the population of interest. If we allow θ to be continuous and thus assume a density function $g(\theta)$, the finite mixture model suggested above is replaced by $p(x) = \int f(x \mid \theta) g(\theta) \mathrm{d}\theta$, which is often referred to as a *compound distribution*. When $g(\theta)$ is modeled without parametric assumptions, $p(x)$ is a *semiparametric mixture model*. For a review of semiparametric mixture models, see Lindsay and Lesperance (1995). Unless otherwise stated, the discussion of mixture models will be confined to finite models of the type shown in Eqn. (2).

Mixture models can be viewed as a semiparametric compromise between a single parametric density ($k = 1$) and a nonparametric model represented by $k = n$, where the mixture components are represented by the kernel method of density estimation.

The presentation and calculation of mixture models are often facilitated by the introduction of unobserved indicators $\zeta_i \in \zeta$, $\zeta_i = (\zeta_{i1}, \ldots, \zeta_{ik})$, which specify the mixture component from which the ith observation arises. Clearly, all the ζ_i's are zero except for that element, q say, indicating that x_i belongs to the qth subpopulation. The distribution of each unobserved vector ζ_i is therefore multinomial $(1, \pi_1, \ldots, \pi_k)$ if the observations are independent, as commonly assumed (see Titterington (1990) for mixture models with dependence structures).

In suggesting a mixture model, it is essential to ascertain that the proposed mixture is identifiable; i.e., that for any one class of models considered there exists a unique characterization. For a discussion of this topic, see Titterington et al. (1985) (see also *Statistical Identification and Estimability*).

1. Applications

The use of mixture models is widespread. For a comprehensive list of applications, see Titterington et al. (1985). Recent areas of application include:

Banking and finance: Watanabe (2000).
Marketing: Kamakura and Wedel (1997)
Mental processes: Yantis et al. (1991)
Insurance: Jasiulewicz (1997)
Demography: Li and Choe (1997)
Health care: Deb and Trivedi (1997)
Inventory management: Karmakar (1994)
Survival analysis: Ng and McLachlan (1998).

Mixture models have proved useful in *robust statistics*, where the mixture structure

$$p(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$$

allows the study of data that mostly arise from $f_1(x)$ with additional 'contaminants' stemming from $f_2(x)$. For a discussion, see Titterington et al. (1985). As an example, normal mixtures can be used to evaluate the degree of robustness of a statistical procedure (like testing or estimation) when the underlying assumption of normality is violated (Andrews et al. 1972). Robustness can be enhanced by using heavy-tailed distributions, like the t-distribution (Peel and McLachlan 2000) (see *Robustness in Statistics*).

Mixture models constitute a useful tool for *clustering* data into 'meaningful' groups. This way a k-component mixture corresponds to a data set comprising of k groups (see *Statistical Clustering*). Once a mixture model is fitted to the data a probabilistic assignment of the data into the k groups can be achieved by allocating an observation to a group to which the posterior membership is highest. Methods of allocation are discussed within the framework of *discriminant analysis* (McLachlan 1992). For more details on mixture models and clustering, see McLachlan and Basford (1988) (see *Multivariate Analysis: Classification and Discrimination*).

Mixture models can treat categorical data too. An important application in this context occurs in *latent class models* which have attracted a lot of attention in the social sciences and biostatistics (see Forman and Kohlmann (1996) for a review). In these models, a random vector x of dichotomous variates follows the mixture model

$$p(x) = \sum_{i=1}^{k} \pi_i f(x \mid \theta_i)$$

where $x_\eta (\eta = 1, \ldots, d)$ the element of x, takes values zero or one, $\theta_{i\eta} (\eta = 1, \ldots, d)$, the element of θ_i, is the probability that the ηth variable is present in class i, and the ith component is given by

$$f(x \mid \theta_i) = \prod_{\eta=1}^{d} \theta_{i\eta}^{x_\eta} (1 - \theta_{i\eta})^{1 - x_\eta}$$

The model suggests the existence of k latent classes, which may or may not have a real physical interpretation. The model generates the d dichotomous variates and explain the covariance structure of the data (see *Factor Analysis and Latent Structure: Overview*).

Mixtures of *generalized linear models* have also been investigated lately. See, for example, Aitkin (1999). These models take the general structure

$$p(x_j \mid \psi) = \sum_{i=1}^{k} \pi_i f(x_j \mid \theta_{ij}, \lambda_i)$$

where the log density of x_j is given by

$$\log f(x_j \mid \theta_{ij}, \lambda_i) = m_j \lambda_i^{-1}[\theta_{ij} x_j - b_i(\theta_{ij})] + c_i(x_j; \lambda_i)$$

This class of mixture models include mixtures of *Poisson regressions*. Such a mixture is adequate to model count data (number of outbreaks of a disease, number of accidents, etc.) where the components of the mixture are Poisson distributions

$$f_j(x \mid \theta_j) = e^{-\theta_j} \theta_j^{x_j} / x_j!, \; x_j = 0, 1, 2, 3\ldots$$

where θ_j, the mean of the underlying Poisson distribution, is a function of k explanatory variables, typically via a log transformation

$$\log(\theta_j) = \beta_0 + \beta_1 x_{j1} +, \ldots + \beta_k x_{jk}$$

Other mixtures of generalized linear models include mixtures of *logistic regressions* and mixtures of *experts model*. These are discussed in McLachlan and Peel (2000) (see *Multivariate Analysis: Discrete Variables (Logistic Regression)*).

2. Parameter Estimation

Learning about the parameters of the model can be achieved by a variety of methods (see *Statistical Identification and Estimability*). *Statistical Identification and Estimability* have become increasingly popular for exploratory data analysis and in the case of mixture data are often used for two purposes:

(a) Suggesting whether the underlying model is a mixture model of a particular type.

(b) Providing crude estimates of the model parameters, including k, the number of components. For more details, see Titterington et al. (1985) and Roeder (1994).

The *method of moments* is one of the oldest methods for estimating mixture parameters. Using the method of moments typically results in consistent estimators which are not necessarily efficient. However, the solution of the method of moments' equations may not be possible, or may be computationally difficult to obtain. For example, in his pioneering paper, Pearson suggested estimating the five parameters of a mixture of two normal densities by attempting to solve a ninth-degree polynomial. Further, estimates produced by the method of moments may not be unique or at all feasible in case they lie outside appropriate regions.

Given n observations from the mixture model in Eqn. (2), estimation by the *maximum likelihood* approach requires the minimization of the likelihood function

$$L(\psi) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} \pi_j f(x_i \mid \theta_j) \right] \tag{3}$$

with respect to the elements in ψ.

The maximum likelihood approach is commonly applied in parametric families both because of the relatively simple computations required and the attractive theoretical properties of the estimates. However, for mixture models, asymptotic properties and computational aspects are not always straightforward. For a discussion of the theoretical properties of the maximum likelihood estimates, see Titterington et al. (1985). As for computational issues, we note that for most mixture models the explicit evaluation of maximum likelihood estimates is usually not possible. However, the observation that mixture data are analogous to incomplete data (where the indicators ζ, assigning an observation to a particular subpopulation, are missing) allows the application of a class of iterative procedures known as *EM algorithms* (expectation maximization), which have proved very useful for mixture data. For more details, see McLachlan and Krishnan (1997). For a review of reliable maximum likelihood algorithms for semiparametric mixture model, see Böhning (1995).

Bayesian inference (see *Bayesian Statistics*) requires the specification of a prior distribution of ψ, $p(\psi)$, which provides the belief about the possible values of ψ prior to observing the data x. Using *Bayes' Theorem* we define the posterior distribution of ψ by

$$p(\psi \mid x) = \frac{L(\psi) p(\psi)}{\int L(\psi) p(\psi) \mathrm{d}\psi}$$

Due to the complicated nature of the likelihood function (Eqn. (3)) implementation of the Bayesian paradigm was, until the early 1990s, computationally prohibitive. Modern Bayesian implementation would require the specification of the prior distribution for the unknown parameters $\psi = (\pi, \theta)$, which is often taken to be the product of independent prior distribution of π and θ. With the aid of the indicators ζ and the prior distribution for ψ, iterative methods can be suggested. The EM algorithm can be used to estimate the mixture parameters, averaging over the indicator variables. For more details, see Gelman et al. (1995, Chap. 16). Another approach, which has been proved highly successful, is based on posterior simulations via *Markov chain Monte Carlo* (MCMC)

methods, (see *Monte Carlo Methods and Bayesian Computation: Overview*). Without dwelling on particular variants of the MCMC, we note that MCMC simulations for estimating the parameters of a mixture model typically alternate between two principal steps (Diebolt and Robert 1998):

(a) Drawing from the posterior distribution of the indicator variables, given the data and the model parameters (as drawn from their posterior distribution).

(b) Drawing from the posterior distribution of the model parameters, given the data and the indicator variables (as drawn from their posterior distribution).

These methods require special attention to issues concerning convergence and the cautious use of improper prior distribution. Robert and Casella (1999) provide further details and Gelman et al. (1995) present numerical illustrations where a mixture model is fitted to reaction time of schizophrenic and non-schizophrenic examinees.

When the number of components in a mixture is unknown, the computational complexity of the model increases dramatically. For extensive discussion, see Titterington et al. (1985) and McLachlan and Peel (2000). One possibility when the order of the model is unknown is to consider two models, one nested within the other, such that under the null hypothesis the model consists of k_1 components and under the alternative hypothesis the same k_1 components with additional k_2 components. It transpires, however, that the application of testing routines in this context is often far from simple, since the classical theory on the asymptotic distribution of the likelihood-ratio statistics fails for testing whether a mixture of k_1+k_2 distributions of the same parametric form reduces to a mixture of k_1 distributions. The evaluation of the P-value of the likelihood ratio test can be carried out, however, by means of bootstrapping as suggested by McLachlan (1987).

From a Bayesian perspective, it is desirable to start with a mixture with a small number of components and increase the number so long as the fit is inadequate. The quality of the fit can be assessed by means of a posterior predictive distribution of a chosen test statistic (Gelman et al. 1995). For alternative MCMC algorithms, see Richardson and Green (1997) and Stephens (2000).

3. Conclusion

The use of mixture models has proliferated in many areas of application, facilitating the introduction of versatile modelling in areas where, until the end of the twentieth century, only simple models were considered. The shift from a single distribution to a mixture of distributions is associated typically with a marked increase in methodological complications. This has attracted a considerable amount of research which aimed at devising tractable methods for estimating the parameters of the model. Extensive bibliographies can be found in the books devoted to finite mixture distributions by Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Lindsay (1995), Böhning (1999), and McLachlan and Peel (2000).

Bibliography

Aitkin M 1999 A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**: 117–28

Andrews D F, Bickel P J, Hampel F R, Huber P J, Rogers W H, Tukey J W 1972 *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, NJ

Böhning D 1995 A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference* **47**: 5–28

Böhning D 1999 *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*. Chapman and Hall, Boca Raton, FL

Deb P, Trivedi P K 1997 Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* **12**: 313–36

Diebolt J, Robert C P 1998 Estimation of finite mixture distributions through Bayesian sampling. *Journal of Royal Statistical Society, Series B* **56**: 363–75

Everitt B S 1996 An introduction to finite mixture distributions. *Statistical Methods in Medical Research* **5**: 107–27

Everitt B S, Hand D J 1981 *Finite Mixture Distributions*. Chapman and Hall, London

Forman A K, Kohlmann T 1996 Latent class analysis in medical research. *Statistical Methods in Medical Research* **5**: 179–211

Gelman A, Carlin J B, Stern H S, Rubin D B 1995 *Bayesian Data Analysis*. Texts in Statistical Science Series. Chapman and Hall, London

Jasiulewicz H 1997 Application of mixture models to approximation of age-at-death distribution. *Insurance, Mathematics and Economics* **19**: 237–41

Kamakura W A, Wedel M 1997 Statistical data fusion for cross-tabulation. *Journal of Marketing Research* **34**: 485–98

Karmakar U S 1994 A robust forecasting technique for inventory and leadtime management. *Journal of Operations Management* **12**: 45–54

Li L, Choe M K 1997 A mixture model for duration data: Analysis of second births in China. *Demography* **34**: 189–97

Lindsay B G 1995 *Mixture Models: Theory, Geometry and Applications,* NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics and the American Statistical Association, Alexandria, VA

Lindsay B G, Lesperance M L 1995 A review of semiparametric mixture models. *Journal of Statistical Planning and Inference* **47**: 29–39

McLachlan G J 1987 On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**: 318–24

McLachlan G J 1992 *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York

McLachlan G J, Basford K E 1988 *Mixture Models. Inference and Applications to Clustering*. Marcel Dekker, New York

McLachlan G J, Krishnan T 1997 *The EM Algorithm and Extensions*. Wiley, New York

McLachlan G J, Peel D 2000 *Finite Mixture Models*. Wiley, New York
Newcomb S 1886 A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics* **8**: 343–66
Ng S K, McLachlan G J 1998 On modifications to the long-term survival mixture model in the presence of competing risks. *Journal of Statistics and Computer Simulation* **61**: 77–96
Pearson K 1894 Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London* **185**: 71–110
Peel D, McLachlan G J 2000 Robust mixture modelling using *t* distribution. *Statistical Computing* **10**: 339–48
Richardson S, Green P J 1997 On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B* **59**: 731–58
Robert C P, Casella G 1999 *Monte Carlo Statistical Methods*. Springer-Verlag, New York
Roeder K 1994 A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association* **89**: 487–95
Stephens M 2000 Bayesian analysis of mixture models with an unknown number of components, an alternative to reversible jump methods. *Ann. Statist.* **28**: 40–74
Titterington D M 1990 Some recent research in the analysis of mixture distributions. *Statistics* **21**: 619–41
Titterington D M, Smith A F M, Makov U E 1985 *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester, UK
Watanabe T 2000 Bayesian analysis of dynamic bivariate mixture models: Can they explain the behavior of returns and trading volume? *Journal of Business and Economic Statistics* **18**: 199–210
Yantis S, Meyer D E, Smith J E 1991 Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin* **110**: 350–74

U. E. Makov

Mnemonics, Psychology of

Mnemonics are memory aids. They have a long history of use, but a short history of research. This article examines a few verbal and visual mnemonic techniques, and three more general mnemonic systems.

1. *Definition and History*

The word *mnemonic* is derived from Mnemosyne, the name of the ancient Greek goddess of memory. It means 'aiding the memory,' and thus could refer generally to any method, technique, or system that improves memory. Typically, however, *mnemonic* refers more specifically to memory aids that most people consider to be rather unusual and/or artificial.

Mnemonics are sometimes referred to as memory 'tricks,' as if they were used in the place of 'real' memorizing. However, mnemonics do not *replace* the basic principles of learning and memory, but rather *use* the principles—such as meaningfulness, organization, association, and attention. One feature of most mnemonics is that they do not have a natural connection with the material to be learned; rather, they impose meaning or structure on material that is otherwise not inherently meaningful or organized (see *Memory: Organization and Recall*). In fact, meaningless, unorganized, or unconnected material is the kind of material for which mnemonics have best shown their value; mnemonics are not needed for material that already has meaning or logical structure. A second feature of most mnemonics is that they involve adding something to the material being learned, and thus are sometimes referred to as 'elaboration'; these elaborations create meaningful associations between what is to be learned and what is already known (see *Elaboration in Memory*). A third feature of most mnemonics is that they focus attention on relevant aspects of the material to be learned, and they can also foster attention by the learner because they tend to be more interesting and more fun than rote learning.

The use of mnemonics is not new, dating back to at least 500 BC (see Yates 1966). However, although people have used mnemonics for at least 2,500 years, almost all of the experimental research on mnemonics has been conducted over less than 40 years. The topic was of considerable interest during the 1800s and very early 1900s, but then for about the next half a century there was virtually no research done on mnemonics. One likely reason was the influence of behaviorism, especially in American psychology—mental processes were not considered to be a very legitimate area for research, as most psychologists concentrated on observable behavior in order to be scientific. A revived research interest in mnemonics in the 1960s accompanied the general acceptability of cognitive processes as a legitimate area for research (see *Cognitive Psychology: History*). A second possible reason why mnemonics might not have been viewed as a legitimate area for scientific inquiry for so many years is that they were (and sometimes still are) associated with sensationalism, showmanship, and commercialism. Researchers might have thought that research on mnemonics would yield little useful knowledge about memory, or thought that such 'gimmicks' were not worthy of serious scientific study.

Since the mid-1960s, there has been a large amount of research on mnemonics—enough to justify a book consisting of 20 review chapters by the mid-1980s (McDaniel and Pressley 1987), and for some recent cognitive psychology textbooks to devote increasing space to mnemonics (see, e.g., 'Mnemonics and experts,' Chap. 10 in Solso 1998). The experimental research has not produced many new mnemonics, but rather has focused mostly on variations of time-honored mnemonic systems, their effectiveness, and new applications for them. The research has found

that mnemonics such as those described in this article have been used effectively by a wide range of people for a wide range of memory tasks.

2. *Verbal and Visual Mnemonic Techniques*

Mnemonics can be either verbal or visual. Verbal mnemonics use words to associate the items to be remembered, while visual-imagery mnemonics make the associations with mental pictures. For example, the words *cat*s and *rats* could be associated either with a sentence such as 'Cats like to chase rats' or with a visual image of cats chasing rats.

2.1 *Verbal Mnemonics*

Examples of verbal mnemonics include such techniques as the first-letter mnemonic, the keyword mnemonic, rhymes and songs, and stories. Research has been done on all of these techniques, especially the first-letter mnemonic and the keyword mnemonic.

Acronyms and acrostics are often called *first-letter mnemonics*. An acronym is a word that is made from the first letters of the items to be remembered. For example, the word HOMES has been used to remember the names of the Great Lakes in the USA (Huron, Ontario, Michigan, Erie, Superior), and the fictional name ROY G. BIV to remember the names of the colors of the rainbow (red, orange, yellow, green, blue, indigo, violet). An acrostic is a series of words, lines, or verses in which the first letters form a word or phrase. For example, an acrostic to remember the names of the Great Lakes is, 'Healthy Old Men Exercise Some,' and to remember the colors of the rainbow, 'Richard of York Gave Battle in Vain.'

Concrete material is easier to remember than abstract material. Therefore, some abstract words can be remembered better by constructing meaningful concrete substitute words to represent the abstract words. This is the first step in the technique that is often called the *keyword mnemonic*. For example, an English speaker learning the Spanish word *pato* for the English word *duck* could use the sound-alike substitute word (keyword) *pot* to represent *pato*, or a person trying to remember the name *Higbee* could use the substitute term *hug bee*. The second step is for the learner to associate the substitute word with the English meaning (e.g., a duck wearing a pot on its head) or with the person (e.g., Mr. Higbee hugging a bee). This association is usually done using visual imagery, as described in the next section. The keyword mnemonic has been found to be effective in several kinds of paired-associate learning tasks in school, and in learning people's names and foreign language vocabulary. (It has been used in Great Britain by Michael Gruneberg (1995) to develop extensive programs for learning several different languages.)

Rhymes and *songs* have been used effectively to impose meaning and structure in relating unrelated items. For example, some people remember which months are the short ones by the rhyme that begins, 'Thirty days has September, April, June and November,' and remember how to spell words with 'ie' in them by the rhyme that begins, 'i before e, except after c.' Many children have learned the letters of the alphabet by singing them in a song. An extensive set of mnemonics using rhymes and songs was developed and used through most of the 1900s by Masachika Nakane in Japan to teach children mathematics, sciences, and languages; Nakane's mnemonics differ from most mnemonic techniques in that they help to teach processes and procedures for performing operations, rather than just facts (Higbee and Kunihira 1985).

Items to be remembered can also be tied together by incorporating them into *stories*. For example, to remember the names of the 13 original states of the United States, some schoolchildren have used a story that begins, 'A lady from *Delaware* bought a ticket on the *Pennsylvania* railroad. She packed a *new jersey* sweater in her suitcase, and went to visit her friend *Georgia* in *Connecticut*.' The story mnemonic can be combined with substitute words, as in a story for remembering the names of the cranial nerves, which begins, 'At the *oil factory* (olfactory) the *optician* (optic) looked for the *occupant* (oculomotor) of the *truck* (trochlear).'

2.2 *Visual-imagery Mnemonics*

Most mnemonic techniques and systems rely heavily on visual imagery, or mental pictures (see *Visual Memory, Psychology of*; *Mental Imagery, Psychology of*). Although there are various theories about why visual imagery can be such a powerful memory aid in learning verbal material, much research has found that it does help.

Some verbal mnemonics can also be represented as visual mnemonics by forming a visual image of the verbal association. Some of the examples in the previous section could be made even more effective for most people by picturing, for example, homes floating on the Great Lakes, or the lady from Delaware taking her trip. The keyword mnemonic is actually both a verbal and a visual mnemonic because, as noted previously, it usually incorporates visual imagery in its second step. All of the mnemonic systems in the next section also use visual imagery.

Three strategies that have been suggested to help make visual associations effective are interaction, vividness, and bizarreness: The learner should form an image of the associated items interacting with each other (not just sitting next to each other), should picture the image as clearly and vividly as possible, and should make the association bizarre or unusual.

All three of these strategies have received research support, with the support being stronger and more consistent for interaction and vividness than for bizarreness.

3. *Mnemonic Systems*

There are at least three differences that might make it useful to distinguish between mnemonic techniques, like those in the previous section, and mnemonic systems. First, techniques are usually single-purpose mnemonics used for remembering specific facts, whereas the mnemonic systems described in this section are more general-purpose mnemonics that can be used repeatedly to remember different sets of material. A second difference between techniques and systems is that for most techniques the material to be remembered also serves as part of the mnemonic to cue memory for other parts of the material, whereas mnemonic systems involve building a mental file of previously-memorized cues that exist independently of the material to be remembered. A third difference is that techniques are more likely to be discovered and used spontaneously by the learner, whereas systems are less likely to be self-generated, and are more likely to require instruction and practice.

Three of the most widely used and studied mnemonic systems are the loci system, the peg system, and the phonetic system. The loci and peg systems have received a large amount of research interest and support, but less research has been done on the phonetic system because of its complexity.

3.1 *Loci System*

The loci system is the most ancient mnemonic system, dating back to about 500 BC, when it was used by Greek orators to remember long speeches without notes. It was *the* mnemonic system until about the middle of the seventeenth century, when other systems began to evolve.

The word *loci* is the plural of *locus*, which means 'place or location.' The loci system consists of two steps. The first step is to thoroughly memorize a series of visual images of familiar locations in some natural or logical order. The locations could be rooms in a house or other familiar building, a series of buildings or other sites on a familiar walk, or even a familiar golf course; some versions have even used parts of the body or parts of an automobile. This series of locations is the mental filing system, which can be used over again for different lists of items. The second step is to associate a visual image of each item to be remembered (substitute words can be used for abstract items) with a location in the series; this is done by placing the items visually, in the order they are to be remembered, one in each location as the learner takes an imaginary walk past the locations. For retrieval, the learner takes a mental trip through the locations and retrieves the item that is associated with each location in order.

3.2 *Peg System*

Direct retrieval of an item at a certain position in a memorized list (for example, the eighth item) is difficult with the loci system, because it is dependent on sequential retrieval—finding the eighth item requires proceeding in order through the mental locations until the eighth location is reached. The peg system consists of a series of prememorized concrete nouns (pegwords) that correspond meaningfully with numbers. Thus, to retrieve the eighth item directly, the learner would just think of the pegword that represents the number 8 and see what item was associated with it.

The peg system can be traced to the mid-1600s, when Henry Herdson developed an extension of the loci system. Herdson dispensed with the spatial locations of the objects and merely used the objects themselves. Each digit was represented by any one of several objects that look like the numbers (e.g., 1 = candle, 8 = spectacles). A system that used rhyming words to represent the numbers was introduced in England in the late 1800s. The nouns rhymed with the numbers they represented so that it was easy to remember what noun represents each number. This is the basis of a current widely-used version of the peg system in which numbers are represented by rhyming pegwords (one = bun, two = shoe, three = tree, etc.). In addition to look-alikes or rhymes, pegwords have also been selected on the basis of a related meaning (one = penny, two = twins, three = tricycle, etc.).

The peg system gets its name from the fact that the pegwords serve as mental pegs or hooks on which the person 'hangs' the items to be remembered. The pegwords are used in the same manner that locations are used in the loci system. The learner associates a visual image of each new item with the image of each pegword in order, and retrieves the items by counting through the pegwords.

3.3 *Phonetic System*

The phonetic system is the most sophisticated and most versatile of the three mnemonic systems, and is the one used most frequently by professional mnemonists and other performers to perform amazing memory feats (cf. Wilding and Valentine 1997). It is also the most complex system, and thus requires the most study and effort to master. However, the phonetic system overcomes a limitation of the peg system by allowing for construction of more than 10 to 20 pegwords (for example, it can be difficult to find words that rhyme with, or look like, 27 or 31), while retaining the peg system's advantage of direct retrieval.

The phonetic system has also been referred to by such terms as figure-alphabet, digit-letter, number-alphabet, hook, number-consonant, and number-to-sound. Its origin has been traced back to 1648, when Winckelman (spelled differently in some sources) introduced a digit-letter system in which the digits were represented by letters of the alphabet. These letters were then used to form words to represent a given number sequence. In early versions of the system the digits were represented by both consonants and vowels, and the letters representing each digit were selected arbitrarily. In subsequent refinements of the system, the digits were represented by consonants only, and the consonants representing each digit were not selected arbitrarily but were selected on the basis of some relationship with the digits they represented (for example 1 = t, which has one downstroke). By the end of the nineteenth century the system had evolved into its present form—the digits were represented not by consonants themselves but by consonant *sounds* (which is why it is referred to here as the 'phonetic' system).

In the phonetic system, each digit from 0 to 9 is represented by a consonant sound or family of similar sounds (for example, 1 = t or d, 2 = n, 3 = m). These consonant sounds are then combined with vowels—which have no numerical value—to code numbers into keywords that represent the numbers (for example, 1 = tie or doe, 2 = no or inn, 12 = tin or dine, 31 = mat or maid). These keywords are memorized, and used as a mental filing system in the same manner that is described above for the peg system's pegwords.

In addition to its use as a mental filing system similar to the loci system and peg system, the phonetic system can also be used to remember numbers by converting them to words (which are more meaningful than numbers for most people). For example, the number 321 could be coded as *minute* or *my net*, and the number 2131 could be coded as *not mad* or as the acrostic '*never take more drugs*.'

This article has provided a comprehensive, but not exhaustive, overview of verbal as well as visual mnemonics, and of mnemonic techniques as well as mnemonic systems. Interest in mnemonics is likely to continue for researchers trying to understand memory better and especially for learners trying to remember better.

See also: Elaboration in Memory; Encoding Specificity in Memory; Expert Memory, Psychology of; External Memory, Psychology of; Memory Retrieval

Bibliography

Bellezza F S 1996 Mnemonic methods to enhance storage and retrieval. In: Bjork E L, Bjork R A (eds.) *Memory. Handbook of Perception and Cognition*, 2nd edn. Academic Press, San Diego, CA

Cornoldi C 1988 Why study mnemonics. In: Gruneberg M M, Morris P E, Sykes R N (eds.) *Practical Aspects of Memory: Vol. 2, Clinical and Educational Implications*. Wiley, Chichester, UK

De Beni R 1984 La ricerca sperimentale sulle mnemotechniche: Una rassegna. (Experimental research on mnemonics: A review). *Girnale Italiano di Psicologia* **11**: 421–56

Gruneberg M 1995 *Linkword Spanish, French, German, Italian*. NTC/Contemporary Books, Lincolnwood, IL

Higbee K L 1979 Recent research on visual mnemonics: Historical roots and educational fruits. *Review of Educational Research* **49**: 611–29

Higbee K L 1996 *Your Memory: How it Works and How to Improve It*, 2nd edn. Marlowe, New York

Higbee K L, Kunihira S 1985 Cross-cultural applications of Yodai mnemonics. *Educational Psychologist* **20**: 57–64

Levin J R 1993 Mnemonic strategies and classroom learning: A twenty-year report card. *The Elementary School Journal* **94**: 235–44

Manalo E 1997 Applications of mnemonics in education: A brief review. In: Habermann G M (ed.) *Looking Back and Moving Forward: 50 Years of New Zealand Psychology*. New Zealand Psychological Society, Wellington, New Zealand

McDaniel M A, Pressley M 1987 *Imagery and Related Mnemonic Processes*. Springer-Verlag, New York

Miller G A, Galanter E, Pribram K H 1960 *Plans and the Structure of Behavior*. Holt, New York

Paivio A 1971 *Imagery and Verbal Processes*. Holt, Rinehart & Winston, New York

Solso R L 1998 *Cognitive Psychology*, 5th edn. Allyn and Bacon, Boston, MA

Wilding J, Valentine E 1997 *Superior Memories*. Psychology Press, Hove, UK

Yates F 1966 *The Art of Memory*. Routledge & Kegan Paul, London

K. L. Higbee

Mobility: Social

Social mobility is the movement in time of social units between different positions in the system of social stratification of a society. The moving units are usually individuals, but may also be groups (families or kinship units). Society is seen as comprising at any given time an ordered system of more or less fixed positions that individuals or groups can hold and potentially move between. The system is stratified (a) to the extent that the positions are more or less advantageous and provide their holders with unequal power, material or symbolic means, and privilege, (b) to the extent that individuals or groups have different chances of access to, or movement between, the positions of unequal advantage. This second aspect represents the core domain in the study of social mobility.

Conventionally, different forms of mobility are studied: horizontal, upward, and downward mobility refer to moves among positions at the same level of advantage, moves up and down along positions

ordered in a hierarchy of advantage. Intergenerational mobility compares individuals' positions in adult life with the positions of their families of origin. Intragenerational (or career) mobility studies position sequences within individuals' own generations, for instance the moves between the successive jobs in working life. Other distinctions frequently made contrast collective vs. individual mobility, structural vs. exchange mobility or absolute rates vs. relative rates of mobility.

1. Research Questions and Traditions of Study

Classical authors have primarily seen social mobility as a mechanism of class formation or status group formation. The basic assumption is that classes, status groups, or other collectivities with some form of sociocultural or sociopolitical identity can only emerge and potentially provide a basis for collective action on condition that they achieve a sufficient degree of 'demographic identity' (Goldthorpe) in their membership, that is, when they include a kernel of members homogeneous in terms of their social origin and stable over time (families across generations or individuals through their working lives). Along these lines, Marx—whose main concern with mobility is otherwise with collective downward mobility of small proprietors and other middle classes into the proletariat—considers individual mobility as a crucial determinant of working class organization in his discussion of the North American class structure. Contrary to Europe, classes in the USA, he argues (much as Sombart does later) , 'have not yet become fixed, but continually change and interchange their elements in a constant flux' (Marx [1852] 1958, p. 225), with the effect of weakening class formation in the working class. Also stressing the connection to social mobility, Max Weber even defines a social class as an ensemble of those class situations among which high interchange of individuals typically takes place, while different classes are distinguished from each other by a lack of mobility between them. Weber's mobility criterion has been used to empirically establish class boundaries by later researchers.

In its later development, research on social mobility has largely turned to the study of social opportunity, aiming to determine the relative fluidity or rigidity of the stratification system concerning access to positions of varying advantage (see *Equality of Opportunity*). Research has in particular concentrated on identifying the degree to which individuals' prospects in life are determined by the social conditions of their family of origin and on specifying the individual, institutional, and societal factors responsible for it. The more strongly the positions persons hold are conditioned by their social provenance, the more rigid and closed is a stratification system; the greater the mobility between social origin and a persons' own position, the more open or fluid that system is.

Issues related to social mobility are often the object of normative and political discourse and are then often interpreted ideologically, for instance as ways of legitimizing social inequalities. In this vein, variants of liberal political philosophy from John Stuart Mill onwards regard unequal conditions as legitimate when they result from (fair) competition or when unequal chances of access to privileged positions are due to unequal achievements. In its extreme form, the functionalist theory of stratification (see *Social Stratification*) holds unequal rewards to be deserved, as they are assumed to be distributed according to the contributions made to the efficient functioning (and ultimately survival) of society and are thought to be necessary to attract the most able and hard-working individuals to the most important positions (see Goldthorpe 1996 for a forceful critique of meritocratic and functionalist assumptions). Though at times influenced by normative preferences, research in the field is analytical and empirical. Its results may, however, serve to examine to what degree and for what reasons social reality conforms to or falls short of particular normative or political goals.

The investigation of social mobility is mainly pursued by way of observation and comparison of national populations. Comparative designs are essential to determine the extent to which fluidity or rigidity of the stratification system is related to other characteristics of societies such as their degree of modernity, their economic and political order (e.g. capitalist market economies vs. state command economies), particular institutions or policies. Research has especially focused on the evolution of social mobility in line with the development of industrial society. While in his early work Sorokin (1927) expected 'trendless fluctuation' rather than a development in a specific direction, later researchers have assumed a historical increase in mobility with the transition from preindustrial to industrial society, either as a precondition for the take-off of industrial development or as its consequence (Lipset and Zetterberg 1959). Others assume a steady increase in mobility as industrialism advances, and an increasing uniformity and convergence of social structures and social mobility in industrial societies, due to standardizing tendencies inherent in advanced technologies and other elements of industrialism. Finally, exceptionally high rates of mobility have been postulated for specific societies, particularly those not burdened by the legacy of the European feudal past or its class-based stratification system, such as the USA, Australia, or Japan. To test such hypotheses, numerous datasets in many different societies have been created, primarily through large-scale population surveys, but also (notably in studies on earlier periods) by using suitable information from official records. As members of social elites tend to be represented in extremely small

numbers only in population surveys, mobility into elite positions is generally not covered (see *Elites: Sociological Aspects*).

For the purposes of structured summary, much of the sociological literature can be represented as a debate between two different paradigms and research traditions. To study mobility, both essentially rely on work positions, that is, the (occupational) jobs individuals hold to earn a living. The status attainment approach, on the one side, sees the principle property of these jobs in terms of their respective location in a finely graduated continuum of rewards related to the position and conveyed to their incumbents. The continuum of rewards is generally conceived in terms of social prestige or socioeconomic status and measured by corresponding scales (see *Social Stratification*).

The class mobility approach, on the other side, conceives the (structure of) positions among which mobility can occur as discrete classes which differ from each other by properties of a qualitatively different nature rather than only along a vertical dimension. In earlier studies, categories consisted of rather broad groupings of occupations defined in an ad hoc manner (sometimes as broadly as agricultural/manual/nonmanual). Newer class schemas are grounded in theoretically comprehensive conceptions of the class structure. In their now widely accepted class schema, Erikson and Goldthorpe (1992) draw on the character of employment relations as core criteria to identify different classes, distinguishing first between classes whose incumbents buy the labor of others (employer), sell their own labor (employee) or do neither (self-employed); among employees, further distinctions are made between different classes of (manual) workers, routine nonmanual employees, and of professionals, administrators, and managers in the so-called service class or salariat, the criterion being the extent to which employment relations are regulated according to a labor contract or a service relationship. Related to the differing conception of the main properties of the positions among which mobility is seen to occur, the two traditions also differ in their theoretical understanding of the mobility process, the methods used, and the main conclusions arrived at.

2. *The Process of Status Attainment*

Research pursued in the status attainment framework is often—though not necessarily—embedded in a more general (liberal) theory of industrialism and shares with it assumptions about the impacts of industrialization on other aspects of social structure, notably social mobility. Industrialization is understood as a pervasive force that sooner or later affects all national societies across the globe. In this process, it is argued, the order of stratification in preindustrial societies in which ascription by family, kin, ethnic ties, gender or status group affiliation determines the position of individuals in society, is increasingly replaced by a new order in which achievement based on individual abilities, qualifications, and work accomplishments play the decisive role in social selection (Blau and Duncan 1967, Treiman 1970). A fundamental trend toward rationalism and universalism is seen to unfold. To survive in markets actors will be forced by the technological sophistication of production processes and by growing market competition to increasingly allocate individuals 'to positions according to what they have shown they can do, rather than according to their social provenance' (Goldthorpe 1996, p. 256). Notably formal educational qualifications are expected to increasingly determine the job level (successively) reached in bureaucratic work organizations. To enhance the development, states expand educational services and subsidize the costs of education, thus providing easier access to education for children in less fortunate economic conditions. As industrialization progresses, social placement of individuals will then be conditioned less by their social origin, and intergenerational social mobility between the positions of parents and those of their children is expected to increase.

These general hypotheses lead to an understanding of social mobility as a process of successive status attainment by individuals, in which later attainment depends on earlier attainment. In line with the conception of the occupational structure as a graded hierarchy, this process is studied with variants of linear regression models (often path or structural equations models) to determine the extent to which the social status or prestige level of a person's job at a given point in his or her career depends on earlier job achievements, educational qualifications, and directly or indirectly on parents' status. In Blau and Duncan's (1967) seminal study *The American Occupational Structure*, education is identified as the main determinant of individuals' occupational careers and as the core link between social origins and attained position. Parental education and occupational status also affect children's occupational status or incomes, but their effects are weaker than those of education and are mainly mediated through education. One far-reaching conclusion has been that under such conditions of status attainment—powered by individual achievement—the very basis for class solidarity and class action disappears. Extensions of the model have studied how other family conditions, personal abilities (intelligence), sociopsychological factors (parental encouragement and support, educational and occupational aspirations, peer contacts) or characteristics of schools intervene in these processes and contribute to generating them (Sewell and Hauser 1975). In evidence from a number of societies at different levels of industrial development, Treiman and Yip (1989) find confirmed the expectation that status attainment

becomes increasingly achievement-based as industrialization unfolds.

The status attainment paradigm has come in for criticism on a number of counts. Its representation of the attainment process is individualistic and one-sidedly supply-driven, concentrating as it does on the way familial resources and individual human capital affect occupational and other outcomes. But the transformation of individual resources into outcomes depends on existing opportunities. The latter are constrained both by the distribution of available positions and by the competition of those contending for the same positions with similar or better resources. Further, the unidimensional representation of positions in terms of status or prestige obscures the way in which opportunities are bound up with structural transformations such as the shifts in forms of employment and between sectors, industries, and occupations. Groups like farmers, skilled industrial workers, small shopkeepers, and service workers hold positions that are heterogeneous in terms of their sociostructural location and affected in different ways by particular state policies and structural transformations. However, in terms of social status or prestige, they all are similarly rated. In the hierarchical framework it thus appears difficult, if not impossible, to trace mobility processes to their causal conditions.

3. *Intergenerational Class Mobility and Social Fluidity*

Proponents of the class framework have attempted to overcome these shortcomings by reorienting the study of mobility to analyze the moves of individuals between a set of social class positions identified within a class structure. These moves are represented in mobility matrices, which, in their elementary form, are bivariate frequency tables with one dimension representing the set of origin classes, the other dimension the set of destination classes, and the cells of the matrix containing the frequencies of transitions observed between origin and destination. While origin and destination can be any two successive positions (for instance in working life), intergenerational comparisons between the class positions of parents and those of their children are of primary interest. Several important strengths result from this reorientation. Compared to hierarchical ratings, the class characteristics allow the derivation of more realistic hypotheses concerning the greater or lesser likelihood of mobility between particular origins and destinations. Structural transformations and changing opportunity structures can be more directly captured with class categories sensitive in this respect. The concomitant development of log-linear models (for early contributions see Goodman 1970, Hauser 1979) has brought crucial improvements in the ability to account for opportunity structures in the statistical analysis of mobility matrices.

In these models the constituent relations in a mobility table are the relative rates of mobility. These are basically a set of odds ratios which measure the relativities existing between different origin classes and the attainment of different destination classes. Odds ratios can be interpreted as measures of the competitive advantages the progeny of one origin class have over those of another origin class in obtaining or avoiding a position in one rather than in another destination class. Odds ratios have the crucial property of being independent of the marginal distributions of the mobility matrix (the relative size of origin classes and destination classes). The latter reflect the macrostructural changes between generations in positions available and can be understood as capturing the opportunity structure. The odds ratios thus express the pattern of association between origin and destination independently of a given opportunity structure. However, as Sørensen (1977) and others have stressed, opportunity structures also depend on how vacant positions available in the market are matched to individual workers, e.g., by creating more or less extended chains of vacancies within and between firms. But still it is the relative rates of mobility—expressed by odds ratios—which are now widely accepted as a proper measure of the degree of fluidity or rigidity of the class structure or the stratification system.

Assessed in terms of these relative rates, class destinations in mature industrial societies continue to depend heavily on class origins. In England or France, for instance, the odds of a son born to a service class father obtaining a service class position rather than one in the unskilled working class is more than 20 times larger than the corresponding odds for an unskilled worker's son. Mobility is more likely to occur for intermediate classes, for example, routine nonmanual employees, while the likelihood of staying immobile in the parental class are larger in the most and least advantageous classes. Class positions that require the ownership of material or financial capital or land are 'inherited' to a clearly larger extent from the parental to the offspring generation than classes not requiring such assets. Short-range mobility occurs more often than long-range mobility. While there is still debate as to whether in the long run social fluidity has increased (most persistently defended by Ganzeboom et al. 1989), most research indicates that there is a high degree of temporal constancy and cross-national commonality in the fluidity pattern of industrial and postindustrial societies.

Explanations of the common characteristics of social fluidity have been proposed in mainly two ways. Researchers emphasizing the hierarchical dimension have stressed the distances (in status or other considered dimensions) between positions as the mobility generating principle (e.g., Hout and Hauser 1992).

The larger the vertical distance between positions, the less mobility there will be between them. Sophisticated distance models have consistently shown considerable distance effects, but they cannot entirely explain the existing pattern. Researchers in the class framework have acknowledged the multidimensional nature of the resources that are available to the offspring of various class origins and that can be used for gaining access into classes of varying advantage and desirability.

In their major comparative study of a broad range of industrial societies, Erikson and Goldthorpe (1992) show that alongside the impacts of vertical 'hierarchy' the fluidity pattern is markedly shaped by 'inheritance' of parental class position (in all classes, but especially in the petty bourgeoisie, among farmers, and in the service class), by the 'sector' division between employment in agriculture and outside agriculture, and by special affinities expressing enhanced mobility among particular classes like the various nonmanual classes. To account for the high stability across time and the marked similarity between countries, Goldthorpe (2000) proposes a general explanation in terms of resource-constrained rational action, much in line with the theory suggested by Boudon (1974). Members of the various social classes pursue the same aims, giving priority to avoiding downward mobility and only in the second place seeking to achieve upward mobility. But in pursuing these same goals and priorities, the different amounts and kinds of resources typically available prompt the working and intermediate classes to use different strategies than the service class. The cumulative effect of different starting positions, resources, and strategies is so pertinent that it produces high stability in the resulting outcomes. If inequalities of opportunity are to decline and social fluidity to increase, it is most likely to occur in a context of markedly reduced inequalities of condition.

High degrees of cross-national commonality and constancy over time, however, are not to be equated with complete uniformity. Some limited, but nevertheless significant differences are found between societies, and change is also found within some societies across time. Erikson and Goldthorpe thus propose accepting, albeit only in a weak form, the hypothesis originally proposed by Featherman et al. (1975), that 'basically the same' relative rates of social mobility exist in all industrial societies with a nuclear family system and a market economy. As changes in time do not seem to form a systematic trend (associated with industrial or some other development), results appear to favor Sorokin's assumption of 'trendless fluctuation'. There is also evidence against the hypothesis of exceptionally high fluidity in the 'new' societies (USA, Australia) and Japan. Nor does social fluidity systematically vary according to the political-economic order of countries. The former communist Eastern European societies show quite similar fluidity patterns to capitalist societies. Instead, deviations from the core pattern appear to arise more from historically contingent conditions resulting from a nation's particular economic, social, and political history or from specific institutional arrangements, for instance in the educational system, in labor market segmentation or in other institutions with an impact on specific elements of the mobility pattern. Periodically higher relative rates of mobility have also been shown to result from massive state intervention, as in the case of the long-term equalizing policies by the social democratic governments in Sweden (Erikson and Jonsson 1996) or the strong discriminatory measures pro and contra particular classes in some phases of the communist era in Hungary.

In contrast to relative rates, absolute rates of mobility refer to the observed absolute number of moves that can be read directly from the mobility matrix, and to proportions such as the percentage of children who have a different (percentage mobile) or the same (percentage immobile) class position as their parents. Other often studied absolute rates are outflow proportions (percentages of offspring of a given origin attaining particular destinations) or inflow proportions (percentages of members in a given destination recruited from the various origin classes). Even though absolute rates are more directly observed, theoretically they are conceived to be implied by the more fundamental conditions of social fluidity and by the opportunity structures captured in the marginal distributions. Under a given set of odds ratios, particular marginal distributions imply a unique constellation of absolute rates. (The distinction between relative rates of mobility and absolute rates has replaced the earlier distinction between exchange mobility and structural mobility, which had similar intentions but was statistically and theoretically unsatisfactory.)

In contrast to relative rates, absolute rates vary considerably between nations and across time. Given only little variation in relative rates, it is mainly variation in the opportunity structure that generates this outcome. In a world of economic and occupational change, mobility is induced as workers are pushed or 'forced' to leave areas of declining employment and gravitate to those which are expanding. Societies differ in the absolute amount of mobility both among themselves and across time because they vary in the pace of change and because differences in class composition exist even among countries with similar levels of development. The absolute numbers of such moves determine the observable mobility experience of a population. They also strongly affect the composition of class membership by persons intergenerationally stable or recruited from other class origins. Most likely they are of particular importance in processes of class formation. Though these issues are of considerable interest, they have been hardly studied.

Similar relative rates alongside varying absolute rates are also found if different population subgroups

are compared. Large differences between men and women, immigrants and natives, or between different races or ethnic groups often exist in absolute rates of mobility into more or less advantageous classes because the groups are constrained to or choose jobs in different segments of the labor market and hence their distribution among classes varies. But the patterns of social fluidity hardly differ between the groups. If they sometimes do, differences tend to be small. As a result of gender-segregated labor markets, for instance, men's and women's absolute chances of access into different classes strongly differ, but these inequalities are generally not intensified by men and women being affected differently by their class origins. As a confirming corollary it is found that in times when married women's labor market participation was low, social fluidity for daughters in the marriage markets has been rather similar to the relative mobility prospects of their brothers in the labor market. Among the highly debated issues concerning gender inequality one important question relates to the proper unit of stratification, class, and mobility analysis. Is it (always) women's or men's own position in the labor market that determines their social status or class affiliation, or is it rather the common position of the family unit? As women's labor force participation has increased and has become more continuous, the most convincing answer is to study the issues separately, to analyze men's and women's labor market careers and social mobility separately, likewise their household division of labor and resource utilization, but also the common position they may share as a family unit.

In the figures on mobility or immobility between social origin and eventual destination, whether viewed in absolute or relative terms, the process of intergenerational social reproduction is only reflected in its global outcomes. It results from various intervening processes, notably educational attainment, transition from education to work and mobility between successive jobs in working life within and between firms. These processes have to be—and are—studied in their own right (see *Life Course: Sociological Aspects*). Of these, the role of education has been scrutinized most intensively. The fact that education is a prerequisite for access to professional and managerial jobs makes it an important channel for upward social mobility, but evidently one which the more advantaged families also make use of to secure favorable positions for the next generation. Further, while education is found to be the main carrier responsible for the hierarchical element in the fluidity pattern, this is much less true—in some cases not true at all—of its role as a mediator of other elements or mechanisms that produce immobility (Ishida et al. 1995). Transfer of capital or business, provision of social contacts, or inculcation of the proper social skills and manners are some of the means prosperous families have to back their children outside school and which can be used to compensate for possible failures in education.

Studies pursued in the class framework so far have not found evidence consistently supporting expectations of a general trend towards increased education-based or 'credentialist' social selection in all industrial or postindustrial societies, neither in the sense of education becoming systematically less dependent on social origins nor in the sense of education consistently becoming a stronger predictor of social destination. As is true for intergenerational social fluidity in general, changes that actually take place in some countries appear to be limited in extent, and the direction they take depends on contingent circumstances specific to these countries.

See also: Class: Social; Equality and Inequality: Legal Aspects; Inequality: Comparative Aspects; Social Mobility, History of; Status and Role: Structural Aspects

Bibliography

Blau P M, Duncan O D 1967 *The American Occupational Structure*. Wiley, New York

Boudon R 1974 *Education, Opportunity and Social Inequality*. Wiley, New York

Erikson R, Goldthorpe J H 1992 *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Clarendon Press, Oxford

Erikson R, Jonsson J O (eds.) 1996 *Can Education be Equalized? The Swedish Case in Comparative Perspective*. Westview, Boulder, CO

Featherman D L, Jones L, Hauser R M 1975 Assumptions of social mobility research in the US: The case of cccupational status. *Social Science Research* **4**: 329–60

Ganzeboom H G B, Luijkx R, Treiman D J 1989 Intergenerational class mobility in comparative perspective. *Research in Social Stratification and Mobility* **8**: 3–55

Goldthorpe J H 1996 Problems of 'meritocracy.' In: Erikson R, Jonsson J O (eds.) *Can Education be Equalized? The Swedish Case in Comparative Perspective*. Westview Press, Boulder, CO, pp. 255–87

Goldthorpe J H 2000 *On Sociology: Numbers, Narratives, and the Integration of Research and Theory*. Oxford University Press, Oxford

Goodman L A 1970 The multiplicative analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association* **65**: 226–56

Hauser R M 1979 Some exploratory methods for modeling mobility tables and other cross-classified data. *Sociological Methodology* **10**: 141–58

Hout M, Hauser R M 1992 Symmetry and hierarchy in social mobility: A methodological analysis of the CASMIN model of class mobility. *European Sociological Review* **8**: 239–66

Ishida H, Müller W, Ridge J M 1995 Class origin, class destination, and education: A cross-national study of ten industrial nations. *American Journal of Sociology* **60**: 145–93

Lipset S M, Zetterberg H L 1959 Social mobility in industrial societies. In: Lipset S M, Bendix R (eds.) *Social Mobility in Industrial Societies*. University of California Press, Berkeley, CA, pp. 11–75

Marx K [1852] 1958 The eighteenth brumaire of Louis Bonaparte. In: Marx K, Engels F (eds.) *Selected Works*. Foreign Language Publications, Moscow

Sewell W H, Hauser R M 1975 *Education, Occupation, and Earnings: Achievement in the Early Career*. Academic Press, New York
Sørensen A B 1977 The structure of inequality and the process of attainment. *American Journal of Sociology* **42**: 965–78
Sorokin P A 1927 *Cultural Mobility*. Free Press, New York
Treiman D J 1970 Industrialization and social stratification. In: Laumann E O (ed.) *Social Stratification: Research and Theory for the 1970s*. Bobbs Merrill, Indianapolis, IN, pp. 207–34
Treiman D J, Yip K-B 1989 Educational and occupational attainment in 21 countries. In: Kohn M L (ed.) *Cross-National Research in Sociology*. Sage, Newbury Park, CA, pp. 373–94

W. Müller

Mode of Production

Karl Marx's term *Produktionsweise* is usually translated into English as 'mode of production,' but it might be more simply rendered as 'way of producing.' Never precisely defined in Marx's work—yet arguably at the center of it—*Produktionsweise* refers to both a form or structure and a way of producing. One might say that Marx's object in *Capital* was to show how the first—a particular structure of unequal power and wealth—was both the cause and the result of the second—men and women pursuing their individual ends under capitalist constraints.

One of the notable aspects of the capitalist mode of production is, according to Marx, its inherent instability. As individuals pursue their projects, there comes a time when the capitalist system can no longer be reproduced. At that point, revolution appears on the horizon, and the entire mode of production will eventually be transformed—capitalism being replaced by socialism, socialism by communism.

Even though Marx referred at times to other modes of production—primitive communist, the ancient, Asiatic, feudal, socialist, and finally modern communist—his only developed work was on capitalism. On what would come after capitalism, he offered only the vaguest of outlines; it would fall to Lenin and others to construct actually existing socialism. And on what came 'before,' Marx sometimes reflected only the European prejudices of the nineteenth century. He thought the Asiatic mode of production was inherently static and would waken to change only through the effects of Western imperialism.

The grand evolutionary teleology in Marx's view of history—the circular but progressive spiral from primitive to modern communism—accounts for much of the political appeal of Marx's ideas. Ironically, that appeal was felt mainly in countries on the periphery of capitalist development; in such locations, the notion of the stages of history appeared to offer a road map to a better future, a way to skip ahead. But this aspect of Marx's work also accounts for its ambivalent reception among social scientists who, even if they acknowledged an inevitable political component in their work, looked to the court of evidence as the final arbiter of knowledge.

As Marxism began to provide the legitimating ideology of socialist states in the twentieth century, it tended to stagnate there. In the capitalist West, however, Marx's ideas received critique and development, principally during periods of crisis in the world capitalist system, namely during the Great Depression of the 1930s and again during the late 1960s and 1970s. During the second period in particular, the notion of mode of production became the subject of wide-ranging discussion.

In the field of anthropology this was particularly true in France where, under the influence of structuralist philosopher Louis Althusser, a wide-ranging conversation took place among scholars such as Claude Meillassoux, Pierre-Philippe Rey, Maurice Godelier, and Emmanuel Terray. Below, I attempt to summarize the lasting legacy of this conversation. I do so, not in the structuralist vocabulary of the time, but in reference to what I regard as the clarifying, later work of British analytical philosopher G. A. Cohen.

What, then, is a mode of production? Perhaps the simplest answer is that it is a model of certain basic regularities that hold across and therefore define historical epochs. At the heart of these regularities is the interrelationship between technology in some broad sense and social stratification. According to Marx, the first—*Produktivkräfte* or productive forces—determines the second—*Produktionsverhältnisse* or relations of production. These two constitute the base of a mode of production, which, in turn, determines the superstructure.

1. Productive Forces

Productive forces are anything that can be used in productive interaction with nature. On this reading, productive forces are not simply raw materials or tools; more inclusively, they are human skills, productive knowledge, and even the technical aspects of human cooperation in the labor process. These factors Marx considered as equally 'material' in that they can all be utilized to produce. But most current readers would probably classify objects, knowledge, and the technical aspects of cooperation as variously material, ideal, and social.

The most striking aspect of productive forces has been their tendency to expand over human history. In the most general sense, the anterior cause of this tendency appears to be located in human nature itself. Humans everywhere develop their powers and skills in a continuous dialectic with nature, and presented with a choice between more and less efficient means of

production, humans everywhere (with certain limited exceptions) choose the first.

If productive forces have tended to expand, this fact can only be established at a world level. Some societies have stagnated; some have even experienced local declines in the level of productive powers. As will become clear below, relations of production can prevent technical development beyond a certain point. For any particular society in stasis, there is no reason to believe that new social relations compatible with the further expansion of forces must develop. In fact, technical knowledge tends to diffuse from place to place, and typically the most 'advanced' societies are not the ones that give rise to the next development. Rather, peripheral groups that are the recipient of the productive knowledge of their more 'advanced' neighbors often leapfrog ahead.

Just what determines the 'level' of development of productive forces should be considered. A society's mass of productive forces is determined by two factors: first, average productivity of labor, and second, population. The product of these two factors yields the level of productive forces. In this formulation, population itself is not a productive force, yet it enters into the determination of the level of productive powers.

2. *Relations of Production*

Relations of production are not, as the English phrase suggests, simply the social relationships formed in the process of production. Rather, they are the *de facto* power relationships that both underlie and are the result of the division of the fruits of a society's total labor. Household head vs. dependents, chief vs. subjects, master vs. slaves, feudal lord vs. peasants, capitalist vs. workers—relations of production are basic asymmetries of power grounded in the organization of material life. In capitalism, relations of production rest upon actual control over productive forces in the process of production, but in other modes of production, this is not necessarily so.

In relation to the last point, consider the contrast between capitalism and precolonial chiefdoms in Africa. Capitalists put the production process into motion; they or their representatives oversee and supervise in order to ensure that a profit is produced at the end of the work cycle. In many chiefdoms, chiefs did not control the production process itself. Yet, having produced, subjects brought tribute, the fruits of their own labors, to the chief. This contrast illustrates the fact that relations of production definitionally correspond to the basic structures of power in a society—however that power is constituted, whether by economic, coercive, or religious means (or some mix of these). As will become clear below, it is the superstructure that clothes and stabilizes power in particular cultural and organizations forms.

Why do productive inequalities occupy a central place in Marx's thought? The answer is that they locate the basic divisions within any society, the lines of potential opposition—of contradiction. Marx saw these as the potential fault lines along which tensions tend to build up, are routinely dissipated by small readjustments, and are sometimes violently resolved by radical realignments. These fault lines are structural; they do not necessarily lead to actual struggle and conflict (indeed, the function of the superstructure is precisely to prevent such occurrences). Nevertheless, contradictions always exist as potentialities; they lie just below the surface.

3. *Forces Determine Relations*

Ordinarily, if a factor X is said to determine factor Y, then X is taken to cause Y. Given X, Y is uniquely specified. Transferred to Marx's scheme above, this understanding of determination yields what has been called vulgar Marxism. Such an interpretation can be supported by some of Marx's own, more incautious summary statements: 'The handmill gives you society with the feudal lord; the steam mill society with the industrial capitalist.' Empirically it is clear that the same technology can exist in different modes of production.

A great deal of effort has been expended by commentators to explicate Marx's notion of determination—from examination of dialectics to, in Engels' phrase, determination only 'in the last instance.' Probably the best way to clarify Marx's claims is to note that they are forms of functional, not causal explanation. Immediately, a distinction needs to drawn between functional explanations and the theoretical school known as functionalism in sociology and anthropology: Marx used functionalist explanations, but he was not a 'functionalist' in the second sense.

Functional explanations have a distinct logical form in which the effects of a trait enter into the explanation of the presence of that trait: for example, when Y's presence is partly explained by some positive consequence Y has for X. In such a case, X can be said to determine or 'select' Y because of Y's effect on X.

Marx believed that the level of productive forces selects relations of production. For any given level of development of forces, there is only a limited set of relations possible—namely, those that are compatible with and in fact promote the further development of forces. Here, X hardly causes Y in a one-way pattern of determination. Rather, the reverse is closer to the truth: it is Y's effect on X that are paramount. Particular relations cause expansion in the forces. Herein lies the rub, however. As forces expand, they reach a level that is no longer compatible with the existing relations of production. At that point, according to Marx, revolution becomes a possibility. In

this way, continuous incremental change (in forces) is mapped onto occasionally discontinuous revolutionary change (in relations) in the base of modes of production.

4. Bases Determine Superstructures

One of the important contributions of G. A. Cohen was a clarification of the notion of superstructure. If Marx claims that the base or, more precisely, relations of production determine the superstructure, then a definition of relations and superstructure must be found in which the two are clearly distinguished. The difficulty is that power as we find it in any social formation is encased in particular cultural and legal concepts, i.e., superstructural notions. It was for this reason that Cohen defined relations of production as the *de facto* distribution of power. Every culturally encoded right, he argued, can be matched with a *de facto* power. And it is *de facto* power that determines culturally specific notions, rather than the reverse.

One important effect of this definitional move is to place the long-term dynamic of a mode of production outside of culture itself. Eventually, culture corresponds to *de facto* power rather than the reverse. But in a shorter-term sense, it places culture at the very center of the analyses of modes of production, for it is precisely the superstructure that defines and stabilizes relations of production. Once again, the logic of functional explanation turns the expectations of ordinary language upside-down. If something is described as having a base and a superstructure, we think of the base as a foundation. It supports the superstructure; without the foundation, the superstructure would fall to the ground. Yet, the logic of functional explanation requires the opposite: it is the superstructure that 'makes' the base.

To adapt a visual image from Cohen (1978), consider the following: Four pieces of wood of the same length are driven a little way into the ground. They stand, but they are easily blown over by a strong wind. A top is attached to the four struts such that the whole becomes a table. The struts, now legs, stand solidly upright. They do not wobble in the wind. Of the tabletop, one can say: (a) it is supported by the legs, and (b) its very existence is required for the legs to be legs. Without the top, the legs would not be legs—only sticks of wood of such-and-such a length. Here we have a building whose base and superstructure relate in the proper way.

The conclusion to this line of the thought (heretical to some Marxists) is that *Capital* is an analysis of the capitalist superstructure—the set of culturally specific and legally encoded meanings and practices that tend to stabilize capitalist relations of production. In a comparative sense, this conclusion is crucial since it sets capitalism on the same analytical plane as any other mode of production. On this view, the analysis of modes of production becomes a matter of locating effective power difference over material production as they are reflected in differential control over the total social product and analyzing the meanings and practices that tend to reproduce such powers.

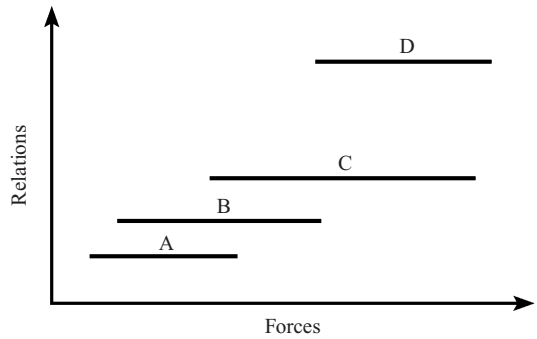

Figure 1
A possible set of forces and relations of production when functional alternatives exist

5. Functional Alternatives

If this reading of Marx's notion of determination solves many conceptual problems, it also creates others in the application of Marx's ideas. Carl Hempel and others have called attention to the difficulties of actually carrying out functional explanations. Not the least of these stem from the problem of functional alternatives. That is, there is often more than one Y (Y_1, Y_2, Y_3 ...) that can fulfill any given function. That being so, the particular presence of Y_1, say, cannot be explained simply by referring to its function, for Y_2 or Y_3 could have served the same need.

The critical question becomes: what is the size of the set of functional alternatives? Implicitly, Marx appears to have believed that alternatives are quite limited—typically to the set of one. Consider such a limiting case. For each level of development of productive forces, there is only one set of relations possible. History therefore proceeds in a series of discrete steps from A to B to C. Both the past and the future are specified—socialism is inevitable because it is supposedly the only economic system consistent with highly developed forces.

A different story emerges if the number of functional alternatives is even moderately few. If several different sets of relations of production are compatible with any given level of forces, then the grand design of world history becomes considerably more complex. One possible trajectory in Fig. 1, for example, would be A to C to B to D. In this reading, the future looks considerably more open. There may indeed be functional alternatives to socialism—if indeed socialism as we have known it is compatible with developed forces.

6. Conclusion

The notion of a mode of production is a deliberate simplification. It provides a model of certain central relationships that define epochal forms. As such, the analysis of modes of production cannot substitute for the examination of particular cases in specific spatial and temporal contexts. In this sense, mode-of-production analyses do not constitute historical explanations.

In a larger sense, however, the concept of modes of production furnishes the necessary scaffolding without which all local historical explanation would be impossible. Comparison and contrast, whether explicitly or implicitly carried out, depend upon a common framework. This is what Marx's notion of ways of producing provides: a unified theory for viewing human society and history.

See also: Communism; Feudalism; Marx, Karl (1818–89); Marxism in Contemporary Sociology; Marxist Social Thought, History of; Socialism: Historical Aspects; Technological Determinism

Bibliography

Althusser L, Balibar E 1970 *Reading Capital* [trans. Brewster B]. Pantheon, New York

Anderson P 1974a *Lineages of the Absolutist State*. New Left Books, London

Anderson P 1974b *Passages from Antiquity to Feudalism*. New Left Books, London

Aston T H, Philpin C H E (eds.) 1985 *The Brenner Debate: Agrarian Class Structure and Economic Development in Pre-Industrial Europe*. Cambridge University Press, Cambridge, UK

Cohen G A 1978 *Karl Marx's Theory of History: A Defence*. Clarendon Press, Oxford, UK

Donham D L 1999 *History, Power, Ideology: Central Issues in Marxism and Anthropology*, 2nd edn. University of California Press, Berkeley, CA

Kahn J S, Llobera J R (eds.) 1981 *The Anthropology of Pre-Capitalist Societies*. Macmillan, London

Moore S 1980 *Marx on the Choice Between Socialism and Communism*. Harvard University Press, Cambridge, MA

D. L. Donham

Model Testing and Selection, Theory of

Probability models represent observations in terms of random variables. To fit a model is to provide an estimate of the probability distribution of these random variables. Goodness-of-fit statistics measure the discrepancy of the observed from the estimated distribution. Goodness-of-fit tests are used to decide whether these discrepancies are significant under the null hypothesis that the model holds. Model selection is the process of selecting one from different families of competing models (see *Computational Approaches to Model Evaluation*).

This article reviews procedures for assessing model fit and techniques for selecting models. Classical chi-square based approaches to model fit provide goodness-of-fit tests for discrete random variables that take only a finite number of different values. For continuous random variables, chi-square based methods are available as well as procedures using order statistics and log-likelihood ratio tests, among others.

Classical approaches to model selection build on goodness-of-fit tests in that they select the most simple model that fits the data. A related approach is based on tests between hypotheses that correspond to the different families of models. Other approaches considered include the Bayesian approach and the discrepancy-based method, in which models are selected with the goal to minimize the expected value of a given discrepancy criterion.

1. Models

In a probability model, observed values x are assumed to be realizations of random variables X with probability distribution Q that is an element of a parameterized family of probability distributions $M = \{P_\theta, \theta \in \Theta\}$, where θ may be a vector with several components and Θ, the parameter space, is the set of values that can be taken by θ. The model is the family of distributions. A model is said to be nested in another model, if its family of distributions is a subset of that of the second model.

2. Goodness-of-fit

Assessing goodness of fit means the statistical testing of models with data. Based on a random sample $x_1, \ldots, x_n$ realizing random variables $X_1, \ldots, X_n$ with distribution Q, it is thus required to test the null hypothesis

$$H_0: Q \in M \tag{1}$$

Most goodness-of-fit tests are parameter-free, that is, the distribution of the test does not depend upon θ, at least asymptotically, that is, as the sample size increases.

2.1 Goodness-of-fit Testing for Discrete Random Variables

Many traditional goodness-of-fit statistics were originally developed for discrete random variables. Dis-

crete random variables take only a finite (or at most countably infinite) number of values. They underlie statistical models of categorical data (Agresti 1990). The probability distribution Q of a discrete random variable X can be specified by its probability function $q(x) = Q(\{x\})$, which assigns the probability of the set $\{x\}$ to x. Likewise, the probability distribution of P_θ is specified by the probability function $f(x \mid \theta) = P_\theta(\{x\})$.

Let N_x be the number of X_i with $X_i = x$, and let n_x be a realization of the random variable N_x. Goodness-of-fit tests are based on asymptotically efficient estimators $\hat{\theta}_n$ of θ such as the maximum likelihood estimator under H_0. The two most widely used tests are given by, respectively, the log-likelihood ratio statistic, where log is the natural logarithm

$$G^2 = -2\sum_x n_x[\log nf(x \mid \hat{\theta}) - \log n_x] \tag{2}$$

and Pearson's chi-square statistic

$$X^2 = \sum_x (n_x - nf(x \mid \hat{\theta}))^2/(nf(x \mid \hat{\theta})) \tag{3}$$

where n is the total number of observations in the sample.

Under H_0 and given suitable regularity conditions, each of the above statistics is asymptotically distributed as χ^2 with degrees of freedom given by the number of different categories minus one less than the number of independent parameters. In practice, it may be necessary to aggregate data from several categories, forming new, composite categories from categories x for which $nf(x \mid \hat{\theta})$ is small. Although the statistics of Equations (2) and (3) are most well known, there are many others (Read and Cressie 1988).

2.2 Goodness-of-fit Testing for Continuous Random Variables

Goodness-of-fit tests for models of continuous random variables can be based on those for the discrete case and on the order statistics. Apart from general-purpose methods, special tests have been developed to test for exponential and normal families of distributions among others (D'Agostino and Stephens 1986).

The goodness-of-fit test for models of discrete random variables can be applied to the continuous case by grouping the values of the random variables by means of a finite number of intervals I_x, $x = 1, \ldots, k$. For this purpose, compute $p(x \mid \theta) = P_\theta(I_x)$; n_x, the number of observations falling into I_x; and, finally, the above test statistics on the basis of these grouped data and probability distributions. Estimation of θ should be based on the grouped data rather than on the original likelihood. Note that a subjective element enters the test through the choice of the intervals I_x. A rational choice is to define I_x so that $P_{\hat{\theta}}(I_x) = 1/k$, $x = 1, \ldots, k$. Although this makes the intervals dependent upon the random sample via $\hat{\theta}$, the asymptotic χ^2-distribution of the above test statistics is not compromised (Moore and Spruill 1975). When the maximum likelihood estimates based on the ungrouped data are available, generalized chi-square statistics such as the Rao–Robson statistic (D'Agostino and Stephens 1986) provide more powerful goodness-of-fit tests and should be preferred.

Given continuous real-valued random variables $X_1, \ldots, X_n$, the so-called order statistics $X'_1, \ldots, X'_n$ correspond to the values of the random variables ordered from smallest to largest and thus, $X'_1 < \ldots < X'_n$ (with probability one). Let $0 < \lambda_1 < \ldots < \lambda_k < 1$, and let n_i be the greatest integer less than or equal to $n\lambda_i + 1$. The λ_i partition the range of cumulative probabilities into a finite number of intervals, and thus a subjective element is implicit in their choice. Denote the cumulative distribution function associated with P_θ by $F(\bullet \mid \theta)$, and define the statistic

$$Y_n^2(\theta) = n\sum_{i=1}^{k}[(F(X'_{n_i} \mid \theta) - F(X'_{n_{i-1}} \mid \theta)) - p_i]^2/p_i \tag{4}$$

where $p_i = \lambda_i - \lambda_{i-1}$, $i = 1, \ldots, k$. If $\hat{\theta}_n$ is an estimator that minimizes $Y_n^2(\theta)$, then under suitable regularity conditions the test statistic $Y_n^2(\hat{\theta}_n)$ is asymptotically distributed as chi-square with $k - q - 1$ degrees of freedom, where q is the number of independent parameters.

2.3 Testing against Specified Alternatives

The above methods are general-purpose tests that aim at realizing test power for detecting a broad range of deviations from the model assumptions. Discussions of diagnostic tests and graphical methods for detecting specific deviations abound in treatments of many particular models. A general and widely applicable principle for deriving such tests is given by Neyman–Pearson log-likelihood ratio tests of nested models. For this purpose, a second, less constrained family of distributions $N = \{P_\eta, \eta \in \Psi\}$ is considered of which the original family M is a subset. For example, the second model may involve the same and additional predictors in the context of multiple linear regression, it may additionally contain interactions of higher order in the context of analysis-of-variance models or generalized linear models, and so forth. A so-called conditional model test is a test of H_0 vs. H_1: H_0: $Q \in M$ vs. H_1: $Q \in N \backslash M$. For this test, both θ and η are estimated by means of the maximum-likelihood method under M and N, respectively, and $-2\log L$ is computed, where L is the ratio of the resulting likelihood values under the two models. Under suitable regularity conditions,

the test statistic is distributed as chi-square with degrees of freedom given by the difference in the number of independent parameters of the two models. The test power of this test is particularly large for detecting violations of constraints that M imposes over and above those imposed by N, but a significant result can also mean that constraints common to M and N are already violated.

3. Model Selection

In model testing, there is one distinguished model, and the objective is to test its validity. In model selection, there are several models, and the objective is to select one which best serves a given function such as estimation of certain common parameters or prediction. Frequently, the different models differ in complexity, usually indexed by the number of independent parameters, and most methods explicitly or implicitly incorporate a trade-off between fit and parsimony. Among the many different approaches which are discussed in this emerging field are model selection by classical hypothesis tests, discrepancy-based methods, bootstrap and cross-validation techniques, Bayes approaches, and others.

3.1 Model Selection by Classical Hypothesis Tests

In nested families of models, procedures for model selection have been proposed based on the classical hypothesis testing framework sketched above. So-called sequential model tests start with the model with the minimal number of constraints. During step k of the procedure, it is assumed that the model with k constraints holds, and is tested against the model with $k+1$ constraints. The sequence of tests is terminated at step l as soon as the first hypothesis with $l+1$ constraints is rejected. In multiple linear regression, the models are defined by the subset of predictors to be included in the equation, and automated sequential selection procedures based on hypothesis tests such as forward and backward stepwise selection have been considered (Jungeilges 1992). For non-nested, disjoint models, selection by hypothesis testing is possible by means of a generalized likelihood ratio approach originally proposed by Cox (1961) or by integrating the models into a joint supermodel first (Atkinson 1970).

3.2 Discrepancy-based Methods

Another principled approach to model selection has been proposed by Linhart and Zucchini (1986). A discrepancy Δ measures the lack of fit of the underlying distribution Q to members P_0 of candidate models. It satisfies $\Delta(\theta) = \Delta(Q, P_\theta) \geqslant \Delta(Q, Q)$ for all θ.

Depending upon the purpose of the model selection exercise, different aspects of model fit may be important and may lead one to consider different discrepancies. Important general-purpose discrepancies are the following, where $F(\bullet)$ is the cumulative distribution function associated with Q and $F(\bullet \mid \theta)$ is the cumulative distribution function associated with P_θ. (For continuous random variables, the probability functions should be replaced by probability density functions, and the sums by integrals.)

Kullback–Leibler discrepancy:

$$\Delta_{K-L}(\theta) = -E_Q \log f(x \mid \theta),$$

Kolmogorov discrepancy:

$$\Delta_K(\theta) = sup_x \mid F(x) - F(x \mid \theta) \mid,$$

Cramér–von Mises discrepancy:

$$\Delta_{C-M}(\theta) = E_{P_\theta}(F(x) - F(x \mid \theta))^2,$$

Gauss discrepancy:

$$\Delta_G(\theta) = \Sigma_x (f(x) - f(x \mid \theta))^2,$$

Pearson chi-squared discrepancy:

$$\Delta_P(\theta) = \Sigma_x (f(x) - f(x \mid \theta))^2 / f(x \mid \theta),$$

Neyman chi-squared discrepancy:

$$\Delta_N(\theta) = \Sigma_x (f(x) - f(x \mid \theta))^2 / f(x).$$

For a given model, the discrepancy due to approximation is defined as the discrepancy $\Delta(\theta_0)$ of that member P_{θ_0} of the model that minimizes $\Delta(Q, P_\theta)$. It is sometimes also called the model bias, the model error, or the degree of misspecification. For a given model and method of estimating θ, the discrepancy between P_{θ_0} and the estimated distribution $P_{\hat\theta}$, $\Delta(P_{\theta_0}, P_{\hat\theta})$, is called the discrepancy due to estimation. It is sometimes also called the estimation variance.

Finally, the overall discrepancy is the discrepancy of the underlying distribution and the fitted distribution, $\Delta(Q, P_{\hat\theta})$. In many of the cases of interest, the overall discrepancy is the sum of the two component discrepancies, but even when it is not, the overall discrepancy is still a monotone function of both component discrepancies.

It is interesting to consider how the component discrepancies behave as model complexity increases. As the model becomes less constrained, the discrepancy due to approximation will tend to decrease, because the model will be able to approximate the true distribution more closely. On the other hand, the

discrepancy due to estimation will tend to increase simultaneously because complex models have a greater propensity to capitalize on chance and to be misled by chance fluctuations in the sample. Thus, the overall discrepancy represents a trade-off between ability to fit and propensity to capitalize on error.

In this approach, selection is based on estimated expected discrepancies, where expectation is taken with respect to the underlying distribution. An estimator of the expected discrepancy is called a criterion and that model is selected that gives rise to the smallest criterion.

The derivation of criteria can be a difficult technical issue. In some contexts, unbiased estimators can be derived in the finite-sample case. For example, in the context of multiple linear regression, the well-known variable selection criteria S_p and C_p can be thus derived as (monotone transformations of) unbiased estimators of discrepancies based on errors of prediction (Linhart and Zucchini 1986). In many instances, unbiased estimators cannot be derived. In these cases, asymptotic methods, bootstrap methods, and cross-validation techniques can be used.

For example, when the underlying distribution is a member of model M, an asymptotically unbiased criterion based on the Kullback–Leibler discrepancy is in fact Akaike's information criterion (divided by $2n$)

$$AIC = -2\log f(x_1, \ldots, x_n \mid \hat{\theta}_n) + 2q \qquad (5)$$

where $f(x_1, \ldots, x_n \mid \theta)$ is the likelihood of the data $x_1, \ldots, x_n$, and estimation is based on the maximum-likelihood method. Even if Q is not in M, $AIC/(2n)$ has been argued to remain a satisfactory approximation of the Kullback–Leibler criterion as long as the model is not grossly misspecified (Linhart and Zucchini 1986).

3.2.1 Bootstrap techniques. Bootstrap techniques provide another means of estimating expected discrepancies which is widely applicable. In the bootstrap method, the unknown distribution Q is replaced by Q_n which assigns probability mass $1/n$ to each observed value x_i, $i = 1, \ldots, n$ (Efron 1982). This makes it possible to compute expected discrepancies when an explicit formula is available, or to use Monte Carlo methods to estimate expected discrepancies to any desired degree. Expectation is with respect to Q_n rather than Q, however, and the bootstrap criterion has a downward bias (Chung et al. 1996).

3.2.2 Cross-validation techniques. In cross validation, the data are divided into two subsamples, a calibration sample of size $n-v$ and a validation sample of size v. The first sample is used to estimate the model, the second is used to estimate the expected discrepancy. Cross-validation techniques for model selection use a small v, typically $v = 1$, but repeat the above steps for all possible subdivision of the sample data into two subsamples of the required sizes. The cross-validation criterion is the average, over these repetitions, of the estimated expected discrepancies. When $v = 1$, this method is called one-item out cross validation. Under many circumstances, the resulting cross-validatory criterion is straightforward to compute and is an estimator of the expected discrepancy for a sample of size $n-v$ (Linhart and Zucchini 1986).

3.3 Bayesian Model Selection

Another principled approach to model selection has been proposed within the framework of Bayesian statistics. In this approach, each model M is seen as potentially containing the true probability distribution, and an *a priori* probability, $P(M)$, for this event is assigned to each model. Based on the data D, the posterior probabilities of the models, $P(M \mid D)$, are then computed, and the model with the largest posterior probability is selected. Another possibility is to avoid selection of one specific model altogether, and to use instead the posterior distribution, which averages over the different models by weighting each model with its posterior probability. This is called model averaging. In addition to an *a priori* probability for each model, an *a priori* probability has to be specified for the distribution of each model's parameters. Computing the posterior probabilities can be a formidable task.

For example, consider the case of just two models M_1 and M_2 and data D. Selection is based on the posterior odds,

$$\frac{P(M_1 \mid D)}{P(M_2 \mid D)} = \frac{P(D \mid M_1)}{P(D \mid M_2)} \frac{P(M_1)}{P(M_2)} \qquad (6)$$

such that M_1 is selected if the posterior odds exceed one, and M_2 is selected otherwise. The first quotient on the right side of the above equation is called the Bayes factor, and it is equal to the posterior odds ratio, if the prior odds ratio is one.

Another index is the Bayes information criterion, BIC

$$BIC = -2\log f(x_1, \ldots, x_n \mid \hat{\theta}_n) + q\log(n) \qquad (7)$$

where $f(x_1, \ldots, x_n \mid \theta)$ is the likelihood of the data $x_1, \ldots, x_n$, and estimation is based on the maximum-likelihood method. It can be shown that selecting the model with the smaller value of the Bayes information criterion is asymptotically equivalent to selecting according to the Bayes factor. Thus, the index BIC can be used to obtain a large-sample approximation of the Bayes factor.

3.4 Model Selection and Testing: Concluding Remarks

Theoretical results obtained in the context of nested families of distributions show that *BIC* identifies the correct model with probability converging to one, as the sample size increases, whereas a similar asymptotic consistency is not given for *AIC*(Jungeilges 1992). However, most distributions underlying real data are likely to be much more complex than the most complex model considered for selection. Under this premise, all of the above selection methods favour the most complex models among the considered models as sample size increases, because estimation variance, but not model bias, converges to zero as n goes to infinity. That is, in large samples, selection is dominated by the degree of misspecification of given models so that non-parsimonious models with more ability to approximate the underlying complex distribution are selected. Similarly, in this situation, any model is eventually rejected in model testing as the test power of model tests for detecting even tiny model violations increases with increasing sample size.

Another class of problems is subsumed under the label model selection bias (Chatfield 1995). Model selection criteria are themselves random variables and thus subject to random error. As a consequence, which model is selected for further inference or prediction is uncertain, and there may be a substantial probability that another model is chosen if the model selection procedure is repeated on the basis of a new random sample. Note that this is a fundamental problem that also applies to criteria based on cross validation.

This has several consequences. The criterion value of the selected model will in general appear too good, because overfitting may have contributed to the model's eventual selection success, and nominal standard errors of parameter estimates and nominal levels of significance tests based on the selected model will be incorrect because they do not take model uncertainty into account. Several approaches are being discussed for dealing with this issue. In Bayesian model selection, model uncertainty can be incorporated explicitly by model averaging (see above). Furthermore, through appropriate bootstrap and cross-validation techniques, model selection can be built explicitly into statistical procedures (Hjorth 1994). Further relevant references are Buckland et al. (1997) and Ye (1998).

See also: Goodness of Fit: Overview; Goodness-of-fit Tests and Diagnostics

Bibliography

Agresti A 1990 *Categorical Data Analysis*. Wiley, New York

Atkinson A C 1970 A method for discrimination between models. *Journal of the Royal Statistical Society* **32**: 323–45

Buckland S T, Burnham K P, Augustin N H 1997 Model selection: An integral part of inference. *Biometrics* **53**: 603–18

Chatfield C 1995 Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society* **158**: 419–66

Chung H-Y, Lee K-W, Koo J-Y 1996 A note on bootstrap model selection criterion. *Statistics and Probability Letters* **26**: 35–41

Cox D R 1961 Tests of separate families of hypotheses. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press. Berkeley, CA, pp. 105–23

D'Agostino R B, Stephens M A 1986 *Goodness-of-fit Techniques*. Marcel, New York

Efron B 1982 *The Jackknife, the Bootstrap and other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia

Hjorth U 1994 *Computer Intensive Statistical Methods. Validation, Model Selection and Bootstrap*. Chapman and Hall, London

Jungeilges J A 1992 *The Robustness of Model Selection Rules*. Lit, Münster, Germany

Linhart H, Zucchini W 1986 *Model Selection*. Wiley, New York

Moore D S, Spruill M C 1975 Unified large-sample theory of general chi-squared statistics for tests of fit. *Annals of Statistics* **3**: 599–616

Read T R C, Cressie N A C 1988 *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer, New York

Ye J M 1998 On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93**: 120–31

K. C. Klauer

Models: Philosophical Aspects

Contemporary philosophers of science argue that models are a major vehicle of scientific knowledge. This applies to highly theoretical inquiry as well as to experimental or otherwise observational research, in both the natural and the social sciences. Making this claim is not yet very illuminating, given that there is a large variety of different kinds of model, and a number of ways in which they function in the service of science.

The ambiguity of the term 'model' and the multiplicity of kinds of model are illustrated by Pierre Duhem's famous comparison of the mind of a continental physicist to that of an Englishman: the former strives for 'theories' that are formulated in 'the clear and precise language of geometry and algebra' and consist of abstract and idealized notions and formulae, while the latter insists on having mechanical 'models' that satisfy 'his need to imagine concrete, material, visible, and tangible things' that are familiar to ordinary experience. While the French or German physicist deals with a formalized theory in his account of electrostasis, the English account is in terms of 'strings which move around pulleys, which roll around

drums, which go through pearl beads, which carry weights; and tubes which pump water while others swell and contract; toothed wheels which are geared to one another and engage hooks. We thought we were entering the tranquil and neatly ordered abode of reason, but we find ourselves in a factory' (Duhem 1954, pp. 70–1). Models in this sense—iconic models—employ analogies to visualize the mechanisms depicted, while formalized theories supposedly lack this property. On the other hand, it is nowadays customary to use the term 'model' also for such formal systems of equations, conspicuously so in the social sciences—and, much of the time, such formal systems are used for representing causal mechanisms. In relation to such a formal system, 'model' is also used for its various interpretations, or just for the interpretation that makes it true. The model muddle needs to be sorted out.

1. Typologies and Distinctions

In ordinary language, the term 'model' is used to stand for many things, such as a type of design (the fashion industry's new spring models; the 1997 model Saab; the German model of industrial relations) and an exemplar: an object proposed or adopted for imitation (a woman displaying clothes in fashion shows; a model wife; a person posing for artists and art students). The term is also used for what are often called scale models or replicas, such as an architect's design of a house presented as a miniature construction (here the thing modeled does not yet exist—the model is used in the process of producing the thing); and three-dimensional depictions of the atom and the solar system exhibited in science museums (here the thing modeled already exists).

In meta-logic, a model of a theory is any set of entities or a structure that satisfies the axioms of the theory. Here 'theory' denotes a formal, uninterpreted system of axioms and deductively implied theorems. In the traditional usage of 'model' in the social sciences, this idea becomes reversed: it is such formalized or semi-formalized systems that are often called by the name 'model'—in contrast to verbal theories.

There are attempts to classify types of model, but none of them is exhaustive. Black (1962) lists four types. Along selected dimensions, *scale models* reduce or magnify the properties of the objects they represent with the purpose of reproducing, in a manipulable or accessible form, important properties of the original thing modeled. While scale models involve change of scale, *analog models* also involve change of medium while reproducing the structure of the original: the model and the thing modeled are isomorphic. *Mathematical models* are the formalized theories of social scientists in which the system modeled is projected upon the highly simplified and abstract domain of sets and functions that can be manipulated by means of mathematical reasoning. Finally, *theoretical models* are simplified systems or structures. Unlike scale models, they are imagined and described but not literally built. In contrast to mathematical models, the properties of a theoretical model are better known than those of the original subject matter that is modeled.

Achinstein (1968) distinguishes between three kinds of model: representational, theoretical, and imaginary. In contrast to Black's typology, there is no distinction here between theoretical (and imaginary) models on the one hand, and mathematical models on the other. What Achinstein calls *representational models* (for reasons that will become clear, this label is unfortunate; a better name would be 'material models' or the like) are three-dimensional physical representations of objects of interest, such as tinkertoy models of molecules, and engineering models of dams and airplanes. The model and the object represented by it are two distinct objects. Engineers call the objects represented by such models prototypes. The model may reproduce all (or only some of) the relevant properties of the object, using one scale uniformly for all properties (such as a model of a bridge in which length, width, and thickness are all reduced uniformly by the same factor of 100); or using different scales for different properties (such as a model of the planetary system in which the sizes of the planets and the distances between them are reduced by different factors); or the relationship between the model and the prototype may be one of analogy (such as using an electrical system as a model of an acoustic system).

A *theoretical model* of an object is a set of assumptions about that object rather than a distinct object. Examples are the Bohr model of the atom, the Crick–Watson model of the DNA molecule, Markov models of learning, and the multiplier–accelerator model of economic growth. Such models are often called mathematical models in the social sciences, as the assumptions of the model frequently are expressed using mathematical equations. Theoretical models characteristically describe an object by ascribing to it an inner structure or mechanism. It is by reference to this structure that theoretical models help explain the behavioral and other properties of the objects so described. A theoretical model in this sense can be understood as a simplifying approximation of the object that is useful for some purposes. The use of such a model characteristically involves the awareness and explicit acknowledgement that the real object is far more complex than its representation in the model: the theoretical model assumes away many complications while highlighting limited aspects of the object. This feature of theoretical models explains why typically it is possible to hold and use a number of different models of the same object: given their characteristic simplifications, they highlight different aspects of the object for different purposes. Theoretical models can be viewed as small-scale theories with a limited scope

of application, drawing from broader theories and from systems of more general and fundamental theoretical principles. They are combinations of such general principles and more specific, locally applicable, auxiliary assumptions.

Finally, many *imaginary models* are similar to theoretical models in that they are constituted by sets of assumptions describing an object or system. The difference lies in the imaginary character of the former: the assumptions of imaginary models are not intended as true or even approximately true descriptions of the features of any real object. Imaginary models describe imaginary worlds. They entail counterfactual conjectures that reveal logically, physically, mentally, or socially possible properties or conceivable behaviors: the imagined object could be one way if it met the conditions laid down in the assumptions—but it does not actually meet them. Imaginary models typically serve as stepping stones directing scientists toward further investigations about the real characteristics of real objects. We could consider Max Weber's notion of an ideal type as being close to the notion of an imaginary model in Achinstein's sense.

Many models involve analogies and metaphors. Given that there are different kinds and aspects of the latter, this gives rise to different kinds of model. Hesse (1963) is a classic discussion of models and analogies. When considering a collection of billiard balls in random motion as a model for a gas, the suggestion is that gas molecules are analogous to billiard balls. This implies that gas molecules and billiard balls share some properties—constituting the positive analogy—while other properties are not shared—they are the negative analogy. The neutral analogy consists of those properties which we do not yet know are shared. Hesse then distinguishes two kinds of model, the first being the conjunction of the positive and the neutral analogy, while the second includes the negative analogy as well.

Harré (1970) proposes a refined classification between three types and further subtypes of models. *Homeomorphs* are models whose subject and source are the same. Scale models are one subtype of homeomorphs (further divided into micromorphs and megamorphs, depending on whether the properties are reduced or magnified). Another subtype consists of teleiomorphs that are improvements on their subjects by way of abstracting (subtracting irrelevant properties, e.g., as in maps) or idealizing (such as in the case of the fashion model). The third subtype is the metriomorph that represents a class by way of averaging, for example (such as the notion of the average family along some dimension of properties, such as the number of children). *Paramorphs* are models whose source and subject are different, thus they connect two or more domains with one another. They may be used heuristically, as in using electric networks as models of hydraulic networks, and in using the latter as models of economic networks; or to create causal hypotheses, such as when using bacteria as models of viruses. Finally, *protomorphs* are kinds of diagrams and geometrical representations. Diagrams of social networks serve as an example of such models.

These examples show that the ontology of models is not uniform: models can be made of material stuff, and they may be linguistic as well as pictorial. Indeed, this last class should not be ignored, as science operates with a variety of visual models: drawings, figures, diagrams, graphs, and other images. Much of causal modeling in social research employs such visual means of representation: various causal chains and loops are represented in terms of boxes and arrows. The use of such models for visual representation does not follow the same rules as the linguistic mode of representation. As soon as a visual model is translated into a mathematical model, the rules of its use change.

Some models are abstract entities, while some others are very concrete. Consider animal models in biomedical research. Under certain conditions, certain animals can be manipulated so as to use them as models for human beings. By experimenting with the animals, medically relevant information is generated about humans. Particular animals are concrete systems that are used as models. In general, any experimental set-up serves as a model of a non-experimental real world system. Such experimental systems are concrete rather than abstract entities. One can also think of numerical computer simulations in these terms: systems of equations are manipulated numerically with the help of a computer to acquire information about some real system that is modeled in such a concrete way.

There is a general sense in which models are constructed. However, a three-dimensional scale model is constructed in a difference way than a theoretical model: it is built, while a theoretical model is imagined and then described. An experimental model is similar to the scale model in this respect. The important difference between an experimental model and a theoretical model is that the elements of the former can be controlled causally, while in the latter, such control is forthcoming only by way of assuming that certain factors are absent, constant, or otherwise neutral.

The concept of model suggests that a model is a 'model of' something. Thus we have models of humans and models of social interaction, models of the atom, and models of the origin of the universe. We also have models of quantum mechanics and models of the general theory of economic equilibrium. In general, we may distinguish between models of theories and models of real systems. The scientific procedure also involves models of data that reduce the complexity of experiential materials to simpler and more manageable portions of information. Based on classifications, simplifications and idealizations of various sorts, sampling, data mining, and other techniques, these are

models of the complete data. Models of data are necessary for the same reason for which models in general are needed: the complete data are too rich and complex to be of any use.

A model *of* something may sometimes also serve as a model *for* design or imitation, as in architecture and social thought. A miniature model of a house and a theoretical model of an institutional structure may serve as models for constructing a concrete house and a concrete structure of social institutions.

2. Theories and Models

Models and theories are viewed as being related in a number of ways, depending on notions of model and theory. It has been suggested that a major difference between a theory and a model is that scientists using either of these hold different attitudes towards them: They take a theory to involve the belief that the system it describes is really governed by the principles suggested by the theory, while such a belief is not involved in the use of models. Scientists take a theory to be (perhaps hypothetically) a true account of a given system, while various models can be used for different purposes in relation to that same system without any commitment as to their truth. One may refer to something as 'just a model,' whereas to characterize it as a 'theory' is to be less modest about it. Another traditional way of stating the idea is to say that, when using a model, the scientist claims that a system behaves *as if it were* as the model represents it, while, when using a theory, the claim is that the system behaves as it does because *it is* (or at any rate might be) the way the theory says it is.

In a similar vein, Achinstein's distinction between theoretical and imaginary models is based on the different attitudes of scientists: theoretical models do, and imaginary models do not, involve ontological commitments as to the reality of the entities and properties depicted. This suggestion implies immediately that different attitudes toward certitude and commitment do not define a theory–model distinction (but rather a distinction among models). On the other hand, Achinstein's distinction, too, is based on differential attitudes. This seems too stark: the differences in the relevant attitudes often seem more blurred. Scientists may not be in agreement regarding the ontological status of a given model. Such attitudes are also often not on–off matters, but rather matters of degree. Thus, no sharp dichotomy may be forthcoming. Moreover, these suggestions presuppose a realist outlook: on an instrumentalist account of theories, there is no difference of any sort, as everything would be an imaginary model. It should be added that, without further specifications, the as-if locution alone is unable to differentiate between the intended two attitudes.

Another popular account is in terms of specificity. A theory is an abstract and general statement about a number of generic entities and their key dependencies. Game theory can be considered a theory in this sense. The neoclassical theory of economic growth is another example. A theoretical model is a more specific and smaller scale version of the theory, involving various idealizations, simplifications and other auxiliary assumptions, and addressing specific issues. A theory serves as a prototype for the construction of such models. One theory can be specified in the form of many theoretical models, using different selections of variables and functional forms between them, for example. The prisoner's dilemma, battle of the sexes, chicken, and other 'games' and their further specific versions are such theoretical models. So are one-sector and two-sector models of growth, with exogenous or endogenous technological progress. An empirical model is one with parameters that can be, or have been, estimated in terms of empirical data. Experimental game models and econometric growth models are examples.

There are further, logically pretheoretic, elements that are often called models. They are world models or metaphysical models, such as models of individuals and models of society. They have the character of general presuppositions or underlying convictions that provide the most abstract conceptual frameworks for more specific intellectual exercises. These include various forms of individualism and collectivism, agency-structure models, visions of society as in harmony and in conflict, and of the development of societies as progressive or repetitive.

The above accounts often regard theory as consisting of statements. This idea has been challenged by those who hold so-called semantic or model-based accounts of theories. Consider an accessible version of such an account of theories as made of models (Giere 1988). Models are taken to be abstract and idealized entities that have only the properties ascribed to them by scientists. Thus scientists refer to planetary systems where planets are dimensionless mass points, and to harmonic motion such as that of a pendulum of a frictionless clock influenced only by a uniform gravitational force presupposing that the Earth is a perfectly homogeneous sphere. Models in this sense can be characterized by linguistic (and other, such as graphical) means, but are not themselves linguistic entities. Models lie somewhere between language and concrete reality.

It is characteristic of such abstract and idealized entities that they perfectly satisfy certain mathematical equations. Thus the simple harmonic oscillator perfectly satisfies the force law $F = -kx$ that serves as its equation of motion. Likewise, the model of perfect competition in economics can be considered an abstract and idealized system that perfectly satisfies 'laws' such as the first and second welfare theorems. The idealizations of the models are needed to ensure

such a perfect fit. Indeed, models in this sense are constructed so as to fully satisfy the key claims of a theory. This way of putting it brings the idea close to the use of the term 'model' in meta-logic where a model of a formal theory is a structure that satisfies the axioms of the theory.

Note that Giere takes models to be abstract entities rather than anything concrete. Using such a model with highly idealizing assumptions, one can derive mathematically statements about the behavior of the abstract entity. The greater the similarity between the abstract system and the concrete system, the closer the fit between the behaviors of the abstract and concrete systems. Giere says he has derived the notion of model in this sense from physics textbooks. However, just as real systems are more complex than abstract ones, so is much of real scientific research more complex than its textbook 'models.' Many models used in science are concrete and do not allow for neat mathematical derivations.

Others challenge such views of theories-as-models and models-as-specifications-of-theories, arguing that models often enjoy a relatively independent role (Morgan and Morrison 1999). The construction of models draws on theories as well as on other sources (such as the data, technological design, and intuitive insight), but because models are relatively independent of each such source, they are able to serve as mediators between them. This applies both to theoretical models and to models of the data and experiments. This further emphasizes the possibility of having a number of (possibly mutually inconsistent) models of a system or phenomenon covered by a theory.

3. *Functions of Models*

It is important to put any thoughts about this issue in the plural: functions of models. Among other things, models help to explain and predict phenomena; construct, interpret and test theories; design and produce technologies, and (material and social) structures. This multiplicity of uses is yet another reason why scientists hold a multiplicity of models of the same system or phenomenon.

It used to be popular to conceive of models—models of theories—as providing interpretations for abstract scientific theories. Theories were taken to consist of formal, axiomatic systems in which theoretical terms—terms not referring to anything observable—figured prominently but lacked empirical interpretation. Models were regarded as visual or otherwise familiar systems that were isomorphic with the theories they help to interpret. Such iconic models provide interpretations for such formal calculi: thus, the system of billiard balls in a box is used to help understand the import of a theory of gases; and the analogy from the solar system is used for interpreting atomic theory.

Many of the earlier accounts of models viewed them as facilitating instruments of scientific cognition, rather than as its embodiments: models as heuristic devices, as aids to imagination, sources of inspiration, pedagogic devices, and the like. On the basis of substantive or formal analogies, models suggest directions for theory construction and for extending the application of a theory to new domains. In Gary Becker's controversial economics of the family, children are represented in analogy to durable consumer goods, thus rendering conventional economic principles applicable to family behavior. Models suggest new questions and new hypotheses as conjectural answers to these questions. The solar system model of the atom is an analog model where the nucleus is like the sun and the electrons are like the planets, orbiting the sun. This model has given rise to fruitful questions and hypotheses, such as those about the shape of the orbits of the electrons and the velocity of their motion. The computer model of the human brain has likewise provided tremendous heuristic services as a source of hypotheses about human intelligence.

One and the same model may serve the purposes of both inventing and testing hypotheses. Consider using animals as analog models in biomedical research whose primary subject consists of human beings. Thanks to the many obvious similarities between such animals and humans, these models provide heuristic guidance by inspiring ideas about the functioning of the human organism. Depending on how similar the underlying causal mechanisms in animal and human organisms are believed to be, analog models may also be used for testing hypotheses about the behavior of human organisms. Rather than testing the impact of a newly developed medicine on humans, it is tested on animal models, then the test outcome is inferred also to apply to humans. This strategy faces the same issues that are generally involved in the use of experimental models to study nonexperimental reality. Are the conditions inside and outside the laboratory sufficiently similar in relevant respects to justify the transference of conclusions from one to the other?

The conventional idea of using theoretical models for testing theories is based on the view of models as specifications of theories. Empirical testing also requires models of the data, those simplified and idealized modifications of the complete set of experiences into the tractable body of relevant evidence that can be compared with theoretical models. Because there are multiple models of a given theory as well as multiple models of the complete data, no test can be perfectly tight and conclusive. There is always slack between the theory and the complete data, thus a failure of a model does not entail the failure of the theory. This observation is further fortified by the idea of models as relatively independent mediators.

Some models serve as maps, describing configurations of phenomena with different scales and degrees of detail, depending on the purpose for which

they are to be used. Some others serve to predict future phenomena by extrapolating past trends into the future. Other prediction models yield predictions of phenomena as responses to changes in exogenous variables. Yet other models are used to analyze the inner mechanisms of the generation of phenomena; often mathematical in form, such analytical models serve as theoretical machineries that are manipulated by the modeler in analogy to the causal manipulation of experimental machineries, with the purpose of identifying relevant causal mechanisms and examining their behavior.

Many models fail, and there may be a temptation to blame the very idea of modeling for this. But it is hard to think of modern science without models. Naturally, there are good models and there are bad ones, and many in between. Good models serve their purposes well, and bad models do not. A failure of a model is not because it is a model, but because it is a bad model for a given purpose.

4. *Models as Representations*

Given the multiplicity of types of model and their uses, one may ask whether there is anything that the various models have in common. A possible answer is that models are used to represent something beyond themselves. Representation has two related aspects: a model represents something else in that it resembles it in some respects, and in that it stands for that other as its 'representative.' Models can represent bits of the world as well as theories in these two ways. Models of X represent X and can be used instead of X itself to acquire information about X. One uses engineering models of airplanes to examine real—perhaps yet unbuilt—airplanes. Likewise, by studying the properties of theoretical models—with different degrees of imaginary contents—one may hope to learn about the real objects or the theories that they represent. One studies the model directly, and by doing so may learn indirectly about what the model represents. Thus, one may study the causal mechanisms of a real economic system by analyzing an imagined model economy populated by Robinson Crusoe and Man Friday; and one may study the biological mechanisms of a disease in human beings by experimenting with concrete (unimagined) rats. Many types of models are on a par in this respect—such as those that are concrete systems, those that are visualizable in terms of common experience, and those that consist of abstract mathematical formulas. What makes a model useful for its purpose is that it can be manipulated to draw informative conclusions about the ultimate target of study.

One may say, as Wartofsky (1979) does, that anything can serve as a model of anything else. This only requires that someone takes one thing to represent another thing in some relevant respects, where relevance is determined by the purpose of the representation. If the computer can be taken as a model of a human brain (and the human brain as a model of a computer) and a set of mathematical equations as a model of the Big Bang, so can a cookbook recipe serve as a model of moussaka, and Adolf Hitler as a model of evil. Models can be linguistic and nonlinguistic, abstract and concrete. One and the same object can be modeled by using a number of other objects. There are many things that can serve as models of the market institution, depending on the purpose of the model: a pair of scissors, a cobweb, a seventeenth-century painting of housewives buying freshly caught fish, a photograph of the New York Stock Exchange, the image of a telecommunication system, a geometrical diagram, or a system of differential equations.

In all such cases, one studies a complex phenomenon or system by representing it with a (much simpler) model. The model has far fewer properties than its subject, many of them idealized. The simple model serves as a substitute for the complex subject. It is within, or in terms of, the model that one recasts whatever questions one has about the complex entity to generate answers to these questions. The traditional philosophical issues about models are both difficult and inescapable. How do models relate to reality? How do the simplified questions and answers within a model relate to the complexities of what is being represented by the model? How do the manipulations of, and speculations about, models help us understand what they represent?

Consider models in one of the senses discussed earlier, as abstract and idealized systems that have only a limited number of idealized properties. It is not surprising that while statements about such models can be perfectly true, they can be also utterly false when applied to real, concrete systems. Real, concrete systems are far more complex than models and therefore behave in ways that cannot be truly represented by those statements. Interpreted as statements about real systems, the laws of science are false—and they remain false even after modifications such as from linear to non-linear equations. But other than this negative feature, how do model systems and real systems relate to one another? Giere suggests that models are representations that are connected to bits of the world by the relationship of similarity: models are similar or dissimilar to real objects in certain respects and to certain degrees. A further component—a theoretical hypothesis—makes claims about such respects and degrees of similarity. Unlike models, hypotheses are linguistic entities that refer to models and how models relate to real systems. Hypotheses in this sense can be true or false, depending on how well they succeed in their task. Models cannot be true or false, Giere believes, since they are not linguistic entities. But this is based on the unnecessarily restricted presupposition that truth bearers have to be linguistic. If we accept that truth and falsehood can be

ascribed to any representations, we can think of a model as possibly being partially true (insofar as it is similar to the real system in certain respects), and as approximately true (depending on the degree of similarity).

In assessing the merits and demerits of a model, the two aspects of representation must be considered jointly. On the one hand, the model should serve as a useful representative of its subject, providing a tractable substitute system that is amenable to a systematic examination yielding definite conclusions. And on the other hand, the model should resemble, or correspond to, its subject in certain respects and degrees closely enough relative to the goal of its purported use. The art of modeling is a matter of striking a balance between these two aspects of representation.

See also: Graphical Models: Overview; Metaphor and its Role in Social Thought: History of the Concept; Narrative Comprehension, Psychology of; Narratives and Accounts, in the Social and Behavioral Sciences; Rhetoric

Bibliography

Achinstein P 1968 *Concepts of Science*. Johns Hopkins University Press, Baltimore, MD

Black M 1962 *Models and Metaphors*. Cornell University Press, Ithaca, NY

Duhem P 1954 *The Aim and Structure of Physical Theory*. Princeton University Press, Princeton, NJ

Freudenthal H (ed.) 1961 *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences*. Reidel, Dordrecht, The Netherlands

Giere R N 1988 *Explaining Science*. University of Chicago Press, Chicago, IL

Harré R 1970 *The Principles of Scientific Thinking*. University of Chicago Press, Chicago, IL

Hesse M 1963 *Models and Analogies in Science*. Sheed and Ward, London

Morgan M, Morrison M (eds.) 1999 *Models as Mediators*. Cambridge University Press, Cambridge, UK

Wartofsky M 1979 *Models. Representation and the Scientific Understanding*. Reidel, Dordrecht, The Netherlands

U. Mäki

Moderator Variable: Methodology

Typically theoretical and empirical models in the behavioral sciences posit that independent variables affect dependent variables. A moderator variable is a variable, which is thought to temper or modulate the magnitude of the effect of an independent variable on a dependent one. Conceptually it is important to differentiate between a moderator and a mediator. A moderator is a variable that affects the magnitude of the relationship between the independent and dependent variables. It identifies the conditions under which, or the type of participant for whom, the effect is likely to be particularly large or particularly small. A mediator is a variable that describes the process that is responsible for the effect of the independent variable on the dependent one. It is a variable that is affected by the independent variable and in turn affects the dependent one, thus being responsible for the effect. (See *Mediating Variable*; also Baron and Kenny 1986).

1. Examples of Hypotheses Involving Moderation

Hypotheses about moderation are ubiquitous in the social and behavioral sciences. The following examples, taken from a variety of fields, provide illustrations. Each example consists of two sentences. The first hypothesizes an effect of an independent variable on a dependent one. The second sentence identifies a hypothesized moderator.

(a) Level of education attained affects lifelong earnings. This relationship is less strong for females than it is for males.

(b) Stressful life events tend to produce psychological problems. These effects are lessened if one has an extensive social support network.

(c) Some forms of instruction lead to better retention than others. These effects are larger among more able students.

(d) Exercise has pronounced health benefits. These benefits are more pronounced among older people.

Note in these examples that moderator variables can take on a variety of forms. They can refer to characteristics of the participants in the research (i.e., their gender, ability, or age); equally plausibly they may characterize the situations or environments which moderate the relationship between the independent and dependent variables. Additionally, their scale of measurement may be nominal (e.g., gender) or may vary more or less continuously (e.g., ability levels).

2. Conceptual and Analytic Issues in Testing Moderation

To say that one variable moderates the effect of another is equivalent to saying that there is a statistical interaction between the two variables. Accordingly, moderation implies interaction, although not all interactions involve moderating variables. The distinction is as follows: moderation means that there is a main effect of the independent variable plus an interaction between the independent variable and the moderator. Overall there is an effect of the independent variable, but the magnitude of that effect depends on the

moderator. An interaction, more generally, makes no claims about the presence or absence of the main effect of the independent variable. It simply indicates that the effect of one variable depends on another.

Because moderation implies a main effect of the independent variable plus an interaction of that variable with the moderator, typically interactions in the presence of moderators are not what are called crossover or disordinal interactions. That is, from one level of the moderator to others, the effect of the independent variable varies in magnitude but not in direction. To illustrate, consider the first of the examples given above, that gender moderates the relationship between educational attainment and lifetime earnings. For both genders, there is presumably a positive effect of educational attainment on earnings, but that positive effect is less strong among females than among males. For no level of the moderator does the relationship between educational attainment and earnings reverse.

Because moderation implies an interaction, it is necessary that the moderator and the independent variable be at least partially independent of each other. Ideally, the two variables would be fully crossed as in an experimental design where both variables are manipulated independently of each other. More typically, while the independent variable may be manipulated, the moderator may be a measured variable that varies more or less continuously. Again, however, to assess moderation it is important to insure that observations from all levels of the moderator occur at every level of the independent variable.

Analytically, moderation is typically assessed via either analysis of variance or multiple regression procedures, with the choice between the procedures traditionally depending on the scale of measurement of the independent and moderator variables. Two effects are of interest. First, the overall main effect of the independent variable is expected. Second, moderation of that effect implies that the interaction between the independent variable and the moderator is statistically significant. The use of analysis of variance to assess moderation is perfectly appropriate when both the independent variable and the moderator are measured on a nominal scale, with a few discrete levels of each variable. With moderators that vary more or less continuously, regression procedures are more appropriate, since substantial loss of power can occur if continuously varying moderators are broken up into discrete levels (e.g., using a median split) to accommodate them in analysis of variance. In general, multiple regression can be used to assess moderation, regardless of whether independent and moderator variables are scaled discretely or more or less continuously. In the case of discrete or nominal scaling, one simply must compute dummy coded or contrast coded variables to be included as predictors. The general model involves including as predictors the independent variable, the moderator variable, and a variable that is the product of the two. The test of the product variable, controlling for the two components of that product, provides a test of the statistical interaction (Cohen 1978). In this analysis, one must be cautious in interpreting the coefficients (and their associated statistical tests) of the component variables (i.e., the independent and moderator variables in a model that includes their product). Unless these variables have been centered on their means (and the product computed from those centered variables), their coefficients will not be readily interpretable as the average or main effect of each variable (Aiken and West 1991, Judd and McClelland 1989).

In some situations, the independent variable is likely to vary within units rather than between them. For instance, the effects of mood as an independent variable may be assessed by naturally occurring variations in mood within individuals over time. In these cases, traditional tests of statistical significance of both the main effects of the independent variable and its interaction with a moderator are likely to be biased unless dependencies in the data are appropriately handled. Repeated measures analysis of variance and, more generally, hierarchical linear models can be used to assess moderation in these sorts of designs.

Tests of moderation with naturally varying, as opposed to manipulated, independent and moderator variables are likely to have relatively low statistical power. Compared to the experimental situation where both variables are manipulated, there is likely to be less variation in the independent and moderator variables in the case where these variables are measured as they naturally occur. As a result the residual variation in their product is likely to be particularly small compared with the situation where both variables are manipulated. Tests of moderation (and interaction more generally) might then be most powerfully conducted by selecting units with relatively extreme values on the two variables (McClelland and Judd 1993).

3. *Interpretive Ambiguities*

As the examples in Sect. 1 illustrated, moderation hypotheses are ubiquitous in the social and behavioral sciences and analytic procedures for testing interactions, as described in the previous paragraphs, are relatively well worked out. However, interpretive difficulties can arise in describing moderation unless strong assumptions can be made about the metric of the dependent variable in any analysis. Nonlinear transformations of the dependent variable may eliminate interactions, especially when those interactions involve moderation of a main effect of an independent variable, rather than a crossover interaction.

Table 1
Means illustrating scale transformation that eliminates moderation

	Latency metric (tenths of sec.)		Speed metric (inverse transformation)	
	No Prime	Prime	No Prime	Prime
Familiar	0.556	0.541	0.0180	0.0185
Unfamiliar	0.833	0.800	0.0120	0.0125

Consider a simple example from cognitive psychology. Participants are timed while identifying letter strings as words or not (a lexical decision task). Words are either preceded by a semantically related prime or not. The main hypothesis is that primes should decrease the latency of the lexical response. Target words vary in their familiarity or frequency in the language. The effect of primes is expected to be moderated by word familiarity, such that prime effects are larger for less familiar target words. Hypothetical latency means (in seconds) are given in the left-hand portion of Table 1. Evidence is found for moderation: the prime versus no prime difference is more than twice as large for unfamiliar words (0.033 seconds) as it is for familiar ones (0.015 seconds). But now consider what happens if we convert the means to represent speed of response (rather than its latency) by taking the inverse of the latencies. In the speed metric, there is absolutely no evidence for moderation due to familiarity.

Which is the correct analysis? What conclusion about moderation is correct? Unfortunately, since there is no way to define the more appropriate response metric, there is no way to answer this question. The lesson is that many interactions in the social and behavioral sciences, where the metrics of the variables are somewhat arbitrary, disappear under nonlinear scale transformations. This is particularly likely to be the case when interactions are not crossover, as in the case of moderation.

To be confident about the presence of moderation, given arbitrarily defined metrics of most dependent variables in the social and behavioral sciences, analyses should be conducted using a variety of different metrics. If interactions prove to be significant under a variety of different nonlinear transformations, then confidence concerning the presence of moderation increases.

See also: Mediating Variable

Bibliography

Aiken L S, West S G 1991 Multiple regression: Testing and interpreting interactions. Sage, Newbury Park, CA

Baron R M, Kenny D A 1986 The moderator-mediator variable distinction in social psychological research. *Journal of Personality and Social Psychology* **51**: 1173–82

Cohen J 1978 Partialed products are interactions; partialed powers are curvecomponents. *Psychological Bulletin* **85**: 858–66

Judd C M, McClelland G H 1989 *Data Analysis: A Model Comparison Approach*. Harcourt, Brace, Jovanovich, San Diego, CA

McClelland G H, Judd C M 1993 Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin* **114**: 376–90

C. M. Judd

Modernity

If it is true that a notion is intuitive knowledge, synthetic and inaccurate enough about one thing, then modernity belongs to this type of mental representation which, as opposed to concept, does not offer clearly defined contours of the abstract object to which it refers. We commonly agree upon what must be understood by 'modern': that which appears, exists, or belongs to the current era or to a recent period, while 'modernism' expresses a preference against all traditionalism. Economists, sociologists, development experts also know what 'modernization' means when expressed in terms of politics and equipment plans. They agree with Jean Fourastié that, in the 1920s, to modernize a factory meant to install machines there and to submit the workers to their rhythms, whereas today 'it is to introduce into it a new working method, based, above all, on the means and faculties of men' (1969, p. 354). All those who deal with religious sciences know about the crisis of 'modernism,' which condenses the doctrines and movements advocating renewal and adaptation of Catholic theology in relation to modern scientific discoveries and the general evolution of the world at the end of the nineteenth century and at the beginning of the twentieth century. Finally, nobody can ignore that 'modern style,' 1900 style, 'noodle' style, is composed of floral ornaments intertwining iris, Virginia creeper and plane tree or chestnut tree leaves.

Numerous semantic samples, which relate to the modern or to what it specifies, could be attached to modernity. In particular, we should mention those which include ideas of custom, of avant-garde, and of doubt, as well as the refusal of traditional values. The complexity—paradoxically, as it would itself become the challenge—of modernity lies in the multiplicity of levels in which it is unfurled and in the uncertainty of the temporal framework which we tend to attribute to it. It appears to show solidarity towards 'progress' made in the field of science and technology; when applied to arts and culture, it can be conceived as being free from established rules, released from former

achievements, and of new creation, original, unpublished. It is possible that modernity saw the light of day in the sixteenth century: did not a French historian entitle one of his works *La Modernité du XVIe siècle* (*The Modernity of the Sixteenth Century*)? Leaving the modern period for the contemporary era, modernity emerges in the nineteenth century, when Chateaubriand created the term in a text dated 19 May 1833 and entitled *Mémoires d'Outre-tombe*, and in which he recounts that, on the outskirts of a town on the Danube,

> Vulgarity and modernity of customs and passports were contrasted with storm, the gothic door, the sound of a horn and the noise of a torrent.

It is also useful, from the beginning, to clear away the confusion which has surrounded the notion of modernity. We will therefore recognize the primacy of the aesthetic dimension which, at first sight, is nothing more than a rather vague constellation of ideas. Seen in this way, one might wonder how sociology should approach the subject, and which works should be consulted when we intend to examine its relationship with modernity.

1. *The Characteristics of Modernity*

Historically, modernity has a double origin: the Renaissance and the Reform. From the beginning, it was the demonstration of a crisis, the symptom of a transformation—and in no way its 'analyzer.' It must be recognized that Jean Baudrillard ably demonstrated that modernity is not an analytical concept, but rather, by converting the crisis into a value, it is a form of regulation—and first and foremost, a way towards civilization—supporting a new mythology. It is not enough to say that the latter is centered upon change; the ideological contents must be defined: these include categorical requirements which institute the 'tradition of novelty.' This argument owes nothing to Christopher Colombus, nor to Galileo Galilei, and very little to Luther and Calvin. The idea of 'rupture,' to which modernity provides the rhetoric is, however, associated with the great discoveries, with the passage from geocentricism to heliocentricism, and to the theses of Wittenberg.

The development of this rhetoric continues through a series of processes: world rationalization, secularization of thought, loss of one's fondness for community links, and assertion of individualism. In France, we are very aware of the strong moments of this movement. First, we have the quarrel of the Ancient and the Modern, in which the champions were, on the one hand, Boileau, Racine, Bossuet, La Bruyère, and, on the other, Quinault, Saint-Evremond, Fontenelle. The last of these was the author of *Digression sur les Anciens et les Modernes* which appeared during the same year (1688) that Perrault began his compilation of *Parallèles des Anciens et des Modernes* (1688–1692). Descartes had already affirmed that there is no reason

> to bow down before the Ancient because of their great age, it is more ourselves who should be called Ancient. The world is now older than before and we have a greater experience of things.

Here we see that the old image of dwarves perched upon the shoulders of giants has reappeared. In fact, it was not only a question of the specific literary value within the works of Antiquity. The opposition was especially interested in the effectiveness of progress of intelligence and morality: the Ancient remained attached to the idea of an original and 'natural' simplicity, while the Modern believed in the beneficial effects of civilization.

Another important step was the period known as the *Aufklärung*, the Enlightenment, the *Siècle des Lumières* which in France was the setting for the adventure of the Encyclopaedia. In a world plunging into obscurity due to the effacing of religious signals, there emerged a willingness to maintain an illuminated centre—and to assume the heritage—far from the models which had merely to be imitated. 'Modernity' provided support for ideas of progress and freedom: '*Laisser faire* (*les hommes*), *laisser passer* (*les merchandises*)' 'Let them do (men), let them pass (merchandise).' This injunction coined by Gournay, became the motto of economical liberalism, an expression of modernity. Free examination, free expansion, free pleasure in fact cumulated their impact at the time of the Enlightenment, but it was not until the next century, with the continuation of the democratic and industrial revolutions, that modernity acquired its full political and economic significance.

Without a doubt, modernity is the inseparable companion of industrialization, of urbanization, and of the introduction of regular salaries, which characterize a productive society wholly placed under the sign of the division of work. It is also an integral part of a progressive movement towards a mass consumer society based upon standardization. The new way of life that it engenders, of which dissociation of work place and living environment, as observed by Max Weber, is just one aspect, corresponds to the organization of the 'modern' state, the centralized, the bureaucratic state and the promotion of the individual as an abstraction, which in turn creates the combination, in the universal de-realization exposed by Marx, of the rationality of the state—henceforth equipped with a constitution—and of private interests—which are linked solidly to ownership. However, the concepts we have just discussed are not, objectively speaking, the most essential part of what we are seeking to understand. This can be found in the anxiety which overtakes the individual freed from the old magic-religious mechanism and which will, in the end, only be alienated by new networks of mediation.

Mobility has in fact reached a state of subjectivity and has rendered the identity of the subject uncertain, so that from then onwards it will be predisposed to instability.

2. *The Attractiveness of Modernity*

If we see modernity man installed in a time-specific, historical era, that is to say in a new temporality which apparently excludes eternity, we are again making him the vector or the product of an ideology of rupture. Within a cultural context where everything seems to be contemporary, current, and immediate, contradictions follow on from each other and opposites unite; e.g., standardization and the search for singularity, social integration and irruption of subjectivity, submission (to norms) and subversion (of forms), and so on. With the emergence of *avant-garde*, i.e., the aggressive emancipation of models, which had been considered legitimate until then, the destruction of traditional frames and perfect examples, we are nevertheless passing from a period of modernity viewed as a rhetoric of rupture to an attractiveness for change.

In a incisive study, *Prosa der Moderne* (1988), Peter Bürger perfectly demonstrated that

> if it is true that philosophical thought is blocked because it is incapable of considering the catastrophe by which it is obsessed, it is time to discover the other medium which allows one to see clearly within ourselves, in other words the prose of modernity.

Thinking laterally, he also gave prominence to how literary knowledge and sociological analysis can cross paths. Among the authors of literary modernity whose writings have the power to take their own presuppositions as a theme, we must place Charles Baudelaire. We will not, however, refer to P. Bürger's subtle comments (1988) about the allegory of the poet of *Fleurs du mal*, we will simply recall how the critic, in *Curiosités esthétiques*, identified modernity by assigning, for the first time, the meaning which remains attached to it today.

In 1863, *Le Figaro* consecrated its famous chronicle to modernity, as seen through the street scenes drawn by Constantin Guys, 'the painter of modern life.' The quality which should exist in the heart of all creation is identified: it is 'movement.' 'And thus he goes, he runs, he seeks' and what is this 'modern' artist looking for? The goal he is aiming for being

> other than the fugitive pleasure of circumstance (…), it is, for him, to release from customs what they can contain as poetical in the historical, to extract the eternal from the transitory.

Modernity 'is the transitory, the fugitive, the contingent, the half of art whose other half is eternal and unchanging.' It is the resolution of a conflict, the reduction of tension between the ever suspect outpouring of emotion and the unjustifiable steadiness of eternal beauty, the reconciliation—thanks to poetical works—of the requirement for emotional authenticity (via the insertion of work in the evanescence of daily life) and of a desire for eternity (which is satisfied by the freeing of the invariable from the contingent).

We can clearly detect the reasons for Baudelaire's strong delight in painting and music, and especially for the canvasses of clashing contrasts by Delacroix and the swelling music of the Wagnerian universe. There is an energy which gives dynamism to these works; it jostles and disrupts the emotional state of the person who contemplates or listens to them. Combining passion and control, Delacroix remains detached from all imitation of nature: he is the 'suggestive' painter and 'what he translates is the invisible, the impalpable, the dream, the nerves and the soul.' On this subject, Baudelaire observes that Breughel, with his dreamlike chaos, and Goya, the painter of the monstrous probable, figure among the initiators of modernity. Modernity is therefore neither topicality, nor contemporaneousness; one is less fond of it at a given time. It becomes an orientation towards a future-past and a discernment in the era of the 'provisionally eternal.'

At the same time, we are experiencing the assumption of imaginary. Towards this— which dilates as the symbolic becomes tense—there opens a space on canvass and score. A constructive imagination, the first constituent, governs its display which does not end with the imagination-deceptions generated by fancy. With passing movements of decomposition and recomposition, the form of the beautiful symbol of modernity is sketched, made up of strange things and devices which render the beautiful the antonym of the real. The caricature of Daumier, the hair-splitting of the universal exhibitions, and the calculated affectation of the dandy are equally make-ups, but not masking, of the real. During the Salon of 1859, Baudelaire gave meaning to the lie which is modernity, when he wrote that, after expressing his enchantment for the theater decor which concerned his artistic dreams

> these things, because they are false, are infinitely nearer to the truth, whereas most of our landscape painters are liars precisely because they forgot to lie.

We can also identify the reasons that led Walter Benjamin to see the passive stroller in the intoxicating crowd as the 'hero of modernity,' in his studies dedicated to Charles Baudelaire, *Ein Lyriker im Zeitalter des Hochkapitalismus* (1938–1939). This is surely the poet of modern society who, in the agitated throng of large towns, feels the loss of experience intensely: for which the instantaneous real-life experience of innumerable shocks was substituted, became permanent and could not be assimilated. In this situation, which is characterized by a new relationship

to time—about which modernity no longer recognizes the accomplishment, having lost all sense of recollection and festive moments—he can respond by miming death. He can also, like Baudelaire, search within the 'correspondences' for an experience which will be pursued and protected from, and far from, the modern clash. In a general way, the loss of the 'aura' will be converted into a poetical experience.

An earlier text by Walter Benjamin, *Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit* (1935) provides a definition for this tendency of modernity to be in constant opposition to uniqueness and distance. The masses appropriate images multiplied by technique; a new, democratic, community is formed, where the concepts of creation, genius, and style fall into disuse as we move from the value of worship to the value of expression. Growing reproducibility leads to the decline in ideas of authenticity and uniqueness, of the authority of the original all condensed into the 'aura,' which becomes a theological survival. In this perspective, modern art trivially signifies that by as many people as possible having access to its products. With the texts united within the title *Das Passagen-Werk* (1927–1940), Benjamin aimed for a deeper comprehension of modernity in the light of historical and dialectical materialism. Taking as examples, arcades, universal exhibitions, Baron Haussman's destruction and constructions, together with photography and the bourgeois interior, he makes the ambiguous demonstration of the process triggered by the fetish for merchandise. The creation of Parisian arcades, covered passages lined with shops and boutiques with their window displays full of goods, is identified as the moment when merchandise and technique seem to be liberated from art in order to appear free of all artifice, but are once again surrounded by borrowed elements. Art becomes technique and, from then on, is placed mainly at the service of the market. It is draped with dreams of the past and transfigures the defaults of the present into works which reveal the utopia of the future. Another art which became very fashionable during the nineteenth century was phantasmagorical art, which consists of showing phantoms through optical illusions in a dark room, leads to the assimilation of modernity with a phantasmagorical universe where signs free of all substance proliferate and where, at the same time, the primacy of form is asserted.

3. *Criticism of Modernity*

The aesthetic dimension of modernity and the society in which it is inscribed are both examined by Georg Simmel (1989–1990). In his opinion, the world of modernity was in opposition to that of classical culture which was considered to be orderly and containing hierarchical values. In this world, where custom was represented as the summit of the change to which it was perpetually dedicated, was born a new living place—the large town inhabited by masses of people—and a new style of behavior appears which was characterized by the modern individual. Individualism, its dominant trait, was no longer seen as a universal archetype, as it was in the ancient Latin and Christian world. The individual was conceived from now on as a totality, which German romanticism had glorified. Similarly, tragedy was no longer, as in ancient times, seen as a confrontation with an external destiny, but interiorised, so that the individual was not only man, but also victim and torturer.

In this configuration, which stems from the movements and the disorder which it designates, 'modernity' appears as a sum of transformations and contradictions that Simmel attempted to disentangle. People can be placed under the sign of conflict, thus providing one of the most important chapters of *Soziologie* in 1908 and were particularly well portrayed in art and religion. In order to persist, it is required to become consciously but implacably subjective within itself: this is considered as the stylization of life expressed through its liberation. Rodin was selected to be the representative of modern art and interpreter of modernity, in so far as his sculpture was the representation of movement—as an expression of the soul—in relation to a renewed conception of temporality. In the same way that time is an addition of separated instants, movement is composed of a succession of instantaneous postures. This idea leads to condemnation of mechanism and of naturalism, to the assertion that art, because it does not have to bear the 'train of reality,' does not have to copy reality. Art is an original conception of the aesthetically modern which is no longer that of the object, but of the subject, with its moods which are expressed in a portrait or a landscape.

All the Simmelian digressions on copy and creation, on distance and proximity, on the tragedy of culture and the intuition of life, are harmonised into a single philosophy of modernity which cannot be separated from general sociology. Lilyane Deroche-Gurcel (1997) perfectly demonstrated their complementarity, beginning with an explanation of how aesthetic points of view and sociological analyses come together for Simmel. Two dissatisfying conceptions of modernity in art are revealed by the author of *Rembrandt—Ein Kunstphilosophischer Versuch* (1916). One provides him with a matrix with as its only formal principle: the beautiful works are the girls and their shape. The other makes it the immediate expression of life: the original creations are born from strict contact with reality. Going beyond the hieraticism of art for art and the outburst of exuberant expressionism, Simmel saw in the interaction of life and form—an interaction which had been accomplished by both Rembrandt and Rodin—the outcome of modernity. It is this conception which led him to general criticisms or criticisms of detail: instant photography which can only

give a fixed expression of a reality, though it pretends to show movement; naturalism, which is constantly at work in the *avant-garde*, closed in by the childish rejection of tradition and the boundless search for novelty, and, finally, all dogmatism.

Taking into account the main characteristic of modernity, that is the multiplication of the interactions among individuals, the dissolution of stable contents into transient elements, Simmel concentrated on studying, not constituted society, but socialization. The substitution of movement for substantiality is in correlation with that of the *Vergesellschaftung* in the study of the *Gesellschaft*. It is also accompanied by the abandoning of the system for function, of the structure for process, of the status for role, behaviour, attitude. It supports the election of the fragment compared with the totality and of the form compared with the contents. In this respect, the sociology of Simmel, more than any other, has a hold on modernity which signifies, apart from the secession with stability, the loss of all substantial contents. Here we can observe the relationship which exists between modernity and melancholy.

What is significant, from this point of view, is the analysis of indifference: Simmel thought of it in terms of melancholy, by observing within certain aspects of socialization (customs, coquetry, adventure, sociability: the politeness of a detachment without despair) the resurging of melancholic syndromes within the species of their form—depersonalization and a-historicity—understood by their sociological meaning. Finally, Simmel underscores the ambivalence of modernity. Though movement in fact appears to be carried off in a flow of tormented forms, it in fact offers itself to the serene contemplation of the amateur who sees in the work of art of modernity, not the deceptive immobilization of life but rather its stylization.

Generally observed from above, modernity could just as well be observed from below, as fragile historical periodizations. Dedicated to post-modernity, the works of Colin Campbell (1987), David Harvey (1989), Sygmunt Bauman (1989, 1992, 1997), Barry Smart (1992) especially, draw in a hollow a representation of modernity. A Fordist method of production, a division of the society into classes and mass consummation would be the last traits to have been presented by modernity—finally discredited in, and by totalitarian barbarism. The postmodern era would therefore begin with the jamming of the effects generated by the rationalization of the world and would be followed by the coming of social configurations which would, from then on, be commanded by new information and communication techniques. Today, the deconstruction of what modernity edified would therefore be put into operation. In the field of urbanism and architecture, the cities of Brasilia and Chandigar show well what modernity meant during the twentieth century: a utopia of simplicity which does not recognize the complexity of life.

Retrospectively conceived in this way, modernity found itself deprived of many of its characteristics which were transferred to the age which closed it down: the mix of the present with the past, of the new with the archaic, of cultural relics with technical gadgets. However, subject to trial by a plurality of readings and interpretations and according to segments of severed time, modernity can less be a rupture than a response to this—by amalgamation. Above all, the postmodern point of view concerning modernity, whose horror is exposed, leaves behind the tragic which profoundly touched it. It is therefore a reducer, the better to reveal the return of complexity which led to its adoption.

In fact, modernity man is seized by this complexity. It is Ulrich, alias Musil, of *Der Mann ohne Eigenschaften* (1931), the man 'without quality' devoted to the juxtaposition of partial histories. Before the series of events crossed by the crisis of sense and before the difference between physical reality and psychic reality which is being increasingly demonstrated, the question is: how does one summarize it? How can we find the balance between the vital, the emotional, the affective and the control, the criticism, the censure? Within the Cacanie, Austria–Hungary mythical, modernity is an impossible synthesis. But it is also the uneasiness to which we are condemned as soon as, in the order of politics, a work interrupted by reduction of admiration or respect renders the calls for trust and esteem hardly audible and credible.

We should also distinguish two registers within the study of modernity. The one which we will qualify as 'ideological' is the most frequented by sociologists. In the ideology of modernity, some see a conservatism via change, others see a spectacular ambiguity and yet others see a culture of daily life. On another level, which, for want of a better term, we will call 'existential,' we must situate all that was produced with the weakening of subjectivity, with the assertion of individualists, with the perversity of human relationships, which are just as much aspects of modernity. It is at this level of analysis that we must position ourselves in order to examine the psychosociological consequences of self-constitution, according to which modernity was unveiled and which will always be, though in changing forms, the return of the same.

See also: Enlightenment; Modernity: Anthropological Aspects; Modernity: History of the Concept; Modernization and Modernity in History; Modernization, Sociological Theories of; Postmodernism: Philosophical Aspects; Reformation and Confessionalization; Renaissance; Simmel, Georg (1858–1918)

Bibliography

Baudrillard J 1968 *Le Système de objets*. Gallimard, Paris
Baudrillard J 1970 *La Société de consommation*. Denoël, Paris

Bauman Z 1989 *Modernity and the Holocaust*. Cornell University Press, Ithaca, NY
Bauman Z 1992 *Intimations of Postmodernity*. Routledge, London
Bauman Z 1997 *Postmodernity and its Discontents*. Cambridge University Press, New York
Bell D 1976 *The Cultural Contradictions of Capitalism*. Heinemann, London [1979 *Les Contradictions culturelles du capitalisme*. Presses Universitaires de France, Paris]
Benjamin W 1955, 1961, 1966 *Schriften, Illuminationen, Angelus Novus*. Suhrkamp, Frankfurt, Germany
Benjamin W 1969 *Charles Baudelaire*. Tiedemann R (ed.) Suhrkamp, Frankfurt, Germany
Burger P 1988 *Prosa der Moderne*. Suhrkamp, Frankfurt, Germany
Campbell C 1987 *The Romantic Ethic and the Spirit of Modern Consumerism*. Oxford, UK, London
Deroche-Gurcel L 1997 *Simmel et la modernité*. Presses Universitaires de France, Paris
Deroche-Gurcel L 2000 Avant-propos and Préface. In: Simmel G *Sociologie—Etudes sur les formes de la socialisation*. Presses Universitaires de France, Paris
Domenach J-M 1986 *Approches de la modernité*. Ellipses, Paris
Fourastié J 1969 *Le Grand Espoir du XXe siècle*.
Giddens A 1994 *Les Conséquences de la modernité*. L'Harmattan, Paris
Habermas J 1985 *Der philosophische Diskurs der Moderne*. Suhrkamp, Frankfurt, Germany
Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK
Jameson F 1991 In: Fish S (ed.) *Post-Modernism—or The Cultural Logic of Late Capitalism*. Duke University Press, Durham, NC
Lipovetsky G 1987 *L'Empire de l'éphémère*. Gallimard, Paris
Lyotard J-F 1985 *La Condition post-moderne*. Editions de Minuit, Paris
Maffesoli M 2000 *L'Instant éternel*. Denoël, Paris
Morin E 1962 *L'Esprit du temps–Essai sur la culture de masse*. Grasset, Paris
Musil R 1931 *Der Mann ohne Eigenschaften*. Rowohlt, Hamburg, Germany, Vol. 1
Rince D 1984 Baudelaire. In: de Beaumarchais J-P, Couty D, Rey A (eds.) *Dictionnaire des Littératures de langue française*. Bordas, Paris
Smart B 1992 *Modern Conditions, Postmodern Controversies*. Routledge, London
Touraine A 1992 *Critique de la modernité*. Fayard, Paris
Vieillard-Baron J-L 1989–1990 Introduction. In: Simmel G (ed.) *Philosophie de la modernité*. Payot, Paris, Vols. 1–2

B. Valade

Modernity: Anthropological Aspects

1. Modernity as Frame and Problem in Anthropology

The idea of modernity has operated as a historical framing for anthropology, but the categories of what is modern (radically new) in the human condition have shifted throughout the history of the discipline. At a simple level, we can say that the social sciences are generally about the study of modern societies, so why would anthropologists discuss modernity as an isolated area? We can say that anthropology is a *modern* study of human existence in which the *anthropos* becomes an object of knowledge and also a technique of modern power. For a discipline that produces its knowledge through an intimate experience of the (exotic) other, the field is particularly unique in that it is both a product, and an interrogator, of modernity.

This relationship of anthropology towards modernity has given anthropology its existential doubling, as an extension, and as an undoing, of the taken for granted aspects of Western modernity. For much of the twentieth century, anthropologists worked within the historical framing of modernity and its others, and used ethnographic findings to question the assumed aspects of Euro-American modernity and progress. The second distinctive aspect of anthropology's relation to modernity is as plural modernisms, as a way to capture the particularistic experiences of nonmodern or subaltern others in encounters with modern Western forms. The most promising approach has been to treat modernity as a specific ethnographic project, one that tracks the spread of political and social rationalities, and their production of new techniques, social forms, and subjects in a variety of ethnographic settings. Anthropology as a field and an ethnographer's craft is perhaps well suited to identify the ever shifting webs of rationalities that shape our plural worlds, and to pose the question of modernity itself as a paradox about humanity.

2. Modernity and its Others

From its very beginning as an academic discipline, anthropology has dealt with modernity as historical encounters between the modern and the nonmodern, and the social consequences that followed from such transformations. Initially driven by an evolutionary logic, social anthropology was dominated by an opposition between the so-called 'primitive' societies characterized by 'premodern' customs regarding magic, sexuality, and exchange which acted as a foil for modern ideas such as rationality, the Oedipus complex, the nuclear family, and the profit motive. There was also a more radical epistemological project to discover a cultural logic embedded within the native social system, giving it a positive valence that established a rationality alternative to that of European cultures. Anthropology thus studied those aspects of other cultures that stripped away the taken for granted nature of Euro-American modernity, and introduced a reflexive element into the 'science' of human knowledges. This implicit critique of the rational and scientific assumptions of European modernity gradually included the indictment of racism, as research

conducted in the shadow of European imperialism exposed its destruction of native cultures and its dependence on racism in colonial systems. Claude Lévi-Strauss ruminated on Europe's destruction of the premodern as inseparable from the loss of its own cultural ideals.

In the United States, the pioneers in anthropology participated in debates over the fate of traditional cultures under modern conditions. Nineteenth century 'salvage anthropology' engaged in a valiant attempt to record dying cultures, sometimes rendered in utopian terms. By the early twentieth century, Franz Boas established *cultural* anthropology as a field that spoke out against the human costs of Western conquest, colonialism, and racism. Boas sought to develop a scientific language of cultural differences and universal human dignity as inescapable aspects of modern humanity. During World War II, Margaret Mead and Ruth Benedict popularized the study of national identities, often contrasting the rationality of American society with premodern others.

In the post-Second World War era, classic social theory in the writings of Marx, Durkheim, Simmel, and Weber replaced earlier evolutionary perspectives on cultural change. Capitalism is now taken as the utterly revolutionary force that has created the modern world by radically breaking with the past. Market civilization tore asunder personal relations, destroyed communal social systems, and introduced egotistical calculations and the modern state to premodern worlds. Thus emerged a strong anthropological tradition to study the varied impact of the capitalist juggernaut on native social forms, subjectivity, and social change.

2.1 Commodity Production and Fetishism

Most scholars would agree that modernity began with the original capitalist transformations in the North Atlantic world, and their effects on the direction and limits of social change for all of humanity. For Marxist anthropologists, the most resonant story was the ways peoples of the new and ancient worlds were displaced and compelled to engage in commodity production. The works of Eric Wolf (1982) and Sidney Mintz (1985) show that exploitative market relations transformed colonized peoples into new kinds of alienated human beings, as commodity relations dissolved preexisting cultural relations among people, uprooting them from former ways of life.

A significant contribution of anthropology is to show that the market system, contrary to Marx and Karl Polanyi, did not gradually replace all other social forms, but rather interacted with other social systems in the differentiation of modern society. The 'modes of production' school demonstrated that kinship- and community-based systems in Africa, Latin America, and the Asia-Pacific became functionally articulated with the capitalist system (Godelier 1977). Ethnographies captured in vivid and concrete detail the parasitic dependence of capitalism on the labor of slaves, sharecroppers, cash-cropping peasants, and women in the developing world. Premodern labor forms and the domestication of women in far away places have historically been part of capitalism, and therefore a part of modernity itself. Forms of servitude such as guest workers and maids, often organized along gender, ethnic, and national lines, are an integral part of advanced capitalist societies. These cultural others furnish the images of the premodern Other to contemporary capitalism's own self-reflection.

The concept of commodity fetishism influenced ethnographic studies of how the market, bureaucracy, and mechanical media erode and reform indigenous notions of sociality and personhood. For over a century, anthropologists have decried the destruction or debasement of native cultures that came with capitalism, Christianity, and foreign rule. Anthropologists have documented the social despair, millennial movements, and rituals of violence inspired by missionaries, travelers, and traders. Another perspective mines the folklore and magical beliefs of colonized peoples for their hidden critique of the objectifying meanings of capitalism. The fetishization of market goods creates conditions for social competition, and gives rise to new hierarchies along lines of town and country, class and gender differences. Across the world, the mass production of culture, including soap operas and beauty contests, engender a variety of cultural configurations, displacing revolutionary sentiments with new desires and confusion.

2.2 Modernity and the Spread of Rationality

The Weberian view in anthropology dealt with the universal diffusion of the peculiar social rationalities of Western capitalism to traditional societies. Like much of the social sciences in the 1960s and 1970s, Weberian-inflected anthropology was framed by a modernization theory which ranked societies according to the degree of their adoption of modern Western values and institutions. In a series of studies on Indonesia, Clifford Geertz (1963) observed that the limited adoption of rational values and entrepreneurial forms such as capital accounting 'stranded' a country seemingly poised to take off into industrial capitalism. Geertz observed that Javanese cultural and economic spheres were firmly integrated by an overriding communal ethos that thwarted the rise of individualism and impersonal relations required by market organization. The Weberian assumption that economic rationalization would be gradually split off from other spheres was itself challenged by studies of advanced capitalist societies. Marshall Sahlins (1976)

argued that in American society, practical reason can be shown to be informed by religious and aesthetic schemes of meaning. But clearly, the project to decenter certain dominant Western conceptions of humanity must be embedded in an analysis of how these schemes of meaning are transformed in relation to the power dynamics of class, gender, and race.

3. The New Modernisms

By the early 1980s, the end of the Cold War spelled the decline of universalizing theories of modernization, and new generations of anthropologists rejected the totalizing, evolutionary assumptions about social change. The writings of Michael Foucault, and the rediscovery of Nietzsche, popularized understanding about the intimate relations between knowledge and power and focused intellectual projects not on things but on the conditions of emergence of new forms that have been built into the conduct of life. Foucault's *The Order of Things* (1970) first sets out the ways knowledge/power schemes shape and normalize objects of knowledge, and his later works have had a powerful effect on anthropology's analytical techniques and self-perception.

But the so-called postmodern challenge, and an influential text, *Anthropology as Cultural Critique* (Marcus and Fischer 1986), became symptomatic of the misrepresentation of the crisis of knowledge as simply a 'crisis of representation.' The self-reflexive turn caused many to scale down their anthropological project to an experiment in writing ethnography. First, a 'politics and poetics' of representation approach seeks to redress the temporal and political inequalities between ethnographic informants and the anthropologist (Clifford and Marcus 1986). Second, feminists—concerned about partial truths and informant empowerment—popularized multivocality to undercut the authority of the author. Third, a radical deconstructionist approach privileges language and writing over the theorizing of social forms. Such ethnographies seek to capture the jarring experiences of colonial and proletarian subjects, both to uncover the hidden 'truths' of subaltern suffering and resistance, and to subvert objectifying categories of description. Fourth, the so-called 'culture/power/history' orientation takes modernity as a historical cultural formation, embedding the study of discourses and cultural politics within the specific political economic contexts. These varied approaches seek to capture the particular conjunctures and disjunctures of discourse, culture, and power in encounters with Western power and knowledge. But unlike experimental ethnographies of earlier generations, the current postmodernist assault on the ideological constitution of Western social science does not offer an alternate method for studying schemes of practice, their logics and limits, and possibilities for transformation. Thus, postmodern approaches represent a detour in the anthropological theory of modernity. They do not develop a positive concept of modernity or analytical categories that can be applied to the systematic study of social forms that will allow for theoretical generalization.

4. Modernity as Political and Social Rationalities

The anthropological focus on modernity as an ethnographic problem in its own right was stimulated by recent European social thought. In American anthropology, with the help of his American interlocutors Herbert L. Dreyfus and Paul Rabinow, Michel Foucault's *oeuvre* came to have a wide-ranging influence on different approaches to questions of modernity. The fundamental question was not to treat modernity as an analytical opposition to given ethnographic realities, but rather to take modernity itself as an ethnographic object, as specific social and political rationalities configuring particular worlds.

Foucault's *Discipline and Punish* (1977) has been especially influential in the anthropological analysis of modern power. We can identify three conceptual clusters: (a) a positive (not repressive) 'capillary' notion of modern power as circulating practices arising from multiple domains; (b) the role of knowledge/power schemes in proliferating social norms and forms that produce subjects; and (c) the specific technologies of governmentality—grids of knowledge/power, mechanisms of surveillance— that govern our everyday behavior and make us particular kinds of modern human beings. Such a reconceptualization of modernity as projects and schemes of power is particularly suited to the ethnographic method of tracking the logics of everyday attitudes, habits, and practices (Bourdieu 1977). It becomes apparent that the goal is not to arrive at some grand theory about modernity, but rather to scrutinize the concrete manifestations of emerging social practices, norms, and cultural politics in relation to the market, the nation-state, and to globalizing forces.

4.1 Alternative Modernities

The Foucault effect on thinking about colonialism and postcolonialism shifts attention from reactionary cultural formations to strategic modern projects. More attention is now given to social rationalities—law, hygiene, sex, and labor disciplines—involved in civilizing colonized societies and subjects. The study of postcolonial development itself is reconceptualized as regimes of discursive power that have material effects on shaping the politics for remaking the Third World (Escobar 1995). The anthropology of imaginary modernities—narratives that depend on the creation of the otherness of their own pasts as well as of the West—show that in the new China, different generations have been shaped by visions of a movement

from a dark past into a bright socialist future. Indeed, for other countries, folk theories of modernity may even incorporate backwardness, thus subverting metropolitan development programs.

The concept of 'alternative modernities' risks reifying native views of modernity, and it will have more purchase if ethnographic attention is also paid to the specific social practices and forms that are distinctively modern. For instance, anthropologists note that the obsessive promotion of marginal cultural practices may engender a 'spectacular modernity' in societies greatly transformed by industrial capitalism and deeply anxious about the loss of national identity.

More grounded approaches have explored the disciplining effects of agricultural programs and bureaucracy in shaping developing societies (Ferguson 1990). Others consider how modernity becomes a state project in developing and postsocialist countries, focusing on the interactions between the market and other social systems that discipline civil society. Such ethnographies demonstrate how the novel forms of governmentality and governance, evolving in relation to globalizing forces, are dialectically linked to self-descriptions of alternative configurations of modernity.

4.2 Modern Schemes and Subjectivity

In *French Modern* (1989) Paul Rabinow opens up the space between the high cultures of modernity and the experiences of ordinary life as a domain of 'middling modernity' where social technologies encounter and regulate everyday sensibility. By seeking the coming into being of modernity in the mundane details of daily life, ethnographers present a complex picture of how the nondramatic norms and forms are constitutive of what it means to be human in particular societies. Perhaps anthropologists have gone further than other social scientists not only in grounding modernity as a particular set of rationalizing norms and practices, but also to specify the locations of different kinds of rationality in modern, universalizing institutions. Ethnographic scrutiny delineates how different regimes of power/knowledge have identified, processed, and constituted particular kinds of modern subjects in runaway factories, hospitals, and scientific laboratories and in utopian cities. Thus, modernity as an object of anthropological inquiry itself is a shifting target, as anthropologists track specific technologies in a range of ethnographic situations, and allow for comparison beyond cultural particularities.

4.3 Cities and the Reconfiguring of Citizenship

The metropolis, as Benjamin and Simmel have noted, is the *locus classicus* of modern sociality, the setting for the most extreme expressions of individual freedom and impersonal objectification. Earlier anthropological studies of urban life focused on the retention of village bonds among migrants or the maintenance of ego-centered networks among working class communities, as a defense against the alienating effects of modernity's flux. Indeed, more and more, cities rather than nations have become the salient sites for examining the gap between political citizenship and the experiences of different categories of persons who have been excluded or subordinated within the ideal national community. A central problem in conceptualizing modernity then is an ethnographic investigation of the mechanisms and routines of governance and the practices of subjects being thus constituted as administrative subjects of particular cities, zones, and nation-states.

There has been a spate of ethnographies on immigrants, drug-addicts, prisoners, street-persons, youths, the working poor, and racial minorities protesting their illusory juridical status and exclusions from the 'liberal community.' Generations of anthropologists have studied black ghetto culture and poor urban neighborhoods, stressing the differential forms of integration into American society. More recent works explore how marginalized urban groups formulate their demands for substantive rights—to schooling, housing, clean water, cultural difference, and respect—from the state and fellow citizens in Western democracies (Holston 1998). A concept of citizenship, not simply as imposed rights for abstract individuals, but as a set of contingent, negotiative relations between local authorities and self-making subjects, suggests a distinctly modern form of state-subject relations and subject-making.

4.4 Transnationalism, Differentiation, and Subject Formation

Of course, the modern condition now includes the proliferation of transnational networks that are transforming earlier understandings about community, mobility, and belonging. In our globalized world, modernity is increasingly understood and experienced in terms of transnational processes and webs of relationships shaped by the intensified flows of people, images, knowledge, and practices in everyday life. Such reconfigurations of social life and connections unsettle earlier anthropological notions of culture, community, and identity, suggesting that it may be possible to talk about a global ecumene spawned by intensified cross-border flows and exchanges (Hannerz 1992). An approach to the cultural dimensions of globalization (Appadurai 1996) suggests that the spatial and imagined coordinates of culture and community are now shaped by diasporic communities that are in the process of becoming 'postnational.' The

category of the transnational imaginary, however, is an indeterminable unit of analysis, and the visions of global companies, multilateral agencies, and diasporic networks are all rather different. The voluntary flows of tourists, intellectuals, expatriates, and businessmen should be theoretically distinguished from the forced homelessness of migrant workers and refugees.

How are such imagined communities constituted in relation to the variety of social systems that have been produced and proliferated by globalization? Accelerated border flows have multiplied specific configurations of subsystems that go to make up particular worlds. While sociologists have focused on the materiality of the new technologies—time–space compression and time–space distanciation—anthropologists seek to sharpen questions about what new modern cultural forms and subjects are emerging in different places, and what alternative cultural logics may be at work (Ong 2000). Transnational institutions and formations privilege new kinds of corporate subjects who internalize the norms of mobility, flexibility, and competitiveness (Martin 1994). A further direction investigates the rationalities that go into the invention of new techniques, practices, objects, places, and subjects, configurations in which the meaning of 'the human' is problematized (Rabinow 1998). Such projects will require us to go beneath the rhetoric of globalization and engage in multisited ethnography that can track the performative and migratory aspects of rationalities—in science, medicine, business, bureaucracy—that are shaping modern subjectivity in a variety of domains. Attention to new forms of governmentality enables us to ask what kinds of rules we have set up for ourselves to be considered modern human beings.

4.5 Living with Uncertainty

But the operation of rationalities is incomplete and contingent, and globalization has proliferated risks—volatile cash flows, unsettling information, diseases, toxic wastes, light arms, and bioscientific innovations—that expose everyone to the irrationalities of modernity. Capitalism has intensified the dominance of space and proliferated the social forms of modernity. Thus, in a world where the globalization of contingency is 'modern society's main attribute' (Luhmann 1998), our thinking needs to go beyond simply the perception and containment of risk. What remains to be investigated is how risks enter into the calculations of everyday life not as an independent but as a dependent variable.

There are signs that the future directions of an anthropology of modernity will concern itself with how societies and individuals live with uncertainties and absorb risks in their everyday routines. What do modern people problematize in their everyday life, and how do they solve problems? In the intersecting realms of health, science, morality, and capitalism, for instance, anthropologists have explored the ways people cope with the AIDS pandemic, the threats of nuclear plants, and the politics and moral perils of human genomic research. The study of divergent human fate as well will attend to how heterogeneous social spaces—global cities, production zones, blighted borderlands, postsocialist formations, refugee camps, and cross-border networks—are experimenting with different ways for ordering human sociality.

While the anthropology of modernity has a shifting target—the particularistic configurations of contemporary social life—there is a fundamental unity in the ethnographic diagnosis of the particular techniques, norms, and practices that go into the constitution of modern human subjects in specific fields of power. The ethnographic method is ideal for identifying new devices, arrangements, and techniques, and for attending to the performative aspects of emerging practices within an actor-network formulation (Latour and Woolgar 1986 [1979]). Our methodology as well must absorb the uncertainties of our human science, and theory must be based in an ever-widening spiral of observations beyond the immediate field situation. The challenge is for anthropologists to wed an ethnographic analysis of the practicable, strategic features we take to be characteristic of modern society with their interpretive, moral aspects.

For all the self-reflection and criticism, the project of anthropology remains one of a rigorous and systematic exposition of emerging worlds. Such a methodology—whether to study the interactions of functions systems, modes of ruling, modalities of subject-formation, biotechnological development, or the management of risks they all entail—allows for linking the particular with the general, the local with the global, the structural with the interpretive dimensions of modernity. By identifying the social and political rationalities constituting our plural worlds, we may ask whether we want to remain the kinds of human beings we have become. Modernity is both the condition and the decision about the future.

See also: Globalization and World Culture; Globalization, Anthropology of; Modernity; Modernity: History of the Concept; Modernization and Modernity in History; Modernization, Sociological Theories of; Multinational Corporations; Rationality in Society; Sex Differences in Pay

Bibliography

Appadurai A 1996 *Modernity at Large*. University of Minnesota Press, Minneapolis, MN

Bourdieu P 1977 *Outline of a Theory of Practice* [trans. Nice R]. Cambridge University Press, Cambridge, UK

Clifford J, Marcus G (eds.) 1986 *Writing Culture*. University of California Press, Berkeley, CA
Escobar A 1995 *Encountering Development*. Princeton University Press, Princeton, NJ
Ferguson J 1990 *The Anti-Politics Machine*. Cambridge University Press, Cambridge, UK
Foucault M 1970 *The Order of Things: An Archaeology of the Human Sciences* [trans. Sheridan A]. Tavistock, London
Foucault M 1977 [1975] *Discipline and Punish: The Birth of the Prison* [trans. Sheridan A]. Pantheon Books, New York
Geertz C 1963 *Agricultural Involution*. University of California Press, Berkeley, CA
Godelier M 1977 *Perspectives in Marxist Anthropology*. Cambridge University Press, Cambridge, UK
Hannerz U 1992 The global ecumene as a network of networks. In: Kuper A (ed.) *Conceptualizing Society*. Routledge, London
Holston J (ed.) 1998 *Cities and Citizenship*. Duke University Press, Durham, NC
Latour B, Woolgar S 1986 [1979] *Laboratory Life*. Princeton University Press, Princeton, NJ
Luhmann N 1998 *Observations on Modernity* [trans. Whobrey W]. Stanford University Press, Stanford, CA
Marcus G, Fischer M 1986 *Anthropology as Cultural Critique*. University of Chicago Press, Chicago
Martin E 1994 *Flexible Bodies*. Beacon Press, Boston
Mintz S 1985 *Sweetness and Power*. Viking, New York
Ong A 2000 *Flexible Citizenship*. Duke University Press, Durham, NC
Rabinow P 1989 *French Modern*. MIT Press, Cambridge, MA
Rabinow P 1998 *The Anthropology of Reason*. University of Chicago Press, Chicago
Sahlins M 1976 *Culture and Practical Reason*. University of Chicago Press, Chicago
Wolf E 1982 *Europe and the People Without History*. University of California Press, Berkeley, CA

A. Ong

Modernity: History of the Concept

In the social sciences, the term 'modernity'—or *modernité*, *modernità*, *Moderne*—in this form has been employed only very recently. Publications with this title have flourished only during the last two decades of the twentieth century. However, the sociological use of 'modernity' could draw on earlier concepts with the same root, in particular 'modern society' and 'modernization,' key terms in sociological work from the 1960s and 1970s drawing on related usage from certainly the nineteenth century onwards. It is therefore against the background of the broader meaning of the term 'modern' that the more recent usage of 'modernity' in the social sciences needs to be understood.

The social sciences of the early post-Second World War decades—often also known under the self-awarded label 'modern social sciences'—worked with the assumption that contemporary Western societies had emerged from earlier social configurations by way of a profound rupture. This rupture, although it could stretch over long periods and occurred in different societies at different points in time, regularly brought about a new set of institutions, most importantly a market-based economy, a democratic polity, and autonomous knowledge-producing institutions developing empirical-analytical sciences. Once such 'modern society' was established, a superior form of social organization was reached that contained all it needed to adapt successfully to changing circumstances. Thus, there would be no further major social transformation.

During the 1980s, it was exactly this key conviction of the modern social sciences that was challenged by the idea of 'post-modernity,' often understood as the assertion that Western societies had recently turned into an entirely new form of social configurations, based on novel forms of social bond. As such, the assertion was most prominently made in Lyotard's (1979) 'report on knowledge' titled *The Postmodern Condition*, but as a hypothesis of an ongoing major social transformation it has guided much sociological research since. In the course of the ensuing debate, further assumptions of the 'modern social sciences' and of their diagnosis of the contemporary social world were questioned. The term 'modernity' entered into the social sciences in response to such questioning.

1. *The Advent of Modernity as a Rupture in Historical Consciousness*

In its derivation from Latin, 'modern' is first of all a temporal term. It refers to the present—to be modern means to be within one's own time—and it implies a rather strong distinction of this present from the past. There are, however, three distinct ways of relating the present to the past by use of this term. First, most neutrally, 'modern' could refer to just what happens to be present: in this sense it would mean nothing else but 'contemporary.' Second, 'modern' could be used to deplore the loss of the greatness of the past. Such usage was common until, and including, the Renaissance. Third, the arrival of the 'modern' could be seen as an accomplishment, an overcoming of the limitations of the past.

In this latter form, the term was made prominent in *La quérelle des anciens et des modernes* in the seventeenth century. From then onwards, most of the uses of the term have retained such strong temporal implication, the later concept of 'modern society' being a key example. Drawing such distinction between eras, however, also demands specification as to how they differ, that is, a conceptualization of what is modern. Such conceptualization regularly transcends historical time, and thus invites analyses that go beyond the initial preference for the present over the

past. In other words, the term inevitably carries a double connotation; it is always both philosophical and empirical, or both substantive and temporal (Yack 1997), or both conceptual and historical (Wagner 2001).

The temporal, historical dimension has been in the focus of recent historiography of concepts. Historians of ideas and concepts have increasingly identified the period between the middle of the eighteenth and the middle of the nineteenth century as the one during which conceptual change in social and political language occurred in such a way in Europe that the horizon of the future was seen as open (Koselleck 1979). Despite the fact that strong distinctions between the past and the present were also made earlier, it is in this period that the rupture in historical consciousness was so profound that it appeared legitimate to see this period as the onset of modernity. This important finding, which has given rise to considerable further research and debate, has two profound implications for social science research on modernity.

First, it suggests that the social sciences co-emerged with modernity. Thus, on the one hand, modernity is the condition for the social sciences, as the reflexive knowledge of the human social world, to emerge and exist. On the other hand, it has become precisely the objective and task of the social sciences to conceptualize and analyze 'modernity' as a historical social formation. In the first meaning, modernity refers to a philosophical, in particular an epistemological condition. In the second meaning, it refers to a historical, empirical instance. Often, social science has been connected to modernity in terms of a correspondence of a form of knowledge to a sociopolitical reality, namely Western, 'modern' societies since the 1800s. Because of this co-emergence, however, it may be more appropriate to see social science as a specific way of interpreting a socio-historical experience, with which it is entangled. Or, to put it the other way round, social science, as manifold as it may often appear, may have occupied but a rather limited space among the available possibilities to theorize modernity. This suspicion has been confirmed in recent debates, as shall be shown below.

Second, while the historiography of concepts identified a significant shift in conceptual language, this finding on its own is insufficient to support the general claim that this period also marks the onset of 'modern society,' that is, of a whole new range of modern social practices and institutions. Sociologists of 'modern society' thus found themselves in the dilemma of having on the one hand to claim a historical break with 'traditional society,' but having on the other hand difficulties in empirically identifying such clear-cut breaks. Thus, spatio-temporal reference points were created and employed in various combinations. First of all, the idea of the emergence of 'modern society' has often referred straightforwardly to the 'history of the West.' No comparable major transformations were long identified in other parts of the world (but see Eisenstadt 1986 and now *Daedalus* 1998, 2000). More specifically, it has referred to the history of Europe, and from some time onwards also of North America. Modern social life then begins at the earliest in the mid- to late fifteenth century with the Renaissance, the invention of movable type and the voyages of discovery. At the latest, it starts in the early twentieth century with modernism in the arts and in architecture and with the new form of self-inspection provided by psychoanalysis. In between, economic modernity is said to begin with the market revolution and the industrial revolution, and political modernity has its take-off with the revolutions in America and France in the late eighteenth century. Scientific and philosophical revolutions towards modernity can be variously dated along this temporal line—from Cartesian rationalism and experimental method to the *fin-de-siècle* critique of science and metaphysics and to the theory of relativity (see, e.g., Toulmin 1990).

2. *Modernity as an Era and a Set of Institutions*

However one looks at this range of dates, a considerable tension between any historical description of a rupture and conceptual understandings of modernity comes immediately to the fore. The conceptual imagery of a 'modern society' characterized by a market-based economy, a democratic polity, and autonomous knowledge-producing institutions developing empirical-analytical sciences sits in an uneasy relation to these historical dates. Were one to insist that the full set of those institutions needs to exist before a society can be called modern, social modernity would be limited to a relatively small part of the globe during only a part of the twentieth century.

This tension between conceptuality and historicity was resolved by introducing an evolutionary logic in societal development. Based on the assumption of a societally effective voluntarism of human action, realms of social life were considered to have gradually separated from one another according to social functions. Religion, politics, the economy, and the arts all emerged as separate spheres in a series of historical breaks—known as the scientific, industrial, democratic revolutions, etc.—that follows a logic of differentiation (Parsons 1964, Alexander 1978). A sequence of otherwise contingent ruptures can thus be read as a history of progress, and the era of modernity emerges by an unfolding from very incomplete beginnings. In this view, indeed, modern society came to full fruition only in the US of the post-Second World War era, but 'modernization' processes were moving towards that *telos* for a long time, and continued to do so in other parts of the world (see *Modernization and Modernity in History*).

In conceptual terms, this perspective on modern social life aimed at combining an emphasis on free

human action with the achievement of greater mastery over the natural and social world. The differentiation of functions and their separate institutionalization was seen as both enhancing human freedom and as increasing the range of action. Thus, it provided a sociologized version of the Enlightenment combination of freedom and reason, or of subjectivity and rationality (e.g., Touraine 1992). Without this double concept being explicated in most of the sociological literature on 'modern society,' it nevertheless can be identified at the root of this conceptualization of modernity. At the same time, it certainly drew on what may be called the self-understanding of historical modernizers. Proponents of what came to be known as the scientific, industrial, and democratic revolutions saw themselves acting in the name of freedom, and they also saw the new institutions they were calling for as providing greater benefits than the old ones.

3. *The Interpretative Approach to Modernity*

The reference to autonomy and mastery has indeed more recently been explicitly used to conceptualize modernity beyond the sociology of modern societies from the 1960s and 1970s (Castoriadis 1990, Arnason 1989, Wagner 1994). Following Castoriadis, modernity can be considered as a situation in which the reference to autonomy and mastery provides for a double 'imaginary signification' of social life. By this term, Castoriadis refers to what more conventionally would be called a generally held belief or an 'interpretative pattern' (Arnason 1989). More precisely, the two components of this signification are the idea of the autonomy of the human being as the knowing and acting subject, on the one hand, and on the other, the idea of the rationality of the world, that is its principled intelligibility. Conceptually, therefore, modernity refers to a situation in which human beings do not accept any external guarantors, that is guarantors that they do not themselves posit, of their knowledge, of their political orders, or of their selfhood.

The difference between this interpretative approach to modernity and former sociological analyses is as follows. The sociology of modern society sought to derive a particular institutional structure from this double imaginary signification. Thus, the social sciences during the nineteenth and twentieth century were often inclined to consider a historically specific interpretation of a problématique as a general trait of modernity. Sociology, for instance, tended to conflate the historical form of the European nation state with the solution to, as it was often called, the problem of social order, which was expressed in the concept 'society' (Smelser 1997, Chap. 3).

When assuming, however, that a modern set of institutions can be derived from the imaginary signification of modernity, it is overlooked that the two elements of this signification are ambivalent each one on its own and tension-ridden between them. Therefore, the recent rethinking takes such tensions to open an interpretative space that is consistent with a variety of institutional forms. The relation between autonomy and mastery institutes an interpretative space that is to be specifically filled in each sociohistoric situation through struggles over the situation-grounded appropriate meaning. Theoretically, at least, there is always a plurality and diversity of interpretations of this space (see Skirbekk 1993).

On this broad basis, alternative possibilities to conceptualize modern social life and to analyze varieties of its institutional forms have been proposed in more recent debates. Before moving to discuss them, the historical critiques of the self-understanding of modernity, and much of the social science that accompanied it, need to be explored as parts of such interpretative struggle over modernity.

4. *The Grand Critiques of Modernity*

A series of major critical inquiries into the dynamics of modernity was elaborated successively from the middle of the nineteenth century up until the 1930s. Their characteristics as grand critiques reside in the fact that they identified basic problems in the practices of modernity, but did not abandon the commitment to modernity as a consequence. They all problematize, although in very different ways, the tension between the unleashing of the modern dynamics of freedom and rational mastery, on the one hand, and its, often unintended, collective outcome in the form of major societal institutions. As such, they provided critical interpretations of the imaginary signification of modernity, and some of the authors partly anticipated the more recent insight that modernity is characterized by a space of interpretative possibilities.

The first of these critiques was the critique of political economy as developed mainly by Karl Marx. In contrast to some of the conservative critics of capitalism, such as the German historical economists who flatly denounced its rationalist individualism, Marx basically adhered to the Enlightenment tradition of individual autonomy. His ideal was 'the free association of free human beings.' In the workings of the 'free' market in capitalism, however, he discovered a societal effect of human economic interaction that asserted itself 'behind the backs' of the actors.

In an economy based on market exchange and forced sale of labor power, relations between human beings would turn into relations between things, because they were mediated by commodities. Driven by laws of abstract value, markets would transform phenomena with a use value into commodities, the sole important criterion of which was the monetary value against which they could be exchanged. The result of such fetishization of products and money and of the reification of social relations would be the

alienation of human beings from their own products, from other human beings and from themselves. In such an alienated condition, the possibility for autonomy and sovereignty of the economic actors would be completely eradicated, though these actors would indeed constantly reproduce these conditions by their own action.

The second grand critique was the critique of large-scale organization and bureaucracy, as analyzed most prominently by Robert Michels and Max Weber. With a view to the enhancement of rational mastery of the world, it postulated the tendency for the formation of stratified bodies with hierarchical chains of command and generalized, abstract rules of action. In the context of a universal-suffrage polity and welfare state, that is in 'large' societies in which all individuals had to be included on a formal, that is, legally equal, basis in all major regulations, such 'iron cages' had emerged as state apparatuses, big industrial enterprises, and mass parties and would spread further in all realms of social life. While such institutions in fact enhanced the reach of human action generally, they limited it to the application of the rules, inside the cage so to say, at the same time.

In these terms, a variant of a critique of conceptions of rationality is the critique of modern philosophy and science, the third grand critique. Weber, too, was aware of the great loss the 'disenchantment of the world' in rational domination entailed, still he understood his own social science in rational and value-neutral terms, as he thought no other approach could prevail under conditions of modernity. In contrast, radical and explicit critiques of science were put forward by others in very different forms. In idealist *Lebensphilosophie* the elaboration of a nonscientistic approach to science was attempted as well as, differently, in early twentieth-century 'Western' Marxism, that is by Max Horkheimer and the early Frankfurt School. In some respects, pragmatism in the US can also be ranged under the critiques of science in as much as a new linkage of philosophy, anthropology and social science was proposed against the unfounded separation of spheres of knowledge in the disciplinary sciences. Such linkage would also bring the sciences back to a concern for the fundamental issues of the contemporary social world.

It was in pragmatism in particular—and in Europe in Durkheim's sociology—that a link between moral philosophy, social science, and politics was maintained, or rather recreated with a view to responding to the contemporary problems of societal restructuring. This link gave rise to the fourth critique, the critique of morality. The problem may be schematically reconstructed as follows. The development of modern society entailed the risk of moral impoverishment, mainly due to two phenomena. The inevitable decline of unquestioned faith eroded a source that could provide foundations for moral behavior. And if it is true that recurring face-to-face interaction is the basis for the solidarity-supporting insight in the human likeness of the other, such kind of interaction would be decreasingly relevant in mass societies integrated on the scale of a nation. The two questions that arise are, first, how to ground ethics at all, when no foundational criteria are universally accepted, and, second, how to develop adequate standards for morality, when social relations are predominantly 'thin' and at the same time widely extended in space and time, that is, to relatively distant others (see now Boltanski 1993). The requirements for ethics have been raised, while the likelihood to agree on any ethics at all may have diminished, in such a view. Again, it is the achievement of reflexively questioning any imposed standards of morality that may subvert the possibility of any standard at all.

Synthetically, then, an argumentative figure emerged as follows. In the historical development of modernity as 'liberal' society, the self-produced emergence of overarching structures, such as capitalism and the market, organization and bureaucracy, modern philosophy and science, and the division of labor, is identified. These structures work on the individual subjects and their possibilities for self-realization—up to the threat of self-cancellation of modernity. The more generalized modern practices will become, the more they themselves may undermine the realizability of modernity as a historical project.

5. *Modernity as an Ethos and an Experience*

The sociological interpretations of modernity provided by these critiques identified the tension between the modern orientations towards autonomy and towards mastery. But they tended to resolve this tension in a rather unambivalent way, namely as the institutionalization of autonomy inevitably leading to forms of mastery that would subject the 'free' human beings. Alienation, atomization, commodification, bureaucratization, and instrumental rationalization would assert themselves as absolutely dominant trends leading to the emergence of 'one-dimensional man' and 'one-dimensional society' (Herbert Marcuse). While this interpretation had some persuasive power, in particular during the first two-thirds of the twentieth century, in its totalizing way of reasoning it underestimated the persistence of the ambivalence of modernity and the possible resurgence of the quest for autonomy. Towards the end of the twentieth century, sociological diagnosis of the present indeed shifted back to an emphasis on individualization, rather than atomization, and reflexivity, rather than rationality (e.g., Anthony Giddens, Ulrich Beck, Alain Touraine).

Although such recent analyses of modernity tend to employ the terminology of a new era (in response to the challenge of 'postmodernity'), they indeed draw implicitly on a different concept of modernity altogether. A common view of the history of social life in

Europe holds that a 'culture of modernity' spread gradually since the 1500s. This 'is a culture which is individualist …: it prizes autonomy; it gives an important place to self-exploration; and its visions of the good life involve personal commitment' (Taylor 1989, p. 305). Such an emphasis on individuality and individualization is quite alien both to the more formalized 'modern' discourses of the individual as in rational choice theory or in liberal political philosophy and to the totalizing critiques of modernity.

In European intellectual and cultural history, there has long been very little connection between these two views of modernity and its inhabitants. Given their interest in institutions and their stability, political philosophy, and social theory proceeded predominantly by presupposition and showed little interest in actual human beings, who tend to be taken into account only as disturbances the more they enter the public scene. In literature and the arts, in contrast, the experience of modernity was in the center and, as an experience, it concerned in the first place the singular human being (Berman 1982). Foucault's lecture 'What is Enlightenment?' very succinctly distinguished between those two readings of modernity. Modernity as an attitude and experience demands the exploration of one's self, the task of separating out, 'from the contingency that has made us what we are, the possibility of no longer being, doing, or thinking what we are, do or think' (Foucault 1984, p. 46). It is counter-posed to modernity as an epoch and a set of institutions, which demands obedience to agreed-upon rules. At least in some writers, like Lyotard, the idea of post-modernity was inspired by such a return to what had been a modern self-understanding since at least the Enlightenment, and much less by the idea of a new era 'after' modernity.

6. *Varieties of Modernity*

The analysis of contemporary Western societies has long been based on the idea that those societies emerged through some rupture with the past. In this sense, the social sciences have long theorized 'modernity,' as the attempt to grasp the specificity of the present, even though the term has been used only rather recently. The dominant strand in the social sciences has aimed at capturing this specificity by structural-institutional analysis. The modern institutions are here seen as the embodiments of the modern promise of freedom and reason. Against and beyond this dominant strand, three different conceptualizations of modernity have been proposed. In parallel to the history of the 'modern social sciences,' the critiques of modernity have provided an alternative institutional analysis, emphasizing the undermining of the promise of autonomy in and through the workings of the modern institutions.

Both of these views recently have been considered too limited in their approach, namely in committing themselves to an overly specific understanding of modernity. The research and theory on 'modernity' that explicitly uses this term is by and large characterized by this insight, having emerged during the past quarter of a century. The interpretative approach to modernity has demonstrated the breadth of possible interpretations of what is commonly understood as the basic self-understanding, or imaginary signification, of modernity. The conception of modernity as an ethos and an experience has underlined the normative and agential features of modernity. In the former sense, it emphasizes the lack of any given foundations and the possibility to push the 'project of modernity' ever further. In the latter sense, it accentuates creativity and openness. In both ways, the experiential understanding complements the interpretative approach by underlining the large, potentially infinite, variety of interpretations of modernity. At the same time, however, it raises the stakes for an analysis of settings of modernity as entire societal configurations.

In attempts to combine these insights without, though, abandoning the objective of analyzing spatio-temporally extended configurations, research interest in what may be called the cultures of modernity has recently increased (Friese and Wagner 2000). Such research investigates the variety of sociohistorical interpretations of the double imaginary signification of modernity and the resources such interpretations draw on and mobilize. Within Western Europe already (e.g., between France and Germany) or within the more broadly defined 'West' (e.g., between Europe and the US), those resources are much richer and much more varied than earlier research has been able to identify (see Lamont and Thévenot 2000, Zimmermann et al. 1999). Both richness and variety increase considerably further as soon as one focuses on the so-called non-Western societies. Research in the 'varieties of modernity' or 'multiple modernities,' aims at analyzing such wider, present, and past, plurality of interpretations of the modern signification (Arnason 1998, Eisenstadt 1998). Such sociologies of modernity break with any reasoning that associates modernization unequivocally with Westernization. Without disregarding the problem of the 'specificity of the West,' that is, the Weberian problématique, interest is accordingly revived in the comparative-historical study of societal configurations.

See also: Cities: Capital, Global, and World; Critical Theory: Frankfurt School; Foucault, Michel (1926–84); Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Modernity; Modernity: Anthropological Aspects;

Modernization and Modernity in History; Modernization, Political: Development of the Concept; Modernization, Sociological Theories of; Religiosity: Modern; Sociology, History of

P. Wagner

Bibliography

Alexander J C 1978 Formal and substantive voluntarism in the work of Talcott Parsons: A theoretical and ideological reinterpretation. *American Sociological Review* **43**: 177–98

Arnason J P 1998 Multiple modernities and civilizational contexts: Reflections on the Japanese experience. Unpublished manuscript

Arnason J P 1989 The imaginary constitution of modernity. In: Busino G et al. (eds.) *Autonomie et autotransformation de la société. La philosophie militante de Cornelius Castoriadis*. Droz, Geneva pp. 323–37

Berman M 1982 *All that is Solid Melts into Air. The Experience of Modernity*. Simon and Schuster, New York

Boltanski L 1993 *La Souffrance à distance*. Métailié, Paris

Castoriadis C 1990 *Le Monde morcelé. Les Carrefours du labyrinthe III*. Seuil, Paris

Daedalus 1998 Early modernities. **127**(3): summer

Daedalus 2000 Multiple modernities. **129**(1): winter

Eisenstadt S N (ed.) 1986 *The Origins and Diversity of Axial-age Civilizations*. SUNY, Albany, NY

Eisenstadt S N 1998 *Antinomien der Moderne*. Suhrkamp, Frankfurt, Germany

Foucault M 1984 What is enlightenment? In: Rabinow P (ed.) *The Foucault Reader*. Penguin, London pp. 32–50

Friese H, Wagner P 2000 When the light of the great cultural problems moves on. On the possibility of a cultural theory of modernity. *Thesis Eleven* **61**(May): 25–40

Koselleck R 1979 *Vergangene Zukunft. Zur Semantik geschichtlicher Zeiten*. Suhrkamp, Frankfurt am Main, Germany

Lamont M, Thévenot L (eds.) 2000 *Rethinking Comparative Cultural Sociology. Polities and Repertoires of Evaluation in France and the United States*. Cambridge University Press, New York

Lyotard J-F 1979 *La Condition postmoderne*. Minuit, Paris [trans. Bennington G, Massumi B 1994 *The Postmodern Condition*. Manchester University Press, Manchester, UK]

Parsons T 1964 Evolutionary universals in society. *American Sociological Review* **29**: June

Skirbekk G 1993 *Rationality and Modernity. Essays in Philosophical Pragmatics*. Skandinavian University Press, Oslo, Norway

Smelser N 1997 *Problematics of Sociology*. University of California Press, Berkeley, CA

Taylor C 1989 *Sources of the Self*. Harvard University Press, Cambridge, MA

Toulmin S 1990 *Cosmopolis. The Hidden Agenda of Modernity*. The University of Chicago Press, Chicago

Touraine A 1992 *Critique de la modernité*. Fayard, Paris

Wagner P 1994 *A Sociology of Modernity. Liberty and Discipline*. Routledge, London

Wagner P 2001 *Theorizing Modernity. Inescapability and Attainability in Social Theory*. Sage, London

Yack B 1997 *The Fetishism of Modernities*. The University of Notre Dame Press, Notre Dame, Canada

Zimmermann B, Didry C, Wagner P (eds.) 1999 *Le Travail et la nation. Histoire croisée de la France et de l'Allemagne*. Editions de la Maison des Sciences de l'Homme, Paris

Modernization and Modernity in History

Modernization and modernity have emerged, particularly after World War II, as key concepts in Western social science. They refer to a long-term process of change primarily in the Western world, but also outside the Euro-Atlantic hemisphere, by which traditional, agrarian societies are transformed into the complex systems of industrial (and possibly democratic) societies typical of the late nineteenth and twentieth centuries ('modernization'). As a result of this process, societies and cultures develop a typical set of characteristics (e.g., in personality structure, organizations and institutions, and systems of belief and symbolic expression) that set them apart from older, premodern types of sociocultural systems ('modernity'). For many centuries, scholars have described and tried to explain—and also, to intellectually support—the process of modernization in terms of concepts such as social differentiation, rationalization, or economic development. From the 1950s onward, such concepts have been more explicitly formulated as social scientific theories of Western development, especially in American social science, but spreading very influentially from there to the academic cultures of other Western countries ('modernization theory'). As modernization describes a process of change in history, research in this field has proved to be a very important, almost paradigmatic area of cooperation and exchange between the social sciences and the discipline of history, and a particular branch of sociology called 'historical sociology' has centered around the concept of modernization as Westernization of societies, mostly during the past three to five centuries. Since the 1980s, modernization theories have increasingly been attacked—and often rejected—for their normative, Western-biased implications. At the same time, however, the concept of modernity has become more prominent in attempts at defining twentieth-century culture and its ambiguities. In addition, the 1990s have witnessed a renaissance of modernization theory in a reformulated framework of cultural skepticism and multiculturalism, an intellectual trend that is also significant for new attempts at 'bringing society back in' as the pendulum swings back from a radical emphasis on 'culture' in history and other social sciences.

1. Concepts of Modernity and Modernization before 1945

The term 'modern' (Latin *modernus*) has been in use since the fifth century AD to distinguish the 'new' or 'current' from the 'old' or 'antique' (Latin *antiquus*), especially as a means of describing and legitimizing new institutions, new legal rules, or new scholarly assumptions (Gumbrecht 1978). Thus, the term 'modern' has been carrying some normative implications from its very inception; to claim modernity for oneself meant a depreciation of the old or traditional and the attempt to win distance from it. In a very general sense and up to our own time, the concept of 'modern/modernity' is anchored in an attempt at defining one's own 'present' by differentiating it from the past. The conflict over the primacy of the new, or modern, over the old, or antique, has paradigmatically been fought out in the French *Querelles des anciens et des modernes* in the late seventeenth century; after the 'Querelles,' the superiority of the modern over the classic, or traditional, has rarely been fundamentally in doubt in the Western tradition of thought. The emphasis on the 'modern' gained additional support in the eighteenth century, when the enlightenment shifted historical thinking from cyclical to linear-progressive models. Since then, and especially after the American and French revolutions, the modern world was considered a world that in a radical and emphatic way was open into the future, a future that was to be attained by a natural progression of the cultural state of society.

In the wake of this fundamental shift, 'modern' was no longer just an opposition to the old or classic, but came to be regarded as the long-term process of change that was called 'modernization' in twentieth-century social science. In that sense, the so-called 'stage theories' of historical development brought forward most impressively by writers of the Scottish enlightenment (Smith, Ferguson, and others) can be seen as the first explicit and scientific modernization theories; at the same time, they were founded on the claim of a Western superiority over colonial or 'savage' societies. During the nineteenth century, when developmental and progressivist thinking became a dominant pattern, the idea of modernization became closely connected to evolutionary theory; the process of modernization was considered an ever-advancing evolution of Western traits leading toward a highly industrialized, rational, and bureaucratic social organization. This link to evolutionary theory prominently figured in the modernization theories of the mid-twentieth century, but has lost much of its appeal since then. Probably the most prominent, and certainly the most influential, variant of nineteenth century (evolutionary) modernization theories has been Karl Marx' and Friedrich Engels' historical-materialist theory of Western development. It provided a consistent view of the long-term processes of change, identified mechanisms of social change and progress (namely, the evolution of the productive forces and the subsequent change of social (class) relations), and it underscored the importance of the transition to a capitalist industrial economy in the nineteenth century. In doing so, it provided a model for later modernization theories, be they Marxist or anti-Marxist, and at the same time established a powerful 'master narrative' of Western history which patterned much historical writing in the twentieth century.

The most important bridge between those 'classical' theories of modernization (from the enlightenment to Marxism) and the more specifically scholarly-scientific approaches to modernization that developed in the mid-twentieth century was built by the founding fathers of modern sociology around the turn of the century, especially by Emile Durkheim and Max Weber. Weber's idea of modernization as occidental rationalization, empirically grounded in his political sociology and his comparative sociology of religion, proved enormously influential in systematic and historical sociology—from Talcott Parsons to Reinhard Bendix and S. N. Eisenstadt—as well as in history. Weber succeeded in combining scope and historical depth with analytical concepts; he evaded the pitfalls of Marxism by giving weight to cultural, religious, and intellectual factors as well as material factors in his view on the specificity of modern Western development. Although the high tide of classical Weberian sociology, and of the adoption of Weberian categories in historical writing, may have been over since the late 1970s or 1980s, his thought remains influential in recent attempts at reformulating the ideas of modernity and modernization, as Weber is now seen less as an advocate of progress and rationalization, and more as a shrewd observer of the deep ambiguities of modernization and modern life.

2. Modernization and Modernization Theory in the Social Sciences after 1945

After World War II, 'modernization theory' in a more limited, very specific sense emerged in the context of American social science, and American postwar political culture in general. With the 'American century' and the vision of a 'Pax americana' reaching its climax, economic, political, and social development elsewhere in the world was increasingly viewed in the perspective of progressive Americanization that would, in the end, lead to a homogeneous system of modern, industrial societies. The modernization of the West, and the US in particular, was regarded a model for the future development of 'Third World' countries, their agrarian societies, and authoritarian political structures. The optimistic mood of social-scientific theories in this vein, their sometimes crude and simplistic assumptions about modernization, and their some-

what intellectually 'imperialist' stance towards non-European cultures, have been much criticized since the late 1960s. In some areas of cultural theory and cultural history, they have been treated with contempt or, at best, with disregard. However, one should not ignore the fact that in the 1950s and 1960s modernization theory marked an important step in the professionalization of concepts of social change and development, and in the thinking about 'modernity' in general. Also, concepts of modernization developed in sociology and political science often radiated beyond the boundaries of their disciplines and proved very influential in other disciplines such as history.

The approaches of modernization theory between the 1950s and 1970s need not be described here in detail (for an overview and literature, see Lerner et al. 1968), but it may be useful to mention some fundamental concepts and assumptions, especially insofar as they inspired empirical work in historical sociology or history proper. The classical, post-1945 modernization theory crystallized at a time when structural-functional theory in sociology stood in high regard, and when social and cultural problems were often discussed in terms of a theory of social systems. Harvard sociologist Talcott Parsons best epitomizes this approach; like many others, he attempted to link the somewhat static assumptions of functionalism with more dynamic Weberian concepts of rationalization and evolutionary change in Western history (see, e.g., Parsons 1966). American modernization theory mostly concentrated on the transition from 'traditional society' to industrial capitalism in the economy, and on the transition to democracy and mass participation in the realm of politics. For Western societies, this implied a strong focus on the nineteenth century, a time period which at that time also became increasingly important in the work of historians in many countries. While quantitative approaches in modernization theory were rarely applied in empirical historical studies, qualitative analytical concepts like theories of economic growth (e.g., Walt W. Rostow, Alexander Gerschenkron) and of political modernization, especially of state-formation and nation-building (e.g., Karl W. Deutsch), proved attractive for historians interested in problems of social and institutional change.

In modernization theories of the 1950s and 1960s, a dichotomy of 'tradition' and 'modernity' often figured prominently; sociologists and political scientists established catalogs with typical features of 'traditional' societies (e.g., agrarian, localistic, and homogeneous) and contrasted them with features of 'modern' societies (e.g., urban, centralistic, and heterogeneous). Though this seemed to confirm a static, even ahistorical bias in modernization theory, the traditional-modern-dichotomy, which in many respects traces back to Ferdinand Toennies' distinction between *Gemeinschaft* and *Gesellschaft*, became a widespread analytical tool in history and historical sociology in the 1960s and 1970s. Concepts such as this, focusing on the transition from tradition to modernity through a series of crises which modernizing societies had to undergo, proved attractive to historians who were interested in the history of revolutions or in the emergence of participatory democracy. They also triggered a wave of empirical work in historical sociology between the late 1950s and the early 1970s in which historical processes of transition to modernity were analyzed, as in Neil Smelser's application of Parsonian theory to English industrialization (Smelser 1959). Other scholars made use of Weberian concepts of comparative change in Western and non-Western (e.g., China, Japan) societies (see, e.g., Eisenstadt 1973, Moore 1966). This line of research has never been completely cut off, but it grew thinner and, in the 1980s and 1990s, stood under an ever larger shadow of cultural theory and anthropology, that is, of disciplines and approaches which fundamentally questioned the assumptions of macrosociological, Western-oriented modernization theory.

3. *Modernization and Modernization Theory in Historical Work*

Historians in many Western countries have found theoretical concepts on modernity and modernization increasingly interesting since the 1950s, and especially during the 1960s and 1970s. There are several reasons for this. First, the postwar modernization theories were more than an intradisciplinary project; they served as an expression of a general mood and intellectual climate—of progress, of Western success, etc.—that did not leave academics from other disciplines, particularly in the humanities, unaffected. Second, the discipline of history at that time was very much in search for new approaches, for new master narratives, which would be able to support the ongoing broad expansion of history into the fields of social and economic history. At the same time, more historians became interested in comparative or international history, and modernization theory offered ways and concepts of pursuing such approaches. Third, the period of the 1950s to the 1970s was characterized by an intense cooperation between historians and social scientists; historians were looking for theoretical models in sociology and political science, while on the other hand, students of these disciplines attempted to treat their subjects—the formation of societies, the building of nation-states, or the emergence of party systems—with more historical depth (Kocka 1996). In a very general sense, in which 'modernization theory' is not understood as a coherent theoretical concept, but more as a diffuse historical 'plot structure' of societal modernization, economic progress, and political democratization, it may indeed be said that 'The quintessential professionalized historical discourse of

the first postwar decades was modernization theory' (Applegate 1999, p. 1162). However, this statement could also be made for much of nineteenth- and early twentieth-century liberal ('Whig') historiography; and even in the 1960s or 1970s, it was always a small minority of historians who made explicit use of social-scientific concepts of modernity and modernization in their empirical work. Also, national cultures and traditions of historiography diverged very much in that regard. While Anglo-American 'empiricist' culture generally worked as a barrier against the intrusion of theoretical concepts in historical writing, German scientific culture, with its strong traditions of fundamental theorizing, encouraged the adaptation of such concepts. American (US) and German historiography can serve as examples for a somewhat closer look at the role of modernization theory in the discipline of history.

3.1 USA

The fact that the US were the seedbed for the most influential modernization theories after 1945 might induce the assumption that concepts of modernization have been widely adopted by American historians, but that is not the case, at least not with modernization theories in a more limited, social-scientific meaning of that term. It is true that American historiography had a long tradition of 'progressivist' interpretation of its nation's history, even before the rise of the 'Progressive Historians' at the beginning of the twentieth century. Indeed, it is safe to say that a progressivist impulse constituted the very beginnings of American national historiography in the early nineteenth century (e.g., George Bancroft). After the nineteenth-century idea of progress had lost much of its original persuasion, modernization theory offered a somewhat more sophisticated replacement. Yet, as Dorothy Ross has maintained, 'Its ideological use, reductionism, and historical determinism made historians wary of it from the start,' and 'Only few historical studies have in fact openly claimed modernization as the theoretical basis of their work' (Ross 1998, p. 93). On the other hand, central assumptions of modernization theory often tacitly informed 1950s historiography, particularly in the then predominant school of 'consensus history,' which sought to establish an image of America as a country of conflict-free modernism. While it can be argued that in this perspective America had never been truly 'traditional' as European feudal societies were, influential consensus historians such as David Potter drew on social-scientific notions of a modernizing American character; Potter (1954), for example, explicitly referred to David Riesman's theory of a transition from 'inner-directed' to 'other-directed' man. Similarly, it is very common among American historians to point to a fundamental dichotomy between *Gemeinschaft* and *Gesellschaft*, and to use this distinction for describing modernization processes, or historical developments that lead in the direction of a growing social complexity and differentiation. In what can be regarded as one of the most important historical case studies on American modernization, Robert Wiebe has argued for a fundamental change in the organizing principles of American society in the decades around 1900, a change which he describes as the transition from 'island communities' with their communal values to a technological and industrialized order shaped by the bureaucratic mentality of a new, professional middle class (Wiebe 1967). Still, American historians have generally been hesitant in explicitly using modernization theories from sociology or political science; even more so since around 1980, when the culturalist turn in the humanities generated additional skepticism about master narratives of progress and linear development.

3.2 Germany

In parts of the West German historical profession during the 1960s, an intellectual climate very favorable to the reception of ideas on modernization in general, and of social-scientific theories of modernization in particular, developed. It was part of a broader, international tendency toward social and economic history in the postwar decades, at the expense of the more traditional political and diplomatic history, but probably even more important were factors specific to German history, and to the German political situation after the Third Reich. As a younger generation of academics began to orient themselves toward the West, especially the US, they opened up for new theoretical and interpretative concepts brought forward by American social science; and modernization theory with its broad perspectives on comparative social change was regarded as one of the most intellectually promising programs (Wehler 1975). The experience of fascism and the Holocaust in Germany lead many historians and social scientists to ask for long-term developments in German history responsible for the German inclination towards authoritarian government in the twentieth century, and the answer was found in failures of the modernization process in Germany especially since the mid-nineteenth century, after the revolution of 1848 and Bismarck's unification of Germany.

In a book enormously influential with the general public and professional historians alike, sociologist Ralf Dahrendorf eloquently elaborated the notion of Germany as a country being unable to adapt its institutional structures and its mentalities to the necessities of a 'modern' society. Paradoxically, Dahrendorf argued, only the results of the National Socialist regime finally threw Germany into modernity

(Dahrendorf 1965). In the following two or three decades, much of nineteenth- and twentieth-century German social history was influenced by an interpretive framework of a German '*Sonderweg*' into modernity, that is, a specific course of German history characterized by deficient modernization, with the transition to democracy and to modern civil values lagging behind other Western countries (see Kocka 1999). As German social historians, more so than their colleagues in the UK, France, or the US, were interested in theoretical concepts of social history and social change, they also undertook efforts at explicitly using social-scientific modernization theories in their empirical work and in their writing of history. The most ambitious and widely discussed project in this vein is Hans-Ulrich Wehler's multi-volume '*Deutsche Gesellschaftsgeschichte*,' a broad attempt at synthesizing modern German history with a focus on social and economic structures, and with a discussion of Germany's specific modernization problems as its red thread and intellectual core (Wehler 1987–95).

In Wehler's work as well as in German history in general, however, Weberian concepts of modernity and of comparative social change increasingly superseded, especially since the 1980s, concepts from contemporary American social science. A Weberian perspective seemed to offer more historical depth, going beyond the assumption of a sharp break between the 'traditional' and the 'modern' world in the nineteenth century. Even more important, as naively optimistic versions of modernization theory became less persuasive in the 1980s, Max Weber's emphasis on the ambiguities of modernization, on the complicated web of gains and losses called forth by the Western rationalization process, looked more attractive to many historians. In this perspective, the failures of German history culminating in the Third Reich were not simply caused by a modernization lag, but were, at least in part, inherent problems and results of modernity itself. This fundamental shift in the assessment of modernity and modernization, however, was not restricted to Germany, but can be traced as a general trend in historical and social-scientific thought since the 1980s.

4. *From Modernization to Modernity: Critique and Reappraisal Since the 1980s*

Since around 1980, postwar concepts of modernization and modernity have come under heavy attack, in history as well as in the social sciences and humanities in general. Within the overall trend toward a more critical, or skeptical, appraisal of modernization in history, it is important to distinguish several strands, even if they often overlap in historiographical practice.

4.1 *The Problem of Early Modern History*

In many Western countries, the history of what in English-language usage is called the 'early modern' period, that is, the history of the sixteenth to eighteenth centuries, has witnessed an unprecedented expansion since around 1970—particularly in Germany and the US, perhaps less so in the UK and France (where the history of this period is called '*histoire moderne*'). As social historians have taken a much closer look at societies which according to social-scientific modernization theory are 'traditional' and 'static' societies, it appears that Western societies before the industrial and democratic revolutions were not so traditional after all, but quite dynamic and expansive. This has been a heavy blow to theoretical and historical assumptions about a sharp break between traditional and modern societies that presumably happened (in North America and Western Europe) in the decades around 1800. It has thus prompted historians to ask for 'modern' structures in earlier times, and for continuities of 'tradition' in the industrial period, with the effect of further questioning historiographical models of modernization as teleological evolution and linear progression. Among interpretations of long-term change in Europe since the Middle Ages, German sociologist Norbert Elias' theory of the civilization process proved very stimulating among historians from the late 1970s, perhaps because his perspective on civilization is compatible with both 'optimistic' and 'pessimistic' viewpoints on modernization (Elias 1976). A 'pessimistic' reading of Elias focuses on the process of growing 'discipline,' self-control, and oppression, in many ways similar to Michel Foucault's understanding of modernity that became influential at about the same time. Especially in French and Italian historiography on the early modern era, modernization was now regarded as a strategic project of literate elites for the suppression of popular culture and ordinary people (see, for example, Muchembled 1978).

4.2 *Cultural Modernity in the Twentieth Century*

Along with the 'culturalist turn' in the humanities, with the emergence of cultural anthropology as a paradigmatic discipline for the whole field of humanities, the core meaning and content of the concept of 'modernity' has shifted from economy to culture. When historians, and scholars in general, talk about modernity, they now hardly ever refer to the economic modernity that originated in the European and North American industrial revolutions, as was the case between the 1950s and 1970s. Instead, they refer to the idea of a specifically modern culture, that is, to the idea that modernity is defined by certain attitudes and belief systems, by models of cultural expression as they

have been shaped, for example, by the mass media. It is important to recognize that the idea of modernity as culture is by no means new; after all, the '*Querelles*' of the seventeenth century were about culture, and a cultural approach to the understanding of modernity is central to writers like Baudelaire in the nineteenth century as well as to many Neo-Marxian strands of thought in the twentieth century, from Gramsci to Adorno and Habermas. In recent years, the current experience of a new information and knowledge revolution has directed attention towards the role of science and knowledge in the formation of modern culture; again, it can be said that modernity now appears to be less a product of economic forces, and more a product of systematic knowledge, its reproduction and its application through the emergence of a system of scientific disciplines.

Two further changes go along with this shift from economy to culture. First, instead of using the term 'modernization' (and hence emphasizing a processual change), many cultural theorists and historians now prefer to talk about 'modernity' or 'modernism' because they have become uncertain whether the notion of the modern and the idea of progress and development should be linked closely to one another, as was the case with the 'older' modernization theories. Second, although 'modernity' may be used very widely to characterize cultural innovation in different historical times, there is a clear propensity to concentrate on the twentieth century, and define the twentieth century as an age of modernity—whereas former modernization theories, as has been pointed out before, tended to see the nineteenth century as the crucial breakthrough period of the modern. Particularly in American and in German historiography, the beginning of the twentieth century is now regarded as a period of fundamental change in society and culture, as a time when modern, urban mass-society as we still today know it, first crystallized. Instead of pointing to industrial production forces, historians have come to view the beginnings of a consumer society as a strong indicator for modernity (see, e.g., Dumenil 1995).

4.3 Contradictions of Modernity

It has already been mentioned that postwar modernization theories have come under attack for their sometimes all too simplistic constructions, for their bias on Westernization, and for their claim that modernization unmistakably effected economic or political 'progress.' As historians have started to ask for the 'social costs' of modernization, for the experience of the 'losers' in this process, the balance sheet began to look different. However, the assumption now is often not just that modernization had some ambivalent consequences or side-effects, but that modernization and modernity in themselves carry some dangerous, anti-civilizatory and eventually catastrophic potentials. Like the concept of 'cultural modernity' sketched before, the notion of inherent contradictions of modernity is very much grounded in the historical experience of the twentieth century: namely, in the experience of dictatorship and totalitarianism, of mass murder and ethnic cleansing. Those dark sides of twentieth century history can no longer be persuasively explained as abnormal aberrations from an otherwise successful path of modernization; modernization and modernity rather look janus-faced; they bear in them the potentials for both civilization and barbarism, and there is no way of neatly separating those two sides. In the 1980s, twentieth century German history provided the first important paradigm for this interpretation. In research about Weimar and Nazi Germany, it turned out that modernism, or the quest for modernization, could be politically reactionary in both its intentions and its effects (Herf 1984). During the 1990s, a fierce debate was fought among German historians over the question of the ostensible 'modernity' of National Socialism (see Frei 1993). At the same time, the Holocaust perpetrated by Nazi Germany came to be seen as a paradigm for general contradictions of the modernization process; it hinted at contradictions not only apparent in German history, but also in the history of Western civilization as a whole. Sociologist Zygmunt Bauman in a widely received book maintained that 'The Holocaust was born and executed in our modern rational society' (Bauman 1989 p. X). Hence, the Holocaust poses a question not just to German history, but to 'Western civilization' as a whole. The Holocaust, in other words, in the 1990s often serves as a metaphor for the potentially murderous implications of modernity and modernization in the Western trajectory.

4.4 From Modernity to Postmodernism?

Instead of reformulating modernization into modernity and pointing at its contradictions, a second way of dealing with the widespread discontents with the postwar modernization theories in the 1980s was to leave the 'modern' behind and claim a new era of 'postmodernism,' and the reign of postmodern intellectual principles in the humanities and social sciences. In the fields of literature, political science, and cultural criticism, the impact of the postmodernist impulse has been enormous, but historical discourse has been much less influenced by it in all major Western countries. Although some claims about the 'end of history' or the 'end of modernity' resonated among historians, hardly any serious attempts were undertaken to conceptualize 'postmodernity' as a distinct epoch in the context of historical periodization. Nor have there been many efforts at describing certain social or cultural phenomena of the twentieth century under the heading of postmodernism in a way

that was comparable to the notion of a modernist culture that emerged during the same time.

As a methodological and theoretical concept, postmodernism was somewhat more successful. Borrowing from literary and cultural theory, historians in the late 1980s and 1990s discussed whether the time-honored claims of historiography to objectivity and factuality could still be sustained, and whether the narrative master-plots of history—of which 'modernization' was one of the most important—should be followed in historical writing at all (for a fair, yet critical discussion of these issues, see Appleby et al. 1994). In its resistance against the modernization master narrative of history, the postmodernist impulse shared some of the more general skepticism about linear-progressive theories of history and social development, but it soon turned out that any contemporary reformulation of modernization concepts, however critical it was considering the ecological crisis, ethnic mass murders, or the challenges of multiculturalism, would be best advised to stick to the category of modernity. For to claim a new, postmodern era did not eventually promise to be helpful in the attempt at explaining the often bitter consequences of modernity and modernization, an attempt that most historians and social scientists still find necessary.

5. *Toward a New Theory of Modernization?*

While historians are still debating the end of modernization, and the demise of the grand narrative of modernization in historical writing, the concept of modernization has witnessed a remarkable renaissance in sociology since the mid- to late-1980s (Mergel 1997, p. 203). The experiences of ecological crisis and catastrophe on the one hand, and of economic globalization and a worldwide diffusion of cultures on the other hand, has made sociologists think anew about societal change on a macro-level, on differences between tradition and modernity, and on the possible course and direction of social evolution at the turn of the millennium. The enormous dynamics of change in contemporary society is now being conceived as an equivalent to the major transition described by older modernization theories as a shift from agrarian to industrial societies. As the classical form of industrial societies is coming to an end, Ulrich Beck has argued, it is becoming 'tradition' itself, a tradition that is now undergoing a process of second-degree rationalization (Beck 1986). Industrial society was only 'half-modern' and is now being dissolved in a process called 'reflexive modernization,' because the modernization process is becoming reflexive by being reproduced and pushed forward in scientific and public discourses about the problems generated by it.

The concept of reflexive modernization has also been suggested by Anthony Giddens in what he calls an 'institutional analysis of modernity with cultural and epistemological overtones' (Giddens 1990, p. I, cf. Beck et al. 1994). For Giddens, the dynamism of modernity derives from (a) a separation of time and space ('time-space-distanciation'); (b) the 'disembedding' of social systems in an increasingly globalized world; and (c) a reflexive ordering and reordering of social relations, affected by continual inputs of knowledge (Giddens 1990, pp. 16–7). In this second stage, or 'radicalized modernity,' social change is not any longer comparable to a car controlled by the driver, but is like a 'juggernaut'—'a runaway engine of enormous power which, collectively as human beings, we can drive to some extent, but which also threatens to rush out of our control and which could rend itself asunder' (Giddens 1990, p. 139). In Giddens' view, globalization is not simply a spreading of Western culture and economy throughout the world; yet he firmly remains committed to the idea of modernity as a Western project.

Other authors, who argue from a 'post-colonial' political and epistemological perspective, are more skeptical about the Westernness of the ongoing modernization process. As traditional spatial boundaries disappear, cultures overlap and identities become hybrid creatures with different cultural backgrounds. The result, then, is not a continuation of Western modernity, but a new plurality of modernities ('multiple modernities') without a clear priority of the West (see Appadurai 1996).

However, as a common denominator of all these approaches, an emphasis on the spatial reordering of social relations and on the complicated relationship between globalization and local identity formation emerges. The new theories about reflexive and global modernization are important because they shift attention from a somewhat static notion of cultural modernity back to modernization as a process in time, and that is, among other things, what makes them particularly promising in the field of history. On the other hand, their understanding of historical processes, especially their assessment of social change in the period of the first (industrial) modernization and prior to it, is often superficial, and sometimes naïve. Historians are just beginning to discuss the new macro-sociological theories of modernization, and perhaps a fruitful dialogue between historians and sociologists will develop in the near future. The idea of modernization has weathered many storms and changed its appearance, but it seems that it still offers an indispensable clue to the historical and sociological analysis of long-term changes in society and culture.

See also: Differentiation: Social; Evolution, History of; Modernity; Modernity: Anthropological Aspects; Modernity: History of the Concept; Modernization, Political: Alternate Paths; Modernization, Political: Development of the Concept; Modernization, Sociological Theories of; Progress: History of the Concept; Social Evolution, Sociology of

Bibliography

Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN

Appleby J, Hunt L, Jacob M 1994 *Telling the Truth about History*. Norton, New York

Applegate C 1999 A Europe of regions: reflections on the history of sub-national places in modern times. *American Historical Review* **104**: 1157–82

Bauman Z 1989 *Modernity and the Holocaust*. Polity Press, Cambridge, UK

Beck U 1986 *Risikogesellschaft: Auf dem Weg in eine andere Moderne*. Frankfurt am Main, Suhrkamp, Germany

Beck U, Giddens A, Lash S 1994 *Reflexive Modernization: Politics, Tradition, and Aesthetics in the Modern Social Order*. Polity Press, Cambridge, UK

Dahrendorf R 1965 *Gesellschaft und Demokratie in Deutschland*. R. Piper, Munich, Germany

Dumenil L 1995 *The Modern Temper: American Culture and Society in the 1920s*, 1st edn. Hill and Wang, New York

Eisenstadt S N 1973 *Tradition, Change, and Modernity*. Wiley, New York

Elias N 1976 *Über den Prozess der Zivilisation*. 2 Vols. [The Civilizing Process]. Suhrkamp, Frankfurt, Germany

Frei N 1993 Wie modern war der Nationalsozialismus? [How modern was Nazism?] *Geschichte und Gesellschaft* **19**: 367–87

Giddens A 1990 *The Consequences of Modernity*. Stanford University Press, Stanford, CA

Gumbrecht H U 1978 Modern, Modernität, Moderne. In: Brunner O, Conze W, Koselleck R (eds.) *Geschichtliche Grundbegriffe*. E. Klett, Stuttgart, Germany, Vol. 4

Herf J 1984 *Reactionary Modernism: Technology, Culture, and Politics in Weimar and the Third Reich*. Cambridge University Press, Cambridge

Kocka J 1996 Annäherung und neue Distanz. Historiker und Sozialwissenschaftler seit den fünfziger Jahre. In: Hettling M, Nolte P (eds.) *Nation und Gesellschaft in Deutschland*. Beck, Munich, Germany

Kocka J 1999 Asymmetrical historical comparison: the case of the German 'Sonderweg'. *History and Theory* **38**: 40–50

Lerner D, Coleman J S, Dore R P 1968 Modernization. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan & Free Press, New York, Vol. 10

Mergel T 1997 Geht es weiterhin voran? Die Modernisierungstheorie auf dem Weg zu einer Theorie der Moderne. In: Mergel T, Welskopp T (eds.) *Geschichte zwischen Kultur und Gesellschaft*. Beck, Munich, Germany

Moore B Jr. 1966 *Social Origins of Dictatorship and Democracy*. Beacon Press, Boston

Muchembled R 1978 *Culture populaire et culture des élites dans la France moderne* (*XVe–XXVe siècle*). Flammarion, Paris

Parsons T 1966 *Societies: Evolutionary and Comparative Perspectives*. Prentice Hall, Englewood Cliffs, NJ

Potter D M 1954 *People of Plenty: Economic Abundance and the American Character*. University of Chicago Press, Chicago

Ross D 1998 The new and newer histories. Social theory and historiography in an American key. In: Molho A, Wood G S (eds.) *Imagined Histories. American Historians Interpret the Past*. Princeton University Press, Princeton, NJ

Smelser N J 1959 *Social Change in the Industrial Revolution: An Application of Theory to the Lancashire Cotton Industry 1770–1840*. Routledge and Paul, London

Wehler H-U 1975 *Modernisierungstheorie und Geschichte*. Vandenhoeck & Ruprecht, Göttingen, Germany

Wehler H-U 1987–95 *Deutsche Gesellschaftsgeschichte*. 3 Vols. C. H. Beck, Munich, Germany

Wiebe R H 1967 *The Search for Order 1877–1920*, 1st edn. Hill and Wang, New York

P. Nolte

Modernization, Political: Alternate Paths

Political modernization has three different, yet related, meanings. In political discourses, it refers to successful and enduring adaptation of a nontraditional political system to societal change in a way which secures system maintenance and political innovation. Political modernization may also be used of political development of a previously nonmodern or semi-modern political system to a higher level of political modernity. The most widespread use of political modernization is that of a long-term, deep-seated transformation of a nonmodern, mostly traditional political system in a post-traditional modern political order.

1. The Term Political Modernization

The defining characteristics of political modernization embrace above all (a) 'rationalization of authority' (i.e., the replacement of traditional political authority by single, secular, national political authority), (b) 'differentiation of new political functions and the development of specialized structures to perform these functions,' and (c) 'mass participation' in politics (Huntington 1966, p. 378). All these characteristics may be framed by a democratic political order, such as in a constitutional democracy, a semi-democratic order or a nondemocratic political system, as in the socialist countries in Central and Eastern Europe.

Political modernization may include the transformation of an autocracy in a democracy, or the transformation of a regime with limited political participation in a political system dominated by mass participation, continuous differentiation, unification and centralization of the political system, and competitive struggles for the acquisition of power. But it has also been common in modernization theories of the 1950s and 1960s to refer political modernization to socioeconomic and political development in nondemocracies, such as in the former socialist countries in Central and Eastern Europe, or in more or less advanced autocracies (see, for example, Rustow 1967, pp. 249–58, 272–82, and Zapf 1991). Despite the diversity of politically modern regimes, the most advanced states of political modernization are commonly attributed to the constitutional democracies of

post-World War II North American or West European origin.

In contrast to premodern political systems, the complexity of the political functions and the complexity of political survival are significantly larger within the context of advanced levels of political modernization. Following Samuel Eisenstadt's contribution to modernization theory (Eisenstadt 1973), this can be attributed largely to the twofold task confronting rulers within advanced political modernity: To cope simultaneously with objective pending problems, including mobilization of economic resources and taxation, and to mobilize permanent political support.

2. Routes to Political Modernization

Many routes lead to political modernization. The most ancient route has been that of Western Europe. That trajectory was mainly derived from, or overlaid with, conflicts of class. A second route was opened through the creation of overseas migration to the New Worlds of North America and Australasia. Political development in the 'colonial zone' (Therborn 1995, p. 6) marked a third path to political modernization. A fourth route has resided in those cases in which a policy of defensive modernization, initially adopted mainly for the purpose of strengthening the 'power state,' was ultimately conducive to democratizing the polity of the country, such as in Japan and South Korea, for example.

A fifth trajectory to political modernization has taken place in Central and Eastern European postcommunist countries. That path involves two modernization processes: First, authoritarian modernization until the fall of the Iron Curtain between Eastern and Western Europe in 1990, and, second, advanced 'modernization of modern states' (Zapf 1991) in the period after 1990. Thus, political modernization exists within diverse political, cultural, and economic contexts. It is by no means confined to socioeconomically developed or relatively rich nations, and it is not an exclusive trademark of democratic countries.

3. Simple and Complex Patterns of Political Modernization

The plurality of routes that led to political modernization, and the heterogeneity of the outcome, have challenged the conventional wisdom of socioeconomic theories of modernization. In a stylized fashion, one can distinguish between theories of 'simple' and 'complex' patterns of political modernization. Theories of simple political modernization patterns tend to regard modernization of the polity as relatively uniform, homogeneous, stable, irreversible, and largely free of major crises. This family of theories also expected more benefits than costs from political modernization. In contrast to this, theories of complex modernization patterns tend to conceptualize the process of political modernization as heterogeneous, pluralistic, precarious, crisis-ridden, and reversible. Moreover, these theories also focus attention on side effects, unintended consequences, and massive social costs associated with political modernization. According to this view, political modernization does not necessarily generate a fairer and more efficient political system.

The shift in focus from simple to complex patterns of political modernization mirrors that political modernization on the one hand, and socioeconomic development as well as political performance on the other, are only loosely interconnected. For example, economic and social development may be associated with modernity or traditionalism, and with democratization or maintenance of authoritarian regimes. The latter is exemplified by the former socialist countries in Central and Eastern Europe; contemporaneous examples include the economically advanced oil-exporting states in the Middle East. Furthermore, democratic modernization can be achieved not only at a relatively advanced stage of socioeconomic development, but also on a far lower level. With development, A. Przeworski and F. Limongi (1997, p. 178) have argued, democracy can flourish in poor countries. This view is evidenced by both the democratization of former Third World countries in the 1970s, 1980s, and 1990s, and the maintenance of democracy in these states. Finally, a relatively advanced stage of socioeconomic development does not necessarily preclude the breakdown of democracy. The collapse of the Weimar Republic in Germany in 1933 is a particularly dramatic example of the vulnerability and fragility of a democracy.

4. Determinants of Successful Paths to Political Modernization

The plurality of routes to political modernity is mirrored by a wide variety of factors that are conducive to, or inhibit, the effort to modernize a country politically. None of the influences on political modernization is related deterministically to successful or unsuccessful outcomes. Among the factors which facilitate political modernization, sustained economic growth deserves to receive the foremost mention, while the level of economic development is, beyond a certain threshold, not decisive. But sustained economic stagnation or a long period of economic decline is likely to block every route to political modernization. The driving forces of successful political modernization include a political culture in which civil and social rights of individuals are emphasized, and where the right to participate in political life is

regarded as a value. Moreover, the chances for successful political modernization tend to be higher in ethnically relatively homogeneous countries, and lower in more heterogeneous states. However, appropriate techniques of compromise building, such as consociationalism, can bridge sharp ethnic differences. The determinants of political modernization outcomes include political action on the part of collective and individual actors, as well as actor constellations. For example, the distribution of power between traditional and modernization-oriented elites is of importance, and so, too, is the distribution of power between elites and counter-elites. Large protest movements demanding political modernization are also relevant players. Supportive factors include coalitions between social classes which are conducive to democracy, such as the coalition between labor and an independent bourgeoisie, or 'red–green' coalitions between labor and independent peasants. The degree of affinity between modernization-oriented elites and major traditional values, as well as the preferences of the most important societal groups, also deserve to receive mention. Furthermore, international factors make a difference. Democratic political modernization, for example, flourishes within the context of a democracy-friendly international environment. In contrast to this, an international environment full of enemies of democracies imposes major restrictions on democratic inauguration and consolidation. Deep-seated cultural traditions have long been regarded as obstacles to political modernization. However, more recent theories point out that cultural traditions can also contribute to political modernization, such as the case of defensive modernization in a country ruled by traditional elites demonstrates (see, for example, the Japanese case in the nineteenth and twentieth centuries).

Like other complex and dynamic processes, political modernization generates different types of outcome—intended and unintended. In contrast to the first generation of modernization theory in the 1950s and 1960s, later generations place more emphasis on unintended, if not self-destructive, outcomes, such as the strengthening of revolutionary tendencies through reforms, or self-destructive costs of rapid modernization, or generation of risks for future generations, such as high ecological costs, or irreversible policy. This tendency manifests itself most clearly in the theory of 'risk society' (Beck 1986) and the theory of 'reflexive modernization' (Beck et al. 1997).

See also: Democracy; Democracy, History of; Development and the State; Development: Social-anthropological Aspects; Development: Socioeconomic Aspects; Differentiation: Social; Modernization, Political: Development of the Concept; Rights; Secularization; State Formation; State, History of; State, Sociology of the

Bibliography

Apter D E 1965 *The Politics of Modernization*. University of Chicago Press, Chicago, IL

Beck U 1986 *Risikogesellschaft. Auf dem Weg in eine andere Moderne*. Suhrkamp, Frankfurt (*Risk Society: Towards A New Modernity*. Sage, London 1992)

Beck U, Giddens A, Lash S (eds.) 1997 *Reflexive Modernization. Politics, Tradition, and Aesthetics in the Modern Social Order*. Polity Press, Cambridge, UK

Eisenstadt S N 1973 *Tradition, Change, and Modernity*. Wiley, New York

Giddens A 1984 *The Consequences of Modernity*. Cambridge University Press, Cambridge, UK

Huntington S P 1966 Political modernization. America versus Europe. *World Politics* **18**: 378–414

Huntington S P 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, CT

Inglehart R 1997 *Modernization and Postmodernization. Cultural, Economic and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ

Lipset S M 1994 *Political Man. The Social Basis of Politics*, 5th edn. Johns Hopkins University Press, Baltimore, MD

Moore B 1965 *Social Origins of Dictatorship and Democracy*. Beacon Press, Boston, MA

Nohlen D, Nuscheler F (eds.) 1993 *Handbuch der Dritten Welt*, 3rd edn. Dietz, Bonn, Vol. 1

O'Donnell G 1973 *Modernization and Bureaucratic Authoritarianism Studies in South American Politics*. Berkeley Institute of International Studies, University of California, CA

Przeworski A, Limongi F 1997 Modernization. Theories and facts. *World Politics* **49**: 155–83

Rueschemeyer D, Huber-Stephens E, Stephens J D 1992 *Capitalist Development and Democracy*. Polity Press, Cambridge, UK

Rustow D A 1967 *A World of Nations. Problems of Political Modernization*. Brookings Institution, Washington, DC

Schmidt M G 2000 *Demokratietheorien. Eine Einführung*. 3rd edn. Leske and Budrich, Opladen, Germany

Therborn G 1995 *European Modernity and Beyond. The Trajectory of European Societies 1945–2000*. Sage, London

Weber M 1972 *Gesammelte Aufsätze zur Religionssoziologie*, 3 vols, 6th edn. Mohr & Siebeck, Tübingen, Germany

Zapf W (ed.) 1991 *Die Modernisierung moderner Gesellschaften*. Campus, Frankfurt, Germany

M. G. Schmidt

Modernization, Political: Development of the Concept

Political modernization, as a concept in the political science literature, embraces today changes in politics and government in individual countries and states derived from major shifts in technology. These changes generally involve significant shifts in the content of work, attitudinal changes, and organizational changes that place pressure on governmental

institutions to develop the capacity to process new demands and issues. While the term is frequently used in discourse regarding economic and political development, by retaining a base in the response to revolutions in technology it is possible to establish greater precision and hence a better understanding of what is entailed in these changes in politics and government.

1. Definitional Problems

Difficulties in distinguishing between political modernization and political development remain very much linked with the revival of the comparative politics field in the 1960s in the shift from the study of political institutions to political behavior. An important subset of this literature is studies of political development and modernization. Lack of precision in the use of 'development' and 'modernization' led to considerable debate during the 1970s. The core issue centered around whether or not this literature was identified with the imposition of Western values and perceptions of politics on the developing world. As a consequence, during the 1980s most standard texts in the field opted for a discussion of political change, reviewed the literature on development and modernization, and focused their attention on distinguishing among democratic and nondemocratic forms of governance. But during the 1990s yet another revolution in technology linked to information technology, the growth of knowledge-based industries, and the globalization of economic processes produced major realignment in politics and economics, the outcome of which has been a revival of interest in political modernization. Today discussions of modernization are closely tied to those cultural, economic, and political changes that are shifting the focus of this discourse away from national states to global perspectives on change in the contemporary world.

2. Evolution of the Concept

Gabriel A. Almond played a major role in comparative politics in the 1960s by calling for more work on and attention to the dynamics of politics in the developing word. Critical of the focus of scholarship at the time on formal institutions in the advanced industrial countries, Almond in particular looked to the discipline of sociology for constructs and ideas more relevant to identifying patterns and processes in politics tied to political behavior. In the outpouring of new research on the developing countries during this decade, an important subset became identified with studies of political development and modernization. While most of this work used these terms interchangeably, a number of scholars attempted to introduce greater precision by arguing that the source of change in politics, economics, and society was technological change and that one needed to clarify exactly what was the nature of this change.

Focusing attention on the industrial revolution and the way in which it had transformed Western European societies and politics, the political modernization literature centered attention on the differential impact of these events on politics outside Europe and North America. A key idea in the political modernization literature of the 1960s and 1970s was that revolutionary changes in the nature of work triggered first by manufacturing and later mass-based industries necessitated fundamental changes in the way in which government was organized and how people thought about politics. Drawing on the work of Marion Levy, C. E. Black, David Apter, and Dankwart Rustow focused on different aspects of modernization. The problem they identified was the fact that the impact of the industrial revolution on traditional societies, far removed from the source of economic transformations, had generated a very different set of dynamics. But scholars differed considerably in how they explained the consequence of these changes. Generally speaking, the literature on political development took a broader focus in seeking to explain the evolution of political institutions, but in the process opened up a debate over the values that guided this research and writing, especially in its reliance on Anglo–US norms. This was especially the case with Gabriel Almond's and Samuel Huntington's writings on development and modernization in the developing countries.

3. Advances in Modernization Theory

The response to the concentration of this new industrial order first in Western Europe, its expansion to North America, followed by the emergence of modern Japan, triggered a debate over the sources of differential development levels. Especially after World War II, it became clear that some national societies were better able than others to meet the new challenges posed by these shifts in technology. Advances in science and technology enabled the early comers to the industrial revolution to respond to new breakthroughs in the organization of work, production, and scientific knowledge which increasingly separated them from the rest of the world. But much of the world had not accompanied these changes and could not easily establish a basis of self-sustaining economic activity that could respond to these changes in technology.

This newer literature, as can be seen in Ronald Ingelhart's work, has sought empirically to establish the interrelationships between cultural, economic, and political change. By comparing and contrasting these interactions in a broader array of countries, this work has moved beyond a simple dichotomy of the world into developed and developing countries. Neverthe-

less, a debate continues over whether or not these perspectives on development (largely heuristic) and modernization (mostly empirical) are closely identified with competing cultural constructs of reality, especially when one enters the political world. Hence, in reading works in this genre it is always useful to keep in mind that many scholars continue to see an overlap among these concepts of development, modernization, and westernization. This concern has continued from the 1960s down through the 1990s, as can be seen in the recent translation of Bertrand Badie's *Etat Importé* (2000).

The political modernization literature, consequently, has gained new life with the new emphasis on the globalization of economic forces and the internationalization of policy issues linked to development and modernization, as cutting edge economic concerns have shifted from mass industries to knowledge-based industries identified with computer-based technologies. In the process a whole new literature has emerged which links structural adjustment economic policies, in the creation of market economies worldwide, with fundamental change in the institutions of state and society. As was true at the outset of this interest in modernization in the 1960s, by the end of the 1990s convergence in economics, politics, and culture has spawned an interdisciplinary literature on globalization. In both contexts, one can identify a continuing subset of concerns with political modernization. But downsizing the state, shifting the focus of attention in modernization initiatives downward and outward, beyond the confines of historic concepts of the national state, has taken on a whole new meaning and set of challenges, as can be seen in the works of Castells and Appadurai. In this context, structural adjustment economic policies have placed enormous pressure on existing national states to reduce public employment while raising the level of expertise required of civil servants, in order to deal with increasingly complex issues in regulatory policy. In this new global setting, political modernization today, likewise, has made it possible for small states, hitherto highly vulnerable politically, to become competitive internationally in those instance where high concentrations of capital, skilled human resources, and innovative leadership can be amassed. Alternatively, one can also identify an even larger number of economic and political restructuring initiatives linked to decentralization and regionalization in larger national states when new partnerships are forged among state and local governments, postsecondary educational institutions, and firms linked to the new technologies.

See also: Democracy; Democratic Transitions; Development: Social; Modernity; Modernity: Anthropological Aspects; Modernity: History of the Concept; Modernization and Modernity in History; Third World.

Bibliography

Almond G A 1970 *Political Development: Essays in Heuristic Theory*. Little Brown, Boston

Almond G A, Coleman J S 1960 *The Politics of the Developing Areas*. Princeton University Press, Princeton, NJ

Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN

Apter D E 1965 *The Politics of Modernization*. University of Chicago Press, Chicago

Badie B 2000 *Etat Importé—The Imported State: The Westernization of the Political Order*. Stanford University Press, Stanford, CA

Black C E 1966 *The Dynamics of Modernization*. Harper and Row, New York

Castells M 2000 *The Information Age: The Rise of the Network Society*. Blackwell Publishers, Malden, MA

Inglehart R 1997 *Modernization and Post-Modernization: Cultural, Economic and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ

Levy M J 1966 *Modernization and the Structure of Societies*. Princeton University Press, Princeton, NJ

Rustow D A 1967 *A World of Nations: Problems of Political Modernization*. Brookings Institution, Washington, DC

Weiner M, Huntington S P 1987 *Understanding Political Development: An Analytic Study*. HarperCollins, New York

L. S. Graham

Modernization, Sociological Theories of

'Modernization,' in social science, refers to the transition from a traditional, rural, agrarian society to a secular, urban, industrial society. Modernization encompasses profound economic, social, political, and cultural changes, but the process of industrialization is at its core. The modernization syndrome includes urbanization, the application of science and technology, rapidly increasing occupational specialization, rising bureaucratization, and rising educational levels. However, the motivating force behind the process is the fact that industrialization is a way to escape the poverty and helplessness of subsistence agriculture. Industrialization makes it possible to enormously increase a society's wealth.

By industrializing, a society could dispel hunger, acquire the new technology now needed to compete militarily, and attain a much longer life expectancy than was possible in pre-industrial society. Moreover, economic development actually seems conducive to human happiness (though only up to a point that today's richest societies passed decades ago). Adopting a life strategy aimed at getting rich becomes compellingly attractive from the perspective of low-income societies, once it has been demonstrated that it can be

done. Although modernization has high costs, from the perspective of most pre-industrial societies, it seemed well worth the price.

Modernization is one of just two giant leaps forward that have transformed the human condition. The agricultural revolution made it possible to support about 100 times as many people per square mile as were possible in hunting and gathering societies, making civilization—that is, settled communities with specialized farmers, artisans, merchants, soldiers and officials—possible. The industrial revolution raised productivity by another 100-fold. But while it took thousands of years for the discovery of agriculture to transform the world, the industrial revolution spread throughout the world in only two centuries.

Karl Marx was the earliest and most influential modernization theorist. His failures as a prophet are well documented, but he correctly foresaw that industrialization would transform the world. When Marx was writing *Das Kapital* in the 1860s, only a handful of societies were industrialized; in the twenty-first century, almost every society on Earth is at some stage of the industrialization process.

Modernization dismantles a traditional world in which the meaning of life is clear. The warm, personal ties of communal societies give way to an impersonal, competitive life style geared to individual achievement. It brings what Max Weber called 'the disenchantment of the world,' eliminating the supernatural forces that preindustrial cultures held responsible for what happens in the world, substituting a scientific interpretation. Industrialization vastly increases productivity, but its early phase gave rise to inhuman working conditions. Marx criticized not only the ruthless economic exploitation of early capitalism, but also the tremendous psychological costs of industrialization, which he referred to as 'alienation;' the industrial worker feels estranged from work because the task has become fragmented, undemanding, and meaningless. Unlike traditional crafts, mass production labor does not challenge one's creative faculties. The industrial worker also feels alienated from the product of work, having no control over how it is manufactured and distributed. A tremendous increase in wealth is achieved at the cost of reducing the human workers to mere 'hands.'

Emile Durkheim argued that modernization brings anomie. The decline of religion and community removes traditional restraints on the individual's aspirations, allowing them to grow without limits. Modernization introduces competition that stimulates unrealistic expectations, without providing the means for their realization. The result is an increase in suicide, crime, and mental disorder.

Weber saw the rationalization of society as an inexorable aspect of modernization; though it facilitated economic growth and public order, rationalization was forcing humanity into a painfully narrow iron cage of bureaucracy and mass production. A single-minded focus on immediate means was driving out the rationality of ultimate ends.

In knowledge societies, this trend may be beginning to reverse itself; a growing segment of the population is concluding that the price is too high. Rationality, science and technology are here to stay, but their relative priority and their authority among the publics of rich societies are declining.

1. *Coherent Syndromes of Change: Why do Things go Together?*

A central claim of modernization theory is that economic development, cultural change, and political change go together in coherent, and to some extent, predictable patterns. Once a society starts to industrialize, a variety of related changes become almost inevitable, such as urbanization and bureaucratization, and eventually, changing gender roles. These consequences are not iron laws of history, but the probability is high that they will occur, once a society has embarked upon industrialization.

Industrialization also brings 'social mobilization,' or rising levels of mass participation. As Daniel Lerner (1958) and Karl Deutsch (1961) pointed out, urbanized populations are easier to organize than a scattered peasantry and a literate public finds it easier to keep in touch with what is happening on the national scene, becoming increasingly articulate. With social mobilization, a growing proportion of the population becomes politically relevant. Representative democracy becomes possible but not inevitable: throughout most of the twentieth century, it was unclear whether fascism, communism, or democracy would triumph. All three of these systems mobilized mass political participation, but the fascist and communist versions did so in the service of one-party authoritarian systems.

Although a variety of social theorists agree that technological and economic changes are linked with predictable patterns of cultural and political change, there has been continuing debate over the causal linkages: does economic change cause cultural and political change, or does it work in the opposite direction?

Marx emphasized economic determinism, arguing that a society's technological level shapes its economic system, which in turn determines its cultural and political characteristics. Given the technological level of the windmill, a society will be based on subsistence agriculture, with a mass of impoverished peasants dominated by a landed aristocracy. The steam engine brings an industrial society in which the bourgeoisie becomes the dominant elite, exploiting and repressing an urban proletariat.

Weber, on the other hand, argued that culture was not just an epiphenomenon of the economic system, but an important causal factor that can shape econ-

omic behavior, as well as being shaped by it. Thus, the emergence of the Protestant Ethic facilitated the rise of capitalism, which contributed to both the Industrial Revolution and the Democratic Revolution.

Some of Marx's successors shifted the emphasis from economic determinism toward a greater emphasis on the impact of ideology and culture. Thus Lenin argued that by itself, the working class would never develop sufficient class-consciousness for a successful revolution; they needed to be led by an ideologically aware vanguard of professional revolutionaries.

Mao emphasized the power of revolutionary thinking even more strongly. Breaking with Marxist orthodoxy, he believed that China need not wait for the processes of urbanization and industrialization to transform her. If an ideologically committed cadre could instill sufficient enthusiasm among the Chinese masses, a communist revolution could succeed even in an agrarian society.

The classic versions of modernization theory were deterministic, with the Marxist version tending toward economic determinism, and the Weberian version sometimes tending toward cultural determinism. A more recent view is that the relationships between economics and culture and politics are mutually supportive, like the various systems of a biological organism. It is pointless to ask whether the behavior of the human body is really determined by the muscular system, the nervous system, or the respiratory system as each plays an essential role, and life ceases if any of them breaks down. Similarly, political and economic systems require a supportive cultural system—otherwise they would need to rely on naked coercion, which rarely endures for long. Conversely, a cultural system that was incompatible with its economic system would be unlikely to survive.

In the course of history, numerous patterns of social organization have been tried and discarded, while other patterns eventually became dominant. At the dawn of recorded history, hunting and gathering societies prevailed, but the invention of agriculture led to their almost total disappearance. They were displaced because agriculture has functional advantages over hunting and gathering. Similarly, within two centuries after it emerged, industrialization was adopted (at least as a goal) by virtually every society on earth.

2. *Religion and Economic Growth*

Changing values can influence economic behavior. The classic example is the rise of the Protestant Ethic, which seems to have played a crucial role in the emergence of capitalism, paving the way for the Industrial Revolution.

Virtually all agrarian societies, including medieval Christian Europe, stigmatized social mobility. The main source of wealth is land, which is in fixed supply and the only way to become rich was by seizing someone else's land—probably by killing the owner. Such violence threatened the survival of any society, and was repressed by norms that emphasized acceptance of the status into which one was born, and stigmatized the economically ambitious. At the same time, traditional societies emphasized duties of sharing and charity—which helped compensate the poor for the absence of social mobility, but further undermined the legitimacy of economic accumulation.

In these societies, social status is hereditary rather than achieved, and the culture encourages people to accept their social position in this life, emphasizing that meek acceptance and denial of worldly aspirations will be rewarded in the next life. Aspirations toward social mobility are sternly repressed. Such value systems help to maintain social solidarity and discourage economic accumulation in a variety of ways, ranging from norms of sharing, to the potlatch and similar institutions in which one attains prestige by recklessly giving away one's worldly goods.

A central element in the rise of modernity was the movement away from inherited status to impersonal, achievement-based roles. For modern economic development to take place, traditional value systems must be shattered.

In Western society, the Protestant Reformation helped break the grip of the medieval Christian worldview on a significant part of Europe. It did not do this by itself. The emergence of scientific inquiry had already begun to undermine this worldview. However, Weber's emphasis on the role of Protestantism captures an important part of reality. Prior to the Reformation, Southern Europe was economically more advanced than Northern Europe. During the three centuries after the Reformation, capitalism emerged—mainly in Protestant countries, and among the Protestant minorities in Catholic countries. Within this cultural context, economic accumulation was no longer despised. Quite the contrary, it was highly respected because it was taken as evidence of divine favor: those whom God had chosen, he made rich.

Protestant Europe manifested a subsequent economic dynamism that was extraordinary, moving it far ahead of Catholic Europe. Shifting trade patterns, declining food production in Southern Europe and other factors also contributed to this shift, but cultural factors played a major role. Throughout the first 150 years of the Industrial Revolution, industrial development took place almost entirely within the Protestant regions of Europe, and the Protestant portions of the New World. This began to change only during the second half of the twentieth century, when those regions that had been most strongly influenced by the Protestant Ethic—and had become economically secure—began to de-emphasize economic growth. They did so precisely because they had become economically secure. At the same time, an entre-

preneurial outlook had emerged in Catholic Europe and (even more strikingly) in East Asia, both of which showed higher rates of economic growth than Protestant Europe. The concept of the Protestant Ethic is outdated if understood as something that can only exist in Protestant countries. However, the more general concept that culture influences economic growth, is a crucial insight.

Modernization involves a shift from a religion-oriented worldview to a rational-legal worldview. Rationalization favors the emergence of a capitalist economy, with its rational organization of labor and its rational calculation of profit and loss. Key components of this process are secularization and bureaucratization.

2.1 Secularization

Religion loses its centrality in the life of the society, losing the legitimating power it has in nonindustrial societies. For Weber, the rise of the scientific worldview—a cognitive change—was the crucial factor that led to the decline of the sacred/pre-rational elements of religious faith. But more recently, the rise of a sense of security among the publics of advanced welfare states has also contributed to the decline of traditional religious orientations. The cognitive interpretation implies that secularization is inevitable: scientific knowledge can diffuse across national boundaries rapidly, and its spread is more or less irreversible. By contrast, the rise of a sense of security among mass publics takes place only after a society has successfully industrialized, and can be reversed by economic decline. Thus, although scientific knowledge has been permeating throughout the world for many decades, religious fanaticism continues to flourish in societies that are still in the early stages of industrialization, and fundamentalist movements continue to emerge among the less secure strata of even the most advanced industrial societies.

2.2 Bureaucratization

Secularization paved the way for another key component of modernization: the rise of 'rational' organizations, based on rules designed to move efficiently toward explicit goals and with recruitment based on impersonal goal-oriented achievement standards. A prerequisite for bureaucratization was the erosion of the belief systems supporting ascriptive traditional authority and zero-sum economies, and their replacement by achievement-oriented, rational, and scientifically-oriented belief systems that supported the authority of large, centralized bureaucratic states geared to facilitating economic growth.

Along with this went a shift of prestige and socioeconomic functions away from the key institutions of traditional society—the family and the church—to the state, and a shift in economic activity from the small family enterprise to mass production that was state-regulated or even state-owned. Globally, it was a shift of prestige and power from society to state.

During the industrializing phase of history, it seemed that social evolution meant ever-increasing subordination of the individual to a leviathan state with superhuman powers. The state would become an omnipotent and benevolent entity, replacing God in a secular world. For most of the nineteenth and twentieth centuries, the dominant trend was a shift from societal authority toward state authority, manifested in the apparently inexorable growth of the economic, political, and social role of government. Even non-Marxist thinkers such as Joseph Schumpeter reluctantly considered the triumph of socialism to be inevitable.

The socialist leviathan state did not turn out to be the wave of the future. Instead, the expansion of the bureaucratic state eventually approached a set of natural limits, and change began to move in a new direction in postindustrial society.

2.3 Individuation

Individuation with industrialization, the erosion of religious social controls opened up a broader space for individual autonomy, but this space was largely taken up by growing obligations to the state. The core function of culture in traditional society was to maintain social cohesion and stability in a steady-state economy. Norms of sharing were crucial to survival in an environment where there was no social security system and here were no unemployment benefits; in bad times, one's survival depended on how strongly the norms of sharing were inculcated. In traditional societies, people feel a strong obligation to help take care of their brothers, sisters, cousins, nieces, nephews, old friends, and neighbors. These norms may make the differences between survival and death. In modern societies, this sense of obligation has greatly eroded.

3. Modernization and Westernization

Much of the early modernization literature equated modernization with Westernization. This emphasizes superficial aspects of modernization and ignores the core process. Wearing Western clothing was not crucial but industrialization was. Though the industrial revolution originated in the West, this process is not inherently Western. The goal of industrialization has now been adopted by almost every society on earth, and non-Western societies show no sign of wishing to abandon it. It is being pursued in the

twenty-first century with more enthusiasm in the non-Western world than in the West.

Even by obvious external indicators, the tendency to equate modernization with Westernization has been obsolete since the 1980s, when Japan became the world's leading automobile producer and attained the highest GNP per capita of any major nation. There is nothing uniquely Western about technology or bureaucratic rationality. China was technologically the most advanced society in the world for most of the past 2000 years, and bureaucracy originated in China.

3.1 A Change in the Direction of Change?

Modernization is not linear. Today, some of the trends linked with the rise of industrial society have reached their limits, and change is taking a new direction in the richest societies. The prevailing direction of development has changed in such fundamental ways that it might be appropriate to describe what is going on at the beginning of the twenty-first century as 'postmodernization,' rather than 'modernization.' Modernization is still at the center of the stage in developing societies. However, in the richest 'industrial' societies, tertiary sector workers have become far more numerous than industrial workers have and in the US a majority of the labor force is employed in the 'knowledge' sector alone.

The urge to survive is common to all creatures, and normally survival is precarious, since the population of any organism tends to rise to meet the available food supply and then is held constant by starvation, disease, or predators. Until very recently, the survival of most human beings was precarious.

The economic miracles and the welfare states that emerged after World War II gave rise to a new stage of history, and ultimately laid the way for the rise of postmodern values. Fundamental changes in formative experiences have given rise to a distinct value system among a growing segment of those raised in advanced industrial societies. The postwar birth cohorts in these societies grew up under conditions profoundly unlike those that shaped previous generations. First, the postwar economic miracles produced levels of prosperity that were literally unprecedented in human history. This alone would tend to encourage a greater sense of economic security, but it interacted with a second factor, the emergence of the modern welfare state. The pie was much bigger than ever before, and it was distributed more evenly and more reliably than before. For the first time in history, a large share of the masses grew up with the feeling that survival could be taken for granted. This led to a process of intergenerational value change from giving top priority to economic and physical security, to giving top priority to self-expression and the quality of life. Traditional cultural norms that limit individual self-expression are fading.

Industrial modernization was linked with a growing faith in the power of science and rational analysis to solve virtually all problems. The Postmodern outlook has a diminishing faith in rationality and a diminishing confidence that science and technology will help solve humanity's problems. There is a growing tendency for emphasis on economic growth to become subordinate to concern for its impact on the environment. Job motivations are also changing, from maximizing income, toward increasing emphasis on the quality of the work experience.

In both traditional and early industrial society, the role of women was largely limited to childbearing and child rearing—two functions that were crucial to societal survival under conditions of high infant mortality and short life expectancy. By the time a woman had borne and raised the four or five children that were needed to replace the population, she was near the end of her life span. Sexual norms encouraged reproduction, but only within the two-parent heterosexual family. Today, with much lower infant mortality, and a much longer life span, postindustrial societies are shifting toward sexual norms that give wider latitude for individual sexual gratification and individual self-expression.

Although the scientific worldview has lost its glamour, a rising sense of security among the economically more advanced societies diminishes the need for the reassurance provided by absolute belief systems. In the uncertain world of subsistence societies, there is a strong psychological need for absolute standards and a sense that an infallible higher power will ensure that things ultimately turn out well. Peace, prosperity, and the welfare state have produced an unprecedented sense of security among the publics of advanced industrial societies, bringing declining acceptance of rigid religious norms concerning sex and reproduction, and a diminishing need for absolute rules. But this also brings a growing concern for the meaning and purpose of life. Although established religious organizations have declined in most advanced industrial societies, we are not witnessing a decline in spiritual concerns, but a redirection of them.

3.2 Limits to the Expansion of the Bureaucratic State

The rise and fall of the Soviet Union illustrates the limits of the centralized, hierarchical state. In its early decades, the USSR was impressively efficient in mobilizing masses of relatively unskilled workers and vast quantities of raw materials to build the world's largest steel mill, the world's largest hydroelectric dam, and to attain one of the fastest rates of economic growth in the world. Though Stalin starved and murdered millions of people, the economic and military achievements of the Soviet State convinced many people throughout the world that this type of society

was the irresistible wave of the future. Soviet economic growth tapered off in the 1970s and stagnated in the 1980s, as a hypertrophied bureaucracy paralyzed adaptation and innovation.

Bureaucracy is inherently deadening to innovation, and this problem became acute once the Soviet Union had moved past the stage of simply importing proven technology from the West. However, the problem also reflected a collapse of motivation and morale. Absenteeism rose to massive proportions, alcoholism became a tremendous problem, and confidence in government eroded until finally the economic and political systems collapsed. Similar manifestations of the diminishing effectiveness of hierarchical, centralized bureaucratic institutions can be seen throughout industrial society. State run economies are giving way to market forces. Old-line political parties and labor unions are in decline and hierarchical corporations are losing ground to more loosely organized and participatory types of organization.

The classic bureaucratic institutions of industrial society are inherently less effective in high technology societies than they were in the earlier stages of the industrial society. They became less acceptable to the publics of postmodern society than they were earlier because of changing values.

The mass production assembly line broke down manufacturing into simple standardized routines that were endlessly repeated. This was marvelously effective in turning out masses of relatively simple, standardized products. The workers became cogs in huge centrally-coordinated machines. In societies of scarcity, people were willing to accept these costs, for the sake of economic gains. In affluent societies, they were less willing to do so.

Postmodern values give a higher priority to self-expression than to economic effectiveness and people are becoming less willing to accept the human costs of bureaucracy and of rigid social norms. Postmodern society is characterized by the decline of hierarchical institutions and rigid social norms, and by the expansion of the realm of individual choice and mass participation.

3.3 Does Modernization Lead to Democracy?

Industrialization leads to urbanization and mass literacy, which facilitate the organization of labor unions, mass political parties, and the enfranchisement of the working class. This process transforms the masses from isolated peasants into organized citizens with the power to bargain for a more equal share of the pie. The institutions that mobilized mass political participation in the late nineteenth and early twentieth century—labor union, church, and mass political party—were hierarchical organizations in which a small number of leaders or bosses led masses of disciplined troops. These institutions were effective in bringing large numbers of newly enfranchised citizens to the polls in an era when universal compulsory education had just taken root and the average citizen had a low level of political skills. However, while these elite-directed organizations could mobilize large numbers, they produced only a relatively low level of participation, rarely going beyond mere voting.

Democracy continues to evolve. The emergence of postindustrial, or 'knowledge' society, favors democratic institutions, partly because these societies require highly educated and innovative workers, who become accustomed to thinking for themselves in daily job life. They tend to transfer this outlook to politics, undertaking more active and more demanding types of mass participation. It becomes increasingly difficult for democracies to limit mass publics to an elite-directed role, and increasingly difficult for authoritarian systems to survive as they face rising mass pressures for liberalization.

See also: Bureaucracy and Bureaucratization; Development, Economics of; Durkheim, Emile (1858–1917); Functionalism, History of; Industrialization; Industrialization, Typologies and History of; Marx, Karl (1818–89); Modernity: History of the Concept; Modernization and Modernity in History; Modernization, Political: Alternate Paths; Modernization, Political: Development of the Concept; Religion and Economic Life; Secularization; Sex Differences in Pay; Weber, Max (1864–1920)

Bibliography

Bell D 1973 *The Coming of Postindustrial Society*. Basic Books, New York

Dahrendorf R 1959 *Class and Class Conflict in Industrial Society*. Stanford University Press, Stanford

Deutch K W 1961 Social mobilization and political development. *American Political Science Review* **55**(3): 493–511

Durkheim É 1984 *The Division of Labor in Society* [1893]. The Free Press, New York

Eisenstadt S 1969 *Comparative Perspectives on Social Change*. Little, Brown, Boston

Giddens A 1981 *The Class Structure of the Advanced Societies*, 2nd edn. Hutchinson, London

Giner S, Archer M (eds.) 1978 *Contemporary Europe: Social Structures and Cultural Patterns*. Routledge & Paul, London

Inglehart R 1997 *Modernization and Postmodernization: Cultural, Economic and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ

Inkeles A, Smith D H 1974 *Becoming Modern*. Harvard University Press, Cambridge, MA

Lerner D 1958 *The Passing of Traditional Society*. Free Press, Glencoe, IL

Meadows D H, Meadows D (eds.) 1972 *The Limits to Growth*. Universe Books, New York

Marx K 1977 *Selected Writings* McLellan D (ed.) Oxford University Press, Oxford, UK

Polanyi K 1944 *The Great Transformation*. Farrar and Rinehart, New York

Schumpeter J A 1947 *Capitalism, Socialism and Democracy*. 2nd edn. Harper Brothers, New York

Spier F 1996 *The Structure of Big History*. Amsterdam University Press, Amsterdam
Weber M [1904] 1930 *The Protestant Ethic and the Spirit of Capitalism*. Parsons T. (trans.) Scribners, New York

R. Inglehart

Modularity versus Interactive Processing, Psychology of

Theories of modularity function to root the structure of mind and behavior in special purpose, neurologically hardwired, faculties of mind. One point of contention is whether a module is a mechanism that produces structure (Fodor 1983), or whether it is an innate knowledge of constraints on the structure of language (Chomsky 1980). A second issue is how and where to draw the boundary between structure that is innately specified and structure that is constructed as a part of a developmental process (Chomsky 1980, Karmiloff-Smith 1992).

The model of interaction that has framed research on modularity is a reduction of the model generally assumed by interactionists. In the initial modularity debate (Piattelli-Palmarini 1980), interactionism was decoded as a 'sensor-motor interaction.' The other part of interactionism is that the study of cognition irreducibly depends on an analysis of a social interaction (e.g., Vygotsky 1978, Mead 1934). The ongoing social interaction within a community of actors, the continuous joint sensemaking of the world, also produces structure. To what degree the function of mind depends on the structures produced by special-purpose neurologically hardwired 'modules,' on structures created as a part of a physical and developmental interaction with the world, or on those structures that are produced by means of a social interaction, is an open research question.

1. *Modularity*

Structuralism developed the idea that the structure of mind (i.e., cognitive structure) is the result of evolutionary and genetically determined biological forces and that the products of human effort reflect the biological basis of structure. An issue for structuralism is to identify where the biology of structure ends and structures created by humankind begin: To what degree is the structure of language, behavior, and society determined by the biological? To what degree is the structure a product of human endeavor? Theories of modularity tie the structural elements of language (Chomsky 1980) and thought (Fodor 1983) to special purpose, neurologically hardwired, faculties of mind. An important issue is to determine to what degree modules result from a gradual constructive and developmental interaction within specific domains of behavior (as opposed to being entirely prespecified and innately determined).

In a debate with Piaget (Piattelli-Palmarini 1980), Chomsky (1980) developed modularity as an innateness argument for a language faculty inherent in the mind. A grammar is a system that specifies the phonetic, syntactic, and semantic properties of an infinite class of potential sentences. Chomsky's position was that there was a special purpose genetically determined language faculty—a mental organ—that specifies a certain class of 'humanly accessible grammars.' The language module is comparable to other organs in the body, e.g., the heart or the liver. It has both general properties that exist across the species (as is the case for any other organ, e.g., the heart) yet has unique characteristics within the individual. It has a place in a system of (mental) structures, and its development within the course of the individual's lifetime has a basis that is genetically determined.

Chomsky's chief argument for the existence of such a language module concerned the learnibility of language. 'Within a given speech-community, children with varying experience acquire comparable grammars, vastly underdetermined by the available evidence' (Chomsky 1980, p. 35). The set of initial assumptions concerning the nature of language—i.e., the grammar—the child brings to language learning is the genetically determined knowledge of the language module. The structure that is attributed to the language faculty 'must be sufficiently rich and specific to account for the attainment of linguistic competence in particular languages, but it must not be so rich and specific that it excludes attested cases' (p. 49).

Fodor (1983) developed the idea of modularity in terms of input modules. An input module produces structure rather than having innate knowledge about the structure of (i.e., grammar for) language. Input modules convert transduced sensory (visual or acoustic) input to a logical form over which the central processors work. 'Whereas transducer outputs are most naturally interpreted as specifying the distribution of stimulations at the 'surfaces' (as it were) of the organism, the input systems deliver representations that are most naturally interpreted as characterizing the arrangement of *things in the world*' (p. 42). An input module is self-contained, informationally encapsulated, and computationally autonomous. It is a special purpose faculty of mind that is neurologically hardwired, genetically determined, and evolutionarily developed. Its operation is mandatory, fast, and domain specific, and it produces 'shallow' outputs in a logical form.

Karmiloff-Smith (1992) combines Piaget's constructivism with Fodor-style modularity of the mind. One key idea is a distinction between the prespecified modules and *modularization* as a process. Karmiloff-

Smith contends that the mind may become modularized as development proceeds; moreover, cognitive development can be domain specific without being strictly modular. The initial hardwiring of an infant is not a genetic blueprint of input module, but rather a domain-specific predisposition, an initial bias, with regard to certain kinds of input. 'The view that I adopt throughout the book is that Nature specifies initial biases or predispositions that channel attention to relevant environmental inputs, which in turn affect subsequent brain development' (p. 5). The language of thought (its structure) results from a process of representational redescription (RR). When first acquired, knowledge is constrained by initial biases and predispositions, and it is implicitly represented. By a process of developmental redescription, that knowledge is gradually transformed into representations that are 'progressively more manipulable and flexible' and ultimately result in the 'emergence of conscious access to knowledge' and explanatory behavior.

2. *Contrasting Modularity and Interactionism*

A significant part of interactionist accounts of cognition concerns the sensor-motor interactions of the individual with her environment; see, for example, Brooks (1991). However, the other part of interactionism is that the study of cognition irreducibly depends on an analysis of a social interaction. A crucial idea is that structure is created as a product of a social interaction, and is modified and accumulates within a community of actors across generations. Thus special purpose, hardwired, evolutionarily designed, input modules can be seen as producing one set of structures and social interaction and cultural history another set of structures (Cole 1996, p. 198–214).

Table 1 shows three contrasting points between the assumptions underlying modularity and interactionism. Modularity accounts assume, as a basic unit of analysis, a reduction of mind to what goes on in the head; for example, language can be analyzed independent of communication. Interactionist accounts assume interaction, especially social interaction, as the basic unit of analysis. 'Mind arises through communication by a conversation of gestures in a social process or context of experience—not communication through mind' (Mead 1934, p. 50). The sources (roots) of intelligence do not only move upwards, from the biological, but also downwards, from social to cognitive (Vygotsky 1978). The study of cognition must be anchored into larger units of analysis than the individual; units that include the sociotechnical (Hutchins 1995a). Cognition is inextricably tied to activity (Suchman 1987); it emerges from an interaction (Agre 1997). It includes the context, and not just its internalization (Lave 1988).

Table 1
Contrasting Modularity and Interactionism

Modularity	Interactionism
Inside-the-head	Social Interaction
Structure as biologically determined	External Structure
Physical Evolution	Cultural History

A second point of contrast concerns the difference between structure as biologically determined and external structures that emerge as a product of human activity. The focus of interactionist accounts is best exemplified by the analyses of the structure of behavior. For an interactionist account, much of the structure of behavior emerges from an interaction, e.g., Garfinkel (1967), Suchman (1987), and Agre (1997). Context, knowledge, and 'seeing' depend on interaction with structures in the world (Lave 1988, Hutchins 1995a, 1995b, Goodwin and Goodwin 1996).

The final contrasting point concerns the historical aspects of cognition. Modularity proponents ground cognition in genetics and physical evolution, and interactionists focus on the aspects of cognition that are the product of a social, cultural, and historic interaction within a community of actors. The full ambit of activities is culturally and historically conditioned, ranging from simple activities like greeting a friend on the telephone to complex cognitive performances like reading a newspaper to physical activities like skiing. In each of these cases, cognitive performance is tied to the performance of this activity by prior generations of individuals within the community. Acquisition of the structure of a given activity depends on the individual internalizing a structure for the activity from somebody else who already knows how to perform the activity (Vygotsky 1978). Examples of models of cultural learning that depend on interaction include Hutchins (1995a), Lave and Wenger (1991), Alterman et al. (1998), and Tomasello et al. (1993).

These differences in assumption lead to some critical differences in viewpoint. Consider the case of language learning, which is critical to Chomsky's (1980) argument for a language module. Chomsky uniformly dismisses the 'performance' aspects of language learning, arguing that children acquire the grammar for a language even though it is 'vastly undetermined by the available evidence.' For interactionists, semantics, reference, and conversation are all jointly engineered among participants in a social interaction (see, e.g., Clark 1996, Hanks 1996, Clark and Wilkes-Gibbs 1986, Sacks et al. 1974). Whether language and its structure can be reduced to an analysis independent of semantics and the social interaction in which language (and language learning) occurs is a point of debate.

With regard to Fodor's version of modularity, the issue pivots over the difference between a 'language of thought' and the structure of thinking. In Fodor's view, central processes work from internal representations. An input module produces a logical form, which

is the language of thought. Contrast that to Vygotsky's view that thought, memory, and planning all begin with a social interaction. The structure of thinking depends on the history of such an activity within a community of actors. Consider the example of literacy. The reasoning processes of literacy are dependent on internalization of the structure of the 'reading' activity. To become literate is to tap into the culture of the activity of 'reading' within a given community (Scribner and Cole 1978). How this practice emerges may differ between communities. Additionally, many interactionists argue against the notion of internal representations of the sort a 'language of thought' argument supports; see a special issue of *Cognitive Science* (1993) for a debate on this issue.

The Karmiloff-Smith (1992) model has interactionist elements, but it does not account for the social elements of the representational structures acquired during the child's development. Consider a task like learning to play the piano. It is true that part of 'learning to play the piano' depends on sensor-motor interactions. But it is also true that it depends on the social interaction with a teacher and the representations provided by the musical score, which is the product of a social interaction. In other words, the child's participation in an ongoing social interaction within a community of actors is a significant factor in the development of high-level representations of the sort that allow for conscious access to knowledge and explanation. Even the more flexible access to knowledge that would be required, say, by jazz improvisation on the piano, is dependent on structures created by a social interaction; e.g., compare bebop to modal forms of jazz improvisation.

See also: Discrete State Models of Information Processing; Information Processing Architectures: Fundamental Issues.

Bibliography

Agre P 1997 *Computation and Human Experience*. Cambridge University Press, Cambridge, UK

Alterman R, Zito-Wolf R, Carpenter T 1998 Pragmatic action. *Cognitive Science* **22**(1): 53–105

Brooks R 1991 Intelligence without representation. *Artificial Intelligence* **47**: 139–59

Chomsky N 1980 On cognitive structures and their development. In: Piattelli-Palmarini M (ed.) *Learning and Language*. Harvard University Press, Cambridge, MA, pp. 35–52

Clark H H 1996 *Using Language*. Cambridge University Press, New York

Clark H, Wilkes-Gibbs D 1986 Referring as a collaborative process. *Cognition* **22**: 1–39

Cognitive Science 1993 Situated action. *Cognitive Science* **17**: 87–116

Cole M 1996 *Cultural Psychology*. Harvard University Press, Cambridge, MA

Fodor J 1983 *The Modularity of Mind*. MIT Press, Cambridge, MA

Garfinkel H 1967 *Studies in Ethnomethodology*. Polity, Cambridge, UK

Goodwin C, Goodwin M H 1996 Seeing is a situated activity: Formulating planes. In: Engestrom Y, Middleton D (eds.) *Cognition and Communication at Work*. Cambridge University Press, Cambridge, UK, pp. 61–95

Hanks W F 1996 *Language and Communicative Practices*. Westview Press, Boulder, CO

Hutchins E 1995a *Cognition in the Wild*. MIT Press, Cambridge, MA

Hutchins E 1995b How a cockpit remembers its speed. *Cognitive Science* **19**: 265–88

Karmiloff-Smith 1992 *Beyond Modularity*. MIT Press, Cambridge, MA

Lave J 1988 *Cognition and Practice*. Cambridge University Press, Cambridge, UK

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK

Mead G H 1934 *Mind, Self & Society*. University of Chicago Press, Chicago

Piattelli-Pamarini M (ed.) 1980 *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky*. Harvard University Press, Cambridge, MA

Sacks H, Schegloff E, Jefferson G 1974 A simplest systematics for the organization of turn-taking in conversation. *Language* **50**: 696–735

Scribner S, Cole M 1978 *The Psychology of Literacy*. Harvard University Press, Cambridge, MA

Suchman L A 1987 *Plans and Situated Actions*. Cambridge University Press, Cambridge, UK

Tomasello M, Kruger A C, Ratner H R 1993 Cultural learning. *Behavioral and Brain Sciences* **16**: 495–552

Vygotsky L S 1978 *Mind in Society*. Harvard University Press, Cambridge, MA

Vygotsky L S 1986 *Thought and Language*. In: Kozulin A (ed.) MIT Press, Cambridge, MA

R. E. Alterman

Monasticism: West and East

1. Introduction

'Monasticism,' derived from the Greek word *monos* meaning 'alone,' is defined inclusively here to incorporate all forms of vocational commitment to religious life. It may embrace the reclusiveness of the solitary desert anchorite, the coenobitical 'flight' of the self-governing rural ashram, monastery, or convent, or the inner-city apostolates of households of professional catechists and ordained welfare workers. Throughout the histories of the major world faiths, countless religious devotees have 'set apart' their lives 'in' but not 'of' the world within religious 'quasisects' adhering to particular theological traditions of 'training the soul.' From the Essenes of ancient Judaism, through the myriad Christian religious orders and religious congregations, to Sufi and Sanusi Islamic brotherhoods, or Buddhist and Hindu monks, mon-

asticism provides contexts in which to study the religious mediation of the social bond and explore the variety of forms in which the quest for spiritual perfection impacts upon lifestyle, social organization, culture, and society.

2. The Religious Life

For Weber, the life of the monk, the religious virtuoso *par excellence*, is in direct contrast to that of the warrior, albeit both vocational forms of life exhibit a primitive communism valorizing a segregationist ascetic ethic of celibacy, poverty, and fraternal solidarity. Warrior-monks—Hospitallers, Templars, Teutonic Knights, Samurai—have existed within the great religious traditions as attempts to integrate the call(ings) to arms and alms. However, the military life consecrated to violence and the religious life dedicated to the peaceful pursuit of an interior spiritual state and the prefiguration of the kingdom of God usually have existed in mutual tension as separate 'ways' (see *Religion: Peace, War, and Violence*). Like militarism, monasticism typically cultivates a compliance culture characterized by pervasive regulatory structures for the discipline of the body together with elaborate command hierarchies demanding uncritical obedience to legitimate superiors ('*alter Christi*,' in the case of Christian religious orders) as a mark of the fidelity and successful socialization of conscripts and volunteers.

Detached from the responsibilities of marriage and family, monastic life provides opportunity for both personal sanctification and an exemplary 'life for others.' Be it through life-long observance of 'the nine canonical hours' or the daily round with the begging bowl of the temporary Buddhist monk, 'the ascetic death-in-life' (Weber 1963), serves to 'mortify the flesh' and 'raise the mind, heart and soul to God.' This quotidian self-discipline develops a unified life of work and prayer where the contemplation of cloister, cell, or convent integrates with the labors of field, study, or hospital ward. Weber (1963) characterized occidental monasticism as 'the disciplined army of a rational bureaucracy of office' and deemed it a mode of 'asceticism directed toward the outer world.' By contrast, 'oriental monasticism' was considered to be 'of magical character,' dominated by the inner-directed mysticism of fakirs, sages, and mystagogues (see *Religious Specialists*).

Becoming a 'religious,' be that a scholar monk-priest, enclaustrated nun, nursing sister, or liqueur-making brother, entails a lengthy process of formation through various graded orders of membership as entrants pass, usually via a formal investiture ceremony ('the taking of the habit'), from novice through to 'full profession.' Under the watchful eye of experienced 'directors of souls,' entrants to 'the vowed life' are absorbed within a distinctive tradition of discipleship having its own idiosyncratic mores, prayer routines, division of labor, and 'heroic' exemplars such as saints and martyrs. Thus, each separate religious 'clan' transmits its own spiritual 'disciplines' and cultural inheritance to incoming generations of recruits whose individual members 'become' religious of a given stamp: Carthusian, Carmelite, Community of the Resurrection, Shinto, Sufi, Tibetan Buddhist, and so on. Distinct 'family resemblances' unite the members of the same monastic dynasties across time and space. This religious cultural capital manifests not only as material symbolic externals such as the cut and color of the diverse religious costumes, but also in the characteristic religious responses associated with the various spiritual 'paths' and their associated forms of life. To the outsider, all monks and nuns may look alike and engage in much the same works of divine service and charity. To the insider, however, a de la Salle Teaching Brother and a Franciscan friar are as different as chalk and cheese. A howling Dervish is not a silent Cistercian Trappist.

3. Civilizational Mission

Civilizations east and west have been marked inerradicably by their monastic patrimonies that have been formative at critical renewal points in the histories of peoples, empires, and cultures. Nisbet (1973) goes so far as to opine: 'Without any question, it seems to me, the monastery, especially as it was conceived by Benedict, is the true source of most of our Western ideas of community achieved through the processes of withdrawal and renewal.' Across the faith traditions east and west, monasteries as centers of 'godliness and good learning' (Leclercq 1961) have been power-houses of liturgical, musical, and architectural creativity (see *Christian Liturgy*; *Symbolism (Religious) and Icon*); sites of practical wisdom in agriculture, animal husbandry, and horticulture (grape growing and bee-keeping being renowned monkish specialisms) (see *Religion and Economic Life*); pre-eminent throughout the religious traditions in the arts and sciences of calligraphy, printing, and librarianship; and, as 'experiments in living,' 'prophetic' centers of innovation in modes of self-governance and welfare administration. Despite a reputation for social conservatism, religious virtuosi from all the world faith traditions across all five continents have participated in diverse independence movements and human rights campaigns.

As a force in the development of the western social order, monasticism has spawned many regional 'styles' with their own distinctive roots and Rules but with considerable syncretism in evidence (Knowles 1969). In the course of time, 'Celtic' monasticism, itself deriving from the Egyptian eremetical exemplars (Antony and Pachomius) fused with the reverence for

learning of the Greek and Frankish scholar-monks, such as Basiland Cassianto develop the high Latin culture that was the main conduit of Christianity from the last days of the Roman Empire to the initial centuries of Christendom in the early Middle Ages. Missionary monks inspired by St. Patrick and St. Columbas took the Christian gospel into the heartlands of European tribal society in the sixth and seventh centuries. From the eighth to the eleventh centuries ('the Benedictine centuries'), the proliferation of daughter monasteries of Benedict's Monte Cassino (founded 529) created the network of organization and intelligence that shaped the character and ethos of medieval European civilization. The religious culture of early modern urban development in the thirteenth and fourteenth centuries was much influenced by the preaching friars (Dominicans, Franciscans, Augustinians) of the Mendicant Orders (Lawrence 1994) (see *Christianity Origins: Primitive and 'Western' History*). Throughout Asia and the Far East the monasteries of Theravada Buddhism had comparable influence (Silber 1995) (see *Buddhism*).

4. *Styles of Life*

The 'contemplative' and 'active' styles of religious life rarely are walled off entirely from each other. The 'world' knocks at the monastery gates in every era. Pontiffs and prelates have sought the support of abbeys and nunneries in the promotion of this or that ecclesial cause from prayers for the crusades to theological backing in the extirpation of heresy. Monarchs and nobles have found ways to secure preference and placement for one of their number in the governmental and suffrage systems of religious life. Paradoxically, these attractions and achievements have at times led to monasticism assuming considerable economic and political power (see *Religion: Mobilization and Power*) be it that of the great medieval religious foundations (Grande Chartreuse, Cluny, Citeaux, Monte Cassino) with their vast benefices and endowments; or the self-governing republic since 1052 of the influential monastery of Mount Athos in northern Greece; or the three centuries of state control assumed by the Dalai Lama, the chief abbot of Tibetan Buddhism; or the impact since 1837 of the Senusi Islamic brotherhood on Sudan and the eastern Sahara. There is a continuous tradition of criticism of the worldly foibles and contradictions of monastic elites down the ages, from Vigilantius (fourth century), through Chaucer, Erasmus, Luther, and Calvin, to Voltaire and modern anticlerical republicanism worldwide. Liberation theology and a spate of contemporary novels and films directed against 'new world theocracies' (the Paraguayan Reductions) and colonial 'foreign missions' sustain the iconoclastic suspicion of monastic imperialism.

Successive waves of new religious foundations have been launched by charismatic (see *Charisma and Charismatic*) religious founders animated by a utopian 'return to sources' (*ad fontes*) from which 'scriptural primitivism,' the combined attack on abuses within the Church and world, have been prepared down the centuries. The inherent tension between its this-worldly and other-worldly parameters accounts in great measure for the transformative capacities of monasticism which owes its survival and dynamism to an ongoing internal reform impulse (the Gregorian Reforms, for instance). Following the upheavals of the Protestant Reformation ('the dissolution of the monasteries') and the French Revolution, Roman Catholicism invested heavily in 'apostolic orders' (Jesuits, Daughters of Charity) and missionary teaching congregations (Sallesians, Marists) which served as 'the pope's battalions' in the Tridentine 'counter-modern' offensive. The nineteenth-century Oxford Movement saw fit to revive monasticism within Anglo-Catholicism (Hill 1973). Come the age, come the religious order. The *aggiornamento* of the Second Vatican Council (1962–5) with its emphasis on 'adaptation to the signs of the times' and 'a preferential option for the poor' (Sweeney 1994) is the most recent expression of the sporadic self-reform tendency in religious life to check the mixed fruits of sanctity's success, social, and theological drift.

Each new age of monastic life and religious congregations addresses the perceived apostolic needs and ecclesiastical interests of the epoch be that for grand monastic citadels, hostels for crusaders and pilgrims, sanctuaries from plague and persecution, hospitals, schools, orphanages, hospices, food kitchens, or way stations for state policies of care in the community. Monasticism serves as a ligature binding civilizations east and west to their pasts and a bridge across which lessons of the religious quest are passed over into civil society. Not surprisingly, militant secular regimes (see *Secularization*) have found pretext to sequester monastic assets and secularise (and worse) religious personnel. Monasticism and the religious orders provide the great world religions with multiple and mixed benefits: an outlet for 'charisms' to flourish then 'routinize'; an opportunity context for a specialized religious 'division of labor'; social space to cultivate a realm of intermediate associational life—collegia—within the complex organic ecclesiastical commonwealth (the *communitas communitatum*) without need for secession from the Mother Church.

5. *Sign of Contradiction*

If, on the one side, the religious institutes represent *gemeinschaft* in its most comprehensive guise (anticipatory democracy, brotherhood of equals, proto-feminist sisterhood), on the other side, monastic life

can exemplify the 'total institution' as escapist flight into infantile dependency and social parasitism. As a 'sign of contradiction,' monasticism and religious life west and east, appear destined to live out these two contrary images and appeals. As many established religious orders enter the 'terminal' stage in the monastic life-cycle (Hostie 1972) marked by membership hemorrhage, aging and low, erratic recruitment (Ebaugh 1993), the third millennium is likely to see novel forms of monasticism—mixed gender lay communities; gay and lesbian institutes; multifaith congregations—bearing witness 'against all reason' and the 'possessive individualism' of secular modern consumer society to the resilience and revitalization (Cada et al. 1979) of 'the consecrated life.'

See also: Asceticism; Buddhism; Catholicism; India, Religions of; Islam: Middle East; Orthodoxy; Pilgrimage; Religious Organizations; Religious Specialists

Bibliography

Cada L, Fitz R, Foley G, Giardino T, Lichtenberg C 1979 *Shaping the Coming Age of Religious Life*. Seabury Press, New York

Ebaugh H R F 1993 *Women in the Vanishing Cloister: Organizational Decline in Catholic Religious Orders in the United States*. Rutgers University Press, New Brunswick, NJ

Hill M 1973 *The Religious Order: A Study of Virtuosi. Religious Life and Its Legitimation in the Nineteenth Century Church of England*. Heinemann, London

Hostie R 1972 *Vie et mort des ordres religieux*. Desclee de Brouwer, Paris

Knowles D 1969 *Christian Monasticism*. Weidenfeld & Nicolson, London

Lawrence C H 1994 *The Friars. The Impact of the Early Mendicant Movement on Western Society*. Longman, London and New York

Leclercq J 1961 *The Love of Learning and the Desire for God: A Study of Monastic Culture*. (Trans. Misrahi C). Fordham University Press, New York

Nisbet R 1973 *The Social Philosophers. Community and Conflict in Western Thought*. Heinemann, London

Silber I F 1995 *Virtuosity, Charisma and Social Order: A Comparative Sociological Study of Monasticism in Theravada Buddhism and Medieval Catholicism*. Cambridge University Press, Cambridge, UK

Sweeney J 1994 *The New Religious Order: A Study of the Passionists in Britain and Ireland, 1945–1990 and the Option for the Poor*. Bellew, London

Weber M (ed.) 1963 *The Sociology of Religion* [Trans. Fischoff E. Intro. Parsons T]. Beacon Press, Boston

W. J. F. Keenan

Monetary Policy

Monetary policy is one of the two principal means (the other being fiscal policy) by which government authorities in a market economy regularly influence the pace and direction of overall economic activity, importantly including not only the level of aggregate output and employment but also the general rate at which prices rise or fall. Indeed, the predominant trend since the mid-twentieth century has been to place increasing emphasis on monetary policy (and correspondingly less on fiscal policy) for these purposes. Governments carry out monetary policy, typically via specialized agencies called central banks, by exploiting their control over the supply of certain kinds of claims against the central bank—hence the label 'monetary'—that enable a country's businesses, banks, and individuals to carry out their day-to-day economic affairs. In most financial systems, banks in particular are legally required to hold claims against the central bank in order to create deposits and make loans, and so the central bank's control over the supply of claims against itself also gives it a form of control over the economy's money and credit in a far broader sense. The evidence from experience, in one country after another, makes clear that the exercise of this control—'monetary policy'—powerfully affects a country's economy, for either good or ill.

1. How Monetary Policy Arises

Providing money for use in everyday transactions has been a commonplace function of governments for well over 2,000 years. Today almost all paper currency is 'fiat money'—in other words, it has value only because the government mandates that within the country's borders it must be accepted as payment in any and all transactions. (Importantly, the government does not mandate, because in a market economy it cannot, the price level at which its currency is accepted.)

In principle, the existence of fiat currency alone would be sufficient to enable a government to carry out a primitive form of monetary policy, distributing paper money to the public, through some device or other, in either small volume or large as it saw fit. In fact, modern central banks conduct monetary policy differently: by exchanging claims against themselves for other claims against the government—typically interest-bearing bills, notes, and bonds—in markets in which these instruments are freely traded. For this exchange to be possible, however, there must be other claims against the government outstanding in the first place.

Hence, monetary policy, as carried out in practice, is made possible by the existence of fiscal policy, in the usual sense of overall government spending and taxing and the government's need to finance any excess of expenditures over revenues by means of borrowing. It is only because on balance over time most governments have spent more than they have taken in—that is, have run a fiscal deficit—that they have a stock of debt obligations outstanding.

2. How Monetary Policy Works

The chief objectives of monetary policy in modern times typically have been to maintain stability of a country's general price level—that is, to prevent either inflation or deflation—and to promote maximum levels of output and employment. Other often accepted goals of monetary policy include maintaining balance in a country's international trade, preserving stability in its financial markets, and fostering increased capital investment so as to enhance its economic growth over time. With the exception of preserving financial market stability, which is usually taken to be secondary, all of these objectives pertain to aspects of an economy's nonfinancial economic activity. By contrast, central banks' monetary policy operations take place exclusively in the financial markets. For monetary policy to be effective therefore requires some process, often called the monetary policy 'transmission mechanism,' by which the purely financial actions taken by the central bank influence the nonfinancial decisions of households and firms.

2.1 Demand for Central Bank Liabilities

The key to how this mechanism works is that, in part for reasons of convenience and in part as a matter of law, participants in a country's nonfinancial economy need to hold claims against its central bank—the outstanding total of which is usually called the country's 'monetary base.'

At the simplest level, currency is usually a claim on a country's central bank. Individuals who buy everyday items for cash, and the businesses with which they deal, therefore need to hold and exchange central bank liabilities. In the past, sudden large increases in the public's demand for currency, when not met by an increase in currency supplied by the central bank, often triggered financial crises and consequent economic downturns. In modern times, however, most central banks have passively supplied whatever changing volume of currency the public seeks. As a result, demand for currency typically plays no significant role in the monetary policy process.

Instead, what matters are the claims that the country's private-sector banks hold in the form of deposits, usually called 'reserves,' at the central bank. For several reasons, the banks' need for these reserves expands or contracts roughly in pace with the overall level of activity taking place in the nonfinancial economy.

First, in many countries banks must, by law, hold such reserves in proportion to the volume of deposits (or if not their total deposits, then some forms of deposits) that they have outstanding. Hence, the ability of the banking system to create deposits for businesses and households to use in executing transactions that they do not make in cash depends on the quantity of reserves that they can obtain.

Second, a bank's ability to make loans depends on its ability to create deposits. A loan is an asset to the bank, while a deposit is the bank's liability. Apart from changes in the bank's capital due to retained profits or new securities issues, its total assets and total liabilities must expand or contract together. The requirement that banks hold reserves in proportion to their deposits, therefore, also means that they must hold reserves in order to advance credit to businesses or households.

Third, although some countries do not impose legal reserve requirements, their banks hold balances at the central bank to use for settling the claims among themselves that arise whenever an account holder at one bank deposits a check drawn on another bank. Even in some countries that do have reserve requirements, a large part of the balances that banks hold at the central bank are likewise held primarily for purposes of settling interbank claims. The standard way of effecting such transfers is to shift reserves, on the books of the central bank, from one bank to the other.

2.2 The Central Bank as Monopolist

The central bank's power to conduct monetary policy stems from its role as the sole source of reserves (or, again, monetary base) to meet this demand. In short, the central bank is a monopolist over the supply of its own liabilities.

2.2.1 Open market operations. The most common procedure by which central banks either increase or reduce the outstanding supply of bank reserves is through 'open market operations'—that is, buying or selling securities (normally the debt obligations of the central bank's own government) in the free market. When a central bank buys securities, it makes payment by increasing the reserve account of the seller's bank. Doing so increases the total volume of reserves that the banking system collectively holds. No bank, nor any other market participant, can add to or take away from the total volume of reserves that all banks together hold.

Expanding or shrinking the total volume of reserves in this way matters because banks can trade reserves among one another in exchange for other assets. Because the central bank pays only a low rate of interest (often zero) on these balances, any bank that has more reserves than it needs typically will try to exchange them for some interest-bearing asset. If the banking system as a whole has excess reserves, however, more banks will seek to buy such instruments than to sell them, thereby bidding up these instruments' price and thus reducing the interest rate on them. The resulting lower return on interest-bearing assets means a reduced opportunity cost of holding

zero-interest reserves. Only when market interest rates fall to the level at which banks collectively are willing to hold all of the reserves that the central bank has supplied will the financial system reach an equilibrium.

Hence, an 'expansionary' open market operation (in which the central bank expands the supply of reserves) creates downward pressure on short-term interest rates not only because the central bank is itself a buyer in the securities market but, more importantly in quantitative terms, because it leads banks to become buyers of securities as well. Conversely, a 'contractionary' open market operation, in which the central bank sells securities, puts upward pressure on short-term interest rates. Under ordinary conditions, a well-functioning central bank following these procedures can pick some short-term interest rate (in many countries the central bank focuses on the overnight interbank lending rate) and establish that rate at whatever level it chooses.

2.2.2 Reserve requirements. In a banking system that imposes reserve requirements, an alternative way for the central bank to achieve the same objective would be to adjust the stated percentage indicating how much in reserves banks are required to hold in relation to their outstanding deposits. Lowering the reserve requirement, and therefore reducing the demand for reserves, has roughly the same effect as an expansionary open market operation, which increases the supply of reserves: either action creates downward pressure on interest rates. Although in principle central banks could carry out monetary policy by either means, in practice most rely primarily on open market operations, using changes in reserve requirements only to achieve more technical objectives concerning the composition of banks' liabilities.

2.2.3 Central bank lending. Another way in which central banks can change the supply of reserves is by lending reserves directly to some bank. In most countries that have reserve requirements, such direct lending of reserves is normally small in scale, and it plays only a minor role in the monetary policy process. Especially in countries that impose no reserve requirements, however, lending directly to banks is an important part of how the central bank supplies reserves.

Because no bank will pay more to borrow elsewhere than the rate at which it can freely borrow from the central bank—and, similarly, will not sell a higher-yielding asset rather than borrow at a lower rate in order to continue holding it—setting the interest rate at which the central bank lends reserves to the banks effectively establishes a ceiling for (risk-free) short-term market rates. In some countries, the central bank confines the market rates on short-term instruments to trade within a narrow range by imposing both a ceiling and a floor: setting an interest rate at which it will lend to banks as well as a (slightly lower) interest rate that it will pay banks on their holdings of reserve balances.

2.3 *Effects on the Nonfinancial Economy*

For monetary policy to achieve its objectives with respect to the nonfinancial economy, there must be some causal process by which the changes that the central bank brings about in short-term interest rates in turn affect real output and employment, or prices and wages, or both. Economic theory, as well as empirical observation, suggests a variety of avenues by which such influences can operate.

2.3.1 Influences on the demand for goods and services. In an economy with well-developed credit markets, households and firms frequently borrow to finance their spending. Fluctuations in interest rates naturally affect the willingness to undertake such expenditures. Although the central bank directly controls only the interest rates on short-term instruments like Treasury bills, the longer-term interest rates applicable to borrowing for these purposes mostly move in the same direction as short-term rates because banks and other investors are able to substitute among different debt instruments in their asset portfolios. Moreover, because banks' ability to create credit depends on having reserves, by decreasing the supply of reserves the central bank can induce banks in particular to cut back on lending in ways that go beyond merely charging a higher interest rate, including rationing credit among would-be borrowers.

The same process of portfolio substitution also leads to downward pressure on the market prices of other assets, most prominently equity securities and real estate. Because these price-sensitive assets bulk large in the total holdings of households (and sometimes firms too) in many countries, changes in wealth also normally influence the demand for goods and services. Yet a further extension of the same process enables monetary policy to influence the value of a country's currency in international markets, which in turn affects demand abroad for the country's exports as well as demand at home for foreign imports.

2.3.2 Demand effects on prices, employment, and wages. If a central bank's monetary policy stimulates

the demand for its economy's goods and services, and there is no immediate matching change in supply, the result will be to create upward pressure on prices. Higher prices in turn lead businesses to seek to produce more. In industries that are labor intensive, increasing production means hiring more workers. Hence, monetary policy also influences employment. In order to attract more workers, however, firms normally have to pay higher wages. Hence, an increase in output and employment due to expansionary monetary policy usually means upward pressure on both prices and wages—in other words, inflation.

2.3.3 Price misperception effects on aggregate supply. Under some economic theories, this tendency for output to increase and prices to rise following expansionary monetary policy has a different interpretation: what matters is that the increased prices due to greater aggregate demand come as a surprise. If a firm mistakenly interprets a rise in the general price level for a rise only in the price of its own specific product, it will, as before, choose to expand production. The immediate effect is the same, but once the firms that react in this way realize their mistake, their incentive to maintain higher production disappears. Under this theory, therefore, monetary policy can still influence both output and prices, but the effect on prices is lasting while that on output is only temporary.

2.3.4 Effects on prices and wages only. Yet a further extension of this same line of theorizing suggests that if people understood that the effects of monetary policy on output and employment are merely temporary—and, importantly, if they also understood what the central bank is doing as it is doing it—there would be no real consequences of monetary policy at all but only the influence on prices and wages. Although constructs that deliver this conclusion appear frequently in theoretical work, evidence that monetary policy does have real effects is sufficiently widespread that current-day interest in this line of thought focuses less on whether monetary policy is 'neutral' in this sense than on the theoretical rationale for understanding why it is not.

3. *The Design of Monetary Policy*

Monetary policy is problematic not only because the central bank's goals are sometimes in conflict but also because the mechanisms by which central bank actions affect the nonfinancial economy mostly involve lags often measured in years rather than weeks or months. Many important aspects of the economic circumstances in which the central bank's actions will be having their effect are, therefore, not just unknown but unknowable when the decisions governing these actions are taken. Much of the study of monetary policy since World War II has focused on how to conduct monetary policy under these handicaps.

3.1 *The Monetary Policy Instrument*

Because of its monopoly position, the central bank can fix the quantity of reserves (or, including currency, the monetary base). Doing so affects the market equilibrium of interest rates on all debt instruments. Alternatively, the central bank can set the interest rate on any one class of debt instrument by continuously supplying whatever amount of reserves is consistent with market equilibrium at the chosen interest rate level, allowing market forces to determine what quantity of reserves it supplies. The market equilibrium also determines the interest rates on all other debt instruments. (A variant of this latter strategy is for the central bank to set the foreign exchange rate of its currency.) By contrast, the central bank cannot directly set the volume of banks' deposits or lending. Nor can the central bank directly determine real output, or prices, or any other aspect of nonfinancial economic activity.

If all other influences bearing on whatever constitutes the ultimate objective of monetary policy were known in advance, it would make no difference whether the central bank conducted policy by fixing the supply of reserves or by setting an interest rate (or the exchange rate). These operating strategies would be equivalent. Because many forces bearing on the central bank's objectives are unpredictable, however, the choice of 'instrument' by which to implement policy matters for the effectiveness of policy. In general, the more uncertainty surrounds the behavior of households and firms in the markets for goods and services—for example, the strength of consumer spending, or of business investment—the more advantage there is to fixing the quantity of reserves. By contrast, the more uncertainty surrounds behavior in the financial markets—households' and firms' demands to hold deposits vs. other assets, their desire to borrow, the willingness of banks to lend, and so on—the more advantageous it is to set the price of reserves (in other words, an interest rate).

Actual practice in this regard has varied over time and across countries. Especially in the 1970s and 1980s, many central banks experimented with strategies based on fixing the quantity of reserves (or of the monetary base), or its growth rate. As of the beginning of the twenty-first century, however, most central banks of large countries operate by setting an interest rate.

3.2 Intermediate Targets

The central bank cannot directly set the amount of either money, in the usual sense that includes both currency and bank deposits, or bank credit. Much empirical work over many years, however, has gone into documenting the relationships between money and either income or prices—especially between money and future income or prices—and there has been some research on comparable relationships for credit. Moreover, data reporting in most countries is such that money and credit are observed before output and prices. To the extent that money, for example, displays a reliable relationship with the aspects of nonfinancial economic activity that the central bank ultimately seeks to influence, conducting monetary policy as if the quantity of money itself, or its growth rate, were the objective of policy—that is, using money as an 'intermediate target'—in effect enables policy to respond to unwanted movements in output or prices before they occur (or at least before they are directly observed). Especially in the 1970s and 1980s, many central banks adopted intermediate targets of this kind.

In time, however, this way of conducting monetary policy fell out of favor for several reasons. First, research showed that using money as an intermediate target amounts to a way of exploiting the information contained in observed movements of money, but that it fails to make use of other information provided by the many different financial and economic variables that are also observable on a timely basis. Further, except under highly special circumstances, using money as an intermediate target fails to use efficiently even the information provided by observations of money itself.

Second, there was never a sound theoretical basis for knowing which measure of money was the right one to target, and even within any one country empirical evidence on which measure had the closest relationship to income and prices was often mixed. Presumptions that different measures of money would show roughly similar growth rates over relevant time horizons proved mistaken.

Third, and most importantly, by the 1990s the empirical relationships between money (or credit) and either income and prices had broken down in many countries. Standard explanations for this breakdown included financial innovations that enabled money holders to switch more easily among different kinds of deposits, or between deposits and other assets; new patterns of wealth holding in the wake of the high inflation of the 1970s and early 1980s; and the advancing integration of world financial markets. But for whatever reason, the previously observed predictive power associated with money disappeared, and without it there was little rationale left for using any measure of money as an intermediate target for monetary policy.

3.3 Rules vs. Discretion

One reason the use of money as an intermediate target had been so attractive was that it supposedly provided an 'anchor' to a country's price level—or, if the central bank was targeting money growth, to the inflation rate. (Once the empirical relationship between money and prices broke down, this supposition bore much less force.) A second reason was that money could potentially serve as the basis for very simple rules for conducting monetary policy. For example, always seeking to have money grow at a constant percentage rate (which some economists have advocated) may be suboptimal, but it is not inherently flawed on theoretical grounds. By contrast, always setting some interest rate at a constant value is inherently unstable.

The long-standing preference among some economists for rules to govern monetary policy reflects, in the first instance, matters less of economics than of political economy. But in the 1970s and 1980s a new line of theory suggested a more specifically economic ground for favoring rules for monetary policy: that the sequential, discretionary decisions made even by fully knowledgeable and well-intentioned central banks were 'dynamically inconsistent' in a way that accounted for the chronic high inflation that by then had appeared throughout the industrialized world. Central banks understandably seek to boost output and employment above the level that market forces would otherwise establish, and under some theories a way to do so is to generate inflation that comes as a surprise to producing firms. But once the central bank has already generated any particular rate of inflation, which is then expected to continue, a yet higher rate is necessary to create the needed surprise. This process goes on until the inflation rate is so high that the cost of any further increases in inflation per se outweighs the consequent gain in output. At that point output reverts to its normal level, but inflation remains high.

Because the dynamic inconsistency argument placed discretionary decision making at the center of the industrial countries' inflation problem, this line of theory pointed to a monetary policy rule as the most straightforward solution. Judged from this perspective, the disinflation that followed was, in most countries, strong counter-evidence. By the early 1990s (earlier still in many countries), inflation had fallen sharply in most industrialized economies. In many this disinflation occurred without any change in central bank procedures, and certainly without the adoption of a monetary policy rule. Even in countries that did change their approach to monetary policy—for example, by adopting an explicit inflation target—the change typically came after inflation had already subsided.

A policy rule is not the only way to overcome dynamic inconsistency, however. Further research showed that the central bank's awareness of the importance of its public reputation ('credibility') acts

in the same way. So does appointing central bank decision-makers whose aversion to inflation is more pronounced than that of the general public. Hence, it remains unclear to what extent theories based on dynamic inconsistency provide a good account of the rise and fall of postwar inflation.

3.4 Central Bank Independence

The debate over rules vs. discretion in monetary policy also focused attention on the differing status of different countries' central banks: whether the central bank has the authority to establish the objectives of monetary policy and/or to carry out policy operations as it sees fit in pursuit of objectives even if they may be determined by other parts of the government, whether the central bank's officials are subject to removal over matters of policy, and so on. Countries where the central bank was more independent, either legally or *de facto*, tended on average to have lower inflation. Some economists inferred from this tendency that the motivation to seek higher output and employment by creating surprise inflation was mostly imposed on the central bank from outside. Hence, a more independent central bank would deliver lower inflation. (At the same time, however, other research pointed out that a key corollary of this line of reasoning—the implied tendency for more independent central banks to find it less costly, in terms of foregone output and employment, to reduce inflation by any given amount—did not correspond to observed cross-country differences.)

One result of this line of research and public discussion was that in the late 1990s several countries formally granted their respective central banks greater independence. Also, the new European Central Bank, established in 1999, has very substantial independence from the governments of the participating member countries.

3.5 Inflation Targeting

By the 1990s the combination of the demise of intermediate targets for monetary policy, the renewed emphasis on interest rate setting, and the failure (by some lights) to impose rules limiting central bank discretion had left a widely perceived vacuum in the structure of most countries' monetary policymaking. The problem of providing an 'anchor' to prices remained. So did the risks associated with step-by-step discretion.

In response, a number of central banks adopted formal 'inflation targets.' Importantly, adopting an inflation target does not necessarily mean eschewing concern for real economic outcomes, nor does it imply the belief that money is neutral in the short or medium run. When a central bank seeks to maintain both low inflation and full-employment output, the relative strength of its preferences between these two objectives determines how rapidly it will seek to return to the targeted inflation rate after some departure has occurred.

One hoped-for advantage of having an explicit inflation target is to facilitate public monitoring and evaluation of the central bank's performance. The underlying presumption is that, on average over time, the central bank can achieve whatever inflation rate it seeks. Hence, failure to satisfy the stated target can be laid at the central bank's door.

4. Outstanding Research Questions

It would be wrong to give the impression that by the end of the twentieth century the combination of experience and research had settled all important questions about monetary policy. Monetary economics is not a laboratory science with the ability to conduct controlled experiments. Moreover, many questions that bear centrally on how best to conduct monetary policy hinge on how an economy's businesses, households, and banks behave in contexts where the changing economic backdrop and changing institutional arrangements matter importantly. Even if such questions appear settled at some time, therefore, the answers do not necessarily remain valid.

4.1 The Transmission Mechanism

A perennial goal of the monetary policy research agenda is to understand the behavioral process by which the central bank's actions influence the non-financial economy. In theory, there are many potential influences at work. But especially for purposes of actual policy implementation, knowing how much of the overall effect of policy is due to each is important. One particular focus of theoretical and empirical research in this area, for example, centers on the distinction between the 'money view,' according to which contractionary monetary policy works by restricting the amount of deposits that banks can create (thereby driving interest rates higher, and thus depressing demand for goods and services), and the 'credit view,' under which what matters is instead banks' reduced capacity to extend loans to firms and households seeking to finance expenditures.

Part of what makes this research program difficult is the tension between the preference for simple theories and the need to take account of the often quite involved circumstances under which banks, businesses and households interact. Extremely simple theories of

monetary policy often deliver conclusions that may be appealing intellectually but that observed experience readily contradicts—for example, that monetary policy affects prices but not real economic activity, or that inflation varies in close proportion to money growth. Theories leading to conclusions more in line with experience tend to be complicated, and the need to draw more finely focused inferences thwarts easy empirical validation.

4.2 Quantitative Measurement

For a central bank charged with carrying out a country's monetary policy, even full knowledge of the qualitative process by which policy works is not sufficient. It is also necessary to have some idea of how much to increase or reduce the supply of reserves, or how far to raise or lower interest rates, in order to achieve any given objective at any particular time. Such knowledge can be no more than a statistical estimate, subject to uncertainty. But having such estimates, and also an informed sense of the associated uncertainty, is crucial.

In recent decades three lines of empirical research have represented attempts to gain such knowledge. Structural economic models—models that use theory to place restrictions on the admissible representation of the behavior under study—allow the use of observed data to measure specific parts of the monetary policy process, and putting together enough such parts produces a quantitative representation of the monetary policy process as a whole. Here theory matters: the value of the resulting estimates, for either the parts or the whole, depends on the validity of the *a priori* restrictions imposed. Most central banks estimate and regularly use structural models of this form, although little such work has taken place in the academic world in recent years.

Especially since the 1970s, vector autoregressions (sets of equations that intensively exploit the raw covariation over time of a small number of economic variables) have been the main vehicle used in academic research for addressing questions about monetary policy at the aggregative level. When used for purposes like assessing the effects of monetary policy, however, even vector autoregressions are not free from restrictions based ultimately on theory—most basically, what (usually small) set of variables to include, and what 'causal ordering' to impose on them. Still richer sets of cross-equation restrictions allow 'structural vector autoregressions' to exploit the covariation in the data yet more fully. As is true for structural models, the resulting estimates depend on the validity of the restrictions imposed.

Beginning in the late 1980s, there has also been renewed interest in combining standard statistical methods with the use of nonquantitative information—drawn, for example, from close reading of minutes of central bank meetings. To date this work has been carried out for only a few countries.

4.3 Phillips Curve Issues

The trade-off between the desire for price stability and the desire for high output and employment has been a persistently important aspect of monetary policy debate, and consequently of monetary policy research, throughout the postwar era. One crucial question is what level of real activity is consistent with maintaining an unchanged inflation rate. Another is how much foregone output or employment is required to bring a high inflation rate down to a lower one (what is the 'sacrifice ratio'?). Although some theories imply that monetary policy is neutral, most experience indicates that there are real costs to disinflation. The question that much research has tried to answer from observed evidence is how much.

Neither the noninflationary level of output (or employment) nor the sacrifice ratio is a natural constant, however. An important part of this same line of research has been to establish what factors, especially including factors subject to influence by national policies, cause each to vary across countries. Possible determinants suggested by familiar theory include labor market institutions governing the difficulty of hiring or firing workers, the generosity of publicly provided unemployment benefits, the economy's openness to international trade, the degree of central bank independence, the central bank's established 'credibility,' and the extent to which monetary policy decisions are transparent to the public. Research on each of these topics has provided mixed results.

Each of these lines of inquiry is implicitly a question about the short, or at most the medium, run. Implicit in the underlying theory is that in the long run output and employment revert to levels determined by resources, technologies, preferences, and other fundamental factors that are independent of monetary policy. Even so, there remains a long-run trade-off between the variability of output and employment and the variability of inflation, which the central bank's approach to monetary policymaking does influence. Seeking to understand how this variability trade-off arises, and exploring the implications for it of different ways of conducting monetary policy, remains an important object of research.

A quite different line of research on these issues, prompted in part by the chronic high unemployment and stagnant growth in so many European countries in the last quarter of the twentieth century, has questioned the standard theory that monetary policy is neutral in the long run, so that a reduction in inflation requires only a temporary loss of output and

employment. If the economy's production process adjusts so that some part of the reduced output or employment never recovers (e.g., because some unemployed workers leave the labor force permanently, or because some productive investments are not undertaken), then the usual interpretation of the sacrifice ratio in terms of strictly temporary real costs of achieving permanent disinflation is invalid.

4.4 Costs of Inflation

Much of the motivation underlying the conduct of monetary policy is to preserve price stability, or at least a low rate of inflation. For some central banks, price stability is the only economic objective formally stated in the bank's charter. For others price stability is one among just a few objectives set forth.

By contrast, economic theory and empirical analysis have never been able to identify just why movements in an economy's overall price level are so harmful. In the simplest theories, inflation that is anticipated in advance has no real effects at all, and unanticipated inflation merely creates transfers from some groups to others. Abandoning simplicity in favor of realism introduces many avenues for inflation to result in welfare-reducing costs, but empirical estimates of the magnitude of such costs are typically small. There is also no evidence that inflation below about 10 percent per annum reduces an economy's growth rate.

Yet opinion surveys as well as election returns give evidence that the public in most countries dislikes inflation, and even accepts sizeable real costs as a sacrifice necessary to reduce inflation. One possibility is that the aspect of inflation that most distresses the public is not economic, in any narrow sense, but instead the concern that if the government is not able to carry out this aspect of its responsibilities effectively—providing a medium of exchange has been a government function almost as long as there have been governments—then perhaps the social order may be at risk in other ways as well. In any case, although understanding the costs of inflation has long been on the research agenda surrounding monetary policy, it remains an even more open question than most.

See also: Central Banks; Consumer Economics; Consumer Psychology; Economics, History of; Keynes, John Maynard (1883–1946); Price Inflation.

Bibliography

Ball L 1994 What determines the sacrifice ratio? In: Mankiw N G (ed.) *Monetary Policy*. University of Chicago Press, Chicago

Barro R J, Gordon D B 1983 Rules, discretion, and reputation in a model of monetary policy. *Journal of Monetary Economics* **12**: 101–21

Bernanke B S, Blinder A S 1988 Credit, money, and aggregate demand. *American Economic Review* **78**: 435–9

Bernanke B S, Gertler M 1995 Inside the Black Box: The credit channel of monetary policy transmission. *Journal of Economic Perspectives* **9**: 27–48

Blanchard O J 1990 Why does money affect output? A survey. In: Friedman B M, Hahn F H (eds.) *Handbook of Monetary Economics*. North-Holland, Amsterdam

Blanchard O J, Summers L H 1986 Hysteresis and the European unemployment problem. In: Fischer S (ed.) *NBER Macroeconomics Annual 1986*. MIT Press, Cambridge, MA

Brayton F, Levin A, Tryon R, Williams J C 1997 The evolution of macro models at the Federal Reserve Board. *Carnegie–Rochester Conference Series on Public Policy* **47**: 43–81

Cukierman A 1992 *Central Bank Strategy, Credibility, and Independence: Theory and Evidence*. MIT Press, Cambridge, MA

Debelle G, Fischer S 1994 How independent should a central bank be? In: Fuhrer J C (ed.) *Goals, Guidelines, and Constraints Facing Monetary Policymakers*. Federal Reserve Bank of Boston, Boston

DeLeeuw F, Gramlich E 1969 The channels of monetary policy. *Federal Reserve Bulletin* **55**: 472–91

Dewald W G 1963 Free reserves, total reserves, and monetary control. *Journal of Political Economy* **71**: 141–53

Driffill J, Mizon G E, Ulph A 1990 Costs of inflation. In: Friedman B M, Hahn F H (eds.) *Handbook of Monetary Economics*. North-Holland, Amsterdam

Feldstein M S 1979 The welfare cost of permanent inflation and optimal short-run economic policy. *Journal of Political Economy* **87**: 749–68

Fischer S, Modigliani F 1978 Towards an understanding of the real effects and costs of inflation. *Weltwirtschaftliches Archiv-Review of World Economics* **114**: 810–33

Friedman B M 1975 Targets, instruments, and indicators of monetary policy. *Journal of Monetary Economics* **1**: 443–73

Friedman M 1968 The role of monetary policy. *American Economic Review* **58**: 1–17

Friedman M, Schwartz A J 1963 *A Monetary History of the United States, 1867–1960*. Princeton University Press, Princeton, NJ

Goldfeld S M, Sichel D E 1990 The demand for money. In: Friedman B M, Hahn F H (eds.) *Handbook of Monetary Economics*. North-Holland, Amsterdam

Kydland F E, Prescott E C 1977 Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* **85**: 473–91

Lucas R E Jr 1973 Some international evidence on output-inflation tradeoffs. *American Economic Review* **63**: 326–34

Meulendyke A 1998 *US Monetary Policy and Financial Markets*. Federal Reserve Bank of New York, New York

Modigliani F, Rasche R, Cooper J P 1970 Central bank policy, the money supply, and the short-term rate of interest. *Journal of Money, Credit, and Banking* **2**: 166–218

Poole W 1970 Optimal choice of monetary policy instruments in a simple stochastic macro model. *Quarterly Journal of Economics* **84**: 197–216

Rogoff K 1985 The optimal degree of commitment to an intermediate monetary target. *Quarterly Journal of Economics* **100**: 1169–89

Romer C D, Romer D H 1989 Does monetary policy matter? A new test in the spirit of Friedman and Schwartz. In: Blanchard O J, Fischer S (eds.) *NBER Macroeconomics Annual 1989*. MIT Press, Cambridge, MA

Sargent T J, Wallace N 1975 'Rational' expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* **83**: 241–54
Sims C A 1992 Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review* **36**: 975–1000
Svensson L E O 1997 Inflation forecast targeting: Implementing and monitoring inflation targets. *European Economic Review* **41**: 1111–46
Taylor J B 1979 Estimation and control of a macroeconomic model with rational expectations. *Econometrica* **47**: 1267–86

B. M. Friedman

Money and Finances, Psychology of

1. Introduction

As a major tool for exchanging goods and services, money plays a crucial role in economic and social life. It serves as a legal tender and a unit for calculation, and allows values to be maintained over time. Tied to these economic functions, money has a large influence on human experience and behavior, delimiting consumer options, life styles, access to education, and even chances on the marriage market. Societies differ considerably regarding the distribution of income among their members and in the extent to which individual economic resources determine access to cultural goods such as education and health care. But individual differences in income almost invariably provide a major indicator of social inequalities. Being one of the multiple markers of social stratification, money carries social prestige and may affect social relations as well as people's self-concepts not only due to its economic function but also through the symbolic value attached to it.

Despite the trend of increasing globalization which is particularly evident in international finances and the increasing importance of stock holdings among private investors in many nations, societies still differ beyond their financial systems in cultural norms and the collective rationale which shape attitudes towards money. Collective experiences with money can affect whole nations, as was evident in the fear of inflation in Germany which followed experiences after World War I. This points to the importance of socialization not only in attitudes towards money but also for expectations concerning the larger economic system. Furthermore, cultural norms guide the display of one's financial resources in public. General financial issues such as prices and the tax system are frequently discussed. However, in many cultures—except for materialistic cultures as, for example, in south Asia—personal financial assets are mostly considered a private issue which is not discussed even with parents or friends. This lack of openness may serve to prevent social envy in the case of personal affluence or help to protect one's social image if personal resources are scarce. However, other social symbols such as clothing and cars are allowed to undermine this social norm.

Given the importance of cultural factors, the research available on psychological aspects of money and finances suffers from a focus on Western systems. Furthermore, psychologists have paid surprisingly little attention to money as a research topic (for an overview see Furnham and Argyle 1998) leaving this domain largely to sociologists and economists. Drawing on relevant research from different disciplines, we will first address the significance of money for personal well-being and social relations. Next, we will turn to beliefs and values connected with money. Thereafter we will address issues concerning the handling of money, looking at correlates and predictors of spending, saving, and donating money, as well as evading taxes. In a final section, selected research on effects of poverty will be summarized.

2. What Money can Buy: the Personal and Social Significance of Money

Although folk wisdom claims that money does not make happy, findings point to an overall low, but significantly positive relationship (about $r = .15$ to $r = .20$, see Furnham and Argyle 1998). Not surprisingly, the prediction of happiness from economic resources improves if other indicators of wealth besides income are also considered. Pointing to the importance of social comparison processes, relative income is more important than its absolute level, because people define themselves as wealthy or poor with regard to others. In countries with high poverty rates and extreme income differences like India or Brazil, the relation between happiness and income is substantially stronger. Across 55 nations, the correlation between income and happiness reaches about $r = .50$ (Diener et al. 1995). Finally, increases in income during prosperous periods have a positive influence on citizens' well-being, whereas income loss during depression periods have the opposite effect. However, the effects of changes in income on changes in satisfaction do not seem to be long-term. People adapt to higher as well as lower levels of income (for an overview see Furnham and Argyle 1998).

Much attention has been paid to the significance of money in family formation and family life (see White and Rogers 2000). Economic assets play a considerable role in marriage decisions, not only among women, who have traditionally been assumed to consider their partner's bread-winning potential, but also among men. Men's earning and employment have consistently been found to increase the likelihood of marriage, and even women's increasing participation in the labor force does not seem to render male income less important. Furthermore, the income of both men

and women, particularly the income differential between husbands and wives, affects the division of household labor between marital partners beyond the time restrictions due to their employment. There is more sharing of household tasks between spouses if wives contribute more to family income (Coltrane 2000). Economic resources also affect family relations. Conflict over financial affairs seems to be among the most important sources of discord in partnerships and family violence has often been linked to the lack of financial resources. As to divorce, findings demonstrate a positive correlation between men's earnings and marital stability (e.g., Hoffman and Duncan 1995). Benefits from women's higher earnings are less clear, because the stabilizing effect may be counteracted by the higher economic independence of the woman which contributes to greater dangers of divorce (the 'independence effect'). Furthermore, the causal link between women's income and marital quality/stability may be bi-directional: Marital discord was found not only to be affected by women's earnings, but also to instigate women's higher labor force participation, thus leading to higher earnings (see White and Rogers 2000).

3. *Attitudes Towards Money*

Attitudes towards money may be understood within a larger system of values which has been broadly distinguished as materialistic or postmaterialistic. Based on Maslow's hierarchy of values, Inglehart (1977) claims that materialistic needs, related to stable economic resources and physical security, are basic and will be foregrounded in a person's value system if they are not sufficiently fulfilled, specifically—coined in a socialization hypothesis—if they were lacking during childhood and adolescence. Conversely, postmaterialistic values, including political participation, the freedom of speech, and self-realization should develop primarily as materialistic needs are satisfied. In line with this assumption, research shows that materialistic values are less foregrounded among the well-educated and those with higher income, and that postmaterialistic values have become increasingly evident among younger generations.

It is remarkable that money is associated with ambivalent characteristics and paradoxical meanings. Money is associated with positive values like security, independence, freedom, prestige, and possessions as well as with negative associations like anxiety, anger, greed, envy, and conflict. People owning much money are likely to be seen as hard-working, successful, and even beneficent, but are also at risk of being regarded as superior, selfish, capitalistic, and hostile. Given these complexities, research has focused on understanding the basic structure of attitudes towards money.

Several different questionnaires have been constructed and psychometrically examined in order to investigate the fundamental dimensions underlying attitudes to money (Furnham and Argyle 1998). For example, the Money Attitude Scale (MAS) differentiates between the factors of power/prestige, retention time, distrust, quality and anxiety, whereas the Money Belief and Behavior Scale (MBBS) measures the factors: obsession, power/spending, retention, security/conservative, inadequacy, and effort/ability. The short version of the Money Ethic Scale (MES) summarizes three factors: budget, evil and success. All the different approaches to the measurement of attitudes to money illustrate that the attitude structure is complex and multidimensional with clearly differentiated factors, reflecting the ambivalence of money. Some of those factors are linked to clinical traits, such as anxiety or obsession, whereas others are related to the retention factor, like saving and investing money. Other factors like power, prestige and spending show that money can also be a means to influence and impress others.

Affected by parental income, education, social class, beliefs, parenting practices, and money habits, attitudes to money are formed early in childhood and remain fairly consistent throughout one's life. Empirical studies show that high-income people tend to think money is less evil and rather reflects one's achievement. Participants with a High Protestant ethic are more likely to budget money properly and tend to see money as evil and representing power. Highly leisure-oriented individuals see money more frequently as good, representing achievement and power/freedom. In contrast to students with credit cards, those without are more likely to score higher on money attitudes of obsession and retention. Other findings illustrate that compulsive spenders have relatively lower self-esteem than normal consumers, believe more that self-esteem can be enhanced by money, and are more convinced that money solves problems.

There are also attitudes towards the notes and coins as the actual currency. Bruner and Goodman (1947) show the influence of subjective values and objective needs on the perception of currencies by asking rich and poor 10 year-old children to match circles of light corresponding to a range of coins. The results illustrate that the size of socially-valued objects such as coins are judged larger than neutral objects like grey discs. Moreover, the greater the value of the coin, the more biased the estimation. This overestimation effect occurs more strongly for poor than for rich children.

4. *Handling Money*

4.1 *Saving, Consuming and Investing Money*

The decision to save money or to spend it for consumption is partly based on rational economic

considerations. However, saving is also a socially accepted goal, formed during socialization and by norms of the social environment. To predict people's saving behavior, economists focus on factors such as the increase of income which leads to a disproportional increase of savings, social comparisons with other people which influence habits of consumption and saving, and saving as postponing consumption in order to guarantee a constant level of lifelong consumption. The framing approach of the 'behavioral life cycling'-hypothesis by Shefrin and Thaler (1988) stresses the importance of individual self-control, self-imposed rules, and prior commitments for explaining saving. Different forms of money, e.g., regular income, assets or windfalls, are subjectively categorized and assigned to different 'mental accounts.' The propensity to save all or part of one's income is assumed to depend on which mental account the income is referred to.

Katona (1975) emphasizes the importance of differentiating between people's *ability* to consume or save and their *willingness* to do so. He has pointed out that people's beliefs about economic issues mediate the amount of money they save. People's consuming or saving behavior is based on their expectations regarding the future development of the economic situation. The Index of Consumer Sentiment (ICS) developed by Katona (1975) predicts turning points in business cycles and expresses changes in optimism or pessimism about the financial future. It was found that total saving goes up when people become more uncertain and worried about the future. One-third of household saving in the USA is explained by the precautionary motive, in particular with regard to health and old age considerations. Because saved money is often used as a protection against economic risks, risk-averse subjects tend to have lower debts and save more money than risk-taking persons do.

Empirical studies show (Furnham and Argyle 1998) that age is linearly related to saving (the older the person, the higher the savings) and education has a curvilinear relation (the best- and the least-educated people save less, the better educated investing more of their money instead of saving it). Savers are more optimistic about economic circumstances and believe in personal control over finances; whereas nonsavers think themselves more as victims of external events, but also seem to blame themselves more when things go wrong. Savers get more social support from friends and relatives than nonsavers and thus more information about how to handle finances. Savers show more consistent shopping habits, like going to a few favorite shops whereas nonsavers tend more to shop around.

Psychological research on financial investments in stock markets focuses on investigating the heuristics people apply, since decisions are highly influenced by the context in which price information is presented and therefore salience and attention effects are important. For example, selective causal attributions provided by the media have a significant influence on investment decisions in financial markets (Andreassen 1987). By explaining recent changes in stock prices with good news after prices rise and with bad news after prices fall, reporters tend to focus on the sets of facts suggesting that change should have happened and omit the set of facts that suggest it should not have occurred. By these attributions expectations are formed that changes will either persist or regress to previous price levels. This can mislead investors and therefore result in wrong decisions.

4.2 Tax Compliance and Evasion

Twenty to 30 percent of taxpayers in Western industrial nations do not declare their income completely. Approximately 10–20 percent of tax is illegally evaded. The majority of people think their tax burden is too high, and that they pay more taxes than other people and even more than those with a comparable income. The lack of transparency of the tax system and people's lack of knowledge about tax issues increases distortions in perception systematically (for overview see Jonas et al. 1999).

The economic approach suggests that paying and evading taxes is a behavior that is very rational, and is influenced only by the benefits and costs associated with tax evasion (like the subjectively-perceived probability of detection, the rates of penalties, the amount of one's income, and the level of marginal tax rate. For an overview see Andreoni et al. 1998). However, tax evasion is not only a matter of rational calculation and utility maximization, but also involves psychological variables like an individual's willingness to evade or comply. The more people experience money and profits in particular as an indicator of success and the more they perceive opportunities to do so, the higher the tax evasion. Negative attitudes towards the government, towards the fairness of the society's resources allocation, towards the legislation, and towards the spending behavior of political actors are also important. Furthermore, tax evasion is lower the more important moral norms and religious values are for a person and the higher her integration in the society. Finally, perceptions of tax fairness, judged on the comparison of the individual's tax burden and the government's provision of public goods as well as the tax burden of other taxpayers, is also an important predictor.

4.3 Donating Money

It is amazing that, on the one hand people are willing to take illegal action to save money by evading taxes but, on the other hand, they voluntarily give money away by donating it to charities. The figures of the

Charities Aid Foundation show for Great Britain that more than two-thirds of the people donate money for charitable causes. However, the relationship between income and donations is quite low (only about r = .13; Furnham and Argyle 1998). Considering the characteristics of people who donate money, it is found that people between the age of 50 and 65 give most, women give more money than men and are also more actively involved in volunteer work. Married people give about 20–40 percent more than singles with a comparable income. Having children leads to an increased willingness to donate (about 50 percent more than people without children). People who are religious give about three times more than people to whom religion is not important. Concerning contextual factors, more money is given if the request comes from a physically attractive person, a similar person, or an expert. More money is given in reciprocal relationships and to one's own group. Touching the potential donors and eye-contact are effective strategies for raising money. Furthermore, role modeling and conformity in groups proved to be crucial factors.

5. *When Money Runs Short: Perspectives on Poverty*

In the past decade, the problem of poverty has become addressed increasingly by social and developmental psychologists. Relevant research dates back to the 1930s when Jahoda et al. (1975, [1938]) introduced the ethnographic approach in studying a whole community hit by unemployment due to plant closing. Showing that different 'reaction types' (e.g., 'unbroken,' 'in despair,') emerged as styles of coping with the increasingly desperate situation, early findings inspired later stress and coping research. Addressing family adaptations to economic loss, Elder (1974) showed that the families' adaptive reactions, along with parents' stress, during the Great Depression, increased the likelihood of conflict and tension within families and shaped children's short- and long-term development.

Present research focuses on relative income poverty, defined by national poverty lines or by a per capita income of less than 50 percent of the average income. In contract with former static approaches, more recent studies consider the duration of poverty, e.g., suggesting that stable poverty is experienced by only a small minority of children in the USA while a large proportion lives for at least two years in poverty. In general, findings show that long-term poverty has more negative effects on well-being and children's behavior than short term deprivation, but the latter is often detrimental, too.

As to the effects of poverty, research pays increasing attention to children who have been pointed out as an important risk group of poverty in many countries, especially if living with a single parent, having two or more siblings, or belonging to ethnic minorities (e.g., Seccombe 2000). Most studies report deleterious effects of poverty on both children's and adults' well-being and health, including lower life expectancy, higher infant mortality, and more acute and chronic health problems. Disadvantages often begin during pregnancy leading to low birth-weight and undernutrition, thus increasing the risk of serious illness. In explaining these findings, impaired health among the poor has been linked to limited access to medical or prenatal care, to less healthy lifestyles, nutritional deficiencies, and low levels of education or literacy, making it difficult to use health services or follow instructions. The negative consequences for the immune system of high levels of emotional stress have also been pointed out.

Furthermore, children living in poverty have more socio-emotional and behavior problems, have more difficulties in peer relationships, and do poorly in school. Many of these effects can be linked to parental stress which results from the economic pressure they face and which spills over into the parent-child relationship leading to less nurturing, more authoritarian, and more inconsistent, harsh physical discipline. Hence, impaired parenting is important in mediating the negative effects of income loss on children. However, individual and family resources, (e.g., emotional stability, a harmonious marriage) may buffer the consequences of income loss for family life and children's development. Increased social support from family and friends has been shown to reduce the stress and violence experienced by poor families, and to improve self-esteem. High rates of poverty in the neighborhood increase the likelihood of children's externalizing problems even beyond the individual family income. Given the spatial segregation of poor families, problems arising from the lack of family and community resources are likely to accumulate.

Finally, it should be noted that running up debts is often used to compensate for a lack of money. Despite the risk of serious financial problems due to long-term obligations, only a few studies in psychology address such issues. Lea et al. (1995) report that debtors have fewer money management facilities than non-debtors, rate their ability at money management lower, and have shorter time horizons. Debt counseling has become an important task in social services provided to the poor.

6. *Conclusion*

Overall, research suggests that money is an important resource which affects many domains in life. Although much of its psychological significance seems to derive from social comparison processes, the lack of money may have far more substantial effects. Attitudes towards money are complex and often ambivalent

and, somewhat similarly, determinants of handling money go beyond economic rationale. It would seem fruitful to include more genuinely psychological perspectives in studying the significance of money in people's life.

See also: Economics and Ethics; Economics, Philosophy of; Finance, Geography of; Money: Anthropological Aspects; Money, Sociology of; Pay and Compensation, Psychology of; Poverty, Culture of; Poverty: Measurement and Analysis; Poverty, Sociology of; Psychology and Economics; Sex Differences in Pay

Bibliography

Andreassen P B 1987 On the social psychology of the stock market: aggregate attributional effects and the regressiveness of prediction. *Journal of Personality and Social Psychology* **53**: 490–96

Andreoni J, Erard B, Feinstein J 1998 Tax compliance. *Journal of Economic Literature* **36**: 818–60

Bruner J S, Goodman C 1947 Value and need as organizing factors in perception. *Journal of Abnormal and Social Psychology* **42**: 33–44

Coltrane S 2000 Research on household labor: Modeling and measuring the social embeddedness of routine family work. *Journal of Marriage and the Family* **62**: 1208–33

Diener E, Diener M, Diener C 1995 Factors predicting the subjective well-being of nations. *Journal of Personality and Social Psychology* **69**: 851–64

Elder G H Jr 1974 *Children of the Great Depression*. University of Chicago Press, Chicago

Furnham A, Argyle M 1998 *The Psychology of Money*. Routledge, London, New York

Hoffman S D, Duncan G J 1995 The effects of income, wages, and AFDC benefits on marital disruption. *Journal of Human Resources* **30**: 19–41

Ingelhart R 1977 *The Silent Revolution. Changing Values and Political Styles among Western Republics*. Princeton University Press, Princeton, NJ

Jahoda M, Lazarsfeld P F, Zeisel H 1975 [1938] *Die Arbeitslosen von Marienthal. Ein soziographischer Versuch*. (The unemployed of Marienthal. A socio-graphic test.) Suhrkamp, Frankfurt, Germany

Jonas E, Heine K, Frey D 1999 *Ein Modell der Steuerzufriedenheit—psychologische Grundlagen (un)ökonomischen Handelns*. (A model of tax contentedness—psychological basis of (un)economic behavior.) In: Fischer L, Kutsch T, Stephan E (eds.) *Finanzpsychologie*. Oldenbourg, Munich, Germany

Katona G 1975 *Psychological Economics*. Elsevier, New York

Lea S E G, Webley P, Walker C M 1995 Psychological factors in consumer debt: money management, economic socialisation, and credit use. *Journal of Economic Psychology* **16**: 681–701

Seccombe K 2000 Families in poverty in the 1990s: Trends, causes, consequences, and lessons learned. *Journal of Marriage and the Family* **62**: 1094–113

Shefrin H M, Thaler R H 1988 The behavioral life-cycle hypothesis. *Economic Inquiry* **26**: 609–43

White L, Rogers S J 2000 Economic circumstances and family outcomes: A review of the 1990s. *Journal of Marriage and the Family* **62**: 1035–51

E. Jonas, S. Walper, and D. Frey

Money: Anthropological Aspects

Money, in various cultures and at various times in history, is defined primarily by what it does and secondarily by what form it takes. In anthropology, money falls into two major categories of cultural constructs: (a) special money which is used in a limited sphere of transactions and is also called special-purpose money, primitive money, or ritual money; and (b) general money used extensively in many social spheres but used primarily for commercial exchange, and also referred to as general-purpose money.

1. Special Money

Many societies use objects to mark ritual transactions in limited spheres of social life such as marriage negotiations, religious offerings, or the acquisition of titles or ritual knowledge. The material item offered may be as varied as mats and bark cloth in Samoa, whale's teeth in Fiji, woven rolls of feathers on Santa Cruz Island, clumps of tobacco among North American Indians or dogs teeth in the Admiralty Islands. However, such items, rarely circulate as a general form of money to buy land, labor, or commodities. The items are usually confined to exchanges within quite limited ritual spheres, but the nature of these spheres varies from one society to another.

The specialized uses of such money can be seen in traditional *Tiv* society in Nigeria where magical rites could only be purchased by means of *tugudu* cloth or brass rods. The famous large stone money of Yap in the Caroline Islands was traditionally only used by aristocratic families in ritual transactions as a means of settling disputes; they were never used for commerce. Indian tribes in the northeastern USA used wampum as a means of sewing large belts with designs on them. These wampum belts were exchanged as a marker of major political events such as the union of tribes in the Iroquois Confederacy. Many forms of money, such as the rolls of feather money on Santa Cruz, were used primarily or exclusively in the sphere of marriage exchanges.

The close association of money with ritual and religious spheres of life persisted into classical times when temples served as the primary mints for making

coins. Earliest coins carried religious images of the gods who protected the money. Athenian coins depicted the city's patron goddess or symbols of her such as the owl. Roman coinage depicted Juno Montea, from whose name derive the modern words of 'money' and 'mint.'

2. General Money

Of the main organizing principles of settled social life, general money was one of the last to develop, long after writing, bureaucracy, priesthood, metallurgy, and kingship. General money arose as a political creation of the state to organize and control commerce and to standardize taxation.

2.1 Origins of General Money

Around 640 BC, the royal family of Lydia began issuing the first coins. Prior to the invention of money, most urban centers were organized and built in a manner that focused on a major temple, palace, or fortress. However, Lydia introduced the first cities centered on a marketplace that became common throughout the Greek world as the *agora*. Herodotus notes that it was not until after the state issued coins in Lydia that its capital, Sardis, became the first city to have a marketplace in its center.

As the state took over the control of money and it became secularized in the domain of politics and commerce, it began to lose its ritual functions, and a great tension between religion and monetary activity began. Beginning in the classical era of Greece and Rome, much religious, ethical, and philosophical literature has been critical of the influence of money and has advocated limits on its use and importance in society. This conflict appeared particularly in most of the religions such as Orthodox Jews' reluctance to use money on the Sabbath, the Biblical accounts of Jesus driving the moneychangers from the temple, or the strict Muslim prohibition on interest. The same distrust of money appeared in the works of Aristotle who distrusted the market and felt that prices should be set not by the scarcity of goods but by the ability of the buyer to pay.

2.2 Functions of General Money

General money is usually defined and explained by what it does within the commercial and financial spheres of life, i.e., by the economic functions that it fulfills in society. These functions include the use of money as a means of exchange, a method of setting or comparing value, a mode of payment, and a store of value. The degree to which items commonly called money actually fulfilled all three functions has varied through history according to the item, the culture, and the social circumstances.

2.2.1 Money as exchange. The commonest function of money for most people is its use as a means of exchange; people use it to buy and sell. The woman, who comes to market to sell eggs with the hope of buying milk, does not have to find a person who only wants to exchange milk for eggs. By using money, the egg seller can sell to anyone who wants eggs and then use that money to buy from any milk seller in the market. The medium of money makes the transactions much faster, easier, and more efficient by separating the individual acts of buying and selling.

2.2.2 Money as a store of value. Money further separates the buying from the selling act by allowing a seller to exchange something for money, and then to keep the money without necessarily buying anything. The egg seller may wish to save up to buy a motorcycle or camel. She could not do this by saving the eggs until she had enough to make such a large purchase, but by converting her weekly production of eggs into money, she can save the money far more efficiently and reliably.

2.2.3 Money as a means of setting value. Just as money is a means of establishing the relative value of eggs, milk, camels, and motorcycles, it also allows a value to be set on a great variety of goods that cannot physically be taken to a market square or store for sale. People can use it to set the value on an hour or month of work or for the sale or temporary use of a building or parcel of land. It can be used to set the value of goods that never exist such as the potential earning power and emotional value of a person killed in an accident or of damage to future sales caused by malicious slander. Money can be used as easily to purchase illegal goods and services as to set the price for the forgiveness of small sins or free a soul from eternal damnation.

2.2.4 Money as a system of keeping accounts. Only 8 to 10 percent of the world's money actually exists in the form of paper or coin. Most of it exists in the form of numbers contained in accounts. The hours of work a person performs is recorded in money terms in an account from which it can be drawn out to pay mortgages, make loans, or be lent out for interest. Money is the conceptual medium that allows this easy movement between hours of work, rent, buying a home, borrowing to take a vacation, or negotiating a contract to kill a rival.

2.3 Forms of General Money

After categorizing money by function, the second most important categorization is according to the form or material from which it is made. How money performs its functions is always limited by the form it takes, and this can be as varied as slips of paper, sea shells, electronic bits, ink entries in a journal, chunks of metal or edible commodities such as salt, rum, or cacao. Money—like the calendar and the system of measurements—is a cultural construct that may have arbitrary aspects, but to function properly, it needs stability and predictability. As long as people agree on the system, time can be as easily measured by the sun as by the moon, and money can be as easily calculated in feathers and shells as in silver grains.

2.3.1 Commodity money. In addition to the items used as special money, societies have used many common commodities for money. These have included rice in the Philippines, tea bricks in Mongolia, coconuts in the Nicobar Islands, furs in Alaska, sugar in the Caribbean, maize in Guatemala, cacao in Mexico, almonds in India, and barley in Babylonia. In the West alone, many different commodities have been used such as salt (from which we derive the word 'salary'), deer skins ('bucks'), and cattle ('fee').

Although commodities serve as a good medium of exchange, they are frequently awkward to transport, and their utility as a store of value depends on their perishability. The salt dissolves in water, the rice rots, and people drink the rum. Commodities usually exist in hierarchies of which metals rank the highest because of their durability. Metals such as copper, tin, bronze, gold, and silver proved to be both a good medium of exchange and a holder of value as well as an easily transported item. Among the metals, silver and gold ranked as the most valuable because of their scarcity combined with the relative ease with which they are mined and used.

Commodity money almost always operates in conjunction with a money economy and not as a self-contained or isolated economic system. When faced with a lack of currency, American colonists, who were already part of an international monetary system, used tobacco in the eighteenth century. Under communist rule in the twentieth century, Romanians used cigarettes, and in wartime a number of commodities from sugar and chocolate to butter and oil circulate as a better means of exchange than the national currency. Because commodity money arises spontaneously in societies accustomed to using money, but temporarily unable or unwilling to use the general money, commodity money will probably continue to play a major, although episodic role, in times of warfare and other forms of social upheaval.

Anthropologists have frequently used the term 'primitive money' to designate both commodity money, which can occur in any type of money system at any time, and ritual or special forms of money, which usually occur in societies that are not fully integrated into the world commercial system. The confusion arose because traders and merchants often transformed special money such as brass rods among the *Tiv* or *Wampum* among North American Indians, into a primitive form of general money in order to connect the previously more isolated group into the emerging world economy.

2.3.2 Paper money. The Chinese invented both paper and printing which were combined to produce the first prototypes for paper money which did not become a general form or money until the Mongol conquests of the thirteenth century. The nomadic Mongols seized on the concept and organization of paper money as more efficient than coins for administering a large empire with a large economy. Merchants and peasants resisted the imposition of paper money, and only the great power of the Mongol rulers and their enforcement of capital punishment made the subject people willing to use it. The Mongol Empire ended by 1368 with the fall of its Yüan dynasty in China. The Ming dynasty stopped printing paper money and soon stopped accepting it in payment. The paper money system collapsed, but it was picked up during the Renaissance in Europe where it had great appeal to the Italian bankers, and was later taken over by central banks issuing paper money backed by the rising power of the state. The development of banking and the newly developed paper money system destroyed feudalism, changing the basic organization of society from one based on heredity and the ownership of land and feudal rights to a system based on money and the ownership of wealth through stocks, bonds, and other paper instruments of ownership.

2.3.3 Electronic money. Throughout the twentieth century, the volume of money flow increased greatly with the introduction of new means of payment such as credit cards, the expansion of banking activities, and the introduction of electronic communications. To speed the movement of money of account, people had to rely increasingly on electronic modes beginning with the telegraph and spreading to the telephone and eventually to computers and satellites. Electronic money operates in the realm of money of account rather than with coins or bills, and it has made the transfer of this type of money much easier and cheaper than the transfer of cash or written documents.

By making money move faster, although not necessarily more efficiently, electronic money created new threats to the economies of the poorer nations. Because of currency speculation or currency flight, the value of the money, and therefore the entire economy, could be crippled within days rather than over the course of the months and years that it traditionally took for inflation to destroy it.

Electronic money produced a further stratification of money systems in which the richer people and richer countries use electronic money, while poorer countries and poor people use cash. Throughout the twentieth century, cash became increasingly the money of poor countries and of poor pockets within the highly developed economies.

3. *The Meaning of Money*

As money has increased in importance in the globalized world economy, social scientists have begun research into the importance of money outside the economic realm. Such inquiries often use the concept of the 'meaning' of money. The 'meaning' of money may be found in the abstract role of money in social institutions, or it may refer to the differential impact of wealth and poverty or the absence and abundance of money in the lives of individual people or groups.

3.1 *Noneconomic Aspects of Money*

Social scientists often analyze the meaning of money in institutions outside of the financial and commercial realms. Such analysis examines the functions, importance, and role in noncommercial spheres of life such as its social or psychological meaning. Zelizer (1994) examines the way people construct different categories of money according to its role in social relations such as courtship, gift giving, and inheritance. Doyle (1992) examines its role in the emotional life of the individual as money relates to levels of satisfaction, accomplishment, and security.

3.2 *Alternative Explanations of the Money Economy*

Anthropologists have researched the different, and often noneconomic, ways that people explain money and commercial activity. Taussig (1980), Gudeman (1986) and White (1994) researched the way people such as Bolivian miners, Colombian peasants, and Turkish factory workers and other groups on the margins of the world system, conceptualize the economic world around them. Humphrey and Hugh-Jones (1992) pursued an analysis of the meanings associated with all forms of economic exchange.

Carrier (1997) showed how the same form of analysis that produces a greater understanding of the meaning of money and markets in simple societies can also be used to understand how people, including scholars and other professionals, in the fully developed Western economies think about and explain the economic world around them. All of these works seek a more holistic analysis of money and how it fits with the political, religious, psychological, and other cultural aspects of social life.

See also: Exchange in Anthropology; Exchange: Social; Markets: Anthropological Aspects; Money and Finances, Psychology of; Money, Sociology of; Primitive Society; Ritual; Symbolism in Anthropology; Trade and Exchange, Archaeology of

Bibliography

Bohannan P J 1959 The impact of money on an African subsistence economy. *Journal of Economic History* **19**: 491–503

Carrier J G 1997 *Meanings of the Market*. Berg, Oxford, UK

Crump T 1981 *The Phenomenon of Money*. Routledge & Kegan Paul, London

Davies G 1994 *A History of Money*. University of Wales Press, UK

Doyle K (ed.) 1992 *The Meanings of Money*. Sage Publications, Newbury Park, CA

Einzig P 1966 *Primitive Money*, 2nd edn. Pergamon Press, London

Gudeman S 1986 *Economics as Culture*. Routledge & Kegan Paul, London

Hart K 1986 Heads or tails? Two sides of the same coin. *Man* **21:** Dec. 636–56

Humphrey C, Hugh-Jones S (eds.) 1992 *Barter, Exchange, and Value: An Anthropological Approach*. Cambridge University Press, Cambridge, UK

Polanyi K 1998 *Primitive, Archaic and Modern Economies*. Beacon Press, Boston

Simmel G 1978 *The Philosophy of Money*, (trans. Bottomore T and Frisby D). Routledge & Kegan Paul, London

Taussig M T 1980 *The Devil and Commodity Fetishism in South America*. University of North Carolina Press, Chapel Hill, NC

Weatherford J 1997 *The History of Money*. Crown Books, New York

White J 1994 *Money Makes Us Relatives*. University of Texas Press, Austin

Zelizer V A 1994 *The Social Meaning of Money*. Basic Books, New York

J. Weatherford

Money, Sociology of

Money is an abstraction that observers make from social interactions. It is a matter of degree, to the extent that interactions transfer rights to goods and services by means of tokens that could also serve transfers of other such rights; we call those tokens money. The more generalizable across social locations,

varieties of goods and services, interaction partners, types of rights, and physical forms of the token itself is that capacity to facilitate transfer of rights, the more readily people recognize the token involved as monetary. International currencies, nationally issued legal tenders, electronic monies, bank accounts, and other highly liquid tokens of transferable rights represent one extreme of a continuum running from such generalized forms to the narrowly limited circuits of such other monies as credits in baby-sitting pools, casino chips, or investment diamonds.

After a review of previous thinking about money in sociology and other social sciences this article outlines an agenda for the sociological analysis of money.

1. Money in the Social Sciences

Since 1900, unfortunately, analysis of money's production and distribution has remained chiefly the preoccupation of economists (see e.g., Eatwell et al. 1989). Other social scientists and historians have concentrated on money's consumption—on monetary practices and their consequences. Within that realm, views of money run from considering money as an impersonal, autonomous force to examining money as a set of culturally and structurally embedded practices. Existing approaches also split over their assessment of money's effects: they range from money as a force for liberation, efficiency, and individuality, to money as a source of corrosion, corruption, and oppression.

In anthropology, for example, ethnographies of nonliterate communities long documented the cultural and social variability of currencies. Yet when it came to modern money, anthropologists typically adhered to economic models of an efficient, fungible, market-based currency. The introduction of modern currencies, in such accounts, not only neutralized but corrupted the more vulnerable cultural or social particularisms of simple societies.

Starting in the 1980s, however, anthropologists began to revise such causal assumptions radically, as they explored the close interplay between monetary transactions and the construction of social relations and meaning systems. A case in point is Jane Guyer's *Money Matters* (1995), a collection of studies on West African currency history challenging the view that generalized European currencies drove out the earlier localized, 'primitive' monies of Europe's African colonies. In a similar vein, Jonathan Parry and Maurice Bloch's influential volume of essays, *Money and the Morality of Exchange* (1989), disputes money's universally homogenizing powers by documenting money's cultural variability with persuasive ethnographic detail. The dominant 'economic mythology' of money, Bloch (1994) tells us, is not only debunked by the exotic cultural distinctiveness of non-Western societies. Even in the most advanced of capitalist states, Bloch insists, people's meaningful social relations mark their monetary transfers (see also Znoj 1998).

Within sociology, such luminaries as Karl Marx, Max Weber, and Georg Simmel made two different contributions to the understanding of money: first, an analysis of monetization as part of the creation of modern economic institutions; and second, an examination of effects of monetization in the contemporary world. The second contribution has had a more durable influence than the first. Classical theories treated money as an impersonal neutral medium of economic exchange suitable for the increasingly rationalized market-driven contemporary world. A money economy, Simmel (1900) argued, increased individual freedom to select terms and partners of economic exchange. At the same time, but more ominously, money inevitably undermined meaningful social ties, reducing interpersonal connections to instrumental calculation.

Sociological theorizing long stayed close to such conceptions of money, framing money not only as separate from but also as a threat to the social sphere (see e.g., Habermas 1989). Even Talcott Parsons' (1967, p. 358) analysis of money as a symbolic language restricted money's symbolism to the economic domain. Since the 1980s, however, North American and European sociologists have challenged the theoretical segregation of economic transactions, portraying economic institutions and activities, including money, as socially contingent phenomena.

Consider, for instance, Wayne Baker's (1987) analysis of the construction of money in financial markets. Rejecting the view that governments and their central banks define and create money, Baker argues that nongovernmental financial institutions play the key part in the creation of money. The more central the location of financiers and their financial organizations within the economy, the more their activity influences monetary flows. On a smaller scale, other sociologists have examined monetary practices, showing their embedding in complex webs of meaningful social relations (Zelizer 1994; for other work in the sociology of money, see Baker and Jimerson 1992, Poggi 1993, Dodd 1994, Mizruchi and Stearns 1994, Carruthers and Babb 1996, Wuthnow 1996, Singh 1997, Ingham 1998, Mongardini 1998, and Zelizer 1998).

Pressing problems for the sociology of money include the following four controversial questions: (a) what is money? (b) where does money come from? (c) how do monetary transactions work? (d) what are their consequences? A useful way to address these problems is to distinguish among three partly independent phenomena—*monetary media*, *transfers*, and *meaningful social relations*—examining their reciprocal interactions (for a similar set of distinctions see Akin and Robbins 1999). Monetary media are representations of rights to goods and services, usually in the form of concrete tokens; transfers are social practices and accounting systems that convey media;

meaningful social relations are connections among persons and groups involved in monetary transactions. Thus, a firm awards vouchers (media) for a paid vacation (transfer) to its top salesman (relationship) while giving a weekly check (medium) summing hourly wages (transfer) to a temporary receptionist (relationship).

1.1 Monetary Media

While classical interpretations portrayed modern money as a single, homogeneous, liquid medium for economic exchange, current analyses map out a continuum involving a wide variety of monetary media. They include highly generalizable tokens such as electronic currencies which transcend social location, multiply interaction partners, activate a variety of rights, and cover a broad array of goods and services. Toward the other end of the continuum monetary media include subway tokens, telephone cards, food stamps, and the now proliferating local currencies, such as the British LETS (local exchange and trading schemes) or the 'Ithaca money' of a college town in New York State. In these days of the euro and electronic currencies, state-backed legal currencies have moved toward an intermediate position on the continuum.

Contrary to the presumption that efficient money automatically follows the expansion of markets, each one of these currencies builds on an extensive process of social reorganization. As Eric Helleiner (1999) has shown, even state-issued legal tenders only displaced a wide variety of other currencies incompletely and through strenuous efforts by nineteenth-century state builders. Similarly the rapid disintegration of the ruble as the money of account in the former Soviet Union demonstrates the complexity and fragility of the social infrastructure supporting any particular currency (Woodruff 1999).

Monetary media, therefore, do not exist in an autonomous economic sphere with its own laws of operation undisturbed by cultural or structural constraints. Which media people adopt, when, and how depends on the type of transfer and social relations involved.

1.2 Transfers

Analysts have often presumed that monetary media circulate in a single form of transfer, the quid pro quo exchange. In fact, monetary media articulate with a great variety of transfers, including donations, bribery, theft, tipping, gambling, loans, and auctions (see Davis 1992).

The broadest distinctions people make separate *compensation* (direct exchange), *entitlement* (rightful claim to a share), and *gift* (one person's voluntary bestowal on another). Money as compensation implies an equal exchange of values, and a certain distance, contingency, bargaining, and accountability among the parties. Examples include the price charged by a painter for a portrait or the annual salary of a college professor. Money as entitlement implies rights not contingent on the recipient's current performance. Cases in point are alimony awards and state-guaranteed family allowances. Money as gift implies intimacy, commitment, and/or inequality plus discretion on the donor's part, as in a parishioner's donation to a church.

Of course finer and multiple distinctions appear within these three categories and at their boundaries. Among gifts, for example, tips, bribes, wedding presents, and charitable contributions differ significantly from each other. At the core of all of these cases, however, is not quid pro quo exchange but the assertion and representation of particular kinds of meaningful social interaction. That is why people often struggle bitterly not only over amounts of money but over forms of payment. For example, controversy over whether welfare recipients should work for their income, receive services in-kind, or benefit from regular cash payments centers on defining the relationship of these recipients to public authorities and donors. Westerners have been debating these issues vigorously for 200 years.

The conjunction of monetary media and transfer routines matters, because together they convey the character of the transaction.

1.3 Social Relations

The full meaning of any monetary transaction, however, also depends on the relationship between the parties. Contrary to the view of money as a great equalizer, people actively distinguish transfers and media according to the social relations involved. Parent–child, priest–congregant, welfare official–aid recipient, legislator–constituent, courting couple—all these relations sometimes involve monetary payments, but each calls for a very different combination of media and transfer routines.

Why do people make such distinctions? This pervasive differentiation of monetary media and transfers serves to support significant differences in the meaning and quality of people's commitments to each other. People work hard to maintain such distinctions: they care greatly about differentiating monies because payment systems are a powerful way in which they mark apart different social ties. Each of these ties has a different quality and each one therefore calls for different forms and rituals of payments. Long-time emigrants, for instance, distinguish carefully between money they send home to support their immediate families, contributions to their home churches, and money paid for taxes in their country of origin.

That is why people will respond with anger, shock, or ridicule to the 'misuse' of monies for the wrong

circumstances or social relations, such as offering a salary to a wife, a gift to a traffic policeman, or a tip to a friend. People care deeply about making such distinctions; the wrong transfer challenges, confuses, or violates the definition of particular social relations.

2. Interaction

To be sure, the three phenomena—monetary media, transfers, and relationships—influence each other. Observers have commonly thought that monetary media predominated, but have divided sharply between the economists' notion that the primary effect of monetization is efficiency and a moral view that sees monetization as creating corruption and desiccation.

Mutual influences are in fact much more varied. They involve a matching process among media, transfers, and relations. The three interact: the establishment of local currencies promotes monetary interchanges with neighbors and local merchants; the creation of well-defined routines for gift giving permits the treatment of gift certificates and crisp new bills as equivalents; the surrounding of rich people with paid service providers fosters the elaboration of routines for the bestowal of tips; and so on through other combinations of media, transfers, and relations.

Each of the three elements, however, brings to any particular interaction its own history and complex array of meanings and symbolism, a fact that limits the number of possible correspondences among the three. Legal tenders associate with particular governments and national histories, systems of pay convey various degrees of honor and commitment, friends bring to any particular monetary transfer memories of previous loans, gifts, and shared expenditures. Close observation reveals people negotiating matches among relations, transfers, and monetary media.

One crucial procedure in this matching we can call *earmarking*: that is, signaling the nature of the relationship and transfer by a particular treatment of the medium. Techniques of earmarking include:

(a) Establishing social practices that sort otherwise identical media into distinct categories. Depending on how it is used, when, and most importantly for what type of social relation, the same physically indistinguishable medium (e.g. a euro or a dollar) can serve as a wage, a bonus, a tip, a gift, an allowance, charity, or a remittance. Each calls for a different set of routines representing its character.

(b) Creation of segmented media in the form of tokens, coupons, scrip, chits, 'time-dollars,' food stamps, affinity credit cards, money orders, vouchers, or gift certificates, which are appropriate for restricted sets of relations and transfers, and in many cases are not legal tender within the larger economy.

(c) Transformation of selected objects into monetary media, as with cigarettes, postage stamps, subway tokens, poker chips, or baseball cards.

Negotiation includes both the creation of consensus about the character of the relation and the transfer, on one side, and the elimination of mismatches, on the other. In fact, legal disputes over contested payments commonly pivot on disagreements not about just one of the elements but about the proper match among medium, transfer, and relation: was that transfer of money from an industrialist's checking account to a politician's campaign fund a loan between friends or a bribe for political services rendered?

3. Challenges

Creation of new forms of money such as international legal tenders, electronic credit, and specialized local currencies will surely transform monetary practices during the twenty-first century. Such innovations set an important challenge to sociological analysis. Students of money face two major tasks. First, to specify regularities in the process of matching monetary media, transfers, and relations, by establishing and explaining correspondences among the three. We have abundant evidence that such correspondences exist, but little knowledge of how they occur. Second, sociologists must examine cases where dramatic changes occur in one of the three phenomena, for example (a) the introduction of a new medium such as the euro, or the dissolution of an established media, such as the ruble, (b) innovation in transfers, such as electronic auctions, or (c) creation of new social relations such as gay families, new international financial institutions, and transnational social movements.

In the long run, sociologists face the further challenge of making a much stronger connection with economists' studies of such classic questions as money supply, inflation, central banking, and international trade. Economics and sociology both suffer from the current segregation between their approaches to money.

See also: Consumer Culture; Consumer Psychology; Consumption, Sociology of; Economic Anthropology; Economic Sociology; Exchange in Anthropology; Exchange: Social; Markets: Anthropological Aspects; Marx, Karl (1818–89); Money: Anthropological Aspects; Simmel, Georg (1858–1918); Trust, Sociology of; Weber, Max (1864–1920)

Bibliography

Akin D, Robbins J (eds.) 1999 *Money and Modernity: State and Local Currencies In Melanesia*. University of Pittsburgh Press, Pittsburgh, PA

Baker W 1987 What is money? A social structural interpretation. In: Mizruchi M S, Schwartz M (eds.) *Intercorporate Relations*. Cambridge University Press, Cambridge, UK, pp. 109–44

Baker W, Jimerson J B 1992 The sociology of money. *American Behavioral Scientist* **35**: 678–93
Bloch M 1994 Les usages de l'argent. *Terrain* **23**: 5–10
Carruthers B G, Babb S 1996 The color of money and the nature of value: greenbacks and gold in postbellum America. *American Journal of Sociology* **101**: 1556–91
Davis J 1992 *Exchange*. University of Minnesota Press, Minneapolis, MN
Dodd N 1994 *The Sociology of Money*. Continuum, New York
Eatwell J, Milgate M, Newman P 1989 *The New Palgrave: Money*. Norton, New York
Guyer J I (ed.) 1995 *Money Matters*. Heinemann, Portsmouth, NH
Habermas J 1989 *The Theory of Communicative Action*. Beacon Press, Boston, Vol. 2
Helleiner E 1999 Historicizing territorial currencies: monetary space and the nation-state in North America. *Political Geography* **18**: 309–39
Ingham G 1998 On the underdevelopment of the 'sociology of money'. *Acta Sociologica* **41**: 3–18
Mizruchi M S, Stearns L B 1994 Money, banking, and financial markets. In: Smelser N, Swedberg R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ and Russell Sage Foundation, New York, pp. 313–41
Mongardini C (ed.) 1998 *Il Denaro Nella Cultura Moderna*. Bulzoni, Rome
Parry J, Bloch M (eds.) 1989 *Money and The Morality of Exchange*. Cambridge University Press, New York
Parsons T 1967 *Sociological Theory and Modern Society*. Free Press, New York
Poggi G 1993 *Money and the Modern Mind*. University of California Press, Berkeley, CA
Simmel G 1900 *The Philosophy of Money* [trans. Bottomore T, Frisby D (eds.) 1978]. Routledge and Kegan Paul, London
Singh S 1997 *Marriage Money: The Social Shaping of Money in Marriage and Banking*. Allen and Unwin, St. Leonards, Australia
Woodruff D 1999 *Money Unmade: Barter and the Fate of Russian Capitalism*. Cornell University Press, Ithaca, NY
Wuthnow R 1996 *Poor Richard's Principle*. Princeton University Press, Princeton, NJ
Zelizer V A 1994 *The Social Meaning of Money*. Basic Books, New York
Zelizer V A (ed.) 1998 Changing forms of payment. *American Behavioral Scientist* **41** (whole issue)
Znoj H 1998 Hot money and war debts: transactional regimes in southwestern Sumatra. *Comparative Studies in Society and History* **40**: 193–222

V. A. Zelizer

Montague Grammar

Montague grammar is a theory of semantics, and of the relation of semantics to syntax, originally developed by the logician Richard Montague (1930–71), and subsequently modified and extended by linguists, philosophers, and logicians. Classical Montague grammar had its roots in logic and the philosophy of language; it quickly became influential in linguistics, and linguists have played a large role in its evolution into contemporary formal semantics. The most constant features of the theory over time have been the focus on truth-conditional aspects of meaning, a model-theoretic conception of semantics, and the methodological centrality of the Principle of Compositionality: 'The meaning of a whole is a function of the meanings of its parts and their mode of syntactic combination' (see Partee and Hendriks 1997, pp. 20–21). This entry presents a brief outline of the origins of Montague grammar, summarizes the basic principles of the classical form of the theory, and sketches some more recent developments and interesting controversies.

1. Montague Grammar in Historical Context

Richard Montague was a logician and philosopher of language whose seminal works (Montague 1970a, 1970b, 1973) founded the theory known after his death as Montague grammar. Montague had been a student of Alfred Tarski, one of the pioneers, with Frege and Carnap, in the model-theoretic semantics of logic. As a logician, Montague built on that tradition and developed an intensional logic with a rich type theory and a possible-worlds model-theoretic semantics, incorporating certain aspects of pragmatics, including the treatment of 'indexical' words and morphemes such as 'I, you' and the present tense. In the late 1960s, Montague turned to the project of 'universal grammar,' which for him meant a theory of syntax and semantics encompassing both formal and natural languages.

Montague's idea that a natural language such as English could be formally described using logicians' techniques was a radical one at the time. Most logicians believed that natural languages were not amenable to precise formalization, while most linguists doubted the appropriateness of logicians' approaches to the domain of natural language semantics.

At the time of Montague's work, Chomskian generative syntax was well established, and linguists were developing and debating approaches to semantics to fit into the enterprise of generative grammar, and questions of the relation of semantics to syntax had become central. The 'linguistic wars' between generative semantics (Lakoff, Ross, McCawley, Postal) and interpretive semantics (Jackendoff, with the support of Chomsky) were in full swing. In the earliest introductions of Montague's work to linguists, including (Partee 1973, 1975) and Thomason's extended introduction to Montague (1974), it was argued that Montague's work offered the potential to accommodate some of the best aspects of both of the warring approaches, with some added advantages of its own.

Early attempts to integrate Montague grammar with transformational grammar (Partee 1975) faced two kinds of obstacle. On the one hand, Montague's

own syntax was quite unconstrained, and in practice combined some systematic use of categorial grammar with some *ad hoc* extensions; it clearly needed overhaul for serious linguistic work. On the other hand, transformational grammars of the period could not be directly coupled with Montague's semantics because many of the posited transformations did not have uniform semantic effects, as Montague's theory required. Much of the subsequent diversity in approaches to formal semantics that incorporate features of Montague's theory reflects different responses to those problems.

Montague's work, and related work by the philosophers Michael Bennett, M. J. Cresswell, David Lewis, Terence Parsons, Robert Stalnaker, Richmond Thomason, and others, and by linguists such as Emmon Bach, Robin Cooper, David Dowty, E. L. Keenan, Lauri Karttunen, Barbara H. Partee, gained influence during the 1970s. Formal semantics became one mainstream approach to semantics within linguistics by the 1980s; major textbooks include Chierchia and McConnell-Ginet (1990), Dowty et al. (1981), Gamut (1991), Heim and Kratzer (1998); a good 'lay introduction' is Bach (1989). See also Lewis (1970), Partee (1976), Partee and Hendriks (1997).

2. *The Theory and Substance of Montague Grammar: Central Principles*

The paper by Montague that has had the most impact on linguists and on the subsequent development of formal semantics was 'PTQ' ('The proper treatment of quantification in ordinary English,' Montague 1973)—short, but densely packed. 'Montague grammar' has often meant what Montague described in the fragment in PTQ, and the extensions of PTQ by linguists and philosophers in the 1970s and 1980s with greater or lesser innovations. But it is the broader algebraic framework of 'UG' ('Universal grammar,' Montague 1970b) that constitutes Montague's theory of grammar. This section therefore begins with the basic principles laid out in UG, concentrating on the principle of compositionality, and then continues with model-theoretic interpretation, type theory, and the 'method of fragments,' including some of the key features of the fragment in PTQ.

2.1 *Universal Grammar: Syntax and Semantics as Algebras, Compositionality as Homomorphism*

Montague's 'Universal grammar' (Montague 1970b) contains the most general statement of Montague's formal framework for the description of language. The central idea is that a grammar should be able to be cast in the following form: the syntax is an algebra, the semantics is an algebra, and there is homomorphism mapping elements of the syntactic algebra onto elements of the semantic algebra. This very general definition leaves a great deal of freedom as to the nature of these algebras. In a logical language, the elements of the syntactic algebra can be the well-formed expressions. But for a natural language, ambiguity makes that impossible, since the homomorphism requirement means that each element of the syntactic algebra must be mapped onto a unique element of the semantic algebra. So for a natural language, the elements of the syntactic algebra are usually taken to be expressions together with disambiguating structural descriptions, typically trees of some sort.

The nature of the elements of the semantic algebra is also open to variation. They may be model-theoretic constructs, as in Montague's own work, Heim's 'file change potentials,' the extensional domains of first-order logic, hypothesized concepts, expressions in a 'language of thought,' or anything else. What is constrained is not the 'substance' of the semantics but properties of its structure and its relation to syntax.

It is the homomorphism requirement, which formalizes the compositionality requirement, that provides one of the most important constraints on UG in Montague's sense. The compositionality requirement, stated in the introductory paragraph, is almost uncontroversial when stated informally. But formalizing it requires an explicit theory of syntax (to specify what is meant by *parts* and *syntactic combination*), an explicit theory of semantics (*meanings*), and an explicit theory of the mapping from one to the other (*is a function of*).

Logicians typically specify the syntax of a formal language as a recursive definition; in that case the requirement of compositionality as homomorphism can be satisfied by giving the semantics in the form of a parallel recursive definition. The 'derivation trees' that correspond to the steps in applying a recursive definition become the elements of the syntactic and semantic algebras, and the homomorphism requirement says that each syntactic derivation must be mapped onto a unique semantic derivation.

The simplest linguistic examples take the elements of the syntactic algebra to be (sub)trees generated by a context-free grammar, and semantic interpretation to specify how the interpretation of a given subtree is computed from the interpretations of its immediate 'daughter' subtrees. Formal semanticists often work either with a 'monostratal grammar' i.e., a grammar with a single level of syntax and no transformations (GPSG, HPSG, categorial grammar), or with a specified level of syntactic representation such as 'LF' in Chomskian syntax.

2.2 *Model Theory: Direct and Indirect Interpretation*

Montague's 'Universal grammar' presents formal frameworks for both 'direct' and 'indirect' semantic

interpretation. The fragment of Montague (1970a) illustrates the direct model-theoretic interpretation of natural language syntactic rules; direct model-theoretic interpretation of natural language is also found in some of the work of Cresswell, Von Stechow, and Kratzer. 'Indirect' interpretation, illustrated in UG and PTQ, and in the works of many others, proceeds via translation into an intermediate language, such as Montague's Intensional Logic. The translation process must then itself be compositional, and the intermediate language is in principle dispensable. Montague viewed the intermediate language as offering increased perspicuity in presentation; linguists tend to want evidence for the psychological reality of some level of 'semantic representation,' but strong evidence for or against such levels is still lacking.

2.3 Type Theory and Intensionality

Linguists before Montague had tried unsuccessfully to give a compositional semantics to natural language syntax, but their efforts were stymied in part by the mismatch between natural language syntax and first-order logic. The richness of Montague's logic, with several crucial features previously unfamiliar to linguists, was a crucial factor in the possibility of giving a compositional semantic interpretation to independently motivated syntactic structure.

Montague's use of a richly typed logic with lambda-abstraction made it possible for the first time to interpret noun phrases (NPs), such as 'every man,' 'the man,' 'a man,' uniformly as semantic constituents, something impossible with the tools of first-order logic. More generally, Montague's type theory represents an instantiation of Frege's strategy of taking function-argument application as the basic 'semantic glue' by which meanings are combined. This view, unknown in linguistics at the beginning of the 1970s, had become widely viewed as standard by the middle of the 1980s.

Intensionality is a key notion in the analysis of referential opacity. The construction exemplified in 'Jones is seeking_____' is *referentially opaque*, since the substitution of one coreferential expression for another in that context does not always preserve the truth value of the whole. It may be true that Jones is seeking the president, and false that Jones is seeking Mary's father even though the president is Mary's father. Montague's intensional logic develops Frege's distinction between *sense* and *reference* and Carnap's distinction between *intension* and *extension* to treat the phenomenon of referential opacity, pervasive in belief-sentences and other propositional attitude constructions.

Montague analyzed intensions as functions from possible worlds to corresponding extensions: propositions as functions from possible worlds to truth values, individual concepts as functions from possible worlds to individuals, properties as functions from possible worlds to sets. Montague proposed that not only propositional attitude verbs such as 'believe' create opaque contexts, but also transitive verbs such as 'seek' and 'need,' modifiers such as 'former' and 'alleged,' and prepositions such as 'about.' Montague's possible-worlds analysis of intensionality opened up new avenues of research as well as leading to interesting controversies and discussions of possible alternative ontological foundations for semantic theory.

2.4 The Method of Fragments: the PTQ Fragment

Each of Montague's three 'grammar' papers (Montague 1970a, 1970b, 1973) contains a 'fragment.' The term was introduced in 'English as a formal language' (EFL), which begins: 'In the present paper I shall accordingly present a precise treatment, culminating in a theory of truth, of a formal language that I believe may be reasonably regarded as a fragment of ordinary English' (Montague 1974, p. 188). The 'method of fragments,' a methodological feature of Montague's work, is the writing of a complete syntax and semantics for a specified subset ('fragment') of a language, rather than writing rules for a single construction while making implicit assumptions about the rest of the grammar.

The fragment of PTQ (Montague 1973), the best known of Montague's works, had a number of noteworthy features. One was the systematic correspondence between syntactic categories and semantic types, an appreciation of which led to a revival of interest in categorial grammar, where such a correspondence is built into the system of categories.

Another innovation was Montague's bottom-up syntax, combining concatenation and transformation operations in single recursive grammatical rules (modeled on the recursive rules of logicians' syntax), thus building well-formed expressions of all sorts of categories directly, rather than building an entire 'deep structure' top-down and then applying transformations bottom-up.

Montague's PTQ contained many other ideas that were new to linguists, some of which had antecedents in the work of other logicians and philosophers. We mention only a few here. The fragment illustrated the use of lambda-abstraction to generate phrasal conjunction directly, without the need for the semantically problematic syntactic transformation of 'conjunction reduction.' Montague introduced a version of Quine's treatment of relative clauses as 'sentential adjective phrases' derived from open sentences by lambda-abstraction, without any kind of 'deletion under identity.' He introduced the binding of bound-variable pronouns by lambda-operators rather than directly by

quantifiers or by their antecedents, a move which has had major consequences for theories of anaphora. Montague used meaning postulates to distinguish entailments due to individual lexical items from entailments due to the semantics of a grammatical construction, and laid the foundations for the strong 'lexicalist' tradition in formal semantics led by Dowty (1979). Details of Montague's analyses have in many cases been superseded, but in overall impact, PTQ was as profound for semantics as Chomsky's *Syntactic Structures* was for syntax. Emmon Bach summed up their cumulative innovations thus: 'Chomsky's thesis was that natural languages can be described as formal systems … "Montague's Thesis" [was that] natural languages can be described as interpreted formal system' (Bach 1989, p. 8).

3. Further Developments and Controversies

3.1 The Impact of Montague's Work on Linguistics

Before Montague, linguists took as the task of semantics the explication of ambiguity, semantic anomaly, and synonymy: the key questions were how many readings a sentence has, and which sentences share readings. But the individuation of 'readings' had always been problematic, so there was often crucial disagreement about data. The introduction of truth conditions as the basic semantic property of a sentence that a semantic theory should capture profoundly affected the adequacy criteria for semantics, and led to a great expansion of semantic research.

Montague had a very abstract theory of syntax (as an algebra), and some specific syntactic ideas; but in general his semantics was far more influential than his syntax. There have been many proposals for different ways to link Montague's semantics, or some variant of it, to different sorts of syntactic theories (see *Syntax–Semantics Interface*). In general, the richness of Montague's semantics eliminates some of the motivation for transformations and some other kinds of richness in syntactic theories, and lends itself to a monostratal syntax with a rich lexical component. Theories which have included a semantics influenced by Montague grammar include extended categorial grammar, GPSG (generalized phrase structure grammar), HPSG (head-driven phrase structure grammar), Heim's file change semantics, dynamic Montague grammar, and to some extent discourse representation theory.

Montague's method of 'fragments' is no longer widely followed except in computational linguistics, where completeness is a necessity. Many aspects of Montague grammar have found fruitful application in computational semantics, and many innovations (such as methods for working with underspecified representations rather than with large sets of meanings of ambiguous expressions) have in turn come from the computational semantics community.

3.2 Resistance, Skepticism, Ongoing Debate

Some aspects of Montague's approach met with resistance or skepticism in the linguistics community, and some aspects were controversial among philosophers. We mention three foundational issues that distinguish Montague grammar, and formal semantics in general, from other approaches to semantics.

Many linguists have resisted the relevance of truth conditions and entailment relations to natural language semantics, and some still do. Some of the objections to truth conditions are countered by arguing that part of human semantic competence is the matching of sentences with their truth conditions relative to possible worlds (including fictional worlds), with no necessary reference to the 'actual world.' Entailment (sentence A *entails* sentence B if sentence B is true in every possible state of affairs in which sentence A is true) is a central semantic concern in logic, and remains so in formal semantics. Cognitive semanticists replace concern with logical entailment by concern with human inference; formal semanticists see the relation of entailment to actual human inference as indirect. But many semanticists agree about the importance of revising the formal logics invented by logicians to model the 'natural logic(s)' implicit in the semantics of natural languages.

Many linguists think of semantics in terms of a 'level of representation' of expressions analogous to a syntactic or phonological level. A representational view of semantics is quite congenial to the popular computational theory of mind. The contrasting model-theoretic view sees semantic interpretation relating expressions to elements of models (possibly mental models) defined in terms of constituents such as possible situations, entities, properties, truth values, etc.

The nonpsychologistic tradition of 'objective' (though abstract) meanings (Frege, Tarski, Carnap, Montague) contrasts with the psychologistic view of meanings 'in the head' (Fodor, Lakoff, Jackendoff, and all psychologists). Do expressions refer to objects or to concepts? Is semantics a branch of mathematics, or is it (as on the Chomskian view of all of linguistics) a branch of psychology? Classical formal semanticists, who take the first disjunct in these choices, follow David Lewis in distinguishing semantics from knowledge of semantics, making semantic competence interestingly different from syntactic competence. Many semanticists seek an integration of cognitive and formal perspectives by studying mind-internal intuitions of mind-external relations such as reference and truth conditions.

3.3 Typology and Natural Language Metaphysics

The issue of 'natural language metaphysics' (Bach 1986) is an important foundational area that connects

cognitive issues with formal semantics and linguistic typology. What presuppositions concerning the constitution and structure of the world as humans conceive it are built into human languages, and how, and which are universal? These questions may concern both semantic structure and semantic content, from the semantic difference between nouns and verbs to the content of color terms or whether time is discrete or continuous. Their investigation may challenge the lines between semantic knowledge and other kinds of knowledge. Formal semantics, following the logical tradition, initially employed relatively 'austere' model structures; recent investigations, particularly into lexical semantics, invite richer models. Typological issues have included the distribution of nominal versus adverbial quantification, the semantics of languages that lack a determiner category, and the range of tense and aspect systems.

3.4 The Naturalization of Formal Semantics

As formal semantics has developed from Montague grammar since the beginning of the 1970s, it has increasingly become a core part of mainstream theoretical linguistics, with continuing important contributions from logicians and philosophers. Areas of early research success tended to be in the 'logical' parts of language: quantification, anaphora, conjunction, tense, and aspect. Later work expanded the domains covered and the range of languages treated. An important development of the 1980s was the work of Kamp and Heim on discourse representation theory and file change semantics. Their work led to changes in theories of indefinite noun phrases, quantification, and anaphora, and more generally led to the replacement of truth conditions by 'context change potential' as the core semantic property of sentences. Later work by them and others extended this 'dynamic' perspective on the interaction of meaning and context, in some cases involving associated innovations in logic. Formal theories of topic-focus structure have been developed.

The basic model theory has been enriched by additional structure on the domain of entities, to model mass nouns and 'plural entities' (Link 1998), and by widespread acceptance of events as basic entities. Kratzer and others have argued for a 'situation semantics' as a more fine-grained model structure, taking situations as parts of worlds. There is increasing attention to details of lexical semantics, and to interfaces of semantics with syntax on the one hand and pragmatics on the other.

Of the key journals in the field, one, *Linguistics and Philosophy*, first published in 1977, maintains the collaborative links between those fields that marked the beginnings of Montague grammar. One, *Natural Language Semantics*, first published in 1993, is more nearly centered within linguistics and includes greater concern with the syntax–semantics interface. A third, *The Journal of Logic, Language, and Information*, first published in 1992, reflects the split found in some European universities, where Chomskian syntax is located in linguistics departments, and formal semantics within philosophy or logic departments. While this entry has emphasized Montague grammar and formal semantics within linguistics, fruitful cross-fertilizations continue to strengthen and enrich the field as they did at its beginnings.

See also: Logic and Linguistics; Semantics; Syntax–Semantics Interface

Bibliography

Bach E 1986 Natural language metaphysics. In: Barcan M R, Dorn G J W, Weingartner P (eds.) *Logic, Methodology and Philosophy of Science VII*. North Holland, Amsterdam, pp. 573–95

Bach E 1989 *Informal Lectures on Formal Semantics*. State University of New York Press, Albany, NY

Chierchia G, McConnell-Ginet S 1990 *Meaning and Grammar. An Introduction to Semantics*. MIT Press, Cambridge, MA

Dowty D 1979 *Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Reidel, Dordrecht, The Netherlands

Dowty D, Wall R E, Peters S, Jr 1981 *Introduction to Montague Semantics*. Reidel, Dordrecht, The Netherlands

Gamut L T F 1991 *Logic, Language, and Meaning. Vol. 1: Introduction to Logic. Vol. 2: Intensional Logic and Logical Grammar*. University of Chicago Press, Chicago

Heim I, Kratzer A 1998 *Semantics in Generative Grammar*. Blackwell, London

Lewis D 1970 General semantics. *Synthèse* **22**: 18–67

Link G 1998 *Algebraic Semantics in Language and Philosophy. CSLI Lecture Notes No. 74*. CSLI Publications, Stanford, CA

Montague R 1970a English as a formal language. In: Visentini B et al. (ed.) *Linguaggi nella Società e nella Tecnica*. Edizioni di Comunità, Milan, pp. 189–224

Montague R 1970b Universal grammar. *Theoria* **36**: 373–98

Montague R 1973 The proper treatment of quantification in ordinary English. In: Hintikka K J J, Moravcsik J M E, Suppes P (eds.) *Approaches to Natural Language*. Reidel, Dordrecht, The Netherlands, pp. 221–42

Montague R 1974 *Formal Philosophy. Selected Papers of Richard Montague* [ed. Thomason R H]. Yale University Press, New Haven, CT

Partee B 1973 Some transformational extensions of Montague grammar. *Journal of Philosophical Logic* **2**: 509–34

Partee B H 1975 Montague grammar and transformational grammar. *Linguistic Inquiry* **6**: 203–300

Partee B H (ed.) 1976 *Montague Grammar*. Academic Press, New York

Partee B H, Hendriks H L 1997 Montague grammar. In: van Benthem J, Ter Meulen A (eds.) *Handbook of Logic and Language*. Elsevier, Amsterdam, pp. 5–91

B. H. Partee

Copyright © 2001 Elsevier Science Ltd. 9999
All rights reserved.

 ISBN: 0-08-043076-7

Monte Carlo Methods and Bayesian Computation: Importance Sampling

1. General Principle

The term importance sampling (IS) designates a method designed to improve the numerical performance (or efficiency) of Monte Carlo (MC) simulation methods for the evaluation of analytically intractable integrals, generally expectations of functions of random variables. Early descriptions of the method can be found, e.g., in Kahn and Marshall (1953), Trotter and Tukey (1956), or in the commonly cited monograph of Hammersley and Handscomb (1964, Sect. 5.4). Examples of applications are the numerical evaluation of Bayesian posterior expectations of functions of parameters of interest—see, e.g., Kloek and van Dijk (1978) or Geweke (1989)—and that of a likelihood function in the presence of unobserved latent variables—see, e.g., Durbin and Koopman (1997) or Richard and Zhang (2000).

The general principle of IS can be outlined as follows. Consider a situation where one has to numerically evaluate an integral of the form

$$I = \int_D \phi(x)\,\mathrm{d}x \tag{1}$$

where ϕ denotes a real function defined on $D \subset R^p$.

A case of special interest is that where ϕ takes the form of the product of a density function μ with support D and a function h whose expectation has to be evaluated on μ, say

$$I = E_\mu[h(x)] = \int_D h(x)\mu(x)\,\mathrm{d}x \tag{2}$$

In the subsequent analysis, μ will be referred to as an 'initial' sampler, to be distinguished from the importance sampler introduced below.

In low-dimensional applications (say $p \leqslant 5$), one might consider using nonstochastic multivariate quadrature rules, as described, e.g., in Stroud and Secrest (1966) or in Golub and Welsch (1969). However, higher-dimensional applications require the use of MC simulation methods. The two most widely used methods are Monte Carlo Markov chain methods (see *Markov Chain Monte Carlo Methods*) and IS which is presented here.

IS requires the selection of a class of samplers, indexed by an auxiliary vector of parameters $a \in A$, say $M = \{m(x \mid a);\ a \in A\}$. While in some applications M might consist of a single sampler, it is generally advisable to select a class of samplers and then, within that class, a specific sampler under an appropriate 'efficiency' criterion. Selection of M and of a sampler within M are discussed further below, once the purpose of IS has been clarified. In the case of formula (2), it will often be the case that M consists of suitable parametric extensions of μ and, therefore, that μ itself belongs to M, i.e., that there exists $a_0 \in A$ such that $\mu(x) \equiv m(x \mid a_0)$ on D. While, as discussed further below, such a condition may be instrumental in the construction of M, it is by no means a requirement of the method.

For any given $a \in A$, the integral in formula (1) can be rewritten as

$$I = \int_D \left[\frac{\phi(x)}{m(x \mid a)}\right] m(x \mid a)\,\mathrm{d}x = E_m\left[\frac{\phi(x)}{m(x \mid a)}\right] \tag{3}$$

In the case of formula (2), this expectation is often rewritten as

$$I = E_m[h(x)\omega(x;\, a)] \tag{4}$$

where

$$\omega(x;\, a) = \frac{\mu(x)}{m(x \mid a)} \tag{5}$$

denotes the weight function associated with the sampler m.

For any given $a \in A$, an MC IS estimate of I is then given by

$$\hat{I}_S(a) = \frac{1}{S}\sum_{i=1}^{S} \left.\frac{\phi(x)}{m(x \mid a)}\right|_{x=\tilde{x}_i(a)} \tag{6}$$

where $\{\tilde{x}_i(a);\ i{:}\,1 \to S\}$ denotes a set of S identically independently distributed (hereafter i.i.d.) draws from the auxiliary sampler $m(x \mid a)$. If $m(x \mid a)$ is such that the sampling variance

$$\sigma^2(a) = \int_D \left[\frac{\phi(x)}{m(x \mid a)} - I\right]^2 m(x \mid a)\,\mathrm{d}x \tag{7}$$

is finite, then a central limit theorem applies, whereby the random variable $\sqrt{S}(\hat{I}_S(a) - I)$ converges in distribution towards a Normal with zero mean and variance $\sigma^2(a)$. Let $\hat{\sigma}_S^2(a)$ denote the MC finite sample variance of the ratio ϕ/m in formula (6). A numerical estimate of the variance of the estimator $\hat{I}_S(a)$ is then given by $(1/S)\hat{\sigma}_S^2(a)$. Note, therefore, that once a sampler has been selected, the numerical accuracy of the corresponding IS estimator of I, as measured by its MC standard deviation is of the order $S^{-1/2}$. Poor selection of a may require a prohibitively large MC sample size in order to achieve sufficient accuracy. Conditions for the finiteness of $\sigma^2(a)$ are discussed, e.g., in Geweke (1996) or Stern (1997). A sufficient condition is that $|\phi/m|$ be bounded above on D. Convergence in the absence of such a condition may be notoriously more difficult to verify.

If there exists $a_* \in A$ such that $\phi(x) = m(x \mid a_*)$ on D,

then $\sigma^2(a_*) = 0$ and $\hat{I}_S(a_*) = I, \forall S \geqslant 1$. More generally, one would aim at selecting a value $\hat{a}$ on A which minimizes the MC sampling variance $\sigma^2(a)$. Since, however, $\sigma^2(a)$ does not have an analytical expression and can only be numerically approximated, the actual choice of $\hat{a}$ will often be based upon more heuristic considerations—some of which are discussed further below—aiming at constructing importance samplers which 'approximate' $\phi(x)$ on D. It should be noted that the term importance sampling emphasizes the fact that $m(x \mid a)$ is precisely meant to oversample in the 'important' parts of the domain D, i.e., those parts which contribute most to the value of integral I.

It is, however, useful to draw users' attention to a potentially misleading interpretation of this notion of region of importance. In practice, it is generally the case that large, possibly infinite, values of $\sigma^2(a)$ will not be generated by values of x in the region of importance (where most importance samplers are expected to perform well by design), but instead by values of x which can be very far away in the tails of $m(x \mid a)$ and which are, therefore, most unlikely to be ever drawn in a finite sample MC simulation. This situation can be particularly pernicious when D is unbounded. Earlier applications of IS to the computation of Bayesian posterior moments did occasionally suffer from such pathologies, especially in situations where noninformative prior distributions were used in combination with a (qualitatively) poorly identified model, see *Statistical Identification and Estimability*. (Note that in such cases, truncation of the support D of the posterior density may not be an acceptable solution, insofar as the moments of interest might be highly sensitive to such a truncation.) A classical example of non-existence of $\sigma^2(a)$, which has been widely cited in the early Bayesian literature on IS, is that where ϕ^{-1} has polynomial tails (e.g., if μ in formula (2) takes the form of a Student-t density kernel) and one selects for M a class of Normal samplers. Nonexistence of $\sigma^2(a)$ may be particularly difficult to detect when it originates from such low-probability events. Indications of a potential problem are the empirical findings that the MC sampling variance of $\hat{I}_S(a)$ does not decrease inversely proportionally to S or, relatedly, in the context of formula (4), that a few draws, often less than a handful, carry excessive relative weights. It is, therefore, absolutely critical to pay very close attention to estimates of the MC sampling variance of $\hat{I}_S(a)$ and/or to the behavior of the weight function in Eqn. (4), at least in those cases where the existence of $\sigma^2(a)$ cannot be formally assessed.

2. *Construction of Importance Samplers*

Thus the two critical issues in the design of a successful IS implementation are: (a) the selection of an appropriate class M of auxiliary samplers; and (b) the selection within that class of an 'efficient' sampler, i.e., one for which the MC sampling variance $\sigma^2(a)$ is as small as possible. A number of truly innovative proposals to that effect are found in the recent econometric and statistical literature, though there remains a significant lack of generic principles for the construction of efficient importance samplers (one recent exception to this general statement is presented on its own in Sect. 3 below). As discussed further in the concluding remarks, this lack of generic principles is perceived by some as a significant impediment to routine use of IS. In the following some of the most useful proposals which have been offered in the late twentieth-century literature are briefly presented.

While Tierney and Kadane (1986) do not discuss IS *per se*, they propose an extension of the Laplace method of integration, as discussed, e.g., in De Bruijn (1958), to construct approximations for Bayesian posterior moments and marginal densities. The very same idea can be used to construct an importance sampler to approximate numerically the exact integrals. See Richard and Zhang (1998) for an application of this principle to a logit model with unobserved heterogeneity. Since, however, the Laplace method essentially serves to construct a local approximation around the mode of $\phi(x)$, it remains important to pay special attention to tail behavior in order to control for $\sigma^2(a)$, as discussed above.

Geweke (1989) pays attention to the minimization of the MC sampling variance $\sigma^2(a)$ within specific families of IS densities, typically multivariate Student-t densities and skewed generalizations thereof, labeled split-t densities. A problem with his approach, which is characteristic of any attempt at minimizing $\sigma^2(a)$, is that the latter depends upon integrals which have themselves to be numerically evaluated (a problem which will be addressed in Sect. 3 below). In the same paper Geweke also discusses more heuristic procedures for tailoring $m(x \mid a)$ in ways which proved quite successful in some Bayesian applications. It should also be noted that Geweke's article includes numerous references relative to MC simulation and numerical integration.

Evans (1991) develops adaptive methods which rely upon earlier draws of x to identify large values of the weight function defined in formula (5) and accordingly to modify the importance sampler. While such adaptive methods can occasionally prove useful, they do suffer from the problem explained earlier, whereby values of x for which the weight function $\omega(x; a)$ is large are typically highly unlikely to be drawn.

Owen and Zhou (2000) discuss improvements of IS technique which are shown to work quite well in low-dimensional applications. One proposal consists of combining IS with the method of 'control variates' to produce an upper bound for the MC sampling variance. (The method of control variates uses knowledge of approximate integrals to I to reduce the MC sampling variance of $\hat{I}_S(a)$. See Hammersley and Handscomb (1964) or Hendry (1984) for detailed

descriptions of the control variate method.) Another extension consists of pooling several IS estimators in order to reduce MC sampling variances.

Durbin and Koopman (1997) apply IS to evaluate the likelihood function of non-Gaussian state space models. Their application is particularly interesting in that it consists of a fairly sophisticated and, in a sense, indirect application of IS in dimensions which are significantly larger than those of the applications discussed above. Let $p(y \mid \lambda, \theta)$ denote a non-Gaussian density for a vector y of observables, conditional on a vector λ of unobserved variables and a vector θ of unknown parameters. Let $v(\lambda \mid \theta)$ denote the density of λ given θ. Finally, let $g(y \mid \lambda, \theta)$ denote a Gaussian approximation to $p(y \mid \lambda, \theta)$. Under g one can use standard Kalman filter techniques, as discussed, e.g., in Harvey (1990), to compute a likelihood function $L_g(\theta)$ and, in parallel, a density $v_g(\lambda \mid y, \theta)$ for the latent variables. It is then trivial to show that the actual likelihood function

$$L_p(\theta) = \int p(y \mid \lambda, \theta)\, v(\lambda \mid \theta)\, \mathrm{d}\lambda \tag{8}$$

can be rewritten as

$$L_p(\theta) = L_g(\theta) \int \frac{p(y \mid \lambda, \theta)}{g(y \mid \lambda, \theta)}\, v_g(\lambda \mid y, \theta)\, \mathrm{d}\lambda \tag{9}$$

The advantage of formula (9) over formula (8) is that it relies upon numerical integration to evaluate departures from the Gaussian likelihood L_g, rather than the likelihood itself. The density v_g in formula (9) plays exactly the role of an importance sampler. Durbin and Koopman (1997) then discuss the construction of the approximating Gaussian model in order to secure maximal numerical accuracy. The interest of that application is to show that the selection of an importance sampler can be approached indirectly via the construction of an operational approximation to a complex model.

Another sophisticated implementation of IS is found in Madras and Piccioni (1999) where the authors propose to provide greater flexibility in the choice of the class of samplers M, often a sticky problem in applications of IS, by allowing $m(x \mid a)$ to be the (implicit) equilibrium distribution associated with a Monte Carlo Markov chain (MCMC) simulator (see *Monte Carlo Methods and Bayesian Computation: Overview*; *Markov Chain Monte Carlo Methods*). The main advantage of this extension lies in the fact, discussed elsewhere, that MCMC simulators are very flexible. Its potential drawback is that its convergence properties are typically more difficult to assess than those of a conventional importance sampler.

An approach which also combines IS and MCMC is used by DiMatteo and Kadane (2001). They employ a standard MCMC technique to compute posterior moments for a given prior. Then, in order to assess robustness relative to the prior, they propose to rerun the analysis under different priors using the initial MCMC sampler as importance sampler.

An important message which emerges from this discussion is that efficient applications of IS have to be carefully tailored to the problem under consideration. This has proved to be a significant obstacle to routine applications of IS and is especially true for Bayesian applications which can produce notoriously 'ill-behaved' and, therefore, difficult to approximate posterior densities. Actually, Bayesian statisticians have largely abandoned IS in favor of MCMC. It should be noted, however, that while MCMC can be more flexible than IS, its convergence is not easily assessed in practice and may require large numbers of MC draws. In fact, both methods can occasionally produce spectacular failures if not carefully designed and controlled.

3. *Efficient Importance Sampling (EIS)*

Expanding upon an earlier contribution by Danielsson and Richard (1993), Richard and Zhang (1998) have recently proposed a generic principle for constructing efficient importance samplers in cases where ϕ is a positive function. For ease of presentation, the method will be introduced here without reference to dimensionality. A straightforward rearrangement of terms in formula (7) produces the following expression:

$$\sigma^2(a) = \int h[d^2(x; a)]\, \phi(x)\, \mathrm{d}x \tag{10}$$

where

$$d(x; a) = \ln \left[\frac{\phi(x)}{m(x \mid a) I} \right] \tag{11}$$

$$h(c) = e^{\sqrt{c}} + e^{-\sqrt{c}} - 2 \tag{12}$$

Note that h is monotone convex on R_+ and that $h(c) \geqslant c$. Minimization of $\sigma^2(a)$ with respect to a requires nonlinear optimization. Consider, therefore, the simpler function

$$q(a) = \int [\ln \phi(x) - c - \ln m(x \mid a)]^2 \phi(x)\, \mathrm{d}x \tag{13}$$

where c stands for an intercept in place of the unknown $\ln I$. Next, one selects an initial sampler, say $m(x \mid a_0)$—it could be $\mu(x)$ in formula (2), or any other sampler based, for example on a local approximation of $\phi(x)$ of the sort used by Tierney and Kadane (1986). An MC approximation to $q(a)$ is then given by

$$\hat{q}_R(a) = \frac{1}{R} \sum_{i=1}^{R} [\ln \phi(x) - c - \ln m(x \mid a)]^2 \omega_0(x) \Big|_{x = \tilde{x}_i^0} \tag{14}$$

where

$$\omega_0(x) = \frac{\phi(x)}{m(x \mid a_0)} \tag{15}$$

and $\{\tilde{x}_i^0;\ i{:}1 \to R\}$ denotes a set of R i.i.d. draws from $m(x \mid a_0)$. If, furthermore, $m(x \mid a)$ belongs to the exponential family of distributions, then minimization of $\hat{q}_R(a)$ with respect to a takes the form of a simple weighted linear least squares problem. It may be advisable to omit the weight function $\omega_0(x)$ if it exhibits excessive variance in cases where $m(x \mid a_0)$ is found to provide a poor global approximation to $\phi(x)$. The method can be iterated by using as initial sampler in any given round the optimized sampler from the previous round. Experience suggests that no more than 3 to 4 iterations are required to produce a stable solution, irrespective of the initial sampler $m(x \mid a_0)$. In order to assess how much efficiency is lost by minimizing $q(a)$ instead of $\sigma^2(a)$, let $\hat{a}$ and a_* denote the minimizers of q and σ^2, respectively. Taking advantage of the convexity of h, Jensen's inequality implies that

$$\sigma^2(\hat{a}) \geqslant \sigma^2(a_*) \geqslant h[q(a_*)] \geqslant h[q(\hat{a})] \tag{16}$$

MC estimates of the two bounds are immediate byproducts of the minimization of $\hat{q}_R(a)$. Unless they were found to be far apart, there would be no reasons for solving the more complex minimization problem associated with $\sigma^2(a)$.

The method is extremely accurate and fast in all cases where the integrand ϕ can be well approximated by a sampler from the exponential family of distributions. The real strength of the method lies in very high-dimensional (1000^+) integration of 'well-behaved' integrands. High-dimensional minimization of $\hat{q}_R(a)$ does require an additional factorization which is not discussed here. The final algorithm is one whereby minimization is recursively carried out one dimension at a time. It has proved useful for likelihood evaluation in dynamic latent variable models such as the one discussed earlier in Durbin and Koopman (1997). Liesenfeld and Richard (2000) successfully apply EIS to estimate parametric as well as semi-nonparametric dynamic stochastic volatility models for three daily financial series (IBM stock prices, S&P 500 stock index, and Dollar/Deutsche Mark exchange rate). Dimensions of integration are of the order of 4700. In each case highly accurate estimates of the likelihood function are obtained with as few as $R = S = 50$ MC draws and each individual likelihood evaluation requires on the order of 8 seconds on a Pentium II 466 MHz notebook.

Such outstanding performance originates from the fact that these high-dimensional integrands turn out to be well-behaved, reflecting the fact that the data happen to be quite informative on the unobserved volatilities. A pilot application to a logit panel model with random effects along both dimensions had proved equally successful (see Richard and Zhang 1998 for details).

4. *Smooth Functional Integration*

In numerous applications, such as likelihood function evaluation, the integral to be evaluated is itself a function of some argument δ, say

$$I(\delta) = \int_{D(\delta)} \phi(x;\delta)\,\mathrm{d}x, \quad \delta \in \Delta \tag{17}$$

The argument δ is then carried along in all the formulae previously discussed. Most importantly, one cannot expect a single sampler to be efficient for all δ's. Hence, IS requires the selection of a particular a, say $a(\delta)$, for each relevant value of δ. If, furthermore, $I(\delta)$ has to be optimized with respect to δ, then smoothness of the functional approximation $\hat{I}_S(\delta)$ becomes critical and cannot be secured with independent pointwise MC estimation, however accurate.

In such cases, it becomes essential that all IS sequences $\{\tilde{x}_i(a(\delta));\ i{:}1 \to S\}$ be obtained by transformation of a common sequence of random draws $\{u_i;\ i{:}1 \to S\}$ from a distribution which does not depend on a, say

$$\tilde{x}_i(a) = H(u_i; a), \quad i{:}1 \to S, \quad \forall a \in A \tag{18}$$

The u_i's, the so-called common random numbers, would typically be uniforms (in which case H denotes the inverse of the distribution function of x given a) or standardized normals (when x given a is itself normally distributed). See, e.g., Devroye (1986) for detailed descriptions of such transformations. Note, in particular, that inversion of a distribution function can be a computer-intensive operation but, nevertheless, is indispensable for smoothness. (The use of carefully designed interpolation techniques can significantly reduce computing times.) It should be noted that smoothness is not only required for numerical convergence of an optimization algorithm but also plays a central role in the derivation of the asymptotic properties of a broad range of simulation estimators. See, e.g., Pakes and Pollard (1989) or Gouriéroux and Monfort (1996).

5. *Conclusion*

IS is one of the key techniques currently available for the evaluation of analytically intractable integrals by MC simulation. It performs at its best when used in combination with 'acceleration' techniques such as,

e.g., control variates, mixtures, common random numbers. Very high-dimensional (1000^+) integration is feasible as long as the integrand is reasonably well-behaved. Further developments are required to guide the design of efficient IS samplers. IS is not an all-purpose technique and lack of flexibility in the construction of samplers may complicate its application to ill-behaved integrands, for which alternative methods of integration such as MCMC might prove more useful. Nevertheless, IS has proved to be a remarkably useful technique for a wide range of applications including very high-dimensional integrations.

Bibliography

Bratley P, Fox B L, Schrage L E 1987 *A Guide to Simulation*, 2nd edn. Springer Verlag, New York

Danielsson J, Richard J F 1993 Accelerated Gaussian importance sampler with applications to dynamic latent variable models. *Journal of Applied Econometrics* **8**: s153–73

De Bruijn N G 1958 *Asymptotic Methods in Analysis*. North Holland, Amsterdam

Devroye L 1986 *Non-Uniform Random Variate Generation*. Springer Verlag, New York

DiMatteo I, Kadane J 2001 Vote tampering in a District Justice election in Beaver County, Pennsylvania. *Journal of the American Statistical Association* **96**: 510–18

Durbin J, Koopman S J 1997 Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* **84**: 669–84

Evans M 1991 Adaptive importance sampling and chaining. *Contemporary Mathematics (Statistical Multiple Integration)* **115**: 137–42

Geweke J 1989 Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**: 1317–39

Geweke J 1996 Monte Carlo simulation and numerical integration. In: Amman H, Kendrick D, Rust J (eds.) *The Handbook of Computational Economics*. North Holland, Amsterdam, Vol. 1, pp. 731–800

Golub G H, Welsch J H 1969 Calculation of Gaussian quadrature rules. *Mathematics of Computation* **23**: 221–30

Gouriéroux C, Monfort A 1996 *Simulation Based Econometric Methods*. CORE Lecture series, Oxford University Press, Oxford, UK

Hammersley J M, Handscomb D C 1964 *Monte Carlo Methods*. Methuen, London

Harvey A C 1990 *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, UK

Hendry D F 1984 Monte Carlo experimentation in econometrics. In: Griliches Z, Intriligator M D (eds.) *The Handbook of Econometrics*. North Holland, Amsterdam

Hesterberg T 1995 Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**: 185–94

Kahn H, Marshall A 1953 Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America* **1**: 263–78

Kloek T, van Dijk H K 1978 Bayesian estimates of equation system parameters: Application of integration by Monte Carlo. *Econometrica* **46**: 1–19

Liesenfeld R, Richard J-F 2000 Univariate and multivariate stochastic volatility models: Estimation and diagnostics. Mimeo, Eberhard-Karls-Universität, Tübingen, Germany

Madras N, Piccioni M 1999 Importance sampling for families of distributions. *The Annals of Applied Probability* **9**: 1202–25

Owen A, Zhou Y 2000 Safe and effective importance sampling. *Journal of the American Statistical Association* **95**: 135–43

Pakes A, Pollard D 1989 Simulation and the asymptotics of optimization estimators. *Econometrica* **57**: 1027–57

Richard J-F, Zhang W 1998 Efficient high-dimensional Monte Carlo importance sampling. Mimeo, University of Pittsburgh, PA

Richard J-F, Zhang W 2000 Accelerated Monte Carlo integration: An application to dynamic latent variables models. In: Mariano R, Schuermann T, Weeks M (eds.) *Simulation-Based Inference in Econometrics: Methods and Applications*. Cambridge University Press, Cambridge, UK

Stern S 1997 Simulation-based estimation. *Journal of Economic Literature* **35**: 2006–39

Stroud A H, Secrest D H 1966 *Gaussian Quadrature Formulae*. Prentice-Hall, Englewood Cliffs, NJ

Tierney L, Kadane J B 1986 Accurate approximations for posterior moments and marginal densities. *Journal of American Statistical Association* **81**: 82–6

Trotter H F, Tukey J W 1956 Conditional Monte Carlo for normal samples. In: Mayer H A (ed.) *Symposium on Monte Carlo Methods*. Wiley, New York

R. Liesenfeld and J.-F. Richard

Monte Carlo Methods and Bayesian Computation: Overview

1. *Introduction*

Monte Carlo simulation methods and, in particular, Markov chain Monte Carlo methods, play a large and prominent role in the practice of Bayesian statistics, where these methods are used to summarize the posterior distributions that arise in the context of the Bayesian prior–posterior analysis. Monte Carlo methods are used in practically all aspects of Bayesian inference, for example, in the context of prediction problems and in the computation of quantities, such as the marginal likelihood, that are used for comparing competing Bayesian models.

Suppose that $\pi(\psi \mid y, \mathscr{M}) \propto p(\psi \mid \mathscr{M}) f(y \mid \psi, \mathscr{M})$ is the posterior density for a set of parameters $\psi \in \mathscr{R}^d$ in a particular Bayesian model $\mathscr{M}$ defined by the prior density $p(\psi \mid \mathscr{M})$ and sampling density or likelihood function $f(y \mid \psi, \mathscr{M})$ and that interest centers on the posterior mean $\eta = \int_{\mathscr{R}^d} \psi \pi(\psi \mid y, \mathscr{M})\, \mathrm{d}\psi$. As an illustration of this general setting, consider binary data $y_i \in \{0, 1\}$ modeled by the probabilities $\Pr(y_i = 1 \mid \psi) = \Phi(x_i' \psi)$, where Φ is the standard normal cumulative distribution function, x_i are covariates and ψ is an

unknown vector of parameters. The posterior distribution of ψ given n independent measurement $y = (y_1,\ldots,y_n)$ is proportional to

$$\pi(\psi \mid y, \mathscr{M}) \propto p(\psi \mid \mathscr{M}) \prod_{i=1}^{n} \Phi(x_i'\psi)^{y_i}(1-\Phi(x_i'\psi))^{(1-y_i)}$$

and the goal is to summarize this density.

Now suppose that a sequence of variates $\{\psi^{(1)},\ldots,\psi^{(M)}\}$ drawn from $\pi(\psi \mid y, \mathscr{M})$ are available and one computes the sample average of these draws. Under any set of conditions that allow use of the strong law of large numbers, it follows that the sample average is simulation consistent, converging to the posterior mean as the simulation sample M goes to infinity. This then is the essence of the Monte Carlo method: to compute an expectation, or to summarize a given density, sample the density by any method that is appropriate and use the sampled draws to estimate the expectation and to summarize the density.

The sampling of the posterior distribution is, therefore, the central focus of Bayesian computation. One important breakthrough in the use of simulation methods was the realization that the sampled draws need not be independent, that simulation consistency can be achieved with correlated draws. The fact that the sampled variates can be correlated is of immense practical and theoretical importance and is the defining characteristic of Markov chain Monte Carlo methods, popularly referred to by the acronym MCMC, where the sampled draws form a Markov chain (see also *Markov Chain Monte Carlo Methods*). The idea behind these methods is simple and extremely general. In order to sample a given probability distribution, referred to as the target distribution, a suitable Markov chain is constructed with the property that its limiting, invariant distribution is the target distribution. Once the Markov chain has been constructed, a sample of draws from the target distribution is obtained by simulating the Markov chain a large number of times and recording its values. Within the Bayesian framework, where both parameters and data are treated as random variables, and inferences about the parameters are conducted conditioned on the data, the posterior distribution of the parameters provides a natural target for MCMC methods.

Markov chain sampling methods originate with the work of Metropolis et al. (1953) in statistical physics. A crucial and landmark extension of the method was made by Hastings (1970) leading to a method that is now called the Metropolis–Hastings algorithm. This algorithm was first applied to problems in spatial statistics and image analysis (Besag 1974). A resurgence of interest in MCMC methods started with the papers of D. Geman and S. Geman (1984) who developed an algorithm, a special case of the Metropolis method that later came to be called the Gibbs sampler, to sample a discrete distribution, Tanner and Wong (1987) who proposed a MCMC scheme involving 'data augmentation' to sample posterior distributions in missing data problems, and Gelfand and Smith (1990) where the value of the Gibbs sampler was demonstrated for general Bayesian inference with continuous parameter spaces.

2. Markov Chain Monte Carlo Methods

2.1 The Metropolis–Hastings Algorithm

This powerful algorithm provides a general approach for producing a correlated sequence of draws from the target density $\pi(\psi)$ that may be difficult to sample by a classical independence method. As before, let $\pi(\psi)$ denote the density that one wishes to sample, where one can assume for simplicity that the density is absolutely continuous; and let $q(\psi, \psi')$ denote a source density for a candidate draw ψ' given the current value ψ in the sampled sequence. The density $q(\psi, \psi')$ is referred to as the proposal or candidate-generating density. Then, the M–H algorithm is defined by two steps: a first step in which a proposal value is drawn from the candidate-generating density and a second step in which the proposal value is accepted or rejected according to a probability given by

$$\alpha(\psi, \psi') = \begin{cases} \min\left[\dfrac{\pi(\psi')q(\psi', \psi)}{\pi(\psi)q(\psi, \psi')}, 1\right] & \text{if } \pi(\psi)q(\psi, \psi') > 0 \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

If the proposal value is rejected, then the next sampled value is taken to be the current value; which means that when a rejection occurs the current value is repeated. Given the new value, the same two-step process is repeated and the cycle iterated a large number of times. A full expository discussion of this algorithm, along with a derivation of the method from the logic of reversibility, is provided by Chib and Greenberg (1995).

The M–H method delivers variates from π under quite general conditions. A weak requirement for a law of large numbers for sample averages based on the M–H output involve positivity and continuity of $q(\psi, \psi')$ for (ψ, ψ'), and connectedness of the support of the target distribution. In addition, if π is bounded then conditions for ergodicity, required to establish the central limit theorem, are satisfied. A discussion of the conditions and the underlying Markov chain theory that justifies these methods was first provided by Tierney (1994) and summarized recently by Robert and Casella (1999). Of course, the variates are from π only in the limit as the number of iterations becomes

large but, in practice, after an initial burn-in phase consisting of (say) n_0 iterations, the chain is assumed to have converged and subsequent values are taken as approximate draws from π. Theoretical calculation of the burn-in is not easy, so it is important that the proposal density be chosen to ensure that the chain makes large moves through the support of the target distribution without staying at one place for many iterations. Generally, the empirical behavior of the M–H output is monitored by the autocorrelation time of each component of ψ defined as

$$\left\{ 1+2\sum_{s=1}^{\mathscr{M}} \left(1-\frac{s}{\mathscr{M}} \right) \rho_{ks} \right\} \tag{2}$$

where ρ_{ks} is the sample autocorrelation at lag s for the kth component of ψ, and by the acceptance rate which is the proportion of times a move is made as the sampling proceeds. Because independence sampling produces an autocorrelation time that is theoretically equal to one, and Markov chain sampling produces autocorrelation times that are bigger than one, it is desirable to tune the M–H algorithm to get values close to one, if possible.

Given the form of the acceptance probability $\alpha(\psi, \psi')$ it is clear that the M–H algorithm does not require knowledge of the normalizing constant of $\pi(\cdot)$. Furthermore, if the proposal density satisfies the symmetry condition $q(\psi, \psi') = q(\psi', \psi)$, the acceptance probability reduces to $\pi(\psi')/\pi(\psi)$; hence, if $\pi(\psi') \geqslant \pi(\psi)$, the chain moves to ψ', otherwise it moves to ψ' with probability given by $\pi(\psi')/\pi(\psi)$. The latter is the algorithm originally proposed by Metropolis et al. (1953).

Different proposal densities give rise to specific versions of the M–H algorithm, each with the correct invariant distribution π. One family of candidate-generating densities is given by $q(\psi, \psi') = q(\psi' - \psi)$. The candidate ψ' is thus drawn according to the process $\psi' = \psi + z$, where z follows the distribution q, and is referred to as the *random walk M–H* chain. The random walk M–H chain is perhaps the simplest version of the M–H algorithm and is quite popular in applications. One has to be careful, however, in setting the variance of z because if it is too large it is possible that the chain may remain stuck at a particular value for many iterations while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution.

Hastings (1970) considers a second family of candidate-generating densities that are given by the form $q(\psi, \psi') = q(y)$. Tierney (1994) refers to this as an *independence M–H chain* because the candidates are drawn independently of the current location ψ. Chib and Greenberg (1994, 1995) implement such chains by matching the proposal density to the target at the mode by a multivariate normal or multivariate-t distribution with location given by the mode of the target and the dispersion given by inverse of the Hessian evaluated at the mode.

2.2 The Multiple-block M–H Algorithm

In applications when the dimension of ψ is large it may be preferable, as discussed by Hastings (1970), to construct the Markov chain simulation by first grouping the variables ψ into smaller blocks. For simplicity suppose that two blocks are adequate and that ψ is written as (ψ_1, ψ_2), with $\psi_k \in \Omega_k \subseteq \mathscr{R}^{d_k}$. In that case the M–H algorithm requires the specification of two proposal densities,

$$q_1(\psi_1, \psi_1' \mid \psi_2); \quad q_2(\psi_2, \psi_2' \mid \psi_1)$$

one for each block ψ_k, where the proposal density q_k may depend on the current value of the remaining block. Also, define

$$\alpha(\psi_1, \psi_1' \mid \psi_2) = \min \left\{ \frac{\pi(\psi_1', \psi_2) q_1(\psi_1', \psi_1 \mid \psi_2)}{\pi(\psi_1, \psi_2) q_1(\psi_1, \psi_1' \mid \psi_2)}, 1 \right\} \tag{3}$$

and

$$\alpha(\psi_2, \psi_2' \mid \psi_1) = \min \left\{ \frac{\pi(\psi_1, \psi_2') q_2(\psi_2', \psi_2 \mid \psi_1)}{\pi(\psi_1, \psi_2) q_2(\psi_2, \psi_2' \mid \psi_1)}, 1 \right\} \tag{4}$$

as the probability of move for block ψ_k conditioned on the other block. Then, in what may be called the multiple-block M–H algorithm, one cycle of the algorithm is completed by updating each block, say sequentially in fixed order, using a M–H step with the above probability of move, given the most current value of the other block.

The extension of this method to more than two blocks is straightforward.

2.3 The Gibbs Sampling Algorithm

Another MCMC method, which is a special case of the multiple-block M–H method, is called the Gibbs sampling method and was brought into statistical prominence by Gelfand and Smith (1990). An elementary introduction to Gibbs sampling is provided by Casella and George (1992). To describe this algorithm, suppose that the parameters are grouped into two blocks (ψ_1, ψ_2) and each block is sampled according to the *full conditional distribution* of block ψ_k, defined as the conditional distribution under π of ψ_k given the other block. In parallel with the multiple-block M–H algorithm, the most current value of the

other block is used in deriving the full conditional distribution. Derivation of these full conditional distributions is usually quite simple since, by Bayes' theorem, each full conditional distribution $\pi(\psi_1 | \psi_2)$ and $\pi(\psi_2 | \psi_1)$ are both proportional to $\pi(\psi_1, \psi_2)$, the joint distribution of the two blocks. In addition, the powerful device of data augmentation, due to Tanner and Wong (1987), in which latent or auxiliary variables are artificially introduced into the sampling, is often used to simplify the derivation and sampling of the full conditional distributions. For further details see *Markov Chain Monte Carlo Methods.*

3. Implementation Issues

In implementing a MCMC method it is important to assess the performance of the sampling algorithm to determine the rate of mixing and the size of the burn-in (both having implications for the number of iterations required to get reliable answers). A large literature has now emerged on these issues, for example Tanner (1996, Sect. 6.3), Cowles and Carlin (1996), Gamerman (1997, Sect. 5.4) and Robert and Casella (1999).

Writing at the turn of the millenium, the more popular approach is to utilize the sampled draws to assess both the performance of the algorithm and its approach to the stationary, invariant distribution. Several such relatively informal methods are now available. Perhaps the simplest and most direct are autocorrelation plots (and autocorrelation times) of the sampled output. Slowly decaying correlations indicate problems with the mixing of the chain. It is also useful in connection with M–H Markov chains to monitor the acceptance rate of the proposal values with low rates implying 'stickiness' in the sampled values and thus a slower approach to the invariant distribution.

Somewhat more formal sample-based diagnostics are also available in the literature, as summarized in the routines provided by Best et al. (1995) and now packaged in the BOA program. Although these diagnostics often go under the name 'convergence diagnostics' they are in principle approaches that detect *lack* of convergence. Detection of convergence based entirely on the sampled output, without analysis of the target distribution, is extremely difficult and perhaps impossible. Cowles and Carlin (1996) discuss and evaluate thirteen such diagnostics (for example, those proposed by Geweke (1992), Gelman and Rubin (1992), Zellner and Min (1995), amongst others) without arriving at a consensus.

While implementing MCMC methods it is important to construct samplers that mix well (mixing being measured by the autocorrelation times) because such samplers can be expected to converge more quickly to the invariant distribution. As a general rule, sets of parameters that are highly correlated should be treated as one block when applying the multiple-block M–H algorithm. Otherwise, it would be difficult to develop proposal densities that lead to large moves through the support of the target distribution and the sampled draws would tend to display autocorrelations that decay slowly. Liu (1994) and Liu et al. (1994) discuss possible strategies in the context of three-block Gibbs MCMC chains. Roberts and Sahu (1997) provide further discussion of the role of blocking in the context of Gibbs Markov chains.

4. Additional Uses of MCMC Methods in Bayesian Inference

MCMC methods provide an elegant solution to the problem of prediction in the Bayesian context. The goal is to summarize the Bayesian predictive density defined as

$$f(y_f | y, \mathscr{M}) = \int f(y_f | y, \mathscr{M}, \theta)\pi(\theta | y, \mathscr{M})\, \mathrm{d}\theta, \quad (5)$$

where $f(y_f | y, \mathscr{M}, \theta)$ is the conditional density of future observations y_f given $(y, \mathscr{M}, \theta)$. In general, the predictive density is not available in closed form. However, this is not a problem when a sample of draws from the posterior density is available. If one simulates

$$y_f^{(i)} \sim f(y_f | y, \mathscr{M}, \theta^{(j)}) \quad (6)$$

for each sampled draw $\theta^{(j)}$, then the collection of simulated values $\{y_f^{(1)}, \ldots, y_f^{(M)}\}$ is a sample from the Bayes prediction density $f(y_f | y, \mathscr{M})$. The simulated sample can be summarized in the usual way by the computation of sample averages and quantiles.

MCMC methods have also been widely applied to the problem of the model choice. Suppose that there are K possible models $\mathscr{M}_1, \ldots, \mathscr{M}_K$ for the observed data defined by the sampling densities $\{f(y | \theta_k, \mathscr{M}_k)\}$ and proper prior densities $\{p(\theta_k | \mathscr{M}_k)\}$ and the objective is to find the evidence in the data for the different models. In the Bayesian approach this question is answered by placing prior probabilities $\Pr(\mathscr{M}_k)$ on each of the K models and using the Bayes calculus to find the posterior probabilities $\{\Pr(\mathscr{M}_1 | y), \ldots, \Pr(\mathscr{M}_K | y)\}$ conditioned on the data but marginalized over the unknowns θ_k (Jeffreys 1961 and Kass and Raftery 1995). Specifically, the posterior probability of $\mathscr{M}_k$ is given by the expression

$$\begin{aligned} \Pr(\mathscr{M}_k | y) &= \frac{\Pr(\mathscr{M}_k) m(y | \mathscr{M}_k)}{\sum_{i=1}^{K} \Pr(\mathscr{M}_l) m(y | \mathscr{M}_l)} \\ &\propto \Pr(\mathscr{M}_k) m(y | \mathscr{M}_k), \\ &\quad (k \leqslant K) \end{aligned} \quad (7)$$

where $m(y \mid \mathscr{M}_k)$ is the marginal likelihood of $\mathscr{M}_k$.

A central problem in estimating the marginal likelihood is that it is an integral of the sampling density over the prior distribution of θ_k. Thus, MCMC methods, which deliver sample values from the posterior density, cannot be used to directly average the sampling density. In addition, taking draws from the prior density to do the averaging produces an estimate that is simulation consistent but highly inefficient because draws from the prior density are not likely to be in high density regions of the sampling density $f(y \mid \theta_k, \mathscr{M}_k)$. A natural way to correct this problem is by the method of importance sampling. If we let $h(\theta_k \mid \mathscr{M}_k)$ denote a suitable importance sampling function, then the marginal likelihood can be estimated as

$$\hat{m}_I(y \mid \mathscr{M}_k) = M^{-1} \sum_{j=1}^{M} \frac{f(y \mid \theta_k^{(j)}, \mathscr{M}_k) p(\theta_k^{(j)} \mid \mathscr{M}_k)}{h(\theta_k^{(j)} \mid \mathscr{M}_k)}$$
$$\theta_k^{(j)} \sim h(\theta_k^{(j)} \mid \mathscr{M}_k) \quad (j \leqslant M) \tag{8}$$

This method is useful when it can be shown that the ratio is bounded, which can be difficult to check in practice, and when the sampling density is not expensive to compute; this, unfortunately, is often not so. Gelfand and Dey (1994) propose an estimator that is based on the posterior simulation output. Another method for finding the marginal likelihood is due to Chib (1995). The marginal likelihood, by virtue of being the normalizing constant of the posterior density, can be expressed as

$$m(y \mid \mathscr{M}_k) = \frac{f(y \mid \theta_k, \mathscr{M}_k) p(\theta_k \mid \mathscr{M}_k)}{\pi(\theta_k \mid y, \mathscr{M}_k)} \tag{9}$$

which is an identity in θ_k.Based on this expression an estimate of the marginal likelihood on the log-scale is given by

$$\log \hat{m}(y \mid \mathscr{M}_k) = \log f(y \mid \theta_k^*, \mathscr{M}_k) + \log p(\theta_k^* \mid \mathscr{M}_k)$$
$$- \log \hat{\pi}(\theta_k^* \mid y, \mathscr{M}_k) \tag{10}$$

where θ_k^* denotes an arbitrarily chosen point and $\hat{\pi}(\theta_k^* \mid y, \mathscr{M}_k)$ is the estimate of the posterior density at that single point. To estimate this posterior ordinate Chib (1995) utilizes the Gibbs output in conjunction with a decomposition of the ordinate into marginal and conditional components. Chib and Jeliazkov (2001) extend that approach for output produced by the Metropolis–Hastings algorithm. A comparison of different estimators of the marginal likelihood is provided by Chen and Shao (1997).

When one is presented with a large collection of candidate models, each with parameters $\theta_k \in B_k \subseteq R^{d_k}$, direct fitting of each model to find the marginal likelihood can be computationally expensive. In such cases it may be more fruitful to utilize model space–parameter space MCMC algorithms that eschew direct fitting of each model for an alternative simulation of a 'mega model' where a model index random variable, denoted as $\mathscr{M}$, taking values on the integers from 1 to K, is sampled in tandem with the parameters. The posterior distribution of $\mathscr{M}$ is then computed as the frequency of times each model is visited. Methods for doing this have been proposed by Carlin and Chib (1995) and Green (1995) which are discussed in article 2.1.76. Both methods are closely related as shown by Dellaportas et al. (1998) and Godsill (1998). The problem of variable selection in regression has also been tackled in this context starting with George and McCulloch (1993).

5. *An Example*

Markov chain Monte Carlo methods have proved enormously popular in Bayesian statistics (see for example, Gilks et al. 1996) where these methods have opened up vistas that were unimaginable fifteen years ago.

As an illustration of how MCMC methods are designed in a Bayesian context, we return to the binary data problem discussed in the introduction and describe the approach of Albert and Chib (1993) that has found many applications. Recall that the posterior distribution of interest is proportional to

$$\pi(\beta \mid y, \mathscr{M}) \propto p(\beta \mid \mathscr{M}) \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{(1-y_i)} \tag{11}$$

where $p_i = \Phi(x_i' \psi)$. To analyze this probit model, and other models for binary and ordinal data, Albert and Chib introduce latent variables and write $z_i = x_i' \psi + \varepsilon_i$ where ε_i is $\mathscr{N}(0, 1)$ and let $y_i = I[z_i > 0]$ where $I[\cdot]$ is the indicator function. This two-equation specification is equivalent to the binary probit regression model since $\Pr(y_i = 1 \mid \psi) = \Pr(z_i > 0 \mid \psi) = \Phi(x_i' \psi)$, as required. The Albert–Chib algorithm now consists of simulating the joint posterior distribution of (ψ, Z), where $Z = (z_1, \ldots, z_n)$. The MCMC simulation is implemented by sampling the conditional distributions $[\psi \mid y, Z]$ and $[Z \mid y, \psi]$. The first of these distributions depends only on Z (and not on y) and is obtained by standard Bayesian results on normal linear models. The second factors into a product of independent distributions $[z_i \mid y_i, \psi]$. The knowledge of y_i merely restricts the support of the normal distribution $z_i \mid \psi$; if $y_i = 0$, then $z_i \leqslant 0$ and if $y_i = 1$, then $z_i > 0$. Suppose that *a priori* $\psi \sim \mathscr{N}_k(\psi_0, C_0)$, where ψ_0 and C_0 are known hyperparameters. Then, a sample of ψ and Z from the posterior distribution is obtained by iterating on the following steps:

(a) sample ψ from the distribution $\mathscr{N}_k(\hat{\psi}, C)$, where $\hat{\psi} = C(C_0^{-1}\psi_0 + \sum_{i=1}^{n} x_i z_i)$, $C = (B_0^{-1} + \sum_{i=1}^{n} x_i x_i')^{-1}$;

(b) sample z_i from $\mathscr{T}\mathscr{N}_{(-\infty,0)}(x_i' \psi, 1)$ if $y_i = 0$ and from $\mathscr{T}\mathscr{N}_{(0,\infty)}(x_i' \psi, 1)$ if $y_i = 1$, where $\mathscr{T}\mathscr{N}_{(a,b)}(\mu, \sigma^2)$ denotes the $\mathscr{N}(\mu, \sigma^2)$ distribution truncated to the region (a, b).

At this time, considerable work on MCMC methods and applications continues to emerge, a testament to the importance and viability of MCMC methods in Bayesian analysis.

See also: Bayesian Statistics; Markov Chain Monte Carlo Methods; Markov Models and Social Analysis

Bibliography

Albert J H, Chib S 1993 Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**: 669–79

Besag J 1974 Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, B Met **36**: 192–236

Best N G, Cowles M K, Vines S K 1995 CODA: Convergence diagnostics and output analysis software for Gibbs sampling. *Technical report*, Cambridge MRC Biostatistics Unit

Carlin B P, Chib S 1995 Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society,* B Met **57**: 473–84

Casella G, George E I 1992 Explaining the Gibbs sampler. *American Statistician* **46**: 167–74

Chen M-H, Shao Q-M 1997 On Monte Carlo methods for estimating ratios of normalizing constants. *Annals of Statistics* **25**: 1563–94

Chib S 1995 Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**: 1313–21

Chib S, Greenberg E 1994 Bayes inference in regression models with ARMA(p, q) errors. *Journal of Econometrics* **64**: 183–206

Chib S, Greenberg E 1995 Understanding the Metropolis–Hastings algorithm. *American Statistician* **49**: 327–35

Chib S, Jeliazkov I 2001 Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* **96**(453): 276–81

Cowles M K, Carlin B 1996 Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* **91**: 883–904

Dellaportas P, Forster J J, Ntzoufras I 1998 On Bayesian model and variable selection using MCMC. *Technical report*, University of Economics and Business, Greece

Gamerman D 1997 *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, London

Gelfand A E, Dey D (1994) Bayesian Model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society* **B56**: 501–14

Gelfand A E, Smith A F M 1990 Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409

Gelman A, Rubin D B 1992 Inference from iterative simulation using multiple sequence. *Statistical Science* **4**: 457–72

Geman D, Geman S 1984 Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**: 609–28

George E I, McCulloch R E 1993 Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**: 881–9

Geweke J 1992 Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo J M, Berger J O, Dawid A P, Smith A F M (eds.) *Bayesian Statistics*. New York: Oxford University Press, pp. 169–93

Gilks W R, Richardson S, Spiegelhalter D J 1996 *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London

Godsill S J 1998 On the relationship between model uncertainty methods. *Technical report*, Signal Processing Group, Cambridge University

Green P J 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–32

Hastings W K 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109

Jeffreys H 1961 *Theory of Probability*, 3rd edn. Oxford University Press, New York

Kass R E, Raftery A E 1995 Bayes factors. *Journal of the American Statistical Association* **90**: 773–95

Liu J S 1994 The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* **89**: 958–66

Liu J S, Wong W H, Kong A 1994 Covariance structure of the Gibbs Sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**: 27–40

Mengersen K L, Tweedie R L 1996 Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics* **24**: 101–21

Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H, Teller E 1953 Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–92

Muller P, Erkanli A, West M 1996 Bayesian Curve fitting using multivariate normal mixtures. *Biometrika* **83**: 67–79

Robert C P, Casella G 1999 *Monte Carlo Statistical Methods*. Springer, New York

Roberts G O, Sahu S K 1997 Updating schemes, correlation structure, blocking, and parametization for the Gibbs sampler. *Journal of the Royal Statistical Society,* B Met **59**: 291–317

Tanner M A 1996 *Tools for Statistical Inference*,3rd edn. Springer, New York

Tanner M A, Wong H W 1987 The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**: 528–40

Tierney L 1994 Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**: 1701–62

Zellner A, Min C K 1995 Gibbs sampler convergence criteria. *Journal of the American Statistical Association* **90**: 921–7

S. Chib

Montesquieu, Charles, the Second Baron of (1689–1755)

1. Biography and Works

Montesquieu, generally held to be the 'chief theoretician of the separation of powers in western constitutional thought' (Stubbe-Da Luz 1998, p. 7), was born in La Brède, south of Bordeaux, on January 18, 1689 as Charles-Louis de Secondat. His family, which owned extensive vineyards, was part of the aristocratic

class, whose pride and independence were to have an influence on Montesquieu's work. His family was also influenced by their experiences of the religious wars which had left particularly traumatic scars in the south of France due to the spread of Protestantism in the area, and to which members of Montesquieu's family had also fallen victim.

After attending the Oratoraner College in Juilly near Paris (1700–5), well-known for its critical orientation, and studying law in Bordeaux and Paris (1705–9), the young Montesquieu inherited his father's post as a councilor of parliament of the province Guyenne in Bordeaux after his father's early death in 1713. In 1721 he inherited the title of Baron of Montesquieu from his uncle, and along with it the post of cabinet chairman (*Président à mortier*) of the *Parlement* (Court of Justice) in Bordeaux. Having been introduced into the Parisian salons in 1719, where he established contacts to the circle of enlightened intellectuals and writers of the time (for instance Fontenelle), he was elected as a member of the *Académie Française* in 1728 after his epistolary novel *Les Lettres Persanes* (1721) was published. His later elections as a member of the Royal Society in London and a corresponding member of the Berlin *Akademie der Wissenschaften* were due to the influence of his work across Europe as well as his personal contacts to intellectual circles in England and Germany.

Parallel to his literary and philosophical work, which took on an increasingly broad extent from the 1730s on, Montesquieu was also active as a judge, and from 1721 to 1725 as cabinet president, at the Parliament of Bordeaux as well as a landowner, supervising all aspects of wine production and finances entailed by his vineyards and property in *La Brède*. From 1728 to 1731 he undertook a three-year tour of Europe, visiting South Germany, Austria, the Netherlands, Italy, and England. His stay in England in particular, during the period of heated debate between the Whig prime minister Walpole and the conservative opposition around Lord Bolingbroke, who were in favor of the separation of powers, proved decisive for the further development of his political thought. His main work *L'Esprit des Loix* took almost 20 years to complete and was published in 1748. In spite of church and state censorship it soon achieved high circulation figures and was translated and circulated throughout Europe. In 1752 Montesquieu published a piece defending his work to the censorship authorities (*Réponses et Explications Données à la Faculté de Théologie*). He died in Paris on February 10, 1755 in the wake of an influenza epidemic.

There are three aspects of his biography and sociocultural background which appear to be of importance for his literary, political, and philosophical work: (a) his origins as a member of the French provincial aristocracy, which had sacrificed a great deal of its political power during the enforcement of absolutist power structures under Louis XIII and Louis XIV and their ministers Mazarin and Richelieu. The provincial parliaments were one of the last traces of the past power of the aristocracy in France, and in their function as high courts could raise a kind of suspense veto against royal laws. The political history of eighteenth century France was marked by a series of conflicts between the Parliaments and the absolutist monarchy, which Montesquieu was able to experience and follow for over 30 years as a member of the *Parliament de Bordeaux*; (b) as an aristocrat and landowner Montesquieu represented the 'revolutionary conservatism' of that fraction of the Enlightenment movement which was a part of the privileged aristocratic class and which attempted to defend and regain its traditional political and social rights, which had been greatly limited since the seventeenth century, against the absolutist state; and (c) Montesquieu's material view of governmental and social conditions, and especially the importance he attributed to climatic and soil conditions, were without doubt decisively influenced by his lifelong experience as a vineyard owner. His thought resulted in a certain sense from the 'philosophy of a vintner' (Stubbe-Da Luz 1998), who knew how important specific climatic and geographical conditions were for the quality and perfect maturity of the fine Bordeaux wines his family had been cultivating for many generations. For example the ice-age pebbles which reflect sunlight onto the chalk, clay, and sand ground of the area around *La Brède* are essential for the quality of the renowned Graves wine grown in this area, which is one of France's best wines 'Climate and soil also allowed,' according to Montesquieu, 'in a similar way the "love of freedom" and the "hate of violence" to develop in very varied intensities' (Stubbe-Da Luz 1998, p. 95).

2. *Thought and Discourse Structures*

Montesquieu's work is complex and covers varied disciplines and forms of discourse: literature, philosophy, aesthetic theory, political theory, and sociology. His epistolic novel *Les Lettres Persanes* (*Persian Letters* 1721) calls attention to the deplorable problems of contemporary French society through the literary medium and from a fictitious perspective, the viewpoint of the Persians Usbek and Rica, traveling through Louis XIV's France in the years 1711 to 1720. The despotic power structures of the French monarchy, the influence of the Church on the state and Court ceremony in particular appears incompatible with the principles of common sense ('*raison*') to the two Persians. Montesquieu's novel became a much-imitated form of societal criticism, which confronts western and non-European societal forms and points of view with one another, and simultaneously applies a radical criticism to the governmental and social structures of absolutist France on the basis of enlightened reason ('*raison*'). Montesquieu's *Lettres*

Persanes are still of relevance today as a matrix for genre and portrayal which non-European writers in particular have adapted productively. Among these writers are the Chinese immigrant to Canada Yin Chen, with her epistolic novel *Lettres Chinoises* (1993), Bernard Dadiéfrom the *Côte d'Ivoire*, who took up Montesquieu's narrative model in his autobiographical epistolic novel *Un Nègre à Paris* (1959) and made a critical reappraisal of his own experiences of a visit to the French metropolis from a non-European perspective, and the Montréal-based Franco-Canadian writer Lise Gauvin, in whose novel *Lettres d'une Autre* (1993) a Persian immigrant describes her Québec experiences using the social-critical and satirical viewpoint of Montesquieu.

His book *Le Temple de Gnide* (1725) is a rather ephemeral work including erotic fantasies. In contrast, his essay *Considérations sur les causes de la grandeur et la décadence des Romains* (1734) and his article *Esthéthique du Goût*, which he wrote for the *Encyclopédie* (1751, published 1756) of Diderot and d'Alembert, bear relation to his historical-philosophical and sociological reflections. The *Encyclopédie* article is a plea for a modern aesthetic that both orients itself by models from antiquity and attempts to go beyond these models. In the *Considérations* Roman history serves as illustrative material for his historical-philosophical reflections on the connection between national greatness and governmental form. The development of despotic regimes since Caesar and Augustus is seen as a fundamental cause of the fall of Roman civilization.

At the center of Montesquieu's ideas is a systematic political and sociological thought structure, which is best manifested in his main work *L'Esprit des Loix* (1748), on which he worked almost uninterruptedly from 1731 to 1747, following his European tour. This work can be divided into three main parts: the 'books' (*Livres*) I–XIII present a theory of the three basic forms of government monarchy, republic, and despotism, in topological form and with reference to Aristotle's *Politeia*; books XIV–XX examine the influence of material factors, particularly of climate and soil quality, on the structure of human societies and their traditions and institutions; finally in books XX–XXVIII the significance of social and economic factors such as trade and commerce, currency, population development, and religion for the traditions and laws of specific societies and their structures of government is analyzed. The final chapter (*Livre* XXIX) illustrates the pragmatic consequences of the consideration of material and social factors for the structure and development of two very different societies, using the examples of the laws of ancient Rome and medieval feudal society.

The sketched summary of contents outlines the basic aims and key terms of Montesquieu's theory of society. In contrast to Jean-Jacques Rousseau, who ruled out empirical and historical facts in his main philosophical work *Du Contrat Social* (1762) ('*Écartons tous les faits*,' 'let us ignore all the facts'), Montesquieu's work is based, as the large collection of notes and material he made when writing *L'Esprit des Loix* show, on an impressive amount of research and source materials on the various societies, which he critically reviewed and connected with his own observations and reflections. It has been proved that ideas from over 600 works can be found in his most important book. From the critical examination of extremely heterogeneous material on completely different societies, natural laws are derived concerning the interdependence of societal structure, form of government, and material and social factors. The republic, for example, is connected with a mild climate and a small territory, despotism is limited in the main to large territories, but simultaneously represents a form of decline of the monarchy. The three forms of governmental structure correspond to dominant values and mental reaction patterns, which influence both the behavior of the rulers and the thinking of the people: in the republic 'virtue' ('*vertu*'), in the monarchy 'honor' ('*honneur*') and in despotism 'fear' ('*crainte*'). Key terms of Montesquieu's thought are '*esprit général*,' '*loix*,' and '*modération*,' two of which appear in the subtitle of the work. Here Montesquieu explains that his book is about 'the connection that the laws must have to the constitution of every government, to the morals, the climate, the religion, the commerce, etc.' The basic term '*loix*' therefore refers both to state laws and to natural structural laws which are derived from the analysis of complex empirical sets of facts (such as climate, soil, religion, etc.) and can explain the specific forms of state legal measures. As '*esprit général*' Montesquieu understands the collective patterns of thought and behavior of a community, which is in it influenced by geographical and historical factors. Montesquieu's term '*esprit*' therefore roughly corresponds to the current anthropological definition of the term 'culture.' The term '*modération*' in turn refers to the normative dimension of Montesquieu's work. His political philosophy, despite his analytical distance, highly values those societies that seek a balance between the various interests of power and societal groups, and are able to create a structural equilibrium ('*équilibre*'). The Roman republic, with its characteristic balance of plebs and privileged classes, the early and high medieval franco-germanic feudal societies, and modern England represent, according to Montesquieu, exemplary models of political and social equilibrium and moderation, despite their differing structures of government and society. The model of the English constitution and the spirit ('*esprit*') behind it is seen to be based in turn essentially on the principle of the separation of powers, the division of executive, legislative, and judiciary, that Montesquieu took over principally from John Locke and developed into a central part of his political theory. Thus Montesquieu presented in *L'Esprit des Loix* not only an ontological

theory of laws and a normative theory of legislation, but also an empirical theory of forms, functioning and development of human societies as well as the political, social, and economical behavior of their members. The relationship of tension between normativity and analysis that is present throughout Montesquieu's *L'Esprit des Loix* points to fundamental contradictions and ambivalence of his work: (a) first the contradiction between a rational universalism on one hand, which is characteristic of the political philosophy of the Enlightenment and regards liberty, the separation of powers and the bicameral system as 'fundamental laws' ('*loix fondamentales*'), and a historical sense of the particular and specific on the other hand, which, according to the German historian Friedrich Meinecke, foreshadows the historicism of the nineteenth century; (b) second this calls attention to the contradiction between sociological and moral approaches, which can be seen for example in the analysis and judgment of the phenomenon of slavery. From a historical-sociological point of view Montesquieu explains slavery using a set of economic, demographic, and climatic factors based on the analysis of overseas plantation societies; from a moral viewpoint however, he condemns slavery as a manifestation of a despotic form of government and society. He sees it as diametrically opposed to the fundamental principle of moderation and the aim of every social and political order, to create a basic consensus and balance of interests between the various interests of power and social groups. According to Stubbe-Da Luz, we can here observe a 'relationship of tension' between Montesquieu the moralist 'who elevated certain moral norms to natural laws' and 'the sociological comparatist tending towards materialism and relativism' (1998, p. 91); (c) third, Louis Althusser referred to the fundamental contradiction between the 'innovative genius' ('*génie innovateur*') of Montesquieu and his 'reactionary opinions' ('*opinions réactionnaires*') in his book *Montesquieu et la Politique* (1959), which is also shown in the juxtaposition of analytical determinism and enlightened moral normativity.

The French political scientist and sociologist Raymond Aron interprets the contradictions and ambivalence of *L'Esprit des Loix* as a result of a profound transition period of European thought, at the center of which is Montesquieu's work, in his book *Les Etapes de la Pensée Sociologique* (1967), the first chapter of which is on Montesquieu. He sees Montesquieu both as one of the last 'classical philosophers' ('*philosophes classiques*,' p. 66) and 'the first sociologist' ('*le premier des sociologues*,' p. 66). His understanding of history, for which the terms rise, greatness, and decadence appear central, is still confined in the pre-enlightenment cyclical theory, while the teleological understanding of progress characteristic of Voltaire's work for example bears no systematic significance in Montesquieu's conception of history and society. His understanding of society and his both analytical and interconnective approach to social and political phenomena however, according to Aron, foreshadow the sociological method of Auguste Comte (who explicitly referred to Montesquieu and viewed him as a forerunner of modern sociology) and of Max Weber and stand for the 'modernity' of his theoretical approach.

3. *Influence and Reception*

Montesquieu was one of the writers and philosophers of the French Enlightenment era who were not only canonized and honored in their own lifetimes, but also experienced a broad international reception. A contributing factor was that both the *Lettres Persanes* and his main work *L'Esprit des Loix* were banned by the state and Church censorship and captured the attention of the enlightened readership for this very reason. D'Alembert, co-editor along with Diderot of the *Encyclopédie*, praised Montesquieu in the foreword (*Discours préliminaire*) of the first volume of the work, writing that his main work was 'denigrated by a few Frenchmen, greeted by the nation and admired by all Europe.' As early as the second half of the eighteenth century, Montesquieu gained posthumous praise that proclaimed him as one of the most outstanding intellectuals of the Enlightenment age. In 1755 Maupertuis of the Berlin *Akademie der Wissenschaften* held a laudatory speech on the deceased Montesquieu; in 1785 the academy of Bordeaux held a competition in honor of Montesquieu's work, in which the later revolutionary Jean-Paul Marat took part. In 1796, after the end of the reign of terror, the French national assembly voted to allot Montesquieu a place in the national temple of fame in the *Panthéon*, an honor otherwise only allowed to Voltaire and Rousseau among the philosophers of the Enlightenment movement during the French Revolution.

Three dimensions of Montesquieu's work have defined its influence and reception since the eighteenth century: first, the fascination that the narrative model of the *Lettres Persanes* held for his contemporaries and later generations, which is shown by the abundance of translations into all important languages and in numerous creative imitations; also the central element of his political theory, and finally the comparatistic method of his societal analysis, which makes him one of the forerunners of modern sociological thought.

The principle of the separation of powers and the theory of intermediary powers and intermediate bodies ('*pouvoirs intermédiares*') between the ruler and the people, which are the most important elements of a moderate monarchy (in contrast to a despotic rule), represent the most broadly quoted and adapted elements of Montesquieu's political philosophy. This philosophy played a particularly outstanding role in

the context of the American independence movement. He was praised as a 'compatriot' during the first continental congress in Philadelphia, and no other modern French book was as widely spread as Montesquieu's *L'Esprit des Loix* in eighteenth century America. Montesquieu's reception in North America was however based on an extremely selective reading of the *Esprit des Loix*, which was also much more widely commentated and criticized in eighteenth century France than it was really read. The *Esprit des Loix* offered itself in the context of the American Revolution as an 'authoritative recipe book and as an ''oracle''' (Stubbe-Da Luz 1998, p. 130). According to the American sociologist James Bryce, no other political principle had a greater influence on the fathers of the American constitution than the principle of the separation of powers advocated by Montesquieu.

Montesquieu's success in North America, particularly his influence on the American constitution of 1789, amplified his resonance in France, Germany, and other European countries. Along with Rousseau, Voltaire, Raynal, and Mably, Montesquieu represented one of the 'fathers of the French Revolution' in the understanding of the revolutionaries of 1789. Writings such as *De l'autorité de Montesquieu dans la Révolution présente* ('Of the authority of Montesquieu in the present revolution') by Grouvelle (1789) are characteristic of this fact. However Montesquieu's work quickly lost its aura and its influence on the discussion on the constitution after the overthrow of the monarchy in August 1792 and the disempowerment of the Girondists in July 1793. Under Jacobean rule he was labeled an 'aristocratic writer,' his grave was plundered, and his remains were scattered into mass graves in the catacombs of Paris in 1794. It was not until the *Directoire* (1795–9) and the constitutions of the Third and Fourth Republics that explicit reference was made to the principle of the separation of powers represented and propagated by Montesquieu.

The political discussion in modern day France makes reference to Montesquieu particularly with the intention of criticizing the state, criticizing forms of state influence on the judicial system, questioning the traditional double role of some politicians (for example as member of parliament and as mayor), and throwing a critical light on the centralist state and administrative structure of France with its deficitary possibilities of autonomy and participation, with reference to the *Théorie des pouvoirs intermédiares* ('Theory of intermediate bodies'). In Germany Montesquieu's constitutional theory, particularly the principle of the separation of powers, had a certain influence on the constitutions of 1849 and 1871, although the writer of *L'Esprit des Loix* could only be quoted to a limited extent, as a French state theoretician in nineteenth century Germany, as his German translator Adolf Ellissen noted in 1844. Montesquieu's emphasis on the plurality and relativity of societies and his equally global and interconnective view of social structures make him the most important forerunner of modern sociology in French-speaking culture, from the point of view of nineteenth and twentieth century sociologists. The French sociologist Raymond Aron went one step further than Auguste Comte and Emile Durkheim, who saw in Montesquieu a mere forerunner, but not a genuine first representative of modern sociology, in his book *Les Etapes de la Pensée Sociologique*. In his study on Montesquieu, *La Politique et l'Histoire* (1959), the French philosopher Louis Althusser emphasized the scientific nature of Montesquieu's theory and his methodical approach in conceding him a central significance in the history of European social sciences.

See also: Absolutism, History of; Balance of Power, History of; Balance of Power: Political; Constitutionalism; Constitutionalism, Comparative; Democracy: Normative Theory; Enlightenment; French Revolution, The; Nostalgic Social Thought; Political History (History of Politics); Political Science: Overview; Power in Society; Power: Political; Republicanism: Philosophical Aspects; Sociology, History of; Theory: Sociological

Bibliography

Althusser L 1959 *Montesquieu. La Politique et l'Histoire*. Presses Universitaires de France, Paris

Aron R 1967 *Les Étapes de la Pensée Sociologique*. Gallimard, Paris

Bryce J 1883 *The American Commonwealth*. Macmillan, London, Vol. 1

Chen Y 1993 *Les lettres Chinoises*. Leméac, Montreal, PQ

Dadié B 1959 *Un nègre á Paris*. Presence Africaine, Paris

Desgraves L 1992 *Montesquieu*. Societes-Verlag, Frankfurt am Main, Germany

Durkheim E 1953 *Montesquieu et Rousseau, Précurseurs de la Sociologie*. Reviére, Paris

Ellissen A 1844 *Montesquieu, Geist der Gesetze. Nebst Destutt de Tracys Kommentar und Noten von Helvetius und Voltaire. Aus dem Französischen von Adolf Ellissen*. Leipzig, Germany, 3 Vols.

Gauvin L 1984 *Lettres d'une Autre*. L'Hesagone, Montreal, PQ

Grouvelle P A 1789 De l'autorité de Montesquieu dans le Révolution présente, Paris

Marat J-P *1883 Éloge de Montesquieu, Présenté à la Académie de Bordeaux le 28 mars 1785*. [de Brézetz A (ed.)] Libourne

Meinecke F 1936 *Die Entstehung des Historismus*. Oldenbourg, Munich, Berlin, 2 Vols.

Montesquieu C 1721 *Lettres Persanes*. Amsterdam (Rouen), 12, 2 Vols., pp. 311, 347

Montesquieu C 1725 *Le Temple de Gnide*. Paris, France 12, **XII**, p. 82

Montesquieu C 1734 *Considérations sur les Causes de la Grandeur des Romains et de leur Décadence*. Amsterdam, **IV**, p. 277

Montesquieu C 1748 *De l'Esprit des loix, ou du Rapport que les Loix Doivent avoir avec la Constitution de Chaque Gouvernement, les Moeurs, le Climat, la Religion, le Commerce, etc. A*

quoi l'auteur a Ajouté des Recherches Nouvelles sur les Loix Romaines, Touchant les Successions, sur les Loix Françaises et sur les Loix Féodales. Geneva, Switzerland, 2 Vols., **XXXII**, p. 522, **IV**, **XVI**, p. 564

Montesquieu C 1750 *Défense de L'Esprit des loix, à Laquelle on a Joint Quelques Eclaircissements*. Geneva, Switzerland, 12°, p. 207

Montesquieu C 1752 Explications données à la Faculté de Theologie sur les propositions quélle a extraites du liver intitule Espirit des Lois et quélle a censurées. In: Caillois R (ed.) *Ôeuvres complicites*. Colliniard, Paris, pp. 1172–95

Morilhat C 1996 *Montesquieu. Politique et Richesses*. Presses Universitariaes de France, Paris, France

Rousseau J J 1762 Du Contrat Social, ou principles du droit politique, Amsterdam

Shackleton R 1961 *Montesquieu. A Critical Biography*. Oxford University Press, Oxford, UK

Spurlin P M 1940 *Montesquieu in America*, 1760–1801. Louisiana State University Press, Baton Rouge, LA

Starobinski J 1957 *Montesquieu par Lui-Même*. Editions du Seuil, Paris, France

Stubbe-Da Luz H 1998 *Montesquieu. Reinbek bei Hamburg*. Rowohlt, Germany

H.-J. Lüsebrink

Mood-dependent Memory

Are events that have been encoded in a certain state of affect or mood (such as sadness) more retrievable in the same state than in a different one (such as happiness)? Stated more succinctly, is memory mood dependent? In search of an answer, many studies have been undertaken since the mid-1970s using a variety of materials, tasks, and mood-modification techniques (for reviews, see Bower and Forgas 2000, Eich and Macaulay 2000a). Though most of the studies published in the 1970s succeeded in demonstrating 'mood dependent memory' (MDM), the majority of those reported in the 1980s failed to do so. So quick and complete was this turnabout that in the view of one leading researcher, Gordon Bower, it took just eight years for MDM to go from being 'a genuine phenomenon' (Bower 1981, p. 134) to 'an unreliable, chance event' (Bower and Mayer 1989, p. 145).

The capriciousness of MDM would be cause for little concern were it not for the fact that the phenomenon has important implications for both clinical and cognitive psychology. Regarding clinical implications, it has long been recognized that severely depressed patients have trouble remembering events that they earlier experienced in a nondepressed state (see Teasdale 1983). Traditional psychoanalytic theory ascribes this problem to a dynamic process of denial, which implies that depressed individuals are *unwilling* to remember prior periods of well- being. However, as Reus et al. (1979) have conjectured, it may be that the patients are *unable* to remember, owing to the shift in their affective state (from a neutral or possibly positive affect to a decidedly negative one). In addition to depression, MDM has been implicated in such diverse clinical conditions as bipolar illness, traumatic amnesia, and dissociative identity disorder (see Johnson and Magaro 1987, Schacter and Kihlstrom 1989).

With respect to cognitive implications, Bower has admitted that when he began working on MDM, he was 'occasionally chided by research friends for even bothering to demonstrate such an "obvious" triviality as that one's emotional state could serve as a context for learning' (Bower and Mayer 1989, p. 164). Though the criticism seems ironic today, it was incisive at the time, for many theories strongly suggested that memory should be mood dependent. These theories included the early drive-as-stimulus views held by Hull (1943) and Miller (1950), as well as such modern innovations as Tulving's encoding specificity principle (Tulving and Thomson 1973) and Bower's own network model of emotion (Bower and Cohen 1982). Thus, the frequent failure to demonstrate MDM reflects badly on many classic and contemporary theories of memory, and it blocks understanding of the basic issue of how context affects learning and remembering (see Davies and Thomson 1988).

Prompted by these clinical and cognitive implications, researchers in the 1990s focused their efforts on identifying why mood dependence sometimes comes and sometimes goes. These efforts have paid off, as it now appears that MDM can be demonstrated in a robust and reliable manner, provided that several critical factors are taken into account. These factors are of two types: one concerned with characteristics of the subjects' encoding and retrieval tasks; the other with attributes of the moods they experience while performing these tasks. The following sections review evidence bearing on both types of factors; unless noted otherwise, this evidence stems from studies in which happy and sad moods were engendered experimentally in 'normal' subjects (university undergraduates or other healthy volunteers) by means of a specific affect-induction technique (examples of which will be given later).

1. *Task Factors*

Regarding task factors, one key consideration is the manner in which retrieval of the target events is tested. According to several accounts (e.g., Bower 1992, Eich 1995a, Kenealy 1997), MDM is more apt to arise when retrieval is mediated by 'invisible' cues produced by the subject than by 'observable' cues provided by the experimenter. Thus, free recall is a much more sensitive measure of mood dependence than are either cued recall or old/new recognition memory.

A second critical task factor is, in effect, the complement of the first. Just as the odds of dem-

onstrating MDM are improved by requiring subjects to generate their own cues for retrieving the target events, so too are these prospects enhanced by requiring subjects to generate the events themselves. In support of this proposition, Eich and Metcalfe (1989) observed a significantly greater mood dependent effect in the free recall of verbal items that subjects had actively generated (e.g., *guitar*, elicited by the request: *Name a musical instrument that begins with g*) in contrast to items that the subjects had simply read (e.g., *gold*, embedded in the phrase: *Gold is a precious metal that begins with g*). This result, which was replicated by Beck and McBee (1995), occurs regardless of whether the overall level of generate-item recall is higher than that of read items—the prototypic 'generation effect' (Slamecka and Graf 1978)—or whether it is lower (as happens when subjects read some target items three times but generate others only once). Moreover, and in line with remarks made in the preceding paragraph, the results of a test of old/new recognition memory, which was administered shortly after free recall, showed no sign of mood dependence for either type of target. Thus, it seems that the more one must rely on internal resources, rather than on external aids, to generate both the cues required to effect retrieval of the target events *and* the events themselves, the more likely is one's memory for these events to be mood dependent.

This reasoning provided the impetus for a series of studies by Eich et al. (1994). During the initial or encoding session of one of these studies (Experiment 2), university students were asked to recollect specific episodes or events, from any time in their personal past, that were called to mind by neutral-noun probes such as *ship* and *street*; every participant recollected as many as 16 different events, each evoked by a different probe. After recounting the particulars of a given event, subjects rated its original emotional valence—that is, whether the event was a positive, neutral, or negative experience when it occurred. Half of the subjects completed this task of autobiographical event generation while they were feeling happy and half did so while sad—affects that had been induced by having subjects ponder mood-appropriate ideas and images while mood-appropriate music played softly in the background.

During the retrieval session, held two days after encoding, subjects were asked to freely recall (i.e., recall in any order, without benefit of any observable reminders or cues) the gist of as many of their previously generated events as possible, preferably by recalling their precise corresponding probes. Subjects undertook this test of autobiographical event recall either in the same mood in which they had generated the events or in the alternative affective state, thus creating two conditions in which encoding and retrieval moods matched (happy/happy, sad/sad) and two in which they mismatched (happy/sad, sad/happy).

Results of the encoding session showed that, on average, subjects generated more positive events (11.1 vs. 6.7), fewer negative events (3.3 vs. 6.8), and about the same small number of neutral events (1.2 vs. 2.0) when they were happy rather than sad. This pattern replicates prior work (e.g., Clark and Teasdale 1982, Snyder and White 1982), and it provides evidence of mood *congruent* memory—the 'enhanced encoding and/or retrieval of material the affective valence of which is congruent with ongoing mood' (Blaney 1986, p. 229).

Results of the retrieval session provided evidence of mood *dependent* memory. Relative to subjects whose encoding and retrieval moods matched, those whose moods mismatched recalled, on average, a smaller percentage of their previously generated positive events (26 percent vs. 37 percent), neutral events (17 percent vs. 32 percent), and negative events (27 percent vs. 37 percent). Collapsing across event types, total-event recall averaged 35 percent in the happy/happy condition, 23 percent in the happy/sad condition, 27 percent in the sad/happy condition, and 34 percent in the sad/sad condition.

This effect does not appear to be a fluke: The same advantage was seen in two other studies in which happy and sad moods were induced through a combination of ideas, images, and music (Eich et al. 1994, Experiments 1 and 3) as well as in three separate studies in which the subjects' affective states were altered by changing their physical surroundings (Eich 1995b). Moreover, similar results were obtained in a recent investigation of bipolar patients who cycled rapidly, and spontaneously, between states of mania or hypomania and depression (Eich et al. 1997). Thus it seems that autobiographical event generation, when coupled with event free recall, constitutes a useful tool for exploring mood dependent effects under both laboratory and clinical conditions, and that these effects emerge in conjunction with either experimentally induced or naturally occurring shifts in affective state.

2. *Mood Factors*

Up to now, discussion has focused on factors that determine the sensitivity of an encoding or a retrieval task to the detection of mood dependence. It stands to reason, however, that no matter how sensitive these tasks may be, the odds of demonstrating MDM are slim in the absence of an effective manipulation of mood. So, what makes a mood manipulation effective?

One consideration is mood strength. By definition, MDM demands a statistically significant loss of memory when target events are encoded in one mood and retrieved in another. It is doubtful whether anything less than a substantial shift in mood could produce such impairment. Indeed, the results of a meta-analysis by Ucros (1989) revealed that the

greater the difference in moods—depression vs. elation, for example, as opposed to depression vs. a neutral affect—the greater the mood dependent effect. In a related vein, Bower (1992) has proposed that MDM reflects a failure of information acquired in one state to generalize to the other, and that generalization is more apt to fail the more dissimilar the two moods are.

No less important than the strength of the moods is their stability over time and across tasks. In terms of demonstrating MDM, it does no good to engender a mood that evaporates as soon as the subject is given something to do, like memorize a list of words or generate a series of autobiographical events. It is possible that some studies failed to find mood dependence simply because they relied on moods that were potent initially but paled rapidly (see Eich and Metcalfe 1989).

Yet a third element of an effective mood is its authenticity or emotional realism. Using the autobiographical event generation and recall tasks described earlier, Eich and Macaulay (2000b) found no sign whatsoever of MDM when undergraduates simulated feeling happy or sad, when in fact their mood had remained neutral throughout testing. Moreover, in several studies involving the intentional induction of specific moods, students have been asked to assess candidly (postexperimentally) how authentic or real these moods felt. Those who claim to have been most genuinely 'moved' tend to show the strongest mood dependent effects (see Eich 1995a, Eich et al. 1994).

Thus it appears that the prospects of demonstrating MDM are improved by instilling affective states that have three important properties: strength, stability, and sincerity. In principle, such states could be induced in a number of different ways; for instance, students might (a) read and internalize a series of self-referential statements (e.g., *I'm feeling on top of the world* vs. *Lately I've been really down*), (b) obtain false feedback on an ostensibly unrelated task, (c) receive a posthypnotic suggestion to experience a specified mood, or, as noted earlier, (d) contemplate mood-appropriate thoughts while listening to mood-appropriate music (see Martin 1990). In practice, however, it is possible that some methods are better suited than others for inducing strong, stable, and sincere moods. Just how real or remote this possibility is remains to be seen through close, comparative analysis of the strengths and shortcomings of different mood-induction techniques.

3. Conclusions

The preceding sections reviewed recent research aimed at identifying factors that play pivotal roles in the occurrence of mood dependence. What conclusions can be drawn from this line of work?

The broadest and most basic conclusion is that the problem of unreliability that has long beset research on MDM is not as serious or stubborn as most investigators believed it to be only a decade ago. More to the point, it now appears that robust and reliable evidence of mood dependence can be realized under conditions in which subjects (a) experience strong, stable, and sincere moods, (b) take responsibility for generating the target events themselves, and (c) also assume responsibility for generating the cues required to retrieve these events.

Taken together, these observations make a start toward demystifying MDM—but only a start. To date, only a few factors have been examined for their role in mood dependence; odds are that other factors of equal or greater significance exist, awaiting discovery. For instance, it is conceivable that mood dependent effects become stronger, not weaker, as the interval separating encoding and retrieval grows longer (see Eich 1995a). Also, given that individual differences in personality have already been shown to play an important part in mood-congruent memory (see Bower and Forgas 2000, Smith and Petty 1995), it is possible—indeed probable—that subject factors also figure prominently in mood dependent memory (see Eich and Macaulay 2000a). And although the literature is replete with clinical conjectures about the role of MDM in various forms of memory pathology, it is lacking in hard clinical data. To date, few studies have sought to demonstrate mood dependence in individuals who experience significant, sometimes extreme, alterations in mood as a consequence of a psychiatric condition, such as unipolar mania (Weingartner et al. 1977), bipolar illness, or dissociative identity disorder (Nissen et al. 1988). As we have seen, recent research involving induced moods in normal subjects has helped establish the reliability of MDM. By exploring MDM within the context of the marked mood swings shown by select clinical samples, it may be possible to establish the phenomenon's generality.

See also: Amygdala (Amygdaloid Complex); Emotion in Cognition; Encoding Specificity in Memory; Episodic and Autobiographical Memory: Psychological and Neural Aspects; Implicit Memory, Cognitive Psychology of; Limbic System; Memory, Consolidation of; Reconstructive Memory, Psychology of

Bibliography

Beck R C, McBee W 1995 Mood dependent memory for generated and repeated words: Replication and extension. *Cognition and Emotion* **9**: 289–307

Blaney P H 1986 Affect and memory: A review. *Psychological Bulletin* **99**: 229–46

Bower G H 1981 Mood and memory. *American Psychologist* **36**: 129–48

Bower G H 1992 How might emotions affect learning? In: Christianson S-A (ed.) *Handbook of Emotion and Memory*. Erlbaum, Hillsdale, NJ, pp. 3–31
Bower G H, Cohen P R 1982 Emotional influences in memory and thinking: Data and theory. In: Clark M S, Fiske S T (eds.) *Affect and Cognition*. Erlbaum, Hillsdale, NJ, pp. 291–332
Bower G H, Forgas J P 2000 Affect, memory, and social cognition. In: Eich E, Kihlstrom J F, Bower G H, Forgas J P, Niedenthal P M (eds.) 2000 *Cognition and Emotion*. Oxford University Press, New York, pp. 87–168
Bower G H, Mayer J D 1989 In search of mood dependent retrieval. *Journal of Social Behavior and Personality* **4**: 121–56
Clark D M, Teasdale J D 1982 Diurnal variation in clinical depression and accessibility of memories of positive and negative experiences. *Journal of Abnormal Psychology* **91**: 87–95
Davies G M, Thomson D M (eds.) 1988 *Memory in Context: Context in Memory*. Wiley, Chichester, UK
Eich E 1995a Searching for mood dependent memory. *Psychological Science* **6**: 67–75
Eich E 1995b Mood as a mediator of place dependent memory. *Journal of Experimental Psychology*: *General* **124**: 293–308
Eich E, Macaulay D 2000a Fundamental factors in mood dependent memory. In: Forgas J P (ed.) *Feeling and Thinking: The Role of Affect in Social Cognition*. Cambridge University Press, New York, pp. 109–30
Eich E, Macaulay D 2000b Are real moods required to reveal mood-congruent and mood dependent memory? *Psychological Science* **11**: 244–8
Eich E, Macaulay D, Lam R W 1997 Mania, depression, and mood dependent memory. *Cognition and Emotion* **11**: 607–18
Eich E, Macaulay D, Ryan L 1994 Mood dependent memory for events of the personal past. *Journal of Experimental Psychology*: *General* **123**: 201–15
Eich E, Metcalfe J 1989 Mood dependent memory for internal vs. external events. *Journal of Experimental Psychology*: *Learning*, *Memory*, *and Cognition* **15**: 443–55
Hull C L 1943 *Principles of Behavior*. Appleton-Century-Crofts, New York
Johnson M H, Magaro P A 1987 Effects of mood and severity on memory processes in depression and mania. *Psychological Bulletin* **101**: 28–40
Kenealy P M 1997 Mood-state-dependent retrieval: The effects of induced mood on memory reconsidered. *Quarterly Journal of Experimental Psychology* **50A**: 290–317
Martin M 1990 On the induction of mood. *Clinical Psychology Review* **10**: 669–97
Miller N E 1950 Learnable drives and rewards. In: Stevens S S (ed.) *Handbook of Experimental Psychology*. Wiley, New York, pp. 435–72
Nissen M J, Ross J L, Willingham D B, MacKenzie T B, Schacter D L 1988 Memory and awareness in a patient with multiple personality disorder. *Brain and Cognition* **8**: 21–38
Reus V I, Weingartner H, Post R M 1979 Clinical implications of state-dependent learning. *American Journal of Psychiatry* **136**: 927–31
Schacter D L, Kihlstrom J F 1989 Functional amnesia. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*. Elsevier, New York, Vol. 3, pp. 209–30
Slamecka N J, Graf P 1978 The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology*: *Human Learning and Memory* **4**: 173–84
Smith S M, Petty R E 1995 Personality moderators of mood congruency effects on cognition: The role of self-esteem and negative mood regulation. *Journal of Personality and Social Psychology* **68**: 1092–107
Snyder M, White P 1982 Moods and memories: Elation, depression, and the remembering of the events of one's life. *Journal of Personality* **50**: 149–67
Teasdale J D 1983 Negative thinking in depression: Cause, effect, or reciprocal relationship? *Advances in Behavior Research and Therapy* **5**: 3–25
Tulving E, Thomson D M 1973 Encoding specificity and retrieval processes in episodic memory. *Psychological Review* **80**: 352–73
Ucros C G 1989 Mood state-dependent memory: A meta-analysis. *Cognition and Emotion* **3**: 139–67
Weingartner H, Miller H, Murphy D L 1977 Mood-state-dependent retrieval of verbal associations. *Journal of Abnormal Psychology* **86**: 276–84

E. Eich

Moral Development: Cross-cultural Perspectives

In the lay person's view of morality, both good and bad as well as right and wrong behavior are considered. Discussions about morality from the psychological viewpoint cannot be separated from the underlying theory, because different theories have different concepts for defining morality. Three main theories in developmental psychology, namely psychoanalytic (Freud), behavioral (Skinner), and cognitive developmental (Piaget), deal with morality, but are concerned with different meanings of morality (Turiel 1998). Freud proposed the concept of conscience, which he defined as the internalization of societal norms. Morality as defined by Skinner consists of behaviors that have been reinforced (positively or negatively) with value judgments associated with cultural norms. Knowledge and judgment about social relationships were considered central to morality in Piaget's theory.

Each of the three theories also differs in describing development (changes over time) in morality. The clearest picture of changes over time in moral judgment is found in Kohlberg's theory of the stages in moral development, which is based on Piaget's theory. Kohlberg's stage theory of the development of moral judgment has triggered research in several different countries and cultures. Because of this, Kohlberg's theory was chosen to discuss the cross-cultural perspectives in moral development. The following discussion, therefore, will cover: (a) the stages in the development of moral judgment and underlying reasons for formulating them, the process and conditions contributing to moral development, and the assess-

ment of moral stages; and (b) a review of perspectives and research results from different cultures that support or do not support Kohlberg's concepts of moral development.

1. Kohlberg's Theory of the Development of Moral Judgment

Kohlberg stated that morality deals with the reasons, considerations, and judgments which underly an individual's decisions about matters that are good or bad, or right or wrong. Kohlberg also proposed six stages of moral judgment, which could be more generally classified into three levels.

1.1 The Stages

The first level of moral judgment, the preconventional level, is generally found in elementary school children. In stage 1, the first stage of this level, children give reasons for their behavior that are consistent with socially acceptable norms because they are told to do so by authority figures, namely parents or teachers. Stage 2, which is the second stage of this level, is characterized by a view that right behavior means acting in one's own interests. The second level of moral judgment is that generally found in society and therefore is known as conventional morality. Stage 3, which is the first stage of this level, is characterized by reasons which seek to do what will gain the approval of others. Stage 4, the second stage of this level, is characterized by an orientation to abide by the law and respond to the obligations of duty. The third level of moral judgment, the postconventional level, is not reached by the majority of adults. Stage 5, which is the first stage of this level, is characterized by an understanding of social mutuality and genuine interest in the welfare of others. Stage 6, the second stage of this level, is characterized by reasoning based on respect for universal principle and the demands of individual conscience. The stages are summarized in Table 1.

The formulation of the stages in the development of moral judgment was based on the principle of morality, namely justice, which is characterized by equality and reciprocity between the self and others or between one's right and obligation in deciding between what is right or wrong. Based on longitudinal studies in various countries, Kohlberg (1969) claimed that the sequence of the stages in moral development is invariant.

1.2 The Process of Moral Development

Kohlberg conceptualized the process of moral development as occurring from the solution of moral cognitive conflicts in the individual. The concept of moral cognitive conflict is analogous to Piaget's concept of disequilibrium in cognitive development. With regard to morally relevant cognitive conflict, Kohlberg states that this will occur when, during interactions with others, individuals experience role-taking opportunities and obtain feedback in the form of reasoning from a higher stage, in this case only one stage higher. As a result they need to reconsider their previous reasoning. If individuals succeed in overcoming their moral cognitive conflict, their stage of moral judgment will progress. However, if they are not able to overcome their moral cognitive conflict, they will remain in the same stage of moral judgment (Setiono 1980). The experience of moral cognitive conflict in the individual can be induced by participation in moral discussions.

Inductive disciplinary technique and democratic/authoritative child-rearing practices both positively influence moral development. Power-assertion, love withdrawal disciplinary techniques, and authoritarian child-rearing practices negatively influence moral development (Salztein 1976, Eckensberger 1994).

1.3 Method of Assessment

Much research has been based on Kohlberg's theory because in addition to providing a detailed conceptual framework, Kohlberg and his colleagues also designed a manual for assessing the stages in moral development. Answers to hypothetical dilemmas typical for each stage of development can be detected through criterion judgments. By comparing an individual's answers to criterion judgments stated in the manual, the individual's stage of moral judgment can be determined. These criterion judgments were formulated on the basis of research using sample children and youth coming from middle- and upper-class families, conducted at the Center of Moral Development, Harvard University, USA.

2. Cross-cultural Perspectives and Research Results

In this section the following issues are discussed from a (cross-) cultural perspective: the nature of moral principles, stages of moral judgment, the process of progression to a higher stage including determining conditions, and methods of assessment. If cross-cultural research findings support Kohlberg's conceptions of moral judgment, it means that there is a universality of the concept; if not, it can be interpreted as the cultural bias in Kohlberg's theory, or the society can be classified as a more traditional one.

2.1 Moral Principle

Various cross-cultural studies on moral development have shown that the justice principle as a core of Kohlberg's theory seems to be deficient (Eckensberger 1996a). There are non-Western moral principles that

differ from Kohlberg's: among others are the Confucian principle of *giri-ninjo* (obligation), which seems to be close to the Indian 'ethic of duty' (Miller 1994, Schweder et al. 1987), the Javanese moral principles of *hormat* (respect to others) and *rukun* (harmonious social relations) (Setiono 1994), the principle of 'collective happiness' of the kibbutzniks (Snarey 1996), and the principle of 'connectedness' of the Chinese (Ma 1988). In addition to the cross-cultural findings, in a study of women's morality, Gilligan (1982) found that responsibility and care are moral principles.

2.2 Stages of Moral Judgment

Cross-cultural research reveals the universal existence of stages 2, 3, and 4, but this was not necessarily the case for stages 1 and 5 (Eckensberger 1996b). Stage 1 was hardly encountered in the USA, Bahamas, Turkey, and Tibet. Surprisingly, stage 5 was also not detected in Kohlberg's latest longitudinal USA data (Colby et al. 1983) but occurred in kibbutzim in Israel (Snarey 1996, Bar-Yam et al. 1996), Germany (Eckensberger 1996b), and China (Lei 1994). In a number of developing countries stage 5 reasoning was absent (Eckensberger 1997a), and this is interpreted as indicating that the justice principle of Kohlberg's theory does not fit with the local ethical principle. Concerning the highest stage of moral judgment, Setiono (1994) argued that Javanese people scored as stage 3 by Kohlberg's manual have the highest stage of moral judgment based on the local moral principle. The nonoccurrence of the higher stages, namely stage 5 or 6, cannot be interpreted as a lack of moral development in that culture. A moral principle which differs from Kohlberg's concept (justice) could be the cause of this nonoccurrence of higher stages. Snarey and Keljo (1991) pointed to *Gemeinschaft*- like postconventional morality as an explanation concerning the absence of postconventional morality in many cultural groups. In contrast to Kohlberg's types, a *gemeinschaft*-like moral voice or type would seem to redefine the same categories as follows:

(a) Hierarchy: any moral hierarchy is derived from the nature of human interconnectedness.

(b) Intrinsicality: means and ends are paradoxically inseparable. Social units are understood as ends in themselves as are individual persons.

(c) Prescriptivity: there is an organismic relationship between right and obligation. The two cannot be seperated. Moral duty is conceived of in vocational terms, in the language of 'ought.'

(d) Universality: judgments are universal based on organismic connectedness, not based on abstract principle.

(e) Freedom: the separation of the external from the internal is a false dichotomy. There is an organic relation between the individual and tradition.

(f) Mutual respect: respect is based on mutual interconnectedness.

(g) Reversibility: role-taking and empathy are central dimensions of moral judgment.

(h) Constructivism: being is given. Rules and laws are derived.

(i) Choice: solutions are chosen in terms of their relationship to and impact on the whole. There is awareness of multiple systems which are connected by a common unity of organic being; actions are rooted in one's place in the larger web of being.

2.3 Progressing to a Higher Stage and its Determinants

(Cross-) cultural research into the process of progressing to a higher stage is represented by the studies of Van Doorn (1982), Asih Menanti (1988), and Hessy (1999). These studies were conducted in Indonesia and were concerned with the existence of role-taking opportunities and moral cognitive conflict and its resolution, in the process of progressing to a higher stage. The results showed that discussions of moral issues and decision-making experiences concerning moral issues are useful for the development of moral judgment. The cross-cultural findings on the process of moral development, including research on social role-taking opportunities, and moral cognitive conflict and its resolution, gave support to Kohlberg's concepts regarding the process of progressing through the stages. Cross-cultural research on the socialization determinants of progression through moral reasoning stages, particularly disciplinary techniques and child-rearing practices, is represented by the study of Mayarina (1984) and Prameswari (1990) in Indonesia and some studies in India (Eckensberger 1994). Cross-cultural research findings indicate that inductive disciplinary techniques and authoritarian/democratic child-rearing practices foster the development of moral reasoning, which provides some support for Kohlberg's assertions regarding the role of cognitive conflict in moral judgment.

2.4 Assessment Method

(Cross-) cultural research using Kohlberg's assessment method revealed that certain responses could not be scored according to the criterion judgment as stated in the manual. Lei and Cheng (1996) pointed out that some of their interviews with individuals in Chinese communities were difficult to analyze, because their subjects referred to different types of collective conflict-solving strategies not mentioned by Kohlberg. Tietjen and Walker (1996) were faced with a similar problem in Papua New Guinea.

According to Eckensberger (1997a), the following criteria should be fulfilled for cross-cultural assessment.

(a) They contain a real moral conflict for the individual.

(b) All the facts mentioned in the dilemma are familiar to the subject.

(c) No situational (social and language) factors mask the individual's moral stage.

3. Future Research Directions

Results of studies about different moral principles in different cultures, and the differences between men and women, give rise to the following questions: should research be based on the differences in moral principles in the collective and individualistic cultural orientation categories existing in different cultures? Should research always specify the subjects as male or female?

Little research on the process of progressing through the stages of moral reasoning and the relationship of moral development to moral behavior has been done. Therefore it would be interesting to do a cross-cultural study on whether there are cultural specifics in regard to the roles of social role-taking opportunities, moral cognitive conflict, and parental discipline in the development of moral reasoning. Is Kohlberg's statement about the consistency between moral reasoning and moral behavior of stage 5 and 6 valid in different cultures? In cultures where stage 5 and 6 do not exist, would this mean that there would be no consistency at all between moral reasoning and moral behavior?

See also: Functionalism in Anthropology; Gender and Feminist Studies in Anthropology; Human Rights, Anthropology of; Hunting and Gathering Societies in Anthropology; Moral Education; Morality (East and West): Cultural Concerns

Bibliography

Asih Menanti 1988 The role of participation in moral discussion on progressing moral stage of university students. Thesis, Padjadjaran University, Bandung, Indonesia

Bar-Yam M, Kohlberg L, Naame A 1997/1980 Moral reasoning of students in different cultural, social and educational settings. In: Eckensberger L H, Zimba R F (eds.) *The Development of Moral Judgment. Handbook of Cross-cultural Psychology, Vol 2. Developmental Psychology*, 2nd edn. Allyn and Bacon, Boston

Colby A, Kohlberg L, Gibbs J, Lieberman M 1983 *A Longitudinal Study of Moral Judgment. Monograph of the Society for Child Development* 48(1–2) Serial no. 200, University of Chicago Press for the Society for Research in Child Development, Chicago

Eckensberger L H 1994 Moral development and its measurement across cultures. In: Lonner W J, Malpass R S (eds.) *Psychology and Culture*. Allyn and Bacon, Boston

Eckensberger L H 1997a Moral development and its measurement across cultures. In: Lonner W J, Malpass R S (eds.) *Psychology and Culture*. Allyn and Bacon, Boston

Eckensberger L H 1997b Research on moral development. In: Eckensberger L H, Zimba R F (eds.) *The Development of Moral Judgment. Handbook of Cross-cultural Psychology, Vol 2. Developmental Psychology*, 2nd edn. Allyn & Bacon, Boston

Gilligan C 1982 *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA

Hessy 1999 The influences of moral discussion on moral judgment of university students in Bandung. Thesis, University of Padjadjaran, Bandung, Indonesia

Kohlberg L 1969 Stage and sequence: The cognitive developmental approach to socialization. In: Goslin D A (ed.) *Handbook of Socialization Theory and Research*. Rand McNally, Chicago

Lei T 1994 Being or becoming moral in Chinese culture: unique or Universal? In: Eckensberger L H, Zimba R F (eds.) *The Development of Moral Judgment. Handbook of Cross-cultural Psychology, Vol 2. Developmental Psychology*, 2nd edn. Allyn and Bacon, Boston

Lei T, Cheng S W 1996 An empirical study of Kohlberg's theory and scoring system of moral judgment in Chinese society. In: Eckensberger L H, Zimba R F (eds.) *The Development of Moral Judgment. Handbook of Cross-cultural Psychology, Vol 2. Developmental Psychology*, 2nd edn. Allyn and Bacon, Boston

Ma H K 1988 The Chinese perspectives on moral judgment development. *International Journal of Psychology* **23**: 201–77

Mayarina I 1984 The relation between parental discipline and moral reasoning of early adolescents. Thesis, University of Padjadjaran, Bandung Indonesia

Salztein H D 1976 The role of disciplinary techniques on moral development. In: Lickona T (ed.) *Moral Development and Behavior*. Holt, Rinehart & Winston, New York

Setiono K 1980 The role of study service experiences on the development of moral judgment and social perspective taking of univerasity students. Dissertation, Padjadjaran University, Bandung, Indonesia

Setiono K 1994 Morality from the viewpoint of Javanese tradition. Paper presented at the Symposium 'Eco-Ethical Thinking from a Cross-Cultural Perspective,' Kirkel, Germany

Snarey J 1996 The social and moral development of kibbutz founders and sabras: a cross-sectional and longitudinal cross-cultural study. In: Eckensberger L H, Zimba R F (eds.) *The Development of Moral Judgment. Handbook of Cross-cultural Psychology, Vol 2. Developmental Psychology*, 2nd edn. Allyn and Bacon, Boston

Snary J, Keljo K 1991 In a Gemeinschaft voice: the cross-cultural expansion of moral development theory. In: Kurtines W M, Gewirtz J L (eds.) *Handbook of Moral Behavior and Development Vol 1: Theory*. Erlbaum, Hillsdale, NJ

Tetjen A, Walker L 1996 Moral reasoning and leadership among men in a Papua New Guinea village. In: Eckensberger L H, Zimba R F (eds.) *The Development of Moral Judgment. Handbook of Cross-cultural Psychology, Vol 2. Developmental Psychology*, 2nd edn. Allyn and Bacon, Boston

Turiel E 1998 The development of morality. In: Damon W (ed.) *Handbook of Child Psychology, Vol 3: Social, Emotional, and*

Personality Development, 5th edn. John Wiley & Son, New York
Van Doorn A 1982 The role of study service experiences in the process of stage progression of moral judgment. Thesis, University of Nijmegen, The Netherlands

K. Setiono

Moral Economy and Economy of Affection

'Moral economy' and 'economy of affection' have different yet related meanings, related in that both assume that human behavior and choice are not driven by material interests alone. The difference lies primarily in the empirical application of the two concepts. Moral economy refers to strategies that peasant farmers in fragile livelihoods adopt to maximize their safety and security. Economy of affection is a broader concept focusing on how individuals, regardless of socioeconomic status, use relations with other persons to protect and promote their own interests.

1. *Definition of the Concepts*

Moral economy and economy of affection are both relatively young concepts in the behavioral and social sciences, having taken root in the literature in the last quarter of the twentieth century. Yet their intellectual origins go back another 100 years in the evolution of economic and social anthropology.

1.1 *Moral Economy*

This term gained prominence with the publication of Scott's study of peasant societies in Southeast Asia (Scott 1976). It assumes that peasants act on the basis of their normative view of the world. Two fundamental values play an important role in a moral economy type of analysis. The first is the ethic of subsistence, which assumes that the right to subsistence is supposed to be assured. The other is the ethic of reciprocity, according to which the respect of the peasants for their elites and the payments made to those elites are supposed to be balanced by a return flow of justice, protection, and subsistence insurance. Moral economy differs from theories stressing individual utility maximization, notably 'rational choice', in three significant respects. The first is that what is collectively rational outweighs individual rationalities. The second is that peasants are risk-averse rather than risk-prone. The third is that the motive for action is not given but needs to be investigated empirically. Like Marxist political economy, moral economy acknowledges the adverse effects that the globalization of economic relations has on peasant livelihoods. It differs from that approach, however, in that it attributes to peasants the ability to defend themselves against the structural forces of world capitalism.

The essence of the moral nature of this economy stems from peasants' responses to the threat of these external forces. These include investment in village insurance and welfare institutions to help assure subsistence for everyone. For example, cultivation on communal lands may be developed as such an insurance mechanism. A local tax may be levied on the village as a whole on the premise that the better off will pay more than the poor. This safety-first maxim is the logical consequence of the ecological fragility of peasant livelihoods and finds expression in a wide array of actual choices, institutions, and values in peasant society.

The peasant community, according to moral economy, then, is not just an association of rural producers, but a ritual and cultural unit which structures choices within certain societally defined limits. It is reinforced by a strong sense of corporate identity, and prescribes certain rights and obligations which all members of the community share. Although inequality exists in the relations of community members to patrons connected with the outside world, relations are still governed by informal bonds of reciprocity and familiarity. Being concerned foremost with subsistence and basic survival, peasants quite understandably are suspicious and fearful of commercial farming techniques and practices which may place them in too much dependence on the capricious fluctuations of markets. At best, they only grow cash crops to fulfill certain obligations required by patrons or the state. In the moral economy the degree to which peasants are willing to participate in the production of high-risk, inedible cash crops is therefore largely dependent upon customary practices and the family unit's ability to assess and bear risk. The moral economy has been criticized for overemphasizing the risk-averse orientation of peasants. Popkin (1979), for example, maintains that individuals are driven by a universal 'investment logic' that applies not only to market but also to nonmarket exchanges. Traditional community institutions that in this perspective promote exploitative patron–client relations, communal interdependence, and stagnant forms of social stratification, operate to constrain the inherent 'rational' inclination of people to seek personal gain and mobility. Therefore, when confronted with opportunities to escape this confining environment, peasants are often willing to gamble on innovations to improve their livelihood. Outside market forces are not seen as inherently threatening, as in moral economy, but as offering potential sources for socioeconomic mobility and advancement.

The moral economy has its intellectual roots in a long tradition of peasant studies which largely sup-

ports its basic premises. Polanyi (1957), inspired by earlier anthropologists like Malinowski and Thurnwald, has argued that institutions through which goods are produced and allocated in what he calls 'primitive and archaic' societies are embedded inextricably in other social institutions. Redfield (1957) and Shanin (1966) have provided a similar perspective when arguing that because peasants are only reluctantly linked up with the state and market, they identify most closely with the local community and its affective bonds and informal institutions. Chayanov (1966), writing on the Russian peasantry in the early years of Communism, emphasizes the strong domestic orientation among these producers. Meeting the needs of the household is the primary concern of the peasant and therefore a determining factor of the amount of labor that the working members of the household are prepared to put in.

1.2 Economy of Affection

The concept of economy of affection was coined by Hyden in a study challenging Marxist interpretations of the political economy in Africa (Hyden 1980). The term has often been confused with moral economy, as both are centered on the ethos associated with subsistence and reciprocity. Their common intellectual origin in substantivist anthropology notwithstanding, economy of affection is not confined to the peasant sector only, but refers more generally to how individuals invest in social relations to actively promote and protect their interests in society.

This term is particularly applicable to socio-economic contexts where neither class solidarity nor institutional trust is prevalent. Seeking out other persons for a favor is natural. So is the extension of benefits to others on the implicit assumption that such a gesture will be met reciprocally at some unspecified time in the future. Building up claims against other persons, as many anthropologists and historians, e.g., Berry (1993), have shown, is often a prerequisite for acquiring material things. African farmers have always gained access to productive resources through such social relations. In practical terms this has led to the proliferation of claims of fixed assets such as land, tree crops, or cleared fields. Property rights in the economy of affection remain subject to renegotiation and multiple claims rather than converging towards private ownership by individuals or closed corporate groups. Social relations in the economy of affection, therefore, tend to be informal. The boundary between 'private' and 'public' typically is blurred.

Affection serves as a driving force in the same way as 'capital' does in a capitalist economy. Wealth and prestige in society are obtained through the clever use of affective relations. If the management of such relations fails, the consequences tend to be the same as when management of capital fails: the system stagnates or gets polarized. Disaffection in various forms, therefore, is the downside of the economy of affection.

Some anthropologists have argued that economy of affection generalizes relations that are more complex than the term can really grasp. They reject the idea that affection is the core of a single social logic that can be applied to the level of a whole society. Like moral economy, economy of affection has also been criticized for exaggerating the 'embeddedness' of individual choice. Those who have examined it in the context of rural Africa, e.g., Kasfir (1986), have argued that peasants there are integrated more closely into the capitalist world economy than economy of affection with its emphasis on other values than economic gain implies. Others have noted that there is no scope for conflict or social contradiction in the economy of affection, thereby rendering it incapable of explaining class stratification and polarization.

2. New and Additional Emphases in Theory and Research

The lifespan of both terms covered here is still relatively short, but it is worth noting that although originally they were developed in the context of studying rural societies, the principles of both moral economy and economy of affection are potentially more broadly applicable. The former speaks to issues of alienation and threats to people's cultural values that can be found in many industrial societies. For example, the threat to their lifestyle that many people concerned with community or ecological values feel in the face of economic development is one area where moral economy intersects with other lines of study, e.g., postmaterialism. Economy of affection is also present in societies regardless of economic development, although it may be more visible in places where modernity is still to be hegemonic. The *mafia* is an example of groups of people operating in accordance with the logic of an economy of affection. Because of the extensive privatization of Western culture, especially in the US, much of what is 'affective' is confined to the private realm. For some neo-institutionalist economists, such social arrangements are viewed as inhibiting the evolution of an efficient market. In this perspective, affective relations become a cost rather than a benefit. While affection has a spontaneous expression in many societies, especially in non-Western countries, it is more formalized and contained in the West. In the US, e.g., people's readiness to express concern and compassion for those beyond the circle of their private lives tends to be confined to voluntary associations.

In spite of being of increasing relevance and interest to scholars studying social and political phenomena in Western societies, the primary value of moral economy

and economy of affection continues to lie in the study of non-Western societies. There are at least three areas, in which the two concepts will serve as helpful analytical tools.

The first are studies focused on human responses to pressures on natural resources. As various factors—e.g., population growth, intensified use, global economic forces—influence the ability of households to sustain their livelihoods, moral economy or economy of affection provides a complementary analytical perspective which is potentially as useful as others. For example, the political economy literature has for a long time been dominated by a liberal, neoclassical, or socialist Marxist perspective on what happens as land use intensifies and becomes more commercialized. The first assumes the ability of individual household members to respond positively to this change in circumstances. The second assumes alienation and marginalization with the ultimate consequence of transforming social consciousness. Neither theory, however, exhausts the options that people are likely to adopt. In recent years, social scientists have come to accept that household members find ways of coping with and adapting to new circumstances using their wit in ways that fit neither neoclassical nor Marxist assumptions about human behavior. Moral economy and economy of affection, respectively, points to how such behavior is induced by communitarian or customary values and by more opportunistic investments in social relations with people that may enhance coping or adaptation.

These two concepts also complement interpretations about human responses to intensified resource use that are associated with Thomas Malthus and Ester Boserup. The Malthusian theory assumes that rapid population growth leads to involution and stagnation because technological change is incapable of keeping pace with the demographic expansion. Boserup (1981), on the other hand, assumes that population growth induces technological change, because people respond to demographic factors in a positive way. Countries with high population density and ongoing high population growth, such as Bangladesh, have been described as Malthusian crises in the waiting. Yet, neither there nor in sub-Saharan Africa, where much the same conditions exist in various places, the scenario does not necessarily fit the theoretical assumptions of either Malthus or Boserup. There is neither doom, as implied by Malthus, nor a technological breakthrough, as suggested by Boserup. Because the household is viewed as more than just a production unit, people find ways of coping with their predicament drawing on the strength of their cultural values. The moral economy or economy of affection offers the possibility of alternative interpretations that are potentially as valid as those mentioned above.

The second area is the study of informal institutions. Moral economy and especially economy of affection is relevant to understanding how people behave in the context of organized action. It is not just the formal aspect of an organization that induces people to act in certain ways. Gouldner's classical study (1954) of patterns of industrial bureaucracy in a gypsum factory shows how informal relations between management and workers—the factory operating as 'a family'—were more effective than what a subsequent manager imposing a formal structure could achieve. Although the economy of affection as a concept in the social sciences was not in use at that time, his study points to what it means in a modern industrial setting.

Being in good terms with others becomes an especially valid concern in precarious situations when sustaining a livelihood on one's own is difficult or impossible. People put their faith in relatives, friends, and sometimes strangers by offering something small—tangible or intangible—in the hope that these people will respond in kind with an even bigger gesture when they are in trouble. Of the two concepts, moral economy is the more communitarian. Economy of affection stresses more individual strategies for how to cope with difficulties by investing in relations—*ad hoc* or more long term—with other people who are in a position to reciprocate in a generous fashion. Social networks of this kind may sometimes be formalized as the case is with self-help groups of various kinds.

The third area is politics. Here the communitarian nature of moral economy means that typically it refers to how social collectivies respond to a perceived common threat. This does not mean that all action is necessarily communal. The point is rather that individuals use shared cultural idioms to defend themselves against external threats, whether they are local landlords or the more impersonal forces of a capitalist economy. Scott's study (1985) of the 'weapons of the weak' is an interesting illustration of how poor people can make sense of their reality without falling into despair thanks to powerful values in the local culture.

Economy of affection is more applicable to the informal aspects of politics that tend to develop in any society, especially where patronage is being used. Here policy matters less than who you are connected with and patronage becomes a means of sustaining relations with a client base as well as with fellow-patrons. This pattern is more prominent in some societies than in others. Studies of politics in Africa, Asia, and Latin America, in particular, have confirmed its permeating influence in those parts of the world. Investments of an affective nature, however, are made not only in vertical relations between patrons and clients. Economy of affection also accommodates the notion that such investments are made in lateral relations among equals with a shared objective. When faced with an external enemy, e.g., people can transcend local identities in order to maximize their gains, as the case was with nationalists in Africa during colonial days. This happens in a less dramatic fashion on a more everyday basis when individuals engage in reciprocal action to develop their community. Such collective action,

especially in non-Western societies, tends to be driven by a social logic other than individual utility maximization.

3. Methodological Issues

Recent developments in the social sciences toward a greater emphasis on the microfoundations of macro action, especially pronounced in the field of economics and political science, pose new challenges to those who are interested in using moral economy or economy of affection as their conceptual tool. In the past, the latter two have been treated as outside the mainstream because they emphasize communitarian or affective values that have not been formalized into theory. With rational choice and institutionalism becoming more dominant, the two terms discussed in this segment, however, do not become less relevant but potentially more so. The one-dimensional interpretation of human behavior and choice associated with individual utility maximization may be theoretically and methodologically convenient, but it leaves out of consideration any alternative logic of action, such as those implied in moral economy and economy of affection. Rational choice, e.g., presupposes a 'civic' logic based on cognition of substantive issues as driving behavior. This is a common, though by no means omnipresent, feature of Western societies, but is much rarer in non-Western societies. The logic of economy of affection is more dominant. Individuals in such societies are driven by a subjective logic based on affection. This means that the issues of what a particular choice means to the relations with other people comes first. Choices are embedded in social relations in ways that are not included in rational choice or game theory.

The challenge before the social sciences is how far an economy of affection can be formalized to prove a more effective rival to rational choice. It lends itself to game-theoretic analysis, although the type of games one plays in the economy of affection differs in its strategic assumptions from those conventionally made in such analysis. Further discussion of this set of issues can be expected in the years ahead.

See also: Alienation, Sociology of; Communalism; Communitarianism, Sociology of; Exchange in Anthropology; Exchange: Social; Institutionalism; Networks and Linkages: Cultural Aspects; Peasants and Rural Societies in History (Agricultural History); Peasants in Anthropology; Social Justice

Bibliography

Berry S 1993 *No Condition is Permanent*. University of Wisconsin Press, Madison, WI

Boserup E 1981 *Population and Technological Change*. University of Chicago Press, Chicago

Chayanov A V 1966 *The Theory of the Peasant Economy*. Irwin, Homewood, IL

Gouldner A 1954 *Patterns of Industrial Bureaucracy*. Free Press, Glencoe, IL

Hyden G 1980 *Beyond Ujamaa in Tanzania*. University of California Press, Berkeley, CA

Kasfir N 1986 Are African peasants self-sufficient? Review article. *Development and Change* **17**(2): 335–57

Polanyi K 1957 *The Great Transformation*. Beacon Press, Boston

Popkin S 1979 *The Rational Peasant: The Political Economy of Rural Society in Vietnam*. University of California Press, Berkeley, CA

Redfield R 1957 The universally variable and the culturally variable. *Journal of General Education* **10**: 150–60

Scott J C 1976 *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia*. Yale University Press, New Haven, CT

Scott J C 1985 *Weapons of the Weak*. Yale University Press, New Haven, CT

Shanin T 1966 The peasantry as a political factor. *Sociological Review* **14**(1): 1–23

G. Hyden

Moral Education

1. What is Educational Moral Progress?

If we ask about moral progress in sociological terms, the most popular answer would be that: during the last few centuries, humans have become more morally rational. Rorty (1998) on the basis of the philosophical reflection of Baier (1987) denies this hypothesis and instead believes that moral progress develops through moral sentiments and moral story telling, in which feelings and sentiments for the weak and the strangers are articulated. If we apply this idea to the educational and developmental field it could mean that moral educational progress also develops through the development of moral feelings; through story telling and possibilities to put oneself in the shoes of others. Or, in simpler terms, reading *Les Misérables* from Victor Hugo would be more effective than rationally solving moral problems. Indeed, it is crucially important, for example, to understand that we cannot stimulate pupils to achieve a higher level of moral judgment (one form of rational progress) simply by imparting knowledge, or by using teaching methods based on practical experience and problem solving. Or indeed, by treating personal experiences in a creative/artistic manner, or by the teaching of strategies designed to improve memory, and so forth. All these things are of course important, but they will not lead to an accommodative transformation of cognitive moral patterns. Nor is it

possible to succeed in achieving a higher level in a short space of time. It is imperative, moreover, to be aware that transformations such as these can be brought about only against the background of development theory and a transformational grammar appropriate to it, involving a constant process of critical arguments. Some knowledge of the transformation of cognitive moral judgment is therefore required if others are to be successfully stimulated to achieve a higher level, and it must also be recognized that human beings are perennially disinclined to take steps to transform and improve their patterns of behavior (to make progress). Knowledge of equal complexity but different in nature is required for the task of creating shared norms, or when a competency to act is aimed for (see later).

2. *Complexity Through Situated Moral Learning*

The foregoing suggests that moral education in situated contexts is a complex and multidirectional phenomenon; that it can provoke a network of suggestions for action, that norms play a role and that justifications are required. If, however, as is currently the case in the USA, a single method is widely regarded as the only correct one, or if teachers subscribe to a single method (e.g., the character education approach, or development-oriented education), then this will inevitably lead to false causal assumptions concerning the antecedents and the consequences of the education in question. When Lemming (1997, p. 47) asserts that the character education movement is eclectic 'both in terms of its psychological premises and its pedagogical practices,' we may counter by stating that whatever system of education has precedence, practice is always eclectic, and that up to now there has been too little research devoted to untangling the separate strands of simultaneous effects. One result of this debate, in the late twentieth century, is that there is no longer just one type of moral education and a pedagogic method appropriate to it. We are attempting rather, to reach different goals at the same time in a consciously differentiated manner, and to stimulate a corresponding learning activity.

But how can we create a synthesis of all these forces? How can we postulate a moral education, which will enable and produce the complex formation of moral sense in a young person? What are the minimal preconditions, and what is the indispensable core of moral education?

3. *The Triforial System of Moral Education: A Theory*

The approach involving three core elements, all of which work at the same time, is designated as 'triforial' because the term suggests that the core elements have something in common, namely their foundation, their support, and their actualization of a moral structure. Based on three-arched windows, a 'triforium' permits different things each to be supported in a different way. A triforium is a kind of gallery in the interior of Romanesque and especially Gothic churches, consisting usually of triple-arched windows running under the roof space of the transept and nave. Despite all the problems of borrowing analogies, the concept of a triforial structure can readily be transferred to the realm of moral education, where the tripartite arched positions signify the three core elements, the structure they support, and the general formation to which the three core elements lead (cf. Klafki 1991). Why do we speak of this 'three-fold' or 'triforial' moral education and its accompanying three-fold pedagogic practice (in which the figure of '3' is used to describe the mere minimum of the many links which are always present in educational modes of action). Here, the problem is formulated in a negative triforial manner:

(a) Moral education is more than a training in weighing up adversarial positions. Although such experiences in critical reasoning stimulate moral judgment, they do nothing more. The moral judgment is merely a precondition for moral action.

(b) On the other hand, one can also say that education designed to foster character and inculcate values simply represents an attempt to influence pupils through persuasion, and as such it often remains blind and unreflective. This can readily be demonstrated through examples of secondary values such as cleanliness, punctuality, discipline, precision, and so forth (see Höffe 1992). These values must always be linked back to higher values (justice, solicitousness, objective neutrality), otherwise they threaten to become fragile, dangerous, or even fundamentalist. An extreme example would be the carrying out of genocide in the name of 'exactness' or 'precision.'

(c) If we simply stimulate moral action, however, then we must always deal with specific situations and use our intuition. But intuitions are frequently unfair, crude, and lacking in regard for others. Thus, sentiments and story telling alone would be one-sided.

Each one of these approaches falls short, either because we lack empathy, or fall victim to a false belief, or to blind action. We are courting the danger of being 'quasimoral,' of thinking and interpreting in a quasimoral manner. A deeper analysis shows that a theory alone only sets its sights on one extreme, and thus the central goals of a comprehensive moral education cannot be achieved. We are therefore in need of an overall theory, which will enable us to combine different goals in the proper manner. Different methodical or rather pedagogic modes of acting must be integrated within this theory in different ways. Instead of pursuing just one of the goals indicated above therefore we need a triforial theory of moral education which will permit at least three central technological frameworks of conditions. Figure 1

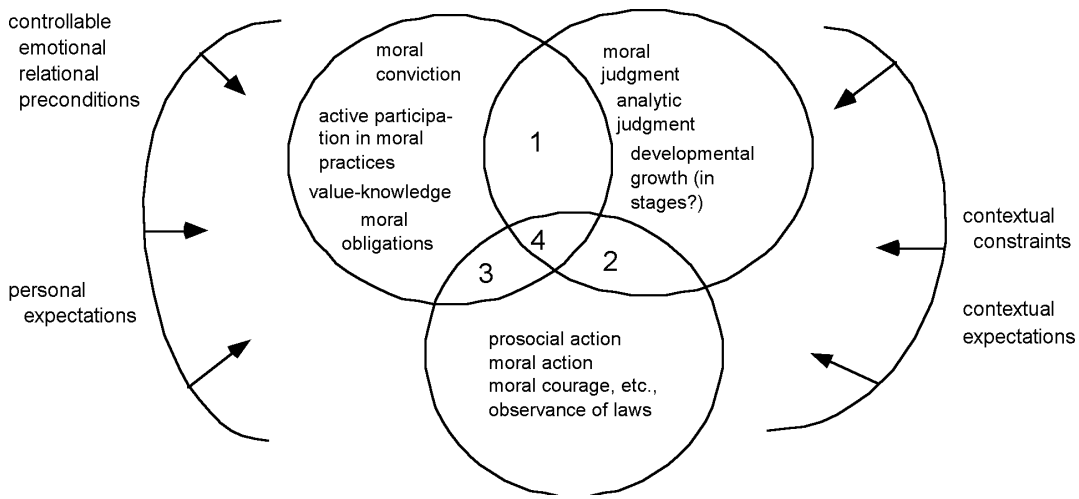

Figure 1
A triforial theory of moral education

shows these three fields, which intersect with each other.

The judgment circle contains moral analyses, moral justifications and a progressive stage-by-stage anchoring of moral thinking. At a given stage we can understand and elaborate the respective cognitive and knowledge-based arguments. The value core (circle 2) contains value knowledge, the intuitive knowledge deriving from moral customs and the respective specific moral culture. This is where moral convictions and moral group identity (unconsciously internalized, or consciously derived form direct participation) come into play. The action circle (3) contains forms of prosocial, moral, and participative behavior. Qualities required here include moral courage, moral sentiments, but also moral performance and the ability to deal in moral terms with concepts of law and justice.

In Fig. 1, we placed particular emphasis on the intersections of the core elements. Intersection 1 relates to content interpreted in a stage-specific manner, e.g., children's narratives are interpreted from the viewpoint of their belonging to a certain stage. Kohlberg's stage 2 stresses the morality of exchange ('you will get something only when you do something for another person'). Conversely, value judgments and culture-bound moral contents enter into the judgment and modify it. Intersection 2 emphasizes the necessity of combining judgment with action. This is the point at which the hiatus between judgment and action becomes apparent. A whole range of models illustrating this connection is available (see *Moralisches Urteil und Handeln*, Garz et al. 1999); more and more variables are being investigated to establish their effect on prompting action, including, for example, stage-level, strength of will, extent of obligations, etc. Moral culture possibly constitutes an important variable affecting moral action (Intersection 3), although there is admittedly a relative lack of research in the area. The subjective acceptance of value responsibility is determined through internal and external pressure against, or for, a certain moral or prosocial course of action. In the absence of a counter-pressure, which resists the direct will, we cannot speak of the acceptance of value responsibility (see Oser 1999, pp. 168–219). The overlapping process represents the moral self of an individual (Damon and Hart 1982). The moral self combines judgment, knowledge (consent), and action in a balanced manner.

It would now be possible to describe the triforial theory from the viewpoint of various aspects of educational theory, such as: (a) different goals, (b) diverse sources of moral thought, (c) a model of moral transformation (moral pedagogy), (d) negative moral thought or behavior, and (e) the measuring of morality. However, due to space restrictions, this cannot be elaborated on here.

4. *Historical Aspects of the Structuralist and Developmental Approach to Moral Education*

The change in the development of moral educational psychology during the last four decades of the twentieth century, is obvious: Up until the 1960s, the teaching of morals was regarded either as the moralization of the relationship between decisions and actions, or as the formation of conscience. Kohlberg (1958), influenced by Piaget and Symbolic Interactionism, presented a dissertation, which became the starting point for a paradigmatic change. From then on, morality was no longer viewed primarily in terms of

conscience or as an internalized emotional state, but rather as highly rational constructions and justifications of potential decisions and actions, justifications, each of which can possess a specific quality.

Kohlberg was aiming to develop a detailed genetic epistemic structure of morals. His work produced the stage-by-stage approach to moral development, a theory which he himself has improved at least three times: Stage 1: reward and punishment morality; stage 2: an eye for an eye, a tooth for a tooth morality; stage 3: good boy–nice girl morality; stage 4: rights and duties morality; stage 5: principles and social contract morality; stage 6: morality of principles held to be universal or based on communicative ethics, together with the assumed ideals of optimal conditions for potential decision-making and taking action.

What then is the significance of this stage structure from the educational point of view, given the fact that it has also been well proven, longitudinally (see Colby et al. 1987). This especially relevant question leads to the catch-phrase 'development as the aim of education,' which has played a considerable role both in research and in the specifics of practical intervention (Kohlberg 1981, 1984). In this view, education means stimulation of development towards the next higher moral stage. Yet, posing this question in no way implies an uncritical acceptance of Kohlberg's stage theory. There are in fact numerous criticisms, such as the assumed general over-orientation towards justice (see Gilligan 1982, 1983), or the criticisms claiming that the lower stages are less egocentric than Kohlberg had assumed (Keller 1996). Critics have argued that Kohlberg failed to observe, fully, the specificity of the social, moral or personal area (Shweder et al. 1987, Turiel 1983, Nucci 1981). The nature of the higher stages has been called into question (see Reichenbach 1996; for earlier criticism, e.g., on measurement problems, see Oser 1981).

(a) Much of this critical debate ignores the fact that Kohlberg conceived of his developmental theory as a 'theory of competence.' In describing competence theories, we are not so much concerned with asking how people form judgments in concrete situations, but rather with the question of how the highest and qualitatively best kind of judgment a person can make, is produced in a variety of general, mostly decontextualized, situations.

(b) It is also frequently overlooked that Kohlberg constantly stressed the limited scope of his scheme. The moral development theory of Kohlberg is in fact powerful precisely because its area of applicability is clearly circumscribed, and in its scope is both highly controllable and comprehensible. The structures are abstract and therefore highly transferable 'creations.' They indicate the extent and nature of the reversible thinking, which people can produce as justifications for their deeds and intentions. The higher the stage, the closer the judgment is to universal principles, the more adequate it is in terms of philosophical theories, and the more the subjects are able to think reversibly in the spirit of the Golden Rule.

(c) Another fact often ignored is that Kohlberg himself raised numerous questions, such as: 'What is the relationship between intelligence and morals?' 'How does judgment relate to action?' 'What role is played by emotional elements under real conditions?' 'Is principled thinking also possible at the lower stages?' 'Is moral judgment better described in categories of a soft or hard stage concept (Kohlberg 1981, 1984)?' Kohlberg was unable to provide answers to many of these questions (see Kohlberg 1995).

Kohlberg established early on, that people in the same age group could reason at different stages in terms of his scheme. It was his student Moshe Blatt (Blatt and Kohlberg 1975) who—on the basis of this finding—successfully attempted to stimulate pupils to attain the next stage upwards. In his study, and much research that followed, it was shown that an adversarial relationship between two values (the selection of the first disadvantages the second), discussed in an open atmosphere, plus the so-called plus-1-convention (confrontation with arguments from a stage or half a stage higher than one's own (see Turiel 1966, Berkowitz 1981)), were adequate for attaining higher stages. This possibility of stimulation provoked a host of intervention studies of uneven quality, which were carried out with various age groups and under different conditions. In 1985, Schläfli et al. published the first meta-analysis, which demonstrates that high quality interventions (i.e., studies with controversial discussions of dilemmas) realized in a medium time span of about 3 to 4 months are especially effective.

All these studies proceed from the *a priori* assumption that the stimulation of development constitutes the most important goal of moral education.

5. *The Judgment–Action Hiatus*

Leading on to a higher stage has the advantage of providing the opportunity to arrive at a judgment level, which is in fact less egocentric, more complex, based on principles, and of greater autonomy. Since higher stages, however, represent nothing other than reversible competence judgments, an exclusive concentration on the structure of moral judgment means than many other things belonging to human morality are not given due attention. The judgment/action connection, for example, is largely unexplained, as are questions regarding the acceptance of responsibility, training in empathy, dealing with 'immoral' behavior, actions expressing regret and emotional moral sensitization. The question of curricular content remains especially unsolved in this theoretical frame. However, at the end of the twentieth century, research has been published nearly on every one of these fields.

From the educational point of view, moral actions and the acceptance of moral responsibility are necessary constituents for the development of a moral self. It must be asked what is achieved by developing a higher judgment without, at the same time, encouraging moral courage, providing moral knowledge and learning to perform moral actions. Kohlberg himself already started out on a path that was decisive for the future of moral education, namely the *contextualization of stimulations for the higher stages* and together with this the *creation of a moral universe* for children and youth. This is in keeping with the 'situated learning movement,' which Kohlberg's work also anticipated. Moral learning can thus be fully integrated learning, when the acceptance of responsibility occurs within contexts and for contexts (Lave and Wenger 1989), otherwise the moral judgment is not capable of transfer and remains simply a judgment. The central idea of situated learning is that the learning situation itself determines the conditions within which acquired knowledge can later be applied (Fischer et al. 1994). In the case of situated learning, action is—so to speak—that part of the problem-solving process, which converts a set of given requirements into knowledge of appropriate action. Thus, people learn how to proceed in order to solve moral problems, and this in turn activates the process of accepting responsibility, which enables moral performance.

The one-sided approach of the Kohlberg early years in practical education needs to be integrated in a more comprehensive model of moral education in which judgment, action and sentiment are connected together in real content based situations.

6. *The Round Table Model*

'Realistic' discourse (Oser 1999) is the term we use to designate the basic model of learning and teaching developed in the spirit of the situated learning movement. It is a model within which the primary aim is not to attain the next higher stage, but which attaches central importance to finding a practical solution acceptable to everyone. It is a dynamic model geared horizontally towards action (and not oriented vertically towards a reversible judgment).

The second element is that the teacher warrants that everyone at the table will have the right to speak and to participate in the process of problem-solving and decision-making. Third, this presupposes that pupils are 'reasonable' and 'capable of responsibility,' even though they may appear to be 'lazy,' 'cheeky,' and 'stupid,' in other words appear superficially, to be irresponsible. The fourth and final point is that the solution must be put into practice and it must be accepted, that for the moment the approved solution is considered the best, and it will hold up over the following days and months, even though better solutions might be theoretically possible.

The basic principle of this model of moral education is that the frictions of everyday life in a school provide moral learning opportunities. Kohlberg (1981) identified three different kinds of moral education:

The first is the technological approach to imparting values, in which it is believed that a value can simply be implanted into a child. This approach fails on account of the danger of indoctrination. The second example is the so-called romantic approach, which just aims to make children consciously aware of values that have always been present. This approach fails because of the problem of arbitrariness or at best, on account of its relativism. The third approach, the progressive approach, is based on developmental psychology and aims to stimulate the progression to a higher stage within the scheme.

The fourth approach, the discourse approach, was not seen by Kohlberg. He did, however, outline the above-mentioned model of a situated learning environment, which reflects features of the discourse approach and which has now been developed further and is known by the name of 'Just Community.' The discourse model thus represents a kind of basis for the 'Just Community School.' The positive approach of the former comes to fruition in this context. Basically, the Just Community School is nothing less than a school-wide 'Round Table.'

7. *The Just Community School*

Kohlberg's idea (1986) that schools could be self-regulating communities, in which the very regulations themselves would provide opportunities for moral learning and the construction of a value system, is a concept which relies on the participation of everyone. This process can be produced only through a comprehensive transformation of the school and can be understood only by means of quality indicators such as those provided by basic research in the field of social and moral development.

In a Just Community School all of the pupils and teachers come together, either at regular intervals or when required by events, in order to solve school-related conflicts in a democratic manner; or to consciously change and shape aspects of school life, including standards and general school culture. The very center of this process is the forum, in which contentious questions are clarified and proposals for change put to the vote, and where every act of voting always implies an important setting of standards within the system of that particular school.

All of these structural devices support a range of principles which enable the formation of social and moral sensitivity in the individual within the group, and which do so in a way superior to other methods.

(a) First of all there exists the permanent possibility of 'putting oneself into the shoes' of others, for

assuming different roles, for defending positions. This in turn provides the social flexibility required in order to reach a nonpartisan decision.

(b) The second principle is 'development as a goal of education.' People attain a higher stage through controversies provided at every level of school life, genuine and contrived, and through discussion of dilemmas, relating to particular fields of study and school events in general. This is where the modalities of a stimulated cognitive imbalance become applicable.

(c) A third principle is that of identification through participation. Decisions become relevant when there is emotional involvement with them. Only those who participate and jointly decide can truly feel responsible for what is to be done.

(d) The fourth principle consists in overcoming the judgment/action hiatus. Basically the action to be decided upon is placed before the judgment (one thinks first of all of solutions and then looks for ways of justifying them), as is customary in, e.g., parliamentary procedure.

(e) The fifth principle finally relates to the reality of life itself. Through the process of co-determination, school-life becomes an authentic entity.

The model of the Just Community School presented here is relatively widespread in the USA, where it has had a longer tradition, but it has already been established in various models in Germany and Switzerland, in a modified and reconsidered form.

8. *Outlook*

One of the first questions to be raised, concerns the communitarism, which can be glimpsed in this model and which indicates that morality emerges differently in each society, that therefore different sources and principles are involved and that hence a degree of relativism must be admitted to. This is indeed true of the process we modeled of the 'Round Table.' Admittedly, we refrain from saying what does not constitute moral content, but simply state how morality comes into being. But this weakness can be countered by arguing that teachers inject their own opinions into the decision-making process, and that this provides for a certain shaping of content. Nor did the developmental model devised by Kohlberg (1995) deal adequately with the universal core of morals. This may not be a satisfactory response, but it does point to a debate which has not yet taken place on the relationship, between the teaching of ethics and a fully integrated intervention such as is represented by this model.

A second criticism is that the model is totally incapable of shedding light on the judgment/action conflict. What is happening is only a social remoulding of what has been decided through the 'Round Table' system. The Just Community, it is assumed, offers no mediation between insight and action. What we have here basically reflects the model of the free market with its optimization of individual and group prosperity, behind which a kind of contract comes into being along the lines of the rational-choice theory. The children and teachers, the argument portends, are doing all these things not from a sense of good will in itself, but merely in order to achieve a better quality of life in school. This argument too needs to be aired thoroughly. At this juncture, it can only be pointed out that educative action is always purpose-driven, and that schools, which develop their own laws, will always 'abuse' the very creation of these laws for educational purposes. This seeming 'abuse', however, is attributed to the law of preserving and furthering culture, and is thus supported from the point of view of social theory.

A third criticism is directed at the incompleteness of moral education in itself. The Just Community meetings do not address the formation of conscience or changes of heart or the desire to atone for something; nor do they deal with emotional contributions to the arguments. The aspect of negative morality is missing with its insistence that only those who have made moral errors can act in a moral manner. Aspects such as these must indeed be integrated and built into the model. Conversely, it can be argued that feelings can contribute to the overcoming of cognitive conflicts only when they are actually in use. And the 'JC' is the only form of organization that in practice, lays claim to feelings, and it does so with neither bigotry nor hypocrisy.

See also: Ethics and Values; Moral Development: Cross-cultural Perspectives; Moral Reasoning in Psychology; Moral Sentiments in Society; Morality (East and West): Cultural Concerns; Religion: Morality and Social Control; Value Pluralism; Values, Anthropology of; Values, Development of; Values: Psychological Perspectives; Values, Sociology of

F. K. Oser

Moral Panics

Students of deviance, social problems, and politics have used the concept of a moral panic to describe sudden concern over a group or activity, accompanied by calls for control and suppression. Out of an infinite range of potential perceived threats, one—which may be neither new nor on the rise—suddenly receives considerable attention. Marijuana use, motorbikes, and rock 'n' roll music are common examples. The news media, public officials, religious leaders, and private 'moral entrepreneurs' are key in focusing public attention on the issue, typically by identifying some recognizable group as 'folk devils'—usually

young people, racial and ethnic minorities, or other relatively powerless groups—responsible for the menace. New political or legal policies are sometimes the result, and new symbols and sensibilities (available as the raw materials for future panics) almost always are.

1. The Concept of Moral Panics

Trying to understand a 1964 incident in which massive police and media attention were given to an evening of rowdy behavior by young people at an English seaside resort, British sociologist Stanley Cohen first formulated the concept of a moral panic in 1972, borrowing the term from his colleague Jock Young. In Cohen's (1972) words,

> A condition, episode, person or group of persons emerges to become defined as a threat to societal values and interests; its nature is presented in a stylized and stereotypical fashion by the mass media; the moral barricades are manned by editors, bishops, politicians and other right thinking people; socially accredited experts pronounce their diagnoses and solutions; ways of coping are evolved or (more often) resorted to; the condition then disappears, submerges or deteriorates and becomes more visible.

The key idea, attractive to cultural constructionists, is that the public reactions and rhetoric are not clearly linked to any objective measure of threat. Scholars who use the concept of moral panic frequently try to show that the rates of the activity are actually declining or at least not increasing, or that those promoting the panic have interests in doing so that are independent of the actual harm done (for instance, police departments may receive more funding if they can persuade the public that Satanic child abuse is rampant). The term 'panic' seems to imply a disproportionate or self-defeating response, although the feelings and thoughts of those panicking will probably be the same regardless of the extent of any real threat. Although some involved in a panic are cynically trying to manipulate public opinion, others (presumably those being manipulated) really do perceive the putative threat as real.

Most scholars have examined panics over aspects of life-style considered—at least by those embracing them, but also usually by the analyst too—to be harmless personal choices, such as clothing styles, sexual choices, or substance use. For them it is easy to show that, under certain value systems, any repressive action is unwarranted. Others have applied the term to actions that are clearly harmful, such as child abuse, but in which the response appears exaggerated or misplaced. For them it is more difficult to show that the response is disproportionate, although they may show that it is ineffective. The term has also been used to describe those opposed to technologies such as nuclear reactors. It is assumed that the emotional dynamics of a panic are similar in these diverse cases.

Many observers have found the concept of moral panics useful because it opens a window onto a society's disagreements over basic values, often intuitively felt ones, as well as onto fears and anxieties that are normally submerged. A panic over child abuse may reveal broader concerns over changes in family structure; one over water fluoridation, fears of government bureaucracy; one over crack cocaine, racist attitudes toward African American urban youths. Cultural meanings rarely reveal themselves so starkly.

2. Moral Panics in History

Human history has been full of rumors, fears, and efforts—often deadly—to constrain or eliminate human sources of perceived problems. When the Black Death came to Europe in the fourteenth century, neither physicians nor priests were of much help, so large numbers of frightened Germans blamed the Jews who lived among them, exterminating most and driving the rest into Central and Eastern Europe. Because few disasters have been as widespread as the plague, few panics have been as broad, but local instances have been common throughout known history.

In the nineteenth century more systematic efforts appeared, especially in the UK and the USA, to make the lower classes conform to the expectations of the middle and upper classes, especially those aspects of life-style considered (by those behind the efforts) to be part of moral character. Drinking, cruelty to animals, neglect of children, pornography, even movies came eventually to be seen as antisocial behaviors that well-intended benevolent societies tried to extirpate, partly for the good of the lower classes themselves. Unionization, rebellion, and urban riots too were seen as little more than a form of crime or bestiality. For more than 100 years, most moral panics consisted of the privileged worried about the less privileged. (Although lower-class panics persisted, such as anti-Catholic riots in the USA.)

For most of this era, intellectuals were more likely to contribute to moral panics than to analyze, and thereby criticize, the panics of others. Few scholars questioned the widespread fears of urban mobs that characterized the middle and upper classes of nineteenth century Europe. In fact, they contributed to them, especially by dismissing 'crowd' behavior as irrational.

There seems to have been some kind of turning point around World War II, since which the sources of moral panics have grown more diverse. Intellectuals and academics, as they grew in numbers in the 1950s, instigated something like mild panics over conformity, the authoritarian personality, and mass society as signs of vulnerability to fascism or communism. Politicians were behind many of the panics over

communist infiltration. Grassroots issue entrepreneurs seem to have motivated the American panic over fluoridation, however. In the USA, the majority of moral panics since the 1970s have been the work of the religious right, whose diligent organizing is one reason the USA seems to have more panics than most other advanced industrial nations. There, panics frequently cluster around sex for pleasure rather than procreation (loose morals, homosexuality, teenage pregnancy, abortion), around race and poverty (crack and other drugs, welfare mothers, urban crime and violence), and around immigration (the English language, standards in the schools, overpopulation, job and wage competition). Each panic carries with it clearly defined folk devils.

Although moral panics are hardly new, they seem to have become more frequent in recent decades, or at least more visible to national audiences. For one thing, advanced industrial societies have a 'social movement sector' of professional activists who act as moral entrepreneurs as they try to attract attention, funds, and volunteers. Such individuals have long existed—think of Thomas Paine or Anthony Comstock—but they now have extensive formal organizations, political networks, and direct-mail fund-raising techniques at their disposal. Second, immigration into many countries has increased in recent decades, accompanied by heightened tensions over group boundaries, especially ethnic ones. Newcomers often have habits that are poorly understood and easily used to stigmatize the group. Increased penetration of the media, especially television, into each and every 'deviant' subculture has further spread awareness of alternative life-styles, especially sexual freedoms. The Protestant right in the USA and elsewhere has discovered the power of life-style panics for raising funds, if not recruiting new members.

3. *The Role of the Media*

Much of the 'panic' in moral panics—the disproportionate nature of the response—is due to framing and amplification by communications media. If they do not report on a phenomenon, or do not frame it as something one should be frightened about, there is rarely a panic. Thus the rise of television everywhere and the wide readership for tabloid newspapers in the UK, for example, have been blamed for an increase in moral panics. In some views, the media reflect the values and interests of economic and political elites, and so focus on the groups and behaviors that frighten these elites. In a less conspiratorial view, the media simply look for unusual life-styles and behaviors as a means to attract viewers, readers, and so on. In doing so, they spread information about what mainstream culture considers deviant. Sometimes, it is deviant enough to ignite panic.

Cohen (1972) originally described three ways in which media coverage contributes to moral panics: in the exaggeration and distortion of what actually has happened; in predictions that the same kind of thing will happen again; and in the framing of key symbols in pejorative ways. To a large extent, these make the event or behavior into news.

4. *Variations in Panics*

A number of variations in culture and social structure affect the shape and outcomes of moral panics. One of the most important is the political, and especially the electoral, system. Where, as in the USA, politicians are left to raise their own funds and do their own campaigning in single-district elections, they may be tempted to play on the fears of the majority, which often have to do with life-styles of minorities. Where, as in Sweden, representation is more proportional and political parties control the chances of individual candidates, party leaders at least have the opportunity to curb efforts at creating moral panics for electoral purposes. Of course, this allows entire parties to play the panic card, as in the case of the National Front in France.

The structure of the media industry also has an effect. In some nations governments exert considerable control over the content of the media—although the widespread relaxation of such control in recent years may have contributed to an increase in moral panics. Their intervention allows governments to curb sensationalism and prevent moral panics, but it also allows them to incite them to their own advantage. On another dimension, in countries with nationally centralized news media, such as the UK, panics can spread more quickly than in countries where the media are more regional and local, as in France or the USA.

Other potential moral entrepreneurs face varying kinds of incentive structures. Local police departments, for instance, may need to appeal to city or regional governments for additional funding. Certain occupations may have opportunities for advancing into new terrain because of moral panics. The more that a taxation or funding system allows input from local units or from the general population, the more tempting it is for those whose funding comes from these sources to use moral panics to increase their budgets.

Some moral panics reflect widespread public anxieties and fears that are easily stirred up through news coverage. Others seem traceable to specific interest groups who will benefit from the panic. Goode and Ben-Yehuda (1994) distinguish three models of moral panics, according to their origin. In one view, moral panics are relatively grassroots in origin, in a second they are organized by elites, and in a third vision they

are promoted by interest groups or social-movement organizations. But there is no reason to think that all panics must come from the same source, so that one model would fit them all. Goode and Ben-Yehuda's typology functions equally well for distinguishing different types of panic as different scholarly models: some panics are indeed more grassroots in origin, others are more elite-directed, and yet others are promoted initially by social-movement groups. In addition, some seem entirely a creation of the news media. These types may differ in their dynamics and outcomes as well as their origins.

As for differences in the general public, many commentators have assumed (more than demonstrated) that there are national differences in susceptibility to 'mass mobilization' in instances such as moral panics. It has been thought that a lively 'civic culture' should prevent the kind of demagoguery often associated with moral panics, and that formal education breeds tolerance for the alternative life-styles that are often demonized in panics. This evidence is not as strong as one might wish. Just the kind of educated, normally tolerant people one might expect to resist moral panics often end up leading them. One of their frequent (if not defining) features is the relative absence of critical, dissenting voices.

5. *Controversies*

The concept of moral panics has not lacked critics. Like all varieties of cultural constructionism, the analysis of moral panics can be turned against those who use it. Just as they find bias and exaggeration in their subjects of study, so their own critics can claim to find the same flaws in research on moral panics. Most scholars who have used the concept have been left-leaning and thus quicker to discern panics on the right, for instance. A more fundamental criticism is that the term panic is necessarily pejorative, assuming rather than demonstrating that those analyzed are not entirely rational. To prove irrationality or disproportionality, analysts must presumably investigate the goals of those accused of panic (often a complex and poorly defined cluster of goals) and show that they are undermining their own progress toward those goals.

Debates over the utility of moral panics as an analytic tool tap into deep disagreements about the rationality of protest and politics. For most of the modern era, students of collective protest viewed it as immature, irrational, or manipulated, a kind of crowd dynamic leading people to do things they would not do otherwise. In this tradition panics, like fads, were analyzed in the same way as movements for unions or civil rights. To critics, moral panics are a holdover from this pejorative vision that dismissed protestors as irrational.

In recent decades a very different view of such activity has arisen, in which collective action is seen as a pervasive and regular form of politics, in which groups and individuals organize to pursue their own interests. They mobilize resources, look for political opportunities in which to act, and formulate rhetoric to appeal to as many potential supporters as possible. The concept of panic seems to undermine this positive view of such activity, but in fact it focuses attention on precisely the dynamics that this rationalist perspective tends to overlook. Issues must be chosen from a myriad of possibilities; grievances must be constructed; emotions must be aroused, often through the demonization of others; previously private issues must be portrayed as a public concern. The dynamics highlighted by research on moral panics, in other words, characterize a great deal of politics both institutionalized and not.

Moral panics remain a common occurrence in modern societies. They generate considerable scholarly interest, although some care must be taken to avoid loaded language and tendentious arguments about them—and some will always think that the term 'panic' itself is necessarily loaded.

See also: Emotions, Sociology of; Mass Communication: Normative Frameworks; Mass Killings and Genocide, Psychology of; Mass Media and Cultural Identity; Media and Social Movements; Moral Sentiments in Society; Psychoanalysis in Sociology; Social Influence, Psychology of; Social Movements, History of: General; Social Movements: Psychological Perspectives; Social Psychology: Sociological; Social Trauma

Bibliography

Beisel N 1997 *Imperiled Innocents: Anthony Comstock and Family Reproduction in Victorian America*. Princeton University Press, Princeton, NJ

Cohen S 1972 *Folk Devils and Moral Panics: The Creation of the Mods and Rockers*. MacGibbon and Kee, London

Goode E, Ben-Yehuda N 1994 *Moral Panics: The Social Construction of Deviance*. Blackwell, Cambridge, MA

Hall S, Critcher C, Jefferson T, Clarke J, Roberts B 1978 *Policing the Crisis: Mugging, the State and Law and Order*. Macmillan, London

Jenkins P 1992 *Intimate Enemies: Moral Panics in Contemporary Great Britain*. Aldine de Gruyter, New York

Jenkins P 1998 *Moral Panic: Changing Concepts of the Child Molester in Modern America*. Yale University Press, New Haven, CT

Springhall J 1998 *Youth, Popular Culture and Moral Panics: Penny Gaffs to Gangsta Rap, 1830–1996*. St. Martin's, New York

Thompson K 1998 *Moral Panics*. Routledge, London

Waddington P A J 1986 Mugging as a moral panic: A question of proportion. *British Journal of Sociology* **37**: 245–59
Watney S 1987 *Policing Desire: Pornography, AIDS and the Media*. Methuen, London

J. M. Jasper

Moral Reasoning in Psychology

The term morality has been used in different ways, and its explanation is subject to a great deal of debate. One controversy pertains to whether morality varies by culture. The position of cultural relativism is that the moral standards of one culture do not apply to other cultures. Various explanations have been provided as to how individuals acquire the morality of their culture, including ones that emphasize the learning of a set of values or the incorporation of cohesive systems of moral thought (see *Moral Development: Cross-cultural Perspectives*). An alternative perspective, with roots in philosophy (e.g., Immanuel Kant, John Stuart Mill, John Rawls) is that morality involves reasoning about right and wrong with regard to social relationships, legal systems, and societies. In psychology, much research has focused on the development of moral reasoning. Morality is defined as entailing understandings of welfare (avoiding harm and promoting good), justice (in the distribution of goods, and equal worth of persons), and rights (group and individual).

1. *Social Interactions and Moral Judgments*

Two developmental psychologists have had a deep and lasting influence on thinking about moral development. One is Jean Piaget, who is regarded as the foremost contributor to developmental psychology more generally. Piaget's (1932) research on children's moral judgments was conducted in the late 1920s when sociologically oriented views of morality were predominant in the social sciences. As an example, the sociologist Emile Durkheim (1961) had proposed that morality constitutes a shared respect for the rules and values of society, and that children must be educated to form a strong emotional attachment to the group and to respect its rules and authority. Piaget, by contrast, examined the ways children construct ideas about rules, authority, fairness, and equalities. The influence of Piaget's theory of moral development on psychology was limited until the early 1960s. Up to that point, greater attention was given to psychoanalytic and behavioristic theories, each of which explained moral development as the incorporation of society's standards. The research of Lawrence Kohlberg (1969) served to refocus attention on moral reasoning.

1.1 *Two Moralities and Two Types of Social Interactions*

Piaget examined the organization of children's thinking and how they interpret events. In his work on moral development he observed children's use of rules in playing games, sharing (distributive justice), punishment (retributive justice), and intentions or consequences in assessing responsibility for transgressions. He proposed a sequence of two broad levels of children's moral judgments connected to two ways they relate to others. In the first, children (approximately 4–7 years) are heteronomous in their moral thinking and have unilateral respect for adults in authority. Heteronomy is characterized by absolutistic judgments, whereby social rules are regarded as fixed and requiring strict adherence. As an example, young children reify game rules, judging that they cannot be altered even if all the participants were to agree to the changes. They also believe that adults' commands must be followed to the letter. Children also judge actions by their material consequences rather than by the actor's intentions.

In part, heteronomous thinking is due to inherent relationships of inequality. Young children attend mainly to the greater power and status of adults—which imbues adults with moral authority. In later childhood there is a shift in moral thinking that occurs with increasing interactions with peers, which provide children with more opportunities to experience mutuality, reciprocity, and cooperation. As a consequence children shift from heteronomy to autonomy, defined as involving participation in the elaboration of norms (as opposed to receiving them ready-made). Through the reciprocity in relationships of equality children construct understandings of justice, mutual respect, and rules alterable by agreement to serve aims of cooperation.

1.2 *Stages in the Development of Moral Judgments*

In extensions of Piaget's research, further emphasis was placed on the need to ground social and behavioral research on definitions of the realm. In the dominant approaches of the time, considerations of the psychological means by which children acquire morality were addressed without formulations of the substantive elements of morality. Kohlberg (1969), especially, was instrumental in combining psychological with philosophical analyses (relying to a fair extent on neo-Kantian philosophers such as Rawls 1971). One reason for stressing inclusion of philosophical formulations was that Kohlberg proposed that children think

systematically about moral matters. He coined the phrase 'the child as a moral philosopher' to convey the idea that they form organized ways of thinking about concepts of justice, rights, and welfare.

In Kohlberg's research children were presented with hypothetical stories that depicted conflicts around issues of life, physical well-being, trust, law, authority, and retribution. (A well-known story depicts the conflict of a man who must decide whether to steal an otherwise unavailable drug to save his wife's life.) Kohlberg formulated a sequence of stages and levels that portrayed children's moral judgments somewhat differently from Piaget's characterizations. In particular, it appeared that young children do not reify rules nor possess an unyielding respect for authority (according to Kohlberg, Piaget had misinterpreted a moral orientation to power and sanctions as reification). Rather, in the first two stages (labeled a 'preconventional' level) moral value is confused with material value, sanctions, power, and attaining people's needs and desires. In early adolescence there is a shift to thinking about morality primarily in terms of social expectations and maintaining the social order by upholding laws and authority (the 'conventional' level). After attaining understandings of morality as rooted in rules or laws and the need to maintain the social order there emerges a way of thinking that includes applying moral judgments to society and its practices, rules, and authority. At the most advanced level (the 'postconventional'), which does not develop at least until late adolescence, concepts of justice, fair treatment, rights, and respect for persons are differentiated from the existing social order. At the most advanced level individuals make moral judgments about society's system of laws and norms.

2. *Distinguishing Morality from Other Domains of Social Judgment*

During the latter part of the twentieth century a great deal of research was influenced by Piaget and Kohlberg. However, the patterns of developmental changes they described have been questioned. In particular, research indicates that young children's moral judgments are neither heteronomous nor based on a confusion of moral value with material value or sanctions. One line of research demonstrates that young children have positive orientations to matters like helping and sharing (see *Prosocial Behavior and Empathy: Developmental Processes*). Another line of research shows that young children develop understandings of welfare and fairness distinct from understandings of other social domains (Turiel 1998). Using research methods that do not pose complex conflicts between moral and social considerations (as was the case in Kohlberg's methods), it was demonstrated that children do not treat all rules as fixed, do not solely have unilateral respect for authority, and do not judge by sanctions.

Children's moral ideas stem from interactions with children and adults that involve observing and experiencing harmful consequences, inequalities, or violations of freedoms. These moral judgments differ from another set of judgments about social uniformities that can be termed 'conventions.' Conventions are socially expected behaviors that coordinate interactions within social systems. Such uniformities in behavior gain their force from being customary and allow for greater efficiency in that people know what to expect of each other. Conventions, such as those pertaining to modes of greeting, forms of address, or dress codes, are judged by children to be alterable by consensus, contingent on common use, rules, or authority expectations, and relative to social contexts. Thus, children judge violations of conventions to be wrong only insofar as there is a governing rule or if it is dictated by an authority (e.g., a conventional rule in a school can be changed by the principal). Similarly, conventions are not generalized across contexts in that it is judged that other cultures can legitimately have a different set of customary behaviors.

In contrast with judgments about conventions, children do not judge moral issues (such as hitting or stealing) by the existence of uniformities, agreements, rules, or common practices. For instance, rules pertaining to moral issues are judged not to be legitimately alterable by general agreement or dependent on what authorities dictate, or relative to the social and cultural context. Maintaining the welfare and rights of persons, as well as promoting fairness in social interactions and cultural practices, guide evaluations of actions, rules, authorities, and the organization of social systems.

Morality is a fundamental domain of social reasoning. Children's judgments within the moral domain, which begin to emerge at a fairly young age, do not entail reification of rules or confusion with other social domains. Changes in moral reasoning occur with age. Young children's moral reasoning is focused primarily on welfare—preventing harm and promoting welfare. Older children and adolescents develop increased understandings of justice and reciprocity among persons, which they coordinate with concepts of welfare.

3. *Social Hierarchy and Social Justice*

The social decisions people make are not always directly in accord with their moral judgments. Moral judgments are applied in concrete situations, in the context of societal arrangements, taking other considerations into account. Another domain (in addition to morality and convention) contributing to social decision-making is the personal sphere (Nucci 1996). Individuals form judgments about personal activities considered out of the boundaries of conventional or

moral regulation. In most societies, an important feature relevant to moral, social, and personal decisions is social hierarchy, by which groups are accorded greater or lesser power, and status. Often, hierarchical distinctions are made on the basis of social class and gender (Nussbaum 1999).

The connections between gender and moral development have engendered intense controversies. Some have proposed that girls and women develop a morality of care, different from a morality of justice of boys and men (Gilligan 1982). Care is based on concerns with interdependence in networks of interrelationships, whereas justice is based on concerns with rules, individual rights, or equal treatment. Others, however, have argued that care and justice are part of the orientations of females and males, but that positions in the social hierarchy distinguish the perspectives of males and females (Nussbaum 1999).

Since women are usually in subordinate positions in the social hierarchy, they often experience injustices and inequalities, and in some cases are victims of oppression. The social system, therefore, can further the personal interests of some (e.g., men) at the expense of others (e.g., women) through social expectations and practices. Such arrangements constitute a source of tension and conflict within social systems. Whereas people in lower positions are part of the social system and do accept norms and practices, at the same time they can be critical of aspects of social arrangements. In many respects, moral reasoning produces social harmony since people are concerned with the welfare of others and promoting fairness. Another side of morality, however, is that it can result in conflicts and efforts at subverting societal arrangements judged to be counter to welfare, justice, and rights.

See also: Cognitive Development in Childhood and Adolescence; Human Rights, Anthropology of; Hunting and Gathering Societies in Anthropology; Infancy and Childhood: Emotional Development; Moral Development: Cross-cultural Perspectives; Moral Education; Moral Sentiments in Society; Morality (East and West): Cultural Concerns; Piaget's Theory of Child Development; Practical Reasoning: Philosophical Aspects; Problem Solving and Reasoning, Psychology of; Social Learning, Cognition, and Personality Development; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Durkheim E 1961 *Moral Education: A Study in the Theory and Application of the Sociology of Education*. Free Press, New York

Gilligan C 1982 *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA

Kohlberg L 1969 Stage and sequence: the cognitive-developmental approach to socialization. In: Goslin D A (ed.) *Handbook of Socialization Theory and Research*. Rand McNally, Chicago

Nucci L P 1996 Morality and the personal sphere of actions. In: Reed E S, Turiel E, Brown T (eds.) *Values and Knowledge*. The Jean Piaget Symposium Series. Erlbaum, Mahwah, NJ

Nussbaum M C 1999 *Sex & Social Justice*. Oxford University Press, New York

Piaget J 1932 *The Moral Judgment of the Child*. K Paul, Trench, Trubner, London

Rawls J 1971 *A Theory of Justice*. Belknap Press, Cambridge, MA

Turiel E 1998 Moral development. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*, 5th edn. Wiley, New York, Vol. 3

E. Turiel

Moral Sentiments in Society

The sociology of moral sentiments has a specific and paradoxical position in the history of sociology. Morality was an important research topic for most classical sociologists of the early twentieth century, whose works attest to the existence of a specific area of research within the discipline. Examples include Durkheim and then Gurvitch in France, Weber in Germany, Hobhouse and Westermarck in the UK, and to a minor extent Parsons in the USA. However, moral sociology has never succeeded in becoming an official subdiscipline of sociology, and it is difficult to trace back the history of its research and present its main paradigms. This situation derives from the lack of visibility of modern moral sociology. If sociology has continually dealt with moral issues, the focus on morality as a specific object of study has been neglected, in contrast to Durkheim's project. However, looking back to the classics reveals the difficulties of the project of a moral sociology.

1. Morality as a Specific Subject Matter of Sociology

What does a moral sociology consist of? For classical sociologists the common idea, which Durkheim has expressed several times, is that moral sociology cannot be the theoretical morality of philosophers. On the other hand, it is not obvious what can replace it. What will be the object of this science? Weber, Durkheim, and Gurvitch have all stated the difficulty of finding the empirical object of a moral sociology. Indeed, moral life is linked to most social activities, without having either a specific institution to represent it, nor a specific instrument to implement it. Besides, it is well known, even without the help of sociology, that moral systems vary according to societies and social groups.

What then will be the contribution of sociological analysis to the understanding of moral phenomena? By revealing the pluralistic and contingent nature of morality, sociology may risk impairing its ability to understand the unity and specificity of these phenomena. Moreover, this object of study has the peculiarity that if it is an experimental fact for the sociologist, it is also a specific object of experience for the individual. Therefore, it is expressed not only in terms of norms, rules, and morals, but also as a set of values, feelings, and ideas, namely as the result of individual experiences. The fact that morality is also embedded in the individual seems to impair the project of a positive moral science. However, we cannot understand the overall phenomenon of morality if we ignore its subjective nature. All these difficulties, which the classical authors have tried to solve, shed light on some aspects of the development of moral sociology and on its current position.

1.1 The Social Nature of Morality

For Durkheim (1960), moral rules are moral facts *par excellence*, or more precisely they are the form under which moral facts are scientific. Such a choice, which allows him to escape the subjectivity of the moral of common sense, explains why moral sentiments have not been, at least in France, the favorite subject of moral sociology. Besides, not only can sociological analysis bring to the fore moral variations according to social groups and societies, but it can also show the intrinsic social nature of morality. Indeed, Durkheim demonstrates that society is the very source of moral authority, which explains why, for individuals, morality can be both desirable and constraining. The merit of this theory is to show the additional understanding that sociology can bring compared to the mere description of the contingency of moral life. It also provides the unity and specificity of morality as a subject matter for sociology. Finally, after revealing the two essential properties of morality, both constraining and desirable, Durkheim (1960) determines the ideal toward which modern societies tend.

However, according to Gurvitch (1960), this normative ambition would have led Durkheim to focus on specific elements of morality, at the expense of the plurality of moral life. He adds that the sociologist, on the contrary, has to consider such diversity and then only reveal the functional correlations between moral systems and social types, leaving to the philosopher the justification side of it. This criticism of the normative feature of Durkheim 's analysis is typical of a continuing tendency that will lead to the following problem: the sociological analysis of moral life may have to resolve itself to being a mere description of the variable, plural, and contingent nature of morality, otherwise it might confuse explanation and justification. But beyond this disagreement, Durkheim and Gurvitch both suggest that the objectivity of morality is due to its strong link, whatever it is, with social life.

Durkheim comes up against another difficulty. In modern societies' secular morality, he discovers a third differentiating characteristic of morality, which he calls 'autonomy' or 'self-determination': individuals want to understand the *raison d'être* of moral rules. How can this be reconciled with the obligatory force of morality? Here, Durkheim (1961) suggests that the third element of morality is only 'the understanding of it,' an 'enlightened assent,' while society is still the author of morality. This solution shows how difficult it is to conceive that the feeling of constraint or obligation, characterizing the obedience to a moral rule, can originate in individuals themselves. If there is a constraint, it means that an external cause has intervened, for example, the group or society. That partly explains the success of analyzes in terms of social causes in modern moral sociology. However, this difficulty will sometimes lead Durkheim (1922) to abandon the term 'moral rules' in favor of 'moral act,' and finally 'moral judgment' (Durkheim 1920), in a posthumous text which should have been the introduction to his book on morality. By the same token, modern moral sociology will deal essentially with moral feelings, no matter what the choice of perspective.

1.2 The Subjective Meaning of Moral Sentiments

The sociological analysis proposed by Weber seems to be able to resolve both difficulties mentioned. Explaining moral activity, for Weber, would consist in understanding its meaning for the individual, which should be explained by finding the meaning of the original moral belief. The notion of 'axiological rationality' (Weber 1949) could be interpreted as meaning that individuals' moral points of view are meaningful for them and therefore are able to foster their social activity. Even if Weber does not express this so clearly, this is what comes out of his sociological analyzes (Weber 1930). In this perspective we understand how the feeling of moral constraint can originate in the individual. However, in other places, Weber (1968) notices that it is not possible to justify rationally the superiority of one ultimate value compared to another. This explains why social sciences cannot solve this sort of conflict. On the other hand, the sociologist can help to make the decision by assessing the degree of coherence of each statement and by revealing the empirical consequences of their application. Thanks to what Weber calls 'comprehensive analysis' the conflict between explanation and justification seems to be more flexible. Sociological analysis can then go beyond a mere descriptive analysis. However, the main idea that modern sociology has retained from Weber is that it is not possible to argue rationally about ultimate values, and in general about

moral opinions. Following this viewpoint, 'axiological rationality' is reduced to the rationality of an action expressing an unquestioning attachment to a value.

2. Morality in Modern Sociology

2.1 Modern Sociological Theory on Moral Sentiments: Causes and Utility

In view of these difficulties, sociological theory about moral feelings has developed in three main directions. A first possibility consists in circumventing the question of the specificity of morality by moving moral phenomena into either the social, cultural, or affective sphere. From then on, moral feelings can be explained in term of causes associated to these latter spheres. This is how models inspired by Durkheim have developed, notably in anthropology, in which individuals recognize as good those values and moral ideas that have been instilled in them by society. According to Geertz (1984), the objective of anthropology is to scientifically develop and establish such a relativist conception of morality. Otherwise the culturalist theory inspires all the sociological approaches that insist on the phenomena of socialization in order to explain the emergence of moral feelings. Among the causalists models are the affectivist theories which state that moral beliefs are rationalizations of subjective feelings. This latter argument is rooted in the positivist tradition according to which only positive beliefs can be rationally explained as only they stem from logical and experimental thought. Such a process has been used by Pareto (1935) when he came to analyze what he calls 'derivations.' Further, it can be found implicitly in many empirical studies of feelings of justice in social psychology.

The analysis of the moral act or moral sentiments through the prism of utilitarian rationality is another major direction taken by modern moral sociology. This instrumental interpretation of practical reason (which has also been prominent in the irrationalist version of Weber's axiological rationality) has found a fruitful source of inspiration in the utilitarian tradition initiated by Bentham and Mill. Such interpretation can be found in the rational choice model which inspires the analyzes of Oberschall (1994a, 1994b) on the emergence of norms, as well as Adams (1963) and Homans's (1974) equity theory regarding feelings of justice. According to Oberschall, norms and morals are 'social inventions' allowing in all cases benefits from social co-operation. Equity theory postulates that feelings of justice express a desire to see the proportionality between contributions and retributions of various individuals being equal. This desire is stimulated by the will to maximize one's personal benefits. Whether the analyzes are in terms of causes or in terms of utilitarian rationality, their common point is to move moral phenomena into other spheres of life or social activity, which eases their explanation but compromises the specificity of the initial object of moral sociology.

Finally, modern moral sociology follows a third direction which no longer consists in moving the object of analysis, but in its dispersion as it belongs to every field of social life. The difficulty in identifying the sector of moral life, as well as the lack of specialized institutions, leads modern moral sociology to abandon morality as a unified subject matter. Moral questions arise when one comes to study religion, family, couples, or health. Similarly, the ethical dimension is used as only one of the properties that characterize one social activity as, for example, political claims or social movements. But in all these analyzes, normative activity in itself and its content remain unexplained.

These three aspects of modern moral sociology's theoretical development can explain its lack of visibility within the discipline. Without any doubt this is why authors such as Isambert et al. (1978) have regretted the absence of ethical sociology.

2.2 Feelings of Justice and Social Values: Plurality and Complexity

In terms of empirical research, notably that carried out in social psychology and to a minor extent in sociology, a considerable amount of literature has been published about feelings of justice in the area of social justice. Using questionnaires, surveys, or experiences, individuals are invited to assess the fairness of different rules of distribution or to build their own distributions. Most analyzes are essentially descriptive and try to follow the variations of feelings of justice according to the social position of individuals, and to the context of distribution or the nature of the resources to be distributed. However, the individuals' evaluations are often analyzed either in an instrumental way, in relation to the objectives groups pursue, or with the help of psychological models. The first approach is illustrated by Deutsch (1986) who examined the impact of a distribution rule on group efficiency, as well as surveys by Kellerhals et al. (1988) about the ability of moral rules to promote or to maintain specific types of social relations. In the second approach some authors refer, for example, to feelings of relative deprivation (Runciman 1966), or to beliefs in a 'fair world' (Lerner and Lerner 1981). Generally, the accumulation and heterogeneity of models, as well as empirical studies themselves, reveal the plurality of criteria and principles of justice used by individuals, and the difficulty of interpreting them with a single theory.

Other important works in the area of empirical research are the quantitative studies by Stoetzel (1983) and Inglehart (1977) who compare systems of values at European and international levels, and the study carried out by Riffault (1994) who describes the

process of increasing diversification of values, as well as the tendencies to majority consensus. These works reveal the individual's ability to recognize both universal and particularistic values. A recent study by Hardin (1995) sheds light on the success of communitarian values.

2.3 Explaining Moral Complexity: The Rationality of Moral Sentiments

Considering the above empirical research, it is hard to adhere to the recent attempts to develop a naturalist explanation of moral feelings, such as the one by Wilson (1993) regarding moral sense, because they risk missing out the moral complexity revealed by the empirical analyzes. Another way to make sense of this complexity consists in concentrating on individuals' points of view and on their reasons, not only the instrumental ones but also the axiological ones, to adhere to moral beliefs in different contexts. It might then be possible to explain the plurality and variability of moral feelings in a way that is not only descriptive. Moreover, analyzing moral feelings with reasons instead of causes opens up an additional and fruitful path in order to resolve the dilemma Durkheim had to face, namely to reconcile the obligatory force of morality for individuals and their moral autonomy. Here, the explanation can be found in the feelings of objectivity and universality that underly the individual's own judgment. In this way, Boudon (1995, 2000) develops a 'cognitivist' model when he gives an account of individuals' feelings of justice by reconstructing the reasons that have convinced them in a given context. In doing so, he comes within the scope of 'comprehensive' sociology as defined by Weber. This perspective has also been implicitly adopted by de Tocqueville (1972) when he sought to understand the moral and religious beliefs of Americans, as well as by Durkheim (1915) when, moving away from the strictly moral area, he explains the primitive's religious beliefs in terms of reasons and asserts that individuals cannot passively adopt the beliefs that society would impose upon them. The sociological analysis of Boltanski and Thévenot (1991) also attest to a renewal of interest in the individual's normative activity. They describe how individuals produce moral justifications in reference to several 'dimensions.' These authors also try to tackle the pluralistic and contextual dimensions of people's evaluations.

Therefore, it seems that an explanation in terms of reasons can give an account of the specificity of morality, without either narrowing it or dissolving it among other objects or social spheres, while in the mean time providing an alternative to Durkheim's solution. One of the issues at stake in moral sociology now is to show how this explanation does not elude the social or collective nature of moral beliefs. From this viewpoint, reflections stemming from the interactionist stream, and notably Simmel's analysis about the objectivity of values, can be a fruitful source of inspiration. The focus put on the individual's point of view also induces moral sociology to renew its empirical research, in a less descriptive way, with a much greater emphasis on individuals' argumentations.

See also: Collective Beliefs: Sociological Explanation; Control: Social; Cultural History; Durkheim, Emile (1858–1917); Industrialization; Integration: Social; Legitimacy; Legitimacy: Political; Moral Education; Morality (East and West): Cultural Concerns; Norms; Parsons, Talcott (1902–79); Religion: Morality and Social Control; Religion, Sociology of; Values, Anthropology of; Values, Development of; Values: Psychological Perspectives; Values, Sociology of; Weber, Max (1864–1920)

Bibliography

Adams C J S 1963 Toward an understanding of inequity. *Journal of Abnormal and Social Psychology* **67**: 422–36

Boltanski L, Thévenot L 1991 *De la justification*. Gallimard, Paris

Boudon R 1995 *Le juste et le vrai*. Fayard, Paris

Boudon R 2000 *The Origin of Values*. Transaction, New Brunswick, NY

Deutsch M 1986 Cooperation, conflict and justice. In: Bierhoff H W, Cohen R L, Greenberg J (eds.) *Justice in Social Relations*. Plenum Press, New York, pp. 3–17

Durkheim E 1915 *The Elementary Forms of the Religious Life*. Allen and Unwin, London

Durkheim E 1920 Introduction à la morale. *Revue Philosophique* **89**: 79–97

Durkheim E 1922 Détermination du fait moral. In: Durkheim E (ed.) *Sociologie et Philosophie*, PUF, Paris [1953 *Sociology and Philosophy*, Cohen, London]

Durkheim E 1960 *The Division of Labor in Society*. Free Press, New York

Durkheim E 1961 *Moral Education*. The Free Press of Glencoe, New York

Geertz C 1984 Distinguished lecture: Anti anti-relativism. *American Anthropologist* **86**: 263–78

Gurvitch G 1960 Sociologie de la vie morale. In: Gurvitch G (ed.) *Traité de sociologie*. PUF, Paris, pp. 137–72

Hardin R 1995 *One for All: The Logic of Group Conflict*. Princeton University Press, Princeton, NJ

Homans G 1974 *Social Behavior: Its Elementary Form*. Harcourt Brace Jovanovitch, New York

Inglehart R 1977 *The Silent Revolution: Changing Values and Political Styles Among Western Publics*. Princeton University Press, Princeton, NJ

Isambert F A, Ladrière P, Terrenoire J-P 1978 Pour une sociologie de l'éthique. *Revue Française de Sociologie* **XIX**: 323–39

Kellerhals J, Coenen-Huther, Modak M 1988 *Figures de l'équité. La construction des normes de justice dans les groupes*. PUF, Paris

Lerner M J, Lerner S C 1981 *The Justice Motive in Social Behavior*. Plenum Press, New York
Oberschall A R 1994a Rational choice in collective protests. *Rationality and Society* **6**: 79–100
Oberschall A R 1994b Règles, normes, morale: émergence et sanction. *L'Année Sociologique* **44**: 357–84
Pareto V 1935 *The Mind and Society*. Harcourt Brace, New York
Riffault H 1994 *Les valeurs des Français*. PUF, Paris
Runciman W G 1966 *Relative Deprivation and Social Justice*. Routledge & Kegan Paul, London
Stoetzel J 1983 *Les valeurs du temps présent: une enquête Européenne*. PUF, Paris
Tocqueville de A 1972 *Democracy in America*. Knopf, New York
Weber M 1930 *The Protestant Ethic and the Spirit of Capitalism*. Allen and Unwin, London
Weber M 1949 *The Methodology of the Social Sciences*. The Free Press, New York
Weber M 1968 *Economy and Society*. Bedminster Press, New York
Wilson J Q 1993 *The Moral Sense*. Macmillan, New York

E. Betton

Morality (East and West): Cultural Concerns

1. Comparison of Three Methodologies

There are three main approaches in the social and behavioral sciences to the comparative study of moral reasoning and conduct in Western and Eastern societies: developmental psychology, anthropological ethnography, and cultural psychology. These three approaches result in different understandings of morality in the West and the East. A comparison of the methodologies of these three approaches is offered before turning to a more detailed examination of research on Western and Eastern moral systems.

1.1 Developmental Psychology

Kohlberg's (1981) theory of moral development and research methods were widely adopted by child psychologists in the 1970s and 1980s to study the cognitive development of moral reasoning and moral judgment in children of varying ages in different cultures. Based on the assumption of cross-cultural universality in moral development, Kohlberg claimed that 'All individuals in all cultures use the same basic moral categories, concepts, and principles, and all individuals in all cultures go through the same order on sequence of gross stage development, though they vary in rate and terminal point of development' (Kohlberg 1971, p. 175). His theory consists of six universal stages of cognitive development in moral judgment.

Kohlberg's theory adopts the perspective of evolutionism, which is based on a normative model and classifies a great diversity of moral discourse from a multitude of cultures into several categories arranged in a formal sequence. According to his theory, the development of moral reasoning follows a universal invariant sequence towards the goal of thinking with universal ethical principles.

A series of empirical studies have been conducted using Kohlberg's standardized moral dilemmas and scoring manuals. After careful examination of the literature, Snarey (1985) found that because stage skipping and stage regressions were rare, progress from Stage 1 to Stage 3 or 4 was virtually universal, but that the presence of Stage 4/5 or 5 was extremely rare in all populations. Nearly all samples from Western middle-class groups and 91 percent of Eastern urban populations exhibited some Stage 6 principled reasoning. However, no tribal or folk cultural groups of the non-Western world showed any postconventional thinking. Postconventional thinking upholds individual rights or universal ethical principles.

In addition, many of the moral reasoning data collected in collectivist or communalistic societies of the non-Western world either could not be scored according to the standardized manual, or could not be explained in the context of Kohlberg's theory. Such anomalistic facts have been used to challenge Kohlberg's theory. Critics charge that the theory was formulated from the perspective of Western ideologies of rationalism, individualism, and liberalism, and that it therefore assigns the values of male, white, American intellectuals as the end point of moral maturation. The theory's ethical objectivism ignores the moral discourse prevailing in many non-Western societies, and fails to recognize their substantial ethical philosophies (Weinreich-Haste and Locke 1983). Imposing Kohlberg's theory as well as his scoring system on the moral reasoning of non-Western peoples may result in systematic bias originating from Western ethnocentrism.

Gibbs (1979) tried to revise Kohlberg's theory in terms of Piaget's phylogenetic perspective. According to Piaget, human intelligence is a holistic phenomenon encompassing social, moral, and logico-physical aspects. Its development follows a standard sequence that is predetermined by epigenetic factors like those of other species. Along with completion of physical maturation and the singular achievement of intellectual development over the course of the adolescent years, there is a progressive ability for people at the highest stage of standard development to reflect on the very conditions of their existence in the world. Reflection on one's own existence may motivate one to define a moral theory to justify one's basic moral

principles from a particular standpoint that is either inside or outside the society.

Cultures show wide variety in their philosophies of morality. These normative philosophies may provide material for second-order thinking or meta-ethical reflection when individuals try to define their own moral theories. It is impossible for individuals to eliminate cultural influences from the contents of their own moral reasoning. Considering this fact, a better way to conduct cross-cultural research on morality would be to use ethnographic analysis to describe the patterns of moral conduct and moral thinking that consistently emerge in the daily social life of people in a particular culture.

1.2 Anthropological Ethnography

Researchers who are interested in studying the ethnography of morality usually adopt the viewpoint of interpretative anthropology, which assumes that culture is an organized collection of symbols with the major function of maintaining social order for normal operation. As derivatives of the cosmology of a given culture, the concepts of morality and personhood, as well as other symbolic systems, can best be understood in the holistic context of that culture.

Many anthropologists have used the ethnographic approach to study indigenous religions, cosmology, personhood, and concepts of morality in non-Western societies (e.g., Edel and Edel 1959, Howell 1997). They tend to adopt the native's point of view to interpret the moral ideas, moral discourse, and moral practices in a particular society. Instead of relying solely on philosophical analysis and anthropological records, they pay special attention to the dynamic interplay between abstract moral principles and empirical facts.

Researchers following this approach tend to adopt the viewpoint of relativism for depicting the uniqueness of each culture. Paradoxically, members of a cultural community tend to insist on the viewpoint of absolutism: They may believe that their moral standard is superior to that of other cultural groups, and use it as a criterion for defining what is meant by *person* or for differentiating their own group from outsiders. Those who violate their moral standards might be condemned as *not a person*. The process of dehumanization may lead to severe interpersonal or even intergroup conflict in extreme cases.

1.3 Cultural Psychology

Cultural psychologists investigate the cultural foundation of psychological diversity in self-organization, cognitive process, emotional function, and moral reasoning among people living in different cultures. In order to understand native peoples' symbolic and behavioral inheritance in terms of their shared framework of interpretation, cultural psychologists also adopt an ethnographic approach, but with the perspective of pluralism. Pluralism assumes that people all over the world are similar in their psychological structure and function, although aspects of their knowledge, thinking, feeling, values, needs, and behavior are activated by the experience of living in a particular cultural community. Cultural psychologists advocate the slogan: 'one mind, many mentalities; universalism without uniformity' (Shweder et al. 1998, p. 871). They work to develop universal theoretical frameworks for interpreting the thinking and behavior of people in different cultures. Although it would be impossible to depict all moral systems in the world, it is possible to describe the sharp contrasts between moral systems of the East and West with the conceptual schemes provided by cultural psychologists.

2. Morality

Cultural psychologists suggest that three main domains of moral discourse exist in human society: the ethics of autonomy, the ethics of community, and the ethics of divinity (Shweder et al. 1997). Because the moral experiences of human beings are so complicated and the recognition of reality so restricted, even though all moral systems contain elements of all three of these domains, any particular moral system will tend to emphasize a particular domain of moral discourse.

2.1 Morality of the West

The contemporary Western moral system is characterized by the ethics of autonomy, which can be traced to the Christian religious tradition. Christianity assumes that the world was created by God, who also created humankind in accordance with His own image. Since all humans were created in God's image, all humans are equal. Every person has the essence of humanity and should obey the divine commandments declared by God.

This concept of person and its underlying discourse of divinity changed drastically after the Renaissance in Europe. The Cartesian philosophy of dualism detached individuals from the outer world where they find themselves. Rationalism emphasized an individual's free will and subjective reasoning. The expansion of modern capitalism also made significant contributions to the rise of individualism.

The Western concept of self formed in the cultural milieu of individualism requires a dualistic view of the world. It draws sharp distinctions between mind and matter, cognition and affection, self and society,

nature and culture, and so on. Self is conceptualized as 'a dynamic center of awareness, emotion, judgment and action organized into a distinctive whole and set contrastively both against other such wholes and against its social and natural background' (Geertz 1974, p. 225). The inner life of the self, with its spontaneity, privacy, uniqueness, genuine feeling, and constancy, has a clear-cut boundary that separates it from the outer social world constituted by mask, role, rule, or context.

In the Western view of personhood, the individual is conceived of as prior to and more fundamental than society, which is treated as an artificial construction designed mainly to satisfy human needs. When a social order can no longer serve the function of realizing this end, it should be redesigned and reconstructed to restore its function. Based on this premise, Western morality is characterized by an emphasis on the rights of individuals. According to Dworkin (1977), all moral codes encompass personal rights, personal duties, and social goals, but they may differ in the priority given to these three categories. Western morality is premised on the concept of natural rights, rather than on natural duties or social goals.

Because individuals are construed as quasi-sacred and absolute, their legitimate demands should be carefully respected. The individual is a monad, and every human group is made up of monads with the same human nature. Individuals' free will and right to liberty are natural, fundamental, and inescapable. By the same token, the duty to respect others' rights is equally natural and inescapable. One's rights are restrained by the identical rights of other individuals, and everybody is encouraged to strive for the rights of all human beings to the greatest overall liberty (Dumont 1986).

The realization of negative duties of omission (e.g., not killing, not cheating, not stealing, etc.) is considered to be the prerequisite of a moral agent. Violating moral codes of this category might be condemned as evil. In contrast, the practice of the positive duties of commission is subject to an individual's right of choice. Individuals who decide to undertake an altruistic act may be admired for their virtue. As a consequence, a certain state of tension is frequently aroused between social demand and the individual's desire for freedom of choice. The autonomous view of self tends to entail the belief that performing social obligations in spite of individual rights may have the negative effect of restricting one's own personal liberties. Even if one can justify working for the benefit of another person, the act of undertaking this responsibility is a resolution of personal choice, rather than an enforced social demand.

As indicated by Gilligan (1982), in contrast to the Kohlbergian ethics of justice for males, the moral reasoning of Western females can be characterized by the ethics of caring and responsibility. However, both the superogatory view of morality held by Kohlberg and the morality-of-caring framework developed by Gilligan are bound to the Western culture of individualism. Both stress freedom of choice and individual responsibility. Commitments to interpersonal caring are viewed as matters of personal decision, rather than as obligations to a particular social target (Miller 1994).

2.2 *Morality in the East*

In sharp contrast to the individual-centered morality of the West, contents of moral systems in the Oriental world, no matter whether they originated from Confucianism, Hinduism, Islam, or Buddhism, can be described as duty-based. Unlike typical Western dualism, most Oriental philosophies advocate a monistic view of origin. That is, one's life is construed as an inseparable part of the world in which one finds oneself, not as an independent entity standing in opposition to the outer world. For example, traditional Chinese philosophy conceives of the formation of all things in the universe as the result of dynamics between the two opposing forces of *ying* and *yang*. A person's life is also produced from the unification of *ying* (female) and *yang* (male). In Confucianism, a family is conceived of as the body of a human being; each role in the family represents a distinct part of the body. For instance, the relationship between parent and child is like that between flesh and bone, while brothers are considered to be the four limbs (Hwang 1999).

The Hindu view of life is also premised on the functional interrelatedness of the universe. From the earliest times of *Rgveda* (about 1500 BC), the Hindus have believed that *Rta* is the governing dynamic principle underlying functional equilibrium of the cosmic order. The famous *purushshukta* of *Rgveda* assumes the functional relatedness of all living organisms. Any existing phenomena may attain a state of *Rta* when each of its parts remains in its proper place and functions in its own law of activity. Humans, animals, and plants are interrelated. They all are part of the cosmic immanent life force. The individual functioning of one in coordination with the individual functions of others results in collaborative equilibrium of the whole universe (Heimann 1964). In the same vein, Hindus conceptualize their society as a huge cosmic organism in which certain functions are allotted to each of the four major parts: the head, arms, trunk, and feet. Each of these four represents one of the four castes of Hindu society. Remaining in the proper function fulfills the goal of smooth collaboration of the whole society.

Instead of drawing the boundary of an individual around the immediate surface of one's physical body, most Oriental philosophies conceive of this boundary as expandable to include one's family or the whole

society. Followers of Oriental philosophy tend to believe that external forces predetermine the social order. Role relationships within one's primary group are especially unchangeable through one's own efforts. Therefore, Oriental moral systems endow mandatory features not only to negative duties of omission, but also to some positive duties of commission. Examples include filial piety in Chinese society, repaying *on* (favors or kindness) in Japanese society, and *utang na loob* (debt inside oneself or a debt of gratitude) in Filipino society. These kinds of positive duties or categorical imperatives urge individuals to fulfill obligations to specific targets in the community. These obligations ensure the smooth functioning of society, rather than protect the rights of individuals. In Dworkin's (1977) conceptual framework, Eastern moralities are based on duties or group goals rather than rights.

Compared with Western ethics of autonomy, ideas about individual rights are relatively uncommon in traditional Eastern moralities. Even moral discourse arguing for negative duties in Eastern moral systems are proposed not on the premise of protecting individual rights, but in consideration of maintaining social order or for religious reasons. For example, the act of killing others violates the principle of benevolence (*ren*). *Ren*, the requirement that all people carry out their moral duty to significant others in their intimate society, is the most fundamental moral rule for maintaining psychosocial homeostasis in Chinese society. Buddhists must adhere to five requirements: do not kill, steal, lie, be lustful, or taste intoxicants or meat. Violation of any of these requirements may result in *karma* bringing automatic retribution in a later life. Muslims also believe that the present life is short and temporary, while the future life is eternal and everlasting. Those who devote themselves to charity, especially those who sacrifice themselves for the almighty Allah, may ascend to Heaven. Those who indulge in evil doings will descend to Hell. This kind of moral discourse indicates that Eastern moral systems entail the ethics of community or the ethics of divinity, rather than the ethics of autonomy.

Moral discourse supporting the ethics of community or divinity in the Oriental world as revealed by sages or prophets are recorded in classics such as *The Analects* of Confucianism, or religious scriptures such as the *Koran* of the Moslem faith, or the *Upanisads* of Hinduism. Intellectuals who are able to interpret those classics or scriptures, like Confucian scholars in traditional China, Brahmans in India, priests in Islamic countries, and monks in Buddhist societies, have occupied high positions in traditional and even contemporary Oriental societies. Their position is similar to that of priests in the Christian world before the European religious Reformation movement.

In contrast, the modern Western view of morality advocates a position of secularism, which assumes that all moral agents are morally equivalent. Anyone can discover the natural law of morality without the assistance of revelations by a prophet or sage.

2.3 Global Ethics

Since the collapse of the Soviet Union and East European communist regimes at the end of the 1980s, most countries have been incorporated into the world economic system of capitalism. The rapid expansion of opportunity for intercultural contact and exchange between people from the East and West has facilitated the formation of such concepts as multiculturalism and global ethics in the age of globalization. As part of the Universal Ethics Project sponsored by UNESCO, a group of international scholars of philosophy have proposed a Universal Declaration of Human Responsibilities and a list of global ethics (Kung and Kuschel 1993). The list consists of four categorical imperatives which are proposed on the basis of a fundamental demand that every human being must be treated humanely:

You shall not kill! Or in positive terms: Have respect for life!

You shall not steal! Or in positive terms: Deal honestly and fairly!

You shall not lie! Or in positive terms: Speak and act truthfully!

You shall not commit sexual immorality! Or in positive terms: Respect and love one another!

A careful examination of the list indicates that it consists of negative and positive duties already present in most cultures of the world, rather than positive duties which are specific to a particular culture in either the East or West.

See also: Ethical Behavior, Evolution of; Ethics and Values; Justice and its Many Faces: Cultural Concerns; Moral Development: Cross-cultural Perspectives; Moral Education; Moral Reasoning in Psychology; Norms; Religion: Morality and Social Control

Bibliography

Dumont L 1986 *Essays on Individualism: Modern Ideology in Anthropological Perspective*. University of Chicago Press, Chicago

Dworkin R 1977 *Taking Rights Seriously*. Harvard University Press, Cambridge, MA

Edel M, Edel A 1959 *Anthropology and Ethics*. Charles C. Thomas, Springfield, IL

Geertz C 1974 From the native's point of view: on the nature of anthropological understanding. In: Basso K, Selby H (eds.) *Meaning in Anthropology*, 1st edn. University of New Mexico Press, Albuquerque, NM

Gibbs J C 1979 Kohlberg's moral stage theory: a Piagetian revision. *Human Development* **22**: 89–112

Gilligan C 1982 *In a Different Voice: Psychological Theory and Woman's Development*. Harvard University Press, Cambridge, MA
Heimann B 1964 *Facets of Indian Thought*. Schocken Books, New York
Howell S 1997 *The Ethnography of Moralities*. Routledge, New York
Hwang K K 1999 Filial piety and loyalty: two types of social identification in Confucianism. *Asian Journal of Social Psychology* **2**(1): 163–83
Kohlberg L 1971 From is to ought: how to commit the naturalistic fallacy and get away with it in the study of moral development. In: Mischel T (ed.) *Cognitive Development and Epistemology*. Academic Press, New York, pp. 151–235
Kohlberg L 1981 *Essay on Moral Development Vol. 1: The Philosophy of Moral Development,* 1st edn. Harper and Row, San Francisco
Kung H, Kuschel K-J (eds.) 1993 *A Global Ethics: The Declaration of the Parliament of the World Religions*. SCM Press, London
Miller J G 1994 Cultural diversity in the morality of caring: individually oriented versus duty-based interpersonal moral codes. *Cross-Cultural Research* **28**(1): 3–39
Shweder R A, Goodnow J, Hatano G, LeVine R, Markus H, Miller P 1998 The cultural psychology of development: one mind, many mentalities. In: Damon W (ed.) *Handbook of Child Psychology: Vol. 1. Theoretical Models of Human Development*. J. Wiley, New York, pp. 865–937
Shweder R A, Much N C, Mahapatra M, Park L 1997 The big three of morality (autonomy, community, divinity) and the big three explanations of suffering. In: Brandt A M, Rozin P (eds.) *Morality and Health*. Routledge, New York, pp. 119–69
Snarey J R 1985 Cross-cultural universality of social–moral development: a critical review of Kohlbergian research. *Psychological Bulletin* **97**(2): 202–32
Weinreich-Haste H, Locke D (eds.) 1983 *Morality in the Making: Thought, Action, and the Social Context*. Wiley, Chichester, UK

K.-K. Hwang

Morphological Case in Linguistics

Case, as traditionally understood, refers to a system of inflection, usually in the form of suffixes, that marks the relation of nouns in a clause. In the Latin sentence *Calpurnia Clodiam vīdit* 'Calpurnia saw Clodia' the fact that Clodia rather than Calpurnia is the one who is seen is signalled by the *-m* suffix on Clodia, rather than by word order as in English. This *-m* is a case suffix.

1. Case, Case Form, and Case Marker

A morphological case system is a system of inflection that marks dependents for the type of relation they bear to their heads. Typically, case marks the relationship of a noun to a verb at the clause level or of a noun to a preposition, postposition, or another noun at the phrase level. Consider the following Turkish sentence:

Table 1
Turkish case system

nominative	*adam*
accusative	*adamı*
genitive	*adamın*
dative	*adama*
locative	*adamda*
ablative	*adamdan*

(1) *Ali* *adam-a* *elma-lar-ı* *ver-di*
Ali.NOM man-DAT apple-PL-ACC give-PAST.3SG
'Ali gave the apples to the man.'

The suffix *-ı* indicates that *elmalar* is the direct object of the verb *vermek* 'to give' and the suffix *-a* indicates that *Adam* is the indirect object. The suffix *-ı* is said to be an accusative (or objective) case marker and the word form *elmaları* is said to be in the accusative case. The suffix *-ı* also indicates that *elmalar* is specific, since in Turkish only specific direct objects are marked as accusative. The suffix *-a* is a dative case marker. *Ali* contrasts with *elmaları* and *adama* in that it bears no overt suffix. It is said to be in the nominative case, which is this sentence indicates the subject.

Turkish has a system of six cases as in Table 1. The locative marks location as in *Istanbul-da* 'in Istanbul,' and the ablative indicates 'from' or 'out of' as in *Ankara-dan* 'from Ankara.' The genitive is used in phrases like *adam-ın ev-i* 'the man's house' where *ın* corresponds to '*s* in English. The suffix *-i* on *ev* 'house' is a third person possessive form translatable as 'his,' 'her,' or 'its'. In Turkish 'the man's house' is literally 'the man's, his house'.

The word forms displayed in Table 1 make up a *paradigm*, that is, they constitute the set of case forms in which the lexeme *adam* can appear. In Turkish one could say that there is only one paradigm in that a single set of endings is found for all nouns, although these suffixes do undergo some phonologically conditioned changes such as undergoing vowel harmony and the like. The locative, for instance, has the form *-da* following stems with back vowels and *-de* following stems with front vowels. A distinction needs to be made between *cases* (of which there are six in a system of oppositions), and the *case markers* or *case forms* through which the cases are realised. A case marker is an affix and a case form is a complete word. In Turkish the case affixes can be separated from the stem, so it is possible to talk about case markers. In some languages, however, it is not possible to isolate a case suffix, so it is necessary to talk in terms of the various word forms that express the cases of the stem. These are case forms.

Table 2
Latin case paradigms

	1	2		3a	3b
	ā-stems	*O*-stems		cons. stems	*i*-stems
	feminine	masculine	neuter		
	domina	*dominus*	*bellum*	*cōnsul*	*cīvis*
	'mistress'	'master'	'war'	'consul'	'citizen'
		singular			
Nominative	*domina*	*dominus*	*bellum*	*cōnsul*	*cīvis*
Vocative	*domina*	*domine*	*bellum*	*cōnsul*	*cīvis*
Accusative	*dominam*	*dominum*	*bellum*	*cōnsulem*	*cīvem*
Genitive	*dominae*	*dominī*	*bellī*	*cōnsulis*	*cīvīs*
Dative	*dominae*	*dominō*	*bellō*	*cōnsulī*	*cīvī*
Ablative	*dominā*	*dominō*	*bellō*	*cōnsule*	*cīvī*, *cīve*
		plural			
Nominative	*dominae*	*dominī*	*bella*	*cōnsulēs*	*cīvēs*
Vocative	*dominae*	*dominī*	*bella*	*cōnsulēs*	*cīvēs*
Accusative	*dominās*	*dominōs*	*bella*	*cōnsulēs*	*cīvīs*, *cīvēs*
Genitive	*dominārum*	*dominōrum*	*bellōrum*	*cōnsulum*	*cīvium*
Dative	*dominīs*	*dominīs*	*bellīs*	*cōnsulibus*	*cīvibus*
Ablative	*dominīs*	*dominīs*	*bellīs*	*cōsulibus*	*cīvibus*

2. *Case Marking Fused with Number and Gender*

The traditional notion of cases and the associated terminology were developed on the basis of Ancient Greek and Latin where it is not possible to separate *number* marking from case marking. This means separate paradigms for the two number categories, singular and plural. Moreover, there are different case/number markers for different stem classes. In Latin five such classes are recognised, or five declensions as they are called. The three major declensions are illustrated in Table 2.

In Latin there is also a three-way gender distinction: masculine, feminine, and neuter. With a few exceptions male creatures are masculine and females feminine, but inanimates are scattered over all three genders (though almost all neuter nouns are inanimate). There is a partial association of form and gender in that *ā*-stems are almost all feminine and *o*-stems mostly masculine (except for a subclass of neuters represented by *bellum* in Table 2). This means that there can be fusion of gender, number, and case. The point is illustrated in Table 2 where we have *domina* 'mistress (of a household)' illustrating feminine *ā*-stems and *dominus* 'master (of a household),' which is based on the same root, representing masculine *o*-stems. As can be seen from Table 2 the word form *domina* simultaneously represents nominative case, feminine gender, and singular number, *dominum* represents accusative case, masculine gender, and singular number, and similarly with other word forms.

Where there is more than one paradigm, it frequently happens that distinctions made in one paradigm are syncretized or neutralized in others. In the traditional descriptions a case is established wherever there is a distinction for any single class of nominals. Informally one might consider that a case is a row in a set of paradigms like those in Table 2. Dealing in cases ('rows') facilitates the description of the relation of form and function. A particular case can be said to have one or more functions without reference to the details of the forms.

The functions of the cases in Latin, and in Ancient Greek, are important in that they have served as a model for applying case labels in other languages such as Turkish. In Latin the nominative encodes the subject and the accusative the direct object. The dative marks the indirect object of a few three-place verbs such as *dare* 'to give' and the complement of a few two-place verbs including *crēdere* 'to believe,' *nōcere* 'to be harmful to,' and *subvenīre* 'to help.' The genitive is used mainly to mark noun phrases as dependents of nouns, in particular to mark the possessor: *cōnsulis equus* 'the consul's horse.'

The ablative in Latin represents the syncretism or merger of three once-distinct cases: the ablative, the locative, and the instrumental so it is not surprising then to find that it expresses source, location, and instrument.

The vocative is anomalous in that it does not mark the relationship of dependents, but is used as a form of address as in *Quō vādis*, *domine*? (whither go.2SG lord. VOC) 'Where are you going, master?' However, the vocative belongs with the other cases since its marking belongs to the same system as the unequivocal cases.

Table 3
Finnish local cases

	LOCATION	SOURCE	DESTINATION
Ø	*-na* essive	*-ta* partitive	*-ksi* translative
INTERIOR	*-ssa* inessive 'in'	*-sta* elative 'from(inside)'	*-(h)Vn, -sVVn* illative 'into'
EXTERIOR	*-lla* adessive 'at'	*-lta* ablative 'from(outside)	*-lle* allative 'to(wards)'

Latin uses a system of prepositions in addition to the case system. Some prepositions, including all those expressing destination, such as *ad* 'to' govern the accusative (*ad urbem* 'to the city'), while others such as *ex* 'out of, from' govern the ablative (*ex urbe* 'from the city').

Table 4
Yuwaalaray case marking

nominative	*dhayn* 'man'	*ngali* 'we two'
ergative/instrumental	*dhayndu*	*ngali*
accusative	*dhayn*	*ngalinya*
genitive	*dhayn.gu*	*ngalingu*
dative	*dhaynda*	*ngalingunda*
ablative	*dhayndi*	*ngalingundi*

3. *The Range of Case Systems*

Case systems differ in terms of the number of distinctions they make and range from a minimum of two as in some Iranian languages, which have a nominative and an oblique, to figures like 15 in Finnish (see Table 3).

Case systems also vary in terms of the type of distinction they make in a given domain. While many languages are like Latin in having one case for the subject, namely the nominative, and another case for the direct object (the accusative), some languages including Basque, northeast Caucasian languages, and many Australian languages have one case for the subject of a transitive verb, which is called the ergative, and another case to cover the subject of an intransitive predicate and the direct object. This case is generally morphologically unmarked, that is, there is no form. Some writers call this case the absolutive; others call it the nominative. The following illustration is from the Australian language Kalkatungu where the ergative is realised by the suffix *-yu* and the absolutive/nominative is unmarked. Note, however, that with the cross-referencing clitics that are attached to the purposive auxiliary a common from represents the subject and a distinct form the object.

(2) *Nhutu* *a=nhurr* *ingka?*
You.NOM PURP = 2PL.SUBJ go
'Are you mob going?'

(3) *Nhutu-yu* *a=nhurr* *nuwa* *puyu?*
you-ERG PURP = 2PL.SUBJ see 2DU.NOM
'Are you mob going to see them two?'

(4) *Nhutu* *a=kutu* *nuwa* *puyu-yu*
you.NOM PURP = 2PL.OBJ see 2DU-ERG
'They two want to see you mob.'

In a number of languages, particularly in Australia, accusative and ergative marking are found in the same language. As pointed out by Silverstein (1976), there tends to be some complementarity with ergative marking confined to nouns and accusative to pronouns or to pronouns plus certain classes of noun such as kinship nouns, personal names or in some instances all human nouns. The following paradigms from another Australian language, Yuwaalaray, show ergative marking with nouns and accusative with pronouns (see Table 4).

Every language has to provide for the expression of a large number of *local notions*. This can be done via adpositions as with English prepositions such as *on*, *to*, *over*, and *under*. Where a number of local notions are expressed via the case system the number of cases naturally rises. In Finnish, for instance, the notions of location, source, and destination can combine with the notions of interior and exterior to produce a set of nine local cases as illustrated in Table 3. Originally there were two systems (location, destination, and source) and (interior, exterior, and neither), but the markers have partially fused. The functions are no longer entirely local either. The partitive is used for the patient if it represents part of a whole or an indefinite quantity, or if the action is incomplete, or if the polarity of the clause is negative. The translative refers to the endpoint of a transformation as in *She turned the frog into a prince*.

Other cases commonly encountered include the comitative, expressing accompaniment (*I went with them*) and instrumental expressing the instrument used to carry out an action (*She wiped the screen with a cloth*).

In the Uralic language a case called *abessive* is found. It means 'without', 'not having.' In Finnish, for instance, *rahta-tta* is (money-AB) 'moneyless.' This case is also found in Australia where it is matched by a 'having' case sometimes called the *proprietive*.

4. Classifying Cases

The Greeks made a primary distinction between the nominative and the other cases, collectively called the oblique cases. This makes sense when one considers the nominative was the form used in isolation whereas the other forms generally required a particular syntactic context. In some languages there is a formal distinction between the nominative and the oblique cases inasmuch as there is a special stem for the obliques. This is the situation in many Dravidian and northeast Caucasian languages, for instance. In most languages the nominative bears no marking as in Turkish (Table 1), but consists of the bare stem. Indo–European case languages are unusual in having a marked nominative in most paradigms (Table 2).

Another distinction that is often made is between grammatical (or syntactic) cases and semantic (or concrete) cases. The grammatical cases are traditionally taken to include the nominative and accusative and often the genitive, and should include the dative and the ergative. The basis for the distinction is often not made clear, but if the criterion for classifying a case as grammatical or syntactic is that it expresses a purely syntactic as opposed to semantic relation, then the dative and ergative should certainly be included since the dative encodes the indirect object and the ergative encodes the 'transitive subject.'

If the distinction between grammatical and semantic cases were to be clearcut, the grammatical cases would encode only purely syntactic relations and the semantic cases would encode only homogeneous semantic relations such as location or source. However, it is common for a syntactic case to encode a semantic relation or role that lies outside of whatever syntactic relation it expresses. In Latin, for instance, the accusative not only expresses the direct object, it also expresses the semantic role of destination as in *Vādō Rōmam* 'I am going to Rome.' On the other hand, there are situations where the so-called semantic cases encode a purely syntactic relation. This commonly occurs with the passive where the demoted subject is often expressed via a semantic case. In Latin the demoted subject is encoded in the ablative as in *vīsus ā cōnsule* 'seen by the consul.'

It is also convenient for the purposes of making certain generalisations to distinguish the cases that encode the complements (including the subject) of typical one-place and two-place transitive verbs. These are the nominative, ergative and accusative, and they are known collectively as the *core* cases and the others as *peripheral* cases. There is another set of terms in use, namely *direct* and *oblique*. These terms are used in the description of Indo–Aryan languages where the direct case covers subjects and objects.

In some treatments the term *local cases* is used to cover cases designating notions of place such as location, source, destination or path (Table 3).

5. Exponence

The relation between the verb and its dependents, whether complements or adjuncts, may be expressed by an adposition, either a preposition as in English or a postposition as in Turkish. In languages with morphological case systems there are usually adpositions as well as cases. In Latin, as we have seen, there are prepositions, and in Turkish there are postpositions such as *sonra* 'after,' which governs the ablative: *tiyatro-dan sonra* (theatre-ABL from) 'after the theatre.' In such languages there are essentially two systems: an inner, restricted system of case and an outer, more elaborate system of adpositions. In some languages such as Japanese there is just one system of postpositions, which includes markers for subject, direct object, indirect object, and topic, as well as for local notions such as 'from.' In such a language the distinction between case and adposition is essentially neutralized. In the following Japanese example *ga* marks the subject, *ni* marks the indirect object, and *o* marks the direct object.

(5) *Sensei ga Tasaku ni hon o yat-ta*
teacher SUBJ Tasaku IO book DO give-PAST
'The teacher gave Tasaku a book.'

Morphological case is almost always realised as a set of suffixes, though there are instances where there is some mutation of the stem or complete suppletion as with the English pronouns: *I/me*, *she/her*, *we/us*, and also *he/him* and *they/'em/them* where the *-m* is an old dative suffix.

Where categories such as number, gender or possessor are marked as well as case and where these categories are not fused, case marking normally follows these other categories as in Turkish : *adam-lar-ım-la* (man-PL-1SG.POSS-LOC) 'with my men' where the number marking and possessor precede the case marking.

Within the noun phrase there are two common distributions for case marking. Either it occurs on all types of nominal or it is occurs on only one word.

In the former type the distribution is traditionally described in terms of dependents showing case concord or agreement with a noun head. In the following example from Ancient Greek *bíos* is a nominative singular form of a second declension (*o*-stem) masculine noun, the nominative indicating that *bíos* is the subject of the predicate. The definite article and the adjective are in the nominative singular masculine form, their *concord* in case, number and gender

indicating that they are dependents of *bíos*,

(6) *Ho* *aneksétastos* *bíos*
the.NOM.SG unexamined.NOM.SG life.NOM.SG
ou *biōtós* *anthrōpōi*
not livable.NOM.SG man.DAT.SG
'The unexamined life is not livable for man.'

This example also illustrates concord between a predicative adjective (*bīotōs*) and the subject (*bíos*). While concord of case, number, and gender of predicative adjectives is the norm in Latin and Ancient Greek, in many other languages predicative adjectives display no concord. This is the situation in German where the predicative adjective consists of the bare stem and is distinct from the various case forms.

Case concord within the noun phrase is found in the older Indo–European languages and also in Balto–Finnic, Semitic, in some Australian languages, and elsewhere, but it is a minority system. In some languages such as Hungarian the concord does not extend to the adjective and in the Indo-Aryan languages adjectives exhibit fewer case distinctions than nouns. In the more conservative Germanic languages, including modern German, the adjective shows fewer case marking distinctions where a determiner is present.

In most languages case concord does not extend to dependents in the genitive case. However, in Georgian a postposed genitive exhibits concord, which results in double case marking: compare *davit-is mama-s* (David-GEN father-DAT) and *mama-s davit-isa-s* (father- DATDavid-GEN-DAT 'to David's father').

Where case marking occurs on only one word in the noun phrase, it is usually on the final word. In many languages, including a vast number in northern Asia and South Asia, the final word is the head noun in the noun phrase. The following example is from the Dravidian language Kannada:

(7) *Naanu ellaa maanava janaangavannu*
I.NOM all human community.ACC
priitisutteene
love.1SG
'I love all mankind.'

In some languages the final word in the noun phrase is not always the head. This is the situation in various Australian and Amazonian languages, and in Basque: *etxe zaharr-etan* (house old-PL.LOC) 'in old houses.'

6. History

Since bound forms in general derive from free forms, it is not surprising that there is evidence of case markers deriving from nouns and verbs, often via postpositions. For instance, the Hungarian illative ('into') case marker *-ba* as in *világ-ba* 'into the world' derives from a noun *bél* 'innards' and the allative case marker *-tharri* in the Australian language Kunggari derives from a verb *tharri* 'to become.'

Case marking may also be lost. This is quite striking in Europe, particularly among the Romance languages. The Latin system of noun inflection illustrated in Table 2 has been lost from the Romance languages except for a reduced system retained in Rumanian. The Germanic languages too have lost case distinctions over the last 2 millennia with English and Afrikaans having lost all case marking on nouns, though German still retains a four-case system: nominative, accusative, genitive, and dative. In languages like English and the Romance languages case distinctions have been retained on pronouns. Case marking can fall prey to phonetic attrition and the availability of prepositions and word order as alternative means of making the relevant distinctions must also be a factor in the loss of case.

See also: Grammatical Gender; Grammatical Number; Morphology in Linguistics

Bibliography

Blake B J 1994 *Case*. Cambridge University Press, Cambridge, U K

Brecht R D, Levine J S (eds.) 1986 *Case in Slavic*. Slavica Publishers, Columbus, OH

Comrie B 1986 On delimiting cases. In: Brecht R D, Levine J S (eds.) *Case in Slavic*. Slavica Publishers, Columbus, OH, pp. 86–105

Mel'čuk I A 1986 Toward a definition of case. In: Brecht R D, Levine J S (eds.) *Case in Slavic*. Slavica Publishers, Columbus, OH, pp. 35–85

Silverstein M 1976 Hierarchy of features and ergativity. In: Dixon R M W (ed.) *Grammatical Categories in Australian Languages*. Australian Institute of Aboriginal Studies, Canberra, Humanities Press, New Jersey, pp. 112–71

B. J. Blake

Morphology in Linguistics

1. Introduction

Morphology is the study of word structure, the way words are formed and the way their form interacts with other aspects of grammar such as phonology and syntax. In many languages words assume extremely complex forms, and propositions requiring a complete sentence in English can be expressed as a single word. In other languages there is virtually no morphology, and notions corresponding to 'plural of' or 'past tense of' are conveyed by separate words. Surveys of the field are found in Carstairs-McCarthy (1992), Matthews (1991), Spencer (1991). Introductions to specific topics can be found in the chapters of Spencer

and Zwicky (1998). The summary given here is expanded in Spencer (2000).

2. The Concept of 'Word'

The ordinary term 'word' corresponds to several distinct notions (see *Word, Linguistics of*). Consider the examples *pick*, *picks*, *picked*, *picking*. In one sense these are four distinct words, but in another sense they are four forms of a single word or 'lexeme' (which we can designate as PICK). The forms *pick*, *picks* are themselves ambiguous between verb forms and noun forms. Thus, there is a sense in which *picks* is one word form but corresponds to two lexemes, the verb PICK and the noun PICK. A more subtle form of ambiguity is illustrated by the lexeme CUT. The form *cut* has two legitimate grammatical descriptions: 'base form of CUT' and 'past tense of CUT.' Thus, *cut* is a single word form of a single lexeme but it represents two distinct 'grammatical words.' Finally, consider the expressions *again* and *a gain*. In normal speech these are completely indistinguishable, yet *a gain* consists of two words, which can be separated with an adjective: *a significant gain*. The two expressions can be called a single 'phonological word.' Normally, we expect a single word form to correspond to a single phonological word, but this is not the case with *a gain*. Another particularly dramatic instance of such a mismatch is illustrated by *Tom's a linguist*. Here *Tom's* is a single phonological word (a single syllable in fact) but corresponds to two distinct words, the noun subject 'Tom' and the main verb 'is.'

3. The Lexeme Concept

The notion of lexeme corresponds to the lexicographer's notion of 'lexical entry' (see *Lexicology and Lexicography*). A lexeme is an abstract concept which unites information about the pronunciation and the meaning of a word, as well as about its grammatical properties. Thus, the lexeme CAT includes a phonological representation (/kat/) and a semantic representation (of whatever it is that 'cat' means). In addition, many would include the information that CAT is of syntactic category 'noun.' In the lexical representation for the verb CUT we would have to include the idiosyncratic information that its past tense/participle forms are irregular. For the (virtually synonymous) verbs STOP and CEASE the lexical representations have to include the information that both take a clause complement whose verb is in the *-ing* form (*They stopped/ceased studying*) and that only *cease* but not *stop* can take an infinitive form (cf. *They ceased to study* but not * *They stopped to study* in the required sense). A lexeme is thus an elaborated sign, a pairing of sound form with meaning (together with other grammatical information).

PICK and WRITE are clearly different lexemes, and each has its set of forms. Consider the words *unpick* or *rewrite*. These are not forms of *pick*, *write*, rather they are newly derived lexemes, formed by adding *un-*, *re-* to the beginning of the base form of the original verb. This adds some component of meaning to the old lexeme. The creation of new, complex lexemes by morphological devices such as this is derivational morphology. The formation of distinct forms of a single lexeme is inflectional morphology. This is a traditional and useful distinction, though one which can be very difficult to draw in practice, even in a language with relatively little morphology like English.

Derivational morphology often changes the part of speech of a word, so that we can create nouns from verbs (*drive—driver*), verbs from nouns (*flea—deflea*), adjectives from nouns (*dirt—dirty*), and so on (see *Word Classes and Parts of Speech*). It is often said that inflection cannot change the part of speech, but this is misleading. Thus, verbs often have participle forms and these behave very much like adjectives, serving to modify nouns, and in many languages agreeing with the noun just like any other adjective (English examples would be forms in *-ing* or *-ed* as *hot running water*, *freshly picked apples*). However, participles are generally regarded as forms of the verb, because they are not seen as creating a new lexeme. However, this is where the traditional distinction becomes difficult to draw, because of equivocation in traditional grammar over the meaning of inflected forms. For instance, it is generally said that *picked* means 'past tense of PICK.' But if that were true, *picked* would have a different meaning from *picks*, which means 'present tense of PICK (with 3rd singular subject).' But now it looks as though inflected word forms undergo a shift of meaning, which should make them into new lexemes.

The solution is to say that inflected forms do not convey meanings directly, but rather are associated with (abstract) features, such as [TENSE: Past] or [NUMBER: Plural]. These features are then interpreted semantically in various (sometimes complex and indirect) ways. Derivational processes, on the other hand, can be regarded as signs, adding their meaning to that of the base lexeme. This is not uncontroversial, however, and some linguists would argue that derivational meanings are every bit as indirect as inflectional meanings, and they would therefore not distinguish inflection from derivation.

For a given language, then, it may be useful to distinguish the inflectional functions (features) from the derivational meanings, though in practice the distinction is not always easy to draw. In many languages alternations such as *drive—driver* are extremely regular, to an extent that they might best be thought of as inflectional. On the other hand, in some languages only a handful of nouns have a plural form (Mandarin Chinese is an example), and this would suggest that plurals in such languages are cases of

derivational morphology rather than inflection. Some argue that we must therefore distinguish two types of inflection, one which is purely featural ('contextual inflection'), such as agreement inflections, and one which is given a direct semantic interpretation ('inherent inflection'), such as plural or past tense inflections. However, it remains very difficult to distinguish inherent inflection from derivation.

4. Morphological Form

Words in English are frequently complex, consisting of identifiable units. One case is that of the compound word, such as *snowball*, *blackbird*, *pickpocket*. These consist of two or more words combined. However, they are not the same as syntactically formed phrases. This is clear from comparing the compound word *blackbird* with the phrase *black bird*. The stress in the compound falls normally on the first element, *BLACKbird*, while in the phrase it falls on the second, *black BIRD*. It is often the case that the meaning of the compound cannot be predicted solely from the meaning of the component words. Thus, a black bird has to be black (even if it's a swan), but a blackbird could be brown, and a raven is not the same as a blackbird. *Snowball* and *blackbird* are kinds of ball and bird respectively. The words *ball*, *bird* are said to be the head of the compound (endocentric compounding). However, a compound such as *pickpocket* is not a kind of pocket and such a compound is not therefore headed (exocentric compounding). Compounding is very productive in English and it is possible to create new compounds in speech. Thus, a book describing the history of tea drinking could be described as the *tea book*. Compounding is also recursive, meaning that the process can feed itself. Thus, the author of the tea book would be the *tea book author*.

Languages differ in the compounds they permit. English has many compounds of the form *train driver* in which *train* appears to be the object of the verb *drive* in *driver* (so-called synthetic compounds). However, it is not normally possible to simply compound a noun with a verb: * *Tom train-drives*. However, just such a process, where the object is 'incorporated' into the verb, is very common in languages outside the Indo-European group.

French has very few compounds of the type *snowball* or *blackbird* but has a great many which resemble *pickpocket*. *Pickpocket* is clearly related to the phrase 'to pick pockets' and French also has a great many other fixed expressions which seem to derive from syntactic phrases, as in *le va-et-vient*, literally 'the comes-and-goes,' 'comings and goings.' This use of 'frozen' bits of syntax is a major way of forming new lexemes in a great many languages.

Compound words are intriguing in that they behave like a single word syntactically even though they are composed of more than one word (both in the sense of lexeme and word form). In many cases they can be difficult to distinguish from phrases or idioms. Other types of complex word are more obviously single word forms. Thus, *picks*, *picked*, *picking* clearly contain a base part or root followed by a suffix. Likewise, we can segment words such as *re-write*, *un-happy*, *pre-war* into a prefix and root. Affixation of this sort is the commonest device for conveying inflection and derivation. In addition to prefixes and suffixes, we sometimes find that simultaneous prefixation and suffixation convey a single grammatical meaning (circumfixation), as with the German past participle ge-…-*t*: In many languages an affix can appear inside the root rather than at its edge. Thus, in Tagalog, a verb root beginning with a vowel takes a prefix *um*- in one of its inflected forms *aral* 'teach' *um-aral*. However, when the verb begins with a consonant the prefix appears as an infix, breaking up the integrity of the root: *sulat* 'write' *s-um-ulat*, *gradwet* 'graduate' *gr-um-adwet*. Another, very common, process is that of reduplication. In many languages a grammatical meaning is signalled by repeating the word, as in Indonesian *anak* 'child' *anak-anak* 'children,' *siapa* 'who,' *siapa-siapa* 'whoever.' In other cases only part of the root is repeated, as in Tagalog *sulat* 'write' *su-sulat* 'will write.'

The lexemes *sing* and *song* are related but by change of the root vowel (ablaut) rather than affixation. Moreover, the verb has past tense and past participle forms *sang*, *sung* with yet different vowels. In Semitic languages such as Arabic and Hebrew it is very common for inflected words to be formed by changing all the root vowels. In the plural of *leaf*, *knife* we see an alternation $f \sim v$: *leaves*, *knives*. In many languages alternations in the final or initial consonant are common and can be used without any affixation to signal grammatical meanings or functions (consonant mutation). The Celtic languages provide well-known examples of this, and other examples of changes in the sound shape being used to convey inflectional or derivational functions can be found in handbooks such as Spencer (1991).

5. The Morpheme Concept

The plural ending in *dogs* is pronounced /z/, in *cats* it is /s/, while in *horses* it is /əz/ (where the symbol /ə/ represents the schwa, or reduced vowel; see *Phonology*). The difference in pronunciation is due to the final sounds of the roots *dog*, *cat*, and *horse*. Despite this difference it is clear that in one sense we are dealing with just one (abstract) suffix with three variant pronunciations. The variation is conditioned solely by the phonology of the root to which the suffix attaches and is a predictable consequence of English grammar. The abstract single plural ending is known as a

Table 1
Verb morphology in Chukchee

Subject	Object		Subject	Object	
1sg	1sg	—	1pl	1sg	—
	2sg	tə-pela-gət		2sg	mət-pela-gət
	3sg	tə-pela-gʔan		3sg	mət-pela-gʔan
	1pl	—		1pl	—
	2pl	tə-pela-tək		2pl	mət-pela-tək
	3pl	tə-pela-nat		3pl	mət-pela-nat
2sg	1sg	ana-pela-gʔe	2pl	1sg	ena-pela-tək
	2sg	—		2sg	—
	3sg	pela-gʔan		3sg	pela-tkə
	1pl	pela-tko-gʔe		1pl	pela-tko-tək
	2pl	—		2pl	—
	3pl	pela-nat		3pl	pela-tkə
3sg	1sg	ena-pela-gʔe	3pl	1sg	na-pela-gəm
	2sg	na-pela-gət		2sg	na-pela-gət
	3sg	pela-nen		3sg	na-pela-gʔan
	1pl	na-pela-mək		1pl	na-pela-mək
	2pl	na-pela-tək		2pl	na-pela-tək
	3pl	pela-nenat		3pl	na-pela-nat

morpheme, *-z*, and the three variants /z/, /s/, /əz/ are called allomorphs of that morpheme. Variation of this sort is referred to as allomorphy.

A long tradition holds that the plural morpheme is a sign consisting of a pronunciation (or rather a set of three pronunciations) and the meaning 'plural.' Thus, the meaning of *cats* is computed by combining the meaning of *cat* with that of 'plural.' Under this view, *-z* is a sign, a pairing of form (/z, s, /əz/) with meaning ('plural'), and this effectively gives the morpheme the same status as a lexeme. Theories which rest on this assumption will be called 'morphemic.'

There are three sets of problems with such a view, all of which feature deviations from the agglutinative process of adding signs together. One is that the form of a morpheme is often difficult to pin down. Thus, to obtain the past tense and past participle of *sing* we have to change the root vowel, *sang*, *sung*, but a vowel change is a process or a relationship between forms rather than a form itself. Even where it seems possible to isolate an affix there may be complications. Thus, in *written* we have an ending /n/ corresponding to the meaning 'past participle' (as in *taken*), but we also change the root vowel (cf. *write* /rait/, *written* /ritn/). Which is 'the' past participle morpheme?

Another problem is that in complex morphological systems the overall meaning is often signalled by the relationships between word forms in the whole paradigm rather than in the meanings of individual morphemes. Russian nouns have six case forms and two numbers, singular and plural, with the case and number being signalled by various endings. The word *komnata* means 'a room' (nominative singular case, which we can think of as the basic form), while *komnati* is the genitive singular case meaning 'of a room.' By coincidence, *komnati* is also the nominative plural case, '*rooms*.' However, for this class of nouns (but not others) the genitive plural 'of rooms' is *komnat*, in which there is no ending, just the bare root. This means we have to countenance a null or zero morpheme, which has no pronunciation at all, just a meaning 'genitive plural (for the *komnata* class).'

More careful consideration shows that null morphemes must be very widespread. Thus, if we are to take seriously the idea that the complete meaning of a word form is given by its component morphemes then we should conclude that the singular *cat* must have a zero singular ending (otherwise, how would we know that the word denoted a single cat?) For very complex languages in which a word might potentially be furnished with a dozen or more affixes we regularly encounter such zero morphemes. This is suspicious, because there are no reliably recorded cases of zero lexemes (i.e., true lexemes whose basic phonological form is null).

A variant of this second problem is encountered in morphologically complex languages. A typical instance can be seen in the verb morphology of Chukchee (spoken in NE Siberia). Verbs agree with both subject and object for person and number, as shown in the paradigm (Table 1) (the symbol *ʔ* is the glottal stop).

The forms *tə-pela-gʔan* 'I left him,' *mət-pela-gʔan*, 'we left him,' *pela-gʔan* 'you (sg.) left him,' *na-pela-gʔan* 'they left him' show that the suffix *-gʔan* means 'him,' and the prefixes *tə-*, *mət-* and *na-* denote subjects, 1sg, 1pl and 3pl respectively, with 2sg remaining unmarked. Thus far the system seems very

regular (agglutinative). However, the form for 'he left him' is *pela-nen* with *-nen* for the expected *-gʔan* ending. Thus, *-nen* signals not just '3sg object' but also '3sg subject.' For 3pl objects ('left them') we have the forms *tə-pela-nat* 'I left them,' *mət-pela-nat* 'we left them,' *pela-nat* 'you (sg.)/he left them,' *na-pela-nat* 'they left them,' showing that *-nat* means 'them.' But again the 3sg subject form has a different suffix: *pela-nenat* 'he left them.' Thus, *-nenat* means both 3sg subject and 3pl object. The forms *pela-tkə* introduces further deviations from agglutination. When we consider the other five tense forms and the imperative and conditional paradigms we often find that the subject/object endings display additional differences. This shows that those endings simultaneously signal information (in part) about the tense and mood of the verb, as well as the person/number of the subject and object. Complexities such as these are the rule rather than the exception with morphologically rich languages, even those often vaunted as highly regular or agglutinating such as Turkish or Hungarian.

A further complication arises from the forms *ena-pela-gʔe* 'you (sg.)/he left me,' *ena-pela-tək* 'you (pl.) left me' and *pela-tko-gʔe* 'you (sg.) left us,' *pela-tko-tək* 'you (pl.) left us.' The suffixes here are not in fact normal realizations of the first person object (these are *-gəm* and *mək* for singular and plural respectively as seen in the forms *na-*, *pela-gəm*, *na-pela-mək* 'they left me/us'). Rather, these endings are borrowed from intransitive verbs. It turns out that these forms are identical to forms of two special voice forms, the antipassive, in which a transitive verb is conjugated as though it were intransitive. The two antipassives are signalled by the prefix *ena-* and the suffix *-tko*. As an antipassive form *ena-pela-gʔe* would mean something like 'you/he were engaged in the process of leaving (someone/something or other).' However, the form has been co-opted for use as a transitive verb form in a systematic fashion. It is quite common in morphologically complex languages for one part of a paradigm to be 'borrowed' by another part, supplanting the forms one might otherwise expect (syncretism). It is particularly difficult to describe syncretism using the morpheme concept, because we would have to provide affixes such as *ena-*, *-tko* with artificial 'meanings' such as 'first person object,' when in fact their use is entirely different. It is simply impossible to state the Chukchee syncretism in morphemic terms, and thus it is impossible to state the essential generalization in a morphemic theory.

Complications such as these have encouraged many morphologists to abandon the morpheme concept, at least for inflection, and adopt the lexeme concept. Inflectional paradigms are treated as the result of applying sets of rules or principles which state the way a given form is related to other forms. Thus, each word form is associated with the set of features which characterize its position in the paradigm of forms. In the very simple case of English nouns we say that there is a feature [NUMBER] with the values [NUMBER: Singular], [NUMBER: Plural]. The feature [NUMBER:Plural] triggers the *-z* affixation process (or the *man-to-men* alternation or whatever). We say then that the *-z* affix realizes, or is an exponent of, that feature. There is no sign 'plural' which is combined with the sign for 'dog.' To handle the Chukchee case we write a battery of rules which are triggered by morphological features such as [Subject = 3sg, Object = 3pl]. These then add the various affixes according to the principle that the most specific rule applies first. Thus, there is a rule governing 3sg Subject/3pl Object which states 'add *-nenat*,' and this is more specific than the general (default) rule governing 3pl Object which says 'add *-nat*.' Similarly, in English the plural form for words such as *man* is specified in lexical entries for those words, and this specification overrides the default, Theories which reject the morpheme concept in this way will be called 'realizational' theories.

Rules systems such as these can easily handle many of the problems posed by zero morphemes. Thus, to represent the fact that the word form *dog* 'means' *dog-in-the-singular*, we just assume a general default rule for English which states 'for any noun the value of [NUMBER] is [NUMBER: Singular].' This default rule will always be more general than even the regular plural rule, and so it will only apply to those noun word forms which are not marked [NUMBER: Plural]. Thus, we can readily specify the feature value of singular 'dog' without resort to suspicious null entities. Likewise, syncretisms can be handled by special 'rules of referral' which would state that the Chukchee second person subject–first person object forms are 'referred' to the corresponding parts of the *ena-* antipassive paradigm.

6. *Productivity and Regularity*

There are various suffixes in English which create de-adjectival property nominalizations, that is, nouns derived from adjectives meaning 'property of being *Adjective*,' including: *-th* (*warmth*, *strength*), *-ity* (*readability*, *curiosity*, *piety*), *-ness* (*goodness*). These differ in important respects. The *-th* suffix is no longer used to form new words; the *-ity* suffix is restricted essentially to words derived from Latin, while the *-ness* suffix is used for all those words which do not take suffixes such as *-th* or *-ity*. In some cases, *-ness* is more or less acceptable even with words which by rights should take a different ending (thus *warmness*, *piousness* are relatively acceptable). In addition, *-th* and *-ity* may induce irregular phonological changes in the stem (*strong* ⇒ *strength*; *curious* ⇒ *curiosity*), while this is never true of *-ness*.

It would appear that *-ness* is the productive suffix in the sense that there is some guarantee that it can be used for new words, whereas this guarantee is lacking

in the case of *-th*, *-ity*. Thus, if a speaker uses (or coins) a new adjective, say, *nerdy*, then-*ness* will be the suffix used to create the nominalization: *nerdiness*. However, it turns out to be very difficult to pin down a watertight notion of productivity, largely because it is very unclear how to measure it. The productivity of an affix or morphological operation depends on the number of bases to which it could attach. Thus, there is a suffix *lʔet* in Chukchee which attaches to nouns such as 'dog' or 'boat' and means 'to travel by means of Noun.' Within the limitations imposed by this meaning the suffix appears to be completely productive even though it only gives rise to a limited number of lexemes. In other cases it is not even quite clear what constitutes a base. Thus, a *-th* suffix (distinct from the nominalizing *-th*) attaches to numerals greater than 'three' to form ordinals: *fifth*, *nineteenth*. But how many such bases are there? Does 'four thousand three hundred and nineteen' constitute a base distinct from 'nineteen'? A further interesting complexity is illustrated by *-ity*. This suffix regularly attaches to adjectives which themselves have been formed from a verb by the suffix *-able* 'such that can be *Verb*-ed,' e.g., *washable*. For this class of adjectives, *-ity* suffixation seems to be almost totally productive, a situation sometimes known as 'relative productivity.'

We may also ask what constitutes the data over which such calculations should be based. For some (written) languages there are very large grammatically analyzed corpora of several million words which allow us to determine how often a particular affix is used. One measure of productivity based on such corpora (developed by Harald Baayen 1991) calculates an index based on the number of new words in the corpus with that affix which only occur once in the entire corpus ('hapax legomenon'). The idea is that a word of such rarity is almost certain to have been created newly by the writer/speaker, so the more such unique words we find with a given affix, the more it would seem that the affix is being used productively. In practice this method provides a useful measure, though it is dependent on the size of the corpus (a small corpus or one which was not sufficiently representative would give misleading results). However, counting words in a corpus is not an ideal procedure because it only indirectly tells us about the grammatical systems which speakers use. Indeed, it is perfectly possible that some morphological operations show different degrees of productivity for different speakers.

Recent psycholinguistic evidence bears on this problem in interesting ways. Subjects can be asked to perform a variety of psycholinguistic tasks on morphologically complex words, such as distinguishing real words from nonsense words as quickly as possible. In some cases the subjects' responses are dependent on the frequency of the target word, on how many other words are similar to it (e.g., how many words rhyme with it) and so on. Such words tend to be those with unproductive, irregular morphology, such as *sang*, *written*. In the case of regular words (such as *walked*), the response pattern is relatively independent of such factors. Some psycholinguists argue that this points to a dual mechanism for processing complex words (see Pinker 1991). The irregular formations are simply stored in memory, which renders them susceptible to the phenomena known to affect the operation of memory, such as frequency of occurrence and analogy with similar forms. The regular formations are created by linguistic (symbolic) rule systems, which are independent of such memory factors. This is a controversial approach to morphology, but it furnishes a different way of looking at the problem of productivity, namely, in terms of regularity, as determined by linguistic analysis and confirmed (or sometimes disconfirmed!) by psycholinguistic testing.

In a realizational model 'productivity' is to be equated with 'regularity.' Thus, the *-ness* suffix is productive precisely because it is the default affix for de-adjectival property nominalizations. This default realization can be overridden by more specific lexical representations such as that for *warmth*. However, realizational theories also make use of the notion of the 'nested default.' For adjectives ending in *-able* the default nominal suffix is *-ity*, not *-ness*. This is a more specific principle than *-ness* suffixation because it is restricted to *-able* adjectives, but it is not (necessarily) a fact which has to be stated for all *-able* adjectives individually. Thus, in principle it can be stated as a linguistic rule, like *-ness* affixation, but of a more restricted kind. A currently open question, which is the subject of ongoing psycholinguistic research, is whether such 'specific' or 'nested defaults' are treated as regular morphology like *-ness* affixation or as irregularities which are simply memorized on a case by case basis, like *-th*.

7. *Evidence for a Distinct Morphology Component*

An important theoretical question in general linguistics is the existence of separate modules of the language faculty with their own properties. Minimally, it would seem that there have to be two modules, a sound module interfacing with speech production and comprehension, and a meaning module interfacing with the conceptual system. Chomsky's (1995) Minimalist Programme explores the idea that only these two interfaces are of significance for the architecture of the language faculty and that the 'computational system' is limited to syntax (see *Syntax*). However, it is apparent that there are principles of organization governing words which are independent of syntax. One clear case is the existence of inflectional classes.

In addition to inflecting for case/number, Russian nouns must belong to one of three genders, masculine,

Table 2
A fragment of the system of Russian noun endings

	Class I (masculine) 'law'	Class II (feminine) 'room'	Class III (feminine) 'bone'	Class IV (neuter) 'place'
Singular				
Nominative	zakon	komnata	kost'	mesto
Genitive	zakona	komnati	kost'i	mesta
Plural				
Nominative	zakoni	komnati	kost'i	mesta
Genitive	zakonov	komnat	kostej	mest

feminine, or neuter (gender is defined in terms of the syntactic phenomenon of agreement, see *Syntax*). The endings which a noun takes depend on the lexically specified inflectional class (declension) to which it belongs. A fragment of the system is given in Table 2.

Inflectional class and gender class are related but are not identical. Ignoring lexically specified exceptions in some of the classes, we can predict the default gender for each class, as shown in Table 2. However, both Class II and Class III contain feminine nouns, which means that it is not possible to predict inflectional class from gender specification (in general, both inflectional class and gender class are arbitrary properties of a lexical item).

The point about these examples is that Russian nouns are organized according to inflectional classes independently of phonology, syntax, or semantics. They are an instance of what Aronoff (1994) calls 'morphology-by-itself.' It is an important question whether such purely morphological classes are learnt, organized, and processed according to specifically linguistic principles, distinct from the principles governing other cognitive capacities. If this proves to be the case, then we will have an important deviation from Chomsky's 'Minimalist' framework, in that such organization is not determined by the requirements of the interface with sound and meaning. On the other hand, if inflectional class organization can be shown to obey principles of associative memory, like irregular past tenses in English then it may be that there is no specifically linguistic morphological module.

See also: Syntax; Word, Linguistics of; Word Meaning: Psychological Aspects

Bibliography

Aronoff M 1994 *Morphology By Itself*. MIT, Cambridge, MA

Baayen R H 1991 Quantitative aspects of morphological productivity. In: Booij G, van Marle J (eds.) *Yearbook of Morphology 1991*. Kluwer Dordrecht, The Netherlands, pp. 109–49

Carstairs-McCarthy A 1992 *Current Morphology*. Routledge, London

Chomsky N 1995 *The Minimalist Program*. MIT, Cambridge, MA

Matthews P 1991 *Morphology*, 2nd edn. Cambridge University Press, Cambridge, UK

Pinker S 1991 Rules of language. *Science* **253**: 530–5

Spencer A 1991 *Morphological Theory*. Blackwell, Oxford

Spencer A 2000 Morphology. In: Aronoff M, Rees-Miller J (eds.) *Handbook of Linguistics*. Blackwell, Oxford

Spencer A, Zwicky A (eds.) 1998 *Handbook of Morphology*. Blackwell, Oxford.

A. Spencer

Morphophonology

In its broadest definition, morphophonology refers to the interaction of word formation (see *Morphology in Linguistics*) with the sound system of language (see *Phonology*). These two linguistic subsystems interact in many different ways in the languages of the world, including phonological alternation of a stem or affix, vowel harmony involving both stems and affixes, spread of a phonological feature as a grammatical marker, combinations of patterns of consonants and vowels in nonconcatenative morphology, and phonologically definable reduplication.

After describing these types of interaction between morphology and phonology, a variety of ways in which they have been analyzed and categorized will be discussed.

1. Interactions between Phonology and Morphology

1.1 Allomorphy and Affixation

Allomorphy, i.e., differing surface forms corresponding to a single morpheme, may involve inflectional and derivational affixes or stems. The surface form may be phonetically or morphologically conditioned.

1.1.1 Phonetically conditioned allomorphy. An example of phonetically conditioned allomorphy in inflection is the productive plural marker of English, pronounced [z], [s], or [əz], depending on the final segment of the stem: *carnation*[z], *violet*[s], *ros*[əz]. Derivational affixes, too, may change shape depending on their phonetic context. For example, the English negative prefix *un-* exhibits assimilation of the nasal segment to the point of articulation of a following stem-initial obstruent: *u*[n]*productive*, *u*[n]*thinking*, *u*[ŋ]*convincing*.

1.1.2 Morphologically conditioned allomorphy. In addition to purely phonetically conditioned alternations, definable classes of morphemes in a language may behave distinctly in terms of their morphophonology. For example, two classes of affix have been identified for English, Class I and Class II (also known as 'non-neutral' and 'neutral,' respectively). Class I suffixes involve a phonological change in the stem relative to an unaffixed stem or relative to the same stem with a Class II suffix. Such changes are shown in (1) with stress placement, vowel quality, stem-final consonants. A capital letter in (1) indicates the stressed syllable.

Unaffixed stem		Class I	Class II	
(a)		-ity	-ness	
	fInal	finAlity	fInalness	
(b)		-ian	-al	
	mAgic	magI[š]ian	mAgi[k]al	(1)

In (1a), comparing *fInal* and *finAlity*, there is a difference in stress, with the first syllable of *fInal* being stressed as compared with the second syllable of *finAlity*; additionally, the vowel quality differs in the second syllable, where the unstressed [ə] of *final* alternates with a full vowel [a] in *finality*. In (1b) there is also a change both in stress placement and in the stem-final consonant ([k] alternating with [š]). In contrast, for the Class II forms, the stress placement and vowel quality are identical to that of the stem.

1.2 Vowel Harmony

Vowel harmony is the requirement in certain languages that all vowels in a particular domain (e.g., word) share particular phonological features, such as rounding, backness, height, etc. Vowel harmony may involve stems alone or stems and affixes. Khalka Mongolian (Eastern Mongolian: Mongolian Republic), for example, has frontness, rounding, and height harmony. In the verbs in (2) note that the infinitive marker, -Vx (where V = a vowel), surfaces as [-ox], [-öx], [-ax] or [ex], depending on the vowel of the stem. That is, the suffix is a single grammatical marker the vowel of which assumes a different phonological shape, depending on the stem vowel.

ot-ox	'to keep watch on'	mid, non-front, round vowels	
örg-öx	'to raise'	mid, front, round vowels	
avr-ax	'to save'	low, non-front, non-round vowels	
mem-ex	'to add'	mid, front, non-round vowels	(2)

In (2) the vowel of the infinitive marker suffix agrees with the stem vowel for frontness, height, and rounding.

1.3 Phonological Feature Spread

Phonological feature spread, similar in certain respects to vowel harmony, involves a phonological feature functioning as a grammatical marker. The difference between vowel harmony and feature spread is that feature spread does not involve a change triggered by a feature already present in the phonological form; the feature has a grammatical function and may occur throughout the word. An example of this is Terena (Arawakan: Brazil), in which the first person singular is marked by nasalization (indicated in the example with a tilde, ~). That is, given a third person singular form containing oral vowels and consonants, the first person form of the same word will exhibit nasalized vowels and consonants (subject to various conditions). Compare the following:

3rd person singular		1st person singular		
ajo	'his brother'	ãj̃õ	'my brother'	(3)

Thus, a phonological feature—here nasalization—fulfills a grammatical function. In the same language a different feature—palatalization—indicates 2nd person.

1.4 Non-concatenative Morphology

In languages with non-concatenative morphology (e.g., Arabic, Hebrew), words consist of combinations of specific consonants and vowels, with for example the consonants contributing the lexical core of the word's meaning while the vowels contribute specific morphosyntactic information.

In Arabic (Semitic: North Africa, Eastern Mediterranean, Arabian peninsula, Iraq, Syria, etc.) the perfective active verb form for the stem meaning 'write' is *katab*, while the corresponding passive is *kutib*; the perfective active form for 'do' is *faʕal* and the corresponding passive is *fuʕil*. Note that for these

words in each case the sequence of consonants and vowels is CVCVC, with the consonantal root providing lexical meaning and the vowels associated with voice, i.e., perfective active or passive.

pattern	root	+ vowel melody *a*	+ vowel melody *u-i*
CVCVC	*ktb*	katab	kutib
	fʕl	faʕal	fuʕil

(4)

1.5 Reduplication

Reduplication involves copying a base (full reduplication) or part of a base (partial reduplication) and affixing that copy to the base. The affixation may involve prefixation, suffixation, or infixation. The copy may be of a phonological constituent, e.g., a syllable, or of a definable part of the base which is not a constituent, e.g., the first two segments of the base. The first kind is known as 'constituent copying,' the second as 'non-constituent copying.'

As an example of constituent copying, consider the following full reduplication in Turkish (Turkic: Europe), in which the entire phonological word is reduplicated. In this illustration the reduplication serves as an intensifier.

čabuk	'quickly'	čabuk čabuk	'very quickly'
yavaš	'slowly'	yavaš yavaš	'very slowly'
iyi	'well'	iyi iyi	'very well'
güzel	'beautifully'	güzel güzel	'very beautifully'

(5)

As an illustration of non-constituent, partial reduplication, consider (6), showing data from Tagalog (Malayo-Polynesian: the Philippines).

takbuh	'run'	tatakbuh	'will run'
lakad	'walk'	lalakad	'will walk'

(6)

In this case the first consonant + vowel sequence of the base is copied and prefixed. The grammatical function of this reduplication is the expression of the future tense.

2. Theories and Analyses

Various theoretical accounts of certain aspects of morphophonology have been proposed in recent years. These include—among others—Lexical Phonology and Morphology, Autosegmental Phonology, Templatic and Prosodic Morphology, and Optimality Theory (see also *Phonology, Nonlinear*). The following sections illustrate how these theories apply to the kinds of data seen above.

2.1 Lexical Phonology and Morphology

In this theory the interaction between morphology and phonology has been modeled in terms of levels of interaction in the lexicon. Developing ideas proposed by Siegel (1974), the two major proponents of this view are Kiparsky (1982) and Mohanan (1986). While there are fundamental differences in the details of the models proposed by Kiparsky and Mohanan, they both share a similar view of lexical organization.

The assumption underlying the model is that morphological processes, e.g., affixation, are interleaved with phonological operations, e.g., stress assignment, and that items exhibiting different behavior may be associated with different levels. As an example, the kind of morphologically conditioned allomorphy seen in (1a) between *finality* and *finalness*, as compared with *final*, arises from the difference in affixation relative to stress assignment, with affixation of Class I affixes at Level I and Class II affixes at Level II. The affixation of *-ity* occurs before stress assignment while the affixation of *-ness* occurs after stress assignment. The result is differing stress placement on the two words and a consequent difference in vowel qualities, comparing the first and second syllables of *finality* and *finalness*.

		final +ity	final +ness
Level I	Class I affixation	final+ity	—
	Stress assignment	finAlity	fInal
Level II	Class II affixation	—	fInal+ness
	output	finAlity	fInalness

(7)

In addition to modeling interaction between word formation and the sound system explicitly, lexical phonology also makes claims and predictions about aspects of language adjacent to morphophonology, for example, morphological productivity and restricted vs. across-the-board application of phonological rules.

Among the unresolved issues surrounding lexical phonology are the number of levels of morphological and phonological interaction required (ranging in the literature from one to five), and the universality of such levels—do all languages have the same number of levels but potentially use only a subset of them?

2.2 Autosegmental Phonology

Autosegmental phonology is typically attributed to Goldsmith (1976), though there are parallels with

Firth (1948), Hockett (1955) and Rischel (1974). The fundamental notion is that there is no necessary one-to-one correspondence between a phonological feature and a segment. That is, adjacent segments may share a single phonological feature or, conversely, two values for the same phonological feature may be associated with a single segment. As an example of the first, English vowels occurring in syllables closed by a nasal stop are nasalized, as in *teen* [tĩn]. In autosegmental terms this can be seen as the spread of the phonological feature for nasality, say [nasal], to the vowel preceding the nasal stop, shown in (8a). The example in (8b) shows two values for the features [continuant] associated with a single segment. The affricate [č], the initial and final sounds in English *church*, consists phonetically of a stop plus a fricative, i.e., a non-continuant sound together with a continuant sound. Phonologically, however, the affricate behaves like any other obstruent consonant: it is a single consonant, not two. Represented autosegmentally, the two values for [continuant] are both associated with the single segment.

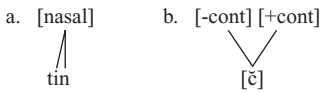

(8)

The application of autosegmental phonology to the morphophonological processes of vowel harmony and phonological feature spread seen above is both straightforward and insightful.

In the case of vowel harmony (see Sect. 1.2 above), the relevant features spread from the stem vowel to the suffix vowel, resulting in the suffix vowel harmonizing with the stem vowel; the dotted line shows that the features in question have spread from the stem vowel:

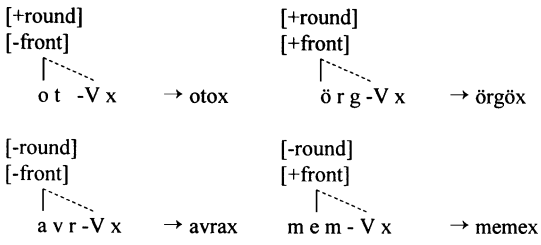

(9)

In the same way, a phonological feature fulfilling a grammatical function, seen in Sect. 1.3 above, may spread throughout a word.

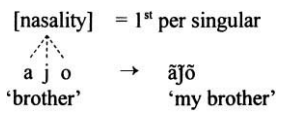

(10)

In addition to the insightfulness of autosegmental analyses of both vowel harmony and feature spread as a grammatical device, such analyses also highlight the interplay between phonological features and word formation.

2.3 Templatic Morphology and Prosodic Morphology

As seen in Sect. 1.4 above, word formation in some languages involves patterns of consonants associated with some meaning which are interleaved with patterns of vowels, themselves associated with some meaning or encoding grammatical information. Noting these recurring patterns and applying autosegmental principles of association, McCarthy (1979) proposed associating specific segments with positions in a template. That is, the recurring patterns are represented as fixed sequences of Cs and Vs onto which specific consonants and vowels are mapped.

Given the *ktb* and *fʕl* roots seen in (4) above, the combination of consonants and vowels—each group of which constitutes a morpheme—can be represented by templates, as in (11), with the specific consonants and vowels mapped onto template positions:

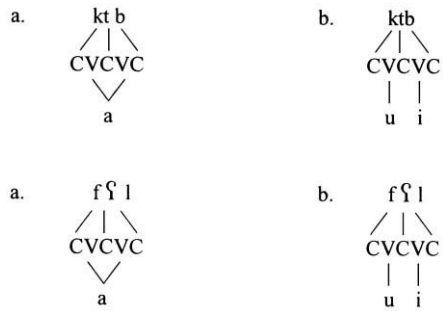

(11)

In related work, McCarthy (1981) and McCarthy and Prince (1986) extended and refined the concept of templates to include not just sequences of Cs and Vs, but prosodic constituents such as syllables or phonological words as well. Consider, for example, the advantage of this with a syllable template: in a given language, there might be five types of syllable, i.e., a valid syllable might consist of V, CV, CVC, CCVC or CCVCC. Thus the notion 'syllable' replaces five different CV templates.

The relevance of both CV templates and prosodic constituent templates to morphophonology is especially clear in reduplication. The Turkish data seen in (5) exemplify full reduplication of the word, a prosodic constituent:

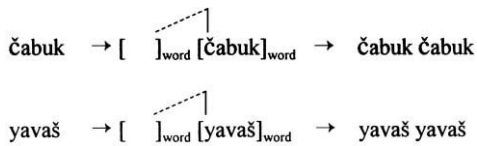

(12)

Reduplication in Tagalog, on the other hand, is clearly non-constituent copying, involving as it does the prefixation of a copy of the first CV sequence of the stem. The copy is not a constituent since universal tendencies of syllabification would suggest that the first syllable of *takbuh* is tak and therefore the first CV sequence of a syllable does not form a constituent.

takbuh	→	CV-takbuh	→	tatakbuh
lakad	→	CV-lakad	→	lalakad
pili?	→	CV-pili?	→	pipili?in

(13)

2.4 Optimality Theory

Prince and Smolensky (1993) have proposed a non-derivational theory of linguistics, relying on putatively universal constraints on linguistic representations to ascertain the wellformedness of those representations. In a least two ways optimality theory has direct bearing on morphophonological analysis. First, the constraint hierarchies themselves constrain both phonological and morphological expression, e.g., prosodic structure constraints favor specific syllable types and alignment constraints ensure that a particular affix appears in the correct position. Second, the interaction of prosodic and morphological constraints allows the analysis of certain kinds of allomorphy.

As an example, Russell, (1997) proposes that regular English plural formation reflects the interaction of three constraints (along with the full range of constraints required independently), (a) a constraint on phonological structure, *SIB-SIB, which disallows the occurrence of two sibilant consonants side by side, (b) a morphological alignment constraint, PL-AFTER-N, which ensures that the plural marker follows the noun stem, and (c) a faithfulness constraint, LEFT-ANCHOR$_{plural}$, which requires that the leftmost segment of the plural suffix should be the same in both the underlying representation and in the surface form, i.e., [z].

Given three candidates for the plural of *rose*, the following tableau represents the evaluation of those candidates by the proposed constraints. Note that the constraints are violable and that the 'most harmonic' or 'optimal' candidate is the one which has the least serious violations of constraints in the hierarchy. As to the formalism of the constraint ranking, a constraint appearing to the left of ≫ is ranked higher than a constraint appearing to the right of ≫ . Constraints of equal importance are separated by a comma. In the tableau these relations are represented by a vertical dashed line (= of equal importance), and a vertical solid line (higher ranked constraint to the left).

Constraint ranking: *SIB-SIB, PL-AFTER-N ≫ LEFT-ANCHOR$_{plural}$

(14)

Tableau for *roses*, evaluating the candidates [rozz], [roz] and [rozəz]

/roz, z/	*SIB-SIB	PL-AFTER-N	LEFT-ANCHOR$_{plural}$
[r o z z]	*!		
[r o z]		*!	
☞ [r o z ə z]			*

(15)

In this example, the correct output form [rozəz], indicated by ☞, is shown to result from the interaction of violable constraints in the evaluation of three suggested surface forms. A violation is marked by an asterisk; a fatal violation (relative to another constraint) is marked by an asterisk followed by an exclamation mark. A shaded cell indicates that because of the violation of a higher ranked constraint, that cell is irrelevant in determining the optimality of the candidate in question. Given the constraint ranking shown, the violation of LEFT-ANCHOR$_{plural}$ by the form [rozəz] is less important than either the violation of *SIB-SIB by *[rozz] or the violation of PL-AFTER-N by *[roz].

Note, though, that the reliance on *SIB-SIB precludes generalizing this analysis to the very similar facts of English past tense formation cf. play[d], hiss[t], batt[∂d], which does not involve a sibilant in the suffix.

2.5 Other Frameworks

While the theories and analyses sketched above have been important in characterizing certain aspects of the interaction between phonology and morphology, they are neither exhaustive nor do they necessarily exclude other approaches. Optimality Theory, for example, as a non-derivational, surface-oriented theory has much in common with the morphological approaches of Bybee (1985). Another important theory is Prosodic Phonology (e.g., Nespor and Vogel 1986), which

focuses primarily on phonology while dealing explicitly with the interaction between phonological and morphological structure.

See also: Morphology in Linguistics; Phonology; Phonology, Nonlinear

Bibliography

Bybee J 1985 *Morphology: A Study of the Relation Between Meaning and Form*. John Benjamins, Philadelphia

Firth J 1948 Sounds and prosodies. *Transactions of the Philological Society* 127–152 [Reprinted in Palmer F R 1970 *Prosodic analysis*. Oxford University Press, London, pp. 1–26]

Goldsmith J 1976 Autosegmental phonology. Ph.D. dissertation, MIT. [Published 1979, Garland Press, New York]

Hockett C 1955 A manual of phonology. *International Journal of American Linguistics* **21**(4): Part 1. Memoir 11

Kiparsky P 1982 From cyclic to lexical phonology. In: van der Hulst H, Smith N (eds.) *The Structure of Phonological Representations*. Foris, Dordrecht, Netherlands, pp. 131–75

McCarthy J 1979 Formal problems in Semitic phonology and morphology. Ph.D. dissertation, MIT. [Published 1985, Garland Press, New York]

McCarthy J 1981 A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry* **12**: 373–418

McCarthy J, Prince A 1986 Prosodic morphology. Ms. University of Massachusetts/Amherst and Brandeis University. [Issued 1996 as Technical Report no. 32, Rutgers Center for Cognitive Science.]

Mohanan K P 1986 *The Theory of Lexical Phonology*. Reidel, Dordrecht, The Netherlands

Nespor M, Vogel I 1986 *Prosodic Phonology*. Foris, Dordrecht, The Netherlands

Prince A, Smolensky P 1993 Optimality theory: Constraint interaction in generative grammar. Ms. Rutgers University and University of Colorado/Boulder. [Issued as Technical Report no. 2, Rutgers Center for Cognitive Science; to appear MIT Press.]

Rischel J 1974 *Topics in Westgreenlandic Phonology*. Akademisk Forlag, Copenhagen, Denmark

Russell L K 1997 Optimality theory and morphology. In: Archangeli D, Langendoen D T (eds.) *Optimality Theory, An Overview*. Blackwell, Oxford, UK, pp. 102–33

Siegel D 1974 Topics in English morphology. Ph.D. dissertation, MIT. [Published 1979, Garland Press, New York]

S. J. Hannahs

Mortality and the HIV/AIDS Epidemic

1. *HIV/AIDS and Mortality*

This article examines the effect that the HIV/AIDS epidemic has had on mortality levels in several regions of the world. We emphasize conditions experienced in countries where the epidemic has had the strongest impact.

2. *An Exceptional Occurrence in the Midst of Unprecedented Progress*

The occurrence of HIV/AIDS is an exceptional event associated with large mortality increases that, paradoxically, takes place at the end of a century characterized by the most significant feat of human history: the unprecedented lengthening of life.

Starting approximately in 1850 the world as a whole began to experience gradual but consistent increases in life expectancy. This process became a true revolution, and in less than two centuries, an infinitesimal fraction in the metric of human evolution, the average survival of the species as a whole was greatly increased. The speed and magnitude of gains in survival varies by regions and, as a result, there is still great heterogeneity in current mortality levels.

A pretransitional mortality regime, typically prevailing during the period before 1850, made human survival a highly uncertain event due to high infant and child mortality and, more generally, to very low levels of life expectancy. Between one in four and one in five live births would not reach their first birthday, whereas a member of the population would, on average, live not longer than 35–40 years. Furthermore, the life expectancy prevailing at any one time was highly variable. This is because mortality levels were very sensitive to climatic changes, variations in economic performance, epidemics, wars, and population movements. Pretransitional mortality regimes were not particularly robust.

A modern, post-transitional, mortality regime is characterized by higher and normally increasing values of life expectancy and by added robustness since the life expectancy that prevails in any one year is subject to less uncertainty than in pretransitional mortality regimes. This does not mean that variability around a trend of increasing life expectancy is not observable. Indeed, the twentieth century has seen its share of conditions that induced deceleration of gains in survival or reversals in an otherwise regular increasing trend. Examples are the flu epidemic of 1918–19, the two World Wars and their aftermath, the Great Depression of 1929, and the China famine of 1956–60. However, these episodes of higher than average mortality are rare. Furthermore, when they occur, they are generally characterized by effects that can be offset in short periods of time. The aftermath of a mortality crisis or of a period of deceleration of gains is usually followed by ressumption of the trend prevalent prior to the events that led to the crisis or deceleration. Thus, modern mortality regimes are less volatile and more resilient, particularly in more developed areas, than they were in the past.

Two exceptions to this absence of volatility stand out. The first is the post-1970 experience of Eastern Europe where the gradual trend toward higher levels of life expectancy has been arrested and, in some cases, reversed. The main factors behind these trends in the

region are not exactly new, but their effects were not felt with full force until the breakup of the Soviet Union. Since the late 1970s and early 1980s, mortality in infancy, early adulthood, and among mature adults has progressed very little, and in some cases has increased.

The second exception is the case of HIV/AIDS. The epidemic has had an impact in many world regions but only in sub-Saharan Africa have its effects being strong enough to thoroughly dismantle a regime of survival gains that took decades to achieve. The effects of the epidemic will mark these societies for a long time to come, certainly for at least the next 50 years. The losses to human life associated with the HIV/AIDS epidemic are formidable, so huge that they dwarf the consequences associated with all other crises in the twentieth century and many of those observed in other centuries.

3. *Characteristics of the HIV/AIDS Epidemic*

HIV is a virus transmitted via heterosexual and homosexual contact, blood transfusions, needle puncturing and, finally, perinatally, from mother to child. The virus is characterized by two main features. The first is that it has a long but variable incubation time during which infected individuals are asymptomatic and can, therefore, unwittingly infect others. The second feature is its ability to completely disable the immune system and, as consequence, to induce extremely high levels of mortality once the disease becomes symptomatic (ARC or AIDS). If HIV's incubation times were short, its impact would be more limited than it actually is. This is because infected individuals would have fewer opportunities to infect others, either as a result of higher lethality or because of more effective societal containment.

The HIV/AIDS epidemic adopts two principal patterns. The first is when the virus spreads mainly in well-bounded but relatively small communities of individuals who practice homosexuality or are intravenous drug users. The second pattern emerges when the epidemic is transmitted via heterosexual contact as well as perinatally. The first pattern has been characteristic of Western and Northern Europe, North America, Australia, and New Zealand. The second pattern established firm roots in sub-Saharan Africa and has been making inroads in India, Thailand, and Cambodia. A hybrid pattern, a combination of the first and second, is characteristic of Eastern Europe, Brazil, Haiti and, to a lesser extent, of other countries in Central America.

HIV prevalence among adults varies widely by major world regions. As of 1995 sub-Saharan Africa experienced levels close to 2.5 percent, whereas all other regions had levels below 1 percent. These figures conceal a great deal of heterogeneity. For example, as of 1997 prevalence levels in sub-Saharan Africa were as low as 1.5 in Benin and as high as 21.5 percent in Zimbabwe. Similarly, whereas in India adult prevalence was below 1 percent, in Thailand it is estimated to exceed 2 percent. Finally, in Brazil levels of adult prevalence are below 0.51, whereas in Haiti they exceed 4.1 percent.

HIV prevalence has been increasing in Asia, sub-Saharan Africa, and Latin America. Instead, after an early start there, levels of prevalence have begun to taper off in North America and Western Europe.

4. *Impact on Mortality Levels*

Other things being equal, the spread of the disease will tend to occur more rapidly and the prevalence levels will attain higher values in populations where the second pattern prevails. This is because the relative size of the population exposed to heterosexual contact (males and females) is usually larger that the relative size of the population exposed to homosexual contacts or intravenous drug use. By the same token, the impact of the HIV/AIDS epidemic on mortality levels will be higher in societies where the second pattern becomes established. That is, equivalent levels of adult HIV prevalence translate into radically different effects on mortality depending on the pattern of spread of the disease. This is because the mortality effect in countries with the second pattern will be felt not just at young-adult ages but also among infants and children. An increase in infant mortality leads to much larger decreases in life expectancy than those induced by equivalent increases of mortality at other ages.

Although cumulated mortality associated with AIDS reflects a formidable toll, the actual impact on life expectancy and crude death rates are sizeable only in sub-Saharan Africa. Thus, while in sub-Saharan Africa death rates will increase by an estimated 13 percent in the year 2005 (Bongaarts 1996), the increase will be of the order of 2 percent in North America, 1 percent in Western Europe, 4 percent in Asia, and 6 percent in Latin America. By the time Asia and Latin America feel most of the epidemic's impact on mortality, their experience will be closer to that of sub-Saharan Africa than to Western Europe or North America.

Despite the strong impact of HIV/AIDS in sub-Saharan Africa, it is unlikely that the increase in mortality referred to above will induce substantial decreases in the natural rate of increase. The exceptions to this are cases where HIV/AIDS also induces substantial reductions in fertility.

5. *Losses in Life Expectancy*

The potential losses of years of life expectancy in sub-Saharan Africa are staggering. During the period 1985–90, the epidemic induced an average loss of

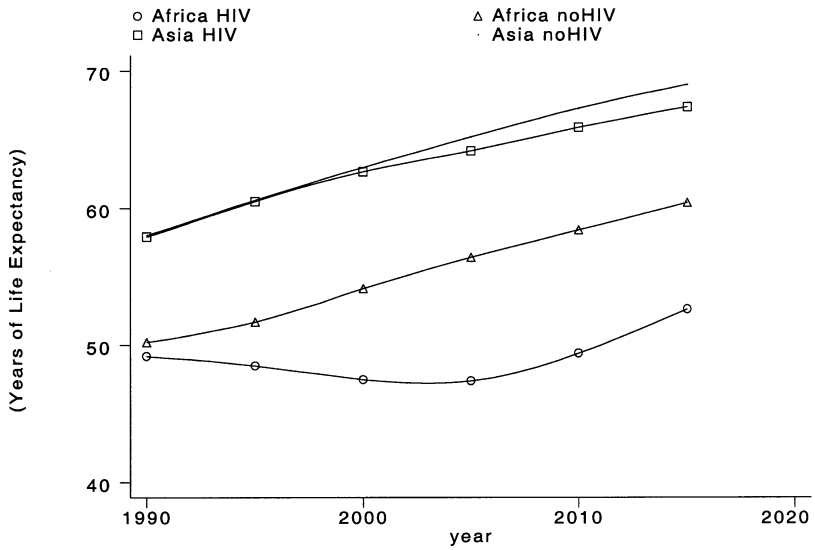

Figure 1
Projected life expectancy with and without HIV, 1990–2020 (average for sub-Saharan Africa and for three countries in Asia (Cambodia, India, Thailand)) (source: United Nations 1998)

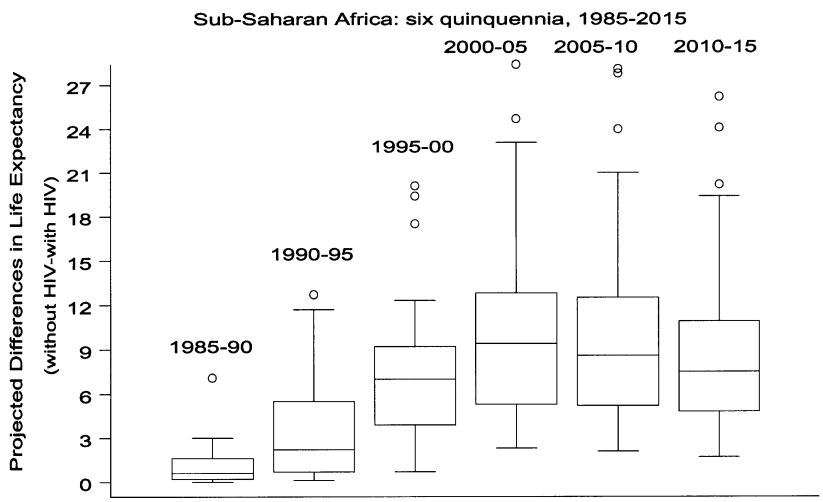

Figure 2
Projected differences in life expectancy

about one year of life expectancy. During the period 2010–15 the average loss is projected to be about 7.8 years (United Nations 1998). Figure 1 displays life expectancies that would be observed in the absence of HIV/AIDS and values projected to occur with HIV over six quinquennia, starting with 1985–90. The figure includes a curve for the average of sub-Saharan Africa and for the average of three countries of Asia (Cambodia, India, and Thailand). Projected values experience a rebound, reflecting decreasing incidence of HIV/AIDS. Though useful, the figure conceals substantial heterogeneity. Thus, in sub-Saharan Africa there are countries such as Bostwana and Zimbabwe where the difference between life expectancy with and without HIV will attain values larger than 15 years before HIV incidence rates reach a maximum. This heterogeneity in sub-Saharan Africa is captured in the box-plots in Fig. 2, which portray the trajectory of the

median, range, first and third quartiles and individual outliers of the distribution of projected values of life expectancy.

Projected levels of life expectancy in the presence of HIV/AIDs will not occur with certainty. This is because we are not quite sure about current levels of prevalence and because we rely on unproven assumptions about the incubation process (timing and about the ultimate fraction of the infected population who actually will contract AIDS).

With this caveat in mind, we note that losses of life expectancy of the magnitude uncovered before and the very long period of time over which they are spread suggest extreme conditions, comparable only to the worst crises episodes experienced by human societies.

6. *Overall Societal Impacts*

Irrespective of what the typical pattern of spread of the epidemic is, all societies affected by it will experience important social, economic, and political disruptions. However, nowhere will the impact be as large as it is estimated to be in areas of sub-Saharan Africa, where levels of adult HIV/AIDS prevalence are over 5 percent. In societies with high prevalence of other infectious diseases, HIV disables individuals before full onset of AIDS. This can have a number of effects, such as reduced productivity in the labor force, school absenteeism, and inability to fulfil paternal or maternal roles as caregivers. The impact on the economy, on physical and human capital will be enormous. Second, the mortality effects will deplete the most productive segment of the population and, in some cases, will be felt particularly hard in social groups who are influential in the political and economic life of a country. Third, and perhaps most disturbingly, the mortality impact will have a multiplying effects on future generations. Indeed, an increase in adult mortality will lead to increased widowhood and orphanhood, family disorganization and disruption of intergenerational transfers (Palloni and Lee 1992). Young children will be deprived of care, will forego education and training in order to survive, and may be even forced to engage in behaviors and activities that increase the risk of contracting the disease.

7. *The Future*

As a result of intense educational campaigns and of changes in individual behaviors, the HIV/AIDS epidemic appears to be tapering off in North America and Western Europe. This is not the case in sub-Saharan Africa, Asia, and Latin America. We expect that incidence will continue to increase, albeit at a decreasing rate, well within the first decade of the twenty-first century before it reaches a maximum and starts to decline. If so, HIV/AIDS will be with us for a long time to come, and its impact on mortality will not cease until past the year 2020. Of all the catastrophic events in the past that altered mortality levels, the HIV/AIDS epidemic will surely stand as one of the most devastating of them all.

See also: AIDS (Acquired Immune-deficiency Syndrome); AIDS, Geography of; HIV and Fertility; HIV Risk Interventions; Mortality, Epidemiological, and Health Transitions; Plagues and Diseases in History; Prostitution; Queer Theory; Rape and Sexual Coercion; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sexual Risk Behaviors; Sexuality and Gender; Sexually Transmitted Diseases: Psychosocial Aspects

Bibliography

Bongaarts J 1996 Global trends in AIDS mortality. *Population and Development Review* **22**: 21–46

Palloni A, Lee J Y 1992 Families, women, and the spread of HIV/AIDS in Africa. *Population Bulletin of the United Nations* **33**: 64–87

United Nations 1998 *The Demographic Impact of HIV/AIDS*. Population Division, United Nations

A. Palloni

Mortality, Biodemography of

Demographers have increasingly incorporated biology into their theoretical models and empirical analyses of both fertility and mortality. This incorporation of biological thinking into the study of mortality and aging has been termed 'biodemography.' Olshanky defines biodemography as the study of 'age patterns of mortality, interspecies comparisons of death rates, and biological explanations of why such patterns exist' (Olshansky 1998, p. 384). Biodemography focuses on whether the senescent process or physiological change with age is immutable and programmed into the species or whether this process can be changed with intervention. Olshansky (1998) phrases the central questions as Why do we age? How do we age? When do we age? While demographers have been interested in the last question, when do we die, without some understanding of why and how this occurs, demographers will not be able to project and predict when.

While it builds on centuries of work by both demographers and biologists, the contemporary ap-

proach to 'biodemography' came into prominence in the 1990s (Olshansky and Carnes 1997). The wide use of the term to describe this subfield of demography was adopted with the publication of the volume of papers by demographers and biologists, *Between Zeus and the Salmon: The Biodemography of Longevity* (Wachter and Finch 1997). This volume, published under the auspices of the National Research Council of the National Academy of Sciences of the United States, provides a good introduction to the breadth of issues addressed in the field.

1. Why Did Biodemography Become a Research Focus in the 1990s?

Trends in mortality over the last half of the twentieth century set the stage for linking demography and biology. The stability of mortality during the late 1950s and most of the 1960s led demographers to predict little future decline in mortality or increase in life expectancy (Crimmins 1981). They reasoned that exogenous causes of death—e.g., infectious diseases—had largely been conquered and that endogenous causes could not be affected by environment and intervention. This view has been shown to be incorrect by the subsequent extensive mortality decline. Dramatic reductions in death rates beginning in the late 1960s were concentrated largely among what might have been considered endogenous causes. These reductions were responsible for the significant increases in life expectancy which occurred during the last three decades of the twentieth century in virtually all low mortality countries of the world (see *Mortality of the Oldest Old*). The rapid decreases in mortality at even the oldest ages contradicted conventional wisdom and left demographers insecure about predicting future mortality level, life expectancy, and age structure of populations. The central questions for demographers interested in mortality at the older ages became: How long will future populations live? Is there a limit to life expectancy? Can the expected length of life or length of the lifespan be changed? The observed increases in length of life at the older ages and the increasing growth in the numbers of older people led to an additional question: Does longer life mean better health?

At the same time as these questions were becoming prominent demographic issues, they were also being raised by biologists interested in aging. While demographers historically have had an empirical approach to examining mortality change, the biological approach to aging has strong theoretical underpinnings. In the past biologists had also emphasized the inevitable nature of the senescent process with aging. More recent biological theories stress a number of mechanisms that can potentially be affected by changes in the physiological makeup of organisms or in the external environment in which the organisms age (Ricklefs and Finch 1995). Both theoretical advances and empirical observations were leading biologists to the same set of questions arising among demographers.

The ability of biologists to examine mortality and aging among other species in experimental settings holds promise for further insight into human aging and has encouraged interest in biodemography. Because experimental conditions, including the characteristics of organisms as well as the environment, can be manipulated, nonhuman populations provide useful information on the potential for changing human longevity. The commonalities between the senescent patterns of other species are being examined for clues to the potential for human life expectancy. The methods of demography can be usefully adapted in analyzing mortality and aging processes in other species; while the experimental design and shorter life span of some species studied in biology can provide answers to the questions of demographers.

Biologists have examined the age patterns of mortality in a variety of organisms and found remarkable similarity across species (Finch 1990, Finch and Pike 1996). While the range of average life expectancy varies from hours to centuries across species, the Gompertz curve, first derived to describe age variation in human mortality, fits the age pattern observed within other species remarkably well. Insights to mortality at the oldest ages are gathered from the fact that for humans and nonhumans, at the end of the species lifespan, mortality is lower than predicted by the Gompertz model (Vaupel et al. 1998).

Another factor that encouraged the collaboration of demographers and biologists in addressing these questions during the 1990s was the mapping of the human genome. Understanding the genetic influences on mortality and morbidity or on susceptibility to diseases will clearly be an important factor in predicting and perhaps changing future mortality and aging (Weiss 1989, Weiss 1990). Currently there is interest in variation in genetic endowment across populations and how this affects differential mortality and how certain genetic characteristics interact with the environment and lifestyle to affect disease onset and longevity (Ewbank 2001). The incorporation of genetic indicators adds an additional dimension to studies of the determinants of both population and individual lifespans and disease experience. The study of factors affecting individual experiences has sometimes been termed 'Medical Demography' which can be regarded as a subset of 'Biodemography.'

2. How Long Will Future Populations Live? /Is There a Limit to Life Expectancy?

The title of the book *From Zeus to the Salmon* represents the range of possible models of species survival. The gods, including Zeus, were immortal;

they did not age after reaching maturity. At the other end of the continuum, salmon die shortly after reproducing. The human species obviously fits somewhere between these two models. Biologists emphasize the link between reproduction and mortality in a way that is novel for demographers but which offers important insights into population potential.

Evolutionary biologists have focused on the role of natural selection in affecting species lifespans. Natural selection promotes characteristics that increase survival up to the end of the reproductive period; it does not eliminate the characteristics of populations that are related to diseases and death occurring after reproduction ends. The characteristics related to the diseases and disabilities that increase in frequency after the end of the reproductive period have thus not been influenced by the process of natural selection (Olshanky et al. 1998). Understanding the influence of the evolutionary process is important for understanding both current and future age patterns of mortality.

Projections of future human life expectancy depend on assumptions about possible reductions in mortality at the oldest ages. Because mortality before old age is already so low, increases in life expectancy will be achieved primarily through declining mortality at the oldest ages. In projecting future mortality at older ages, a central question arises: Is there an age beyond which humans cannot live? If there were such an age, mortality would be 100 percent at this age and there would be no survival beyond this age.

The empirical evidence on this point for humans is provided by both individual and population experience. Jeanne Calment, a French woman, lived to the oldest documented age—122. This provides the evidence that humans can live to at least that age. Data for populations from a number of countries indicates that mortality decline and the extension of life expectancy have occurred even at the oldest ages (Kannisto et al. 1994). There is no empirical evidence that we are reaching an age where there can be no further decrease in mortality or a limit to lifespan.

As noted above, the empirical evidence is that at the oldest ages, the age-specific increase in mortality is less than that predicted by the exponential Gompertz model. One explanation of lower mortality among the oldest old is the heterogeneity that occurs in population with regard to survival enhancing characteristics (Vaupel 1997). As mortality is examined at increasingly older ages, the population is increasingly selected for characteristics that affect survival which may reflect heterogeneity in initial endowment in the population. Of course it could also reflect heterogeneity in behavior and environment related to survival.

Demographers studying health and mortality in old age have also turned to increasingly complex models of individual life chances. In developing models to explain individual differences in age at death or disease onset, demographers have incorporated explanatory variables from other disciplines including biology. For instance, there has been recognition that early lifespan environmental circumstances may have significant effects on late life health outcomes. Early life includes uterine development as well as childhood conditions. Finch notes that the 'the ova from which we arose were found in our mother's ovaries while she was a fetus in our maternal grandmother's uterus, which allows environmental effects to span at least three generations in mammals' (Finch 1997, p. 245). This leads to a lifecycle view of individual aging prospects and the incorporation of indicators of early lifecycle events in explanations of old age heterogeneity in health outcomes (Barker 1998, Elo and Preston 1992).

3. *Can the Expected Length of Life be Changed?*

Recent work with a variety of species has clarified that there is plasticity to the aging process among a variety of nonhuman species, which leads to the expectation that human lifespan will be changed by similar influences. Longevity changes can be induced with environmental, behavioral, and physiological changes. For instance, food restriction in rodents has been shown to increase life expectancy; this provides clues to the way metabolic changes in humans may related to the aging process (Finch 1997). Experiments on fruit files in which the population was selected according to reproductive ability have been used to show that the longevity of the species can be changed and that evolutionary limits to lifespans do not exist (Finch 1997).

One of the central questions being addressed by biologists of aging is whether there is a set of genes that are more generally related to longevity. What are called longevity assurance genes (lag genes) have been identified in some nonhuman species. Manipulation of the genetic makeup of populations of organisms has resulted in experimental extension of average life expectancy in a number of species. Specific genes have been identified in some species that control some of mechanisms theorized to cause the aging process. For instance, one biological theory of aging emphasizes the role of oxidative stress as a cause of progressive physical deterioration with age. Genetic defects in the oxidative-stress response mechanisms of organisms have been linked to longer lifespans in a number of species including worms, flies, and mice. The success at finding similar classes of genes across species including mammals provides rationale for assuming that similar mechanisms will be identified in humans.

Knowledge of the human genome may allow us to increase both average life expectancy and potential life span. Longevity is partially under genetic control and the coming decades will provide extensive information on genetic determinants of both specific disease processes and possibly the aging process itself. Estimates derived from twin studies of human populations are

that about 20–30 percent of the variation in age at death after adulthood can be explained by genetic makeup (Vaupel at al. 1998). Genes that cause age-related diseases are being identified with increasing frequency. Specific genes related to familial types of breast cancer, colon cancer, and Alzheimer's disease are now known. One genetic allele (ApoE2) appears to be protective of a variety of conditions that are part of the senescent process including cardiovascular deterioration and cognitive loss (Ewbank 2001).

One insight for the general pattern of mortality gained from experiments with genetically identical individuals is that the species age curve of mortality tends to retain its general shape even among genetically and environmentally fixed groups. Demographers might have expected constant genetic endowment and environment to eliminate the variability in mortality or to cause mortality for all individuals in the population to occur at an identical age. In other words, to cause the survival curve to be perfectly rectangular. The fact that the age curve of mortality retains its original shape means that additional factors affect variability in survival. Interesting hypotheses are suggested by Vaupel's discussion of bioreliability (1997) and Finch's (Finch and Tanzi 1997) notion of stochastic biological process.

The empirical results of demographers based on human populations and the experimental efforts of biological researchers both lead to the conclusion that there are no observable fixed limits to species' lifespans. At this point we must conclude that the length of the lifespan can be influenced by a variety of factors both internal to the species and occurring in the environment in which species exist. This leads to the conclusions that there is dramatic potential for change in the human lifespan.

4. *Does Longer Life Mean Better Health?*

With recognition of the substantial mortality declines at the oldest ages and knowledge of the impending dramatic increase in the size of older populations projected for the low mortality part of the world, interest has turned to the expected length of healthy life as well as the expected total length of life. Demographers and policy makers wanted to know whether the increased length of life had been accompanied by increases in the length of healthy life. This discussion was initially generated by the introduction of the 'Compression of Morbidity' hypothesis by Fries (1980). Fries posited a fixed limit to life expectancy and assumed that as it was approached morbidity would be compressed into a shorter period at the end of life (see *Compression of Morbidity*).

The question of the relationship between longer life and healthy life does not arise in a disease regime dominated by infectious diseases but only when health is dominated by the chronic 'incurable' conditions of older age. Mortality is the end of a process of individual health change that begins with the onset of diseases and conditions. It is possible to intervene in that process in a variety of places: to prevent disease, to prevent progression of disease, to prevent the disability arising from disease, or to prevent death among those with severe disability and disease. The point of change determines the answer to the question of whether longer life will be accompanied by improved health. If the reason for longer life is that disease has been prevented, then declining mortality will be accompanied by lengthened healthy life. If mortality decline occurs because severely ill people are being kept alive, the length of healthy life will not be longer and populations will not be healthier.

The length of healthy life can be measured by combining mortality and morbidity schedules using a life table approach to estimate years of life in good health and in poor health. This combination of mortality and morbidity is appropriate for measuring health status in populations with low mortality which are dominated by chronic disease. The empirical evidence for a number of countries showed that initially mortality decline at older ages was accompanied by increasing expected life with disease and moderate disability. For countries where change can be examined, the length of non-disabled life appears to have increased. Extending healthy life and improving population health in the future is likely to be accomplished by delaying the onset of diseases and conditions associated with aging (Bone et al. 1995).

5. *Future Directions in Biodemography*

This is a field whose boundaries are changing constantly with relevant scientific advances. It will become an increasing focus of demographers interested in both mortality and morbidity as the implications of the mapping of the human genome are developed. Demographers are beginning to integrate both biological and genetic information into surveys designed to examine individual variation in the length of life and the disease process (Finch et al. 2001).

See also: Aging and Health in Old Age; Aging, Theories of; Life Expectancy and Adult Mortality in Industrialized Countries; Mortality of the Oldest Old; Mortality: The Great Historical Decline

Bibliography

Barker D J 1998 *Mothers, Babies, and Health in Later Life* 2nd edn. Churchill Livingston, Edinburgh

Bone M R, Bebbington A C, Jagger C, Morgan K, Nicolaas G 1995 *Health Expectancy and Its Uses*. HMSO, London

Crimmins E M 1981 The changing pattern of American mortality decline. 1947–1977, and its implications for the future. *Population and Development Review* **7**: 229–54

Elo I T, Preston S H 1992 Effects of early-life conditions on adult mortality: A review. *Population Index* **58**: 186–212
Ewbank D 2001 The potential contribution of genetic information to demographic research. In: Finch C, Vaupel J W, Kinsella K (eds.) *Cells and Surveys: Should Biological Measures Be Included in Social Science Research?* National Academy Press, Washington, DC
Finch C E 1990 *Longevity, Senesence, and the Genome*. University of Chicago Press, Chicago
Finch C E 1997 Comparative plasticity in human aging and lifespans. In: Wachter K W, Finch C E (eds.) *Between Zeus and the Salmon: The Biodemography of Longevity*. National Academy Press, Washington, DC
Finch C E, Pike M C 1996 Maximum life span predictions from the Gompertz mortality model. *Journal of Gerontology* A-Biol **51**: B183–94
Finch C E, Tanzi R E 1997 Genetics of aging. *Science* **278**: 407–11
Finch C E, Vaupel J W, Kinsella K 2001 *Cells and Surreys: Should Biological Measures Be Included in Social Science Research?* National Academy Press, Washington, DC
Fries J 1980 Aging, natural death and the compression of morbidity. *New England Journal of Medicine* **31**: 407–28
Kannisto V, Lauritsen J, Thatcher A R, Vaupel J W 1994 Reductions in mortality at advanced ages: Several decades of evidence from 27 countries. *Population and Development Review* **20**: 793–810
Ricklefs R E, Finch C E 1995 *Aging: A Natural History*. Scientific American Library: New York
Olshansky S J 1998 On the biodemography of aging: A review essay. *Population and Development Review* **24**: 381–93
Olshansky S J, Carnes B A 1997 Ever since Gompertz. *Demography* **34**: 1–15
Olshansky S J, Carnes B A, Grahn D 1998 Confronting the boundaries of human longevity. *American Scientist* **86**: 52–61
Vaupel J W 1997 Trajectories of mortality at advanced ages. In: Wachter K W, Finch C E (eds.) *Between Zeus and the Salmon: The Biodemography of Longevity*. National Academy Press, Washington, DC
Vaupel J W, Carey J R, Christensen K, Johnson T, Yashin A, Holm N, Iachine I, Kannisto V, Khazaeli A, Liedo P, Longo V, Zeng Y, Manton K, Curtsinger J 1998 Biodemographic trajectories of longevity. *Science* **280**: 855–60
Wachter K W, Finch C E (eds.) 1997 *Between Zeus and the Salmon: The Biodemography of Longevity*. National Academy Press, Washington, DC
Weiss K M 1989 Are the known chronic diseases related to the human lifespan and its evolution?*American Journal of Human Biology* **1**: 307–19
Weiss K M 1990 The biodemography of variation in human frailty. *Demography* **27**: 185–206

E. M. Crimmins

Mortality Crossover

A mortality crossover is said to occur when mortality curves of two populations or population subgroups intersect. At younger ages, age-specific death rates for one population exceed those of the other, with a gradual narrowing of differentials as age advances, until at the oldest ages the death rate differentials are reversed. Such intersection of mortality curves has provided empirical support for theories of mortality selection and 'the survival of the fittest' mechanism in old age mortality. Others have argued that the observed mortality crossovers are simply a consequence of age misreporting.

1. *Empirical Evidence*

One of the most widely noted mortality crossovers is that for African Americans and whites in the United States. Throughout the twentieth century, mortality estimates based on vital statistics and census data have consistently shown black death rates to exceed white rates until some age above the mid-seventies at which point white rates exceed black age-specific death rates. The greatest relative disadvantage for African Americans is observed in middle age, and the age at which the two rates cross has increased over time. Investigations of cause-specific mortality have revealed a similar crossover in cause-specific death rates, although the relative differentials by age and the age at which a crossover occurs appear to vary by cause (Elo and Preston 1997, Manton et al. 1979, Manton and Stallard 1997).

Evidence from linked data sources, where information for deaths and population at risk comes from a single source, also suggests a racial crossover in mortality. Kestenbaum (1992) has conducted a careful analysis of mortality based on Medicare and Social Security data for African and white Americans at ages 85 and above. Black mortality exceeded or was the same as white mortality up to age 87 for males and up to age 88 for females, at which ages the rates crossed over. Similarly, mortality estimates based on the National Longitudinal Mortality Study and small-scale, longitudinal data sources show a crossover in black and white death rates at the oldest ages (Elo and Preston 1997).

A mortality crossover is also observed for Hispanics and non-Hispanic whites in the United States, but at a younger age than for blacks and whites. Hispanic Americans have higher mortality than non-Hispanic whites at younger ages and lower mortality after about age 40–45 with the proportionate gap between Hispanic and white rates increasing with age. At least some of the Hispanic advantage appears to be due to the exceptionally low mortality of the foreign-born. Hispanic populations in Latin America are also known for having unusually low death rates at older ages relative to their level of mortality at younger ages (Elo and Preston 1997, Dechter and Preston 1991).

Empirically observed mortality crossovers are not limited to the United States. Coale and Kisker (1986)

compared death rates in selected life tables for developed and developing countries with a similar life expectancy at age five. In each pair of life tables considered, death rates at younger ages in the developing country exceeded those of the developed country, while at the oldest ages the relative differentials were reversed. Based on indicators of data quality, the authors concluded that the observed crossovers were due to poor data quality at the oldest ages. Nam et al. (1978) investigated mortality crossovers by comparing life table death rates for pairs of national populations where one population was considered low and the other high income. Of the 1,035 pairs of mortality curves examined, in only 16 percent of the original pairs were the authors' findings consistent with their original hypothesis with lower income population having higher mortality at younger ages but lower mortality at older ages.

More recently, Manton and Vaupel (1995) have suggested that a crossover in death rates at the very oldest ages also occurs between white Americans and populations in England, France, Sweden, and Japan. In Japan and the European countries, life expectancy at birth is notably higher than in the United States. Death rates, however, begin to converge around age 65, and above age 80 the US rates appear to fall below those of the European countries and Japan.

2. *Proposed Explanations*

Most observers attribute mortality crossovers either to selective survival and unobserved heterogeneity or to poor data quality at the oldest ages.

2.1 *Selective Survival and Unobserved Heterogeneity*

Manton and Vaupel have been at the forefront of advancing theories involving selective survival and unobserved heterogeneity (Vaupel et al. 1979, Manton and Stallard 1981, Manton et al. 1979). In a heterogeneous population where individuals have different susceptibilities to mortality that are fixed at birth, frailer members of a cohort die at younger ages leaving a more robust group of survivors at successively older ages. If we further consider two population subgroups that have the same distribution of frailty, but one is subject to more adverse conditions in early life, then its frailest members are eliminated more quickly than among the advantaged group. If such selection is appropriately balanced, it is theoretically possible that a crossover of the mortality curves for the two population subgroups is observed at some advanced age (Manton et al. 1979, Olshansky 1995).

Although the above mechanism is feasible, it requires that genetic endowments for longevity are fixed at birth and remain constant over the lifespan. However, lifestyle and environmental factors are likely to influence individual as well as population longevity. To address this possibility, Manton and his colleagues have extended the selection hypothesis to acknowledge the potential impact of environmental factors on differential rates of population aging (Manton et al. 1979). The authors have sought support for this hypothesis from epidemiological evidence on disease progression among whites, Hispanics, and African Americans (Manton and Stallard 1997).

The crossover in black and white death rates at older ages has provided much of the empirical support for the mortality selection and survival of the fittest explanations. Given large socioeconomic differentials between whites and blacks and adverse conditions African Americans experienced earlier this century, it is conceivable that selective survival has been more pronounced for blacks than whites. Evidence from other populations, however, suggests that cohorts who experienced unfavorable conditions in early life also tend to experience elevated mortality at older ages (Elo and Preston 1992, Mosley and Gray 1993).

2.2 *Poor Quality Data and Age Misreporting*

The second explanation for mortality crossovers is that they are produced by poor data quality. It is hypothesized that age misreporting in death statistics and/or censuses biases mortality estimates downwards at the oldest ages. Age overstatement has been considered the most common form of age misreporting, although other patterns have also been recorded (Ewbank 1981, Preston et al. 1999). Coale and Kisker (1986) concluded that the mortality crossovers they observed between developed and developing country populations were caused by poor data quality. Age misreporting has also been implicated in the black/white mortality crossover in the United States and the low mortality among Hispanics at older ages. Other potential explanations for the low mortality of Hispanics include the positive selection of migrants on such characteristics as healthiness and the difficulty of following the foreign-born in US death records (Elo and Preston 1997).

Preston et al. (1999) have shown that various patterns of age misreporting—net age overstatement, net age understatement, and symmetric age misreporting—bias mortality estimates downwards at the oldest ages. The downward bias is produced by all methods of mortality estimation the authors considered: conventional methods derived from vital statistics and censuses; longitudinal studies where age is identified at baseline; variable-*r* procedures based on age distribution of the population; variable-*r* pro-

cedures based on age distribution of deaths; and extinct generation methods. Thus, it is not unreasonable to propose that age misreporting at the oldest ages has led to downwardly biased death rates at the oldest ages and spurious mortality crossovers.

The most extensive studies of the potential impact of age misreporting on mortality crossovers pertain to the racial crossover in the Unites States. There is a general agreement that age misreporting has seriously biased African American death rates downwards at the oldest ages. The first study to document inconsistencies in age reporting for African Americans and whites in vital statistics and census data was the 1960 Matched Records study, which linked death certificates registered in May to August in 1960 to the 1960 Census of Population. For less than half of the nonwhite male and female matched cases was the age in the two sources identical. Age reporting was far better for whites. After the data were corrected for age misreporting, the racial crossover in mortality moved from the age interval 75 to 79 to the last open-ended age interval, 85 and older (Kitagawa and Hauser 1973).

Two recent matched-record studies of death certificates from the 1980s for whites and African Americans to early census records for these same individuals when they were children or adolescents continue to show substantial age misreporting for African Americans in death statistics at the oldest ages, but minimal problems for whites. Thus, conventionally constructed death rates from vital statistics and census data for African Americans continue to be seriously understated at the oldest ages. In contrast, death rates for whites do not show a systematic bias. A comparison of mortality rates for blacks and whites based on corrected data continue to reveal a mortality crossover for males in the age interval 95+ and for females at ages 90–94 (Preston et al. 1996, Hill et al. 2000). Age-specific death rates for this comparison are, however, based on somewhat unconventional variable-*r* procedures and on a very small number of deaths for African Americans at the oldest ages. They do not provide conclusive evidence of whether the racial crossover in mortality is real.

The affirmation of the reliability of mortality estimates for whites in turn lends support to the findings by Manton and Vaupel that white mortality in the United States at the oldest ages is lower than in Japan and Europe. These findings raise compelling questions about the possible underlying factors. The authors suggest several possibilities, including differences in health care, mortality selection, and the potential role of immigration among others. It is unlikely that a single explanation will account for this finding or the convergence of death rates at the oldest ages under all circumstances. To gain greater insights into the variability in the aging process, it is important to explain not only those instances where crossovers occur, but also when they have failed to materialize.

See also: Aging and Health in Old Age; Ethnicity, Race, and Health; Ethnicity, Sociology of; Mortality of the Oldest Old

Bibliography

Coale A J, Kisker E E 1986 Mortality crossover: Reality or bad data? *Population Studies–Journal of Demography* **40**: 389–401

Dechter A, Preston S H 1991 Age misreporting and its effects on adult mortality estimates in Latin America. *Population Bulletin of the United Nations* **31/32**: 1–17

Elo I T, Preston S H 1992 Effects of early-life conditions on adult mortality: A review. *Population Index* **58**(2): 186–212

Elo I T, Preston S H 1997 Racial and ethnic differences in mortality at older ages. In: Linda G, Soldo M, Soldo B J (eds.) *Racial and Ethnic Differences in Health of Older Americans*. National Academy Press, Washington, DC, pp. 10–42

Ewbank D C 1981 *Age Misreporting and Age Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis*. Committee on Population and Demography, Report No. 4. National Academy Press, Washington, DC

Hill M E, Preston S H, Rosenwaike I 2000 Age reporting among white Americans aged 85+: Results of a record linkage study. *Demography* **37**: 175–86

Kestenbaum B 1992 A description of the extreme aged population based on improved Medicare enrollment data. *Demography* **29**: 565–80

Kitagawa E M, Hauser P M 1973 *Differential Mortality in the United State: A Study of Socioeconomic Epidemiology*. Harvard University Press, Cambridge, MA

Manton K G, Poss S S, Wing S 1979 The black/white mortality crossover: Investigation from the perspective of the components of aging. *The Gerontologist* **19**(3): 291–300

Manton K G, Stallard E 1981 Methods for evaluating the heterogeneity of aging processes in human populations using vital statistics data: Explaining the black/white mortality crossover by a model of mortality selection. *Human Biology* **53**: 47–67

Manton K G, Stallard E 1997 Health and disability differences among racial and ethnic groups. In: Martin L G, Soldo B J (eds.) *Racial and Ethnic Differences in Health of Older Americans*. National Academy Press, Washington, DC, pp. 43–105

Manton K G, Vaupel J W 1995 Survival after the age of 80 in the United States, Sweden, France, England, and Japan. *New England Journal of Medicine* **333**: 1232–5

Mosley W H, Gray R 1993 Childhood precursors of adult morbidity and mortality in developing countries: implications for health programs. In: Gribble J, Preston S H (eds.) *The Epidemiological Transition*: *Policy and Planning Implications for Developing Countries*. National Academy Press, Washington, DC, pp. 69–100

Nam C B, Weatherby N L, Ockay K A 1978 Causes of death which contribute to the mortality crossover effect. *Social Biology* **25**(4): 306–14

Olshansky S J 1995 Introduction: New developments in mortality. *The Gerontologist* **35**(5): 583–7

Preston S H, Elo I T, Rosenwaike I, Hill M 1996 African-American mortality at older ages: Results of a matching study. *Demography* **33**(2): 193–209

Preston S H, Elo I T, Stewart Q 1999 Effects of age misreporting on mortality estimates at older ages. *Population Studies–Journal of Demography* **53**: 165–77

Vaupel J W, Manton K G, Stallard E 1979 The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**(3): 439–54

I. T. Elo

Mortality Differentials: Selection and Causation

Differentials in mortality by socioeconomic status and the nature of social relationships have been well established. In general, persons of higher socioeconomic status (SES) and persons who are more socially integrated experience lower death rates than their respective counterparts. These associations have been found across time, place, and age. Three sets of hypotheses have been proposed by researchers to explain these patterns. The first encompasses *causal* mechanisms through which socioeconomic status and social relationships potentially affect health status and the risk of dying. A second type of explanation, referred to as *selection* or *reverse causation*, pertains to a set of pathways whereby unhealthy individuals may reduce their social position or become socially more isolated as a consequence of their inferior health status. A third explanation relates to *artifactual* or *spurious* mechanisms, such as measurement error.

1. SES and Mortality

Differences in rates of morbidity and mortality by SES have been identified for the vast majority of illnesses, along a social gradient: persons of higher SES fare better than those lower in the social hierarchy (Adler et al. 1993). These differences (a) are present regardless of whether socioeconomic status is defined in terms of income (or other measures of wealth), education, or occupation; (b) exist at every level of the social hierarchy (i.e., not simply between the poor and the nonpoor); and (c) have been documented in a large number of studies in both industrialized and developing countries, based on vital registration, census, and survey data. An individual's health has been shown to relate not only to absolute income, but also to relative income: greater income inequality in a society is associated with higher mortality. Socioeconomic inequalities in overall mortality are typically smaller among women than men, although there are exceptions for some causes of death (Mackenbach et al. 1999).

Many of the studies have been carried out in the United States or Britain. One that has received wide recognition is the Whitehall Study of British Civil Servants (Marmot et al. 1995). This study has demonstrated that the mortality gradient is present even within a relatively homogeneous group: civil servants in one type of occupation (stable office jobs) and one geographical location (London), but in different grades of employment. Numerous studies in Britain and the US have shown that both countries have experienced widening mortality differentials in recent decades (Pappas et al. 1993).

Many recent investigations have focused on assessing the extent to which causal rather than selective or spurious mechanisms drive the observed socioeconomic inequalities in health. To reduce the impact of selection bias in the statistical analysis, researchers have relied largely on prospective (longitudinal) data; these data generally allow the analyst to measure preexisting conditions that may confound the association between SES and health, and to relate changes in SES to changes in health. There is a general consensus that the observed disparities in health are not driven by *social selection*, also referred to as *social drift*. Specifically, while there is some downward drift in SES among persons in poor health (e.g., schizophrenics are apt to lose economic resources), this type of selection is unlikely to be an important part of the association between SES and a wide range of health indicators (see, for example, Davey Smith et al. 1994, Marmot et al. 1995, Macintyre 1997). The findings also suggest that *indirect selection* mechanisms, i.e., background (or *spurious*) factors, such as height, that affect both social and health status later in life, are unlikely to be of major importance because the relationships between indicators of SES and health persist in the presence of controls for these background variables (Marmot et al. 1995). Artifactual mechanisms (e.g., errors of measurement, such as census undercounts, 'numerator-denominator' problems, such as inconsistencies in reports between registration and census data, or inappropriate measures of mortality or SES) are also not considered to be a powerful explanation of the observed associations.

The rejection of both selection and artifactual explanations has led researchers to conclude that the persistent mortality and morbidity differences by SES result largely from causal processes. Many mechanisms, some of which are interrelated, may be operating. These include differences by socioeconomic status in: (a) access to medical care, both preventive and curative; (b) access to information regarding health risks and health care; (c) patterns of health risk behaviors (such as smoking, drinking, unhealthy diet, and inadequate exercise); (d) exposure to environments that are not conducive to good health and longevity (e.g., poor housing conditions, occupational hazards, pollution, and crime); (e) exposure to stressful situations; (f) access to resources that mediate the

physiological consequences of stress; (g) ability to control one's environment and feel secure about one's position; and (h) the availability of social relationships and support.

Researchers who have focused on a single mechanism have concluded that *multiple* pathways almost certainly underlie the observed associations. For example, although the higher death rates experienced by persons of lower SES in several British and American studies have been shown to result in part from higher rates of smoking, poorer diets, and inadequate levels of exercise, these differences in behavior were insufficient to account for the observed mortality differentials (Marmot et al. 1995, Macintyre 1997). Similarly, the widening of mortality differentials by SES between the 1960s and 1980s in Britain and the US—despite the presence of the National Health Service in the former and Medicaid in the latter—has been taken as evidence that the explanation goes beyond inequities in access to medical care.

2. Social Relationships and Mortality

A related area of research has examined the associations between *social relationships* and health and mortality. These undertakings have concluded that persons who are less socially integrated have higher rates of illness and lower life expectancy than their counterparts. The associations are present for a wide range of causes of death. The possible interpretations of these findings mirror those described above, namely causal vs. selection-type arguments. Causal explanations encompass a broad set of pathways described below. Selection-type arguments posit that unhealthy people are less likely to establish and maintain social relationships or that some other (often difficult to measure) variables reduce a person's ability to have positive relationships and maintain good health.

As with studies of SES, prospective studies have provided much more convincing evidence than cross-sectional or retrospective studies regarding the importance of causal mechanisms in producing the association between social relationships and health. Findings from surveys have been strengthened by experimental and clinical research which has demonstrated that variations in the degree of social contact have psychological and physiological effects in both animals and humans (House et al. 1988).

2.1 Marital Status and Mortality

The earliest research in the area of social connection and health—dating as far back as the mid-1800s—focused largely on marital status. Since that time, hundreds of studies, primarily in industrialized countries, have demonstrated that married persons have greater longevity and experience better health than single, divorced, and widowed persons. These results hold across time, place, gender, and age, and the differentials have been consistently larger for men than for women. The dominant explanations once again can be classified as causation or selection.

Causal hypotheses, often termed *marriage protection* in this literature, encompass a broad range of social, psychological, economic, and environmental benefits that are presumed to be associated with having a spouse. For example, the presence of a spouse is hypothesized to: (a) result in greater emotional and social support; (b) lead to higher income; (c) facilitate access to medical information and health services; (d) constrain risk-taking behavior and encourage healthier lifestyles; (e) act as a buffering mechanism in the presence of stress; and (f) substitute for inferior or inaccessible formal health care. Moreover, the termination of marriage—either by divorce or death of a spouse—is believed to lead to higher risks of dying because of the stresses associated with the loss of a spouse (e.g., *bereavement effects* in the case of widowhood) and the accompanying decline in social and economic resources.

The state of being unmarried may be associated with health disadvantages that extend beyond the absence of an intimate partner: e.g., in societies where marriage is an integral part of one's social status, persons (especially women) who remain single or get divorced are likely to face discrimination and, consequently, poorer health.

Marriage selection arguments, by contrast, are based on the premise that mentally and physically healthier persons are more likely to marry in the first place, less likely to become widowed or divorced, and more likely to remarry if a previous marriage dissolves. The factors that drive the selection criteria may be far from obvious. For example, spouses may be selected for better health not only through the direct exclusion of mentally and physically ill persons from first marriage or remarriage, but also through a range of criteria that govern mate selection (i.e., *indirect selection*) such as income, physical appearance (e.g., height, weight, and beauty), health-related habits (e.g., smoking and drinking), and emotional stability.

In line with studies of SES and social connection, analyses focusing on marital status have typically concluded that the associations with mortality are produced predominantly by various benefits associated with having a spouse and not by marital selection. While this conclusion may generally be correct, many of the inferences drawn by researchers have not been justifiable. In addition, selection appears to be the dominant source of the observed differentials in at least one population (described below).

Two types of problems of inference have been noted. The first is that demographers have often based their arguments on cross-sectional data. In particular,

some demographers have reached conclusions about the importance of marital selection on the basis of aggregate patterns of mortality by marital status, e.g., age patterns of mortality differentials and the direction and strength of the relationship between the magnitude of the mortality differential and the relative size of the single population. Results from a simple simulation model demonstrate that the resulting inferences are not justifiable because both causal and selection processes could lead to the same aggregate patterns of mortality (Goldman 1993a).

The second type of weakness pertains to unjustifiable inferences derived from prospective data. Although these data have permitted researchers to conclude that the health benefits of marriage (or of higher socioeconomic status) persist in the presence of controls for selection, they have also led some to argue that selection mechanisms are unimportant. The latter conclusion has generally been unwarranted: existing statistical analyses frequently include controls for selection processes, but most were not designed to *measure the impact* of selection. Only in the past few years have demographers used longitudinal data to directly assess health-related selection into first marriage and into divorce. These studies have demonstrated that health-related characteristics and behaviors are used (explicitly or implicitly) as selection criteria for marriage and divorce. For example, in the US, users of illicit drugs, heavy drinkers, and obese persons have lower marriage rates than their respective counterparts (Fu and Goldman 1996).

Perhaps because demographers (like other social scientists) have been most interested in identifying causal mechanisms, they have failed to recognize that the impact of selection on mortality differentials by marital status is likely to vary across time and place. A recent study of atypically high mortality among never-married Japanese demonstrates this point clearly. Data from a variety of sources strongly suggest that the Japanese arranged marriage process, which was still widely prevalent in the mid-1900s, appears to be largely responsible for the huge disadvantages in longevity (10–15 years) faced by single Japanese during the period 1950–80. Several characteristics of arranged marriages are particularly noteworthy in this context: the importance of mental and physical health in the choice of a potential spouse; the inclusion of family, friends, go-betweens, and even private detectives in the selection of a mate; and, the efficiency of the marriage process (Goldman 1993b).

3. Future Directions

In spite of extensive research focusing on mortality differentials by socioeconomic status and by the nature of social relationships, relatively little is known about the relative importance of different pathways in producing the associations. Two current trends in data collection are likely to increase scientists' understanding of these linkages in the future. The first is the increasing richness and detail of longitudinal surveys in the social sciences. The second is the inclusion of biomedical markers (e.g., obtained from blood or urine samples or physician's exams) in social and demographic surveys. By identifying some of the biological pathways relating aspects of the social environment to physiological outcomes (such as neuroendocrine response, immune function, and cardiovascular function), social scientists may be more successful in elucidating the complicated linkages that underlie the observed mortality differentials.

See also: Divorce, Sociology of; Mortality, Biodemography of; Mortality, Epidemiological, and Health Transitions; Mortality: The Great Historical Decline; Social Support and Health; Social Support and Recovery from Disease and Medical Procedures; Social Support and Stress; Socioeconomic Status and Health; Socioeconomic Status and Health Care

Bibliography

Adler N E, Boyce W T, Chesney M A, Folkman S, Syme S L 1993 Socioeconomic inequalities in health: No easy solution. *Journal of the American Medical Association* **269**: 3140–5

Davey Smith G, Blane D, Bartley M 1994 Explanations for socio-economic differentials in mortality evidence from Britain and elsewhere. *European Journal of Public Health* **4**: 131–44

Fu H S, Goldman N 1996 Incorporating health into models of marriage choice: Demographic and sociological perspectives. *Journal of Marriage and the Family* **58**: 740–58

Goldman N 1993a Marriage selection and mortality patterns: Inferences and fallacies. *Demography* **30**: 189–208

Goldman N 1993b The perils of single life in contemporary Japan. *Journal of Marriage and the Family* **55**: 191–204

House J S, Landis K R, Umberson D 1988 Social relationships and health. *Science* **241**: 540–5

Macintyre S 1997 The Black report and beyond what are the issues? *Social Science and Medicine* **44**: 723–45

Mackenbach J P, Kunst A E, Groenhof F, Borgan J, Costa G, Faggiano F, Jozan P, Leinsalu M, Martikainen P, Rychtarikova J, Valkonen T 1999 Socioeconomic inequalities in mortality among women and among men: An international study. *American Journal of Public Health* **89**: 1800–6

Marmot M, Bobak M, Davey Smith G 1995 Explanations for social inequalities in health. In: Amick B C, Levine S, Tarlov A, Walsh D C (eds.) *Society and Health*. Oxford University Press, Oxford, UK

Pappas G, Queen S, Hadden W, Fisher G 1993 The increasing disparity in mortality between socioeconomic groups in the United States, 1960 and 1986. *New England Journal of Medicine* **329**: 103–9

N. Goldman

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Mortality, Epidemiological, and Health Transitions

Demographers attempt to understand past and contemporary trends in mortality, disease, and health within the frameworks of three transitions. The mortality transition is the process whereby mortality rates declined from the high levels of the past to the low levels that they exhibit in rich countries today. The epidemiological transition refers to shifts in causes of death from infectious and communicable diseases when mortality is high to degenerative diseases when mortality is low. The health transition refers to the changes over time in a society's health. Health-transition research addresses the cultural, social, and behavioral determinants of health, and thus has a wider focus and a broader disciplinary base than has until recently been usual in demographic research on mortality.

1. *The Mortality Transition*

The mortality transition and the fertility transition together make up the so-called demographic transition. This term is used to describe the historical process whereby fertility and mortality rates declined from the high and approximately compensating levels that they exhibited in past times to the low and approximately compensating levels that they exhibit in rich countries today, with the intervening period generally exhibiting higher rates of population growth than occurred in either the earlier or later periods because mortality rates tended to fall before fertility rates. The intermediate period has tended therefore to be one of marked population increase.

The Western-bloc countries of Europe, the English-speaking countries (the United States, Canada, Australia, and New Zealand), and Japan now enjoy the longest life expectancies ever recorded. Japan appears currently to be leading the field, with the most recent official statistics (for 1993) suggesting an average life expectancy exceeding 82 years for women and 76 years for men (United Nations 1996). Apart from Japan, countries with life expectancy at birth for females of 80 years or more include (in declining order) France, Australia, Sweden, Norway, and the Netherlands; those with life expectancies for men of 74 years or more include Sweden, Australia, Norway, the Netherlands, and Singapore. These figures represent increases in life expectancy of about 25 years over the last century (Preston and Haines 1991).

It is salutary to recall that past levels of infant and child mortality in today's low-mortality countries were brutally high. It has been estimated, for example, that not until the early eighteenth century did fewer than 30 percent of children of the Genevan bourgeoisie die within the first five years of life, and not until the middle of that century did fewer than 20 percent die (Hill 1995), yet at the same time in France the comparable proportion dying was close to one-half (Vallin 1991). In 1780 in Sweden almost one-third of children died before they reached their fifth birthday; this proportion had fallen to about 23 percent by 1870, about 15 percent by 1900, and less than 3 percent by 1950 (Keyfitz and Flieger 1968). As we begin the twenty-first century, it is well under 1 percent (Keyfitz and Flieger 1990).

One interesting aspect of the historical mortality transition in the West is that its occurrence was not recognized at the time. Not until the beginning of the twentieth century, for example, was it recognized that mortality rates in Britain were really declining rather than just fluctuating, but the onset of decline is now placed a century and a half earlier, in the middle of the eighteenth century (Fogel 1997). Indeed, by the middle of the eighteenth century, mortality was declining not just in Britain but throughout northwestern Europe.

The discovery that the twentieth-century mortality transition represents the continuation—admittedly the most rapid—of a process that had started considerably earlier complicates the task of explaining what brought it about. Various studies attributed the mortality transition to public-health reforms, advances in medical treatment, improved hygiene, and rising standards of living, or a combination of these, but none of these explanations has escaped criticism (Fogel 1997). McKeown (1976), perhaps the best known of the critics, reacted in particular against claims that what had been most important were advances in medical treatment and improvements in public health. McKeown shows that death rates from respiratory tuberculosis, for example, declined steadily in England and Wales at least from the middle of the nineteenth century although the bacillus was identified only in 1882, effective chemotherapy was introduced only in 1947, and BCG vaccination used on a substantial scale only from 1954. McKeown makes similar arguments for other infectious diseases as well, showing that in England and Wales rates of death from pulmonary infections (bronchitis, pneumonia, and influenza) and, for children under 15, from whooping cough, measles, scarlet fever, and diphtheria were all falling before the causal organism was identified, and certainly before effective treatment or immunization was introduced. McKeown's alternative suggestion was that the primary determinant of the mortality transition must have been increasing affluence which brought with it improved nutrition, but as Preston and Haines (1991) and Fogel (1997) have noted, and as McKeown (1978) himself admits, his arguments are both circumstantial and based more on the elimination of other factors that may have been important—through what Schofield and Reher (1991, p. 8) term a 'rather liberal use of reductive logic'—than on demonstrations of improv-

ing nutrition and of a link between better nutrition and lower mortality.

Perhaps McKeown's greatest contribution has been in stimulating closer examination of this issue. Interesting and persuasive cases have been made for the importance of improvements in water supply and sewerage (Preston and van de Walle 1978) and more generally of public-health measures designed to ameliorate living conditions in crowded urban areas (Szreter 1988), of the control of disease-spreading insects (Riley 1986), of improvements in housing (Burnett 1991), and, around the turn of the twentieth century, of improvements in domestic childcare (Preston and Haines 1991). Finally, the importance of improved nutrition has at last been demonstrated more directly. Improvements in nutritional status appear to explain about nine-tenths of the mortality decline in France and England between the last quarter of the eighteenth century and the third quarter of the nineteenth, but only about half of the decline between the third quarter of the nineteenth century and the third quarter of the twentieth (Fogel 1997). 'Only half,' however, is an impressive proportion, particularly given the progress that medicine made during this century, most significantly with the development of antibiotics and of vaccines for major childhood and other diseases.

In the earliest formulations of the demographic transition, in the 1940s and 1950s, the driving force behind the declines in both mortality and fertility was identified as modernization although this term, apparently in the belief that it was self-evident, was not defined. In addition, links were suggested between these declines: people were said to begin to limit the size of their families after they recognized that fewer of their children were dying and hence that they did not need to produce so many children in order to ensure a sufficient number of survivors.

With the growth of demography as a discipline and the proliferation of analytical methods and of collections of demographic data to which to apply them it has become clear that the reality is far more complex. Mortality decline has not always preceded fertility decline, for example, nor has fertility always declined after mortality did, and both declines have been observed in the absence of social and economic modernization. Thus, even if demographic transition remains 'one of the best-documented generalizations in the social sciences' (Kirk 1996, p. 361), it is also true that the demographic transition is both 'less a theory than a body of observations and explanations' (Caldwell 1996a, p. 175).

2. *The Epidemiological Transition*

Although proposed as a theory, the heart of Omran's epidemiological (or epidemiologic) transition is probably best characterized, like the mortality transition, as an empirical generalization. The core proposition is that when mortality is high during 'the age of pestilence and famine' (Omran 1971, p. 516), causes of death are dominated by infectious and communicable diseases; then, after an intervening 'age of receding pandemics,' low mortality brings 'the age of degenerative and man-made diseases' (Omran 1971, p. 517). Analyses of more recent developments in mortality in the United States have led some observers to posit a fourth period, 'the age of delayed degenerative diseases,' during which the age at which degenerative diseases become lethal is postponed to such an extent that life expectancy is propelled into or even beyond the eighth decade of life (Olshansky and Ault 1986).

The developed market-economy countries, with the longest life expectancies in the world, exhibit what Omran (1971, p. 533) called the 'classical or western model' of the epidemiological transition. These are countries in which the transition, at least as originally formulated by Omran, is virtually complete. They may be contrasted with the developing countries, whose epidemiological transition Omran (1971) designated a quarter of a century ago as 'contemporary' or 'delayed.' In many instances, however, their subsequent mortality declines are now more accurately termed 'accelerated': the onset of mortality decline may have been delayed relative to that in the West, but the pace of subsequent decline in many regions has been more rapid than it was in the West. Indeed, some of the greatest success stories, as well as some of the greatest failures (as discussed below), have been registered in the developing countries.

Although no country in sub-Saharan Africa has failed to achieve at least some decline in infant and child mortality, this region remains the most resistant to intervention. The few apparent successes (such as Botswana) are counterbalanced by the experience of a large group of countries that have experienced disappointingly little decline at all (such as the former Zaire). Of 15 countries for which child-mortality estimates were available in 1985 four had a probability of child mortality of 200 or more per 1,000 (i.e., at least 200 out of 1,000 children did not survive to their fifth birthday), six were in the range 150–199, three were in the range 100–149, and only two fell below 100 (Ewbank and Gribble 1993).

These levels, and the life expectancies they imply can be interpreted in different lights. One view is that there has been progress. Thus, Preston (1995, p. 34) observes that Africa's regional life expectancy of 52 years 'would have been the envy of Europe at the turn of the century.'

Another view is that progress has been slow. Vaccines exist for measles, diphtheria, whooping cough, tetanus (the latter three being combated with DPT vaccine), and tuberculosis, but these diseases, and especially measles, still remain major killers of Africa's children (as does tuberculosis of her adults) (Ewbank and Gribble 1993). Except for smallpox, for which vaccination was available already in the nine-

teenth century, Europe exhibited her 'African' levels of child mortality when there were neither vaccines to prevent the major childhood diseases nor modern drugs, most importantly antibiotics, with which to combat them. Where child mortality has fallen in sub-Saharan Africa it is largely because of public-health programs, notably the World Health Organization's Expanded Programme on Immunization, no parallel version of which existed in Europe at the turn of the century.

It would be wrong to give the impression that among the developing regions it is primarily sub-Saharan Africa where these diseases remain to a greater or lesser extent resistant to attempts at intervention, and where reduction of mortality continues to pose a difficult problem. High levels of infant and child mortality still prevail in much of South Asia, especially Bangladesh, and also in Southeast Asia, especially Indonesia.

It would also be wrong to give the impression that the major problem in these regions is early mortality. Rather, these are regions that suffer not just from high mortality but from poor data, especially on adults. Demographic surveys, as they have developed since the 1970s through the experience of the 40-odd national fertility surveys of the World Fertility Survey, and from the mid-1980s to the present with the even broader coverage of the Demographic and Health Surveys, have become increasingly good at measuring the mortality of children from information supplied by their mothers, but in the absence of good systems of vital registration in the vast majority of these countries, or of reliable ways of estimating adult mortality by more indirect means, too little is known about the mortality of adults in most of the countries that are classed as 'developing.' One can say only that it is too high. In many countries the evidence for this is that causes of adult death are still dominated by infectious and communicable diseases: that people do not yet have the luxury of surviving such diseases for long enough to succumb to the degenerative ones.

While the epidemiological transition is generally described in terms of declining mortality's being associated with a shift in the disease burden, one can reverse the description and say that as the burden shifts from infectious to degenerative diseases, mortality falls. This re-statement highlights a number of important considerations. One is that death rates from infectious diseases are capable of reaching levels far higher than those ever achieved by degenerative diseases and neoplasms. Although extremely high levels of mortality are not sustainable for extended periods since survivors of infectious diseases may be better equipped to withstand the next onslaught, and since diseases themselves die out if they kill all their hosts, the mortality rates experienced by pre-transitional populations, particularly in years of severe epidemics, are strikingly high. Indeed, one of the features of the epidemiologic transition and, indeed, of the mortality transition, is a pronounced reduction in the variability of death rates from year to year (Schofield and Reher 1991, Fogel 1997).

Another consideration is that, although infectious diseases can be lethal at any age, their effects tend to be particularly virulent among the young. In contrast, the major targets of degenerative diseases are older people. Reducing the mortality rate from an infectious disease is likely therefore to have a much more pronounced effect on overall mortality than would reducing the mortality rate from a degenerative disease because, given the characteristic shape of the population pyramid, there are more people at risk of death from an infectious disease than from a degenerative one. Moreover, the effects of such a disease shift are not confined to mortality rates since infectious diseases eliminate a proportion of the very young who would otherwise go on to bear their own children, but degenerative diseases affect those who have started, or even completed, their own families. In illustration, Keyfitz (1977) has shown that even when approximately the same number of deaths occur from malaria and heart disease, eliminating malaria has four times the effect on subsequent population increase as eliminating heart disease.

3. *The Health Transition*

Dissatisfaction with the gains from conventional demographic investigations of mortality and mortality change led to the coining of a new term, health transition. The use of the word 'health,' which was deliberate, signaled a commitment to focus not just narrowly on death and illness but more broadly on the positive state of health and well-being, an altogether larger condition than survival or even the absence of ill-health; while the use of the word 'transition' was intended to carry echoes of demographic transition, mortality transition, and epidemiological transition. The term 'health transition'—'the changes over time in a society's health' (Caldwell 1996b, p. 356)—might seem self-explanatory and, indeed, is used by some researchers as a virtual synonym for the mortality transition. Nevertheless, as originally formulated, the term was meant to refer also to the determinants of changing health and, in particular, the cultural, social, and behavioral determinants. Health-transition research therefore does not necessarily have at its core a collection of empirical data (although, this being said, most analysts support at least some of their arguments with standard demographic measures). Rather, the research seeks to explain how cultural, social, or behavioral factors have contributed to a particular health situation, whether static or dynamic, and frequently concentrates on a search for determinants of health improvement that may be susceptible to intervention. The reasons for looking beyond strictly medical and economic conditions are twofold. First,

these two areas have already received a great deal of research attention. Second, the policy implications of research that finds a beneficial effect on health of an increase in a country's health budget or its industrial or tertiary sector is in many cases likely to be of limited practical importance.

A major stimulus to the growth of health-transition research was Caldwell's (1986) article that attempted to explain the mortality transition that had occurred in Sri Lanka, Costa Rica, and the state of Kerala (which is in population terms as big as many countries). Each population remained poor, nor was public expenditure on the provision of medical services such as to preclude emulation by other societies at a similar level of economic development. What these very different populations shared was a respect—indeed a demand—for education, a relatively high degree of female autonomy, a fairly open society, and a tradition of government intervention to improve both health and nutrition. In suggesting the possibility of improved health and lower mortality in societies that have not passed through the economic and social transformations experienced by contemporary low-mortality countries, this and other key pieces of research have directed the attention of a sizeable group of researchers back to the beliefs, attitudes, and behavior of the individuals whose collective experience is ultimately reflected in morbidity and mortality statistics.

4. *Three Transitions or One?*

In considering what is meant by the mortality transition, the epidemiological transition, and the health transition the question arises of whether the approaches they imply are best viewed as distinct or as complementary. The question is important because the framework in which past and contemporary developments in population health, ill health, and mortality are viewed is likely to dictate the form of analysis, and hence the form of its conclusions. The framework of the mortality transition is likely to emphasize the identification of trends in mortality rates, of differences between rates that obtain in different populations, or of heterogeneity in rates within a population. That of the epidemiological transition is likely to stress the contributions of shifts in different causes of death to shifts in overall mortality, or even to concentrate solely on the evolution of rates of death according to a particular cause. Finally, the health-transition approach is likely to focus more closely on the role of individual agency, and to broaden the field of investigation from death, disease, or ill health to beliefs about health and illness, attitudes to modern preventative and curative health services, and social or cultural barriers or enabling factors that variously inhibit or promote the use of such services.

Despite the fact that the three approaches appear to be distinct, it can be difficult to state categorically which one of the three is best suited to the investigation of a particular research issue. Take, for example, work on within-population heterogeneity in the risk of death. Well before the advocacy of the health-transition approach, researchers had recognized that mortality rates varied according to socioeconomic characteristics either, in the case of adult mortality, of the individuals themselves or, in the case of infants and children, of their parents. Indeed, even Omran (1971, p. 527), whose description of the epidemiological transition is most usually interpreted as referring merely to a shift in causes of death as mortality rates fall, proposed that the shifts in patterns of health and disease that characterize the epidemiological transition are 'closely associated with the demographic and socioeconomic transition that constitute the modernization complex.' More recently, Mensch et al. (1985, p. 286) concluded that 'membership in a social group that is more advantaged or more modern is associated with lower child mortality, whether the variable of interest is education, income, urban residence, housing type, father's occupation, or one of the many other variables examined.' Many of the characteristics that were identified as associated with lower child mortality have been the focus of subsequent research that can probably best be described as falling within the health-transition framework. The difficulty is that a large-scale demographic sample survey or census is a good vehicle through which to identify the existence of heterogeneity in the chance of survival, but it is ill-suited to explaining how such heterogeneity has been created. Conversely, a smaller investigation that employs a variety of data collection techniques, probably involving considerable direct participation by the principal investigators, is a good vehicle through which to understand the immediate determinants of individual health-enhancing behavior, but is ill-suited to quantifying the extent to which different forms of individual behavior lead to different health outcomes.

For people who would wish to apply research findings to health policy this distinction is of more than academic significance. A study of the former type, involving quantitative analysis of a large-scale demographic survey, perhaps identifies maternal schooling as an important correlate of child survival, and reports that ten years of schooling is associated with an average reduction of child mortality of one-third. But what is the policy implication? A huge injection of (possibly scarce) government funds into mass education, especially for girls? Conversely, although a study of the latter type, involving close analysis of the pathways that individuals follow in the quest for health, can result in no such dramatic statistic, it may reveal something about how maternal education operates to improve child survival. It may tell us, for example, whether what is most important is

having grown up in a family that bothered to send its daughters to school, or the actual experience of having gone to school, or the confidence instilled by the experience of schooling, or the content of what was taught, or simple literacy. The implications for health interventions of each of these would be very different.

It is likely, then, that the studies that will be the most revealing and certainly the most useful to the makers of health policy are ones that combine the different approaches (see also Cleland and van Ginneken 1988). Moreover, despite the fact that most research that explicitly declares itself to fall under the health-transition rubric is concerned with health in the developing world, the quest for a better understanding of the behavioral and social determinants of health is a worthwhile endeavor in any population, whatever its mortality statistics. An obvious candidate for such investigation is that of the recent reversal of the mortality transition in the countries of the former Soviet bloc, most notably in Russia herself. Another is that of the widening gap between the death rates of manual and non-manual workers in Sweden, the country where it was believed until recently that mortality differences according to social class probably no longer existed (Vågerö and Lundberg 1995).

See also: Health in Developing Countries: Cultural Concerns; Mortality, Biodemography of; Mortality: The Great Historical Decline; Population Dynamics: Theory of Stable Populations; Public Health

Bibliography

Burnett J 1991 Housing and the decline of mortality. In: Schofield R, Reher D, Bideau A (eds.) *The Decline of Mortality in Europe*. Clarendon Press, Oxford, UK

Caldwell J C 1986 Routes to low mortality in poor countries. *Population and Development Review* **12**: 171–220

Caldwell J C 1996a Demographic transition. In: Kuper A, Kuper J (eds.) *The Social Science Encyclopedia*. Routledge, London

Caldwell J C 1996b Health transition. In: Kuper A, Kuper J (eds.) *The Social Science Encyclopedia*. Routledge, London

Cleland J G, van Ginneken J K 1988 Maternal education and child survival in developing countries: The search for pathways of influence. *Social Science and Medicine* **27**: 1357–68

Ewbank D C, Gribble J N (eds.) 1993 *Effects of Health Programs on Child Mortality in Sub-Saharan Africa*. National Academy Press, Washington, DC

Fogel R W 1997 New findings on secular trends in nutrition and mortality: Some implications for population theory. In: Rozenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*, Vol. 1A. Elsevier, Amsterdam

Hill K 1995 The decline of childhood mortality. In: Simon J L (ed.) *The State of Humanity*. Blackwell, Oxford, UK

Keyfitz N 1977 *Applied Mathematical Demography*. Wiley, New York

Keyfitz N, Flieger W 1968 *World Population: An Analysis of Vital Data*. University of Chicago Press, Chicago

Keyfitz N, Flieger W 1990 *World Population Growth and Aging. Demographic Trends in the Late Twentieth Century*. University of Chicago Press, Chicago

Kirk D 1996 Demographic transition theory. *Population Studies* **50**: 361–87

McKeown T 1976 *The Modern Rise of Population*. Edward Arnold, London

McKeown T 1978 Fertility, mortality and causes of death. An examination of issues related to the modern rise of population. *Population Studies* **32**: 535–42

Mensch B, Lentzner H, Preston S 1985 *Socio-economic Differentials in Child Mortality in Developing Countries*. United Nations, New York

Olshansky S J, Ault A B 1986 The fourth stage of the epidemiologic transition: The age of delayed degenerative diseases. *The Milbank Quarterly* **64**: 355–91

Omran A R 1971 The epidemiologic transition: A theory of the epidemiology of population change. *Milbank Memorial Fund Quarterly* **49**: 509–38

Preston S H 1995 Human mortality throughout history and prehistory. In: Simon J L (ed.) *The State of Humanity*. Blackwell, Oxford, UK

Preston S H, Haines M R 1991 *Fatal Years: Child Mortality in Late Nineteenth-Century America*. Princeton University Press, Princeton, NJ

Preston S H, van de Walle E 1978 Urban French mortality in the nineteenth century. *Population Studies* **32**: 275–97

Riley J C 1986 Insects and the European mortality decline. *The American Historical Review* **91**: 833–58

Schofield R, Reher D 1991 The decline of mortality in Europe. In: Schofield R, Reher D, Bideau A (eds.) *The Decline of Mortality in Europe*. Clarendon Press, Oxford, UK

Szreter S 1988 The importance of social intervention in Britain's mortality decline c. 1850–1914: A re-interpretation of the role of public health. *Social History of Medicine* **1**: 1–37

United Nations 1996 *1994 Demographic Yearbook*. United Nations, New York

Vågerö D, Lundberg O 1995 Socio-economic mortality differentials among adults in Sweden. In: Lopez A, Caselli G, Valkonen T (eds.) *Adult Mortality in Developed Countries: From Description to Explanation*. Clarendon Press, Oxford, UK

Vallin J 1991 Mortality in Europe from 1720 to 1914: Long-term trends and changes in patterns by age and sex. In: Schofield R, Reher D, Bideau A (eds.) *The Decline of Mortality in Europe*. Clarendon Press, Oxford, UK

G. Santow

Mortality of the Oldest Old

Remarkable progress has been made since 1950 in reducing the mortality of the oldest old, i.e., people 80 years old and older. Death rates increase roughly exponentially with age at adult ages but after age 80 the pace of increase decelerates and mortality may level off or even decline after age 110. This article reviews theoretical concepts and empirical findings about the causes and demographic consequences of mortality reduction and mortality deceleration.

1. Reductions in Mortality after Age 80

For females in developed countries, central death rates in the period from 1900 to 1950 at ages 85, 90, and 95 were about 0.2, 0.3, and 0.4. By the close of the twentieth century the corresponding death rates were about 0.1, 0.2, and 0.3. Fig. 1 shows the trend in mortality decline since 1950 for female octogenarians in four representative countries: France, Japan, Sweden, and the United States. In Japan the pace of improvement has been particularly rapid.

Male mortality is higher than female mortality even at ages above 100. Analysis of Western European data indicates that the annual average rate of improvement in female death rates since 1950 declined from about 2 percent at age 80 to about 1 percent for centenarians. For males the rate of improvement was roughly constant at oldest-old ages at about 1 percent per year.

Changes in the pace of oldest-old mortality improvement from 1950 to the late 1990s have been analyzed for Western Europe and Japan. For octogenarian females, the rate of reduction, which was about 1 percent per year in the 1950s, accelerated to more than 2 percent per year in the 1990s. Rates of reduction for nonagenarian females and octogenarian and nonagenarian males also roughly doubled, from levels of well under 1 percent per year in the 1950s to levels well above 1 percent per year in the 1990s. Female death rates in Japan since 1970 and France since 1980 have declined at annual rates of about 3 percent for octogenarians and 2 percent for nonagenarians. There is no correlation between levels of mortality and rates of reduction in countries with low oldest-old mortality.

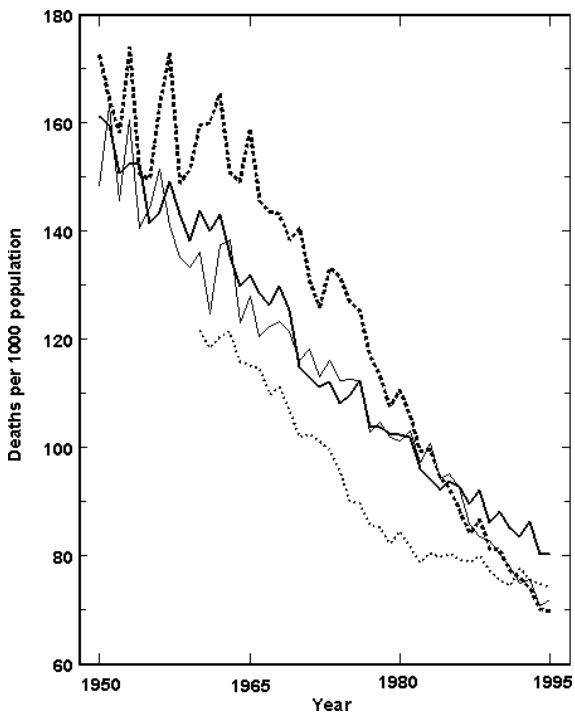

Figure 1
Death rates for female octogenarians from 1950 to the late 1990s are shown for France (thin solid line), Sweden (bold solid line), the United States (dotted line), and Japan (bold dashed line). The US data may be unreliable before 1970

Almost all available data on oldest-old mortality pertain to countries in Europe or to the United States, Canada, Australia, New Zealand, and Japan. Little is known about oldest-old mortality in developing countries, in part because age reports are unreliable. The major exception is China, where acceptably reliable data on the majority Han population has been compiled. In 1990 Han Chinese death rates for men and women in their 80s and 90s were somewhat higher than in Sweden or Japan.

As shown in Fig. 1, oldest-old mortality in the United States has been low compared with most other developed countries but improvements have been slower. The apparent US advantage was generally considered, until recently, to be an artifact of age misreporting. New evidence now indicates that US data for whites are acceptably reliable at least up through ages in the late 90s and at least since 1980. Death rates for the US white population closely match those shown in Fig. 1. Reliable data indicate that oldest-old death rates in Iceland are about the same as shown in Fig. 1 for the United States. It appears that oldest-old death rates in Canada, Australia, and New Zealand may follow a similar pattern.

In Eastern Europe, reductions in oldest-old mortality were relatively small between 1960 and the end of the 1990s, and in some countries in some decades mortality increased, particularly for males. East Germany followed this general pattern until reunification with West Germany in 1990. Thereafter substantial improvements were achieved.

2. Growth of Oldest-old Populations

In most countries, both developed and developing, the population above age 80 is rapidly growing, in part because of mortality reductions after age 80. Consider, for example, England and Wales, where the remaining life expectancy of 80-year-old females around 1950 was approximately six years. In the late 1990s, the corresponding figure was nine years and the population of female octogenarians in England and Wales was roughly half again as big as it would have been if mortality after age 80 had remained at 1950 levels. More than a half million females age 80+ were alive in England and Wales who would have been dead if mortality after age 80 had not been reduced.

In Western Europe in the late 1990s there were between two and three times as many nonagenarians

and about eight times as many centenarians as there would have been if mortality after age 80 had stayed at 1960 levels. In Western Europe, Japan, and China, the population of centenarians has been doubling every decade in recent decades, largely because of mortality improvements after age 80. (Concerns about age misreporting preclude reliable estimation of numbers of centenarians in most other parts of the world, including the United States.) Long-term, highly reliable data for Denmark illustrate the trend. In the decade of the 1870s, an average of three people per year attained age 100. In 1970, this number increased to 43 and in 1999 to 254.

3. *The Advancing Frontier of Survival*

Increases in record lifespans are also largely attributable to improvements in survival at the highest ages. In Sweden in the 30 years between 1860 and 1889, no one survived past age 105. The maximum attained lifespan then gradually rose, reaching age 112 in 1994. It is possible in Sweden and other countries with populations of a few million people that no one celebrated their 100th birthday before 1800. Many accounts exist of people living far longer at earlier times, but these accounts are generally fallacious, as are most accounts of centenarians today in most poor countries and some richer ones. A few scattered centenarians may have lived before the nineteenth century, perhaps a handful or less at any point in time. This contrasts with the 100,000 or more centenarians who celebrated the dawn of the twenty-first century.

There were almost certainly no true super-centenarians (individuals aged 110 or above) prior to the mortality decline of the past two centuries. The first reasonably well-documented case of a super-centenarian is Katherine Plunket, who died at the age of 111 in 1932 in Northern Ireland. Jeanne Calment was the first carefully verified instance of a person reaching age 120; she died at the age of 122 years and 5 months in 1997 in France.

4. *Age Trajectories of Oldest-old Mortality*

Reliable data on mortality from age 80 to the highest age attained are available for Japan and 13 Western European countries. When these data are pooled, it is possible to accurately estimate the age trajectory of human mortality up to age 105. Reasonable estimates can be made up to age 110 and shakier guesstimates can be boldly ventured up to age 122 by fitting curves to the data. Fig. 2 shows that mortality does not increase exponentially after age 80. Mortality decelerates. Whether mortality is slowing increasing, level, slowly decreasing, or rapidly decreasing after age 110 is uncertain. A logistic curve that fits the data well from age 80 to 105 indicates that death rates may reach a plateau. A quadratic curve fit to the data at ages 105+ suggests a decline in mortality after age 110.

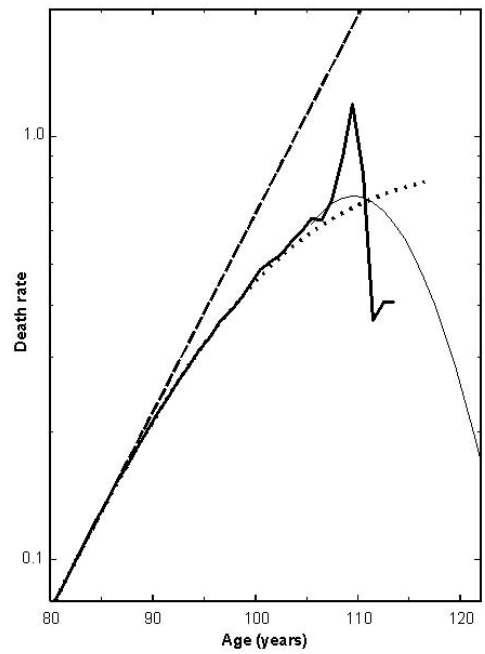

Figure 2
Death rates, on a log scale, in an aggregation of 14 countries (Japan plus 13 Western European countries) with reliable data, over the period from 1950 to 1997 for ages 80 and above are shown by the bold line. The last observation is Jeanne Calment's death at age 122, but data are so sparse at the highest ages that the trajectory of mortality is too erratic to plot. Although the graph is based on massive data, some 287 million person-years at risk, reliable data were available on only 82 people who survived past age 110. The exponential (Gompertz) curve that best fits the data at ages 80–84 is shown by the dashed line. The logistic curve that best fits the data is shown by the dotted line. Finally, a quadratic curve (that is, the logarithm of the death rate as a quadratic function of age) was fit to the data at ages 105 and higher; it is shown by the thin solid line (source: Vaupel et al. 1998)

Studies of large populations of yeast, nematode worms, and several species of insects demonstrate that death rates rise at younger ages and level off and in some cases decline at older ages. This pattern also holds for various makes of automobiles.

5. *Theories about Oldest-old Mortality*

Thus, the two main research findings in the 1990s concerning oldest-old mortality were: (a) death rates after age 80 declined substantially since 1950, at an accelerating pace, and (b) the increase in mortality

with age decelerated at advanced ages for humans, various other species, and even automobiles. These two findings are perplexing. What biological charter permits us (or any other species) to live long post-reproductive lives? A canonical gerontological belief posits genetically determined maximum lifespans. Most sexually reproducing species show signs of senescence with age, and evolutionary biologists have developed theories to account for this. The post-reproductive span of life should be short because there is no selection against deleterious mutations that are not expressed until ages when reproduction and nurturing have ceased. It should be increasingly difficult to reduce death rates with advancing age and mortality should accelerate with age as reproductive activity declines. Because these hypotheses are wrong, theory has to be rethought.

Vaguely related to the biological theory of aging are three widely-held notions:

(a) Deaths at older ages are essentially due to old age, and nothing can be done about old age.

(b) The typical human organism is not constructed to survive much past age 80 or 90.

(c) Causes of death at younger ages are largely extrinsic but causes of death at older ages are mostly intrinsic, and it is very difficult to reduce intrinsic causes of death.

These notions are dubious, ill-defined speculations that are inconsistent with available evidence.

Three biodemographic concepts—mortality correlation, heterogeneity in frailty, and induced demographic schedules—point to promising directions for developing theory.

5.1 Mortality Correlation

Demographers have long known that death rates at different ages are highly correlated across populations and over time. In addition to environmental correlations, there may be genetic correlations: mutations that raise mortality at older ages may do so at younger ages as well. Postreproductive lifespans might be compared with postwarranty survival of equipment. Although living organisms are vastly more complex than manufactured products, they too are bound by engineering constraints that may impose mortality correlations. The trajectory of mortality for automobiles suggests the possibility that both mortality deceleration and mortality correlation are general properties of complicated systems.

5.2 Heterogeneity in Frailty

All populations are heterogeneous. Some individuals are frailer than others, innately or because of acquired weaknesses. The frail tend to suffer high mortality, leaving a select subset of particularly strong or resistant survivors. This creates a fundamental problem for analyses of oldest-old mortality: as a result of compositional change, death rates increase more slowly with age than they would in a homogeneous population.

The leveling off and even decline in mortality can be entirely accounted for by frailty models in which the chance of death for all individuals in the population rises at a constant or increasing rate with age. On the other hand, mortality deceleration could also result from behavioral and physiological changes with age. Research is needed on the relative importance of compositional change vs. individual change in determining oldest-old mortality patterns.

5.3 Induced Demographic Schedules

A key construct underlying evolutionary theory is the Lotka equation, which determines the growth rate of a population or the spread of an advantageous mutation (Charlesworth 1994). The simplistic assumption in the Lotka equation that fertility and survival schedules are fixed is surely wrong. Environments are uncertain and changing. An individual does not face the fixed fertility and survival schedules of stable population theory (Keyfitz 1968), but dynamically adopts alternative schedules as the environment and the individual's capabilities change. For these and other reasons, Lotka-based theory needs re-thinking. The Lotka equation, however, does capture a fundamental insight: it is reproductive success that is optimized by evolution, not longevity. Deeper understanding of survival at older ages depends on intensified research into the interactions between fertility and survival.

6. Determinants of Oldest-old Mortality

Oldest-old mortality is lower for women than men, people born in this century rather than earlier, and people born in rich countries rather than poor ones. Smoking is a health hazard at all ages. Diet is important. The significance of medical treatment and other health care interventions may grow with age.

About a quarter of the variation in adult lifespans appears to be attributable to genetic variation among individuals (Vaupel et al. 1998). Several genes, including the *ApoE* gene, with polymorphisms that influence oldest-old mortality have been uncovered and there may be hundreds of others that substantially affect oldest-old mortality (Vaupel et al. 1998).

Research in progress suggests that an additional quarter of the variation in lifespans after age 50 may attributable to nongenetic characteristics that are more or less fixed by the time a person reaches age 50. In addition to educational achievement and socio-

economic status, month of birth and mother's and father's age at a person's birth influence oldest-old mortality. Women and men who have a baby in their 40s suffer lower mortality at oldest-old ages. Nutrition and infections *in utero* and early in life have major effects on late-life mortality (Barker 1992, Doblhammer 1999).

7. *Forecasts of Oldest-old Mortality*

Almost all population forecasts made by governmental organizations have assumed slow rates of mortality improvement. Because death rates over the past century have continued to fall at most ages and have been reduced at an accelerating pace at older ages, past forecasts have seriously underestimated progress in increasing survival.

In developed countries large increases in life expectancy depend on substantial reductions in oldest-old mortality. If progress is slow, then period life expectancy at birth in 2050 may be less than 85 in most developed countries. On the other hand, if improvements in French mortality continue in the future at the same pace as in recent decades, half of the baby girls born in France at the end of the twentieth century may survive to celebrate the dawn of the twenty-second century. Note the difference between these two statements. The first pertains to period life expectancy for both sexes combined. The second pertains to median cohort lifespans for French females. Much of the apparent disagreement among demographers about the duration of life in the future arises because different measures and populations are being compared. Real disagreements, however, exist about whether the pace of mortality improvement will persist, further accelerate, or decline over the coming century (Wilmoth 1998).

See also: Adult Mortality in the Less Developed World; Aging and Health in Old Age; Life Expectancy and Adult Mortality in Industrialized Countries; Mortality, Biodemography of; Mortality: The Great Historical Decline; Social Support and Health; Socioeconomic Status and Health

Bibliography

Barker D J P 1992 *Fetal and Infant Origins of Adult Disease*. BMJ Press, London

Charlesworth B 1994 *Evolution in Age-Structured Populations*. Cambridge University Press, Cambridge, UK

Christensen K, Vaupel J W 1996 Determinants of longevity: genetic, environmental, and medical factors. *Journal of Internal Medicine*. **240**: 333–41

Doblhammer G 1999 Longevity and month of birth: evidence from Austria and Denmark. *Demographic Reseach* **1**: 3 (text available at http://www.demographic-research.org)

Hill M E, Preston S H, Rosenwaike I 2000 Age reporting among White Americans aged 85+. *Demography*. **37**: 175–86

Jeune B, Vaupel J W (eds.) 1995 *Exceptional Longevity: From Prehistory to the Present*. Odense University Press, Odense, Denmark (text available at http://www.demogr.mpg.de)

Jeune B, Vaupel J W (eds.) 1999 *Validation of Exceptional Age*. Odense University Press, Odense, Denmark (text available at http://www.demogr.mpg.de)

Kannisto V 1994 *Development of Oldest-Old Mortality, 1950–1990: Evidence from 28 Developed Countries*. Odense University Press, Odense, Denmark (text available at http://www.demogr.mpg.de)

Kannisto V 1996 *The Advancing Frontier of Survival*. Odense University Press, Odense, Denmark (text available at http://www.demogr.mpg.de)

Keyfitz N 1968 *Introduction to the Mathematics of Population*. Addison-Wesley, Reading, MA

Manton K G, Vaupel J W 1995 Survival after the Age of 80 in the United States, Sweden, France, England and Japan. *New England Journal of Medicine*. **333**: 1232–5

Robine J M (ed.) 2001 Special Issue on Longevity. *Population*

Robine J M, Vaupel J W, Jeune B, Allard M 1997 *Longevity: To the Limits and Beyond*. Springer Verlag, Berlin

Thatcher A R, Kannisto V, Vaupel J W 1998 *The Force of Mortality at Ages 80 to 120*. Odense University Press, Odense, Denmark (text available at http://www.demogr.mpg.de)

Vaupel J W, Carey J R, Christensen K, Johnson T E, Yashin A I, Holm N V, Iachine I A, Kannisto V, Khazaeli A A, Liedo P, Longo V D, Zeng Y, Manton K G, Curtsinger J W 1998 Biodemographic trajectories of longevity. *Science*. **280**: 855–60

Wachter K W, Finch C E 1997 *Between Zeus and the Salmon: The Biodemography of Longevity*. National Academy Press, Washington, DC

Wilmoth J R 1998 The future of human longevity: a demographer's perspective. *Science*. **280**: 395–7

J. W. Vaupel

Mortality: The Great Historical Decline

1. *Features of the Decline*

The great mortality decline started in Western and Northern Europe in the middle of the eighteenth century or perhaps earlier, leveled off for a couple of decades in the mid-nineteenth century, and then continued again. Life expectancy at birth rose from some 35 years to more than 70 years over this period. The development in North America was similar. Typically, mortality began to decline some 50 to 150 years before the industrial revolution and before living standards started to improve monotonically, and some 100 to 150 years before marital fertility started its own long-term decline. Exceptions are England, where the decline in death rates started at about the same time as the industrial revolution during a period of falling real wages, and France, where fertility started declining at about the same time as mortality, and also well before

living standards improved. The timing of the great mortality decline is strikingly similar in the countries of Western and Northern Europe despite differences in economic structure and development. It could also be surprisingly simultaneous in different regions in the same country despite large differences in economic development.

In countries where we have long historical series of age-specific mortality rates (England, France, and some others including the Nordic countries), we know that the decline began with a fall in the rates of infant and child mortality. This was due mainly to a reduction in deaths from smallpox, which was a very common childhood disease in the eighteenth century. The increase in life expectancy during this period was due almost entirely to the decline in infant and child mortality.

While infant mortality continued to decline in Sweden throughout the nineteenth century, it leveled off in England and remained stable until end of the nineteenth century; then dropped rapidly as it did in all other Western countries. Childhood mortality in both countries actually increased in the middle of the nineteenth century before it started a persistent decline. This exemplifies two patterns of the historical infant and childhood mortality decline: a northern pattern with declines essentially throughout the nineteenth century and a Western one with a leveling off during the nineteenth century and a new decline after the 1880s (Perrenoud 1984).

Adult and old age mortality started to decline slowly at the beginning of the nineteenth century, or possibly earlier for England. The decline was generally more pronounced in the latter part of the nineteenth century and accelerated after World War I along with mortality at all other ages. The decline slowed down for adults and the elderly around 1950 but accelerated again in the 1970s.

With the great mortality decline followed a change in the composition of the causes of death, away from pestilence, over receding pandemics, to man-made diseases. Omran (1971), in his theory of epidemiological transition, distinguished between three development patterns: the classical or Western pattern, described above, the accelerated pattern, and the delayed pattern. They are distinguished by differences in timing and speed. The decline that took 200 years in the West started 150 to 200 years later in the Third World but then took less than 50 years to complete. There are still many countries where the great mortality decline has been delayed and where it remains to reap the gains from low mortality.

2. *Explanations of the Mortality Decline*

An understanding of the great mortality decline can be based on two approaches, namely *period and cohort explanations*. One should also make a distinction between factors that rely almost entirely upon the actions of man and those beyond deliberate human control.

2.1 *Period Explanations*

A widely accepted multifactorial explanation (United Nations 1953) is based on period factors that depend upon our activities such as public health reforms, advances in medical knowledge, improved personal hygiene, and rising income and standards of living. This type of explanation is consistent with demographic transition theory and with the theory of the epidemiological transition.

The demographic transition theory was once one of the most widely accepted 'truths' in social sciences and one of the main pillars of the UN paradigm of how to promote developing countries. It stated that mortality was high and fluctuated strongly before the demographic transition. The transition was said to begin with a decline in the level of mortality and with a dampening of its fluctuations. This was regarded as a consequence of a modernization of the economies (the agricultural and industrial revolutions). A rapid increase in population size would follow because fertility adjusted much later due to a slow response in institutions and norms that influence fertility (Davis 1945, Notestein 1953).

The idea of a stable level of mortality prior to modernization is a central feature in the theory, but before the mid-eighteenth century the mortality level was far from stable, as was shown for England by Wrigley and Schofield in their classic book (1981). Similar patterns had been demonstrated by many other scholars as well, but only for individual parishes or institutions, or for smaller regions like southern Sweden (Bengtsson and Oeppen 1993) or northern Italy (Galloway 1994). Since the mortality level prior to the great mortality decline was not stable, it is not appropriate to use it to date the beginning of the demographic transition. Another finding that questions the demographic transition theory is the long-term movement in fertility in the past. Since such changes took place in the past, one cannot justify the assumption that fertility norms changed slowly and kept fertility high for many decades after the start of the demographic transitions (Bengtsson and Ohlsson 1994).

On a broader front, McKeown (1976) questioned the multifactorial explanation of the great mortality decline. He argued that a single factor, namely better nutrition, could explain almost the entire great mortality decline. His criticism was based on a study of cause-specific mortality in England and Wales from 1838 to 1947, where he observed that two-thirds of the mortality decline was due to a reduction in infectious diseases. In later work he also analyzed mortality rates and economic development for other countries and

further back in time, though not in such detail as for England and Wales.

McKeown argued that medical advances had little influence on mortality trends before the breakthrough of sulfonamides and antibiotics in the 1930s and 1940s. In earlier periods the only disease that one could cure was diphtheria. Very few deaths, however, were due to diphtheria, and it was already retreating when the antitoxin came in use around 1900. For periods before 1838, McKeown argued that inoculation and vaccination for smallpox had little or no impact on the general mortality decline. Vaccination started in England at the end of the eighteenth century, but it was not widespread until after 1840, at which time it became available at public expense. Others have drawn the same conclusion for Sweden, where inoculation never became widespread. Only about 50,000 individuals were inoculated between 1750 and 1800 (Sköld 1997). If inoculation provided total safety against death from smallpox, some 7,500 would have been saved, assuming a fatality rate of 15 percent. This saving should be compared with the 300,000 who died from smallpox in Sweden during this period. Thus, even though inoculation saved some lives and possibly slowed the spread of the disease, it was not a major factor behind the initial mortality decline.

A similar story can be told about vaccination. In all the Nordic countries, the smallpox mortality declined before vaccination started in the first years of the nineteenth century and well before vaccination became common in the 1820s. Even though the level of mortality in smallpox was low already when vaccination started, the disease continued to be fatal for many decades. There were outbreaks in Finland and Sweden as late as in the 1880s, most likely because not all children were vaccinated (for an overview of the Nordic countries, see Bengtsson 1992). In England, vaccination did not become widespread until the latter part of the nineteenth century. Thus neither inoculation nor vaccination was of major importance for the initial mortality decline but it is likely that vaccination ensured that the disease never came back once it had essentially been eradicated.

The improvement of personal hygiene may have had some effect on mortality in England and Wales after about 1880, when a decline in intestinal infections coincided with substantial improvements in water supply and sewage control, which must have reduced the incidence of waterborne infections. Since changes in mortality from intestinal infections were only a smaller part of the general mortality decline prior to, say, 1870, the general reduction must have been due mostly to factors other than improved personal hygiene, according to McKeown. For the same reason the type of public health measures discussed above must have had little impact on the great mortality decline prior to 1870 and only a partial effect afterwards. Other public health measures, such as the breastfeeding campaigns that took place in Sweden in the 1830s and later, were not discussed by McKeown. Fridlizius (1984) has argued for Sweden that these campaigns might very well have had an impact on infant mortality, but that they started much later than the decline in mortality. Incidentally, childhood mortality went up somewhat during the period when breastfeeding campaigns started.

Helleiner argued in 1957 that the population of Western Europe must have increased from the mid-eleventh century to the late thirteenth century and from the mid-fifteenth century to the end of the sixteenth century. The population increase observed in the eighteenth century should therefore not be unique except that mortality decline started from a higher level and went on longer than before. Helleiner and others scholars have argued that these changes in mortality were spontaneous (natural), and in 1973 the UN added natural factors as a fifth determinant of the great mortality decline. Later, others like Fridlizius (1984), Perrenoud (1984), and Schofield (1984) have asserted that a change in the virulence of pathogens initiated the great mortality decline. The basic idea is that the virulence of pathogens changes spontaneously over time. Virulence in an organism is generally understood as its ability to overcome host defenses. At this point it is important to note that some pathogens develop more quickly in a malnourished host while others do not depend on such circumstances to produce very high mortality. (Influenza epidemics, for example, are neither created nor sustained in malnourished persons, and they never were.)

McKeown's main argument on this issue, though not proved, is that the initial mortality development during the latter part of the eighteenth century was an integrated part of the great mortality decline, which continued for the next two centuries and therefore cannot be due to a spontaneous reduction in the virulence of pathogens. According to McKeown, what is left is nutrition as an explanation not only for the decline in infectious diseases from the mid-nineteenth century onwards but also for the initial phase. This reasoning is part of his attempts at finding one explanation for the entire mortality decline.

In criticism, Fogel (1994) noted that McKeown only discussed the nutritional intake, the diet, and not the claims from the body to maintain itself and build up cells. Thus, McKeown only took into account gross nutrition, not net nutrition, and the latter must be more closely related to health and mortality. We will come back to this issue shortly.

While Helleiner regarded the virulence hypothesis as a period explanation outside of man's control, Kunitz (1986) did not agree. He argued that the disappearance of epidemics, such as plague, typhus, and smallpox, is associated with the growth of stable governments, the expansion and integration of nation states, and the emergence of disciplined armies in the seventeenth and eighteenth centuries. These advancements made it possible to control epidemics by means

of isolation and better hygiene. Population growth and loss of regional and local isolation then converted epidemics into relatively more benign childhood diseases. The mortality in the latter diseases varied depending on nutritional status and the availability of adequate nursing, as did mortality in the nonspecific pneumonia–diarrhea complex, which became relatively more important when epidemic diseases declined. Poverty, crowding, poor sanitation, and hazardous work conditions became more important as determinants of death, particularly in the cities. As a result, the eighteenth century saw a divergence in the life expectancy in the upper and lower classes, according to Kunitz. His empirical evidence for this conclusion is rather weak since it is only based on a comparison of aggregate infant mortality rates in England, France, Sweden, Germany, and Spain, and with rough estimates of the standard of living in these countries at the end of the eighteenth century.

The first part of Kunitz' explanation, namely the control of epidemics by means of isolation and better hygiene, is much in line with Easterlin's arguments. He argued that technology (in a broad sense) is the main determinant of the great mortality decline (Easterlin 1999). Easterlin argued that that the mortality decline from the beginning of the nineteenth century onwards is not due to economic growth and market developments but to public policy initiatives and new institutions based on new knowledge of disease. This could explain the simultaneous mortality patterns among countries despite differences in economic development.

2.2 Cohort Explanations

Cohort explanations of the mortality decline refer to features that initially affect only certain young age groups but that may have long-lasting effects for the groups involved. We mainly have improvements in childhood conditions in mind, or even conditions during the fetal stage that have lasting effects on health and on the lifespan. Net nutrition is seen as the main determinant of the development of cells, which is most rapid during the fetal stage and which diminishes gradually until the body is fully developed at an age around 20, or a little later. Net nutrition is what is left for the development of cells after nutritional claims needed for other life-supporting claims and claims to sustain work are accounted for. A low net nutrition could thus either be due to a low nutritional intake or to additional claims coming from diseases. Also, many diseases not only make their own claims on nutrients but also make it more difficult for the body to absorb nutrients in general. This is the case for infectious diseases. If cells and organs do not develop, children may become stunted, become shorter in general and less healthy. Thus we can basically differentiate between two types of cohort explanations for the mortality decline, namely (a) increased nutritional intake during the fetal stage and/or early years of life, and (b) decreased claims during the fetal stage or early years of life due to a reduction in diseases in the mother or the child.

The importance of early childhood conditions for later life has probably been well-known since time immemorial. The idea is that each generation seems to carry along the same relative mortality from childhood to old age. Both medical and historical research on this matter has expanded rapidly over the last couple of decades (for an overview, see Elo and Preston 1992).

In this connection, the work by Barker (1994 and elsewhere) has been of major importance. He has summarized the medical evidence for the importance of the nutrition of the fetus and newborn infant for the health of the adult. Fogel (1996) has advocated these ideas in historical research perhaps more than anyone else. In addition, Preston and van de Walle (1978) for urban France, and Fridlizius (1989) for Sweden, emphasize the importance of cohort factors for the mortality decline.

Steckel (1983) and Fogel (1996 and previous work) use final body height as a measure of net nutrition and health during childhood. Individuals who had well-nourished and healthy mothers and were well-nourished themselves during the fetal stage have a lower risk of death during infancy. If they are well-nourished and healthy their cells and organs develop better and they reach greater heights and live a longer life. Since net nutrition and not gross nutrition determines health there is no direct link between gross nutrition and height. Improvements in health and height may either be a result of better nutrition (better diet), or fewer claims due to diseases. Thus a decline in the prevalence of smallpox, for example, has a positive effect on height and extends the lifespan, everything else being equal. The problem is that it is difficult to evaluate how much of the improvement in health is due to diet and how much is due to lower claims. Calculating diets for premodern populations is a difficult task (Fogel 1996), and it is even more difficult to calculate claims. Still, historical records show similarities between trends in height and gross consumption (Fogel 1994, 1996), indicating that trend in claims has been of minor importance. If that is the case, then McKeown's focus on gross nutrition can be justified.

Undernourishment, whether due to a low or badly composed intake, or due to an increased claim because of diseases, may lead to a stunting of height or weight and illness, diseases and mortality later in life instead of increasing mortality in the immediate future. The immediate relationship—i.e., the period link between economy and mortality—is therefore much weaker than what Malthus believed, according to Fogel. The rather weak short-term relationship often found between prices and deaths for many European countries, shown by Lee (1981, 1993), Galloway (1988) and

others, supports this interpretation (Fogel 1994). Thus cohort factors matter more to the mortality decline than period ones, according to Fogel.

A warning against an uncritical use of cohort-mortality notions is implied in the findings by Fridlizius (1989), who has shown that the mortality of the Swedish cohorts born in the beginning of the nineteenth century deviated greatly from what would have been anticipated given their mortality at younger ages. He called them *deformed cohorts* and saw the excess mortality for men as due to high levels of alcoholism during the 1830s and 1840s. Such life-styles lifted their mortality above what would otherwise be their 'programmed' level. Other analyses of aggregate cohort mortality data also bring any strong cohort effect of conditions in early life into question.

While Fogel was mainly concerned with chronic malnutrition, we may also note that temporary disturbances in nutrition, in particular during some periods at the fetal stage, may also have long-lasting effects on health, since later improvements cannot always compensate for prior loss (Barker 1994). Therefore, temporary malnutrition due to lack of nutrients or to an increased disease load, whether of mother or of child, may have long-lasting effects on health and lifespan. It has, for example, been shown for Sweden that the disease load in the first year of life has a strong impact on mortality from airborne infections and other diseases in later life (Bengtsson and Lindström 2000). Consequently, elimination of temporary malnutrition, which is still a major problem in many countries of the world, may cause another long-term cohort decline in mortality.

2.3 Mortality Fluctuations

Another feature of the great mortality decline, as observed by demographers, is that the mortality fluctuations became much smaller in the nineteenth century than they were previously. Its consequences for the balance between the population and the economy in the preindustrial society, and also its effects on the mortality decline, have been controversial. There are two schools of thought, and they have reached very different conclusions. According to those who analyze mortality crises (Meuvret 1946, Goubert 1965), the reduction in such crises was a major factor behind the secular decline in mortality. This view gained support during the 1960s and 1970s. Wrigley and Schofield (1981) showed, however, that the elimination of crises accounted for only a small fraction (about 10 percent) of the secular decline in mortality in their analysis based on a large national sample of English parishes. Many others also obtained similar results (see Ortega Osona 2000 for an overview). This view is now questioned by Ortega Osona, who has found that the conclusion of Wrigley and Schofield holds true for England but not for Northern Italy and New Castile, where the reduction of the variability in mortality is important for the initial decline in mortality.

3. Summary

Few scholars today will argue that a single factor is the main determinant of the great mortality decline. Of course, it is no coincidence that the greater resources made available by the transformation of our economies in the eighteenth and nineteenth centuries runs parallel with the great mortality decline and with the fertility transition as well. In a millennial perspective, these events took place at about the same time. That is not to say that there is a close relationship between the economy and the great mortality decline. On the contrary, economic growth is not likely to be even a major determinant, neither before the decline nor in its initial stages, and perhaps far less than expected during its later stages. Instead, the causes are multifactorial and vary from the start to the end of the decline.

Parts of the initial decline may very well be the results of pure luck—a spontaneous decline of the aggressiveness of smallpox pathogens that were a part of an old demographic pattern and not a result of a modernization process. The development and compulsory use of vaccine surely prevented the disease's return. Improvements in nourishment and in the care of mothers and children had long-lasting effects on the lifespan. Advancements in water supply and sanitation, as well as better housing, contributed to the decline from the second part of the eighteenth century onwards. Medical development in the twentieth century prolonged life. The fact that health is determined by net nutrition, intake minus claims, and that claim is due partly to diseases, makes it difficult to evaluate the determinants of the mortality decline. The influence of conditions in childhood, as well as temporary ones on mortality at older ages, also adds to the complexity. The impact of each individual factor is therefore very difficult to measure and requires a long series of high quality data. Several variables, like the virulence of pathogens and the claims on nutrients due to disease, can at best be estimated indirectly, if at all. The analyses based on highly aggregate longitudinal data have identified the problems and directed us towards the solutions. It is, however, more doubtful as to whether solutions will be found at that level of analysis. Perhaps the use of longitudinal microdata will serve as an important complement, as it has in other areas.

Thus the cause of the great mortality decline is multifactorial and the importance of the various factors changes over time. The factors are also involved with each other. For example health is both a cause and a consequence of net nutrition, and this

makes it extremely difficult to measure the impacts of each factor. Still it seems very unlikely today that the initial part of the great mortality decline was a result mainly of an increasing standard of living and of intervention by man. From the nineteenth century onwards, we are certain that improved nourishment and various interventions have an impact on the decline. One might therefore argue that the initial decline is not a result of modern developments but rather is a part of an old pattern, where the level of mortality varied. If this pattern would have continued, one could have anticipated an upswing in mortality during the latter part of the eighteenth century, and in fact such an upswing started in several European countries around the middle of the century, although it was only of limited strength. At present, we know that improved nourishment, personal care, and various public health measures mattered. Thus, we may speculate that it was not the initial mortality decline that was a result of the development of modern society, but rather the absent increase. This leads us to a new generalization of the demographic transition, where the pretransition stage is characterized by changing levels of mortality, not controlled by man, and population growth varies as do nuptiality and fertility. From around the middle of the eighteenth century this pattern was broken and man determined levels of mortality as well as fertility.

See also: Cohort Analysis; Demographic Transition, Second; Infant and Child Mortality: Central and Eastern Europe; Infant and Child Mortality in the Less Developed World; Malthus, Thomas Robert (1766–1834); Mortality, Biodemography of; Mortality, Epidemiological, and Health Transitions; Public Health

Bibliography

Barker D J P 1994 *Mothers, Babies and Disease in Later Life*. BMJ, London

Bengtsson T 1992 Den demografiska utvecklingen i de nordiska länderna 1720–1914. *Lund Papers in Economic History* **15**

Bengtsson T, Lindström M 2000 Childhood misery and disease in later life: The effects on mortality in old-age of hazards experienced in early life, southern Sweden, 1760–1894. *Population Studies* **54**(2): 263–77

Bengtsson T, Oeppen J 1993 A reconstruction of the population of Scania 1650–1760. *Lund Papers in Economic History* **32**

Bengtsson T, Ohlsson R 1994 The demographic transition revised. In: Bengtsson T (ed.) *Population, Economy, and Welfare in Sweden*. Springer-Verlag, Berlin, pp. 13–36

Davis K 1945 The world demographic transition. *Annals of the American Academy of Political and Social Sciences* **237**: 1–11

Easterlin R A 1999 How beneficent is the market? A look at the modern history of mortality. *European Review of Economic History* **3**(3): 257–94

Elo I, Preston S 1992 Effects of early-life conditions on adult mortality: A review. *Population Index* **59**(2): 186–212

Fogel R W 1994 The relevance of Malthus for the study of mortality today: Long-run influences on health, mortality, labour force participation, and population growth. In: Lindahl-Kiessling K, Landberg H (eds.) *Population, Economic Development, and the Environment*. Oxford University Press, Oxford, pp. 231–84

Fogel R W 1996 The escape from hunger and premature death 1700–2100: Europe, America and the Third World. The 1996 Ellen McArthur Lectures, Cambridge

Fridlizus G 1984 The mortality decline in the first phase of the demographic transition: Swedish experiences. In: Bengtsson T, Fridlizius G, Ohlsson R (eds.) *Pre-Industrial Population Change: The Mortality Decline and Short-Term Population Movements*. Amqvist & Wiksell International, Lund, Sweden, pp. 71–114

Fridlizius G 1989 The deformation of cohorts: Nineteenth-century mortality decline in a generational perspective. *Scandinavian Economic History Review* **37**(3)

Galloway P R 1988 Basic patterns in annual variations in fertility, nuptiality, mortality and prices in pre-industrial Europe. *Population Studies* **42**: 275–302

Galloway P R 1994 A reconstruction of the population of North Italy from 1650 to 1881 using annual inverse projection with comparison to England, France, and Sweden. *European Journal of Population* **10**: 223–74

Goubert P 1965 Recent theories and research in French population between 1500 and 1700. In: Glass D V, Eversley D E C (eds.) *Population in History: Essays in Historical Demography*. Arnold, London, pp. 457–73

Helleiner K F 1957 The vital revolution reconsidered. *Canadian Journal of Economics and Political Science* **23**: 1–9

Kunitz S J 1986 Mortality since Malthus. In: Coleman D, Schofield R S (eds.) *The State of Population Theory: Forward from Malthus*. Blackwell, Oxford, pp. 279–302

Lee R D 1981 Short-term variation: Vital rates, prices and weather. In: Wrigley E A, Schofield R S (eds.) *The Population History of England, 1541–1871: A Reconstruction*. Arnold, London, pp. 356–401

Lee R D 1993 Inverse projections and demographic fluctuations: A critical assessment of new methods. In: Reher D, Schofield R S (eds.) *Old and New Methods in Historical Demography*. Oxford University Press, Oxford, pp. 7–28

Meuvret J 1946 Les crises de subsistence et la demographie de la France d'Ancien Régime. *Population* **1**: 643–50

McKeown T 1976 *The Modern Rise of Population*. Arnold, London

Notestein F W 1953 Economic problems of economic change. *Proceedings of the Eighth International Conference of Agricultural Economics*. London

Omran A R 1971 The epidemiological transition. A theory of the epidemiology of population change. *Millbank Memorial Fund Quarterly* **49**(4): 509–38

Ortega Osona J A 2000 Determinants of mortality variability in historical populations and its behavioural and aggregate consequences. In: Bengtsson T, Saito O (eds.) *Population and Economy. From Hunger to Modern Economic Growth*. Oxford University Press, Oxford, pp. 279–300

Perrenoud A 1984 The mortality decline in a long term perspective. In: Bengtsson T, Fridlizius G, Ohlsson R (eds.) *Pre-Industrial Population Change: The Mortality Decline and Short-Term Population Movements*. Almqvist & Wiksell International, Lund, Sweden, pp. 41–70

Preston S H, van de Walle E 1978 Urban French history in the nineteenth century. *Population Studies* **32**

Schofield R S 1984 Population growth in the century after 1750: The role of mortality decline. In: Bengtsson T, Fridlizius G, Ohlsson R (eds.) *Pre-Industrial Population Change*. Almqvist & Wiksell International, Lund, Sweden, pp. 17–40

Sköld P 1997 *The Two Faces of Smallpox. A Disease and its Prevention in Eighteenth- and Nineteenth-Century Sweden*. Report no. 12 from the Demographic Data Base, Umeå University, Sweden

Steckel R 1983 Height and per capita income. *Historical Methods* **16**

United Nations 1953 The determinants and consequences of population trends. *Population Studies* **17**

United Nations 1973 The determinants and causes of population trends. *Population Studies* **50**

Wrigley E A, Schofield R S 1981 *The Population History of England, 1541–1871: A Reconstruction*. Arnold, London

T. Bengtsson

Motherhood: Economic Aspects

The role of the female parent is crucial for social reproduction and the economic status of women. Motherhood, though particularly demanding while children are 'dependant,' is a lifelong commitment. The activities and responsibilities involved vary across and within societies, and depend on their prevailing division between mothers and fathers. Although mothers' direct contribution to childrearing is of central importance, it has received less attention than the indirect impact of motherhood on women's paid work. This question arises particularly in industrialized economies, where paid workplace and home are separate. It is relevant to the distribution of earning and purchasing power between men and women, especially those in couples. If motherhood impedes women's accumulation of human capital, or if decisions about having children are influenced by the earning opportunity cost, resource allocation is affected (Cigno 1991). This entry focuses mainly on the impact of motherhood on paid work in contemporary developed market economies, but starts with the topic of mothers' unpaid work.

1. *The Mother's Unpaid Job*

The unpaid rewards and penalties of motherhood may have escaped attention from economists (Rathbone 1917 is an exception) because of the difficulty of valuing them on a money metric. Such quantification could seem inappropriate for a 'labor of love', but the element of selflessness in the 'job description' need not mean that it deserves no credit. The cartoonist, Posy Simmonds (1979) points this out in text for a spoof military-style recruitment poster:

> A mother provides a vital service to the country which often goes *unrecognised* and *unrewarded*. Motherhood costs spare time and the chance to travel. But it does bring much in return: the chance to develop skills and resourcefulness. Above all, it is a *challenge* and a chance to *serve* the country in a worthwhile way?…Today's young mothers have got a *job* to do, whether it's holding the fort single handed, being in the thick of action in a playgroup or fighting dangerous household germs?…*It's a good life in the Mothers, and it's a REAL job too*.

In most societies the mother's 'job' is a multifold combination of manual and managerial tasks, but the 'training' is largely informal and starts when she is a little girl. Besides producing and nurturing infants, duties include taking a frontline interest in child health and education. Mothers usually play a decisive role in the allocation of consumption to their offspring. They also have an important role in a child's psychological and social development. Mothers' contribution to the emotional work of maintaining a family and its social network may last throughout their lives. Motherhood is a necessary condition for becoming a grandmother, in which role many women also contribute to family welfare.

The greater is a mother's own human capital and economic autonomy, the greater is her potential to achieve good outcomes for her children. In poor countries, there is evidence of this in positive association of maternal education and child survival. In developed countries, there is evidence linking cognitive development of children to, *inter alia*, indicators of maternal inputs.

1.1 *Domestic Production and the Domestic Division of Labor*

The New Home Economics (e.g., Becker 1991) recognizes that unpaid time in the home, combined with goods purchased on the market, yields valuable 'home production.' Among home outputs, such as meals and household repairs, are the production and maintenance of the next generation and their human capital. More children require a greater input of parental time. Alternatively, time can be used to improve the 'quality' of a given number. The parent primarily involved is normally the mother. In two-parent families, fathers generally spend less time than mothers in childrearing and other domestic tasks, though they generally provide more cash. In one-parent families, usually lone mothers, all the parenting, and most of the provisioning, falls on one parent.

The tendency for home production to be primarily mothers' business—rather than fathers'—is attributed partly to biology, for example, because of women's ability to breastfeed. Becker (1991) argues that it is reinforced by the scope for specialization (and trade) between partners. The gains are increased if the market pays higher rates to men than to women. The inevitability of such specialization is debatable (Joshi 1998), and its attractiveness to a mother depends on the security of her claim on the father's earnings. In

many parts of the world this has never been very reliable, and in others the structure of paternal financial obligations is weakening (Folbre 1994).

1.2 Other Unquantified Benefits and Costs of Motherhood

Before turning to the more quantifiable costs of motherhood, its benefits must be acknowledged. Mothers (fathers too) usually benefit directly from a loving relationship with their children, and from taking a pride in them. Under some circumstances children provide help, social status and security, or assure a lineage. The kin group, indeed wider society, can benefit from effective child rearing, and so do children themselves. Most of these items may be considered 'noneconomic', but they underlie the 'demand' for children despite their costs. Some of the costs of childrearing are also fairly intangible, such as the worry, fatigue, and the risk of grief. In poor countries there are also still serious health hazards associated with childbirth.

2. Earnings Costs of Motherhood

Besides the direct material costs of material expenditure on children, there are indirect costs of mothers' forgone earnings. Where maternal duties displace productive work, motherhood has an economic 'opportunity' cost. If a mother's earnings would be higher without children, there is an earnings shortfall due to motherhood. Conversely, if she maintains the higher earnings trajectory by not having children, some of her output has been gained at the expense of forgone motherhood.

The time costs of motherhood are not only diverted from paid work. Leisure, sleep and freedom of movement may also be affected. Forgone earnings (and forgone nonpecuniary benefits of employment) do not arise in cases where the woman is not expecting to do paid work in any case, or where she can combine productive work with taking care of her children. Various arrangements for nonmaternal care permit mothers to spend time in paid work, reducing or eliminating the lost earnings, though childcare may have to be purchased.

Thus, the prospect of forgone earnings varies by context, as well as the number and timing of births. It becomes important where childless women would be lifelong full-time members of a labor force, where workplaces are unsuited to simultaneous childcare, and where care by the mother is the major mode of care for pre-school children. If hours of employment are inflexible, mothers may not be able to earn at all, but where nonstandard hours are available, mothers' earnings may be modified. The earnings costs of motherhood are amplified if the opportunity to accumulate earning power through experience and training depends on continuous full-time employment.

2.1 An Example from the USA

Calhoun and Espenshade (1988) attempt to quantify lifetime foregone earnings for US women on the basis of data observed mainly in the 1970s. Allowing for effects on participation, hours and rates of pay, they estimate average forgone gross earnings, for white and black women, by number, timing and spacing of births, and educational level. The central estimate for white women with two children is about 10 percent of potential earnings over the ages 25 to 54, five times greater than the cost for blacks, who reduce market hours less. The estimated earnings costs rise with the woman's education, and, roughly proportionally, with the number of children. Deferring childbearing reduces the earning loss somewhat for whites, but had the opposite effect for blacks, while effects of birth spacing are minor. This study illustrates the diversity of earnings effects implied in one country by the labor market patterns of a particular period. It also points out that the estimate of US$25,000 forgone per (white) child (1981 prices), is very much less than the direct expenditure costs, which the authors put around US$80,000 over the first 18 years. Among other expenditures, this includes the purchase of daycare services associated with the employment of US mothers of preschool children.

2.2 A British Example

A similar exercise was attempted for the UK by Joshi (1990), but estimating typical, rather than average profiles. Based on behavior in 1980, the illustrative mother would remain completely out of employment while she had children under school age. She would return to employment part-time initially and when she eventually resumed full-time, her rate of pay would fall short of her childless counterpart, assumed to earn full-time continuously, and to gain increments thereby. An additional feature of the British model is an extra pay penalty in part-time jobs. Though a common solution to reconciling domestic and market claims on time, these are generally low status, low skill, dead-end jobs, off any career track. Downward occupational mobility on re-entering employment was also common. The estimated foregone earnings are illustrated for two children. On central assumptions, the sum forgone was around half potential lifetime earnings after childbearing at age 25. In further contrast to the US estimates, the costs did not increase in proportion to the number of births. There appeared to be economies of scale in having extra children, when the

pre-school care is self-provided. If births are reasonably closely spaced, the next child can, for some years, 'free ride' on the fact that the mother would be at home anyway. Another contrast to the US estimates is that the earnings costs in the UK are of a similar order of magnitude to direct costs (Davies and Joshi 1995). This partly reflects more purchased child care in the US but perhaps also more medical expenditure.

2.3 Foregone Earnings in Some European Countries

Estimates of typical lifetime earnings in other European countries (Joshi and Davies 1992), based also on data from around 1980, showed further international contrasts. While the trajectories of interrupted lifetime earnings were quite similar to the British in West Germany, the Swedish scenario showed a much smaller dip in lifetime earnings. Swedish institutions permit mothers to maintain labor force attachment through parental leave when their children are infants and, while their children are small, to work reduced hours, at no hourly pay penalty and with cheap, high-quality daycare services. The results for France polarized into those for the high skilled, where mothers of one or two maintain almost continuous full-time employment, and a low-skill scenario involving long absence from paid work. In the first case, there was little loss of earnings until the third child; in the other, the earnings costs were greatest for the first child. This exercise serves to illustrate the importance of public policies in shaping these patterns. Where daycare is subsidized and publicly provided, as in France and Sweden, mothers' engagement in paid work is less perturbed than where mothers must make and fund their own arrangements on the market or informally.

2.4 British Forgone Earnings Updated

The picture of the earnings losses of the 'typical' British mother became out-dated. By the 1990s the maternal employment had risen. 'Family friendly' employment practices, such as flexible hours and career breaks, started to gain currency. Maternity leaves permitted career women to return to their full-time jobs, and they turned increasingly to the market for daycare services from nannies, nurseries or child-minders. Very little of this was subsidized by either employers or state, until the very end of the decade. The US, by contrast, was already subsidizing daycare through tax deductions and Welfare-to-Work Programmes. Informal care by grandmothers, fathers and neighbors remains important but unsubsidized and unregulated in the UK. Davies et al. (2000) present an update of the estimates of foregone earnings in the UK, comparing the pictures generated by employment patterns around 1980 and 1994. The illustrative individuals are also given three educational levels. A polarization among mothers emerges again. For women with low skill, the loss of earnings has if anything increased, for the graduates the costs have diminished, close to zero for one or two children. For the mid-skill woman, forgone earnings have diminished moderately, reflecting her increased propensity to earn while her children are small. In contrast with the earlier US results, forgone earnings vary inversely with education.

2.5 Effects of Motherhood on the Rate of Pay

There have also been studies in several of the differences in rates of pay received by mothers and other women. Such a 'family gap' may become more salient as a source of low pay among women as the gender pay gap narrows. It may arise partly from superior initial earning power of women without children, but also because of interrupted employment experience, and the current constraints of maternal responsibility on productivity or bargaining power. Among seven industrial countries studied in the 1990s, the UK had the largest unadjusted family gap (mothers aged 24–44, 15 percent below the hourly pay of their childless contemporaries, particularly attributable to the part-time penalty in that county (Harkness and Waldfogel 1999)). This contrasted with a wage lead for mothers in Australia, negligible differences in Sweden, Finland and Canada, and more modest gaps in Germany (3 percent) and the US (9 percent). Regression estimates which control for education, age, ethnicity and place of residence find negative effects of increasing number of children on wages in all countries, greatest in the UK, where one child lowers the predicted wage by 9 percent, two children by 28 percent and three by 35 percent. The estimates in other countries are considerably less, for the impact of two children the runners up are 12 percent in Germany, 11 percent in Australia and 8 percent in the US. Countries with more labor force withdrawal by mothers tend to have the larger family gaps in women's pay. The link between employment interruption and low wages for employed mothers is also apparent when comparing women within countries (Waldfogel 1997, Joshi et al. 1999). These analyses suggest that maintaining employment continuity through taking maternity leave protects women from the pay penalties of motherhood, but this is a recent phenomenon. Thus the pay penalties of motherhood are not universal or inevitable. They are affected by decisions of private and social policy.

2.6 Mothers' Forgone Income

Whatever the size and composition of forgone earnings their cost is not necessarily borne only by the mother. In the first place, if she is taxed on her earnings, the state loses taxes. If she pools her net

income with a partner, this pooling will spread the rest of the loss. Her partner is deprived of some share of the cash that would otherwise be earned. Davies and Joshi (1995) show that in the central British case (1980-based) the earnings costs of motherhood are spread roughly equally between the state, the woman and her partner (assuming partners split net income). If the partner does not contribute to the woman's support, or is absent, she bears the earnings cost net of taxes and state benefits paid for children. Lone mothers face a double demand on their time as sole caregivers and sole breadwinners. Societies have tried various packages of policy to relieve this squeeze—state benefits, state childcare, collection of support from fathers, fostering and the support of the extended family.

Where pension entitlements are earnings-related, some of the earnings loss is deferred to old age. Unless pension systems deliberately compensate childrearing, mothers who have had interrupted earnings tend to face old age on lower pension than childless women and men, and their savings may also be lower. This shortfall of income in old age may be recompensed, in cash or services, by the children, depending on their earning power and whether such a claim on them is acknowledged. The issue of mothers' incomes in old age will remain important into the twenty-first century.

3. Conclusions

The costs of rearing children may not all fall on mothers, they are usually shared by fathers and often at least partly subsidized by state services and transfers. They may be cut at the expense of the quality of the upbringing children receive, or, alternatively, if there are efficiency gains from rationalizing adult time inputs. International experience demonstrates that there is already a variety of ways in which paid work and parenthood are combined.

Although it is generally believed that better earning opportunities for women discourage childbearing, or lead to its postponement, the material reviewed here suggests that the terms on which motherhood and paid work may be combined can vary greatly, and hence the tradeoff between them is not immutable. It is affected by the organization of employment, the organization of state services for childcare and education, by the structure of state transfer systems and by the contributions made by fathers in time and cash. A mother's time is an important input into childrearing, but it is not the only one, and women's contribution to rearing children has never been their only one to the economy.

See also: Childbearing, Externalities of; Divorce and Gender; Family and Gender; Family Bargaining; Family Law; Family Theory: Economics of Intergenerational Relations; Family Theory: Economics of Childbearing; Family Theory: Economics of Marriage and Divorce; Family Theory: Feminist–Economist Critique; Fatherhood; Feminist Economics; Feminist Legal Theory; Fertility of Single and Cohabiting Women; Gender and Feminist Studies in Economics; Gender and the Law; Gender, Economics of; Household Production; Land Rights and Gender; Lone Mothers in Affluent Nations; Lone Mothers in Nations of the South; Male Dominance; Marriage and the Dual-career Family: Cultural Concerns; Motherhood: Social and Cultural Aspects; Partnership Formation and Dissolution in Western Societies; Poverty and Gender in Affluent Nations; Repartnering and Stepchildren; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender; Teenage Fertility

Bibliography

Becker G 1991 *A Treatise on the Family: Enlarged Edition*. Harvard University Press, Cambridge, MA

Blau F D, Ehrenberg R G (eds.) 1997 *Gender and Family Issues in the Workplace*. Russell Sage Foundation, New York

Calhoun C A, Espenshade T J 1988 Childbearing and wives' foregone earnings. *Population Studies – Journal of Demography* **42**: 5–37

Cigno A 1991 *The Economics of the Family*. Clarendon Press, Oxford, UK

Davies H, Joshi H 1995 Social and family security in the redress of unequal opportunities. In: Humphries J, Rubery J (eds.) *The Economics of Equal Opportunity*. Equal Opportunities Commission, Manchester, UK, pp. 313–44

Davies H, Joshi H, Peronaci R 2000 Forgone income and motherhood what does recent British data tell us? *Population Studies – Journal of Demography* **54:** 293–305

Folbre N 1994 *Who Pays for the Kids? Gender and the Structures of Constraint*. Routledge, London

Harkness S, Waldfogel J 1999 The family gap in pay: evidence from seven industrialized countries. *CASE paper* 30. London School of Economics, UK

Joshi H 1990 The cash opportunity costs of childbearing: an approach to estimation using British data. *Population Studies – Journal of Demography* **44**: 41–60

Joshi H 1998 The opportunity costs of childbearing: more than mothers' business. *Journal of Population Economics* **11**: 161–83

Joshi H, Davies H 1992 Daycare in Europe and mothers' foregone earnings. *International Labour Review* **131**: 561–79

Joshi H, Paci P, Waldfogel J 1999 The wages of motherhood: better or worse? *Cambridge Journal of Economics* **23**: 543–64

Rathbone E 1917 The remuneration of women's services. *Economic Journal* [Reprinted in *Population and Development Review* 1999 **25**: 145–158]

Simmonds P 1979 *Mrs Weber's Diary*. Jonathan Cape, London

Waldfogel J 1997 Working mothers then and now: a cross cohort analysis of the effects of maternity leave on women's pay. In: Blau F D, Ehrenberg R G (eds.), *Gender and Family Issues in the Workplace*. Russell Sage Foundation, New York, pp. 92–126

H. Joshi

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Motherhood: Social and Cultural Aspects

To understand motherhood one needs to look at how it is personally experienced, socially organized, and talked about publicly and among scholars. Changes in women's lives, contested understandings of gender, sexuality, family, new reproductive technologies, and political debates on the welfare state, and on population, development, and environment show that there is little agreement on the meaning of motherhood. In the scholarly literature, the recognition of diversity has replaced the search for generalizing accounts. In many accounts, the framing has shifted subtly from that of oppression to motherhood as cultural resistance. The assumed determinacy of biology has been abandoned, although issues of the body and biology return again and again to unsettle social and cultural explanations. Methodologically, research shifted from outsider, expert stories to a focus on experiences of mothering and of being mothered. Publicly, talk of motherhood is framed by the idea of choice. Scholars pursue increasingly complex questions about motherhood through hotly debated theories of embodiment and biology, sex and gender, subjectivity and relationships, connectedness and autonomy, and the ethics of care, crossing disciplinary borders, and blurring the lines between science, philosophy, and literature. The area is truly interdisciplinary. For some, the social and cultural dimensions of motherhood are produced discursively. The concept of motherhood is now often mined as much for its symbolic and ideological significance as for an empirical referent.

Scholarly theories shape cultural meanings of motherhood and the politics of motherhood shapes scholarship. The story told here is a situated account that centres feminist work and directs attention to important connections rather than providing an exhaustive review. The empirical relationship between changes in how motherhood is organized, experienced, and debated should be located in different regions' own historical contexts. Yet, in all disciplines, discussion of motherhood has been shaped by feminist politics and scholarship. Feminist scholarship developed in the context of universalizing and biologically-based accounts in which motherhood appeared as natural, unchangeable, private, and thus outside politics and culture (de Beauvoir 1960, O'Brien 1981). Historically, dualistic understanding of the relationship between nature and culture in the West had meant that motherhood and nature were associated culturally and devalued empirically (O'Brien 1981, Rothman 1989). If motherhood was outside politics and culture, there was little to discuss or change. In practice, poor or minority women's fertility and mothering had long been of public and political interest. Feminism, New Left, Gay and Lesbian, and other social and counter cultural movements challenged the separation of the social and natural, public and private, turning motherhood, family, sexuality, and culture into political issues and giving motherhood the new contested cultural meanings it has today.

How did the social meanings of motherhood become so contested? British scholars (Boulton 1983, Oakley 1979) explained that thinking had been dominated by biologically and psychoanalytically oriented theories that saw mothering as largely instinctual, the fulfillment of mature femininity, and adaptively tailored to the needs of the uniquely dependent human infant. These views were inaccurate, they argued, and helped make mothering oppressive. Historical research became a way of testing and challenging taken-for-granted meanings of motherhood. It shows, for example, that motherhood was not always emotionally or socially central in mothers' lives (Badinter 1981), and anthropological research documents the cross cultural variability in birth mothers' responsibility for children. Then why does an expectation of 'intensive mothering' (Hays 1996) organize many Western mothers' experience? Cross-culturally and historically, motherhood and children have enormously variable social meanings. One can think of children as 'cultural objects,' and thus the meaning of those who mother these cultural objects varies with the meaning of children. One can see pressure towards intensive mothering as rooted in the cultural idealization of children as sacred and innocent (Zelizer 1985) that emerged in the nineteenth century. Intensive mothering, others add, serves both capitalism's interests, and men's. Behind the general idealization, children of different race/ethnic mothers can be valued differently in terms of investment in their health, education, or welfare. Children's different social worth and their mothers' differently valued maternal identities may be inscribed in social policy (Solinger 1992). More disruptive to common sense understandings is that giving birth does not necessarily mean that a woman is socially or legally recognized as a mother. Women are often encouraged to think about pregnancy and maternity as part of an achieved identity which only some women have the right to claim and which it is their responsibility to get right. Scholarship also shows that popular understanding of motherhood in terms of choice is misleading, and hides the lack of choice open to many women.

Sociological research first exposed mothers' role conflict and role strain (Bernard 1974, Lopata 1971) and later showed the new forms of inequality mothers face as they manage both paid work and motherhood (Hochschild 1989). Mothering can be understood as unpaid, culturally devalued reproductive and caring labor in a society characterized by a separation of public and private spheres and dominated by market exchange value. But this leaves mothering as a relationship under-theorized. Moreover, the framing doesn't always fit. It misses the interdependence of public and private spheres and the fact that caring isn't always private but may be low paid work done by

race/ethnic minority persons. Rich (1976) distinguished between the institution of motherhood that was oppressive, and the relationship of mothering which was a potential source of joy and strength. Psychoanalytic and psychological accounts had theorized the mother–child relationships. But they had tended to universalize a culturally specific view of child development that emphasize the need for uninterrupted nurturance from a solitary psychological, always available, parent (usually the mother) over the importance of a network of caring relationships in children's lives. More sociological accounts challenged these ideas for their sexism, ethnocentricism, and unrealisability. Ironically, in the late nineteenth and early twentieth centuries, the privileging of the mother–child bond was part of cultural changes that enabled mothers to challenge fathers' ownership of children, allowing them to claim custody in cases of separation and divorce, and later opening professional opportunities for women in social work and child welfare. By the 1980s, feminist analyses (Benjamin 1988, Chodorow and Contratto 1992) were explaining that prevailing theories of child development contributed to a cultural fantasy of the perfect mother that devalued the actual work mothers did and encouraged mother blaming. From Benjamin (1988) one learned that children's recognition of the mother's autonomous subjectivity was important for healthy child-development; thus mothers' demands for some autonomy were not merely the demands of selfish women. Cross-cultural research and the experiences of race/ethnic minorities in the US (Collins 1994, Glenn et al. 1994) also show that shared caring, mothering other women's children, or being an economic provider are considered part of good mothering in different contexts. To make intensive mothering central to social definitions of good mothering was to privilege idealizations of white middle-class family life and was race and class biased. Research shows that prevailing academic theories about good mothering become embodied in policies and laws which regulate and control mothers, especially those who don't fit the white, middle-class, heterosexual family model on whose idealized experience the theories were first based.

Early research exposed the oppressions in mothers' lives that cultural romanticization made invisible (Bernard 1974, Firestone 1970, Lopata 1971, Oakley 1979, Rich 1976). Research also showed that many women in Western industrialized countries were marrying later, having fewer children, having them later in life and thus spending less of their lives actively involved in mothering, living in a far greater variety of family arrangements (Eichler 1983, Stacey 1990), yet retaining motherhood as a central identity in their lives (McMahon 1995). Also, more women were having children outside legal marriage, fewer 'unwed' mothers were giving children up for adoption and childlessness was talked about as a valid option. In the US, the economic costs of parenthood have risen and are increasingly privately borne by women. Most Canadian and US mothers of preschool children are now employed outside the home, but still do most of the work of mothering, which is framed by an ideology of intensive mothering. Mothers often embody the cultural tensions of their societies, trapped in the contradictions between the ideals, values, and logic that organize the world of paid work and those that are believed to organize family life. Whether they are mothers or not, women remain defined in relation to motherhood. Despite romantic ideologies, mothers and their children may be treated as of little public value, given little support, be isolated or marginalized, or blamed for a variety of society's ills. The gender-neutral term of parent often replaces the use of mother or father. However, in the context of unequal responsibility for the care of dependent children and unequal access to the resources necessary to meet these responsibilities, Fineman (1995) explains, the gender neutrality of social policy discourse and of family law harms mothers and children. Mothering may be seen as maternal practice and a positive discipline that is oriented towards the preservation of children's lives, the fostering of emotional and intellectual development, and preparing children for their roles in life by making them socially acceptable (Ruddick 1989). Thus, it is useful to look at mothering as a relationship of care and responsibility rather than a genetic one. By centring the experiences of white middle-class mothers in heterosexual families in emerging studies of mothering, however, there was danger of creating new kinds of universalizing and ideological accounts. Until recently, the diversity of mothering remained largely invisible. Only slowly, for example, did mothers' experiences under slavery, of leaving their own children to care for privileged women's children, of mothering in a racist society, or in war zones, or civil war, of migration and immigration, of shared community responsibility for childrearing, of extreme poverty and infanticide, of lesbian mothers' struggles for custody, that is, of the roles of race, class, heteronormativity, and poverty in organizing mothering enter dominant understandings of motherhood.

Feminism earned the misleading reputation of being anti-motherhood (Umansky 1996) and of promoting selfishness. How could one value mothering and also criticize it—even if the criticisms were really of compulsory motherhood and the patriarchal conditions under which women mothered? Chodorow (1978) clearly values motherhood yet asks why women want to mother. She argues that the proclivity to nurture becomes embedded in women's personality because, unlike boys for whom separation and differentiation from the mother are central to psychosexual identity development, girl's identity develops around a continuity of connectedness with the mother. Girls, therefore, grow up with greater relational capacities

and needs than men and with psychological senses of self-in-relationship that directs them towards motherhood. She explains that men need to become equally engaged in parenting for their own growth and development, so that men and women can be both nurturant and autonomous and the world a better place. Critics argue that this reduces social change to psychological rather institutional transformation. Chodorow identifies the social value of care and of subjectivities grounded in connection and relationships rather than in masculinist values of separation and autonomy. French feminism/postmodern feminism attempts a different kind of appreciation of the maternal. The importance of a potentially libertory prelinguistic, imaginative, chaotic, pre-oedipal phase of union with the mother is recovered from an abject existence and the privileged role of the oedipal phase in child development is recast as entry into the regulatory symbolic order, a repressive world of logocentric, phallocentric language culturally symbolized in and by 'the father.' In other words, new scholarship profoundly challenges the cultural and academic devaluation of the maternal.

The 1980s brought increasingly positive experience-centered accounts of mothering. Literary work by African-American and other ethnic/race minority women in the US honored their mothers and their own experiences as daughters. Research on the social, not just private, value of caring and caregiving proliferated. For Ruddick (1989), maternal practice could develop alternative values of care and nurturance capable of resisting oppression and conflict. Motherhood, as the physical embodiment of connectedness, challenged the dominant political culture of separation and individualism that underlies the commodification of human relationships, Rothman (1989) argued. Mothering allowed oppressed communities to survive and struggle against inequality (Collins 1994). The political potential of motherhood to rehabilitate social relationships of care and connectedness (Fineman 1995, Rothman 1989), to culturally resist commodification (Hays 1996), and build community (Umansky 1996) have been persistent themes in feminist work.

At first the move towards a positive re-evaluation of mothering and unpaid caregiving positioned gender as the key explanatory variable over race, class, or heterosexuality. By virtue of being mothers, do women develop special knowledge and moral qualities? Mothers' own accounts often express this belief. Were these more positive scholarly accounts of motherhood essentialist and did they universalise maternal nurturance as constitutive of (all) women's identities and reproduce ideological stories that oppressed mothers? Although research showed that mothers tend to experience their mothering as an expression of love, mothers can have multiple and different relationships with children, whether instrumental, contractual, caring, moral, or even violent. The question lingers, however, as to whether there is something unique in the activities of caring, empathy, attentiveness, and responsiveness to children (or dependent others) that could develop social values of caring, enrich democratic society, and make the world a better place. The difficulties of juggling optimistic and negative interpretations of motherhood, of emphasizing the struggle for equality in terms of sameness or recognizing difference (the differences mothering makes in women's lives, or differences among mothers, for example), of minimizing sex differences or reclaiming motherhood as special, remain a central tension or paradox in feminist analyses of motherhood.

New theories about sex and gender, about situated, embodied subjectivities, about socially constructed emotions and experience problematize motherhood. In new research, minority women's experience is increasingly centered, difference is depathologized, gender and the body retheorized, sex/gender systems analyzed as power regimes, motherhood seen as a gender masquerade, and mothers' sexual identities reclaimed. More is known about the diversity of mothering; about how race/ethnic minority mothers teach their children to resist and survive in a racist society, how mothers of children killed in war may claim political agency and public voices as mothers, and how desperately poor mothers in a Brazilian suburb 'let go' and neglect sickly children. We learn that mothers of disabled infants may feel themselves blamed for producing a less than perfect baby and define their children as special gifts, and that while new reproductive technologies may change women's experience of pregnancy, surrogate mothers and gamete donors and their recipients can resist economic and technological definitions of their unusual relationship by defining the exchange as the gift of life. Young alone mothers can resist the cultural association of heterosexuality and motherhood and reject the social disparagement of their maternity and see themselves as mature and responsible parents. Lesbian mothers, who may have to struggle for custody of their children, can resist but also reproduce conventional understandings of motherhood. Mothers struggling with postpartum depression can use illness to politicize motherhood, and unsettle prevailing theories about social movements. Understandings of motherhood are expanded by learning that nonbirth mothers may 'othermother' children from their communities, infertile women may be blamed for their infertility, prospective adoptive parents may try to naturalize adoption by waiting for their baby 'out there,' and childless women are often still seen as selfish. In contrast to the diversity and despite gender neutral talk of parenting, mothers are still held responsible for their children in ways fathers are not and, unlike in many Scandinavian countries, there is still little social policy support for childrearing in the US. Scholarly analysis often inverts popular fears about mothers becoming too individualistic by asking if the real risk to a caring society may come from the individualism

that is embedded in social and economic restructuring in accordance with market forces.

Despite motherhood's aura of intimacy, its meanings are political and metaphoric. Talk about motherhood carries coded messages about race, class, gender and sexuality, epistemology, political order, the meaning of the good society, and so on. The abortion debates in the US were about the meaning of motherhood and its place in women's lives (Ginsberg 1989). Moral panics about unwed mothers in Great Britain were about immigration, race, patriarchal heterosexuality, public responsibility for child rearing, and the social role of the State (Phoenix et al. 1991, Silva 1996). Lone mothers were often blamed for Britain's 'moral decay' and we can see the withdrawal of public support for lone mothers and shifting costs to absent fathers as part of the dismantling of the welfare state.

The motherhood of women in the poorer nations of the South is often talked about in terms of population, development, and environment. Population control policies embedded in development aid packages may organize maternity. Some feminists from the South (Akher 1999) turn to community and women's indigenous gendered knowledge rather than to notions of individual rights for liberatory analyses of motherhood. Others caution that although birth control may be coercively encouraged by some patriarchal nation states in the South and financial and development organizations in the North, it is a mistake to see modern technologies as inevitably patriarchal and unhelpful to women or to dismiss contraceptive technologies, individual choice or reproductive rights as western impositions. It might be better to recast Western rights discourse in a more 'referential universe,' and thereby transform the classical liberal rights model so that it emphasizes the social nature of rights and the multiple nature of different individuals' identities (Correa and Petchesky 1994).

Motherhood is a difficult issue not only because of its unstable empirical and politicized meanings but also because representations of motherhood, even in academic work, are often projections of individual and collective identities, what we feel is important to society, and the kind of persons we like to think women (and men) are, or are capable of being. Thus, despite all the changes, talk of motherhood still comes very close to the heart.

See also: Androgyny; Families and Households, Formal Demography of; Family and Gender; Family Bargaining; Fatherhood; Feminist Theory; Gender and Feminist Studies; Gender, Class, Race, and Ethnicity, Social Construction of; Gender Ideology: Cross-cultural Aspects; Lone Mothers in Affluent Nations; Lone Mothers in Nations of the South; Male Dominance; Masculinities and Femininities; Motherhood: Economic Aspects; Nation-states, Nationalism, and Gender; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender; Sex Differences in Pay; Sex Segregation at Work; Sexuality and Gender; Small-group Interaction and Gender; Social Movements and Gender; Social Welfare Policies and Gender; Time-use and Gender

Bibliography

Akher F 1999 Reproductive rights: A critique from the realities of Bangladeshi women. *Reproduction* 1: On-line journal. Global Reproductive Health Forum@Harvard www.hsph.harvard.edu/organizations/healthnet/sasia/index.html

de Beauvoir S 1960 *The Second Sex*. Four Square Books, London

Benjamin J 1988 *The Bonds of Love: Psychoanalysis, Feminism, and the Problem of Domination*. Pantheon Books, New York

Bernard J 1974 *The Future of Motherhood*. Dial Press, New York

Badinter E 1981 *Mother Love: Myths and Reality*. Macmillan, New York

Boulton M G 1983 *On Being a Mother*. Tavistock, London

Chodorow N 1978 *The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. University of California Press, Berkeley, CA

Chodorow N, Contratto S 1992 The fantasy of the perfect mother. In: Thorne B (ed.) *Rethinking the Family: Some Feminist Questions*, rev. edn. Northeastern University Press, Boston

Collins P 1994 Shifting the centre: Race, class and feminist theorizing about motherhood. In: Glenn E N, Chang G, Forcey L R (eds.) *Mothering: Ideology, Experience, and Agency*. Routledge, New York, pp. 45–65

Correa S, Petchesky R 1994 Reproductive and sexual rights: A feminist perspective. In: Sen G, Germain A, Chen L C (eds.) *Population Policies Reconsidered: Health, Empowerment, and Rights*. Harvard University Press, Boston

Eichler M 1983 *Families in Canada Today*, 2nd edn. Gage, Toronto

Fineman M A 1995 *The Neutered Mother, the Sexual Family and Other Twentieth Century Tragedies*. Routledge, New York

Firestone S 1970 *The Dialectic of Sex*. Bantam, New York

Ginsberg F 1989 *Contested Lives: The Abortion Debate in an American Community*. University of California Press, Berkeley, CA

Glenn E N, Grace C, Forcey L R (eds.) 1994 *Mothering: Ideology, Experience, and Agency*. Routledge, New York

Hays S 1996 *The Cultural Contradictions of Motherhood*. Yale University Press, New Haven and London

Hochschild A, Machung A 1989 *The Second Shift: Working Parents and the Revolution at Home*. Viking, New York

Lopata H 1971 *Occupation Housewife*. Oxford University Press, New York

McMahon M 1995 *Engendering Motherhood: Identity and Self-transformation in Women's Lives*. Guilford, New York

Oakley A 1979 *From Here to Maternity: Becoming a Mother*. Penguin, Harmondsworth, UK

O'Brien M 1981 *The Politics of Reproduction*. Routledge and Kegan Paul, Boston

Phoenix A, Woollett A, Lloyd E 1991 *Motherhood: Meanings, Practices and Ideologies*. Sage, London
Rich A 1976 *Of Woman Born: Motherhood as Experience and Institution*. Norton, New York
Rothman B K 1989 *Recreating Motherhood: Ideology and Technology in a Patriarchal Society*. Viking, New York
Ruddick S 1989 *Maternal Thinking: Toward a Politics of Peace*. Beacon, Boston
Silva E B (ed.) 1996 *Good Enough Mothering*. Routledge, London
Snitow A 1992 Feminism and motherhood: An American reading. *Feminist Review* **40**: 32–51
Solinger R 1992 *Wake up Little Susie: Single Pregnancy and Race before Roe vs. Wade*. Routledge, New York
Stacey J 1990 *Brave New Families: Stories of Domestic Upheaval in Late Twentieth Century America*. Basic Books, New York
Umansky L 1996 *Motherhood Reconceived: Feminism and the Legacies of the Sixties*. New York University Press, New York
Zelizer V 1985 *Pricing the Priceless Child*. Basic Books, New York

M. McMahon

Motion Perception Models

A motion perception model is a theory or computation that relates visual stimuli or visual scenes that contain motion to the motion perceptions and motion-related actions of observers. We restrict ourselves here to formal mathematical and computational models. We can know about another person's perceptions only by his or her observable responses. Therefore, the input to a motion model is a quantitative description of a stimulus or scene, the output is a predicted response. In a mathematical motion model, a stimulus (such as a moving sinewave grating) is described by an equation. For example, a mathematical theory might make a prediction, for sinewave gratings, of the probability that an observer could correctly discriminate a particular leftwards-moving grating from a rightwards-moving grating.

In a computational theory, the stimulus typically is described in terms of an x, y, t cube (see Fig. 1a). For example, the scene facing a pilot landing an airplane can be described in terms of the luminance at each point (x, y) of the pilot's visual field as a function of time t. The theory might predict the pilot's perceived orientation and velocity relative to a runway. In a psychophysical experiment (in an aviation simulator), the perception would be measured by asking the pilot to set a marker to the spot at which he believes the aircraft is heading at the moment of testing. In a more complex simulation, the theory would predict the manipulation of the controls in a real or simulated aircraft.

Motion theories themselves break down into components: pure motion processing, and subsequent perceptual, decision, and motor components. Even the pure motion processing is now believed to consist of several stages: motion direction is computed first, then velocity, then more complex motion computations that incorporate velocity at many locations to derive velocity gradients—curl, shear, and divergence—and finally, structure-from-motion, a computation in which the 3-D structure of an object or environment is derived from the complex properties of the 2-D motion field.

1. Historical Background

The formal psychophysical study of motion perception is usually traced to Exner (1875), who produced a spark at one spatial location x_1 followed after a short interval Δt by a spark at an adjacent location x_2. For a wide range of distances Δx between the locations and of time intervals Δy, observers perceived motion between the two locations. This is often called apparent motion (in contrast to real motion), because there is no instantaneous movement of the stimulus. Exner's observations, elaborated by Wertheimer (1912), ultimately became a cornerstone of Gestalt Psychology as the Phi phenomenon: The whole, which includes a perception of motion, is more than the sum of the parts—two static flashes.

One might have expected that demonstrations of stroboscopic motion in the 1830s, which preceded Exner, and the later advent of motion pictures early in the twentieth century would have demystified apparent motion. It is perfectly obvious that as the number of frames per second in a motion picture increases, its representation of motion more and more closely approximates real motion. Phi is a special case of only two frames of a movie that depicts motion. Although a simple theory that encompasses both sampled (apparent) and continuous (real) motion was published by Fourier (1822), it was not applied to motion until more than 150 years later.

2. Fourier Analysis of Motion Stimuli

Fourier analysis is now regarded as the default motion model, or better perhaps, as the default description of motion itself. As noted above, when color is neglected, a stimulus is characterized by the luminance falling at a point x, y at time t in the 2-D visual field. Suppose the x dimension is sampled at n_x points, the y dimension at n_y points and the t dimension at n_t points. Then it requires $N = n_x \times n_y \times n_t$ points to describe the visual stimulus. In computer terminology, one would say the n_x, n_y, n_t stimulus cube consists of N voxels (units of volume). Whereas visual space is three-dimensional (it takes three numbers to describe the location of a point in 3-D space), the dimensionality of the stimulus cube is N, because N numbers are required to describe it.

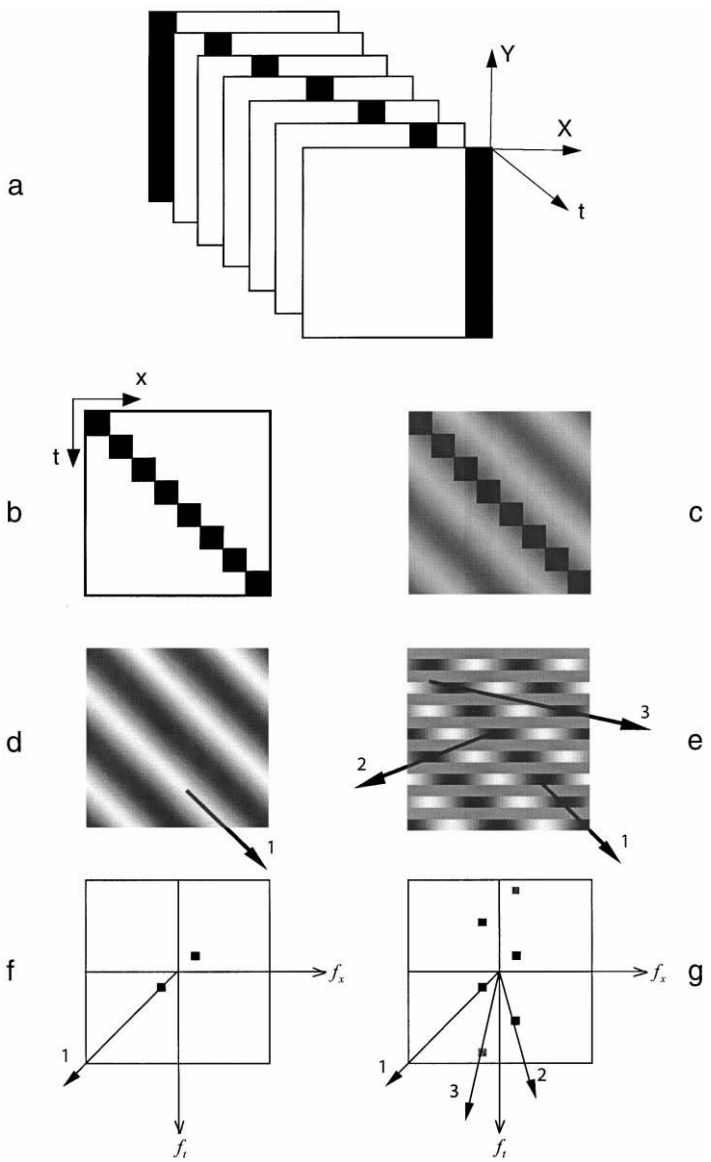

Figure 1
(a) A stimulus cube illustrating eight successive frames in the left-to-right movement of a vertical black bar. (b) An x, t cross section of (a). (c) Same as (b) with a sinewave (the dominant sinewave component in (a) superimposed. (d) x, t cross section of a sinewave grating that is moved continuously. The arrow represents the direction of movement. (e) x, t cross section of a sinewave grating that is sampled every 90 degrees. The arrows represent directions of movement. (f) Fourier amplitude spectrum of (d). The axes represent the frequencies f_x and f_t of the Fourier components. The weight of the points represents the amount of the indicated component. The arrow(s) represent the Fourier component(s) corresponding to the direction(s) of movement as (e). (g) Fourier amplitude spectrum of (e). Axes similar to (f) (after Chubb and Sperling 1985 and 1989)

A discrete Fourier transform describes an n_x, n_y, n_t stimulus cube as the sum of $N = n_x \times n_y \times n_t$ sine and cosine component waves. The Fourier representation is equivalent to a rotation in N-dimensional space. That is, the same space-time stimulus is simply viewed from a different vantage point; all the relational properties within the stimulus remain unchanged. The Fourier representation of a sampled motion stimulus

(e.g., a Phi stimulus or a motion picture) has all the same components as the continuously moving stimulus plus some extra components that represent the sampling function (Fig. 1). Watson et al. (1986) propose a useful rule of thumb to determine whether an observer can discriminate a continuously moving from a sampled stimulus: Discrimination requires that at least one of the extra temporal frequencies introduced by sampling is less than 30 Hz and less than 30 cycles per degree (cpd).

The Fourier representation of the typical Phi apparent-motion stimulus contains many sampling-produced frequencies below 30 Hz and 30 cpd. Therefore, according to the Fourier model, Phi is easily discriminable from a continuously moving stimulus. But it is perceived as moving because it also contains the same motion components as a continuously moving stimulus.

3. *Reichardt Detector for First-order Motion*

The Fourier model is a descriptive theory. It does not specify the actual processes by which motion is perceived. The first computational motion process theory was proposed by Reichardt (1957, 1961) to account for motion-induced responses of the beetle, *Chlorophanus*. Reichardt's theory was adapted for human vision by van Santen and Sperling (1984, 1985). According to their theory, corresponding to every small neighborhood of the visual field, there is an array of Reichardt detectors of different scales (from small to large) and different orientations (from 0 to 179 degrees) that compute motion direction in this neighborhood. Motion is detected according to a voting rule, that is, a rule for combining the outputs of the detectors. The voting rule that seems to be used for computing motion direction when there are many oppositely directed motions (as is typical in sampled motion), is to choose the direction that has greatest number of Fourier components above threshold, i.e., the greatest number of active Reichardt detectors (Sperling et al. 1989).

Figure 2 shows an elaborated Reichardt detector (ERD). SF_A and SF_B represent spatial filters in adjacent spatial locations. To detect left-to-right motion, the output of SF_A is delayed by filter TF. When an object traveling in the external world from SF_A to SF_B reaches SF_B with the same delay as TF, then the direct signal from SF_B and the delayed signal from SF_A both arrive at the comparator at the same time. Basically, a delay-and-compare computation is at the core of all motion theories; in the Reichardt detector the comparison operation is multiplication (covariance).

It is assumed that the inputs to a Reichardt detector are point contrasts. That is, the contrast of the mean luminance l_0 is taken as zero. A point contrast less than

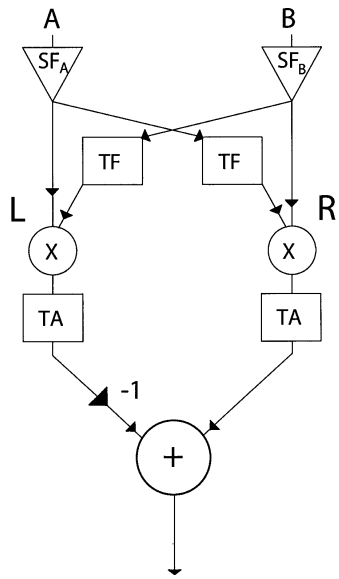

Figure 2
Elaborated Reichardt Detector (ERD). It computes motion direction from two inputs that sample the visual display in two spatial areas A and B. SF_A and SF_B denote the linear spatio-temporal filters (receptive fields) that may have different spatial distributions. In the R ('right') subunit of the detector, the output of SF_A at A is delayed by the temporal delay filter TF and then multiplied ($\times$) with the direct output from SF_B. In the L subunit of the detector, the output of SF_B is delayed by the temporal delay filter TF and then multiplied with the direct output of SF_A. The sign of the difference between the outputs of L and R subunits determines the perceived direction of motion. Outputs greater than zero indicate stimulus motion from A to B; outputs less than zero indicate stimulus motion from B to A (reproduced by permission of the Optical Society of America from Lu and Sperling 2001)

0 means the luminance at that point is $< l_0$, a positive point contrast means the luminance at that point is $> l_0$. If both the straight-through input and the delayed input are positive, or both are negative, then the multiplication produces a positive output indicating motion (in this case) from left to right. If either input is zero, the comparison produces a zero output, and if one input is positive, and the other negative, the ERD signals motion in the reverse direction.

The Reichardt detector has two subunits, one which detects left-to-right motion, the other which detects right-to-left motion. Ultimately, the outputs of these two subunits are subtracted. Thus, a positive output for the Reichardt detector of Fig. 1 indicates left-to-right motion, a negative output indicates right-to-left motion. The subtraction is critical for filtering out certain nonmotion signals that stimulate the subunits.

For example, a simple flickering light (which contains no net motion) stimulates each subunit equally. These subunit outputs are canceled in the final subtraction so that the output of the Reichardt detector is zero for nonmoving stimuli.

In the visual system, prior to motion detection, there is very powerful contrast gain control (Lu and Sperling 1996). To observe the properties of the Reichardt detector in human psychophysics requires using stimuli with sufficiently low contrast (e.g., less than a few percent) so that they are relatively undistorted by the gain control. With such stimuli, several very surprising, counterintuitive predictions of the Reichardt detector have been verified (van Santen and Sperling 1984). For example, adding a uniform-field flicker to a moving stimulus has absolutely no effect on its visibility unless the flicker happens to be of the same temporal frequency. On the other hand, when the flicker has the same temporal frequency as the motion stimulus, depending on the relative phases, the flicker can either enhance or reverse the direction of apparent motion. The successful prediction of a number of such psychophysical results has established the Reichardt computation as the algorithm of human motion-direction detection.

In addition to the Reichardt detector, two apparently different theories of motion have been proposed: motion energy detection (Adelson and Bergen 1985) Hilbert transforms (Watson and Ahumada 1985). These have been shown to be computationally equivalent to the Reichardt detector (Adelson and Bergen 1985, van Santen and Sperling 1985). Consequently, there is only one theory of human motion-direction detection, although it can be framed in different ways.

4. *Velocity Detection*

The output of the Reichardt detector gives only a measure of motion-strength in a particular direction, not a measure of velocity. Although velocity cannot be derived from an individual Reichardt detector, it can be derived from the outputs of an array Reichardt (or equivalent) detectors. Each individual Reichardt detector is optimally stimulated by a particular velocity; an appropriate voting rule enables an array of Reichardt detectors to signal the most likely velocity.

5. *Second-order Motion*

It had long been suspected that motion detection may involve different systems. Following the description of Phi motion, introspectionists proposed other categories (α, β, and γ: Kenkel 1913; Δ: Korte 1915), in addition to Phi to describe experiential aspects of motion perception (for a review, see Boring 1942).

Figure 3
Space-time representations of drift-balanced and microbalanced stimuli that selectively stimulate the second-order motion system. The overlay shows Hubel-Wiesel receptive fields oriented at +45 degrees and −45 degrees and illustrates that both have exactly the same expected outputs. Detection of left-to-right motion (or upper-left to lower-right slant) requires second-order motion (or second-order texture processing). (Adapted from Chubb and Sperling 1989.)

Late in the twentieth century, a number of dual process theories of motion were proposed, the most influential being the short- and long-range motion systems proposed by Braddick (1974). However, it was not possible to evaluate such theories because no explicit motion computation had been proposed, merely incidental properties, such as sensitivity to binocular versus monocular presentation.

With the establishment of the Reichardt model for motion perception, it quickly became obvious that observers perceived strong apparent motion in some stimuli that were completely ambiguous to Reichardt detectors and which (equivalently) contained no useful Fourier motion components. Chubb and Sperling (1988) originally characterized such stimuli as having two properties, drift-balanced and microbalanced. Originally, these stimuli were said to activate a non-Fourier motion system. Later the terminology of Cavanagh and Mather (1989) was adopted. 'First-order motion' refers to motion that can be detected by Reichardt detectors; 'second-order motion' refers to apparent motion that is invisible to Reichardt detectors.

In a drift-balanced stimulus, for every Fourier component that represents motion in one direction, there is another Fourier component, with the same expected amplitude, that represents motion in the opposite direction. In other words, a drift-balanced stimulus cannot convey Fourier motion. A microbalanced stimulus is a stimulus that remains drift-balanced when viewed through any space-time separable window. In other words, every neighborhood or combination of neighborhoods of a microbalanced stimulus is itself drift-balanced. An immediate corollary is that a microbalanced stimulus is ambiguous for every Reichardt detector or combination of detectors. Fig. 3 shows two examples of microbalanced stimuli in

which apparent motion is easily perceived. Indeed, observers cannot discriminate when they perceive motion whether the motion computation is a first-order or a second-order computation.

The model for the detection of second-order motion is the same as the Reichardt model for the detection of first-order motion with one exception. Prior to motion detection, the stimulus is processed by a texture grabber—a linear filter plus a rectifier—that produces a positive output that is proportional to the amount of texture present in a neighborhood. Thus the first-order motion system computes the motion conveyed by photons, the second-order motion system computes the motion conveyed by features (the unit of texture).

6. *Third-order Motion*

There is now a large class of exotic stimuli that are invisible to first- and second-order motion computation in which observers clearly see motion (Lu and Sperling 1995b). The basis for this motion computation appears to be figure-ground. The clearest example is a sequence of frames, each of which contains a square that is perceived as figure and in which the square translates in a consistent direction (Fig. 4). In frame 1, the figure is a red square on a green background, in frame 2, a high contrast texture on a low contrast background, in frame 3 a stereoptically produced square that appears closer to the observer than the background, and so on. The only common element linking these frames to permit a motion computation is that the area perceived as figure moves in a consistent direction.

It has been assumed that the actual motion algorithm, i.e., a Reichardt computation, is the same for all motion systems, only the preprocessing of the input differentiates them. The first-order motion system computes motion more-or-less directly from the raw representation of the stimulus (the luminance cube, $l(x, y, t)$). Second-order motion operates on an input that represents the 'amount of' features $f(x, y, t)$ at each location x, y, t. Third-order motion computes motion from a salience field $s(x, y, t)$ in which positive values of $s(x, y, t)$ denote 'figure' and zero denotes 'ground.'

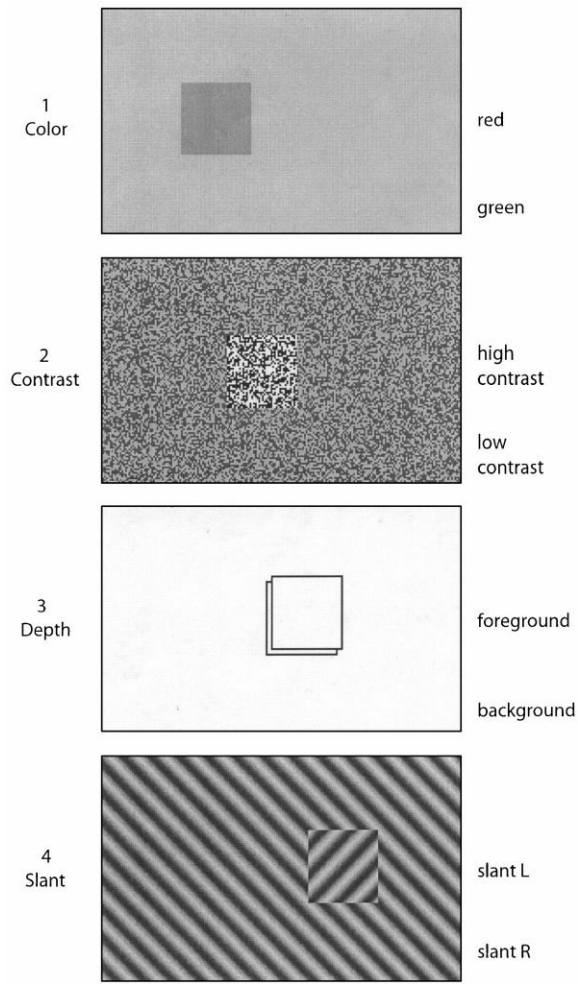

Figure 4
A sequence of frames for stimulating only the third-order motion system (after Lu and Sperling 2001)

7. *Properties of the Three Motion Systems*

Most natural motion, as well as most stimuli used in motion experiments, stimulates all three motion systems. Although it can be difficult to stimulate only one system, such experiments have revealed the properties of each of the three systems. First- and second-order motion seem to be primarily monocular (computed separately for stimuli arriving in each eye) and quite fast with a 'corner' frequency of 10–12 Hz. The third-order system is relatively slower, (corner frequency typically 4 Hz) and exclusively binocular, and is indifferent to the eye of origin. When a monocular image contains a third-order stimulus, but the binocular image does not, the motion cannot be processed by the third-order system (Solomon and Morgan 1999). On the other hand, only the third-order system can decode isoluminant chromatic motion (Lu et al. 1999a, 1999b), and motion in dynamic random-dot stereograms. Unlike first- and second-order motion, third-order is influenced strongly by attention (Lu and Sperling 1995a). Brain trauma can selectively destroy first-order (but not second-order) motion perception in one hemifield (Vaina et al. 1998) or second-order (but not first-order motion) perception (Vaina and Cowey 1996).

8. Structure from Motion, KDE

When a twisted wire is rotated in front of a projection lamp so that it casts a shadow on a screen, the 2-D shadow appears to be a dynamically rotating 3-D object. Originally labeled the kinetic depth effect (KDE) by its discoverers (Wallach and O'Connell 1953), the process of the recovery of the 3-D shape is now known as 'structure from motion.' The shadow demonstration shows that cues such as texture, shading, and object familiarity are not necessary for the recovery of 3-D shape. Ullman (1979) provided the first formal theory for the recovery of 3-D structure from 2-D motion. He was able to prove that three independent views of four identifiable noncoplanar points were sufficient to recover the 3-D structure of a rigid, rotating object. Subsequently, numerous algebraic algorithms have been proposed for the recovery of a rigid 3-D structure from n views of m points. In a technical sense, these algorithms do not involve motion, merely different views. In a practical sense, they fail to recover even slightly nonrigid motion or rigid motion from imperfect data.

An alternative algorithm for deriving 3-D structure from 2-D motion uses a 'hill climbing' optimization procedure. The x, y coordinates of the structure are given in the image, only the z coordinate (the distance from the observer) is unknown. In frame 1, z values are randomly assigned to each point. In general, frame 2 will be inconsistent with a rigid rotation of the object defined by frame one. Therefore, the z-values in frame 2 are perturbed in such a direction as would have made the rotation from frame 1 to frame 2 more rigid. By successive improvements, such an algorithm can converge to a rigid object when one is presented, and can also accurately represent a slowly changing nonrigid object (Ullman 1984).

Imagine black spots painted on a perfectly transparent surface, such as a mound. When such a mound is stationary and viewed from above, it is perceived as a flat, spotted plane. As soon as it starts to move, the 3-D shape of the surface is instantly apparent. This is another example of the KDE. However, suppose that as the surface moves, in each new frame, half the dots are replaced, so that dots have a lifetime of only two frames (Sperling et al. 1989). The two-frame lifetime would cause most algorithms that depend on frame-to-frame dot correspondence to fail. On the other hand, a motion flow-field can be calculated perfectly well in such a display. And observers perfectly well perceive the 3D depth. Indeed, complex properties of the motion flow-field (such as shear) are sufficient to enable (computational) recovery of latent 3D object structure (Koenderink and van Doorn 1986).

9. Heading and Self-motion

When we move in a textured environment, the self-movement generates a motion flow-field on the retina. For the special case in which we fixate on the point towards which we are moving, the flow-field is one of divergence from a focus of expansion, and the focus itself is the point towards which we are moving. However, if we happen to be looking elsewhere, matters are much more complicated. These complications can be mathematically resolved, and it can be shown that such flow-fields contain sufficient information to determine the heading direction (Longuet-Higgins and Prazdny 1980) and other useful quantities, such as time to contact (Lee 1980).

The computation of structure from motion (and presumably also heading direction) has been shown to rely almost entirely on the first-order motion system (Dosher et al. 1989). This is probably a matter of spatial resolution. The computation of complex properties of a flow-field (such as shear and divergence) requires accurate comparison of velocity in adjacent neighborhoods. Computing small velocity differences between large velocities demands more resolution that is available in the second- and third-order motion systems.

See also: Fourier Analysis; Motion Perception: Psychological and Neural Aspects; Sensation and Perception: Direct Scaling; Statistical Pattern Recognition; Vision for Action: Neural Mechanisms; Vision, High-level Theory of; Vision, Low-level Theory of; Vision, Psychology of; Visual Space, Geometry of

Bibliography

Adelson E H, Bergen J K 1985 Spatio-temporal energy models for the perception of apparent motion. *Journal of the Optical Society of America A: Optics and Image Science* **2**: 284–99

Boring E G 1942 *Sensation and Perception in the History of Experimental Psychology*. Appleton-Century-Crofts, New York

Braddick O 1974 A short-range process in apparent motion. *Vision Research* **14**: 519–29

Cavanaugh P, Mather G 1989 Motion: The long and the short of it. *Spatial Vision* **4**: 103–29

Chubb C, Sperling G 1988 Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *Journal of the Optical Society of America A: Optics and Image Science* **5**: 1986–2006

Dosher B A, Landy M S, Sperling G 1989 Kinetic depth effect and optic flow: 1. 3D shape from Fourier motion. *Vision Research* **29**: 1789–813

Exner S 1875 Experimentelle Untersuchung der einfachsten psychischen Processe. *Pflugers Archiv für Gesamte Physiologie des Menschen und der Tiere* **11**: 403–32

Fourier J B J 1822 *Théorie analytique de la chaleur*. Chez Firmin Didot, Paris

Kenkel F 1913 Untersuchungen über Zusammenhang zwischen erscheinungsgross und Erscheinumngsbewegung beim einigen sogenannten optischen Tauschungen. *Zeitschrift Psychologie* **61**: 358–449

Koenderink J J, van Doorn A J 1986 Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America A* **3**: 242–9
Korte A 1915 Kinematoskopische Untersuchungen. *Zeitschrift Psychologie* **72**: 193–206
Lee D M 1980 Visuo-motor coordination in space-time. In: Stelmach G E, Requin J (eds.) *Tutorials in Motor Behavior*. North-Holland, Amsterdam, pp. 281–95
Longuet-Higgins H C, Prazdny K 1980 The interpretation of a moving retinal image. *Proceedings of the Royal Society of London B* **208**: 385–97
Lu Z-L, Lesmes L A, Sperling G 1999 The mechanism of isoluminant chromatic motion perception. *Proceedings of the National Academy of Science* **96**: 8289–94
Lu Z-L, Lesmes L A, Sperling G 1999 Perceptual motion standstill in rapidly moving chromatic displays. *Proceedings of the National Academy of Science* **96**: 15374–9
Lu Z-L, Sperling G 1995a Attention-generated apparent motion. *Nature* **377**: 237–9
Lu Z-L, Sperling G 1995b The functional architecture of human visual motion perception. *Vision Research* **35**: 2697–722
Lu Z-L, Sperling G 1996 Contrast gain control in first- and second-order motion perception. *Journal of the Optical Society of America A*: *Optics and Image Science* **13**: 2305–18
Reichardt W 1957 Autokorrelationsauswertung als funktionsprinzip des zentralnervensystems. *Zeitschrift für Naturforschung* **12b**: 447–57
Reichardt W 1961 Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In: Rosenblith W A (ed.) *Sensory Communication*. Wiley, New York
van Santen J P H, Sperling G 1984 Temporal covariance model of human motion perception. *Journal of the Optical Society of America A*: *Optics and Image Science* **1**: 451–73
van Santen J P H, Sperling G 1985 Elaborated Reichardt detectors. *Journal of the Optical Society of America A*: *Optics and Image Science* **2**: 300–21
Solomon J A, Morgan M J 1999 Dichoptically canceled motion. *Vision Research* **39**: 2293–7
Sperling G, Landy M, Dosher B A, Perkins M E 1989 The kinetic depth effect and the identification of shape. *Journal of Experimental Psychology: Human Perception and Performance* **15**: 426–40
Ullman S 1979 *The Interpretation of Visual Motion*. MIT Press, Cambridge, MA
Ullman S 1984 Maximizing rigidity: The incremental recovery of 3-D structure from rigid and non-rigid motion. *Perception* **13**: 225–74
Vaina L M, Cowey A 1996 Impairment of the perception of second order motion but not first order motion in a patient with unilateral focal brain damage. *Proceedings of Royal Society of London B* **263**: 1225–32
Vaina L M, Makris N, Kennedy D, Cowey A 1998 The selective impairment of the perception of first-order motion by unilateral cortical brain damage. *Visual Neuroscience* **15**: 333–48
Wallach H, O'Connell D N 1953 The kinetic depth effect. *Journal of Experimental Psychology* **45**: 205–217
Watson A B, Ahumada A J 1985 Model of human visual-motion sensing. *Journal of the Optical Society of America A* **1**: 322–42
Watson A B, Ahumada A J, Farrell J E 1986 The window of visibility: A psychophysical theory of fidelity in time-sampled motion displays. *Journal of the Optical Society of America A* **3**: 300–7
Wertheimer M 1912 Über das Sehen von Scheinbewegunen und Scheinkorpern. *Zeitschrift Psychologie* **45**: 205–217

G. Sperling

Motion Perception: Psychological and Neural Aspects

Visual motion perception involves much more than just seeing movement. Though people are not generally aware of it, a great deal also goes on 'behind the scenes,' involving a series of elaborate computational stages that not only detect image movements, but also analyze image velocity patterns to extract surprisingly complex and useful information about the outside world. For example, watching a person walk through a crowd requires calculations, probably done in early stages of cortical processing, that would cripple the fastest of modern computers. The goal of this review is to describe the crucial role motion processing plays in visual perception, to tell how our present understanding of the visual motion system came about, and to suggest exciting future directions that motion perception studies are likely to take.

1. *Background*

Intuitively, it would seem very easy to detect motion. As a car drives past us, we note that its position is not constant; therefore, it must be moving. But the apparent simplicity of this is misleading. Tracking an object involves first recognizing it as an object, a daunting challenge even under stationary conditions. To measure motion by continuously updating a position estimate would require us to parse an image (find the objects), recognize an object (while it is moving), make sure it is the same object as in the last sample, calculate its position, then find the derivative with respect to time (or at least divide by the time since the last update). Even then, how would we know that snow is blowing towards the East, or that bubbles on the surface of a stream flow from left to right? In such cases, there would be no objects to recognize and track.

Fortunately, there is a way of detecting motion that is as simple and robust as sensing the orientation of a line. It does not require object recognition, and it does not require us to know anything about position or time. Wertheimer was perhaps the first to suggest this kind of fundamental motion perception: the raw sense of movement, irreducible to separate senses of object, space, and time (Wertheimer 1912). We will see below that the visual motion system does indeed work like

this. First, however, it is important to understand the fundamental and perhaps unexpected role visual motion plays in our daily existence.

2. Functions of Motion Processing

Arguably the most basic judgment one has to make about the world is what is moving and what is not. This is the most obvious role for visual motion processing. Less obvious, perhaps, is the need for accurate velocity measurements when our eyes track a moving object; it simply does not work to keep adjusting the eye position according to where the target appears on the retina. Even simple things like catching a baseball, filling a cup of coffee, crossing the street, or even interpreting facial expressions rely on motion processing (Zihl et al. 1983). If such functions seem specialized, consider the complexity of static vision, where we recognize colors, shapes, textures, and 3D form. It is no more remarkable that the visual system should have developed specialized processing for motion, which, through evolution, has shown itself to be rich in information.

In some types of motion processing, the quantity of interest is not really motion at all. For example, as one moves through the environment, the visual motion system analyzes the expanding image on the retina to compute the direction of self-motion, which is (or can be) a static thing. Also, motion cues give rich information about where object boundaries are, and about the relative distance of objects in a scene. Humans can even use expanding image cues to predict their precise moment of collision with an object; this remarkable computation is equivalent to taking the spatial derivative of a temporal derivative (the expansion rate) and does not require knowledge of target distance or approach speed! There are other important roles for visual motion; the point is that visual motion is a valuable source of information about the environment, and about our relationship to the environment, and our visual system has clearly evolved to exploit it.

3. Anatomy of Motion Processing

Image contrasts are captured at high resolution in the retina and sent back, via a relay in the lateral geniculate nucleus, to the primary visual cortex (V1). In primates, V1 neurons are the first in the visual hierarchy to show directionality: a heightened response to objects moving in a particular direction. These neurons are mainly packed into layer 4B, which makes a direct projection to the next cortical stage, the middle temporal area (MT). MT is now understood to be a specialized motion-processing area. Ascending projections out of layer 4 are atypical, usually coming instead from layers 2/3 and 6, and this suggests the importance of a fast, specialized motion pathway. MT itself projects to the medial superior temporal area (MST), another motion processing area whose function will be discussed in Sect. 6.

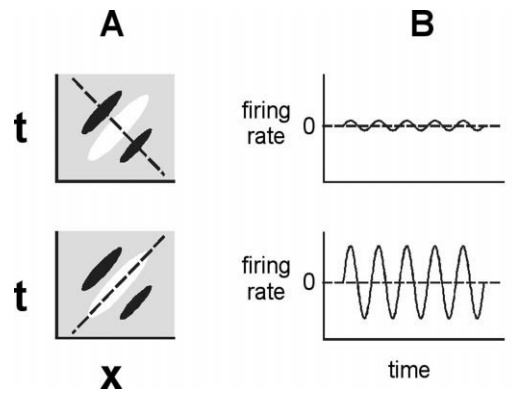

Figure 1
Space-time filter tuned for rightward motion

The motion pathway described above represents only the core of what has so far been discovered. Other areas, such as V2, V3, and VIP, clearly have roles in motion processing, but they are less understood. The goal of the present review is explain basic principles of motion perception; thus, it will suffice to concentrate on the basic pathway described.

4. Theory of Motion Detection

Reichardt proposed a model of motion detection that supposes a pair of spatially offset contrast detectors as the input (Reichardt 1961). When these are activated, they send their outputs to a multiplier, but not at the same speed. One of the connections to the multiplier is presumed to have some kind of delay. Therefore, if this slow line is triggered first, and the quicker line is triggered second, the two signals may arrive simultaneously at the multiplier, creating a very large, multiplicative response. There is no other way to get this strong response; one simply has to trigger the detectors with the appropriate timing. Therefore, the Reichardt model is a velocity detector. (Note: we use the term velocity in the vector sense here, indicating both direction and speed).

Later theoretical work by several groups led to a different conceptualization, which we will call the spatiotemporal energy model (Adelson and Bergen 1985). It is based on the idea of a linear space-time filter. In the same way that a neuron's spatial receptive field can be thought of as a weighting function (a region of varying sensitivity) over space, we can define

a space-time field as a weighting function over space and time. Referring to Fig. 1, the neuron that owns this space-time receptive field will be excited when the white parts of the field are contacted and suppressed when the dark parts are contacted. If we place a bar, representing stimulus contrast, with the appropriate angle, it can be made to cover the white parts only; i.e., it is a purely excitatory stimulus (Fig. 1A). Now consider that any oblique orientation in space-time represents motion: a finite change in position over a finite change in time. Thus, the obliquely oriented lobes of the space-time receptive fields can be thought of as velocity detectors.

It is now generally believed that in primates, a subset of V1 neurons act as velocity detectors. Various experiments have demonstrated some degree of linearity in these detectors (Reid et al. 1987), which argues against the strict Reichardt model since its directional output occurs at a nonlinear (multiplicative) stage. But the Reichardt model and the spatiotemporal energy model are not really so different; in fact with minor modifications to the Reichardt model, the models produce equivalent output (Adelson and Bergen 1985, van Santen and Sperling 1985). The differences are in the order of computational steps.

Note that even the spatiotemporal energy model is only linear in its earliest stages; at some point a nonlinearity must occur. A sine-wave grating moving in the optimal (preferred) direction across a linear space-time filter generates large outputs (Fig. 1B), but these oscillate around zero as the positive and negative phases of the wave move through the filter. The total output—the integrated area—is zero. Some kind of nonlinear operation is needed to prevent this; squaring, for example, works well (thus the 'energy' nomenclature). Physiologically, the nonlinearity is probably a combination of response rectification, threshold and saturation, gain normalization, and perhaps other factors.

5. *Integration of Local Velocities*

One of the most important revelations in the study of motion perception has been that motion processing goes well beyond simple detection (though detection is clearly not simple at all). Theorists have known this for many years; anyone trying to create a computerized motion processor realizes that local motion cues have to be integrated to reduce noise and to solve the aperture problem (see below), and this integration creates new problems because separate objects risk being mixed together. Independent visual cues such as disparity (Bradley et al. 1995) or color (Croner and Albright 1999) may be used to sort motion signals from different objects, but then there is the issue of how these cues are incorporated; and so on. Thus,

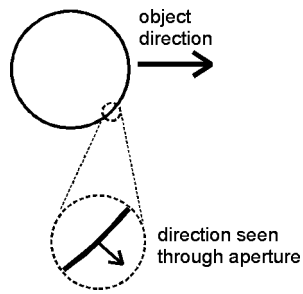

Figure 2
Aperture problem

before discussing the fancier aspects of motion processing, such as kinetic depth and heading computation, we must first examine the integration steps.

Whether through direct projection or via V2 (or other areas), much of the output of V1 directional cells ends up in area MT. Whereas about one quarter of the neurons in V1 are direction-selective, the great majority of cells in MT are direction-selective; thus, MT as an area is clearly preoccupied with visual motion. Still, V1 is much bigger—about 20 times bigger—so there are more directional neurons overall in V1. What, then, is the function of MT?

It is, at least in part, to combine local velocity samples intelligently. V1 neurons have tiny receptive fields, so when a moving edge appears, it always looks like it is moving perpendicularly to the edge (Fig. 2), even when it is not. This dilemma, called the aperture problem, could be resolved with large receptive fields, but then motion detection could only be done at coarse resolution. The solution is to combine the outputs of many directional V1 cells. Individually their outputs are meaningless; there is no probabilistic relationship between edge orientation and object direction (although, in the absence of noise, a single direction sample does narrow the range of possible object directions to 180°). Collectively, however, the local velocity samples (the V1 outputs) can give accurate information about the overall direction of movement.

The manner in which MT neurons integrate input has been the subject of some debate. One possibility is simply to average the local velocities; in many cases, at least, this gives a good approximation of the object's direction. But to the extent that visible edge orientations are not evenly distributed about the object direction, the vector average tends to be wrong. Another possibility is to find an explicit solution, which is in theory given by the direction and speed of any two vector samples. The geometric form of this calculation is called the intersection of constraints (IOC). Or, one might approximate the IOC by emphasizing higher speeds; since the apparent speed of local velocity samples falls off to the extent that their apparent direction differs from the object direction,

this would have the effect of favoring motion signals that really indicate the object's direction. Finally, one could isolate features, such as line intersections or texture changes, and track them. Features always appear to move in the object direction so they are a trustworthy source of motion information. Note that the vector average differs from the other computations in that it is linear.

Movshon and colleagues constructed a stimulus that can logically be seen as either a single, moving plaid, or as two sine gratings sliding over each other. They discovered that some MT neurons were capable of finding the overall plaid direction, which requires integrating information from the individual gratings. They called these 'pattern cells' (Movshon et al. 1985). From those studies, we cannot tell how the integration is done, only that it is done. But recently Pack and Born created stimuli that appear to be moving in the wrong direction if one simply takes their vector average. Responding to these stimuli, MT neurons computed the correct direction, or at least came close to it, which indicates one of the three nonlinear computations mentioned above. Remarkably, this semi-exact solution required some 140 msec to compute, and in the meantime, the neurons generated a vector average result. Thus, MT neurons compute a fast, approximate solution followed by a slower, more accurate solution.

Any operation by MT that combines multiple inputs risks crossing object boundaries. Since it is generally useless to know the 'overall' direction of independent objects, it is critical that MT confine its integrative operations to a particular, coherently moving object. Several lines of evidence suggest that MT neurons constrain their integration in this way. First, MT neurons have antagonistic regions surrounding their (classical) receptive fields that tend to suppress responsiveness when stimuli are uniform in their direction, speed, and/or disparity (depth) (Allman et al. 1985, Bradley and Andersen 1998). That is, neurons tend to be more active when their receptive fields detect discontinuities, and thus probable object boundaries. Second, the mutual cancellation of preferred- and non-preferred direction signals in MT (suggesting that they are being integrated) decreases or disappears when the movements occur at different depths (Bradley et al. 1995). This makes sense in that cues from different depths are probably from different objects. Finally, the ability of MT neurons to extract coherent motion signal from noise is facilitated by color cues that separate the noise from the genuine signal (Croner and Albright 1999).

6. *Higher Motion Processing*

One of the more useful quantities one can compute from visual motion is depth, or relative distance from the observer. This is called kinetic depth, or structure-from-motion. Overall, it is based on lawful relationships between retinal motion and 3D image structure, presumably discovered by the brain sometime during its evolution. These relationships, for now, are understood to be of three basic types, though their neural mechanisms may not be distinct. In one type, the presence of different directions is sufficient to evoke the perception of a depth slice between them, particularly if the signals are spatially overlapped (like screen doors sliding over each other). This effect probably occurs in MT, possibly through competitive circuitry between neurons tuned for different directions (Bradley et al. 1998). The depth-slice effect is logical when one considers that if two objects move through the same part of a 2D image, one is probably in front of the other; otherwise they would collide.

Another form of kinetic depth involves large speed gradients. Looking out the window of a moving car, for example, invokes a strong sense of near and far, and this is largely due to the speed difference between the immediate foreground, which rushes by, and the distant background, which appears to move slowly. Orban and colleagues found MT neurons that respond selectively to this type of stimulus (Xiao et al. 1997). Interestingly, the neurons specifically sense the speed gradient that spans their receptive field center and surround, an important result because it is among the most sophisticated functions yet found for a surround. Finally, localized speed gradients give a sense of curvature; this fact is exploited in CAD programs that rotate an object on the screen to give a better sense of its 3D structure (Wallach and O'Connell 1953). It is not known where these localized speed gradients are computed.

Another remarkable output of the visual motion system is the direction of self-motion, or heading. Gibson realized that when an observer moves through the environment, the retinal image expands, and the focus (origin) of this expansion corresponds to the direction of heading (1950). His suggestion that humans sense heading in this way was not initially accepted, however, because of the problem that when the eyes move—as in tracking any object not on the motion path—the expansion focus shifts in the direction of eye movement, and thus no longer corresponds to the heading. But Warren, Banks, and their colleagues later showed that subjects do in fact rely on the expansion focus to know their direction of heading, even during eye movements. Remarkably, they use an internal signal linked to the eye movement to approximate where the focus would have been without the eye movement, thus recovering the heading.

Following work by Saito et al, Duffy and Wurtz, and others, it became clear that heading computation might be carried out in the dorsal part of MST, called MSTd (Duffy and Wurtz 1991, Saito et al. 1986). The primary reasons were that: (a) MSTd neurons have large receptive fields, often covering most of the visual

image; (b) many MSTd neurons are particularly responsive to expanding images; and (c) expansion-selective MSTd neurons are sensitive to the position of an expansion focus. The role of MSTd in heading perception was confirmed when MSTd neurons were shown to adjust their focus tuning in a way that compensates for eye movements (Bradley et al. 1996), paralleling the psychophysical results, and when it was discovered that microstimulation of MSTd neurons produces biases in heading percepts (Britten and Wezel 1998). The combination of theoretical, psychophysical, and neurophysiological efforts that led to our current understanding of heading perception is a good example of how neuroscience benefits from the convergence of ideas.

Other kinds of information are extracted from visual motion cues in the primate brain. For example, motion cues suggest object boundaries; they provide an estimate of time-to-collision; and they provide retinal slip information needed to track moving objects. Unfortunately, these cannot all be discussed thoroughly here.

7. *Neural Correlates of Motion Perception*

Studies by Newsome and colleagues have clearly demonstrated the role of MT and MST in motion perception (Celebrini and Newsome 1994, Newsome et al. 1989). These researchers varied the signal-to-noise ratio of an image moving right or left and required monkeys to report the overall direction. They found a clear correlation between the direction reported by the monkey and the activity of direction-selective cells in MT and MST. But for some stimuli, the S/N ratio was actually zero; i.e., they contained no information about direction. The monkeys were required to answer anyway. Remarkably, in such cases, neural activity was still correlated with perception.

One possibility is that stochastic changes in MT/MST activity resulted in slight directional imbalances; for example, on a given trial, right-tuned cells might have been accidentally slightly more active than left-tuned cells, leading to a net sense of rightward motion. The other possibility is that monkeys, through cognitive mechanisms we can only speculate about, imagined right or left motion, and this in turn influenced MT/MST activities. In either case, the inseparability of direction judgment and MT/MST activity leaves little doubt about the crucial role of these areas in motion perception.

Reinforcing these findings, microstimulation studies have shown that direction judgments by monkeys can be reproducibly and predictably biased by stimulating MT or MST neurons with a particular direction preference (for example, stimulating cells selective for rightward motion tends to make monkeys see rightward motion) (Celebrini and Newsome 1995, Salzman et al. 1992). And lesion studies suggest that MT, at least, is crucial for normal motion perception. Monkeys with MT ablation, for example, have difficulty tracking moving targets Dursteler et al. 1987), and motion perception is profoundly disturbed in humans with bilateral damage to the MT/MST region Zihl et al. 1983).

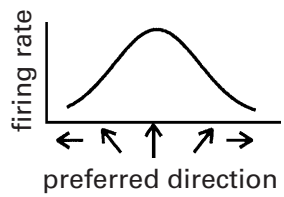

Figure 3
Hypothetical MT population response to upward motion

None of the above proves that motion perception comes from MT or MST. In fact, there is no reason to assume that perception occurs in any specific place. But the evidence presented above does suggest that in the processing that occurs between retinal motion and behavioral output, MT and MST have crucial roles.

8. *Future Questions*

Research on visual motion perception has advanced rapidly in the last few decades, from basic detection studies in the 1960s (Hubel and Wiesel 1968) to recent results concerning heading perception (Bradley et al. 1996), decision centers (Gold and Shadlen 2000), solutions to the aperture problem (Pack and Born 2001), and attentional gating of direction signals (Treue et al. 1999). As exciting as these developments have been, it is important not to leave the most basic questions behind, as many of these have yet to be answered. For instance, what kind of neural circuitry accounts for MT and MST response properties? How is speed measured? And what are the roles of directional neurons in areas outside MT and MST? Some of the most important results will probably derive from such basic studies.

It also seems that exciting things are about to happen in the study of population coding. MT and MST clearly play important roles in motion processing, which is to say that they compute something useful. This implies that the output of these areas—whatever it is they derive from their input—takes some recognizable form. To suggest otherwise would be to imply that other parts of the brain that use this information, such as the smooth-pursuit system and the decision centers that lead to direction judgments, understand a magical code that is not knowable to us. Rejecting this, the logical question is: what is the code? That is, in what format is the result of the neural computations expressed?

In the case of unidirectional percepts, that information looks something like a bell-shaped curve that expresses neurons' responses as a function of their preferred direction (Fig. 3). The idea is simply that neurons whose preferred direction matches the stimulus direction should have the strongest response. Whether this is interpreted internally in terms of the distribution mean or peak remains to be determined. But more basic questions apply to multidirectional percepts. When two movements occur near each other in an image, how does the brain know this? Because of noise and the aperture problem, the presence of different direction signals in the same part of an image never comes as a surprise and certainly is not a reliable indicator that independent movements have occurred.

Could we peer into MT, MST, or other areas and know whether an animal perceives one direction or two? Currently, no, because we do not know what to look for; we do not know the code. But recent studies suggest that we are on the way. Treue et al. have evidence that the actual shape of the MT response distribution may be decomposed by the visual system in deciding how many moving objects are present (2000). Pursuing a different hypothesis, Castelo-Branco et al. discovered synchronized oscillations in MT cortex of anesthetized cats in response to moving patterns that normally look coherent, but not with stimuli that look like two separate movements (2000). Recent preliminary results from awake monkeys, in contrast, showed no such oscillations (Thiele 2000). As channel capacity for multi-neuron recording continues to increase, the future holds great promise for unraveling some of the basic principles of neural coding. Much of this insight will no doubt come from studies of visual motion perception.

See also: Motion Perception Models; Neural Representations of Direction (Head Direction Cells); Neural Representations of Intended Movement in Motor Cortex; Perception and Action; Vision for Action: Neural Mechanisms

Bibliography

Adelson E H, Bergen J R 1985 Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A* **2**: 284–99

Allman J, Miezin F, McGuinness E 1985 Direction-specific and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). *Perception* **14**: 105–26

Bradley D C, Andersen R A 1998 Center-surround antagonism based on disparity in primate area MT. *The Journal of Neuroscience* **18**: 7552–65

Bradley D C, Chang G C, Andersen R A 1998 Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature* **392**: 714–7

Bradley D C, Maxwell M, Andersen R A, Banks M S, Shenoy K V 1996 Mechanisms of heading perception in primate visual cortex *Science* **273**: 1544–7

Bradley D C, Qian N, Andersen R A 1995 Integration of motion and stereopsis in middle temporal cortical area of macaques. *Nature* **373**: 609–11

Britten K H, van Wezel R J A 1998 Electrical microstimulation of cortical area MST biases heading perception in monkeys. *Nature Neuroscience* **1**: 59–63

Castelo-Branco M, Goebel R, Neuenschwander S, Singer W 2000 Neural synchrony correlates with surface segregation rules. *Nature* **405**: 685–9

Celebrini S, Newsome W T 1994 Neuronal and psychophysical sensitivity to motion signals in extrastriate area MST of the macaque monkey. *The Journal of Neuroscience* **14**: 4109–24

Celebrini S, Newsome W T 1995 Microstimulation of extrastriate area MST influences performance on a direction discrimination task. *Journal of Neurophysiology Online* **73**: 437–48

Croner L J, Albright T D 1999 Segmentation by color influences responses of motion-sensitive neurons in the cortical middle temporal visual area. *The Journal of Neuroscience* **19**: 3935–51

Duffy C J, Wurtz R H 1991 Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli. *Journal of Neurophysiology Online* **65**: 1329–45

Dursteler M R, Wurtz R H, Newsome W T 1987 Directional pursuit deficits following lesions of the foveal representation within the superior temporal sulcus of the macaque monkey. *Journal of Neurophysiology Online* **57**: 1262–87

Gibson J J 1950 *The Perception of the Visual World*. Houghton-Mifflin, Boston

Gold J I, Shadlen M N 2000 Representation of a perceptual decision in developing oculomotor commands. *Nature* **404**: 390–4

Hubel D H, Wiesel T N 1968 Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology* **195**: 215–43

Movshon J A, Adelson E H, Gizzi M, Newsome W T 1985 The analysis of moving visual patterns. In: Chagas C, Gattass R, Gross C (eds.) *Study Group on Pattern Recognition Mechanisms*. Pontifica Academia Scientiarum, Vatican City, Rome, pp. 117–51

Newsome W T, Britten K H, Movshon J A 1989 Neuronal correlates of a perceptual decision. *Nature* **341**: 52–4

Pack C C, Born R T 2001 Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature* **409**: 1040–2

Reichardt W 1961 Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In: Rosenblith W A (ed.) *Sensory Communication*. Wiley, New York

Reid R C, Soodak R E, Shapley R M 1987 Linear mechanisms of directional selectivity in simple cells of cat striate cortex. *Proceedings of the National Academy of Sciences* **84**: 8740–4

Saito H, Yukie M, Tanaka K, Hikosaka K, Fukada Y, Iwai E 1986 Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *The Journal of Neuroscience* **6**: 145–57

Salzman C D, Murasugi C M, Britten K H, Newsome W T 1992 Microstimulation in visual area MT: Effects on direction discrimination performance. *The Journal of Neuroscience* **12**: 2331–55

Thiele A, Stoner G 2000 Neural synchrony in macaque area MT does not correlate with plaid pattern coherence. *Society of Neuroscience Abstracts* **30**: 777–8

Treue S, Hol K, Rauber H J 2000 Seeing multiple directions of motion-physiology and psychophysics. *Nature Neuroscience* **3**: 270–6
Treue S, Martinez Trujillo J C 1999 Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**: 575–9
van Santen J P, Sperling G 1985 Elaborated Reichardt detectors. *Journal of the Optical Society of America A* **2**: 300–21
Wallach H, O'Connell D N 1953 The kinetic depth effect. *The Quarterly Journal of Experimental Psychology* **45**: 205–17
Wertheimer M 1912 Experimentelle Studien über das Sehen von Beuegung. *Zeitschrift für Psychologie* **61**: 161–265
Xiao D K, Marcar V L, Raiguel S E, Orban G A 1997 Selectivity of macaque MT/V5 neurons for surface orientation in depth specified by motion. *European Journal of Neuroscience* **9**: 956–64
Zihl J, von Cramon D, Mai N 1983 Selective disturbance of movement vision after bilateral brain damage. *Brain* **106**: 313–40

D. C. Bradley

Motivation and Actions, Psychology of

Three sets of phenomena have traditionally been of concern for research on motivation and action: (a) the selection of a certain course of action, (b) its energization, and (c) its regulation. Taking this comprehensive perspective, many different kinds of behavior (e.g., helping others, aggression, intergroup relations, achievement) can be analyzed from a motivational viewpoint. In the following sections, selected concepts are discussed that characterize present-day research on motivation: (a) motives and needs, (b) expectations, attributions, and control beliefs, and (c) goal setting and goal striving.

1. *Motives and Needs*

McClelland (1985) distinguished three basic groups of motives: the achievement motive, the power motive, and the affiliative motives. As food is the reward or incentive for hunger, so is improving one's performance on a given task the incentive for the achievement motive. The incentive of the power motive is having impact, control, or influence over another person, a group, or the world at large. Finally, the incentives for the affiliative motives extend to sexual pleasures (sexual motive), being together with people (need for affiliation), and experiencing harmony, concern, and commitment (intimacy motive). All of these motives may entail a fear or avoidance component. Trying to meet a standard of excellence may not be motivated solely by hope for success, but also by fear of failure, and spending one's spare time affiliating with others may not be determined solely by the anticipated positive feelings of togetherness, but also by strong fear of rejection.

In principle, all humans share these various motives, although with different strengths. Motive strength can be assessed by exploring both the array of situations a person interprets in terms of a given motive (e.g., a person high in need for power interprets all kinds of situations as power-related) and the intensity of the anticipated affect associated with having acquired respective incentives. Commonly this is done with the Thematic Apperception Test (TAT) which contains pictures of scenes loosely related to the motive measured. Recent research has linked the activation of different motives to different hormonal responses that in turn facilitate motive-specific behaviors (McClelland 1995).

Scoring high on a certain motive implies a recurrent concern for acquiring the respective incentives. For instance, people high on the affiliation motive perform affiliative acts frequently and energetically, readily perceive affiliative cues in the environment, and quickly detect affiliative networks. Also, predictions of the professional success of managers are strikingly accurate, particularly if one considers the motive dispositions in achievement (high), power (high), and affiliation (low) in concert. However, attempts to predict behaviors from motives commonly fail when engagement in these behaviors is based on conscious reflection. When it comes to choosing between courses of action, tasks of different difficulty, or persisting on a given task versus leaving the field, people deliberate on the feasibility and desirability of the alternative courses of action.

2. *Expectations, Attributions, and Control Beliefs*

One of the first attempts to integrate cognitive aspects of motivation was made by Atkinson (1957) in his risk-taking model. He proposed that the subjective probability of success and the task's incentive value conjointly affect task choice, both variables being influenced by the perceived difficulty of the task. Whereas easy tasks lead to a high subjective probability of success (direct function), they also possess low incentive value (inverse function) because the anticipated affect associated with success (pride) is lowest for easy tasks. The reverse is assumed for difficult tasks. Atkinson suggested that multiplying probability of success and incentive value will give a good estimate of whether a person will choose to work on a task, especially when the obtained score is weighted by the approach (hope for success) and avoidance (fear of failure) components of the person's achievement motive. Research supports the model for predictions on task choice, but the model fails to account for the quantity and quality of task performance once people have started to work on the

chosen tasks (see *Group Decision Making, Social Psychology of*).

Elaborations of the model (Heckhausen 1991) added further expectation-related concepts (e.g., the expectation that successful task performance will lead to the anticipated incentives) and differentiated various incentives (e.g., extrinsic side effects, such as when an achievement task has affiliative benefits). Atkinson's model has also been elaborated by attribution theorists (Weiner 1992) who attempted to understand changes in expectations and incentive value in terms of the attributions made for past performances. Moreover, Weiner discovered that the approach component of the achievement motive (hope for success) is associated with attributing failure to luck or lack of effort and success to ability, whereas the avoidance component is linked to attributing failure to lack of ability and success to luck (see *Academic Achievement Motivation, Development of*).

Recognition of the motivational importance of expectations and attributions was the starting point of the cognitive revolution in the psychology of motivation which also has introduced the concept of control beliefs. Self-efficacious individuals hold the firm belief that they possess the potential to execute the kinds of behaviors that a given task demands (Bandura 1997). People acquire this belief by reflecting on their own relevant past behaviors, observing the behaviors of similar others, being evaluated by significant others (e.g., teachers), and observing their own physiological reactions when challenged by a given task. High self-efficacy beliefs are associated with choosing aspiring goals, exerting strong efforts to attain these goals, and persisting in the face of obstacles and hindrances (see *Control Behavior: Psychological Perspectives*).

3. Goal Setting

The most recent advance in the psychology of motivation and action is research on goal pursuit. Research on goals addresses the question of how well people translate their desires and beliefs into action.

3.1 Determinants of Goal Setting

Research on the determinants of goal setting discovered that people differ in their preference for setting goals with certain structural features or contents. For example, people who generally think about their actions in concrete versus abstract terms also prefer to set themselves concrete versus abstract goals, respectively. People who construe their self as an ideal (which they intrinsically desire to attain) set promotion goals (i.e., goals focusing on accomplishments and aspirations), whereas people who construe their self as an ought which they feel compelled to reach set prevention goals (i.e., goals that focus on safety and responsibilities, Higgins 1997). It also matters what kind of implicit theories people hold on the nature of personal attributes. If people believe that ability is fixed and cannot easily be changed, they choose performance goals (i.e., goals that serve the purpose of finding out how capable one is). If, however, people believe that ability can be improved, they choose learning goals geared at finding out how one can successfully carry out the task at hand (Dweck 1999). People's needs, wishes, and higher order goals also influence the type of goals that are set (Ryan et al. 1996). Moreover, people's concept of what they could possibly become (i.e., the possible self) provides thematic conceptions of what future selves they may strive for.

3.2 Processes of Goal Setting

Recent research (Oettingen 2000) demonstrates that the perceived feasibility of goal attainment does not always determine people's goal setting. When people fantasize about a desired future, they set themselves goals independent of perceived feasibility. Thus people who indulge in fantasies about a desired future commit themselves to goals irrationally: they are too committed when probabilities of success are low, and not committed enough when probabilities of success are high. Such irrational goal commitments are also observed with people who are caught up in ruminations about aspects of the present reality that stand in the way of reaching one's fantasies. Only when people mentally contrast their positive fantasies with present reality does their goal setting reflect perceived feasibility. Strong goal commitments emerge when perceived feasibility is high, and no goal commitment is found when perceived feasibility is low.

Goals may also become activated outside of awareness (Bargh and Chartrand 1999). Strong mental links develop between the cognitive representations of situations and the goals that people chronically pursue within these situations. As a consequence of this repeated and consistent pairing in the past, such goals become automatically activated when the person enters the relevant situation. The automatically activated goal then guides behavior, without the individual choosing or intending the respective goal-directed line of action (see *Automaticity of Action, Psychology of*).

4. Goal Striving

4.1 Goal Content Effects

Successful goal striving is determined by how goals are framed and what contents they specify. The following structural features of goals are important: challenging goals that are spelled out in specific terms lead to a higher attainment rate than modest specific goals or challenging but vague ('Do your best!') goals. Proximal goals that relate to what the individual does in the near

present or will do in the future are superior to distal goals that point far into the future. *Promotion goals* geared at accomplishment facilitate goal pursuit, whereas *prevention goals* geared at acquiring security hamper goal pursuit. Learning goals lead to better performances than performance goals, as the former allow for a more effective coping with failure than the latter. Performance goals are less detrimental, howver, when they are framed as approach goals (e.g., I want to get good grades) as compared to avoidance goals (e.g., I do not want to get bad grades).

Moreover, the thematic content of goals matters. Goals covering issues of autonomy, competence, and social integration are said to further intrinsic goal pursuit which in turn promotes creativity, higher cognitive flexibility, and greater depth of information processing. Side-effects of intrinsic goal pursuits are an increased subjective well-being and life satisfaction (see *Intrinsic Motivation, Psychology of*). Effects of goals on subjective well-being are also influenced by how well people's goal contents match the strengths of their motives of achievement, affiliation, power, and intimacy (Brunstein et al. 1998).

Other parameters of goal content and goal structure have also been found to relate to subjective well-being. Strong goal commitment, for instance, favors subjective well-being, but only when the probability of success is perceived as high. The strongest predictor of positive well-being is the proportion of intimacy goals in the total of personal goals a person holds, whereas the proportion of achievement and power goals tends to be related to negative well-being. However, the strength of these effects seems to be contingent on positive and negative life events in domains relevant to these goals. Moreover, the level of goal specification also affects subjective well-being. High-level goals (e.g., bring happiness to those around me) tend to be associated with psychological distress (anxiety, depression). Low-level goals (e.g., get along with my brother) on the other hand, have been linked to greater levels of psychological well-being, but also to more physical illness. A high proportion of avoidance goals (e.g., not being late) as compared to approach goals (e.g., being on time) impedes psychological and physical well-being. Finally, a lack of integration of the many goals people hold in terms of experiencing much goal conflict or a strong fragmentation of the self has also been linked to low subjective well-being (Emmons 1996).

4.2 Planning

Having set a goal is just a first step toward goal attainment, commonly followed by a host of implemental problems that need to be solved. Research on implemental mind-sets (Gollwitzer 1990) has shown that planning the implementation of a set goal creates a cognitive orientation that facilitates getting started with goal-directed actions. Implemental mind-sets prevent distraction by irrelevant information and promote processing of information related to the implementation of set goals. Moreover, desirability-related information is processed partially, favoring pros over cons, and the analysis of feasibility-related information is optimistic.

Set goals commit people to attaining the specified future (outcome or behavior), but they do not commit people to when, where, and how they want to attain it. Planning one's goal pursuit via forming implementation intentions that take the form of 'If I encounter situation x, I will perform the goal-directed behavior y' (Gollwitzer 1999) promote the attainment of difficult to reach goals (e.g., healthy eating). As implementation intentions spell out links between situational cues and goal-directed behavior, the control of goal-directed behavior is delegated to environmental cues (e.g., good opportunities). The situational cues specified in implementation intentions are more easily detected, remembered, and more readily attended to than comparable nonintended situations. The goal-directed behavior specified in implementation intentions is initiated immediately and effortlessly in the presence of the critical situational cues, without necessitating a conscious intent. The task of planning can also be approached in a more reflective way, however, as is entailed in mental simulations that explore possible routes to achieving one's goal (process simulations, Taylor et al. 1998). If such process simulations are applied repeatedly, they further the attainment of set goals.

4.3 Action Control Strategies

Successful goal attainment implies that a currently pursued goal has to be shielded from competing goals (e.g., the goal of making a phone call from the competing goal of tidying up one's messy desk). Various control strategies can be differentiated (Kuhl and Beckmann 1994), such as attention control or emotion control. Whether and how effectively these strategies are used depends on the control mode of the individual. An action-oriented person concentrates on the planning and initiation of goal-directed action; responds flexibly to contextual demands; and uses control strategies effectively. A state-oriented person, however, cannot disengage from incomplete goals and is thus caught up in persevering thoughts related to aversive experiences or future successes. Also, state-oriented individuals readily misperceive assigned goals as self-generated, and the degree of such false self-ascriptions is closely associated with reduced enactment of self-chosen as compared to assigned goals.

Successfully resolving goal conflicts is not only an issue of shielding an ongoing goal pursuit from competing goal pursuits (Cantor and Fleeson 1994). There is also the possibility of creative integrations, where new goals are formed which serve both of the conflicting goals (e.g., affiliation and achievement

goals can be reconciled by taking on civic responsibility). Moreover, in an attempt to meet higher order goals (e.g., graduating from high school) people can strategically link behavioral goals that on the surface appear in conflict (e.g., when affiliating with people and studying are reconciled by studying in groups).

4.4 Mobilization of Effort

People can secure goal attainment not only by planning and shielding off distractions, but also by increasing effort. A person's readiness to exert effort turns out to be directly determined by the perceived difficulty of the task at hand (Wright 1996). As the perceived difficulty increases so does the person's effort expenditure, unless the task is recognized as unsolvable. But there is a second limit to the linear increase of effort expenditure in response to heightened task difficulty: a person's potential motivation.

Potential motivation is determined by need-related variables (i.e., strength of the related need or higher order goal, the incentive value of the task, and the instrumentality of task completion for need satisfaction or attainment of the higher order goal). If the level of potential motivation is low, people do not find it worthwhile to increase effort when an easy task becomes more difficult. This is because the upper limit of effort expenditure (suggested by the potential motivation) is low and thus reached quickly.

4.5 Discrepancy Reduction

Even when failure occurs people do not give up on their goal pursuits. Rather, they experience a discrepancy that needs to be closed. According to Bandura (1997), goals only specify the conditions that allow for a positive or negative self-evaluation. If the set goal is attained through one's actions, a positive self-evaluation prevails; whereas staying below one's goals leads to a negative self-evaluation. The individual thus is seen as pushed by the negative self-evaluation associated with the discrepancy, and pulled by the anticipated positive self-evaluation that is intrinsically linked to closing the gap between the status quo and the goal (i.e., the performance standard). This implies that goals stimulate effortful action toward goal attainment only when people recognize a discrepancy between the status quo and the set goal. Bandura therefore proposes giving frequent feedback as a powerful means of stimulating goal pursuit.

Carver and Scheier (1998) propose a different discrepancy reduction theory of goal pursuit. Based on cybernetic control theory, the central conceptual unit of their analysis is the negative feedback loop. Carver and Scheier highlight the hierarchical organization of goal pursuit and thus assume a cascading loop structure. Goal-directed behavior is usually regulated at the middle level ('Do-goals') with action at higher levels ('Be-goals') suspended until the individual becomes self-aware. When discrepancies on the 'Be-level' or the 'Do-level' are discovered, lower level goals or behaviors geared at discrepancy reduction are triggered. A positive affective response as a consequence of goal attainment is not assumed, nor is the detection of a discrepancy assumed to be associated with negative affect. Rather, the speed of progress in discrepancy reduction is seen as the source of positive or negative feelings.

Research on identity goals demonstrates, however, that people do not necessarily have to move downwards (i.e., to lower level goals) when trying to close goal discrepancies. When it comes to 'Be-goals' that specify a desired identity (such as being a good scientist) there are many different, alternative ways to indicate to oneself and others that one possesses the aspired identity. If one has failed to attain an indicator or has discovered that an indicator is out of reach (e.g., having productive students), one can compensate by striving for alternative indicators (e.g., presence at conferences).

5. Conclusion and Future Perspectives

Research on motivation and action traditionally focused on identifying the determinants of motivation. This search has moved from the affective determinants of motives, needs, and incentives to more cognitive determinants, such as expectations, attributions, and control beliefs. With the recent focus on goals (Gollwitzer and Bargh 1996), the volitional issue of the regulation of goal-directed behavior has become prevalent, and the human being is conceived as a flexible strategist. This perspective leads to a focus on the analysis of reflective and reflexive psychological processes that guide the successful setting and implementing of goals (see *Self-regulation in Adulthood*).

Even though research on goals has won momentum in recent years, there is a host of issues that have not yet received much theoretical and empirical attention. One of these is the issue of goal conflict. Future research will have to discover how goal conflicts emerge and how they affect thoughts, feelings, and actions. Moreover, different ways of resolving goal conflicts (e.g., creative integrations versus disengagement from one of the conflicting goals) need to be distinguished. It does not suffice to simply analyze the determinants and consequences of goal conflict resolution. Observing that disengagement from goals is triggered, for instance, by a lack of opportunities to pursue the goal, or observing that any disengagement from goals is accompanied by ruminative thought as well as frustrated and depressed affect (Klinger 1975), is an important first step. The next question is, How can people effectively self-regulate disengagement from goals. As any self-regulation is taxing in the sense that subsequent self-regulation becomes less effective, research on goals will need to discover self-regulatory

strategies of goal setting, and goal implementation that are easy to perform.

See also: Academic Achievement Motivation, Development of; Action Theory: Psychological; Activity Theory: Psychological; Attitudes and Behavior; Automaticity of Action, Psychology of; Control Behavior: Psychological Perspectives; Group Decision Making, Social Psychology of; Intrinsic Motivation, Psychology of; Self-regulation in Adulthood

Bibliography

Atkinson J W 1957 Motivational determinants of risk-taking behavior. *Psychological Review* **64**: 359–72

Bandura A 1997 *Self-efficacy: The Exercise of Control*. W. H. Freeman, New York

Bargh J A, Chartrand T L 1999 The unbearable automaticity of being. *American Psychologist* **54**: 462–79

Brunstein J C, Schultheiss O C, Grässmann R 1998 Personal goals and emotional well-being: The moderating role of motive dispositions. *Journal of Personality and Social Psychology* **75**: 494–508

Cantor N, Fleeson W 1994 Social intelligence and intelligent goal pursuit: A cognitive slice of motivation. In: Spaulding W (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, Vol. 41, pp. 125–79

Carver C S, Scheier M F 1998 *On the Self-regulation of Behaviour*. Cambridge University Press, Cambridge, UK

Dweck C S 1999 *Self-theories. Their Role in Motivation, Personality, and Development*. Psychology Press, Philadelphia, PA

Emmons R A 1996 Striving and feeling: Personal goals and subjective well-being. In: Gollwitzer P M, Bargh J A (eds.) *The Psychology of Action: Linking Cognition and Motivation to Behavior*. Guilford Press, New York, pp. 313–37

Gollwitzer P M 1990 Action phases and mind-sets. In: Higgins E T, Sorrentino E M (eds.) *Handbook of Motivation and Cognition: Foundations of Social Behavior*. Guilford Press, New York, Vol. 2, pp. 53–92

Gollwitzer P M 1999 Implementation intentions: Strong effects of simple plans. *American Psychologist* **54**: 493–503

Gollwitzer P M, Bargh J A 1996 *The Psychology of Action: Linking Cognition and Motivation to Action*. Guilford Press, New York

Heckhausen H 1991 *Motivation and Action*. Springer-Verlag, Berlin

Higgins E T 1997 Beyond pleasure and pain. *American Psychologist* **52**: 1280–1300

Klinger E 1975 Consequences of commitment to and disengagement from incentives. *Psychological Review* **82**: 1–25

Kuhl J, Beckmann J 1994 *Volition and Personality: Action versus State Orientation*. Hogrefe & Huber, Seattle, WA

McClelland D 1985 *Human Motivation*. Scott, Foresman, Glenview, IL

McClelland D 1995 Achievement motivation in relation to achievement-related recall, performance, and urine flow, a marker associated with release of vasopressin. *Motivation and Emotion* **19**: 59–76

Oettingen G 2000 Expectancy effects on behavior depend on self-regulatory thought. *Social Cognition* **18**: 101–29

Ryan R M, Sheldon K M, Kasser T M, Deci E L 1996 All goals are not created equal: An organismic perspective on the nature of goals and their regulation. In: Gollwitzer P M, Bargh J A (eds.) *The Psychology of Action: Linking Cognition and Motivation to Behavior*. Guilford Press, New York, pp. 7–26

Taylor S E, Pham L B, Rivkin I D, Armor D A 1998 Harnessing the imagination: Mental simulation, self-regulation, and coping. *American Psychologist* **53**: 429–39

Weiner B 1992 *Human Motivation*. Sage, Newbury Park, CA

Wright R 1996 Brehm's theory of motivation as a model of effort and cardiovascular response. In: Gollwitzer P M, Bargh J A (eds.) *The Psychology of Action: Linking Cognition and Motivation to Behavior*. Guilford Press, New York, pp. 424–53

P. M. Gollwitzer and G. Oettingen

Motivation: History of the Concept

Evolutionary theory, the study of learning, and the psychoanalytic study of mental illness have been committed to see motivation as a primary cause of behavior, and this is also true of behavioral biology and physiological psychology, as they prefer to think in terms of homeostasis (Cofer and Appley 1964). When hunger occurs, for instance, behavior is instigated (e.g., finding and eating food) that is instrumental to rectifying the imbalance caused by tissue needs and deficits. The named theoretical perspectives have barricaded the simple insight that behavior can occur under externally applied forces as well (e.g., a shove). Even more importantly, the physical structure of the organism, its sensory and perceptual capacities, its cognitive and motor abilities, and so forth, qualify as causal factors. Even habits, once formed, can be seen as such.

Research on motivation has recognized early on that motivation can only be one of the many causes of behavior, and this extends to the causation of affect and cognition as well. Accordingly, the leading question became: What are the aspects of cognition, affect, and behavior that benefit most from a motivational analysis? Traditionally, the following issues have been addressed: First, there is the question of the facilitation and energization of certain responses. This issue has been analyzed by the classic proponents of learning theory (Hull 1943, Spence 1956) who suggested that motivational processes as such (assumed to be rooted in general drive or arousal states) do not necessarily control or guide specific forms of behavior but may at times invigorate innate or learned associative tendencies. Second, thoughts, feelings, and actions are often characterized as guided, directed, goal-oriented, persistent, or purposive. Such qualities relate to making choices, but also to short-term or long-term efforts in implementing the choices made.

Researchers like Atkinson (1957) and McClelland (1955) focused on this issue and suggested a motivational analysis.

To demonstrate how the concept of motivation has changed in history, this article describes how two central issues (i.e., basic needs and action control) have been addressed differently by the traditional and the modern psychology of motivation.

1. Basic Human Needs

Whereas learning theorists, following the lead of Hull, conceived of need as a general, content-free drive, personality psychologists were eager to put content into the concept of need. Sigmund Freud was rather parsimonious by only suggesting two basic needs, namely the life and the death instinct. McDougall (1932) listed 18 basic needs which he referred to as instincts (e.g., curiosity, self-assertion, submission). Murray's (1938) *Explorations in Personality* contains a catalog of 20 psychogenic needs (e.g., need for nurturance, need for succorance) and Maslow's (1954) *Motivation and Personality* proposed a hierarchy of needs whereby the lower need categories related to deficiency needs (e.g., hunger, safety) and the higher ones to growth needs (e.g., need to achieve, need to realize one's potential).

The named approaches excelled in generating sophisticated definitions and descriptions, but did not provide reliable instruments for need assessment. They also failed to answer the question of which of the many needs are activated in a given situation and how such activation processes run off. Moreover, they did not explore the origins of individual differences in the strength of these needs. In later years, Atkinson (1957) and McClelland (1985) addressed these problems with respect to the need for achievement and the power motive.

Still, the question remains: Which of the many conceivable human needs is the most basic? Baumeister and Leary (1995) have recently raised this issue by applying criteria such as universality, nonderivativeness, satiation, and substitution. According to their analysis, the need to belong in the sense of forming and maintaining strong, stable interpersonal relationships is suggested to be a most basic need. However, this need seems to be in tension with another basic need, the desire to distinguish oneself and be different from others, as people are found to compete even with their intimates when it comes to performing well on dimensions that are of high personal relevance (Tesser 1988). Indeed, the need for high self-esteem has been suggested as the ultimate human motive, because it buffers fear of death which is a consequence of humans being conscious of their mortality. Such awareness engenders abject terror that needs to be managed as it could paralyze the ability to act (Pyszczynski et al. 1997).

Recent research on individual differences in terms of motivational orientations has replaced the need (motive) construct with constructs that describe a general goal orientation, such as personal projects, personal strivings, life tasks, or identity goals. Such personal strivings (Emmons 1996) are more limited in scope (e.g., be a honest person) and can be characterized in terms of expectancies of success, complexity, high vs. low level of abstraction, avoidance vs. approach orientation, degree of conflict between each other, and integration vs. fragmentation. The named parameters of personal strivings have been observed to relate to measures of psychological and physical well-being.

2. Action Control

Early theories portrayed the human as a machine-like reactive organism compelled to act by internal and/or external forces beyond our control (e.g., instincts, needs, drives, incentives, reinforcers, etc.). Prototypical theories are the psychoanalytic theory of Freud, Hull's learning theory (and the reformulations by his students), or Lewin's field theoretical approach. These theories imply that if one just pushed the right button, motivation would result. There is no room for conscious reflections and attempts towards self-regulation. Instead, motivational forces transmit their energy outside of awareness, establishing a state of balance or equilibrium (referred to as arousal reduction, self-preservation, or need satisfaction).

More modern theories construe human beings as all-just and all-knowing final judges of their actions. Expectancy-value theories (e.g., Atkinson 1957) assume that people choose goals in a rational way, based on the comprehensive knowledge of the probability of goal attainment and the goal's expected value. It is proposed that the subjective probability of success and the incentive value of having performed a task (i.e., pride or shame) conjointly affect task choice, both variables being influenced by the perceived difficulty of the task. Elaborations of this model (Heckhausen 1991) added further expectation-related concepts (e.g., action-outcome expectancies), and differentiated various aspects of the incentive value (e.g., extrinsic side effects). Attribution theories (e.g., Weiner 1992) propose that the motivational determinants of a person's behavior are causal explanations of prior action outcomes. People are seen as amateur scientists who systematically explore the causes of their behavior. The type of causes discovered are expected to affect a person's readiness to engage in these or related behaviors by influencing affect and expectations.

More recently, the motivational importance of control beliefs has been analyzed. According to Bandura's (1997) self-efficacy theory, self-efficacious individuals hold the firm belief that they possess the potential to execute (i.e., control) the kinds of behaviors that a given task demands. People acquire

such beliefs by reflecting on their own relevant past behaviors, observing the behavior of similar others, being evaluated by significant others (e.g., teachers), and observing their own physiological reactions when challenged by a given task. High self-efficacy beliefs are associated with choosing aspiring goals, exerting strong effort to attain these goals, and persisting in the face of obstacles and hindrances.

Present theories of motivation go beyond conceptualizing humans as all-just and all-knowing. Human beings are construed as flexible strategists. The focus is on the different tasks a person has to perform when transforming wishes into actions (Gollwitzer 1990). When choosing goals, people try to live up to the ideal of being an all-knowing and all-just person by processing all the available information in an impartial manner. However, when the implementation of an already set goal is at issue, people are determined, become partial, and the desirability and feasibility are seen in the most positive light.

Recent research on goals focuses on the determinants and processes of goal setting as well as goal implementation. With respect to goal setting, for instance, it has been discovered that people who construe their self as an ideal which they intrinsically desire to attain, set themselves promotion goals focusing on establishing and keeping positive outcomes, whereas people who construe their self as an ought which they feel compelled to reach, set themselves prevention goals focusing on avoiding and getting rid of negative outcomes (Higgins 1997). Moreover, people can regulate the process of goal setting in a more or less productive manner, by the way they think about the future outcomes they want to attain. When the desired future is mentally contrasted with negative aspects of impeding reality (e.g., effectively mastering a project is mentally contrasted with obstacles standing in its way), relevant expectations of successfully realizing one's fantasies become activated and used. Accordingly, people form goal commitments in a rational manner (i.e., form strong goal commitments when expectations of success are high, and leave the field when probabilities of success are low). When people only dream about positive future outcomes or solely ruminate about the negative reality, however, respective goal commitments are moderate and expectancy-independent (Oettingen 2000).

Regarding the determinants of successful goal implementation, how goals are framed makes an important difference. For instance, when achievement goals are framed as learning goals (i.e., goals geared at trying to learn more about how one can successfully carry out the task at hand) as compared to performance goals (i.e., goals geared at trying to find out how capable one is), failure experiences are coped with more effectively and thus people are more likely to reach their goals (Dweck 1999). It also matters how goal pursuit is regulated by the individual. For instance, people can protect an ongoing goal pursuit from distractions by making plans on how to deal with them, or they can plan the details (when, how, and where) of the initiation and execution of the goal-directed behavior ahead of time, so that any potentially disturbing self-states (e.g., being tired) can no longer interfere (Gollwitzer 1999). People can also step up their efforts when hindrances are encountered and turn to substitute goals if increased effort still fails to guarantee goal attainment. Other effective action control strategies relate to the regulation of one's emotions, the perceived attractiveness of the goal, and to creating an environment that offers good opportunities for making progress toward goal attainment (Kuhl and Beckmann 1994).

This recent revival of research on the self-regulation of goal pursuit (Mischel et al. 1996) is reminiscent not only of the mentalists' analysis of willing (James 1890), but also of German will psychology (Ach 1935, Lewin 1926) before the heyday of behaviorism. William James pointed out that any self-regulation either has to do with strengthening a weak tendency to perform a desired behavior (i.e., issues of the obstructed will) or with weakening a strong tendency to perform an unwanted behavior (i.e., issues of the explosive will). James' analysis of willing is based on the assumption that behavior can potentially be regulated by a person's resolutions (or intentions, subjective goals), even though in certain situations and at certain times this may be difficult.

Kurt Lewin's experimental work on the willful control of behavior also offers ideas on how such control may come about, and the same is true for the research of Narziss Ach. Lewin suggested that goals assign a valence to objects and events in people's social and nonsocial surroundings. In Lewin's example of a person who intends to mail a letter, a mail box entices the person to deposit the letter much as food entices a hungry person to eat. As needs can be satisfied by various types of behaviors which may all substitute for each other in reducing need tension (e.g., eating fruit, vegetables), many different goal-directed behaviors qualify for satisfying the quasi-need associated with a set goal. Lewin's tension state metaphor thus effectively accounts for the flexibility of goal striving.

Ach's approach to the analysis of willing was different. He assumed that the linking in one's mind of an anticipated situation to a concrete intended behavior creates what he called a 'determination' which in turn automatically triggers the intended action when the specified situation is encountered. The strength of the determination was not assumed to relate to the importance of the person's intention or goal, but rather to the degree of concreteness when specifying the situation and to the intensity of the act of willing. For Ach, concepts like need relate to the importance of goals and thus were assumed to be critical for choosing between goals, whereas the implementation of set goals was an issue of willing. The suggested distinction between motivational issues

of goal choice and volitional issues of goal implementation is reflected in recent research on goals that distinguishes between goal setting and goal implementation.

See also: Action Planning, Psychology of; Attributional Processes: Psychological; Behaviorism; Behaviorism, History of; Cognitive Psychology: History; Free Will and Action; Mental Representations, Psychology of; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction; Motivation, Neural Basis of; Personality Structure; Personality Theories; Self-regulation in Adulthood; Self-regulation in Childhood

Bibliography

Ach N 1935 *Analyse des Willens*. Urban und Schwarzenberg, Berlin

Atkinson J W 1957 Motivational determinants of risk-taking behavior. *Psychological Review* **64**: 359–72

Bandura A 1997 *Self-efficacy: The Exercise of Control*. W. H. Freeman, New York

Baumeister R F, Leary M R 1995 The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin* **117**: 497–529

Cofer C N, Appley M H 1964 *Motivation: Theory and Research*. Wiley, New York

Dweck C S 1999 *Self-theories. Their Role in Motivation, Personality, and Development*. Psychology Press, Philadelphia

Emmons R A 1996 Striving and feeling: Personal goals and subjective well-being. In: Gollwitzer P M, Bargh J A (eds.) *The Psychology of Action: Linking Cognition and Motivation to Behavior*. Guilford Press, New York

Gollwitzer P M 1986 Action phases and mind-sets. In: Higgins E T, Sorrentino E M (eds.) *Handbook of Motivation and Cognition: Foundations of Social Behavior*. Guilford Press, New York, Vol. 2

Gollwitzer P M 1999 Implementation intentions: Strong effects of simple plans. *American Psychologist* **54**: 493–503

Heckhausen H 1991 *Motivation and Action*. Springer-Verlag, Berlin

Higgins E T 1997 Beyond pleasure and pain. *American Psychologist* **52**: 1280–1300

Hull C L 1943 *Principles of Behavior*. Appleton-Century, New York

James W 1890 *The Principles of Psychology*. Holt, New York

Kuhl J, Beckmann J 1994 *Volition and Personality: Action Versus State Orientation*. Hogrefe and Huber, Seattle, WA

Lewin K 1926 Vorsatz, Wille und Bedürfnis. *Psychologische Forschung* **7**: 330–85

Maslow A H 1954 *Motivation and Personality*. Harper & Row, New York

McClelland D C (ed.) 1955 *Studies in Motivation*. Appleton-Century, New York

McClelland D C 1985 *Human Motivation*. Scott, Foresman, Glenview, IL

McDougall 1932 *The Energies of Men: A Study of the Fundamentals of Dynamic Psychology*. Methuen, London

Mischel W, Cantor N, Feldman S 1996 Principles of self-regulation: The nature of will power and self-control. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology: Handbook of Basic Principles*. Guilford Press, New York, pp. 329–60

Murray H A 1938 *Explorations in Personality*. Oxford University Press, New York

Oettingen G 2000 Expectancy effects on behavior depend on self-regulatory thought. *Social Cognition* **18**: 101–29

Pyszczynski T A, Greenberg J, Solomon S 1997 Why do we need what we need?: A terror management perspective on the roots of human social motivation. *Psychological Inquiry* **8**: 1–20

Spence K W 1956 *Behavior Theory and Conditioning*. Yale University Press, New Haven, CT

Tesser A 1988 Toward a self-evaluation maintenance model of social behavior. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 21, pp. 181–22

Weiner B 1992 *Human Motivation*. Sage, Newbury Park, CA

P. M. Gollwitzer and G. Oettingen

Motivation, Learning, and Instruction

In popular language, motivation is equated with goal-directed behavior and is, as such, easily understood. However, educational researchers who wanted to include this seemingly clear construct into their models of learning and instruction soon discovered that it is a blanket term which refers to a variety of interrelated self-perceptions and affects, including outcome expectations, self-efficacy, goal orientation, goal setting, perception of control, interest, self-concept of ability, goal intentions, goal striving, persistence, and effort expenditure. Each of these facets of the motivation construct has been intensively researched within its own conceptual niche, a situation which has resulted in a kaleidoscope of overlapping constructs and complementary measurement instruments. Quite clearly, the presence of too many motivation-related constructs has hindered the integration of the concept within models of learning and instruction. This article begins with a brief historical tour of the motivation theories that are relevant to the study of learning and instruction. Next, some recent developments in motivation research will be discussed. These developments have been instrumental in giving motivation the prominent place in pedagogical practice and instruction models that is long overdue.

1. A Brief History of Motivation

An examination of the history of motivation reveals that the significance of the motivation construct was acknowledged in the 1940s and 1950s, when interpretation problems plagued the study of goal achievement. Researchers argued that individuals have access to a

large amount of prior knowledge which facilitates the formation of expectancies about various goals. In the late 1950s Atkinson (1958) published his famous 'expectancy by value' model in which he described action tendencies as the outcomes of the interactions between goals, expectancies, and motives related to success and failure. The theory of need achievement is, in fact, one of the first comprehensive theories of human motivation. It was expanded and refined several times. Heckhausen (1977) introduced some extrinsic and content-related components into the model, including novelty and task specificity, and illustrated how these components may contribute to students' action tendencies in a learning context.

Another comprehensive theory of motivation, namely the theory of intrinsic motivation, developed in parallel to the theory of need achievement. It is important to note that in both these theoretical frameworks motivation refers to a process whereby goal-directed behavior is instigated and sustained. However, the assumptions made about the reasons why students make conscious and deliberate efforts to attain specific goals are at variance. Indeed, Deci (1975) and DeCharms (1976) defined intrinsic motivation as motivation that is inherent in the activity being pursued and the meaning assigned to it. A striking finding reported by Harter (1981) is that intrinsic motivation, as an important energizer for learning, is being lost as children move up in the elementary school system. Deci and Ryan (1985) identified a number of factors in the learning environment that are supportive of intrinsic motivation, and they put forward principles to foster this type of motivation.

For quite some time, proponents of the two theoretical traditions have competed for dominance in the field of education, but they were not successful in convincing educational psychologists that their constructs should be incorporated into theories of learning and instruction. Fortunately, in the 1980s, the two motivation theories were integrated into a new theoretical framework, displaying a theoretical richness that has enlightened researchers investigating cognitive skills and school achievement. Findings stemming from this new research tradition, called goal orientation, substantially contributed to our knowledge of student motivation in the life classroom. It became clear that the quality of the learning process depends to a large extent on the students' goal orientation, more specifically on the type of criteria students use to determine the success of a particular learning activity (Dweck 1986, Nicholls 1984). Students who perceive school tasks as opportunities to learn predominantly use mastery criteria to determine the merit of their work (task orientation); they focus on cues in the learning situation that invite them to build up their competence and confidence and solicit feedback to assess their progress. In contrast, students who rely on performance criteria to determine their successes view school tasks as opportunities to demonstrate their (in)ability.

Several research groups argued that students' goal orientation is a prominent motivational unit, since it is directly linked to cognitive engagement and strategy use. It was demonstrated that students of different age groups who score high on task orientation select deeper processing strategies than students high on ego orientation (e.g., Meece et al. 1988). These results were applauded by researchers who study the development of 'higher order' skills, mainly because they saw for the first time that motivational beliefs directly affect cognitive strategy use. They finally had a motivational unit on a par with cognitive strategy use. However, equivalent positive relationships were demonstrated between cognitive strategy use and other motivational beliefs including interest, intrinsic motivation, self-efficacy, and self-concept of ability. The profusion of motivation constructs which were introduced into the educational literature created complexity, ambiguity, and tension, not least because researchers found it hard to choose the most appropriate motivation constructs to integrate into their models. They also lost their way in the forest of motivation questionnaires. This situation of abundance, linked to the fact that researchers in the field of metacognition had renamed several motivation constructs and grouped them together under the heading of 'metacognitive experiences,' turned the field of metacognition and learning strategies into a jungle where motivation researchers and metacognitive specialists compete for dominance.

2. *Impact of Motivation on Instruction*

How do different instructional practices influence student motivation? Although many studies have been designed to answer this question, much more research is necessary to understand how instruction variables interact with student and group characteristics. One of the difficulties in answering the question is that teachers use an enormous repertoire of strategies to create different learning settings. A helpful categorization of learning settings was put forward by Ames and Ames (1984), who made a distinction between competitive, individualized, and cooperative learning settings and reported typical motivational practices which impact on students' goal orientation. Ames (1992) argued that motivation practices which make social comparisons salient in the classroom, such as using ability groupings, evaluation of individual abilities, and providing public feedback, elicit ego-focused cognitions and affects. There is evidence that teachers can play down ego-focused learning goals by focusing on what students have already mastered rather than on what has gone wrong. The extent to which teachers succeed in creating a mastery-oriented learning setting, where everybody who puts in effort can experience

some degree of success, is considered to be an indication of one's professional expertise (Brophy 1999).

Moreover, research on motivation (see e.g., Pintrich and Schunk 1996, Stipek et al. 1998, Wigfield et al. 1998) suggests that students are more motivated when teachers (a) supply interesting tasks, meaning that they present an optimal level of challenge to the learners, (b) put greater emphasis on process and on seeking alternative solutions, (c) encourage and reward effort, (d) put emphasis on social goals, (e) avoid the use of controlling techniques, (f) manage to embed newly learned cognitive and motivation strategies within a variety of tasks and activities so that students can make use of these strategies within a wide range of application contexts. Although these principles are highly recommendable, there is still much to be learned about the motivational practices for students who prefer to work individually and for those who like to work in groups. Emerging findings justify the conclusion that teachers and students have different views on the effectiveness of motivational practices (see Thorkildsen et al. 1994). For example, not all teachers agree that drawing students' attention away from ability judgments is the way to proceed. There are also large groups of students who still think that teachers should take the responsibility for their learning, meaning that they should provide instructional support and motivate them to keep going. Researchers should further explore the degree to which various learning settings support students' values and basic psychological needs (needs for competence, autonomy, and social relatedness, cf. Deci and Ryan 1990).

3 *Bridging the Gap between Motivation and Strategy Use*

At this point, it is important to refer to three recent developments in motivation research that have the potential of providing more structure to the study of student motivation and of bridging the gap between research on motivation and the literature on learning and instruction. First, a clear distinction was made between two intricately related aspects of goal-directed behavior, namely motivation and volition. The second development refers to new measurement instruments which emanated from process-oriented models of student motivation and allowed researchers to assess student motivation *in situ*. The third development embodies a new framework for studying the interrelations between motivation and metacognition, called self-regulated learning.

3.1 *Making a Distinction between Motivation and Volition*

Several authors (e.g., Kuhl and Goschke 1994, Snow et al. 1996) argued that it is necessary to separate 'motivation' from its twin construct 'volition.' Snow et al. reviewed the literature on motivation and volition and concluded that, consonant with the bulk of theory and research, the motivation construct can best be conceptualized in terms of decision making and choice with respect to individual goals. It refers to a 'commitment pathway' which results when students contemplate or deliberate over options to enactment and implementation. They argued that volition begins when the goal has been set. Kuhl and Goschke (1994) explained that in the postdecisional phase, students mainly have to inhibit counter-intentional impulses and protect their intention from competing action tendencies. Following this line of argument, the study of motivation for learning is restricted and expanded at the same time: it is confined to the predecisional phase and expanded to include all the 'commitment pathways' that are, or are not, being established in relation to events, tasks, and activities that are part of a student's life.

3.2 *Measuring Motivation, as Commitment Pathways, in Relation to Learning*

Traditionally, motivation was measured as a personality characteristic, meaning that it was equated with relatively stable beliefs that are important reasons for action. This approach to motivation led to the development of trait questionnaires that measure students' inclination to act as an enduring characteristic, thus categorizing them into either one of two opposite categories, such as 'intrinsically motivated vs. extrinsically motivated students,' 'success-oriented students vs. failure-oriented students,' 'ego-oriented students vs. task-oriented students.' Several authors cautioned against generalizations based on 'dispositional' motivational characteristics and warned about the use of favorable and unfavorable traits for learning. For example, Pintrich and Garcia (1991) provided evidence that it is incorrect to suggest that extrinsic motivation is an unfavorable trait for all students. These authors observed that the association between intrinsic motivation and the use of a deep level processing style was only evident in those students who scored low on extrinsic motivation. Extrinsic motivation also showed a positive association with deep level processing, but this relationship was only evident in those students who were low on intrinsic motivation.

It is important to realize that the trait approach to motivation describes general commitment pathways based on enduring motivational characteristics, thus ignoring the fact that students may have a tendency to react in a specific way to particular domains of knowledge (e.g., mathematics, science, language learning), based on their theory of mind and self. In the terminology of Cantor et al. (1986) students form subcategories of the self, related to various domains. Higgins (1987) referred to similar motivational units,

contrasting 'the actual self' to the 'ideal self' and the 'ought self.' These different aspects of the self are viewed as self-directive standards or acquired self-guides.

A subset of students' theory of self, namely their motivational beliefs (e.g., self-efficacy beliefs, values, expectancies, wishes), are activated when they are confronted with tasks, learning activities, or specific courses and interact with these concrete aspects of reality (e.g., with mathematics as a subject-matter domain), thus giving meaning and value to all learning activities that are part of that domain (Boekaerts 1996). It is clear that students' motivational beliefs about the various subject-matter domains taught in school and about specific leisure activities may be rather weak at the beginning of primary school. However, these beliefs grow steadily once students realize that different academic subjects and leisure activities have their own inherent structure, rules, and standards. Once they are established, domain-specific motivational beliefs give rise to domain-specific motivation or domain-specific commitment pathways.

Although it is feasible that a positive inclination towards homework will be translated into domain-specific motivation to do homework for mathematics, there is no guarantee that students will form a domain-specific commitment pathway. This will depend on the unique way in which they have mentally represented that subject-matter domain, particularly in terms of motivational beliefs. In a similar vein, it is not certain that students who have favorable domain-specific motivational beliefs will always display that behavior in a given situation. Indeed, domain-specific motivational beliefs activated from memory interact with current learning opportunities giving rise to 'situation-specific motivational beliefs' or 'appraisals.' In other words, students who are invited to invest effort in a specific mathematics problem-solving activity may or may not translate their domain-specific needs, expectations, and wishes into concrete behavioral intentions. Behavioral intentions generated in the situation are called goal intentions. They reflect the individual's decision to use his or her personal resources to reach the goal. In learning situations, these goal intentions are referred to as 'learning intentions' because they denote the students' intentions to engage in a specific learning activity, given local conditions (context-specific commitment pathways).

The point being made is that one can only study student motivation as an actual state of the learner when students are invited to observe the unique properties of a current situation; only then will their unique cognitions and feelings surface and reveal context-sensitive or insensitive behavior. The experience sampling method makes it possible to record such 'student by situation transactions' while students are actually participating in unfolding learning episodes. Boekaerts (1999), Krapp (1999), Vermeer et al. (2000), and Volet (1997) studied the impact of situational factors on student motivation in concrete learning situations. Preliminary evidence suggest intraindividual and gender differences in appraisals and goal intention.

In conclusion, regularities in student behavior, defined in terms of stable motivational characteristics should not be confused with (a) students' tendency to react favorably or unfavorably in relation to a domain of study, and (b) with their sensitivity, given local conditions. The distinction made between inclination, tendency, and sensitivity allows researchers to describe the time frame of different 'commitment pathways.' Both domain-specific motivation and situation-specific motivation can be viewed as explicit behavioral intentions. However, only motivation measured at the subordinate level (in the situation), is a deliberate choice to initiate a specific task or activity, because it is based on a student's appraisal of (a) the learning situation, given the local context, and (b) rivaling tendencies.

3.3 Studying Metacognition and Motivation as Interwoven Aspects of Self-Regulated Learning

It has already been emphasized that researchers who investigate the use of learning strategies have largely ignored motivational constructs in their theories and models, while researchers working within the motivation tradition have always assumed that motivation for learning is associated with higher school performance, without bothering to examine whether highly motivated students have access to better learning strategies. Fortunately, these two lines of research have merged. However, researchers still concentrate too much on academic goals and devote little attention to the social and well-being goals that children and adolescents also pursue in a school setting. Yet, as Wentzel (1996) argued convincingly, the social worlds of youngsters, and thus their social goals, play a dominant and influential role in their lives at school. Wentzel drew attention to social motivation, documenting that there are strong links between academic and social components of motivation (e.g., wanting to help others, to share information with classmates, to take responsibility for the learning of others).

Since the early 1990s, researchers have become fully aware of the multiple goals that students pursue in a classroom context and also of the fact that these goals may be in harmony or disharmony. The unique ways in which students steer and guide their behavior to meet these multiple goals are viewed as indicators of their self-regulatory skills. Self-regulated learning offers a new perspective on the study of motivation for learning, since it encourages researchers to go beyond the mere description of activity in the one control system (cf. Boekaerts and Niemivirta 1999). When researchers are prompted to study cognitive and

metacognitive processes in close connection with motivation, mutual influences are the focus of attention, and emerging properties of the system as a whole will be noted. This new approach will hopefully tell us how different forms of control are realized in a classroom context, how feedback processes operate in different control systems, and how students coordinate the interacting processes.

4. Conclusion

The study of motivation has always been the study of inferred causes for goal-directed behavior, but several shifts in focus have been noted: a shift from needs to cognitions, from extrinsic to intrinsic values, from performance-oriented criteria for success to activity-oriented criteria. Recent developments in theory and instrument construction announce a new era for research on motivation for learning. The construct of self-regulated learning inspires researchers to explore the factors that facilitate adaptive motivational patterns that are instrumental to learning in school and beyond (lifelong learning). In focusing on self-regulated learning, motivation researchers have also ventured beyond the traditional motivation agenda, investigating motivation control as an important aim in itself. They advocate that the ability to elicit favorable motivational beliefs in learning situations, but also at choice points in one's career and life, is an important resource.

See also: Academic Achievement Motivation, Development of; Intrinsic Motivation, Psychology of; Motivation and Actions, Psychology of; Motivation: History of the Concept; School Achievement: Cognitive and Motivational Determinants; Schooling: Impact on Cognitive and Motivational Development; Self-concepts: Educational Aspects; Self-efficacy: Educational Aspects

Bibliography

Ames C 1992 Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology* **84**: 261–71

Ames C, Ames R 1984 *Research on Motivation in Education: Vol. 1. Student Motivation*. Academic Press, San Diego, CA

Atkinson J W 1958 Towards experimental analysis of human motivation in terms of motives, expectancies, and incentives. In: Atkinson J W (ed.) *Motives in Fantasy, Action and Society: A Method of Assessment and Study*. Van Nostrand, Princeton, NJ, pp. 839–62

Boekaerts M 1996 Personality and the psychology of learning. *European Journal of Personality* **10**: 377–404

Boekaerts M 1999 Motivated learning: studying student situation transactional units. *European Journal of Psychology of Education* **14**(1): 41–55

Boekaerts M, Niemivirta M 1999 Self-regulated learning: Finding a balance between learning goals and ego-protective goals. In: Boekaerts M, Pintrich P R, Zeidner M (eds.) *Handbook of Self-regulation*. Academic Press, San Diego, CA, pp. 417–46

Brophy J 1999 *Teaching*. Educational Practices Series-1. International Academy of Education, Brussels, Belgium

Cantor N, Markus H, Niedenthal P, Nurius P 1986 On motivation and the self-concept. In: Sorrentino R M, Higgins E T (eds.) *Handbook of Motivation and Cognition*. Guilford Press, New York, pp. 96–121

DeCharms R 1976 *Enhancing Motivation in the Classroom*. Irvington, Halstead-Wiley, New York

Deci E L 1975 *Intrinsic Motivation*. Plenum, New York

Deci E L, Ryan R 1985 *Intrinsic Motivation and Self-determination in Human Behavior*. Plenum, New York

Deci E L, Ryan R 1990 A motivational approach to self: Integration in personality. In: Dienstbier R A (ed.) *Nebraska Symposium on Motivation: Vol. 38. Perspectives on Motivation*. University of Nebraska Press, Lincoln, NE pp. 237–338

Dweck C S 1986 Motivational processes affecting learning. *American Psychologist* **41**: 1040–8

Harter S 1981 A model of mastery motivation in children: Individual differences and developmental change. In: Pick A (ed.) *Minnesota Symposium on Child Psychology*. Erlbaum, Hillsdale, NJ

Heckhausen H 1977 Achievement motivation and its constructs: A cognitive model. *Motivation and Emotion* **1**: 283–329

Higgins E T 1987 Self-discrepancy: A theory relating self and affect. *Psychological Review* **94**: 319–40

Krapp A 1999 Interest motivation and learning – An educational psychological perspective. *European Journal of Psychology of Education* **14**: 23–40

Kuhl J, Goschke T 1994 A theory of action control: Mental subsystems, modes of control, and volitional conflict resolution strategies. In: Kuhl J, Beckmann J (eds.) *Volition and Personality: Action Versus State Orientation*. Hogrefe & Huber, Seattle, WA, pp. 127–54

Meece J L, Blumenfeld P C, Hoyle R H 1988 Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology* **80**: 514–23

Nicholls J G 1984 Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review* **91**: 328–46

Pintrich P R, Garcia T 1991 Student goal orientation and self-regulation in the college classroom. In: Maehr M L, Pintrich P R (eds.) *Advances in Motivation and Achievement: Goals and Self-regulatory Processes*. JAI Press, Greenwich, CT, pp. 371–402

Pintrich P R, Schunk D 1996 *Motivation in Education: Theory, Research, and Applications*. Merrill, Englewoods Cliffs, NJ

Snow R E, Corno L, Jackson D 1996 Individual differences in affective and conative functions. In: Berliner D C, Calfee R C (eds.) *The Handbook of Educational Psychology*. Prentice-Hall, London, pp. 243–310

Stipek D, Givvin K D, Salomon J M, MacGyvers V L 1998 Can a teacher intervention improve classroom practices and student motivation in mathematics? *The Journal of Experimental Education*. **66**: 319–37

Thorkildsen T A, Nolen S B, Fournier J 1994 What is fair? Children's critiques of practices that influence motivation. *Journal of Educational Psychology* **86**: 475–86

Vermeer H J, Boekaerts M, Seegers G 2000 Motivational and gender differences: Sixth-grade students' mathematical

problem-solving behavior. *Journal of Educational Psychology* **92**: 308–15
Volet S E 1997 Cognitive and affective variables in academic learning: The significance of direction and effort in students' goals. *Learning and Instruction* **7**: 235–54
Wentzel K R 1996 Social and academic motivation in middle school: Concurrent and long-term relations to academic effort. *Journal of Early Adolescence* **16**: 390–406
Wigfield A, Eccles J S, Rodriguez D 1998 The development of children's motivation in school contexts. *Review of Research in Education* **23**: 73–118

M. Boekaerts

Motivation, Neural Basis of

This article will describe the organization of the neural systems and their interactions that are critical for the expression of motivated behaviors. Presented first is an influential model of how motivated behaviors are organized temporally that will provide the background for discussing the structure of underlying neural circuits. It will emphasize that any model or circuit analysis must ultimately account for the complexity of all behavioral components, and must explain how temporal sequences (ethograms) of behavioral episodes are organized and controlled. The neural systems thought to be critical for generating motivated behaviors are then discussed. At this point however, it is worth considering that, although the forebrain plays a critical role, there is probably no single neural system that puts the 'motivation' into 'behavior.'

1. Temporal Organization of Motivated Behaviors

Motivated behavior is initiated by the neural interaction of *interosensory* information, *exterosensory* information, and *arousal state*. The outcome of this integration determines the value of the '*drive*' associated with a particular behavior, and is used to select the most appropriate behavioral action, which is then expressed as a *procurement* (*or appetitive*) *phase* where the goal object is actively sought out (Toates 1986). Procurement behaviors include foraging, are individualized for the particular situation, can be quite complex, and are under cognitive control. When the goal object has been located, a *consummatory phase* allows the animal to interact directly with the goal object with more stereotypic rhythmic movements—licking, chewing, copulating, etc. Reward/aversion functions, together with learning and memory processes are also critical during the procurement and consummatory phases; a previously rewarded or aversive experience of a particular goal object, remembering where it is located, and how to get there are important considerations. Finally, as the consummatory phase continues, a variety of interosensory feedback signals are generated that increase the probability that it will be terminated (commonly called satiety), presumably through the action of inhibitory networks. However, termination may also arise from separate exterosensory signals (e.g., the presence of a predator) that can override the ongoing behavior at any time, so allowing the animal to switch immediately to one that is more appropriate (McFarland and Sibly 1975).

2. Neural Systems

At the simplest level, four broad-ranging neural systems are concerned with generating motivated behaviors: those involved with the transduction and processing of sensory signals; those that control arousal state; those that process the types of information that neurally represents sensory objects; and those involved with motor control (Fig. 1). Using this scheme, four principal sets of inputs activate

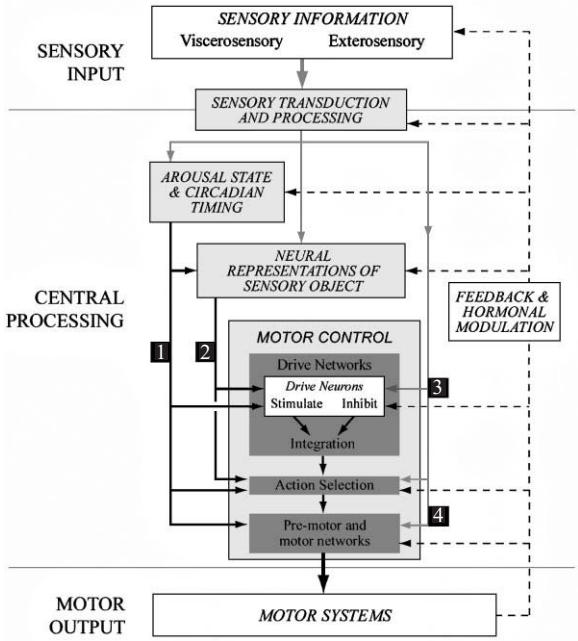

Figure 1
A schematic representation of the neural systems involved with generating motivated behaviors. Points 1, 2, 3, and 4 represent various inputs to motor control networks that activate behaviors. Inputs 1 and 2 are associated with complex anticipatory behaviors, 3 with homeostatically-generated behaviors; all three of which are motivated. Input 4 can only generate reflex (i.e., nonmotivated) actions (abbreviations: Stimul., stimulatory; Inhib., inhibitory; Disinhib., disinhibitory)

motor systems and generate motivated behaviors. The most complex motor control processes are those that generate *anticipatory behaviors* and their accompanying autonomic and neuroendocrine motor events. In some instances, information derived from the systems controlling arousal state—for example, the *circadian clock*—may provide the predominant signals (input 1, Fig. 1). But in many instances this type of anticipatory control derives from the interactions between processed sensory information and those forebrain systems concerned with encoding *object representation*, particularly learning and memory, reward, and spatial orientation. The integrated output of these regions then influences the motor control systems (input 2, Fig. 1). However, increased drives for motivated behavior can also be produced by internally-generated *deficit signals* (e.g., the thirst arising from dehydration, or the hunger from starvation) that engage motor control networks more directly (input 3, Fig. 1). Finally, simple *reflex actions* (e.g., reflex rejection of unpalatable foods or fluids) are generated by more or less direct sensory inputs to the pre-motor and motor networks with little higher-order processing (input 4, Fig. 1). But of course, these reflex actions lack any motivated character.

The notion of drive and the reduction in the level of specific drive states following particular behaviors has been a very influential if somewhat controversial concept in neuroscience. The idea that some neural mechanism selects the behavior that is the most appropriate to reduce the drive state that has the highest value is a useful one from the perspective of delineating neural systems. In this way, drive is determined by an interaction between arousal state, those interosensory signals that encode internal state, and sensory object representations that effectively encode the current incentive value of a particular goal object (see Toates 1986). In this way, inputs 1, 2, and 3 to motor control networks (Fig. 1) can be thought of as 'drive-determining' interactions. Finally, the integration of this information within motor control systems governs which action is selected, and thereby which particular behavior is expressed.

2.1 Sensory Information

The neural systems that control motivated behaviors are regulated by a host of sensory inputs, which can be categorized either as interosensory signals that encode internal state or exterosensory inputs that encode features of the goal object such as smell, taste, temperature, tactile properties, and appearance. Each of these sensory signals has specific transduction mechanisms and 'labeled line' access to central processing networks found throughout the brain. Although a great deal of important sensory processing occurs within the telencephalon—particularly sensory cortex—the preliminary sensory processing that occurs subcortically can have important implications for the control of motivated behaviors, for example the altered gustatory processing that occurs in the hindbrain of hyponatremic animals is an important adjunct to increased sodium appetite.

Some sensory signals can access command networks directly, as typified by the drinking initiated by the effects of increasing plasma osmolality or angiotensin II (A-II; see *Thirst and Drinking*), or the deficit-induced feeding activated by those adiposity signals (primarily leptin and insulin) that have direct hypothalamic actions (Elmquist et al. 1999). In both these cases, it is not clear whether the initial sensory processing of deficit signals utilize the sensory object representation networks, but the fact that hypothalamic regions involved with this type of sensory transduction project directly to regions concerned with ingestive motor control, suggests that they may not.

2.2 Arousal State Control and Behavioral Timing

Parts of the brain provide critical timing information and control arousal state so enabling motor command networks to generate the types of actions that are anticipatory. This system includes the suprachiasmatic nucleus in the hypothalamus, which generates the circadian timing signal that entrains virtually all neural activity within limits determined by the prevailing photoperiod. In terms of arousal state, catecholamine cell groups in the hindbrain (e.g., the locus coeruleus), histaminergic neurons in the tuberomammillary nucleus, the ventrolateral preoptic nucleus, and the recently identified hypocretin/orexin neurons in the LHA also supply information that is of critical importance for controlling arousal state (see *Arousal, Neural Basis of*; *Circadian Rhythms*; *Sleep: Neural Systems*; *Suprachiasmatic Nucleus*).

2.3 Neural Representation of Sensory Objects

Some processes important for motivational motor control generate neural representations of sensory objects. These include the following: learning and memory mechanisms in the telencephalon and cerebellum; reward/aversion systems in the midbrain ventral tegmentum, parts of the basal forebrain (particularly the nucleus accumbens), amygdala, and parts of the cortex, particularly prefrontal regions; and systems in the hippocampus and parts of the cortex responsible for allocentric and egocentric spatial representation. A great deal of exterosensory information is processed through these networks, which collectively assign what has been called '*incentive value*' to a particular goal object (Toates 1986). The neural pathways mediating the interactions be-

tween the object representation and motor networks are not fully understood, but sets of bidirectional connections between the hypothalamus and cortical structures such as the prefrontal cortex and hippocampus, as well as subcortical regions such as the amygdala, septal nuclei, bed nuclei of the stria terminalis, and basal ganglia are all likely to be critical in the integrative operations that designate and coordinate the full spectrum of motor events associated with motivated behaviors (Saper 1985, Risold et al. 1997, Swanson and Petrovich 1998) (see also *Amygdala (Amygdaloid Complex)*; *Cognitive Control (Executive Functions): Role of Prefrontal Cortex*; *Learning and Memory, Neural Basis of*; *Neural Representations of Direction (Head Direction Cells)*; *Reinforcement: Neurochemical Substrates*).

2.4 Motor Control

Fig. 1 shows that motor control systems operate at three levels: drive networks that set up and coordinate specific motor events; action selection systems concerned with the planning, selection, and the moment-to-moment execution of particular motor actions; and the premotor/motor neuron networks that actually execute them motor actions (see *Behavior, Hierarchical Organization of*). Furthermore, Fig. 1 shows a clear distinction between drive networks and object recognition systems. This derives from the fact that object representation systems are goal nonspecific, whereas drive networks are explicitly concerned with specific behaviors. For example, an animal uses the same networks in the telencephalon for spatial navigation whether it is looking for food, a mate, or shelter, whereas parts of the hypothalamus are concerned exclusively with feeding or sexual behavior. However, in terms of the organization of other neural structures (e.g., cell groups in the basal ganglia), this distinction may not be so clear-cut, and it may prove difficult to determine experimentally within which system a particular set of neurons belongs.

2.4.1 Drive networks. The concept of drive networks evolved from the idea of discretely localized hypothalamic satiety and hunger 'centers' popular during the 1950s and 1960s. The center concept was synthesized by Eliot Stellar into a scheme for explaining the role of the hypothalamus in controlling motivated behaviors (Stellar 1954). Elaborating these pioneering studies since has replaced the idea of isolated 'centers' with a scheme where sets of more widely distributed but highly interconnected motor drive networks direct the somatic, autonomic, and neuroendocrine motor responses that accompany particular behaviors. In accord with Stellar's original idea, each individual drive network contains neurons that either stimulates or inhibits a particular motor event. The exact nature of the expressed behavior, or whether it is expressed at all, is then determined by the integrated output of these neurons. Furthermore, it is likely that a separate series of circuits is concerned with disinhibiting particular motor actions. Although the location of these disinhibitory circuits is unclear, Swanson (2000) has suggested that neurons in the pallidum may function in this way.

Determining the precise anatomical substrates of the drive networks has proved quite difficult. However, a wealth of data cataloging the stimulatory or inhibitory effects of neuropeptides on particular motor functions clearly concurs with earlier lesion and stimulation experiments that have identified the hypothalamus as a key locus for the components of specific drive networks (see *Hypothalamus*). Many groups have identified which hypothalamic neurons synthesize these neuropeptides and have determined where they project. One critical point that has emerged from these studies is that placing individual hypothalamic cell groups within one or other of drive network is rather difficult. This is because traditionally-defined cell groups such as the lateral hypothalamic area (LHA), arcuate (ARH), or paraventricular (PVH) nuclei appear to contain elements that, in terms of function, belong to more than one type of drive network; it seems unlikely that there is a tight 'one cell group–one network' relationship in the hypothalamus.

A well-documented example of a stimulatory network is the one activated by circulating A-II to stimulate drinking. Results from many studies have shown that A-II is detected by a central sensory transducer, the subfornical organ (SFO), which then directly and specifically stimulates water intake. The SFO provides efferents—most of which also contain A-II (Swanson 1987)—to a relatively limited set of structures, including parts of the prefrontal cortex, substantia innominata, medial preoptic area, bed nuclei of the stria terminalis, zona incerta, PVH, supraoptic nucleus, and LHA (Swanson 1987); presumably it is these regions that constitute part of the stimulatory network that initiates the motor aspects of drinking (see *Thirst and Drinking*). Neuropeptide Y (NPY) neurons in the ARH contribute to a well-known example of a stimulatory eating mechanism; results from many studies show that NPY contributes to a circuit that directly stimulates food intake (see *Hunger and Eating, Neural Basis of*).

In terms of inhibitory networks, feeding again provides good examples of hypothalamic components. For example, the peptide α-MSH is synthesized in ARH neurons from the proopiomelanocortin (POMC) gene and provides an inhibitory signal to feeding that acts at melanocortin 4 receptors expressed by neurons in the LHA and PVH (Elmquist et al. 1999). POMC neurons in the ARH and retrochiasmatic area also express cocaine- and amphetamine-regulated transcript. This widely expressed

neuropeptide will inhibit eating without affecting spontaneous drinking when injected introcerebroventricularly.

Finally, interactions between different motivational drive systems, particularly in the hypothalamus are of paramount importance. For example, the effects of starvation are not limited just to increasing the drive to eat, they also reduce reproductive capacity. Similarly, dehydration leads to severe anorexia as well as increasing the drive to drink (Watts 2001). This cross-behavioral coordination is achieved by hormonal modulation acting together with the divergent neuroanatomic outputs from individual drive networks.

2.4.2 Premotor and motor networks. The brain has three sets of motor neurons with which to effect both motivated behavioral action and the internally-directed motor actions that accompany it: α-motor neurons in the ventral horn of the spinal cord that control the striate musculature and hence the expression of all behavior; postganglionic sympathetic and parasympathetic motor neurons that control autonomic motor function; and sets of parvicellular and magnocellular neuroendocrine motor neurons in the hypothalamus that control pituitary function. Motivated behaviors involve the coordinated activation, to a greater or lesser degree, of all three motor systems (see *Autonomic Classical and Operant Conditioning*; *Neuroendocrinology*).

Sets of premotor networks directly control oscillatory and more complex patterns of motor neuron firing. Simple rhythmic movement patterns develop from an interaction between oscillatory rhythm generators, which directly involve the motor neurons, and networks of premotor central pattern generators located somewhat more distally in the spinal cord and hindbrain. A critical feature of these pattern generators is that they are capable of producing rhythmic output without sensory input. In turn, pattern generator output is modulated further by afferents from those parts of the appropriate drive networks in the diencephalon and telencephalon. These often highly varied inputs provide the critical command and contextual information from object representational networks for selecting the most appropriate motor program at any particular time. Collectively, all these components begin to account for how motivated behavior is organized.

2.5 Feedback and Humoral Modulation

Feedback is a critical feature of behavioral control, and sensory signals encoding the magnitude and consequences of generated motor actions can control the length of a motivated behavioral episode. For example, postabsorptive humoral feedback (e.g., increasing CCK or decreasing plasma osmolality) and interosensory signals (e.g., gastric distension, oropharyngeal metering) lead to the termination of ingestive behaviors and subsequent behavioral refractoriness. Finally, rather than acting as feedback control a further set of regulatory signals are more modulatory and can influence a variety of neural structures at all brain levels. Steroid hormones, particularly gonadal steroids, are important components of this type, and many familiar motivated behaviors are critically regulated by this type of signal.

See also: Amygdala (Amygdaloid Complex); Arousal, Neural Basis of; Circadian Rhythms; Hunger and Eating, Neural Basis of; Learning and Memory, Neural Basis of; Reinforcement: Neurochemical Substrates; Reinforcement, Principle of; Sex Hormones and their Brain Receptors; Sexual Behavior: Sociological Perspective; Sleep: Neural Systems

Bibliography

Elmquist J K, Elias C F, Saper C B 1999 From lesions to leptin: Hypothalamic control of food intake and body weight. *Neuron* **22**: 221–32

McFarland D J, Sibly R M 1975 The behavioural final common path. *Philosophical Transactions of the Royal Society of London, Series B* **270**: 265–93

Risold P Y, Thompson R H, Swanson L W 1997 The structural organization of connections between hypothalamus and cerebral cortex. *Brain Research Reviews* **24**: 197–254

Saper C B 1985 Organization of cerebral cortical afferent systems in the rat. II. Hypothalamocortical projections. *Journal of Comparative Neurology* **237**: 21–46

Stellar E 1954 The physiology of emotion. *Psychological Reviews* **61**: 5–22

Swanson L W 1987 The hypothalamus. In: Bjorklund A, Hökfelt T, Swanson L W (eds.) *Handbook of Chemical Neuroanatomy*. Elsevier, Amsterdam, pp. 1–124

Swanson L W 2000 Cerebral hemisphere regulation of motivated behavior. *Brain Research* **886**: 113–64

Swanson L W, Petrovich G D 1998 What is the amygdala? *Trends in Neuroscience* **21**: 323–31

Toates F 1986 *Motivational Systems*. Cambridge University Press, Cambridge, UK

Watts A G 2001 Neuropeptides and the integration of motor responses to dehydration. *Annual Review of Neuroscience* **24**: 357–84

A. G. Watts

Motivational Development, Systems Theory of

Systems theory provides a powerful method for the description of systems in which feedback controlled regulation processes occur. Since human goal-directed behavior is regulated by such processes, systems theory is also very useful for psychological research. One of the most elaborated psychological models based on

systems theory, the Zurich Model of Social Motivation by Bischof (1985, 1993), will be presented below, together with some empirical evidence for its validity.

1. Systems Theory

Systems or control theory originated from practical, engineering-related questions concerning regulation processes. Its aim is to analyze the structure of systems and to provide adequate, preferably quantitative, models for them. The basic unit of systems control is a negative feedback loop which consists of the following elements: a detector which provides the system with information about the environment (input), a reference value which indicates what the input ideally should be, and an output function that causes the system in case of a discrepancy between input and reference value to 'behave' in a way that affects the environment in order to reduce the discrepancy (for details see Bischof 1985, 1995).

The notion that processes such as these not only occur in machines, but also can be found in living systems such as animals or human beings was first introduced into the life sciences by Wiener (1948). Almost at the same time, it was also formulated in German by von Holst and Mittelstaedt (1971). Since then, this idea has appeared in different theories. The models based on systems theory, however, vary considerably in their degree of quantification, their complexity, and their generalizability.

One simple, general, and well-known example of a negative feedback loop used to characterize human behavior is the TOTE unit described by Miller et al. (1960). A more complex attempt to explain the general principles of goal directed behavior in terms of systems theory stems from Carver and Scheier (1998). Examples of mathematically formulated control systems and their experimental verification can be found in Powers (1978).

Since systems theory not only is especially apt for the description of purposive behavior, but also allows for the description of short-term as well as long-term regulation processes, models or theories dealing with more specific topics can be found in the fields of motivational as well as developmental psychology. Examples in the area of motivational psychology are the control theory model of work motivation by Klein (1989) and the control system model of organizational motivation by Lord and Hanges (1987). The former describes individual work motivation as a function of feedback and several cognitive processes. The latter also deals with task performance, but can be applied to individual motivation as well as interpersonal activities in organizations. A more comprehensive model, which is at least partly based on a systemic approach, is presented by Dörner (1994, 1998), who intends nothing less than to discover the 'construction plan' of the human soul. In contrast to these authors, Hyland (1988) uses his motivational control theory as a meta-theoretical framework to integrate different, already existing, motivational theories which—according to him—reflect different aspects of a single underlying process which can again be described with a negative feedback loop.

In developmental psychology, a dynamic systems approach is used, for example, by Fogel and Thelen (1987) in their theory of motor development, even if they do not put it into terms of systems theory. The models of Burke (1991) and Bell (1971) are more explicitly based on systems theory. The former describes the process of identity formation as a negative feedback loop. The latter develops a control system to explain reciprocal socialization effects between children and their caregivers. A much broader model which is concerned with the general principles of human development is the developmental systems theory by Ford and Lerner (1992). The best known example for a systems approach in developmental psychology, however, is probably Bowlby's (1969) theory of infant attachment, a phenomenon which he viewed in feedback terms, although he never ventured to formulate it in a mathematical model. This was eventually performed by Bischof (1975, 1985, 1993), who provided a completely quantified model of social motivation. This is referred to as the Zurich Model of Social Motivation (or Zurich Model) and will be described in the following section.

2. The Zurich Model of Social Motivation

The Zurich Model of Social Motivation by Bischof (1985, 1993, Fig. 1) arose from the want of a quantitative model able to explain and predict processes of social distance regulations. Such regulations have been described by attachment theorists such as Bowlby (1969) and also occur during individuation processes in puberty and adolescence. In short, the model postulates three basic motivational systems: the security, the arousal, and the autonomy system. All of these are represented as negative feedback loops and are considered homeostatic. In addition, there is a fourth system, the coping system, which deals with cases in which one of the other systems gets blocked and cannot operate in its usual manner. These four systems will be described qualitatively, although there also exists a mathematical description (Bischof 1993, Gubler and Bischof 1991, Gubler et al. 1994).

2.1 The Security System

Security is defined as the feeling of intimacy, warmth, protection, and so forth that one obtains when one is close to a familiar person, typically one's primary caregiver. This feeling has its source in three input variables, all of which are sensed by a specific detector. The detector for familiarity provides the information

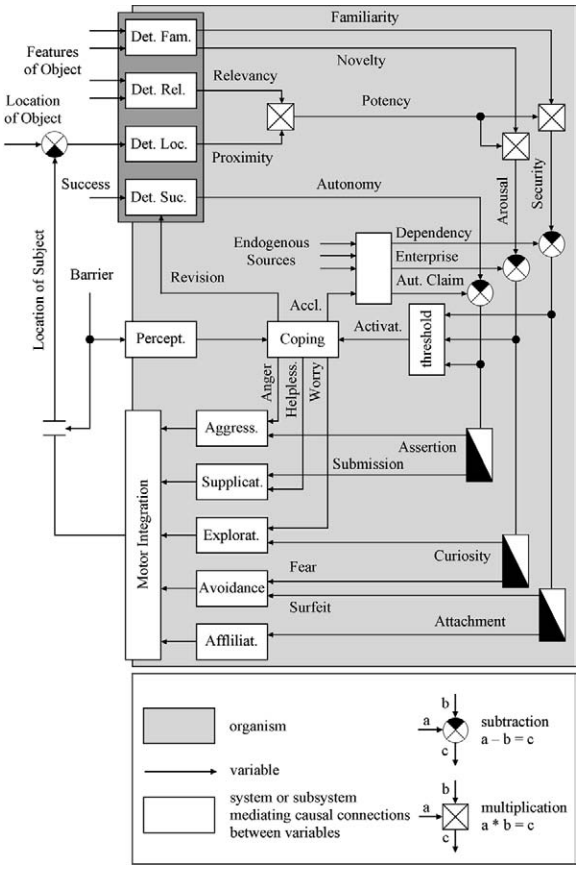

Figure 1
The Zurich Model of Social Motivation (after Bischof 1993, p. 13)

as to whether a certain object is familiar, in contrast to being new, unknown, or discrepant from one's expectancies. The detector for relevancy assesses the rank and the 'conspecificity' of the object in question. The detector's output is highest when the perceived object is an adult, high-ranking conspecific, and can be reduced, for example, by submissive behavior of the object or by substitution of a human adult by another object such as a child or an animal. The third detector gives information about the distance between oneself and the perceived object. The closer one is positioned to a familiar object the more security one obtains. Hence, the construct of security refers to a hypothetical emotional variable monotonically increasing with familiarity, relevancy, and proximity. Since proximity is the only one among those three variables which the subject can control, the obvious way to provide a homeostasis of security results in distance-regulating behavior.

The reference variable of the security system is the degree of dependency the subject feels. As long as dependency exceeds security, an appetence for security is maintained. This induces the subject to show attachment behavior, that is, to reduce the distance to a person who is able to provide security. The opposite situation, frequently encountered in puberty, results in an aversion against security and consequently in an avoidance of familiar persons (surfeit behavior).

2.2 *The Arousal System*

Arousal is defined as feelings such as interest, fascination, curiosity, as well as feelings of alarm or fear which can occur when one encounters a nonfamiliar person. Basically, the arousal system can be thought of as analogous to the security system, and moreover it can utilize the same three detectors. Like security, arousal is hypothesized to increase with the object's proximity and relevancy. Since the familiarity detector also provides information about how novel an object is, arousal can be connected to the output of this detector as well. Thus, the paradigmatic case for the arousal system is the encounter with a high-ranking, adult human stranger.

The reference value of the arousal system is labeled enterprise. Again, a distinction between an aversive and an appetent behavioral response can be made. Aversion against arousal and subsequent fear behavior result from arousal exceeding enterprise. If arousal falls short of enterprise, an appetence for arousal develops which leads to exploratory behavior.

The security and the arousal systems do not work independently of each other. Both systems respond to the same stimulus situations. Thus, the quantitative processing of the inputs involved leads automatically to effects such as fear-evoking stimuli intensifying attachment behavior towards an available caregiver, or the caregiver serving as a security base for exploratory ventures. Furthermore, the two systems do not develop independently during ontogeny. The reference variables—dependency and enterprise—are assumed to be negatively correlated, and to covary with the individual's age and maturity. Young infants are strongly dependent and only modestly enterprising. They need the presence of a security-providing caregiver in order to start exploring the environment and to acquire competence. Later in development, usually culminating around early adolescence, this pattern changes: growing independence and enterprise now demand and allow for the separation from the familiar partners and the establishment of new relationships. Finally, after adolescence, the two reference values normally level out.

2.3 *The Autonomy System*

The autonomy system can be regarded as analogous to the security and the arousal systems. The input here,

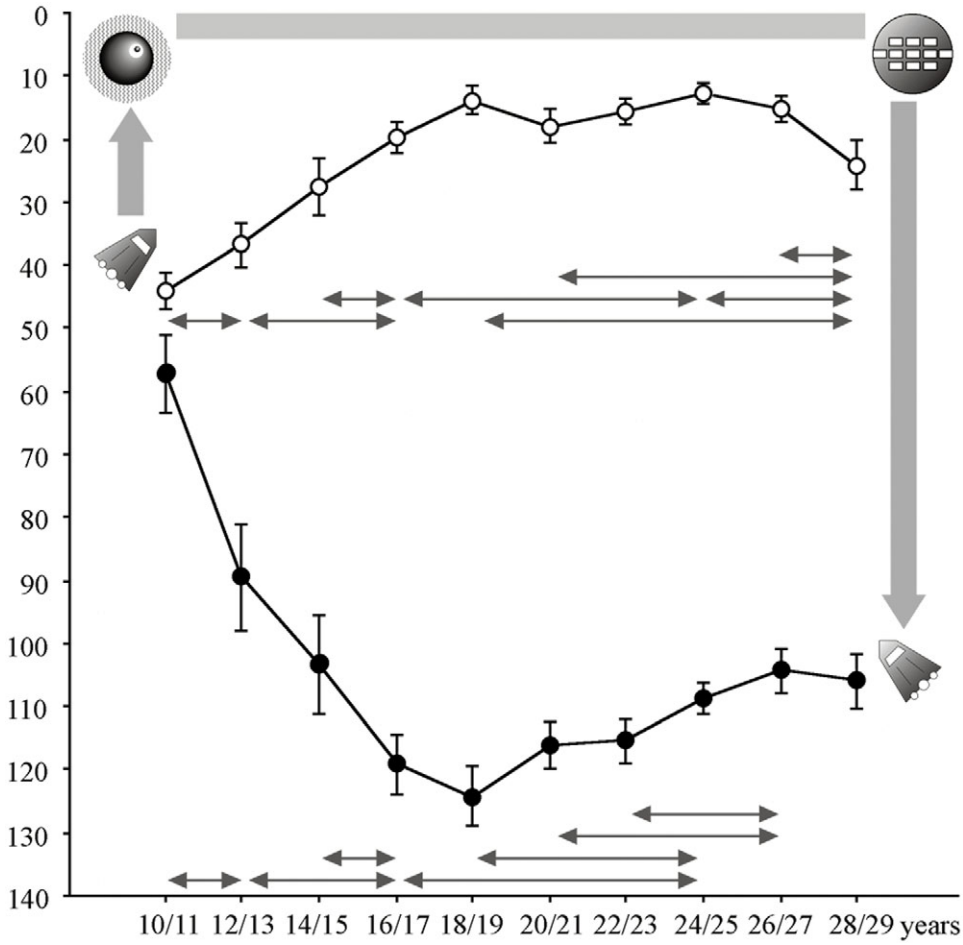

Figure 2
Mean proximity to the security base (black points) and maximal proximity to the source of arousal (white points) in virtual distance units. Error bars indicate standard errors. Horizontal arrows show the nearest age groups which differ with $p \leqslant 0.05$

however, stems from a different source. The detector which provides the amount of felt autonomy senses experiences of success such as obtaining obedience, acknowledgement, or admiration from other people, solving difficult problems, or being competent in certain subjects. Therefore, autonomy can be described as feeling competent, strong, and powerful, and having influence, prestige, and authority.

The reference variable of the autonomy system is called 'autonomy claim' and is closely related to the power and the achievement motives, as well as the need for prestige and an aspiration for self-confidence. In the case of a discrepancy between the feeling of autonomy and the autonomy claim, however, the model postulates the occurence of behaviors stemming clearly from the domain of power and dominance. If individuals feel an appetence for autonomy, they are assumed to behave assertively by becoming threatening, demanding, or even aggressive. In the opposite case, if individuals have too much autonomy and feel aversive, they will show submissive behaviors such as behaving humbly or servilely.

The autonomy claim is not independent of the other two reference variables, dependency and enterprise. On the contrary, the autonomy claim is postulated to underlie the connection of the security and the arousal system by affecting dependency negatively and enterprise positively.

2.4 *The Coping System*

The coping system operates in cases where one of the other systems gets blocked by an (internal or external)

obstacle or barrier. The coping system provides three external and two internal coping strategies, external strategies being active attempts to change the situation and internal ones being efforts to adapt oneself to the situation. The three external coping strategies are labeled invention (searching for a detour), aggression (attempting to destroy the barrier), and supplication (begging someone else to remove the barrier). The two internal strategies are referred to as acclimatization (adaptation of the reference value to the input) and revision (adaptation of the perception of the situation).

The model does not predict which one of these coping responses is performed in a given context since this is primarily a result of the individual's learning history. However, it is assumed that the internal strategies will succeed if the external strategies do not lead to a success.

3. Empirical Evidence

The assumptions made by the Zurich Model of Social Motivation were tested by means of several behavior observations and estimation studies. Both types of studies were conducted with either very young children of about one year of age or with older participants ranging from late childhood to adulthood. Since the Zurich Model contains several ideas of attachment theory, settings similar to the Strange Situation Test by Ainsworth et al. (1978) were used for the younger children. For the older participants, a spaceship simulator was developed in which they experienced a space flight which was also designed according to the Strange Situation.

In this simulator, the participants were seated in a cockpit where the virtual space was visible on a computer screen. There were several instruments to manipulate in order to move, orient oneself in space, and protect the spaceship from dangers. The effects of the participant's actions were visible on the screen and the participants could hear sounds, such as their gun firing or the noise of the accelerating engine. In addition, the simulator was moved according to the steering maneuvers of each participant, which ensured a high emotional involvement.

During their virtual flights the participants were confronted with two different objects. One object was considered familiar (either the earth or a home base) and served as an orienting point in space as well as protection. The proximity of the participants to this security base served as a measure for their dependency. A novel object—a strange planet—also appeared during the flight. This planet was fascinating as well as potentially dangerous to the participants. The proximity to this source of arousal was used to assess the enterprise of the participants.

In the late 1980s and throughout the 1990s 14 studies were conducted with this spaceship simulator. Participants were for the most part schoolboys and girls as well as students between 9 and 33 years (N = 704, mean age = 21.2 years). The data obtained in these studies (Fig. 2) strongly support one of the main theoretical assumptions made by the Zurich Model: dependency and enterprise correlate and take an inverse developmental course. The time of the biggest difference is attained at about 18 years, when dependency reaches its minimum while enterprise is at its maximum.

In the spaceship simulator, as well as the laboratory settings used for the younger children, the behavior of the participants was registered continuously. This allowed for investigation methods exceeding simple comparisons of groups by analyzing the distance regulation processes of single participants. In the estimation studies mentioned above, it was possible to simulate the actual behavior of the participants by means of a parameter estimation based on the mathematical formulation of the Zurich Model (Fig. 3).

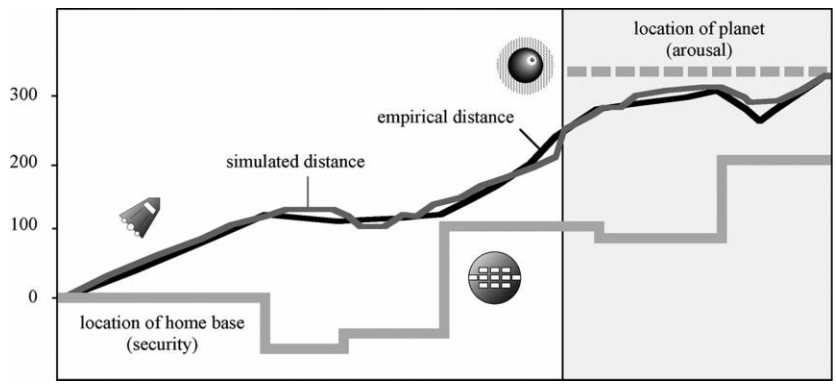

Figure 3
Example of simulated and real distance regulation behavior of one participant (after Gubler et al. 1994, p. 115). The vertical scale indicates virtual distance units, the horizontal scale represents the time of the flight

These studies gave further evidence for the validity of the model (for details see Gubler and Bischof 1991, Gubler et al. 1994).

4. Future Directions

While the functioning of the security and the arousal systems is sufficiently supported by empirical evidence, there are still several open questions for future research. Although the existing data also support the assumptions made about the autonomy system, the causal link of the autonomy claim to dependency and enterprise still has to be established. Since the Zurich Model was originally conceived from the point of view of a child, more adult motives such as altruism have to be included as well. Furthermore, the study of the developmental course taken by the different reference variables or motives also has to be extended to older age groups in order to get a model explaining the motivational development of the whole life span.

See also: Academic Achievement Motivation, Development of; Intrinsic Motivation, Psychology of; Motivation and Actions, Psychology of; Motivation: History of the Concept; Motivation, Learning, and Instruction

Bibliography

Ainsworth M D S, Blehar M C, Waters B, Wall S 1978 *Patterns of Attachment: A Psychological Study of the Strange Situation*. L. Erlbaum, Hillsdale, NJ

Bell R Q 1971 Stimulus control of parent or caretaker behavior by offspring. *Developmental Psychology* **4**: 63–72

Bischof N 1975 A systems approach towards the functional connections of attachment and fear. *Child Development* **46**: 801–17

Bischof N 1985 *Das Rätsel Oedipus—Die biologischen Wurzeln des Urkonfliktes von Intimität und Autonomie*. Piper, Munich, Germany

Bischof N 1993 Untersuchungen zur Systemanalyse der sozialen Motivation I: Die Regulation der sozialen Distanz—Von der Feldtheorie zur Systemtheorie. *Zeitschrift für Psychologie* **201**: 5–43

Bischof N 1995 *Struktur und Bedeutung. Eine Einführung in die Systemtheorie*. Huber, Bern, Switzerland

Bowlby J 1969 *Attachment and Loss: Vol. 1. Attachment*. Basic Books, New York

Burke P J 1991 Identity processes and social stress. *American Sociological Review* **56**: 836–49

Carver C S, Scheier M F 1998 *On the Self-regulation of Behavior*. Cambridge University Press, Cambridge, UK

Dörner D 1994 Eine Systemtheorie der Motivation. In: Kuhl J, Heckhausen H 1996 (eds.) *Enzyklopädie der Psychologie: Motivation, Volition und Handlung*. Hogrefe, Göttingen, pp. 329–57

Dörner D 1998 *Bauplan für eine Seele*. Rowohlt, Reinbek

Fogel A, Thelen E 1987 Development of early expressive and communicative action: Reinterpreting the evidence from a dynamic systems perspective. *Developmental Psychology* **23**: 747–61

Ford D H, Lerner R M 1992 *Developmental Systems Theory*. Sage, Newbury Park, CA

Gubler H, Bischof N 1991 A systems' perspective on infant development. In: Lamb M E, Keller H (eds.) *Infant Development: Perspectives from German-speaking Countries*. L. Erlbaum, Hillsdale, NJ, pp. 35–66

Gubler H, Bischof N 1993 Untersuchungen zur Systemanalyse der sozialen Motivation II: Computerspiele als Werkzeug der motivationspsychologischen Grundlagenforschung. *Zeitschrift für Psychologie* **201**: 287–315

Gubler H, Paffrath M, Bischof N 1994 Untersuchungen zur Systemanalyse der sozialen Motivation III: Eine Ästimationsstudie zur Sicherheits- und Erregungsregulation während der Adoleszenz. *Zeitschrift für Psychologie* **202**: 95–132

Hyland M E 1988 Motivational control theory: An integrative framework. *Journal of Personality and Social Psychology* **4**: 642–51

Klein H J 1989 An integrated control theory model of work motivation. *Academy of Management Review* **14**: 150–72

Lord R G, Hanges P J 1987 A control system model of organizational motivation: Theoretical development and applied implications. *Behavioral Science* **32**: 161–78

Miller G A, Galanter E, Pribram K H 1960 *Plans and the Structure of Behavior*. Holt, Rinehart, & Winston, New York

Powers W T 1978 Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology. *Psychological Review* **85**: 417–35

von Holst E, Mittelstaedt H 1971 The principle of reafference: Interactions between the central nervous system and the peripheral organs. In: Dodwell P C (ed.) *Perceptual Processing: Stimulus Equivalence and Pattern Recognition*. Appleton, New York, pp. 41–72

Wiener N 1948 *Cybernetics. Communication and Control in theAnimal and the Machine*. Wiley, New York

M. E. Schneider

Motor Control

The rich varieties of human and non-human animal movement are the final product of motor mechanisms within the nervous system and the muscles they innervate. The selective contraction and relaxation of these muscle groups support the organism (posture), mediate internal functions such as heartbeat, breathing, and peristalsis of the digestive system, and propel the creature through its external world. The *motor neurons* typically work in concert with interneurons and sensory mechanisms that are orchestrated together through intricate patterns of activity. The motor control pathways in whole organism actions read-out these patterns from many parts of the brain

and its contact with the sensory world. It has been argued (Lashley, 1951 Sperry 1952) that movement, and thus motor control, is the most fundamental property of nervous systems. Without movement there is no behavior as the term is normally defined. Movement also provides overt and objective measures of the rich symphonies of behavior we see in the natural world.

1. Commonalities and Specializations in Mechanism

Motor control in all vertebrate species (animals with backbones) operates from a similar ground plan, but with important differences in detail. For all vertebrates spinal and brain stem mechanisms mediate a number of basic forms of movement, such as locomotion and postural adjustment (see *Behavior, Hierarchical Organization of*; *Eye Movement, Control of (Oculomotor Control)*). In mammals, and especially in humans, higher cortical areas along with the *basal ganglia* and *cerebellum* contribute importantly to the acquisition, planning, flexibility, and fine-tuning of movement sequences (Fig. 1). Major properties of these structures are discussed below. In invertebrates (animals without backbones) the basic building blocks of motor control are diverse and distinct from those found in vertebrate species, yet there is a remarkable convergence in terms of overall solutions to motor control. For example, relatively small groups of neurons, and in some species single cells, generate rhythms of an endogenous (intrinsic to the cells) rhythms without the need for molding by afferent influences. These *central pattern generators* were first documented clearly in invertebrates, and then determined for vertebrates (Fentress 1976, Pearson 1993; see *Comparative Neuroscience*). In the intact animal these patterns of spontaneous activity are typically modulated by other neural and endocrine influences, however.

Most investigators today believe that many of the most important questions of motor control rest at the interface between intrinsic properties of neurons and their sensitivity to extrinsic factors. Motor commands may remain latent until *triggered* by descending influences, rather as dropping a needle onto a phonograph record triggers patterns of sound programmed within the chosen record. The analogy to music turns out to be a useful one, for it reminds us that we wish to account for the notes (basic units) of movement, the construction of chords of simultaneously activated motor elements, melodies in time, rhythmic principles, volume (intensity) and timbre (fine tuning), and even possible transpositions from one moment to the next. Questions of movements guided toward specific features of the environment are also of obvious importance, and are discussed in other chapters (see *Behavior, Hierarchical Organization of*; *Neural Representations of Direction (Head Direction Cells)*; *Navigation in Virtual Space: Psychological and Neural Aspects*; *Navigation in Spatial Environments, Neural Basis of*; *Neural Representations of Intended Movement in Motor Cortex*).

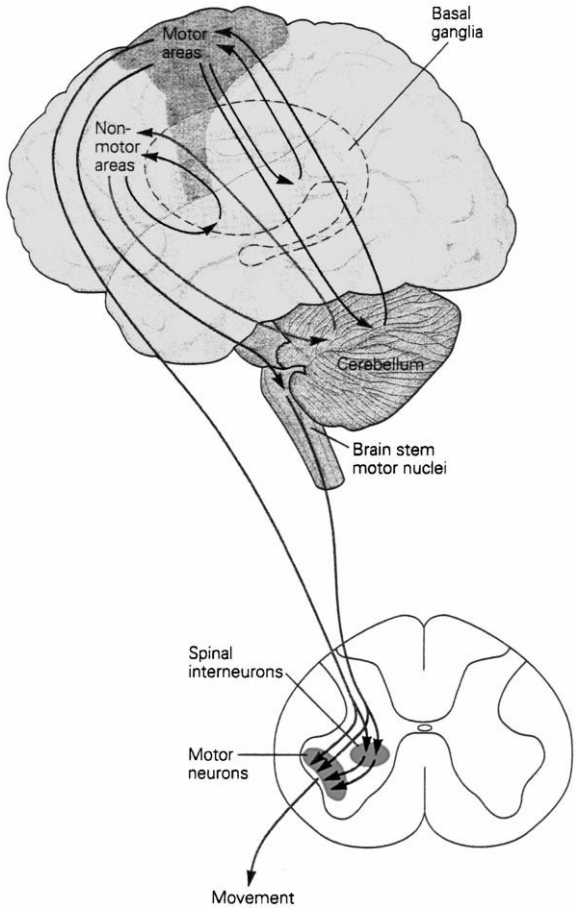

Figure 1
The motor systems have three levels of control—the spinal cord, brain stem, and forebrain—organized both serially and in parallel. Source: Kandel et al. 2000

2. Early Models and Current Perspectives

Early models of vertebrate movement stressed individual reflexes and reflex chains as the basic building blocks in the organization of action patterns (Sherrington 1906). Spinal reflexes operate through relatively simple chains of neurons, where sensory and interneurons link up with motor neurons that in turn project to a select group of muscles. Interest in reflex behavior (such as the well-known knee jerk reflex elicited by your local physician) remains intense, and reflexes are clearly essential for the production of posture, basic movements, and defensive reactions.

Their speed and reliability of production reflect what are often called 'automated' properties of motor control, that is movement profiles that are rapid, essentially ballistic, and not mediated by conscious processes. Even these reflexes can be modulated by higher neural activities, however. As in nearly all aspects of neural function it is thus a matter of looking at intrinsic circuits *together with* extrinsic influences. These occur at many levels of neural control.

2.1 Movement Units and Their Limits

One of the basic challenges in the study of motor control is the dissection of fundamental movement units. In this terminology a *unit* is a relatively invariant pattern of muscular contractions that are typically elicited together. Reference to one of the most complex forms of human movement control, speech, is illustrative (Neville 1995). When we speak we articulate a broadly invariant set of phonemes (basic sound patterns) in sequence. This allows for the articulation of speech to be relatively 'thought free' in normal discourse, and allows the listener to decode the message through a complex array of perceptual and higher-order cognitive processes (see *Syntactic Aspects of Language, Neural Basis of*; *Motor Control*; *Sign Language: Psychological and Neural Aspects*; *Lexical Processes (Word Knowledge): Psychological and Neural Aspects*; *Speech Production, Neural Basis of*). The example is useful in that in the smooth flow of speech individual phonemes are often 'co-articulated' with the fine details of one sound unit reflecting the production of preceding and subsequent sound patterns. So even here the 'intrinsic' properties of a speech unit can be sensitive to extrinsic influences as defined by other speech units. This is what gives speech its smooth and flowing nature (as opposed, for example, to most computer generated 'speech').

Thus, in some sense movement and their motor control 'units' are abstractions that frequently blur at the edges, and which can be defined from a number of complementary perspectives (Golani 1992). In complex forms of movement, such as playing a sport or musical instrument, or dancing with a moving partner, many neural pathways are orchestrated together in dazzling patterns that are widely distributed, serially ordered, and parallel in their operations. It is for this reason that the very distinction between motor control systems and other properties of the nervous system are often difficult to untangle (see *Vision for Action: Neural Mechanisms*; *Cognitive Control (Executive Functions): Role of Prefrontal Cortex*).

2.2 Movement Notation

A number of movement notation systems have helped clarify the organization of motor patterns at many different levels. For example, the Eshkol–Wachman movement notation system (Golani 1992) has helped investigators separate specific parameters of movement, such as movements in a plane (flat surface) and those that involve rotations or conical movements around a single limb. These notational systems, derived from human choreography, also remind us that movement is in its essence both relational and dynamic. Its description also depends upon the frame of reference adopted. For example, we can move our bodies with respect to internal coordinates and yet maintain a fixed relationship to external events, such as a moving social partner. Often the units of control involve multiple muscular systems, such as when we produce relatively invariant motions during the final phases of feeding or brushing our teeth via complementary adjustments of individual muscular acts (such as those involved in arm movements or turning the head).

2.3 Dynamic Systems

Many workers today have chosen to supplement the language of motor units and experimentally isolated mechanisms with insights derived from *dynamic systems theory* (Kelso 1995, Thelen 1985). These approaches remind us that behavior (movement, and thus motor control) is fundamentally imbedded in time, and involves the cooperation of a rich multitude of processes whose boundaries are often difficult to define. Further, there has been an increased appreciation for the flexibility of underlying mechanism operations. Even in invertebrates circuits that contribute to a given movement may be shared and change their internal properties from one occasion of expression to the next (Marder and Calabrese 1996). This 'remolding' of circuit properties adds importantly to the challenges of deep understanding of the processes of motor control.

3. Diversity and Plasticity in Human Movement

The human species is especially remarkable in its ability to master skills of an enormous diversity. One needs to reflect just for a moment on actions such as speech, sports, dance, driving an automobile, or musical performance to appreciate this fact. Indeed while each species is the master of its own behavioral niche, it is in humans that we find the most rich variety of forms of expression. Our nervous systems, throughout much of our lives, are remarkably *plastic*, meaning that they can be refined and remolded into an often dizzying array of specialized talents (see *Neural Plasticity in Auditory Cortex*; *Neural Plasticity*; *Dysexecutive Syndromes*; *Neural Plasticity in Visual Cortex*).

We have all had the experience that in the early learning of a new motor skill our attention is fully absorbed. Later we produce the actions in an automated manner, without any conscious awareness. When we seek to modify this skill, or apply it to a new context, the *executive* parts of our nervous system are recaptured, with thought and planning once again playing a major role. It is now well established that different neural circuits are employed in the early and late stages of skill acquisition (Posner et al. 2001).

3.1 Basic Pathways

It remains possible to separate particular neural pathways that contribute to different properties of movement. The human motor system can be visualized in part as being composed of descending pathways derived from the cerebral cortex and brainstem that are fundamental to the suprasegmental control of movement. The term 'segmental' refers to spinal segments that control different muscle groups in an orderly way. Suprasegmental thus refers to actions that incorporate muscles of different extremities such as the head, arms, trunk, and legs. The *basal ganglia* and *cerebellum* contribute to the planning, sequencing and fine-tuning of movement. The *basal ganglia* are embedded deep within the forebrain and also include midbrain pathways. These pathways do not connect directly to motor pathways, but exert important indirect influences that are mediated through cortical and thalamic loops. Disorders of the basal ganglia greatly compromise effective movement, such as seen in Parkinson's Disease with its characteristic tremor and poverty of movement, and Huntington's Disease with its characteristic choreoform (sporadic and ballistic) movements. Each of these diseases is also marked by cognitive (thought, planning) deficits as the disease progresses, highlighting once again how motor systems are intimately linked with other neural functions.

The *cerebellum* has multiple sensory and motor functions, among which the timing and fine-tuning of motor control are among the most apparent. It is now known that lesions of the lateral cerebellum also compromise the perception of temporal events (Ivry 1996), thus illustrating yet again how distinctions between motor and other neural functions can be misleading. Individual circuits, and even individual neurons, can mediate a variety of functions (see *Cerebellum: Associative Learning*; *Cerebellum: Cognitive Functions*).

The spinal cord, which receives descending inputs from the brain, is neatly organized with anatomically arranged segments influencing and controlling actions of the upper and lower body. These richly interconnected pathways work together with the neuromuscular system which is also composed of nuclei of the motor cranial nerves and the muscles that they innervate.

At every level of organization major afferent pathways project to the motor control mechanisms, but their influence is a function of such factors as speed of movement and the extent to which movement patterns have become automated (as in well practiced tasks). Often the influence of afferent pathways is relatively small in well practiced and simple automated movements, but sensitivity to sensory influences can attain major importance when actions have to be adjusted, such as in the confrontation with unexpected circumstances.

Activities from extrinsic sources are gated to varying degrees as a function of the demands placed upon the organism at the moment. Routine activities run more or less 'autonomously,' but are able to be corrected if circumstances change. In this way the production of movement sequences can be both automated and responsive to novel circumstances. There are many known human motor diseases that can selectively affect both the automated performance of movement and the ability of the individual to adjust to novel situations (Linas 2001). A skillful neurologist can, by examining detailed properties of movement performance, often pinpoint parts of the nervous system that are damaged.

Much of what we call voluntary movement (as opposed, to the elicitation of simple reflexes) is mediated by connections that in a sense 'wander' through a complex neuronal maze from the cerebral cortex to the spinal cord. Thus, *the primary motor cortex*, which is somatotopically organized as is the somatic sensory cortex (meaning that different body parts are connected to specific neural regions), influences specific muscle groups. The cortex itself is layered, with neurons forming groups that are rather like a series of pies stacked upon one another. It is in what anatomists call layer V (counting from the top of the cortex) of the primary motor cortex that one finds projections directly to motor neurons and interneurons, in the ventral ('bottom') level of the spinal cord, via what is called the corticospinal tract. In humans there are approximately one million axons (those processes of nerves that carry signals out from the neuron's cell body) and just under 50 percent of these originate from the motor cortex. Other parts of the cortex, such as the premotor areas and the supplementary motor areas, also exert important and distinct influences upon our movement skills. Thus our cortices, the highest parts of the human brain, connect us more or less directly to the world through movement.

From the cortex the neural axons pass through white matter of the sub-cortex, through the internal capsule and cerebral peduncles on their way to the spinal cord, and hence the muscles that make movement possible (see Fig. 1). The term 'white matter' refers to the fact that these neurons have a fatty insulating sheath that promotes rapid and reliable transmission of signals from one neuron to another.

These tracts in turn form the medullary pyramids deep within the brain (a region called the medulla). The projection encompasses what has classically been called the pyramidal tract. In addition to the pyramidal tract there is the extrapyramidal system, which is composed of other nuclei, such as those found in the basal ganglia. These extrapyramidal systems are for the most part more indirect in their motor control functions than are the pyramidal pathways, but their importance should not be underestimated. For example, linkages between emotional and cognitive processes and the motor system are largely mediated through these extrapyramidal systems.

The left side of the brain primarily is involved with movements on the right side of the body, and vice versa, but the crossing of fibers is incomplete. Ultimately, the coordination of movements from each side of the body involves unified actions of neural circuits that are in themselves primarily concerned with one body side.

The corticospinal tract itself makes monosynaptic (i.e., direct) connections with motor neurons, which in turn control muscular contractions. There are also many connections to interneurons in the spinal cord. The importance of these indirect connections is that they allow a number of influences from other brain regions that are essential for adaptive motor performance. Throughout, the activities of the corticospinal tract and related pathways incorporate a rich variety of descending modulating influences.

3.2 Movement Disorders and Development

In addition to the often tragic consequences of neural damage (e.g., stroke) the neurologist and behavioral neurobiologist can gain important insights into the foundations of motor control by examining the developmental progression of motor skills. There are predictable milestones in the construction of human motor skills. A relatively recent perspective has stressed the theme of 'ontogenetic adaptations' in motor control (Oppenheim 1982). In the past it was customary to think of nervous systems starting out as 'imperfect' adult systems—rather like a linear progression from less than fully adequate to fully adequate. This view has now changed. Think, for example, of the newborn infant who must swallow (which involves a complex coordination of 20 plus muscles in the throat; Doty 1976), and breathe and suckle. There is no room for error or practice. Indeed, many motor circuits are formed in a *prefunctional* manner. This means that the circuits are established without the benefit of overt motor practice. At a more detailed neuronal level these circuits develop in what has been called a *retrograde* manner, with the establishment of final neuromuscular and other motor circuits preceding the establishment of higher-order interneuronal influences and thus responses to sensory events from the internal and external environments (Fentress and Gadbois 2001).

This reminds us of the work of ethologists (e.g., Hinde 1970, Lorenz 1950, Tinbergen 1963) who stressed the 'instinctive' nature of many actions in animals and humans. The basic idea is that many of the fundamental properties in animal and human behavior are a relatively direct reflection of our genetic make-up. People, for example, speak; other species do not. Many basic human and animal capacities for movement emerge relatively independently of specialized experiences. In some deep sense we are indeed 'preprogrammed' to be human, just as the bird, and bat, and insect are preprogrammed to be members of their own species. What is perhaps most remarkable about humans is that our genetic structure allows for enormous flexibility in the detailed organization of our movements. In that respect we stand alone among the animal species. Yet, for all animal species the delicate interplay between genetic and experiential factors (broadly defined, across many levels of organization) remains a major area of importance for further investigation. An essential point is that even the most basic processes of differentiation in development involve genes and experience working in concert (Hinde 1970, Edelman 1987).

How is it that we learn motor skills, and how are these skills mediated? At the present time our knowledge of these essential questions is very limited. Certainly many of the mechanisms we have touched upon contribute richly. Thus, cerebellar circuits are important in the establishment of newly acquired reflexes as well as more complex cognitive skills (Thompson, Thach). The basal ganglia as well as cortical mechanisms also contribute to perfections of motor performance through experience (Graybiel). The importance of cortical mechanisms is perhaps predominant, but a challenge for future research is to determine how various neural mechanisms operate together at various stages of skill acquisition and performance (see *Neural Plasticity in Auditory Cortex*; *Neural Plasticity*; *Neural Plasticity of Spinal Reflexes*; *Neural Plasticity in Visual Cortex*).

4. The Need to Combine Mechanistic and Systems Perspectives

A challenge of increasing importance in both the neural and behavioral sciences is how to capture the dynamic as well as modular properties of motor performance. Excellent overviews that cross-complementary levels of organization can be found in Gazzinga 2000, Kandel et al. 2000. Most investigators in the neurobehavioral sciences today take the position that any level of analysis must make contacts with levels that are both more molar and more molecular.

5. Challenges for Future Research

Human motor control systems are awesome in their richness. It is through these motor control systems that we interact as individuals, yet individuals deeply connected to other human individuals and to our animal brethren by solutions that evolution has enabled us with. We talk, sing, and dance. In spite of recent progress at many levels of research, precisely how we do these basic things remains deeply mysterious. Yet, piece by piece, we are beginning to understand the operation of individual mechanisms and their coordinated actions. For example, the rapid progress in brain imaging techniques has helped us understand how the final processes of motor control are connected to higher-order cognitive functions (see *Vision for Action: Neural Mechanisms*; *Cognitive Control (Executive Functions): Role of Prefrontal Cortex*; *Functional Brain Imaging*; *Neural Representations of Intended Movement in Motor Cortex*). The further understanding of motor control processes, including not just their performance on a short-term basis, but also their development and evolution, offers riches into the nature of the human psyche itself.

In recent years a number of model systems for motor control have been offered (see *Computational Neuroscience*). One of the most promising is that there are variable foci of activity that occur in conjunction with variable expanses of collateral inhibition, that is, those events that block competing activity (Fentress 1976). For example, both the cerebellum and basal ganglia have lateral inhibitory (collateral) circuits that suppress action properties that would otherwise interfere with the primary focus of adaptive movement. There is now growing evidence that these collateral pathways can expand in their extent of influence as the core of motor performance becomes more tightly focused (Fentress 1976). What this means is that systems in the control of movement can be both broadly and much more finely tuned as circumstances dictate. Extrinsic influences are either accepted or rejected as the needs of the organism change.

In terms of the establishment of motor behavior during development, a number of genetic analysis that focus upon specific neural circuits are now being combined with experiments designed to clarify the roles and limits of specific sources of ontogenetic experience. The old 'nature vs. nurture' debate is now considered *passé* by most active investigators; the challenge is to see how genes and experience work as partners during the construction of brains and their behavioral manifestations.

In motor control we have much to learn about the detailed operation of specific mechanisms, but equally much to learn how these specific mechanisms become established and integrated into the rich symphonies of human and animal action. Whatever the details that will be revealed by future research, it is clear that motor control mechanisms not only reflect what we mean by behavior, but also give us insights into the inner workings of nervous systems that remain a major source of inspiration and mystery for all those who are fortunate enough to study them.

Major components of the human motor system. Spinal cord, brain stem, and motor areas of the cerebral cortex work together with direct and indirect connections of the basal ganglia and cerebellum in the planning and execution of integrated movement patterns.

See also: Basal Ganglia; Behavior, Hierarchical Organization of; Cerebellum: Associative Learning; Cerebellum: Cognitive Functions; Eye Movement, Control of (Oculomotor Control); Motor Control Models: Learning and Performance; Motor Cortex; Neural Representations of Intended Movement in Motor Cortex; Neuromuscular System; Short-term Memory, Cognitive Psychology of

Bibliography

Doty R W 1976 The concept of neural centers. In: Fentress J C (ed.) *Simpler Networks and Behavior*. Sinauer Associates, Sunderland, MA, pp. 251–65

Edelman G M 1987 *Neural Darwinism: The Theory of Group Selection*. Basic Books, New York

Fentress J C 1976 Dynamic boundaries of patterned behavior: Interaction and self-organization. In: Bateson P P G, Hinde R A (eds.) *Growing Points in Ethology*. Cambridge University Press, Cambridge, UK, pp. 135–69

Fentress J C, Gadbois S 2001 The developments of action sequences. In: Blass E (ed.) *Developmental Psychobiology*. Kluwer/Plenum, New York, pp. 403–41

Gazzaniga M S (ed.) 2000 *The New Cognitive Neurosciences*, 2nd edn.

Golani I 1992 A mobility gradient in the organization of vertebrate movement: The perception of movement through symbolic language. *Behavioral and Brain Sciences* **15**: 249–66

Graybiel A M 1995 Building action repertoires: Memory and learning functions of the basal ganglia. *Current Opinion in Neurobiology* **5**: 733–41

Hinde R A 1970 *Animal Behavior*, 2nd edn. McGraw-Hill, New York

Ivry R B 1996 The representation of temporal information in perception and motor control. *Current Opinion in Neurobiology* **6**: 851–7

Kandel E R, Schwartz J H, Jessell T M 2000 *Principles of Neural Science*, 4th edn. McGraw-Hill, New York

Kelso J A S 1995 *Dynamic Patterns: The Self-organization of Brain and Behavior*. MIT Press, Cambridge, MA

Lashley K S 1951 The problem of serial order in behavior. In: Jeffries L A (ed.) *Cerebral Mechanisms in Behavior*. John Wiley and Son, New York, pp. 112–36

Llinas R R 2001 *I of the Vortex: From Neurons to Self*. MIT Press, Cambridge, MA

Lorenz K Z 1950 The comparative method in studying innate behavior patterns. *Symposia of the Society for Experimental Biology* **4**: 221–68

Marder E, Calabrese R L 1996 Principles of rhythmic motor pattern generation. *Physiological Reviews* **76**: 687–717

Neville H J 1995 Developmental specificity in neurocognitive development in humans. In: Gazzinga M (ed.) *The Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 219–31
Oppenheim R W 1981 Preformation and epigenesis in the origins of the nervous system and behavior: Issues, concepts and their history. *Perspectives in Ethology* **5**: 1–100
Pearson K G 1993 Common principles of motor control in vertebrates and invertebrates. *Annual Review of Neuroscience* **16**: 265–97
Posner M I, Rothbart M K, Farah M, Brucer J (eds.) 2001 The developing human brain. *Developmental Science* **4**: 253–387
Sperry R 1952 Neurology and the mind–brain problem. *American Scientist* **40**: 291–312
Thach W T 1996 On the specific role of the cerebellum in motor learning and cognition: Clues from PER activation and lesion studies in man. *Behavioral and Brain Sciences* **19**: 411–31
Thelen E 1985 Developmental origins of motor coordination: Leg movements in human infants. *Developmental Psychobiology* **18**: 1–22
Tinbergen N 1963 On aims and methods of ethology. *Zeitschrift für Tierpsychologie* **20**: 410–33

J. C. Fentress

Motor Control Models: Learning and Performance

Modeling the way in which humans learn to coordinate their movements in daily life or in more demanding activities is an important scientific topic from many points of view, such as medical, psychological, kinesiological, cybernetic. The article analyzes the complexity of this problem and reviews the variety of experimental and theoretical techniques that have been developed for this purpose.

With the advent of technical means for capturing motion sequences and the pioneering work of Marey and Muybridge in this area, the attempt at describing, modeling, and understanding the organization of movement has become a scientific topic. The fact that human movements are part of everyday life paradoxically hides their intrinsic complexity and justifies initial expectations that complete knowledge could be achieved simply by improving the measurement techniques and carrying out a few carefully designed experiments. Unfortunately, this is not the case. Each experiment is frequently the source of more questions than answers and thus the attempt to capture the complexity of purposive action and adaptive behavior, after a century of extensive multidisciplinary research, is far from over.

The conventional view is based on a separation of perception, movement, and cognition and the segregation of perceptual, motor, and cognitive processes in different parts of the brain, according to some kind of hierarchical organization. This view is rooted in the empirical findings of neurologists of the nineteenth century, such as J. Hughlings Jackson, and has a surprising degree of analogy with the basic structure of a modern PC that typically consists of input and output peripherals connected to a central processor. Perhaps the analogy with modern technology justifies why this old-fashioned attitude still has its supporters, in spite of the massive empirical and conceptual challenge to this view and its inability to explain the range of skills and adaptive behaviors that characterize biological organisms.

Let us consider perception, which is the process whereby sensory stimulation is translated into organized experience. That experience, or percept, is the joint product of the stimulation and of the process itself, particularly in the perception and representation of space. An early theory of space perception put forth by the Anglican bishop G. Berkeley at the beginning of the eighteenth century was that the third dimension (depth) cannot be directly perceived in a visual way since the retinal image of any object is two-dimensional, as in a painting. He held that the ability to have visual experiences of depth is not inborn but can only result from logical deduction based on empirical learning through the use of other senses.

The first part of the reasoning (the need of a symbolic deductive system for compensating the fallacy of the senses) is clearly wrong and the roots of such misconception can be traced back to the neoplatonic ideas of the Italian Renaissance, in general, and to Alberti's window metaphor, in particular. Also, the Cartesian dualism between body and mind is just another face of the same attitude and such *Descartes' error*, to quote Damasio (1994), is on a par with the *Berkeley's error* above and is at the basis of the intellectualistic effort to explain the computational complexity of perception which characterizes a great part of the classic artificial intelligence approach. However, the latter part of Berkeley's conjecture (the emphasis on learning and intersensory integration) is surprisingly modern and agrees, on one hand, with the modern approach to neuropsychological development pioneered by Piaget and, on another, with the so-called *connectionist* point of view, originated in the 1980s as a computational alternative to classic artificial intelligence (see *Artificial Neural Networks: Neurocomputation*).

An emergent idea is also the motor theory of perception, well illustrated by Berthoz (1997); that is, the concept that perception is not a passive mechanism for receiving and interpreting sensory data but is the active process of anticipating the sensory consequences of an action and thereby binding the sensory and motor patterns in a coherent framework. In computational terms, this implies the existence in the brain of some kind of 'internal model,' as a bridge between action and perception. As a matter of fact, the idea that the instructions generated by the brain for controlling a movement are utilized by the brain for interpreting the sensory consequences of the movement is already present in the pioneering work of Helmholtz and von Uexküll and its influence has

resurfaced in the context of recent control models based on learning (e.g., Wolpert and Kawato 1998). The generally used term is 'corollary discharge' (von Holst and Mittelstaedt 1950) and implies an internal comparison between an outgoing signal (the *efferent copy*) and the corresponding sensory *re-afference*: the coherence of the two representations is the basis for the stability of our sensorimotor world. This kind of circularity and complementarity between sensory and motor patterns is obviously incompatible with the conventional reasoning based on hierarchical structures. A similar kind of circularity is also implicit in Piaget's concept of 'circular reaction,' which is assumed to characterize the process of sensorimotor learning; that is, the construction of the internal maps between perceptually identified targets and the corresponding sequence of motor commands.

An additional type of circularity in the organism/environment interaction can be identified at the mechanical interface between the body and the outside world, where the mechanical properties of muscles interact with the physics of inanimate objects. This topic area has evolved from the Russian school, with the early work on the nature of reflexes by I. P. Pavlov and the subsequent critical re-examination by P. K. Anhokin and N. Bernstein. In particular, we owe to Bernstein the seminal observation (the *comparator model*) that motor commands alone are insufficient to determine movement but only identify some factors in a complex equation where the world dynamics has a major influence. This led, among other things, to the identification of muscle stiffness as a relevant motor parameter and the formulation of the theory of equilibrium-point control (Feldman and Levin 1995, Bizzi et al. 1992).

In general, we may say that, in different ways, Helmholtz's corollary discharge, Piaget's circular reaction, and Bernstein's comparator model are different ways to express the *ecological nature* of motor control; that is, the *partnership* between brain processes (including muscles) and world dynamics. This concept is graphically sketched in Fig. 1. On the other hand, these general ideas on motor control could not provide, immediately, mathematical tools of analysis from which to build models and perform simulations. The art and science of building motor control models is a later development and has been influenced by the methods designed by engineers in the field of automatic control and computer science. Most of the techniques are based on linear approximations and explicit schematizations of the phenomena. In particular, two main concepts can be singled out for their influence on the study of motor control: the concept of feedback and the concept of motor program. Their level of influence in brain theory is certainly determined by the tremendous success of these techniques in the modern technological world. However, their applicability to what we may call the *biological hardware* is questionable for two main reasons: (a) Feedback control can only be effective and stable if the feedback delays are negligible, but this is not the case for biological feedback signals, where transduction and transmission delays add up to tens of milliseconds; (b) The concept of motor program implies a sequential organization, which only can be effective if the individual steps in the sequence are sufficiently fast, and this is in contrast with the parallel, distributed processing of the brain made necessary by the relative slowness of synaptic processing. In fact, the recognition of the limits of the analytic–symbolic approach has motivated, since the late 1980s, a re-evaluation of earlier approaches under the new light of connectionist thinking.

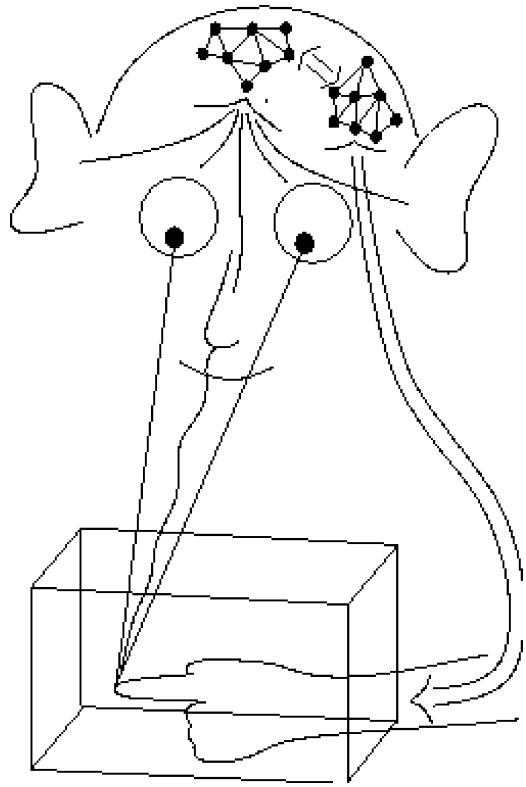

Figure 1
Ecological nature of the sensorimotor control system

1. Trajectory Formation

The computational process that is necessary for realizing a planned motor task has been the subject of a great deal of research in robotics. As sketched in Fig. 2, five main subproblems can be identified: (a) planning, (b) trajectory formation, (c) inverse kinematics, (d) inverse dynamics, and (e) actuator control. In the robotic approach, the blocks correspond to different procedures to be executed in a sequence. However, this

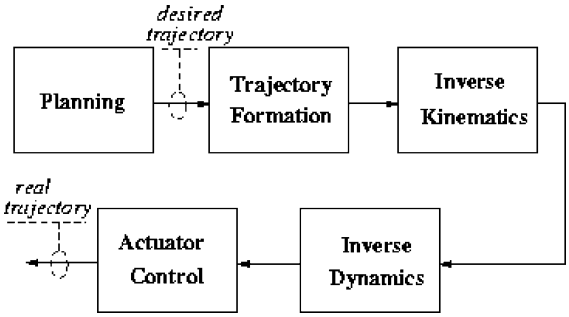

Figure 2
Computational steps in robotic control models

is only an implementation strategy, quite natural for the usual engineering hardware, but probably inappropriate for the biological hardware. The first two processes correspond to the intuitive idea of imagining and then drawing a figure on a piece of paper. This implies selecting the initial point, the size and orientation of the figure, and then breaking down the complex trajectory of the pen that realizes such intention into a sequence of pen strokes, smoothly joined together in a quasi-continuous curve. However, for the pen to faithfully follow the intended path, the brain must first figure out the patterns of angular rotations of the different joints, the so-called inverse kinematic problem. This is a difficult computational task because the human arm (and the body in general) has a high degree of 'kinematic redundancy,' in the sense that there is an excess number of degrees of freedom and thus the same pen stroke can be realized by means of an infinite number of joint rotation patterns. The biological solution to this problem is characterized by an invariant spatiotemporal structure whose main feature is that the elementary strokes are approximately straight, with a symmetric *bell-shaped speed profile* (Morasso 1981), and whose global smoothness is well approximated by a criterion of 'minimum jerk' (Flash and Hogan 1985). In any case, the computations described above only cover the first three blocks of Fig. 2. The real challenge is in the inverse dynamics block and there are reasons to assume that the controllable compliant properties of muscles can relieve the brain of at least part of the computational burden, which must be explicitly shouldered by conventional robotic systems because torque motors have zero stiffness.

2. *The Theory of Equilibrium-point Control*

This theory is dependent on the elastic properties of muscles, that are well captured by the so called λ-model (Feldman and Levin 1995). In this model, λ is the controllable parameter that sets the activation

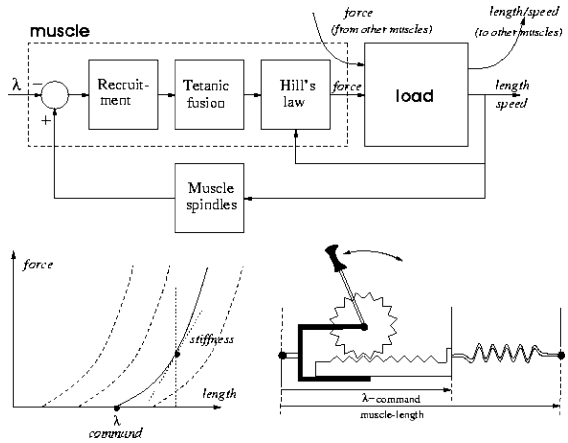

Figure 3
The λ-model of muscle control. Stiffness is the slope of the length–tension curve. Mechanical feedback (Hill's law) is instantaneous; spinal feedback (through muscle spindles) has a significant delay

threshold of the monosynaptic stretch reflex and thus determines the 'rest length' of the 'muscle spring' (Fig. 3). Its value is specified by supraspinal motor commands and expresses a combination of the desired levels of muscle length and stiffness. By setting the λ-commands of all the muscles, the brain implicitly codes an equilibrium point, determined by the fact that at this point the spring actions cancel each other. In this way, movement follows as a mechanical consequence of the force field, without a continuous intervention of the brain. In Fig. 3, the 'Recruitment' block corresponds to the exponential family of curves (shifted to the left according to the rack and pinion mechanism, operated by the λ-command); 'Tetanic fusion' refers to the relatively slow force build-up for sudden changes of motor commands; 'Hill's law' is the nonlinear dependence of muscle's force on the muscle's speed of contraction. Moreover, the brain can take advantage of a second type of redundancy, the 'muscle redundancy': the excess number of muscles in relation to the number of degrees of freedom. The important feature is that muscles are not linear springs but are characterized by length–tension curves of exponential type, whose stiffness depends on the difference between the λ-command and the actual muscle length. Therefore, the brain has the opportunity of independently controlling two variables: (i) the global equilibrium point, by setting the λ-commands equal to the muscle lengths of a desired posture (the reciprocal λ-commands); (ii) the global stiffness of the joints, by adding to the previous pattern a set of coactivation λ-commands (the stronger the coactivation, the stronger the stiffness).

The implications of this model are extremely relevant but the interpretation of the experimental data is

still controversial. There is no doubt that muscle stiffness can be seen as a kind of implicit feedback mechanism that tends to overcome the action of external and internal disturbances and loads, such as the action of gravity and the intrinsic dynamics of the body masses. The big question is concerned with the quantitative and functional relevance of this effect. For some researchers, muscle stiffness is all that is needed, without any kind of internal dynamic model. In this view, the brain is only supposed to generate equilibrium-point trajectories (the reciprocal commands) and to set up an appropriate level of co-activation. A very important feature of this model is that it assigns a computational role to the muscles, in addition to its obvious executive action. The defenders of this view also point out that an explicit feedback control action, supposedly mediated by proprioceptive signals and the corresponding reflex mechanisms, is ruled out by the large delays in the feedback loops, which are known to be of the order of tens of milliseconds and thus sufficient to cause a feedback control solution to be unstable. In spite of its elegance and appealing 'ecological' nature, this extreme form of equilibrium-point control model would only be plausible if the empirical values of muscle stiffness, at equilibrium as well as during movement, were strong enough in relation to the usual dynamics of body motion. The problem is that this is a difficult type of measurement and there is not yet a complete agreement on the available data; however, it is fair to say that, although stiffness is certainly relevant as a co-factor, its natural level is probably not enough to fully counteract the body dynamics, at least in more demanding dynamic tasks. For example, its role is likely to be much greater in the case of handwriting movements, which involve relatively small masses, than in the case of sway movements in upright standing, which are affected by the whole body mass.

3. Dynamic Compensation

The alternative solution is some combination of 'feedforward' and (explicit) 'feedback' control, in addition to the implicit feedback provided by muscle stiffness. In general, a feedforward control model is based on the anticipatory computation of the disturbances that will be encountered by a system when it attempts to carry out a desired plan of motion. This computation is complex and requires learning; a good example is given by the 'feedback error learning' model (Kawato and Gomi 1992) where the feedforward block is trained according to the residual errors of an underlying feedback controller. Figure 4 shows a modified version of the model that takes into account the intrinsic feedback due to the elastic muscle properties, ignored in Kawato's original model. In the figure, the feedback error (the discrepancy between the desired and real trajectories) is used as the learning

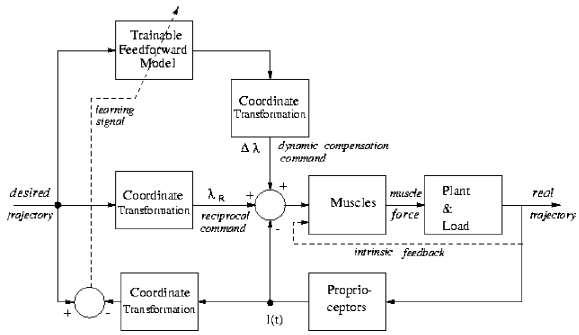

Figure 4
Feedback error learning

signal of the trainable feedforward model, which gradually takes over the responsibility of counter-balancing the dynamic disturbances, thus acquiring an internal model of body dynamics that generates an additional type of motor command: the *dynamic compensation* λ-command, to be added to the *reciprocal* and *coactivation* commands defined above (Morasso and Sanguineti 1997).

This example introduces the next big issue for understanding the trend in motor control modeling, the issue of motor learning, which has been deeply influenced by the advent in the early 1980s of connectionist theories and techniques.

4. Learning Paradigms in Neural Networks

At the core of the theories of neural network models is the attempt to capture general approaches for learning from experience tasks that are too complex to be expressed by means of explicit or symbolic models (Arbib 1995). The mechanism of learning and memory has been an intriguing question after the establishment of the neuron theory at the turn of the nineteenth century (Ramón y Cajal 1928) and the ensuing conjectures that memories are encoded at synaptic sites (Hebb 1949) as a consequence of a process of learning. In accordance with this prediction, synaptic plasticity was first discovered in the hippocampus and nowadays it is generally thought that LPT (long-term potentiation) is the basis of cognitive learning and memory, although the specific mechanisms are still a matter of investigation.

Three main paradigms for training the parameters or *synaptic weights* of neural network models have been identified: (a) 'Supervised learning,' in which a teacher or supervisor provides a detailed description of the desired response for any given stimulus and exploits the mismatch between the computed and the desired response or error signal for modifying the synaptic weights according to an iterative procedure. The mathematical technique typically used in this type of learning is known as *back propagation* and is based

on a gradient–descent mechanism that attempts to minimize the average output error; (b) 'Reinforcement learning,' which also assumes the presence of a 'supervisor' or teacher but its intervention is only supposed to reward (or punish) the degree of success of a given control pattern, without any detailed input–output instruction. The underlying mathematical formulation is aimed at the maximization of the accumulated reward during the learning period; (c) 'Unsupervised learning,' in which there is no teacher or explicit instruction and the network is only supposed to capture the statistical structure of the input stimuli in order to build a consistent but concise internal representation of the input. The typical learning strategy is called 'Hebbian,' in recognition of the pioneering work of D. O. Hebb, and is based on a 'competitive' or 'self-organizing' mechanism that uses the local correlation in the activity of adjacent neurons and aims at the maximization of the mutual information between stimuli and internal patterns.

5. *Adaptive Behavior and Motor Learning*

The neural machinery for learning and producing adaptive behaviors in vertebrates is sketched in Fig. 5, which emphasizes the recurrent, nonhierarchical flow of information between the cerebral cortex, the basal ganglia/thalamus, and the cerebellum. A growing body of evidence has been accumulated in recent years that challenges the conventional view of segregated processing of perceptual, motor, and cognitive information (Shepherd 1998). For example, it was usually considered that basal ganglia and cerebellum were specialized for motor control and different cortical areas were devoted to specific functionalities, with a clear separation of sensory, motor, and cognitive areas.

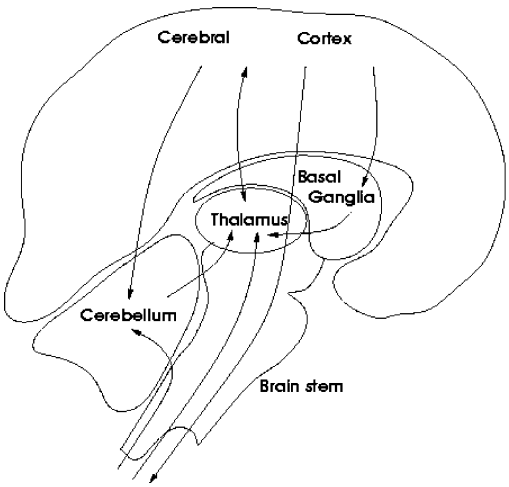

Figure 5
Information flows in motor control and learning

This is not now the conventional wisdom and the emerging picture is that the three main computational sites for adaptive behavior are all concerned with processing sensorimotor patterns in a cognitive–sensitive way but are specialized as regards the learning paradigms and the types of representation: (a) The cerebral cortex appears to be characterized by a process of unsupervised learning that affects its basic computational modules (the micro-columns that are known to have massive recurrent connections). The function of these computations might be the representation of non-linear manifolds, such as a body schema in the posterior parietal cortex (Morasso and Sanguineti 1997). This view seems to be contradicted by the 'fractured shape' of cortical maps (Rizzolatti et al. 1998), but the apparent contradiction may only be a side effect of the basic paradox faced by the cortical areas: how to fit higher dimensional manifolds (such as a proprioceptive body schema) on to a physically flat surface; (b) The cerebellum is plausibly specialized in the kind of supervised learning exemplified by the feedback error learning model. Moreover, the cerebellar hardware (characterized by a large number of microzones, comprising mossy-fiber input, bundles of parallel fibers, output Purkinje cells, with teaching signals via climbing fibers) is well designed for the representation of time series, according to a sequence-in sequence-out type of operation (Braitenberg et al. 1997); (c) The basal ganglia are known to be involved in events of reinforcement learning that are required for the representation of goal-directed sequential behavior (Sutton and Barto 1998).

6. *A Distributed Computational Architecture*

Figure 6 summarizes many of the points outlined above and outlines in which direction we may expect future research to be directed. It must be emphasized that in spite of its apparent simplicity the sketched model is extremely complex from many points of view: (a) it is nonlinear; (b) it involves high-dimensional variables; (c) it has a coupled dynamics, with internal and external processes; (d) it is adaptive, with concurrent

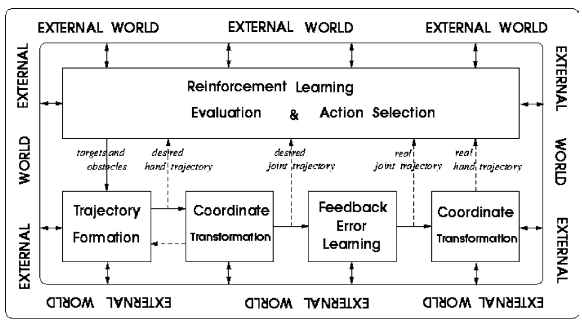

Figure 6
A distributed motor control computational architecture

learning processes of different types. No simulation model of this complexity has been constructed so far, partly because the mathematical tools for its design are only partially available.

However, there is a need for improving our current level of understanding in this direction because this is the only sensible way for interpreting the exponentially growing mass of data coming from new measurement techniques, such as advanced brain imaging. As a matter of fact, since the time of Marey, better measurement techniques of movement analysis have required better and better models of motor control, and vice versa.

In the figure, the reader may recognize the presence of computational modules already considered: 'trajectory formation' and 'feedback error learning.' The latter, in particular, obviously is characterized by a supervised learning paradigm and thus we may think that its main element (the 'trainable feedforward model') is implemented in the cerebellar circuitry. The learning signal, in this case, is the discrepancy between the desired trajectory (the motor intention) and the actual trajectory, determined by the combined body–environment dynamics and measured by different proprioceptive channels. In a sense, the brain acts as its own supervisor, setting its detailed goal and measuring the corresponding performance: for this reason it is possible to speak of a *self-supervised paradigm*.

The underlying behavioral strategy is an active exploration of the 'space of movements' also known as 'babbling,' in which the brain attempts to carry out randomly selected movements that become the teachers of themselves.

On the other hand, the trajectory formation model cannot be analyzed in the same manner. It requires different *maps* for representing task-relevant variables, such as the position of the objects/obstacles in the environment, the position of the body with respect to the environment, and the relative position of the body parts. Most of these variables are not directly detectable by means of specific sensory channels but require a complex process of 'sensory fusion' and 'dimensionality reduction.' This kind of processing is characteristic of associative cortical areas, such as the posterior parietal cortex which is supposed to hold maps of the body schema and the external world (Paillard 1993) as a result of the converging information from different sensory channels. The process of cortical map formation can be modeled by competitive Hebbian learning applied both to the thalamo-cortical and cortico-cortical connections: the former connections determine the receptive fields of the cortical units whereas the latter support the formation of a kind of high-dimensional grid that matches the dimensionality of the represented sensorimotor manifold. In a cortical map model, sensorimotor variables are represented by means of 'population codes' which change over time as a result of the map dynamics. For example, the 'Trajectory formation' module of Fig. 6 can be realized by means of a cortical map representation of the external space that can generate a time varying population code corresponding to the 'desired hand trajectory.' Another map can transform the 'desired hand trajectory' into the corresponding 'desired joint trajectory,' thus implementing a transformation of coordinates from the 'hand space' to the 'joint space.' This kind of distributed architecture is necessary for integrating multisensory redundant information into a task-relevant, lower-dimensional representation of sensorimotor spaces. On top of this computational layer, that operates in a continuous way, there is a layer of reinforcement learning that operates mostly by trial and error through two main operating modules: an 'actor' that selects a sequence of actions and a 'critic' that evaluates the reward and influences the action selection of the next trial. In the model of Fig. 6 it is the reinforcement learning module that decides the allocation of 'targets and obstacles' for the next trial.

See also: Artificial Neural Networks: Neurocomputation; Cerebral Cortex: Organization and Function; Classical Mechanics and Motor Control; Motor Control; Motor Cortex; Motor Skills, Psychology of; Perception and Action; Self-organizing Dynamical Systems

Bibliography

Anokhin P K 1974 *Biology and Neurophysiology of Conditioned Reflexes and their Role in Adaptive Behaviour*. Pergamon Press, Oxford, UK

Arbib M A 1995 *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA

Bernstein N A 1957 *The Coordination and Regulation of Movement*. Pergamon Press, Oxford, UK

Berthoz A 1997 *Le sens du mouvement*. Édition Odile Jacob, Paris

Bizzi E, Hogan N, Mussa Ivaldi F A, Giszter S F 1992 Does the nervous system use equilibrium-point control to guide single and multiple movements? *Behavioral and Brain Sciences* **15**: 603–13

Braitenberg V, Heck D, Sultan F 1997 The detection and generation of sequences as a key to cerebellar function: Experiments and theory. *Behavioral and Brain Sciences* **20**: 229–45

Damasio A R 1994 *Descartes' Error. Emotion, Reason and the Human Brain*. Putnam Press, New York

Feldman A G, Levin M F 1995 The origin and use of positional frames of references in motor control. *Behavioral and Brain Sciences* **18**: 723–45

Flash T, Hogan N 1985 The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience* **7**: 1688–1703

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

von Holst E, Mittelstaedt H 1950 Das Reafferenzprinzip. Wechselwirkungen zwischen Zentralnervensystem und Peripherie. *Naturwissenschaften* **37**: 464–76
Jeannerod M 1988 *The Neural and Behavioral Organisation of Goal-directed Arm Movements*. Clarendon Press, Oxford, UK
Kawato M, Gomi H 1992 A computational model of four regions of the cerebellum based on feedback error learning. *Biological Cybernetics* **69**: 95–103
Marey E J 1894 *Le mouvement*. Édition Masson, Paris
Morasso P 1981 Spatial control of arm movements. *Experimental Brain Research* **42**: 223–7
Morasso P, Sanguineti V 1997 *Self-organization, Cortical Maps and Motor Control*. North Holland, Amsterdam
Muybridge E 1957 *The Human Figure in Motion*. Dover Press, New York
Paillard J 1993 *Brain and Space*. Oxford University Press, Oxford, UK
Piaget J 1963 *The Origin of Intelligence in Children*. Norton Press, New York
Ramón y Cajal S 1928 *Regeneration in the Vertebrate Central Nervous System*. Oxford University Press, Oxford, UK
Rizzolatti G, Luppino G, Matelli M 1998 The organization of the cortical motor system: New concepts. *Electrencephalography and Clinical Neurophysiology*. **106**: 283–96
Shepherd G M 1998 *The Synaptic Organization of the Brain*. Oxford University Press, Oxford, UK
Sutton R S, Barto A G 1998 *Reinforcement Learning*. MIT Press, Cambridge, MA
Wolpert D M, Kawato M 1998 Internal models of the cerebellum. *Trends in Cognitive Science* **2**: 338–47

P. G. Morasso

Motor Cortex

The term motor cortex has both a broad and a narrow sense. The broad meaning refers to a number of functionally distinct cortical areas, which collectively control voluntary action (Wise 1999). The narrow sense, adopted here, pertains exclusively to one of those areas, the *primary* motor cortex (Fig. 1). The primary motor cortex comprises only a small proportion of the cerebral cortex but plays a crucial role in controlling movement. Some of its neurons have axons that terminate directly on motor neurons in the spinal cord, giving motor cortex the most direct possible access to muscles (Porter and Lemon 1993).

1. *History of Motor Cortex*

The motor cortex has a short history by evolutionary standards. Before the advent of mammals, about 200 million years ago, the motor cortex did not exist. Only mammals evolved a motor cortex in the sense used here. Accordingly, it probably functions to supplement the remainder of the vertebrate motor system (see *Cerebellum*: *Cognitive Functions*; *Motor Control*), which evolved more than 500 million years ago.

In contrast to its relatively short evolutionary history, on a geological timescale, the motor cortex has a long intellectual history by the standards of neurobiology. In the fourth and fifth centuries BC, Hippocratic physicians observed that damage to one side of the brain caused motor disabilities on the opposite side of the body (Lassek 1954). In the nineteenth century, G. Theodor Fritsch and Eduard Hitzig in Germany and David Ferrier in England confirmed experimentally the existence of a localized motor cortex. They showed that electrical stimulation of the motor cortex in dogs and monkeys produced movements on the opposite side of the body. Although condemned by antivivisectionists of the time as inapplicable to humans, this late nineteenth-century animal research established the principles of cortical organization for all mammals, including our species.

2. *Description of Motor Cortex*

Like other areas of the cerebral cortex, the motor cortex consists of several cell layers. However, the motor cortex exceeds other areas in thickness. Four other characteristics collectively identify the motor cortex: electrically stimulating it evokes movements more easily than for other cortical areas; it lacks a layer common to most cortical areas (layer 4); it receives relatively direct inputs from the basal ganglia and cerebellum; and it lies immediately in front of the somatosensory cortex. The neurons that send axons directly to the spinal cord all arise from layer 5, including some of the largest neurons in the cerebral cortex. Named after Vladimir Betz, these giant neurons can reach 60 μm in cell-body width in monkeys and 120 μm in people. Cells in the motor cortex send outputs to and receive inputs from somatosensory and premotor cortex. Subcortical targets include the basal ganglia, cerebellum, red nucleus, and other parts of the brain stem and spinal motor system. According to contemporary theory, individual cortical cells function as components of distributed neural networks incorporating a number of anatomically distinct structures, rather than as part of a localized 'center' performing a certain function.

2.1 *Nature of the Body Map*

By noting the progression of epileptic seizures from one body part to another, the British neurologist J. Hughlings Jackson surmised that the motor cortex was arranged in a rough map of the body. Electrical brain stimulation confirmed that conclusion. Nearest the top of the brain (Fig. 1), the leg is *represented*. Representation, in this sense, implies that this part of motor cortex processes information about leg move-

ments. The leg representation sends some of its axons to motor neurons that control leg muscles. Adjacent to the leg representation, in roughly the middle of the hatched zone in Fig. 1 (from top to bottom), lies an arm representation. The part of motor cortex that controls the face, head, tongue, and lips is situated yet further from the top of the brain. The finer organization of the motor cortex remains controversial and, although many detailed maps have been proposed, none have stood the test of time. In this respect, the organization of motor cortex contrasts with that of the somatosensory, visual, and auditory areas of the cerebral cortex, which have finer and more orderly maps of the receptors for touch, sight, and hearing, respectively.

3. *Function of Motor Cortex*

The earliest knowledge about the function of motor cortex arose from observing the effects of damage to it or its main output pathway, the *pyramidal tract*. Unfortunately, the medical literature often lacks an accurate assessment of the exact site and extent of brain damage. It is clear, however, that a number of severe motor disabilities follow such damage, including weakness, difficulty in controlling the muscles individually, and loss of fine motor control. It is commonly held that the motor cortex functions mainly to permit motor 'fractionation,' i.e., the independent control of muscle groups that usually work in concert. In this sense, the motor cortex can be viewed as supplementing the subcortical motor system. According to this idea, motor fractionation adds an important degree of flexibility to the motor system. Although in many primates, including humans, independent movement of the fingers typifies fractionation, the ability to fractionate movements is by no means confined to the digits. Whatever the motor cortex does, it does for the body as a whole.

In the early 1960s, the American neurophysiologist Edward Evarts developed a method for recording the activity of individual neurons in the motor cortex of awake, behaving monkeys. He began by investigating a question first posed by Jackson in the nineteenth century. Does the motor cortex control muscles or movements? This seemingly innocuous question engaged researchers for many years. It was finally resolved by dividing the question into two components, which yielded contradictory answers. One question concerns the *addressing* of muscles by motor cortex. Do cells in motor cortex address individual muscles (the 'muscle' view) or several muscles, which may act together to produce a coherent action (the 'movement' view). Early brain-stimulation studies suggested single-muscle addressing. However, with more careful research the multiple-muscle movement hypothesis prevailed for the specific issue of addressing. The other question concerns the signal communicated to these muscle groups. Does motor cortex control the level and pattern of force generated by muscles (the 'muscle' view) or, alternatively, joint angles, limb position, and the direction of movement (the 'movement' view)? Neurophysiologists have found that most cells in the motor cortex control muscle activity rather than movement. Motor cortex emits a force signal, which usually causes movement, but may instead generate force without movement (e.g., when one pushes against an immovable object).

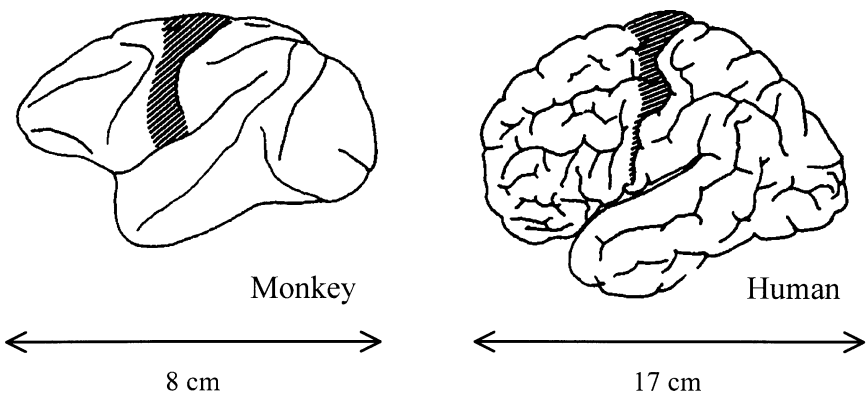

Figure 1
The cerebral cortex of a macaque monkey (left) and human (right), to different scales. Each of the lines within the outline depicts the typical location of the sulci, or grooves, in the cortex. The arrows show the distance from the front to the back of the cortex. The zone with the oblique hatching is the primary motor area. Source: Passingham (1993), by permission of Oxford University Press. Passingham R E 1993 *The Frontal Lobes in Voluntary Action*. Oxford University Press, Oxford, UK

A smaller number of cells control a combination of force and movement.

3.1 Population Coding

Individual neurons have their greatest activity for force in one direction, with systematically less activity as the direction of force diverges from that preferred direction. The cells are thus said to be *tuned* for direction. On the assumption that these cells generate force in their preferred direction, it is possible to estimate a collective force signal emanating from a neuronal population. This computation, termed the *population vector*, matches closely the direction of movement and force generated in a straight line, as well as both spiral and sinusoidal trajectories (Georgopoulos 1991).

3.2 Motor Cortex, Volition, and Awareness

Voluntary actions are those that are learned, attended, and based on a comparison among alternatives (Passingham 1993). Other kinds of action include reflexes and other innate behaviors. It seems likely that the motor cortex controls voluntary actions, which are typically based upon consciously perceived information. However, this contention implies neither that it lacks a role in reflexes nor that all voluntary movements depend upon consciousness.

Recently, the distinction between explicit (or declarative, perceived) knowledge and implicit (or procedural, unperceived) knowledge has gained wide acceptance. The former has access to conscious awareness; the latter does not. Both explicit and implicit knowledge can guide voluntary action. For example, people with 'blindsight' can point to a visual stimulus while denying that they see it (Weiscrantz 1986). Similarly, one well-studied neurological patient can orient her hand in order to put it through a narrow slot, but cannot report the orientation of the same slot (Milner and Goodale 1995). These are examples of implicit, unperceived knowledge guiding voluntary action. It seems likely that the motor cortex is involved in both explicitly and implicitly guided voluntary movements.

It has become increasingly clear that the processing of sensory information for perception differs from that for movement. In some respects, the performance of the motor system surpasses the neural systems underlying perception. For example, people can make movements that accurately match the size of an object to be grasped, but report incorrectly the size of the same object due to a visual illusion (Milner and Goodale 1995). The motor system thus has access to somewhat different sensory information than the brain systems underlying perception. As a consequence, one cannot assume that people's actions invariably reflect their perceived knowledge.

3.3 Motor Learning and Plasticity

Motor learning and memory include both adaptation and skill acquisition. Adaptations involve changes in motor performance without an enhancement in capabilities. For example, if a prism shifts the visual image of a target, it takes several attempts to correct the error induced in reaching toward that target. Such adaptation leaves the system able to reach or throw accurately, but no more so than before adaptation. By contrast, learning a skill involves obtaining a new capability or improved performance, beyond the system's prior limits. Evidence from brain-imaging studies suggests that the motor cortex changes the way it processes information during both skill learning (e.g., learning a movement sequence) and adaptation (e.g., tracking a rotating target). Other methods lead to the same conclusion. During sequence learning, for example, the movements caused by stimulating motor cortex become larger and easier to evoke (Hallett 1999). And during motor adaptation, the activity levels and other properties of neurons change.

Nearly 1 percent of the population has suffered a stroke causing motor disability. Many of these patients survive for years, 35 percent for a decade or more. Recovery of function usually proceeds slowly and often remains incomplete. Recent research has shown that on account of the brain's plasticity, important organizational changes occur after damage to motor cortex. Practice that leads to such plasticity enhances recovery of function (Nudo et al. 1997). Future research will likely focus on drug treatments and other therapies aimed at regulating cortical plasticity. The advent of treatments to ameliorate damage to the motor cortex could have effects resembling the development of correctional optics in the thirteenth century. Eyeglasses extended the productivity of artisans by freeing them from the limitations of middle-age farsightedness (Landes 1998). Extending the productivity of patients with motor cortex damage could have similar benefits for both the individual and society.

See also: Cerebellum: Cognitive Functions; Motor Control; Motor Control Models: Learning and Performance; Motor Skills, Psychology of; Pre-motor Cortex

Bibliography

Georgopoulos A P 1991 Higher order motor control. *Annual Review of Neuroscience* **14**: 361–77

Hallett M 1999 Plasticity in the human motor system. *Neuroscientist* **5**: 324–32

Landes D S 1998 *The Wealth and Poverty of Nations: Why Some are so Rich and Some are so Poor*. 1st edn. W W Norton, New York

Lassek A M 1954 *The Pyramidal Tract: Its Status in Medicine*. Thomas, Springfield, IL
Milner A D, Goodale M A 1995 *The Visual Brain in Action*. Oxford University Press, Oxford, UK
Nudo R J, Plautz E J, Milliken G W 1997 Adaptive plasticity in primate motor cortex as a consequence of behavioral experience and neuronal injury. *Seminars in Neuroscience* **9**: 13–23
Passingham R E 1993 *The Frontal Lobes in Voluntary Action*. Oxford University Press, Oxford, UK
Porter R, Lemon R 1993 *Corticospinal Function and Voluntary Movement*. Clarendon Press, Oxford, UK
Weiscrantz L 1986 *Blindsight. A Case Study and Implications*. Oxford University Press, Oxford, UK
Wise S P 1999 Motor cortex. In: Adelman G, Smith B H (eds.) *Encyclopedia of Neuroscience*, 2nd edn. Elsevier, Amsterdam

S. P. Wise

Motor Skills, Psychology of

The history of the psychology of motor skills starts at the end of the nineteenth century. Among the landmarks are Woodworth's studies of aimed movements, which established a fundamental distinction between a first 'ballistic' movement phase and a later visually guided homing-in phase. Another landmark are Bryan and Harter's 'studies of the telegraphic language,' which established the notion that during the acquisition of motor skills elementary units are integrated to form higher-level units. In the early twentieth century, Thorndike's 'law of effect' motivated a number of studies on the effects of knowledge of results on skill acquisition, and later Hull's concepts of reactive and conditioned inhibition formed the theoretical foundation for studies of massed and distributed practice. Among the discoveries emanating from European schools of psychology were the covariations between components of motor skills and the almost constant duration of movements when identical figures of different sizes are drawn.

From the 1960s on there was a noticeable increase in research on motor skills. Concepts of cognitive psychology turned out to be better suited for the analysis of motor behavior than, for example, concepts of behaviorism. In addition, many researchers came to entertain the idea that cognitive processes primarily serve the control of action rather than cognition *per se*. Finally, there was a merging of different academic disciplines such as neurobiology, physiology, neurology, biomechanics, psychology, and physical education to form what is called 'movement science' or 'kinesiology.'

The psychology of motor skills can be organized in two fundamentally different ways. One organization is according to different skills such as standing, walking, speaking, and writing. The other relates to basic concepts that are applicable to real-life skills, but typically studied by means of artificial tasks that are particularly suited for their purpose. Here some major concepts will be described.

1. Autonomous and Stimulus-based Processes of Motor Control

There is a kind of classical controversy which, historically, came in at least two different formats. In both formats it is about the question of how critical sensory stimulation is for the production of movement. In the historically earlier format, the alternative conceptions were that of a reflex chain, according to which stimuli which arise from one elementary movement trigger the next elementary movement, and that of autonomous activity of the central nervous system in generating the basic motor patterns. In the historically later format, the reflex chain was replaced by the notion of a continuous processing of sensory feedback.

The controversy about the roles of autonomous and stimulus-based processes in motor control has dissolved. The surviving problem is how sensory information is integrated with autonomous processes. Basically, sensory information can be used continuously as a feedback signal or as a reference signal, intermittently for the updating of parameters of autonomous processes or for the triggering of new autonomous processes, and finally sensory information can be used before a movement is initiated.

There is good reason to assume that autonomous or feed-forward processes of motor control run in parallel to closed-loop processes. Such an arrangement can actually be identified anatomically in the spinal control of muscle length, and it helps in understanding some fundamental characteristics of human movements. First, closed-loop systems have a tendency to become unstable when the feedback is processed with some delay and the gain in the loop is sufficiently high. In biological systems feedback is always processed with some delay, and the high velocities of human movements would require a high gain of a closed-loop system. Nevertheless, instabilities are extremely rare. With open-loop control in parallel with closed-loop control, the closed-loop system can operate with a small gain. Hence instabilities can be avoided, but the inaccuracies of open-loop control can nevertheless be compensated.

Second, primate movements are astonishingly little affected by elimination of peripheral feedback, even by more or less complete deafferentation. Such observations suggest that peripheral feedback is not processed at all. On the other hand, the processing of peripheral feedback is clearly indicated by strong effects of

distorted feedback signals. Among the best-known of such effects is stuttering, which can be induced by delayed feedback of speech signals. Such apparently contradictory findings can be predicted from a parallel arrangement of open- and closed-loop control. With this arrangement there will always be effects of distorted feedback, but the elimination of sensory feedback will affect movements only to the extent that open-loop control is inaccurate.

Sensory information is used not only as a feedback signal, but also as a reference signal which indicates desired movement characteristics. Only rarely does sensory information specify the time course of a movement in all details. More frequently, only certain characteristics, such as target positions for aiming movements, are specified, and sometimes there is essentially no sensory specification of desired movement characteristics. Consider a movement aimed at a certain target position. Before autonomous processes of motor control can be initiated, parameters related to movement characteristics such as amplitude and direction must be adjusted according to the position of the target, and other movement characteristics such as velocity must be specified according to other criteria (which often seem to be related to some kind of cost associated with performance). Such preparatory processes are often called 'programming,' and the autonomous processes are referred to as 'motor programs.' One of the continuing problems in movement research is to identify the defining characteristics of motor programs.

2. *Visuo-motor Coordination*

Movements have to be adapted to the spatial layout of the environment. Transforming the visually perceived position of an object in space into appropriate motor commands which guide the hand to that particular position is a nontrivial problem. On the one hand it involves transformations between different coordinate systems, and on the other it involves transformations of kinematic movement characteristics into the appropriate torques at the joints and finally muscle activity. Typically these transformations involve 'ill-posed' problems in that there is no unique solution except when additional boundary conditions are introduced. In addition, the solutions vary across persons and across time. For example, when the task is to place the finger in a certain spatial position, the appropriate joint angles at the shoulder, elbow, and wrist depend on the lengths of the limb segments of the moving person; what exactly the required torques will be depends on the passive torques such as those which result from the action of gravity; what exactly the required innervation will be depends on the current state of the muscles, e.g., whether they are fatigued or not. Therefore, to take account of all such factors, visuo-motor coordination should exhibit a fair degree of flexibility.

Perhaps the most straightforward way to explore the flexibility of visuo-motor coordination is to modify the relation between visually perceived positions and the movements required to reach these positions. For example, spectacles with wedge prisms can be used which displace the visual world. When instructed to point (without seeing the hand) at a target which is visually displaced to the right, humans will point to the right of the target. When they are allowed to see the error at the end of each movement, it will gradually disappear—and this is a matter of minutes, not of hours. After removal of the displacing prisms, open-loop pointing exhibits a negative after-effect—the movements end to the left of the target.

The recalibration of visuo-motor coordination induced by the exposure to prismatic displacement is a fairly rapid process. Even more rapid is the recalibration of the relation between kinematic movement characteristics and the (active) torques required. For example, upward and downward movements of the hand are essentially identical kinematically, but vastly different in terms of muscle activity (and active torques). Adjustment to different weights that have to be lifted is almost instantaneous, in particular when the weight can be inferred from the appearance of the object before the movement is initiated. However, there seems to be at least one situation to which full adaptation is never achieved, namely microgravity. Microgravity changes the 'normal' relation between mass and weight, and here the adaptive capabilities of humans seem to fail. This may not be particularly surprising because gravitational acceleration has been a constant during human evolution, and there has never been any evolutionary pressure toward developing adaptive capabilities with respect to changes in gravitational acceleration.

Spectacles can be used to induce more severe distortions of the visual world than just displacing it by a few degrees. For example, the visual world can be rotated or inverted, or left and right can be exchanged. Adaptation to such distortions takes days rather than minutes. In spite of the full visuo-motor adaptation to severe distortions of the visual world, its appearance seems to remain different, although it is not exactly clear in what ways. The difference between visuo-motor and perceptual adaptation was perhaps the first indication that sensory information for (conscious) perception of the world and motor control can be at least partially dissociated.

In the late twentieth century, both clinical and experimental evidence has been accumulated to indicate that the visual world in which we move is not exactly the world which we experience. Among the clinical evidence is the phenomenon of 'blindsight': patients are unable to perceive the location of a light spot in a certain part of their visual field, e.g., in the

right hemifield, but nevertheless they can point to it with some degree of accuracy. Among the experimental evidence is the demonstration that induced motion apparently is 'unknown to the hand.' A target can be made to appear moving by way of moving a frame that surrounds it. When asked to point at the one or at the other end point of the induced motion (without seeing the hand), humans point at more or less the same position. When the induced motion is canceled by a real motion so that the target appears stationary, pointing after the target has disappeared at one or the other end point of the real motion ends at different positions. Although it is clear that 'visual information for action' can be separated from 'visual information for perception,' the precise characterization of the different visual systems involved is still a topic of research.

3. *Motor Timing*

Some movements have to be adapted to temporal characteristics of environmental events. An example is catching or hitting a ball. Perhaps of greatest importance is the timing of movements in playing music. These examples suggest a distinction between extrinsic and intrinsic timing. In both cases timing is of critical importance for the skill, but only in the former case it is the temporal relation with respect to some environmental event which is critical, whereas in the latter it is only the temporal relation among the movements of a sequence. Even in tasks without explicit timing requirements, temporal regularities can be observed.

A great deal of effort has been invested in identifying the visual information involved in extrinsic timing, in particular with respect to approaching objects. There is indeed visual information about time to contact with an object. In principle any ratio of a distance and its first derivative with respect to time provides an estimate of the time remaining until the distance becomes zero, provided that the first derivative is constant. Such ratios can also be computed for variables which can be sensed directly, e.g., the size of the retinal image of an object and its rate of expansion when the object approaches. This ratio in fact is involved in the extrinsic timing of, e.g., catching, but other information is likely to be involved in addition.

A kind of mixture of extrinsic and intrinsic timing can be found in the simple experimental task of tapping in synchrony with a metronome. Although this is basically an extrinsic-timing task, the timing requirements are highly regular so that intrinsic mechanisms come into play. One of the fundamental findings in synchronization tasks is the so-called negative asynchrony: the taps lead the pacing signals by about 20–50ms, depending on conditions. This asynchrony is likely to come about by different delays for processing the pacing signals and the tap feedback, so that synchronization at some central level of representation requires asynchrony at the peripheral level where the measurements are made. Negative asynchrony turns into positive synchronization errors when the frequency of the pacing signals is slowed, so that the timing of the clicks can no longer be anticipated with sufficient accuracy.

Mechanisms of intrinsic timing can be studied best in tasks without external pacing. In the simplest task, a tapping rate, which is established with the support of pacing signals, has to be continued without pacing. Among the basic results is a linear increase in variability with the mean intertap interval and a negative serial correlation between adjacent intervals, but a zero correlation between nonadjacent intervals. Hence there is a tendency for shorter and longer intertap intervals to alternate. These findings gave rise to a two-level model of motor timing which allows one to decompose the total variability of the intertap intervals into a central and a peripheral component, ascribed to a central timekeeper and motor delays, respectively. The two components are affected differentially by different task conditions and also by different kinds of brain lesions. The two-level model has been generalized to more complex unimanual and bimanual rhythms and has proven to be a useful framework for the analysis of motor timing.

For many motor skills there are no explicit timing requirements, but nevertheless there are constraints on temporal characteristics so that the timing is nonarbitrary. One of the well-researched constraints is the relation between velocity and curvature in drawing figures. A classical observation is that there is a constant velocity in drawing circles and a variable velocity in drawing ellipses; it is not only that velocity increases with the radius of curvature (as a power function with the exponent 1/3), but also that the radius of curvature changes with instructed changes in velocity. For example, when one attempts to draw a circle with high velocity in the more vertical segments and low velocity in the more horizontal segments, the result is likely to be a vertically oriented ellipse.

4. *Motor Coordination*

Movements are spatially and temporally adjusted not only to the environment, but also to concurrent movements of other parts of the body. Motor coordination can be in the service of the performed task. For example, in producing a certain lip aperture the positions of the upper and lower lips are correlated across a series of attempts, so that variation in the position of one lip is compensated for by the covariation of the position of the other. 'Compensatory

covariation' seems to be a fairly robust and general mechanism for producing stable movement outcomes.

Particularly intriguing is the coordination of the two hands in the service of task performance. In right-handers at least there is a clear specialization of hand use in that manipulations are assigned to the dominant (right) hand, while a supportive function is assigned to the nondominant (left) hand. The supportive function can be a simple hold function as in threading: when the needle is in the left hand, the thread will be moved toward the needle, and when the thread is in the left hand, the needle will be moved, but in each case the right hand is moved toward the left one. In more complex cases, the left hand provides something like a moving reference for the movements of the right hand, as in using a shovel or broom.

Constraints on motor coordination can also impede task performance. This is not very often noted in everyday life, except perhaps when playing a musical instrument, but it is easy to design experimental tasks which cannot be performed as instructed. For example, when concurrent movements of the two hands with different temporal and/or spatial characteristics are instructed, they tend to become too similar. This assimilation indicates the existence of cross-talk between concurrent processes of motor control that is hard or impossible to overcome intentionally.

The exploration of the limits of bimanual skills has served to identify a number of constraints on coordination. The most conspicuous ones are a temporal constraint, which results in a tendency toward identical timing of concurrent left- and right-hand movements, a spatial constraint, which results in a tendency toward identical movement amplitudes, a force constraint, and a symmetry constraint. The last constraint has been studied most extensively. For example, it is an easy task to produce symmetric oscillations of the two hands. In contrast, the production of asymmetric oscillations is an easy task only when the frequency is low; with a sufficiently high frequency, asymmetric oscillations tend to switch to symmetric ones.

There are some indications that cross-talk effects in bimanual movements are asymmetric, perhaps related to the functional specialization of the two hands in bimanual actions. For example, when one hand produces a steady beat and the other hand taps as rapidly as possible or with a certain rhythm, the bimanual task is much easier when the steady beat is assigned to the left hand than when it is assigned to the right hand. Less consistently found is a smaller assimilation of the amplitudes of concurrent movements with the long amplitude being assigned to the left hand and the short amplitude to the right hand than with the opposite assignment of amplitudes. These differences are consistent with the suggestion that in bimanual tasks the left hand is specialized for long-amplitude, low-frequency movements, but the right hand for small-amplitude, high-frequency movements.

See also: Automaticity of Action, Psychology of; Classical Mechanics and Motor Control; Motor Control; Motor Control Models: Learning and Performance; Motor Cortex; Neural Representations of Intended Movement in Motor Cortex

Bibliography

Heuer H, Keele S W (eds.) 1996 *Handbook of Perception and Action. Vol. 2: Motor Skills*. Academic Press, London

Rosenbaum D A 1991 *Human Motor Control*. Academic Press, San Diego, CA

Schmidt R A, Lee T D 1999 *Motor Control and Learning. A Behavioral Emphasis*, 3rd edn. Humans Kinetics, Champaign, IL

H. Heuer

Movement Theory and Constraints in Syntax

1. Movement in Syntactic Theory

Movement theory is a 'trademark' of Generative-Transformational Grammar (see *Generative Grammar*; *Syntax*; *Linguistics: Incorporation*). In this line of inquiry, certain sentences are generated, not just by combining words or phrases in some appropriate order, but also by moving them from one position to another. Thus the sentences (1)–(3) below are not constructed by directly 'merging' the leading items *will*, *what*, and *John* respectively at the sentence-initial position, but by first introducing them in their 'original' positions (marked by the underscores) and then moving them to the initial position:

(1) Will John __ buy this toy?
(2) What has John put __ on the table?
(3) John was arrested __ yesterday.

In the Principles and Parameters Theory (see *Linguistics: Theory of Principles and Parameters*) as implemented in the Government-Binding-Barriers framework (Chomsky 1981, 1986), the computation system of a grammar generates underlying Deep-Structure (DS) representations in accordance with the X′ Theory that satisfy the selectional requirements of lexical items (see *Valency and Argument Structure in Syntax*), and transformational operations (particularly, movement operations) convert them to Surface-Structure (SS) representations. These SS representations are in turn mapped onto two levels that form the interface between grammar and other cognitive performance systems, PF (Phonetic Form) on the one hand and LF (Logical Form) on the other. (See *Syntax–Semantics Interface*; *Syntax–Phonology Interface*.) Thus, for sentence (2) above, the computation system will carry out movement operations (as

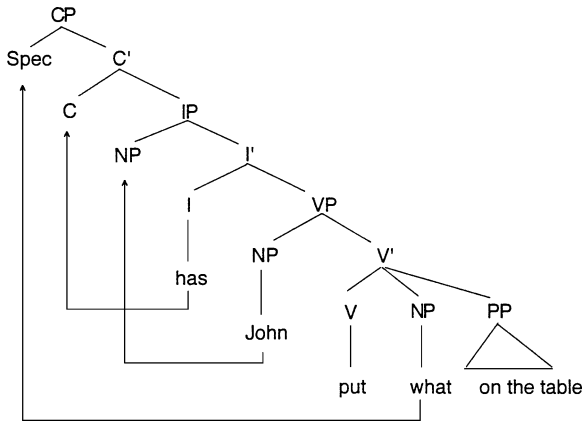

Figure 1
Tree diagram showing derivation of 'What has John put on the table?'

indicated by the arrows) that turn its DS to its SS representation:

It is assumed that the mapping SS → PF, and especially SS → LF, also involve critical movement operations.

2. *Motivations for Movement*

In some cases the postulation of movement can be justified by intuitive observations. The verbs *put* and *arrest* are transitive in traditional terms, requiring NP objects following them (compare, for example **John put on the table*; **Bill arrested*; an asterisk preceding a sentence indicates ungrammaticality). In sentences (2) and (3), no NP object appears in the 'expected' postverbal position, despite their grammaticality. This paradox is explained if one says that *put* does have a postverbal object at DS, though the object is fronted at SS (as depicted in Fig. 1). The postulation of movement thus enables one to preserve generalizations about the predicate-argument structure of verbs and other lexical items (see *Valency and Argument Structure in Syntax*) that explain grammatical well-formedness.

In other cases, the movement hypothesis has enabled generative syntacticians to unearth interesting phenomena that had escaped introspection in previous descriptions, and gain valuable insight into syntactic structure. For example, it is now widely assumed that, despite their surface similarity, derivation of sentence (4a), but not of (4b), involves movement:

(4) a. Harry seems to be honest.
    b. Harry tried to be honest.

In particular, in (4a) but not (4b), the subject of the main verb *seem* originates as the subject of *be honest* and raises to the higher subject position at SS. Postulating Subject Raising in (4a) but not (4b) explains several differences between *seem*-type and *try*-type sentences that surface upon closer examination, including e.g., *The shit seems to be about to hit the fan* (under the idiomatic interpretation) but **The shit tries to be about to hit the fan*, etc. Similarly interesting facts have been explained concerning *Mary is likely to leave* vs. *Mary is eager to leave* vs. *Mary is easy to leave*.

In yet other cases, the existence of movement could not have been detected by intuitive observations of the type one can make concerning sentences (1)–(3). One early rule, proposed by Chomsky (1955) and popularly known as Affix Hopping, concerned tense marking in English finite sentences. The fact is that tense marking seems to 'migrate' among auxiliary elements and the main verb: it may occur on a modal (*Monica* ***would*** *sing*, but **Monica will sang*), on a perfective or progressive aspect (*Monica* ***has*** *sung*, *Monica* ***is*** *singing*, but **Monica have sings*, **Monica have is singing*, etc.), or directly on the main verb (*Monica* ***sang***). A context-free phrase-structure grammar would be hard put to describe this range of sentences in any simple way. Chomsky proposed treating the tense inflectional suffix as an independent syntactic item in DS representation (as a leftmost constituent of the Auxiliary) which, by Affix Hopping, 'docks' on whatever first auxiliary or verbal element follows it. Affix Hopping vastly simplifies the grammar of tense-marking, at the same time capturing the observed 'migration' (hopping) property insightfully.

The primary argument for movement is that it allows one to simplify grammatical descriptions, capture significant generalizations, and uncover deep insight into the nature of human language as a 'mirror of the mind.' Such considerations justify the movement hypothesis not only for more observable cases like sentences (1)–(3) and (4a), but also many surprising or less observable cases. (Some versions of generative syntax incorporate other devices to explain the 'movement phenomena,' including Lexical Functional Grammar (see *Lexical Functional Grammar*), head-driven phrase structure grammar, categorical grammar, Dependency Grammar.)

3. *Evolution of Movement Theory*

3.1 *Early Movement Theory*

The early years following Chomsky's proposal of Affix Hopping saw an entire field of syntax scholars devoted to grammatical description applying the transformational model. New syntactic phenomena were unearthed or became interesting enough theoretically to call for an analysis, and new rules were added to the repertoire of a grammar's transformational com-

ponent. In addition to Subject-Auxiliary Inversion, *Wh*-Movement, Passivization and Subject-Raising illustrated in sentences (1)–(3) and (4a) respectively, and the aforementioned Affix Hopping, other movement rules were also proposed, including Relativization (as shown in sentence (5)), Topicalization (6), *Tough*-Movement (7), Extraposition from NP (8), etc.

(5) This is the person *who* I have always wanted to meet__.
(6) *This sort of movie*, I would never watch __.
(7) *John* is tough for us to deal with __.
(8) A book has just appeared *on syntactic theory*.

In order to ensure descriptive adequacy, each rule proposed was formulated in such a way as to (a) apply to only the input strings that meet a certain Structural Description (SD), (b) effect precisely the Structural Change (SC) as desired, and (c) only do so under any additional conditions as may be imposed. Thus the Passivization rule of English was defined in terms of the SD of an input structure that contains a postverbal object NP and the instruction for SC that it will (among other things) front the NP to the subject position. Additional conditions are noted for the applicability of the rule, among them the requirements that the main verb be put in the passive participle form, that the movement be carried out to an immediate local subject position, etc. Other rules were defined with similar explicitness and in detail. Because the investigations were carried out in the generative (i.e., formal) approach, linguists working on syntactic phenomena were required to formulate precise hypotheses about the form and functioning of movement transformations they proposed. The explicit (and therefore falsifiable) formulations led to unprecedented expansion of our explicit knowledge of natural language syntax.

3.2 *Modularity: From Rules to Principles*

While the detailed formulations of movement rules helped ensure observational and descriptive adequacy of proposed grammatical descriptions, linguists realized increasingly that the adequacy was achieved at the cost of explanatory adequacy and conceptual beauty. For one thing, the detailed statement for each movement rule obscured the fact that many movement rules have similar properties. By stipulating these properties repeatedly in the statement of every rule, the theory was highly uneconomical and failed to capture the generality of such properties. In fact, some properties often included in the statement of movement rules are not peculiar to movement, but hold generally of other syntactic phenomena as well. To capture the correct generalizations, then, it was necessary to separate out general properties from the statement of specific rules.

One general property of movement that was repeatedly stipulated in the formulation of each rule is that movement is generally *upward*, i.e., that it moves elements to a 'c-commanding' position. Defined in terms of tree diagrams like Fig. 1, a node α c-commands β iff neither α nor β dominates (contains) the other, and the first branching node dominating α dominates β (see *Constituent Structure: Syntactic*.) In Fig. 1, each instance of movement is upward to a position c-commanding the original position. This 'upward movement requirement' should be stated once and for all movement rules, rather than separately for each rule. In fact, Chomsky (1975) noted that this requirement is just a special case of a more general principle that also governs anaphora in natural language. For example, the interpretation of a reflexive pronoun in English requires that it have an antecedent that c-commands it: *John criticized himself*, but **John's mother criticized himself* (the latter is ungrammatical because the intended antecedent *John* is not high enough to c-command *himself*). To capture this common property between movement and anaphora, Chomsky proposed that, when a category is moved, it leaves a 'trace' behind, co-indexed with the moved category (as in *John*$_i$ *was arrested* t_i *yesterday* for sentence (3)). The trace, being phonetically empty and having no inherent reference, depends on the moved category as its 'antecedent,' much as a reflexive depends on another noun phrase for its reference. The 'upward movement requirement' thus reduces to a general requirement governing anaphoric dependency, and may be eliminated from the statement of each movement rule.

Another common property of movement rules is that they operate only locally, as the next section will show. A third general property is that movement seems to operate in a structure-preserving manner, so that a head (e.g., verb) can only move to a head position, an argument NP to an argument position (A-position), and an operator to an operator position (A′-position). (See again Fig. 1, illustrating *has* undergoing head-movement, *John* undergoing A-movement, and *what* undergoing A′-movement.) The discovery of these and other general properties finally led linguists to the view that most details formerly attributed to individual rules should be dissociated from them and be ascribed to independent modules of grammar each with general applicability, and that all movement rules can be reduced to one rule of maximal simplicity: Move α, α a category. The result is a highly modular theory of movement: Universal Grammar consists of only one general movement rule, Move α, and a number of independent general principles whose interaction among themselves and with Move α produce a complex array of syntactic phenomena. There is then no Passivization or Relativization as a movement rule *per se*, these being simply special instances of Move α. This modular view of grammar culminated in the theory of

Government and Binding in Chomsky (1981) and subsequent works, and is still the prevalent view of current researchers in generative grammar.

Under this highly modular view of movement, there is little or nothing left to do in the formulation of the movement rule itself. At the same time, increasing attention has shifted to the study of other modules of grammar, in particular the study of general locality constraints on movement, and on the distribution of empty categories created by movement.

4. Constraints on Movement

In generative linguistics, a grammar G of a given language L is a hypothesis about the linguistic phenotype of the speakers of L. Such a hypothesis is descriptively adequate only if it defines all and only the (potential and actual) grammatical sentences of the language. A linguistic theory, a hypothesis of the human linguistic genotype, is likewise adequate only if it defines all and only the possible grammars that are descriptively adequate. In the early stages of generative grammar study, linguists were occupied primarily with providing their grammars with enough descriptive powers to generate all grammatical sentences. In most subsequent developments to date, however, most attention has been directed toward restricting the expressive powers of their grammars. The study of movement constraints was motivated not only on empirical grounds of descriptive and explanatory adequacy, but also because it has proven a very productive probe into the nature of the human language faculty.

4.1 A-over-A and Ross Constraints

The need for a general, powerful constraint on movement rules was noted quite early by Chomsky (1964), who proposed the A-over-A Principle to exclude movement of a phrase out of a larger phrase of the same type. The importance of movement constraints gained widespread attention with the emergence of Ross (1967), who showed the A-over-A to be inadequate and proposed, in its place, a module of grammar containing a set of 'island constraints.' According to Ross, although movement rules may appear to displace constituents over an arbitrary distance (e.g., *What$_i$ did Harry say that Maria bought t$_i$?*), they must be constrained sharply so as not to extract elements out of certain domains that he called 'islands.' The following examples (sentences (9)–(14)) respectively) illustrate Ross's island constraints (in each case the relevant 'island' is marked by square brackets): Coordinate Structure Constraint (CSC), Complex NP Constraint (CNPC), Sentential Subject Condition (SSC). Other familiar constraints include the Subject Condition (SC), the Adjunct Condition (AC), and the *Wh*-Island Condition (WIC):

(9) * What$_i$ did [Harry buy t$_i$ and Maria buy a pencil]? (CSC)
(10) * What$_i$ did Harry talk to [the woman who bought t$_i$]? (CNPC)
(11) * What$_i$ was [that Maria had bought t$_i$] surprising? (SSC)
(12) * Who$_i$ did [a picture to t$_i$] please you? (SC)
(13) ?* What$_i$ did Harry go home [before Maria bought t$_i$]? (AC)
(14) ?* What$_i$ did Harry wonder [who bought t$_i$]? (WIC)

4.2 Subjacency, CED, and the ECP

The Ross-style constraints effectively reduced the power of grammar but, as given, they amounted to a list of environments that form 'islands' to the exclusion of those that do not, and thus failed to explain their clustering. Chomsky (1973) proposed subsuming most of Ross's island constraints under one general Subjacency Condition, which prohibits extraction of a constituent across two or more 'bounding nodes' (NP or S). Huang (1982) argued that the SSC/SC and AC reflect a general complement vs. non-complement asymmetry and are best collapsed under one general Conditon on Extraction Domains (CED), which provides that extraction is possible only from 'properly governed' domains (i.e., complements). (See also Kayne 1981.) The notion of proper government also plays a crucial role in defining the Empty Category Principle (ECP), which excludes long extraction of a subject or an adjunct, as shown in sentences (15)–(16):

(15) * This is the person who$_i$ John wondered [whether t$_i$would fix the car].
(16) * This is how$_i$ John wondered [whether Bill would fix the car t$_i$].

The ECP requires that each trace be governed by a lexical head, or extraction must be local (short distance). Long extraction is excluded in sentences (15) and (16) because subjects and adjuncts are not governed by a lexical head.

4.3 Barriers, Economy, and Minimalism

The program of unifying (and thus explaining) movement constraints continued in Chomsky (1986), where a single notion of *barrierhood* was proposed to tie together these conditions, whereby the theory of bounding (Subjacency and CED) allows movement to cross at most one barrier (at a time), and the theory of government (ECP) requires that a trace not be separated from its governor by any barrier at all. The *barriers* framework also provided a tool for measuring the degree of (un)grammaticality in incremental terms, capturing the fact that a movement crossing n barriers produces worse results than one that crosses $n-1$

barriers. The various constraints, defined in terms of barrierhood, then reduce to the notion that grammars select movements that are local, or as local as possible.

The notion of locality led to the idea of minimality and economy. Rizzi (1990) argued that most of the locality constraints can be reduced to the proposal that each movement operation must land at the closest potential landing site. The general success of reducing movement constraints to minimality is an important step that led Chomsky to other considerations of economy (minimality being one type of economy) and to the design of the Minimalist Program, which aims to employ no theoretical apparatus or levels of representation beyond 'virtual conceptual necessity,' and only derivations that occur with 'least effort' or as a 'last resort.' In Chomsky (1995), minimality is incorporated directly into the definition of Attract F, a reformulation of Move α.

Although the early A-over-A Principle proved to be inadequate as a unifying general constraint on movement, the various constraints are unified finally under one notion, this time couched in a general theory of economy.

5. *Covert Movement: The Syntax of LF*

In the Principles and Parameters theory, it is assumed that movement may be overt (if it occurs before SS—or Spell-Out), or it may be covert (if it occurs in LF). In the examples we have seen, movement is overt, with phonetically visible effects. It has been assumed widely that certain elements may move covertly as derivations enter LF, producing no phonetic effect. Examples include sentences containing quantifiers such as *everyone*, *some student*, *most people*. Chomsky (1976) argued that quantifiers are located covertly on the (left) periphery of the clause containing them, binding a variable in their surface position. Thus the surface sentence *John hates every politician* has the predicate–calculus style of representation *For every x, x a politician, John hates x*. May (1977) suggested that this representation may be obtained by an instance of Move α that raises the quantifier leftward, in a style reminiscent of 'quantifier lowering' rules proposed by generative semanticists in the early 1970s. Important arguments include generalizations concerning the possibilities of interpreting pronouns as 'bound variables' (Higginbotham 1980).

Another sort of covert movement is that of *Wh*-phrases in languages in which they occur typically in their 'original' positions. Unlike English-type languages, Chinese (and many other languages) leave their question words *in situ* when forming *Wh*-questions. Thus *What did you buy?* in English is simply *ni mai-le shenme* 'you bought what' in (Mandarin) Chinese. Note that the Chinese sentence (17) below is ambiguous between a direct and indirect question reading, corresponding to two distinct forms in English, as shown in sentence (18):

(17) ni jide Zhangsan mai-le shenme?
you remember Zhangsan bought what

(18) a. What$_i$ did you remember that John bought t$_i$? (direct question)
b. You remembered what$_i$ John bought t$_i$. (indirect question)

Despite the overt differences, Huang (1982) argued that the two languages (and all languages) have essentially the same LF representations. In particular, although *Wh*-phrases do not move overtly in Chinese, they do so covertly in LF. The ambiguity of sentence (17) follows as a result of the fact that *shenme* 'what' may be moved to the subordinate clause below *jide* 'remember' or to the main clause, deriving LF representations not unlike sentences (18a and b). In addition to capturing certain cross-linguistic generalizations, this hypothesis is supported by the fact that interpretation of *in-situ Wh*-questions is subject to some restrictions that govern overt movement, e.g., the ECP.

As an instance of Move α, LF movement is not limited to A′-movement like quantifer-raising or *Wh*-movement. Given the *Wh*-movement parameter (some languages move *Wh*-phrases overtly, others do it in LF), proposals exist in the current literature that both A-movement (e.g., object shift) and head-movement (e.g., V-to-I or N-to-D raising) may occur overtly in some languages but covertly in others.

6. *Current Issues and Future Prospects*

Recent and current works in the Minimalist Program have been guided by the desire to limit theoretical apparatus to a minimum and assume only rules that apply 'least effort' and as a 'last resort.' The last requirement has led scholars to require that no rule, not excepting Move α, can apply optionally. Each operation must be forced (or 'triggered') by grammatical elements that need to be 'checked off' or which would otherwise cause a derivation to 'crash.' Under the idea of 'triggering,' Move α is reformulated in Chomsky (1995) as a general rule of 'attraction' by the triggering element (Attract F). Considerations of economy favor the assumption that while overt movement unavoidably involves whole words or phrases, covert movement may simply move features. The overt–covert parameter continues to play an important role in accounting for cross-linguistic variation, now seen as a function of the variation in the nature of the triggers (the 'attractors') across languages.

The postulation of movement in syntactic theory has proven a productive probe into the inner workings of the human language faculty. Many discoveries have been made and questions answered since Affix Hopping came into being. The idea of 'covert movement' in LF opened up new possibilities and raised new

questions. In current Minimalist inquires, ever more emphasis is given to the role of movement, with new consequences that will continue to intrigue future researchers.

See also: Generative Grammar; Linguistics: Incorporation; Syntactic Aspects of Language, Neural Basis of; Syntax

Bibliography

Chomsky N 1955 *The Logical Structure of Linguistic Theory*. Harvard University and MIT Press, Cambridge, MA; published in part in 1975 by Plenum Press, New York

Chomsky N 1964 *Current Issues in Linguistic Theory*. Mouton, The Hague, The Netherlands

Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Chomsky N 1973 Conditions on transformations. In: Anderson S R, Kiparsky P (eds.) *A Festschrift for Morris Halle*. Holt, Rinehart, and Winston, New York

Chomsky N 1975 *Reflections on Language*. Pantheon Books, New York

Chomsky N 1976 *Essays on Form and Interpretation*. Elsevier, Amsterdam

Chomsky N 1981 *Lectures on Government and Binding*. Foris Publications, Dordrecht, The Netherlands

Chomsky N 1986 *Barriers*. MIT Press, Cambridge, The Netherlands, MA

Chomsky N 1995 *The Minimalist Program*. MIT Press, Cambridge, MA

Higginbotham J 1982 Pronouns and bound variables. *Linguistic Inquiry* **11**: 679–708

Huang C-T J 1982 *Logical Relations in Chinese and the Theory of Grammar*. Ph.D. dissertation, MIT, Cambridge, MA; published 1998 by Garland Publishing, New York

Huang C-T J 1994 Logical form. In: Webelhuth Q (ed.) *Handbook of Government and Binding Theory and the Minimalist Program*. Blackwell, New York

Kayne R 1981 ECP extensions. *Linguistic Inquiry* **12**: 93–133

May R C 1977/1990 *The Grammar of Quantification*. Ph.D. dissertation, Garland, New York

Rizzi L 1990 *Relativized Minimality*. MIT Press, Cambridge, MA

Ross J R 1967 Constraints on variables in syntax. Ph.D. dissertation, MIT, Cambridge, MA; published as *Infinite Syntax* 1986 by Ablex, Norwood, NJ

C.-T. J. Huang

MRI (Magnetic Resonance Imaging) in Psychiatry

1. Introduction

Although computerized tomography has been available for imaging studies in psychiatry for a number of years, the advent of MRI led to significant improvements in structural and subsequently also in functional imaging due to higher resolution and high-quality gray–white matter segmentation. The first studies using structural MRI were performed at the beginning of the 1980s and opened up new avenues for the investigation of functional and neurochemical properties of normal functions and pathological alterations of the brain.

This article will provide an overview over the different aspects of MR technology in psychiatry. Considering the limitations of space, no comprehensive presentation of the studies performed so far and no complete references can be given. Instead, milestones and representative examples of studies and results will be presented. For more comprehensive information, the reader is referred to suitable monographs and review articles (Mazziotta et al. 2000, Toga and Mazziotta 2000).

2. Structural Magnetic Resonance Imaging (MRI)

2.1 Technical Foundations

The MRI technique is based on the physical phenomenon that atomic nuclei with an odd number of nucleons (protons plus neutrons) behave like electrically charged spinning tops that produce a magnetic field. If a strong external magnetic field is now imposed, the nuclei align either parallel or antiparallel to the field, and precess around their axis at a specific (Larmor) frequency. External electromagnetic radiofrequency (RF) pulses can move the spinning nuclei out of their initial alignment. When the RF pulse is removed, the nuclei return to their equilibrium state. This relaxation is accompanied by a RF signal, which is detected by RF coils. Localization of the signal is accomplished by frequency and phase encoding by means of external gradients. The acquired information is then transformed into different gray levels and displayed as a MR image.

2.2 Findings

2.2.1 Schizophrenia. Numerous structural MRI studies suggested that the occurrence of ventricular enlargement is one of the most consistent findings in schizophrenia. Additionally, reductions in the size of the temporal lobes, the amygdala-hippocampal complex, the frontal lobes, parietal lobes, and the corpus callosum were described. On the other hand, volume increases of the subcortical nuclei have been found.

Despite these findings, there is still no general consensus on the presence of localized structural

abnormalities. Chua and McKenna (1995) conclude from their detailed review of structural MRI studies that the findings concerning the temporal lobe limbic structures as most consistent. In general, the described structural MRI findings support the notion of schizophrenia as a disturbance of fronto-temporal neuronal integrity.

It has been reported that the structural brain changes in schizophrenia do not correlate with the duration of the illness. Therefore, it seems likely that the structural alterations do not merely reflect advanced atrophic processes. Rather, they seem to stem from pathogenic factors which are most active either prior to or at the time of onset of the disease and do not significantly progress depending on the duration of the illness. However, this concept of schizophrenia is still controversially discussed and newer studies are suggesting progressing neuropathological alterations in schizophrenia after onset of the disease (DeLisi et al. 1997).

2.2.2 Affective disorders. Global cerebral atrophy and increased ventricular–brain ratio (VBR) in unipolar as well as bipolar depressed patients have been described but these findings are less consistent than those reported for schizophrenic patients. MRI studies in bipolar patients revealed evidence of global cerebral atrophy including lateral ventricular enlargement and increased VBR.

Reductions in the prefrontal lobe volume especially in the subgenual prefrontal cortex have been reported (Drevets et al. 1998). Recurrent major depression associated with hypercortisolism induced neurotoxicity might lead to hippocampal atrophy, which was indicated by the finding of bilaterally reduced hippocampal and amygdala core nuclei volumes in patients with major depression. Additionally, depressed patients have been shown to have smaller caudate and putamen volumes bilaterally. Late onset depressives demonstrated significantly more left medial temporal lobe atrophy than early onset patients. (See *Depression*.)

2.2.3 Dementia. CT and MRI have delineated structural changes in Alzheimer's disease (AD), including cerebral atrophy associated with widening of cortical sulci, ventricular enlargement, and deep white-matter lesions with periventricular distributions. Using MRI technology, it could be shown that significant brain atrophy is present early in the disease process and a direct correlation between the atrophy and the severity of cognitive impairment could be observed. Magnetic resonance imaging has illustrated this relationship as greater involvement of mediotemporal areas (Murphy et al. 1993).

Regarding the differential diagnosis of Alzheimer's disease from vascular dementia (VD), greater amounts of white matter and periventricular hyperintensities have been suggested to be indicative of VD.

3. Functional Magnetic Resonance Imaging (fMRI)

3.1 General Aspects

Aside from the above presented structural information, the MRI technology has been extended to the investigation of various functional aspects of the central nervous system and related neuropsychiatric disorders. These variations of functional MRI (fMRI) are either based on measures of regional differences in oxygenated blood (BOLD contrast), the change in regional blood flow (perfusion fMRI), the random movement of water molecules (diffusion weighted fMRI), or the quantity and distribution of neurochemical compounds in the brain (MRI spectroscopy). Based on the fact that changes in neuronal activity result in focal changes of perfusional parameters, blood oxygenation, and metabolism, fMRI lends itself to the study of brain activation under behavioral, neuropsychological, and emotional tasks. The technical aspects and representative findings will be discussed below.

3.2 Blood Oxygen Level Dependent Contrast (BOLD) fMRI

3.2.1 Technical foundations. Studies of brain activation based on the contrast agent-free BOLD technique are commonly referred to as functional MRI (fMRI). The BOLD technique uses the endogenous MRI contrast agent deoxyhemoglobin as the source of the signal. The ferrous iron on the heme of deoxyhemoglobin is paramagnetic but diamagnetic in oxyhemoglobin. Cerebral blood flow studies showed that cerebral activation leads to a local increase in oxygen delivery beyond the actual metabolic demand. This results in a higher concentration of diamagnetic oxyhemoglobin and a decrease of the paramagnetic deoxyhemoglobin in the activated brain regions. Consequently, the existing field inhomogeneities between intra- and extravascular space diminish and the T2* signal increases.

The most commonly used imaging sequences for BOLD-fMRI employ echo-planar imaging (EPI). The technique is based on the application of only one RF pulse followed by multiple reversals of the frequency-encoding readout gradient. The advantages of this sequence include high acquisition speed and high signal-to-noise ratio. fMRI using scanners with field

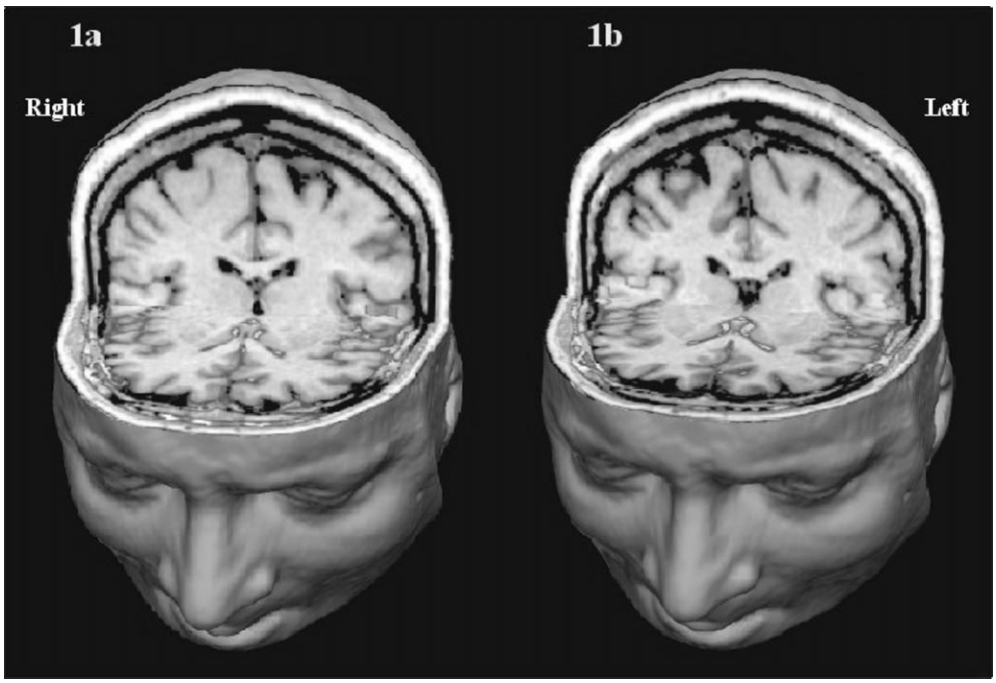

Figure 1
Increase in BOLD fMRI signal in Heschl's gyrus during hallucinations (a) and during acoustic stimulation (b). (Reproduced with permission of Cell Press from Dierks T et al. 1999.)

strengths greater than 1.5 T lead to improvement of the signal-to-noise ratio. Faster computer algorithms are currently starting to allow for near real time fMRI studies in which the activation maps are calculated during the study.

3.2.2 Paradigm design. The experimental designs for fMRI studies are either established as blocked or as event related designs. In the case of the blocked design, a specific activation condition and a baseline condition are presented continuously over a certain period of time. The underlying assumption is that the cognitive demand imposed by the activation condition will be added to the baseline activity without any significant interaction ('pure insertion'). Event related designs try to model the hemodynamic response and signal changes associated with individual trials.

3.2.3 Data processing and analysis.

(a) *Realignment*. The observed signal change in fMRI activation studies at 1.5 T are usually in the range of 1–5 percent. Postprocessing and statistical procedures enable the detection and interpretation of these signals. In order to avoid movement, physical head constraints such as headholders or bite-bars are used. Additionally, computerized algorithms for realignment of the acquired images are applied.

(b) *Spatial normalization*. Spatial normalization of the images to a standard template brain is often performed prior to statistical analysis to allow signal averaging across subjects and to report the resulting activation foci in terms of standard coordinates.

(c) *Statistics*. Various parametric and nonparametric methods are adopted for the analysis of fMRI datasets. In blocked paradigm designs, significant activation states in terms of differences between the baseline and the activation conditions can be detected by a t-test for each individual voxel. The time series of the ideal response function can be modeled for example by a box-car function and voxels which correlate significantly with this response function are regarded as activated. An extension of this statistical approach is based on the General Linear Model.

In event-related designs, a modeling of the individual hemodynamic response function is employed and those voxels that are best predicted by a particular response function in relation to the known onsets of the stimulus events will be identified.

Functional connectivity has been defined as the correlation of activity among brain regions. Additionally, the term 'effective connectivity' has been introduced; this is the influence an area has, directly or indirectly, on another area. Functional connectivity

can be described in terms of path coefficients (Büchel et al. 1999).

3.2.4 Findings.

(a) *Schizophrenia*

(i) *Verbal fluency*. Since disturbances in language perception or production are related to both positive and negative symptoms of schizophrenia, the investigation of neural circuits involved in language processing are likely to be of importance for the pathophysiology of schizophrenia. Functional activation studies using language activation paradigms lead to some new concepts of the underlying networks. It appears that the cerebral structures involved in language processing exceed the 'classical' language areas such as Broca's and Wernicke's area. Especially the extent of frontal, temporal, and parietal areas associated with spontaneous language production is noteworthy (Schlösser et al. 1998). Schizophrenic patients demonstrated significantly less left frontal activation and greater left temporal activation than normal controls during a word fluency task (Yurgelun Todd et al. 1996).

(ii) *Working memory*. There is a vast body of literature suggesting cognitive deficits in schizophrenia especially in the domains of attention, short-term memory, and executive functions. The fMRI technique has been used to demonstrate alterations in the activation pattern of higher cortical association areas and related subcortical neuronal networks in schizophrenic patients while performing working memory tasks. It has been demonstrated that schizophrenic patients revealed less activation in the frontal cortex than a group of healthy controls (Callicott et al. 1998).

(iii) *Emotional processing*. Schizophrenic patients as compared to normal controls did not demonstrate amygdala activation during induced sadness despite matched ratings to normal controls (Schneider et al. 1998).

(iv) *Hallucinations*. Aside from the application of external activation tasks or events, spontaneously occurring psychopathological symptoms could be directly visualized. An increase of the BOLD fMRI signal in Heschl's gyrus could be observed during hallucinations (Fig. 1(a)). Activation at the same location could also be identified during acoustic stimulation (Fig. 1(b)), suggesting the involvement of primary auditory areas in auditory verbal hallucinations (Dierks et al. 1999).

(b) *Obsessive-compulsive disorder*

Symptom provocation paradigms with individually tailored stimuli in patients with obsessive compulsive disorder led to defined activations in different brain areas. Whereas a patient group revealed activations in medial orbitofrontal, lateral frontal, anterior temporal, anterior cingulate, and insular cortex, as well as caudate, lenticulate, and amygdala, normal control subjects did not show any significant activations (Breiter et al. 1996). These results point towards a physiological overactivation of an orbitofrontal-caudate circuit.

(c) *Substance abuse*. Individually tailored cues have also been used to study different aspects of the functional neuroanatomy of the reward system in the context of drug craving. Cocaine-associated cues were applied to map activation in regions thought to be related to craving (e.g., prefrontal and cingulate cortices).

(d) *Dementia*. While performing an auditory stimulation task, predominant activation foci were seen in the inferior and middle frontal gyrus (left greater than right) in healthy controls but not in patients with mild AD. The latter group also showed additional activation suggesting compensatory recruitment of locally expanded foci and remote regions, for example, right frontal activation (Saykin et al. 1999).

(e) *fMRI and pharmacology*. There is considerable interest in using fMRI for the study of pharmacological influences on functional brain networks. The studies reported have either investigated the direct effect of a drug challenge on the BOLD signal or they have examined the modulatory influence of a pharmacological intervention on brain activation patterns.

After acute cocaine infusion, a negative signal change was noted in the amygdala, which correlated with craving ratings, whereas activation in the ventral tegmentum, pons, basal forebrain, caudate, cingulate, and most regions of lateral prefrontal cortex correlated with cocaine induced euphoria. The beneficial effects on brain activation patterns of treatment with atypical in comparison to typical antipsychotic drugs has been demonstrated both with motor stimulation (Braus et al. 1999) and a working memory task (Honey et al. 1999. (See *Schizophrenia*.)

3.3 Perfusion fMRI

In analogy to radioactive tracers in emission tomography, 'MR tracers' have been used to image brain perfusion. Two different approaches are currently adopted. The first is based on the application of paramagnetic contrast agents, such as gadolinium-DTPA. The second approach employs steady-state arterial spin labeling (ASL) techniques, which 'label' arterial blood as it moves up through the brain and then follow the effect of the labeled blood on the amplitude of the MRI signal from defined slices. Using a flow imaging technique, parieto-occipital and temporo-occipital to whole slice signal intensity ratios were reported to be significantly lower in AD patients than healthy controls. Additionally, parieto-occipital hypoperfusion correlated with dementia severity.

3.4 Diffusion-weighted fMRI

Diffusion-weighted images are MRI images with signal intensities sensitized to the random motion of

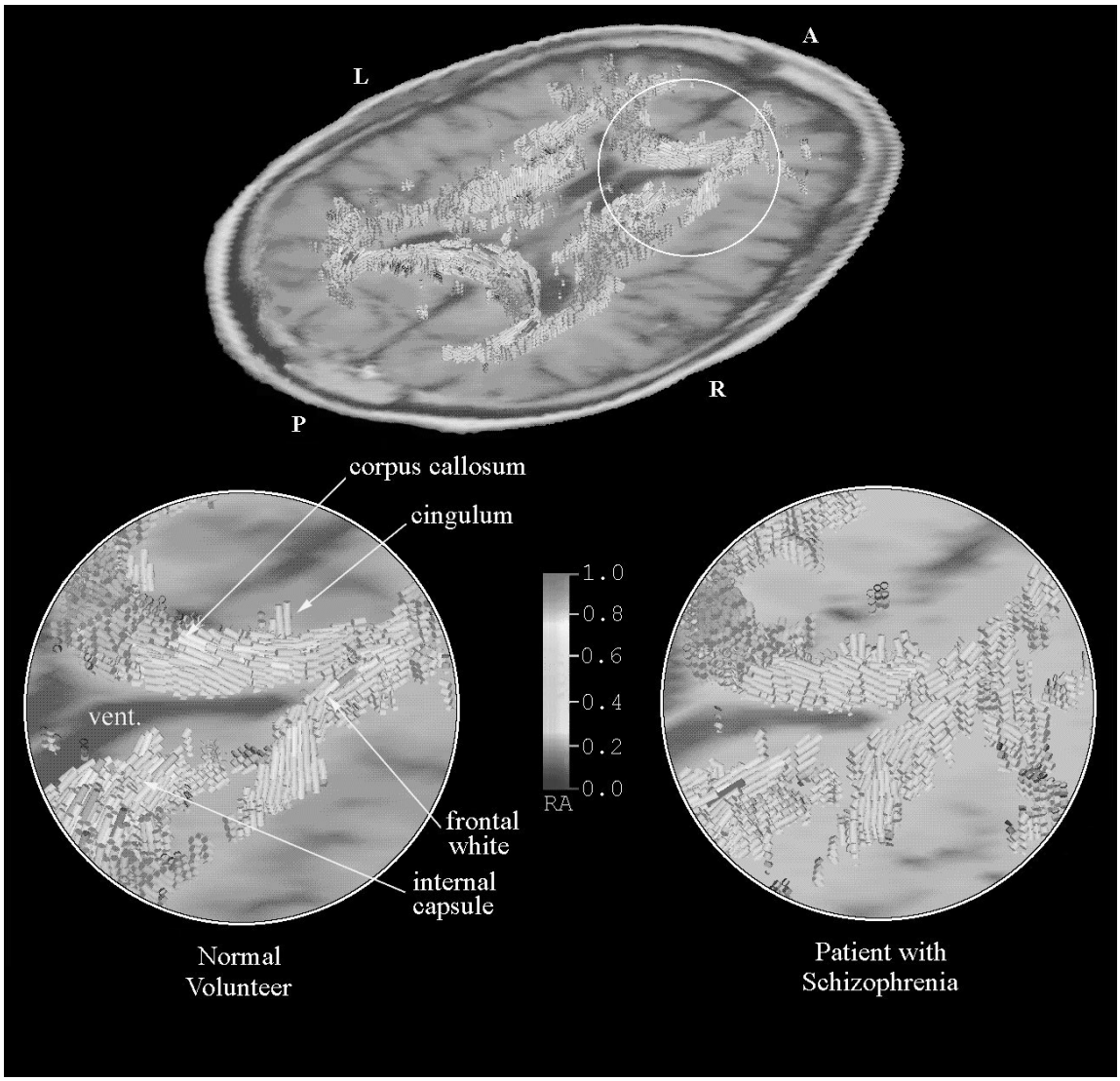

Figure 2
Diffusion tensor images with relative anisotropy represented as small 3-D color-coded tubes. Bottom left and right images show close-ups, in a normal volunteer and a schizophrenic patient, respectively, of the right frontal region where statistically significant anisotropy differences were found. (Reproduced with permission of Lippincott Williams & Wilkins. Buchsbaum et al. 1998.)

water molecules. In tissues that have a random microstructure, the measured apparent diffusion coefficient (ADC) seems to be the same in all directions (isotropic diffusion), while in tissues that have a regularly ordered microstructure, such as in parallel axons, the measured ADC varies with tissue orientation (anisotropic diffusion).

Diffusion tensor analysis allows advanced quantification of the directionality of diffusion anisotropy. The six independent scalar elements of the diffusion tensor contain the information required to characterize diffusion in all directions.

Applying diffusion weighted imaging to a psychiatric population, significantly lower diffusion anisotropy in the white matter of the prefrontal cortex in schizophrenic patients than in normal controls was observed (Buchsbaum et al. 1998). Figure 2 displays diffusion tensor images of a normal volunteer and a schizophrenic patient, respectively, illustrating these findings.

3.5 Magnetic Resonance Spectroscopy (MRS)

Aside from structural and functional imaging with MRI as described above, magnetic resonance spectroscopy (MRS) allows for the *in vivo* investigation of various biochemical substances including membrane compounds, substrates of energy metabolism, neurotransmitters, and also brain concentrations of pharmacological agents.

3.5.1 Technical foundations. The basis of MRS is widely the same as for MRI imaging as outlined above. As mentioned above, the different nuclei have a defined resonance frequency and as the nuclei reorientate themselves along the static field they send out a signal at this frequency, thus giving rise to the specific MRI signal. MRS is based on the fact that not all nuclei of one type resonate at the same frequency even if a homogeneous magnetic field without any externally applied gradients is present. This is due to the fact that the nuclei find themselves in small local variations of the magnetic field caused by differing chemical environments. This small variation in the resonant frequency is called chemical shift. Whereas MRI and fMRI do not specifically use this chemical shift information, with MRS the full range of different frequencies is usually resolved and displayed as a frequency spectrum with different peaks and valleys.

The most important applications of MRS in psychiatry are ^{1}H-MRS and ^{31}P-MRS. In ^{1}H-MRS, the peaks of the spectrum include N-acetyl aspartate (NAA), creatin-phosphocreatin (Cr-PCr), quaternary ammonium methyl groups (containing choline and other compounds), myo-inositol (containing myo-inositol monophosphate and glycine), glutamine, glutamate, aspartate, gamma-aminobutyric acid (GABA), and lactate. NAA is an amino acid present in high concentrations in neurons and is regarded as a neuronal marker although the exact role of the substance is not yet established. Since the signal of these molecules is embedded in the signal of abundant water molecules in the brain, the water signal has to be 'suppressed' by certain techniques.

^{31}P-MRS allows the measurement of brain membrane phospholipids and high-energy phosphate metabolism. The phosphorus spectrum includes the three peaks from adenosine triphosphate (ATP, beta-ATP and alpha-ATP), phosphocreatine (PCr), phosphodiesters (PDE), phosphomonoesters (PME), and inorganic phosphate (P_I). The peaks of PME and PDE are indicators of membrane synthesis and a decrease in the PME/PDE ratio has been suggested to be indicative of a decreased rate of phospholipid synthesis. Other nuclei which are of importance for MRS in psychiatry are ^{7}Li, ^{13}C, and ^{19}F.

3.5.2 Findings.

(a) *Schizophrenia.* Decreased levels of PME and P_I and increased levels of ATP and PDE in the dorsolateral prefrontal cortex could be found in drug-naïve first episode schizophrenic patients. Additionally, PME levels were significantly decreased in frontal regions of schizophrenic patients who had high scores on negative symptom rating scales. However, these and subsequent findings are not without controversy. Decreased levels of PDE in schizophrenic patients compared to normal controls were described (Volz et al. 1998b). Major improvements and changes in the MRS techniques including different localization methods might be possible explanations for the divergent findings.

Using a technique with multiple slices consisting of several single-volume elements, it could be demonstrated that NAA-relative signal intensity reductions in schizophrenia appear to be remarkably localized, involving primarily the hippocampal region and the dorsolateral prefrontal cortex (Bertolino et al. 1998).

Memory function in patients with schizophrenia were found to be related to temporal but not frontal Cr content as determined with MRS, a pattern that was not apparent among controls. Ratios of NAA to choline-containing compounds were reported to be significantly reduced in the bilateral basal ganglia regions of medicated schizophrenic patients compared to normal subjects, suggesting that neuronal dysfunction may be observable subcortical structures as well.

(b) *Affective disorders.* Bipolar I patients showed increased PME and intracellular pH levels in the manic or depressed state while the concentration of these metabolites was decreased in the euthymic state. Bipolar II patients had decreased PCr levels in the depressed, manic, and euthymic state. The observed changes were not related to cerebral ^{7}Li concentration as determined by MRS. In euthymic patients with bipolar disorder, the Cho/Cr-PCr peak ratio was higher than that in normal controls which could be interpreted as a result of a membrane breakdown that may occur in the basal ganglia of patients with bipolar disorder. In patients with unipolar major depression, increased PME and decreased ATP values in the frontal lobe could be detected (Volz et al. 1998a). With ^{1}H-MRS, decreased levels of NAA/Cr-PCr were found in the dorsolateral prefrontal cortex of bipolar patients, suggesting decreased neuronal density or neuronal dysfunction in the DLPFC region.

(c) *Dementia.* Alzheimer's disease has been shown to be associated with an increase in brain PME early in the course of the disease and followed by an increase in PDE and PCr later on. The increase in PME might reflect early abnormalities in membrane metabolism while the increase of PDE and PCr shows neuronal degeneration and death. Recent studies suggest an increase in the PME/PDE ratio of approximately 50 percent, but no differences in absolute values of PME

and PDE between patients with Alzheimer's diseases and normal controls could be established. However, in other studies in patients with mild to moderate AD, no differences in any phosphorus metabolite concentrations or ratios could be detected, nor any correlations of ^{31}P metabolite levels with severity of dementia.

With ^{1}H MRS, decreased levels of NAA and increased myo-inositol could be found. The change in myo-inositol might represent alterations of the polyphosphate second messenger system. Cognitive dysfunction of demented patients was significantly correlated with reductions of NAA, but not with increases of myo-inositol (Heun et al. 1997).

(d) *Treatment studies*. It could be demonstrated that the treatment response to lithium in manic patients with bipolar disorder is more closely related to the lithium concentration in the brain than to the concentration in serum.

It remains unclear whether neuroleptic treatment might cause additional changes in the brain metabolic patterns. Associations between the chlorpromazine equivalent neuroleptic dose received by patients and NAA contents in different brain structures have been described, suggesting a favorable long-term effect of neuroleptic treatment on neuronal function. In schizophrenic patients treated with atypical antipsychotics, decreased levels of Cho were found in the left basal ganglia, while increased levels of NAA were found in the left frontal cortex. Few MRS follow-up studies after neuroleptic treatment have been reported so far. A ^{1}H-MRS spectroscopy study demonstrated a renormalization of increased pretreatment (GABA+Glu)/Cr ratio after neuroleptic treatment which may implicate the recovery of normal neuronal function in neurotransmitters.

4. Outlook

In the past years, applications of MRI in psychiatry have experienced a tremendous growth both in research and clinical practice. The MRI imaging modalities reviewed above have to be regarded as complementary with respect to the acquired information. For example, combining fMRI and MRS will allow for the detection of metabolite patterns within areas of altered functional activation. fMRI in combination with DTI will help to address the issue of functional and effective connectivity by mapping both regional intercorrelations of BOLD signal changes and structural connections in terms of long fiber tracts. MRI-based methods will further benefit from multimodal imaging approaches in conjunction with imaging techniques based in the field of nuclear medicine or electrophysiology, such as positron emission tomography (PET), electroencephalography (EEG), or magnetoencephalography (MEG).

The availability of modern technologies such as MRI has to a significant extent contributed to the recent progress in the assessment of the biological basis of psychiatric disorders and the development of successful treatment strategies. MRI technology has already proved itself and will continue to be a valuable 'window to the brain.'

See also: Cortical Activity: Differential Optical Imaging; Differential Diagnosis in Psychiatry; Functional Brain Imaging; Magnetoencephalography; Mental and Behavioral Disorders, Diagnosis and Classification of; Syndromal Diagnosis versus Dimensional Assessment, Clinical Psychology of; Topographic Maps in the Brain

Bibliography

Bertolino A, Callicott J H, Elman I, Mattay V S, Tedeschi G, Frank J A, Breier A, Weinberger D R 1998 Regionally specific neuronal pathology in untreated patients with schizophrenia: a proton magnetic resonance spectroscopic imaging study. *Biological Psychiatry* **43**: 641–8

Braus D F, Ende G, Weber-Fahr W, Sartorius A, Krier A, Hubrich-Ungureanu P, Ruf M, Stuck S, Henn F A 1999 Antipsychotic drug effects on motor activation measured by functional magnetic resonance imaging in schizophrenic patients. *Schizophrenia Research* **39**: 19–29

Breiter H C, Rauch S L, Kwong K K, Baker J R, Weisskoff R M, Kennedy D N, Kendrick A D, Davis T L, Jiang A, Cohen M S, Stern C E, Belliveau J W, Baer L, O'Sullivan R L, Savage C R, Jenike M A, Rosen B R 1996 Functional magnetic resonance imaging of symptom provocation in obsessive-compulsive disorder. *Archives of General Psychiatry* **53**: 595–606

Büchel C, Coull J T, Friston K J 1999 The predictive value of changes in effective connectivity for human learning. *Science* **283**: 1538–41

Buchsbaum M S, Tang C Y, Peled S, Gudbjartsson H, Lu D, Hazlett E A, Downhill J, Haznedar M, Fallon J H, Atlas S W 1998 MRI white matter diffusion anisotropy and PET metabolic rate in schizophrenia. *Neuroreport* **9**: 425–30

Callicott J H, Ramsey N F, Tallent K, Bertolino A, Knable M B, Coppola R, Goldberg T, van Gelderen P, Mattay V S, Frank J A, Moonen C T, Weinberger D R 1998 Functional magnetic resonance imaging brain mapping in psychiatry: methodological issues illustrated in a study of working memory in schizophrenia. *Neuropsychopharmacology* **18**: 186–96

Chua S E, McKenna P J 1995 Schizophrenia—a brain disease? A critical review of structural and functional cerebral abnormality in the disorder. *British Journal of Psychiatry* **166**: 563–82

DeLisi L E, Sakuma M, Tew W, Kushner M, Hoff A L, Grimson R 1997 Schizophrenia as a chronic active brain process: a study of progressive brain structural change subsequent to the onset of schizophrenia. *Psychiatry Research* **74**: 129–40

Dierks T, Linden D E J, Jandl M, Formissano E, Goebel R, Lanfermann H, Singer W 1999 Activation of Heschl's gyrus during auditory hallucinations. *Neuron* **22**: 615–21

Drevets W C, Ongur D, Price J L 1998 Neuroimaging abnormalities in the subgenual prefrontal cortex: implications for the pathophysiology of familial mood disorders. *Molecular Psychiatry* **3**: 220–6

Heun R, Schlegel S, Graf Morgenstern M, Tintera J, Gawehn J, Stoeter P 1997 Proton magnetic resonance spectroscopy in dementia of Alzheimer type. *International Journal of Geriatric Psychiatry* **12**: 349–58

Honey G D, Bullmore E T, Soni W, Varatheesan M, Williams S C, Sharma T 1999 Differences in frontal cortical activation by a working memory task after substitution of risperidone for typical antipsychotic drugs in patients with schizophrenia. *Proceedings of the National Academy of Science USA* **96**: 13432–7

Mazziotta J C, Toga A W, Frackowiak R S J (eds.) 2000 *Brain Mapping. The Disorders*. Academic, San Diego, CA

Murphy D G M, DeCarli C D, Daly E, Gillette J A, McIntosh A R, Haxby J V, Teichberg D, Schapiro M B, Rapoport S I, Horwitz B 1993 Volumetric magnetic resonance imaging in men with dementia of the Alzheimer type: correlations with disease severity. *Biological Psychiatry* **34**: 612–21

Saykin A J, Flashman L A, Frutiger S A, Johnson S C, Mamourian A C, Moritz C H, O'Jile J R, Riordan H J, Santulli R B, Smith C A, Weaver J B 1999 Neuroanatomic substrates of semantic memory impairment in Alzheimer's disease: patterns of functional MRI activation. *Journal of the International Neuropsychology Society* **5**: 377–92

Schlösser R, Hutchinson M, Joseffer S, Rusinek H, Saarimaki A, Stevenson J, Dewey S L, Brodie J D 1998 Functional magnetic resonance imaging of human brain activity in a verbal fluency task. *Journal of Neurology Neurosurgery and Psychiatry* **64**: 492–8

Schneider F, Weiss U, Kessler C, Salloum J B, Posse S, Grodd W, Muller Gartner H W 1998 Differential amygdala activation in schizophrenia during sadness. *Schizophrenia Research* **34**: 133–42

Toga A W, Mazziotta J C (eds.) 2000 *Brain Mapping. The Systems*. Academic, San Diego, CA

Volz H P, Rzanny R, Riehemann S, May S, Hegewald H, Preussler B, Hubner G, Kaiser W A, Sauer H 1998a 31P magnetic resonance spectroscopy in the frontal lobe of major depressed patients. *European Archives of Psychiatry and Clinical Neuroscience* **248**: 289–95

Volz H P, Rzanny R, Rossger G, Hubner G, Kreitschmann Andermahr I, Kaiser W A, Sauer H 1998b Phosphorus magnetic resonance spectroscopy of the dorsolateral prefrontal region in schizophrenics—a study including 50 patients and 36 controls. *Biological Psychiatry* **44**: 399–404

Yurgelun Todd D A, Waternaux C M, Cohen B M, Gruber S A, English C D, Renshaw P F 1996 Functional magnetic resonance imaging of schizophrenic patients and comparison subjects during word production. *American Journal of Psychiatry* **153**: 200–5

R. Schlösser

Muller, Hermann Joseph (1890–1967)

Geneticist Hermann Joseph Muller [or H. J. Muller as he identified himself in his articles] was born in New York on December 21, 1890. His Catholic paternal grandparents had emigrated to the United States from Coblenz in the Moselle Valley of Western Germany after their unsuccessful support of the Revolution of 1848. Muller's first cousins also became well-known: Alfred Kroeber as an anthropologist, and Herbert J. Muller as a literary critic. Muller's maternal grandparents were partly Jewish, partly Episcopalian and had emigrated from England. Muller was raised as a Unitarian but considered himself an atheist or humanist. Muller was educated at Columbia College, receiving a BA in 1910 and his Ph.D. in 1916. He was recruited by Julian Huxley, then the first Chair of the Zoology Department at the Rice Institute. From 1918 to 1920 he filled in for his mentor, Thomas Hunt Morgan while Morgan was on sabbatical from Columbia. In 1920 Muller joined the faculty at the University of Texas, where he did his Nobel prize experiments, and then left in 1932 for an eight-year sojourn in Europe, where he spent a Guggenheim Fellowship year in Berlin, became a visiting investigator at Leningrad (1933–4), headed a laboratory in Moscow (1935–7), and then spent a year in Spain as a volunteer during the Spanish Civil War. He was a guest investigator at the University of Edinburgh from 1937 to 1940 and then returned to the United States as an *ad interim* professor at Amherst College for the war years. He was recruited by Indiana University in 1945, won the Nobel Prize in medicine in 1946, and spent the rest of his career, until retirement in 1964, at Indiana University. He died in Indianapolis, Indiana, April 5, 1967. He was married first to Jessie (née Jacobs) Muller, a professor of mathematics by whom he had a son David, who also became a professor of mathematics. He was married a second time to Dorothea [Thea] (née Kantorowicz) Muller, a physician. Their daughter, Helen, became a professor of public health (Carlson 1982).

1. Major Contributions to Genetics

Muller's most noted contribution was the induction of mutations in fruit flies using x-rays (Muller 1928); this was the sole research organism of his career. Others had tried earlier, with no success or with ambiguous results. It was Muller's gift for designing genetic stocks that clinched the case and opened the field of radiation genetics. Muller was also well regarded for contributing, with Morgan, Alfred Henry Sturtevant and Calvin Blackman Bridges to the founding of classical genetics, that body of work that brought together the theory of the gene, the chromosome theory of heredity, breeding analysis, cytogenetics, and genetic mapping, along with explanations for new phenomena such as X-linked inheritance and nondisjunction of chromosomes.

Muller considered his greatest achievement to be the recognition of the gene as the basis of life through its unusual property of copying its errors, offering this as a mechanistic basis for the variations that make

evolution possible. He also clarified what he called gene-character relations, showing by experimental means that complex character traits in fruit flies, such as variable wing shape, could be analyzed successfully into their genetic and environmental components (Altenburg and Muller 1920). Muller devoted considerable attention to estimating the size and shape of genes, the number of genes in a chromosome or in an organism's genome, and the properties of genes when studied through their various mutations. He made substantial contributions to human genetics, including estimates of genetic damage from mutations, the amount of mutations on average each person carries, and the fate of those mutations over generations (Muller 1962).

2. *Social Controversies over Genetics*

Muller was not typical of his generation of scientists. He took a position on social issues that arose from the impact of genetics on society. He is perhaps best known for first raising a concern about the harmful effects of radiation on a population exposed to different doses of radiation (Muller 1957). He interpreted the symptoms of radiation sickness from the Hiroshima and Nagasaki bombings as a consequence of induced chromosome breaks leading to a cycle of breakage, reunion of broken pieces, and subsequent re-breakage resulting in cell death, a phenomenon he had independently worked out in fruit flies while in Edinburgh and that Barbara McClintock had worked out in maize while then at the University of Missouri. Muller fought vigorously to protect the public from the hazards of nuclear radiation at a time when those in the health professions, industry, and the military sought to ignore all except the highest doses of exposure to radiation.

During the Cold War he was accused by the Soviet Union of having been an American spy sent to wreck Soviet genetics, at the same time that he was regarded by the Atomic Energy Commission and the Eisenhower administration as a Soviet spy sent back to the US to wreck the development of the hydrogen bomb program by alarming the public over the dangers of nuclear radiation. Muller had debated with Trofim D. Lysenko in 1936 in Moscow and had had to flee the USSR after denouncing Lysenko as a charlatan. Muller's dubious distinction on both sides of the Cold War stems from his initial attraction to socialism and communism (although he never joined the Communist Party), and his denunciation of communism after his disillusionment, having seen what life was like under the Stalinist USSR. Although he loathed Stalinism he never abandoned the ideals of socialism by which social justice would be part of society rather than left to chance.

3. *Muller's Conflicting Appraisals of Eugenics*

Muller is least well regarded for his views on eugenics. Muller was idealistic and shared many of Francis Galton's views on the need for a positive eugenic program in which the ablest males would be encouraged to produce a more than average number of children. Muller advocated this through what he later called germinal choice, which would use frozen sperm retained for a generation before being thawed and used (Muller 1961). He opposed the American eugenics movement that was led by Charles Benedict Davenport and Harry Laughlin in the United States as racist, sexist, and elitist. He denounced that movement in a speech at the third, and last, International Congress of Eugenics held in New York in 1932 (Muller 1934). Muller had naively believed that in the USSR he would find a society in which merit, not class, would determine who were the genetically well-endowed individuals who could contribute to sperm banks. He published a popular book, *Out of the Night* (1935), and had it translated into Russian and sent it to Premier Stalin. The arrest and execution of two of Muller's students, and the hostile response of the pro-Lysenko audience at the debates over Soviet genetics, made him seek a graceful exit by joining the International Brigade.

Muller failed all his life to convince either scientists or the public that eugenics was necessary to compensate for the accumulation of mutations in human populations that were no longer subject to the pressures of natural selection (see Muller 1950). However, he shunned coercive eugenics policies and those based on compulsory sterilization of what were then called 'the unfit.'

4. *Muller's Philosophical Views on Science and the Gene*

Muller's genetic insights on the gene were resisted in his own day, when he was considered a zealot and the term genetic determinism had not yet been coined. His view of the gene as the basis of life has permeated the thinking of molecular biologists who see in DNA the same basis for life as he did in the gene (see Muller 1929). By 'the gene as the basis of life' Muller intended scientists to recognize that all other components of life are made by genes and that only genes had the capacity to retain their copying capacity, despite being altered by the mutation process. This phrase also implied that all of evolutionary life could be traced back to the first replicating molecule with this capacity. Muller never ignored the environment in the expression of genetic traits. He realized that a gene could be modified by other genes, and by such environmental components as temperature, humidity, and nutrition; this was evidenced in the various traits he studied in fruit flies.

He made similar arguments for the expression of genetic diversity conditioned by genetic modifiers as well as environmental factors, both uterine and extra-uterine, in the expression of human traits such as intelligence, creativity, and personality. He was one of the first to study a pair of separated identical twins (in 1925) and to recognize how difficult it would be to isolate all the major factors, genetic and environmental, that distinguished the careers, aptitudes, and lives of these two subjects (Muller 1925).

Muller's philosophy of science was that of a reductionist. He believed science could interpret nature, that there was an external reality, and that both theory and experimentation were essential for science to be effective. He rejected mysticism, holism, vitalism, and the belief that scientific reality was a social construction. At the same time he was aware of the biases that society could impose on complex issues often thought by non-scientists to be 'scientific', and he resisted and rebuked those who believed that intelligence testing was free of bias or that eminence or success in the financial world (or in a capitalist society in general) was largely a matter of merit rather than privilege and unequal opportunities. These views made Muller seem, to his critics, contradictory, a turn-coat, paranoid, or muddled in his thinking.

See also: Behavioral Genetics: Psychological Perspectives; Epidemiology, Social; Ethical Issues in the 'New' Genetics; Eugenics as the Basis of Population Policy; Eugenics in Asia; Eugenics in Europe; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionary Epistemology; Evolutionary Selection, Levels of: Group versus Individual; Genetic Factors in Cognition/Intelligence; Genetic Screening for Disease-related Characteristics; Genetic Studies of Behavior: Methodology; Genetics of Complex Traits Through the Life Cycle; Human Genome Diversity and its Value; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation

Bibliography

Altenburg E, Muller H J 1920 The genetic basis of truncate wing—an inconstant and modifiable character in *Drosophila*. *Genetics* **5**: 1–59

Carlson E 1982 *Genes, Radiation, and Society: The Life and Work of H J Muller*. Cornell University Press, Ithaca, NY

Muller H J 1925 Mental traits and heredity. *Journal of Heredity* **16**: 433–48

Muller H J 1928 The problem of genic modification. In: *Verhandlungen des V. Internationalen Kongresses fur Vererbungswissenschaft, Berlin, 1927*. Supplementband I der *Zeitschrift fur Induktive Abstammungs-und Vererbungslehre*, pp. 234–60

Muller H J 1929 The gene as the basis of life. In: *Proceedings of the International Congress of Plant Sciences, August* 1926. Ithaca, NY, pp. 897–921

Muller H J 1934 The dominance of economics over eugenics. In: *A Decade of Progress in Eugenics: Proceedings of the Third International Congress of Eugenics, New York*. Williams and Wilkins, New York, pp. 138–44

Muller H J 1935 *Out of the Night*. Vanguard Press, New York

Muller H J 1950 Our load of mutations. *American Journal of Human Genetics* **2**: 111–76

Muller H J 1957 Damage from point mutations in relation to radiation dose and biological conditions. In: *Effect of Radiation on Human Heredity*. World Health Organization, Geneva, pp. 25–47

Muller H J 1961 Human evolution by voluntary choice of germ plasm. *Science* **134**: 643–9

Muller H J 1962 *Studies in Genetics*. Indiana University Press, Bloomington, IN

E. A. Carlson

Multi-attribute Decision Making in Urban Studies

Multi-attribute decision making refers to situations in which several measures of effects or consequences are taken into account in choosing what to do. Planners confront such situations in choosing among projects, plan alternatives, or policies and in assessing environmental consequences or land use suitability. It is also important to distinguish multi-attribute decision making—a single decision maker considering multiple attributes—from collective choice—a collection of decision makers with different preferences reaching a common choice. Multi-attribute decision theory is highly developed, but is seldom fully applied in planning practice. This article identifies the theory, current contenders, and one way in which the theory is applied in practice.

1. Prescriptive Decision Theory

The most broadly accepted prescriptive decision theory is Multi-attribute Utility Theory (MUT). It builds on utility theory from economics, statistics, and operations research. Differences between observed decision making and the prescriptions of MUT suggest specific opportunities to make better decisions. The arguments that MUT requires too much information and is too difficult to understand emphasize the need to derive operational methods from theory to reduce information requirements and ease application.

1.1 Multi-attribute Utility Theory

A multi-attribute problem can be described as a matrix in which rows are alternatives, columns are attributes, and each cell is a measure of the effect of an alternative with respect to an attribute. To illustrate, the decision

might be to choose among three sites (A, B, and C) for a landfill by considering four attributes: costs to the local government in dollars, number of persons displaced, index of groundwater contamination, and capacity of site in cubic meters. Each of the three sites is described in terms of these measures. The problem is then to choose the best alternative, based on the attributes. Hill (1968) and Litchfield et al. (1975) are classic descriptions of such matrices for use in planning. Attributes should be chosen so as to be independent of each other in value terms, meaning that the value of a level on one attribute does not depend on the level of another attribute. Carefully designed questions can, however, elicit valid value judgments under certain types of attribute interdependence (Keeney and Raiffa 1976).

Choosing an alternative requires two types of judgments from the decision maker: transformations of effects into value measures, and transformations of value measures for each attribute into a common value measure across all attributes. The theory behind such judgments is most thoroughly developed in Keeney and Raiffa (1976).

The value-scale transformation recognizes that measures of effects are not the same as measures of value. The unit of value can be set arbitrarily, often by setting the difference between the lowest level of an effect and the highest level of that effect equal to a value of 1.0 unit. This unit of value is valid for the entire range of the effect only if the value transformation is linear (a constant rate of change in value with change in effect). For example, a decision maker might judge that an increase in the groundwater contamination index yields a constant decrease in value regardless of the level of contamination. More likely, a decision maker would decide that when the index gets very high, there is little further change in value because the groundwater is already so contaminated that it makes little difference. In this latter case the transformation is nonlinear, and the decision maker must estimate a function that describes the rate of change in value with respect to the rate of change in the index for each level of the index. In practice, such functions are estimated from a sample of levels of the index and extrapolated for other levels. Other methods for such judgments are described below.

To choose the best alternative, considering all attributes, the units of value for each attribute must be transformed to a common unit of value. Although a decision maker might try to judge each attribute directly in the same unit of value, this is extremely difficult to do and is confounded by the usual case in which the value transformation is nonlinear. Instead, decision makers are asked to assess weights that transform the units of value judged for each attribute into a common unit.

Multi-attribute utility theory is consistent with utility theory incorporating uncertainty. Thus value transformations can be used in conjunction with subjective expected utilities to consider uncertain outcomes with respect to alternatives. For a textbook treatment of decision analysis in these terms see Kirkwood (1997). Value scales can be estimated so as to take into account variation among decision makers in relative aversion to taking risks on outcomes.

1.2 Prescriptive Theory and Behavioral Theory

Over the past 20 years, research in psychology, most prominently by Kahneman and Tversky (e.g., Kahneman et al. 1982), has identified several biases in the way people make decisions. Biases are distortions from what a decision maker would want to have chosen after fully reflecting on the character of the bias in a particular situation, but the biases persist even for people who are well aware of them as general behaviors. These biases are very predictable in form and direction of error, though not in precise quantity.

Recognizing these biases allows decision makers to follow routines that are likely to counter them (von Winterfeldt and Edwards 1986). For example, estimates expressed numerically tend toward previously considered numbers (called 'anchoring and adjustment'). A counter to this bias would be to have the decision maker estimate value tradeoffs more than once and ask the tradeoff questions in different sequences each time.

1.3 Contending Approaches

Other approaches to multi-attribute decision making are based on fundamentally different theories (Lai and Hopkins 1989). The most salient current contender is the Analytic Hierarchy Process (AHP). The Analytic Hierarchy Process (AHP), developed by Saaty (1980), uses pairwise comparison questions to elicit a matrix of judgments of the relative preference between each pair of alternatives with respect to each attribute, and a matrix of judgments of the relative importance of each pair of attributes. For each matrix, the eigenvector associated with the maximum eigenvalue is the set of value scores for alternatives or the set of weights for the attributes. These two types of matrices correspond to the transformation of effects into value scales and to the transformation of value scales to a common unit. Saaty's explanation of an underlying theory of judgment for AHP is rejected by those working in the MUT tradition (Dyer 1990) and defended by advocates of AHP (Harker and Vargas 1990). The debate centers on the meaning of weights in AHP and some implications of these meanings: the possibility of rank reversals among previously included alternatives in AHP when additional alternatives are considered, and the problem of assessing weights based on vague notions of importance without regard to units of measure of effects or units of value.

Applications of AHP continue to appear in the planning literature. One of the attractions of AHP is that the judgments required seem to be easier to make, but this occurs, at least in part, because the difficulties of nonlinear value transformations, interdependent attributes, and careful consistency among measurement units are submerged and ignored. One of the major dilemmas in multi-attribute decision making is finding techniques that people will use but that do not submerge the actuality of hard tradeoff questions.

2. Methods in Use

Fully developed applications of MUT that justify value functions and weights and include uncertainty and sensitivity analysis seldom occur in urban planning. Kirkwood (1997) gives examples of applications in related fields. Most applications in planning make the simplifying assumptions of linear value functions and independent attributes, but without justifying them. Making these assumptions results in 'rating and weighting,' in which a unit of value for each attribute and weights for each attribute are sufficient to arrive at aggregate values measured in a common unit.

Complete estimation of value functions and weights for MUT or AHP is a daunting task. In most decision situations, however, good choices can be made without estimating the value functions and weights. Instead estimates can be made of only those aspects of these functions necessary to distinguish among available alternatives. 'Computation of Equivalent Alternatives' (CEA) (Stokey and Zeckhauser 1978), which is also called 'even swaps' (Hammond et al. 1999) takes advantage of this idea.

Starting from a multi-attribute matrix of alternatives, CEA focuses on making a small number of tradeoff judgments just sufficient to discover an alternative that dominates all others. The approach has the distinct advantages of requiring significantly fewer hard tradeoff judgments, while at the same time keeping the concrete performance measures salient and allowing for valid value transformations, even when attributes are nonlinear with respect to value and attributes are interdependent. The CEA procedure is a sequence of judgments that creates hypothetical alternatives that are equivalent in value to available real alternatives but have different combinations of levels on attributes.

For example, consider an instance in which alternative A is preferred to alternative B on the cost attribute but not preferred on the groundwater contamination attribute. Ask the decision maker to imagine a hypothetical alternative that is equivalent in value to alternative A, but has a cost equal to the cost of alternative B. For the hypothetical alternative to be equivalent in value to alternative A while increasing the cost, the contamination index of alternative A must be improved by an amount judged by the decision maker. If this improvement in the contamination index is less than the difference in contamination index between alternatives A and B, then B is preferred to the hypothetical alternative because they have equal costs, and B is preferred on ground water contamination. And, because the hypothetical alternative was created as equivalent in value to A, B is also preferred to A. This procedure can be repeated to discover preferred alternatives among many alternatives with many attributes, but without specifying complete value transformation functions or weights. The judgments are always made for specific combinations of levels of attributes so they do not assume linear value transformations or independence of attributes. Lee and Hopkins (1995) suggest tools to reduce the number of judgments necessary in applying CEA.

3. Problems and Future Work

The dilemma for multi-attribute decision making is how to devise methods that are easy enough to use to achieve common use in practice without creating methods that hide the choices rather than help decision makers face choices. Future research will focus on efficient techniques for framing such judgments for decision makers so as to ensure validity, require as small a number of judgments as possible, counter persistent biases, and give decision makers confidence that they understand what they are doing as they use the techniques.

See also: Decision-making Systems: Personal and Collective; Planning Ethics; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Strategic Planning

Bibliography

Dyer J S 1990 Remarks on the analytic hierarchy process. *Management Science* **36**: 249–58

Hammond J S, Keeney R L, Raiffa H 1999 *Smart Choices: A Practical Guide to Making Better Decisions*. Harvard Business School Press, Boston, MA

Harker P T, Vargas L G 1990 Reply to 'Remarks on the analytic hierarchy process' by J. S. Dyer. *Management Science* **36**: 269–73

Hill M 1968 A goals-achievement matrix for evaluating alternative plans. *Journal of the American Institute of Planners* **34**: 19–29

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York

Keeney R, Raiffa H 1976 *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York

Kirkwood C W 1997 *Strategic Decision Making: Multiobjective Decision Analysis with Spreadsheets*. Wadsworth, Belmont, CA

Lai S-K, Hopkins L D 1989 The meanings of tradeoffs in multi-attribute evaluation methods: A comparison. *Environment and Planning B: Planning and Design* **16**: 155–70

Lee I, Hopkins L D 1995 Procedural expertise for efficient multiattribute evaluation: A procedural support strategy for CEA. *Journal of Planning Education and Research* **14**: 255–68
Litchfield N, Kettle P, Whitbread M 1975 *Evaluation in the Planning Process*. Pergamon Press, Oxford, UK
Saaty T L 1980 *The Analytic Hierarchy Process*. McGraw-Hill, New York
Stokey E, Zeckhauser R 1978 *A Primer for Policy Analysis*. W. W. Norton, New York
von Winterfeldt D, Edwards W 1986 *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, UK

L. D. Hopkins

Multicultural Education

1. The Goals and Scope of Multicultural Education

Multicultural education is an educational reform movement whose major aim is to change schools, colleges, and universities so that students from diverse racial, ethnic, cultural, language, social class, and religious groups will experience equal educational opportunity. A major tenet of multicultural education is that students whose cultural and language characteristics are highly consistent with those of mainstream educational institutions are more likely to experience academic success than students whose cultural and language characteristics conflict with those of mainstream institutions.

Another important goal of multicultural education is to help students from all cultural and social groups, including mainstream groups, to develop democratic racial and ethnic attitudes so that they can function effectively in culturally diverse democratic societies. Most nation-states and societies throughout the world are characterized by cultural, ethnic, and language diversity. One of the challenges to pluralistic democratic nation-states is to provide opportunities for cultural and ethnic groups to maintain components of their community cultures while at the same time constructing a nation-state in which diverse groups are structurally included and to which they feel allegiance. This challenge is becoming intensified as democratic nation-states such as the USA, Canada, Australia, and the UK become more diversified, and as racial and ethnic groups within these nations become involved in cultural and ethnic revitalization movements. An important aim of multicultural education is to help future citizens develop the knowledge, attitudes, and skills needed to function effectively within their ethnic and cultural communities as well as within the nation-state. Kymlicka (1995) describes the opportunity for citizens to function both within their cultural communities and within the national civic culture as 'multicultural citizenship.'

There is consensus among multicultural researchers and theorists about overarching goals. However, there is less agreement, in both practice and theory, about which ethnic, social, and cultural groups should be included in educational reforms related to diversity. The groups that are the focus of multicultural education vary to some extent across nations and across regions within nations. When multicultural education emerged in nations such as the USA, the UK, and Canada in the 1970s and 1980s, the focus was on ethnic- and language-minority groups, on immigrant groups, and on groups that had been historically victimized by institutionalized discrimination and racism. As the field of multicultural education has developed, especially within the USA, increased emphasis has been given to issues related to gender and social class. A few theorists in the USA include sexual orientation in their conceptualization of multicultural education. However, it receives little attention in most multicultural education courses, programs, and projects. In most nations, such as the USA, Canada, Germany, Australia, and the UK, multicultural education usually focuses on racial, ethnic, and language groups (Banks and Banks 1995). Gender, class, and special education are receiving increased attention in some societies.

2. The Historical Development of Multicultural Education

Multicultural education emerged in response to the ethnic revitalization movements and immigration that occurred in the 1960s and 1970s in Western democratic nations such as the USA, Canada, the UK, France, Germany, and Australia. When these movements emerged, the Western nations were characterized by ethnic, cultural, language, and racial diversity. This diversity resulted from several historical developments. The nations in Western Europe had long-standing linguistic and cultural minorities, such as the Basques in France and Spain, the Germans in Denmark, and the Danes and Sorbes in Germany.

The newer nations such as the USA, Canada, and Australia were populated by many different racial, ethnic, cultural, and religious groups from Europe. The ethnic and cultural diversity within these new nations was enriched by the native peoples that the Europeans displaced, such as Indians in the USA and Canada, the Aborigines in Australia, the Maori in New Zealand, and by Blacks from Africa. Large numbers of immigrants and refugees from nations throughout the world settled in these nations in search of economic opportunity as well as political and religious freedom.

Ethnic and cultural diversity increased in European nations such as the UK, France and The Netherlands after World War II as groups from their former colonies immigrated to these nations to improve

their economic conditions. Significant numbers of immigrants from the West Indies, India, and Pakistan settled in the UK. Many Algerian immigrants, as well as immigrants from Cambodia, Laos, and Vietnam, settled in France. Immigrants from Morocco and Surinam settled in The Netherlands. In the post-World War II period many migrant workers from the Mediterranean nations in Europe settled in Western European nations such as France, the UK, Germany, and Switzerland, in search of greater economic opportunities. They provided desperately needed unskilled labor that was not available in these nations.

3. *The Rise of Ethnic Revitalization Movements*

The democratic ideologies institutionalized within the major democratic Western nations and the wide gap between these ideals and realities were major factors that resulted in the rise of ethnic revitalization movements in nation-states such as the USA, Canada, and the UK. These nations share a democratic ideal, a major tenet of which is that the state should protect human rights and promote equality and the structural inclusion of diverse groups into the fabric of society. These societies are also characterized by widespread inequality and by racial, ethnic, and class stratification. The discrepancy between democratic ideals and societal realities, and the rising expectations of structurally excluded racial, ethnic, and social-class groups, created protest and revival movements within the Western democratic nations during the 1960s and 1970s.

The civil rights movement of the 1960s and 1970s in the USA was one of the most significant ethnic revitalization movements. It echoed throughout the world. Ethnic revitalization movements also occurred in Canada, Australia, and the UK. A major goal of the civil rights movement in the USA was to improve the academic achievement of African American students in US schools, colleges, and universities. Leaders of the civil rights movement demanded more African American teachers, changes in textbooks and curriculum that would incorporate more content about African American history and culture, and more African American teachers and college professors.

The success of the African American led civil rights movement encouraged other ethnic-minority groups in the USA and throughout the world to start similar protest movements and to make similar demands. Consequently, schools, colleges, and universities responded, with varying levels of success and effectiveness, to the demands made by Mexican Americans, American Indians, Asian Americans, and other ethnic minorities in the USA. In Australia, Aboriginal groups participated in social protest for equal rights. Blacks in the UK used similar strategies to focus attention on the problems of West Indian students in British schools. A report that highlighted the over-representation of Black students in schools for the 'educationally sub-normal' (Coard 1971) was especially important in stimulating action in some schools, colleges, and Local Education Authorities (LEAs) in the UK.

4. *The Development of Paradigms and Explanations*

Many of the ethnic and linguistic minority groups who are indigenous to the Western nation-states or who are immigrants experience significant academic problems in the schools of their societies. There are notable exceptions to this observation, such as the Chinese, Japanese, and Cubans in the USA, the Asian Indians in the UK, and the Chinese and Greeks in Australia. A number of explanations have been developed by scholars and researchers to explicate why ethnic- and language-minority students often fail to achieve academic and social success in schools and to guide educational reform. These explanations are considered *paradigms* in this article because they have policy, programmatic, and research implications. A paradigm consists of the laws, principles, explanations, and theories of a discipline or field of study. It guides policy and action (Banks 2001).

The discussion that follows describes four influential paradigms and explanations that have been constructed by theorists and researchers to explicate the low academic achievement of low-income and ethnic-minority students since the field of multicultural education emerged in the 1970s. Another goal of these explanations is to guide school reform that will enable students from diverse groups to experience academic and social success. The explanations and paradigms that have been constructed by educators and social scientists are influenced in complex ways by their life experiences, values, and personal biographies. Consequently, the explanations that have been constructed about minority student achievement differ in important ways in part because of the social, political, and economic locations of the researchers who produced them. In other words, the explanations reflect the positionality of the researchers who constructed them.

4.1 *The Cultural Deprivation or Culture of Poverty Explanation*

Cultural deprivation and culture of poverty theorists believe that low-income students achieve poorly in school because the socialization in their families and communities does not equip them with the knowledge, skills, attitudes, and cultural capital essential for academic success in mainstream society. Unlike genetic theorists such as Herrnstein and Murray (1994), who believe that low-income students and ethnic-minority students do not achieve well in school because of their genes, these theorists believe that low-

income students can achieve if they are provided with early childhood experiences that will compensate for their family and community socialization.

A cultural deprivation curriculum intervention prototype is the DISTAR program developed by Bereiter and Englemann (1966). It is designed to help low-income students develop reading and writing skills using behavioral modification teaching techniques. *The Culturally Deprived Child* by Riessman (1962) was an influential book that exemplifies the cultural deprivation explanation. The culture of poverty concept is epitomized in *La Vida: A Puerto Rican Family in the Culture of Poverty* by Lewis (1965). The cultural deprivation and culture of poverty explanation has re-emerged today as 'children at risk.'

Critics of the cultural deprivation/culture of poverty explanation describe how it essentializes the characteristics of specific groups, blames the victims for their marginalized status, and does not focus on the ways in which these groups are victims of political and socioeconomic structures. It focuses on changing students rather than changing schools or the socio-political structure.

4.2 The Cultural Difference Explanation

These theorists challenge the genetic and cultural deprivation explanations. They argue that low-income students and ethnic-minority students have rich cultures. However, these students do not achieve well in school because of the cultural discontinuity between the school and the home. Consequently, to improve the academic and social achievement of ethnic-minority and low-income students, the school should be reformed in ways that draw upon the strengths of these students. Their languages, cultural styles, and ways of communicating should be incorporated into the curriculum, into teaching, and into the school culture.

Philips (1983) found that American Indian students participated more frequently in class discussions when instruction reflected the group-oriented participation structures used in Indian communities. In her research in the Kamehameha Early Education Program in Hawaii, Au (1980) found that the reading achievement of Native Hawaiian students was significantly increased when teachers incorporated traditional Hawaiian story structures into their teaching.

Lee (1993) found that the academic achievement of African American students increased when teachers incorporated African American language components and styles into instruction. Research summarized in August and Hakuta (1998) indicates that Latino students learn English more effectively when teachers reinforce and build upon their first language rather than when they attempt to eradicate it. Geneva Smitherman, William Labov, and John Baugh, as well as other linguists, have written about the importance of teachers recognizing and understanding the form of English spoken by many African American students (called Black English or Ebonics), respecting it, and incorporating it into instruction.

Critics of the cultural difference explanation point out that it fails to explain why some immigrant groups, whose cultures differ substantially from the mainstream group, are successful in mainstream schools, such as many groups of Asian immigrants in the USA, Australia, and Canada.

4.3 The Cultural Ecology or Caste Theory Explanation

Ogbu (1992), an anthropologist, is the major proponent of this theory. Ogbu distinguishes two kinds of minority groups within a society: *voluntary* and *involuntary* or *caste-like* minorities. He argues that voluntary minorities, such as Asians in the USA and in the UK, view their new nations as places of opportunities and hope. They consequently internalize the educational aspirations of the mainstream groups within their adopted societies and experience educational success.

Caste-like minorities, such as the Aborigines in Australia, Blacks in the USA, and Indians in Canada, view their native lands as oppressive because of the institutionalized discrimination they have experienced. They develop fictive kinship and collective opposition as a survival strategy. Involuntary minorities believe that if they internalize the achievement ethos of the dominant society, they will violate fictive kinship and 'act White.' Their perception of the school becomes a self-fulfilling prophecy. They perceive structural and institutional barriers to academic achievement and success in society long after they have been eliminated.

Critics of this explanation point out that it explains the academic failure of involuntary minority groups but does not explain those who are successful. It does not explain variation across and within groups or describe the complex ways in which institutionalized racism influences student achievement. Critics also note that the cultural ecology explanation has limited implications for action and reform that will improve schools.

4.4 The Critical Theory Explanation

Post-Marxists in orientation, proponents of this explanation, such as Giroux (1992) and McLaren (1997), argue that the latent function of schools in capitalist societies is to educate different social-class and racial groups for their positions within the social, economic, and political structures. The curriculum and the school reproduce the social structure. Consequently the school is doing the job that it is designed to do: to differentiate students on the basis of race and class and to provide them with an appropriate education for

their social and economic status. Theorists who write about multicultural education using a critical theory perspective often criticize other explanations and paradigms within the field. They call their work 'critical multicultural education' to distinguish it from other paradigms. Critical multiculturalists argue that other explanations in multicultural education ignore structural inequality and institutionalized racism. In nations such as the UK and Canada, critical approaches to multicultural education that focus on racism are called 'antiracism education.'

In theory, antiracism education differs significantly from other paradigms in multicultural education. However, in practice it is often difficult to distinguish antiracism education from other forms of education related to diversity. Critical multicultural educators often essentialize and distort other forms of multicultural education and provide few practical guidelines for teachers and other practitioners.

5. *The Need for a Comprehensive Paradigm to Guide Educational Reform*

Each of the explanations discussed above can contribute to the development of educational research, policy, and reform that can help low-income and minority students to increase their academic achievement, and all students to become effective citizens of pluralistic democratic nation-states. However, neither is a sufficient explanation of the problems and opportunities that result from diversity. Each is a middle-range or partial theory or explanation that can contribute to a needed comprehensive theory. A comprehensive theory must incorporate insights and research from each of the explanations discussed above. A comprehensive theory or explanation must view the school as a complex social system that has norms, values, roles, and goals like other cultural systems. It must also conceptualize the school within a social, political, and economic context that limits the role of the school in important ways. However, it must recognize that both teachers and students have agency and are not made powerless by social and economic structures. Teacher and student resistance must be an integral part of a comprehensive theory to guide educational reform.

Because schools are embedded within a social, political, and economic context, school reform requires teachers, students, and sociopolitical structures to change. Social and political structures will need to become more liberatory. Teachers and students must acquire some of each other's views, perceptions, and ethos as they interact. Both teachers and students will be enriched by this process of acculturation and cultural exchange. The academic achievement of students from diverse groups will be enhanced because their cultures and perspectives will be enriched and legitimized in the school.

See also: Equality of Opportunity; Ethnic Identity, Psychology of; Ethnicity, Sociology of; Gender, Class, Race, and Ethnicity, Social Construction of; Multiculturalism; Multiculturalism and Identity Politics: Cultural Concerns; Multiculturalism, Anthropology of; Multiculturalism: Sociological Aspects

Bibliography

Au K H 1980 Participation structures in a reading lesson with Hawaiian children. *Anthropology and Education Quarterly* **11**: 91–115

August D, Hakuta K (eds.) 1998 *Educating Language-minority Children*. National Academy Press, Washington, DC

Banks J A 2001 *Cultural Diversity and Education: Foundations, Curriculum and Teaching*, 4th edn. Allyn & Bacon, Needham Heights, MA

Banks J A, Banks C A M (eds.) 1995 *Handbook of Research on Multicultural Education*. Macmillan, New York

Bereiter C, Englemann S 1966 *Teaching Disadvantaged Children in the Preschool*. Prentice-Hall, Englewood Cliffs, NJ

Coard B 1971 *How the West Indian Child is Made Educationally Sub-normal in the British School System*. New Beacon Books, London

Giroux H A 1992 Post-colonial ruptures and democratic possibilities: Multiculturalism as anti-racist pedagogy. *Cultural Critique* **21**: 5–39

Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York

Kymlicka W 1995 *Multicultural Citizenship: A Liberal Theory of Minority Rights*. Oxford University Press, New York

Lee C D 1993 *Signifying as a Scaffold for Literary Interpretation*. National Council of Teachers of English, Urbana, IL

Lewis O 1965 *La Vida: A Puerto Rican Family in the Culture of Poverty—San Juan and New York*. Random House, New York

McCarthy C 1990 *Race and Curriculum*. Falmer Press, London

McLaren P 1997 Decentering whiteness: In search of a revolutionary multiculturalism. *Multicultural Education* **5**: 4–11

Ogbu J U 1992 Understanding cultural diversity and education. *Educational Researcher* **21**: 5–14

Philips S U 1983 *The Invisible Culture: Communication in Classroom and Community on the Warm Spring Indian Reservation*. Longman, New York

Riessman F 1962 *The Culturally Deprived Child*. 1st edn. Harper, New York

J. A. Banks

Multicultural Feminism: Cultural Concerns

'Multicultural feminism' is a term used increasingly to designate a growing literature and practice of feminism that seeks to supplement, if not supplant, the discourse of liberal feminism, which promotes gender equality

and individual human rights for all women, with one that advocates their empowerment within the specific contexts of their cultures and societies. The contextual dimensions of the lives of women that multicultural feminists particularly attend to include: histories of external as well as internal domination and subordination; cultural legacies of empowerment alongside those of disempowerment; and specific effects of the present global economy on particular societies and the differently situated women that they enclose (Volpp 1996).

Because no term satisfactorily captures all these dimensions, the designation 'multicultural feminism' is by no means settled but loosely shares the field with other terms like 'cross-cultural' or 'transnational' feminism (Grewal and Kaplan 1994, Mohanty 1988). In addition, feminists particularly troubled by liberal feminism's lack of attention to the inequities that women disadvantaged by hierarchies of race and class suffer have called for a 'revolutionary' or 'materialist' feminism to transform existing national and global structures of power, which they see as accommodating only a minority of relatively privileged women (Hooks 1984, Delphy 1980). In any event, insofar as most feminist critics of liberal feminism argue for a close analysis of the concrete, complex, and variable contexts of women's lives, the term 'multicultural feminism,' expansively understood, adequately if imperfectly covers the terrain of their concerns.

It should be noted that multicultural feminism eschews, in the main, the universal versus relativist debate that marked the discussion of human rights at the end of the twentieth century (Brems 1997). It insists, instead, on context, not as a device for evading discussions of the human rights of women, but as the necessary framework within which to conduct them. Indeed, multicultural feminism mediates two sets of tensions: the one that has lately arisen in countries of the North between the value of individual human rights and that of multiculturalism (Taylor 1994), and that which, in countries of the South, too often pits the discourse of nationalism against that of feminism (Chatterjee 1993).

1. *The Critique of Liberal Feminism*

The dominant feminist discourse in the world today is properly termed Western or liberal because, like the general human rights matrix that surrounds it, the discourse springs from, and is marked by, the European Enlightenment, which, as is well known, consecrated the value of individual liberty. To that, the American and French Revolutions added the concept of equality. In reality, the promise of these two concepts was realized in the eighteenth and nineteenth centuries in parts of the West for only a section of its population: propertied men. It took the liberal feminist movement the better part of the twentieth century to extract the fruits of that promise for some women as well: those from dominant classes and/or ethnic or racial groups. Multicultural feminism, which found its voice in the last quarter of the twentieth century, now comes along to challenge liberal feminism on two principal counts: its identification with a privileged constituency in the North, and its concomitant tendency to conflate liberty with individualism, and equality with uniformity.

Liberal feminists in the North, for their part, have begun to voice the concern that the cultural diversification which globalization has brought to their countries could work to erode the rights of women therein. Many believe that non-Western immigrant communities oppress or discriminate against women, and fear that proponents of multiculturalism will persuade the liberal state to shield such practices from the reach of laws solicitous of individual autonomy and gender equality. By contrast, feminists who reside in the South tend to support, rather than fear, the increasing juxtaposition of liberal and communitarian discourses in their countries. This, no doubt, is because they experience their societies as multidimensional and hence capable of evolution where needed; and also because juxtaposition sanctions their own selective adoption of liberal feminist tenets (Toubia 1998).

A recent essay by Susan Moller Okin (1999) in the USA sums up the anxiety that the subject of multiculturalism evokes in some liberal feminists. Okin fears that Americans, whom she describes as prone to romanticize egalitarianism, will be tempted to extend the principle of equality, which currently protects individuals only, to safeguard cultural groups as well. She advocates strongly against such extension where groups invoke culture to rationalize their control over the bodies and lives of women. Denial of protection for such groups' practices, Okin continues, must be absolute, even if it results in the extinction of their cultures.

Third World patriarchies, she asserts, forbid women to be independent from men, to be lesbian, and to refuse to bear children. They are, furthermore, sites of clitoridectomy, polygamy, forced marriages, and other practices that Okin finds sexist and abhorrent. The French government, in her view, erred gravely when it allowed African men to enter France with more than one wife, all of whom the state consequently recognized as legitimate spouses. Okin believes that the liberal state cannot thus validate polygamy for immigrants, and not for the native-born, without undermining its own existence. Finally, she accuses proponents of multiculturalism of seeing culture in monolithic terms whereas, she asserts, it constitutes a field in which women contest their oppression.

All in all, Okin's essay displays, in forthright fashion, the reductionism and presumption that non-Western feminists decry in much of Western feminism's gaze on its Other. To begin with, the approach conflates diverse Third World histories and cultures. It then reduces these to what Westerners perceive as their

most 'exotic' practices. Finally, it prescribes a repudiatory role for the West, and for Third World women, vis-à-vis the miscast practices and, if need be, their associated cultures. At its most problematic, the approach re-enacts the gestures of ignorance and arrogance, as well as the strategy of divide-and-rule, that marked the West's colonial and missionary enterprises in much of the non-Western world through the mid-twentieth century.

Not surprisingly, Third World feminists, whether in the South or in diasporic communities in the North, regularly challenge Western feminism's depiction of their cultures, to correct the record certainly, but also to distance themselves from a caricatural vision and an improbable program that too often engender animosity rather than support for the advancement of women in their communities. Chandra Talpade Mohanty (1988), for example, rejects liberal feminism's fundamental assumption that a universal patriarchy has operated across all time and space to subordinate women, and that the phenomenon remains most entrenched in the South. She and others observe that human cultures are far too complex to all fit under the rubric of patriarchy. In the case of the Indian subcontinent, Mohanty notes, the world of women instead abounds with legacies and experiences of domination as well as subordination that fluctuate with age, marital status, presence of children, proximity to kin, and other factors that condition that of gender. Commenting on the Middle East, Deniz Kandyoti (1991), Lila Abu-Lughod (1998) take issue with the prevailing Western view that an undifferentiated Islam oppresses women there. They write that the latter's situation in fact varies significantly with their countries' specific histories—indigenous, colonial, and postcolonial. Marnia Lazreg (1994) likewise emphasizes that Algeria is more than Islam. It is not religion, she points out, but changing historical circumstances, that most influence the production and reproduction of gender difference and inequality.

The general multicultural critique, then, is that liberal feminism habitually essentializes the cultures of Third World communities in its account of the oppression of their women (Harris 1990, Romany 1991). In the process, it assigns univalent meanings to the unfamiliar cultural symbols and practices that it encounters. Thus, all veiling, all instances of purdah, all forms of female genital modification are read automatically to signify the subordination of women, when they may in fact alternately or additionally represent an assertion of identity, an opportunity for female bonding, and a rite of passage to maturity respectively. In consequence, liberal feminism dismisses both the complexity of culture, and the agency of Third World women (Moraga and Anzaldúa 1983). Misperception and reductionism of course characterize all cultural groups' construction of their Other; however, the power to affect the life of the Other through such distortion is not given equally to all. To the extent that women in the North hold greater power in the international arena than women in the South to affect the future of all women, the former's errors in its construction of the lives of the latter are not merely regrettable, but sometimes downright dangerous.

Azizah al-Hibri (1996) illustrates this in the context of international human rights conferences. She relates that, to the dismay of women from the South, women from the North used their organizational know-how at the 1981 UN Mid-decade for Women Conference in Copenhagen, the 1993 World Conference on Human Rights in Vienna, and the 1994 International Conference on Population and Development in Cairo to dominate their proceedings and speak in the name of all women. In Copenhagen, women from the North announced that the gravest concerns of Third World women were veiling and clitoridectomy; in Cairo they said that these were contraception and abortion. When women from the South spoke on their own behalf in Cairo, they asserted, instead, that their highest priorities were peace and development! Nevertheless, liberal feminists' single-minded reproductive agenda in Cairo, which many in the developing countries saw as racist, had the spillover effect of delegitimizing some women from the South who participated in that conference in their own countries. More, it provoked a conservative backlash in a number of Muslim countries so that women there, al-Hibri writes, subsequently found it more difficult to obtain abortions.

The tension noted above between North and South women highlights the individual versus communitarian character of their respective agendas. For liberal feminists, who enjoy substantial peace and usually also independent access to material wealth in their societies, a focus on a woman's right to control her body and sexuality, and otherwise enjoy personal autonomy, may indeed make sense. For Third World women however, most of whom, in tandem with their communities, suffer strife and/or poverty, the collective pursuit of peace and development appear far more urgent (Davies 1983/1987). Moreover, in all cultures, sexuality is fraught with profound social as well as personal implications; and invites strong, albeit divergent, social control. As such, sexuality represents a particularly difficult subject for transnational discussion and action. Western feminists' early and disproportionate emphasis on the subject thus erected barriers rather than bridges to women in the South.

Multicultural feminists also critique the West's essentialist gaze on its at-home Other (Lâm 1994, Lorde 1984, Razack 1998). Sherene Razack identifies the cultural reductionism at work in Canada when cases of gender violence involving persons of color are discussed. While violence against women perpetrated by white men is generally depicted as manifestations of individual deviance, that by men of color is typically vested with cultural significance. Furthermore, although class- and race-based inequities contribute to the genesis of violence in communities of color, this

dimension of the problem is regularly ignored in the public discourse on the subject. The cultural story of female vulnerability, that is, trumps that of political injustice.

Audrey Lorde, in her influential work *Sister Outsider*, faults white middle-class liberal feminists in the USA for failing to educate themselves on the situation and needs of women who do not share their circumstances. She rejects the notion that the victim class must educate the victimizer class, and insists instead that the latter itself seek out knowledge and undertake self-reflection. Victims, she writes, need to conserve their energies for self-organizing. The liberal feminist call for universal sisterhood, Lorde points out, assumes a homogeneity of experience that simply does not exist. Black women, for example, bear the special responsibility of distinguishing between the oppressor's need to oppress black men from their own legitimate conflicts with them. It is a tension that confounds women who do not share a history of devastation with their men—be it of slavery, genocide, colonialism, the Holocaust, or apartheid. Lorde calls on these women to appreciate the viewpoint of those who live the tension, and to respect the solidarity that they extend to victimized men.

Third World feminists often hear, in liberal feminism's call for global sisterhood, intimations of the divide-and-conquer campaign that the West waged in colonial times (Ahmed 1992, Gandhi 1998, Lazreg 1994, Spivak 1998). Leela Gandhi finds echos of that strategy in certain moves sanctioned by liberal feminism in the North: the essentializing of Third World women as marginal and victimized; the use of liberal feminist criteria to validate civilizing missions to the Other; liberal feminists' execution of such missions. Bluntly put, where white men once purported to save brown women from brown men, Gandhi and others now see white women doing the same.

Marnia Lazreg recalls that Westerners regularly used the 'condition of women' in the colonies to illustrate the 'doomed' nature of native cultures. Yet, in Algeria, she writes, Frenchmen regularly showed contempt for those Algerian women who shed tradition; they referred to them as prostitutes, and otherwise unveiled their own hidden perspective on women. Many Algerians thus saw in the colonists' text for the advancement of Algerian women a pretext for the fragmentation of their country's revolutionary independence movement. Leila Ahmed adds that the West continues to profoundly distrust Islam, and consequently to exaggerate, and exacerbate, the gender divide in Muslim countries. For her part, Gayatri Chakravorty Spivak, in a well-known essay, finds Western feminists to be troublingly fixated on the figure of the Third World Woman. They have reified the latter, she writes, into a veritable paradigm in the West's recreation of a new Orientalism. Once again, Spivak notes, the locating of ignominy at the margin supports self-exaltation at the center. Already, under colonialism, white women acted as agents of the civilizing mission of empire. This gave them a public significance in the colonies that they did not previously enjoy in the metropolis, but which they managed to successfully transport back there.

Notwithstanding the civic personhood and individual rights that Western women gained from the modernization that colonialism helped to underwrite in their countries, Laura Nader (1989) observes that they became the first victims of the social transformation involved. For modernity, on a first order, doubles the work of women by adding wage labor to unpaid domestic work. It also nuclearizes the family, thereby disconnecting women from their extended kin and increasing their dependence on their husbands. The road away from tradition then, Nader writes, leads to alienation rather than liberty. Modernity, moreover, replaces the rough equality that prevailed between preindustrial extended households with a hierarchy of wealth that pits individuals directly against each other. Women entering the modern world thus find themselves trapped by an amplified male authority even as they also confront sudden aloneness and insecurity.

Societies in the North, where modernity has had a longer run than in the South, in time developed laws and institutions that mitigate, for women positioned to take advantage of them, the harshness that modernity otherwise inflicts. Liberal feminists tend to be so positioned, and to forget those who are not. The criticism that Okin levels at the French government's recognition of polygamous immigrant families, for example, and her readiness to assign certain cultures to oblivion, exhibit such forgetfulness, which multicultural feminists challenge. They question a feminism that would split up an existing polygamous family when this would likely leave some co-wives and their children without support. They reject a feminism that would accept the extinction of a whole culture to suppress some perceived misogynist practices therein when the women of that culture, like the men, draw the very shape and value of their personalities and lives from the totality of its signification, such that its eradication must somehow also implicate their own existence.

While multicultural feminists emphasize the constitutive nature of culture, they by no means assert its homogeneity, determinism, coherence, or even ahistorical validity. What they do assert is that, apart from gross violations of human rights, the pace and direction of change in a culture must be determined by the relevant cultural community itself. Moreover, they are confident that impulses for change exist in any cultural community precisely because a culture, unlike a cult, is not built around uniform values and limited interests but rather accommodates an ongoing adjustment of the wide-ranging and sometimes conflicting needs and desires of the many individuals and groups that make up the community. Finally, internal

impulses for change are also triggered by external stimuli (which differ from dictates), of which there are plenty in today's closely connected world.

The effect that the interplay of internal and external forces can have on cultural change, or its arrest, is illustrated by a set of events occasioned by the influx of Somalian immigrants into Seattle in the mid-1990s (Coleman 1998). Sometime in 1996, a number of Somalian women about to give birth at a state hospital in that city asked its staff to circumcise their soon-to-be-born sons as well as daughters. In the case of the latter, they requested a small cut to the prepuce or hood over the clitoris, but no removal (and hence no scarring) of tissue. The women related that the procedure would replace the more extensive traditional cutting that they would otherwise have to procure, given their religious beliefs and cultural practices, for their daughters outside the hospital and, possibly, outside the USA. While reluctant at first, the hospital in time agreed to perform the procedure under local anesthesia for girls old enough to add their consent to those of their parents.

In explaining its decision, the hospital stated that the procedure represented a far safer alternative to the more invasive traditional methods that the Somalian girls would otherwise be subject to—illegally in the USA, or legally in Somalia if they were sent back there for the purpose, as many were. Nevertheless, when news of the hospital's decision broke, reaction from feminist ranks in the USA was swift, vehement, and condemnatory. A female member of the US Congress who was prominent in liberal feminist circles publicly pronounced the proposed procedure barbaric, baffling, and horrifying in addition to illegal. The media largely concurred. Pressed from many sides, the hospital reversed its decision. Lost in the brouhaha was the considered voice of an earlier group of Somalian immigrants in the USA who supported the proposed procedure because they saw it as a bridge over which this generation of Somalians could walk on their way to a future in the USA where their own children would likely not re-enact the practice of female genital modification.

2. *The Critique of Nationalism*

Although multicultural feminists appreciate the constitutive nature of culture, they also recognize its frequent misuse, particularly in the discourse of nationalism, to rationalize the oppression of women. The ideology of nationalism, which promotes the welding together of ethnicity and power in the form of the nation-state, clearly fueled the Third World's bid to liberate itself from the West in the last century. At the same time, the ideology imposed serious constraints on a number of groups, notably indigenous or tribal peoples, minorities, and women. The constraints were prompted by the anticolonial movement's need to unify peoples and glorify histories as it sought to generate and amass the political energy required to evict colonizers and build independent modern states. Unification, nationalism taught, demanded the suppression of ethnic difference. The glorification of the past, in turn, provoked an exaggeration of gender distinctions (Chatterjee 1993). That is, the more colonialism, and later independence, drew Third World men out of their traditional milieu into the ambivalent workplace of an imposed and later borrowed modernity, the more it fell to women to assure the reproduction of what passed, in an already irreversibly transformed society, for the authenticating and reassuring domain of tradition.

Third World womanhood, as a result, became a potent symbol of the sacred 'us' versus the profane 'them' (Lazreg 1994). Where colonialism had invented the lowly status Third World woman, then, nationalism now imagined a noble Third World figure. In both versions, reification distorted culture and hid from view the concrete situations of real women. Religious or cultural fundamentalists in the Third World today who police the behavior of its women are thus engaging, as Arab feminists regularly point out, not in a return to tradition, but in the postcolonial construction of an alternate modernity which these feminists strongly contest (Abu-Lughod 1998, al-Hibri 1996, Kandyoti 1991, Lazreg 1994). In the case of Islam, the feminists emphasize, the tradition of *ijtihad* in fact invites every Muslim, male and female, to conscientiously engage in the personal interpretation and adaptation of religious texts to their lives which are understood as unfolding in social contexts that are complex and fluid. Though religious law, or *shari'a*, binds, its meaning is thus not determined by an ultimate clerical authority, as is Catholic doctrine for example, but results from consultation, or *shuria*. The emancipation of contemporary Muslim women, then, entails their reclaiming of *ijtihad* and *shuria* (Majid 1998).

Some Arab feminists are also skeptical of calls for the emancipation of women that their governments sometimes issue (Kandyoti 1991). They hear it as the state's attempt to substitute its power over women for that of kin groups, to mobilize cheap female labor, and to satisfy the international financial institutions (IFIs) of the global capitalist economy. Indeed, where Third World states early on invited women into the modern arena by enrolling them in national schools, it was generally done for the purpose of advancing the state, and not women: the latter were taught to become scientific managers of homes that would turn out compliant citizens. In the same vein, earlier colonial administrations had wanted to free women from their traditional contexts so that colonizers could train them to mold the future subjects of the colonies to the colonizers' liking.

The view that global capitalism, rather than local patriarchies, is the primary cause of the present

poverty and subordination of women, particularly in the Third World, received a late hearing in academia (Boserup 1989, Enloe 1990, Jayawardena 1986). Multicultural feminists today, nevertheless, generally insist on this perspective. They point to the present super-exploitation of women in garment, electronic, and tourist (including prostitution) industries in every region of the globe. Their campaign on behalf of women consequently focuses less on promoting individual rights to liberty and equal treatment than on challenging the political and economic structures, national and international, that break down local communities and cultures in order to make way for a single global economy to freely exploit, oppress, and degrade women. In important respects, the multicultural feminist campaign thus recapitulates the multidimensional fight waged 100 years ago in the West against the excesses of laissez-faire capitalism. Only this time, the source of excess is ubiquitous, and far more powerful: laissez-faire global capitalism (Basu 1995, Davies 1983/1987, Vinding 1998).

3. *The Multicultural Feminist Perspective*

Multicultural feminism is neither a unified theory nor a single political movement. It is a name for a perspective on the situation of women that highlights the contexts of their lives and supports their self-directed liberation from subordination, whatever its origin. Third World women in the South and in diasporic communities in the North are the principal, but not exclusive, architects of this perspective. Men from the South, together with men and women from the North, have offered important formulations of, and support for, the perspective (An-Naim 1991, Minow 1990, Taylor 1994). Multicultural feminists' emphasis on context leads them to value both the dignity of the individual and her need for the support of a meaningful community (which they see in any event as mutually constitutive), especially in light of the spreading global capitalist economy which commodifies individuals and scatters communities. Multicultural feminists thus seek out diverse and flexible ways in which women can be supported in their communities. In the area of individual human rights, for example, they support the formulation of core values, but generally find the rigid allocation of specific rights and responsibilities premature in a world of diverse cultures and unequal power. The rights of Third World women cannot be fully honored, they further hold, if the status of their communities in the global econony and polity is not elevated. In sum, multicultural feminists support the empowerment of women in the concrete, if evolving, realities of their lives.

See also: Female Excision: Cultural Concerns; Feminist Political Theory and Political Science; Feminist Theory; Feminist Theory: Liberal; Gender and Feminist Studies in History; Multiculturalism; Multiculturalism, Anthropology of; Multiculturalism: Sociological Aspects; Race and Gender Intersections

Bibliography

Abu-Lughod L 1998 Introduction: Feminist longings and postcolonial conditions. In: Abu-Lughod L (ed.) *Remaking Women: Feminism and Modernity in the Middle East*. Princeton University Press, Princeton, NJ

Ahmed L 1992 *Women and Gender in Islam*. Yale University Press, New Haven, CT

al-Hibri A 1996 Who defines women's rights? A Third World woman's response. *Human Rights Brief* **2**(1)

An-Naim A 1991 *Human Rights in Cross-cultural Perspectives: A Quest for Consensus*. University of Pennsylvania Press, Philadelphia, PA

Basu A (ed.) 1995 *The Challenge of Local Feminisms: Women's Movements in Global Perspective*. Westview Press, Boulder, CO

Boserup E 1989 *Women's Role in Economic Development*. Earthscan, London

Brems E 1997 Enemies or allies? Feminism and cultural relativism as dissident voices in human rights discourse. *Human Rights Quarterly* **19**(1)

Chatterjee P 1993 *The Nation and its Fragments*. Princeton University Press, Princeton, NJ

Coleman D L 1998 The Seattle compromise: Multicultural sensitivity and Americanization. *Duke Law Journal* **47**

Davies A M (ed.) 1983, 1987 *Third World—Second Sex: Women's Struggles and National Liberation*, Vols. 1 and 2. Zed Books, London

Delphy C 1980 Towards a material feminism. In: Marks E, de Courtwron I (eds.) *New French Feminisms*. University of Massachusetts Press, Amherst, MA

Enloe C 1990 *Bananas, Bases, and Beaches: Making Feminist Sense of International Politics*. University of California Press, Berkeley, CA

Gandhi L 1998 *Postcolonial Theory*. Columbia University Press, New York

Grewel I, Kaplan C 1994 Introduction: Transnational feminist practices and questions of postmodernity. In: Grewel I, Kaplan C (eds.) *Scattered Hegemonies*. University of Minnesota Press, Minneapolis, MN

Harris A 1990 Race essentialism in feminist legal theory. *Stanford Law Review* **42**

Hooks B 1984 *Feminist Theory: From Margin to Center*. South End Press, Boston

Jayawardena K 1986 *Feminism and Nationalism in the Third World*. Zed Books, London

Kandyoti D (ed.) 1991 *Women, Islam, and the State*. Temple University Press, Philadelphia

Lâm M C 1994 Feeling foreign in feminism. *Signs*: *Journal of Women in Culture and Society* **19**(4)

Lazreg M 1994 *The Eloquence of Silence: Algerian Women in Question*. Routledge, New York

Lorde L 1984 *Sister Outsider*. The Crossing Press, Trumansburg, NY

Majid A 1998 The politics of feminism in Islam. *Signs*: *Journal of Women in Culture and Society* **23**(2)

Minow M 1990 *Making All the Difference*. Cornell University Press, Ithaca, NY
Mohanty C T 1988 Under Western eyes: Feminist scholarship and colonial discourses. *Femininist Review* **30**
Moraga C, Anzaldúa G (eds.) 1983 *This Bridge Called My Back: Writings by Radical Women of Color*. Kitchen Table: Women of Color Press, New York
Nader L 1989 Orientalism, occidentalism, and the control of women. *Cultural Dynamics* **2**(3)
Okin S M 1999 *Is Multiculturalism Bad for Women?* Princeton University Press, Princeton, NJ
Razack S 1998 *Looking White People in the Eye: Gender, Race, and Culture in Courtrooms and Classrooms*. University of Toronto Press, Toronto, ON
Romany C 1991 Ain't I a feminist? *Yale Journal of Law & Feminism* **4**(23)
Spivak G C 1998 Can the subaltern speak? In: Nelson C, Grossberg L (eds.) *Marxism and the Interpretation of Culture*. University of Ilinois Press, Urbana, IL
Taylor C 1994 The politics of recognition. In: Gutmann A (ed.) *Multiculturalism*. Princeton University Press, Princeton, NJ
Toubia N (ed.) 1998 *Women of the Arab World: The Coming Challenge*. Zed Books, London
Vinding D (ed.) 1998 *Indigenous Women: The Right to a Voice*. International Work Group for Indigenous Affairs, Copenhagen, Denmark
Volpp L 1996 Talking 'culture': Gender, race, nation, and the politics of multiculturalism. *Columbia Law Review* **96**

M. C. Lâm

Multiculturalism

Multiculturalism has come to refer to (a) the state of a society or the world in which there exists numerous distinct ethnic and cultural groups seen to be politically relevant; and (b) a program or policy advocating or promoting such a society. The first is a factual claim, the second evaluative. But like so many claims in the social sciences, the empirical claim involves much more than empirical judgements, and the evaluative claim is contested by as many who promote it (and certainly does not follow from the empirical claim). Controversy has raged over what the best response to the fact of cultural diversity in the world might be, and it is not too much to suggest that the social and political issues raised by multiculturalism are among the most pressing and important for the new millennium. Coming to grips with these issues requires asking some difficult questions about the nature of culture and its relation to human welfare, and especially what political consequences follow in light of the answers given. Political philosophers aspire to deal with the challenges posed by multiculturalism *fairly*; the problem is, no one is quite sure what this actually entails.

1. The Relevance of Culture

Human societies have always been culturally diverse, and certainly most organized political societies. Yet for much of the history of Western political philosophy, the ideals of political life which have been promulgated have presumed that citizens shared not only a common political status, but a common descent, language, and culture. Aristotle, for example, one of the central figures in the history of Western political thought, presumed this despite himself being a 'resident alien' of Athens. John Stuart Mill, champion of modern liberty and individual autonomy, thought that free institutions were next to 'impossible' among a people 'without fellow feelings,' or who read and spoke different languages. He argued that the boundaries of a state should coincide 'in the main with those of nationalities' (Mill 1975). The presumption that states should be culturally homogeneous was inspired in part, no doubt, by straightforward prejudice against foreigners and outsiders, but also linked to a series of issues that continue to bedevil contemporary political philosophers. Among these are: does justice require a common culture in order to motivate the goodwill and sacrifices it typically demands of citizens? Can a political society be stable if too much emphasis is placed on the diversity of its peoples and cultures? Can democratic institutions work, and be seen to be legitimate, if the citizens of a state speak different languages, have different world views, and even perceive the history of their nation differently? Finally, if individuals are fundamentally equal, can we balance the demands of equality with those of particular cultures?

Political philosophers from Aristotle to Mill (and beyond) have tended to answer these questions in the negative; too much cultural diversity threatens to undermine the possibilities for a peaceful, prosperous, and ultimately just political community. The problem is, the challenges of cultural diversity are impossible to avoid. The history of the twentieth century, and especially of recent decades, makes this abundantly clear; witness the re-emergence of nationalism, the spread of violent ethnic conflict, and the extraordinary movement of peoples fleeing cultural and economic persecution.

Before tackling these questions directly, however, it is important to get a sense of what is actually meant by 'culture' and how it has come to (re)occupy a preeminent position in discussions concerning distributive justice and the emergence of 'the politics of difference.'

1.1 Diversity of What?

Theorists concerned with multiculturalism agree that culture matters, but disagree profoundly over how and why it does. Liberal political philosophers, in particular, have struggled to reconcile their commitment

to the welfare of individuals with what seems to be the collective nature of cultural identification. But first, when political philosophers talk about cultural diversity, what are they are actually talking *about*?

The concept of 'culture' has a complex history and, as a result, a set of meanings sometimes at odds with each other (Moody-Adams 1997). At a basic level, culture refers to the particular set of habits, beliefs, and customary practices of a people or society; to its way of life. In the nineteenth century, culture was often tied closely with quasibiological notions of race and ethnicity, and thus to judgements about the inherent superiority of one form of culture over another—a tendency that continued to manifest itself in murderous ways throughout the twentieth century. Linking culture to biological or quasiscientific characteristics has thus come to been seen as extremely dubious, involving both bad science and questionable ethics. Instead, anthropologists (amongst others) have argued that culture should be understood as a set of ultimately malleable and negotiable cognitive and social frameworks governing human behavior which are complex, dynamic rather than static, and patched together from a variety of sources rather than woven of the same cloth (Geertz 1973, Clifford 1988, Moody-Adams 1997).

But how are cultures individuated politically? Some insist that the groups which matter politically are those grounded in *ethnicity* (stripped of any biological or scientific overtones): a people united by certain prepolitical allegiances of ancestry, collective memory, homeland, or culture (Smith 1986). For others, cultural diversity refers to a range of different groups—whether based on ethnicity, race, socioeconomic class, language, lifestyle, or sexual orientation—who are disadvantaged relative to a dominant political culture manifested in the main institutions and 'public reason' of the state in which they reside (Young 1990). Those shared characteristics or practices, often only perceived as being shared in light of 'mainstream' oppression or rejection, are said (for better or worse) to partly constitute the culture of a group.

Another important account of cultural diversity is provided by Will Kymlicka (1995a, 1998, cf. 1995b). He argues that multiculturalism, in fact, involves two different kinds of cultural diversity. The first is a diversity of 'societal cultures.' The second is that created by individual and familial migration. A societal culture provides its members 'with meaningful ways of life across the full range of human activities, including social, educational, religious, recreational, and economic life, encompassing both public and private spheres' (1995a, p. 76). It consists of not only collective values and memories, but usually a shared language and territory (1995a, pp. 76–80). Note that these are *cultural* groups rather than ones based on race or descent. Membership is thus defined in terms of integration into a cultural community rather than in terms of descent. Aboriginal 'Nations' and the Quebecois, for example, are societal cultures but also ones in which there is a high degree of racial and ethnic intermixing and thus great internal diversity. Although a state may contain a dominant societal culture, it might also include other minority ones. Kymlicka calls these groups 'national minorities': at the time of their incorporation into a larger state they possessed (a) a distinct 'social vocabulary' embodied in a full set of social practices and institutions; (b) were territorially distinct; and (c) self-governing. The capacity and motivation to form such a culture, for Kymlicka, means it is characteristic of 'nations' or 'peoples.' Migrants, on the other hand (and refugees; although they present a difficult case), have essentially chosen to give up their societal culture when deciding to migrate (1995a, pp. 100–1). So multiculturalism involves (at least) both 'polyethnic' diversity (arising from individual and familial migration) and 'multinational' diversity (arising from the incorporation of previously self-governing cultures into a larger state). Religious groups who seek out marginalization from mainstream society present something of a special case, which will be discussed below. Furthermore, there are groups which cut across ethnic and national lines, such as women, gays and lesbians, and the disabled who, at different times, make claims based on their disadvantage or difference relative to a dominant societal culture. These claims tend to have more in common with 'polyethnic' rather than 'multinational' claims, however, since they usually concern the terms of integration into society rather than a demand for the preservation or protection of a distinct society (cf. Kymlicka 1998, pp. 90–103).

The reason why so much hangs on how a culture is to be individuated is because of what is said to follow in terms of the *rights* members of that culture—or indeed the group itself—can claim. For Kymlicka, 'national groups' are entitled to strong group rights, i.e., those that entail varying degrees of self-government and protection from the decisions of the wider political community. Polyethnic rights are important but much weaker; they are concerned mainly with ensuring that the terms of integration into the dominant societal culture are as fair as possible.

Any set of distinctions generates anomalies and hard cases, and Kymlicka's is no different. Some groups, like certain religious groups, seem to fulfil the criteria for being a 'societal culture' and yet are usually not granted rights associated with national minorities. In some cases however, some 'ethno-religious' sects—such as the Amish or Hutterites in North America—have been granted exemptions from general laws in order to enable them to maintain certain key features of their religious life. But these are historically special cases; they usually involve a theology which seeks to minimize contact with the modern world. Other groups, such as refugees and certain migrant groups (e.g., some Muslim communi-

ties in Great Britain), do not fit comfortably into either category; they aren't national minorities but they also seem more distinct and thus possibly vulnerable to collective harm than the discussion of 'polyethnic' rights might suggest. But how is it that rights follow from claims about culture in the first place?

1.2 Cultural Rights

What is striking about recent debates among liberals about multiculturalism is how liberalism's traditional concern with individual well-being has been linked to the value of cultural membership—an ostensibly 'collectivist' value. Not all cultures are individualistic of course, but liberals are generally concerned with how culture matters for the lives of individuals. But how exactly does culture matter in this way? And even if it does, is it something in relation to which rights should be, or even could be, assigned?

Consider three basic arguments:

(a) *The argument from cultural diversity*: Respect for cultural pluralism entails preserving existing cultures. Pluralism is valuable because it provides individuals with alternative ways to live, promotes critical reflection on the culture within which one currently lives and thus promotes change and growth within cultures generally.

(b) *The historical injustice argument*: Certain group rights are justified insofar as they are aimed at redressing or compensating for a history of past injustice against particular historical communities at the hands of dominant majorities.

(c) *The value of cultural membership argument*: Certain group rights are justified insofar as membership in a culture has value for individuals, and there is a link between the welfare of the group and welfare of its members. Group rights contribute to the preservation of that culture and thus to the welfare of its members.

All three arguments have played a prominent role in recent debates. However, (a) seems incomplete, since the promotion of cultural diversity *per se* does not necessarily justify the protection of any *particular* culture. Furthermore, if more cultural diversity is better than less, then either all cultural practices are multiculturalism-enhancing and deserving of toleration or support, or only some are. The former cannot be true (since surely some cultural practices are not so deserving), and the latter forces us to identify those properties which make some cultural practices valuable and others not. This is no easy task. This leads to consideration of (b) and (c). Perhaps the most powerful case has been one made in relation to (c) and so that is a good place to begin. The historical injustice argument will be returned to in Sect. 3.

One way in which cultures matter to individuals is that they constitute an important aspect of their identity, or as Kymlicka puts it, provide a meaningful 'context for choice': 'freedom involves making choices amongst various options, and our societal culture not only provides these options, but also makes them meaningful to us' (1995a, 1995b). So in addition to the familiar liberal emphasis on individuals having the freedom of expression, movement, and association, they also need access to a societal culture. A societal culture is a basic 'primary good' for all individuals, and thus relevant to a liberal theory of justice (on primary goods see Rawls 1971).

A variation of the above argument (though made independently of it) is the relation said to exist between cultural identity and individual self-respect or self-esteem. People's self-respect can be bound up with the way in which the group of which they are a member is perceived by others and by society at large. Lack of esteem for the group can translate into lack of esteem on the part of individual members and this constitutes harm, the remedy of which may justify granting some form of group rights or rights of self-determination (Taylor et al. 1992, Margalit and Raz 1990, on the 'social bases of self-respect' see Rawls 1971, pp. 303, 440).

But do individuals require access to a *particular* culture, one they can call their own? There are at least three objections to this strategy. Some argue that liberals should not be in the business of preserving particular cultures, since the kind of society they should be promoting is one in which individuals are free to move and adapt *between* cultures as much as within them. So as long as individuals have access to *a* societal culture—and especially one made up of a diversity of cross-cutting cultural influences—their 'context for choice' is secure. If there is a cultural structure that liberals should promote, it is a cosmopolitan one, a common but diverse 'mélange' of cultural meanings conditioned by the basic liberal freedoms of conscience, expression, association, and exit (Waldron 1992, Turner 1993).

A variation on the cosmopolitan argument is to place greater emphasis on the value of diversity or difference itself. There is no identity without difference, and thus to talk of 'societal cultures,' or even a singular 'cosmopolitan culture,' is to risk missing the inherent multiplicity and hybridity of political and cultural identities. Individuals and groups are constituted by a diversity of moral sources, and the central task of justice should be to remain 'critically responsive' to this fluidity and diversity rather than reinforce 'found' or extant cultural structures (Connolly 1995, Bhabha 1994).

Finally, granting special rights to cultural groups on the grounds of their value for individuals involves a delicate balancing act between the rights of individuals and the measures needed to ensure the survival of the culture. The dangers here are ones liberals should be especially sensitive to. For some, even to talk about 'societal cultures' or groups as meriting political recognition is to engage in dubious metaphysics. The

basic units of social reality cannot be group-based since groups are simply the product of interactions between individuals; there are only associations constituted by individuals and not vice versa (Kukathas 1997). Thus there are no grounds upon which to provide political or legal recognition for groups on the basis of their playing a 'constitutive' role in individuals' personal identity.

2. *The Politics of Multiculturalism*

Disagreement over the social ontology of multiculturalism only intensifies when the more concrete matter of deriving political and institutional arrangements from them arises. The 'politics' of multiculturalism refers to the way in which the consequences of social and cultural diversity are thought to necessitate social and political change.

If the argument is that cultures matter to individuals because they are connected to the capacity for living a good life—whether in terms of providing the 'social bases of self-respect' or as a meaningful 'context for choice'—then does it follow that the *state* has a direct interest in recognizing or promoting a societal culture (or set of cultures)?

For cosmopolitans it does not follow at all. People are sure to value their cultural attachments, but the state should steer clear of tampering with these commitments. It should treat them the way it treats religious beliefs, as a private matter about which the state is officially neutral. If it promotes anything, it should be the ability of people to live according to whatever way of life they wish drawn from the many possibilities on offer, consistent with the basic rights of others to do likewise. So the state should remain 'indifferent' towards and generally tolerant of the associations individuals wish to form or join. The costs of a 'politics of difference'—of recognizing and encouraging individuals to organize politically around their cultural identities—outweigh any potential benefits. Politics can be a source of oppression as much any other social practice, especially given the potential for cultural elites or others to gain control of the resources and institutions set up to protect cultural minorities and use them to take advantage of more vulnerable members and/or enrich themselves (Kukathas 1997, p. 149, Hardin 1995). If individuals are being harmed as a result of their cultural attachments, then the remedy is to ensure a more consistent application of basic liberal rights and protections, rather than to modify them in light of cultural demands.

The latter view has been referred to as pursuing a policy of 'benign neglect.' Cultures neither deserve or need state assistance to survive, and if they do, probably are not worth preserving anyway. This view has also been associated with a 'civic' as opposed to 'ethnic' account of the state. A civic state is one in which citizens are members in virtue of a shared commitment to constitutional principles or values that are independent of particular cultural attachments (Habermas 1993). An ethnic state, on the other hand, presupposes a set of values grounded in a shared cultural framework based on common descent or ethnicity. Hence also the appeal to civic and ethnic nationalisms, the former involving 'constitutional' patriotism grounded in cosmopolitan principles independent of ethnocultural and historical features, and the latter, a particularist patriotism grounded in 'blood and soil.'

Practically, however, the pursuit of 'benign neglect' is extremely difficult to accomplish. States play a crucial role in coordinating social and economic activity and it is inevitable that in the process, certain cultural presumptions and practices will be favored over others. Almost every state, for example, has to establish internal boundaries, a common educational system, an 'official' language (or set of languages), common holidays, and enforce a basic legal framework within which citizens will have to act. In other words, the idea that government could be neutral when it comes to the treatment of all cultural groups is not only difficult to imagine but 'incoherent' (Kymlicka 1995a, p. 108). States have an inescapable cultural component built into them. Moreover, those states which are often invoked as exemplars of civic rather than ethnic nationalism or as having pursued a policy of 'benign neglect' have in fact possessed a distinct societal culture made up of more than just allegiance to cosmopolitan principles and values. Some groups have suffered harm as a result of their difference from this dominant societal culture. Some of these states have granted relatively strong group rights to various cultural groups in light of their disadvantage or via historical agreements in the course of nation building. (All of this seems true of the United States and Canada, for example, two apparently quintessential 'civic' nations.)

Focus on the *harm* people have suffered in relation to their particular cultural attachments has featured prominently in recent debates, mainly because of the appalling treatment just about every state has at some point handed out to minority groups. This touches on the historical injustice argument mentioned in Sect. 1.2. Protecting societal cultures may be justified as a means of redressing or compensating members of that culture for a history of past discrimination or oppression at the hands of a dominant majority or minority (Buchanan 1991). However, there are questions surrounding this argument too. First, would citizens have to agree on a specific account of their nation's history in order to fix responsibility for these past injustices? This is difficult to imagine in many places where intercultural conflict has been particularly brutal. Mutual loathing and distrust undermines cool and impartial reflection on the past.

Second, why should the viability of group rights depend on the group having been oppressed or discriminated against? Tying the granting of group rights to the existence of (past or present) discrimination by or exclusion from a dominant societal culture risks misunderstanding the motivation of groups seeking such rights.

Some societal cultures, it might be argued, are owed special rights independently of any historical pattern of disadvantage they may have suffered (as is argued is the case with Aboriginal peoples; see Kymlicka 1995a, p. 142, Tully 1995). Group rights, in other words, are not necessarily about compensation for exclusion from or discrimination by a dominant culture, but rather about the intrinsic commitment members of particular groups have towards their societal culture(s).

But what kind of rights follow specifically in light of the arguments canvassed above? It is here that the generality of normative discussions about cultural rights runs up against the complexity of the world in which they are claimed. For it is difficult to imagine a generalizable set of rights that can be applied to all cultural groups equally, given the complex interplay of history, context, and principle in each case. The tensions inherent in the desire to protect both societal cultures and individual well-being emerge with particular force here. The very measures needed to protect a societal culture might conflict with the actual freedoms and capacities of the individuals in whose name the protection is being offered in the first place (Hardin 1995, Okin 1998). This is particularly true for those more vulnerable members of a societal culture. Having suffered past injustice is no guarantee that relations *within* that group will be just. As a result, it is probably best to think of cultural rights as involving a bundle of different kinds of claims rather than as a set of universalizable principles, albeit always conditional on them not harming the basic interests of individuals. Considered as such they might include, among other things: the right to be exempted from certain laws or government policies; assistance for the maintenance of linguistic, educational, or cultural activities; various forms of recognition (legal, moral, and political); 'affirmative action' programs, representation rights, self-government rights (within a state or federation); and ultimately secession (Levy 1997; on secession see Buchanan 1991). Some of these claims involve negative rights of noninterference, some positive rights of assistance, and others a combination of the two. Some might be compatible with individual rights (wieldable by individuals as members of that group) and some held exclusively by the group (wieldable by an agent or representative of that group). Much will depend on the context in which they are appealed to. Many will inevitably be controversial.

Aboriginal peoples, for example, have a strong claim for self-government rights because they hold land (or have strong claims to land), possess a distinct societal culture, were previously self-governing, and have suffered a long history of oppression and discrimination at the hands of most settler states. Despite the fact that some Aboriginal practices entail significant restrictions on individual liberty (such as rules governing property holding, for example), most liberal theorists think the value and importance to Aboriginal people of maintaining their distinct cultural structure outweigh some of these concerns, or at least that fears of potential abuse are overstated (Kymlicka 1995a, Tully 1995, but cf. Okin 1998). Needless to say, when concerning matters such as family law or criminal punishment, opinions begin to differ (Okin 1998, Pogge 1997, Turpel 1991). Self-government rights are the source of considerable debate and social division in countries where they have been claimed or granted. They present a significant challenge to traditional conceptions of equal citizenship and the principle of equality before the law.

Claims by religious groups, on the other hand, are not generally thought of in the same way. They might possess a distinctive societal culture, but save in exceptional circumstances they are not usually thought of as thereby possessing rights to a particular territory or self-government. Their claims will more likely have to do with exemptions from certain state laws, or for state support of educational and cultural activities. But in states where religious groups do have extensive exemptions and privileges against the general population (as in Israel, for example), even these 'weaker' claims can cause considerable social tensions (especially when combined with claims for self-government on the part of another historical community or societal culture living on the very same territory).

3. *The Future of Multiculturalism*

It is often suggested that the seemingly irreversible forces of globalization will result in the eventual eradication of the local or particular in favor of a global, though not necessarily cosmopolitan, Western consumerist culture. Or that with the end of the Cold War and the 'triumph' of liberal democracy, the great ideological battlelines of the past will be tamed and redrawn within a generally liberal framework (Fukayama 1992). Recent developments in world politics, however, and especially the explosion of ethnic conflict in Africa, Europe, and South-East Asia suggest otherwise. The local and particular has come back with a vengeance. Furthermore, within the established liberal democracies, many are questioning the programs established to promote multicultural identities, especially where they involve providing groups with extra resources or support to maintain their way of life, or as compensation for past discrimination and injustice. This is particularly evident in disputes surrounding the claims of Aboriginal peoples in Canada and Australasia, and reconsideration of 'affirmative action' programs and differen-

tiated citizenship rights in the USA and Europe. The future of multiculturalism, in other words, might appear rather bleak.

Some have laid the blame for disenchantment with multiculturalist policies at the feet of those who were loudest in insisting on the 'recognition of difference' in the first place. Overemphasis on difference has encouraged (perhaps unwittingly) the reification of ethnic and cultural modes of identification, instead of taming them and making them more open-ended and dynamic (Turner 1993, Hollinger 1995). The constant demand for greater inclusiveness and plurality in every sphere of a nation's cultural life—from its school curricula and national history, to its political parties and assemblies—has sown the seeds of resentment and distrust rather than cross-cultural understanding (Schlessinger 1991; cf. the essays in Taylor et al. 1992, Connolly 1995, Tully 1995, Kymlicka 1998). The constant focus on cultural, ethnic, and historical identities risks undermining, so it is argued, the necessary conditions for a stable and ultimately just political society, i.e., commitment to a common set of basic individual rights that are justifiable independently of any particular ethnocultural framework (Ignatieff 1998).

Is the future multicultural? In many ways it must be, for the only way to rid societies and states of their diversity is to forcibly remove it, and that entails injustices of the worst kind (with which everyone today is all too familiar). The case against multiculturalism as undermining political stability and the conditions for justice is weak, for it is clearly *not* the case that culturally homogeneous societies are the most just. The common culture of justice is multicultural since the people to whom justice must be done and the societies within which it must be done are multicultural. Perhaps what is needed is to combine the attempt at identifying those basic harms no one should suffer from as result of their cultural attachments, with the desire to respect and appreciate the different ways people come to value and hold their beliefs about the world in 'thick' social and cultural frameworks (Miller and Walzer 1995, Moody-Adams 1997). Paradoxically, the desire to respect difference and acknowledge alternative sources of moral commitment may indeed lead to the emergence of a kind of cosmopolitan culture 'multiculturalists' are often said to abhor. But if so, this will not entail having moved 'beyond multiculturalism,' whatever that means. Insofar as diverse cultures are incorporated into the public life of a state in a diversity of ways, and people continue to value the cultures within which and between which they live, this marks the beginnings of a new kind of public culture, one in which the grooves etched by the demands of justice begin to track calls for the engagement with and respect of difference.

See also: Democracy: Normative Theory; Discrimination; Discrimination: Racial; Diversity and Disagreement; Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity, Sociology of; Ethnonationalism: Cultural Concerns

Bibliography

Bhabha H K 1994 *The Location of Culture*. Routledge, London

Buchanan A 1991 *Secession: The Legitimacy of Political Divorce*. Westview Press, Boulder, CO

Clifford J 1988 *The Predicament of Culture*. Harvard University Press, Cambridge, MA

Connolly W E 1995 *The Ethos of Pluralization*. Minnesota University Press, Minneapolis, MN

Fukayama F 1992 *The End of History and the Last Man*. Free Press, New York

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Habermas J 1993 Struggles for recognition in constitutional states. *European Journal of Philosophy* **1**: 128–55

Hardin R 1995 *One for All; The Logic of Group Conflict*. Princeton University Press, Princeton, NJ

Hollinger D 1995 *Postethnic America: Beyond Multiculturalism*. Basic Books, New York

Ignatieff M 1998 *The Warrior's Honour: Ethnic War and the Modern Conscience*. Chatto & Windus, London

Kukathas C 1997 Liberalism, multiculturalism and oppression. In: Vincent A (ed.) *Political Theory: Tradition and Diversity*. Cambridge University Press, Cambridge, UK

Kymlicka W 1995a *Multicultural Citizenship*. Oxford University Press, Oxford, UK

Kymlicka W (ed.) 1995b *The Rights of Minority Cultures*. Oxford University Press, Oxford, UK

Kymlicka W 1998 *Finding Our Way: Rethinking Ethnocultural Relations in Canada*. Oxford University Press, Toronto, ON

Levy J 1997 Classifying cultural rights. In: Shapiro I, Kymlicka W (eds.) *Ethnicity and Group Rights*. New York University Press, New York

Margalit A, Raz J 1990 National self-determination. *Journal of Philosophy* **87**(9): 439–61

Mill J S 1975 On representative government. In: Wolkeim R, Mill J. S (eds.) *Three Essays*. Oxford University Press, Oxford, UK

Miller D, Walzer M 1995 *Pluralism, Justice and Equality*. Oxford University Press, Oxford, UK

Moody-Adams M M 1997 *Fieldwork in Familiar Places: Morality, Culture & Philosophy*. Harvard University Press, Cambridge, MA

Okin S M 1998 Feminism and multiculturalism: some tensions. *Ethics* **108**: 661–84

Pogge T 1997 Group rights and ethnicity. In: Shapiro I, Kymlicka W (eds.) *Ethnicity and Group Rights*. New York University Press, New York

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Schlessinger A 1991 *The Disuniting of America: Reflections on a Multicultural Society*. Whittle Communications, New York

Smith A D 1986 *The Ethnic Origins of Nations. Basil*. Blackwell, Oxford, UK

Taylor C, Gutmann A, Rockefeller S C, Walzer M, Wolf S 1992 *The Politics of Multiculturalism*. Princeton University Press, Princeton, NJ

Tully J 1995 *Strange Multiplicity: Constitutionalism in an Age of Diversity*. Cambridge University Press, Cambridge, UK

Turner T 1993 Anthropology and multiculturalism: what is anthropology that multiculturalists should be mindful of? *Cultural Anthropology* **8**: 411–29
Turpel M 1991 Home/land. *Canadian Journal of Family Law* **10**: 17–40
Waldron J 1992 Minority cultures and the cosmopolitan alternative. *University of Michigan Journal of Law Reform* **25**: 751–93
Young I M 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NY

D. Ivison

Multiculturalism and Identity Politics: Cultural Concerns

Multiculturalism is a social perspective that is committed to publicly recognizing and respecting many cultures and cultural identities. Identity politics is one means by which members of a society strive for public recognition of their cultures and cultural identities. One conception of multiculturalism that animates identity politics is that people primarily identify as members of groups defined by ethnicity, race, gender, religion, or sexual orientation, and therefore can be expected to pressure their government for greater positive political recognition of, and public benefits for, their own group. Another conception of multiculturalism is critical of identity politics because it rejects the idea of a one-to-one correspondence between a person and a particular cultural identity and views individual identities as themselves multicultural.

1. *History*

Although societies have long been multicultural, the use of multiculturalism as a term in social and political discourse is relatively recent.

1.1 *Canada*

The term was first prominently used in a 1965 preliminary report of the Canadian Royal Commission on Bilingualism and Biculturalism to suggest that Canada recognizes many ethnic cultures other than the English and French, even though its official languages are English and French. The report introduced the recurring image of a 'mosaic' society in which citizens could maintain distinctive ethnic cultures within a larger societal framework (which in Canada is bilingual), as contrasted to that of a 'melting pot' society in which immigrant cultures are expected to blend into a mainstream culture over time.

Three general kinds of multiculturalism can be distinguished, which coexist in Canada (Kymlicka 1998). Ethnic multiculturalism is directed toward immigrant groups and reflected in the 1971 introduction of a formal multicultural policy in Canada, directing all parts of the federal government to respect the diversity of all ethnic groups in their policies and practices. Bilingual multiculturalism is recognized, among other places, in the Official Languages Act. Nationalist multiculturalism recognizes a special set of rights for indigenous peoples. The Quebecois also claim a kind of nationalist multiculturalism. All three multiculturalisms are recognized in the 1982 Canadian Charter of Rights, which guarantees minority language education within the country's bilingual framework and special treaty rights of indigenous peoples.

In 1988, the Canadian Parliament unanimously passed the most thoroughgoing national statement of multiculturalism, the Canadian Multiculturalism Act. The Act's opening indicates that the aim of all three kinds of multiculturalism, taken together, is to recognize both the civic equality and the cultural diversity of all Canadians:

> The Government of Canada recognizes the diversity of Canadians as regards race, national or ethnic origin, colour and religion as a fundamental characteristic of Canadian society and is committed to a policy of multiculturalism designed to preserve and enhance the multicultural heritage of Canadians while working to achieve the equality of all Canadians in the economic, social, cultural, and political life of Canada.

The Canadian experience also indicates that all three kinds of multiculturalism can become public policy in a single country, although not without controversy, consistent with maintaining social unity.

1.2 *The USA and the Globalization of Multiculturalism*

In the late 1970s, the term multiculturalism began to be used in the USA and subsequently throughout the world to refer to a commitment to making societal cultures more inclusiveness of various kinds of diversity. Popular culture increasingly recognized the contributions of members of groups that historically had been socially marginalized. In 1976, for example, Alex Haley's *Roots: The Saga of an American Family*, a best-selling novel and popular television program that vividly portrayed the history of an African-American family, ignited great public interest in the genealogies of identity groups.

In the 1980s and 1990s, with the rapid expansion of world commerce and communications, multiculturalism became recognized as a global phenomenon in two senses: first, the societal cultures of many more countries became self-consciously multicultural, and second, the influences on diverse societal cultures became more apparently global. The globalization of multiculturalism in the second sense raised the question of whether the culture of commercialism and

consumerism would dominate more indigenous and diverse cultural influences within societies. Some critics of globalization argued that it had led to greater cultural homogenization, which threatened the mosaic of cultural diversity within societies that many consider the hallmark of multiculturalism.

The recognition of diverse cultures within societies nonetheless continued throughout the 1980s and 1990s. Within the USA, a primary focus of concern about multiculturalism in this period was the content of public school curricula. A widely publicized New York State task force appointed by the Commissioner of Education issued a report, *A Curriculum of Inclusion* (1989), which proposed that 'all curricular materials be prepared on the basis of multicultural contributions to all aspects of our society.' A 'Committee of Scholars in Defense of History' publicly attacked the report for 'contemptuously dismiss[ing] the Western tradition' and reducing 'history to ethnic cheerleading.' A subsequent report by the New York State Social Studies Review and Development Committee, *One Nation, Many Peoples: A Declaration of Cultural Interdependence* (1991), came under similarly sharp attack for supporting what the historian Arthur Schlesinger, Jr., called the cult of ethnicity' (Schlesinger 1992). Schlesinger voiced a general criticism of multiculturalism for portraying the USA as a 'nation of groups' in which ethnicity is the defining life experience rather as a 'nation of individuals.' Yet advocates and critics of multiculturalism alike agreed with one of its central aims: to publicly recognize the social contributions of members of minority groups and women in the history curricula of schools. Many history textbooks were consequently rewritten and social studies curricula throughout the country were revised to become more socially inclusive (Fullinwider 1996).

Another controversial set of policies, loosely called affirmative action, has become widely associated with multiculturalism and identity politics in the USA. Although early affirmative action policies were supported by both Democratic and Republican presidents, Lyndon Johnson and Richard Nixon, these policies soon became a source of partisan controversy. Affirmative action was defended as a partial means of correcting the competitive disequilibrium created by past discrimination against women, African-Americans, Latino-Americans, Native Americans, Asian-Americans, and other disadvantaged groups. Special efforts were made by government agencies and contractors, large corporations, and selective colleges and universities to recruit members of these groups into the competitive social positions from which they had previously been excluded, either by *de jure* or *de facto* means, or often by both. By the 1980s, affirmative action policies had helped to transform the face of many public and private institutions. By the 1990s, affirmative action policies became a lightning rod for controversy over identity politics in the USA.

Two kinds of affirmative action policy, which parallel two kinds of multiculturalism, were distinguished in a seminal US Supreme Court case, *Regents of the University of California v. Allan Bakke* (the *Bakke* case), decided in 1977. One kind aims for proportional representation: it sets separate quotas for women and disadvantaged minorities (the 'protected' groups) so as to ensure that a proportional number are chosen for competitive positions, regardless of the relative quality of members of the protected and unprotected groups in the pool. A second kind aims for nondiscrimination: it sets 'targets' that are to be achieved only if a sufficient number of applicants in the protected groups qualify by the relevant criteria, which count race, ethnicity, and gender as one but only one among many qualifications. The pivotal opinion in the *Bakke* case, that of Justice Powell, found the use of quotas unconstitutional (in the absence of a 'compelling' state interest), but the use of race as one among many qualifications constitutional. As of January 2000, *Bakke* remains the reigning constitutional law of the land, but the 5th District Court of Appeals in the case of *Hopwood v. Texas* (1996) decided that the use of race even as one factor among many in university admissions was unconstitutional.

2. *Competing Conceptions*

Different perspectives on various dimensions of multiculturalism—personal identity, public recognition, and standards of assessment—help elucidate the controversies that multiculturalism generates and the policies that it supports.

2.1 *Two Perspectives on Identity Formation of Individuals*

One perspective on multiculturalism is based on the idea that personal identities are not just influenced but *constituted* by cohesive cultural communities. On this view, communally constituted identities call for special political recognition in the form of exemptions from otherwise legitimate laws, or more broadly in the form of rights of self-governance for the group as a whole. Indigenous groups in multicultural societies have a right to be governed by their own group rather than by the larger political unit whose laws would otherwise govern anyone who lived within its territorial boundaries. The argument generated by this perspective is that in the absence of group rights to self-governance, the larger political unit will violate the cultural identities of the members of indigenous groups. The core premises are that inherited cultural identities are essential attributes of individuals, these identities must be politically respected, and political respect will be forthcoming only by a government that itself repre-

sents the particular cultural identity of the group. In the absence of political recognition of indigenous group rule, members of the indigenous group will experience a loss of identity, an injury to who they are, and therefore an injustice.

A second multicultural perspective supports some of the same policy conclusions as the first but for different reasons. The second perspective defends individuality. It rejects the idea that inherited cultural identities are 'essential' attributes of individuals, and therefore argues against the view that personal identity either is or should be 'tightly scripted,' as the philosopher K. Anthony Appiah puts it, by a person's cultural inheritance (Appiah 1996). To endorse a tight one-to-one correspondence between personal identity and cultural inheritance is to subscribe to what the literary critic Henry Louis Gates, Jr. calls 'cultural geneticism' (Gates 1993). The historically inherited identities of African Americans and women, both perspectives agree, are not changeable merely at whim or will. But the perspective that defends individuality argues that inherited group identities can be altered, and are often altered, by human agency over time. This perspective on multiculturalism, by contrast to the first, therefore views change in personal identity over time not necessarily as a 'loss' or an injustice, but rather as a possible—and often desirable—manifestation of individual freedom.

The second perspective, because it defends individuality at its core, makes a more moderate case for group rights than the first. The political philosopher Will Kymlicka argues that among the basic human (or primary) goods is having a cultural community as a 'context of choice' (Kymlicka 1989). It does not follow that unwanted changes in cultural communities are unjust. The identities of Quebecois and Inuit Indians in Canada have changed significantly over time, partly in response to the political environment that is beyond any individual's control, or even beyond a single group's control. But Quebecois and Inuit Indians as individuals have not therefore lost their personal identities. Unwanted change in a person's identity in itself is insufficient to constitute an injustice on this perspective. A more limited claim is therefore made on behalf of multiculturalism, personal identity, and group rights: groups that were involuntarily incorporated into a larger political unit—often through colonial conquest—have a right to govern themselves at least partly independently of that larger political unit.

The claim that indigenous peoples have a right to collective self-governance is an extension of the basic democratic right of self-governance, applied to indigenous people. It is therefore a claim that can be exercised consistently with a rejection of the idea that personal identity is completely constituted by group inheritance or membership. The second perspective therefore supports a correspondingly more qualified conception of a group right. A group that has historically been taken over by force may legitimately exercise rights of collective self-governance against the larger political unit, but it must not exercise its group rights in a way that limits the individual rights of its members. On this perspective, the claims of 'tradition' or 'cultural purity' must not take priority over those of individual rights (Kymlicka 1998).

The second multicultural perspective addresses itself to the concern that group rights are a way of subjugating the less powerful members of a group, often women and children, to the more powerful, and therefore should be opposed (Okin 1999). This concern still leaves open the question as to whether some individual members of the indigenous groups in question, if given the choice as individuals, would choose to be governed by the larger or smaller political unit. In the case of children, the question also remains as to what governing unit best serves their interests. These questions do not lend themselves to easy political answers because political governance is necessarily a matter of collective agency even if human identities and interests are individualized.

Historical evidence makes clear that multiculturalism is misconceived if it is premised on the idea that people have singular or unchangeable group identities. Individual identities themselves are multicultural and they are alterable over time. Not only societies but individuals are shaped by more than a single culture. And they are not only externally shaped, but also internally altered, by human agency. In short, people are multicultural agents of change. The contrast between cultural identities that are completely constituted by an inheritance and those that are completely voluntary is therefore a false dichotomy. Most cultural identities are neither *given* like hair color nor *chosen* like high cuisine. People select, interpret, and evaluate stories, histories, and customs in an attempt to make the best out of the various cultures given to them. They also interpret and evaluate the institutions, laws, practices, and procedures of the society that they inherit. The range of moral responses to a multicultural inheritance includes obedience, criticism, reform, civil disobedience, self-imposed exile, and revolution. All are part of the human potential for cultural identities that are morally reflective and multicultural. In even minimally free societies, such identities turn out to be very diverse.

2.2 Two Modes of Public Recognition

Should multicultural societies publicly recognize the diverse cultural identities of individuals, and if so, how? Many public schools in the USA once assigned American history texts that referred to Native Americans as savages, and were almost entirely devoid of voices of African-Americans and women. Their curriculum also commonly included Protestant prayers

that all children were expected to reverently recite. These curricular practices reflect two different problems to which two different modes of public recognition are a response. Although the modes of public recognition differ, they are united by the single principled aim of treating individuals with diverse cultural identities as civic equals.

The history books illustrate the first problem: exclusion and therefore public disrespect of the experiences of historically disadvantaged and therefore marginalized groups. The mode of response to this problem that is consistent with civic equality is public recognition of the experiences of oppressed groups. The forced recitation of Protestant prayers in public schools illustrates the second problem: public intolerance of minority beliefs and practices. The mode of response to this problem that is consistent with civic quality is not public recognition but toleration, agreeing to disagree about beliefs and practices that are considered part of the basic liberty of individuals. Toleration substitutes for imposing a single comprehensive system of beliefs and practices on all students, which would disregard their religious or other spiritual convictions. Public recognition and toleration are different modes of response to multiculturalism, depending on what is at stake; they need not compete with one another.

A multiculturalism based on respecting individuals as civic equals supports public recognition of the social contributions of diverse groups when the absence of public recognition—as in the case of history textbooks that exclude the contributions and experiences of oppressed minorities or women—is publicly disrespectful and discriminatory. When cultural beliefs and practices are matters of personal conviction rather than matters of public concern, these beliefs and practices call for toleration rather than public recognition. Whether and how individuals worship God, for example, calls for toleration, not public recognition of the value of divine worship or its absence among individuals. These are the two modes of response that are consistent with a perspective that tries to secure both the basic liberty and the civic equality of individuals with culturally diverse identities.

Rather than viewing these two modes of responses as both appropriate reactions to different problems that arise within a multicultural society, some debates about multiculturalism instead pose a dichotomous choice between toleration and public recognition. Either a society tolerates cultural differences among groups by privatizing them, or it respects cultural differences by publicly recognizing them. The first response is often identified with the values of liberal democracy and the second response is often identified as critical of liberal democratic values. But as the educational examples suggest, the choice of whether to tolerate or to publicly recognize cultural differences does not necessarily reflect a choice between different underlying values. Depending on the circumstances, each response can be justified by the same aim of treating all individuals as civic equals.

Neither response—public recognition or toleration—is justifiable if extended to some cultural practices, such as slavery, torture, human sacrifice, child prostitution, and female circumcision (clitoridectomy), that violate individual rights or otherwise oppress individuals. The fact that such an oppressive cultural practice is valued by some or even many members of a society is not in itself a sufficient reason to defend either its toleration or public recognition when the basic human rights of individuals are at stake (Taylor 1994). A separate question still arises as to how oppressive cultural practices that violate human rights are best changed, and by whom, in light of the social support they continue to receive, often by the most powerful members of a society.

2.3 *Two Standards of Assessment*

Controversies over multiculturalism raise a recurrent philosophical question about standards for assessing whether cultural practices are good or bad, consistent with basic human rights or oppressive. Does multiculturalism presume that the standards for assessing specific cultural practices are themselves culturally specific? Or are there universal standards of assessment? Cultural relativism suggests that moral responses to cultural practices necessarily depend on one's specific cultural identity, and proclamations of cross-cultural standards are themselves culturally specific. Universalism suggests that there are moral responses to cultural practices that span a wide range of cultural identities and that defend a set of human needs and capacities, such as freedom from want and for personal and political self-determination. On this view, these responses are common to all human beings who live under favorable social conditions (Gutmann 1993).

Cultural relativism suggests that standards of morality and justice are relative to particular cultural understandings. It follows that if cultural understandings are radically different, then some cultural practices—which seem oppressive to members of one culture—can be justified to members of another culture. On this view, a multicultural society should allow different cultural groups to be governed by their own cultural understandings. Culturally diverse people can relate to each other, at best, by tolerating their different cultures. They can live peacefully and indifferently side by side, however, only if their own cultural understandings are compatible with establishing a *modus vivendi* with other cultures. In the absence of an appreciation of the virtue or usefulness of cross-cultural toleration, multicultural politics on this relativist view becomes a war of all cultures against all.

Whereas relativists criticize universalists for the moral hubris that is a source of colonialism and imperialism, universalists criticize relativists for condoning the oppression of vulnerable individuals—women, children, and disadvantaged minorities—in the name of respecting cultural practices that by the relativists' own moral standards are unworthy of respect. Less widely noted is the fact that cultural relativism calls for the same critical response as does universalism by all people whose own moral standards—whether relative or universal—consider torture, slavery, religious intolerance, and gender and racial discrimination to be oppressive practices.

The convergence between relativist and universalist standards of assessment is made possible by three common features of diverse cultures. First, cultures contain moral standards by which their cultural practices themselves can be criticized (Walzer 1987). Second, cultures are not internally univocal in their moral standards (Gutmann 1993). Third, human beings are not completely constituted by their cultures (Kateb 1992). They can and do think beyond their cultural inheritances. Multiculturalism therefore need not await a resolution of the age-old debate between relativism and universalism to support a set of moral standards for assessing cultural practices, and to permit proponents of multiculturalism to oppose cultural practices that oppress individuals or otherwise violate human rights.

See also: Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Civil Liberties and Human Rights; Cultural Rights and Culture Defense: Cultural Concerns; Race Relations in the United States, Politics of

Bibliography

Appiah K A 1996 Race, culture, identity: misunderstood connections. In: Appiah K A, Gutmann A (eds.) *Color Conscious*. Princeton University Press, Princeton, NJ

Fullinwider R K 1996 *Public Education in a Multicultural Society: Policy, Theory, Critique*. Cambridge University Press, New York

Gates Jr. H L 1993 *Loose Canons*. Oxford University Press, New York

Gutmann A 1993 The challenge of multiculturalism in political ethics. *Philosophy & Public Affairs* **22**(3): 171–206

Gutmann A 1999 *Democratic Education*. Princeton University Press, Princeton, NJ

Joppke C, Lukes S (eds.) 1999 *Multicultural Questions*. Oxford University Press, Oxford, UK

Kateb G 1992 *The Inner Ocean: Individualism and Democratic Culture*. Cornell University Press, Ithaca, NY

Kymlicka W 1989 *Liberalism, Community, and Culture*. Oxford University Press, Oxford, UK

Kymlicka W 1995 *Multicultural Citizenship: A Liberal Theory of Minority Rights*. Oxford University Press, Oxford, UK

Kymlicka W 1998 Ethnic associations and democratic citizenship. In: Gutmann A (ed.) *Freedom of Association*. Princeton University Press, Princeton, NJ, pp. 177–213

New York State Commissioner's Task Force on Minorities 1989 *A Curriculum of Inclusion*, New York State Education Department, Albany, NY

New York State Social Studies Review and Development Committee 1991 *One Nation, Many Peoples: A Declaration of Cultural Independence*. New York State Education Department, Albany, NY

Okin S M 1999 *Is Multiculturalism Bad for Women*? Princeton University Press, Princeton, NJ

Schlesinger Jr A M 1992 *The Disuniting of America*. Norton, New York

Taylor C 1994 The politics of recognition. In: Gutmann A (ed.) *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton, NJ

Walzer M 1987 *Interpretation and Social Criticism*. Harvard University Press, Cambridge, MA

Young I M 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

A. Gutmann

Multiculturalism, Anthropology of

In the years since the widespread social unrest that characterized the 1960s in the United States and Western Europe, multiculturalism has become a common idiom for discussing social inequality and locating oneself politically in the US, other 'white settler' societies, and to a lesser extent Western Europe. In the US, which has been the most prominent and influential locus of multiculturalism, the term succeeded a similar use of civil rights in the decades following the Second World War. To begin with, then, multiculturalism serves as a one-word label or code-word—used by advocates and critics alike—to index a certain, if somewhat loose, genre of progressive politics emanating from, but not confined to, a particular set of liberal democratic and robustly capitalist states, most notably the US.

The politics signaled by the term multiculturalism has at least two notable features. First, it frames issues of social justice and domination in terms of identity-groupings, that is, groupings of persons that correspond to socially prevalent notions of human kinds. Foremost among these identity-groupings, in affirmations of multiculturalism, have been groupings defined by 'race' and 'ethnicity,' with groupings based on 'gender' and 'sexuality' the next most prominent. Second, multiculturalism has promoted an understanding of 'the political' that foregrounds issues of representation, particularly in the senses of depiction and proportionate access to public institutions and space.

In addition, like so much other talk about persons and social relations at the end of the twentieth century, multiculturalism as an idiom reflects a complex movement of ideas between the social and behavioral sciences, on the one hand, and the world those disciplines inhabit and study, on the other. More specifically, multiculturalism represents a major instance of the public circulation of the key term of American anthropology, culture—though contained within a compound term that was not itself prominent in any social or behavioral science discipline prior to the compound's wider circulation. Multiculturalism is, in short, a hybrid concept, born out of the too little examined interaction between social scientific knowledge and its social reception and circulation.

1. *Multiculturalism in Time and Place*

Multiculturalism takes the form of, and thus characterizes itself as, an affirmation of plurality, announcing both the recognition and championing of diversity. Deployed as a predicate about the US and other Western societies, the designation multicultural rhetorically contests the more established identification of these societies as 'white' and, in the case of societies in Europe, as exclusively 'European.' At a descriptive level, then, multiculturalism acts to disrupt the set of presuppositions that allow whiteness to go without saying when speaking about an American (or an Australian or a Canadian) and that make black European sound oxymoronic and white European sound redundant. At the same time, in normative terms, multiculturalism speaks against the notion that dominant kinds, most notably white folk, are the primary agents of history, producers of 'great works,' and the holders, as it were, of the intellectual property rights over most every human institution of national and universal worth. In this sense, the singularity contested by multiculturalism is the singularity of monopolistic claims to agency and worthiness that differentially distribute status and cultural capital along the lines of the most legible distinctions of 'identity' in Western societies and, concomitantly, the larger global order.

1.1 *Multiculturalism and Prior Critiques of Eurocentrism*

In attacking the attribution of unmarked (whether national or universal) status to dominant identities, multiculturalism converges with, and to some extent has borrowed from, an earlier set of critical analyses of ideologies and practices of domination in colonial societies (e.g., Fanon 1963). Yet when compared to these earlier critiques of monopolistic claims to worthiness on behalf of whiteness (that is, the critiques of Eurocentrism), multiculturalism exhibits at least two distinctive features or tendencies.

First, given that whiteness and persons of European ancestry are fixtures of Western societies (or more precisely, are imagined to be), in multiculturalist discourse has not taken the form of an affirmation of an alternative cultural singularity for these societies absent whiteness (as is the case in various instances of anticolonial nationalisms). Rather, it has taken the form of an affirmation of cultural plurality (multiculturalism), in which the ideal is less the excision or repatriation of the 'Euro-' element than its repositioning as just one among many cultures.

Second, by comparison to earlier critiques of European colonialism, multiculturalism has exhibited a greater tendency to supplement concerns about Eurocentrism with attention to socially visible axes of stratification other than race, though in limited and uneven ways, as we shall see. This relative potential for analogic extension, that is, for recognizing oppressions and cultures defined in other terms, is both signaled and rhetorically enabled by the very open-endedness and lack of specificity of multi- as the term placed into opposition to singularity/Eurocentrism.

1.2 *From Civil Rights to the Politics of Identities and Representation*

When considered in relation to the historically antecedent Civil Rights movement in the US, what is arguably most striking about multiculturalism is the relative valorization and amplifying of 'identity' consciousness. In the key legal statutes and cases of the Civil Rights movement, attending to identity distinctions, such as observable signs of racial identity, was condemned as the very root of unjust discrimination—as when an employer judges a job applicant on the basis of 'color' rather than 'merit' or 'ability.' Indeed, one of the lasting legacies of the Civil Rights era in the US is the legal principle that employers are permitted to take social identity into account only when it can be established that social identity is a '*bona fides* occupational qualification,' as when a synagogue is hiring a religious instructor. Moreover, that 'race' could ever be so regarded was diminished almost to the vanishing point in US law, for it was held that resistance to nonwhites serving in certain positions, for example, as policemen in predominantly white communities, did not make nonwhites less qualified to serve in those positions, even if it made it more difficult for them to do. In sum, in the defining legal battles of the Civil Rights era, social injustice was defined as discrimination on the basis of group identity, as distinct from a judgment based on an individual's abilities and accomplishments.

While advocates of multiculturalism do not repudiate this understanding of 'discrimination,' they

add to it in ways that are transformative. They argue that prevalent and institutionalized standards of worthiness are themselves fundamentally biased, due to systematic privileging of dominant identity-groupings in the social recognition and definition of merit. Here too, as in the US Civil Rights movement, we see that social injustice is defined in a context of meritocratic ideals and institutions claiming allegiance to such ideals. In the case of multiculturalism, however, unjust discrimination is found to be embedded deeply in institutionalized standards of merit. Persons of color, women, and other subordinated identity groupings are disadvantaged because standards of judgment are, on a systematic basis, disguised measures of 'whiteness' and 'maleness,' rather than the 'neutral' and 'objective' measures they are represented as being. Concomitantly, these same persons suffer harm because they find 'achievement' defined in institutional and public settings as something alien to their 'identity' and 'culture.' The comments of Dale Fleming, the director of the Native American Center of the Stockton Unified School District in California, capture well this multiculturalist vision of both injustice and the social reproduction of identity-group stratification: 'The culture is the base … [I]f the kids feel bad about themselves or can't deal with their own history or who they are, then school is real difficult' (quoted in Gallagher 2000, p. 37). By contrast, even the most intensely class-disadvantaged white male, multiculturalism tells us, does not experience the acquisition of school knowledge—be it in the form of 'correct' speech, knowledge of Shakespeare, or mastery of calculus—as contradicting, violating, or masking 'who they are'; instead, they experience this more happily as individual 'achievement.'

From such a diagnosis of the manifold injuries of identity, it follows that social injustice cannot be achieved without a systematic revaluation of subordinated identities. On this view, in contrast to the judicial lessons of the Civil Rights era in the US, identities and their observable signs should not be disregarded categorically in the quest for social justice. On the contrary, overcoming existing injustice requires a compensatory identity consciousness.

1.3 Forms of Multiculturalism Emanating from the US

In this historical context, as it has emanated from the US, multiculturalism is more a struggle over the differential distribution of cultural capital to identity groupings than over legal rights, financial capital, or material resources. Concomitantly, we find an intense politicization of schooling, as well as of museums, performance venues, and sites of commemoration. That a politics located in these sites has been embraced as a means of pursuing progressive agendas no doubt registers the extent to which such agendas have been blocked in electoral politics, the courts, and labor negotiations in the wake of the 1960s, particularly in the US.

Within US higher education, the effort to redress injuries of identity has been associated primarily with three overlapping projects: the increased hiring of faculty and administrators who are not white males; the increased representation on syllabi of works authored by persons of color and women, particularly in humanities courses; and increased curricular attention to histories and experiences of nondominant identity groups. In lower grades of schooling in the US there has been less organizing to change the identity demographics of classroom teachers, primarily because the lower status of K-12 teaching has for some time meant that this occupational niche has been relatively open to both women and persons of color. In other words, here 'diversity' preceded multiculturalism, if for reasons that confirm multiculturalism's general understanding of social stratification. In addition, in the lowest levels of schooling, the precursor to the humanities is primarily the teaching of language skills, both oral and written. In this context, what counts as 'diversity' is not works authored by persons of subordinated identities, but language correlated with and indexical of those identities. As such, multiculturalism in these grade levels has been concerned with both multilingual education and classroom uses of nonstandard forms of the national language.

Multiculturalism in museums and other sites of commemoration and performance, both in the US and elsewhere, are recognizable variants of these school forms. In art museums, for instance, multiculturalism means collecting and exhibiting more works by artists who are persons of color and/or female; producing and trumpeting exhibitions of the 'heritage' of subordinated identity-groupings; and actively pursuing 'new' (here a code-word for 'nonwhite') audiences.

1.4 Forms of Multiculturalism Emanating from outside the US

As multiculturalism has increasingly circulated both into and out of Western Europe, there have been notable supplements to and shifts away from multiculturalism as it has emanated from the US. What most distinguishes the two cases is that in the US, 'diversity' is understood to be something longstanding, going back many centuries to the beginnings of European and then African migration to the Americas. By contrast in the European case, multiculturalism primarily registers a response to 'diversity' understood as something historically recent, resulting from significant numbers of 'new' (here again, a raced term) immigrants from Africa, Asia, and the Caribbean. This historical consciousness reflects the success of European nationalist ideologies in producing the reality effect of primordial homogeneity. The conse-

quent foregrounding of 'new immigrants' in Western European multiculturalism has meant, in turn, a distinctive concern with legal issues regarding citizenship and immigrant workers' rights. In addition, debates about multiculturalism in both Germany and France have been haunted by, and deployed as a means of shaping, collective memories of the holocaust. The major exception to these general trends within Western Europe is Spain, where the relative weakness of Spanish nationalism has meant that regional identities are the primary focus of debates about 'pluralism' in society and in relation to the state (Greenhouse 1998).

Other national traditions of multiculturalism similarly reflect their own patterns of erasing and remembering difference, relative to both whiteness and the nation. In Australia, Canada, and New Zealand, for instance, the relatively greater social visibility of persons of 'Native' ancestry—by comparison to the US—has brought to the fore legal issues involving land and natural resources, often in the context of disputes over past treaties.

1.5 Marketing Multiculturalism

The use of multiculturalism by arts institutions to pursue 'new' audiences suggests ironic affinities with commercial efforts to expand consumption by reaching and forging 'new' markets parsed along the lines of already established social identities. Indeed, what is perhaps the only other widely recognized instance of 'multiculturalism' is also the least political and, arguably, the most international: marketing. Of course, the linking of brands to legible social identities has been a staple of marketing from the time it first became mass, a century or so ago. By contrast, what has been most distinctive in recent marketing labeled 'multicultural' has been advertisements in which the desirability of a product is located in figures of multiracial comingling, though of uniformly affluent consumers. This fetishism of multiracial, intraclass sociality has been aptly dubbed 'the Benneton effect,' after the Italian garment corporation that used it to such commercial advantage (Chicago Cultural Studies Group 1992, p. 552).

1.6 The Mutual Abstraction of Class Relations and Cultures

While such marketing of affluent cosmopolitanism is quite far from the serious political struggles of multicultural educators and other activists, the very legibility of these tableaus of class-homogenous consumers as 'multicultural' raises the vexed issue of the relationship between multiculturalism and class, or put slightly differently, between identity politics and class politics.

In multicultural talk, 'race, gender, and class' trip off the tip of the tongue as a triplet. Yet class has clearly played a different and much less significant role in multiculturalism than have race and gender, or even sexuality (Ortner 1998). If one surveys US art museums over the decades of multiculturalism, for example, one finds few if any exhibitions of the art of working-class persons, in contrast to the by-now-numerous exhibitions of art by African–Americans, Chicanos, and women—exhibitions framed (and marketed) in precisely these identity terms. Similarly, though Studs Terkel's *Working* (1974) has been taught in a great number of US college and university courses, its inclusion on syllabi is neither presented nor conventionally read as enacting a multicultural agenda. Class positions are not, in short, among the cultures of multiculturalism. On the contrary, the cultures of multiculturalism are figured as occupying such positions, understood in the minimal sense of ranked slots along a scale of wealth, or somewhat more complexly, these cultures are figured as having a particular statistical distribution along such a scale. Moreover, multiculturalism offers a vision of a just future in which extant cultures will no longer be stratified—as if the same cultures would exist if the political-economic relations between them were transformed. Such a view obscures, and thereby risks giving safe haven to, the mutual constitution of class and identity distinctions. We would most emphatically find a different constellation of class relations were classes not mediated by imagined distinctions of kind (i.e., by 'identities'), just as vice-versa, we would find a significantly different system of identities were identities not used to organize and ground class relations.

1.7 Essentializing and Multiculturalism

The broader point here is that the cultures of multiculturalism characteristically are figured as stable elements, and not as contingent formations. Indeed, what 'race,' 'ethnicity,' and 'gender' have in common, and what distinguishes them from 'class,' is that the former, unlike the latter, are recognized conventionally as inherent or unalterable characteristics, which is to say as components of 'identity' or 'kind,' rather than of situation. So too, with the next most common foundation of cultures in multiculturalism: sexuality. This, then, is the principle circumscribing the set of cultures in multiculturalism: cultures correspond to divisions of humanity (of a suprafamilial scale) that are presumed to be fixed, even natural.

This objectifying and essentializing delineation of cultures can be observed in other aspects of multiculturalism as well. For example, looking at multiculturalist discussions of history, we find contemporary identity groupings projected indefinitely, and thus anachronistically, into the past. This mode of commemoration obscures the very historicity

of extant identity groupings in order to establish 'their' agency and define 'their' heritage. So too, multiculturalism often replicates dominant assumptions that allot expressive forms to persons on the basis of observable signs of identity, so that 'jazz' but not 'Beethoven' is spoken of as belonging to 'the culture' of a musician who is 'black'—without regard to her personal taste, musical training, or primary socialization to music.

Overall, it would be fair to say that multiculturalism contests social domination of extent identity groupings, but does not interrogate the presuppositions and contingencies of their existence. Indeed, some multiculturalists have criticized such interrogation, particularly from anthropologists, for undermining the political project of valorizing currently subordinate identity-groupings (e.g., Trask 1991).

1.8 'Neoconservative' Attacks

Much of multiculturalism's notoriety both in and beyond the US is due to attention given to it by prominent neoconservative pundits and politicians. Like multiculturalism, the term neoconservative (or neocon) registers a political stance adopted in the wake of 'the 1960s' across Western societies but emanating in particular from the US. The self-proclaimed newness in this political label indexes distance from the discredited racism and anti-Semitism of the 'old' conservatism, as well as opposition to 'sixties radicalism,' itself a codeword for a decline in patriotism, a rise in secularism, and experimentation with gender roles, sexuality, and/or illicit drugs. Certainly, the multicultural debate was intensified and internationalized in 1989 when Nobel Laureate Saul Bellow was quoted, famously and widely, as saying that he would study Zulu culture when 'the Zulus have a Tolstoy'—a quip in which 'Zulus' was at once legible as a metonymic substitute for peoples of color and derisively suggestive of primitivism.

Neocons have offered two main, and related, criticisms of multiculturalism. First, they have argued that demands for curricular diversification promote social divisiveness ('Balkanization') and threaten national unity; and second, that the celebration of diversity promotes ethical 'relativism' (typically depicted by means of Eskimo euthanasia and/or sub-Saharan clitoridectomy).

In terms of the first criticism, what is perhaps most striking is the extent to which neocons have been particularly exercised about curricular shifts in higher education. One thinks, for example, of William Bennett, the Secretary of Education in the administration of US President George H. Bush, insisting that Stanford University had jeopardized America's 'common culture' by its decision in 1989 to drop from its graduation requirements a year-long survey course on the history of Western Civilization. Even allowing for the hyperbole of politicians, what is fantastic about Bennett's view is that the pyramidal structure of schooling means that higher education in general, and education at Stanford in particular, has never been a means of producing a 'common culture' at the national or societal level (Guillory 1993). Through either intent or confusion, then, neocons have steadfastly misrepresented multiculturalism's challenge to an ongoing class project—the teaching of 'great works' to elites—as a threat to the nation.

The second neocon criticism recycles a workhorse of modern conservative thought—the bogey of secular modernity's loss of moral certainty. In this regard, what is striking is that multiculturalists have exhibited little interest in relativizing ethical norms, as opposed to norms of canon formation or even norms of speech and writing. Indeed, multiculturalism has in many ways been strikingly conventional in its ethics. Multiculturalists have not, for instance, championed the rights of new immigrants in Western democracies to practice polygamy, even as they have embraced the extension of monogamous marriage to same-sex couples; nor have multiculturalists attacked the obvious ethnocentrism of laws that criminalize the consuming of pet animals that are elsewhere edible. Similarly, if in France and Germany the wearing of scarves by Islamic school girls has been defended in the name of 'multiculturalism,' multiculturalists in these European states have with rare exceptions steered away from seeking legal protection for clitoridectomies, traditional or not.

Finally, when the specter of relativism is raised, it is important to look back at modernity's horrors and ask, 'which should humanity fear more: modernity's relativism or its absolutisms?' It was, after all, racial anthropology, not its Boasian cousin, that did service in Hitler's Germany. In this regard, it is worth recalling an observation made by Benedict at an early moment in the Nazi rise to power: 'The recognition of cultural relativity carries with it its own values, which need not be those of absolutist philosophies' (1934, p. 278; see also Geertz 2000, Chap. 3).

2. Multiculturalism and Anthropological Knowledge

As multiculturalism gained attention in both the academy and the mass media, many anthropologists responded with an edgy defensiveness. Sometimes in writing and more often in hallway talk, it was said that anthropology knew best about 'culture', and that anthropology had been multicultural *avant la lettre*. Yet though too often concerned to protect disciplinary authority, anthropologists have nonetheless provided insightful, and largely sympathetic, criticisms of multiculturalism, focussing in particular on both its in-

attention to global diversity and its essentializing tendencies, strategic or not (e.g., Domínguez 1995, Urciuoli 1999, Yanagisako 1995).

At the same time, anthropologists largely have overlooked the most fundamental issue of theory raised by multiculturalism: whether human sociality has its origins in sameness, as seems so obviously true in a world populated by identities, or in difference, as argued by Lévi-Strauss in his discussions of reciprocity, thought, and myth. Similarly, at a different level of abstraction, cultural-social anthropologists have neglected to bring to the multiculturalist debate relevant ethnographic examples of social orders that normalize forms of social diversity that our world pathologizes. Thus, for example, crucially important struggles over bilingual education and over the proper educational roles of standard and nonstandard forms of speech have proceeded undisturbed by the salient evidence of the human capacity to live in multilingual social orders, as documented in, say, numerous ethnographic studies of Amazonia (e.g., Jackson 1983).

Finally, though it has gone largely unacknowledged within the discipline, it is worth observing that the use of culture in multiculturalism has itself contributed to the discipline's own heightened care and scrutiny when using the culture concept in recent decades. For many anthropologists, observing essentializing uses of culture in multicultural discourse has yielded a pained but fruitful recognition of essentializing practices within anthropology. Thus, the public circulation of culture has itself impacted on the discipline, just as the earlier remaking of culture by the Boasians provided a conceptual resource that contributed to and helped authorize the public perception of Western societies as 'multicultural'—if in a fashion outside of disciplinary control. In the end, this may be the greatest lesson of multiculturalism for anthropology: The disciplinary knowledge we strive to produce cannot be retained and stabilized. If that knowledge is at all 'good to think,' it will be put to some social use, and as a result, when we look again, the world we observe will not be the same.

See also: Cultural Assimilation; Cultural Critique: Anthropological; Cultural Expression and Action; Cultural Rights and Culture Defense: Cultural Concerns; Culture: Contemporary Views; Culture, Sociology of; Multicultural Feminism: Cultural Concerns; Multiculturalism; Multiculturalism and Identity Politics: Cultural Concerns; Multiculturalism: Sociological Aspects; Primary Elections; Social Stratification Thought

Bibliography

Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston

Chicago Cultural Studies Group 1992 Critical multiculturalism. *Critical Inquiry* **18**: 530–55

Domínguez V (ed.) 1995 (Multi) Culturalisms and the baggage of 'race.' *Identities* **1**(4): 297–300

Fanon F 1963 *The Wretched of the Earth*. Grove, New York

Gallagher B 2000 Teaching (native) America. *The Nation* **270**: 26–38

Geertz C 2000 *Available Light: Anthropological Reflections on Philosophical Topics*. Princeton University Press, Princeton, NJ

Greenhouse C (ed.) 1998 *Democracy and Ethnography: Constructing Identities in Multicultural Liberal States*. SUNY Press, Albany, NY

Guillory J 1993 *Cultural Capital*. University of Chicago Press, Chicago

Hollinger D 1992 Postethnic America. *Contention* **2**: 79–96

Jackson J 1983 *The Fish People: Linguistic Exogamy and Tukanoan Identity in Northwest Amazonia*. Cambridge University Press, Cambridge, UK

Kymlicka W 1995 *Multicultural Citizenship: A Liberal Theory of Minority Rights*. Clarendon Press, Oxford, UK

Ortner S B 1998 Identities: The hidden life of class. *Journal of Anthropological Research* **54**(1): 1–17

Rosaldo R 1997 Cultural citizenship, inequality and multiculturalism. In: Flores W, Benmayor R (eds.) *Latino Cultural Citizenship: Claiming Identity, Space and Rights*, Beacon Press, Boston, pp. 27–38

Smith R T 1996 On the disutility of the notion of 'ethnic group' for understanding status struggles in the modern world. In: *The Matrifocal Family: Power, Pluralism, and Politics*. Routledge, New York, Chap. 11

Terkel S 1974 *Working*. Pantheon, New York

Trask H 1991 Natives and anthropologists: The colonial struggle. *The Contemporary Pacific* **3**: 159–77

Urciuoli B 1999 Producing multiculturalism in higher education: Who's producing what for whom? *Qualitative Studies in Education* **12**(3): 287–98

Yanagisako S 1995 Transforming Orientalism: Gender, nationality, and class in Asian American studies. In: Yanagisako S, Delaney C (eds.) *Naturalizing Power: Essays in Feminist Cultural Analysis*. Routledge, New York, pp. 275–98

D. A. Segal

Multiculturalism: Sociological Aspects

Theories, interpretations of reality, and concepts of polity travel from continent to continent. One expression of 'globalization' is an increased readiness of observation and imitation of successful institutional inventions coming up in some part of the world or in some social system and being adopted in another context, be it in the sphere of politics, economics, science, education, or the arts. John W. Meyer and colleagues (Thomas et al. 1987, Meyer et al. 1992) analyse a 'global diffusion' of problem-solving strategies in various fields, a phenomenon that is more probable in the modern functionally differentiated

society (Luhmann 1982) which is not fenced by national borders. The world seems to move towards the predicted 'world society' (Stichweh 2001) in which new ideas can be communicated rapidly supported by new technologies—sometimes by unexpected recipients and, now and then, with surprising results. One of these diffused concepts is 'multiculturalism,' which has spread throughout the Western world and lately has moved to Eastern and Central Europe.

1. The Term 'Multiculturalism'

Descriptions of society depend on their reference to 'grand theories.' Paradigmatic changes of such descriptions emerge out of a combination of social mobilization, the spirit of the times, scientific innovation, and/or of new political interests. That was true for the concept of 'class society' replacing the traditional notion of 'social stratification,' it is true for the opposed distinction of 'functionally differentiated society,' and it can be attributed to fashionable terms like 'civil society,' 'postmodern society,' 'risk society,' 'information and knowledge society,' 'event society,' and many other labels created frequently through the cooperation of the social sciences and the mass media. Each of them introduces a new semantics of self-description of modern society meant as an answer to perceived change or as a result of changed theoretical perception.

A change of reference also occurred with the term 'multicultural society' coined in the 1970s. Already in the 1980s the concept could look back to a formidable career in the public discourse. With very rare exceptions, there was no idea of ethnopluralism before that time in the anthropological literature. Consequently all categorial ingredients of this concept like 'cultural pluralism,' 'cultural identity,' and 'cultural conflict,' but also the main issue of 'ethnicity,' which has replaced 'race,' are scientific inventions no older than 100 years (Dittrich and Radtke 1990, Sollors 1989). They are used to describe or conjure new constellations of cultural diversity in nationally framed regional parts of the world society. In these cases various local, collective, or individual actors (social scientists, political philosophers and politicians, professional educators and majority/minority speakers), start to describe the relations between social groups along an emphasized distinction of ethnically coded 'cultures.'

Again, it can be left open if the reality has changed or if its construction has promoted the predicted changes causing interest groups to perceive each other in ethnic terms. Disregarding the controversies about 'realism' versus 'nominalism' and the relation of 'observer' and 'reality' (Searle 1995) one can state: 'about' with the phenomenon of the multicultural society is a new semantic of 'multiculturalism' as a mode of public communication about the reality of modern society using the distinction of 'culture.' Many such redefinition attempts conceptualizing society as a composition of ethnic communities can be empirically traced with regard to chronology and locality around the world. They are an object of sociological analysis.

2. The Occurrence of Multiculturalism in Western Democracies

Campaigning for multiculturalism has had various causes and different initials. The distinction of *cultures* (in plural) is primarily used in more or less culturally homogeneous nation states who claim responsibility for a certain territory and the people living there, when they are challenged by desired or undesired mass immigration. Multiculturalism, then, is initiated to counteract the threat of disturbing and destabilizing the existing social order and under certain circumstances the racial/ethnic stratification and the balance of power and influence, as well. Multiculturalism is either, presented as a governmental *top down* strategy in the sense of a social technology to manage ethnic conflict, to maintain social integration, and guarantee social peace by means of mass communication and education. Or in reverse, and in response to official multiculturalism, the concept occurs in such immigration situations as a *bottom up* strategy of ethnic community building and ethnic mobilization to prevent the loss of traditional privileges or fight anticipated or real discrimination.

Implementing multiculturalism as well as the chance of ethnic mobilization presupposes indispensable and established political pluralism with the formal guarantee of equal rights and equality of chances to all members of the state. Additionally there is a third motive, confronted with their own history of slavery and their suppression of the indigenous 'first nation' people, the former colonist societies, who are now constitutionally bound to human rights and the promise of social justice, pick up the strategy of managing diversity and fighting discrimination. With multiculturalism they seek to redeem their self-obligations by rounding off the political pluralism, which forms the basis of their constitutional self-image, with a cultural pluralism. The aim is to recognize the existence of previously neglected primordial ethnic bonds that, despite all prophecies, do not seem to be losing their power and thus have remained a permanent challenge to the national state and its policies.

Examples for these different versions of multiculturalism within democratic nation states can be drawn from North America, Australia, and Western Europe. After probably first being launched in 1971 in Canada as an explicit political program by the government, it then was imitated in Australia where the Australian Council on Population and Ethnic Affairs passed a document on 'Multiculturalism for all Australians' in 1982. In the USA, this phenomenon had *avant la lettre*

already taken the form of an 'ethnic revival' movement during the early 1970s.

In Canada the origins of multiculturalism are strictly related to the Quebec question with the periodically renewed secession threat. To water down the Anglo-French confrontation and integrate the bilingual federation, a policy was passed declaring the conservation of cultural variety of all Canadians to be a duty of the state (see Breton 1986). The policy of recognition of cultural identity was then extended to all groups of immigrants as well to the marginalized indigenous peoples who suffer from racism and social exclusion.

In Australia multiculturalism was part of a governmental effort to restrain the traditional widespread and deep-rooted racism in the white majority population. In order to come to a settlement with the aborigine people and above all to facilitate economically inevitable non-white immigration from the surrounding Asian neighbor states, this program also raised ethnic variety to be a perpetual, legitimate characteristic of Australian society that should be protected and promoted by political means (see Castles 1996).

In the case of the USA, which is quite different from the other examples, multiculturalism started as the answer of white immigrants from southeast Europe to the equal rights movement of the black population until then excluded by racial apartheid and segregation. It was a 'multiculturalism from below' that was initially staged as a social mobilization by the 'PIGS' (Poles, Italians, Greeks, and Slovaks). The switch from a 'black' policy of equal legal rights to an originally 'white' campaign of multiculturalism was later also taken up by blacks and other minority groups as a mode of identity policy. Afterwards it became a battlefield in education and a new medium in the struggle for distribution equality ('affirmative action') covering, as well as ethnicity, also race, gender, and sexual orientation as distinction markers.

In all these cases the emerging positive perception of ethnic difference was accompanied and reinforced by the insight within the social sciences that the 'melting pot' was no longer an adequate description of the reality of the migration society (Glazer and Moynihan, 1963; Novak, 1972) and should be replaced by the notion of the 'salad bowl' focussing on the importance of primordial ethnic bonds and the evidence of ethnic community for social life and individual wellbeing. The recognition of ethnic identity became an issue in the public fields of education, health, and the media.

This shift of self-description was received with open arms in northwestern Europe, too. It was first adopted in Great Britain which had become, during the transition process from Empire to Commonwealth of Nations, the goal of huge immigration flows from the Indian subcontinent, Africa, and the Caribbean. The claim for recognition of ethnic diversity and the acceptance of ethnic mobilization was welcomed as a new mode of dealing with social conflicts and unequality (Rex 1996). For some time multiculturalism had a particularly decisive influence on the discussions in the field of education. Here, 'multicultural education' could be interpreted as an attempt by the moral part of society to process the undeniable fact of racist discrimination and violence against immigrants.

The idea of multiculturalism simultaneously was embraced in the Netherlands (which traditionally describes itself as a 'pillar society' of different groups) as well as in some Scandinavian countries. In the context of highly developed welfare states multiculturalism is a new means to present oneself as an addressee of special aid whether in the field of labor, health, or education and also a new field for state activity to promote social equality and justice (Entzinger 1998, Ring 1998).

In West Germany (FRG), multiculturalism in the 1980s served as a field of symbolic policy. The government had tried stubbornly to deny that Germany has, ever since it existed, been an immigration country and refused to allow citizenship to more than seven million immigrants from southern Europe ('guest worker'), to talk about a 'multicultural reality' that needed to be perceived and accepted as such was a benevolent attempt to initiate a change of policy towards political, social and cultural interpretation. In the end, multiculturalism turned out to be a compromise between the necessity of mobility/immigration and the desire for ethnic homogeneity which ended up in 'intercultural education' for all with a revision of school curricula and textbooks (Radtke 1994).

The purpose of this short retrospective sketch of the origins and variations of multiculturalism, which could also cover the specific reasons for its non-reception in southern Europe and France, is merely to demonstrate how the concept fundamentally changes its meaning within each historically and politically shaped national context. There is no unified concept of multiculturalism which has developed a set of approved and coherent principles for the management of ethnic conflicts but there are different local adaptations and instrumentalizations which react to local needs.

3. *The Short Cult of Multiculturalism and its Linkage with 'Postmodernism' in the USA*

Multiculturalism, which had so far mostly been restricted to the discussion of ethnic traditionalism, ethnic community building, curriculum revisions, lessons in native languages, headstart courses, or cookery and dance groups, during the 1990s allied itself surreptitiously with the fashionable discourse on postmodernism (Giroux 1993). According to Ihab Hassan, postmodernism at that time influenced the public discussion like a 'mysterious, if ubiquitous

ingredient—like raspberry vinegar, which instantly turns any recipe into *nouvelle cuisine*' (Hassan 1986, p. 508). This is exactly what happened to multiculturalism when it was presented with a postmodern 'dressing' as the proposed pattern for the society of the new century.

The collapse in the ideologies of the new left, following the fall of the socialist idea as a real utopia after the end of the cold war in 1989, was pushing leftist intellectuals into the arms of the liberalism and pluralism that they had previously criticized so strongly. They seem to have found these palatable in the fashionable form of postmodern relativism. Impressed by the experience of the leveling out and homogenization of both life and thought, the postmodern discourse for a short period has said goodbye to all forms of universalism, totalization, and unity, and welcomed with a somewhat heavy heart particularity, fragmentation, and variety. Postmodernists admired the increasing plurality in the world with the accompanying autonomy of language games and communities with shared interpretations or meanings with cultural traditions of equal value.

The shift toward postmodernism was motivated by the theoretical vacuum that arose with the disintegration of socialism. However, left-wing thought is moralistic, whereas postmodernism offers only an aesthetic criticism. The gap for a few moments was closed by multiculturalism, for it is there that the idea of (cultural) variety can be linked to the highly moral aspiration to protect (ethnic) minorities and their right to self-expression. The element of solidarity that can be rescued by transferring it into a multicultural rhetoric makes it attractive for moralists even when they have to realize that they are aligning themselves with neo-conservative ideologies emerging with the concept of 'communitarism.'

4. *Theoretical and Political Controversies on Multiculturalism*

Multiculturalism in its theoretically most advanced version was elaborated in disintegrating Canada. Based on ethics and the normative ideas of human rights and the right of self-determination of the individual, political philosophers like Taylor (1992) and Kymlicka (1989, 1995) try to deduce collective cultural rights from the individual right of recognition and identity which is said to be realized only with respect to one's *own* culture. Both Taylor and Kymlicka, trying to reconcile liberalism and communitarism, link their sense of ethnic embeddedness of the individual in 'rich cultural structures' with the idea of political pluralism and a regime of tolerance and acceptance (Walzer 1997). The ongoing controversy again follows the mentioned lines of constructivism versus realism.

Critics of Taylor and Kymlicka (collected in Gutman 1992), among them Habermas and Walzer, have argued that even though liberal multiculturalism with its postmodern overtones joins political pluralism in opposing any form of totalitarianism, now identified as ethnocentrism, it cannot avoid the reification and essentialization of culture and identity. Liberal multi culturalism, too, translates the concept of a plurality of interests into a plurality of cultural and ethnic origins. For the empty space between the state and the individual, it does not offer an autonomous group organized to lobby for limited (economic, political, social) interests, but proposes that this function be taken by the posited community of those who consider that they share some fixed values and at the same time common characteristics such as religion, language, history, or origins. These primordial features are highly attractive and have enormous suggestive and deterministic power. This is especially true in the modern functionally differentiated society where the individual has no secure and firm position but is involved only temporarily with parts of the person in several social systems simultaneously. But ethic distinctions make the resolution of regular social conflict rather unlikely. Once one has identified with one's *own* culture, one cannot compromise or simply walk away from it in the same way that one can resign from a trade union or an action group organized to oppose the construction of a highway.

This results in two systematic difficulties for any amalgamation of political pluralism with cultural pluralism: ethnic communities cannot function as *pouvoirs intermédiares* in the sense of Montesquieu. They are not able to guarantee the minimal democratic consensus that is indispensable in pluralistically constituted societies in order to safeguard the procedural rules used in the process of balancing interests. The organizational principle of such communities is exclusiveness. As far as issues of cultural identity, religious norms, or even ethnic origins are concerned, there are no compromises. Differences become irreconcilable.

Even if there is a strong individual need for community building and social integration, multiculturalism is a regressive proposal for dealing with the problems of modern society. It is regressive in the psychoanalytical sense of going back to an earlier stage of psychogenetic development that recreates the basic triad of native *child* of a country, *father*land, and *mother*tongue. And it is regressive in the historical sense in that it maintains a political pattern of differentiation that first arose during the making of the nation state in the nineteenth century, which is, in a functionally differentiated and globalized society paradoxically up to and out of date simultaneously.

Schlesinger (1991) has pointed to the costs of multiculturalism for liberal democracies. In those states in which the program of multiculturalism has been raised to official policy, there have been few

notable positive effects but very clear side effects. Controversial debates on alarming trends are reported from the USA in the 1990s. At universities and colleges, where the citizen's rights movements once arose, an ethnic and racial separatism could be observed that covered not only student life on campus, but was also increasingly expanding into teaching. One pursued 'ethnic studies,' returns to seating arrangements according to 'ethnics' or 'race,' and so forth—not just in the dining hall, but also in seminars and in the library. Whereas the efforts toward 'multiculturalism' were expected to lead to ethnic harmony, one actually finds a disuniting (self-imposed) segregation and a growing, malignant racism advanced with the pathos of self-determination, identity, and emancipation (see Aufderheide 1992).

This trend is the consequence of the ambivalence of particularism and universalism written into multiculturalism. In particular, multiculturalism tries to legitimize the role of ethnic borders as a form of exercising rights of cultural self-determination. Through this, it revitalizes patterns of differentiation that have to fill the gap in the functionally differentiated society. They therefore have not lost their subjective attractiveness. By dressing them with science, upgrading them to analytical categories, and making them politically acceptable, politicians and social scientists construct a 'multicultural society' that establishes itself as a social reality to the extent that the participants in the social process direct their practices toward this orientation.

5. *Export of Multiculturalism to Eastern and Central Europe*

Albeit the 'ethnic' and 'race' relations in none of the Western liberal democracies are in an exemplary status, the concept of multiculturalism is attentively observed in so-called transforming societies in Eastern and Central Europe, and is also offered as a model of problem solving by international organizations (Organization for Security and Cooperation in Europe, OSCE, United Nations, UN, Kosovo Force, KFOR) and Western foundations sponsoring conferences and publications. Whenever ethnically defined nation building by previously oppressed national minorities takes place, this is the hour of birth of new ethnic minorities along language or religion lines. The status of these new groups may rapidly become critical. Multiculturalism is introduced to the successor nations of the former Soviet Union (especially the Baltic States and in the Caucasus) and in former Yugoslavia (especially Bosnia and Kosovo) with the aim to contribute to the pacification of ethnic conflicts arising in the aftermath of the eroding Empires or Unions being replaced by new ethnic nation states. Even if the intention is not to propose the transplantation of institutions and policies from West to East (Kymlicka 2000), multiculturalism in fact is meant and received as an ideological instrument to implement the founding principles of Western political pluralism: moderate coercion to promote national identity, which is reduced to few principles of political culture laid down in the constitution; strict division of 'public' and 'private' spheres; freedom of ethnic mobilization and community building; full citizenship for all ethnic groups with a low threshold of access; intercultural openness; commitment to the recognition of differences of all kind. These principles represent the program of Western pluralism only but not the reality of interethnic relations or even the state of conflict management in these countries. In addition, the probability is rather low that they will work in a climate just after a long-lasting period of totalitarian suppression or civil war. Nevertheless, they seem to be without any alternative as long as ethnically defined nation states freed ethnic minorities.

See also: Diversity and Disagreement; Ethnic Conflicts; Ethnic Groups/Ethnicity: Historical Aspects; Globalization and World Culture; Multiculturalism; Multiculturalism and Identity Politics: Cultural Concerns; Multiculturalism, Anthropology of; Pluralism; Postmodernism: Philosophical Aspects

Bibliography

Aufderheide P (ed.) 1992 *Beyond PC: Towards a Politics of Understanding*. Graywolf, Saint Paul, MN

Breton R 1986 Multiculturalism and Canadian nation-building. In: Cairns A, Williams C (eds.) *The Politics of Gender, Ethnicity and Language in Canada*. University of Toronto Press, Toronto, Canada

Castles S 1996 Immigration and multiculturalism in Australia. In: Bade K J (ed.) *Migration, Ethnizität, Konflikt*. Rasch Universitätsverlag, Osnabrück, Germany, pp. 251–71

Dittrich E J, Radtke F-O (eds.) 1990 *Ethnizität: Wissenschaft und Minderheiten*. Westdeutscher Verlag, Opladen, Germany

Entzinger H 1998 Zu einem Modell der Inkorporation von Einwanderern: das Beispiel der Niederlande. In: Bommes M, Halfmann J (eds.) *Migration in Nationalen Wohlfahrtsstaaten*. Rasch Universitätsverlag, Osnabrück, Germany, pp. 105–22

Giroux H A 1993 *Living Dangerous: Multiculturalism and the Politics of Difference*. Peter Lang, New York

Glazer N, Moynihan D P 1963 *Beyond the Melting Pot: The Negroes, Puerto Ricans, Jews, Italians, and Irish of New York City*. Harvard University Press, Cambridge, MA

Gutman A (ed.) 1992 *Multiculturalism and the Politics of Recognition*. Princeton University Press, Princeton, NJ

Hassan I 1986 Pluralism in postmodern perspective. *Critical Inquiry* **12**: 503–20

Kymlicka W 1989 *Liberalism, Community, and Culture*. Oxford University Press, Oxford, UK

Kymlicka W 1995 *Multicultural Citizenship: A Liberal Theory of Minority Rights*. Oxford University Press, Oxford, UK

Kymlicka W 2000 Nation-building and minority rights: Comparing West and East. *Journal of Ethnic and Migration Studies* **26**: 183–212

Luhmann N 1982 *The Differentiation of Society*. Columbia University Press, New York
Meyer J W, Kamens D H, Benarot A, Yun-Kyung C, Suk-Ying W 1992 *School Knowledge for the Masses: World Models and National Primary Curricular Categories in the Twentieth Century*. Falmer Press, London
Novak M 1972 *The Rise of the Unmeltable Ethnics: Politics and Culture in the Seventies*. Macmillan, New York
Radtke F-O 1994 The formation of ethnic minorities and the transformation of social into ethnic conflicts in a so-called multi-cultural society: The case of Germany. In: Rex J, Drury B (eds.) *Ethnic Mobilization in a Multi-Cultural Europe*. Avebury, Aldershot, UK, pp. 30–7
Rex J 1996 Ethnic minorities in the modern nation state. In: *Working Papers in the Theory of Multiculturalism and Political Integration*. Macmillan/St. Martin's Press, London/New York
Ring H 1998 Einwanderungspolitik im schwedischen Wohl-fahr-ts-staat. In: Bommes M, Halfmann J (eds.) *Migration in nationalen Wohlfahrtsstaaten*. Rasch Universitätsverlag, Osnabrück, Germany, pp. 239–49
Schlesinger Jr A 1991 *The Disuniting of America: Reflections on a Multicultural Society*. Whittle Direct Books Knoxville, TN
Searle J R 1995 *The Construction of Reality*. The Free Press, New York
Sollors W 1989 *The Invention of Ethnicity*. Oxford University Press New York/London
Stichweh R 2001 *Die Weltgesellschaft*. Suhrkamp, Frankfurt am Main, Germany
Taylor C 1992 *Multiculturalism and The Politics of Recognition*. Princeton University Press, Princeton, NJ
Thomas G M, Meyer J W, Ramirez I, Boli J 1987 *Institutional Structure: Constituting State, Society, and the Individual*. Sage Publications, Newbury Park, CA
Walzer M 1997 *On Toleration*. Yale University Press, New Haven

F.-O. Radtke

Multidimensional Scaling in Psychology

Multidimensional scaling (MDS), as defined in this article, is a family of models and methods for representing proximity data in terms of spatial models in which proximities (e.g., similarities or dissimilarities of pairs of stimuli or other objects) for one or more subjects (or other sources of data) are related by some simple, well-defined (e.g., linear or monotonic) function(s) to (Euclidean or other) distances between those objects in the multidimensional spatial representation, whose (original or rotated) coordinate axes are then 'interpreted' in terms of meaningful psychological (or other) dimensions or attributes.

In *nonmetric* MDS only the *rank orders* of proximities are used. In one widely cited example, the map of 10 US cities was reconstructed so well, using the rank orders of pairwise distances from a table of airline distances, that the map could be almost perfectly superimposed on a standard map of the United States, once the scale factor was adjusted appropriately, and a rigid rotation was applied to the MDS map to align its coordinates with the north–south and east–west coordinates of the standard map.

Rotation of coordinates is not a trivial issue (at least when Euclidean distances are assumed), but is essential to interpretation of coordinates as meaningful dimensions. Certain coordinate systems correspond to 'natural' easily interpretable dimensions, whereas the particular coordinate system provided by an MDS method may be highly arbitrary, requiring rotation for reasonable interpretability. While rotation of a two- or three-dimensional representation can often be accomplished by visual inspection, in higher dimensionalities rotation usually becomes prohibitively difficult—which is probably why there are few examples of what is called 'two-way' MDS in more than three dimensions. We later discuss how 'three-way,' or individual differences, MDS—INDSCAL (IN dividual Difference multidimensional SCALing) in particular—can solve this rotational problem.

We now focus on MDS models and methods for one-mode (e.g., stimuli), two-way (i.e., pairwise) proximity data, which can arise from direct human judgments of dis/similarity, from confusability of pairs of stimuli, or from various other types of derived measures of dis/similarity (e.g., a 'profile dissimilarity' measure computed between stimuli over rating scales, or measures of similarity derived from word association tasks). Matrices of correlations can be used as measures of similarity, so that MDS can also be viewed as an alternative to factor analysis for multidimensional analysis of correlational data.

These proximities comprise a two-way square table (or matrix)—usually (but not always) symmetric. We call MDS analysis of such data 'two-way' MDS. An important distinction is between *metric* MDS and *nonmetric* MDS. In metric approaches, proximities are assumed measured on at least *an interval scale*. Nonmetric MDS assumes proximities measured only on *an ordinal scale*.

The earliest approach to MDS, now often called the 'classical' metric (two-way) MDS method (Torgerson 1958), stems back to theoretical work by Young and Householder in 1938, applied by Richardson in 1938 to unidimensional visual stimuli. This 'classical' approach assumes a linear relationship between proximities and (Euclidean) distances.

Nonmetric MDS assumes ordinal scale proximities, and a merely monotonic function relating proximities to distances. The earliest proposal for nonmetric MDS was made by Coombs (1964) but did not result in a practical algorithm. The first practical computer implemented method was 'analysis of proximities' proposed by Shepard in 1962. This was followed by a more mathematically rigorous approach, which remains the prototypical approach to nonmetric MDS, proposed by Kruskal in 1964, which we describe

below, as currently implemented in the computer program KYST2-A.

Kruskal assumes proximities between stimuli j and k, δ_{jk}, are related to distances, d_{jk}, in an underlying psychological space via a function, f transforming proximities into (approximate) distances $\hat{d}_{jk}$, so that $\hat{d}_{jk} = f(\delta_{jk})$. In the case of a nonmetric analysis, the function f can be any monotonic function. Kruskal's approach allows fitting a spatial representation in R dimensions, allowing any of a general class of metrics (or distance functions), called 'Minkowski-p' (or, by mathematicians, L_p) metrics, of the form:

$$d_{jk} = \left(\sum_{r=1}^{R} \left| x_{jr} - x_{kr} \right|^{p} \right)^{\frac{1}{p}}$$

where p is generally assumed $\geqslant 1$. The Euclidean metric, in which $p = 2$, is the default option in Kruskal's (and most other) algorithm(s). Kruskal uses a well known numerical optimization procedure called the gradient method, combined with a method of least squares monotone regression (for fitting the monotonic function f) to fit the best R-dimensional spatial representation by minimizing a measure of 'badness of fit' called STRESS, defined (in its original, and most widely used form) as:

$$\text{STRESS} = \sqrt{\left(\frac{\sum_{jk} (d_{jk} - \hat{d}_{jk})^2}{\sum_{j,k} d_{jk}^2} \right)}$$

where the summations are over all values of j and k for which data are provided (so that, unlike 'classical' MDS, Kruskal's algorithm easily handles missing data). More generally a *weighted* generalization in which each term in the summations in both the numerator and denominator is weighted by nonnegative weights input by the user (or calculated as a function of δ_{jk}) can be substituted for (unweighted) STRESS.

If the user is not certain of the appropriate dimensionality, solutions may be obtained for a range of values of number of dimensions, R, and a plot of STRESS vs. dimensionality, plus criteria based on interpretability, used to choose the 'best' value of R. See Kruskal and Wish (1978) for further discussion, and a general overview of metric and nonmetric two-way MDS.

While Euclidean distances are most often assumed, other values of p in the Minkowski-p (or L_p) metric can be used. In practice, the only alternative value of p that has been used to any extent is $p = 1$. This leads to what is often called the 'city block' or 'Manhattan' metric, for which there is considerable theoretical basis, especially in the case of what Shepard and others have called 'analyzable' or 'nonunitary' (often, but not necessarily, more 'conceptual' than 'perceptual') stimuli. However, as discussed in detail by Arabie et al. (1987) it is inadvisable to use Kruskal's algorithm (or other similar algorithms) to fit a city block model, because of a particularly severe '*local minimum*' *problem* (which can also less seriously affect fitting Euclidean models—but this can be dealt with reasonably effectively by procedures described by Kruskal and Wish 1978 and elsewhere). Kruskal's approach (and many others usually characterized as 'nonmetric' MDS) also allow(s) the function, f, to be linear, in which case the analysis is a metric one, assuming interval or even ratio scale data (depending on whether the linear function includes an intercept, or additive constant term, or not). Other functions are also allowed.

Other approaches to two-way metric or nonmetric MDS have been proposed by Guttman, ('Smallest Space Analysis'), by Takane et al. (ALSCAL), Ramsay (MULTISCALE), and Heiser and de Leeuw (SMACOF). Ramsay's MULTISCALE is a maximum likelihood approach, which is more soundly based statistically than other approaches. Winsberg and Carroll (1989) proposed a 'quasi-nonmetric' approach to fitting an extended Euclidean model (assuming dimensions 'specific' to individual stimuli in addition to common dimensions underlying all the stimuli) which is also fitted via maximum likelihood procedures.

1. *Individual Differences Multidimensional Scaling*

The first approach to individual differences MDS was the 'points of view' approach proposed by Tucker and Messick in 1963. This was followed by the now dominant approach, generally called INDSCAL (Carroll and Chang 1970), for *IN*dividual *D*ifferences multidimensional *SCAL*ing. INDSCAL, which stands ambiguously for a *model*, a *method*, and a specific *computer program* implemented by Carroll and Chang, accounts for individual differences in proximity data via a model assuming a common set of stimulus/object dimensions, while assuming that different subjects (or other data sources) have different patterns of saliences or importance weights for these common dimensions. In the case of similarity judgments by human subjects, the INDSCAL model makes the quite plausible assumption that each subject has a different set of scale factors (amplifications or attenuations) applied to a set of 'fundamental' psychological stimulus dimensions common to all individuals. It is as if each subject has a set of 'gain controls,' one for each of the fundamental stimulus dimensions, and that these are differently adjusted for each subject. These different weights might be due to genetic differences, to environmental factors related to different experiential histories, or, most likely, to a combination of nature (genetics) and nurture (environment).

One strength of INDSCAL is 'dimensional uniqueness'—the property that the (estimated) stimulus dimensions are uniquely determined by similarity judgments of subjects (based on the patterns of individual differences in proximities). Equally importantly, the saliences, or perceptual importance weights of dimensions, can provide quite useful individual differences measures for subjects.

The data for individual differences MDS generally comprise a number of symmetric proximity matrices, one for each subject (or other data source). Such data are generally designated as two-mode (e.g., stimuli and subjects) but three-way (e.g., stimuli × stimuli × subjects). A 'mode' is a specific set of entities (e.g., stimuli). The number of 'ways' can be thought of as the number of 'directions' in the data table (e.g., rows, columns, and 'slices' for a three-way table, where both rows *and* columns correspond to the stimulus mode, while 'slices' correspond to the subject mode). The purpose of INDSCAL analysis is, given such three-way proximity data, to solve simultaneously for the stimulus coordinates and the subject weights so as to optimize the fit of the INDSCAL model to the (transformed) proximity data.

The INDSCAL method makes *metric* assumptions and entails a kind of three-way generalization of the 'classical' method for two-way MDS. The most efficient approach for implementing this analysis is provided by a program called SINDSCAL written by Pruzansky in 1975, which improves on the Carroll and Chang INDSCAL *program* for fitting the INDSCAL *model* via the INDSCAL *method*, in the specific case of three-way *symmetric* proximity data.

2. The INDSCAL Model and Method

Let $\delta_{jk}^{(i)}$ be the measure of proximity between stimuli/objects j and k for subject/source i. The INDSCAL model assumes that proximities are approximately related to generalized (weighted) Euclidean distances via a function f_i (different for each subject or source), so that :

$$f_i(\delta_{jk}^{(i)}) \cong d_{jk}^{(i)} = \left[\sum_{r=1}^{R} w_{ir}(x_{jr} - x_{kr})^2 \right]^{\frac{1}{2}}$$

where '$\cong$' means 'equals, except for (unspecified) error terms.'

The INDSCAL method, most effectively implemented in Pruzansky's SINDSCAL program, assumes f_i is linear for all i, and uses a generalization of the approach used in the 'classical' metric approach to two-way MDS. Where the latter uses the SVD (singular value decomposition) to derive coordinates of n stimuli on R dimensions from a *single* derived matrix of 'scalar products,' INDSCAL/SINDSCAL uses a generalization of the SVD called CANDECOMP (for

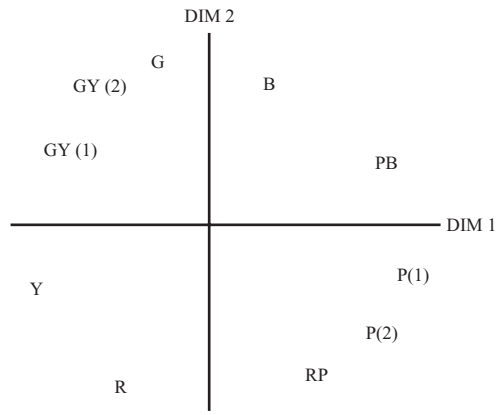

Figure 1
The 'group' stimulus space. Stimuli are Munsell color chips, with saturation and lightness held constant. Dimension 1 (DIM 1) can be interpreted as blue vs. yellow, and dimension 2 (DIM 2) as red vs. green. (Coding for hues: R = red; Y = yellow; GY(1) and GY(2) = green yellow, GY(2) containing more green; G = green; B = blue; PB = purple blue; P(1) and P(2) = purple, P(2) containing more red; RP = red purple)

*CAN*onical *DECOMP*osition of N-way tables) simultaneously to fit coordinates of the R dimensions (x_{jr}, for $j = 1, 2, \cdots n$ and $r = 1, 2, \cdots R$) for the n stimuli *and* weights for m (w_{ir}, for $i = 1, 2, \cdots m$ and $r = 1, 2, \cdots R$) to a three-way array of derived scalar products. The details of the INDSCAL method for fitting the INDSCAL model are described by Carroll and Chang (1970). Other methods mentioned below for fitting the INDSCAL model make various other assumptions about the relations between proximities and weighted Euclidean distances.

In addition to accounting for individual differences among subjects or other data sources via differential patterns of weights for the R dimensions, INDSCAL has the important property—often called 'dimensional uniqueness'—that the dimensions are—under very general conditions—uniquely oriented, and not subject to rotation. A set of necessary and sufficient conditions for this property to hold is that the subject weights are *not* all equal, or proportional, to one another for different subjects, for *any subset of* two or more *dimensions*.

This is true because a rotation of coordinate axes in the 'group' stimulus space can be shown *not* to yield the same set of weighted distance matrices (even if subject weights are recomputed for the rotated dimensions) as the 'preferred' orientation determined by the procedure fitting the INDSCAL model, so that the measure of goodness of fit to the proximity data will not be as good as for this 'preferred' orientation. This statistical uniqueness of INDSCAL dimensions rein-

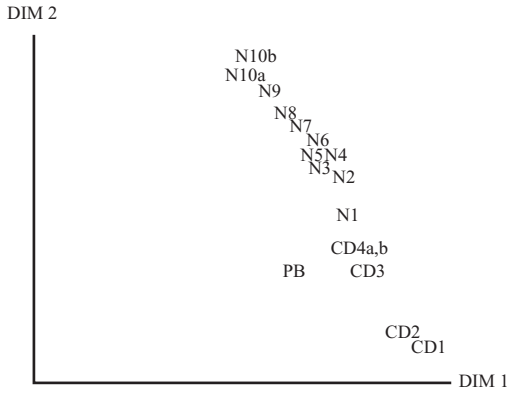

Figure 2
The subject space (CD1 through CD4 are color deficient. N1 through N10 are normal. Replications of two S's are indicated by small-case letters)

forces the fact that INDSCAL dimensions can be considered to be psychologically 'fundamental.' This theoretical and statistical uniqueness of INDSCAL dimensions has been supported in numerous applications in which INDSCAL dimensions, even in fairly high dimensionalities (up to nine dimensions in some cases), are almost always interpretable *without* rotation.

We illustrate INDSCAL by an example involving 'points of view' analysis of data on color perception due to Helm. In 1962 Helm and Tucker analyzed these data via this model, and found that about ten 'points of view' were necessary to account adequately for all the individual differences among subjects. The INDSCAL analysis accounted quite nicely for the individual differences among the 14 subjects in terms of a two dimensional solution, illustrated in Fig. 1 easily interpretable in terms of two dimensions that accord very well with physiological and psychophysical evidence, strongly suggesting the existence of 'blue–yellow' and 'red–green' receptors. Dimension one corresponds quite well to a 'blue vs. yellow' (or 'purple–blue vs. green–yellow') dimension, and dimension two to a 'red vs. green' (or 'purple–red vs. blue–green') dimension, as can be seen in Fig.1.

Helm included four subjects with varying degrees of red-green color deficiency, labeled (in the 'subject space' shown in Fig. 2) as subjects CD1 through CD4 (a and b in the last case, since the data for this subject were collected twice, about a month separating the two replications). As can be seen in Fig. 2, all the red–green color-deficient subjects had smaller weights for the red–green dimension than any of the normal subjects (coded N1 through N10 a and b). This means that the red–green dimension is less salient perceptually to the four subjects with variants of the most common types of anomalous color vision (protanomalous or deuteranomalous vision—two types of deficiency in distinguishing red from green), as would be predicted on theoretical grounds. In what is often called the 'private perceptual space,' constructed by applying the (square roots of the) subject weights, for each subject, as rescaling factors to the dimensions, the space for color-deficient subjects is compressed in the 'red–green' direction, so that for them red and green are perceived as being much closer (more similar) relatively speaking than is the case for the subjects with normal color vision.

While in one sense this INDSCAL analysis merely confirms what was learned by over a hundred years of careful physiological and psychophysical studies, had we *not* been aware of these studies these results might well have been helpful to scientists seeking to understand more fully the nature of both normal and anomalous color vision. In many other domains of study (see Wish and Carroll 1974 or Arabie et al. 1987 for examples) psychologists and other behavioral or social scientists *have* in fact learned a great deal about both the psychological (or other) dimensions, and the nature of individual differences underlying perception of important classes of stimuli (or of measured proximities of other 'objects').

Various alternative nonmetric and metric algorithms have been developed for fitting the INDSCAL model. Most prominent are Takane et al.'s ALSCAL (in its three-way implementation) and Ramsay's MULTISCALE procedure, which provides methods for fitting the three-way INDSCAL model based on a maximum likelihood criterion. Carroll and Winsberg (1995) generalized their approach based on the extended Euclidean model to an extended INDSCAL model, called EXSCAL, fitted by what is called a 'quasi-nonmetric' approach, using a maximum likelihood procedure. More recently a three-way version of SMACOF, called PROXSCAL, has been developed and is distributed by Heiser and colleagues, and has been available in SPSS, version 10 or later, since mid-2000. This is probably the best approach for a purely nonmetric analysis of three-way proximity data in terms of the INDSCAL model.

3. *Broad Definition of Multidimensional Scaling*

In its broadest definition, MDS includes a wide variety of geometric models, not only spatial distance models, for representation of psychological or other data. These can include *discrete* geometric models such as tree structures (often associated with hierarchical clustering), overlapping or nonoverlapping cluster structures, or more general cluster or network models. While MDS is typically associated with continuous spatial distance models for representation of proximities, the broad definition also includes other spatial models, such as what are called the vector or the

unfolding/ideal point models for representation of individual differences in preference (or other dominance) data, or two- or three-way factor/components analysis models—many of which are nondistance based. For further information on MDS, in both its narrow and broader definitions, as well as for further details and/or applications of two-way and three-way MDS, see Carroll (1976, 1980), Shepard (1978), Carroll and Arabie (1980, 1998), Meulman (1992), Klauer and Carroll (1995), Carroll and Chaturvedi (1995), Carroll and Wish (1974a, 1974b), De Soete and Carroll (1992, 1996), and/or Wish and Carroll (1974).

See also: Factor Analysis and Latent Structure: Overview; Procrustes Analysis; Psychometrics; Scaling: Multidimensional; Unfolding and Vector Models

Bibliography

Arabie P, Carroll J D, DeSarbo W S 1987 *Three-Way Scaling and Clustering*. Sage, Newbury Park, CA

Carroll J D 1976 Spatial, non-spatial and hybrid models for scaling. (Presidential Address for Psychometric Society). *Psychometrika* **41**: 439–63

Carroll J D 1980 Models and methods for multidimensional analysis of preferential choice (or other dominance) data. In: Lantermann E D, Feger H (eds.) *Similarity and Choice*. Hans Huber, Bern, Switzerland

Carroll J D, Arabie P 1980 Multidimensional scaling. In: Rosenzweig M R, Porter L W (eds.) *Annual Review of Psychology*. Annual Reviews, Palo Alto, CA

Carroll J D, Arabie P 1998 Multidimensional scaling. In: Birnbaum M H (ed.) *Handbook of Perception and Cognition. Vol. 3: Measurement, Judgment and Decision*. Academic Press, San Diego, CA

Carroll J D, Chang J J 1970 Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart–Young' decomposition. *Psychometrika* **35**: 283–319

Carroll J D, Chaturvedi A 1995 A general approach to clustering and multidimensional scaling of two-way, three-way, or higher-way data. In: Luce R D, D'Zmura M, Hoffman D D, Iverson G, Romney A K (eds.) *Geometric Representations of Perceptual Phenomena*. Erlbaum, Mahwah, NJ

Carroll J D, Wish M 1974a Models and methods for three-way multidimensional scaling. In: Krantz D H, Atkinson R C, Luce R D, Suppes P (eds.) *Contemporary Developments in Mathematical Psychology*. W H Freeman, San Francisco, CA, Vol. 2

Carroll J D, Wish M 1974b Multidimensional perceptual models and measurement methods. In: Carterette E C, Friedman M P (eds.) *Handbook of Perception*. Academic Press, New York, Vol. 2

Coombs C H 1964 *A Theory of Data*. Wiley, New York

De Soete G, Carroll J D 1992 Probabilistic multidimensional models of pairwise choice data. In: Ashby F G (ed.) *Multidimensional Models of Perception and Cognition*. Erlbaum, Hillsdale, NJ

De Soete G, Carroll J D 1996 Tree and other network models for representing proximity data. In: Arabie P, Hubert L J, De Soete G (eds.) *Clustering and Classification*. World Scientific, River Edge, NJ

Klauer K C, Carroll J D 1995 Network models for scaling proximity data. In: Luce R D, D'Zmura M, Hoffman D D, Iverson G, Romney A K (eds.) *Geometric Representations of Perceptual Phenomena*. Erlbaum, Mahwah, NJ

Kruskal J B, Wish M 1978 *Multidimensional Scaling*. Sage, Beverly Hills, CA

Meulman J J 1992 The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika* **57**: 539–65

Shepard R N 1978 The circumplex and related topological manifolds in the study of perception. In: Shye S (ed.) *Theory Construction and Data Analysis in the Behavioral Science* 1st edn.. Jossey-Bass, San Francisco, CA

Torgerson W S 1958 *Theory and Methods of Scaling*. Wiley, New York

Wish M, Carroll J D 1974 Applications of individual differences scaling to studies of human perception and judgment. In: Carterette E C, Friedman M P (eds.) *Handbook of Perception: Psychophysical Judgment and Measurement*. Academic Press, New York

J. D. Carroll

Multimodal Therapy in Clinical Psychology

The outcomes of several follow-up inquiries pointed to the importance of breadth if treatment gains were to be maintained, and led to the development of the multimodal approach. Emphasis was placed on the fact that at base, we are biological organisms (neurophysiological/biochemical entities) who behave (act and react), emote (experience affective responses), sense (respond to tactile, olfactory, gustatory, visual, and auditory stimuli), imagine (conjure up sights, sounds, and other events in our mind's eye), think (entertain beliefs, opinions, values, and attitudes), and interact with one another (enjoy, tolerate, or suffer various interpersonal relationships). By referring to these seven discrete but interactive dimensions or modalities as Behavior, Affect, Sensation, Imagery, Cognition, Interpersonal, Drugs/Biologicals, the convenient acronym BASIC I.D. emerges from the first letter of each one.

The BASIC I.D. or multimodal framework rests on a broad social and cognitive learning theory (Bandura 1977, 1986, Rotter 1954) because its tenets are open to verification or disproof. Instead of postulating putative complexes and unconscious forces, social learning theory posits testable developmental factors (e.g., modeling, observational, and enactive learning, the acquisition of expectancies, operant, and respondent conditioning, and various self-regulatory mechanisms). While drawing on effective methods from any discipline, the multimodal therapist does not embrace divergent theories but remains consistently within

social-cognitive learning theory. The virtues of technical eclecticism (Lazarus 1992, Lazarus et al. 1992) vs. the dangers of theoretical integration have been emphasized (e.g., Lazarus 1995, Lazarus and Beulter 1993). The major criticism of theoretical integration is that it inevitably tries to blend incompatible notions and only breeds confusion.

1. *The Holistic and Comprehensive Nature of the Multimodal Orientation*

The polar opposite of the multimodal approach is the Rogerian or Person-centered orientation that is entirely conversational and virtually unimodal. While the relationship between therapist and client is highly significant and sometimes 'necessary and sufficient,' in most instances, the doctor-patient relationship is but the soil that enables the techniques to take root. A good relationship, adequate rapport, a constructive working alliance are 'usually necessary but often insufficient' (Fay and Lazarus 1993).

Many psychotherapeutic approaches are trimodal, addressing affect, behavior, and cognition—ABC. The multimodal approach provides clinicians with a comprehensive template. By separating sensations from emotions, distinguishing between images and cognitions, emphasizing both intraindividual and interpersonal behaviors, and underscoring the biological substrate, the multimodal orientation is most far-reaching. By assessing a client's BASIC I.D. one endeavors to 'leave no stone unturned.'

The elements of a thorough assessment involve the following range of questions:

(a) What is this individual doing that is getting in the way of his or her happiness or personal fulfillment (self-defeating actions, maladaptive behaviors)? What does the client need to increase and decrease? What should he/she stop doing and start doing? (B)

(b) What emotions (affective reactions) are predominant? Are we dealing with anger, anxiety, depression, combinations thereof, and to what extent (e.g., irritation vs. rage; sadness vs. profound melancholy)? What appears to generate these negative affects—certain cognitions, images, interpersonal conflicts? And how does the person respond (behave) when feeling a certain way? It is important to look for interactive processes—what impact does various behaviors have on the person's affect and vice versa? How does this influence each of the other modalities? (A)

(c) Are there specific sensory complaints (e.g., tension, chronic pain, tremors)? What feelings, thoughts and behaviors are connected to these negative sensations? What positive sensations (e.g., visual, auditory, tactile, olfactory, and gustatory delights) does the person report? This includes the individual as a sensual and sexual being. When called for, the enhancement or cultivation of erotic pleasure is a viable therapeutic goal. (S)

(d) What fantasies and images are predominant? What is the person's 'self-image?' Are there specific success or failure images? Are there negative or intrusive images (e.g., flashbacks to unhappy or traumatic experiences)? And how are these images connected to ongoing cognitions, behaviors, affective reactions, etc.? (I)

(e) Can we determine the individual's main attitudes, values, beliefs, and opinions? What are this person's predominant shoulds, oughts, and musts? Are there any definite dysfunctional beliefs or irrational ideas? Can we detect any untoward automatic thoughts that undermine his or her functioning? (C)

(f) Interpersonally, who are the significant others in this individual's life? What does he or she want, desire, expect, and receive from them, and what does he or she, in turn, give to and do for them? What relationships give him/her particular pleasures and pains? (I)

(g) Is this person biologically healthy and health conscious? Does he or she have any medical complaints or concerns? What relevant details pertain to diet, weight, sleep, exercise, alcohol and drug use? (D)

The foregoing are some of the main issues that multimodal clinicians traverse while assessing the client's BASIC I.D. A more comprehensive problem identification sequence is derived from asking most clients to complete the Multimodal Life History Inventory (Lazarus and Lazarus 1991). The Inventory is useful for speeding up routine history taking and readily provides the therapist with a BASIC I.D. analysis. Most psychiatric outpatients who are reasonably literate will comply.

2. *Placing the BASIC I.D. in Perspective*

In multimodal assessment, the BASIC I.D. serves as a template to remind us to examine each of the seven modalities and their interactive effects. It implies that we are social beings who move, feel, sense, imagine, and think, and that at base we are biochemical-neurophysiological entities. The seven modalities are by no means static or linear but exist in a state of reciprocal transaction.

Why bother to work multimodally—why involve the entire BASIC I.D. when feasible? Follow-up studies that have been conducted since 1973 consistently have suggested that durable outcomes are in direct proportion to the number of modalities deliberately traversed. Although there is obviously a point of diminishing returns, it is a multimodal maxim that the more someone learns in therapy, the less likely they are to relapse. In this connection, circa 1970, it became apparent that *lacunae* or gaps in people's coping responses were responsible for many relapses. This occurred even after they had been in various (non-multimodal) therapies, often for years on end. It is

important to emphasize that it takes up very little extra time to assess and ameliorate the most salient problems across a patient's BASIC I.D. Follow-up indicates that this ensures far more compelling and durable results (Lazarus 2000). Brief multimodal therapy has been fully explicated (Lazarus 1997). MMT takes Paul's (1967, p. 111) mandate very seriously: 'WHAT treatment, by WHOM, is most effective for THIS individual with THAT specific problem and under WHICH set of circumstances? But in addition to techniques of choice, the multimodal clinician tries to be an authentic chameleon who also asks about relationships of choice. Decisions regarding different relationship stances or styles include when and how to be directive, supportive, reflective, cold, warm, tepid, gentle, tender, tough, earthy, chummy, casual, informal, or formal.

3. Some Specific Features of Multimodal Therapy

3.1 Bridging

A strategy that is probably employed by most effective therapists can readily be taught to novices via the BASIC I.D. format. It has been referred to as Bridging. Let's say a therapist is interested in a client's emotional responses to an event. 'How did you feel when your parents showered attention on your brother but left you out?' Instead of discussing their feelings, the client responds with defensive and irrelevant intellectualizations. 'My parents had strange priorities and even as a kid I used to question their judgment. Their appraisal of my brother's needs was way off—they saw him as deficient whereas he was quite satisfied with himself.' Additional probes into their feelings only yield similar abstractions. It is often counterproductive to confront the client and point out that they are evading the question and seem reluctant to face their true feelings. In situations of this kind, bridging is usually effective. First, the therapist deliberately tunes into the client's preferred modality—in the case, the cognitive domain. Thus, the therapist explores the cognitive content. 'So you see it as a consequence involving judgments and priorities. Please tell me more.' In this way, after perhaps a 5–10 minute discourse, the therapist endeavors to branch off into other directions that seem more productive. 'Tell me, while we have been discussing these matters, have you noticed any sensations anywhere in your body?' This sudden switch from Cognition to Sensation may begin to elicit more pertinent information (given the assumption that in this instance, Sensory inputs are probably less threatening than Affective material). The client may refer to some sensations of tension or bodily discomfort at which point the therapist may ask him or her to focus on them. 'Will you please close your eyes, and now feel that neck tension. (Pause). Now relax deeply for a few moments, breathe easily and gently, in and out, just letting yourself feel calm and peaceful.' The feelings of tension, their associated images, and cognitions may then be examined. One may then venture to bridge into Affect. 'Beneath the sensations, can you find any strong feelings or emotions? Perhaps they are lurking in the background.' At this juncture it is not unusual for clients to give voice to their feelings. 'I feel a lot of anger and sadness.' By starting where the client is and then bridging into a different modality, most clients then seem to be willing to traverse the more emotionally charged areas they had been avoiding.

There are two other specific MMT procedures that should be mentioned. The first is called 'Tracking the ''Firing Order'' of Specific Modalities,' and the other is 'Second-order BASIC I.D. Assessments.'

3.2 Tracking the 'Firing Order'

A fairly reliable pattern may be discerned behind the way in which people generate negative affect. Some dwell first on unpleasant sensations (palpitations, shortness of breath, tremors), followed by aversive images (pictures of disastrous events), to which they attach negative cognitions (ideas about catastrophic illness), leading to maladaptive behavior (withdrawal and avoidance). This S-I-C-B firing order (Sensation, Imagery, Cognition, Behavior) may require a different treatment strategy from that employed with say a C-I-S-B sequence, a I-C-B-S, or yet a different firing order. Clinical findings suggest that it is often best to apply treatment techniques in accordance with a client's specific chain reaction. A rapid way of determining someone's firing order is to have them in an altered state of consciousness—deeply relaxed with eyes closed—contemplating untoward events and then describing the sequence of feelings, thoughts, and other reactions.

3.3 Second-order BASIC I.D. Assessments

The initial Modality Profile (BASIC I.D. Chart) translates vague, general, or diffuse problems (e.g., depression, unhappiness, and anxiety) into specific, discrete, and interactive difficulties. Techniques—preferably those with empirical backing—are selected to counter the various problems. Nevertheless, treatment impasses arise, and when this occurs, a more detailed inquiry into associated behaviors, affective responses, sensory reactions, images, cognitions, interpersonal factors, and possible biological considerations may shed light on the situation. This recursive application of the BASIC I.D. to itself adds depth and detail to the macroscopic overview afforded by the initial Modality Profile. Thus, a second-order assessment with a client who was not responding to antidepressants and a

combination of cognitive-behavioral procedures revealed a central cognitive schema 'I am not entitled to be happy' that had eluded all other avenues of inquiry. Therapy was then aimed directly at addressing this maladaptive cognition.

4. Research Findings

Williams (1988) in a carefully controlled outcome study, compared multimodal assessment and treatment with less integrative approaches in helping children with learning disabilities. Clear data emerged in support of the multimodal procedures. Kwee (1984) conducted a treatment outcome study on 84 hospitalized patients suffering from obsessive-compulsive disorders or phobias, 90 percent of whom had received prior treatment without success, and 70 percent of whom had suffered from their disorders for more than 4 years. Implementing multimodal treatment regimes resulted in substantial recoveries and durable 9-month follow-ups. This has been confirmed and amplified by Kwee and Kwee-Taams (1994). Recently, the vast literature on treatment regimens, be they journal articles or entire books, has accentuated multimodal therapy—although the authors often label their work multidimensional, multimethod, or multifactorial.

Herman (e.g., 1992, 1994) has conducted a good deal of research into the reliability and validity of various multimodal assessment instruments and their applications to several areas and clinical conditions.

Nevertheless, MMT is so broad based, so flexible, and so personalistic that tightly controlled outcome research is virtually impossible. There is, at best, suggestive evidence, rather than hard data, to confirm the clinical impression that covering the BASIC I.D. enhances outcomes and follow-ups. It should be understood that MMT endeavors first and foremost to apply empirically supported methods whenever feasible (most of which are drawn from the field of cognitive-behavior therapy (CBT), because careful studies have found it to be one of the very few orientations with well-established empirically supported treatments).

5. Concluding Comment

The field of CBT—from which MMT draws the vast majority of its techniques and procedures—has grown increasingly broad-based over the past two decades. Many of the methods advocated by MMT in the 1970s that were viewed askance by behaviorally oriented theorists and practitioners are now mainstream methods within a cognitive-behavioral framework. These include various methods of cognitive restructuring, mental imagery, and sensory awareness techniques. The entire field of psychological treatment seems to have shifted away from the exclusive focus on intrapsychic phenomena that dominated the profession during the 1950s, toward a perspective that emphasizes breadth.

Perhaps MMT has played a role in facilitating what might become a twenty-first century direction in psychological treatment. The results of the Delphi poll (Norcross et al. 1992) portend systematic eclecticism's continued domination of the field. One may predict that treatments of choice for selected clinical disorders will become standard practice, and that psychological therapies will be matched increasingly to client variables beyond diagnostic labels. Indeed, diagnostic categories seldom have relevance for the specific, moment-to-moment decisions made by practitioners. It is probable that programatic goals will be designed on the basis of a guiding theoretical definition of what is clinically relevant and meaningful. In the most general terms, one can predict confidently that the twenty-first century will witness unimodal schools of psychological thought replaced by multimodal, multifaceted, and multidimensional perspectives.

See also: Behavior Therapy: Psychological Perspectives; Cognitive Therapy; Eclectic Psychotherapy: Psychological Perspectives; Person-centered Psychotherapy; Psychological Therapies: Emotional Processing; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy Integration; Therapist–Patient Relationship

Bibliography

Bandura A 1977 *Social Learning Theory*. Prentice Hall, Englewood Cliffs, NJ

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall, Englewood Cliffs, NJ

Fay A, Lazarus A A 1993 On necessity and sufficiency in psychotherapy. *Psychotherapy in Private Practice* **12**: 33–9

Herman S M 1992 A demonstration of the validity of the multimodal structural profile inventory through a correlation with the vocational preference inventory. *Psychotherapy in Private Practice* **11**: 71–80

Herman S M 1994 The diagnostic utility of the multimodal structural profile inventory. *Psychotherapy in Private Practice* **13**: 55–62

Kwee M G T 1984 *Klinische Multimodale Gegragtstherapie*. Swets & Zeitlinger, Lisse, Holland

Kwee M G T, Kwee-Taams M K 1994 *Klinische Gedragstherapie in Nederland & Vlaanderen*. Eubron, Delft, Holland

Lazarus A A 1967 In support of technical eclecticism. *Psychological Reports* **21**: 415–6

Lazarus A A 1992 Multimodal therapy: Technical eclecticism with minimal integration. In: Norcross J C, Goldfried M R (eds.) *Handbook of Psychotherapy Integration*. Basic Books, New York, pp. 231–63

Lazarus A A 1995 Different types of eclecticism and integration: Let's be aware of the dangers. *Journal of Psychotherapy Integration* **5**: 27–39

Lazarus A A 1997 *Brief but Comprehensive Psychotherapy: The Multimodal Way*. Springer, New York
Lazarus A A 2000 Multimodal therapy. In: Corsini R J, Wedding D (eds.) *Current Psychotherapies*. Peacock, Itasca, IL
Lazarus A A, Beutler L E 1993 On technical eclecticism. *Journal of Counseling and Development* **71**: 381–5
Lazarus A A, Beutler L E, Norcross J C 1992 The future of technical eclecticism. *Psychotherapy* **29**: 11–20
Lazarus A A, Lazarus C N 1991 *Multimodal Life History Inventory*. Research Press, Champaign, IL
Norcross J C, Alford B A, DeMichele, J T 1992 The future of psychotherapy: Delphi data and concluding observations. *Psychotherapy* **29**: 150–8
Paul G L 1967 Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology* **31**: 109–18
Rotter J B 1954 *Social Learning and Clinical Psychology*. Prentice-Hall, Englewood Cliff, NJ
Williams T A 1988 A multimodal approach to assessment and intervention with children with learning disabilities. Ph.D. thesis, University of Glasgow

A. A. Lazarus

Multinational Corporations

The multinational corporation is a business organization whose activities are located in more than two countries and is the organizational form that defines foreign direct investment. This form consists of a country location where the firm is incorporated and of the establishment of branches or subsidiaries in foreign countries. Multinational companies can, obviously, vary in the extent of their multinational activities in terms of the number of countries in which they operate. A large multinational corporation can operate in 100 countries, with hundreds of thousands of employees located outside its home country.

The economic definition emphasizes the ability of owners and their managerial agents in one country to control the operations in foreign countries. There is a frequent confusion that equates the ability to control with the flow of capital across national borders. Since Hymer's thesis (1976), it is recognized widely that capital flow is not the distinguishing characteristic of a multinational corporation (see *International Business*). Capital can flow from one country to another in expectation of higher rates of return. However, this flow may be invested in the form of bonds, or in equity amounts too insignificant to grant control to foreign owners. In this case, this type of investment is treated as a 'portfolio' investment. The central aspect of 'direct investment' is the ownership claim by a party located in one country on the operations of a foreign firm or subsidiary in another. The multinational corporation is, thus, the product of foreign direct investment that is defined as the effective control of operations in a country by foreign owners.

The economic definition, however, does not capture the importance of the multinational corporation as the organizational mechanism by which different social and economic systems confront each other. The multinational corporation, because usually it develops in the cultural and social context of one nation, exports its organizational baggage from one institutional setting to another. In this regard, it plays a powerful role as a mechanism by which to transfer organizational knowledge across borders. However, while being foreign implies that it might serve the valuable role of importing new practices, its foreign status also implies that its practices are likely to conflict with existing institutions and cultural norms. Moreover, since multinational corporations are often large, they pose unusual challenges to national and regional governments who seek to maintain political autonomy and yet are often anxious to seek the investment, technology, and managerial skills of foreign firms.

There are, thus, economic and sociological definitions of the multinational corporation that differ, and yet complement, each other. In the economic definition, the multinational corporation is the control of foreign activities through the auspices of the firm. In the sociological definition, the multinational corporation is the mechanism by which organizational practices are transferred and replicated from one country to another.

1. *History*

The multinational corporation is defined in some sense arbitrarily by where frontiers are drawn. In ancient Greece, these frontiers were the borders among city–states. In imperial Rome, the new administrative units of an expanding empire and its external boundaries defined the borders. Political borders correspond only approximately to the distribution of cultural and ethnic dispositions. The history of the multinational corporation is tied closely to the origins of trade in and between cultural communities, and these communities remain important in many sectors in the modern economy.

Early trade was often characterized by the difficulty of transacting across borders. Trading has always instigated from the unequal and varied distribution of resources across geographies. The trading of salt was an important factor in most continents and still can be found in rather ancient forms in east Africa. Towns, such as Salzburg, owed their origins to their fortuitous access to salt mines.

This unequal distribution incited traders to travel long distances and undergo unusual risks for the hope of gain. The dilemma is quite apparent. How can markets for trade develop between distant cultures? Indeed, there is good evidence that early trading was quite precarious. Brown (1969) notes that Hermes was not only the god of messengers but also of theft and

trade. He was often worshipped at borders, marking the place of trade. Herodotus describes the practice of the 'silent trade' of goods at boundary markings in which one party would deposit their goods at the border, the second party would later deposit a corresponding offer. The original party might accept this bid, take the goods, and leave their initial offer. Clearly, only a strong demand for unusual foreign goods could sustain such a treacherous institution of transaction.

Not surprisingly, Curtin (1984) doubts the accuracy of this account. Instead, he stresses the importance of ethnic groups, such as the Gujerati or overseas Chinese, or the force of imperialism and military power as ways in which 'cross-cultural trading' was permitted to develop. But such solutions remained fragile, with foreign enclaves and ethnic groups subject to local hatred and military rule often confronting competition from other powers.

Yet, trade persisted through the centuries. Indeed, the demand for foreign goods was so strong that incredible profits could be realized by international trade. Since transportation was poor, and technology to hold inventories rudimentary, an agent in the right spot at the right time could earn fortunes. And in this laid the great problem. For in times of great demand, prices could rise astronomically, as Braudel (1973) documents in his history of material life. With such distances between the agent and the principal in the exporting country, an agent could easily, and often did, disappear with the profits.

The early solutions were many and provide insight into one of the great properties of the multinational corporation: its ability to organize transactions within its own organizational boundaries. Great fairs were one solution, where principals met their customers in one place. But fairs were intermittent and often distant from final markets. The further development of Roman commercial law during the medieval and Renaissance periods strengthened the liability of the principal and also legal recourse in principal–agent conflict. However, the application of law between two parties in different legal jurisdictions has always been a difficult form of dispute resolution. In his study of the Jewish Magrebi traders of the twelfth century, Greif (1989) places considerable emphasis upon the value of membership in ethnic 'trading communities.' Since ethnic enclaves in foreign sites could monitor the activities of agents, principals could be informed of malfeasance. Agents could then be excluded from future contracts within the community, thus deterring dishonest behavior. Partnerships were also another way in which agents were also bonded to the firm.

However, another solution was the company, a word whose roots come from the Latin meaning 'with bread' which implies the familial origins of this business organization. The company permitted in time the investing of capital in its ownership by outsiders in promise of future dividends. The capital might be raised for a single voyage or, eventually, could be maintained in the company in the form of stock ownership. With the evolution of limited liability law in the nineteenth century, and the diminished role of the state in restricting the growth of the joint stock company, this organizational form expanded rapidly in many European countries as well as in the United States. Along with the growth of equity markets, the financial resources to invest overseas became more available due to advances in the banking system and also in bond markets. Indeed, the great capital needs of the railway industry created an international market for the sale and purchase of railroad bonds. However, international loans and bonds were risky and it was not uncommon for sovereign governments to default, including state governments within the United States.

The increasing wealth of western countries, along with constraints on the speed by which national industries could absorb new loans, encouraged massive foreign investments by the end of the nineteenth century and early twentieth century. This was a period of globalization in terms of the percentages of capital outflows to total capital accumulation. Great Britain is estimated to have exported some 25 percent of its capital prior to World War I and French capital exports were often greater. This capital went to countries hungry to finance their industrial expansion, including the newly industrializing countries of Russia and especially America.

The British were unusual in directing a considerable amount of their outward capital flows in the form of direct investments. A great deal of this investment went into their colonies. Prior to the mid-1800s, companies operating overseas engaged largely in wholesale operations; they did not run factories, operate mines, or own agriculture. The British direct investments were, thus, different than before, as British came to own and run local operations. Moreover, the British invested in South and North America.

The 'free-standing' company, as labeled by Wilkins (1988), was a peculiar form of investment that was quite prominent in many British colonies. This firm raised capital in the domestic financial market where its administrative office was located, but it operated no domestic activities. All operations were overseas. This form of company represents the advantage of western countries in being able to raise capital from efficient financial markets. However, it posed the same problem as facing the medieval agent relationship, namely, how could foreign investors be confident that they would recover profitably their investments if the assets were located overseas, often in countries with very poor legal infrastructures? Wilkins found that oversight occurred through the executive boards of the companies, where prominent people located in the home market faced the loss of reputation if the company should prove to be dishonest.

The United States offers an unusual case. A rapidly growing country, it imported more capital than it

exported up to World War I. However, whereas it imported largely portfolio investments, its outward flows were dominated by foreign direct investments. In other words, American companies showed an early penchant for expanding overseas (Wilkins 1970). The Singer Company built within two decades of its founding a large factory employing thousands of workers in Scotland. Oil companies, Kodak, Westinghouse, Ford, and mining and agricultural companies all invested overseas. Companies in oil, mining, and agriculture often invested in poorer markets where there were resources to be found. These early investors were often involved inextricably in the politics of the foreign governments, and the American military itself intervened aggressively numerous times in the Caribbean, Mexico and Central America, and South America. As in the British case, the history of American direct investment occurred in the context of an expanding military presence of the home government. Moreover, since many of these countries were poor, the multinational corporations responded to the demands of the host nation, especially in the form of concessionary contracts, to provide public services, e.g., hospitals, roads, and power (Robinson 1964). This complicated legacy of the early history of the multinational corporation created hostility on the part of the local population that persisted throughout most of the twentieth century.

It is important to underscore that the multinational corporation usually evolved in the context of specific national institutions. As many others have pointed out, the multinational corporation is a growing firm whose organizational borders have spilled across borders. Moreover, since this large firm is usually tied to a larger domestic network of suppliers and customers, its expansion overseas is accompanied by the co-investments of these other members. This is a pattern seen in American investments in the United Kingdom in the 1950s and repeated by Japanese multinational corporations investing in the United Kingdom in the 1980s and 1990s (Dunning 1993).

Chandler (1990) noted that these multinational corporations reflected the national characteristics of management. In comparing the cases of the largest firms in the United Kingdom, Germany, and the United States, Chandler found that differences in managerial capabilities, reflecting national institutions, explain their success and failure patterns. He particularly criticized the managerial capabilities of British firms, a point not shared by some British historians (Hannah 1999). But more importantly, Chandler's thesis assumed that size itself constituted the realization of scale and scope economies instead of the outcome of success and growth. This observation is especially important for understanding the lack of large multinational corporations in Italy or in Taiwan, both of which have very successful small firm economies but do not have multinational corporations comparable to other countries of similar levels of economic wealth. Yet, both countries are relatively wealthy and successful, and their many small companies have achieved high rates of exporting. Even in the case of the United States, the evidence implies that American firms, large and small, came to Europe, riding on the back of the national organizing principles of standardization in work methods (Kogut 1992). Chandler's larger point of the effect of national systems on firm capabilities is largely accepted; his belief that large firms reflect better management because they achieve scale is disputed far more.

2. *Organization of Multinational Corporations*

The national origins of the multinational corporation influence their subsequent organizational evolution. In the United States, the great post-World War II expansion of multinational corporations coincided with the diffusion of their adoption of the multidivisional structure. Indeed, the refinement of 'organizational technologies' permitted American firms to manage their rapidly growing operations on a worldwide basis.

The young American multinational corporation began with little knowledge and experience of the foreign market. Initial foreign sales are exports executed within the existing organizational structure. These structures could be functional (e.g., organized by production and sales) or divisional (e.g., organized by geographic area or product division). As foreign sales increased, an international division was created to make the sale and to provide customer support. In consequence, the domestic organizational unit would sell to the international division its products at an internal 'transfer price.' As this price was set by the domestic market, it was unlikely to reflect the competitive conditions in the foreign country. Moreover, the products were often not designed to the tastes of foreign demand. Because of this internal conflict, the American multinational corporation would replace the international division by transferring responsibilities to area divisions (e.g., Europe, Asia, South America) or to product divisions (Stopford and Wells 1972). While these structures diminished the internal conflict, they ran the hazard that the divisional managers across areas did not cooperate or that managers across product divisions tended to fall back upon focusing on the home market.

European firms grew up in a different environment. Europe was a highly unstable continent during the twentieth century, with significant political and economic conflicts. Borders sometimes changed, and they also marked the necessity to pay commercial duties. The gradual creation of the European Union in the last third of the twentieth century eliminated tariffs and lessened national differences. This history marked their firms. The organizational technologies in most of

these countries were quite different to those found in the United States. Holding company structures were common, banks or other financial institutions played powerful roles in ownership and cross holdings, and firms were tied together in complex financial webs. These structures, often called 'mother–daughter' organizations, reflect the linguistic terms used in northern Europe to describe headquarters and subsidiary. These mother–daughter structures consisted often of a headquarters that held a portfolio consisting of ownership control over dozens of companies, many of them in the same sector but located in different countries. Reporting relationships among these companies and headquarters was less formal than found in American companies, and control was often exercised through the movement of managers who were lifetime employees. These structures were often seen as inefficient, especially compared to the size and organization of American multinationals that were investing rapidly in Europe at this time. European critics called for the restructuring of European firms and industry along American lines.

The structure of Japanese companies reflected also their historical origins. In Japan, the trading company performed an extraordinary role in the exports during the first two-thirds of the twentieth century. These companies started primarily as wholesalers, but over time developed their own industrial companies located in Japan and eventually overseas. Though the forced imposition of American antitrust law after World War II disrupted these patterns, trading companies and other affiliated company networks, called *keiretsu*, reconstituted the earlier pattern of constellations of domestic companies competing in home and, through their trading companies, in foreign markets. As in the American and European cases, Japanese firms varied in their internal structures, with functional and even factory-based organization being the predominant structures. Still, a dominant trend emerged over time as the role of trading companies declined, especially for the exportation of industrial and high technology goods. This trend consisted of the appendage of an international division to the existing structure (Suzuki 1991).

These different national structures confronted an increasingly more globally integrated environment in the course of the twentieth century. This environment posed the classic organizational problem of balancing integration and differentiation. For the multinational corporation, this problem posed itself as meeting the demands to achieve global scale against the needs of national markets and governments. Given different national principles of organizing the activities of multinational corporations, there was not a rapid convergence to a global structure. The growth of multinational corporations, especially in Europe, was stymied by these traditions. Continuing national differences and the failure to develop harmonized European corporate law deterred the emergence of pan-European multinational corporations. Many of the most important cross-European mergers in the 1970s failed within a decade. Instead, American multinational corporations, less bound by these separate national traditions, were able to create integrated European strategies.

Nevertheless, it was in Europe where new organizational structures developed to resolve the conflict between integration and differentiation. Multinational corporations from smaller countries offered the laboratory experiments that influenced the evolution of multinational corporations from larger nations. Because of the small size of many European countries, large firms in Sweden, Switzerland, and the Netherlands quickly became multinational corporations, and these corporations had far more assets and employees outside their home countries than within. Over the course of the last two decades of the twentieth century, organizational structures developed, labeled as transnational corporations, that distributed global responsibilities to national subsidiaries, created global teams and projects, and encouraged the transfer of 'best practices' across borders.

3. *Diffusion of Organizational Knowledge*

In its evolution the multinational corporation is not without serious contradictions. Evolving from its national context, the multinational corporation employs large numbers of employees of diverse nationalities and ethnicities. Westney (1993) notes that a subsidiary is, thus, caught between the institutional pressures to conform to the company norms and values, as well as to the cultural and social influences of its local national environment. At the heart of the evolution of the multination corporation, thus, lies the tension between national institutions and the fragile emergence of a global culture.

The international evolution of the organizational structures of American multinationals mirrored, as we noted above, the broader diffusion of organizational technologies in the home market. The initial investments by a firm took place, often, on the basis of opportunity and the extension to familiar countries. Much like the ethnic trading communities dating back to the earliest times, the inexperienced multinational corporation preferred countries that are culturally similar to what their managers know at home (Johanson and Vahlne 1978). In these countries, they often established foreign enclaves where their expatriate managers could live in the simulated familiarity of their home environments.

The relevance of the second definition of foreign direct investment as the transfer of organizational knowledge is critical to understanding the powerful conflicts posed by the multinational corporation. The multinational corporation, competing often on superior technologies and managerial capabilities, serves

as a conduit of knowledge across borders. Some kinds of technologies can be purchased on markets and the flow of licensing payments attached to the sale of the right to use a technology constitutes a nontrivial flow in international balance of payments. Technology can also flow by the movement of people. Just as English craftsmen were imported by continental Europe in the early industrial period, there exists an international market for managers and skilled workers.

However, some knowledge is embedded in organizations and can only be transferred as an organization across borders. The issue is more than whether knowledge is tacit or explicit, for this pertains to the knowledge held by individuals too. The central feature of the diffusion of knowledge by the multinational is that it transfers the knowledge of how to organize and how to coordinate people and across divisions. These organizing principles then provide the capabilities for the firm to achieve quality, speed products to the market, or lower costs (Kogut 1992). For this the multinational corporation is required, which through its own activities or in joint ventures with host companies, transfers the organizational knowledge.

The global market for knowledge extends also to consultants. The large American consulting practices often have their origins in their acquisition of American organizing principles that they then transferred around the world. The international diffusion of the divisional structure, for example, reflects the knowledge acquired in American consulting practices that was then sold in abroad. Channon (1973), for example, observes that half of the firms he observed as adopting the divisional structure relied upon the consulting services of the same American firm. Similarly, British and then American banks spread throughout the world on servicing the needs of their expanding home clients, and then on transferring their practices into these countries (Jones 1993).

It is important to realize that once this transition period passed, the multinational corporation and multinational service firms, such as those engaged in consulting, banking, and advertising, had permanently and historically changed the global economy. No longer was the flow from the US to Europe, or from one ethnic community to another. The network of connections developed by the multinational corporation permitted the flow of ideas and practices among and between countries. Thus, whereas it took half a century for American practices to diffuse to Europe, the introduction of quality circles from Japan happened within a decade or two of their original innovation. In Europe, Japanese innovations were often diffused by American firms, including consulting companies.

However, this image of the free flow of knowledge needs to be strongly conditioned on the continuing importance of national institutions. Nations consist of defined cultures and economic and social institutions, such as unions, financial systems, and religious values. These institutions interact and their complex interactions causally influence behaviors. For example, the German system of centralized bargaining, enterprise clubs, strong banks, and social welfare has, until recent times, composed a national configuration of institutions that influences the capabilities of resident firms to be able to manufacture high quality goods for export markets (Streeck 1995, Soskice 1990). The introduction of Japanese and American teams confronts in Germany the presence of existing work councils. These councils are fairly rigid features of the German environment. Since Japanese methods may not be effective without teams as complementary factors in organization, this institutional refusal can effectively deter the diffusion of these methods. In short, national organizing principles are embedded in the wider social institutions. Inconsistency between these institutions and principles, then, is an empirical question of the bargaining strength of vested powers.

4. *Economics and Politics of Multinational Corporations*

Since the multinational corporation is definitionally equivalent to foreign direct investment, theories of foreign direct investment must account for why one country invests in another and why this investment is carried out within organizational boundaries of a firm (see Buckley and Casson 1976, see *Foreign Investment: Direct*). In distinguishing between portfolio and direct investment, Hymer noted that firms operate at a disadvantage in foreign markets and hence they must have an offsetting competitive advantage to compete overseas. These advantages for overseas investments are the same ones that allow a firm to compete and grow in the home market. These observations have important implications. The first is that direct investment is the growth of the firm across borders and hence the firm expands internationally on what it has learned at home. This observation is the basis for the evolutionary theory of the firm. The second observation that Hymer made is that firms that expand overseas, because they have competitive resources, are also likely to be large and to belong to oligopolistic industries.

In these observations, we can understand the ambivalence expressed in popular and policy debates regarding the multinational corporation. Competition among multinational corporations often is the extension of their home domestic and oligopolistic rivalry that spills across national borders. In many global industries, the same company names dominate each country's list of the largest firms inside their national frontiers. No matter if it is Poland or France, Singapore or Mexico, the same multinational corporations will be found in the local oligopolistic industries (e.g., consumer goods or automobiles).

Because they are large even in their home markets, investments by multinational corporations can have a large impact on a host country (Caves 1974).

As a consequence, the multinational corporation has often been the subject of debates concerning national sovereignty and welfare. In recent decades, acquisitions have generally been the primary way by which multinationals invest in wealthy foreign countries, where the vast proportion of direct investment is concentrated. Given the size of a multinational corporation and occasional national importance of the targeted acquisition, even wealthy countries frequently evidence discomfort, if not outright public hostility, to multinational investments. Moreover, multinational corporations are sometimes the vehicles for foreign policies of their home or host country. The decision, for example, of the US to embargo technology and investment flows to Cuba, the former Soviet Union, Iran, and other countries periodically has caused conflict with other countries.

Multinational corporations are especially problematic in developing countries. By definition, developing countries are relatively poor, thus both in need of capital and yet concerned over their loss of independence. As discussed above, the history of multinational corporations in developing countries is marked by its origins in policies of imperialism and colonialism. Especially in Latin America, where a school of thought labeled *Dependencia* has been influential, the concern over dependence on the United States resulted in efforts to curb the power of the multinational corporations by restricting the amount of equity ownership a foreign firm could hold in a domestic company or by prohibiting investment in certain sectors. Mexico's constitution forbids foreign investment in the oil industry; Brazil pursued for a long time a policy to restrict foreign participation in the electronics industry.

The other side of the coin is that multinational corporations bring investment and technology to the foreign country. Vernon (1966) hypothesized that innovations start in wealthy countries. As the market is saturated and as oligopolistic rivalry increases, multinational corporations are pushed out from their home markets to expand abroad in new markets and to locate less expensive places. Thus, Vernon seized both sides of the debate, recognizing the value of the transfer of technology but also emphasizing the oligopolistic nature of multinational investment.

It is, in fact, difficult to draw simple conclusions regarding the relationship of foreign investment and national growth. Countries such as Singapore, Malaysia, and Thailand have encouraged foreign direct investment actively. The growth in China's coastal sector is indisputably linked to the massive investments by multinational corporations. However, historically Japan and Korea have pursued more cautious policies regarding investments by multinational corporations. In these countries, the state has often negotiated the terms for entry by multinational corporations, sometimes requiring licensing to domestic competitors as a price. The efficacy of such policies for these countries is much disputed. However, for many other countries, the intervention of the government in demanding licenses unquestionably leads to internal corruption and to insufficient domestic competition.

There are many channels by which a country can absorb foreign technology and managerial techniques. Most of the evidence shows, however, that prohibitions on the in-flows of direct investment can be very costly for many countries. With their domestic industries still to be developed, a developing country requires substantial investment. Some countries, primarily in Asia, have been able to achieve very high savings rates to finance their industries without direct investment. Moreover, high savings rates, plus political stability, create growth, and growth attracts foreign portfolio capital. A poor country that prohibits foreign direct investment but does not have high rates of saving is entirely dependent upon portfolio capital. The history of debt and currency crises in the 1990s convinced many poor countries that foreign direct investment was a preferable means of attracting capital, because it could not be easily pulled out of a country on short-notice in response to a financial crisis.

However, multinational corporations also respond to the volatility in the global market. This volatility derives from changes in exchange rates, politics, and productivity. Once having achieved sufficient experience and having established subsidiaries around the world, the multinational corporation might choose to close a plant in one location and open plants in new locations. Of course, such actions might provoke a response by labor, but historically, labor has been organized by national, not by international, organizations (Martinelli 1975). Yet, there is also the possibility that locations lose some kinds of plants but gain more sophisticated investments. Cantwell (1999) proposed that some regions and countries pull multinational investments. Yet, it has long been noticed that foreign direct investment among developed countries flows to high cost locations. Regions such as Silicon Valley, Baden-Wuertemberg, and Singapore attract multinational investments not because wages are low, but because productivity levels are high and workers are well trained. In many cases, developing countries have given rise to their own multinational corporations acting in the region and sometimes globally (Lall 1983). In this sense, the multinational corporation acts as a training center in the developmental strategies of emerging economies.

5. *Globalization*

The peculiar conflict in the world economy is the growing trend toward economic integration without a

concomitant growth in global political and social institutions to regulate and arbitrate this trend. Multinational corporations are rapidly adopting advanced information technologies to increase the efficiency and capabilities of their operations. Valuable information, such as software or financial services, is transmitted digitally. Governments are often unable to tax the value of these services or control their content. Moreover, information technologies, often supported by private telecommunication networks, permit, for example, a unit in Germany to control the manufacturing operations located in Brazil.

Digital technologies permit a high degree of integration and, in the short run, aggravate the gap between countries that are rich and those that are poor and do not have the infrastructural capabilities. However, these technologies are beginning to have profound effects on some regional economies. In parts of China, India, Israel, and elsewhere, advanced satellite transmission transmits digitally encoded work between their sites and other organizational units of multinational corporations located elsewhere. Whereas before the Indian engineer from Bangalore might have tried to migrate to the US to bring his human capital to a more attractive labor market, the increasing wages and job prospects in India are encouraging many to stay at home and participate digitally in the world economy.

These trends have unclear effects on multinational corporations. They permit more easily the development of projects that are in continual development, as work passes from one unit to the next as the day advances. They also allow more easily the coupling of less expensive labor in one country with more expensive skilled labor in another.

At the same time, the origins of the multinational corporation laid in its ability to organize labor on a worldwide basis on principles other than ethnic and cultural identities. This organization has never been without cost and risk. The growth of a world digital economy permits alternative ways by which labor can cooperate and be coordinated on a world basis. The intriguing question at this point in history is whether the multinational corporation, though still a vital presence in the world economy, nevertheless will recede relatively in importance as information technologies reduces the meaning of geographic distance.

See also: Development and the State; Diaspora; Globalization: Geographical Aspects; Globalization: Legal Aspects; Globalization: Political Aspects; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; International Business; International Law and Treaties; International Marketing; International Organization; International Trade: Economic Integration

Bibliography

Braudel F 1973 *Capitalism and Material Life, 1400–1800* (trans. Kochan M). Harper and Row, New York

Brown N O 1969 *Hermes the Thief: The Evolution of a Myth*, 2nd edn. University of Wisconsin Press, Madison, WI

Buckley P J, Casson M 1976 *The Future of the Multinational Enterprise*. MacMillan, London

Cantwell J 1989 *Technological Innovations and Multinational Corporations*. Blackwell, London

Caves R E 1974 Causes of direct investment: Foreign firms. Shares in Canadian and United Kingdom manufacturing industries. *Review of Economic Statistics* **56**: 79–93

Chandler A D 1990 *Scale and Scope: The Dynamics of Industrial Capitalism*. Belknap Press, Cambridge, MA

Channon D F 1973 *The Strategy and Structure of British Enterprises*. Harvard University, Boston

Curtin P D 1984 *Cross-cultural Trade in World History*. Princeton University Press, Princeton, NJ

Dunning J 1993 The governance of Japanese and US manufacturing affiliates in the UK: Some country-specific differences. In: Kogut B (ed.) *Country Competitiveness: Technology and the Organizing of Work*. Oxford University Press, Oxford, UK

Greif A 1989 Reputation and coalitions in medieval trade: Evidence on the Maghribi traders. *Journal of Economic History* **39**: 857–82

Hannah L 1999 Marshal's 'trees' and the global 'forest': Were 'Giant Redwoods' different? In: Lamereaux N R, Raff D M G, Temin P (eds.) *Learning by Doing in Markets, Firms, and Countries*. University of Chicago Press, Chicago pp. 253–86

Hymer S 1976 The international operations of national firms: A study of direct investment. Ph.D. thesis, Massachusetts Institute of Technology

Johanson J, Vahlne J-E 1978 A model for the decision making process affecting pattern and pace of the internationalization of the firm. In: Ghertman M, Leontiades J (eds.) *European Research in International Business*. North Holland, Amsterdam pp. 9–27

Jones G 1993 *British Multinational Banking, 1830–1990*. Oxford University Press, New York

Kogut B 1992 National organizing principles of work and the erstwhile dominance of the American multinational corporation. *Industrial and Corporate Change* **1**: 285–325

Lall S 1983 *The New Multinationals: The Spread of Third World Enterprise*. Wiley, Chichester, UK

Martinelli A 1975 Multinational corporations, national economic policies and labor unions. In: Lindberg L, Alford R, Crouch C, Offe C (eds.) *Stress and Contradiction in Modern Capitalism*. Heath, Lexington, MA pp. 425–43

Robinson R D 1964 *International Business Policy*. Holt, Rinehart, and Winston, New York

Soskice D 1990 Wage determination: The changing role of institutions in advanced industrialized countries. *Oxford Review of Economic Policy* **6**: 36–61

Stopford J M, Wells L T 1972 *Managing the Multinational: Organization of the Firm and Organization of the Subsidiaries*. Basic Books, New York

Streeck W 1995 German capitalism: Does it exist? Can it survive? In: Crouch C, Streeck W (eds.) *Modern Capitalism or Modern Capitalisms?* Pinter, London

Suzuki Y 1991 *Japanese Management Structures, 1920–1980*. St. Martin's Press, New York

Vernon R 1966 International investment and international trade in the product life cycle. *Quarterly Journal of Economics* **80**: 190–207

Westney D E 1993 Institutional theory and the multinational corporation. In: Westney D E, Ghoshal S (eds.) *Organization Theory and the Multinational Corporation*. St. Martin's Press, New York

Wilkins M 1970 *The Emergence of Multinational Enterprise: American Business Abroad from the Colonial Era to 1914*. Harvard University Press, Cambridge, MA

Wilkins M 1988 The free-standing company, 1870–1914: An important type of British foreign direct investment. *Economic History Review* **39**: 259–82

B. Kogut

Multiple Imputation

Missing data are ubiquitous in social science research. For example, they occur in a survey on people's attitudes when some people do not respond to all of the survey questions, or in a longitudinal survey when some people drop out. A common technique for handling missing data is to impute, that is, fill in, a value for each missing datum. This results in a completed data set, so that standard methods that have been developed for analyzing complete data can be applied immediately (see *Nonsampling Errors* and *Statistical Data, Missing*).

Imputation has other advantages in the context of the production of a data set for general use, such as a public-use file. For example, the data producer can use specialized knowledge about the reasons for missing data, including confidential information that cannot be released to the public, to create the imputations; and imputation by the data producer fixes the missing data problem in the same way for all users, so that consistency of analyses across users is ensured. When the missing-data problem is left to the user, the knowledge of the data producer can fail to be incorporated, analyses are not typically consistent across users, and all users expend resources addressing the missing-data problem.

Although imputing just one value for each missing datum satisfies critical data-processing objectives and can incorporate knowledge from the data producer, it fails to achieve statistical validity for the resulting inferences based on the completed data. Specifically, for validity, the resulting estimates based on the data completed by imputation should be approximately unbiased for their population estimands, confidence intervals should attain at least their nominal coverages, and tests of null hypotheses should not reject true null hypotheses more frequently than their nominal levels. But a single imputed value cannot reflect any of the uncertainty about the true underlying value, and so analyses that treat imputed values just like observed values systematically underestimate uncertainty. Thus, using standard complete-data techniques will result in standard error estimates that are too small, confidence intervals that undercover, and *P*-values that are too significant; even if the modeling for imputation is carried out carefully. For example, large-sample results by Rubin and Schenker (1986) show that for simple situations with 30 percent of the data missing, single imputation under the correct model followed by the standard complete-data analysis results in nominal 90 percent confidence intervals having actual coverages below 80 percent. The inaccuracy of nominal levels is even more extreme in multiparameter problems (Rubin 1987, Chap. 4), where nominal 5 percent tests can easily have rejection rates of 50 percent or more when the null hypothesis is true.

Multiple imputation (Rubin 1987) retains the advantages of single imputation while allowing the data analyst to obtain valid assessments of uncertainty. The key idea is to impute two or more times for the missing data using independent draws of the missing values from a distribution that is appropriate under the posited assumptions about the data and the mechanism that creates missing data, resulting in two or more completed data sets, each of which is analyzed using the same standard complete-data method. The analyses are then combined in a simple generic way that reflects the extra uncertainty due to having imputed rather than actual data. Multiple imputations can also be created under several different models to display sensitivity to the choice of missing-data model.

1. *Theoretical Motivation for Multiple Imputation*

The theoretical motivation for multiple imputation is Bayesian (see *Bayesian Statistics*), although the procedure has excellent properties from a frequentist perspective. Formally, let Q be the population quantity of interest, and suppose the data can be partitioned into the observed values $\mathbf{X}_{\mathbf{obs}}$ and the missing values $\mathbf{X}_{\mathbf{mis}}$. If $\mathbf{X}_{\mathbf{mis}}$ had been observed, inferences for Q would have been based on the complete-data posterior density $p(Q \mid \mathbf{X}_{\mathbf{obs}}, \mathbf{X}_{\mathbf{mis}})$. Because $\mathbf{X}_{\mathbf{mis}}$ is not observed, inferences are based on the actual posterior density $p(Q \mid \mathbf{X}_{\mathbf{obs}})$, which can be expressed as

$$p(Q \mid \mathbf{X}_{\mathbf{obs}}) = \int p(Q \mid \mathbf{X}_{\mathbf{obs}}, \mathbf{X}_{\mathbf{mis}})\, p(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}})\, \mathrm{d}\, \mathbf{X}_{\mathbf{mis}} \qquad (1)$$

Equation (1) shows that the actual posterior density of Q can be obtained by averaging the complete-data posterior density over the posterior predictive distribution of $\mathbf{X}_{\mathbf{mis}}$. In principle, the set of multiple imputations under one model are repeated independent draws from $p(\mathbf{X}_{\mathbf{mis}} \mid \mathbf{X}_{\mathbf{obs}})$. Thus, multiple imputation allows the data analyst to approximate Eqn. (1) by separately analyzing each data set completed by imputation and then combining the results of the separate analyses.

2. Analyzing a Multiply-imputed Data Set

The exact computation of the posterior distribution (Eqn. (1)) by simulation would require that an infinite number of values of $\mathbf{X}_{\mathbf{mis}}$ be drawn from $p(\mathbf{X}_{\mathbf{mis}} | \mathbf{X}_{\mathbf{obs}})$. Fortunately, simple approximations work very well in most practical situations with only three to five imputations.

2.1 Inferences for Scalar Q

Suppose, as is typically true in the social and behavioral sciences, that if the data were complete, inferences for Q would be based on a point estimate $\hat{Q}$, an associated variance/covariance estimate U, and a normal reference distribution. When data are missing and there are M sets of imputations for the missing data, the result is M sets of complete-data statistics, say, $\hat{Q}_m$ and U_m, $m = 1,\ldots, M$.

The following procedure is useful for drawing inferences about Q from the multiply imputed data (Rubin 1987, Rubin and Schenker 1986). The point estimate of Q is the average of the M completed-data estimates,

$$\bar{Q} = \sum_{m=1}^{M} \hat{Q}_m / M$$

and the associated variance estimate is

$$T = \bar{U} + (1 + M^{-1})B$$

where $\bar{U} = \sum_{m=1}^{M} U_m / M$ is the average within-imputation variance, and $B = \sum_{m=1}^{M} (\hat{Q}_m - \bar{Q})^2 / (M-1)$ is the between-imputation variance (cf. the usual analysis of variance breakdown in *Linear Hypothesis*). The approximate reference distribution for interval estimates and significance tests is a t distribution with degrees of freedom

$$\nu = (\nu_m^{-1} + \hat{\nu}_{\text{obs}}^{-1})^{-1}$$

where $\nu_m = (M-1)\hat{\gamma}^{-2}$, $\hat{\nu}_{\text{obs}} = \nu_{\text{com}}(1-\hat{\gamma})(\nu_{\text{com}}+1)/(\nu_{\text{com}}+3)$, $\hat{\gamma} = (1+M^{-1})B/T$ is the approximate Bayesian fraction of missing information, and ν_{com} is the complete-data degrees of freedom (see Barnard and Rubin 1999 for the derivation).

2.2 Significance Tests for Multicomponent Estimands

Multivariate analogues of the expressions given in the previous section for scalar Q are summarized by Rubin (1987, Sect. 3.4). Methods also exist for likelihood-ratio testing when the available information consists of point estimates and evaluations of the complete-data log-likelihood-ratio statistic as a function of these estimates and the completed data.

With large data sets and large models, such as in the common situation of a multi-way contingency table, the complete-data analysis might produce only a test statistic and no estimates (cf. *Multivariate Analysis: Discrete Variables (Overview)*). Rubin (1987, Sect. 3.5) provided initial methods for situations with such limited information, and improved methods exist that require only the M complete-data χ^2 statistics (or equivalently the M complete-data P-values). These methods, however, are less accurate than methods that use the completed-data estimates.

3. Creating Multiple Imputations

Ideally, multiple imputations are M independent random draws from the posterior predictive distribution of $\mathbf{X}_{\mathbf{mis}}$ under appropriate Bayesian modeling assumptions. Such imputations are called repeated imputations by Rubin (1987, Chap. 3). In practice, approximations often are used and work well.

3.1 Modeling Issues

An initial modeling issue in creating imputations is that the predictive distribution for the missing values should be conditional on all observed values. This consideration is particularly important in the context of a public-use database. Leaving a variable out of the imputation model is equivalent to assuming that the variable is not associated with the variables being imputed, conditionally given all other variables in the model; imputing under this assumption can result in biases in subsequent analyses, with estimated parameters representing conditional association pulled toward zero. Because it is not known which analyses will be carried out by subsequent users of a public-use database, ideally it is best not to leave any variables out of the imputation model. Thus, it is desirable to condition imputations on as many variables as possible and to use subject-matter knowledge to help select those variables that are likely to be used together in subsequent complete-data analyses, as by Clogg et al. (1991).

Imputation procedures can be based on explicit models or implicit models, or even combinations (see, e.g., Rubin 1987, Chap. 5, Rubin and Schenker 1991). An example of a procedure based on an explicit model is stochastic normal regression imputation, where imputations for missing values are created by adding normally distributed errors to predicted values obtained from a least-squares regression fit. A common procedure based on implicit models is hot-deck impu-

tation, which replaces the missing values for an incomplete case by the values from a matching complete case, where the matching is carried out with respect to variables that are observed for both the incomplete case and complete cases.

The model underlying an imputation procedure, whether explicit or implicit, can be based on the assumption that the reasons for missing data are either ignorable or nonignorable (see Rubin 1976, and *Statistical Data, Missing*). The distinction between an ignorable and a nonignorable model can be illustrated by a simple example with two variables, X and Y, where X is observed for all cases whereas Y is sometimes missing. Ignorable models assert that a case with Y missing is only randomly different from a complete case having the same value of X. Nonignorable models assert that there are systematic differences between an incomplete case and a complete case even if they have identical X-values. An important issue with nonignorable models is that, because the missing values cannot be observed, there is no direct evidence in the data to address the assumption of nonignorability. It can be important, therefore, to consider several alternative models and to explore sensitivity of resulting inferences to the choice of model. In current practice, almost all imputation models are ignorable; limited experience suggests that in major surveys with limited amounts of missing data and careful design, ignorable models are satisfactory for most analyses (see, e.g., Rubin et al. 1995).

3.2 Incorporating Proper Variability

Multiple imputation procedures that incorporate appropriate variability across the M sets of imputations within a model are called 'proper' by Rubin (1987, Chap. 4) where precise conditions for a method to be proper are also given; also see Meng (1994) and Rubin (1996). Because, by definition, proper methods reflect sampling variability correctly, inferences based on the multiply-imputed data are valid from the standard repeated-sampling (i.e., design-based) frequentist perspective.

One important principle related to incorporating appropriate variability is that imputations should be random draws rather than best predictions. Imputing best predictions can lead to distorted estimates of quantities that are not linear in the data, such as measures of variability and correlation, and it generally results in severe underestimation of uncertainty and therefore invalid inferences.

For imputations to be proper, the variability due to estimating the model must be reflected along with the variability of data values given the estimated model. For this purpose, a two-stage procedure is often useful. Fixing parameters at a point estimate (e.g., the maximum likelihood estimate), across the M imputations and drawing $\mathbf{X}_{\text{mis}}$ from this distribution-generally leads to inferences based on the multiply-imputed data that are too sharp. First drawing the parameters from their joint posterior distributions is preferable.

The two-stage paradigm can be followed in the context of nonparametric methods such as hot-deck imputation as well as in the context of parametric models with formal posterior distributions for parameters. The simple hot-deck procedure that randomly draws imputations for incomplete cases from matching complete cases is not proper because it ignores the sampling variability due to the fact that the population distribution of complete cases is not known but rather is estimated from the complete cases in the sample. Rubin and Schenker (1986, 1991) discussed the use of the bootstrap (see *Resampling Methods of Estimation*) to make the hot-deck procedure proper, and called the resulting procedure the 'approximate Bayesian bootstrap.' The two-stage procedure first draws a bootstrap sample from the complete cases and then draws imputations randomly from the bootstrap sample. Thus, the bootstrap sampling from the complete cases before drawing imputations is a nonparametric analogue of drawing values of the parameters of the imputation model from their posterior distribution before imputing conditionally upon the drawn parameter values. Heitjan and Little (1991) illustrated the use of bootstrap sampling to create proper imputations with a parametric model.

3.3 Choice of M

The choice of M involves a tradeoff between simulating more accurately the posterior distribution and using a smaller amount of computing and storage. The effect of the value of M on accuracy depends on the fraction of information about the estimand Q that is missing (Rubin 1987, Chaps. 3, 4). With ignorable missing data and just one variable, the fraction of missing information is simply the fraction of data values that are missing. When there are several variables and ignorable missing data, the fraction of missing information is often smaller than the fraction of cases that are incomplete because of the ability to predict missing values from observed values; this is often not true with nonignorable missing data, however.

For the moderate fractions of missing information (< 30 percent) that occur with most analyses of data from most large surveys, Rubin (1987, Chap. 4) showed that a small number of imputations (say, $M = 3$ or 4) results in nearly fully efficient estimates of Q. In addition, if proper multiple imputations are created, then a substantial body of work suggests that the resulting inferences generally have close to their nominal coverages or significance levels, even when

the number of imputations is moderate (Clogg et al. 1991, Heitjan and Little 1991, Rubin and Schenker 1986 and additional publications cited by Rubin 1996).

3.4 Use of Iterative Simulation Techniques

Recent developments in iterative simulation, such as data augmentation (Tanner and Wong 1987) and Gibbs sampling (Gelfand and Smith 1990), can facilitate the creation of multiple imputations in complicated parametric models (see *Markov Chain Monte Carlo Methods*). Consider a joint model for $\mathbf{X}_{\mathbf{obs}}$ and $\mathbf{X}_{\mathbf{mis}}$ governed by a parameter, say θ. The data augmentation (Gibbs sampling) procedure that results in draws from the posterior distribution of θ produces multiple imputations as well. (Schafer 1997) developed algorithms that use iterative simulation techniques to multiply impute data when there are arbitrary patterns of missing data and the missing-data mechanism is ignorable.

4. Some Recent Applications of Multiple Imputation in Social-science Research

To illustrate settings in which multiple imputation can be useful, brief descriptions of several recent applications of multiple imputation to social-science research are now given. References to many other examples are provided by Rubin (1996).

4.1 Multiple Imputation to Achieve Comparability of Industry and Occupation Codes Over Time

The first large-scale application of multiple imputation for public-use data involved employment information from the decennial census. In each census, such information is obtained from individuals in the form of open-ended descriptions of occupations. These descriptions are then coded into hundreds of industries and occupations, using procedures devised by the Census Bureau. Major changes in the procedures for the 1980 census resulted in public-use databases from the 1980 census with industry and occupation codes that were not directly comparable to those on public-use databases from previous censuses, in particular, the 1970 census. This lack of comparability made it difficult to study such topics as occupation mobility and labor force shifts by demographic characteristics.

Because of the size of the 1970 public-use databases (about 1.6 million cases) and the inaccessibility of the original descriptions of occupations, which were stored away on microfilm, it would have been prohibitively expensive to recode the 1970 data according to the 1980 scheme. There existed, however, a smaller (about 127,000 cases) double-coded sample from the 1970 census, that is, a sample with both 1980 and 1970 occupation codes, that had been created for other purposes. Based on Bayesian models estimated from the double-coded sample, five sets of 1980 codes were imputed for the 1970 public-use database as outlined by Clogg et al. (1991). Simulations described by Rubin and Schenker (1987) support the validity of multiple-imputation inferences in this application. Finally, analyses described by Schenker et al. (1993) demonstrated potential gains in efficiency from basing analyses on the multiply imputed public-use sample rather than on the smaller double-coded sample, even though the actual 1980 codes are known on the latter and unknown on the former.

4.2 Multiple Imputation for Missing Data in a National Health Survey

The National Health and Nutrition Examination Survey (NHANES) is conducted by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of the United States population and important subgroups. The data from NHANES are obtained through household interviews and through standardized physical examinations. NCHS has recently made available a database from the third NHANES, in which multiple imputation based on complicated multivariate models and iterative simulation techniques was used to handle missing data. This multiple-imputation project is significant because it demonstrates the feasibility of generating proper multiple imputations for a public-use sample obtained by a complex survey design with many observations and variables, varying rates of missingness on key variables of interest, and various patterns of nonresponse.

Documentation available with the database contains descriptions of the models and procedures used to multiply impute data for the third NHANES, as well as descriptions and examples of how to analyze the multiply-imputed data. Earlier work on the project is summarized by Schafer et al. (1996). In particular, a simulation study of the validity of multiple-imputation interval estimates was conducted based on a hypothetical population constructed from data from the third NHANES as well as earlier versions of NHANES. Results indicated that the imputation procedures were successful in creating valid design-based repeated-sampling inferences.

4.3 Multiple Imputation in a Study of Drinking Behavior with Follow-up for Nonresponse

The Normative Aging Study was a longitudinal study of community-dwelling men conducted by the Veterans Administration in Boston. In a survey on drinking behavior conducted as part of this study, a large fraction of the men surveyed provided essentially complete information, whereas for a smaller fraction, there was background information available but no information on drinking behavior. Because drinking

behavior is a sensitive subject, there was concern that such nonresponse might be nonignorable.

In a follow-up survey, information on drinking behavior was obtained from about one-third of the initial nonrespondents. Glynn et al. (1993) multiply imputed data on drinking for the roughly two-thirds of the initial nonrespondents who remained, under the assumption that, in the stratum of initial nonrespondents, given the follow-up information as well as the background information that had been collected previously, nonresponse was now ignorable. It was found that inferences about the effect of retirement status on drinking behavior (adjusting for age) were very sensitive to whether the multiply-imputed data or only the data for the initial respondents were used.

4.4 A Contest: Multiply Imputing the Consumer Expenditure Survey

Raghunathan and Rubin (1997) describe a simulation study to examine the frequentist properties of various methods of generating multiple imputations that was funded by the United States Bureau of Labor Statistics (BLS). The simulation study used real data from the Consumer Expenditure Survey (CES), which is the most detailed source of expenditures, income, and demographics collected by the US government. The objective of the simulation study was to compare frequentist performance of imputation-generation methods that differed by whether they fully accounted for parameter uncertainty and/or whether they conditioned on all observed variables. The recommended approach in the multiple-imputation literature is to generate imputations that account for all sources of uncertainty and condition on all observed information. Some BLS staff questioned this recommendation, preferring to use simpler methods for generating imputations that do not account for parameter uncertainty and exclude some variables, in this case measures of expenditure, because of concerns about endogeneity.

The artificial population from which the simulation samples were drawn consisted of complete reporters from the CES. BLS staff generated 200 independent samples, each of approximate size 500 from the constructed population, then imposed an ignorable monotone missing-data mechanism on each sample, resulting in 200 incomplete samples. Two variables in the samples had missing data: ethnicity and income. The incomplete samples were then sent to the multiple-imputation team, who were unaware of the exact nature of the missing-data mechanism, for analysis. All the variables used in the missing-data mechanism as well as some other variables were provided to the multiple-imputation team without noting which were which.

The multiple-imputation team generated five imputations for each incomplete sample under a variety of approaches: accounting for parameter uncertainty and using all variables, accounting for parameter uncertainty but excluding some variables (expenditure information), ignoring parameter uncertainty and using all variables, and ignoring parameter uncertainty and excluding some variables. The multiply-imputed samples were sent to BLS for analysis. The results were that the recommended method for generating imputations, accounting for parameter uncertainty and including all variables, produced confidence intervals that had the nominal coverage whereas all of the other methods produced intervals that undercovered, with the method that ignored parameter uncertainty and excluded expenditure information having the worst performance.

4.5 Multiple Imputation in the Survey of Consumer Finances

Kennickell (1998) reported on the use of multiple imputation for handling missing data and for disclosure limitation in the Survey of Consumer Finances (SCF). The SCF is a triennial survey conducted by the US Board of Governors of the Federal Reserve System in cooperation with the Statistics of Income Division of the Internal Revenue Service. The SCF is a major source of detailed asset, liability, and general financial information and includes over 2,700 questions. The SCF is focused on the details of households' finances, and due to the sensitivity of this topic it suffers from substantial unit and item nonresponse rates. In addition, because the SCF asks many questions about monetary amounts, the skip patterns in the SCF are very complex, resulting in complicated response pat terns. These factors make direct analysis of the incomplete SCF data extremely difficult for general users.

To combat the problem of missing data in the SCF, multiply-imputed versions of the SCF have been and continue to be released to the public. The SCF multiple imputations were generated using a novel variation of the Gibbs sampling algorithm that combines great modeling flexibility with computational tractability. Multiple imputation is also being used for disclosure limitation of income information in the SCF Kennickell (1997), based on ideas put forth by Rubin (1993) for protecting sensitive data by creating via imputation completely artificial populations.

4.6 Handling 'Don't Know' Survey Responses in the Slovenian Plebiscite

The Republic of Slovenia separated from Yugoslavia on October 8, 1991. One year before, Slovenians voted on independence through a plebiscite. Of eligible voters, 88.5 percent voted in favor of independence. To predict the results of the plebiscite, the Slovenian

government inserted questions into the Slovenian Public Opinion (SPO) Survey that was conducted four weeks prior to the plebiscite. The SPO is a face-to-face survey of approximately 2,000 voting-age Slovenians containing questions on many aspects of Slovenian life, including education and health. It has been conducted one to two times each year for the past 30 years. The plebiscite-related questions inserted into the SPO had three possible responses: 'yes,' 'no,' and 'don't know'(DK). In some surveys DK is a valid response, but in the plebiscite everyone votes either 'yes' or 'no' to the issue of independence, either by voting in person or by staying home (effectively a 'no' vote). Hence, the DK responses to the SPO survey question can be viewed as missing or masked data, masking either a 'yes' or 'no' plebiscite response.

Rubin et al. (1995) analyzed the SPO results with the goal of predicting the true percentage of voters in favor of independence. They treated the DK responses as missing data and considered ignorable and nonignorable missing-data mechanisms. Their prediction from a straightforward ignorable model was very accurate, while that from their most realistic nonignorable model was substantially far from the truth. The results are particularly interesting because the predictions were obtained before the true percentage was known and indicate that simple ignorable models can produce better answers than more complex and seemingly more plausible nonignorable models.

5. *Conclusion*

Multiple imputation is a general approach for handling the pervasive and often challenging problem of reaching statistically valid inferences from incomplete data (see *Statistical Data, Missing*). It was derived from a Bayesian perspective but produces inference procedures that have excellent frequentist properties. Generating multiple imputations can be difficult, and fortunately computer software is appearing that aids in this task (see below), but analyzing a multiply-imputed data set was designed to be almost as easy as analyzing a complete data set. A multiple-imputation inference is obtained by applying a complete-data inference procedure to each of the multiple data sets completed by imputation and then combining these estimates using simple combining rules. To learn more about multiple imputation (see Rubin 1987, 1996 and Schafer 1997).

There is currently only a limited amount of software for generating multiple imputations under multivariate complete-data models and for analyzing multiply-imputed data sets (i.e., completing the data sets, running the complete-data analyses, and combining the complete-data outputs), but the situation appears to be rapidly improving. For generating imputations, software to implement the methodology developed by Schafer (1997) has been written for the S-PLUS (Mathsoft 2001) statistical package and is freely available on the Internet. This software includes programs for multiple imputation in the contexts of incomplete multivariate normal data, incomplete categorical data, and incomplete data under the general location model allowing both categorical and multivariate normal variables.

A revised and expanded version of this software, including facilities for easily analyzing multiply-imputed data, has been incorporated into recent release of S-PLUS. SOLAS (Statistical Solutions 2000), the first commercial product for multiple imputation, combines a graphical user interface and integrated environment for analyzing multiply-imputed data with flexible and efficient methods of generating imputations that exploit the computational and modeling advantages of near monotone missing-data patterns. SAS Institute is developing procedures for analyzing and generating multiply imputed data under a variety of parametric models. These procedures are expected to be available in the very near future. The web site www.multiple-imputation.com contains a fairly complete list of software for multiple imputation and serves as source for news and events associated with multiple imputation.

Bibliography

Barnard J, Rubin D B 1999 Small-sample degrees of freedom with multiple imputation. *Biometrika* **86**: 948–55

Clogg C C, Rubin D B, Schenker N, Schultz B, Weidman L 1991 Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* **86**: 68–78

Gelfand A E, Smith A F M 1990 Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409

Glynn R J, Laird N M, Rubin D B 1993 Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association* **88**: 984–93

Heitjan D F, Little R J A 1991 Multiple imputation for the fatal accident reporting system. *Applied Statistics* **40**: 13–29

Kennickell A B 1997 Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In: *Record Linkage—1997: Proceedings of an International Workshop and Exposition*. US Office of Management and Budget, Washington, DC, pp. 248–67

Kennickell A B 1998 Multiple imputation in the Survey of Consumer Finances. In: *ASA Proceedings of the Business and Economic Statistics Section*. American Statistical Association, Alexandria, VA, pp. 11–20

Mathsoft 2001 *S-PLUS User's Manual, Version 6*. StatSci, a division of Mathsoft Inc., Seattle, WA

Meng X-L 1994 Multiple-imputation inferences with uncongenial sources of input (Disc: P558–573). *Statistical Science* **9**: 538–58

Raghunathan T E, Rubin D B 1997 Roles for Bayesian techniques in survey sampling. In: *Proceedings of the Survey Methods Section of the Statistical Society of Canada*. Statistical Society of Canada, McGill University, Montreal, PQ, pp. 51–5

Rubin D B 1976 Inference and missing data. *Biometrika* **63**: 581–90

Rubin D B 1987 *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York
Rubin D B 1993 Comment on 'Statistical Disclosure Limitation'. *Journal of Official Statistics* **9**: 461–8
Rubin D B 1996 Multiple imputation after 18+ years (Pkg: P473–520). *Journal of the American Statistical Association* **91**: 473–89
Rubin D B, Schenker N 1986 Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**: 366–74
Rubin D B, Schenker N 1987 Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *Journal of Official Statistics* **3**: 375–87
Rubin D B, Schenker N 1991 Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine* **10**: 585–98
Rubin D B, Stern H S, Vehovar V 1995 Handling 'Don't know' survey responses: The case of the Slovenian Plebiscite. *Journal of the American Statistical Association* **90**: 822–8
Schafer J L 1997 *Analysis of Incomplete Multivariate Data*. Chapman and Hall, New York
Schafer J L, Ezzati-Rice T M, Johnson W, Khare M, Little R J A, Rubin D B 1996 The NHANES III multiple imputation project. In: *ASA Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, VA, pp. 28–37
Schenker N, Treiman D J, Weidman L 1993 Analyses of public use Decennial Census data with multiply imputed industry and occupation codes. *Applied Statistics—Journal of the Royal Statistical Society Series C* **42**: 545–56
Statistical Solutions 2000 *SOLAS for Missing Data Analysis 2.0*.Statistical Solution Ltd., Cork, Ireland
Tanner M A, Wong W H 1987 The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**: 528–40

J. Barnard, D. B. Rubin, and N. Schenker

Multistate Transition Models in Demography

Demographic multistate models deal with the transitions that members of a population may experience during their lifetimes as they move from one state of existence to another. These states may be demographic ones, such as family status or occupational status, or geographic ones, such as region of residence or place of work, or more general social groups like classes in school, electoral constituencies, and so on.

1. Increment-decrement Tables

During the 1960s and the 1970s, demographic work on multistate life tables produced a generalization of classical mathematical techniques used in demography for more than 300 years (see *Life Table*). These new models can be viewed as superimposing a set of two or more life statuses on the single dimension of the classical life table, which only has a live-to-dead transition. For example, they can incorporate several forms of decrement from an initial state, such as mortality by cause of death in multiple-decrement tables. They can also chain together a series of multiple-decrement tables without permitting re-entries into a state previously occupied, such as transitions to rising educational levels in hierarchical tables. They finally lead to more general nonhierarchical increment-decrement tables, which accommodate re-entrants into states and differentiate interstate moves by both origin and destination (Land and Rogers 1982), as in multiregional migration models.

These models describe the dynamics of populations where individuals in a given state are assumed to be homogeneous: the probability to move from one state to another is the same for all individuals in the same state. Population heterogeneity can be accounted for to some extent by increasing the number of possible states. However, the introduction of a growing number of subpopulations in order to ensure homogeneity in each state quickly leads to group sizes so small as to rule out all analysis. What is more, one can never be sure of having taken into account every factor of heterogeneity.

1.1 A Markov Chain Formulation

Mathematics allows us to develop tools to describe transitions between the various observed states. One considers an individual life course as the result of a stochastic process occurring in the given state space. The time dimension may represent biological age or duration since some initial event, for example, time since entry into the labor force. This process is said to satisfy the Markov property if its present state is all that counts in predicting its future: what other states were occupied previously by the individual has no effect on the coming move. The future is independent of the past except through the present. The process is not always stationary, as transition probabilities may depend on time. This leads to time-continuous Markov chain models with a finite state space.

Consider a single sample path, and let $I_i(t) = 1$ if it is in state i at time t, with $I_i(t) = 0$ otherwise. One defines the transition intensity from state i to state j at time t as

$$\mu_{ij}(t) = \lim_{\Delta t \to 0} \frac{P(I_i(t+\Delta t) = 1 \mid I_i(t) = 1)}{\Delta t} \qquad (1)$$

For increment-decrement tables only aggregate data on flows between states are used; one does not use

individual trajectories in the calculations. In some cases transition intensities may be specified as given functions that are known except for the value of parameters that must be estimated from the observed data. In other cases one uses approximations such as piecewise constant intensities provided time intervals are available over which the intensities can be taken as constant. Once these transition intensities are estimated, one may find the corresponding transition probabilities for every suitable time period, for example, the probability that an individual in state i at time t_1 will be in state j at time t_2. It is also useful to estimate expected sojourn times in each state and many other summaries of behavior. Finally, these analyses may be completed using regression models with aggregate characteristics for each state (see *Event-history Analysis in Continuous Time*).

The availability of aggregate or contextual data may invite the analyst to examine the effect that characteristics of the groups studied have on group behavior. The aggregate or contextual characteristics are interpreted as a set of constraints that each subpopulation imposes on its members (or corresponding opportunities) and which influences their behavior. For example, such an analysis may perhaps reveal a positive association between the rate of unemployment in a region and its out-migration rate. But there is a real danger of inferring from such an observation that individuals of this population who are unemployed have a higher probability of out-migrating from a region, whereas all that is identified by the model is that a high rate of unemployment is accompanied by a high rate of out-migration, whether the individuals involved are economically active, unemployed, or inactive. This type of mistaken inference from group-level associations to individual behavior is known as the ecological fallacy. Committing this fallacy is avoided in event-history analysis, which is discussed later.

1.2 More Complex Models

The Markov chain assumption is restrictive and constitutes a rough approximation for many demographic processes. For example, in migration analysis one needs to account for duration dependence in the propensity to move. There are also preferred patterns of moves between regions; for instance, return moves back to a region of origin are preferred to moves to other regions. Such features can be taken into account using more complex models.

The first way to analyze non-Markovian processes is to expand the state space so that the process in the new space is Markovian. Such an extension, however, again leads to an increase of the amount of data necessary to estimate the large number of transition intensities beyond what is possible with the usual data sets. A second possibility lies in the use of a semi-Markov process. In such a model the probability of moving at a given time can bear any arbitrary relation to the duration of stay (Ginsberg 1971). Nevertheless, any extension leaves out some property of the studied process and even more complex models are needed. For example, in the case of migration a semi-Markov process is unable to take into account the higher probability of return moves unless one introduces information about previously occupied states.

2. Multistate Event-history Models

Increment-decrement tables are computed primarily from aggregate data. It will normally be preferable to use individual-level data instead. This approach allows us to examine how events that occur to an individual can influence their future life course and how personal characteristics can give individuals behavior patterns that are different from those of other individuals. Event history models (see *Event-history Analysis in Continuous Time*) have been developed to deal with such information, mainly during the 1980s.

2.1 Interaction Between Events

Again a mathematical formulation helps us analyze the issues involved (Andersen et al. 1993). Let $P_i(t)$ be the number of individuals observed in state i just before time t and let $M_{ij}(t)$ be the number of observed transitions from state i to state j during the time interval $[0, t]$. Then the vector $M = \{M_{ij}, i \neq j\}$ is a multivariate counting process where M_{ij} has intensity process $\mu_{ij}(t)P_i(t)$. Here $\mu_{ij}(t)$ is the corresponding transition intensity.

With these tools it is possible to follow each individual and to study interactions between various processes. Consider the links between family formation and urban migration in France, for example. In this country, the metropolitan areas are centered on Paris, Lyons, and Marseilles, and using data for the years 1925 to 1950 Courgeau (1989) followed women who lived in these areas or outside them on their 14th birthday. For these women, childbearing intensities for birth orders higher than one turned out to depend strongly on whether they had migrated or not. For women from outside the metropolitan areas, migration to such an area reduced their fertility by one third. Conversely, migration to a nonmetropolitan area brought about a sharp rise in fertility for women from the cities. After migration, cumulative childbearing intensities were 1.4 to 2 times higher than for those who remained in the metropolitan area where they started out. Migration to a metropolitan area attracted women whose fertility before migration was already similar to that prevailing in the urban area,

and they displayed selective behavior. Migration to a less urbanized area attracted women whose fertility before migration was similar to that of other women in the urban area, and the women showed adaptive behavior.

Conversely, the birth of successive children also influenced women's mobility; there is reciprocal dependence. The intensity of moving to a metropolitan area was reduced after each successive birth, but that of moving in the opposite direction was increased slightly by these births.

This analysis illustrates the complex dependencies that can be observed between various dimensions of behavior, in this case childbearing and migration. We can speak of local dependence when there is a one-sided influence of one process on the other without reciprocity (Schweder 1970), and of total independence if at any time each event occurs independently of the other, though this rarely occurs in practice. For even more complex dependencies, see Courgeau and Lelièvre (1992).

It is possible to enlarge the ambit of the analysis to cover situations with even more states. For example, one can consider studying fertility occurring in various types of partnership (marriage or cohabitation), or interactions between fertility and changes of residence.

2.2 Population Heterogeneity

Suppose that each individual carries a collection of possibly time-dependent characteristics which may change at distinct points in time due to outside influences and also on the influence of the individual's own previous life history. The effect of these characteristics on the events of interest is investigated via a time-dependent regression model, which may be a generalization of the Cox (1972) model and may allow us to model the hazard or intensity using fixed and time-dependent covariates.

Consider an individual k with a vector of possibly time-varying characteristics $Z^k(t)$ at time t. We write his hazard rate (or intensity) for a transition from state i to state j as

$$\mu_{ij}^k(t) = \mu_{ij}^0(t)\exp\{\beta_{ij}Z^k(t)\} \tag{2}$$

where the baseline hazard (intensity) $\mu_{ij}^0(t)$ is the same for all individuals and each β_{ij} is a vector of parameters that measure the impact of the covariates. Individual-level covariates may be binary (being married or not), polytomous (being (a) economically active and employed, (b) unemployed, or (c) economically inactive), a feature that can be treated as continuous (such as income), or even more diverse. An analysis based on such a model can reveal whether individuals who are unemployed have a higher probability of migrating out of a region than employed individuals, for instance, a result which cannot be shown with aggregate data.

Courgeau (1989) used such a model to reconsider the interaction between the birth of the third child and mobility, taking into account individual characteristics. Some of these characteristics turned out to be affected by the migration process in France. Being the daughter of someone engaged in agriculture had an effect on such a birth irrespective of whether the woman lived in a metropolitan area or not, for instance. Before migration this would delay the birth, but after a move the effect disappeared. Other aspects of behavior were not affected by this migration. For example, if a woman had a large number of siblings, she would be more likely to have a third child irrespective of whether she came from a rural or an urban background. Once these different effects were taken into account, the migration to metropolitan areas continued to induce a marked reduction in the chance of giving birth to a third child, whereas migration to less urbanized areas always raised fertility.

However, as one can never be sure to have collected information on all aspects that might explain a given behavior, unobserved heterogeneity may be an issue. When it is independent of the characteristics observed, it reduces the absolute value of the estimates of regression parameters β_{ij} that represent the effect of observed characteristics, but it does not change their sign (Bretagnole and Huber-Carol 1988). When the observed characteristics turn out to have a significant effect, one can then conclude that this will not be modified by independent unobserved heterogeneity; but when no significant effects are found, a real effect can be disguised by unobserved heterogeneity.

There are various ways of accounting for unobserved heterogeneity, but a common procedure is to posit a set of individual-level scalar random effects V_{ij}^k that are added in the exponent of the intensity function, as in

$$\mu_{ij}^k(t) = \mu_{ij}^0(t)\exp\{\beta_{ij}Z^k(t) + V_{ij}^k\} \tag{3}$$

These random variables are then supposed to represent selectively and/or heterogeneity in the population by picking up effects of covariates that are not included in the intensity-regression analysis.

Even when these powerful tools are used, care is needed, given that one frequently specifies behavior as influenced only by the characteristics of the individual. The danger here is of committing the atomistic fallacy if no attention is paid to the context in which human behavior occurs. In point of fact, this context often does have an influence, in which case it is fallacious to consider individuals in isolation from the constraints and opportunities imposed by the society and milieu in which they live. Contextual variables may be aggregates of individual level characteristics (such as the percent married in the individual's region of residence) or general contextual characteristics (such as the number of hospital beds in the same region or its political color). More complex analytical procedures

can also be employed (such as the average income or the correlation between income and matrimonial status).

Formally, contextual variables appear in the formulas above simply as factors $Z^k(t)$ whose values are common to all individuals k who live in the same context.

3. Contextual and Multilevel Analysis

The need to introduce the influence of contextual characteristics has given rise to the idea of working simultaneously on different levels of aggregation with the aim of explaining a behavior which is treated as individual rather than aggregated as in increment-decrement tables. This removes the risk of committing the ecological fallacy, since a characteristic on the aggregate or contextual level is only used to measure an influence that may be different from its equivalent at the individual level. It is introduced not as a substitute but as a feature of the subpopulation, a feature that influences the behavior of an individual who belongs to it. And the atomistic fallacy is also eliminated once the context in which the individual lives is specified correctly in the analysis.

The main limitation of contextual analysis is that the behaviors of individuals in the same context are treated as independent. Such an assumption is usually not true, as the behavior of an individual in a group often depends on other individuals in the same group. Think of members of a family or household, or citizens of a town. One may need to take into consideration the observed social structure of a group and the interactions that exist between its members as well as changes in such interactions over time (Lelièvre et al. 1998). Ignoring within-group dependence may lead to estimated variances of contextual effects that are biased downward, making confidence intervals too narrow. To handle this problem one extends the contextual model, following a development that took place during the late 1980s and in the 1990s. Here is a quick sketch of the main principles of such an extension.

Suppose that individuals are organized into groups and that individual k in group h has transition intensities of the form

$$\mu_{ij}^{kh}(t) = \mu_{ij}^{0}(t)\exp\{\beta_{ij}Z^{kh}(t) + V_{ij}^{kh}\} \qquad (4)$$

Observed dependences may be picked up through the specification of the $Z^{kh}(\cdot)$. To model unobserved dependence between individuals, introduce a set of group-specific random variables W_{ij}^{h} and replace Eqn. (4) by

$$\mu_{ij}^{kh}(t) = \mu_{ij}^{0}(t)\exp\{\beta_{ij}Z^{kh}(t) + V_{ij}^{kh} + W_{ij}^{h}\} \qquad (5)$$

Modern software like aML (see Lillard and Panis 2000) can handle analyses of such models. There is much interest currently in how one can identify sensible distributions for the heterogeneity variables $\{V_{ij}^{kh}\}$ and $\{W_{ij}^{h}\}$ ('mixing distributions') (see Horowitz 1999).

Multilevel analysis can be extended to more than two levels. The simplest and more widely used structure is hierarchical: each level is formed by grouping together of all units of the previous level: pupils are grouped in classes, which in turn are in schools, which are perhaps divided between the state and private sector, and so on. The relationship between the levels can be more complicated: individuals can be grouped in towns of ascending size order, but distinguished also between administrative, industrial, and tourist centers, for example. In this case there is a cross-classification that includes towns being classified by size and by function, say. It is of course, possible to have relationships that combine hierarchical and cross-classifications. For example, individuals may be classified by type of residential neighborhood and by type of place of work (cross-classification), which are themselves placed in a hierarchical classification of parishes and regions.

It is not necessary to study a single process alone, as when moves between different states of one kind are considered independently of other changes occurring simultaneously in other life domains. Dynamic multilevel event-history analysis can be used to study several interrelated processes at the same time. For instance, the study of fertility in the various regions of a country can be carried out, in principle, simultaneously with a study of migration between the regions, and this may be useful since some women change residence between the regions during their reproductive period.

4. Conclusion

Improvements in multistate analysis such as that sketched here are the subject of current research. One can expect much future development in the epistemological bases, methods of measurement, and procedures of analysis of a dynamic theory of human behavior along these lines.

See also: Ecological Fallacy, Statistics of; Event Sequence Analysis: Methods; Regression Discontinuity Design

Bibliography

Andersen P, Borgan O, Gill R, Keiding N 1993 *Statistical Models Based on Counting Processes*. Springer-Verlag, New York

Bretagnole J, Huber-Carol C 1988 Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics* **15**: 125–38

Courgeau D 1989 Family formation and urbanization. *Population: An English Selection* **1**: 123–46

Courgeau D 1995 Migration theories and behavioural models. *International Journal of Population Geography* **1**: 19–27

Courgeau D, Baccaïni B 1998 Multilevel analysis in social science. *Population: An English Selection* **10**: 39–71

Courgeau D, Lelièvre E 1992 *Event History Analysis in Demography*. Clarendon, Oxford, UK
Cox D 1972 Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* **B34**: 187–220
Ginsberg R 1971 Semi-Markov processes and mobility. *Journal of Mathematical Sociology* **1**: 233–62
Goldstein H 1995 *Multilevel Statistical Models*. Arnold, London
Hoem J, Funck Jensen U 1982 Multistate life table methodology: A probabilistic critique. In: Land K, Rogers A (eds.) *Multidimensional Mathematical Demography*. Academic Press, New York
Horowitz J L 1999 Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica* **76**(5): 1001–28
Keilman N 1993 Emerging issues in demographic methodology. In: Blum J, Rallu L (eds.) *European Population, II—Demographic Dynamics*. John Libbey/INED, Paris
Land K, Rogers A (eds.) 1982 *Multidimensional Mathematical Demography*. Academic Press, New York
Lelièvre E, Bonvalet C, Bry X 1998 Event history analysis of groups. The findings of an on-going research project. *Population: An English Selection* **10**: 11–37
Lillard L A, Panis C W A 2000 *aML Multilevel Multiprocess Statistical Software, Release 1.0*. EconWare, Los Angeles
Schweder T 1970 Composable Markov processes. *Journal of Applied Probability* **7**: 400–10

D. Courgeau

Multivariate Analysis: Classification and Discrimination

The need for classification arises in most scientific pursuits. Typically, there is interest in classifying an entity, say, an individual or object, on the basis of some characteristics (feature variables) measured on the entity. This classification is usually undertaken in two main contexts. In the first context, there is a finite number, say g, of predefined distinct populations, categories, classes, or groups, and the entity to be classified is assumed to belong to one (and only one) of these g possible groups. In order to assist with the construction of a classification rule for this purpose, there are usually available so-called training data from each group; that is, these training data comprise the features measured on some entities that are classified with respect to the g underlying groups. In statistical terminology, this classification process is referred to as *discriminant analysis*, while it is referred to as *supervised learning* or *learning with a teacher* in pattern recognition and machine learning.

In the second context the classification of some entities is undertaken on the basis of their feature data in the absence of any prior knowledge on the existence of distinct groups. In this case, the primary purpose of the classification is to put the entities into a number of clusters so that entities within a cluster tend to have more similar features than those in different clusters. In statistics, this process is referred to as *cluster analysis*; see *Statistical Clustering*. In this article, attention is confined to discriminant analysis.

1. Discriminant Analysis

It is supposed that an entity belongs to one of g possible groups, denoted by $G_1,\ldots, G_g$, and that there is a (feature) vector $\boldsymbol{x}$ containing p variables that can be measured on the entity. The group-membership of the entity is denoted by the categorical variable z, where $z = i$ implies that the entity belongs to G_i ($i = 1,\ldots, g$). In this notation, the topic of discriminant analysis may be viewed as being concerned with the relationship between z and the feature vector $\boldsymbol{x}$. Within this broad topic, there is a spectrum of problems that corresponds to the inference-decision spectrum in statistical methodology. At the decision end of the scale, the group membership of the entity is unknown and the intent is to make an outright assignment of the entity to one of the g possible groups on the basis of $\boldsymbol{x}$. That is, in terms of the present notation, the problem is to estimate z solely on the basis of $\boldsymbol{x}$. An example in which an outright assignment is required concerns the rejection or acceptance of loan applicants by a financial institution. For this decision problem there are two groups with, say G_1, referring to applicants who will service their loans satisfactorily, and G_2 to those who do not. The feature vector $\boldsymbol{x}$ for an applicant contains information such as age, income, and marital status. A rule based on $\boldsymbol{x}$ for allocating an applicant to either G_1 or G_2 (that is, either accepting or rejecting the loan application) can be formed from an analysis of the feature vectors of past applicants from each of the two groups.

At the other extreme end of the spectrum, no assignment of the entity to one of the possible groups is intended. Rather the problem is to draw inferences about the relationship between z and the feature variables in $\boldsymbol{x}$. An experiment might be designed with the specific aim to provide insight into the predictive structure of the feature variables. For example, a political scientist may wish to determine the socio-economic factors that have the most influence on the voting patterns of a population of voters. Between these extremes lie most of the everyday situations in which discriminant analysis is applied (Hand 1997, McLachlan 1992, Ripley 1996).

2. Allocation Rules

Let $r(\boldsymbol{x})$ denote an allocation rule, where $r(\boldsymbol{x}) = i$ implies that an entity with feature vector $\boldsymbol{x}$ is to be assigned to the ith group G_i. The allocation rates associated with this rule $r(\boldsymbol{x})$ are denoted by $e_{ij}(r)$,

where $e_{ij}(r)$ is the probability that a randomly chosen entity from G_i is allocated to G_j $(i, j = 1,\ldots, g)$. For a diagnostic test using the rule $r(\boldsymbol{x})$ in the context where G_1 denotes the absence of a disease or condition and G_2 its presence, the error rate $e_{12}(r)$ corresponds to the probability of a false positive, while $e_{21}(r)$ is the probability of a false negative. The correct allocation rates $e_{22}(r)$ and $e_{11}(r)$ are known as the sensitivity and specificity, respectively, of the diagnostic test.

Decision theory provides a convenient framework for the construction of discriminant rules in the situation where an outright allocation of an unclassified entity is required. For given $\boldsymbol{x}$, the expected loss or risk for allocation performed on the basis of the rule $r(\boldsymbol{x})$ is equal to $\Sigma_{i=1}^{g}\tau_i(\boldsymbol{x})Q[i, r(\boldsymbol{x})]$, where, for any u and v, $Q[u, v] = 0$ for $u = v$ and 1 for $u \neq v$, and where $\tau_i(\boldsymbol{x})$ is the posterior probability that the entity belongs to G_i which, by Bayes' theorem, can be expressed as

$$\tau_i(\boldsymbol{x}) = \pi_i f_i(\boldsymbol{x})/f(\boldsymbol{x}) \quad (i = 1,\ldots, g) \tag{1}$$

In Eqn. (1), $f_i(\boldsymbol{x})$ denotes the density of the feature vector $\boldsymbol{X}$ in group G_i, $f(\boldsymbol{x}) = \Sigma_{i=1}^{g}\pi_i f_i(\boldsymbol{x})$ is the unconditional density of $\boldsymbol{X}$, and π_i denotes the prior probability that the entity belongs to G_i. The costs of misallocation have been assumed to be zero here for a correct allocation and unity for a misallocation from one group to another.

An optimal rule of allocation can be defined by taking it to be the one that minimizes the conditional risk at each value $\boldsymbol{x}$ of the feature vector. In decision-theory language, such a rule is said to be a Bayes' rule. It follows that the Bayes' rule $r_o(\boldsymbol{x})$ assigns the entity to the group to which it has the highest posterior probability of belonging under the specified utility function.

3. *Sample-based Allocation Rules*

In practice, the group-conditional densities $f_i(\boldsymbol{x})$ are usually unknown. A basic assumption in discriminant analysis is that in order to estimate the unknown group-conditional densities there are entities of known origin on which the feature vector $\boldsymbol{X}$ has been recorded for each. These data are referred to in the literature as either the initial, reference, design, training, or learning data.

The initial approach to the problem of forming a sample discriminant rule, and indeed to discriminant analysis in its modern guise, was by Fisher (1936). In the context of $g = 2$ groups, he proposed that an entity with feature vector $\boldsymbol{x}$ be assigned on the basis of the linear discriminant function $\boldsymbol{a}^T\boldsymbol{x}$, where $\boldsymbol{a}$ maximizes an index of separation between the two groups. The index was defined to be the magnitude of the difference between the group-sample means of $\boldsymbol{a}^T\boldsymbol{x}$ normalized by the pooled sample estimate of its assumed common variance within a group.

To consider the formation of sample rules in general, we need to consider the sampling designs under which the training data $\boldsymbol{T}$ have been realized. There are two major schemes: joint or mixture sampling and z-conditional or separate sampling. They correspond respectively to sampling from the joint distribution of $\boldsymbol{Y} = (\boldsymbol{X}^T, Z^T)^T$ and to sampling from the distribution of $\boldsymbol{X}$ conditional on z. The first design applies to the situation where the feature vector and group of origin are recorded on each of n entities drawn from a mixture of the possible groups. Mixture sampling is common in prospective studies and diagnostic situations. In a prospective study design, a sample of individuals is followed and their responses recorded. With separate sampling in practice, the feature vectors are observed for a sample of n_i entities taken separately from each group G_i, and so it is appropriate to retrospective studies.

An obvious way of forming a sample-based allocation rule $r(\boldsymbol{x}; \boldsymbol{t})$ is to take it to be an estimated version of the Bayes' rule $\tau_o(\boldsymbol{x})$, where the posterior probabilities of group-membership $\tau_i(\boldsymbol{x})$ are replaced by some estimates $\hat{\tau}_i(\boldsymbol{x}; \boldsymbol{t})$ formed from the training data $\boldsymbol{t}$. A common approach, referred to as the sampling approach, is to use the Bayes' formula (Eqn. (1)) to formulate the $\tau_i(\boldsymbol{x})$ through the group-conditional densities $f_i(\boldsymbol{x})$. With the fully parametric approach to this problem, the group-conditional densities $f_i(\boldsymbol{x})$ are assumed to have specified functional forms except for a finite number of parameters to be estimated. A commonly used parametric family for the $f_i(\boldsymbol{x})$ for continuous feature data is the normal with either a linear or quadratic rule in $\boldsymbol{x}$ being obtained depending on whether the group-covariance matrices are taken to be equal or unequal. A compromise between linear and quadratic discriminant analyses is regularized discriminant analysis (Friedman 1989).

There is also the direct approach (the diagnostic paradigm) to the estimation of the $\tau_i(\boldsymbol{x})$, using either nonparametric estimates as with nearest-neighbor methods or parametric estimates via the logistic model. With the latter approach, it is the ratios of the group-conditional densities that are being modeled. The fundamental assumption of the logistic approach is that these ratios are linear, which is equivalent to taking the log (posterior) odds to be linear. The linearity here is not necessarily in the basic variables; transforms of these may be taken.

Another method for the direct modeling of the group-posterior probabilities is to use neural networks (Bishop 1995), which have been coming under increasing attention by statisticians; see Ripley (1996). For example, with a two-layered feed-forward neural network with m hidden units in the case of $g = 2$ groups, $\tau_1(\boldsymbol{x})$, is modeled as $\tau_1(\boldsymbol{x}) = \phi(w_{20} + \Sigma_{k=1}^{m} w_{2k} u_k)$, where $u_k = \phi(w_{10k} + \Sigma_{h=1}^{p} w_{1hk}(\boldsymbol{x})_h)$ is the value of the kth hidden unit, and the w_{1hk}, w_{2k} are parameters known as weights; $\phi(u)$ is the so-called activation function, often taken to be the

sigmoid function, $\phi(u) = 1/(1+e^{-u})$. Neural networks are fitted mostly using *back propagation*, which is essentially gradient descent using the chain rule, but the EM algorithm has been used too recently. By the inclusion of an appropriate number of hidden units in the hidden layer (or layers), neural networks can handle nonlinear boundaries between the groups.

An advantage of the logistic and neural network-based rules is that they can handle discrete feature variables. Also, the multinomial distribution can be used to model the group-conditional probabilities for discrete feature variables. But unless the number m of multinomial cells is very small or the sample size n is extremely large, some of the cells will contain no observations and so some form of smoothing has to be applied. The nonparametric estimates of the multinomial cell probabilities under the assumption of independence of all the p feature variables may impose too high a degree of smoothing on the multinomial estimates. Hence the aim of smoothing is to provide a degree of smoothing intermediate to that obtained under the extremes of the full multinomial and independence models.

To illustrate the construction of sample-based allocation rules, we consider the construction of a discriminant rule for the problem of assessing loan applications, as mentioned in Sect. 1. Here the applicant for a loan is a company and the feature vector contains information on its key financial characteristics. For the purposes of our illustration, we consider the case where there are $p = 2$ feature variables which are the first two principal components of three financial ratios concerning the debt, liquid assets, and working capital of a company. In Fig. 1, we have plotted the boundary

$$\hat{\tau}_1(\boldsymbol{x}_j; \boldsymbol{t}) = \hat{\tau}_2(\boldsymbol{x}_j; \boldsymbol{t})$$

for the sample version of the Bayes' rule (the sample linear discriminant rule) formed under the assumption that the feature vector has a bivariate normal distribution with a common group-covariance matrix in each of the two groups G_1 and G_2, corresponding to good (satisfactorily serviced) and bad (unsatisfactorily serviced) loans. It can be seen that the overall error rate of this rule reapplied to the training data (its apparent error rate) is not as low as that of the quadratic rule formed without the restriction of a common group-conditional covariance matrix for the feature vector $\boldsymbol{X}$. With its curved boundary, the quadratic rule misallocates only four (all bad) loans. One should be mindful, however, that the apparent error rate provides an optimistic assessment of the accuracy of a discriminant rule when it is applied to data not in the training set (see next section). We have also plotted in Fig. 1 the boundary of the rule obtained by modeling the distribution of the feature vector $\boldsymbol{X}$ by a two-component normal mixture, which results in the misallocation of one bad loan at the expense of misallocating two good loans. The latter rule is an example of a more flexible approach to classification than that based on the assumption of normality for each of the group-conditional distributions, as to be considered now.

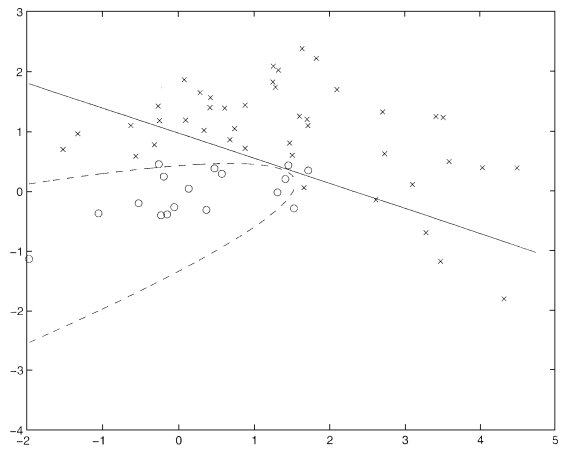

Figure 1
Boundaries of linear rule (solid line), quadratic rule (dashed line), and mixture model-based rule (dotted line); crosses and circles denote good and bad loans, respectively

4. *Flexible Discriminant Rules*

A common nonparametric approach to discriminant analysis uses the kernel method to estimate the group-conditional densities $f_i(\boldsymbol{x})$ in forming an estimate of the Bayes' rule. More recently, use has been made of finite mixture models, mainly normal mixtures, to provide flexible rules of discrimination (Hastie and Tibshirani 1996). Mixture models provide an extremely flexible way of modeling a density function (McLachlan and Basford 1988, McLachlan and Peel 2000), and can be fitted in a straightforward manner via the EM algorithm (McLachlan and Krishnan 1997, McLachlan et al. 1999). Among other work on flexible discrimination, there is the FDA (flexible discriminant analysis) approach based on nonparametric regression (see Hastie and Tibshirani 1996). The generic version of FDA based on smoothing splines proceeds by expanding the predictors in a large (adaptively selected) basis set, and then performing a penalized discriminant analysis in the enlarged space using a linear discriminant or a normal mixture model-based rule. The class of nonlinear regression methods that can be used include additive models, the multivariate adaptive regression spline (MARS) model, projection pursuit regression, and neural networks. In machine learning, there has been increasing attention in the case of two groups given to nonlinear rules based on

the foundations of support vector machines, as developed by Vapnik (2000). With this approach, the initial feature space is mapped into a higher dimensional space by choosing a nonlinear mapping and then choosing an optimal separate hyperplane in the enlarged feature space.

A rather different approach to the allocation problem as considered up to now, is to portray the rule in terms of a binary tree. The tree provides a hierarchical representation of the feature space. An allocation is effected by proceeding down the appropriate branches of the tree. The classification and regression tree (CART) methodology of Breiman et al. (1984) has contributed significantly to the growing popularity of tree classifiers; see also the C4.5 decision-tree learning algorithm of Quinlan (1993). In the context of tree classifiers in particular, there has been growing interest in the use of boosting, which is one of the most important recent developments in classification methodology. Algorithms like the Adaboost algorithm of Freund and Schapire (1996) and the bagging algorithm of Breiman (1996) can often improve performance of unstable classifiers like trees or neural nets by sequentially applying them to reweighted versions of the training data and taking a weighted majority vote of the sequence of classifiers so formed. The test error of the weighted classifier usually does not increase as its size increases, and often is observed to decrease even after the training error reaches zero. Friedman et al. (2000) have exhibited the link of boosting to statistical ideas on additive fitting.

A guide to the performance of a given discriminant rule can be obtained by applying it to the training from which it was formed and noting the proportion of these observations misallocated (the apparent rate). However, it is well known that the apparent error rate gives too optimistic a view on how the rule might perform on future cases; see McLachlan (1992, Chap. 10) for an account of error-rate estimation. One way of correcting for this bias is to use cross-validation, but it produces an estimate that often is more highly variable than the apparent error rate. This led Efron (1983) to consider his so-called .632 estimator, which is a weighted sum of the apparent error rate and the bootstrap error rate at an original data point not in the training set $\boldsymbol{t}$; see Efron and Tibshirani (1997) for an improved version known as the .632+ estimator. The bootstrap methodology can be used also to estimate the variability in the errors and in other quantities of interest, such as the estimated parameters in the form of the estimated rule.

An excellent summary of available software and algorithms for discriminant analysis has been given by the Panel on Discriminant Analysis, Classification and Clustering (1989); see also Huberty (1993) and Michie et al. (1994).

In conclusion, discriminant analysis is concerned with the construction of allocation rules for the assignment of unclassified entities to distinct groups in the situation where there are data on entities of known group of origin (training data). In the past, linear rules such as those based on Fisher's linear discriminant function were commonly used in practice. In recent times with the wide availability of high speed computers, there has been increasing attention given to the development of more complex nonlinear rules such as those based on flexible parametric models, neural networks, support vector machines, and classification trees.

See aslo: Multivariate Analysis: Discrete Variables (Logistic Regression); Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Discrete Variables (Overview); Multivariate Analysis: Overview

Bibliography

Bishop C M 1995 *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK

Breiman L 1996 Bagging predictors. *Machine Learning* **24**: 123–40

Breiman L, Friedman J H, Olshen R A, Stone C J 1984 *Classification and Regression Trees*. Wadsworth, Belmont, CA

Efron B 1983 Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**: 316–31

Efron B, Tibshirani R 1997 Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* **92**: 548–60

Fisher R A 1936 The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**: 179–88

Freund Y, Schapire R E 1996 Experiments with a new boosting algorithm. In: Saitta L (ed.) *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kaufmann, San Francisco, pp. 148–56

Friedman J H 1989 Regularized discriminant analysis. *Journal of the American Statistical Association* **84**: 165–75

Friedman J H, Hastie T, Tibshirani R 2000 Additive logistic regression: a statistical view of boosting. *Anuals of Statistics* **28**: 337–407

Hand D J 1997 *Construction and Assessment of Classification Rules*. Wiley, New York

Hastie T, Tibshirani R 1996 Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society* B **58**: 155–76

Huberty C J 1994 *Applied Discriminant Analysis*. Wiley, New York

McLachlan G J 1992 *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York

McLachlan G J, Basford K E 1988 *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York

McLachlan G J, Krishnan T 1997 *The EM Algorithm and Extensions*. Wiley, New York

McLachlan G J, Peel D 2000 *Finite Mixture Models*. Wiley, New York

McLachlan G J, Peel D, Basford K E, Adams P 1999 The EMMIX software for the fitting of mixtures of normal and *t*-components. *Journal of Statistical Software* **4**(2)

Michie D, Spiegelhalter D J, Taylor C C (eds.) 1994 *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Chichester, UK

Panel on Discriminant Analysis, Classification, and Clustering 1989 Discriminant analysis and clustering. *Statistical Science* **4**: 34–69
Quinlan J R 1993 *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA
Ripley B D 1996 *Pattern Recognition and Neural Networks*. Cambridge University Press, New York
Vapnik V N 2000 *The Nature of Statistical Learning Theory*, 2nd edn. Springer, New York

G. McLachlan

Multivariate Analysis: Discrete Variables (Correspondence Models)

1. Introduction

Correspondence models extract information on the association between categorical variables. As in correspondence analysis (see *Scaling: Correspondence Analysis*), correspondence models are based on a canonical parametrization of the joint distribution of the variables. Traditional correspondence analysis (Benzécri 1973, Greenacre 1984, Hill 1974, Nishisato 1980, Gifi 1990) is concerned with methods used for description and detection of patterns of association between variables via graphical displays, and does not consider the role of random sampling in the generation of the observed data from a population with an associated probability model. In contrast, correspondence models are explicitly concerned with inferential methods for the study of probability models which are fitted to observed data under the assumption that the data have been obtained by random sampling (Goodman 1985, 1986, Gilula and Haberman 1986, 1988). The parameterizations used in correspondence models may still be used to construct descriptive measures and graphs; however, unlike in traditional correspondence analysis, correspondence models involve formal statistical methods such as tests of goodness of fit and approximate confidence intervals for parameters. Development of correspondence models is most straightforward in the case in which the relationship of two categorical random variables is studied (Sect. 2). When more than two variables are involved, correspondence models may involve a division of the categorical variables into two groups (Sect. 3). Correspondence models based on multiple correspondence analysis are much more difficult to construct (Sect. 4).

2. Bivariate Correspondence Models

Let X and Y be two categorical random variables with respective ranges consisting of I and J categories indexed from 1 to I, and from 1 to J, respectively. The joint probability distribution of X and Y may be characterized by the I by J matrix p^{XY} of joint probabilities $p_{ij}^{XY} = P(X = i, Y = j)$, $1 \leqslant i \leqslant I$, $1 \leqslant j \leqslant J$. The marginal probability distributions of X and Y may be characterized by the marginal probabilities $p_i^X = P(X = i)$, $1 \leqslant i \leqslant I$, and $p_j^Y = P(Y = j)$, $1 \leqslant j \leqslant J$. For simplicity, assume that $p_i^X > 0$ for $1 \leqslant i \leqslant I$ and $p_j^Y > 0$ for $1 \leqslant j \leqslant J$. The entire family of correspondence models for bivariate distributions is based on the canonical decomposition (Hirschfeld 1935, Fisher 1940)

$$p_{ij}^{XY} = p_i^X p_j^Y \left(1 + \sum_{m=1}^{M} \rho_{im} \mu_{im} \nu_{jm} \right) \tag{1}$$

where $M = \min(I-1, J-1)$,
$1 \geqslant \rho_1 \geqslant \rho_2 \geqslant \cdots \geqslant \rho_M \geqslant 0$

$$\sum_{i=1}^{I} p_i^X \mu_{im} = \sum_{j=1}^{J} p_j^Y \nu_{jm} = 0, 1 \leqslant m \leqslant M \tag{2}$$

and

$$\sum_{i=1}^{I} p_i^X \mu_{im} \mu_{im'} = \sum_{j=1}^{J} p_j^Y \nu_{jm} \nu_{jm'} = \begin{cases} 1, 1 \leqslant m = m' \leqslant M \\ 0, 1 \leqslant m \leqslant m' \leqslant M \end{cases} \tag{3}$$

In (1), (2), and (3), for each integer m from 1 to M, the parameter ρ_m is the mth canonical correlation, the score μ_{im} is the mth canonical score of category i of variable X, $1 \leqslant i \leqslant I$, and the score ν_{jm} is the mth canonical score of category j of variable Y. Let the mth canonical variable μ_{Xm} associated with X be the random variable with value μ_{im} for $X = i$, and let the mth canonical variable μ_{Ym} associated with Y denote the random variable with value ν_{jm} for $Y = j$. Then μ_{Xm} and ν_{Ym} both have mean 0 and variance 1, and ρ_m is the correlation of the canonical variables μ_{Xm} and ν_{Ym}. In addition, if $1 \leqslant m < m' \leqslant M$, then the canonical variables μ_{Xm} and $\mu_{Xm'}$ of X are uncorrelated, and the canonical variables ν_{Ym} and $\nu_{Ym'}$ of Y are uncorrelated.

For some M^* between 1 and M, graphical display of category i of X may be based on the observed scores μ_{im} for $1 \leqslant m \leqslant M^*$ and graphical display of category j of Y may be based on the observed scores ν_{jm} for $1 \leqslant m \leqslant M^*$. An alternative approach uses graphical displays for row category i based on the correspondence scores $\rho_m \mu_{im}$ for $1 \leqslant m \leqslant M^*$ and graphical displays for column category j based on the correspondence scores $\rho_m \nu_{jm}$ for $1 \leqslant m \leqslant M^*$.

Formal inference for canonical models has generally been based on maximum likelihood. Computation of maximum-likelihood estimates has been attempted by use of a cyclic algorithm (Goodman 1985) and by use of a scoring algorithm for parameters subject to constraints (Gilula and Haberman 1986). Goodness of fit has typically been examined by use of likelihood-ratio and Pearson chi-square statistics. In typical cases, a chi-square approximation can be applied when the model is true. If the model only assumes that $\rho_m = 0$

Table 1
Cross-classification of origin by principal worry

	Principal worry[a]							
Origin	ENR	SAB	MIL	POL	ECO	OTH	MTO	PER
A/A	61	70	97	32	4	81	20	104
E/A	104	117	218	118	11	128	42	48
fA/A	8	9	12	6	1	14	2	14
fE/A	22	24	28	28	2	52	6	16
fI/I	5	7	14	7	1	12	0	9

Source: Guttman (1971).
[a]Columns are: ENR, enlisted relative; SAB, sabotage; MIL, military situation; POL, political situation; ECO, economic situation; OTH, other; MTO, more than one worry; and PER, personal economics. Rows are: A/A, Africa or Asia; E/A, Europe or America; fA/A, father from Africa or Asia; fE/A, father from Europe or America; and fI/I, father and self from Israel.

Table 2
Estimated canonical parameters for model with two pairs of scores

Principal worry	Score 1	Score 2
ENR	−0.0573	−0.2380
SAB	−0.0096	−0.2880
MIL	−0.5121	−1.0690
POL	−1.2270	0.6685
ECO	−0.6866	0.0811
OTH	0.0692	1.7794
MTO	−0.5203	−1.1601
PER	2.4428	−0.2684
Origin	**Score 1**	**Score 2**
A/A	1.3472	−0.4105
E/A	−0.8663	−0.4678
fA/A	1.1632	0.7179
fE/A	−0.2830	2.6321
fI/I	0.4120	0.8064

for $m > M^*$, then the chi-square approximations have $(I-M^*-1)$ $(J-M^*-1)$ degrees of freedom.

As a statistical model, (1) is a *saturated* model, for no restrictions are imposed on the joint probabilities p_{ij}^{XY}. Goodman (1985, 1986) and Gilula and Haberman (1986) consider models in which the canonical correlations ρ_m are assumed 0 for $m > M^*$ for some M^* between 0 and M. The case of $M^* = 0$ leads to the conventional model of independence of the variables X and Y. The case of $M^* = M$ is the saturated model in which no restrictions are imposed. The most common nontrivial case has $M^* = 1$. Models are also considered in which linear restrictions are imposed on the scores μ_{im} and ν_{jm} for $m \leqslant M^*$.

As an example, consider the data in Table 1 (Guttman 1971) on the origin of Israeli adults versus their principal worry. Several cells in the table have small values, so some care must be exercised in interpretation of customary large-sample approximation. Nonetheless, it is worth noting that the model with a single nonzero canonical correlation ($\rho_m = 0$ for $m > 1$) yields a likelihood-ratio chi-square of 29.35 on 18 degrees of freedom, so that the fit is relatively poor. On the other hand, the model with two nonzero canonical correlations ($\rho_m = 0$ for $m > 2$) yields a likelihood-ratio chi-square of only 6.49 on 10 degrees of freedom, so that the fit is quite satisfactory. The estimated canonical correlations are 0.245 and 0.124. These results suggest a modest relationship between the variables under study. The maximum-likelihood estimates of the canonical scores are provided in Table 2. Fig. 1 provides correspondence scores for the column variable, and Fig. 2 provides correspondence

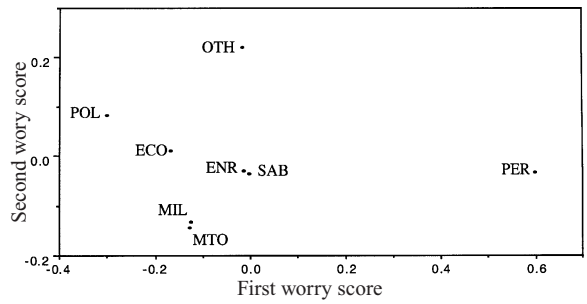

Figure 1
Correspondence analysis scores for worry

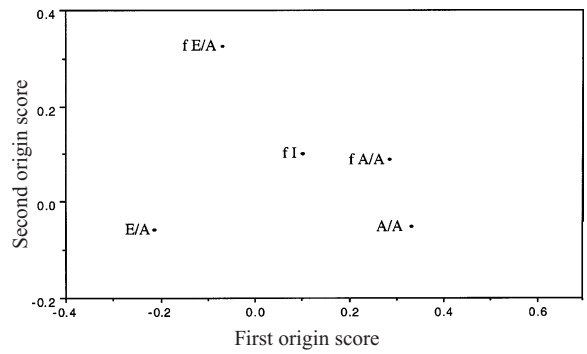

Figure 2
Correspondence analysis scores for origin

scores for the row variable. Results suggest a contrast between personal worries and other concerns, especially political ones, with subjects of Asian or African background more likely to express personal worries than subjects of European or American background. The second dimension of background appears to involve a contrast between subjects born in Israel with fathers of European or American background and other subjects. Subjects born in Israel with fathers from Europe or America are more likely to cite more than one worry than are other subjects.

3. Two Groups of Categorical Variables

Extension of correspondence models to apply to more than two categorical variables is a nontrivial matter. The simplest approach is appropriate if the variables can be divided into two groups (Gilula and Haberman 1988). In this case, each group of variables can be regarded as a categorical variable with many categories, and correspondence analysis methods for the bivariate case can be applied. For example, in addition to the categorical random variables X and Y of Sect. 2 let an additional categorical random variable Z be defined with values from 1 to $K \geqslant 2$. As an example, Gilula and Haberman (1988) consider an example with X equal to education level, Y equal to religion, and Z equal to attitude toward legal nontherapeutic abortions. Each variable in this example involves a classification into three categories, so that I, J, and K are 3. In general, let p_{ijk}^{XYZ} be the joint probability that $X = i$, $Y = j$, and $Z = k$. Then

$$p_{ijk}^{XYZ} = p_{ij}^{XY} p_k^Z \left(1 + \sum_{m=1}^{M} \rho_m' \mu_{ijm}' \nu_{km}' \right), \tag{4}$$

where M is the minimum of $IJ-1$ and $K-1$, the ρ_m' are nonnegative and nonincreasing in m,

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij}^{XY} \mu_{ijm}' = \sum_{k=1}^{K} p_k^Z \nu_{km}' = 0, \quad 1 \leqslant m \leqslant M, \tag{5}$$

and

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij}^{XY} \mu_{ijm}' ? \mu_{ijm'}' = \sum_{k=1}^{K} p_k^Z \nu_{km}' \nu_{km'}' = \begin{cases} 1, 1 \leqslant m = m' \leqslant M, \\ 0, 1 \leqslant m < m' \leqslant M. \end{cases} \tag{6}$$

Various restrictions may be imposed on the canonical parameters. For example, if Y and Z are conditionally independent given X, then $\rho_m = 0$ for any $m > (-1)$ and μ_{ijm}' depends only on i for $\rho_m > 0$. For other examples, see Gilula and Haberman (1988).

4. Multiple Correspondence Analysis

Multiple correspondence analysis, another approach to extension of correspondence analysis to the study of two or more categorical variables, appears in Guttman (1941). Multiple correspondence analysis can be regarded as a special case of correspondence analysis. For illustrative purposes, the case of two and three variables will be considered.

Consider the case of two categorical random variables X and Y defined as in Sect. 2. In a multiple correspondence analysis of X and Y, a canonical analysis is used to predict the joint variable (X, Y) from either X or Y, where X or Y is chosen at random. Formally, (X, Y) is predicted by the pair (U, V), where V is a random variable independent of X and Y such that V assumes the values 1 and 2 with equal probability. The variable U is X if V is 1, and U is Y if V is 2. when $V = 1$, then $U = X$, so that prediction of (X, Y) reduces to prediction of Y by X. When $V = 2$, then $U = Y$, so that prediction of (X, Y) reduces to prediction of X by Y. The canonical analysis in this case is closely related to conventional canonical analysis of the variables X and Y.

For multiple correspondence analysis of the three categorical variables X, Y, and Z of Sect. 3 the joint variable (X, Y, Z) is predicted by random choice of a marginal variable X, Y, or Z. In this case, let V be a random variable independent of X, Y, and Z which assumes value k with probability $\frac{1}{3}$ for $1 \leqslant k \leqslant 3$. Let U be the random variable which is X for $V = 1$, Y for $V = 2$, and Z for $V = 3$. Multiple correspondence analysis considers correspondence analysis of (X, Y, Z) and U. If $V = 1$, then $U = X$ and prediction of (X, Y, Z) reduces to prediction of (Y, Z) by X. If $V = 2$, then $U = Y$ and prediction of (X, Y, Z) reduces to prediction of (X, Z) by Y. If $V = 3$, then $U = Z$ and prediction of (X, Y, Z) reduces to prediction of (X, Y) by Z. The analysis differs from that in Sect. 3 in that it is symmetrical in the three variables. Thus in the example of Gilula and Haberman (1988), the analysis looks at educational level, religion, and abortion attitude symmetrically. This approach may or may not be appropriate, depending on the context. For instance, one may reasonably argue that abortion attitude reflects educational and religious background, but neither educational background nor religious background is appreciably affected by abortion attitude.

5. Conclusion

Correspondence models compete in practice with a number of other statistical methods in common use. In their development of canonical models, both Goodman (1985) and Gilula and Haberman (1986) compare canonical models with association models in the case of two variables. Here association models involve

a loglinear parameterization similar to the parameterization used in correspondence models. Interpretations involve loglinear interactions rather than the correlations used in correspondence models, but it remains the case that scores are attached to variable categories. As noted in both papers, association models in which scores are fixed rather than estimated are special cases of conventional log-linear models (see *Multivariate Analysis: Discrete Variables (Loglinear Models)*). Goodman (1985) also discusses the relationship of parameters in canonical models to parameters in correspondence models. Gilula and Haberman (1986) and van der Heijden, Gilula and van der Ark (1999) consider the relationship of canonical models to latent-class models (see *Factor Analysis and Latent Structure: Overview*). In the case in which more than two variables are analyzed by division of the variables into two groups, Gilula and Haberman (1988) compare canonical models to association models in cases in which more than two variables are present.

The canonical correlations used in bivariate correspondence analysis may be used to derive formal tests of independence that compete with conventional chi-square statistics and may be employed to ascertain the nature of departures from independence. For details, see Haberman (1981).

A variety of alternative interpretations of multiple correspondence analysis are available. For example, see Tenenhaus and Young (1985) and Gifi (1990).

See also: Factor Analysis and Latent Structure: Overview; Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Overview; Scaling: Correspondence Analysis

Bibliography

Benzécri J-P 1973 L'Analyse des Données: 2. *L'Analyse de Correspondances*. Dunod, Paris

Fisher R A 1940 The precision of discriminant functions. *Annals of Eugenics* **10**: 422–29

Gifi A 1990 *Non-linear Multivariate Analysis*. Wiley, New York

Gilula Z and Haberman S J 1986 Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association*. **81** 780–6

Gilula Z, Haberman S J 1988 The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association* **83**: 760–71

Goodman L A 1985 Analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics* **13**: 10–69

Goodman L A 1986 Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review* **54**: 243–309

Greenacre M J 1984 *Theory and Applications of Correspondence Analysis*. Academic Press, London

Guttman L 1941 The quantification of a class of attributes: a theory and method of scale construction. In: Horst P (ed.) *The Prediction of Personal Adjustment*. Social Science Research Council, New York, pp. 319–48

Guttman L 1971 Measurement as structural theory. *Psychometrika* **36**: 329–47

Haberman S J 1981 Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. *Annals of Statistics* **9**: 1178–86

Hill M O 1974 Correspondence analysis: a neglected multivariate method. *Applied Statistics* **23**: 340–54

Hirschfeld H O 1935 A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society* **31**: 520–4

Nishisato S 1980 *Analysis of Categorical Data: Dual Scaling and Its Applications*. University of Toronto Press

Tenenhaus M, Young F W 1995 An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* **50**: 91–119

Van der Heijden P G M, Gilula Z, van der Ark L A 1999 An extended study into the relationships between correspondence analysis and latent class analysis. In: Sobel M, Becker M (eds.) *Sociological Methodology*. Blackwell, Cambridge pp. 147–86

Z. Gilula, S. J. Haberman, and P. G. M. van der Heijden

Multivariate Analysis: Discrete Variables (Logistic Regression)

Logistic regression is used widely to examine and describe the relationship between a binary response variable (e.g., 'success' or 'failure') and a set of predictor variables. In common with linear regression (e.g., *Linear Hypothesis: Regression (Basics)*), the primary objective of logistic regression is to model the mean of the response variable, given a set of predictor variables. However, what distinguishes logistic regression from linear regression is that the response variable is binary rather than continuous in nature. This has a number of consequences for modeling the mean of the response variable. In this article the main features of logistic regression are described and some aspects of interpretation of logistic regression are illustrated with an example.

1. Introduction to Logistic Regression

In this section we introduce some notation and consider regression models for a binary response variable. Let Y denote a binary response variable, that is, a discrete response variable having only two categories, for convenience often referred to as 'success' or 'failure.' For example, Y might indicate an

individual's political party preference (Democrat or Republican, in the US), an individual's opinion on the death penalty (favor or oppose), or an individual's self-reported happiness (happy or unhappy). Denoting the two possible outcomes for Y by 0 and 1, the probability distribution of the response variable is the Bernoulli distribution, with $Pr(Y=1)=\pi$ (and, correspondingly, $Pr(Y=0)=1-\pi$). The primary goal of logistic regression is to describe the effect on π of changes in a set of predictor variables. For ease of exposition, we will first consider the simple case where there is only a single predictor variable, x. Generalizations to more than one predictor variable will be considered later in the article.

In logistic regression the main analytic goal is to investigate the relationship between π and x. Recall that with a binary response, the mean of Y is $\pi = Pr(Y=1)$ that is, the mean of Y is simply the probability that Y takes on the value 1. Since linear models (e.g., see *Linear hypothesis: Regression (Basics)*) play such an important and dominant role in applied statistics, it may at first seem natural to assume a linear model relating the mean of Y to x

$$E(Y|x) = \pi = \beta_0 + \beta_1 x \tag{1}$$

However, this linear model for the probabilities has one obvious difficulty. Expressing π as a linear function violates the restriction that probabilities must lie within the range from 0 to 1. That is, for sufficiently large or small values of x, the linear model given by Eqn. (1) will yield predicted probabilities outside of the range from 0 to 1. A further difficulty with the linear model for the probabilities is that we often expect there to be a nonlinear relationship between π and x. For example, a 0.2 unit increase in π might be considered more 'extreme' when $\pi = 0.1$ than when $\pi = 0.5$. In terms of ratios, the change from $\pi = 0.1$ to $\pi = 0.3$ represents a three-fold ($\frac{.3}{.1}$) or 300 percent increase, whereas the change from $\pi = 0.5$ to $\pi = 0.7$ represents only a 40 percent (or $\frac{.7}{.5}$) increase. In a sense, the units of measurement for a probability or proportion are often not considered to be constant over the range from 0 to 1. The linear probability model given by Eqn. (1) simply does not take this into consideration when relating π to x.

To circumvent some of the aforementioned difficulties with the linear probability model, a nonlinear transformation usually is applied to π and the transformed probabilities are related linearly to x. In particular, a transformation of π, say $g(\pi)$, is chosen so that it maps the range of π from (0,1) to $(-\infty, \infty)$. Since there are many possible transformations, $g(\pi)$, that achieve this goal, this leads to an extensive choice of models that are all of the form

$$g(\pi) = \beta_0 + \beta_1 x \tag{2}$$

However, the most commonly used in practice are

(a) logit or logistic function: $g(\pi) = \log[\pi/(1-\pi)]$;

(b) probit or inverse normal function: $g(\pi) = \Phi^{-1}(\pi)$, where Φ is the standardized normal cumulative distribution function; and

(c) complementary log–log function: $g(\pi) = \log[-\log(1-\pi)]$.

We note that all of these transformations are very closely related when $0.2 < \pi < 0.8$, and in a sense only differ in the degree of 'tail-stretching' outside of this range. Indeed, for most practical purposes it is not possible to discriminate between a data analysis that is based on, for example, the logit and probit functions. To discriminate empirically between probit and logistic regression would, in general, require very large numbers of observations. However, the logit function does have a number of distinct advantages over the probit and complementary log–log functions that probably account for its more widespread use in practice. Later in this article we will consider some of the advantages of the logit or logistic function. Finally, we remark that the logit function is the so-called natural or canonical link function, that is, the logit is a special function that can be derived from the Bernoulli distribution, expressed in exponential family form. Although the choice of canonical function can result in some minor mathematical simplifications, in general the selection of a suitable function should primarily be made on scientific grounds, that is, based on the goodness-of-fit of the resulting model and the ease of interpretation of the model parameters.

When the logit or logistic function is adopted, the resulting model

$$\log[\pi/(1-\pi)] = \beta_0 + \beta_1 x \tag{3}$$

is known as the logistic regression model. Recall that if π is the probability of success, then $(\pi/1-\pi)$ is known as the odds of success; the latter term will be familiar to readers belonging to the gambling fraternity. For example, if the probability of success is 0.8 then the odds of success is 4 (or $\frac{0.8}{0.2}$) to 1, that is, the probability of success is 4 times as large as the probability of failure. Thus, the logistic regression model assumes a linear relationship between the log odds (of success) and x. Note that this simple model can be expressed equivalently in terms of π

$$\pi = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{4}$$

While the latter expression may appear to be somewhat more daunting, we simply wish to emphasize that Eqns. (3) and (4) are completely equivalent ways of expressing the logistic regression model. Equation (3) describes how the log odds, $\log(\pi/1-\pi)$, has a linear or straight-line relationship with x. Equation (4), on the other hand, describes how π has an S-shaped relationship with increasing values of $\beta_1 x$; although, in

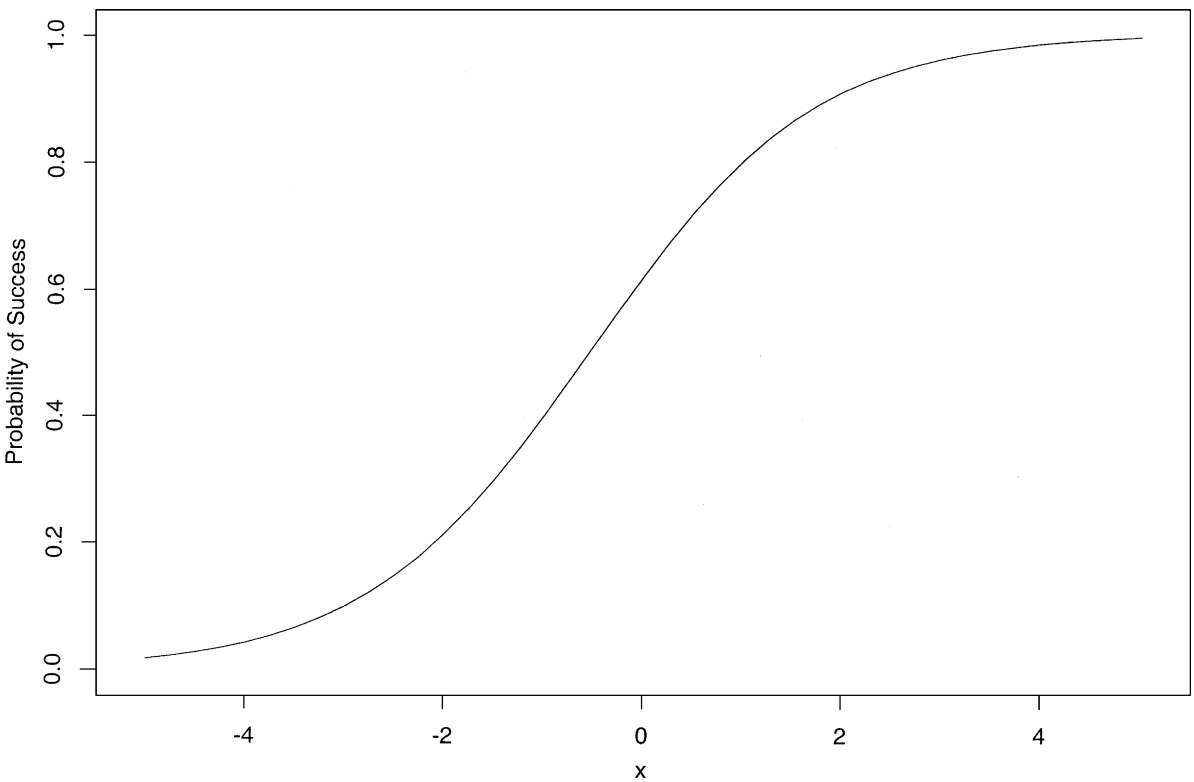

Figure 1
Plot of logistic response function

general, this relationship is approximately linear within the range $0.2 < \pi < 0.8$ (see Fig. 1 for a plot of π vs. x when $\beta_0 = 0.5$ and $\beta_1 = 0.9$). Observe that the expression on the right of Eqn. (4) cannot yield a value that is either negative or greater than 1, that is, the logistic transformation ensures that the predicted probabilities are restricted to the range from 0 and 1. Finally, note that

$$1-\pi = \frac{1}{1+\exp(\beta_0+\beta_1 x)} \tag{5}$$

so that the odds, $(\pi/1-\pi)$, is simply $\exp(\beta_0+\beta_1 x)$, and the log odds can be expressed as $l = l(x) = \log(\pi/1-\pi) = \beta_0+\beta_1 x$.

2. *Interpretation of Logistic Regression Coefficients*

Next we consider the interpretation of the logistic regression coefficients, β_0 and β_1, in Eqn. (3). In simple linear regression, recall that the interpretation of the slope of the regression is in terms of changes in the mean of Y for a single unit change in x. Similarly, the logistic regression slope, β_1, in Eqn. (3) has interpretation as the change in the log odds (of success) for a unit change in x. Equivalently, a unit change in x increases or decreases the odds of success multiplicatively by a factor of $\exp(\beta_1)$. Also, recall that the intercept in simple linear regression has interpretation as the mean value of the response variable when x is equal to zero. Similarly, the logistic regression intercept, β_0, in Eqn. (3) has interpretation as the log odds (of success) when $x = 0$; alternatively, if we exploit the expression on the right of Eqn. (4), $[\exp(\beta_0)/1+\exp(\beta_0)]$ is the probability of success when $x = 0$.

For the special case where the predictor variable x is dichotomous, taking values of 0 and 1, the logistic regression slope, β_1, has a simple and very attractive interpretation. Consider the two possible values for π when $x=0$ and $x = 1$. Let $\pi(x = 0)$ denote the probability of success when $x = 0$, that is,

$$\pi(x = 0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \tag{6}$$

(obtained by plugging-in or substituting the value 0 for x in the expression on the right of Eqn. (4)). Similarly,

		Y		
		1	0	
x	1	$\pi(x=1) = \frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)}$	$1-\pi(x=1) = \frac{1}{1+exp(\beta_0+\beta_1)}$	1.0
	0	$\pi(x=0) = \frac{exp(\beta_0)}{1+exp(\beta_0)}$	$1-\pi(x=0) = \frac{1}{1+exp(\beta_0)}$	1.0

Figure 2
Cross-classification probabilities for logistic regression of Y on x

let $\pi(x = 1)$ denote the probability of success when $x = 1$, that is

$$\pi(x = 1) = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} \tag{7}$$

(obtained by plugging in the value 1 for x in the expression on the right Eqn. of (4)). Then, we can construct a 2×2 table of the cross-classification of Y and x (see Fig. 2). Note that the log of the odds ratio (or cross-product ratio) in this 2×2 table is equal to β_1, that is,

$$\begin{aligned}\log(\text{odds ratio}) &= \log\left\{\frac{\pi(x=1)\times[1-\pi(x=0)]}{\pi(x=0)\times[1-\pi(x=1)]}\right\}\\ &= l(x=1)-l(x=0)\\ &= (\beta_0+\beta_1)-\beta_0 = \beta_1\end{aligned} \tag{8}$$

Thus, $\exp(\beta_1)$ has interpretation as the odds ratio of the response for the two possible values of the predictor variable.

The odds ratio has many appealing properties that probably account for the widespread adoption of the logistic function (and, hence, logistic regression) over other functions such as the probit or complementary log–log functions. First, note that the odds ratio does not change when rows and columns of the 2×2 table are interchanged. This implies that it is not necessary to distinguish which variable is the response and predictor variable in order to estimate the odds ratio. Furthermore, a very appealing feature of logistic regression is that the odds ratio, $\exp(\beta_1)$, is equally valid whether the study design is prospective, cross-sectional, or retrospective. That is, logistic regression provides an estimate of the same association between Y and x regardless of whether we (a) randomly sample individuals having $x = 0$ and $x = 1$ and subsequently classify them according to whether $Y = 1$ or $Y = 0$; (b) randomly sample individuals from a population and cross-classify them according to whether $x = 0$ or $x = 1$ and whether $Y = 1$ or $Y = 0$; (c) randomly sample individuals having $Y = 0$ and $Y = 1$ and subsequently classify them according to whether $x = 1$ or $x = 0$. In all three study designs logistic regression provides an estimate of the same association between Y and x.

Finally, in medical studies where Y typically denotes the presence or absence of a disease or disorder, the odds ratio is often interpreted as an approximation to the relative risk of disease. The relative risk of disease is simply the ratio of the probability of disease ($Y = 1$) when a risk factor is present ($x = 1$) to the probability of disease when a risk factor is absent ($x = 0$), that is, the relative risk of disease is simply $[\pi(x = 1)/\pi(x = 0)]$. When the disease is rare, and π is reasonably close to 0 in both of the risk groups (often known as the 'rare disease' assumption), the odds ratio provides a close approximation to the relative risk. Retrospective designs (case (c) above) are especially common in health and medical studies where the possible outcomes of interest are very rare. Although the relative risk cannot be estimated from a retrospective study, the odds ratio can be used to provide an approximation to the relative risk.

3. *Multiple Logistic Regression*

So far, we have only considered the simple case where there is a single predictor x. Next, we consider the extensions of Eqns. (3) and (4) to the case where there are two or more predictors. When there are many predictor variables, the logistic regression model becomes

$$\log[\pi/(1-\pi)] = \beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p \tag{9}$$

where $x_1, x_2,\ldots, x_p$ are the outcomes of p predictor variables. Note that $x_1, x_2,\ldots,x_p$ may represent p

distinct predictor variables; alternatively, they may represent q distinct predictor variables (where $q < p$) in addition to a subset of interactions among the predictors. The logistic regression coefficients in Eqn. (9) have the following interpretations. The logistic regression intercept, β_0 , now has interpretation as the log odds (of success) when all predictor variables assume the value zero, that is, when $x_1 = x_2 = \ldots = x_p = 0$. Alternatively, $[\exp(\beta_0)/1+\exp(\beta_0)]$ is the probability of success when $x_1 = x_2 = \ldots = x_p = 0$. Each of the logistic regression slopes, β_j (for $j = 1,\ldots, p$), has interpretation as the change in the log odds (of success) for a unit change in x_j given that all of the other predictor variables remain constant. We note that this is completely analogous to the interpretation of the regression coefficients in multiple linear regression. Thus, holding the remaining predictors at some fixed set of values and not allowing them to vary with any changes in x_j a single unit increase in x_j is predicted to increase or decrease the log odds of success by an amount β_j. Equivalently, a single unit increase in x_j increases or decreases the odds of success multiplicatively by a factor of $\exp(\beta_j)$. Finally, we remark that the multiple logistic regression model given by Eqn. (9) can also be expressed equivalently in terms of π

$$\pi = \frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p)}. \quad (10)$$

In Sect. 2 we remarked that one of the appealing properties of logistic regression is that the odds ratio can be estimated from either a prospective or retrospective study design. This result readily generalizes when x is quantitative rather than dichotomous, and also when there are two or more predictor variables. Thus, in a retrospective study $\beta_1,\ldots, \beta_p$ can be estimated by analyzing the data as if they had been obtained from a prospective study. That is, the data from a retrospective study can be treated as prospective data in a logistic regression analysis where the binary outcome of interest is modeled as the response variable. A formal proof of this result can be found in Farewell (1979) and Prentice and Pyke (1979). The only parameter that cannot be estimated with data from a retrospective study is the intercept, β_0, since it is determined by the proportions of 'successes' ($Y = 1$) and 'failures' ($Y = 0$) selected into the study. In many studies, however, the intercept is not of direct scientific interest.

4. *Estimation*

In this section we discuss very briefly estimation of the logistic regression coefficients. This section is somewhat more technical in nature than the previous sections and can been skipped-over on a first reading of this article. Recall that in linear regression we estimate the linear regression coefficients using the method of least squares. That is, the least squares criterion chooses values for the linear regression coefficients that minimize the sum of squared deviations of the observed Y from the predicted values of Y, say $\hat{Y}$, under the assumed regression model. The least squares method yields estimates of the regression coefficients that are also the maximum likelihood (ML) estimates when the errors have a normal distribution with constant variance. We use the more general method of maximum likelihood estimation (e.g., *Likelihood in Statistics*) for logistic regression.

The essential idea behind maximum likelihood is quite simple, and in fact is conveyed by its name. The method chooses values of the logistic regression coefficients that are most likely (or most probable) to have generated the observed data. This is achieved by maximizing the likelihood function for the data. The likelihood function is simply the probability of the response variables evaluated at the observed values, regarded as a function of the unknown logistic regression coefficients. Construction of the likelihood function requires that we must make an assumption about the probability distribution of Y. Fortunately, with a binary response the selection of a probability distribution is straightforward as there is only a single distribution to choose from, namely the Bernoulli distribution.

In the following we assume that there are independent observations of the response variables Y_i and the predictor variables $x_{i1}, x_{i2},\ldots, x_{ip}$ available on N individuals. The N binary responses, $Y_i(i = 1,\ldots, N)$, can be regarded as independent Bernoulli random variables, with $\pi_i = E(Y_i)= Pr(Y_i = 1)$. Since the N binary responses are independent the joint probability of $(Y_1,\ldots, Y_N)$ is the product of the N Bernoulli functions. Thus, the likelihood function can be expressed as the product

$$\prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad (11)$$

It is this function, or equivalently, the logarithm of the likelihood function, that must be maximized. Note that the likelihood is a function of the unknown logistic regression coefficients, since

$$\pi_i = \frac{\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip})}{1+\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip})} \quad (12)$$

The maximum likelihood estimates of the logistic regression coefficients are obtained by substituting the above expression for π_i into the likelihood function and finding the values of the logistic regression coefficients that produce the largest value for the likelihood function. Ordinarily, the likelihood function has only

a single maximum. Maximization of the likelihood, or the log likelihood, requires an iterative procedure and this has been implemented in many of the commercially-available statistical software packages (e.g., SAS and SPSS). A description of the iterative procedure can be found in, for example, McCullagh and Nelder (1989). Finally, estimates of the standard errors of the estimated logistic regression coefficients can readily be obtained using the method of maximum likelihood estimation.

5. *Example*

Next we consider an application of logistic regression to illustrate how the model can be used in practice. Table 1, adapted from Clogg (1982), contains data from 1518 individuals who participated in the US General Social Survey (Davis 1977). The data comprise a pair of 2×4 cross-classifications on the relationship between reported happiness (not too happy vs. pretty or very happy) and completed years of schooling (less than 12, 12, 13–16, 17 or more), by the gender of the respondent. The goal of the analysis is to describe how reported happiness is related to years of schooling and the gender of the respondent. In particular, it is of interest to determine whether reported happiness is related to the amount of schooling received and whether this relationship is different for males and females.

We first consider a logistic regression model that includes effects of years of schooling, gender and their interaction (model 1):

$$\log[\pi/(1-\pi)] = \beta_0 + \beta_1 G + \beta_2 YSC_1 + \beta_3 YSC_2 + \beta_4 YSC_3 + \beta_5 G \times YSC_1 + \beta_6 G \times YSC_2 + \beta_7 G \times YSC_3 \quad (13)$$

where $G = 1$ if female and $G = 0$ if male, and YSC_1, YSC_2, and YSC_3 are three indicator variables for the 4 levels of years of schooling. That is, $YSC_1 = 1$ if years of schooling < 12 and $YSC_1 = 0$ otherwise; $YSC_2 = 1$ if years of schooling is 12 and $YSC_2 = 0$ otherwise; $YSC_3 = 1$ if years of schooling is 13–16 and $YSC_3 = 0$ otherwise. Maximum likelihood estimates of the regression coefficients for model 1 are given at the top of Table 2.

Table 1
Cross-classification of US sample according to reported happiness and years of schooling, by gender

		Years of school completed			
Gender	Reported happiness	< 12	12	13–16	⩾ 17
Males	Not too happy	40	21	14	3
	Pretty or very happy	213	177	167	54
Females	Not too happy	62	26	12	3
	Pretty or very happy	242	283	171	30

Table 2
Logistic regression parameter estimates and standard errors under three models for the probability of reporting happiness

Model	Parameter	Estimate	S.E.	Z	P
Model 1	Intercept	2.8904	0.5932	4.87	<.0001
	G	−0.5878	0.8477	−0.69	0.4880
	YSC_1	−1.2180	0.6177	−1.97	0.0486
	YSC_2	−0.7587	0.6365	−1.19	0.2332
	YSC_3	−0.4114	0.6552	−0.63	0.5300
	$G \times YSC_1$	0.2772	0.8766	0.32	0.7519
	$G \times YSC_2$	0.8435	0.9021	0.94	0.3498
	$G \times YSC_3$	0.7656	0.9408	0.81	0.4158
Model 2	Intercept	2.6728	0.4272	6.26	<.0001
	G	−0.0898	0.1625	−0.55	0.5805
	YSC_1	−1.1278	0.4375	−2.58	0.0099
	YSC_2	−0.3362	0.4512	−0.74	0.4562
	YSC_3	−0.0618	0.4696	−0.13	0.8952
Model 3	Intercept	2.4732	0.1491	16.59	<.0001
	G	−0.1085	0.1616	−0.67	0.5022
	YSC_1	−0.9178	0.1606	−5.71	<.0001

Note that since all of the predictor variables are discrete, this is a saturated model and will perfectly fit or predict the data in Table 1. Examination of Table 2 reveals that the three estimated coefficients for the interaction of gender and years of schooling are relatively small when compared with their respective standard errors. A formal test of the interaction can be constructed by fitting a logistic regression model that omits the interaction and comparing the likelihoods for the two models. Next, we consider a logistic regression model that includes only the effects of years of schooling and gender (model 2):

$$\log[\pi/(1-\pi)] = \beta_0 + \beta_1 G + \beta_2 YSC_1 + \beta_3 YSC_2 + \beta_4 YSC_3 \quad (14)$$

Maximum likelihood estimates of the logistic regression coefficients in model 2 are given in the middle of Table 2. Comparing model 2 to model 1 in terms of their likelihoods, we can formally test the null hypothesis of no interaction, H_0: $\beta_5 = \beta_6 = \beta_7 = 0$. The likelihood-ratio chi-squared statistic is 2.99 on 3 degrees of freedom ($p \approx 0.4$). Thus, we cannot reject H_0: $\beta_5 = \beta_6 = \beta_7 = 0$, and model 2, without an interaction between gender and years of schooling, is defensible. This implies that the relationship between reported happiness and years of schooling is the same for males and females. Examination of the estimated regression coefficients for years of schooling in model 2 suggests that a further simplification of the model might be possible. The effect of years of schooling on reported happiness appears to be greatest for those individuals with less than 12 years vs. those having 12 or more years of education. Thus, we next consider the following logistic regression model (model 3):

$$\log[\pi/(1-\pi)] = \beta_0 + \beta_1 G + \beta_2 YSC_1 \quad (15)$$

Note that in this model we have combined the levels of years of schooling completed to form a dichotomy: less than 12 years vs. 12 or more years. Maximum likelihood estimates of the regression coefficients in model 3 are given at the bottom of Table 2. Comparing model 3 to model 2 in terms of their likelihoods provides a formal test of the null hypotheses, H_0: $\beta_3 = \beta_4 = 0$. The likelihood-ratio chi-squared statistic is 1.46 on 2 degrees of freedom ($p \approx 0.5$) and thus the null hypothesis cannot be rejected. The estimated logistic regression coefficients for model 3 indicate that the overall reported rates of happiness are very similar for males and females but that the rates of reported happiness are substantially lower for those with less than 12 years of education. In particular, those with more than 12 years of schooling are 2.5 (or exp[0.9178]) times more likely to report being happy (rather than unhappy) when compared to those with less than 12 years of schooling. That is, the ratio of odds of reported happiness is 2.5 when those with more than 12 years of schooling are compared to those with less than 12 years of schooling. Finally, the nature of this relationship between years of schooling and happiness is the same for males and females.

6. *Extensions of Logistic Regression to Clustered Data*

So far, we have considered logistic regression for a single binary response. However, multivariate binary response data commonly arise in a number of applications in the social and behavioral sciences. That is, two or more measurements of the response are often obtained in a block or cluster and the binary responses within a cluster are expected to be positively correlated. When this occurs the responses from any pair of members of the same cluster are expected to be more closely related than the responses from a pair belonging to different clusters. Some common examples where data arise in clusters include repeated measures or longitudinal studies and studies on families, communities, or other naturally occurring groups. The important aspect of all these studies is that the binary responses within a cluster cannot be regarded as independent of one another.

There may be many reasons for the correlation among cluster members. For example, when the cluster comprises all the siblings within a family the correlation among siblings may be due to shared (or at least similar) genetic, environmental, and social conditions. In a longitudinal study, where the responses within a cluster represent measurements taken at different occasions, the binary responses are expected to be positively correlated simply because they have been obtained on the same individual (or cluster). Whatever the underlying reasons for the correlation, failure to account for it in the analysis can lead to invalid inferences. That is, the standard application of logistic regression in this setting is no longer appropriate.

There are two general approaches for handling the analysis of clustered binary data. The first is to consider models for the joint distribution of the cluster of binary responses that explicitly account for the within-cluster correlation. There is an extensive statistical literature on this topic and the interested reader in referred to a review article by Pendergast et al. (1996). For the most part, these models can be demanding computationally and few have been implemented in commercially-available statistical software. An alternative approach is simply to ignore the correlation among members of a cluster. That is, the analysis proceeds naively as though the binary responses within a cluster can be regarded as independent observations, but later a correction is applied to ensure that valid standard errors are obtained. Note that in this 'naive' approach that ignores the within-cluster

correlation the estimated logistic regression coefficients are valid, but their nominal standard errors are not. However, valid standard errors can be readily obtained using the well known empirical variance estimator, first proposed by Huber (1967). Thus, the analysis proceeds in two stages. In the first stage, the correlation among binary responses within a cluster is simply ignored and standard logistic regression is used to obtain estimates of the logistic regression coefficients. In the second stage, valid standard errors for the estimated logistic regression coefficients are obtained using an alternative, but widely implemented, variance estimator that properly accounts for the correlation among the binary responses. The chief advantage of this approach is that it can be implemented readily using standard statistical software for logistic regression.

7. *Conclusion*

In this article we have described the salient features of logistic regression and considered various aspects of interpretation of logistic regression coefficients. We remarked that the many appealing properties of the odds ratio probably account for the widespread adoption of the logistic function, and hence logistic regression, over alternative functions such as the probit or complementary log-log functions. A more detailed discussion of some of the issues presented here can be found in the textbooks by Agresti (1996), Kleinbaum (1994), and Hosmer & Lemeshow (1989).

We conclude this article by noting that logistic regression is a special case of a broad and unified class of statistical models known as generalized linear models. Other familiar examples of generalized linear models include linear regression and analysis of variance (ANOVA) models for a continuous response, log-linear models for count data, and probit regression for a binary response. Readers wishing to know more about this connection are referred to the textbooks by Dobson (1990) and McCullagh & Nelder (1989).

See also: Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Linear Hypothesis: Regression (Quantile); Multivariate Analysis: Discrete Variables (Loglinear Models)

Bibliography

Agresti A 1996 *An Introduction to Categorical Data Analysis*. Wiley, New York

Clogg C C 1982 Some models for the analysis of association in multiway cross-classifications having ordered categories. *Journal of the American Statistical Association* **77**: 803–15

Davis J A 1977 *Codebook for the 1977 General Social Survey*. National Opinion Research Center, Chicago

Dobson A J 1990 *An Introduction to Generalized Linear Models*. Chapman and Hall, New York

Farewell V T 1979 Some results on the estimation of logistic models based on retrospective data. *Biometrika* **66**: 27–32

Hosmer D W, Lemeshow S 1989 *Applied Logistic Regression*. Wiley, New York

Huber P J 1967 The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA, Vol. 1, pp. 221–33

Kleinbaum D G 1994 *Logistic Regression: A Self-learning Text*. Springer-Verlag, New York

McCullagh P, Nelder J A 1989 *Generalized Linear Models*. Chapman and Hall, New York

Pendergast J F, Gange S J, Newton M A, Lindstrom M J, Palta M, Fisher M R 1996 A survey of methods for analyzing clustered binary response data. *International Statistical Review* **64**: 89–118

Prentice R L, Pyke R 1979 Logistic disease incidence models and case-control studies. *Biometrika* **66**: 403–12

G. M. Fitzmaurice and N. M. Laird

Multivariate Analysis: Discrete Variables (Loglinear Models)

Loglinear models are a class of probability models commonly used to analyze relationships involving discrete variables. They include models for independence and conditional independence of polytomous variables, logit models for prediction of binary variables, multinomial response models for prediction of multinomial variables, and Poisson regression models. They are used both in contingency tables in which all variables under study are categorical, and in cases in which responses are categorical variables but some predicting variables may be continuous. Although numerous probability models in the statistical literature prior to 1969 are examples of loglinear models, an explicit description of a probability model as a loglinear model is not readily found in statistical journals or published books prior to Bishop (1969) and Bishop and Fienberg (1969).

The cell frequencies are random variables with positive and finite expected values, and the logarithms of the expected values of the cell frequencies satisfy a linear model. The frequencies may be independent Poisson random variables (Poisson sampling) or they may have a joint multinomial distribution with known sample size (multinomial sampling). In the case of multinomial sampling, to each cell in the table of frequencies corresponds a cell probability. The linear model for the logarithms of the expected values of the cell frequencies is defined in such a manner that the logarithms of the cell probabilities satisfy the same linear model. In a more complex case (product multinomial sampling), the table of frequencies consists of independently distributed subtables, with each

subtable having an independent multinomial distribution with known sample size. Again, each cell in the table of frequencies corresponds to a cell probability for a multinomial distribution, and the linear model is defined so that the logarithms of the cell probabilities satisfy the same linear model as do the logarithms of the expected values of the cell frequencies.

1. Examples

Several examples of loglinear models provide an indication of their scope.

Example 1. The additive loglinear model for a two-way contingency table. As in Bishop and Fienberg (1969), consider a contingency table with frequencies n_{ij}, $1 \leqslant i \leqslant I$, and $1 \leqslant j \leqslant J$, where there are $I \geqslant 2$ rows and $J \geqslant 2$ columns in the table. Let each frequency n_{ij} be a random variable with expectation $m_{ij} > 0$. In an additive loglinear model, it is assumed that the logarithms $\mu_{ij} = \log m_{ij}$ satisfy the additive model

$$\mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B \tag{1}$$

for $1 \leqslant i \leqslant I$ and $1 \leqslant j \leqslant J$, where λ, λ_i^A, and λ_j^B are unknown parameters such that

$$\sum_{i=1}^{I} \lambda_i^A = \sum_{j=1}^{J} \lambda_j^B = 0 \tag{2}$$

Interpretation of this model depends on the sampling assumptions concerning the frequencies. In the multinomial case, the n_{ij} have a joint multinomial distribution with known sample size $N > 0$ and unknown probabilities $p_{ij} > 0$. Thus $m_{ij} = Np_{ij}$ and

$$\log p_{ij} = \kappa + \lambda_i^A + \lambda_j^B \tag{3}$$

where $\kappa = \lambda - \log N$. Thus, the logarithms $\log m_{ij}$ of the cell frequencies satisfy the same additive loglinear model as the logarithms $\log p_{ij}$ of the cell probabilities. The probabilities p_{ij} define the joint probability distribution of a row variable A and a column variable B such that $1 \leqslant i \leqslant I$, $1 \leqslant j \leqslant J$, and p_{ij} is the probability that $A = i$ and $B = j$. The additive model is then equivalent to the model that A and B are independent, so that $p_{ij} = p_i^A p_j^B$, where p_i^A is the marginal probability that $A = i$ and p_j^B is the marginal probability that $B = j$.

In the row-multinomial case, the I rows n_{ij}, $1 \leqslant j \leqslant J$, are mutually independent for $1 \leqslant i \leqslant I$, and the ith row has a multinomial distribution with known sample size $N_i^A > 0$ and probabilities $p_{j \cdot i}^{B \cdot A} > 0$. The mean $m_{ij} = N_i^A p_{j \cdot i}^{B \cdot A}$, so that

$$\log p_{j \cdot i}^{B \cdot A} = \kappa + \kappa_i^A + \lambda_j^B \tag{4}$$

where

$$\kappa = \lambda - \eta,\ \eta = I^{-1} \sum_{i=1}^{I} \log N_i^A \tag{5}$$

and

$$\kappa_i^A = \lambda_i^A - \log N_i^A + \eta \tag{6}$$

Once again, the logarithms $\log P_{j \cdot}^{B \cdot A} i$ of the probabilities satisfy the same additive linear model as the logarithms $\log m_{ij}$ of the expected values of the cell frequencies. The additive model is equivalent to the homogeneity model that, for any fixed column j, $p_{j \cdot i}^{B \cdot A}$ is constant for each row i.

In the column-multinomial case, J columns n_{ij}, $1 \leqslant i \leqslant I$, are mutually independent for $1 \leqslant i \leqslant I$, and the jth column has a multinomial distribution with sample size $N_j^B > 0$ and probabilities $p_{i \cdot j}^{A \cdot B} > 0$. The mean $m_{ij} = N_j^B p_{i \cdot j}^{A \cdot B}$, so that

$$\log p_{i \cdot j}^{A \cdot B} = \kappa + \lambda_i^A + \kappa_j^B \tag{7}$$

where

$$\kappa = \lambda - \eta, \quad \eta = J^{-1} \sum_{j=1}^{J} \log N_j^B \tag{8}$$

and

$$\kappa_j^B = \lambda_j^B - \log N_j^B + \eta \tag{9}$$

The additive model is equivalent to the homogeneity model that, for any fixed row i, $p_{i \cdot j}^{A \cdot B}$ is constant for each column j.

Rasch (1960) considers an additive loglinear model for a contingency table in which each frequency n_{ij} has an independent Poisson distribution with mean m_{ij}. This model is applied to analysis of a reading test in which I subjects read J passages, and n_{ij} is the number of errors made by subject i while reading passage j. The parameter—λ_i^A provides a measure of relative subject ability, and λ_j^B measures relative item difficulty. This model is a special case of a general class of models encountered in the study of item responses (see *Factor Analysis and Latent Structure: IRT and Rasch Models*).

The additive loglinear model leads to the quasi-independence model of Goodman (1968) and Bishop and Fienberg (1969) when (1) is only required to hold for pairs (i, j) in a subset k of the pairs (i, j) such that i and j are integers, $1 \leqslant i \leqslant I$, and $1 \leqslant j \leqslant J$.

Example 2. Logit models. As in Bishop (1969) and Haberman (1974, Chap. 8), a logit model (see *Multivariate Analysis: Discrete Variables (Logistic Regression)*) may be described by a frequency table with elements n_{ij}, $1 \leqslant i \leqslant 2$, and $1 \leqslant j \leqslant J$. Covariates X_{jk},

$1 \leqslant k \leqslant K$, correspond to each column j. Each column n_{ij}, $1 \leqslant i \leqslant 2$, has an independent multinomial distribution with known sample size N_j^B and unknown positive probabilities $p_{i \cdot j}^{A \cdot B}$. It is assumed that the expected values $m_{ij} = N_j^B p_{i \cdot j}^{A \cdot B}$ have respective logarithms $\mu_{ij} = \log m_{ij}$ which satisfy the linear model

$$\mu_{1j} - \mu_{2j} = \log p_{i \cdot j}^{A \cdot B} - \log p_{2 \cdot j}^{A \cdot B} = \eta + \sum_{k=1}^{K} \beta_k X_{jk} \quad (10)$$

for some unknown parameters η and β_k. Thus the logarithms $\log {}_{j \cdot i}^{B \cdot A}$ satisfy the same linear model as do the logarithms $\log m_{ij}$ of the expected cell frequencies. Equivalently, if δ is a Kronecker function, then

$$\mu_{ij} = \alpha_j + \eta \delta_{i1} + \sum_{k=1}^{K} \beta_k X_{jk} \delta_{i1} \quad (11)$$

for unknown parameters α_j, η, and β_k. Thus

$$\log p_{i \cdot j}^{A \cdot B} = \gamma_j + \eta \delta_{i1} + \sum_{k=1}^{K} \beta_k X_{jk} \delta_{i1} \quad (12)$$

where

$$\gamma_j = \alpha_j - \log N_j^B \quad (13)$$

Example 3. Interactions in a three-way table. As in Birch (1963), consider a three-way table of frequencies n_{ijk}, $1 \leqslant i \leqslant I$, $1 \leqslant j \leqslant J$, $1 \leqslant k \leqslant K$, with respective positive expectations m_{ijk}. Assume that I, J, and K are at least 2. The logarithm $\mu_{ijk} = \log m_{ijk}$ can always be written as

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \quad (14)$$

where the unknown parameters λ, λ_i^A, λ_j^B, λ_k^C, λ_{ij}^{AB}, λ_{ik}^{AC}, λ_{jk}^{BC}, and λ_{ijk}^{ABC} satisfy the constraints

$$\sum_{i=1}^{I} \lambda_i^A = \sum_{j=1}^{J} \lambda_j^B = \sum_{k=1}^{K} \lambda_k^C = \sum_{i=1}^{I} \lambda_{ij}^{AB} = \sum_{j=1}^{J} \lambda_{ij}^{AB} = \sum_{i=1}^{I} \lambda_{ik}^{AC}$$
$$= \sum_{k=1}^{K} \lambda_{ik}^{AC} = \sum_{j=1}^{J} \lambda_{jk}^{BC} = \sum_{k=1}^{K} \lambda_{jk}^{BC} = \sum_{i=1}^{I} \lambda_{ijk}^{ABC}$$
$$= \sum_{j=1}^{J} \lambda_{ijk}^{ABC} = \sum_{k=1}^{K} \lambda_{ijk}^{ABC} = 0 \quad (15)$$

As in the analysis of variance, the parameters λ_i^A, λ_j^B, and λ_k^C may be termed main effects, the parameters λ_{ij}^{AB}, λ_{ik}^{AC}, and λ_{jk}^{BC} may be termed two-factor interactions, and λ_{ijk}^{ABC} is a three-factor interaction. In the simplest case, the table is a cross-classification obtained from a random sample of size $N \geqslant 1$ in which the categorical random variables A, B, and C are observed. Here A has values from 1 to I, B has values from 1 to J, and C has values from 1 to K. The count n_{ijk} is the number of observations with $A = i$, $B = j$, and $C = k$. The probability that $A = i$, $B = j$, and $C = k$ is p_{ijk}, so that $m_{ijk} = Np_{ijk}$ and

$$\log p_{ijk} = \kappa + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

where $\kappa = \lambda - \log N$. If all two-factor and three-factor interaction parameters are zero, so that the logarithms μ_{ijk} satisfy the additive model

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

then A, B, and C are mutually independent. Note that the logarithms $\log p_{ijk}$ also satisfy the additive model

$$\log p_{ijk} = \kappa + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

If λ_{ik}^{AC}, λ_{jk}^{BC}, and λ_{ijk}^{ABC}, are zero, so that

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}$$

then C is independent of the joint variable (A, B). If λ_{jk}^{BC} and λ_{ijk}^{ABC} are zero, so that

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC}$$

then B and C are conditionally independent given A. In a model of no-three-factor interaction, it is assumed that the three-factor interaction terms λ_{ijk}^{ABC} are all 0, so that

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} \quad (16)$$

Models of interactions in three-way tables can be applied with other sampling schemes. For example, in the Poisson case, each frequency n_{ijk} has an independent Poisson distribution with mean $m_{ijk} > 0$.

Prior to Birch (1963), the model of no-three-factor interaction can be encountered with somewhat different parameterizations. The models for interactions considered here are examples of hierarchical log-linear models. In these models, the restriction is imposed that higher-order interactions are only included in models if corresponding lower-order interactions are also included. Thus if λ_{ijk}^{ABC} is in the model, then all other interactions are included. If λ_{ij}^{AB} is included, then λ, λ_i^A,

and λ_j^B must also be included. This family of models is developed in Birch (1963). The models, which are analogous to linear models encountered in the analysis of factorial experiments, have been extensively used in the analysis of multivariate categorical data. For example, see Goodman (1970, 1971a, 1971b), Haberman (1974, Chaps. 5 and 7, 1978, 1979), and Bishop et al. (1975).

2. *Maximum Likelihood*

The method of maximum likelihood is commonly used for estimation of parameters in a loglinear model. As shown in Haberman (1973b), maximum-likelihood estimates of the expected values of the cell frequencies are uniquely defined if they exist, and they exist if all the observed cell frequencies are positive. Maximum-likelihood estimates of expected cell frequencies can also exist if some observed cell frequencies are zero. Haberman (1973b) provides a necessary and sufficient condition for existence of these maximum-likelihood estimates. These conditions do not depend on the particular sampling method employed. In addition, if maximum-likelihood estimates of expected values of cell means do exist, then the estimates do not depend on the sampling method. Thus the same results are obtained for Poisson sampling as are obtained for multinomial sampling.

Three basic methods are available for computation of maximum-likelihood estimates. The simplest method, if available, involves use of closed-form expressions for the maximum-likelihood estimates. *This approach is available in Example 1*. The estimates $\hat{m}_{ij}$ are defined if the row marginal totals $n_i^A = \Sigma_j n_{ij}$ and the column marginal totals $n_j^B = \Sigma_i n_{ij}$ are all 0. In this case, the overall total $N = \Sigma_i \Sigma_j n_{ij}$ is positive, and $\hat{m}_{ij} = n_i^A n_j^B / N$. Closed-form maximum likelihood estimates may be obtained for a large class of log-linear models which are commonly termed decomposable models. See Bishop and Fienberg (1969), Goodman (1968, 1970, 1971a), and Haberman (1974, chaps. 5 and 7). These loglinear models normally correspond to combinations of hypotheses concerning equiprobability, conditional independence, and independence. They are closely related to graphical models (Darroch et al. 1980). See also *Graphical Models: Overview*.

A traditional approach to calculation of maximum-likelihood estimates is based on the iterative proportional fitting procedure developed in Deming and Stephan (1940) for use with adjustment of census sample data to conform with census population totals. *This approach is readily applied in Example 3*. The algorithm generates a series of approximations to the maximum likelihood estimates by means of simple multiplicative adjustments. This approach was used in Goodman (1968, 1970), Bishop (1969), and Bishop and Fienberg (1969) to study hierarchical loglinear models. The convergence properties of the algorithm, together with generalizations that apply to all loglinear models, are examined in Darroch and Ratcliff (1972) and Haberman (1974, Chap. 3). A disadvantage of iterative proportional fitting is that it does not help in determination of confidence intervals for parameters.

Several versions of the Newton–Raphson algorithm (Haberman 1974, 1978, 1979) can be used to obtain maximum-likelihood estimates. The Newton–Raphson algorithm applies to all loglinear models and is currently used in the majority of computer packages which calculate maximum-likelihood estimates for loglinear models. In some instances, the computer packages for loglinear models are contained in packages for generalized linear models (see *Linear Hypothesis*). The algorithm normally converges rapidly, and by-products of the calculations may be used to estimate the asymptotic covariance matrix of the model parameters.

Large-sample properties of maximum-likelihood estimates for loglinear models have been extensively studied under a variety of conditions. For example, if all expected cell frequencies becomes large and if the model is true (Haberman, 1977), then, for each cell in the table, the ratio between the estimated expected cell frequency and the actual expected cell frequency converges in probability to 1, and a normal approximation is available for the distribution of a linear combination of maximum-likelihood estimates of logarithms of expected values of cell frequencies. This normal approximation is readily employed to construct approximate confidence intervals for linear combinations of logarithms of expected values of cell frequencies. Haberman (1977) also examines much more complex cases under which normal approximations can be used.

In cases in which doubts arise concerning the accuracy of the normal approximation used to construct confidence intervals, it is possible, at least in principle, to obtain exact confidence intervals for linear combinations of logarithms of expected values of cell frequences. The basic approach required is given in Haberman (1974, Chap. 1); however, actual use of exact confidence intervals often involves substantial computational challenges as well as a number of problems concerning the precise way to define an exact confidence interval. Agresti (1992) provides an extensive review of the issues as well as numerous references.

3. *Hypothesis Tests*

In the case of loglinear models, the most common tests of goodness of fit are likelihood-ratio and Pearson chi-square statistics, although other alternatives exist. These statistics apply to the null hypothesis that the model holds and to the alternative hypothesis that the

model is not valid. Central chi-square approximations for the Pearson and likelihood ratio chi-square statistics are available if all expected values of cell frequencies are large and if the model holds. If the model is not valid but all expected cell frequencies are large, then noncentral chi-square approximations may be used.

Cressie and Read (1984) propose a power divergence family of test statistics that includes both the Pearson and likelihood-ratio chi-square and maintains the basic properties of these statistics for the case of large expected cell frequencies. See *Goodness of Fit: Overview*.

Two loglinear models may be compared by use of likelihood-ratio chi-squares. Let M and M' represent two loglinear models such that M implies M'. Let L^2 be the likelihood-ratio chi-square statistic for goodness of fit of M, and let L'^2 be the likelihood-ratio chi-square statistic for goodness of fit of M'. The difference $L^2 - L'^2$ provides a likelihood-ratio test for the null hypothesis that M holds and the alternative hypothesis that M' holds but M is false. If the expected counts are all large and M holds, then $L^2 - L'^2$ has an approximate central chi-square distribution. If the expected counts are large but M does not hold, then a noncentral chi-square approximation may be considered.

In cases in which the accuracy of large-sample approximations is questionable, exact tests of hypotheses are also available, although it is typically the case that the computational labor involved is considerable. Agresti (1992) provides an extensive review of the literature concerning exact tests.

4. *Residual Analysis*

Residual analysis provides a basic supplement to formal hypothesis tests. As in regression analysis, residual analysis of loglinear models permits a thorough search for discrepancies between the data and the model. As in regression analysis, residual analysis can involve use of graphs to consider patterns of residuals and to detect residuals of unusual size. A basic development of residual analysis in loglinear models is found in Haberman (1973a). The raw residual for a cell is the difference between the observed and estimated expected cell frequency. The variability of the raw residual is dependent on the expected cell count, so that it is common to try to rescale the raw residual. A simple rescaling is provided by the standardized residual. The standardized residual for a cell is the ratio of the raw residual and the square root of the estimated expected cell frequency. The sum of the squares of the standardized residuals yields the Pearson chi-square statistic. If the model holds and if the expected cell frequencies are large, then the standardized residuals have approximate normal distributions with means of 0 and variances no greater than 1. The adjusted residual is the ratio of the raw residual and the estimated standard deviation of the raw residual. If the model holds and if the expected cell frequencies are large, then the adjusted residuals have approximate standard normal distributions.

Alternative definitions of residuals are also available. The Freeman-Tukey residuals (Bishop, Fienberg, and Holland 1975, pp. 136–141) for a cell with observed count n and estimated expected count $\hat{m}$ issatisfy

$$z = \sqrt{n} + \sqrt{n+1} - \sqrt{4\hat{m}+1}$$

The Poisson deviance residual D for this cell satisfies the conditions that

$$D^2 = 2[n\log(n/\hat{m}) - n + \hat{m})]$$

and D has the same sign as the raw residual $n - \hat{m}$. These residuals are similar to the standardized residual. The sum of squares of the Poisson deviance residuals is equal to the likelihood-ratio chi-square statistic. For discussion, see Pierce and Schafer (1986).

5. *Model Selection*

A common issue in the study of loglinear models is the selection of models. This issue is complicated in many common cases due to the very large number of loglinear models which may be considered. As a consequence, problems of multiple tests arise (see *Hypothesis Tests, Multiplicity of*). Consider a case in which $c \geqslant 2$ loglinear models M_a are under study for $1 \leqslant a \leqslant c$, and let L_a^2 be the likelihood-ratio chi-square statistic for goodness of fit of model M_a. Approximate significance levels may be obtained for each L_a^2, and, for cases in which M_a implies M_b, an approximate significance level may be obtained for $D_{ab} = L_a^2 - L_b^2$. Based on these results, various strategies can be devised to select models (Bishop, Fienberg, and Holland 1975, Chap. 4, Goodman 1971b, 1973, Haberman 1978, Chap. 3). Measures of quality of fit have also been proposed by Goodman (1971b), Akaike (1974), and Schwarz (1978).

6. *Conclusion*

An extensive literature exists concerning loglinear models. Bishop, Fienberg, and Holland (1975) provide a basic overview of the subject with an emphasis on hierarchical loglinear models. Haberman (1974) provides a rigorous mathematical treatment of loglinear models, while Haberman (1978, 1979) examines such models in the context of research in the social sciences.

Goodman (1968, 1970, 1971a, 1971b, 1973) provides a fundamental series of papers on the use of hierarchical loglinear models. For further consideration of logit models, see *Multivariate Analysis: Discrete Variables, Logistic Regression*.

Bibliography

Agresti A 1992 A survey of exact inference for contingency tables. *Statistical Science* **7**: 131–77

Akaike H New look at the statistical identification. *IEEE Trans. Auto. Control* **19**: 716–23

Bartlett M S 1935 Contingency table interactions *Supplement to the Journal of the Royal Statistical Society* **2**: 248–52

Birch M W 1963 Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B* **25**: 220–33

Bishop Y M M 1969 Full contingency tables, logits, and split contingency tables. *Biometrics* **25**: 383–99

Bishop Y M M, Fienberg S E 1969 Incomplete two-dimensional contingency tables. *Biometrics* **25**: 119–28

Bishop Y M M, Fienberg S E, Holland P W 1975 *Discrete Multivariate Analysis: Theory and Practice*. MIT, Cambridge, MA

Cressie N, Read T R C 1984 Multinomial goodness-of-fit tests. *Journal of the Royal Statistics Society, Series B* **46**: 440–64

Darroch J N, Lauritzen S L, Speed T P 1980 Markov fields and loglinear interaction models for contingency table. *Annals of Statistics* **8**: 522–39

Darroch J N, Ratcliff D 1972 Generalized iterative scaling for loglinear models. *Annals of Mathematical Statistics* **43**: 1470–80

Deming W E, Stephan F F 1940 On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* **11**: 427–44

Goodman L A 1968 The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. *Journal of the American Statistical Association* **63**: 1091–131

Goodman L A 1970 Multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association* **65**: 226–56

Goodman L A 1971a Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. *Journal of the American Statistical Association* **66**: 339–44

Goodman L A 1971b Analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* **13**: 33–61

Goodman L A 1973 Guided and unguided methods for the selection of models for a set of T multidimensional contingency tables. *Journal of the American Statistical Association* **68**: 165–75

Haberman S J 1973a Analysis of residuals in cross-classified tables. *Biometrics* **29**: 205–20

Haberman S J 1973b Loglinear models for frequency data: sufficient statistics and likelihood equations. *Annals of Statistics* **1**: 617–32

Haberman S J 1974 *The Analysis of Frequency Data*. University of Chicago Press, Chicago

Haberman S J 1977 Loglinear models and frequency tables with small expected cell counts. *Annals of Statistics* **5**: 1148–69

Haberman S J 1978 *Analysis of Qualitative Data: Volume 1. Introductory Topics*. Academic, New York

Haberman S J 1979 *Analysis of Qualitative Data: Volume 2. New Developments*. Academic, New York

Haberman S J 1981 Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. *Annals of Statistics* **9**: 1178–86

Norton H W 1945 Calculation of chi-square for complex contingency tables. *Journal of the American Statistical Association* **40**: 251–8

Pierce D A, Schafer D W 1986 Residuals in generalized linear models. *Journal of the American Statistical Association* **81**: 977–86

Rasch G 1960 *Probabilistic Models for Some Intelligence and Attainment Tests*. Nielsen and Lydiche, Copenhagen

Roy S N, Kastenbaum M A 1956 On the hypothesis of no 'interaction' in a multi-way contingency table. *Annals of Mathematical Statistics* **27**: 749–57

Schwarz G 1978 Estimating dimension of a model. *Annals of Statistics* **6**: 461–4

S. J. Haberman

Multivariate Analysis: Discrete Variables (Overview)

This article describes statistical methods for multivariate data sets having discrete response variables. Section 1 summarizes some traditional ways of analyzing discrete data, such as analyses of two-way contingency tables. As in the multivariate analysis of continuous variables, modeling approaches are vital for studying association and interaction structure. Models for a categorical response variable, discussed in Section 2, describe how the probability of a particular response outcome depends on values of predictor variables. These models, of which logistic regression is most important, are analogs for categorical responses of ordinary regression. The loglinear model, discussed in Section 3, provides an analog of correlation analysis; it analyzes association patterns and strength of association. Section 4 discusses sampling assumptions for analyses for discrete data, and Section 5 mentions other types of models for discrete data. Section 6 describes recent advances for analyzing data resulting from repeated or clustered measurement of discrete variables. Section 7 provides some historical perspective.

1. Traditional Analyses of Discrete Multivariate Data

Most commonly, discrete variables are categorical in nature. That is, the measurement scale is a set of categories, such as (liberal, moderate, conservative)

for political philosophy. In multivariate analyses with two or more categorical variables, a *contingency table* displays the counts for the cross-classification of categories. For instance, a contingency table might display counts indicating the number of subjects classified in each of the $9 = 3 \times 3$ combinations of responses on political philosophy and political party orientation (Democrat, Republican, Independent). Those combinations form the cells of the table.

Two categorical variables are *independent* if the probability of response in any particular category of one variable is identical for each category of the other variable. This is an idealized situation that is rarely true in practice, although like any statistical model it is sometimes adequate for describing reality. For instance, some surveys of American college students have shown that opinion about whether abortion should be legal is essentially independent of gender; that is, the percentage favoring legalization is about the same for male and female students.

Treating independence as a simple model for a contingency table, we can test its fit using statistics that compare the observed cell counts in the table to those expected (predicted) by that model. The most commonly used test statistics are

Pearson statistic

$$= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Likelihood – ratio statistic

$$= 2\sum \text{observed} \log\left(\frac{\text{observed}}{\text{expected}}\right) \tag{1}$$

These both have large-sample chi-squared distributions. For r rows and c columns, the degrees of freedom (df) for the test equal $df = (r-1)(c-1)$, and the P-value is the right-tail probability above the observed test statistic value. These chi-squared statistics also can compare counts to expected values for more complex models than independence.

In conducting goodness-of-fit tests, one does not seriously entertain the possibility that the model truly holds; any model is false and provides a simple representation that only approximates reality. However, when the chi-squared statistics do not show significant lack of fit, benefits of model parsimony accrue from using the model fit, which smooths the sample cell proportions somewhat. Interpretations are simpler, and estimates of population characteristics such as the true cell proportions are likely to be better using the model-based estimates.

Multidimensional contingency tables result from cross-classifying more than two categorical variables, such as for studying the association separately at levels of control variables race, gender, or educational level. The model of *conditional independence* states that two categorical variables are independent at each level of a third; that is, ordinary independence applies in each *partial table*. Large-sample chi-squared tests of fit of this model, such as the Cochran–Mantel–Haenszel test, combine the information from the partial tables. Most are designed to be relatively powerful when the strength of the association is similar in each partial table. The model of *homogeneous association* states that the strength of association between two variables is identical at each level of a third variable. Goodness-of-fit tests of these models were originally formulated for stratified two-by-two tables, but generalizations exist for stratified r-by-c tables with ordered or unordered categories

Goodness-of-fit tests have limited use, and in many circumstances it is more informative to estimate the strength of the associations. Goodman and Kruskal (1979) discussed a variety of summary measures. Those for ordinal variables, such as their *gamma* measure, are correlation-like indices that fall between -1 and $+1$. More attention has been focused on measures that arise as parameters in models. The odds ratio, discussed in the next section, is a parameter in logistic regression and loglinear models and is applicable for two-way or multi-way tables.

For estimating parameters in models for discrete data, the method of *maximum likelihood* estimation is the standard. The probability function for the data (e.g., Poisson, binomial), expressed as a function of the parameters after observing the data, is called the *likelihood function*. The maximum likelihood estimates are values for the parameters for which this achieves its maximum; that is, these estimates are the parameter values for which the probability of the observed data takes its maximum value. Except in a few simple cases, numerical methods are needed to find the maximum likelihood estimates and their standard errors. Software is reasonably well developed in the major statistical packages. It is often most natural to analyze discrete data using procedures for generalized linear models that permit non-normal responses, such as the Poisson and binomial. Examples include PROC GENMOD in SAS and the 'glm' function in S-Plus. For instance, for SAS see Allison (1999) and Stokes et al. (1995), for S-Plus see Venables and Ripley (1999), and for Stata see Rabe-Hesketh and Everitt (2000).

2. *Regression Modeling of a Categorical Response*

For multidimensional contingency tables and other forms of multivariate data, modeling approaches are vital for investigating association and interaction structure. *Logistic regression* is an analog of ordinary regression for binary response variables, and it has extensions for multicategory response variables. Like

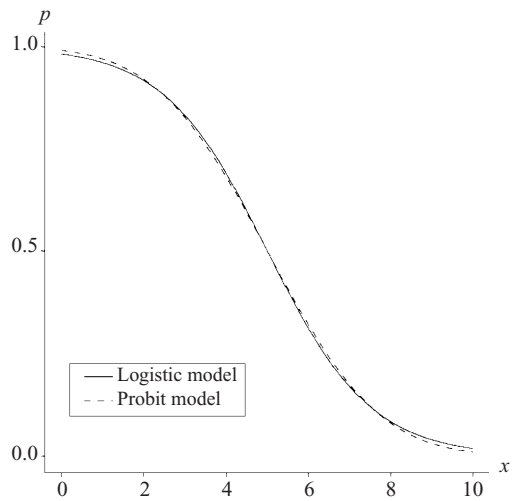

Figure 1
Plot of logistic regressional model $\log[p/(1-p] = 4.0-0.8x$ and the probit regression model probit(p) $= 2.35-0.47x$

ordinary regression, logistic regression distinguishes between response and predictor variables. The predictor variables can be qualitative or quantitative, with qualitative ones handled by dummy variables.

Let p denote the probability of outcome in a particular response category (e.g., favoring the Democratic rather than Republican candidate for President). For the observations at particular fixed values of predictor variables, the number of responses in the two outcome categories is assumed to have the binomial distribution. With a set of predictors x_1, $x_2, \ldots, x_p$, the logistic regression model is

$$\log\left(\frac{p}{1-p}\right) = \alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p. \quad (2)$$

The ratio $p/(1-p)$ is called the *odds*, and the natural logarithm of the odds is called the *logit*.

One reason for using the logit instead of p itself on the left-hand side of this equation is to constrain predicted probabilities to fall between 0 and 1. With a single predictor, for instance, the *linear probability model* $p = \alpha+\beta x$ permits p to fall below 0 or above 1 for many values of x; by contrast, the formula $\log[p/(1-p)] = \alpha+\beta x$ maps out an S-shaped function for which p increases or decreases according to whether $\beta > 0$ or $\beta < 0$ but always stays bounded between 0 and 1. For instance, for the decision to buy/rent a home, suppose the relationship between $p =$ the probability of buying and $x =$ annual family income (in ten thousands of dollars) satisfies the logistic regression model $\log[p/(1-p)] = 4.0-0.8x$. Figure 1 shows the plot of p for x between 0 and 10 (i.e., below \$100,000). The plot is roughly linear for p between 0.2 and 0.8.

The *probit model* is another model that has this S-shape. The function probit $(p) = \alpha+\beta x$ has a graph for p (or for $1-p$ when $\beta < 0$) that has the shape of a normal cumulative distribution function with mean $-\alpha/\beta$ and standard deviation $1/|\beta|$. The probit model can be derived by assuming an underlying latent normal variable (e.g., Long 1997). In practice, the probit model gives a very similar fit to the logistic regression model. To illustrate, Fig. 1 also shows the plot of the equation probit $(p) = 2.35-0.47x$, which has the shape of a normal cumulative distribution function with mean 5.0 and standard deviation 2.1; for instance, a decrease in p of 0.68, from $p = 0.84$ to $p = 0.16$, occurs between $x = 5.0-2.1 = 2.9$ and $x = 5.0+2.1 = 7.1$. Although the fits are very similar, estimates of α and β are on a different scale for probit and logistic models, and usually logistic parameter estimates are about 1.6 to 1.8 times the probit parameter estimates.

With multiple predictors, the coefficient β_k of x_k in the logistic regression model describes the effect of x_k, controlling for the other predictors in the model. The antilog of β_k is an odds ratio representing the multiplicative effect of a one-unit change in x_k on the odds. For instance, in a model for the probability in favoring the Democratic candidate for President, suppose $\beta_k = 0.40$ for the coefficient of $x_k =$ gender, with $x_k = 1$ for females and 0 for males. Then, controlling for the other variables in the model, for females the odds of favoring the Democratic candidate equal $e^{0.40} = 1.49$ times the odds for males; that is, the odds of favoring the Democrat are 49 percent higher for females than for males.

For quantitative predictors, simple interpretations that apply directly to the probability p rather than to the odds $p/(1-p)$ result from straight-line approximations to the logistic S-shaped curve for p. When p is near 0.5, $\beta_k/4$ is the approximate change in p for a one-unit change in x_k, controlling for the other predictors. To illustrate, suppose $\beta_k = -0.12$ for the coefficient of $x_k =$ annual income (in ten thousands of dollars) in the model for the probability of favoring the Democratic candidate. Then, for an increase of \$10,000 in income near where $p = 0.5$, the probability of support for the Democratic candidate decreases by about $0.12/4 = 0.03$, other variables being kept constant. To test the effect of a term in the model, the ratio of the estimate of β_k to its standard error has an approximate standard normal distribution, under the hypothesis of no effect ($H_0: \beta_k = 0$).

Using the estimated parameters from the logistic regression model, one can also estimate the probability p for given settings for the predictors, as

$$\hat{p} = \frac{\exp(\hat{\alpha}+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\cdots+\hat{\beta}_p x_p)}{1+\exp(\hat{\alpha}+\hat{\beta}_1 x_1+\hat{\beta}_p x_p+\cdots+\hat{\beta}_p x_p)} \quad (3)$$

When all predictors are categorical, chi-squared tests analyse the adequacy of the model fit by comparing the observed cell counts to values expected under the model.

Generalizations of logistic regression apply to categorical response variables having more than two categories. A multicategory extension of the binomial distribution, called the *multinomial*, applies to the response counts at each setting of the predictors.

For nominal responses, the generalized model forms logits by pairing each response category with a baseline category. For a r-category response, the model has form

$$\log\left[\frac{P(Y=j)}{P(Y=r)}\right] = \alpha_j + \beta_{1j}x_1 + \beta_{2j}x_2 + \cdots + \beta_{pj}x_p,$$
$$j = 1, 2, \ldots, r-1.$$

A separate set of parameters applies for each response category contrasted with the final one. For instance, if Y refers to preferred candidate and has categories (Democrat, Republican, Independent), the effect of gender on the choice between Democrat and Independent may differ from its effect on the choice between Republican and Independent. The choice of category for the baseline (the denominator of the logits) is arbitrary. Regardless of this choice the same inference occurs about comparisons of models with and without certain predictors and about estimated probabilities in the r response categories. For ordinal responses, the *cumulative logit model* describes the effects of predictors on the odds of response below any given level instead of above it (i.e., the model refers to an odds for the cumulative probability; see McCullagh 1980). It has form

$$\log\left[\frac{P(Y \leqslant j)}{P(Y > j)}\right] = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$
$$j = 1, \ldots, r-1. \quad (4)$$

For further details about logistic regression models, see *Multivariate Analysis: Discrete Variables (Logistic Regression)*; also Agresti (1996, Chap. 5), Collett (1991) Hosmer and Lemeshow (2000) and Long (1997).

3. *Loglinear Models*

Logistic regression resembles ordinary regression in distinguishing between a response variable Y and a set of predictors $\{x_k\}$. *Loglinear models*, by contrast, treat all variables symmetrically and are relevant for analyses analogous to correlation analyses, studying the association structure among a set of categorical response variables.

For multidimensional contingency tables, a variety of models are available, varying in terms of the complexity of the association structure. For three variables, models include ones for which (a) the variables are mutually independent, (b) two of the variables are associated but are jointly independent of the third, (c) two of the variables are conditionally independent, given the third variable, but may both be associated with the third, (d) each pair of variables is associated, but the association between each pair is homogeneous at each level of the third variable, and (e) each pair of variables is associated and the strength of association between each pair may vary according to the level of the third variable (Bishop et al. 1975).

Relationships describing probabilities in contingency tables are naturally multiplicative, so ordinary regression-type models occur after taking the logarithm, which is the reason for the term *loglinear*. To illustrate, in two-way contingency tables independence between row variable X and column variable Y is equivalent to the condition whereby the probability of classification $P(X=i, Y=j)$ in the cell in row i and in column j depends only on the marginal probabilities $P(X=i)$ and $P(Y=j)$,

$$P(X=i, Y=j) = P(X=i)P(Y=j) \quad (5)$$

for all i and j. For a sample of size n, the expected count $\mu_{ij} = nP(X=i, Y=j)$ in that cell therefore satisfies

$$\log\mu_{ij} = \log[nP(X=i)P(Y=j)]$$
$$= \log(n) + \log[P(X=i)] + \log[P(Y=j)]. \quad (6)$$

This has the loglinear form

$$\log\mu_{ij} = \alpha + \beta_i + \gamma_j \quad (7)$$

for which the right-hand side resembles a simple linear model—two-way analysis of variance (ANOVA) without interaction. Alternatively this has regression form by replacing $\{\beta_i\}$ by dummy variables for the rows times parameters representing effects of classification in those rows and replacing $\{\gamma_j\}$ by dummy variables for the columns times parameters representing effects of classification in those columns.

Similarly, loglinear model formulas for more complex models such as those allowing associations resemble ANOVA models except for predicting the logarithm of each cell expected frequency rather than the expected frequency itself. Dummy variables represent levels of the qualitative responses, and their interaction terms represent associations. The associations are described by odds ratios. Logistic regression models with qualitative predictors are equivalent to certain loglinear models, having identical estimates of odds ratios and identical goodness-of-fit statistics. For ordinal variables, specialized loglinear

models assign ordered scores to the categories and have parameters describing trends in associations.

For further details about loglinear models, see *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)*; also Agresti (1996, Chap. 6), Bishop et al. (1975), and Fienberg (1980).

4. Sampling Assumptions

Inferential analysis of multivariate discrete data, like that of any type of data, requires assumptions about the sampling mechanism that generated the data. Often, such as in most surveys, the overall sample size is fixed and the multinomial distribution applies as the probability distribution for the counts in a contingency table. In many cases, certain marginal totals are also fixed, such as in experimental designs in which the number of subjects at each experimental condition is fixed. In addition, in regression-type models such as logistic regression, it is common to treat the counts at the combinations of levels of the predictor variables as fixed even if they are not fixed in the sampling design. In these cases, at each such combination the response is usually assumed to have a binomial distribution when the response is binary and a multinomial distribution when the response is multicategory; this replaces the usual normal response assumption for regression models for continuous responses. Other sampling possibilities, less common in practice, are that the sample size is itself random (i.e., no counts are fixed, such as in observing categorical outcome measures for all subjects who visit some clinic over a future time period) or, at the other extreme, all marginal totals are fixed; these lead to Poisson and hypergeometric distributions.

Although some methods depend on the sampling assumption, estimates and test statistics for many analyses are identical for the common sampling designs. For instance, for two-way contingency tables, the chi-squared statistic, its degrees of freedom, and its large-sample P-value are the same when only the total sample size is fixed, when only the row totals are fixed, when only the column totals are fixed, or when both row and column totals are fixed.

Under standard sampling assumptions, traditional methods of inference for discrete data use statistics having approximate chi-squared or normal distributions for large samples. Alternative methods are available for small samples. Fisher's exact test, due to the famous statistician R. A. Fisher, is a test of independence in 2-by-2 contingency tables. Analogs of this test are available for larger tables and for testing conditional independence and homogeneous association in three-way tables (characterized in terms of homogeneity of odds ratios) as well as testing fit of more complex logit and loglinear models. Small-sample confidence intervals for parameters such as proportions, difference of proportions, and odds ratios, result from inverting exact tests for various null values of parameters. Much of the current research in small-sample inference uses simulation methods to closely approximate the true sampling distributions of relevant statistics (e.g., Forster et al. 1996). StatXact (Cytel Software 1999) is statistical software dedicated to small-sample discrete-data methods. Some of these analyses are also available in modules of software such as SAS and SPSS.

5. Alternative Model Types

For contingency tables relating several response variables, alternative statistical approaches to loglinear models can describe association among the variables. Foremost among these is *correspondence analysis*, a graphical way of representing associations in two-way contingency tables that has been particularly popular in France (e.g., Benzécri 1973; see *Multivariate Analysis: Discrete Variables (Correspondence Models)*. The rows and columns are represented by points on a graph, the positions of which indicate associations between the two. Goodman (1986) developed a model-based version of this approach. He showed it has equivalences with *correlation models*, which determine scores for categories that maximize correlations between categorical variables. He also showed approximate connections with a set of *association models* that he helped to develop, some of which have loglinear model form but with scores or score-like parameters assigned to levels of ordinal variables. Clogg and Shihadeh (1994) surveyed such models.

Another graphical way of representing a multivariate relationship uses *graphical models* that study conditional independence structure among the variables. Roughly speaking, the graph has a set of vertices, each vertex representing a variable. An edge connecting two vertices represents a conditional association between the corresponding two variables. For instance, for four variables, the graph in Fig. 2 portrays a model that assumes that X and Y are independent and that X and Z are independent, conditional on the remaining two variables, but permits association between W and X and between each pair of variables in the set $\{W, Y, Z\}$. The four edges in the graph, connecting W and X, W and Y, W and Z, and Y and Z, represent pairwise conditional associations. Edges do not connect X and Y or X and Z, since those pairs are conditionally independent. Many loglinear models are graphical models. For details, see Lauritzen (1996) and *Graphical Models: Overview*.

Logistic/probit regression models and loglinear models, together with long-established regression and ANOVA methods for normal response variables, are special cases of a broad family of *generalized linear*

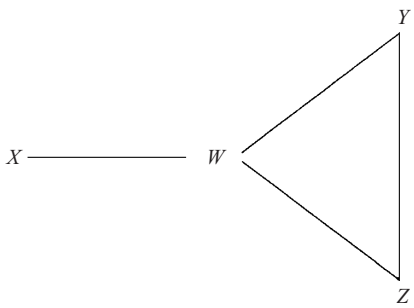

Figure 2
A graphical representation of four variables, with conditional independence between X and Y and between X and Z

models (Gill 2000). These models are characterized by a choice of distribution for the response variable (e.g. binomial, Poisson, normal), a function of the mean to be modeled (e.g., logit of the mean, log of the mean, the mean itself), and the variables that enter the linear model formula as predictors.

For binary data, the model using the logit transform of the probability p is the most popular generalized linear model. For count data, generalized linear models usually assume a Poisson response, sometimes called *Poisson regression models*. However, more variability often occurs in the counts than the Poisson distribution allows (i.e. there is said to be *overdispersion*). The Poisson distribution forces the variance to equal the mean; according to it, for instance, at a given mean for Y the variance of the conditional distribution of Y cannot change as predictors are added to or deleted from the model. Methods of dealing with overdispersion include using discrete distributions for which the variance can exceed the mean, such as the *negative binomial*, and adding an extra error term to the model (a *random effect*) to account for unexplained variability due to factors such as unmeasured variables and measurement error.

Extensions of generalized linear models deal with more complex situations. For instance, sometimes the response is a mixture of discrete and continuous, such as when a certain proportion of the responses take value 0 but the rest can be any positive real number (e.g., responses to 'How much time do you spend on exercise each week?'). Sometimes observations are censored—we know only that the response falls above some point or below some point, such as when we count events of a certain type over time but cannot observe beyond the end of the experimental period (Andersen et al. 1993). Models can then focus on the continuous time to response of an event or the discrete counting of events. Depending on the application and the sort of censoring, the models have different names, including survival models, event-history models, and tobit models (e.g., Long 1997).

6. *New Methods for Repeated Measurement Data*

For standard inference with the usual sampling assumptions, the methods discussed above treat separate observations in a sample as independent. In the social and behavioral sciences, multivariate data often results from repeated measurement (e.g., longitudinal) studies. For such studies, this independence assumption is violated. More generally, data sets that consist of clusters of observations such that observations within clusters tend to be more alike than observations between clusters require specialized methods. This was an active research area in the last 15 years of the twentieth century. New methods make it possible to extend traditional repeated measures ANOVA to discrete data and to permit various correlation structures among the responses.

For clustered data, recently developed methods enable modeling binary or categorical responses with explanatory variables, for instance using the logit transform. Two approaches have received considerable attention. In one, models address marginal (separate) components of the clustered response, allowing for dependence among the observations within clusters but treating the form of that dependence as a nuisance that is not a primary object of the modeling. The other approach uses cluster-specific random effects as a mechanism for inducing the dependence.

The marginal modeling approach has the advantage of relative computational simplicity. Estimates are obtained using the methodology of *generalized estimating equations* (see Diggle et al. 1994). With this quasi-likelihood method, rather than fully specify joint distributions for the responses, one need only model how the variance depends on the mean and predict a within-cluster correlation structure. A simple version of the method exploits the fact that good estimates of regression parameters usually result even if one naively treats the component responses in a cluster as independent. Although those parameter estimates are normally adequate under the naïve independence assumption, standard errors are not. More appropriate standard errors result from an adjustment using the empirical dependence in the sample data. Although one makes a 'working guess' about the likely correlation structure, one adjusts the standard errors to reflect what actually occurs for the sample data. An exchangeable working correlation structure under which correlations for all pairs of responses in a cluster are identical is more flexible and usually more realistic than the naïve independence assumption.

With models that apply directly to the marginal distributions, the regression effects are population-averaged. For instance, a fitted probability estimate refers to *all* subjects at the given level of predictors, and the model does not attempt to study potential heterogeneity among the subjects in this probability.

Alternatively, one can model subject-specific probabilities and allow such heterogeneity by incorporating subject or cluster random effects in the model. For instance, a model might contain a cluster-specific intercept term that varies among clusters according to some normal distribution. The estimated variance of that distribution describes the extent of heterogeneity. This type of model is called a *generalized linear mixed model* since it contains a mixture of fixed and random effects. Specifying this form of model also has implications about the effects in a marginal model, but the form of model may differ. For instance, with random effects in a logistic regression model, the corresponding model for marginal distributions is only approximated by a logit model and the effects have different size (Diggle et al. 1994).

An alternative approach of dealing with the cluster terms is the fixed effects one of item response models such as the *Rasch model* (see *Factor Analysis and Latent Structure: IRT and Rasch Models*). In some applications with a multivariate response, it is natural to assume the existence of an unmeasured *latent variable* that explains the association structure. Conditional on the value of that latent variable, the observed variables are independent. The observed association structure then reflects association between the observed variables and that latent variable. When the latent variable consists of a few discrete classes, such as when the population is a mixture of a few types of subjects, then the resulting model is called a *latent class model*. This is an analog of factor analysis in which the observed and latent variables are categorical rather than continuous. See Heinen (1993) and *Factor Analysis and Latent Structure: Overview*.

7. Historical Notes

The early development of methods for discrete multivariate data took place in England. In 1900, Karl Pearson introduced his chi-squared statistic (although in 1922 R. A. Fisher showed that Pearson's *df* formula was incorrect for testing independence) and G. Udny Yule presented the odds ratio and related measures of association. Much of the early literature consisted of debates about appropriate measures of association for contingency tables. Goodman and Kruskal (1979) surveyed this literature and made significant contributions of their own.

In the 1950s and early 1960s, methods for multidimensional contingency tables received considerable attention. These articles were the genesis of substantial research on logistic regression and loglinear models between about 1965 and 1975. Much of the leading work in that decade took place at the universities of Chicago, Harvard, and North Carolina. At Chicago, Leo Goodman wrote a series of groundbreaking articles for statistics and sociology journals that popularized the methods for social science applications. Simultaneously, related research at Harvard by students of Frederick Mosteller (such as Stephen Fienberg) and William Cochran and at North Carolina by Gary Koch and several students and coworkers was highly influential in the biomedical sciences.

Perhaps the most far-reaching contribution was the introduction by British statisticians John Nelder and R. W. M. Wedderburn in 1972 of the concept of generalized linear models. This unifies the primary categorical modeling procedures—logistic regression models and loglinear models—with long-established regression and ANOVA methods for normal-response data.

Texts covering logistic regression are, from least to most advanced, Agresti and Finlay (1997, Chap. 15), Agresti (1996), Hosmer and Lemeshow (2000), Collett (1991), Long (1997), Agresti (1990). Texts covering loglinear models, from least to most advanced, are Agresti and Finlay (1997, Chap. 15), Agresti (1996), Fienberg (1980), Clogg and Shihadeh (1994), Agresti (1990), Bishop et al. (1975). Andersen (1980) and Long (1997) discuss some alternative methods for discrete data.

See also: Experimental Design: Overview; Multivariate Analysis: Classification and Discrimination; Multivariate Analysis: Discrete Variables (Correspondence Models); Multivariate Analysis: Discrete Variables (Loglinear Models)

Bibliography

Agresti A 1990 *Categorical Data Analysis*. Wiley, New York

Agresti A 1996 *An Introduction to Categorical Data Analysis*. Wiley, New York

Agresti A, Finlay B 1997 *Statistical Methods for the Social Sciences*, 3rd edn. Prentice-Hall, Upper Saddle River, NJ

Allison P D 1999 *Logistic Regression Using the SAS System*. SAS Institute, Cary, NC

Andersen E B 1980 *Discrete Statistical Models with Social Science Applications*. North Holland, Amsterdam

Andersen P K, Borgan Ø, Gill R D, Keiding, N 1993 *Statistical Models Based on Counting Processes*. Springer, New York

Benzécri J-P 1973 *L'analyse des Données*. Dunod, Paris

Bishop Y M M, Fienberg S E, Holland P W 1975 *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA

Clogg C C, Shihadeh E S 1994 *Statistical Models for Ordinal Variables*. Sage, Thousand Oaks, CA

Collett D 1991 *Modelling Binary Data*. Chapman and Hall, London

Cytel Software 1999 *StatXact, Version 4*. Cambridge, MA

Diggle P J, Liang K-Y, Zeger S L 1994 *Analysis of Longitudinal Data*. Oxford University Press, Oxford

Fienberg S E 1980 *The Analysis of Cross-classified Categorical Data*, 2nd edn MIT Press, Cambridge, MA

Forster J J, McDonald, J W, Smith, P W F 1996 Monte Carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society*, *Series B* **58**: 445–53

Gill J 2000 *Generalized Linear Models: A Unified Approach*. Sage, Thousand Oaks, CA

Goodman L A 1986 Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review* **54**: 243–70
Goodman L A, Kruskal W H 1979 *Measures of Association for Cross Classifications*. Springer, New York
Heinen T 1993 *Discrete Latent Variable Models*. Tilburg University Press, The Netherlands
Hosmer D W, Lemeshow S 2000 *Applied Logistic Regression*, 2nd edn., Wiley, New York
Lauritzen, S L 1996 *Graphical Models*. Oxford University Press, Oxford
Long J S 1997 *Regression Model for Categorical and Limited Dependent Variables*. Sage, Thousand Oaks, CA
McCullagh P 1980 Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* **42**: 109–42
Rabe-Hesketh S, Everitt B 2000 *A Handbook of Statistical Analyses Using Stata*, 2nd edn. CRC Press, London
Stokes M E, Davis C E, Koch G G 1995 *Categorical Data Analysis Using the SAS System*. SAS, Cary, NC
Venables W N, Ripley B D 1999 *Modern Applied Statistics with S-Plus*, 3rd edn. Springer, New York

A. Agresti

Multivariate Analysis: Overview

1. Introduction

Multivariate analysis is conceptualized by tradition as the statistical study of experiments in which multiple measurements are made on each experimental unit and for which the relationship among multivariate measurements and their structure are important to the experiment's understanding. For instance, in analyzing financial instruments, the relationships among the various characteristics of the instrument are critical. In biopharmaceutical medicine, the patient's multiple responses to a drug need be related to the various measures of toxicity. Some of what falls into the rubric of multivariate analysis parallels traditional univariate analysis; for example, hypothesis tests that compare multiple populations. However, a much larger part of multivariate analysis is unique to it; for example, measuring the strength of relationships among various measurements.

Although there are many practical applications for each of the methods discussed in this overview, we cite some applications for the classification and discrimination methods in Sect. 6.5. The goal is to distinguish between two populations, or to classify a new observation in one of the populations. Examples are (a) solvent and insolvent companies based on several financial measures; (b) nonulcer dyspeptics versus normal individuals based on measures of anxiety, dependence, guilt, and perfectionism; (c) Alaskan vs. Canadian salmon based on measures of the diameters of rings. For other such applications, see Johnson and Wichern (1999).

Multivariate analysis, due to the size and complexity of the underlying data sets, requires much computational effort. With the continued and dramatic growth of computational power, multivariate methodology plays an increasingly important role in data analysis, and multivariate techniques, once solely in the realm of theory, are now finding value in application.

2. Normal Theory

2.1 Distribution Theory

2.1.1 Multivariate normal distribution. In much the same way that the univariate normal distribution is central to statistics, the multivariate normal distribution plays a similar central role. Let $\mathbf{X} = (X_1,\ldots, X_p)$ be a p-dimensional vector of the multiple random variables measured on an experimental unit. The probability density, $f(\mathbf{x})$, of $\mathbf{X}$ is multivariate normal, denoted $N(\boldsymbol{\mu}, \Sigma)$, if

$$f(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\{-(1/2)(\mathbf{x}-\boldsymbol{\mu})\,\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})'\}$$

for $-\infty < x_i < \infty$, $i = 1,\ldots, p$, where $\boldsymbol{\mu} = (\mu_1,\ldots, \mu_p)$ is the p-dimensional population mean vector $(-\infty < \mu_i < \infty, i = 1,\ldots, p)$ and Σ is the $p \times p$ population (positive definite) variance–covariance matrix. When Σ is singular, the probability density function $f(\mathbf{x})$ does not exist as noted, but there are other meaningful ways to define the multivariate normal distribution.

The contours of constancy of $f(\mathbf{x})$ are ellipsoids centered at $\boldsymbol{\mu}$ with axes determined by the eigenvectors of Σ. Any subset of $X_1,\ldots X_p$ itself has a multivariate normal distribution and any set of linear combinations has a multivariate normal distribution. Also the conditional probability distribution of any subset given another subset has a multivariate normal distribution.

2.1.2 Sample statistics and their properties. Suppose that a random sample $\mathbf{X}_1,\ldots, \mathbf{X}_N$, of size N, is observed from a multivariate normal distribution. The natural estimators of $\boldsymbol{\mu}$ and Σ are, respectively, the sample mean vector $\bar{\mathbf{X}}$ and the sample variance–covariance matrix $S = \sum_{i=1}^{N}(\mathbf{X}_i-\bar{\mathbf{X}})'(\mathbf{X}_i-\bar{\mathbf{X}})/(N-1)$.

These estimators possess many of the properties of the univariate sample mean and variance. They are independent and unbiased; furthermore, $\bar{\mathbf{X}}$ and $[(N-1)/N]S$ are maximum likelihood estimators. The

sampling distribution of $\bar{\mathbf{X}}$ is $N(\boldsymbol{\mu}, \Sigma/N)$, and the sampling distribution of $(N-1)S$ is called the Wishart distribution with parameter Σ and $N-1$ degrees of freedom, as obtained by John Wishart in 1928.

For the univariate normal distribution, when the variance σ^2 is known, the mean $\bar{x}$ is an optimal estimator of the population mean μ in any reasonable sense. It was shown in 1956 by Charles Stein that when $p > 2$, and the underlying multivariate distribution is $N(\boldsymbol{\mu}, I)$, one can improve upon $\bar{\mathbf{X}}$ as an estimator of $\boldsymbol{\mu}$ in terms of its expected precision, because surprisingly, even with the identity matrix I being the population variance–covariance matrix, there is information in S to improve estimation of $\boldsymbol{\mu}$.

Authoritative references to the multivariate normal distribution and its properties include Anderson (1984) and Tong (1990).

2.1.3 Characteristic roots. When the sample covariance matrix S has a Wishart distribution, many invariant tests are functions of the characteristic roots of S. This led to the study of the joint distribution of the roots, which was obtained simultaneously by several authors in 1939. Characteristic roots of a sample covariance matrix also play a role in physics and other fields.

The basic distribution theory for the case of a general diagonal population covariance matrix was developed by Alan James in 1956 and requires expansions in terms of zonal polynomials. For details of the noncentral distribution of the characteristic roots and zonal polynomials see Mathai et al. (1995).

2.2 Inference

Inference concerning $\boldsymbol{\mu}$ when Σ is known is based, in part, upon the Mahalanobis distance $N(\bar{\mathbf{X}}-\boldsymbol{\mu})\Sigma^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})'$ which has a χ^2_N distribution when $\mathbf{X}_1, \ldots \mathbf{X}_N$ is a random sample from $N(\boldsymbol{\mu}, \Sigma)$. When Σ is not known, inference about $\boldsymbol{\mu}$ utilizes the Mahalanobis distance with Σ replaced by its estimator S. The distribution of the quantity $N(\bar{\mathbf{X}}-\boldsymbol{\mu})S^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})'$ was derived by Harold Hotelling in 1931 and is called Hotelling's T^2. He showed that $(N-p)T^2/(N-1)p$ has a standard F-distribution, $F_{p,N-p}$.

To test, at level α, the null hypothesis $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against the alternative $H_A: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ based upon multivariate normal data with unknown Σ, one rejects H_0 if $N(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)\, S^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)'$ exceeds $(N-1)p/(N-p)F^{\alpha}_{p,N-p}$, where $F^{\alpha}_{p,N-p}$ denotes the upper α-critical value of $F_{p,N-p}$. Elliptical confidence intervals for $\boldsymbol{\mu}$ can be constructed as $\{\boldsymbol{\mu}: N(\bar{\mathbf{X}}-\boldsymbol{\mu})\ S^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})' \leqslant ((N-1)p/(N-p)) \times F^{\alpha}_{p,N-p}\}$. Useful rectangular confidence intervals for μ are more difficult to obtain and typically rely on suitable probability inequalities (see Hochberg and Tamhane 1987). For the univariate normal distribution, one-sided hypothesis testing straightforwardly follows from two-sided testing. However, various one-sided tests for the multivariate case are much more complicated. A particular class of techniques that is useful in this case are those of order-restricted inference.

The comparison of two multivariate normal population means, with common but unknown variance–covariance matrix, can be based upon a two-sample version of Hotelling's T^2 statistic.

2.3 Linear Models: Multivariate Analysis of Variance (MANOVA)

To compare k (> 2) multivariate normal population means (with common unknown variance–covariance matrix), one needs to utilize generalizations of Hotelling's T^2 statistic, similar to the relationship of the F-statistic for multiple populations to the two-sample t-statistic for univariate data.

To understand more generally multivariate analysis of variance, consider for the univariate setting the analysis-of-variance (ANOVA) for a given hypothesis.

Source	Sum of squares	Degrees of freedom	F-statistic
Hypothesis	SS_H	d_H	$F = (SS_H/d_H)/(SS_E/d_E)$
Error	SS_E	d_E	

The null hypothesis reference distribution for the F statistic is F_{d_H,d_E} and the alternative distribution is a non central F-distribution; these result from SS_H and SS_E being independent random variables with $SS_E \sim \sigma^2\chi^2_{d_E}$ and under H_0, $SS_H \sim \sigma^2\chi^2_{d_H}$. For example, in a one-way layout, SS_H is the sum-of-squares for treatment, and SS_E is the usual sum-of-squares for error.

The general structure of an ANOVA table for multivariate data of dimension p is parallel in structure; S_H becomes H, a $p \times p$ matrix of sum-of-squares and cross-products, and SS_E becomes E, a $p \times p$ matrix related to the variance–covariance matrix of the multivariate error vector. The random matrices H and E are independent; E has a p-dimensional Wishart distribution with d_E degrees of freedom, and under H_0, H has a p-dimensional Wishart distribution with degrees of freedom d_H.

However, now the ratio of sums of squares is translated to the eigenvalues of HE^{-1}. Because there is no uniformly best function of these eigenvalues, various alternative functions of the product, sum, or maximum of the eigenvalues have been proposed. The best-known statistics are the Wilks lambda, Hotelling-Lawley trace, Pillai trace, and Roy maximum, and are provided in standard MANOVA statistical packages.

A classical archeological example from 1935 consists of four series of Egyptian skulls ranging from the predynastic, sixth to twelfth dynasties, twelfth and thirteenth dynasties, and Ptolemaic dynasties. For each of the multiple skulls in the four series, there were measurements on breadth, length, and height. A MANOVA showed that the dynasties were not homogeneous. Subsequent analyses were carried out using multiple comparisons. The MANOVA of only two dynasty comparisons reduces to the two-sample Hotelling's T^2.

3. Relationships

3.1 Regression

Regression analysis is designed to predict a measure Y based on concomitant variables $\mathbf{X} = (X_1, \ldots, X_p)$. It was historically traditional to assume that $(Y, \mathbf{X})$ have a joint normal distribution with zero means and $(p+1) \times (p+1)$ covariance matrix of $(Y, \mathbf{X})$: $\mathrm{Var}(Y) = \sigma_{00}$, $\mathrm{Covar}(Y, \mathbf{X}) = \mathbf{\Sigma}_{12}$, $\mathrm{Var}(\mathbf{X}) = \mathbf{\Sigma}_{22}$. Then the conditional mean of Y given $\mathbf{X} = \mathbf{x}$ is $\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\ \mathbf{x}'$, which is a linear function of the form $\beta_1 x_1 + \ldots + \beta_p x_p$; the conditional variance is $\sigma_{00} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}'$, which is independent of $\mathbf{x}$, a fact that is critical in the analysis.

Although the above description involves the multivariate normal distribution, the subject of regression is usually treated more simply as ordinary linear regression involving least squares. See *Linear Hypothesis: Regression (Basics)*.

The more general regression model consists of a $p \times n$ matrix Y in which the n columns are independently distributed, each having an unknown $p \times p$ covariance matrix Σ, and the expected value of Y is a function of unknown parameters, namely, $EY = X_1 B X_2$, where X_1 and X_2 are $p \times q$ and $r \times n$ matrices, respectively, and B is a $q \times r$ matrix of unknown parameters which is to be estimated. This model comprises what is called the multivariate general linear model (see Anderson 1984 for details).

3.2 Correlation Hierarchy: Partial, Multiple, Canonical Correlations

If $(\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is a p-dimensional random vector with covariances σ_{ij}, the population correlation coefficient between X_i and X_j is defined as $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$, with the sample analogue correspondingly defined. The term co-relation was used by Sir Francis Galton in 1888. Karl Pearson introduced the sample correlation coefficient in 1896; alternative names are product moment correlation or Pearson's correlation coefficient. Because $-1 \leqslant \rho_{ij} \leqslant 1$ and $\rho_{ij} = \pm 1$ only when X_i and X_j are perfectly linearly related, the correlation has become a common measure of linear dependence.

For normally distributed data, the distribution of the sample correlation was obtained in 1915 by R. A. Fisher. The test statistic $\sqrt{N-2}\,r/\sqrt{1-r^2}$ can be used to test that the population correlation is zero; under this hypothesis, this statistic has a Student's t distribution with $N-2$ degrees of freedom. Fisher later obtained a transformation to $z = \tanh^{-1}(r)$, which he showed to have an approximate normal distribution with mean $\tanh^{-1}(\rho)$ and variance $(N-3)^{-1}$. Consequently, the z transformation provides confidence intervals for ρ. The definitive analysis of the distribution of r and z is given by Hotelling (1953).

The population multiple correlation coefficient arises when $X_2, \ldots, X_p$ are used to predict X_1, and is defined as the population correlation between X_1 and the best linear predictor of X_1 based on $X_2, \ldots, X_p$. A similar definition holds for the sample multiple correlation. The population partial correlation is the correlation between a pair of variables, say, X_1 and X_2, conditional on the remainder being fixed, that is, $X_3, \ldots, X_p$ being fixed, with a similar definition holding for the sample. See Anderson (1984) for results concerning the distributions of the sample multiple correlation coefficient and the sample partial correlation coefficient, as well as related techniques for statistical inference, in the case of normally distributed data. Canonical correlations are used to measure the strength of the relationship between two sets of variables and are described more fully in Sect. 6.2.

3.3 Measures of Dependence

Although Pearson's correlation coefficient is a natural way to measure dependence between random variables X and Y having a bivariate normal distribution, it can be less than meaningful for non-normal bivariate distributions. Popular and useful population measures for continuous random variables include Kendall's $\tau (= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x, y) f(x, y)\,dx\,dy - 1)$ and Spearman's $\rho\ (= 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (F(x, y) - F_X(x)F_Y(y)) f_X(x) f_Y(y)\,dx\,dy)$, where $F(x, y)$, $F_X(x)$, $F_Y(y)$ are, respectively, the bivariate distribution function of X and Y, and the two marginal distribution functions of X, Y; and where $f(x, y)$, $f_X(x)$, and $f_Y(y)$ are their corresponding density functions. Kendall's τ can be written as $4\,E(F(X, Y)) - 1$, where the expectation is with respect to the joint distribution of X, Y and Spearman's rho can be written as $12\,E[F(X, Y) - F_X(X)F_Y(Y)]$, where the expectation is with respect to independent X and Y, with distributions $F_X(x)$ and $F_Y(y)$. The sample estimates of Kendall's τ and Spearman's rho are given in Agresti (1984). Other more theoretical approaches to measuring dependence are based upon maximizing the correlation, $\rho(a(X), b(Y))$, over certain classes of functions $a(x)$ and $b(y)$. A classical reference is

Goodman and Kruskal (1979); see also *Multivariate Analysis: Discrete Variables (Overview)*.

A related notion is the canonical decomposition of a bivariate distribution into weighted sums of independent-like components (see Lancaster 1969).

4. Methods for Generating Multivariate Models

Many newer multivariate distributions have been developed to model data where the multivariate normal distribution does not provide an adequate model. Multivariate distributional modeling is inherently substantially more difficult in that both marginal distributions and joint dependence structure need to be taken into account. Rather straightforward univariate properties or characterizations often have many multivariate versions; however, numerous multivariate concepts have no univariate analogues. Some of the more commonly used approaches to developing models are given in the following.

4.1 Characterizing Properties

One common approach to generate multivariate distributions is to develop extensions of a specific univariate property. For instance, symmetry of the univariate distribution, that is, $f(x) = f(-x)$, can lead to the family of reflection symmetric or multivariate spherical distributions. Here $f(x_1,\ldots, x_p) = g(|x_1|,\ldots, |x_p|)$ or $f(x_1,\ldots, x_p) = g(x_1^2+\cdots+x_p^2)$ for a suitable function g. The latter family, especially when incorporating covariates, naturally leads to the class of elliptically contoured multivariate distributions (e.g., Fang and Anderson 1990).

The lack-of-memory property, which characterizes the univariate exponential distribution, has been generalized to the multivariate case, and this characterization yields the Marshall–Olkin multivariate exponential distribution. Because this distribution has both a singular and a continuous part, other purely continuous multivariate distributions with exponential marginals have been developed.

Other properties that have been used to develop multivariate distributions are independence of certain statistics, identical distribution of statistics, closure properties, limiting distributions, differential equation methods, series expansions, and conditioning methods. In some instances physical models, such as shock models, provide a method for extending univariate distributions to the multivariate case.

4.2 Mixture Methods

Mixture methods have provided successful extensions for models with non-negative random variables. The introduction of a dependence structure can be accomplished by various notions of mixing. For example, if U, V, W are independent random variables; then the 'mixtures' (a) $X_1 = U+V$, $X_2 = U+W$, (b) $X_1 = \min(U, V)$, $X_2 = \min(U, W)$, (c) $X_1 = U/V$, $X_2 = U/W$ generate dependent bivariate random variables. In particular, a bivariate binomial distribution, a bivariate exponential distribution, and a bivariate Pareto or F distribution can be so generated. An inherent problem with multivariate versions obtained from mixture methods, as well as other generating techniques, is that the number of parameters so introduced can be as large as 2^p, thereby necessitating large sample sizes for estimation. For a comprehensive treatment of bivariate distributions see Hutchinson and Lai (1990).

4.3 Copulas

Another approach to providing non-normal multivariate models is based on copulas or uniform representations. We provide details for bivariate random variables; the more general case is discussed in Nelsen (1999) and Joe (1997). This approach first requires specification of the univariate marginal distribution functions $F_1(x)$, $F_2(y)$, and then the joint distribution is $F(x, y) \equiv U(F_1(x), F_2(y))$, where $U(s, t)$ is a two-dimensional copula. There are a number of technical conditions that the function $U(s, t)$ must satisfy, but the basic idea is that $U(s, t)$ is essentially the distribution function of a two-dimensional random vector whose every component marginally has a uniform distribution on [0, 1]. There are many different families of copulas; for the bivariate case, examples of copulas include the standard normal copula,

$$\text{(a) } U_N(s, t) = (1-\rho^2)^{-1/2} \exp[-\{\rho^2(u^2+v^2) - 2\rho uv\}/(2(1-\rho^2))]$$

where $u = \Phi^{-1}(s)$, $v = \Phi^{-1}(t)$ and Φ^{-1} is inverse standard normal distribution function; and the Farlie–Gumbel–Morgenstern copula,

$$\text{(b) } U_{FGM}(s, t) = 1+\theta(2s-1)(2t-1).$$

Based upon copulas like these, different multivariate distributions can be generated. For instance, the standardized bivariate normal distribution function with correlation ρ can be expressed as $U_N(\Phi(x), \Phi(y))$, whereas $U_N(1-e^{-x}, 1-e^{-y})$ has normal-like dependence structure but standard exponential distributions as its marginals. A fairly standard technique for creating copulas is to begin with any joint distribution $F(x, y)$ and then compute $F(F_1^{-1}(x), F_2^{-1}(y))$, which under nonstringent technical conditions will be a copula.

Every bivariate copula $U(s, t)$ with $0 \leqslant s, t \leqslant 1$ satifies $U^-(s, t) \equiv \min(s, t) \leqslant U(s, t) \leqslant \max(s+t-1, 0) \equiv U^+(s, t)$. Both $U^-(s_1, s_2)$ and $U^+(s_1, s_2)$ themselves

are copulas, and are known as, respectively, the lower and upper Fréchet–Hoeffding bounds.

The probabilistic theory concerning copulas has been fairly extensively developed, but the applications to data analysis are much more limited. Copulas can be readily extended to higher dimensions. However, bounds become more complex. For more details see Rüschendorf et al. (1996).

5. *Multivariate Probability Computational Techniques*

5.1 *Multivariate Bounds for Probabilities*

Inequalities or bounds for multivariate probabilities are important tools in simultaneous inference and for approximating probabilities. We consider three basic types of inequalities, noting that for each there are numerous generalizations and extensions.

Perhaps the oldest method is based on Boole inequalities which provide relations between probabilities of events. The Fréchet–Hoeffding bounds can be viewed as extensions of Boole's inequality. In a statistical context, the lower bound of the Boole inequalities are called Bonferroni inequalities, and are often used when dealing with multiple comparisons. These bounds have many extensions; see Hochberg and Tamhane (1987), Galambos and Simonelli (1996), and *Hypothesis Tests, Multiplicity of*.

A second set of inequalities are called Chebyshev inequalities. For bivariate random variables X_1, X_2 with $EX_i = \mu_i$, $\mathrm{Var}(X_i) = \sigma_i^2$ and correlation

$$\rho,\ P\{|X_1-\mu_1| \geqslant \sqrt{k_1}\sigma_1,\ |X_2-\mu_2| \geqslant \sqrt{k_2}\sigma_2\} \leqslant \frac{k_1+k_2+\sqrt{(k_1+k_2)^2-4\rho^2k_1k_2}}{2k_1k_2}$$

thereby extending the classical Bienaymé–Chebyshev univarite inequality.

Another class of inequalities are multiplicative inequalities that hold for certain classes of random variables, namely,

$$P\{X_1 \in A_1,\ldots, X_m \in A_m\} \geqslant \prod_{i=1}^{m} P\{X_i \in A_i\}$$

When $X_1,\ldots, X_m$ are independent normal variates with respective means $\mu_1,\ldots, \mu_m$, and common variance σ^2, then

$$P\{\cap\, |X_i-\mu_i| \geqslant k_i s\} \geqslant \prod_{1}^{m} P\{|X_i-\mu_i| \geqslant k_i s\}$$

where s^2 is an independent estimate of σ^2. This inequality yields a bound for the joint probability of a simple multivariate t-distribution in terms of probabilities of univariate t-distributions.

When $X_1,\ldots, X_m$ are correlated normal variates with a correlation matrix $R = (\rho_{ij})$ with $\rho_{ij} \geqslant 0$, then the one-sided orthant is bounded

$$P\{\cap\, (X_i \leqslant a_i)\} \geqslant \prod_{1}^{m} P\{X_i \leqslant a_i\}$$

The two-sided version of the preceding holds for $|X_i|$ without any non-negativity restrictions on the correlations.

For discussions of multivariate inequalities see Dharmadhikari and Joag-dev (1988) and Tong (1980).

5.2 *Approximations of Probabilities*

There are a number of different numerical and stochastic techniques that are in use to obtain numerical approximations to multivariate probability statements. These techniques are useful in a variety of circumstances, for example, computing p-values for complicated multivariate data, obtaining power curves for multivariate tests, and computing multivariate posterior distributions.

Basic techniques for multivariate probability approximations rely on standard numerical integration principles; here $P\{\mathbf{x} \in A\} = \int\cdots\int_{\mathbf{x}\in A} f(\mathbf{x})d\mathbf{x}$ is approximated by $\sum_{\{x_i\}} w_i f(\mathbf{x}_i)$, where w_i are weights and $\{x_i\}$ is a finite set of points specified in A. Elementary Monte Carlo integration techniques involve random sampling N points x_i, $i = 1,\ldots, N$ according to $f(\mathbf{x})$ and estimating $P\{A\}$ by $N^{-1} \sum_{i=1}^{N} I(\mathrm{x}_i \in A)$, where $I(\mathbf{x}_i \in A) = 1$, if $\mathbf{x}_i \in A$, and $= 0$, otherwise. Some methods for computing multivariate normal probabilities first transform the random variables, and rewrite $P\{A\} = \int\cdots\int_{\|\mathbf{z}\|=1} F(\mathbf{z})dz$, where $F(\mathbf{z})$ is a function defined by Σ and the region A. Then $P\{A\}$ is estimated by $N^{-1}\sum_{i=1}^{N} F(\mathbf{z}_i)$ where the $\mathbf{z}_i$ are randomly chosen from the surface of a sphere in p-dimensional space.

Monte Carlo importance sampling techniques assume a convenient density $g(\cdot)$ whose support includes the support of the density $f(\cdot)$, and $g(\cdot)$ is chosen to expedite convergence. Then points $\mathbf{y}_1,\ldots, \mathbf{y}_N$ are random samples according to $g(\mathbf{y})$ and $P\{A\}$ is estimated by $\mathrm{N}^{-1}\sum_{i=1}^{N} I(\mathbf{y}_i \in A)[f(\mathbf{y}_i)/g(\mathrm{y}_i)]$.

More recently, Markov chain Monte Carlo methods have been developed and have proven to be particularly useful for high dimensional random vectors. The common aspects of these types of techniques is the definition of a Markov chain $\mathbf{x}_1$, $\mathbf{x}_2,\ldots$ which converges to a stationary Markov chain where $\mathbf{x}_i$ has marginal distribution $f(\mathbf{x})$. Again, $P\{A\}$ is approximated by $(N-m+1)^{-1} \sum_{i=m}^{N} I(\mathbf{x}_i \in A)$, where $\mathbf{x}_m,\ldots, \mathbf{x}_N$ is a suitably chosen sequence of elements generated by the Markov chain. For much more complete details, see *Monte Carlo Methods and Bayesian Computation: Importance Sampling*; *Markov Chain Monte Carlo Methods*; *Monte Carlo Methods and Bayesian Computation: Overview*.

6. Structures and Patterns

6.1 Principal Components

A correlation matrix represents the interdependencies among p measures, which may be likened to a connected network. The removal of one of two closely connected variables (that is, highly correlated) takes no account of how these variables are connected to the remaining measures. Principal components introduced by Harold Hotelling in 1933 (see *Hotelling, Harold (1895–1973)*, methodology that transforms the original correlated standardized variables $z_1,\ldots, z_p$ into new variables $z_1^*,\ldots, z_p^*$ (called principal components), which are uncorrelated. These new variables are linear combinations of the original variables. The variance of the standardized z_i is 1, and denote the variance of z_i^* by τ_i^2; then the total variance $p = \Sigma\tau_i^2$ remains fixed for the two systems of variables. By ordering the variances $\tau_{(1)}^2 \geqslant \tau_{(2)}^2 \geqslant \cdots \geqslant \tau_{(p)}^2$, and examining the cumulative proportions $\tau_{(1)}^2/p$, $(\tau_{(1)}^2+\tau_{(2)}^2)/p,\ldots, (\tau_{(1)}^2+\cdots+\tau_{(p)}^2)/p = 1$, we obtain what is often called the 'proportion of variance explained by the principal components.' In this way we may obtain a subset of principal components that explain most of the variance, thereby providing a parsimonious explanation of the data. In effect, principal components provides a method of data reduction. The challenge in using this approach is deciding how many principal components are reasonably required and how to interpret each component.

6.2 Canonical Variables

As noted, principal components provides a procedure for the reduction of a large number of variables into a smaller number of new variables. When the variables fall into two natural subsets, as for example production and sales variables, or multiple measurements on two siblings, then canonical analysis provides another exploratory data reduction procedure. If the two sets are labeled $X_1,\ldots, X_p$ and $Y_1,\ldots, Y_q$ $(p \leqslant q)$, canonical analysis attempts to simplify the correlational structure consisting of $(p+q)$ $(p+q-1)/2$ correlations to a simpler correlational structure of only p correlations.

This reduction is accomplished by forming linear combinations (U_i, V_i), $i = 1,\ldots, p$, of the X_i's and Y_j's, respectively, with the property that all the variances are 1, $\rho(U_i, V_i) = \rho_i$ and all other correlations are zero. The new variables are called canonical variables, and the correlations ρ_i are called the population canonical correlations.

This procedure is equally applicable to sample data. Often the first few sample canonical correlations are large compared to the remaining ones, in which case the first few canonical variables provides a data reduction procedure, which captures much of the structure between the two sets of variables.

6.3 Latent Structure and Causal Models

Latent structure models refers to a set of models that attempts to capture an understanding of causality, and hence are sometimes referred to as causal models. The term is not well-defined and at its broadest includes factor analysis, path analysis, structural equation models, correspondence analysis, loglinear models, and multivariate graphical analysis. One feature of some of these models is the description of a set of observables in terms of some underlying unobservable random quantities. Depending on the context, these unobservable variables have been called latent variables, factors, or manifest variables. See also *Factor Analysis and Latent Structure: Overview*.

Graphical models often provide a visual understanding of relationships. Their origin arose in a number of scientific areas: as path analysis from genetic considerations, as interactions of groups of particles in physics, as interactions in multiway contingency tables. They are particularly useful in multivariate models because variables are related in a variety of different ways, as, for example, conditionally, longitudinally, or directly. Graphical models can be used in conjunction with specific models, for example, factor analysis, and structural equation models. For details concerning graphical models see Lauritzen (1996) and *Graphical Models: Overview*.

6.3.1 Factor analysis. Factor analysis is one of the oldest structural models, having been developed by Spearman in 1904. He tried to explain the relations (correlations) among a group of test scores, and suggested that these scores could be generated by a model with a single common factor, which he called 'intelligence,' plus a unique factor for each test.

Factor analysis has been used in two data analytic contexts: in a confirmatory manner designed to confirm or negate the hypothesized structure, or to try to discover a structure, in which case the analysis is called exploratory. See also *Latent Structure and Casual Variables*; *Factor Analysis and Latent Structure, Confirmatory*; *Factor Analysis and Latent Structure: IRT and Rasch Models*; *Factor Analysis and Latent Structure: Overview*.

6.3.2 Path analysis. The origins of path analysis is attributed to Sewall Wright, who from 1918 to 1934 developed a method for studying the direct and indirect effects of variables. Herein, some of the variables are often taken as causes, whereas others are taken as effects. However, because the linear relations exhibit only correlations, the view of causality remains hypothetical.

Path diagrams are typical graphical devices to show the direct and indirect variables. For example, in a study on the relation between performance and sat-

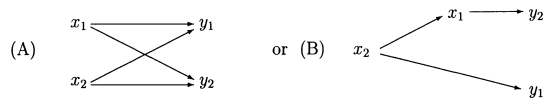

Figure 1
Examples of path structures

isfaction, let x_1 be achievement motivation, x_2 be verbal intelligence, y_1 be performance, and y_2 be job satisfaction. There are several path diagrams that could exhibit these connections.

In Fig. 1A, variation in x_1 and x_2 affect both outcomes y_1 and y_2, whereas in Fig. 1B, verbal intelligence affects performance directly, and affects job satisfaction through achievement motivation. Additionally, see *Latent Structure and Casual Variables*.

6.3.3 Linear structural equations (LISREL). Structural equation models, or econometric models, were developed early on to provide explanations of economic measures. Variables whose variability is generated outside the model are called exogenous and variables explained by exogenous variables or other variables in the model are called endogenous.

A standard basic model is, for vectors $\boldsymbol{x} = (x_1,\ldots, x_q)$ and $\boldsymbol{y} = (y_1,\ldots, y_p)$ of observable variables, $y = \boldsymbol{\eta}\Lambda_1 + \boldsymbol{\varepsilon}$, $\boldsymbol{x} = \boldsymbol{\xi}\Lambda_2 + \boldsymbol{\delta}$, connected by $\boldsymbol{\eta}\Delta = \boldsymbol{\xi}\Gamma + \boldsymbol{\zeta}$ where Λ_1, Λ_2 are, respectively, $m \times p$ and $q \times n$ matrices of regression coefficients; $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are vectors of unobservable dependent variables, Δ is an $m \times m$ regression matrix of the effects of endogenous on endogenous variables, Γ is an $m \times n$ regression matrix of the effects of exogenous on endogenous variables, and $\boldsymbol{\varepsilon}$, $\boldsymbol{\delta}$, and $\boldsymbol{\varepsilon}$ are residual vectors. For more details see *Latent Structure and Casual Variables*.

6.4 Clustering

Often, subsets of individuals are grouped by a known characteristic, as for example, by sex, by race, by socioeconomic status. In other instances, prespecified grouping characteristics may not be available. This is analogous to thinking of a galaxy of stars to be separated into groups. The methodology of clustering may be thought of as a data reduction procedure, where the reduction is to group individuals. This is in contrast to principal components or canonical analysis, in which the reduction is based upon combining variables. Alternatively, we may think of clustering as creating a classification or typology of individuals or items. For example, we may think of creating regions based on a set of demographic variables, or a set of products based on how they are used.

Suppose that the data consists of p measurements taken on each of n individual (objects). For each pair of individuals, we form a 'distance' measure, where 'distance' is flexibly defined. Thus we create an $n \times n$ matrix of 'distances' for which there are a variety of algorithms that govern the creation of subgroups, in part because there is no universal definition of a cluster, nor a model that defines a cluster. Intuitively, we think of a cluster as being defined by points that are relatively 'closer' to one another within the cluster, and 'further' from points in another cluster. See *Statistical Clustering* for more details.

6.5 Classification and Discrimination

Classification or discriminant analysis is another classically important problem in which multivariate data is traditionally reduced in complexity. Suppose that there are k populations of individuals, where for each individual we observe p variables $(X_1,\ldots, X_p)$. Based upon historical or 'training' data where we observe a sample of n_i individuals from population i, $i = 1,\ldots, k$, we want to develop a method to classify future individuals into one of the k populations based upon observing the p variables for that new individual. Although fairly general procedures have been developed, it has been traditional to assume that for population i, $\mathbf{X} \sim N(\boldsymbol{\mu}_i, \Sigma)$, $i = 1,\ldots, k$. Then $\boldsymbol{\mu}_1,\ldots, \boldsymbol{\mu}_k$, Σ are estimated from the historical or training data. For a new individual, with data $\mathbf{Z}$ we want to optimally classify that individual into exactly one of the k populations.

The case $k = 2$ yields a linear classifier involving $\mathbf{Z}$ called Fisher's linear discriminant function, and the case $k > 2$ involves k linear functions of $\mathbf{Z}$.

To handle settings where we do not assume underlying normal populations, there is the technique of logistic discriminant analysis, which again results in a linear discriminant function for two populations. More recent techniques allow nonlinear discriminant functions and nonparametric discriminant functions having a monotonicity property. The classification and regression-tree (CART) approach yields a branching decision tree for classification; see Breiman et al. (1984). Decision tree analyses have found many uses in applications, e.g., for medical diagnostics. For further information on discrimination see *Multivariate Analysis: Classification and Discrimination*.

6.6 Multidimensional Scaling

From a geographical map, one can determine the distances between every city and every other city. The technique, called multidimensional scaling (MDS), is concerned with the inverse problem: Given a matrix of distances, how can we draw a map? The distances are usually in p-dimensional space, and the map is in $q < p$ dimensional space, so that MDS can be thought of as a dimension reduction technique. Further, $q = 2$ is the most natural dimension for a map. However, we know from the many projection procedures for drawing a map of the globe ($p = 3$) in $q = 2$ dimensions that the resulting maps differ in how one perceives distances. For more details see *Scaling*: *Multidimensional*.

6.7 Data Mining

Data mining refers to a set of approaches and techniques that permit 'nuggets' of valuable information to be extracted from vast and loosely structured multiple data bases. For example, a consumer products manufacturer might use data mining to better understand the relationship of a specific product's sales to promotional strategies, selling store's characteristics, and regional demographics. Techniques from a variety of different disciplines are used in data mining. For instance, computer science and information science provide methods for handling the problems inherent in focusing and merging the requisite data from multiple and differently structured data bases. Engineering and economics can provide methods for pattern recognition and predictive modeling. Multivariate statistical techniques, in particular, clearly play a major role in data mining.

Multivariate notions developed to study relationships provide approaches to identify variables or sets of variables that are possibly connected. Regression techniques are useful for prediction. Classification and discrimination methods provide a tool to identify functions of the data that discriminate among categorizations of an individual that might be of interest. Another very useful technique for data mining is cluster analysis that groups experimental units which respond similarly. Structural methods such as principal components, factor analysis, and path analysis are methodologies that can allow simplification of the data structure into fewer important variables. Multivariate graphical methods can be employed to both explore databases and then as a means for presentation of the data mining results.

A reference to broad issues in data mining is given by Fayyad et al. (1996). Also see *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)*.

7. Further Readings

This article and related entries provide only a brief overview of a vast field. For further general readings see Anderson (1984), Dillon and Goldstein (1989), Johnson and Wichern (1999), Morrison (1990), Muirhead (1982); for detailed developments see the *Journal of Multivariate Analysis*.

Bibliography

Agresti A 1984 *Analysis of Ordinal Categorical Data*. Wiley, New York

Anderson T W 1984 *An Introduction to Multivariate Statistical Analysis*, 2nd edn. Wiley, New York

Breiman L, Friedman J H, Olshen R A, Stone C J 1984 *Classification and Regression Trees*. Wadsworth, Belmont, CA

Dharmadhikari S W, Joag-dev K 1988 *Unimodality, Convexity, and Applications*. Academic Press, Boston

Dillon W R, Goldstein M 1989 *Multivariate Analysis: Methods and Applications*. Wiley, New York

Fang K-T, Anderson T W (eds.) 1990 *Statistical Inference in Elliptically Contoured and Related Distributions*. Allerton Press, New York

Fayyad U M, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds.) 1996 *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence/MIT Press, Cambridge, MA

Galambos J, Simonelli E 1996 *Bonferroni-type Inequalities with Applications*. Springer-Verlag, New York

Goodman L, Kruskal W 1979 *Measures of Association for Cross Classifications*. Springer-Verlag, New York

Hochberg Y, Tamhane A C 1987 *Multiple Comparison Procedures*. Wiley, New York

Hotelling H 1953 New light on the correlation coefficient and its transform. *Journal of the Royal Statistical Society, Series B* **15**: 193–225

Hutchinson T P, Lai C D 1990 *Continuous Bivariate Distributions, Emphasizing Applications*. Rumsby Scientific Publishing, Adelaide, South Australia

Joe H 1997 *Multivariate Models and Dependence Concepts*. Chapman & Hall, London

Johnson R A, Wichern D W 1999 *Applied Multivariate Statistical Analysis*, 4th edn. Prentice-Hall, New York

Lancaster H O 1969 *The Chi-squared Distribution*. Wiley, New York

Lauritzen S L 1996 *Graphical Models*. Oxford University Press, Oxford, UK

Mathai A M, Provost S B, Hayakawa T 1995 *Bilinear Forms and Zonal Polynomials*. Springer Verlag, New York

Morrison D F 1990 *Multivariate Statistical Methods*, 3rd edn. McGraw-Hill, New York

Muirhead R J 1982 *Aspects of Multivariate Statistical Theory*. Wiley, New York

Nelsen R B 1999 *An Introduction to Copulas*. Springer, New York

Rüschendorf L, Schweizer B, Taylor M D (eds.) 1996 *Distributions with Fixed Marginals and Related Topics*. Lecture Notes – Monograph Series, Vol. 28, Institute of Mathematical Statistics, Hayward, CA

Tong Y L 1980 *Probability Inequalities in Multivariate Distributions*. Academic Press, New York

Tong Y L 1990 *The Multivariate Normal Distribution*. Springer-Verlag, New York

I. Olkin and A. R. Sampson

Mumford, Lewis (1895–1990)

Lewis Mumford was one of the seminal cultural critics to emerge from the United States during the twentieth century. Although known chiefly for his work in architecture and urbanism, he made significant contributions to art and literary criticism, biography and autobiography, the history of technology, sociology, and political science. Largely self-taught, Mumford

defined himself as a 'generalist,' one who crossed disciplines and epochs in pursuit of an 'organic' equilibrium in human affairs. This search for 'eutopia' or the good place—to use the archaic spelling and meaning that he preferred—preoccupied him over the course of a writing career that spanned an astonishing six decades.

1. *Life and Work*

1.1 *Background and Intellectual Development*

Born on October 19, 1895 in Flushing, Long Island (now part of New York City), Mumford was reared in a middle-class, albeit fatherless, household on the Upper West Side of Manhattan. His youth was unremarkable, save for a series of extended walks around New York City he undertook in the company of his stepgrandfather. He was graduated from Stuyvesant High School in 1912, but ill health forced him to withdraw from the City College of New York before completing his baccalaureate degree. Rather than languish, Mumford undertook a regimen of self-study and self-improvement under the direct influence of Sir Patrick Geddes, the Scottish biologist, sociologist, and townplanner whose ideas Mumford first encountered while enrolled in a college biology course. Mumford began corresponding with Geddes in 1918, and the latter acted as an intellectual mentor and surrogate father to the young man.

Geddes, a remarkable polymath who disregarded disciplinary boundaries, advanced a Spencerian view of evolutionary change in society based upon the integrated study of a given region, including its physical surroundings, the population that lived there, and the work they performed. Over the next several years, Mumford heeded Geddes's call for 'regional survey,' the collection of detailed topographical and architectural data, beginning with the New York metropolitan area and, as his health improved, continuing up and down the eastern seaboard of the US. This activity, so reminiscent of Mumford's childhood walks, helped to focus his interest in the study of architecture and city planning, while still embracing Geddes's emphasis on the social, economic, and cultural aspects of regional development.

At about the same time, Mumford discovered Ebenezer Howard's garden city, which, in its compact form and its balance between urban and rural modes of living, dovetailed almost seamlessly with Geddes's regionalism. Among the numerous other writers who influenced Mumford's development at this early stage were the popular novelist H. G. Wells, the playwright George Bernard Shaw, the Arts and Crafts polemicists John Ruskin and William Morris, and the American Transcendentalists Ralph Waldo Emerson and Henry David Thoreau. By 1919 Mumford had established himself as a promising journalist and editor in New York and, briefly, in London. In 1922, he published his first book, *The Story of Utopias*, a survey of utopian literature in which many of the overarching themes of his career were adumbrated: the investment of artistic form with social meaning; the harnessing of technology for the common good; and the restructuring of the built environment on a human scale in balance with the surrounding region. That same year he married Sophia Wittenberg (1899–1997), who was to be his life-long intellectual confidante and amanuensis.

Mumford's subsequent career can be divided roughly into three chronological stages: an early stage during which he concentrated on American cultural history and criticism (c. 1923–31); a mature stage when his focus broadened to include all of Western civilization (c. 1932—70); and a late stage when he reexamined his life and work in a series of autobiographical collections of writings, leading to an incomplete autobiography (c. 1971–82).

1.2 *Early Career*

The 1920s found Mumford in the company of architects and planners as well as writers. In 1923, he joined the Regional Planning Association of America, a group dedicated to the improvement of housing, the development of garden cities, and the promotion of regionalism. The Association achieved modest successes in their planned communities at Sunnyside Gardens in the New York City borough of Queens and at Radburn in suburban New Jersey. The Appalachian Trail, a 2,000-mile wilderness trail along the mountain spine of the eastern US is the Association's most lasting achievement. Although he was not a designer, Mumford acted effectively as the Association's secretary and publicist.

Mumford continued to earn his living as journalist, writing chiefly criticism on a vast range of subjects for such periodicals as the *Freeman,* the *Journal of the American Institute of Architects,* the *American Mercury, Commonweal,* and the *New Republic*. Under the influence of his close friend and fellow writer, Van Wyck Brooks, Mumford published a series of short, but insightful books on American culture in an effort to recover what Brooks termed the 'usable past.' *Sticks and Stones* (1924) outlines what was at the time an unorthodox social history of American architecture. *The Golden Day* (1926) examines five writers from the mid-nineteenth century—Emerson, Thoreau, Nathaniel Hawthorne, Herman Melville, and Walt Whitman—whom Mumford believed had created a specifically American literature in touch with the regional landscape and who might be emulated by his contemporaries. Mumford's first foray into biography, *Herman Melville* (1929), was a direct outgrowth of *The Golden Day* and one of the first studies to wrestle with Melville's troubled psyche. *The Brown Decades* (1931) culminated this early period, and it

was Mumford's attempt to redress some critical oversights in both *Sticks and Stones* and *The Golden Day*, particularly the cursory treatment of architects H. H. Richardson, Louis Sullivan, and Frank Lloyd Wright. Although surpassed by later, more scholarly works, these four books made a significant contribution to the fledgling discipline of American Studies. Moreover, with the unifying color metaphor of its title and the elegance of its prose, *The Brown Decades* endures as a major work of American letters.

1.3 Mature Career

The mature stage of Mumford's career extended from the early 1930s to the early 1970s, with only a brief interruption around World War II. During the Depression, Mumford continued to make his living as a journalist, but he confined his output to a few selected journals and magazines. In 1931, he began an extraordinary 30-year tenure as the architecture critic for the *New Yorker*. His column, 'The Sky Line,' was an important platform from which he mediated between the often conflicting currents of European and American modernism. Many of his architecture columns have been collected and published in *From the Ground Up* (1956), the *Highway and the City* (1963), and *Sidewalk Critic* (1998). Virtually unknown today is his short but influential tenure as the magazine's art critic (1932–7). From cubism to surrealism, Mumford embraced most currents of modern art as they took tentative root in the museums and galleries of New York City and elsewhere, although he turned a blind eye to nonrepresentational experimentation. That artists might lead a cultural renewal in post-World War II society forms the major theme of *Art and Technics* (1952), a published collection of lectures delivered at Columbia University.

Mumford, however, wanted to be known primarily as a writer of important books. Fresh from his modest success with the American 'usable past' project, he embarked on an ambitious three-part—later expanded to a four-part—study of Western civilization that became known as *The Renewal of Life*. Inspired by Geddes, the series surveys medieval and postmedieval European history and culture from distinct perspectives utilizing an implicitly Christian paradigm. In Mumford's view, the Middle Ages represented an organic unity that was disrupted by the industrial revolution and its dehumanizing effects. Modern technology, coupled with rational planning and cultural renewal, he believed, could restore that lost unity in the present.

The first volume, *Technics and Civilization* (1934) examines technological developments from the turret windmills of sixteenth-century Holland to the hydroelectric dams of the Tennessee Valley Authority. Only through careful control, Mumford warns, can technology continue to improve the quality of modern life. *The Culture of Cities* (1938), the second and most widely acclaimed volume of *The Renewal of Life*, traces concurrent urban developments, concluding with an eutopian vision of a balanced regional landscape. Following the book's publication, Mumford was regarded internationally as an authority on urbanism, his status reified by the placement of his photograph on the cover of *Time* (April 18, 1938). This success of this book in turn led Mumford to publish *City Development* (1945), a collection of his planning reports and essays.

The last two volumes of The Renewal of Life, *The Condition of Man* (1944) and *The Conduct of Life* (1951), were met by less critical and popular success as they were in some ways eclipsed by the events of the world war that they bracketed. With tempered optimism, Mumford in these volumes surveys the impact of politics, science, philosophy, and religion on the human personality and argues that the renewal of the individual is the necessary prelude to the renewal of society as a whole. Spread over a period of 17 years, The Renewal of Life as a series had a less-lasting impact than such contemporary multivolume studies of Western civilization as Oswald Spengler's *The Decline of the West* and Arnold J. Toynbee's *A Study of History*, even though its first two volumes remain landmark studies in their respective areas. World War II transformed Mumford on a personal and professional level, while raising his political profile significantly. Early to recognize the threat of both Fascism and Nazism, he was one of the first left-leaning American intellectuals to advocate US intervention in the war, a point which he argued forcefully and eloquently in *Men Must Act* (1939). *Faith for Living* (1940) offered a hopeful vision of a world freed from the totalitarian menace. Tragedy struck Mumford and his wife Sophia in 1944 when their only son—named, significantly, Geddes—was killed in action on the Italian front. Mumford channeled his grief into the writing of *Green Memories* (1947), a moving biography of his son that opens a fascinating window into the complexities of parenting in the middle of the twentieth century. Alarmed by the US detonation of the atomic bomb and its potential for global destruction, Mumford, the interventionist, became—once the threat of totalitarianism had passed—Mumford the pacifist. He became an ardent foe of the arms build-up of the Cold War, as advances in technology once again threatened to overwhelm human reason. *In the Name of Sanity* (1954) contains many of his essays on this topic. This same pacifism lay at the core of his vocal and written opposition to the Vietnam War during the 1960s.

Dissatisfied with the reception of The Renewal of Life, Mumford spent the next two decades revising basic assumptions and searching further into the past for clues into modern behaviors. *The Transformations of Man* (1956) was his attempt to condense the core message of The Renewal of Life into a single volume. His conclusion that 'One World man'—one who

recognizes and synthesizes the rich traditions of all cultures around the globe—was about to emerge in the latter part of the twentieth century has proven to be both prescient and elusive. This reevaluation of past themes continued with *The City in History* (1961), Mumford's best-known book and the winner of the National Book Award for nonfiction in 1962. What he intended as a mere revision of *The Culture of Cities*, gradually evolved into an expansive, mostly new survey of Western urbanism, beginning with the Paleolithic cave and concluding with the sprawling twentieth-century metropolis. Missing was his earlier, eutopian vision of a new regional order; instead he forecast the inevitable decline of cities unless technology, especially the automobile, was brought under control by designers and planners. The book solidified Mumford's reputation as America's senior authority on urbanism, although his presumptions regarding the efficacy of rational planning on city design were soon challenged by younger writers such as Jane Jacobs. The success of *The City in History* spurred Mumford to publish *The Urban Prospect* (1968), yet another collection of his essays on urbanism. Building on the momentum of *The City in History*, and at a stage in life when many of his contemporaries would have been content to garner praise and awards, Mumford reposited his views on the history of technology in what may be his most ambitious work, the two-volume *Myth of the Machine*. In volume one, *Technics and Human Development* (1967), he put forward what he believed was his most original insight: that the ancient rulers of the eastern Mediterranean world exerted insidious control over human labor through the employment of massive work units or 'megamachines.' Although he argued in volume two, *The Pentagon of Power* (1970), that modern technology made such human megamachines obsolete, the modern megamachines all-embracing power, left unchecked by human reason, would eventually lead to global destruction. Critical and popular reaction to *The Myth of the Machine* was mixed, but Mumford's critique of the political and scientific establishment, which he identified as masterminding the modern megamachine, won him wide admiration among a younger, more reform-minded generation.

1.4 Late Career

Mumford spent the last two decades of his life engaged in various activities related to the creation of his considerable public legacy. He donated his personal papers to the Van Pelt Library of the University of Pennsylvania in Philadelphia and his graphite, ink, and watercolor sketches—many of them dating to the regional surveys of his youth—to Monmouth College, now Monmouth University, in West Long Branch, New Jersey. Several collections of his earlier writings were published: *Interpretations and Forecasts: (1922–1972)* (1973); *Architecture as a Home for Man* (1975); and *The Lewis Mumford Reader* (1986). Two autobiographical collections—*Findings and Keepings* (1975) and *My Works and Days* (1979)—preceded the publication of his autobiography, *Sketches from Life* (1982), which covered only the first half of his life. In addition, Mumford's correspondence with several key friends and colleagues has been edited and published, providing a fuller understanding of his formidable but engaging personality. He died quietly at his country home in Leedsville near Amenia, New York on January 26, 1990.

2. Critical Assessment

Most of Mumford's books remain in print, and his name appears with increasing frequency in both popular and scholarly contexts. Nevertheless, the critical evaluation of his life and work is just beginning.

Figure 1
Lewis Mumford. Courtesy of the estate of Lewis and Sophia Mumford

Mumford's wide-ranging legacy as a generalist has been difficult for any one scholar to address; some have selected one aspect of his work, while others have joined together in delivering group assessments. Examples of the latter include the special issue of the American journal *Salmagundi* (Summer 1980) and Hughes and Hughes's *Lewis Mumford: Public Intellectual* (1990), a collection of essays that focus on Mumford's lifelong involvement with technology. As its title suggests, Miller's *Lewis Mumford: A Life* (1989) provides a comprehensive and often insightful overview of Mumford's life. Novak focuses more narrowly on one aspect of Mumford's writing in *The Autobiographical Writings of Lewis Mumford* (1988). In *Beloved Community*, Blake (1990) examines Mumford's relationship to the 'young Americans,' a group of New York intellectuals searching for cultural renewal in the aftermath of World War I. Mumford's contributions to architecture and urbanism are the subject of Wojtowicz's *Lewis Mumford and American Modernism* (1996), while Tschachler discusses Mumford in a German context in *Lewis Mumford's Reception in German Translation and Criticism* (1994). Numerous dissertations, moreover, have addressed one or more aspects of Mumford's work. As yet, however, little has been written about Mumford outside of the US even in Great Britain, where many of his books were published in separate editions.

See also: Architecture; Art and Culture, Economics of; Art History; Art, Sociology of; Civilization, Concept and History of; Civilizational Analysis, History of; Civilizations; Cultural Studies: Cultural Concerns; Cultural Studies of Science; Fine Arts; History of Technology; Political Science: Overview; Regional Planning, History of; Urban Life: Future Developments; Urban Places, Planning for the Use of: Design Guide; Urban Planning: Methods and Technologies; Urban Planning (Western), History of; Urban Policy: North America; Urban Studies: Overview; Utopias: Social

Bibliography

Blake C N 1990 *Beloved Community: The Cultural Criticism of Randolph Bourne, Van Wyck Brooks, Waldo Frank, and Lewis Mumford*. University of North Carolina Press, Chapel Hill and London

Hughes T P, Hughes A C (eds.) 1990 *Lewis Mumford: Public Intellectual*. Oxford University Press, New York

Miller D L 1989 *Lewis Mumford: A Life*. Weidenfeld and Nicolson, New York

Mumford L 1922 *The Story of Utopias*. Boni and Liveright, New York

Mumford L 1931 *The Brown Decades: A Study of the Arts in America, 1865–1895*. Harcourt, Brace and Company, New York

Mumford L 1934 *Technics and Civilization*. Harcourt, Brace and Company, New York

Mumford L 1938 *The Culture of Cities*. Harcourt, Brace and Company, New York

Mumford L 1944 *The Condition of Man*. Harcourt, Brace and Company, New York

Mumford L 1951 *The Conduct of Life*. Harcourt, Brace and Company, New York

Mumford L 1956 *The Transformations of Man*, 1st edn. Harper, New York

Mumford L 1961 *The City in History: Its Origins, Its Transformations, and Its Prospects,* 1st edn. Harcourt, Brace & World, New York

Mumford L 1967 *The Myth of the Machine: I. Technics and Human Development*. Harcourt, Brace & World, New York

Mumford L 1970 *The Myth of the Machine: II. The Pentagon of Power*. Harcourt Brace Jovanovich, New York

Mumford L 1973 *Interpretations and Forecasts: 1922–1972: Studies in Literature, History, Biography, Technics, and Contemporary Society,* 1st. edn. Harvest H B J (ed.) Harcourt Brace Jovanovich, New York

Mumford L 1982 *Sketches from Life: The Autobiography of Lewis Mumford: The Early Years*. Dial Press, New York

Mumford L 1986 *The Lewis Mumford Reader*. Miller D L (ed.) Pantheon, New York

Novak F G 1988 *The Autobiographical Writings of Lewis Mumford: A Study in Literary Audacity*. University of Hawaii Press, Honolulu

Tsachler H 1994 *Lewis Mumford's Reception in German Translation and Criticism*. University Press of America, Lanham, MD, New York and London

Wojtowicz R 1996 *Lewis Mumford and American Modernism: Eutopian Theories for Architecture and Urban Planning*. Cambridge University Press, Cambridge, UK

R. Wojtowicz

Music: Anthropological Aspects

Anthropological aspects of music include all those elements pertaining to music as a social and cultural phenomenon. Currently, research on these aspects is concentrated in ethnomusicology. Other disciplines, such as musicology, cultural studies, and popular music studies also increasingly take into account anthropological aspects, but have been slow in adopting more context-oriented approaches and ethnographic field methods. Anthropologists and social scientists in turn have generally shown only moderate interest in what the members of the societies they study hear, sing, or play. Moreover, many wrestle with the problem of how to integrate the study of music as a social phenomenon with music as sound. Within ethnomusicology there exists a considerable plurality of paradigms, models, and methodologies, but the majority of scholars have approached their respective (geographic or historical) area of study from of one or

several of the following perspectives: (a) music as culture and the mediation of social process; (b) music as discourse and the nature of musical meaning: (c) music and identity; and (d) music and modernization.

1. *Music as Culture and the Mediation of Social Process*

Ethnomusicologists in the past have tended to treat music as somehow separate from culture, almost as a thing in itself. They assumed that culture was separated into different spheres such as politics, economics etc., each of which could be brought into some sort of relationship with music. According to one approach inspired by British social anthropology, ethnomusicology was the 'study of music in culture' (Merriam 1964) with culture being little more than some ill-defined social context and music having the function of somehow maintaining the coherence of society. Another approach, modeled on the ideas of Ruth Benedict among others, saw each society as being constructed around a fundamental principle or pattern which was then imparted to various cultural realms such as religion, art, music, etc. Thus, Charles Keil, in a classic study of African music, found among the Tiv of Nigeria an 'expressive grid' that manifested itself in the circles, patterns, and structures of Tiv handicrafts, architecture, and song (Keil 1979).

More recently, scholars have begun to study music as culture. Influenced by the work of anthropologist Clifford Geertz and what has been called the 'interpretive turn' in the humanities and social sciences, researchers are increasingly thinking about music not so much as a secondary reflection of a more essential social structure or reality than itself being a social activity mediating other social processes. Instead of gaining its meaning by somehow becoming part of a context, music itself is a context through which other forms of activity, equally primary, achieve meaning (Herndon and McLeod 1981).

This new direction had two major consequences for the anthropological study of music. The first is that music makers and their music become dynamic forces, actively intervening in the shaping of cultures that are no longer seen as more or less coherent totalities or realities. The focus on culture as a heterogeneous field of conflicting forces shifts the attention from music as a product, a static expression of collective identity—as in 'the music of the Zulu,' 'music in village X,' etc.—to music as a process, as a mediating agent in the overall process of cultural negotiation. Or, as prominent British ethnomusicologist John Blacking put it, music was not only 'humanly organized sound,' but also 'soundly organized humanity' (Blacking 1973).

The second consequence concerns the nature of scholarly discourse and knowledge about 'other' musics. For the reformulated, critical disciplines truth could no longer be something that was be found from a neutral and distanced position. Following Edward Said's argument in his seminal book *Orientalism* (1979) the point of scholarship is not to deliver the 'true' Other. Difference can no longer denote the distance between cultures such as the West and the Orient, just as it no longer refers back to an overarching scholarly (presumably Western and objective) frame of reference. Rather, difference, in the new disciplines, means the interplay of a multitude of competing and juxtaposed discourses. An example of this approach is Martin Stokes's study of Turkish *arabesk* music (1992). This style of music popular with the marginalized urban masses is more than the result of the stylistic development of Turkish folk music. It is part of a 'cultural field' in which different musical styles are best understood as 'different ways of talking about music' (p. 50).

2. *Music as Discourse and the Nature of Musical Meaning*

The third consequence, finally, of the 'interpretive turn' in ethnomusicology is the idea that musical meaning is no longer seen as being intrinsic to its sonic structure—a viewpoint still prevalent in musicology. Rather musical meaning, like the meaning of a novel or a poem, materializes only in the act of experiencing music. Like a 'text' a piece of music must be 'read.' And just like the meanings of a written text are the result of a multiplicity of discourses that include the author's own voice, the meanings of a musical performance are constantly changing and emergent, depending on the social structures and power relationships they help to define. The shift in musical scholarship from structural analysis to the interpretation of music as discourse occurred parallel to another significant development in anthropology and sociology in the 1970s: the emergence of theories and ethnographies of performance. On the basis of Erving Goffman's work and Dell Hymes's pioneering essays on the ethnography of speaking, anthropologists like Victor Turner, Richard Bauman, and James Fernandez began to examine the way in which performance contexts, events and genres shaped communication and social action.

They were followed by numerous ethnomusicologists for whom the question of music's meanings and affective power became intimately tied to its real 'life,' its relations with other forms of behavior and expressive culture, the interaction between performers and their audiences, as well as the unpredictability and often improvisatory, playful nature of music making. But whereas scholars such as Turner stressed the communal, ordering functions of performance, researchers are now more interested in the potential of musical performance for prompting change and resistance.

In line with these developments, scholars have also increasingly turned to the human body as the primary site for the construction of musical meaning. While earlier studies were based on a notion of the body as a biological given and, hence, stressed the physiological aspects of music making, recent studies posit the body as a cultural phenomenon, no less susceptible to and dependent on meanings than other spheres of human practice (Muller 1999). In this light, future research will increasingly have to concentrate on specifically musical processes of socialization and the formation of selves.

3. *Music and Identity*

The waning appeal of discourses that take stable notions of music and culture for granted have led scholars to re-examine the relationship between identity and music in a new, rather more skeptical light. Ironically, it is because of the fact that the growing isolation of the individual in modern society has generated new forms of collective identification—new nations, 'neo-tribes' and new religions—that the time-honored interest in nationalism and ethnicity as the chief bases for music making has experienced something of a revival.

Of all the articulations of identity currently studied by music scholars the modern nation-state and various types of nationalist discourse have probably commanded the most attention. Although studies of small-scale, mostly rural communities remain important locales for the ethnographic study of music, it is the links between the nation-state, nationalism, and the various contestations and entanglements with the nation that have become the object of the most sustained scrutiny undertaken to date. A rich literature on the new popular musics of the 'Third World'—and to a lesser degree on the advanced countries of the West—demonstrates how musical practices often disrupt the putatively homogeneous times and spaces of the nation-state, but at other times exist quite comfortably alongside nationalist agendas. In most cases, however, both positions blend into each other, often producing new exoticisms and articulations of racial or ethnic exclusivity. Israeli rock influenced by Arabic music, for instance, celebrates an 'other' quite blatantly at odds with the images of backwardness that the Israeli majority projects of the Palestinian minority, yet at the same time takes a stable 'oriental' identity as its reference point (Regev 1986). Conversely, some musics have also been shown to lend themselves to more pliant forms of identity. This is perhaps more typical of the conflicted relationship between the nation-state, nationalist ideology and cultural practices in diasporic communities. Thus, the music of black Americans is a product of the slaves' encounter with the West and at the same time it has emerged in opposition to modernity (Gilroy 1993). It may thus serve as a model for what Gilroy calls anti-antiessentialism, a critique of the oppressive history and racial injuries of the 'host' country that refuses to let go of memories of collective identity even where it resists grounding them in a direct genealogy rooted in the African past.

In some contexts, the nation-state may also come to the rescue of cultural diversity. Tanzania, for instance, is one of several countries that have tried successfully to promote 'national' musics by reserving some 85 percent of the music content on its national broadcasting service for local music products. But here, as in many other cases, the definition of what counts as representative of Tanzanian culture rested in the hands of the Swahili-speaking minority dominating the state apparatus (Wallis and Malm 1992, pp. 113–14).

Race and ethnicity are perhaps the second most studied forms of identity in ethnomusicology today. Deeply intertwined with nationalism, but also often standing in opposition to nationalist discourses, racial and ethnic identities are frequently seen to be most effectively asserted in the symbolic domain, such as in music. There are many reasons for this close nexus between ethnicity and music, but one of them is the fact that assertions of racial pride are voiced from marginalized social groups whose access to other power bases such as economic wealth is limited.

Processes of class formation, by contrast, and the musical components of class consciousness have lost a great deal of their former appeal, primarily because the evidence gathered does not sustain the Marxist notion of a material base determining a suprastructure of ideology and culture. This lack of evidence was particularly pronounced in those areas where working-class cultures with pristinely defined contours failed to develop as in much of the 'Third World,' or where they were about to disintegrate as was the case in most of the advanced economies of post-World War II Europe and North America. By contrast, recent years have seen a growing interest in the formation of somewhat fuzzier identities and the permeation of cultures considered to have once been 'premodern' with bourgeois values and practices. Under the impact of A. Gramsci's notion of 'hegemony' and Pierre Bourdieu's concept of 'habitus,' ethnomusicologists are beginning to reassess colonization and Westernization as projects that entail domination and acquiescence and that do not always occur along neatly separated lines of class or racial identity.

The notion, finally, that all identities are also gendered identities has greatly invigorated anthropological research on music and dance (Cowan 1990). Although the idea has been familiar to ethnomusicologists for some time—much more so than to musicologists whose subject always seemed to be an unquestioned male one—recent scholarship has emphasized the negotiated character of sexual identity and the way in which music intervenes in the construction of such identities (Koskoff 1987). As a recent study of the music of Maghrebian immigrants in

France has shown, for example, *rai* performers and audiences construct complex identities around notions of sexuality, romantic love and adolescence that are inspired by Western liberal ideals of gender equality (Gross et al. 1996). At the same time these performers reproduce more traditional images of Arabness. While female singers often advocate more freedom for women, the jackets of their cassettes reproduce clichés of a more subdued kind of femininity and the ideal Arab woman.

Finally, an important focus that emerged from the current interest in identity and that undermines even further conventional notions of stable collective and individual identities—and the merely passive role of music in 'expressing' these—is the rather diverse set of practices circumscribed by terms such as diaspora, travel, migrancy, immigration, or displacement. In the recent past, all these practices have been dealt with to some extent by ethnomusicologists (Turino 1993), but musicologists and students of popular music have been less anxious to explore the interchange between music and the shifting boundaries of communities and audiences caused by large-scale movements of people. Furthermore, the spaces people traverse as tourists, migrants, refugees or touring musicians are all too often depicted as though they were bounded sites that are already filled with sound and much less as multiply overlapping contact zones in which identities and musics are always blurred and emergent.

4. *Music and Modernization*

Most scholars now agree that musics everywhere have been drawn into the maelstrom of modernization and, more often than not, have themselves become effective agents of modernization. While these scholars have as yet to arrive at a shared definition of modernization and modernity, it seems that the debate has centered on a series of intertwined issues and concepts. The first of these issues is globalization. Much of the discussion about globalization grew out of the concern with displacement and entailed a vigorous rethinking of the interrelationship between music and locality (Stokes 1994). While previous scholarship tended to assume a relatively static fit between place, spatial identity and music, an equation that often enabled scholars to produce richly detailed, empirically grounded narratives of local identity and music-making (Finnegan 1989), some researchers have gone one step further in examining locality as something that needs to be created in the first place. In a study of Nigerian *fúji* music and *oríkì* praise poetry, Karin Barber and Christopher Waterman (1995) show that while performers of these genres draw on a multiplicity of external sources to imagine densely local worlds, the emergent sense of locality is not simply stemming from any given body of materials or traditions. Rather, the 'ultimate goal of any performance is to intensify the presence, image and prospects of local actors' (p. 243). Although it cannot be ruled out that Yoruba musical productions of place may be driven by the same agendas that underlie movements of national and ethnic purity, Barber and Waterman insist on the hybrid underpinnings of such assertions of local identity.

Moreover, the spaces that scholars of popular music study and work in are no longer necessarily only such rather abstract entities as cities and countries. Increasingly, other spaces, such as studios, malls, music departments and taxis, are becoming sites of inquiry.

Another aspect of modernization that dominates current debates is the power of capitalism and commodification to transform 'traditional' cultures, and the capacity of the Western mass media to shape the collective imagination of large populations. Some critics such as French sociologist Jean Baudrillard have tended to think about global cultural flows in rather uniform terms, positing a regime of simulations and simulacra as an all-embracing system. In line with this reasoning, some music scholars have taken a rather sinister view of global musical production at the end of the twentieth century. Steven Feld draws on Gregory Bateson's term 'schismogenesis' to think through a set of escalating relationships marking the production and consumption of music on a global scale. Schismogenesis is a process through which the interplay between essentially dissimilar but mutually appropriative actions may lead to a closer symbiotic interdependence of both sides, to the point that they may even become incapable of self-correction and caught up in closed circuits of repetition (Feld 1994). Ultimately, what results from this scenario is a gray out of a new type. As discourses of authenticity and difference assert themselves and activities of commercial appropriation in turn get more overt and outrageous, some kind of 'fusion of the parties for mutual business gain' becomes likely (p. 273).

Others scholars, by contrast, appear to privilege 'Third World' cultural practices with antihegemonic agency *per se*, stressing the potential of music to maintain the heterogeneity of the world's cultures in the face of Western media dominance (Lipsitz 1994). The majority of scholars, however, have taken a more nuanced position toward the role of culture in reflecting and mediating global relations of power, numerous theoretical interventions having emerged in recent years that seek to provide new conceptual tools and models with which to grasp the twists and turns of global musical flows (Manuel 1993). Mark Slobin, for instance, in an attempt to account for the impossibility of representing both global generality and local specificity, has depicted the interplay of various 'micromusics' in the West as a kind of 'interculture,' a constant oscillation between the pressures of the 'supraculture' and the contention of various subcultures (Slobin 1993).

Finally, a handful of studies have attempted to view the making of globally connected identities as a two-way process in which center and periphery are constantly confused, and in which the West is just as much an 'other' of Africa or Asia as Africans and Asians have been the 'others' of Europeans (Erlmann 1999). This includes greater attention to the agency of those (predominantly Western) 'global' figures such as Paul Simon, Peter Gabriel, Quincy Jones, and Mickey Hart—whose practices and discourses are more than just manifestations of some anonymous capitalist world system or the global music industry.

See also: Cultural Expression and Action; Folklore; Globalization and World Culture; Music as Expressive Form; Music Perception; Popular Culture

Bibliography

Barber K, Waterman C 1995 Traversing the global and the local: Fuji music and praise poetry in the production of contemporary Yoruba popular culture. In: Miller D (ed.) *Worlds Apart. Modernity through the Prism of the Local*. Routledge, London

Blacking J 1973 *How Musical is Man?* University of Washington Press, Seattle, WA

Cowan J K 1990 *Dance and the Body Politic in Northern Greece*. Princeton University Press, Princeton, NJ

Erlmann V 1999 *Music, Modernity and the Global Imagination. South Africa and the West*. Oxford University Press, New York

Feld S 1982 *Sound and Sentiment. Birds, Weeping, Poetics and Song in Kaluli Expression*. University of Pennsylvania Press, Philadelphia, PA

Feld S 1994 From schizophonia to schismogenesis: On the discourses and commodification practices of 'world music' and 'world beat.' In: Keil C, Feld S (eds.) *Music Grooves*. University of Chicago Press, Chicago

Finnegan R 1989 *The Hidden Musicians: Music-making in an English Town*. Cambridge University Press, Cambridge, UK

Gilroy P 1993 *The Black Atlantic. Modernity and Double Consciousness*. Harvard University Press, Cambridge, MA

Gross J, McMurray D, Swedenburg T 1996 Arab Noise and Ramadan Nights: Rai, Rap, and Franco-Maghrebi Identities. In: Lavie S, Swedenburg T (eds.) *Displacement, Diaspora, and Geographies of Identity*. Duke University Press, Durham, NC

Herndon M, McLeod N 1981 *Music as Culture*. Norwood, Darby, PA

Keil C 1979 *Tiv Song*. University of Chicago Press, Chicago

Koskoff E 1987 *Women and Music in Cross-Cultural Perspective*. Greenwood Press, New York

Lipsitz G 1994 *Dangerous Crossroads: Popular Music, Postmodernism and the Poetics of Place*. Verso, London

Manuel P 1993 *Cassette Culture. Popular Music and Technology in North India*. University of Chicago Press, Chicago

Merriam A P 1964 *The Anthropology of Music*. Northwestern University Press, Evanston, IL

Muller C 1999 *Rituals of Fertility and the Sacrifice of Desire: Nazarite Women's Performance in South Africa*. University of Chicago Press, Chicago

Regev M 1986 The musical soundscape as a contested area: 'Oriental music' and Israeli popular music. *Media, Culture and Society* **8**: 343–56

Said E 1979 *Orientalism*. Vintage Books, New York

Slobin M 1993 *Subcultural Sounds. Micromusics of the West*. Wesleyan University Press, Hanover

Stokes M 1992 *The Arabesk Debate: Music and Musicians in Modern Turkey*. Clarendon Press, Oxford, UK

Stokes M 1994 *Ethnicity, Identity and Music: The Musical Construction of Place*. Berg, Oxford, UK

Turino T 1993 *Moving Away From Silence. Music of the Peruvian Altiplano and the Experience of Urban Migration*. University of Chicago Press, Chicago

Wallis R, Malm K 1992 *Media Policy and Music Activity*. Routledge, London

V. Erlmann

Music as Expressive Form

Across the spectrum of the social sciences (and overlapping with musicology and cultural studies), scholars have employed a wide range of theoretical and methodological strategies in their efforts to reveal the mechanisms that link musical and social processes. While work prior to the 1970s outlined comprehensive and ambitious programs of sociomusical study, in the past three decades sociomusical analysis has followed two, more delineated, paths of growth. The first is concerned with *contexts* of music production, distribution, and consumption, the second, with the interpretation of musical *texts* and their social meanings. In recent years, the two paths have converged in the study of music in everyday life, and in ways that have forged new and more empirically grounded answers to questions originally broached during the early and middle twentieth century.

1. Classic Approaches

The most ambitious paradigm to have been developed within sociomusical studies during the twentieth century—T. W. Adorno's—is, arguably also the most controversial. Adorno's periodic resurgence within and across the social sciences and musicology, however, is testament to the enduring heuristic value of his work. Adorno's perspective is distinguished, above all, by its comprehensive vision, and for the central place it accords to music within modern (and often repressive) culture and social formation.

For Adorno, music was linked to and able to inculcate cognitive habits, modes of consciousness, and historical developments. Music's compositional processes—its degree of conventionality, its interrelation of musical parts or voices, its use of concord and dissonance—could serve, within the Adornoian purview, as means of socialization. For Adorno, nothing was more serious or more urgent than sociomusical study: linked to a long-term and a developmental historical trajectory, music for Adorno could 'aid

enlightenment.' Conversely, in the age of 'Total Administration,' music was also a medium that 'trains the unconscious for conditioned reflexes' (Adorno 1976, p. 53). Jazz, Tin Pan Alley, and nearly every other popular music genre were, in Adorno's view, linked to forms of regression and infantile dependency.

In contrast to Max Weber, who primarily was concerned with the origins of musical–technical practices distinct to the West, Adorno's focus was psychosociological, at times music-psychoanalytic. It was directed to the question of music's ideological component, and to music's ability not only to reflect but (in line with classic philosophers such as Plato and Aristotle) to instigate or reinforce forms of consciousness and social structures. At the same time, the absence in his work of an empirical methodology by which his claims could be demonstrated has frustrated critics from his own day to the present. By the late 1970s, Adorno, and the 'grand' paradigm represented by his work, was set aside (with the lone exception within musicology of Rose Subotnik). It is currently undergoing a resurgence within the social sciences.

2. *Contexts of Musical Production, Distribution, and Consumption*

By 1989, new perspectives, established initially in reaction to grand programs such as Adorno's, were firmly established. These perspectives and the various publications that issued from them drew upon close empirical work—ethnographies, cultural and social histories, quantitative surveys, and studies of music-producing organizations. In an essay on ethnomusicology and sociology, Becker described scholars working in the new mode as not 'much interested in "decoding" art works…(but rather)…prefer to see those works as the result of what a lot of people have done jointly.'

The years between roughly 1978 and the middle 1990s were enormously productive for grounded sociomusical enquiry. In retrospect, the contributions of these years may be seen to fall into three broad categories: (a) conditions of production, (b) the construction of musical value and reputation, and (c) musical tastes, consumption, and social identity.

2.1 *Conditions of Musical Production*

Studies of music support have highlighted many of the ways in which the content of musical works is shaped in relation to musicians' working conditions. Elias's study of Mozart, for example, suggests Mozart's compositional scope was hampered by his position between two patronage modes and his inability to escape the shackles of aristocratic control. Similarly, Becker's study of dance musicians (published in *Outsiders*) documents how career patterns and occupational opportunities are shaped by patrons and by the need to find a fit between musicians' aspirations and tastes and what the public will tolerate. Cerulo's (1984) study of composition in six countries during World War II took the production approach into the realm of composition itself. It showed how the networks and music-occupational worlds within which composers operated affected their compositional choices. Combat Zone composers, Cerulo argues, were more likely to engage in idiosyncratic compositional practice, not as a matter of ideological or expressive intent, but because they were isolated from their professional communities, and it was this separation from the wider music world that 'caused the unravelling of the normative prescriptions that govern techniques of composition' (1984, p. 900).

Not only are individual composition practices affected by production organization, so too is the selection of compositions ultimately produced and marketed. Peterson and Berger (1990) illustrated this point in a highly influential study that reveals innovation as enabled and constrained by infrastructural features of the pop music industry. Their work suggested that innovation in pop is cyclical and arises from competition between large record companies and their smaller rivals, showing that diversity in musical forms is inversely related to the degree of market concentration.

Peterson and Berger's study set the scene from the 1970s onward for the concern, within popular music studies, with the production system. More recent studies have highlighted the impact of music industry occupational stratification on the types of music that are produced: women and unfamiliar styles and artists, for example, are marginalized. These forms of musical-gender segregation may be seen within musical production in pedagogical settings also, particularly with regard to instrument choice.

2.2 *The Construction of Musical Value and Reputation*

The stratification of composers, styles, and genres is a rich seam of sociomusical research. Historical studies have helped to unveil the strategies by which the musical canon and its hierarchy of 'Master (sic) Works' was constructed and institutionalized during the nineteenth century in Europe (Weber 1978) and America (DiMaggio 1982). An aesthetic and music-occupational movement, the fascination with 'high' music culture during the nineteenth century was simultaneously a vehicle for the construction of class and status group distinction.

More recently, focus on distinction between 'high' and 'low' musical forms has widened to include questions of authenticity (Peterson 1997), focusing in

particular on the practices and strategies through which particular versions of aesthetic hierarchies are stabilized. Other work focuses on musical reputation and emphasizes the simultaneous production of works and evaluative criteria, and the role played by patrons and the public in this process, particularly in relation to 'revolutionary' stylistic developments (DeNora 1995). Within the study of musical value and its construction, attention has been devoted as well to the role played by technology in stabilizing or undercutting value and to the configuration of the listener as a historical type.

2.3 *Musical Taste, Consumption, and Identity*

Studies of musical value and its articulation address, by definition, the matter of how music is appropriated and how music consumption is linked to status definition. This program is overt in the work discussed in the previous section. It is buttressed in turn by a wide band of quantitative studies of arts consumption that document links between musical taste and socioeconomic position. In recent years, perhaps particularly in the USA, the traditional highbrow/lowbrow division of musical taste has been transformed in favor of an omnivore–univore model that suggests elite music consumers participate in and state preferences for a variety of musical genre while members of nonelite occupational groups exhibit more restricted taste preferences and defend those preferences more vehemently (Peterson and Simkus 1992).

Pioneering work by Willis (1978) and Frith (1981) inaugurated a strong tradition within popular music studies of music consumption ethnography. This work has been linked to broader concerns with subculture and style. Its strength has been to highlight music as a medium for the production of group identity. The particular contribution of these studies arises from their focus on the connections or 'articulations' made—by individuals and groups—between styles of action on the one hand and musical styles and forms on the other. Through articulation work, music is rendered emblematic of subcultural values and also constitutes and elaborates those values. The concern with 'articulation' has been subject to renewed attention in recent years in ways that have drawn it beyond its original subcultural confines (see Sect. 4.1).

3. *The Meanings Of Musical Texts*

The focus on music's social contexts has contributed substantially to sociomusical understanding, in particular by theorizing and documenting in concrete terms some of the mechanisms by which music is linked to and affected by its social contexts. At the same time, critics of context studies, some of them internal, have suggested that these studies look everywhere but at the musical forms (or, more broadly, the artworks) themselves.

The concern with musical texts has tended to follow one of two basic theoretical strategies. On the one hand, scholars have pursued the question of music's capacities as a communicative and representational medium, in particular whether music can be said to possess grammatical and semantic properties and its capacity both to signify and refer to social and conceptual matters. This question continues to receive attention from psychologists and philosophers, as signaled, for example, by a special interdisciplinary symposium on *Music and Meaning* published in *Musicae Scientiae* (Symposium 1997). It connects most closely with some aspects of so-called 'new musicology' and the concern there with music's social and cultural significance. On the other hand, sociologically minded scholars have suggested that while musical texts may appear to lend themselves to some interpretations more readily than others, their semiotic possibilities remain, at least in principle, open. For this reason, they argue, interpretive music analysis simultaneously constitutes the object of musical analysis. Sociological work has thus been concerned with musical meaning as an emergent phenomenon. It has focused on how texts connect, function, and are deployed within wider clusters of materials and conventions of reception such as the listening space, its social relations, interpretative choices involved in its performance (such as phrasing, dynamics, tempo, and tone color), and its accoutrements such as program notes, critical discourse, and technology. Others have pointed to music's ability to delineate external meaning through performance.

These interactionist developments have run counter to earlier attempts to theorize music as homologous with deep mental structures or with social structural epochs. They also point away from laboratory-based studies of music perception (Deutsch 1982) which, as some have suggested, while valuable, are nonetheless far removed from forms of music listening that occur in daily life (Sloboda 1985, pp. 152–4 and *passim*). They do not dismiss the text or reduce it to readings but rather are concerned with delineating musical meaning as involving much more than individual readers and bounded texts (Martin 1995). As such, interactionist approaches to music's meaning point the way toward a growing trend within sociomusical studies in favor of music as it occurs, is used and has effects in naturally occurring settings.

4. *New Directions—Music in Naturally Occurring Settings*

More recently, sociomusical analysis has been marked by two significant—and to a large extent, pandisciplinary—developments. The first is a shift away from the text/context dichotomy. The second is an

expanding interdisciplinary common ground upon which the question of musical ecology has been mooted and explored. Both trends have helped to highlight music's active role in relation to human social agency and both document the ways in which music is used in social life. Both also point to a future for sociomusical studies wherein the (traditionally anthropological) matter (Merriam 1964) of how music functions in practice across a range of social settings will be more fully illuminated.

4.1 Overcoming the Text/Context Dichotomy in Music Studies

The divide, within sociomusic studies, between work dealing with texts vs. those dealing with contexts left little space for questions concerning music's social powers, and its role, *pace* Adorno, as an active ingredient in social life. In recent years, this gap has been bridged by work devoted to music and the emotions, the body and conduct style.

A common thread running through this work is its concern with music as a resource for social action and for agency broadly conceived. Within social movement theory, for example, music has been conceptualized as providing exemplars or models within which social action and movement activity is constructed and deployed. In this respect, music provides, as earlier ethnographers of musical subcultures suggested, a resource for articulating meanings that apply beyond the sphere of music itself. This perspective concentrates on the interpenetrative relation between music and human activity. This pragmatic and, conceptually, reflexive theory of the music–society nexus circumvents text/context dichotomies, conceiving of music as a referent or model *for* (rather than a signifier *of*) action, feeling, and thought.

This focus on music as resource has been applied fruitfully to the question of subjectivity and its cultural/social construction (Gomart and Hennion 1999, DeNora 2000). There, music is portrayed as a resource for the production and auto-production of emotional stances, styles and states in daily life and for the remembering of emotional states. Although this work clearly connects with pioneering efforts within social psychology (Sloboda 2000), it also indicates an explicitly sociological focus on self-regulatory strategies in—and oriented to—particular social contexts. It reveals some emotional management and self-production as social processes across a range of circumstances.

Concurrently, there has been a renewed interest in music's effect on and relation to the body. This work has sought to document corporeal states (e.g., energy, coordination, entrainment, and bodily self-awareness) as they are mediated by musical materials (e.g., rhythm, cadence, tempo, motive, style, and genre). It downplays a conception of music as a 'stimulus' and highlights instead music's capacity (according to how it is heard and appropriated) to 'afford'—provide resources for and to enable—forms of corporeal organization and states of being. In its focus on music's connection to modes of being and modes of attending to the social environment, it connects with Schutz's classic emphasis on the phenomenological dimension of music making (Schutz 1964). These issues are currently being illuminated through studies of music and bodily activity, in particular in relation to fitness where music may be seen to function as a 'prosthetic' and/or body-modifying technology (DeNora 2000, Chap. 4). Music's capacity as an affordance structure and its organizing powers with respect to the body is perhaps most concretely illustrated in medical-based studies of music therapy.

4.2 Interdisciplinary Convergences

In recent years, psychology, sociology and ethnomusicology have developed common ground with regard to the study of musical ecology. This focus reconnects with Blacking's (1977) path-breaking attempt to draw sociomusic studies well beyond culturally specific notions of musical works and their reception and to move toward more universal questions concerning soundscape and the non-propositional bases of human social action. Such a program entails a shift away from 'compositionality' and toward the links between conduct and musical ecologies.

Work on musical ecology has stemmed from two distinct disciplinary sets of concerns, social psychological and sociological. The former has relied primarily upon environmental experimental procedures and quantitative modes of investigation. It has been concerned with music's influence on behavior. Some of this work has emphasized music's primary characteristics such as tempo and its relation to embodied activities such as walking, browsing, eating, or drinking. Other studies, particularly those concerned with musical 'atmospherics' in organizational settings have emphasized how, through its stylistic connotations, music may provide a set of cues about 'appropriate' forms of action and so serve as a resource for the creation of scenic specific forms of action and interaction (North and Hargreaves 1997).

Sociologically oriented work has been concerned more broadly with music's heightened profile within managerial and consumer cultures. It has posed theoretical questions concerning the aesthetic and specifically musical means by which actors and their bodies are socialized (and seek to socialize each other) across a range of settings, public, private, collective, and intimate. To the extent that sociomusical analysis has addressed and continues to address questions of conduct and aesthetic consciousness (emotion and embodiment), it promises to enhance social science understandings of the aesthetic bases of social action and organization. In this regard, sociomusical studies

have been restored to the heart of social science investigations, the position they occupied in social theoretical work from Plato through Adorno.

See also: Censorship and Transgressive Art; Cultural Expression and Action; Cultural Policy: Outsider Art; Expressive Forms as Generators, Transmitters, and Transformers of Social Power; Music: Anthropological Aspects; Music Perception; Popular Culture

Bibliography

Adorno T W 1976 *Introduction to the Sociology of Music* (trans. Blomster W). Seabury, New York

Becker H S 1989 Ethnomusicology and sociology: A letter to Charles Seeger. *Ethnomusicology* **33**: 275–99

Cerulo K A 1984 Social disruption and its effects on music—an empirical analysis. *Social Forces* **62**

DeNora T 1995 *Beethoven and the Construction of Genius*. University of California Press, Berkeley, CA

DeNora T 2000 *Music in Everyday Life*. Cambridge University Press, Cambridge, UK

Deutsch D 1982 *The Psychology of Music*. Academic Press, New York

DiMaggio P 1982 Cultural entrepreneurship in nineteenth-century Boston: The creation of an organizational base for high culture in America. *Media, Culture and Society* **4**: 35–50, 303–22

Frith S 1981 *Sound Effects: Youth, Leisure, and the Politics of Rock 'n' Roll*. Pantheon, New York

Gomart E, Hennion A 1999 A sociology of attachment: Music amateurs, drug users. In: Law J, Hazzart J (eds.) *Actor Network Theory and After*. Blackwell, Oxford, UK

Hennion A 1993 *La Passion Musicale*. Metaille, Paris

Martin P J 1995 *Sounds & Society: Themes in the Sociology of Music*. Manchester University Press, Manchester, UK

Merriam A 1964 *The Anthropology of Music*. Northwestern University Press, Chicago

North A, Hargreaves D 1997 Music and consumer behavior. In: Hargreaves D, North A (eds.) *The Social Psychology of Music*. Oxford University Press, Oxford, UK

Peterson R 1997 *Creating Country Music: Fabricating Authenticity*. University of Chicago Press, Chicago

Peterson R, Berger D 1990 Cycles in symbol production: The case of popular music. In: Frith S, Goodwin A (eds.) *On Record: Rock, Pop and the Written Word*. Routledge, London

Peterson R, Simkus A 1992 How musical tastes mark occupational status groups. In: Lamont M, Fournier M (eds.) *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago

Schutz A 1964 *Making Music Together*. Collected Papers, Vol 2. Martinus Nijhoff, The Hague

Slobada J 1985 *The Musical Mind: The Cognitive Psychology of Music*. Clarendon Press, Oxford, UK

Sloboda J 2000 Everyday uses of music listening. In: Yi S W (ed.) *Music, Mind and Science*. Seoul National University Press, Seoul, South Korea

Weber W 1992 *The Rise of Musical Classics in Eighteenth Century England: A Study In Ritual, Canon, And Ideology*. Clarendon Press, Oxford, UK

Willis P 1978 *Profane Culture*. Routledge, London

T. DeNora

Music, History of

'History of music' signifies the sum of all musical or music-related events in as much as they have been documented or handed down through time. In particular, it refers to the historiography of music (music history) and is a synonym for the discipline of musicology. In recent decades it has come to designate that part of musicology referred to as 'historical musicology.'

1. History and Representation

The panorama of cultures, with their divergent concepts of music and history, are too multitudinous to be adequately and equitably considered here. A review of the world's oral and written musical traditions, whether they be Indian, American-Indian, Chinese, Australian-Aboriginal, Japanese, Ewe, Near-Eastern, or Central European, would only culminate in the emptiest of generalizations to compare their individual courses. For reasons of expediency, therefore, the following remarks refer to the history of music as perceived from the perspective of Western culture and are restricted to its tradition.

Three characteristics of music determine and modify questions concerning historical data, documents, and sources, modulating thereby the various approaches to the history of music. (a) By necessity, the transitory nature of music as a performed art form does not engender an immutable artifact for permanent view, and, in contrast to the visual arts, is dependent on reproduction and representation. This contributes to the tendency to perform and listen to music as though it would be permanently present. In fact, performing or listening to music from an earlier period decomposes the historical distance between the period in question and the present time. This impulse to dispossess music of its history may in some instances lead to the radical inclination to regard music as a natural phenomenon rather than as a historical one. (b) Musical notation was an early invention. Two different kinds were already established in Greek antiquity, most likely for educational purposes and not for the preservation of music for future generations or for performances. This also holds true for the earliest stages of polyphony in the Middle Ages, which went hand in hand with the beginning of notational practice

and ultimately led to the types of scores now familiar to us. Before printing was invented, it can scarcely be claimed that musical notation produced works as texts comparable to written texts, such as the *The Odyssey* or *The Iliad*. And while illiteracy and education are by nature contradictory, musical illiteracy has remained a socially tolerated phenomenon; someone who is unable to read music, will not be ostracized from an elitist social group. In popular music circles musicians have never been required to possess the ability to decipher musical notation. (c) With Edison's invention of the phonograph, voices and sounds could be acoustically documented for the first time. Due to the omnipresence of music in today's popular culture and the recording industry, one can easily forget that this 'kopernikanische Wendung' in music history took place just one century ago. In 1886 Friedrich Nietzsche asserted that music is the latecomer of every culture (*Menschliches*, *Allzumenschliches* II, Sect. 171). Although this corresponds to the romantic conviction that music is the youngest of the fine arts it seems to contradict the Western tradition that ranked music theory among the earliest individual academic disciplines, together with arithmetic, geometry, and astronomy. This was the case in the Greek *enkyklios paideia* and later in the *quadrivium* of the medieval *septem artes liberales*. To the degree that music is more often associated with emotional and spiritual qualities than with intellectual or physical ones, the reconstruction of the history of sound is no less precarious than the history of the emotions themselves. Both fields are predominantly dependent on a systematic or experimental approach rather than on a historical one. And in contrast to archeological sites or the visual arts, the historical foundations of music are ultimately of less interest within the context of the essentially personal need for music.

2. Considering Music History and the Beginning of Music

History has always played a certain role in speculations about the theoretical basis of music. The historical account of music in (pseudo-)Plutarchus's *De musica* from ca. AD 100 weaves together mythological and historical names (that cannot be clarified in each instance) before turning to the matter of 'old music,' regarded as holy and good, and its succession by 'new music,' seen as degenerate and decadent. Aristophanes and other contemporaries maintained that this transition took place in the late fifth century BC. Twentieth-century historians still adhered to this model. Greek myths and medieval legends based on biblical narrative that once explained and interpreted the origins of music have today been replaced by 'scientific' theories that biologically and genetically construe the earliest stages of the development of music, whether beyond written traditions or in paleontological respects. One may ask whether modern speculations as to music's prehistoric inception are more reliable than Ovid's report of Pan's invention of the syrinx (*Metamorphoses* I, 689 ss.) that signaled the beginning of another *ars nova*, or more insightful than the lemma handed down since late antiquity that states Pythagoras as the *inventor musicae*, claiming that he was to have found an explanation for the phenomenon of consonances and therefore to have 'invented' music theory.

By 1700, book titles espousing the history of music appeared throughout Europe (e.g., W. C. Printz, *Historische Beschreibung der Edelen Sing- und Kling-Kunst*, Dresden, 1690; G. A. Bontempi, *Historia musica*, Perugia, 1695; P. Bourdelot and J. Bonnet, *Histoire de la musique*, Paris, 1715). In all cases, the point at which the history of music is seen to begin is dependent upon the historiographical premise that has been taken. Some scholars proceed from early organized polyphony around AD 900; others go back to monophony (Gregorian chant). The belief that ancient Greek music need not necessarily belong to music history's canon is in large part due to the decrease in the knowledge of the Greek language in the twentieth century, with the result that it is subsequently relegated to the fields of ethnomusicology or anthropology (cf. West 1992, p. 3). Thus, what was still regarded to be the foundation of European culture and humanistic tradition in the nineteenth century has been cut off from the history of Europe.

3. The Discipline and its Methods

The rise of historicism in the nineteenth century brought music history (together with art history) as an academic field into the university, a development that began primarily in German-speaking countries. The history of music quickly became an essential component of both the theoretical and practical study of music. Together with specialized research institutes, the academic discipline increased in German universities after World War I, while the wave of emigration following 1933 played an exceptional role in its advancement in the USA. Since World War II the discipline has grown inside and outside Europe, and today is established in universities worldwide. Aside from biographical studies and philological efforts (such as editing scores and theoretical texts on sources), music history's chief methodological approach has been to examine the history of style, or more recently, genre. This is narrowed down further to the history of composition or individual works. Special attention has been given to the history of reception throughout the last decades. Even with respect to essentially synesthetic productions, such as opera or stagings within the multimedia landscape, the score

remained the music historian's most important source at hand. Up to the present time, the predominant individuals of interest for historians have been composers, regardless of how prominent a performer, whether singer, conductor, or instrumentalist, may be. This is also the case within the history of performance practice (impulses coming from the UK), which attempts to reconstruct performance style within any given age independently of the work of individual interpreters. Compared to music history's concentration on composers and compositions, adjacent fields such as the history of music theory and of music aesthetics, and the archeology and iconography of music have assumed less prominent positions. This holds true for the relationship between the history of music and history as well. For example, while writings on the history of sacred Christian music are often aligned with related areas such as the history of liturgy, churches, and religion, contextualizations within social and political history remain rare. Supported by the romantic notion of *Kunstreligion* (the substitution of religion by the arts), the ideology that positions music in an apolitical, internalized, subjective domain strictly separated from social and political conditions has not yet been fully overcome.

3.1 Boundaries

The history of music has pertained and, in large part, still pertains to Western music, if not exclusively. The periodization of earlier epochs (the Middle Ages, the Renaissance, and Baroque) has remained comparatively stable; thereafter, less so. The history of music is patterned after conceptualizations similar to those used in general historiography (whether in the framework of *Geistesgeschichte*, social, or cultural history); the philosophy of history; and the fundamental concept of permanent progress (innovation and its aesthetic equivalent, originality) and cyclic forms (such as ascent, climax, and decline). In spite of how commonplace the expression 'occidental music' (*abendländische Musik*) is, a history of European music in which local, regional, and national traditions are treated as a unity within diversity—in accordance with the standards of the modern European Union—has yet to be written. Historians who specialize in the musical culture of their own regions often lay claim to the concept of occidental music, and yet upon closer scrutiny, the traditions of other countries are comfortably overlooked. In the subtext one can ascertain that the inclusion of music from the Byzantine empire (with consequences for Eastern European music) and from the Americas within the occidental circle is questioned. The post-Wagnerian composer Hans Pfitzner (1926, p. 193) compared music's historical course to the upbringing of a boy whose first nurse was Dutch, who then lived for a season in Italy before making his home in Germany around 1750. Music historiography, dominated by the ascendance of nationalism in the nineteenth century, has been influenced by national historiography. Some music historians' nationalistic stance relies on a conception of music history in which consecutive composers are threaded onto a string of pearls, neither allowing for diversity or pluralistic considerations.

3.2 'Relative' Autonomy

Music historians who view the history of music within the scope of art history tend to view favorably the idea of a '"relative autonomy" of music history' (cf. Dahlhaus 1983, Chap. 8). This concept focuses on questions of style, genre, and composition without giving credence to their contextual arena—an idealistic interpretation of art music that boasts of a pragmatic concentration on musical components at the cost of music's isolation from society, economy, and politics. The idea of autonomy maintains that (a) art is independent of life and (b) music history is made up of a sequence of masterpieces written by heroic composers of genius, whose musical innovations and originality are synonymous with progress. This progress is often associated with decline in differing degrees up to the present time: while contemplating the grand summits of art music, no thought should be squandered on the valleys. As can be expected, Marxist writers have explicitly opposed this view (cf. Knepler 1977, p. 462 ss.). Art and social histories were often strictly and programmatically opposed to each other; particularly in the tradition of Hegel, dominant in German musicology, this led to a disengagement between music history and history in general. The historiography of Western music has remained the historiography of an elitist art form. In spite of its universal popularity, Beethoven's music has never belonged to the realm of popular culture (one notable exception being the *Ode to Joy*, used as the anthem for the European Union). Popular music—written for the day, not for eternity—has socially, politically, and historically been the most effective type of music throughout the ages. Music historians seldom participate in the methodological discourses of historians. Likewise, the historian's rare stand on musical matters, owing to a lack of technical skills or detailed knowledge, is similar to the philosopher who eschews discussions on music in their aesthetic writings. None of the famous historians, therefore, focused on music or, particularly, music history as a methodological paradigm, and in spite of the general increase of interdisciplinary research, historical musicology and history rarely cooperate.

3.3 Recent Developments

Music history, perceived either as the history of Western music in general or its various national entities, came to be questioned in the second half of the

twentieth century. In the USA a discussion about whether its music history should solely be defined from the European perspective has emerged. A front-page article of the *New York Times* entitled 'Ibn Batuta and sitar challenging Columbus and piano in schools' on April 12, 1989 asked whether university professorships for piano should be replaced by those for sitar. Jazz and other forms of popular culture excluded from traditional historiography have been seen by some to pose a threat to the canon of classical music heard in Carnegie Hall and at the Metropolitan Opera. 'Eurocentrism' in music has come under critical scrutiny in recent years. Ethnomusicologists increasingly bemoan the dominance, even hegemony, of Western music as it spreads globally through radio and recordings. This new wave of cultural colonialism is seen as a potential annihilator of the last traces of 'original' music cultures. In extreme cases, ethnomusicologists deny history in music altogether if the related culture in question does not reflect the concept of history, or where a lack of observable change has led to an 'almost absolute stability of the musical cultures of non-Western societies,' a view modified by major scholars (Nettl 1983, p. 172). Although 'world music' is the slogan of the day, the idea of a universal history of music is seen to be obsolete or, due to the plethora of materials, unmanageable. The relevant UNESCO project begun in 1979, *Music in the Life of Man*, retitled in 1989 *The Universe of Music. A History*, has produced programmatic statements rather than results. Conversely, the controversial search for 'universals' in music is generally conducted in a non-historical fashion. 'New musicology,' which began in the USA in the early 1980s, directs its attention to gender issues and the music of minority groups, and emphasizes a contemporary approach to criticism in which the role of history has yet to be fully defined. The phrase 'new historicism' has entered into the discussion of late (cf. Stanley 2001).

4. Identity Exchanges and Modern Mythology

The current global fashion and passionate search for cultural and national identities presents some dubious consequences for music history. If one attempts to decipher what is specifically 'Italian' or 'American' in a piece of music, the answer will remain arbitrary by definition, since sounds cannot signify unequivocal and indubitable meanings. An example of one such identity crisis can be found in Austria. Viennese classical music, denoting the music of Haydn, Mozart, and Beethoven, has long been considered to be a paradigm of, if not synonymous with, German music. In 1947 Albert Einstein stated that the 'concept of German music' itself has come to rest on the works of these three masters (cf. Einstein 1992, p. 261). Austrian historiographers of music have now come to emphasize their own musical identity as being distinctly different in substance from a German one. This recent dichotomy is all the more curious since Austrians have frequently been eager to discern the German identity in their Austrian composers. Now Beethoven regains the status of an alien immigrant at best (cf. Flotzinger 1997, col. 1201). Hitler's former Reich's Music Chamber President, Richard Strauss, revealed yet another astonishing conviction shortly after World War II. On May 20, 1946 he admitted to his devoted friend and historiographer, Willi Schuh, that in all likelihood cultural historians, German patriots, and the Jewish press would be less than pleased with his tenet that 'political Germany *had* to be destroyed after it had fulfilled its world mission: the creation and perfection of German music' (Strauss 1969, p. 89). Of course, the composer regarded his own music as this ultimate fulfillment, as the end of music history. Uniting the idea of the end of art and the end of politics, his vision placed all future music in a state of a *post histoire*.

Nowadays, music composed just a few years after Strauss' remarkable statement, whether serial, electronic, or chance, is considered classic. The histories of all musics have been subject to various readings throughout the course of time, often stemming from a national, ethnic, or otherwise political intention. The genesis of one of the most significant twentieth-century forms, jazz, has been alternately depicted as originating in Africa, Europe, and America. And it is in the New World that the search for a national musical identity during the first half of the twentieth century failed to produce a national school of composition. 'Each composer found his own way,' and 'perhaps this is the way it should be in America' (Hamm 1983, p. 459). Perhaps this maxim could on occasion be applied to the realm of music history as well.

See also: Art: Anthropological Aspects; Dance, Anthropology of; Music: Anthropological Aspects; Music as Expressive Form; Music Perception; System: Social

Bibliography

Allen W D 1962 *Philosophies of Music History. A Study of General Histories of Music 1600–1960*. Dover Publications, New York

Dahlhaus C 1983 *Foundations of Music History*. Cambridge University Press, Cambridge, UK

Eggebrecht H H 1980 Historiography. In: Sadie S (ed.) *The New Grove Dictionary of Music and Musicians*. Macmillan, London, Vol. 8, pp. 592–600

Eggebrecht H H 1991 *Musik im Abendland*. Piper, Munich, Germany

Einstein A 1947 *Music in the Romantic Era*. Norton, New York [original reprinted in 1992 as *Die Romantik in der Musik*, Metzler, Stuttgart, Germany]

Flotzinger R 1997 Österreich. In: Blume F (ed.) *Die Musik in Geschichte und Gegenwart*. 2nd edn. Metzler, Stuttgart, Germany, Vol. 7, Col. 1176–223
Hamm C 1983 *Music in the New World*. Norton, New York
Knepler G 1977 *Geschichte als Weg zum Musikverständnis*. Verlag Philipp Reclam, Leipzig, Germany
Nietzsche F 1878–1879 *Menschliches, Allzumenschliches*. Chemnitz (2nd edn. 1886) Vol. 11, pp. 546–61
Nettl B 1983 *The Study of Ethnomusicology*. University of Illinois Press, Urbana, IL
Pfitzner H 1926 *Gesammelte Schriften*. Benno Filser-Verlag Augsburg, Germany, vol. 1
Stanley G 2001 Historiography. In: Sadie S (ed.) *The New Grove Dictionary of Music and Musicians*, 2nd edn. Macmillan, London
Strauss R 1969 *Briefwechsel Mit Willi Schuh*. Atlantis, Zurich
Treitler L 1989 *Music and the Historical Imagination*. Harvard University Press, Cambridge, MA
West M L 1992 *Ancient Greek Music*. Clarendon Press, Oxford

A. Riethmüller

Music Perception

Music consists of sound organized in time, in terms of the auditory dimensions of pitch, loudness, timbre, and location. That organization is usually intended to be, and many claim must be, perceived to constitute music. Music is organized within a cultural context of shared patterns that lead the listener to develop perceptual frameworks that reflect them. Those frameworks develop throughout a lifetime of musical experience, both listening and playing, and serve to guide expectancies and facilitate the perceptual processing of expected events when they occur. From this perspective there should be individual differences in the way music is heard, depending on cultural and personal background. All listeners share the human auditory system, but that basic endowment is overlaid by perceptual learning. Music is constrained by what can be perceived by the human ear, and also by the wide range of patterns accessible to human cognition.

1. Historical Context

The modern history of the study of music perception starts with Helmholtz ([1877] 1954). Helmholtz was concerned chiefly with the organization of the pitch material in music, and much of his pioneering work was devoted to the sensory encoding of pitch. Pitches can be organized in temporal succession, called 'melody,' and in simultaneous groups of different pitches, called 'chords' and 'harmony.' Among the many problems addressed by Helmholtz was that of harmony; namely, how the properties of chords emerge from those of their component tones.

Among the many ways in which chords can differ in sound is along the dimension of consonance vs. dissonance. Consonant combinations of pitches are stable and strike the ear as relatively simple (e.g., a major chord such as C-E-G). Dissonant combinations sound relatively unstable, restless, and complex (e.g., a diminished-seventh chord such as B-D-F-Ab. It is useful to distinguish between two types of consonance: 'tonal,' and 'musical' or 'esthetic' (Dowling and Harwood 1986). 'Tonal consonance' arises from the structure of the human auditory system and refers to a quality of an isolated chord, often described in terms such as 'smooth' vs. 'rough,' 'pleasing' vs. 'unpleasant,' etc. 'Esthetic consonance' refers to the relative stability of a chord in musical context, and depends upon culturally conditioned perceptual learning. Esthetically consonant sounds occur at points of rest and stability in the musical structure; esthetically dissonant sounds are unstable and require resolution. The distinction between tonal and esthetic consonance is illustrated by the judgment of a single tone. A single tone represents the maximum of tonal consonance, but single pitches can differ in esthetic consonance. If you sing 'Twinkle, Twinkle' and stop on the penultimate note, you can observe the instability of that note and its requirement for resolution to the final note. It is as if it were being pulled down to the final note. That is esthetic dissonance. The final note will be no more tonally consonant, but will be esthetically consonant.

In European music the dimensions of tonal and esthetic consonance are closely correlated. Tonal dissonance is used to emphasize the instability of esthetically dissonant chords, and vice versa. Helmholtz was thus led to approach the problem of harmony in terms of tonal consonance, with little consideration of esthetic consonance. He had observed that as two pure tones are gradually separated in frequency a beating sensation occurs. The frequency of the beats is equal to the difference in frequency between the two tones. Helmholtz thought that the roughness of sensation arising from the beats between adjacent tones was the source of dissonance. However, Helmholtz realized that that was not the whole story. If beats of 30–35 per second were the only source of dissonance, then a constant frequency difference between tones should result in the same degree of dissonance across the auditory spectrum, from the bottom of the piano keyboard to the top. Helmholtz's observations contradicted this, showing that at least in the range of 250–1000 Hz (from middle C up two octaves) a constant ratio between frequencies (and hence a constant musical interval) results in the same dissonance. Helmholtz ([1877] 1954, Chap. 8) thought that distance along the basilar membrane in the inner ear between the excitations produced by the two tones was involved, but could not specify how. This problem was solved nearly a century later (Plomp and Levelt 1965).

In spite of his emphasis on physiology and on culturally independent aspects of music perception, Helmholtz ([1877] 1954) provided a careful review of musical structure, especially scale structure, both in the history of European music and in non-Western cultures. Helmholtz, however, remained attached to a culturally universal, physical explanation of scale structure, namely that the pitches of the tonal scale were such as to produce coincidences of upper harmonics; that is, that their frequencies should represent simple integer ratios. Since pure tones have no upper harmonics, Helmholtz found melodies made of pure tones puzzling. Then as now, the cross-cultural evidence contradicts the integer ratio hypothesis for scale construction.

Helmholtz was concerned with music perception as a fruitful field for exploring human cognition. In one of his most interesting passages he speculated on why music relies on scales with their sets of discrete steps of pitch, instead of using continuous variations of pitch (like a fire siren). He suggested that discrete steps give the auditory system a way of measuring distances of pitch, which would remain more ambiguous with continuous variations. He was probably on the right track, given what we know of limitations on human cognition.

2. *The Impact of Behaviorism*

Early in the twentieth century US psychology turned away from the study of mental life to the study of behavior. 'Mentalistic' language, including references to human experience, was forbidden in the major journals, and there was a tendency to study only problems that could be addressed with nonhuman organisms. This discouraged the study of perception, which cannot avoid reference to the psychological dimensions of experience (cf. Boring 1933), and of music and language. Though the Gestalt theorists emphasized musical examples, taking melody as a prototype of a meaningful perceptual whole with emergent properties beyond those of the notes that make it up, they performed few empirical studies. Werner's (1948) studies of the listener's adaptation to 'micromelodies'—melodies constructed of tiny pitch intervals—are an intriguing exception.

The most prominent figure in music perception in the first part of the twentieth century was Carl Seashore (1938). Seashore followed the nineteenth-century tradition of studying aspects of music perception thought to be culture-free, and studied them in a way that did not counter the behaviorist paradigm. For example, Seashore made an extensive study of the rate and extent of the vibratos (minute rapid frequency and intensity variations) of professional performers.

Seashore is best known for his tests of musical talent, whose dissemination was facilitated by the development of phonograph recordings. The battery of tests was focused on those component sensory abilities thought to be essential for the development of musical skill, and included judgments of pitch, intensity, melody ('tonal memory'), rhythm, timbre, and consonance. As an example of the effort to make the tests 'culture-free,' the test of melodic judgment (in which the listener has to say whether the two melodies of a pair are the same or different) used nontonal melodies. That is, the melodic patterns did not conform to the tonal scale structure of any culture. The child acquires most of the tonal structure of the culture by the age of 10 (Dowling 1999). Whatever the test was measuring, it was not the efficacy with which the child had acquired the essentials of the tonal scheme. Seashore's measures have rather modest validity in terms of predicting musical achievements, while tests that take cultural aspects into account have better success (Shuter-Dyson 1999). (For a fuller account of the contributions of Seashore and his place in the history of music psychology, see Gjerdingen 2000.)

3. *The Return of Consciousness*

In the 1950s US psychology, stimulated by discoveries in the area of language and by the advent of the computer, returned to the study of human cognition (see *Cognitive Psychology: History*). Psychologists began to think in terms of cognitive frameworks or 'schemata' that are acquired through the listener's acculturation, that guide expectancies and perception, and that facilitate memory and thought. This approach was given new impetus by Francès ([1958]) who concentrated on the framework governing pitch structure. Since then psychologists have concentrated on describing the cognitive frameworks that underlie the mental representations of pitch and time in music.

The cognitive frameworks involved in music perception are largely implicit, embodying procedural knowledge concerning the structure of music. They are acquired very slowly over a lifetime of listening and performing. These frameworks are automatically invoked, so that what the listener perceives is already interpreted via the cultural framework. The frameworks for music are thus similar to those of language. When we listen to speech in a language we know we hear a meaningful succession of words and phrases. The interpretation of the sounds and their meaning occurs automatically.

3.1 *Pitch*

An experiment by Francès ([1958], Exp. 2) illustrates several aspects of the implicit framework for musical pitch. As noted above, pitches in a musical context differ in stability. The tonic note *do* in a *do-re-mi* scale is very stable; melodies can end on it giving a

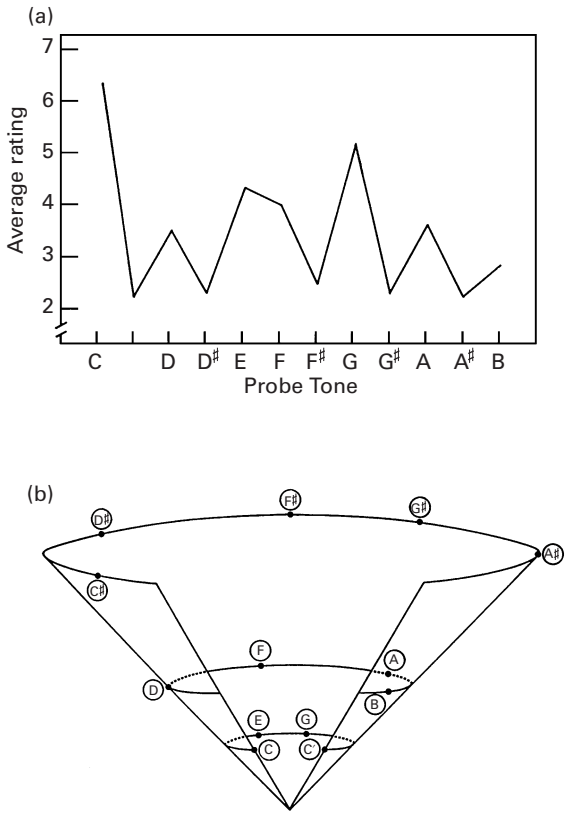

Figure 1
(a) The profile of ratings of the 12 semitones used in European music when they were presented in a tonal context in the key of C major. (b) Multidimensional scaling solution in three dimensions representing judgments of the relatedness of pairs of pitches presented in a C-major context. The greater the distance, the less closely related the pitches

sense of closure. Other pitches in the scale are less stable, and have tendencies that pull them toward more stable pitches. In the previous example, when you sing 'Twinkle, Twinkle' and end on the penultimate note, not only does the song sound unfinished, but that note has a very strong tendency pulling it down to resolve on the final note. To show that such tendencies have real perceptual effects, Francès lowered ('flatted') the frequencies of two notes on the piano. He then placed those notes in different tonal contexts, invoking different tonal tendencies, to see whether listeners would notice the mistuning. For example, one of the flatted notes was Ab (also called G#). Francès placed the flatted note in pieces in the keys of C minor and E major. In C minor the flatted Ab is the sixth degree of the scale, with a strong downward tendency. (The scale goes: C-D-Eb-F-G-Ab.) In that case, listeners did not notice the flatting. However, when the very same note was the third degree (G#–E-F#-G#) in E major (with an upward tendency), the flatting became very noticeable.

Francès's experiment illustrates the importance of the tonal framework in the perception of musical pitch in that listeners hear pitches in terms of their stability in the tonal context. It also shows that the tonal framework operates implicitly. Listeners did not hear a mistuned note, and then think, 'Yes, but this is a flatted note in a context where the note already has a downward tendency, so that's all right.' They simply did not notice the flatting where it went in the contextually determined direction. Francès's experiment also shows that the tonal framework is not just a static system for categorizing pitches, but rather represents interrelationships of stability and instability, attraction and repulsion.

The mental representation of the tonal system has been explored extensively using an approach developed by Krumhansl and Shepard (1979). On each trial of a typical study the listener hears a strongly tonal context, for example, a scale (in the key of C major: C-D-E-F-G-A-B …). The context is followed by a test pitch, one of the 12 semitones in the European chromatic scale. The listener judges how well the test pitch fits the context. Typical results are shown in Fig. 1a. Listeners, especially listeners with some musical training, give high ratings to the most stable pitches in the musical key context: the members of the tonic triad (in the key of C: C, E, and G). The next highest ratings go to the other members of the musical scale: D, F, A, and B. The lowest ratings go to those pitches that fall outside the key—that would be left out of melodies such as 'Twinkle, Twinkle' and 'Happy Birthday' in C major: C#, D#, F#, G#, and A#.

With a converging method Krumhansl (1990) showed that the tonal hierarchy governs relations among pitches. She had listeners judge the relatedness of pairs of pitches presented in a tonal context and used the ratings to infer relative psychological distances in the mental representation of the tonal hierarchy. The more closely related a pair of pitches was judged to be, the smaller the distance between them. Krumhansl used multidimensional scaling to achieve a three-dimensional representation of the tonal hierarchy (see *Mental Representations, Psychology of*), resulting in the conical pattern shown in Fig. 1b. Note that the most stable pitches are tightly clustered near the apex of the cone, with the other scale pitches at an intermediate distance, and nonscalar pitches in the outer circle. Both of the patterns in Fig. 1 show the effect of musical structure on the way pitches are heard in context.

Studies of melody recognition have provided converging evidence for the tonal framework. Both musicians and nonmusicians find it easier to remember tonal melodies, and notice changes in them, than nontonal melodies (Dowling 1991). Melodies constructed of more stable pitches are easier to encode in

memory and to retrieve. Further, listeners notice alterations in melodies more easily when they introduce nonscalar pitches than when they stay within the scale framework (Dewar et al. 1977). Listeners perceive and remember more accurately melodies that conform to the cultural norms, and quickly notice deviations from those norms.

Musicians and nonmusicians typically perform in qualitatively similar ways respecting the tonal framework in melody recognition, though musicians usually perform better overall. However, even a few years of music lessons changes the encoding of pitches. Dowling (1986) presented listeners with brief melodies in the midst of a chord progression that centered the melody on the first degree of the scale (the tonic) or the fifth degree. When tested later they had to say whether the melody had been altered (with one pitch changed). The test melodies either retained the same context, or had the context shifted. With shifted context, all the scale values of the pitches changed, even though the melodic pattern remained the same. That is, a melody that had been centered on the tonic was now centered on the fifth: What had been the first degree became the fifth, the second degree became the sixth, etc. The context shift did not affect nonmusicians' performance; however, it did affect that of listeners with about five years of music lessons in their youth. Those listeners performed well when the context was the same, but fell to chance when the context was changed. Their moderate training led them to encode pitches in terms of their place on the scale framework, and when those encodings did not match the test items, they were lost (see *Encoding Specificity in Memory*).

This illustrates again the implicit nature of the framework: people with five years of music lessons are not aware of encoding pitches in terms of their *do-re-mi* values. If they had explicit access to that encoding they would be able to write down the melodies they hear, which they cannot do. (Professional musicians, having explicit control over application of the scale framework, perform well whether or not the context is changed.)

Thus whether measured via perceptual judgments or the detection of changes in melodies, the tonal scale framework that listeners acquire through a lifetime of experience with the music of their culture guides perception and perceptual expectancies.

3.2 Time

A framework for time is as important as a framework for pitch. Time in music is structured by the overlay of more or less complex rhythmic patterns on a regular beat pattern (Dowling and Harwood 1986). In the absence of a regular beat pattern listeners' temporal judgments are not very precise, but with a beat they become so. Bharucha and Pryor (1986) found that listeners were much better at detecting disruptive pauses in a tone sequence when a regular beat was present than when it was not. Dewitt and Samuel (1990) found that pitch alterations are easier to detect when the target pitch is part of a regular, predictable sequence.

Converging evidence comes from a study by Povel and Essens (1985), who had listeners tap along with complex rhythmic patterns. They found that listeners were much more successful following the rhythm when the pattern agreed with an underlying beat structure than when it did not.

The relationship between the various levels of rhythmic organization in music, and between the musical organization and the perceptual frameworks stored in the brain, is one of the most exciting areas of developing research. This is an area where theoretical developments have kept pace with empirical discoveries, and an area of broad psychological interest, since a central issue involves the encounter between the biological clocks of the brain and more 'rational' structures in the environment. Large and Jones (1999) provide a lucid and stimulating account of this fast-developing area.

3.3 Perception and Memory

One conclusion that has become more and more compelling since the early 1950s is that there is not a sharp line of demarcation between the phenomena of perception and memory. Crowder (1993) provided a nice demonstration that the same brain processes are involved in both perception and memory. First he showed that when the listener judges the pitch of a test tone, the response is faster and more accurate when the test tone has the same timbre as the cue. That is, if we present a middle C played by a trumpet, then the judgment of the pitch of a comparison tone close to middle C is facilitated if it also is played by a trumpet. Then Crowder replaces the cue with a pure tone (like a flute sound), and performance declines. However, if he told the listener to imagine the pitch of the pure tone played on the trumpet, and tested with a trumpet tone, performance again improved (but not if the listener was misinformed and imagined a guitar). The listener's recall of the trumpet timbre before the test has the same facilitating effect as an actually perceived trumpet. Crowder concludes that memories are embodied in the residue of the neural activity associated with perception (and not, for example, stored in a separate module).

4. Conclusion

We have seen considerable growth over the past 150 years in our understanding of the perceptual processing of pitch in musical contexts. This understanding extends to both the sequential organization of pitches (melody) and their organization into simultaneous combinations (consonance and dissonance—har-

mony). This progress has been accompanied by a broadening understanding of the importance of acculturation through a lifetime of perceptual learning. One consequence of the role of acculturation is the wide range of individual differences in music cognition, varying with culture, subculture, and training.

During the coming decades I expect that there will be commensurate progress in our understanding of rhythm and time. The present convergence of cogent theories with sophisticated experimentation in that area provides ample grounds for optimism. For both pitch and time these advances will shed increasing light on human cognition in general.

See also: Audition and Hearing; Auditory Models; Music: Anthropological Aspects; Music as Expressive Form; Music, History of; Neural Plasticity in Auditory Cortex

Bibliography

Bharucha J J, Pryor J H 1986 Disrupting the isochrony underlying rhythm: An asymmetry in discrimination. *Perception & Psychophysics* **40**: 137–41

Boring E G 1933 *The Physical Dimensions of Consciousness*. The Century Co., New York

Crowder R G 1993 Auditory memory. In: McAdams S, Bigand E (eds.) *Thinking in Sound: The Cognitive Psychology of Human Audition*. Oxford University Press, Oxford, pp. 113–45

Dewar K M, Cuddy L L, Mewhort D J K 1977 Recognition memory for single tones with and without context. *Journal of Experimental Psychology: Human Learning & Memory* **3**: 60–7

Dewitt L A, Samuel A G 1990 The role of knowledge-based expectations in music perception: Evidence from musical restoration. *Journal of Experimental Psychology: General* **119**: 123–44

Dowling W J 1986 Context effects on melody recognition: Scale-step versus interval representations. *Music Perception* **3**: 281–96

Dowling W J 1991 Tonal strength and melody recognition after long and short delays. *Perception & Psychophysics* **50**: 305–13

Dowling W J 1999 The development of music perception and cognition. In: Deutsch D (ed.) *The Psychology of Music*, 2nd edn. Academic Press, San Diego, CA, pp. 603–25

Dowling W J, Harwood D L 1986 *Music Cognition*. Academic Press, Orlando, FL

Francès R 1958 La perception de la musique. Vrin, Paris [1988 *The Perception of Music*. Erlbaum, Hillsdale, NJ]

Gjerdingen R 2001 The psychology of music. In: *The Cambridge History of Western Music Theory*. Cambridge University Press, Cambridge, UK

von Helmholtz H L F [1877] 1954 *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Dover, New York

Krumhansl C L 1990 *Cognitive Foundations of Musical Pitch*. Oxford University Press, New York

Krumhansl C L, Keil F C 1982 Acquisition of the hierarchy of tonal functions in music. *Memory & Cognition* **10**: 243–51

Krumhansl C L, Shepard R N 1979 Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception & Performance* **5**: 579–94

Large E W, Jones M R 1999 The dynamics of attending: How people track time-varying events. *Psychological Review* **106**: 119–59

Plomp R, Levelt W J M 1965 Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America* **38**: 548–60

Povel D J, Essens P 1985 Perception of temporal patterns. *Music Perception* **2**: 411–40

Seashore C E 1938 *Psychology of Music*. McGraw-Hill, New York

Shuter-Dyson R 1999 Musical ability. In: Deutsch D (ed.) *The Psychology of Music*, 2nd edn. Academic Press, San Diego, CA, pp. 627–51

Werner H 1948 *The Comparative Psychology of Mental Development*. Follett, Chicago

W. J. Dowling

Myrdal, Gunnar (1898–1987)

Gunnar Myrdal (1898–1987), Nobel Laureate in economics and social scientist of encyclopaedic range, is best known for his two monumental studies *An American Dilemma* (1944) and *Asian Drama* (1968). He was a founding member of the Stockholm School of Economics and one of the architects of the Swedish welfare state. Throughout his life Myrdal was also deeply concerned with the role that values play in the social sciences.

1. Early Life and Career as an Economist

Karl Gunnar Pettersson—the father later changed the name of the children to 'Myrdal'—was born on December 6, 1898 in Skattungsbyn in the province of Dalarna, Sweden. His mother, Anna Sofia Pettersson (born Karlsson), came from a family of farmers and was softspoken, religious, and bitterly unhappy in her marriage. The father, Carl Adolf Pettersson, also came from a family of farmers but was secular, restless, and upwardly mobile. The father was a building constructor, and in 1904 he moved the family to Stockholm. Myrdal went to a very good high school, Norra Real, and was mainly interested in the natural sciences. He, however, was also deeply influenced by the ideas of the Enlightenment, to which he was introduced by one of his teachers. All through his life he would display a passionate faith in the power of reason.

After having graduated with excellent grades in 1918, Myrdal became a student of law at Stockholm University (*Stockholms Högskola*). The decision to study law had been influenced by his father, but also

by Myrdal's own conviction that law would help him better understand how society operates. Whatever conservatism and class prejudices still dominated his mind, and which he had picked up from his family, were soon to be challenged by Alva Reimer (1902–86), whom he met in the summer of 1919 and married in 1924. It was also Alva Myrdal who prodded her husband to study economics, when he had received his law degree in 1923 and felt deeply depressed.

While Myrdal had never enjoyed studying law, he quickly took to economics. In 1927 he got his doctorate through a dissertation for which Gustav Cassel had been the advisor, *Prisbildningsproblemet och föränderligheten* (*The Problem of Price Formation and Change*). In this work Myrdal attempted to develop the beginnings of a dynamic theory of price formation by introducing the idea of expectations. In 1930 he published *Vetenskap och politik i nationalekonomien* (*The Political Element in the Development of Economics*), in which he argued that economists often confuse values and facts in their studies, something which endangers both economics as a science and its potential use for social reform. Based on an article from 1932, Myrdal published a booklength study in 1939 entitled *Monetary Equilibrium*. In this work he attempted to further develop Knut Wicksell's theory of monetary equilibrium by introducing the notion of expectations. In the early 1930s Myrdal also produced a couple of writings for the Swedish Social Democratic government in which he advocated an underbalanced budget as a way of countering a downturn in the business cycle. These works were based on Myrdal's discussions with his friend Erik Lindahl and the Swedish finance minister Ernst Wigforss.

All of these writings, except for *Vetenskap och Politik*, are today regarded as important contributions to the Stockholm School of Economics, which is usually defined as a Keynesian type of economics, based, however, on Knut Wicksell's ideas. The value of these works is today mainly historic. They nonetheless played a crucial role in the development of Myrdal's ideas, especially in his attempt to construct a dynamic theory of causation, based on Wicksell's ideas. *Vetenskap och Politik*, on the other hand, has aged rather well and represents a still valuable contribution to the history of economic ideas. While Myrdal would always approvingly refer to its main thesis—that neoclassical economics is deeply biased through its origin in the natural law tradition and in utilitarian philosophy—he would later criticize his youthful assumption that it was possible to separate out a core of economics, which was free from all values.

1.1 Political Activities and An American Dilemma

Both Gunnar and Alva Myrdal decided in the early 1930s to become actively involved in politics and were to play a crucial role in the construction of the Swedish welfare state through their advocacy of 'social engineering.' The timing to do so was perfect, since in 1932 the Social Democratic Party had won the election and was eager to transform the poor and class-ridden Sweden into a real 'people's home' (*folkhem*). Gunnar Myrdal not only helped to formulate the economic aspects of the new policy that now came into being, but also its social policy. This he did, in close collaboration with Alva Myrdal, through an avalanche of writings as well as his participation in various royal commissions, appointed by the government to investigate some specific issue. The basic message of the two Myrdals was that social policy had to be fundamentally changed, from focusing on symptoms, to becoming 'prophylactic.'

Gunnar and Alva Myrdal presented their vision for what a new and modern Sweden would look like in *Kris i befolknigsfrågan* (*Crisis in the Population Question*, 1934), which became a political bestseller. 'All of society', they here argued, would have to be changed, so that 'the quality of people' could be improved and better adjusted to the demands of modern, industrial society: living standards, consumption patterns, the way children are brought up, the structure of education, and much, much more. Though permeated by the authors' generosity and enlightenment, *Kris i befolkningsfrågan* also evinces an enormous belief in the authority of science and experts. The Myrdals' advocacy of sterilization of certain categories of people, who were deemed unfit for the new society, has also been criticized.

By the mid-1930s Myrdal's work had become so dominated by practical and political concerns that he longed for a return to academic studies; and when he in 1937 was asked by the Carnegie Corporation if he would lead a study on the blacks in the United States, he accepted. During the next few years Myrdal travelled around in the United States like a latter-day Tocqueville, interviewing people from all strata—'I'm from Sweden and I don't know anything about the race problem. What can you tell me?' Through the backing of the Carnegie Corporation Myrdal put together a small circle of collaborators, and he also engaged, in one way or another, some 70 other scholars, including Samuel Stouffer, Louis Wirth, and E. Franklin Frazier. Fearing many times that he would never be able to synthesize the enormous amount of material that he had commissioned, in 1944 Myrdal had the satisfaction of seeing *An American Dilemma* appear in print in two volumes of more than 1,400 pages.

Myrdal's work was much commented upon, and most of what was said was positive. *An American Dilemma* was actually to dominate the discussion of ethnicity in the United States from the end of World War II till the 1960s, and it also played an important role in the decision of the Supreme Court in 1954 to declare segregated schools unlawful (*Brown vs. Board*

of Education). From the mid-1960s till the 1970s Myrdal's book was criticized for depicting certain aspects of black life as pathological and also for not advocating change from the bottom up; and today *An American Dilemma* is seen primarily as being of historical interest in ethnic studies.

Myrdal's basic thesis—that there is a moral dilemma in the heart of the average person between 'the American Creed' and everyday behavior towards blacks—may seem old-fashioned and passé today. This, however, should not prevent one from realizing the value of other parts of *An American Dilemma*, especially its appendices on values and on cumulative change (Appendices 1–3). In the first two of these appendices Myrdal argues that social scientists should always state their value premises since these deeply influence the analysis. In Appendix 3 Myrdal presents the core of his own social science approach, which is centered around Wicksell's theory of unstable equilibria or what Myrdal calls 'the principle of cumulation.' An initial change, Myrdal explains, can set off further changes; and these may then move the whole system either in an upward direction or in a downward direction ('vicious circles').

2. *Activities After World War II and Becoming an Institutionalist*

After *An American Dilemma* Myrdal returned to a combination of politics and applied research till the mid-1950s, serving first as Minister of Trade in Sweden (1945–7) and then as Executive Secretary in the Economic Commission for Europe of the United Nations (1947–57). Neither of these two experiences were particularly successful for Myrdal, who was more of an intellectual than a politician; and he soon longed to return to academic studies. The topic that he chose to focus on this time was underdevelopment, and a fine introduction to his early efforts in this direction can be found in *Rich Lands and Poor* (1957). It is in this work, for example, that Myrdal introduces the idea that a strong economic development in one area may either impede or accelerate economic growth in nearby areas ('backwash effects' and 'spread effects'). His most sustained effort to understand underdevelopment, however, took 10 years to produce and appeared in 1968 in three volumes of more than 2,000 pages: *Asian Drama*.

Like *An American Dilemma*, *Asian Drama* has been much debated. The initial reaction was many times positive, but today Myrdal's work is less often referred to. One reason for this is the fact that his strong advocacy of large-scale state planning and state intervention as well as his clear stance for what he calls 'The Modernization Ideals' do not go very well with today's emphasis on market forces and the prevalent scepticism towards Western ideals. Still, several parts of *Asian Drama* are still of interest, such as the notion of 'the soft state,' the critique of various aspects of neoclassical economics, and Myrdal's attempt to use Wicksell's theory of cumulative causation to construct a theory of institutionalist economics (see especially Appendices 1–3).

After *Asian Drama* Myrdal kept writing books and articles, and he also gave lectures all over the world. In 1969 he published a small and still useful volume entitled *Objectivity in Social Research*, and in 1974 he was awarded the Nobel Prize in Economics, together with Friedrich von Hayek. Autobiographical elements can be found in *Against the Stream* (1973) as well as in *Hur styrs landet?* (*How is the Country Governed?* 1982). His reference to the superiority of institutionalist economics over neoclassical economics became ever stronger; and the essay *Institutional Economics* (1978) is especially recommended. An attempt to write a sequence to *An American Dilemma*, entitled *An American Dilemma Revisited*, had, however, to be called off in 1985, due to illness. Gunnar Myrdal died on May 17, 1987 in Stockholm. To his very last day he remained a firm believer in the ideals that he shared so strongly with his beloved wife Alva Myrdal, and which for him were symbolized by the one word with which *An American Dilemma* ends: *Enlightenment*.

See also: Constitutionalism; Development, Economics of; Economic Growth: Theory; Race: History of the Concept; Racial Relations; Racism, History of; Racism, Sociology of; Southeast Asia: Sociocultural Aspects

Bibliography

Andersson S 1986 Alva och Gunnar Myrdals arkiv. *Arbetarhistoria* **10**: 75–8

Assarsson-Rizzi K, Bohrn H (eds.) 1984 *Gunnar Myrdal: A Bibliography, 1919–1981*, 2nd edn. Garland Publishers, New York

Jackson W A 1990 *Gunnar Myrdal and America's Conscience: Social Engineering and Racial Liberalism, 1938–1987*. University of North Carolina Press, Chapel Hill, NC

Myrdal A, Myrdal G 1934 *Kris i befolkningsfrågan (Crisis in the Population Question)*. Albert Bonniers Förlag, Stockholm

Myrdal G 1927 *Prisbildningsproblemet och föränderligheten (The Problem of Price Formation and Change)*. Almqvist and Wicksell, Uppsala, Sweden

Myrdal G 1930 *Vetenskap och politik i nationalekonomien*. P. A. Nordstedt and Söners Förlag, Stockholm [trans. 1953 by Paul Streeten as *The Political Element in the Development of Economics*, Harvard University Press, Cambridge, MA]

Myrdal G 1939 *Monetary Equilibrium*. Hodge, London

Myrdal G 1944 *An American Dilemma: The Negro Problem and Modern Democracy*. Harper, New York

Myrdal G 1957 *Rich Lands and Poor*. Harper, New York

Myrdal G 1968 *Asian Drama: An Inquiry into the Poverty of Nations*. Twentieth Century Fund, New York

Myrdal G 1969 *Objectivity in Social Research*. Pantheon Books, New York

Myrdal G 1973 *Against the Stream: Critical Essays on Economics*. Pantheon Books, New York

Myrdal G 1978 Institutional Economics. *Journal of Economic Issues* **12**: 771–83
Myrdal G 1982 *Hur styrs landet? (How is the Country Governed?)*. Rebén and Sjögren, Stockholm
Myrdal G 1987 *Historien om An American Dilemma (An American Dilemma Revisited)*. SNS, Stockholm

R. Swedberg

Mysticism

1. *Definition of Mysticism*

Mysticism consists of a practice of religious ecstasies, i.e., of religious experiences during alternate states of consciousness, together with whatever ideologies, ethics, rites, myths, legends, magics, etc., are related to the ecstasies. 'Ecstasy' may be defined as any experience of involuntary belief in the reality of the numinous, i.e., of the holy or demonic. Whenever the numinous is as manifestly real as perceptible reality is during waking sobriety, consciousness has been sufficiently altered that the experience is intrinsically religious (Merkur 1992).

Since the paradigm shift promoted by the work of Katz (1978, 1983), mysticism is no longer defined as the ideology and practices surrounding the experience of union with the Absolute (Underhill 1910), or ultimate reality, etc. There is abundant evidence that the people who traditionally have been called mystics have enjoyed many varieties of religious ecstasies. Comparative studies of mysticism currently include the vision quests and shamanism, spirit possessions and spirit mediumship, prophetism and seers, etc., of the oral religious traditions, as well as the visions, voices, mystical unions, enthusiasms, etc., of the scripture-based religions.

Mystical experiences are as varied as the dreams of sleep. Not only personal matters, but also culturally and historically variable religious ideologies contribute to the contents of mystical experiences and their postexperiential interpretations. People in all cultures may potentially experience all possible varieties of religious experience. The world's ecstatic religious traditions each tends to accommodate a considerable diversity of ecstatic experiences, even when certain types are privileged. Conformity is required in matters of doctrine and behavior much more than in experience. The vagueness or nonspecificity of many traditional doctrines permits them to be applied simultaneously to diverse religious experiences.

No sociological research has yet been done that conforms with mysticism's current definition. Collating existing research on shamanism, spirit mediumship, prophetism, mysticism in the book religions, etc., in search of common denominators, discloses few commonalities. People who make religious uses of alternate states of consciousness take their religion more personally and seriously than the general population. They tend to be convinced, and not only hopeful, about the reality of their objects of belief. Religion is not only a social institution for them; it is an integral component of the cosmos. In Allport's (1959) terms, their religiosity is 'intrinsic,' and not only 'extrinsic.'

In other respects, mystics are diverse. Some mystics are religiously and/or socially conservative, but others are revolutionary. Some seek to transcend all perceptible phenomena; others address the immanence of spiritual phenomena within material reality. Some shun their bodies and engage in asceticism. Others celebrate both their bodies and the physical world. Some seek to withdraw from society, but others participate within it. Some mystics prioritize a single type of mystical union, e.g., nothingness, or all-being, or identity with God; but others favor complex systems of multiple types of religious ecstasy that include both visions and unions.

The term 'mysticism' derives historically from the noun *mystes*, 'mystic,' meaning the possessor of a secret, and denoting an initiate of a classical Greek or Hellenistic Mystery. The Mysteries were rites of initiation that induced religious experiences. Technical terms of the Mysteries were taken over by Plato, Philo, and the Christian Church Fathers, who applied the terms to their own practices of religious experiences. By the early medieval period, Christianity officially denied possessing any secrets that had not been publicly divulged through Christ. Hindu yoga, Buddhist *vipassana* meditation, and Confucian meditation, among other mystical traditions, similarly teach their alternate-state practices openly. At the opposite extreme, Taoism, Hindu and Buddhist tantra, theosophical Sufism, the Kabbalah, and Western alchemy are so very secretive that they encode their teachings in secret languages, whose meanings cannot be understood by noninitiates.

2. *Incidence and Varieties of Mystics*

In a survey of 488 societies worldwide, Bourguignon (1973) found that 90 percent had institutionalized one or another religious use of alternate states of consciousness. In different surveys in Western cultures, between one-third and one-half of the population has been found to have had one or more mystical experiences (Greeley 1975). A considerably smaller part of the population pursues mystical experiences on a frequent or habitual basis. Many factors have contributed to the distinctive historical forms that different mystical traditions have taken.

2.1 *Shamans*

A shaman may be defined as a religious functionary who enters ecstasy and, by commanding the aid of helping spirits, intermediates between the social group

and the numina (Hultkrantz 1973). Shamanism has its origin and major distribution in hunter cultures. Like hunters, shamans' ecstasies are highly mobile. In their visions, shamans may travel out of their bodies to the sky, netherworld, or distant overland locations. Shamans also summon helping spirits to them, sometimes from considerable distances, and they send them abroad on distant errands. Shamans relate to their helping spirits much as hunters relate to half-wild dogs.

A shaman typically is paid for services rendered to a client, such as healing. In their ecstasies, shamans may also find game animals, placate angry deities, guide new ghosts to the afterlife realm, and hex enemies in war. In some hunter cultures, vision quests are pursued more generally by part or all of the society. In these cases of 'democratized shamanism,' shamanism occurs as a private practice, and only mystics who are exceptionally gifted emerge as professionals.

Shamanism persists in agricultural communities in varied forms. Democratized shamanism persisted as a practice among warriors in Indo–European cultures (e.g., Glosecki 1989). It was also adapted to group possession states in the classical Mysteries of Dionysus. In late antique and medieval Europe, shamans who lacked traditional social locations within their communities emerged as itinerant seers (Brown 1981, Davidson 1989), magicians (Smith 1978), and witches (Ginzburg 1983). Where seers have a clientele, magicians and witches often engage in shaman-like conjuring of spirits, but privately rather than professionally.

2.2 Spirit Mediums

A spirit medium may be defined as a religious functionary who enters ecstasy and, through states of possession, serves as a vehicle for the manifestation of a numen to the social group. Spirit mediumship has its major distribution in clan-based farming villages, both in Africa (Walker 1972) and in Southeast and East Asia (Elliott 1955, Paper 1995). The sedentary life style of farming villages is reflected in the lodging of spirits within mediums. Mediums and their spirits perform no errands that require travel outside the village. Much spirit mediumship is intimately connected with the worship of ancestral ghosts. Deference to family elders, which is economically motivated by the ownership and inheritance of arable land, is continued after the elders' deaths through the agency of mediums, whom the ancestral ghosts possess. The deference is reflected, among other manners, in mediums' claims that all of their activities while possessed are performed by the possessing spirits, while they themselves are passive.

Spirit mediums may also be possessed by gods and spirits that are not ancestral ghosts. Spirit mediumship also occurs in democratized forms, where large numbers of people within a social group engage communally in possession states, and only a few show sufficient virtuosity to function as professionals. Spirit mediumship and possession may persist into more complex societies, e.g., in the theurgy of late antique Hellenism (Luck 1989), enthusiasm in Protestant Christianity (Knox 1950), and modern New Age channeling.

2.3 Prophets

A prophet may be defined as a religious functionary who enters ecstasy, learns of the will of the god, and later communicates the divine will to the social group. Prophets have their major distribution in kingdoms, whose gods are imagined as kings, and whose prophets are imagined as heralds, spokespeople, or ambassadors of the gods (Cross 1973). Prophetism had its origin and classical location in urban societies with differentiated labor and trades, both in the ancient Near East and East Asia. Prophetism can be conceptualized as an adaptation of ecstasy to monarchic politics, with the prophet's god outranking the king. Prophets' ecstasies favor verbal revelations from their gods, but may include shamanic and mediumistic forms that have been adapted to monarchic politics.

Anthropologists also use the term 'prophet' in reference to shamans and spirit mediums who have been influenced by missionaries' accounts of Biblical prophets and developed hybrid types of mysticism that reflect their postcontact situations.

2.4 Mystics in Scripture-based Religions

Because book religions invest religious authority in scriptures and their official interpreters, they marginalize professional ecstatics. Mysticism is then often pursued by a leisure class, which may embrace poverty in order to support its practices, as among Hindu forest dwellers, and in Buddhist and Christian monasticism. In other cases, e.g., Taoism, Sufism, and Kabbalah, mysticism has been pursued by individuals sufficiently affluent to devote leisure time to it.

Mystics in the book religions often avoid conflict with religious authorities by favoring varieties of ecstatic experience that are socially and political powerless. Unitive experiences, which are not useful for professional ecstatics, may consequently be privileged. However, other ecstasies are often pursued as well.

Mystics in the book religions may be public religious functionaries, e.g., as leaders or teachers of other mystics. Their leadership roles among nonmystics tend to be limited to religious revival movements, in which mystics proselytize ecstasy among their co-religionists.

Although most mystics are sociopolitically conservative, a minority are radical and potentially revolutionary. In such cases, mystics in the book religions typically seek social roles appropriate to prophets. Rather than seeking ecstasies that are accommodating, inconspicuous, and innocuous, they seek inspirations that claim authority to transform their religions and/or societies. The would-be prophets may then emerge as charismatic founders of sects, schisms, and/or cults. An apocalyptist may be conceptualized as a would-be prophet in a book religion, who seeks the radical reformation of society, often as an aftermath of catastrophe and revolution. Apocalyptists may found or belong to millennial movements.

3. *The Contents of Mystical Experiences*

All ideas and practices that may be entertained and performed during normal waking sobriety can also be thought and done during at least many varieties of ecstasy. Ecstasies facilitate unconscious manifestations that tend not to occur otherwise; but they do not necessarily preclude the activities of normal consciousness. Mysticism may embrace the whole of non-ecstatic religiosity, together with an additional component i.e. contributed by the ecstatic experiences.

Ecstasies may be subdivided on psychological criteria into two major categories. Some ecstasies involve narratives, as do daydreams and night dreams. The narratives may consist of visual images, with characters, sceneries, dialogue, etc. They may instead consist exclusively of words. Narrative ecstasies can also have physical dimensions and take form as automatic speech, automatic writing, automatic dramatization (i.e., possession), and other types of behavior. Somatizations (stigmata) may also occur.

The second category of ecstasies may be termed unitive experiences. All unitive ecstasies involve a sense of separate things becoming unified. What is unified varies. In normal waking sobriety, our sense of self is elastic. Our senses of our 'selves' is limited to our bodies when naked; extends to clothing when dressed; and may include, e.g., even the fenders of a vehicle when we are driving. Romantic partners, children, family, ethnic group, and nation may also be included in the normal fluctuation of the sense of self. Unitive ecstasies include exceptional instances of the elasticity of the self. In some instances of mystical union, the self is co-extensive with the entire physical cosmos; in others, with infinite timeless being; in others again, with God; in still others, with ultimate nothingness. Other commonly encountered unitive ecstacies include a sense of identity with some one particular thing in the cosmos.

Unitive experiences are always unconsciously dyadic. Just as self exists and is defined psychologically only in relation to others, so mystical unity exists and is defined experientially only in relation to the absolutely Transcendent. One cannot experience self as transcendent, unless the ideas of the self and the Transcendent have been brought to union. Conversely, experiences of self and a numinous being in a manifestly dyadic relationship may be categorized as unitive experiences. Self and other are united in a unifying relationship during experiences of the immediate presence of a numinous power, or absolution by a numinous power, etc.

So-called paranormal phenomena, such as precognition, clairvoyance, and meaningful coincidences ('miracles') are further varieties of unitive experience. They unite subjective fantasy with the objectivity of perceptible reality, imparting objective reality to imagined ('extra-sensory') perceptions (Merkur 1998, 1999).

4. *The Varieties of Alternate States*

Most mystical experiences occur during states of trance. Religious trances are instances of either self-hypnosis or hetero-hypnosis. In other cases, mysticism depends on trances that occur as symptoms of hysteria, epilepsy, and other pathologies. In all cases, suggestion—either autosuggestion or hetero-suggestion—is responsible for directing the trance to serve religious purposes (Arbman 1963).

An important aspect of the trance state is its ability to reify the contents of consciousness (Shor 1959). For example, a trance state is all that is necessary in order to transform a visual daydream into a waking hallucination. Instead of seeming to be imaginary, daydream images may seem during trance to be objectively existing, external realities. Trance states may similarly transform speculative ideas into delusions that have the force of firm convictions.

Trances are often termed 'dissociative states' because the reification of fantasies leads to a corresponding devaluation and/or denial of sense perceptions. Mystics who use trances uncritically may develop dissociative religious doctrines that reject the perceptible world in favor of transcendent realms or states, otherworldliness, asceticism, devaluation of mortal life, community, and the physical world, and asocial or anti-social behavior (Furst et al. 1976). Mystics may, however, use trance states critically, much as hypnotherapists do.

Some mystical experiences occur in a different class of alternate states that may be termed 'reveries.' Nonecstatic forms of reverie include daydreams, creativity, and hypnagogic states between waking and sleeping. Reveries of ecstatic intensity include 'peak experiences,' which typically occur spontaneously and last only a few seconds or minutes. Sensory deprivation is a means to induce reverie states that may reach ecstatic intensity. Psychedelic drugs, i.e., psilocybin, mescaline, LSD, and THC, the active factor in marijuana, similarly induce reverie states, as distinct from trances (Merkur 1993).

Both narrative and unitive ecstasies may occur in either trances or reveries. Reveries differ from trances, among other manners, in the ecstatic's knowledge during reverie that his or her experiences are subjective and intrapsychic, rather than objective and external. They are self-evident as heuristic phenomena. Religious reveries seem subjectively to be apperceptions of truth through the medium of imagination. The experiences take form as pseudo-hallucinations and illusions, where the same narrative and unitive experiences have the force of hallucinations and delusions when they occur during trances.

5. *Psychological Transformation*

Mystics traditionally have sought transformations that were metaphysical in nature: magical empowerment, salvific perfections of their souls, etc. Since the 1970s, transpersonal psychologists have argued that mystical transformations have powerful psychological benefits, consistent with Maslow's (1964) concept of 'self-actualization.' Some mystical practices have been shown to be beneficial developmentally (Sacks 1979, Emavardhana and Tori 1997), but others may be harmful.

See also: Charisma and Charismatic; Christianity: Evangelical, Revivalist, and Pentecostal; Healing; Millennialism; Monasticism: West and East; Prophetism

Bibliography

Allport G W 1959 Religion and prejudice. *Crane Review* **2**: 1–10

Arbman E 1963 *Ecstasy or Religious Trance*. Bokforlaget, Stockholm

Bourguignon E 1973 *Religion, Altered States of Consciousness, and Social Change*. Ohio State University, Columbus, OH

Brown J P 1981 The Mediterranean seer and shamanism. *Zeitschrift für die Alttestamentliche Wissenschaft* **93**: 374–400

Cross F M 1973 *Canaanite Myth and Hebrew Epic*. Harvard University Press, Cambridge, MA

Davidson H E (ed.) 1989 *The Seer: in Celtic and Other Traditions*. John Donald, Edinburgh, UK

Elliott A J A 1955 *Chinese Spirit–Medium Cults in Singapore*. London School of Economics, London

Emavardhana T, Tori C D 1997 Changes in self-concept, ego defense mechanisms, and religiosity following seven-day Vipassana meditation. *Journal for the Scientific Study of Religion* **36**: 194–206

Furst S S, Leavy S A, Lewis R C, Lubin A J, Ostow M, Zales M R 1976 *Mysticism: Spiritual Quest or Psychic Disorder?* Group for the Advancement of Psychiatry, New York

Ginzburg C 1983 *The Night Battles* (trans. Tedeschi J, Tedeschi A). Johns Hopkins University Press, Baltimore, MD

Glosecki S O 1989 *Shamanism and Old English Poetry*. Garland, New York

Greeley A M 1975 *The Sociology of the Paranormal*. Sage, Beverly Hills, CA

Hultkrantz Å 1973 A definition of shamanism. *Temenos* **9**: 25–37

Katz S T (ed.) 1978 *Mysticism and Philosophical Analysis*. Sheldon, London

Katz S T (ed.) 1983 *Mysticism and Religious Traditions*. Oxford University Press, Oxford, UK

Knox R A 1950 *Enthusiasm*. Oxford University Press, Oxford, UK

Luck G 1989 Theurgy and forms of worship in Neoplatonism. In: Neusner J, Frerichs E S, Flesher P V M (eds.) *Religion, Science, and Magic*. Oxford University Press, New York

Maslow A H 1964 *Religions, Values, and Peak Experiences*. Penguin, Harmondsworth, UK

Merkur D 1992 *Becoming Half Hidden*. Garland, New York

Merkur D 1993 *Gnosis*. State University of New York Press, Albany, NY

Merkur D 1998 *The Ecstatic Imagination*. State University of New York Press, Albany, NY

Merkur D 1999 *Mystical Moments and Unitive Thinking*. State University of New York Press, Albany, NY

Paper J 1995 *The Spirits are Drunk*. State University of New York Press, Albany, NY

Sacks H L 1979 The effect of spiritual exercises on the integration of self-system. *Journal for the Scientific Study of Religion* **18**: 46–50

Shor R E 1959 Hypnosis and the concept of the generalized reality orientation. *American Journal of Psychotherapy* **13**: 582–602

Smith M 1978 *Jesus the Magician*, 1st edn. Harper & Row, San Francisco

Underhill E 1910 *Mysticism*. New American Library, New York

Walker S S 1972 *Ceremonial Spirit Possession in Africa and Afro-America*. Brill, Leiden, Germany

D. Merkur

Myth in Religion

Definitions of myth vary. Some definitions focus on the content of myth. Others focus on the function. Defined by content, myth would be a belief about something significant. The subject of myth might then be specified, such as the world, society, or god. Myth here could be a sheer conviction, such as the American 'rags to riches' myth, or could take the form of a story. Defined by function, myth would accomplish something significant for adherents. The function might likewise be specified, such as explaining the world, supporting society, or encountering god.

1. *Theories of Myth*

To study myth is to apply to it theories from the disciplines of anthropology, sociology, psychology, literature, religious studies, and philosophy. There are no theories of myth itself. What unites the study of myth across the disciplines are the questions asked: what is the origin, the function, and the subject matter of myth? Theories differ not only in their answers to

these questions but also in the questions they answer. Some theories concentrate on the origin; others, on the function; still others, on the subject matter. In addition, some theories judge myth false; others judge it true; still others take no stand.

Theories of myth are hardly a modern invention. They go all the way back to the Presocratics (see Chase 1949, Chap. 1, de Vries 1961, Chap. 1). Only in the last century and a half, however, have theories become scientific, for only since then have there existed the social sciences. Some social scientific theories may have counterparts in earlier theories (see Feldman and Richardson 1972), but earlier theorizing was largely speculative and philosophical. Modern theorizing is far more empirical and testable. Even modern theories from literature, religious studies, and philosophy reflect the influence of the social sciences.

The most straightforward way to compare theories of myth is by discipline. But the differences among disciplines are often blurry. For example, the theories of E. B. Tylor, Bronislaw Malinowski, and Claude Lévi-Strauss are as much psychological as anthropological. It is more telling to compare theories by the questions they ask and by the answers they give.

1.1 Myth as Primitive Science

An apt starting point is the theory of the pioneering English anthropologist E. B. Tylor (1832–1917). For Tylor, myth originates and functions to explain physical events. Myth is the 'primitive' counterpart to science, which is exclusively modern. 'Modern myth' is a contradiction in terms. By science, Tylor means natural, not social, science. The events explained by myth are primarily those in the external world such as the rising and setting of the sun, though also explained are human events such as birth and death. The payoff of myth is wholly intellectual: it is not control over the world—the position of James Frazer (1854–1941)—but knowledge of the world: 'Man's craving to know the causes at work in each event he witnesses ... is no product of high civilization, but a characteristic of his race down to its lowest stages' (Tylor 1871, I, pp. 368–9).

For Tylor both myth and science originate through observation, inference, and generalization. Both cover the same range of phenomena. Both are meant literally. The difference is in the explanations themselves. Myth, a part of religion, attributes events to the decisions of gods. Science attributes events to impersonal processes. Myth and science duplicate each other in function but exclude each other in content. One must therefore choose between them. While myth is as rational as science, it is false. Primitives create myth rather than science because they think less critically than moderns. Tylor died far too early to envisage a stage post the modern one.

In the twentieth century, Tylor's theory has been spurned on many grounds: for pitting myth against science and thereby precluding modern myths, for subsuming myth under religion and thereby precluding secular myths, for deeming the function of myth intellectual, for deeming myth false, and for reading myth literally. Postmodernism would dismiss myth as explanation and would treat myth as sheer story. Nevertheless, Tylor's theory remains central to the study of myth, and all other theories can be seen as rejoinders to it. One response to Tylor has been to take the function of myth as other than explanatory, in which case myth runs askew to science and can therefore coexist with it. Another response to Tylor has been to read myth other than literally, in which case myth does not even refer to the physical world and can therefore likewise coexist with science. The most radical response to Tylor has been to alter both the explanatory function and the literal reading of myth.

1.2 Myth as Other Than Explanatory in Function

The most influential reinterpreters of the function of myth have been Bronislaw Malinowski, Claude Lévi-Strauss, and Mircea Eliade. It is not clear whether for Malinowski (1884–1942), the Polish-born anthropologist, moderns as well as primitives have myth. It is clear that for him primitives have science as well as myth, so that myth cannot be the primitive counterpart to modern science. For Malinowski, myth is no more applied science—Frazer's view (Frazer 1911–15)—than theoretical science—Tylor's view. Primitives use science to explain and, more, to control the physical world. They use myth to reconcile themselves to aspects of the world that cannot be controlled.

Myth reconciles humans to the travails of life by rooting those travails in the primordial actions of gods or humans. Humans age because a god or human once did something that brought old age irremediably into the world: 'the longed-for power of eternal youth and the faculty of rejuvenation which gives immunity from decay and age, have been lost by a small accident which it would have been in the power of a child and a woman to prevent' (Malinowski 1926, p. 104). Myth pronounces the world not the best possible one but, in the wake of irreversible events, the only possible one. Still, the world becomes less capricious than it would otherwise be. In tracing back the origin of phenomena, myth does explain those phenomena, but as a means to an end rather than, as for Tylor, the end itself.

Whereas for Tylor myth deals primarily with physical phenomena, for Malinowski it deals equally with social phenomena like customs and laws. Myth still functions to reconcile humans to the unpleasantries of life, but now to unpleasantries that, far from unalterable, can be cast off by members of society. Myth spurs members to accept the impositions of society by tracing them, too, back to a hoary past, thereby conferring on them the clout of tradition. Myths say, do this because this has always been done. In the case

of physical phenomena, the beneficiary of myth is the individual. In the case of social phenomena, it is society itself. The modern counterpart to myths of social phenomena, if moderns do not have myth, is ideology.

The most influential contemporary theorist of myth, the French anthropologist Claude Lévi-Strauss (b. 1906), at first glance seems a throwback to Tylor. For Lévi-Strauss, myth is an exclusively primitive enterprise yet a rigorously intellectual one. Lévi-Strauss denounces nonintellectualists like Malinowski as vigorously as Malinowski denounces intellectualists like Tylor. Indeed, in declaring that primitives, 'moved by a need or a desire to understand the world around them, ... proceed by intellectual means, exactly as a philosopher, or even to some extent a scientist, can and would do' (Lévi-Strauss 1978, p. 16), Lévi-Strauss seems indistinguishable from Tylor. Yet he is in fact severely critical of Tylor. For Lévi-Strauss, primitives think differently from moderns, not fail to think as well.

Primitive, or mythic, thinking is concrete. Modern thinking is abstract. Primitive thinking focuses on the observable, sensory, qualitative aspects of phenomena rather than, like modern thinking, on the unobservable, nonsensory, quantitative ones. Yet myth for Lévi-Strauss is no less scientific than modern science. It is simply part of the 'science of the concrete' rather than part of the science of the abstract, to which it is less at odds than askew. Myth *is* primitive science, but it is not inferior science.

Myth is an instance of thinking *per se*, modern and primitive alike, because it classifies phenomena. According to Lévi-Strauss, all humans think in the form of classifications, specifically pairs of oppositions, and project them onto the world. Many cultural phenomena express oppositions. Myth is distinctive in resolving oppositions: 'the purpose of myth is to provide a logical model capable of overcoming a contradiction' (Lévi-Strauss 1958, p. 105). Myth resolves a contradiction by providing either a mediating middle term or an analogous, but more easily resolved, contradiction. Either tactic narrows and thereby alleviates the contradiction, but neither fully resolves it.

Like the contradictions expressed in other phenomena, those expressed in myth are apparently reducible to the fundamental contradiction between 'nature' and 'culture.' Humans experience themselves as at once animal-like and civilized. This split stems from the projection onto the world of the oppositional character of the mind. By diminishing that opposition, myth makes life more bearable and, even more, solves a logical conundrum.

In calling his interpretation of myth 'structuralist,' Lévi-Strauss distinguishes it from 'narrative' interpretations, which adhere to the plot of myth. Lévi-Strauss dispenses with the plot of myth and locates the meaning of myth in the structure. The plot is that element—say, event—A leads to event B, which leads to event C. The structure, which is identical with the expression and dimunition of contradictions, is either that events A and B constitute an opposition mediated by event C or that events A and B are as opposed to each other as events C and D, an analogous opposition, are opposed.

Lévi-Strauss confines himself to primitive myths, but other structuralists analyze modern myths. In *Mythologies* (1972) the French semiotician Roland Barthes (1915–80) takes as myths various cultural artifacts and shows how they serve to justify the bourgeois outlook of postwar France. The function of myth here is not intellectual but ideological. Myth has nothing to do with natural science. Whereas Lévi-Strauss largely analyzes myths independent of their social context—the grand exception is his analysis of the myth of Asdiwal—others inspired by him have tied myths to their contexts. For the classicists Jean-Pierre Vernant (1983), Marcel Detienne, Pierre Vidal-Naquet, and Nicole Loraux, the relationship between myth and society is much more malleable and ironic than it is for Malinowski or even Barthes. Myth can as readily challenge as bolster existing ideology.

Unlike Malinowski, the Romanian-born historian of religions Mircea Eliade (1907–86) has no hesitation in making one function of myth explanatory. For him, as for Tylor, myth attributes events to the decisions of gods. Like Malinowski, Eliade includes social phenomena as well as physical phenomena. Explanation for Eliade is both an end in itself and a means to another end, which is magically to return to the time of the myth, the time of the origin of whatever phenomenon it explains. In this 'primordial time' gods are nearest to humans, as in the biblical case of 'the Lord God['s] walking in the garden of the cool of the day' (*Genesis* 3:8). The return to primordial time reverses the everyday separation from gods, a separation that is equivalent to the fall, and is regenerative spiritually: 'what is involved is, in short, a return to the original time, the therapeutic purpose of which is to begin life once again, a symbolic rebirth' (Eliade 1968, p. 8). The ultimate benefit of myth is proximity to gods, one or more.

Eliade ventures beyond the other respondents to Tylor in proclaiming myth panhuman rather than merely primitive. He cites modern novels, plays, and films with the mythic theme of yearning to escape from the everyday world into another, often earlier one. If even professedly atheistic moderns have myths, then myth must be universal. How modern myths, which do not involve gods, provide entree to gods, Eliade never discloses.

1.3 Myth as Other Than Literal in Meaning

The most prominent reinterpreters of not the function but the meaning of myth have been the German New Testament scholar Rudolf Bultmann (1884–1976) and

the German-born philosopher Hans Jonas (1903–93). Both were followers of the early Martin Heidegger and offer existentialist readings of myth. While they limit themselves to their specialties, Christianity and Gnosticism, they presuppose a theory of myth *per se*.

Bultmann acknowledges that, read literally, myth is about the physical world and is incompatible with science. But unlike Malinowski and Eliade as well as as Tylor, Bultmann reads myth symbolically. In Bultmann's excruciatingly misleading phrase, one must 'demythologize' myth, by which he means not eliminating, or 'demythicizing,' myth but on the contrary extricating its true, symbolic subject matter. Once demythologized, myth is no longer about the external world but is instead about the place of human beings in that world: 'the real purpose of myth is not to present an objective picture of the world as it is, but to express man's understanding of himself in the world in which he lives. Myth should be interpreted not cosmologically, but anthropologically, or better still, existentially' (Bultmann 1953, p. 10). Myth depicts the human condition.

Read literally, the New Testament in particular describes a cosmic battle of good and evil anthropomorphic gods for control of the physical world. These gods intervene miraculously not only in the operation of nature but also in human lives. Humans are therefore not responsible for their actions. Demythologized, the New Testament still refers in part to the physical world, but now to a world ruled by a single, nonanthropomorphic, transcendent God. Because God does not act directly in the world and because no evil powers exist, humans are free and therefore responsible for their actions. Because a literal interpretation reduces humans to the pawns of larger, competing forces, it concentrates on those forces, which means on the world. Because a symbolic interpretation pronounces humans free, it concentrates on the actions they choose in response to the world.

Demythologized, myth ceases to be purely primitive, as for Tylor, and becomes universal, as for Eliade. Myth ceases to be false, as for Tylor, and becomes true. Whereas Eliade invokes the existence of modern myths as *ipso facto* evidence of the compatibility of myth with science, Bultmann actually tries to reconcile myth with science. Whereas Eliade claims that moderns have myths of their own, Bultmann claims that moderns can retain biblical myths.

Bultmann's boldest response to Tylor is to circumvent the function of myth. In translating the meaning of myth into terms acceptable to moderns, he sidesteps the issue of why moderns, even if they can have myth, need it. Unlike other symbolic interpreters of myth such as the religious philosopher Paul Ricoeur (1967) and the philosopher Philip Wheelwright (1968), Bultmann never maintains that the meaning of myth is untranslatable into nonmythic terms. Since he takes the meaning of myth from philosophy, he can hardly do so. Why, then, is myth needed? Bultmann never says.

Like Bultmann, Jonas seeks to show that ancient myths have an eternal meaning that speaks to moderns. For both Bultmann and Jonas, myth describes the alienation of humans from the world as well as from their true selves prior to their acceptance of God. Because Gnosticism, unlike mainstream Christianity, is radically dualistic, humans remain alienated from the physical world and from their bodies even after they have found the true God. Indeed, Gnostics find the true God only by rejecting the false god of the physical world.

Unlike Bultmann, who wants to bridge the divide between Christianity and modernity, Jonas acknowledges the divide between Gnosticism and modernity. In Gnosticism, the state of alienation is temporary. In modern, secular existentialism, alienation is permanent. Alienation *is* the human condition, not a fall from it. What, then, Jonas 'demythologizes' is not the source of alienation or the solution to it, for there is neither source nor solution, but the fact of alienation. He translates Gnostic myths into existentialist terms not to make Gnosticism palatable to moderns but only to show the similarity between the Gnostic and the existentialist outlooks: 'the essence of existentialism is a certain dualism, an estrangement between man and the world .... There is only one situation ... where that condition has been realized and lived out with all the vehemence of a cataclysmic event. That is the Gnostic movement' (Jonas 1963, p. 325). Like Bultmann, Jonas bypasses the function of myth and confines himself to the meaning.

1.4 Myth as Both Other than Explanatory and Other than Literal

The most radical departures from Tylor transform both the explanatory function and the literal meaning of myth. The most influential theorists here have been psychologists, the Austrian Sigmund Freud (1856–1939) and the Swiss C. G. Jung (1875–1961). For both, the subject matter of myth is the unconscious, and the function of myth is to express the unconscious.

Because the unconscious for Freud is composed of repressed sexual and aggressive drives, myth functions to release those drives, but in a disguised, vicarious way, so that the creator and the user of a myth need never confront its meaning and thereby their own nature. Myth, like other aspects of culture, serves simultaneously to reveal and to hide its unconscious contents. The classical psychoanalytic study of myth is that of Freud's one-time disciple, the fellow Austrian Otto Rank (1884–1939). Focusing on myths of male heroes, Rank sees the myths as providing an unconscious, vicarious fulfillment of, above all, Oedipal drives. Myth serves neurotic adult males fixated at their Oedipal stage: 'Myths are, therefore, created by

adults, by means of retrograde childhood fantasies, the hero being credited with the myth-maker's personal infantile [i.e., childhood] history' (Rank 1914 p. 82).

Contemporary Freudians like Jacob Arlow (b. 1912) take myth positively rather than negatively. For them, myth helps solve the problems of growing up rather than perpetuate them, is progressive rather than regressive, and abets adjustment to society and the outer world rather than childish flight from both. Myth serves less to vent repressed drives than to sublimate them. Consequently, 'myth can be studied from the point of view of its function in psychic integration—how it plays a role in warding off feelings of guilt and anxiety, how it constitutes a form of adaptation to reality and to the group in which the individual lives, and how it influences the crystallization of the individual identity and the formation of the superego' (Arlow 1961, p. 375). Myth serves everyone, not only neurotics.

Jungians have taken myth positively from the outset. For them, the unconscious expressed in myth is not the Freudian repository of repressed, antisocial drives but a storehouse of innately unconscious 'archetypes,' or sides of the personality, that have simply not had an opportunity at realization: 'contents of an archetypal character … do not refer to anything that is or has been conscious, but to something essentially unconscious' (Jung 1968, p. 156). Myth is one means of encountering this Jungian, or 'collective,' unconscious. The function of myth is therefore less release, as for classical Freudians, than growth, as for contemporary ones. But growth means less adjustment to the outer world, as for contemporary Freudians, than cultivation of the 'inner world.' The payoff is self-realization. Some Jungians and Jungian-oriented theorists like the American Joseph Campbell (1904–87) so tout the payoff of myth that it becomes a panacea for all humanity's problems (see Campbell 1949). Jung himself never goes this far. For Jung, myth works best as part of therapy.

For even contemporary Freudians, myth harks back to childhood. For Jungians, myth points forward. Myth especially serves adults already settled in the outer world but severed from the unconscious. Myth is to be read symbolically, as for Freudians, but not because its meaning has intentionally been disguised. Rather, the unconscious speaks a language of its own. Contemporary Jungians like James Hillman (b. 1926) seek to rectify what they consider the monotheistic orientation of Jung's psychology and in turn his approach to myth (see Hillman 1975).

1.5 Recent Theories of Myth

The most notable recent theories of myth have been variations on myth as primitive science, theoretical and applied. Cognitivists led by the French anthropologist Pascal Boyer (1994) analyze the mental processes that shape thinking—processes far more rigid than Tylor's open-ended sequence of observation, inference, and generalization. Cognitivists investigate the constraints on mythic explanations more than, like Tylor, the explanations themselves. The German classicist Walter Burkert (b. 1931) and the French literary critic René Girard (b. 1923) have revived the theory that ties myth to ritual. Whereas for Frazer and his English follower Lord Raglan (1885–1964) myth is the script for the ritualistic killing of the king, whose death and replacement magically ensure the rebirth of crops, for Burkert myth reinforces the ritual that commemorates the past hunting of animals. The function of myth is not physical but psychological and social: to cope with the guilt and anxiety that members of society feel toward their own aggression and to unite society by turning that aggression onto outsiders (see Burkert 1996). Whereas for Frazer and Raglan (1936) the king is heroic because he is willing to die for the sake of the community, for Girard the hero, who can range from the most marginal person to royalty, is killed as a scapegoat for violence in society. Rather than directing the ritualistic killing, as for Frazer and Raglan, myth for Girard arises afterwards to cover it up by making the victim first a criminal and then a hero. The function of myth is social: to preserve the ethos of sociability by hiding not only the killing but, more deeply, the violence endemic to society (see Girard 1972).

See also: Eliade, Mircea (1907–86); Magic, Anthropology of; Malinowski, Bronislaw (1884–1942); Myths and Symbols: Organizational; Religion, Sociology of; Ritual; Symbolism in Anthropology; Symbolism (Religious) and Icon; Totemism

Bibliography

Arlow J A 1961 Ego psychology and the study of mythology. *Journal of the American Psychoanalytic Association* **9**: 371–93

Barthes R 1972 *Mythologies* [trans. Lavers A] Hill and Wang. Cape, London, New York

Boyer P 1994 *The Naturalness of Religious Ideas*. University of California Press, Berkeley, CA

Bultmann R 1953 New testament and mythology (1941). In: Bartsch H-W (ed.) *Kerygma and Myth* [trans. R H Fuller]. SPCK, London, Vol. 1

Burkert W 1996 *Creation of the Sacred*. Harvard University Press, Cambridge, MA

Campbell J 1949 *The Hero with a Thousand Faces*. Pantheon Books, New York

Chase R 1949 *Quest for Myth*. Louisiana State University Press, Baton Rouge, LA

de Vries J 1961 *Forschungsgeschichte der Mythologie*. Alber, Freiburg

Eliade M 1968 *The Sacred and the Profane* [trans. W R Trask]. Harvest Books, New York

Feldman B, Richardson R D 1972 *The Rise of Modern Mythology, 1680–1860*. Indiana University Press, Bloomington, IN

Frazer J G 1911–15 *The Golden Bough*, 3rd ed. Macmillan, London, Vol. 5

Girard R 1972 *Violence and the Sacred* [trans. P Gregory]. Athlone Press, London; Johns Hopkins University Press, Baltimore, MD
Hillman J 1975 *Re-visioning Psychology*. Harper & Row, New York
Jonas H 1963 *The Gnostic Religion*, 2nd edn. Beacon Press, Boston
Jung, C G 1968 *The Archetypes and the Collective Unconscious*, 2nd edn. Princeton University Press, Princeton, NJ
Lévi-Strauss C 1958 The structural study of myth (1955). In: Sebeok T A (ed.) *Myth*. Indiana University Press, Bloomington IN (reprinted in Lévi-Strauss 1963 *Structural Anthropology* [trans. Jacobson C, Schoepf B G]. Basic Books, New York)
Lévi-Strauss C 1978 *Myth and Meaning*. University of Toronto Press, Toronto, ON
Malinowski B 1926 *Myth in Primitive Psychology*. Kegan Paul, London; Norton, New York (reprinted in 1948 *Magic, Science and Religion and Other Essays*. Free Press, Glencoe, IL)
Raglan Lord 1936 *The Hero*. Methuen, London (reprinted in Rank O et al. 1990 *In Quest of the Hero*. Princeton University Press, Princeton, NJ)
Rank O 1914 The myth of the birth of the hero [trans. Robbins F, Jelliffe S E]. *Journal of Nervous and Mental Disease Publishing* (reprinted in Rank et al. O 1990 *In Quest of the Hero*. Princeton University Press, Princeton, NJ)
Ricoeur P 1967 *The Symbolism of Evil* [trans. Buchanan E]. Harper & Row, New York
Segal R A 1999 *Theorizing about Myth*. University of Massachusetts Press, Amherst, MA
Tylor E B 1871 *Primitive Culture*. Murray, London, 2 Vols.
Vernant J P 1983 *Myth and Thought Among the Greeks* [trans. not given]. Routledge & Kegan Paul, London
Wheelwright P 1968 *The Burning Fountain*, rev. edn. Indiana University Press, Bloomington, IN

R. A. Segal

Myths and Symbols: Organizational

The constructs 'organizational myth' and 'organizational symbol' first made their appearance in organizational and managerial studies in the second half of the 1970s, at a time of renewed and rapidly spreading interest in the cultural study of organizations. This interest was manifest in two main currents of study: the macroperspective of 'new institutionalism' (Meyer and Rowan 1977, Powell and DiMaggio 1991), which explored the ways in which cultural values shape organizational forms—and induce the organizations that populate a particular 'field' (a territory, an industrial sector, or a distinct sphere of institutional life) to grow similar rather than dissimilar—and the microperspective of 'organizational symbolism' (Dandridge et al. 1980, Pondy et al. 1983, Turner 1990), which viewed individual organizations as 'cultures' characterized by distinct paradigms, to be then analyzed not only in their instrumental and economic aspects but also in their ideational and symbolic ones, using holistic and interpretative research models (see *Organizational Culture*). In what follows, described first are the essential features of organizational symbolism, since this has been the intellectual context within which the constructs analyzed have been largely developed. Then presented is a typology of the symbolic languages to be observed in organizations, with brief descriptions of their main functions. Discussed finally is the function performed by the saga—defined as a 'collection of myths'—in shaping organizational identity, and the function of 'mediatory myths' in managing the dialectical relation between organizational ideology and operational necessity.

1. Organizational Symbolism

Born at the end of the 1970s as a marginal and nonconformist phenomenon, this intellectual movement rapidly grew into one of the main currents of thought in organizational studies. It produced an extraordinary flourishing of studies and research, gaining unusual popularity even outside the academic sphere. Most notable among the numerous surveys of this body of literature are the following: Alvesson and Berg (1992), a particularly thorough treatment which reviews the various theoretical perspectives that have developed internally to the movement, and discusses the relevance of symbolism to managerial practice; Trice and Beyer (1993), which provides a detailed picture of the material set out in numerous empirical research studies; and Jones (1996), which gives an overview of the main methodological issues connected with the study of organizational symbolism.

Although the approach employed by the movement's exponents was often presented as radically innovative, it was in fact deeply rooted in the social sciences. Despite their frequent use of anthropological terms and concepts, scholars of organizational cultures more frequently acknowledge their intellectual indebtedness to certain currents of sociological thought—in particular to phenomenological sociology, symbolic interactionism, ethnomethodology, and above all to the great ethnographic tradition of urban and occupational sociology (e.g., Whyte 1955, Becker 1951, Roy 1961). However, although an approach inspired by the same theoretical ideas had already been used in the study of formal organizations (see, in particular, Selznick 1949, Gouldner 1954, Blau 1955, Dalton 1959; Crozier 1963), until the early 1970s organizational and managerial studies were indubitably dominated by a rationalist and positivist paradigm which viewed the analysis of relations among objective and measurable variables as the central concern of research, and substantially neglected the notion of culture. Ironically, many of these analyses were conducted in the name of Max Weber (Ouchi and Wilkins 1985), regarded as the founding father of organizational theory, but they ignored Weber's invitation to see organizations as the expression of

cultural values, interpreting his bureaucratic model instead as the organizational model *par excellence* and with no alternatives (Clegg 1995). In one of the first essays devoted expressly to organizational symbolism, Dandridge et al. (1980) wrote that: '…a survey of major texts within the field of organizational behavior establishes clearly that there are virtually *no* references to the phenomenon that is the subject of this paper…For all practical purposes…it is as if the phenomenon did not exist or was not important.' In effect, in those years the statement that organizations have their own myths and symbols would have seemed at most a paradox, almost an oxymoron, to the extent that formal organizations—especially economic ones—were regarded as utilitarian forms of social aggregation in which individual and collective behaviors were governed mainly by norms of instrumental rationality (see *Organization: Overview*). If modern organizations are, by definition, the domain of a legal-rational ethos, and characterized by pragmatic behaviors shaped by the practical goals to which they are directed, how can these organizations possibly harbor expressive and disinterested behaviors shaped by impulses and emotions? And if these forms of behavior do exist, how can they be anything but secondary, irrelevant, or at any rate dysfunctional and therefore to be eliminated? Organizational symbolism challenges these prevalent opinions and invites us to read organizations in a completely different light. As Turner (1990) observes, '…the current growth of interest in organizational symbolism and corporate culture points to the end of an era. Looking anew at the organizational world we can see that it is a sensual and emotional realm, replete with its own ceremonies, rites and dramas.'

The cultural approach asserts that every productive practice is determined equally by the practical requirements from which it springs and by particular visions of the world. No productive practice is only and exclusively such, for it is simultaneously a symbolic practice, a way to appropriate reality by imposing one's own specific view of what reality is. Accordingly, scholars of organizational cultures have explored the interweaving of technical requirements and expressive needs, the ritualistic significance of apparently rigorously pragmatic behaviors and processes, the overlapping of ideological prejudice and mythic knowledge with scientific expertise. In other words, they have shown that culture and utility are woven together in even the most strictly utilitarian forms of social aggregation. From this perspective, therefore, organizations—and their environments—are conceived as 'symbolic fields': that is, as intersubjectively negotiated systems of meaning recognizable in a coherent set of symbols or conventional representations of important and complex aspects of organizational and social reality. Decoding an organization's symbolic field involves analysis of its answers to such fundamental questions as, for example, the reliability or unreliability of human beings, the symmetry or asymmetry of social relations, the type of relationship—dominance, harmony, or subordination—with the external environment, the nature of reality and of truth, and so on.

2. *Organizational Symbols: Types and Functions*

The definition of 'symbol' most widely used by scholars of organizational cultures is the one provided by Cohen (1976): 'Symbols are objects, acts, relationships or linguistic formations that stand ambiguously for a multiplicity of meanings, evoke emotions, and impel men to action.' Organizational symbols normally are grouped into three categories corresponding to the three main systems for expressing beliefs and cultural values within an organization: verbal language, behavioral language, and the language of objects.

2.1 *Verbal Symbols*

Principal among verbal symbols is language itself, with its ability to categorize and articulate experience using specific 'jargons,' which may differ even among organizations operating in the same context and speaking the same language. The category of verbal symbols also comprises linguistic artifacts like legends, stories, slogans, the names given to places, persons, events, and myths. The latter can be defined as dramatized accounts of events whose veracity is asserted as dogma, or is taken for granted, and which have the effect of legitimating and making desirable the behavior enacted in the events narrated and the ideas from which such behavior springs. To the extent that the cadence of organizational life is set by the alternation of crucial experiences—which jeopardize its identity and sometimes its very survival—with periods of relative stability, myths are often the idealization of the specific ways in which those difficulties have been overcome, and they shape the routines that orient behavior during periods of stability. Consequently, the myths most frequently encountered in organizations relate to their birth (myth of the origin) and to the transformations induced by traumatic events (the death of the founder, for example, or a slump in the market). In all organizations, the beliefs and values that myths embody exert a pervasive influence on decisions (see *Organizational Decision Making*), and it is difficult not to discern an element of 'mythic knowledge' even in decisions apparently taken solely on technical-scientific grounds. However, it has been pointed out that the more decisions are operational, and the effects of actions measurable, the more decisions are taken following principles of instrumental rationality, and the broader the range of options and viable choices available. By contrast, the more choices are general

and tied to the base values on which the organization grounds its identity, the more restricted the range of options, and the more mythical, prescientific and unamenable to the scrutiny of reason and experience becomes the knowledge utilized (Gagliardi 1986).

2.2 Symbolic Actions

Of particular importance among the symbolic manifestations represented by acts and styles of behavior are rites and ceremonials as emotionally charged collective actions—usually performed in a rigorously prescribed sequence—by means of which an organizational community celebrates its successes, heroes, and organizational ideals. Various classifications of organizational rites—which assume a wide variety of connotations—have been proposed, as well as analysis of their manifest and covert cultural consequences (Trice and Beyer 1984). Among the most frequently observable rites are:

(a) rites of passage (e.g., training seminars as a prelude to a promotion), which ease the transition of individuals to new roles and status;

(b) rites of enhancement (like the awarding of bonuses for outstanding performance), which consolidate the social identities of the rewarded and induce others to emulate them;

(c) rites of degradation (like the replacement of senior managers for some reason deemed unworthy), which solemnly and publicly reaffirm the importance of the social roles compromised by the behavior of the persons degraded;

(d) rites of integration (like the celebrations organized by businesses on special occasions like Christmas or the anniversary of the company's foundation), which encourage shared feelings of equality and participation in a common enterprise, temporarily suspending the norms which sanction differences in power and status, but implicitly reaffirm the adequacy of those rules in day-to-day life;

(e) rites of renewal (like the periodic drawing up of strategic plans), which reassure the organization's members that it is keeping up with the times and that present problems will be overcome, but at the same time bolster the legitimacy of existing systems of power and authority.

2.3 The Language of Artifacts

A third system of symbols is constituted by material artifacts, or the intentional products of human action which exist in the organization independently of their creator and which can be perceived by the senses: products, images, buildings, furnishings, arrangements of physical space. Given the durability of physical matter, artifacts are able tenaciously and incessantly to transmit particular messages and cultural stimuli, thereby encouraging the diffusion and sharing of special 'modes of feeling' in ways that are all the more efficient because they evade intellectual control (Gagliardi 1990).

3. The Role of the Saga in the Process of Institutionalization

The organizational myths of origin and transformation are often blended together in a 'saga' or an account of an 'extraordinary' series of events which traverses and unifies the various phases of the organization's evolution. In a saga, historical facts are embellished by subsequent re-elaborations, inconsistencies are eliminated or justified, and the rational explanation of how certain means lead to certain ends acquires an affective connotation which transforms the workplace into a loveable and loved institution. The saga therefore performs an essential role in the process of institutionalization: that is, in the process by which the organization, from a sterile mechanical apparatus for the efficient distribution and coordination of tasks, becomes imbued with values and acquires a distinctive 'character.' This character on the one hand underpins the collective identity and sentiment of belonging of the organization's members, while on the other it legitimates the organization in the social environment (Selznick 1957). 'Whether developed primarily by management or by employees, the story helps rationalize for the individual his commitment of time and energy for years, perhaps for a life-time, to a particular enterprise. Even when weak, the belief can compensate in part for the loss of meaning in much modern work, giving some drama and some cultural identity to one's otherwise entirely instrumental efforts. At the other end of the continuum, a saga engages one so intensely as to make his immediate place overwhelmingly valuable. It can even produce a distortion, with the organization becoming the only reality, the outside world becoming illusion, (Clark 1972).

To the extent that the saga generates a shared conviction that the organization is unique because it has been the protagonist of an unusual enterprise crowned by extraordinary success—a conviction which induces individuals to remain within the system, endeavoring to improve it rather than pursuing their interests elsewhere—the saga may constitute a precious organizational resource. Indeed, ensuring the stable participation and constant commitment of its members is one of the crucial problems faced by any co-operative system constructed to pursue specific ends.

4. Mediatory Myths

In that, as mentioned in Sect. 1, the dialectical relation between expressive needs and pragmatic exigencies unavoidably conditions the lives of organizations, a

further crucial organizational problem is the management of the contradictions, tensions and psychological costs that derive from that relation: mediatory myths are discursive formulations which make contradictions acceptable.

Unlike the myths of origin and transformation, which mainly orient actions to be undertaken, mediatory myths provide retrospective justification for previous actions and states of affairs. The need to justify concrete action arises, first, from the fact that different values may conflict in specific situations—and it is difficult to comply with one principle without being in breach of another—and second from the fact that actions always tend, actually or apparently, to fall short of the ideals that have inspired them, so that there is invariably a gap between 'what we should do' and 'what we are able to do,' between the purity of the ideal and the impurity of the concrete organizational practice. Mediatory myths harmonize conflicting principles, not by means of logical arguments but by means of assertions of fact, and they resolve or reconcile contradictions by simply denying and inverting them, or by holding irreconcilable alternatives in suspension. Abravanel (1983) reports an interesting example of a mediatory myth: that of the 'wild ducks' institutionalized at IBM when T. J. Watson Jr. was chief executive officer. The contradiction between the ideology encouraged by Watson, namely that the company's employees were free to express their individuality, and a corporate reality characterized by rigid training and socialization practices was resolved by institutionalizing the myth that the employees were like wild ducks: they could not be tamed but nevertheless flew in formation and did not choose their migratory route.

See also: Competitive Strategies: Organizational; Corporate Culture; Information and Knowledge: Organizational; Innovation: Organizational; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; Leadership in Organizations, Sociology of; Learning: Organizational; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations, Metaphors and Paradigms in; Organizations, Sociology of; Rational Choice and Organization Theory; Strategy: Organizational; Symbolism (Religious) and Icon

Bibliography

Abravanel H 1983 Mediatory myths in the service of organizational ideology. In: Pondy L R, Frost P J, Morgan G, Dandridge T C (eds.) *Organizational Symbolism*. JAI Press, Greenwich, CT

Alvesson M, Berg P O 1992 *Corporate Culture and Organizational Symbolism: An Overview*. Walter de Gruyter, New York

Becker H S 1951 The professional dance musician and his audience. *American Journal of Sociology* **57**: 136–44

Blau P M 1955 *The Dynamics of Bureaucracy*. Wiley, New York

Clark B R 1972 The organizational saga in higher education. *Administrative Science Quarterly* **17**: 178–84

Clegg S R 1995 Of values and occasional irony: Max Weber in the context of the sociology of organizations. In: Bacharach S B, Gagliardi P, Mundell B (eds.) *Studies of Organizations in the European Tradition*. JAI Press, Greenwich, CT

Cohen A 1976 *Two Dimensional Man: An Essay on the Anthropology of Power and Symbolism in Complex Society*. University of California Press, Berkeley, CA

Crozier M 1963 *Le phénomène bureaucratique*. Editions du Seuil, Paris (1964 *The Bureaucratic Phenomenon*. University of Chicago Press, Chicago)

Dalton M 1959 *Men Who Manage*. Wiley, New York

Dandridge T C, Mitroff I, Joyce W F 1980 Organizational symbolism: A topic to expand organizational analysis. *Academy of Management Review* **1**: 77–82

Gagliardi P 1986 The creation and change of organizational cultures: A conceptual framework. *Organization Studies* **2**: 117–34

Gagliardi P 1990 *Symbols and Artifacts: Views of the Corporate Landscape*. Walter de Gruyter, New York

Gouldner A W 1954 *Patterns of Industrial Bureaucracy*. Free Press, New York

Jones M O 1996 *Studying Organizational Symbolism: What, How, Why?* Sage, Thousand Oaks, CA

Meyer J W, Rowan B 1977 Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology* **2**: 340–63

Ouchi W G, Wilkins A L 1985 Organizational culture. *Annual Review of Sociology* **11**: 457–83

Pondy L R, Frost P J, Morgan G, Dandridge T C (eds.) 1983 *Organizational Symbolism*. JAI Press, Greenwich, CT

Powell W W, DiMaggio P J 1991 *New Institutionalism in Organizational Analysis*. University of Chicago Press, Chicago

Roy D R 1961 Banana time: Job satisfaction and informal interactions. *Human Organization* **18**: 158–68

Selznick P 1949 *TVA and the Grass Roots*. University of California Press, Berkeley, CA

Selznick P 1957 *Leadership in Administration*. Harper & Row, Evanston, IL

Trice H M, Beyer J M 1984 Studying organizational cultures through rites and ceremonials. *Academy of Management Review* **4**: 653–69

Trice H M, Beyer J M 1993 *The Cultures of Work Organizations*. Prentice Hall, Englewood Cliffs, NJ

Turner B A 1990 The rise of organizational symbolism. In: Hassard J, Pym D (eds.) *The Theory and Philosophy of Organization. Critical Issues and New Perspectives*. Routledge, London

Whyte W F 1955 *Street Corner Society: The Social Structure of an Italian Slum*. University of Chicago Press, Chicago

P. Gagliardi

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

N

Narrative Comprehension, Psychology of

Narratives are pervasive in people's lives: fairy tales, news reports, novels, historical accounts, people's communications about their personal experiences, and so on, all take the form of narratives. They also come to us via a wide range of media. The majority of psychological investigations of narrative comprehension have focused on narratives in textual form, but narratives occur in auditory (e.g., oral narratives, interpersonal discourse), visual (e.g., silent movies), and mixed forms (e.g., movie/television, comics) as well. Finally, we process narratives for many different reasons: entertainment, transmittal of social and cultural information and traditions, scientific communication (e.g., history), legal arguments, reading instruction, to convince, to elicit desired actions, to provide coherence to our existence, etc. This article provides an overview of the psychological processes that are involved in the comprehension of narratives and in the construction of a coherent memory representation. An understanding of these processes is important not only because it provides insights into narrative comprehension itself (which is of theoretical and practical—e.g., educational—interest) but also because it constitutes a window to the mind, a view of complex human cognitive behavior.

1. What Does it Mean to Comprehend a Narrative?

Narratives relate how a series of events, fictitious or real, unfold. In doing so, they provide more than a simple list of the individual events: they also indicate how these events are interrelated and interdependent (etymologically, the roots of the term 'narrative' and its cognate 'story' are ancient forms of 'knowing/understanding,' indicating the interpretative and explanatory nature of narratives). Analogously, the goal of narrative comprehension is to construct a memory representation in which the various individual parts of the narrative are connected in meaningful ways (e.g., see *Mental Representations, Psychology of*). Psychological examination of narrative comprehension includes consideration of the nature of this representation, of the processes that take place during its construction, and of the connection between these processes and the resulting representation.

2. Comprehension of Narrative Texts

Narrative comprehension is investigated most extensively in the domain of reading (van den Broek 1994, van Dijk and Kintsch 1983; see also *Memory for Text*). There are several reasons for this. Reading is one of the most uniquely human and complex of all cognitive activities, with enormous social and educational importance. Moreover, texts allow researchers considerable experimental control over the materials they present to their participants.

2.1 Memory Representation of Narratives: Inferences and Coherence

When reading is successful, the result is a coherent and usable mental representation of the text. This representation resembles a network, with nodes that depict the individual text elements (e.g., facts, events, settings) and connections that depict the meaningful relations between the elements. These relations are what make a representation coherent and usable. Many possible relations exist, but two types have been found to be particularly important for establishing coherence: referential and causal/logical. Referential relations enable readers to keep track of elements such as objects, people and events mentioned in several places within the text (Kintsch and van Dijk 1978; see also *Propositional Representations in Psychology*). Causal and logical relations enable readers to identify how different events or facts depend on each other (Schank and Abelson 1977, Graesser and Clark 1985, Trabasso and van den Broek 1985). For example, when reading 'The lady gave the waiter 10 Euros. He returned to give her the change,' referential coherence is established by recognizing that 'he' in the second sentence refers to the waiter and 'her' to the lady. Causal/logical coherence is established by inferring that the waiter is returning change because the lady consumed something in an establishment where the waiter works, and because the lady had given him 10 Euros for a purchase of a lesser amount.

Several properties of coherent representation of narrative texts are to be noted. First, the identification

of relations requires readers to make an inference, that is, to integrate the textual information with their semantic knowledge (or background knowledge). In the above example the requisite semantic knowledge (e.g., that 'she' tends to refer to a female protagonist) is likely to be readily available to most readers, but in other situations it is not. For example, many readers have difficulty recognizing the causal relation between the following two sentences: 'The presence of the moon exerted a strong gravitational pull on the earth. Thus the moon contributed to the emergence of life on earth'; even though the referential relation is direct (i.e., 'moon') and the causal relation is explicitly marked by linguistic means ('thus'). The reason is that most readers lack the semantic knowledge necessary for the inference: The moon's gravitational pull is instrumental in creating the earth's magnetic field which, in turn, shields the earth's surface from lethal cosmic radiation. The semantic knowledge that readers bring to bear includes both individual facts (as in the above example) and generalized assemblies of knowledge (e.g., the sequence of events in 'going to a restaurant'; see *Schemas, Frames, and Scripts in Cognitive Psychology*).

Second, many of the referential and causal/logical relations that readers must infer are much more complex than the simple ones in the 'lady and the waiter' example. Relations may extend over long distances in the text, may be difficult to interpret, or may require coordination of numerous pieces of information. Third, the text's linguistic structure may aid the reader by explicitly marking a relation, thereby making the inference more obvious. Linguistic markers take various forms, ranging from inflections to explicit connectives (as the word 'thus' in the above example). Fourth, the number of relational inferences required for a coherent representation usually is fairly large, even if many are generated without effort or awareness by skilled readers.

Considerable empirical evidence supports the key role that referential and causal/logical networks play in successful comprehension. The evidence is too extensive to be reviewed in its entirety (for a review, see van den Broek 1994), but some examples may be helpful: (a) the more relations events have, the more frequently they are recalled and the more important they are judged by readers; (b) answers to questions about the text tend to follow the connections captured by the networks; (c) when reminded, after reading, of an event or fact from the text, readers spontaneously and automatically bring to mind other elements directly related to it but not unrelated ones from the same text; and (d) the time it takes to read a sentence is proportional to the number of inferences needed for comprehension. Thus, referential and causal/logical relations form the backbone of the mental representation of narrative texts.

Although referential and causal/logical relations are the most prevalent in narratives, other types of relations such as spatial and temporal ones can also be generated. These relations are less central to comprehension, however, and generally are only included in the memory representation if they are functional in establishing referential or causal/logical coherence.

The network of nodes and semantic relations constitutes the fundamental representation of a narrative text. It also forms the basis for further organization within the narrative. This is the case because, within a network, events are clustered into episodes, each of which revolves around a protagonist's goal. The nature of episodes was first explored in the 1970s, with the development of story grammars (e.g., Mandler and Johnson 1977). Following anthropological work on folktales, story grammarians noted that story statements can be categorized into a small number of content types and, furthermore, that the different types are sequenced according to a small number of grammatical rules. These story grammars differ in detail and use different labels to designate categories, but they show substantial overlap. In general, stories are described as consisting of a number of Setting statements and one or more episodes. The *Setting* statements provide the backdrop for the remainder of the narrative by describing main characters, objects, temporal and geographic information, and so on. Episodes describe the events and actions of the narrative proper. Each episode is organized around a goal and its outcome. Within an episode, events differ in function. In the simplest episode, an *Initiating Event* describes the occurrence of an event that sets the train of action in motion. It elicits an *Internal Response*, describing the reactions of the protagonist, which in turn causes the implicit or explicit establishment of a *Goal* on the part of the protagonist. The goal motivates one or more *Actions*, which lead to the *Outcome* of the episode. The outcome results in a *Reaction*, describing the protagonist's reaction to the success or failure of the outcome. Most narratives have multiple instantiations of each category, with the grammatical rules dictating their order. Although story grammars now generally are considered derivative from the various semantic relations that exist between narrative elements, they serve an important function in drawing attention to the semantic structure of narratives.

The episodes in a narrative are themselves interrelated, forming more complicated macrostructures (see *Macrostructure in Discourse Comprehension, Psychology of*). A particularly important episode structure results when episodes arrange into a hierarchy, with one goal (or outcome) causing the establishment of another goal. This happens, for example, when one goal is pursued in order to attain another: the former is subordinate to the latter, its superordinate. Goals higher in the hierarchy form the main ideas of the narrative, with the most superordinate goal representing the narrative's theme. Other macrostructures exist: linear (e.g., streams of consciou-

sness), intertwined (e.g., maintaining different story lines), competing strands (e.g., some genres of detective stories), and so on.

2.2 The Comprehension Process: Attention and Inferences

The memory representation of a text is the result of inferential processes that take place during reading. These processes allow the reader to establish coherence as he or she proceeds through the text. In general, the more relations a reader detects, the more coherent the text representation is and the better comprehension will be. Limitations in the reader's information processing system restrict the generation of large numbers of inferences, however. Extensive inferential processes would overwhelm the reader's attentional capacity (see *Working Memory, Psychology of*) or slow down reading to a snail's pace. Instead, only a subset of all possible inferences is generated.

Which subset is generated depends on a number of factors. The primary of these factors concerns the *function* of a possible inference in establishing coherence. At each reading cycle (i.e., each new sentence), the comprehender attempts to make sense of the information explicitly stated in the current sentence by inferring connections to his or her background knowledge and/or statements in the prior text. Thus, the most important types of inferences are those that provide coherence to the text, in particular those that resolve referential ambiguity and provide causal and logical explanations. These are required for comprehension. Results from a large number of studies indicate that these coherence-building inferences are routinely made during reading. For example, when readers are interrupted during reading by the presentation of a word, they recognize or speak it aloud more quickly if the word corresponds to one of these inferences than if it does not. Thus, the concept represented by this word apparently already was activated (i.e., inferred) before the interruption. Similar evidence indicates that other types of inferences (e.g., spatial, temporal, instrumental) usually are made only when they also provide referential or causal coherence or when the reader has special goals in reading.

Other factors determining whether a particular inference is made are the extent to which the semantic information in the text and the linguistic cues provide sufficient constraints for the potential inferences, the reader's background knowledge relevant to the inference, the reader's attentional resources, and any special objectives in reading the text.

2.3 The Construction of a Memory Representation

The processes that take place during comprehension ultimately lead to a memory representation of the text. Several theoretical models capture the mechanisms by which this is accomplished (see van Oostendorp and Goldman 1999 for descriptions of several prominent models; see also *Text Comprehension: Models in Psychology*). These models differ in their details, but they are remarkably similar in their general description of the process by which a memory representation emerges from the processes during reading.

Central is the notion that reading involves a continuous and dynamic fluctuation of activations. This is a direct consequence of a reader's limited attentional capacity and of the inferential processes described above. At each reading cycle, new concepts are (re)activated, some old ones are retained, and others are removed from the focus of attention. Sources of activations are (a) the currently read text, (b) carry-over from the preceding cycle, (c) the reader's semantic knowledge, and (d) the reader's episodic memory representation of the text read so far. Carry-over from the preceding cycle is limited, and primarily includes theme or goal information. Background knowledge is accessed either through an automatic activation of concepts associated with the ideas in the current sentence or through a more strategic search for relevant information (see *Knowledge Activation in Text Comprehension and Problem Solving, Psychology of*). Finally, information from preceding sentences can be reactivated by retrieval ('reinstatement') from the episodic memory representation of the prior text or by consultation of the actual text ('looking back').

With each cycle, the memory representation of the narrative is updated. New text elements are added and existing ones increased in strength as a function of their activation. In general, the more often and the more strongly an element is activated over the course of reading, the more strongly it will feature in the final representation of the narrative. Furthermore, when two text elements are activated simultaneously, meaningful relations between them can be identified. The more often and more strongly two text elements are coactivated, the stronger their relation in the memory representation will be.

In this fashion, the memory representation gradually evolves out of the processes that take place during reading. The relation between process and product is reciprocal: at each cycle, the activation of concepts is partly determined by the memory representation that resulted from the preceding cycles; the new activations, in turn, update this representation.

2.4 Variations and Veracity in Narrative Comprehension

Individual readers differ in attentional capacity, cognitive skills (e.g., coherence-building strategies), personal and cultural background knowledge, and so on. These variations alter the comprehension processes and, hence, the representation of the narrative. Moreover, readers also differ in terms of their *standards for*

coherence. The present account has focused on the standards of referential and causal/logical coherence, which are basic and shared by most fluent readers. However, other—stricter, more lenient, or simply different—standards may be employed. For example, different representations result when people are asked to read a narrative taking place in a house with the perspective of a potential homebuyer or that of a burglar (Bransford and Johnson 1972). Moreover, standards vary as readers have different goals for reading the narrative (e.g., for a test, for entertainment, to learn about the spatial layout). Similar variations on the above description of narrative processing and representation occur within individuals from one reading situation to the next.

Further, a representation that is internally coherent need not be identical with that intended by the author of the narrative. Indeed, the overlap between intended and obtained representations is likely to be partial at best. In some situations (e.g., educational contexts) comprehension may require that the overlap is substantial, whereas in others (e.g., entertainment) it may not.

3. Comprehension of Narratives in Other Contexts

Narratives exist in many contexts outside reading. Examples are television and movie viewing, oral communication, historical accounts, and so on. The comprehension of narratives in these contexts has received far less attention than that during reading. Moreover, most extant research has focused on the product of comprehension (the representation) rather than on the process. Nevertheless, the available evidence suggests that the processes involved in the comprehension and representation of narratives are remarkably similar across contexts. For example, causal relations have been found to play a central role in the processing of television and movie narratives, of orally presented narratives, and even in the interpretation of arguments in courts.

4. Future Directions

The psychological investigation of narrative comprehension is likely to proceed in several directions. One direction is the further development of the computational models of reading comprehension described above. Aided by increasingly powerful computers, these models are likely to address hitherto underexplored issues such as the mechanism by which semantic knowledge is recruited during comprehension, how semantic knowledge itself is altered by the comprehension of individual texts, and so on. A second direction is the investigation of the neurological bases for these and other comprehension processes. Such investigations will be focused not only on *where* in the brain certain processes take place, but also on the theoretical implications of such determination.

Third, the affective component of narrative comprehension will be investigated and the results integrated with those of the cognitively oriented models developed so far. Clearly, narratives elicit strong emotions on the part of the comprehender; indeed, such emotions are part of what makes narratives interesting. A final direction will be the extension of the detailed models developed for narrative comprehension to comprehension in other domains, both in reading and in other contexts. Thus, the study of narrative comprehension truly becomes a study of the human mind in general.

See also: Inferences in Discourse, Psychology of; Inferences, Psychology of; Models, Metaphors, Narrative, and Rhetoric: Philosophical Aspects; Narrative, Sociology of

Bibliography

Bransford J D, Johnson M K 1972 Contextual prerequisites for understanding: some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior* **11**: 717–26

Graesser A C, Clark L F 1985 *Structures and Procedures of Implicit Knowledge*. Ablex, Norwood, NJ

Kintsch W, van Dijk T A 1978 Towards a model of text comprehension and production. *Psychological Review* **85**: 363–94

Mandler J, Johnson N S 1977 Remembrance of things parsed: story structure and recall. *Cognitive Psychology* **9**: 111–51

Schank R C, Abelson R 1977 *Scripts, Plans, Goals, and Understanding*. Erlbaum, Hillsdale, NJ

Trabasso T, van den Broek P 1985 Causal thinking and the representation of narrative events. *Journal of Memory and Language* **24**: 612–30

van den Broek P 1994 Comprehension and memory of narrative texts: inferences and coherence. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, pp. 539–88

van Dijk T A, Kintsch W 1983 *Strategies of Discourse Comprehension*. Academic Press, New York

van Oostendorp H, Goldman S R (eds.) 1999 *The Construction of Mental Representations During Reading*. Erlbaum, Hillsdale, NJ

P. W. van den Broek

Narrative, Sociology of

Narratives are ways of linking together two or more events that have taken place in the past, making sense of historical or fictional personae. There exists a

multitude of genres—such as films, comic strips, Homeric epics, creation myths, novels, fables, case studies, hagiographies, and biographies—that, in turn, draw on diverse academic traditions ranging from legal studies, mass communication, and history to literature and hermeneutics. This review explores recent work on narratives, focusing on discursive or conversational approaches to narratives as elements of social action. In contrast to an analysis of fixed narratives, an analysis of narrative action does not merely involve issues related to domination such as the nature of grand narratives in relation to that of taboo or subordinated narratives. It also involves the conversational work of narrators and their co-participants.

1. The Narrative Turn

During the last two decades of the past century, social theory has undergone something of what could be called a narrative turn. In the post-war period, Wittgenstein's work on language games has broken new ground in legitimizing the study of language use. Today, there is a focus, not only on language games and fuzzy concepts, but on personal texts, life-stories, and inter-textuality. Some of this new emphasis can be traced to Bakhtin, Kristeva and the importation from literary theory into the social sciences. At the beginning of the twenty-first century, narration and narrative modes are cast as key elements in several theories oriented toward the epistemology of social action (e.g., Bakhtin 1984, Foucault 1997, Rorty 1980).

All such narrative models position language and the role of narration in the foreground during the formation of minds. In fact, narratives are the basic building blocks for communities, in that the very notion of community is often based on historically grounded, or oppositional, narratives. Similarly, social sciences are based on local ways of building narratives and on the type of language varieties involved (Rorty 1980). Key concepts in narrative models of self-identities are analyzed in a recent review by Ochs and Capps (1996). Some contemporary models are deliberately 'mindless' in that they are oriented towards procedural aspects of narration: how narratives evolve sequentially and how they are co-construed or deconstrued as part of social action (Drew 1992, Sacks 1992, Schegloff 1997).

Many mechanisms in the orchestration of narration were first described by Sacks (1992), who investigated the interactional features of a variety of narrative practices, such as telephone calls to a suicide center, problem formulations in therapy, and jokes among friends. Sacks' (1992) lectures on conversation—recorded between 1964 and 1972—primarily discussed the turn-taking machinery, the dialogical and sequential organization of conversational turn-taking. Yet his lectures also involve entitlement to stories, tellability, and what he called 'a technology for building stories.' More recently, it has been discussed how 'talk at work'—that is, dialogic mechanisms for story-telling and speaking at large—can be seen in courtrooms and other institutional settings (Drew 1992, Lynch and Bogen 1996). Such processual interactional analyses of narratives all diverge from traditional narrative analyses that have normally been focused on static structural aspects of narratives (plots, abstracts, resolutions, etc), not on the story telling.

In folkloristic and ethnographically oriented sociology, life stories have been key methods in the formulation of reflexive analyses of everyday action. Traditionally, interviews have been the main tools for eliciting life stories or story fragments. However, what has been analyzed is normally the life story as such, not the story telling.

In the 1990s, anthropological work on life narratives has been directed both to issues about how narratives are linked to the formation of selves in different cultures and to procedural aspects; that is, to story-telling. In their analyses of agoraphobia in ordinary family conversations, Capps and Ochs (1995) illuminate how panic is displayed and construed in the everyday narratives of a family where one member suffers from agoraphobia. Through her mundane panic stories, the mother in the family reveals her fears, but she also passes on patterns for how to construe the world as a fearful place.

The dinner-time narrations in the agoraphobic family can be seen to generate fragments of life stories. Analyses of such story-telling practices are particularly illuminating in that they combine life-story material with natural data that is not primarily elicited for the benefit of the interviewer.

One of the main methodological controversies in recent work on narration concerns the relative role of text and context. The work of Foucault, for instance, as well as ethnographic and traditional sociological work, is primarily focused on contextual aspects of discourse. In contrast, conversation analysis is primarily oriented to the text as such—and to intersubjectivity (the dialogic production of narratives). Context is focused only in those cases in which it is invoked by the participants themselves (cf. Schegloff 1997). In conversation analysis and the growing field of ethno-methodology, phenomena such as dominant narratives are thus primarily seen as participants' phenomena. It is only if domination is invoked by the participants themselves that it is, in fact, taken up in the researcher's analyses.

2. Technologies for Building Stories

Much research on narratives has centered on the structure of plots in relation to various related elements (such as background, future events, introduction of main characters), and the temporal order of the

narrative has long been a key element of narrative orgnization as in Labov's (1972) classical analyses of oral narratives of Afro-American inner-city youth. Minimally, a narrative in Labov's scheme involves two temporally ordered events. In the more elaborate type of fight stories and other personal narratives that he elicited from inner-city boys, it is often possible to discern a more complex pattern involving abstract, orientation, complicating action, evaluation (what's the point?), resolution, and coda (is there a signal that the story is over?).

Yet, the sequential order of story telling need not coincide with the temporal organization of the tale. In a critical discussion of narratives and cognition, Edwards (1997) demonstrates ways in which script models and other structural models (for example, Labovian schemes) in fact do not hold up in the face of the complexities of mundane interaction; they collapse when they are applied to real life on-line narration in interactional contexts.

Interactional analyses of story-telling dialogues are therefore important tools for sharpening our reasoning on narration as social action. Such work focuses on participants' orientations to co-participants and to the narrated events, rather than primarily or only on the temporal framework of the story. The narrator's stance or point of view can often be detected in linguistic cues such as word order, tense aspect marking, case marking, verb voice, and pronoun choices (cf. Ochs and Capps 1996). In some cases, linguistic resources can be seen to index stances that are closely related to conversational genres or institutional voices (Bakhtin 1984, Lynch and Bogen 1996). In interactional work on narration, genre is therefore not merely seen as a background variable for talk, but rather as the outcome of the interaction as such. In many institutional encounters, for instance, it is the powerful party who engages in authoritative moves, such as reformulating what the story is about or doing so-called formulations (Sacks 1992). Conversely, this means that someone may position himself/herself as a person in power through exploiting discursive devices that are associated with institutional authority.

In several institutional practices—e.g., medical consultations, trials, and psychotherapy—an important part of conversational action is to identify what is the problem. Joint or competing problem formulations are therefore key elements in such institutional narrations (e.g., Aronsson and Cederborg 1994, 1997). Narrations that are related to self presentations entail substantial moral work in that they project different positions in given social orders. Yet, in conversational studies of narration, control is not merely related to fixed positions in hierarchical organizations. In Goffman's (1971) analyses of social interaction and territoriality in public places, narration is principally studied in terms of remedial work, as in apologies and accounts. More recently, others have explored defensive aspects of narration in more depth, specifying how participants present accounts of their actions, denying, justifying, or excusing themselves in the light of implicit or explicit blame (Buttny 1993).

3. *Narrations as Story-telling Contests*

In adversarial courtroom proceedings, attorneys can be seen to employ questions as tools for eliciting the winning story. Trials can thus be considered story-telling contests, in which each attorney must ask questions in such a way that his/her client tells the most convincing and victorious story. Credibility is thus largely a matter of narrative design, performed by the witness or defendant and masterminded by the attorney.

Such story-telling contests underlie various actions in adversarial courtroom contexts. Witnesses are regarded as members of either side's team (the prosecution or defense side), which means that, for instance, rape witnesses/victims are treated as prosecution team members, who are at times exposed to discrediting and damaging questions. In an analysis of an American rape trial, Drew (1992) showed how such trials may yet involve substantial resistance and strategic defense on the rape victim's part. In his detailed analyses of an extended episode from a rape trial, he showed that the rape victim employed alternative descriptions as well as other discursive devices in her story (of being merely an acquaintance to the rapist) as opposed to that of the prosecution (of her being a lover or willing partner). Yet, the two competing stories were told as fragments, sequentially revealed in the attorney's questions and in the witness' defensive and oppositional responses.

An early study of oppositional work in legal settings can be seen in Cicourel's (1968) monograph on juvenile delinquents and the conversational work of parents, probation officers, and various experts such as psychiatrists. Cicourel did not detail the conversational work as such, but he discussed the defensive interactional work of parents, showing, for instance, how middle-class parents recruited outside assistance from experts such as therapists and teachers, working on recategorizing their sons as depressive rather than as delinquent, or as temporarily influenced by peers rather than as active and wilful. The ultimate written police record categorizations of the same delinquent actions were therefore partly the result of extended parent–expert negotiations.

Competing narratives can, of course, also be found outside of legal settings. In an analysis of clients' voices in family therapy, Aronsson and Cederborg (1994) demonstrate how family therapy episodes often involve underlying story-telling contests in which one side ratifies the same side story, whereas the other side's story is deauthorized in that the story point is challenged or undermined. The therapist plays a key role as an intermediary—an orchestrator of who talks

to whom about what—and as arbitrator about when a narrative is ultimately finished. Therapists' control in family therapy is more concerned with issuing story-telling rights than with, for instance, sheer dominance over turn taking. Such implicit norms concern what is reportable or tellable, as well as story boundaries, that is, when a story is to be told as well as implicit norms for when it is finished. In another analysis of family therapy narratives, a love story was jointly construed and deconstrued by a young girl, on the one hand, and by her mother and therapist, on the other (Aronsson and Cederborg 1997). Here, it was shown how past and future time was construed in radically different ways by the two sides. An important part of the therapeutic work was to realign the two parties' diverging time categorizations with respect to parameters such as 'suitable age,' 'proper timing,' and 'normal duration and frequency.' The girl's story was, however, told as a series of fragmented justifications and excuses; that is, defensive accounts, whereas her mother's story was told as an authoritative problem formulation.

In his micro-sociological analyses of parental story telling in the face of alternative legal versions, Cicourel (1968) shows that detailed ethnographic data need not exclude a concern for macrosociological issues such as class. Detailed analyses of institutional encounters simultaneously involve micro- and macro-concerns. This means that the radical split between micro and macro is partly an artificial controversy inextricably linked to monolithic models of institutions and the type of work that institutions do.

Moreover, the cited work on defense and resistance (Aronsson and Cederborg 1994, 1997, Cicourel 1968, Drew 1992) in the face of authoritative legal and clinical narratives demonstrates that simple models of domination or cultural reproduction (e.g., mechanistic Foucauldian analyses) do not fully explore what happens in adversarial contexts. There is also a need for a more complex analysis of adversarial talk that could account for participants' resistance.

4. *Technologies of Selves*

On a historical societal level, Foucault has discussed how dominant stories have been employed in the missionization and colonialization of selves. In his discussion of technologies of selves, Foucault (1997) presupposes that our public selves are intimately linked to conversational practices. In his analysis of the historical roots of self-technologies, these proposed technologies were closely linked to self-reflexivity practices and to early forms of self-confessions. The self was narrated in more or less monological formats, e.g., in letters to friends, as in the case of early Greco-Roman philosophy, or in confessions in monasteries in early Christian spirituality. In the first case, the focus was on deeds and actions (what I have done in relation to what I should have done) and in the case of early Christian spirituality on thoughts (what I have been thinking in relation to what I should have been thinking).

In modern society, confessions have become integrated into what Foucault calls truth regimes that organize institutional practices in normative patterns for what should be told to whom and in what ways. In modern life, there are several discursive genres that can be related to the formation—and reflection—of moral order and selves in a Foucauldian universe. Interviewing, teaching, counselling, and therapy can be seen as prototypical discursive technologies for eliciting what Foucault calls confessions and authoritative responses to such confessions, e.g., judgments, evaluations, celebration, and disciplining (Foucault 1997). Such genres are important in modern discursive practices, but Foucault himself does not detail how such identity work is done on a conversational level.

In the realm of interactional theories, however, the micro-politics of identity work in institutional settings has been documented in a series of analyses of contemporary institutional narratives. There are many genres of such narratives: couple therapy (Buttny 1993), family therapy (Aronsson and Cederborg 1994, 1997), and, of course, witnesses' and defendants' narratives in court (Drew 1992, Lynch and Bogen 1996). All these genres have implications for identity formation, both for displaying identities and for creating identities. This Janus-face nature of conversations and selves is also an essential feature of ethnographic and conversational work on narratives and self-identities.

The intersubjective architecture of confessions has been discussed in Sacks' (1992) lectures on conversation. In telling a first story, a speaker may invite a second story, and categorizations and first assessments often precede second assessments. In everyday conversations between friends, disclosures of secrets or taboo or delicate information may similarly invite disclosures by co-present parties. Disclosures similarly tend to assume a sequential structure, involving fragmented and gradual revelations. Similarly, Goodwin (1990) demonstrated the complex sequential structure of young girls' multi-party story telling about other girls by showing how they would animate characters and use extensive reported speech in casting an absent girl as blameworthy, while simultaneously securing the alignment of co-present parties. Ultimately, the blame narrative was therefore a jointly created team product, rather than a story told by one person alone.

The written record is obviously an important parameter in technologies of selves. Cicourel (1968) revealed ways in which written police records were ultimately the products of talk; that is, the products of extended oral negotiations. Recently, Lynch and Bogen (1996, p. 203) have specified many ways in

which courtroom dialogue may be 'on the verge of breaking into writing,' as in the Iran–Contra hearings in which prepared written statements, documentary exhibits, citations of prior testimonies, and written notes played a prominent role in the defendant's 'spontaneous' oral response to questions in the public hearing.

Written narratives tend to look more objective or factual than oral narratives. Yet, many written documents obviously emanate from talk-in-interaction. The delicate interplay between written and spoken language can also be seen in other institutional settings than courtrooms. In his work on answering patterns at a suicide prevention center, Sacks (1992), for instance, showed that the professional's answering sheet influenced the order in which s/he asked questions, illustrating how the design of written forms and records is important for the formation of oral narratives.

5. *Collusion, Genres, and Narrative Roles*

Narratives in interaction are often contested, as language use is never disinterested—narrators always direct their narratives to someone, and they wittingly or unwittingly speak on somebody's behalf (Bakhtin 1984). In Bakhtin's epistemology of a multi-voiced or polyphonic universe, multi-party encounters always involve potentials for secret alliances and collusions. Such collusion can be seen in several different institutional contexts, such as in asymmetrical encounters when there are dominant or authoritative voices, on the one hand, and subordinate voices, on the other.

In detailed analyses of small claim court interactions, it has been shown that narrative genres at times have collusive qualities in that relationship-oriented and informal narratives need not be more egalitarian and democratic than more formal rule-oriented narratives (Conley and O'Barr 1990). In fact, inexperienced persons are often more at a loss in informal contexts, in that they are required to construe legal narratives on their own, whereas formal settings offer more support from legal professionals.

In a Foucauldian universe, education is seen as one of the technologies of the self. Yet education is not merely an issue for formal schooling, but also for informal schooling in family contexts. In work on family dinner conversations, it has been shown that narratives are deeply embedded in family politics that are, in turn, shaped by cultural norms for participation. In their work on American family dinner conversations, Capps and Ochs (1995) demonstrated recurring narrative patterns, reflecting silent opposition and underlying family hierarchies.

Silent opposition can also be found in public events. In legal contexts, witnesses are expected to be compliant. Yet in police interrogations, for instance, compliant witnesses may, through their collaboration, produce the very evidence that is needed for an eventual accusation (Lynch and Bogen 1996). Institutional confessions or disclosures, therefore, often involve substantial downgrading of agency by the accountable person through impersonal constructions, vague or evasive responses, and minimizations of the offensive act.

In their in-depth analyses of the Iran–Contra hearings, Lynch and Bogen (1996), for instance, demonstrate how the defendant Oliver North cleverly exploited his response turns, changing the genre of the trial durings its six days on television. Through his military politeness, minimal responses, and strategic repetitions, he displayed an attitude of apparent compliance and respectfulness to the great media audience—the popularity of which was also reflected in an audience response of flowers, supportive letters, and successive transformation of an alleged villain of war into a public hero. Moreover, the two authors present fine-grained analyses of the conversational management of staged spontaneity and strategic 'forgetting' of crucial events. For instance, Oliver North acknowledged lying in the past, but claimed that he would now tell the truth 'the good, the bad, and the ugly' (Lynch and Bogen 1996, p. 38). On a deeper level, however, it is shown how a well-prepared witness like Oliver North skillfully avoided responding to the important questions, undermining the interrogator's narrative, thereby indirectly mocking his interrogators. The trial was thereby changed into public entertainment; what Foucault would have called a historical spectacle as opposed to a serious courtroom interrogation.

In order to understand a story fully, it is necessary to know something about by whom the story is told and to what audience. In work on narration, analyses of subversion are closely related to dramaturgical aspects of story telling: as in a focus on voices and multivoicedness (Bakhtin), or a focus on participation frameworks; that is, on speaker/listener roles (Goffman). Subversive action has been at the heart of Bakhtinian and Goffmanian work on negotiation, whereas it has been a less central concern in some applications of conversation analysis (that do not focus on participants' hidden agendas) or in some historical Foucauldian applications (that do not attempt to account for participants' agency).

Dominance and power are classical sociological themes, but in the work of narratives in interaction, collusion and subversion are equally important phenomena. In many ways, narratives can be seen as battlefields for self-ascriptions and self-projections even if some of the battles are indirect or hidden. In his classical work on asylums, Goffman (1961) discussed both the type of categorization work done by total institutions in their diagnoses of patients and the patients' resistance to stigmatization. In some recent work on youth sub-cultures, it has also been demonstrated how adolescent participants are acutely aware

of categorizational aspects of interview talk, and how they discursively resist such categorizations (Widdicombe and Wooffitt 1995). Following Bakhtin and Goffman, many students of narration have studied silent resistance or secret alliances and other types of collusion in social interaction. More recently, however, subversion—located in ordinary narrations found in mundane interactions—has been analyzed in more detail, and it has been brought out into the open.

See also: Conversation Analysis: Sociological; Conversation: Linguistic Aspects; Foucault, Michel (1926–84); Narratives and Accounts, in the Social and Behavioral Sciences; Postmodernism in Sociology; Postmodernism: Philosophical Aspects

Bibliography

Aronsson K, Cederborg A-C 1994 Voice and orchestration in family therapy: Conarration and voice in family therapy. Voicing, devoicing, and orchestration. *Text* **14**: 345–70

Aronsson K, Cederborg A-C 1997 A love story retold: Moral order and intergenerational negotiations. *Semiotica* **114**: 83–110

Bakhtin M 1984 *Problems of Dostoevsky's Poetics*. Manchester University Press, Manchester, UK

Buttny R 1993 *Social Accountability in Communication*. Sage, London

Capps L, Ochs E 1995 *Constructing Panic: The Discourse of Agoraphobia*. Harvard University Press, Cambridge, MA

Cicourel A 1968 *The Social Organization of Juvenile Justice*. John Wiley, New York

Conley J M, O'Barr W M 1990 *Rules versus Relationships: The Ethnography of Legal Discourse*. University of Chicago Press, Chicago

Drew P 1992 Contested evidence in courtroom cross-examination: The case of a trial for rape. In: Drew P, Heritage J (eds.) *Talk at Work: Interaction in Institutional Settings*. Cambridge University Press, Cambridge, UK

Edwards D 1997 *Discourse and Cognition*. Sage, London

Foucault M 1997 Technologies of the self. In: Rabinow P (ed.) *Michel Foucault, The Essential Works 1: Ethics, Subjectivity and Truth*. Penguin Press, Harmondsworth, UK

Goffman E 1961 *Asylums. Essays on the Social Situation of Mental Patients and Other Inmates*. Penguin Books, Harmondsworth, UK

Goffman E 1971 *Relations in Public. Microstudies of the Public Order*. Allen Lange, London

Goodwin M H 1990 *He-said-she-said: Talk as Social Organization among Black Children*. Indiana University Press, Bloomington, IN

Labov W A 1972 *Language in the Inner City*. University of Pennsylvania Press, Philadelphia, PA

Lynch M, Bogen D 1996 *The Spectacle of History: Speech, Text, and Memory in the Iran-contra Hearings*. Duke University Press, Durham

Ochs E, Capps L 1996 Narrating the self. *Annual Review of Anthropology* **25**: 19–43

Rorty P 1980 *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ

Sacks H 1992 *Lectures on Conversation: Vols. I and II*. Basil Blackwell, Oxford, UK

Schegloff E A 1997 Whose text? Whose context? *Discourse & Society* **8**: 165–87

Widdicombe S, Wooffitt R 1995 *The Language of Youth Subcultures: Social Identity in Action*. Harvester Wheatsheaf, Hemel Hempstead, UK

K. Aronsson

Narratives and Accounts, in the Social and Behavioral Sciences

Diaries, narratives, and accounts represent methods used in the social and behavioral sciences for investigating people's perceptions and subjective experiences (Josselson and Lieblich 1993). Diaries are written or recorded (audio or video) records of events, as well as related thoughts and feelings, often recorded on a daily basis. Diaries reflect text material in the form of stories (e.g., what happened to me today, or my interactions with my friend today) that are defined in terms of the ideas of accounts and narratives. Diaries often ask quite specific questions such as what did you do at 11 a.m. today? They may only ask a series of such questions or may involve questions for more lengthy involved responses. The latter are accounts and narratives and are the main focus of this discussion.

1. Definition and Scope

Accounts and narratives are terms referring to storylike compositions (that also may be written, or presented in another media format), that usually pertain to descriptions of and explanations for events observed or participated in by the individual. They also may include raw emotional expressions (e.g., 'Oh no, I can't believe it happened' as a reaction to hearing about a major loss occurring in someone's life). Some scholars (e.g., Schonbach 1990) have contended that accounts only pertain to justifications that people offer for consequential events in their lives (e.g., 'I had the accident because I was distracted by a large building I had not seen before on that highway'). Other scholars, however, have defined accounts more generally as the storylike responses to life events that may emphasize explanation and description, without special emphasis on justification. In the latter vein, accounts and narratives refer virtually to the same ideas and methodological approaches.

The following example is an excerpt from an account of a 63-year-old woman regarding how she coped with the loss of her husband and became

involved with another man who appreciated her continued reflections about her deceased husband (from Harvey 2000, p. 163):

> [After her husband's death] I had to help myself … I worked on myself. I privately reflected long and hard about our life together. But I joined support groups (e.g., Young Energetic Widowed Singles) and met the man I intend to marry in one of them. He had lost his wife in 1989, about the same time I lost my husband. We have been dating about 2 years. We have recovered from our grief and have a happy relationship. It's a real blessing. We talk often about our deceased spouses and how they would approve of our new relationship. We have shared our grief and stories, and are not afraid to shed tears … But we go on and appreciate each day in our new life. You have a right to have and express your memories of your former life. You have to talk, talk, talk—never stop.
> We are now sharing our experiences with other widows and widowers.
> We hope to help them in their grief and help them be successful in making their new lives …

Scholars also have used narratives and accounts as theoretical concepts. For example, Shotter (1984), Harvey et al. (1990), and Schonbach (1990) have produced theoretical conceptions that emphasize the role of narratives or accounts as mental and verbal constructions of meaning. These scholars argue that people develop these constructions in attempting to make sense of the world. In the above excerpt, the woman is reporting how she and her new lover have constructed meanings about the losses of their spouses and how they continue to develop and communicate those meanings. Central to their relationship and coping are their stories of loss, the regular acts of constructing those stories, and their sharing of stories with one another. There is a valuing of stories and confiding them to one another and to others.

Harvey et al. (1990) contend that people formulate many accounts about the matters of moment in their lives (e.g., their families, careers) and that these accounts become subaccounts of an overall life story master account. This master account, for instance, might contain stories of all of a person's close romantic relationships, that then form the overall picture in that person's mind of her or his romantic career.

Harvey et al. (1990) define people's act of formulating accounts as 'account-making' (similar to Bruner's 1990 concept of 'acts of meaning' in the construction of narratives). Harvey et al. (1990) argue and provide evidence supporting the position that account-making is particularly important to people in dealing with major loss events in their lives. This activity facilitates development of a sense of control, clarification, and expression. In the above excerpt, the couple emphasizes how their accounts about their deceased spouses have helped them to develop new identities and new lives that are positive for their health, but that do not discard the realities and impacts of their past relationships.

An influential theory of emotional expression articulated by Pennebaker (1990) also has stressed the value of confiding stories of loss as a way of coping with major loss. Pennebaker (1990) argues and provides much evidence that such expression relieves the behavior inhibition of emotions and that such inhibition is conducive to negative psychological and physical health developments.

The value of narratives and accounts is that they provide relatively in depth and naturalistic reports of human experience (Berscheid 1994). People may not feel as constrained in their reports when producing accounts and narratives as they do in responding to standardized measures or structured questionnaires (Bruner 1990).

2. *Typical Topics Examined Using These Methods*

Increasingly, these methods are used to examine many types of social and psychological phenomena. An example of the use of this technique is in the study of couples' retrospective reports of the breakdown of their close relationships (Weiss 1975). Couples may be asked to provide their versions of the events leading to the breakdown and sometimes are asked to try to report their partner's versions as well. These reports may be obtained by investigators from both members of a terminating relationship during the early separation period through final divorce or dissolution. The couple may be asked to write regularly over this period about their views of why their relationship is ending, new events occurring related to their feelings about the relationship, and future expectations for new close relationships (Bochner et al. 1997).

Another common topic for investigation using narratives and accounts pertains to people's experience of loss and trauma (Harvey 2000). For example, persons who have experienced the death of close others or personal losses due to illness or injury may be asked to report the events thoroughly and then their emotional reactions, as well as the reactions of close loved ones. An interesting question in such studies involves the extent to which close others were empathic and supportive of a person experiencing major loss. It has been found that people who receive a non-empathic, unresponsive reaction from close others to whom they tell their stories of loss will be negatively affected psychologically by those reactions, such that they may not again attempt to confide in close others about the losses (Harvey et al. 1990, Stroebe et al. 2000).

3. *Methodology Issues*

The methodology emphasizing narratives and accounts (which are classified as qualitative evidence)

often also includes structured questionnaires asking about thoughts, feelings, or behavior, as well as personality measures (with both of these latter types of evidence classified as quantitative in nature).

While refinements in such subjective report techniques are being regularly developed, their use dates back to the beginning of fields such as anthropology, psychology, and sociology (Howard 1991). Case study reports that are published in these fields historically have involved emphasis on subjective reports and narrative type evidence.

These subjective report methodologies have been well received for their depth of information and perspective on individuals' thoughts and feelings. They are also often favored by research participants who wish to 'tell their story' about some area of their life. On the other hand, such reports have been criticized because the evidence obtained may not be easily replicable by other investigators. Further, such evidence often requires coding, a step that introduces subjective decision making by independent raters evaluating the written, spoken, or video material.

The contemporary study of subjective reports increasingly is focusing on the process of production of such reports and the qualities of those reports. For example, there are theories of how people develop accounts of losses in their lives (Bochner et al. 1997). Reports that show a style emphasizing active decision making and positivity have been found to be associated with greater psychological and physical health in the aftermath of experiencing a major loss such as the death of a spouse (Bonanno and Kaltman 1999).

4. *Probable Future Directions*

As shown by an increasing number of users of these techniques, accounts and narratives can be used to study diverse human social phenomena that otherwise would not be easily amenable to scientific study (e.g., Baumeiser 1997). The value of these ideas also is embraced strongly by scholars stressing how people socially construct key elements of their lives, such as their marriages. As such, the future of accounts and narrative ideas and techniques is exceedingly bright. This future should involve greater integration across fields of work including psychology, sociology, communication studies, family studies, social work, and ethnography and cultural anthropology in general. Beyond this scholarly interface, many biographers, journalists, and investigative reporters use accounts and narratives. A future direction may involve bridges between this more popular writing with scientific writing emphasizing these ideas (Coles 1989). Methodologists also are increasingly showing the merit of different techniques of coding and thereby enhancing the case for generality of accounts and narrative evidence.

On the theoretical side of the accounts-narrative topic, researchers such as Nolen-Hoeksema et al. (1997) have pointed to the negative consequences of people's continued ruminating and whining about their losses and distress. Thus, a question for future work concerns the optimal degree of narrative work and confiding in others about such issues as major losses. It is likely that people may burden close others with repeated story-telling, unless their relationship is built on acceptance of that activity and it is reciprocal in nature.

Overall, both as methodological and theoretical ideas, narratives and accounts have become increasingly accepted during the last two decades of the twentieth century and have enriched the understanding of human behavior in the literature of the social and behavioral sciences.

See also: Case Study: Methods and Analysis; Case-oriented Research; Narrative Comprehension, Psychology of; Narrative, Sociology of

Bibliography

Baumeiser R F 1997 *Evil: Inside Human Cruelty and Violence*. W. H. Freeman, New York

Berscheid E 1994 Interpersonal relationships. *Annual Review of Psychology* **45**: 79–129

Bochner A P, Ellis C, Tillmann-Healy L M 1997 Relationships as stories. In: Duck S (ed.) *Handbook of Personal Relationships*, 2nd edn. Wiley, West Sussex, UK, pp. 307–324

Bonanno G A, Kaltman S 1999 Toward an integrated perspective on bereavement. *Psychological Bulletin* **125**: 760–76

Bruner J 1990 *Acts of Meaning*. Harvard University Press, Cambridge, MA

Coles R 1989 *The Call of Stories*. Houghton Mifflin, Boston

Harvey J H 2000 *Give Sorrow Words: Perspectives on Loss and Trauma*. Brunner/Mazel, Philadelphia

Harvey J H, Weber A L, Orbuch T L 1990 *Interpersonal Accounts: A Social Psychological Perspective*. B. Blackwell, Oxford

Howard G S 1991 A narrative approach to thinking, cross-cultural psychology, and psychotherapy. *American Psychologist* **46**: 187–97

Josselson R, Lieblich A (eds.) 1993 *The Narrative Study of Lives*. Sage, Newbury Park, CA

Nolen-Hoeksema S, McBride A, Larson J 1997 Rumination and psychological distress among bereaved partners. *Journal of Personality and Social Psychology* **72**: 855–62

Pennebaker J 1990 *Opening Up*. Morrow, New York

Schonbach P 1990 *Account Episodes: The Management or Escalation of Conflict*. Cambridge University Press, Cambridge, UK

Shotter J 1984 *Social Accountability and Selfhood*. B. Blackwell, Oxford

Stroebe M S, Hansson R O, Stroebe W, Schut H (eds.) 2000 *Handbook of Bereavement*. Cambridge University Press, New York
Weiss R S 1975 *Marital Separation*. Basic Books, New York

J. H. Harvey

Nation: Sociological Aspects

1. *Nation: A Late Political Meaning*

Etymologically speaking, a Nation, from the latin *natio*, designates all those linked by birth. In modern language, the word which best expresses that original meaning of Nation is the English word 'native.' In ancient Rome, the word had no particular political meaning. It seemed even to characterize human groups devoided of political structures. During the Middle Ages, the word Nation obtained a linguistic meaning. This explains why students at the first great European universities were regrouped in Nations according to the language they spoke and sometimes even clashed along those lines!

At the Renaissance, during the process of State building, the Nation will again endorse a more political bent until the French Revolution: for a long time, the Nation will be identified with those benefiting from a political status because they were close to the crown (the power) or were represented in those predemocratic parliaments, from which the rest of the population was excluded (Schulze 1996)

Only during modern times will the Nation take its present and more popular meaning. The concept will even absorb almost all the different denominations characterizing political societies. Just think about the 'League of Nations,' or the 'United Nations Organization': this will testify equally to the might of the word and to its largely undetermined character, quasi-assimilating the concept of Nation with any kind of statist political organization.

2. *From Nation to the Sociology of Nation*

The Nation as we conceive it today is a product of the nineteenth century, which makes it the main support of the constitution and development of modern States by using the famous 'principle of nationalities.' However, sociology will take her own good time to discover the Nation.

The founding fathers of sociology who carved their imprint in the nineteenth century, whether Karl Marx, Max Weber, or Emile Durkheim, were far from giving a central place to the Nation in their analysis of modern society, contrarily to historians or philosophers, or even artists, who started discussions, as witnessed by diverging opinions from Ernest Renan, de Fustel de Coulanges, Fichte, or Herder. Sociology is based on 'society' whose meaning remains largely undefined. As her object research, she deals with 'the State,' the 'People,' but not the 'Nation.' However, these notions are often described under characteristics of a modern Nation—as without doubt by Weber in his analysis of society—but never does the concept of Nation emerge.

Marx and Engels deal a little more with the notion of Nation, but they consider it more as an archaic form than as a new reality pregnant with history. The next generation of Marxist authors will be more concerned by the notion, but for practical more than theoretical reasons, given the attractiveness of nationalistic slogans for the working class. We owe to Bauer, and to the 'Austro-Marxism,' a more sophisticated conceptualization (Bauer 1907). Staline himself will try to define exactly what he means by Nation in a work where he uses its nonpolitical aspects as a basis—a restrictive one—for the bolshevik policy of recognition of nationalities (Staline 1937).

As pointed out by Hobsbawm, the multiplication of works on the Nation dates back mainly to the period 1968–1988 where more has been written on the subject 'than at any other period, even twice as long' (Hobsbawm 1990). The concept of Nation gets then a truly sociological meaning, though its definition has opened new conflicts that are still hotly alive today.

However, in modern times, Nation will be unanimously recognized as 'the' political community that will ensure the legitimacy of the State over its territory, and moreover transform the State into the State of all its citizens. The notion of 'nation-state' illustrates that new alliance between State and Nation. Nationality will bind the citizen to the State, a bond that will be more and more tied to the advantages of a social policy in as much as the Welfare State will develop.

Nobody will deny either that the Nation finds its ideology in nationalism, where Nation and State necessarily are intertwined into a new entity dominating all others. Nationalism will not only mark the building of Nations, but also provoke the reaction of those populations who feel suffocated by this process and want to build their own nation-state. The violence of nationalism will depend on the surrounding circumstances and evolve according to the obstacles found on its way.

Where true nation-states will emerge, the Nation will be the point of convergence of three historical processes. The first one is the development of States with growing sovereignty over territories whose boundaries are more and more clearly defined, and with growing autonomy vis-à-vis the nonterritorial power of the Church, whose prestige and influence declined after the Reformation; this process being faster in Northern Europe than in the South. The next step is the emergence of democracy in the aftermath of the French revolution, transfering sovereignty from

the State to the 'People' and regrouping all citizens under principles of equality and solidarity. Finally, the development of capitalism increases communications between individuals, social homogeneity, educational and cultural standardization, and by its very crises leads to the strengthening of States' power and the development of the Welfare State endorsing an ever increasing part of the life of each citizen.

The differences of evaluation among authors focus on the importance to be given to the ties linking the modern Nation to the former 'protonational' entities. These differences concern the process of construction of those Nations and the role of history in that process. For sociologists like Gellner, Hobsbawm, and Anderson, modern Nation is first a rupture, even sometimes a complete one, with the past. This 'modernist' definition is at loggerheads with the definitions, called 'primordialist' by Gellner, of which the best known representative is Smith, a former pupil of Gellner.

3. *Two Concepts of Nation?*

For the modernists, the rise of the Nation can only be understood as the product of State action, and more particularly, according to the famous hypothesis of Gellner, as the result of the nationalist action of its elites (Gellner 1983). Nationalism gives birth to Nation and not the opposite. This voluntarist action is made possible, and even necessary, by modernization, that is, by the transition from agrarian society to industrial society and capitalism. Modernization destroys old social hierarchies and loyalties therein associated, and entails homogeneization of language and culture throughout the State territory in order to allow both for mobility of workers and for mass production. Whereas agrarian societies possessed stable hierarchies where culture was linked to social status, modern society, by destroying these hierarchies, transform culture into a kind of personal appropriation of each individual and creates a longing for harmony between individual culture and the environment: 'What automatically makes people into nationalism' (Gellner 1999).

In that sense, Nation is a truly modern invention which takes the form of an 'imaginary community' (Anderson 1983). Indeed, nation-states have a past, at least most of them, but modernization is a phenomenon modifying so radically social practices that this past is practically irrelevant for the construction of modern Nations. Or, at best, it is so remolded that it looses any true historical characteristic. One example, pointed out by Gellner, of a Nation created *quasi ab nihilo* is the case of Estonia. Whereas that Nation has no past to speak of, its elites managed to give it life, to develop a distinctive culture and to impose it among the other Nations. In this manner, Estonia is an ideal type of a Nation in the modern acceptation of the word.

At the opposite end, for the 'primordialists,' the development of the nation-state, and therefore of the Nation, is only understandable by analyzing its historical roots. Smith calls this approach 'ethno-symbolist.' One cannot but study the role of myths and prenational symbols in order to distinguish nationalisms from one another, and above all to explain why nationalism will be virulent in one case and not in the other (Smith 1986). The existence of competing nationalisms in Greece cannot be understood if one looses sight of 'its dual heritage of byzantine imperial orthodoxy and classical democratic antiquity.' 'Modernism tells half the story,' writes Smith. But Gellner answers that 'if it tells half the story, that is enough for me!' (Smith 1999, Gellner 1999).

In this perspective, Hroch offers an authoritative developmentalist typology of national movements. These movements evolve in three phases. First is that of purely cultural, linguistic, or folklore-related aspirations. Second, that of the rise of militant groups promoting the 'national idea' and starting social trouble around the national theme. Third, one that finds a spreading support among the mass for the new national aspirations. The transition from the second to the third phase will often be marked by a 'nationally relevant conflict of interest' (Hroch 1985).

What are the factors prior to the Nation that can legitimize its rise? The most frequently mentioned ones are language and religion. Language will be particularly important as a factor when it has been threatened or has been a basic motor of identification. This is the case, for instance, in Belgium, where Flemish nationalism concentrates in a language which has been the object of assimilation by a dominant language (Delfosse 1994). However, language is not necessarily the essential legitimizing source of nationalism and Nation. The French Revolution does not establish a nation-state on a linguistic basis: French language is deemed indispensable to proper revolutionary ideals, but not considered an exclusive and sacred good. It is a kind of tool of citizenship, and, through it, of the Nation. Completely different is the importance of language in the creation of the German or Italian nation-state. Before its founding, language was the only unifying factor of scattered populations and therefore took a more affective meaning than in the French case.

Another factor of importance is religion. The Reformation played a key role in the rise of modern nation-states by emancipating them from the power of the Church, on the spiritual level as well as on the territorial level. However, for a sociologist like Rokkan, religious criteria seem of less importance than language. For him, language has a more defined territorial characteristic than religion and the notion of Nation is linked strongly with a territorial foundation. That last factor is the decisive one: there is no Nation without territorial identity. Therefore one must call upon intricate processes, ethnicocultural,

religious, linguistic, and politico-administrative ones that can go back a long way (Flora 1999).

Other determinants could be named here, but it seems almost impossible to find a unique definition of the objective elements that would characterize the Nation. The primordialist school would tend to offer a specific approach of the creation of each particular Nation. Indeed, Nations have been built upon criteria much too different to allow for the constitution of a 'hard core,' even if modernization in each case has added its own particular homogenizing effect.

This controversy between modernists and primordialists, that is, between those for whom nationalism created the Nation and those who look for links between nationalisms and pre-existing national characteristics to explain their different development and intensity, that controversy in sociological literature is far from ending. For Hroch, these two approaches are not necessarily incompatible for there is interaction and not dominance between Nation and nationalism (Hroch 1999). In this perspective, we can borrow Disraeli's lapidary expression that applies so well to national construction: 'of art and time.'

4. *The Future of the Nation*

Since the nineteenth century, the world has become one of Nations whose number is ever increasing in the framework of the United Nations. Some of them have been the result of Europe's export of her own model of nation-state through colonialism. That model, in spite of it artificial character, has been largely 'instrumentalized' by movements of national liberation in order to consolidate their State authority (Badie 1992). National demands inside even the best established nation-states, directed against their domination (the so-called 'internal colonialism'), never stopped and the 'desire for a State' sometimes led to ruptures, sometimes to reorganizations of States on the regionalist or federalist models.

Today, however, nobody can deny that changes are at work on a world scale that put into jeopardy the future of the so-called modern Nation. 'Mondialization' and most of all 'globalization' send a major challenge to most of the nation-states. Whereas nation-states drew most of their power from their quasi-absolute control over their territory, and particularly over their national economy, globalization deprives them increasingly of their real competences as well as their resources, while severely affecting their legitimacy, even though this globalization is on a 'shifting scale' according to different domains. At the same time, that evolution has been facilitated largely by the collapse of the Soviet Empire, which however gave free rein to an unexpected 'pent-up nationalism.'

An additional phenomenon of 'cultural melting pot,' together with the prevalence of English as an international language and the mixing of populations and cultures, be it on the virtual level with Internet or in reality with population migrations and immigration to rich countries, the whole constitutes an evolution whose effects are particularly difficult to integrate in homogeneous cultures typical of classical nation-states.

Habermas states the problem in terms of a relatively simple alternative, at least on the politicoeconomic level: either the most threatened nation-states will withdraw in neoprotectionist tentatives bound to fail, or they will try to reach 'the same level as the markets,' which entails their entering some sort of supranational integration. The world of culturally homogeneous nation-states will fade away leaving the stage for new mixed entities, particularly on the level of cultural homogeneity (Habermas 2000).

The big question whether the process of globalization will dissolve nation-states into new and necessarily pluralistic entities, or will give rise to renewed nationalistic demands challenging it, or, yet, whether the two phenomena will go hand in hand, remains largely open, even if what can be observed goes in the sense of the third alternative. In this sense, the European Union represents the most advanced form of supranational integration in the world. Since Europe is also the cradle of modern Nations, an analysis of the processes taking place there is of the utmost interest for the future of the Nation.

Comparative sociological studies show that nationalism in European nation-states has lost some of its acuity which favors European construction. However, reactions of identity withdrawals are manifest from time to time, illustrating thus the double nature of the process at work.

'Eurobarometers' surveys show that 'national pride'—a dimension of nationalism—clearly has decreased in intensity since the 1980s. Those who are very proud of their Nation shift from 41 percent in 1982 to 28 percent in 1994, and this trend is present in all the countries, with the exception of Ireland and Greece. Moreover, it is clearly an effect of era and not of generation (Belot and Tournier 1998).

This does not prevent reactions of national reassertion reappearing at the same time in European countries, which shows that evolutions are not linear. Witness the sometimes important scores—though always in minority—of those 'populist nationalist' parties. The highest score reached so far in national elections is that of the Austrian FPÖ with 26.9 percent of votes (at the last legislative elections in autumn 1999). However, it should be pointed out that these reactions concern less European integration *per se* as the waves of non-European immigration which have marked most of the European States. Apart from the European Union and from an evolution of nationalism perhaps growing dull with time and with the painful experience of past struggles, the thesis of the irresistible vitality of Nation feeds on the renewed violent outbreaks of center-European nationalisms growing

on the rubbles of Soviet domination. However, some authors, like Hobsbawm, question the 'naturalness' of this process. For him, it is more the collapse of the USSR than the strength of nationalist movements that made this evolution possible. Alone they would never have been able to impose themselves without first finding the way wide open (Hobsbawm 1990).

Another trend, also stemming from the globalization of problems, is a potential threat to the nation-states: that of the multiplication of effective actors on the international scene.

Beside Nations, some of them lacking in consistency and some others emerging, and in addition to classical international organizations and forms of supranational integration at various levels of development, we find entities as diverse as multinational corporations, non governmental organizations of all types, terrorist groups, new organizations based on language, religion, sport activities, or other, and even entire civilizations if we are to follow the hotly contested predictions of Huntington on the future clashes on a world scale (Huntington 1993).

We find ourselves thus in a world which is not anymore that of a quasi-exlusively 'concert of Nations,' but in an environment where diverse entities emerge, making it much more complicated. That world could be in transition towards new dominant structures; it can also wallow in the trough of a relatively chaotic situation, which, for some like Bull, could be identified with a 'new Middle Age,' characterized by multiple loyalties towards entities more or less defined by territories (Bull 1995).

Whether this strange 'back to the future' takes place or not, one thing is sure: the future of the Nation is uncertain.

See also: Citizenship, Historical Development of; Citizenship: Political; Community Sociology; Democracy; Democratic Theory; Globalization: Political Aspects; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Identity: Social; National Character; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Nations and Nation-states in History; Religion: Nationalism and Identity; State and Society; State Formation; State, History of; State, Sociology of the

Bibliography

Anderson B R O'G 1983 *Imagined Communities*. Verso, London

Badie B 1992 *Etat importé* [*The Imported State*]. Fayard, Paris

Bauer O 1907 *Nationalitätenfrage und die Sozialdemocratie* [*National Question and Social Democracy*]. Verlag der Wiener Volksbuchhandlung, Vienna

Belot C, Tournier V 1998 Les jeunes, l'Europe et la nation. In: Bréchon P, Cautrès B (eds.) *Les enquêtes Eurobaromètres. Analyse comparée des données socio-politiques* [*The youngs, Europe and the nation*. In: *The Eurobarometer's Surveys. A Comparative Analysis of Socio-political Data*]. L'Harmattan, Paris, pp. 115–31

Billig M 1995 *Banal Nationalism*. Sage, London

Breuilly J 1992 *Nationalism and the State*. Manchester University Press, Manchester, UK

Bull H 1995 *The Anarchical Society. À Study of Order in World Politics*. Macmillan, London

Delannoi G 1999 *Sociologie de la nation. Fondements théoriques et expériences historiques* [Sociology of Nation. Theory and Historical Experiences]. Colin, Paris

Delfosse P 1994 Nationalisme et parlementarisme. *Res Publica* **36**(2): 179–219

Deutsch K 1953 *Nationalism and Social Communication*. MIT Press, Cambridge, MA

Flora P (ed.) 1999 *State Formation, Nation-Building, and Mass Politics in Europe: The Theory of Stein Rokkan*. Oxford University Press, Oxford, UK, pp. 64–7

Gellner E 1983 *Nations and Nationalism*. Oxford University Press, Oxford, UK

Gellner E 1999 Adam's navel: 'Primordialists' versus 'Modernists.' In: Mortimer E (ed.) *People, Nation and State. The Meaning of Ethnicity and Nationalism*. Tauris, London, pp. 31–6

Habermas J 2000 *Après l'Etat-nation. Une nouvelle constellation politique* [After the Nation State. A New Political Constellation]. Fayard, Paris

Hall J A (ed.) 1998 *The State of the Nation. Ernest Gellner and the Theory of Nationalism*. Cambridge University Press, Cambridge, UK

Hoffmann S 1995 Obstinate or obsolete? The fate of the Nation State and the Case of Western Europe. *Daedalus* **95**: 867

Hobsbawm E J 1990 *Nations and Nationalism since 1780: Programme, Myth, Reality*. Cambridge University Press, Cambridge, UK

Hobsbawm E J, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Hroch M 1985 *Social Preconditions of National Revival in Europe*. Cambridge University Press, Cambridge, UK

Hroch M 1999 Real and constructed: The nature of the nation. In: Mortimer E (ed.) *People, Nation and State. The Meaning of Ethnicity and Nationalism*. Tauris, London, pp. 91–107

Huntington S P 1993 The clash of civilizations. *Foreign Affairs* **72**: 22–49

Inglehart R 1977 *The Silent Revolution: Changing Values and Political Styles among Western Publics*. Princeton University Press, Princeton, NJ, p. 337

Mortimer E (ed.) 1999 *People, Nation and State. The Meaning of Ethnicity and Nationalism*. Tauris, London

Renan E 1882/1990 What is a nation? In: Bhabha H K (ed.) *Nation and Narration*. Routledge, London

Schulze H 1996 *Etat et Nation dans l'Histoire de l'Europe* [State and Nation in the History of Europe]. Seuil, Paris

Smith A D 1971 *Theories of Nationalism*. Duckworth, London

Smith A D 1986 *The Ethnic Origin of Nations*. Blackwell, Oxford, UK

Smith A D 1999 The nation: Real or imagined. In: Mortimer E (ed.) *People, Nation and State. The Meaning of Ethnicity and Nationalism*. Tauris, London, pp. 36–43

Staline I 1937 *Le marxisme et la question nationale et coloniale* [Marxism and the National and Colonial Question]. Sociales, Paris

Tilly C (ed.) 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ

A.-P. Frognier

National Character

National character is an expression which describes forms of collective self-perception, sensibility, and conduct which are shared by the individuals who inhabit a modern nation-state. It presupposes the existence of psychological and cultural homogeneity among the citizens of each country, as well as the idea that each nation can be considered a collective individual, with characteristics analogous to the empirical individuals who are its inhabitants. The noun *character* seeks to describe a universal aspect of social life—simultaneously an internal dimension to the existence of individuals and an external one, observable through collective behavior. The adjective *national* situates this universal aspect of social life in the specific context of those social units we call nations.

Since it was first formulated in Europe during the second half of the eighteenth century, the idea that the inhabitants of each nation possess psychological and cultural characteristics in common, which confer upon them identity and distinguish them from others, has always been a part of the ideology and practice of nationalism. Hence its double status, being simultaneously a descriptive and a prescriptive expression. On the one hand it describes communities of custom, forms of being and feeling, on the other it serves to make legitimate claims for sovereignty and forms of political action. One refers to the essence of each people or nation, the other supposes that collective action can serve to favor or even to mold the national character of a political community in the making.

During the 1940s and 1950s, the expression national character gained scientific legitimacy within the so-called North American anthropological school of Culture and Personality, to the point of designating an object of study and a university specialization: national character studies. From then on, it has been at the center of various academic controversies, where many have questioned its theoretical status, attributing its scientific shortcomings to its close relationship with the essentialism characteristic of nationalist ideologies. Yet recently, social theory studies interested in understanding the social force of feelings of national belonging and the new forms of nationalism have turned once again to the expression. What distinguishes this trend from national character studies is the fact that there is no attempt to attribute any theoretical status to the expression; instead, current studies are more concerned with the expression national character as an object, a practical category used in the discourse and action of the social agents and groups which one seeks to understand (for an argument in favor of the study of the categories of national identification as 'practical categories' see Brubaker 1996).

1. *National Character and Nation-states*

Questions dealing with national self and unself have been historically associated with the origin of nation-states and with international relations. While the processes leading to the emergence of new nation-states served to substitute older national elites for new ones, redefining the borders between countries along the way, the belief in the singularity of *one* national character, also demanded the recognition of a plurality of national characters from which it became necessary to differentiate.

One of the most detailed examples of the relationship between we-ideals and we-images in the formation of nation-states can be read in Norbert Elias' studies of the Germans (see Elias 1994, 1996). Discussions concerning 'national character' in Germany emerged within the context of an encounter between new intellectual figures and a public capable of understanding and identifying with them. The poetry and philosophy of the second half of the eighteenth century (authors such as Herder, Schiller, or Goethe and movements such as *Sturm und Drang*) celebrated the ethic and aesthetic community grounded in the concept of *Kultur*, whose principal, distinguishing element was the use of the German language in opposition to the French spoken in the German courts by the social elites of the time. One century later, when the heirs to this intellectual tradition, backed by sections of the military and the old imperial nobility, established themselves within the German nation-state, words such as *Kultur* (alongside words such as *Volksgeist* and *Zeitgeist*) had already become 'nationalized,' thus designating the we-ideals and the we-images of social groups in a larger scale and describing forms of collective self-perception and the 'German National Character.'

The diffusion of the nation-state as a means for organizing social life (see, for example, Balibar 1991 [1988] and Anderson 1983) brought about similar processes in other areas of the world. At the start of the twentieth century, and particularly after World War I, the end of the old Empires served to reinvigorate nationalist ideologies in central and eastern Europe. Ideas about the national character of the Hungarians, Poles, Czechs, Romanians, Croats, and Serbs began to inform the unification policies of new states. By then, however, debates about 'national character' were no longer confined to European

intellectual circles. In America too, the most renowned authors were those who succeeded in describing the traits of the 'spiritual community' to their compatriots, who then recognized themselves in these. Their texts and ideas served as a guide for state politics, spreading beyond the literate elite and onto other social groups in the process of nation-making (particularly through the expansions of literacy and the school system). In Spanish-speaking America one of the best examples may be the Mexican writer Manuel Gamio, author of well-known statements on what it means to be Mexican, centered on the relations between the descendants of pre-Hispanic peoples and the Spanish (Gamio 1916). In Portuguese-speaking America the best example is Gilberto Freyre, remembered as the author of one of the most influential representations of Brazilian national character, based on the idea of miscegenation between Whites, Blacks, and Amerindians (see Freyre 1933). Among the many similarities between the two authors is the fact that during the 1920s both resided in the USA and studied under the German anthropologist Franz Boas at Columbia University. Boas is largely recognized as one of the intellectual fathers of the Culture and Personality school which later served as an inspiration behind national character studies (on the Culture and Personality School see the articles collected in Stocking 1986).

2. *National Character Studies: A War-time Anthropology*

The Culture and Personality School was marked by two main ideas. The first, based on the concept of primary socialization, sought to explain the relationship between types of culture and types of personality based on the experiences of childhood (see Mead 1928). The second, centered around the idea of patterns of culture, intended to describe and psychologically profile cultural configurations and the types of personality associated with them: each culture, although made up of fragments from different origins, developed a specific 'style,' capable of bringing about its unification as a synthetic totality and, at the same time, distinguishing it from other cultural units (see Benedict 1934). Even though the relationship between the concept of patterns of culture and the theories concerning German national character were very explicit (in her books, for example, Benedict mentions Dilthey and Spengler), the use of the concept was primarily restricted to primitive or simple (i.e., non-national) social forms, the traditional object of anthropological inquiry.

The use of these ideas in attempts to understand cultural configurations and personalities within national communities was, in fact, closely tied to the experience of the extreme and intense form of international relations which is war. For it was immediately after North-American intervention in World War II that a group of anthropologists (foremost among them were Gregory Bateson, Ruth Benedict, Geoffrey Gorer, Clyde Kluckhon, Ralph Linton, Margaret Mead, Rhoda Métraux, and George P. Murdock) became involved with the Office of War Information, intending to place the theories and methods of social anthropology at the service of the war effort. Its first goal became to produce a statement about national character which would help maintain the high morale of Americans during the conflict. Soon afterwards, its activities turned towards attempts at understanding allied and, specially, enemy nations.

The end of the war led to what would later be considered the most convincing evidence of the efficacy of national character studies. After Japan's surrender, the American government decided to maintain the Japanese emperor in power, as had been recommended by the Office of War Information in a memorandum signed by Gorer. This helped to ensure stability in postwar Japan and to guarantee that, despite Hiroshima and Nagasaki, a new state of equilibrium between the two countries could be established. This concurrence between US foreign policy and the anthropological analysis of the place of the Emperor in Japanese culture has been cited ever since as a fine example to the kind of 'applied anthropology' which national character studies allows.

But it was in the early 1930s that anthropological solutions to 'practical' problems began to be elaborated in the USA. Up until World War II, applied anthropology had been a matter for internal affairs, aimed at solving problems derived from American cultural and social diversity. In shifting the field of applied anthropology to external affairs, national character studies crossed a significant threshold in the history of anthropology. According to the division of labor established between the disciplines at the time, anthropology had been confined largely to the study of social forms different from those of the observer, and thus different from national societies and cultures. Now, for the first time, anthropologists began to deal explicitly with nation-states as an object of study. The discussions which then followed went beyond the concerns of anthropologists alone, and began to interest all those who sought to comprehend national forms of social and cultural life (for an analysis of the relations between theory and politics in national character studies, see Neiburg and Goldman 1998).

First the USA and the Americans, then Japan and the Japanese, Germany and the Germans, Great Britain and the British; a little later (already in the context of the Cold War) Russia and the Russians, Poland and the Poles (see, for example, Gorer and Rickman 1950)—all came to be treated as cultural worlds susceptible to analysis on the basis of the same categories used in the study of so-called 'simple societies,' those to which, until then, the majority of anthropologists had restricted their attention. Fron-

tiers between countries were treated as cultural, and cultural frontiers were treated as national, thereby delimiting new objects for analysis. The context of war and of conflicts between nations demanded, however, a methodological 'development': 'studies of culture at a distance' (see Mead and Métraux 1953, Mead 1953). Anthropologists, prevented from carrying out the traditional task of prolonged participant observation in the field, would now interview immigrants, analyze novels or films made in other countries, or examine media productions, so as to reach those cultural universes which were always qualified as *national*.

So began a tradition of research based on empirical sources which can still be recognized today, and not only in anthropology but also in other disciplinary traditions interested in the study of national cultures or societies (notably the so-called cultural studies). Yet it is interesting to draw attention to the paradoxical nature of the idea of 'studies of culture at a distance' for, in studying nation-states, researchers were socially and culturally 'closer' than traditional anthropologists in relation to their objects. Thus the fact that traditional methods and practices of anthropology (such as participant observation) were rarely made use of seems to have more to do with the specific properties of the object of study (or with the nature of the relations between researcher and object) than with the supposed impossibility of direct access to it.

3. *National Character Reframed*

In the theoretical debates concerning the expression 'national character' we can recognize some of the main questions surrounding contemporary discussions on the status of the nation-state as an object of the social sciences. One finds arguments in favor of the need to renew theoretical models in order to conceptualize empirical objects which need to be thought of as a historical novelty—such as those connected with processes of decolonization, localization, and transnationalization of social and cultural life—or denunciating the supposed limitations behind theoretical models whose origin would be a mere reflection of the world of nation-states which these models were developed to explain in the first place.

The main objections to national character studies can be summed up in three points. First, cultural patterns (obtained through the direct observation of empirical behavior) may become the principles through which one then seeks to explain these very patterns—a vicious circle which appears to be characteristic of culturalism in general (for a critique, see Kuper 1999). Second, observed behavior in a fraction of the society analyzed would then be abstracted and attributed to all nationals, without a critical enquiry into the social position occupied by that fraction within national society. As such, the idea of national character would serve to disguise a more profound sociological reality; it would become an 'ideology' in the Marxist sense. Finally, in its rush to discuss racist theories, the Culture and Personality school (and, by extension, national character studies) may have shifted from a biological essentialism to a cultural or psychological one, which may then lead it to neglect other dimensions of social life, fundamental to an understanding of national realities and the elaboration of feelings of national belonging—such as institutional aspects related to the presence of the state, or the relations of differentiation, force, and domination between social groups (see Leite 1969 [1954]).

In an article first published in 1942, Gregory Bateson outlined what remains the most convincing argument in favor of studies of national character. First, he argued that the fact that national characters are constructs does not mean that they do not exist and therefore do not serve to emphasize cultural and social differences. Second, he affirmed that to speak of national character does not imply cultural uniformity, but can, on the contrary, stress the existence of a system of differentiations organized in relation to a pattern (a set of stereotypes) recognized as being dominant. In this sense, national borders (political, cultural, and 'character-based') would serve as a good example of how something which is socially generated can lead to certain effects, acting on the forms of self-perception of the very communities which they differentiate and constitute, ultimately affecting ways of feeling, living styles, and moral configurations (for a similar argument see M. Mauss' 1969 [1921]) classical text; good examples of anthropological studies heading in this direction include Handler (1988) on the objectification of national cultures and Herzfeld (1993) on the social productivity of national stereotypes).

Feelings of national belonging and national forms of organizing social life are, like the Roman god Janus, two-faced. One face is that of the solidarity between nationals, between those who stand united through a relatively homogenous social and cultural base. The other reveals itself at the national borders and the attempts at imposing them and re-enforcing them, through exclusion and nationalist violence. The sociological and anthropological understanding of both faces—as well as of the social force of feelings of national belonging, which are, in effect, the basis for their existence—constitute one of the central themes in contemporary social science. A careful examination of the diverse contents of the discussions brought about by the idea of national character shows us how an advance on this understanding must pass through three theoretical frameworks: we must situate feelings of national belonging alongside other modes of social belonging and identification (ethnic, racial, regional, transnational) as a means of discovering what they may owe to (and reveal about) the universal properties and forms of social life; we cannot ignore the historical and sociogenetical processes behind the formation of

categories of identity and the homogenization of forms of being and feeling; and, finally, it is essential to situate the comprehension of national characters as a dimension of the processes which objectify realities (cultures, societies, identities, stereotypes, etc.), both national and international.

See also: Benedict, Ruth (1887–1948); Cultural History; Cultural Variations in Interpersonal Relationships; Culture and Emotion; Education: Anthropological Aspects; Elias, Norbert (1897–1990); Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender Ideology: Cross-cultural Aspects; Mead, Margaret (1901–78); Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Sociology of; Psychological Anthropology; Religion: Nationalism and Identity; Sexual Orientation: Biological Influences; Values, Anthropology of

Bibliography

Anderson B 1983 *Imagined Communities*. Verso, London

Banac I, Verdery K (eds.) 1995 *National Character and National Ideology in Interwar Eastern Europe*. Yale Center for International and Area Studies, New Haven, CT

Balibar E 1991 [1988] The nation form: history and ideology. In: Balibar E, Wallerstein I (eds.) *Race, Nation, Class*. Verso, London

Bateson G 1978 [1972] Morale and national character. In: *Steps to an Ecology of Mind*. Granada, London

Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston

Benedict R 1946 *The Chrysanthemum and the Sword*. New York and Houghton Mifflin, Boston

Brubaker R 1996 *Nationalism Reframed. Nationhood and the National Question in the New Europe*. Cambridge University Press, Cambridge, UK

Elias N 1994 (1978) *The Civilizing Process: The History of Manners and State Formation and Civilization*. Blackwell, Oxford, UK

Elias N 1996 *The Germans. Power Struggles and the Development of Habitus in Nineteenth and Twentieth Centuries*. Columbia University Press, New York

Freyre G 1933 *Casa-Grande & Senzala (Fromação da Familía Brasileira sob o Regime de Economia Patriarcal)*. Maia & Schmidt Ltda. Rio de Janeiro (published in English 1946 *The Masters and the Slaves (A Study in the Development of Brazilian Civilization)*. Translated by Samuel Putnam, Afred A. Knopf, New York)

Gamio M 1916 *Forjando Patria: Pro nacionalismo*. Porrua Hermanos, México

Gorer G 1943 Themes in Japanese culture. *Transactions of the New York Academy of Sciences* **2**(5): 106–124

Gorer G 1948 *The American People*, 1st edn. Norton, New York

Gorer G 1955 *Exploring English Character*. Cresset Press, London

Gorer G, Rickman J 1950 *The People of Great Russia*. W. W. Norton, New York

Handler R 1988 *Nationalism and the Politics of Culture in Quebec*. University of Wisconsin Press, Madison, WI

Herzfeld M 1993 *The Social Production of Indifference*. University of Chicago Press, Chicago

Kuper A 1999 *Culture—The Anthropologists' Account*. Harvard University Press, Cambridge, MA

Leite D M 1969 [1954] *O Caráter Nacional Brasileiro. História de uma Ideologia*, 2nd edn. Livraria Pioneira Editora, São Paulo, Brazil

Mauss M 1969 [1921] La Nation. In *Œuvres III*. Minuit Paris, pp. 571–625

Mead M 1928 *Coming of Age in Samoa*. Penguin, Harmondsworth, UK

Mead M 1965 [1942] *And Keep Your Powder Dry. An Anthropologist Looks at America*. Morrow, New York

Mead M 1953 National character. In: Tax S (ed.) Appraisal *Anthropology Today*. University of Chicago Press, Chicago

Mead M, Métraux R (eds.) 1953 *The Study of Culture at a Distance*. University of Chicago Press, Chicago

Neiburg F, Goldman M 1998 Anthropology and politics in studies of national character. *Cultural Anthropology* **13**: 1

Smith A D 1986 *The Ethnic Origins of Nations*. Blackwell, Oxford, UK

Stocking Jr G W (ed.) 1986 *Malinowski, Rivers, Benedict and Others. Essays on Culture and Personality*. University of Wisconsin Press, Madison, WI

F. Neiburg

National Health Care and Insurance Systems

All national governments have some policies regarding healthcare for their citizens. National Health Care/Insurance Systems (NHC/I) are distinguished by two characteristics. First, the government sponsors or coordinates a system of healthcare service delivery or insurance. Second, this system aspires to guarantee the benefits of modern standards of medical care to all, or virtually all, citizens, and comes reasonably close to achieving that goal.

Government public health measures such as sanitation are not NHC/I because they do not focus on individual medical services. National systems to provide subsets of care, or care to subsets of the population, such as primary care clinics for the poor, are not NHC/I because they do not provide the full range of modern medical care to virtually all citizens. In any given instance, also, there may be disagreements about whether the aspiration is achieved: for instance, whether 'waiting lists' mean that adequate care is not guaranteed. Certainly no system meets all possible demands. Each nation also will have its own definition of the necessary extent of medical benefits. Nevertheless, the definition here fits common and scholarly understandings.

All advanced industrial countries, save the USA, have NHC/I systems. A number of countries that are nearing that status, such as South Korea and Taiwan, have created systems that may be said to meet the standard. However, in many countries that have aspired to meet the standard, such as much of Latin America and the former Soviet bloc, healthcare systems fall sufficiently short that most observers would say the aspiration is not reality.

NHC/I systems are of interest to the social sciences because they are large portions of modern economies, with direct impact on every person's life.

Care may be guaranteed, as in Germany, by providing insurance. The government will regulate or provide incentives to create healthcare providers, but the national system mainly reimburses individuals or pays directly for the cost of care provided from outside the system.

Alternately, as in the UK, the government may operate a care delivery system and guarantees access, with low or no fees, to that system. Normally, however, some providers (such as general practice physicians) will contract with, rather than be employed by, the system.

Although structured and supported by national governments, the key operating bodies in these systems rarely are units of that government. In Canada, each province acts as health insurer to its citizens, within the five principles of the Canada Health Act. In Germany, most individuals are insured through one of hundreds of sickness funds, nonprofit public law bodies. In Japan, one major fund is run by the national government, but most citizens are in funds sponsored by corporations for their employees, or by local governments. In Sweden, county governments manage delivery systems.

These approaches may be combined in varied ways. In Australia, state governments operate public hospitals, but the national government insures physician and other ambulatory services. In the Netherlands, sickness funds cover most medical expenses but government insurance covers 'extraordinary' costs such as long-term care.

1. *Similarities*

In spite of their differences, NHC/I systems tend to share fundamental attributes. Any national healthcare or insurance scheme is a system of shared savings, to which many people contribute so that, if any one happens to get sick, they will not have to rely on their own resources to pay for their medical bills. Hence, whether specifically a 'care' or 'insurance' system, all socialize risk, and so insure.

Except in Switzerland, most citizens are compelled to affiliate with some, usually a specific, insurer or delivery system. In Germany and the Netherlands, upper-income individuals are allowed to purchase private insurance; in Switzerland, three-quarters of citizens are not compelled to have insurance, but all have strong financial incentives to insure.

The systems uncouple access to care from ability to pay for it. The elderly, having far more expensive needs than the average, always pay less than their costs. A family with five children will pay no more than a childless couple with the same income. High-earners pay more than low-earners.

All systems remove or drastically reduce price constraints on consumption of healthcare services, so require alternate methods to allocate scarce resources. These methods are the object of intense political conflict.

Whether formally part of a nation's 'social security' scheme or not, all these systems are part of a nation's social protections. Healthcare, however, is always much more administratively difficult and contentious than systems such as pensions and unemployment compensation, because of the need to work with and/or regulate healthcare providers.

2. *Coverage, Benefits, Equity*

The scope of protection varies across countries. Basic hospital and physician services are always covered. Pharmaceutical benefits, however, are part of the core package in Germany and Japan but not in Canada. Coverage of dental and long-term care varies widely. Abortion is covered in some countries but not others.

All systems protect beneficiaries from excessive financial loss, but they vary as to the extent of costs that individuals may face, and the equality of potential burdens. In Canadian provinces, everyone has the same mandatory insurance for the basic benefit package, but some services, such as outpatient pharmaceuticals, are not covered for all citizens. In Germany, about a fifth of the population is not required to join the sickness funds, half of these have private insurance instead, and individual sickness funds have had different risk profiles, so some have had higher costs than others. In Germany and Canada, individuals face no or trivial charges for most covered care; in Japan and France, the basic insurance requires that they pay 20 or 30 percent of charges for many services.

All NHC/I systems provide more equal access to care than in the USA, and all provide a better guarantee than those same countries had before their systems were developed or completed. Nevertheless, no system creates equal access.

Access may vary for geographic reasons. Rural areas almost always have inferior hospital care. Capital cities and university centers tend to be advantaged. One of the arguments for a national healthcare, rather than insurance, system is that direct control of delivery should make redressing supply

imbalances easier; but experience shows that promise is hard to keep.

Nor can the NHC/I scheme address all inequalities in resources. This is not simply a matter of finance: 'the well-educated and assertive patient is best able to make a particular system work at some level of satisfaction' (Rosenthal 1992). But monetary resources affect ability to pay cost sharing, to access national systems, and to obtain uncovered services.

The role of private insurance within each system significantly influences equity of access. Where private insurance is only allowed for uncovered services, as in Canada, it only affects access to those services. When private insurance pays for cost sharing in a NHI scheme, as in France, its holders will face fewer cost constraints than those without private insurance, and may consume more within the basic scheme. In Australia and the UK, people may buy private insurance for the same services covered by the national scheme. They do so to get access to private hospitals or 'private' beds in public hospitals, so as to 'jump the queue' for relatively elective procedures.

In short, NHC/I systems are meant to guarantee a social minimum, not equality of care. Each system has chosen, and may modify, a level of equity. All include a norm that applies to the great bulk of the population, some 'safety valve' by which a privileged minority can do better, and some shortages for the minority who face special barriers such as language and location.

3. *Analytic Issues and Approaches*

Many questions about the development and behavior of NHC/I systems are part of the study of the welfare state more broadly, or of the sociology, economics, and politics of medical care. Nevertheless, certain issues are peculiarly salient to the study of NHC/I systems. Any state's effort to structure healthcare so as to provide a social minimum of care, at reasonable cost, with enough acquiescence from medical providers to have a functioning system, raises the following concerns:

Reasons for, and how to limit, growth in spending.

The effects of different ways of organizing the guarantee, such as insurance versus care systems, or national versus local government influence

Governance of the medical profession, given the state's interest in equity and cost control. A related concern is how relations between the paying organizations and the medical profession affect the relationship between physicians and patients

Effects of NHC/I policies on health outcomes, with special interest in 'health' as opposed to 'delivery of medical services.'

In addition to these particular concerns, the study of any phenomenon yields a literature on the origin, development, and internal dynamics of any given system or the phenomenon more broadly. This includes discussions of individual cases (e.g., Eckstein 1958, Stone 1980) or comparative studies (e.g., Glaser 1991, White 1995).

Social science's contributions to answering such questions are mixed. Often there are too few observations relative to the number of variables to have convincing results. For instance, there are at least as many ways of organizing the guarantee as there are systems. Different analytical perspectives, such as the attitudes of economists and sociologists towards markets, also can lead to different conclusions.

Nevertheless, significant contributions include empirical work that documents social ills, dating back to the nineteenth Century. During the post-World War II period when most advanced industrial societies were completing their national guarantees, much of the study of NHC/I systems could be viewed as, in Richard Titmuss' (1966) term, the study of 'social administration.' This meant application of the perspectives of various disciplines: sociology to explain the division of labor in medicine and its effects on administration and production; an eclectic economics to explain cost and efficiency trends; and public administration theory to organization issues.

More recent evolution of NHC/I systems has created a role for the social sciences in management and policy choice. Budgeting and policy debates create a demand for evaluation of the return on alternative social investments. These normally measure effects on the population as a whole, or particular groups. Therefore governments have encouraged the development of both health economics and health services research (HSR).

Studies that show medically unexplained variation across regions in rates of different procedures have been used to de-legitimize physicians' claims to expertise. The economic and HSR perspectives more generally suggest that resources should be allocated not by professional assessment of individual patients' needs, but in accordance with statistical studies of populations. The most prominent forms of social science analysis of NHC/I systems (or medical care in general) thus serve payers' interests in challenging the medical profession's control of resource allocation (Belkin 1998).

In resisting the population focus, medical providers are not simply representing their parochial interests. NHC/I systems offer protection at moments of fear and pain, especially relief from worry about paying the bills when simply getting well seems challenging enough. If medical procedures offer even minimal hope at a desperate time, then the case for NHC/I systems is that a politically decisive coalition does not want to live in a society that divides its citizens into those with hope and those without. From this perspective the emphasis on population averages, or even health outcomes, misrepresents the function of NHC/I systems.

4. Origins, Evolution, and Prospects

Economists tend to explain government involvement in healthcare finance and delivery as a response to 'market failures' (Arrow 1963). Yet it is not an example of government replacing the market, because much healthcare historically has been financed and provided by neither. European sickness funds are the descendants of guild sickness and burial expense arrangements, in which membership in a continuous economic community (not something that could be purchased) was the basis for shared protection. Workers without that basis formed 'friendly societies.' Religious orders organized and staffed hospitals, while physicians provided some services in these hospitals without charge. When government was involved, it was largely local government, particularly public hospitals in urban centers.

NHC/I systems were created as governments transformed and expanded these varied arrangements into 'stable systems of social protection for their entire populations' (Glaser 1997). Industrial workers were included first. The German Kaiser's speech to the Reichstag on November 17, 1881, in which he proposed compulsory sickness insurance for urban workers, is conventionally viewed as the beginning of this development, but there were precedents of a sort within Prussia (Zollner 1982).

The Kaiser's Chancellor, Otto von Bismarck, conceived health insurance as part of a package of social protections that would bind urban workers, who might otherwise be a revolutionary threat, to the state. NHC/I often was a reformist initiative, promoted by conservative forces interested in buttressing the developing capitalist economic order, while radical socialists viewed it as a palliative, not fundamental reform (Eckstein 1958).

While new funds had to be created to cover workers whose workplaces or crafts could not support insurance, the early mandatory systems were largely based on pre-existing funds. Thus workers were gathered into different risk pools, with varied income and cost profiles that meant insurance was more of a burden for some than others. Funds had substantial independence in their governance and operation.

NHC/I systems then expanded by bringing successive social groups into the mandatory framework. This took a very long time: in Germany, agricultural workers were not included until 1972. Retirees generally were not in the original arrangements, and their inclusion required special financing. Where some citizens have never been included in the mandatory scheme, as in Germany and the Netherlands, governments over time have subjected the remaining private insurance to extensive regulation, so as to minimize adverse selection between the sectors. Inclusion of the more resistant groups (farmers, white-collar employees) has often required some concessions about financial terms. At the same time, one of its purposes has been to include higher-income groups whose contributions could be used to help subsidize the remainder. Since the 1970s, rising costs per patient and economic change have made many employer- or locality-based funds impractical, and the hundreds or thousands of funds in given countries have consolidated significantly. Hence there has been a continual evolution towards more aggregated pooling of risk.

Governments have also become more directly involved in managing the supply side of care. That is most evident in the national healthcare systems; indeed, the British opted for the National Health Service, rather than extending insurance, in large measure because it seemed that only state power could remedy shortages and maldistribution in the supply of physician and hospital services (Titmuss 1966, Eckstein 1958). Central control has been reduced by the historic role of municipal or state governments in hospital finance. Yet as local financing sources become less adequate to the costs of modern medicine, national governments, even in sickness fund systems, have increasingly used control of capital or regulation to allocate healthcare capacity.

When mandatory insurance covered only a portion of workers, and funds were substantially self-managed, funds in many countries sought to control costs by entering into selective contracts with providers—an approach nearly unanimously condemned, by physicians, as 'contract medicine.' As coverage was expanded, 'closed panels were eliminated in all countries,' in order to win acquiescence from physicians. 'All national health insurance laws eventually guaranteed the subscriber's free choice of primary doctor' (Glaser 1991).

Except where coverage was not yet universal, as in Australia, cost control has dominated the policy agenda since the early 1970s. Where the system is financed from tax dollars, the argument has been that rising costs, by increasing public deficits, would depress economic growth. Where finance is mainly by payroll contributions (as in the sickness fund systems), the argument has been that these proportional charges on top of wages depress hiring, so create unemployment.

Governments therefore have taken an increasingly active role in cost control, even where the funds are technically independent (as in Germany). They have limited capital investment; created or encouraged standard fee schedules and restrained those fees; created measures to limit the quantity of services, or compensate with lower fees; imposed budgets on institutions whose total expenses could be capped, such as hospitals; and imposed new cost-sharing or even 'de-listed' a few services. These measures significantly slowed cost increases during the 1980s, but increasing economic pressures meant that, even where costs grew very little as a share of GDP, political pressure to restrain costs further remained intense.

As these measures have increased dissatisfaction within the medical profession, whose incomes are at stake, they also have been highly unpleasant for the regulators. The result in the 1990s has been three common initiatives, pursued to varying degrees in many countries.

The 'outcomes movement' seeks to save by providing only 'appropriate' care, on the theory that there is lots of unnecessary, inappropriate care. By requiring some oversight of the individual decisions of clinicians, however, even the hard-to-criticize emphasis on results poses political challenges. In an ideal world, the medical profession would eliminate such 'waste' through collective self-regulation, but such corporatist self-governance is difficult even in Germany.

'Decentralization' is touted as a way to save by adjusting capacity more closely to the 'needs of the community,' usually assumed to mean less hospital care and more 'community-based care.' But people in communities tend to like hospitals, and the mantra that 'hospitals are expensive places' so money could be saved by substitution, though an old chant, may often be false.

Finally, there have been many initiatives to use 'market mechanisms' of some sort—basically selective contracting—to control costs. There is little evidence that these have saved money, perhaps because they have not been allowed to operate in ways that would threaten the underlying socialization of risk and costs that is fundamental to NHC/I systems (Saltman 1995).

These three reform approaches contradict earlier trends. Decentralization and selective contracting, in particular, might ultimately threaten the risk-pooling that is the basis for NHC/I systems. Are these common reform proposals part of a broader trend that threatens survival of the NHC/I systems? Perhaps, but there have been countervailing trends, such as the elimination of barriers that segregated lower-income Germans into more expensive funds, in the 1990s. Perhaps the much-touted reform trends are just failed stratagems by governments to control costs but avoid blame. If so, NHC/I systems are likely to continue to evolve in unique ways in each country, but remain basic aspects of virtually all advanced industrial societies.

See also: Health Care Markets: Theory and Practice; Health Care Systems: Comparative; Health Insurance: Economic and Risk Aspects

Bibliography

Arrow K J 1963 Uncertainty and the welfare economics of medical care. *American Economic Review* **53**: 940–73

Belkin G S 1998 The technocratic wish: Making sense and finding power in the 'managed' medical marketplace. In: Peterson M A (ed.) *Healthy Markets? The New Competition in Medical Care*. Duke University Press, Durham, NC

Eckstein H 1958 *The English Health Service: Its Origins, Structure, and Achievements*. Harvard University Press, Cambridge, MA

Evans R G 1984 *Strained Mercy: The Economics of Canadian Health Care*. Butterworth, Toronto, ON

Freddi G, Bjorkman J W 1989 *Controlling Medical Professionals: The Comparative Politics of Health Governance*. Sage Publications, London

Glaser W A 1991 *Health Insurance in Practice: International Variations in Financing, Benefits, and Problems*. Jossey-Bass, San Francisco

Marmor T R, Kudrle R T 1981 The development of the welfare state in North America. In: Flora P, Heidenheimer A (eds.) *The Development of Welfare States in Europe and America*. Transaction Books, New Brunswick, NJ

Rosenthal M M 1992 Systems, strategies, and some patient encounters: A discussion of twelve countries. In: Rosenthal M M, Frenkel M (eds.) *Health Care Systems and Their Patients: An International Perspective*, Westview Press, Boulder, CO

Saltman R B 1995 The role of competitive incentives in recent reforms of northern European health systems. In: Jerome-Forget M, White J, Wiener J M (eds.) *Health Care Reform Through Internal Markets: Experience and Proposals*. IRPP and Brookings, Washington, DC

Saltman R B, Figueras J 1997 *European Health Care Reform: Analysis of Current Strategies*. WHO Europe, Copenhagen, Denmark

Stone D A 1980 *The Limits of Professional Power: National Health Care in the Federal Republic of Germany*. University of Chicago Press, Chicago

Titmuss R M 1966 *Essays on 'the Welfare State'*, 2nd ed. Unwin University Books, London

White J 1995 *Competing Solutions: American Health Care Proposals and International Experience*. Brookings Institution, Washington, DC

Zollner D 1982 Germany. In: Kohler P A, Zacher H F (eds.) *The Evolution of Social Insurance 1881–1981. Studies of Germany, France, Great Britain, Austria, and Switzerland*, St Martin's, New York

J. White

National Innovation Systems

National Innovation Systems (NIS) are nationally distinct configurations of institutions and organizations that structure the development, diffusion, and use of new technologies, products, and processes in different ways. They typically include the institutions and organizations undertaking research and development, state science and technology policies, the financial system, the education and training system, labor market institutions, and the organization and strategies of dominant firms.

1. The Development of the Concept of National Innovation Systems

National innovation systems became the focus of extensive research and comparison in the mid-1980s as the result of four important factors. First, both policy makers and academics had become increasingly dissatisfied with the earlier linear model of knowledge-push innovation, which viewed technical change as being largely driven by research in the pure and applied sciences. Science and technology became understood as separate arenas of knowledge production and evaluation that developed interdependently in a variety of ways, and much technical change occurs within production organizations and technical communities (Freeman and Soete 1997).

This intellectual shift was reinforced, second, by an increasing awareness of the importance of informal, tacit knowledge in developing and implementing technological changes among economists, sociologists, and other researchers studying innovations. This, in turn, emphasized the highly contextual and local nature of much technological knowledge and its dependence on historically determined skills and search routines. Rather than being seen as a publicly available and readily accessible pool of formal knowledge, technology became viewed as more institutionally variable and costly to transfer and to use.

Third, the rise of Japan as a leading industrial power, together with the rapid economic development of newly industrialized countries such as South Korea and Taiwan and the relative decline of the US economy in the 1980s, stimulated considerable interest in alternative ways of organizing capitalist economies, especially variations in the role of interfirm networks and the coordinating role of the state (see, for example, Orru et al. 1997, Whitley 1992). Together with a number of studies of different forms of European corporatism and the organization of business interests (see, for example, Streeck and Schmitter 1985), this growing concern with the differences between national and regional systems of economic coordination and control, and their institutional interdependences, highlighted the considerable variations in innovative performance between the major industrial economies.

This broad concern with national success in developing and implementing new technologies led, fourth, to more systematic comparisons of national variations in sectoral specialization and technological development as policy makers and researchers began to study how and why firms from different kinds of countries were more effective in some industries and technologies than others. These variations have increasingly been seen as depending on societal institutions as well as upon the formal system of knowledge production and diffusion.

These shifts in perceptions and concerns among researchers, policy advisors, and policy makers encouraged a number of studies of NIS in the Americas, Europe, and Pacific Asia in the late 1980s and early 1990s, as well as leading to a series of debates about their nature and significance. Two collective volumes edited by Lundvall (1992) and Nelson (1993) brought many of these contributions together and effectively defined this new field of research. A later collection edited by Edquist (1997) provided some more systematic cross-sectoral and cross-national analyses of innovation patterns and their institutional connections.

There remain, however, considerable variations in how NIS are understood and how they are related to variations in technological specialization and performance. Beyond the broad consensus that they encompass additional organizations and agencies to those engaged in formal research and development (R&D) activities, they range from the whole network of institutions whose activities affect technological development to those more narrowly and directly concerned with innovation in firms. Most of the descriptions of NIS in Nelson's (1993) book, for example, focus on the location of R&D activity, sources of its funding, characteristics of key firms and dominant industries, and the roles of universities and state science and technology policies. Broader institutional influences and social structures such as labor market institutions and the skill credentialing system are typically not discussed.

This variety of conceptualizations of NIS extends to the types of innovations produced by them. While many contributors focus on technological change and ignore social and organizational innovations, others explicitly refer to new ways of organizing economic activities, even if in practice they tend to restrict their concrete analyses to technical developments. Clearly, if all forms of novelty in the economic sphere are to be included in the analysis of innovations, most aspects of national economies are involved in NIS. On the other hand, much recent work has emphasized the importance of organizational change in realizing the benefits of technological improvements, and numerous studies of 'learning organizations' (Dodgson 1993) have highlighted the importance of organizations' abilities to change their structure and behavior if they are to remain competitive. Just as different institutional arrangements encourage different patterns of technical change, so too they vary in their support for organizational flexibility and risk sharing, and thus in firms' ability to reshape themselves in rapidly changing economic and technological environments.

A further aspect of innovations that has become more central to the analysis of NIS in the 1990s has been their technological and market radicalness or discontinuity. Different types of innovation have come to be seen as more or less likely to develop in contrasting institutional environments. In particular, the prevalent pattern of incremental innovations in assembly and machinery industries in the more collab-

orative capitalisms of postwar Germany and Japan has been contrasted with the ability of the US economy to generate radical innovations that have led to the development of new industries, such as biotechnology and computer software. In such comparisons, NIS are viewed as distinctive configurations of institutions that encourage particular kinds of innovation strategies while inhibiting other ones, and so resulting in contrasting patterns of technological development.

The concept of NIS has become significant in understanding how and why market economies develop different patterns of sectoral and technological specialization and varying rates of economic growth. As such, it contributes to the general analysis of capitalist development, especially the comparative study of differences between types of capitalism and the role of technical change in these. The importance of innovation—broadly conceived—and technological development for economic growth also means that NIS are an important concern for policy makers and advisors as well as for students of technical change. The realization that innovation patterns vary significantly between countries as a result of institutional differences, such that innovation-led economic growth involves more than narrowly economic policies, has accordingly stimulated considerable interest in NIS among policy makers, especially in the linkages between institutional arrangements, state policies, and innovation patterns.

2. *Major Components of National Innovation Systems*

The comparative analysis of innovative performance and technological specialization across subnational regions, countries, and supranational regions involves a considerable variety of institutions, organizations, and collective actors. In addition to the public and private organization of R&D and allied science and technology policies, these include more general state structures and policies, the system governing the flow of investment capital and the evaluation of financial performance, the skill production, evaluation and control system, and the nature of firms' governance structures and organizational capabilities. These components of NIS will now be further discussed.

Universities, state research laboratories (including those incorporated into national academies of science), corporate and collaborative R&D efforts, and similar organizations for producing formal scientific and technological knowledge constitute the formal research systems of industrial societies. These vary considerably between nation states in terms of: (a) the balance between public and private provision and funding, (b) the kinds of industries that conduct R&D and use its outputs, (c) the predominant kinds and objectives of research conducted in separate organizations, and (d) the overall degree of integration and coordination of research and innovative activities carried out by different organizations. Where organizational boundaries and goals are fluid and often overlap, as in the USA, knowledge transfer and the recombination of different kinds of knowledge for novel purposes are easier to accomplish than where organizational objectives and boundaries are strongly differentiated and maintained, as seems to be the case for universities and firms in Japan since the war.

A related important component of NIS that varies between countries is the style and focus of state science and technology policies. Two major types can be distinguished. First, mission-oriented approaches concentrate resources on a small number of technologies and public goals—such as military or health ones—and attempt to achieve significant breakthroughs in these. Second, diffusion-oriented policies are more concerned with the development and diffusion of a range of new technologies that improve firms' competitiveness in existing industries. Both public resources and the state's coordinating activities are more directed towards upgrading existing technologies and developing new products that build upon current capabilities in this approach than they are to achieving radical innovations that transform entire industries.

These contrasting state science and technology policies are linked to more general attributes of state structures and behavior that affect innovation patterns. In particular, the overall capacity and willingness of the state to coordinate and direct economic development constitute key features of market economies that structure firms' management of risks and innovation strategies. Developmental states are those that take a leading role in organizing new industries and managing declining ones through the control of technology licenses, guidance of bank credit allocation, coordinating investment plans, and influencing market entry and exit. Japan, South Korea, and France resembled this type for parts of the postwar period.

Regulatory ones, in contrast, see their roles more in terms of maintaining general rules of the competitive game while remaining aloof from particular market processes and industry developments. The UK and USA represented this type for much of the twentieth century. In addition, states vary considerably in their involvement of intermediary associations, such as trade unions and employers' groups, in policy development and implementation, as well as their more general encouragement of collaboration between competitors, suppliers, and customers. This, in turn, affects the willingness of firms to share risks and information in the development of new technologies.

A crucial institution in the development of new technologies is the system governing the flow of capital and the terms on which different economic agents are able to obtain investment funds. One important distinction is between capital market based financial systems, such as those of the UK and USA, and credit-

based ones, such as those of most continental European countries, Japan, and South Korea. In the former, capital is largely allocated through competition between suppliers and users of funds in large, liquid stock markets. These markets facilitate investor entry and exit to and from businesses by providing low cost transfers of property rights and establishing a strong market for corporate control.

In contrast, credit-based financial systems allocate capital through intermediaries such as banks and insurance companies, sometimes under state guidance, and such capital markets as exist are small and illiquid. Entry and exit are more difficult in these circumstances as there is no large, liquid, and anonymous market for trading financial claims, which means that investments, once made, are relatively long term. Investors thus are locked in to the destinies of their major clients to a greater extent in these systems. This encourages more mutual dependence and risk sharing between credit providing organizations and their larger customers, but inhibits radical changes in industrial structure and investments in highly novel and risky ventures. As a result, the significant development of the venture capital industry in high technology sectors in the USA during the last two decades or so of the twentieth century has been less evident in credit-based financial systems.

Increased awareness of the importance of process innovation and organizational learning in developing and implementing new technologies has focused attention on the education and training system as a key component of NIS. This is significant in two major ways. First, it determines the general level of workforce skills in an economy, and therefore the ease with which new technologies can be developed, adopted, and changed. Second, it structures the kinds of skills that are available to develop and implement particular kinds of innovations. A broad distinction has been drawn between education systems that focus on academic skills to the exclusion or downgrading of practical ones, such as those in France and Japan, and those that invest considerable resources in training children and adults in manual and nonmanual skills that are prestigious and highly valued by employers.

The dual training systems of Germany and some Scandinavian societies exemplify the latter. They typically involve close cooperation between employers, unions, and local state bodies and integrate formal training in technical schools and colleges with practical apprenticeships in firms. These education and training systems provide highly skilled workers who are able to learn new technologies and adapt to innovations. When combined with corporatist systems of labor relations and restrictions on managers' ability to hire and fire workers, such training systems develop considerable interdependences between employers and skilled workers so that they have common interests in skill upgrading and continuous improvement of products and processes. They thus enable and encourage incremental innovation strategies. Conversely, more fluid external labor markets facilitate the circulation of skills and knowledge, and enable new firms to recruit different kinds of skills more easily than they could in corporatist societies.

A further important feature of education and training systems affecting innovation patterns is of course their development of scientists and engineers, or formal knowledge producers and developers more generally. Both the sheer number of doctoral researchers and their distribution between fields are significant aspects of education systems that vary between countries. The USA, for example, has developed a large number of doctoral programs in most scientific disciplines since the war that produce many researchers every year. In contrast, the development of research training programs at the doctoral level is much less significant in Japan. The proportion of graduates and postgraduate students in the physical, biological, and engineering sciences also varies between Britain, where many more natural scientists than engineers graduate each year, and most continental European countries and Japan, where the reverse is the case. The increasing importance of formal, codified knowledge in technical change means that the production of knowledge producers has become more significant in many industries, both to generate new knowledge and to absorb, evaluate, and use new knowledge produced by others.

In addition to influencing directly the pace and type of technological change through the generation, diffusion, and use of different kinds of knowledge, these institutions and agencies in the business environment also affect patterns of innovation through their connections with forms of economic organization and types of firm. Market economies differ considerably in the extent to which economic activities are integrated through non-spot-market means, and how they are, from the industrial districts of North East and Central Italy and other European societies to the highly coordinated postwar Japanese economy, dominated by large firms, vertical partnerships, and horizontal inter-market groups. Associated with these different types of economic coordination and control, or business systems, are considerable variations in prevalent firm type (Whitley 1999). Dominant kinds of companies in twentieth century market economies vary from the small owner-controlled, highly specialized enterprise to the large, diversified managerially integrated corporation that has developed in the USA and the less vertically integrated and horizontally diversified Japanese large firm. These different business systems and firm types develop interdependently with distinctive societal institutions and tend to follow contrasting innovation strategies and patterns of technological change.

Opportunistic firms in fragmented business systems, for instance, typically compete by being highly responsive to changes in market demand and rarely

commit themselves to any one technology or market for very long. Innovations made by these kinds of firms thus tend to be customer driven and imitative rather than very novel. Isolated hierarchies in compartmentalized business systems, on the other hand, develop complex organizational capabilities for introducing and marketing standardized products for large anonymous markets, but without sharing authority and risks with business partners and employees. They compete by combining mass production with mass marketing and driving down unit costs through formalized routines and systematized procedures. Innovations introduced by these kinds of firms thus tend to be restricted to products and processes that are highly standardized and capable of being sold to a mass market without being continuously improved or adapted to different users' tastes.

Cooperative hierarchies in coordinated business systems, in contrast, rely more on risk sharing with partners and skilled workers to develop high quality products for demanding customers and to upgrade these continuously. Information in these kinds of firms and business systems is widely diffused and used to improve products and services so that innovations are often quite complex and risky, but essentially build on existing organizational capabilities and commitments. They are, then, incremental in terms of current organizations and alliances, while sometimes being technically discontinuous.

3. *Current Issues and Developments*

Two major concerns about the concept of NIS arise from the focus on national units of analysis, on the one hand, and the idea that institutions, firms, and technologies constitute coherent systems, on the other hand. Many have questioned the appropriateness of concentrating on the nation state as the primary level for studying technological development, as opposed to the sector or firm, and others have queried the basis for assuming that NIS are systemic. These issues are connected to broader questions about the analytical purpose of the concept and the sorts of explanations it could usefully provide. In particular, it is not always clear how differences between NIS account for variations in technological change, or indeed exactly what the concept is intended to explain.

At least three distinct kinds of intellectual enterprise can be distinguished in considering how national and systemic are patterns of innovation. First, there is the comparative analysis of formal innovative activity at the national level, especially the organization and funding of designated R&D activity. This has been the focus of numerous surveys since the war, and reflects the prevalence of the science push, linear view of innovation. Since much of this activity is undertaken in state-supported and/or controlled organizations and has been the object of many state science and technology policies, it is not surprising that it mostly deals with national units of analysis.

Second, there is the comparative analysis of innovative performance and technological development that focuses on how and why different societies, states, and regions vary in their sectoral specialization and apparent success in developing different technologies over time. Insofar as these patterns vary systematically between nation states because of national differences in dominant institutions and policies, then the national level of analysis is clearly the appropriate one. It should be noted, however, that the intellectual focus on NIS in this institutionalist approach depends on innovation patterns, business systems, and institutions varying interdependently and significantly between countries. Where they differ more between regions within states, or between types of institutional environments that are characteristic of a number of states, such as the legal and financial systems of the USA and UK, on the one hand, and those of many continental European countries, on the other hand, NIS may be less important than regional or transnational ones.

Third, many studies of technological change and industrial development focus more on sectoral differences than national ones. These concern both technological and economic characteristics of industries. Technological systems vary considerably on two dimensions. First, they differ in the degree of interdependence of their components, or the extent to which changes in any one part of the system affect the operation of other ones, and so whether innovations are systemic or modular. Second, they differ in their causal complexity and therefore in the amount of knowledge required to make them work effectively. These differences have significant implications for coordination and control mechanisms that in turn are dependent upon institutional arrangements for their development and reproduction. As a result, different kinds of national business environments encourage the development of industries with varied degrees of technological integration and complexity (Kitschelt 1991).

Economic aspects of sectoral technological regimes comprise four basic factors according to Malerba and Orsenigo (1996): opportunity conditions, appropriability conditions, the cumulativeness of technological development, and the nature of the knowledge involved. The first refers to the likelihood of significant innovations arising from investments in R&D and other forms of research, the second to the feasibility of appropriating gains from innovations, the third to the dependence of any one innovation on previous innovations and competences, while the knowledge base of a sector varies in its complexity, tacitness, specificity, and decomposability. These aspects of different technological regimes generate distinctive patterns of industrial organization and innovation between sectors that vary in terms of firm size, the concentration

of innovative activities in a few firms and geographical dispersion.

In both instances of this sectoral approach, it is assumed that particular features of production chains require—through some functional logic not always explicitly stated—specific kinds of coordination and control mechanisms such that different sectoral types develop distinctive patterns of innovation and economic organization that fit different institutional contexts. Thus, the critical unit of analysis and determining factors are sectoral, not national, in this approach. Each type of industry is characterized by one dominant kind of innovative pattern and organization that is standardized across countries such that institutional variations affect the relative significance of different sectors, but not their specific pattern of coordination and control. National innovation systems from this point of view, then, are simply combinations of institutions, agencies, and policies that fit some industries more than others and so encourage nationally distinctive patterns of sectoral and technological specialization. They do not in themselves generate different kinds of innovation and technical change that compete for sectoral dominance, and so are not especially systemic.

Contrarily, the institutional approach focuses on national, or regional, or transnational innovation systems as more coherent units of analysis because distinctive institutional arrangements are seen as important influences on the kinds of firms that develop, the strategies they pursue, and the capabilities they develop. Different types of business environment, public research systems etc., generate different innovation strategies in systematic ways, in this view, and so constitute more distinctive and cohesive innovation systems than those described by the sectoral approach. Patterns of technological development are here considered to vary more between institutional contexts, and less between sectors, than in the sectoral approach.

The question of how national and systemic are NIS depends, then, on whether they are conceived as varying national combinations of standardized sectoral innovation systems and technological imperatives, or as distinctive configurations of interdependent institutions, agencies, policies, business firms, and patterns of innovations that become established and reproduced most coherently at the national level in twentieth-century capitalisms. Clearly, these latter configurations are more systemic than the former ones, but need not necessarily be more nationally specific since that depends on how cohesive and distinctive are the key institutions and forms of economic organization at the national level.

This contrast in conceptions of NIS is reflected in researchers' intellectual goals and preferred modes of explanation. The sectoral approach attempts to show how the coordination requirements of different industries are more or less fulfilled by the dominant institutions and policies of each nation state to produce nationally distinctive patterns of innovation and technological specialization. More institutionally focused oriented researchers, in contrast, focus on how particular types of institutions and agencies develop interdependently to generate distinctive innovation strategies and outcomes such that contrasting types of innovation and technical change are systematically produced by different societal arrangements.

4. *Future Directions*

This divergence in intellectual goals and strategies is likely to continue. Policy-oriented researchers, whether at the regional, national, or supranational levels, will focus on how public agencies and policies do, or do not, appear to enhance innovative performance in different contexts. National innovation systems here are summaries of particular structures and policies that seem to be connected to innovative outcomes and are, in theory, amenable to policy interventions. Levels of analysis are determined more by units of political and administrative action than by theoretical frameworks.

Researchers analyzing different forms of economic organization and the institutional structuring of innovations are more interested in contrasting systems of innovation than in describing different national patterns of innovative activity. They will therefore continue to see NIS as particular kinds of interdependences between institutions, firms, and technical change that are more or less approximated by those manifested in specific countries. Comparative analyses of NIS in this approach focus on how distinctive configurations of institutions develop separate patterns of innovation and technical specialization, and so reproduce different varieties of capitalist economies. Future work will concentrate on identifying the processes through which different kinds of institutions and organizations generate contrasting patterns of technical change in different state and regions.

Finally, analysts of sectoral innovation systems and technological regimes remain primarily interested in the logics through which different kinds of industries and technologies develop distinct governance structures and forms of economic organization. NIS are here constituted by the interaction between nationally specific institutions and policies and these logics resulting in nationally varied patterns of innovations. Institutions and policies in this approach are construed as the contextual contingencies that modify and guide the realization of various sectoral logics. Comparative research will therefore continue to focus on how different countries develop contrasting technologies and industrial structures as the result of such intersections.

The concept of NIS will remain, then, rather eclectic and diffuse, reflecting continuing divergences in re-

searchers' priorities and concerns. As long as the nation state remains the primary unit of political competition and controller of key resources, and national institutions and policies remain distinctive, NIS are likely to continue to be analyzed and compared. However, different researchers will also continue to understand the term in different ways and use it for different purposes.

See also: Evolution: Diffusion of Innovations; History of Science; History of Technology; Innovation, Theory of; Science and Technology, Social Study of: Computers and Information Technology; Scientific Academies, History of; Technological Innovation; Technology Transfer; Universities and Science and Technology: Europe; Universities and Science and Technology: United States

Bibliography

Dodgson M 1993 Organizational learning: A review of some of the literatures. *Organization Studies* **14**: 375–94

Edquist C (ed.) 1997 *Systems of Innovation: Technologies, Institutions and Organizations*. Pinter, London

Freeman C, Soete L 1997 *The Economics of Industrial Innovation*. Pinter, London

Kitschelt H 1991 Industrial governance structures, innovation strategies, and the case of Japan: Sectoral or cross-national comparative analysis? *International Organization* **45**: 453–94

Lundvall B-A (ed.) 1992 *National Systems of Innovation*. Pinter, London

Malerba F, Orsenigo L 1996 Technological regimes and firm behavior. In: Dosi G, Malerba F (eds.) *Organization and Strategy in the Evolution of the Enterprise*. Macmillan, London

Nelson R (ed.) 1993 *National Innovation Systems*. Oxford University Press, Oxford, UK

Orru M, Biggart N, Hamilton G 1997 *The Economic Organisation of East Asian Capitalism*. Sage, Thousand Oaks, CA

Streeck W, Schmitter P (eds.) 1985 *Private Interest Government*. Sage, London

Whitley R 1992 *Business Systems in East Asia: Firms, Markets and Societies*. Sage, London

Whitley R 1999 *Divergent Capitalisms: The Social Structuring and Change of Business Systems*. Oxford University Press, Oxford, UK

R. D. Whitley

National Liberation Movements

National liberation movements (NLMs) constitute a large spectrum of socio-political movements which share in common the target of struggling for establishing an independent state for what they consider their nation within the borders of a territory recognized by the international community. These movements also share that they privilege the struggle for the liberation of their nation over the many other targets of other social movements, and therefore call their people to unite around that priority. The very existence of such movements assumes that the nation either must be united with a view of constituting a single state, or must separate from a plurinational state in which it is included and that is in their view dominated by another nation, or has to emancipate from a colonial status and achieve political independence of that territory. NLMs have indeed been among the major political forces which operated mainly in Europe in the nineteenth century and later, in the twentieth century, expanded to Asia and Africa. They have shaped major changes in the political geography of the modern world and possibly will continue to do so in some areas.

There are, of course, different approaches to the study of this set of important social phenomena. Some scholars take for granted the pre-existence of the nation which these movements mobilize, and therefore focus on the analyses which help in understanding the reasons for their success (or failure), their strategies, tactics and ability to mobilize objective realities operating within the society or in the field of international politics which may support their project.

Other scholars focus rather on the relationship between the deployment of NLMs and the major changes which singularize the process of modernization of the nineteenth and twentieth centuries. Along with Eric Hobsbawn, Immanuel Wallerstein, and the Austro-marxists, the author will adopt that specific approach for the analysis of the NLMs. Comodification of production, industrialization, urbanization, expansion of education—and all these dimensions of the process of modernization are deeply related to that of capital accumulation—played, in that respect, a decisive role in the crystallization of a new type of community, usually called nation. The new economy and social life implied an intensified use of written communications and therefore gave to the language used for education and in the workplace a role far greater than it had been the case in premodern societies.

1. Centers and Peripheries in the Modern World System

Nations as well as social classes are products of modern times, even if one can discover ancestors to them in the past, which anyway have been radically reshaped by the dominant rationale of capital accumulation operating simultaneously at the various local ('national') and global levels. In that respect, the history of capitalism is that of a continuous conflict between the logic of capital accumulation and that

through which various other social interests express themselves. NLMs have their place in that historical frame, just as the other social movements, particularly those of the working classes. In fact, these two families of movements–labor and national liberation—have represented, throughout at least the nineteenth and twentieth centuries the major antisystemic forces, compelling capital accumulation to adjust to the social interests that they meant to represent.

National liberation refers to 'nation' and that term presupposes certain articulations between this reality, real or alleged, and other realities such as the state, the world system of states, the local economy as related to the world economy, social classes. We owe these concepts and their articulation into a system to the different social theories developed in the light of the nineteenth century European historical experience. Within this framework, the elaboration of two sets of theories took place—as it turned out, in counterpoint to one another: on the one hand, socialism and the theory of class struggle: on the other hand, nationalism and the theory of class integration into the nation-state. Both theories take account of many aspects of the immediate reality which is marked both by social struggles ending in deep transformations of capitalism and even in revolutions, and by struggles between nation-states ending sometimes in war as well as national liberation movements ending in many cases in the creation of new independent states. For protagonists of these theories, they have proven to be potent guides to action.

The expansion of capitalism on a world scale has always been, and continues to be, based on unequal economic and therefore also political relations between those areas which constituted themselves as full fledged centers of the system controlling the global process of accumulation and those areas which have been submitted to the logic of that accumulation i.e., the peripheries. Therefore, the relationship between state and nation on the one hand and the economic system on the other is not similar in the centers and in the peripheries of global modern capitalism. In the centers there is a rough coincidence between the state (whether national or multinational) and an autocentric productive system; in the peripheries such a coincidence is always lacking. That coincidence between the state, the nation, the high level of capitalist transformation and the dominant position of such states in the world capitalist system which led these central partners to form 'autocentered' interdependent economic units enjoying a high degree of autonomy vis-à-vis each other, defines a possible field for 'national' economic policy. The instruments of this policy—the national centralized monetary system, customs, the network of material infrastructures in transport and communications, the unifying effects of a 'national' language, the unified administrative system, and so on—enjoy a definite autonomy in relation to the constraints imposed by an economy operating worldwide. Relations between classes, however wrought, are relegated to and by the national state.

True, there is no Wall of China to separate these national systems from the world system that they constitute. Internal social relationships are partly dependent upon positions occupied by the national states, in the world hierarchy. If social relations permit it, these states can improve their position by pursuing coherent national policies. This effectiveness in turn facilitates social compromise and, without 'abolishing the class struggle,' puts definite limits to conflicts.

However, at closer inspection it appears that these circumstances, pervasive but limited in time to the nineteenth century, are even more limited in space. Around a few 'model' nation-states, the world of the capitalist system—structured by different pasts which in turn lost their legitimacy and efficacy—remains undefined in the light of an uncertain and obscure future. Therefore, the problem changes when we quit the limited framework imposed by the central bourgeois nation-states. For this forces us to examine more closely 'regions' whether they are organized into the states or not. Regions are peripheral in relation to continuously expanding capitalist reproduction. On this level there is only a central state, i.e., a state which masters external relations and submits them to the logic of autocentered accumulation. On different levels, there are only 'countries,' which are administered from outside as colonies or semicolonies; these appear to be independent but incapable not only of molding the outside according to their needs but also of avoiding their drift and shaping from outside.

The consequence of that fundamental and continuous contrast between centers and peripheries immanent to the history of really existing global capitalist expansion is that 'national liberation movements' have been continuously produced by the uneven development associated with that expansion. These movements have always pursued the same major target, that of 'catching up' with the advanced centers, through constructing in their turn a modernized independent autocentered nation-state able to act as an autonomous participant in the global capitalist system. They always have nurtured a basic ideology of modernization and national building. In their struggle against the dominant forces in the global system, NLMs have also often attempted to build 'international common fronts' associating the movements involved. Mutual sympathy, common strategic and tactic approaches helped in constructing such common fronts. These common fronts often survived the victory of the national liberation movements, i.e., after those movements had conquered political independence. Yet these fronts were also gradually eroded by the very success of the NLMs. The new states pursued their own policy targets within the global system, which led to their insisting on their particular interests and specific strategic and tactical policies aiming at helping them to achieve them. The similarity with the

other major antisystemic movement—that of the working classes and socialism which have built their own successive 'internationals'—is striking.

2. *Stages of Deployment of the NLMs*

Capitalist global expansion has moved through successive stages. For each of those stages specific forms of the relationship between centers and peripheries can be identified. The NLMs produced in each stage have therefore been confronted with specific challenges.

During the mercantilist stage of capitalist expansion (1500–1800) the American continent was conquered and indigenous populations subjugated when not exterminated, Africans were imported as slaves, while migrants from Europe either created autonomous small production systems (in New England, for instance) or constituted themselves as the local ruling class operating as a transmission belt for the metropoles dominant powers (slave owners and masters of encomiendas). The NLMs in the late eighteenth century (the 'American revolution') and in the early 1800s (in Latin America) were led by those local ruling classes of European origin, disregarding the interests of subjugated peoples (black slaves and indigenous Indians). Haiti is in this respect an exception. The common front which was constituted with a view to excluding European powers from the continent was quickly captured by the new capitalist center which emerged in the USA (that is the meaning of the Monroe declaration in 1823). The pattern of an uneven relation was gradually constructed in that frame between the US center and its Latin American periphery. It was only much later, in the twentieth century, that nationalist bourgeois movements started questioning that uneven relation, along in some cases with a revival of indigenous Indian peoples movements. The Neozapatistas (Chiapas) movement in Mexico is the most recent expression of that unfinished struggle for national liberation and social progressive change.

A second wave of NLMs developed throughout the nineteenth century in central, eastern, and southern Europe, associated with the expansion of capitalism beyond its north-western European corner of origin. In two cases, Germany and Italy, these movements led to the constitution of new nations, which became new capitalist centers. Further to the east the three multiethnic autocratic empires (the Russian, the Austro-Hungarian, and the Ottoman) were assaulted by the expansion of capitalism, moved by a conjunction of internal forces (produced by industrialization and the formation of modern social classes—a bourgeoisie and a working class) and the impact of external forces dominating the global system. The dilemma with which those societies were confronted was the following: could capitalist development flourish within the frame of such large states, becoming multinational and democratic, or should they be dismantled on a more or less national base? In these countries more than elsewhere the various class struggles (between the autocratic leading groups of the old regime and the new bourgeoisie, between the latter and the working class, between the large landowners—often of different ethnic origin—and their peasants), the political struggles (for democracy, for socialism) and the 'national question' overlapped and interacted in a complex manner. The major organizations within the socialist movement were rather in favor of maintaining the large multinational state (so was the meaning of the struggle of the Austro-marxists for cultural autonomy). However, the peasant and bourgeois movements favoured 'national liberation first,' and that option was finally the one which succeeded in imposing itself.

Simultaneously, during the second half of the nineteenth century, there was an extraordinary expansion of global capitalism which conquered Asia and Africa and colonial (or at least semicolonial) systems were established there in most cases. The national question was moving out of Europe and national liberation movements of a new type started being constituted, inspired by the local elites and fractions at least of the new bourgeoisies. The ancestors of those NLMs which played a decisive role in the making of the twentieth century could be identified at least, in China, India, Persia, Egypt, and Syria, where it animated cultural revivals (such as the Arab Nahda) and the first political modern organizations. These new NLMs related to class issues in a variety of ways, but in general they gradually became what has been labeled later 'antifeudal–anticompradore' political movements, associating in their struggles for reconquerring the lost national independence and modernization the 'external enemy' ('imperialism') and its local allies (precisely the 'feudals' and 'compradore'). Yet the dominant force among the international socialist movement—the social democracy of the Second International—ignored those new NLMs, overlooked their importance for the future and shared the dominant European imperialist ideology of the 'civilisation mission' of colonialism.

The colonization of Asia and Africa has nationally produced a wave of most important NLMs which finally won their struggle for national independence after World War II and played an important role in the changing of the face of the world since. The NLMs of Africa and Asia were very different in outlook. The past and recent history of their peoples, the social forces, the political and ideological parties and trends which they represented, their visions of the future of the societies to be constructed or to be reconstructed, and its relationships with the West were all subjects on which they differed. However, what brought them together was their common project, which included the completion of the political decolonization of Asia and Africa. Beyond, the Afro-Asian States which emerged out of the victories of NLMs after World

War II all considered that the political independence they had regained was only the means, the end being the achievement of economic, social and political liberation. In the opinion of by far the majority 'development' would be possible through interdependence within the world economy. The NLMs and then after the new states they created never intended to 'leave the system,' or 'delink' even if they did not all share the same strategic and tactical vision of 'development.' But, to varying degrees, they considered that the building of an independent, developed economy and society (even within global interdependence) implied a certain amount of 'conflict' with the West in its dominance (the radical wing thought it necessary to put an end to the control of national economies by the capital of foreign monopolies). Moreover, anxious to safeguard the independence which they had regained, they refused to take part in the military struggle at planet level and to serve as bases for the surrounding of the socialist countries which American hegemony was attempting to impose. However, they also thought that if they refused to be part of the Atlantic military alliance, this did not imply that it would be necessary for them to seek the protection of the latter's adversary, namely the USSR, whence 'neutralism' and 'non-alignment.'

Wilson's 'principle of nationalites' (a state for each nation), which was restricted to Europe after World War I, was thereafter partly recognized to Asian and African peoples by the UN after World War II. Yet it took several more decades for the NLMs to achieve their target. That success was favored to a certain extent by the bipolarity which characterized the international system in those times.

The coming together of the Afro-Asian states had already begun by the setting up of the Arab–Asian group within the United Nations in the late 1940s with a view to defending the independence of the colonies in arms. In 1955, the foremost Heads of State of the countries in Asia and Africa which had regained their political independence met for the first time at Bandung. Nasser, Nehro, Sukarno and Chou En Lai were to have a considerable influence on their times, even although the experience of the new powers they represented and to a large extent inspired was very recent. India and Indonesia had been independent for barely 10 years, communist China for not quite 5 and in Egypt the monarchy had disappeared from the scene only 3 years previously. The struggle for the achievement of the historical task of independence was not over: the first war in Vietnam was barely over when the second one loomed on the horizon, the Korean war had just ended in a status quo, the Algerian war was at its height, the decolonization of Africa south of the Sahara had not even been contemplated, the Palestinian drama was still in its first stages, South Africa was closing itself in hard apartheid. Bandung had the effect of reinforcing this rapprochement and of giving an impetus to this struggle. Three years later, in independent Accra, Kwame Nkrumah declared that 'Africa must unite.' But it was only to do so one independence had been achieved and after the failure of Nkrumah's pan-africanism, followed by the demonstration of the impotence of the two groups formed as a result of the Congo crisis (the Casablanca group and the Monrovia group between 1960 and 1963), in the minimal form of the setting up of the OAU in 1963.

From summit to summit, throughout the 1960s and 1970s, 'non-alignment' was nevertheless gradually to change direction from a position of a front of political solidarity based on the support for national liberation movements and struggles and the refusal of military pacts, to a position of building a common front with a view to reinforcing collective economic demands from the North. It inspired the 'decades of development' in the United Nations system, the creation of UNCTAD in 1965 and a whole series of global negotiations on crucial economic global issues. The struggle for a 'new international economic order' which began in 1975 after the October 1973 war and the revision of oil prices was the culmination of this development and the beginning of its end.

The West had no intentions of accepting the spirit of Bandung as a foregone conclusion on either the political level or on that of the economic struggle. The project of society which constituted a kind of basic charter for the NLMs and thereafter inspired the policies of the Asian and African states did exist, although it was often implicit and imprecise. This project, setting aside its various concrete manifestations and the specificities of its expression at national level, can be defined in terms of the following elements: (a) the desire to develop the forces of production and to diversify production (in particular to industrialize); (b) the desire to ensure that the nation-state controls and directs the process, (c) the belief that 'technical' models constitute 'neutral' data which, once mastered, only have to be reproduced; (d) the belief that the initiative did not imply its being initiated by the people in the first instance but only their support for state actions; (e) the belief that the process is not fundamentally in contradiction with the participation in exchange within the world capitalist system, even if it implied conflicts at times with the latter.

The circumstances of capitalist expansion between 1955 and 1970 to some extent encouraged the crystallization of this project. The implementation of this national bourgeois project implies that the hegemonic national, bourgeois class, by means of its state, controls a series of processes which include as a minimum:

(a) the management of the reproduction of the labor force, which implies relatively full and balanced development so that local agriculture, for example, is capable of providing the elements required for this reproduction in sufficient quantity and at suitable prices to ensure the valorization of capital;

(b) the control of national natural resources;

(c) the control of local markets and the capacity of entering the world market on competitive conditions;

(d) the control of the financial circuits enabling the centralization of surpluses and the directing of its productive usages;

(e) the control of the technologies being used at the level of the forces of production concerned.

Seen from this angle, the experiences of the Third World inspired by the ideology of NLMs can be divided into two categories: that of the countries which have attempted nothing more than the acceleration of growth without considering the implementation of the conditions outlined above (Ivory Coast, Kenya, Pakistan, Saudi Arabia, etc.) and that of the countries which have attempted to implement the conditions outlined (Nasser's Egypt, Algeria, Tanzania, India, Brazil and South Korea, etc.). It is obvious that the classification does not necessarily distinguish between regimes concerned with the implementation of a modicum of social justice and reforms, in particular agrarian reforms (e.g., Nasser's Egypt and South Korea) and those which have had no hesitation in accepting the intensification of social inequality (e.g., Brazil). Nor does it necessarily distinguish between the attitudes towards transnational capital (Brazil and Kenya both encourage the latter, but the former attempts to integrate it into its own national policy, whereas the latter simply adjusts to its demands), nor even between the question of the political relationship of conflict or alliance with the East and the West. It is true that there are correlations but the variations in composition as a function of the concrete situation make a special case of each of the Third World countries. However, over and above the variety of experiences, the fact remains that the most coherent implementations have taken place when there has been the conjunction of an intense nationalist struggle and a powerful social revolution. Led by communist parties, these NLMs in China, Vietnam, and Cuba are the most distinct examples of such a radicalization.

It is no longer possible to disregard the inadequacies of all these attempts, which collapsed when circumstances became favorable to their overthrow or erosion. The food and agricultural crisis, the external financial debt, the increased technological crisis, the external financial debt, the increased technological dependence, the fragility of the capacity to resist possible military aggressions, the advance of waste on the lines of the capitalist consumption models and of their ideological and cultural content, point to the historical limits of the attempts. Even before the present crisis presented the possibility of an 'offensive from the West' which would succeed in reversing the trends, these inadequacies had already led to an impasse.

In this sense, we may consider that the NLM project under consideration deserves to be described as a national, bourgeois project. The project for the society that the NLMs had in mind was, as we have shown above, that of creating a new state integrated in the global capitalist system. In that perspective the ruling class imagined itself becoming similar to that which controls the process of accumulation in the capitalist centers, and therefore it deserves to be qualified as 'bourgeoisie' (including state bourgeoisie forms), although of course it constituted itself out of different historical origins. Moreover, it has proved to be impossible. In this sense, history has proved that the national bourgeoisie is incapable, in our times, of implementing what it achieved elsewhere, in Europe, in North America and in Japan in the nineteenth century. A concrete examination of the history of NLM ideologies and the policies they inspired in the Third World independent countries would illustrate that fact: that of the uninterrupted sequence of attempts by the national bourgeoisie, of their repeated failures and of the submission to the demands of subordination which followed on each occasion; in Latin America since the nineteenth century (we shall only refer here to the more recent examples of the Mexican revolution between 1910 and 1920 and of Peronism in Argentina), in India (e.g., the change in direction since Nehru's 'first plan' until the return to power of the right after the first failure of the Congress), as in numerous Arab and African countries.

The favorable economic circumstances at the end of World War II were indeed exceptional and could not last. On the economic level, strong growth in the 'North' facilitated 'adjustment' in the South. On the political level, peaceful co-existence along with the rise of Soviet power, and its combination with the decline of the old French and British colonialisms and the rise of the Afro-Asian struggles for independence made the Soviet alliance truly efficient.

3. *New Challenges, Problems, and Perspectives*

The page of that history of NLMs has now been turned and the global capitalist system has entered in a new stage of its expansion. In this respect it is often said that any attempt at a national development has become irrational and impossible, that the challenges are basically global and therefore cannot find their solutions but at that level, in other words that there is no more national capital, national bourgeoisies, but only a globalized capital managed by a globalized bourgeoisie. Yet the system is not moving toward reducing global polarization, but is rather deepening it. Uneven development between centers and peripheries and therefore nations and peoples not only continues but aggravates, albeit in new conditions. The global capital and bourgeoisie—if existing—would be restricted to the central triad, whereas in the

peripheries they are bound to be compradore transmission belts of domination of the first. Therefore, the national question remains on the agenda; it continues to be a real challenge. What can be said in that respect is that the change marks the end of an epoch which in the case of Asia, Africa and, Latin America can be called the century of the national bourgeoisie, in the sense that it has precisively been marked by successive national bourgeoisie edification attempts. The Third World bourgeoisie now finally sees its own development in terms of the comprador subordination imposed upon it by the expansion of transnational capitalism. A new wave of NLMs should therefore be expected, different in their social targets and methods from the previous ones. The crystallization of the alternative new national popular (not bourgeois) strategy involves a merging of three conditions. First, delinking in the sense of a strict submission of external relations in all areas to the logic of internal choices taken without consideration of criteria relating to world capitalist rationality. Second, a political capacity to operate social progressive reforms through advances of democracy. This political capacity is a condition of delinking—since existing hegemonic classes have no interest in it. Delinking without that transfer of political hegemony has little chance of emerging, and if it did emerge under such conditions, it would lead to an impasse. Third, a capacity for technological absorption and invention, without which the autonomy of decision making acquired could not be realized.

Thus defined, the conditions for a positive response to the challenge of history appear severe, and any merging of such conditions remote. That is not surprising. The response to new challenges never comes spontaneously. It takes time for the social structures to stabilize, for the new forms of social movements to produce efficient strategies, for them to formulate legitimate ideologies in keeping with the real challenges. In the meantime, what we have is rather chaos, and therefore disarray and nostalgias of the past, in other words, false responses to real problems. It is in this frame that we can observe revivals of movements which label themselves 'national liberation,' usually on the basis of ethnicity and religion, denying importance to any other dimension of social reality, such as class. Often presented by the dominant ambiguous ideologies as 'primordial' in the sense of being the expression of a transhistorical fundamental 'difference' and 'specificity,' these movements are in fact more of the nature of being a symptom of the crisis rather than an answer to it.

See also: Capitalism: Global; Colonization and Colonialism, History of; Environmental History; Imperialism, History of; Imperialism: Political Aspects; Nationalism: General; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: East Asia; Nationalism, Historical Aspects of: South Asia; Nation-states, Nationalism, and Gender; Social Movements, History of: General; Third World

Bibliography

Albertini M 1978 *L'Etat National*. Fédérop, Paris

Amin S 1976 *Unequal Development*. Monthly Review Press, New York, Chap. 5

Amin S 1994*a* *L'Ethnie à l'Assaut des Nations*. L'Harmattan, Paris

Amin S 1994*b* *Re-reading the Post War Period*. Monthly Review Press, New York, Chap. 5

Amin S 1997 *Capitalism in the Age of Globalization*. Zed Books, London, Chap. 4

Anderson B 1983 *Imagined Communities*. Verso, London

Bauer O 1907 *Die Nationalitäten frage und die Sozialde mobratis*. Vienna

Benot Y 1975 *Les Independences Africaines Ideologie et Realites*. Maspero, Paris

Blunt J M 1987 *The National Question*, 2nd edn. Zed Books, London

Bottomore T, Goode P 1978 *Austro Marxism*. Clarendon Press, Oxford, UK

Cohen M 1994 *Ethniciti Politique*. L'Harmattan, Paris

Davis H B 1967 *Nationalism and Socialism*. Monthly Review Press, New York

Davis H B 1978 *Toward a Marxist Theory of Nationalism*. Monthly Review Press, New York

Gibson R 1972 *African Liberation Movements*. Oxford University Press, Oxford, UK

Haupt G et al. (ed.) 1974 *Les Marxistes et la Question Nationale 1848–1914: Etudes et Textes*. F. Maspero, Paris

Hobsbawn E J 1990 *Nations and Nationalism since 1780*. Cambridge University Press, Cambridge, UK

Lenin V I 1970 *Imperialism, the Highest Stage of Capitalism*. Progress Publishers, Moscow

Riazanov 1923 *Le Komintern et l'Orient*

Strasser J, Pannekock A 1974 *Nation et Lutte de Classe*. Union Generale d'Editions, Paris

Wallerstein I 1974 *The Modern World Systems*. Academic Press, New York, 3 Vols.

Wallerstein I 2000 *The Essential Wallerstein*. New Press, New York

S. Amin

National Security Studies and War Potential of Nations

Klaus Knorr wrote *The War Potential of Nations* in 1956. This book marks a key transition between different approaches toward the study of national security. Through much of the nineteenth and into the first half of the twentieth century, the security of a nation depended heavily on the balance between its own war potential and that of its adversaries and allies. In the second half of the twentieth century that

relationship weakened in the context of cold war bipolarity and a massive nuclear standoff between superpowers. The relationship is changing again in fundamental ways at the start of the twenty-first century. The war potential of nations surely will continue to affect national security, but it will not nearly determine it. More importantly, almost all the terms in that statement are coming to mean very different things than they did in 1956.

1. Historical Perspectives

In the second half of the nineteenth century, 'war potential' was a narrow conceptual proxy for 'measuring' the military power of nations. The foundations of war potential were the economic and technological resources that could be mobilized for military purposes. This made good analytic sense in an era of industrial warfare—when nations maintained in peacetime a small fraction of the military capabilities that they could mobilize for war by simply redirecting production. World Wars I and II were won by nations with greater war potential, not by those with the advantage of existing forces at the start of the war.

Knorr understood well the promise—both analytic and practical—of being able to measure this potential. If national security were some function of objective measures of power, then parsing out the problem (at least) would be made a much simpler task for scholars and policy makers. He distinguished among three critical components of war potential—economic and technological capacity, administrative skill, and the political foundations of military power. The idea was to calculate war potential: *if* each component could be measured with reasonable precision *and* some arguments about their interactions could be built.

Knorr assessed this enterprise carefully enough to see clearly its inherent limitations. Except in the very few cases where nations use sheer force simply to seize something, the purpose of deploying military power was to influence the behavior of another actor. In all cases of influence there clearly were objective components to power. But these were always and everywhere put into play in the context of relationships or anticipated relationships that had many other components of power. Without using formal models or even 'soft' game-theory metaphors, Knorr wrote in rather precise analytic terms about the importance of information assymetries, domestic political constraints, reputational effects, and the like in relationships of mutual coercion. He argued strongly that net assessment (the detailed counting and analysis of military forces) was by itself a very weak tool. Without reference to particular objectives and contingencies, to study and measure the war potential of nations would be a fragile approach to thinking about and planning for national security.

These arguments today seem obvious, almost trivial. What was less obvious to Knorr, and remains more difficult today, is how to move the intellectual and policy debate forward. The consolidation of a bipolar power distribution during the middle years of the Cold War, along with the deployment of massive nuclear arsenals that essentially overwhelmed any realistic measures of 'war potential,' made that problem more immediate.

2. Twentieth-century Strategy

Thomas Schelling's 1960 work *The Strategy of Conflict* proposed one kind of solution. The new studies of strategy focused much less on the war potential of nations and much more on the war potential of *situations*. It did this by abstracting away from precise measures of capabilities and even more importantly from precise understandings of the intentions and objectives of an adversary—a major analytic move during a period of time when American policy makers (at least) felt they had a poor understanding of the USSR.

At a time when net assessment was becoming intellectually stagnant (and arguably boring), the turn to 'strategic studies' opened up a new and extremely interesting intellectual agenda for national security scholars to explore. It also pulled into the picture a new set of tools that were developing rapidly in mathematics and economics—in particular, mixed motive game-theory models. The combination of newly configured research questions and fast-improving tools led to an explosion of work that clarified powerfully the strategic logic of situations like nuclear deterrence, brinkmanship crises, and coercive diplomacy.

This knowledge came at a cost, of course. The power of abstract-deductive theory to produce generalizable arguments rests in part on a set of assumptions about the characteristics of the actors or players in the 'game.' It has always been a central disagreement among scholars whether (more precisely, when and under what conditions) the benefits of this approach are worth the cost.

The end of the Cold War sharpens both the intellectual and practical consequences of that disagreement. Bipolarity was a significant clarifying force in international politics—while the Cold War was never as 'simple' as some nostalgic views would have it, it did create a marked explicitness about the major lines of conflict that the twenty-first century so far lacks. The national security environment for many nations, and surely for the USA, is now much more fine-grained than it was in 1965.

3. Future Directions

The problem with flattening out the past is that it makes it harder to see the future in multiple dimensions without becoming frightened and confused. It may be

that the USA has enjoyed a post-Cold War honeymoon since 1990 and that the future will be increasingly challenging and unstable. Of course, there are periods of modern history where linear extrapolations of contemporary trends would have worked tolerably well as a foundation for planning the future—for example, between 1950 and 1975. There are other periods—for example 1925 to 1950, or 1975 to 2000—where discontinuous changes in world politics, technology, social forces, and the economy radically displaced what were assumed to be core 'realities' of security planning. The period around the start of the twentieth century was such a period. Between 1895 and 1920, applications of the internal combustion engine, electricity, and the airplane—along with the rise of the German nation, social Darwinist ideologies, and decolonization—recast fundamental features of the security environment for nations.

Such periods present two distinct kinds of challenges: to master the technology, and to understand the social, economic, and military implications and possibilities that come with it. National security planners recognize the new uncertainties of threats that are smaller, more diffuse, and likely coming from many different directions. From a theoretical perspective, the strategic approach (to build increasingly sophisticated, generalized models of situations) will need at a minimum to be supplemented with renewed attention to the actors and players in the game.

4. *Questions of National Security*

The historic literature on war potential of nations is a reasonable place to start. It forces certain questions: What is war going to be? What potentials are relevant? And are nations still the appropriate central focus for the investigation? The answers to these questions were never clean and simple, and they are becoming far more complex in ways that matter deeply for thinking about national security.

4.1 *The Meaning of 'Security'*

One important complicating factor is simply that security has come to mean much more to many people than the core issues of territorial integrity and political autonomy. The discourse of international politics (both popular and scholarly) now embraces concepts like economic security and environmental security. There may be instrumental political reasons to import metaphors from the security debate into these other important areas of national interest, but there are intellectual costs as well.

The broadening of the realm of 'security' makes thinking and theorizing about it as a dependent variable of anything a much more complicated task. The war potential of other nations would play a small (although not insignificant) part in an analysis of US environmental security. Environmental interests are not necessarily any less important than security interests. But the clarity of focusing on security in a narrow sense as territorial integrity and political autonomy is diluted by stretching the definition. Realists argue that security is a prerequisite for almost anything else that nations want to achieve in world politics. Nonrealists may disagree with the vehement emphasis on security. But few doubt that military power is still a critical restraint on the freedom of nations to act internationally, and particularly to act in ways that threaten the vital interests of the possessor of military power.

4.2 *Globalization and Mobility*

A second complicating factor comes from a set of transformations that fall under the rough heading of 'globalization.' More precisely, the ability of political units to produce and deploy organized violence for political ends (an updated notion of war potential) is being deeply affected by changes in the global economic environment.

The simplest way to think about globalization is to see it as a story about the causes and consequences of an increase in *mobility*. For most of human history, neither goods, nor capital, nor most people, nor many ideas moved very far from their place of origin. Over the last several hundred years, mobility of many (although not all) things has increased—unevenly across time and with some setbacks.

Political economists recognize that mobility makes possible the expansion of markets beyond the physical and other borders that previously contained them. There is simply no serious theoretical framework in which security is unaffected by this change. The war potential of nations effectively stretches beyond national borders in a profound way. This is made more conspicuous by the 'spin-on' phenomenon. The commercial sector now leads the military sector in most key technologies that underpin military power. And the logic of the commercial sector is decidedly not national. The so-called revolution in military affairs (RMA), and particularly the increasing information intensity of military operations (which is part of that revolution), cuts in the same direction. Information is inherently more mobile than tanks, airplanes, or almost any physical asset. Of course, it is also harder to secure in most settings than an aircraft carrier or a nuclear missile buried deep underground.

The most significant exception to increased mobility is people. People do not move across borders as readily as other technological and economic resources. This may be a key national factor of power that nations still control or contain—although it is not clear for how long that will remain the case. In any case, national security concerns about the increased

mobility of people may complicate economic globalization as well as vice versa. Technology can be anywhere, but people still have to be somewhere, and where they are may matter a great deal for individual nation states. The current US lead in information technology owes a great deal to Indian, Taiwanese, and Israeli engineers and entrepreneurs who have immigrated to the USA (or come temporarily on H1B visas) and maintain strong network links to evolving technology clusters in their home nations. At some point nations will experience in a much more acute way the tension between the mobility of people who carry key technological talent and the 'national' security implications of allowing that talent to cross borders freely.

In addition to mobility (which has increased markedly in the past, most dramatically during an earlier period of 'globalization' at the end of the nineteenth century), the next phase of globalization is also about a dramatic increase in *speed*. The digital revolution is reducing the cost of moving a piece of information around the world in real time to near zero. Big, heavy things move much faster as well because they can be moved more intelligently with the use of information. This is unsettling to humans, who think and act slowly. It is even more unsettling to political organizations, because aggregating a bunch of slow humans into a polity usually makes them even slower. An increasing number of those polities now have some form of democratic decision-making system which has the effect (indeed, is designed explicitly in some ways) to slow things down even more.

This disconnect in speed will have profound implications for national security. There is a straightforward and narrow version of the argument. If wars will be faster—that is, time limited by military technology and by the constraints of democratic control—then existing forces may once again become a more important determinant of who wins than 'war potential' in a deeper sense. Yet there are inherent strains—and not just fiscal ones—for democracies maintaining large existing forces during peacetime.

There is a broader version of the argument as well. Balances of power have been an important stabilizing factor in international politics because balancing behavior has prevented the consolidation of hegemony by an aggressor nation. But balances do not form quickly. Napoleon nearly conquered the European landmass before an effective balance formed against him. The Grand Alliance did not fully come together to oppose Hitler until 1942—quite late in an offensive that actually began almost a decade before. When competing systems move at increasingly asymetric speeds the results are likely to be unstable and surprising. Balance of power both as theory and practice has been an important part of the relationship between national security and the war potential of nations. Yet balances may become less stable and more volatile in the future than nations expect.

4.3 *Size of the War Machine*

A third complicating factor, loosely connected to globalization arguments, is the issue of *size*. The tools of large-scale organized violence are getting smaller. Nuclear weapons were the first modern step in this process, but they were hard to engineer and build, even for large nations. Biological weapons are different: a biological agent can be microscopic. More importantly, the equipment to produce custom-designed virulent biological agents will soon be available in suitcase size for US$10,000 or less. An effective computer virus may be even easier and cheaper to manufacture. In more than 50 years no subnational group has yet gained possession of and 'used' a nuclear weapon (including in a bargaining situation) for political purposes. Almost certainly it will not be possible to say the same of the next generation of weapons of mass destruction or disruption 50 or far fewer years hence.

5. *Conclusions*

Perhaps this kind of capability does not represent war potential in a traditional sense. One conclusion to draw from this is that it makes sense fundamentally to rethink the term itself, as well as the range of actors who can possess war potential. Changes in mobility, speed, and size have dramatically increased the porosity of national borders. Given that porosity, behavior that in the past might have been thought of as terrorism could increasingly become much more like war, in the sense that the goals of the behavior might not simply be revenge or the placing of pressure on a particular, discrete policy. The goal—realistically—might be to undermine the foundations of a nation or bring down a government. In that scenario, each important term in the phrase 'national security and the war potential of nations' will have taken on a very different meaning.

See also: Cold War, The; Defense, Economics of; Geopolitics; Globalization: Political Aspects; International Relations:Theories; Military and Politics; Military Geography; Military History; Military Psychology: United States; Military Sociology; War: Causes and Patterns; Warfare in History

Bibliography

Art R J, Waltz K N 1999 *The Use of Force: Military Power and International Politics*, 5th edn. Rowman and Littlefield, Lanham, MD

Bull H 1977 *The Anarchical Society: A Study of Order in World Politics*. Columbia University Press, New York

Katzenstein P J, Social Science Research Council (US) Committee on International Peace and Security 1996 *The Culture*

of National Security: Norms and Identity in World Politics, New Directions in World Politics. Columbia University Press, New York
Kennedy P M 1987 *The Rise and Fall of the Great Powers*, 1st edn. Random House, New York
Knorr K 1956 *The War Potential of Nations*. Princeton University Press, Princeton, NJ
O'Rourke K H, Williamson J G 1999 *Globalization and History: The Evolution of a Nineteenth-century Atlantic Economy*. MIT Press, Cambridge, MA
Ruggie J G 1996 *Winning the Peace: America and World Order in the New Era*. Columbia University Press, New York
Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA
Walt S M 1999 Rigor or rigor mortis? Rational choice and security studies. *International Security* **23**(4): 49–80
Waltz K N 1964 *Man, the State, and War: A Theoretical Analysis*. Columbia University Press, New York
Weber S 2001 *Globalization and the European Political Economy*. Columbia University Press, New York

S. Weber

National Socialism and Fascism

Italian fascism and National Socialism signify a new type of political mobilization which came into being after the end of World War I. Both movements were ideologically influenced by integral nationalism and racism, but drew their specific political impetus by exploiting the widespread anticommunism and anti-socialism among the bourgeois strata of postwar society. Italian fascism in particular was taken as a model by comparable movements all over Europe, but, with the exception of that in Spain, only Italian fascism and National Socialism were able to establish a dictatorship.

1. *Italian Fascism*

1.1 *Origins*

The origins of Italian fascism are closely connected with the extreme legitimacy crisis of the liberal regime facing protest strikes and factory occupations by rural and industrial workers. Hence, Giovanni Giolitti tolerated the use of violence against organized labour by the fascist *Squadre d'azione*, which resulted in the emergence of a dualistic power structure, reflected in Gabriele d'Annuncio's expedition to Fiume in 1919 and the actionism of the Black Shirts who terrorized the countryside and socialist municipalities. Mussolini used the opportunity to create the veterans' organization of the *Fascio di Combattimento* and to shape it into a non-partisan mass movement which he reorganized in November 1921 as *Partido Nazionale Fascista* (PNF). By holding out the threat of the revolutionary potential of quadrism and unleashing the march on Rome on October 27, 1922, Mussolini became prime minister in a right-liberal coalition cabinet, although the PNF controlled only a minority of seats in the parliament.

1.2 *Fascist Dictatorship*

Mussolini used his new position to establish his personal dictatorship by skilfully extending the prerogatives of the prime minister through legal means. By an ostensibly manipulative reform of the electoral system he attained an overwhelming parliamentary majority for the governing coalition in the 1924 elections. By a series of 'fascist decrees' the civil liberties, the freedom of the press, and the space of action for the parliamentary opposition which formed the antifascist Aventin were severely curtailed. The ensuing Matteotti crisis provided a pretext to dissolve the opposing parties and to promulgate a one-party state relying on the extraparliamentary power of the Fascist Great Council. Mussolini thus managed to maneuver both the cabinet and the parliament—which in 1938 was reshaped on a corporatist basis—into the background. The introduction of the fascist calendar and the adoption of the fascist emblems as state symbols completed the creeping process of shaping a fascist state. Mussolini achieved this by carefully evading open clashes with the constitution and achieving a *modus vivendi* with the conservative elite and big business.

Mussolini's tactical flexibility climaxed in concluding the Lateran treaties which, by comprehensive concessions in social and cultural matters, secured the support of the Catholic church. The combination of social propaganda and compulsion, as reflected in the program of '*dopo lavoro*,' together with co-operation with big business, enabled him to neutralize the syndicalistic trade unions. The PNF, which in 1943 comprised 4.75 million members combined heterogeneous social and political groups and exhibited strong inter-regional tensions. By isolating the party radicals under Roberto Farinacci, who endorsed the principle of forming a cadre party, the PFN became more or less bureaucratized and was controlled through local and regional administration. Thus, despite terrorist police and the abolition of civil rights by the *lege fascissimi* in 1925/26 the system was not completely totalitarian, particularly in comparison to the Nazi regime. The monarchy, the army and certainly the Catholic church preserved a certain independence, and the fascistization of the administrative apparatus was only partially attained. Mussolini's rule relied primarily on a continuous organization of consensus—his domestic policy lacked any clear direction and ideological consistency. The myth of the indispensable role of the Duce for Italian national survival kept everything together, but the disastrous consequences of Mussolini's alliance with

Hitler raised resistance among not only the conservative elite and the military, but also his own chieftains, who eventually used the instrument of the Great Council to depose him in July 1943.

1.3 The Alliance with Hitler

Mussolini's ambition to establish a Mediterranean empire, as manifested in the Abyssinian and the Ethopian campaigns, compromised Italy's former relations with Great Britain and France and, in 1936, implied a movement towards the Third Reich. Through the promulgation of the Axis and, in 1937, by Italy joining the Anti-Comintern-Pact, Mussolini, who over-estimated Italian military and economic capabilities, became completely dependent upon Hitler (and definitively so by signing the 'Steel Pact' in 1939). His decision, in June 1940, to enter the campaign against France, while waging parallel wars against Albania and Greece and starting an offensive in North Africa overstrained Italian resources and made the German intervention unavoidable. Despite these military disasters, Mussolini insisted in participating actively in the Soviet campaign. Under German pressure, he initiated racial legislation and eventually acquiesced in the deportation of Italian Jews to the annihilation camps. His fascist regime in the Salo Republic (formerly Repubblica Sociale Italiana) depended completely on German military protection and was marked by an escalating use of terror. In some respects it reverted to the radical origins of the fascist movement, before it fell asunder on April 28, 1945.

2. National Socialism

2.1 Origins

The origin of the NSDAP (National Socialist German Workers Party) dates back to the period immediately after World War I, and to the Soviet Republic in Munich in April 1919. The DAP was founded by Anton Drexler and Karl Harrer who came from the *völkisch* 'Thule Society,' a clandestine affiliation of the Pangerman League which, at the same time, formed the German Völkish Protection League (*Deutsch-völkischer Schutz- und Trutzbund*) in order to mobilize racial antisemitism against the Socialist left.

Adolf Hitler acting, until the Fall of 1919, as a political informant of the Bavarian Reichswehr, shaped the early Nazi party—hitherto an insignificant splinter group among a myriad of similar *völkish* associations—into a fascist party by enforcing the leadership principle and streamlining the party's propaganda. Hitler achieved unrestricted leadership in 1921 and subsequently eliminated intraparty elections and debates, focusing the party's activity solely on propaganda which exploited the public phobia against the Versailles peace treaty as well as the growing resentments of parts of the constituency against the Weimar Republic.

With its stronghold in Bavaria, the Nazi party remained dependent upon right-wing conservative groups, which prepared a march on Berlin following the model of Benito Mussolini in October 1923, with the aim of establishing a national dictatorship under Erich von Ludendorff, the former war hero and leader of the *voelkish* wing of the German Nationalists. Hitler, who had gained the nominal political leadership of the right-wing opposition and its paramilitary units, ultimately decided to act independently from his conservative partners, but his attempt to proclaim his German dictatorship and to gain the approval of the national camp through his propaganda march to the Feldherrnhalle on November 9, 1923 failed. Neither the prohibition of the party and the SA (Stormtroopers), nor Hitler's indictment for high treason and imprisonment in the Landsberg penitentiary until December 1924 did anything to quieten the activities of his followers.

After his release, Hitler, faced with competition from successor organizations, decided to dissolve the party and to reinvent it on the basis of the unrestricted leadership principle, excluding any intraparty elections and restricting its activities exclusively to unrelenting organization and agitation. Hitler succeeded, in 1926, in neutralizing the opposition of the North-Western group of the *Gauleiter* (district leaders), which was controlled by Gregor Straßer, and in preventing any alteration of the original party program of 1920, although it did not comply with the requirements of the stabilization period and did not drop its pseudo-socialist ingredients. The left wing led by Otto Straßer therefore left the party in 1931.

Despite its enormous propaganda efforts, until 1928 the NSDAP remained an insignificant splinter group without any real hold on the industrial working class until 1928. Not before the spring of 1929, in conjunction with the accelerating erosion of electoral support for the bourgeois middle parties, did the NSDAP become a serious factor in German politics, especially after its breakthrough in the September 1930 elections. The sudden rise of the party into a fully fledged mass movement depended on the mobilization of divergent elements of the electorate. These comprised at first mainly protest voters, pensioners and peasants, then especially new or previous non-voters, but by 1932 the NSDAP was an omnibus party now also drawing support from upper middle strata. The high turnover of members and voters made it vulnerable in the event of a political stabilization.

2.2 The Seizure of Power

The success of the Nazi party at the polls appeared spectacular, but at its peak in July 1932 it attracted no

more than 37.3 percent of the vote and, in November, suffered a visible setback while, in the March 5, 1933 elections, it obtained a slight majority only combined with the conservative German Nationalist Party. Thus, Hitler's nomination as Reich chancellor in the 'Cabinet of National Concentration' on January 30, 1933 by Reich president Paul von Hindenburg was far from being the outcome of the democratic process, being due to informal pressures exerted by conservative interest groups and the army leadership on the Reich president whose spokesman, Franz von Papen, believed that he could control Hitler within the Cabinet by a majority of conservatives. Papen convinced the reluctant president that the Hitler cabinet was virtually a majority government pretending that the catholic Centre Party would eventually join the coalition.

Hitler by scheduling premature elections used his majority to get rid of the bourgeois encirclement, compelling his conservative partners to accept the Enabling Law of March 23, 1933. This law, together with the emergency powers of the Decree for the Protection of People and State issued on February 28 after the burning of the Reichstag, invalidated civil liberties and legalized sanctions against the socialist and communist opposition, thereby paving the way for unrestricted dictatorship.

In the first phase of his chancellery, Hitler was preoccupied with the consolidation of his power through a systematic policy of *Gleichschaltung* (coordination) of public and private institutions: subduing them to party control and placing leading Nazis in high administrative positions without changing the governmental apparatus and its conservative persuasion—except for the instalment of Joseph Goebbels as Minister for Public Enlightenment and Propaganda. The radical wing of the party, especially the huge paramilitary organization of the SA under Ernst Röhm, who called for a 'second revolution' and demanded to interfere in the economy as well as taking control of the armed forces, were neutralized by the events of June 30, 1934, when Hitler under the pressure of the SS, Hermann Göring and the army command ordered the immediate liquidation of the Supreme SA leaders. This action led his conservative partners to believe that he was basically a moderate and would put an end to the party's influence on the political system.

2.3 The Nazi State

The process of consolidation of power came to an end with Hitler's decision to combine the presidency and the chancellery after Hindenburg died in August 1934. This was particularly significant because Hitler thus became the official leader of the armed forces which, in order to prove their loyalty, swore an oath of personal allegiance to the Fuehrer. At the same time, Hitler ceased to attend the Reich cabinet and eventually prohibited cabinet meetings altogether, thereby depriving the Reich of its old-established governing body, without replacing it by comparable advisory boards on the party level. In fact, he showed less and less interest in day-to-day politics and refrained even from interfering in internal feuds so long as his personal prestige was not affected by them. Instead of using governmental institutions, Hitler preferred to rule on the informal basis of oral information supplied by various—mostly irresponsible—advisers, whereas the Reich chancellery, the task of which was above all to coordinate the department, was circumvented by ad hoc agencies established interposed between party and state—such as the SS empire, the German Labour Front and Hermann Göring's Four-year Plan Office.

Contrary to Goebbels' propaganda slogans, the Nazi system of government was not characterized by any 'unity of will' or clear-cut subordination patterns, but by the principle of far reaching freedom of action for individual office-holders as long as they did not interfere with Hitler's personal prestige. This system of 'government without administration' instigated a quasi-socio-Darwinist struggle between competing agencies and never-ending personal rivalries among the chieftains, a feature which during the combat period had intensified the party's dynamism. In the regime phase, where mere mobilization was no longer sufficient, this system turned out to be counterproductive and utterly detrimental for any rational governmental process, creating a process of cumulative radicalization. In the long run, the system was destined to bring about the complete dissolution of any regular decision-making process and an escalating tendency to self-destruction, which accelerated particularly after 1943 under the pretext of implementing the party revolution that had been erroneously stopped in 1933/34 for tactical reasons.

2.4 The Politics of Military Aggression and Racial Annihilation

Analogous to his domestic policy, Hitler dealt with foreign affairs by adapting a 'trial and error' method. Initially he pursued the path of a revision of the Versailles treaty, although being resolved from the start to establish a German hegemony over Europe beyond the territorial limitations of Versailles and to conquer 'living space' (*Lebensraum*) in Eastern Europe. His first steps, however, consisted in breaking the diplomatic isolation of the German Reich which had been increased by her retreat from the Geneva disarmament conference and the League of Nations in October 1933. By signing the German–Polish non-aggression treaty which ran counter to previous conservative foreign policy, and by concluding a naval agreement with Great Britain, Nazi foreign policy

succeeded in undermining the Stresa front. Despite conflicting interests with respect to Austria's independence, Mussolini entered the Axis alliance in 1936 and joined the Anti-Comintern Pact thereafter.

The domestic corollary of the transition to active foreign policy lay in stepping up the clandestine rearmament which was already on the way before the seizure of power. Financial and economic policies both gave overall priority to rearmament, culminating in 1936, when the four-year plan required the German army to be capable of waging war within a period of four to six years. The remilitarization of the Rhineland as well as the introduction of conscription in 1935 were also accepted by the Western powers, despite the fact that they obviously violated the demilitarization provisions of Versailles.

The Western powers still believed it was possible to reach an agreement with Hitler and tame his territorial ambitions, even after having acceded to the annexation of the Austrian Republic in March 1938. Not November 1937, but the spring of 1938 became the climacteric of the progress towards open aggression. Whereas Hitler had played a rather passive role in the Austrian crisis (which had been pushed ahead by Hermann Göring), he was resolved to settle the Sudeten conflict by military means, and his decision to take over the war ministry himself enabled him to act. Neville Chamberlain's decision to continue the policy of appeasement and Mussolini's mediation forced the dictator to reluctantly accept the provisions of the Munich conference in September 1938, however, any expectation that the cession of large parts of Czechoslovakia would curb Hitler's aggressive appetite proved to be utterly wrong—as was proved by the conquest of independent Czechoslovakia and the establishment of the Reich protectorate Bohemia–Moravia in March 1939.

2.5 *World War II*

Historians disagree as to whether the British guarantee of Polish independence was an adequate response to Hitler's expansionist policy because it reduced the already limited readiness of the Polish government to join negotiations over the 'Corridor' and Danzig. At any rate, it confirmed Hitler in his conviction that he had to extinguish Poland first before a military conflict with the Western powers could be envisaged. Backed by the German–Soviet non-aggression pact, he gambled on the likelihood that Britain and France would disregard their obligations toward the Polish Republic and eventually took the risk of a war on two fronts—something he had always vowed to avoid.

Despite the tremendous military success of the German armed forces—especially the occupation of the Scandinavian states and the unexpectedly rapid victory over France—the long-range perspective was more than uncertain. After Winston Churchill had rejected Hitler's allegedly generous peace offer and Germany had lost the air battle over the English Channel, her strategic situation could not but deteriorate despite short-term military successes all over Europe and North Africa. Hitler tried to overcome the strategic stalemate by attacking the Soviet Union in June 1941, thereby overextending the German military resources and manpower. The defeat of the German army before Moscow destroyed any expectation of an imminent collapse of the Soviet Union; and in other theatres of war, especially in North Africa, Greece, and Yugoslavia, the Germans were forced onto the defensive.

2.6 *Racial Annihilation*

Jews were the foremost target of Nazi violence, especially after the abolition of organized labour and the replacement of the trade unions by the German Labour Front. Between the boycott of April 1 to the Nuremberg Laws in September 1935, and the pogrom of November 1938, antisemitic hardliners pressed for the disappearance of the Jews from Germany; but, despite the expropriation, social segregation, and lawlessness of the Jews, the number of Jewish emigrants remained insignificant compared with the assimilation of the Austrian, Czech and Polish Jewry into the German power bloc. Thus, enforced emigration of Jews reached an impasse in 1939, although emigration continued to be fostered by the regime until October 1941. The Polish and Russian campaigns gave rise to plans for establishing Jewish reservations, starting with the Nisco project and followed by the Madagascar plan, but all these schemes came to nothing because of opposing domestic and strategic interests or the ongoing war.

Hitler was resolved to wage the campaign against the Soviet Union as a war of unrestricted racial annihilation. A war which would lead not only to the defeat of the Bolshevik regime, but also to the complete destruction of Russian statehood and to a systematic ethnic cleansing of wide parts of the Soviet territory in order to provide living-space for German and Germanic settlers—a goal explicitly pursued by Heinrich Himmler's 'General Plan East.' In conjunction with this, the deportation and elimination of the indigenous Jewish population was put in motion on the basis of a close cooperation between special mobile killing units, established by the SS and the Wehrmacht in the occupied territories. The killing instructions of the task units originally embraced only Soviet functionaries and Jews in leading positions, but were subsequently extended to the murder of whole Jewish populations. The majority of indigenous Jews had already been liquidated by October 1941, when the establishment of annihilation camps and the use of gas

vans created the foundations for systematic implementation of the 'Final Solution,' including Jews in the General Gouvernement and other parts of German-dominated Europe.

The implementation of the Holocaust, that was argued upon at the Wanusee Conference on January 20, 1942, resulted not so much from a premeditated concept, but from an interaction between proceedings on the local level and the Reich Main Security Office. There was no need for any formal order by the dictator whose part consisted mainly in instigating radicalization, without being directly involved in the concrete destruction measures for which the close cooperation of the SS, the Wehrmacht and civil administration was indispensable and which caused the murder of at least 5.5 million Jews. Official secrecy regarding the killing operations did not prevent news about the atrocities being widely distributed, although a concrete picture of the systematic liquidation was not attained before the war ended.

2.7 The Dissolution of the Third Reich

The capitulation of the 6th army in Stalingrad on January 31, 1943 marks the definite turn of tide against the Axis powers. Because of its lack of manpower, ammunition and technical equipment the Wehrmacht did not regain the strength necessary to continue strategic operations, and was forced onto the defensive. The breakthrough of the Red army at the middle sector of the Eastern front in the summer of 1944 opened its road into the Reich despite unremitting German attempts to stabilize the front. After the Allied invasion of France and the Allied air offensives the military situation became hopeless, and it was just a question of time until the Reich was defeated.

The artificial prolongation of the fight relied upon the total suppression of domestic opposition, since the army had lost its relative autonomy after the abortive putsch of July 20, 1944. Unlike in Italy, there was no institution left which could take an initiative. The renewed efforts of fanatical party functionaries to take over public administration and to urge for fighting to the last man by means of a progressively tenuous hold over propaganda, and a rule of terror by martial law, signified the last phase of the war—during which Hitler and his chieftains in Berlin had clearly lost any grasp on reality. The end of the Third Reich comprised not only total military defeat, but also the complete collapse of its institutional framework.

3. European Fascist Movements

Fascist movements originated in most parts of Europe during the interwar period, but only in exceptional cases were they able to exert any great political influence or to establish—as in Italy and Germany—a regime of their own. Most of these fascist parties never developed into mass movements, especially within the rural societies of Southeastern and Eastern Europe. Only the Romanian 'Iron Guard' and the Hungarian Nemzeti Akarat Partja ('Arrows of the Cross') achieved temporarily significant popular support comparable to the PFN. In two cases, fascists were able to rule for a limited period: the Ustasha under Ante Pavelic in Croatia in 1941; and the Arrows of the Cross under Ferenc Salazi in Hungary in 1944—both relying on the German occupants. The Hlinka Guards in Slovakia were amalgamated in the authoritarian Rassemblement of Josef Tiso and lost their independence. In Austria competing fascist movements—National Socialists and the Heimwehr (Home Guard)—attracted considerable portions of the electorate, but were suppressed (like the National Socialists) or became absorbed by the state-controlled Vaterländische Front (Patriotic Front).

In Britain, France, Belgium, and the Netherlands, as well as in Norway, fascist groups played an insignificant role. In Spain, the influential fascist Falange was tamed by the authoritarian rule of General Franco and lost its original impetus. Similar developments prevailed in Southeastern Europe where parliamentary democracy was replaced by authoritarian regimes with varying populist leanings. The majority of the fascist movements remained splinter parties and stood in opposition not only to the left, but also to the ruling conservative forces and could share power only with the support of the German occupants. Those fascist groups that sympathized with National Socialism—such as the Partie Populaire Française of Jacque d'Doriot in France, the Front Populaire des Rex of Léon Degrelle in Belgium, the National-Socialistische Beweging of Adriaan Anton Mussert in the Netherlands, and the National Samling of Vidkun Quisling in Norway—exerted only marginal influence, and this was even more true for Oswald Mosley's British Union of Fascists.

3.1 Structural Elements

Italian Fascism provided the model for similar movements, especially for National Socialism which copied techniques, symbols and political strategies from its Italian counterpart. Although some historians reject the use of a general typology of fascism including National Socialism, the similarities between them cannot be disregarded, although the extremely destructive and inhuman character of the Nazi experiment appears to be exceptional. The varieties of fascism are characterized by their rigorous antiliberal and anticommunist stance as well as by their extreme nationalism and racism, their specific populist ideology, and voluntarist and decisionist political style. They can be distinguished from right-wing bourgeois parties by their techniques of exploiting resentments and prejudices of the electorate, especially nationalism

and antisemitism or—as in the case of Italy—racism, for mass mobilization, and of avoiding any choice between contradictory programmatic targets, thus replacing actual political issues by purely propagandistic slogans, although many of their goals are similar to those of the bourgeois right.

The 'cult of the will'—the arbitrary pursuit of utopian visions first articulated by Georges Sorel—and the myth of the unity of the people ('people's community') are specifics for the fascist variety of politics, that added a new element to the political process. Whereas the antiliberal and anti-enlightenment impetus is a general criterion, there exist differences with regard to the degree of parasitic disruption of the inherited state apparatus which culminated in the National Socialist regime in Germany, caused by their technique of mobilizing non-synchronous popular resentments and prejudices.

Fascist movements tend to replace liberal organizational principles by means of amorphous personal legacies culminating in the leadership cult. Mass mobilization is regarded by them as an end in itself, even at the price of a chameleon-like propaganda. As genuine protest movements, their membership and constituency are characterized by an over-representation of youth, of marginal social strata, and of social outcasts. Frequently modernization crises provide a fertile ground for emergent fascist movements. They are specifically a phenomenon of the inter-war period, although a series of successor organizations did emerge after 1945, such as Peronism in Argentina.

See also: Authoritarian Personality: History of the Concept; Authoritarianism; Contemporary History; Dictatorship in History; Ethnic Cleansing, History of; First World War, The; Holocaust, The; Nationalism, Historical Aspects of: The West; Parties/Movements: Extreme Right; Second World War, The; Totalitarianism; Totalitarianism: Impact on Social Thought

Bibliography

Bach M 1990 *Die charismatischen Führerdiktaturen. Drittes Reich und Italienischer Faschismus im Vergleich ihrer Herrschaftsstrukturen*. Baden-Baden, Germany

Bessel R (ed.) 1996 *Fascist Italy and Nazi Germany: Comparisons and Contrasts*. Cambridge, UK

De Felice R 1972 *Le interpretazioni des fascismo*. Bari, Italy

Dipper C 1998 *Faschismus und Faschismen im Vergleich. Wolfgang Schieder zum 60. Geburtstag*. Greifswald, Germany

Kershaw I 1985 *The Nazi Dictatorship. Problems and Perspectives of Interpretation*. Arnold, London

Laqueur W Z 1991 *Fascism. A Reader's Guide, Analyses, Interpretations, Bibliography*. Aldershot, UK

Mantelli B 1994 *Il nascita des fascismo*. Milan, Italy

Nolte E 1968 *Die Krise des liberalen Systems und die faschistischen Bewegungen*. Piper, Munich, Germany

Payne S G 1995 *A History of Fascism, 1914–1945*. University of Wisconsin Press, Madison, WI

Rees P 1984 *Fascism and Pre-Fascism in Europe, 1890–1945. A Bibliography of the Extreme Right*. Harvester, Brighton, UK

Rogger H, Weber E (eds.) 1964 *The European Right. A Historical Profile*. Stanford, CA

Schieder W 1983 *Faschismus als soziale Bewegung. Deutschland und Italien im Vergleich*. Göttingen, Germany

Weber E 1964 *Varieties of Fascism. Doctrine of Revolution in the 20th Century*. London

Wippermann W 1983 *Europäische Faschismus im Vergleich 1922–1982*. Frankfurt, Germany

Woolf S J 1981 *Fascism in Europe*. London

H. Mommsen

National Traditions in the Social Sciences

To a very considerable extent the social sciences in each country nowadays are a local reproduction of the common global pattern of the social sciences. Further, it is clear that the achievement of a universal social science is a widely shared goal amongst social scientists from all nations, although some pursue more nationalistic agendas. Nevertheless, the national context within which social scientists operate continues to have a considerable influence on the content and form of social science work. Some national contexts have been more important than others in the production of high-quality social science: for example, in their study of nineteenth-century social science innovations, Deutsch et al. (1986) suggest that, while their set of 64 discoveries is equally split between those made in Europe and those made in the USA, there is a marked switch over time, with most of the European contributions being made in the first half of the period (up to World War II), but the period since being overwhelmingly dominated by discoveries advanced by US-based social scientists.

In some periods, for some areas of knowledge, and for some geographical areas the national intellectual context is particularly active in shaping social knowledge production, and it is possible clearly to identify national traditions of social science which share particular features and which may have an accepted label. 'It is a common observation that national intellectual traditions exist in the social sciences. These traditions are often remarkably resistant to processes of "universalisation" or "modernisation" and even provide intellectual resources from which new social theorising can continuously emerge' (Wagner et al. 1991b, p. 341). The imprint of past national traditions can endure. Behind each of the present universalistic social science disciplines the historical imprint of those particular national traditions which were formative in establishing that discipline can often still be glimpsed. Finally, on occasions, a national approach is so strongly and explicitly developed by a more or less

coherent social grouping of social science scholars that we can reasonably talk about a national school, which is often centered on the capital of the nation: for example, the German School of History (founded by Ranke), the French School of Sociology (led by Durkheim), the Stockholm School of Economics, or the Prague School of Linguistics.

Attention in this article is mainly addressed to the rather more central nations such as France, Germany, the UK, and the USA. Little is written about national traditions of social science that might contrast different countries in Latin America, Africa, or Asia. Since national traditions are more often identifiable in the half-century between about 1875 and 1925, or more generally the period bounded by a further 25 years on either side, attention is more directed to this period. A further limitation is that the focus centers only on major social science disciplines.

Acquiring a thorough comparative knowledge of national social science traditions is hampered by several difficulties. Appropriate information is routinely provided by 'country reports,' in which the state of scholarly activity in some particular area of knowledge is described, fructified perhaps with some interesting but unsystematic observations about the linkages between the content described and the national context in which it was produced. Sometimes such country reports are published alongside each other in a 'regional report' or a 'world report.' Less frequently, developments across the range of the social sciences are depicted for a particular country or region. Such literature tends to be largely descriptive, and often entirely speculative. The question of national traditions in social science is more often raised than satisfactorily answered.

Nevertheless, more recently a rather more theoretically informed sociology of social science has begun to generate appropriate studies on the comparative history of the social sciences (e.g. Wagner et al. 1991a, 1991b). This newer approach draws particularly from theoretical advances in the sociology of science, and also from developments in social theory. Rather than presenting a picture of the universally uniform, unilinear, and progressive evolution of each of the social science disciplines, somewhat more nuanced treatments are now more prevalent. These stress ruptures and discontinuities, interlinkages, and oppositions, and variations over time and space.

Levine's study of national traditions in sociology provides a definition that can be more generally applied to other social sciences. He sees traditions as inter- and intragenerational conversations amongst intellectuals in a particular country, which tend to share particular assumptions about social reality. In such conversations writers mainly cite fellow nationals.

> More important, over the generations they reproduce what are palpably national characteristics. Moreover, when they engage in dialogue with parties from other national traditions, they do so, openly or by implication, in a more contrastive mode—as when Durkheim explicitly contrasts his French discipline with British and German traditions (Levine 1995, p. 99).

The rise and fall of national traditions are shaped by various cultural, ideological, political, institutional, cognitive, and social factors, which will be reviewed in the remainder of this article.

The broadest influence is undoubtedly that of national cultural traditions. Cultures which are much more likely to foster the successful development of social science are those stressing the importance, in both the natural and social realms, of the acquisition of explicit, rational scientific principles and of empirical fact-finding, and, even more strategically, the importance of developing systematic ways of interfacing the rational and the empirical. An important influence in each national culture on the development of social science has been those models of natural scientific inquiry advocated in that national culture by philosophers or natural scientists, which have been admired by social scientists.

Aspects of religious thinking in a country also affect the development of social science in that country. Prior to the Reformation and Enlightenment, secular social thinking was often discouraged. Moreover, in Protestant countries, close relations developed between intellectuals and churches. Intellectuals were harnessed in the conflict with Catholic ideas and politics; there was more room for debate since (some) Protestant religions were not anchored by a central dogma; and since Protestant clergymen could raise families, intellectual dynasties could be more readily formed. Thus, in the UK, and also The Netherlands and Scandinavia, scientific innovation was linked with religious debate. However, in France secularization of intellectual culture took place with support from the state and the court, and did not involve the development of scientific thought, since literary genres were dominant. Therefore there was scientization without secularization, in contrast to secularization without scientization (see Heilbron 1995, pp. 63–4). The universities in both Italy and Spain tended to be circumscribed by the Catholicism of their societies, and were fairly moribund intellectually, except for occasional intellectual forays animated by links with Parisian thought.

Each culture tends to make similar ontological assumptions about the nature of social reality, and how it might be known and changed. Such assumptions underpin much social science work in that country. In particular, cultural choices tend to be made between stressing the individual level as ontologically prior (as in the UK), or the collective level (France), and between an objective approach (as in both the UK and France) compared to a more subjective approach (as in Germany).

Levine identifies seven national traditions in sociology, although his characterizations are extendable to the social sciences more generally, as follows:

(a) Aristotle, representing the Hellenic tradition, is concerned with the way in which different constitutional arrangements of societies (that is, city-states) shape the conditions for achieving human virtues, and thus human happiness, and with the achievement of that potential arising from the deliberative judgments of the participants (see *Aristotelian Social Thought*).

(b) The UK tradition emphasizes the naturally sociable disposition of individual humans: for example, the preoccupation with the properties, rights, liberties, and utilities of individual actors, and then the conduct of the actors (see *Utilitarian Social Thought, History of*).

(c) The French tradition emphasizes that 'society formed a supra-personal entity with properties and needs of its own, above all needs for normative regulation and solidaristic integration,' and that this was a natural phenomenon (see *Functionalism, History of*).

(d) The German tradition emphasizes human freedom as an essential element of moral decision making, and also affirms the creative power of language as a distinctive feature of human action and morality.

(e) The transnational Marxian tradition draws on German idealism, French socialism, and English economics to provide an ideologically committed and yet scientific macrosocial science of the development of capitalism and other modes of production (see *Marxist Social Thought*, *History of*).

(f) The Italian tradition seeks laws about societies as natural systems, with the prominent features of these systems being the functioning of ruling classes or elites.

(g) American pragmatism stresses social activism disposed to solve problems on an experimental, commonsense, pragmatic basis (see *Pragmatist Social Thought, History of*).

Different nations each house much the same range of ideological or political perspectives, but some perduring national emphases can be found. For example, French ideology more often stresses radical change, drawing on its rationalist heritage, whereas in the UK and the USA the emphasis is more on reform, flowing from a strain toward empiricism. In Germany ideology is often idealist, humanist, and antipositivist. Such emphases in political ideologies tend to flavor national social science traditions.

Each culture understands the relationships between its component 'social fields' rather differently. Although most modern societies have in common a set of separate economic, political, and social realms (together with many minor arenas of social life), how each is constituted and what its relationships are to other spheres may differ considerably, and this has consequences for social science work in that society. Each discipline tends to have an interactive relationship with one of the social realms in that society, with the social science discourse being in part 'constitutive' of that realm, and in part being shaped by it. The way a society is organized in terms of its social fields may guide the development of particular national traditions of social science over many generations. For example, in France, the state has long been a very central and powerful institution, although there has been a strong discouragement of scholarship in political science. In France, society ('the social') became distinct from church and politics, and then the economy, but in Germany the distinction was resisted. The possibilities for the development of sociology in each country were shaped by these differences.

Cultures can differ in how they set the boundaries of the social sciences. In Germany, *Sozialwissenschaften* is a broad conceptualization, whereas in the Anglo-American world a sharper distinction is usually drawn between the social sciences and the humanities. France is more complex, with economics located within faculties of law, whereas the other human sciences are located within the broader humanistic framework of faculties of letters (see *Universities, in the History of the Social Sciences*). Which particular disciplines are separately identified has also differed, with some continental European university systems often presenting more 'policy-orientated' types of knowledge such as demography, criminology, and public administration. More recently, that division of labor among which has developed in US universities has successfully diffused worldwide, and has become the norm against which the array of disciplines in other countries is measured.

During the period of German university development from the late 1700s through to the mid-1800s literary and artistic intellectuals were not accommodated in universities. One result of this appears to have been that 'the independent non-university intellectuals became hostile to the new professorial form of knowledge production' (Olson 1993, p. 24). This involved a Romanticist rejection of cold hard facts, which developed alongside, and in reaction to, the professionalization of knowledge, and provided an alternative and oppositional stock of intellectual resources, which has most recently been drawn on in the development of postmodernism.

The policy process is different in different states. In the nineteenth-century UK, fact-finding was institutionalized in the role of Royal Commissions, inspectorates, and social-reform research associations, but these were weakly linked (at least as far as formal ties are concerned) with the processes of policy making. However, social network ties amongst various members of the elite could provide channels for information to percolate to policy makers. Links between social researchers and policy makers ever since have remained generally at arm's length. In Germany, there was also considerable state involvement in statistics collection. In contrast, in The Netherlands, a tradition

of longer-term policy making developed, and this more farsighted orientation was able to articulate with a slew of empirically orientated research approaches, such as sociography and social geography, which were particularly well developed there. The national political culture of states leads to differences in their interest in social research, and their capacity to promote it and then to utilize its findings varies considerably (for an earlier review, see Cherns 1974).

Although the attention of historians of social science is particularly drawn to examining theoretical traditions, there are also research traditions (or more broadly methodological traditions) which may not be at all strongly linked with the more theoretical traditions. The state, as the key institution in developing statistical information, is particularly important in shaping the types of empirical research and methodological developments that eventuate in a particular country. But there are other influences. Levine (1995, p. 276) suggests that:

> The empirical traditions also bore the mark of national dispositions, if not in such a pronounced form as the philosophical ones. England led the way with social surveys, systematic investigations of living and working conditions, mainly of members of the working class. France and Italy pursued the collection of national social statistics, work that enabled Durkheim to lead off so impressively with his analysis of divorce rates, educational levels, mental illness data, religious affiliation and the like in *Suicide*. Germany pioneered the experimental manipulation of subjects and also the systematic collection of ethnographic data in broadly defined culture areas. The US pioneered in producing census data and later in systematic gathering of information through personal documents and direct observation as well as interviews.

Later, in the USA, an array of methodological developments occurred. For example, content analysis was developed, especially in the context of World War II and the Cold War, when direct access to totalitarian countries was denied, and so more indirect means of study were especially required. The methodological emphasis of US social science led to the very considerable systematization of social research methods in the USA in the mid-nineteenth century, including the successful forging of a marriage between social research and statistics, which had not been achieved in the UK at the turn of the century, when social research on one hand and mathematical statistics on the other had developed alongside each other.

National social science traditions may be shaped by differences in internal organization amongst social scientists themselves. In 'state-centered societies' (such as France or Germany) social scientists tend to consider themselves already organized by virtue of their university posts, with further organization amongst their colleagues not being so necessary; whereas in 'stateless societies' (such as the UK or the USA) there has been a conception of a profession of knowledge producers, which has led to considerable associational activity for the promotion of the social sciences.

In many countries there has been a development of semiprofessions, at least partly based on social science knowledge (social workers, planners, nurses, librarians, psychologists, economists, educators, and teachers). Their differential association with social sciences in each national setting has undoubtedly had particular affects.

A range of coalition partners, which differ in particular circumstances, may support particular developments of social science knowledge; especially where there is a cognitive and moral affinity. Such partners may include political parties, trade unions, probusiness groupings, welfare reform groups, or more generally social movements. Think tanks have been set up to mobilize social science knowledge for more specifically ideological purposes, especially in support of the doctrines of neoliberalism. There can be a 'discursive affinity' involving an overlapping of key concerns, and some basic similarity in cognitive assumptions and terms. In such circumstances, the social science work provides some of the conceptual elaboration and/or the social information required to support the program of its ally, while the ally may assist in providing identification of what issues are seen as problematic, empirical material (for example, access to research sites) and may assist in mobilizing resource support, as well as in providing more general legitimacy. This relationship is often strengthened when the partner obtains parliamentary power, or, more so, is in government. Social movements can be important in linking national traditions: for example, Marx bequeathed his writings to the German socialist movement (which gained a parliamentary foothold in the 1880s and 1890s) which harbored them, in turn making them available for widespread trade-union and academic utilization at the turn of the twentieth century and beyond.

The era of nationalism occurred over the same period as that of the most powerful discipline building among the social sciences. It is not unusual, then, to posit links between these two processes. Nation building was especially important in Germany and Italy, which were each culturally and intellectually dominated by their 'national questions' as they groped toward a newly united nationhood. The effect of nationalism seemed to provide a strong, if diffuse, stimulus to the development of the social sciences (see *Disciplines*, *History of*, *in the Social Sciences*).

The effects of different institutional forms and the material basis of the provision of resources has been especially important for fuelling the differential development of social science. A range of institutional forms have been relevant at different times and in different places: coffeehouses, salons, associations, university teaching departments, and research institutes (see Coser 1965, Wuthnow 1989) Salons and coffeehouses can be significant for the flexible inno-

vation of new ideas (as in France and the UK). Universities can be important for providing a more secure and enduring environment, with systematization required for teaching, as well as being rewarded as a goal. University settings propel specialization and divisions of labor, especially in the form of separate disciplines. However, university teaching departments are not necessarily appropriate institutions to support larger-scale research, so that the tackling of larger topics, and also policy research, often requires the development of research centers where a specialized division of labor can be built up and resources for particular products can be mobilized (see *Specialization and Recombination of Specialties in the Social Sciences*).

Different national university systems have provided different contexts for the development of social science. The German universities were reformed from the 1820s on and launched a range of more systematically based scientific work, especially in philology and then extending especially into history, which was placed on a far more scientific footing. The French system, which was not revivified until the 1870s, was (and still is) highly centralized, which can mean the rapid institutionalization of a particular area of knowledge, although the centralization can prove intellectually stultifying in the longer run as it can cramp innovations. As with natural science, the US university system, in which university presidents have strong power to develop new areas, where there can be fierce competition for prestige amongst institutions, and where (at least in larger universities) the appointment of several full professors in each department fosters a democratic climate and a diversity of lines of research, seems to have been a particularly successful environment within which social science has flourished. Although these were pioneered in Germany, the USA has also been especially important in the development of specialized social-research centers. One particular design feature which supports innovation seems to have been the importance of role hybrids—those with one foot in practical concerns and the other in a setting allowing for systematization of ideas are often especially innovative situations. US universities have tended to be better in providing such links.

The institutions within which social science development takes place may also have an effect on the dominant cognitive style through which social knowledge is produced and debated. German scholars were often ensconced in universities, whereas French scholars debated within salons. As a result, a typical style tended to develop: 'Whereas the German intellectual was systematic, scholarly, even pedantic, the French intellectual tended to be orientated to science as well as to political controversy and to be brilliant and lucid as well as facile and flowery in exposition' (Collins 1994, p. 14).

National traditions are more likely to develop where the subject matter is strongly localized: for example, the study of languages seems to have nurtured a considerable degree of longevity of localized scholarship. In contrast, more abstractly formulated social sciences are more likely to float free from national contexts.

Three broad modes of formation of social science can be distinguished (Wagner et al. 1991b):

- comprehensive social science perspectives, which reject specialization and professionalization: for example, idealism, social romanticism, German historical economics, classical sociology;
- formalized disciplinary discourses, which are appropriate to the internal organization of universities: for example, neoclassical economics, legal theory of the state, and formal sociology; and
- pragmatically specializing professions: codified professional training.

The first type of broad theorizing developed only in continental European countries during the late nineteenth century, where scholarship spanned both university chairs and independent intellectuals. More formalized disciplines developed as the social sciences became institutionalized within universities, especially in Germany and then in the USA. The third mode developed especially in the USA.

In discussing national traditions in social science, distinctions are often drawn between the English-speaking world (with further differentiation between the USA and the UK) and continental Europe (in turn divided between the north and south). Other countries fall into yet further patterns. In various of the smaller European (and other First World) countries, various early forms of social science developed, often with significant local figures, not yet entirely attuned to a social science framework. Sometimes these were institutionalized. Such local developments tended to be overwhelmed following World War II by the international spread of social science from its US bases. Local variants of social science were then reestablished, usually with a strong US flavor. Gradually, especially out of the pressures for change in the late 1960s and early 1970s, a little more local room for maneuver developed, although usually this did not lead to reconnections with the very early forerunners of national traditions.

The situation of national traditions in developing countries is broadly similar, although more muted, and with a later timescale. Some had local traditions, although usually these were very limited. Again, expansion and Americanization were late postwar phenomena. Although social science establishments now exist, they are smaller and less self-sufficient than in the previous group of countries.

In sum, during the period of the most active development in our current social science disciplines, innovative work was largely carried out in particular national contexts, albeit within a context of the considerable international exchanges of views. These innovative breakthroughs were often centered in

national traditions extending back some centuries. Although the high period of the play of national traditions has largely passed and social science now seems to be relatively universalistic, national contexts still continue to shape the work of the social scientists within each national community.

See also: Academy and Society in the United States: Cultural Concerns; Area and International Studies: Sociology; Nation: Sociological Aspects; Pragmatism: Philosophical Aspects; Qualitative Methods, History of; Quantification in the History of the Social Sciences; Quantitative Cross-national Research Methods; Science and Social Movements; Scientific Academies in Asia; Social Science, the Idea of; Theory: Conceptions in the Social Sciences; Traditions: Social; Universities, in the History of the Social Sciences

Bibliography

Cherns A B (ed.) 1974 *Social Science Organization and Policy*. Mouton, The Hague, The Netherlands

Collins R 1994 *Four Sociological Traditions*. Oxford University Press, New York

Coser L A 1965 *Men of Ideas: A Sociologist's View*. Free Press, New York

Deutsch K, Platt J, Rader J 1986 *Advances in the Social Sciences, 1900–1980: What? Who? When? How?* Lanham University Press, MD

Heilbron J 1995 *The Rise of Social Theory*. Polity Press, Cambridge, UK

Levine D 1995 *Visions of the Sociological Tradition*. University of Chicago Press, Chicago

Olson R N 1993 *The Emergence of the Social Sciences, 1642–1792*. Twayne, New York

Wagner P, Weiss C , Wittrock B, Wollman H (eds.) 1991a *Social Sciences and Modern States: National Experiences and Theoretical Crossroads*. Cambridge University Press, Cambridge, UK

Wagner P, Wittrock B, Whitley R (eds.) 1991b *Discourses on Society*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Wuthnow R 1989 *Communities of Discourse: Ideology and Social Structure in the Reformation, the Enlightenment and European Socialism*. Harvard University Press, Cambridge, MA

C. Crothers

Nationalism and Expressive Forms

Nationalism is a complex, often problematic concept for social scientists. But among existing definitions of the term, one finds three critical areas of agreement. First, the rise of nationalism is historically specific. Nationalism emerged in the late eighteenth century, appearing first in Europe, then in North and South America. In the twentieth century, nationalism spread to many Asian countries and to the new independent nations of Africa. Second, nationalism exemplifies an ideological movement, promoting the autonomy, unity, and sovereignty of those gathered in a single territory. A nation's citizens are believed to be joined not by ethnicity, but by a single public culture and a set of shared political goals. Finally, nationalism evokes a strong collective sentiment. While ruling elites may sow the first seeds of this fervor, it is argued that all citizens come to experience a shared collective identity and to embrace a common national purpose. (Calhoun 1993 or Hutchinson and Smith 1994 offer thorough discussions of nationalism.) The third element of nationalism, a strong collective sentiment, provides the focus for this entry. This article explores the most powerful expressions of nationalistic sentiment—a nation's symbols. The sections which follow highlight the ways in which national symbols bring nationalism alive.

1. National Symbols: The Nation Made Real

Ostensibly, national symbols may seem little more than calling cards by which nations identify themselves to one another. But closer reflection reveals that national symbols accomplish much more. Just as a weapons objectify feelings of anger, or romantic musical interludes objectify feelings of love, national symbols concretize the highly abstract; they make tangible the subjective nature of the nation—its moods, desires, and goals. Historian Michael Walzer (1967, p. 194) writes, 'The nation is invisible; it must be personified before it can be seen, symbolized before it can be loved, imagined before it can be conceived…national symbols provide a starting point for such political thinking.'

One can think of national symbols as modern totems. Like the signifiers of the primitives, these symbols merge the mythical, sacred substance of a collective with a specified, manifest form—one that is grounded in the everyday experience of sight, sound, or touch. By blending subject and object, national symbols move beyond simple representation. In a very real sense, national symbols become the nation. Recall Durkheim's (1915, p. 251) observations: 'The soldier who dies for his flag, dies for his country; but as a matter of fact, in his own consciousness, it is the flag that has the first place.'

2. Symbols as Expressions of Nationalism

As tangible entities, national symbols become the perfect vehicle with which to rally the citizenry. These symbols take on various social functions that heighten the experience of nationalism and facilitate its shared expression.

2.1 Crystallizing National Identity

National symbols vivify nationalistic sentiments by crystallizing national identity. In this regard, Anthony

Giddens (1985, pp. 216–19) identifies national symbols as the content of nationalism. These symbols establish the distinctiveness, the cultural autonomy of a population. They become the basis of a unique conceptual community.

To accomplish crystallization, political leaders give national symbols high exposure. Governments equip official institutions with national flags and emblems. Anthems, mottoes, and pledges of allegiance are taught in the primary grades and recited daily through the completion of secondary education. In many nations, citizens sing the anthem or salute the flag at the onset of concerts, sporting events, and other public assemblies. Political leaders often erect national monuments and designate historical sites. Such areas commemorate past strengths or sufferings of the national body. Citizens are encouraged to visit such sites and share in the sacred ideals that they enshrine. Often, national leaders create special holidays that revolve around such shrines. For example, each April, the Japanese celebrate Yasukuni Matsuri, a holiday during which citizens are encouraged to visit Tokyo's Yasukuni shrine and honor Japanese soldiers who died for their country (Gregory 1975, p. 153). On Memorial Day in the United States, flowers and flags decorate the gravesites of military casualties, with citizens engaged in parades or memorial services designed to honor the dead. By interjecting national activities into the routine calendar of events, national leaders disrupt citizens' individualistic orientations; they refocus citizens on collective life and recharge their nationalist sentiments.

2.2 Nationalism Through Bonding

National symbols can inspire nationalistic sentiment by creating bonds between citizens. This bonding power emanates from the symbols' sacred nature. As citizens join in veneration of the symbol, individual interests become secondary to those of the nation. Collective reverence toward national symbols brings individuals out of themselves and into communication with others. At such moments, national symbols bind citizens in a shared consciousness, liking them despite differences in wealth, social standing, power, or age. 'By uttering the same cry, pronouncing the same word, or performing the same gestures in regard to these (symbolic) objects, individuals become and feel themselves to be in unison' (Durkheim 1915, p. 262).

National anthems can unite citizens every time they are performed, bringing citizens together (physically and/or mentally) in patriotic communion. During World War II, the BBC promoted this goal, airing weekly broadcasts of the Allied Power's national anthems. European listeners (numbering in the millions) reported a strengthened resolve and increased camaraderie after hearing such broadcasts. In targeted nations, the broadcasts reportedly increased citizens' nationalistic fervor (Mead 1980, p. 46). A similar phenomenon can be witnessed when citizens sing the anthem or salute the flag at the onset of sporting events or international competitions. In attending to the symbol, a seemingly disconnected crowd is momentarily united as they collectively applaud their national essence.

National heroes also bond citizens in ways that intensify nationalistic sentiments. Heroes become civil deities capable of uniting citizens in common beliefs. For decades, Vladimir Lenin served this purpose in the Soviet Union. Ironically, his death best marks the initiation of this power. As a man, Lenin was admired, but as a symbol he became the paste of a society. A posthumous tribute in the Soviet Newspaper *Pravda* highlights the point: 'Lenin lives in the hearts of every member of our party. Every member of our party is a small part of Lenin. Our whole Communist family is a collective embodiment of Lenin' (*Pravda*, 23 January 1924, quoted in Lane 1981, p. 216). In passing from life to death, Lenin transcended the rigidity of observable human behavior. Thus in death, his image is freely molded to suit the evolution of each nationalist cause. Barry Schwartz (1987, 1996) documents a similar phenomenon in researching American heroes such as Washington and Lincoln, Schwartz argues that, in life, these men were undeniably important. But in death, these men became divine and extraordinary. Images of Washington and Lincoln have effectively drawn millions together in 'a moral communion.' The commemoration of these heroes becomes much more than the celebration of men. Rather, it defines and personifies the nation for which the men stand. In the moral communion of which Schwartz speaks, these heroes become the common food, the mystic substance that unites citizens in their experience of the nation.

2.3 Nationalism Through Honor

National symbols heighten nationalistic sentiments when they are used to honor the efforts of citizens. The process creates a symbiotic relationship between the living nation and the symbolic nation. Valorous citizens breathe life into the symbol by providing concrete examples of that for which the symbol stands. At the same time, the efforts of ordinary citizens become basked in the symbol's sacred aura. During the moments in which citizen and symbol are linked, the nation becomes 'real.' At these moments, nationalism becomes a lived experience.

Nations honor citizens in a wide variety of arenas. In world competitions such as the Olympics or the World Cup, nations honor winning athletes by playing their national anthem and hoisting their national flag. Similarly, nations welcome military victors against the backdrop of their national anthem and flag. To

commemorate their bravery, nations decorate military heroes with medals that display their nation's symbols. When citizens die in service of their country, nations drape the national flag over their coffins to signify their heroism. Nations memorialize patriotic sacrifice with national monuments: *L'Arc de Triomphe* in France, the *Völkerschlachtdenkmal* in Germany, the *Lenin Mausoleum* in the Soviet Union, and the *Vietnam War Memorial* in the United States. National holidays also serve the honoring function. Jose Marti Day in Cuba honors the famous patriot and his struggles (January 28). Tomb-Sweeping Day in Taiwan honors the nation's dead (April 5). Matilda Newport Day honors the Liberian patriot and the ideals for which she fought (December 1). In all of these examples, national leaders light the fires of nationalism with the best and strongest of the nation's citizens.

2.4 Symbols and Patriotic Action

National symbols often generate deep emotion. In so doing, these symbols can motivate patriotic action. National symbols become what Durkheim called a 'rallying center' for citizens. They can spur nations to victory and comfort nations in defeat. They can buoy citizens and heighten their resolve when national resistance must be sustained. As such citizens fights to protect their symbols just as they would fight to protect themselves. In such fights rests the supreme expression of nationalism.

National leaders recognize the motivational power of national symbols. They often create or adopt such symbols with motivational goals in mind. For example, in 1792 Mayor Dietrich of Strasbourg directed composer Claude-Joseph de L'isle to 'produce one of those hymns which conveys enthusiasm to the soul of the people' (Sousa 1889, p. 99). The result was *La Marseillaise*. When the tune made its way through France, it was credited with motivating volunteer battalions to descend on Paris (Nettle 1967, p. 69). In a similar way, Philippine General Emilio Aguinaldo implored native musician Julian Felipe to write an anthem that would rally his fellow Filipinos against the forces of Spain. When the battle was over, the General believed that such an anthem would further solidify the new Philippine regime. Each time it was heard, the song would rekindle the patriotism and loyalty that secured victory (Nettle 1967, p. 168).

In many nations, national martyrs become motivational symbols. Although prominent when alive, such individuals can become inspirational after death. They stimulate deep emotional commitments—commitments that often lead to extraordinary action. Consider Filipino activist, Benigno Aquino. While living, Aquino and his followers energetically opposed the Marcos regime. Yet, only after his assassination (August 1983) were Aquino's followers able to effectively unite and oust Marcos from power. Similarly, Father Miguel Hidalgo y Castilla preached revolution to the Indian peasants in Mexico for years. Yet, only with his execution in 1811 did his words gain the power to trigger the 1814 insurrection.

2.5 Symbols, Nationalism, and Legitimacy

Strong nationalist sentiments are contingent on the legitimacy of a nation's authority. National symbols play an important role in establishing and maintaining that legitimacy. When nationhood ensues, flags, crests, mottoes, and other national symbols facilitate nationalistic sentiments by finalizing an entity's new status. For example, when over 100 nations celebrate an Independence Day, they institutionalize the birth of a national entity. Similarly, the inclusion of a nation's flag at the United Nations marks the acceptance of a new, full-fledged member in the 'national club.'

Symbols also legitimate changes to an established nation's internal authority structure. During such transitions, symbols are often recreated or altered so as to reify the shift in power. For example, upon declaring independence, Venezuela added the phrase '19 de Abril de 1810—Independencia; 20 de Febrero de 1859—Federacion; Republica de Venezuela' to its coat of arms. By linking the independence movement to the past history of the arms, the new Venezuelan government was more easily woven into the existing national identity (Mucha 1985, pp. 38, 176, 181). Similarly, changes made in 1991 to the Croatian national flag became a powerful signal of the nation's new independent status. While part of the Yugoslavian republic, Croatia's national flag displayed a large red star over a red–white–blue tricolor field. At independence, the government replaced the red star with five small shields representing different parts of the country (Crampton 1992). In this way, Croatia downplayed its communist past and presented a new self-study.

National leaders sometimes use national symbols to cloak their policies in a nationalistic sentiment. By merging their persona and plans with the nation's symbols, they seek legitimacy for themselves and their administrations. It is not surprising, then, that when rulers address their nations, national symbols are often central to the setting. The US President frames his public addresses with the national flag and our national emblem, the eagle. Portraits of Mao Zedong still provide a backdrop for the addresses of many Chinese leaders. Heads of state often have a song, seal, or flag which becomes a permanent accompaniment to their presence—*Hail to the Chief* in the US, *The Royal Fanfare* in Iraq, or the President's flag in Gabon, Korea, Liberia, Peru, etc. National rulers often are honored on the country's postage stamps or national currency. Leopold III, for example, was

commemorated on Belgian coins. The accession of Jigme Wangchuk was depicted on Bhutan's currency. Jawaharlal Nehru could be found on certain Indian coins. Francisco Franco was pictured on Spanish currency. Combining leader with symbol renders the leader the standard bearers of all that the nation's citizens value and believe. Leaders, like symbols, become the subject of nationalistic fervor.

2.6 Bottom-up Nationalism

When symbols are used to crystallize identity, create bonds, honor, motivate action, or legitimate, their ability to spur nationalistic sentiment travels downward through the social system. The power of symbols is directed from those that govern to the governed. Conversely, when symbols become tools of popular protest, the general public takes command of them. Such conditions present popularistic or 'bottom-up' expressions of nationalism.

Bottom-up nationalism is most likely to emerge when leaders and their constituents become glaringly divided. When leaders violate the public trust, citizens may reclaim the nation's sacred symbols and return them to 'virtuous' hands. French resisters, for example, utilized this strategy during the Second Empire. While Napoleon III outlawed *La Marseillaise*, deeming it too dangerous for public consumption, resistors embraced the anthem anew, resolving to forward it as a reminder of revolutionary ideals. These citizens sang *La Marseillaise* at both underground settings and public protests. In this way, the anthem was dubbed the 'true' nationalist message, directly challenging the official government position (Mead 1980, p. 54). The American flag served a similar role in the 'Bonus March' of 1932. Throughout this protest, American veterans consciously cloaked their grievances in the American flag. They carried the flag during rallies; they posted it on their temporary settlements. The flag became aligned with the veterans' fight. By taking command of the flag, the veterans placed the formal government leaders in opposition to the will of the people. Thus, when President Hoover ordered federal troops to forcibly evacuate the veterans, millions of Americans were aghast. Physical attacks on the veterans represented attacks on the flag, and thus, antinationalist actions.

Bottom-up nationalism often involves the juxtapositioning of symbols. In such cases, a 'grass roots' symbol is positioned against one strongly promoted by ruling elites. In this way, citizens force a confrontation between what 'is' and what 'should be.' Chinese students in Tiananmen Square used the Goddess of Democracy in this way. This 35 foot sculpture, modeled after the Statue of Liberty, sat in opposition to a massive portrait of Mao Zedong. Juxtaposing statue and portrait pitted democracy against communism. The strategy initiated a new nationalistic surge in the face of old routines.

When the schisms between national leaders and their constituents become severe, the general population may actually reject its nation's symbols. Under such conditions, the role of national symbols can change significantly. The symbols, like those who control them, become the target of protest. Thus, the expression of nationalism can involve some form of attack on national symbols. For example, during the 1989 revolution, Romanian protesters frequently cut the Communist ornament from the center of the nation's flag. This attack symbolically aborted the reigning leadership, leaving room for the general populous to reclaim its nation. A. Codrescu (1991, p. 67) a journalist and Romanian exile, wrote 'It is through that hole that I am returning to my birthplace.' Recent independence movements in the Baltic republics (1990–1) involved similar attacks. In Lithuania, citizens danced on a toppled statue of Lenin. In Russia, plans were made to dismantle Lenin's burial shrine. Defiling these symbols crystallized citizens' rejection of the standing authorities. Further, the unity and intensity that characterized such attacks located the center of nationalism in the hands of the people.

3. Designing Nationalistic Expressions

National symbols represent the most powerful expressions of nationalism. Consequently, one might expect that the design of such symbols would be grounded in each nation's unique, indigenous characteristics. Research shows that a nation's indigenous characteristics can indeed govern the selection of symbol *content*—i.e., the particular notes or lyrics that appear in an anthem, the particular colors or emblems chosen for a flag, etc. Yet, when one studies the *structure or design* of national symbols—i.e., the way in which notes are combined to create an anthem melody, the ways in which colors and shapes are combined to create a flag design—the influence of a nation's indigenous characteristics prove far less important.

Cerulo (1995) conducted a detailed study of the creation and adoption of national anthems and flags. She explored the different factors that guided the selection of symbol content vs. symbol design. Her research suggests that symbol content forwards the 'raw materials' of a nation's identity. As such, content generally is guided by each nation's unique cultural history. In contrast, symbol structure represents the strategy or 'voice' by which nations choose to express their essence. In reviewing the more than 180 anthems and flags currently in existence, Cerulo showed that variations in such strategies of expression are a 'supranational' phenomenon.

Some nations choose to express their identities via basic structures—designs built upon concise, restricted

forms of expression. Other nations express their identities using embellished structures—elaborate, dense, and complex forms of expression. Differences in structure are systematically linked to the broader social conditions from which symbols emerge. Those creating and adopting national symbols locate their nations within certain economical, political, and social 'maps.' They then choose their strategy of expression with reference to those who share similar locations in these domains.

For example, a nation's economic position within the world arena is related to the structure of its anthem and flag. Nations at the core of the world system—those with the greatest power and resources—favor very basic symbol designs. In contrast, nations that lack wealth and power adopt symbols with highly embellished designs. A nation's political system (authoritarian vs. democratic) is also associated with the structure of its anthem and flag. While nations with authoritarian leadership favor basic, sparse symbol structures, nations with democratic systems tend to adopt more elaborate, embellished structures. Specific social events surrounding the adoption of symbols impacts symbol designs as well. Events that unify the collective in a common focus—i.e., independence movements, nationalistic periods, victory in war, etc.—prove to be associated with the adoption of basic symbol designs. When symbols are adopted during periods that fragment the population's focus—i.e., bloody coups, economic depressions, lost wars, natural disasters, or revolutions—embellished symbol designs are favored.

The patterns identified here should be thought of as norms of symbol design. Cerulo's work explains the ways in which a core economic location *calls for* the adoption of basic symbolic expression; she outlines the ways in which a democratic political structure *presents the expectation* for embellished symbolic expression, etc. Further, her data suggest that nations who violate these expressive norms in creating and adopting anthems and flags often find themselves with symbols that fail to fully connect with the citizenry. When symbol structures fail to conform to design expectations, citizens often disengage from the symbols.

This line of research presents an interesting thesis. All national symbols are used quite similarly in the expression of nationalism (i.e., crystallizing national identity, bonding and honoring citizens, motivating patriotic action, legitimating the nation, or providing citizens with a means of protest). Yet, the specific design of national symbols can make them more or less effective in achieving their goal. When it comes to the expression of national identity, where a national audience 'comes from' (i.e., its cultural roots) may be less important than where that audience 'is' (i.e., its experience at the social moment during which symbols emerge). Put another way, the ultimate 'success' of national symbols may be more strongly influenced by a citizenry's community of experience than to its community of record.

See also: Collective Beliefs: Sociological Explanation; Collective Identity and Expressive Forms; Cultural Expression and Action; Cultural Psychology; Cultural Rights and Culture Defense: Cultural Concerns; Cultural Studies: Cultural Concerns; Culture and Emotion; National Character; National Socialism and Fascism; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Sociology of; Nations and Nation-states in History; Nation-states, Nationalism, and Gender; Political Culture; Religion: Nationalism and Identity; Symbolism (Religious) and Icon; Xenophobia

Bibliography

Calhoun C 1993 Nationalism and ethnicity. *Annual Review of Sociology* **19**: 211–39

Cerulo K A 1995 *Identity Designs: The Sights and Sounds of a Nation*, Rose Book Series, American Sociological Association. Rutgers University Press, New Brunswick, NJ

Codrescu A 1991 *A Hole in the Flag: A Romanian Exile's Story of Return and Revolution*, 1st edn. W. Morrow, New York

Crampton W 1992 *Pocket Guide to Flags*. Crescent Books, New York

Durkheim E 1915 *The Elementary Forms of Religious Life* (trans. Swain J W). Free Press, New York

Giddens A 1985 *The Nation-State and Violence*. University of California Press, Berkeley, CA

Gregory R W 1975 *Anniversaries and Holidays*, 3rd edn. American Library Association, Chicago

Hutchinson J, Smith A D 1994 *Nationalism*. Oxford University Press, Oxford, UK

Lane C 1981 *The Rites of Rulers*. Cambridge University Press, Cambridge, UK

Mead R 1980 National anthems. In: Sade S (ed.) *The New Grove Dictionary of Music and Musicians*. Macmillan, London, Vol. 13, pp. 46–75

Mucha L 1985 *Webster's Concise Encyclopedia of Flags and Coats of Arms*. Crescent Books, New York

Nettle P 1967 *National Anthems*, 2nd edn. (trans. Gode A). Frederich Ungar, New York

Schwartz B 1987 *George Washington: The Making of an American Symbol*. Free Press, New York

Schwartz B 1996 Memory as a cultural system: Abraham Lincoln in World War II. *American Sociological Review* **61**(5): 908–27

Sousa J P 1889 *National, Patriotic, and Typical Aires of All Nations*. Horry Colemon, Philadelphia

Walzer M 1967 The facsimile fallacy. *American Review of Canadian Studies* **12**(2): 82–6

K. A. Cerulo

Nationalism: Contemporary Issues

'Nationalism' is the *belief* that one's own land and people are unique and superior to all others, and/or

a *policy* that systematically advantages one's own nation. Frequently the two, belief and policy, are linked, occurring simultaneously.

Nationalism's manifold dimensions are captured well by the *Oxford English Dictionary* (OED): 'Devotion to one's nation; national aspiration; a policy of national independence.' This concatenation of concerns has fueled modern research classics such as C. Hayes (1931), H. Kohn (1944), and E. Gellner (1983). Yet even such a definition is inadequate, for it ignores the role nationalism has played in changing the modern world.

More helpful but still not without problems is K. Deutsch's (1963) concept of nationalism. It speaks of a *people* as a social grouping linked communicatively, a *nationality* as a people seeking control over its members, a *nation* as a nationality seeking self-control, and a *nation-state* as a nation with its own political apparatus. The center point is the nation, which the OED describes as 'an extensive aggregate of persons, so closely associated with each other by common descent, language, or history, as to form a distinct race or people usually organized as a separate political state and occupying a definite territory.' Does this tell us what we want to know about, for instance, the Yugoslav revolution of the late 1990s and the policy of ethnic cleansing?

1. Conceptualizing Nationalism

Fundamental to the basic concept of nationalism is, first of all, the observation that the nation into which we were born makes a difference. Not all nations are born equal. Moreover, the social values that nations offer to their citizens—such as preferred jobs and access to higher education—are distributed unequally. In the international arena, some nations are 'haves' and some are 'have nots.'

Second, one's responsibility as a member of a nation is to improve its competitive position. But note: while the first postulate is uniform across a nation or people, the second offers variable individual acceptance of the nation's policies, instruction, or behaviors.

Third, nationalists may at least in principle undertake whatever actions they deem appropriate for improving their people's access to wealth, power and other values.

In such an approach the term *nationalism* becomes ambiguous, and its meaning a matter of time and place. Even so the idea of nationalism is useful as an analytic marker. Europe's march across history before the fourteenth century frequently repeated the social mobilization sequence: 'people→nationality→nation →nation-state.' Using nationalism as a concept reminds us of its many dimensions, among them national consciousness, national character, ethnonationalism, national cleansing, and National Socialism. It offers a framework for differentiating among population groups.

2. Historical Origins

Although the English-language usage of the word 'nationalism' dates back only to the 1840s, its earlier presence as a concept had deep roots. An oft-cited point of origin occurred a half-century earlier—1776 when colonial Americans bolted from their English sovereigns, 1787 when they met in Philadelphia to pen a Constitution, and 1789 when French revolutionaries drafted their Declaration of the Rights of Man.

2.1 An Inchoate Concept

The historical record merits more attention, however. Consider the rich medieval fabric with centuries of unified governance and political stability. Charles IV's Golden Bull (1356), by removing Papal power from the Holy Roman Empire, contributed to a process that eventually killed the HRE itself. The end of the Hundred Years War (1453) sealed the division of England and France.

During the Renaissance (fourteenth to seventeenth centuries) the cosmic drama shifted from gods to humankind, and a concomitant sense of personal responsibility. Also deserving our attention are Martin Luther who tore apart a united Christendom (1517), Philip II's Escorial (1563–84) which gave Spain a residence *cum* administrative center, and arts and literatures intoning national pride. Who has not thrilled to Shakespeare's account of Prince Hal exhorting his troops: 'We few, we happy few, we band of brothers'?

But all this was prototypical, *not* full-scale nationalism as we know it today. The next two and a half centuries would further secularize Europe and shore up this popular albeit inchoate concept. Thoughtful observers and policy makers, like Plato's cave dwellers, could only dance around the idea's images, unaware that thoughts dancing through their minds could transform European and, by now, North American politics.

2.2 Belonging: Romantic Nationalism

The nineteenth century was one of romantic nationalism. Policy makers were interested in the idea of nationalism and in understanding the ways in which it might be used to accomplish their own national goals. Among these goals was of course military defense. By and large, however, political leaders were more preoccupied with traditional ways of conducting warfare, and dubious when 'psychological' variables were brought into play.

Had the nationalist movement been presented as an organization table, then surely a major item would have aimed at understanding and communicating information about the status of the national entity itself, whether that entity was a people, nationality, nation, nation-state, or something else. Another major

item would have been proselytizing: bringing more recruits into the national fold, expanding the entity's structure, and propagandizing. Yet another item would have involved the emerging world of advertising: symbols, heroes, flags and anthems, operas and plays, and common histories, as well as formal presentations.

What made this late nineteenth century nationalism 'romantic'? Among other factors two stand out: undigested sentiments endorsing the glories of nationhood in general, and unbridled acceptance of their own nationalizing efforts. Nationalism became viewed as the modern society's universal solvent, ready on call to remedy all of society's ills. Such conditions may produce an *élan* that improves the quality of a society. Indeed, *élan* and competence are both necessary for effective governance. But eventually such political thinkers as President Woodrow Wilson turned against the tide of popular sentiment: they saw nationalism as a universal evil.

2.3 Aggressive Nationalism

Beginning in the mid-nineteenth century nationalism became more aggressive. Each kind of nationalism, we might assume, aimed at improving the lot of its own members. The critical question became the treatment of territorially linked individuals or groups who did not share much else in common with the dominant national population. Other societies had for centuries tolerated this apparent anomaly: thus under the Roman empire the famed legions relied on some Romans but mostly on foreigners recruited abroad to fight for Rome. Europe's nineteenth century romantic nationalists also treated nonconformers with benign neglect.

Writers favoring aggressive nationalism took a different stance, one that neither sought nor tolerated any more than a modicum of diversity. If, they reasoned, we are responsible only for our own people, and if territory and population are conterminous, then we have no obligation to sustain or protect aliens in our midst. If our nation-state continues to house 'The Stranger within our Gate' (in Rudyard Kipling's poetic vision), perhaps because of long-accepted customs or usages, then we are justified in doing whatever is necessary to restore our social balance.

How can nation-states deal with 'alien' minorities (or even majorities) in their midst? First, individual options include bureaucratic hassle, hostile customs officers, burdensome taxation, visa denial—and a whole host of other boundary maintaining techniques designed to intimidate the timid; 'we don't want you here in *our* country.' A second level of anti-outsider actions might seek to isolate alien groupings in ghettos or deny their children access to public education. A third technique expels outsiders. Ousting the *personae non grata* has been used since ancient days.

An extreme hallmark of modern aggressive nationalism is *genocide*, which OED characterizes as 'The deliberate and systematic extermination of an ethnic or national group.' Examples include formal war crimes charges after the American Civil War, the US Army's extermination of American Indians, concentration camps in Cuba during the Spanish–American war, Turkey's murder of Armenians in the 1910s, Germany's mass murder of Jews and others, and internal mass murders in Yugoslavia, Laos, and Rwanda. All have made today's world more dangerous.

3. The Study of Nationalism

Nationalism has now and again captured the attention of scholars and others wanting to understand the concept or perhaps employ its policy opportunities. Some research has been suggestive, some others less—in part because the topic itself is a moving target that is approached in very different ways.

3.1 Analytic Approaches

Early writings on nationalism were not systematic but rather

(a) descriptive accounts, such as St. John de Crèvecoeur's (1782) *Letters from an American Farmer* and A. Tocqueville's (1835) *Democracy in America*, which sagely characterized American ways. The model they set up became the pathway for a plethora of subsequent journalists, historians, academics, and other writers. An important twentieth-century innovation was work in

(b) cultural anthropology by such specialists as R. Benedict (1946) and M. Mead (1942) who were trained to study the ways in which institutions and cultural patterns shape the individual. Social scientists like S. Kracauer (1947) and H. Lasswell et al. (1952) used

(c) content analysis to explore how culture and such artifacts as movies and novels inform us about a society's values, perspectives, and norms. By the mid-1930s researchers were using

(d) survey research to study 'modal personality structures'—relatively enduring personality characteristics and patterns among a society's adults (Inkeles and Levinson 1954). These and other research techniques have deeply enriched our ability to understand social processes affecting modern nationalism.

3.2 Research Pitfalls

Analysts using these various approaches have unquestionably enriched our understanding of nationalism. Certain elements, however, have occasionally produced little more than discord. The sharpest split has

pitted historians against data-oriented social scientists. Consider 'consciousness.' Attention to *national consciousness* stems from the plausible and hence oft-heard notion that individual and national consciousness are tightly linked to each other, and that its presence has a political or otherwise social impact on the rest of us.

Is this hypothesis warranted? The social scientist may have little difficulty describing *individual consciousness* (for instance, as manifest by an individual's perspectives). The good-natured social scientist may even tolerate rhetorical sallies suggesting that people who have been together for a long time may develop a *group consciousness*. But when the analysis moves from loose words to precise understandings, and tries to merge activities at different *levels of analysis* (individual, family, small group, religious sect, and so forth), social scientists are likely to cry 'Foul!' At this point the analysis is committing the logical error that W. Robinson (1950) identified as an ecological fallacy. The idea of national character encounters a similar problem.

4. Contemporary Issues in Nationalism

What progress have we seen regarding the various dimensions of nationalism? The question is particularly complex because theoreticians and practitioners use the relevant terms in different fashions. For instance, what is 'national' about Hitler's National Socialism? (For that matter, what is its 'social' or 'socialist' content?) It was a term of uncertain meaning that caught hold in a particular place at a particular time. The fact that it caught hold, however, ensures that international studies will continue to wrestle with it. Similar concepts include 'ethnonationalism,' 'ethnic cleansing,' and 'nation-state.'

4.1 Ethnonationalism

'Ethnonationalism' is a neologism seeking clarity in the worldwide discussion of nationalism. The Greek word '*ethno*-,' which means 'race,' 'culture,' or 'people,' has been added to the root 'nationalism.' This makes explicit the link between the nation's location and ethnicity—without, however, addressing such central issues as the nonscientific nature of words such as 'race.'

4.2 Ethnic Cleansing

The notion of ethnic cleansing captured the world's attention in the 1990s. It is a policy that 'cleanses' one's own national entity (such as ethnic group) by systematically expelling or even killing residents who are not part of that dominant entity. What is the desired outcome? Purity of race is a major goal. The idea is that, if our ethnic group were truly free, and if it were permitted to exert its own control over our nation-state, then we would have a 'pure' society that could realize our goals. Critics view ethnic cleansing as a euphemism for an inefficient *genocide*.

4.3 Nation-state

A dominant norm organizing international life is the nation-state. Nineteenth-century thinkers and policy makers treated nationalism as the mantra of the new age. 'Nationalists' moved around the world encouraging other nationalist groups—writers (the Grimm brothers with their 'purely German' fairy tales), composers (Tchaikovsky's *Marche slave*, so powerful that if threatened to disrupt public order when it was first presented), and many more. Liberals such as Woodrow Wilson who had originally favored the nationalist movement soon came to believe that its excesses were the causes of rather than the solution to a continuing international disorder. The post-1945 euphoria spawned by the United Nations movement in turn gave way to the nationalist, divisive orientation of Cold War politics.

The early 1990s once again found national leaders in the international arena making moves toward conciliation, while at the same time we have seen an increase in the number of national entities around the globe. The point is twofold. First, the idea of nationalism is alive and well, and seems likely to persist in the new millennium. Second, even steps toward interstate unity do not mean the end of the nation-state. Thus European unity does not necessarily mean abandoning France or Germany.

See also: American Revolution, The; Ethnic Cleansing, History of; Ethnocentrism; European Integration; French Revolution, The; Genocide: Anthropological Aspects; Nation: Sociological Aspects; Nationalism: General; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Nations and Nation-states in History; Religion: Nationalism and Identity; Xenophobia

Bibliography

Benedict R 1946 *The Chrysanthemum and the Sword: Patterns of Japanese Culture*. Houghton Mifflin, Boston, MA

Deutsch K W 1963 *Nationalism and Social Communication: An Inquiry into the Foundations of Nationality*. The Technology Press of The Massachusetts Institute of Technology, Cambridge, MA

Gellner 1983 *Nations and Nationalism*. Basil Blackwell, Oxford, UK

Hayes C J H 1931 *The Historical Evolution of Modern Nationalism*. R. R. Smith, New York

Inkeles A, Levinson D J 1954 National character: The study of modal personality structure and sociocultural systems. In:

Allport G W, Lindzey G (eds.) *Handbook of Social Psychology*. Addison-Wesley, Cambridge, MA, Vol. II, pp. 977–1020
Kohn H 1944 *The Idea of Nationalism: A Study of Its Origins and Background*. Macmillan, New York
Kracauer S 1947 *From Caligari to Hitler: A Psychological History of the German Film*. Princeton University Press, Princeton, NJ
Lasswell H D, Lerner D, de Sola Pool I 1952 *The Comparative Study of Symbols: An Introduction*. Stanford University Press, Stanford, CA
Mead M 1942 *And Keep Your Powder Dry: An Anthropologist Looks at America*. W. Morrow and Co., New York
Robinson W S 1950 Ecological correlation and the behavior of individuals. *American Sociological Review* **15**: 351–57
St. John de Crèvecoeur J H 1782 *Letters from an American Farmer*. John Exshaw, Dublin
de Tocqueville A 1835–40 *Democracy in America*. Saunders & Otley, London

R. Merritt

Nationalism: General

Most theorists are agreed that in the specific form in which we know it today, nationalism is closely related to the modern state and modern politics. Its rise and spread is linked to the need in modern industrial societies for a standardized language, a homogeneous national culture, and an imagined horizontal community. The modern idea of popular sovereignty was expressed in the identity of the people with the nation-state. This in turn legitimized the rise of anticolonial movements as a necessary unfolding of modernity. But at the turn of the millennium, the relations of nationalism to the new global order, on the one hand, and to the particularisms of ethnicity, gender, and class, on the other, are facing new challenges.

1. The Modernity of Nationalism

There is a debate about nationalism's age. Some argue that nations are old, built as they often are around histories of ethnic, linguistic, religious, or dynastic solidarity going back several hundred years before the modern era (Armstrong 1982, Smith 1986). European countries such as France, Portugal, Denmark, Sweden, or Russia and Asian countries such as China, Japan, or Korea are often portrayed as having old and continuous histories in which a well-defined national identity can be easily perceived. The Jewish people are said to have been a nation, distinguished by ethnicity and religion, since Biblical times. Other analysts hold that even when nations claim ancient lineages through such ethnic and cultural histories, the particular combination of elements that gives to a people the distinct form of a nation is always a product of the modern era (Gellner 1983, Hobsbawm 1990). Various dates have been suggested as the point of emergence of nationalism as a modern phenomenon. Some say it was the French Revolution, others the independence movements in Latin America, while still others think that it was the German Romantic movement which supplied the crucial ingredient. But it is generally agreed that by the early nineteenth century, most of the elements that would characterize nationalism were in place within the emergent discourses of politics and society in the Western world. Henceforth, nationalism would be a constituent aspect of modernity itself.

One of the doctrines that nationalism has turned into a feature of our ordinary common sense is that the world is divided into nations, so that every person, under normal circumstances, is born into a nationality. This common understanding is bolstered by such everyday visual representations as multicolored maps of the world, identity documents such as passports, and international gatherings such as the United Nations or the Olympic Games. These representations are built on another modern doctrine that has become a truism—except for Antarctica, the entire landmass on earth is divided up into national territories, each of which is under the unique sovereign jurisdiction of a state. The two doctrines together add up to a moral claim—that the proper form of modern politics is to be found where the political unit of the state coincides with the cultural form of the nation.

It is not difficult to see the strong connections between these doctrines of nationalism and some of the basic tenets of modern political theory. The definition of national sovereignty as the unique jurisdiction of a single state authority over a territory clearly demarcated on maps and recognized by other state authorities had emerged in Europe in the seventeenth century. What was added in the eighteenth century was the revolutionary idea that the true locus of this sovereignty is the people. Following the course of antiabsolutist thinking in the period of the French and the American revolutions, just as there emerged the idea of the citizen as the autonomous bearer of rights, so also was it asserted that the people, self-constituted and self-determining, are autonomous. This doctrine would supply the moral justification for the nineteenth-century democratic national movements in the territories of the Habsburg, the Ottoman and the Czarist empires. In the twentieth century, with the rise and success of anticolonial movements in Asia and Africa, this would become recognized as the right of self-determination of nations.

2. The Morality of Nationalism

Although nationalism shares its genealogy with the modern state and its politics, it has not always found moral approval. The grounds of criticism, however, usually lie in ideological debates that are part and parcel of modern politics itself. A common criticism is based, for instance, on the distinction between those nationalisms that are more voluntarist in their membership, built around institutions that are more like civic associations, and those that are more organic where membership is ascribed by birth into a collective bound by historic ties of cultural community. The former is said to be the more desirable type, since it respects the right of the individual to choose his or her nationality; the latter is likely to place greater value on the community than on the individual and could thus easily lead to xenophobia and authoritarian politics. This is a distinction that is at the heart of liberal theories of nationalism and usually classifies nations like France, Britain, the Netherlands, Belgium, and the USA as belonging to the first type and those of Central and Eastern Europe, Russia, the Middle East, and most of Asia to the second (Kohn 1944).

Civic nationalism is meant to think of the political community of the nation as a voluntary fraternity of citizens whose rights of equality and liberty find fruition in the nation-state. The organic idea of the nation, on the other hand, seeks a more natural basis of community, such as, very influentially, in the German Romantic notion of a folk united by the historic associations of a common language and shared myths. The former, the liberal would say, tends to be universalist and inclusive in its political vision, the latter particularist and exclusive (Greenfeld 1992, Tamir 1993). In actuality, of course, even liberal states subscribing to the universalist values of civic association have gone through much conflict before extending their citizenship to those who were not regarded as part of the 'original' political community—Catholics in the UK, for instance, or Afro-Americans in the USA—and the debate still goes on in many Western countries over the status of the new immigrants whose race, language, or religion marks them out as different from the 'normal' member of the political community.

Of modern political ideologies, Marxism in its early years strongly upheld the universalist idea of the political community and regarded nationalism as a limited and transient phenomenon which, if espoused by the working class, would impede the truly emancipatory path of proletarian revolution. The emerging democratic movements in the territories of the Austro-Hungarian empire, however, impelled the Austrian Marxist Otto Bauer to the view that social democrats should support the demand for independent nation-states. Later, in the context of the nationalities question in Russia and the discussions in the Communist International on the colonial question, Lenin argued for the right of self-determination of nations and Stalin proposed a set of objective criteria such as shared territory, language and history by which nationalities were to be identified (Nimni 1991).

3. The Sociology of Nationalism

Quite apart from the doctrinal terms in which nationalists describe their project, sociologists have tried to provide an account of the sociohistorical conditions that produce nationalism. Ernest Gellner has identified these as the conditions of modern industrial society which requires free movement of labor and goods throughout the national economy, a general and universal schooling in a standard national language and a rhetoric of equal access to social mobility. The complexity, interdependence, and mobility of industrial life places an unprecedented emphasis on complex, precise, and context-free communication. This tends to destroy both the exclusiveness of premodern elite high cultures and the provincialism of local peasant cultures and to produce in their place an increasingly standardized and homogeneous national culture. The transition to some form of modern industrial society, Gellner argues, is unstoppable. Consequently, nationalism too is an ubiquitous feature of our contemporary world (Gellner 1983).

If nationalism is indeed the product of certain structural features of modern industrial society, why is it that its particular cultural content is so often the focus of such intense ideological passion and contestation? Benedict Anderson has sought to supply the link between the material and the ideological. The nation, he says, is an *imagined* political community. It is imagined, first, in the sense that although no single member personally knows all of the other members of the nation, yet in their minds they all live in community. It is imagined, second, as a sovereign community, in line with the modern political idea that the emblem of freedom is the sovereignty of the people. It is also imagined, third, as a limited and particular community, because humankind is divided into many nations. And finally, it is imagined as a community—one that creates a deep, horizontal comradeship. This imagining of the nation, however, becomes possible only through certain specifically modern technologies and institutions. The most crucial is the technology of printing which, in combination with the new production system of capitalism, gives rise to a publishing industry that produces, through newspapers, textbooks, novels, and bureaucratic documents, a national literature in a standardized national language. Newspapers and novels, in particular, make possible the imagining of lives lived in simultaneity with other real and fictional members of the community, convert events happening to other people at other places into events in the shared life of the community, and thus transform an imagined solidarity into the homogeneity

of everyday lives. Censuses, maps, museums, and memorials are other important institutions that contribute to the imagining of nations (Anderson 1991).

4. Nation and State

The modern form of the nation is both universal and particular. The universal dimension is represented by the notion of 'the people' as the original locus of sovereignty in the modern state and of all humans ('man,' in the original eighteenth-century revolutionary declaration) as bearers of rights. If this was held to be universally true, how was it to be realized? By enshrining the specific rights of *citizens* in a state constituted by a particular people, namely, a *nation*. Thus, the nation-state became the particular, and normal, form of the modern state. It has been argued that if the basic framework of rights in the modern state was defined by the twin ideas of freedom and equality, the often contradictory pulls of these two ideas had to be mediated by two further concepts: those of property and community. Property sought to resolve the contradictions between freedom and equality at the level of the individual in relation to other individuals. Community was where the contradictions were sought to be resolved at the level of the whole fraternity. Along the dimension of property, the particular resolutions might be more or less liberal; along the dimension of community, they might be more or less communitarian. But it was within the specific form of the sovereign nation-state that the universal ideals of modern citizenship were expected to be realized (Balibar 1993).

It needs to be emphasized that, just as the sovereign nation-state was a sign of the fulfillment of the universal moral values of modern citizenship, so also was it a concrete entity defined by the practical considerations of viable statehood. The former provided the ideological force in the nineteenth century to the revolutionary democratic movements in many regions of the Habsburg and the Ottoman empires as well as to the movements for German and Italian national unification. But the latter became operative when judging the claims of groups that were considered too small or territorially dispersed or politically immature to have viable nation-states. Thus, the Welsh or Sicilians or Bretons, or the many small ethnic groups in the Balkans, or a European colony such as Ireland, were considered unfit for independent statehood. This highlights the fact that no matter what the specific cultural content that defines nationhood in a particular nation-state, the requirement of undivided and viable territorial sovereignty will always produce national minorities within that state. It also points to another feature that developed through the nineteenth century to become a basic principle of the modern political world: Sovereign statehood depended crucially on recognition from other states within an international states system.

The late nineteenth century was a period of the democratization of politics in the Western world, with the expansion of the suffrage and a wider sense of involvement of the working class in national politics, even if in a role of opposition to the ruling classes. Alongside, there also operated a process of producing national citizens, through a rapidly expanding system of schooling, the imposition of standardized forms of the national language, and not the least, by inculcating a popular sense of patriotism by celebrating the achievements of its state leaders and armies. There is little doubt that at the turn of the century, popular nationalism in the UK, France, Germany, Russia, Italy, Japan, and the USA—regarded as the great powers—was deeply implicated in imperial ambitions, colonial expansion, and military glory.

5. Anticolonial Nationalism

The emancipatory appeal of modern nationalism lay in its association with the universalist moral claims of the modern state. But, as we have pointed out before, it was also necesarily conditioned by particular claims to viable statehood within an international states system. The anticolonial settler and creole nationalisms of North and South America in the late eighteenth and early nineteenth centuries were, in many ways, of the same type as the democratic nationalisms of nineteenth-century Europe in that the antiabsolutist element was far more predominant and unambiguous than the anticolonial. The interesting and tragic contrast was the failed Haitian revolution of the 1790s, launched by Black slaves and ex-slaves inspired by the emancipatory message of the French revolution, and crushed under orders from the revolutionary regime in Paris (James 1963, Trouillot 1995). The world was not yet ready for the real wave of anticolonial nationalism.

That would begin a hundred years later. Many of the apparatuses characteristic of the modern state, including its bureaucracies, legal and judicial systems, economic institutions, censuses, communication networks, etc. were introduced into European colonies in Asia, Africa, and the Caribbean as technologies of colonial governance. In some British and French colonies, this even included limited representative institutions in which native elites could participate. Equally significant was the introduction into these countries of modern education in European languages which opened up for a new group of indigenous intellectuals the possibility of imagining a modern nationhood for their peoples, free from colonial bondage. Curiously, the emancipatory claims of modern statehood were used in the nineteenth and twentieth centuries by colonial apologists to justify Western rule over the rest of the world: those backward regions, they said, had to come under Western tutelage in order to introduce them to modernity and to prepare them for responsible government. The nationalist

movements of Asia and Africa, by contrast, argued that it was precisely colonial rule that stood in the way of their becoming modern and self-governing nations, fully responsible for their own destinies. Not only was self-government demanded, in accordance with the universalist modern doctrine, as a natural right of every nation, but a new economic argument was added to assert that modern colonialism was inevitably accompanied by the economic exploitation of the colony, and hence, political sovereignty was a necessary condition for the removal of poverty and economic backwardness.

By the time of World War II, anticolonial nationalism had become a significant political force in many countries of South and South-East Asia, the Middle East and North Africa. While aspiring to fulfill the project of modernity in their own countries, anticolonial nationalist leaders also launched powerful movements of social and cultural reform, but away from the supervision of the colonial state, attempting to assert their sovereignty over an inner sphere of national life even before the political battle over the economy and the state had been won (Chatterjee 1986, 1993). This was the period when standard national languages were institutionalized and national literatures created and canonized, national histories constructed and archives assembled, and religious doctrines and social practices submitted to a thorough modernist scrutiny and reform. The political process of decolonization became rapid and unstoppable after the end of World War II. In many colonies of Asia, Africa, and the Caribbean, independent states were created through negotiated transfers of power; in others, national liberation was won after armed struggle. It was often argued that independence achieved by peaceful negotiation produced socially conservative and politically reformist regimes, whereas when it was won through armed struggle, the structures of colonial rule, themselved founded on violence, had a greater chance of being socially and politically transformed (Fanon 1963).

6. *Nation and Culture*

One of the consequences of the new theoretical understanding of nationalism, introduced especially by Benedict Anderson, has been the explosion of scholarly interest in the deep involvement of nationalism in areas of culture that apparently have little to do with the political contest over nation-states. The histories of language, literature, historiography, art, music, theater, cinema, sports, and such other modern forms of cultural production are being examined for their many connections with nationalist projects—in their formal properties, their changing contents, and in the institutions that promote and foster them. Alongside, there has emerged a new literature on family, gender, sexuality, religious practices, food, clothes, and other aspects of everyday life that can be shown to have been shaped and transformed by nationalist cultural projects. A recurrent theme in many of these studies is the active role of the nationalist imagination in the gendering of numerous institutions and practices of modern life. Thus, a common feature is the identification of the inner, the domestic, and the spiritual as the domain of the feminine and that of the outer, the public, and the material as that of the masculine. Another is the imagining of women as the bearers of national tradition. There are interesting variations in these structural relations of gender in the case of anticolonial nationalisms (Sangari and Vaid 1988, Abu-Lughod 1998).

The other aspect of the relation between culture and nationalism that has received much attention in recent years is that of ethnic politics or, in some descriptions, the politics of identity. A distinction has been made, most notably by Anderson (1998), between nationalism in its classical version and the politics of ethnicity. The former, he argues, imagines community in the form of 'unbound serialities,' i.e., the everyday universals of modern social thought such as nations, citizens, workers, bureaucrats, intellectuals, etc. This form of imagining of universal community is potentially liberating. Ethnic politics, on the other hand, is produced by the 'bound serialities' of governmental classification, such as censuses. They operate by integers—one can be counted only as one or zero, for example, Black or non-Black, Muslim or non-Muslim, German or non-German, and all partial or mixed affiliations must be ruled out. Such serialities are constricting and inherently conflictual. Others argue that the supposed distinction between classical nationalism and ethnic politics is once again a version of the liberal distinction between good and bad nationalism and that they represent not so much a difference of type as the inadequacy of the classical nation-form to properly represent within itself the full range of aspirations for community in modern political life.

It has also been said that the apparent proliferation of the politics of identity, occurring in the era of economic globalization, is a symptom of the historical demise of the nation-state. The breaking down of trade barriers between nations, the global spread of the new communications technology, and the unprecedented flow of international migrants are making the nation-form largely redundant (Appadurai 1997). This is accompanied by the increasing assertiveness of a global civil society that often considers the sovereignty of existing nation-states as impediments to the realization of universal ethical goals such as human rights or development. These conditions, the argument goes, are producing new postnational formations that may well supplant in future the nation-state as the normal form of the modern political community.

See also: Colonialism, Anthropology of; Colonialism: Political Aspects; Colonization and Colonialism, History of; Globalization and World Culture; Nation: Sociological Aspects; National Character; National Liberation Movements; Nationalism: Contemporary Issues; Nationalism, Sociology of; Nations and Nation-states in History

Bibliography

Abu-Lughod L 1998 *Remaking Women: Feminism and Modernity in the Middle East*. Princeton University Press, Princeton, NJ

Anderson B 1991 *Imagined Communities*, rev. edn. Verso, London

Anderson B 1998 *The Spectre of Comparisons*. Verso, London

Appadurai A 1997 *Modernity at Large*. University of Minnesota Press, Minneapolis, MN

Armstrong J A 1982 *Nations Before Nationalism*. University of North Carolina Press, Chapel Hill, NC

Balibar E 1993 *Masses, Classes, Ideas: Studies on Politics and Philosophy Before and After Marx*, Swenson J trans. Routledge, New York

Chatterjee P 1986 *Nationalist Thought and the Colonial World: A Derivative Discourse?* Zed Books, London

Chatterjee P 1993 *The Nation and Its Fragments: Colonial and Postcolonial Histories*. Princeton University Press, Princeton, NJ

Fanon F 1963 *The Wretched of the Earth*. Grove Press, New York

Gellner E 1983 *Nations and Nationalism*. Blackwell, Oxford, UK

Greenfeld L 1992 *Nationalism: Five Roads to Modernity*. Harvard University Press, Cambridge, MA

Hobsbawm E 1990 *Nations and Nationalism since 1780*. Cambridge University Press, Cambridge, UK

James C L R 1963 *Black Jacobins*. Random House, New York

Kohn H 1944 *The Idea of Nationalism*. Macmillan, New York

Nimni E 1991 *Marxism and Nationalism*. Pluto, London

Sangari K, Vaid S (eds.) 1988 *Recasting Women*. Kali for Women, New Delhi, India

Smith A 1986 *The Ethnic Origins of Nations*. Blackwell, Oxford, UK

Tamir Y 1993 *Liberal Nationalism*. Princeton University Press, Princeton, NJ

Trouillot M-R 1995 *Silencing the Past: Power and the Production of History*. Beacon Press, Boston

P. Chatterjee

Nationalism, Historical Aspects of: Africa

Most analyses of the phenomenon of nationalism in Africa were made recently, in the period after independence was achieved. Studies focus on nationalist processes comparable to those which have taken place in Europe, most of them starting in the nineteenth century. The analyses usually begin with the era when emerging states started to demand international recognition, as other nations had done earlier. Only recently have a few political scientists and historians given a modest amount of attention to a long history of nationalist movements in Africa (Davidson 1992, Berman and Lonsdale 1992). Others have stressed the catalytic role played by the colonial phase (Ranger 1968, Johnson 1971, Mamdani 1996). But Africanist literature has focused primarily on 'ethnic' problems, portraying them as quite different from—even antagonistic to—a 'nationalist' approach. Until recently, no investigator had pointed out that ethnicity can be a function of nationalism as much as nationalism was derived from ethnic awareness (Anderson 1983, Berman and Lonsdale 1992, Turner 2000). The standard view must be considerably revised, because it is the offshoot of a long tradition of biases inherited from the colonial past of ethnographic research (Hobsbawm and Ranger 1983, Amselle and Mbokolo 1985, Coquery-Vidrovitch 1995).

African political history began very early, several centuries before the colonial episode. Therefore, we designate endogenous nationalism antedating colonization as 'older nationalism.' We will see that in precolonial days, what we could regard as legitimate small 'nation-states' did indeed exist, although they were not founded on the same presuppositions as in the Western world. These preexisting elements were the foundation on which the process of 'modern' nationalism was, more or less convincingly, grafted in the form of independence as defined by international settlements and the United Nations.

With minor exceptions, the African continent has remained independent until very late in history, at least vis-à-vis European powers. Certain limited points on the coast were occupied at a very early date: the port of Loanda in the west and Mozambique Island in the east by the Portuguese; the port of Mombasa on the Kenyan coast, disputed among the Arabs and Portuguese from the sixteenth century; the Senegal island of Saint Louis at the mouth of the river by the French; and Cape Town by the Dutch, starting in the middle of the seventeenth century. A series of coastal outposts was taken in the seventeenth and eighteenth centuries by the British, including Cape Coast Castle on the Gold Coast and, above all, the small Crown colony created in 1807 at the tip of the Sierra Leone Peninsula to receive slaves 'emancipated' from ships engaged in smuggling them. However, at least until the 1830s—when the conquest of Algeria by the French began and the occupation of the interior plateaux of southern Africa was commenced by the Boers, soon followed by the British on the coast of Natal (1844)—almost the total population of African peoples lived in independence, except for a conglomeration of east coast territories dominated by the Sultanate of Oman and Zanzibar.

1. Older Forms of Nationalism

In the past, we meet two forms of 'nationalism': the erection of actual small nation-states built over a long course of time around a well-defined plan, and the irruption of conquest states, which hardly had the time to establish corresponding nations and instead triggered revolts of preexisting nationalities.

These conquest states resulted from a series of jihads (holy wars of Islam), which succeeded one another with no interruption in the northern hemisphere from the middle of the eighteenth to the last third of the nineteenth century. They did not arise from nationalism in the usual sense of the term, but they forged a strong cultural bond. This was established in western Africa around the Pular culture (Robinson 1985). Futa Djallon is a good example. In this mountainous region containing the sources of the Niger, Senegal, and Gambia rivers, the Muslim Fulbe (advocates of slavery, including shepherds, farmers, and scholars) gained the upper hand over local farmers and indigenous religions. Locally, the main innovation of the Fulbe was using the Arabic alphabet to transcribe the literature of the Pular language. This led to the popularization of a culture based on educated people who had an awareness of solidarity. Pular migrations all over West Africa helped propagate their feeling of belonging to a common culture. The Pular community, founded by Usman dan Fodio in 1804 in northern Nigeria, was thus both an ideological and an economic venture. Its specificity has survived to the present day. The community marvelously managed to resolve the problem common to all Pular theocracies: how to convert a tolerated minority religion, only adopted by chiefs when it seemed to them to be useful, to a religious and imperial state ideology; how to turn a small, strict, geographically isolated Muslim group into a driving force to obtain recognition of western Sudan as Dar Al Islam.

2. Precolonial Nation-states

Several animist or Muslim states were constituted as a nation much earlier. The oldest examples of these are ancient Egypt and, to a certain extent, Ethiopia from the beginning of one era. North Africa was reconfigured by the Ottoman Empire: Morocco, Algeria, and Tunisia already had their borders long before the nineteenth century; the first as an independent country, the other two as Turkish provinces. Tunisia and Morocco became protectorates (in 1883 and 1912), Algeria a French *département*. Nationalism was molded in preexisting frameworks in conjunction with the anticolonial struggle, although it triggered reactions in the form of regionalisms founded on linguistic and cultural claims. Berber speakers, vigorously repressed by speakers of Arabic in contemporary Algeria, are a case in point.

Outside Islam, the west coast of Africa was experiencing the proliferation of small nation-states at least since the sixteenth century—the term is not exaggerated, as Basil Davidson (1992) has demonstrated, *mutatis mutandis*—if we take this to mean political formations of people who have shared a common history and most often the same language for several generations, or even several centuries and, recognizing the same political system under the rule of a sovereign or chief acknowledged by all, and working to a greater or lesser extent under this authority in the same economic network of productive and commercial activities. Since Benedict Anderson (1983), we do not hesitate to speak of the genesis of nationalism outside the Western world.

The emergence (toward the end of the seventeenth century), the rise (from 1760 to 1840), and the decline (in the second half of the nineteenth century) of small political organizations focused on slaving—coinciding chronologically with the parallel evolution of the Atlantic slave labor trade, which itself developed simultaneously with the expansion of sugar cane plantations from Brazil to the Caribbean—was no mere coincidence. The establishment of many political organisms correlates closely with the era's dominant international market, at least until the beginning of the nineteenth century. The system combined war and trade. During each dry season, the army foraged for prisoners to supply objects of the slave trade. During the rainy season, the soldiers became peasants again, coming back home to toil their lands working in subsistence agriculture.

The kingdom of Abomey (in today's Benin, formerly Dahomey) established a strong political community with a common language and culture, and each generation of children of slaves born on Dahomean soil became Dahomean. This national unity explains why the conquest of Dahomey by the French (1890–1894) was the toughest colonial war and the longest in the region. The people of Dahomey gave mass support to the resistance put up by the last king, Behanzin. By the same token, in the heart of present-day Ghana, the Ashanti Empire, annexed in 1897, did not accept British sovereignty until the Crown resolved, in 1928, to recall the Asantehene and reinstate the sovereign on his Golden Stool, the symbol of national unity.

We could go on and on citing cases of small nation-states which took shape from the eighteenth century on (Coquery-Vidrovitch 1999). Other kingdoms appeared during the same period in Central Africa, in the rift lakes region of Rwanda, Burundi, Ankole, and Buganda, where a long collective history had endowed the inhabitants with a common language and, consequently, a common culture. Nationalist excesses in the form of ethnic purges are merely a recent manifestation of a long-standing demographic and political disequilibrium aggravated by the traumas of colonial and postcolonial history.

At the turn of the nineteenth century, the small, militarized Zulu kingdom of Shaka appeared in the northern part of what is now Natal. This realm never submitted to British domination. The last revolt there erupted in 1906. Using this glorious past to justify their political demands paved the way for young militant Christians to found the nationalist Inkatha Party in the early 1920s.

3. Genesis of Modern Nationalisms

Militant nationalists of the early years of independence attuned their brand of emerging nationalism to the numerous revolts which had developed nearly everywhere during the colonial period. Nevertheless, we should not confuse what have been called 'primary resistance movements' with modern nationalism (Ranger 1968). Ranger postulated that, slowly but surely, inhabitants of a territory delimited by a colonial power will accept this entity and claim possession of it. By now this 'nationalization' has taken place almost everywhere, despite often poorly understood reactionary movements. A contrasting example is Eritrea, which won its independence for the very reason that Eritrean nationalism has been constructed for a century outside Ethiopia. Ethiopia triumphed in 1896 over the Italian invader, while the province of Eritrea became an Italian colony in 1890, then, as a result of World War II, a British protectorate from 1940 to 1960. The Eritreans never recognized their territory's reversion to Ethiopia, arranged by Western negotiations in 1960.

In sub-Saharan Africa, nationalism engendered by colonialization did not always appear where preexisting nation-states were confirmed. An example is the Fante nationalism, which developed in the second half of the nineteenth century on the Ghanian coast. Fante communities which had previously existed constituted a sprinkling of small districts, united solely by language and their occupations because they had long been agents of European merchants on the coast and on privileged terms with the English. Exposed to Western influences more than the rest of the population, they had created a Christianized Creole culture, which made them early adherents of political modernism and aligned them against European intervention from the start. The movement was triggered by British interference starting in 1843. The chiefs, dispersed until then, reacted in 1852 by sanctioning an internal confederation inspired by Western ideas. The Fante Confederation (1867–1873) provided for a constitutional regime, with a national assembly of chiefs and an executive parliament. These ideas were propagated by the publication of *West African Countries and People* by the Sierra Leonean physician James Africanus Horton. From 1865 on, to the dismay of the British, Horton advocated the idea of African independence, arguing that the country was more advanced than Liberia 20 years earlier.

4. A Moderate Colonial Protonationalism

Beginning at the end of the nineteenth century, new economic and social conditions fostered by the conquest provoked an anticolonial reaction of the 'evolved' victims of colonialization. These protests had the paradoxical effect of reinforcing their territorial cohesion within borders which were frequently conceived and imposed entirely by outsiders.

The initial militant nationalists were graduates of mission schools. The first of these in Sierra Leone were descendants of freed slaves, subsequently known as Krio, which describes both a language and a culture. Others were natives of the Ghanian coast and the mouth of the Niger, often scions of historically dominant classes. A prime example is the Brew clan, founded by an eighteenth-century Irish slaver. His African descendants devoted themselves to trading—first to the slave trade and, when that became impossible, to export commodities. In the second half of the nineteenth century, when British competition was making grave inroads thanks to colonialization, these cultivated bourgeois still sent their children to study in England, or at least to mission schools on the coast. They bred lawyers and journalists. A dozen of these intellectuals earned PhDs. One Brew was a member of the delegation the Fante sent to Queen Victoria in 1895, when they still believed it was possible to set up a less closed political system than colonial law and order (Priestley 1969). The next year, three Tswana chiefs traveled from the other end of the continent to London to try to persuade the queen to help their country preserve its independence.

5. South Africa

African nationalism manifested itself most conspicuously in South Africa. We have already mentioned Zulu regionalism. Another important point is the length of the so-called Xhosa Wars, expressions of resistance which flared up periodically from the end of the eighteenth century to the end of the nineteenth century, reflecting the maturation of a Xhosa awareness among peoples inhabiting Eastern Cape Province. But it was the injection of mining capital (diamonds were discovered in 1867, gold in 1886) that elicited modern forms of black nationalism in the Rand compounds and in Johannesburg. The African National Congress (ANC), ever since the main protagonist of South African nationalism despite several periods of severe repression, was founded in 1912. From the start, militant South African nationalists were radicals and often had close clandestine ties to the Communist Party, but most of them were graduates of institutions of higher learning created from the middle of the nineteenth century by various Christian missions (Lovedale, Healdtown, St. Matthew). The first militant feminists appeared on the scene in South Africa and West Africa even before the 1950s. They

included Ray Alexander (1941), Ida Mtwana, who assumed leadership of the Women's League of the ANC in 1947, and Liz Abraham, who is renowned for her struggle against apartheid (Walker 1982).

6. *Evolution From the Years Between the World Wars Until Decolonialization*

It was not until the twentieth century, between the two world wars, that nationalist demands really began to be voiced. This is because severe censorship was never loosened by the French, except by the Front Populaire government (1936). In British Africa the turning point was not until the World War II, when trade unions were first permitted, even before non-colonial political parties. Although its way of life and thinking was more akin to European intellectuals than to the people in local rural communities, the young African bourgeoisie did not wait until granted civil rights before demanding, sometimes vigorously, that it be allowed to participate in public affairs. The spearhead of the movement was British West Africa, first centered on Sierra Leone, where an indigenous press developed at an early stage. In the French colonies, political life went on in classical French style in the 'Four Communes' of Senegal (Saint-Louis, Gorée, Rufisque and Dakar), whose inhabitants became French citizens in 1916 (Johnson 1971). The colony of Dahomey was also a hotbed of protest. There, a merchant bourgeoisie developed alongside the low-ranking civil servants with a French education. It had connections to prominent Afro-Brazilian families and to aristocratic circles of the former kingdoms of Abomey and Porto Novo, and it possessed sufficient financial resources to launch local publications. *La voix du Dahomey*, a newspaper founded in 1928 and existing into the 1950s, was its principal mouthpiece. The colonial government was unable to ban the paper, despite a highly publicized trial during the Front Populaire administration.

Nationalists expressed a desire to take the reins of government into their own hands in British-ruled West Africa at an early date. During the 1920s, they obtained the right to participate in municipal government to the extent of voting in local elections. In French colonies, nationalist aspirations were offset for a long time by the French model of assimilation tried out in Senegal: acquiring the citizenship of the colonial power was a way of obtaining rights, based on equality and individual freedom, i.e., everything normally denied to natives, even assimilated groups. In both cases, the political ideal remained reformist. It did not challenge the colonial system. The idea was to fight colonial abuses and to collaborate with the system. This collaboration became a fact after the World War II. Virtually alone, Dr. Danquah demanded the status of a dominion at the Gold Coast Youth Conference in 1940. His platform was still moderate. He advocated the democratization of the legislative council, which then only had nine Africans among the 29 elected members. He proposed a two-chamber system modeled after the British Parliament, one elected and the other composed of chiefs, and he demanded that Africans be granted the right to apply for civil service jobs.

Models, institutions, and trappings of the colonial government fascinated the future leaders of the newly independent states for a long time (Werbner 1998). Anticolonial radicalism was limited to a very small minority of young Communists, whose militancy had little chance to express itself in Africa. The initial impulse came from George Padmore, a Marxist from Trinidad, who founded the Negro Workers Bureau in Hamburg and contacted Lagos militants, undoubtedly thanks to dockers in the port. It was virtually impossible in French Africa and not much easier in Sierra Leone, where the most radical militant was the journalist Wallace Johnson, who founded the Sierra Leone Youth League (SLYL) on the eve of the war.

The climate changed only slowly after the World War II. In the French sphere, the nationalist ideas of the elite were sublimated under the cult of Negritude, which was less of a political movement than an ethical ideal. This concept, launched in 1939 by the Antilles poet Aimé Césaire, author of the explicit indictment *Discours sur le colonialisme*, was adopted for the most part after 1947 by the Senegalese poet and politician Léopold Sédar Senghor. In British Africa, intellectuals sued for divorce earlier. They were, nevertheless, likewise steeped in Western culture (the Ghanian Kwame Nkrumah and the Nigerian Azikiwe, future statesmen, had left on the eve of the war to study in the United States, and in 1938 the Kenyan Jomo Kenyata wrote the first thesis in anthropology on his country). But since in the 1950s the British were not yet prepared to budge, the nationalists hardly had any choice but to accept 'multiracialism': one 'race,' one vote, or in Kenya the same number of seats for each of three groups—Whites, Indians, and Blacks—which were nowhere near equal in number.

7. *The Role of National Liberation Movements*

Much has been written about the Mau Mau rebellion, the 1952–1955 brush fire that swept most of Kenya. It has been argued that the movement was primarily a Kikuyu affair. But the severity of the repression and the blindness of the colonial authorities, whose minds were set in the belief that opposition politicians were the main instigators, contributed to cementing a national, anticolonial sentiment which forced Britain, although victorious on the field of battle, to grant independence (1963). Berman and Lonsdale have written a superb work on the subject (1992).

The anticolonial struggle also formed a national awareness in former Portuguese colonies. After inde-

pendence (1975), resistance against apartheid South Africa and white Rhodesia sufficed to forge a nation state in Mozambique, a more than 3,000-kilometer-long territory which was very disparate at the outset. However, in Angola, the southern Savimbi regionalism was constructed on a partially real, partially manipulated precolonial Ovambo past. This is the case with many contemporary tribalist manifestations; they are exploited primarily for political ends by ambitious, unscrupulous manipulators. For the past 15 years or so it has been demonstrated that one should not lump ethnic group and nation together (Hobsbawm and Ranger 1983, Amselle and Mbokolo 1985). But contemporary propaganda resuscitates strong sentiments handed down from the past—for example, the resentment of ravaged peoples against the nations which ravaged them, going back to the black slave trade, as was the case in Angola, or that of social groups dominated and exploited by large landowners of the period, as in the case of the Hutu and Tutsi in Rwanda.

Are these regionalisms of today tomorrow's nationalisms? This is unlikely, because in most of the examples we have cited all sides claim the entirety of the country involved. Once again, the struggle against foreign power sparks nationalism. A typical case is the Congo (former Zaire), where international observers forecast an inevitable explosion. Ironically, it was the widely hated dictator Mobutu who sanctified the national unity of the former Belgian Congo in the name of the national hero and founding father Patrice Lumumba (Turner 2000).

The risks of regionalist explosions seem to be less of a threat today than do conflicts arising from a chauvinistic approach to nation-states that have become established despite arbitrary delineation imposed more than a century ago (Bayart 1996, GEMDEV 1996). The recent Ivory Coast conflict based on an exclusive, xenophobic *Ivoirité* is one of the most ominous signs.

See also: African Studies: Culture; African Studies: History; African Studies: Religion; Central Africa: Sociocultural Aspects; Colonization and Colonialism, History of; East Africa: Sociocultural Aspects; National Liberation Movements; Nationalism: Contemporary Issues; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Nation-states, Nationalism, and Gender; Postcoloniality; Southern Africa: Sociocultural Aspects; West Africa: Sociocultural Aspects

Bibliography

Amselle J-L, Mbokolo E 1985 *Au Coeur de l'Ethnie*. Decouvert, Paris

Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Bayart J-F 1989 *L'État en Afrique: La Politique du Ventre*. Paris, Fayard. Trans. 1993 *The State in Africa: The Politics of the Belly*. Longman, London

Bayart J-F (ed.) 1996 *La Greffe de l'État*. Karthala, Paris

Berman B, Lonsdale J 1992 *Unhappy Valley: Conflict in Kenya; 1. State and Class; 2. Violence and Ethnicity*. James Currey, London

Coquery-Vidrovitch C 1995 De la Nation en Afrique Noire. *Le Débat* **84**: 121–38

Coquery-Vidrovitch C 1999 *L'Afrique et les Africains au 19è Siècle*. Colin, Paris

Davidson B 1992 *The Black Man's Burden: Africa and the Curse of the Nation-State*. Times Books, New York

GEMDEV (ed.) 1996 *Les Avatars de l'État en Afrique*. Karthala, Paris

Hobsbawm E, Ranger T 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Horton J A 1868 *West African Countries and People*. London

Johnson G W 1971 *The Emergence of Black Politics in Senegal: The Struggle for Power in the Four Communes, 1900–1920*. Stanford University Press, Stanford, CA

Mamdani M 1996 *Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism*. Princeton University Press, Princeton, NJ

Priestley M 1969 *West African Trade and Coast Society: A Family Study*. Oxford University Press, London

Ranger T O 1968 Connections between primary resistance movements and modern mass nationalism. *The Journal of African History* **9**(3): 437–54; **9**(4): 631–42

Robinson D 1985 *The Holy War of Umar Tal: The Western Sudan in the Mid-nineteenth Century*. Clarendon Press, Oxford, UK

Turner T 2000 *Ethnogénèse et Nartionalisme en Afrique Centrale: Aux Racines de Patrice Lumumba*. L'Harmattan, Paris

Walker C 1982 *Women and Resistance in South Africa*. Onyx Press, Cambridge, MA

Werbner R 1998 *Memory and the Postcolony: African Anthropology and the Critique of Power*. Zed Books, London

C. Coquery-Vidrovitch

Nationalism, Historical Aspects of: Arab World

For almost a century, Arab nationalism was one of the major political currents in the Arab world; in the 1950s and 1960s, it was undisputedly the dominant one, albeit never uncontested. Like in other parts of what was then still called the Third World, its intellectual origins in Europe did not hinder its spread and success as a principal source of anti-imperialism. Like other nationalisms, it covered a wide range of ideas, images, and aspirations that appealed to and were supported by different social groups and strata; it built on existing self-views and affinities, forming various blends of local, regional, and transnational allegiances, reflecting different shades of religious and secular orientation. Critical scholarship has absorbed key assumptions of the theoretical literature on nationalism and ethnicity more generally, which can be

read as a refutation of, or counter discourse to, nationalist claims and convictions: it stresses the constructedness of Arab like any other identity or identification; its hybrid character rather than the purity of race, culture, language, and design nationalists insist on; fragmentation rather than the unity of intent, action, and objective nationalists assert; fuzziness rather than the clarity of program, boundaries, and orientation they proclaim (Jankowski and Gershoni 1991, Khalidi et al. 1991, Gelvin 1998). Arab nationalism thus emerges as multifocal, multifaceted, and multilayered. Measured against the level of theoretical sophistication, knowledge of the historical evolution of Arab nationalism remains highly uneven, and in several respects quite inadequate.

1. *Contingency and Continuity*

There was no unilinear movement, or organic growth, from cultural revival (Arabism) to Arab nationalism first in greater Syria and then the remainder of the Arab world, to pan-Arabism and Arab unity. From the beginning, Arab activists interacted, and competed, with advocates of alternative frames of reference that emerged simultaneously in the Middle East and North Africa: if Ottomanism died with the collapse of the Empire, territorial nationalism and political Islam found equally strong resonance among the area's population. While in the Middle East, as in most other parts of the world, there existed a tradition of local pride and attachment that was also reflected in a specific literary genre (*fada'il*) listing the beauties, monuments, and personalities of a specific town or region (*watan*, comparable to the German notion of *Heimat*), this sense of belonging did not necessarily translate into patriotism or nationalism, be it defined in terms of territory (*wataniyya*) or a people/nation (*qawmiyya*). Arabism never constituted more than one option of cultural and political orientation. To insist on choice and contingency, however, cannot imply denying all kinds of continuity between Arabism, Arab nationalism, and pan-Arabism.

1.1 *Arabism*

The intellectual and social roots of Arabism can be traced back to the second half of the nineteenth century, when in the Arab provinces of the Ottoman Empire, more particularly in certain cities of Egypt and the Fertile Crescent, small circles of scholars and intellectuals began to call for a revival or renaissance (*nahda*) of the Arab language, literature, and culture. The Nahda, which in later nationalist writing was represented as an 'awakening' of the Arab nation (*al-umma al-arabiyya*), or of 'Arab consciousness' (*al-wa'i al-arabi*), involved religious scholars, writers, journalists, government officials, and members of the liberal professions, reflecting the spread of education, literacy, and new means of communication, first and foremost an Arab press, that resulted in the creation of a public sphere ranging from clubs and associations of all kinds to salons, tea and coffee houses, theaters, etc. The revival of Arab(-Islamic) culture and literature (in written standard Arabic, *fusha*, not the vernacular used by the vast majority of Arabic-speakers) was to be combined with modern knowledge coming from Europe. While new literary forms gradually developed (novels, theater plays, later films), poetry retained its central role for cultural and political expression in the Arabic language. The long-held view (e.g., Antonius 1938) that Arab cultural revival was largely the work of Arab Christians who, through bonds of faith and their education in European schools had been more receptive to European thought including nationalism as it was embraced by their coreligionists in the European provinces of the Ottoman Empire, has since been challenged and revised (Khalidi et al. 1991, Tauber 1993, Nafi 1998). The prominent role of Christians in the early Arab press cannot be denied. However, education on modern lines was not restricted to foreign schools, but equally adopted at Ottoman state schools and private institutions that were established by members of the local elites, Muslims, Christians, and Jews. Most important, Arab cultural revival was at that stage intimately linked to Islamic reform and maintained its link to Islam even later. A second view has been modified (Kayali 1997): that Arabism evolved as a reaction to a Turkification drive pursued by the Young Turks after their 'revolution' of 1908, resulting in the marginalization of the Arabic language in official use and of Arab delegates in Ottoman representative bodies. The use of Turkish in Ottoman courts, schools, and government offices had been made obligatory since the 1870s. The Young Turks merely accelerated and intensified the centralization policies of the Ottoman reform era (Tanzimat), and as a result Turkish was more widely used in the Arab provinces of the Empire. At the same time, Arabs became less well represented in parliament, the higher echelons of the bureaucracy, and at court. With the decline of a legal and educational system that was based almost exclusively on Islamic learning, whose normative texts were in Arabic, knowledge of the 'sacred language' lost much of its prestige and usefulness for a career in the Ottoman public service. Modernization, then, contributed in more than one way to the rise of Arabism and, later, of Arab nationalism.

Arab political societies that from the 1880s onwards, formed and met in Damascus, Baghdad, Beirut, Istanbul, or Paris mostly worked for greater autonomy within the Ottoman Empire, for constitutional reform or decentralization, not for political independence (Tauber 1993, Nafi 1998). Geographically, their horizon was limited to 'greater' or 'natural Syria' (*bilad al-Sham*); Egypt with its distinctive tradition and emerg-

ing national movement, Sudan, the Maghreb, and the Arab Peninsula, Hijaz excepted, were not part of their vision. Membership in secret societies such as al-Ahd (The Covenant, 1909), al-Fatat (The Young, ca. 1911) or the Ottoman Administrative Decentralization Party (1912) was small; an Arab Congress held in Paris in 1913 was attended by some 25 delegates. Similar to Iranian, Turkish, or Russian opposition groups of the same period, they were made up of urban intellectuals and military officers. In contrast to the literary clubs and associations that also included some women, or addressed themselves to a female audience, they appear to have been exclusively male. Their impact on local society and politics was slight. Still, the secrecy surrounding them and rumours of broad popular support alerted official circles in Istanbul at a time when nationalist feeling was rising not only in the Balkans and among other Christian subject populations like the Armenians, but also in predominantly Muslim territories such as Egypt.

1.2 Arab Nationalism

During World War I, the vast majority of Ottoman subjects remained loyal to the Empire. Separatism remained a minority trend among its Muslim population: in 1915, Arab activists in Damascus contacted the Sharif of Mecca, Husain ibn Ali, a descendant of the Prophet Muhammad in the Hashemite line and *de facto* of the holy city, who, after some hesitation, signaled his willingness to rebel, in the name of Islam, against the sultan-caliph. In a separate move, Husain obtained British declarations of support for Arab independence under Hashemite leadership in exchange for military revolt. The Ottoman authorities reacted harshly to all stirrings of autonomous political activity including Arabism, which under the circumstances amounted to high treason, and in 1915/16 hanged dozens of activists, creating the first martyrs to the Arab cause. In June 1916, Husain declared the Arab Revolt and after the fall of Medina in November 1916, proclaimed himself 'king of the Arab lands.' In the course of 1917/18, British, Arab, and Jewish forces jointly conquered Palestine and Syria. In October 1918, an 'Arab government' was installed in Damascus under Faisal ibn al-Husain, one of the Sharif's sons, who with a group of Arab nationalists of diverse origins, attempted to take control of a country ravaged by war, famine, and disease. In March 1920, a Syrian General Congress representing Arab nationalists from Syria, Palestine, and Lebanon proclaimed the independence of 'greater Syria' with Faisal as king. Within Syria proper, their policies were partly assisted and partly resisted by urban, tribal, and village-based popular committees defending their own understandings of collective interest (Gelvin 1998).

Britain and France were in the meantime taking measures to implement their different wartime arrangements, most of them secret, which included not only the correspondence with the Sharif, but also the Sykes–Picot agreement of May 1916 regarding their respective spheres of interest in the Arab provinces of the Ottoman Empire, and the Balfour Declaration of November 1917 promising British support for the establishment 'in Palestine' of a 'national home' for the Jewish people. The terms of these commitments had somehow to be reconciled; Britain and France did so at the expense of their Arab allies. As a result of World War I, more Arab lands were under European control than ever before, be it as provinces (Algeria), colonies (Aden), protectorates (Morocco, Tunisia, Egypt, Kuwait), or mandates (Lebanon, Syria, Palestine, Transjordan, Iraq) under the newly created League of Nations. Only the Zaidi Imamate of Yemen and the Saudi kingdom in Najd and Hijaz gained or retained full independence. The war and its aftermath defined some of the themes that were gradually woven into the 'grand narration' of Arab nationalism: the crucial role of 'martyrs' for national liberation; the 'Great Arab Revolt' or 'revolution' (*thawra*) of 1916–18 (to be followed by similar 'revolutions' in Egypt 1919, Iraq 1920, Syria 1925–27, the Egyptian officer coup of July 1952, as well as a series of 'corrective revolutions' in the 1970s and 1980s); exposure to 'conspiracies' and intimately linked to it, 'betrayal' of the Arabs by the West.

Within the framework of the emerging Arab state system, Arabism evolved into a political movement of growing impact and complexity, providing one of the major sources of anticolonial struggle and sentiment. The interwar period witnessed the rise of new social and political forces alongside the urban notables that still dominated politics and society (at least outside tribal areas): new urban middle class parties that could not be considered mere covers for elite families or factions; paramilitary youth movements inspired by the fascist model; Islamic movements of social and political reform; socialist and communist groups, some with links to trade unions, of various backgrounds and affiliations—all of them responding to Arab references, inflecting their emphases, adapting or rejecting them. If there was a hegemonic formulation of (politicized) Arabism (*al-uruba*) it was provided by *Sati al-Husry* (1880–1968) who as a high government official was able to influence public education in Iraq (then still under British mandate) (Cleveland 1971, Tibi 1981). Visibly influenced by the German romantic school, and Herder in particular, Husry defined the Arab nation on the basis of language, history, and destiny, embedding his vision in a narrative of pristine glory, foreign-machinated decline and eventual awakening, with a heavy emphasis on the us/them dichotomy, 'them' being the Arabs' enemies and oppressors from the Byzantines and Crusaders to the Ottoman Turks, Zionist Jews, and European colonial powers. The criteria of inclusion and exclusion remained malleable and contested: Arabness could be inter-

preted in secular terms to include all men and women of Arabic language and culture regardless of their religious or ethnic background, and thus provide the basis of common nationhood and citizenship; but it was impregnated so strongly with references of a specifically Sunni Muslim nature that it could also be given a distinctly religious (Sunni Muslim) coloring. At the same time it contained enough references to the superiority of the Arab race and genius to make it possible for Kurds, Berbers, or Jews to be marginalized, if not entirely excluded, from the national community.

1.3 Pan-Arabism

After World War II, two dimensions were added to nationalist discourse and strategy: Arab socialism addressing the issue of how to organize the growing number of independent Arab states, and pan-Arabism (*al-qawmiyya al-arabiyya*) aiming to transcend the existing regional order. Previous projects of political union had been associated mostly with members of the Hashemite dynasty, who, after losing Syria to the French and Mecca, Medina and the rest of Hijaz to the Saudis, reigned in Iraq and Transjordan; and they had been restricted to the Fertile Crescent and Arabia (Porath 1986). The 1930s saw the gradual diffusion of Arab identification to other parts of the Middle East and North Africa. The conflict over Palestine, which in 1936–39 erupted in the Arab revolt or 'revolution,' contributed to widening the sense of a shared Arab destiny. Intellectuals, scholars, and politicians in Egypt, Sudan, and the Maghreb who had not defined themselves previously in terms of Arabness, began to increasingly add an Arab dimension to their agenda, or even adopt pan-Arabism as their ideology. In the late 1940s, the Arab Socialist Ba'th Party presented the new and enlarged credo of Arab nationalism, revolving around three concerns: unity (*al-wahda*) of the Arab nation both internally and across the existing borders that had been 'artificially' created by the colonial powers and had to be eliminated to create one single Arab state stretching from the Indian to the Atlantic Ocean (*min al-nahr ila l-bahr*); freedom (*al-hurriyya*) from foreign rule and reactionary feudal oppression at home; (Arab) socialism (*al-ishtirakiyya*) which was to be distinguished from Marxism or communism by its belief in God and the rejection of class struggle (Abu Jaber 1966). The Arab League founded, with British support, in 1945, could at best offer partial fulfillment of this vision: it worked for political, economic, and cultural cooperation among existing Arab states, not for their unification, and it did not adopt Arab socialism (Maddy-Weitzmann 1991).

In the course of the 1950s, Syria, which until then considered itself the 'beating heart of Arabism,' was overshadowed by Nasserist Egypt as self-proclaimed leader of the Arab nation. The call for Arab unity and solidarity evoked extraordinary responses from broad sections of the population all across the Arab world (the Arab 'masses' constantly evoked in Arab nationalist rhetoric), testifying to the power nationalism, and Arab nationalism in particular, had acquired regardless of its foreign origins. Appeals to Arab solidarity had the obvious function of legitimizing, *inter alia*, a 'fair share' in 'Arab' oil wealth. The politics of Arab unity equally bespoke the will to control and interfere, in the name of shared Arab concerns, in what rivaling national elites viewed as their domestic affairs. The heyday of pan-Arabism (ca. 1956–67) was also a time of relentless competition among Arab states and intrastate actors (Kerr 1971). All experiments of merger and fusion (some of them, like Iraq's incorporation of Kuwait in 1990, based on military conquest) failed, most notably the union between the two champions of pan-Arabism, Egypt and Syria, that only lasted for three years (1958–61). The failure to solve the Palestine issue, epitomized in the Arab defeat of 1967, was therefore not the only factor weakening pan-Arabism as a political movement of mass mobilization. Yet while, from the 1970s, the 'end of pan-Arabism' was proclaimed repeatedly (Ajami 1981), not least by the adherents of political Islam and many foreign analysts, reference to Arab identity, interest, and solidarity continued to play a significant role in Middle Eastern politics even at the beginning of the twenty-first century.

See also: Islam: Middle East; Islamic Fundamentalism; Middle East and North Africa: Sociocultural Aspects; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Sociology of; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Politics; Religion: Nationalism and Identity

Bibliography

Abu Jaber K S 1966 *The Arab Socialist Ba'th Party*. Syracuse University Press, Syracuse, NY

Ajami F 1981 *The Arab Predicament. Arab Political Thought and Practice since 1967*. Cambridge University Press, Cambridge, UK

Antonius G 1938 *The Arab Awakening*. Librairie du Liban, Beirut, Lebanon

Cleveland W L 1971 *The Making of an Arab Nationalist: Ottomanism and Arabism in the Life and Thought of Sati'al-Husri*. Princeton University Press, Princeton, NJ

Gelvin J L 1998 *Divided Loyalties, Nationalism and Mass Politics in Syria at the close of the Empire*. University of California Press, Berkeley, CA

Hourani A H 1962 *Arabic Thought in the Liberal Age 1789–1939*. Oxford University Press, Oxford, UK

Jankowski J, Gershoni I (eds.) 1997 *Rethinking Nationalism in the Arab Middle East*. Columbia University Press, New York

Kayali H 1997 *Arabs and Young Turks*. University of California Press, Berkeley, CA
Kerr M H 1971 *The Arab Cold War. Gamal 'Abd al-Nasir and His Rivals, 1958–1970*, 3rd edn. Oxford University Press, Oxford, UK
Khalidi R, Anderson L, Muslih M, Simon K S (eds.) 1991 *The Origins of Arab Nationalism*. Columbia University Press, New York
Khoury P S 1987 *Syria and the French Mandate: The Politics of Arab Nationalism, 1920–1945*. Princeton University Press, Princeton, NJ
Maddy-Weitzmann B 1991 *The Crystallization of the Arab State System: 1945–1954*, 1st edn. Cambridge University Press, Cambridge, UK
Nafi B M 1998 *Arabism, Islamism and the Palestine Question 1908–1948*. Ithaca Press, London
Porath Y 1986 *In Search of Arab Unity, 1930–1945*. Cass, London
Tauber E 1993 *The Emergence of the Arab Movements*. Cass, London
Tibi B 1981 *Arab Nationalism: A Critical Enquiry*. St Martin's Press, New York

G. Krämer

Nationalism, Historical Aspects of: East Asia

Nationalism in East Asia challenged a broad international system known as the Chinese World Order. In particular, nationalism in modern Japan played the most important role in dissolving this system.

1. *The Chinese World Order*

The scholar J. K. Fairbank and others (1968) have noted that the Chinese World order of premodern East Asia was a far-reaching international system characterized by the existence of an emperor considered to have been designated by heaven because of his innate virtue. He was referred to as a 'son of Heaven' and had the 'Mandate of Heaven' to rule. The emperor was considered to rule the world (*tianxia*, 'all under heaven') as a 'child of Heaven,' and his subjects ostensibly enjoyed the benefits of benevolent rule reminiscent of the 'Golden Age' of Chinese civilization described in ancient sacred texts. In reality, the power of the Chinese emperor had a limited reach, and there were a number of foreign tribes on the outskirts of the clearly defined spheres of Chinese civilization. Under the Chinese World Order, however, all subjects were ideally expected to be protected by the emperor, regardless of differences in their race, ethnicity, place of birth, or language—similar to the corpus christianum of the Middle Ages in Europe.

One basic principle of the Chinese World Order was that it attempted to enfold these foreign tribes in its system and stave off invasion through tributary missions and other rituals. Thus, many foreign rulers from surrounding areas shared and benefited from the common idea of a Chinese World Order. They used the system to their own advantage by sending tributary missions to Chinese dynasties which in turn gave them authority and protection in their own areas. Chinese emperors entered into feudalistic relationships with these rulers under this tributary system, providing them with royal titles, official seal and a calendar system in return for their loyalty. Foreign rulers, in turn, benefited by securing long-lasting political relations with Chinese dynasties through tributary missions. In addition, the tributary system played a significant role in functioning as a broad network of exchange of people, commodities, and information among different regions (Hamashita 1990).

There has been academic dispute over when and how the Chinese World Order was developed and maintained. Some argue that it can be dated as far back as the 'Spring and Autumn' and 'Warring States' periods (770–221 BC), while others claim that the prototype of the system can be identified in the age of the Sui-Tang dynasties (581–907), with refinement of the system being completed as late as the Ming-Qing period (1368–1911). These theories notwithstanding, certain characteristics are indisputable: the Chinese World Order constituted one of the primary ways for China to regard its world, and it had a great impact in connecting many disparate and broad regions through their common use of the Chinese written language.

2. *The China-centered View of China's Neighbors*

The philosophy of China as the center of the world (made evident by the Chinese character word for China, 'Middle Kingdom'), which constituted the Chinese World Order, was not always used solely by Chinese emperors. For example, in seventh century Japan, which was undergoing the formation of a state, the self-consciousness of the people as an independent state was enhanced by their admiration of China and their emulation of Chinese imperial political systems. Later, in the seventeenth century, when the Ming dynasty was destroyed by a non-Han ethnic minority (the Manchus of the Qing Dynasty), a powerful new discourse arose in tributary countries such as Korea and Vietnam which claimed that they, not the Manchus, were the legitimate successors to the Chinese dynasty, given their long history of tributary relations with China.

In Japan, where the tributary system never took root, a number of intellectuals expressed their outright hostility against Chinese civilization, which had en-

tered Japan through the widespread use of Chinese characters and Chinese literary texts. For instance, in the eighteenth and nineteenth centuries some Confucian scholars claimed that the term 'Middle Kingdom' was meant to refer only to Japan. Other nativist scholars studying Japanese classical literature during the Tokugawa period (1600–1868) claimed Japan was the 'land of the gods,' effectively placing Japan above China. Anti-Chinese sentiment against Qing (Manchu) China prevailed in the peripheral countries such as Korea, Japan, and Vietnam, and played an important role in forming a political and societal basis for national self-consciousness, because of the natural hostility of these nations towards the 'usurper' dynasty (Arano 1988).

Early modern Japan engaged in the process of forming an independent national consciousness earlier than other East Asian nations. Never having been a tributary of China, Japanese scholars were able to form their own ideology of international relations during this period of relative stability (Toby 1984). In addition, Japan successfully maintained an individual route for economic trade with Western countries during the peaceful Tokugawa era. In this respect, it can be said that the Japanese state in the early modern age had already begun to separate itself from a China-centered Asia well before the nineteenth century, when the formation of a modern state along the lines of Western nation states was begun.

3. *The Western Threat and the Acceptance of International Law*

The Chinese World Order underwent a tremendous change in the nineteenth century when Western nations began incursions into Asia in search of trading opportunities. During the Qing dynasty, a series of wars between Western nations and China in 1840–2 and 1858–60 compelled a small number of intellectuals to promote the urgent necessity of strengthening state defenses. The treaties concluded between China and the Western powers forced China's involvement in the Western nation state system. As symbols of this, the first bureau of foreign affairs, the *Zongli-Yamen* (Office for General Management), was established in the capital city in 1861, and the exchange of permanent missions also began in the 1870s (Bannno 1964, 1973).

However, high-ranking Qing government officials as well as officials responsible for making foreign policy actually regarded these treaties as an extension of their tributary policy; therefore, they did not regard these new relations with the West as relations among equal, independent nation states, let alone with China as a subordinate. Although Chinese territory was repeatedly invaded, few Chinese were able to reach across societal divisions to form a consciousness of national crisis and national identity. As long as the conventional, historical Chinese view of the world (with China at the center) remained alluring, it was difficult for the Chinese to rally around a nationalism which would necessitate that they actually view their nation as one of many nations in the world.

Meanwhile, nationalism was growing rapidly in Japan—perhaps partially because Japan had never been free of a consciousness of itself as on the periphery of dominant Chinese civilization, and so had long tried to carve out an independent identity. The catalyst for the rapid growth of nationalism and nation-strengthening movements was the Opium War (1840–2) in China, which called Japan's attention to the danger of the Western powers and their interest in East Asia. Japanese nationalism was aimed at strengthening and modernizing Japan along the lines of a Western-style nation state, and to encourage Japan to abide by the system of international law which theoretically supported the formation of a nation state. As soon as the first international treaty was concluded between the USA, the UK, and Russia in 1854, a political movement called *sonno joi* ('revere the emperor, expel the barbarian') was launched to overthrow the bakufu and establish the Emperor as the head of the state. This movement led to the Meiji Restoration of 1868 (see *Meiji Revolution, The*). With the aim of defending itself against the threat of the Western powers, the Meiji government energetically promoted policies aimed at modernizing Japan's industries, military power, and constitutional system. The end goal was to establish a fully-fledged modern nation state, with the emperor at its head, which would earn the respect of the Western nations.

Meiji statesmen and bureaucrats learned the principles of sovereignty and the territory through their studies of international laws (*bankoku koho* in Japanese, *wanguo gongfa* in Chinese) which were introduced from China during the late Tokugawa period. Their main goal was to abolish the 'unequal treaties' which Japan had been forced to sign with the West in the 1850s (see *Peacemaking in History*). Japan also sought influence in some East Asian territories in which China had been dominant, such as Taiwan and Korea. Japanese expansionism brought about increasing regional tensions with the Qing dynasty over influence in Taiwan, Ryukyu and Korea, and the tension finally culminated in military clashes between the two countries.

4. *The Sino–Japanese War of 1894–5 and the Advent of Modern Nationalism*

The Sino–Japanese War of 1894–5 was a major catalyst in promoting the growth of nationalism in Japan and transforming the East Asian world order. This war, which broke out over struggles between the two countries over supremacy in Korea, culminated in

Japan's unexpectedly thorough victory over China. The tributary system was finally dissolved with the granting of the unconditional independence of Korea, which had been Qing China's closest tributary state. Taiwan also became a colony of Japan. The result was a new East Asian world order with Japan at its center, replacing China.

The mobilization of the Japanese people during the Sino–Japanese (1894–5) and Russo–Japanese (1904–5) Wars marked a decisive turning point for Japan and the formation of national self-identity. In the wake of the Russo–Japanese War, modern Japan began to expand its image of empire with a view to invading the Chinese continent, since Japan was given rights and interests in Korea and Manchuria as part of the spoils of war. The war united patriotic militarists and also promoted an unprecedented sense of alliance among people in Japan from different regions or social classes.

China's defeat during the Sino–Japanese War was a tremendous shock to the Chinese. Moreover, the sense of crisis in China was intensified by the fact that a small country which had long been regarded as a backward area unexpectedly defeated the Chinese army with military power adopted from Western civilization. It was at this momentous time that a true sense of Chinese nationalism first appeared as the old Chinese World Order finally was dissolved. In particular, the news on the military defeat of China rendered the serious impact on lower level bureaucrats and progressive intellectuals, and they attempted to promote nation state building along the lines of Meiji Japan. Although this reform movement (*bianfa*) failed in 1898 under the pressure of conservatives, the Qing government attempted their own topdown reforms, and the transition to a modern constitutional state was launched on a large scale during the first decade of the twentieth century.

During this period, efforts to build up a nation state were also being made in the former tributary countries such as Korea and Vietnam, especially among the lower stratum of elite. In Vietnam and Korea, nationalism came into being through anti-imperialist movements against France in the case of Vietnam and against Japan in the case of Korea. Taiwan's nationalistic movement was born when the Taiwanese opposed Japanese colonial discrimination during the 1920s under the influence of Wilsonian ideas of national self-determination prevailing since World War I. Decades later, a separate form of nationalism in Taiwan gained strength that was based upon a consciousness of Taiwan as separate from the Chinese continental government.

5. *Nationalism and Asianism*

As mentioned earlier, modern Japan was the first East Asian country to create a unified nation state successfully, and it struggled to build up a national identity under the strong outside pressures of Western forces and with the new common national goal of 'leaving Asia, entering Europe' (*Datsu-a nyu-ou*). However, the nationalism of the expanding Japanese empire that had already incorporated Okinawa and Hokkaido in its territories took on a new form when Taiwan and Korea were colonized by Japan in 1895 and 1910, respectively. A new discourse of Asianism emerged that had the goal of opposing Western imperialism against all Asian nations by uniting the Asian nations under Japanese leadership. This ideology led to new complications owing to the variety of ethnicities and national loyalties that existed in East Asia (Takeuchi 1963).

One example of an Asianist thinker was the Meiji intellectual Tokutomi Soho (1863–1957), who promoted an idea of democracy among the common people (*heimin-shugi*) that also stressed the importance of a strong power in Asia to oppose Western powers. He was particularly provoked by the Triple Intervention of Russia, France and Germany in 1895, which was perceived as the ultimate example of Western arrogance against Japan.

Asianism would only succeed if it accommodated itself to the new regional order, which was constantly changing in response to factors such as the gradual worsening of Japanese–Chinese relations and global changes in the wake of World War I. The possibility of Sino–Japanese cooperation against the West seemed stronger with the publication of works such as Sun Yat-sen's *Greater Asianism* in 1924 (Sun 1986).

However, in spite of the lofty ideals of the building of a new East Asian order, the Manchurian Incident of 1931 and the start of the Sino–Japanese war in 1937 meant that the autonomy and unification of the Asian nations was not brought to fruition. Certainly, right after the outbreak of hostilities, the idea of 'East Asian Cooperativism' differed from other philosophies such as liberalism or communism as it attempted to promote a principle that would unify Asia during this period of total war. Such ideas were an important test of the possibilities of redefining the traditional boundaries of the state and its people. They were also a criticism of self-serving Japanese policies towards China and of the general attitude towards Chinese nationalism. These were a new manifestation of the possibilities of the principle of Asianism (Hasikawa and Matsumoto 1970).

However, aggression continued to maintain a stubborn relationship to unity within Asianist thought. The constantly expanding battlefront and the continued use of military aggression against Asian peoples eventually destroyed the hope of Asian unification. Euphemistic slogans such as 'Liberation of the [Asian] people' or 'Building a new order' inevitably became fodder for Japan to deceive itself and justify aggressive war against its neighbors. The wartime concept of the 'Greater East Asian Co-Prosperity Sphere' (*Dai to-a kyoei ken*) was another manifestation of pan-Asian-

ism. In spite of the lofty principle of Asian cooperation behind the idea, proponents of this concept were never able to take on a point of view critical of Japan's increasing aggression and imperialism. It was not until Japan's defeat in the Second World War in 1945 that these discourses justifying Japanese imperialism disappeared.

6. Ethnic Nationalism in the Periphery

The Qing dynasty dissolved in the wake of the Revolution of 1911. The new Republic of China aimed at establishing a highly centralized state that would embark on a number of industrial, societal, and educational reforms. Rural skirmishes with warlords and regional military cliques occupied the energies of the government, however, and the unification of the nation was not achieved until the Nanking Nationalist Government was established in 1928. Mongolia and Tibet both virtually declared independence from China on the grounds that the former tributary relationship was severed when the Chinese emperor abdicated in 1912. Undoubtedly, the Chinese central government did not approve of independence, but it was difficult for the government to rule these areas in an effective way since they needed to maintain diplomatic relations with countries such as the USSR and the UK, which enjoyed a number of concessions in those regions. The tendency for non-Han regions to distance themselves from the central government grew stronger (Mancall 1984). One such example could be found in the independence of the 'Outer Mongolia' in 1924.

During the 1930s and 1940s, there were a number of political movements for independence by Muslims of Turkish descent in the area of East Turkestan (Xinjiang province), as well as military disputes between Tibetans and Chinese in the co-habitated areas of Tibet. Although areas such as Inner Mongolia, East Turkestan, and Tibet were merged into China under direct governmental control when the Peoples' Republic of China was established in 1949, ethnic nationalism in these areas reportedly has become increasingly vigorous in recent years (Mori 1998).

See also: East Asian Studies: Politics; East Asian Studies: Society; Historiography and Historical Thought: East Asia; Nationalism and Expressive Forms; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Nations and Nation-states in History; Nation-states, Nationalism, and Gender; Xenophobia

Bibliography

Arano Y 1988 *Kinsei Nihon to Higashi Ajia* [*The Early Modern Japan and East Asia*]. Yoshikawa Kobunkan, Tokyo

Banno M 1964 *China and the West, 1858–1861: The Origins of the Tsungli Yamen*. Harvard University Press, Cambridge, MA

Banno M 1973 *Kindai Chugoku Seiji Gaikoshi* [*The History of Foreign Relations of Modern China*]. Tokyo Daigaku Shuppankai, Tokyo

Fairbank J K (ed.) 1968 *The Chinese World Order; Traditional China's Foreign Relations*. Harvard University Press, Cambridge, MA

Hamashita T 1990 *Kindai Chugoku no Kokusaiteki Keiki: Choko Boeki Shisutemu to Kindai Ajia* [*International Moments of Modern China: Tributary System and Modern Asia*]. Tokyo Daigaku Shuppankai, Tokyo

Hasikawa B, Matsumoto S 1970 *Kindai Nihon Seiji Shisoshi* [*The History of Political Thought in Modern Japan*]. Yuhikaku, Tokyo

Mancall M 1984 *China at the Center: 300 Years of Foreign Policy*. Free Press, New York

Mori K 1998 *Shuen kara no Chugoku: Minzoku Mondai to Kokka* [*Ethno-nationalism in Contemporary China*]. Tokyo Daigaku Shuppankai, Tokyo

Sun Yat-sen 1986 *Sun Zhongshan Quanji* [*Collected Works of Sun Yat-sen*], Zhonghua Shuju, Beijing, China, Vol. II

Takeuchi Y (ed.) 1963 *Ajia Shugi* [*Asianism*]. Chikuma Shobo, Tokyo

Toby R P 1984 *State and Diplomacy in Early Modern Japan: Asia in the Development of the Tokugawa Bakufu*. Princeton University Press, Princeton, NJ

Y. Murata

Nationalism, Historical Aspects of: South Asia

Nationalism, which first developed in the West as a conscious political ideology, spread to other parts of the world in the course of the nineteenth and twentieth centuries. It sustained political and social-cultural movements of great power in South Asia, i.e., the Indo-Pak subcontinent. The ethnic diversity of the region, the responses to colonial rule which subsumed and overlapped with reactions to the encounter with Western civilization, as well as the shifting definitions of national identity determined by the region's history both under colonial rule and after decolonization, gave nationalism in the area a very distinctive and at times highly idiosyncratic character. The purpose of this article is to identify these diversities, explain their nature and origin, and trace the outlines of their history over time. It will also discuss the wide variety of approaches to the phenomenon which define and explain it in fundamentally different ways.

1. Origin and Development of Nationalism in South Asia

1.1 The Pan-Indian Scene in the Nineteenth Century

One major consequence of the British rule in India was the unification of the subcontinent under one cen-

tralized political and administrative authority, even though two-fifths of the territory remained under the rule of Indian princes who accepted British paramountcy but enjoyed autonomy in all internal matters. A second development, crucial to our understanding of nationalism in the region, was the growth of a class of people who assisted in the functioning of colonial rule as middle- or low- level functionaries or found new means of livelihood like law, modern medicine, journalism, or teaching in the new Western-style educational institutions. This 'colonial middle class' soon had a shared language in English and were participants in a common intellectual discourse, facilitated by the new print culture and the modern means of transport and communication. Their colonial experience and exposure through modern education to Western history and social-political ideas laid the intellectual basis for nationalism and also helped create a shared body of expectations and grievances, which were sharpened by the experience of racist discrimination (see Sarkar 1989).

A new upsurge in literary and intellectual creativity and movements of social reform, described by some as India's nineteenth century Renaissance, laid the foundations of what may be described as protonationalist ideology in the first half of the nineteenth century (Majumdar1962–3). The researches of Western, especially British, Orientalists fed into these new concerns helping forge the belief in a great and unique past and thence the yearning for a comparable future glory. One essential component of this glorious future was to be a seat of honor in the comity of nations. The latter implied sovereignty, though clear articulation of such bold expectations was late in coming: In the last decade of the nineteenth century only a small group of militant nationalists began to think in such terms (Majumdar 1962–3). The educated Indians, a small portion of the total population, who began to share the newly emerging political consciousness accepted for the most part the fact of British rule as an enduring political reality which would persist for an indefinite period. The responses to this awareness ranged from enthusiastic faith in the redeeming and providential character of British rule, through hopeless resignation, to a desperate longing to try to end it: nationalist ideology could be and initially was, for the most part, strongly proempire.

1.2 The Dominant Strand in Indian Nationalism

All-India nationalist consciousness and ideology found an institutional and programmatic base in the Indian National Congress (INC) established in 1885. Initially blessed by the colonial government as a 'safety valve' for letting out any accumulated grievance against the ruler, the INC soon came to be suspect as a potentially seditious organization, despite its repeated avowal of loyalty to the empire at its annual meetings. But soon a militant wing did develop within the Congress and its most prominent leader, B. G. Tilak, declared in his journal, *Kesari*, that freedom was the birthright of humankind (Wolpert 1989). When the first political murders took place, the government believed that Tilak was the source of inspiration for these acts. Militancy found expression in a mass agitation when the government decided to partition the province of Bengal in 1905 hoping, *inter alia*, to weaken the political base of the nationalistic Bengali Hindu middle class whom they identified as highly 'anti-British' (Sarkar 1973). The consequent resentment also fed into the movement for violent action against the rulers which led to local and India-wide conspiracies during World War I and after. The conflict between the 'extremists' and 'moderates' induced a split in the Congress (1907) which weakened the organization for years to come (Argov 1967).

The advent of Gandhi, a world figure since his nonviolent campaign in South Africa against the regime's racist policies, led to a transformation of nationalist agenda after the war. Gandhi, despite his pacifism, had cooperated with the war effort until 1918. But violent repressive actions in the Punjab in 1919 and the way the British-sponsored enquiry into these events tried to play these down were major factors in his decision to launch a nonviolent campaign of noncooperation: The aim was to win *swaraj*, i.e., self-rule (Brown 1972). A major section of Muslim opinion in India had been alienated by the way the Allies treated their religious head or *khalifa*, the Sultan of Turkey. The mass movement launched by Gandhi brought in these disgruntled Muslims who sought justice for the Khalifa (Hasan 1991). One strikingly new feature of this agitation, known as the Noncooperation-Khilafat Movement, was the technique of *satyagraha* or nonviolent struggle which had strong ethical overtones. For Gandhi, the struggle was not merely a campaign to achieve independence but an 'experiment with truth,' the moral purpose being for him at least as important as the political goal, if not more so (Gandhi 1948). Following a very violent action by the supporters of the movement, he suspended it in 1922. This brought to an end the Khilafat Movement as well, and with it the unity in a common struggle which brought the nationalists and militant Muslims together.

Gandhi led one more mass agitation, the Civil Disobedience Movement (1930–4) and in 1937, when the British granted a new constitution to India which conceded a fair measure of power to elected governments in the provinces, the INC gained a majority in most of the provinces, but had very limited support in the Muslim majority provinces. During World War II, India was declared a belligerent without consultation with the Indian political parties and the Congress ministries in the provinces resigned. A British mission led by Sir Stafford Cripps failed to secure the INC's cooperation in the war effort. The Congress lead-

ership, including Gandhi, who were planning to launch another mass agitation, were imprisoned in 1942 and a largely spontaneous mass movement, known by a slogan devised by Gandhi, 'Quit India,' erupted. Meanwhile, in 1940 the Muslim League, which eventually won the support of the majority of Muslims in India, launched a campaign for an autonomous Muslim homeland, Pakistan. The Congress leaders were released in 1945 at the end of the war and negotiations with all political parties led to transfer of power and the partition of India into two successor states, India and Pakistan in 1947 (Moore 1974, Mansergh 1970).

While the INC under Gandhian leadership provided a common platform for nationalists of many hues, Indian nationalism was never a seamless web. Firstly, within the Congress, from a very early stage there were the moderates and extremists, i.e., those who preferred the path of constitutional agitation leading to autonomy within the empire and the others who preferred a more militant program aiming at independence. Later, the 'Right' and 'Left' within the Congress represented these two different approaches, except that Gandhi, usually identified with the Right, preferred caution even though he led two nationwide agitations. 'Right' and 'Left' came to subsume a more basic ideological division: the former was in favor of moderate socioeconomic reforms, while the latter wanted a radical program of change undermining the existing status quo. The Congress developed a socialist wing in the 1930s and even the newly born Communist Party of India was allowed to function from within the Congress until 1942 when they opposed the Quit India movement (Majumdar 1962–3). The tension between the Congress Right and Left reached a breaking point when the youthful leader, Subhashchandra Bose, felt compelled to leave the party and form a separate organization, the Forward Bloc. Later he left India during the war and, with Japanese help, formed the Indian National Army (INA). The INA, which participated in the war against the British in the eastern theater of war made a tremendous impact on Indian nationalist consciousness. The British had to abandon the trial of INA prisoners of war in the face of such strong sentiments (Gordon 1990)

Secondly, while the Muslim support for the Congress dwindled especially after 1928 when an All Party Conference failed to secure agreement between the nationalists and Muslim parties seeking future safeguards for Muslims in independent India, sections of Muslims remained staunchly loyal to the nationalist cause within or outside the Congress (Hasan 1991).

Thirdly, India's multiethnicity meant that modern political consciousness had more than one focus. 'Nationalism' often meant a strong sense of local cultural identity and a consequent pride in an imagined ethnic past. In many cases, pan-Indian nationalism could accommodate such 'local nationalisms' as an integral part of the larger identity. There was no conflict between a Bengali, Assamese, or Gujrati sense of past glory and present loyalty to the language-based culture and similar sentiments focused on Indianness and an Indian political identity (McCully 1940, Brass 1974). The Congress successfuly accommodated regional aspirations and managed interregional conflicts of aspiration both before and after the transfer of power. Only in specific historical circumstances, especially after the transfer of power, did such potentialities of tension and alienation reach unmanageable levels.

1.3 Muslim Nationalism and Foundation of Pakistan

The INC and Indian nationalism failed in their effort to absorb one religion-based ethnic identity. In precolonial times, the Hindu–Muslim divide, so called, was a social rather than a political fact. Under the Muslim kings, Hindus always constituted a large if not the larger proportion of the ruling class. In the great rebellion of 1857, Hindu and Muslim soldiers, peasants, feudatories, and princes fought together with the ostensible purpose of restoring Mughal rule. In the colonial era, the rulers' policy of securing alliances through the distribution of such resources as they controlled, especially education, jobs, and seats on representative bodies, were often based on an identification of the communities as basic political constituencies. This fact helped politicize community identities: In the case of Muslims this identity was seen to have a pan-Indian dimension. The cultural revival in nineteenth century India focused among other things on a romanticized vision of Islam's past glory in and outside India (Pakistan Historical Society 1957–[1970]).

As Indian nationalism's image of the past focused initially on Hindu achievements and often portrayed the rule of the medieval Muslim dynasts as an age of slavery for the Hindus, the political agenda of the Pakistan movement postulated the theory of two nations, ignoring the multiple components of both the Hindu and the Muslim communities in India, as also the absence of any pan-Indian consciousness in either community in the precolonial period. This two-nation theory was endorsed by some Hindu historians as well who treated the history of the Indo-Islamic era as a period of foreign rule (Hardy 1972, Page 1987, Majumdar 1962–3).

1.4 Nationalism in Postcolonial South Asia

In divided India, the power of nationalism in the Indian Union was manifest in the success in bringing all the princely states within the union's boundaries into the union with the active support of the local populations and the very survival of the union for more than 50 years to date. The local aspirations,

often manifested as centrifugal tendencies, have been successfully negotiated for the most part. In the Punjab, the aspirations of the Sikh community developed into a movement for independence, but a policy of severe repression and conciliation of more moderate elements has proved successful (Brass 1994). In the north-east, several local groups continue to fight for secession from India as do the Muslim separatists in Kashmir. But the forces of commitment to a united India appear to be stronger than centrifugal tendencies. A number of ethnic movements in different parts of India are concerned more with the formation of autonomous states within India than with dreams of independence (Baruah 1999).

In Pakistan, the ascendancy of the Punjab and the ruling elite's refusal to accommodate the aspirations of the Bengalis in East Pakistan produced a nationalist movement first based on their linguistic-cultural identity. In 1971, the East Pakistan-based Awami League won an absolute majority in Parliament, but their demand for virtual autonomy was answered by massive repression. This led to a war of liberation, and Indian intervention contributed to defeat of the Pakistani army. The independent state of Bangladesh was thus born. Some saw in this event the inevitable triumph of ethnicity-based centrifugal tendencies, but perhaps the hamhanded policies of Pakistan's military government were more to blame. The movements for autonomy or secession in Sindh, Baluchistan, and the North West have also perhaps more to do with cynical misrule than spontaneous ethnic aspirations (Khan 1985).

2. *Differing Interpretations of Nationalism in South Asia*

There is no consensus on the nature and dynamics of nationalism in South Asia. The imperialist perception repeatedly emphasized the multiethnicity of South Asian society, implying that a nation state was not viable in the context of this 'society of societies.' In this view, nationalism was the false aspiration of a cynical elite in India who tried to mislead the masses, normally loyal to the imperial connection (Strachey 1888, Chirol 1910). Opposed to this was the nationalist self-perception as the upholders of the natural rights of the Indian people: They wrested India's independence from the rulers through sustained struggle; a reluctant colonial power conceded it in slow stages in response to successive waves of mass agitation, unique in their allegiance to an ideology of nonviolence (Sitaramayya 1946–7). An organization originally confined to the elite gradually brought in the masses through successful propaganda by word and action. The imperialists, by contrast, projected the notion that they gradually trained up the Indian people in the art of self-government. The delay in handing over power was entirely due to the internal squabbles of Indian politics, especially the irreconcilable aspirations of Hindus and Muslims which eventually led to the partition of India (Coupland 1944).

More recent interpretations have strong resonances of these older views. A group of historians in Cambridge explained the dynamics of Indian politics in terms of collaboration with the Raj and competition between various indigenous groups for a share in the resources made available to the collaborators. The 'have nots' of such shares took to agitation and nationalism was the rhetoric of their frustrated aspirations intended to win support (Gallagher et al. 1973). The Marxists explain Indian nationalism as the agenda of the colonial bourgeoisie whose expectations remained unfulfilled under colonial rule. While they successfully secured the support of peasants and workers, their agenda both before and after independence denied the masses their legitimate rights (Dutt 1940). One version of this interpretation suggests that the class basis of the nationalist movement did expand over time, bringing in the vast majority of the Hindu population, but it excluded significant sections of the masses, especially those at the bottom rungs of Indian society (Pandey 1978).

The emphasis on ethnicity as a factor which will inevitably undermine national cohesion features prominently in the writings of some observers like Neville Maxwell, who saw in the emergence of Bangladesh the beginning of the process of disruption leading to the inevitable breakup of India and Pakistan along ethnic lines. Radical writers like Partha Chatterjee provide indirect support for this view in stating that the basic loyalty of the masses is to the communities who should be empowered. The state, in this view, is an institution superimposed by the dominant classes to exploit the underprivileged (Chatterjee 1993). This perception is a logical corollary of the thesis developed by the historians and social scientists associated with the publications entitled the *Subaltern Studies*. This thesis suggests that all studies of nationalism hitherto have focused on the elite groups denying the underprivileged their agency, their role as subjects in their own right. The notion that the elite leadership mobilized the ignorant populace is shown to be false: The masses interpreted the messsage of nationalism in their own light and participated in the action only when they felt it to be in their interest. And their primary loyalty was always to their community.This thesis has been criticized, among other things, on the ground that the idea of a pan-Indian nation did not come spontaneously to the masses: they learnt it from the nationalist leadership and their propaganda by word and action.

3. *Open Questions and Prognosis for the Future*

Some features of nationalism in South Asia are indeed unique. The Indian Union, the world's largest democracy representing a population of which 40 percent are

illiterate, is also the largest multiethnic state in the history of the world. The comparison with the USSR is invalid, because that state was successor to the Tsarist empire and hence did not represent any voluntary union. The Indian state, on the other hand, was the union of willing partners who voted for such a union through a Constituent Assembly. And the remarkable fact is that despite tensions the union has held together and representative democracy, a system unknown to the region's political traditions, has survived for half a century. In Pakistan, the creation of a state based on the identity of faith was an innovation. That identity has not, however, been able to hold the state together and the democratic principle has also been violated repeatedly by the army. In Bangladesh, there is a persistent tension between the language-based Bengali identity and the emphasis on Islam preferred by significant groups in the country (Ziring 1992). In both countries, however, the electorate have repeatedly proved their allegiance to the imported ideal of representative democracy.

Crystal-gazing into the future of the South Asian states is not a very rewarding exercise. If there are threats to their survival and risks of breakup into smaller units, one needs to ask whether these arise from irreconcilable ethnicities or from the contingencies of bad government and the exploitation of particular groups by others. How far the problems of poverty and ignorance have inhibited the growth of national consciousness among the masses also needs to be analyzed. Another question which will have to be addressed is the implication of globalization for nationalist ideologies and the nation state itself.

See also: Area and International Studies: Development in South Asia; Nationalism: General; Nationalism, Sociology of; Nation-states, Nationalism, and Gender; Religion: Nationalism and Identity; Religious Nationalism and the Secular State: Cultural Concerns; South Asia: Sociocultural Aspects; South Asian Studies: Culture; South Asian Studies: Politics

Bibliography

Argov D 1967 *Moderates and Extremists in the Indian Nationalist Movement*. Asia Publishing House, Bombay, India

Baruah S *India against Itself: Assam and Politics of Nationality*. University of Pennsylvannia Press, Philadelphia, PA

Brass P R 1974 *Language, Religion and Politics in North India*. Cambridge University Press, London

Brass P R 1994 *The Politics of India since Independence*. Cambridge University Press, Cambridge, UK

Brown J 1972 *Gandhi's Rise to Power—Indian Politics 1915–1922*. Cambridge University Press, Cambridge, UK

Chatterjee P 1993a *Nationalist Thought and the Colonial World*. University of Minnesota Press, Minneapolis, MN

Chatterjee P 1993b *Nation and its Fragments: Colonial and Post-Colonial Histories*. Princeton University Press, Princeton, NJ

Chirol V 1910 *Indian Unrest*. Macmillan, London

Coupland R 1944 *Constitutional Problem in India*. Oxford University Press, New York

Dutt R P 1940 *India To-Day*, 2nd edn. Manisha, Calcutta, India

Gallagher J, Johnson G, Seal A 1973 *Locality, Province and Nation. Essays on Indian Politics, 1870–1947*. Cambridge University Press, Cambridge, UK

Gandhi M K 1948 *Gandhi's Autobiography: Story of My Experiments with Truth*. Public Affairs Press, Washington, DC

Gordon L 1990 *Brothers Against the Raj: A Biography of Subhas and Sarat Bose*. Viking, New Delhi, India

Hardy P 1972 *The Muslims of British India*. Cambridge University Press, Cambridge, UK

Hasan M 1991 *Nationalism and Communal Politics in India, 1918–1928*. Manohar, New Delhi, India

Hashmi T I 1992 *Pakistan as Peasant Utopia. The Communalization of Class Politics in East Bengal 1920–1947*. Westview, Boulder, CO

Khan A (ed.) 1985 *Islam, Politics and the State. The Pakistan Experience*. Zed Books, London

Mansergh N (ed.) 1970 *Transfer of Power*, 12 vols. HMSO, London

Majumdar R C 1996 *History of Freedom Movement*, 3 Vols. Classical Publishing Company, Calcutta, India

Majumdar R C (ed.) 1962–3 *British Paramountcy and Indian Renaissance*. Bharatiya Vidya Bhavan, Bombay, India

McCully B 1940 *English Education and the Origins of Indian Nationalism*. New York

Moore R J 1974 *Crisis of Indian Unity*. Clarendon Press, Oxford, UK

Page D 1987 *Prelude to Partition: The Indian Muslims and the Imperial System of Control, 1920–1932*. Oxford University Press, Delhi, India

Pakistan Historical Society 1957–[1970] *History of Freedom Movement, Being the Story of Muslim Struggle for the Freedom of Hind-Pakistan, 1707–1947*. Pakistan Historical Society, Karachi, Pakistan

Pandey G 1978 *The Ascendancy of the Congress in Uttar Pradesh 1926–34. A Study in Imperfect Mobilization*. Oxford University Press, Delhi, India

Sarkar S 1973 *Swadeshi Movement in Bengal 1903–08*. People's Publishing House, New Delhi, India

Sarkar S 1989 *Modern India, 1885–1947*, 2nd edn. Macmillan, Basingstoke, UK

Sitaramayya P 1946–7 *History of the Indian National Congress*. Padma Publications, Bombay, India

Stratchey J 1888 *India*. London

Wolpert S 1989 *Tilak and Gokhale: A Revolution and Reform in the Making of Modern India*. Oxford University Press, Delhi, India

Ziring L 1992 *Bangladesh from Mujib to Ershad. An Interpretative Study*. Oxford University Press, Karachi

T. Raychaudhuri

Nationalism, Historical Aspects of: The West

The article starts from the premise that nationalism has be to studied in relation to nation: an historical

aspect of nationalism is the formation of a modern nation and the creation of nation-states in Europe. The study of the phenomenon, however, is complicated by disparate use of the concept nationalism by various researchers. 'Nationalism' has thus become a hazy and not very suitable tool for analysis, which is also reflected in numerous misunderstandings in evaluations of both the positive and the negative roles of nation and nationalism in history. This article documents that, despite the considerable differences in terminology and evaluation, there are roughly five fundamental causal relations which the various research schools consider crucial for nation—and nationalism—forming: (a) state-political historical foundations; (b) ethnic roots; (c) processes of modernization, including industrialization, institutional education, and social communication; (d) nationally relevant conflicts of interest and unevenness of development; and (e) psychological and emotional manipulation.

The complicated process of the modern-nation-forming and the onset of the manifestations of nationalism has been, quite understandably, a subject of scholarly research ever since the nineteenth century. This research has had to struggle with a number of theoretical and semantic problems, and has not achieved consensus over the fundamental issues of terminology. The question of what role was played by 'nationalism' has been answered differently by different authors, primarily depending on how they defined nationalism, nation, nationality, national identity, ethnic group—and relations between these categories. Therefore, it is most important to investigate through what terms the authors analyzed the processes and how they defined the terms.

1. The Problem of Definition

The fundamental problem of terminology derives from the fact that it concerns a process resulting in the formation of new communities, and that these communities used a term for themselves which had already originated in the Middle Ages, but was given a different content in different political and cultural conditions. On the one hand, 'nation' in English included all inhabitants living under one government and ruled by the same laws. The first edition of *Grande encyclopédie* also defined '*la nation*' in similar terms, but later the use of a common language was added to this characterization. On the other hand, 'nation' in German was traditionally primarily associated with language, culture, and also sometimes with a shared past. That is why anglophone researchers find it hard to accept speaking about a fully fledged nation in cases where there is no nation-state (Kemiläinan 1964).

Another difficulty comes from the fact that 'nationalism' entered the academic vocabulary quite late: it was a neologism originally used in politics and, therefore, laden with political connotations. Nevertheless, the term is also derived from 'nation,' which means that it becomes related to the linguistic connotations mentioned. It follows that anglophone literature will associated nationalism, one way or other, with a state and/or a struggle toward statehood.

1.1 Views on Nation, Nationality, and Nationalism in the Nineteenth Century

The study of nation and national consciousness developed alongside the nation-forming processes at least form the mid-nineteenth century and was influenced by the following circumstances: (a) the previous linguistic tradition; (b) the axiom of primordiality, that is by the idea that the nation *an sich* had existed since the Middle Ages and that people gradually realized their objective belonging to it; (c) the degree of nation-forming of a given nation: English and French researchers presumed the existence of a nation as a state-nation, while for German, Italian, and 'East-European' writers a nation was not commonplace and they were still looking for a suitable definition. It is not a coincidence that the largest numbers of the studies of nation can be found in the German-speaking territory. It needs to be added that English and later American authors also looked for a definition of a nonstate community. For the most part, they used the words 'nationality' or 'peoples.'

The large number of studies written up until World War II can be divided into two main streams:

(a) Nation was defined by objective, empirically researchable, characteristics. These usually included, first of all, cultural and linguistic characteristics in combination with various other features—from political ties to territorial, economic, and 'blood' relations. Nevertheless, already at the beginning of the twentieth century some authors concluded that it was impossible to find a universally applicable combination of characteristics, including all entities considered nations at the time. That is why a differentiation between a state and a cultural nation gained support (Meinecke 1907). The criterion was whether the community which called itself a nation was a state or not. Another approach was to define nation as a 'community with a shared past,' which originally applied only to the members of the ruling classes and élites, but which integrated the entire population in the course of historical development since the Middle Ages (Bauer 1907).

(b) Nation was defined subjectively—its members were aware of their belonging to the nation and wanted to belong to it. This concept could be expressed in terms of agitation—as a will to form a nation—or existentially, as a sort of a 'plebiscite'—as a decision to belong to an existing national community (Renan

1888). Typically, the subjectivist concept of nation originally appeared in essays rather than in empirically researched studies.

The dividing line between both concepts is not clear cut. On the one hand, supporters of the objectivist definition realized more and more strongly that a nation could not exist without national consciousness. On the other hand, the subjectivist definition did not preclude a primordialist concept of a nation.

It is symptomatic that national consciousness, 'nationalism,' became a tool of scholarly analysis as late as in the period between the two World Wars. The pioneer of the study of nationalism, Carlton J. Hayes (1926), presumed that it was necessary to define the object of nationalism, and attempted to define both nation and nationality through objective features.

1.2 The Definition of Nation and Nationalism in the Second Half of the Twentieth Century

The sad experience with nationalism during World War II for Europe effected the approaches of historians and social scientists to the study of nation. The political acuteness of this study increased and also with it the political commitment of the authors. Although opinions continued to vary, most researchers agreed on a number of points in setting priorities in research and in discarding insupportable views:

(a) An overwhelming majority of researchers more or less distanced themselves from the view that nation is a timeless category and a blood-related community;

(b) Agreement was also reached concerning the impossibility of defining nation only through cultural and linguistic characteristics;

(c) Nation was usually accepted as a distinct community only if evidence was available that the members of the community (or at least the prevailing majority of them) considered themselves a nation;

(d) Hence the growing emphasis on the study of 'nationalism,' as a subjective manifestation or as a precondition for an objective existence of a nation.

Nevertheless, conservative historians perpetuated for a while the traditional idea that a modern nation was a continuation of a medieval nation. After all, studies of national (or proto-national) identities in the Middle Ages and the Early Modern Age are a fairly common occurrence in Europe.

In the conditions of political bias against German nazism and Italian fascism, the atmosphere in Europe was not favorable for applications of the term nationalism as a neutral tool for historical analysis as it was developed by Hayes (1931). On the contrary, the numbers and importance of authors who considered nationalism as a negative extravagance which needed to be overcome increased. According to authors such as Friedrich Hertz (1945), Edward H. Carr (1945), and later also Elie Kedourie (1960) and Kenneth Minogue (1967), nationalism was invented by irresponsible educated individuals, spread by equally irresponsible demagogues and politicians, and should be eradicated as a dangerous legacy of the past.

A polarized and modified variant of this concept as promoted by Hans Kohn in his numerous publications became more influential. In his view, two kinds of nationalism were to be distinguished: 'western,' which was rationalist, constitutional, liberal, democratic, and developed in France and the UK; and 'eastern,' which was irrational, mythical, racist, and reactionary. Although this 'dichotomy' (Snyder 1954) was persuasively criticized many times (Kemiläinen 1964), it had a number of adherents.

Most historians and social scientists sought acceptable ways of defining nation as an objectively existing social group during the postwar decades. The approach that it was possible to develop an authoritative set of 'features' every nation had to have survived only as the official (although not Stalinist) doctrine of Soviet research. On the other hand, a number of authors were trying to find a more free definition by combining several kinds of relations. Among these relations, Boyd Shafer (1955) distinguished 'illusions,' such as economic needs, language, shared history, and objective relations, such as shared government, shared cultural, and social institutions. The sociologist Pitrim Sorokin (1947) saw nation as a 'multibound group' formed by a set of individuals 'linked by two or more unibound ties.' These could be a territory, a language, a religion, and/or physical conditions. A similar view among sociologists is held by Smith (1991) who characterized nation as a population set sharing the same myths, same historical territory, mass culture, and economy. Another variant of this perspective was a concept of nation as a social group bound by a changing combination of several kinds of historically strengthened, mutually substitutable relations, while civic equality of all members of the group and the consciousness of their shared history were necessary and non-substitutable ties (Hroch 1985).

The problems with finding a common, generally acceptable definition of nation also lead to giving up such a definition and favoring the subjective definition of nation: 'a nation exists when a significant number of people in a community consider themselves to form a nation' (Seton-Watson 1977). This change was often associated with the 'discovery' of Renan's concept of nation. The result was an increased frequency of use of the term nationalism. The object of this nationalism was often not defined, or it was characterized as 'super-personal entity' towards which a nationalist individual feels 'unconditional loyalty' (Eugen Lemberg 1964).

This view gradually radicalized: nationalism began to be considered a given reality, while nation was regarded a construct, a myth, or an invention. The theoretical authority became Benedict Anderson

(1983)—through no initiative of his own. He called nations 'imagined communities,' by which he meant that a nation as a group could only exist if its members were able to imagine that they belonged to a community of people, of whom the majority they did not know, and would never know. This view, however, can also be interpreted that nation is just an illusion and an invention, therefore, the 'necessity' of its formation can be doubted. The constructivist concepts of nation do not solve the problem of definition, they merely move it to another perspective: they also have to explain the term nationalism in its historical context. Equally as in the case of nation, there is not a consensus among researchers on the term nationalism either. Quite the contrary, the terminological confusion is augmented by the haphazard use of the term nationalism by researchers and without attempts at definition.

However, we can also find important exceptions. John Breuilly (1982) defines nationalism as a politics based on the premise that a nation exists, and which identifies itself with this nation. It demands political autonomy for the nation, usually in the form of a state. Gellner's authoritative definition of nationalism as an effort to obtain the attributes of a state for own ethnic entity appears around the same time. Later, Gellner substantially modified his definition: 'Nationalism is a political principle which maintains that similarity of culture is the basic social bond' (Gellner 1997). His preference for the term nationalism does not mean a denial of an objective existence of a nation: his attitude is a mediating one between a subjectivist and an objectivist notion of nation.

Attention will now be paid to those authors who considered nation as objectively existing, 'essential.' Karl W. Deutsch (1953) made the first significant attempt at a new approach. He defined nation as 'a community of complementary habits of communication.' This communication community then forms a basis for a shared national culture and also nationalism. This approach was later more precisely defined as a group of people capable of better mutual than 'outward' communication and co-operation, and which is politically organized and autonomous. Members of the nation then associate community consciousness with certain values and with culture. Thus a foundation was created for the development of nationalism. Another step in this direction is the definition of nation as a 'social formation' which is 'integrated through the means and relations of communication, production, organization, and exchange' (James 1996). It needs to be mentioned that the majority of these authors did not know that they were drawing on a concept of Bauer (1907).

The view that nation is primarily defined by statehood is widespread mainly among anglophone authors. J. Breuilly, Stein Rokkan, and Charles Tilly should be mentioned, and also Louis Snyder, and Anthony Giddens, who maintains that a nation only exists 'when a state has a unified administrative reach over the territory over which its sovereignty is claimed' (Giddens 1987). Similarly, Eric Hobsbawm concludes that nation is 'a social entity only insofar as it relates to a certain kind of modern territorial state' (Hobsbawm 1989). In these schools of thought, nationalism was closely connected with the state.

Defining nation by statehood runs against an important empirical fact that communities which have had a long way to go to statehood, and sometimes have not even sought it, call themselves 'nations.' Especially in English there is a problem with finding an adequate term for this case.

1.3 Debate over the Evaluation of Nationalism

As mentioned in Sect. 1.2, the term nationalism had been a political term until the inter-war period. It was used in both negative and positive senses: referring either to a negative and an anti-social attitude (especially in the political language of the socialist movement), or to a commendable and noble love for one's nation (in the political language of some right-wing movements).

Some authors writing within the Marxist tradition associated nationalism with negative social phenomena. E. Hobsbawm and, less distinctly, B. Anderson are examples of this perspective. It is also true, however, that this same tradition made the important attempt to characterize the 'Janus-face' of nationalism: as a phenomenon which could have a negative and oppressive impact on the one hand, and a positive impact as a liberating force in the struggle against national or colonial oppression on the other (Nairn 1981).

The distinction between the two kinds of nationalism is typical for a large number of researchers, and it is foreshadowed by Hans Kohn's dichotomy mentioned above. The distinction between nationalisms according to the country or the national context was also used later.

Another dichotomy is a distinction between nationalisms according to time periods. In 1956, the influential German historian Theodor Schieder (1991) evaluated nineteenth-century nationalism as a progressive and creative factor, while he criticized the manifestations of the destructive side of nationalism, especially in the first half of the twentieth century. This distinction was later repeated by other authors. Ironically, the official Soviet scholarship also recognized two similar phases, although it did not explicitly use the term nationalism.

A majority of contemporary authors recognize the double-face of nationalism in history. For example, Louis Snyder (1976) considered nationalism as moral, as well as immoral, human as well as inhuman—according to these criteria, it could be evaluated as positive in some cases and as negative in others. He

saw the main danger of nationalism in its association with collectivism against individualism. The French philosopher Etienne Balibar (1991) sees the 'good' side of nationalism in that it contributes to the development of state and a community of citizens in general, it derives from the love of people, and it is tolerant. In contrast, the 'bad' side of nationalism is in its expansiveness, hatred, and intolerance. Another French scholar, the political scientist Jean Leca, considered the distinction between the varying impacts of nationalism to be a matter of definition. It is symptomatic that the positive potential of nationalism is mostly appreciated by non-Anglo-American authors: apart from the French mentioned, also the Catalunian historian Josep Llobera (1994) and the Israeli historian Hedva Ben-Israel can be mentioned.

Another group of authors considers the broad meaning of the term nationalism to be a source of imprecisions and misunderstandings, and recommends the use of the term only to refer to national egoism, preference of own nation over others, and placing one's own national interest over the interests of an individual (Peter Sugar, Leonard Doob, Otto Dann). Some recommended the use of other terms, such as patriotism (Doob, Samuel 1989), love for one's country, national consciousness, and national feelings for the positive manifestations of national identity. Another variant is the recommendation to preserve nationalism as the neutral term and find other terms, such as ethnocentrism L. Le Vine, D. Cambell, ethnonationalism Martijn Roessingh, xenophobia or chauvinism, for the negative manifestations.

A real terminological potential can come from those researchers who take the historical, that is developmental, perspective. They follow from the fact that attitudes called nationalism developed from certain previous, developmentally 'lower' attitudes, perceptions, and relations, which are sometimes called identities, at other times loyalty, or primordial sentiments. The term loyalty is rarely used in historic research, but it belongs to the terminology of cultural anthropology. It is characterized as 'a relation of an individual towards pre-existing given groups, phenomena values—in the social reality in which the individual lives' (Geertz 1973).

The term identity now belongs to the common terminology of a number of historical analyses. It is used most consistently by Anthony Smith, who maintains that nationalism needs to be studied as a certain consequence, or as another phase in the development of national identity. At the same time the consciousness of each individual is determined by multiple identities, which are more or less interrelated. Unlike the term nationalism, the term national identity is not laden or obscured with superficial journalism in the vocabulary of historic analysis. A number of contemporary scholars and projects promote the term. For example, the Swedish historian Sven Tägil (1984) suggests a distinction between 'identity,' referring to the simple consciousness of belonging to a group, and 'identification,' referring to the active and committed participation in the work for the nation.

The debate over definitions and evaluation should not be seen as mere academic disputes or noncommittal intellectual exercises. The debate is not self-serving, it is a discussion over whether and how, a particular term or concept is suitable for explaining the fact that after 200 years of dramatic and controversial development in Europe, in most cases, the nation-forming process resulted in mono-ethnic nation-states.

2. *Why Nation and Nationalism? Attempts at a Historical Explanation*

Every survey of views on the genesis of nations and nationalism is confronted with the contradiction between those who consider nation an invention or a construct—and therefore the product of nationalism—and those who see nation as a result of objective nation-forming processes. Judging from the verbal expressions, the conflict is irreconcilable. However, is that really so, or is the conflict not so fundamental—as Paul James (1996) argues?

The work of Ernest Gellner illustrates the productivity of the approach which tries to see reality behind words and aphoristic statements. His apodictic statement about nationalism creating nations is often cited and made absolute by 'modernists.' Few of them realized that Gellner's definition only moved the causal explanation to a different level: the point was to explain why nationalism had gained such a response. At this level, Gellner works mostly with objective social and cultural factors. Similarly, it clearly follows from the statements of Benedict Anderson that nations as 'imagined communities' could only be formed thanks to particular objective social preconditions.

Therefore, a causal interpretation, does not necessarily have to dramatically contrast the nation-forming process with nationalism. That is especially true in light of the vast differences in the understanding of the term nationalism by the various authors. The flood of 'theories of nationalism' seems quite justly confusing at first sight, but a closer examination will reveal that most accounts attempting historical explanations of nationalism repeat the same basic relations and factors, although using different vocabulary and concepts. The differences between the authors are then mostly in the emphasis on a particular kind of relation. This article will therefore not deal with the authors but rather it will characterize their evaluation of the relevance of the particular relations and contexts: (a) historical background; (b) ethnic and cultural 'roots' and, within these, also denominational relations; (c) the process of modernization, that is, particularly industrialization—and the ensuing changes of social

structure, social communication, and the unevenness of the modernization process; (d) conflicts of interests; and (e) psychological relations.

2.1 Historical Preconditions or the Legacy of the Past

The view that nation-forming was an inevitable result of the historical conditions of the past centuries is now considered a traditionalist residue. Despite that, historians as well as social scientists still consider the past to be an important component of the causal interpretation of the nation-forming process and nationalism. They differ in the strength of the influence they ascribe to history and, particularly, in the analytical perspective. To put it simply, the importance of history is seen from two perspectives: from the perspective of institutions or other objectively operating results of the past development, and from the perspective of 'collective memory' or the construct of national history.

In the former the issue is how institutions, relations, and stereotypes surviving from the past influenced nation-forming: it refers to the state, legislation, and also estates, judiciary, borders, internal administration, the extent of centralism and others. The notion that the 'right' to the recognition of national existence is conditioned by the existence of a 'national' state from the Middle Ages gave rise to the nineteenth-century distinction between 'historical' and 'non-historical' nations, that is, nations with or without their own (political) history. Although this distinction is no longer considered relevant, a number of authors still work implicitly with the idea that some nations (such as the French, the Dutch, and the Portuguese) are a given, 'natural,' and organic part of the state system, while others formed somewhat by 'accident' (the Estonians, the Slovenes, the Bulgarians, and many others). Authors (for example, Shafer, Breuilly, Hobsbawm, Rokkan, Llobera, Tilly, and Barth) who define nation by statehood or understand nationalism as a struggle for statehood, usually assign an important role to political or state history.

A particular problem in analyzing the role of history is that people expressed views and stereotypes which many researchers would today call 'nationalist'—that is, manifestations of national consciousness—already in the Middle Ages and especially in the Early Modern Period—a long time before the modern nation-forming processes. Modern historians agree that the pre-Modern manifestations of national consciousness did not include the entire community but only the privileged classes—the nobility and also sometimes the higher burgher categories (Hagen Schulze, Norbert Elias, and others). Bauer was already aware of this in his periodization of national development.

'Collective memory' is the other perspective on the effects of the past on nation-forming. Authors working within discourse analysis particularly emphasize that the strength of past statehood is not important, but the way the past was understood, interpreted, and worked with—and therefore also manipulated in the national interest. Hobsbawm's phrase 'invented tradition' (1983) is usually quoted in this context, and it usually refers to skepticism toward any real historical continuity. At the same time, it is unthinkable to invent the past without any real evidence. On the other hand, it is important to study the creating of the image, the construct of 'national history' under the conditions of nation-forming.

2.2 Ethnic Roots of Nationalism

It is understandable that most authors who associate nation and nationalism with state do not ascribe much importance to the ethnic and cultural component of the nation-forming process. Some even consider ethnic consciousness neutral, and not automatically a source of nationalism.

On the other hand, a number of authors consider ethnic identities and cultural particulars an important, if not a key precondition of a development toward a modern nation. Anthony Smith (1991, 1986) paid the greatest attention to this relation. He differentiated between the degrees of ethnic identity and considered that an 'ethnie' could also develop its collective memory. Ethnic groups could also enter into conflictual relations, which in many aspects forestalled national conflicts (Horowitz 1985). A number of other authors considered ethnic roots of national existence important, despite their differences in terminology: sometimes they spoke of ethnonationalism, sometimes of nationality, and at other times of primordiality or primordial ethnicity.

Unlike the general role of ethnicity, religion is only mentioned in concrete cases (the Balkans, Ireland), and it is not listed among the necessary preconditions of the nation-forming processes. Nevertheless, several authors acknowledge the indirect role of religion as a modernizing factor in education and communication, particularly as an agent of reformation and counter-reformation (Anderson (1983), Armstrong (1982) and others). The significance of overcoming religiosity with the rise of secularism is also acknowledged. At the same time, some authors see nationalism as a 'national,' secularized religion.

2.3 Nation and Nationalism as Manifestations or Consequences of Modernization

The modern nation-forming doubtlessly occurred during the period of civil revolutions, industrializations, and the beginnings of the free market. That is, during the period called modernization by some, overcoming of traditional society by others, and the transition from feudal to capitalist society by other authors. Most authors (unlike, for example, Kohn)

consider this simultaneous occurrence more than a coincidence, but they construct the causal relations in different contexts. Already in 1953, Deutsch saw the industrial revolution and the onset of the free market as crucial for the spread of nationalism. Later, he considered nationalism a response of people, on the one hand, to the possibilities of improving their status thanks to social mobility, and on the other hand, to the loss of ties to the traditional lordships and settlements.

Gellner's thesis (1983) about industrialization as the fundamental impulse for the birth of nationalism cannot therefore be seen as original. Gellner, however, does not conceive of industrialization in the literal sense referred to by historians, that is, as the introduction of the industrial mode of production. Rather, he means market economy and social mobility in general as factors opening the way for the principles of equality of citizens and their new role in the transformation of culture. Michael Mann (1988) and other authors, such as Etienne Balibar and Immanuel Wallerstein (1991) work with a similar, although more differentiated, thesis. Carolyne M. Vogler (1985) also shares this perspective when she maintains that the 'disjunction between economics and politics' occurred, and that significantly contributed to the dissolution of multinational states and the creation of nation-states. In contrast to the thesis of the role of industralization, other authors emphasize general modernization changes ensuing from the progressing liberation of the market and people (Hroch (1985), Llobera (1994)).

The emergence of the new forms of enterprise (that is, capitalist enterprise) and the related social mobility are specifically reflected in increasing social communication, that is, a system of an increasing amount of information being exchanged among an increasing number of people. As mentioned, Karl W. Deutsch ascribed the key importance to this circumstance on the formation of nation and nationalism, but his one-sided emphasis on communication was not generally accepted. His views were, nevertheless, confirmed within another terminological framework: the thesis about the crucial role of education in the people's ability to accept the notion of a nation as an abstract community became the central argument of Anderson. It was already advocated earlier by Gale Stokes (1974) who proposed that the acceptance of such a notion depended not only on elementary literacy, but also on schooling until the age when a child developed the ability of abstract thinking. The German sociologist Berndt Giesen (1991) connected both of these perspectives: in the conditions of the new type of communication between addressees who did not know each other, a new style of communication and new coding was necessary. The idea of nation became such a code for encoding the new living conditions of modern society.

Numerous researchers argue that the spread of nationalism was boosted by the deepening of the unevenness of development—another consequence of modernization and 'industrialization.' A specific form of unevenness was a 'cultural division of labor between the core and the periphery,' which resulted in the inferior social role of the periphery. This led to the 'internal colonialism' as a breeding ground of national and regional movements opposing the 'center' (Hechter 1975). Mann (1988) and Nairn (1981) also considered uneven development and the associated sentiments of oppression, important factors of nation-forming processes. The view of the nation-mobilizing role of unevenness ('backwardness') has been criticized frequently and its general validity has been rejected.

2.4 *Nationally Relevant Conflicts of Interests*

This brings the discussion to the factor which—again veiled by the various terms—plays an important role for perhaps still the majority of approaches seeking the explanation of nation-forming and also of the spread of nationalism in objective social changes. The thesis of internal colonialism alone suggested that the conflict between the falling-behind periphery and the center could be expressed in the categories of national interest. In other words, uneven development could be counted among nationally relevant conflicts.

A whole range of conflicts from wars to socially and status-motivated tensions can be seen as conflicts of interest. This fact makes any generalization difficult, but it also allows for different interpretations. The most frequently listed conflict in the explanations of nationalism is the conflict ensuing from uneven development. A social conflict between groups or classes which, at the same time, belonged to different ethnicities is theorized relatively seldom in contemporary research (Miroslav Hroch, Paul James, Tom Nairn, Eric Hobsbawm), as is the idea of the struggle against an old society in the national interest (Boyd Shafer).

2.5 *Nationalism as a Result of Psychological, Cultural, and Emotional Manipulation*

Most researchers working on the explanations of nationalism also look for its sources in the realm of psychology and emotions. The individual models, however, differ significantly according to which aspect of human psyche they consider important. Insecurity of a human individual, the need to belong to a group and to identify with it, are some of the most frequently quoted aspects. The insecurity is then seen as being caused by perceived (or real) endangerment, while the source of endangerment can be either external danger or internal insecurity resulting from the perception of uprootedness, dissolution of old securities, old faith, etc. (Miroslav Hroch, Gale Stokes, Anthony Smith, Bernd Giesen, Ernest Gellner). Nationalism is sometimes considered a sick, pathological reaction to such

perceived endangerment (Committee on International Relations 1987), and at other times a response to the inferiority complex (Lemberg 1964). Anderson speaks of the need to overcome the fear of an individual death by identifying with an immortal nation.

Following partially from Geertz's primordial ties, some researchers tried to apply biological research on animal behavior to nationalism, and explain nationalist attitudes as the 'territorial imperative' (Snyder 1975). Nevertheless, this approach has remained marginal in the overall context.

'All the stronger' has recently become the effort to explain nationalism as a cultural construct deriving from the activities aiming at the emotional mobilization of human psyche. Thus national rituals are becoming a subject of intense research (George Mosse, 1973), as well as symbols, celebrations in the name of nation, and pilgrimages to places commemorating the national past (Nora 1986)—in short, activities deliberately aimed at the emotional element of the human psyche.

Overall, differences in the explanations of nation-formation are not as fundamental as it may appear on the surface from the different terminology and the verbal radicalism of some authors. The understandable efforts for originality and impressive contribution are often a façade of new expressions repeating older conclusions and findings.

Most of the studies on the historical origins of nationalism are flawed in that they ignore the fact that the process of nation-forming was present in two basic types in Europe: the first started with state-nations which had existed since the Middle Ages and began to transform into modern nations at the end of the eighteenth century or later. The point of departure of the second type were non-dominant ethnic groups which inhabited the territories of large multi-ethnic empires, did not have ruling élites, had their own cultural traditions in a codified language, and had either disrupted or no tradition of statehood. That is why it is important to distinguish between nationalism (whether considered as a part of national identity or merely as its nationally egoistic variant) developed in the conditions of state-nations and the kind of nationalism known from national movements. The latter type again differs depending on whether it developed under conditions of national agitation or a mass national movement. In any case, nationalism should never be viewed as a theoretically sound 'ideology' comparable to liberalism, clericalism, or socialism. It has always been a mixture of variously structured stereotypes which, in the environment of forming a fully fledged nation, had the ability to attach itself to any political ideology and live with it in a symbiotic or parasitic relationship.

(Translated by Libora Oates-Indruchová)

See also: Clan; Cold War, The; Conservatism: Historical Aspects; Environmental History; Ethnic Groups/Ethnicity: Historical Aspects; Medicine, History of; National Socialism and Fascism; Nationalism and Expressive Forms; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: Arab World; Nationalism, Sociology of; Nations and Nation-states in History; Nation-states, Nationalism, and Gender; Power in Society; Primary Elections; Rationality in Society; Religion: Nationalism and Identity; Revolutions, History of; State Formation; State, History of; Western European Studies: Culture; Western European Studies: Gender and History

Bibliography

Anderson B R 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Armstrong J A 1982 *Nations before Nationalism*. University of Carolina Press, Chapel Hill, NC

Balibar E, Wallerstein I 1991 *Race, Nation, Class: Ambiguous Identities*. Verso, London, New York

Bauer O 1907 *Die Nationalitätenfrage und die Sozialdemokratie* [The issue of nationality and social democracy]. Verlag der Wiener Volksbuchhandlung, Vienna

Carr E H 1945 *Nationalism and After*. Macmillan, London

Breuilly J 1982 *Nationalism and the State*. St. Martin's Press, New York

Brubaker R 1996 *Nationalism Reframed: Nationhood and the National Question in the New Europe*. Cambridge University Press, Cambridge, NY

Connor W 1994 *Ethnonationalism: The Quest for Understanding*. Princeton University Press, Princeton, NJ

Committee on International Relations, Group for the Advancement of Psychiatry 1987 *US and THEM: the Psychology of Ethnonationalism*. Brunner/Mazel, New York

Deutsch K W 1953 *Nationalism and Social Communication: An Inquiry into the Foundations of Nationality*. Technology Press of the Massachusetts Institute of Technology, Cambridge, MA

Doob L W 1964 *Patriotism and Nationalism: Their Psychological Foundations*. Yale University Press, New Haven, CT

Gellner E 1983 *Nations and Nationalism*. Basil Blackwell, London

Gellner E 1997 *Nationalism*. Weidenfels & Nicolson, London

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Giddens A 1987 The nation-states and violence. In: Giddens A *Social Theory and Modern Sociology*. Stanford University Press, Stanford, CA

Giesen B (ed.) 1991 *Nationale and kulturelle Identität: Studien zur Entwicklung des kollektiven Bewusstseins* [*National and Cultural Identity: Studies on the Development of Collective Consciousness*]. Suhrkamp, Frankfurt am Main, Germany

Hayes C J H 1926 *Essays on Nationalism*. Macmillan, New York

Hayes C J H 1931 *The Historical Evolution of Modern Nationalism*. R R Smith, Inc., New York

Hechter M 1975 *Internal Colonialism: the Celtic Fringe in British National Development 1536–1966*. Routledge & Kegan Paul, London

Hertz F 1945 *Nationality in History and Politics: a Psychology and Sociology of National Sentiment and Character*. Kegan Paul, Trench, Trubner & Co., London

Hobsbawn E J 1990 *Nations and Nationalism since 1780: Programme, Myth, Reality*. Cambridge University Press, Cambridge, NY
Hobsbawm E J, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK
Horowitz D L 1985 *Ethnic Groups in Conflict*. University of California Press, Los Angeles
Hroch M 1985 *Social Preconditions of National Revival in Europe: a Comparative Analysis of the Social Composition of Patriotic Groups among the Smaller European Nations*. Cambridge University Press, Cambridge, NY
James P 1996 *Nation Formation: Towards a Theory of Abstract Community*. Sage, London
Kedourie E 1960 *Nationalism*. Hutchinson, London
Kemiläinen A 1964 Nationalism: problems concerning the word, the concept and classification. *Studia Historica Jyväskylänsia III*, Jyväskylä, Finland
Kohn H 1944 *The Idea of Nationalism: a Study in its Origins and Background*. Macmillan, New York
Lemberg E 1964 *Nationalismus I.II*. Rowohlt Reinbeck, Hamburg, Germany
Llobera J R 1994 *The God of Modernity: the Development of Nationalism in Western Europe*. Providence, Oxford
Mann M 1988 *States, War and Capitalism: Studies in Political Sociology*. Blackwell, Oxford, UK
Meinecke F 1907/1911 *Weltbürgertum und Nationalstaat* [The Weltbürgertum and the Nation-State]. R. Oldenbourg, München/Berlin, Germany
Minogue K R 1967 *Nationalism*. Batsford, London
Mosse G L 1973 *The Nationalization of the Masses*. Howard Fertig, New York
Nairn T 1981 *The Break-up of Britain: Crisis and Neonationalism*, 2nd exp. edn. NL Bond/Verso, London
Nora P 1986 *Les Lieux de Mémoire* [*The Locations of Memory*], *II. 1–3 La Nation*,. Gallimard, Paris
Renan E 1888/1947 Qu'est-ce qu'une nation? (What is a nation?] In: *Oeuvres completes I*. Calman Lévy, Paris
Samuel R (ed.) 1989 *Patriotism: The Making and Unmaking of British National Identity*. Routledge, London/New York
Schieder T 1991 *Nationalismus und Nationalstaat: Studien zum nationalen Problem im modernen Europa*. Vandenhoeck & Ruprecht, Göttingen, Germany [Nationalism and nation-state: studies of the national problem in modern Europe]
Schulze H 1996 *Staat und Nation in der europäischen Geschichte* [*State and Nation in European History*]. Verlag C. H. Beck, Munich, Germany
Seton-Watson H 1977 *Nations and States: An Inquiry into the Origins of Nations and the Politics of Nationalism*. Methuen, London
Shafer B C 1955 *Nationalism: Myth and Reality*. Harcourt, Brace and Comp., New York
Smith A D 1986 *The Ethnic Origins of Nations*. Blackwell, Oxford, UK
Smith A D 1991 *National Identity*. Penguin Books, Harmondsworth, UK
Snyder L L 1954 *The Meaning of Nationalism*. Rutgers University Press, New Brunswick, NJ
Snyder L L 1975 Nationalism and the territorial imperative. *Canadian Review in Studies on Nationalism* **3**: 17–30
Snyder L L 1976 *Varieties of Nationalism: a Comparative Study*. Dryden Press, Hinsdale, IL
Sorokin P 1947 *Society, Culture and Personality: Their Structure and Dynamics, A System of General Sociology*. Harper, New York
Stokes G 1974 Cognition and the function of nationalism. *Journal of Interdisciplinary History*. **4**: 525–42
Tägil S (ed.) 1984 *Regions in Upheaval. Lund Studies in International History*. Essette Studium, Solna, Sweden, Vol. 22
Vogler C M 1985 *The Nation State: The Neglected Dimension of Class*. Brookfield Gower, Aldershot

M. Hroch

Nationalism, Sociology of

The study of nationalism is a fairly recent development in sociology, a discipline which took for granted the nation-state as the territorial framework of its main subject of investigation—modern societies. Nevertheless, several sociological approaches to the study of nationalism can be distinguished: Weberian, structural-functionalism, modernization theory, neo-Marxism, and a variety of comparative-historical and social-constructionist approaches.

One recurrent theme in sociological debates about nationalism concerns the nature of the relation (historical and conceptual) between several interrelated phenomena: ethnic group, nation, and nationalism. Each one of these phenomena has been seen either as a historical antecedent of the others or somehow accorded priority. Thus, while one of the most influential theorists argues that 'nationalism created nations' (Gellner), another sees nationalism as a 'subjective reflection' of the group reality of the nation on the level of social consciousness (Hroch). In addition, whereas some emphasize the 'ethnic origins of nations' (Smith), others argue that 'ethnic nationalism' is a latter-day ideological development (Greenfeld). A second line of sociological debate has separated modernization theorists and neo-Marxists who explain the emergence of nationalism by reference to economic factors or the functional requirements of industrial society, and historical sociologists who question this view on a variety of grounds. An important obstacle to sociology in this field is imposed by the subject itself: as the most salient form of particularism in the modern world, nationalism defies easy generalization.

1. *Nationalism and Classical Sociology*

Nationalism was not a central concern in classical sociology. This curious lacuna must be understood in light of the history of a new discipline preoccupied with demarcating its own field of study—society as a reality *sui generis* (Durkheim). A second reason might be sought in the intellectual bias of sociology itself:

political phenomena were seen as derivative of more fundamental social realities. Finally, sociological interest was consumed by the dramatic transformations caused by industrialism and capitalism. As a result, concentration on the problems of industrial society overshadowed the concern with its overall territorial framework—the modern nation-state. The rise of imperialism and mass politics at the turn of the century changed this equation.

1.1 Max Weber

Weber's views on the nation must be understood in this historical context. For him, the unprecedented fusion of culture and politics in the modern world offers the main clue to understanding the nation: its *differentia specifica* as a social group lies in the striving for territorial political power on the basis of a shared culture. This symbiosis of culture and power in the nation-state resulted from a twofold dependency: if the state's legitimacy increasingly depended on nationalist appeals in the age of mass politics, the nation needed the state for the protection of its unique culture. By combining the status concerns of intellectual strata as the bearers of national culture with the *Realpolitik* ambitions of political elites, nationalism provided the state with a new source of legitimacy, greatly increasing its mobilization potential. The sacralization of the state as the guardian of cultural values, in turn, inevitably transformed national conflicts, both within and between states, into 'a struggle of life and death.'

Second, Weber attempted to explain national solidarity. In contrast to ethnic groups whose members must share a 'subjective belief in common descent,' national solidarity does not depend on a myth of common origin (Switzerland), or 'objective' markers of status differentiation (race, religion, language). Thus, the loyalty of German-speaking Alsatians to France was derived from the historical experience of a revolutionary regime which abolished feudalism. Conversely, the ethnic and linguistic affinity of Serbs and Croats did not obliterate their national differences. Consequently, Weber argued that common historical experiences and shared memories play a decisive role in forging the nation as a community of 'political destiny.'

Weber's third point concerns the mass appeal of nationalism: as the only source of status superiority available to the masses in a secularized world (ethnic or national honor), nationalism cuts across class lines in defiance of socialist expectations. An additional reason for nationalism's appeal to the lower strata has to do with status compensation: in many instances, the denigration of 'inferior' ethnic or racial groups by the lower classes of the dominant group (such as the emancipated black Americans by the poor whites in the southern USA) is the very precondition of the latter's status superiority. These observations do not amount to a theory of nationalism; however, Weber's main intuition—that the fusion of culture and power constituted the *differentia specifica* of the nation as a new type of sociological community—became a cornerstone of all later theories of nationalism.

1.2 Emile Durkheim and Marcel Mauss

In *The Elementary Forms of Religious Life* (1995) Durkheim notes the similarity between the collective effervescence of religious rituals and revolutionary-patriotic festivals: in the Jacobin phase of the French Revolution, the secular values of fatherland, liberty, and reason became the core values of a revolutionary-patriotic religion whose object of worship was society itself. For Durkheim, this self-worship of society was another confirmation of his theory of religion; the connection with nationalism as a 'secular faith' remained unexplored in his work. On the other hand, as a committed advocate of secular schools in the Third Republic, Durkheim envisaged an important role for civic patriotism in 'moral education.' During World War I, Durkheim favorably contrasted this ideal of civic patriotism with expansionist German nationalism, whose excesses he interpreted as a sign of collective *anomie*.

A Durkheimian view of the nation was elaborated further by Mauss. Mauss saw nationalism as a modern phenomenon originating in the doctrine of popular sovereignty espoused by the American and French revolutionaries. Sociologically, nationalism's success as a movement was predicated on the presence of three general factors: (a) a strong central power which successfully broke down traditional affiliations of tribe, clan, and estate; (b) a morally integrated society consisting of citizens who identified with the state and its laws; and (c) economic unity. Conversely, the belief in a unique racial or civilizational heritage was a retrospective elaboration of nationalists: its importance lay in cementing an already existing social bond on the level of consciousness. For Durkheim and Mauss, the nation was first and foremost a community of citizens: their understanding of 'ethnic nationalism' was limited.

2. Structural-functionalism

A new sociological interest in nationalism developed as a result of the proliferation of independent states after World War II. The dominant sociological school of the postwar decades—structural-functionalism—viewed nationalism as a by-product of the structural strains caused by industrialization.

Thus, Smelser saw nationalism as a response to the state of collective *anomie* induced by structural dif-

ferentiation (division of labor, role specialization, etc.). As a secular ideology of industrialization, nationalism overcomes the resistance of traditional structures, mobilizes new constituencies, and gives meaning to the sacrifices required for social change. Conversely, once differentiation proceeds apace, nationalism can take the form of a nativist reaction against the corroding influences of further modernization.

This logic was generalized by Shils, who built on Durkheim's idea about secular nationalism as the functional equivalent of religion: the transition from loosely integrated traditional societies to the modern nation-state is predicated on the gradual penetration of the various social and geographical peripheries (kinship, ethnic, religious, and regional) by the 'charismatic center,' seen as the embodiment of a society's core values. By allowing the members of society to partake in the institutionalized charisma of a sacralized political center through the exercise of citizenship rights, the nation-state succeeds in breaking the power of tradition. This sacralization of the secular is of special significance in transitional societies: by transposing the charisma of the tribe or religion onto the *nation*, the leaders of new states are able to attract mass support and legitimate social change.

Functionalist explanations were criticized on the following grounds: (a) there is no inevitable correlation between structural differentiation and nationalism; (b) teleology: the 'end-state' is posited by the observer and used to explain the past; (c) the desirability of meeting select 'functional goals' is subject to elite contest; and (d) nationalism is a secular ideology more similar to other modern ideologies than to religion.

3. *Modernization Theory*

A significantly modified functionalist modernization approach was developed subsequently by Karl Deutsch and Ernst Gellner. The formulations of these two theorists have remained influential to this day.

3.1 *Karl Deutsch*

Deutsch (1966) begins with a distinction between society and culture. 'Society' refers to the division of labor and social stratification; 'culture,' in contrast, is a set of 'stable, habitual, preferences' based on values inculcated through socialization. Societies 'produce, select, and channel goods and services'; culture 'stores' *information* about the past, and recombines it in the present, enabling political authority to function. The technological changes and complex division of labor induced by the Industrial Revolution resulted in a growing need for expanding channels of cultural communication. Simultaneously, the breakdown of traditional status barriers and exposure to the insecurity of the market raised the aspirations of the socially mobilized masses, already agitated by democratic ideas. These developments favored the emergence of the nation as a *horizontal functional group* linked by complementary communication channels, typically language (e.g., Czechs versus Germans in Bohemia), but also shared memories or historical experiences (e.g., Switzerland). As *the* ideology of popular sovereignty, nationalism held a special attraction for the socially mobilized masses, whose claims to status equality and upward social mobility it helped justify.

Deutsch's theory is empirical and predictive: the complementarity of communication channels can be assessed by social-psychological indicators; rates of social mobilization can be measured by urbanization, mass literacy, newspaper circulation, etc.; and assimilation depends on the extent and content of communication between in- and out-groups. The theory runs counter to the standard predictions of Parsonian structural-functionalism: the expansion of communications strengthens nationalism, not universalism.

3.2 *Ernst Gellner*

Gellner (1983) starts with an ideal type of agrarian society: their central feature is an ideologically buttressed functional division of labor which separates the 'high culture' of a hereditary administrative-military ruling class and universal clerisy from the 'low cultures' of socially isolated and illiterate peasant communities. This status segregation has a cognitive dimension as well: the world is experienced as culturally and ontologically heterogeneous.

A profound change occurs with the modern cognitive revolution: the world is now seen as a coherent whole subject to universal laws expressed in a unitary linguistic idiom. The social correlate of this cognitive revolution is industrialism. If intellectual progress presupposes the perpetual exploration of reality, the idea of unlimited growth demands the constant redrawing of traditional social boundaries and roles in line with the functional requirements of the division of labor. These requirements can be met only by a common linguistic idiom transmitted through standardized education. Literacy in a shared language prepares individuals for new functional roles, increases their mobility prospects, and facilitates communication among strangers in an impersonal world. The emergence of a shared culture favors nationalism as a political principle which holds that the state must rest on the foundation of national culture; the state, in

turn, acquires a new source of legitimacy (compare Weber). Thus, industrialism is a necessary condition of nationalism. Ethnicity is secondary—it is nationalism which invents nations.

Finally, nationalism disseminates through uneven development. In a typical imperial (e.g., Habsburg) situation an ethnic division of labor is present: the carriers of high culture belong to one ethnic group (or groups), those of low culture to another (or others). The social exclusion of the aspiring intelligentsia of the subordinate group leads it to adopt 'ethnic nationalism' as a strategy of collective mobility (e.g., Czech nationalism). In contrast, 'diaspora nationalism' develops among politically excluded 'high-culture' pariah groups (e.g., Jews). The ideal case is a mature industrial society in which elites and masses share a standardized idiom and consider themselves conationals in a common state.

Gellner's elegant theory has enjoyed enormous influence. It has been criticized on several grounds: (a) the theory does not specify a causal (group) agent: nationalism is seen as a by-product of the impersonal process of industrialization; (b) nationalism can precede industrialization (e.g., Balkan nationalism); (c) nations can precede nationalism (see Sect. 4); and (d) nations are rooted in premodern, ethnic identities (see Sect. 5).

4. *Neo-Marxism*

The classics of Marxism failed to develop a consistent sociological theory of the nation (Szporluk 1988). Although partially filled by Austro-Marxism (Bauer, Renner) and the Bolshevik ideologists (Lenin, Stalin), this lacuna was addressed seriously only after World War II. A strong impetus to the Marxist sociology of nationalism was given by Immanuel Wallerstein's world systems theory: in the underdeveloped periphery of the world capitalist system, class struggle (indigenous proletariat and peasantry versus the 'comprador' bourgeoisie) and national conflict (the exploited nation versus the world capitalist metropole) were superimposed upon each other.

4.1 *Class and Peripheral Nationalism*

Hechter (1975) extended the metropole-periphery logic to ethnic conflict in the UK. In defiance of functionalist logic, the incorporation of the (Celtic) periphery by the (English) center did not result in the creation of a homogenous national society, but a dynamic of 'internal colonialism' instead. The subordination of the underdeveloped periphery was ensured by its economic dependence on a developed center and a rigid ethnic division of labor which enabled the English to monopolize power and status privileges. The persistence of ascriptive status identities and the superimposition of class and ethnic conflicts run counter to the expectations of modernization theory: instead of integration and consensus, uneven development leads to the resurgence of peripheral nationalism.

Studying the same case, Nairn (1977) came to a different conclusion: peripheral 'neonationalism' could be caused by 'relative overdevelopement' as well. Thus, the resurgence of Scottish neonationalism was the result of the relative economic decline of the English center and new opportunities opened by oil exploitation in the North Sea. A newly awakened Scottish bourgeoisie in search of profit appealed to nationalism in its struggle against a parasitic center. In a larger historical perspective, nationalism was the product of the ideological reaction of the awakened intelligentsia of the semiperiphery to uneven development and imperialist exploitation of their countries (not yet 'nations') by the centers of world capitalism (England and France). The nationalism of the center, in turn, developed in a 'dialectical countermobilization' against its semiperipheral counterparts.

4.2 *Class and Nation. The Social Base of Nationalist Movements*

The relationship between class and nation is a central concern in Hroch's (1985) study of the developmental phases and social base of nineteenth-century national movements in small European nations. For Hroch, the nation is an 'objective social category' rooted in the memory of a shared *ethnic* past, dense linguistic and sociocultural ties, and a conception of citizenship. National awareness and nationalism are only a 'subjective reflection' of this reality on the level of consciousness (contrast Gellner).

Hroch distinguished three phases in the development of national movements: in phase A (ethnographic) linguists and folklorists rediscover a shared past; in phase B (agitation) intellectuals promote national awareness among the masses; and finally, in phase C, a mass national movement arises. While the social base of nationalist movements varied, especially in phase B (teachers in Slovakia, the petty bourgeoisie in Norway), the social mobilization of the bourgeoisie and peasantry was a necessary condition for success in phase C: an independent nation requires 'a fully developed class structure.' Another necessary condition of nationalist movements was the superimposition of ethnic and class conflicts: only when membership in a small nation appeared as a decisive disadvantage could 'social contradictions' be expressed in national terms. Hroch did not conflate class and nation; however, he did argue that a successful national movement must be based on distinct class interests.

Neo-Marxist approaches have been criticized for: (a) an unwarranted 'materialist' reduction of national

to class conflicts; (b) poor generalizability (internal colonialism); and (c) inability to explain the nationalism of the 'metropole' of the world system (England, France), and/or dominant nations (e.g., Russia).

5. Comparative-historical Sociology

Historical sociologists have questioned functionalist and Marxist explanations in two ways: (a) by emphasizing the premodern, ethnic roots of nationalism; and (b) by demonstrating that nationalism predates industrialization.

5.1 Ethnic Groups and Nations

The first type of argument is developed by Smith (1986). Smith rejects the view that nations are based on *primordial* ties (e.g., race); however, he states that they *are* rooted in premodern *ethnie*. *Ethnie*, in turn, are not based on ascriptive criteria, but upon intergenerationally transmitted 'constitutive myths' (compare Weber). Typically, such constitutive myths include or imply a collective name, a myth of common descent, a shared historical narrative (legends, epics), belief in a unique culture, association with a territory, and solidarity vis-à-vis out-groups. The transformation from *ethnie* to nations occurred under the impact of the triple revolution of administrative centralization, capitalism, and 'cultural coordination.' Capitalism helped bind social classes into a unified economic community; political centralization, warfare, and cultural coordination (the rise of vernacular languages and mass culture) forged the link between state and nation. Whereas nation-building paths varied in accordance with historical timing, religious identifications, the character of the constitutive myth, and its elite carriers (aristocracy, bourgeoisie, intelligentsia), two broad trajectories emerged as a result of uneven development: (a) the early 'Western' path in which the nation was defined as a territorial community of citizens bound by laws; and (b) the late 'German' (East European) route in which *ethnie* served as the basis for statehood. However, Western nations also required a 'constitutive myth': premodern cultural identities were crucial everywhere, and industrialization was only one factor which helped transform *ethnie* into nations.

5.2 Nationalism as Ideology

Another approach is exemplified by Greenfeld (1992): nationalism is an antecedent of industrialization. Greenfeld traces the idea of the nation to a long process of semantic transformation: originally applied to 'groups of students' in medieval universities and the *elite* of a 'community of opinion,' in sixteenth-century England the 'nation' was extended to include the whole people. This momentous change from the medieval (people = rabble) to the modern *Weltanschauung* (people = elite) was caused by the structural crisis of feudal society. In the first nation, England, the decimation of the aristocracy and the decline of the clergy (War of the Roses, separation from Rome) shattered traditional status barriers, opening channels of social mobility to gentry and educated commoners. The profound 'status inconsistency' felt by members of the elite led them to 'discover' the nation: the view that status inequality was secondary to membership in a community of citizens resolved their identity crisis and legitimized social mobility across status lines. English nationalism culminated in popular sovereignty and an *individualistic* and *civic* ideological self-definition: the 'people' was a composite of individuals, and membership was open to all citizens (likewise the USA).

The diffusion of nationalism was a consequence of England's hegemonic status. Thus, the eighteenth-century French nobility, squeezed between the absolute monarchy and wealthy commoners, imitated its English counterpart and competitor, finding in nationalism a solution for its own identity crisis. However, while membership in the French nation was open to all (as in civic nationalism), the nation came to be seen as expressing a single collective will (*volonté générale*)—thus, French nationalism was collectivist and civic. A further semantic transformation occurred in Germany and Russia, where the term was associated with *ethnicity*, the unique character of *a people*—nationalism became both collectivist and *ethnic*. The motor force behind this change was *ressentiment*—the profound existential envy felt by the elites of follower societies toward the 'superior' West. Hence the compensatory glorification of the *Volk* with all its authoritarian implications.

Greenfeld's conclusions are original: (a) democracy emerged from individualist and civic nationalisms; (b) ethnic nationalism is a later ideological elaboration; (c) the historical carrier of nationalism is not the bourgeoisie, but aristocracy or intelligentsia (Germany); and (d) capitalism and democracy could flourish only once the idea of the nation delegitimized traditional status barriers.

Greenfeld's theory avoids 'materialism' and teleological functionalism; however, it has been criticized on both theoretical and empirical grounds: (a) the dichotomy between Western, civic (whether individualist or collectivist) and Eastern, ethnic nationalisms is far from clear-cut, as Western nations were based on an ethnocultural foundation as well; (b) early national consciousness should not be equated with a fully developed ideological nationalism; (c) the original ideological code of every nationalism is subject to reinterpretation and change; and (d) cultural determinism—some nations (the USA, England, France) appear destined for democracy, while others seem doomed to authoritarianism (Germany, Russia).

5.3 Nationalism as Cultural Idiom

The continued salience of 'civic' vs. 'ethnic' national self-definitions is documented by Brubaker (1992). In France, the republican tradition defined the nation as a territorial 'community of citizens': membership was not tied to ethnicity. In Germany, a territorially divided nation was defined as a 'community of descent': all (and only) Germans were members. The persistence of these *cultural idioms* (not instrumental interests) helps to explain the difference in immigration policies: the French is assimilationist and inclusive, the German 'ethnic' and exclusive.

6. Social-constructionist Approaches

6.1 Nationhood as Institutionalized Form

In a subsequent work, Brubaker (1996) challenges the fallacy of sociological 'group realism' from the standpoint of the new sociological institutionalism: instead of treating the nation as a 'substantial entity, collectivity, or community' which can be 'objectively defined,' sociologists should study 'nationhood' from the point of view of its political and cultural institutionalization, both within and among states. Thus, the dramatic collapse of the USSR has to be understood as a consequence of the ethnoterritorial institutionalization of nationhood both on the group (ethnic federalism) and individual (personal identity document) levels. Such institutionalization turned ethnicity into an 'organizing principle of social classification,' constituting the interests of social groups and individuals, serving as a basis for political action, and informing practical conduct in everyday life. In this sense, Brubaker argues, 'nationalism' is not so much 'engendered by nations,' as induced by particular institutional arrangements.

6.2 Nation as Invented Tradition

The 'socially constructed' character of all collective and, *ipso facto*, national identity is emphasized by a variety of contemporary authors. One influential version of the argument is developed by Hobsbawm and Ranger (1983): the nation is largely a product of 'invented tradition.' By 'invented tradition' is meant 'a set of practices, normally governed by overtly or tacitly accepted rules and of a ritual or symbolic nature, which seek to inculcate certain values and norms of behavior by repetition, which automatically implies continuity with the past' (Hobsbawm and Ranger 1983, p. 1). The modern nation is a particularly good example of an invented tradition for, although the nation-state with its well-demarcated territory and standardized national language is an eminently modern phenomenon, nationalists typically claim it be a 'natural' community rooted in the 'remotest antiquity.' In actual fact, national symbols, flags, languages, and histories are a product of nationalist 'social engineering,' while the connection with the remote past is retrospectively established by nationalist imagination. It follows from this that the main task of the historian and sociologist is to study those symbolic and discursive practices that helped establish the particular nation's idea of itself as a 'natural' community in the first place.

6.3 Nation as Imagined Community

A related argument is developed by Anderson (1991), who sees the nation as a new type of limited and sovereign 'imagined community.' The nation is an *imagined* community because most of its members will never know each other; it is imagined as a *limited* community because it is externally bounded by other nations; and it is seen as *sovereign* because it is historically rooted in the ideas of internal and external freedom (popular rule and national independence). Finally, the nation is imagined as a *community* because it is 'always conceived as a deep, horizontal comradeship' (Anderson 1991, p. 7). The emotional resonance of the nation stems from the fact that it replaces kin and religion as the main loci of individual loyalty and self-sacrifice in the modern era.

According to Anderson, the emergence of the nation as a new type of imagined community was predicated on the presence of certain definite cultural and material conditions: (a) linguistic diversity; (b) the increasing relevance of vernacular languages and literacy after the Protestant Reformation; and (c) 'print-capitalism'—the production of books, newspapers, and journals for mass cultural consumption. Once set in motion, print-capitalism gave an unprecedented stimulus to the development of nationalist discourse, albeit under a variety of historical circumstances that require separate study.

Social-constructionists correctly attacked the fallacy of group realism. However, their views can be subjected to the following criticisms: (a) while the nation might be 'invented,' this social construction must be based on a preexisting reality, be it the memory of historical statehood (e.g., Croatia, Poland), ethnicity (e.g., Germany), or a more complex combination of various factors (e.g., Austria); (b) in view of the above, modern national identity might be 'invented' but it is not 'fabricated'; (c) political factors (warfare, state centralization), and elite (or class) interests played an equally important role in the emergence of nationalism as 'print-capitalism'; and (d) the notion of 'imagined community' applies to other kinds of communities besides the nation (e.g., class, religious denomination).

7. Conclusion

Despite the fact that nationalism did not stand at the center of sociological interest, especially in the classical epoch of the discipline, sociologists have studied the topic from a variety of perspectives. Even so, a considerable area of agreement has emerged. For one, sociologists agree that the nation is not based on 'primordial ties'; rather its emergence as a 'community of citizens' was related to the rise of the modern state (administrative centralization), the novel idea of popular sovereignty, industrialization, and the expansion of cultural communications (vernacular language literacy). Uneven development was crucial for the dissemination of nationalism, whether on account of the 'ethnic division of labor,' the reaction of the 'exploited periphery,' or *ressentiment*. The salience of *ethnicity* (itself not an ascriptive but a 'subjective' category) was greatest in areas where class and ethnic conflicts were superimposed (e.g., Habsburg empire), where the putative nation lacked a state (e.g., Poland), and where ideology conceptualized the nation as *Volk* (e.g., Germany). Conversely, ethnicity was secondary in established states (e.g., England), or where shared memories and experiences favored other identifications (e.g., Switzerland). Finally, the consolidation of national identity everywhere involved a considerable effort in the 'invention of tradition.' The intimate relation between modernity and nationalism defies liberal and Marxist expectations of a 'unified world'; this may well be the most provocative insight of the sociology of nationalism.

These areas of agreement point to a considerable, if still relatively modest progress in the sociology of nationalism. Still, many other areas await further investigation: the relations between religion and nationalism, nationalism and collective violence, nationalism and collective memory, diaspora and homeland nationalism, nationalism and gender, and the nationalist narrative. The reluctance of sociologists is understandable—to an extent every nationalism tells its own story, and this presents obstacles to generalization. Yet, as Max Weber demonstrated, the uniqueness of a phenomenon neither diminishes its importance nor precludes sociological investigation: a conceptually informed comparative-historical sociology can develop middle-range theories while taking into account the peculiarities of individual instances.

See also: Deutsch, Karl W (1912–92); Gellner, Ernest (1925–95); Ideology: Political Aspects; Marxism in Contemporary Sociology; Nation: Sociological Aspects; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: East Asia; Nationalism, Historical Aspects of: South Asia; Nationalism, Historical Aspects of: The West; Nations and Nation-states in History; Rational Choice Theory in Sociology; State, Sociology of the; Weber, Max (1864–1920)

Bibliography

Anderson B 1991 *Imagined Communities. Reflections on the Origins and Spread of Nationalism*. Verso, London

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Brubaker R 1996 *Nationalism Reframed. Nationalism and the National Question in the New Europe*. Cambridge University Press, Cambridge, UK

Connor W 1994 *Ethno-nationalism. The Quest for Understanding*. Princeton University Press, Princeton, NJ

Deutsch K 1966 *Nationalism and Social Communication*. MIT Press, Cambridge, MA

Durkheim E 1995 *The Elementary Forms of Religious Life*. Free Press, New York

Gellner E 1983 *Nations and Nationalism*. Cornell University Press, Ithaca, NY

Greenfeld L 1992 *Nationalism. Five Roads to Modernity*. Harvard University Press, Cambridge, MA

Hall J A (ed.) 1998 *The State of the Nation. Ernst Gellner and the Theory of Nationalism*. Cambridge University Press, Cambridge, UK

Hechter M 1975 *Internal Colonialism*. California Press, Berkeley, CA

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Hroch M 1985 *The Social Preconditions of the National Revival in Europe*. Cambridge University Press, London

Mauss M 1969 La nation. In: Mauss M (ed.) *Oeuvres*. Minuit, Paris, Vol. 3, pp. 573–639

McCrone D 1998 *The Sociology of Nationalism*. Routledge, London

Motyl, Alexander et al. 2001 *Encyclopedia of Nationalism* (2 Vols). Academic Press, San Diego, CA

Nairn T 1977 *The Break-up of Britain*. Humanities Press, London

Schnapper D 1998 *The Community of Citizens. On the Modern Idea of Nationality*. Transaction Publishers, New Brunswick, NJ

Shils E 1975 *Center and Periphery. Essays in Macrosociology*. Chicago University Press, Chicago

Smelser N 1966 *Essays in Sociological Explanation*. Prentice-Hall, Englewood Cliffs, NJ

Smith A 1986 *The Ethnic Origin of Nations*. Blackwell, Oxford, UK

Smith A 1998 *Nationalism and Modernism*. Routledge, London

Szporluk R 1988 *Communism and Nationalism. Karl Marx vs. Friedrich List*. Oxford University Press, Oxford, UK

Weber M 1978 *Economy and Society*. University of California Press, Berkeley, CA

V. Vujacic

Nations and Nation-states in History

This survey starts by asking whether nations existed before the establishment of nation-states. It then considers the formation of national states before the development of nationalist movements. The next two

sections look at nation-state formation in and beyond Europe around 1850 to 1970. A final section considers the place of nations in a world of nation-states.

1. Nations Before Nation-states?

Nationalists answer this question in the affirmative. In the beginning was the nation. As the nation acquired consciousness of itself, there arose explicit demands and movements (nationalism) seeking autonomy, usually a nation-state. The nationalist sequence is: nation > nationalism > nation-state.

Nationalist historiography locates the 'nation' far in the past. Hermann Arminius is celebrated by German nationalists for his victory over Roman legions in AD 9. The feudal kingdoms succeeding the Roman Empire are presented as national—western Frankish kingdom/France, Anglo-Saxon kingdoms/England. The Capetian and Norman dynasties found the modern French and British nation-states. Nationalist etymologies cite references to English, French, Germans, Italians, etc.

This view has superficial plausibility for cases like England. Nationalist historians of subordinate cultural groups (e.g., Slavs) or stateless, nonliterate societies, have difficulties. Political fragmentation presents problems, even given evidence of cultural identity (e.g., Italy, Germany). There were premodern societies which stressed ethnic/religious unity between elites and non-elites against foreign rule (e.g., Jewish resistance to Rome). However, dominant states were marked by divisions between elites of priests, warriors, and officials, usually serving a monarch, and the mass of subjects. Elites referred to ethnicity, but this was not sustained and subordinated to religious and dynastic values. Subjects were transferred from one ruler to another without reference to their 'nationality.' Ethnicity had a marginal role in premodern politics. Indeed, the puzzle would be if elites wished to identify with those they ruled.

A more significant debate concerns the relationship of premodern ethnicity to modern nations and nation-states. One view is that modern nationality is based upon ethnicity. An opposing view is that ethnicity is little more than an arsenal upon which nationalists can draw for historical myths, which admittedly are important. There is agreement that nationalism and the nation-state are modern. However, just *how* modern, what the relationship is between state, nationalism, and nation-state, and why the nation-state has become the dominant political unit of modern times, are matters of debate (Armstrong 1982, Hastings 1997, Smith 1986).

2. Nation-states Before Nationalism

In Europe at around 1500 there was no idea that the political order was only legitimate if national. Germany and Italy were fragmented. By 1600 some voices called for national unity in Italy and the United Provinces against foreign domination. Nevertheless, political loyalty was overwhelmingly to existing states with their republican (city-state), confessional, and dynastic legitimations.

Dynastic empires ruled over different ethnic groups. Henry VIII described his kingdom as an 'empire,' ruling Wales and Ireland, and claiming parts of France. The Habsburg dynasty ruled Austria, much of south-eastern Europe, Spain, Belgium, the Netherlands, and South America. The Valois and Bourbon dynasties claimed territories beyond those of contemporary France, itself inhabited by different language groups. By the seventeenth century the Romanovs were extending their rule beyond the core of Russia, west and south into Slav regions, north and east into Asiatic ones. Opposition came not from 'Slavs' or 'Asians,' but from other dynasties, e.g., the elective Kings of Poland (Saxons). The core of the Hohenzollern dynasty, Brandenburg-Prussia, included Polish speakers. The Ottoman Empire ruled Slav populations, Muslim and Christian.

The national idea was unimportant. Historians write of long-term 'winners' emerging as the core of later nation-states (England, France, Spain, Prussia, Russia, and Piedmont). However, these had no stronger, more consciously 'national' characteristics than long-term 'losers' like Poland-Lithuania, Burgundy, Britanny, Habsburg Austria, the Ottoman Empire, city-states, and small princedoms. Beyond Europe there was little one could call a national state. (Possible exceptions are Ethiopia, and Japan.) Often there was no distinct state but systems of lineal rule described as stateless or tribal. Dynastic empires ruled over many ethnic groups, e.g., China and India.

Only after 1500 does the national idea start to legitimize states and political movements. In England the construction of a common legal and administrative system from the late Anglo-Saxon period, the growing importance of Parliament, island situation, increased emphasis upon the vernacular language, and the national emphasis of the Protestant Reformation made the national idea salient. Shakespeare in *Henry V* (1599) delineates Irish, Welsh, and Scottish stereotypes, though all support the king. (Shakespeare only constructed English stereotypes at an elite level in contrast to other Europeans, e.g., the suitors for Portia's hand in *The Merchant of Venice*.) The enemy was France, but hostile stereotypes were confined to the elite, not the people. The play includes Henry wooing a French princess; his marriage and military campaigns are designed to maintain power in France. Ethnic and national ideas are prominent but subordinated to the monarchical principle. If England was the most 'advanced' case, this suggests the marginality of the idea elsewhere in 1600.

Four processes increased the importance of national arguments: confessional conflict, absolutism, new political opposition, and intensification of interstate

conflict. All developed alongside an extension of European influence beyond Europe and the growth of a commercial economy which undermined privilege and brought into politics 'middling sorts' of people—financiers, lawyers, merchants, commercial farmers, and manufacturers.

The Reformation challenged the dominance of Papacy, Catholic clergy, and the use of Latin. Translations and mass circulation of the Bible into vernacular languages, because individuals must take responsibility for their souls rather than leaving it to priests, gave a national flavor to Protestantism. Luther challenged the alien Papacy as a German, and Henry VIII as an Englishman. In Catholic states the prince took control over the church; in Spain and France this led to a 'national' church. Popular religious movements share affinities with later nationalist movements. Nevertheless, religion mattered more than nationality, and only when combined, usually under princely control, did national language figure prominently in politics.

Princes extended their powers by sale of office, appointment of extraordinary commissioners, legal codification, subordination of church, greater taxation, and the establishment of standing armies; processes summarized as 'absolutism.' There was nothing intrinsically 'national' in this, but where the core territory could be regarded as national, e.g., Spain, England, and France, this was justified by national arguments.

This affected political opposition. By 1620 the English Parliament presented itself as a defender of national freedoms. Parlements and provincial estates in eighteenth century France depicted the crown as despotic and themselves as guardians of national liberties. These arguments could escape beyond elite control. In England in the 1640s, national claims were extended to the 'poorest he' by radicals in the Putney debates. In France, radical values were projected upon the Estates-General summoned in 1789, soon converted into the 'National Assembly.' Monarchy was redefined as at best the servant of the nation, at worst its enemy.

Conflict between states had national consequences. Wars in late seventeenth- and eighteenth- century Europe were 'cabinet wars'; the War of Spanish Succession, Austrian Succession, etc. However, they had a global dimension. The central struggle was between France and Britain for world hegemony, using other states as proxies. The struggle extended to India and North America. The wars mobilized the expanding resources of commercializing, imperial economies. They engaged the attentions of an ever-larger part of the population. Commercial growth was associated with the rise of a print culture. In England the electorate increased in size. In France the crown and its opponents resorted to pamphleteering, appealing to 'public opinion,' and presenting political conflicts in national terms.

These national arguments accepted the territory of the state but questioned its legitimacy.

> The Nation is essentially the source of all sovereignty; nor can any individual or any body of men, be entitled to any authority which is not expressly derived from it.' (Article III of the French Declaration of the Rights of Man and Citizen, 1789)

The nation was the body of citizens within the existing state. For radicals this meant all adult males. If ideology was logical this would mean that the national idea was purely political. However, cultural elements were soon introduced.

The idea of a state occupying a national territory associates national virtues with the state. As the French revolutionary state plunged into crisis, it took up these associations. The ethnic and linguistic characteristics of the regions which rebelled against the Jacobins, let alone of opposing states, were linked to counter-revolution:

> We have observed that the dialects called Bas-Breton, the Basque dialect, and the German and Italian languages have perpetuated the reign of fanaticism and superstition, revered the domination of priests and aristocrats, and favoured the enemies of France. (Speech in the National Assembly by Barère, quoted in Carlton J. Hayes, *The Historical Evolution of Nationalism* (New York, 1931, p. 65)).

Opponents of the Jacobins presented the 'true' France as Catholic and monarchical in which the historic provinces and the estates enjoyed autonomies and privileges. Edmund Burke depicted nations as carriers of traditions which bound together generations and inhibited the arbitrary and destructive use of reason to alter institutions.

This greater use of national arguments was reinforced by the making of sharp distinctions between 'state' and 'society' and between state boundaries. Such changes became apparent during the Napoleonic wars. The French revolution abolished distinctions of privilege in France and exported this principle beyond France. This did not herald an era of freedom and equality. New governments were often more autocratic than their predecessors and political rights were restricted to a small minority. Nevertheless, equality before the law was introduced and a clear distinction was made between the state as public power and society as private. When governments appealed for support, they appealed to that 'whole' society rather than privileged minorities.

Opponents of Napoleon noted he could tap popular enthusiasm and linked this to new freedoms achieved in France. Reformers advocated the removal of privilege and the extension of political rights to strengthen support for the state. In politically fragmented regions, some invoked the 'nation' as a unifying and mobilizing principle. The nation was presented as a cultural entity from which should rise

nation-states, unlike France where the state created national unity. The idea was weak at first, though stronger in Germany than Italy, probably because in Germany there were national traditions such as the Holy Roman Empire and two powerful states—Prussia and Austria—which had an interest in appealing to the national idea.

Napoleon's defeat was inflicted by the conservative regimes he had not conquered—Russia and Britain—rather than by national movements in areas he had. Even there dynasties, state patriotism, hatred of the French, religious convictions, and opportunism better explain effective resistance. After 1815 there were attempts to restore privileges and old political arrangements. However, the map of Europe after 1815 was closer to that created by Napoleon than what had existed before. States were fewer and larger, more rationally and bureaucratically organized, with clearer boundaries. Governments could suppress national and liberal oppositions, but not the ideas those movements embodied. Commercialization, even without industrialization, increased social and territorial mobility which undermined attempts to restore privileges. Without those props, monarchies had to shift towards constitutional forms of rule. New politicians emerged who appealed to the people or the nation (Anderson 1974, Colley 1992, Dann and Dinwiddy 1988, Greenfeld 1992, Marcu 1976).

3. *Nation-state Formation in Europe*

Nation-state formation took place in three different ways: within preexisting territory (reform), by bringing together smaller states (unification), and by breaking away from a larger state (separation). These are artificial distinctions, but they draw attention to types of political movements and processes and provide a framework for analysis (Breuilly 1993).

3.1 *Reform*

France and Britain had a national character by 1800, although rebellions in Ireland and the Vendée indicated that not all subjects agreed. Napoleon introduced legal, administrative, and other reforms which over the nineteenth century overcame regional antagonisms, turning 'peasants' into Frenchmen. Vital was the development of mass education and politics. This took some time, marked by intense conflicts. Nevertheless, by 1900, although disagreements remained about the values to attach to the nation-state, all key political actors subscribed to the national idea and the priority of defending France against foreign threats.

In Britain, imperial growth and the stress on Protestant resistance to Catholic threats shifted the national idea from an English to a British level. The Reform Act of 1832 doubled the electorate, established the House of Commons as the national center of power, and entrenched the two-party system. Following further franchise extensions universal manhood suffrage was instituted in 1918, and universal adult suffrage in 1930. By 1870 elementary education was made compulsory and in the last decades of the nineteenth century there emerged a mass print media. Peasant and Catholic Ireland responded with separatist movements, but in the rest of the kingdom the effect was to integrate politics and sentiments on national lines.

Similar processes can be observed in other 'state-nations' such as Spain, Belgium, and the Netherlands (after the separation of the two territories in 1830), and the Scandanavian states. However, sometimes this was accompanied by regional resistance which took on a national form, e.g., Catalonia and the Basque regions of Spain (Hobsbawm 1987, Pugh 1997, Weber 1979).

3.2 *Unification*

German and Italian unification was marked by four features: a dynasty opposed to unification; a state promoting unification; a zone of small states; neutrality or support from major powers. The growth of constitutional and popular politics meant political movements increasingly appealed to the nation. Such movements were more important in Germany with supra-state institutions (Confederation of 1814; Customs Union of 1834) and more advanced economic development. Even here, national movements lacked mass support or access to state power.

The opposing dynasty was the Habsburg Empire. The revolutions of 1848 made that opposition very clear and revealed the weakness of popular nationalism. Elite nationalists turned to Prussia and Piedmont. The support of Louis Napoleon, British neutrality, and the weakening of Austrian and Russian capacity to resist change after the Crimean War—all created the conditions for effective action against Austria. Cavour, Prime Minister of Piedmon, allied with France against Austria in 1859. The decrepit Kingdom of the Two Sicilies collapsed in the face of invasion by Garibaldi, extending unity to the south. The new state had little elite unity or popular support. As D'Azeglio put it: 'We have made Italy, now we must make Italians.'

Prussia became capable of acting against Austria after the 1859 war, Austrian crisis, and military modernization in Prussia. Yet Prussia as a conservative dynasty was reluctant to take up the national idea with its radical implications. However, Bismarck, Prime Minister of Prussia from 1862, was an unconventional conservative who believed the national idea could serve the Hohenzollerns. In alliance with Austria he exploited the mistakes of Danish nationalism to defeat Denmark in war in 1864 and bring

Schleswig-Holstein under direct Austro-Prussian control. He maneuvered a war with Austria, helped by an alliance with Italy (engineered by Louis Napoleon) which was swiftly won. Prussian control was extended over northern and central Germany in alliance with liberal nationalists who helped to organize this political creation. French opposition to extension of control over south Germany was removed in the war against France of 1870 to 1871. Italy exploited the wars of 1866 and 1870 to 1871 to bring the Papal States and Rome into the nation-state.

The greater role of 'positive' factors of Prussian leadership and a national liberal movement compared to the 'negative' factors of state collapse and international support meant regionalism was more quickly overcome compared to the Italian case. Nevertheless, Germany in 1871 was reluctant to embrace the national idea, as evidenced by the absence of a national anthem, the refusal to adopt the national colors of black-red-gold, and the rapid move into opposition of popular parties representing workers and Catholics. Both Germany and Italy required the same processes of industrial and urban growth, mass education and politics, as noted for France and Britain, before a genuine nation-state could develop (Breuilly 1996, Hearder 1983).

3.3 Separation

Military defeats weakened Habsburg capacity to resist demands for internal change. In 1867 a new constitution gave autonomy to Hungary, converting the Magyars into a prop of the empire while depriving non-Magyars of support from Vienna. This intensified nationalist opposition, but Magyars maintained control, providing the outlet of individual assimilation into the dominant culture. In the western half of the empire a Czech national movement had compelled the dynasty by the 1880s to make concessions at the expense of Germans. Nationalist ambitions were checked by a combination of state power, reform, and the sense that full independence would create a small state at the mercy of powerful neighbors. Consequently, although national conflicts intensified in the Habsburg Empire, they did not directly threaten the empire before 1914.

By contrast, in the Ottoman Empire in Europe national opposition was less the product of conflicts between ethnic/cultural groups, and more the work of Christian elites seeking to replace waning Ottoman power. With assistance from European powers, motivated by self-interest and sentimental identification with Christian and national principles, this enabled the creation of separate states: Greece, Romania, Serbia, and Bulgaria. This stimulated a Turkish nationalist response which challenged the imperial Ottoman elite and in turn reinforced nationalist opposition, not only in Ottoman Europe but also amongst Arab subjects. Serbia supported national movements amongst 'its' national counterparts in the Habsburg Empire which in turn stimulated Vienna to crush Serbia.

In Russia, there was an intensifying interaction between 'official' Russian nationalism and non-Russian movements. These remained weak, especially after the most significant, that of the Poles, was repressed in 1863. In Germany there was Polish and Danish resistance to incorporation into a nationalizing state (Haddad and Ochsenwald 1977, Okey 2000).

3.4 Political Transformation

The trigger for transformation did not come from direct nationalist challenges to Habsburg and Romanov power, but from their responses to the new states in former Ottoman Europe. Austria sought to exploit the assassination of Archduke Ferdinand in June 1914 to bring Serbia to heel and Russia supported Serbia. Austria only acted as she did because of assurances of German support. Russia mobilized her alliance with France. Britain entered the war to prevent German domination in Europe. The success of nation-building, especially in Britain, France, and Germany, meant public opinion powerfully supported their nation-states.

Nationalist movements had little impact on the war. Rather, the defeat of Russia in 1917, and Germany, Austria, and Turkey in 1918 catapulted nationalism into prominence. The national idea was the legitimating principle of the peace settlement. Germany lost her Danish, Polish, and French acquisitions. The Ottoman and Habsburg empires were transformed into a series of nation-states in Europe and Allied mandates in the Middle East. Bolshevik Russia maintained control of non-Russian peoples, but incorporated the national principle into its ideology and institutions.

The new states were the product of great power diplomacy. They included national minorities. There were ethnic conflicts, even massacres. The result was a series of weak states which appealed to the core nationalism, persecuted national minorities, and whipped up discontent amongst their 'own' national minorities in other states.

These problems were compounded by global depression and the recovery of Germany and the USSR. The Allies could not longer guarantee the post-1918 order in east-central Europe. Between 1936 and 1941 Germany and the USSR destroyed that order. With the German invasion of the USSR, the way was opened to the fantastic project of racial empire in Europe.

After Allied victory in 1945 the national principle was too entrenched to permit any alternative to the nation-state, although the principle was not applied to Germany, which was partitioned and occupied. However, states were grouped into supranational blocs under US and USSR leadership. Many observers thought the era of nationalism (especially when

equated with fascism) and sovereign nation-states had ended in Europe, even if nation-state and national sentiment were firmly established (Brubacker 1996, Dunn and Fraser 1996, Payne 1996).

4. Nation-state Formation Beyond Europe

Nationalism and nation-state formation beyond Europe were responses to Europe. The earliest movements by settlers in the Americas utilized eighteenth century appeals to political liberty and self-determination, but not the nationalist idea of cultural uniqueness. Americans did not yet see themselves as culturally distinct from Europeans. The formation of distinctive national cultures was here, as in Europe, a nineteenth century development associated with commercial, industrial, and urban growth. Similar patterns, though not accompanied by political revolt against the imperial state, mark other states dominated by large-scale white settlement. In Australia, New Zealand, and Canada this led to the peaceful achievement of Dominion status within the British Empire, later membership of the British Commonwealth and *de facto* independence, although disputes continue about remaining political ties as evidenced by the 1999 referendum in Australia on the British queen remaining head of state. In Canada and South Africa the story was complicated by conflict between two different streams of European settlement—French and English, and English and Dutch, respectively. In Canada this conflict remains politically important. In South Africa the conflict became violent at the start of the twentieth century, but after the formation of an autonomous South African state in 1910 was increasingly replaced by conflict along race lines. In the Americas, Australia, and New Zealand the precolonial populations were marginalized; in South Africa this was not possible.

Neither was it possible where European domination was less through settlement than government and trade. Officials, missionaries, and merchants outnumbered white farmers, entrepreneurs, and workers. Colonial states were established in India and Africa. In large parts of Asia, notably China, power was indirectly exercised through client rulers.

There is a continuum rather than a sharp distinction between 'direct' and 'indirect' rule, 'formal' and 'informal' empire. Even an elaborate imperial bureaucracy as in India depended upon informal agreements with indigenous groups. Areas of indirect rule were often marked by economic exploitation and cultural domination. Nevertheless, to understand the formation of nation-states the distinction is important. Arbitrary boundaries imposed by Europe largely persist in the twenty-first century.

This suggests that the nation-state was the inheritor of the colonial state. Political movements formed within the colonial state, whether to oppose or collaborate. Such movements, using modern methods of political rallies, electoral contests, and print media, broke from precolonial patterns, appealing to the 'nation,' as subjects of the colonial state and as bearers of a distinct culture to set against European claims to civilizational and racial superiority.

Formal empire was largely constructed between 1858 (India) and 1900. Territory was transferred from the defeated to the victors of the World War I. There were hardly any nationalist movements in these vast regions. Some opposition developed in the inter-war period. Just as World War I created the basis for independence claims in Europe, so too did World War II beyond Europe. Japanese success in Southeast Asia and China abruptly destroyed myths of white superiority although only to be replaced by a new and brutal imperial rule. Nationalist movements were divided as whether to ally with the old or the new imperial power.

World War II destroyed the empires of the defeated powers and fatally weakened those of the European victors. The USA and USSR had no overseas empire and promoted anti-imperialist values. Imperial rule was restored by European countries, but it was desperately weak. The UK had to quit India in 1947. Arab nationalism threatened British and French mandates in the Middle East. A uniquely complicating factor was the foundation of Israel in 1948. The crisis of 1956, when Israel, the UK, and France sought control of the Suez Canal, led to the withdrawal of US support for her western allies and defeat for the UK and France.

From 1956 European withdrawal from Africa and the Middle East was just a matter of time. Nationalism became a politics of inheritance. Where political structures had promoted indigenous elites using modern political methods, decolonization was negotiated. Where this was blocked by white settlers or overdependence on direct economic exploitation, decolonization took place in chaos and violence, often with fateful consequences for the independent state. Decolonization is not the same as removal of cultural or economic power. Indeed some regard decolonization as an imperialist trick. New forms of nationalism developed, stressing economic independence and cultural authenticity rather than 'mere' political self-determination.

In the Middle East, Southeast Asia, and China, the story is complicated by the emergence of the USA as an imperial power, even if indirectly exercising power. This was partly due to fear of communist expansion, partly support for client states (e.g., Israel), and partly to protect economic interests. Whatever the reasons it stimulated anti-American nationalism. This remained tied to existing state boundaries. Nationalists in Tunisia, Libya, Algeria, Egypt, Syria, and Iraq espoused Arab nationalism, but jealously maintained state independence. Vietnamese anti-imperialists turned to communism, but never ceased to distrust

China. This is not to argue that ancient national rivalries predate and trump other sentiments, but rather that the colonial state and its successor provided the framework around which nations formed.

Important regions not subject to formal empire were Japan and China. In Japan, a unique transition took place from feudal/decentralized to imperial/centralized state, similar to the European pattern and partly in reaction to Western threats. The assertion of autonomy was couched in national terms. Japanese modernization was sufficiently advanced to lead to successful war against Russia in 1904. This is perhaps a unique case of modern nation-state formation without European involvement, although Japan's post-1945 history has been shaped by defeat and American occupation.

The subjugation of China to indirect Western rule produced nationalist rejection of the corrupt empire and alien influences, leading to the formation of a republic in 1911. However, nationalists had little control over much of China which was run by local warlords and this inhibited appeals to the 'people.' That appeal came from communists, although the 'people' was constructed in class, not national, terms. The Japanese invasion of China tilted the balance of power from nationalists to communists, enabling the communists to appeal to anti-Japanese sentiment as well as class interest. Thus, Chinese communism had a nationalist element which in turn shaped the state it established (Anderson 1998, Brown 1985, Darwin 1988, Wilson 1977).

5. *Nations in a World of Nation-states*

After 1945 the nation-state appeared both less and more important than previously. More important as the accepted political unit and because forced population transfers, mass murder, and boundary changes since 1918 had created more ethnically homogenous populations; less important because of the discrediting of nationalism and the emergence of supranational blocs.

The result was 'banal nationalism.' National identity was presented as something 'natural,' separate from 'politics.' National rivalry was displaced form politics into cultural contests. How far such sentiments can be politically mobilized is questionable; after a few years of casualties in Vietnam it was difficult to sustain popular support for the US war effort. Nevertheless, sentiments of national identity are transmitted every day through mass media. The banality of the idea means no political force can disavow national values, but equally they have little specific political content.

There are two major exceptions to this. First, the emergence of regional national movements in Europe and Canada. Some have gained regional autonomy, often the goal rather than the dangers of independence, and something central governments felt able to concede, often as part of a more general decentralizing political process. It is difficult to generalize about these movements which do not appear as part of a general trend, but represent specific problems in individual nation-states.

Second, the collapse of the Soviet Union. Warsaw Pact states gained real independence, whether retaining communist rule or not. Non-Russian republics in the USSR declared independence. Regions within Russia asserted claims against Moscow. Russian nationalists resented the loss of imperial power and sought to retain and regain territory and offered help to 'Russians abroad.' Has the collapse of the USSR unleashed a new wave of nationalism?

Nationalism contributed little to the collapse of the USSR which was due to economic crisis and destabilizing political reforms. National movements were a response to that instability. The issue was less about national independence than the form of the post-Soviet state. Answers varied along a continuum from old-style communism through reform communism, takeover by popular leaders, dissenting intelligentsia or nationalists to disappearance (East Germany). Each can change, e.g., as electorates shift between market-oriented liberal reformers and protectionist, socialist restorers. The nation-state is the political unit within which these politics are practiced, including seeking closer association with Western Europe.

In non-Russian republics the first nationalist outbursts leading to independence subsequently subsided. In the Ukraine and Moldova ethnic nationalism gave way to a stress on the national character of the territory and all its inhabitants. Russian minority problems in states like Estonia became matters of negotiation in ways pessimists had not thought possible. In Russia major conflicts seem likely to revolve around economic, rather than nationalist issues.

That does not mean nationalism is unimportant. There are non-Russian areas where military-based movements claim autonomy by violent means. Russian nationalism has been inflamed by Western intervention. Festering disputes remain between Russia and former republics. Muslim sentiments can fuel separatist movements, especially in mountainous and inaccessible areas. Nevertheless, in much of the region, political geography seems fairly settled.

Just as one can exaggerate nationalist conflict in Eastern Europe, so one can underestimate it in Western Europe. Many features of the 'classic' nation-state are being superseded. European Union (EU) nationals are semi-citizens in other EU states. Settled immigrant workers possess legal and social rights. National governments shift from policies of cultural assimilation or exclusion to those of multiculturalism. Governments cede powers to EU institutions. Nevertheless, the most important EU institution is the Council of Ministers, i.e., delegates of national governments. While most people remain largely confined to their national language and territory, they will regard

themselves 'naturally' as national, something routinely reinforced through mass media. Nationalism as the basis for the organized pursuit and use of state power may have passed its peak. Nations as communities and nation-states as their protectors are still regarded as natural and remain of central importance in the modern world (Billig 1995, Christian 1997, Kupchan 1995).

See also: Age, Sociology of; European Integration; Nationalism: General; Nation-states, Nationalism, and Gender; State and Local Taxation; State: Anthropological Aspects; State Formation; State, Sociology of the; States and Civilizations, Archaeology of

Bibliography

Anderson B 1998 *The Spectre of Comparisons: Nationalism, Southeast Asia and the World*. Verso, London and New York

Anderson P 1974 *Lineages of the Absolute State*. New Left Books, London

Armstrong J A 1982 *Nations before Nationalism*. University of North Carolina Press, Chapel Hill, NC

Balakrishnan G (ed.) 1996 *Mapping the Nation*. Verso, London and New York

Billig M 1995 *Banal Nationalism*. Sage, London/Thousand Oaks/New Delhi

Breuilly J 1993 *Nationalism and the State*. Manchester University Press and Chicago University Press, Manchester and Chicago

Breuilly J 1996 *The Formation of the First German Nation-state, 1800–1871*. Macmillan, London

Brown J 1985 *Modern India: The Origins of Asian Democracy*. Oxford University Press, Oxford, UK

Brubacker R 1996 *Nationalism Reframed: Nationhood and the National Question in the New Europe*. Cambridge University Press, Cambridge, UK

Christian D 1997 *Imperial and Soviet Russia: Power, Privilege, and the Challenge of Modernity*. Macmillan, London

Colley L 1992 *Britons: Forging the Nation 1707–1837*. Yale University Press, New Haven, CT and London

Dann O, Dinwiddy J 1988 *Nationalism in the Age of the French Revolution*. Hambledon, London

Darwin J 1988 *Britain and Decolonisation: The Retreat from Empire in the Post-war World*. Macmillan, London

Dunn S, Fraser T G (eds.) 1996 *Europe and Ethnicity: The First World War and Contemporary Ethnic Conflict*. Routledge, London

Greenfeld L 1992 *Nationalism: Five Roads to Modernity*. Harvard University Press, Cambridge, MA

Haddad W, Ochsenwald W (eds.) 1977 *Nationalism in a Non-national State: The dissolution of the Ottoman Empire*. Ohio State University Press, Columbus, OH

Hastings A 1997 *The Construction of Nationhood: Ethnicity, Religion, and Nationalism*. Cambridge University Press, Cambridge, UK

Hayes C J 1931 *The Historical Evolution of Nationalism*. Smith, New York

Hearder H 1983 *Italy in the Age of the Risorgimento, 1790–1870*. Longman, London

Hobsbawm E 1987 *The Age of Empire 1875–1914*. Weidenfeld and Nicolson, London

Kupchan C (ed.) 1995 *Nationalism and Nationalities in the New Europe*. Cornell University Press, Ithaca, NY and London

Mann M 1993 *The Sources of Social Power: Vol. 2: The Rise of Classes and Nation-states, 1760–1914*. Cambridge University Press, Cambridge, UK

Marcu E D 1976 *Sixteenth Century Nationalism*. Abaris Books, New York

Okey R 2000 *The Hapsburg Monarchy 1780–1918: The Improbable Empire*. Macmillan, London

Payne S 1996 *A History of Fascism, 1914*. University of Wisconsin Press, Wisconsin

Pugh M 1994 *State and Society: British Political and Social History 1870–1992*. Edward Arnold, London

Pugh M 1997 *State and Society: British Political and Social History 1870–1992*.

Seton-Watson H 1997 *Nations and States: An Inquiry into the Origins of Nations and the Politics of Nationalism*. Methuen, London

Smith A D 1986 *The Ethnic Origins of Nations*. Blackwell, Oxford, UK

Smith A D 1998 *Nationalism and Modernism*. Routledge, London

Weber E 1979 *Peasants into Frenchmen: The Modernization of Rural France 1870–1914*. Chatto & Windus, London

Wilson H S 1977 *The Imperial Experience in Sub-Saharan Africa*. University of Minnesota Press, Minneapolis, MN

J. Breuilly

Nation-states, Nationalism, and Gender

1. The State

The state is a form of centralized sociopolitical organization with a territorial base marked by institutionalized inequalities in gender and in access to wealth, power, and prestige among social groups. According to anthropologists, the state first emerged in Mesopotamia in the fourth millennium BC. Though numerous definitions of the state exist, most scholars would agree that states have (a) a ruling elite; (b) a bureaucracy which carries out administrative tasks; (c) a fiscal system which oversees production, distribution, and consumption and collects taxes; (d) a legal system; (e) a monopoly of the legitimate uses of force; (f) an armed force which keeps order, defends against external encroachment, and protects the wealth, power, and privilege of the elite; (g) a set of dominant cultural norms, values, beliefs, symbols, and institutions which promote the legitimacy of state power and of institutionalized inequalities while seeming to represent the interests of all social groups. Modern states are nation states.

2. The Nation and the Ethnic Group

Nations emerged in the eighteenth century. Anderson (1991) in a widely influential work defines the nation as a cultural construction, as the dominant form of

imagining political community in modern states. Constructed as limited in its membership and sovereign over its territory, the nation is imagined as a political community characterized by a deep, horizontal comradeship. The nation largely has been theorized as part of the public political sphere, in which women, from the point of view of nonfeminist scholars, have had only a marginal presence; hence, most influential male theorists of nationalism have ignored gender.

The notion of the nation is not to be conflated with that of the ethnic group. Following Weber (1978), an ethnic/racial group is a sub- or transnational status group formed around a subjective belief in common descent and in similarities of a cultural and/or physical type. Most contemporary nation states are multiethnic/racial. Despite this, many nationalisms have stressed the ethnically/racially homogenous character of the imagined community, whether in terms of 'shared blood' or common linguistic and cultural heritage. This occurs when a particular ethnic/racial group tries to seize or hold on to power, wealth, and privilege by defining the political community in ethnically/racially exclusive terms. Historically, this has often been the case, leading many scholars to argue that there is a strong link between nationalism and ethnicism/racism.

The image of horizontal comradeship among nationals has been key to legitimizing the power, wealth, and privilege of elites and making the inequalities of state societies less visible. One of the salient contradictions of nation states is that they are hierarchical societies which are imagined as political communities of equals. This contradiction is culturally mediated by national symbols that fuse emotional, sensory, and cognitive meanings and hence are able to induce powerful ideas, attitudes, feelings, and dispositions to action in nationals. These include the willingness to kill and die (as well as to bear children) for the nation.

3. *Nations and Gender*

The most powerful national symbols draw on the domain of the family and on tropes of common descent such as shared blood to create the structure of feeling that is nationalism. Membership in nations is most commonly obtained through birth; thus nation states actively regulate sexuality, reproduction, gender, and kinship. The nation cannot be theorized simply in relation to the public sphere; it must also be understood in relation to the domestic sphere. This insight has been at the core of the new scholarship on gender and nationalism, which emerged in the 1980s and burgeoned after 1995. Most of this research has focused on women.

3.1 *Women as Biological and Cultural Reproducers of the Nation*

Nation states manage sexuality, birth rates, family forms, and property transmission in specifically gendered ways—through laws, through cultural norms and values, and even through violence. Women's generative capacities, rights, and roles are regulated by nation states so as to produce suitably socialized citizens and workers. State interventions can range from the promotion of heterosexuality and of specific family forms, to forced sterilization, to the restriction of contraception and the banning of abortion, to material and other incentives to multiply or to limit fertility. The most well known case of reproductive limitation is that of China's 'one child' family policy. In China, where sons are preferred to daughters, females are the main object of abortion and infanticide. Ironically, the effects of the one-child policy include trafficking in women as well as children. Since the mid-1980s, Chinese newspapers have been reporting that organized rings of peddlers are abducting women from the poorer interior provinces and selling them as brides in the richer coastal areas; children have also been kidnapped and sold. In Japan, by contrast, the government and private companies are offering financial and other incentives to women to bear children as the low birth rate creates a concern for the future of the nation's demographic profile and economy. In cases where ethnic/racial exclusivism defines the national project, different reproductive policies are applied to women from ethnic/racial majorities than to minorities. For example, government policy in Singapore during the 1980s encouraged upper- and middle-class professional women of the dominant Chinese ethnicity to marry and bear children as a patriotic duty while working-class women of minority Malay and Indian origin were offered cash awards to restrict their childbearing. In welfare states characterized by public patriarchy—a taking over by the state of some of the functions of domestic patriarchy—poor women, often from ethnic/racial minorities, are encouraged to limit childbearing; this has been the case in the US.

Biological and cultural reproduction cannot be separated. As the bearers of children, women have also been considered their main socializers, held responsible for the transmission of the 'mother tongue' as well as the continuity of nationalist values, symbols, and rituals. But the regulation of women's fertility and maternity by nation states is not just about the quantity and 'quality' of the population or the enculturation of children. It is also about the maintenance of internal stratification, including the class system. Stratification in the public sphere depends on stratification in the domestic sphere and on the creation of differences between and among men and women. Guatemala's class and race hierarchy, for instance, has been maintained through a system of

stratified reproduction in which elite women of European descent bear children in legal monogamous unions with men of their own class and race, lower class 'mixed' *Ladinas* have extralegal liaisons with men of higher status and class position, while Mayan women marry men within their own communities. Maintaining internal stratification by privileging whiteness and cultural ties to Europe and North America in a country where Maya are numerically the majority has made nation building in Guatemala a fraught and violent process. From the point of view of the elite, the Maya can only be integrated into the nation through de-Indianization; this has entailed a series of policies ranging from outright genocide to the militarization of the countryside to forced resettlement. Yet Mayan women endeavor to cope collectively with the challenges of reproducing their communities in circumstances in which fear, as Green (1999) puts it, has become 'a way of life.'

3.2 Women and Men as Symbols of National Boundaries and Identity

The unity of the political community has been represented in gendered forms. Collective territory, collective identity, and womanhood are often conflated in nationalist discourses. As place of origin and source of being, the nation is often represented as a woman: Mother Russia, Mother India, *La Patrie* in France. However, the nation as active subject is most often gendered as male. In Mexico, for example, *La Patria* is represented as a Mother but *El Pueblo* (The People), the collective protagonist of the national, is masculine as are the founders of the nation, its Fathers. In some cases, such as that of Australia, the feminine is marginal to the representation of the nation altogether; Australia is imagined as a political community of virile mates. Likewise, the Greek nation is represented as a group of brothers related through the male line. In the case of Singapore under Prime Minister Lee Kuan Yew, the state constructed the nation not as a fraternity but as a patriarchy; the articulation of the traditional family with the modern state there has rendered natural an omnipotent government, a State Fatherhood. It would be important to know what difference, if any, iconographies of the nation as a fraternity versus a patriarchy make to the status of women.

Tropes of gender are often central to the construction of legitimate forms of authority in nation states. The redefinition of authority in the public sphere that was key to the nationalist project of Mexico's nineteenth century liberals entailed a corresponding transformation of authority in the domestic sphere. The protection of women became linked to the liberal state's project to build a nation founded on the principles of reason and modernity as well as to the defense of the *Madre Patria* symbolized by the feminine. Wife beating became a sign of backwardness and a criminal offense. Yet though reforms in Mexican law accorded women certain protections they also restricted their freedom in numerous ways, particularly if they were married.

The symbol of nation-as-woman puts real women in a double bind as they are simultaneously sacralized and profaned, honored and dishonored, protected and restricted. In order to position themselves in the international field, for instance, postcolonial nations have negotiated the contradictions between tradition and modernity, between sovereignty and westernization, by making women the symbols of a purportedly pure national tradition. Anticolonial Indian nationalists, for example, identified Indian subjectivity with a spirituality, which they considered the guarantee of India's superiority to and independence from the West. As the symbols of this internal spirituality, women were to be protected from modernization so that they could transmit the nation's spiritual essence and cultural heritage; this had a negative impact on women's autonomy.

Women can also become the symbolic bearers of modernity. During the Turkish Revolution of 1917, the unveiling of women became an important symbol of Turkey as a modern nation state. The same effort marked the Shah of Iran's project to build a modern nation; the subsequent fundamentalist revolution, by contrast, insisted on the reveiling of women as a symbol of the return to tradition.

3.3 Gender, Nationalist Violence, and Warfare

Scholars have found that an intimate relationship exists between warfare and nation building with distinct consequences for men and women. The equation of the bearing of arms with masculinity and citizenship is one factor determining the less than full inclusion of women in many nations. Most nation states and nationalisms actively promote a different relationship to violence for men and women. If the motherland is feminine, the state and its warriors as active agents are masculine. Boys and men are socialized for warfare through sports and military training and cultural constructions root masculinity in the moral and physical capacity to exercise force and take life in the name of the nation as the state deems fit. Women, by contrast, are linked symbolically to peace, and are restricted in their participation in the military though they have played key noncombatant roles as nurses, support personnel, and conduits for intelligence and arms. The gender division of labor in the military tends to be more rigid than in the civil labor market. Even in nation states such as Israel and the US where women have a greater role in the military, they still do not have parity with men.

Many nationalist discourses have relied on the trope of sacrifice in order to secure men and women's gender specific cooperation in wars. Another key trope has

been that of rape—of the nation as a vulnerable maternal body in danger of being violated by foreign men; in this sense, men are called to sacrifice themselves in war in order to protect women and children.

The special burden of representation that women carry as symbols of the nation and bearers of national honor and identity can give women a limited protection from violence in some cases. For example, the Argentineian Madres de la Plaza de Mayo drew on the moral authority of the nation-as-mother to challenge the state's campaign of terror against its citizens. But this special burden can also make women the victims of particular forms of sexualized violence such as rape, as demonstrated in warfare after the partition of India, within Guatemala, during the Mexican Revolution, during World War II, and most recently in the Balkans. Women's bodies are often the terrains over which wars are fought—both symbolically and literally.

During periods of warfare, women and even men's generativity can come more closely under state control. The *Lebensborn* program in Germany, for example, encouraged SS men to father as many children as possible with Aryan women they were not expected to marry; the resulting progeny were raised by the state. At the same time, a program of involuntary sterilization prevented those considered unfit from reproducing. In Serbia during the 1990s, a coercive pronatalism was linked to explicitly ethnonationalist warfare, generating feminist protest.

3.4 Nation States, Women, and the Labor Market

Through legislation as well as cultural institutions, nation states have shaped women's involvement in the labor market. Policies restricting women's labor market participation and tacitly supporting differential compensation along gender lines have been enacted, at one time or another, in most nations. Despite the gains that women have made in industrial nations, their earnings and employment opportunities still lag behind those of men; in developing countries gender discrimination in the labor market is more pronounced. On a world scale, women constitute a disproportionate share of the poor. In particular, households without an adult male earner are poor, so the rise of female-headed households has thrown more women and children into poverty. The causes of female poverty are often seen in terms of women's individual characteristics. This fails to recognize that the rise of single motherhood, whatever its causes, has allowed men, collectively, to benefit from women's reproductive labor without sharing in the costs of raising children. It also fails to recognize the role of labor market discrimination in women's low earnings.

Nation states have not just limited women's labor force participation; they have also encouraged it, albeit in highly restricted ways. This has often been the case in times of war as epitomized by the image of Rosie the Riveter during World War II in the United States. More recently, nation states have been promoting women's employment in the global apparel, footwear, and electronics industries as key to national development. For example, in South Korea under President Kim, young, single women were encouraged to be 'dutiful daughters' by becoming part of the manufacturing labor force and accepting low wages and poor working conditions in the name of national development. In Mexico's export processing zones, the hiring of women for work in apparel and electronics assembly plants is fueled by gender stereotypes which proclaim women to be more docile and to have more nimble fingers. In both these cases, nationalist rhetoric has promoted an image of working women allied with modernity that confers agency in name only.

3.5 Feminism and Nationalist Movements

Anticolonial and other nationalist movements have often construed subordination as emasculation; hence the redefinition of gender roles has been seen as central to the rebuilding of society. A debate exists in the literature as to whether nationalist movements have been supportive of or inimical to women's interests. Following Jayawardena (1986) some argue that at the end of the nineteenth and beginning of the twentieth century, feminist goals were compatible with those of anti-imperialist liberation movements in Asia. This position also highlights the fact that there have been many feminisms, including those without a Western origin. By contrast others argue that fundamentalist and ethnically exclusionary forms of nationalism are reinscribing new forms of patriarchy under the banner of tradition. A particularly clear instance of the range of positions is provided by the case of postcolonial Indian nationalism: while there are those who maintain that Gandhi's concern with remasculinizing men led to a new subordination of women, others insist that Gandhi took over the women's movement in India but gave it a new impetus by celebrating women's strength. Work on the Nation of Islam in the US indicates that the emancipatory language of racial redemption draws on images of gender which construct an insurgent black masculinity at the expense of black women who are to give up action in the public sphere in order to become men's help mates in the domestic sphere. This implies a constriction not an expansion of women's citizenship.

3.6 Women as Citizens

Women have had a contradictory position in most nation states. On the one hand, they are symbols of the nation. On the other, they are often excluded or

marginalized from the collective we of the body politic and endowed with less than full citizenship rights. While some scholars have argued that women were simply latecomers to citizenship rights, others have insisted that their exclusion was integral to a construction of state citizenship as 'the rights of Man.'

Recent scholarship emphasizes that a nonsexist notion of citizenship can only be developed if citizenship is defined in relation to both the public and domestic spheres and considered to include not only political but also social and civic rights and duties. The need to move beyond the nation state as the unit of analysis and to examine the ways the global political economy is creating new contradictions has become more pressing; social rights which have been integral to national welfare states have come under threat in ways that disproportionately affect women and put into question the viability of nation based identities.

See also: Chinese Revolutions: Twentieth Century; Colonialism, Anthropology of; Cuban Revolution, The; Development and the State; East Asian Studies: Gender; Economic Development and Women; Economic Globalization and Gender; Female Genital Mutilation; Gender and Feminist Studies in Anthropology; Gender and Feminist Studies in Geography; Gender History; Gender Ideology: Cross-cultural Aspects; Labor Markets, Labor Movements, and Gender in Developing Nations; Latin American Studies: Gender; Lone Mothers in Nations of the South; Mexican Revolution, The; National Liberation Movements; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: East Asia; Nationalism, Historical Aspects of: The West; Nations and Nation-states in History; Near Middle East/North African Studies: Gender; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender; Southeast Asian Studies: Gender; Vietnamese Revolution, The; Women's Suffrage

Bibliography

Alonso A 1995 *Thread of Blood: Colonialism, Revolution and Gender on Mexico's Northern Frontier*. University of Arizona Press, Tucson, AZ

Anderson B 1991 *Imagined Communities*, rev. edn. Verso, London

Balakrishnan G, Anderson B (eds.) 1994 *Mapping the Nation*. Verso, New York

Ginsburg F D, Rapp R 1995 *Conceiving the New World Order: The Global Politics of Reproduction*. University of California Press, Berkeley, CA

Green L 1999 *Fear as a Way of Life: Mayan Widows in Rural Guatemala*. Columbia University Press, New York

Jayawardena K 1986 *Feminism and Nationalism in the Third World*. Zed, London

Kim S 1997 *Class Struggle or Family Struggle? The Lives of Women Factory Workers in South Korea*. Cambridge University Press, Cambridge, UK

Parker A, Russo M, Sommer D, Yaeger P (eds.) 1992 *Nationalisms and Sexualities*. Routledge, New York

Peterson V S 1995 The politics of identity and gendered nationalism. In: Neack L et al. (eds.) *Foreign Policy Analysis: Continuity and Change in Its Second Generation*. Prentice-Hall, Englewood Cliffs, NJ

Silverblatt I 1988 Women in states. *Annual Review of Anthropology* **17**: 427–60

Service E 1975 *Origins of the State and Civilization*. Norton, New York

Weber M 1978 *Economy and Society*, 1st edn. University of California Press, Berkeley, CA

Williams B (ed.) 1996 *Women Out of Place: The Gender of Agency and the Race of Nationality*. Routledge, New York

Yuval-Davis N 1997 *Gender and Nation*. Sage, London

Yuval-Davis N, Anthias F (eds.) 1989 *Woman–Nation–State*. Macmillan, Basingstoke, UK

A. M. Alonso

Natural Concepts, Psychology of

Concepts are crucial for intelligent thought and action. They allow people to treat a set of objects as equivalent and apply their prior knowledge about this set of objects to any new instance. A *concept* is a mental representation of a class of entities (such as robins or actors). In this article, the focus is on *natural concepts*, mental representations of those classes that occur in nature (such as robins, maples, or sapphires).

1. Concepts and Their Functions

A concept is a mental representation, and *category* refers to the set of items selected by a concept. The concept allows people to tell if a new item is a member of the category (*classification*). Thus, one can have the concept 'dog' and use that concept to know that a particular object belongs to the category of dogs.

Classification is an important function of concepts, but it is not the only one. Concepts are the representations that people use to think. They are necessary for many cognitive activities, such as for predictions, comprehension, reasoning, communication, and building new, more complex, concepts. In fact, it is these functions that make concepts so essential. For example, the concept 'dog' allows one to classify a new item as a dog. However, classifying something as a dog is not that useful unless one has lots of other knowledge about dogs that can be applied. If that knowledge allowed a prediction of whether the dog might bite, this would be important for knowing how to act. Similarly, one's knowledge about dogs may allow one to understand why a dog chased a toy or to

reason about the relation of dogs to other mammals. In addition, because people have much common knowledge and often refer to concepts by the same words, people can learn from others' experiences. Finally, having a mental representation of dog means that one can combine it with other concepts to get new concepts, in what is called *conceptual combinations*, such as 'couch dog.'

2. Conceptual Structure

A central question about natural concepts is their *conceptual structure*, how the information about each concept is organized and used for its various functions. To begin, what knowledge permits the classification of a new item?

2.1 The Classical View

The earliest answer, going back to at least Aristotle, is that conceptual knowledge includes a definition to identify new members. This classical view asserts that there is a set of simple features that define each concept—if an object has all of these features then it is a member of the category, otherwise it is not. For example, a square is a two-dimensional, four-sided, closed figure with equal angles and equal sides. All objects that have those features are squares, and any object that is missing one or more of these features is not a square. There are two main types of problems with this classical view as a theory of people's conceptual structure. First, it fails to account for the fact that some members of a category are viewed as better, or more typical members, than other members of the category. Thus, 'robin' is a typical bird, whereas 'ostrich' is not. These effects are seen in both time to verify category membership and in ratings of typicality. If both birds meet the definition, then something else must be determining this goodness-of-category membership. Second, although it is not difficult to figure out the definition for concepts such as square, it is much more difficult to understand what the defining features are that allow us to classify an item as a member of a natural category such as birds. Clearly, many features are important to being a bird, but no-one has been able to characterize a set of observable features that would allow accurate classification of all birds without including some non-birds. Even bird-crucial features such as feathers are problematic, since a bird that lost its feathers, perhaps owing to disease, would still be a bird.

2.2 Prototypes and Family Resemblance

Natural concepts often have highly inter-related feature representations, even if there is no single set of features possessed by all members of the category. For example, birds have feathers and lay eggs, but in addition they have wings, beaks, tend to fly, and so on. Thus, the members of the category tend to have many similar features, even if none of them perfectly separates every bird from every nonbird. This idea led to the suggestion that the conceptual structure might consist of a summary representation, or *prototype*, that captures the most commonly occurring feature values for the category members (Posner and Keele 1970, Rosch 1978). To classify a new instance, one compares it with the prototypes and chooses the most similar prototype, with more important features given more weight. There are no necessary or sufficient features for membership, but prototypical features each add evidence that it is a member of that category. Thus, although an ostrich does not fly or live in trees, it can be classified as a bird because it has many of the other features in the bird prototype (such as has feathers, has wings, lays eggs, etc.).

This view addresses both the problems mentioned with the classical view. First, typicality is directly accounted for—more typical members have more of the prototypical features. For example, a robin might have all of the features, whereas an ostrich has only some. Second, there are no defining features, but rather a set of typical features that tend to be true of birds.

How are the members of the category related to the prototype? Although few members may have all the prototype features, all members will have some of the features. This type of structure is called *family resemblance* because it is similar to the feature distribution one would see with an extended family (Rosch and Mervis 1975; see Wittgenstein 1953 for an earlier treatment in philosophy). There will be some family features that occur often across the whole family (e.g., tall, red hair, blue eyes, small nose), but some family members will have more of them than others.

The prototype view has much intuitive appeal and accounts for many of the empirical findings. However, there are a set of problems that all relate to the idea that the prototype view is too simple—it does not take into account much of the other knowledge people have about the concepts, the category members, and the relation of this concept to other concepts. People often know not just the prototypical features, but the range of common feature values and whether there is a correlation among various features (such as size of birds and whether they sing). See Smith and Medin (1981) for a thorough examination of the classical and prototype views.

2.3 The Theory-based View

The *theory-based view* argues that concepts must be integrated with other knowledge. First, the representation of concepts is not a simple list of prototypical

features but includes the relations among features (such as correlations) and explanations as to how the features are related (for example, how wings, feathers, and hollow bones enable flying, or how flying enables having nests in trees). Second, the concepts are explicitly related to other concepts, because part of our understanding of any concept relies upon its relation to much else that we know. An understanding of a natural concept, such as 'dog,' requires one to relate it to what we know about living organisms, pets, digestive systems, and so on. The point of the theory-based view is that our conceptual knowledge requires embedding in these complex knowledge relations so that we can quickly and reliably use our knowledge for classification and other category functions; Medin (1989) explains this view in more detail and points out problems with other views that are dealt with by the theory-based idea.

Most importantly, this view helps understand a central issue in natural concepts, *conceptual coherence*—what makes categories members of a concept all go together? Instead of relying on similarity to the prototype, the theory-based view argues that there is a set of underlying (perhaps unobservable) features and explanations that lead to items being members of the same category. Thus, although birds have some similarities to one another, the real reason 'birds' is a coherent concept is because people believe birds to have some basic underlying commonality (perhaps genetic structure). The importance of this distinction is illustrated by porpoises—although they look like many fish, they are classified as mammals.

Despite the advantages of the theory-based view in bringing to bear much of our other knowledge, there is concern about the complexity of the view and that it has not been fully described yet. Many researchers believe that this complexity is necessary for integrating conceptual structure work with other work on concepts (some of which is described below), but also believe that the prototype view is a good, if simplified, description for many cases. There is also some question as to whether even all natural concepts have some underlying commonality that is necessary and sufficient for category membership (Malt 1994).

3. *Organization Among Concepts*

The focus has been on the structure of each concept, but the relations among concepts is an essential part of our knowledge. An important aspect of the relations among concepts also is the level of abstraction at which a particular item may be classified. For example, when people see an item, they are likely to classify it as an 'apple' not as a 'fruit' or a 'Granny Smith' apple. Rosch and her colleagues showed that there was one level, the *basic level*, which is psychologically privileged. See Lassaline et al. (1992) for a review.

The evidence for this basic level is very strong. Not only do people choose to name objects at this level, they are faster to verify an item is at this level (an apple) than at either the more abstract (fruit) or more specific (Granny Smith apple) level. In addition, people often learn these concepts earlier and a variety of measures show it is the most abstract level at which objects share many common features.

The interpretation for the basic level is that it represents a compromise between two opposing goals of the categorizer: the goal to have the most information possible and the goal to have the categories be as discriminable (distinctive) as possible. The informativeness leads to more specific levels, but the distinctiveness favors more abstract levels. The basic level is a compromise—getting more specific (such as Granny Smith apple) does not add much information but does greatly lower the distinctiveness (since it must be distinguished from other apples). Getting more abstract (fruit) greatly reduces the informativeness.

Are there exceptions to the basic level? Yes, and they help to clarify how such a level may come about. Very atypical members may be classified more quickly at a more specific level. For example, a sports car is classified more quickly as a sports car than as a car. The claim is that the atypical items have many features that distinguish them from other members of the basic-level category (such as other cars), so the most informative and distinctive level may be at a more specific level. For natural concepts, think of an atypical bird such as penguin or ostrich. Second, the basic level seems to change with expertise. Dog trainers can classify the breeds they work with as quickly at the more specific level, such as collie, as at the basic level of dog. Again, distinctiveness appears to be an important part of the explanation—with extended experience, experts learn additional features distinguishing members of the basic level.

4. *Uses of Concepts Beyond Classification*

Much of the research on natural concepts has focused on classification, which is why it has been the main focus of this article. However, as mentioned earlier, the importance of concepts is that they play a role in many cognitive activities. When people classify an object, they can access knowledge about the concept to help in various types of reasoning and understanding. Much of this work is reviewed by Medin and Atran (1999).

4.1 *Inferences from a Concept*

Even young children will use category membership to make inferences about a new item. Gelman and Markman (1986) showed children pictures of two animals (such as a flamingo and a bat) and taught the children novel properties about each animal. For example, they might learn that a flamingo feeds its baby mashed-up food and that a bat feeds its baby

milk. They then were shown a picture of a blackbird that was perceptually similar to the bat, but was labeled as a bird. The children thought the blackbird would feed its baby mashed-up food, just like the flamingo. Categories provide a basis for inference.

Not only are categories important for inference when people know the category, but they are important for inference even when there is uncertainty about the category. Suppose you are in the woods and see an animal, which you think is probably a skunk, but could also be a small dog, squirrel, or raccoon. What should you infer about the probability it will bother you? One possibility is that you could estimate the probabilities that each of those animals would bother you and weight those estimates by the likelihood of each animal to obtain an overall probability. The second possibility is that you assume it is the most likely type of animal, in this case a skunk, and make your estimate based only on that category. Although the first way seems more rational, it comes at the cost of much increased effort, even in this simple case. In fact, under many circumstances, people choose the second approach, just relying on the most likely category.

This research shows the importance of conceptual knowledge in making inferences both when the category membership of an object is known and when it is uncertain.

4.2 Inference from Relations Among Concepts

Concepts are not represented separately from our other knowledge. As mentioned in Sect. 3, an important relation among concepts is the level of abstractness. The previous section examined the inferences made from a single category, whereas this section examines how inferences might be made from one category to other categories. For example, suppose you are told that a new Property X is true of robins and of bluejays. Do you think it would also be more likely to be true of sparrows or of geese? Most people answer sparrows—in general the more similar the first categories are to the last category, the stronger the inference. In addition to similarity, other factors affecting the strength of the inference are typicality and diversity. The more typical the first category, the stronger is the inference to the more abstract category—so if you know Property X is true of robins, you think it is likely to be true of all birds, compared with when you know the property is true of ostriches. Diversity refers to how varied the first categories are, or how much of the more abstract category they cover. For example, wolf and elephant are very different so if you know Property X is true of both of them, you are likely to believe it is true of all mammals. However, if you know Property X is true of two similar mammals, such as wolf and dog, then you are less likely to believe it is true of all mammals.

People use their knowledge of these hierarchies to make a variety of inferences. Some cross-cultural work points out similarities and differences between cultures that vary in their knowledge of biological categories. In particular, a Mayan culture in which the people spend much time with local mammal species shows much greater differentiation of the categories and also uses more knowledge of the ecology in making their inferences compared with American college students.

4.3 Conceptual Representations Are Affected by Use

Conceptual representations have large effects on a variety of conceptual uses, such as inferences, understanding, etc. In addition, it is important to remember that these conceptual representations are constructed partly in response to how they are used. This point is easiest to see in some work examining the organization of complex domains. When different types of tree experts sorted a large number of tree names, landscapers' sortings were greatly influenced by landscaping properties of trees, such as their shadiness. Ross and Murphy (1999) had people sort a large number of food names. Although the dominant influence was standard food categories such as breads, fruits, and meats, there was a strong influence on food categories that arise from human activity, such as breakfast foods, snacks, and desserts. Hence one should not think of conceptual representation as mirroring the outside world. Our representations depend on the world and on what we do.

See also: Concept Learning and Representation: Models; Conceptual Change, Cognitive Psychology of; Connectionist Models of Concept Learning; Early Concept Learning in Children; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Knowledge Representation; Mental Representations, Psychology of; Propositional Representations in Psychology; Scientific Concepts: Development in Children

Bibliography

Gelman S A, Markman E M 1986 Categories and induction in children. *Cognition* **23**: 183–208

Lassaline M E, Wisniewski E J, Medin D L 1992 Basic levels in artificial and natural categories: Are all basic levels created equal? In: Burns B (ed.) *Percepts, Concepts, and Categories*. Elsevier, Amsterdam, pp. 327–78

Malt B C 1994 Water is not H_2O. *Cognitive Psychology* **27**: 41–70

Medin D L 1989 Concepts and conceptual structure. *American Psychologist* **44**: 1469–81

Medin D L, Atran S (eds.) 1999 *Folkbiology*. MIT Press, Boston

Posner M I, Keele S W 1970 Retention of abstract ideas. *Journal of Experimental Psychology* **83**: 304–8

Rosch E 1978 Principles of categorization. In: Rosch E, Lloyd B (eds.) *Cognition and Categorization*. Erlbaum, Hillsdale, NJ, pp. 27–48
Rosch E, Mervis C B 1975 Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* **7**: 573–605
Ross B H, Murphy G L 1999 Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology* **38**: 495–553
Smith E E, Medin D L 1981 *Categories and Concepts*. Harvard University Press, Cambridge, MA
Wittgenstein L 1953 *Philosophical Investigations*. Macmillan, New York

B. H. Ross

Natural Environmental Psychology

Natural environmental psychology examines the influences of natural forces on individuals. Thus, it includes examination of extraterrestrial influences, weather, nature, and natural disasters on the behavior, thoughts, stress, and well-being of persons.

1. *Influences from Beyond the Atmosphere*

Contemplation of the vastness of space, the stars and planets produces awe. The Earth is a spin-off of some ancient cosmic event, and humans therefore are composed of highly evolved star dust. In that sense, deep space has everything to do with human existence. The end of the planet, and human life, may result from some gigantic celestial collision. Beyond these titanic effects, does deep space have any continuing psychological influence in everyday life?

1.1 *Cosmic Rays*

Cosmic ray activity has been mildly associated with psychiatric hospital admissions and behavioral disturbances in schizophrenics. In one study, hospital admissions increased with cosmic ray activity; behavioral disturbances increased for some individuals with schizophrenia but decreased for others. Possibly, although more research certainly is necessary, the flow of neutrons from cosmic rays affects schizophrenia.

1.2 *The Stars and Planets*

The best-known putative influence of the stars and planets on human psychology is that posited by astrology. Many people around the world treat astrology merely as a sort of entertainment, but a substantial minority, including some world leaders, grant it more serious predictive status. That in itself is psychologically interesting. However, there are virtually no published scientific studies of astrology's claims, so scientists cannot grant it any credence.

A single published study did examine the relation between the positions of planets at birth and subsequent personality. Among eminent French scientists, a significantly high proportion of introverts were born when Saturn had just risen or just passed overhead, and extroverts were more likely to be born when Mars or Jupiter has just risen or passed overhead.

Nevertheless, from a scientific perspective, one would have to propose a plausible mechanism by which the stars and planets affect personality, and this has not been done. Furthermore, neither this nor the cosmic ray study has been replicated, and may represent dramatic instances of chance findings. Perhaps many similar studies were done without the discovery of statistically significant findings, and therefore went unpublished.

2. *The Sun and the Moon*

2.1 *The Sun*

The sun ultimately powers almost everything on earth, either directly through solar power or indirectly through the use of stored solar power. Coal, gas, and oil are plant-stored solar power; hydroelectric power is based on the sun lifting water into the atmosphere so to produce rain; most food is produced through photosynthesis. In other very basic ways, of course, the sun influences everyday human behavior. Sunny weather encourages gardening, going to the beach, and playing summer sports; the lack of it encourages people to stay indoors or to engage in winter sports.

Sunshine may induce good moods, which may result in more generosity; tipping, in one study, increased when the weather was sunny. However, sunlight can improve the spirits too much for some people: the incidence of mania rises in the long days of summer.

2.2 *The Moon*

The belief that the moon influences strange behavior goes back at least to 400 BC. One study showed that 49 percent of undergraduates agree with the statement, 'Some people behave strangely when the moon is full,' and even higher percentages of police officers believe that the full moon affects people.

Numerous published studies report significant connections between phases of the moon and such behaviors as calls to counseling centers, admissions to psychiatric hospitals, and homicide. However, rig-

orous examination of the evidence suggests that, once again, such studies represent chance outcomes that are to be expected when many studies are conducted (Rotton and Kelly 1985).

At least the 'lunacy hypothesis' (as opposed to astrology) has been accompanied by several plausible causal mechanisms. One is that the moon's gravitational influence on bodily fluids causes 'biological tides' which result in odd behaviors. However, the moon's gravitational force on humans is more than 5,000 times weaker than that of the Earth; even a large building tugs on bodily fluids harder than the moon. The moon's gravitational effect is far too weak to affect bodily fluids. A subsequent analysis examined eight other possible causal mechanisms and concluded that none were credible.

Individuals may continue to believe in lunar effects for two reasons. First, the full moon's brighter light may keep light sleepers awake, leading to sleep deprivation, which in turn may produce unusual behavior. Second, unusual behaviors may occur about as frequently when the moon is not full, but the salience of the full moon, together with legends about its powers, may lead people incorrectly to assign it a causal role.

3. *Climate, Weather, and Atmosphere*

Many people believe that weather influences them. One survey found that 71 percent of people asked said that their mood was affected by that day's weather, and 39 percent said it affected their mood 'strongly' or 'very strongly.'

Investigating the objective effects of weather variables is difficult because they cannot be experimentally manipulated except in a climate chamber, and then there are problems with the realism of the setting. Some researchers use special statistical methods to probe the truth of the matter.

Based on the best research, however, weather does affect individuals both indirectly and directly. For example, warm weather brings people outside to work and play, and they injure themselves more often than in cold weather. People stay indoors more when it is cold, which induces 'cabin fever' and depression and leads to more calls to help lines.

But temperature may also have more direct influences on human behavior. As the temperature moves from comfortable to hot, collective violence such as riots, domestic violence, assaults, and rape increase (Anderson 1989). Higher temperatures seem to increase crimes against persons, but not against property. This relation holds only until the 85–90°F (30–33°C) range; aggression then decreases because the temperature becomes too hot to function easily.

Cold temperatures appear to facilitate aggressive acts that are swift and more specifically directed toward ambient events, such as insults. Cold is not strongly related to crime except robbery; this may represent a survival strategy for people who desperately need food or shelter. Other meteorological components such as wind and precipitation have not received enough research attention to draw firm conclusions.

Combinations of weather components may have their own effects; researchers report that assaults and family disturbances tended to occur just after days that were warmer and dryer, with less wind. Another kind of weather combination is 'bad' weather, which usually means cold, wet, and windy conditions with fluctuating barometric pressure; 'good' weather is warm, dry, and calm with stable barometric pressure. Among preschoolers, good weather, and weather changing from good to bad, are associated with more playing with objects than other children; bad weather and weather changing from bad to good are associated with more playing with other children than objects.

4. *Nature the Restorer*

The belief that nature is good for people is ancient. Long ago, when the largest cities were small towns by today's standards, the hanging gardens of Babylon and walled gardens in Mesopotamia were nurtured so those early city dwellers could maintain some contact with nature.

4.1 *Looking at Nature*

People much prefer, in general, to look at nature scenes than urban scenes. Even when nature is potentially dangerous, many people want to experience it; in one survey, 42 percent of those who visited the outdoors responded positively to the idea of hiking in an area which was posted with a sign warning of bears in the area, and 86 percent said they would like to see a grizzly bear in the wild—from a safe distance.

Merely looking at nature is restorative. In a classic study, surgery patients who had a view of nature from their hospital room windows (as opposed to a view of a brick building) recovered faster: they had shorter hospital stays, fewer problems noted by nurses, and fewer postoperative medical problems.

4.2 *Being in Nature*

Nature restores people by facilitating cognitive freedom, ecosystem connectedness, escape from routine, challenge, growth, guidance of young people, a renewed social life, and health. Compared with those who took a nonwilderness vacation or no vacation, backpackers who took a wilderness vacation felt better and performed better on a proofreading task after they returned. Interestingly, the backpackers actually felt worse immediately after returning, but better some time later. Possibly they were a bit depressed at first about the return to society!

Wilderness experiences appear to foster self-actualization, the fulfillment of one's potential for personal growth and psychological health. In a large-scale survey, individuals who spent time in wilderness areas were more self-actualized than those who did not.

4.3 Mechanisms of Restoration

The way in which nature restores functioning has been the subject of some debate. One theory, the mental fatigue approach, asserts that nature is inherently fascinating, that it compels involuntary attention (Kaplan and Kaplan 1989). This is restorative because ordinarily, in the workaday world, people are required to actively focus their attention much of the time. Focused attention is taxing. Nature provides a setting for less-taxing involuntary attention; being in nature gradually refreshes individuals.

A second theory, the biophilia approach, holds that humans evolved for about two million years in natural environments, and have only lived in cities for a tiny fraction of that time (Ulrich 1993). Genetically, people are much more adapted to natural than built settings. Being in nature is like going home, genetically, and this is very restorative.

5. Natural Disasters

Nature often has negative effects on people, too. Fires, storms, floods, avalanches, earthquakes, and animal attacks kill about 250,000 people each year around the world (Burton et al. 1978).

The psychological dimensions to these calamities may be divided into the predisaster phase, which includes preparation and thinking about impending disasters, the period during which the disaster strikes, including thinking and behavior during disasters, and the postdisaster phase, including mental health and coping with the aftermath.

5.1 Before the Disaster

Before a disaster, one focus is on attitudes toward the possibility of disaster, which include knowledge and perception of the risk, degree of trust in authorities, concern, fear, or anxiety.

Generally, greater concern, less trust in officials, greater perceived risk, increased anxiety, and environmentalism are interrelated. Those who are fearful or concerned about one natural hazard are likely to be concerned about other hazards. These clusters of thoughts and feelings are similar across cultures. Women report greater fears, and members of families and younger people usually report more concern or fear than single or older people.

Officials must be careful with warnings; one prediction of a quake that did not happen produced mild but widespread stress disorders in children. Warnings also must be carefully planned because people respond differently depending on which medium carries the warning, because warnings carried by radio, newspapers, and TV vary in memorability, vividness, and imaginability.

Closeness to the site of the hazard also affects concern. Three forms of closeness may be identified: (a) physical proximity, (b) place attachment or length of residence in the threatened community, or (c) closeness in the sense of having experienced the hazard before. Living close generally leads to greater concern and less trust of official warnings. However, concern does not always rise directly with closeness; in fact the reverse sometimes occurs. Particularly when a hazard is difficult to control, residents tend to deny or downplay it. Residents who have lived in an area longer tend to believe the probability of a future disaster is greater. Experience with hazards obviously affects attitudes toward them; in one study, people who had experienced an earthquake were more concerned about future quakes than those without experience. More educated people tend to have fewer fears and concerns about hazards.

When people in an affected area discuss hazards, perceived risk may be socially amplified, that is, risk estimates can be increased, decreased, or shaped by the social, cultural, psychological, and institutional context in which risk is discussed.

Often, serious conflicts arise between hazard experts and laypersons. This happens for at least five reasons. The two groups often (a) use different terminologies, (b) try to solve different problems, (c) disagree about which solutions are feasible, (d) perceive the facts differently, and (e) whereas experts consider risks and benefits separately, most laypersons do not separate the two. In general, experts employ technical models for estimating risk and make decisions that place more value on rationality, efficiency, and expertise, whereas laypersons base their risk estimates on a democratic model that places more emphasis on personal, experiential, and social values.

Disaster preparedness is important but often inadequate. One reason is bounded rationality, the tendency to perceive and adopt a narrower than necessary range of preparations. Individuals also usually prefer crisis response patterns, that is, to wait until disaster strikes before preparing. Finally, individuals routinely misperceive risks from environmental hazards.

5.2 During and Soon After the Disaster

What is known about behavior, thoughts, and feelings as the disaster unfolds? Most individuals do *not* panic during natural disasters: there are many well-documented cases of victims acting as calmly, rationally and effectively as possible.

After the first seconds of shock, people move to protect what is dearest to them. Some, of course, do collapse completely or act in ways that endanger their own and others' lives. However, one frequent immediate postdisaster phenomenon is the therapeutic community, when even unacquainted people work hard together.

Also, before a disaster, many individuals believe they will be able to control dangerous events—an illusion of control. When fate spares a person, even when neighbors are struck, this illusion is strengthened.

5.3 *After the Disaster*

Certain postdisaster cognitive tendencies are unfortunate. For example, some nonvictims see *victims* as blameworthy, perhaps because they live in (or failed to move from) the affected area.

Unless the disaster is extreme, most residents do not move; ties to jobs, family, friends, and the physical setting result in place attachment, which tends to keep people in a place even after a disaster.

One interesting postdisaster behavioral change is the rate of marriage among survivors. A study in the USA of six large-scale disasters found little evidence of change, but the huge 1976 Tangshan earthquake in China had an important effect. Remarriage after the death of a spouse is traditionally discouraged in China, but 2,000 of 3,000 widowed people in Tangshan remarried within a few years, primarily to provide economic security and to ease personal suffering.

The major outcomes of disasters are, of course, stress and other mental health problems. Many victims show signs of post-traumatic stress disorder, but depression, phobias, and medically unexplained symptoms such as abdominal pain, vomiting, nausea, fainting, amnesia, and even paralysis are also seen.

In many ways, victims avoid the overwhelming reality facing them. For example, those who survived the November 1985 volcano and mudslide that killed about 22,000 people in Armero, Colombia often didn't want to know the full extent of the tragedy and resorted to primitive or 'magical' thinking styles (Cohen 1987).

Many studies of mental health after major disasters have been conducted. The results of 52 studies were quantitatively summarized in a meta-analysis. Compared to normal conditions, a 17 percent increase in the rate of psychopathology was found in the wake of disaster (Rubonis and Bickman 1991). Of course, this figure is averaged across all the studies. Mental health problems are greater when more people are killed. For example, in the Armero disaster major depression and post-traumatic stress disorder affected about 50 percent of the survivors.

Fortunately, mental health problems appear to decline with time. However, this does not occur fully or completely. Among victims of the Armero disaster, the rate of emotional distress over a four-year period declined from 65 to 31 percent. One could say that stress declined by 50 percent, but one could also say that after four years, almost a third of the survivors were still suffering. Psychological testing more than 12 years after the great Tangshan earthquake showed that survivors were still affected.

How can stress be relieved? Survivors must cope somehow. Some do so by using problem-focused coping, some by emotion-focused coping, and others use denial. Each style may help, depending on the nature of the disaster. When the situation is very difficult to change, a problem-focused approach may increase distress. Generally, support from family and friends helps, but may help only some—for example, according to one study, men who were not severely stressed and are offered support by nonfamily members who are their seniors.

Survivors appear to follow a three-part social stage model of coping. After the disaster, victims openly talk and think about it for about two weeks. In the next stage, lasting about six weeks, they inhibit talking about the disaster but continue to think about it. During this stage, some indicators of distress increase because victims are thinking but not talking. Finally, an indefinite adaptation stage arrives, during which victims neither talk nor think about the disaster. This does not mean that all the disaster's effects are over; symptoms may remain under the surface for years.

Psychologists have tried to model the ways that authorities and victims will behave during a disaster. Just as one might study how a new jet design might react by using a wind tunnel, social scientists work at predicting behavior during disasters.

6. *Conclusions*

Reports that extraterrestrial forces from cosmic rays, planets, and the moon affect human behavior probably are false-positive errors; no plausible mechanisms by which such forces would influence people have been advanced. Earthly forces do have important effects on people. Aggressiveness rises with temperature to a point and then declines. Ozone levels may influence violence as well as health. Electromagnetic radiation appears to affect public fears more than it affects health.

Nature can be both restorative and destructive. More research is needed, but it seems to restore by refreshing attention capacity, improving moods, facilitating cognitive freedom, increasing a sense of ecosystem connectedness, providing an escape from daily life, presenting achievable challenges, offering a setting for renewed social life, and health through exercise. Even merely viewing nature apparently is restorative.

Attitudes about environmental risk vary with gender, education, proximity, disaster experience, and media exposure. Once formed, attitudes toward hazards are, in turn, related to the quality of disaster preparedness. Laypersons and experts differ dramatically in their perception of risk; efforts toward mutual understanding are necessary. Bounded rationality, under-preparation, crisis response, and risk misperception often blunt effective personal action before and during disasters. During disasters, losses may be high but individuals generally act in relatively rational ways. The outcome of disaster is stress in many forms. Some forms of coping reduce stress better than others. Community attachment and economic factors often prevent residents from vacating high-risk areas; they would rather see governments engage in massive public works to protect their homes.

See also: Bounded Rationality; Coping across the Lifespan; Disasters, Coping with; Environmental Psychology: Overview; Military and Disaster Psychiatry; Nature, Concepts of: Environmental and Ecological; Post-traumatic Stress Disorder

Bibliography

Anderson C A 1989 Temperature and aggression: Ubiquitous effects of heat on occurrence of human violence. *Psychological Bulletin* **106**: 74–96

Burton I, Kates R W, White G F 1978 *The Environment as Hazard*. Oxford University Press, New York

Cohen R E 1987 The Armero tragedy: Lessons for mental health professionals. *Hospital & Community Psychiatry* **38**: 1316–21

Kaplan R, Kaplan S 1989 *The Experience of Nature: A Psychological Perspective*. Cambridge University Press, New York

Rotton J, Kelly I W 1985 Much ado about the full moon: A meta-analysis of lunar-lunacy research. *Psychological Bulletin* **97**: 286–306

Rubonis A V, Bickman L 1991 Psychological impairment in the wake of disaster: The disaster–psychopathology relationship. *Psychological Bulletin* **109**: 384–99

Ulrich R S 1993 Biophilia, biophobia, and natural landscapes. In: Kellert S R, Wilson E O (eds.) *The Biophilia Hypothesis*. Island Press, Washington, DC

R. Gifford

Natural Law

Natural law doctrines have in common the idea that universally valid legal and moral principles can be inferred from nature—human nature in particular. These doctrines vary widely in content and method. In the West, where such doctrines have principally evolved, natural law began with Aristotle; proceeded to Rome under Stoic tutelage; merged with Judeo-Christian theology in the works of St. Augustine and St. Thomas Aquinas; was later adopted in the Reformation and Enlightenment; provided an ideological basis for the American and French Revolutions; and continues to exert influence at the beginning of the twenty-first century.

Natural law, which combines epistemological and normative elements, has been a continuing concern of philosophers and theologians in the western world. Natural law theorists characteristically have asserted that there are ways of knowing how society should be ordered and that this knowledge can be derived from an understanding of the world and of humanity. Many natural law theorists have attributed the natural order of the universe to God, but secular philosophers have also adopted natural law assumptions.

1. History

Natural law thinking originated in Athens with Aristotle and the Stoics. In the Athenian *polis*, unity and diversity provided the thoughtful philosophers of the Academy and the early historian-anthropologists with the central question: how can human society be organized to ensure order and virtue? This question has remained central in natural law theories, even though the assumptions and the terms have changed over time.

Fundamental changes in natural law ideas have occurred as societies have posed new problems. Natural law thinking in pre-Christian Rome, responding to the diversity of the Empire, led to the distinction between the law of nations and the natural law. This limited the scope of natural law, while preserving the Ciceronian idea of fundamental principles. In the Christian era, natural law was seen as justifying the authority of the church temporal, throughout the decline of Rome and the chaos of the Holy Roman Empire. When the authority of the church was challenged at the Reformation, major theoretical developments reflected the changes. Samuel Pufendorf (1673) and Hugo Grotius (1625) used natural law doctrine to justify the authority of nation states and to order their relations in war and peace. And when monarchical authority, especially the divine right of kings, came into question, natural law doctrine was used to justify new democratic doctrines in both France and America during and after their popular revolutions. In the modern world, natural law ideas continue to draw on the tradition and to apply it to problems posed by a world order struggling to be born. Some illustrative detail follows.

1.1 Athens

Most natural law theorists acknowledge Aristotle as the first philosopher to have raised central questions in this area. While he did not develop a systematic philosophy of natural law, he did discuss several of the

issues that have concerned natural law theorists in the millennia that followed. A key question that concerned Aristotle was whether there is any single moral standard by which humans can be guided. Aristotle proposes that virtue is the standard which, if fully practiced, can provide the highest satisfaction in life. The term virtue is used with respect to both individual character and proper relations with others. In relations with others, complete virtue is described as 'justice.' It is complete, Aristotle says, 'because he who possesses it can employ his virtue in relation to his neighbors and not merely by himself; for there are many people who are capable of exhibiting virtue at home, but incapable of exhibiting it in relation to their neighbors' (1987, p. 146).

The polity is the principal means by which virtue is promoted. While acknowledging that socialization helps, Aristotle maintains that 'We shall need laws to teach us what is right' (1987, p. 355) and these laws must be backed by sanction. He further argues that law must take account of the special circumstances in each polity, but should be directed toward the single goal of encouraging virtue in the citizens.

Aristotle declines to say whether virtue is sent by the gods or whether it is the result of human 'learning and discipline of some kind' (1987, p. 29), but even if the latter were true, 'it would appear that that which is the prize and end of virtue is the supreme good and is in its nature divine and blessed.' It is clear from the context that the duty of generating virtue-promoting law rests in his view on legislators and that they rarely succeed. While his is an appeal for a more active state, Aristotle recognizes that some of the responsibility in the event of legislative default should be assumed by every citizen 'to further the cause of virtue in his own children and friends, or at least to set before himself the purpose of furthering it.'

Aristotle suggests that the administration of justice is in the hands of the magistrate. That role makes the magistrate 'the guardian of justice and, if of justice, then of equality.' To play that role properly, the magistrate 'labors in the interest of others, which is the reason why justice is called the good of others.' Since the magistrate does not assign to himself a larger share of the good, 'some reward must therefore be given him in the shape of honour or privilege' (1987, p. 166).

Thus Aristotle raised a set of questions that set the stage for natural law concerns. These ideas were carried forward by the Stoics in a way that emphasized individual virtue, rather than interpersonal relations, the law of the state, or the relation to God. In the subsequent development of the idea, however, the Aristotelian questions continued to play an important part.

1.2 Rome

If the Stoics can claim an important role in the history of natural law thinking, it is largely through their influence on the Roman aristocracy. Cicero (106 to 46 BC) was the great orator of the Republican period and Marcus Aurelius (121 to 180 AD) was the Emperor who recorded his Stoic philosophy. Their views of natural law reflected the culture of the period in which they made their major philosophical contributions.

Cicero sought in nature a standard by which the law of any given community should be judged. Recognizing that the actual law of nations fell far short of this standard, he urged the use of reason as the method by which wise men as legislators and magistrates could make law and justice approach the ideal standard that inheres in nature. In his view, only one standard of law can exist. Positive laws that depart from this standard are not to be considered law at all. 'It may thus be clear,' he declared, 'that in the very definition of the term "law" there inheres the idea of choosing what is just and true' (1928, pp. 317–391).

Cicero's formulation of natural law emphasized its universality. It was the task of Roman law, in his view, to take account of the distinctive customs and laws of societies while developing a standard that would apply to all nations. 'True law is right reason consonant with nature, world-wide in scope, unchanging and everlasting.' He goes on to say that, 'We may not oppose or alter that law, we cannot abolish it, we cannot be freed from its obligations by any legislature, and we need not look outside ourselves for any expounder of it. The law does not differ from Rome and Athens, for the present and for the future, but one eternal and unchanging law will be valid for all nature and for all times' (cited by Haskell 1964, pp. 62–63).

That idea continued to be influential in Rome, through the time of Marcus Aurelius, and beyond. During the Empire, however, Stoicism served primarily as a moral guide to individuals, rather than as a system of social thought. If natural law concepts survived, they were relegated during the Imperial period to a lesser role in thought and action. Instead, the law of nations developed to take into account the diversity of custom, politics, and military security that served to provide a legal basis for maintaining the hegemony of Rome.

Although the Stoic school effectively ended about the time of Marcus Aurelius, the concept of natural law continued to live in Roman culture. It was in fact revived in an important way in the works of St. Augustine, who utilized the Stoic ideas of individual virtue and the Aristotelian–Ciceronian theme of justice to be achieved through a universally valid legal order. These two themes he saw as elements that must serve the will of God. And the instrument of God's will on earth was to be the Christian commonwealth, under the guidance of the Christian Emperor.

Augustine's formulation has had a profound impact on natural law thinking. Unlike Aristotle, he had little doubt that natural law reflected the will of God. Drawing on the Hebrew and Greek Scriptures, he

emphasized the concept of one God whose moral directives were clearly manifest in the Ten Commandments, the words of the prophets, and the teachings of Jesus. He was able to use the Greek and Roman ideas in a way that subsumed individual virtue and social justice as part of God's will on earth. His synthesis, surviving the collapse of the Roman Empire, became the guiding philosophy of medieval Europe. Eight centuries later, it provided a major foundation on which St. Thomas Aquinas constructed his cogent and influential conception of natural law.

1.3 Thomas Aquinas and the Scholastic Tradition

In the medieval period, universities developed in Europe that revived the traditions of learning that had virtually been lost after the fall of Rome. Paris in the thirteenth century was the most active center of intellectual ferment in all of Europe and Thomas Aquinas (1225–1275) had access there to the writings of Aristotle, the teachings of the Stoics, Cicero's orations, Marcus Aurelius' *Meditations*, and Augustine's *Confessions* and *City of God*. In addition, Thomas had the writings of two Muslim scholars, Avicenna (980–1037) and Averroes (1126–1198), and the Jewish scholar Maimonides (1135–1204). All of these thinkers helped him to advance an intellectual synthesis that fit his times and has caught the attention of scholars and seekers ever since.

As the Renaissance gathered strength, there was tremendous excitement over the possibilities that emerged. With the growth of commerce came the awareness of other civilizations, ancient and modern. Into the cities came the most inquisitive along with the most acquisitive. There arose an intellectual and social ferment in the cities that provided opportunities for diverse ideas and actions. That newly found freedom raised again the kinds of questions that had been formulated and considered in Athens and Rome. It was in this context that Aquinas rendered his synthesis of Greek and Christian philosophy.

At the core of Thomas' social philosophy was the idea of natural law. This concept provided him with the strength of two traditions that had been brought together by Augustine. His system assumed that the one God was the ultimate authority. For the individual believer, that provided spiritual support for creative philosophical thought.

It also helped to gain acceptance by the church of doctrines that might otherwise have been viewed as heretical. If God's law is recognized as the ultimate authority, Thomas' systematic theology encouraged the effort to approximate God's will by finding in nature the kind of law that should apply universally. Human law, he acknowledged, can only approximate this effort, but the effort must be made.

In this system of thought, many of the ancient elements are present. The Stoic idea is reflected in that responsibility devolves on the individual to reason, to understand, and to implement God's will. But the system depends on more than individual virtue. In addition, it requires the construction of laws that constrain evil and encourage the good according to universally valid principles. Those tenets are more Aristotelian than Stoic. Moreover, Thomas offers the hope of a natural law that has a place for everyone, or at least for all who have faith in God—a belief that follows pre-Reformation Christian teaching more than Greek or Roman philosophy.

Several points inhere in Thomas' synthesis. Natural law exists in nature. It is not overridden or imposed by God the Creator. It is for humans to understand the law of nature, and to implement it in the real world through practical reason. In this way, the individual is virtuous and human law comes closer to the law of nature.

The fundamental assumption is that we must continually strive through the use of reason to approximate The Heavenly City. The conception of a just society is one that is fixed and permanent. It is the duty of the virtuous man to understand that conception and to help implement it. Thomas sought in this way to help stabilize the society through his integration of philosophy and theology. Many subsequent commentators have generally used his scholastic concepts in this way.

The strength and durability of the scholastic tradition is manifest in the work of Francisco Suarez, a Jesuit, who following Aquinas covered many of the same questions in his extensive work called *De legibus* (*Concerning the Laws*) published in 1612. One point of interest in Suarez's work is the suggestion that circumstances can affect the natural law standard. Thus, if a person and his family are starving, that condition makes a difference in what is right for that person to do. Suarez would have the right of property overridden if the need is great enough in such circumstances. Additionally, Suarez contended that natural law standards are not limited to the obligatory and the forbidden. He also talked of the permitted, a concept that explicitly restored discretion to the virtuous person. Further, he would temper the absolute power of the king by reference to the preferences of their subjects. In these regards, Suarez showed a concern for freedom, equality, discretion, and judgment, but without explicitly departing from the scholastic position.

1.4 Changing Conceptions

A major change in natural law thinking resulted from the work of Hugo Grotius (1583–1645). Writing in Protestant Holland, he undertook to justify the Dutch challenge to the Spanish Empire. Focusing on the relations between sovereign states, Grotius developed

new norms which focused on the rights of each nation to defend its interests, if necessary, through the use of war. To justify that position, he emphasized the theme of freedom. Individuals, he maintained, should be free to pursue their own interests even if they conclude that those interests could best be served by accepting slavery. Nations also must be free to choose whether to fight an enemy or to capitulate. In the rational pursuit of their self-interest, rational actors must be free to judge what was best for them. That freedom should be exercised, however, within certain limits, if the legitimate interests of all nations were to be protected. Grotius spelled out when wars were justified and how they should be fought. In so doing, he expanded the area covered by natural law and proposed new ways of justifying its provisions.

It was not Grotius' goal to abolish war, but to outline a normative setting within which the parties could contend. His method provided ground rules for war, and for resuming peace. In recognizing the legitimate interests of contending parties, he provided a way out of interminable wars and enduring enmities. In this way, he opened the door to peacemaking. As war becomes increasingly violent in the contemporary world, Grotius' suggestions for limiting or preventing war through international law become increasingly relevant. Natural law philosophy could play a part in supporting, if not steering, movement toward a world rule of law.

Following Grotius, many writers utilized the basic idea of natural law to justify their views of the world. Among the prominent authors whose works fit under this heading are Samuel Pufendorf, Thomas Hobbes, John Locke, Jean-Jacques Rousseau, Adam Smith, Edmund Burke, and John Stuart Mill. Each of them used the concept distinctively, but in so doing they contributed to the view that there exists in nature some basis for law that deserves attention. As scholastic philosophy lost its exclusive ownership of this idea, others were able to use it to their ends. This trend was expedited by the social and intellectual movements we know as the Renaissance, the Reformation, and the Enlightenment.

2. *Natural Law as a Factor in Social Movements*

The conviction that there are principles of law to be discerned in nature can strengthen and consolidate social movements. Examples of such effects are found repeatedly in British history, in the French Revolution, in the German reform movements of 1832 and 1848, and in the movement led by Garibaldi for Italian unification. Even when the term natural law is not explicitly used, some belief systems claim valid support for major social change on grounds that a new order is required by principles rooted in physical, human, or societal nature. In traditional Chinese culture, for instance, there is a widely held belief that a particular regime must be replaced if (as evidenced in natural disasters such as floods or earthquakes) the mandate of heaven has been withdrawn from the current dynasty. Another example is the Marxist idea that proletarian revolution inevitably leads toward a more just social order. (For a natural law approach to Marxism, see Bloch 1977.)

One of the best-known examples is found in the American Revolution. It is clear from many historical sources that the leaders and popular supporters of the American Revolution used natural law to justify their beliefs. English legal tradition from the time of the Magna Carta had been suffused with the idea that valid law must be consistent with nature. Moreover, the common law continued to provide affirmation of the view that justice inhered in nature and that it was the duty of reasonable people to ascertain how the law of nature applied to any given situation. Against that background, it comes as no surprise that the colonists reacted to the injustices (as they saw them) of the English king as violations of the laws of nature. In that light, they declared themselves independent of Great Britain and claimed 'the separate and equal station to which the laws of nature and Nature's God entitle them.'

Through the years, natural law has demonstrated remarkable vitality. When a new normative consensus is forming, natural law can be used to strengthen the force of agreement. A recent example is found in Martin Luther King's letter to white clergy from the Birmingham jail, in which he explained his justification for breaking the law in opposing segregation. Adherence to just laws he strongly favored, but laws 'that relegate persons to the status of things' are unjust laws. 'To put it in the terms of St. Thomas Aquinas, an unjust law is a law that is not rooted in eternal and natural laws.'

As an example, 'an unjust law is a code inflicted upon a minority which that minority had no part in creating because they did not have the unhampered right to vote' (King 1963). The letter contributed to the passage of the Voting Rights Act in the following year.

The natural law theme often surfaces when societies experience new needs whose solution requires a wide consensus implemented by law. In recent years there has been a growing awareness of the dangers of violence in a nuclear world. The answer to such dangers might well be found in a world rule of law. But such a legal order, to be legitimate, must be supported by a widespread consensus. Natural law concepts might, if adapted to our times, help fill that need.

There are signs of a revival of interest in natural law that address this problem. One example is found in the work of John Finnis and Germain Grisez, who have become leading figures in the study of natural law. In Finnis and Boyle (1987), they explore the morality of nuclear deterrent policy using natural law as the basis

of their critique. In this and their other writings, Finnis and Grisez modify the scholastic tradition that is their starting point and move on to a more general formulation than could have been predicted by Aquinas. In some sense, their work returns to the Aristotelian starting points while retaining the innovations that have attached to natural law over its entire history.

As humanity struggles with serious new challenges, we can expect to see more examples of this intellectual history unfolding. Still needed is a way of integrating the natural law philosophy with the insights and methods of the social and behavioral sciences. This article has described the origins and developments of natural law thinking as it has occurred in different social settings. Social and behavioral scientists might well ask whether the idea has enough vitality and relevance to contribute to the evolution and legitimization of a world rule of law.

See also: Aristotelian Social Thought; Aristotle (384–322 BC); Hobbes, Thomas (1588–1679); Justice and Law; Law and Morality: A Continental-European Perspective; Law and Morality: An Analytic Perspective; Locke, John (1632–1704); Mill, John Stuart (1806–73); Rousseau, Jean-Jacques (1712–78)

Bibliography

Aquinas, Thomas 1988 *On Law, Morality, and Politics*. Hackett, Indianapolis, IN

Aristotle 1987 *The Nichomachean Ethics*, trans. Weldon J E C. Prometheus Books, Amherst, NY

Augustine 1958 *The City of God*, Bourke V J (ed.). Image Books, New York

Bloch E 1986 *Natural Law and Human Dignity*. MIT Press, Cambridge, MA

Kretzmann N, Stump E (eds.) 1993 *Cambridge Companion to Aquinas*. Cambridge University Press, Cambridge, UK

Cicero M T 1991 [1928] *De legibus*. Packard Humanities Institute, Los Altos, CA

Declaration of Independence of the United States of America

Finnis J, Boyle J M Jr. 1987 *Nuclear Deterrence, Morality and Realism*. Clarendon Press, Oxford, UK

Grotius H 1625 *De jure belliac pacis libritre (The Rights of War and Peace)*. N. Buon, Paris

Haskell H J 1964 *This Was Cicero*. Premier, New York

King M L Jr. 1994 *Letter from the Birmingham Jail*, 1st edn. Harper, San Francisco, CA

Pufendorf S 1673 (1991) *On the Duty of Man and Citizen According to Natural Law*. Cambridge University Press, Cambridge, UK

Suarez F 1944 *Selections from Three Works*. Oxford University Press, Oxford, UK

R. D. Schwarz

Natural Selection

It is a well-known fact that the majority of resources providing the energy for life on our earth are limited. On the other hand, organisms have the potential for exponential reproduction. This is easily seen by the fact that organisms produce more offspring than can survive. As a consequence, populations are eventually limited in size because of restriction in access to resources as, for example food, except if some other factor of mortality as predation keeps them below their environmental carrying capacity.

Due to the excess in fecundity of organisms living in an environment with inadequate resources to support all the young, organisms compete over such resources. Competition can be direct, by aggressive interactions or by defending territories, or indirect, by eating food that is no longer available to other individuals. Organisms compete especially with conspecifics, because conspecifics are most similar in their needs and use of resources. Such inter- and intraspecific competition is an important although not the only factor promoting evolution.

Individuals that are better able to cope with such competition because, for example, they can more efficiently digest food than other individuals in the same population, use more energy for reproduction, and may leave more offspring surviving into the next generation. If the ability to digest food is heritable, it will be passed on to the offspring and also improve the progeny's reproductive efficiency. Because only successful competitors survive and reproduce, any character that improves such rivalry will spread in a population. This is in brief how natural selection works, which is a basic mechanism of evolution (recent textbooks on evolution by natural selection are Ridley 1996, Futuyma 1998, Stearns and Hoekstra 2000).

1. The Concept of Natural Selection

Charles Darwin (1809–82) was among the first to formulate a scientific theory how evolution works. As an explanation for how the diversity of life with so many species and large variations in characters came to exist, he formulated the concept of natural selection (Darwin 1859). Natural selection was his central principle of how evolutionary change occurs, and it is still considered to be of paramount importance in evolutionary biology. Other factors definitely also play an important role in evolution, as, for example, genetic drift or chance events, but natural selection of hereditary variations usually is raised as the first and main hypothesis to explain even complex morphological, physiological, or behavioral characteristics of organisms.

Darwin based his theory on three general observations:

(a) Variation: Members of a species typically do not look alike but differ in some characters.

(b) Heredity: Offspring are more similar to their parents than to other conspecifics, suggesting that parents pass on characters to the progeny.

(c) Differential reproduction: Some parents leave more progeny than others, due to some inherited characters.

Darwin had developed his theory without knowing the process of how exactly offspring inherit their parents' characters. Gregor Mendel published his discovery of genes and of the principles of genetics as a contemporary of Darwin. However, it was not before the 1930s until the 'Modern Synthesis' combined the contributions of genetics, systematics, and paleontology into a new neo-Darwinian Theory (Young 1994). Due to the growth of genetics, we nowadays have detailed knowledge about how genes encode the information that is given from one generation to the next. Genes are copied and transmitted to the offspring when organisms reproduce. The variation that Darwin observed, thus, is based on genetic differences among individuals. Within populations, genes typically occur in two or more alternative forms, so-called alleles. Different alleles produce different forms of the same protein and result in different morphological, physiological, biochemical, behavioral, or other properties of an organism that develop through action of genes and environment. In other words, differences in genotype (the genetic constitution of an individual) result in different phenotypes.

As already emphasized by Darwin, given the three conditions as described before, natural selection always operates. If different genes or differences in hereditary characters have an influence on the survival or reproduction of an individual, such characters also promote with a high probability the individual's offspring survival or reproduction. As a consequence, an allele responsible for improved reproduction becomes more common in the population and will increase in frequency in the next generation.

Natural selection operates whenever genotypes differ in their relative contribution of offspring to the next generation. Generally speaking, natural selection acts on differences in individual lifetime reproductive success, and thus results in the maximization of an individual's biological fitness. Selection on individuals will favor alleles that help to build an organism with a relatively high probability to survive and reproduce successfully in its environment. Traits that benefit an organism to maximize its biological fitness function as adaptations.

2. *The Importance of the Environment*

When natural selection works, a modified trait that confers higher fitness replaces another version that now yields lower fitness. This process of individual selection takes place in a situation with numerous selective factors, and most traits of organisms, actually are adaptations to the prevailing environmental conditions. The ecological environment includes both abiotic features, as climate, salinity, soil type, or availability of water, and biotic factors, as food supply, prey, predators, parasites, or conspecifics. The features of the environment that are important differ from species to species because of their different evolutionary histories.

Given that the relationship between genetic differences and reproductive success in the natural environment remains constant, alternative alleles at a specific gene locus are completely replaced. Environmental conditions, however, can change unpredictably and over long periods of time. Furthermore, catastrophic geologic events as mass extinctions, even if rare, impose strong forces and have a major influence on the course of evolution. This is why selection pressure is not constant over many generations and why scientists typically study the adaptive value of a trait by focusing strictly on its current fitness effects.

3. *Genetic Variation Arises through Mutations*

According to the concept of individual selection populations should differ in traits. An organism is a complex system, the development of which is controlled by a genetic program embodied in its DNA, the genetic material. When the DNA is replicated within the cell, errors in copying occur from time to time. Each copying error is a mutation that results in an altered genotype. If such an alteration changes the bodily structure or the phenotype of an organism, natural selection begins to operate on such trait.

4. *Natural Selection Explains Evolution and Adaptation*

The process of natural selection explains both evolution and adaptation. It is 'a process by which the forms of organisms in a population that are best adapted to the environment increase in frequency relative to less well-adapted forms over a number of generations' (Ridley 1996). This fact is illustrated beautifully in long-term field studies of a group of finches living on the Galapagos Archipelago. This group of birds became known as Darwin's finches because Charles Darwin had been the first to study them when he visited the islands during his voyage on the *H. M. S. Beagle* (1831–36). He had been impressed by the way how different species of finches differed in the size and shape of the beak. Already in the 1940s,

David Lack identified 14 separate species of finches that differ in the beak, body size, and male and female plumage. Otherwise, the birds are so similar in appearance and in behavior that he was convinced that they were all descended from a single ancestral species that colonized the Galapagos Islands.

Most of these species can be found on several islands. The island forms of the same species often also differ in the same characters that distinguish different species: in beak, body size, and plumage. Lack concluded that 'new species have arisen when well-differentiated island forms have later met in the same region and kept distinct' (Lack 1947). Members of the ancestral species have colonized one island after the other. After having arrived on an island they reached species level due to geographic isolation from other members of the original founder population. After they reached species level they recolonized the islands from which they had come.

Lack also provided evidence that the main beak differences between the species are adaptations to differences in diet based on the observation of clear correlations between a species' beak size and its method of feeding. Seed-eating species had heavy finch-like beaks, the flower-probing species had rather long beaks, and so on.

The correlation with diet was evident even in closely related species where the beaks differ only in size, as in the three species of seed-eating ground-finches of the genus *Geospiza*. These species eat many of the same seeds; however, they also show a clear preference for different sized seeds in proportion to the size of their beak. Detailed studies by Peter Grant and his co-workers have not only supported David Lack's main conclusion but also shown that variable environmental conditions result in changes in mean beak size within populations even over short periods of time as few generations (Grant 1986, Grant and Grant 1989). During longer periods of droughts populations of the large ground-finch *Geospiza fortis* declined fairly dramatically on the island Daphne. This was because the birds feed on seeds produced by the island's plants, and seed production generally is reduced during dry periods. During the next year, however, the scientists observed an increase in the average size of the birds, indicating that survival was not random but that large ground finches survived better than small ones. The supply of smaller and softer seeds, which the finches can feed on easily, soon had been exhausted and the birds were forced to handle larger and harder seeds. Larger birds with larger beaks are better able to cope with the large seeds. As a consequence of strong natural selection in favor of the larger birds, mean body size and beak size in the population of *Geospiza fortis* on Daphne had increased from one season to the next. During wet periods as during El Niño years, however, with an excess production of small plant seeds, small sized birds with relatively small beaks had an advantage. As a consequence, natural selection on body size and beak size fluctuates in ground finches, dependent on the climatic condition and food availability during a given season.

5. *Modes of Selection*

The Galapagos ground finches are an example of directional natural selection. During dry periods natural selection produces an increase in body size because larger finches have a higher fitness. If, however, intermediate phenotypes have the highest fitness, this mode of selection is called stabilizing or balancing selection. Here, natural selection acts against changes of a specific character and keeps the population constant over time. Birth weight in humans is an example of stabilizing selection that is generally a common mode of selection in nature. Babies with a lower or higher than average body weight at birth do not survive as well as babies of an average birth weight. A third possibility is for natural selection to be disruptive when both extremes of a phenotype are favored over the intermediate types. Such disruptive selection may work in sexually dimorphic species, with males being much larger than females (see Ridley 1996).

Furthermore, there are cases where the fitness of a phenotype depends on its frequency in the population, so that the fitness of an individual with a specific character is not constant in a given environment but is frequency dependent. In many frogs, males aggregate during the breeding season near or at ponds and call for mates, usually during the night. Calling males attract females that come to the pond to lay their eggs. Some males, however, are silent and simply remain near others that are calling and try to grasp and mate with a bypassing female that approaches a caller. Calling is energetically expensive, especially when producing low-frequency calls that are most attractive to females. Furthermore, calling males sometimes engage in aggressive fights over access to females. Males that are silent do not have to pay these costs. However, the highest benefit is expected for a silent male in a large chorus of calling conspecifics that attract a lot of females. If too many males are silent, on the other hand, no or only few females are attracted to the site. As a consequence, the reproductive success of a silent male is dependent on the frequency of calling males in its environment.

Frequency-dependent selection, thus, can result in populations containing two—or even more—alternative characters or traits. A form of natural selection that is called heterozygous advantage (or overdominance for fitness), however, can also maintain such a polymorphism. It is the case if individuals carrying two different alleles at the specific gene locus (heterozygote individuals) have a higher fitness than individuals being homozygoes for either of the alleles.

As a consequence, both alleles are maintained in the population by the superior survival or reproduction of the heterozygote. A well-documented example of heterozygous advantage at a single gene locus is sickle-cell hemoglobin in humans. It was one of the first polymorphisms that were understood in terms of its selection process. Hemoglobin occurs in different variants with Hb^A being the normal type. Persons homozygous for Hb^S, however, suffer severe anemia and usually die before they reach adulthood. Heterozygotes have slight anemia, but in parts of Africa where falciparum malaria is common, they have a higher survival probability than homozygotes for normal hemoglobin Hb^A. Other examples of heterozygous advantage are rather rare, and this mode of selection does not seem to be a major cause of polymorphisms in nature.

6. *Adaptation as a Compromise Between Different Selection Pressures*

Natural selection has produced many adaptations that benefit an individual organism. Many evolutionary and behavioral scientists all over the world have collected a wealth of information about how individual selection has shaped the life of such diverse organisms as elephants, toads, sticklebacks, dung flies, or mites (recent text books summarizing such examples are Krebs and Davies 1993, Alcock 1998). Nevertheless, when studying such adaptations, however, it is important to realize that selection from different environmental factors simultaneously affects an individual's behavior. Searching for food in an environment with predators certainly requires different behavioral rules than feeding in a safe place. Adaptations, therefore, typically have to be understood as a compromise between different needs of an organism.

Michael Ryan and his co-workers studied the mating behavior of the Tungara frog (*Physalaemus pustulosus*) on Barro Colorado Island, Panama. Males of this species are active during the night and produce rather complex calls, consisting of 'whining' components that are followed by another component that can be described as 'chucks.' Females prefer males that call and especially those that produce chucks. Nevertheless, some males do not produce chucks especially when they are alone or in small groups. Given that a male's attractiveness for females depends on chucks, why don't they always produce complete calls? The reason is that Tungara frogs have a special enemy, a frog-eating bat. Fringe-lipped bats (*Trachops cirrhosus*) track their prey, especially males, by their calls during the night, and they can better localize males that produce whine–chuck combinations in comparison to males with incomplete calls. The risk of being eaten by a predator is lower for frogs in large choruses because a bat has more individuals to choose from. Males calling alone or in small groups, however, suffer much higher costs in terms of predation. Thus, by producing only whining calls, which are less attractive to females but also harder for bats to locate, they can increase the probability to survive. During the next night, they might end up in a larger chorus and have a better chance to reproduce successfully (Ryan 1990, Ryan et al. 1990).

7. *Sexual Selection as a Special Case of Natural Selection*

Given that individuals ultimately compete over reproduction, selection is typically strong on any characters that influence an organism's access to mating partners. Charles Darwin introduced the term 'sexual selection' to describe 'the advantage which certain individuals have over others of the same sex and species solely in respect of reproduction' (Darwin 1871).

From a biological point of view, males and females in sexually reproducing species differ in the size and number of their gametes. Males typically produce a relatively large number of tiny, mobile sperm cells that contain mainly DNA and some mechanism to provide energy so that sperms can move towards an egg within the female reproductive tract. Females, on the other hand, typically produce much larger gametes. Such egg cells are immobile, and besides genetic material they are also equipped with proteins and carbohydrates as energy stores for the embryo. As a consequence, the difference between the sexes in the size of the gametes can be expressed as a difference in parental investment, the time, energy, and risk that a parent expends in each progeny. This difference starts at the level of the gametes, and often is amplified by nourishment of offspring during pregnancy and maternal care even after the birth of the offspring. The energy available for reproduction, however, is generally limited for each organism. Therefore, investment in one offspring reduces the parent's ability to invest in other or future progeny (Trivers 1972).

A male's reproductive success is not so much limited by its sperm production but rather by its access to females. Thus, males are expected to increase their reproductive success by mating with as many females as possible. The fitness of a female, however, that invests relatively more in each offspring, is likely to be more reduced by unproductive matings. Females, therefore, are expected to be more selective in mating with males. As a consequence, sexual selection is typically stronger among males, which compete against each other for access to females. This pattern is reversed in some species such as seahorses (*Syngna-*

thidae) or phalaropes (*Phalaropodidae*) where males care for the progeny.

Sexual selection occurs in two types, intrasexual and intersexual selection. Intrasexual selection refers to males competing among each other for access to females. In red deer (*Cervus elaphus*) males aggregate in autumn during the rut—the breeding season—and engage in roaring contests and parallel marches. Males gain information about the size and fighting capacity of a competitor during such contests. Only males in very good condition can sustain roaring at a high rate for several minutes, and parallel marches allow judging each other's body size. Smaller males usually leave the area after such contests and avoid energy-consuming fights that they are likely to lose. Males that are similar in size and condition, however, engage in vicious fights with antler-clashing and shoving matches until the less competitive one flees the ground. The winning male acquires the group of females the deer were fighting over, and copulates with them as soon as they are in oestrous (Clutton-Brock et al. 1982).

Sometimes competition among males over reproduction can be more subtle. In species where fertile females mate with several males as in many insects, males have evolved extraordinary devices so that their sperm has an advantage over those of others in fertilizing the female's eggs. Male damselflies have a penis with lateral horns that enable a male to scrub out a female's genital tract. This behavior allows removing sperm from other competitors already stored in the female's sperm storage organ before releasing its own sperm into the female (Waage 1979).

Characters as large size sperm-removing devices or large horns or antlers confer an advantage over access to receptive females and, thus, are adaptive. There are other elaborate male characters, however, as the peacock's train that cannot be explained by that mechanism. In peacocks or in some African widow birds, males have extraordinarily long tail feathers that even seem to lower survival because they reduce maneuverability, the power of flight, and they make the bird more conspicuous for predators. Furthermore, these male characters do not contribute an advantage in brood care because males do not provide resources or other care for the progeny.

It is difficult to explain the existence of such characters by natural selection. Darwin was well aware of that problem and suggested that the reduced survival of males with such handicaps is more than compensated by the bearers increased advantage in reproduction because of female choice. Males with exaggerated traits exist because females prefer to selectively mate with males with longer or brighter or more colorful feathers or other characters.

Darwin introduced the term intersexual for this second type of sexual selection to emphasize that such characters exist because males compete over being chosen or preferred by females. The idea that female choice of males can explain extraordinary characters like the peacock's train is intriguing, nevertheless the question arises why females should do so.

In species where males provide parental care, male exaggerated traits might enable females to identify large and competitive males that successfully protect and care for the progeny. In pied flycatcher (*Fidecula hypoleuca*), females prefer males with black and white plumage over those with a duller, browner appearance. To test whether such preference evolved because black-and-white males are better able to care for the progeny, Sætre and co-workers removed the females and measured offspring growth reared exclusively by either black-and-white or dull males. Indeed black-and-white fathers could feed their young better which consequently had a higher weight than offspring cared for by dull males only (Sætre et al. 1994).

In species where males do not offer resources or parental care, the following three main hypotheses currently are discussed why females choose males with exaggerated traits.

First, females prefer the trait because it is an indicator of male health. Hamilton and Zuk (1982) have suggested that in birds bright or colorful plumage is a sign of health, because sick or parasitized individuals cannot produce such a trait. A similar argument can be applied for other extravagant traits or behaviors in other animals. Male appearance, thus, informs a female about the potential sexual partner's health or parasite load. If resistance to disease or parasitic infection is heritable, females pass on the males' genes for resistance to their offspring when picking a male with characters that indicate that they are parasite-free. This evolutionary mechanism results in the selection of 'good genes' because sons and daughters may inherit their father's viability advantages. Experimental evidence from many species supports the hypothesis of 'good genes selection' by female preference of exaggerated traits, as in sticklebacks or barn swallows (Milinski and Bakker 1990, Møller 1994).

Second, females prefer males with an exaggerated trait because, again, it is an indicator of male health. Females use this information to mate with males that are less likely to transmit lice, mites, fleas, or bacterial pathogens which typically are harmful to the females themselves, or to their progeny. According to this 'healthy male hypothesis,' choosy females benefit because they may avoid contagious diseases and parasites.

Third, in the 1930s the evolutionary biologist R. A. Fisher suggested the idea of 'runaway selection' to explain genetic benefits for females with preferences for exaggerated traits in males (Fisher 1930). A choosy female benefits because her sons inherit a trait that makes them very attractive to females, and because her daughters will prefer males with such an attractive trait. The idea that female mate choice genes and genes for the preferred male trait are inherited together is the

basis for the concept of a runaway process: as new mutations occur, ever more extreme female preferences and male characteristics spread together in a population. According to this theory, the adaptive value to choosy females holds even if the male exaggerated trait actually reduces the survival probability of the bearer of the trait. The runaway process stops only if the trait or behavior gets too costly or risky. In other words, if the bearer of the trait has such a low survival probability that he dies before he even reproduced, natural selection will counterbalance sexual selection for exaggerated, attractive characters (for a detailed review of the underlying arguments based on mathematical models see Andersson 1994).

Natural selection and sexual selection operate fundamentally the same way. Both forms of selection require that individuals that differ in heritable characters differ in the number of surviving offspring. Although sexual selection is just a special case of natural selection, Darwin already intended to emphasize the selective consequences of sexual interactions among conspecifics by introducing the distinction between the two processes. Sexual selection explains characters that cannot be understood as adaptations in dealing with other aspects of an individual's environment, such as access to food or avoiding predators or parasites.

8. *Altruism as a Challenge to the Concept of Individual Selection*

Conspecifics are a major environmental factor for many organisms because they are not only partners to mate with—in sexual species—but also partners in competitive and cooperative interactions. It is the special focus of scientists in the field of sociobiology and behavioral ecology to study such interactions among organisms (Wilson 2000). The interesting aspect here is to understand how individuals in social species influence each other's reproduction. In this context special emphasis is on any behavior where individuals help others to raise young. Biologists term a behavior 'altruistic' if it increases the number of offspring produced by the recipient and includes fitness costs—a decreased number of young produced—for the acting individual. In many animal species individuals invest time and energy in the care of nonoffspring. Such behavior is common in social insects, where many or most members of a society may even spend their entire life in caring for and rearing the young of another female (Hölldobler and Wilson 1990). Helping behavior or allparental care also occurs in cooperative breeding birds and mammals (Brown 1987, Stacey and Koenig 1990, König 1997, Solomon and French 1997). The provision of care to nonoffspring in animal societies has attracted substantial scientific attention because of its apparent contradiction to the concept of individual selection. Can natural selection ever favor altruistic behavior that decreases the reproduction of the actor?

In the 1960s, Hamilton (1964) raised the point that relatedness can be of paramount importance for the evolution of such altruistic behavior. Hamilton realized that an individual can maximize its fitness not only by direct reproduction but also indirectly by helping kin to produce and/or to rear additional offspring. With a certain probability—dependent on the degree of relatedness between the donor and the recipient of help—such offspring will carry copies of alleles identical by descent to the helping individual. John Maynard Smith named this genetical mechanism kin selection (Maynard Smith 1964). According to Hamilton's idea, altruistic behavior is more common in social groups of related individuals than in groups consisting of unrelated conspecifics.

Besides this mechanism of kin selection, altruistic behavior can also evolve via mutualism (each partner gains fitness benefits by performing the altruistic act) or reciprocity (as long as the altruistic act is reciprocated at some later date, both partners will gain fitness benefits; Dugatkin 1998). Both mechanisms allow for cooperation among unrelated individuals, however, require specific conditions so that the altruistic act is stabilized within groups.

Natural selection has not only resulted in many adaptations that benefit individuals in competitive interactions. It has also produced highly cooperative behaviors as the phenotypically altruistic behavior of nonoffspring care, which nevertheless is genetically 'selfish' and, thus, adaptive.

See also: Adaptation, Fitness, and Evolution; Cultural Evolution: Overview; Darwin, Charles Robert (1809–82); Darwinism; Evolution, History of; Evolution: Optimization; Evolutionary Selection, Levels of: Group versus Individual; Evolutionary Theory, Structure of; Human Behavioral Ecology; Natural Selection; Optimization in Evolution, Limitations of

Bibliography

Alcock J 1998 *Animal Behavior: An Evolutionary Approach*, 6th edn. Sinauer, Sunderland, MA

Andersson M 1994 *Sexual Selection*. Princeton University Press, Princeton, NJ

Brown J L 1987 *Helping and Communal Breeding in Birds*. Princeton University Press, Princeton, NJ

Clutton-Brock T H, Guinness F S, Albon S D 1982 *Red Deer. Behavior and Ecology of Two Sexes*. Edinburgh University Press, Edinburgh, UK

Darwin C R 1859 *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Murray, London

Darwin C R 1871. *The Descent of Man, and Selection in Relation to Sex*. Murray, London

Dugatkin A L 1998 *Cooperation among Animals*. Oxford University Press, Oxford, UK

Fisher R A 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK
Futuyma D J 1998 *Evolutionary Biology*, 3rd edn. Sinauer, Sunderland, MA
Grant P R 1986 *Ecology and Evolution of Darwin's Finches*. Princeton University Press, Princeton, NJ
Grant R, Grant P R 1989 *Evolutionary Dynamics of a Natural Population. The Large Cactus Finch of the Galapagos*. University of Chicago Press, Chicago
Hamilton W D 1964 The genetical evolution of social behaviour (I and II). *Journal of Theoretical Biology* **7**: 1–16, 17–52
Hamilton W D, Zuk M 1982 Heritable true fitness and bright birds: A role for parasites? *Science* **218**: 384–7
Hölldobler B, Wilson E O 1990 *The Ants*. Springer, Berlin
König B 1997 Cooperative care of young in mammals. *Naturwissenschaften* **84**: 95–104
Krebs J D, Davies N B 1993 *An Introduction to Behavioural Ecology*, 3rd edn. Blackwell, Oxford, UK
Lack D 1947 *Darwin's Finches*. Cambridge University Press, Cambridge, UK
Maynard Smith J 1964 Group selection and kin selection. *Nature* **201**: 1145–7
Milinski M, Bakker T O M 1990 Female sticklebacks use male coloration in mate choice and hence avoid parasitized males. *Nature* **334**: 330–3
Møller A P 1994 *Sexual Selection and the Barn Swallow*. Oxford University Press, Oxford, UK
Ridley M 1996 *Evolution*, 2nd edn. Blackwell, Oxford, UK
Ryan M J 1990 Signals, species, and sexual selection. *American Scientist* **78**: 46–52
Ryan M J, Fox J H, Wilczynski W, Rand A S 1990 Sexual selection for sensory exploitation in the frog *Physalaemus pustulosus*. *Nature* **343**: 66–7
Sætre G-P, Dale S, Slagsvold T 1994 Female pied flycatchers prefer brightly coloured males. *Animal Behaviour* **48**: 1407–16
Solomon N G, French J A 1997 *Cooperative Breeding in Mammals*. Cambridge University Press, Cambridge, UK
Stacey P B, Koenig W D 1990 *Cooperative Breeding in Birds*. Cambridge University Press, Cambridge, UK
Stearns S C, Hoekstra R F 2000 *Evolution. An Introduction*. Oxford University Press, Oxford, UK
Trivers R L 1972 Parental investment and sexual selection. In: Campbell B (ed.) *Sexual Selection and the Descent of Man*. Aldine, Chicago, pp. 136–79
Waage J K 1979 Dual function of the damselfly penis: Sperm removal and transfer. *Science* **203**: 916–8
Wilson E O 2000 *Sociobiology. The New Synthesis*, 25th Anniversary Edn. The Belknap Press of Harvard University Press, Cambridge, MA
Young D 1994 *Die Entdeckung der Evolution*. Birkhäuser, Basel

B. König

Nature, Concepts of: Environmental and Ecological

1. Introduction

Concepts of environmental/ecological nature can be minimally understood as ideas about the physical world that accompany actual or potential sets of practices. Though scarcely if at all mentioned in the 1937 and 1968 editions of this encyclopedia, the contemporary significance of the topic in environmental/ecological analysis is a testimony to the increasing sophistication of the field as well as to the intellectual climate at the end of the twentieth century—in both cases, fostering an emphasis (some would say overemphasis) on the meaningful basis of human–environment relations and not solely their practical effects and amelioration.

The social and behavioral sciences have entered into these discussions from a variety of epistemological positions, especially in recognition of human-induced environmental change and the changing notions of nature underlying the natural sciences. Concepts of nature emerge as, implicitly or otherwise, concepts of nature-human relations (Williams 1980); ultimately, one of two broad positions—human distinctiveness from nature (the prevailing position), or human unity with nature—lies just beneath the surface. In addition, the unstated question '*Whose* concepts of nature?' must be acknowledged, pointing to the considerable potential variety among individuals, societies, and epochs, and thus to the limited though privileged voice of any writer seeking to provide a definitive treatment.

The word 'nature' itself is etymologically rich, suggesting its conceptual breadth over time. Nature comes from the Latin *natura*, which is derived from the verb 'to be born' (e.g., *natal* comes from the same root). There have been three different though interrelated and progressive senses of the English use of the word nature through time (Williams 1983). From the thirteenth century on, nature meant the essential quality or character of something, such as the nature of a person or of mortality. Beginning with the fourteenth century, the word was also used to represent the inherent force directing the world and human beings, as in 'the way of nature.' Not until the seventeenth century—relatively recently in English language usage—did the word nature also mean the physical world as a whole, the meaning adopted in this article.

Thus, historical usage suggests an important development from nature as quality or process to nature as an all-encompassing thing, with profound implications for the way both people ('human nature') and the biophysical world ('Nature') are viewed.

2. Concepts of Concepts of Nature

Thinking about concepts of nature raises several important philosophical and theoretical issues. Is, for instance, the whole project guilty of idealism by subtly imputing causal priority to consciousness versus practice? One example could be the thesis proposed by Lynn White (and debated endlessly thereafter) that

ideas inherent in Judeo-Christianity are the root of contemporary environmental crisis (White 1967). Or, as a related issue, does this project tend to assume the existence of an excessively homogeneous and rather simplified set of meanings among people in its general statements? And to what extent does it foster a politically naïve view of social relations by suggesting that concepts are largely a matter of free individual choice? Additionally, are concepts of nature of primarily evolutionary origin, as has been suggested with respect in the biophilia hypothesis (Kellert and Wilson 1993), or is their basis cultural? The list of potential issues goes on and on.

One issue in particular has dominated recent discussion and debate over concepts of nature: the question of their epistemological status. Do concepts of nature reflect, more or less faithfully, the real state of nature? Or are concepts of nature better understood as rooted in and constructed out of particular places, peoples and times? The two positions may not appear to be entirely contradictory, but they have served as opposing poles to which the bulk of recent intellectual positions have stubbornly clung, and indeed their respective abstract and practical commitments often differ markedly. The debate between realist and constructivist positions is beneficial because it forces practitioners of the environmental/ecological sciences to think carefully about the knowledge they generate, and specifically how to unify or harmonize knowledge related to human–environment interaction drawn from the natural sciences on one hand, and the social sciences on the other.

2.1 Realism and Constructivism

Realism is the default epistemological position on concepts of nature taken in lay life and the bulk of scholarly work. Realism involves the assertion that reality exists and is knowable. In terms of concepts of nature, most realists would argue that they can be evaluated in terms of their truth-content, and that in the most common and general account this truth-content can be derived from comparing truth-statements to reality to see whether or not they faithfully correspond (see *Realisms and their Opponents: Philosophical Aspects*; *Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects*). As applied to environmental issues, realism generally underscores efforts to separate 'ecofacts' from 'ecofiction' (Baarschers 1996) in developing a scientifically grounded basis for environmental policy.

Realism is so diffuse a premise in the physical and biological sciences and their public uptake as to sound obvious, without conceivable opposition. Yet the realist position has been challenged in the last few decades. The constructivist challenge lies in the combination of two recent movements in the social sciences (see *Theory: Sociological*): (a) an interest in meaning (vs. action or behavior alone) as an object of analysis and interpretation (vs. explanation or prediction) as a goal of analysis, and (b) a turn toward greater reflexivity and self-criticality. Combined, these movements have resulted in a desire among social scientists and humanists to examine the cultural and related political contexts of the truths underlying contemporary environmentalism, and to question the notion that natural science as the site of production of such truths stands aloof from these contexts.

Constructivism (also called social constructivism or constructionism) is not, as its detractors charge, a metaphysical position that reality only exists in the form of ideas in our heads, but rather an epistemological position asserting that there is no other way to make sense of this reality than by invoking ideas, and that these ideas have a significant though generally overlooked human (cultural, political, and so forth) dimension to them. Far from reflecting—more or less accurately—reality, knowledge and truth are thus constructed (hence the term constructivism) on this account, and differences in truth-statements arise not simply out of variably-accurate reflections of reality, but rather from different human contexts (interests, values, systems of meaning, and so forth). To a constructivist, the difference between ecofacts and ecofiction may indeed legitimately be a matter of opinion, as the rules designating the dividing-line are themselves constructed by people. In effect, the constructivist turns the realist question 'What is the truth about this environmental issue' into '*Whose* truths about this environmental issue are being portrayed as *the* truth?'—clearly a frustrating and deeply worrisome turn to the realist, for whom defense of truth plays a far purer motivational role.

2.2 The Debate and Possible Rapprochement

The terms of the debate between realists and constructivists have not been equal in numbers or clout, with realists generally playing the status quo, Goliath role, and constructivists the upstart, David role. Indeed, the very paucity of environmental publications on this debate underscores the implicit dominance of realism as a mode of understanding concepts of nature as derived from natural science and appropriated by the public and decision-makers. With beginnings in the 1980s and an outpouring of publications in the 1990s, scholars of a more constructivist bent set out to question the realist premises underscoring environmental science and environmental concern (e.g., Williams 1980, Wilson 1992, Bennett and Chaloupka 1993, Evernden 1993, Milton 1993, Simmons 1993, Cronon 1995). The arguments were varied, but at their heart was an understanding of environmental 'truths' so deeply woven into history, politics, and culture that

the whole project of environmentalism-founded-on-facts was challenged.

The response by environmental realists—the 'silent majority'— was for the most part disregard. Only a few publications (e.g., Soulé and Lease (1995) have been explicitly devoted to countering constructivism in the environmental arena. Yet, as the roots of constructivism lie outside of the environmental arena *per se*, so has the debate primarily taken place on terrain outside of environmental/ecological sciences, most notably revealing itself as the so-called 'Science Wars' battle (Gross et al. 1996, Ross 1996), in which environmental issues have been but one part of a much larger epistemological conflict over the status and privilege of science as a preferred method for obtaining truth. The argument of realists in these contexts has been to question what they believe are the relativistic or 'anything goes' implications of constructivism, and to reassert the primacy of reason as a guiding principle for the discovery and application of truth—a position that clearly has strong resonance in the environmental arena as well (Ehrlich 1996, Lewis 1996).

Although the premises of realism and constructivism seem quite distinct, their more naïve or strident versions share an important epistemological flaw in hyperseparating the object of knowledge from the knowing subject. Any attempt at gaining knowledge necessarily involves the interaction of subjects (e.g., scientists) and objects (e.g., biophysical nature). The debate between realism and constructivism reveals that concepts of nature have unfortunately suffered from accounts that deny or downplay this relationality, thus being too closely attached either to the object of knowledge, on the realist account, or the subject of knowledge, on the constructivist account.

Thankfully, concerns with subject/object dualism are diffuse, and have led some scholars to define forms of environmental realism and constructivism based more on a relational sense of subject/object epistemological interaction (Plumwood 1993, Hayles 1995, Soper 1995, Proctor 1998). Nonetheless, the institutionalized hegemony of objectivist approaches will no doubt encourage Davids to slay this Goliath rather than facilitate the much needed discussion between the two, working towards fashioning concepts of concepts of nature that satisfy the demands of relationality.

3. *Nature in the Era of Environmental Change*

The late twentieth century ushered fully into public consciousness an era of environmental change (Dunlap et al. 1993), increasingly at regional and global scales, with profound implications for popular concepts of nature and their influence by the burgeoning institution of science (see *Human–Environment Relationships*).

3.1 *From Plato to Princeton*

Popular concepts of environmental change build upon a long history of interest in the ways societies have conceived and transformed nature, and concern stretching back as far as Plato (Wall 1994). Clarence Glacken has traced three persistent ideas as they weave through several millennia of western history from the fifth century B.C.E. to the eighteenth century C.E.: the idea of a designed nature, of the influence of nature on culture and of culture on nature (Glacken 1967); the last of these three ideas links with current discussions over anthropogenic environmental change. Glacken argues that this idea long existed in embryonic form attached to the notion of a designed earth: the human race could be seen as fulfilling its 'God-given mission of finishing the creation, bringing order into nature, which God, in giving him mind, the eye, and the hand, had intended that he do' (p. viii). Yet it did not come into its own until relatively recently in western history: Glacken cites the first detailed treatment as George Perkins Marsh's *Man and Nature* (1864). From Marsh one can draw a direct link to the present, through the mid-1950s volume *Man's Role in Changing the Face of the Earth* (based on a 'Marsh festival' symposium at Princeton; see Thomas 1956). Its contemporary successor, affectionately known as ET, *The Earth as Transformed by Human Action* (Turner et al. 1990), was the first volume in which environmental change was systematically studied at regional and global scales, and represents a transition between the intellectual concerns culminating in Princeton and the concerns of recent global change science.

This increasing scholarly emphasis dating from the mid-nineteenth century onwards mirrors increasing rates of transformation of nature: fully half to nearly all of the extent of major forms of environmental change since the beginnings of human history have occurred in this relatively short period (Kates et al. 1990). Indeed, the 'domination of nature' thesis elaborated in the mid-twentieth century by critical theorists (Horkheimer and Adorno 1972, Leiss 1972), and recently expanded by ecofeminists, social ecologists, and others, could only have arisen in such an era of rampant industrial transformation of nature. G. P. Marsh's masterful account reminded his sanguine nineteenth-century contemporaries of ancient ravages upon the biophysical world, as well as the dangers inherent in more recent proposed and completed anthropogenic transformations (Marsh 1864). In the face of this tremendous human impact, however, Marsh's moral sensibility was complex; ultimately, Marsh ended with the rhetorical question 'whether man is of nature or above her,' though the text suggests his mind had long been made up in favor of human distinctiveness from nature. As with Marsh, the Princeton symposium featured a paired celebration and condemnation of the human transformation of

nature, with final essays betraying a strong flavor of concern. From Plato to Princeton, then, accounts of the human transformation of nature were straightforwardly normative as well as scientifically descriptive, with the role and accountability of humans front and center.

3.2 Global Change

Though interest and concern over the impact of humans on nature has grown in recent decades, the bulk of scholarly research has largely displaced humans from the center of the problematic in favor of earth processes, resulting not so much in a fuller understanding of the human role in transforming the earth as a more complex account of the earth system which long precedes, and dynamically responds to, human perturbations. Physical (and, to a lesser and more recent extent, life) scientists have played a major role. Atmospheric research in particular has, for political and scientific reasons, been central, beginning with the International Geophysical Year of 1957–8 (Fleagle 1994). More recently, the International Geosphere-Biosphere Program (IGBP) was created to synthesize information and promote systematic integration of models of earth processes.

Yet climate change has received primary emphasis: the US Global Change Research Program (USGCRP), for example, was launched in 1989, with roughly two-thirds of its annual funding emanating from the National Aeronautics and Space Administration (NASA) and the same amount devoted to space-based data gathering and data management, as compared to less than 10 percent of the USGCRP budget devoted to social science dimensions. Indeed, the bulk of 'human dimensions of global change' research has emphasized the human *response* to global environmental (primarily climate) change, and most studies which have included humans as causes of environmental change greatly simplify their social dynamics, modeling them primarily as biophysical agents. The tremendous influence of *Climate Change: The IPCC Scientific Assessment* (Houghton et al. 1990), published around a fever-pitch period of scientific predictions and policy pronouncements on the specter of global warming, further confirmed this bias.

The distance between Marsh and global change research is unmistakable, but it is not merely one of increased scientific sophistication. Also at work is a mode of understanding environmental change—a concept of nature and nature–society relations—where moral complexities yield to biophysical complexities (which have in many cases yielded to climatic complexities!), at ever-greater scales of abstraction from specific human social contexts and landscapes. Brian Wynne has noted (1994 p. 171):

> As the geopolitical reach of environmental science has become more and more expansive, its intellectual temper has become more reductionist … Whereas [the 1987 Brundtland Commission] articulated a basic political, moral and social framework from which to define policies for environmentally sustainable global development … IPCC began from a scientific origin—defining and managing a sustainable climate—from which should be derived the necessary social, economic, and other policies for survival.

3.3 Evaluation: Nature vs. Culture

Have human impacts on nature been good or bad? Concepts of nature concern not only what nature and nature–society relations *are* but what they *ought to be*, and it would be a mistake to infer from the above that moral questions have vanished from the temporal horizon. Yet, as with the global aspirations of environmental change research, they have taken on a particular scale. As Anne Buttimer argues (1993 p. 216):

> With [the predominance of rational-scientific inquiry] has come a peculiar paradox: a penchant for describing the earth in literal, materialist, and reductionist terms, on the one hand, yet a penchant for a totalizing, generalizing approach to normative action on the other.

Perhaps sweeping normative generalizations about nature are not new, as Marsh's *Man and Nature* suggests; yet simply considering the prominence of planetary-level concern in major events (Earth Day; the UN 'Earth Summit' of 1992) and popular publications (e.g., *Earth in the Balance* [Gore 1993], *Healing the Planet* [Ehrlich and Ehrlich 1991], and countless others), it is clear that these generalizations are increasingly broad in their audience and spatial scale. The desire to 'think globally' is laudable; what is more problematic is the gulf between nature and culture implied in the process by which people have come to think globally. The popularization of global environmental change research in policy issues such as global warming has led to a predominantly nature-based point of departure in understanding and evaluating human impacts marked by a good deal of doomsaying, which in turn has fostered a counter-moment of cornucopians championing human ingenuity (see *Cassandra/Cornucopian Debate*).

Similarly, environmental ethicists have struggled over whether concern for nature ought to be rooted in values connected to culture (anthropocentrism) or nature (biocentrism and ecocentrism), with similar doses of skepticism of one for the other (Rolston 1988, Callicott 1989, Norton 1991, Weston 1994). Still others such as deep ecologists have challenged this hyperseparation of nature and culture, replacing it with a monist sense of Self as including but moving far beyond the individual person to encompass all of nature—a 'solution' to the problem of nature/culture dualism most will find, practically-speaking, unintelligible (see *Ecology*, *Deep*). It is apparent that the nature–culture dualism implicit in this era of en-

vironmental change has led to conceptual troubles among those who have attempted to define an alternative to human distinctiveness vis-à-vis nature.

4. Nature and Metaphor in Natural Science

The reality of environmental change and the role of scientists in explicating it to the public have given them a particularly influential voice in defining nature. Yet the very metaphors that underscore research on nature in physical and biological sciences have increasingly been challenged by scientists themselves, with important potential implications for how people will understand nature in future.

The Greek root of metaphor means to transfer or carry, and suggests how a metaphor transfers a complex reality onto a simple, meaningful substrate. Metaphor is the inescapable way of viewing the world *as if* it were a this or a that; particular metaphors convey certain benefits as well as liabilities. Metaphor is the *a priori* ground upon which facts are assembled; to assert the metaphorical basis of facts is not to deny the possibility of interrogating their truth-status as much as to admit their dependence upon simple (metaphorical) truths for interpretation.

4.1 The Mainstay: Mechanism

Mechanism has, by all accounts, been the dominant metaphor for nature over the last three centuries (Merchant 1980, Buttimer 1993). Mechanism is, quite simply, a way of understanding the complexities of the biophysical world by thinking of this world as if it were a machine or mechanical system. The rise of mechanistic thinking has generally been attributed to Cartesian metaphysics, Newtonian physics, and parallel movements in the development of science, rationality and nature-society relations during the seventeenth century and afterward. Mechanism as a metaphor for nature has had many significant impacts—in fact, it is closely allied with the subject/object and culture/nature divides noted above. One of key importance to environmental/ecological science concerns implications for determinism and control: mechanism presumes that nature, like a machine, follows strict deterministic rules, and that to the extent that scientists can uncover these rules nature can be controlled or managed predictably. In short, mechanism paints a picture of nature as subject to manipulable necessity versus capricious chance, a quite fundamental premise for science, and more particularly its application to engineering and technology.

4.2 New Metaphors?

As with any metaphor, mechanism has conveyed both benefits and liabilities. With the tremendous power of technology in controlling and harnessing nature for human ends has come a whole host of goods and evils, as requires little recounting here (see *Environmental Sciences*). In particular, the mechanistic assumption that nature is more a creature of necessity than chance has increasingly been challenged by the failures and surprises of its technological applications (McPhee 1989).

Though natural scientists have by and large been comfortable with adopting the language of mechanism in their work, they have increasingly been concerned about its overly deterministic impulse, a confidence in completely explaining nature and deriving predictive capacity that sounds more resonant with the heady days of science and technology in the early to mid-twentieth century than the present. Some of the most important debates among ecologists in the twentieth century, in fact, betray the tension between necessity and chance, as in that between the Clementsian notion of tightly-coupled communities of species governed by equilibrium and the Gleasonian notion of these communities as loose affiliations of individual species which simply happen to be together at a particular place and time (Barbour 1995).

In recent years ecology seems to be moving from mechanism and its steady-state models toward an appreciation of nature's inherent dynamism and surprise (Botkin 1990). More broadly, Robert Ulanowicz (1997) has proposed to replace mechanistic determinism with a probabilistic view (derived in part from Karl Popper) that suggests how strict adherence to either necessity or chance denies the ways in which ecosystems develop certain consistencies or order out of chaos. Many natural scientists concerned with modifying determinism have been inspired in part by complexity theory, which has demonstrated the potential asymmetry between explanation and prediction in that simple rules can generate unpredictable behavior ('determinate complexity') in ecological systems (May 1973). Complexity theory has been extended to particular concerns in the ecological sciences, such as questions related to species diversity and conservation (Pimm 1991; Levin 1999), as a manner of reframing mechanistic determinism and its implications for predictability and control.

On the popular front, mechanism has not so much been regarded as a metaphor as an increasingly indefensible 'truth' about nature that must be supplanted by a truer metaphor. Indeed, mechanism has never fully displaced other metaphors in the biological sciences, such as the notion of nature as organism that has inspired the 'Arcadian' tradition of a holistic, spiritual, anti-technological ecology (Worster 1977). Organicism has increasingly been presented as the perfect alternative—witness, for instance, the enormous popularity of the notion of the earth as a self-regulating 'organism' or Gaia (see *Gaia*). Yet a more thoroughgoing animistic (i.e., nonmechanistic, though not necessarily organismically integrated) notion of

nature would necessitate a fundamentally different way of knowing nature—primarily though encounter, 'feel,' and empathy—and a whole different derivative language, with its own set of inherent strengths and liabilities, the latter of which are unmentioned among its champions (e.g., Abram 1996). In addition, advocates of the need to move beyond mechanism frequently adopt mechanistic language to describe their position: for instance, Botkin states 'We can leave behind the metaphors of the machine … and can arrive … at a new organic view of the Earth, a view in which we are part of a living and changing system whose changes we can accept, use, and control' (Botkin 1990 p. 189). Just as necessity and chance are both important ingredients in understanding and living with nonhuman nature, the metaphor and language of mechanism can never be fully supplanted by some nonmechanistic alternative. Ultimately, concepts of nature that derive from more sophisticated interweavings of necessity and chance than either classic mechanism or its supposed alternatives will afford better understandings of biophysical reality.

5. *Conclusion*

What has been covered above is admittedly limited. There will be other debates around other dualisms and derivative boundaries related to concepts of nature, such as the distinction between what is natural versus artificial in both the biophysical and human domains (Robertson et al. 1996), especially given the implications of genetic research and technology. These divisions are not unrelated: all derive in part from the question of the status of humans vis-à-vis nature. The prevalent notion of human distinctiveness from nature has fed the hyperseparation between nature and culture, between nature as object of knowledge and humans as knowing subject, between nature as guided by brute necessity and humans as guided by free will. These schisms have accompanied the human transformation of nature, most scientific and popular ideas of nature, and indeed even the problematic division between the natural and social sciences. The alternative notion of human unity with nature, arising in cases such as deep ecology, biophilia, ecocentrism, Arcadian ecology, and Gaia theory, are limited in their own right as noted above. In some ways, however, both human distinctiveness from, and unity with, nature are defensible positions; the choice is thus less a clear either/or than a paradoxical both/and.

It is unlikely that epistemological debates over concepts of nature and conflicts over the human transformation of nature will soon end; nor will the emerging concepts of nature in the natural sciences go unnoticed. These issues will matter greatly for the practice and uptake of environmental/ecological science. Concepts of nature are not mere epiphenomena in the social scientific analysis of the environment; they matter through and through. Thus it also matters to formulate better concepts of nature than those that have predominated over the last few centuries. In this regard, the primary impediment encountered is the influence of pervasive and persistent dualisms—dualisms which are observed not only in defense of status quo scientific and popular concepts of nature but in opposition as well, either by flip-flopping from one pole to the next or by implicitly adopting the language of one pole in defense of the other. Yet dualism is not going to be replaced by some grand monistic synthesis as much as a more diffuse recognition of the historically-embedded asymmetries and hyperseparations of subject and object, culture and nature, and necessity and non-necessity that have largely precluded more considerate, relational, and ultimately reflexive accounts of nature.

See also: Critical Realism in Geography; Environmentalism: Philosophical Aspects; Nature–Society in Geography; Social Constructivism

Bibliography

Abram D 1996 *The Spell of the Sensuous: Perception and Language in a More-Than-Human World*, 1st edn. Pantheon Books, New York

Baarschers W H 1996 *Eco-facts and Eco-fiction*. Routledge, London

Barbour M G 1995 Ecological fragmentation in the fifties. In: Cronon W (ed.) *Uncommon Ground: Toward Reinventing Nature*, 1st edn. W.W. Norton and Company, New York, pp. 233–55

Bennett J, Chaloupka W (eds.) 1993 *In the Nature of Things: Language, Politics, and the Environment*. University of Minnesota Press, Minneapolis

Botkin D B 1990 *Discordant Harmonies: A New Ecology for the Twenty-first Century*. Oxford University Press, New York

Buttimer A 1993 *Geography and the Human Spirit*. The Johns Hopkins Press, Baltimore, MD

Callicott J B 1989 *In Defense of the Land Ethic: Essays in Environmental Philosophy*. State University of New York Press, Albany, New York

Cronon W (ed.) 1995 *Uncommon Ground: Toward Reinventing Nature*, 1st edn. W. W. Norton & Company, New York

Dunlap R E, Gallup Jr. G H, Gallup A M 1993 *Health of the Planet: Results of a 1992 International Environmental Opinion Survey of Citizens in 24 Nations*. The George H. Gallup International Institute, Princeton, NJ

Ehrlich P R, Ehrlich A H 1991 *Healing the Planet: Strategies for Resolving the Environmental Crisis*. Addison-Wesley Publishing Co. Inc., Reading, MA

Ehrlich P R, Ehrlich A H 1996 *Betrayal of Science and Reason: How Anti-environmental Rhetoric Threatens our Future*. Island Press, Washington, DC

Evernden N 1993 *The Social Creation of Nature*. The Johns Hopkins University Press, Baltimore, MD

Fleagle R G 1994 *Global Environmental Change: Interactions of Science, Policy, and Politics in the United States*. Praeger, Westport, CT

Glacken, C J 1967 *Traces on the Rhodian Shore: Nature and Culture in Western Thought from Ancient Times to the End of*

the Eighteenth Century. University of California Press, Berkeley, CA
Gore, A 1993 *Earth in the Balance: Ecology and the Human Spirit*. Plume Books, New York
Gross P R, Levitt N, Lewis M W 1996 *The Flight from Science and Reason*. New York Academy of Sciences, New York
Hayles, N K 1995 Searching for common ground. In: M E Soulé, G Lease (eds.) *Reinventing Nature? Responses to Postmodern Deconstruction*. Island Press, Washington, DC, pp. 47–63
Horkheimer M, Adorno T W 1972 *The Dialectic of Enlightenment*. Herder and Herder, New York
Houghton J T, Jenkins G J, Ephraums J J (eds.) 1990 *Climate Change: The IPCC Scientific Assessment*. Cambridge University Press, Cambridge, UK
Kates R W, Turner B L II, Clark W C 1990 The great transformation. In: B L Turner II, W C Clark, R W Kates, J F Richards, J T Mathews, W B Meyer (eds.) *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere over the Past 300 Years*. Cambridge University Press, Cambridge, UK, pp. 1–17
Kellert S R, Wilson E O 1993 *The Biophilia Hypothesis*. Island Press, Washington, DC
Leiss W 1972 *The Domination of Nature*. George Braziller, New York
Levin S A 1999 *Fragile Dominion: Complexity and the Commons*. Perseus Books, Reading, MA
Lewis M W 1996 Radical environmental philosophy and the assault on reason. In: P R Gross, N Levitt, M W Lewis (eds.) *The Flight from Science and Reason*. The New York Academy of Sciences, New York, pp. 209–30
Marsh, G P 1864 *Man and Nature; or Physical Geography as Modified by Human Action*. Scribner and Co., New York
May R M 1973 *Stability and Complexity in Model Ecosystems*. Princeton University Press, Princeton, NJ
McPhee J 1989 *The Control of Nature*, 1st edn. Farrar Straus Giroux, New York
Merchant, C 1980 *The Death of Nature: Women, Ecology, and the Scientific Revolution*, 1st edn. Harper & Row, San Francisco
Milton K (ed.) 1993 *Environmentalism: The View from Anthropology*. Routledge, London
Norton, B G 1991 *Toward Unity Among Environmentalists*. Oxford University Press, New York
Pimm S L 1991 *The Balance of Nature? Ecological Issues in the Conservation of Species and Communities*. The University of Chicago Press, Chicago
Plumwood V 1993 *Feminism and the Mastery of Nature*. Routledge, London
Proctor J D 1998 The social construction of nature: relativist accusations, pragmatist and critical realist responses. *Annals of the Association of American Geographers* **88**(3): 352–76
Robertson G, Mash M, Tickner L, Bird J, Curtis B, Putnam T (eds.) 1996 *Future Natural: Nature, Science, Culture*. Routledge, London
Rolston III H 1988 *Environmental Ethics: Duties To and Values in the Natural World*. Temple University Press, Philadelphia, PA
Ross A (ed.) 1996 *Science Wars*. Duke University Press, Durham, North Carolina
Simmons I G 1993 *Interpreting Nature: Cultural Constructions of the Environment*. Routledge, London
Soper K 1995 *What is Nature? Culture, Politics and the Non-human*. Blackwell Publishers Ltd., Oxford, UK
Soulé M E, Lease G (eds.) 1995 *Reinventing Nature? Responses to Postmodern Deconstruction*. Island Press, Washington, DC
Thomas W L (ed.) 1956 *Man's Role in Changing the Face of the Earth*. The University of Chicago Press, Chicago
Turner II B L, Clark W C, Kates R W, Richards J F, Mathews J T, Meyer W B (eds.) 1990 *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere Over the Past 300 Years*. Cambridge University Press, Cambridge, UK
Ulanowicz R E 1997 *Ecology, The Ascendent Perspective*. Columbia University Press, New York
Wall D (ed.) 1994 *Green History: A Reader in Environmental Literature, Philosophy and Politics*. Routledge, London
Weston A 1994 *Back to Earth: Tomorrow's Environmentalism*. Temple University Press, Philadelphia, PA
White Jr L 1967 The historic roots of our ecologic crisis. *Science* **155**(March): 1203–7
Williams R 1980 Ideas of nature. In: Williams R (ed.) *Problems in Materialism and Culture*. Verso Press, London, pp. 67–85
1983 *Keywords: A Vocabulary of Culture and Society*. Oxford University Press, New York
Wilson A 1992 *The Culture of Nature*. Blackwell, Cambridge, UK
Worster D 1977 *Nature's Economy: A History of Ecological Ideas*. Cambridge University Press, Cambridge, UK
Wynne B 1994 Scientific knowledge and the global environment. In: Redclift M, Benton T (eds.) *Social Theory and the Global Environment*. Routledge, London, pp. 169–89

J. D. Proctor

Nature–Society in Geography

Nature–society or human–environment relationships have been part of geography since antiquity. The modern foundation of their study in geography, however, was established in nineteenth century Germany. The themes of human impact on and adjustment to the physical environment were articulated by German geographers, with various claims that these relationships, broadly interpreted, constituted the identity of the discipline. For a brief period in the early twentieth century, a particular definition of the relationship (environmental determinism) formally dominated geographic education in the USA. The association of determinism with geography relegated nature–society studies to the margins of the discipline. By the middle of the century a spatial–chorological identity for geography was largely unchallenged, despite various alternative nature–society visions to determinism. By the late 1970s, however, nature–society studies had returned to geography in a significant way, building from interests that helped to give rise to cultural ecology and risk-hazard studies as well as several other subfields of study. Entering the twenty-first century, nature–society geography has grown signifi-

cantly in the number of its practitioners and research influence, exploring the range of perspectives found in the human sciences and humanities. Its rebirth, which coincides with renewed public interests in the nature–society condition, resurrects the various identity issues raised in Germany in the nineteenth century and challenges the discipline to situate its spatial–chorological and nature–society visions equitably.

1. Roots of the Theme in Geography

Ancient geography (Greek *geographia*, meaning writing about the earth) involved descriptions of the earth, serving to compile knowledge of places not experienced in everyday life and as found in travel and exploration. All places contain 'nature,' and most are occupied or used in some capacity by our species. It is not surprising, therefore, that 'complete' geographies documented the human and environmental character of places, at least as this understanding could be derived from observing places (Glacken 1967). This substantive interest, in various intellectual forms, remains throughout the history of geography, although it continually struggles with other identities of the discipline in the search for its intellectual niche.

2. German Foundations of Modern Geography

The nineteenth-century organization of knowledge within European academe shaped modern geography. This reshaping took place at time in which the formal study of geography was especially vibrant in Germany, and the competing views expounded by German scholars during that century remain essential, if contested, elements of the discipline's identity in the twenty-first century.

2.1 The Spatial–Chorological Science

Immanuel Kant (1724–1803) laid the foundation for a spatial or chorological identity: geography provided integrative understanding through a focus on the spatial attributes and relations of phenomena as they were arranged in areas, regions, or space (Tatham 1951, p. 40). This vision was elaborated in three ways by Karl Ritter (1779–1859), Ferdinand von Richthofen (1833–1905), and Alfred Hettner (1859–1941), among others: geography was the science of areal or regional differentiation (and, hence, the need to bound or demarcate areas to inform understanding), the history and particularities of places (chorography), and distribution studies. These various identities, collectively labeled chorology, tended to stress geography as a discipline based largely on a shared approach to study, more so than its substance. This 'unique' quality, shared with history, the chronological science according to Kant, marked geography as an exception to the 'systematic' disciplines. This status is one with which geography has consistently struggled.

2.2 The Nature–Society Science

Building on the natural histories of Comte de Buffon (1707–88), Johann R. Forster (1729–98), and Johann G. A. Forster (1754–95), an alternative identity of the discipline was championed by other nineteenth-century geographers, if not always explicitly. Alexander von Humboldt (1769–1859) set the stage for this vision of geography through such works as *Cosmos* (*Sketch of a Physical Description of the Universe*), which documented the 'unity in diversity' in nature, or how diverse phenomena, including humans, give rise to ordered, functioning landscapes (akin to the modern use of landscape in ecology). Humboldt was not especially interested in defining geography or articulating approaches within it, but he recognized three divisions of knowledge about nature, one of which he called geognosy (earth science) and which dealt with the spatial distribution and relationships of phenomena (Dickinson 1969, p. 24). This recognition and his research approach, which involved bounding the landscape (creating an areal unit) in question, led some to the conclusion that Humboldt's vision of geography was consistent with chorology. This conclusion, however, trumpets the spatial qualities of his approach over the substance of his interest, landscape order and function.

Unlike Humboldt, Frederich Ratzel (1844–1904) was keenly interested in the identity of geography and championed the view that the discipline should be a systematic science, complete with a phenomenon of study. This phenomenon was the nature–society relationship and constituted the identity of the relationship (Dickinson 1969, p. 65, Tatham 1951, p. 64). In *Anthropogeography: Introduction to the Application of Geography to History* (1882), Ratzel advanced a science of reciprocal relationships between nature and society in which the distribution of cultures (his students employed 'cultural area' before the term emerged in anthropology) was predicated on a Darwinian-influenced notion of the capacity of different societies to adapt to their physical surroundings, also referred to as landscape. Ratzel focused on human adjustments to and manipulation of natural resources and environmental conditions, although subsequently others would reverse this orientation.

Other German geographers of the time followed partially, if not fully, the Ratzelian view. Otto Schlüter (1872–1952) found fault with the environmental orientation in the emerging nature–society thought (Dickinson 1969, p. 126), but saw the scope of geography as understanding the 'visible' landscape, not as a bounded area but as the nature–society relationships shaping it (Livingstone 1993, p. 264). Albrecht Penck (1858–1945), among others, found

merit in this view, favoring a landscape-led geography, and his meaning of landscape appeared to be closer to Schlüter's than to that of an areal unit (Dickinson 1969, p. 110). From this point onwards, landscape would hold different meanings to different geographers. Regardless of the specific definition used, the landscape concept invariably implied a collection of organic and inorganic phenomena whose interrelationships gave rise to some meaningful areal identity on the earth's surface, be it a coastline, system of cultivation, or urban settlement.

2.3 Ties Between the Two Views

In spite of a nonunified identity for geography, the German founders of the modern discipline linked chorology and nature–society. The nature–society proponents did so by focusing their relationships on landscapes, replete with its various meanings, which required defined areal units. Exemplary of these ties was the Ratzel-influenced French geography of the time. Vidal de la Blache (1845–1918) and, to a lesser extent, his student, Jeans Brunhes (1869–1930), organized the French discipline around the nature–society relationship, emphasizing a 'possibilist' position toward it—human choices to environmental settings. The French approach, however, focused on these relationships as they were manifested in small areas and reported on them through regional monographs. In this way the relationship in question was tied to area or region, akin to the landscape. German chorologists, in contrast, invariably called for the examination of nature–society conditions as part of the study of the place, area, or region. Ritter's own work, *Erdkunde*, is exemplary, as registered in the subtitle (*Earth Science: in Relation to Nature and the History of Man*) and in the text: 'The earth and its inhabitants stand in the closest reciprocal relations, and one cannot be truly presented in all its relationships without the other' (in Tatham 1951, p. 44).

3. The Geographic Influence or Geographic Factor

The nature–society position vied for the sole identity of the discipline as it emerged in the USA during the early part of the twentieth century. There, geography's birth was profoundly influenced by William Morris Davis (1850–1934), a geomorphologist of no small repute, who created a geography program within geology at Harvard and was the force behind the creation of the Association of American Geographers (1904), serving as its President three times. He increasingly came to view geography as the study of inorganic (nature) controls over the organic (society). In his 1905 Presidential Address to the Association of American Geographers, he not only made the case for the study of this relationship as the defining element of geography, but also was totally dismissive of the chorological position, disposing of it in two brief sentences.

> There is so little support for this narrow view of the subject [location] to be found in modern geographical books that it need not be further considered [and] … if geography were only the science of distribution, that is, the regional aspect of other subjects, it would be hardly worth while to maintain the study of geography apart from that of the subject whose regional aspect it considers (in Johnson 1954, p. 9).

The strength of Germanic influences on Davis is unclear. He referenced and critiqued German geographers, but his views of geography may have been more strongly affected by the Darwinian and neo-Lamarckian fervor of the time and the search for organic analogy (Livingstone 1993, pp. 203–5). His position was shared, even anticipated, by a number of highly productive geographers who were well versed with the German traditions, none more important than Ellen Churchill Semple (1863–1932) of Clark University and University of Chicago. Mentored by Ratzel, she became enamored with the theme of *geographic influence* or the *geographic factor* as articulated in *Influences of Geographic Environment* (1911). She was joined in this vision in the USA by such luminaries as Ellsworth Huntington (1876–1947) at Yale, a prolific author who sought to understand the history of regions through the lens of climate (e.g., *Civilization and Climate*, 1915), and abroad by Griffith Taylor (1880–1963), among others, who sought to explicate the controlling constraints of environment on human occupation of different lands (e.g., *Australia, a Study of Warm Environments and Their Effect on British Settlement*, 1941).

Taken as a whole, this group of geographers paid little, if any, attention to alternative or linked views of the identity of the discipline. Enhanced by the elite private institutions in which they were positioned, their moment of power penetrated the very foundations of American education. The Conference on Geography adopted the Davisian identity of geography, which in turn was accepted by the Committee of Ten of the National Educational Association (1892), whose role it was to define precollege educational programs and college entrance requirements (James and Martin 1981, p. 290). The discipline was thus defined as the study of inorganic controls over the organic, or the geographic influence/factor, and its K-12 textbooks were organized accordingly.

While these influences and events propelled nature–society geography to the forefront of the American educational academy, the moment was short lived. The geographic factor narrowed the nature–society relationship to the search for the causal power of nature over society—an extreme form of human adjustment or paucity thereof to nature. Alternative views of the relationship, such as the possibilism of the French geographers or by the American, George

Perkins Marsh (*Man and Nature or, Physical Geography as Modified by Human Action*, 1864) on the environmental impacts of human activity were unheeded. The geographic factor became *environmental determinism* which, in its lay forms, not only led to unfounded simplifications about nature's influence on humankind but also were drawn upon as support for social and political views proposing the superiority of northern environments and peoples over all others.

The extremes of determinism devastated geography and seriously wounded the nature–society position. The stigma attached to the geographic factor was so strong that the nature–society position was relegated to margins of the discipline by the 1940s and was even dismissed by some historians of the field (Livingstone 1992, p. 6). For much of the middle of the twentieth century, the central intellectual debate in the discipline centered on the kind of spatial–chorological science geography should be: a place-based and areal differentiation science (Hartshorne 1961), or a science of spatial relationships (National Research Council (U.S.) 1965). By the 1960s, the latter focus gained intellectual hegemony among the leadership of UK and US geography, remaining in power for two decades before it succumbed to challenges that reinserted the role of the contingencies of place and regions (see *Geography*). These challenges notwithstanding, geography's identity remains linked primarily to the 'why of where' question and much of human geography dismisses nature–society studies or sees them as falling outside the 'human' domain of the discipline's study (see *Geography*).

4. *Alternatives to the Geographic Influence*

If the 'battle' for the identity of geography were lost by proponents of the geographic influence, strong geographic interest in nature–society themes was not. At least two alternatives to environmental determinism arose before its collapse, each providing the foundation for contemporary nature–society studies in geography. Carl O. Sauer (1889–1975) sketched his alternative vision in his 1925 essay, 'The Morphology of the Landscape.' Inspired by German scholarship, Sauer sought to bring the discipline back to the landscape (an area of interrelated phenomenon), whose qualities as a whole (form, structure, and function) constituted a reality not captured by the constituent parts separately (in Leighly 1963, p. 321). The landscape was a nature–society relationship, and the morphology was a method of synthesis, a 'massing and ordering phenomena as forms … integrated into structures' (p. 326). He argued that there was 'a strictly geographic way' of entertaining these structures, and it was through the material imprints that society made upon area (p. 326). Thus Sauer's nature–society relationship was the human impact on the 'physical whole' which created the landscape.

Whatever Sauer's views on the ties between nature–society and spatial–chorological geography, he and the research prodigy of the 'Berkeley school' produced a large number of works focused on the human use of and impress upon the earth, including Sauer's own role in *Man's Role in Changing the Face of the Earth* (Thomas 1956). Exemplary was Philip Wagner's *The Human Use of Earth: an Examination of the Interactions Between Man and His Physical Environment* (1960), which provides only passing references to alternative identities of geography before exploring a discipline devoted to the human-made (artificial) landscape. Such nature–society interests were taken up outside the USA under the rubric of cultural and/or historical and geography (e.g., Darby 1940, Williams 1974).

Harlan Barrows (1877–1960) laid the foundation for the other nature–society theme in his 1922 Presidential Address to Association of American Geographers. He proposed geography as the study of human ecology or the adjustment that humankind makes to the occupied environment (Barrows 1923). Here was an explicit statement that nature–society relationships were the foundation of discipline, and that this foundation need not be deterministic. Barrows' call had no immediate discernible impacts, but some 30 years later his former student, Gilbert F. White (1911–), returned to Chicago to pioneer risk-hazards geography. White was captured by the service role that academic research could provide for solving social problems and was not particularly responsive to academic debates about the discipline's definition. Nevertheless, the defining quality of his interests and the geographic work he advanced was a form of nature–society relationship, specifically human adjustments to environmental perturbations, articulated in Ian Burton, Robert W. Kates, and Gilbert F. White, *The Environment as Hazard* (1978).

Thus geography entered the 1960s with two strong clusters of research on nature–society themes: human use and impact on nature, and human responses and adjustment to environmental change and perturbations. Arguably, a large number, perhaps even a majority, of these practitioners were drawn to geography primarily out of its nature–society traditions and the paucity of them in most other disciplines (Kates 1987). For the most part, however, the nature–society practitioners were content to do their work and made minimal attempts to challenge the identity of geography as a chorological science.

5. *Contemporary Nature–Society Subfields in Geography*

By the 1970s, nature–society research in geography was prospering, but under newly emerging and/or expanding subfields to which geographic organizations were partitioning themselves. These subfields

were initially stimulated by systems and ecosystems approaches combined with theory-driven, behavioral analysis (Gaile and Willmott 1989, Turner 1997). The *environmental hazard* subfield was in full swing, having taken on virtually every type of environmental perturbation to human occupation and use of the environment, from drought to volcanic eruptions. Research began to expand into non-Western world contexts, and a search for generic templates of human responses to such perturbations was under way. Hazards research spawned risk assessment and technological hazards and, more recently, risk studies have led to vulnerability assessments (e.g., see *Environmental Risk and Hazards*; *Environmental Vulnerability*). A complementary subfield, *environmental perception and behavioral geography*, moved towards issues of the bounding the human–environment experience and its meaning for hazard adjustments and landed resource use more generally.

The *cultural ecology* subfield in geography arose from the cross-fertilization of practitioners whose legacy may be seen in the substantive interests of the landscape but with the goals and aims articulated by Harold C. Brookfield in Australia on household and community behavior and Karl W. Butzer in the USA (via Canada and Germany) on systems (e.g., see *Ecology, Cultural*). Increasingly, the work of cultural ecologists in geography and anthropology blurred as attested in work by, for example, Philip Porter and Robert McC. Netting or William M. Denevan and Donald Lathorp

By the 1980s, both environmental hazard and cultural ecology were strongly affected by Marxist and critical approaches, and in the 1990s by social theory broadly conceived, with various impacts on the two subfields. *Political ecology* emerged (Blaikie and Brookfield 1987), directing attention to local nature–society conditions as a product of social differentiation (e.g., power, wealth, class, gender) and of linkages to distant decisions and processes. Many proponents were interested in both the drivers and responses to environmental change and linking the two (Watts 1983), dissolving some of the distinction that existed between cultural ecology and risk-hazard research. In addition, political ecology research in its various renditions broke down the partitions in nature–society research between 'pure' and 'applied' work (e.g., see *Feminist Political Ecology*; *Ecology, Political*).

Contemporary agriculture and rural land use and *water resource* subfields emerged to tackle contemporary nature–society problems, primarily in the developed world (e.g., see *Water Resources*). Both experienced similar intellectual influences that gave rise to political ecology. The newest nature–society subfield in geography, the *human dimensions of global environmental change*, emerged in the 1990s as a hybrid of the other subfields—one that recognized the new interdisciplinary research efforts on this theme globally and the strong influence of nature–society geographers on the development of this theme. This subfield, if you will, returns to the contemporary versions of the problems registered in the early 1800s by Forster, Humboldt, and Marsh (Turner et al. 1990, O'Riordan 1997).

Mirroring the cacophony of perspectives playing among the human sciences and humanities, nature–society practitioners have explored a wide range of frameworks for understanding (Turner 1997): post-positivistic science to critical and social theory to the search for meaning in the symbolic landscape (e.g., see *Cultural Landscape in Environmental Studies*). An appreciation for such diversity and the ability to move among the various frameworks have long been hallmarks of geographic work. During these episodes of diverse explorations, however, nature–society geography has tended to lose sight of its substance of study and focus instead on the explanatory form or theory invariably drawn from elsewhere. In such moments, the framework or theory is used, almost uncritically, to explain a nature–society relationship, rather than exploring the insights gained from the relationship to inform and restructure the explanation.

Much nature–society geography has now refocused, informing concepts and problems in the way that its intellectual ancestors might have imagined, if not the kinds of questions and issues addressed. That work, embedded firmly in the social and behavioral sciences, is strongly empirical, and increasingly quantitative and focused on the material relationships between humankind and the physical environment, be it human-induced land degradation or equity in opportunities to respond to it. Explanatory models seek to treat more equitably than before actor-based and structural-based approaches, and consider the roles of idiosyncratic agents and historicity (path dependence). Additionally, the environment as an active force within the relationship has been reinstated (Zimmerer and Young 1998), albeit examined through such themes as complexity and co-evolution. Finally, the former barriers between human use and impact studies and human response and adjustment studies have all but disappeared. Nature–society practitioners must now take care not to build new barriers through the blinders of favored perspectives, but to seize the moment of 'hybrid explanatory vigor' to formulate improved understanding of the nature–society relationship.

6. *Geography as Nature–Society Science?*

Do these trends suggest that geography is poised to redirect its identity to nature–society relationships? Probably not, although Stoddart (1987) came very close to making such an appeal, and Peet (1998) argues for a variant of such a move. To adopt this identity, retain the spatial–chorological one, or, better, unite the two, requires resolution of several intellectual and pragmatic considerations.

6.1 What Other Disciplines Concede

May (1970, p. 186) noted that no other discipline claims the nature–society relationship and most concede the subject to geography. Indeed, three of five texts granted the status as paradigmatic shifts in geographic practice by Harvey and Holly (1981) belong to Ratzel, Vidal, and Sauer, all of which can be interpreted as nature–society visions as much as, if not more so, than spatial–chorological visions. May's observations, however, ran counter to those by Hartshorne (1961, pp. 124–5) who argued that the other human sciences 'must' address nature–society relationships in order '… to understand the different social conditions of different lands and places.' Save for anthropology and the economics of resource use, however, the human sciences have largely ignored these relationships, at least until the recent rise of public and governmental concerns about them. If Hartshorne erred, was May correct? Perhaps, especially considering the strong role that nature–society geographers have played in the creation of interdisciplinary research institutes and international organizations dedicated to environmental studies and, along with physical geographers, the considerable impact that they have had, far out of proportion to their numbers, in defining much of the international agenda dealing with global environmental change, especially its human dimensions. Yet the very interdisciplinary qualities of these activities raises questions about how much the 'core' of nature–society studies are embedded in any one discipline.

6.2 Unified Understanding

Does a nature–society identity provide unity to geographical understanding? Again, Hartshorne was adamant that it did not, drawing on Hettner's argument that the 'logical unitary structure of geography [is] … destroyed' [by the dualism inherent in the nature–society position], whereas 'if the concept of nature and man proceeds from the chorological viewpoint, it is homogeneous in all important points' (in Hartshorne 1961, p. 123).

Hartshorne provided little support for Hettner's claim, as if the weight of history were sufficient to prove the point. But this history is not necessarily borne out in contemporary geography in which the divisions between human and physical geography have grown (Stoddart 1987), and frameworks for unification drawn from the spatial–chorological position do not necessarily speak to nature–society interests. Contemporary attempts to articulate the unity of geography through the nature–society position are few (but see below). In principle, however, there is no reason why a focus on the relationship should create the dualism noted, and various reasons why it should lead to more integrative or unified views of humankind and the physical world. Indeed, this last rationale is precisely that used to justify the investment in the many interdisciplinary human–environment programs, centers, and institutes emerging worldwide.

6.3 Nature–Society and Geography as Practiced

What passes for accepted practice in large measure shapes the identity of disciplines. Pattison (1964) identified four traditions or domains of research practiced in geography in the mid-1990s, which are still appropriate in the twenty-first century, allowing for name changes: mapping sciences, physical or environmental geography, space–society studies (human, including regional, geography), and nature–society (or human–environment) studies. Privileging the nature–society domain as the identity of geography encounters problems rationalizing the existence of the other three domains as independent subjects of study. Barrows and other proponents of the nature–society position of yesteryear noted this predicament and dealt with it by dismissing the other domains, save as they were needed to understand nature–society relationships.

More recently, Peet (1998, pp. 1–3) tied a nature–society defined geography to spatial–chorological geography by insisting that nature–society relationships can only be understood in real-world, relational space. In this interpretation, space is the 'veneer' of nature–society relationships, such that to examine spatial relationships axiomatically involves nature–society relationships. Alternatively, Sack (1997) explored a way in which nature and society can be reconciled without the reductionism employed by natural science approaches and some social science approaches. This treatment is accomplished by focusing on places because, according to Sack, place (a bounded area) is the instrument that humankind uses to combine nature and society, or the forces and process of each. Without place, understanding privileges nature, society, or a particular explanatory form drawn from one or the other. With place, however, a 'less privileging' understanding could be possible. Should this understanding be forthcoming, perhaps the long-standing tensions between the two visions of geography might dissipate.

7. Nature–Society Geography Entering the Twenty-first Century

What is and ought to be humankind's relationship with nature? This question remains one of the most fundamental and vexing confronting our species. Geography has long claimed much of this question as its own, and perhaps has devoted more attention to its various elements than any other formal discipline of study. This tradition, strongly muted during the mid-twentieth century, had revived significantly by the

1990s, perhaps reflecting geography's self-imagine vis-à-vis environmental determinism, the rise internationally of multidisciplinary concerns about the human–environment condition, and the emergence of new perspectives and methods for nature–society studies. Rebuilding on its past, nature–society geography may offer the largest number and array of researchers from any discipline experienced in bridging the natural and human science components of these concerns. So reinvigorated, the contributions of the nature–society domain of geography appear to stand on equal footing with those of all other geographic domains of study (National Research Council 1997).

See also: Cultural Geography; Environmental Determinism; Geographic Thought, History of; Geography; Sauer, Carl Ortwin (1889–1975)

Bibliography

Barrows H H 1923 Geography as human ecology. *Annals of the Association of American Geographers* **13**: 1–14

Blaikie P, Brookfield H 1987 *Land Degradation and Society*. Methuen, London

Darby H C 1940 *The Draining of the Fens*. Cambridge University Press, Cambridge, UK

Dickinson R E 1969 T*he Makers of Modern Geography*. Praeger, New York

Gaile G L, Willmott C J (eds.) 1989 *Geography in America*. Merrill Pub. Co., Columbus, OH

Glacken C J 1967 *Traces on the Rhodian Shore: Nature and Culture in Western Thought from Ancient Times to the End of the Eighteenth Century*. University of California Press, Berkeley, CA

Hartshorne R 1961 [1939] *The Nature of Geography: a Critical Survey of Current Thought in the Light of the Past*. Association of American Geographers, Lancaster, PA

Harvey M E, Holly B P (eds.) 1981 *Themes in Geographic Thought*. Croom Helm, London

James P E, Martin G J 1981 *All Possible Worlds: A History of Geographical Ideas*, 2nd edn. Wiley, New York

Johnson D W (ed.) 1954 *Geographical Essays by William Morris Davis*. Dover, New York

Kates R W 1987 The human environment: The road not taken, the road still beckoning. *Annals of the Association of American Geographers* **77**: 525–34

Leighly J (ed.) 1963 *Land and Life: A Selection from the Writings of Carl Ortwin Sauer*. University of California Press, Berkeley, CA

Livingstone D N 1993 *The Geographical Tradition: Episodes in the History of a Contested Enterprise*. Blackwell, Oxford, UK

May J A 1970 *Kant's Concept of Geography and Its Relation to Recent Geographical Thought*. University of Toronto Press, Toronto, Canada

National Research Council (U.S.) Committee on Geography 1965 *The Science of—Geography: report*. National Academy of Sciences–National Research Council, Earth Science Division, Washington, DC

National Research Council (Rediscovering Geography Committee, Board on Earth Sciences and Resources, Commission on Geosciences, Environment and Resources) 1997 *Rediscovering Geography: New Relevance for Science and Society*. National Academy Press, Washington, DC

O'Riordan T (ed.) 1997 *Ecotaxation*. St. Martin's Press, New York

Pattison W D 1964 The four traditions of geography. *Journal of Geography* **63**: 211–16

Peet R 1998 *Modern Geographical Thought*. Blackwell, Oxford, UK

Sack R D 1997 *Homo Geographicus: A Framework for Action, Awareness, and Moral Concern*. Johns Hopkins University Press, Baltimore, MD

Stoddart D R 1987 To claim the high ground: Geography for the end of the century. *Transactions of the Institute of British Geographers* **12**: 327–36

Tatham G 1951 Geography in the nineteenth century. In: G Taylor (ed.) *Geography in the Twentieth Century*. Methuen, London, pp. 28–69

Thomas W M (ed.) 1956 *Man's Role in Changing the Face of the Earth*. University of Chicago Press, Chicago

Turner B L II 1997 Spirals, bridges, and tunnels: Engaging human-environment perspectives in geography? *Ecumene* **4**: 196–217

Turner B L II, Clark W C, Kates R W, Richards J F, Mathews J T, Meyer W B (eds.) 1990 *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere Over the Past 300 Years*. Cambridge University Press, Cambridge, UK

Watts M 1983 *Silent Violence: Food, Famine and Peasantry in Northern Nigeria*. University of California Press, Berkeley, CA

Williams M 1974 *The Making of the South Australian Landscape: A Study in the Historical Geography of Australia*. Academic Press, London

Zimmerer K S, Young K R (eds.) 1998 *Nature's Geography: New Lessons for Conservation in Developing Countries*. University of Wisconsin Press, Madison, WI

B. L. Turner II

Navigation

'Navigation,' from the Latin *navis*, a ship, and *agere*, to drive, originally referred to the art of sailing a vessel from one point to another, but is now also applied to guiding aircraft and spacecraft. Beginning with early evidence for deep sea voyaging, this article examines the worldwide development of navigation and touches upon the implications for humankind of navigating freely around the globe and into space.

1. Early Navigation in the Pacific

Although most accounts of the development of navigation begin with early seafaring in the Mediterranean and the Indian Ocean, recent research points to the Pacific as the region where humans first sailed routinely

far out to sea. Around 1500 BC Stone Age seafarers began expanding eastward from island Southeast Asia toward the open Pacific, where, over the next 2,500 or so years, they and their descendants discovered and settled all the inhabitable islands of remote Oceania, the vast island world composed of eastern Melanesia, Micronesia, and Polynesia.

Sailing eastward into the ocean was a formidable task as the trade winds blow from the east and inter-island distance increase until they reach more than 3,500 km in the far reaches of Polynesia. To accomplish this, pioneering seafarers sailed in outrigger and double canoes, exploited seasonal wind shifts, and navigated without instruments by observing the stars, sun, and other celestial bodies, ocean swells, winds, birds, and other phenomena. They oriented on a conceptual compass of up to 32 points labeled with the rising and setting points of key stars and constellations or the distinctive winds that blew from particular directions. Although they might outline their compass on the ground for teaching purposes, navigators sailed only with its image in mind, along with memorized compass bearings to and from the islands within their voyaging range. They steered on stars and sun when low on the horizon, or on the dominant swells and prevailing wind, all calibrated to their compass points. Their system did not include latitude and longitude, although some navigators may have judged north-south position by what stars passed directly over them, and from the height of Polaris and other circumpolar stars. Progress en route was estimated in terms of the shifting bearings of an unseen reference island off to one side of the course, and other mental plotting methods. To detect land before it could be seen, navigators looked for land-nesting birds fishing out to sea, disruptions in the ocean swells caused by islands, and other signs. That virtually every Oceanic island is part of an archipelago both facilitated discovery of new islands and made navigating between widely separated ones easier. A navigator could make landfall on any island within the chain containing the target island, then reorient toward the final destination.

However, although this Pacific initiative demonstrates that long, navigated voyages can be made without instruments, charts or mathematical calculations, it did not lead to the global, science-based navigation methods employed today.

2. *Navigating Offshore Europe, Asia, and Africa*

While canoe voyagers were expanding over the Pacific, other maritime peoples were sailing along the shores of the linked continents of Africa, Europe, and Asia, and to and from adjacent islands. Sea traders and raiders had long sailed in the Mediterranean, the Indian Ocean, and the China Sea. By Roman times—if not before—trade was flowing between the Mediterranean world and China via the Indian Ocean. Judging from Greek, Arab, and Chinese sources, although much voyaging was coastal, and thus involved navigating primarily by landmarks and soundings, some routes took mariners well away from land, and therefore required at a minimum the ability to sail by steady winds and observations of the stars. Voyages made between Arabia and India by means of alternating monsoon winds provide a case in point.

3. *Deep Sea Navigation and Global Voyaging*

Although Herodotus mentioned a Phoenician expedition made around Africa, and much later Viking sailors were able to travel from one island to another across the North Atlantic, regular sailing beyond the linked continents did not begin until the fifteenth century AD, when the Portuguese developed methods for deep-sea navigation. Early in that century Portuguese mariners began to explore down the African coast, but found that whereas it was easy to sail south with the prevailing northerly winds, tacking back against these was most difficult. Experience gained in sailing to and from the Azores and islands scattered down the African coast had acquainted them with the global winds, the easterly trade winds of the tropics and the westerlies of the temperate latitudes. In sequence, these offered a way to sail back from Africa with fair winds all the way. By heading directly out to sea, these mariners learned that they could sail in a long slant northwest across the easterly trade winds, and, upon reaching the belt of westerlies, could then turn east to sail downwind to Portugal.

However, methods worked out in the Mediterranean for navigating with compass, rudimentary charts, and dead reckoning (estimating position by recording compass heading and distance run) proved inadequate on these long, roundabout crossings. Accordingly, the navigators adapted an astronomical technique for measuring the height of Polaris above the horizon to monitor their progress northward until they were directly west of their destination in Portugal. Before departing from, for example, Lisbon, they marked the altitude of Polaris as seen from that port on a quadrant, a simple instrument for measuring the height of stars above the horizon. (As Polaris was then located about 3.5 degrees from the north celestial pole, with corrections its angular altitude indicated the latitude of the observer.) Then, on the return voyage as they headed west from Africa and then started curving to the north, they watched Polaris rise higher and higher in the sky. Using their quadrant they determined when Polaris had risen to the same height as seen from Lisbon, and then turned east to sail downwind with the westerlies, keeping the Pole Star at the same height in order to reach the port directly.

The crisis caused when, upon crossing the Equator, Polaris dropped below the horizon was resolved in the 1480s. In one of the first examples of a government

bringing science to bear on an applied problem, King João II enlisted astronomers, mathematicians, and other scholars to develop a method whereby the height of the sun could be used in place of that of Polaris to determine latitude. With the resultant 'Rule of the Sun,' and tables of the sun's daily position at various latitudes, navigators were able to sail farther and farther down the African coast. In 1497–8 Vasco da Gama sailed in a long arc across the trade winds to the South Atlantic westerlies, then turned east to sail around Africa and up its east coast. From there, under the direction of a local navigator (arguably either an Arab or a Gujerati), he reached India, thereby opening a direct route from Europe to this fabled source of spices, jewels, and other riches.

In 1492 Christopher Columbus used techniques learned from the Portuguese to sail with the trade winds to the Bahamas (which he insisted were outliers of Asia), and then returned by heading north to catch the westerlies. Then in 1522 Juan del Cano brought the *Victoria* back to Spain, completing the first circumnavigation begun by his fallen commander, Ferdinand Magellan. At last the 'world encompassed,' thanks to Iberian seamanship in exploiting global winds and in applying navigational innovations pioneered in other seafaring traditions. For example, a good case can be made that the Chinese developed the magnetic compass, and employed it at sea at least several centuries before its appearance in the Mediterranean in the twelfth century. Furthermore, although the Portuguese use of the height of Polaris to judge progress northward from Africa was a genuine breakthrough, Arab navigators had long before mastered the art of sailing across the Indian Ocean with the aid of Polaris measurements to locate headlands and ports.

4. *Further Navigational Improvements and the World System*

The Iberians exploited their maritime achievements by establishing maritime empires, but soon were outgunned and surpassed in both nautical innovation and techniques of colonial exploitation by the Dutch and English. As ship design and seamanship advanced, quantitative navigation evolved through the further application of mathematics and astronomy, and developments in instrumentation and cartography. Astronomical tables multiplied, eventually leading to the publication from 1678 onwards of nautical almanacs. Instruments for measuring the altitude of celestial bodies improved, leading to the reflecting octant and sextant in the eighteenth century. Magnetic variation was mapped globally, and improved compasses and ways of measuring distance sailed were introduced, along with charting innovations, including the Mercator projection that enabled mariners to plot long courses with reasonable accuracy.

The last great innovation of this era of sail was the marine chronometer. Whereas navigators could determine latitude, they could only estimate their longitude, resulting in uncertainty, delay, and sometimes maritime tragedy. Because the earth turns on its axis, to determine the longitude of a ship, that is its position east or west of a reference point such as England's Greenwich Observatory, navigators had to keep the time of that reference point. With the development in the late eighteenth century of marine chronometers that could keep time accurately on transoceanic voyages, and their widespread adoption in the nineteenth century, they could at last calculate longitude. The celestial navigation system that had been evolving over the centuries was finally complete. Given clear weather, a navigator could expeditiously and safely direct a ship to any port over the globe. With the conversion from sail to steam, the nautical means for efficient global maritime trade were in place, and the modern world system erected upon it boomed.

5. *Navigating Electronically*

However, the weak point in navigating by magnetic compass, chronometer, and stellar observations remained the changing weather. When clouds, rain, or fog prevented navigators from seeing the stars they had to fall back upon dead reckoning, which, even with instruments for measuring direction of travel and distance run, suffered from cumulative error. Radio direction finding pioneered in the 1910s was an early but only partial solution. However thankful fogbound navigators were to gain radio bearings, these were not very accurate, and available only along the coasts of industrial nations. The first truly all-weather position finding system, LORAN, was developed during World War II for wartime navigation. The reception of synchronized pulses from two pairs of transmitters located at carefully surveyed positions resulted in accurate navigational fixes. The subsequent development of inertial navigational systems was stimulated by further military requirements: guiding missiles and providing bombers and submarines with position information needed for targeting. Employing accelerometers and gyroscopes to determine changes in velocity and direction, these systems provide a continuous indication of position, subject, however, to cumulative error. The most accurate and revolutionary of the electronic navigational technologies is the Global Positioning System (GPS), a US military system made accessible to civilian users, though initially degraded in accuracy. (Russia has its own *Glonass* system.) The GPS system depends upon a constellation of 24 satellites in low earth orbit, which transmit signals that cheap hand-held receivers can process to determine latitude, longitude, and altitude with great accuracy. It was fully operational by the

early 1990s, and is now widely used around the world, though not without some misgivings over military control.

6. Global Navigation and Beyond

Efforts to develop techniques of wayfinding that began when seafarers first sailed beyond sight of land resulted eventually in a global system of navigation whereby mariners employed the starry points of the celestial sphere to guide their vessels. Now ships, aircraft, and even lone hikers are guided precisely over the globe within an artificial sphere of orbiting satellites. Sophisticated navigational tools also direct robotic and crewed spaceships through interplanetary space, and eventually will be employed to guide them to the stars. Considering the globalizing impacts on human society that have flowed from navigating with increasing accuracy around the planet, thought must now be given to what transformations await humankind as they spread beyond the natal planet.

See also: Polynesia and Micronesia: Sociocultural Aspects

Bibliography

Casson L 1991 *The Ancient Mariners, Seafarers and Sea Fighters of the Mediterranean in Ancient Times*, 2nd edn., Princeton University Press, Princeton, NJ

Finney B 1998 Traditional navigation and nautical cartography in Oceania. In: Woodward D, Lewis G M (eds.) *The History of Cartography: Cartography in Traditional African, American, Arctic, Australian, and Pacific Societies*. University of Chicago Press, Chicago Vol. 3, pt. 2

Finney B R, Jones E M (eds.) 1985 *Interstellar Migration and the Human Experience*. University of California Press, Berkeley, CA

Getting I A 1993 The Global Positioning System. *IEEE Spectrum* **30**(12): 36–47

Hourani G F 1963 *Arab Seafaring in the Indian Ocean in Ancient and Early Medieval Times*. Khayats, Beirut, Lebanon

Lewis D 1972 *We the Navigators*. University of Hawaii Press, Honolulu, HI

Nadvi S S 1966 *The Arab Navigation*. Sh Muhammad Ashraf, Lahore, Pakistan

Needham J 1971 *Science and Civilization in China, Vol. 4 Physics and Physical Technology, Part III Civil Engineering and Nautics*. Cambridge University Press, Cambridge, UK

Parry J H 1981 *The Discovery of the Sea*, 1st edn., University of California Press, Berkeley, CA

Taylor E G R 1956 *The Haven-Finding Art, A History of Navigation from Odysseus to Captain Cook*. Hollis & Carter, London

Waters D 1992 Columbus's Portuguese Heritage. *The Mariner's Mirror* **78**: 385–405

Williams J E D 1992 *From Sails to Satellites, the Origin and Development of Navigational Science*. Oxford University Press, Oxford, UK

B. Finney

Navigation in Spatial Environments, Neural Basis of

1. Navigational Abilities are Thought to Depend on an Internal Representation Known as a Cognitive Map

Many animals, including humans, show a remarkable ability to navigate through their environment (e.g., Gallistel 1990). For example, in a familiar town it is often possible to plan an efficient route between two places, such as the bank and the grocery store, even if one has never previously traveled directly between the two. It has been postulated that this ability depends on an internal, abstract, map-like representation of the large-scale environment within which we conduct our daily activities (Tolman 1948, O'Keefe and Nadel 1978). This postulated internal representation, sometimes referred to as a cognitive map, presumably provides us with the ability to have a sense of our current position and directional heading within our environment, as well as the relative position of other important locations (work, home, school, etc.).

2. The Hippocampal Formation has been Implicated as a Possible Location for the Cognitive Map

Studies of individuals who have sustained focal brain damage due to illness or injury have revealed certain brain regions that may contain the postulated map-like representations. For example, experimental studies in rats have suggested that a brain region known as the hippocampal formation may be critical, since damage to this area causes deficits on spatial learning tasks (O'Keefe and Nadel 1978). In addition, humans with damage to this region sometimes show selective deficits in topographical orientation (Habib and Sirigu 1987, Maguire et al. 1996a). These individuals are unable to learn their way around in a new environment, and when tested formally, they are impaired in making correct judgments about the relative distance or direction between different locations (Maguire et al. 1996a). This topographical disorientation can sometimes appear in the absence of other detectable mnemonic or perceptual deficits. That is, the people can recognize and name landmarks, and can even learn and remember the identity of new landmarks, but they cannot learn effectively about the spatial relations between them, or how to navigate from one to the other.

Complimentary evidence for the importance of the hippocampal region has come from brain imaging studies that employ navigation tasks (Aguirre et al. 1996, Maguire et al. 1996b, 1997, 1998). In one study

(Maguire et al. 1997), experienced London taxi drivers were asked to plan routes from one location to another within the city. For comparison, the drivers were also asked to perform several control tasks, such as describing famous landmarks, or recounting the plot sequence from well-known films. Relative to the control tasks, the route-planning task caused relatively selective activation of the right hippocampus.

Despite the apparent importance of the hippocampal formation for learning about new environments, recent evidence suggests that navigational memory for environments learned long ago may reside in cortical areas other than the hippocampus. For example, a study by Teng and Squire (1999) reported on an individual who had sustained massive bilateral damage to the temporal lobes, due to an episode of encephalitis. As is well documented, large temporal lobe lesions like this cause a global amnesia for episodes, events, and places which are experienced after the damage, as well as for those which occurred for some time prior to the damage. Of relevance here, the patient reported by Teng and Squire (1999) had normal recall for the urban area in which he had grown up, but had moved away from, nearly 40 years earlier. He was able to accurately describe how to travel from one location to another within this environment, and when asked to imagine himself positioned in a certain location and directional orientation, he could accurately point in the direction of various other landmarks. Thus, these data indicate that the ability to recall and utilize topographic information for an environment learned long ago must reside in some area other than (or in addition to) the hippocampus.

3. *The Retrosplenial (Posterior Cingulate) Cortex has also been Implicated as a Possible Location for the Cognitive Map*

If topographical information for an environment learned long ago eventually becomes independent of the hippocampal formation, this raises the question as to what other brain region may take over these memories. Recent evidence has suggested that the retrosplenial cortex may be one important region. Takahashi et al. (1997) have reported on three patients who showed selective topographic amnesia after sustaining damage to the right retrosplenial region. Importantly, the disorientation experienced by these people was equally severe for both familiar as well as new environments.

Additional evidence for the importance of the retrosplenial cortex comes from the brain imaging study outlined above (Maguire et al. 1997); retrosplenial cortex was another brain area which was selectively activated during the taxi drivers' route planning.

4. *Investigations of Single Cell Activity in the Hippocampal Formation and Retrosplenial Cortex have Revealed Clues about the Cellular Basis of the Postulated Cognitive Map*

In order to gain insight into the detailed nature of the postulated cognitive map, many investigators, beginning with O'Keefe and colleagues (e.g., O'Keefe and Dostrovsky 1971, O'Keefe and Conway 1978) have looked at the activity of individual neurons in the relevant brain regions of experimental animals, to see what kinds of spatial signals may exist. These investigations have revealed two main types of spatial cell.

4.1 *Place Cells*

One type, known as place cells, fires whenever the rat is in a particular location. For example, one cell may fire at a high rate whenever the rat is in the northeast corner of a square enclosure, and be completely inactive when the rat is in any other location. Each of these place cells has its own region of high firing so that, as a whole, the population of hippocampal place cells provides an ongoing representation of the animal's current location; each location the rat occupies is coded by activity in a unique subset of the hippocampal cells. Cells of this general type have been found throughout the hippocampal formation (O'Keefe and Dostrovsky 1971, Sharp and Green 1994, Taube 1995) and also in the retrosplenial cortex (Cho and Sharp 2001).

It is not completely clear how these location-specific signals are generated. One possibility could be that they simply respond to the presence of landmarks (environmental stimuli) which are available in the preferred region. This possible explanation is supported by early data which showed that the location of the hippocampal place cell firing fields can be controlled by the location of environmental cues. For example, in one study (Muller and Kubie 1987) cells were recorded while rats navigated in a gray cylindrical apparatus. The wall of the cylinder was equipped with a single, white cue card that acted as a landmark. When the card was rotated during separate probe sessions, the location of each of the place cell fields rotated as well, by the same amount.

Additional data, however, showed that the place cell signals are more complicated than suggested by this simple sensory explanation. For example, in the study outlined above (Muller and Kubie 1987) it was found that when the white card was removed entirely, the place cells could still retain location-specific firing patterns. In this case, each place cell showed a firing field that was the same size, shape, and distance from the wall as it was in the presence of the white card. However, the angular location of the fields was unpredictable in this case. During any one recording

session, each cell fired in just one location, but the location that it chose could not be predicted in advance.

Data of this sort suggest that the cells utilize a process known as path integration (ded reckoning), at least in part, to generate the place cell signals (e.g., McNaughton et al. 1995). Path integration refers to the use of one's own movement to track position. The behavior of the place cells in the no-card probe sessions could be explained using this idea by suggesting that, first, when the animal is introduced into the chamber at the start of the session, the place cell firing pattern is randomly set to represent some particular location. Thus, since there is no polarizing cue, when the rat is initially placed down next to the cylinder wall, place cell activity appropriate to, for example, the northwest region of the inner wall may be randomly 'cued' up. Then, after this initialization process, when the animal begins to move, the system somehow uses information about the speed and direction of movement to continuously update the representation of the animal's current position.

The fact that the place cells are controlled by both environmental landmarks and path integration is remarkably compatible with introspective evidence about our own navigation. Thus, in a familiar environment, if one becomes momentarily confused about one's location, it is usually possible to simply look around for a familiar landmark, to become reoriented. However, if familiar landmarks become unavailable (such as might happen after dark, or when one has ventured into new territory), then it is still possible to have some sense of where one is, just based on information about where the journey began, and the subsequent direction and distance traveled.

4.2 Head Direction Cells

The other type of spatial cell is known as the head direction cell (Ranck 1984, Taube et al. 1990a). Each of these cells fires whenever the rat faces one particular direction. For example, one cell may fire whenever the rat faces southeast, while another fires only when the rat faces north, etc. Thus, these cells are the compliment of the place cells, in that they provide an ongoing representation of the current directional heading, rather than location. These cells were originally discovered in the subicular complex of the hippocampal formation (Ranck 1984), and have since been discovered in several other areas as well, including the retrosplenial cortex (Chen et al. 1994, Cho and Sharp 2001). As is the case for the place cells, the head direction cells have also been found to utilize both visual landmarks as well as path integration to keep track of the animal's directional heading (Taube et al. 1990b).

5. Neural Network Models have been Developed to Explain the Firing Properties of the Place Cells and Head Direction Cells

Several neural network models have been developed to attempt to replicate the spatial firing characteristics of place cells and head direction cells.

5.1 Models of Place Cells

Many of the place cell models have focused on the fact that these cells are influenced by environmental cues, and they attempt to explain the properties of place cells as due to convergent sensory inputs (e.g., O'Keefe and Burgess 1996, Sharp 1991, Treves et al. 1992, Zipser 1985).

One obvious limitation of these models is that they do not account for the path integration properties shown by these cells. For this, McNaughton and colleagues have developed a model in which movement information (along with environmental cues) serves as an input to the hippocampal system (McNaughton et al. 1995, Samsonovich and McNaughton 1997). Specifically, the place cells themselves project onto another postulated cell layer which also receives input about the current direction and speed of movement. Thus, each cell in this layer fires to a particular combination of place, movement, and direction. These cells then project back onto the place cells themselves, and are connected so that they induce firing in the appropriate 'next' set of place cells (i.e., those which represent the location which will result from the current location, direction, and speed of movement).

5.2 Models of Head Direction Cells

Several models have been developed to simulate the characteristics of head direction cells (Redish et al. 1996, Sharp et al. 1996, Skaggs et al. 1995, Zhang 1996). These models all have the same basic properties. They all begin with the idea that the head direction cells themselves are linked together to form a stable attractor. In this network, cells that are meant to represent different directions (e.g., north vs. south) have strong, mutual, inhibitory connections, while cells meant to represent similar directions are linked through excitatory connections. This connectivity ensures that, at any one time, the system will settle into a pattern of activity in which cells which represent one directional range will be active, while all other cells are silent.

To implement the path integration abilities, the system also receives information about angular head movement (such as might be provided by the vestibular system, or motor command signals). Cells carrying

this information project onto another postulated layer which also receives input from the head direction cells themselves. Each of these cells, thus, fires for a unique combination of current directional heading and angular movement. This combination allows for the prediction of the next directional heading (e.g., a current heading of north, along with a 90-degree clockwise turn will result in a directional heading of east). These cells then project back onto the head direction cells in such a way that they drive the next appropriate set of head direction cells.

The head direction cells are also assumed to receive input from sensory cells, so that they can be 'set' by environmental cues, as described above.

6. *There is Still Much to be Learned about the Cognitive Map*

The discovery of place cells and head direction cells constitutes a major step toward an understanding of the neural/computational basis of navigation. However, there are critical aspects of the story which are yet unknown. In particular, it is not known how these place and direction signals are used to generate actual goal-directed behavior.

Although there has been remarkably little empirical work on this problem, several neural modeling approaches have been used to attempt to provide a conceptual framework for this aspect of navigation.

In one type of model it is assumed that the information necessary for route planning resides in the connectivity of the place cells themselves (Blum and Abbott 1995, Hetherington and Shapiro 1993, Muller et al. 1996). It is assumed that as an animal explores an environment, place cells which fire successively along particular paths taken by the animal will become strongly connected, due to a Hebb-like synaptic plasticity. In the Muller et al. (1996) model, this connectivity is used to find the shortest path between the place cell representations of the animal's current position and a desired goal. In the Blum and Abbott (1995) model, the strong connection between two cells along a pathway is assumed to cause the firing of place cells at any one moment to 'jump ahead' a bit, so that cells corresponding to the 'next' location are active. Moreover, this connectivity between place cells is presumed to be enhanced by previously experienced reward. During recall, the rat selects an optimal trajectory by traveling through the path represented by the activity of the currently active place cells. Due to the experience on previously rewarded trials, these place cells represent the next step in the optimal path toward reward.

One limitation of these models is that they do not specify how the activity of the place cells themselves can be translated into actual behavior. Additional neural machinery would be necessary to calculate how to move from the current position to the one indicated by the trajectory-related place cell activity.

A model which attempts to provide the necessary route-generating machinery has been provided by Brown and Sharp (1995). In this model, both place cells and head direction cells project onto a layer of motor cells which drive behavior. Initially, the strengths of these connections are set randomly. When the rat encounters reward, however, recently active connections between place/head direction cells and motor command cells are strengthened. This means that behaviors that lead to the goal from a particular position are strengthened.

One limitation of this model is that it does not work over long distances. This is true since the only connections that can be strengthened are those that were active a short time before the animal reached the goal.

A postulated solution to this problem has been provided by Foster et al. (2000). In this model, a temporal distance learning module has been added. Over the course of the animal's exploration this model eventually learns the reward value of each step the animal takes, regardless of where it is in the environment. The output of this module is then used to guide the actual motor learning.

7. *Summary*

In summary, the remarkable navigational abilities demonstrated by both animals and humans are thought to depend on a sophisticated, internal representation of the environment, sometimes referred to as a cognitive map. Two brain regions that have been implicated as possibly containing the cognitive map are the hippocampal formation and the retrosplenial cortex. It appears that these two regions somehow work together to help orchestrate navigational abilities. Specifically, the hippocampus seems to play a necessary role during initial learning about a new environment, while the retrosplenial cortex may be necessary at all stages. The discovery of place cells and head direction cells in both these regions has provided some clues about the details of the cognitive map, but it is still not clear how these cells actually help to guide navigational behavior.

See also: Cognitive Maps; Ethnic Conflict, Geography of; Mental Maps, Psychology of; Navigation; Navigation in Virtual Space: Psychological and Neural Aspects; Perception of Extrapersonal Space: Psychological and Neural Aspects; Reading, Neural Basis of; Spatial Cognition

Bibliography

Aguirre G K, Detre J A, Alsop D C, D'Esposito M 1996 The parahippocampus subserves topographical learning in man. *Cerebral Cortex* **6**: 823–29

Blair H T, Sharp P E 1996 Visual and vestibular influences on head direction cells in the anterior thalamus of the rat. *Behavioral Neuroscience* **110**: 1–18
Blum K I, Abbott L F 1995 A model of spatial map formation in the hippocampus of the rat. *Neural Computation* **8**: 85–93
Brown M A, Sharp P E 1995 Simulation of spatial learning in the Morris water maze by a neural network model of the hippocampal formation and nucleus accumbens. *Hippocampus* **5**: 189–97
Chen L L, Lin L H, Green E J, Barnes C A, McNaughton B L 1994 Head direction cells in the rat posterior cortex. I. Anatomical distribution and behavioral modulation. *Experimental Brain Research* **101**: 8–23
Cho J, Sharp P E 2001 Head direction, place, and movement correlates for cells in the rat retrosplenial cortex. *Behavioral Neuroscience* **115**: 3–25
Foster D J, Morris R G M, Dayan P 2000 A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* **10**: 1–16
Gallistel C R 1990 *The Organization of Learning*. MIT, Cambridge, MA
Habib M, Sirigu A 1987 Pure topographical disorientation: A definition and anatomical basis. *Cortex* **23**: 73–85
Hetherington P A, Shapiro M L 1993 A simple network model simulates hippocampal place fields: II. Computing goal-directed trajectories and memory fields. *Behavioral Neuroscience* **107**: 434–43
Maguire E A, Burgess N, Donnett J G, Frackowiak R S J, Frith C D, O'Keefe J 1998 Knowing where and getting there: A human navigational network. *Science* **280**: 921–24
Maguire E A, Burke T, Phillips J, Staunton H 1996a Topographical disorientation following unilateral temporal lobe lesions in humans. *Neuropsychologia* **34**: 993–1001
Maguire E A, Frackowiak R S J, Frith C D 1996b Learning to find your way: A role for the human hippocampal formation. *Proceedings of the Royal Society of London B* **263**: 1745–50
Maguire E A, Frackowiak R S J, Frith C D 1997 Recalling routes around London: Activation of the right hippocampus in taxi drivers. *Journal of Neuroscience* **17**: 7103–10
McNaughton B L, Barnes C A, Gerrard J L, Gothard K, Jung M W, Knierim J J, Kudrimoti H, Quin Y, Skaggs W E, Suster M, Weaver K L 1995 Deciphering the hippocampal polyglot: The hippocampus as a path integration system. *Journal of Experimental Biology* **199**: 173–85
Muller R U, Kubie J L 1987 The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience* **7**: 1951–68
Muller R U, Stead M, Pach J 1996 The hippocampus as a cognitive graph. *Journal of General Physiology* **107**: 663–94
O'Keefe J, Burgess N 1996 Geometric determinants of the place fields of hippocampal neurons. *Nature* **381**: 425–28
O'Keefe J, Conway D H 1978 Hippocampal place units in the freely moving rat: Why they fire where they fire. *Experimental Brain Research* **31**: 573–90
O'Keefe J, Dostrovsky J 1971 The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely moving rat. *Brain Research* **34**: 171–75
O'Keefe J, Nadel L 1978 *The Hippocampus as a Cognitive Map*. Oxford University Press, Oxford, UK
Ranck J B Jr 1984 Head-direction cells in the deep cell layers of the dorsal presubiculum in freely moving rats. *Society of Neuroscience Abstracts* **10**: 599
Redish A D, Elga A N, Touretzky D S 1996 A coupled attractor model of the rodent head direction system. *Network* **7**: 671–86
Samsonovich A, McNaughton B L 1997 Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience* **17**: 5900–20
Sharp P E 1991 Computer simulation of hippocampal place cells. *Psychobiology* **19**: 103–15
Sharp P E, Blair H T, Brown M 1996 Neural network modeling of the hippocampal formation spatial signals and their possible role in navigation: A modular approach. *Hippocampus* **6**: 720–34
Sharp P E, Green C 1994 Spatial correlates of firing patterns of single cells in the subiculum of the freely moving rat. *Journal of Neuroscience* **14**: 2339–56
Skaggs W E, Knierim J J, Kudrimoti H S, McNaughton B L 1995 A model of the neural basis of the rat's sense of direction. In: Tesauro G, Touretzky D S, Lean T K (eds.) *Advances in Neural Information Processing Systems 7*. MIT Press, Cambridge, MA
Takahashi N, Kawamura M, Shiota J, Kasahata N, Hirayama K 1997 Pure topographic disorientation due to right retrosplenial lesion. *Neurology* **49**: 464–69
Taube J S 1995 Place cells recorded in the parasubiculum in freely moving rats. *Hippocampus* **5**: 569–83
Taube J S, Muller R U, Ranck J B Jr 1990a Head direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience* **10**: 420–35
Taube J S, Muller R U, Ranck J B Jr 1990b Head direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. *Journal of Neuroscience* **10**: 436–47
Teng E, Squire L R 1999 Memory for places learned long ago is intact after hippocampal damage. *Nature* **400**: 675–77
Tolman E C 1948 Cognitive maps in rats and men. *Psychological Review* **55**: 189–208
Treves A, Miglino O, Parisi D 1992 Rats, nets, maps, and the emergence of place cells. *Psychobiology* **20**: 1–8
Zhang K 1996 Representation of spatial orientation by the intrinsic dynamics of the head direction cell ensemble: A theory. *Journal of Neuroscience* **16**: 2112–26
Zipser D 1985 A computational model of hippocampal place fields. *Behavioral Neuroscience* **99**: 1006–18

P. E. Sharp

Navigation in Virtual Space: Psychological and Neural Aspects

1. Introduction

The terms 'Virtual Reality,' 'Virtual Space,' and 'Virtual Environment' (VE) are used interchangeably to describe a computer-simulated place or environment with which users can interact via an interface. In its most advanced (and currently, putative) form, a VE would mesh seamlessly with the user's perceptions such that they would be unable to tell it apart from their perceptions of the real world. Currently the most

common form of VE uses a simple interface consisting of a computer screen for vision, speakers or headphones for audition, and hand-operated input devices such as keyboards or joysticks for motion. More advanced systems might include a head-mounted display (HMD) and motion tracking of body movements. This allows separate views to be delivered to each eye, providing a more convincing experience of visual depth; movements of the head are converted to changes of view so that the eye-screens can be updated consistently with this movement. In all VEs there is a model of the user's body that can be moved by manipulating the control interface. As the user's virtual body moves, the simulated sensory inputs are updated to reflect its current position.

In addition to visual simulation, many VR systems are now able to accurately represent the location of auditory sources using hardware specifically designed for the task. Haptic feedback can also be available to some extent. This involves applying forces to the user's body to simulate the presence of objects in the environment. However, haptic systems tend to impose considerable constraints on the user's freedom of movement and are generally only used on particular joints for specific purposes. For example, the thumb and forefinger may be attached to a device that can apply pressure to signal the size of objects being grasped in the VE.

2. *Why Use VR?*

The advantages of VR depend very much on the use to which it is put; in architecture, for example, entire buildings can be experienced as a simulation without a single brick being laid. In psychology there are numerous benefits. Within a VE the stimuli are completely under the control of the experimenter, and responses can be conveniently recorded. This is especially useful in studies involving large spaces, circumventing the need for large-scale motion tracking and the practical problems of controlling a large environment. Also, in a VE the normal rules of physics need not apply, so that impossible manipulations of space and objects can occur. For example, non-Euclidian geometry can be simulated (e.g., Ruddle, in press); objects can be repositioned and physical dimensions changed instantaneously. Finally, one of the most exciting advantages of VR in neuroscience is that it can be used to simulate movement through space while a subject's head is kept still for the purposes of functional neuroimaging.

3. *'Immersion' and the Correspondence Between Real and Virtual Behavior*

A major concern in VR is the issue of immersion. This refers to the degree to which users accept the VE as a space in which they can act, and how convincing a simulation of reality the VE provides. There is evidence that variations in the VR interface have effects on the performance of the user. Chance et al. (1998) examined subjects' ability to judge direction to target locations in a virtual maze—a task requiring a series of turns and translations to be integrated into a representation of displacement (often referred to as path integration). Their interface had three modes: a motion-tracking, head-mounted display mode in which the subjects actually walked and turned physically (Walk mode); a mode in which movements were made using a joystick (Virtual Turn mode); and a mode in which turns were physically enacted but translations were not (Real Turn mode). The experimenters found significantly better performance in the Walk mode than in the Virtual Turn mode, suggesting that, for the purpose of path integration, immersion is best achieved using real walking to control movement in the VE.

Alternatively, research on imagined movements by Decety and Jeannerod (1995) suggests that, even when the environment is simulated entirely using visual imagery, actions are performed in a similar way to that in a computer-mediated VE. In this study, subjects imagined themselves moving through previously experienced virtual doorways of varying widths, and the duration of each action was related to the door width in the same way as during VR trials. Similarly Ghaem et al. (1997) found a good correspondence between real and imagined walking on a route through Paris in terms of the speed of movement. However, the validity of generalizing between virtual and real spaces remains a major concern. A number of studies have considered this issue. Henry and Furness (1993) found that architects consistently underestimated the size of VEs. Conversely, Ruddle et al. (1997) found subjects' performance on judgments of distance and direction in a large virtual building to be compatible with performance in the same real building. Witmer et al. (1996) compared route and building layout knowledge in subjects trained in a real building, the same building in VR, and the building learned from photographs and verbal descriptions. They found that when subjects were tested in the real building, virtual exposure gave better performance than verbal/pictorial, but neither was as effective as learning in the real building. In another experiment, Witmer and Klein (1998) found that subjects underestimate distances in both real and virtual spaces, but more so in VR. Also, Waller et al. (1998) found that, given long exposure, VE training of an environment eventually surpassed real-world training in tests of route learning. Finally, Klatzky et al. (1998) found a systematic overestimation of turn amplitude in simulated environments, which was not present in the real world.

Evidence from studies of normal subjects is mixed, then, but it appears that spatial cognition in VEs is at least closely related to that in the real world. Clinical

studies support this: Brooks et al. (1999) showed that practice in a virtual simulation of a rehabilitation unit transferred to real-world route-finding performance in a 53-year-old patient with anterograde amnesia.

4. The Psychology of Navigation in Virtual Space

A popular model of the acquisition of spatial information, the landmark–route–survey, or LRS, model holds that spatial information is acquired in stages during exploration. Landmark information is acquired first, followed by information concerning specific routes before, finally, a survey-like spatial representation is formed. (Siegel and White 1975, Thorndyke and Hayes-Roth 1982).

Alternatives to this model have been suggested by experiments using virtual spaces. Aginsky et al. (1997) examined driving behavior in VR and propose that there are two parallel strategies—visually dominated and spatially dominated. The former, they say, relies on wayfinding using visual decision points (landmarks, essentially), which are not integrated into a survey representation. In the later strategy, survey knowledge is used from the very start, rather than learned in stages. Colle and Reid (1998) agree with this; in their study of learning metric knowledge in small spaces, they found local survey knowledge was learned rapidly without recourse to landmarks and routes. They propose a dual mode model where local survey knowledge is achieved rapidly, but larger-scale layouts are learned according to the LRS model.

VR has also been used to probe the nature of the formation of the representation of space. Gillner and Mallot (1998) examined the use of information restricted to the views of landmarks and movements local to choice points and found that, at least for some subjects, the maze could be learned from this restricted information alone, and more abstract survey information could be extracted.

5. Strategy and Gender

Virtual studies of navigation have revealed interesting sex differences in navigational strategies. Astur et al. (1998) produced a virtual replication of the Morris water maze: a pool surrounded by landmarks in which a platform that is hidden under the water must be found). They found that males navigated to the platform consistently better than females. In another virtual maze study, Moffat et al. (1998) found that males made fewer errors and completed maze tasks more quickly than females. These findings are typical of the well-known gender differences in spatial cognition (e.g., Silverman and Eals 1992). We note that gender differences in navigation often reflect differences in strategy, as shown by Sandstrom et al. (1998) in another virtual water maze. In their task the location of the goal could be determined either relative to visual landmarks provided around the maze, or relative to the geometry of the room surrounding the maze. Males would use either landmarks or geometry as available whereas females tended to predominantly use the landmarks, thus performing worse when the landmarks were removed. In looking at gender differences, the use of VR itself tends to introduce a confounding factor: males tending to be more familiar with navigation in VR by virtue of having spent more time playing video games. Interestingly, Viaud-Delmon et al. (1998) found no difference in ability to estimate the amplitude of a passive turn while blindfolded, but did find males more likely to recalibrate their estimations after receiving distorted visual feedback while being turned in VR.

6. Neural Aspects of Navigation in Virtual Space

6.1 Background

Navigation requires many different processes, such as generation of body movements, recognition of landmarks, sense of direction, memory for routes and memory for survey-type knowledge. Thus disentangling the neural correlates of navigation will not be easy. Before considering the uses of VR in this pursuit, we briefly summarize the background to this area.

Electrophysiological recording in awake behaving animals has revealed several types of representation of spatial location in the brain. Neurons in posterior parietal areas have been found to represent the location of stimuli or actions in reference frames related to parts of the body ('egocentric' reference frames). For example, the firing rates of neurons in the posterior parietal cortex of monkeys may reflect the location of a stimulus relative to its eye or head or body. They also appear to encode the information necessary to translate between reference frames (e.g., given a location relative to the eye, computing that location relative to the head. See Andersen et al. 1985). Of particular interest with respect to navigation, 'place cells' in the rat hippocampus represent the rat's current location with respect to its environment (an 'allocentric' reference frame (e.g., O'Keefe and Nadel 1978). These cells become active whenever the rat is in a particular portion of an open environment, independent of the rat's orientation. The complementary representation, of heading direction independent of location ('head-direction cells'), was found in an area close to the hippocampus, the presubiculum.

Consistent with the finding of place cells, and the idea that the hippocampus maintains a survey-type representation or 'cognitive map' of the environment (O'Keefe and Nadel 1978), lesions to the hippocampus

impair a rat's ability to navigate to the hidden platform in the Morris water maze (Morris et al. 1982).

The egocentric parietal and allocentric hippocampal systems undoubtedly work together, not least because all sensory information enters the brain in an egocentric form and the read-out of an allocentric map must at some stage be translated into egocentric motor commands such as turning left or right. Indeed, there is evidence that the firing of neurons in area 7a of parietal cortex (the part that projects to parahippocampal cortex) also reflects stimulus locations relative to the monkey's environment (Snyder et al. 1998). The complementary roles of these two brain regions may also be understood in terms of the time scales over which they support behavior. The hippocampus is involved in long-term memory, whereas the parietal lobe is involved in perception and action over short time scales. It makes sense to use an allocentric representation to store spatial locations over the long term, as the position of the body will have changed before the information is to be used. However, to act on a spatial location (e.g., to reach with the hand or look with the eye) an egocentric representation (e.g., the position relative to the hand or eye) is more useful (see Burgess et al. 1999).

The rat's place cell representation of space appears to be derived from an appreciation of the relative positions of the walls of its enclosure (O'Keefe and Burgess 1996), presumably represented in the parahippocampal cortices that provide the input to the hippocampus. Functional neuroimaging in humans has also implicated parahippocampal cortex in the perception of space generated by visual stimuli such as enclosing walls (Epstein and Kanwisher 1998). Lesions to the medial temporal lobe (particularly those including parahippocampal cortex) can impair the recognition of landmarks or scenes (see Farrell 1996, Aguirre and D'Esposito 1999). Finally, we should note that navigation to a place can be contrasted with simply learning a sequence of body turns. In rats, learning to go to a place is often learned rather quickly, such that the rat will approach the learned place whether this requires a left or right body turn. However, after several trials of making a particular turn, the turn response will tend to predominate: the rat will make that turn whichever place it leads to. Interestingly, while hippocampal lesions tend to impair place navigation, lesions to the caudate nucleus (part of the basal ganglia) prevent development of a predominant body-turn response (Packard and McGaugh 1996), consistent with a role in storing well learned routes.

6.2 VR and Functional Neuroimaging

The recent advent of functional neuroimaging provides a new opportunity for VR in investigating the neural basis of behavior. This technique involves measuring some metabolic correlate of brain activity while the subject performs various cognitive tasks. The subject's head must be kept still during this process, which presents a problem for investigating many real-world behaviors, including navigation. However, VR can allow a realistic impression of navigation via a screen in front of the eyes and a keypad or joystick by the side, even while the head is fixed. An obvious limitation of this technique is that it excludes the proprioceptive and vestibular inputs experienced in real navigation.

There have now been several studies combining VR and functional neuroimaging, most making use of the development of realistic first-person perspective computer games. These have provided a cheap and effective means of producing VR displays in modifiable environments rendered at higher frame-rates and with a richer selection of textures than all but the most expensive commercial VR packages. Aguirre et al. (1996) created a simple maze of five dungeon-like corridors, each identified by an object placed at the end. This was used with fMRI to scan subjects while they first explored the maze and then navigated from the ends of four of the corridors to the end of the fifth. Their main finding was greater activation of bilateral parahippocampal areas associated with both exploration and navigation relative to a control condition of moving down an endless corridor. Maguire et al. (1998a) found activation of the right parahippocampus when subjects were scanned using PET as they explored a maze of similarly simple form, containing several rich textures and objects, relative to a control condition of viewing static and moving textures. Interestingly, no such activation was found when subjects explored a featureless maze without rich textures or objects. This difference may arise from neural involvement in encoding of object and texture information within this task, or from the greater level of immersion generated by the extra detail. Aguirre and D'Esposito (1997) also showed the parahippocampus to be involved in recognition of a location within a virtual town and indicating the direction to one location from another. Inferior parietal areas were also activated in these tasks, with relatively more parahippocampal activation in the place recognition task and relatively more parietal activation in the direction task (consistent with a role in translating between allocentric and egocentric reference frames).

Maguire et al. (1998b) directly investigated the brain regions involved in navigation by making subjects find their way between locations within a complex texture-rich VR town while in a PET scanner. This town was created to appear as lifelike as possible and to include many different possible routes between any two locations. The right parahippocampus and hippocampus were activated by navigation relative to following a route of arrows through the town. The subject's accuracy of navigation was found to correlate with activation of the right hippocampus and the right

inferior parietal cortex. Again, these results are consistent with the recall of allocentric information from long-term storage in the hippocampus, followed by translation into an egocentric representation of movement direction. Interestingly, speed of virtual movement correlated with activation in the caudate nucleus, and performance of novel detours caused activation of frontal areas (often associated with planning).

A recent study by Grön et al. (2000) used a VR maze in conjunction with fMRI to investigate the neural bases for such differences. Their results are interesting in that they show different brain areas active in men and women as well as different performance. Both groups showed parietal, parahippocampal and right hippocampal activation associated with navigation. However, in men, the left hippocampus was also activated, whereas in women right parietal and right prefrontal areas were activated. The authors suggest that the differences reflect a gender bias in strategy, with women relying more heavily on landmarks, and men using geometric information as well.

Finally, another recent study used VR in combination with electroencephalography of the human brain. One of the most striking signals in the electroencephalogram (EEG) of rats is the appearance of a roughly sinusoidal oscillation of around 6 to 10 Hz (known as the 'theta' rhythm, Vanderwolf 1969) whenever the rat is involved in displacement movements of the head. However, this rhythm had not been reported in humans. Kahana et al. (1999) recorded EEG from subdural electrodes in epileptic patients while moving through a virtual space. They reported finding episodes of theta rhythm during this virtual movement, perhaps providing a link between the electrophysiological mechanisms of rats and humans in spatial navigation.

7. Conclusions

The increasing openness and sophistication of commercial VR software and computer games such as Quake III and Unreal open the possibility of further experiments involving good immersion and an ability to make subtle manipulations of the environmental space. This will allow the design of behavioral and functional neuroimaging paradigms that address our processing of large-scale space in ways not possible in reality. Thus, we foresee a bright future for the application of VR in elaborating the neural basis of every day tasks such as spatial navigation.

See also: Navigation; Navigation Maps; Navigation in Spatial Environments, Neural Basis of; Neural Representations of Direction (Head Direction Cells); Spatial Learning and Memory, Neural Basis of; Virtual Reality; Virtual Reality, Psychology of

Bibliography

Aginsky V, Harris C L, Rensink R A, Beusmans J M 1997 Two strategies for learning a route in a driving simulator. *Journal of Environmental Psychology* **17**: 317–31

Aguirre G K, D'Esposito M 1997 Environmental knowledge is subserved by separable dorsal/ventral neural areas. *Journal of Neuroscience* **17**(7): 2512–18

Aguirre G K, D'Esposito M 1999 Topographical disorientation: a synthesis and taxonomy. *Brain* **122**: 1613–28

Aguirre G K, Detre J A, Alsop D C, D'Esposito M 1996 The parahippocampus subserves topographical learning in man. *Cerebral Cortex* **6**(6): 823–9

Andersen R A, Essick G K, Siegel R M 1985 The encoding of spatial location by posterior parietal neurons. *Science* **230**: 456–8

Astur R S, Ortiz M L, Sutherland R J 1998 A characterization of performance by men and women in a virtual Morris water task: A large and reliable sex difference. *Behavioural Brain Research* **93**(1–2): 185–90

Brooks B M, McNeil J E, Rose F D, Greenwood R J, Attree E A, Leadbetter A G 1999 Route learning in a case of amnesia: The efficacy of training in a virtual environment. *Neuropsychological Rehabilitation* **9**(1): 63–76

Burgess N, Jeffery K J, O'Keefe J (eds.) 1999 *The Hippocampal and Parietal Foundations of Spatial Cognition*. Oxford University Press, Oxford, UK

Chance S S, Gaunet F, Beall A C, Loomis J M 1998 Locomotion mode affects the updating of objects encountered during travel: The contribution of vestibular and proprioceptive inputs to path integration. *Presence* **7**: 168–78

Colle H A, Reid G B 1998 The room effect: Metric spatial knowledge of local and separated regions. *Presence* **7**: 116–28

Decety J, Jeannerod M 1995 Mentally simulated movements in virtual reality: Does Fitts's law hold in motor imagery? *Behavioural Brain Research* **72**: 127–34

Epstein R, Kanwisher N 1998 A cortical representation of the local visual environment. *Nature* **392**(6676): 598–601

Farrell M J 1996 Topographical disorientation. *Neurocase* **2**: 509–20

Ghaem O, Melle E, Crivello F, Tzourio N, Mazoyer B, Berthoz A, Denis M 1997 Mental navigation along memorized routes activates the hippocampus, precuneus, and insula. *Neuroreport* **8**(3): 739–44

Gillner S, Mallot H 1998 Navigation and acquisition of spatial knowledge in a virtual maze. *Journal of Cognitive Neuroscience* **10**: 445–63

Grön G, Wunderlich A P, Spitzer M, Tomczak R, Riepe M W 2000 Brain activation during human navigation: Gender-different neural networks as substrate of performance. *Nature Neuroscience* **3**: 404–8

Henry D, Furness T 1993 Spatial perception in virtual environments: Evaluating an architectural application In: *IEEE, 1363*, S. 33–40

Kahana M J, Sekuler R, Caplan J B, Kirschen J M, Madsen J R 1999 Human theta oscillations exhibit task dependence during virtual maze navigation. *Nature* **399**: 781–4

Klatzky R L, Loomis J M, Beall A C, Chance S S, Golledge R G 1998 Spatial updating of self-position and orientation during real, imagined, and virtual locomotion. *Psychological Science* **9**: 293–8

Maguire E A, Burgess N, Donnett J G, O'Keefe J, Frith C D 1998a Knowing where things are parahippocampal in-

volvement in encoding object locations in virtual large-scale space. *Journal of Cognitive Neuroscience* **10**: 61–76
Maguire E A, Burgess N, Donnett J G, Frackowiak R S, Frith C D, O'Keefe J 1998b Knowing where and getting there: a human navigation network. *Science* **280**(5365): 921–4
Moffat S D, Hampson E, Hatzipantelis M 1998 Navigation in a 'virtual' maze: Sex differences and correlation with psychometric measures of spatial ability in humans. *Evolution and Human Behavior* **19**: 73–87
Morris R G M, Garrud P, Rowlings J N P, O'Keefe J 1982 Place navigation impaired rot, with hippocampal lesions. *Nature* **297**: 681–83
O'Keefe J, Nadel L 1978 *The Hippocampus as a Cognitive Map*. Clarendon Press, Oxford, UK
O'Keefe J, Burgess N 1996 Geometric determinants of the place fields of hippocampal neurons. *Nature* **381**(6581): 425–8
Packard M G, McGaugh J L 1996 Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory* **65**(1): 65–72
Ruddle R A (in press) Navigating overlapping virtual worlds: Arriving in one place and finding that you're somewhere else. In: Freksa C, Brauer W, Habel C, Wender K F (eds.) *Lecture Notes in Computer Science*. Springer
Ruddle R A, Payne S J, Jones D M 1997 Navigating buildings in 'desk-top' virtual environments: Experimental investigations using extended navigational experience. *Journal of Experimental Psychology: Applied* **3**: 143–59
Ruddle R A, Payne S J, Jones D M 1998 Navigating large-scale 'desk-top' virtual buildings: effects of orientation aids and familiarity. *Presence* **7**: 179–92
Sandstrom N J, Kaufman J, Huettel S A 1998 Males and females use different distal cues in a virtual environment navigation task. *Cognitive Brain Research* **6**: 351–60
Siegel A W, White S H 1975 The development of spatial representations of large scale environments. In: Reese H (ed.) *Advances in Child Development and Behavior*. Academic Press, New York, Vol. 10, pp. 10–55
Silverman I, Eals M 1992 Sex differences in spatial abilities: Evolutionary theory and data. In: Cosmides L, Tooby J (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, New York, pp. 487–503
Snyder L H, Grieve K L, Brotchie P, Andersen R A 1998 Separate body- and world-referenced representations of visual space in parietal cortex. *Nature* **394**(6696): 887–91
Thorndyke P W, Hayes-Roth B 1982 Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology* **14**: 560–89
Vanderwolf C H 1969 Hippocampal electrical activity and voluntary movement in the rat. EEG. *Clinical Neurophysiology* **26**: 407–18
Viaud-Delmon I, Ivanenko Y P, Berthoz A, Jouvent R 1998 Sex, lies and virtual reality. *Nature Neuroscience* **1**(1): 15–16
Waller D, Hunt W, Knapp D 1998 The transfer of spatial knowledge in virtual environment training. *Presence* **7**: 129–43
Witmer B G, Bailey J H, Knerr B W, Parsons K C 1996 Virtual spaces and real world places: Transfer of route knowledge. *International Journal of Human Computer Studies* **45**: 413–28
Witmer B G, Klein P B 1998 Judging perceived and traversed distance in virtual environments. *Presence* **7**: 144–67

N. Burgess and J. A. King

Nazi Law

1. Basic Principles

The basic Nazi principles of the organization of law and state can be characterized by an ambivalent juxtaposition of the leader principle (*Führerprinzip*) and a racial–ethnic substance of the people's community (*Volksgemeinschaft*). On the one hand we find the 'Führer' as the head of a unitary, authoritarian, and militaristic state whose decisions are law. On the other hand we have the homogeneous nation of the German people. On the political level there is no separation of powers, no federal structure, only one party; no democratic, parliamentary, and pluralistic struggle for power. The leader acts beyond the ligatures of the rule of law. On the societal level class struggle is replaced by racial integration. Foreign elements—non-Aryans, political enemies such as communists and social democrats, asocials, criminals, etc.—are excluded by legal measures, by eugenic procedures, or by the arbitrary power of the police and the SS. The common good is authoritatively defined according to the 'sane feeling of the people' (*gesundes Volksempfinden*). The distinction between state and society, the boundaries between public and private, are dissolved by the totalitarian equalization and unification (*Gleichschaltung*) of intermediary organizations such as parties, trade unions, the media, universities, arts, etc. (only the churches could maintain some kind of relative autonomy). This system of paradoxical 'substantial decisionism' perished in the horrible collapse after its belligerent expansion.

2. Legislation

On the ideological level of legal philosophers legislation was understood—and in fact it was a misunderstanding of Savigny's notion of the *Volksgeist*—as the 'expression' of the popular substance, i.e., of a homogeneous *Volksgemeinschaft*. But how did the 'law in the books' come into the books in the first place? Legislative products, even during the Nazi era (1933–45), constituted a compromise of conflicting forces. So one has to look behind propaganda and find out how they came into being. But there exist only a few studies on the origin of Nazi legislation, e.g., in the field of marriage law and labor law.

The mere counting of the number and size of legal regulations shows that in Germany during the twentieth century the absolute maximum of legislative regulation—as far as the number is concerned—was reached in 1938 and 1939 (mainly measures to prepare for a war economy); a finding that might be of interest for those who talk about today's 'regulation flood.'

This vast amount indicates an instrumental use of norms, not an 'expression' of a *völkische* (racial–ethnic) will. As far as the forms of regulation are concerned, we find a variety of norms ranging from traditional laws to single orders of the Führer, even unpublished ones, sometimes circulated via radio broadcasts.

2.1 Public Law

In public law it was for several years an open question whether the Weimar constitution was still valid. Although it was never formally invalidated it was displaced by new 'basic laws,' including the program of the Nazi Party (NSDAP). Basic rights were eliminated through the *Verordnung des Reichspräsidenten zum Schutz von Volk und Staat* of February 28, 1933. There was no separation of powers after the *Ermächtigungsgesetz* of March 24, 1933, which empowered the government or single ministries and other authorities to issue laws. The federal structure of the Weimar Republic was destroyed. Political power was centralized in the Reich authorities with Hitler on top as the Führer and Reichskanzler in one person after August 1, 1934. The will of the Führer was the supreme and unlimited source of law. His will could be manifested in a variety of norms (*Gesetz*, *Verordnung*, *Erlaß*, *Einzelbefehl*, *Gesamtauftrag*, etc.). On the level inferior to the Führer there existed no clear competences, but rather a structural chaos that has been labeled a polycratic regime.

2.2 Penal Law

In penal law the principle of *nulla poena sine lege* (no punishment without a law) was abandoned by introducing the 'analogous' application of the law according to the *gesundes Volksempfinden*, and replaced by the principle of *nullum crimen sine poena* (no crime without a punishment). Various retroactive laws were issued. The penal system no longer focused on the criminal act defined as an offense against the *Volksgemeinschaft*. Instead the 'types' of the criminal (*Tätertypen*) became important. There was the *Gewohnheitsverbrecher*, i.e., the habitual criminal against whom the Nazis introduced the 'three strikes rule'; the violent criminal (*Gewaltverbrecher*); the murderer as a distinct type (*Mörder*); and the 'people's blight,' i.e., a person like a pestilential insect that has to be exterminated (*Volksschädling*). Instrumental use of penal law such as the *Kriegssonderstrafrechtsverordnung* (1939), or in its extreme instance the *Polenstrafrechtsverordnung* (1941), becomes normatively unlimited. It is characterized by extremely vague norms that confer broad judicial discretion, and by a terroristic severity of sanctions.

2.3 Civil Law

In civil law there were only minor changes in the legislation, except in agricultural law and in family law. Using old rules the judiciary invented new doctrinal formulae that enabled the judges to intervene in the formerly private sphere of autonomous individuals by the interpretation of a contract as a *Gemeinschaftsverhältnis* (relationship of a community, a organic union). Through the 'Law ordering national labor' (*Gesetz zur Ordnung der nationalen Arbeit*) of January 20, 1934, the autonomy of trade unions and employer organizations to reach collective agreements was abolished. Strikes were illegal. The collective actors were integrated in the *Deutsche Arbeitsfront* and put under the control of state executors. The worker became a popular figure of Nazi propaganda, while the labor market, including wages and mobility, was intensively regulated long before World War II. In other spheres of the economy the state favored the interests of factions of big business and their trusts; state intervention increased during the war.

3. The Legal Staff

During the first years after 1933 court organization was not substantially changed. Special courts (*Sondergerichte*) (1933), the *Volksgerichtshof* (1934), and the *Reichskriegsgericht* (1936) were established. Under the new regime about 15 percent of judges and prosecutors were dismissed on political and/or racial grounds. For various reasons the caseload decreased. A substantial change took place in the form of *Verreichlichung*, i.e., the centralization of the court administration under the department of justice of the Reich that after 1935 controlled the jurisdiction of the courts. After 1936 the judiciary lost its competence in 'political' issues; mainly matters concerning the (secret) police. In an agreement with the ministry of justice in 1937 the Gestapo gained permission to torture ('*verschärfte Vernehmung*'). Step by step the courts were excluded from prosecuting and sentencing Jews, homosexuals, and other enemies of the German Volk. In an agreement between Thierack (Minister of Justice) and Himmler (Chief of the Police and of the SS, respectively) of September 1942, 'asocial elements,' Jews, gypsies, Russians, Ukrainians, and so on, had to be handed over from the prisons to Himmler's myrmidons.

All these measures were executed, accepted, or tolerated by members of the legal staff. A basic problem in dealing with Nazi law is how to explain the degree of conformity among judges, prosecutors, attorneys, and legal scholars. One explanation given after 1945 was that they acted on the basis of a 'positivistic' attitude by simply adhering to the law. Instead, it was a positive attitude to the new regime

that was hailed by the organizations of judges and prosecutors. Judges were overzealous in their duty, e.g., in dismissing Jews from their legal positions as employees or tenants. Faced with the atrocities against Jews and political opponents, they displayed a passive onlooker's attitude. Courts used racist arguments without being obliged to do so. They imposed excessively severe sentences: estimates of the number of executions, including those imposed by military courts, range from 20,000 to 50,000. There was almost no judicial opposition; instead there was widespread voluntary obedience.

4. *Legal Culture*

There exists very little evidence about 'legal culture' in the sense of knowledge and opinions about law among the populace during the Nazi era. Systematic examination of facts that illustrate people's attitudes towards legal norms, judgments, legal institutions, and professional groups does not reveal a homogeneous *Volksgemeinschaft*, but a variety of differing opinions and attitudes. Many findings, for example reports by the *Sicherheitsdienst* of the SS, indicate that a deeply rooted trust in the courts (and in the Nazi regime in general) persisted until 1945. Only a small fraction of social problems or conflicts reach courts; criminal prosecution and private invocation of the courts are extremely selective, so it is significant that a considerable proportion of the activities of the Gestapo were provoked by private informers.

5. *Normality and Terror*

It would not be appropriate to look at Nazism only in light of the holocaust, without considering everyday law and the administration of justice. It is important not to neglect the routine work, the 'normality' of the judicial system with its millions of cases and unpublished decisions. A high degree of normality, of daily routine, is indispensable in such a system in order to secure the conformity of the legal staff, while continuous intervention would lead to a feeling of uncertainty.

In terms of legislative activity in Nazi Germany, a huge amount of routine work was done after 1933. There were not only the Nuremberg Laws (September 15, 1935), including the 'Law for the protection of German blood and German honor' (*Gesetz zum Schutz des deutschen Blutes und der deutschen Ehre*), and the 'Reich citizenship law' (*Reichsbürgergesetz*), but also regulations concerning chimney sweeps, seeds, and wine corks.

An essential feature of the Nazi regime was the combination of open terror and everyday routine. Legally 'normal' and terrorist measures alternated, and terrorism occurred under the guise of the 'normal.' Three events might serve to exemplify this:

(a) The murder of Röhm and his followers in June and July 1934, which was followed by doctrinal justification as an act of government self-defence.

(b) The elimination and extermination of persons of so-called 'inferior quality,' starting with the 'Law on the prevention of offspring afflicted with a hereditary disease' (*Gesetz zur Verhütung erbkranken Nachwuchses*) of July 14, 1933, introducing coercive sterilization, followed by the 'Law on the protection of the hereditary health of the German people' (*Gesetz zum Schutze der Erbgesundheit des deutschen Volkes*) of October 18, 1935 (interdiction of marriage in the case of certain diseases), and finally the transition to secret 'euthanasia' actions (authorized by Hitler on September 9, 1939), started in January 1940.

(c) The prosecution and extinction of Jews.

6. *Law and the Holocaust*

The legal persecution of Jews advanced step by step, interrupted and radicalized by brutal, direct measures. First came the 'Law for the reconstitution of officialdom' (*Gesetz zur Wiederherstellung des Berufsbeamtentums*) of April 7, 1933, excluding Jews—with a few exceptions—from the civil service. The concept 'Jew' was not defined according to criteria of 'race'; instead the religion of the grandparents was adopted as a criterion that could be used administratively. Step by legal step, the scope of social activities and the positions available to Jews were reduced. In September 1935 (through the Reich citizenship law) Jews were denied full citizenship. The new crime of '*Rassenschande*' (prohibition of sexual intercourse between Jews and non-Jews) was introduced (in the 'Law for the protection of German blood and German honor'). Legal 'aryanization' of Jewish property and exclusion from professions were pursued continuously by the legislators. The yellow star (*Judenstern*) was introduced in September 1941. In December 1941, Poles and Jews in the occupied territories in Eastern Europe were put under special penal regulations, with the threat of capital punishment for almost every kind of disobedience. After the January 1942 Wannsee Conference, at which the plans for the extermination of the Jews were formulated, their further persecution and extinction was by no means carried out extralegally, but in a legal–bureaucratic manner. From July 1, 1943, crimes committed by Jews were persecuted and sanctioned by the police only, and no longer by courts. The last published directive of November 11, 1944, barred Jews in Berlin from using heated public rooms.

Apart from this kind of legally sanctioned persecution, there were repeated 'wild' actions, starting with the antisemitic excesses in March and April 1933.

Such persecutions reached their first climax in the antisemitic pogroms in November 1938, followed by a legal sequel (obligation to pay a tribute; collection of insurance benefits). The same pattern holds for the legal exclusion of other victims of Nazism: Sinti, Roma and other '*Fremdvölkische*,' homosexuals, criminals, the insane, asocials, communist, social democrats, Jehovah's Witnesses, etc.

Injustice did not arise all of a sudden, but gradually. The step by step character of the prosecution of people not belonging to the *Volksgemeinschaft* contributed to the high degree of conformity among the legal staff. By accepting the first steps the official becomes incapable of resisting the further steps that might seem unjust to him (there were almost no women among the legal staff).

7. *Hitler and the Law*

The remarks about law in Hitler's *Mein Kampf* indicate a purely instrumentalist understanding: law is deprived of any intrinsic value and only conceived as an instrument to promote the Aryan race, to suppress and exclude all kinds of enemies and to secure the life of the German people. A few deprecatory remarks made by Hitler about jurists are reported. Sometimes he intervened arbitrarily in court cases, most famously in the Schlitt case (in 1942), which he used to have himself declared—for the second time after the Röhm murders—the supreme judge. After 1942 he was no longer concerned with individual cases.

8. *Continuity and Discontinuity, Modernity and Regression*

The Nazi era must neither be seen as entirely continuous with or totally isolated from what came before and what happened afterwards, i.e., its Weimar prehistory and the coming to terms with Nazism after 1945. The moments of continuity include both personnel and substantive factors. Personnel issues include the continued activity—and loyalty—of most lawyers within the judiciary and among legal scholars before and after 1933 as well as (at least in the Western zones of Germany) after 1945. Substantively, in the legal domain the Nazis and their legal lackeys were not very inventive. With the possible exception of agricultural law—and after 1939 the regulations concerning the war economy—they retained the traditional legal corpus. The vast majority of legal norms remained valid after 1933, but they were given new interpretations. In large measure they continued trends that had been evident before 1933, e.g., in penal law. With regard to eugenics, too, the Nazis referred back to much older concepts and also claimed to rely on similar practice abroad, e.g., in the USA and Scandinavia. In civil law, too, one can find continuity, e.g., in the so-called 'materialization' tendencies in private law, the limitating of contractual autonomy.

In some areas, the Nazi period achieved solutions for old disputes (e.g., in the Corporation Act (*Aktiengesetz*) of 1937). From a wider historical perspective some reforms can be interpreted as 'progressive': protective regulations concerning pregnancy and motherhood as well as juvenile workers, regulation of working hours, etc. In divorce law (1938), the grounds for divorce were broadened from the guilt of one spouse to include marriage breakdown (this again was an idea dating from before 1933). An innovation in legal doctrine was E. Forsthoff's concept (introduced in 1938) of '*Leistungsverwaltung*,' reflecting the change from the traditional intervention state (*Eingriffsverwaltung*) to a state administration that provides, in addition, welfare services and infrastructure.

The mixture of normality and terror, together with strong element of continuity, helps explain the high degree of conformity among the legal staff as well as among the general populace.

9. *Justifications*

All of the Nazi regime's official legal measures—starting from the *Machtergreifung* between January and March 1933—were justified by legal scholars, in legal philosophy as well as in legal doctrine. C. Schmitt, for example, justified the Röhm murders in 1934; K. Larenz and others provided reasons for legal discrimination against Jews. It was no problem to accept eugenics, the presidents of the regional appeal courts and the chief public prosecutors were informed by the ministry of justice about the 'euthanasia' program in April 1941; in turn, they showed no opposition. For the excessive use of the death penalty one could refer to the saying that hard times require hard measures.

10. *Was it Law?*

In general, we may doubt—from a legal–philosophical point of view—whether Nazi law was law at all, whether there existed a legal system under the Nazis. It is not necessary to rely on natural law in order to deny the term law to what the Nazis called law, as did Radbruch after 1945. Ernst Fraenkel (Fraenkel 1941) called the regime as a 'dual state,' combining 'normative state' and a 'prerogative state.' The normative state, i.e., traditional state agencies, in particular the courts, acted 'according to the law' in dealing with nonpolitical issues, while the prerogative state, the

'organs of dictatorship,' acted 'according to given conditions' in what they defined as political matters. While the normative state contributed to the rational calculability required in a capitalist society, the prerogative state aimed at the oppressive creation of the *Volksgemeinschaft*, the dissolution of the institutions of the working class and the persecution of political foes. Fraenkel finished his manuscript in 1938 before he left Germany. In American exile Franz Neumann (Neumann 1942/44) came to the conclusion that Nazi Germany was a 'non-state.' Conflicts between various power groups, i.e., parts of the state administration, party cliques, big business interests, the military forces, and the SS–police complex resulted in a chaos of competences and a polycentric dissolution of legality as the mode of state organization.

In his pragmatic jurisprudence Lon Fuller (Fuller 1969, p. 39) describes 'eight distinct routes to legal disaster': no rules at all but only *ad hoc* decisions; failure to publicize; retroactive legislation; failure to make rules understandable (e.g., extremely vague rules); contradictory rules; rules that require conduct beyond the powers of the affected party; too frequent changes in the rules; no congruence between the rules as announced and their actual administration. According to Fuller's 'eight ways to fail to make law,' the Nazis did not create a bad system of law; instead, their attempts resulted in 'something that is not properly called a legal system at all' (Fuller 1969, p. 33).

See also: Anti-Semitism; Compliance and Obedience: Legal; Dictatorship; Dictatorship in History; Holocaust, The; Judaism and Gender; National Socialism and Fascism; Rechtsstaat (Rule of Law: German Perspective); Rule of Law; Totalitarianism; Totalitarianism: Impact on Social Thought

Bibliography

Fraenkel E 1941 *The Dual State: A Contribution to the Theory of Dictatorship*. Oxford University Press, New York (original German version *Der Doppelstaat*. In: Fraenkel E 1999 *Gesammelte Schriften Band 2*. Nomos, Baden-Baden, Germany)

Fuller L L 1969 *The Morality of Law*, rev. edn. Yale University Press, New Haven, CT

Gruchmann L 1988 *Justiz im Dritten Reich 1933–1940. Anpassung und Unterwerfung in der Ära Gürtner*. Oldenbourg, Munich, Germany

Hirsch M, Majer D, Meinck J (eds.) 1984 *Recht, Verwaltung und Justiz im Nationalsozialismus*. Bund, Cologne, Germany

Müller I 1987 *Furchtbare Juristen. Die unbewältigte Vergangenheit unserer Justiz*. Kindler, Munich, Germany (trans. into English, 1991 *Hitler's Justice: The Courts of the Third Reich*. Harvard University Press, Boston)

Neumann F 1942/44 *Behemoth. The Structure and Practice of National Socialism 1933–1944*. Oxford University Press, New York

Rüthers B 1989 *Entartetes Recht*. Beck, Munich, Germany

Stolleis M 1998 *The Law under the Swastika*. University of Chicago Press, Chicago

H. Rottleuthner

Near Middle East/North African Studies: Culture

Culture in the postcolonial Middle East and North Africa is the ensemble of literary and artistic production that reflects and shapes the norms and values of daily life. Fiction, drama, film making, and art play a greater role in the people's lives than they do in Western Europe and the United States. Because they are often viewed as spokespersons for the people to the regime, public intellectuals have a moral authority that promises influence, imposes responsibility, but also invites manipulation by those in power. In this article the evolving relationship between cultural producers, society, and the state is examined.

1. The Impurity of Postcolonial Authenticity

The Middle East and North Africa (MENA) have come to be conjoined as the Arab world, with Turkey, Iran and, more recently, Israel as non-Arab neighbors. Scholars have traditionally treated the region as a separate and more or less isolated block, yet it never was. What is significant about this area is that it frames the southern and eastern shores of the Mediterranean. It is thus geographically, historically, and culturally connected with Europe. Edward Said argues convincingly in his influential *Orientalism* (1978) that eighteenth and nineteenth century military and political incursions from the northern rim were accompanied by an equally violent discursive campaign. Apparently objective orientalist scholarship in most disciplines in fact served the political ends of their governments and shaped domestic practices and policies. Portrayals of MENA as a landscape empty of people, sexually available and civilizationally fertile if carefully husbanded, have come to be understood as colonizing projections. Yet orientalists' analyses had the force of truth at the time, and they were remarkably influential. Arab, Turkish, and Iranian intellectuals writing during the nineteenth and early twentieth centuries often built on such positivist foundations. 'Progress' became the *leitmotiv*. Failure to construct an active national subject was predicated on lack of knowledge and the inability to borrow creatively from the West.

Contact with the cultures of Western Europe in the modern period came through various channels. French, British, and Italian colonizers brought new technology and new ideas which highlighted the

difference between advanced cultures and those which world processes were bypassing. MENA rulers sent scholarly delegations to Europe to study science and to translate key documents. Exposure to French, English, Russian, and even American literature during protracted stays in the European capitals changed intellectuals' notions of the writing and function of literature: it should reflect the new realities in the people's lives. By the early twentieth century, MENA writers were beginning to publish their own works of fiction and drama. Egypt provided the regional milieu for literary experimentation and excitement. Literary salons convened men and sometimes even women to discuss the new trends in politics and literature. At a time when middle and upper class women were confined to their homes, these gatherings of intellectuals were almost the only places where women could appear with men.

The 150 years stretching from the French occupation of Egypt at the end of the eighteenth century until World War II, known in the Arab world as the *nahda* or renaissance, has been the subject of much research by those within the region. More recently, scholars of colonialism and postcolonialism, many of whom come from MENA but have been educated elsewhere, have subjected this period of initiation into a Western-style modernity to careful scrutiny. Their insider–outsider perspective provides a new lens through which to analyze the texts of orientalists like the linguist William James and to view the process of Westernization.

After World War II, most MENA countries achieved their independence. With the exception of Palestine, which was transformed into the home of Jews who had been flooding into the region out of fascist Europe, liberation struggles gave birth to autonomous regimes throughout the region, but always under the watchful eye of the former rulers and in the shadow of globalization, a new form of US-centered imperialism.

The historical fluidity of connections across the Mediterranean fostered persisting ties and facilitated the formation of new ones from further afield. Analysts from with the region articulated this moral tug-of-war with Western Europe and the US existentially; questions concerning authenticity, modernity, and tradition became central to the assertion of a new identity. However, the dream of a postcolonial return to an authentic, pre-European past was shattered by the reality of independence. MENA intellectuals came to the realization that cultural authenticity was not a pure haven but rather a present reality that involved an acknowledgment of all the strands of past experience, however painful.

2. *Westtoxification*

European imperial rule had introduced European culture and its values and, like a poison, it was coursing through the veins of the people. How to find an antidote? During the 1930s, some writers began to look for an indigenous, premodern greatness that might rival the cultures and power of the Europeans and thereby instill pride in their people. The Pharaonic, Babylonian, Ugaritic, and more recent Ottoman civilizations enjoyed a cultural revival in the 1930s and 1940s. Intellectuals questioned where MENA fitted in the modern world. In *The Future of Culture in Egypt* (1938), the Egyptian writer and educator Taha Husayn (1889–1973) connected his native land with a Mediterranean future. Arab critical schools such as Diwan (1920s) and Apollo (1930s) brought together poets and critics schooled in England and France who saw the need to knit Western theory together with local literary traditions. The publications of these schools produced a new space, that of the public sphere. Discussions about national identity and the new roles women and men should play in the future became matters of general concern.

European scholarship has made much of this period of intellectual ferment which many tried to organize through clear periodization according to genre categories such as neoclassical, romantic, and sociorealist. Some MENA intellectuals turned their attention to the negative effects of rapid change. The Saudi exile Abd al-Rahman Munif (b. 1933), who now lives in Syria, situated his five-volume novel *Cities of Salt* (1980s) in the recently oil-rich desert states. A petroleum economist, he understood the environmental and demographic impact of oil. The lure of easy profit has displaced tribal solidarity with individual enterprise. This multivolume book earned its author the displeasure of his Saudi government which deprived him of citizenship.

Then, there are those who absolutely reject Western cultural influence. The Iranian writer Al-e Ahmad (1923–69), the son of a Muslim Shiite cleric, coined the term 'westtoxification.' The title of a book he wrote in 1962, it was used as a slogan against the Shah. For him, as for many of his contemporaries, the cultural, economic, and political subordination of MENA countries to the West was anathema. The future had to be modern and Islamic, it must not be Westernized. It was this anti-occidentalism that led to the Islamic Revolution of 1978 (Talatoff 2000). Scholarship of this period spanning the independence struggles and the Islamic Revolution has emphasized growing self-confidence and identification with movements opposed to EuroAmerican neocolonialism worldwide.

3. *War Stories*

Conflict and violence are so much a part of MENA history and culture that no consideration of its culture can afford to ignore them. The Palestinian struggle against Israeli hegemony since 1948, the Algerian war of independence (1954–62), the Lebanese civil war

(1975–92), the Iranian Revolution (1978–9), the Iran–Iraq war (1980–8), the Gulf War (1991) are only the best known of the MENA conflagrations. Each war produced its own coherent body of literature with women as actively engaged as the men. Women's versions of their experiences of national violence have challenged the familiar frame narrative that characterizes men's war stories. They have insisted on the importance of their contributions to the national welfare. In the 1980s, the Algerian novelist and historian Assia Djebar (b. 1938) started to publish her semi-autobiographical quartet. Through the eyes of women she radically revises Algerian colonial history, linking the stories of neglected nineteenth century women resisters with those of living women who had been active and then silenced during the war of independence from the French, with her own story. The recent attacks by Islamic fundamentalists on intellectuals, and particularly women, in the Algerian capital seem to perpetuate a war that was not resolved in the early 1960s.

In Lebanon, the Beirut Decentrists, a school of women writers, wrote about the civil war. Some analysts have argued that political anarchy relaxed social norms so that the prohibition of women writing about war, an experience they have always been said not to have had, was lifted. Beirut became emblematic of postmodern wars that erase the boundaries between binaries, like home and front, that have been so necessary to the effective telling of the War Story (Cooke 1997). Critical of those who participated in the senseless fighting, they also exposed the literary recklessness of those who wanted to represent the war as having any moral sense. Emily Nasrallah (b. 1938) wrote about the devastating effects of the civil war and the massive emigration it precipitated. Hanan al-Shaykh's (b. 1941) *The Story of Zahra* (1980) traces the growth in consciousness of a woman whom the war empowers. She fights in her own limited but intense way to do something to end at least part of the madness.

While the war in Lebanon was at its height, the Iran–Iraq war broke out. Scholars from the region as well as those from outside have pointed out that Saddam Hussein, like many dictators, has had an ambivalent relationship with the intellectual elite of his country. He needed them and therefore he feared them. Throughout the war, he coerced the writers and artists to glorify the war in ink and paint. The Ministry of Culture established literary series and organized festivals in honor of the war to which scholars of Arabic literature from all over the world were invited. Many Iraqi writers and artists did as they were told. Some did not. Under the eye of the censor, some writers managed to articulate criticism of a war they were supposed to praise (Cooke 1997). Not only writers and artists, but literary critics also were commissioned to describe and analyze the cultural effects and products of the war. So prolific were they that by the late 1980s books of criticism about the literary criticism of war literature were being published.

Ironically, the Islamic government of Iran was doing exactly the same as its enemy, namely sponsoring praise of the war and censuring whatever was seen to be oppositional to its specifically Islamic interests. The case of Salman Rushdie and the 1989 fatwa pronounced against him for his 'heretical' *Satanic Verses* (1988) has entered the annals of world history. Within Iran, some writers complied with the government's will; others protested. A. Rahmani's *Short Hike* acts out a moment of defiance between an Islamic guard and the woman he has just reprimanded for revealing a strand of hair. Restrictions forced women to speak out (Milani 1992). Many went into exile. London, Paris, Los Angeles, and New York have become MENA exile capitals. Observing their native lands from a distance, creative artists and critics are publishing works of political, social, and cultural analysis as well as fiction and memoirs.

4. *Moral Authority*

Political independence in MENA countries demanded political and moral vision that would owe as little as possible to European norms. Intellectuals set out to 'revolutionize the revolution.' In 1946, the Iran-Soviet Society sponsored the first congress of Iranian writers which called for the centrality of ideology to literature. Seven years later in Lebanon, the first editorial of the influential monthly periodical *Al-Adab* declared that all forms of writing must be politically engaged. The time of art for art's sake had passed to be replaced by a revolutionary ethic in all creative activities.

MENA writers, and more recently filmmakers, have a moral authority rarely found in the US and Western Europe. Since books, plays, paintings, and films may assume the force of a political manifesto, governments as well as nongovernmental associations such as Islamist groups try to monitor cultural production before, during, and after its completion. Intellectuals judged to be dissident may be eliminated, exiled, or co-opted. For most MENA intellectuals prison is a badge of honor, exile a terrible necessity, co-optation an ever-threatening possibility.

Co-optation is a widespread phenomenon in MENA, making arduous the decision to stay in one's country. Filmmaking is particularly susceptible. Like theater, film watching is a communal experience and as such it allows individuals to share with an anonymous group a sense of justified anger, or to laugh at the expense of oppressors. The power of such cathartic moments has not been lost on repressive regimes: national film institutes sponsor, censor, and retain monopoly of the finished product. Since the Islamic Revolution, Iranian filmmakers, like the internationally recognized Mohsen Makhmalbaf, have had to

submit their scripts, rushes, and final films to the government censors, who are especially concerned with Islamizing the representation of women. During the latter part of Hafiz Assad's rule, Syrian filmmakers were sponsored to produce films that were remarkably critical of the regime but that were rarely shown.

Films like Muhammad Malas' *The Night* (1991) are only shown during international film festivals. People know of their existence and long to watch them. In the meantime, the dissenting individual may be given a government post or some other form of visibility that demands a measure of criticism of the government that is sponsoring the criticism. Such 'commissioned critics' provide an outlet for the people's frustration and anger with a repressive regime, thereby helping to fashion a facade of democracy. In his *Historical Miniatures* (1993), the Syrian playwright Saadallah Wannus (1941–98) examines the triangular relationship between state, society, and public intellectuals through the controversial meeting of the fourteenth century Mongol tyrant Tamerlane with the North African historian Ibn Khaldun.

The project of cultural production from inside authoritarian regimes is fraught with danger: just how much criticism is too much, how much too little? Nevertheless, despite the dire consequences for articulating opposition, public intellectuals rarely flinch from what they consider to be their responsibility. Some are killed for their words. Many are imprisoned. So many have published books about their prison experiences that it is possible to talk about a body of MENA prison literature that is beginning to receive the critical attention it deserves (Abrahamian 1999, al-Musawi 1999). The fact that intellectuals have had such experiences and that they then publish them gives them a moral authority that others, whose writings may be as 'good,' may never achieve. Research on intellectuals' prison experiences and writings has tended to be conducted outside the country. The dictates of Aesopian language demand that the text not be decoded in the place where it is produced for fear of further political repercussions. Michel Foucault's *Discipline and Punish* (1977) has influenced scholarship on prison writing throughout the MENA region.

5. *Exile*

For some the cost of staying at home is too much to bear, and so they leave. Citizens of repressive regimes flee in order to find a place where they might breathe and speak freely. They have congregated in major cities like Paris, London, Buenos Aires, New York, and Los Angeles, where they have founded publishing houses and launched newspapers and magazines.

The Palestinian case has become emblematic of the dilemma of many MENA intellectuals: they cannot live at home, they scarcely survive outside. Should the committed intellectual fight or write or do both or neither? In 1963 Ghassan Kanafani (1936–72) published his *Men in the Sun*, which was turned into a film entitled *The Dupes*. This novella explores the possible outcome of the decision to leave one's home and to seek refuge in Arab countries. The fugitives suffer multiple humiliations before they perish on their way to Kuwait.

After the Arab defeat by the Israelis in 1967, Arab as well as Turkish and Iranian intellectuals tried to understand what had gone wrong. Scholarship in the region focused on a loss of spiritual purpose and guidance. Literary production echoed this existentialist trend. The Egyptian Nobel laureate Naguib Mahfouz immediately abandoned his long, socio-realist novels to write short stories and novellas that exposed the angst of a tortured hero in a kafkaesque world.

Some Palestinians tried to stay in Israel where they were promised that all who lived on the land of Israel and spoke its language were its citizens. Anton Shammas' (b. 1941) autobiographical *Arabesques*, hailed by some as a masterpiece of Hebrew prose, has not yet been accepted as part of the canon of Hebrew literature. No matter how hard they tried to assimilate, or how excellent their command of Hebrew, Palestinians could never become Israeli citizens in the full sense of the word.

In a world whose borders are increasingly porous, where migration is a necessary part of many people's lives, indigenous scholars, writers, and artists are imagining a new kind of cultural nationalism that allows those who are living far from their land of birth or ancestry to retain very real ties of identity. Literary associations in MENA are slowly recognizing that their writers may not live in the country that they claim as their nation.

6. *Globalization and Islamization*

Since the late 1980s there has been a shift in the relations between the people and MENA governments as global pressure forces an acknowledgment of human rights and civil society. Western governments have predicated aid on demonstrable attention to individual and group rights to life, freedom, and to nongovernmental political participation. Although the motivations of the human rights enforcers are often questionable, they have contributed to changing the norms of interaction between state and society.

Religion has come to play an increasing role in MENA societies as Islamist groups have assumed responsibility for social welfare from below. The outcome is not always benign. Islamist groups are struggling against their governments but also against secular intellectuals to impose an Islamic state. In

1990s Algeria, Islamists increasingly targeted intellectuals who they charged with secularism and perversion of the people's morals. In 1992, Egyptian fundamentalists assassinated Farag Foda, a journalist, for his critique of their discourse, especially in his *Before the Fall* (1992). In Iran, public intellectuals are under constant attack by the government and vigilante groups. Moral censure of literature has escalated to sanction for deadly assaults on the authors.

Although the most visible aspects of the Islamization of local institutions have been the violent reordering of society according to so-called Islamic norms and values, this is not the whole story. A corollary development has been a shift in public discourse. In countries whose regimes follow Muslim law, this discursive transformation has been imposed by fiat. Elsewhere Muslim ideologues have been the catalysts. Scholars, especially Western-trained anthropologists, have asserted that the wide distribution of cassette tapes of charismatic preachers' Friday sermons provide even the illiterate with a new vocabulary and syntax. Whether the Islamization of the public sphere is regarded as a good or a bad thing, it has created a new kind of public sociability that entails a mandate for public intellectuals, even the most secular of them, to learn its language.

The most radically affected by this development have been women, and particularly those with a feminist agenda. Feminists from within the region are analyzing the situation, forming contingent coalitions and proposing new ways to challenge the policing of women's actions, behavior, and speech. They have recognized the dangers inherent in Islamists' declarations that women are the symbolic and moral center of their society. What they do and how they act have become men's affairs. It is within this context that the veil, abandoned, or even banned in most places during the first half of the twentieth century, has re-assumed such central importance. In places like Afghanistan, Iran, Algeria, and Sudan, the unveiled woman may be reflexively labeled wanton. Although the veil is not Islamic, it is seen by many to be highly recommended. There are several problems with this highlighting of the veil as an emblem of the society's moral rectitude. Feminists in the late nineteenth and early twentieth century fought against the veil as a symbol of men's ownership of women. To go back to wearing this patriarchally loaded symbol is to renounce the gains of their grandmothers. Another problem with the blanket mandate that women veil is that not all women in MENA countries are Muslims.

Women scholars and activists from the region are responding to Islamist pressure. Even when they do not accept the label of feminist because of its overdetermined connection with Westernization, they are learning the rules of Islamic discourse and legal practice so as to challenge them effectively. In *The Veil and the Male Elite* (1987), the Moroccan sociologist Fatima Mernissi (b. 1938) reveals the weakness at the heart of the most trusted religious and legal texts. The Egyptian writer, physician, and activist Nawal El Saadawi (b. 1930) has called for the 'Lifting of the Veil from the Mind' in her Arab Women's Solidarity Association (founded 1982), in her lectures given all over the world, and in novels like *The Fall of the Imam* (1987), which earned her a place on the Islamists' death list. At the other end of the political spectrum are Islamist women like Zaynab al-Ghazali (b. 1912) whose prison memoirs, *Days from my Life* (1977), enact a model of women's empowerment within orthodox Islam. These Islamic feminists have learned how to respond to the silencing gestures of men in power by developing a multiple consciousness of how they are perceived, named, and controlled. From these discontinuous vantage points they are producing a multiple critique that keeps incompatible audiences in tension with each other in such a way that the speaking subject can hold on to her agency (see Cooke 2001).

7. *Translation and Recognition*

MENA literature, art, and films are now enjoying international recognition, as they enter the mainstream of world culture. In 1988, Mahfouz became the second MENA writer to win the Nobel Prize in literature having been preceded by the Israeli Shmuel Yosef Agnon (1888–1970) in 1966. In 1989, the Moroccan Tahar Benjelloun won the Prix Goncourt, the first Arab to be awarded this coveted French prize. In 1997, the Turk Yashar Kemal was awarded Italy's most prestigious prize, the Nonino Literary Award. In that same year four Iranian films were awarded major prizes at Cannes, Locarno, Tokyo, and Montreal.

Arabic, Hebrew, Persian, and Turkish fiction and poetry are being translated into European languages. In 1981, the Palestinian poet and literary critic Salma al-Khadra' al-Jayyusi founded her Project for Translation from Arabic (PROTA) in the United States. Whereas the first translations found homes in marginal presses dedicated to publishing works by writers from the global South, by the late 1980s major European, American, and even Japanese publishing houses started to adopt literary series as well as individual authors.

The growth of the translation industry is significant for the readers who now have unprecedented access to the cultural imaginary of people they had previously known through the stories of scholars and travelers only. It is important above all for the intellectuals who are writing with the awareness that their works will be translated, that what they say about their culture is being consumed far from the local readership they had originally targeted. Writing thus entails consequences that must be anticipated. MENA intellectuals are seeing their work becoming part of a global project in which they should play an increasingly visible role.

8. Conclusion

The legacy of the twentieth century in MENA countries is a whirlwind of conflicting realities and emotions: colonialism, independence, wars, spiritual search, religious self-assertion, despair, defiance, hope, capitulation, resistance, incorporation into the global system, and marginalization from its benefits. MENA cultural producers have had to shoulder a huge responsibility as they confront the contradictions marking their era and imagine different futures. Many have paid the price of articulating their conscience in the belief that sacrifice is necessary and may even be productive. Culture in the MENA region is a political arena where women and men strive to contest the actions, ideologies, and policies of oppressive regimes and world systems. It remains to be seen what will be the effect of the information revolution on MENA countries which have committed themselves to resisting Westernization and who now find themselves compelled to go on-line.

See also: Christianity Origins: Primitive and 'Western' History; Islamic Fundamentalism; Near Middle East/North African Studies: Politics; Near Middle East/North African Studies: Religion; Near Middle East/North African Studies: Society and History; Orientalism; Practice: Anthropological Aspects

Bibliography

Abrahamian A 1999 *Tortured Confessions. Prisons and Public Recantations in Modern Iran*. University of California Press, Berkeley, CA

Badawi M M (ed.) 1992 *Modern Arabic Literature*. Cambridge University Press, Cambridge, UK

Barakat H 1993 *The Arab World. Society, Culture, and State*. University of California Press, Berkeley, CA

Cooke M 1997 *Women and the War Story*. University of California Press, Berkeley, CA

Fanon F 1967 *Black Skin, White Masks*. Grove Press, New York

Halman T S (ed.) 1982 *Contemporary Turkish Literature*. Associated University Presses

Milani M 1992 *Veils and Words. The Emerging Voices of Iranian Women Writers*. Syracuse University Press, New York

Modern Turkish drama 1976 *An Anthology of Plays in Translation*. Bibliotheca Islamica

al-Musawi M J 1999 *Infirat al-'aqd al-muqaddas. Mun'atifat al-riwaya al-'arabiya ba'da Mahfuz* (*The Dissolution of the Sacred Contract. Trends in The Arabic Novel after Mahfuz*). Al-hay'a al-misriya al-'amma li al-kitab, Cairo

Said E 1978 *Orientalism*. Pantheon, New York

Talatof K 2000 *The Politics of Writing in Iran. A History of Modern Persian Literature*. Syracuse University Press, New York

Women claim Islam 2000 *Creating Islamic Feminism Through Literature*. Routledge

Yudkin L I 1984 *1948 and After: Aspects of Israeli Literature*. University of Manchester Press, Manchester, UK

M. Cooke

Near Middle East/North African Studies: Economics

The study of Middle East and North African economies is a relatively recent endeavor. Charles Issawi, the leading Middle East Economic historian, asserts that prior to World War II, the only Middle Eastern country whose economy had been adequately studied was Egypt (Issawi 1990). Since then the field has grown considerably but until the late 1970s, a handful of topics dominated scholarly research in the economics of the Middle East and North Africa. These included the economics of oil, the role of the state in the economy, and international labor migration, all of which are clearly interrelated. With the exception of North African migration flows to Europe, international migration flows in the region were a response to the sudden increase in labor demand in labor scarce but oil-rich countries during the oil boom era. Moreover, oil rents, which flow primarily into state coffers, centralize economic resources in the hands of the state and significantly increase its role in the economy. However, the dominant role of the state was not limited to oil-rich countries. The rise of nationalist governments throughout the region in the post-independence period resulted in significant state intervention, even in countries where oil has played a minimal role, such as Turkey, Syria, and Yemen.

Since the decline in oil prices in the 1980s and the consolidation of the so-called Washington consensus, the state-led development model pursued by most Middle Eastern and North African countries became increasingly untenable and movement toward greater economic liberalization and market-based reforms began in earnest. Accordingly, the focus of scholarly work in Middle East economics shifted to the study of economic reform and structural adjustment, privatization, and trade and financial liberalization, including the prospects for greater trade integration with the European Union through the Euro–Mediterranean partnerships. The 1993 Oslo accords and the prospects of settling the Arab–Israeli conflict brought forth a wave of research on the economics of Middle East peace, but interest in this topic fluctuates with the ebbs and flows of the Arab–Israeli conflict. Finally, the onset of the demographic transition in the region and the growing importance ascribed to human development in the international development literature has led to a renewed interest in the study of labor markets, education, human resource development, and gender.

1. The Economics of Oil

Oil continues to play a very important role in the region's economies, either directly through the sale of

oil and gas or indirectly through the receipt of labor remittances from migrant workers in oil-rich countries. Although oil-related revenues can potentially provide the necessary resources to finance infrastructural and educational investments, they can also cause significant economic distortions that could potentially hamper the long-term development of economies receiving such revenues. Since much of the revenue from the sale of oil accrues as rent, that is the difference between the product's market price and the opportunity costs of the resources that go into its production, they also give rise to rent-seeking behavior on the part of various social actors. When the rents accrue to the state, the focus of the political process becomes the allocation of the rent to various interest groups, rather than the creation of new wealth. Middle Eastern and North African states used their oil rents to intervene massively in their economies by setting up large-scale state-owned industries, providing highly subsidized services and consumer goods, instituting generous safety nets, expanding public employment schemes, and protecting local industries and workers from international competition (Karshenas 1990, Askari et al. 1997).

Besides its effect on the behavior of governments, the sudden increase in the flow of external resources in an economy can lead, simply through the working of market forces, to a phenomenon known as the 'Dutch disease,' in reference to the problems experienced by the Dutch manufacturing sector after the discovery of North Sea oil. The increase in demand brought about by the windfall leads to an increase in the price of non-tradables, such as construction and services, whose supply is constrained by domestic production capacity, relative to that of tradables, such as manufactured and agricultural products, which can be imported from abroad. This shift in relative prices leads to the stagnation and diminished international competitiveness of the non-oil traded goods sectors (Corden 1984, Collier and Gunning 1999). The Dutch disease is often blamed for the contraction of agriculture throughout the region in the 1970s and the region's increased dependence on food imports (Richards and Waterbury 1996). The effects of the Dutch disease on manufacturing were often counteracted with increased protection, making these manufacturing industries even less competitive internationally.

2. *International Labor Migration*

Research on international labor migration in the region was most active from the mid 1970s to the late 1980s, the period when migration flows to the oil-rich countries increased significantly. The primary focus of the research was the investigation of the determinants of migration and migrant remittances (Adams 1993, Wahba 1991, El-Sakka and McNabb 1999) and the assessment of the effects of international migration on the economies of sending countries (Fergany 1988, Adams 1991, Roy 1991). Arguments for the positive impact of migration and remittances include the obvious effects on foreign exchange earnings, as well as its effects on lowering unemployment rates, increasing per capita income and savings rates, and improved skill levels of returning migrants. Arguments against include its effects on skill shortages, increased conspicuous consumption, higher income inequality, and increased dependency on a risky source of foreign exchange. However, the evidence presented on either side of the arguments can generally be characterized as weak. With a few exceptions (Fergany 1988, Adams 1991, 1993), most studies rely on aggregate figures from official sources rather than microdata from household surveys.

3. *The Dominant Role of the Middle Eastern State*

Although oil clearly contributed to the expanding role of the state in the region's economies, the dominant role of the state in economic affairs preceded, in many countries, the oil boom era. Like states elsewhere in the developing world, the post-colonial Middle Eastern state took upon itself the task of inducing the structural transformation of the economy from agriculture to industry through direct state intervention. The accession to power of modernizing nationalist leaders, mostly with a military background, in Turkey, Egypt, Iraq, Syria, Libya, Tunisia, Yemen, and Algeria led to massive state intervention in the form of nationalization of large private firms as well as some entire sectors, such as utilities, banking, and insurance. Governments in the region also invested in heavy industry, erected high trade barriers, and sought to control agricultural prices and delivery systems. This state-led import substitution industrialization model was first instituted in Turkey, without an explicitly socialist ideology, and then adopted and expanded upon in a number of Arab countries and justified by a home-grown brand of socialism that came to be known as Arab Socialism (Richards and Waterbury 1996, Hopfinger 1996). An excellent comparison of Turkey's and Egypt's development trajectories during the twentieth century is offered by Hansen (1991).

The statist model of development resulted in some early successes in raising per capita income and achieving a structural transformation toward more industrialized economies. However, it soon ran into the usual bottlenecks that haunt import substitution industrialization regimes, namely the shortage of foreign exchange and excessively high capital-output ratios. The 1967 Arab–Israeli war diverted attention to the military buildup in Egypt and Syria and temporarily suspended the economic development agenda in these countries.

4. Structural Adjustment and Economic Liberalization

The first and second oil shocks of 1973 and 1979 forced many developing countries to abandon the state-led import substitution industrialization model and undertake far-reaching economic liberalization programs. Oil rents allowed most countries in the Middle East and North Africa to avoid instituting painful structural reforms. The earliest reformers, Jordan, Morocco, Tunisia, and Turkey, were all countries with limited oil revenues. The collapse of oil prices in the 1980s forced the other countries of the region to introduce drastic cuts in public investments, reduce public sector hiring, and curtail safety nets, with a few, such as Egypt and Algeria, adopting wide-ranging macroeconomic stabilization, structural adjustment, and privatization programs (Hopfinger 1996, Harik and Sullivan 1992, Harik 1997, Niblock and Murphy 1993, Celasun 2000).

A further impetus for pursuing economic reforms comes from prospects for greater trade liberalization and exposure to international competition as a result of accession to the World Trade Organization and the implementation of the Euro–Mediterranean partnership agreements. These agreements that 12 countries in the region have either negotiated or are in the process of negotiating with the European Union, emerge from a Euro–Mediterranean conference held in Barcelona in 1995. The partnership aims at creating free trade areas for all industrial products between the countries of the region and the European Union. The partnership agreements would also significantly increase aid flows to the region to facilitate the introduction of structural reforms. The prospects of intensified international competition and more open economies in the region has spawned an active literature on the consequences of these developments on the region's economies (Galal and Hoekman 1997, Handoussa and Reiffers 2000).

5. The Economics of Middle East Peace

A stream of research that investigates the economic consequences of Middle East peace on the region's economy has ebbed and flowed since the early 1990s with the vagaries of Arab–Israeli peace negotiations. An influential collection that brings together authors from both sides of the conflict was published in 1993, just before the signing of the Oslo Peace accords (Fischer et al. 1993). The authors investigated the likely size of a peace dividend that would result from lower military spending in the region, the impact of increased intra-regional trade flows and the possibility of greater foreign direct investment made possible by open borders and communications links. More recent reassessments (Roy 1999, Diwan and Shaban 1999) show the adverse consequences of an incomplete settlement on the fledgling Palestinian economy.

6. Demography, Labor Markets and Human Resources

A combination of demographic forces, human resource and labor market policies, and external economic conditions has saddled the region with some of the highest unemployment rates in the world, with most of the unemployed being concentrated among educated youths. The rapid decline in mortality with continued high fertility, characteristic of the early stages of the demographic transition, has contributed to very high rates of growth in the youth population. Although fertility has begun to decline fairly rapidly in many countries of the region, the number of young job seekers will continue to grow for some time. While educational attainment has increased significantly during the oil boom period, the educational systems of the region were primarily oriented to preparing workers for careers in the civil service. Such careers were virtually guaranteed for workers attaining a secondary level of education. Moreover, because public sector jobs provided compensation levels and working conditions superior to those obtainable in the private sector, educated new entrants had an incentive to queue for them, driving up unemployment rates (Shaban et al. 1995, Assaad 1997). The slowdown in government hiring in the 1980s simply resulted in longer queues and sent unemployment rates soaring. In the Gulf countries, the situation was exacerbated by the fact that public sector employment was used as a mechanism to distribute the oil rents to nationals, with the private sector relying exclusively on foreign workers. These dualistic labor markets resulted in the paradoxical situation of having high unemployment rates in economies that suffer from severe labor shortages (Al-Qudsi 1985, Askari et al. 1997). As a result of these developments, labor and human development issues have moved to center stage on the region's policy agenda.

Research on labor market and human resource issues in the region has been hampered by the severe limitations of access by researchers to the micro-data sets that have become indispensable for undertaking state-of-the-art empirical research in this field. Although the general rule is still that survey data sets collected by statistical agencies are off limits to researchers, some improvement in data availability is occurring in Egypt, Turkey, Iran, Morocco, Tunisia, and the West Bank and Gaza. The improved availability of microdata has allowed a significant research agenda to develop on gender issues in the labor market, determinants of fertility decline, the impact of the demographic transition, human capital accumulation and returns to schooling, youth labor market

trajectories, poverty, and the optimal targeting of social assistance.

7. Developments in the Institutional Support for Economic Research in the Region

The support system for undertaking economic research on the region and for regional researchers has greatly improved in recent years with the founding of the Economic Research Forum for the Arab Countries, Iran and Turkey (ERF). ERF is a regional non-governmental organization based in Cairo, Egypt, which was established in 1993 to produce quality research, inform the policy debate, and build research capacity through conferencing, research competitions, dissemination of interim research results, and publication. The ERF publishes an annual volume entitled *Economic Trends in the MENA Region*, a working paper series, and a regular newsletter and maintains a record of economic research on the region. It also sponsors the publication of edited volumes on specific issues, such as trade liberalization, human development strategies, population issues, and the privatization of state-owned enterprises. Finally, with funding from the World Bank and the European Union, ERF administers regular peer-reviewed competitions for research funds. Besides the ERF, a number of country-specific non-governmental research centers have become increasingly active in economic research in the region, including the Egyptian Center for Economic Studies, the Royal Scientific Society in Jordan, the Lebanese Center for Policy Studies, the Arab Planning Institute in Kuwait, and the Centre de Recherche en Economie Appliqué pour le Dévéloppement in Algeria. In North America, the Middle East Economic Association, which is affiliated to the American Economic Association, continues to hold its annual meetings, whose proceedings are published electronically. It also publishes *Research in Middle East Economics*, an annual periodical.

8. Conclusion

In response to developments in the region and in the world economy, the scope of scholarship in Middle East economics has broadened considerably in recent years. Although the longstanding interest in the role of the state, oil, and international migration remains, there is a growing interest in issues relating to structural adjustment, trade liberalization, and financial reform in the context of greater integration with Europe, in particular, and the rest of the world economy, in general. Demographic forces related to the fertility transition and the increased focus on human development worldwide have led to expanded interest in labor market analysis, human resource development, and poverty. This trend is being further encouraged by the greater availability of microdata from household surveys in some of the countries of the region. Finally, there are indications that governance and institutional reform will be the next set of issues to attract significant attention from researchers working on the economics of the region.

See also: International Migration: Mediterranean Basin and the Middle East; Near Middle East/North African Studies: Geography; Near Middle East/North African Studies: Politics; Near Middle East/North African Studies: Society and History

Bibliography

Adams R H 1991 The effects of international remittances on poverty, inequality, and development in rural Egypt. Research report 86, International Food Policy Research Institute, Washington, DC

Adams R H 1993 The economic and demographic determinants of international migration in rural Egypt. *Journal of Development Studies* **30**(1): 146–67

Al-Qudsi S 1985 Earnings differences in the labor market of the Arab gulf states: the case of Kuwait. *Journal of Development Economics* **18**(1): 119–32

Askari H, Vahid N, Mohamed J 1997 *Economic Development in the GCC: The Blessing and the Curse of Oil*. JAI Press, Greenwich CT

Assaad R 1997 The effect of public sector hiring and compensation policies on the Egyptian labor market. *World Bank Economic Review* **11**: 85–118

Celasun M 2000 (ed.) *State-Owned Enterprises in the Middle East and North Africa: Privatization Performance and Reform*. Routledge, New York

Collier P, Gunning J W (eds.) 1999 *Trade Shocks in Developing Countries*. Oxford University Press, New York

Corden W M 1984 Booming sector and Dutch disease economics: Survey and consolidation. *Oxford Economic Papers* **36**: 359–80

Diwan I, Shaban R A (eds.) 1999 *Development under Adversity: The Palestinian Economy in Transition*. The World Bank, Washington, DC

El-Sakka M I T, McNabb R 1999 The macroeconomic determinants of emigrant remittances. *World Development* **27**(8): 1493–502

Fergany N 1988 *Sa'yan wara'a al-rizq: dīr asah mayd¯an¯iyah 'an hijrat al-Misr-iy-in lil-'amal f¯i al-aqt¯at al-'Arab¯iyah* (*In Search of Livelihood*). Center for Arab Unity Studies [In Arabic], Beirut, Lebanon

Fischer S, Rodrik D, Tuma E (eds.) 1993 *The Economics of Middle East Peace: Views from the Region*. MIT Press, Cambridge, MA

Galal A, Hoekman B (eds.) 1997 *Regional Partners in Global Markets: Limits and Possibilities of the Euro-Med Agreements*. Center for Economic Policy Research, London, UK and the Egyptian Center for Economic Studies, Cairo, Egypt

Handoussa H, Reiffers J-L 2000 *The Euro-Mediterranean Partnership in the Year 2000*. Second FEMISE Report. Institut de la Méditerranée and Economic Research Forum for the Arab Countries, Iran and Turkey

Hansen B 1991 *Egypt and Turkey*. Oxford University Press, Oxford, UK

Harik I F 1997 *Economic Policy Reform in Egypt*. University Press of Florida, Gainesville, FL
Harik I, Sullivan D J (eds.) 1992 *Privatization and Liberalization in the Middle East*. Indiana University Press, Bloomington, IN
Hopfinger H 1996 *Economic Liberalization and Privatization in Socialist Arab Countries: Algeria, Egypt, Syria and Yemen as Examples*. Justus Perthes Verlag, Gotha, Germany
Issawi C 1990 Studying Middle Eastern economies: the early years. *Middle East Economic Association Newsletter* **1**(1): 2
Karshenas M 1990 *Oil State and Industrialization in Iran*. Cambridge University Press, Cambridge, UK
Niblock T, Murphy E (eds.) 1993 *Economic and Political Liberalization in the Middle East*. St Martin's Press, New York
Richards A, Waterbury J 1996 *A Political Economy of the Middle East*. Westview Press, Boulder, CO
Roy D A 1991 Egyptian emigrant labor: Domestic consequences. *Middle Eastern Studies* **27**: 551–82
Roy S (ed.) 1999 *The Economics of Middle East Peace: A Reassessment*. *Research in Middle East Economics*. JAI Press, Greenwich, CT Vol. 3
Shaban R, Al-Qudsi S 1995 The challenge of unemployment in the Arab region. *International Labour Review* **134**(1): 65–81
Wahba S 1991 What determines workers' remittances? A framework for examining flows from migrant workers, with a focus on Egypt's experience in the 1980s. *Finance and Development* **28**: 41–4

R. Assaad

Near Middle East/North African Studies: Gender

The study of women and gender has opened up a new perspective in our approaches to Middle Eastern history and society. This article examines the interactions between women's issues and the major cultural and political movements, namely national independence, social reform, and modernization movements in the nineteenth century and Islamist movements revived by the advent of Islamic Revolution in Iran in 1979. It observes changing approaches to the study of women and history in the Middle East, and their implications for gender studies in general.

'Gender' studies began to appear in the 1980s, first in the United States and soon afterwards in Europe and parts of the Middle East, in the wake of the new discipline of women's studies, which was itself an academic and intellectual outcome of the feminist movements of the 1960s and 1970s. It was a new attempt to reveal unspoken inequalities and power relations between men and women. In contrast to the biological differences highlighted by reference to 'sexes,' the term 'gender' depicts socially constructed, culturally specific forms of masculinity and femininity that are subject to social interpretation, negotiation, and change. Furthermore, gender is used, not exclusively in relation to women's issues, but as a constitutive element of all aspects of human society and organizations, such as the family, economy, science, medicine, state, and politics. The analysis of the ways in which different aspects of social life, history, and culture are gendered, namely embedded by culturally perceived differences within a field of power relations, has become a major concern for social sciences. Few concepts have transcended cultural and geographical boundaries so powerfully during the 1980s and 1990s as gender. This has contributed to pioneering work on gender within the field of Middle Eastern history (Keddie and Baron 1991). Gender awareness has changed ways of studying the history of colonization, state building, nationalism, and Islamic identity in the Middle East. At the same time, tracking the distinct trajectories and identity formations followed by Middle Eastern women, both Muslim and non-Muslim, renders a new map for critical reflection in gender studies. As concepts travel in two directions, they create new avenues of learning, dialogue, and interaction between the West and other cultures, ranging from academy and law to citizenship. This new dialogical concern revitalizes an interest in the study of local cultures and in cross-cultural comparative perspectives.

The study of women in the Middle East posits gender as a central issue, an agenda-setting issue for broader historical, political, and cultural processes that have shaped these societies. For at least a century, the problematic link between modernization and cultural identity in the Middle East put women at the heart of political and societal debates. Women's position and agency in society were central in the shaping of movements of modernization, state-building processes, nationalist ideology, and Islamist protest movements. Discussions on the appropriate position for women in Muslim countries continue to underpin any political discourse aiming at the reorganization of these societies. Women are situated at the crossroads between two distinct sets of values and affinities, namely between universalism, secularism, and feminism on the one hand, and cultural authenticity, religion, and Islam on the other. Or, to put it succinctly, in Middle Eastern societies, gender issues are embedded in civilizational issues in the sense that women's position in society is the determining factor both in framing and intertwining what were posited as the existing dualities, such as traditional/modern, religion/secularism, Islam/the West, equality/difference, and interior/exterior. Studies on women in the Middle East follow common and overlapping trends with Western feminist thought but also reflect the distinctive centrality of gender issues in Middle Eastern societies.

1. *Women's Civilizing Mission*

The social position of women began to be discussed with the onset of modernization attempts in the nineteenth century. As modernization seemed to

imply Westernization, the ultimate question that reformists needed to answer was whether Islam was compatible with Western civilization and more particularly with women's rights and freedom. Different answers to that question—ranging from national modernists and Islamist reformists to religious conservatives—continue to put their stamp on the intellectual and political history of the Middle East. Analysis of these historical debates from the point of view of women's questions has shown discursive parallels in countries such as Egypt, Turkey, and Iran (Ahmed 1992, Badran 1995, Gole 1996, Najmabadi in Kandiyoti 1991). For progressive nationalists, the emancipation of women from 'erroneous' religious prescriptions and traditional ways of life was the prerequisite for national progress. Islamist reformists asserted that Islam was not a barrier to progress and that it was possible in Islam to embrace modern ideals, such as gender equality and human rights. But for conservatives, any change in the direction of Western modernity, especially pertaining to women's position in society, would have meant a loss in cultural identity and consequently, to preserve moral integrity and social cohesion, gender relationships were to be framed according to the Islamic religious law (*Shari'a*) as they understood it.

As markers of cultural identity, women therefore set the agenda for different and competing social projects. Furthermore, the most disputed issues throughout the nineteenth and twentieth centuries, namely polygamy, veiling, gender segregation, women's education, employment, and participation in public life, remain on the public agenda of Middle Eastern countries in the 1980s and 1990s.

The Middle Eastern model of early modernization, namely the era of colonial, post-colonial, and nationalist state building, can best be summarized by a formula that equates women's emancipation and national progress. With striking similarities between Turkey, Iran, and Egypt, and also with the North African countries, women's public visibility and gender equality are considered to be the yardsticks of progress and civility. Turkish Kemalist nationalism articulated and implemented gender equality as an intrinsic part of its modern identity and societal reforms. Women gained full political rights in 1934, and the removal of family law from the jurisdiction of religion with the enactment in 1926 of a secular Civil Code (adapted from Swiss law) in Turkey can be considered as the most radical example of engagement in secular feminism. In Iran, too, women's rights, free education, and compulsory unveiling in 1936 were implemented by the secular nationalist politics of Reza Shah. In Egypt, alongside nationalism, the Islamic reformist movement (especially in the works of Muhammed Abduh and Qasim Amin) paved the way for women to gain access to public life, to claim a public voice, and to demand educational, work, and political rights. Veiling was criticized as a sign of women's confinement and as an obstacle to women's participation in public life. Thus, the removal of the veil and the head covering of women in public became a symbol for women's emancipation as well as national progress. A well-known Egyptian feminist, Huda Sharawi, announced the start of the feminist movement when, returning from an international feminist meeting in Rome in 1923, she disembarked from her train in Cairo and unveiled her face in public. Therefore, public visibility of women was put forward by the first wave of feminist activism and writing in the Middle East. A pre-Islamic past, whether Anatolian, Zoroastrian, Pharaonic, or Berber, was idealized as egalitarian in gender relations and celebrated in the construction of a new nationalist idiom. Women, especially upper and middle class women, mostly coming from a traditional universe but exposed to modern ideals and education, were called to participate in this nation-building process, as citizens, educators, and role models for a 'new life' in the family as well as in the public sphere. The limits of women's participation in public life and social agency were framed within the terms set by the nationalist movements whether backed up by secular or religious reformist discourse. Feminist scholarship in the Middle East revealed these issues pertaining to gender and nationalism and developed in distancing itself from and in criticism of nationalist idiom (Kandiyoti 1991, Moghadam 1994).

2. *Women and Islamist Movements*

The revival of Islamist movements marks the second turning point in contemporary Middle Eastern history where women's issues and agencies once again play a central role, but this time as markers of Islamic difference rather than as transmitters of Western-centered modernity.

Although Islamic movements are far from being monolithic, one can still discern common features in the contemporary Islamist movements that have developed since the end of the 1970s in almost all Muslim countries. Their aim is to transform the whole of society along Islamic lines, ranging from government, law, and science to ways of life, gender relations, and faith. The Iranian Islamic revolution in 1979 provided the political source of reference and legitimacy for such an ideal. Actors of Islam, men and women alike, come mostly from provincial towns and lower social classes, but are an outcome of urban migration and modern education. The pressures generated by the encounter with a modern universe create a need to reinterpret religion and re-adapt gender roles and divisions. A call for a return to the original, unadulterated, fundamental sources of Islam (the Qur'an and the Sayings of the Prophet) has engendered the revival of debates on Islamic knowledge. The interpretation of religious precepts and texts is opened up

to public debate in which not only the religious specialists (*ulema*), but also new Islamist figures, both men and women, participate. In a paradoxical way, contemporary Islamism has instigated the democratization of religious knowledge whereby various actors of Islamism can lay a claim on the interpretation of religion and Islamic jurisprudence. Issues related to the segregation of, and discrimination against, women, such as compulsory veiling (*hijab*); religious marriages and polygamy; women's obedience and legal subordination to the husband's authority; inequalities in divorce, custody, or inheritance; and interdictions put on Muslim women's access to religious positions, praying in mosques, marriage to non-Muslim men are no longer issues settled under the monopoly of religious *ulema* but are the subjects of public controversy. Fatima Mernissi has made a pioneering contribution to understanding the specificities of Islamic constructions of female sexuality and spatial social organization, and giving voice to religious women figures in public life and history (Mernissi 1975). Other academic work followed and contributed to new readings of women's position in religious and legal texts (Stowasser 1994, Yamani 1996, Mayer 1999).

In its criticism of essentialism and reflections on the constructed nature of identity and difference politics, feminist theory has contributed to the development of new approaches to Islamic identity. Islamist identity is distinct from Muslim identity to the extent that it expresses a refutation of given definitions, accepted ways of 'being a Muslim,' 'born being Muslim' categories, handed down from generation to generation, and asserts a collective and countercultural attitude. Islamism expresses the problematization of Muslim identity both by a return to the origins of religion and by a non-assimilative attitude towards Western modernity. Islamism shares a similar pattern of collective and personal empowerment with other contemporary social movements in the West, namely that of appropriation of their difference (Muslim, woman, or black), the very difference which accounted for their exclusion and stigmatization. Islamism is a collective manifestation of an exacerbated form of Muslim difference. A search for an alternative in Islam endows especially newly urbanized and educated Muslim actors with a very strong quest for an authentic indigenous identity. The covering of women, the 'veiling,' is the most visible symbol of this claim for Islamic authenticity and difference.

In all Muslim societies, the Islamist movements gained public visibility through the veiling of women. On the one hand, the veil recalls the traditional definitions of Muslim womanhood based on values of modesty and virtue, the segregation of sexes, and the interdictions put on women with respect to their participation in social and public life. But, on the other hand, the new veiling signifies the political participation and the active voluntary reappropriation of Islam identity. As such, the new veiling does not represent the traditional images of Muslim women: uneducated, docile, passive, devoted to family life. On the contrary, young, urban, educated groups of Muslim girls are politically demanding and publicly visible. The way they readopt the Islamic outfit is different also from its traditional outlook.

Traditional Islamic ways of covering change from one region to another, manifesting diverse forms in different national settings (the Iranian *chador*; the 'kerchief' covering the lower part of the face in North Africa; the face mask of the Gulf region; the Afghan *burqa*). In contrast, contemporary Islamic dress is a new style, almost like a uniform, independent of the national context. Although there is a wide range of styles in Islamic dress, it can be distinguished by a headscarf totally covering the hair and falling to the shoulders, combined with a sort of long and ample overcoat (made of pale-colored industrial fabric) to hide the forms of the body. Overall, the Islamic outfit is a modern reappropriation and reinterpretation of Islamic modesty; the silhouette of women with widened shoulders, carrying big bags, walking in the streets, taking buses, going to the universities, hints at the emergence of a new profile of Muslim women: that of urban, combative women pursuing a career and mixing with men. In short, women are actors of Islamist movements, but they are also following a multifold life and professional strategies, and making part of a modern universe. And this, of course, cannot avoid creating a tension between Islamic precepts, which define the 'ideal Muslim woman' as a homemaking, self-sacrificing mother and wife, and the new individual and professional aspirations of these women. Acquired public visibility, social mixing with men, and intellectual and professional aspirations force them into a tension between their collective role as agents of Islam, as markers of 'authentic Islamic identity,' and their emerging individualist and feminist consciousness (Gole 1996). Paradoxical it may appear, in promoting the veil as the symbol of their cultural difference, but Islamic movements unintentionally encourage women's participation and visibility in public life, which, in turn, constitutes a subversive potential for gender consciousness and activism.

Women voice publicly their demands for access to education, employment, and political life (Adelkhah 1991). In the post-Khomeini Iran, women's periodicals proliferate in which women defend gender equality in their readings of religious texts and criticize Islamic jurisprudence. They argue for women's rights such as their eligibility to engage in religious studies, to hold juridical positions, to participate in the lawmaking process, to stand for public office, to have custody of children after divorce, to have equal consideration with men in testimony. In these periodicals, Iranian women argue for a women's rights perspective in their critical readings and interpretations of Islamic law and power.

In Egypt, where there has been a rich tradition of modernist Islamic thought, there is an increasing concern for the feminist consciousness to be religiously rooted in Islam. Whereas in Turkey, where secularism and Islam were radically separated, Islamic women's claim to have the right both to wear the headscarf and to enter into secular public spaces (to attend the Universities, to be admitted to public service, to be elected to the Parliament) addresses such questions with new urgency to Islam, but also and foremost to secular definitions of feminism and politics. The homogeneous and secular constitution of the public sphere is thus challenged by the emerging claims of religion and difference.

3. Gender and Multiple Modernities

The entrance of the newly veiled women into the public sphere across the Middle East by political, educational, and professional means stands in continuity with the process begun by the upper and middle class women who took off the veil almost a century ago. Both figures of women are products and agents of social change, transforming the relations between tradition and modernity, religion and secularism. If the first generation of women acted for accommodation and aspiration to modernity, framed by nationalism, the second generation searches resources in religion to act upon modernity. Islamism is yet another moment in the indigenization of modernity in the Middle East; it represents a critical reflexive mode of the appropriation of modernity.

Religious and cultural traditions, rather than being immutable, static, and marginalized, emerge as a vibrant force in the processing of modernity (Fernea 1994, Gocek and Balaghi 1994). Traditions are selectively reappropriated and transformed in the light of modern dilemmas that women face at the forefront. Women's employment, political engagement, public presence, have an impact on the necessity to reconsider the family sphere, the place of motherhood in society, the values of modesty and primacy of community. Rather than assimilating to the liberal values of modernity and encouraging every move from the private to the public in the name of liberation, Islamic politics advances cultural difference as a form of resistance and as a way to articulate an alternative project. Women's bodies are sites where this politics of difference is inscribed; the new veiling displays the embodied difference. In a similar way that the modernist experience was carried out on unveiled bodies and secularized ways of life, the Islamist difference is crystallized through women. New Muslim subjectivities are worked out by politics of difference, celebrating and disciplining religious self and bodily practices in a modern universe. Muslim women are situated at the crossroads between assimilating modernity and subjugating traditions; gender consciousness provides women with a critical reflexivity, a compass to map out a new trajectory. The interplay between Islam and modernity blurs the distinctions between secular and religious, private and public. New definitions of self, modern and Muslim, are emerging from this conflictual but creative encounter. Women's agency is decisive in determining the course of this ongoing, dynamic process.

The new combinations and crossbreeding between cultural difference and modernity, Islam and gender equality necessitate a new perspective for understanding simultaneous but different experiences of modernity. Approaches in terms of insufficient and delayed modernity do not account for the centrality of women in the Middle East. Studying gender in the Middle East inevitably reinforces local and a cross-cultural comparative perspective.

See also: Islam and Gender; Islamic Fundamentalism; Nationalism, Historical Aspects of: Arab World; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Economics; Near Middle East/North African Studies: Politics; Near Middle East/North African Studies: Religion; Near Middle East/North African Studies: Society and History

Bibliography

Abu-Lughoud L (ed.) 1998 *Remaking Women. Feminism and Modernity in the Middle East*. Princeton University Press, Princeton

Adelkhah F 1991 *La revolution sous le voile*. Karthala, Paris

Ahmed L 1992 *Women and Gender in Islam*. Yale Univesity Press, New Haven, CT

Badran M 1995 *Feminists, Islam and Nation. Gender and the Making of Modern Egypt*. Princeton University Press, New Jersey

Fernea E (ed.) 1994 *Women and the Family in the Middle East: New Voices of Change*. University of Texas Press, Austin, TX

Gocek F M, Balaghi S (eds.) 1994 *Reconstructing Gender in the Middle East. Tradition, Identity and Power*. Columbia University Press, New York

Gole N 1996 *The Forbidden Modern. Veiling and Civilization*. University of Michigan Press, Ann Arbor, MI

Haddad Y Y, Esposito J L (eds.) 1997 *Islam, Gender and Social Change*. Oxford University Press, Oxford, UK

Kandiyoti D (ed.) 1991 *Women, Islam and the State*. Macmillan, London

Keddie N R, Baron B (eds.) 1991 *Women in Middle Eastern History. Shifting Boundaries in Sex and Gender*. Yale University Press, New Haven, CT

Mayer A E 1999 *Islam and Human Rights. Tradition and Politics*. Westview, Boulder, CO

Mernissi F 1975 *Beyond The Veil: Male–Female Dynamics in a Modern Muslim Society*. Rev. edn. Indiana University Press, Bloomington, IN

Moghadam V M (ed.) 1994 *Gender and National Identity. Women and Politics in Muslim Societies*. Zed Books, London

Najmabadi A (ed.) 1990 *Women's Autobiographies in Contemporary Iran*. Harvard University Press, Cambridge, MA

Stowasser B F 1994 *Women in the Qur'an, Traditions, and Interpretation*. Oxford University Press, Oxford, UK
Yamani M 1996 *Feminism and Islam. Legal and Literary Perspectives*. Ithaca Press, UK

N. Göle

Near Middle East/North African Studies: Geography

Middle East research has always been first and foremost 'geography,' as have most of the other regional and thematic sectors of human geography referred to here. In other words, it has always stayed closely bound to the theoretical and methodological developments of the discipline as a whole. It is only within this framework that a regional category like 'Middle East research' can be described as 'specific' or 'different.' This integration is crucial to the understanding of many essential developments. Whether one thinks of the decreasing importance attached to map drawing as the geographer's methodological ideal, the decreasing relevance of studies which restrict themselves to the morphology of the cultural landscape, or of critical reflections on the meaning of 'culturally specific,' the essential impulses have always come from other sub-disciplines or from geography's sister social sciences. Two characteristics of the institutional organization and staff composition of geography as a university discipline have contributed substantially to the perpetuation of this integration:

(a) Essentially, geographical Middle East research has always been established at university institutes, rather than as independent 'centers of geographical Middle East research.' Thus, the institutional conditions have hindered fundamentally disparate developments.

(b) Only very few geographers have restricted their own scientific work to a single regional specialization such as 'the Middle East,' and the creation of specific chairs for Middle East research has been rare. Most scientists are also concerned with their own country of origin or with cross-regional topics, and indeed some have questioned the validity of the regional (as opposed to the sectoral) approach to scientific study in general. This common attachment to different sections of the discipline has also contributed significantly to the fact that geographical Middle East research has never developed along a fundamentally different path from the general discipline, whether conceptually or methodologically.

1. Concepts and Methods of Geographical Middle East Research

Despite spatial proximity and an age-old tradition of economic and intellectual exchange, natural and cultural differences are factors that were decisive in establishing scientific interest in the Middle East. Although the various early reports by famous travelers may be invaluable as sources, if the strict criteria of methodical and systematic data collection and analysis are adhered to, then it is not possible to speak of 'geographical Middle East research' before the time of World War II. Many important earlier works must primarily be assigned to sister disciplines such as archaeology. 'Natural difference' and primarily 'cultural difference'—with their thematic foci, conceptual premises, and methodical approaches changing considerably over time—have to date defined a loose framework for geographical Middle East research and given rise to two 'axes of tension,' at least one of which can be found in nearly all works:

(a) Without necessarily implying acceptance of an over-simplifying geographical determinism, the question as to how the specific natural environment in the Middle East affects space-shaping human activities has played an important role in many now 'classical' themes in Middle East research. This is especially apparent in research on nomadism or in agricultural geography, which deals with different patterns of land use and have to take into consideration the consequences of resource scarcity as well as distinguishing features of the agricultural social structure. This linking of specific Middle Eastern natural factors with the creation of spatial human environments—a link that is central to geographic interest—should not necessarily be seen as direct and causal, but instead regarded as intrinsic to the origin of whole social systems and societies. This subject attracts little attention today.

(b) A second axis of tension can be seen in the degree of significance attributed to the cultural foundations of space-shaping activity. A typical example can be seen in urban geography, where—expressed crudely—the Middle Eastern city is seen with the culturalist eye as 'Middle Eastern' and with the universal-functionalist eye as 'city.' Said's (1978) criticism of 'Orientalism' directly affected some geographical models like that of the 'cultural continents,' re-emphasizing the difficulties of an appropriate approach to 'culture.' Even so his theses had far less impact among Middle East geographers than among human geographers in general where—predominately in the Anglo-Saxon world—an intensive discussion over the accusation of 'orientalism' was embedded within a broader debate on concepts of 'culture.' The general concentration of scientific studies on elements of material culture contributed significantly to the low level of resonance met by Said's criticism, as did a questionable confidence in more 'objective'—compared with other social sciences—research methods (Fisher 1981, p. 433).

A distinct empirical approach, which has contributed to a partial preservation of an empirical-descriptive attitude to materials, has always been

characteristic of geographical Middle East research. Even today, it is almost impossible to find dissertations that are not based on first-hand fieldwork; and as long as interest in the historical and political aspects of the discipline remains underdeveloped, a broad and comprehensive treatise on the history of geographical Middle East research will not be written. It is only since the 1970s that it has been possible for the researcher to supplement his or her own independently collected material with publicly available data published by local authorities.

Apart from the decreasing interest in environmental factors of influence, one dominant 'global trend' can be recognized despite all diversity and difference. This is a decreased occupation with examinations of material culture, which were often primarily historically-genetically oriented. In place of this, focus is shifting to people, identities, patterns of action, and everyday life as well as to social groups, institutions, and social change. This thematic shift applies in general to all of the concrete fields of research mentioned below and has several causes. After a phase of extensive occupation with material culture, a basic consensus had been achieved on many open questions. Thus, it was possible to move on to consider variables that represent in different ways the basis for the origin and development of material culture. A second reason lies in the increasing public interest in problems of intercultural understanding. Geography's sister disciplines, with the exception of anthropology, had also paid little attention to this topic and consequently a whole new field of research unfolded. Finally, the increasing acceptance of qualitative research methods created a suitable environment to look at the social foundations of human space-shaping activities. Whereas interviews in the early phases of geographical Middle East research served primarily as a supplement to drawing maps and therefore concentrated on 'hard' facts, today the whole spectrum of methods of empirical social research, quantitative as well as qualitative interviews, or participant observation, are important. Values, norms, subjective meanings, and references of sense that guide space-shaping activities have now been able to move into the foreground of analysis and evaluation. Working with interpreters has thus become increasingly difficult and more and more geographers have acquired their own knowledge of foreign languages.

2. *Main Thematic Foci*

The spectrum of thematic foci of geographical Middle East research is very broad and has been molded by different national traditions. These in turn have been decisively influenced by the historical roles as colonial powers of various countries in the Middle East, or by the absence of colonies in the case of Germany. Close contact to other human sciences is characteristic of many larger research projects, where the common denominator is often an approach that is 'empirical' and oriented on 'the present,' combined with a mutual 'interest in material culture.' Consequently, the dominant partners have been anthropology and sociology, archaeology and, in part, architecture/town planning. Added to this are the important roles played by history and politics as well as by Islamic sciences and economics. Such cooperation, and the resulting familiarity with the approaches of other disciplines, has also contributed to the broadness of the field of geographical research topics. Sweeping statements about 'research foci' or 'thematic shifts' are thus only possible with reservations.

In the early stage of modern geographical Middle East research (up to the 1930s), extensive field trips to often largely unfamiliar target areas usually sought to collect as much information on different thematic topics as possible. Already in its infancy, this methodology generated descriptions of single cities and regions that—positively formulated—were characterized by an 'integrating perspective' but at least partly represented encyclopedic collections of facts. Even in 2000 (especially in Germany, but also in France), regional-geographical studies without scientific claim but based on material which has been collected and analyzed scientifically and methodologically make up a large part of all Middle East-related monographs. They are an indicator of the continued demand for comprehensible basic knowledge and constitute a medium of transfer between an institutionalized Middle East research system and an interested public.

Moving away from this array of works dedicated to single regions, the geographical perception of the Middle East as the Lebensraum for the three lifestyles (*'genres de vie'*), 'farmer,' 'urbanite,' and 'nomad,' was already guiding research before this trilogy had been explicitly expressed. 'Patterns of agricultural land use by settled farmers,' 'city and urbanization,' and 'transformation of mobile lifestyles and economic systems' became the early foci of geographic Middle East research. Of course, this description is a crude one, which lumps together extremely heterogeneous works within the same division:

(a) Works on agriculture and land use represent the earliest focus of geographical research; within French geography the link with the role as a colonial power is especially apparent. The temporal-spatial organization of agrarian land use, underlying social structures, and the relationships between rural and urban space as well as strategies, successes, and problems of agrarian reforms were among the subjects of important publications. Questions about the natural limits of habitable areas and shifting borders between the Lebensraum of farmers and nomads were often examined from a historical point of view and with some overlap with the field of physical geography. In particular, British geography dealt with questions of

water use and irrigation, providing an anchor point for the discussion of water problems in the context of international conflicts, a theme that is now very important. As one of the most important practically oriented research projects, the 'Khabura project' in Oman—led by the University of Durham—became well known.

(b) As population growth accelerated and migration to urban areas began, the 'urbanization' theme complex became increasingly important. Today, this field probably enjoys the most extensive corpus of geographical work. Urban geographers found it much easier to draw on works done in other social sciences than did their colleagues in the field of agricultural geography. Internal division into different quarters, the function and organization of the bazaar or suq, changing social structures in the old towns, the organization and change of urban crafts, and the cultural-specific character of Middle Eastern cities in general are among the most important topics. Questions of town planning were often touched on here, but were rarely central. Urban geographical works and more broadly designed studies on settlement systems often have a strong historical bias and thus include archive research alongside fieldwork. Though functional and socio-spatial differences between old and new towns are often addressed briefly, the majority of works deal with the old cities without thereby reflecting their real importance today. Marginalized areas, bidonvilles, and gecekondus have become the subject of geographic interest since the 1980s and mainly in conjunction with the increasing concern in developmental problems.

Nomadism as a way of life and a type of economic system which is unfamiliar to central Europe and which underwent substantial changes during the twentieth century was also the subject of early scrutiny. Areas of distribution and regionally differing organizational forms, relationships with town dwellers and farmers, and the historical importance of nomadism were among the focal points of research, as were the threat to the nomadic lifestyle and the forced settlement of nomads.

A significant proportion of studies do not fit into this classification, and this is especially true for recent publications. They either cut across the categories and belong to all of them in various ways or represent enlargements of various kinds.

The group of works that were written from the point of view of economic and social development or developmental deficits is especially heterogeneous. They adopt widely different theoretical approaches and partly overlap with the study of culturally specific strategies of economic action already established in other social sciences. The spectrum extends from works on modernization and social change up to concrete development and planning projects. Overall, however, the occupation with modern economic sectors began late and hesitantly.

As a reaction to the explosive increase in international travel, research projects were set up which dealt with the numerous and diverse problems and opportunities of international tourism in the destination countries. Only later did the various forms of internal tourism in the Middle East receive attention.

(c) Projects on international migration were also less the result of processes within the discipline than of changes in the research subject area. They deal with motives and consequences in the countries of origin as well as in the destination regions, and thus sometimes have the character of 'Middle East research outside the Middle East.'

(d) Questions of water use and distribution, which are gaining increasing importance in many international political conflicts and consequently meet rising public interest, have formed a growing proportion of recent geographical studies.

(e) Across all these thematic fields, a trend towards a strengthened focus on life world and routines of everyday life can be seen. Works dealing with gender-specific spatial activities and studies on the character of urbanity or public space fall into this category.

3. *The Institutional Basis of Geographical Middle East Research*

There have always been various factors determining which geographers (for the first decades of the twentieth century it is perhaps better to use the broader term 'geo-scientists'), from which countries, to what extent, and with which regional foci are engaged in Middle East research. This depends on the one hand on political constellations and the sharing out of colonial spheres of influence at the start of the twentieth century, and on the other to the financial resources and staff composition of particular faculties. As a whole, most impulses for geographical Middle East research came from French and German scientists. British and American studies were of secondary importance, as were Italian and Israeli works.

After World War II, more and more research by geographers from the Middle Eastern countries themselves were published, although these scholars had often been educated in Europe or in the United States and thus did not approach the subject from a completely new direction. Russian research concentrated mainly on the central Asian parts of the former Soviet Union, and Japanese geographers only began to work more frequently in the Middle East in the late 1980s. In addition, individuals from many other countries have carried out research projects without it being possible to talk about a real tradition of geographical Middle East research. Even in countries with a relatively well-established Middle East geography like France or Great Britain, however, the Middle East as a research area has played a more or less marginal role within the whole of the discipline. Fisher (1981, p. 433) for

instance estimates the number of British geographers for whom the Middle East is the main area of interest at no more than 10–20 persons, which would equate to about 1 percent of the members of the Institute of British Geographers (IBG). Only in Germany does the geography of the Middle East carry a somewhat greater weight.

The fortunes of centers of Middle East research within these countries vary naturally in accordance with staff fluctuations, project financing, and indeed is substantially dependent on the political situation in the Middle Eastern countries under examination, as this can sometimes impede or even prohibit fieldwork. Lyon, Paris, and Tours should be mentioned as centers of French research; Berlin, Bonn, Erlangen-Nuremberg, Mainz, and Tübingen are important in Germany; Durham and London in Britain; while Tucson/Arizona and Austin/Texas lead in the United States. A regional specialization within the Middle East which has lasted into the early 21st century can be especially well seen in the case of French Middle East geography and its concentration on the Maghreb states. Up to World War II, German geographers worked mainly in Turkey, in Persia, and in southern Arabia. Similar specialization is less pronounced in other countries, and what little demarcation did exist gradually disappeared after World War II.

Interdisciplinarity has already been mentioned as characteristic of geographical approaches in general, and this is equally valid for the institutionalization of forums of scientific exchange. On an international as well as on the national level, geographers are organized in associations together with Middle East scientists of other disciplines: on the European level in EURAMES (European Association for Middle Eastern Studies), in France in AFEMAM (Association Française pour l'Etude du Monde Arabe et Musulman), in Germany in DAVO (Deutsche Arbeitsgemeinschaft Vorderer Orient), in Britain in BRISMES (British Society for Middle East Studies), and in the United States in MESA (Middle East Studies Association) and IAMES (International Association of Middle East Studies). There is however no international association exclusively for Middle East geographers. The situation is similar with journals: periodicals which concentrate thematically on the Middle East certainly exist at a lot of faculties, but articles intended for a broader readership are published either in specialist geographical publications without regional foci or in regionally focused periodicals with an interdisciplinary character. One internationally distributed journal has existed since 1998: the *Arab World Geographer*. This admittedly does not concentrate exclusively on the publication of articles concerning Middle East geography, but its thematic focus lies firmly in this area, and it might be that this publication will prove to be an important forum for Middle East geographers. The fact that this journal was created by Arab geographers for Arab geographers but is published in English and has an international editorial advisory board may be seen as confirmation of a long history of intensive contact between scientists studying the same culture from within and without.

Summarizing, it can be recorded that there are no indications for a development conceptually or institutionally separating geographical Middle East research from the discipline geography as a whole. A distinct empirical and interdisciplinary approach as well as a thematic focusing on economic and social development, problems, and opportunities of international and internal tourism, the acceleration in migration, the growth of cities, and the scarcity, use, and distribution of resources are characteristic for geographical research in other regions of the world too. The diminishing importance of isolated analyses of material culture and a trend towards a strengthened concentration on life-world and routines of everyday life go along with a rise in theoretical reflections and are still unbroken. The fundamental question about the future of a constituting 'object' for geographical Middle East research is closely bound to the discussion about the relativization or increasing relevance of cultural regionalization under conditions of 'globalization' and has of course to be left open here.

See also: Near Middle East/North African Studies: Economics; Near Middle East/North African Studies: Politics; Near Middle East/North African Studies: Society and History; Orientalism

Bibliography

Ehlers E 1984 German geography of the Middle East: Trends and prospects. *Middle East Studies Association Bulletin* **19**(2): 183–95

Fisher W B 1981 Progress in the Middle East. *Progress in Human Geography* **5**: 432–8

Fisher W B 1983 Middle East. *Progress in Human Geography* **7**: 429–35

Said E W 1978 *Orientalism*. Routledge Kegan Paul, London

Wirth E 1988 German geographical research in the Middle East and North Africa. In: Wirth E (ed.) *German Geographical Research Overseas: A Report to the International Geographical Union*. Deutsche Forschungsgemeinschaft/Institute for Scientific Cooperation, Bonn, Germany, pp. 93–132

P. Lindner

Near Middle East/North African Studies: Politics

This article provides a selective overview of the development of Middle East politics as a field of study. It identifies a number of the more prominent themes

and questions that have shaped the study of Middle East politics, with a particular focus on research trends in comparative politics and comparative political economy. The article also emphasizes how the relationship between the study of Middle East politics and the social sciences has changed over time, and points to some current tensions in that relationship. It concludes with a brief, speculative observation about where the relationship between the social sciences and the study of Middle East politics is headed at the turn of the twenty-first century.

The study of Middle East politics has deep historical roots in the social sciences. The Middle East—defined here to include the region that stretches from Morocco in the west to Afghanistan in the east, and from Turkey in the north to the Arab Peninsula in the south—has figured prominently in a tradition of political and sociological-comparative research stretching from Max Weber to Karl Wittfogel, and Immanuel Wallerstein. Broadly defined, this tradition has sought to understand why capitalism and democracy emerged first in the West, and how the initial appearance of these forms of economic and political organization in the West affected the developments of non-Western parts of the world. For well over a century, social scientists have drawn, in particular, on the experience of the Ottoman Empire to study political institutions, social structures, and economic arrangements perceived as explicit counterpoints to those that sparked the rise of capitalism and the transition from absolutist to republican systems of rule in the West (Anderson 1979).

This tradition, in which the experiences of the Ottoman Empire and the Arab states that formed in its wake are used as counterpoints to European experiences—often negative counterpoints—has had a lasting impact on the study of Middle East politics. It established a focus on absolutism, 'predatory rule,' despotism, and authoritarianism as central themes in social science research on the Middle East. The legacy of this tradition has been decidedly mixed. On one hand, authoritarianism and personalistic modes of governance remain prominent issues in the study of Middle East politics today—appropriately so given political conditions in the region (Heydemann 1999, Weeden 1999). Less positively, this legacy is expressed through research in which individual behavior and collective outcomes are understood in terms of deeply embedded cultural constraints—an essentially authoritarian political culture, for example—that reinforce patrimonialism, undermine prospects for democracy, and impede the development of institutions capable of securing economic growth. Such culturalist assumptions reinforce the impression of Arab (or Muslim) politics as exceptional. They diminish possibilities for integrating Middle Eastern cases into broader comparative frameworks, or applying to social and political processes in the Middle East theories and methods that have been used productively in other settings. In practice, moreover, culturalist research has rarely escaped the tautological quality of many theories that rely on culture as an explanatory variable. Largely for this reason, culturalist arguments are less prominent in the study of Middle East politics today than they were in the 1960s and 1970s.

After World War II, with the rise of the USA as a superpower, and the not unrelated emergence of comparative politics as an established subfield of political science in the USA, the study of Middle East politics made significant contributions to a number of important research programs in the social sciences. These included theories of political development, political decay, and praetorianism (military rule), as well as theories of dependent and peripheral development (Keyder 1987, Abdel-Malek 1968). As early as 1956, for example, economic historian Charles Issawi studied the relationship between economic development and democracy in the Middle East, predating the more general work on this topic by Seymour Martin Lipset (1960). Daniel Lerner's *The Passing of Traditional Society: Modernizing the Middle East* (1958) also influenced the subsequent development of modernization theory. His study was cited by Lipset, and was noted, as well, in Almond and Coleman's *The Politics of the Developing Areas* (1960), a volume that exerted considerable influence in the USA on the study of politics in the developing world for much of the 1960s and 1970s, both as a model to be followed and as the object of criticism. Samir Amin's *Accumulation on a World Scale: A Critique of the Theory of Underdevelopment* (1974) was an influential contribution to debates about the role of the industrialized North in imposing conditions of underdevelopment on the postcolonial states of the South. Edward Said's *Orientalism* (1978) originated outside the social sciences, but has had a profound and widespread impact on the study of Middle Eastern politics and is central to the rise of postmodernism within the social sciences more generally. Said's work, along with the writings of postmodern theorists such as Michel Foucault, has generated significant interest in discursive and interpretive methods in the study of Middle East politics (Mitchell 1988), and revitalized research on power and authority in relations between the Middle East and the West. Other important areas of research in which the Middle East has figured prominently include the study of nationalism, state formation and state–society relations in the developing world; religion and politics; and social revolution—themes that continue to inform research on Middle East politics today.

For much of the postwar period, therefore, the study of Middle East politics, especially in the USA, developed in a close and interactive relationship with the field of comparative politics, and with other disciplines in the social sciences and humanities. During these years scholars of Middle East politics, like their colleagues who worked on other regions, made considerable progress in exploring such concerns

as the experience of colonialism and how it influenced the trajectory of political and social life after independence; state formation, the construction of national societies, and the formation of national markets; the political economy of industrialization in the Middle East, and the region's incorporation into a global capitalist system; the role of new military elites and the rise of single-party authoritarian regimes; the incorporation of workers and peasants into the political arena and its effects on coalition building and regime formation; the rise of import substitution industrialization as the dominant economic development strategy in the region and the growing role of the state in the economy, as well as the relationship between private sectors and regimes that often pursued, and coercively implemented, highly redistributive social policies; the effects of regime type on a wide range of economic, social, and political processes; the rise of state-corporatist models of interest representation, and more recently the revitalization of 'civil' societies as sites of political contestation.

While these issues were prominent in social science research on Middle East politics within the USA, and provided the empirical foundations for theoretical developments in the field, research programs in Europe followed a somewhat different trajectory, less with respect to themes and issues—European scholars, like their colleagues in the USA, engaged issues concerning the state, modes of governance and power, the politics of economic development, and processes of social formation—than to methods and conceptual approaches. Disciplinary boundaries loomed less large among European scholars of Middle East politics, who preserved closer links between history and political science than became the case in the USA, especially from the 1980s onward. It is also the case, however, that European traditions of linguistic–philological and Orientalist scholarship, as well as the legacy of having ruled the Middle East as colonial powers, were more evident in the concerns of European research programs than in those that developed within the USA. In addition, during the postwar period, the work of European scholars reflected a more assertive engagement with various strands of Marxist political theory than was evident in the Cold War atmosphere of the USA. This was manifest most productively in the development of a rich tradition of labor studies among European, especially French, scholars of Middle East politics. More recently, the presence of growing Middle Eastern communities within Europe has sparked considerable interest in issues such as the dynamics of migration from Muslim societies to Europe; the experience of immigrant life among Muslims in Europe; the politics and political economy of a greater Mediterranean region that links, in particular, North Africa and southern Europe; as well as the impact of expanding Muslim populations on social relations and political processes in largely Christian European states.

Within the Middle East, as well, research trends were and remain heavily influenced by the experience of colonialism. Higher education systems tend to be organized along lines similar to those found in the country that exercised colonial authority in a given place, with scholars in North Africa and Lebanon socialized into French modes of training and scholarship—as well as French academic networks—while training, research, and professional socialization of scholars in Egypt and Jordan reflects (though to a more modest degree) the influence of British rule. However, the most powerful influence on the study of contemporary Middle Eastern politics by scholars working in the Middle East is the repressive political environment found in most Arab states, and the resulting lack of financial support, data resources, and independent institutions capable of functioning as sponsors and promoters of social science research. The research environment in the region varies, however, and scholars in the Arab world continue to work, against the odds, to produce high-quality social science. In recent years, active and vibrant research communities in Morocco, Turkey, Egypt, and Lebanon (among others) have established significant research programs on democratization, Islam and politics, civil society and the role of new social movements, the politics of economic liberalization, as well as regional and international relations. The state of social science research on Middle East politics in Israel reflects the influence on Israeli scholars of trends in American social science, including an emphasis on quantitative and formal methods. Work on Middle East politics also benefits from a more open research environment than elsewhere in the region. However, Israeli research programs also reflect the effects of the protracted and persistent Arab–Israeli and Palestinian–Israeli conflicts, both as a filter that shapes the perspectives of researchers, and as a focal point of research programs themselves.

Since the early 1990s, and still in keeping with broader trends in the social sciences, scholars of Middle Eastern politics have focused attention on issues of gender, economic and political informality, local level politics, identity politics, civil society and the dynamics of social movements (Singerman 1995, Norton 1996). They have sought to integrate the experience of the Middle East into larger debates in the social sciences around such questions as authoritarianism and democratization (al-Naqib 1991, Salame 1994) and the comparative political economy of development (Waterbury 1993, Waldner 1999). Inevitably, however, the study of Middle East politics has also emphasized that which is distinctive about the region. Three aspects of Middle East politics have received particular attention from researchers over the past two decades.

First, the relationship between Islam and politics, and the dynamics of politics in Muslim societies more generally, have been and remain subjects of con-

siderable interest (Eickelman and Piscatori 1996). These became especially prominent in the wake of the Iranian Revolution of 1979 and the increasing visibility of militant Islamist movements and organized Islamist political parties from Morocco to Afghanistan. Second, the social, economic, and political consequences of the oil boom of the 1970s on Middle Eastern states and societies has also sparked a considerable body of research in Middle East comparative politics and comparative political economy that focuses on the fiscal sociology of the state. Perhaps the most enduring contribution of this research can be found in theories of the rentier or distributive state (Luciani and Beblawi 1987, Chaudhry 1997) that explore the political, institutional, and social implications of circumstances in which states and governments generate a significant share of their revenue from the sale of natural resources, and are not dependent on the direct taxation of their citizens. Whether representative forms of governance can emerge where taxation (the direct extraction of resources from citizens) is low remains a matter of considerable interest in the study of Middle East politics. With the decline of oil prices in the 1980s and 1990s, the question of whether, and under what conditions, states that have not historically developed the capacity to tax might do so has also received some attention.

Third, and not entirely unrelated to work on the political economy of oil, researchers in a number of disciplines and fields, writing from a variety of (often contending) perspectives, have developed research programs focused on the state in the Middle East. Again, this work has been situated largely within political science, and in the subfields of comparative politics and international relations, but has been a core concern of scholars in the Middle East and in Europe, as well as in the USA. Here too, moreover, the state as a conceptual category has been approached from a wide range of perspectives, including many that have subjected the category to sustained critical scrutiny. Within comparative politics, one persistent debate has addressed the question of the exceptionalism of the Arab state; whether culture and history (the absence of an experience comparable to the Enlightenment, and the resulting lack of differentiation between religion and state are most often cited in this regard) combined to produce a distinctive form of state, one that defies ready comparison with states in other regions and, moreover, requires distinctive conceptual tools and a special analytic vocabulary if we are to make sense of it. Increasingly, however, researchers tend to operate on the contrary assumption and view the Arab state as simply one variant among a range of state forms. States in the Middle East are shaped by and subject to the same social, economic and political processes that affect states everywhere, and there is a recognition that these processes are filtered through and experienced within particular historical contexts, institutional legacies, and cultural frames of reference that give state formation in the Middle East a distinctive quality (Zubaida 1989, Ayubi 1995). This starting point permits the development of research programs that draw on and contribute to general theoretical frameworks, while accounting for both intra- and interregional variations.

In addition, questions about and critiques of the state in the Middle East have also served as the basis for a number of productive research programs. A concern with the fixity or fluidity of state borders in the case of Israel and Palestine served as the basis for Lustick's general theory of how state boundaries change (1993). Mitchell (1991) has used Middle Eastern cases to explore how the impression of a boundary separating the state and society operates to construct stateness as a general attribute, and how this attribute is mirrored in (and, in his view, distorts) research on the state in political science. Migdal (1987) also drew on a nuanced understanding of the state in the developing world, including the Middle East, to explore how societies maintain a measure of autonomous power in the face of states that seem to possess an extraordinary capacity to intervene in and shape societies.

In international relations, the contingent quality of state sovereignty in the Middle East, and the impact on state consolidation of regimes that were at times highly responsive to transnational, pan-Arab ideologies, pose important challenges to neorealist theories of the international system. Pointing to the fluid, fragmented, and contested quality of stateness in the Middle East, scholars of international relations, both those more sympathetic and more critical of neorealism, have used Middle Eastern cases to question the neorealist assumption that states are like units that operate as unified actors intent on maximizing their security in an anarchic international system. They have worked to develop theories of inter-state behavior that account for variability in the nature of state interests, and extend more weight to economic considerations, and to the role of norms, ideas, and institutions as determinants of policy making (Barnett 1998, Mufti 1996).

At the beginning of the twenty-first century, however, the study of Middle East politics stands in uneasy relationship to the social sciences. Within political science, sociology, and certain subfields of economics, significant research programs have emerged to explore processes of democratization, the transition from state- to market-based modes of economic governance, and the deepening of economic globalization. Yet as noted earlier, the Middle East remains largely authoritarian, despite limited political openings in some states. The transition to markets has been partial and slow. Linkages between domestic and global markets vary considerably but typically are weak. Economic performance also varies widely, and though

trends in the 1990s were positive, overall performance levels remain low. The effect of these conditions has been to limit opportunities for integrating Middle East cases into some current research agendas, notably but not exclusively in the subfield of comparative politics. In contrast to general trends in the discipline, therefore, the study of Middle East politics at the dawn of the twenty-first century has tended to focus less on democratization and democracy than on understanding how persistent authoritarian regimes adapt to changing domestic and international conditions; less on the transition to and consolidation of markets than on the limits of economic reform; less on the impact of global economic integration than on barriers to globalization; less on developmental success than on persistent obstacles to development. Work on these questions has been profoundly important in explaining political outcomes in the Middle East and holds out significant theoretical value for the disciplines. Yet they are, nonetheless, reflective of how the everyday, lived experience of politics in the Middle East has moved the study of politics along paths that differ from those that are grounded in the experiences of other regions in the developing world.

Methodological issues have also constrained the inclusion of Middle East cases into broader research agendas. Generally, the study of Middle East politics rests on the use of case-specific, macro-historical methods rather than the formal or quantitative methods that are now more broadly applied to the study of politics in other regions. The reasons for this are several, but the methodological orientation of research on Middle East politics results, in no small measure, from the lack of data resources that are needed to undertake quantitative research and from a research climate that is not conducive to data collection and dissemination. Thus, while the continuing development of various institutionalist approaches holds out considerable potential for linking region and discipline, and for cross-regional comparative research that incorporates Middle East cases, prospects for building robust research programs that make use of quantitative and formal methods for the study of Middle East politics are limited, at least at present.

Despite these concerns, current tensions in the relationship between the study of Middle East politics and the social science disciplines should not be overstated. Certainly, it should not be permitted to overshadow the presence of a large number of creative, analytically focused, and highly productive research programs both inside and outside the USA. As noted, researchers on Middle Eastern politics have made good use of and contributed to the development of new institutionalist methods. Comparative macro-historical research coexists comfortably with case-specific but theoretically generalizable research programs that apply a wide range of methods—from archival, discursive, and interpretive to public opinion polling; from network and social movement theories to moral economy and rational choice—all of which are being used to ask important questions about contemporary political life in the Middle East and to engage the analytic concerns of the disciplines. Through attention to the dynamics of political institutions and institutional change; the organization and dynamics of local-level politics and social networks; business–state relations; the relationship between gender, power, and politics; the impact of new information technologies on political life; the formation and transformation of identities; new approaches to the study of culture and politics, and to the role of norms and ideas in political behavior—a list that is far from complete or comprehensive—scholars of Middle Eastern politics are pursuing research programs that not only resonate across regional boundaries but express increasingly sophisticated theoretical perspectives. On this basis, there is reason for optimism that the relationship between the study of Middle Eastern politics and the social sciences is now (again) becoming interactive, reciprocal, and mutually constructive.

See also: Colonialism: Political Aspects; Islam: Middle East; Nationalism, Historical Aspects of: Arab World; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Economics; Near Middle East/North African Studies: Religion; Near Middle East/North African Studies: Society and History; Orientalism

Bibliography

Abdel-Malek A 1968 *Egypt, Military Society: The Army Regime, The Left, and Social Change Under Nasser*. Vintage Books, New York

Almond G A, Coleman J (eds.) 1960 *The Politics of the Developing Areas*. Princeton University Press, Princeton, NJ

Amin S 1974 *Accumulation on a World Scale: A Critique of the Theory of Underdevelopment*. Monthly Review Press, New York

Anderson P 1979 *Lineages of the Absolutist State*. Verso Editions, London

Ayubi Nazih N M 1995 *Over-stating the Arab State: Politics and Society in the Middle East*. I B Tauris, London/New York

Barnett M N 1998 *Dialogues in Arab Politics: Negotiations in Regional Order*. Columbia University Press, New York

Chaudhry K A 1997 *The Price of Wealth: Economies and Institutions in the Middle East*. Cornell University Press, Ithaca, NY

Eickelman D F, Piscatori J P 1996 *Muslim Politics*. Princeton University Press, Princeton, NJ

Heydemann S 1999 *Authoritarianism in Syria: Institutions and Social Conflict, 1946–1970*. Cornell University Press, Ithaca, NY

Issawi C P 1956 Economic and social foundations of democracy in the Middle East. *International Affairs* (January): 27–42

Keyder C 1987 *State and Class in Turkey: A Study in Capitalist Development*. Verso, London

Lerner D 1958 *The Passing of Traditional Society: Modernizing the Middle East*. Free Press, New York
Lipset S M 1960 *Political Man: The Social Bases of Politics*. Doubleday, Garden City, NY
Luciani G, Beblawi H (eds.) 1987 *The Rentier State*. Croom Helm, London
Lustick I S 1993 *Unsettled States, Disputed Lands: Britain and Ireland, France and Algeria, Israel and the West Bank–Gaza*. Cornell University Press, Ithaca, NY
Migdal J S 1987 *Strong Societies and Weak States: State–Society Relations and State Capabilities in the Third World*. Princeton University Press, Princeton, NY
Mitchell T 1988 *Colonising Egypt*. Cambridge University Press, New York
Mitchell T 1991 The limits of the state: Beyond statist approaches and their critics. *American Political Science Review* **85**(1): 77–96
Mufti M 1996 *Sovereign Creations: Pan-Arabism and Political Order in Syria and Iraq*. Cornell University Press, Ithaca, NY
al-Naqib K H 1991 *al-Dawlah al-tasallutiyah fi al-Mashriq al-'Arabi al-mu'asir: dirasah bin'iyah muqaranah*. Markaz Dirasat al-Wahdah al-'Arabiyah, Beirut, Lebanon
Norton A R (ed.) 1996 *Civil Society in the Middle East*. Brill, Leiden, Holland/New York
Said E W 1978 *Orientalism*. Vintage Books, New York
Salame G (ed.) 1994 *Democracy Without Democrats?: The Renewal of Politics in the Muslim World*. I B Tauris, London
Singerman D 1995 *Avenues of Participation: Family, Politics, and Networks in Urban Quarters of Cairo*. Princeton University Press, Princeton, NY
Waldner D 1999 *State Building and Late Development*. Cornell University Press, Ithaca, NY
Waterbury J 1993 *Exposed to Innumerable Delusions: Public Enterprise and State Power in Egypt, India, Mexico, and Turkey*. Cambridge University Press, Cambridge/New York
Weeden L 1999 *Ambiguities of Domination: Politics, Rhetoric, and Symbols in Contemporary Syria*. University of Chicago Press, Chicago
Zubaida S 1989 *Islam, the People and the State: Essays on Political Ideas and Movements in the Middle East*. Routledge, London/New York

S. Heydemann

Near Middle East/North African Studies: Religion

1. *The Changed Setting for Religious Pluralism*

Islam is the predominant religion in all the countries of the Middle East and North Africa with the obvious exception of Israel, where non-Jews make up 20 percent of the population. There are significant Christian and Jewish minorities throughout the region and the lives, doctrines, and practices of the carriers of all these faiths are intricately linked. As with Islam, the practice and understanding of Judaism and Christianity in the region have been highly diverse both in history and in the present, showing an ongoing creative tension between universalism and particularism. The labels 'Muslim,' 'Christian,' and 'Jew'—not to mention Zoroastrianism—aggregate the identities of distinct and often autonomous sectarian groups. Christians form significant minorities in Egypt, Lebanon, Iraq, Syria, and Turkey, and European settlers gave the North African states sizable Christian populations until the end of the colonial era in the mid-twentieth century.

At the beginning of the twentieth century, Christians and Jews were widely distributed throughout the Middle East and North Africa (Planhol 1997). Until the late 1940s in northern Iraq, for example, Sunni Arab Muslims formed a majority of the population in Mosul, with Jews living in both Mosul and surrounding agricultural communities. Shiʿi Turkomen tribes resided in outlying villages, as did communities of Monophysite, Assyrian, and Nestorian Christians. Likewise, Lebanon has 23 recognized Christian and Muslim religious sects, each with separate community identities. At the beginning of the twentieth century there were also thriving Jewish communities distributed widely throughout the region. Some of these—notably in Morocco—remained vital for decades after the declaration of Israel as a state in 1948 (Courbage and Fargues 1995).

Theologians and ideologues often highlight the distinctiveness of their respective faiths, but Muslims, Christians, and Jews in the region often share implicit background understandings and practices. Sacred spaces shared by all three faiths may be found primarily only in Jerusalem and nearby places such as Hebron/al-Khalil, but shared Muslim and Jewish places of regional pilgrimage dot the Moroccan landscape and are equally venerated by Muslims and Jews. Such shared shrines continue to play a vital role in popular belief and practice. Thus in North Africa, Muslim saint festivals (*musim*s) and pilgrimages to regional shrines (*ziyara*s) are paralleled by festivals for Jewish saints (*hillula*s).

Variations in doctrine, religious practice, and traditions of shared memory also distinguish Muslim communities from one another. These include Sunni Muslims, the Shiʿa (or 'sectarians,' who comprise about 10 percent of Muslims worldwide and are further divided into numerous groups), and Ibadi Muslims of Oman and parts of Tunisia, Libya, and Algeria. The intellectual and educated elite often deny or downplay the significance of local, regional, and sectarian differences in faith and practice, but these practices often are powerfully linked to the local social order and continue to play a major role in the popular imagination.

'Modernity' in the form of uniform state schooling and mass communications often alters these practices, but does not erode them. Nonetheless, rising educational levels and the proliferation of communications media, including satellite television unrestrained

by state authorities, are tilting the balance in favor of ideas and practices that can be explained, defended, and foregrounded in public.

2. Explaining Religious Diversity and Change

Like Judaism and Christianity, Islam is rich and varied in its creativity and cultural expression. It is a challenge for Muslim believers and scholars, both Muslim and non-Muslim, to account for the variations of belief and practice as they have been expressed for over 1,400 years in richly diverse cultural and historical traditions. As a response to this complexity, some scholars dismiss as non-Islamic or as incorrect those practices they consider not in accord with 'central' Islamic truths, even though people who hold such beliefs consider themselves fully Muslim. Some deal with these issues by seeing 'foundation' texts at the core of each religious tradition—the Qur'an in the case of Islam—and then seeing how successive generations understand and articulate them. If such texts and ideas are too closely bound to their original settings, no matter how persuasive for their original audience, they risk failing to attract wider audiences. Such texts must remain relevant to the social and political contexts of succeeding generations. The risk of such an approach is to encourage the assumption that the 'foundation' text itself is unproblematic, making successive interpretations the equivalent of substituting values in a mathematical equation in order to perceive unity.

When first introduced in the late 1940s, the notion of 'Great' and 'Little' traditions offered one way of coming to terms with the universal and the particular in religious expression and practice. The Great/Little tradition contrast rekindled an interest in how popular understandings of religion—the 'little,' localized, traditions—were related to elite, literary ones. It also encouraged attention to the specifics of who advocated or explained particular religious orientations and practices, whether they were influential in their own society or marginal, appealing primarily to a particular social group or category. In some settings, individuals claiming descent from the family of the Prophet Muhammad were given precedence. The concept of the Great/Little tradition encouraged exploring the links among religious ideas and practices, authority, and influence.

Often, however, the concept was employed only to juxtapose statements of 'essential' Islamic principles, as elaborated in standard texts and by educated Muslims, with inventories of local religious practices. For example, the study of trance and possession states (called *zar* cults in the Sudan and East Africa), often indicative of the affective side of religious traditions, was often frequently relegated to the 'little' tradition side of the divide and attributed primarily to women. Yet such ideas and practices also shed light on ideas of the person and how both women and men in societies with such practices express and explore alternative social realities (Boddy 1989). Mosque-centered activities such as Friday prayers, almost the exclusive domain of men, often have been seen as part of the 'great' tradition. Women, who often predominate in visits to local shrines and in *zar* cults, are then seen as sustainers of 'local' traditions. This gendered division is made all the more plausible by the fact that until recent decades women had lower levels of education and mobility than men. An alternative explanation, however, is to see mosque-oriented and shrine-oriented practices as integrally complementary, and which must be 'read' together as a unit in order to understand religious practice (Tapper and Tapper 1987).

In North Africa, Hammoudi (1993) suggested the explanatory utility of linking the canonical Muslim sacrificial feast associated with the annual pilgrimage to Mecca with practices such as *Bûjlûd*, a carnival-like masquerade in Morocco's Berber highlands that follows a few days after the canonical feast. Colonial ethnographers regarded the two events as separate, as do Muslim intellectuals. In the masquerade, a young male villager dons the skin of a sacrificial animal and assumes the character of *Bûjlûd* ('the wearer of the skins'), and with his cohorts cavorts noisily throughout the village, sacrilegiously mocking established religious and social conventions, entering homes, and bantering obscenely with women. Hammoudi argues that the masquerade evokes the contradictions implicit in the cultural classifications that regulate ordinary behavior, including gender roles, directing caricature and self-directed irony against local social hierarchies, tradition, and innovation. The canonical Muslim sacrifice, on the other hand, embodies the ideal community. The canonical sacrifice and masquerade inherently conjoin ritual piety and impiety, embodying an ongoing contest over significance and interpretation in which existing and alternative social orders are embodied and intertwined.

3. Ritual and Belief in Morocco

In the late 1960s, Morocco became 'good to think with' for understanding religion in complex societies, just as the Trobriand Islands had earlier entered the anthropological canon for understanding exchange systems and India for understanding caste and stratification. The ethnographies of Morocco's colonial era did not lend themselves to comparative studies or lead to disciplinary classics. The publication of Clifford Geertz's *Islam Observed* (1971), originally a series of lectures delivered to historians of religion, marked a breakthrough in linking religious studies in general to the ethnographic study of specific cultural traditions. Geertz took for granted the unity of the Islamic

tradition, but emphasized the highly divergent ways in time and space that it articulated with social experience. His point of departure was Islam as experienced, interpreted, and practiced from the fifteenth century to the present in Indonesia and Morocco, the two antipodes of the Muslim majority world. For Geertz, Islamic symbols and meanings are culturally defined and socially shared, like all other systems of meaning. If commonsense gives shared meaning to immediate practical experience, religion offers a higher and more general interpretive order. These two systems of meaning—commonsense and religion—are intricately intertwined, both shaped by historical experience and giving shape to it.

A key element of Geertz's argument was to explore the constructed meanings of sainthood and Sufism (mysticism) in the two contrasting settings. In Morocco, individuals with highly varied attributes—reputations for piety, scholarship, miracle working, or other distinctions—were popularly imputed with saintly authority. Since the current Moroccan social order took hold in the fifteenth century, Geertz argued, the cultural order in which Moroccan ideas of saintly authority was sustained was characterized by a 'strong man' politics of zealous warriors, rulers, and saints enforcing their will and inducing clients to act on their behalf. Their authority was charismatic, based on personal dominance. In contrast, the Indonesian view of sainthood is hierarchical. Saints incrementally achieved prominence through years of meditation and Indonesia's settled agricultural villages placed high value on cooperation and order.

In the contexts of both Indonesia and Morocco, the idea of saintly authority took significantly divergent forms because of the contrasting nexus of commonsense perceptions of the world and the different historically constructed senses of meaning through which Indonesians and Moroccans interpreted Islam. These commonsense perceptions and background understandings form the necessary point of departure for understanding religious belief, experience, and practice. Indeed, the perspectives of religion and commonsense are intertwined in a dynamic, often unstable, dialectic (Zein 1977).

Theologians and 'common' people interpret religious texts, including the 'foundation' texts of religious traditions such as the Qur'an, in multiple and contested ways. Thus, the most common term for saint in contemporary Morocco is 'pious one' (*salih*), although 'marabout,' the term used by Geertz and which has made its way into French and English usage, is also understood. 'Pious one' conveys the multivalence of Moroccan usage. For educated Moroccans, the term is reserved solely for individuals of outstanding piety, and many speak of saint, or maraboutic, cults as 'pre-Islamic' vestiges. For many other Moroccans, however, both in towns and in rural areas, the 'pious ones,' living or dead, are imputed with mystical powers, including the ability to secure positive outcomes for individuals and groups who have maintained a connection with them through gifts or donations to their descendants (Eickelman 1976). Moroccan monarchs and senior government officials continue publicly to visit the shrines of pious ones and make gifts toward their maintenance. Elsewhere in the Muslim world there is a similar interplay between different notions of piety, and there is a similar dialectic of ideas of sainthood among Jews of North African origin (Weingrod 1990).

The emphasis Geertz placed on the cultural dimension in understanding religious experience suggested an alternative to characterizing religious traditions in terms of an essential core prevalent in all places and for all historical periods. His approach instead stresses how systems of religious meaning and practice, like all cultural experience, are contested and dynamic, and therefore not reduced to a single, cohesive set of principles. Some conceptions of Islam are significantly intertwined with, although not limited to, the local social order. However, other conceptions are more universalistic and global, amenable to articulation with a wide range of contexts. These universalistic and particularistic conceptions are copresent and in dynamic tension with one another.

Geertz's approach, like other efforts to understand religious expression and practice in Moroccan society, including Ernest Gellner's *Saints of the Atlas* (1969, see also Gellner 1981) had significant implications for understanding religious expression in earlier historical periods and outside the Muslim majority world. For example, Brown (1981) notes that his understanding of the moral order of early Latin Christianity, the 'lasting, urgent problems of any civilized community—friendship and protection, justice, power and mercy'—were in part inspired by ethnographies of contemporary North African saints. Similarly, Cornell (1998) shows how medieval Moroccan arguments on the nature of sainthood—cultural understanding of and debates among mystics in both urban and rural settings about concepts related to God's love and authority from the eleventh through sixteenth centuries—contributed to the Sufi tradition elsewhere in the Muslim world. His argument also compares the North African experience with sainthood as understood and practiced in Europe in the same period. Writings on Judaism in North African and Middle Eastern societies likewise indicate a reemphasis on the cultural dimension and explore the significance of local cultural contexts (Deshen 1989, Goldberg 1996).

4. *The Iranian Revolution and Religious Politics*

The 1978–9 Iranian revolution was one of several major events that indicated the continued significance of religion in politics and public life, notwithstanding its neglect in most mid-twentieth century discussions concerning modernization and modernity. In Iran

itself, the revolution led to the creation of a new, rapidly changing religious idiom for interpreting religious, social, and political events (Khosrokhavar 1997). The debates in Iran continue unabated over the form that an Islamic state should take and the role that religion should play in society, including gender relations (e.g., Mir-Hosseini 1999). Successive incremental decisions and the unanticipated outcomes of earlier ones meant that the revolution looked different in the decades after the event both to many of those who participated in it and to the majority of Iranians not even born when it happened. Some observers use the term 'post-Islamism' to describe Iranian society in the 1990s, emphasizing the significantly different way that many Iranians, including many clerics, began to argue publicly for a more limited role of religious institutions in Iranian politics and public life (Khosrokhavar and Roy 1999, Adelkhah 2000).

In spite of contemporaneous hopes for some and fears for others at the time of the Iranian revolution, it did not trigger similar political movements in other Muslim majority countries. It did, however, compel regimes elsewhere in the Middle East, both secular (including Egypt, Turkey, Syria, and Algeria) and religiously oriented (including Morocco and Saudi Arabia) to become more attentive to popular understandings of social justice and religion in society. From the first decades of the twentieth century, new religious movements emerged that had distinctly modern political forms. Egypt's Muslim Brotherhood, founded in the late 1920s and regarded by the mid-twentieth century as Egypt's only genuinely popular mass political movement, foreshadowed the growth of similar later Islamic movements. Distinctly modern in organizational form, with a clear hierarchy and (at least at the outset) formal criteria for membership, the Brotherhood's leadership rested not with traditionally educated religious scholars but with those effective in organizing their followers and articulating their values.

By the 1960s, however, the Muslim Brotherhood and similar organizations looked conservative in contrast to the proliferation of movements advocating the Islamization of politics and society. These 'Islamist' movements ranged from small, radical groups such as the one responsible for the assassination of Egypt's Anwar Sadat in 1981 to larger ones sometimes organized like political parties, such as Hizbullah of Lebanon, the Virtue Party of Turkey, or the Islamic Salvation Front of Algeria. Some of these Islamist movements had their origins as 'spokesmen for the despised,' and had Jewish and Christian analogues (Appleby 1997, Marty and Appleby 1995). In general, the influence of extremists, having taken off in the late 1960s, had waned by the 1990s (Kepel 2000). Most movements work within existing national boundaries, even if they sometimes receive external or transnational support. Some movements have advocated violence to achieve their goals, but others have been strong advocates of tolerance and civil society, and many movements have altered as they enter the mainstream of national political life.

5. *Religion as System and Ideology*

A distinguishing feature of the latter half of the twentieth century is the role of mass education and mass communications in changing how religious knowledge and practice are produced and understood in the Middle East. In the Arab world, for example, large numbers of people now have the level of literacy needed to follow religious and political debates from multiple sources and the command of formal language needed to participate effectively in public life. The religious lessons learned from state schooling may be regarded with skepticism, but like other aspects of state schooling, they inculcate the habits of thought and language needed to learn about and elaborate alternative views of the role of religion in society.

Ironically, the ideas of many radical Islamists mirror competing secular ideologies in form and structure. The writings of Sayyid Qutb, executed by Nasser's Egypt in 1966, attracted a wide audience among white-collar workers, professionals, and students, the beneficiaries of modern education. In Arabic and in translation, Qutb's writings continue to have widespread appeal because they offer 'Islamic' explanations for contemporary political and economic developments and for the perceived injustices of existing regimes. More contemporary religious ideologues, like Morocco's 'Abd al-Salam Yasin, argue that contemporary Muslim societies have been de-Islamized by imported ideologies and values. Yet the style and content of these writers, like others, imply an audience already familiar with the secular, imported ideologies against which they argue. Implicitly, they also convey a sense of Islam as an ideology and objectified system of meaning, which can be conveyed as one subject among others in schools, and in catechism-like pamphlets. In this sense, religion forms one of several systems of meaning in modern society, even as some believers argue that Islam is a 'complete' way of life.

The awareness of alternatives to immediate social circumstances is facilitated by the increased availability of contemporary forms of communication, ranging from the press and broadcast media to fax machines, audio- and videocassettes, increasingly ubiquitous telephones, and the Internet. In an earlier era, state authorities easily regulated the media. By the end of the twentieth century, the top-down asymmetry of broadcast radio and television were reversed by newer media that are readily adaptive to horizontal communications and difficult for states effectively to control. These developments have contributed to the fragmentation of religious and political authority by allowing large numbers of people to participate in an increasingly public debate over the symbolic language

of Islam (Eickelman and Anderson 1999). Combined with political change and the growth of diaspora communities, these developments indicate a changed social and cultural environment for religious expression and practice in which a mass and anonymous public has given way to one characterized by an increased multiplicity of horizontal relationships. The shift to a network society has created communities and communicative spaces in which large numbers of people, while not abandoning their ties of family and community, discover and create alternative ways of thinking about the role of the sacred in society and politics.

See also: Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns

Bibliography

Adelkhah F 2000 *Being Modern in Iran*. Columbia University Press, New York

Appleby M (ed.) 1997 *Spokesmen for the Despised: Fundamentalist Leaders of the Middle East*. University of Chicago Press, Chicago

Boddy J 1989 *Wombs and Alien Spirits: Women, Men, and the Zar Cult in Northern Sudan*. University of Wisconsin Press, Madison, WI

Brown P 1981 *The Cult of the Saints: Its Rise and Function in Latin Christianity*. University of Chicago Press, Chicago

Cornell V J 1998 *Realm of the Saints: Power and Authority in Moroccan Sufism*. University of Texas Press, Austin, TX

Courbage Y, Fargues P 1995 *Christians and Jews under Islam*. Tauris, London

Deshen S 1989 *The Mellah Society: Jewish Community in Sherifian Morocco*. University of Chicago Press, Chicago

Eickelman D F 1976 *Moroccan Islam: Tradition and Society in a Pilgrimage Center*. University of Texas Press, Austin, TX

Eickelman D F, Anderson J 1999 Redefining muslim publics. In: Eickelman D F, Anderson J (eds.) *New Media in the Muslim World: The Emerging Public Sphere*. Indiana University Press, Bloomington, IN

Geertz C 1971 *Islam Observed: Religious Development in Morocco and Indonesia*. University of Chicago Press, Chicago

Gellner E 1969 *Saints of the Atlas*. University of Chicago Press, Chicago

Gellner E 1981 *Muslim Society*. Cambridge University Press, Cambridge, UK

Goldberg H (ed.) 1996 *Sephardi and Middle Eastern Jewries: History and Culture*. Indiana University Press, Bloomington, IN

Hammoudi A 1993 *The Victim and its Masks: An Essay on Sacrifice and Masquerade in the Maghreb*. University of Chicago Press, Chicago

Kepel G 2000 *Jihad, expansion et déclin de l'Islamisme*. Gallimard, Paris

Khosrokhavar F 1997 *Anthropologie de la révolution Iranienne: La rêve impossible*. L'Harmattan, Paris

Khosrokhavar F, Roy O 1999 *Iran: Comment sortir d'une révolution religieuse*. Seuil, Paris

Marty M, Appleby R S (eds.) 1995 *Fundamentalisms Compared*. University of Chicago Press, Chicago

Mir-Hosseini Z 1999 *Islam and Gender: The Religious Debate in Contemporary Iran*. Princeton University Press, Princeton, NJ

Planhol X 1997 *Minorités en Islam: Géographie politique et sociale*. Flammarion, Paris

Tapper N, Tapper R 1987 The birth of the Prophet: Ritual and gender in Turkish Islam. In: *Man* (n.s.) **22**: 69–92

Von Grunebaum G E (ed.) 1955 *Unity and Variety in Muslim Civilization*. University of Chicago Press, Chicago

Weingrod A 1990 *The Saint of Beersheba*. State University of New York Press, Albany, NY

Zein A H 1977 Beyond ideology and theology: The search for the anthropology of Islam. *Annual Review of Anthropology* **6**: 227–54

D. F. Eickelman

Near Middle East/North African Studies: Society and History

The term 'Middle East' first came into use early in the twentieth century, and after World War II it gradually supplanted the nineteenth century term 'Near East' as a geopolitical and geocultural designation. Both terms have been repeatedly criticized for their Eurocentric origin and point of emphasis, yet the 'Middle East' continues to survive as a term for lack of better alternatives. It supplanted older designations such as the Levant, which designated the eastern Mediterranean and included present-day Turkey and Greece. Geographically, the Middle East encompasses West Asia, including Iran and Turkey, and North Africa, including the Sudan. Some scholarly journals extend its boundaries further to include Afghanistan and the horn of Africa. While in modern history the region has been divided among more than two dozen States, many historians and social scientists assume its various parts to contain roughly homogenous cultural patterns and sociological profiles, shaped by a shared history. Middle Eastern culture is frequently discussed in religious rubrics, and religiosity (specifically Islamic) is frequently thought to capture what is essential about Middle Eastern culture in general. An earlier tradition in Western social science regarded Middle Eastern society culturally as tradition-bound and sociologically as a tribally identified and ethnically segmented premodern mosaic kept together normally by an authoritarian State.

The social scientific study of the region we now call Middle East is actually quite old, although its history is very uneven. Ibn Khaldun (d. 1406), who spent most of his life in North Africa, was the first person anywhere to use the term 'sociology,' which he understood as the science of the life cycles of urban civilization. His theories, which emphasized the role of

social solidarity and the weakening of that solidarity over time as civilization became more entrenched and oriented toward surplus and luxury, anticipated by centuries European theories of civilizational cycles, expressed most loudly by Arnold Toynbee and Oswald Spengler. Ibn Khaldun's work must also be understood in the context of his time, which was characterized by conflicting claims of legitimacy, unstable political orders, recurrent Mongol invasions, the political fragmentation of the Muslim world, and the demonstrated incapacity of sovereigns to claim monopoly over the loyalties of the faithful everywhere, as had been roughly the case in the first three centuries of Islam.

An important feature of Ibn Khaldun's perspective consisted of combining the study of society and history, in the underlying belief that a sociology of civilization could only be based on an adequate philosophy of history. This continues to be the spirit characterizing much of the sociological investigation done in the Middle East itself, where many practitioners are just as likely to be historians or philosophers, and where questions of culture, religion, heritage, and modernity are inseparable from a myriad of contemporary social issues.

The modern professional study of the Middle East began in the nineteenth century from two directions. In the East, there arose an intellectual movement which came to be known as the *nahda*, whose significance has been pointed out in a recent flow of important studies by Albert Hourani, Hisham Sharabi, Aziz al-Azmeh, and Sadiq Jalal al-'Azm, among many others. That movement was essentially an eclectic mix of reflexive native commentators whose thought was in tune with urban middle-class sentiments and in communication with European modernity. They sought to revive what they regarded to be a dormant spirit of science and liberalism in Islam, while seeking politically to modernize the Ottoman State. In Europe, by contrast, the study of the Middle East was far more scholastic and confined to a small circle of orientalist scholars who, in the course of the nineteenth century, became increasingly better acquainted with the classic texts and archaeology of the region, especially following Napoleon's Egyptian campaign. The orientalist approach was highly anti-sociological and anti-theoretical. It focused almost exclusively on ancient history and disregarded contemporary society, which was thought to be largely static in terms of its cultural patterns, and thus accessible through knowledge of the classic texts founding it. Influential trends within orientalism viewed oriental society culturally as static, anti-rational, unphilosophical and anti-scientific; sociologically as mosaic of floating tribal and ethnic solidarities; and politically as predisposed to authoritarianism.

This approach gained further ascendance within sociology itself with Max Weber's ventures into the study of oriental cultures, including Islam. In common with the reigning orientalist tradition, Weber's approach to the question as to why Muslim societies (along with Buddhist and Hindu ones) had not developed capitalism as Europe had done, referred to the Qur'an and sayings attributed to Muhammad, and assumed the cause of social change (or lack of it) to be internal rather than external to the culture. While Weber was more interested in social theory than were orientalist scholars, who were preoccupied with philological and textual concerns, Weber nonetheless replicated the basic image of oriental society as a self-enclosed and distinct cultural cocoon, interpretable by its classical texts and definable in terms of its difference from the European experience of liberal capitalism.

The subsequent rise of modernization theory, closely associated with the Parsonian paradigm in sociology, set aside for some time the question of culture, emphasizing instead the normative convergence of societies along the path of modernization. Daniel Lerner's *The Passing of Traditional Society* (1958) is a classic work in this genre with respect to the Middle East. While steeped in orientalist assumptions that Middle Eastern society, culture, and personality was largely tradition-bound and anti-rational, Lerner's work sought to detect venues of the modernist challenge arriving to society from without. Like many theorists of modernization, Lerner understood it as simply 'Westernization.' Another important text of modernization theory was C. A. O. van Nieuwenhuijze's *The Sociology of the Middle East* (1971), a detailed and lengthy study which portrayed Middle Eastern society as a set of parallel societies converging upon a cultural core. This further justified the old view of culture as an independent variable when approaching Middle Eastern societies.

With decolonization, the increased presence of the US in the region, the Cold War, the creation of the State of Israel, the rise of Arab nationalism and the Arab defeat in 1967, Middle East studies itself became a veritable battleground. A new generation of scholars was determined to contest what they regarded to be a distorted image of the region and its peoples in both academic and public settings. The rallying cry was issued by Edward Said's *Orientalism* (1978), which immediately galvanized the field and continues to be referenced today as a founding text in postcolonial studies generally. A sophisticated critical appraisal of the academic and high culture representations of largely Middle Eastern societies, *Orientalism* was conceived by Said as part of a trilogy, which included two subsequent books on the media coverage of Islam and the Palestine question. Said sought to reveal classical orientalism, which modernization theory did little to contest, as the intellectual regurgitation of common stereotypes based less on social and historical realities than on the tendency of European intellectuals to imagine a distinction between Europe and the rest of the world.

While Said's analysis was directed largely at mainstream trends in Western scholarship of the Middle East, another set of scholars set out around the same time to contest the heritage of orientalism in left-leaning scholarship. The main point of contention here was Karl Marx's thesis of an 'Asiatic mode of production,' distinct from European feudalism in its propensity to stagnate after a founding phase, needing a strong central State for its maintenance, and incapable of producing mercantile surplus that could found a basis for capitalism. While his method was fundamentally different from Weber's, Marx nonetheless reached the same conclusion regarding the foreignness of capitalism to the orient in general, and thus positing the need for an external stimulus, such as European colonialism, to introduce the bases of bourgeois accumulation and, in turn, the precondition of proletarian socialism. In spite of the brevity and inconclusive nature of Marx's own remarks on this theme, which addressed sporadically the encounter between India and Algeria with European colonialism, his assessment of the lack of internal capacity of oriental societies for capitalism and thus revolution has reigned well into our times. For example, Shlomo Avineri, a Marxist Israeli political theorist, has argued that Israel, by introducing into 'traditional' Middle Eastern society the same forces of destruction of 'Asiatic' patterns, would also add a revolutionary stimulus to the sum total of productive forces in the region.

This line of thinking was contested in a number of widely-referenced works from the recent past. They include Maxime Rodinson's *Islam and Capitalism* (1973), which took both Weber and Marx to task for failing to appreciate the paramount role of the urban merchant class throughout the economic history of the Middle East, for their propensity for partial appraisal of Islamic economic doctrines, for their tendency to confuse religion and culture, and for treating Islam, in particular, as a monolithic force. Rodinson argued that the Muslim world in general had a historically strong merchant capitalist class, and that, if anything, Islam was more of an aid than an obstacle to its interests. Mahmood Ibrahim took the argument further, arguing in *Merchant Capital and Islam* (1990) that the rise of Islam itself must be understood in terms of its connection to the interests of merchant capital and long-distance trade. The argument was also made in Janet Abu-Lughod's *Before European Hegemony* (1989) where, using a world-system perspective, she shows both the centrality of the Middle East to premodern world systems and the evolution of common modes of capitalist thinking and planning among commercial classes throughout the world well before the rise of the West.

The academic attacks on the leftist variety of orientalism were sealed in a number of other works which sought to demonstrate the capitalist vitality of Middle Eastern societies in early modern times. These include Peter Gran's *Islamic Roots of Capitalism: Egypt, 1760–1840* (1979) and Surayah Faroqhi's *Towns and Townsmen of Ottoman Anatolia* (1984). In *Marx and the End of Orientalism* (1978), Bryan Turner attempts to revalidate Marxist analysis and the mode-of-production approach to Middle East studies by repudiating the Asiatic mode of production thesis altogether and presenting a more tenable historical and theoretical alternative. An overarching treatment of the subject matter from a much broader perspective is Halil Inalcik's *Economic and Social History of the Ottoman Empire: 1300–1914* (1994).

Those studies also occasion the beginning of the application of a political economy framework to the study of Middle Eastern history and society. In his pioneering *The Middle East in World Economy: 1800–1914* (1981) and subsequent works, Roger Owen identifies major global economic factors at work in the final century of Ottoman domination; factors which reshuffled the social structure within the region. Owen argues that Europe's increased demand for agricultural commodities during that period strengthened coastal parts of the Middle East at the expense of internal ones, and produced small local elites tied to European commerce. Those elites failed, nonetheless, to invest back in agriculture and industry, thus causing a development in the direction of dependency rather than modern, locally-rooted capitalism. Owen's thesis marks an avowed application of dependency theory, hitherto common in Latin American studies, to Middle Eastern studies. A similar approach which situates the history of Iran, omitted from his analysis by Owen, in the context of the world system is John Foran's *Fragile Resistance* (1993). Other political economists such as Yusif Sayigh (1979), Abbas Alnasrawi, and Jacqueline Ismail do the same for other parts of the region.

Some works in this genre stand out because they attempt to connect political economic analysis to historical formations in the culture at large. In *The Arab Nation* (1978), Samir Amin outlines the basis of a pan-Arab nation in a historical, far-flung urban network maintained by a commercial warrior class, that generated the surplus upon which it depended from long-distance trade rather than from the adjoining peasantry, which remained unincorporated. This feature meant that the fate of the Arab nation would always be dependent on factors external to it, and thus explained why the Arab nation readily entered the modern era in the logic described by dependency theory.

Of particular concern in studies of the political economy of the Middle East is the question of the nature and role of the State. The fact that for much of its history the Middle East was governed by loose, territorially expansive States rather than confined city-states, meant that local civil society, social networks and even identities could exist and evolve at a distance from State doctrines. The historian Roy Mottahedeh,

for example, has shown how medieval Muslim society in Buyid Iran could function without the trappings of the State as such. While in general the term 'nation-state' would not adequately describe Middle Eastern polities, there were notable attempted exceptions. These involved projects aiming at building nation-states along the European model, as in nineteenth century Egypt and Kemalist Turkey following the dissolution of the Ottoman empire in the aftermath of World War I. The State Mehmed Ali set out to build in Egypt early in the nineteenth century paralleled the European State in its extension of the range of its social roles, its curtailment of clerical prerogatives, the strengthening of direct state presence in the countryside, its reliance on regular levies and taxes, and the introduction of educational and health reforms. As in Europe, that kind of State began to assume for itself many functions that had historically been within the domain of civil society.

But such state experiences tended to be the exceptions, and even the Egyptian State was unable to demonstrate credibly its sovereignty in the face of colonialism. On the other extreme of experiences of state formation is the Persian Gulf region, whose political history until modern times had known only ephemeral States at best. In *State and Society in the Gulf and the Arab Peninsula* (1990) Khaldoun H. al-Naqeeb identified the pre-eighteenth century political economy of the Gulf as a 'natural economy,' because all its components, including long-distance trade across the Indian Ocean, lacked the services and supervision of a centralizing state apparatus. Al-Naqeeb's thesis was that such a natural economy was eventually destroyed by British imperialism, which supported the rise of new dynasties and authoritarian, rentier States. Such States thus had to be run (until today) as official or unofficial protectorates.

This reality, coupled with the newness of all Arab States in the Levant and North Africa, has led some commentators like Michael Hudson in *Arab Politics: The Search for Legitimacy* (1977) to identify the question of legitimacy as the central problem of postcolonial polities in the region. This stance essentially dismisses the relevance of much of the common practice in political science writing on the region, which generally takes States as normative categories of analysis, focuses attention on diplomatic history and games theory, and pays little attention to ideologies, social movements, political parties and organizations, and also to such questions as political alienation and, above all, legitimacy.

New directions of research thus began in the early twenty-first century to address more complicated issues regarding the nature of the State in the region. At least three directions of research can be identified in this connection. One involves projects oriented toward studying the historical sociology of the State in the Middle East, spearheaded by Elia Zureik, Talal Asad, Huri Islamoglu-Inan, and Khaled Fahmy. Another concerns civil society which, in the context of Middle Eastern studies, is usually seen as both an antidote to the State and a factor in democratizing it. Major works in this direction are led by Saad Eddin Ibrahim's series of reports on civil society and democracy in various countries of the region, and similar collections of reports edited by Jillian Schwedler and Augustus R. circa 1995. A third direction of research focuses on state rituals and symbolism, a good example of which is Peter Chelkowski and Hamid Dabashi's study of the Iranian post-revolutionary State, *Staging a Revolution* (1999). Also significant in this genre is Lisa Wedeen's *The Ambiguities of Domination* (1999), which employs ethnographic methods to study the nuances of popular responses to state symbolism in Syria. Earlier important studies in this genre by Sami Zubaida (1989), Ernest Gellner (1981) and Clifford Geertz focus on the state appropriation of popular beliefs, and demonstrated how the State tends to cause religious ideology and symbols to become more rigid and orthodox when it adopts them for its own use.

These studies have helped move the center of gravity in the sociological study of the Middle East away from a focus on formal institutions and historical continuities, to a focus on the determining relations between social structures and social movements. Hanna Batatu's classic *The Old Social Classes and the Revolutionary Movements of Iraq* (1978), as well his more recent work on Syria, combine Marxist and Weberian perspectives to show how modern social movements could be understood in terms of the transformation in the basis of status groups from tribal, religious, and official dignitaries to landed aristocratic and merchant capitalist classes.

Batatu's landmark study hails from a new trend which rejects a well-established view of religion as preconstitutive and internal to the culture. That view is common in orientalist tracts, but also in a tradition of sociological commentary ranging from Reuben Levy to S. N. Eisenstadt, and is well represented in the contemporary *Fundamentalisms Project*, edited by Martin Marty and Scott Appleby. More recent trends by such anthropologists as Talal Asad (1993) and Ralph Coury treat both Islam and culture as dependent variables, and stress the historical dynamism and multiplicity of both. With respect to modernity, Talal Asad shows how religion itself can foster modernizing tendencies, and Ernest Gellner, who had otherwise stressed the similarities across Muslim societies and the preconstitutive role of religion, also comes to argue the special closeness of Islam to the spirit of modernity, and its ability to encompass various types of social structures.

Many observers, however, continue to be deeply interested in the dynamics operating between more permanent historical social structures and contemporary ideologies and social movements. Sami Zubaida's work stands out in this regards as an exposition of the complexity of determining forces,

and for demonstrating the need for an integrated analytical grid of cultural, sociological, comparative, and historical knowledge to apprehend the nuances of modern movements. Other scholars emphasize the determining role of specific sub-State social formations, such as the Ottoman *Millet* system, whereby various religious communities related to the State as recognized and legally autonomous corporate bodies. Main research questions in this regard concern the extent to which such a system gave rise to or inhibited the full expression of modern movements, such as nationalism, sectarianism, or secularism.

Scholars of the region are increasingly inclined to regard seemingly 'premodern' identities expressed in the political culture of the region as actually *modern* inventions. They also tend to stress the rootedness of modern social movements in external stimuli, such as colonialism and foreign influence. An important recent example of this scholarly trend is Ussama Makdisi's *The Culture of Sectarianism* (2000). Focusing on nineteenth century Lebanon, Makdisi shows how confessional identities along which the modern Lebanese system was built were manifestations of various processes of modernity rather than expressions of reaction to it, and how they were fostered by Western intervention rather than emerging on their own. Elizabeth Thompson's *Colonial Citizens* (2000) shows how it was the intrusion of colonial powers, rather than indigenous traditionalism, that had been detrimental to both women and liberal thought generally in Syria and Lebanon. Lacking roots in the colonies and conceiving of them as largely tradition-bound territories, colonial powers weakened indigenous liberalism, because the latter adhered to anti-colonial nationalism, and strengthened, through its compensatory alliance with them, the more conservative forces, whose influence had been declining before the advent of colonialism.

Many commentators also view the emergence of modern nationalism, in particular, in the context of colonialism rather than that of indigenous developments. Because of its extremity, the case of the Palestinians is usually cited as the most evident, but by no means the only, illustration of this centrality of external intervening factors to the framing of modern identities. An important study in this regard is Rosemary Sayigh's *Palestinians: From Peasants to Revolutionaries* (1979), which uses oral history to show the growth of nationalism due to external colonizing pressure. A more recent work along these lines is Rashid Khalidi's *Palestinian Identity* (1997), which shows the gradual emergence of that identity in reaction to Jewish nationalism in Palestine. Similarly, in examining the Palestinian *Intifida*, Samih Farsoun and Jean Landis reject the relevance of most trends in social movement theory, emphasizing instead the global context of Palestinian struggle.

The pivotal geopolitical status of the modern Middle East, which entailed massive foreign intervention and a plethora of movements of resistance as well as political fragmentation, gave rise to a rich tradition of engaged scholarship on social movements. Khalil Nakhleh, (1977) a Palestinian sociologist, has rejected outright any possibility of non-reflexive, detached, 'normal,' or 'nonnative' sociology under current circumstances in the region. Other scholars have sought to outline oppositional social movements in terms which endow them with an under-explored sense of agency. Examples of this approach include Joel Beinin and Zachary Lockman's *Workers on the Nile* (1987), and Anouar Abdel-Malek's *Contemporary Arab Political Thought* (1983), which moves away from cultural and structural determinism altogether and stresses the priority of agency under modern circumstances.

Apart from questions of colonial encounter and political economy, an important line of emphasis in studying social movements in the Middle East concerns how they make use of cultural symbols. Julia Clancy-Smith's *Rebel and Saint* (1994) shows how resistance to colonialism in North Africa used icons and authorities from traditional settings whose function in society had until then been quite apolitical. In a nuanced analysis of the new veiling and work environments in Cairo, Arlene Elowe Macleod (1991) shows how the veil today, far from being simply a patriarchal imposition, is actually used proactively by women to express various messages combining both accommodation and protest.

These studies stem from a very weighty orientation in Middle East studies, particularly in Middle Eastern countries, toward questions of heritage and its connection to modernity. Many important contemporary commentators, such as Fazlur Rahman, Seyyed Hossein Nasr, Syed M. N. al-Attas, and Ismail al-Faruqi have revisited classical Islamic philosophy and science in a quest to reveal a specifically Islamic philosophy of knowledge and science in tune with modernity. More politically-inclined contemporary philosopher-activists (frequently Shi'ites), such as Ali Shariati, Mohammed Baqr al-Sadr, and Morteza Motahhari, have also revisited classical sources extensively in order to show the compatibility of Islam with modern ideologies, notably socialism and anti-colonial nationalism. The same is true of Tayyib Tazzini's and Hussein Muruwwah's secular reading of the heritage. The voluminous contributions of the latter, who was killed by Amal fighters in Beirut in 1985, were geared toward showing how classical Islamic philosophy and thought had always housed 'materialist' perspectives and methods in tune with Marxist dialectics.

Many other commentators are equally interested in voluminous, encyclopedic surveys of classical thought, but emphasize as organizing principles discursive, semiological, and hermeneutic rather than sociohistorical approaches. These include Muhammad 'Abid al-Jabiri's analysis over many volumes of the struc-

tures of knowledge in Arab culture and the poet Adonis's (Ahmad 'Ali Sa'id) equally critical study of imitation and originality in Arab culture, *The Permanent and the Changeable* (1974). While many of these studies seek to discover through historical investigation a root cause of contemporary 'stagnation' or 'backwardness,' many more seek to reveal through the same method the dynamism and adaptability of the heritage to various social structures and ideological claims. This latter approach to the heritage had required an abandonment of the philological and preconstitutive approach of orientalism. Among the most important founders of this critical approach is Muhammad Arkoun, whose pioneering studies of the Qur'an have emphasized the value of other than purely philological approaches to the holy text, such as anthropological, sociological, and semiological approaches. Mohammed Bamyeh's *The Social Origins of Islam* (1999) builds on that tradition of combined methods of reading, arguing that the growth of Islam was precisely due to an ability exhibited at its founding phase to accommodate the worldviews of different and even discordant social classes and types of social structure. Abdul-Kabir al-Khatibi echoes a similar theme when he argues that Islam must be seen as a sort of 'compromise' religion, and that its historical vitality stems from that attitude rather than from an alleged orthodox base and orientation.

In light of these new perspectives, the question arises again as to why this demonstrated dynamism of the heritage has not lent itself to a more vocal communion with liberal European modernity. Many responses to this question tend to dismiss it outright, citing as its faulty premise a Eurocentric tendency to outline various historical trajectories as more sharply differentiated than they actually were. In his critique of Eurocentrism, Samir Amin builds on and extends Martin Bernal's thesis in *Black Athena* (1987). Amin contests European fabrication of a monopolistic claim on Greek heritage, as well as the construction of European history through a direct line that flows from Greece to Rome. This perspective, according to him, ignores the larger Mediterranean contexts of classic civilization and the continuity and further development of Greek political and philosophical thought, as well as mathematics, medicine, and science in the Islamic world during Europe's Middle Ages.

Criticism of the tendency to view Middle Eastern or Islamic history in terms too distinct from other world histories has also been levied by the urban historian and sociologist Janet Abu-Lughod (1989) who, after her initial study of Cairo, came to repudiate the notion that there was anything specifically Islamic about the 'Islamic city.' For others, such as Jabiri and Hichem Djait, (1974) the point became less relevant, as they identify the important lines of demarcation with respect to modernity not along large East/West types of categories. Rather, they prefer subregional categories, such as Arab Mashreq (Western Asia and Egypt), vs. Arab Maghreb (rest of North Africa). Djait argues that the relatively more intense and enduring colonialism experienced by the Maghreb caused a more cultural proximity to modernity, since all illusions about self and society were destroyed through a colonial uprooting that had not been as severe in the Mashreq. The latter, being politically more advanced, had obtained its independence earlier than the Maghreb and proceeded to be governed by dictatorial regimes. That condition, according to Djait, favored the construction of more illusions about self and society in the Mashreq, and thus more relative distance from the spirit of modernity.

Within discussion of the proper attitude to modernity, the question of gender has had a particular resonance. The feminist sociologist Fatima Mernissi attained fame by blaming Islam itself for perpetuating a retrograde attitude toward women. An alternative trajectory within feminist writings on Islam, represented in the works of Leila Ahmed, Nawal Sa'dawi and Denise Spellberg, offers more complex readings of the role of women in Islamic history and places more emphasis on the agency of women. More recent trends have de-emphasized the alleged role of traditional religion even further as an explanatory factor. Deniz Kandiyoti's (1991) work has contributed to shifting attention away from religion as such and into patriarchal structures as sources of the oppression of women. Hisham Sharabi's pioneering study *Neopatriarchy: A Theory of Distorted Change in Arab Society* (1988) singles out structural changes within the Arab World as well as elsewhere within the past century that had led to 'neopatriarchy' instead of modernity. Neopatriarchy is explained by Sharabi as a form of patriarchy produced by dependent capitalism and social change lacking in a sense of authenticity. Thus neopatriarchy tends to distort modernization and nervously presents itself as the only valid option.

Given the magnitude of social, political, economic, and cultural transformation in the Middle East over the twentieth century, many social scientists will find some literary works very useful in apprehending change in certain areas. Three particular novelists stand out for their ability to reconstruct history and society in more overarching and encyclopedic ways than has so far been possible in the social sciences, and thus are worthy of greater attention. Adbelrahman Munif's (1989) multivolume epic *Cities of Salt* reconstructs in elaborate and patient detail the coming into being of Saudi Arabia in the 1930s, following the discovery of oil in the country by Western companies. The novel documents transformations in social relations and customs, the introduction of wage labor on top of an austere, semi-nomadic economy, the emergence of a dynasty and a State, the resulting tensions between authenticity and inauthenticity in public imagination, and other fascinating details which are usually glossed over in more dispassionate sociological and historical writings.

Gamal Ghitani's historical novel *Zayni Barakat* (1990), set at the time of Ottoman takeover of Egypt from the Mamluks early in the sixteenth century, reconstructs elements of continuity and change in political structure as well as in the relationship between spiritual and political authority over the time-span of the two systems. Finally Nobel laureate's Naguib Mahfuz's prolific career is dedicated to an elaborate and highly nuanced reconstruction of social psychology and its connection to mores and traditions. Characters in Mahfuz's works, who frequently mimic middle- and lower-middle class inhabitants of Cairo's neighborhoods, are usually shown as agents of their own destinies rather than as victims of established structures, and are capable of using a wide reservoir of proverbs and customary allowances of deviance to justify their actions. This holds true even in their attitude toward fate. In analyzing one famous story by Mahfuz, Halim Barakat shows that when characters claim what they are doing to be ordained by some fate, they are usually using the language of fatalism as a tool of self-assertion, rather than as an explanation of their powerlessness with respect to established patterns of relations and social expectations. At both the macro and micro levels, these works combine research orientations central to understanding Middle Eastern society and history. These orientations involve political economy, social and institutional structures and mentalities, and the flexible morality governing everyday, street-level interactions.

1. Conclusion

The study of society and history of the Middle East has gradually moved from focus on homogeneity and continuity to more elaborate discussions of the impact of many determining factors. It is now commonly accepted that due to its historical centrality to historical world systems and its contemporary geopolitical significance, Middle Eastern society and history cannot be understood apart from global trends which influence the region and are influenced by it as well. The studies are most revealing when they go above or below contemporary state jurisdictions and emphasize instead the historically autonomous mode of self-organization of civil society, cultural communities, travel, and trade. Middle Eastern society has historically been defined more so by the hybridity and diversity resulting from the above factors, than by the social engineering of States.

See also: Islam and Gender; Islam: Middle East; Islamic Fundamentalism; Modernization, Sociological Theories of; Nationalism, Historical Aspects of: Arab World; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Politics; Near Middle East/North African Studies: Religion; Orientalism

Bibliography

Abu-Lughod J L 1989 *Before European Hegemony*. Oxford University Press, New York

Al-Azmeh A 1993 *Islams and Modernities*. Verso, London

Al-Naqeeb K H 1990 *Society and State in the Gulf and Arab Peninsula*. Routledge, London

Amin S 1978 *The Arab Nation*. Zed Press, London

Arkoun M 1982 *Lectures du Coran*. Editions G.-P. Maisonneuve et Larose, Paris

Asad T 1993 *Genealogies of Religion*. Johns Hopkins University Press, Baltimore, MD

Bamyeh M A 1999 *The Social Origins of Islam*. University of Minnesota Press, Minneapolis, MN

Barakat H 1990 Beyond the always and the near: a critique of social psychological interpretations of Arab society and culture. In: Sharabi H (ed.) *Theory, Politics and the Arab World: Critical Responses*. Routledge, New York

Batatu H 1978 *The Old Social Classes and the Revolutionary Movements of Iraq*. Princeton University Press, Princeton, NJ

Beinin J, Lockman Z 1987 *Workers on the Nile: Nationalism, Communism, Islam, and the Egyptian Working Class, 1882–1954*. Princeton University Press, Princeton, NJ

Chelkowski P, Dabashi H 1999 *Staging a Revolution: The Art of Persuasion in the Islamic Republic of Iran*. New York University Press, New York

Clancy-Smith J A 1994 *Rebel and Saint: Muslim Notables, Populist Protest, Colonial Encounters*. University of California Press, Berkeley, CA

Djait H 1974 *La Personnalité et le Devenir Arabo-islamiques*. Editions du Seuil, Paris

Fahmy K 1997 *All the Pasha's Men: Mehmed Ali, His Army, and the Making of Modern Egypt*. Cambridge University Press, Cambridge, UK

Faroqhi S 1984 *Towns and Townsmen of Ottoman Anatolia*. Cambridge University Press, Cambridge, UK

Foran J 1993 *Fragile Resistance: Social Transformation in Iran From 1500 to the Revolution*. Westview Press, Boulder, CO

Ghitani G al- 1990 *Zayni Barakat*. Penguin Books, New York

Gellner E 1981 *Muslim Society*. Cambridge University Press, Cambridge, UK

Gran P 1979 *Islamic Roots of Capitalism: Egypt, 1760–1840*. University of Texas Press, Austin, TX

Hourani A 1991 *History of the Arab Peoples*. Belknap Press of Harvard University Press, Cambridge, MA

Hudson M 1977 *Arab Politics: The Search for Legitimacy*. Yale University Press, New Haven, CT

Ibrahim M 1990 *Merchant Capital and Islam*. University of Texas Press, Austin, TX

Inalcik H 1994 *An Economic and Social History of the Ottoman Empire: 1300–1914*. Cambridge University Press, Cambridge, UK

Islamoglu-Inan H (ed.) 1987 *The Ottoman Empire and the World-Economy*. Cambridge University Press, Cambridge, UK

Kandiyoti D (ed.) 1991 *Women, Islam, and the State*. Temple University Press, Philadelphia, PA

Khalidi R 1997 *Palestinian Identity*. Columbia University Press, New York

Lerner D 1958 *The Passing of Traditional Society*. Free Press, Glencoe, IL

Makdisi U S 2000 *The Culture of Sectarianism*. University of California Press, Berkeley, CA

Macleod A E 1991 *Accommodating Protest: Working Women, the New Veiling, and Change in Cairo*. Columbia University Press, New York
Mottahedeh R P 1980 *Loyalty and Leadership in an Early Islamic Society*. Princeton University Press, Princeton, NJ
Munif A 1989 *Cities of Salt*. Vintage Books, New York
Nakhleh K 1977 Anthropological and sociological studies on Arabs in Israel: critique. *Journal of Palestine Studies* **6**(4): 41–70
Nieuwenhuijze C A O van 1971 *Sociology of the Middle East*. Brill, Leiden, The Netherlands
Owen R 1981 *The Middle East in World Economy: 1800–1914*. Methuen, London
Rodinson M 1973 *Islam and Capitalism*. Pantheon Books, New York
Said E W 1978 *Orientalism*. Pantheon Books, New York
Sayigh R 1979 *Palestinians: From Peasants to Revolutionaries*. Zed Press, London
Sharabi H 1988 *Neopatriarchy: A Theory of Distorted Change in Arab Society*. Oxford University Press, New York
Thompson E 2000 *Colonial Citizens: Republican Rights, Paternal Privilege, and Gender in French Syria and Lebanon*. Columbia University Press, New York
Turner B 1978 *Marx and the End of Orientalism*. Allen & Unwin, London
Wedeen L 1999 *The Ambiguities of Domination: Politics, Rhetoric, and Symbols in Contemporary Syria*. University of Chicago Press, Chicago
Zubaida S 1989 *Islam, the People and the State: Essays on Political Ideas and Movements in the Middle East*. Routledge, London
Zureik E 1981 Theoretical considerations for a sociological study of the Arab State. *Arab Studies Quarterly* **3**(3)

M. Bamyeh

Needham, Joseph (1900–95)

Joseph Needham was not only an excellent scientist, but also an eminent historian. He became famous for his impressive achievement embodied in the massive series *Science and Civilisation in China* (SCC), published in many volumes since 1954. His work on this subject has not only deeply influenced research on the history of China, but also studies in the history of science, technology, and medicine in general. Although the exploration of the history of traditional Chinese science, technology, and medicine had begun already before him, he is certainly the scholar who contributed most to the institutionalization of these studies, resulting, both in East and West, in the foundation of specific academic research disciplines and institutes. One of his basic convictions was that *modern* science and technology, as they arose in the West, owed much to ideas and inventions originating in East Asia or China. He conceived of modern science as an outcome of an universalistic process, like a sea into which the rivers of various cultures had flowed, making their contributions. In his opinion, the reason that modern science did not arise in China was because it was hindered by external factors, that is, the structure of Chinese feudal-bureaucratic society.

1. Life

Needham was born in London on December 9, 1900, the only son of a physician and a musically talented mother. In 1918 he went to Gonville and Gaius College in Cambridge to study biochemistry, where he became a student and collaborator of Frederick Gowland Hopkins, a subsequent Nobel Prize winner in medicine. In 1924, he married the biochemist Dorothy Moyle. In 1931 Needham published *Chemical Embryology* in three volumes, a description, definition, and historical account of a new scientific discipline in which he had increasingly specialized. This was followed, in 1942, by *Biochemistry and Morphogenesis*. Already at an early stage in his career, Needham engaged in public debates about the relationships between science, religion, philosophy, history, and politics. In the period from 1925 to 1946 he wrote or edited such works as *Science, Religion, and Reality* (1925), *The Sceptical Biologist* (1926), *The Nazi Attack on International Science* (1941), *History is on Our Side* (1946), etc., the latter two attacking the Nazi abuse of science.

In 1937 a group of young Chinese scientists came to Cambridge, among them Lu Gwei-Djen, who later became one of his collaborators and, two years after the death of Dorothy Moyle in 1987, his second wife. These Chinese scientists aroused in him a deep interest in and fascination for Chinese culture and language. Needham started to learn Chinese privately with the help of a Cambridge professor of Sinology. In 1942 the British government commissioned him to go to China to establish the Sino-British Science Co-operation Office in Chongqing, the city to which the Chinese government had withdrawn in the face of the Japanese invasion. He spent more than three years in China, where, apart from his official duties, he started to collect information, literature, and source material on the history of science, technology, and medicine in China. From 1946 to 1948 Needham was appointed director of the department for science at UNESCO in Paris. In 1952 he was severely criticized for joining an international scientific committee, which on the invitation of the governments of North Korea, the People's Republic of China, and the communist-sponsored World Peace Council was commissioned to investigate charges of US biological warfare in the Korean War. The committee concluded that the US military had experimented with biological weapons, and Needham was subsequently reproached as having been deluded by communist propaganda and his love for China. As a consequence, he was shunned by many in the academic world, and was not permitted to visit the US in the coming years.

After his return to Cambridge in 1948 Needham all but gave up his scientific research, and turned instead to his great project SCC. The first volume was published in 1954. In the course of time, more than twenty collaborators from all parts of the world joined Needham to participate in this project. In 1968 Needham founded in Cambridge the East Asian History of Science Trust to preserve his library and continue his work. By 1987, funds had been raised to build a home for his library, the Needham Research Institute, which was finally completed in 1991. Together with his first wife he was elected Fellow of the Royal Society (science) in 1941. Thirty years later, he also became Fellow of the British Academy (humanities). In 1968 the Society for History of Science and the Society for History of Technology honored him with the George Sarton Medal and the Leonardo da Vinci Medal, respectively. He was bearer of the Order of the Brilliant Star conferred on him by the People's Republic of China. Finally, in June 1992, his achievements were given official recognition when he was made Companion of Honour by Her Majesty the Queen. He died in Cambridge on March 24, 1995, at the age of 94.

2. *Contribution to Knowledge and Intellectual Environment*

Based on his ecumenical and universalistic view of the history of science and technology, in which the Old World must be taken as a whole, one of Needham's crucial and seemingly paradoxical problems was why modern science, the mathematisation of hypotheses about Nature, with all its implications for advanced technology, made its meteoric rise only in the West at the time of Galileo. Of equal importance for him was the question of why it was that between the second century BC and the sixteenth century AD East Asian culture was much more efficient than the European West in applying human knowledge of Nature to useful purposes. How could it have been that the weakness of China in theory and geometrical systematization did not prevent the emergence of technological discoveries and inventions often far in advance of contemporary Europe? What were the inhibiting factors in the Chinese case that prevented a rise of modern science in Asia analogous to that which took place in Europe from the sixteenth century onwards? And finally, how was it that Chinese backwardness in scientific theory coexisted with the growth of an organic philosophy of Nature, closely resembling that which modern science (in Whitehead's sense) had adopted after three centuries of mechanical materialism?

In order to find answers to these questions, Needham carried out enormous comparative research, not only dedicated to internalist interpretations of science and technology, but also taking into account influences from philosophy, language, religion, and ideology as well as social, economic, and geographical conditions. It was likewise important for him to fix dates of discoveries and inventions, aiming to 'titrate' the great civilizations one against another and to find out and give credit where credit was due. In many cases where Needham and his collaborators were able to ascertain an earlier date of a discovery or invention in China than in the West, he concluded that this must have been a case of direct borrowing, indirect transmission, or diffusion stimulus, even if no other concrete evidence was available. Basically, he shifted the onus of proof onto those people who doubted that such a transfer had taken place. Generally speaking, however, Needham was never thinking in simplistic terms, as he did, for instance, admit the possibility of independent discoveries and inventions. From very early on he concluded that until the Scientific Revolution in the late Renaissance the West was not only deeply influenced in the field of technology by Chinese or East Asian technological inventions and scientific discoveries, but was also deeply affected by them in its social structure. For instance, the adoption in the West of the Chinese invention of gunpowder and firearms dealt a fatal blow to castles and hence to Western military aristocratic feudalism.

Needham drew a clear distinction between modern science on the one hand and ancient and medieval sciences on the other. Modern science, as it developed only in Western Europe at the time of Galileo, is based on five fundamental characteristics, namely the application of mathematical hypotheses to Nature, the full understanding and use of the experimental method, the distinction between primary and secondary qualities, the geometrisation of space, and the acceptance of a mechanical model of reality. Primitive and medieval sciences distinguish themselves clearly from the modern type, as, for instance, their intrinsic and essential vagueness always made them incapable of proof or disproof. As far as numerical figures entered into them, these numbers were manipulated in forms of numerology or number-mysticism constructed *a priori*, and not employed as the stuff of quantitative measurements compared *a posteriori*.

During his years of study and research in Cambridge, Needham's worldview was characterized by four conceptual clusters that substantially influenced his work for SCC:

(a) Needham had a strong antipathy against the mechanical concepts upheld by many physicists of his time. His philosophy of science was influenced by Whitehead's organicism. He was, however, not a rigid defender of Neo-vitalism, but together with Lancelot Hogben argued for its 'mechanical' or 'materialistic' aspects. Needham considered Whitehead's theory of organicism a synthesis of mechanism and vitalism.

(b) Needham espoused the idea of evolution as a concept of wide application. For him evolution was not only a phenomenon of the biological order, but

could also be observed in social organization. Social organizations necessarily followed a linear course of evolution, for which the Marxist stage theory of social development was an indispensable explanative tool. Social development cannot be explained by biological theory, but needs its own unique approach. Needham's concept of evolutionary development of social organizations did, however, not deal with a socio-Darwinist 'survival of the fittest,' but aimed at overcoming the excesses of capitalism. 'Collectivism,' as it would manifest itself in a worldwide communist society, was seen by Needham as the next higher stage of social organization.

(c) Needham sympathized with Marxist dialectic materialism. Most of the researchers in Hopkins's laboratory were politically left wing. This was typical of many British biologists of the twenties and thirties, like J. B. S. Haldane, J. G. Crowther, and J. D. Bernal. During his youth Needham was impressed by the events of the Russian Revolution. Moreover, in 1925 he read Engels's *Dialektik der Natur* [Dialectics of Nature] which fascinated him. During one of his stays at the Roscoff Marine Biological Station Needham and his wife met Louis Rapkine, a biochemist born in Lithuania who grew up in Canada. Rapkine introduced Needham to the Marxist classics. Moreover, he learned much of him about the living conditions of the lower strata of the working society for which Needham had always entertained great sympathy. Like Julian Huxley and the adherents of a 'scientific humanism,' Needham expanded his interests from biological problems into the social field. Due to his rational, antireductionist attitude towards science, he was convinced that scientists also had social and ethical responsibilities. Democratic, socialist, and populist elements were the main constituents of his political stance. Needham, however, did not count himself an orthodox Marxist because he was not dogmatic with regard to the question of which aspects of human action constituted foundation or superstructure. He belonged to a generation of the 1920s who adhered to Marxism because of ideology and romantic ideals, in contrast to the generation of the 1930s which was characterized by a more realistic attitude towards politics and tactics. In his later years Needham never gave up his hopes for the realization of an 'ideal socialism'.

(d) Needham stressed synthesis over analysis. It was his basic conviction that dialectic materialism was the synthesis between mechanical materialism and idealism. Religion was for him something that need not be got rid of in favor of materialism, but which should be absorbed by it. On the one hand, he stated in 1936 that communism possessed a higher sense of the numinous than any of the traditional religions, and that, on a higher stage, a synthesis has to be carried out between religion and Marxism. Communism could provide the moral theology of our time. On the other hand, Christianity was necessary for him to help redeem the ruthlessness of science. In an autobiographical account published in 1973 under the pseudonym 'Holorenshaw,' Needham informed the public about three antitheses which had determined his life. The first is the antithesis between deep religiousness and science, the second between deep religiousness and his sympathy for socialism, especially in its Marxist form, and the third referring to the differences between Eastern and Western cultures. He declared it to be his duty to transform these antitheses into syntheses. This also shows that Needham was not an adherent of scientism. He was of the opinion that different molds of human experience (e.g., religion, art, history, philosophy, and science) coexist and all have claims for validity, and that science could not do without ethics.

During their tracing of the history of discoveries and inventions Needham and his collaborators unearthed an impressive array of early Chinese achievements in technology and science. One example is the mastery of cast iron in China, a millennium and a half before it was known in the West. This Chinese expertise which started in the fourth century BC and which required substantial empirical knowledge of the properties of iron was based on a highly developed bronze casting technology. Cast iron was not only mass-produced for weapons, agricultural tools and cooking utensils, but was also put to such innovative uses as chain suspension bridges (first century AD) and deep-drilling for brine (eleventh century AD) and, later, natural gas. The double-acting piston bellows producing a constant flow of air both during the pushing and pulling of the piston was probably a crucial device enabling the Chinese to achieve superiority in metallurgy for so many hundreds of years. In metallurgical processing also the crank and the piston rod were developed. In agriculture, row cultivation and intensive hoeing were adopted from about the sixth century BC. By the third century BC frame-plows were equipped with cast-iron plowshares and moldboards which substantially reduced friction and allowed the cultivation of heavy and waterlogged soils. These plows were far in advance to the ards used in the Western world.

About the same time, the equine breast-strap and collar harnesses emerged by which the efficiency of animal power was improved. Other agricultural innovations adopted centuries prior to the West included various types of rotary-fan winnowing machines for the efficient separation of grain and chaff as well as multitube seed drills especially in row cultivation. Both these technical devices were known since the second century AD. All these inventions spurred agricultural production and provided the economic basis for social differentiation and state building.

Among the nautical innovations of the Chinese, apart from the well-known compass, the sternpost rudder (first century AD) and watertight compartments in ships (second century AD) should be

mentioned. The Chinese were also great masters in the building of irrigation works, canals, and canal pound-locks. Some of these ancient infrastructure achievements can still be admired today, like the Dujiangyan irrigation system north of Chengdu in Sichuan Province built in the third century BC. Some of these feats of engineering, however, fell into disuse and were forgotten, like the modern-type canal pound-locks with two gates of the tenth century, which were later replaced by the much more ordinary flash-gates with one gate only.

In the military field, Needham noted that gunpowder had arisen from the investigations of Taoist alchemists in about AD 850. About two centuries later it was used for fireworks, bombs, grenades, land mines, and sea mines. In the eleventh and twelfth centuries gunpowder served the production of rockets, and only about a century later it provided the propelling force in guns, cannons, and mortars. In China, however, gunpowder and firearms were not further developed to occupy a position as dominant as that which they did later in European warfare.

Woodblock printing on paper and silk arose in China around the seventh century AD. This allowed the mass production of all kinds of scriptures and texts. However, only few experiments were made with movable-type printing. Paper, another revolutionary product, was invented in China already in about the first century BC. It spread from China to Korea and Japan in the seventh century and to Central Asia, the Arabs and Europe one century later. Another famous Chinese invention that in the eighteenth century was imitated in the West was porcelain.

The decimal system was already in use in archaic China. It also seems that the Chinese were the first to use a blank space for zero and to recognize and use negative numbers as well as decimal fractions. They were also in the possession of a refined value for *pi* (third century AD) and were the first to note the hexagonal structure of snowflakes (second century BC), to mention only the most important inventions and discoveries. Many of the Chinese findings described by Needham and his collaborators in SCC have stood the test of time and thus have to be considered historical facts.

3. *Assessment and Impact*

Needham's work in the framework of SCC was frequently the object of criticism that can be summarized as follows: because Needham took modern science in its final result as a guideline, there is a tendency in his work to highlight those Chinese developments and events best suited to be interpreted as antecedents to modern science. Those currents, however, which did not show universalistic characteristics were neglected by him or even considered obstructive to the development of modern science, without his having undertaken a serious attempt at interpreting them as culturally specific and integral systems of thought and action. It has been pointed out that in his earlier works his linguistic capabilities were not of the highest standard, sometimes reading modern scientific ideas and virtues into traditional Chinese texts. Some reproached him for not having differentiated clearly enough between science and technology, because he sometimes regarded technology as nothing more than applied science. Clear conceptual differentiation was also lacking with regard to such concepts as the Scientific Revolution and the Industrial Revolution, as well as invention and innovation.

Others criticized the encyclopedic structure of SCC. The first volumes, at least, were said to be largely characterized by a timeless, even static, sense of history, to provide only a blurred picture of the world of thought and social interaction in the fields of science and technology, and to be inadequate in pointing out temporal and spatial differences. Discussions arose also with regard to individual findings of Needham. This in particular concerned his history of traditional Chinese astronomical clockworks and their linkwork escapement mechanism that he interpreted as the missing link in the development from clepsydras to mechanical clocks, the production of crystalline synthetic sexual hormones from the urine of pubescent boys by Chinese alchemists, or his identification of traditional Chinese concepts of Nature as precursors of Whitehead's organicist view. Moreover, it was shown that cross- and intracultural transmissions of ideas and devices may have been more complex processes than imagined by Needham.

More fundamental criticism addressed Needham's explanations of the socioeconomic and political reasons why the Scientific Revolution did not take place in China. Nathan Sivin repudiated this as an inadmissible contrafactual question that would diminish in interest as soon as it became clear in which social, economic, political, and cultural environments, and for which purposes traditional Chinese scientists carried out their activities. Benjamin Nelson stated that Needham had underestimated the complexity of social, political, cultural, and intellectual processes leading to the Scientific Revolution in the West. Needham's juxtaposition of Europe's mercantile bourgeoisie and commercial capitalism as conducive to the rise of modern science and China's Confucian 'feudal bureaucracy' as obstructive was considered by many critics as simplistic—as was his characterization of the Taoists as promoters of science and the Confucians as exerting a rather negative influence on it. In conclusion, Needham's work is often considered to be a mixture consisting to a large part of solid findings, plausible, but not always incontestable interpretations, personal speculations, and, sometimes, errors.

In spite of these criticisms, Needham's work undoubtedly embodies a number of tremendous achieve-

ments. First of all, due to his work not only is much more now known about the history of science, technology, and medicine in China, but it also triggered a dramatic increase in the research activities and institutionalization of the academic disciplines concerned. At the same time, a specialization took place, as it is hardly possible anymore for an individual researcher to deal with all the various disciplines to the same depth. Furthermore, as Nelson pointed out, Needham was much more interested in historical processes than Max Weber, one of the great scholars who had attempted to explain the uniqueness of the Western rise of modern science. Needham's work also underlines the importance of intercivilizational encounters and points to the inevitable drive towards universality and thus the overcoming of all local particularities. This is another difference to Weber who was much more interested in a differential typology and the establishment of ideal types, and thus more or less treated each civilization as a discrete unit. These and other insights of Needham therefore provide important elements for a comparative historical and differential sociology of civilizational complexes and intercivilizational encounters (Nelson 1986). And although much criticism was voiced against his notion of the universality of modern science because potentially this might be invoked to justify a blanket modernization, one should not forget that Needham not only ascribed social values to this notion, but also perceived it as a sharp tool of critique directed against Eurocentric internalism.

See also: Biomedical Sciences and Technology: History and Sociology; Chemical Sciences: History and Sociology; China: Sociocultural Aspects; Chinese Law; Chinese Revolutions: Twentieth Century; East Asia: Environmental Issues; East Asian Studies: Culture; East Asian Studies: Economics; East Asian Studies: Gender; East Asian Studies: Geography; Historiography and Historical Thought: Christian Tradition; Historiography and Historical Thought: East Asia; History of Science; History of Science: Constructivist Perspectives; History of Technology; Human Sciences: History and Sociology; Nationalism, Historical Aspects of: The West; Physical Sciences: History and Sociology; Southeast Asian Studies: Health; Southeast Asian Studies: Politics; Southeast Asian Studies: Religion; Urban Planning (Chinese), History of

Bibliography

Blue G 1997 Joseph Needham: A publication history. *Chinese Science* **14**: 90–132

Brook T 1996 The sinology of Joseph Needham. *Modern China* **22**(3): 340–8

Elvin M (ed.) 1980 Symposium: The work of Joseph Needham. *Past and Present* **87**: 17–53

Gazagnadou D 1991 *Joseph Needham, un taoïste d'honneur, autobiographie. De l'embryologie à la civilisation chinoise*. Éditions du Félin-UNESCO, Paris

Habib S I, Dhruv R (eds.) 1999 *Situating the History of Science: Dialogues with Joseph Needham*. Oxford University Press, Delhi, India

Holorenshaw H [pseudonym of Needham] 1973 The making of an honorary Taoist. In: Teich M, Young R (eds.) *Changing Perspectives in the History of Science: Essays in Honour of Joseph Needham*. Heinemann, London

Li G H, Zhang M W, Cao T Q (eds.) 1982 *Explorations in the History of Science and Technology in China: A Special Number of the 'Collections of Essays on Chinese Literature and History,' Compiled in Honour of the Eightieth Birthday of Dr. Joseph Needham*. Shanghai Classics Publishing House, Shanghai, China

Nakayama S, Sivin N (eds.) 1973 *Chinese Science: Explorations of an Ancient Tradition*. MIT, Cambridge, MA

Needham J 1970 *Clerks and Craftsmen in China and the West: Lectures and Addresses on the History of Science and Technology*. Cambridge University Press, Cambridge, UK

Needham J 1979 *The Grand Titration: Science and Society in East and West*, 2nd edn. Allen & Unwin, London

Needham J et al. 1954–2001 *Science and Civilisation in China*. 16 and more vols. Cambridge University Press, Cambridge, UK

Nelson B 1986 *Der Ursprung der Moderne: Vergleichende Studien zum Zivilisationsprozeß*. Suhrkamp, Frankfurt, Germany

Sivin N 1983 Chinesische Wissenschaft: Ein Vergleich der Ansätze von Max Weber und Joseph Needham. In: Schluchter W (ed.) *Max Webers Studie über Konfuzianismus und Taoismus: Interpretation und Kritik*. Suhrkamp, Frankfurt, Germany

Sivin N 1988 Science and medicine in imperial China—the state of the field. *Journal of Asian Studies* **47**(1): 41–90

Vogel H-U 1993 *Naturwissenschaften und Technik im vormodernen China: Historie, Historiographie und Ideologie*. Studienbrief für die FernUniversität—Gesamthochschule Hagen, Fachbereich Erziehungs-, Sozial- und Geisteswissenschaften

H. U. Vogel

Negation, Linguistics of

The term 'negation' refers to a kind of meaning, most intuitively paraphrased in terms of a speaker denying or contradicting something.

(1) The King of France is *not* bald.

If one has any occasion to use sentence (1), one will use it to deny or contradict the view that the King of France is bald. One communicates this negative meaning with a linguistic sign, namely, with the word *not*. This word can also be called a 'negation,' or a 'negative marker' or 'negator.' Obviously, the meaning and the form are related: one could say that the English word *not* is dedicated to the expression of a negative meaning. However, the dedication is not

complete. Not every use of a negative marker actually conveys a negative meaning.

(2) Je crains que Jean *ne* vienne.
I fear that Jean NEG come. SUBJ.PRS.3SG
'I fear that John is coming.'

[Abbreviations: ACC accusative; AFF affirmative; COMP completive; DISTPST distant past; FFNLV fixed form of negative lexical verb; GEN genitive; NEG negation; OBJ object; PFV perfective; PL plural; POT potential; PRS present; PRT particle; PST past; SBJ subject; SG singular; SUBJ subjunctive.]

In French, as in sentence (2), the complement of a verb of fearing may contain the negative element *ne*, but what the speaker fears in (2) is not that John would not be coming, but rather that he *would*. Conversely, not every transmission of a negative meaning needs a unique dedicated negative marker. In (3) the same negative meaning allows two strategies, and one is not more dedicated than the order.

(3) (a) I saw *nothing*.
(b) I did*n't* see *anything*.

In what follows, the focus will first be on meaning, and then on form.

1. Meaning

1.1 Scope and Focus

In Shakespeare's *Julius Caesar*, Brutus has to explain to the citizens of Rome why he killed Caesar. His speech includes the following words:

(4) Not that I loved Caesar less, but that I loved Rome more.

Sentence (4) starts with the negator *not*, followed by *that I loved Caesar less*, which is a complete statement or proposition. This proposition is said to be 'within the scope of negation.' It is a statement that some citizens of Rome might have made in the context of Brutus' murder of Caesar: if one murders someone, one probably does not care for the victim too much. We here hit upon a typical feature of negation: when a proposition is negated, it is usually entertained in the context or even mentioned explicitly.

Sentence (4) illustrates another typical feature: After the negative sentence, i.e., after establishing that lack of love was not that motive for the assassination, Brutus goes on to identify the real motive, the love of Rome. Without the part of the sentence that starts with *but*, the audience would not be satisfied. In general, people are more interested in what is the case than in what is not the case, and in the context at hand, the Romans are more interested in what motivated Brutus to kill Caesar than in what did not motivate him.

Finally, note that though *not* has scope over an entire statement, the negation does not really concern each constituent. The *but that I loved Rome more* part still involves Brutus as a subject and it still involves Brutus' love for someone or something. What is denied is only the degree of Brutus' love for Caesar, compared to his love for Rome. If one wants, one could call the degree of love the '*focus*' of the negation, but note that the focus is not actually reflected in the syntax. In (4) the syntax is vague with respect to the focus of the negation. And even when we integrate the negation into the sentence, as in (5), the focus may still be vague syntactically.

(5) I did not love Caesar less,
(a) … I loved Rome more.
(b) … Cassius did.
(c) … I loved Antony less.

Even though in (5) the negation is associated syntactically with the verb, it does not necessarily negate the verb. It remains a 'sentence negation.' This is different in (6), which is part of what Mark Antony has to say to the Romans:

(6) I come to bury Caesar, *not* to praise him.

In (6) the scope is still a whole sentence: Mark Antony denies the truth of *I came to praise Caesar*. But this time the focus of the negation is much clearer: it is the constituent *to praise Caesar*. This use of the negation can be called 'constituent negation.'

The study of negation and its relation to the truth of the sentence is not only a concern for linguists. It has been on the agenda of logic for as long as the discipline has existed. Within logic, one can approach negation from the perspective of scope and then one is necessarily doing propositional logic, and negation will be a propositional operator. When one looks at the focus of negation, however, one can enter the internal structure of the proposition and can enrich the study of quantifiers (*all, some*) with the study of negative quantifiers (*not all* or *no, not some*). A term for the part of the scopal proposition that is out of the focus of negation is 'presupposition.' The term was introduced into twentieth-century philosophy of language by Strawson (1950). It was a much debated notion in the 1970s, and attempts were made to turn it into a pivotal notion in the study of the interface between semantics and pragmatics (e.g., Levinson 1983).

1.2 Contrariness and Contradiction

(7) (a) Mark Antony was *unhappy*.
(b) Mark Antony was *not happy*.

In both versions of the sentence in (7) arguably we are dealing with constituent negation: in both, the speaker denies the appropriateness of the predicate 'happy.' However, the two versions can differ in meaning. In (7a) Mark Antony is reported as being sad. (7b) is vague: Mark Antony could indeed be unhappy too, but he could also be neither happy nor unhappy. Both 'unhappy' and 'neither happy nor unhappy' are instantiations of 'not happy,' and 'not happy' is called the 'contradictory negation' of 'happy.' On the other hand, 'unhappy' is the 'contrary' negation of 'happy.'

Like the distinction between sentence negation and constituent negation, the one between contradiction and contrariness has existed since the beginnings of

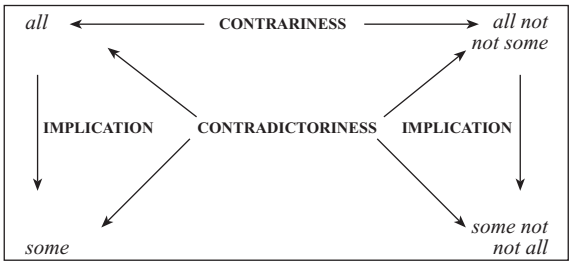

Figure 1
Aristotelian square—Square of Oppositions

logic, most prominently since the formulation of Aristotle's Square of Oppositions.' A version of this square, as applied to quantifiers, is shown in Fig. 1.

Just as 'unhappy' is the contrary of 'happy,' 'all not' is the contrary of 'all,' and just as 'not happy' and 'happy' exhaust all logical possibilities and are thus each other's contradictories, so 'not all' and 'all' are contradictories.

Notions such as contrariness and contradictoriness have been a concern to both logicians and linguists. What interests linguists is that these notions offer a framework for the study of lexicalization patterns. Just as English *unhappy* is a lexicalization of the contrary of *happy*, English *all* has its contrary lexicalized as *no*. The contradictory 'not happy' does not allow for a lexicalization, and neither does 'all not.' What these examples illustrate is that contrariness lexicalizes more easily than contradictoriness, which is not to say that the latter does not lexicalize at all (thus English *non*-, different from both *un*- and *in*-, often produces contradictories, e.g., *nonscientific*). Linguists have also made it clear that contradictory notions often change their meanings into contrariness. A classic case, known since Tobler (1882), is French *il ne faut pas*. It used to mean 'it is not required,' which is the contradictory of 'it is required,' but it now only means 'it is forbidden,' which is the contrary of 'it is required.' The linguist who has done most for making the logical tradition relevant for the linguistics of negation, and for discussing issues of lexicalization and polysemy, is Laurence R. Horn (esp. 1989). Horn has also looked at what has come to be known as 'negative raising' from this perspective. What happens with negative raising is that a negation whose meaning is felt to belong to a complement clause is actually expressed with the complement-taking main verb.

(8) (a) I do*n't* suppose that you believe him.
(b) I suppose that you do*n't* believe him.

Using a transformational perspective, the negation can be seen as being 'raised' from the lower clause (*you believe him*) to the higher clause (*I suppose*). The point that is relevant here is that the two versions of the sentence can be characterized in terms of the distinction between contrariness and contradictoriness. The raised (8a) is literally a contradictory negation (either one supposes something or one does not), but it is normally equivalent to the non-raised (8b), and it then conveys a contrary meaning (instead of supposing *p* and supposing *not p*, one could also suppose neither *p* nor *not p*).

2. Form

The distinction between *unhappy* and *not happy* has a formal dimension, too. In *unhappy* the negation is expressed with morphology, more specifically with a prefix. In *not happy* the negative marker is an independent word, an adverb or particle—one can speak about 'syntactic negation.' Semantics and form are again related. With *unhappy* vs. *not happy*, affixal negation is associated with contrariness, and syntactic negation is associated with contradictoriness. The relation is not very tight, however. The syntactically negated *il ne faut pas que*, after all, expresses a contrary meaning in present-day French. Conversely, in *nonscientific*, the prefixal *non*-strategy results in contradictoriness.

Cross-linguistically, negating with morphology and negating with a separate syntactic element are the two basic strategies. Sound or word order differences may have a function too, but they rarely or never mark negation by themselves. Both morphological and syntactic negation exhibit subtypes. The linguist's increased understanding of the typology of morphological and syntactic negation is related directly to the increased importance of typology in general, in the wake of Greenberg (1966). Important works in the typology of negation are Dahl (1979), Dryer (1988), Forest (1993), Honda (1996), Kahrel (1996), Miestamo (2000), and, somewhat indirectly, Haspelmath (1997).

As to syntactic negation, there are two subtypes: either the negator is an uninflected element (adverb or particle) or it is a verb (and, depending on the criteria used, an auxiliary verb). English *not* is an example of an uninflected element. For an example of a negative verb, one can turn to the Tungusic language, Evenki. Evenki has a negative verb, whose stem is *e*-, which shows agreement with the subject in person and number, and which furthermore inflects for tense, aspect, and mood. It combines with what Nedyalkov (1994) calls the 'fixed form of the negated lexical verb' (FFNLV).

(9) Si min-e *e-che-s* doldi-ra.
You I-ACC NEG-PST-2SG hear-FFNLV
'You did not hear me.' (Nedyalkov, 1994 p. 16.)

Negative verb forms are arguably also the English form *don't*, as in (10), or the classical Latin negative imperative construction with *noli* in (11).

(10) I *don't* like reggae.
(11) *Noli venire*.
'Don't come.'

The Evenki negative verb differs from the English and Latin ones, however, in that a morphological

analysis of the latter will allow one to identify a morphological negator, too. The element *don't* contains the morphological (enclitic) negator *n't*, and *noli* is the imperative of the negative *nolle* 'not want,' which is a contraction of *velle* and a proclitic or prefixal *ne*-. This does not detract from the fact that *don't* and *noli* count as syntactic negators. To negate (12) and (13), the morphological negator cannot attach to the main verb, nor can it be replaced by a negative particle.

(12) (a) I like reggae.
(b) *I liken't reggae.
(c) *I like not reggae.
(d) *I not like reggae.

(13) (a) Veni.
'Come'
(b) *Neveni.
(c) *Non veni.
(d) *Veni non.

It has been argued that if a language has a special negative verb, frequently it will be an existential verb or a verb of wanting, and that these specialized constructions may lose their verbal properties and thus turn into particles (Payne 1985 p. 222, Croft 1991). It has also become clear (esp. through Honda 1996) that when a language employs a special verb to express negation, this special verb does not itself have to be negative. In the Carib language, Hixkaryana, negation can be expressed with an affix on the verb, but this construction further requires a positive copula 'be.'

(14) (a) Koso w-ono-ye
deer 1SBJ/2OBJ-eat-DISTPST
'I ate the deer.'
(b) Koso y-ono-*hra* w-exe-ye.
deer 3 eat-NEG 1-be-DISTPST
'I didn't eat the deer.' (Derbyshire 1979 p. 48.)

A literal paraphrase of (14b) is 'I was not/without eating the deer,' and there is a sense in which the sentence is not really negative.

Note that each of the syntactic negators illustrated thus far precedes the main verb. There seems to be a consensus that languages generally prefer to have their syntactic negation in front of the main verb (Dryer 1988, Dahl 1979). However, a postverbal position is by no means rare, as in Dutch (15), or as when negation is sentence-final, i.e., post-everything, and hence also postverbal, e.g., in the Chadic language, Ngizim, in (16).

(15) Ik hou *niet* van reggae.
I hold NEG of reggae
'I don't like reggae.'

(16) Ii naa duuka *bai*.
I have horse NEG
'I don't have a horse.' (Thompson 1998 p. 328)

Syntactic negation may also involve a part that precedes and a part that follows. This is often called 'discontinuous' or 'embracing' negation. It is illustrated with French (17), be it that the preverbal part *ne* is a clitic is thus intermediate between morphology and syntax.

(17) Je *ne* le lui donne *pas*.
I NEG it him give NEG
'I don't give it to him.'

Discontinuous negation is particularly perspicuous and well-studied in European languages (Bernini and Ramat 1996). The *locus classicus* is Jespersen (1917), and what has come to be known as 'Jespersen's cycle' is the view that languages may go from a preverbal system to a postverbal system via an embracing negation stage, in which the postverbal element functions originally as a reinforcement.

Whether the Negation-before-Verb preference also holds for morphological negation is not clear (Dahl 1979 p. 82 vs. Bybee 1985 p. 177). In any case, there are again three types: (a) a prefixal type, already illustrated with Latin *noli*, (b) an affixal type, as in English *isn't*, and (c) a circumfixal type, as in Cairene Arabic.

(18) *Ma*-katab-m̄-š.
NEG-write. PFV-3PL-NEG
'They didn't write.' (Bernini and Ramat 1996 p. 47.)

A peculiar manifestation of morphological negation is the 'zero morpheme.' In South Dravidian languages, negation can be expressed by the mere absence of a tense suffix. In Havyaka Kannada, spoken in the coastal areas of the Indian state of Karnataka, these negative forms are frequent and have furthermore acquired a modal meaning. The Havyaka verbal paradigm makes a distinction between indicative and subjunctive, and between past and nonpast. In (16a–b) the verbal forms have a tense mood slot filled by the nonpast indicative and subjunctive forms. In (16c) this slot is unfilled, and this in itself marks the structure as negative subjunctive (van der Auwera and Bhat 1999).

19 (a) ma:śTrakko kate o:du-tt-avu
teachers story read-IND.NPST-3PL
'The teachers are reading a story.'
(b) ma:śTrakko kate o:du-g-u
teachersI story read-SUBJ.NPST-3PL
'The teachers may be reading a story.'
(c) ma:śTrakko kate o:d-Ø-avu
teachers story read-SUBJ.NEG-3PL
'The teachers may not be reading a story.'

Zero negatives may be rare, but they are not exclusive to Dravidian. They have also been reported for the Tupí language Karitiána, spoken in the Amazon Basin. According to Landin (1984), affirmative sentences mark verbs with a preverbal particle or clitic expressing affirmation. In negative sentences, however, there is no such prefix or clitic.

(20) (a) Y ta-oty-j ỹn
I AFF-bathe-tense I
'I will bathe.'
(b) Y oty ỹn
I bath I
'I will not bathe.' (Landin 1984 p. 237.)

Note that the verbs in the Karitiána and Havyaka Kannada negative sentences lack tense markers. This touches upon a large area of investigation: the relation between negation and other categories. A forceful illustration of the intricacy of this relation is provided by the Mixtecan language, Copala Trique. The illustration makes the following points: (a) a language may have more than one negative marker, (b) the choice between them may be related to the choice of other categories, in this case a tense–aspect–mood (TAM) category, and (c) the compositionality of joining the negative marker and the TAM marker need not be very transparent. In (21a) the prefix *g-* marks completive TAM. The language has two negative markers, *nẹ*^{3} and *zẹ*^{4}. The construction in (21a) combines only with *zẹ*^{4}, but then the TAM changes: as (21b) shows, the completive becomes what Hollenbach (1976 p. 126) calls 'potential'—the superscripts in the examples mark tone contours.

(21) (a) g-uc̆uh^{34} žini^{3} yuwe^{21} âh
COMP-lay boy palm.mat PRT
'The boy laid the palm mat down.'
(b) zẹ^{4} g-uc̆uh^{34} žini^{3} yuwe^{21} âh
NEG COMP-lay boy palm.mat PRT
'The boy won't lay the palm mat down.'

To reach the meaning 'The boy didn't lay down the palm mat,' one has to start from a potential aspect.

(21) (c) nẹ^{3} g-uc̆uh^{4} žini^{3} yuwe^{21} âh
NEG POT-lay.POT boy palm.mat PRT
'The boy didn't lay the palm mat down.

Negation may also relate to the shape and the meaning of referring expressions. In Russian, for instance, objects that take Accusative cases in positive environments may allow or prefer Genitive case in negative environments.

(22) (a) On soxranil podlinnik pis'm-a
he preserved original.ACC letter-GEN
'He preserved the original of the letter.'
(b) On ne soxranil
he NEG preserved
podlinnik/podlinnik-a pis'ma
original.ACC/original-GEN letter-GEN
'He didn't preserve the original of the letter.' (Timberlake 1993 p. 869.)

Or consider English (23): the word *something* is normally replaced by (c) *did not … anything* or by (d) *nothing*, a strategy that seems typical for Europe (Kahrel 1996, Haspelmath 1997), and even by (e) *didn't … nothing*, which is substandard in English but grammatical in many other languages.

(24) (a) I heard something.
(b) ?I didn't hear something.
(c) I didn't hear anything.
(d) I heard nothing.
(e) I didn't hear nothing.

This leads to a topic that has received a lot of attention from semanticists, namely, 'positive polarity,' the preference words and constructions may exhibit for nonnegative environments (as with *something* in (23a)) and especially 'negative polarity,' the preference for negative (and quasi-negative) constructions (as with *anything*) (van der Wouden 1997).

See also: Linguistics: Overview; Syntax–Semantics Interface

Bibliography

Bernini G, Ramat P 1996 *Negative Sentences in the Languages of Europe*. Mouton de Gruyter, Berlin

Bybee J L 1985 *Morphology. A Study of the Relation Between Meaning and Form*. Benjamins, Amsterdam

Croft W 1991 The evolution of negation. *Journal of Linguistics*. **27**: 1–27

Dahl Ö 1979 Typology of sentence negation. *Linguistics* **17**: 79–106

Derbyshire D 1979 *Hixkaryana: Lingua Descriptive Studies*. North-Holland, Amsterdam

Dryer M S 1988 Universals of negative position. In: Hammond M, Moravcsik E, Wirth J (eds.) *Studies in Syntactic Typology*. Benjamins, Amsterdam

Forest R 1993 *Négations: Essai de syntaxe et de typologie linguistique*. Klincksieck, Paris

Greenberg J H (ed.) 1966 *Universals of Grammar*. MIT Press, Cambridge, MA

Haspelmath M 1997 *Indefinite Pronouns*. Oxford University Press, Oxford, UK

Hollenbach B E 1976 Tense-negation interplay in Copala Trique. *International Journal of American Linguistics* **42**: 126–32

Honda I 1996 *Negation: A Cross-linguistic Study*. PhD diss. University of New York, Buffalo, NY

Horn L R 1989 *A Natural History of Negation*. University of Chicago Press, Chicago

Jespersen O 1917 *Negation in English and Other Languages*. A. F. Host, Copenhagen, Denmark

Kahrel P 1996 *Aspects of Negation*. PhD diss. University of Amsterdam

Landin D J 1984 An outline of the syntactic structure of Karitiâna sentences. In: Dooley R A (ed.) *Estudos Sobre Linguas Tupi do Brasil*. Summer Institute of Linguistics, Brasilia, Brazil

Levinson S C 1983 *Pragmatics*. Cambridge University Press, Cambridge, UK

Miestamo M 2000 Toward a typology of standard negation. *Nordic Journal of Linguistics* **23**: 65–88

Nedyalkov I 1994 Evenki. In: Kahrel P, Van den Berg R (eds.) *Typological Studies in Negation*. Benjamins, Amsterdam

Payne J R 1985 Negation. In: Shopen T (ed.) *Language Typology and Syntactic Description. Vol. 1. Clause Structure*. Cambridge University Press, Cambridge, UK

Strawson P F 1950 On referring. *Mind* **69**: 320–44

Thompson S A 1998 A discourse explanation for the cross-linguistic differences in the grammar of interrogation and negation. In: Siewierska A, Song J J (eds.) *Case, Typology and Grammar. In Honor of Barry J. Blake*. Benjamins, Amsterdam

Timberlake A 1993 Russian. In: Comrie B, Corbett G C (eds.) *The Slavonic Languages*. Routledge, London
Tobler A 1882 *Vermische Beiträge zur französischen Grammatik. Erste Reihe*. Hirzel, Leipzig, Germany
van der Auwera J, Bhat D N S 1999 Havyaka Kannada: Negation and modality. *Indian Linguistics* **60**: 1–26
van der Wouden T 1997 *Negative Contexts: Collocation, Polarity and Multiple Negation*. Routledge, London

J. van der Auwera

Neglect, Neural Basis of

1. The Term 'Hemispatial Neglect'

One of the most dramatic deficits in spatial perception and action is hemispatial neglect (neglect for short), a neurobehavioral disorder in which a patient is unable to acknowledge, report, or otherwise make explicit use of information falling in the visual hemispace contralateral to the responsible lesion (Bartolomeo and Chokron 2001, Bisiach and Vallar 2001, Driver 1998, Vallar 1998). The most frequently affected neural region associated with neglect is the inferior parietal lobule although other cortical and subcortical sites have also been implicated (Karnath et al. 2001). Neglect occurs more commonly and with greater severity after right- than left-hemisphere lesions in humans and the differential involvement of the right hemisphere in processes of attention and spatial representation is consistent with recent neuroimaging studies (Corbetta et al. 2000). Because of this asymmetry, neglect will be referred to as 'left-sided' throughout this section.

1.1 The Phenomenon of Hemispatial Neglect

Neglect can be severe and debilitating in everyday life; individuals may fail to notice objects on the left of a scene, may ignore words on the left of a page or food on the left of a plate, and typically, may omit to copy features on the left of a figure while preserving the corresponding features on the right, as shown in Fig. 1 Neglect is also standardly elicited in tasks that require the patient to cross out items that appear randomly scattered on a piece of paper or to mark the midpoint of a single horizontal line.

Neglect is not restricted to the visual domain; patients may not orient to contralateral auditory information, may not detect a contralateral somatosensory stimulus and may even ignore contralateral odors. Neglect is also not restricted to sensory input; patients with neglect may plan and even execute movements more poorly to the contralateral than ipsilateral side and this holds both for upper limb movements (Heilman et al. 1987, Mattingley and Driver 1997, Mattingley et al. 1998) as well as for eye movements (Gainotti 1993). The relative paucity of eye movements to the left is illustrated in Fig. 2 in which we see, relative to control subjects, both fewer and more brief eye fixations contralesionally as well as a simultaneous increase in the number and duration of right-sided fixations.

The failure to process information on the contralateral side cannot be attributed to a primary motor impairment nor to a hemianopia; even when infor-

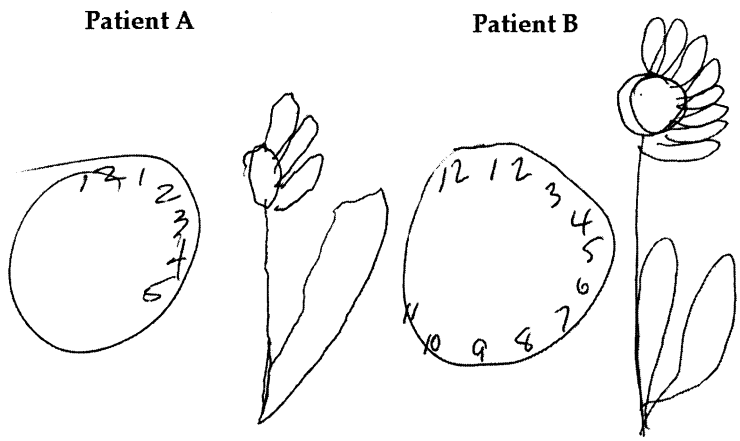

Figure 1
Copy of a daisy and a clock by two patients with left-sided neglect following a lesion to the middle cerebral artery of the right hemisphere

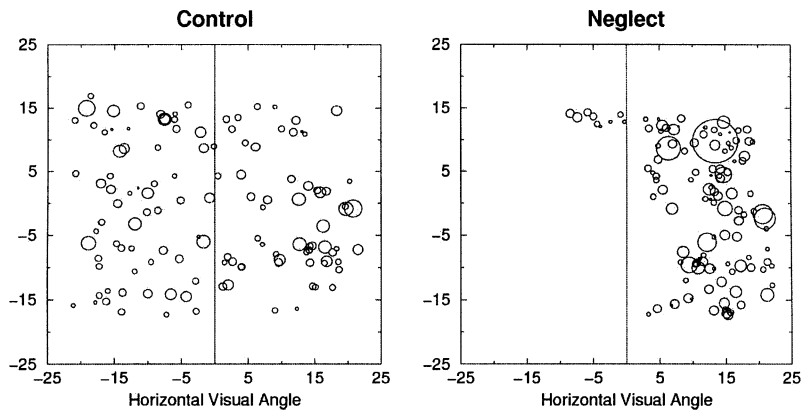

Figure 2
The location (in x and y coordinates) and duration (reflected as size of the circle) of fixations for a normal control subject and a typical patient with hemispatial neglect during search for the letter 'A' in a random array of letters covering approximately 45° of horizontal visual angle. Subjects wore a scleral contact annulus in one eye while they viewed the target display and eye position was sampled every 5 ms. Whereas the normal subject searches roughly equally across the entire space, the patient makes few eye movements into the contralateral space and those that are made are extremely brief in duration whereas ipsilesional fixations are more numerous and longer in duration (Behrmann et al. 1997)

mation is presented to the intact visual field of a neglect patient, information in the contralateral hemispace is still neglected (Làdavas et al. 1990). Patients may even show neglect in the absence of visual input; for example, patients ignore the left side of a scene constructed in visual mental imagery (Bisiach and Luzzatti 1978) and fail to search the left of a dark room (Hornak 1992). Finally, patients can be cued to process information on the contralesional side (Milner et al. 1993, Riddoch and Humphreys 1983) through verbal or other visual instructions. Taken together, these findings all rule out a primary sensory deficit as being causally linked to neglect although a field defect can exacerbate any tendency towards neglect (Doricchi and Angelelli 1999).

Neglect occurs most frequently along the horizontal plane although 'altitudinal neglect,' in which information appearing in the upper rather than lower portion of an array (or vice versa) is ignored, may also occur. Neglect may also occur along the radial dimension although this is thought to be tied to the need for a motor response and is less commonly seen when perception alone is tested.

1.2 Related Disorders

There are a number of deficits related to neglect. 'Extinction to double simultaneous stimulation' refers to a phenomenon in which patients fail to report the stimulus on the contralesional side when two stimuli are presented simultaneously to the ipsilateral and contralateral side of space. When only a single stimulus is presented, even to the contralateral side, it can be detected and reported. Although some consider extinction to be a mild form of neglect, their exact relationship remains controversial. A number of recent studies have suggested that extinction arises from a competition for attention and activation between two stimuli (di Pellegrino et al. 1997); unilateral brain damage represents a reduction in the competitive weights on the affected side and, as a consequence, stimuli appearing on that side are only weakly activated and hence dominated by stronger ipsilateral stimuli. Extinction can be observed both when the two stimuli are from the same sensory modality (e.g., visual) or even when they are from two different modalities (e.g., visual and tactile) (also see Làdavas et al. (1998), and for related work in monkeys, Graziano and Gross (1998)).

A second, related phenomenon is 'allesthesia,' in which a patient is able to detect the contralesional stimulus but refers its position to the intact side. Asked where the stimulus had appeared, the patient usually locates the detected stimulus on the ipsilesional side. Finally, 'anosagnosia,' a disorder in which patients are unaware of their limited representation of the environment and consequently minimize or deny that they have any impairment, may co-occur with neglect and may adversely affect attempts at rehabilitation.

2. Underlying Mechanisms

A number of general interpretations have been advanced to account for the varied manifestations of neglect. For example, some have suggested that neglect arises from a deficit in orienting spatially to the

contralesional side with a gradient of attention biasing attention to the ipsilateral side. Others have argued that the problem is one in which there is a rotation or translation of the egocentric frame of reference towards the ipsilesional side. Yet a third suggestion, the premotor theory of neglect, is that there is an imbalance between the spatial representations that control motor programs, and this imbalance gives rise to an attentional deficit. Although described as alternatives, these differing interpretations may not be mutually exclusive and may refer to different aspects of a single deficit. What is clear at present is that, although they may diverge in detail, these explanations share a common perspective and that is that neglect is a higher-order spatial impairment.

One view of neglect, formalized most recently by Pouget and Driver 2000) is that neglect reflects a partial loss or dysfunction of cells in the parietal cortex and other related neural areas. Because there is a certain number of neurons for representing information in different lateral positions, damage to these neurons gives rise to a pathological gradient with a smooth decline for increasingly leftward locations and an off-center peak reflecting the peak activation on the ipsilateral side. A gradient has been incorporated into several neural network models that simulate neglect and these models have successfully characterized many aspects of neglect including extinction, neglect on line-bisection and line cancellation tasks, and neglect in reading (Mozer and Behrmann 1990, Mozer et al. 1997, Pouget and Sejnowski 1999).

3. *Spatial Reference Frames*

One question that has been addressed in the context of neglect concerns the nature of the spatial representations which, when damaged, give rise to the disorder. Specifically, the approach adopted to address this issue has involved determining what constitutes 'left' and 'right.' Thus, when the patients ignore information on the left, what is it left of? Because 'left' and 'right' cannot be defined in absolute terms, they must always be qualified with respect to a set of coordinates or spatial reference frame, i.e., relative to an origin and axes. It is these spatial coordinates that are used for organizing spatial information, and for establishing spatial relationships between objects in the scene and between the observer and the objects. By examining what reference frame(s) determines the midline such that information to the left of it is neglected, we may obtain an understanding of how spatial position is coded in the parietal cortex and associated regions.

A major distinction between the types of spatial reference frames has been between those that define spatial position egocentrically or with respect to the viewer and those that define position allocentrically, irrespective of the position of the viewer. At least four different egocentric frames have been identified, with the midline centered on: (a) the eyes or vertical meridian of the visual field, (b) the head, (c) the trunk, or (d) the longitudinal axis of the limb that is involved in executing an action, such as the arm. Other studies have shown that patients neglect information on the left, defined allocentrically, with respect to the gravitational upright or environment (or scene) or with respect to the midline of individual objects or perceptual units in the scene.

Several important insights concerning the spatial coordinates used for coding left and right have been obtained recently. Among these is the finding that spatial position may be coded simultaneously with respect to multiple frames of reference, both egocentric and allocentric. As depicted in Fig. 3, a patient, seated in front of a large screen, reports the identity of letters from a display. In the first condition (panel A), the patient's head, eyes, and trunk are all aligned with the midline of the scene and the patient reports fewer letters from the left than right. Because, under this condition, the midlines of the different reference frames are all aligned, one cannot distinguish the relative contribution of these various egocentric and allocentric reference frames to the observed neglect. To deconfound this, in a second condition, the viewer's midline is rotated out of alignment from the midline of the scene and now the upper and lower portions of the display coincide with the viewer's left and right, respectively. Interestingly, in this second condition (panel B), the patient continues to report less information from the left than right of the midline of the allocentric scene-based frame but also reports fewer letters from the top than bottom of the display, implicating a viewer-based reference frame. Indeed, in that quadrant (upper left) in which both allocentric and egocentric neglect co-occur, there is the greatest neglect with the relationship between these two reference frames apparently being additive.

Several recent studies have also documented the presence of neglect defined with respect to the midline of individual objects in the scene. In one such study, neglect patients saw computer displays consisting of an object (barbell) in which the two circles were differently colored (Fig. 4, panel (A)) (Behrmann and Tipper 1994). The barbell appeared in the center of the computer screen (or entirely on the right for patients with a left field defect) and then, after a short delay, on two-thirds of the trials, a target or probe in the form of a small white dot appeared in the center of either the left or the right circle. The patients pressed a key as soon as they detected the presence of this target. As expected, in this 'static' condition, patients were poorer at detecting the left than the right target. Because the left and right of the barbell object are aligned with reference frames defined by the viewer and scene, a decoupling condition was introduced. In this second 'rotating' or moving condition, the midline of the barbell was dissociated from various other reference frames in the following way: first, the barbell

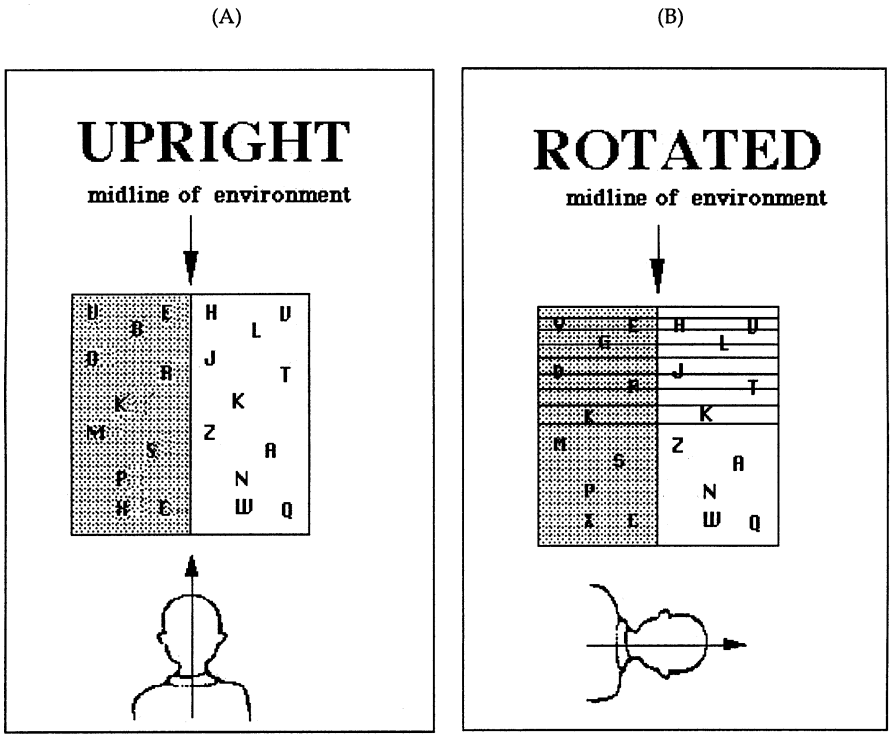

Figure 3
Left-sided neglect in environment- and viewer-based frames. Neglect patients report fewer letters from the shaded region in panel (A). They report fewer letters from the shaded region in panel (B) (environment frame) but also fewer letters from the striped region i.e., the left of the viewer's midline (viewer-frame)

was presented on the computer screen and then, after a short delay, it was rotated 180° around the midpoint of the connecting horizontal rod while the patient watched. When the rotation was complete and the barbell had reached its stationary position, a short delay ensued and then the target appeared on the left or right. Note that, as shown in Fig. 4, the display in which the target finally appears is identical in the static and rotating condition. However, in the rotation condition, the original left circle of the barbell is now on the right side of the computer screen and the original right circle is now of the left side of the screen. This arrangement allows us to determine whether the patients now neglect targets appearing on the 'good' right side of space (but on the left of the object) and show good detection of targets appearing on the 'poor' left side of space (but on the right of the object).

In contrast with the static condition, in the rotating condition, patients are faster at detecting the left- than right-sided target (panel B). The explanation for this paradoxical reversal of neglect is that patients neglect the information that falls to the left of the object's midline and, conversely, attend well to the information that falls to the object's right. Even when the object moves into a new spatial location or undergoes a complete rotation, information on its left is still at a disadvantage compared with information on its right. This pattern cannot be attributed to an artifact of eye movement; even when eye movement is controlled, the same pattern is observed (Tipper and Behrmann 1996). It also is not obviously attributable to the covert tracking of the left or right of the object. In fact, relative to the static condition, the facilitation for targets on the left and inhibition of targets on the right in the rotation condition critically depend on the two circles forming a unified object and are not evident when the two circles are not joined. If the result were simply a function of the patients' covertly tracking say the right circle of the barbell (to which their attention is drawn initially), a similar finding should be obtained even when the two circles are not connected. These findings lead to the conclusion that spatial positions are also defined with respect to an object-based reference frame and information to the left of this frame may be neglected. These findings are compatible with results from a single neuron recording study in monkeys which shows directional selectivity for neurons in the parietal cortex in a task in which saccades are made to the relative left or right of a stimulus bar (Olson and Gettner 1998).

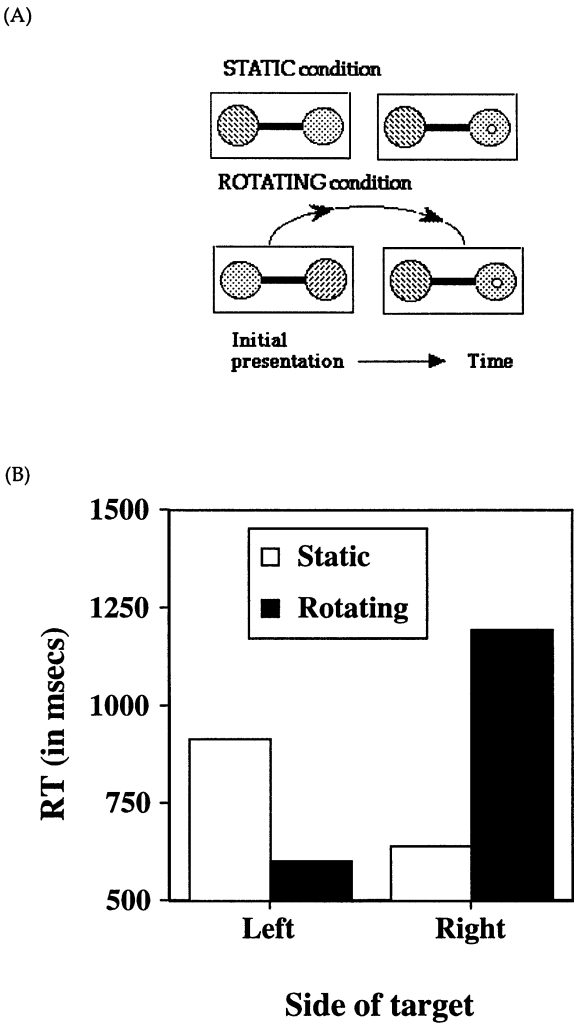

Figure 4
Depiction of experimental displays for static and rotating conditions (panel A) and results reflecting object-based neglect (panel B); adapted from Behrmann and Tipper (1994)

Interestingly, this object-based neglect can occur simultaneously with neglect defined with respect to the environment or to the viewer (Behrmann and Tipper 1999). In an experiment using similar displays but with two grey squares appearing simultaneously with the barbell, one on the left and one on the right, and which remain stationary throughout all conditions, object-based neglect is observed (as above, see Fig. 4). More importantly, detection of the probe in the left square is poorer than in the right square, revealing the presence of neglect for information on the left not only in an object-defined frame but also (and concurrently) in a frame defined by the midline of the viewer or scene. Finally, the extent of neglect in one of these reference frames can be exacerbated by presenting the probe more frequently within this one frame and less often in other reference frames. Taken together, these results suggest two important conclusions: first, patients can neglect the left of the environment and/or viewer and, at the very same time, can neglect the left of an object, even when the left of the object is on the right side of space (as demonstrated by the rotating barbell). Second, the distribution of visuospatial attention in these two frames of reference is flexible, and may be influenced by task demands or contingencies. These findings suggest that the parietal lobe can code spatial information relative to multiple spatial reference frames simultaneously and flexibly.

Studies concerned with reference frames have also been conducted using single unit recording techniques with awake, behaving monkeys either with or without lesions. Although there does not appear to be a direct analog of hemispatial neglect in monkeys, and anatomical and cytoarchitectonic homologies between the posterior parietal cortex in monkeys and humans are not perfect (Crowne and Mah 1998), several experimenters have described the occurrence of visual extinction after lesions to the posterior parietal areas (Gaffan and Hornak 1997, Lynch and McLaren 1989, Rizzolatti et al. 1985). Converging evidence is also obtained from recent neurophysiological studies with nonlesioned monkeys in which the spatial coordinates mediated by the parietal cortex are explored (for recent reviews see Andersen et al. 1997, Colby and Goldberg 1999).

4. *Implicit Visual Information Processing*

One of the most provocative claims in the research on neglect is that, despite the loss of phenomenal awareness of the contralateral information, these patients appear to have processed a considerable amount of contralateral information implicitly and this information can be used to guide their behavior (Driver and Vuilleumier 2001). An early observation by Kinsbourne and Warrington (1962) suggested that these patients 'knew' more than their overt responses suggested. Thus, in their reading errors, patients with neglect dyslexia often retained word length, for example, producing MILKMAN for MAILMAN. Since that time, several innovative paradigms have been developed to explore this issue. It has been shown, for example, that patients can make same/different judgments about two pictures, one presented to each visual field, significantly above chance even though they are unable to identify the left-sided pictures, can make meaningful and consistent preference ratings on aspects on the left of stimuli even though they fail to report the presence of the information on which their apparent preference ratings were made, and can show

facilitation in a priming task when the prime appears on the left and the probe centrally (McGlinchey-Berroth 1997).

5. Perspectives on Rehabilitation

A diverse set of strategies have been adopted in studies which attempt to rehabilitate patients with neglect. These include training visual information and visual search, tactile exploration, actively or passively moving the contralesional arm and imagining mental scenes. Caloric vestibular stimulation in which iced water is introduced to the left ear and stimulates the vestibular apparatus has been shown to lead to a dramatic, albeit, short-term improvement in the disorder. Vibration of the left neck muscles (which produces an afferent that is similar to twisting the trunk midline leftward) has also been shown to be successful. Finally, a recent study, in which patients wore optical prisms that shifted the visual scene to the right, showed that patients were increasingly able to reach towards the contralesional side. The neural basis of the improvements remain poorly understood and may involve plasticity in some brain areas although the parietal lobe is notoriously nonplastic (Kolb and Cioe 1998). It is also possible, however, that whatever change is observed may arise from involvement of other postural alterations (such as the shift of the viewer-midline more leftward).

6. Concluding Remarks and Future Directions

As is clear from this review, hemispatial neglect is a heterogeneous and complicated disorder that defies simple explanation. For example, there remains a pressing need to understand what gives rise to this disorder, and what frames of reference are used to code spatial position. Additionally, we need to know how structured and unified behavior emerges if position is defined with respect to multiple coordinates. One possible explanation is that the contribution of each set of coordinates can be differentially weighted according to the task at hand. Consistent with this task dependency is the fact that the activity in neurons in the posterior parietal cortex differs depending on the task the animal is intending to perform (eye movements vs. reaching; Snyder et al. 1997). An additional pressing issue concerns the fate of the neglected information; whether this information is processed normally and in a rich and detailed fashion but is somehow not available for conscious or explicit report remains debatable and highly controversial. Promising studies using functional magnetic resonance imaging have shown differences in brain activation in the striate and extrastriate regions of the contralateral hemisphere for stimuli that have been extinguished compared with those that are reported (Driver and Vuilleumier 2001, Marzi et al. 2000, Rees et al. 2000, Vuilleumier et al. 2001).

A final issue to resolve concerns the role of the parietal/temporal cortex in relation to other cortical regions that are involved in spatial processing. The parietal/temporal cortex does not have a monopoly on spatial processing and regions, such as the frontal cortex and parahippocampal regions, also contribute importantly to spatial representations. How these different areas work together to give rise to our unitary visual experience remains unclear but neuropsychologic studies with patients suffering from hemispatial neglect allow us to observe some of the representations and processes that ultimately contribute to our unified spatial experience. These findings together with those obtained from functional imaging studies with normal subjects, single-unit recordings from nonhuman primates and computational analyses continue to provide new insights into this perplexing disorder.

See also: Brain Damage: Neuropsychological Rehabilitation; Parietal Lobe; Spatial Cognition

Bibliography

Andersen R A, Snyder L H, Bradley D C, Xing J 1997 Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* **20**: 303–30

Bartolomeo P, Chokron S 2001 Levels of impairment in unilateral neglect. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*. Elsevier Science, Amsterdam, Vol. 4, pp. 67–98

Behrmann M, Barton J J S, Watt S, Black S E 1997 Impaired visual search in patients with unilateral neglect: An oculographic analysis. *Neuropsychologia* **35**: 1445–58

Behrmann M, Tipper S P 1994 Object-based attentional mechanisms: Evidence from patients with unilateral neglect. In: Umilta C, Moscovitch M (eds.) *Attention and Performance XV: Conscious and Nonconscious Information Processing*. MIT Press, Cambridge, MA, pp. 351–75

Behrmann M, Tipper S P 1999 Attention accesses multiple reference frames: Evidence from neglect. *Journal of Experimental Psychology: Human Perception and Performance* **25**: 83–101

Bisiach E, Luzzatti C 1978 Unilateral neglect of representational space. *Cortex* **14**: 129–33

Bisiach E, Vallar G 2001 Unilateral neglect in humans. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*. Elsevier Science, Amsterdam, pp. 459–502

Colby C L, Goldberg M E 1999 Space and attention in parietal cortex. *Annual Review of Neuroscience* **22**: 319–49

Corbetta M, Kincade J M, Ollinger J M, McAvoy M P, Shulman G L 2000 Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience* **3**: 1–6

Crowne D P, Mah L W 1998 A comparison of hemispatial neglect from posterior parietal and periarcuate lesions in the monkey. *Psychobiology* **26**: 103–8

di Pellegrino G, Basso G, Frassinetti F 1997 Spatial extinction on double asynchronous stimulation. *Neuropsychologia* **35**: 1215–23

Doricchi F, Angelelli P 1999 Misrepresentation of horizontal space in left unilateral neglect: Role of hemianopia. *Neurology* **52**: 1845–52
Driver J 1998 The neuropsychology of spatial attention. In: Pashler H (ed.) *Attention*. Psychology Press, Hove, UK, pp. 297–340
Driver J, Vuilleumier P 2001 Perceptual awareness and its loss in unilateral extinction. *Cognition* **79**: 39–88
Gaffan D, Hornak J 1997 Visual neglect in the monkey. *Brain* **120**: 1647–57
Gainotti G 1993 The role of spontaneous eye movements in orienting attention and in unilateral neglect. In: Robertson I, Marshall J C (eds.) *Hemispatial neglect*. Lawrence Erlbaum, London, pp. 107–22
Graziano M S, Gross C G 1998 Spatial maps for the control of movement. *Current Opinion in Neurobiology* **8**: 195–201
Heilman K M, Bowers D, Valenstein E, Watson R T 1987 Hemispace and hemispatial neglect. In: Jeannerod M (ed.) *Neurophysiological and Neuropsychological Aspects of Spatial Neglect*. Elsevier Science, Amsterdam, pp. 115–50
Hornak J 1992 Ocular exploration in the dark by patients with visual neglect. *Neuropsychologia* **30**: 547–52
Karnath H O, Ferber S, Himmelbach M 2001 Spatial awareness is a function of the temporal not the posterior parietal lobe. *Nature* **411**: 950–3
Kinsbourne M, Warrington E K 1962 A variety of reading disability associated with right hemisphere lesions. *Journal of Neurology, Neurosurgery and Psychiatry* **25**: 339–44
Kolb B, Cioe J 1998 Absence of recovery or dendritic organization after neonatal posterior parietal lesions. *Psychobiology* **26**: 134–42
Làdavas E, di Pellegrino G, Farnè A, Zeloni G 1998 Neuropsychological evidence of an integrated visuotactile representation of peripersonal space in humans. *Journal of Cognitive Neurosciences* **10**: 581–9
Làdavas E, Petronio A, Umilta C 1990 The deployment of visual attention in the intact field of hemineglect patients. *Cortex* **26**: 307–17
Lynch J C, McLaren J W 1989 Deficits of visual attention and saccadic eye movements after lesions of parietooccipital cortex in monkeys. *Journal of Neurophysiology* **61**: 74
Marzi M, Girelli M, Mimussi C, Smania N, Maravita A 2000 Electrophysiological correlates of conscious vision: Evidence from unilateral extinction. *Journal of Cognitive Neuroscience* **12**: 869–77
Mattingley J B, Driver J 1997 Distinguishing sensory and motor deficits after parietal damage: An evaluation of response selection biases in unilateral neglect. In: Thier P, Karnath H O (eds.) *Parietal Contributions to Orientation in 3D Space*. Springer, Heidelberg, Germany, pp. 309–38
Mattingley J B, Husain M, Rorden C, Kennard C, Driver J 1998 Motor role of human inferior parietal lobe revealed in unilateral neglect patients. *Nature* **392**: 179–82
McGlinchey-Berroth R 1997 Visual information processing in hemispatial neglect. *Trends in Cognitive Sciences* **1**: 91–7
Milner A D, Harvey M, Roberts R C, Forster S V 1993 Line bisection errors in visual neglect: Misguided action or size distortion? *Neuropsychologia* **31**: 39–49
Mozer M C, Behrmann M 1990 On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience* **2**: 96–123
Mozer M C, Halligan P W, Marshall J C 1997 The end of the line for a brain-damaged model of hemispatial neglect. *Journal of Cognitive Neuroscience* **9**: 171–90
Olson C R, Gettner S N 1998 Impairment of object-centered vision following lesions of macaque posterior parietal cortex. *Society of Neuroscience Abstracts* **24**: 1141
Pouget A, Driver J 2000 Relating unilateral neglect to the neural coding of space. *Current Opinion in Neurobiology* **10**: 242–9
Pouget A, Sejnowski T 1999 A new view of hemineglect based on the response properties of parietal neurones. In: Burgess N, Jeffery K J, O'Keefe J (eds.) *Spatial Functions of the Hippocampal Formation and Parietal Cortex*. Oxford University Press, Oxford, UK, pp. 127–46
Rees G, Wojciulik E, Clarkje K, Husain M, Frith C, Driver J 2000 Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. *Brain* **123**: 1624–33
Riddoch M J, Humphreys G W 1983 The effect of cueing on unilateral neglect. *Neuropsychologia* **21**: 589–99
Rizzolatti G, Gentilucci M, Matelli M 1985 Selective spatial attention: One center, one circuit, or many circuits. In: Posner M I, Marin O S M (eds.) *Attention and Performance XI*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 251–65
Snyder L H, Batista A P, Andersen R A 1997 Coding of intention in the posterior parietal cortex. *Nature* **386**: 167–70
Tipper S P, Behrmann M 1996 Object-centred not scene-based visual neglect. *Journal of Experimental Psychology: Human Perception and Performance* **22**: 1261–78
Vallar G 1998 Spatial hemineglect in humans. *Trends in Cognitive Sciences* **2**: 87–96
Vuilleumier P, Sagiv N, Hazeltine E, Poldrack R A, Swick D, Ragal R D, Gabrieli J D E 2001 Neural fate of seen and unseen faces in visuo spatial neglect: A combined event-related functional MRI and event-related potential study. *Proceedings of the National Academy of Sciences* **98**: 3495–3500

M. Behrmann

Negotiation and Bargaining: Organizational Aspects

Negotiation and bargaining is a process in which two or more parties seek a mutual agreement through an explicit or implicit exchange of views. The focus of this article is on international negotiation and bargaining with a special emphasis on the role of international institutions in facilitating cooperative bargains. In particular, the article discusses institutional design in connection with the nature of the problem that parties face, viewed through the lens of the provision of different types of goods.

1. *The Diversity of Bargaining Situations*

Reaching a mutual agreement may take very different roads depending on the characteristics of the bargaining situation. This section discusses the key distinction

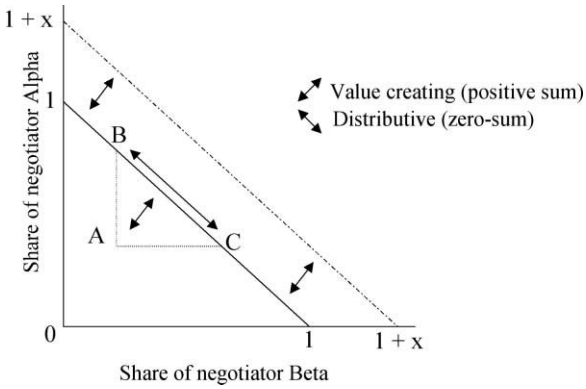

Figure 1
A mixed-motive bargaining situation

between positive vs. zero-sum situations. It then relates this basic difference to a typology of goods that parties may seek to provide through a process of negotiation.

1.1 Pure Conflicts of Interest: Zero-sum Situations

A widespread assumption is that bargaining is, foremost, a process in which parties seek to reconcile contradictory interests and values. The aim is to agree on a particular split of the differences between parties. Such situations have been labeled 'zero-sum situations' and are particularly salient when negotiators value outcomes in relative terms. In such cases, any gain by one party comes at the expense of a loss by another party. Bargaining is essentially a process of (re)distribution. In international politics, an archetype of a zero-sum situation is the division of a given territory (for instance the carving-up of the Antarctic into different national sectors). In economics, the process of allocating market shares under nongrowing market assumptions is also a zero-sum situation.

1.2 Mixed-motive Situations

Whereas distribution is often a key concern of bargaining, parties may also share common interests. Agreements can create value, as well as distribute it. Negotiation situations with both distributive and creative dimensions are labeled 'mixed-motive situations.' Figure 1 illustrates the two dimensions for a hypothetical bargaining case between negotiators Alpha and Beta aiming at establishing market shares.

Moving along the line [1,1] corresponds to the case of pure conflict of interests mentioned in Sect. 1.1. Negotiator Alpha and Beta fully control a steady market and have to mutually agree on a distribution of market shares. Consider, however, the case where Alpha and Beta are not the only market actors and thus will not be able to fully control the market unless they come to a mutual agreement. In Fig. 1, point A is the outcome in case of no agreement, with Alpha and Beta joint market shares lower than one. Bargaining aims at reaching total control of the market, thus creating value collectively. The collective gain must be redistributed, along segment [B, C] in Fig. 1. Another mixed situation arises when negotiators seek to expand the total size of the market, from 1 to (1+x) in Fig. 1. This does not eliminate the initial distributive aspect of bargaining but makes the potential conflict less severe.

In sum, in mixed-motive situations there may be a trade-off between individual and collective rationality. Mixed-motive situations dominate when negotiators not only care about bargaining outcomes in relative terms but also in absolute terms.

Table 1
Type of goods

		Jointness? YES	Jointness? NO
Exclusion possible?	NO	Public	Common pool resources
	YES	Inclusive club	Private

1.3 Bargaining Situations and Types of Goods

Bargaining situations may be characterized by different combinations of positive and zero-sum features. In particular, it is useful to distinguish four types of objects or goods that negotiators may deal with, namely public goods, common pool resources (CPRs), inclusive club goods, or private goods (see Aggarwal 1998, Cornes and Sandler 1996, Snidal 1979).

Differences among goods can be characterized along two dimensions: jointness, which refers to the extent to which goods are affected by consumption; and by the possibility of exclusion, which refers to whether noncontributors to the provision of the good can be kept from consuming it (see Table 1).

The two dimensions—jointness and exclusion—largely influence the mixture of interests in a bargaining process. In the case of public goods, there is an important value-creating potential—given that actors can consume the same amount of the good without affecting the consumption of others (as in the case of national defense). Yet, the possibility of cost sharing creates a distributive aspect to this good. Given the impossibility of excluding noncontributors, there is in fact a dual distributive problem—among those who pay for the good and between those who pay and those who do not.

Inclusive club goods refer to the case of goods that exhibit jointness (not diminished, by use), but where exclusion is possible (for instance satellite transmission of television). In this situation, bargaining reveals a strong value-creating feature and the problem of distributing costs is less problematic than in the case of public goods because of the possibility of excluding noncontributors from the consumption of the good. The availability of exclusionary mechanisms may help generate revenue, which further reduces the initial problem of cost distribution.

The positive potential of cooperation to collectively produce private goods tends to be complicated by the distributive problem created by the lack of jointness (for instance, the provision of a system of irrigation that delivers private amounts of water to different field owners). Yet, the availability of exclusionary mechanisms can alleviate the problem of distributing the costs. By contrast, in the case of common pool resource goods, bargaining is essentially a distributive process among those that are willing to pay for the good and between them and the non-contributors.

2. *Bargaining Situations and the Role of Institutions*

When and how do institutions affect patterns of negotiations? We begin with a definition of institutions, and then turn to an examination of the role they might play. The most readily evident role for institutions could be for situations of market failure, described nicely by the Prisoners' Dilemma. In such cases, institutions may help to overcome suboptimality and ensure that actors reach a mutually superior outcome by cooperating. But as noted in Sect. 2.2, other bargaining situations might also benefit from institutions. Section 2.3 briefly examines some theories that help explain the bargaining process leading to the formation of institutions themselves.

2.1 *Definitions: Meta-regimes, Regimes, and International Institutions*

International institutions have been defined in many ways, and considerable disagreement remains over the differences among international organizations, international institutions, and international regimes. International organizations are generally seen to be more formally specified institutions, with a secretariat, permanent office, and the like. International regimes have been defined as sets of principle, norms, rules, and decisions upon which actors' expectations converge (Krasner 1983). It is useful to distinguish between principles and norms on the one hand, and rules and procedures on the other. This allows us to delineate two aspects of institutions: meta-regimes and regimes (Aggarwal 1998). Whereas meta-regimes represent the principles and norms underlying international arrangements, international regimes refer specifically to rules and procedures. Regimes can be examined in terms of their characteristics: their strength, nature, and scope. Strength refers to the stringency of the multilateral rules that regulate national behavior; nature (in an economic context) refers to the degree of openness promoted by the accord; and scope refers to two aspects: (a) the number of issues incorporated in the regime, or issue scope; and (b) the number of actors involved (bilateral or multilateral), or institutional scope. To summarize, international institutions consist of meta-regimes and regimes; if they are highly formalized, we can refer to them as international organizations.

2.2 *Institutional Designs for Different Goods*

Institutions affect the strategies available to actors and the payoffs linked to various bargaining outcomes. Systematic and specific links between bargaining situations and institutions have been explored through strategic analysis of actors' behavior (Aggarwal and Dupont 1999; for earlier efforts, see Snidal 1985, Stein 1982). Aggarwal and Dupont (1999) model negotiation processes over the provision of diverse types of goods with game-theoretic structures that correspond to different assumptions about the costs and benefits of these goods and variation in actors' capabilities. They derive a variety of games (including the situations known as Prisoners' Dilemma, Chicken, Assurance, Harmony, Deadlock, or Battle of the Sexes; for a taxonomy of situations see Rapoport and Guyer 1966) that can be solved using the Nash equilibrium concept (Nash 1951). Issues of institutional design are then discussed in connection with the evaluation in terms of efficiency of the predicted outcome(s) of the various games. This analysis yields the following findings on the role of institutions in bargaining situations.

First, institutions may help secure positive outcomes in bargaining processes. When actors have large resources and when they face public and CPR goods, conflict may prevent the provision of such goods. Institutions can play the role of a third party in such cases and enforce a mutual agreement, making the parties more willing to settle on a bargaining outcome in the first place. To successfully overcome the tendency of players to defect, institutions should be strong and formalized. When actors have fewer resources, free riding is less of an option and the problem is one of developing mutual confidence in the ability of the other actors to implement a bargain. In such cases, institutions may provide insurance schemes against defection.

Second, institutions may reduce distributive problems. They may eliminate some sharply asymmetric outcomes and may provide focal point solutions for

both cost sharing (which applies to all types of goods) and benefit splitting (when there is no jointness of supply, that is, for private and CPR goods). Institutions with a firmly and widely established meta-regime tend to perform these tasks well.

Third, institutions are not useful only for their help in the initial provision of various types of goods, but also may provide ongoing benefits. For instance, institutions may be useful even if there is some powerful political entrepreneur, or hegemon, that provides public goods. Hegemons wish to coerce weaker actors to move away from free riding on the unilateral provision of goods and wish to have them carry some of the burden. They try to do so by linking the provision of the good with other issues. This can be done through institutions that connect different issues.

Fourth, institutions help clarify the mechanism of exclusion. Goods are rarely naturally given but must often be constructed. Actors in international relations use institutions to construct and enforce exclusionary mechanisms. Careful institutional design might therefore transform problematic goods such as public or CPR goods into less problematic bargaining situations such as club goods.

2.3 Bargaining to Create Institutions

While institutions may help facilitate or alter the costs and benefits arising from the bargaining process, institutions themselves are the result of a bargaining process. We can consider three schools of thought that have been used to explain the formation of institutions (see Aggarwal 1998 for references on these three schools of thought). First, from a neorealist power-based tradition, states in the international system compete for security in an anarchic international system. For analysts in this school, international institutions are simply reflections of the existing balance of power, and institutional outcomes have distributional consequences, that is, benefits from cooperation may be unequal.

The neoliberal tradition, on the other hand, emphasizes the incentives that states have to cooperate. In the neoliberal framework, states create 'rules of the game' (i.e., institutions) based on certain agreed-upon norms. Institutions foster sustained cooperation as well as lower information and transaction costs in a particular issue-area. International institutions are seen to be 'sticky'—that is, existing institutions constrain future institutional change—and new institutions must be designed to be either 'nested' within, parallel, or independent from existing institutions (Aggarwal 1998). Nested institutions are hierarchically ordered. Parallel institutions refer to an institutional division of labor, while independent institutions are simply unconnected since their mandates do not overlap.

The cognitive perspective stresses the role of expert consensus and the interplay of experts and politicians in the formation of institutions. Those within the cognitive tradition are concerned with how the development of knowledge can foster learning on a particular issue area that may drive policy innovation.

3. Future Research Directions

The issue of institutional design, both at the international and domestic level, will remain a key research and policy challenge for the future. Analysts are increasingly recognizing the need to precisely specify the types of bargaining situations for which institutions might be needed. Rather than debating if institutions are necessary or useless, future work will provide us with a more precise understanding of the nature of goods and their relationship to institutions. Together with insight into the factors that contribute to the creation of institutions themselves, we will achieve a better understanding of the conditions under which institutions facilitate and contribute to the bargaining process.

See also: Conflict/Consensus; Conventions and Norms: Philosophical Aspects; Dispute Resolution in Economics; Game Theory; Game Theory: Non-cooperative Games; International Arbitration; International Communication: History; Negotiation and Bargaining: Role of Lawyers; Prisoner's Dilemma, One-shot and Iterated

Bibliography

Aggarwal V K 1998 Reconciling multiple institutions: Bargaining, linkages, and nesting. In: Aggarwal V K (eds.) *Institutional Designs for a Complex World: Bargaining, Linkages, and Nesting*. Cornell University Press, Ithaca, NY

Aggarwal V K, Dupont C 1999 Goods, games, and institutions. *International Political Science Review* **20**: 393–409

Cornes R, Sandler T 1996 *The Theory of Externalities, Public Goods, and Club Goods*. Cambridge University Press, New York

Kaul I, Grinnberg I, Stern M A (eds.) 1999 Global Public Goods. Oxford University Press, New York

Krasner S (ed.) 1983 *International Regimes*. Cornell University Press, Ithaca, NY

Nash J 1951 Non-cooperative games. *Annals of Mathematics Journal* **54**: 286–95

Rapoport A, Guyer M 1966 A taxonomy of 2x2 games. *General Systems*. : 203–14

Snidal D 1979 Public goods, property rights, and political organization. *International Studies Quarterly* **22**: 532–66

Snidal D 1985 Coordination vs. prisoner's dilemma: Implications for international cooperation. *American Political Science Review* **79**: 923–42

Stein A A 1982 Coordination and collaboration: Regimes in an anarchic world. *International Organization* **36**: 294–324

V. K. Aggarwal and C. Dupont

Negotiation and Bargaining: Role of Lawyers

Negotiation is an interactive communication process that occurs whenever we want something from someone else or another person wants something from us (Shell 1999). Often treated by law schools, legal scholars, and lawyers as an alternative dispute resolution or 'ADR' process, see *Mediation, Arbitration, and Alternative Dispute Resolution (ADR)*, negotiation is the ADR process that lawyers in America and around the world are likely to employ most often. Whether functioning as litigators trying civil cases; as transactional or corporate lawyers putting together deals; as policymakers or advisors proposing and revising rules and regulations; or as prosecutors or defense counsel litigating criminal cases, lawyers negotiate.

1. Negotiation Models

Scholars have identified two basic models of, or approaches to, negotiation. The first, and perhaps most familiar, is variously labeled the 'adversarial' or 'distributive,' or positional model of negotiation (Gifford 1985, Lewicki et al. 1994). This adversarial model of negotiation posits that negotiation is a zero-sum game in which any gain one party receives is necessarily at the expense of the other party. Because it views negotiation as a zero-sum game, the adversarial model counsels negotiators to attempt to accumulate as many of those gains as possible through the use of such tactics as extreme opening positions, few (and small) concessions, commitments to positions, and the withholding of relevant information (Gifford 1985). The adversarial model, in short, places a primacy on claiming value in negotiation.

The downside to the adversarial model, at least according to its critics, is that it is often harmful to the relationship between the parties and can easily lead to impasse or to negotiated agreements that leave value on the table. To address these shortcomings, scholars have developed a second approach to negotiation, currently fashionable in American law schools (as well as business, public policy, and international relations schools), called the 'problem-solving' or 'integrative,' or 'principled' model of negotiation (Fisher et al. 1991, Menkel-Meadow 1984, Walton and McKersie 1965). The problem-solving model of negotiation posits that negotiation is not necessarily a zero-sum game. Rather, most negotiations are (or at least have the potential to become) collaborative problem-solving exercises in which the parties work together to devise creative, 'win-win' outcomes. Given this view of negotiation, the problem-solving model counsels negotiators to focus on the parties underlying interests in negotiation, to devise creative options that help satisfy both parties, interests, and to develop and use legitimate, objective criteria that both parties view as applicable to the negotiation (Fisher et al. 1991). The problem-solving model, in short, emphasizes both creating and claiming value in negotiation.

Neither of the primary negotiation models is sufficient to capture the complexity of legal negotiations as a positive matter, and in all likelihood, neither is sufficient as a prescriptive guide, either. Lawyer-negotiators no doubt do, and probably should, borrow liberally from either or both negotiation model(s) depending upon the nature of the negotiation, the relationship between the clients, the practice area, the norms in the community, and relevant legal standards and rules.

Available empirical evidence suggests, however, that lawyer-negotiators tend to rely somewhat more heavily on one model or the other when negotiating. According to one empirical study, for instance, lawyer-negotiators evaluating their adversaries rated 65 percent of them as cooperative or problem-solving negotiators, 24 percent as competitive or adversarial negotiators, and only 11 percent as incapable of categorization (Williams 1983). A more recent empirical study of American civil litigators found that most of them express a preference for the problem-solving model of negotiation but tend to behave more consistently with the adversarial model of negotiation (Heumann and Hyman 1997).

2. Negotiation in the Legal System

Lawyers perform myriad roles in America and around the world. Broadly speaking, though, lawyers' work typically falls into one of two categories: dispute resolution or deal making. Negotiation is central to both.

2.1 Dispute Resolution

Many lawyers function as dispute resolvers. Civil litigators and criminal defense attorneys, for instance, are dispute resolvers. Dispute resolvers spend much of their professional lives helping their clients settle disputes through negotiation.

Most disputes do not require lawyers, however, because most disputants settle their disputes on their

own through negotiation (Miller and Sarat 1980–1). Once disputants have invoked the formal legal system by hiring a lawyer and filing suit, their negotiation is no longer shaped merely by informal norms but also by the law and legal system. The disputants and their lawyers negotiate, in other words, 'in the shadow of the law' because the consequence of failing to reach a negotiated agreement is possible imposition of a binding decision by a third party, i.e., a trial (Mnookin and Kornhauser 1979).

Trials, however, are quite rare in the American civil justice system because most disputants settle their cases through negotiation. The conventional explanation of this phenomenon is economic. According to the economic model of suit and settlement, disputants and their lawyers seek to maximize value in litigation. To do so, they calculate the expected value of trial (i.e., anticipated judgment times the probability it will occur less trial costs), compare it to any settlement offer on the table, and select whichever promises to give them more value. Because the costs of trial are so much greater than the costs of settlement, most disputants opt to settle their cases through negotiation because settlement generally promises more value than trial. Trial, according to this economic account of litigation, results only when the disputants and their lawyers develop widely divergent predictions about the likely outcome at trial. In the absence of widely divergent predictions, disputants will settle their cases because they expect that they will maximize their outcomes through settlement (Priest and Klein 1984). As noted below, however, settlements may break down for other reasons as well.

2.2 Deal Making

Other lawyers function as deal makers. Corporate lawyers, mergers and acquisitions specialists, and securities lawyers, among others, are deal makers. Deal-making lawyers are 'transaction cost engineers' who use differences between the parties (e.g., one party wants to buy a good, while one wants to sell), non-competitive similarities (e.g., licensee and licensor, both of whom want a technology to succeed), economies of scale (e.g., pooling production or consumption facilities to lower costs of a good) and the like to create value for their clients (Gilson 1984, Gilson and Mnookin 1995). Deal-making lawyers also strive to claim value for their clients by negotiating advantageous divisions of the value they have helped to create.

3. Barriers to Negotiation

Whether functioning as dispute resolvers or deal makers, lawyer-negotiators are likely to find that predictable barriers or impediments arise during the negotiation process to prevent them from obtaining optimal negotiation outcomes for their clients. These barriers or impediments fall roughly into three categories: (a) strategic; (b) psychological; and (c) structural (Arrow et al. 1995, Mnookin 1993).

3.1 Strategic Barriers

Strategic barriers arise where lawyer-negotiators consciously adopt strategies and tactics, like hard bargaining techniques, that backfire. Strategic barriers are likely to arise because lawyer-negotiators and their clients find themselves trapped in the so-called negotiator's dilemma (Lax and Sebenius 1986). The negotiators dilemma, like the prisoner's dilemma upon which it is based, focuses on the tension between competitive or value-claiming moves on the one hand and cooperative or value-creating moves on the other.

Every negotiation involves something of value, which the parties can simply divide (i.e., value claiming) or which they can attempt to expand and divide (i.e., value creating and claiming). Negotiators who make competitive or value-claiming moves may succeed in obtaining relatively more of the existing value, but their competitive moves are likely to inhibit the cooperation needed to create additional value for the parties. The negotiator's dilemma refers to negotiators' tendency to focus narrowly on claiming value at the expense of creating additional value for the parties to divide.

The negotiator's dilemma is inherent in negotiation, but it is particularly likely to confound lawyer-negotiators and their clients under certain conditions. For example, the negotiator's dilemma is more likely to impede successful negotiation when the parties face informational asymmetries because value creation depends upon common information (Gilson and Mnookin 1995). When information is asymmetric, parties must agree to share their private information to create value, but parties may be reluctant to do so for fear that their adversaries will exploit the information they have divulged. This is particularly likely to be a problem in litigation.

3.2 Psychological Barriers

Psychological barriers arise due to predictable cognitive and motivational biases that influence the way lawyer-negotiators and their clients interpret information, assess their adversaries, and experience feelings of gain or loss (Korobkin and Guthrie 1994, Ross and Stillinger 1991). Three of the many psychological barriers that may arise include self-serving biases, reactive devaluation, and loss aversion.

Lawyer-negotiators and their clients are likely to fall prey to self-serving biases when making judgments about ensuing negotiations because people generally have difficulty interpreting information in a neutral fashion. To illustrate, researchers asked experimental

subjects to assess damages in a tort case arising from an automobile-motorcycle collision in which the plaintiff had sued the defendant for $100,000. The researchers assigned some subjects to play the role of plaintiff and others the role of defendant, but they provided both sets of subjects with identical information about the case. Despite reviewing identical information, subjects interpreted it in self-serving ways. When asked to predict how much the judge would award, the subjects evaluating the case from the perspective of the plaintiff predicted that the judge would award $14,527 more than the defendant-subjects predicted. When asked to identify what they perceived to be a fair settlement value for the case, plaintiff-subjects selected a value $17,709 higher than the value selected by defendant-subjects (Loewenstein et al. 1993). These self-serving assessments are 'an important determinant of bargaining impasse' (Babcock and Loewenstein 1997).

Not only are lawyer-negotiators and their clients likely to interpret their own positions in self-serving ways, they are also likely to discount or devalue their adversaries' positions due to a phenomenon psychologists have labeled reactive devaluation. For example, psychologists asked two groups of Stanford University undergraduates to evaluate divestment proposals during a period of heated debate over Stanford's divestment policy in South Africa. The researchers told one group that the university had proposed a modest 'deadline' plan rather than the more aggressive 'specific divestment' plan it had considered proposing. The researchers told the other group that the university had proposed the specific divestment plan rather than the deadline plan. Eighty-five percent of the subjects told that the university had proposed the deadline plan rated the specific divestment plan more favorably, but only 40 percent of the subjects told that the university had proposed the specific divestment plan rated it more favorably (Ross and Stillinger 1991). The subjects devalued both proposals simply because they had been offered.

Reactive devaluation is an even more serious problem when the parties distrust one another, as is often the case in negotiation. In another study, for instance, psychologists asked three groups of American subjects in the mid-1980s to evaluate a nuclear disarmament proposal. One group learned that President Ronald Reagan had offered the proposal, another learned that neutral 'strategy analysts' had offered the proposal, and a third group learned that Mikhail Gorbachev, the Soviet leader at the time, had offered the proposal. Ninety percent of subjects rated the proposal as evenhanded or favorable to the USA when attributed to Reagan, and 80 percent did so when the proposal was attributed to neutral strategy analysts. When attributed to Gorbachev, however, only 44 percent rated it evenhanded or favorable to the USA (Ross and Stillinger 1991). In short, 'the very offer of a particular proposal or concession' especially if the offer comes from an adversary 'may diminish its apparent value or attractiveness in the eyes of the recipient' (Ross 1995).

Reactive devaluation may be exacerbated by loss aversion, another significant psychological barrier that can arise during the negotiation process. Psychologists have found that negotiators do not experience equivalent gains and losses the same way. Rather, losses loom larger 'generally twice as large' as corresponding gains (Kahneman et al. 1991, Kahneman and Tversky 1984). Loss aversion may manifest itself in negotiation in a couple of different ways.

First, loss aversion is likely to affect negotiators' valuations of goods in transactional negotiations. Sellers are likely to place a higher value on goods than buyers because sellers experience the transfer of the good as a loss while buyers experience the transfer of the good as a gain. In an often-replicated study, researchers divided experimental subjects into a sellers group and a buyers group. The researchers gave members of the sellers group a coffee mug and the members of the buyers group $6.00 in cash. The researchers asked the sellers to indicate the minimum amount they would accept to sell the coffee mug, asked the buyers to indicate the maximum amount they would pay for a mug, and informed all of the subjects that they would execute trades based on the responses given. The researchers found that the two groups valued the mugs quite differently. Sellers, who would experience a loss if they parted with their mug, assigned a median price of $5.25, while buyers, who would experience the attainment of the mug as a gain, assigned a median price of $2.25–$2.75 (Kahneman et al. 1991). In short, sellers tend to value goods much more highly than buyers, which may prevent lawyer-negotiators from obtaining optimal outcomes for their clients in transactional negotiations.

Second, loss aversion may influence the way negotiators perceive concessions. Negotiators view concessions they make as losses but concessions they receive from the other side as gains. Because losses are weighted more heavily than gains, negotiators are likely to view the concessions they make as more significant than the equivalent concessions their adversaries make. This 'concession aversion' may also undermine successful negotiation (Kahneman and Tversky 1995).

3.3 Structural Barriers

The third type of barrier that may arise in negotiation is structural. Structural barriers are a product not of individual negotiators' calculations or cognitive processes but rather of the way in which conflict and conflict resolution are managed. The most significant structural impediment to successful negotiation faced by lawyer-negotiators and their clients arises out of the lawyer–client relationship.

In the lawyer–client relationship, the client is the principal who possesses substantive decision-making authority, while the lawyer is an agent who is ethically obligated to carry out the client's wishes. The lawyer–client relationship is a classical principal-agent relationship (Pratt and Zeckhauser 1985). The problem with any principal–agent relationship is that incentives are seldom perfectly aligned between principal and agent. This misalignment of incentives can have a detrimental impact on negotiation.

To give one concrete illustration, consider a typical personal injury lawsuit in the USA. In such a case, the plaintiff is usually a one-shot player with limited resources who hires a lawyer on a contingency fee basis, while the defendant is often a repeat-player organization of some type that hires a lawyer on an hourly-fee basis (Galanter 1974, Gross and Syverud 1996). Principal–agent problems may arise for both plaintiffs and defendants.

The plaintiff's lawyer who takes the case on a contingency basis recovers a fee only if the plaintiff succeeds in obtaining a settlement or judgment. If the plaintiff is successful, the plaintiff's lawyer will receive a percentage of the recovery. Because of this financing scheme, the plaintiff's lawyer may have an incentive to settle quickly so as to maximize his hourly return, but the plaintiff may fare better if the plaintiff's lawyer settles later or takes the case to trial. Suppose, for instance, that a plaintiff's lawyer on a 25 percent contingency settles a case for $10,000 after 10 h of work. In this instance, the plaintiff receives a net recovery of $7500, and the plaintiff's attorney receives $2500 for his efforts. On an hourly basis, the plaintiff's attorney earns a $250 h wage. Now suppose, instead, that the plaintiff's lawyer settles the case for $16,000 after conducting 40 h of work. In this instance, the plaintiff does better, obtaining a net recovery of $12,000. By contrast, the plaintiff's attorney does worse, earning only $100 h for his efforts (i.e., $4000 for 40 h work). In short, the plaintiff's attorney would do better to settle earlier, while the plaintiff would do better to settle late.

Defense counsel working on an hourly-fee basis faces a very different incentive structure than the plaintiff's lawyer. Defense counsel's hourly wage is constant regardless of the number of hours worked. Thus, defense counsel has a financial incentive to work the case as long as possible because that will maximize her earnings. Her client, by contrast, is likely to fare better if the case settles earlier. Suppose, for instance, that defense counsel charging $200 h settles a case for $10,000 after working 10 h on the case. In this instance, settlement costs the defendant $12,000 (i.e., $10,000 to plaintiff and $2000 to defense counsel) and earns defense counsel $2000. Now suppose, instead, that defense counsel settles for $6000 after spending 40 h on the case. In this instance, settlement costs the defendant more (i.e., $6000 to plaintiff and $8000 to defense counsel) but earns defense counsel more (i.e., $8000 vs. $2000). In short, defense counsel would do better to settle later, while defendant would do better settling earlier.

4. *The Future of Negotiation Research*

Because of its importance in law, business, foreign affairs, and interpersonal relationships, negotiation will continue to attract the attention of scholars and practitioners trying to shed light on the process. Three research trends are likely. First, scholars will propose additional models of negotiation that combine or challenge the existing adversarial and problem-solving models of negotiation. Second, scholars will explore opportunities that new communications technologies create (as well as problems they cause) for negotiation. Finally, scholars will continue to conduct empirical and experimental investigations into barriers to optimal negotiation outcomes. Ideally, scholars will not only identify additional barriers but will develop some appreciation of the magnitude of these barriers, learn how the barriers might interact with one another, and devise prescriptions for overcoming the barriers. Continued investigation of the negotiation process is important because negotiation will continue to be the process by which most disputes are resolved and most deals get done.

See also: Contracts: Legal Perspectives; Dispute Resolution in Economics; Lawyers; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Negotiation and Bargaining: Organizational Aspects; Parties: Litigants and Claimants; Plea Bargaining

Bibliography

Arrow K J, Mnookin R H, Ross L, Tversky A, Wilson R (eds.) 1995 *Barriers to Conflict Resolution*. W.W. Norton, New York

Babcock L, Loewenstein G 1997 Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives* **11**: 109–19

Fisher R, Ury W with Patton B (ed.) 1991 *Getting to Yes: Negotiating Agreement Without Giving In*, 2nd edn. Houghton-Mifflin, Boston

Galanter M 1974 Why the 'Haves' come out ahead: Speculations on the limits of legal change. *Law & Society Review* **9**: 95–7

Gifford D G 1985 A context-based theory of strategy selection in legal negotiation. *Ohio State Law Journal* **46**: 41, 48–52

Gilson R J 1984 Value creation by business lawyers: Legal skills and asset pricing. *Yale Law Journal* **94**: 239–313

Gilson R J, Mnookin R H 1995 Foreword: Business lawyers and value creation for clients. *Oregon Law Review* **74**: 1, 7–11

Gross S R, Syverud K D 1996 Don't Try: Civil jury verdicts in a system geared to settlement. *U.C.L.A. Law Review* **44**(1): 1–64

Heumann M, Hyman J M 1997 Negotiation methods and litigation settlement methods in New Jersey: 'You can't always get what you want'. *Ohio State Journal on Dispute Resolution* **12**: 253,255–56

Kahneman D, Knetsch J L, Thaler R H 1991 Anomalies—The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives* **5**: 193–206
Kahneman D, Tversky A 1984 Choices, values, and frames. *American Psychologist* **39**: 341–50
Kahneman D, Tversky A 1995 Conflict resolution: A cognitive perspective. In: Arrow K J, Mnookin R H, Ross L, Tversky A, Wilson R (eds.) *Barriers to Conflict Resolution*. W.W. Norton, New York, pp. 44, 54–7
Korobkin R, Guthrie C 1994 Psychological barriers to litigation settlement: An experimental approach. *Michigan Law Review* **93**: 107–92
Lax D A, Sebenius J K 1986 *The Manager as Negotiator: Bargaining for Cooperation and Competitive Gain*. The Free Press, New York, pp. 38–41
Lewicki R J, Litterer J A, Minton J W, Saunders D M 1999 *Negotiation*. 2nd edn. Irwin, Burr Ridge, IL, pp. 62–77
Loewenstein G, Issacharoff S, Camerer C, Babcock L 1993 Self-serving assessments of fairness and pretrial bargaining. *Journal of Legal Studies* **22**: 135–59
Menkel-Meadow C 1984 Toward another view of legal negotiation: The structure of problem solving. *U.C.L.A. Law Review* **31**: 754–842
Miller R E, Sarat A 1981 Grievances, claims, and disputes: Assessing the adversary culture. *Law and Society Review* **15**: 525–66
Mnookin R H 1993 Why negotiations fail: An exploration of barriers to the resolution of conflict. *Ohio State Journal on Dispute Resolution* **8**: 235
Mnookin R H, Kornhauser L 1979 Bargaining in the shadow of the law: The case of divorce. *Yale Law Journal* **88**: 950–97
Pratt J W, Zeckhauser R (eds.) 1985 *Principals and Agents: The Structure of Business*. Harvard Business School Press, Boston, MA
Priest G L, Klein B 1984 The selection of disputes for litigation. *Journal of Legal Studies* **13**: 1–55
Ross L 1995 Reactive devaluation in negotiation and conflict resolution. In: Arrow K J, Mnookin R H, Ross L, Tversky A, Wilson R (eds.) *Barriers to Conflict Resolution*. W.W. Norton, New York, pp. 26, 28
Ross L, Stillinger C 1991 Barriers to conflict resolution. *Negotiation Journal* **7**: 389–404
Shell G R 1999 *Bargaining for Advantage: Negotiation Strategies for Reasonable People*. Viking, New York, p. 6
Walton R E, McKersie R B 1965 *A Behavioral Theory of Labor Negotiations: An Analysis of a Social Interaction System*. McGraw-Hill, New York, pp. 126–83
Williams G R 1983 *Legal Negotiation and Settlement*. West Publishing, St. Paul, MN, p. 18

C. Guthrie

Negotiations: Statistical Aspects

1. Introduction

The term 'negotiation analysis' encompasses an emergent body of work that seeks to generate prescriptive theory and advice for parties wishing to negotiate in a 'rational' manner, without necessarily assuming that the other side(s) is a strategically sophisticated, rational actor.

To what bodies of theory and empirical work might one turn for advice in a negotiation over selling a car or forging a global warming treaty? Decades of psychological studies of people negotiating offer many powerful insights but generally lack a prescriptive theory of action. Instead, this behavioral approach has largely sought accurate descriptions and analytic explanations of how negotiators do act. What they should do often remains an implicit or *ad hoc* implication of this careful work.

Decision analysis, the practically oriented cousin of decision theory, would seem a logical candidate to fill this prescriptive void. It suggests a systematic decomposition of the problem: structuring and sequencing the parties' choices and chance events, then separating and subjectively assessing probabilities, values, risk, and time preferences. A von Neumann–Morgenstern expected utility criterion aggregates these elements in ranking possible actions to determine the optimal choice. This approach is well-suited to decisions 'against nature,' in which the uncertainties—such as the probability that a hurricane will strike Miami in August—are not affected by the choices of other involved parties anticipating one's actions.

Yet when decision-making is interactive—as is true in negotiation, where each party's anticipated choices affects the other's and vice versa—assessment of what the other side will do qualitatively differs from assessment of 'natural' uncertainties. Of course, the theory of games was developed to provide a logically consistent framework for analyzing such interdependent decision-making. Full descriptions of the courses of action open to each involved party are encapsulated into 'strategies.' Rigorous analysis of the interaction of strategies leads to a search for 'equilibria,' or plans of action such that each party, given the choices of the other parties, has no incentive to change its plans. A great deal of analysis by game theorists seeks conditions for unique equilibria among such strategies (von Neumann and Morgenstern 1944, Luce and Raiffa 1957, Aumann 1989).

Game theory has been especially useful for understanding repeated negotiations in well-structured financial situations. It has offered useful guidance for the design of auction and bidding mechanisms, has uncovered powerful competitive dynamics, has usefully analyzed many 'fairness' principles, and now flourishes both on its own and in applications such as microeconomic theory. Despite signal successes, the dominant game-theoretic quest to predict equilibrium outcomes resulting from the strategic interactions of fully rational players often lacks prescriptive power in negotiations.

Three major aspects of mainstream game theory, discussed at length in Sebenius (1992), contribute to this 'prescriptive gap.' First, on standard assumptions,

there are often numerous plausible equilibrium concepts, each with many associated equilibria—and no *a priori* compelling way to choose among them. Second, even where one party wishes to act rationally, the other side may not behave as a strategically sophisticated, expected utility-maximizer—thus rendering conventional equilibrium analyses less applicable. A large and growing body of evidence—especially in 'behavioral game theory' and experimental economics—suggests that people systematically and significantly violate the canons of rationality. Third, the elements, structures, and 'rules' of many negotiating situations are not completely known to all the players, and even the character of what is known by one player may not be known by another. The frequent lack of such 'common knowledge' limits—from a prescriptive standpoint—much equilibrium-oriented game analysis, though recent work has begun to relax the widespread understanding that common knowledge of the situation was essential to game models (Aumann and Brandenburger 1995). Even where it is possible to bend such a situation into the form of a well-posed game, and gain insights from it, the result may lose prescriptive relevance.

2. *The Negotiation Analytic Response*

If descriptive psychological approaches to negotiation lack a prescriptive framework, if decision analysis is not directly suited to interactive problems, and if traditional game theory presupposes too much rationality and common knowledge, then 'negotiation analysis' represents a response that yokes the prescriptive and descriptive research traditions. Using Howard Raiffa's (1982) terms, unlike the 'symmetrically prescriptive' approach of game theory, wherein fully rational players are analyzed in terms of what each should optimally do given the other's optimal choices, the 'asymmetrically prescriptive/descriptive' approach typically seeks to generate prescriptive advice to one party conditional on a (probabilistic) description of how others will behave. This need not mean tactical naiveté; as appropriate, the assessments can incorporate none, a few, or many rounds of 'interactive reasoning.'

Works that embody the spirit of this work can be found as early as the late 1950s. While Luce and Raiffa's *Games and Decisions* (1957) was primarily a brilliant synthesis and exposition of game theory's development since von Neumann and Morgenstern's (1944) classic work, Luce and Raiffa began to raise serious questions about the limits of this approach in analyzing actual interactive situations. Perhaps the first work that could be said to be in the spirit of 'negotiation analysis' was *The Strategy of Conflict* by Thomas Schelling (1960). Its point of departure was game-theoretic but it proceeded with less formal argument and the analysis had far broader direct scope. Though nominally in the behavioral realm, Walton and Mckersie's *Behavioral Theory of Labor Negotiation* (1965) drew on Schelling's work as well as rudimentary decision and game theories.

The first overall synthesis of this emerging field appeared with Raiffa's (1982) *The Art and Science of Negotiation*, elaborated in Raiffa (1997). Building on Sebenius (1984), this approach was systematized into an overall method in Lax and Sebenius' *The Manager as Negotiator* (1986). *Negotiation Analysis* (1991), edited by Peyton Young, furthered this evolving tradition, which was characterized and reviewed in Sebenius (1992). Further contributions in the same rationalist vein were included Zeckhauser et al. (1996), Arrow et al. (1995), and, adding insights from organizational and information economics, in Mnookin et al. (2000).

Meanwhile, another group of researchers was coming to a negotiation analytic view from a behavioral point of departure. With roots in the cognitive work of behavioral decision theorists, e.g., Bell et al. (1988), behavioral scholars began in the early 1990s to explicitly link their work to that of Raiffa and his colleagues. In particular, Neale and Bazerman's *Cognition and Rationality in Negotiation* (1991), Bazerman and Neale (1992), and Thompson (2001) pulled together and developed a great deal of psychological work on negotiation in an asymmetrically prescriptive/descriptive framework. These efforts began systematically to build up more structure on what had been, in the works of Raiffa et al. a largely *ad hoc* descriptive side of the ledger.

3. *Elements of a Negotiation Analytic Approach*

Full negotiation analytic accounts (e.g., Sebenius 1992, 2000) generally consider the following basic elements: the actual and potential parties; their perceived interests; alternatives to negotiated agreement; the linked processes of 'creating' and 'claiming' value; and the potential to 'change the game' itself.

3.1 *Parties*

In the simplest negotiation, two principals negotiate with each other. Yet potentially complicating parties such as lawyers, bankers, and other agents may be present as may multiple internal factions with very different interests. The crucial first step for an effective negotiation analysis is to map the full set of potentially relevant parties in the context of the decision processes.

3.2 *Interests*

The next step is to probe deeply for each relevant party or faction's underlying interests and to assess their

tradeoffs carefully. When individuals or groups with different concerns constitute a negotiating 'side,' it is no longer in general possible to specify overall tradeoffs; however, carefully tracing which set of interests is ascendant according to the internal bargaining process of given factions may continue to provide insights. Raiffa (1982) offers an extended discussion of assessing tradeoffs in negotiation, building on extensive work by Keeney and Raiffa (1976, Sebenius 1992). This assessment is radically subjective in the sense that less tangible concerns for self-image, fairness, process, precedents, or relationships can have the same analytic standing as the 'harder' or 'objective' interests such as cost, time, and quality that are common to traditional economic approaches.

It is often useful to distinguish the full set of the parties' underlying *interests* from the *issues* under negotiation, on which *positions* or stands are taken. The connections among positions on issues and interests are rarely simple. Sometimes a focus on deeper interests can unblock a stubborn impasse over incompatible positions that relate only partially to the parties' real concerns; in other cases, emphasizing interests will only generate hopeless conflict when mutually beneficial agreement on certain overt positions could be reached.

3.3 Alternatives to Negotiated Agreement

People negotiate in order to satisfy the totality of their interests better through some jointly decided action than they could otherwise. Thus, for each side, the basic test of a proposed joint agreement is whether it offers higher subjective worth than that side's best course of action absent agreement. In examining a negotiation, one should analyze each party's perceptions of its own—and the others'—valuations of their alternatives to negotiated agreement.

Alternatives to agreement may be certain, with a single attribute: an ironclad competing price quote for an identical new car. They may be contingent and multi-attributed: going to court rather than accepting a negotiated settlement can involve uncertainties, trial anxieties, costs, time, and precedents that contrast with the certain, solely monetary nature of a pretrial accord. No-agreement alternatives may also involve potentially competing coalitions, threats, and counterthreats or the necessity to keep negotiating indefinitely. Various analytic tools from decision analysis to optimal search theory may aid in assessment of such alternatives (see Sebenius 1992 for references).

While this evaluation provides a strict lower bound for the minimum worth (the 'reservation price') required of any acceptable settlement, alternatives to agreement also play tactical roles. The more favorably negotiators portray their best alternative course of action and willingness to 'walk away,' the smaller is the ostensible need for the negotiation and the higher the standard of value that any proposed accord must reach. Moves 'away from the table' that shape the parties' alternatives to agreement can strongly affect negotiated outcomes. Searching for a better price or another supplier, cultivating a friendly merger partner in response to a raider, or preparing an invasion should talks fail, may influence negotiations more than sophisticated tactics employed 'at the table' such as clever opening offers or patterns of concession.

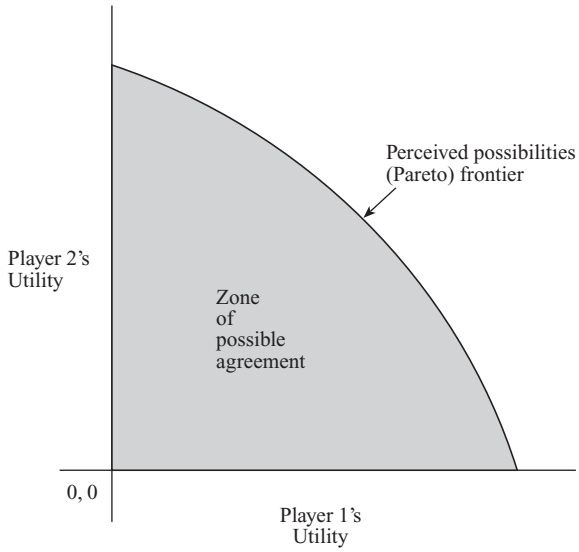

Figure 1
Representation of the zone of possible agreement

3.4 Representing the Structure

Imagine that two negotiators have thought hard about their underlying interests in different possible settlements of the apparent issues. Further, suppose that they have a relatively clear, if possibly changing, assessment of their tradeoffs and have compared them to the value of their best no-agreement alternatives. Each has a sense of any 'rules of engagement' that structure their interaction. From the viewpoint of each party, a set of possible agreements has been envisioned. Assume that an analyst were privy to the results of these evaluations by each side, along with the (likely asymmetric) distribution of information about interests, beliefs, no-agreement options, and possible actions; these evaluations need not be common knowledge of the parties. The situation might be familiarly represented as in Fig. 1.

The origin represents the value of failing to reach agreement: each side's best alternative to agreement implies the location of this point. The 'Pareto frontier' in the northeast part of the graph represents the evaluations of the set of those possible agreements on the issues that could not be improved on from the standpoint of either party without harming the other.

In general, neither side knows the location of the frontier, only theoretically that it is there. The entire shaded region—bounded by the two axes and the frontier—is the 'zone of possible agreement.' In general, each party has its own perceptions of it. (In a purely distributive bargain, with no room for joint gains beyond the fact of agreement, the shaded region would collapse to a diagonal frontier.) Since this representation is quite general, it can in principle encompass the whole range of possible interests, alternatives, and agreements.

3.5 From Structure to Outcome: Favorable Changes in the Perceived Zone of Possible Agreement

This is the point at which game-theoretic and negotiation-analytic approaches tend to diverge. A game theorist would typically tighten the above specification of the situation, presume common knowledge of the situation and strategic rationality of the parties, and, by invoking a rigorous concept of equilibrium, investigate the predicted outcome(s) of the interaction. Indeed, as Rubenstein's insightful commentary noted, 'for forty years, game theory has searched for the grand solution' that would achieve a

> prediction regarding the outcome of interaction among human beings using only data on the order of events, combined with a description of the players' preferences over the feasible outcomes of the situation. (1991, p. 923)

Even with as powerful and ubiquitous a concept as that of the Nash equilibrium in non-cooperative games, however, it is often impossible, even with the imposition of increasingly stringent requirements or refinements to limit a game's equilibrium outcomes to a unique or even small number of points. Often there is an infinitude of such outcomes. As Tirole noted when explaining why 'we are now endowed with nearly a dozen refinements of perfect Bayesian equilibrium,' the

> leeway in specifying off-the-equilibrium-path beliefs usually creates some leeway in the choices of equilibrium actions; by ruling out some potential equilibrium actions, one transforms other actions into equilibrium actions. (1988, p. 446)

Frequently this implies an infinite number of perfect Bayesian equilibria.

Despite insights into how rational players might select from among multiple Nash equilibria (Harsanyi and Selten 1988), the rationale for a particular choice may ultimately seem arbitrary. As Kadane and Larkey remark,

> we do not understand the search for solution concepts that do not depend on the beliefs of each player about the others' likely actions, and yet are so compelling that they become the obvious standard of play for all those who encounter them. (1982, pp. 115–6)

This seems especially apt in light of their observation that 'solution concepts are a basis for particular prior distributions' and hence

> the difficulty in non-zero sum, N-person game theory of finding an adequate solution concept: no single prior distribution is likely to be adequate to all players and all situations in such games. (1982, pp. 115–6)

Since each party should accept any settlement in the zone of possible agreement rather than no agreement, Schelling (1960) made the potent observation that the outcome of such a situation could only be unraveled by a 'logic of indeterminate situations.' Without an explicit model or formal theory (equilibrium-based or other) adequate to map structure and tactics onto bargaining outcomes with confidence, how can an individual negotiator or interested third party decide what to do? In the (often implicit) view of many negotiation analysts, the negotiator's subjective distribution of beliefs about the negotiated outcome conditional on using the proposed tactics must be compared with his subjective distribution of beliefs about the outcome conditional on not using them. The tactic is attractive if the former distribution gives him higher expected utility than the latter.

Such 'improvement' has a subjective basis analogous to the Rothschild and Stiglitz (1970) characterization of a subjectively perceived 'increase' in risk. Specifying these distributions may require an internalized and subjective model of the bargaining process, since no such general model exists; where there is a well-developed and applicable game-theoretic model, of course, it should be used. Of course, the 'better' the empirical and theoretical basis for the assessment, the 'better' the subjective distributions of outcomes.

Much negotiation analytic work consists in improving the basis for assessing such outcome distributions. To do this, analysts pay attention to the special structure and dynamics that derive from the joint decision-making among parties, some of whom may be 'boundedly' rational.

3.6 Improving the Basis for Assessing Outcome Distributions I: Fundamental Processes of Negotiation

At bottom, negotiation processes involve both (a) actions to enhance what is jointly possible through agreement ('creating value'), and (b) actions to allocate the value of agreement ('claiming value').

(a) *Creating Value*. In most negotiations, the potential value of joint action is not fully obvious at the outset. 'Creating value'—that is, reaching mutually beneficial agreements, improving them jointly, and preventing conflict escalation—requires an approach often associated with 'win-win,' 'integrative,' or 'variable sum' encounters: to share information; communicate clearly; be creative; and productively chan-

nel hostilities. Yet regardless of whether one adopts a cooperative style or not, it is useful to have an analytic guide as to the underlying substantive bases for joint gains and how to realize them by effective deal-crafting.

First, apart from pure shared interests, negotiators may simply want the same settlement on some issues. Second, where economies of scale, collective goods, alliances, or requirements for a minimum number of parties exist, agreement among similar bargainers can create value. Third—and most interestingly—though many people instinctively seek 'common ground' and believe that 'differences divide us,' it is often precisely the differences among negotiators that constitute the raw material for creating value. Each class of difference has a characteristic type of agreement that makes possible its conversion into mutual benefit. For example, differences in relative valuation suggest joint gain from trades or from 'unbundling' differently valued attributes. Differences in tax status, liquidity, or market access suggest arbitrage. Complementary technical capacities can be profitably combined. Probability and risk aversion differences suggest contingent agreements or bets. Differences in time preference suggest altering schedules of payments and other actions. Sebenius (1984) formally characterizes such optimal betting, risk sharing, and temporal reallocations. Much work on optimal contracting (e.g., Hart 1995) as well as classical economic study of gains from trade and comparative advantage directly bear on the kinds of agreements that create value on a sustainable basis. Sebenius (1992) and Lax and Sebenius (2000) offer extensive illustrations of these 'deal-crafting' principles in practice.

(b) *Claiming Value*. Crucial aspects of most negotiations, however, are primarily 'distributive,' 'win-lose,' or 'constant-sum,' that is, at some points in the process, increased value claimed by one party implies less for others. Several broad classes of tactics used for 'claiming value' in these kinds of bargains have been explored (e.g., Schelling 1960, Walton and McKersie 1965, Raiffa 1982, and Lax and Sebenius 1986). Such tactics include: shaping perceptions of alternatives to agreement; making commitments; influencing aspirations; taking strong positions; manipulating patterns of concessions; holding valued issues 'hostage'; linking issues and interests for leverage; misleading other parties; and, exploiting cultural and other expectations. By means of these tactics, one party seeks advantage by influencing another's perceptions of the zone of possible agreement.

3.7 Managing the Tension Between Creating and Claiming Value: The Negotiators' Dilemma

In general, the benefits of cooperation are not fully known at the outset of a negotiation. Colloquially, the parties often do not know how large a value pie they can create. The way in which they attempt to expand the pie often affects its final division, while each side's efforts to obtain a larger share of the pie often prevents its expansion in the first place—and may lead to no pie at all, or even to a fight. Conclusion: creating and claiming value are not in general separable processes in negotiation. This fact undermines much otherwise useful advice (that, for example, presumes 'win-win' situations to have no 'win-lose' aspects, or 'integrative' bargains to be unrelated to 'distributive' ones).

Each negotiating party tends to reason as follows: if the other parties are open and forthcoming, I can take advantage of them and claim a great deal of value; thus I should adopt a value-claiming stance. By contrast, if the other parties are tough and adopt value-claiming stances, I must also adopt such a stance in order to protect myself. Either way, a strong tendency operating on all parties drives them toward hardball. Since mutual hardball generally impedes value creation, competitive moves to claim value individually often drive out cooperative moves to create it jointly. Outcomes of this dynamic include poor agreements, deadlocks, and conflict spirals. This tendency, closely related in structure to the famous prisoner's dilemma, was dubbed the 'Negotiator's Dilemma' (Lax and Sebenius 1986, Chaps. 2 and 7). Much negotiation advice is aimed at productively managing the inherent tension between creating and claiming value, especially to do so on a sustainable basis (Lax and Sebenius 1981, Raiffa 1982).

3.8 Improving the Basis for Assessing Outcome Distributions II: Behavioral Insight

A fully rational 'baseline' analysis helps to understand the possible responses of a rational other side. Urging consistent, if not fully rational, behavior on the subject of one's advice is often wise. After all, well-structured, repeated negotiations may penalize departures from rational behavior. Yet many negotiating situations are neither well-structured, repeated, nor embedded in a market context. While negotiators normally exhibit intelligent, purposive behavior, there are important departures from the 'imaginary, idealized, super-rational people without psyches' (Bell et al. 1988, p. 9) needed by many game-theoretic analyses.

Considerable empirical work—such as that cited above by Bazerman, Neale, Thompson, and their colleagues—offers considerable insight into how people actually behave in negotiations. Excellent reviews of the psychological side of negotiation can be found in Bazerman et al. (2000) and, focusing especially on developments on the social psychological side, in Bazerman et al. (2000). Complementing this work is the burgeoning research in experimental economics (Kagel and Roth 1995) and what Colin

Camerer (1997) described as a 'behavioral game theory.' This work blends game-theoretic and psychological considerations in rigorous experimental settings. Two related levels are consistently important, the individual and the social:

(a) *Individual negotiating behavior*. As negotiators, people have different histories, personalities, motivations, and styles. Systematic characteristics of individual cognitive processes can both help and hinder the process of reaching agreement. For example, negotiators may be anchored by irrelevant information, subject to inconsistencies in the way they deal with uncertainty, hampered by selective perception, obsessed by sunk costs and past actions, prone to stereotyping and labeling, susceptible to influence tactics and manipulation by the way in which equivalent situations are framed, liable to see incompatible interests when joint gains are possible, and use a variety of potentially misleading heuristics to deal with complexity, ambiguity, and conflict.

(b) *Social behavior in negotiation*. In groups of two or more, especially where there is some perceived conflict, a variety of social psychological dynamics come into play that may enable or block negotiated outcomes. For example, a powerful norm toward reciprocity operates in most groups and cultures. Tentative cooperative actions by one side can engender positive reactions by the others in a cycle that builds trust over time. By contrast, social barriers can involve aspects of the interactive process that often lead to bad communication, misattribution, polarization, and escalation of conflict, as well as group dynamics that work against constructive agreements. Such dynamics may be especially pronounced when negotiations involve players of different genders, races, or cultures. Group dynamics can involve pressures for conformity, a hardening of approach by a representative before an 'audience' of constituents, bandwagon effects, and the individual taking cues from the behavior of others to decide on appropriate actions in the negotiation.

This experimentally based work is developing an empirical grounding for the behavioral basis of much *a priori* theorizing in economics and game theory. For negotiation analysis, these experimental approaches to actual behavior help to remedy a key defect of prior game theoretic work. While for the most part not prescriptively framed, this body of work also provides rigorous evidence and theory on how people in fact are likely to behave—to inform assessments of outcome distributions and against which to optimize as appropriate.

3.9 Changing the Game

Much existing theory proceeds from the assumption of a well-specified and fixed situation within which negotiation actions are taken. Yet purposive action on behalf of the parties can change the very structure of the situation and hence the outcomes. Often actions can be understood as a tacit or explicit negotiation over what the game itself will be. First investigations into this phenomenon gave rise to the terms 'negotiation arithmetic' or 'adding' and 'substracting' issues and parties (Sebenius 1984). This means that a perfectly legitimate and potentially valuable form of analysis may involve a search for ways to change the perceived game—even though the menu of possibilities may not be common knowledge.

Issues can be linked or separated from the negotiation to create joint gains or enhance leverage. Parties may be 'added' to a negotiation to improve one side's no-agreement alternatives, as well as to generate joint gains or to extract value from others. Though perhaps less commonly, parties can also be 'subtracted'—meaning separated, ejected, or excluded—from larger potential coalitions. For example, the Soviets were excluded from an active Middle East negotiating role in the process leading up to the Camp David accords that only involved Israel, Egypt, and the United States. The process of choosing, then approaching and persuading, others to go along may best be studied without the common assumption that the game is fully specified at the outset of analysis; Sebenius (1996) dissects and offers many examples of the process sequencing to build or break coalitions. Walton and McKersie (1965) focused on how negotiators seek to change perceptions of the game by what they called 'attitudinal restructuring.' In the context of competitive strategy and thinking, Brandenburger and Nalebuff (1996) develop a powerful, analogous logic for 'changing the game' that provides an overall approach and many ingenious examples of this phenomenon.

Refer back to Fig. 1, an improvement in Party One's no-agreement alternative shifts the vertical axis to the right, leaving the bargaining set generally more favorable to that side. If Party Two's no-agreement alternative worsens, the horizontal axis shifts down, worsening its prospects. A successful commitment to a bargaining position cuts off an undesired part of the zone of possible agreement for the party who makes it. A new, mutually beneficial option (e.g., suggestion of a contingent, rather than an unconditional, contract) causes the frontier to bulge upward and to the right, reducing the parties' 'conflict of interest.' When issues change or other basic aspects of the game vary, each side's perceptions of the basic picture in Fig. 1, the zone of possible agreement, will be transformed. These possibilities add evolutionary elements to the analysis.

3.10 The Approach as a Whole

Figure 1 can now be seen to summarize visually the extended negotiation analytic 'model' of possible joint

action. Parties determine the axes; interests provide the raw material and the measure; alternatives to agreement imply the limits; agreements hold out the potential. Within this configuration, the process consists of creating and claiming value, which gives rise to characteristic dynamics, yet, the elements of the interaction may themselves evolve or be intentionally changed. In this sense, the elements of the approach form a logically consistent, complete whole oriented around perceptions of the zone of possible agreement.

In the skeptical view of Harsanyi, this negotiation analytic approach might boil down to

> the uninformative statement that every player should maximize expected utility in terms of his subjective probabilities without giving him the slightest hint of how to choose these subjective probabilities in a rational manner. (1982)

Yet, as described above, distinct classes of factors have been isolated that appear to improve subjective distributions of negotiated outcomes. Understanding the dynamics of creating and claiming value can improve prescriptive confidence. Psychological considerations can help as can cultural observations, organizational constraints and patterns, historical similarity, knowledge of systematic decision-making biases, and contextual features. Less than full-blown game-theoretic reasoning can offer insight into strategic dynamics as can blends of psychological and game-theoretic analysis. When one relaxes the assumptions of strict, mutually expected, strategic sophistication in a fixed game, Raiffa's conclusion is appealing: that some 'analysis—mostly simple analysis—can help' (1982, p. 359).

4. Conclusions and Further Directions

Naturally, there are many other related topics ranging from game-theoretic concepts of fairness for purposes of mediation and arbitration to various voting schemes. More elaborate structures are under study. For example, where negotiation takes place through agents, whether lawyers or diplomats, or where a result must survive legislative ratification, the underlying structure of a 'two-level game' is present (see Putnam 1988). Negotiations also take place in more complex multilevel and coalitional structures (Raiffa 1982, Lax and Sebenius 1986, 1991, Sebenius 1996). In many cases, the individual negotiation is not the appropriate unit of analysis. Instead, the study of 'negotiation design'—for example, of international conferences, alternative dispute resolution procedures, grievance negotiation processes—seeks to maximize the chances of constructive outcomes from different structures and processes (Sebenius 1991).

While game theorists and behavioral scientists will continue to make valuable progress in understanding negotiation, a complementary prescriptive approach has been developing that conditions its prescriptions on the likely behavior of the other side, fully 'rational' or not, and regardless of whether the 'game' is fixed and entirely common knowledge. In describing the logic of negotiation analysis and the concepts and tools that can facilitate it, this discussion has not stressed the many useful ideas that arise from focusing on interpersonal and cultural styles, on atmosphere and logistics, on psychoanalytic motivation, on communication, or on other aspects. Yet because the logic is general, it can profitably accommodate insights from other approaches as well as from experience. The basic elements of this logic—parties' perceptions of interests, alternatives, agreements, the processes of creating and claiming value, and changing the game—become the essential filters through which other factors must be interpreted for a meaningful assessment of the zone of possible agreement and its implications for the outcome.

See also: Game Theory; Game Theory and its Relation to Bayesian Theory; Game Theory: Noncooperative Games; Negotiation and Bargaining: Organizational Aspects; Negotiation and Bargaining: Role of Lawyers; Organizations: Negotiated Orders; Trade Unions, Economic Behavior of

Bibliography

Arrow K J, Wilson R, Ross L, Tversky A, Mnookin R H 1995 *Barriers to Conflict Resolution*. W W Norton, New York

Aumann R J 1989 Game theory. In: Eatwell J, Milgate M, Newman P (eds.) *Game Theory*. Norton, New York, pp. 1–53

Aumann R J, Brandenburger A 1995 Epistemic conditions for Nash equilibrium. *Econometrica* **63**: 1161–80

Bazerman M H, Curhan J, Moore D 2000 The death and rebirth of the social psychology of negotiations. In: Fletcher G, Clark M (eds.) *Blackwell Handbook of Social Psychology: Interpersonal Processes*. Blackwell, Oxford, UK, pp. 196–228

Bazerman M H, Curhan J, Moore D, Valley K L 2000 Negotiations. *Annual Review of Psychology*. Annual Reviews, Palo Alto, CA

Bazerman M H, Neale M A 1992 *Negotiating Rationally*. Free Press, New York

Bell D E, Raiffa H, Tversky A 1988 *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Cambridge University Press, Cambridge, UK

Brandenburger A M, Nalebuff B J 1996 *Co-opetition*. Currency Doubleday, New York

Camerer C F 1997 Progress in behavioral game theory. *Journal of Economic Perspectives* **11**(4): 167–88

Harsanyi J C 1982 Subjective probability and the theory of games: Comments on Kadane and Larkey's paper. *Management Science* **28**(2): 120–4

Harsanyi J C, Selten R 1988 *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA

Hart O 1995 Firms, contracts, and financial structure. *Clarendon Lectures in Economics*. Oxford University Press, Oxford, UK
Kadane J B, Larkey P D 1982 Subjective probability and the theory of games. *Management Science* **28**(2): 113–20
Kagel J, Roth A E 1995 *The Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ
Keeney R, Raiffa H 1976 *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley, New York
Lax D A Sebenius J K 1981 Insecure contracts and resource development. *Public Policy* **29**: 417–36
Lax D A, Sebenius J K 1986 *The Manager as Negotiator: Bargaining for Cooperation and Competitive Gain*. Free Press, New York
Lax D A, Sebenius J K 1991 Thinking coalitionally: Party arithmetic, process opportunism, and strategic sequencing. In: Young H P (ed.) *Negotiation Analysis*. University of Michigan Press, Ann Arbor, MI, pp. 153–93
Lax D A, Sebenius J K 2000 Basic dealcrafting: Principles for creating value in negotiation (under review). Harvard Business School, Boston, MA
Luce R D, Raiffa H 1957 *Games and Decisions*. Wiley, New York
Mnookin R H, Peppet S R, Tulumello A S 2000 *Beyond Winning: Negotiating to Create Value in Deals and Disputes*. Harvard University Press, Cambridge, MA
Neale M A, Bazerman M H 1991 *Cognition and Rationality in Negotiation*. Free Press, New York
Putnam R D 1988 Diplomacy and domestic politics: The logic of two-level games. *International Organization* **4Z**: 427–60
Raiffa H 1982 *The Art and Science of Negotiation*. Belknap Press, Harvard University Press, Cambridge, MA
Raiffa H 1997 *Lectures on Negotiation Analysis*. PON Books, Cambridge, MA
Rothschild M, Stiglitz J E 1970 Increasing risk, I: A definition. *Journal of Economic Theory* **2**: 225–43
Rubinstein A 1991 Comments on the interpretation of game theory. *Econometrica* **59**: 909–24
Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA
Sebenius J K 1984 *Negotiating the Law of the Sea: Lessons in the Art and Science of Reaching Agreement*. Harvard University Press, Cambridge, MA
Sebenius J K 1991 Designing negotiations toward a new regime: The case of global warming. *International Security* **15** (4): 110–48
Sebenius J K 1992 Negotiation analysis: A characterization and review. *Management Science* **38**(1): 18–38
Sebenius J K 1996 Sequencing to build coalitions: With whom I should talk first? In: Zeckhauser R, Keeney R, Sebenius J K (eds.) *Wise Decisions*. Harvard Business School Press, Cambridge, MA, pp. 324–48
Sebenius J K 2000 Dealmaking essentials: Creating and claiming value for the long term. Harvard Business School Publishing, Boston, Item 2-800-443
Thompson L 2001 *The Mind and Heart of the Negotiator*, 2nd edn. Prentice-Hall, Upper Saddle River, NJ
Tirole J 1988 *The Theory of Industrial Organization*. MIT Press, Cambridge, MA
von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ
Walton R, McKersie R 1965 *A Behavioral Theory of Labor Negotiations*. McGraw-Hill, New York
Young H P (ed.) 1991 *Negotiation Analysis*. University of Michigan Press, Ann Arbor, MI
Zeckhauser R, Keeney R, Sebenius J K (eds.) 1996 *Wise Decisions*. Harvard Business School Press, Cambridge, MA

J. Sebenius

Neighborhood: General

The term neighborhood is rooted in the verb neighbor. 'To neighbor' has a double meaning: (a) to live in vicinity, one of another; (b) to be friendly, to render mutual favors or assistance (Webster 1983). This double meaning expresses the essence of the neighborhood: continuous physical proximity among people, together with some special attitudes, such as friendliness, and/or special behaviors, such as mutual assistance.

This article is intended to serve students and scholars of the social and behavioral sciences, as well as professional planners and designers who are interested in the social context of their work. It is divided into four sections: the first presents an analysis of neighborhood characteristics and neighborhood change by social scientists throughout the twentieth century; the second deals with the neighborhood and its design in the eyes of planners and architects in the same period; the third focuses on poor neighborhoods and the initiatives to regenerate them, in which considerable scholarly, professional and financial resources have been invested by almost every country; the last section discusses the role of neighborhoods in the society of the twenty-first century.

1. *The Neighborhood Concept in the Social Sciences*

Neighborhoods have been regarded by many scholars as a basic pattern of social organization, which is universal and timeless. Cooley (1909) argued that the neighborhood is 'a natural phenomenon' that exists wherever a group of people produces a dwelling place. Archeologists identified clusters of living that may be considered neighborhoods in the kingdoms of the ancient world. The ancient neighborhoods were arranged by family or clan lines, then later on by religious, ethnic, or professional lines. In modern times, the divisions are largely by cultural–ethnic origin and/or social status.

Neighborly relations among people who lived in one another's vicinity, including intimate acquaintance and characterized by many interactions, were the normal way of life in rural communities prior to the industrial era (Tonnies 1887/1957). Sociologists of the early twentieth century, especially those from the Chicago School, claimed that such *primary relations*

among neighbors are essential for maintaining the social order, because they create an effective informal control system (Cooley 1909, Park et al. 1925/1967). According to this view, isolation from the neighborhood portends an individual's alienation, and disintegration of neighborhoods threatens the social order (Durkheim 1893/1964).

The socioeconomic process of urbanization that followed the technological–economic process of industrialization changed people's way of life. The easy means of communication and transportation in the city, which enabled individuals to live at the same time in several social worlds, tended to destroy the permanency and intimacy of the neighborhood. Wirth (1938) argued that the size, density, and heterogeneity that characterize cities create changes that reduce primary relationships, including neighborly relationships, and increase loneliness and social deviance. Wirth represents the nostalgic yearning for healthy village life and the antiurban tone that was prevalent among writers of his generation.

Empirical sociological and anthropological studies of urban neighborhoods in the middle of the twentieth century portrayed a less pessimistic and more diverse urban picture (Glass 1948, Janowitz 1952, Jacobs 1961). Within large industrial cities the researchers found residential areas that were 'Gemeinschafts' in Tonnies's typology; the intimate spontaneous relationships between the neighbors in such areas were based on proximity as well as on kinship and/or ethnic identity and/or sharing a work place or other common denominator. Gans (1962) provided a vivid analysis of the life of Italian-Americans in a Boston neighborhood. He gave these people, who managed to preserve their special family life, peer groups, and communal institutions within the modern city of Boston, the name of 'urban villagers.' Suttles (1972) argued that in the modern city one can find very different types of neighborhoods, from the ones that are based on a form of organization that is almost a primary group, to the expanded communities of limited liability.

A special stream of research was devoted to the impact of *architectural design* on social attitudes and behavior in the neighborhood context. Researchers frequently found strong relationships between architectural design and social variables: proximity and planned paths seemed to influence close friendship (Festinger et al. 1950), and racially mixed housing seemed to reduce prejudice (Deutsch and Collins 1951). Large project size and height of buildings in turn increased the number of crime events (Newman 1972). Later critics claimed that these studies failed to control for important intermediate variables, especially those related to the sociocultural and the socioeconomic characteristics of the residents. A researcher especially aware of these variables is Keller (1968). She joined those who believed that the physical environment strongly influences the social relations within it, and concluded that architects are obliged to tailor the design according to the variety of cultural values and socioeconomic status of the users.

In contrast to the unidirectional view on the influence of the built environment on people, the school of *man–environment research*, consisting mainly of psychologists who cooperate with physical planners, emphasizes a two-way man–environment interaction. Unique to this group of researchers is their focus on personal/subjective definitions of neighborhoods. The various subjective definitions fall into two categories, physical and social, and neighborhoods are most clearly defined when social and physical space coincide, i.e., when physical boundaries, social networks, local facilities and special symbolic and emotional connotations are congruent in people's minds (Rappoport 1977).

In addition to their traditional social and psychological roles, urban neighborhoods function as administrative units for various ends, including political organization and the provision of social and commercial services. For census purposes, residential areas are divided into statistical units, the borders of which were frequently drawn by the boundaries of traditional neighborhoods. Social scientists have made intensive use of this statistical division, especially for developing empirical *social area analysis*, which concentrates on changes over time in residential areas.

In the second half of the twentieth century, scholars interested in urban processes focused much of their work on analyzing neighborhood change (Schwirian 1983).

The Chicago School of sociologists (Cressy 1938) used analogies from plant and animal ecology to invent the *invasion–succession model*, which describes processes of neighborhood population change. 'Invasion' illustrates the entrance of a different social group—racially, ethnically, or socioeconomically different—to a populated area, and the creation of competition and conflict within the invaded neighborhood that ends either by withdrawal of the invaders or by their triumph in the form of 'succession.' Taeuber and Taeuber (1965) used this model to produce a comprehensive study of racial neighborhood change in American cities. Several economists adopted the model and added a considerable amount of supporting empirical evidence, especially for explaining neighborhood deterioration (Grigsby et al. 1987).

Another model of neighborhood change—*the life-cycle model*—was suggested by Hoover and Vernon (1959), who identified five stages: development, transition, downgrading, thinning-out, and renewal. These stages are related to the quantity and status of the residents, the intensity and kind of land use, and the quality of housing. Not all neighborhoods go through each stage; some may skip one or stay indefinitely in another one.

The political economy school offers a different approach. They argue that the fate of a neighborhood is determined by the interests and the complex relation-

ships among the political and economic public institutions, and strong actors in the business market (banks, developers, and others) (Logan and Molotch 1987). Most of these forces play from outside the borders of the neighborhood, and hence the ability of local people to influence them is very limited.

In contrast to the structural–functional approaches that emphasize external-to-the-neighborhood social, economic and organizational forces, the recent *theory of the social movements* focuses on internal actors, on the individual human agent and his/her search for identity and meaning. In the context of neighborhoods, the researchers of the social movements are interested in organizations of local residents. The residents organize to fight for maximization of their quality of life, sometimes even when it is not congruent with maximization of economic profits. They may find identity and meaning by joining neighborhood movements and promoting bottom-up change, frequently challenging top–down planning by politicians, economists and designers (Hasson 1993).

2. *Models of Neighborhoods in the Eyes of Planners and Architects*

Modern urban planning took shape in the last quarter of the nineteenth century, against the background of the socioeconomic movements of industrialization and urbanization. Large populations of workers and their families lived in extremely poor and crowded conditions in the industrial cities. The first planned neighborhoods in Europe were workers' neighborhoods, built by owners of industrial plants or by philanthropic organizations.

All the early city planning initiatives—the City Beautiful, the City Efficient and the Garden City—sought to make the city and its neighborhoods safe, sanitary, economically efficient, and socially attractive. The leading idea was to use physical planning to create a socially better world. This central motive accompanied the efforts of architects and planners throughout the twentieth century. Several critics say that this is clear evidence that the road to hell is paved with good intentions.

The prominent figure at the starting point is Ebenezer Howard. His small book, *Garden City of Tomorrow* (1902/1985), influenced twentieth-century urban planning more than any other publication. He advocated an urban environment that combines the advantages of the city, such as modern occupations and services, with those of the village—human scale and many local social contacts. His model covered the entire scale from the dwelling through the neighborhood to the region. Howard's ideas had a decisive impact on the New Towns movement, which gave rise in the middle of the twentieth century to 30 new towns in Great Britain, a similar number in Israel, several *villes nouvelles* around Paris and Stockholm, as well as to a few new towns in the USA. Within each of these, some version of Howard's plan for the 'ward' (his version of a neighborhood) was implemented. However, none of the many towns and neighborhoods that followed Howard's physical suggestions (at least partly) followed his social vision. As Hall (1988, p. 87) put it, 'they see him as a physical planner, ignoring the fact that his garden cities were merely the vehicles for a progressive reconstruction of capitalist society into an infinity of cooperative commonwealths.'

In the USA, a rather similar (physically) model for a neighborhood was suggested by Clarence Perry (1929). The size of his *neighborhood unit* was to be set by the catchment area of the local elementary school that functions as its focus. The unit was to serve as a building-block of the modern city. Like the English precedent, the American model included prescriptions for residences, streets, open spaces and a range of services, all designed to serve a social vision. But the American version was far less revolutionary than the English one; it did not aspire to change the social order, just to make it attractive and friendly.

A strong supporter of the concept of the neighborhood unit was the journalist–sociologist Lewis Mumford (1954). He saw neighborhood planning as an opportunity and a leverage for fostering feelings of belonging and for supporting amiable behavior among neighbors. He contributed to the idealization of the neighborhood as a framework for preserving positive social values and institutions. This idealization was among the factors that convinced the American Public Health Association to adopt the neighborhood unit model as the basis for planning residential environments. Subsequently, it was adopted by other professional organizations and public agencies and guided much of the post-World War II construction of middle-class suburbs. All this happened in spite of the severe criticism of the model that had been prevalent since the 1940s. Social scientists disagreed with its premises and planners questioned its unintended consequences (Banerjee and Baer 1984). Even though theoretically the model called for social mix and *social integration*, in practice it encouraged segregation. An example is the early settlement in the neighborhoods of the new town of Radburn (New Jersey, USA); in 1934 there were no blue-collar workers among the residents, and the realtors kept out Jews and blacks as well (Schaffer 1982).

Back in Europe, what was planned and built in the middle of the century was highly influenced by the architects of CIAM (*Les Congres Internationaux d'Architecture Modern*), an organization that was established in 1928 and existed for 30 years (Kallus and Law-Yone 1997). These architects were deeply influenced by rigid socialistic concepts on the one hand and by new technical options opened through the industrial advancement on the other. They created the international style of architecture with the prin-

ciples of minimalism (planning a house by the minimal objective needs of health and convenience of the residents—*existensminimum*), standardization (disregarding differences of culture, climate, etc.) and separation of land uses. The most famous architect who promoted these concepts was Le Corbusier. Thousands of neighborhoods, which were built in Europe following World War II were influenced by these principles. The only problem was that people disliked the minimal and standard blocks that the architects built for them. Better-off households left them as soon as they could afford it. The unattractive design, together with severe maintenance problems, played a central role in the processes of social and physical deterioration and segregation that affected much postwar residential construction, mainly in Europe.

A recent planning trend, which opposes the standardization of international architecture as well as the typical design of neighborhood suburbs in the USA, is the *New Urbanism*. It developed in the 1980s as an umbrella term that encompasses two design schemes: the TND (Traditional Neighborhood Development) and the TOD (Transit-Oriented Development). The common principles of both are: small (compact) size that creates walkable distances (typically no more than a quarter of a mile from center to edge); clearly defined edges and centers; mixed land uses (residences, shops, schools, workplaces) and mixed 'market segments'; and priority given to public buildings and spaces (Katz 1994). In some magic way, this design is expected not only to create a distinct physical unit but also some social identity—a community. Underneath the detailed design instructions of the New Urbanism lies the same old belief of architects that by means of a proper design of the physical environment we can create a better social world.

In addition to a common belief in *physical determinism*, most of the neighborhood models suggested by architects are characterized by disregard of the users' voice. For Le Corbusier this was obvious; he argued that planning is too important to allow ignorant people be involved in it. Other architects continued the paternalistic approach of the professionals who preceded them, without considering issues of *residents' participation*. One of the exceptions was the English architect John Turner (1976), the father of 'housing as a verb' rather than a noun. He argued that when dwellers control the major decisions concerning the design, construction, and management of their home and neighborhood, both the process and the environment produced stimulate individual and social wellbeing. However, Turner based his approach on his experience with communities in which there was some form of commitment and partnership between the members, mostly traditional communities. Is residents' participation in designing and managing the built environment applicable also to regular neighborhoods in urban Western societies?

A positive answer to this question evolved within a new breed of planners that emerged on both sides of the Atlantic in the middle of the twentieth century. In the new discipline of Urban and Regional Planning, scholars and practitioners of the social sciences displaced architects as leaders. Many of them, especially those who organized under the umbrella of 'progressive planning,' supported cultivation of community networks, listening carefully to people and involving the less organized groups, notably those in disadvantaged neighborhoods (Rohe and Gates 1985, Carmon 1990, especially articles by Bratt and by Fainstein). The planners contributed to a process that made residents' involvement in decision-making regarding their homes and neighborhoods a common practice, at least in cases of neighborhood renewal.

3. *Neighborhood Regeneration as a Focus of Research and Planning Initiatives*

The scholarly studies of neighborhoods by social scientists and the noble intentions of physical planners did not prevent the development of one of the most severe social problems of the twentieth century—the concentration of multifaceted distress, including poverty and crime, in certain urban neighborhoods. Some parts of the problem are residues of the extremely poor residential areas of the nineteenth century, but the main part is the fruit of socioeconomic changes and planning failures of the last century.

In all the developed and many of the developing countries there were deliberate efforts to get rid of the poorest slum areas, especially where they were in the heart of cities, and/or to improve the living conditions of people in distressed urban neighborhoods. Below is a review of the history of such interventions and the impact of social studies on policy changes. Usually, this history is presented country by country. Here the focus is on the shared experience. The similarities are attributed to resemblance in socioeconomic and sociopolitical developments in Western countries after World War II, and in addition, to international transfer of knowledge among social scholars who attained a significant influence on decision-making with regard to neighborhood policies. Three successive 'generations' of neighborhood policies are identified, which occurred at about the same time in Western countries on both sides of the Atlantic (Carmon 1997).

3.1 *First Generation: The Era of the Bulldozer*

This phase was characterized by physical determinism and emphasis on the built environment.

Intolerable housing conditions in old and very old neighborhoods in the growing cities, coupled with the wish to make 'better use' of central urban land and

drive the poor out of sight, gave birth to the idea of *slum clearance*. Large-scale clearance programs were implemented in the UK (1930–39) (1954–1970s), the USA (1937–64), Canada (1948–68), France (1958–75), and many other countries. In the European states, the public authorities managed both the clearance and the provision of housing to the relocatees. The result was that most of the displaced households were rehoused. In the USA, by contrast, concentration and clearance of land sites was generally done by public agencies, while the new construction was in the hands of private entrepreneurs. As a consequence, the number of apartments demolished under the aegis of the urban renewal programs in the USA was much greater than the number of units built. The slum areas were frequently replaced by shopping centers, office buildings, and cultural and entertainment centers, all of which were in high demand in the boom years that followed World War II. The few housing developments built were generally designated for people with higher socioeconomic status than those who were relocated. Gans (1965) found that between 1949 and 1964 only one half of one percent of all expenditures by the American federal government for urban renewal was spent on relocation of the families and individuals removed from their homes.

Despite the significant differences in the nature of activity between countries, the criticism voiced against most of the clearance and redevelopment projects was similar (Wilmott and Young 1957). The executors were criticized for ignoring the heavy psychological cost of forced relocation, the social cost of the destruction of healthy neighborhoods and communities, and the cultural cost of erasing old urban fabrics. In many cases where new residential neighborhoods were built, the planners and designers were blamed for building inhuman blocks which were unfit for family life, and certainly not suitable for poor families. The bulldozer approach as a leading regeneration strategy was condemned and disqualified wherever it was applied.

3.2 Second Generation: Neighborhood Rehabilitation

This phase saw a more comprehensive approach to neighborhood renewal; one that emphasized social considerations.

A salient expression of the great change between the first and second generation of neighborhood remedies is found in the Model Cities program in the USA (middle 1960s). This was a series of comprehensive rehabilitation programs, aimed at improving the existing housing and its environments (instead of demolishing them), while simultaneously, treating the social problems of the population by adding social services and bettering their quality. Many of the new programs attempted to involve local residents in their decision-making processes and made 'maximum feasible participation' a leading slogan of the period.

Similar comprehensive neighborhood programs were implemented in Canada—the Neighborhood Improvement Program in 322 local authorities, starting 1973, France—Neighborhood Social Development in 150 neighborhoods, starting 1981, and Israel—Project Renewal in 150 neighborhoods, starting 1977. In Britain, as in Sweden, the Netherlands, and West Germany, the activities for housing improvements were not integrated into the same programs as plans for improving the social services, but in many cases, the two kinds of new initiatives were implemented simultaneously in the same neighborhoods (see Alterman and Cars 1991).

Despite the abundance of goodwill and the large sums expended in the various countries, evaluation studies of these programs with a social emphasis showed frequently negative and at best mixed results. Several studies found that the programs had not been implemented and the money had been spent on issues and sites that were not part of the official plans. Where implemented, the programs frequently benefited people (local residents) but not their neighborhoods, i.e., they could not change the negative image and the low status of the target neighborhoods (Carmon and Baron 1994). This might have been expected: the researchers could have known that the status of a neighborhood is dependent on the socioeconomic status of its residents much more than on the level of housing and social services. Hence, programs that work with existing low-status populations and prevent the entry of 'strangers' of somewhat higher socioeconomic rank, can do a lot to better the situation of their clients but can hardly improve the image of their neighborhoods.

3.3 Third Generation: Revitalization, Especially in City Centers

This phase is characterized by a businesslike approach that emphasized economic development. The slowdown of the economy in the 1970s and the criticism of the large public programs of the second generation influenced neighborhood policies: governments retreated from deep involvement and deteriorated areas were open to new initiatives. These took two main forms: public–individual partnerships and public–private partnerships. Both emphasized economic investments and profits.

Public–private partnerships are usually large real estate projects, in which private developers receive special sanctions from the local authority in exchange of some designated contribution to 'the public interest.' Among the famous ones are London Docklands in the UK and Quincy Market in Boston, USA. Many of these projects took place within or in the vicinity of distressed neighborhoods. Nevertheless, incumbent

residents seldom benefited from them and more often were hurt, directly or indirectly (Smith 1996).

Within the group of *public–individual partnerships*, the most salient phenomenon is gentrification, the entrance of upper middle-class residents into low-income neighborhoods, usually in the vicinity of vibrant central business districts of cities. In spite of the frequently found and often denounced consequence of displacement of poor veteran residents, gentrification processes have been encouraged by city councils in Europe and America, especially through enabling regulations and tax deductions. Another process in this group is housing and neighborhood upgrading by incumbent residents. Local people invest their own resources in improving their living environment, and often manage to receive at least some assistance from voluntary and public bodies. Last but not least is neighborhood upgrading by immigrants. Waves of new urban and educated immigrants have been settling in deteriorated urban areas in the developed countries, where they could find cheap dwellings, and gradually they have contributed to revitalization of these areas. This is a rather new phenomenon, which may be growing with the expected increase in international migration in the coming years. What is common to the three processes is that they usually start as spontaneous private investments, which later receive some support from local public agencies. Together they are changing parts of our old cities and neighborhoods, but at the end of the twentieth century, they are not much more than islands of revitalization within seas of decay.

4. *Looking into the Future: Neighborhoods in the Twenty-first Century*

This article began with the statement that neighborhoods have been part of the human mode of life since the dawn of civilization. The ancient Arab proverb 'better a close neighbor than a remote brother' illustrates the important social functions that neighborhoods and neighbors had in traditional societies. With the transition to the industrial era and the development of communication technologies, people became much more mobile, their world widened considerably, and they became less dependent on physical proximity and on constant relationships. The decreasing importance of distance has been enhanced with the development of telecommunication in the postindustrial era. Do these changes mean that the time of neighborhoods as territories that combine physical and social closeness is over?

Our answer is that neighborhoods will continue to exist in the twenty-first century and to fulfil important social functions for some of the people at least some of the time. In general, those who will tend to be attached to neighborhoods are people at both ends of the life-cycle—children and the elderly—and at both ends of the social ladder—the disadvantaged and the well-to-do. This answer is not based on a belief in some moral commitment to neighbors and to members of the surrounding community, which is one of the principles of the current social movement of communitarianism (Etzioni 1994), but rather on analysis of social reality.

Analysis shows that housing proximity is associated, or has the potential of being associated, with the existence of important joint interests. This is very clear among minorities and immigrant groups, which tend to cluster geographically and create urban enclaves; in their enclaves they can support one another, defend their specific interests and, when they feel necessary, organize to fight contradicting interests (Abrahamson 1996). This is also common among the most exclusive neighborhoods of affluent citizens who create gated communities (Blakely and Snyder 1997). Mobile middle-class households may not be inclined to be part of neighborhoods or communities, but where and when they feel a need to promote a local interest, and especially when change in the social or physical environment is perceived as a threat to the status of their home, they too function as a community of interests on a local basis.

In the twenty-first century, human beings will continue to have emotional needs in addition to materialistic needs, along with the need to have close personal contact with others. For people whose mobility is limited, either because modern communication technologies have not reached their environment, or because they are too young, too old, or too poor to use the technologies, the neighborhood will continue to be a significant source for fulfilling functional as well as social and emotional needs. As for the more privileged people, the new technologies enable them to choose and also to change their choices: to belong to a neighborhood/community only, or to a virtual community only, or to both, or to none; to make certain choices in adolescence and change them once or more than once as an adult or an old person.

Hence, neighborhoods will continue to exist and researchers will continue studying them. One may expect future research into the dual face of neighborhoods/communities, as integrating units from the point of view of their residents and as separating forces from the point of view of the city. Another fruitful direction of research is related to the role of neighborhoods in the development of civic societies, and to the impact of new communication technologies on this and other neighborhood roles. Finally, in a postindustrial era with increasing disparities between the haves and the have-nots, social studies may try not only to understand neighborhoods but also to contribute their share in the efforts to improve the life of disadvantaged people who live in them.

See also: Community, Social Contexts of; Community Sociology; Community Studies: Anthropological; Neighborhood Revitalization and Community De-

velopment; Residential Concentration/Segregation, Demographic Effects of; Residential Segregation: Geographic Aspects; Urban Geography; Urban Poverty in Neighborhoods; Urban Sociology

Bibliography

Abrahamson M 1996 *Urban Enclaves: Identity and Place in America*. St. Martin's Press, New York

Alterman R, Cars G (eds.) 1991 *Neighborhood Regeneration: An International Evaluation*. London, Mansell

Banerjee T, Baer W C 1984 *Beyond the Neighborhood Unit*. Plenum Press, New York

Blakely E J, Snyder M G 1997 *Fortress America: Gated Communities in the US*. Brookings Institute Press, Washington, DC

Carmon N (ed.) 1990 *Neighborhood Policy and Programmes: Past and Present*. Macmillan, London

Carmon N 1997 Neighborhood regeneration: The state of the art. *Journal of Planning Education and Research* **17**: 131–44

Carmon N, Baron M 1994 Reducing inequality by means of neighborhood rehabilitation. *Urban Studies* **31**: 1465–79

Cooley C H 1909 *Social Organization*. Scribner and Sons, New York

Cressy P 1938 Population succession in Chicago: 1898–1930. *The American Journal of Sociology* **44**: 59–69

Deutsch M, Collins M E 1951 *Interracial Housing*. University of Minnesota Press, Minneapolis, MN

Durkheim E 1893/1964 *The Division of Labor in Society*. Free Press, New York

Etzioni A 1994 *The Spirit of Community: The Reinvention of American Society*. Touchstone, New York

Festinger L, Schachter S, Back K 1950 *Social Pressure in Informal Groups: A Study of Human Factors in Housing*. Harper, New York

Gans H 1962 *The Urban Villagers: Group and Class in the Life of Italian Americans*. Free Press, New York

Gans H J 1965 The failure of urban renewal: A critique and some proposals. *Commentary* **2**: 29–37

Glass R 1948 *The Social Background of a Plan: A Study of Middlesbrough*. Routledge & Kegan Paul, London

Grigsby W, Baratz M, Galster G, Maclennan D 1987 The dynamics of neighborhood change and decline. *Progress in Planning* **28**: 1–76

Hall P 1988 *Cities of Tomorrow*. Basil Blackwell, Oxford

Hasson S 1993 *Urban Social Movements: The Protest of the Second Generation*. SUNY Press, Albany, NY

Hoover E M, Vernon R 1959 *Anatomy of a Metropolis*. Harvard University Press, Cambridge, MA

Howard E 1902/1985 *Garden Cities of Tomorrow*. Attic Books, London

Jacobs J 1961 *The Death and Life of Great American Cities*. Random House, New York

Janowitz M 1952 *The Community Press in an Urban Setting*. Free Press, Glencoe, IL

Kallus R, Law-Yone H 1997 Neighborhood—the metamorphosis of an idea. *Journal of Architectural and Planning Research* **14**: 107–24

Katz P 1994 *The New Urbanism: Towards and Architecture of Community*. McGraw-Hill, New York

Keller S 1968 *The Urban Neighborhood: A Sociological Perspective*. Random House, New York

Logan J, Molotch H 1987 *Urban Fortunes*. University of California Press, Berkeley, CA

Mumford L 1954 The neighborhood and the neighborhood unit. *Town Planning Review* **24**: 250–70

Newman O 1972 *Defensible Space: Crime Prevention through Urban Design*. Collier Books, New York

Park R E, Burgess E W, McKenzie R 1925/1967 *The City*. University of Chicago Press, Chicago

Perry C A 1929 *The Neighborhood Unit: A Scheme of Arrangement for the Family Life Community*. Regional Plan of New York and its Environs, New York

Rappoport A 1977 *Human Aspect of Urban Form: Towards a Man–Environment Approach in Urban Form and Design*. Pergamon Press, Oxford

Rohe W M, Gates L B 1985 *Planning with Neighborhoods*. University of North Carolina Press, Chapel Hill, NC

Schaffer D 1982 *Garden Cities for America: The Radburn Experience*. Temple UP, Philadelphia

Schwirian K P 1983 Models of neighborhoods. *Annual Review of Sociology* **9**: 83–102

Smith N 1996 *The New Urban Frontier: Gentrification and the Revanchist City*. Routledge, London

Suttle G 1972 *The Social Construction of Communities*. University of Chicago Press, Chicago

Taeuber K E, Taeuber A R 1965 *Negroes in Cities*. Aldine Press, Chicago

Tonnies F 1887/1957 *Community and Society*. Michigan State University Press, East Lansing, MI

Turner J F C 1976 *Housing by People: Towards Autonomy in Environments*. Pantheon Books, New York

Webster 1983 *Webster's New Universal Unabridged Dictionary*, 2nd edn. Dorset and Baber. Simon and Schuster, Cleveland, Ohio

Wilmott P, Young M 1957 *Family and Kinship in East London*. Routledge and Kegan Paul, London

Wirth L 1938 Urbanism as a way of life. *American Journal of Sociology* **44**: 1–24

N. Carmon

Neighborhood Revitalization and Community Development

1. Introduction

Neighborhood revitalization efforts and programs, both public and private, will be reviewed. These include early efforts at elimination of slums led by settlement houses, housing reformers, and social workers. This was followed by government programs like housing and health codes, public housing, urban renewal (both slum clearance and rehabilitation), model cities, neighborhood planning and redevelopment, and, most recently, the empowerment and enterprise zone program. Various techniques of neighborhood revitalization will be reviewed, as well as critiques of their impact. Likewise, community development programs, both public and private, and the role of community development corporations will be reviewed. The primary emphasis will be on the United States.

2. The Progressive Era: Tenement Reform and Settlement Houses

From the mid-nineteenth century through the beginning of the First World War, the population of US cities swelled as a result of mass immigration, primarily from Europe. Most poor immigrants were concentrated in ethnic enclaves. In an era in which there were no governmental social and welfare services other than the poorhouse, poor immigrants lived in overcrowded and unsanitary housing and neighborhood conditions. In cities like New York, they lived in tenements in wretched conditions.

Around the turn of the century, the Progressive era saw the rise of reform movements. These included the efforts of housing reformers to address the conditions prevalent in slum neighborhoods. New York City was in the forefront, led by a Tenement House Committee of reformers like Lawrence Veiller. The journalist Jacob Riis exposed slum conditions in his graphic descriptions and photographs of 'how the other half lives.' This led to the enactment of the New York Tenement House Law of 1901, designed to eliminate the worst abuses. It became a model for similar legislation in other states. However, reformers like Veiller were loath to invoke the government and were especially opposed to government-subsidized housing for the poor. Instead, they eventually turned to the adoption of municipal housing and health codes, setting minimum regulatory standards for residential housing (Friedman 1968). A few philanthropists sponsored limited dividend model housing for workers but this failed to prove to be a viable alternative to the housing provided by private landlords (Birch and Gardner 1981).

During this same period, a settlement house movement began, based upon similar previous efforts in England. Reformers supported by private philanthropy sought to alleviate social problems in slum neighborhoods. They established settlement houses to provide social services to the poor. Largely apolitical, they sought to uplift the poor, while trying to improve their living conditions. Leaders included Jane Addams in Chicago and Robert Woods in Boston (Fisher 1984, Halpern 1995).

3. Slum Clearance: Public Housing

It was not until the Great Depression of the 1930s that public housing appeared in the US, decades after its adoption in Europe. The Depression exacerbated already existing problems in poor urban neighborhoods. The reforms of the Progressive Era had done little to alleviate the underlying causes of the slums and substandard housing. A new generation of activist housing reformers argued for intervention of the federal government to replace slum housing with government-financed, owned and operated public housing. This was a radical departure from US policy.

Reluctantly, the New Deal administration of President Franklin D. Roosevelt began a program of building public housing as part of a public works slum clearance program aimed at provided work to unemployed workers in the building trades. The real estate industry fought hard against the creation of public housing, arguing instead that private landlords should provide for the housing needs of low-income persons. The 1937 legislation provided federal financial assistance to those local governments which chose to participate in a program to provide decent, affordable housing for the poor. Local public housing authorities were to construct and operate the public housing projects. The real estate lobby was successful in limiting public housing to the poor and imposing limits on the amounts allowed for public housing units. In addition, public housing could only be built to replace substandard units. These limitations would later prove fateful in undermining the goals of the reformers and dooming the expansion of public housing. Despite the attempts of a few localities, racial segregation of public housing prevailed, reflecting racist attitudes and existing housing and neighborhood patterns.

The public housing program began slowly, was interrupted by the emergency of World War II, and barely survived conservative opposition in the housing act of 1949. It was adopted primarily by large cities with a preponderance of the poor, which received most of the funding and engaged in the construction of large-scale public housing projects. Even with its limitations, public housing was a great improvement over the slum housing that it replaced. In its early stages, public housing served as housing for the working poor.

However, as the public housing grew older, its population in the 1960s became poorer (Meehan 1979). They lived in projects that were often physically isolated, in poor condition, and racially segregated—despite increased federal subsidies and court mandates. Except for housing for the elderly poor, new construction of public housing virtually ended in the 1970s. Instead, the federal government in the US switched to providing subsidies to enable poor tenants to rent private rental units. In the 1990s, some of the most notoriously unsuccessful public housing projects were either demolished or reduced in scale. Despite the bad reputation of several examples nationally, for the most part public housing in urban areas did contribute to the elimination of the worst slum conditions.

4. Slum Clearance: Urban Renewal

Planning for postwar redevelopment in US cities began in the early 1940s. Liberals favored large-scale slum clearance, replacing them largely with public housing (Gelfand 1975). In 1949, Congress approved an urban redevelopment program which would eliminate slums and 'blight.' However, as it unfolded, prime downtown real estate became the focus of

urban renewal in many cities. Instead of public housing, this land was re-used for commercial and industrial facilities, luxury housing, and institutional uses. Despite assurances of relocation and replacement housing, hundreds of thousands of poor slum dwellers saw their neighborhoods and housing destroyed but their housing conditions worsened instead of improving as little new low-rent public housing was constructed on urban redevelopment sites (Greer 1965). New York City under the leadership of Robert Moses provided some of the most egregious examples of this process, although Moses also built many of the best public housing projects in the US (Caro 1974).

The program was renamed 'urban renewal' in 1954. The newly elected Republican administration of President Dwight Eisenhower, while not opposed to urban renewal, was concerned about its cost to the federal government and the possibility that much of the cleared land was not proving to be attractive to private, for-profit developers. It tried to change the focus of urban renewal in part to conservation and rehabilitation, which it saw as less expensive. Urban renewal clearance and redevelopment policies would later be condemned as bad planning by critic Jacobs (1961) in a famous critique.

In the 1930s, the realtors opposed federal intervention in the housing market and proposed neighborhood associations as the primary vehicle for neighborhood conservation. This was aimed primarily at homeowners' associations. This often had an exclusionary aim, with white homeowners organizing to prevent any entry by African Americans and other minorities. This is also the era in which the newly created Federal Housing Administration (FHA) adopted the racist standards of the real estate industry to rate neighborhoods for the purpose of making and insuring home mortgage loans. This became the basis for the 'redlining' of older urban neighborhoods and contributed to the exodus from these neighborhoods to the suburbs after World War II.

Many urban neighborhoods, mostly poor, also were destroyed to make way for central city links to the interstate highway system created in 1956. As with urban renewal, promises of decent, affordable replacement housing often went unfulfilled. Not all neighborhoods designated as slums for urban renewal or highway construction were necessarily blighted (Gans 1962). Where residents did organize to oppose the demolition of their neighborhood, typically they failed because they did not have political influence and had no legal standing in court to challenge these plans.

This finally changed in the late 1960s, following summers of riots in US cities from 1965 through 1968. Citizen participation finally was recognized and neighborhood residents at least had an opportunity to present their views. Advocacy planning arose to provide neighborhood residents with the capacity to develop alternate plans. The courts finally recognized the right of those to be displaced to challenge plans, typically on the ground that there was insufficient replacement housing available to meet federal requirements (Hartman 1974). Community resistance led to the demise of large-scale clearance for urban renewal and highways in the central cities by the early 1970s.

5. *Model Cities*

It was felt that displacement of poor and minority residents for these purposes contributed to the resentment that sparked so many of the urban riots of the 1960s. Resident participation in planning led to more serious consideration of rehabilitation of blighted neighborhoods rather than clearance and redevelopment. The first major example of this policy shift was embodied in the Model Cities program launched by a newly created US Department of Housing and Urban Development (HUD) in 1967. In contrast to urban renewal slum clearance, this reaction to the urban riots proposed to concentrate federal, state, and local resources to improve targeted poor neighborhoods. Housing rehabilitation, social services, job training, and community participation were hallmarks of this approach. Due to insufficient funds, intergovernmental coordination problems, and a change from the Johnson to Nixon administrations after the 1968 presidential election, Model Cities failed to fulfill its promises (Haar 1975).

6. *Community Development Block Grants*

In 1974, the urban renewal and Model Cities programs, among others, were ended and replaced by the Community Development Block Grant (CDBG) program. The CDBG program was intended to give local governments more discretionary control over spending and to involve more communities (Hays 1995). However, CDBG funds were supposed to benefit primarily low- and moderate-income residents. This caused controversy in many localities and often led to wider dispersal of funds, to the detriment of those neighborhoods most in need of assistance (Rich 1993). This era marks a shift in federal policy that persisted through the Clinton administration, emphasizing devolution of responsibility for many federal aid programs and resulting in cutbacks in many HUD housing assistance programs.

7. *Federal Recognition of Neighborhoods*

Federal policy since the 1930s did not usually recognize neighborhoods. HUD and its predecessor agencies made grants and loans to cities and states, developers and lenders, and provided insurance to individuals (homebuyers). In 1977, President Jimmy Carter made the unprecedented move of appointing

an Assistant HUD Secretary for Neighborhoods and created a small program which for the first time awarded grants directly to neighborhood development organizations (Mayer 1984). In 1979, Carter appointed a presidential commission to advise him on neighborhood policies and programs but then disregarded its recommendations. His successor President Ronald Reagan immediately dismantled Carter's modest neighborhood initiatives at HUD.

8. *Redlining and Community Reinvestment*

One of the major reasons for the decline of older urban neighborhoods was their 'redlining.' Lenders, insurers, and realtors avoided those neighborhoods which had a significant minority population on the basis that their presence jeopardized real estate values and would cause white flight. Federal agencies, most notably the FHA, followed suit. This in turn almost assured that eventually these neighborhoods would become all or predominantly minority-occupied. 'White flight,' triggered by unscrupulous real estate practices, then occurred in many neighborhoods (Jackson 1985).

Chicago, where such segregationist practices as 'racial covenants' originated and long the most racially segregated city and metropolitan area in the US, was the focus of organizing in the 1970s against lender disinvestment in urban neighborhoods. Neighborhood organizers demanded that lenders cease redlining and invest in neighborhoods. Racial discrimination in housing and lending had been previously outlawed, although enforcement of this legislation has yet to end such practices. The antiredlining movement spread and eventually led to the enactment of federal legislation in 1975 and 1977 requiring federally regulated lenders first to disclose generally where their loans were made and then to serve all of the communities within their territory. The US Congress rejected more restrictive proposals. The movement then used this legislation to document lender discrimination against older neighborhoods and to require lender commitments to again lend in these neighborhoods. This has led to major lender commitments in many cities (Squires 1992, Schwartz 1998). This is a critical component of neighborhood revitalization, especially when the federal government was reducing its role in aiding urban areas.

9. *Gentrification of Urban Neighborhoods*

Beginning in the late 1970s, a contrasting phenomenon to disinvestment began. In many large cities, middle- and upper-income newcomers began to discover and move into older neighborhoods in decline near downtowns (Palen and London 1984). These 'urban pioneers' renovated buildings which were typically in disrepair. As their numbers increased, these neighborhoods saw other changes, particularly in the quality of services, both private and public, reflecting the demands of the affluent newcomers.

City governments viewed this process as benign, with improvements in the physical quality of neighborhoods and increased tax revenue. However, in many of these neighborhoods, lower-income residents were displaced, directly or indirectly, because they were relocated for building renovation or could not afford rehabilitated housing. There was also often a clash of lifestyles between the residents and the newcomers. While the magnitude of gentrification-based displacement was debated (Schill and Nathan 1983), many localities did institute land use and housing regulations to provide some protection to residents against displacement and to ensure the replacement of lower-priced housing units (Keating and Smith 1996).

10. *The Emergence of Community Development Corporations*

In the 1960s, the US Congress funded directly community development corporations (CDCs). They are nonprofit community-based organizations operating in distressed, mostly urban and central city neighborhoods. Many originally derived from community organizing. Many are church-sponsored. They attempt to address both social and physical problems. They have expanded greatly in number, totaling several thousand by the 1990s (Vidal 1992). In larger cities, there are now citywide networks of CDCs.

CDCs engage in a variety of activities, typically housing and commercial development, job training, and, to a lesser degree, social services. They are funded from a wide variety of sources, including all levels of government, philanthropic foundations, and corporations (Vidal 1992). The impact of CDC activities has been rated positively (e.g., Rohe 1998, Walker 1993).

However, it must be conceded that, despite the often heroic efforts of CDCs and companion social service agencies, measurable levels of distress such as poverty and crime have increased rather than decreased in many urban neighborhoods since the advent of CDCs (Rusk 1999). CDCs are simply not enabled or funded to deal with the entire myriad of social and physical problems in distressed urban neighborhoods. Ultimately, governments in partnership with non-profit organizations, residents, and so-called 'intermediary' funders are responsible for the revitalization of neighborhoods.

11. *Neighborhood Revitalization Techniques*

A variety of techniques has been used in the revitalization of neighborhoods. Resident attitudes toward their neighborhood have been surveyed (e.g.,

Ahlbrandt and Cunningham 1979). Neighborhoods have been classified according to their physical and social characteristics (Keating and Smith 1996). Public programs have often targeted neighborhoods, based upon their profile and degree of need for assistance. Housing has been newly constructed and rehabilitated, both for lower- and moderate-to-middle-income range occupants (Rohe et al. 1998, Walker 1993). The low-income housing tax credit, which enables private corporations to profitably invest, is a critical factor in the provision of much affordable housing. In badly deteriorated neighborhoods, an urban homesteading approach has been tried to encourage self-help rehabilitation (Varady 1986). Existing and new businesses have been subsidized to improve their buildings, expand their operations, and assist in the training and employment of neighborhood residents. Neighborhood plans have been developed, with and without resident participation, to identify serious problems and programs and policies to alleviate them (Rohe and Gates 1985). A favored technique is the mapping of neighborhood assets as cornerstones for neighborhood revitalization. Neighborhood organizing remains a key ingredient in successful neighborhood redevelopment, with block clubs often being the cornerstone of neighborhood organizations (Medoff and Sklar 1994). Neighborhood organizing can be confrontational or cooperative, depending upon the circumstances and the style of organizing (Fisher 1984).

12. *The People vs. Place Policy Debate*

A major debate in neighborhood revitalization has been termed people vs. place. In 1968, the national commission on urban riots posed the dilemma of whether to rebuild poor ghetto areas or to try to disperse and integrate their mostly minority residents into mainstream society. Three decades later, this debate continues. Downs (1994) has posed alternate strategies aimed either at improving the social fabric and economic and physical conditions of impoverished inner city neighborhoods or providing mobility in work and housing to enable their residents to move to better neighborhoods in central cities and the surrounding suburbs. While there have been a few small-scale mobility experiments, more emphasis has been placed on the revitalization of inner city neighborhoods with the goal of their improvement for the existing residents.

The latest example in the United States of the latter approach is embodied in the Empowerment Zone and Enterprise Community program launched in 1994. It is a belated response to the 1992 civil disturbance in South Central Los Angeles. Using a combination of employer tax credits, subsidized loans for businesses, and special funding for social services, including job training, the aim is to provide employment primarily in designated high poverty areas in more than one hundred communities (Gale 1996).

13. *Neighborhood Revitalization in Canada and Western Europe*

Urban revitalization and regeneration programs in Canada and Western Europe are similar to approaches in the US described above. However, the context is much different. Governmental systems typically feature national and regional policies, programs, and agencies. Municipal authorities may have less authority than in the US. With few exceptions, the levels of poverty and distress do not compare with the US, with the exception of industrial cities in sharp decline. Finally, there has not been nearly as much racial conflict in the cities of Canada and Western Europe as in the US, again with some exceptions. For a comparative perspective on urban regeneration in North America and Europe, see Judd and Parkinson (1990). For a comparative perspective on neighborhood revitalization in Australia, Canada, and Western Europe, see Gale (1984).

See also: Community Aesthetics; Community Economic Development; Development and Urbanization; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Strategic Planning; Urban Growth Models; Urban Life and Health

Bibliography

Ahlbrandt R S, Cunningham J V 1979 *A New Public Policy for Neighborhood Preservation*. Praeger, New York

Birch E L, Gardner D S 1981 The seven percent solution: a review of philanthropic housing, 1870–1910. *Journal of Urban History* **7**: 403–38

Caro R A 1974 *The Power Broker: Robert Moses and the Fall of New York*. Alfred A. Knopf, New York

Downs A 1994 *New Visions for Metropolitan America*. The Brookings Institution, Washington, DC and Lincoln Institute of Land Policy, Cambridge, MA

Fisher R 1984 *Let the People Decide: Neighborhood Organizing in America*. Twayne, Boston

Friedman L M 1968 *Government and Slum Housing: A Century of Frustration*. Rand, McNally, Chicago

Gale D E 1984 *Neighborhood Revitalization and the Post-industrial City*. Lexington Books, Lexington, MA

Gale D E 1996 *Understanding Urban Unrest: From Reverend King to Rodney King*. Sage, Thousand Oaks, CA

Gans H J 1962 *The Urban Villagers: Group and Class in the Life of Italian–Americans*. Free Press, New York

Gelfand M I 1975 *A Nation of Cities: The Federal Government and Urban America, 1933–1965*. Oxford University Press, New York

Greer S 1965 *Urban Renewal and American Cities: The Dilemma of Democratic Intervention*. Bobbs-Merrill, Indianapolis, IN
Haar C M 1975 *Between the Idea and the Reality: A Study in the Origin, Fate and Legacy of the Model Cities Program*. Little, Brown, Boston
Halpern R 1995 *Rebuilding the Inner City: A History of Neighborhood Initiatives to Address Poverty in the United States*. Columbia University Press, New York
Hartman C 1974 *Yerba Buena: Land Grab and Community Resistance in San Francisco*. Glide Publications, San Francisco
Hays R A 1995 *The Federal Government and Urban Housing: Ideology and Change in Public Policy*. State University of New York, Albany, NY
Jacobs J 1961 *The Death and Life of Great American Cities*. Random House, New York
Jackson K T 1985 *Crabgrass Frontier: The Suburbanization of the United States*. Oxford University Press, New York
Judd D, Parkinson M 1990 *Leadership and Urban Regeneration: Cities in North America and Europe*. Sage, Newbury Park, CA
Keating W D, Smith J 1996 Neighborhoods in transition. In: Keating W D, Krumholz N, Star P (eds.) *Revitalizing Urban Neighborhoods*. University Press of Kansas, Lawrence, KS
Mayer N S 1984 *Neighborhood Organizations and Community Development: Making Revitalization Work*. The Urban Institute Press, Washington, DC
Medoff P, Sklar H 1994 *Streets of Hope: The Fall and Rise of an Urban Neighborhood*. South End Press, Boston
Meehan E J 1979 *The Quality of Federal Policymaking: Programmed Failure in Public Housing*. University of Missouri Press, Columbia, MI
Palen J J, London B 1984 *Gentrification, Displacement and Neighborhood Revitalization*. State University of New York Press, Albany, NY
Rich M J 1993 *Federal Policymaking and the Poor: National Goals, Local Choices, and Distributional Outcomes*. Princeton University Press, Princeton, NJ
Rohe W M 1998 Do community development corporations live up to their billing? A review and critique of the research findings. In: Koebel C T (ed.) *Shelter and Society: Theory, Research and Policy for Nonprofit Housing*. State University of New York Press, Albany, NY
Rohe W M, Gates L B 1985 *Planning with Neighborhoods*. University of North Carolina Press, Chapel Hill, NC
Rohe W M, Quercia R G, Levy D K, Biswas P 1998 *Sustainable Nonprofit Housing Development: An Analysis of Maxwell Award Winners*. Fannie Mae Foundation, Washington, DC
Rusk D 1999 *Inside Game Outside Game: Winning Strategies for Saving Urban America*. Brookings Institution Press, Washington, DC
Schill M H, Nathan R P 1983 *Revitalizing America's Cities: Neighborhood Reinvestment and Displacement*. State University of New York Press, Albany, NY
Schwartz A 1998 Bank lending to minority and low-income households and neighborhoods: Do community reinvestment agreements make a difference? *Journal of Urban Affairs* **20**: 269–301
Squires G D (ed.) 1992 *From Redlining to Reinvestment: Community Responses to Urban Disinvestment*. Temple University Press, Philadelphia, PA
Varady D P 1986 *Neighborhood Upgrading: A Realistic Assessment*. State University of New York Press, Albany, NY
Vidal A C 1992 *Rebuilding Communities: A National Study of Urban Community Development Corporations*. New School for Social Research, New York
Walker C 1993 Nonprofit housing development: Status, trends, and prospects. *Housing Policy Debate* **4**: 361–414

W. D. Keating

Network Analysis

1. Definition

Network analysis (NA) is a set of integrated techniques to depict relations among actors and to analyze the social structures that emerge from the recurrence of these relations. The basic assumption is that better explanations of social phenomena are yielded by analysis of the relations among entities. This analysis is conducted by collecting relational data organized in matrix form, rather than surveying social characteristics of individuals sampled from a larger population. Traditional sample surveys isolate each actor and reduce their social relations to individual characteristics, the purpose being to analyse statistical relations among variables. These causal relations can be only viewed as useful indicators of underlying mechanisms, which are in fact any kind of encounter and exchange among real actors displaying everyday life strategies. NA concentrates its analytical attention directly on these phenomenological encounters, while traditional surveys can only identify causal connections and are unable to explain operating mechanisms (Collins 1988). For instance, a statistical relation between social origin and status attainment fails to depict the actual mechanism of social mobility; rather, it is merely a signal of social relations among the respondent's relatives and within their broader social milieu. It is the shape and intensity of the network of these relations that explains the persistence of, or change in, social status across generations. Although deterministic approaches usually emphasise that NA enables study of how the social structure of relationships around a person group, or organization affects behaviors and attitudes, social networks are in their turn constructed by individual or corporate actors, to the extent that they are able to mobilize social resources for the purpose of choosing partners or counterparts and organizing coalitions. In short, structurally bounded purposive actions may affect the social structure and vice versa.

If actors are depicted as nodes, and their relations as lines among pairs of nodes, the concept of social network changes from being an evocative metaphor to an operative analytical tool which utilizes the mathematical language of graph theory and of matrix and relational algebra.

Since the mid-1970s, the development of algorithms integrated into statistical packages, and the availability of greater computational capacity in the decade

that followed, enabled the successful application of these techniques to structures larger and more complex than the personal networks studied previously. The foundation of specialized journals (*Social Networks* and *Connections*) and the publication of NA manuals has helped to institutionalize this approach, which finds increasing acceptance in the literature of such disciplines as sociology, anthropology, social psychology, history, and political science, while statistics and mathematics contribute constantly to the development of analytical tools. The activity of an interdisciplinary community of scholars is evidenced by the founding of the International Network for Social Network Analysis in the late 1970s.

These developments have induced some scholars, mostly structural analysts, to identify a new scientific paradigm. Indeed, NA has features typical of what Kuhn described as a paradigm: the formation of a community of scholars which shares a special language, publishes journals, and meets regularly at conferences to discuss specialized topics concerning applications of technical tools and their theoretical interpretation. Also evident is the cumulative nature of this activity, in spite of the heterogeneity of the disciplines involved and of its fields of application. However, NA suffers from a lack of shared theoretical assumptions, and of theories derived from these assumptions. The idea that society is itself a set of relations, and that individual actors are embedded in social networks, is still too loosely defined for it to become a theory. For these reasons, NA has been sometimes called a technique in search of a theory. One may accordingly regard it as a set of techniques with a shared methodological perspective, rather than as a new paradigm in the social sciences. These techniques allow to undertake substantive research which yields good descriptions of complex structures and generates theoretical interpretations.

2. *Origins and Developments*

The ideas that social phenomena can be interpreted as networks of relations, that society itself can be interpreted as a network, and that social action can be explained as the outcome of the constraints imposed and the opportunities afforded by social relations, have been recurrent since Simmel's formal sociology and his studies of the relations between actors in dyads and triads. Nevertheless, it is only since the 1930s that attempts have been made in various disciplines to move from theoretical programs to empirical applications (Scott 1991).

A first approach, influenced by *Gestalt* theory, was developed in the US by a group of social psychologists who had emigrated to that country from Germany. Lewin, Moreno, and Heider stressed the role played by the structure of interpersonal relations in the construction of cognitive schemes, and they studied information flows and the formation of shared ideas. The main tool that they used for the description of empirical networks was the sociogram, together with analytical concepts like symmetrical-asymmetrical relation, centrality and periphery, reciprocity and cognitive balance. Another approach was directly linked with structural functionalism, and it was used to recast the classic research study on work organization in the wiring room of the Western Electric plant at Hawthorne. In that study, observation techniques drawn from anthropology had evidenced the importance of informal relations in work performance. Some years later Homans reorganized the same data in matrix form to conduct analysis very similar to what came to be called the 'block model' in the 1970s.

The so-called Manchester School of Social Anthropology, whose most outstanding representatives were Gluckman, Barnes, Mitchell, and Bott, also started from structural functionalism. But in contrast to the traditional emphasis on social integration and cohesion, it developed NA concepts to study conflict and power relations. Its shift of research subject from African communities to next-door-neighbor relations in European cities led to the discovery of the importance of informal relations in urban settings. The Manchester School applied concepts like density, connectivity, reachability, as well as parameters related to the intensity and strength of ties, but its exclusive concern with informal relations and its merely descriptive approach have contributed to the decline of the school since the 1970s.

By this time the main technical developments had moved to Harvard, where White devised mathematical concepts of structural analysis, following the pioneering work of Harary and Cartwright during the 1950s. Thereafter, the mainstream program would be to model the formal properties of social structures on the basis of graph theory and matrix algebra. The approach was, therefore, analytical and deductive rather than descriptive and inductive. Different types of graphs were defined and their properties studied. Field research and applications generated statistical debate on centrality, centralization, and the identification of cliques and other types of subgroups, while important concepts such as 'structural equivalence' were proposed.

The main theoretical assumption that fostered these technical developments was structural analysis, which theorized that an actor's behavior and attitudes can be interpreted in terms of their structural position in the network. Since the end of the 1970s, however, this deterministic approach has been criticized, prompting a second generation of authors to investigate also the structural opportunities available to actors when they use relations that can be mobilized from specific positions. In sociology especially, this reorientation has yielded results pertaining both to substantive

theories and to analytical models, usefully employed to suggest explanations of social phenomena.

As regards the former of these domains, different analytical types of ties, as well as concepts such as point centrality, network centralization, and structural balance have been proposed, together with numerous criteria for subgroup detection (see *Social Network Models: Statistical*). The concept of structural equivalence introduced by Lorrain and White (1971) has been inspired directly by role theory. As regards substantive research, a number of middle range theories can be grouped into four different fields of empirical application:

(a) The relation between network structure and actors' behavior and attitudes has evidenced that the density of ego-centered networks is correlated with mental disorders (Kadushin 1982); the efficiency of task-oriented groups depends on their structure (Bavelas 1950); members of cohesive networks tend to share common opinions (Epstein 1969); innovation diffusion is influenced by personal relations among actors (Coleman et al. 1957).

(b) The role played by networks in preserving and creating social inequalities has been investigated by research on social mobility, where macro-outcomes at the level of social stratification and microstrategies are analyzed jointly by means of network techniques. Moreover, studies on the labor market have evidenced that people acquire information about job opportunities mainly on informal relations, and 'strong' or 'weak' interpersonal ties determine differential chances of entering employment (Granovetter 1973).

(c) The function played by networks in economic behavior has inspired studies of interlocking directorates (Stokman et al. 1985) (see also *Interorganizational Relationships and Networks*), and theoretical interpretations of the market as a structure of roles in which the strategy of each producer is based on the observed actions of all others (White 1981).

(d) The relational concept of power derived from Weber is particularly suitable for NA applications in the study of political phenomena. It has been found that in negatively connected exchange networks power is not a function of centrality, and exchange choices may generate structural transformations in the overall network because of changes in power dependency relations (Cook et al. 1983).

3. *Merits, Limits, and Perspectives*

NA techniques allow researchers to specify empirical indicators of relational concepts and to control field hypotheses through the definition and measurement of traditional catch-all concepts like social structure and cohesion. The development of network analytical tools has helped to highlight the structures underlying the apparently inextricable complexity of relations among actors in everyday life. From this point of view, the NA approach has the same analytical purpose as other techniques, for instance, those developed in social mobility studies, whose main goal is to simplify descriptions and to standardize measures for use in comparison. Nevertheless, NA offers a wider range of potential applications than traditional surveys, given that it is hard to conceive a social phenomenon that cannot be analyzed as a set of relations among individual or corporate actors. Moreover, research problems can be addressed at both the micro and macro levels, without changing the research design and techniques, because corporate actors can simply be interpreted as networks of networks.

Major drawbacks counterbalance these advantages, however. It has often been pointed out that NA's abundance of analytical and technical apparatus contrasts with the poverty of its theoretical and substantive results, which are mostly descriptive, or indeed artifacts generated by simplified simulations. While these criticisms are perhaps applicable to surveys as well, the relative scarcity of empirical results in NA largely is determined by the specific difficulties of field research. Unlike traditional surveys based on standardized questionnaires on the opinions of anonymous respondents, the collection of relational information involves sampling problems as well as difficulties relating to the reliability of intrusive questions on personal contacts. This is why network analysts often prefer to collect data from documents and repertoires rather than ask direct questions. Nevertheless, scholars recognize that NA has contributed to the development of a number of empirically-grounded middle-range theories in a variety of fields. These new perspectives have in their turn encouraged the emergence of theoretical concepts based on relational assumptions. A promising concept is 'social capital' (Coleman 1988), which can be better formalized in network terms.

The availability of a complete set of structural indicators encourages network comparison in order to explain differences in space and changes over time. Two tendencies are now emerging. One is the development of statistical models which permit comparison between actual networks and models generated by probabilistic assumptions to test goodness of fit. The other is the integration of static structural data with additional survey-derived information. The consequent systematic combination of relational information with individual attributes of nodes and ties has yielded promising results, thereby demonstrating that NA and traditional surveys are not necessarily mutually exclusive techniques of research.

See also: Group Processes in Organizations; Network Models of Tasks; Networks and Linkages: Cultural Aspects; Organization: Informal; Simmel, Georg (1858–1918)

Bibliography

Bavelas A 1950 Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America* **22**: 271–82

Coleman J S 1988 Social capital in the creation of human capital. *American Journal of Sociology* **94**: 95–120

Coleman J S, Katz E, Menzel H 1957 The diffusion of an innovation among physicians. *Sociometry* **20**: 253–70

Collins R 1988 *Theoretical Sociology*. Harcourt Brace Jovanovich, San Diego, CA

Cook K S, Emerson R M, Gillmore M R, Yamagishi T 1983 The distribution of power in exchange networks: Theory and experimental results. *American Journal of Sociology* **89**: 275–305

Epstein A L 1969 Gossip, norms and social network. In: Mitchell J C (ed.) *Social Networks in Urban Situations*. Manchester University Press, Manchester, UK

Granovetter M 1973 *The Strength of Weak Ties*. University of Chicago Press, Chicago

Kadushin C 1982 Social density and mental health. In: Marsden P V, Lin N (eds.) *Social Structure and Network Analysis*. Sage, London

Lorrain F, White H C 1971 Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology* **1**: 49–80

Marsden P V, Lin N (eds.) 1982 *Social Structure and Network Analysis*. Sage, Beverley Hills, London, New Delhi, India

Scott J 1991 *Social Network Analysis. A Handbook*. Sage, London

Stokman F N, Ziegler R, Scott J (eds.) 1985 *Networks of Corporate Power*. Polity, Cambridge, UK

White H C 1981 Where do markets come from. *American Journal of Sociology* **87**: 517–47

A. M. Chiesi

Network Models of Tasks

Processes executed to perform a task can often be represented in a task network, as in Fig. 1. Vertex o is the start of the task (typically, stimulus presentation). Each arc represents a process. Associated with each arc is a real number (or random variable), the duration of the process. The processes leaving a vertex all begin when a release condition is met by the processes entering the vertex. A process leaving an AND gate vertex begins execution when all the processes entering the vertex are completed. A process leaving an OR gate vertex begins execution as soon as any process entering the vertex is completed. Vertex r is the end of the task (typically, response execution). The processes are partially ordered by their precedence relation, and so can be represented in a directed acyclic graph. Because mental processes are unobservable, the cognitive psychologist faces an inverse problem, namely, to infer the task network and the durations of the processes from response times measured under various conditions.

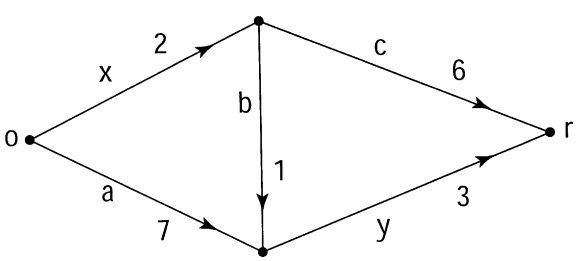

Figure 1
Task network. This is a Wheatstone bridge

1. Representation of Processes

The bar chart (or Gannt chart) in Fig. 2 illustrates a model of the psychological refractory period, adapted from Davis (1957) and used recently by Pashler (1998). Two stimuli, o_1 and o_2, are presented, requiring responses r_1 and r_2, respectively. For stimulus o_1, perceptual processing is denoted a_1, central processing is denoted b_1, and motor preparation is denoted c_1. Processes for stimulus o_2 are denoted analogously. The interval between the onsets of the stimuli, the stimulus onset asynchrony, is denoted SOA. When process durations are fixed numbers, the length of the bar representing a process indicates its duration. The arrow from the end of b_1 to the beginning of b_2 indicates that b_1 must be completed before b_2 can start, in accordance with Welford's (1952) single-channel theory.

Process durations are easily apprehended from the bar chart, but precedence relations between the processes are not. The bar chart in Fig. 2 can be converted readily to a task network, as in Fig. 3. Precedence relations are seen more easily in the task network, at the cost of losing a visual representation of the process durations.

One use of a task network in scheduling theory is to calculate the time required to complete the task, that is, the response time, from known durations of directed paths through the network. A directed path from vertex u to vertex z is formed by starting at u, going

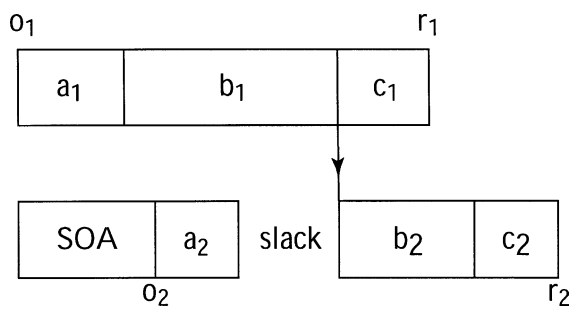

Figure 2
Bar chart

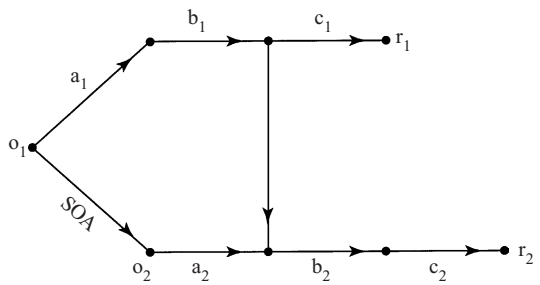

Figure 3
Bar chart, redrawn as task network

along an arc leaving u to another vertex v, and so on, ending at vertex z. There is no directed path from the starting vertex of a process to itself, that is, no directed path is a cycle. Hence a task network is a directed acyclic network. The duration of a path is the sum of the durations of all the arcs on it. If the vertices are AND gates, the network is called a PERT network, and the longest path from o to r is called the critical path. The time to complete the task is the duration of the critical path. If the vertices are OR gates, the time to complete the task is the duration of the shortest path from o to r.

2. *Constructing the Network*

The key idea for constructing the network is selective influence, pioneered by Sternberg (1969) in his additive factor method. Suppose one experimental factor (say, degrading stimulus quality) prolongs one of the processes, leaving the others unchanged. Suppose another experimental factor (say, increasing response complexity) prolongs another process, leaving the others unchanged. Suppose the processes required for the task are executed one after the other, in series. Then the increase in response time produced by prolonging both processes will be the sum of the increases produced by prolonging them individually (see *Additive Factor Models*).

When process durations are stochastically dependent, factors selectively influencing processes can have unexpected consequences (Townsend and Thomas 1994). The complications arise when a factor prolongs a process p, and the duration of process p is correlated with the duration of a process q; then the duration of process q is changed when the level of the factor is changed. See Dzhafarov (1999) for a general interpretation of selective influence under these circumstances.

2.1 *Prolonging a Single Process*

Suppose, for a moment, the process durations are real numbers. When a process p is prolonged by an amount Δp, the response time need not increase by Δp. To see this, suppose (until noted otherwise) all vertices are AND gates. In Fig. 1, process x can be prolonged by up to 2 units without increasing the response time. The longest time by which a process p may be prolonged without delaying the response is called the total *slack* for p, denoted $s(p, r)$. The first $s(p, r)$ units of the prolongation are used to overcome the total slack for p, and the remainder increases the response time. Let $T(0)$ denote the response time when no processes are prolonged and let $T(\Delta p)$ denote the response time when process p is prolonged by Δp. Then $T(\Delta p) = T(0) + \max\{0, \Delta p - s(p, r)\}$. Suppose now the process durations and the prolongation of p are random variables. For expected values,

$$E[T(\Delta p)] = E[T(0)] + E[\max\{0, \Delta p - s(p, r)\}].$$

2.2 *Prolonging Two Processes*

When each of two factors influences selectively a different process, the effects depend on how the two influenced processes are related. (The importance of process arrangements has been advocated continuously by Townsend.) Relations between processes are determined by directed paths. If there is a directed path from the head of a process p to the tail of a process q, then p must be completed before q can start, and p precedes q. Processes p and q are called *sequential* if either p precedes q or q precedes p. If processes p and q are not sequential, they are *concurrent*. Concurrent processes p and q can be executed in any order. That is, p and q may be executed simultaneously, or p may be completed before q starts, or q may be completed before p starts.

Assume (a) no process precedes itself (*irreflexivity*), and (b) if a process p precedes a process q, and q precedes a process t, then p precedes t (*transitivity*). In other words, precedence is a partial order.

For simplicity, again consider process durations and prolongations as fixed numbers. Let the response time when p is prolonged by Δp and q by Δq be denoted $T(\Delta p, \Delta q)$; the notation is analogous for other conditions. Suppose p and q are concurrent processes. The response time when both p and q are prolonged is the larger of the response time when p is prolonged and the response time when q is prolonged, that is,

$$T(\Delta p, \Delta q) = \max\{T(\Delta p, 0), T(0, \Delta q)\}.$$

Generally, the times are random variables with expected values,

$$E[T(\Delta p, \Delta q)] = E[\max\{T(\Delta p, 0), T(0, \Delta q)\}].$$

A quantity sensitive to the relation between two prolonged processes is the interaction contrast. For concurrent processes p and q, this quantity is negative or zero, i.e.,

$$E[T(\Delta p, \Delta q)] - E[T(\Delta p, 0)] - E[T(0, \Delta q)] + E[T(0, 0)] \leqslant 0.$$

Suppose p and q are sequential processes. When p and q are both prolonged the response time is a complicated function of their respective prolongations, Δp and Δq. Fortunately, when the prolongations are long enough to overcome the relevant slacks, the interaction contrast can be shown to take the following simple form:

$$E[T(\Delta p, \Delta q)] - E[T(\Delta p, 0)] - E[T(0, \Delta q)] + E[T(0, 0)] = E[s(p, r) - s(p, q)]$$

where $s(p, q)$ is the slack from p to q, that is, the longest time by which p can be prolonged without delaying the start of process q.

Ordinarily, the total slack for p is greater than or equal to the slack from p to q, and consequently the interaction contrast for sequential processes is zero or positive. The one exception is processes on opposite sides of a Wheatstone bridge; see x and y in Fig. 1. For such processes, the interaction contrast is zero or negative. In Fig. 1, $s(x, r) = 2 \leqslant s(x, y) = 4$ (considering process durations as fixed numbers), and the interaction contrast is negative. Processes on opposite sides of a Wheatstone bridge can mimic concurrent processes, because both arrangements of processes lead to negative interaction contrasts. The main difference is that for the Wheatstone bridge, when prolongations are long, the value of the interaction contrast does not change as the prolongations become still longer. This is not true for the interaction contrast corresponding to concurrent processes.

When vertices are OR gates, the main difference is a reversal of signs (Schweickert and Wang 1993, Townsend and Nozawa 1995). For concurrent processes, the interaction contrast is positive or zero. For sequential processes not on opposite sides of a Wheatstone bridge, the interaction contrast is negative or zero. For sequential processes on opposite sides of a Wheatstone bridge, the interaction contrast is positive or zero.

2.3 Constructing the Partial Order

Because concurrent and sequential processes behave differently, one can determine which pairs of processes are concurrent and which are sequential. This information is sufficient for constructing a directed acyclic graph. One begins with a *comparability graph* in which each process is represented by a vertex, and two vertices are joined by an undirected edge if and only if the corresponding processes are sequential. All possible directed acyclic graphs compatible with the comparability graph can then be constructed with the transitive orientation algorithm (Golumbic 1980).

2.4 Estimating Process Durations

If process duration distributions are exponential or general gamma, a reasonable assumption, then the response time can be derived from an order-of-processing (OP) diagram (Fisher and Goldstein 1983). Figure 4 gives the OP diagram for a PERT network in which processes a and b are concurrent and both are followed by process c. Each vertex in the OP diagram is a state, in which certain processes are under way (upper list) and certain others are completed (lower list). The task begins with a and b both under way. In the following states, either a is completed first or b is completed first. Following either of these states, process c is under way. In the final state, process c is completed, and the task is finished.

Suppose processes have independent exponential distributions, with the mean of process a denoted $1/\lambda_a$, and so on. The expected value of the response time can be read off the OP diagram. The first state is completed as soon as the faster of processes a and b is completed. The expected value of the minimum of two independent exponentially distributed processes is itself exponentially distributed, with mean equal to the sum of the means, so the expected value of the duration of the first state is $1/(\lambda_a + \lambda_b)$. The probability a is the first process to finish is $\lambda_a/(\lambda_a + \lambda_b)$. Because b is exponentially distributed, the expected value of the time remaining for b to complete given that a has already completed is $1/\lambda_b$. After repeated application of these rules, the expected value of the response time is the sum over all states of the expected duration of each state multiplied by the probability of reaching that state. For Fig. 4,

$$E[T] = 1/(\lambda_a + \lambda_b) + [\lambda_a/(\lambda_a + \lambda_b)](1/\lambda_b) + [\lambda_b/(\lambda_a + \lambda_b)](1/\lambda_a) + (1/\lambda_c).$$

To finish, one can use a nonlinear parameter estimation program to estimate the unknown rates, given observations from enough conditions.

2.5 A Generalization

OP diagrams can represent more general arrangements of processes than those in directed acyclic networks. Consider Fig. 2: if b_1 and b_2 were sequential, but either could be executed first, the situation cannot be

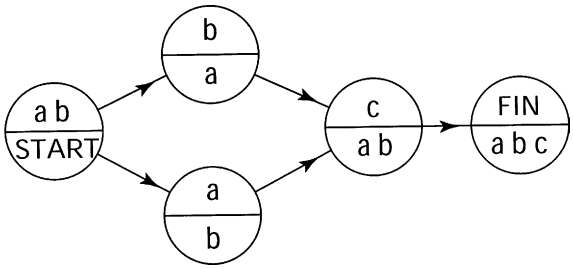

Figure 4
Order-of-processing diagram

represented by a single directed acyclic network, but can be represented by one OP diagram.

See also: Cognitive Modeling: Research Logic in Cognitive Science; Decision and Choice: Random Utility Models of Choice and Response Time; Discrete State Models of Information Processing; Stochastic Dynamic Models (Choice, Response, and Time)

Bibliography

Davis R 1957 The human operator as a single channel information system. *Quarterly Journal of Experimental Psychology* **9**: 119–29

Dzhafarov E N 1999 Conditionally selective dependence of random variables on external factors. *Journal of Mathematical Psychology* **43**: 123–57

Fisher D L, Goldstein W M 1983 Stochastic PERT networks as models of cognition: Derivation of the mean, variance, and distribution of reaction time using order-of-processing (OP) diagrams. *Journal of Mathematical Psychology* **27**: 121–51

Golumbic M 1980 *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, New York

Pashler H 1998 *The Psychology of Attention*. MIT Press, Cambridge, MA

Pinedo M 1995 *Scheduling: Theory, Algorithms, and Systems*. Prentice Hall, Englewood Cliffs, NJ

Schweickert R 1978 A critical path generalization of the additive factor method: Analysis of a Stroop task. *Journal of Mathematical Psychology* **18**: 105–39

Schweickert R, Wang Z 1993 Effects on response time of factors selectively influencing processes in acyclic task networks with OR gates. *British Journal of Mathematical and Statistical Psychology* **46**: 1–30

Sternberg S 1969 The discovery of processing stages: Extensions of Donders' method. In: Koster W G (ed.) *Attention and Performance II*. North-Holland, Amsterdam, pp. 276–315

Townsend J T, Nozawa G 1995 Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology* **39**: 321–59

Townsend J T, Thomas R D 1994 Stochastic dependencies in parallel and serial models: Effects on systems factorial interactions. *Journal of Mathematical Psychology* **38**: 1–34

Welford A T 1952 The 'psychological refractory period' and the timing of high-speed performance.' A review and a theory. *British Journal of Experimental Psychology* **11**: 193–210

R. Schweickert

Networks and Linkages: Cultural Aspects

For individuals in complex modern societies, greater variety in the kinds of people one knows goes with greater variety in the kinds of culture one knows. Both network and cultural variety are useful resources, and both are related to socioeconomic status, hence both can be seen as forms of social and cultural capital. The nature of social and cultural capital depends on the wider structural context, and hence varies over time and space.

Though there are many kinds of networks, this article only considers personal networks, or all the people that a person knows. Since people learn and use culture in social relationships, network analysis has much to offer to the sociology of culture, using powerful ideas that go back as far as Simmel (1955).

1. Network Variety and Cultural Variety

The variety in a person's network is correlated with the variety in that person's culture. Network variety means knowing many different kinds of people; cultural variety means knowing about many different forms or genres of culture. Each kind of variety feeds the other. On the one hand, people with a wide range of cultural tastes and knowledge can easily find something in common with many different kinds of people, and shared interests are attractive, so culturally diversified people make new acquaintances and friends of many kinds. On the other hand, if one knows many kinds of people they will stimulate interest in and knowledge of their many tastes. DiMaggio (1987) anticipated this relationship, while Erickson (1996) showed evidence for it.

Clearly this relationship hinges on suitable definitions of network and cultural variety: network variety must be some kind of variety that leads to contact with diverse cultures, and cultural variety must include forms of culture popular with people in different structural locations. This is always possible because boundaries between different social locations thrive on and maintain subcultural differences. Though possible, it is essential, and requires some care.

Research so far on network variety has focused on the extent to which a person has contacts in many different occupations. Given the importance of work in modern societies, and the known connections between occupation and taste, this is a useful though

incomplete starting point. Reasons for occupational differences in culture include cultural selection in getting into an occupation, the role of occupational boundaries in maintaining cultural differences, and the cultural demands of work with others outside one's occupation. First, people who enter different occupations have different cultures to begin with. Education, a strong predictor of tastes, is also a strong predictor of the kind of work people get. When employers select people with a type and level of education seen as appropriate for a job, they also select people who have the cultural profile that goes with that kind of education. People also try to get work that suits their tastes. People are more likely to get jobs in milieus where their kind of taste and style is prevalent (so that they show well in interviews, looking like people who will fit in and work well with those already in the occupational setting). Second, occupations have distinctive locations in the social structure of work, that is, niches in particular organizations, organizational subunits, sectors of the economy, labor markets, and so on. In turn, different segments of the working world develop their own subcultures, as can easily be seen by comparing the style of dress conventional for (say) a financial district, a hospital, and a bohemian artists' quarter. Third, occupations differ in the kinds of outsiders that people have to work with, and hence the kinds of cultural displays people are encouraged to make in order to get along with outsiders. Peterson and Simkus (1992) divided occupations into 'status groups' with 'the same job conditions and the same level of requirements for social and cultural skills' (Peterson and Simkus 1992, p. 155), and showed striking differences in tastes for music and other leisure activities. Using a different occupational grouping, Bourdieu (1984) also showed substantial differences in the tastes of people in different occupational groups in France.

Since people in different occupations do have different cultural repertoires, contact with people in many different occupations should include exposure to many kinds of culture. But how can we measure the extent to which people have such occupationally varied contacts? It is surprisingly easy to do so. Following a format pioneered by Lin and Dumin (1986), Erickson (1996) developed a list of nineteen occupations, asked respondents whether they knew anyone in each of these, and counted the number of different occupations in which a respondent did know someone: this simple count indexes network variety. Though the strategy is easy to execute, the choice of occupations is critical. Since Erickson was studying the private security industry in Toronto, she selected occupations that varied in prestige and power (e.g., company owner, locksmith) and that were important to the industry in various ways (e.g., electricians, important as skilled labor for alarm installation, and insurance brokers, who are an important source of investigation contracts). Measured in this way, network variety was strongly related to knowledge of several different cultural genres, even controlling for several known correlates of culture (ethnicity, gender, family background, education, and class). Studies of other social settings would call for different, locally appropriate choices.

Future work should not only use carefully considered forms of occupational variety, but should also use entirely different kinds of variety with respect to other social cleavages. Each kind of cleavage goes with its own cultural contrasts. Both gender and ethnicity are important forms of inequality and division, and each goes with a pattern of cultural difference that is its own, different from the contrasts between classes or occupational groups (Hall 1992). Owners and managers know more about business magazines than simple employees; men know more about sports and women know more about novels; nonwhites and immigrants know less about locally famous artists than do native born whites (Erickson 1996). One could ask people whether they know anyone in each of a set of ethnic groups, chosen to vary widely in their cultures. It would be of little use to ask whether people know anyone in very large ethnic groups, or know men or women, since almost everyone will. It would be more useful to find out whether people know a wide variety of men, women, or members of large ethnic groups. Thus one might ask whether people know any men, any women, or any members of large ethnic groups in each of a set of occupations. Those who know (say) women in many lines of work should have wide and strong access to the kinds of culture women prefer. Whatever the details of the procedure, we need research on network diversity of nonoccupational kinds and the linkages of such network diversities with culture.

Research so far on cultural diversity has focused on liking of, participation in, and knowledge of various genres that people take for granted as being different kinds of cultural things. For example, Peterson and Simkus (1992) look at liking of different kinds of music such as classical or country music and arts participation such as going to museums, while Erickson (1996) looks at knowledge such as recognition of sports stars or successful artists. One issue here is the theoretically important kind of relationship to culture: should we be interested in what people know, like, do, own, or what? Probably all need attention, since many forms of culture reflect and maintain the boundaries between social groups (Lamont and Lareau 1988). Another kind of issue is the choice of genres to consider. Theoretically, the choice of genres depends on the cleavages and processes of interest. For example, Bourdieu (1984) gives extensive attention to taste related to the fine arts, because he thinks this is a crucial aspect of the high status cultural capital distinguishing elite people from others. Meanwhile, Erickson (1996) examined some genres correlated with class (art, business magazines) to explore culture

related to domination between classes *and* a genre not related to class (popular sports) to explore culture related to coordinating the activities of people in different classes. She argues that company owners and managers need both to dominate and to coordinate as they work, and hence need a range of cultural resources. Examining other cleavages, such as gender or ethnicity, calls for attention to other genres that divide or unite different groups.

2. Network Variety, Cultural Variety, and Social Class

The relationship between network variety and cultural variety is important in part because both variables are strongly linked to class, as causes and as consequences. Again, however, it is essential to think carefully about *which* aspects of networks and cultures have strong and important relationships with inequality. One lively current debate concerns the importance of variety vs. the importance of high status in both networks and cultures. Research on France tends to emphasize high status, while research on Canada and the United States emphasizes variety, perhaps because of the greater centralization and hierarchy in France. For France, Bourdieu (1984) argues that the culture of elites becomes defined as superior, more distinguished culture. People with higher status family backgrounds learn higher status cultural resources; these resources help them to succeed in school and in work, and these successes reinforce cultural advantages. Higher status people not only have higher status tastes, but they disdain and reject the tastes of those with lower social standing. Bourdieu has less to say about how networks are related to social status, but tends to underline the extent to which people associate with others of their own cultural level, suggesting that social status is correlated with the status of one's contacts as well as the status of one's culture. Some research on social networks in France shows that people have friends who are very like them in occupation, education, and income (Degenne and Forse 1999, pp. 32–5).

Bourdieu's model implies that culture is clearly ranked in status, that higher status people have higher status culture, and that higher status people reject lower status taste. Peterson and Kern (1996) argue that this 'elite to mass' or 'snob to slob' pattern may well have held in North America a century ago, when elite groups used high status culture to distinguish themselves from ambitious waves of lower status immigrants, but this pattern does not hold now. Instead, higher status people are 'omnivores' who like many kinds of culture, including a mixture of genres of high, medium, and low status. Lower status people are 'univores' who like one or a few genres. For example, lower status people often like just one kind of music, while higher status people like many; lower status people usually have a favorite kind of music, while higher status people often do not, valuing variety instead. Similarly, Erickson (1996) found that higher status people know more about each of several genres, whether these were cultivated ones like art or popular ones like sports. Thus North American research indicates that social status is correlated with the variety of culture people take an interest in, not with culture's prestige.

The real degree of difference between France and North America is still a matter of debate. There is little strictly comparable research. Further, it is not clear what truly comparable research would be like. One cannot simply use the same research tools in different countries, since different cultural items can play equivalent roles in two different societies. For example, each national elite may follow a different prestigious newspaper with special attention, or may share a common knowledge of a different set of classic novels. Holt (1998) works on translating the spirit of Bourdieu's work into theoretically, not literally comparable research in the United States. Lamont (1992) used in-depth interviews to compare the ways that middle class American and French men use culture to mark status boundaries. Much more of this careful comparative work needs to be done.

However, there is some agreement that the game of culture is at least partly different in France and North America, and that this is in part because of social structural differences. For example, Lamont (1992) found that her French respondents more often used culture (in the sense of cultivation) to separate themselves and their class from others, while her North American respondents used socioeconomic standing more. She attributes this difference in part to the greater centralization of the French educational system, mass media, cultural institutions, and informal intellectual networks, which all contribute to a much clearer and more socially consequential ranking of culture. Lamont also discusses the lower rates of geographic and class mobility in France, which contribute to greater cultural distancing between different classes.

Like culture, social networks are related to social status as both cause and consequence; as for culture, the important aspect of networks may be high status contacts or varied contacts. Lin (1999) provides a useful overview of research in North America, Europe, and Asia. Young adults from higher status families begin their working lives with more diversified networks including more high status contacts. Richer networks lead to better jobs. Higher status jobs then lead to still more diversified networks, since higher status jobs include contacts with a wider range of people, and to more high status contacts, since higher status jobs include work with others at high levels. Thus although higher status people do have more high status contacts, especially among those they are closest to, they do not restrict themselves to high status contacts. Instead, the higher the status the more wide-

ranging the network, with contacts over much of the occupational prestige hierarchy. In turn, both high status contacts and contact variety can be useful in getting a better job: higher status contacts are gatekeepers for higher level jobs, and varied contacts provide a wide range of job opportunities, as well as being a job qualification for the kinds of upper level work that include extensive networking. Unfortunately France is not one of the countries Lin (1999) reviewed, so it is not clear whether high status contacts may have a relatively large role (compared to network variety) in the centralized French system as a whole, though high status contacts do play a major role in elite life (e.g., Kadushin 1995).

3. *Variations Between Different Kinds of Societies*

Comparison of France and North America suggests that the large-scale organization of social structure and culture affects the interconnections of networks, culture, and socioeconomic standing at the level of individuals. DiMaggio (1987) provides a more general argument concerning these linkages. At the level of communities and societies, DiMaggio (1987) argues that the organization of culture mirrors the organization of social relationships. He defines organizations of cultural elements as 'artistic classification systems (ACSs): the way that the work of artists is divided up both in the heads and habits of consumers *and* by the institutions that bound the production and distribution of separate genres' (DiMaggio 1987, p. 441). A genre is one particular kind of culture, as defined by the people in a society or community, and an ACS is the total ensemble of recognized genres. ACSs vary on four dimensions, three of which are important here: differentiation, hierarchy, and boundary strength.

Differentiation is the number of distinct genres in an ACS. Using music as an example, people may just distinguish music from other arts like story telling; or may distinguish several varieties of music like jazz and classical; or may make finer subdivisions such as Baroque and other subtypes of classical music. ACS differentiation grows with the complexity of social structure, because each social subgroup is a distinct social setting in which unique tastes can develop, and members of each subgroup have some interest in developing unique tastes to express their distinctive identity and mark their group boundary. Complexity of social structure includes all socially consequential ways in which different groups form: heterogeneity (such as multiple ethnic groups), status diversity (many different levels of inequality), and multiple forms of inequality not well related to each other (so people share a similar level of education with one set of people, a similar occupational prestige level with another set, and so on).

Hierarchy is 'the degree to which genres are ordered hierarchically by prestige as opposed to perceived as different but of equal value' (DiMaggio 1987, p. 447). Because genres gain prestige through association with high status groups, genres are more clearly ranked in prestige to the extent that a community has clearly ranked social groups: groups with a single kind of ranking, with great inequality from top to bottom, and with little interaction between different levels.

Boundary strength is the extent to which people treat cultural boundaries as important, for example by rejecting art works that are hybrids of distinct traditions. Cultural differences seem more important to the extent that social boundaries are important, with distinct groups having limited interaction across their borders. Deep social divisions encourage the separate development of cultures that can become quite different objectively. Strong social boundaries also encourage taking these cultural differences seriously, since culture expresses group identities and differences.

In these terms, the contrast between France and North America is one small part of a broader hypothesis: social status is more highly correlated with high status taste and high status networks, and high status tastes and contacts are more important to gaining social status in societies with ACSs and social structures that are less differentiated, more hierarchical, and have greater boundary strength. In such societies high status taste is cultural capital and high status contacts are social capital, where 'capital' is something one can invest in (for example, through education or astute networking) and from which one can gain returns in the form of class location. But in societies with ACSs and social structures that are more differentiated, less hierarchical, and have weaker boundaries, individuals must relate well to many different kinds of other people to succeed, especially at the higher social status levels where many disparate contacts are necessary. In such societies cultural capital is varied culture plus the ability to switch on the kind of culture appropriate to the circumstances, while social capital is a wide variety of contacts including but not limited to high status contacts.

Fully testing these idea would require extensive work on the social and cultural structures of many varied communities, a daunting task not yet performed (though see Blau 1989 for some comparisons of 125 cities in the United States).

4. *Variations Over Time*

Since societies vary over time, so should networks, culture, and their connections. The earliest theorist of historical change, and still one of the most interesting, is Simmel (1955). Simmel compares social structures of many kinds over many centuries; for simplicity, this article considers only the contrast between very simple

societies and complex modern ones. In simple societies such as small, largely independent communities, each community contains people who vary fully in age, status, and other important social attributes. But each of these different people knows most or all of the others, so each has more or less the same network and is exposed to the same influences. Given a small population, each person must interact with each of the others a good deal in varied ways, so the shared influences are strong. Thus people can develop little individuality or differentiation in their social experiences or their tastes. In more complex modern societies, people can selectively affiliate with others who share their particular tastes and interests, as in joining a voluntary association dedicated to one's profession or one's hobby. Within a special interest group, relationships are narrow, focused on the special interest itself, so they strongly reinforce that interest without affecting other tastes or interests much. Each member can and often does have other interests which can be quite different from the outside interests of fellow members. Indeed, Simmel portrays the modern person as one who lives at the unique intersection of his or her own particular collection of formal or informal groupings devoted to a range of particular tastes. This uniqueness of social and cultural influence is the critical foundation of modern individualism. Though Simmel does not emphasize this, it is clear that this modern condition does not hold to the same degree for everyone in a modern society: the multiplicity of affiliations grows with social status. For example, research in dozens of countries shows that those with higher education or occupation prestige belong to a greater number of voluntary associations. Thus class predicts variety of affiliations and hence variety in networks, variety in tastes, and individuation, a pattern echoing the discussion of the North American pattern above.

Changes over smaller ranges of time are also of interest. Most strikingly, we know very little about changes over time in people's lives, since there are no substantial studies of networks, culture, and class over time. This leaves us unsure of the causal connections among these; we suspect that each of the three affects the others, but have no evidence that this is the case, nor do we know the details of the mutual influences. For example, are some effects faster than others? Does a change in job alter networks immediately as former work colleagues are lost and replaced while a change in networks takes some time to lead to a new job? Is the pace of change faster at some times in life, especially perhaps in the turbulent and changeful transition from youth to adulthood? We also know little of the processes by which networks affect cultures, or culture affects networks, or each shapes and is shaped by class; most of our research shows strong and interesting correlations but only offers speculation about how these correlations develop over time. These are exciting directions for future work.

See also: Cultural Expression and Action; Cultural Variations in Interpersonal Relationships; Culture, Sociology of; Network Analysis; Networks: Social; Simmel, Georg (1858–1918); Social Networks and Gender

Bibliography

Blau J 1989 *The Shape of Culture: A Study of Contemporary Cultural Patterns in the United States*. Cambridge University Press, Cambridge, UK

Bourdieu P 1984 *Distinction* (trans. Nice R). Routledge and Kegan Paul, London

Degenne A, Forse M 1999 *Introducing Social Networks*. Sage Publications, London

DiMaggio P 1987 Classification in art. *American Sociological Review* **52**: 440–55

Erickson B 1996 Culture, class, and connections. *American Journal of Sociology* **102**: 217–51

Hall J R 1992 The capital(s) of cultures: A nonholistic approach to status situations, class, gender, and ethnicity. In: Lamont A, Fournier M (eds.) *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago

Holt D B 1998 Does cultural capital structure American consumption? *Journal of Consumer Research* **25**: 1–25

Kadushin C 1995 Friendship among the French financial elite. *American Sociological Review* **60**: 202–21

Lamont M 1992 *Money, Morals, and Manners: The Culture of the French and the American Upper-Middle Class*. University of Chicago Press, Chicago

Lamont M, Lareau A 1988 Cultural capital: Allusions, gaps, and glissandos in recent theoretical developments. *Sociological Theory* **6**: 153–68

Lin N 1999 Social networks and status attainment. *Annual Review of Sociology* **25**: 467–87

Lin N, Dumin M 1986 Access to occupations through social ties. *Social Networks* **8**: 365–86

Peterson R A, Kern R M 1996 Changing highbrow taste: From snob to omnivore. *American Sociological Review* **61**: 900–7

Peterson R A, Simkus A 1992 How musical tastes mark occupational status groups. In: Lamont A, Fournier M (eds.) *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago

Simmel G 1955 The web of group affiliations. In: Simmel G (ed.) *Conflict and the Web of Group Affiliations*. Free Press, Glencoe, IL

B. H. Erickson

Networks: Social

Social network analysis in general studies the behavior of the individual at the micro level, the pattern of relationships (network structure) at the macro level, and the interactions between the two. The analysis of the interaction structures that is involved in social network analysis is an important element in the analysis of the micro–macro link, the way in which

individual behavior and social phenomena are connected with one another. In this perspective, social networks are both the cause of and the result of individual behavior. Social networks provide and limit opportunities of individual choices, whereas at the same time individuals initiate, construct, maintain, and break up relationships and by doing so determine the global structure of the network. However, network structure as it exists is seldom constructed consciously by its individuals. It is often the unintended effect of individual actions and can as such be called a spontaneous order.

1. Theoretical Perspectives

Which network structures and positions create strong opportunities or, on the contrary, strong constraints depends on the instrumental value of the relationships under study. *Social capital* is the opportunity structure created by social relationships (Lin 1982, Coleman 1990, Burt 1992). Social capital gives individuals access to resources of others that can be exploited for the realization of their goals. The amount of social capital depends on the amount of these resources, their value for the goal realization of the individual, and the willingness of others to mobilize them (Flap 1999). The value of the resources for the individual strongly depends on functional interdependence, the willingness of others to mobilize resources for the individual on their perception of the interdependence (i.e., the cognitive dependence). Differences in functional and cognitive interdependence explain why certain researchers emphasize other network structures and positions as contributing to an individual's social capital. For example, Coleman (1990) stresses the importance of large and dense networks for control in his study of school communities. Large and dense networks create shared information, high visibility, and common norms in a community. Burt (1992), on the other hand, stresses the importance of unique and nonoverlapping relationships for acquiring unique information in organizations, giving individuals a better chance to find creative solutions for problems and thus providing them better opportunities for career.

Functional and cognitive interdependencies differ particularly between strong and weak ties (Granovetter 1973). Strong ties are valued in themselves. The ties are not primarily instrumental for the attainment of other goals. Their value is based on the other individual as a person and the quality of the relationship with that individual. Family and friendship ties are typical examples of such relationships. Strong ties tend to be reciprocal, transitive, and clustered. Strong ties give a sense of belonging to a group and the group often has priority above the individual and individual relationships. Sharing is often based on need and norms tend to promote equality (Lindenberg 1998). Creating negative attitudes towards other groups often helps to strengthen the predominance of the group, which may give strong negative externalities for society as a whole (think of gangs and other criminal organizations). Weak ties are valuable long-term relationships but their value is primarily instrumental, related to higher ordered goals, goals not primarily located in the relationship or individuals themselves. Weak ties tend to be less clustered; the group is less dominant and often only vaguely delineated. Reciprocity based on equity norms prevails.

The fact that social networks create conditions for cooperation through information and sanctioning is due to exchange processes that create win–win situations (Homans 1950, Blau 1964). A fruitful and very promising approach is the study of the effects of social networks in noncooperative game theory (Raub and Weesie 1990, Flache and Macy 1996, Bienenstock and Bonacich 1992). Network exchange theory specifically investigates the effects of network structures on the choice between alternative exchanges and on exchange rates (Willer 1999). Major effects are particularly due to possibilities for social actors to exclude others. Exchange network theory illustrates again that effects of network structures are context-sensitive and cannot be generalized without taking the context and substance into account. From this perspective it is to be regretted that the splitting of private resources is the dominant situation considered in exchange network theory. The integration of exchange theory and social networks has also been proven to be very successful in the field of policy networks. Most of these models build on Coleman's social exchange model and confine exchanges to influence network relationships (Laumann et al. 1987). These models make the step from micro behavior to macro effects explicit and are able to predict outcomes of decisions, to derive the power of social actors and the value of decisions. Later models try to solve a number of remaining theoretical issues (Stokman and Van den Bos 1992, Pappi and Henning 1998).

2. Network Analysis: Concepts and Techniques

In the social network literature the emphasis lies on the description of static network structure. The best illustration of this can be found in the handbook by Wasserman and Faust (1994) and the computer package *Ucinet* (developed by Borgatti, Everett, and Freeman) that contain most current static network methods. A large number of other network studies examine the effects of the network structure on the behavior and attributes of the network members, the effects of the macro structure on micro behavior. In these studies, the network is considered given and constant, and the ways in which this network influences processes or individuals in the network is

examined. In many of these studies, only relationships of the individuals under investigation are collected (at most combined with their perception of the relationships among their network members). These *ego-centered network studies* examine the effects of differences in size and composition of the personal networks and the multiplexity of an individual's personal relations (Wellman and Berkowitz 1997). Examples of effects studied are an individual's social well-being, social support, health, labor market position, career. Only recently has more attention been given to the evolution of social networks and the interplay and feedback mechanisms between macro structure and microbehavior.

In social network analysis a large number of measures have been developed to characterize and compare network structures and positions in networks. Depending on what determines differences in opportunity structures, the analysis can be focused on differences in centrality, on the investigation of strongly connected clusters, of positions that are structurally equivalent in networks, or of unique positions. Other measures enable the comparison of network structures as a whole, e.g., the investigation of their effectiveness for goal achievement. In addition, statistical network models can be used for testing structure against null models, for parameter estimation, and, more recently, for testing network effects of different incentive structures.

2.1 Representation of Social Networks

Most social network measures start from a representation of points and lines. The points usually represent individuals or other social actors like organizations or positions in organizations. Lines represent relationships between social actors. Any type of dyadic relationship can be represented, but most common are communication, friendship choices, advice, trust, influence, and exchange relationships. Most of these relationships are not necessarily reciprocal. In that case, directed lines can be used. Vertices are sometimes used for points, arcs for directed lines, and edges for undirected ones. This representation makes it possible to apply graph theory, a branch of discrete mathematics. Many concepts and theorems of graph theory can meaningfully be applied in social network analysis (see particularly Harary et al. 1965). Connecting positive or negative signs to directed lines enables the representation of positive and negative ties. Heider's representation of the cognitive balance theory as a signed graph with three points is well known. An individual under study can have a positive or negative relationship to another person. Both persons can have a positive or negative attitude toward another object (e.g., a third person or a certain activity). The cognitive system of the individual under study is in balance if and only if all three relationships are positive or two of them are negative. Harary et al. (1965) generalized the idea of balance to a whole social system. In a valued graph, values are attached to the (directed) lines, and in a multi graph different types of relationships are distinguished.

Major computational possibilities became possible by the representation of a social network in a matrix. Moreno (1934) introduced such a representation for friendship choices in a school class. The rows and columns represent the points, and the cells the relationships from the row to the column point. In an adjacency matrix only ones (for a directed line from the row point to the column point) or zeros (no such line) are used. In a signed graph positive ties are represented by $+1$ and negative ones by -1. In a valued graph the values of the ties can be given. Stacked matrices are used to represent multi graphs. Modern relational databases make a more efficient representation of social networks possible in one or more tables for points (with the possibility of adding many variables for the points) and a table for each type of relationship (possibly with many relationship variables). The most important measures and approaches in the different classes are summarized below.

2.2 Actor Centrality

Depending on the instrumental value of the social network under investigation, central social actors may have better opportunities to realize their goals than less central ones. But also the type of centrality that is relevant for differences in opportunity structures may vary. Assuming a social network that represents communication, Freeman (1978) classifies the many centrality measures into three classes. A fourth class measures prestige or status. Measures based on the degree, the number of points with which a point is directly connected, indicate the communication activity of a point. In directed networks, centrality in terms of outdegree and indegree should be distinguished. In friendship choice networks the number of choices received (indegree) generally indicates centrality (popularity); in influence networks centrality is based on the number of outgoing relationships (outdegree). Degree-based measures indicate local centrality, as the global structure of the network is not taken into account (think of the éminence grise's important tie to the throne). Distance-based measures indicate the relative proximity of points with other points in the network and the extent to which a point can communicate with other points independently of others. In this context, the distance from a point to another point is the minimum number of ties that must be used to transmit a message to that point, the length of a shortest path. Betweenness or rush is the third type of centrality and measures how important a point is for the transmission of information between other points. Betweenness measures assume that information is mainly transmitted through shortest paths, connections based on the lowest number of con-

secutive ties. Status-based measures take all direct and indirect connections into account. They were originally developed to indicate centrality in influence networks.

2.3 Network Centrality

Again depending on the instrumental value of the social network under investigation, certain network structures are more effective than others. Three dimensions of social network structure dominate in the literature. The first is network *density*, the number of existing relationships relative to the possible number. Dense networks are more important for control and sanctioning than for information. Dense networks tend to generate a lot of redundant information. They generate many constraints and are inefficient for creative new solutions (Burt 1992). Comparing densities of networks of quite different sizes is difficult, as large networks tend to be sparse. A comparison in terms of mean degrees is a better alternative in such cases. The second is network *centralization*. In connected networks, high centralization corresponds with a high variance of the degrees of the points. Snijders (1981) derived the maximal possible values of the variance, given the number of points or given the number of points and lines. He also derived the expected variance for different null models under the same two conditions. This makes a good comparison of centralization possible in networks of different sizes and densities. The third is *clustering* and *segmentation*. A simple measure of segmentation is S_3, the number of pairs of points at distance two or higher divided by the number of pairs of points at distance three or higher (Baerveldt and Snijders 1994). Closely connected to the question of clustering and segmentation is the detection of dense clusters in a network.

2.4 Detection of Equivalent Sets

Whereas in the detection of dense clusters the emphasis lies on the relationships, analysis of equivalence is focused on the detection of similar positions of points in the network. The basic idea behind equivalence is that two or more points with a similar position in the network have the same opportunity structure. Two points are structurally equivalent if and only if they have identical sets of points to which they have outgoing relationships and from which they have incoming relationships. If more types of ties are considered, this should be the case for each type. Two points are regularly equivalent if the points with which they are directly connected are themselves equivalent rather than identical. For example, in a hierarchical organization, chiefs of different production units supervise different individuals, but their place in the organization is identical. They probably have the same incentive structure. In these cases regular equivalence represents the theoretical notion better than structural equivalence. Depending on the definition of the desired equivalence, many types of regular equivalence have been distinguished. Everett and Borgatti (1994) give a general definition of equivalence and show that all types, including structural equivalence, are specifications of a common class. As strict structural or regular equivalence is unlikely to occur in reality, methods have been developed to detect equivalent sets based on similarity measures between positions. Methods differ in terms of the type of similarity measure used and in terms of the grouping method to arrive at a partition of the points in more or less equivalent sets.

2.5 Statistical Models for Static Network Analysis

Statistical models for social networks are more difficult to derive because the relationships between the points cannot be treated as independent observations. Unbiased estimation of parameters is only possible by taking these dependencies into account. Consequently, maximum likelihood estimation is often not possible and estimation methods have to be used that are based on computer-simulated approximations. Dependencies between relationships also complicate the choice of null models. An illustration of the latter is the statistical analysis of local structure in networks, known as the triad count. The triad count aims to investigate which types of triads are overrepresented and which ones are underrepresented in a network. For example, if friendship choices tend to be transitive, then triads with intransitive choices tend to be underrepresented while triads with transitive choices tend to be overrepresented. From empirical studies, it is well known that individuals tend to differ systematically in terms of the number of friendship choices they make (activity) and receive (popularity). Moreover, friendship choices tend to be reciprocal, i.e., mutual choices tend to be overrepresented. Ideally, testing transitivity in friendship networks should be tested against a null model that conditions for the outdegrees, indegrees, and the number of reciprocal choices in a network. Recent statistical models, based on simulated approximations, do so, but in the earlier models either null models were used that only take the number of mutual choices into account (e.g., the MAN null model) or the outdegrees and/or indegrees of the points. Any linear combination of triad types can be tested against these null models, thus enabling the testing of different characteristics of local structure (Holland and Leinhardt 1975).

Most recent statistical models start from the assumption that the relationship between any two points A and B can be described by four states: a reciprocal choice, an asymmetric choice from A to B, an asymmetric choice from B to A, and no choice. The four states can be seen as a realization of a stochastic process. They estimate for each point a sender (ac-

tivity) and a receiver (popularity) parameter and for the network as a whole a density and a reciprocity parameter. Later models also include the estimation of higher order parameters, like triadic dependencies as e.g., transitivity (Wasserman and Pattison 1996). Subsequently, a type of logistic regression makes it possible to explain variation of these parameters over individuals and subgroups (Lazega and Van Duijn 1997). For example, assume girls (sex equals 1) tend to make more friendship choices than boys (sex equals 0). A positive sender effect of sex implies a larger probability of a relationship from a high scoring point on that variable (a girl) than from a low scoring point (boy) to some other point. Density and reciprocity effects are often defined by the (absolute) differences of the scores of the points on the variable. If boys and girls tend to make more friendship choices within their own groups, we find a negative effect of the absolute difference of the sex scores on the density parameter.

2.6 Evolution and Network Dynamics

In network evolution, two processes take place simultaneously. On the one hand, social actors shape the network by initiating, constructing, maintaining, and breaking up relationships. On the other hand, attributes (behavior, opinions, attitudes) of social actors are partly shaped by their relationships. Increasingly, Markov models are used for the analysis of network change, and spatial autocorrelation models for influence processes. Integration of the two processes requires further integration of theory and statistical testing. The first steps in this direction have been taken.

Holland and Leinhardt (1977) introduced Markov processes as the general framework for stochastic models of network evolution. The basic idea of Markov models is to conceive the social network structure as changing from one state into another over time. The unit of analysis is usually the dyad with its four possible states (see above). The parameters that govern the process concern the likelihood of transition from one of these four states into another. The original Markov models assume that the parameters are stationary over the whole process and that the population is homogeneous. Recent models have considerably increased the analytic possibilities of Markov models by eliminating these strongly limiting assumptions. Now, change parameters may well be dependent on the stage of network development and different for pairs within and between subgroups (Leenders 1996). An important next step is Snijders' integration of Markov models with random utility models, thus realizing a much stronger link between theory and statistical testing (Snijders 1996). In these models, random utility modeling is used to derive which network characteristics or interactions between individual and network characteristics are likely to produce high utility and thus are likely to govern network change. For each of these utility components, parameters are estimated indicating their strength.

Spatial autocorrelation models, such as those proposed by Doreian et al. (1984), Friedkin and Johnsen (1997) and Leenders (1997), are often used to model the influence process. In the most commonly used spatial autocorrelation models, it is assumed that influence on an individual characteristic (like behavior, opinion, or attitude) is only partly determined by the relationships in a network (estimated by the parameter α). The other part depends on other individual characteristics, like one's own background characteristics. The network part of social influence is assumed to be determined by a matrix of weighted influence relations where the total incoming influences on an actor sum to 1.

3. Conclusion

Since the early 1970s there has been impressive cumulative progress in social network analysis and research. From a rather isolated field, strongly oriented to descriptive structural and static analysis, social network analysis has grown into a well-embedded field, widely accepted as highly important for solving central theoretical problems of cooperation and coordination. With the growing importance of social networks in the information society with virtual communities developing in many segments of society, its importance and contributions to theoretical solutions can only grow. Issues of scope are challenging. Objects of study will vary from small group networks to social networks of billions of points. New techniques to visualize networks are also challenging. These allow certain structural characteristics to become visible and the effects of changes on these characteristics to become transparent. These developments can be followed by linking with the international virtual community of INSNA (the International Network for Social Network Analysis) and the links to be found there (http://www.heinz.cmu.edu/project/INSNA).

See also: Coleman, James Samuel (1926–95); Cooperation: Sociological Aspects; Exchange: Social; Game Theory; Influence: Social; Integration: Social; Interorganizational Relationships and Networks; Macrosociology–Microsociology; Network Analysis; Networks and Linkages: Cultural Aspects; Rational Choice Theory in Sociology; Social Capital; Social Networks and Gender; Solidarity, Sociology of; Structure: Social; Trust, Sociology of

Bibliography

Baerveldt C, Snijders T A B 1994 Influences on and from the segmentation of networks: Hypotheses and tests. *Social Networks* **16**: 213–32

Bienenstock E J, Bonacich P 1992 The core as a solution to exclusionary networks. *Social Networks* **14**: 231–44
Blau P 1964 *Exchange and Power in Social Life*. Wiley, New York
Burt R S 1992 *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA
Coleman J S 1990 *Foundations of Social Theory*. The Belknap Press, Cambridge, MA
Doreian P, Teuter K, Wang C 1984 Network autocorrelation models: Some Monte Carlo results. *Sociological Methods & Research* **13**: 155–200
Everett M G, Borgatti S P 1994 Regular equivalence: General theory. *Journal of Mathematical Sociology* **19**: 29–52
Flache A, Macy M W 1996 The weakness of strong ties: Collective action failure in highly cohesive groups. *Journal of Mathematical Sociology* **21**: 3–28
Flap H D 1999 Creation and returns of social capital. A new research program. *La Revue Tocqueville* **20**: 1–22
Freeman L C 1978 Centrality in social networks—conceptual clarification. *Social Networks* **1**: 215–39
Friedkin N E, Johnsen E C 1997 Social positions in influence networks. *Social Networks* **19**: 209–22
Granovetter M S 1973 The strength of weak ties. *American Journal of Sociology* **78**: 1360–80
Harary F, Norman R Z, Cartwright D 1965 *Structural Models. An Introduction to the Theory of Directed Graphs*. Wiley, New York
Holland P W, Leinhardt S 1975 Local structure in social networks. In: Heise D R (ed.) *Sociological Methodology—1976*. Jossey-Bass, San Francisco, pp. 1–45
Holland P W, Leinhardt S 1977 A dynamic model for social networks. *Journal of Mathematical Sociology* **5**: 5–20
Homans G 1950 *The Human Group*. Harcourt Brace, New York
Laumann E O, Knoke D, Kim Y H 1987 Event outcomes. In: Laumann E O, Knoke D (eds.) *The Organizational State. Social Choice in National Policy Domains*. The University of Wisconsin Press, Madison, WI, pp. 343–73
Lazega E, Van Duijn M 1997 Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data. *Social Networks* **19**: 375–97
Leenders R T A J 1996 Evolution of friendship and best friendship choices. *Journal of Mathematical Sociology* **21**: 133–48
Leenders R T A J 1997 Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. In: Doreian P, Stokman F N (eds.) *Evolution of Social Networks*. Gordon and Breach, Amsterdam, pp. 165–84
Lin N 1982 Social resources and instrumental action. In: Marsden P V, Lin N (eds.) *Social Structure and Network Analysis*. Sage, London, pp. 131–45
Lindenberg S 1998 Solidarity: Its microfoundations and marco dependence: A framing approach. In: Doreian P, Fararo T J (eds.) *The Problem of Solidarity: Theories and Models*. Gordon and Breach, Amsterdam, pp. 61–112
Moreno J L 1934 *Who Shall Survive?: Foundations of Sociometry, Group Psychotherapy, and Sociodrama*. Nervous and Mental Disease Publications, Washington DC
Pappi F U, Henning C H C A 1998 Policy networks: More than a metaphor? *Journal of Theoretical Politics* **10**: 553–75
Raub W, Weesie J 1990 Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology* **96**: 626–54
Snijders T A B 1981 The degree variance: An index of graph heterogeneity. *Social Networks* **3**: 163–74
Snijders T A B 1996 Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology* **21**: 149–72
Stokman F N, Van den Bos J M M 1992 A two-stage model of policy making: With an empirical test in the U.S. energy policy domain. In: Moore G, Whitt J A (eds.) *Research and Society, Vol. 4: The Political Consequences of Social Networks*. JAI Press, Greenwich, CT, pp. 219–53
Wasserman S, Faust K 1994 *Social Network Analysis. Methods and Applications*. Cambridge University Press, Cambridge, UK
Wasserman S, Pattison P 1996 Logit models and logistic regressions for social networks. I. An introduction to Markov graphs and *p**. *Psychometrika* **61**: 401–25
Wasserman S, Weaver S O 1985 Statistical analysis of binary relational data: Parameter estimation. *Journal of Mathematical Psychology* **29**: 406–27
Wellman B, Berkowitz S D (eds.) 1997 *Social Structures: A Network Approach*. JAI Press, Greenwich, CT
Willer D (ed.) 1999 *Network Exchange Theory*. Praeger, Westport, CT

F. N. Stokman

Neumann, John von (1903–57)

John von Neumann was born Janos Neumann on December 28, 1903. He was the eldest son of Miksa (Max) Neumann and Margit (Margaret) Kann. The family prospered in the last decades before World War I. Max Neumann was a lawyer who was well known in banking circles, and he was later a banker. He was subsequently ennobled, hence the 'von' in John's name.

Max Neumann instilled his ardent love of high culture into his children. In the atmosphere in which John spent his childhood, his phenomenal precociousness in mathematics was soon discovered. His gymnasium teacher pointed out to John's father that there was no point in John's studying mathematics in school. Mature mathematicians were engaged as private tutors. His first paper, written jointly with one of them, appeared in 1922. John was recognized as a professional mathematician before he was 18.

After receiving his *matura*, von Neumann enrolled as a mathematics student in the University of Budapest. However, he did not attend classes and he only appeared at the end of each semester to take examinations. He was simultaneously enrolled first at the University of Berlin, Germany then in the *Eidgenossische Technische Hochschule* in Zurich, Switzerland. He received his doctorate in mathematics from the University of Budapest and the degree of *Diplomingenieur* in chemistry from the *Eidgenossische Technische Hochschule* almost simultaneously. The study of chemistry was probably motivated by the increasing importance of its applications in military technology. Toward the end of his life, von Neumann's most

brilliant contributions were closely related to the development of weaponry.

Von Neumann's professional career falls roughly into three periods. He spent five years in Weimar, Germany (the period between the monarchy and the Third Reich) as *Privatdozent* in Berlin and Hamburg. Recognizing the impending Nazification of Germany, he accepted a visiting professorship at Princeton University and emigrated. Subsequently he joined the newly founded Institute for Advanced Studies. The depression decade roughly marks the second period, when he made his most brilliant contributions to mathematics and to mathematical physics. In the last period, emphasis shifted to applied mathematics, particularly to aerodynamics and the physics of explosions and implosions, the development of nuclear weapons, and of electronic computers. His academic affiliation remained with the Institute of Advanced Studies. However the locale of his most intense preoccupation was Washington, where his contributions to US military might were fully appreciated. On one of his last public appearances President Eisenhower presented him with the Medal of Freedom.

In Norman Macrae's biography of von Neumann (1992), the last chapter is aptly entitled 'With Astounding Influence.' The final scene is at the Walter Reed Hospital in Washington.

The great and important now flocked to his bedside. It was known that Johnny's illness might lead to his talking in his sleep, and Colonel Ford was told to place soldiers to ensure that he did not shout out military secrets. Von Neumann died on February 8, 1957.

An obituary by S. Ulam, a close friend and colleague, in the *Bulletin of the American Mathematical Society* lists titles of 125 publications and of 17 abstracts of papers presented to the American Mathematical Society (1958). The list ranges over all major fields of mathematics and some of physics, for example quantum theory and statistical mechanics. Von Neumann's doctoral dissertation was on axiomatization of mathematics based on set theory, a subject he taught at age 11 to 12-year-old future Nobel laureate Eugene Wigner. This branch of mathematics seems to provide access to the way von Neumann viewed the functioning of the human mind at its best.

A prevalent view of mathematics (among non-mathematicians) is that mathematical concepts are essentially abstractions of experiences with the physical world. Von Neumann has on occasions expressed that view; and this seems strange, since his earliest dedication, manifested in his doctoral dissertation, was to the foundations (or axiomatization) of mathematics, which represents, perhaps more than any other branch, a complete break with concrete, sense-induced experiences. All universally recognized mathematical knowledge is gained from deduction alone, which is from chains of reasoning about what one assertion implies about another. Empirics plays no role whatsoever in this process.

David Hilbert, arguably the most outstanding mathematician of the nineteenth century, 'bequeathed' to mathematicians 23 problems to be solved in the course of the twentieth century. One of them was the 'axiomatization' of all mathematics, that is, a follow-up of Euclid's feat in geometry. This task appeared challenging, because of difficulties that arose in the highly original work of Georg Cantor on the arithmetic of so-called 'transfinite' numbers. It seemed as if paradoxes or, perhaps, even gross-contradictions appeared in that theory. The only way to clear them up was to axiomatize the entire realm of mathematics on the foundation of its most basic concept—that of a set, thus reducing all mathematical propositions to statements about relationships among sets.

Hilbert, who was markedly enthusiastic about Cantor's work, was convinced that such axiomatization could be accomplished. Other distinguished mathematicians, notably Hermann Weyl and L. E. J. Brouwer, thought it could not and recommended a revision of mathematics that would 'delegitimize' some of its most important branches (mostly dealing with continua). Von Neumann sided with Hilbert (who, in turn, recognized young von Neumann as a rising star) and undertook the task, which Bertrand Russell and Alfred North Whitehead had undertaken at the turn of the century and failed. Von Neumann also failed. In 1931 the problem was solved, if an impossibility proof can be regarded as a solution, by Kurt Gödel, who showed that any mathematical system of complexity comparable to that of elementary arithmetic or greater must contain propositions that cannot be shown to be either true or false. Yet it was von Neumann's unsuccessful effort to solve the problem bequeathed by Hilbert that in a way helped him lay the theoretical foundations of computer technology. He adopted Gödel's cataloguing of logical statements as numbers by cataloging computer instructions as numbers.

Von Neumann's preoccupation with the foundations of mathematics suggests that his philosophy of scientific cognition (it seems he recognized no other kind) was based on 'reductionism.' The merging of conceptual repertoires and methodologies of two disciplines, whereby one with fewer concepts absorbs the one with more, exemplifies reductionism. The best known example is the reduction of chemistry to physics. For example, the fundamental conceptual building blocks of classical chemistry were originally the elements. These differed qualitatively. Their combinations into compounds and, conversely, the break up of compounds into elements were the principal events constituting chemical processes. With the discovery of the periodic table and of the internal structure of atoms, the properties of elements were described in quantitative terms (e.g., numbers of electrons, protons, etc., characterizing the atoms of elements), and the basic events were explainable by physical laws. Thereby, much of chemistry was 'reduced' to physics, a more rigorous science. Eventu-

ally physiology was reduced to biochemistry, parts of psychology to physiology. In economics, reduction of macroeconomics (behavior of large economic systems) to microeconomics, that is, the behavior of *homo economicus* (a postulated 'atom' of economic activity) is in the spirit of reductionism.

Von Neumann's work in this spirit was particularly fruitful. For example the conception of a human brain as a computer is certainly a reductionist step, since computer science is much more rigorous, that is, more mathematical than traditional cognitive psychology or the physiology of the nervous system. In general, the 'mathematization' of any discipline is in the spirit of reductionism. So it was with von Neumann's most important contribution to the social sciences, the discipline of which he undisputedly laid the foundation, namely, the theory of games.

Game theory can be clearly seen as a branch of decision theory in which more than one rational decision-maker (or 'actor') are involved. The prototypes of games examined in the theory comprise so-called 'games of strategy' such as chess, go, bridge, poker, etc., that is, contests in which facility in formal reasoning is decisive. Of special interest are 'two-person constant sum games,' in which the sum of the payoffs to the respective players is constant in all outcomes: the more one wins, the less the other wins or the more he loses. Hence, the interests of the two players are diametrically opposed.

It is in the spirit of reductionism that von Neumann built up the theory of economic behavior, on a foundation provided by the theory of the two-person constant sum game. In his treatment of so-called 'cooperative' n-person games ($n > 2$), some of the players may form a coalition and coordinate their strategies so as maximize their joint payoff. It is then in the interest of the remaining players to form their own coalition. If the n-person game is constant sum, then the game between the two coalitions is essentially a two-person constant sum game. If the sum is not constant, a 'dummy' player can be added whose winnings or losses in each outcome compensate for the differences, making the game constant sum. In this way, a theory of the 'cooperative' n-person game reduces to a game of completely opposed interests.

Von Neumann's great achievement was the proof of the so-called 'minimax theorem.' It asserts that every two-person constant sum game, in which a finite number of strategies are available to each player has a 'solution,' expressible as the outcome resulting from an independent choice by each player of a particular strategy, pure or mixed. A pure strategy is essentially an adoption of a plan to make a particular 'move' among those permitted by the rules in each of the possible situations that may occur in the course of the game. A mixed strategy is a probabilistic mixture of pure strategies. This concept was already introduced by E. Borel in 1921, but illustrated only in special cases. Von Neumann proved the general existence theorem in 1928. *Theory of Games and Economic Behavior*, co-authored with Oscar Morgenstern (1944) very soon attracted the attention of the American military establishment as a possible foundation of a mathematically precise rational conduct of conflict.

The minimax corresponds to what military strategists call the 'worst-case scenario,' the assumption that the enemy doesn't necessarily act so as to achieve the best possible outcome for himself but rather to impose the worst possible outcome on you. In situations analogous to the two-person constant sum game the two goals are identical, since confrontation of enemies is most naturally modeled by a constant sum game. The situation in so called 'neoclassical' economics is similar. At the foundation of its reductionist theoretical framework is 'methodological individualism,' which P. H. Wicksteed (1933), perhaps to avoid the pejorative connotations of 'egoism,' called 'nontuism'—the overriding goal of maximizing one's own payoff. As a result of impersonal market forces (Adam Smith's 'invisible hand') an economic system so modeled is shown with mathematical rigor to tend toward an 'equilibrium,' assuming this to be the best attainable condition for everyone concerned. In this way von Neumann's minimax theorem appeared to provide not only an idealized model of 'rationally conducted conflict' but also a justification of a Hobbsian 'competition of every one with every one else.'

To the extent that some references are made in military doctrines to human behavior, these can be arguably included in the subject matter of the social sciences. Economics is unequivocally regarded as a branch of social science, indeed the most theoretically rigorous one. Thus, von Neumann's and Morgenstern's treatise can be justifiably celebrated as a most impressive contribution to social science. Paradoxically, an even more impressive contribution can be discerned in the pursuit of answers to questions that were not raised in the treatise but were inspired by it, namely, in the course of analyzing non-constant sum games.

As has been said, the US military establishments quickly recognized the potential of game theory in developing a 'science of rationally conducted conflict.' Generous support of further research was not long in forthcoming. Workers in the field were invited to the think tank established by the Rand Corporation. Attention was focused on experimental verification of theoretical predictions generated by game theory. Note that laboratory experimentation with a real economic system is practically impossible, while such experimentation on people playing specially designed games is easy, cheap and, most importantly, reproducible.

The results of the early experiments with non-constant sum games were disappointing. They in no way confirmed systematically the predictions of game theory based on the minimax principle. Neither did

they systematically refute them in a way that would suggest a generalization of the theory to explain the discrepancy. However, a result was observed demonstrating a 'serendipity' effect—finding something more valuable than what was originally sought. Namely, a game was discovered that was soon called Prisoner's Dilemma, now a byword in the vocabularies of economics, social psychology, sociology, political science, management science, peace research, peace education, and philosophy. The game was discovered by M. M. Flood (1952) and M. Dresher and christened by A. W. Tucker, followers of von Neumann's leads. The name of the game stems from the following anecdote.

Two men are arrested on suspicion of burglary. There is not enough evidence to convict them unless at least one confesses. They can, however, be convicted of a lesser crime—possession of stolen goods, which carries a lighter sentence. If only one confesses, he is not prosecuted, while the other is given the most severe sentence. If both confess their sentence is reduced. If neither confesses, both get the lightest sentence (for the lesser crime). It turns out that it is in the interest of each to confess regardless of whether the other confesses or not. However, if both confess, both are given longer prison terms than if neither confesses.

The paradox brings out the distinction between individual and collective rationality. There are innumerable situations where this distinction is dramatically demonstrated. If a fire breaks out in a crowded theater, it is in every one's interest to get to an exit as quickly as possible. If all act 'rationally,' all may perish. It is in the interest of each fishing fleet to maximize its catch. If each acts 'rationally' there may soon be no fish to catch. It is in each of two hostile powers' interest to be stronger than the other is. An escalating arms race may lead to a disastrous war, which both wanted to avoid.

There is no indication that von Neumann was aware of the 'paradox of rationality' revealed by Prisoner's Dilemma, although the basic idea was expressed over a century before the publication of *Theory of Games and Economics Behavior* (1944) by von Neumann and Morgenstern. Having determined the equilibrium point of a market controlled by a duopoly, A. A. Cournot wrote in 1838:

> … si les deux producteurs *s'etaient entendue* [emphasis in the original], la valeur de ... [price calculated on the basis of the agreement] aurait assigné à chacun un revenu plus grand que celui qu'ils obtiendront avec la valeur de … [price determined by the duopolists independently] Cournot 1838.

According to Merril Flood, who called von Neumann's attention to the paradox, the latter admitted that the Prisoner's Dilemma was a challenge to the minimax principle but did not take seriously the results of the Flood-Dresher experiment, which did not support the minimax (equilibrium) solution.

Two years after von Neumann's death, systematic experimentation with games similar to the Prisoner's Dilemma began in the USA and spread worldwide. It gave rise to a current contrary to that originating in the so-called 'rationality paradigm' which sought to reduce all aspects of human behavior to manifestations of self-interest and all 'rational' behavior to cost-benefit analysis. Von Neumann evidently thought within the framework of that paradigm. In his opinion, the fact that 40,000–50,000 annual traffic fatalities are tolerated in the USA indicates that this is an acceptable price to pay for the privilege of driving private automobiles. In the same spirit, he advocated acceleration of the arms race, on the grounds that if the Soviet Union got ahead, it would be more likely to launch a preventive nuclear strike against the USA. Apparently the opposite hypothesis (that the party behind might be more likely to launch a preventive war) did not occur to him.

How, then, can the suggestion be justified that the theory of games launched a movement in the social sciences transcending the reductionist model of the human condition? Recall the seemingly unassailable finality of the Newtonian paradigm resting on the metaphysics of a Euclidian three-dimensional space, independent flow of time, and an all-pervasive ether, defining an absolute system of coordinates. This paradigm was overthrown in the course of a few years by the relativity and quantum revolutions. The point is that the success of non-Newtonian physics was due to the clear axiomatization (except for implicit metaphysical postulates, such as independence of space and time) of the *Newtonian* paradigm. It was clear just what 'axioms' had to be dropped, generalized, or supplanted by others to account for phenomena inexplicable within the old paradigm. For example, the (implicit) postulate of an absolute system of coordinates was supplanted by the (explicit) postulate of the constancy of the speed of light with respect to all systems of coordinates. In quantum mechanics, the independence of physical events from acts of observation (implicit in the Newtonian system) was supplanted explicitly by the Uncertainty Principle, incompatible with that assumption. Thus, the 'revolution' triggered by the Prisoner's Dilemma was fueled by the clear axiomatization of the rationality paradigm as manifested in von Neumann's and Morgenstern's theory of games. In particular, Pareto-optimality (maximization of common rather than individual gain) supplanted the 'non-tuism' of neoclassical mathematical economics.

Also noteworthy is the 'reflexive' model of man developed by Vladimir Lefebvre (1998), in which the (implicit) assumption of 'non-tuism,' namely that the thought processes of interacting individuals are independent, is supplanted by a 'mirror image' assumption. An actor is assumed to be able to mentally reconstruct another actor's thoughts and feelings and, regarding him as his own 'mirror image,' acts ac-

cordingly. In the light of this model the cooperative solution of the Prisoner's Dilemma and similar games is rigorously deduced.

In this way, Max Weber's conception of 'empathic (sic!) understanding' as an epistemological principle in the social sciences was revived. The dogma of 'value-free' social science was revealed as an analog of the Newtonian axiom of independence of observations from the act of observation. It appears that social science, no matter how rigorous, cannot be value free, because social self-observation plays a crucial part in its epistemology and inevitably generates value-oriented conclusions.

See also: Altruism and Self-interest; Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence in Cognitive Science; Artificial Intelligence: Search; Axiomatic Theories; Decision and Choice: Economic Psychology; Decision and Choice: Paradoxes of Choice; Decision-making Systems: Personal and Collective; Decision Theory: Bayesian; Decision Theory: Classical; Economics, History of; Game Theory; Game Theory and its Relation to Bayesian Theory; Game Theory: Noncooperative Games; Group Decision Making, Cognitive Psychology of; Individual/Society: History of the Concept; Individualism versus Collectivism: Philosophical Aspects; Interest, Psychology of; Military Psychology: United States; Philanthropy, Economics of; Prisoner's Dilemma, One-shot and Iterated; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory: Cultural Concerns; Reduction, Varieties of; Science, Technology, and the Military

Bibliography

Borel E 1921 La théorie des jeux et les équations intégrales à niveau symmétrique. *Comptes rendus Academie des Sciences* **173**: 1304–8

Cournot A A 1838 *Recherches sur les principes mathematiques de la richesse*. Hachette, Paris

Flood M M 1952 *Some Experimental Games. Research Memorandum RM-789*. Rand Corporation, Santa Monica, CA

Heims S J 1982 *John von Neumann and Norbert Wiener*. MIT, Cambridge, MA

Lefebvre V A 1998 Sketch of reflexive game theory. *Proceedings of Workshop on Multi-reflexive Models of Agent Behavior ARL-SR-63*. Army Research Laboratory

Luce R D, Raiffa H 1957 *Games and Decisions*. Wiley, New York

Macrae N 1992 *John von Neumann*. Pantheon Books, New York

Poundstone W 1992 *Prisoner's Dilemma*. Doubleday, New York

Rapoport A, Chammah A M 1966 *Prisoner's Dilemma*. University of Michigan Press, Ann Arbor, MI

Ulam S 1958 John von Neumann. *Bulletin of the American Mathematical Society* **64**(3), Part 2: 1–49

Von Neumann J 1925 Eine Mathematisierung der Mengenlehre. *Acta Universitatas Szegedsienses* **1**: 219–40

Von Neumann J 1928 Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* **100**: 295–320

Von Neumann J 1932a *Mathematische Grundlagen der Quantenmechanik*. Springer, Berlin

Von Neumann J 1932b Zur Operatorenmethode in der klassischen Mechanik. *Annals of Mathematics* **33**: 587–62

Von Neumann J 1966 (posthumous) *Theory of Self-reproducing Automata*. University of Illinois Press, Urbana, IL

Von Neumann J, Fakete M 1922 Über die Lage der Nullstellen gewisser Minimalpolynome. *Berichte des Deutschen Mathematischen Vereins* **31**: 125–38

Von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

Von Neumann J, Burks A W Goldstein H H 1946 Report prepared for US Army Ord. Dept. under Contract W-36-034-ORD-7481

Wicksteed P H 1933 *The Common Sense of Political Economy*. Routledge, London

A. Rapoport

Neural Circuits: Models of Emergent Functions

Information about the external world and internal actions is represented in the brain by the activity of large networks of heavily interconnected neurons, that is, neural circuits. To be efficient, this representation requires neuronal circuits to perform certain functions which cannot be performed by single neurons. In this article, some of the current models of emergent functions of neural circuits are described.

1. Single Neurons and Neural Networks

To some extent, a neural circuit can be viewed as a complex electronic device. Every part of the device (e.g., transistor) may possess rather complex characteristics, yet only the whole device can perform the function for which it is designed. This distinction is even more striking in the case of neural circuits, where the parts of the device, the single neurons, in many cases exhibit a remarkable degree of uniformity. For example, pyramidal cells of the hippocampus, which carry the representation of space, are similar to each other in terms of their anatomical and physiological properties. It appears therefore that a major factor that determines the emerging functional properties of neural circuits is the connections between the neurons in the network. The relation between connectivity and function is the main problem addressed by the models which are described below.

A single neuron exchanges information with other neurons by generating a sequence of sharp repolarizations of its membrane, called action potentials, which propagate along the neuron's axon until they reach the target neurons. Since every action potential has a largely stereotyped shape, it does not contain any specific information about the state of the neuron. Instead, it is believed that all of the information is

represented in the neuron's *spike train*, that is, the sequence of times at which action potentials are emitted. In most of the models, this is further simplified into a *firing rate*, that is, a rate with which a neuron is emitting its action potentials (some of the models however consider the timing of every single action potential to be informative; see, e.g., Hopfield 1995). Once the action potential of a neuron arrives at its destination, its effect can be either increased or decreased depolarization of the target neuron's membrane potential. These two opposing effects are called excitation and inhibition, respectively. In many parts of the brain each target neuron is subject to excitation and inhibition effects from thousands of other neurons, and responds by generating its own spike train in a particular time window (Abeles 1991).

The amplitudes of the excitatory and inhibitory influences differ widely and define the strengths of the corresponding connections. It is the distribution of the connection strengths across the neural network which enables it to perform a particular function. Since the strength of the connections is not a fixed quantity but can be modified as a result of development and/or experience, this mechanism also allows for the acquisition of a certain function with experience, which therefore constitutes the neural basis of learning and memory (see, e.g., *Implicit Learning and Memory: Psychological and Neural Aspects*)

2. Main Network Paradigms

Understanding the emergence of functions from connectivity in neural networks is a subject of active research. This research has resulted in the formulation of numerous models, each characterized by various degrees of generality and biological realism. Since the emphasis is put on the role of connectivity, a highly simplified description of single neurons is adopted in most of the models. Usually, a single neuron is modeled as an input–output element which integrates excitatory and inhibitory influences from other neurons in the network and transforms them directly to its firing rate

$$I_i(t) = \sum_j J_{ij} R_j(t)$$
$$R_i(t+\mathrm{d}t) = g(I_i(t) + I_i^{ext}(t)) \qquad (1)$$

where R_i stands for the rate of the i-th neuron in the network and J_{ij} denotes the strength of the connection between neurons j and i. The total input which the neuron receives is divided in two components: I_i, which is due to the influences coming from the other neurons in the same network, and I_i^{ext}, resulting from the activity in other areas of the brain. Finally, the input is translated into the neuron's rate at the next time step by a gain function $g(I)$. Typically, the gain function is a continuous, monotonically increasing function of its input, such as, for example, a sigmoid or a threshold-linear function. In some of the more realistic models, a spike train is produced as an output. This approach is broadly called *connectionism* (McClelland et al. 1986).

As an individual neuron can be considered as an input–output element transforming inputs into output rate, so the whole network can be viewed as transforming its external inputs, I_i^{ext}, into the firing rates of the neurons. Specifying the rates of all of the neurons in the network defines the corresponding *network state* of activity. The additional complexity brought by the presence of the set of connection strengths however makes the global transformation from external inputs to the network states much richer than the one occurring at the single neuron level. In particular, the network can even display the dynamic behavior which is to a large degree autonomous and is only partially influenced by external inputs. It is this property which enables new functional behaviors to emerge in neural circuits.

Some of the major network paradigms that were formulated in the framework of connectionist modeling are described below. They are classified according to the architecture of the connections. In the *feed-forward networks*, the neurons can be divided in an ordered sequence of distinct layers, such that all of the connections originating from the neurons of a particular layer only target the layers which are higher up the stream. Importantly, neither feed-back connections from upper to lower layers, nor recurrent connections within the layers are allowed. The lowest layer usually contains the neurons which are receiving the external inputs, while the highest layer is the one where the output neurons are located. Another type of network which will be described is the *recurrent network* where information from a single neuron can return to it via the connections which form so-called *feedback loops*. The division of the network into distinct layers is therefore not possible, and the whole network must be considered as one layer in which both input and output neurons are located.

2.1 Feed-forward Neural Networks

In these networks, information flows in one direction from the input layer to the output layer, without feedback loops. This arrangement of connections does not allow the network to exhibit an autonomous dynamics, and the activity of output neurons is therefore completely determined by the inputs in a finite number of time steps. In other words, a feed-forward network realizes a certain mathematical function on its input

$$O = f(I) \qquad (2)$$

The precise form of this function depends on the architecture of the network (the number of layers, the number of neurons in each layer, etc.) and the

distribution of strength of connections between the layers. The simplest feed-forward network is a famous *perceptron*, which is also the first studied neural network (Rosenblatt 1962). It consists of an input layer and a single output neuron. The mathematical function which it realizes has, therefore, the following form

$$o = g\left(\sum_i J_i I_i\right) \tag{3}$$

It was shown that only a limited family of functions can be realized with the simple perceptron, but adding intermediate layers of neurons expands it to include practically every mathematical function. Effective learning algorithms for feed-forward networks have been developed, the most widely used being the so-called back propagation of errors. This allowed numerous practical applications of feed-forward networks.

Another prominent feed-forward network paradigm, which also contains some elements of recurrent processing, is the *self-organizing feature map*, developed by Kohonen (1984). It attempts to account for the ubiquitous presence of feature maps in the neocortex of the mammalian brain, ranging from the retinotopic map in the visual cortex to the topographic representation of the body in the somato-sensory cortex. From the functional perspective, the network allows extraction of low-dimensional features from the set of multidimensional inputs, and maps them to an array of output neurons arranged in a low-(typically one or two-) dimensional sheet in a topographic manner, such that the distance relations in the space of the inputs is represented as distance relationships between the neurons in the output layer.

This is achieved by numerous presentations of inputs and gradual update of feed-forward connections. At every presentation of the (high-dimensional) input, I, the excitation received by each neuron, occupying the site r at the output layer, is given by the scalar product $\sum_i J_{ri} I_i = J_r I$, where J_r is a vector of the neuron's connection strengths. The crucial step of the learning algorithm is finding the neuron which receives the maximum excitation for a given input. After this is done (let's says s is the site with the maximum excitation), feed-forward connections are modified in a Hebb-like manner, but only for the output neurons which are located in the neighborhood of s

$$\delta J_r = \varepsilon I h(r-s) \tag{4}$$

where the function $h(r-s)$ is such that it selects the neighboring sites which are participating in the current update, for example

$$h(r-s) = \exp\left(-\frac{\|r-s\|^2}{\sigma^2}\right) \tag{5}$$

This learning rule guarantees that neighboring neurons tend to update their connections in response to similar inputs, and thereby develop the selectivity to similar features. Self-organizing feature maps have been applied successfully to model the appearance of ocular dominance and orientation selectivity maps in the primary visual cortex (Erwin et al. 1995).

2.2 Recurrent Neural Networks

In these network models, the emphasis is put on the functions which emerge from the recurrent connections between the neurons in a single layer. As opposed to feed-forward networks, recurrent networks may have nontrivial autonomous dynamics, since activity can propagate along the feedback loops independently of the external input. The most famous and well studied model of this type is associative memory (see *Implicit Learning and Memory: Psychological and Neural Aspects*) This paradigm goes back to the classical ideas of Hebb on neuronal ensembles and synaptic plasticity (Hebb 1949), and was reinvigorated in the highly influential paper by Hopfield (1982). It is based on the observation that under some choice of the connections, the network described by Eqns. (1) can sustain various nontrivial activity states without the presence of external inputs. On the intuitive level, this can come about from the mutual excitation within a group of neurons, resulting in the activity reverberating across this group without ever extinguishing itself. On the mathematical level, this property manifests itself in the presence of many stable stationary solutions to the nonlinear Eqns. (l) with different subgroups of neurons having nonzero rates. The corresponding network states are often called *attractors*. Every attractor state can be reached autonomously by the network provided that the initial state has a large enough degree of overlap with it, that is, sufficiently many shared neurons with nonzero rates. This process can be identified with an associative recall of stored memories from initial cues containing only partial information. For instance, a telephone number can be recalled after several first digits are provided. As shown in the original paper by Hopfield, the sufficient condition for the network to have stationary attractors is the symmetry of the connections, that is, the condition that any pair of neurons has reciprocal connections of equal strength. More importantly, for any 'random' choice of putative memory states, R_i^μ, one can propose the distribution of connections which makes them attractor states of the network

$$J_{ij} = \sum_\mu (R_i^\mu - \langle R \rangle)(R_j^\mu - \langle R \rangle) \tag{6}$$

where summation is performed over the set of putative memories. This construction also suggests a way in which connections could be modified during the

learning. For each new memory, the corresponding state should be imposed on the network by external inputs.

Then, for every pair of neurons, the strength of the corresponding connection is modified by adding the product of the terms, each of which depends on the activity levels of the respective neurons. At the end, the distribution of the connection strengths depends on all of the memories, but does not depend on the order in which they were stored in the network. This *learning algorithm* was first envisioned by Hebb and subsequently gained neurophysiological confirmation. Although some of the features of the Hopfield model, such as the symmetry of connections, are not biologically plausible, subsequent studies showed that they can be relaxed without destroying the performance of the model (Amit 1989, Hertz et al. 1991).

More recently, associative memory networks were directly confronted with experimental results on delayed memory tasks in monkeys. In these experiments, the task of the monkey is to compare the sample and the match stimuli which are separated by the delay period of several seconds. In several cortical areas, neurons have been found which exhibited elevated firing rate during the delay period, which was selective for the sample stimulus, but only if this stimulus was seen many times in the past during training. These experiments provided a wealth of quantitative information about the persistent activity patterns in the brain, which were reasonably reproduced in the model (Amit 1995).

Nonstationary states of recurrent networks can also be of functional significance. In particular, introducing temporal delays in the connections, one can obtain the solutions where the network exhibits a set of transitions between the attractors stored in a temporal sequence (Kleinfeld and Sompolinsky 1988). This property can be used, for instance, in storing a sequence of motor commands associated with a particular action. On the more general level, it was shown that recurrent networks can learn all kinds of time-dependent patterns, both periodic and aperiodic.

Another class of recurrent neural networks was developed in order to model the representation of continuous features rather then discrete sets of memories. In particular, coding of the direction of hand movements in the motor cortex, orientation of stimuli in the visual cortex, and spatial location of the animal and direction of gaze were all proposed to involve the networks with *continuous sets of attractor states*. This property of the network arises from a special structure of connections, in which neurons with preference for similar features are strongly connected. For example, for neurons in the visual cortex which prefer stimuli oriented at the angles $\theta_i(-90° < \theta_i < 90°)$, the distribution of connection strengths was proposed to be of the form

$$J_{ij} = J_2 \cos(2 * (\theta_i - \theta_j)) - J_0 \qquad (7)$$

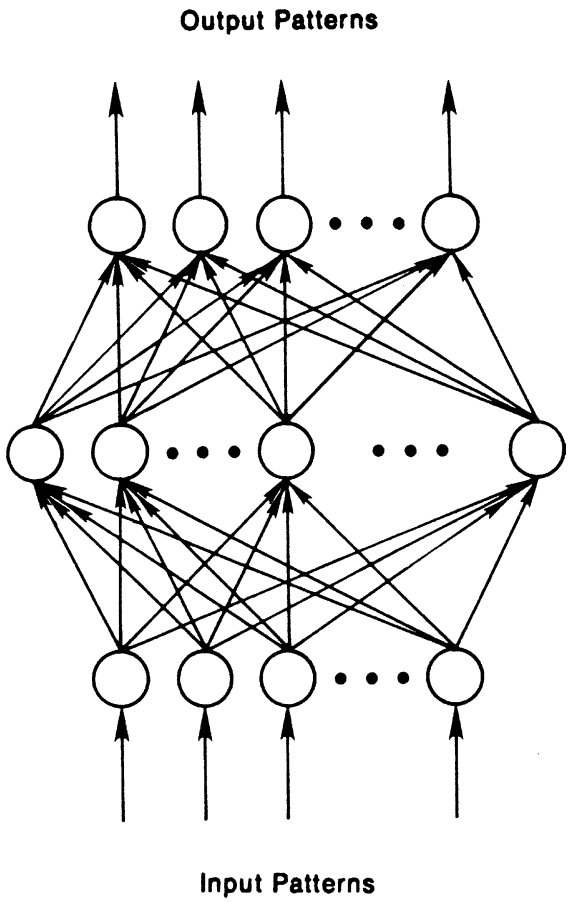

Figure 1
A multilayer, feedforward neural network. Information is coming to the input layer and then flowing towards the output layer without encountering feedback loops (reproduced by permission of the Center for the Neural Basis of Cognition from *Parallel Distributed Processing: Exploration in the Microstructures of Cognition, Vol. 1: Foundations*, 1986, p. 320)

This structure of connections combines the effects of synergy between neurons with similar preferences (first, excitatory term) and global competition for activity (second, inhibitory term). As a result, the network possesses a set of marginally stable attractors, such that in every attractor the activity is localized on a group of neurons with similar preferences, the rest of the network being in the quiet state. External input serves to evoke a particular attractor out of a set possessed by the network, for example, oriented visual stimulus evokes the activity profile centered at the corresponding neuron. The precise shape of the profile of activity is determined by properties of single neurons (responsiveness) and parameters of interactions, but largely is independent on the variations in

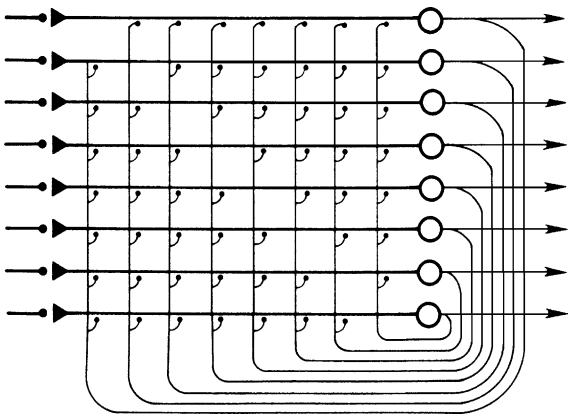

Figure 2
A recurrent neural network, consisting of a small ensemble of eight processing units. In addition to external inputs and outputs, each unit has feedback connections to other units in the network, as indicated by the branches of the output lines that loop back onto the input lines leading into each unit (reproduced by permission of the Center for the Neural Basis of Cognition from *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models*, 1986, p. 174)

the input which evokes it. This property of networks with continuous attractors may account for such peculiar effects as invariance of orientation tuning to the contrast of the stimuli (Sompolinsky and Shapley 1997), persistence of space-related activity in hippocampus in total darkness (Tsodyks 1999), and stability of gaze direction in the darkness (Seung 1996), which are difficult to explain in the more traditional, feed-forward models (see Figs. 1 and 2).

See also: Artificial Neural Networks: Neurocomputation; Cognitive Functions (Normal) and Neuropsychological Deficits, Models of; Linear Algebra for Neural Networks; Network Models of Tasks; Neural Networks and Related Statistical Latent Variable Models; Neural Networks: Biological Models and Applications; Neural Networks, Statistical Physics of; Neural Systems: Models of Behavioral Functions

Bibliography

Abeles M 1991 *Corticonics*. Cambridge University Press, New York

Amit D J 1989 *Modeling Brain Function*. Cambridge University Press, New York

Amit D J 1994 The Hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioural Brain Science* **18**: 617–26

Erwin E, Obermayer K, Schulten K 1995 Models of orientation and ocular dominance columns in the visual cortex: A critical comparison. *Neural Computation* **7**: 425–68

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Hertz J, Krogh A, Palmer R G 1991 *Introduction to the Theory of Neural Computation*. Addison-Wesley, New York

Hopfield J J 1982 Neural networks and physical systems with emergent collective properties. *Proceedings of the National Academy of Science USA* **79**: 2554–8

Hopfield J J 1995 Pattern recognition computation using action potential timing for stimulus representation. *Nature* **376**: 33–6

Kleinfeld D, Sompolinsky H 1988 Associative neural network model for the generation of temporal patterns. Theory and application to central pattern generators. *Biophysics Journal* **54**: 1039–51

Kohonen T 1984 *Self-organization and Associative Memory*. Springer, Heidelberg, Germany

McClelland J L, Rumelhart D E, PDP Research Group (eds.) 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA

Rosenblatt F 1962 *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan, Washington, DC

Seung H S 1996 How the brain keeps the eyes still. *Proceedings of the National Academy of Science USA* **93**: 13339–44

Sompolinsky H, Shapley R 1997 New perspectives on the mechanisms for orientation selectivity. *Current Opinion in Neurobiology* **7**: 514–22

Tsodyks M 1999 Attractor neural network models of spatial maps in hippocampus. *Hippocampus* **9**: 481–9

M. Tsodyks

Neural Development: Mechanisms and Models

How a single cell develops into a richly structured organism, capable of processing highly complex information about the world around it, is one of the most fascinating questions of the twenty-first century. A crucial step in the development of such intelligent organisms is the formation of the nervous system. Billions of neurons must be generated, move to appropriate locations, become specialized to perform particular tasks, establish connections with the correct set of other neurons, and then sculpt their initial connectivity patterns into precisely the right configurations to subserve sophisticated information processing tasks. The last years of the twentieth century have seen a dramatic increase in our knowledge of the genes, molecules, and patterns of neural activity that control these key developmental events. Theoretical models can help organize this data, provide rigorous analyses of the key mechanisms at work, and make novel and experimentally testable predictions.

1. Overview of Neural Development

The first stage of neural development (Gilbert 2000) begins in humans three weeks after conception with the differentiation of the neural plate, a sheet of cells on the surface of the embryo. This folds in on itself to form the neural tube and the neural crest. The neural tube gives rise to the central nervous system and parts of the peripheral nervous system, while the neural crest gives rise to parts of the peripheral nervous system and some non-neural tissues. In the second stage of neural development the neural tube expands as cells proliferate and migrate to their final destinations. Cells become specified as particular types of neurons or glia by both genetic programs and factors within their environment. In stage three the initial connectivity between neurons is established. Axons elongate and are guided to their appropriate targets by a variety of molecular cues. When they reach their targets, a synapse forms between the tip of the axon and the local region of the dendrite it contacts. By the end of this stage the large-scale structure of the nervous system is established. In the fourth stage this initial crude architecture is refined by cell death, and by the gradual modification of synaptic strengths based on patterns of neural activity. These patterns of activity can be both transduced from the environment via the sense organs and generated internally via spontaneous activity. This activity molds the brain's synapses to best match the structure of the environment. It should be noted that, although these four stages have here been presented as distinct, in reality they overlap and intermingle. In addition, experimental evidence increasingly suggests that many of these processes continue to occur to some extent in the adult brain: not just the continued modification of synaptic strengths within a fixed architecture, but also the generation of new neurons, their differentiation and migration, and the formation of new connections. The majority of computational modeling work has so far been focused on understanding stage four, activity-dependent refinement of connections. There are no models addressing stage one (neural tube formation), and only very few addressing stages two (migration and differentiation), and three (axon guidance and synaptogenesis). These will be briefly discussed first, before concentrating on models of stage four.

2. Models of Pattern Formation and Regionalization

A basic problem at stage two of development is to understand how an initially spatially uniform array of cells becomes spatially nonuniform, with different neurons committed to different fates. A very influential class of mathematical models of this type of general pattern formation process are 'reaction–diffusion' systems, first proposed by Turing (1952), and subsequently developed by Meinhardt (1982). In the simplest version, each cell in the array is imagined to be continually releasing two chemicals, an 'activator' and an 'inhibitor,' which diffuse freely through the array. The activator causes cells to produce both more activator and more inhibitor, while the inhibitor causes cells to produce less activator. If the inhibitor diffuses faster than the activator, the end result is the formation of a stable, non-uniform pattern consisting of small regions with high concentrations of activator surrounded by regions with high concentrations of inhibitor. Although this is mathematically elegant and has the potential to explain many aspects of biological pattern formation (Meinhardt 1982 Murray 1993), there are actually very few biological systems where reaction–diffusion has been definitively established as the pattern formation mechanism employed. In fact, experimental evidence increasingly suggests that a rather different process may be more important for regional specification of the nervous system. Homologues of many of the genes that control nervous system regionalization in mammals were first identified in the formation of the segmented bodies of fruit flies. Recent molecular data suggests that, in fruit flies, an initial anterior–posterior molecular gradient is converted into differential levels of expression of particular genes via different thresholds for their activation. This leads to differential levels of expression of further genes, and so on until a complex mosaic of regions is formed. This was first discussed theoretically as a mechanism for regionalization by Wolpert (1969), and it seems highly likely that these types of mechanisms are at work in the regionalization of the mammalian nervous system (Rubenstein and Beachy 1998) rather than reaction–diffusion. Computational modeling has only recently begun to address how complex networks of interacting proteins can lead to the formation of spatial patterns (e.g., Reinitz and Sharp 1995).

3. Models of Axon Guidance

A basic problem at stage three of development is to understand how an extending axon navigates over long distances through a complex molecular environment to find an appropriate target. Among the cues involved are molecules that promote and inhibit axonal growth, and molecules that pull axons in particular directions when they are present in a concentration gradient (Tessier-Lavigne and Goodman 1996). Axons detect the distribution of molecules in their environment via a dynamic sensing device at their tip called the growth cone (see Fig. 1). This draws the axon up or down gradients of molecules such as netrin-1 by sensing a concentration difference across its spatial extent. Crucial to this process are filopodia (see Fig.1), long and dynamic protrusions from the growth cone. Modeling work has approached the growth cone in several different ways, attempting to understand the minimum gradient steepness that

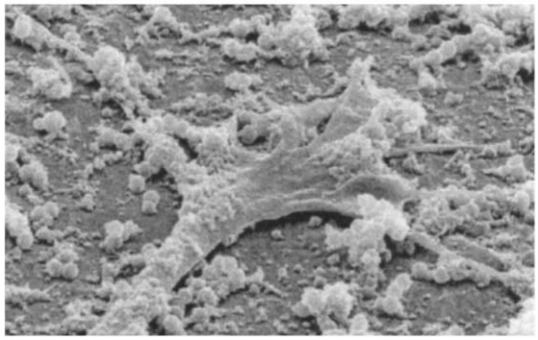

Figure 1
Electron micrograph of a growth cone at the end of an axon (here the growth cone is growing over a surface irregularly covered with coated beads). The long finger-like protrusions are filopodia. Growth cones are typically about 0.01 mm across. (Source: Rosentreter et al. 1998, *Journal of Neurobiology* **37** 541–62, ©John Wiley and Sons)

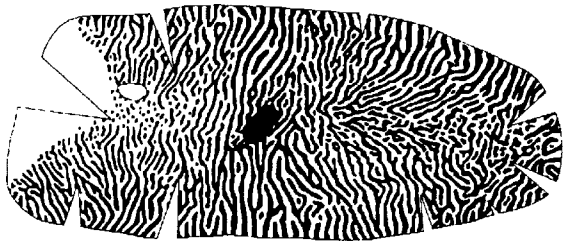

Figure 2
The pattern of ocular dominance columns from the primary visual cortex of a macaque monkey. White represents regions of cortex dominated by input from one eye, black the other eye. Each individual (including each human) apparently has a different ocular dominance pattern that is unique as a fingerprint. The width of individual columns is 0.5–1 mm. (Source: Levay et al. 1985 *Journal of Neuroscience* **5** 486–501, © Society for Neuroscience. Reprinted by permission of Wiley–Liss, Inc., a Subsidiary of John Wiley and Sons, Inc.)

growth cones can detect (Goodhill and Urbach 1999), movement of filopodia (Buettner 1995), how filopodia could be generated as the result of a reaction–diffusion mechanism inside the growth cone (Meinhardt 1999), and how molecules released by growth cones themselves could help guide them to their targets (Hentschel and Van Ooyen 1999). Besides modeling the growth cone, computational researchers have also considered the gradient shapes that could be present *in vivo* and how these shapes constrain the maximum distance over which axons could be guided (Goodhill 1998). A very popular system for studying axon guidance by gradients is the growth of retinal axons into the optic tectum to form a topographic map. Sperry (1963) suggested that this process could be implemented by a matching process between molecular gradients in the retina and the tectum. Following Sperry many experiments were performed to test this hypothesis, and a number of theoretical models were proposed based on concepts of both gradient matching and competition between retinal axons for space in the tectum (e.g., Prestige and Willshaw 1975). More recently gradients of Eph receptors have been identified in the retina, and gradients of their ligands the ephrins in the tectum, which could be a molecular substrate for Sperry's hypothesis (Flanagan and Vanderhaeghen 1998). So far, however, theoretical models have not kept pace with these developments (Goodhill and Richards 1999).

4. *Models of Activity-dependent Refinement of Connections*

A basic problem at stage four of development is to understand how the initial pattern of synaptic strengths established by stage three of development is refined and sculpted by neural activity. It is to this stage that theoretical modeling has so far been most extensively applied, perhaps because of the link between experimentally determined mechanisms of synaptic plasticity and learning algorithms for artificial neural networks (Hertz et al. 1991). The key organizing principle is the Hebb rule (Hebb 1949), which can be summarized as 'cells that fire together wire together.' In other words, the strength of a synapse tends to increase when the firing of the presynaptic neuron is correlated with the firing of the postsynaptic neuron. Hebb's qualitative statement can be made mathematically precise in several different ways.

Perhaps the paradigm example of the application of Hebbian learning rules to understanding neural development is modeling of the development of receptive fields and maps in the mammalian visual system. Visual information passes initially from the retina, via the lateral geniculate nucleus (LGN), to the primary visual cortex (V1). Receptive fields of neurons in V1 can be crudely characterized as preferring (being most strongly excited by) input from one small region of visual space, more from one eye than the other eye ('ocular dominance'), and in the form of an edge or bar of light of a particular orientation (Hubel 1988). Nearby neurons in V1 usually prefer similar positions in space, similar orientations, and have a similar eye preference, but there are also occasional discontinuities where these preferences change abruptly. This leads receptive field properties to be laid out across the cortex in the form of maps with very particular spatial properties (a map of ocular dominance is shown in Fig. 2). These properties of V1 have been appealing to theoretical modelers because they seem to progress to this final, highly organized state from a far less

organized initial state, and this process of 'self-organization' depends on both the presence and correlational structure of neural activity in the retina in a way broadly consistent with the Hebb rule. The first mathematical model of receptive field development in V1 was put forward by von der Malsburg (1973), and since then a large number of variations on the Hebbian theme have been proposed (reviewed in Swindale 1996). A key element of all of the models addressing map formation is the assumption of some kind of lateral interaction between nearby cortical neurons. This is usually short-range excitation balanced against longer-range inhibition, which gives these models a mathematically somewhat analogous form to the reaction–diffusion models that have been applied to spatial pattern formation at much earlier stages of development.

A key feature of these visual maps is their plasticity in response to the correlational structure of input from the environment. The paradigm example is the change in the ocular dominance map when one eye is deprived of visual input. In this case the regions of the cortex dominated by the open eye expand at the expense of the regions dominated by the closed eye (Hubel 1988 Katz and Shatz 1996). These and other experiments indicate that axons from the two eyes compete for space in the primary visual cortex. This can be modeled theoretically in a similar way to the competition for space between individual retinal axons in topographic map formation in the optic tectum (Goodhill 1993). In other words, there are active processes by which the developing nervous system adapts its structure to make the best use of the available environmental input. This also shows up in, for instance, increases in the size of the cortical representation of body surfaces that are highly stimulated (Buonomano and Merzenich 1998). Interestingly, much of this kind of plasticity in the nervous system is limited to certain temporal windows in development called critical periods. For instance, the critical period for the plasticity of ocular dominance columns in kittens is between three and six weeks after birth: changing the visual input before or after this time has relatively little effect. Critical periods also exist for the development of other parts of the nervous system, but what controls the timing of critical periods is unknown. Possibilities include a simple genetically-controlled gating mechanism that allows plasticity only for a limited period, or a less rigid mechanism whereby the dynamics of the learning algorithm naturally lead to a state that is relatively stable, as in some theoretical models (e.g. Linsker 1986).

However, despite the presence of critical periods, the structure of the nervous system continues to evolve for a very long time after the initial processes of development are complete. For instance, a recent study has shown continued growth in the brains of children up to at least 15 years old (Thompson et al. 2000). Although it used to be thought that the neurons born initially during development are the total complement available, recent research indicates that new neurons may be being born in specialized regions and integrated into the nervous system throughout life (Gage 2000). Theoretical modeling has barely begun to address such phenomena. Neural development will clearly continue to be an expanding frontier of scientific enquiry for some time to come.

See also: Artificial Neural Networks: Neurocomputation; Brain Development, Ontogenetic Neurobiology of; Cognitive Development in Infancy: Neural Mechanisms; Computational Neuroscience; Connectionist Models of Development; Hebb, Donald Olding (1904–85); Neural Networks: Biological Models and Applications; Neural Plasticity; Neural Plasticity in Auditory Cortex; Neural Plasticity in Visual Cortex; Neural Systems: Models of Behavioral Functions; Prenatal and Infant Development: Overview

Bibliography

Buettner H M 1995 Computer simulation of nerve growth cone filopodial dynamics for visualization and analysis. *Cell Motility and the Cytoskeleton* **32**: 187–204

Buonomano D V, Merzenich M M 1998 Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience* **21**: 149–86

Flanagan J G, Vanderhaeghen P 1998 The ephrins and eph receptors in neural development. *Annual Review of Neuroscience* **21**: 309–45

Gage F H 2000 Mammalian neural stem cells. *Science* **287**: 1433–8

Gilbert S F 2000 *Developmental Biology*. Sinauer, Sutherland, MA

Goodhill G J 2000 Topography and ocular dominance: A model exploring positive correlations. *Biological Cybernetics* **69**: 109–18

Goodhill G J 1998 Mathematical guidance for axons. *Trends in Neurosciences* **21**: 226–31

Goodhill G J, Richards L J 1999 Retinotectal maps: Molecules, models, and misplaced data. *Trends in Neurosciences* **22**: 529–34

Goodhill G J, Urbach J S 1999 Theoretical analysis of gradient detection by growth cones. *Journal of Neurobiology* **41**: 230–41

Hebb D O 1949 *The Organization of Behaviour*. Wiley, New York

Hentschel H G E, Van Ooyen A 1999 Models of axon guidance and bundling during development. *Proceedings of the Royal Society of London B* **266**: 2231–8

Hertz J, Krogh A, Palmer R G 1991 *Introduction to the Theory of Neural Computation. Lecture Notes in the Santa Fe Institute Studies in the Sciences of Complexity*. Addison-Wesley, Redwood City, CA

Hubel D H 1988 *Eye, Brain and Vision*. Freeman, New York

Katz L C, Shatz C J 1996 Synaptic activity and the construction of cortical circuits. *Science* **274**: 1133–38

Levay, Connolly M, Houde J, Van Essen D C 1985 *Journal of Neuroscience* **5**: 486–501

Linsker R 1986 From basic network principles to neural architecture. *Proceedings of the National Academy of Sciences USA* **83**: 7508–12, 8390–4, 8779–83
Malsburg C von der 1973 Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* **14**: 85–100
Meinhardt H 1982 *Models of Biological Pattern Formation*. Academic Press, London
Meinhardt H 1999 Orientation of chemotactic cells and growth cones: models and mechanisms. *Journal of Cell Science* **112**: 2867–74
Murray J D 1993 *Mathematical Biology*, 2nd edn. Springer, Berlin
Prestige M C, Willshaw D J 1975 On a role for competition in the formation of patterned neural connexions. *Proceedings of the Royal Society of London B* **190**: 77–98
Reinitz J, Sharp D H 1995 Mechanism of *eve* stripe formation. *Mechanisms of Development* **49**: 133–158
Rosentreter S M, Davenport R W, Löschinger J, Huf J, Jung J, Bonhoeffer F 1998 *Journal of Neurobiology* **37**: 541–62
Rubenstein J L, Beachy P A 1998 Patterning of the embryonic forebrain. *Current Opinions in Neurobiology* **8**: 18–26
Sperry R W 1963 Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proceedings of the National Academy of Sciences USA* **50**: 703–10
Swindale N V 1996 The development of topography in the visual cortex: A review of models. *Network* **7**: 161–247
Tessier–Lavigne M, Goodman C S 1996 The molecular biology of axon guidance. *Science* **274**: 1123–33
Thompson P M, Giedd J N, Woods R P, MacDonald D, Evans A C, Toga A W 2000 Growth patterns in the developing brain detected by using continuum mechanical tensor maps. *Nature* **404**: 190–3
Turing A M 1952 The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B* **237**: 37–72
Wolpert L 1969 Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology* **25**: 1–47

G. J. Goodhill

Neural Networks and Related Statistical Latent Variable Models

1. Introduction

This article presents statistical features associated with artificial neural networks, as well as two other related models, namely mixtures-of-experts and hidden Markov models.

Historically, the study of artificial neural networks was motivated by attempts to understand information processing in biological nervous systems (McCulloch and Pitts 1943). Recognizing that the mammalian brain consists of very complex webs of interconnected neurons, artificial neural networks are built from densely interconnected simple processing units, where each unit receives a number of inputs and produces a single numerical output. However, while artificial neural networks have been motivated by biological systems, there are many aspects of biological systems not adequately modeled by artificial neural networks. Indeed, many aspects of artificial neural networks are inconsistent with biological systems. Though the brain metaphor provides a useful source of inspiration, the viewpoint adopted here is that artificial neural networks are a general class of parameterized statistical models consisting of interconnected processing units.

Artificial neural networks provide a general and practical approach for approximating real-valued (regression problems) and discrete-valued (classification problems) mappings between covariate and response variables. These models, combined with simple methods of estimation (i.e., learning algorithms), have proven highly successful in such diverse areas as handwriting recognition (LeCun et al. 1989), spoken word recognition (Lang et al. 1990), face recognition (Cottrell 1990), text-to-speech translation (Sejnowski and Rosenberg 1986), and autonomous vehicle navigation (Pomerleau 1993).

An artificial neural network consists of multiple sets of units: one set of units corresponds to the covariate variables, a second set of units corresponds to the response variables, and a third set of units, referred to as hidden units, corresponds to latent variables. The hidden units (or, more precisely, their associated hidden or latent variables) mediate the nonlinear mapping from covariate to response variables. Based on observed pairings of covariate and response variables, the back-propagation algorithm is commonly used to estimate the parameters, called weights, of the hidden units.

Some mixture models have statistical structures that resemble those of artificial neural networks. Mixture models are multiple-component models in which each observable data item is generated by one, and only one, component of the model. Like artificial neural networks, mixture models also contain latent or hidden variables. The hidden variables of a mixture model indicate the component that generated each data item.

Mixtures-of-experts (ME) models combine simple conditional probability distributions, such as unimodal distributions, in order to form a complex (e.g., multimodal) conditional distribution of the response variables given the covariate variables. They combine properties of generalized linear models (McCullagh and Nelder 1989) with those of mixture models. Like generalized linear models, they are used to model the relationship between covariate and response variables. Typical applications include nonlinear regression and binary or multiway classification. However, unlike standard generalized linear models they assume that the conditional distribution of the responses is a finite mixture distribution. Mixtures-of-experts are 'piecewise estimators' in the sense that different mixture components summarize the relationship between covariate and response variables for different subsets of the observable data items. The subsets do not, however, have hard boundaries; a data item might simultaneously be a member of multiple subsets. Because

ME assume that the probability of the response variables given the covariate variables is a finite mixture distribution, they provide a motivated alternative to nonparametric models such as artificial neural networks, and provide a richer class of distributions than standard generalized linear models.

Hidden Markov models (HMMs), in contrast with ME models, do not map covariate to response variables; instead, they are used to summarize time-series data. To motivate HMMs, consider the following scenario (Rabiner 1989). A person has three coins. At each time step, she performs two actions. First, she randomly selects a coin by sampling from a multinomial distribution. The parameters of this distribution are a function of the coin that was selected at the previous time step. Next, the selected coin is flipped in order to produce an outcome. Each coin has a different probability of producing either a head or a tail. A second person only observes the sequence of outcomes. Based on this sequence, the person constructs a statistical model of the underlying process that generated the sequence. A useful statistical model for this scenario is a hidden Markov model.

HMMs are mixture models whose components are typically members of the exponential family of distributions. It is assumed that each data item was generated by one, and only one, of the components. The dependencies among the data items are captured by the fact that HMMs include temporal dynamics which govern the transition from one component to another at successive time steps. More specifically, HMMs differ from conventional mixture models in that the selection of the mixture component at time step t is based upon a multinomial distribution whose parameters depend on the component that was selected at time step $t-1$. HMMs are useful for modeling time series data because they explicitly model the transitions between mixture components, or states as they are known in the engineering literature. An advantage of HMMs is that they can model data that violate the stationarity assumptions characteristic of many other types of time series models. The states of the HMM would be Bernoulli distributions in the above scenario. More commonly HMMs are used to model sequences of continuous signals, such as speech data. The states of an HMM would be Gaussian distributions in this case.

Artificial neural networks and mixture models, such as mixtures-of-experts and hidden Markov models, can be seen as instances of two ends of a continuum. Artificial neural networks are an example of a 'multiple-cause' model. In a multiple-cause model, each data item is a function of multiple hidden variables. For instance, a response variable in a neural network is a function of all hidden variables. An advantage of multiple-cause models is that they are representationally efficient. If each hidden variable has a Bernoulli distribution, for example, and if one seeks N bits of information about the underlying state of a data item, then a model with N latent variables is sufficient. In the case when there is a linear relationship between the latent and observed variables, and when the variables are Gaussian distributed, such as is the case with factor models, then the equations for statistical inference (i.e., the determination of a probability distribution for the hidden variables given the values of the observed variables) are computationally efficient. However, a disadvantage of multiple-cause models is that inference tends to be computationally intractable when there is a nonlinear relationship between latent and observed variables, as is frequently the case with neural networks.

Consequently, researchers also consider 'single-cause' models, such as mixture models. Mixture models are regarded as instances of single-cause models due to their assumption that each observable data item was generated by one, and only one, component of the model. An advantage of this assumption is that it leads to a set of equations for performing statistical inference that is computationally efficient. A disadvantage of this assumption is that it makes mixture models inefficient from a representational viewpoint. Consider, for example, a two-component mixture model where each component is a Gaussian distribution. This model provides one bit of information regarding the underlying state of a data item (either the data item was sampled from the first Gaussian distribution or it was sampled from the second Gaussian distribution). Suppose that one seeks N bits of information about the underlying state of a data item. A mixture model requires 2^N components in order to provide this amount of information. That is, as the amount of information grows linearly, the number of required components grows exponentially.

As discussed at the end of this article, a recent trend in the research literature is to consider novel statistical models that are not purely single-cause and are not purely multiple-cause, but rather combine aspects of both.

2. Neural Networks

A simple (and historically important) type of neural network is the single-layer perceptron presented in Fig. 1. This single-layer perceptron receives a vector of inputs, computes a linear combination of these inputs, then outputs a $+1$ (i.e., assigns the case represented by the input vector to group 2) if the result exceeds some threshold and -1 (i.e., assigns the case to group 1) otherwise (the output of a unit is often also called the unit's activation). Specifically, given the covariate values (i.e., inputs) $x_1,\ldots, x_n$, the output of the single-layer perceptron is given by

$$y(x_1,\ldots, x_n) = \begin{cases} +1 \text{ when } w_0 + w_1 x_1 + \cdots + w_n x_n > 0 \\ -1 \text{ otherwise} \end{cases} \tag{1}$$

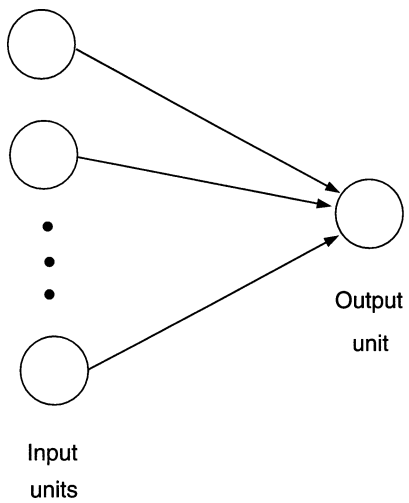

Figure 1
A single-layer perceptron

where each w_i is a real-valued coefficient or weight that determines the contribution of the covariate x_i to the output of the single-layer perceptron. The quantity w_o, termed the bias, is the negative of the threshold that the weighted combination of inputs $w_1x_1 + \cdots + w_nx_n$ must exceed for the perceptron's output to equal unity. Rosenblatt (1962) presented a simple algorithm to estimate the weights of a single-layer perceptron. He also proved that if the data used to estimate the weights (i.e., train the perceptron) are drawn from two groups that lie on opposite sides of a hyperplane (i.e., the data are linearly separable), then his estimation algorithm converges and positions the linear decision surface defined by the perceptron between the two classes. Haykin (1999) presents a proof of this Perceptron Convergence Theorem, as well as other estimation procedures for single-layer networks, e.g., least-squares.

Alternatively, the single-layer perceptron may be written as

$$y(x_1,\ldots, x_n) = g\left(w_0 + \sum_{j=1}^{n} w_j x_j \right) \tag{2}$$

where $g(a)$ (i.e., known as the activation function) equals $+1$ when $a > 0$ and -1 otherwise. Note that were we to take $g(a) = a$, we would have the classical linear regression function. Taking

$$g(a) = \frac{1}{1+\exp(-a)} \tag{3}$$

(i.e., the logistic activation function), results in logistic regression (Ripley 1996). Assuming the data in each class follow a multivariate normal distribution with common variance-covariance matrix, by an application of Bayes' Theorem, the posterior log-odds of class two versus class one is given by this logistic regression model.

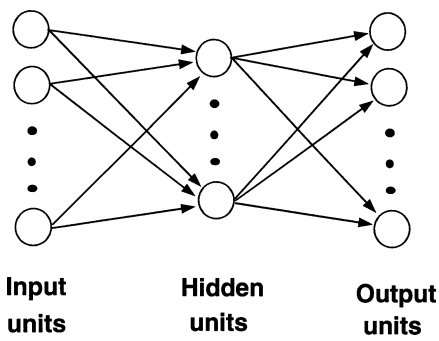

Figure 2
A multilayer network

By the late 1960s mounting evidence suggested that the single-layer perceptron had limited representational abilities. The work of Minsky and Papert (1969) demonstrated that the single-layer perceptron cannot solve problems which are not linearly separable, such as the Boolean exclusive-OR problem (Bishop 1995). As a result, work on neural networks diminished in the 1970s.

The multilayer perceptron, also known as the multi-layer feedforward network, combined with the back-propagation learning algorithm (Rumelhart et al. 1986) represented a landmark in the neural network literature. By moving to a multilayer network, one can model very general mappings to a good approximation. The back-propagation learning algorithm provides a fast mechanism for training the multilayer network relative to other neural network algorithms.

Typically, the multi-layer network consists of a set of sensory units called an input layer, one or more hidden layers of computational nodes, and an output layer of computational nodes. Figure 2 presents a multilayer perceptron with one hidden layer. The input signal enters the net through the input layer and propagates forward through the highly connective network and emerges at the output end as a response signal. At each unit of the network, the unit's activation is computed as a function of the inputs and associated weights specific to that unit. In this presentation, connections are made from every unit in one layer to every unit in the next layer, with no other connections permitted. The output of the network in Fig. 2 can then be mathematically expressed as

$$y_k(x_1,\ldots, x_n) = h\left(w_{k0} + \sum_{j=1}^{n} w_{kj} g\left(w_{j0} + \sum_{i=1}^{n} w_{ji} x_i \right) \right) \tag{4}$$

In practice, the function g is often taken to be the logistic function (Eqn. 3). The function h can be taken as the identity function in the case of regression. Every bounded continuous function can be approximated with arbitrary accuracy (for a finite norm) by a network with two layers of computational units, where logistic activation functions are used at the hidden layer, and linear activation units are used at the output layer, provided the number of hidden units is sufficiently large (Hornik 1991, see also Ripley 1996).

The back-propagation algorithm of Rumelhart et al. (1986) is a nonlinear least-squares algorithm. Given a training set $(\mathbf{x}_j, t_j), j = 1,\ldots, n$, the weight vector $\mathbf{w}$ is chosen to minimize the error function

$$E(\mathbf{w}) = \sum_j \| t_j - y(\mathbf{x}_j; \mathbf{w}) \|^2 \tag{5}$$

where $y(\mathbf{x}_j; \mathbf{w})$ is the output of the network. In short, the key idea underlying the back-propagation algorithm is that the chain rule from calculus can be used to compute the partial derivatives of the error function with respect to the weights of the output units and also with respect to the weights of the hidden units (assuming that the hidden units use a differentiable activation function such as the logistic function). Moreover, the back-propagation algorithm computes these derivatives in a computationally efficient manner due to its recursive nature. Given these partial derivatives, the method of steepest descent (Ripley 1996) yields the update rule:

$$w_{ij} \leftarrow w_{ij} - \eta \frac{\partial E}{\partial w_{ij}} \tag{6}$$

where η is a learning rate parameter. See Cheng and Titterington (1994) or White (1989) for further discussion.

The development of the back-propagation algorithm revived the widespread study of artificial neural networks because it provides a practical procedure for estimating the weight values of hidden units. Whereas single-layer networks are limited to computing linear (or linearly separable) functions, multi-layer networks containing hidden units can compute highly nonlinear unctions. Hidden units should be regarded as instances of latent variables that underlie the mapping from covariate to response variables. Indeed, it is sometimes the case that investigators interpret the weight values of hidden units that result from the application of the back-propagation algorithm as indicating something fundamental about the properties of this mapping.

At the start of training, the weights of a multilayer network are set to random values. At each iteration of the training process, the back-propagation algorithm is applied in order to modify the weight values. The training process may contain many iterations. The issue of when to stop the training process is critical—it is well recognized that one does not want to overfit the training sample leading to a network with poor predictive ability on new data (i.e., it does not generalize well). Haykin (1999) discusses the use of a validation curve to stop the training process.

It is sometimes the case that, in order to prevent a network from overfitting the training sample, the back-propagation algorithm is used with an error function that includes a term (often referred to as a regularization term) that increases with the magnitude of the network's weights. In other words, the goal of the training process is to find a set of weights that are small in magnitude and that allow the network to approximate the mapping from covariate to response variables. The use of a regularization term is often interpreted with respect to the statistical metrics of bias and variance (Geman et al. 1992). Compared to networks trained to minimize an error function that does not contain a regularization term, networks trained with such a term tend to have larger bias and smaller variance. The regularization term alters the training process so that some weight values are more likely to result than other weight values. Consequently, the outcome of the training process is constrained so that this outcome is not due solely to the particular training sample.

3. *Mixtures-of-experts*

We now turn to the first of two types of mixture models reviewed in this article. To motivate the mixtures-of-experts model, assume that the process generating the data is decomposable into a set of sub-processes defined on (possibly overlapping) regions of the covariate space. For each data item, a sub-process is selected, based on the covariate $\mathbf{x}^{(t)}$, and the selected sub-process maps $\mathbf{x}^{(t)}$ to the response $\mathbf{y}^{(t)}$. More precisely, data are generated as follows. For each covariate $\mathbf{x}^{(t)}$,

a label i is chosen from a multinomial distribution with probability $P(i|\mathbf{x}^{(t)}, V)$, where $V = [\mathbf{v}_1,\ldots, \mathbf{v}_I]$ is the matrix of parameters underlying the multinomial distribution;

a response $\mathbf{y}^{(t)}$ is generated with probability $P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_i, \Phi_i)$, where U_i is a parameter and Φ_i represents other (possibly nuisance) parameters, for $i = 1,\ldots, I$.

To more explicity relate this approach to the generalized linear models framework, it is assumed that the conditional probability distribution $P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_i, \Phi_i)$ is a member of the exponential family of distributions such as Gaussian, Poisson, binomial, or multinomial distributions (McCullagh and Nelder 1989). The expected conditional value of the response $\mathbf{y}^{(t)}$, denoted $\boldsymbol{\mu}_i^{(t)}$), is defined to be a user-specified function f of $\mathbf{x}^{(t)}$ and parameter matrix U_i. The quantities $\eta_i = f^{-1}(\boldsymbol{\mu}_i)$ and Φ_i are, respectively, the

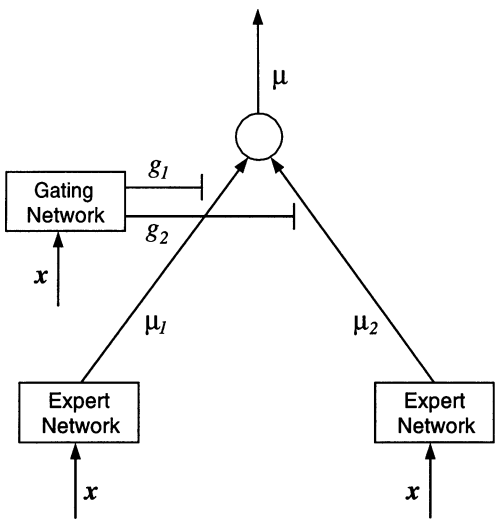

Figure 3
A mixture-of-experts model

natural parameter and dispersion parameter (McCullagh and Nelder 1989) of the response's conditional distribution $P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_i, \Phi_i)$. The total probability of generating $\mathbf{y}^{(t)}$ from $\mathbf{x}^{(t)}$ is then given by the mixture density

$$P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)},\Theta) = \sum_i P(i|\mathbf{x}^{(t)}, V)P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)},U_i, \Phi_i) \tag{7}$$

where $\Theta = [\mathbf{v}_1,\ldots, \mathbf{v}_I, U_1,\ldots, U_I, \Phi_1,\ldots, \Phi_I]^T$ is the matrix of all parameters. Assuming independently distributed data, the total probability of the dataset χ is the product of T such densities, with the likelihood given by:

$$L(\Theta|\chi) = \prod_t \sum_i P(i|\mathbf{x}^{(t)},V)P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)},U_i,\Phi_i) \tag{8}$$

Figure 3 presents a graphical representation of the ME model. The model consists of several modules referred to as 'expert networks.' Because each expert network is a generalized linear model, it calculates a weighted sum of the covariate variables and then forms its output by using this sum as the parameter of a monotonic and differentiable function, such as the linear or logistic activation functions discussed earlier. Expert networks approximate the data within each region of the covariate space: expert network i maps its input, the covariate vector $\mathbf{x}$, to an output vector $\boldsymbol{\mu}_i$. It is assumed that different experts are appropriate in different regions of the covariate space. Consequently, the model requires a module, referred to as a 'gating network,' that identifies for any covariate $\mathbf{x}$, the expert or blend of experts whose output is most likely to approximate the corresponding response vector $\mathbf{y}$. The gating network is also a generalized linear model. Its outputs are a set of scalar coefficients g_i that weight the contributions of the various experts. For each covariate $\mathbf{x}$, these coefficients are constrained to be nonnegative and to sum to one. The total output of the model, given by

$$\boldsymbol{\mu} = \sum_{i=1}^{n} g_i \boldsymbol{\mu}_i \tag{9}$$

is a convex combination of the expert outputs for each $\mathbf{x}$.

From the perspective of statistical mixture modeling, the gating network is identified with the selection of a particular subprocess. That is, the gating outputs g_i are interpreted as covariate-dependent, multinomial probabilities of selecting subprocess i. Different expert networks are identified with different sub-processes; each expert models the covariate-dependent distributions associated with its corresponding subprocess.

Note that each expert multiplies the covariate vector $\mathbf{x}$ by a matrix of parameters. (The vector $\mathbf{x}$ is assumed to include a fixed component of one to allow for an intercept term.) For expert i, the matrix is denoted as U_i and the resulting vector is denoted as $\boldsymbol{\eta}_i$:

$$\boldsymbol{\eta}_i = U_i \mathbf{x} \tag{10}$$

The quantity $\boldsymbol{\eta}_i$ is mapped to the expert output $\boldsymbol{\mu}_i$ by the user-specified function f.

For example, in regression problems, f may be taken as the identity function (i.e., the experts are linear). Moreover, the probabilistic component of the model may be Gaussian. In this case, the likelihood is a mixture of Gaussians:

$$L(\Theta|\chi) = \prod_t \sum_i g_i^{(t)} |\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}[\mathbf{y}^{(t)}-\boldsymbol{\mu}_i^{(t)}]^T \Sigma_i^{-1}[\mathbf{y}^{(t)}-\boldsymbol{\mu}_i^{(t)}]} \tag{11}$$

The expert output $\boldsymbol{\mu}_i^{(t)}$ and expert dispersion parameter Σ_i are interpreted as the mean and covariance matrix of expert i's Gaussian distribution for response $\mathbf{y}^{(t)}$ given covariate $\mathbf{x}^{(t)}$. The output of the entire model $\boldsymbol{\mu}^{(t)}$ is, therefore, interpreted as the expected value of $\mathbf{y}^{(t)}$ given $\mathbf{x}^{(t)}$.

For binary classification problems, f may be the logistic function. Other problems (e.g., multiway classification, counting, rate estimation, and survival estimation) may suggest other choices for f.

The gating network first computes the intermediate variables ξ_i as the inner product of the covariate vector $\mathbf{x}$ and the vector of parameters $\mathbf{v}_i$:

$$\xi_i = \mathbf{v}_i^T \mathbf{x} \tag{12}$$

The ξ_i are then mapped to the gating outputs g_i by using a generalization of the logistic function:

$$g_i = \frac{e^{\xi_i}}{\sum_j e^{\xi_j}} \tag{13}$$

In the neural network literature (e.g., Bridle 1989) this function is known as the 'softmax function.' Note that the inverse of this function is the canonical link for a multinomial response model (McCullagh and Nelder 1989).

4. Hidden Markov Models

The second type of mixture model reviewed in this article is hidden Markov models. As discussed in Sect. 1, HMMs are mixture models whose components are typically members of the exponential family of distributions. They are useful for summarizing time-series data because they include temporal dynamics which govern the transition from one mixture component to another at successive time steps.

An HMM with Gaussian states is characterized by four quantities (Rabiner 1989). The number of states in the model is the first quantity. The second quantity is the state transition probability matrix $\mathbf{A}$, where matrix element a_{ij} is the probability of moving from state i to state j. The third quantity is the set of emission parameters $B = \{\boldsymbol{\mu}_i, \Sigma_i\}$ where $\boldsymbol{\mu}_i$ is the mean of the Gaussian distribution associated with state i, and Σ_i is its covariance matrix. The set of initial state probabilities $\pi = \{\pi_i\}$ is the fourth quantity, where π_i is the probability that the sequence of states starts in state i.

Given an observation sequence $\chi = \mathbf{x}^{(1)},\ldots, \mathbf{x}^{(T)}$, and a model parameterized by Θ, the matrix of all parameters, we would like to evaluate the likelihood $L(\Theta|\chi) = p(\chi|\Theta)$. The most straightforward way is to enumerate every possible state sequence of length T (Rabiner 1989). Let $Q = q^{(1)},\ldots, q^{(T)}$ denote one such fixed-state sequence where $q^{(t)}$ indicates the state at time t ($q^{(t)}$ is a vector whose i^{th} element is set to 1 when the HMM is in state i and whose other elements are set to 0). Assuming statistical independence of observations, the probability of observation sequence χ given state sequence Q and model parameters Θ is

$$p(\chi|Q,\Theta) = \prod_{t=1}^{T} p(\mathbf{x}^{(t)}|q^{(t)}\Theta) \tag{14}$$

$$= \prod_{t=1}^{T} b_{q^{(t)}}(\mathbf{x}^{(t)}) \tag{15}$$

where $b_{q^{(t)}}(\mathbf{x}^{(t)})$ is the probability of producing observation $\mathbf{x}^{(t)}$ when the HMM is in the state indicated by $q^{(t)}$ at time t. The probability of state sequence Q is

$$P(Q|\Theta) = \pi_{q^{(1)}} \prod_{t=2}^{T} a_{q^{(t-1)}q^{(t)}} \tag{16}$$

The probability of the observation sequence χ is a mixture distribution given by the product of $p(\chi|Q, \Theta)$ and $P(Q|\Theta)$ summed over all possible state sequences Q:

$$p(\chi|\Theta) = \sum_{\text{all } Q} p(\chi|Q,\Theta)P(Q|\Theta) \tag{17}$$

Unfortunately, computing the probability of the observation sequence in this manner is often computationally expensive. A more efficient algorithm, known as the forward–backward algorithm, is reviewed in Rabiner (1989).

One way of finding maximum likelihood estimates of the values of the parameters A, B, and π is to perform gradient ascent on the log likelihood function. A second way is to use the EM algorithm, known in the HMM literature as the Baum–Welch algorithm (Rabiner 1989). This algorithm may be motivated as follows. Suppose that at each time step, the state of the HMM is known (in fact it is not known, but suppose for the moment that it is known). In this case, parameter estimation is easy. Estimates of the elements of the state transition matrix $\boldsymbol{A}$ are obtained by counting the number of times that the HMM moves from one state to the next state. If n_{ij} is the number of times that the HMM moves from state i to state j, then the estimate of the probability of moving from state i to state j is the fraction $n_{ij}/\Sigma_k n_{ik}$. Estimates of the emission parameters B are obtained by recording the statistics of the data items that were observed in each state. For example, if each state is a Gaussian distribution, then the estimates of the parameters associated with state i, namely the mean $\boldsymbol{\mu}_i$ and covariance matrix Σ_i, are the sample mean and sample covariance matrix based on those data items observed when the HMM was in state i. Multiple observation sequences are needed in order to estimate the initial state probabilities π_i. In practice, multiple sequences are often available.

However, this parameter estimation strategy is generally not available, because we do not know the state of the HMM at each time step. Nonetheless, it is possible to approximate this strategy by using Bayes' rule to compute the posterior probabilities that the HMM is in each state at each time step given the observation sequence and the current estimates of the parameter values (see Rabiner 1989 for details).

5. Extensions and Related Work

To conclude this article, we briefly review some recent trends in the research literature. Both artificial neural networks and mixture models, such as mixtures-of-experts and hidden Markov models, can be regarded as special cases of a broader class of statistical models known as graphical models. For example, an artificial

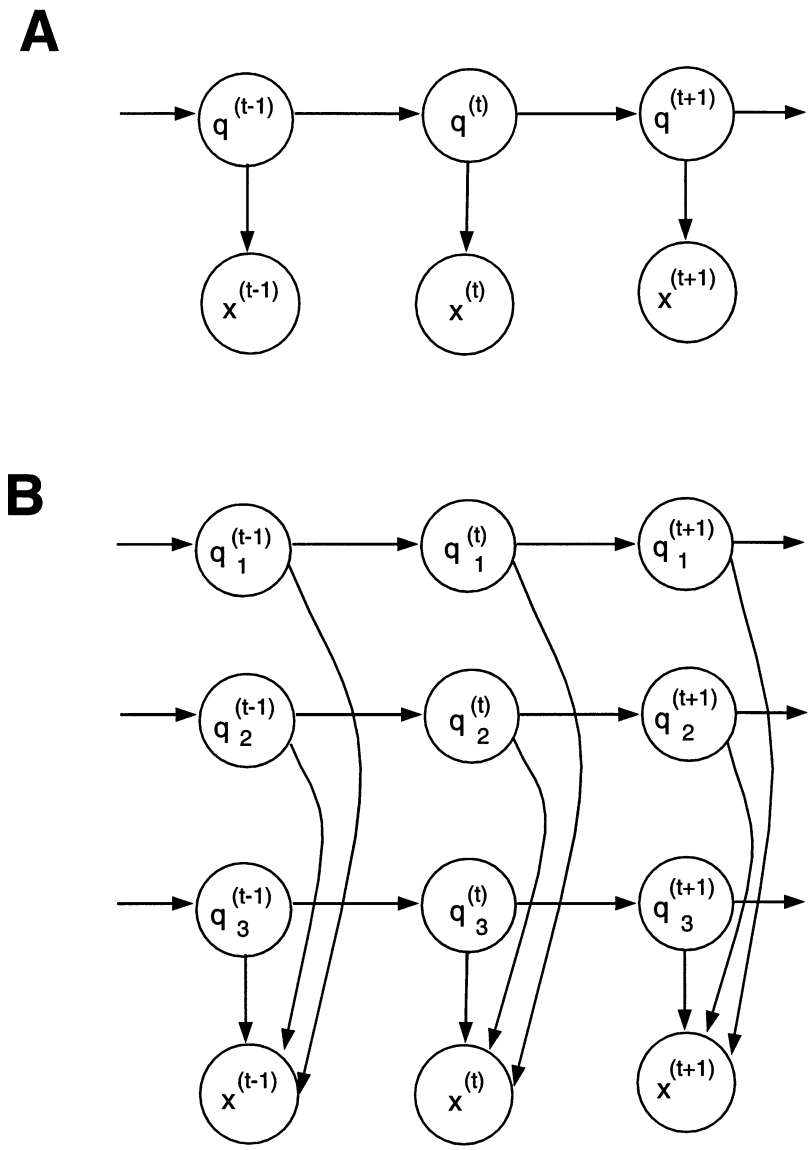

Figure 4
(A) Illustration of an HMM. The data item at time t, $x^{(t)}$, depends on the state at time t, $q^{(t)}$. (B) Illustration of a factorial HMM. The data item at each time step still depends on the state against that time step, but now the state is factored into multiple-state variables.

neural network may be considered to be an instance of a directed acyclic graph in which the units of the network are the nodes of the graph, and the connections between units are the directed edges of the graph. If one adopts this perspective, then one can apply methods developed for the analysis of graph structures to the problem of statistical inference in artificial neural networks. This perspective is developed in Jordan (1998).

Jordan and Jacobs (1992) proposed hierarchical mixtures-of-experts (HME) as a multilevel extension of mixtures-of-experts. An EM algorithm for estimating the parameter values of an HME was presented by Jordan and Jacobs (1994). They found that an HME trained with the EM algorithm learned approximately two orders of magnitude faster than a multilayer perceptron trained with the back-propagation algorithm on a relatively difficult robot dynamics task.

Bayesian inference for both ME and HME based on Markov chain Monte Carlo methodology (see Tanner 1996) is discussed in Peng et al. (1996) and Jacobs et al. (1997). An HME architecture trained via an MCMC sampler produces both an estimate of a correct response and also a measure of how confident it is in its estimate. This confidence measure is useful for obtaining robust performance. Peng et al. found that an HME architecture trained with the MCMC sampler showed excellent performance on a spoken vowel classification task. Recently, researchers have studied the analytical properties of ME and HME. Jiang and Tanner (1999a) consider the approximation rate of HME and Jiang and Tanner (1999b) consider statistical consistency issues for HME trained via maximum likelihood estimation procedures. Recent work has considered extensions of ME and HME to be mixtures of models other than generalized linear models such as mixtures of factor models (Ghahramani and Hinton 1996) and mixtures of principal component analyzers (Tipping and Bishop 1997).

As mentioned earlier, a recent trend in the research literature is to consider novel statistical models that are not purely single-cause and are not purely multiple-cause, but rather combine aspects of both. One example is the development of novel kernel-based methods for regression and classification, such as radial basis function networks (Poggio and Girosi 1990) or support vector machines (Vapnik 1995). A second example is the development of novel hidden Markov models. Conventional HMMs assume that the data item at each time step is sampled from a single mixture component. This is illustrated in Fig. 4, Panel A where the data item at time t, $x^{(t)}$, depends on the state at time t, $q^{(t)}$. As discussed earlier, although statistical inference in single-cause models is computationally efficient, single-cause models are inefficient from a representational viewpoint. Ghahramani and Jordan (1997) recently developed a multiple-cause version of an HMM referred to as a factorial hidden Markov model (fHMM). In this model, the data item at each time step still depends on the state at that time step, but now the state is factored into multiple-state variables (see Fig. 4, Panel B). Due to the combinatorial nature of the state representation, exact statistical inference in this model is computationally inefficient. Nonetheless, approximate inference is computationally feasible (see Ghahramani and Jordan 1997 for details). When the factorial hidden Markov model was used to summarize a time series based on Bach's chorales, it was found that the model captured statistical structure in the dataset which a standard HMM did not.

See also: Bayesian Graphical Models and Networks; Graphical Methods: Analysis; Graphical Methods: Presentation

Bibliography

Bishop C M 1995 *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK

Bridle J 1989 Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In: Fogelman-Soulie F, Hérault J (eds.) *Neurocomputing: Algorithms, Architectures, and Applications*. Springer-Verlag, New York

Cheng B, Titterington D M 1994 Neural networks: A review from the statistical perspective. *Statistical Science* **9**: 2–54

Cottrell G W 1990 Extracting features from faces using compression networks: Face, identity, emotion, and gender recognition using holons. In: Touretzky D (ed.) *Connectionist Models: Proceedings of the 1990 Summer School*. Morgan Kaufman, San Mateo, CA

Geman S, Bienenstock E, Doursat R 1992 Neural networks and the bias/variance dilemma. *Neural Computation* **4**: 1–58

Ghahramani Z, Hinton G E 1996 The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1. Department of Computer Science, University of Toronto

Ghahramani Z, Jordan M I 1997 Factorial hidden Markov models. *Machine Learning* **29**: 245–75

Haykin S 1999 *Neural Networks*, 2nd edn. Prentice Hall, Upper Saddle River, NJ

Hornik K 1991 Approximation capabilities of multilayer feedforward networks. *Neural Networks* **2**: 359–66

Jacobs R A, Peng F, Tanner M A 1997 A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks* **10**: 231–41

Jiang W, Tanner M A 1999a On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural Computation* **11**: 1183–98

Jiang W, Tanner M A 1999b Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *The Annals of Statistics* (in press)

Jordan M I 1998 *Learning in Graphical Models*. MIT Press, Cambridge, MA

Jordan M I, Jacobs R A 1992 Hierarchies of adaptive experts. In: Moody J E, Hanson S J, Lippmann R P (eds.) *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann Publishers, San Mateo, CA

Jordan M I, Jacobs R A 1994 Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6**: 181–214

Lang K J, Waibel A H, Hinton G E 1990 A time-delay neural network architecture for isolated word recognition. *Neural Networks* **3**: 33–43

LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W, Jackel L D 1989 Backpropagation applied to handwritten zip code recognition. *Neural Computation* **1**: 541–51

McCullagh P, Nelder J A 1989 *Generalized Linear Models*. Chapman and Hall, London

McCulloch W S, Pitts W 1943 A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–33

Minsky M L, Papert S A 1969 *Perceptrons*. MIT Press, Cambridge, MA

Peng F, Jacobs R A, Tanner M A 1996 Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association* **91**: 953–60

Poggio T, Girosi F 1990 Networks for approximation and learning. *Proceedings of the IEEE* **78**: 1481–97

Pomerleau D A 1993 Knowledge-based training of artificial neural networks for autonomous robot driving. In: Connel J, Mahadevan S (eds.) *Robot Learning*. Kluwer Academic, Boston, MA

Rabiner L R 1989 A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**: 257–86

Ripley B D 1996 *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK

Rumelhart D E, Hinton G E, Williams R J 1986 Learning internal representations by error propagation. In: Rumelhart D E, McClelland J L, and the PDP Research Group (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Volume 1: Foundations)*. MIT Press, Cambridge, MA

Rosenblatt F 1962 *Principles of Neurodynamics*. Spartan Books, Washington, DC

Sejnowski T J, Rosenberg C R 1986 NETtalk: A parallel network that learns to read aloud. Technical Report JHU/EECS-86/01, Johns Hopkins University. Reprinted in Anderson J A, Rosenfeld E (eds.) *Neurocomputing: Foundations of Research*. MIT Press, Cambridge, MA

Tanner M A 1996 *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd edn. Springer-Verlag, New York

Tipping M E, Bishop C M 1997 Mixtures of probabilistic principal component analysers. Technical Report TR-NCRG/97/003. Neural Computing Research Group, Aston University, Birmingham, UK

Vapnik V 1995 *The Nature of Statistical Learning Theory*. Springer-Verlag, New York

White H 1989 Learning in artificial neural networks: A statistical perspective. *Neural Computation* **1**: 425–64

M. A. Tanner and R. A. Jacobs

Neural Networks: Biological Models and Applications

Beginning with the seminal work of McCulloch and Pitts in the 1940s, artificial neural network (or connectionist) modeling has involved the pursuit of increasingly accurate characterizations of the electrophysiological properties of individual neurons and networks of interconnected neurons. This line of research has branched into descriptions of the nervous system at many different grains of analysis, including complex computer models of the properties of individual neurons, models of simple invertebrate nervous systems involving a small number of neurons, and more abstract treatments of networks involving thousands or millions of neurons in the nervous systems of humans and other vertebrates. At various points in the history of neural network research, successful models have moved beyond the domain of biological modeling into a variety of engineering and medical applications.

1. Modeling the Computational Properties of Neurons

Although the idea that the brain is the seat of the mind and controller of behavior is many centuries old, research into the computational properties of interconnected neurons was largely absent until the 1940s. McCulloch and Pitts (1943) initiated the field of neural network research by investigating networks of interconnected neurons, with each neuron treated as a simple binary logic computing element. In this model, the axon of a neuron carries the cell's binary output signal. This axon forms a set of synaptic connections to the dendrites, or inputs, of other neurons. The axonal signal corresponds roughly to the voltage level, or *membrane potential*, of the neuron. The total input to a neuron is the sum of its synaptic inputs, and if this sum exceeds a certain threshold, the McCulloch–Pitts neuron produces an output of 1. Otherwise the cell's output is 0. This binary conception of a neuron's output was based on observations of 'all or nothing' spikes, or *action potentials*, in the membrane potentials of biological neurons.

The McCulloch–Pitts model of the neuron was formulated when our physiological understanding of neurons was rather limited. Although simple binary neurons are still used in some neural network models, many later models treat a neuron's output as a continuous, rather than binary, function of the cell's inputs. This change was motivated by neurophysiological observations that the frequency of neuron spiking, rather than the presence or absence of an individual spike, is the more relevant quantity for measuring the strength of signaling between neurons. Perhaps the most widely used formulation of a neuron's output in current neural networks is the following equation:

$$y_j = f\left(\sum_i y_i z_{ij} - \theta_j\right)$$

where y_j is the output (spiking rate) of a neuron labeled j, z_{ij} is the strength of the connection (synapse) from neuron i to neuron j, θ_j is the firing threshold for neuron j, and the *output function* $f(x)$ is typically an increasing function of x, with $f(x) = 0$ for $x < 0$. In words, if a neuron's total input ($\Sigma_i y_i z_{ij}$) is below the firing threshold, then the neuron's output is zero, and if the input exceeds the firing threshold, then the output is a positive value related to the difference between the total input and the threshold. Different models use different forms of $f(x)$, with common choices including threshold linear and sigmoidal output functions.

A significantly more sophisticated account of neuron dynamics was formulated by Hodgkin and Huxley (1952), who won the Nobel prize for their experimental and modeling work elucidating the relationship between the ionic currents flowing in and

out of a neuron and the membrane potential of the neuron. The behavior of networks of Hodgkin–Huxley-like neurons, or *shunting neural networks*, has been studied in some detail (e.g., Grossberg 1980), and these networks have formed the basis of a number of models of biological nervous systems.

The models described above are *point* models of a neuron in that they treat the electrical properties of the neuron as uniform across the neuron's membrane; in other words, they treat neurons as if they did not have a spatial extent. However, the membrane potential of a real neuron varies as a function of position on the membrane. For example, the membrane potential at a distal dendrite of a neuron can differ substantially from the membrane potential at the cell body or along the axon. Sophisticated *compartmental* models of neurons, which treat the neuron as a collection of interconnected electrical subcircuits (compartments), have been developed in recent years. Each subcircuit in a compartmental model corresponds to a different portion of a real neuron, such as a single dendrite, and large-scale computer simulations are used to simulate the membrane potential of a single neuron. Although compartmental models provide a more accurate description of single neuron dynamics than the point models used in most neural networks, the complexity of this type of model has prevented its use in networks containing more than a handful of neurons. (See Arbib 1995 for more information on single neuron models and neuron simulators.)

2. Learning in Neural Networks

The connections between cells in an artificial neural network correspond to synapses in the nervous system. In the earliest neural network models, the strengths of these connections, which determine how much the presynaptic cells can influence the activity of the postsynaptic cells, were kept constant. However, much of the utility of neural networks comes from the fact that they are capable of modifying their computational properties by changing the strengths of synapses between cells, thus allowing the network to adapt to environmental conditions (for biological neural networks) or to the demands of a particular engineering application (for artificial neural networks). A major challenge for computational neuroscientists has been to develop useful algorithms for changing the weights in a neural network in order to improve its performance based on a set of training samples.

Training a neural network typically consists of the presentation of a set of input patterns alone, or the presentation of input/output pattern pairs, to the network. During the presentation of each pattern or input/output pair, the weights of the synapses in the network are modified according to an equation that is often referred to as a *learning law*. In a *supervised learning* network, training consists of repeated presentation to the network of input/output pairs that represent the desired behavior of the network. The difference between the network's output and the training output represents the performance error and is used to determine how the weights will be modified. In a *self-organizing* network, the weights are changed based on a set of input patterns alone, and the network typically learns to represent certain aspects of the statistical distribution of the training inputs. For example, a self-organizing network trained with an input data set that includes three natural clusters of data points might learn to identify the data points as members of three distinct categories.

A variety of learning laws have been developed for both supervised and self-organizing neural networks. Most of these learning laws fall into one of two classes. The origins of the first class, *associative* or *Hebbian learning laws*, can be traced to a simple conjecture penned by the cognitive psychologist Donald Hebb (1949): 'when an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.' With this statement, Hebb gave birth to the concept of Hebbian learning, in which the strength of a synapse is increased if both the pre- and postsynaptic cells are active at the same time. Remarkably, Hebb's conjecture, which was made before the development of experimental techniques for collecting neurophysiological data concerning synaptic changes, has proven to capture one of the most commonly observed aspects of synaptic change in biological nervous systems, and variations of the Hebbian learning law are used in many current neural network models. It should be noted, however, that these learning laws are considerably simplified approximations of the complex and varied properties of real synapses (e.g., see *Synaptic Efficacy, Regulation of*; *Long-term Potentiation (Hippocampus)*; *Long-term Potentiation and Depression (Cortex)*; *Long-term Depression (Cerebellum)*; *Long-term Depression (Hippocampus)*).

In a neural network using Hebbian learning, a synapse's strength depends only on the pre- and postsynaptic cell activations, rather than on a measure of the network's performance error. Hebbian learning laws are thus well-suited for self-organizing neural networks.

A second common class of neural network learning laws, *error minimization learning laws*, are commonly employed in supervised learning situations where an error signal can be computed by comparing the network's output to a desired output. Whereas Hebbian learning laws arose from psychological and neurophysiological analyses, error minimization learning laws arose from mathematical analyses aimed at minimizing the network's performance error, usually through a technique known as *gradient descent*. The network's performance error can be represented

as a surface in the space of the synaptic weights. Valleys on this error surface correspond to synaptic weight choices that lead to low error values. Ideally, one would choose the synaptic weights that correspond to the global minimum of the error surface. However, the entire error surface cannot be 'seen' by the network during a training trial; only the local topography of the error surface is known. Gradient descent learning laws change the synaptic weights so as to move down the steepest part of the local error gradient.

One of the first gradient descent learning laws was developed by Widrow and Hoff (1960) for a simple one-layer neural network, the ADALINE model, which has found considerable success in technological applications such as adaptive noise suppression in computer modems. However, one-layer networks have limited computational capabilities. Gradient descent learning has since been generalized to networks with three or more layers of neurons, as in the commonly employed back-propagation learning algorithm first derived by Paul Werbos in his 1974 Harvard Ph.D. dissertation and later independently rediscovered and popularized by Rumelhart et al. (1986).

3. Common Neural Network Architectures

The cells in a neural network can be connected to each other in a number of different ways. A *feed-forward network* is one in which the output of a cell does not affect the cell's input in any way. Probably the most common artificial neural network architecture is the three-layer feed-forward network, first described by Rosenblatt (1958). In this architecture, an input pattern is represented by cells in the first layer. These cells project through modifiable synapses to the second layer, which is often referred to as the hidden layer since it is not directly connected to the input or output of the network. The hidden layer cells in turn project through modifiable synapses to the output layer. The back-propagation learning algorithm is a common choice for supervised learning in three-layer feed-forward networks.

Recurrent or *feedback networks* can exhibit much more complex behavior than feed-forward networks, including sustained oscillations or chaotic cycles of cell activities over time. In their most general form, the cell outputs of one layer in a recurrent network not only project to cells in the next layer, but they can also project to cells in the same layer or previous layers. Variants of the back-propagation algorithm have been developed for training multilayer recurrent networks.

One important and heavily studied recurrent network architecture is the *self-organizing map*, which was first formulated to account for neurophysiological observations concerning cell properties in primary visual cortex (von der Malsburg 1973). The basic principle of a self-organizing map is as follows. Cells in the input layer, sometimes referred to as a subcortical layer, project in a feed-forward fashion to the cells in the second, or cortical, layer, through pathways that have modifiable synapses. Cells in the cortical layer are recurrently interconnected, and they compete with each other through inhibitory connections so that only a small number of the cortical layer cells are active at the same time. These 'competition winners' are typically the cells that have the most total input projecting to them from the subcortical layer, or the cells that lie near cells with a large amount of input. The connections between the subcortical cells and the active cortical cells are then modified via an associative learning law so that these cells are even more likely to become active the next time the same input pattern is applied at the subcortical layer. The net effect over many training samples is that cells that are near to each other in the cortical layer respond to similar input patterns (a property referred to as *topographic organization*), and more cortical cells respond to input patterns that are frequently applied to the network during training than to rarely encountered input patterns. Although originally formulated as recurrent networks in which the cells in the cortical layer project to each other in a recurrent fashion, simplified feed-forward versions of the self-organizing map architecture that approximate stable behavior of the recurrent system have been developed and thoroughly studied (e.g., Kohonen 1984).

A related neural network architecture that has also been used to explain a number of neurophysiological observations is the *adaptive resonance theory* (ART) architecture (Grossberg 1980). In this network, an additional set of recurrent connections project from the cortical layer back down to the subcortical layer. The top-down projections emanating from a cortical cell embody the subcortical pattern that the network has learned to expect when that cortical cell is activated. These learned expectations can be used to correct coding errors before learning has taken place in the bottom-up pathways, thereby providing a more stable cortical representation. In addition to their use in biological modeling, neural network systems based on the ART model have been applied to a number of pattern recognition problem domains (Carpenter and Grossberg 1991).

4. Specialized Neural Models of Biological Systems

The neural networks described so far are 'general-purpose' models in that the same architecture is used to attack a variety of biological modeling or engineering problems. In addition to these models, many specialized models of particular neural circuits have been proposed. Among the earliest were models of cerebellum function, proposed by researchers such as Marr and Albus beginning in the late 1960s. The

cerebellum has a very regular and well-characterized anatomical structure, and cerebellar physiology has been heavily studied in recent decades (e.g., see *Cerebellum: Cognitive Functions*; *Cerebellum: Associative Learning*; *Long-term Depression (Cerebellum)*). Different cells and synapses in the cerebellum have different properties, and neural models of the cerebellum typically incorporate these differences. Relatively primitive invertebrate neural circuits, such as heartbeat oscillators, have also been the focus of numerous biologically specialized neural network models, as have vertebrate circuits such as the superior colliculus, hippocampus, basal ganglia, and various regions of cortex.

Another type of specialized biological model approximates the function of entire behavioral systems involving large-scale networks of the human brain. Individual cells in these models often correspond to relatively large brain regions, rather than to single neurons or distinct populations of neurons. These models often combine aspects of different neural network architectures or learning laws. The DIVA model of speech production (Guenther 1995), for example, combines several aspects of earlier neural network models into an architecture that learns to control movements of a computer-simulated vocal tract. The model has been shown to provide a unified account for a wide range of experimental observations concerning human speech that were previously studied independently. Other models of this type address various aspects of human cognition, movement control, vision, audition, language, and memory.

5. Pattern Recognition Applications

Neural networks are capable of learning complicated nonlinear relationships from sets of training examples. This property makes them well suited to pattern recognition problems involving the detection of complicated trends in high-dimensional datasets. One such problem domain is the detection of medical abnormalities from physiological measures. Neural networks have been applied to problems such as the detection of cardiac abnormalities from electrocardiograms and breast cancer from mammograms, and some neural network diagnostic systems have proven capable of exceeding the diagnostic abilities of expert physicians. Supervised learning networks have been applied to a number of other pattern recognition problems, including visual object recognition, speech recognition, handwritten character recognition, stock market trend detection, and scent detection (e.g., Carpenter and Grossberg 1991). For further reading on neural networks and their biological bases, see Anderson et al. (1988), Arbib (1995), and Kandel et al. (2000).

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Connectionist Approaches; Connectionist Models of Concept Learning; Connectionist Models of Development; Connectionist Models of Language Processing; Learning and Memory: Computational Models; Neural Circuits: Models of Emergent Functions

Bibliography

Anderson J A, Pellionisz A, Rosenfeld E (eds.) 1988 *Neurocomputing*. MIT Press, Cambridge, MA

Arbib M A (ed.) 1995 *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA

Carpenter G A, Grossberg S (eds.) 1991 *Pattern Recognition by Self-Organizing Neural Networks*. MIT Press, Cambridge, MA

Grossberg S 1980 How does a brain build a cognitive code? *Psychological Review* **87**: 1–51

Guenther F H 1995 Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review* **102**: 594–621

Hebb D O 1949 *The Organization of Behavior: a Neuropsychological Theory*. Wiley, New York

Hodgkin A L, Huxley A F 1952 A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology* (*Lond.*) **117**: 500–44

Kandel E R, Schwartz J H, Jessell T M (eds.) 2000 *Principles of Neural Science*, 4th edn. McGraw-Hill, New York

Kohonen T 1984 *Self-organization and Associative Memory*. Springer-Verlag, New York

Malsburg C von der 1973 Self-organization of orientation sensitive cells in the striata cortex. *Kybernetic* **14**: 85–100

McCulloch W S, Pitts W 1943 A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematics and Biophysics* **5**: 115–33

Rosenblatt F 1958 The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**: 386–408

Rumelhart D E, McClelland J L, PDP Research Group 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA

Widrow B, Hoff M E (eds.) 1960 *Institute of Radio Engineers Western Electric Show and Convention*, Institute of Radio Engineers, New York. Reprinted in Anderson J A, Pellionisz A, Rosenfeld E (eds.) 1988 *Neurocomputing*. MIT Press, Cambridge, MA

F. H. Guenther

Neural Networks, Statistical Physics of

1. Introduction

Neural computing is a field in computer science that aims at constructing computational models that follow an analogy with neural networks in the human brain. These models are often called artificial neural networks and can be thought of as a collection of computational units called neurons that are connected in some way. Each neuron has an internal state that is

computed by applying a non-linear response function to a weighted summation of its inputs and a threshold. By continuously updating the states of the individual neurons the network may organize itself to exhibit a global behavior resembling that of an associative memory or of a self-learning system.

Here, we consider the special class of neural networks in which individual neurons have only two possible output values. Two well-known examples of this class of two-state neural networks are Hopfield networks (Hopfield 1982) and Boltzmann machines (Ackley et al. 1985). Boltzmann machines can be viewed as an extension of Hopfield networks in the following two ways. First, the greedy response function of a Hopfield network is replaced in a Boltzmann machine with a stochastic response function. Second, the relatively simple Hebbian learning rule in a Hopfield network is replaced in a Boltzmann machine with a more powerful stochastic learning algorithm. Both extensions are achieved through the use of stochastic neurons.

Two-state neural networks have a resemblance to many-particle systems as studied in statistical physics. This analogy can be used to analyse the behavior of these neural networks using models and methods derived from statistical physics. Here, we present some basic results on the storage capacity and convergence behavior of two-state neural networks. For a more elaborate treatment of this subject the reader is referred to the book by Hertz et al. (1995).

2. Two-state Neural Networks

A two-state neural network consists of a set $\mathscr{U}$ of two-state neurons that are interconnected by weighted symmetric connections. A neuron u can be in one of two states, either it is firing, corresponding to state '1,' or it is not firing, corresponding to state '-1.' A configuration k is a $|\mathscr{U}|$-dimensional vector that describes the global state of the network. The state of an individual neuron u is given by the component $s_u(k)$. Let $\mathscr{R} = \{-1, 1\}^{|\mathscr{U}|}$ be the set of all possible configurations. A weight $w_{u,v} \in \mathbb{R}$ is associated with the connection between a neuron u and a neuron v. By definition, we assume $w_{u,u} = 0$ for each $u \in \mathscr{U}$ and $w_{u,v} = w_{v,u}$ for each pair $u, v \in \mathscr{U}$. Each neuron u has a threshold $b_u \in \mathbb{R}$.

Self-organization in a two-state neural network is achieved by allowing neurons to change their states, from '-1' to '1' or vice versa. Let the network be in configuration k, then a state change of neuron u results in a configuration l, with $s_u(l) = -s_u(k)$ and $s_v(l) = s_v(k)$ for all neurons $v \neq u$. A possible state change of a neuron u is related to its net input $z_u(k)$, defined as the sum of its weighted inputs minus its threshold, i.e.,

$$z_u(k) = \sum_{v \in \mathscr{U}} w_{u,v} s_v(k) - b_u. \tag{1}$$

Here, we consider instantaneous asynchronous state changes, i.e., state changes that occur one by one and take zero time to execute. The state change rule of a Hopfield network is given by

$$\begin{aligned} &\text{select randomly a neuron } u; \\ &s_u(k) := \operatorname{sign}(z_u(k)) \end{aligned} \tag{2}$$

where $\operatorname{sign}(u) = 1$ if $u > 0$ and $\operatorname{sign}(u) = -1$ otherwise. So once the neuron that may change its state is selected, the state change is completely deterministic.

The state change rule of a Boltzmann machine is given by

$$\begin{aligned} &\text{select randomly a neuron } u; \\ &r := \text{random number in } [0, 1); \\ &\text{if } r < f_\beta(2z_u(k)) \text{ then } s_u(k) := 1 \text{ else } s_u(k) := -1 \end{aligned} \tag{3}$$

where $f_\beta(y) = 1/(1+e^{-\beta y})$ is the logistic activation function. Hence, the selected neuron goes to state '1' with probability $f_\beta(2z_u(k))$ and to state '-1' with probability $1-f_\beta(2z_u(k))$. The factor '2' in these expressions is not essential; it only simplifies forthcoming formulas. Note that the state change mechanism of a Boltzmann machine is not deterministic. The net input determines only the probability that the next state is '-1' or '1.' It follows from the form of the logistic activation function that a large positive net input leads with a high probability to state '1,' and a large negative net input leads with a high probability to state '-1.' The parameter $\beta > 0$ determines the steepness of the logistic function f_β. For very large β values, f_β is almost a step function, which implies that a positive net input almost always leads to state '1' and a negative net input almost always leads to state '-1.' Hence for $\beta \to \infty$, a Boltzmann machine behaves like a Hopfield network.

An important role in the analysis of the dynamics for both types of two-state neural networks is played by the energy function $E: \mathscr{R} \to \mathbb{R}$, defined by

$$E(k) = -\tfrac{1}{2} \sum_{u,v \in \mathscr{U}} w_{u,v} s_u(k) s_v(k) + \sum_{u \in \mathscr{U}} b_u s_u(k). \tag{4}$$

The term energy function stems from the analogy with physical systems, as treated in Sect. 5. The energy is a global measure indicating to what extent the neurons in the network have reached a consensus about their individual states relative to the weights and biases. Let the network be in configuration k, and suppose a state change of a neuron u results in a configuration l, with $s_u(l) = -s_u(k)$ and $s_v(l) = s_v(k)$ for all neurons $v \neq u$. A simple computation shows that the energy difference between the new and old states is given by

$$E(l) - E(k) = 2 s_u(k) z_u(k). \tag{5}$$

3. Hopfield Networks

A configuration of a Hopfield network is called stable if no neuron can change its state anymore. From Eqn. (1) we infer that a stable configuration k satisfies

$$s_u(k) = \mathrm{sign}(z_u(k)) \tag{6}$$

for all neurons u. It is easy to show that a state transition of a Hopfield network always leads to a decrease in the energy E. Hence, for any start configuration, the network always reaches a stable state by repeated application of the state change mechanism. Goles-Chacc et al. (1985) showed that finding a stable state may require a number of transitions that cannot be bounded by a polynomial in the number of neurons. This result has been generalized by Schäffer and Yannakakis (1991) who showed that the problem of finding stable states in Hopfield networks is PLS-complete. This result implies that it is unlikely that algorithms exist that find a stable state in a Hopfield network with a worst-case running time that can be bounded by a polynomial in the size of the network.

Since a Hopfield network always converges to a stable configuration, it can be used as an associative memory, in which the stable configurations are the stored patterns. The idea is that, starting with a corrupted pattern as initial configuration, repeated application of the state change mechanism will lead to a stable configuration, which is hopefully the original pattern. The question is how the weights and thresholds must be chosen to obtain a given set of stable configurations. Let l be a pattern that we want to store in a Hopfield network, i.e., l is a $|\mathscr{U}|$-dimensional vector with components $s_u(l)$. It can be shown that l is a stable configuration if the weights are chosen as $w_{u,v} = s_u(l)s_v(l)$ for all $u \neq v \in \mathscr{U}$ and the thresholds as $b_u = 0$ for all $u \in \mathscr{U}$. If we want to store a set $\mathscr{L}$ of patterns, then an appropriate choice for the weights is

$$w_{u,v} = \frac{1}{|\mathscr{U}|} \sum_{l \in \mathscr{L}} s_u(l)s_v(l) \tag{7}$$

for all $u \neq v \in \mathscr{U}$ with biases $b_u = 0$ for all $u \in \mathscr{U}$. This formula, which is a variant of the rule of Hebb (1949), however, may result in configurations that are not stable in the sense defined above. Moreover, it may introduce stable configurations that do not belong to the set of patterns; these are called spurious configurations. The capacity of this type of associative memory, i.e., the number of patterns that can be stored in a Hopfield network of given size, is considered in Sect. 7.

4. Boltzmann Machines

Suppose a Boltzmann machine goes from configuration k to configuration l with the state change rule as given Sect. 2. This means that first a neuron u is selected, and second, the state of u is changed. It is easily seen that each possible state change of u, from state '–1' to '1' or from state '1' to '–1', has probability $f_\beta(-2s_u(k)z_u(k))$. Hence, using Eqn. (5) the probability of a state change of neuron u equals

$$\mathbb{P}\{\text{accept a state change of neuron } u \,|\, \text{current configuration is } k\} = \frac{1}{1+\exp(\beta(E(l)-E(k))} \tag{8}$$

Clearly, state changes that decrease energy have a higher probability than state changes that increase the energy. Note the difference with the Hopfield network, where state changes that increase energy are not possible.

The process that governs state changes in a Boltzmann machine is known as simulated annealing. It was first proposed by Metropolis et al. (1953) for computing low-energy states of solids. Kirkpatrick et al. (1983) first used the method to find near-optimal solutions to hard combinatorial optimization problems. Ever since, simulated annealing has become a well-known heuristic for finding approximate solutions of combinatorial optimization problems. The method originates from the simulation of physical annealing processes. The physical temperature T is related to the parameter β by $T = \frac{1}{\beta}$.

The mathematical treatment of simulated annealing uses the theory of Markov chains; for a detailed description we refer to Aarts and Korst (1989). The theory of Markov chains can also be used to describe the state changes of the neurons in a Boltzmann machine. The resulting iterative procedure can be described as a sequence of Markov chains, where each chain consists of a sequence of trials, and the outcome of a given trial only depends on the outcome of the previous trial in a probabilistic sense. A trial consists of two steps: given a configuration k, first a neighboring configuration l is generated, determined by a neuron $u \in \mathscr{U}$ that proposes a state change, and second, it is evaluated whether or not l is accepted; see Eqn. (9). If it is accepted, the outcome of the trial is l, otherwise it is k. Aarts and Korst (1989), among others, showed that for a Boltzmann machine with acceptance probability given by Eqn. (9) the following holds:

(a) The probability $q_\beta(k)$ of obtaining a configuration k after a sufficiently large number of trials carried out at a fixed value of β is given by

$$q_\beta(k) = \frac{\exp(-\beta E_k)}{Z(\beta)} \tag{9}$$

where $Z(\beta) = \sum_{l \in R} \exp(-\beta E_l)$ is a normalization constant that equals the well-known partition function in statistical physics. The expression of Eqn. (9) is often referred to as the equilibrium distribution of the corresponding Markov chain, and it is known in

statistical physics as the Boltzmann distribution, which explains the name 'Boltzmann machine.' The process of reaching the equilibrium distribution is called equilibration. It implies that low-energy configurations have a larger probability of occurring than high-energy configurations.

(b) For $\beta \to \infty$. Eqn. (9) reduces to a uniform distribution over the set of configurations with minimum energy. Or in other words, if β approaches ∞, the Boltzmann machine finds a configuration with minimum energy with probability one. This property makes it possible to use Boltzmann machines to obtain near-optimal solutions of combinatorial optimization problems.

Finally, we mention that the randomized state-equilibration method described above can be used to formulate a more sophisticated learning algorithm for the Boltzmann machine than the simple Hebbian rule that is applied in Hopfield networks. This algorithm was introduced by Ackley et al. (1985), and it follows an approach in which the weights are iteratively adjusted to minimize an information theoretic measure that quantifies the difference between the network performance in the clamped phase and the free running phase. To this end one distinguishes between visible neurons and hidden neurons. The objective for the Boltzmann machine is to learn a given behavior on its visible neurons. In the clamped phase the visible neurons are fixed to a required behavior, and only the hidden neurons can evolve. In the free running phase all neurons can evolve. The difference in behavior between these two phases is used to adapt the weights and thresholds of the Boltzmann machine. A more detailed presentation of the algorithm is considered beyond the scope of this paper, but can be found in Aarts and Korst (1989).

5. *The Physics Analogy*

The analogy between two-state neural networks and statistical many-particle systems is probably best explained from the well-known spin glass models in statistical physics (Amit et al. 1985). In a spin glass model one considers an ensemble of spin particles having spin 'up' or 'down,' i.e., $S_i = \pm 1$. The corresponding Hamiltonian is given by

$$H = -\tfrac{1}{2}\sum_{i,j} j_{ij} S_i S_j + h_0 \sum_i S_i \tag{10}$$

where the J_{ij} denotes the coupling strengths of the spin-spin interaction between two particles i and j, and h_0 denotes the strength of an external magnetic field.

The analogy with two-state neural networks is evident. The spins play the role of the states of the individual neurons and the Hamiltonian is the equivalent of the energy. The spin glass model described above is a well-studied system in statistical physics. Hence, its analogy to two-state neural networks permits the use of such concepts as phase transitions, equilibrium dynamics, and mean field theories in the analysis of the behavior of two-state neural networks; see for instance Amit et al. (1987).

As an example we consider the computation of statistical averages at equilibrium. For instance, using Eqn. (9) the average activity $\langle s_u \rangle_\beta$ of neuron u of a Boltzmann machine at temperature $T = \frac{1}{\beta}$ can be expressed as

$$\langle s_u \rangle_\beta = \frac{\sum_{k \in \mathscr{R}} s_u(k) \exp(-\beta E_k)}{\sum_{i \in \mathscr{R}} \exp(-\beta E_l)} \tag{11}$$

where $\langle \cdot \rangle_\beta$ denotes the average over all configurations in equilibrium at a given β-value. This expression can also be written as

$$\langle s_u \rangle_\beta = \frac{\partial F}{\partial b_u} \tag{12}$$

where F denotes the free energy given by

$$F = -\beta \log \sum_{k \in \mathscr{R}} \exp(-\beta E(k)) = -\frac{1}{\beta} \log Z(\beta). \tag{13}$$

It is a standard result from statistical physics that all macroscopic quantities at equilibrium can be calculated by taking the appropriate derivatives of the free energy.

6. *Mean Field Approach*

The idea of the mean field approach is to model statistical equilibrium from the point of view of the mean field experienced by a neuron. This means that we approximate the actual net input $z_u(k)$ of a neuron by the mean net input, i.e., the mean field, given by

$$\langle z_u \rangle_\beta = \sum_{v \in \mathscr{U}} w_{u,v} \langle s_v \rangle_\beta - b_u. \tag{14}$$

The state change rule Eqn. (3) yields that a neuron with this net input has probability $f(\langle 2z_u \rangle_\beta)$ of being in state '1' and probability $1 - f(\langle 2z_u \rangle_\beta)$ of being in state '–1.' Using the logistic activation function we find that the mean activation of the neuron satisfies

$$\langle s_u \rangle_\beta = \tanh\left(\beta\left(\sum_{v \in \mathscr{U}} w_{u,v} \langle s_v \rangle_\beta - b_u\right)\right). \tag{15}$$

Clearly, this is a system of $|\mathscr{U}|$ non-linear equations with $|\mathscr{U}|$ unknown variables, but they are at least deterministic. The solvability of the equations depends on the weights and the biases. For some specific values the solution is simple. For instance, in the case where $w_{u,v} = w_0/|\mathscr{U}|$ for all $u \neq v \in \mathscr{U}$ and $b_u = 0$ for all $u \in \mathscr{U}$, Eqn. (15) reduces to

$$\langle s_u \rangle_\beta = \tanh(\beta w_0 \langle s_u \rangle_\beta) \tag{16}$$

for all $u \in \mathscr{U}$. This corresponds to a very simple model of a ferromagnetic system for which in the limit $\beta \to \infty$ all particles (neurons) align in the same state, either spin up or spin down.

7. Storage Capacity

The mean field approach presented above can also be used to analyze the storage capacity of two-state neural networks. We suppose that we want to store a set $\mathscr{L}$ of (random) patterns by choosing weights and thresholds as in Sect. 3. Hence, Eqn. (15) reduces to

$$\langle s_u \rangle_\beta = \tanh\left(\frac{\beta}{|\mathscr{U}|} \sum_{v \in \mathscr{U}} \sum_{l \in \mathscr{L}} s_u(l) s_v(l) \langle s_v \rangle_\beta \right). \quad (17)$$

The question is whether these mean field equations have retrieval states, i.e., solutions that are approximately equal to one of the configurations l from $\mathscr{L}$. It turns out that this depends on the temperature $T = 1/\beta$ and on the quantity $\alpha = \frac{|\mathscr{L}|}{|\mathscr{U}|}$, which is the quotient of number of patterns that must be stored and the number of neurons in the network. The region in the T, α plane where retrieval states occur has been calculated by Amit et al. (1987), in the limit for $|\mathscr{U}| \to \infty$. The result is plotted in the phase diagram of Fig. 1. The figure shows four regions with different storage behavior. In region D there are no stable configurations at all. In all other regions there exist stable configurations. In region C the only stable states are the so-called spin glass states. These states are spurious, i.e., they have no relation with the patterns that must be stored. In regions A and B the network functions as a memory, i.e., there exist stable states that have a large overlap with the patterns from $\mathscr{L}$. In both regions there are, however, also spurious configurations. In region B the spurious configurations have on average lower energy values than the patterns; in region A the situation is the reverse. Consequently, the network works best in region A, because there the probability of finding a stable configuration that corresponds to pattern is maximum. The figure shows that the ability to store patterns is bounded by a critical value α_c, whose value is roughly equal to 0.138. Furthermore, for T-values greater than 0 the number of storable patterns decreases.

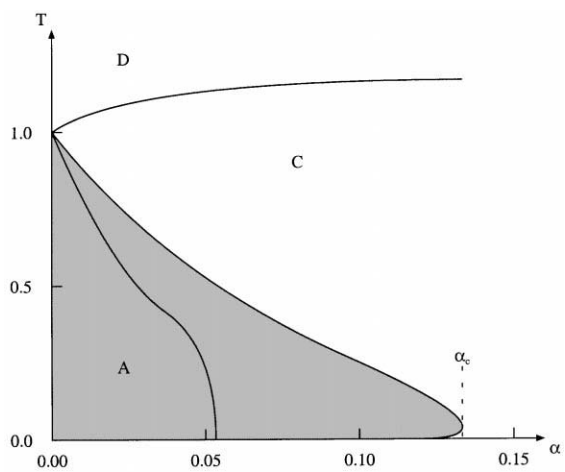

Figure 1
Phase diagram

8. Conclusion

We have analysed a special class of two-state neural networks using methods from statistical physics of many particle systems leading to specific results on the asymptotic behavior and storage capacity of these networks. The analysis can be further extended to include other behavioral properties such as phase transitions, entropy, and ensemble averages. We finally mention that there exists a one-to-one relation between two-state neural networks and many combinatorial optimization problems; the travelling salesman and the quadratic assignment problem probably being the best known examples. The physics analogy again allows analysis of specific properties of these combinatorial optimization problems, including ultrametric structures and probabilistic value analysis. For examples the reader is referred to Laarhoven and Aarts (1987).

See also: Artificial Neural Networks: Neurocomputation; Hebb, Donald Olding (1904–85); Linear Algebra for Neural Networks; Statistical Pattern Recognition

Bibliography

Aarts E H L, Korst J H M 1989 *Simulated Annealing and Boltzmann Machines*. Wiley, Chichester, UK

Ackley D H, Hinton G E, Sejnowski T J 1985 A learning algorithm for Boltzmann machines. *Cognitive Science* **9**: 147–69

Amit D J, Gutfreund H, Sompolinsky H 1985 Spin-glass models of neural networks. *Physical Review A* **32**: 1007–18

Amit D J, Gutfreund H, Sompolinsky H 1987 Statistical mechanics of neural networks near saturation. *Annals of Physics* **173**: 30–67

Goles-Chacc E, Fogelman-Soulie F, Pellegrin D 1985 Decreasing energy functions as a tool for studying threshold networks. *Discrete Applied Mathematics* **12**: 261–77

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Herz J A, Krogh A, Palmer P G 1995 *Introduction to the Theory of Neural Computation*. Addison Wesley

Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA* **79**: 2554–8

Kirkpatrick S, Gelatt Jr. C D, Vecchi M P 1983 Optimization by simulated annealing. *Science* **220**: 671–80

Laarhoven P J M van, Aarts E H L 1987 *Simulated Annealing; Theory and Applications*. Reidel, Dordrecht, The Netherlands

Metropolis N, Rosenblut A W, Rosenblut M N, Teller A H, Teller E 1953 Equations of state calculations for fast computing machines. *Journal of Chemical Physics* **21**: 1087–92
Schäffer A A, Yannakakis M 1991 Simple local search problems that are hard to solve. *SIAM Journal on Computing* **20**: 56–87

E. Aarts, H. ten Eikelder, and J. Korst

Neural Plasticity

Neural plasticity refers to the ability of neurons to change in form and function in response to alterations in their environment. Neurons function as parts of local circuits in the brain, and each neuron can change its functional role in a circuit by altering how it responds to inputs or influences other neurons. Neuronal interactions for computational functions largely depend on synaptic relationships, and neural plasticity largely results from changes in the strengths of synaptic connections between neurons and the formation of new connections. Neural plasticity occurs as neurons respond to the activities of neighboring neurons that are spontaneously active or are activated by events in the external environment, and to trophic and guidance factors released in the local environment. Neural plasticity is essential for the normal development of brain circuits, creating the differences in those circuits that make us individuals. Neural plasticity mediates the acquisition of knowledge and skill, and brain repair after injury. Plasticity may also lead to misperceptions and pain, and maladaptive behavior. An understanding of neural plasticity might help us promote this plasticity when it is useful, and control this plasticity when it has unfortunate consequences.

1. *Neural Plasticity During Brain Development*

The basic circuitry of the brain gradually emerges during development. While the mature brain remains plastic, developmental changes are especially impressive. For example, a short period of sensory loss, such as produced by having an eye covered by a patch early in childhood, can have a devastating effect on vision for that eye. The brain circuits for vision become dominated by the seeing eye so that connections and synaptic strengths change, deprived neurons fail to grow and axon branches withdraw, and a deficit in vision for the closed eye occurs that persists after the eye patch is removed (Antonini and Stryker 1993). A similar period of deprivation in adults has little effect. The developing brain is especially plastic because it is using information from the environment to help construct itself (Rauschecker 1991, Schatz 1990). Slight changes in the instructions from the environment, and the timing of those instructions can lead to quite different outcomes. In the mature brain, the circuits have been built, and further modifications are usually less profound.

One of the important accomplishments of developmental plasticity is to create functionally adaptive circuits in the brain. Developmental plasticity refers to the capacity of the brain to have different developmental outcomes. The development of some brain circuits, some of those in simple invertebrates for example, may be highly fixed. Thus, the developing circuits usually achieve the same adult form. But, this is not necessarily the case for the complex circuits of the higher brain centers of mammals, especially those in humans with long periods of development. The human brain is highly plastic because the brain circuits are not fully specified by the genes. All of the necessary information for proper development does not reside in the genes or gene products. Thus, it is necessary to use information from the environment to direct development. Neurons are designed to respond to sensory stimuli and each other with action potentials or spikes. Self-generated spontaneous activity and evoked activity are very important in mediating normal development (Schatz 1990). Neurons also influence the development of other neurons through the release of trophic and guidance factors (Connor and Dragunow 1998).

Consistencies in the internal and external environments during development typically lead to the development of similar brains with very similar brain circuits. Because the developmental outcomes can be so similar, the role of the environment may not always be obvious. Thus, one way to study developmental plasticity is to rear animals in abnormal environments (Rauschecker 1991). As examples, the development of the visual system can be altered by rearing mammals in the dark, with one eye closed, or under strobe lights. These experiments indicate that sensitivity to the environment is an important component of the normal development of brain circuits. In an abnormal environment, development can be misdirected, often with unfortunate functional consequences.

Sensitivity to the environment also allows the developing nervous system to become tailored to specifics of the environment of the individual. The newly hatched bird imprints on its parents, rather than birds in general, although even a moving box or balloon will do in the absence of tending parents. The young bird can become completely attached to such obviously false parents. The child that practices piano or basketball becomes the great piano or basketball player. Humans learn the sounds of their own language, and as adults find it hard to process the sounds of other languages. Thus, neural plasticity allows the brain to adapt to an uncertain environment, and permits specializing the formation of neural circuits within individuals for individual needs.

Because neuronal circuits are designed to be altered by aspects of the environment at certain times in development, or later during critical points in the

lifecycle, there are changes in circuits that are possible only at certain times, and these changes are often difficult or impossible to reverse. These times are called critical or sensitive periods. Eye patching for a period of days to weeks, for example, will greatly alter brain circuits, only during a fixed early period of visual development. Earlier or later periods of deprivation have little effect. There may be a critical or sensitive period for song birds learning songs, for human language to be easily and flawlessly acquired, and for learning perfect pitch.

In summary, neural plasticity is essential for normal brain development, especially for the development of large complex brains that mature slowly. The normal development of the human brain is especially dependent on neuronal plasticity. Many circuits require information from the environment to be designed properly. In addition, neuronal plasticity allows the development of the individual, and individual skills and abilities.

2. *Neural Plasticity and the Lifecycle*

Neural plasticity also allows brain circuits to function differently at different phases of the lifecycle. Hormones play a major role in adjusting neural circuits. Song birds, for example, adjust and even add new neurons to the neural circuits used in producing and acquiring songs used during the breeding season (Konishi 1985). At puberty, humans and other mammals alter brain circuits that mediate aspects of sexual behavior, and later brain circuits are adjusted to promote maternal and paternal behaviors. Thus, plasticity permits rapid and functionally important changes in brain circuits as life styles change with puberty and parenthood.

3. *Neural Plasticity and Learning*

Learning is based on neural plasticity. Each individual acquires a great deal of useful information about the environment through experience, and improves and perfects a great number of motor skills. The importance of learning varies with species, and it is probably most important in humans. There are several types of learning, but they all depend on modifying the functional circuits of neurons.

The neuronal mechanisms of learning are not fully understood, although they have been under intensive study for many years. It is likely that two types of induced changes in synaptic effectiveness between neurons are important in most learning. When a source of input to a neuron or group of neurons is intensely stimulated for a brief period of time, later stimulations might be more effective and produce a greater response. This change in synaptic effectiveness, which may persist for minutes to hours or longer, has been called long-term potentiation (LTP; see Bliss and Lomo 1973). Of course, new learning involves forgetting, so that connection strengths need to be weakened as well. Patterns of less intense, more distributed stimulation may reduce synaptic effectiveness. This reduction is called long-term depression (LTD; see Nelson and Turrigiano 1998). Possibly, LTP depends on reversing LTD.

Both LTP and LTD are likely to be critical components of memory formation and the acquisition of skills. LTP and LTD have been most studied in the hippocampus, both in a tissue slice preparation, which offers clear technical advantages, and in the intact animal. LTP and LTD are components of what has been called Hebbian plasticity (see Brown et al. 1988, Rauscheker 1991) after the neuropsychologist, Donald Hebb. In an effort to account for learning, Hebb theorized that the repeated activations of local neuron circuits changed the effectiveness of pathways so that some inputs came to activate the circuits more easily and more effectively. If neuron A takes part in the firing of neuron B, the future effectiveness of neuron A in firing neuron B is increased. The cellular mechanisms of Hebbian plasticity have been studied, and they involve the activation of the NMDA (N-methyl-D-aspartate) glutamate receptor and a calcium influx into the post synaptic neuron when pre synaptic activity is paired with post synaptic depolarization (Bear 1995, Rauschecker 1991). The hippocampus is known to have a critical early role in the formation of long-term memories, but it is not generally thought to be the site of permanent storage. The amygdala encodes the emotional component of memories formed during fear conditioning, but like the hippocampus, the neuronal plasticity that occurs in the amygdala may modulate the consolidation of memories formed elsewhere.

Most memories and skills depend on neocortex. LTP and LTD also have been demonstrated in neocortex circuits where these types of synaptic plasticity are likely to be important in the consolidation and storage of information.

4. *Neural Plasticity and Brain Repair*

Neural plasticity is also important in brain reorganization and repair after injury. The effects of sensory loss have been extensively studied in the sensory systems of adult mammals (Kaas 1991, Kaas et al. 1999). Sensory loss of one type or another commonly occurs during adult life. For example, humans typically lose auditory receptors and, especially, high-frequency hearing as they age. Studies in monkeys, cats, and rodents show that auditory cortex reorganizes after such partial hearing loss so that more neurons are devoted to processing the remaining inputs. The functional significance of this reorganization is not clear, but it may help individuals adjust to the hearing loss and more effectively use the remaining inputs. Similar changes occur when small portions of the retina are damaged. The regions of visual cortex that are deprived of their

normal sources of activation by such damage soon come to be responsive to surrounding intact parts of the retina.

Finally, any restricted loss of tactile inputs from the skin, such as those caused by nerve damage, is followed by a reorganization of somatosensory cortex so that deprived neurons become responsive to preserved tactile afferents. Such findings indicate that the brains of even adult mammals have considerable capacity for functional reorganization so that neurons that are deprived of a normal source of activation gradually acquire new sources of activation. Unresponsive neurons are not of much use, and the reactivation of deprived neurons, as a common feature of brain damage, may greatly contribute to behavior recovery and compensation. The mature brain may well be designed to compensate for some neuronal loss and damage, and reorganization may lead to more effective use of remaining neurons and neural circuits.

Humans and other mammals compensate remarkably well from cortical damage that is relatively restricted. In part, this may be the result of reorganization within the damaged area of cortex. Sensory and motor representations in cortex reorganize after damage to recover some of the lost parts of the representation. Other intact areas of cortex may reorganize to help compensate for the loss. During some period of brain development, parts of the brain may be capable of regenerating lost neurons (Kolb et al. 1998).

Impressive recoveries after brain damage may reflect design features involving plasticity, or even the fortunate utilization of mechanisms built in by selection for normal development and adult learning. However, plasticity is not always beneficial. The plasticity that is there for one set of purposes may be evoked in other sets of circumstances where the results are detrimental and unfortunate. As one example, individuals with amputated limbs typically feel that the missing limb is still there (Ramachandran and Blakeslee 1998). This may be the result of neurons in the brain, deprived of normal sources of activation by afferents from the missing limb, becoming more sensitive to intrinsic signals or other inputs. The growth of new connections in the brain stem and cortex seems to add the further complexity of trigger zones, locations on the face and stump that, if touched, evoke a sensation in the missing limb. Fortunately, this type of misperception is not very disturbing. However, severe pain also might be felt in the missing limb, apparently as a result of the plasticity of circuits in the thalamus. This pain is very difficult to treat since it is not generated in the normal way.

As another example of maladaptive plasticity, individuals that overpractice a skill may develop motor disorders or dystonias (Byl et al. 1996). Neural plasticity is essential for developing motor skills, but musicians and writers may overpractice or overuse a skill so that performance deteriorates rather than improves. The practice may coactivate neurons so frequently that they become so strongly interconnected that they fail to respond independently when needed in other neuronal circuits. Thus, the musician loses control of the skilled hand movements needed to play, and the writer develops writer's cramp. Recoveries depend on other training and experiences that rebalance the motor circuits.

5. *Mechanisms of Neural Plasticity*

Brain circuits are plastic because new connections form, unused connections are lost, and existing connections increase or decrease in effectiveness. Developing brains are especially plastic because it is a time of great neuronal growth. Activity-dependent genes (Kaczmarek and Chaudhuri 1997) within neurons function to promote and direct new growth, or stop growth, and cause retractions of axons and dendrites and even cell death. The developing nervous system over-produces neurons and connections, and uses activity patterns induced in neurons by the external and internal environments to select the useful, more active neurons and connections and eliminate other less active neurons and connections. By altering activity patterns through lesions, drugs, and experience, outcomes can be changed so that normally lost neurons are saved, normally lost connections are preserved, and even new growth occurs. One of the early discoveries was that rats reared in a complex rather than a simple environment (Greenough et al. 1985) developed a thicker cortex with larger neurons and more complex dendritic arbors. As a more unusual, more dramatic, example of developmental plasticity, rodents reared after a major loss of auditory inputs to the forebrain grow visual inputs into the auditory thalamus, so that auditory cortex is activated by visual stimuli (Sur et al. 1988). The potential for the growth as well as the loss of new connections is great in the developing brain.

Neuronal growth and death can be a feature of normal lifestyle changes as well. Neurons may increase in size and connectional complexity in brain regions related to sexual behavior or bird song during appropriate parts of the breeding season, and then shrink and disconnect later (Konishi 1985). Neurons may even be born and added to such circuits, to later die. While neuronal growth and even death are major components of programmed plasticity for expected developmental and lifecycle events, much less new growth or retraction might be expected at other times in the mature brain, which has developed brain circuits for a period of stable performance. Indeed, except for learning and learning mechanisms, the mature brain has been considered to be rather fixed in organization and not plastic. Now, considerable evidence has accumulated to show that the functional organization of the mature brain is quite mutable, and that at least some of the changes depend on new growth. In cases

of severe sensory loss, such as with brain or spinal cord lesions, or peripheral nerve damage, growth of axons over considerable distances occurs to reactivate deprived neurons in the spinal cord, brainstem, and cortex. While extensive new growth in the mature brain is often likely to be detrimental, and glial cells in the central nervous system produce factors that inhibit such new growth, the local growth of axons and dendrites is probably ongoing and common. In addition, the cellular factors for inhibiting more extensive growth may be overcome by extreme deprivation. It remains to be demonstrated how much new growth contributes to behavioral recoveries after brain damage, but a major contribution by new growth seems likely. In addition, new neurons appear to be added to neocortex and other parts of the forebrain throughout life (Gould et al. 1999, Van Praag et al. 1999). While neurogenesis had been thought to be limited to the period of brain development, recent evidence for the addition of neurons to the mature primate brain raises the possibility that such new neurons contribute to brain plasticity.

A common and possibly more powerful factor in neural plasticity is increasing or decreasing the effectiveness of existing connections. This can be done by increasing or decreasing the number, sizes, and positions of synapses, the amount of neurotransmitter released, and the receptor sites responsive to the transmitter. The important roles of LTP and LTD have already been discussed. These mechanisms may be potentiated by the appropriate release of neuromodulators that increase neural activity for short but behaviorally significant periods of time. Thus, stressful and other motivating events activate brainstem neurons that release the neurotransmitter, acetylcholine (ACh) throughout the brain where synaptic strengths are likely to be modified. This makes the changes in neural circuits more likely, so that the motivating events can be retained in memory and skills can be acquired more readily. The permissive effects of ACh release on neural plasticity can even be obtained artificially by electrically stimulating the brainstem neurons that project to cortex and release ACh during learning (Kilgard and Merzenich 1998).

There are many other ways of promoting neural plasticity. Neurons tend to self-regulate through the actions of activity-dependent genes to reach a set level of activity. Neurons with lowered levels of activity tend to express lower levels of the inhibitory neurotransmitter, GABA, or become less sensitive to GABA by making fewer GABA receptors (Hendry and Jones 1998). More active neurons make more GABA and more GABA receptors. Thus, deprived neurons become more sensitive and respond to inputs that were previously subthreshold and ineffective, while overactive neurons lose responsiveness to weak but previously effective inputs. Other activity-dependent genes probably regulate neuron growth through the release of trophic and guidance factors.

6. *Conclusions*

Neural plasticity is important for the normal development and maintenance of neural circuits. This plasticity adjusts mature brain circuits to mediate individual perceptual and motor skills, and adaptive adjustments to an uncertain environment. Neural plasticity is especially important in storing information. Plasticity also functions in brain repair. Because some outcomes of neuronal plasticity are highly desirable, while others are unfortunate, a comprehensive understanding of the mechanisms of plasticity and how to control them would be clinically and therapeutically beneficial.

See also: Long-term Depression (Cerebellum); Long-term Depression (Hippocampus); Long-term Potentiation and Depression (Cortex); Long-term Potentiation (Hippocampus); Neural Plasticity in Auditory Cortex; Neural Plasticity in Visual Cortex; Neural Plasticity of Spinal Reflexes

Bibliography

Antonini A, Stryker M P 1993 Rapid remodeling of axonal arbors in the visual cortex. *Science* **260**: 1819–21

Bear M F 1995 Mechanism for a sliding synaptic modification threshold. *Neuron* **15**: 1–4

Bliss T V P, Lomo T 1973 Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* **232**: 331–56

Brown T H, Chapman P F, Kairisc E W, Keenan C L 1988 Long-term synaptic potentiation. *Science* **242**: 724–28

Buonomano D V, Merzenich M M 1998 Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience* **21**: 149–86

Byl N N, Merzenich M M, Jenkins W M 1996 A primate genesis model of focal dystonia and repetitive strain injury: I: Learning-induced dedifferentiation of the representation of the hand in the primary somatosensory cortex in adult monkeys. *Neurology* **47**: 508–20

Connor B, Dragunow M 1998 The role of neuronal growth factors in neurodegenerative disorders of the human brain. *Brain Research Reviews* **27**: 1–39

Gould E, Reeves A V, Graziano S A, Gross C G 1999 Neurogenesis in the neocortex of adult primates. *Science* **286**: 548–52

Greenough W T, Hwang H M, Gorman C 1985 Evidence for active synapse formation or altered postsynaptic metabolism in visual cortex of rats reared in complex environment. *Proceedings of the National Academy of Sciences USA* **82**: 4549–52

Hendry S H, Jones E G 1988 Activity-dependent regulation of GABA expression in the visual cortex of adult monkeys. *Neuron* **1**: 701–2

Kaas J H 1991 Plasticity of sensory and motor maps in adult mammals. *Annual Review of Neuroscence* **14**: 137–67

Kaas J H, Florence S L, Jain N 1999 Subcortical contributions to massive cortical reorganizations. *Neuron* **22**: 657–60

Kaczmarek L, Chaudhuri A 1997 Sensory regulation of immediate-early gene expression in mammalian visual cortex: Implications of functional mapping and neural plasticity. *Brain Research Reviews* **23**: 237–56

Kilgard M P, Merzenich M M 1998 Cortical map reorganization enabled by nucleus basalis activity. *Science* **279**: 1714–18
Kolb B, Gibb R, Grazyna G, Whishaw I Q 1998 Possible regeneration of rat medial frontal cortex following neonatal frontal lesions. *Behavioral Brain Research* **91**: 127–41
Konishi M 1985 Birdsong: From behavior to neuron. *Annual Review of Neuroscience* **8**: 125–70
Nelson S B, Turrigiano G G 1998 Synaptic depression: A key player in the cortical balancing act. *Nature Neuroscience* **1**: 539–41
Ramachandran V S, Blakeslee S 1998 *Phantoms in the Brain*. Morrow, New York
Rauschecker V P 1991 Mechanisms of visual plasticity: Hebb synapses, NMDA receptors, and beyond. *Physiological Reviews* **71**: 587–615
Schatz C J 1990 Impulse activity and the patterning of connectivity during CNS development. *Neuron* **5**: 745–56
Sur M, Garraughty P E, Roe A W 1988 Experimentally induced visual projections into auditory thalamus and cortex. *Science* **242**: 1437–41
Van der Loos H, Dovfl J 1978 Does the skin tell the somatosensory cortex how to construct a map of the periphery? *Neuroscience Letters* **7**: 23–30
Van Praag H, Kempermarin G, Gage F H 1999 Running increases cell proliferation and neurogenesis in the adult mouse dentate gyrus. *Nature Neuroscience* **2**: 266–70

J. H. Kaas

Neural Plasticity in Auditory Cortex

This article covers the topic of neural plasticity in the auditory cortex, particularly with reference to learning and memory in adult subjects. As used here, the term 'neural plasticity' refers to systematic long-term (minutes to months) changes in the responses of neurons to the same physical stimulus (e.g., a tone), due to experience.

Neural plasticity in the auditory cortex is interesting not only in itself but also as a case study in the intersection of two scientific fields that had developed quite separately, those of sensory physiology and the neurobiology of learning and memory. Furthermore, this topic provides a clear example of how assumptions constrained thought and experiment for most of the twentieth century. Finally, because the majority of studies on learning and sensory systems have been carried out in the auditory cortex, the development of new ideas and findings in the auditory cortex has direct implications for studies of neural plasticity in adult somatosensory and visual cortices (see *Neural Plasticity in Visual Cortex*).

The history of neuronal plasticity in the auditory cortex involves four stages: (a) sensory-motor framework excluding the auditory cortex from learning and memory; (b) delineation of cortical responses to sounds in anesthetized animals; (c) revelation of plasticity of responses to sounds during learning; and (d) discovery that learning systematically changes the frequency tuning of neurons in the auditory cortex.

1. *Traditional Functional Organization of the Cerebral Cortex*

Attempts to understand the auditory cortex (as well as many other brain systems and structures) have their origin in the nineteenth century, within the framework of a sensory-motor conception of the nervous system. In the 1820s, the French physiologist F. Magendie, building on the partially correct work of the English anatomist C. Bell, showed that the dorsal and ventral roots of spinal nerves have different functions, sensory and motor, respectively. In the 1840s, Marshall Hall, an English physiologist, discovered that the sensory and motor nerves, together with the internal substance of the cord, constituted a behavioral structural–functional unit, that of the spinal reflex. The question of whether or not the entire nervous system was organized on sensory-motor principles formed much of the research program for the rest of the century (Young 1970).

It was generally believed that the cerebral cortex was not electrically excitable and operated only in a holistic manner, as it was widely thought to be the location of the unitary self or soul. However, in 1870 Fritsch and Hitzig, two German medical students, used focal electrical stimulation to discover the motor cortex. Also during the last quarter of the nineteenth century, David Ferrier, an English physiologist, used brain lesions to approximately locate the auditory cortex in the temporal lobe, by showing impaired hearing in animals. Thus by the beginning of the twentieth century the sensory-motor principle had been extended to the cerebral cortex. However, many regions of the cortex appeared to be neither sensory nor motor. About 1910, several of these areas were labeled the 'association cortex' by the anatomist W. LeGros Clark. The dominant implication of this formulation was that sensory and motor processes converged in these areas to form learned associations. Implicit in this schema was the concept that sensory (and motor) regions of the cortex were not sites of learning and memory.

2. *Auditory Cortical Responses to Acoustic Stimuli*

The availability of sensitive electronic amplifiers in the 1930s enabled scientists to record brain potentials that were elicited by controlled sensory stimuli, hence initiating the modern field of sensory neurophysiology. In the United States, recordings in the 1940s by Woolsey and Walzl in the cat and by Tunturi in the dog delineated auditory cortical fields. In particular, their studies revealed that the primary auditory cortex exhibits a 'tonotopic map,' i.e., that cells best tuned to one frequency were located next to cells that were tuned to an adjacent frequency. In effect, the primary auditory cortex has an organization for acoustic frequency that is similar to that of a keyboard or a

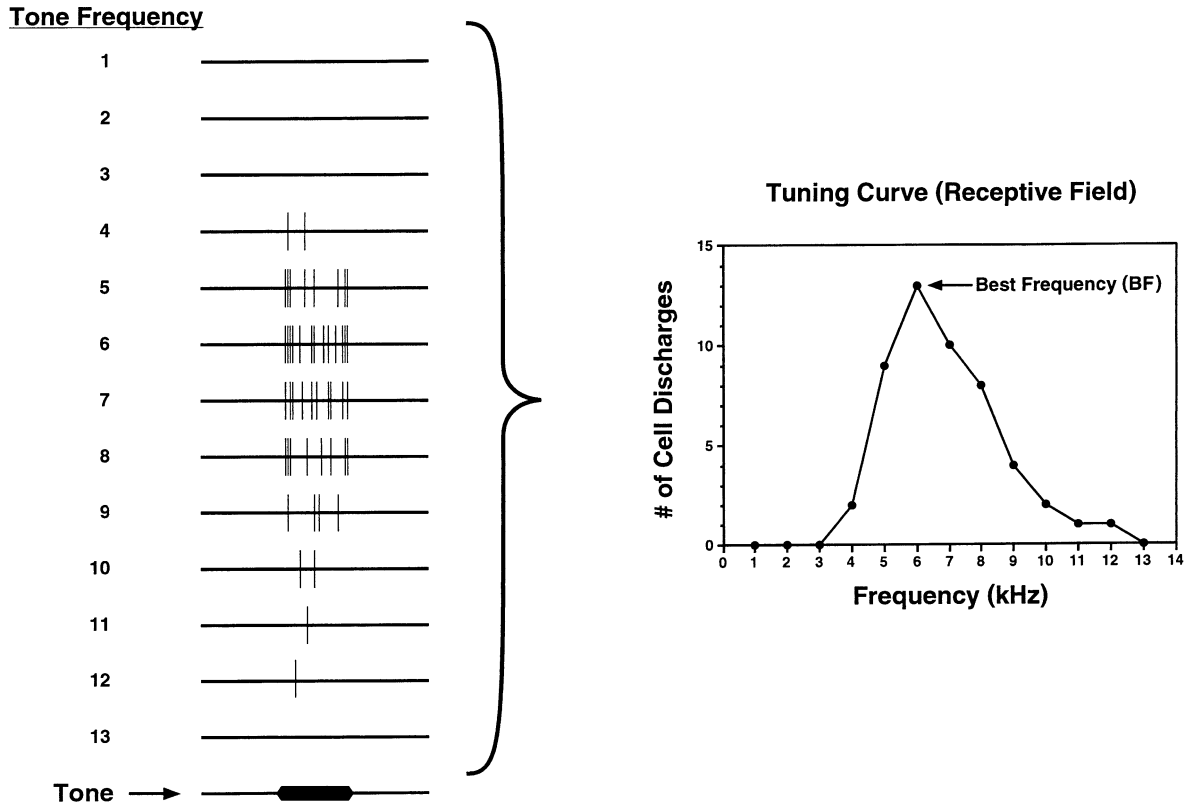

Figure 1
An example of a frequency-receptive field. The left panel shows neuronal discharges against tones of various frequencies. The right panel shows the tuning curve plot of these data. The frequency of maximal response is called the 'best frequency'

harp. Subsequent more detailed studies showed that single neurons exhibited tuning functions, responding best to one frequency (the 'best frequency') and less strongly to other, higher and lower adjacent frequencies. Tuning curves are also called 'frequency receptive fields' because they show the regions of acoustic frequency to which neurons are receptive, i.e., to which they respond (Fig. 1).

3. *Initial Behavioral Studies of Cortical Plasticity*

While auditory neurophysiology was determining how the auditory cortex responds to sound, the field of learning and memory independently began to study how responses of the brain change when a stimulus becomes behaviorally important or loses its importance. Robert Galambos and his colleagues in Washington, DC performed a seminal experiment in 1955. Cats were classically conditioned by pairing an auditory conditioned stimulus (CS) with a puff of air to the face (e.g., see *Classical Conditioning, Neural Basis of*). Evoked potentials elicited by the CS became larger during conditioning; complementary studies of habituation, i.e., repeated presentation of a tone without reinforcement, by Irwin Westenberg, later revealed the opposite effect, a decrement of cortical responses to the repeated stimulus. Over the next decades, numerous additional studies in various animals and training situations demonstrated that learning systematically alters evoked potentials and neuronal discharges in the auditory cortex elicited by conditioned stimuli or habituated stimuli.

The findings *in toto* showed that the responses of the auditory cortex to sounds were affected not only by the physical sounds themselves, but also by the learned psychological or behavioral importance of acoustic stimuli. These findings clearly showed the error of the traditional belief that sensory cortices had purely sensory functions and were not regions directly involved in learning and memory.

However, the documentation of neural plasticity had little effect on the field of sensory neurophysiology, probably for several reasons. First, learn-

ing studies used only one or two tones, which were not interesting to sensory workers who used many stimuli. Second, there appears to have been conceptual confusion in which sensation was equated with perception so that the constructive aspect of the latter was not appreciated; hence plasticity of sensory responses was mistakenly seen as incompatible with perceptual accuracy. Neither was sensory cortical plasticity initially influential within the field of learning and memory itself. This relative neglect appeared to reflect the traditional belief in association cortex as sites of learning and memory on the one hand and emphasis on the hippocampus and other noncortical structures on the other hand.

4. *Receptive Field Plasticity in Learning and Memory*

The first evidence that learning actually changes receptive field (RF) tuning was obtained in 1986 at the University of California Irvine. As this initial experiment concerned so-called 'associative' areas of the auditory cortex, rather than the tonotopic auditory cortex, it did not address the question of whether learning actually changed tuning in the primary sensory cortex. Then in 1990 Bakin and Weinberger recorded RFs from cells in the primary auditory cortex before and immediately after conditioning to a tonal conditioned stimulus. They found that responses to the CS frequency increased whereas responses to the best frequency (BF), and many other frequencies, decreased. These opposing changes were sufficient to shift frequency tuning toward or to the frequency of the CS (Fig. 2). Tuning shifts were caused only by association between a tonal CS and an unconditioned stimulus because sensitization training was ineffective. Also, this plasticity was not caused by arousal to the CS frequency because RF shifts are still observed after training when subjects are placed under general anesthesia, a state lacking arousal to stimuli.

Several follow-up experiments showed that this neural plasticity was highly specific to the frequency of the CS; responses to tones a small fraction of an octave away were unchanged or decreased. Also, discrimination training (one tone reinforced, another tone not

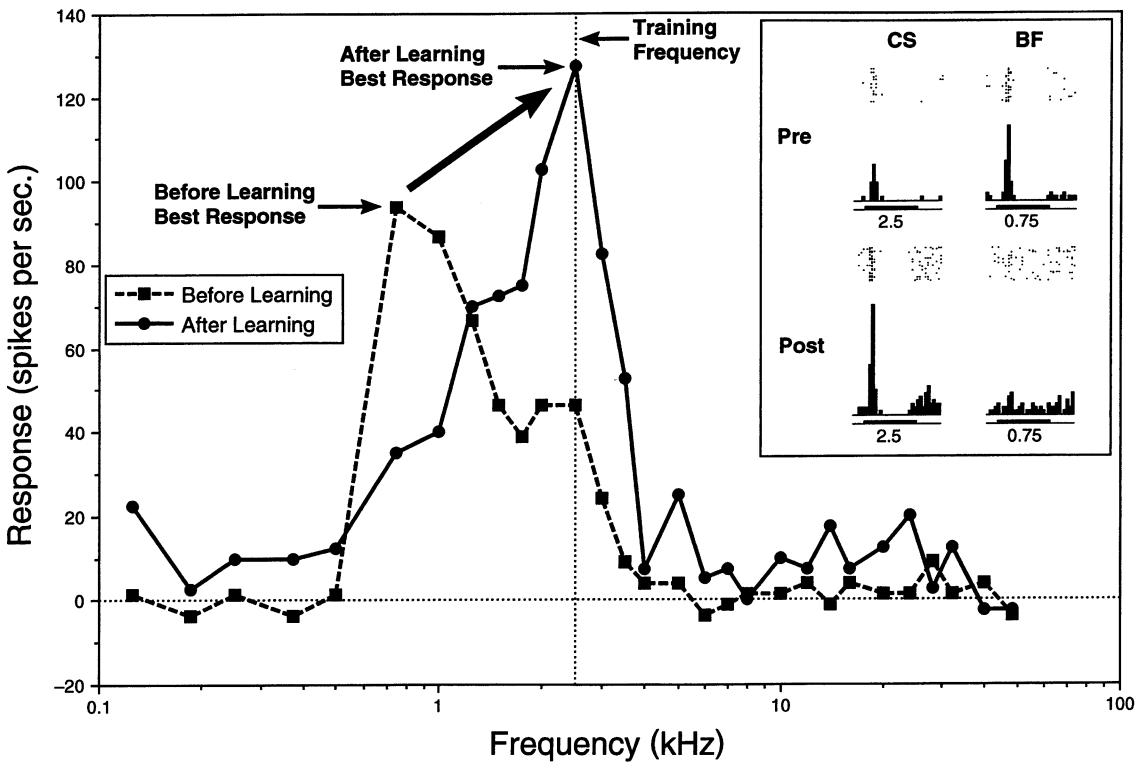

Figure 2
Learning shifts receptive field tuning to the frequency of the conditioned stimulus. The best frequency before learning was 0.75 kHz. The subject was trained with a CS of 2.5 kHz. After training the tuning shifted so that 2.5 kHz became the best frequency. The inset shows the opposite changes in number of discharges against each frequency before and after learning

reinforced), produced increased responses to the former and decreased responses to the latter. Further, RF plasticity developed within only five trials of conditioning and could be retained as long as eight weeks after training. Moreover, instrumental avoidance training also produced the same shifts in tuning, both for one-tone training and for two-tone discrimination training. Habituation produced the opposite type of RF plasticity, a decrease in response that was highly specific to the frequency of the repeated tone. Finally, RF plasticity was found to exhibit a key characteristic of memory, that of consolidation, i.e., for periods of hours to days after a single training session, the RF plasticity became stronger and more specific. Thus, neural plasticity of receptive fields in the primary auditory cortex shares all of the major characteristics of associative memory; in addition to being associative, RF plasticity is highly specific, discriminative, develops very rapidly, is very long lasting and shows consolidation. Therefore, RF plasticity may be considered to be 'physiological memory' (Weinberger 1998).

RF plasticity has been reported in many laboratories and several species and similar effects have been reported in studies of brain metabolism in animals. Moreover, brain-imaging studies also have found RF plasticity in the human auditory cortex during learning (Morris et al. 1998).

5. *Future Directions*

The mechanisms of RF plasticity are unknown but are currently under investigation (Edeline 1999). There is strong evidence that the induction of RF plasticity involves the release of a modulatory transmitter, acetylcholine, into the auditory cortex from the *nucleus basalis* of the basal forebrain. Responses of its cells to a CS increase during conditioning. Moreover, pairing a tone with stimulation of the *nucleus basalis* produces the same type of tuning shifts observed during normal behavioral learning and pharmacological blockade of cholinergic muscarinic receptors in the auditory cortex prevents such tuning shifts.

The functional role of receptive field plasticity in the auditory cortex is also under investigation. One hypothesis is that it represents a memory code for the psychological importance of stimuli; that is, the brain stores the acquired significance of sensory experiences by increasing the number of neurons that become tuned to the relevant stimulus. Such a memory code could explain why one is more likely to hear one's own name in a noisy room than a random name. Also, it could explain why the loss of memory in aging and brain degeneration is less severe for the most important memories of a lifetime. Alleviation of deafness by electrical stimulation of the inner ear ('cochlear implant') involves new learning and thus probably RF plasticity. Future studies of neural plasticity in the auditory cortex are likely to be directed toward its cellular and molecular mechanisms, other sensory systems, and psychological and behavior functions.

See also: Auditory System; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Neural Plasticity; Neural Plasticity in Visual Cortex

Bibliography

Edeline J M 1999 Learning-induced physiological plasticity in the thalamo-cortical sensory systems: A critical evaluation of receptive field plasticity, map changes and their potential mechanisms. *Progress in Neurobiology* **57**: 165–224

Morris J S, Friston K J, Dolan R J 1998 Experience-dependent modulation of tonotopic neural responses in human auditory cortex. *Proceedings of the Royal Society (Series B)* **265**: 649–57

Weinberger N M 1998 Physiological memory in primary auditory cortex: Characteristics and mechanisms. *Neurobiology of Learning and Memory* **70**: 226–51

Young R M 1970 *Mind, Brain and Adaptation in the Nineteenth Century. Cerebral Localization and its Biological Context from Gall to Ferrier*. Clarendon Press, Oxford, UK

N. M. Weinberger

Neural Plasticity in Visual Cortex

1. *Introduction*

Plasticity in the visual cortex refers to visual cortical neurons altering their response properties according to long-lasting changes of visual input. Within visual cortical areas, neurons exhibit a high degree of feature selectivity and are grouped into small columns according to their response preferences. These columns are located at regularly spaced intervals covering the whole cortical representation of the visual field with a modular system of feature selective neurons (see *Topographic Maps in the Brain*; *Visual System in the Brain*; *Occipital Lobe* (*Visual Cortex*): *Functional Aspects*; *Cortical Activity*: *Differential Optical Imaging*). The selectivity of these cells and their modular arrangement is thought to emerge from interactions in the network of specific thalamocortical and corticocortical connections. Therefore, plasticity involves processes ranging from modification of synaptic coupling strength to rearrangement of neuronal circuits like outgrowth and elimination of connections.

In early postnatal development, both thalamocortical and corticocortical connections are known to

be especially malleable by visual experience during a phase called the critical or sensitive period (see also *Sensitive Periods in Development*, *Neural Basis of*). In the adult, experience-dependent plasticity of thalamocortical connections seems to be very limited. However, some plasticity of corticocortical and in particular long-range intra-areal connections persists ranging from synaptic gain changes to even axonal outgrowth (see also *Motor Cortex*).

Furthermore, there is accumulating evidence that certain neuronal response properties are dynamic rather than fixed entities since they can change within minutes to hours even in the adult brain. Therefore, the transition between the visual perception process, short-term adaptation to visual context, learning and long-term adaptive changes might be continuous.

In this article, mainly experimental evidence of plasticity in primary visual cortex of cats or monkeys during development and in the adult will be reviewed. Links to the human will be suggested.

2. Neural Plasticity in the Visual Cortex During Development

After birth, the visual cortex of most mammals is still in a period of growth and change. The majority of neuronal response properties and connections develop only postnatally and maturation in the visual cortex seems to proceed in two distinct steps (Frégnac and Imbert 1984, Singer 1995, Schmidt et al. 1999). In the first phase, mechanisms inherent to the cortex establish a crude framework of interconnected neural modules which exhibit the basic but still immature traits of the adult state. Relevant mechanisms in this phase are assumed to consist of molecular cues and patterns of spontaneous neural activity in cortical and corticothalamic circuits. Experience-independent development is prolonged beyond eye opening in most species. It might serve to prevent substantial activity-dependent remodeling before the sensory periphery and the motor control of the eyes have matured sufficiently. In a second phase, the sensitive period, the primordial layout becomes refined under the control of visual experience establishing a fine-tuned network of connections and mature response properties. Neuronal connections and cortical response properties are especially susceptible to visual experience during the latter phase. One of the main functions attributed to the sensitive period is to consolidate already-established connections and to specify still-proliferating connections according to functional criteria. This process might serve to adapt the developing organism ideally to the requirements of the outside world (Singer 1995, Schmidt et al. 1999).

However, the individual outcome is not always favorable. Very early, clinical observations indicated that the ability to see in the adult might be irreversibly impaired if the unrestricted use of vision during early childhood is not assured (von Senden 1932) (see also *Visual Development*: *Infant*). Congenital or early childhood cataract, corneal blurring after eye infections, and large uncorrected refractive errors prevent the development of normal vision of the affected eye when the problem is not treated within the first years of life. Misalignment of the two eyes by either strabismus or anisometropia, a refraction asymmetry between the eyes, can lead to permanent suppression of one eye resulting in poor vision (amblyopia), although the optical apparatus is intact (Duke-Elder and Wybar 1973, von Noorden 1990). Milder consequences of these disturbances are a lack of binocular functions such as stereo vision. Structural alterations reflecting observations in the human were studied in animal experiments in monkeys and cats pioneered by Hubel and Wiesel (1963).

2.1 Animal Models: Plasticity of Thalamocortical and Intracortical Connections During Development

Thalamocortical connections and especially their specific alignment onto cortical neurons determine the selectivity of neuronal responses regarding ocularity and, partly, stimulus orientation.

The segregation of eye-specific thalamocortical afferents into ocular dominance columns seems to reach a limited level without vision. Similarly, neuronal responses to oriented contours and the organization of orientation-selective neurons into modular cortical maps improve independently of visual experience in the first phase of development (Frégnac and Imbert 1984, Schmidt et al. 1999). Besides thalamocortical connections, intracortical circuits and especially long-range intra-areal connections, which run horizontally through the cortex, influence neuronal response properties. In the adult, their terminal arbors typically cluster at regularly spaced intervals to reciprocally interconnect modules of neurons with similar response selectivity for orientation. During development, these connections also pass a first experience-independent phase in which they increase in length and gradually gain a crudely clustered termination pattern from an initially exuberant distribution (for references, see Schmidt et al. 1999). These initial events in the development of modular cortical structures and the initial emergence of neuronal response properties can only be blocked by silencing cortical activity and are therefore mainly dependent on molecular cues and patterns of spontaneous activity within the cortex.

However, in the second sensitive phase of development, already established response properties can even deteriorate markedly (for reviews, see Sherman and Spear 1982, Frégnac and Imbert 1984, Henry et al. 1994, Schmidt et al. 1999) if unimpaired binocular pattern vision is not assured. Likewise, long-range

horizontal intra-areal connections do not elongate further and refine into their specific adult pattern if animals are deprived of vision until they are adult.

This is strong support for the hypothesis that during the sensitive period, visual experience guides the continuing reorganization process and that developmental plasticity is needed to integrate newly formed connections into an already existing but still growing network according to functional criteria. A couple of paradigms illustrating the strong influence of pathological visual experience on cortical structures shall be mentioned.

If one eye is deprived from contour vision during the sensitive period, most of the neurons which would normally respond equally to stimulation of both eyes will be driven almost exclusively by the non-deprived eye (Wiesel and Hubel 1963; for a review, see Sherman and Spear 1982). The functional shift starts shortly after deprivation and subsequently is followed by a removal of the synaptic terminals and terminal arbors of the deprived eye (for a review, see Schmidt et al. 1999). Ocular dominance columns formed by eye-specific afferents will shrink and thalamic relay cells will decrease in size when driven by the deprived eye (Sherman and Spear 1982). These alterations are fully reversible during the sensitive period but no longer in adulthood. Since they allow only very limited visual function of the affected eye, the importance of early interference with inborn eye diseases like ptosis, a weakness of the upper lid occluding the pupil, or cataract (opacity of the lens), becomes obvious.

Another important paradigm demonstrating the influence of visual input on both thalamocortical and intra-areal connections is strabismus. Since the optical axes of the two eyes are not aligned, activity delivered from normally corresponding retinal loci to the cortex is uncorrelated. As a consequence, all cortical neurons become almost exclusively responsive to stimulation of only one eye, and thalamocortical afferents of the two eyes are completely non-overlapping (for a review, see Stryker 1991). Furthermore, long-range intra-areal connections, which normally are not eye-specific, come to connect only cortical neurons driven by the same eye (for a review, see Singer 1995). It is likely that binocular fusion and therefore stereo vision in strabismics are disturbed because parts of the visual system, thalamocortical, and intra-areal circuits are completely separated by eye.

An example, particularly emphasizing the instructive role of vision, is selective rearing in an environment displaying contours of only one orientation which will enhance the number of neurons in the adult responding to that particular stimulus orientation (Henry et al. 1994).

2.2 Mechanisms of Developmental Plasticity

Use- or experience-dependent changes of connectivity during development can be interpreted to follow Hebbian (1949) modification rules (Rauschecker and Singer 1981, Stryker 1991, Singer 1995) (see also *Long-term Potentiation and Depression* (*Cortex*)). According to Hebb's idea, which was modified by Stent (1973) for activity-dependent synaptic plasticity, the synaptic coupling of an afferent fiber and a postsynaptic target neuron is strengthened when they are activated at the same time. Synaptic connections to cells which fire uncorrelated or are inactive are destabilized. The deprivation and strabismus-induced alterations in ocular dominance development can be therefore seen as the result of an activity-dependent competition between the eye-specific afferents for postsynaptic neurons in the cortex. Cortical target neurons seem to strengthen their connection to only the sort of afferents providing the input which matches other local activity patterns. Postsynaptic activation is presumably required for changes to occur (Rauschecker and Singer 1981). Metabolically, changes involve glutamatergic synaptic transmission via NMDA receptors, are gated by neuromodulators such as norepinephrine, acetylcholine and serotonin, and are also dependent on local concentrations of neurotrophic substances (Berardi et al. 2000).

3. Neural Plasticity of Visual Cortex in the Adult

There is also evidence for plasticity of cortical functional properties and connections in the adult brain. It is well established that representations in the primary visual cortex can be reorganized reversibly as well as permanently following visual conditioning, restricted deafferentiation, or electrical stimulation.

3.1 Retinal Lesions

Visual cortical map reorganization was first demonstrated after retinal lesions in both monkey and cat primary visual cortex (see also *Neural Plasticity*). When cortical neurons are deprived of all their normal input by binocular lesions affecting corresponding retinal loci, they acquire new or modified receptive fields driven by inputs from intact retinal regions surrounding the lesions (Kaas 1991, Buonomano and Merzenich 1998, Gilbert 1998).

In the primary visual cortex of cats, substantial reorganization on a small spatial scale occurs within hours, similar to observations in the somatosensory system (Buonomano and Merzenich 1998). After recovery periods of more than 2–3 months, the majority of the neurons which lost their visual input start to respond to parts of the visual field to which they did not respond before. Finally, orientation and direction selectivity can recover to normal levels but responsiveness might remain reduced and receptive field borders less defined (for a review, see Gilbert 1998).

Obviously, cortical rearrangements caused by retinal lesions could serve to mask the actual loss of visual input and to create a continuous visual perception. This is observed mainly with lesions in the peripheral visual field, e.g., with postinfectional retinal scars or after preventive retinal laser coagulation in certain eye diseases (diabetes, myopia).

3.2 Visual Conditioning

As also observed in monkey and human somatosensory and motor cortex, representational plasticity in the visual system can occur without lesioning the sensory periphery (retina). After several months, neurons of adult monkeys fitted with prisms that reverse the visual field developed new receptive fields in the ipsilateral hemifield that normally only activates neurons in the contralateral cortex. Obviously, lasting changes in the visual input—not necessarily lesions—are sufficient to induce rearrangements in cortical representations. Context-dependent dynamic alterations of neuronal responsiveness and receptive field size occur also on a very short time scale. When stimulated with an artificial scotoma, a gray occluder covering receptive fields of the recorded neurons surrounded by a texture or dynamic noise, neurons with receptive fields inside the scotoma start to respond as if there were no occluder at all after a few seconds or even increase their receptive field such that they can be excited by surrounding stimuli after minutes (Gilbert 1998). Such short-term changes certainly operate within a transitional zone between mere adaptation and short-term plasticity, since they are fully reversible within seconds. They are thought to correspond to the psychophysical phenomenon of filling-in which is the basis for perceiving continuous images despite physiological disruptions like, for example, the blind spot (Ramachandran 1993).

3.3 Neuronal Substrate of Adult Cortical Plasticity

For several reasons, the observed long- and short-term modifications in the adult visual cortex are thought to be mainly mediated by long-range intra-areal connections. First, their large horizontal extent and characteristic functional selectivity makes them ideally suited to transmit feature-specific information over a large region of visual space and thus act within visual context. Second, after retinal lesions, reorganized receptive fields match the former response properties to oriented contours of that cortical region. However, thalamocortical connections do not sprout into the silenced cortical area. Therefore, the orientation-specific information can only be conveyed by intracortical orientation-specific connections. Third, collaterals of intracortical connections have indeed been demonstrated to branch and give rise to new synaptic boutons within the cortical region representing the scotoma (Gilbert 1998).

3.4 Mechanisms of Adult Plasticity

Short-term modifications such as enhanced neuronal responsiveness and receptive field expansion are thought to be mainly achieved by rapid modulations of the efficacy of preexisting connections enabling subthreshold excitatory inputs to effectively drive cortical neurons by either suppressing inhibitory or potentiating excitatory connections (Gilbert 1998).

After retinal lesions, a cascade of metabolic changes seem to precede long-term reorganization (for a review, see Eysel et al. 1999). In the weeks following the immediate expansion of receptive fields, decreased GABA-mediated inhibition within and a shift of hyperexcitability from its border towards the silenced site are observed until normal activity relations are restored. These changes are accompanied by increased spontaneous and visually driven activity and enhanced ability to induce long-term potentiation (LTP) (see also *Long-term Potentiation and Depression* (*Cortex*); *Neural Plasticity*). Mechanisms using LTP could lead to long-lasting changes in the efficacy of existing synapses of long-range horizontal connections and the additional secretion of neurotrophic substances might finalize the local cortical reorganization through formation of new synapses and sprouting of axon collaterals.

See also: Neural Plasticity; Occipital Lobe (Visual Cortex): Functional Aspects; Sensitive Periods in Development, Neural Basis of; Visual Development: Infant; Visual Perception, Neural Basis of

Bibliography

Berardi N, Pizzorusso T, Maffei L 2000 Critical periods during sensory development. *Current Opinion in Neurobiology* **10**: 138–45

Buonomano D V, Merzenich M M 1998 Cortical plasticity: from synapses to maps. *Annual Reviews in Neuroscience* **21**: 149–86

Duke-Elder S, Wybar K 1973 *System of Ophthalmology. Ocular Motility and Strabismus*. Kimpton, London, pp. 290–17

Eysel U T, Schweigart G, Mittmann T, Eyding D, Qu Y, Vandesande F, Orban G, Arckens L 1999 Reorganization in the visual cortex after retinal and cortical damage. *Restorative Neurology and Neuroscience* **15**: 153–64

Frégnac Y, Imbert M 1984 Development of neuronal selectivity in primary visual cortex of cat. *Physiology Reviews* **64**: 325–434

Gilbert C D 1998 Adult cortical dynamics. *Physiology Reviews* **78**: 467–85

Hebb D O 1949 *The Organization of Behavior. A Neuropsychological Theory*. Wiley, New York

Henry G H, Michalski A, Wimborne B M, McCart R J 1994 The nature and origin of orientation specificity in neurons of the visual pathways. *Progress in Neurobiology* **43**: 381–437

Kaas J H 1991 Plasticity of sensory and motor maps in adult mammals. *Annual Review of Neuroscience* **14**: 137–67

Ramachandran V S 1993 Behavioral and magnetoencephalographic correlates of plasticity in the adult human brain. *Proceeding of the National Academy of Sciences USA* **90**: 10413–20

Rauschecker J P, Singer W 1981 The effects of early visual experience on the cat's visual cortex and their possible explanation by Hebb synapses. *Journal of Physiology* **310**: 215–39

Schmidt K E, Galuske R A W, Singer W 1999 Matching the modules: Cortical maps and long-range intrinsic connections in visual cortex during development. *Journal of Neurobiology* **41**: 10–17

Sherman S M, Spear P D 1982 Organization of visual pathways in normal and visually deprived cats. *Physiology Reviews* **62**: 738–855

Singer W 1995 Development and plasticity of cortical processing architectures. *Science* **270**: 758–64

Stent G S 1973 A physiological mechanism for Hebb's postulate of learning. *Proceeding of the National Academy of Sciences USA* **70**: 997–1001

Stryker M P 1991 Activity-dependent reorganization of afferents in the developing mammalian visual system. In: Lam D M-K, Shatz C J (eds.) *Development of the Visual System*. MIT Press, Cambridge, MA, pp. 267–87

von Noorden G K 1990 *Binocular Vision and Ocular Motility. Theory and Management of Strabismus*. Mosby, St Louis, MO

von Senden M 1932 *Raum- und Gestaltauffassung bei operierten Blindgeborenen vor und nach der Operation*. Barth, Leipzig, Germany

Wiesel T N, Hubel D H 1963 Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of Neurophysiology* **26**: 1003–17

K. E. Schmidt

Neural Plasticity of Spinal Reflexes

The mammalian spinal cord has long been viewed as a passive conductor of information from the sensory receptors of the body to the processing areas of the brain, and of motor commands from the brain to the reflex centers of the cord, thence to the effector organs. Spinal reflexes traditionally have been viewed as hard wired, nonchanging, behavioral patterns evoked by external stimuli or descending influences from the brainstem or cortex. New data show that spinal reflex circuits can undergo alterations that may underlie a variety of behavioral and sensory changes. Insights into the actual neural mechanisms of these changes are beginning to be understood. These insights are having major impacts on pain control and spinal injury rehabilitation.

1. *Organization of the Spinal Cord*

The spinal cord receives from the peripheral nervous system most of the sensory information from both the musculoskeletal system and the visceral organs. Likewise, most motor control passes from the brain through the spinal cord and out to the striate muscles of the musculoskeletal system, and smooth muscles and other effectors of the visceral systems.

Two basic components make up the spinal cord. One is the 'long-lines' that transmit information from the input levels to the brain, and other spinal levels. These fibers generally are myelinated axons travelling from the cell bodies of origin in the spinal cord to brainstem and cortical levels, and convey sensory information. Likewise, motor and effector organ information is sent from the cerebral cortex and subcortical areas to the spinal cord along myelinated fiber tracts. These myelinated ascending and descending tracts form the white (colored by the white myelin coatings around the axons) outer aspect of the spinal cord.

The second, more complex area of the cord is the central portion, an 'H' shaped inner core, the central gray. Most incoming sensory information travels via the sensory nerves into the top or dorsal aspect of the central gray, where the nerves synapse on interneurons. The interneurons lack the myelin coatings of the long-lines, hence the gray color. The interneurons of the central gray are specialized for collecting and integrating information, and passing it to other areas of the nervous system. It is the interneurons that provide most of the processing power for reflex interactions and for actual alterations of spinal function now being shown.

2. *Research into Spinal Neural Plasticity*

Two main areas of research have evolved to investigate possible spinal neural plasticity. The first is the area of behavioral indicators of spinal cord plasticity. Various behavioral tasks and measures have shown alterations in reflex functions that imply underlying neural alterations. In some cases, these studies have led to investigations of underlying neural mechanisms. The second area involves the use of nociceptive (pain) inputs to alter spinal function in studies aimed at investigating underlying neural plasticity. These lines of investigation have produced a body of knowledge that is now being applied to human spinal cord injury rehabilitation and to pain control with encouraging results.

3. *Behavioral Indicators of Spinal Neural Plasticity*

For many years the role of spinal circuitry in reflex behaviors has been recognized. However, spinal reflexes, such as arm withdrawal when a finger is burned, were thought to be basically hard-wired, unchanging circuits whose activity might be increased or decreased by descending influences from the brain but that did not actively participate in brain mediated changes like learning. One of the early questions flowing from the

search for neural mechanisms of learning and memory was that of what was the smallest unit of the nervous system that could support learning. Since the spinal reflex arcs were easily accessible, simple units of neural circuitry, they quickly attracted the attention of early investigators looking for neural mechanisms of behavioral change.

3.1 Sensitization and Habituation

Researchers have long recognized that spinal reflexes could show some activity alterations. Sherrington (1906) described spinal fatigue or decreases in spinal reflex activity secondary to repeated activation in spinalized animals. The decrease in spinal function was transient, however, and recovered soon after stimulus cessation. The converse process was also recognized; with more intense stimulus inputs, the reflex response could increase, rapidly returning to baseline after the stimulus was discontinued. Groves and Thompson (1970) summarized much of this work and proposed a neural model. With repetitive inputs to the spinal circuits, relatively low intensity stimuli would activate 'H-type' interneurons that decreased their activity, resulting in decreased output. Stronger stimuli would also activate an additional circuit whose interneuron activity would increase, resulting in increased output. The resultant behavior would be an algebraic summation of the two output streams and could be decreasing behavior (habituation), no change, or increasing behavior (sensitization).

A variant of sensitization studied for many years is termed spinal fixation (Patterson 1976, 2001). A strong stimulus given to a spinalized animal for 15–45 minutes will produce an alteration of spinal postures that are retained for weeks. This nonassociative behavioral alteration has been shown to be stimulus intensity dependent and to occur as a neural excitability alteration in the interneuron circuits of the cord. It appears to be a longer-term variant of sensitization that occurs with high level nociceptive stimuli. Its behavioral consequences remain obscure, but it may be a process underlying some forms of chronic pain syndromes.

3.2 Classical Conditioning

Classical conditioning is recognized as the simplest form of associative learning. An association between a signaling stimulus (conditioned stimulus or CS) and a response producing stimulus (unconditioned stimulus or UCS) forms when the CS is presented shortly before UCS onset. The CS gradually comes to elicit a response (CR) similar to that evoked initially by the UCS. A considerable body of research beginning in the 1930s (see Patterson 1976) attempted to demonstrate that spinal reflex circuits show the associational learning of classical conditioning. While beset with theoretical and methodological difficulties, the evidence supported the ability of spinal circuits to support long-lasting (days) changes due to temporal association. Other data (e.g., Beggs et al. 1985) indicate that classical conditioning procedures produce a variety of long-term neural alterations closely approximating associative learning in the intact animal. There is some suggestion that the ability of the spinal cord to sustain this neural plasticity decreases for several days after spinal transection, but may return within a few weeks, presumably after neural reorganization.

3.3 Instrumental Conditioning

When the behavior to be learned is not simply evoked by a UCS, but is the result of some activity of the organism that is rewarded or punished by a stimulus, the learning is attributed to instrumental learning (see Grau et al. 1998). A typical spinal instrumental conditioning situation involves the omission of a nociceptive stimulus if the spinalized animal's leg does not fall below a preset level. The leg gradually comes to rest at a higher level (increased muscle tension) with repeated trials. A particularly interesting variant of spinal instrumental conditioning has been shown by Wolpaw (Carp and Wolpaw 1994). Intact monkeys were conditioned instrumentally to increase or decrease the excitability of a simple spinal circuit. The conditioning procedure involved no nociceptive (pain) stimuli, unlike most spinal conditioning procedures, and took weeks to complete, as opposed to hours for most spinal conditioning. After conditioning was complete, spinal transection, isolating the spinal circuits, did not abolish the spinal excitability changes. Thus, the alterations produced by training procedures involving stimuli processed at cortical levels, but primarily activating spinal circuits, resulted in neural alterations at the spinal level.

The ability of spinal circuits to respond to positive (reward) or negative (punishment) contingencies by increases or decreases in behavioral responses indicates that the range of neural plasticity in spinal cord circuits is not limited to simple stimulus driven pattern changes. Thus, the current data strongly suggest that spinal circuits can undergo both short- and long-term alterations that affect behavior.

4. Neural Mechanisms of Spinal Cord Plasticity

The search for mechanisms of spinal neural plasticity has been heavily driven by the need to understand certain types of chronic pain syndromes. With the realization that spinal circuits can be active participants in information processing, and can both amplify and decrease information transmission, it has become apparent that spinal circuits may also affect nociceptive or pain signal transmission from the body to the brain, where pain perception occurs. Many pain states are now recognized as having a component of a 'central pain state,' essentially a component that is

generated in the spinal circuitry itself (see, e.g., Dubner and Gold 1999). Since pain is perceived at the conscious level, neural inputs to the brain coming from an injured body structure are perceived as pain in that structure. If these brain inputs were generated not in the peripheral structure, but in the spinal centers between the body and the brain, the perceived pain would still be felt as occurring in the body. Thus, a central pain state can occur with little or no actual physical damage or ongoing input from the body.

Many investigations of spinal plasticity secondary to nociceptive inputs have been in spinal or intact animals with an irritant injected into a leg joint to produce inflammation that lasts for hours or days (e.g., Woolf and Costigan 1999). The long-lasting nociceptive inputs terminate primarily on the interneurons in the upper layers of the spinal gray matter. The impulses are transmitted to lower order interneurons and thence to motor outputs as well as to the long-line axons for transmission to the brainstem and cortex.

The effects of even fairly short bursts of high intensity nociceptive inputs to the spinal interneurons can be dramatic. Increases in interneuron excitability similar to that shown with the behavioral studies of sensitization occur quickly and appear to be alterations in the presynaptic terminals of the input neurons coming from nociceptors at the injury site. The terminals release more neural transmitter with each activation of the afferent neuron, thus increasing the overall activation of the spinal interneurons. With continued inputs, alterations occur that increase the sensitivity of the interneuron membranes to transmitters. In addition, other substances such as nitric oxide (NO) and substance *p* (a neurotransmitter) are released into the general area of the activated interneurons and act to increase the excitability of other interneurons. Thus, not only the directly involved circuits, but neighboring circuits can become over activated. Also, certain genes of the interneurons will be activated, producing increases in numbers of membrane NMDA receptors that further increase the sensitivity of the interneurons to neurotransmitters (see, e.g., Dubner and Ruda 1992). Further nociceptive input begins a biochemical reaction that results in the death of inhibitory interneurons that normally dampen the activity of the spinal networks. Additionally, over days, there appears to be sprouting of processes from the excitatory interneurons to other interneurons forming new excitatory synapses. The overall result of this neural plasticity is to temporarily, then permanently, increase the overall excitability of the spinal circuits responsible for processing nociceptive input.

5. *Implications for Spinal Function*

Evidence from behavioral and learning studies of spinal processes shows that spinal reflex circuits have an inherent capacity for alterations subsequent to stimulus inputs. The understanding of underlying neural mechanisms from these studies seem to implicate the same types of processes that have been found to underlie alterations seen with pain inputs. From these data, some generalizations can be made. Spinal reflex circuits have an inherent capacity for change due to both simple activation and in response to certain stimulus patterns, such as temporal proximity. Nociceptive inputs are especially potent in producing excitability changes in neural function, but nonnociceptive inputs, over a longer time, can also produce alterations. The alterations include short-term changes lasting from seconds to minutes, longer lasting alterations measured in hours or days, intermediate term changes lasting days and weeks, and finally, semipermanent to permanent changes.

The implications for understanding pain generation and control as well as spinal cord injury rehabilitation are dramatic (e.g., Wickelgren 1998). Central pain is one of the most intractable medical problems known. Without a peripheral site of pain generation, common methods of pain reduction are almost useless, and medications affecting central pain perception (e.g., morphine) often leave the patient severely incapacitated. Understanding the mechanisms underlying spinal circuit excitability alterations will allow the central changes to be stopped or aborted. Learning how to alter excitability of reflex patterns by behavioral interventions will allow increasingly effective rehabilitation for humans with spinal cord injuries to a degree not even imagined a few years ago. Coupled with a greater understanding of how to encourage spinal neurons to regenerate through areas of injury and even how to transplant new neural tissue into injured areas, these techniques are changing the way spinal injury and chronic central pains states are treated.

See also: Autonomic Classical and Operant Conditioning; Classical Conditioning, Neural Basis of; Habituation

Bibliography

Beggs A L, Steinmetz J E, Patterson M M 1985 Classical conditioning of a flexor nerve response in spinal cats: Effects of tibial nerve CS and a differential conditioning paradigm. *Behavioral Neuorscience* **99**(3): 496–508

Carp J S, Wolpaw J R 1994 Motoneuron plasticity underlying operantly conditioned decrease in primate H-reflex. *Journal of Neurophysiology* **72**: 431–42

Dubner R, Gold M 1999 The neurobiology of pain. *Proceedings of the National Academy of Science* **96**(July): 7627–30

Dubner R, Ruda M A 1992 Activity-dependent neuronal plasticity following tissue injury and inflammation. *Trends in Neuorscience* **15**(3): 96–103

Grau J W, Barstow D G, Joynes R L 1998 Instrumental learning within the spinal cord. I. Behavioral properties. *Behavioral Neuroscience* **112**: 1366–86

Groves P M, Thompson R F 1970 Habituation: A dual-process theory. *Psychological Review* **77**: 419–50

Patterson M M 1976 Mechanisms of classical conditioning and fixation in spinal mammals. In: Riesen A H, Thompson R F (eds.) *Advances in Psychobiology*. Wiley, New York, Vol. 3
Patterson M M 2001 Spinal fixation. In: Patterson M M, Grau J W, Wolpaw J R, Willis W D, Edgerton R E (eds.) *Spinal Cord Plasticity: Alterations in Reflex Function*. Kluwer, Norwell, MA
Sherrington C S 1906 *The Integrative Action of the Nervous System*. C. Scribner's Sons, New York
Wickelgren I 1998 Teaching the spinal cord to walk. *Science* **279**(16 Jan): 319–21
Woolf C L, Costigan M 1999 Transcriptional and posttranscriptional plasticity and the generation of inflammatory pain. *Proceedings of the National Academy of Science* **96**(July): 7723–30

M. M. Patterson

Neural Representations of Direction (Head Direction Cells)

1. The Head Direction Signal

To navigate successfully animals require knowledge about their location and directional heading before embarking on a route. While location information may be provided by place cells in the hippocampus which discharge with respect to the animal's location (see *Navigation in Spatial Environments, Neural Basis of*), another group of limbic system neurons discharge in relation to the animal's directional heading in the azimuthal plane, independent of the animal's behavior and location (Taube et al. 1990a, Muller et al. 1996, Taube 1998). For example, a head direction (HD) cell might discharge whenever the animal points its head northeast, no matter where it is in the environment. The direction at which the cell fires maximally is referred to as the preferred firing direction. The position of the animal's trunk does not affect the cell's firing, nor does pitch or roll of the animal's head up to 90 degrees. HD cell firing is motion-independent—cells will discharge whether the rat is moving or remains still, as long as the animal is pointing its head in the proper direction. Very little, if any, adaptation of firing rate is noted when the head remains fixed at the preferred firing direction.

HD cells are tuned to different directional headings, and within a particular brain area the population of these cells encodes all 360 degree directions. Each HD cell can be described by its tuning curve, constructed by plotting firing rate as a function of directional heading. Figure 1 shows a typical tuning curve for an HD cell that fired maximally at 258 degrees. For each HD cell several properties can be characterized from its tuning curve: (a) the cell's preferred firing direction; (b) the maximal (peak) firing rate of the cell at the preferred direction; and (c) the angular range over which cell firing is elevated from background firing rate. Some HD cells have low peak firing rates (5–10 spikes sec^{-1}) while other HD cells have high peak firing rates (100 spikes sec^{-1}). Intermediate peak firing rates are also observed. The functional significance of different cells having different peak firing rates is unknown. The firing rate of each HD cell decreases linearly as one moves away from the preferred firing direction. The range of directional headings where firing is elevated above background firing levels is about 90 degrees for most cells. The firing characteristics of an HD cell usually remain stable indefinitely, as long as the environmental parameters remain unchanged.

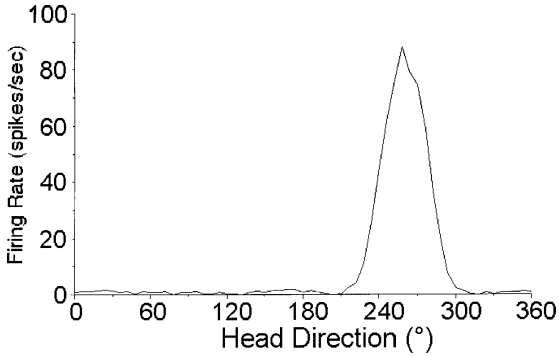

Figure 1
Firing rate versus head direction plot for a typical HD cell

2. Where are HD Cells Located in the Brain?

HD cells were originally identified in the dorsal presubiculum of rats. They have subsequently been identified in several limbic system areas in the rat, including the anterior dorsal thalamic nucleus, lateral dorsal thalamic nucleus, lateral mammillary nuclei, and posterior cingulate cortex (Taube 1998). Note that many of these areas are considered part of the classical Papez circuit. HD cells have also been identified in one nonlimbic structure—the striatum. In addition, HD cells have been reported in the presubiculum of nonhuman primates. With few exceptions, HD cells in each of the different brain areas have similar tuning curves, and behave similarly following manipulations of the animal's environment. However, HD cells in the anterior dorsal thalamus and lateral mammillary nuclei appear to discharge optimally to the animal's future directional heading, while HD cells in the dorsal presubiculum discharge maximally to the animal's current directional heading (Blair and Sharp 1995). Similarly, in contrast to HD cells in the dorsal presubiculum, HD cell firing in the anterior dorsal thalamus and lateral mammillary nuclei is secondarily correlated to the rat's angular head velocity, where

firing rates within the cell's preferred direction are higher for faster head turns than slower head turns (Taube 1995). It is also noteworthy that there are distinct populations of cells in the lateral mammillary nuclei that, independent of directional heading, encode for either pitch of the animal's head or angular head velocity (Stackman and Taube 1998).

3. HD Cell Responses in 3-D

How HD cells respond in three dimensions is understood less well. HD cells will continue to discharge in the vertical plane, if the vertical locomotion begins with the rat's orientation corresponding to the preferred firing direction (Stackman et al. 2000). But if the vertical plane locomotion begins with the animal's head orientation outside the cell's preferred direction, then the cell will not discharge on the vertical wall until the animal turns its head to the proper orientation. Thus, HD cells are capable of maintaining their discharge in planes outside Earth horizontal and their firing is dependent on the directional orientation of the rat as it moves out of the horizontal plane. One model consistent with these findings is that HD cells define the horizontal reference frame as the animal's plane of locomotion, but this leaves open the question of which cue source(s) determine the horizontal reference frame: vestibular, somatosensory, visual, motor cues, or some combination of these sources. It is not known how HD cells respond when an animal locomotes upside down on a ceiling. In terms of height, if an animal locomotes back and forth between two horizontal surfaces that are separated by about one meter in height, the preferred firing direction is the same on both surfaces, but the peak firing rate increases a little on the higher surface.

4. Both Landmark Cues and Internally-derived Self-motion Cues Affect HD Cell Firing

While HD cells bear a similarity to a navigational compass, they are not dependent on the Earth's magnetic field. They respond to several different types of sensory stimuli, some of which are external to the animal and others of which are self-generated. Visual landmarks that are stable over time play the predominant role in controlling the HD cell's preferred firing direction. For example, if an animal is first disoriented and then returned to a familiar environment where all the visual landmarks have been rotated, the HD cell's preferred firing direction will also be shifted by a similar amount (Taube 1995). However, HD cells continue to respond normally when the animal moves about in darkness or when the prominent visual cues have been removed, indicating that self-motion cues are capable of maintaining cell firing as well as the cell's preferred firing direction (Goodridge et al. 1998). HD cells' discharge is *not* environment- or context-specific, and they will discharge at some direction in any environment. If an animal locomotes from a familiar to a novel environment, the preferred firing direction will be similar in both environments (Taube and Burton 1995). These findings indicate that HD cells receive information concerning the animal's movements, either through the vestibular or proprioceptive systems, or through motor corollary discharge. This movement information is integrated over time so that the animal can keep track of its directional heading given an initial reference point. When an animal is passively transported from a familiar to a novel environment, the preferred firing direction of a cell will often shift substantially, indicating the important role that motor systems play in maintaining a stable preferred firing direction (Taube 1998). Under this condition, the animal is deprived of normal motor, proprioceptive, and kinesthetic cues which accompany self-locomotion, but vestibular cues are left intact. Thus, while the vestibular system may play a crucial role in the generation of the HD cell signal (see below), these findings indicate that vestibular cues alone are *not* sufficient to allow maintenance of a stable preferred direction and suggest a critical role for motor information.

Most studies have shown that when spatial information derived from landmarks conflicts with information derived from internally generated self-movement, the landmark information usually predominates and overrides the spatial signals developed through self-generated movements (Goodridge and Taube 1995). Thus, when a salient, familiar cue is reintroduced into an animal's environment into a position that conflicts with the cell's current preferred firing direction, HD cells will shift their preferred direction to reflect their originally established orientation with the cue.

When two or more HD cells are monitored simultaneously in the same animal, the effects of an environmental manipulation on the preferred direction for one cell are similar to the responses observed in the other cell(s) (Taube et al. 1990b). This finding indicates that afferent input driving one HD cell will similarly influence other HD cells, and indicates that HD cells within a particular brain area behave as a network and that their preferred directions always remain a fixed angle apart (in-register) from one other.

5. Generation of the Directional Signal

Given that HD cells have been identified in several brain areas, many of which are interconnected with one another, the question arises as to which areas are critical for generating the directional signal. By recording from one brain area containing HD cells, and lesioning various afferent areas, a number of studies have concluded that (a) the anterior dorsal thalamus, but not the hippocampus or lateral dorsal thalamus, is critical for the generation of directional activity in the

dorsal presubiculum; (b) lesions of the lateral mammillary nuclei disrupt the HD cell signal in the anterior dorsal thalamus; and (c) lesions of the vestibular labyrinth abolish all direction-specific firing in areas containing HD cells (Stackman and Taube 1997, Taube 1998). It is also noteworthy that lesions of the anterior dorsal thalamus or the dorsal presubiculum do *not* abolish location-specific firing in hippocampal place cells, indicating that spatial representations concerning location can be generated independently of HD cell information.

The vestibular lesion data suggest that the neural code for directional bearing is critically dependent upon vestibular information, and that the loss of HD cell information may account for the orientation and navigational deficits observed following vestibular dysfunction in humans. Researchers have suggested that vestibular information is conveyed to the brain areas containing HD cells via a subcortical pathway starting in the medial vestibular nuclei and then projecting to the dorsal tegmental nucleus directly (or indirectly via the nucleus prepositus), and then to the lateral mammillary nuclei, which in turn project to the anterior dorsal thalamus. In this scheme, the anterior dorsal thalamus is viewed as a convergence site where subcortical information representing self-movement is combined with spatial information concerning landmarks that originate from projections in the posterior cingulate cortex. Attempts to model HD cell firing have used attractor-type neural networks (Skaggs et al. 1995). These models have been able to replicate most, but not all, the empirical data.

In animals with lesions of the hippocampus each HD cell's preferred firing direction remained stable across days when the animal was placed into a novel environment (Golob and Taube 1997). This result suggests that extrahippocampal structures are capable of creating and maintaining a novel representation of the animal's spatial context. This representation shares features in common with mnemonic processes involving episodic memory, which are thought to require an intact hippocampus.

6. *Relationship of HD Cell Activity to Behavior*

One important issue concerning HD cells is what role they play in an animal's behavior. Are they actively involved in guiding navigational behavior? Evidence suggesting a link between HD cells and behavior has been produced using a radial arm maze in a spatial reference memory task (Dudchenko and Taube 1997). Animals were trained to run down a particular maze arm relative to a visible landmark cue to obtain a water reward. Following rotation of the cue, the animal's maze arm selection and the HD cell's preferred direction shifted in tandem relative to the salient visual cue. Furthermore, when the HD cell's preferred direction did not shift relative to the cue, neither did the animal's behavioral response. When an HD cell was recorded over several days as the animal learned the task, there was little change in the firing properties of the cell. These findings indicate that the coding of directional heading information is independent of the neural systems involved in task acquisition, and that over time, an animal learns to use information provided by the directional system to guide its behavioral responses.

Recent experiments that have tested this hypothesis more rigorously have surprisingly found that HD cell activity and behavioral choice are *not* linked together on two different types of spatial tasks (Golob et al. 2001). One task was a spatial reference memory task where animals were required to generalize appropriately between a square and a rectangular enclosure to receive a reward. The other task involved a spatial working memory task. In both cases HD cell activity in the anterior dorsal thalamus did *not* correlate very well with the animal's behavioral responses. While it appears that information derived from HD cells is not used in these two types of spatial tasks, it is likely that future studies will find that HD cell information will be more utilized in spatial tasks that require path integration or navigation over larger distances. Clearly, additional experiments that explore this issue are warranted.

7. *Summary*

In summary, HD cells represent a high-level spatial signal that encodes the animal's directional heading. HD cells have been identified in several brain areas—mostly within the limbic system, but the functional role served by each brain area remains to be elucidated. HD cells receive multimodal sensory and motor information. It appears that the signal arises from the lateral mammillary nuclei, or structures afferent to it, and that both vestibular and motor information play important roles in generating the signal. Visual landmarks can exert control over the preferred firing direction of an HD cell. The nature of the HD cell signal makes it an ideal candidate for studying the neural mechanisms underlying navigation.

See also: Motion Perception Models; Motion Perception: Psychological and Neural Aspects; Navigation in Spatial Environments, Neural Basis of; Spatial Cognition

Bibliography

Blair H T, Sharp P E 1995 Anticipatory head direction signals in anterior thalamus: Evidence for a thalamocortical circuit that integrates angular head motion to compute head direction. *Journal of Neuroscience* **15**: 6260–70

Dudchenko P A, Taube J S 1997 Correlation between head direction cell activity and spatial behavior on a radial arm maze. *Behavioral Neuroscience* **111**: 3–19

Golob E J, Taube J S 1997 Head direction cells and episodic spatial information in rats without a hippocampus. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 7645–50

Golob E J, Stackman R W, Wong A C, Taube J S 2001 On the behavioral significance of head direction cells: Neural and behavioral dynamics during spatial memory tasks. *Behavioral Neuroscience* **115**: 285–304

Goodridge J P, Taube J S 1995 Preferential use of the landmark navigational system by head direction cells. *Behavioral Neuroscience* **109**: 49–61

Goodridge J P, Dudchenko P A, Worboys K A, Golob E J, Taube J S 1998 Cue control and head direction cells. *Behavioral Neuroscience* **112**: 749–61

Muller R U, Ranck J B Jr, Taube J S 1996 Head direction cells: Properties and functional significance. *Current Opinion in Neurology* **6**: 196–206

Skaggs W E, Knierim J J, Kudrimoti H S, McNaughton B L 1995 A model of the neural basis of the rat's sense of direction. In: Tesauro G, Touretzky D S, Leen T K (eds.) *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, Vol. 7, pp. 173–80

Stackman R W, Taube J S 1997 Firing properties of head direction cells in rat anterior thalamic neurons: Dependence upon vestibular input. *Journal of Neuroscience* **17**: 4349–58

Stackman R W, Taube J S 1998 Firing properties of rat lateral mammillary nuclei single units: Head direction, head pitch, and angular head velocity. *Journal of Neuroscience* **18**: 9020–37

Stackman R W, Tullman M L, Taube J S 2000 Maintenance of rat head direction cell firing during locomotion in the vertical plane. *Journal of Neurophysiology* **83**: 393–405

Taube J S 1995 Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *Journal of Neuroscience* **15**: 70–86

Taube J S 1998 Head direction cells and the neurophysiological basis for a sense of direction. *Progress in Neurobiology* **55**: 225–56

Taube J S, Burton H L 1995 Head direction cell activity monitored in a novel environment and during a cue conflict situation. *Journal of Neurophysiology* **74**: 1953–71

Taube J S, Muller R U, Ranck J B Jr 1990a Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience* **10**: 420–35

Taube J S, Muller R V, Ranck J B Jr 1990b Head direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. *Journal of Neuroscience* **10**: 436–47

J. S. Taube

Neural Representations of Intended Movement in Motor Cortex

The motor cortex consists of a number of functionally and connectionally interrelated areas that participate in the planning and production of voluntary movements. Functions like conditional association, coordinate transformation, and linking of movements occur in multiple parietal, cingulate, and frontal areas. There is a general organization that abstract aspects of planning and spatial organization occur in parietal cortex, while sequencing and execution of movement occurs in frontal areas. However, these functions are extensively intermingled across the set of motor areas, so that intended movements are products of continual interactions across a broad region of cerebral neocortex.

1. *What is the 'Motor Cortex'?*

The cerebral neocortex is frequently considered to be the primary site where voluntary movements are planned, initiated and, to some extent, learned. These ideas have been supported by studies using a variety of methods, including single and multiple neuron recording, lesion effects, and human functional brain imaging (Sanes and Donoghue 1997). This body of work has significantly advanced our understanding of the means by which plans for voluntary movement are assembled and used for muscle control. This article will define the term motor cortex and the cortical areas that comprise it, then describe how various motor areas participate in voluntary movement. Our present knowledge leads to the conclusion that a vast cortical territory cooperates in fashioning voluntary movements from external cues or mental intentions.

The motor cortex comprises a collection of more or less distinct areas or 'fields,' which are united by function and connectivity. A field is typically, but not exclusively, included as motor if neurons in that area modulate their discharge in close conjunction with the contraction of voluntary muscles, it receives inputs from other motor structures in the central nervous system (CNS) and it contributes output through the corticospinal tract. More detailed differences in the functional properties, connections, and cytoarchitecture distinguish one motor area from its neighbors. Ongoing anatomical and functional studies of frontal and parietal cortex add to a progressively increasing list of motor separate cortical fields; probably more than a dozen can be identified. The blending of sensory and motor properties across the neocortex means that many cortical fields could be classified within several different domains, pointing out the vagaries of forcing divisions in a region of the CNS that is inherently integrative. Among motor fields, oculomotor and somatomotor control areas are often separated. This review deals with those areas of cortex that control the somatic musculature, in particular those involved in arm reaching movements, which have been most thoroughly studied.

1.1 *Subdivisions of Motor Cortex*

Figure 1 summarizes the regions where neurons related to arm reaching are likely to be present in the human brain. Localization of motor areas in humans is inexact because their identification is based upon data

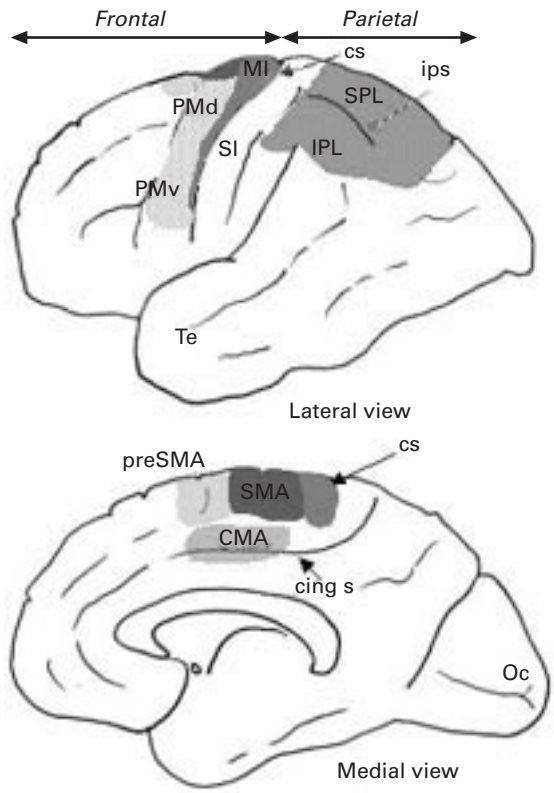

Figure 1
Human cortical motor areas. The general regions where neurons related to intended movements are shaded in a lateral (top) and medial (bottom) view of the human brain. The frontal and parietal lobes are separated by the central sulcus (cs): Primary motor cortex lies in the precentral gyrus anterior to the cs, while the somatic sensory cortex (SI) is situated in the postcentral gyrus posteriorly. The superior and inferior parietal lobes (SPL, IPL), separated by the intraparietal sulcus (ips), probably contain multiple motor areas based on data from non-human primates. Frontal motor areas include the premotor ventral (PMv) and dorsal (PMd), supplementary and presupplementary areas (SMA, preSMA). The cingulate cortex region also contains multiple motor areas (CMA), which are located along the cingulate sulcus (cing s) (Oc, occipital lobe; Te, temporal lobe)

obtained directly from single neuron recordings in non-human primates. Similar data is difficult to obtain in the normal human brain. Localization is made by correlating areas of blood flow changes from imaging techniques in humans with a presumed homolog in other species. Traditionally, the primary motor cortex (MI, Brodmann's area 4), which is located on the precentral gyrus of many primates (including humans), is considered a primary site for movement generation. It is ranked 'primary' largely because movements of the opposite side of the body can be elicited by the lowest levels of electrical stimulation anywhere in the cortex. Among cortical areas, MI also has the bulk of the direct projections to the spinal cord, including some direct projections to alpha motor neurons ($\sim$ 5 percent). In addition, MI neurons are active with voluntary movement, and lesions of MI lead to paralysis of the contralateral musculature. Lying in front of MI are a collection of frontal lobe non-primary areas, where stimulation thresholds are higher, neurons may have more complex properties, and lesions result in movement defects other than frank paralysis, such as incoordinations called apraxia. While there is some diversity in the naming of these other motor areas, a dorsal and ventral premotor (PM), a supplementary motor (SMA), and pre-supplementary motor area (pSMA) are commonly recognized as distinct areas; each of these may be subdivided further (Geyer et al. 2000). More recently it has become apparent that the cortex on the medial wall of the cerebral hemisphere along the cingulate sulcus is related to movement planning and preparation (CMA; Strick et al. 1998), which also appears to contain multiple subdivisions.

The parietal lobe can be subdivided into a number of areas associated with somatic sensation in its anterior part (SI or primary somatic sensory cortex), and for all senses more posteriorly. The posterior parietal cortex is divided into superior and inferior lobes which, on the basis of non-human studies, is likely to contain multiple somatic motor areas. The parietal lobe is heavily engaged in spatially guided actions. In the anterior parietal cortex, the postcentral gyrus contains a collection of four distinct somatic sensory areas—Brodman's areas 3b,1, 2, and 3a. The former two areas seem mainly concerned with processing of touch, while areas 2 and 3a process deep or kinesthetic information from the joints, muscle spindles (muscle length and velocity), and golgi tendon organs (muscle tension receptors). Neurons related to the production of arm movement are also in this region, particularly in area 2. More posterior areas of parietal cortex (both IPL and SPL) contain areas related to eye and arm movement, often blended with various types of visuo-spatial or other sensory modality related neurons. In monkeys, group areas related to reaching and grasping movement are located in the intraparietal sulcus; whether such areas are in the same location in humans is not resolved. Most posteriorly in the parieto-occipital area, the PO region of monkeys is an additional arm related area; also termed V6 for its visual responsiveness (Burnod et al. 1999).

1.2 Connections of Motor Areas

Motor areas are densely interconnected with each other and with other cortical areas; the thalamus and

the basal ganglia, brainstem and, for most areas, the spinal cord. All motor cortical areas are interconnected within at most two or three synaptic relays, but this is also true of most of the CNS, making connectivity difficult to be used as a sole differentiator of dedicated motor areas. Direct connections exist from parietal to frontal lobes with a general pattern that regions close to the central sulcus (dividing SI and MI, see Fig. 1) in the parietal cortex project to regions close to the central sulcus in the frontal cortex, creating a rough pattern of mirror symmetry of connections (see review of Geyer et al. 2000). It is essential to emphasize the vastly parallel nature of possible interactions among motor areas, rather than any simplistic view of a stepwise serial process from one to the next area. Beyond the clear picture of vast interconnectivity of areas, studies of the timing of neurons during movement planning and action demonstrate that neurons in most or all of the motor areas interact extensively (Wise et al. 1997).

2. *Cortical Areas Engaged in Movement Imaging Methods*

A broad perspective of the activation patterns of motor areas can be gained from brain imaging studies. Both positron emission (PET) and functional magnetic resonance imaging (fMRI) studies in humans show that primary motor cortex is engaged when voluntary actions are produced. By contrast, premotor and supplementary motor areas can be activated simply by planning motor actions, although they are also engaged during movement production (Sanes 2000). Recently, human cingulate cortex has been tied to movement based on activation in PET and fMRI studies. Cingulate motor areas (CMA, Fig. 1) can be activated by grasp and reaching actions; some functions may be segregated within rostral and caudal cingulate areas (Picard and Strick 1996). Other results suggest that cingulate cortex may play a role in performance monitoring and in selection of actions (Carter et al. 2000). Similarly, the parietal lobe is heavily engaged in voluntary arm movements when viewed by brain imaging methods. Numerous areas of the parietal lobe activate with arm movement (Connolly et al. 2000), suggestive of multiple subdivisions, as reported in monkeys (see later).

3. *Activity of Single Neurons in Cortical Motor Areas*

Single neuron recordings in nonhuman primates give the most direct evidence for the role of cortical areas in movement. In addition, they verify conclusions based upon the indirect PET and fMRI measures. Neuronal activity has been measured during movement planning by instructing movement first, then, after a delay, requesting its enactment (an instructed delay task) during movement sequences and after various aspects of kinematics and dynamics have been varied. Neuron activity has been correlated with direction, amplitude, various types of loads, sequence structure, and sensory-motor contingencies and attention. Rather than describe the patterns of neural activity in each area under each of these conditions, only general patterns of engagement of each area will be described. One of the most pronounced features of motor cortex neurons is the relationship of firing modulation to direction of a reaching movement.

Neurons in frontal and parietal areas are directionally tuned so that they fire most to one direction and less so to other directions, and typically are least modulated in the direction opposite to their preferred direction. Neurons in motor areas are also modulated by force, and kinematic features (amplitude, position, and velocity changes) (Kalaska et al. 1997, Fu et al. 1993, Ashe and Georgopoulos 1994). Each of these features is correlated with neuronal activity, but causal links to the actual computation of these aspects of movements are less certain. The current experimental data cannot fully dissociate whether or where conversion from an abstract notion of movement in space to muscle activity patterns occurs. However, the relatively weaker influence of movement force on neural activity in the parietal areas suggests that motor plans are more abstract here, compared to the frontal lobe.

Correlation of force and neural activity is common in MI, but many neurons are influenced by kinematics as well, so that there is no clear cut neuronal segregation of these features. The activity of neurons during motor preparatory delays has provided further evidence for a role in planning actions. Neurons with delay activity are found across motor cortex, but they are more prevalent in the nonprimary regions. However, other work also suggests that even primary motor cortex may be involved in cognitive aspects of a motor plan (Carpenter et al. 1999), so it is difficult to accept a simple notion that plans are formulated in nonprimary cortex and then passed to MI for command generation. Rather, the set of areas continually cooperate throughout the planning and production of movement. This does not negate the finding that certain functions predominate in specific subdivisions of the motor cortex.

4. *Parietal–Frontal Engagement in Movement*

It is clear from electrophysiological, lesion, and imaging data that the computations concerning the spatial coordinates of movement are processed in the parietal lobe, while movement execution is more weighted towards frontal areas; planning occurs across the parietal–frontal network (Kalaska et al. 1997). Knowledge of arm position in space is available from

tactile and prophoceptive information in SI, while the location of visual cues is delivered from occipital visual areas. There is strong evidence that the posterior parietal cortex combines signals related to arm position and planned arm movements, as well as for gaze direction and location of sensory cues. (Batista and Andersen 2001, Burnod et al. 1999). In parietal cortex multiple frames of reference are present, although how these are interconverted is not clear (Boussaoud and Bremmer 1999). Further, planning for visually guided actions begins very early in 'visual' pathways, such as PO, only two synapses from primary visual cortex.

Nonprimary motor areas within the frontal lobe appear to be engaged in planning processes, based on the observation that activity in these regions occurs during instructed delay periods. Frontal areas are important for linking stimuli to movements or to link various movements together in space or time. These more complex actions seem less coded in primary motor cortex, where neuronal activity is more tightly coupled to the actual performance of intended arm movements. However, neurons related to planning and production of motion can be found intermingled throughout motor cortex; it appears as if only the relative proportions change across motor areas. There is considerable evidence that supplementary motor cortex (SMA) is important in formulating sequences of movements. Neurons in SMA can be selective for particular movement sequences or transitions and these same neurons may not be active when the same movements are made in other sequences (Shima and Tanji 2000). The pre-SMA, lying just in front of SMA, appears to be critical for learning new sequences, but the pre-SMA's role seems to fade as sequences are learned (Hikosaka et al. 1996). The dorsal premotor cortex PMD is also engaged in forming conditional associations between sensory cues and movement. While the entire PM cortex seems to be engaged in planning in various coordinate systems, the ventricle premotor cortex (PMv) seems to be a region for sensorimotor processing for stimuli near the body (Graziano and Gross 1998).

The primary motor cortex is often considered as the final common pathway and home of the upper motor neuron for movement commands to be issued. The widespread origin of the corticospinal path from most of the motor cortex coupled with the broad activation of other motor areas during movement is sufficient to reject these ideas in favor of a more distributed role for motor cortex in movement. Nevertheless, the preponderance of neurons closely coupled to the details of muscular action and the ability to activate muscles easily with electrical stimulation from MI suggest a special role for MI in voluntary movement. As stated earlier, MI neurons modulate their activity with nearly every aspect of movement including force, velocity, acceleration, and distance. It remains a considerable challenge to identify whether MI represents movements in terms of muscle activity or more abstract variables (Todorov 2000). On the one hand, up to 5 percent of MI's output directly contacts alpha motor neurons so that it provides a very powerful direct route to the muscles. On the other hand, some MI neurons are active with very abstract aspects of motor plans, such as the serial order of stimuli, suggesting that it participates in the cognitive events leading to action.

5. *Conclusions*

The combination of cortical motor areas combine signals related to proprioceptive and tactile signals of arm position, vision, gaze angle, attention and motor plan, which conveys a patterned outflow of information that guides somatic musculature for action.

The parietal lobe deals with the spatial aspects of movement: keeping track of the configuration of the arm in space, its relationship to the direction of gaze and to visual cues. In addition, it participates in formulating upcoming actions, possibly in a form more abstract than found in the frontal lobe. The frontal lobe motor areas contribute to learning and assembling sequences of motor actions and are also involved in non-spatial motor planning where arbitrary cues (or perhaps thoughts) are linked to particular motor actions. However, parietal–frontal motor areas also combine many types of signals, such as gaze or attention, which may have overt, or sometimes more subtle or conditional influences upon the activity of motor cortex neurons. Finally, this complex of areas is nearly simultaneously engaged in many forms of motor action so that they act in concert when intention leads to action. The essence of understanding how an intention becomes an action depends on further understanding of the functional interactions among the collection of motor areas.

See also: Classical Mechanics and Motor Control; Eye Movement, Control of (Oculomotor Control); Motor Control; Motor Control Models: Learning and Performance; Motor Cortex; Motor Skills, Psychology of

Bibliography

Ashe J, Georgopoulos A P 1994 Movement parameters and neural activity on motor cortex and area 5. *Cerebral Cortex* **4**: 590–600

Batista A P, Andersen R A 2001 The parietal reach region codes the next planned movement in a sequential reach task. *Journal of Neurophysiology* **85**: 539–44

Boussaoud D, Bremmer F 1999 Gaze effects in the cerebral cortex: reference frames for space coding and action. *Experimental Brain Research* **128**: 170–80

Burnod Y, Baraduc P, Battaglia-Mayer A, Guigon E, Koechlin E, Ferraina S, Lacquaniti F, Caminiti R 1999 Parieto-frontal coding of reaching: an integrated framework. *Experimental Brain Research* **129**: 325–46

Carpenter A F, Georgopoulos A P, Pellizzer G 1999 Motor cortical encoding of serial order in a context-recall task. *Science* **283**: 1752–7

Carter C S, Macdonald A M, Botvinick M, Ross L L, Stenger V A, Noll D, Cohen J D 2000 Parsing executive processes: strategic vs. evaluative functions of the anterior cingulate cortex. *Proceedings of the National Academy of Science USA* **97**: 1944–8

Connolly J D, Goodale M A, Desouza J F, Menon R S, Vilis T A 2000 Comparison of frontoparietal fMRI activation during anti-saccades and anti-pointing. *Journal of Neurophysiology* **84**: 1645–55

Fu Q G, Suarez J I, Ebner T J 1993 Neuronal specification of direction and distance during reaching movements in the superior precentral premotor area and primary motor cortex of monkeys. *Journal of Neurophysiology* **70**: 2097–116

Geyer S, Matelli M, Luppino G, Zilles K 2000 Functional neuroanatomy of the primate isocortical motor system. *Anatomy and Embryology* **202**: 443–74

Graziano M S, Gross C G 1998 Spatial maps for the control of movement. *Current Opinion in Neurobiology* **8**: 195–201

Hikosaka O, Sakai K, Miyauchi S, Takino R, Sasaki Y, Putz B 1996 Activation of human presupplementary motor area in learning sequential procedures: a functional MRI study. *Journal of Neurophysiology* **76**: 617–21

Kalaska J F, Scott S H, Cisek P, Sergio L E 1997 Cortical control of reaching movements. *Current Opinion in Neurobiology* **7**: 849–59

Picard N, Strick P L 1996 Motor areas of the medial wall: a review of their location and functional activation. *Cerebral Cortex* **6**: 342–53

Sanes J N 2000 The relation between human brain activity and hand movements. *Neuroimage* **11**: 370–4

Sanes J N, Donoghue J P 1997 Static and dynamic organization of motor cortex. *Advances in Neurology* **73**: 277–96

Shima K, Tanji J 2000 Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology* **84**: 2148–60

Strick P L, Dum R P, Picard N 1998 Motor areas on the medial wall of the hemisphere. *Novartis Foundation Symposium* **218**: 64–75

Todorov E 2000 Direct cortical control of muscle activation in voluntary arm movements: a model. *Nature Neuroscience* **3**: 391–8

Wise S P, Boussaoud D, Johnson P B 1997 Caminiti Premotor and parietal cortex: corticocortical connectivity and combinatorial computations. *Annual Review of Neuroscience* **20**: 25–42R

J. Donoghue

Neural Representations of Objects

1. How Small Numbers of Neurons Code Information About Objects

One of the most commonly measured units of representation in the brain is the neuron. Somewhat surprisingly, functional information about objects seems to be coded at this incredibly fine scale. Single-unit recordings of neurons in the visual cortex reveal activity that is often tuned to individual objects or classes of objects. That is, a particular stimulus object presented in the receptive field of a given neuron will produce significantly greater activity (more spikes per second) in that neuron as compared to any other tested stimulus (typically a varied assortment of natural and artifactual objects). For example, neurons in many cortical extrastriate areas (the part of the primate brain that is thought to support high-level vision, including object recognition) in the monkey brain are highly selective to primate faces and, sometimes, even individual faces, and recent evidence has revealed similar selectivity for many classes of familiar objects (Sheinberg and Logothetis 2001). There is similar evidence for cortical neurons specifically tuned to individual nonface objects, at least when the stimuli are highly familiar to the monkey (Logothetis and Sheinberg 1996). At the same time, few researchers believe that single neurons actually code for individual objects. Rather, it is commonly held that large populations of neurons represent objects or object classes in a distributed fashion. Some evidence for this hypothesis is provided by single-unit studies that systematically decompose the response of a given neuron by exploring which particular features of an object are actually driving neural activity. It is often the case that the maximum response is maintained when only schematic elements of the object are presented, suggesting that single neurons must work in concert to represent an object (Tanaka 1996).

To the extent that individual neurons participate in the representation of a given object, there remains the question of the format of this representation (Marr 1982). Object representations can be more or less invariant over changes in the appearance of the object in the image (see *Object Recognition: Theories*). Ideally, perception and recognition should remain robust when an object is moved, rotated, or illuminated differently. Most object-tuned neurons exhibit some invariance: the level of response to a preferred object is not dramatically affected by changes in the size or the spatial position of an object for which a given neuron is selective. On the other hand, these same neurons often show great sensitivity to changes in orientation, viewpoint, and illumination direction. For example, a large percentage of object-tuned neurons are also 'view-tuned' in that they show the highest level of activity to specific objects in specific viewpoints. Face-selective neurons may respond most strongly when presented with a frontal view of a face, but show a progressively diminished response as the face rotates away from this viewpoint. Similarly, in the visual cortex of monkeys taught to recognize novel objects, neurons that become selective for these objects are typically highly tuned to the particular viewpoints used during training. These findings are consistent with much of the behavioral data on object recogni-

tion, where recognition performance is often strongly dependent on the familiarity of a given viewpoint (Tarr and Bülthoff 1998). Thus, at the finest scale of analysis, single neurons participate in the representation of objects in a manner that is invariant over size and location, but specific to viewpoint and illumination.

1.1 Hierarchical Processing in the Visual Cortex

How can single neurons or populations of neurons code the complex patterns and feature conjunctions present in most objects? Many models of object recognition assume that the answer is that object- or feature-tuned neurons are the end-point of a hierarchy of progressively more and more complex visual codes (see *Visual System in the Brain*). Supporting this approach, neurons in the earliest visual areas (see Occipital lobe) respond to very simple generic properties of objects, for instance, oriented edges or blobs of color. As one moves 'up' the visual pathway (see Dorsal and ventral visual systems) the coding becomes more and more elaborate, with neurons at one level having responses that generally code more complex properties of objects than the level(s) below them (Van Essen and Maunsell 1983). The principle behind this hierarchy is straightforward: from one level to the next, neurons respond only if they receive input from multiple neurons, thereby creating a code that is necessarily more complex than the individual inputs. This many-to-one principle not only pools information in the straightforward sense, but because of the pattern of connectivity from one layer to the next, assembles new and more complex features as one moves up the hierarchy. Given the strong evidence for this sort of feed-forward processing in the visual system, many computational models, even those that are quite dissimilar in other respects, have adopted a hierarchical processing architecture to assemble complex object features (Hummel and Biederman 1992). Although the hierarchical approach is appealing, some computational theorists have pointed out that constructing complex features in this fashion may lead to overly specific object representations. For instance, imagine using the features present in the image to derive a very precise description of an object. If the object is then rotated in depth how does one combine the new very precise description with its predecessor? To address this need for invariance over changes in the image, some computational models have adopted a hierarchical architecture in which there is an alternation between conjunctions of features—the standard hierarchical model—and disjunctions of features (Riesenhuber and Poggio 1999). For example, one layer might consist of neurons that respond only if they receive input from a *collection* of neurons coding for simpler features, while the next layer might consist of neurons that respond if they receive input from *any* of several neurons. This sort of hierarchical coding is exemplified by a single neuron that codes for a three-dimensional object regardless of its orientation, but only by virtue of the fact that this view-invariant neuron is driven by any one of several view-tuned neurons, each coding for a different viewpoint of the same object (Logothetis and Sheinberg 1996).

1.2 The Building Blocks of Object Representation

The hierarchical organization of the visual cortex suggests that objects are represented in terms of a 'vocabulary' of features constructed from simpler features (Marr 1982). This assumption implies that somewhere between orientation-tuned neurons in early vision and object-tuned neurons at the highest levels of the visual system, are neurons representing the critical features that distinguish one object from another. Under one view, these critical features should look something like object parts; intuitively the first level of decomposition when we look at objects (Marr 1982). There is, however, little evidence to indicate that neurons coding for parts as we know them are the precursors to object-tuned neurons. When the responses of neurons in intermediate and high-level visual areas are systematically decomposed, they show consistent responses to patterns that are more complex than edges, but less complex than objects. However, these 'features' appear arbitrary, coding for odd, almost random shape properties. For example, one neuron might respond most strongly to a lollipop shape, while another neuron might respond to a eight-point star. While such features may appear to be random or, at best, schematic versions of certain objects, they are not arrayed haphazardly across the visual cortex. Rather, as with most visual areas, there is a columnar organization in which neurons within a column all code for the same feature and adjacent columns tend to code for similar features, albeit with columnar boundaries sometimes representing dramatic changes in feature selectivity (Tanaka 1996). It is worth noting that although such features might provide the 'building blocks' for representing objects, there is, at present, no plausible account that can explain how this occurs or why these particular features are used.

2. How Large Numbers of Neurons Code Information About Objects

As discussed above, despite neurophysiological results that seem to indicate that objects are represented by single neurons, there is little fondness for this stance. Many researchers have observed that single-unit methods limit the direct measurement of simultaneous

Figure 1
A novel object known as a 'Greeble'

neural activity to small numbers of units and it is only by considering single neuron responses in the context of population codes that we can understand the true nature of object representation (Booth and Rolls 1998). Thus, to the extent that a given neuron plays a larger role in the representation of one object over all other objects (or at least tested objects), a single-unit approach may bias the experimenter to interpret the neural response as 'tuned' for one object. The critical point is that this preference is not equivalent to 'representation.' Even if a neuron does not respond maximally to a given object, its partial (or even low) response may be part of the code by which a complex object is represented. Thus, individual objects and classes of objects may be represented by ensembles of 1,000s or 1,000,000s of neurons—something impossible to assess completely using neurophysiological measurements. In contrast, neuroimaging (Positron Emission Tomography—'PET'; and functional Magnetic Resonance Imaging—'fMRI') indirectly measures the conjoint activity of large numbers of neurons; the smallest unit of measurement being about a $3\,mm^3$ 'voxel' which encompasses approximately 1–2 million neurons.

2.1 Category Selectivity in the Visual Cortex

The neural representation of objects at this larger scale, as measured by PET or fMRI, reveals a degree of organization in extrastriate areas that mirrors the selectivity observed at the single-neuron level. That is, much as individual neurons respond preferentially to individual objects, localized regions of the visual cortex respond preferentially to classes of objects. Moreover, as with neurophysiological methods, the clearest preferences are obtained with highly-familiar object classes. A region of the visual cortex known as mid-fusiform gyrus (midFG) shows significantly higher activity when observers view faces as compared to when they view common objects (Kanwisher 2000). Similar selectivity for small regions of the cortex near the midFG has been found for letters, places, houses, and chairs. Thus, object classes appear to be represented minimally by large numbers of neurons in localized regions of the extrastriate cortex.

However, even this level of analysis may be misleading. Evidence for localized category-selective cortical regions comes from neuroimaging methods that compare the activation observed for one object class, e.g., faces, to a second object class, e.g., flowers. In truth, viewing any class of objects produces a pattern of activity across much of the ventral temporal cortex that is different from the activation pattern obtained for any other class. These differences, however, are often subtle and relatively small compared to the large differences seen between highly-overlearned categories such as faces and less familiar objects. Such differences may be critical elements in the complete neural code for objects; if so, objects and classes of objects may be represented as large-scale networks of neurons distributed over much of the visual cortex.

2.2 The Origins of Category Selectivity

If category selectivity is not a marker for the representation of objects *per se*, how can we explain the preferential responses obtained in localized regions of the visual cortex? One interpretation is that the visual cortex is organized into dedicated *modules* that are prewired to be selective for the geometries of particular object classes (Kanwisher 2000). Alternatively, category selectivity may reflect particular computations that, with experience, become automatically executed on distributed representations of objects and classes (Tarr and Gauthier 2000)—the 'Process Map' hypothesis. Again, the most salient example is faces: we are trained from birth to recognize faces at the individual level. The result of this lifetime experience is that, as adults, we are experts at processing faces and, by default, apply a degree of perceptual analysis that is not necessary for recognizing objects at more categorical levels. This interpretation suggests that the same computational resources should be applied to any object class (not just faces) for which observers are experts at recognizing individuals. Several studies bear out this prediction. Neuroimaging reveals that both car and bird experts show selectively higher activation in midFG for their domain of expertise. Similarly, for novel classes of objects (e.g., 'Greebles'; see Fig. 1), trained experts, but not untrained novices, exhibit increased midFG activation for the trained object class (Tarr and Gauthier 2000). Apparently, visual

experience with many similar objects 'tunes' midFG, as well as other category selective extrastriate areas, to respond automatically when those or other objects from the class are seen again (see *Neural Plasticity*). Consequently, category selectivity is perhaps best understood as a consequence of how experience reorganizes the visual recognition process, rather than an indicator of how objects themselves are neurally represented.

3. The Functional Role of Localized Brain Regions in Object Representation

Neuroimaging methods are useful for localizing *where* in the brain particular computations occur, but are less informative regarding the different roles these neural substrates play in producing behavior. In contrast, neuropsychological methods allow us to explore deficits in behavior as a direct consequence of brain injuries to particular brain regions. Given the neuroimaging results discussed in the previous section, damage to category-selective areas such as midFG should impair face recognition, and, perhaps, performance in any difficult individual-level recognition task. Such neuropsychological cases have been documented: Patients who suffer damage to extrastriate visuals areas often lose the ability to recognize many or all object classes—a syndrome known as *agnosia* (see *Agnosia*). Although these patients are grouped under a single label, agnosia includes damage to many different cortical regions and many different recognition deficits.

Accounts of category selectivity as revealed by neuroimaging argue for either modular subsystems organized by geometrically-defined object classes or for a flexible process map of distributed representations and processing mechanisms that are recruited by expertise and individual-level recognition. If the modularity argument is correct, then damage to the module specialized for faces should only result in deficits at face recognition but no deficits in nonface object recognition. Conversely, damage to the module specialized for objects or classes of objects should only result in deficits in object recognition, but not face recognition. The logic of the double dissociation method is clear: if two distinct brain regions perform independent functions, then lesions to one region should not affect the functioning of the other region and vice versa. Alternatively, if the process map model is correct and processing mechanisms can be recruited for the recognition of any object class, there should be common deficits in face and object recognition when there is damage to the ventral temporal cortex. It is critical, however, that the recognition tasks be equated in such comparisons: spurious differences may be found simply because different computational components of the recognition process are being recruited for different tasks.

3.1 Patterns of Sparing and Loss in Visual Object Recognition

Given knowledge about the location of the brain damage in an individual, a clear pattern of sparing and loss may help reveal the functional role of intact and damaged brain regions (see *Cognitive Neuropsychology, Methodology of*). Patients with injuries to their ventral temporal cortex often suffer a specific form of agnosia referred to as *prosopagnosia* (see *Prosopagnosia*) in which face recognition appears to be disproportionally impaired relative to object recognition (Farah 1990). At the same time, there is one case study in which the brain-injured patient appears to have intact face recognition, but severely impaired object recognition. Although this pattern is consistent with the modularity account, further examination of prosopagnosia does not support this conclusion. When patients with apparent face recognition deficits perform difficult individual-level discriminations with nonface objects (e.g., snowflakes), they show equal impairment with faces and nonfaces (Gauthier et al. 1999). Thus, neuropsychology yields a mixed bag of results supporting both modular and nonmodular accounts.

See also: Agnosia; Binding Problem, Neural Basis of; Face Recognition Models; Face Recognition: Psychological and Neural Aspects; Neural Synchrony as a Binding Mechanism

Bibliography

Booth M C A, Rolls E T 1998 View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex* **8**: 510–523

Farah M J 1990 *Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision*. MIT Press, Cambridge, MA

Gauthier I, Behrmann M, Tarr M J 1999 Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience* **11**: 349–370

Hubel D H 1995 *Eye, Brain, and Vision*. Freeman, San Francisco, CA

Hummel J E, Biederman I 1992 Dynamic binding in a neural network for shape recognition. *Psychological Review* **99**: 480–517

Kanwisher N 2000 Domain specificity in face perception. *Nature Neuroscience* **3**: 759–63

Logothetis N K, Sheinberg D L 1996 Visual object recognition. *Annual Review of Neuroscience*. Annual Reviews, Palo Alto, CA

Marr D 1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco, CA

Palmer S E 1999 *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, MA

Riesenhuber M, Poggio T 1999 Hierarchical models of object recognition in cortex. *Nature Neuroscience* **2**: 1019–1025

Sheinberg D L, Logothetis N K 2001 Finding familiar objects in real-world scenes: The role temporal cortical neurons in natural vision. *Journal of Neuroscience* **21**: 1340–1350

Tanaka K 1996 Inferotemporal cortex and object vision. *Annual Review of Neuroscience*. Annual Reviews, Palo Alto, CA
Tarr M J, Bülthoff H H 1998 *Object Recognition in Man, Monkey, and Machine*. MIT Press, Cambridge, MA
Tarr M J, Gauthier I 2000 FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience* **3**: 764–69
Van Essen D C, Maunsell J H R 1983 Hierarchical organization and functional streams in the visual cortex. *Trends in Neuroscience* **6**: 370–375

M. J. Tarr and Q. C. Vuong

Neuronal Synchrony as a Binding Mechanism

Neuronal systems have to solve immensely complex combinatorial problems and require efficient binding mechanisms in order to generate representations of perceptual objects and movements. In the context of cognitive functions, combinatorial problems arise from the fact that perceptual objects are defined by a unique constellation of features, the diversity of possible constellations being virtually unlimited. Combinatorial problems of similar magnitude have to be solved for the acquisition and execution of motor acts. Although the elementary components of motor acts—the movements of individual muscle fibres—are limited in number, the spatial and temporal diversity of movements that can be composed by combining the elementary components in ever-changing constellations is again virtually infinite. In order to establish neuronal representations of perceptual objects and movements the manifold relations among elementary sensory features and movement components have to be encoded in neural responses. This requires binding mechanisms that can cope efficiently with combinatorial complexity. Evolved brains have acquired an extraordinary competence to solve such combinatorial problems and it appears that this competence is a result of the evolution of the cerebral cortex.

1. *Two Complementary Binding Strategies*

In the primary visual cortex of mammals, relations among the responses of retinal ganglion cells are evaluated and represented by having the output of selected arrays of ganglion cells converge in diverse combinations onto individual cortical neurons. Distributed signals are bound together by selective convergence of feed forward connections (Hubel and Wiesel 1962). Through iteration of this strategy neurons in higher cortical areas acquire increasingly sophisticated response properties, representing complex constellations of features and, in the special case of faces, even whole perceptual objects (Gross 1992, Tanaka 1997).

However, this strategy—of binding features together by recombining input connections in ever-changing variations and representing relations explicitly by responses of specialized binding cells—results in a combinatorial explosion of the number of required binding units. This problem is further accentuated by the fact that perceptual objects are often defined by conjunctions of features encoded in different sensory systems; they have specific visual, haptic, or acoustic properties. Moreover, if binding is achieved solely by the formation of conjunction-specific neurons, difficulties arise if novel objects with unfamiliar feature constellations need to be represented. It has been proposed, therefore, that the cerebral cortex uses a second, complementary strategy, which permits utilization of the same neurons for the representation of different objects (Hebb 1949). Here, the particular constellation of features characterizing perceptual objects is thought to be represented by the joint and coordinated activity of a dynamically associated ensemble of cells, whereby each of the participating neurons represents explicitly only one of the elementary features that characterize a particular perceptual object. Different objects can then be represented by recombining in various constellations neurons tuned to elementary features, whereby a particular cell can participate at different times in different assemblies. This sharing of neurons economizes on neuron numbers and copes effectively with the huge combinatorial space occupied by real-world objects.

For assembly coding two constraints need to be met, however. First, a selection mechanism is required that permits dynamic, context-dependent association of neurons into distinct, functionally coherent assemblies. Second, grouped responses must get labeled so that they can be distinguished by subsequent processing stages as components of one coherent representation and do not get confounded with other unrelated responses.

Such labeling can be achieved by jointly raising the saliency of the selected neurons, because this ensures that they are processed and evaluated together at the subsequent processing stage. Neuronal systems have three options to increase the relative saliency of selected responses: first, non-grouped responses can be inhibited; second, the amplitude of the selected responses can be enhanced; and third, the selected cells can be made to discharge in precise temporal synchrony. Inhibition simply eliminates potentially confounding signals, while enhanced firing raises the impact of the grouped responses through temporal summation and synchronization through spatial summation of synaptic potentials (Singer 1999). Selection of responses by enhancing discharge rates is common in attentional gating (Luck et al. 1997) but it has certain disadvantages when used as the only mech-

anism for the distinction of assemblies (von der Malsburg 1981, Singer 1999). Discharge rates of feature selective cells can vary over a wide range when stimuli change and since these modulations of response amplitude would not be distinguishable from those signaling membership of a cell in an assembly, ambiguities could arise. Problems could also arise when simultaneous maintenance of different assemblies is required at the same processing stage. This may be the case when different objects and their relations to one another need to be represented simultaneously. If the corresponding assemblies were defined solely by the increasing discharge rate of their member neurons it would be impossible to distinguish which of the enhanced responses belonged to which assembly.

These ambiguities can be overcome if the selection and labeling of responses is achieved through synchronization because it can be adjusted independently of rate fluctuations, and relations between responses can be defined with very high temporal resolution. Synchronization enhances only the saliency of those discharges that are precisely synchronized and generate coincident synaptic potentials in target cells at the subsequent processing stage. Hence the selected event is the individual spike or a brief burst of spikes. Thus, different relations can be defined on short timescales and hence different assemblies can alternate with one another at fast rate within the same neuronal network without getting confounded. The rate at which assemblies can be multiplexed is limited only by the duration of the interval over which synaptic potentials are integrated. If this interval is in the range of 10 to 15 ms, different assemblies can alternate with a frequency up to 40 Hz and hence several different feature constellations can be corepresented within perceptually relevant time windows.

2. *Synchrony as a Signature of Relatedness*

The discharges of neurons in the cerebral cortex reflect with great precision the temporal structure of stimuli. Synchronously presented stimuli evoke synchronous responses. This makes it possible to examine with psychophysical experiments whether synchronized responses are indeed interpreted as a signature of relatedness. The results indicate that the synchronous responses evoked by simultaneously presented stimuli are indeed bound perceptually and get segregated from the temporally offset responses evoked by asynchronously presented pattern elements. In the visual system the temporal resolution with which synchronous responses can be segregated from temporally offset responses is considerably better than 10 ms (for review see Singer 1999). This indicates that precise temporal relations among the discharges of distributed neurons are exploited for perceptual grouping and that offset intervals are resolved that are considerably shorter than those that can be consciously perceived.

In the mid-1980s it was discovered that neurons distributed across the visual cortex synchronize their responses when responding to a single visual object (Gray and Singer 1987, 1989). This synchronization was found to be due to dynamic interactions among reciprocally-coupled cortical neurons. It was often associated with rhythmic firing in the frequency range of 40 Hz—the so-called gamma frequency band—and was as precise as the synchrony induced by external, temporally modulated stimuli. Based on the theoretical considerations exposed above it had been proposed that this internal synchronization could serve to select and bind responses in all those cases where selection and binding by rate enhancement might lead to ambiguity. Since then numerous experiments have been performed in order to test predictions derived from this hypothesis (for review see Singer 1999). In particular, evidence needed to be obtained that response synchronization is not simply an epiphenomenon of neuronal operations but does have a function in sensory and motor processing. This requires demonstration that (a) synchronous activity has a stronger impact in target structures than non-synchronous activity, (b) synchrony can be established and changed sufficiently rapidly to be compatible with processing speed, and (c) synchronization patterns are related closely to cognitive processes requiring dynamic selection and binding of responses. Such is the case for feature binding and perceptual grouping, for response selection by attention, for the establishment of novel conjunctions between stimuli or between stimuli and responses in learning paradigms, and for the coordination of complex movements.

3. *Rapid Synchronization*

Both simulation studies with spiking neurons and recordings from the brain revealed that networks can undergo very rapid transitions from uncorrelated to synchronized states (e.g., Gerstner 1996). One factor contributing to very rapid synchronization is the coordinated fluctuations of neuronal excitability. Evidence from the visual cortex indicates that ongoing activity is not random but exhibits correlated fluctuations. Cortical neurons with colinearly aligned receptive fields tend to engage in coherent excitability fluctuations due to strong reciprocal coupling. These fluctuations act like a dynamic filter and cause a virtually instantaneous synchronization of the very first discharges of responses (Fries et al. 2001a). The spatio-temporal patterns of these spontaneous fluctuations of excitability reflect the architecture and the actual functional state of the intracortical association connections which mediate the reciprocal interactions among distributed cortical neurons and are held responsible for the grouping and binding of responses. Spontaneous activity can thus be considered as the reflection of a continuous readout of the grouping criteria that reside in the functional architecture of

cortical connections. Because of this anticipatory patterning, grouping by synchronization can be extremely fast and occur as a function of both the prewired associational dispositions and the current functional state of the cortical network.

4. Relation to Perception

Stimulus-induced synchronization does not simply reflect anatomical connectivity but depends critically on the configuration of the stimuli. It is the result of a dynamic and context-dependent self-organization process that follows, so far without exception, the Gestalt rules that govern perceptual grouping. In the visual domain this has been shown for the criteria of continuity, proximity, similarity in the orientation domain, colinearity, and common fate (Gray et al. 1989, Castelo-Branco et al. 2000).

Most of the early experiments on response synchronization were performed in anesthetized animals but more recent evidence indicates that internally generated synchrony is even more pronounced in the awake brain, exhibiting similar sensitivity to context (Kreiter and Singer 1996), and becoming particularly precise when the EEG is desynchronized (Munk et al. 1996) and the animals are attentive (Roelfsema et al. 1997, Steinmetz et al. 2000, Fries et al. 2001b).

Direct relations between response synchronization and perception have been found in cats that suffered from strabismic amblyopia, a developmental impairment of vision associated with suppression of the amblyopic eye, reduced visual acuity and disturbed perceptual grouping in this eye. Quite unexpectedly, the discharge rates of individual neurons in the primary visual cortex fail to reflect these deficits (see Roelfsema et al. 1994 for references). The only significant correlate of amblyopia at the level of the primary visual cortex is the drastically reduced ability of neurons driven by the amblyopic eye to synchronize their responses (Roelfsema et al. 1994), and this accounts well for the perceptual deficits: reduced synchronization leads to reduced saliency of responses, and this could explain the suppression of signals from the amblyopic eye. If synchronization has a function in response selection and binding, disturbed synchronization could explain the reduced visual acuity and the impaired figure-ground segregation in amblyopic vision.

Another close correlation between response synchronization and perception has been found in experiments on binocular rivalry (Fries et al. 1997). A highly significant correlation existed between changes in the strength of response synchronization in primary visual cortex and the outcome of rivalry. Responses to stimulation of the eye that won in interocular competition that were perceived by the animal increased their synchronicity upon presentation of the rival stimulus to the other, losing eye, while the reverse was true for cells driven by the eye that became suppressed. Interestingly, however, there were no rivalry-related changes in the response amplitudes of the neurons.

These results support the hypothesis that precise temporal relations between the discharges of spatially distributed neurons play a role in cortical processing and that synchronization may be exploited to jointly raise the saliency of the responses selected for further processing, i.e., for the dynamic binding of distributed responses.

The example of rivalry also illustrates how synchronization and rate modulation depend on each other. The signals from the suppressed eye failed to induce tracking eye movements and were not perceived by the animal, suggesting that the vigorous but poorly synchronized responses in primary visual areas eventually failed to drive the neurons responsible for the execution of eye movements. Direct evidence for a loss of neuronal responses in higher cortical areas to stimuli presented to the suppressed eye has been obtained in awake monkeys (Leopold and Logothetis 1996). Thus, changes of synchronicity result in changes of response amplitudes at subsequent processing stages. This convertibility provides the option to use both coding strategies in parallel in order to encode complementary information.

5. Epilogue

The discovery of stimulus-induced response synchronization and the proposal that synchronization might serve as a neuronal code for the definition of relations have solicited highly controversial debates in the scientific community. Most of the arguments raised in this context have been summarized recently in a series of reviews that have been published in the journal *Neuron* (1999, Vol. 24, pp. 7–125). These contributions focus on the putative role of synchronization in visual perception, as does this article. However, evidence has been obtained that response synchronization is an ubiquitous phenomenon not only in the cerebral cortex and its many subdivisions but also in the nervous systems of invertebrates. As in the visual cortex these synchronization phenomena are often associated with an oscillatory modulation of the underlying discharge patterns, whereby the oscillation frequencies tend to concentrate in the beta- and gamma-frequency band (20–60 Hz). Due to effective summation of currents associated with synchronous activity, the episodes of response synchronization can be detected with non-invasive recording techniques such as electroencephalography (EEG) and magnetoencephalography (MEG). Following the demonstration of stimulus and task-dependent synchronous oscillations in the visual system, numerous EEG and MEG studies have been performed on human subjects, in search of correlations between the occurrence of response synchronization on the one hand and cognitive or executive functions on the other. The results of these studies are compatible with the view that an

oscillatory patterning of neuronal activity in the beta- and gamma-frequency band and the synchronization of this oscillatory activity are associated with a wide diversity of higher brain functions, such as perceptual grouping, object recognition, short-term memory, associative learning, selective attention, and calibration of motor functions (for review see Tallon-Baudry and Bertrand 1999, Singer 1999).

Likewise, close relations have been found between patterns of response synchronization and cognitive functions in the insect olfactory system (reviewed in Singer 1999). Apart from the investigations in insects, most of the evidence for a functional role of synchronization is still correlative in nature because it is difficult to selectively interfere with synchronization without affecting other response variables. It is undeniable, however, that the new approach—to register neuronal activity simultaneously from different sites and to search for temporal relations among the activity of distributed neuronal populations, rather than measuring only the rate variations of individual cells—provides new and fascinating insights into the dynamics of neuronal processes. In a number of investigated paradigms, changes in temporal correlations among distributed neuronal responses turned out to be a better predictor of stimulus configurations, of attentional states and of the outcome of a perceptual process than the rate changes of individual neurons. This suggests that the nervous system exploits the option to encode information in the precise temporal relations between responses of distributed neurons. It is highly likely that massive parallel recording from large numbers of neurons will reveal synchronization to be merely a special and rather simple state in a highly complex state space, and it is one of the challenges for systems physiology to decipher the alphabet of these temporal codes.

See also: Binding Problem, Neural Basis of; Feature Representations in Cognitive Psychology; Neural Representations of Intended Movement in Motor Cortex; Neural Representations of Objects

Bibliography

Castelo-Branco M, Goebel R, Neuenschwander S, Singer W 2000 Neural synchrony correlates with surface segregation rules. *Nature* **405**: 685–9

Fries P, Neuenschwander S, Engel A K, Goebel R, Singer W 2001a Rapid feature selective neuronal synchronization through correlated latency shifting. *Nature Neuroscience* **4**: 194–200

Fries P, Reynolds J H, Rorie A E, Desimone R 2001 Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **291**: 1560–63

Fries P, Roelfsema P R, Engel A K, König P, Singer W 1997a Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 12699–704

Gerstner W 1996 Rapid phase locking in systems of pulse-coupled oscillators with delays. *Physical Review Letters* **76**: 1755–8

Gray C M, König P, Engel A K, Singer W 1989 Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**: 334–7

Gray C M, Singer W 1987 Stimulus-specific neuronal oscillations in the cat visual cortex: A cortical functional unit. *Society of Neuroscience Abstracts* **13**: 404.3

Gray C M, Singer W 1989 Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* **86**: 1698–702

Gross C G 1992 Representation of visual stimuli in inverior temporal cortex. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **335**: 3–10

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Hubel D H, Wiesel T N 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* (*London*) **160**: 106–54

Kreiter A K, Singer W 1996 Stimulus-dependent synchronization of neuronal responses in the visual cortex of awake macaque monkey. *Journal of Neuroscience* **16**: 2381–96

Leopold D A, Logothetis N K 1996 Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* **379**: 549–53

Luck S J, Chelazzi L, Hillyard S A, Desimone R 1997 Neural mechanisms of spatial selective attention in areas V1, V2 and V4 of macaque visual cortex. *Journal of Neurophysiology* **77**: 24–42

Munk M H J, Roelfsema P R, König P, Engel A K, Singer W 1996 Role of reticular activation in the modulation of intracortical synchronization. *Science* **272**: 271–4

Roelfsema P R, Engel A K, König P, Singer W 1997 Visuomotor integration is associated with zero time-lag synchronization among cortical areas. *Nature* **385**: 157–61

Roelfsema P R, König P, Engel A K, Sireteanu R, Singer W 1994 Reduced synchronization in the visual cortex of cats with strabismic amblyopia. *European Journal of Neuroscience* **6**: 1645–55

Singer W 1999 Neuronal synchrony, a versatile code for the definition of relations? *Neuron* **24**: 49–65

Steinmetz P N, Roy A, Fitzgerald P J, Hsiao S S, Johnson K O, Niebur E 2000 Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature* **404**: 187–90

Tallon-Baudry C, Bertrand O 1999 Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences* **3**: 151–62

Tanaka K 1997 Mechanisms of visual object recognition: Review of monkey and human studies. *Current Opinion in Neurobiology* **7**(4): 523–9

von der Malsburg C 1981 The Correlation Theory of Brain Function. Internal Report 81–2, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany. [Reprinted 1994. In: Domany E, van Hemmen J L, Schulten K (eds.) *Models of Neural Networks II*. Springer-Verlag, New York, pp. 95–119]

W. Singer

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Neural Systems and Behavior: Dynamical Systems Approaches

Humans move with ease over rough terrain (and so do many other species), identify and orient toward perceived objects in cluttered environments, or maintain objects and the intention to reach for them in memory while sensory information about the objects is no longer directly available and other tasks must first be accomplished. How do nervous systems generate persistent patterns of behavior in the face of very complex sensory environments, and given the continuous scale of possible behaviors within which any particular action or representation must be stabilized? Dynamical systems approaches address the fundamental conflict between stability and flexibility of nervous function. Stability is the property of a behavioral or neural state to persist in the face of systematic variation or random perturbation. All observable behavioral or neural states possess stability to some extent, as loss of stability leads to change of state. The generation of behavior or of neural patterns of activity is flexible in the sense that a variety of behavioral or neural patterns may be assembled from a multitude of sensory, motor, and neural systems in response to the sensory situation, the inner state of the neural or behavioral system, and its recent history. Flexibility may therefore involve pattern change and hence require the release of behavioral or neural states from stability. Dynamical systems approaches provide methods with which stable states can be identified and characterized as 'attractors.' The coexistence of multiple attractors and changes in the number and nature of attractors, as sensory information, environmental conditions, or task requirements vary, are important in dynamic systems approaches because they demarcate different elementary behaviors or functional modules.

1. *Stability*

The functioning of technical systems (for example, maintaining an airplane on course) must be stable against random perturbations (for example, from turbulence). Stability may be brought about by the physical properties of the system (for example, by the viscous properties of air), by control systems designed specifically to measure deviations from the desired state and change actuator parameters to minimize such deviations, or by a human operator acting as a control system. The theoretical framework of control theory has been, and is being, used to understand simple aspects of nervous function, an approach known as *biological cybernetics*. To analyze the capacity of the housefly to orient toward a moving visual stimulus, for instance, Reichardt and Poggio (1976) identified the control dynamics linking the detected motion stimulus to the input to the fly's flight motor. Such cybernetic analysis is aimed at identifying relevant sensory signals, the motor system, and the control circuitry transforming sensory signals into motor parameters. This approach is successful as long as the number of components is small, their function is independent of each other, and a single behavioral pattern is generated.

2. *Flexibility*

Any behavior involves many sensory subsystems, many motor degrees of freedom, and takes place in complex environments. Certainly true of daily life human activities, this also holds for laboratory experiments, even when they seem to limit the number of relevant components. The simple behavioral task, for instance, of pressing a key in response to a visual stimulus, actually involves many sensory subsystems (including proprioception, haptics, reafference, and vision) that contribute to the estimation of the finger's current position, with additional subsystems (maintaining reference frames for eye, head, and body position, different visual features including depth and motion) contributing to the processing of the visual stimulus. Those subsystems go unnoticed because the information they provide is coherent. This is, in part, a property of the environment (in which, for instance, haptic and visual cues about finger position co-vary), and, in part, due to the experimental task setting which generates a unique relationship between sensory stimulation and motor action.

A unique match of sensory onto motor configurations is the exception more than the rule, however. Even Reichardt and Poggio's (1976) houseflies were more flexible than that: When two separate patches of the visual array moved, the flies steered toward an averaged direction as long as the two patches lay in similar directions, but locked on to only one of the two patches when these were at larger angular separation. Frequent random switches occurred between epochs in which the flies tracked one or the other visual target selectively. More generally, not all possible couplings between sensory and motor degrees of freedom can be effective all the time. The selective activation of particular couplings between sensory and motor components, and the release from stable coupling of others, is a necessary prerequisite for the generation of flexible behavior. Initiating a motor act, for instance, requires releasing a fixation or postural system from stability.

3. *The Dynamical Systems Approach*

Broadly speaking, dynamical systems approaches (Port and van Gelder 1995) generalize cybernetic ideas so as to encompass flexibility. The specific variant reviewed in the rest of this article, a dynamical systems approach emphasizing attractor dynamics and their instabilities, grew out of analogies between pattern

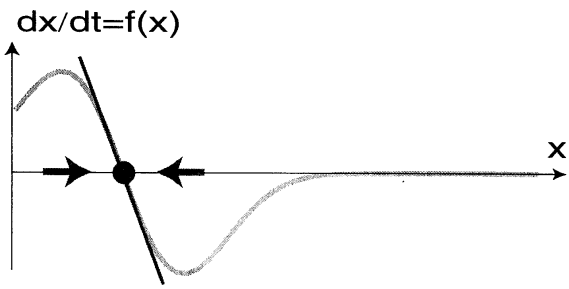

Figure 1
The temporal evolution of a behavioral or neural variable, x, is characterized by a dynamical system, if for any value of x there is a unique rate of change, $\mathrm{d}x/\mathrm{d}t = f(x)$, such as plotted here (light grey line). A fixed point, at which the rate of change is zero, is an attractor (filled circle), if the system converges to the fixed point from nearby states. The negative slope of the rate of change at the intersection with the x-axis indicates stability: initial states below the fixed point have positive rate of change and thus grow toward the fixed point; initial states above the fixed point have negative rate of change and thus decrease toward the fixed point (arrows). The linear approximation (straight line in grey, tangent to $\mathrm{d}x/\mathrm{d}t$ at the fixed point) characterizes the system well close to, but not far from, the attractor

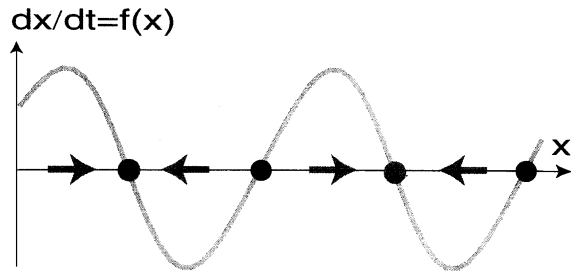

Figure 2
Only nonlinear dynamical systems can have multiple attractors. In the example shown, there are two attractors (circles filled in grey) and two repellors (circles filled in black). One of the repellors forms the attractor boundary such that solutions starting from its left converge to the left-most attractor, solutions starting from its right converge to the right-most attractor

formation in nonequilibrium systems (Haken 1983) and patterns of coordination in biological movement (Schöner and Kelso 1988), but has since been extended to address many aspects of action, cognition, and perception (e.g., Thelen and Smith 1994, Giese 1999). A link to dynamical systems ideas in neural network theory (Amari 1977, Grossberg 1980) has been established through the concept of dynamical fields (Schöner et al. 1997).

3.1 Behavioral or Neural Variables

Variables that are very closely linked to the peripheral motor or sensory subsystems are often not suited to describe behavioral or neural patterns, which may involve many different such subsystems or even different combinations of such subsystems. The dynamical systems approach employs, instead, variables that characterize directly the behavioral or neural patterns themselves by differentiating the different observed forms of the patterns and quantifying deviations from behavioral or neural states. When, for instance, Reichardt and Poggio's (1976) houseflies switch stochastically between two patterns of locking on to either of two moving targets, their heading direction in an outer reference frame is a useful variable in terms of which both states and deviations from them can be described, even though this variable is not directly available to sensor or motor systems. Other examples of behavioral variables that are useful to characterize multiple behavioral states are the relative phase between different limbs or articulators (Schöner and Kelso 1988), the position and velocity of an end-effector in outer (task) space to characterize the formation of trajectories in goal-directed movements (Saltzman and Kelso 1987), or the directions of reaching movements in outer space to characterize how movement targets are represented (Georgopoulos et al. 1986).

3.2 Behavioral or Neural Dynamics

Any behavioral or neural system is exposed to multiple sources of potential perturbation. Stability, resistance to change, is thus a necessary prerequisite for a pattern to be reproducible and observable. Even when transient patterns are generated (in response to stimulation, for instance) it is the asymptotic state and its stability properties that determine the form of such transients. This stability requirement is central to the dynamical systems approach, which aims to characterize and analyze behavioral and neural systems in terms of its stable states.

Given behavioral or neural variables, the approach can be formalized in terms of dynamical systems that describe the evolution in time of the system. The dynamical system is characterized by its asymptotically stable states or *attractors* to which the variables converge in time (Fig. 1). Linear dynamical systems have exactly one state, and linear systems theory provides formal methods for how to identify the dynamical system given such measures as the means, variances, and co-variances of the behavioral or neural variables. For a system to have more than one attractor, the dynamics must be strongly nonlinear (Fig. 2). There is no formal procedure for identifying such systems. One heuristic approach is to characterize

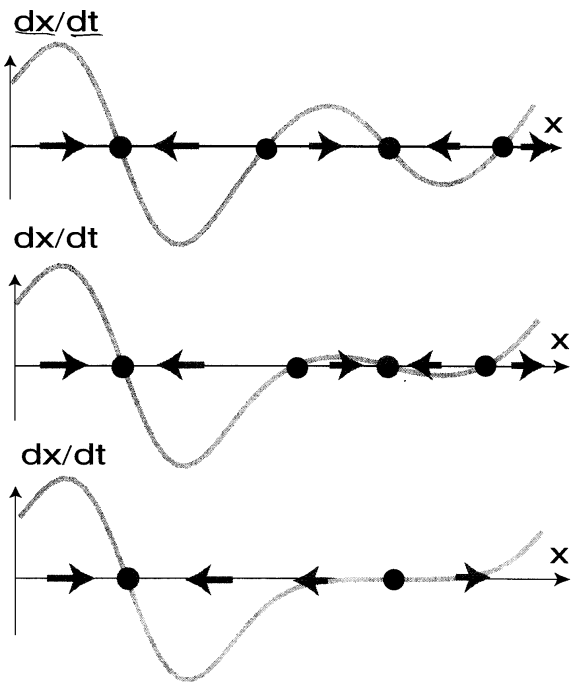

Figure 3
Changes in the number of attractors occur through bifurcations. In the pitch-fork bifurcation illustrated here (from top to bottom), as the right-most attractor becomes less stable (negative slope of rate of change at fixed point approaches zero), the flanking repellors approach, until their collision removes the attractor, leaving a single repellor behind

the different attractors by local linear models which can by fused into a nonlinear model by taking the limited range of validity of the linear models into account (Schöner et al. 1995). Noise induced switches between different attractors provide additional information about the nonlinear dynamics.

The strongest tool for the identification of such nonlinear dynamics is the observation of bifurcations, also called instabilities. At bifurcations, the number or nature of attractors changes (Fig. 3). The detection of bifurcations as sensory information, task constraints, or motor parameters vary helps identify qualitatively different contributions to the behavioral or neural dynamics. States that may coexist multistably or are linked through bifurcations reflect separate elementary behaviors or neural patterns. The various quadrupedal gaits, for instance, are different elementary behaviors rather than variants of the same basic behavior, because changes from one gait to another may take the form of bifurcations. Kelso and colleagues studied a laboratory model of such patterns of coordination (Schöner and Kelso 1988) when they asked human participants to move two homologous limbs rhythmically (e.g., fingers, arms, legs). In-phase and anti-phase patterns of coordination were identified as two elementary behaviors, which coexist bistably at low frequency of the rhythmic movement. Anti-phase coordination (syncopation) becomes unstable in a pitchfork bifurcation (Fig. 3) when movement frequency is increased. Other domains in which this form of analysis can be applied include the integration of multiple sources of sensory information to control action (such as in posture), the establishment of stable coupling between perceived events and action (such as when catching a ball), or the control of timing itself (such as when generating polyrhythms). At the neural level, central pattern generators and their interactions with sensory and motor subsystems can be analyzed this way. This is of particular interest in view of the ubiquitous multifunctionality of such circuits.

3.3 Dynamic Fields

What would be a dynamical systems description of the behavior of generating, in response to a coded stimulus, a goal-directed movement to one of three previously learned, but unmarked, targets? The range of possible movements may be characterized in terms of an appropriate behavioral variable, e.g., the direction in which the target lies from the initial end-effector position. But how to represent the three possible values of this variable? Dynamical variables have a unique value at all times, and that value must change continuously over time. The missing element is some method for representing variable and graded amounts of information. The classical concept of activation, when combined with the notion of a behavioral variable, may supply this missing element. For each potential value of the behavioral variables, an activation variable represents the amount of information about this value: Absence of information is indicated by low levels of activation, and presence of information by large levels of activation. This leads to a field of activation variables defined over the space spanned by the behavioral variables. A single peak of activation represents information about a unique value of the behavioral variables (that over which the peak is localized: Fig. 4(a)). Multiple possible values are represented by multiple peaks, and graded distributions of activation represent graded information about the values of the behavioral variables. In neurophysiological terms this resembles the notion of 'space code,' in particular, when topography leads to regular neighborhood relationships within a neural network.

Ideas about attractor dynamics and bifurcations can be 'lifted' to this level of description by postulating that such activation fields evolve toward attractor solutions of a dynamical system, which stabilizes localized peaks of activation (Schöner et al. 1997). Interpreted as neural networks, such dynamical system

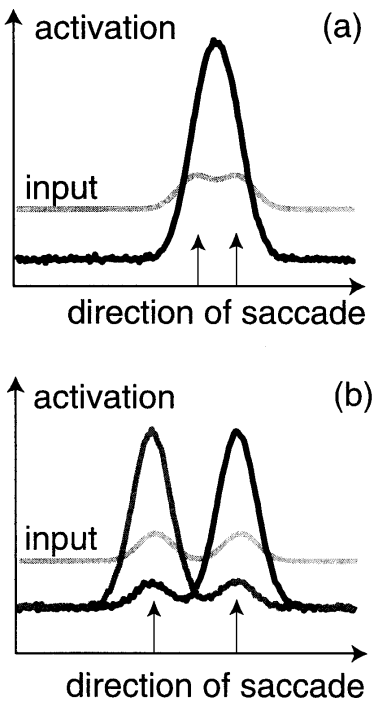

Figure 4
Two visual targets (arrows) generate bimodal input (light grey) into the dynamic field representing the direction of planned saccades. (a) At small angular separation, the field dynamics stabilizes a single peak of activation (black), representing a saccade to an averaged direction. (b) At larger angular separation, one of two activation patterns emerges, each with a single peak located either over the left-most target (dark grey) or the right-most target (black). The transition between these regimes is a bifurcation of the field dynamics

receive input from sensors or other representations that specify locations at which peaks may be localized. In addition, strong internal interactions within the dynamic field stabilize individual localized peaks even when input is distributed across the field. Such strongly interactive fields are thus not primarily input-driven. Dynamical neural networks of this kind have been studied as models of cortical function (Amari 1997, Grossberg 1980).

As a simple example consider saccadic eye movements generated toward visual targets. When two potential targets are presented, two types of responses are possible. For narrowly spaced targets, an eye movement toward an average position between the two targets is generated. Beyond a critical angular separation of the two targets, a bimodal response pattern is obtained in which on some trials the saccade goes to one target, and on other trials, to the other target (Ottes et al. 1984). A dynamic field representing the direction of the planned saccade receives input from the target locations. When this input consists of two narrowly spaced peaks, a single peak of activation localized over an averaged direction is generated (Fig. 4(a)). When the two input peaks are more separate, mutual inhibition makes it impossible for peaks to become activated at both input sites. The field dynamic is bi-stable and generates a single peak, positioned either at one or at the other stimulated location. Which of the two targets is selected depends on the pattern of prior activation, on fluctuations, or on small differences in stimulation between the two sites.

Through the notion of dynamic fields, the dynamical systems approach can be extended from the domain of sensori-motor patterns toward problems in perception (Giese 1999) and representation (Thelen et al. 2001). A dynamic field may, for instance, change through a bifurcation from an input-driven mode, in which activation remains stable only in the presence of appropriate input, to a short-term memory mode, in which input-induced activation remains stable once input has been removed. The dynamic field can be linked to the notion of population coding in neurophysiology (Georgopoulos et al. 1986) by interpreting distributions of population activation as fields of activation over a relevant behavioral variable, to which neurons are tuned (Erlhagen et al. 1999).

4. *Implications*

If behavioral or neural patterns emerge as stable states from underlying dynamical systems, as postulated in dynamical systems approaches, then the characterization of this dynamics is sufficient for an understanding of the behavioral or neural system. No additional need for computation or specification arises.

With origins in the domains of motor behavior and sensori-motor coupling, the more specific dynamical system approach, emphasizing attractors and their bifurcations, is now being probed as a wider framework for the understanding of nervous function also at the level of perceptual and cognitive representations. Multistability and bifurcations demarcate qualitatively different modules or elementary behaviors, which are the units of analysis of the dynamical systems approach. Because these units reflect both the structure of the neural or behavioral system and that of its environment, the dynamical systems approach may play a natural role in the current research program aimed at understanding how cognition is embodied by systems whose structure reflects functional adaptation to an environment and a specific developmental and evolutionary history.

See also: Motor Control; Neural Systems: Models of Behavioral Functions; Self-organizing Dynamical Systems; Systems Modeling

Bibliography

Amari S 1977 Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics* **27**: 77–87

Erlhagen W, Bastian A, Jancke D, Richle A, Schöner G 1999 The distribution of neuronal population activation (DPA) as a tool to study interaction and integration in cortical representations. *Journal of Neuroscience Methods* **94**: 53–66

Georgopoulos A P, Schwartz A B, Kettner R E 1986 Neural population coding of movement direction. *Science* **233**: 1416–19

Giese M A 1999 *Dynamic Neural Field Theory of Motion Perception*. Kluwer, Dordrecht

Grossberg S 1980 Biological competition: Decision rules, pattern formation, and oscillations. *Proceedings of the National Academy of Sciences* (*USA*) **77**: 2338–42

Haken H 1983 *Synergetics: An Introduction*, 3rd edn. Springer Verlag, Berlin

Ottes F P, van Gisbergen J A M, Eggermont J J 1984 Metrics of saccade responses to visual double stimuli: Two different modes. *Vision Research* **24**: 1169–79

Port T, van Gelder R (eds.) 1995 *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, MA

Reichardt W, Poggio T 1976 Visual control of orientation behavior in the fly: I. A quantitative analysis. *Quarterly Reviews in Biophysics* **9**: 311–75

Saltzman E, Kelso J A S 1987 Skilled actions: A task dynamic approach. *Psychological Review* **94**: 84–106

Schöner G, Kelso J A S 1988 Dynamic pattern generation in behavioral and neural systems. *Science* **239**: 1513–20

Schöner G, Kopecz K, Erlhagen W 1997 The dynamic neural field theory of motor programming: Arm and eye movements. In: Morasso P G, Sanguineti V (eds.) *Self-Organization, Computational Maps and Motor Control*, Psychology Series, Vol. 119. Elsevier–North Holland, Amsterdam, pp. 271–310

Schöner G, Dose M, Engels C 1995 Dynamics of behavior: Theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems* **16**: 213–45

Thelen T, Schöner C, Scheier C, Smith L B (2001) The dynamics of embodiment: A field theory of infant perseverative reaching. *Brain and Behavioral Sciences* **24**: 1–33

Thelen E, Smith L B 1994 *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA

G. Schöner

Neural Systems: Models of Behavioral Functions

The nervous system plays a role in guiding most animal and human behavior. Extensive data has been gathered on certain aspects of behavior, including perception, learning and memory, and motor function, but theories of these behaviors are not sufficiently constrained by behavioral data alone. Research on the mechanisms of behavior can only converge to a unique solution if theories include anatomical and physiological constraints on the substrates of behavior. Extensive data has demonstrated the physiological and anatomical properties of individual neurons and large-scale brain structures. However, our understanding of specific behaviors in terms of the complex interactions within neural networks remains incomplete. The development of theoretical models and computer simulations has greatly assisted research on the neural basis of behavior.

Neural models can be designed with different levels of physiological and anatomical detail. Less detailed models place less demand on computing resources, and have greater flexibility for rapidly addressing a range of behavioral tasks. However, these less detailed models do not address various constraints available from physiological and anatomical data, and therefore may be less accurate in their simulation of the actual function of neural systems. Examples of some of these types of models are presented in *Learning and Memory: Computational Models*, and *Cognitive Functions (Normal) and Neuropsychological Deficits, Models of*. This article will focus on more detailed models of the dynamics of individual brain regions and neural circuits involved in specific behaviors—models which draw explicitly on available data from neuroanatomy and neurophysiology. These studies all fall within the field of computational neuroscience (see *Computational Neuroscience*). At this level of analysis, considerable work has focused on specific brain regions, such as the primary visual cortex or the hippocampus. The relative focus on these particular regions arises from the more extensive experimental data available in these regions.

1. Neural Models of Visual Cortex and Perception

Extensive anatomical and physiological research on the primary visual cortex in cat and monkey provides excellent data to constrain models addressing a range of different functional properties of primary visual cortex. This work has contributed important theoretical concepts to the empirical field—including the concept of self-organization of feature detectors and topographic maps. Self-organization refers to a mechanism in which a population of units with random initial connectivity compete among themselves to represent salient features of the input patterns. This results from synaptic learning rules which contain some dependence upon presynaptic and postsynaptic activity, combined with some mechanism for competition between the growth of different synapses (such that growth in one set of inputs results in weakening of another set of inputs). This basic principle has been used in numerous models of the development of neurons selective to specific stimulus features. In addition, numerous models have addressed the development of the functional topography of primary visual cortex, including the mapping of adjacent locations on the retina to adjacent locations

in the primary visual cortex (retinotopy), and the distribution of ocular dominance and orientation selectivity (see review in Erwin et al. 1995). Predictions from these models have led to specific experiments on the role of neural activity in cortical development, and on the specific properties of the long-term modification of synaptic strength. Other models of primary visual cortex have addressed issues such as the mechanisms for direction selectivity of motion-sensitive neurons (Somers et al. 1995) and the physiological basis of psychophysical phenomena such as illusory contours (Grossberg and Williamson 2001).

Less data are available on the anatomical structure of regions of the cortex involved in the processing of more complex visual features. However, physiological data demonstrate stimulus selective properties of some of these regions, which have been analyzed in neural models. For example, the task of reaching for an object guided by vision cannot be done on the basis of where the object falls on the retina alone. Computing the correct reaching direction requires a complex computation involving awareness of eye direction and head direction. A neural network given input combining retinal position with eye position develops units with selectivity for specific combinations, which corresponds to unit selectivity of neurons in the parietal cortex, and provides enhanced understanding of how these representations are essential to guiding behavior (Mazzoni et al. 1991, Pouget et al. 1999).

2. *Neural Models of Hippocampus and Prefrontal Cortex in Memory*

The extensive physiological, anatomical, and behavioral data on rat hippocampal formation has led to a similar focus of computational modeling on this structure. Some detailed models of hippocampal function address the network basis of physiological phenomena, without focusing on behavioral manifestations of these phenomena (Traub et al. 1992). However, numerous groups have developed detailed functional models of memory encoding and retrieval within the hippocampal formation, with explicit functional roles assigned to the dentate gyrus, region CA3, and region CA1 (Marr 1971, McNaughton and Morris 1987, Treves and Rolls 1994, O'Reilly and McClelland 1994, Hasselmo and Wyble 1997; see review in Hasselmo and McClelland 1999). These models utilize several common features which stand as important theoretical contributions to the understanding of hippocampal function.

One key idea which has emerged from this work is the idea that the overlap of the patterns of neural activity representing different inputs must be minimized in order to avoid interference between stored patterns; many models assign this function of pattern separation to the dentate gyrus. Another common principle in models of the hippocampus concerns the use of excitatory recurrent connections in region CA3 to encode different components of an episodic memory. In most of these models, associations between different elements of an episode involve activity co-occurring during a short time window in different populations of neurons in region CA3. For example, in learning an arbitrary association between two words in a paired associate memory task (e.g., dishtowel–locomotive), the co-occurrence of presynaptic activity in one population with postsynaptic activity in another population results in strengthening of the excitatory connection between the two populations. Subsequently, activation of one population by a cue word (e.g., dishtowel) causes activity to spread across the strengthened excitatory connection to activate the second population (representing the word 'locomotive')—allowing retrieval of the episodic association.

Modeling these associative properties of region CA3 demonstrated the dynamical requirement that the excitatory afferent input should be stronger during encoding, whereas the excitatory recurrent connections should be stronger during retrieval. Competing models have obtained these dynamics through the use of different fiber systems (Treves and Rolls 1994) or modulatory regulation of synaptic strength (Hasselmo and Wyble 1997). Models have also generated functional hypotheses concerning the convergence of input from region CA3 and the entorhinal cortex on region CA1. This convergence could allow retrieval from region CA3 to be compared with current sensory input and goal representations from the entorhinal cortex to determine whether the retrieval is valid on the basis of its match with the cue information. Many of the initial models of hippocampal memory function focus on static patterns of activity representing locations or individual memories, but newer models focus on encoding and retrieval of sequences of activity (Levy 1996).

Physiological data and lesion data suggest that associations initially encoded in the hippocampus gradually guide changes in semantic representations in the neocortex. Computational modeling demonstrates a functional necessity for this two-stage process (McClelland et al. 1995), illustrating that episodic memory can be readily formed with a single learning event, whereas semantic memory requires interleaved activation of multiple different episodic representations in order to prevent new information from distorting previous semantic representations.

Numerous cognitive processes require maintenance of information in working memory. For example, performance of simple delayed matching to sample tasks requires mechanisms for maintaining information during the delay period. This maintenance of information could be mediated by sustained neural activity within cortical structures, and neural models have demonstrated mechanisms for this sustained neural activity on a network level—using excitatory

feedback (Lisman et al. 1998), and on a single neuron level—using self-sustaining membrane currents (Fransen et al. 1999).

Detailed physiological models have also been developed on other aspects of learning and memory function for which extensive physiological and anatomical data have become available. For example, detailed neural models of the role of the cerebellum in classical conditioning have been developed (Medina and Mauk 2000), and other neural models of conditioning are presented in *Learning and Memory: Computational Models* and *Learning and Memory, Neural Basis of*. Modeling has also elucidated mechanisms for operant conditioning, in which the activity of dopamine neurons originally triggered by reward can shift to respond to earlier stimuli which are predictive of that reward. This process provides mechanisms for selective strengthening of connections mediating goal-directed behavior (Schultz et al. 1997).

3. *Models of Small Neural Circuits*

Another approach to understanding the neural basis of behavior focuses on simpler neural circuits in nonmammalian preparations. Nonmammalian preparations often have smaller numbers of neurons than mammalian structures, and sometimes these neurons are more accessible to detailed physiological recording than those in mammalian preparations. The smaller number of neurons allows the creation of models incorporating specific features of individual neurons, and analysis of the dynamical interactions of these modeled neurons (Calabrese and De Schutter 1992). Nonmammalian preparations have numerous neural specializations that may result in fundamental differences between the neural mechanisms in these animals compared to mammals. However, important convergent properties have been described in other species, including principles of self-regulation, in which individual neurons adapt their cellular properties so that they maintain a similar level of activity, even after the circuit interactions have been perturbed (Golowasch et al. 1999). With regard to motor function, these smaller scale and more extensively described neural circuits provide the opportunity for more detailed analysis of the neural circuits underlying the generation of specific motor patterns (Cohen et al. 1992, Calabrese and De Schutter 1992).

4. *Models which Include Ongoing Interaction with the Environment*

The models described above provide effective mappings of data on input parameters to specific behavioral output, but they do not include an alteration of the environmental input which depends upon the current output of the model. This can only be obtained with explicit robotic implementations, or with representations of a virtual organism moving through a virtual environment. Ultimately, accurate representation of the temporally continuous interaction with the environment during behavior will require models with environments influenced by the model output.

A few examples have explicitly demonstrated the ongoing interaction of a model with its environment. For example, modeling of spatial navigation requires an ongoing interaction of the model with the environment, as the movement of an animal through the environment under the control of the model will result in constant updates of its position. Several simulations have demonstrated how the place cell representations formed in the hippocampal formation can be used to guide movement through an environment (Sharp et al. 1996, Burgess et al. 1997, Redish and Touretzky 1998). These models provide excellent examples of the necessity for a constant interactive updating of parameters including the sensory input from the environment, and the proprioceptive knowledge of current heading (head direction). Neural models also demonstrate mechanisms for maintenance of this proprioceptive information, using attractor dynamics within regions encoding the head direction (Sharp et al. 1996).

The production of complex sounds or movements requires constant monitoring of feedback from the sensory periphery and from the environment in order to monitor the appropriate process of production. Sophisticated models have moved beyond a simple feed-forward output and have addressed the continuous dynamical interaction of motor output and sensory input (Troyer and Doupe 2000).

5. *Summary*

In the realm of experimental data, there continues to be a gap between the knowledge of anatomical and physiological properties in neural systems and the knowledge of how behavior could be mediated at a neural level. Neural models of behavior will provide an essential theoretical link between these different levels of analysis—similar to the role of theoretical physics in understanding the properties of physical systems. Neural models of behavior have advanced dramatically in recent years, both in their level of anatomical and physiological detail, and in their capacity to flexibly address a broader scope of behavioral data. Computational models have already provided important theoretical concepts to the empirical work. As empirical techniques become more sophisticated for analyzing complex systems, computational models will become increasingly important in understanding the theoretical principles involved in the neural basis of behavior.

See also: Artificial Neural Networks: Neurocomputation; Hippocampus and Related Structures; Learning and Memory: Computational Models; Neural

Networks: Biological Models and Applications; Neural Systems and Behavior: Dynamical Systems Approaches; Prefrontal Cortex; Visual System in the Brain; Visual Systems: Dorsal and Ventral

Bibliography

Burgess N, Donnett J G, Jeffery K J, O'Keefe J 1997 Robotic and neuronal simulation of the hippocampus and rat navigation. *Philosophical Transactions of the Royal Society of London Series B—Biological Sciences* **352**(1360): 1535–43

Calabrese R L, De Schutter E 1992 Motor-pattern-generating networks in invertebrates: Modeling our way toward understanding. *Trends in Neuroscience* **15**(11): 439–45

Cohen A H, Ermentrout G B, Kiemel T, Kopell N, Sigvardt K A, Williams T L 1992 Modelling of intersegmental coordination in the lamprey central pattern generator for locomotion. *Trends in Neuroscience* **15**(11): 434–8

Erwin E, Obermayer K, Schulten K 1995 Models of orientation and ocular dominance columns in the visual cortex: A critical comparison. *Neural Computation* **7**(3): 425–68

Fransen E, Wallenstein G V, Alonso A, Dickson C T, Hasselmo M E 1999 A biophysical simulation of intrinsic and network properties of entorhinal cortex. *Neurocomputing* **26–27**: 375–80

Golowasch J, Casey M, Abbott L F, Marder E 1999 Network stability from activity-dependent regulation of neuronal conductances. *Neural Computation* **11**: 1079–96

Grossberg S, Williamson J R 2001 A neural model of how horizontal and interlaminar connections of visual cortex develop into adult circuits that carry out perceptual grouping and learning. *Cerebral Cortex* **11**(1): 37–58

Hasselmo M E, McClelland J L 1999 Neural models of memory. *Current Opinion in Neurobiology* **9**: 184–8

Hasselmo M E, Wyble B P 1997 Simulation of the effects of scopolamine on free recall and recognition in a network model of the hippocampus. *Behavioral Brain Research* **89**: 1–34

Levy W B 1996 A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus* **6**: 579–90

Lisman J E, Fellous J M, Wang X J 1998 A role for NMDA-receptor channels in working memory. *Nature Neuroscience* **1**(4): 273–5

Mazzoni P, Andersen R A, Jordan M I 1991 A more biologically plausible learning rule than back propagation applied to a network model of cortical area 7a. *Cerebral Cortex* **1**(4): 293–307

McClelland J L, McNaughton B L, O'Reilly R C 1995 Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**: 419–57

McNaughton B L, Morris R G M 1987 Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neuroscience* **10**: 408–15

Medina J F, Mauk M D 2000 Computer simulation of cerebellar information processing. *Nature Neuroscience* **3**: 1205–11

O'Reilly R C, McClelland J L 1994 Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus* **4**(6): 661–82

Pouget A, Deneve S, Sejnowski T J 1999 Frames of reference in hemineglect: A computational approach. *Progress in Brain Research* **121**: 81–97

Redish A D, Touretzky D S 1998 The role of the hippocampus in solving the Morris water maze. *Neural Computation* **10**(1): 73–111

Schultz W, Dayan P, Montague P R 1997 A neural substrate of prediction and reward. *Science* **275**: 1593–9

Sharp P E, Blair H T, Brown M 1996 Neural network modeling of the hippocampal formation spatial signals and their possible role in navigation: A modular approach. *Hippocampus* **6**(6): 720–34

Somers D C, Nelson S B, Sur M 1995 An emergent model of orientation selectivity in cat visual cortical simple cells. *Journal of Neuroscience* **15**(8): 5448–65

Traub R D, Miles R, Buzsaki G 1992 Computer simulation of carbachol-driven rhythmic population oscillations in the CA3 region of the in vitro rat hippocampus. *Journal of Physiology (London)* **451**: 653–72

Treves A, Rolls E T 1994 Computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**(3): 374–91

Troyer T W, Doupe A J 2000 An associational model of birdsong sensorimotor learning I. Efference copy and the learning of song syllables. *Journal of Neurophysiology* **84**(3): 1204–23

M. E. Hasselmo

Neuroanatomical Methods

Neuroanatomy investigates neural structure, with the motivation of deducing and formulating general organizational principles, mechanisms, and structural–functional correlations. It includes several subdisciplines, which deal with normal organization, experimentally or pathologically induced changes, developmental processes, or comparative species differences.

Data analysis relies heavily on visualization, at either light or electron microscopic magnifications (respectively, up to $\times 1000$ or $\times 100\,000$). Sometimes, exciting observations can result from 'just looking.' One example could be the discovery of the 'barrel field' of rodent somatosensory cortex, where there is a one-to-one mapping of snout whiskers and cortical cell groups ('barrels'). Much more typically, however, the relationship between structure and function is not straightforward. Only after years of work with combined anatomical and other techniques, can moderately good understanding be achieved; for example, how neuron geometry influences the conduction of the nerve impulse. For questions as complicated as 'what are the neural substrates of higher order complex behaviors,' our understanding is still at an early stage.

Because of the need for microscopic magnifications, neuroanatomical techniques are often invasive and require chemical fixation of brain tissue, thin sectioning via microtome, and histological processing for various markers. Recently, non-invasive imaging techniques such as functional magnetic resonance (fMRI)

are becoming more commonly available. Used in combination with sophisticated stimuli, these provide unprecedented visualization of functional networks in the human brain.

In this article, we give a brief overview of basic techniques that are used to investigate neural structure at different levels of organization. The first five sections cover single neurons, neuron populations, connections, microcircuitry, and functional groups. Each section includes some treatment of historical developments, examples of applications, and comments on limitations. The sixth section is on neuroanatomy of the human brain, and the last section reviews some of the most recent technical developments. For further details, we refer the interested reader to the more specialized works listed in the references.

1. Nerve Cell

The structure of nerve cells ('neurons') is complex. There is the cell body or soma and, for most neurons, two types of specialized processes (Fig 1). Dendrites (Greek *dendron*, tree) are thin-caliber extensions of the cell body that receive input, via synaptic contacts, from other neurons. The dendritic tree or arbor is confined mainly to the immediate vicinity of the cell body and, depending on the type of neuron, does not extend beyond 0.25–2.00 mm. The axon (Greek *axon*, axis) is a single, thinner filament-like process, specialized for conduction of the nerve impulse. It conveys output from a given neuron by making synaptic contact with hundreds or even thousands of other neurons. For some neurons (interneurons or local circuit neurons), the axon arbor is small and about the same diameter as the dendritic tree ($\sim$ 0.25–0.50 mm). For projection neurons, however, the axon can be quite long, in the range of many centimeters. Examples of long distance projections are axons crossing between the two cerebral hemispheres in the corpus callosum, and axons originating from neurons in motor cortex and projecting into the spinal cord.

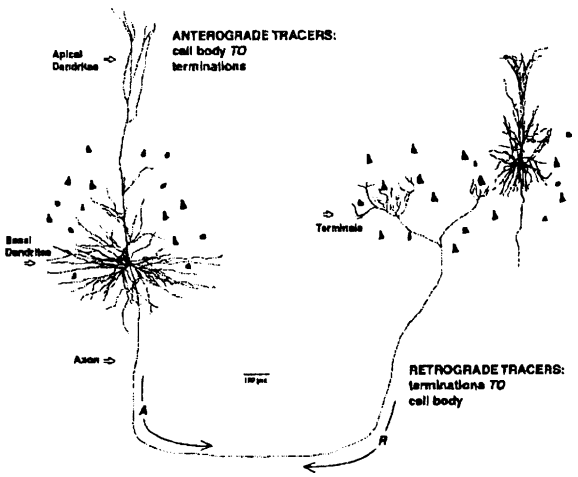

Figure 1
This schematic diagram illustrates the main parts of a representative neuron (the neocortical pyramidal cell). Two neurons are shown in Golgi-like detail, including dendrites, axons, and terminals (only part of the descending axon is drawn for the neuron on the right). The two neurons are shown together with more numerous cell bodies (triangles and ovals), such as would be seen in a Nissl stain. Note the divergence of the terminal arbor, which makes synaptic contact with hundreds of other neurons. This diagram also illustrates how experimental tracer substances are transported anterogradely (A) or retrogradely (R) along the axon

Understanding even the basic structure of the nerve cell was long hindered by the delicacy of brain tissue and the complex shapes of the dendritic and axonal specializations. Early histologicial techniques could not distinguish the relationship of the soma, dendrites, and axons; and one interpretation, current in the late nineteenth century, was that nerve 'cells' were nodes, perhaps serving a nutritive function, embedded within a continuous reticulum of nerve fibers (Van der Loos 1967).

The first routine demonstration of nerve cell morphology was almost accidentally achieved by a silver dichromate stain, where small blocks of tissue (not more than 5–10 mm in one dimension) are successively immersed in solutions of potassium dichromate and silver nitrate. The procedure was originally intended as a way to preserve and harden nerve tissue so that it could be thinly sliced and viewed under the light microscope. Unexpectedly, owing to the argyllophilic properties of nerve cells, the procedure also resulted in the deposition of black silver particles throughout the soma and dendritic arbor. This is the 'Golgi stain,' discovered by Camillo Golgi in 1873. It was used to spectacular advantage over the next decades by both Golgi and Ramon y Cajal, who shared the Nobel Prize in 1906 for their major contributions to functional neuroanatomy.

The mechanism of the Golgi stain is to this day not well understood, as only a small proportion of nerve cells are silver-impregnated, and which cells are stained is not predictable. Axons are even less reliably stained. They cannot be visualized at all with this method once they become myelinated, and any Golgi study of axons must be carried out on young brains.

Since its discovery, the Golgi stain has remained important in investigations of normative structure and of structural changes associated with development, pathology, or plasticity. It is also useful for investigations of neuronal structure in postmortem human brains. It has, however, otherwise been superceded by more controllable techniques.

Figure 2
An electron micrograph of a calbindin-positive dendrite ('D') in the rat hippocampus. The dendrite is labeled with black particles, after immunocytochemical processing for calbindin. It is surrounded by multiple synaptic contacts (two of which are indicated as 'S')

One newer technique is intracellular injections of tracers (such as biocytin, neurobiotin, or horseradish peroxidase). This method often requires electrophysiological localization of the target neuron, and thus gives precise data about the physiological response properties of a neuron in correlation with its morphology. The soma and entire dendritic arbor are demonstrated, and at least the proximal portion of the axon. Intracellular filling can be achieved either *in vivo* or *in vitro*, and has been a basic tool in functional classifications of neuron types; for example, the different subcategories of bipolar, amacrine, and ganglion cells in the retina. The principal limitation of this technique is the difficulty in making up a large sample size.

For fine cellular and subcellular details, the high magnification of electron microscopy is required (see Peters et al. 1991, Guillery 2000).

2. *Neuronal Populations*

Nissl stains (named for the histologist Franz Nissl) bind to acidic elements within the cell, and reveal cell bodies and sometimes the proximal dendrites. They reveal all the cells in a given piece of tissue, including the glia. They are used to help demarcate subcortical nuclei and cortical areas. They are also commonly used for deriving cell counts ('stereology').

Early investigations of brain structure relied extensively on Nissl stains, sometimes in conjunction with the complementary myelin stains. Differences in cell density and size formed the basis for the 'architectonic' maps of cerebral cortex, constructed by Brodmann, O. and C. Vogt, Von Economo, and other workers in the first part of the twentieth century (see Braak 1980). These parcellations have been subjected to considerable dispute and even ridicule; but have

gained renewed interest and relevance, especially because of the focal activation patterns resulting from functional imaging experiments (fMRI and PET).

Another, newer method for revealing specific subpopulations is immunocytochemistry. There is now a large assortment of commercially available antibodies; for example, against structural proteins, such as neurofilaments or microtubule-associated-proteins; against transmitters (glutamate; GABA; dopamine) or neuropeptides and calcium-binding proteins (somatostatin, calbindin, parvalbumin, calretinin); against synaptic components (synaptophysin, synapsin); against receptors (glutamate; GABA). Immunocytochemisty involves a series of steps, where thinly cut tissue sections ($< 50\,\mu m$) are transferred through solutions containing different concentrations of a primary antibody and then a secondary antibody (for example, anti-GABA made in rabbit, and anti-rabbit made in goat, mouse or some other host animal). The secondary antibody might carry a fluorescent tag, so that the end result is visible by fluorescent microscopy. Since a fluorescent signal tends to fade, another procedure uses a secondary antibody tagged with biotin. After several other steps, this results in a brown or black reaction product that labels the original protein.

Histochemistry is a somewhat related technique, where various chromogen markers (frequently diaminobenzidene dihydrochloride, DAB) are deposited as the end product of an enzymatic reaction, and thus reveal cells and/or processes positive for certain substances. Acetylcholinesterase reactions, for example, have been used to chart neurons and/or terminations positive for the neurotransmitter acetylcholine. (Antibodies against acetylcholine are now more commonly used for this purpose.) Cytochrome oxidase reactions show neurons and/or processes with high levels of this mitochondrial enzyme, which is part of the Krebs cycle and therefore an indicator for high levels of metabolic activity.

3. *Connectivity*

Neurons receive connections from and project to hundreds or thousands of other neurons; and a key step in understanding neural function is characterizing these connections. A surprising number of connections were correctly inferred by close analysis of Golgi-stained tissue, by Cajal and Golgi. Other early observations came from pathological conditions or 'experiments of nature.' For example, neuron death in motor cortex after large lesions or strokes results in degenerated corticospinal axons. This could be detected by myelin stains in postmortem histological preparations and used to correctly infer the subcortical pathways. Or localized lesions could be made in experimental animals. In some conditions, this would result in cell shrinkage or cell loss (retrograde or 'Wallerian' degeneration) in other interconnected structures; for example, thalamic neurons might degenerate if their cortical targets were destroyed. Because neurons often send axon collaterals to two or more structures, however, degeneration would not always result, and this method was notoriously difficult to interpret.

The first controlled method for investigating long-distance connections again capitalized on the argyllophilic properties of neurons. Degenerating fibers were produced in animal models by making small lesions in particular structures by suction or heat. (These and other experimental procedures are performed with animals that have been deeply and humanely anesthetized, and are carefully carried out under sterile conditions. Protocols are always subject to institutional approval and monitoring.) Without the nutritive support from the parent cells, over a short time, there is large-scale degeneration of axons originating from the lesioned area. After the brain is perfused with fixatives and thinly sectioned ($< 50\,\mu m$), the tissue is reacted in a series of solutions. Degenerating axons are selectively seen as silver blackened particles, and can be followed, under the light microscope, from the lesion site to its target structures.

In the 1970s, the lesion-degeneration method was replaced by several much superior techniques, which were based on the physiological processes of axon transport (Fig 1). Anterograde tracers, such as ^{3}H-amino acids and WGA-HRP were injected in small, microliter amounts, taken-up by cell bodies and dendrites, incorporated into various protein compartments, and transported down the axon to its distal terminations. Retrograde tracers, such as horseradish peroxidase (HRP), were similarly injected in microliter amounts, taken up by nerve terminals, and transported 'upstream' through an axon to the cell body. Anterogradely labeled terminations are detected as brown or blue dots (WGA-HRP reacted histochemically with DAB or TMB as chromogen); or, in autoradiography (photographic film or emulsion developed by radioactive tritium), as black silver grains. Retrogradely labeled cell bodies appear colored as brown, black, or blue after histochemical processing. These are both sensitive, reliable techniques, which are used to determine the projections originating *from* a given structure (anterograde tracers: structure A $\rightarrow$ targets x,y,z) and those structures which project *to* a given region (retrograde tracers: structures x,y,z $\rightarrow$ target A). Additional information is provided concerning the locations, number, and sometimes type of parent neurons (retrograde tracers) and their terminations (anterograde tracers).

HRP and WGA-HRP are still used in connectivity studies, along with several other newer retrograde tracers (fluorescent dyes, such as 'fast blue' and

'diamidino yellow,' and a modified form of cholera toxin). Autoradiography has itself been largely replaced by other anterogradely transported tracers such as biocytin, kidney bean lectin, biotinylated dextran amino, or fluoro-ruby. These newer anterograde tracers are non-radioactive and therefore easier to use. (They are visualized by a final histochemical reaction with DAB or other chromogen, after histological reactions with a complex called avidin-biotinylated HRP.) They also have a higher resolution, producing a Golgi-like image of the entire axon and its terminations (in contrast with the cloud silver grains in autoradiography).

These techniques have been used to chart the basic connectivity ('wiring diagram') of different brain regions and functional pathways (see also, Kobbert et al. 2000). The dataset is still incomplete, however, and very little is known about connectional interactions, either structurally or functionally. Another problem is that the available techniques have a necessary but artificial duality implied by the paradigm of 'injection sites' and 'projection zones.' This tends to overlook the possible importance of indirect pathways and loops.

4. *Microcircuitry*

Knowing the overall connectivity (inputs and outputs) of different brain structures is important, but this needs to be followed-up by other studies to define the functional microcircuitry at the synaptic level. In some structures—such as the retina, cerebellum, spinal cord, sensory thalamic nuclei, basal ganglia—there is a considerable amount of detailed information about synaptic types, and postsynaptic locations; but in many structures, such as primate association cortex, this information is very incomplete.

Investigations at the level of microcircuity are difficult for several reasons. First, there must be a way of identifying presynaptic inputs and their postsynaptic cellular targets. Even now, despite the availability of a much larger choice of labels, experiments are usually limited to two or at most three labels. Second, microcircuity depends on a confident identification of synaptic contacts. What look like close contacts between terminal specialization and neurons at the light microscopic level, frequently are not actual contacts. These need to be verified at the higher magnifications of electron microscopy or, with special controls, with confocal microscopy. These methods are extremely labor intensive and slow. Third, in many structures, the circuitry may be dynamic rather than fixed. A typical neocortical neuron, for example, has been estimated to receive 6000 inputs. The combination and integration of this connectional set can differ, depending on factors such as internal state, external context, and the nature of a stimulus. A common distinction is between 'driving' and 'modulatory' inputs. In some instances, these can be distinguished on the basis of synaptic location and/or size: large synapses located on or near the cell soma are likely to be more 'effective' than small synapses located at dendritic tips. There are, however, many exceptions to this principle, and considerable further work is necessary. Combined anatomical–physiological studies in tissue slices are one current approach to structural–functional characteristics at the microcircuitry level.

5. *Functional Networks*

There are several anatomical techniques for visualizing the activity of large populations of neurons. In animal models, one of the most direct methods is 2-deoxyglucose (2-DG) autoradiography. In this technique, anesthetized animals are given defined stimuli, such as monocular visual stimulations, repeated somatosensory stimulations, or pure auditory tones. Neurons that are selectively sensitive to these stimuli are expected to require higher amounts of glucose to support their higher activity. If the subject is intravenously injected with radioactive deoxyglucose, the label should be taken up by the more responsive neurons. The network of responsive neurons is visualized after histological processing. This technique was developed by L. Sokoloff and collaborators in the late 1970s, before the use of PET scans or fMRI. It offers excellent resolution and also is compatible with other experimental manipulations.

Increasingly, however, 2-DG has been replaced by other activity markers. Early immediate genes like c-fos, jun, and zipf268 can be 'turned on' by defined stimuli (Morgan and Curran 1995). The responding neurons can be visualized by *in situ* hybridization for the mRNA, or by immunocytochemistry against the protein products. As compared with 2-DG, these newer methods have better resolution at the cellular level; and for immunocytochemistry, they do not involve radioactive tags. Another advantage is that *in situ* hybridization can be used together with immunocytochemistry to compare the responses to two different events at different times (respectively, very short for mRNA expression, and slightly longer for protein expression). The disadvantage of these methods, which they share with other activity-related markers, is the necessity for rigorous controls that the signal is in fact linked to the specific stimulation paradigm and not activated by some other side condition.

At longer time scales, of days or months, many other substances can change in response to environmental or other manipulations. These include cytochrome oxidase, a mitochondrial enzyme associated with metabolic demand; the inhibitory transmitter GABA; and various receptor molecules. These changes are assayed by histological processing and compared with baseline levels in normal controls.

In this section, it is worth mentioning as a comparison, multiple electrode arrays. These arrays are another means of investigating functional networks, although they are still limited in terms of how many neurons can be monitored (< 100). In contrast with more anatomical techniques, they have the advantage of providing physiological response properties, but are so far less advantageous for localization.

6. Neuroanatomy in Humans

Because so much of neuroanatomy requires invasive techniques and terminal experiments, most data have been derived from animal models. More direct research on the human brain would be preferable, given the likely species differences even between humans and subhuman primates, and the specialization of the human brain for language, abstract calculation, and other higher cognitive functions.

Recently, the development of noninvasive fMRI imaging has allowed rapid progress in the investigation of functional networks in the human brain (see Frackowiak et al. 1997). These experiments reveal brain areas activated or deactivated by controlled stimuli. Typically, the results are tabulated as lists, however; and the interpretation of network relationships and anatomical substrates is still difficult, depending, for validation, on extrapolations from animal experiments.

Standard connectivity experiments, which require injection of tracer substances into brain tissue and perfusion of the brain for histological processing, are obviously not possible. Two modifications have been used with limited success. One is to stain for degenerating fibers (following the early lesion-degeneration techniques) in postmortem brains that had sustained localized damage at some time before death (e.g., by stroke or loss of vision). The callosal connections between cortical visual areas have been studied in this way, and the resulting patterns validated to some extent by comparisons with animal models (Zilles and Clarke 1997).

A second tracer that can be used in postmortem tissue is carbocyanine dye, which is lipophilic and will be transported by diffusion along fiber pathways. DiI has been used in several studies to demonstrate intrinsic connections. An interhemispheric asymmetry has recently been suggested in the columnar structure of left and right human temporal cortex, on the basis of differential patterns of DiI-labeled intrinsic connections (Galuske et al. 2000, Gazzaniga 2000). Unfortunately, DiI travels over only a relatively short distance (about 5.0 mm), and even this takes weeks or months. Thus, unless some modification can be found, it is not suited for investigation of the long extrinsic connections.

Golgi stains can be applied to human tissue; but, as in animal brains, these work in a somewhat arbitrary fashion, and, in particular, are not satisfactory for studies of long distance connections.

In contrast with the severe limitations of connectivity experiments, immunocytochemical techniques have been applied to human tissue with considerable success. Antibodies against neurofilaments, neurotransmitters, and many receptors have been used almost as routinely as in animal preparations, although the rather long postmortem delays before tissue fixation (3–18 h) are not optimal.

EM has also been carried out in postmortem human tissue or biopsy specimens. There is the same problem, of suboptimal tissue preservation, but comparisons with perfused animal preparations have provided validation of results.

7. Modern Techniques

There have been several spectacular advances in recent years, first, in instrumentation. Laser scanning confocal microscope (LSCM), by greatly improving the resolution of images stained with fluorescent dyes, has positively impacted on immunohistochemistry, *in situ* hybridization histochemistry, intracellular labeling, and tract tracing. The remarkable theoretical and technical progress in this field is based on an insight of Marvin Minsky on the confocal optical system in 1957 (for an overview of confocal microscopy, see Paddock 1999, Pawley 1995). In brief, signals are acquired from tissue sections from a focally illuminated scanning plane, through small pinhole apertures positioned on the microscope axis. Using lasers as a light source, sharp images can be obtained and processed by appropriate image-analyzer software for LSCM. This results in the clear demonstration of neuronal morphology, including such fine structures as dendritic spines and synaptic contacts, even though actual synaptic contacts should be confirmed electron microscopically. Reconstructed 3-dimensional and multiple color images from LSCM data are now routinely used.

Another recent development is the 2-photon laser scanning microscope, which utilizes near infrared laser sources with long wavelengths. Deep penetration of light into the tissue and low toxicity of lighting, are the main advantages in 2-photon microscopy.

There has also been continued progress with neuroanatomical techniques, in part, prompted by applications of molecular biology. *In situ* hybridization histochemistry, for example, detects histologically the expression of genes for functional and compositional molecules in neural tissue. The phenomenon was described by Spiegelman and his collaborators in 1964, i.e., hybridization, between matched nucleotide sequences of dissociated single strands of DNA and RNA in *in vitro* experiments; and the methods have been rapidly developed since then. Detection probes for mRNAs, fragments of complementary oligonucleotides (cDNAs or cRNAs), are constructed first.

These are labeled with radioisotopes (^{32}P, ^{35}S and ^{3}H) or non-radioactive molecules, such as biotin and digoxigenin, which are 'reporters' that can describe precisely the location of hybridized nucleotides in question. Hybridized signals in cell bodies are examined with a conventional light or fluorescent microscope. For example, gene expression for receptor subunits of the inhibitory neurotransmitter γ-aminobutyric acid (GABA), is detected in almost all neurons. By combining *in situ* hybridization with immunohistochemistry, simultaneous detection of a particular gene expression and protein products of other genes in the same tissue can be achieved.

Gene manipulation has also contributed substantially to the development of new types of cell labeling and tract tracing techniques. Genes of extrinsic origin that do not exist in an experimental animal can be used as 'reporters' when they are manipulated so as to be co-expressed with genes that are expressed in that animal. Green fluorescent protein (GFP), whose gene was originally cloned from a jellyfish (*Aequorea victoria*), is available for such a purpose (Praser et al. 1992). Reporter genes can be introduced at an appropriate site of a nucleotide sequence, for example, near a promoter region that is specific for certain types of cells, so that only a targeted subpopulation of cells is labeled with GFP. This technique has been modified to develop an effective transneuronal tract tracing method (Yoshihara et al. 1999). In this modification, not the GFP gene, but the gene for a plant lectin WGA (wheat germ agglutinin) was introduced in transgenic mice with promotor genes (e.g., cerebellar Purkinje cell specific L7). Gene-transferred Purkinje cells produce WGA in the cell bodies, and transport it through the axons to reach target cells. Target cells in the second and third order relay structures are labeled trans-synaptically one after another through a series of neural pathways. This technique is applicable for an analysis of selective connectivity of a particular type of neurons, and for actually visualizing chains of interconnected neurons.

In addition to reporter gene manipulations described here, various virus vectors (adenovirus, pseudorabies, herpes simplex virus etc.) are also available for introduction of extrinsic genes (Yuste et al. 2000). With these new methods, it is likely that tract tracing for targeted cells will be refined in considerable detail.

See also: Cerebral Cortex: Organization and Function; Synapse Ultrastructure

Bibliography

Bassel A, Hayashi M, Spiegelman S 1964 The enzymatic synthesis of a circular DNA–RNA hybrid. *Proceedings of Natural Academy of Science USA* **52**: 796–804

Braak H 1980 *Architectonics of the Human Telencephalic Cortex*. Springer, Berlin

Frackowiak R S J, Friston K J, Firth C D, Dolan R J, Mazziotta J C 1997 *Human Brain Function*. Academic Press, San Diego, CA

Galuske R A W, Schlote W, Bratzke H, Singer W 2000 Interhemispheric asymmetries of the modular structure in human temporal cortex. *Science* **289**: 1946–9

Gazzaniga M S 2000 Regional differences in cortical organization. *Science* **289**: 1887–8

Guillery R W 2000 Early electron microscopic observations of synaptic structures in the cerebral cortex: A view of the contributions made by George Gray. *Trends in Neuroscience* **23**: 594–8

Kobbert C, Apps R, Bechmann I, Lanciego L, Mey J, Thanos S 2000 Current concepts in neuroanatomical tracing. *Progress in Neurobiology* **62**: 327–51

Morgan J I, Curran T 1995 Immediate—early genes: Ten years on. *Trends in Neuroscience* **18**: 66–7

Paddock S W 1999 *Methods in Molecular Biology, Confocal Microscopy Methods and Protocols*. Humana Press, Totowa, NJ, Vol. 122

Pawley J B 1995 *Handbook of Biological Confocal Microscopy*, 2nd edn. Plenum Press, New York

Peters A, Palay S L, de Webster H 1991 *The Fine Structure of the Nervous System*, 3rd edn. Oxford University Press, New York

Prasher D C, Eckerode V K, Ward W W, Prendergast F G, Cormier M J 1992 Primary structure of the Aequorea victoria green-fluorescent protein. *Gene* **111**: 229–33

Van der Loos H 1967 The history of the neuron. In: Hyden H (ed.) *The Neuron*. Elsevier, Amsterdam

Yoshihara Y, Mizuno T, Nakahira M, Kawasaki M, Watanabe Y, Kagamiyama H, Jishage K, Ueda O, Suzuki H, Tabuchi K, Sawamoto K, Okano H, Noda T, Mori K 1999 A genetic approach to visualization of multisynaptic neural pathways using plant lectin transgene. *Neuron* **22**: 33–41

Yuste R, Lanni F, Konnerth A (eds.) 2000 *Imaging Neurons*. Cold Spring Harbor Press, Cold Spring Harbor, NY

Zilles K, Clarke S 1997 Architecture, connectivity, and transmitter receptors of human extrastriate visual cortex: Comparison with nonhuman primates. In: Rockland K S, Kaas J H, Peters A (eds.) *Cerebral Cortex:Extrastriate Cortex in Primates 12*. Plenum Press, New York, pp. 637–742

K. S. Rockland and T. Hashikawa

Neuroendocrinology

Neuroendocrinology is the study of interactions between the nervous and endocrine systems. The nervous system is composed of the brain, spinal cord, ganglia, and nerves. Neural cells communicate directly with one another (and with cells of sensory and effector tissues) by means of neurotransmitters. The endocrine system is composed of ductless glands that release hormones that act systemically. Neuroendocrinology therefore investigates reciprocal influences of local and widespread systems for signaling in animals. This

article will outline the progression of research in neuroendocrinology from the analysis of regulation of the pituitary as an output of the nervous system to various influences of hormones on the nervous system.

1. Definition of 'Neuroendocrinology'

The term 'neuroendocrinology' is composed of several morphemes. The prefix 'neuro-' is derived from words for 'nerve' or 'sinew,' and so refers to a nervous system with physical connections between the parts. The term 'endocrine' is a combination of 'endo-,' meaning 'within' and '-crine,' from a Greek word for 'separate.' Therefore, the endocrine system is composed of glands that are separate from the tissues they influence within the body. The suffix '-ology,' of course, signifies a body of knowledge. Therefore, the most literal meaning of neuroendocrinology is the study of the interactions between the nervous and the endocrine systems.

It should be emphasized that the nervous and endocrine systems share common signaling functions, in that the cells of both systems release chemical messengers that bind to sensitive cells. The binding of the chemical messengers initiates a response only in target cells having the appropriate receptor. In the case of the nervous system, the chemical messenger is a neurotransmitter that is released and acts primarily at an area of close cellular contact, a synapse. The cells of the nervous system are localized in centralized structures such as the brain and spinal cord and make numerous synapses with one another and with other cells by means of extended cellular processes (axons or dendrites). In the case of the endocrine system, hormones are released from the cells of a gland and are carried through the circulatory system to act on cells throughout the body. It has become increasingly clear that the fields of neuroscience and endocrinology are not so distinct as once thought, and there are now known to be many overlapping characteristics and functions of neural and endocrine systems.

2. Historical Development of Research in Neuroendocrinology

Neuroendocrine research has progressed from an original focus on the endocrine system as a specialized output of the nervous system to a consideration of the complex feedback influences of endocrine glands on the nervous system. A primary emphasis of neuroendocrinology has been an analysis of the regulatory interplay between the hypothalamus and pituitary, and glands in other places of the body. Additionally, neuroendocrinology has investigated the direct neuronal control of glands other than the pituitary. The field has further expanded to include hormonal influences on brain behavioral systems.

2.1 Neurosecretion

For many years, studies on invertebrates have suggested that neuronal cells secrete substances systemically (Speidel 1919). The hormonal function of neural cells was termed 'neurosecretion' (Scharrer and Scharrer 1954), and the idea of neurosecretory cells acting as an interface between the nervous and endocrine systems gave rise to the field of neuroendocrinology. Lesions of the median eminence of the hypothalamus were found to influence the activity of the pituitary gland profoundly, and the effects of the lesions were mimicked by blocking the blood flow from the hypothalamus to the pituitary (see Harris 1972). These results suggested that hypothalamic neurons secrete factors directly into the blood, and are therefore neurosecretory.

Many hormonal factors have now been isolated from the hypothalamus and it has become clear that most of these factors are short peptides (see Akil et al. 1999, Frohman et al. 1999). Characterization of the release of neuropeptides became an early major focus of neuroendocrinology (see Table 1). There are two distinct systems for neuropeptide secretion related to pituitary function. The posterior lobe of the pituitary (or neurohypophysis) consists of terminals of neurosecretory cells with somas in the hypothalamus (see Akil et al. 1999, Frohman et al. 1999). The secretions of the posterior pituitary are neuropeptides that enter the circulation. The anterior lobe of the pituitary (or adenohypophysis) is an endocrine gland. Neuropeptides from the median eminence are carried through the portal blood supply to the anterior pituitary and control the release of anterior pituitary hormones. In turn, several of the anterior pituitary hormones regulate the functions of glands in other parts of the body. Such glands are said to form part of a hypothalamo-pituitary axis of hormonal regulation, as in, for example, the 'hypothalamo-pituitary-thyroid axis.'

2.2 Sympathetic Nervous System

The output of the nervous system is typically described as having a somatic portion, controlling voluntary muscle contractions, and an autonomic division, regulating activity of the viscera (see Dodd and Role 1991). Activation of the sympathetic nervous system due to injury or perceived threat causes effects classified as a 'fight-or-flight' response. These effects include an increased heart rate, respiration, blood pressure, and glucose mobilization. Axons from the sympathetic division of the autonomic nervous system exit the spinal cord and synapse in a chain of ganglia alongside the vertebrae or in prevertebral ganglia closer to the tissue innervated. The postganglionic cells release norepinephrine at synapses in smooth muscle, cardiac muscle, or exocrine glands. Additionally, some fibers

Table 1
Summary of roles of some neurosecretory peptides in relation to pituitary function

Hypothalamic Neuropeptide(s)	Pituitary Hormones(s)	Lobe	Peripheral Target Tissue	Process Regulated at Peripheral Target Tissue	Process Regulated by Peripheral Hormone
prolactin inhibitory factor (PIF), dopamine	*prolaction*	ant.	mammary glands	lactation	
growth hormone-releasing hormone	*growth hormone* (GH or *somatotropin*)	ant.	various	growth	
(GHRH); *somatostatin* (SS or *SRIH*)			liver	secretion of *insulin-like growth factor-I* (IGF-I)	growth
thyrotropin-releasing hormone (TRH)	*thyroid-stimulating hormone* (TSH or thyrotropin)	ant.	*thyroid gland*	secretion of thyroid hormones	basal metabolic rate; development
corticotropin-releasing hormone (CRH)	*adrenocorticotropic hormone* (*ACTC* or *corticotropin*)	ant.	*adrenal* cortex	secretion of adrenal steroids	reaction to stress; electrolyte balance
gonadotropin-releasing hormone (GnRH)	*gonadotropins*: *luteinizing hormone* (LH); *follicle-stimulating hormone* (FSH)	ant.	gonads	gamete formation; secretion of sex steroids	reproduction
melanocyte-stimulating hormone releasing factor (MRF); inhibiting factor (MIF)	*melanocyte-stimulating hormone* (MSH);	inter-med.	melano-cytes	skin pigmentation	
antidiuretic hormone (ADH or *vasopressin*)	same	post.	kidneys	fluid retention	
oxytocin	same	post.	mammary glands, uterus	*milk letdown*; labor	

from the spinal cord directly innervate the adrenal medulla through the greater splanchnic nerve. The endocrine cells of the adrenal medulla secrete epinephrine and norepinephrine systemically, much as the postganglionic sympathetic neurons release norepinephrine at synapses.

2.3 Temporal Aspects of Neural Regulation of the Endocrine System

A common feature of neurosecretion is the striking rhythmicity of the release of hormone (see Akil et al. 1999, Frohman et al. 1999). Levels of circulating hormones controlled by a hypothalamo-pituitary regulation typically fluctuate with a daily, or 'circadian,' rhythm of approximately 24 hours. Superimposed on the circadian rhythm is a more rapid 'ultradian' pattern of discrete pulses of hormone, which is also regular in rhythm. The period of the ultradian secretion of hormone varies by species. In humans, the typical pattern of secretion of anterior pituitary hormones is a series of spikes at 90–120 minute intervals, with the maximal concentration of the spikes varying systematically according to the time of day. The temporal pattern of pulsatile and circadian secretions of hormones is thought to be the result of the pattern of activity of hypothalamic neurosecretory cells. A major function of the brain in regulating the endocrine system may therefore be in the timing of hormonal fluctuations. The suprachiasmatic nucleus of the hypothalamus has a clock-like function in the

brain, and there are probably other circadian and ultradian clocks in the brain.

The pineal gland has the unique role of integrating sensory and endocrine functions. In lower vertebrates the pineal is actually a photosensitive third eye (Dodt and Meiss 1982). In mammals, the photoreceptive capacity seems to have been lost, but the role of the pineal is nonetheless linked to photic input. The innervation of the mammalian pineal includes multisynaptic inputs from the retina, especially by way of the suprachiasmatic nucleus of the hypothalamus, and the hormone of the pineal, melatonin, is secreted predominantly at night.

2.4 *Feedback Regulation of the Nervous System by Peripheral Hormones*

Neuroendocrinology is not only the study of endocrine outputs from the nervous system, but also examines the effects of peripheral hormones on the brain. A central theme of neuroendocrinology is the importance of negative feedback processes, wherein an output of a system inhibits the controller of the system. In most of the hypothalamo-pituitary axes, there is a demonstrated effect of the hormone of the peripheral gland on the neurosecretory cells in the hypothalamus (see Akil et al. 1999, Frohman et al. 1999). Growth hormone (GH), for example, inhibits the release of growth hormone-releasing hormone (GHRH) and increases release of somatostatin. Because of the feedback of GH on the hypothalamus, secretion of GH decreases as levels of the hormone become elevated, and the circulating levels of GH are controlled. In an analogous fashion, thyrotrophin-releasing hormone (TRH) release is very sensitive to circulating unbound levels of thyroid hormone. The feedback of sex steroids on the hypothalamic secretion of gonadotrophin-releasing hormone (GnRH) is extraordinarily complex, and shifts from inhibitory to stimulatory just before ovulation in females (see Fink 1979).

2.5 *Effects of Hormones on the Developing Brain*

A major effect of the endocrine system on the brain is seen during development. Thyroid hormones have long been known to have an essential role in brain development, and hypothyroidism during development can lead to serious disturbances, such as cretinism. These effects are reflected in structural deficits in certain types of brain neurons (see Oppenheimer and Schwartz 1997). Perinatal steroids, primarily androgens, play a key role in the determination of gender differences in brain structure and function (see Cooke et al. 1998).

2.6 *Regulation of Behavior by Hormones*

It has become increasingly clear that hormones influence a variety of brain functions in adulthood, and behavioral neuroendocrinology has become a fertile area of research. Reproductive behaviors of both genders are influenced by gonadal steroids, due to direct effects on the brain (see Beach 1967). Thyroid hormone effects in adult brain are illustrated by a variety of clinical signs and symptoms of thyroid disease, and there are a variety of indications that thyroid hormones are important in the etiology of depression (see Henley and Koehnle 1997). Peptide hormones including corticotrophin-releasing factor (CRF), GHRH, SS, GnRH, arginine vasopressin, oxytocin, vasotocin, and adrenocorticotropic hormone (ACTH) can influence behaviors including feeding, reproductive behavior, and behavioral responses to stress (see Koob 1992, Le Moal et al. 1992).

3. *Recent Trends of Neuroendocrine Research*

3.1 *Novel Mechanisms of Action of Hormones in the Brain*

In peripheral tissues and in the brain, a well-defined mechanism of action of steroid and thyroid hormones is through a family of nuclear receptors that alter gene expression. Recent studies have indicated that several steroids can affect binding and activity membrane receptors for neurotransmitters such as γ-aminobutyric acid (GABA) (see McEwen 1998). Thyroid hormones have also been shown to have effects at membrane receptors for GABA (Martin et al. 1996). A current focus of neuroendocrine research is the functional significance of multiple distinct mechanisms of brain actions of a given hormone in the brain.

3.2 *Convergence of Research on Molecular Signals*

Gradually, distinctions between hormones and neurotransmitters have eroded. Some neuroactive steroids are known to be synthesized from cholesterol in the brain and may have important direct signaling functions within brain tissue (see McEwen 1998). Several lines of evidence suggest that thyroid hormones may also have neurotransmitter-like functions in the brain (Dratman and Gordon 1996, Martin et al. 1996).

Neuropeptides that had been originally studied in relation to pituitary function are now known to be present in a variety of brain areas (Mains and Eipper 1998). Many other neuropeptides originally classified as gastric hormones are also found in brain areas other than the hypothalamus. These neuropeptides probably

do not enter the circulation as hormones but may be released locally and act within brain tissue, like neurotransmitters. These neuropeptides appear to modulate the actions of other, more typical, neurotransmitters acting on the same cell and are sometimes termed 'neuromodulators.'

Cytokines are peptides, including interleukins and tumor necrosis factors, which have important roles in immune response. These molecules are also found in the brain and may have neuromodulatory roles (Vitkovic et al. 2000). A rapidly developing area of research is the role of cytokines in the regulation of sleep (Krueger et al. 1999). It seems likely that other signaling factors with actions that have been characterized extensively in one system will turn out to have wide-ranging effects in neural tissue.

4. *Future Directions of Neuroendocrine Research*

During the history of neuroendocrinology, a host of factors have been found to carry signals from the brain to the endocrine system and from the endocrine system to the brain. The process of identifying and cataloging factors and their receptors has been facilitated by molecular biological techniques, and seems likely to accelerate rapidly when data from human and other genome sequences becomes widely available. As the list of signaling factors and their receptors approaches completion, neuroendocrine research should turn to the complex issue of how the influences of these factors could be integrated within various tissues. For example, one probable direction of future research is the analysis of the mechanisms and functional roles of the pulsatile and circadian rhythms of hormonal secretions. A surprising number of the signaling factors have similar roles in nervous, endocrine, immune, and other systems. Future researchers may eventually study chemical signals in biological systems as a single unified discipline, classifying signals and receptors by their structure, gene family, or presumed evolutionary origins, instead of according to the secretory or target tissues.

See also: Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of; Hypothalamus; Sex Hormones and their Brain Receptors

Bibliography

Akil H, Campeau S, Cullinan W E, Lechan R M, Toni R, Watson S J, Moore R E 1999 Neuroendocrine systems I: Overview: thyroid and adrenal axes. In: Kandel E R, Schwartz J H, Jessell T M (eds.) *Fundamental Neuroscience*. Academic Press, San Diego, CA

Beach F A 1967 Cerebral and hormonal control of reflexive mechanisms involved in copulatory behavior. *Physiological Reviews* **47**: 289–316

Cooke B, Hegstrom C D, Villeneuve L S, Breedlove S M 1998 Sexual differentiation of the vertebrate brain: principles and mechanisms. *Frontiers in Neuroendocrinology* **19**: 323–62

Dodd J, Role L W 1991 The sympathetic nervous system. In: Kandel E R, Schwartz J H, Jessell T M (eds.) *Principles of Neural Science*. Appleton and Lang, Norwalk, CT

Dodt E, Meissl H 1982 The pineal and parietal organs of lower vertebrates. *Experientia* **38**: 996–1000

Dratman M B, Gordon J T 1996 Thyroid hormones as neurotransmitters. *Thyroid* **6**: 639–47

Fink G 1979 Feedback action of target hormones on the hypothalamus and pituitary with special reference to gonadal steroids. *Annual Review of Physiology* **41**: 571–80

Frohman L, Cameron J, Wise P 1999 Neuroendocrine systems II: Growth, reproduction, and lactation. In: Kandel E R, Schwartz J H, Jessell T M (eds.) *Fundamental Neuroscience*. Academic Press, San Diego, CA

Harris G W 1972 Humours and hormones. *Journal of Endocrinology* **53**: 2–23

Henley W N, Koehnle T J 1997 Thyroid hormones and the treatment of depression: An examination of basic hormonal actions in the mature mammalian brain. *Synapse* **27**: 36–44

Koob G F 1992 The behavioral neuroendocrinology of corticotropin-releasing factor, growth hormone-releasing factor, somatostatin, and gonadotropin-releasing hormone. In: Nemeroff C B (ed.) *Neuroendocrinology*. CRC Press, Boca Raton, FL

Krueger J M, Obal F, Fang J D 1999 Humoral regulation of physiological sleep: cytokines and GHRH. *Journal of Sleep Research* **8**: 53–9

Le Moal M, Mormede P, Stinus P 1992 The behavioral endocrinology of arginine vasopressin, adrenocorticotropic hormone and opioids. In: Nemeroff C B (ed.) *Neuroendocrinology*. CRC Press, Ann Arbor, MI

Mains R E, Eipper B A 1998 Peptides. In: Siegel G J, Agranoff B W, Albers R W, Fisher S K, Uhler M D (eds.) *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*. Lippincott Williams & Wilkins, Philadelphia

Martin J V, Williams D B, Fitzgerald R M, Im H K, VonVoigtlander P F 1996 Thyroid hormonal modulation of the binding and activity of the GABAA receptor complex of brain. *Neuroscience* **73**: 705–13

McEwen B S 1998 Endocrine effects on the brain and their relationship to behavior. In: Siegel G J, Agranoff B W, Albers R W, Fisher S W, Uhler M D (eds.) *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*. Lippincott Williams and Wilkins, Philadelphia, PA

Oppenheimer J H, Schwartz H L 1997 Molecular basis of thyroid hormone-dependent brain development. *Endocrine Reviews* **18**: 462–75

Scharrer E, Scharrer B 1954 Hormones produced by neurosecretory cells. *Recent Progress in Hormone Research* **10**: 183–232

Speidel C C 1919 Gland-cells of internal secretion in the spinal cord of the skates. *Carnegie Institute of Washington Publications* **13**: 1–31

Vitkovic L, Bockaert J, Jacque C 2000 'Inflammatory' cytokines: Neuromodulators in normal brain? *Journal of Neurochemistry* **74**: 457–71

J. V. Martin

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Neurogenetics and Behavior

'Neurogenetics' refers to the study of the development and/or function of the nervous system using identifiable gene variants. The nervous system comprises the peripheral sense organs, the sets of excitable and supporting cells by which these communicate with the spinal cord and brain, the spinal cord and brain themselves (collectively called the central nervous system or CNS), and the sets of excitable and/or secretory cells by which the CNS communicates with muscles, tissues, and organs. The nervous system is where evolution, development, and individual experience combine to produce behaviors at the level of a whole organism. Thus, neurogenetics includes the study of the behavioral consequences of gene variants that affect nervous system development and/or function.

1. Historical Varieties of Behavioral Neurogenetic Studies

There are three identifiably different paradigms that emerged in prior behavioral neurogenetic work. The first documented use of 'neurogenetic' was by the World Federation of Neurology, which hosted a series of international congresses 'for Neuro-genetics and Neuro-ophthalmology' in the 1960s. Here, as in most present-day use by the medical research community, the term refers to discovering relationships between gene sequence variants and 'inherited' neurological diseases:

> Just as our understanding of pathology transformed neurology in the nineteenth century, new genetic knowledge has permeated many aspects of neurology today. Indeed, genetic methods have changed our basic understanding of neurologic disease processes, and the definition of diseases themselves has undergone changes as DNA based testing is substituted for clinical, morphological or biochemical definitions as a gold standard for analysis (Pulst 2000, p. ix).

Although this medical usage of 'neurogenetic' is fairly recent, the work it refers to has continuity with pre-1940s research in both biology and medicine. The prominent involvement of hereditarian views (and prewar researchers such as Ernst Rudin, a pioneer of genetic psychiatry) in eugenics movements and Nazi racial and genocidal policies made the subject of genetics, the nervous system, and behavior a rather uneasy one in medical circles for a period of time after 1945. Neurogenetic approaches to behavioral pathologies experienced a resurgence in the late 1950s and early 1960s, and developed an individual identity within the larger study of hereditary disorders.

A second neurogenetic paradigm was articulated outside of medicine in the late 1940s when the developmental biologist Paul Weiss (1898–1989) organized a meeting for the new discipline of 'Genetic Neurology.' In the preface to an edited volume of the same name, he explained the purview of this field:

> The days of March 21–25, 1949 … may at some future time be remembered as the birth date of a new biological discipline—'genetic neurology'—birthdate in the sense that an organism, which had for some time been in the making, came to light, gave evidence of its vitality, acquired a name, and set out for a career of its own. Neuroanatomy, neurophysiology, and neuropsychology deal essentially with the nervous system in its final mature state. Genetic neurology encompasses all those processes that lead up to that state ('neurogenesis'), maintain it in its integrity, and restore it after disruption. The attribute 'genetic,' accordingly, refers to 'genesis,' i.e., development in the widest sense, and not to 'genetics' in the narrower modern sense of 'inheritance,' although it naturally includes the latter (Weiss 1950, p. v).

This developmental paradigm has a different emphasis from the 'medical' one: genes are not unifying explanations, disease predictors, or gold standards for classification, but rather tools for learning about development. Behavior becomes incorporated in the 'developmental' paradigm in the following way. Since all genetic changes to the nervous system do not necessarily lead to readily observable anatomical, biochemical, or physiological differences in individuals, it is necessary to have some way of assessing 'invisible' aspects of neural function. Behavior provides one convenient assay of this type, making it possible to identify subtle changes in development and/or ongoing function of the nervous system correlated with genetic changes. Even when behavioral assays are used, the goal of the developmental agenda is typically to understand the developmental processes particular gene products participate in rather than the genesis of the 'assay' behaviors themselves.

The third, 'behavioral' paradigm resembles the first, and shares a common history with it. Here the focus is on understanding the contribution of genetic differences to non-pathological individual differences in behavior:

> More recently, it has become increasingly accepted that genetic variation makes an important contribution to differences among individuals in the normal range of behavior as well as for abnormal behavior. Moreover, many behavioral disorders, especially common ones, may represent variation at the extremes of the same genetic and environmental factors that are responsible for variation within the normal range (Plomin 1999).

As in the 'medical' paradigm, the emphasis of the 'behavioral' paradigm is on genes as unifying explanations for individual differences, as predictors for such differences, and as potential avenues for manipulating the characteristics of individuals. In spite of their procedural and explanatory similarities, it is

important to note that the 'medical' and 'behavioral' paradigms have different aims. The 'medical' paradigm is fundamentally about understanding and alleviating dysfunction. The 'behavioral' paradigm is about explaining, predicting (and potentially manipulating) any differences in behavior between individuals or groups. These both contrast with but do not necessarily conflict with the 'developmental' paradigm, which is about understanding developmental processes within individuals that have an impact on behaviors.

2. *Behavior as an Object of Neurogenetic Studies*

Whatever their aims and assumptions, all neurogenetic paradigms use behavior as a measurable object to which neurobiological and genetic variation can be related. However, behavioral variation does not manifest itself in predefined units that have a direct relationship with underlying biological processes. Researchers must make decisions about how they are going to classify and measure behavior in order to make it suitable for subsequent biological analyses. Approaches to behavioral biology that grew out of different research traditions (such as medicine, experimental psychology, and zoology) have divided behavioral variation differently to suit their particular interests, and these categorization schemes continue to change over time. The most important attributes of behavioral categorization with respect to biology are whether the behavioral phenomena of interest can be measured directly or not, and whether these phenomena are measured at the level of individuals or groups of individuals.

2.1 *Direct Output vs. Inferential Linkage*

Two broadly defined methods of objectifying behavior for biological studies can be identified. The first, direct output, grew out of an interest in mechanistically explaining externally observable behavioral acts. What events go on in a brain when a particular behavior such as a yawn or the production of fluent speech is generated? Notice that such acts can either be narrowly or loosely defined (a particular type of yawn versus all yawning, or the generation of a particular word or class of words versus any vocal emission), so that these measures can either deal with individual acts or some population of acts. The thing to be explained is always the thing that is physically being measured.

Yet there are many behavioral phenomena that are more abstract than a concrete motor act, for instance, behavioral states such as hunger and thirst, or emotional states such as happiness or sadness, or finding stimuli pleasant or distasteful. While it is sometimes possible to find a combination of externally measurable variables that will uniquely characterize a state (as in the case of sleep), it is more usual that there is no combination of behavioral measures that will mark one of these phenomena off from all others. Rather, it is a combination of both a subject's behavior and the context in which that behavior occurs that indirectly defines this type of behavioral variation.

The second method for objectifying behavior, inferential linkage, takes its starting point from these facts. Investigators have an underlying model (not always made explicit) that defines the relationship between externally observable variables and the underlying phenomena of interest. For instance, operant conditioning paradigms use motor responses like bar pressing to study learning in non-human animals. A neurogeneticist who adopts this paradigm is not interested in the neurobiology of bar pressing *per se*, but in brain processes that subserve learning (particularly the influence of prior history on the brain, and what happens in the brain up to the time a decision is made to press or not press the bar). Implicitly or explicitly, such an investigator has a schema for how these learning processes are linked to bar pressing, allowing bar pressing to be a proxy for learning. The same applies to measurements of human behavior from written or oral questionnaires or tests. The thing that is to be explained is not what is directly measured. Another way of stating this is that there are intervening variables between the object being measured and the object of interest.

The distinction between these two methods is important because of the readiness with which biological variables can be causally related to behaviors. Neurogenetic studies using direct outputs are studying gene effects on the object of interest. Unfortunately, there are many behavioral phenomena that cannot be studied this way. Studies using inferential linkages are thus a necessity, but they do not provide an immediate and direct causal link between the behavioral phenomenon of interest and biological variables. Biological effects on processes that intervene between the object of study and what is actually measured must first be ruled out.

2.2 *Individual Behavior vs. Group Behavior*

Most behavioral scientists are used to thinking of behavior as a product of a single individual's nervous system that can be validly measured and analyzed within that individual. For instance, one can measure the extent to which a behavioral trait occurs repeatedly in the same individual by presenting him or her with the same context multiple times, and using this set of measures as a basis for documenting differences in behavior among different individuals. In this case, the performance of each individual is preserved and individual differences can be directly assessed.

However, for the purpose of assessing population expectations in behavioral patterns, it may be desirable to measure a behavior with respect to groups rather than single individuals. In this case, individuals are classified according to some predetermined schema and the behavioral responses of different individuals in each group are measured in the same context. Individuals are typically assessed only once or a few times, and the data of interest are differences in the distribution of group scores. Individual deviations from group means (or from the grand mean of all groups) have been suggested as measures indicating individual differences. Such group measures have been very popular in neurogenetic studies, where individuals can be classified according to the identity of particular genetic variants to look for behavioral differences, or sorted according to behavior to look for genetic differences.

From the point of view of biology, individual and group measures are not equivalent. The vast majority of behaviors that are studied biologically are not collectively produced by the linked physiological and biochemical activity of many nervous systems simultaneously. Neurogeneticists who are interested in disease, in development, or in the causation of behavior all need to find out how gene variants in individual bodies make differences in developmental decisions. While accepting that there are many arenas that group measures of behavior are good and appropriate for, their application to neurogenetic research is limited. They can be useful in determining whether a correlation exists between variation in a particular gene sequence and behavioral variation at a population level, but they cannot be directly converted into the currency of developmental events that go on in individuals.

2.3 Limits on Behavioral Measurement Constrain Causal Attribution

To many scientists and non-scientists alike, it may appear that behaviors are relatively simple and straightforward to measure, while neurochemistry, neuroanatomy, gene sequences, and gene expression patterns are vastly more complicated. It is frequently argued that causal attributions in neurogenetic work are limited by an ability to quantify genetic or neurobiological variation, or limitations in the statistical techniques used to relate behavioral variation to neurobiology and genetics. In reality, just the opposite is the case. Measures of neurobiological and genetic variables are typically subjected to closer critical scrutiny than behavioral measures. They are typically also direct measures. Since so many of the behaviors people study neurogenetically can only be measured via inferential linkages, the characteristics of the behavioral variables themselves are typically the factors which most constrain causal attributions.

One example is provided by work on rodent learning that uses the amount of time it takes animals to find a platform hidden under the surface of a deep pool of water as a behavioral variable. During some number of training sessions, individual animals can presumably use a constellation of spatial cues in the room surrounding the pool to memorize where the platform is. However, there are many other factors that will determine whether animals can utilize these training opportunities in the same way, and also determine the speed with which an animal familiarized with this situation will find the platform.

To take a hypothetical illustration, animals which are more prone to panic attacks may find being tossed into a pool of water quite stressful, and may spend a good deal of time swimming around the edges of the pool trying to escape. Thus, they may not have the same learning opportunities as calmer animals, and, even if this difference in opportunity to learn is controlled for, they may still be slower because it takes them a while to calm down enough and remember that there is a platform to be looked for. Other animals may have motor problems with coordinated swimming and be hesitant to leave the side of the pool or simply be slower swimmers. There are many reasonable scenarios by which animals whose genotypes differ may show reliable differences in this behavioral task, all unrelated to their spatial learning ability. Researchers must conduct numerous tests to rule out these alternative explanations before accepting that a neurobiological or genetic linkage is with learning ability rather than with stress, motor coordination, or a host of other possible explanations.

Because there is currently no systematic way of advancing reasonable alternative explanations, or of deciding how many potential alternatives there are, this open-ended process can take a long time. When one is finally sure of a linkage between a behavioral process and genetic and/or neurobiological variation, one must then make sure that individual differences in behavior can be sufficiently well quantified to enable the biological linkages between genes, neurobiology, and behavior to be adequately explored.

3. Current Trends in Neurogenetic Studies

Study designs commonly utilized in neurogenetics can be divided according to whether they involve correlations between behavior and genetic variation based on classifications of subjects into groups, or study of the behavioral outcome of controlled genetic and/or neurobiological manipulations. Correlative studies can be further subdivided by whether they classify subjects by gene variants and examine behavioral differences, or whether they classify subjects by behavioral variants and examine genetic differences. Manipulative studies can directly alter single genes (gene deletions, gene insertions, and changes in gene

expression), groups of genes (inactivation or loss of parts of, or whole, chromosomes), and/or manipulations on brain cell groups (substitutions, deletions, and chemical alterations of function).

3.1 Correlative Studies: Candidate Behaviors and Candidate Genes

Correlative designs currently dominate neurogenetic studies in both humans and animals. In the first of these study designs, individuals are classified according to some behavioral attribute and the genomes of individuals in the different groups are searched for sequence variation that correlates with group membership. For example, individuals are administered a psychological or IQ test, and a breakpoint is determined whereby the subjects are divided into 'high-score' and 'low-score' groups. The DNA sequences of group members are subsequently searched for sequence variants that statistically differ according to group membership. The sequence variants can either be part of a DNA sequence coding for a gene product, part of a DNA sequence that either controls the expression of a nearby gene or contributes to the local 3-dimensional geometry of a stretch of DNA, or part of a sequence that has no identifiable function. A common variant of this procedure is to take extremes of these two groups, say subjects who scored in the upper and lower 10% of the test distribution, and compare their DNA.

These approaches are sometimes referred to as 'quantitative trait loci' (QTL) approaches because when one takes group measures of the behavioral performances one can generate estimates of the amount of behavioral variation in the population that is removed ('explained') by variation in particular gene sequences. Such estimates are not necessarily very informative: different genes may affect behavioral traits through the same biological pathways, so one does not expect the variance contributions of all genes to add to 100%. The point of these approaches is not to generate tallies of explained variance, but rather to discover genes that putatively influence the behavioral trait of interest.

A more classical research design used in animal studies is to induce genetic changes in reproductive cells of parents, mate them, and use a particular behavior as a 'screen' for animals containing gene sequence changes that affect the criterion behavior. While formerly these approaches relied upon chemical mutagens or radiation-induced damage to chromosomes, it is now possible via molecular biological techniques to target changes to particular genes, insert these into cells in embryos, and implant the embryos in foster mothers ('transgenic' techniques). When such a design is applied with a particular behavior as the object of study, it becomes a variant of the 'candidate behavior' approach.

The second type of correlative study is quite literally the inverse of the first design. Individuals are classified by which sequence variants of particular genes they possess, and are measured for one or a collection of behavioral attributes. Group behavioral measures are then screened for covariation with group membership. In this method one typically knows something about the molecule whose gene sequence is used for classification, and where in the brain this molecule is typically found. Educated guesses can be made about the kinds of developmental and cellular processes this molecule might participate in, and hence the range of behaviors one might expect to be affected by structural variants in the molecule. This approach is potentially capable of providing faster reliable links between biology and behavior than the candidate behavior approach.

3.2 Caveats Governing the Interpretation of Correlative Studies

Correlative studies have an appealingly simple design that belies the complexity of their interpretation. Three major interpretational difficulties are:

(a) The illusion of homogeneity in group composition. Correlation designs can give the illusion of a 'one-to-one' correspondence between trait and behavioral variation when this may not be the case. Because one has chosen variation in a limited number of gene sequences or behavioral traits as a basis for group classification, this does not mean that groups could not also systematically differ in other gene sequences and other behavioral traits. This problem would cause correlative studies to produce spurious positive correlations between a particular genetic variant and a particular behavioral variant. The same problem could also cause investigators to underestimate the complexity of the genetic system contributing to behavioral variation significantly, since an effect due to multiple co-varying genes would be attributed to a single gene. The candidate behavior approach is particularly prone to underestimating the complexity of a gene's effects in development via some fundamental cellular process, because it focuses on linkages between a single gene and a single behavior. This may cause investigators to misconstrue seriously the developmental and functional pathways that a particular gene contributes to, if they fail to arrive at a true accounting of the multiple co-varying behavioral systems a gene product may affect.

(b) Non-genetic explanations for the correlation. There may also be correlations between non-genetic biological or experiential factors and group membership. It is frequently the case that correlational designs are applied to groups of subjects with uncontrolled (and unknown) prior developmental histories. Thus, while there may indeed be correlations between genetic and behavioral variables, the genetic

variation may be causally unrelated to the behavioral variation.

(c) Misidentification/mislabeling of the behavioral dimension that co-varies with gene sequences. As described in Sect. 2.3, when using behavioral measures that depend on inferential linkages in a candidate behavior approach, it is possible to misidentify the domain specificity of a behavioral effect completely if only one or a small number of similar behavioral measures are employed. This misidentification has serious consequences in a biological research program, where domain assignments (for instance, 'learning' as versus 'fear responses') have implications for which specific brain regions and biochemical pathways one implicates in a behavioral difference, based on the historical body of neurobiological work.

Correlational studies generate possible linkages between particular genes and particular behaviors, but cannot be said to constitute a demonstration of such a linkage. Because they are the only neurogenetic studies possible in humans, they will undoubtedly remain a staple of research in the medical and behavioral paradigms. It is important that their conclusions be constrained by their limitations.

3.3 Manipulative Studies in Genes and Cells and their Limitations

Manipulative studies are only possible in the domain of animal work. They are the best means available for definitively linking changes in particular biological pathways with particular behavioral differences. Manipulative studies are especially important for mapping out the full range of biological processes that particular gene variants participate in, which is essential for understanding the linkage between gene variation and behavioral variation.

The most basic level at which such manipulations can be made is to single genes. As discussed in Sect. 3.1, it is now possible to make controlled changes to single genes and place these altered genes into animals. In addition to using this approach to correlate gene changes simply with behavioral changes, one can also systematically alter the structure of molecules, the timing of when they are produced, and the places where they are produced in an animal's nervous system. Such manipulations make it possible to move into the realm of identifying and characterizing the fine structure of the developmental and cellular processes in which gene products participate.

There have traditionally been genetic techniques available in animal species such as fruit flies for manipulating the expression of whole groups of genes in particular cells of the nervous system. This makes it possible to produce so-called 'chimeric' animals, where some cell types express one set of genes, while other cell types express another set. Chimeras provide an important tool for examining developmental interactions at the level of cells, and techniques for producing them in both mammals and birds have made it possible to examine developmental differences in cells that have many genetic differences. In this way, complex behavioral differences that may be due to many gene changes in many different cell types, such as behavioral differences among species of animals, become amenable to empirical investigation.

Manipulative studies can suffer from the same interpretational difficulties as correlational studies. The major difference is that in manipulative studies, one has no excuse for failing to control for non-genetic explanations, for failing to measure enough different organismal attributes to avoid mislabeling or misidentification of the relevant behavioral dimensions, and for failing to look carefully enough at genetic and behavioral variables to avoid any illusions of homogeneity in characterizing genotypic or behavioral groups. There are some known technical problems with particular transgenic manipulations in rodents that can compromise causal attributions of gene effects on behavior (see Gerlai 1996); some of these problems may be remedied by future technical developments. The major limitation to manipulative studies comes from our limited ability to measure many different biological and behavioral attributes in a short period of time in single individuals. This limits our ability to understand cellular and developmental processes that are important for brain development and brain function in general, and for cells that participate in the formation and function of particular behavioral circuits.

4. Probable Future Directions

Clearly, an important development for the field of behavioral neurogenetics will occur when researchers take appropriate account of the limitations imposed by their behavioral measurements, and make research decisions and interpretations in line with these limitations. New developments in automated recording and analysis of natural behaviors will not necessarily advance this agenda, because the majority of 'behaviors' that are studied must be indirectly inferred. It is rather an increasing sophistication in defining the domain specificity (or, just as importantly, the domain generality) of particular linkages to biology that will most clearly benefit the field.

Up to the present time, behavioral neurogenetic research has been dominated by paradigms that emphasize genes as unifying explanations or predictors for behavioral dysfunction or differences. Yet the greatest untapped potential contribution of behavioral neurogenetics lies in the developmental domain. While Paul Weiss' *Genetic Neurobiology* has largely lain dormant for more than 40 years, it is starting to be

realized in the work of developmentally oriented neurogeneticists. They believe that to understand evolutionary, developmental, and functional contributions of molecules and cells to behavior, one must, as Weiss first articulated, understand the developmental processes the molecules and cells participate in, as well as those processes which 'maintain it (the nervous system) in its integrity, and restore it after disruption.'

Such 'process-oriented' views toward understanding both behavior and brain function have been slow to gain acceptance, because for a good part of this past century, scientists have thought of the brain as a stable structure with unvarying neuronal connections set down during a defined developmental period. Since the 1980s the brain has come to be appreciated as an entity in which apparent stasis at more global levels is produced by competing dynamic processes at the levels of cell groups and individual cells. This has fostered the perceived value of relating global products of nervous systems like brain structures and behaviors to the dynamic processes that give rise to and maintain them.

The change in the emphasis of behavioral neurogenetic work from finding molecules to fitting them into developmental and functional pathways also entails a shift in the technical emphasis of the 'genetic' part of the research, a change that is already underway. Investigators are beginning to focus more on where and when in nervous systems genes are expressed during development, and how gene expression in cells is modified in development and in adulthood.

4.1 Gene Expression Correlations and Manipulations

Many non-biologists envision genes purely as molecules of heredity whose reason for being disappears (everywhere but in the germ cells) once an organism develops. Yet genes are a fundamental part of the physiological machinery in cells, especially in nerve cells which must constantly manufacture the products they need to communicate with other cells, to receive communications from other cells, and to change their size, shape, and other aspects of their morphology. While analogies between nerve cells and electrical components can be terribly misleading, it would not be an error to regard nerve cells as akin to circuits containing tunable resistors, inductances, and capacitors, which enable them to alter both their sensitivity to incoming signals and the signals they send in return. To change any aspect of their physiology or morphology (changing their cellular 'behavior'), nerve cells need to alter their patterns of gene expression. They need to be able to make these changes repeatedly throughout the entire life of the individual in whose brain they are found.

Behavioral neurogeneticists are just starting to do correlative and manipulative work with gene expression patterns. Such research is likely to provide important empirical and theoretical insights in three major areas: developmental and functional pathways that behaviorally important brain circuits utilize; how external stimuli and an individual's own behavior can affect the development and function of his or her brains; and how 'inborn' behavioral influences are manifest in nervous systems and how these meld with ongoing behavioral development and function.

4.2 Developmental and Functional Pathways and Interactions

By examining gene expression pattern changes in response to manipulation of the structure or expression of 'targeted' genes, much new information should be gained about how gene products interact during development and ongoing adult function. This will facilitate the construction of functionally coherent biochemical and physiological 'pathways' that are vital for understanding cellular interactions and decision-making processes that help construct and run behavioral circuitry in the brain. Such work may also lead to an increased appreciation of the role that stochastic variation in cell interactions may play in generating behavioral variation among individuals. The 'cause' of such behavioral variety would not necessarily be identifiable with variation in any particular gene product involved in a pathway, but rather as a joint property of the structure of molecular constituents and the types of interactions in which they are involved. A better appreciation of the mechanics underlying stochastic contributions to development may also provide profound insights into dysfunction and radically alter the way we identify and assess developmental 'risk' factors.

4.3 How Stimuli and Behavior Feedback on Brain Development and Function

A second way that developmental and functional studies of gene expression may advance the field of behavioral neurogenetics is by providing a new arena in which to examine environmental and social influences on brain development. It would be especially enlightening to identify the pathways through which an organism's own behaviors may provide feedback that alters the development and function of its own brain. There has been a long and controversial history of the idea that organisms are actively involved in the regulation of their own development, just as there has been a tendency in behavioral neurogenetic work to disregard the biological pathways that mediate environmental influences on development. By being able

to study and to manipulate the gene expression responses of cells and cell groups, we have much more powerful and sensitive assays to bring to bear on these questions.

4.4 'Inborn' Influences: How Development, Evolution, and Ongoing Function are Integrated

The study of species differences has proven to be a powerful method for learning about developmental mechanisms in biology for well over a century. The application of gene expression studies (together with the other manipulative neurogenetic techniques described above) toward the problem of inborn species differences in behavior is likely to produce powerful insights into the developmental pathways that make different members of a species behaviorally similar. Of equal importance will be the light such research sheds on developmental pathways that make individuals in one species all different from members of another species in the same way.

Comparative work is also expected to provide unique information on the mechanisms evolution uses to change brains in order to change behavior. Such mechanisms are likely to involve changes to both the autonomous characteristics of cells and changes in cellular interactions. Many of these changes may depend not on changes to the structure of gene products, but rather on changes in 'silent' parts of DNA sequences that influence the context and timing of gene expression. Understanding how such changes are accomplished evolutionarily may provide new technologies for dealing with neural dysfunction caused by injury and disease.

Finally, behaviors that show inborn species differences typically also have individually distinctive components that are malleable in the lifetimes of individuals. Comparative work could produce major insights into the ways that dynamic developmental and functional processes within and between neural cells integrate development and evolution with ongoing activity.

See also: Behavioral Genetics: Psychological Perspectives; Behavioral Neuroscience; Developmental Behavioral Genetics and Education; Memory: Genetic Approaches; Memory in the Fly, Genetics of; Mental Illness, Genetics of

Bibliography

Breakefield X O (ed.) 1979 *Neurogenetics: Genetic Approaches to the Nervous System*. Elsevier, New York

Gerlai R 1996 *Seminars in Neuroscience* **8**: 153–61

Greenspan R J, Kyriacou C P (eds.) 1994 *Flexibility and Constraint in Behavioral Systems*. Wiley, New York

Hall J C, Greenspan R J, Harris W A 1982 *Genetic Neurobiology*. MIT Press, Cambridge, MA

Kelner K, Benditt J (eds.) 1994 Genes and behavior: A special report. *Science* **264**: 1685–739

Pfaff D W, Berrettini W H, Joh T H, Maxon S C (eds.) 2000 *Genetic Influences on Neural and Behavioral Functions*. CRC Press, New York

Plomin R 1999 *Nature* **402**: C25–C29 , p. C25

Pulst S M (ed.) 2000 *Neurogenetics*. Oxford University Press, Oxford, UK

Wahlster D 1999 Single gene influences on brain and behavior. *Annual Review of Psychology* **50**: 599–624

Weiss P (ed.) 1950 *Genetic Neurology*. The University of Chicago Press, Chicago

E. Balaban

Neuromuscular System

The neuromuscular system links the central nervous system to the peripheral nervous system and is composed of a neural circuit including motor neurons in the spinal cord, sensory neurons in the dorsal root ganglion, and skeletal muscle fibers (Fig. 1A). The neuromuscular system is essential to movements of the body, the control of posture and breathing. Electrical impulses triggered from the motor neuron propagate along the motor axon and result in the release of the neurotransmitter, acetylcholine, from the motor nerve terminal at the synaptic contact, called the neuromuscular junction (Fig. 1B, C). Acetylcholine molecules diffuse across a small gap and bind acetylcholine receptors on the muscle fiber. The binding results in the opening of ion channels, thereby changing the muscle membrane potential. The membrane depolarization triggers a release of internal calcium ions that leads to a cascade of events resulting in muscle contraction (Fig. 1D). The muscle spindles and Golgi tendon organs are sensory organs in the skeletal muscle that detect changes in the muscle fiber length and tension (Fig. 1E). The information is conveyed by electrical impulses propagating along sensory fibers of dorsal root ganglion neurons, which, in turn, innervate motor neurons in the spinal cord. The activation of motor neurons generates electrical impulses propagating along the motor axon and thereby completing the circuit of the knee-jerk reflex in the neuromuscular system (Fig. 1A). Motor neurons in the spinal cord also receive synaptic inputs from the brain to perform voluntary movements.

1. The Neuromuscular Junction (NMJ)

The NMJ (Fig. 1B) plays an indispensable role in transmitting information from the motor neuron to the skeletal muscle (Salpeter 1987). The NMJ, also called the endplate, is a specialized region on the muscle surface, on which the motor nerve terminal

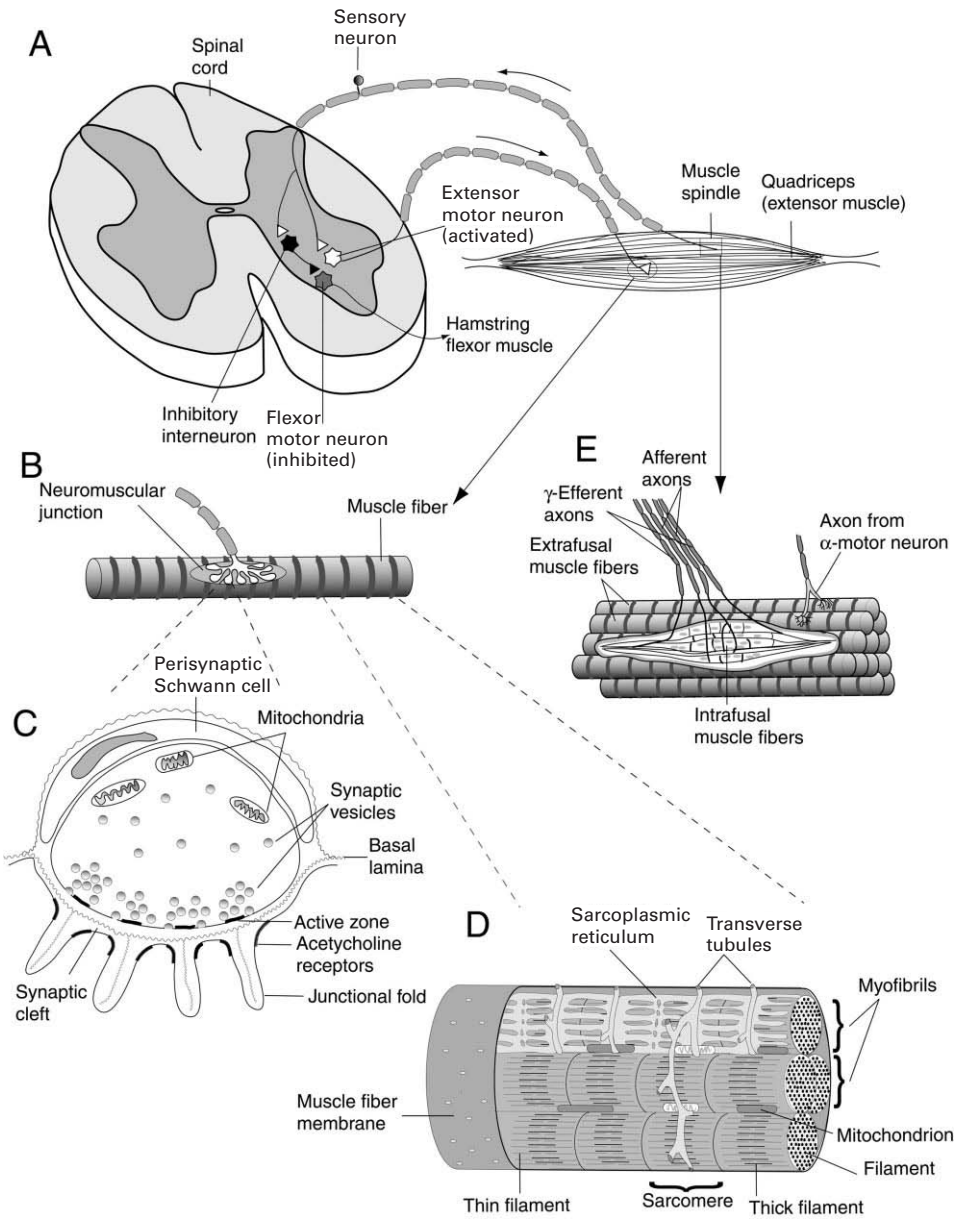

Figure 1
Schematic representation of the neuromuscular system (A–E). (A) The circuitry of the neuromuscular system involves motor neurons in the spinal cord, skeletal muscle, sensory receptors, and sensory neurons. Stretching of the extensor muscle activates the muscle spindle and generates electric impulses in the sensory neuron, which, in turn, excites the motor neuron that innervates the extensor muscle, causing muscle contraction. The sensory neuron acts indirectly, through the inhibitory neuron, to inhibit the motor neuron that innervates the flexor muscle. Thus, the knee-jerk reflex involves a coordinated contraction of the extensor muscle and relaxation of the flexor muscle. (B) A light microscopic representation of the neuromuscular junction where an axon innervates a single muscle fiber. (C) An electron microscopic representation of a cross-section of the neuromuscular junction. (D) An electron microscopic representation of a longitudinal section of a segment of a single muscle fiber. (E) A diagram of a muscle spindle showing intrafusal muscle fibers innervated by afferent and efferent axons

makes a synaptic contact with a gap of approximately 50 nm, called the synaptic cleft (Fig. 1C), between these two cellular elements. Besides the nerve terminal and the muscle fiber, the third cellular element at the NMJ are glial cells, called perisynaptic Schwann cells (also known as terminal Schwann cells, Fig. 1C), which cap the nerve terminal. At the adult vertebrate NMJ, each skeletal muscle fiber is, in general, innervated by a single motor axon. Due to its relative simplicity and accessibility, the NMJ has served as a model system for investigating the structure, function, and development of chemical synapses (Sanes and Lichtman 1999).

1.1 The Motor Nerve Terminal

Although the motor axon is wrapped by myelin sheath formed by Schwann cells, the terminal region is only capped by perisynaptic Schwann cells without myelin formation (Fig. 1B). In mammalian muscles, the nerve terminal branches into multiple grape-like varicosities, called synaptic boutons, which occupy a circumscribed region, approximately 50 μm in diameter, or less than 0.1 percent of the muscle surface area. At the frog NMJ, in contrast, the end of the motor fiber branches out into several elongated branches, each of which can extend several hundred μm long. The diameter of individual boutons or terminal branches is approximately 1–2 μm. Electron microscopy reveals mitochondria and conspicuous synaptic vesicles (Fig. 1C), approximately 20–50 nm in diameter, inside the nerve terminal. Synaptic vesicles tend to cluster around a thickened membrane specialization called the active zone (Fig. 1C) on the side of the nerve terminal facing the muscle fiber. The active zone is believed to be the site of transmitter release where the opening of synaptic vesicles during intense stimulation can be observed with electron microscopy (Heuser and Reese 1977). Freeze-fracture electron microscopy shows that the active zone is composed of two double-rows of large intramembrane particles ($\sim$10 nm in diameter), located precisely opposed to the opening of junctional folds where acetylcholine receptors are clustered on the muscle surface (Fig. 1C). These large intramembrane particles are thought to correspond to the location of voltage-gated calcium channels that allow Ca^{2+} entry to trigger exocytosis of synaptic vesicles and release of the transmitter. The precise alignment of voltage-gated calcium channels at the active zone and acetylcholine receptors at the junctional fold plays a strategic role in the speedy action of the transmitter-receptor interaction. The length of the active zone correlates with the efficacy of synaptic transmission.

1.2 Transmitter Release

The neurotransmitter, acetylcholine, is stored in synaptic vesicles. There are about 10^6 synaptic vesicles inside the nerve terminal at the frog NMJ. A variety of proteins on the synaptic vesicle and on the nerve terminal membrane govern the process of transmitter release. Following nerve stimulation, the conductance of voltage-gated calcium channels increases and the intracellular calcium concentration rises transiently. It is believed that some synaptic vesicle proteins serve as Ca^{2+} sensors and trigger the exocytosis of synaptic vesicles. Exocytosis involves binding between some integral proteins in the vesicle membrane, called v-SNARES, and some in the nerve terminal membrane, called t-SNARES (Sudhof 1995). Their binding results in fusion between the synaptic vesicle membrane and the nerve terminal membrane at the active zone. Some of these SNARES are the targets of neurotoxins. For example, botulinum toxins cleave t-SNARES, causing transmitter release blockade and muscle paralysis.

The transmitter can be released spontaneously or evoked by nerve stimulation (Katz 1966). The spontaneous exocytosis of synaptic vesicles produces miniature endplate potentials (mepps), which occur randomly with an average frequency of about 1 Hz and amplitude of 0.5 mV. It is believed that the mepp is caused by the release of approximately 10,000 acetylcholine molecules stored in one synaptic vesicle. The nerve stimulation evokes nearly synchronous release of many synaptic vesicles and produces endplate potentials (epps). The amplitude of epps is dependent on Ca^{2+} influx in a non-linear manner such that a two-fold increase in Ca^{2+} influx can increase the evoked transmitter release up to 16-fold. In a normal saline solution, each nerve impulse can trigger a synchronous release of about 200 synaptic vesicles and result in an epp that is large enough to trigger an action potential and muscle contraction. Following transmitter release, synaptic vesicle membranes are retrieved through a process called endocytosis that involves membrane invagination and formation of an endosome, a membrane compartment inside the nerve terminal (Heuser and Reese 1977). New synaptic vesicles bud off the endosome, refilled with acetylcholine. The local recycling of synaptic vesicles takes about 30 sec to 1 min to complete and provides the nerve terminal with a continual supply of synaptic vesicles even under intense stimulation.

1.3 Acetylcholine Receptors at the NMJ

The acetylcholine receptors (Unwin 1993) in skeletal muscles belong to the nicotinic type because nicotine derived from the tobacco plant activates these receptors. This is in contrast to the muscarinic type of acetylcholine receptors found in the brain that are activated by muscarine derived from a poisonous red mushroom. Both mepps and epps are caused by the binding of acetylcholine to nicotinic acetylcholine receptors on the muscle fiber. The acetylcholine receptors are densely packed, approximately 10,000 per μm^2, at the endplate region, compared with fewer

than 10 per μm^2 on the rest of the muscle membrane. Each acetylcholine receptor at the adult NMJ consists of five subunits: two α-subunits, one β-, one δ-, and one ε-subunit. All of these five subunits contribute to form the receptor channel. As a consequence of two molecules of acetylcholine binding to the α-subunits, the receptor channel opens and allows ion flow that results in depolarization of the muscle membrane.

Prior to innervation, acetylcholine receptors are distributed throughout the muscle membrane at a density of approximately 1000 per μm^2 during development. A protein, called agrin, triggers the aggregation of acetylcholine receptors at the developing NMJ (McMahan 1990). Agrin is synthesized by the motor neuron, released from the motor nerve terminal, and incorporated into the extracellular matrix (see Section 1.5) in the synaptic cleft. Agrin interacts with several molecules, including dystroglycan and a muscle-specific tyrosine kinase, located on the muscle membrane, and stimulates the aggregation of acetylcholine receptors, which are also associated with a cytoplasmic protein termed rapsyn. Motor neurons also enhance the synthesis of acetylcholine receptors through a molecule named neuregulin, which interacts with a set of receptor tyrosine kinases in the muscle membrane (Fischbach and Rosen 1997). After nerve injury, the density of acetylcholine receptors in the muscle membrane outside the NMJ increases markedly, a phenomenon termed denervation supersensitivity. Direct muscle stimulation can prevent or reverse denervation supersensitivity. Thus, the motor nerve plays a role not only in the aggregation and synthesis of acetylcholine receptors, mediated by agrin and neuregulin, respectively, but also in the nerve activity-mediated suppression of acetylcholine receptor expression in the extrajunctional region of the muscle (Sanes and Lichtman 1999). As a result of these effects of the motor nerve terminal, acetylcholine receptors are concentrated at the endplate region where they can readily bind acetylcholine molecules released from the terminal.

1.4 The Perisynaptic Schwann Cell (PSC)

The PSC is a glial cell that overlays the nerve terminal (Fig. 1C). The role of the PSC had not been studied extensively until the 1990s. Neural activity can cause an elevation of internal Ca^{2+} concentration and regulate gene expression in the PSC. Conversely, activation of the PSC can regulate the amount of transmitter released. Thus, modulation of the efficacy of neuromuscular transmission likely involves reciprocal interactions between the nerve terminal and the PSC. Complex neuron-glia interactions have also been found at the NMJ after nerve injury. PSC processes sprout profusely beyond the original endplate region following nerve injury. As reinnervation progresses, nerve terminals closely follow the preceding PSC sprouts. It is believed that these PSC sprouts play a role in guiding and leading nerve regeneration and sprouting after nerve injury (Son et al. 1996). Schwann cells express many trophic factors that may be important for neuronal survival and axonal growth. Thus, it is likely that the PSC plays a role in the modulation of synaptic function, as well as in the formation and maintenance of synaptic connections.

1.5 The Extracellular Matrix at the NMJ

Besides agrin, there are many important molecules that reside in the synaptic cleft, which is filled with amorphous material called the extracellular matrix or basal lamina (Fig. 1C). Some of these molecules are large glycoproteins, such as laminins, which may promote nerve growth, adhesion between nerve and muscle, or differentiation of the nerve terminal (Sanes and Lichtman 1999). An important enzyme located in the extracellular matrix is acetylcholinesterase, which hydrolyzes acetylcholine molecules that do not bind to acetylcholine receptors. The prompt removal of acetylcholine from the synaptic cleft by acetylcholinesterase prevents desensitization of acetylcholine receptors caused by the continuous presence of acetylcholine. Thus, acetylcholinesterase plays an essential role in ensuring proper transmission between nerve and muscle (Salpeter 1987).

2. Skeletal Muscle Contraction

A typical mammalian skeletal muscle is composed of hundreds of individual muscle fibers that are approximately 50–100 μm in diameter and 2–6 cm in length. Each muscle fiber is a multi-nucleated muscle cell and is made up of myofibrils, the contractile elements that carry out the work of contraction. Each myofibril consists of a chain of repeated contractile units called sarcomeres (Fig. 1D). In each sarcomere, thin and thick filaments run parallel to one another with varying degrees of overlap depending on the contractile state. The thin filament is made primarily of actin molecules and the thick filament of myosin molecules. According to the sliding filament model, Ca^{2+} triggers interactions between actin and myosin molecules and results in the sliding of thin and thick filaments past one another, thereby causing muscle contraction. Following nerve stimulation and synaptic transmission, electrical impulses generated in the muscle fiber spread rapidly into membrane invaginations, called transverse tubules, which extend into sarcomeres and form contacts with sarcoplasmic reticulum, a modified endoplasmic reticulum that sequesters and releases Ca^{2+}. It is thought that depolarization of the transverse tubule results in the opening of Ca^{2+} release channels located on the sarcoplasmic reticulum membrane. As a result, Ca^{2+} is released from the sarcoplasmic reticulum into the

cytosol and initiates muscle contraction. Muscle relaxation occurs when Ca^{2+} ions are pumped back into the sarcoplasmic reticulum. Thus, Ca^{2+} plays important roles not only in transmitter release but also in excitation-contraction coupling in the neuromuscular system (Engel and Franzini-Armstrong 1994).

3. *Sensory and Motor Controls of the Neuromuscular System*

The neuromuscular system is under both voluntary controls and reflex controls. The motor neurons in the spinal cord receive synaptic connections from other neurons in the central nervous system for voluntary control of body movements. The spinal motor neurons are also involved in reflex controls to maintain proper posture and to prevent muscle injury from excessive contraction. Two sensory receptors, the muscle spindle and the Golgi tendon organ, provide information on muscle length and tension, respectively, to motor neurons for reflex controls (Engel and Franzini-Armstrong 1994). The message may excite or inhibit motor neurons depending on the extent of the stretch or the tension, and on the type of muscle involved.

Muscle spindles are embedded in, and run parallel to the long axis of, the skeletal muscle (Fig. 1E). These spindle-shaped sensory organs consist of a group of intrafusal muscle fibers, which are modified muscle fibers lacking myofibrils in the central region. Thus, the centers of the intrafusal fibers are not contractile, in contrast to the regular skeletal muscle fibers, which are also called the extrafusal fibers. The ends of the intrafusal fibers are innervated by nerve terminals of small-diameter motor neurons, called γ motor neurons. This is in contrast to the regular extrafusal fibers, which are innervated by the large α motor neurons. Activation of γ motor neurons causes shortening of the polar regions of the intrafusal fiber and thus enhances the sensitivity of the muscle spindle to stretch. In addition to motor innervation, intrafusal fibers are also wrapped by nerve endings of sensory (afferent) fibers. The cell bodies of sensory fibers reside in the dorsal root ganglion.

The role of the muscle spindle is exemplified in the knee-jerk reflex (Kandel et al. 2000). When the quadriceps (extensor) muscle in the thigh is stretched, for example, by tapping the kneecap with a small rubber hammer, the stretching activates the sensory nerve endings in the muscle spindle. This stimulation generates receptor potentials and activates the sensory neuron, which propagates electrical impulses towards the spinal cord. The sensory axon makes excitatory synapses directly onto the α motor neuron that innervates the extrafusal fibers of the quadriceps muscle (Fig. 1A). Thus, the stretch of the muscle spindle in the quadriceps muscle activates the α motor neuron, which, in turn, causes the quadriceps to contract. As a result, the lower leg swings forward. This simple reflex, which involves only one synapse between the sensory neuron and the motor neuron, is called the monosynaptic stretch reflex. In addition to the monosynaptic pathway, the sensory neuron acts indirectly, through an inhibitory neuron, to inhibit the motor neuron that innervates the opposing hamstring (flexor) muscle (Fig. 1A). The knee-jerk reflex involves a simultaneous activation of both pathways and results in a coordinated contraction of the extensor quadriceps muscle and relaxation of the flexor hamstring muscle.

Whereas muscle spindles are most sensitive to stretch, Golgi tendon organs are most sensitive to tension. Golgi tendon organs reside at the junction between the tendons and muscle fibers, and are thus arranged in series to the skeletal muscle. Each tendon organ, about 1 mm long and 0.1 mm in diameter, consists of unmyelinated nerve endings that intertwine among collagen fibers encapsulated by connective tissue. Excessive muscle contraction activates Golgi tendon organs, which inhibit α motor neurons that innervate the activated muscle. Thus, the Golgi tendon organ provides negative-feedback control, which prevents extreme muscle contraction that might damage the muscle fibers.

4. *Toxins that Affect the Neuromuscular System*

Because of the importance of the neuromuscular system to vital functions, such as respiration and escape from predators, it is not surprising that nature has evolved animals and plants that produce a variety of poisons to block neuromuscular transmission and thereby paralyze their prey. For example, α-bungarotoxin, a peptide found in the venom of the banded krait *Bungarus multicinctus*, blocks neuromuscular transmission by irreversibly binding to nicotinic acetylcholine receptors (Chang 1979). Because of its irreversible binding, α-bungarotoxin has been a useful tool to study the function and distribution of acetylcholine receptors at the NMJ. Another acetylcholine receptor blocker is curare, a mixture of plant toxins used for poison darts by South American Indians. Other toxins block neuromuscular transmission by affecting transmitter release. For example, fish-hunting marine cone snails disable their prey by injecting an array of toxins, known as conotoxins, some of which block voltage-gated calcium channels (Olivera et al. 1994). The venom in the female black widow spider contains α-latrotoxin that causes a massive release of acetylcholine at the NMJ. Botulinum toxin and tetanus toxin from bacterial growth in infected food or tissues cleave some proteins associated with synaptic vesicle release and thus disrupt neuromuscular function. Another mode of interfering with neuromuscular transmission is to inactivate acetylcholinesterase. For example, many deadly nerve gases are inhibitors of acetyl-

cholinesterase and cause paralysis of the respiratory muscle through the accumulation of acetylcholine molecules, thereby desensitizing and inactivating acetylcholine receptors.

5. *Diseases of the Neuromuscular System*

Diseases of the neuromuscular system (Engel and Franzini-Armstrong 1994) include disorders of the NMJ, degeneration of motor neurons in the spinal cord, and muscular dystrophies. The most common disorder affecting transmission at the NMJ is myasthenia gravis. This disease is characterized by weakness of limb muscles and cranial muscles. Myasthenia gravis is an autoimmune disease in which patients produce antibodies against their own nicotinic acetylcholine receptors at the NMJ. These antibodies inhibit neuromuscular function by reducing the number of functional receptors, or by interfering with the binding of acetylcholine to the receptors. Another neuromuscular disorder is the Lambert-Eaton myasthenia syndrome, also an autoimmune disease in which patients generate antibodies against voltage-gated calcium channels at the motor nerve terminal, thereby disrupting transmitter release. The best-known motor neuron disorder is amyotrophic lateral sclerosis, also known as Lou Gehrig disease. It is not known why only motor neurons, but not sensory neurons, die in this devastating disease. As a result of the motor neuron degeneration, patients experience weakness of the arms and legs, and eventually suffer fatal failure of respiratory muscles. Muscular dystrophies, characterized by muscle weakness and loss of muscle fibers, are myopathic diseases. One severe form of these diseases is Duchenne muscular dystrophy, which is an X-linked inherited disease causing symptoms only in boys. The disease is attributed to a deficiency in dystrophin, a large protein located on the inner surface of the muscle membrane. Dystrophin links actin molecules and an array of glycoproteins that form a complex near the muscle membrane. Absence of dystrophin or some of these dystrophin-associated glycoproteins causes a disruption of the muscle membrane and thereby muscle degeneration.

See also: Classical Mechanics and Motor Control; Motor Control; Motor Control Models: Learning and Performance; Neural Plasticity of Spinal Reflexes; Somatosensation; Synapse Formation; Synapse Ultrastructure; Synaptic Transmission

Bibliography

Chang C C 1979 The action of snake venoms on nerve and muscle. In: Lee C Y (ed.) *Snake Venoms: Handbook Experimental Pharmacology*, Vol. 52. Springer, Berlin, pp. 309–76

Engel A G, Franzini-Armstrong C 1994 *Myology*. McGraw-Hill, New York

Fischbach G D, Rosen K M 1997 ARIA: a neuromuscular junction neuregulin. *Annual Review of Neuroscience* **20**: 429–58

Heuser J E, Reese T S 1977 Structure of the synapse. In: Kandel E R (ed.) *Handbook of Physiology: A Critical, Comprehensive Presentation of Physiological Knowledge and Concepts, Sect. 1, The Nervous System. Vol. 1, Cellular Biology of Neurons*, Part 1, American Physiological Society, Bethesda, MD, pp. 261–94

Kandel E R, Schwartz J H, Jessell T M 2000 *Principles of Neural Science*, 4th edn. McGraw-Hill, New York

Katz B 1966 *Nerve, Muscle and Synapse*. McGraw-Hill, New York

McMahan U J 1990 The agrin hypothesis. *Cold Spring Harbor Symposium Quantitative Biology* **5**: 407–18

Olivera B M, Miljanich G P, Ramachandran J, Adams M E 1994 Calcium channel diversity and neurotransmitter release: The ω-conotoxins and ω-agatoxins. *Annual Review of Biochemistry* **63**: 823–67

Salpeter M M 1987 *The Vertebrate Neuromuscular Junction*. Alan R. Liss, Inc, New York

Sanes J R, Lichtman J W 1999 Development of the vertebrate neuromuscular junction. *Annual Review of Neuroscience* **22**: 389–442

Son Y J, Trachtenberg J T, Thompson W J 1996 Schwann cells induce and guide sprouting and reinnervation of neuromuscular junctions. *Trends in Neurosciences* **19**: 280–5

Sudhof T C 1995 The synaptic vesicle cycle: a cascade of protein-protein interactions. *Nature* **375**: 645–53

Unwin N 1993 Neurotransmitter action: opening of ligand-gated ion channels. *Cell* **72**(*Suppl.*): 31–41

C.-P. Ko

Neurons and Dendrites: Integration of Information

1. *Introduction and Scope*

At the beginning of the twenty-first century, neural tissue remains arguably the most powerful substrate for information processing in existence. Unfortunately, the principles underlying brain function are still known only in rough outline. To unlock the many secrets of the brain, and to help explain the panoply of neural functions ranging from simple motor reflexes to conscious thought and action, neuroscientists have amassed an extensive body of experimental data bearing on the properties of nerve cells (neurons) and their interconnections (synapses), which together form the building blocks of neural tissue. At the same time, the functional significance of the neurobiological data has been explored in computer modeling studies, which have often been used to test hypotheses inaccessible to direct experimental approaches. This article summarizes what is known about the intracellular physiology of individual nerve cells, and interprets these data in relation to the information processing

functions of single neurons. For comprehensive treatments of many of these same issues, see Stuart et al. (1999) and Koch (1999).

Given the morphological and functional diversity of neurons across the invertebrate and vertebrate phyla, the present article is focused on the principal neuron types of the vertebrate central nervous system (CNS), though much of what is discussed applies to neurons in general. This focus is justified, first, by the fact that neurons of the vertebrate CNS have been among the most frequent objects of correlated physiological, behavioral, and modeling studies, and second, by the fact that these cells are those which most directly underlie the cognitive and behavioral phenomena with which this volume is concerned.

2. *Basic Construction of a Neuron*

The typical neuron of the vertebrate CNS is composed of (a) a cell body, or soma, which contains the nucleus and other molecular machinery present in any living cell, (b) the dendritic tree, which emanates from the cell body and acts as the main input-receiving surface of the neuron, (c) the axon which carries output signals generated by the neuron over long distances to other neurons, and (d) a large number of synapses, the specialized sites where axons make contacts with the dendrites and cell bodies of other neurons for subsequent processing (Fig. 1).

The axon, a thin fiber emanating from the cell body, is the cell's final common output used to convey messages to other neurons. The axon carries trains of 'action potentials' (APs), all-or-none voltage spikes 60–100 mV in height and 1–2 ms in duration which propagate undecremented along the length of the axon at speeds up to 100 m/s. The single axon of a neuron, which may travel over distances ranging from millimeters to meters depending on the neuron, and which may branch profusely at one or more sites along its run, gives rise to many thousands of synaptic terminals impinging on the dendrites and cells bodies of hundreds to thousands of other neurons. Neurons can fire APs at rates up to 1,000 impulses per second, though the maximum firing rate for most neurons in the CNS lies in the range of 100–200 Hz. APs are triggered when the membrane potential at the site of spike initiation reaches a firing threshold. In a typical neuron, this threshold is around -50 mV, that is, about 20 mV more positive than the cell's resting potential of -70 mV. While APs are thought to be triggered in the axon near the cell body, the electric currents which flow there represent the combined influences of many widely scattered synaptic inputs acting on the soma/dendritic membrane.

The dendritic tree is the main input-receiving domain of the neuron, generally accounting for more than 95 percent of the cell's total surface area, and receiving many thousands of synaptic inputs from other cells. With a few notable exceptions, the den-

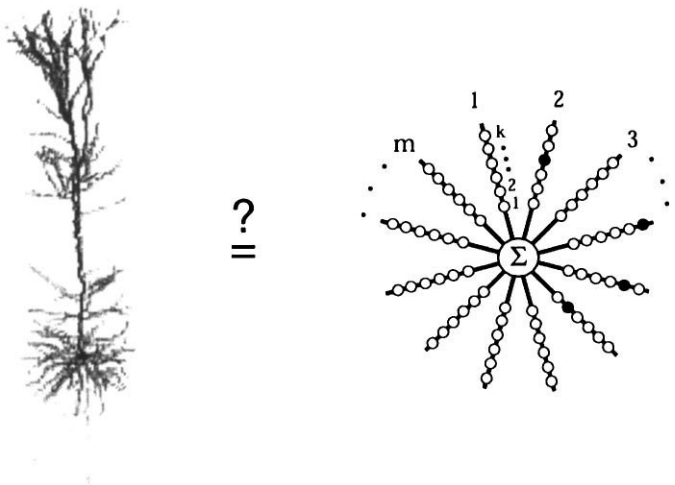

Figure 1
(Left). Pyramidal cell from rat neocortex, adapted with permission from Lübke et al. (1996). Dendrites and cell body are shown in dark (green); cell body is dark pyramidal shape. Tiny protrustions covering dendritic branches (barely visible in figure) are 'spines,' the sites of contact of most excitatory synapses; one pyramidal cell may contain tens of thousands of spines. Inhibitory synapses are less numerous, occur primarily on dendritic shafts rather than spines, and are concentrated near the cell body. Axon is shown in faint (red). Small boutons throughout axonal arborization denote sites of synaptic contact with dendrites of other cells. (Right). Model neuron with m identical branches containing k synapses each.

drites of CNS neurons are purely input structures (Shepherd 1998). A synaptic input to a dendrite can be categorized roughly as either 'excitatory' or 'inhibitory,' meaning that when the synapse is activated, the membrane on the 'postsynaptic,' that is, dendritic, side of the synaptic junction is driven transiently to a more positive (excitatory) or more negative (inhibitory) potential. Most of the contacts on a dendritic tree are excitatory (80–90 percent); inhibitory contacts can occur anywhere, but tend to be concentrated on or around the cell body. A synaptically evoked change in membrane potential can have a time course ranging from one to hundreds of milliseconds depending on the type of synapse. In addition, synaptic inputs can trigger 'active' dendritic responses, including complex spike-like events involving both fast and slow components (e.g., Kamondi et al. 1998, Golding et al. 1999). Thus, the rules which govern how excitatory and inhibitory synaptic inputs are combined and transformed within the dendritic tree to produce outbound action potentials can be quite complex. This process commonly is referred to as synaptic or dendritic 'integration,' and is the focus of the remainder of this article.

3. *The Membrane Potential*

To understand how synaptic and other electrical events combine to generate APs, it is important to understand the basic electrical properties of 'neurites' (axons and dendrites).

The resting membrane potential of a neuron, as for any living cell, is established by molecular pumps which actively maintain concentration gradients of sodium (Na^+) and potassium (K^+) ions across the cell membrane. Relative to the extracellular fluid, the inside of the cell contains a high concentration of K^+ ions and a low concentration of Na^+ ions. The constant pressure on these electrically charged ions to flow across the membrane down their respective concentration gradients (Na^+ in, K^+ out), is analogous to the pressure electrons feel which propels them from the negative to the positive terminal of a battery given a suitable path for current. The electrical pressure on each of these ions can thus be represented by a battery, whose voltage conveys the magnitude and direction of the ion's concentration gradient; the fact that the Na^+ and K^+ batteries point in opposite directions will turn out to be critical for electrical signaling.

Since the membrane lipid bilayer is itself impervious to ions, currents can only flow through 'ion channels,' specially designed pores in the membrane formed by large, membrane-spanning protein complexes. Ion channels are generally quite selective as to which ions may flow through them, giving rise to the nomenclature 'sodium channels,' 'potassium channels,' and so on. Furthermore, certain ion channels are tonically open, while others can open and close under the control of various signals. For example, some channels open and close in response to changes in the membrane potential, such as occurs in the production of action potentials, others in response to chemical signals, such as occurs in the production of synaptic potentials.

The resting membrane potential of a neuron is a weighted average of ionic battery values, where the weight given to each ion's battery is determined by the membrane's total electrical conductance g_I to ion I through all open channels

$$V_m = \frac{g_{Na}E_{Na} + g_K E_K + \dots}{g_{Na} + g_K + \dots} \tag{1}$$

The usual resting potential around $V_m = -70\,\text{mV}$ is dominated by a K^+ conductance, whose associated battery value is $E_K = -90\,\text{mV}$. An additional small resting Na^+ conductance, whose associated battery value $E_{Na} = +60\,\text{mV}$ or higher, pulls the resting potential some 20 mV in the positive direction.

4. *Electrical Signaling Results from Conductance Changes*

Electrical signaling in a neuron is caused by transient changes in the conductances of various ion channels, which cause transient fluctuations in the membrane potential about its resting value. At every moment in time, the membrane potential moves towards the equilibrium value given by Eqn. (1)—not instantaneously—but at a rate limited by the need to charge or discharge the membrane capacitance. The classical signal generated by a neuron is the action potential. When a synaptic or other injected current drives the membrane potential to a threshold of approximately $-50\,\text{mV}$, a population of voltage-dependent Na^+ channels, normally closed at rest, opens abruptly to a large conductance value resulting in a large influx of current. As predicted by Eqn. (1), this leads to a rapid voltage upswing as the membrane potential spikes toward the Na^+ battery potential of $+60\,\text{mV}$. Within a fraction of a millisecond, however, the Na^+ channels begin to close, or 'inactivate,' of their own accord. At the same time, voltage-dependent K^+ channels sense the voltage upswing on the leading edge of the AP, and open to produce a large outward current. The combined effects of Na^+ channels closing and K^+ channels opening results in an abrupt downswing in the membrane potential to terminate the AP.

Active electrical responses of the neuronal membrane have historically been most closely associated with the long-range signal transmission function of the axon. It is now apparent, however, that the dendrites of many neuron types contain voltage-dependent

channels capable of generating spikes which are independent of the cell body (Amitai et al. 1993, Golding and Spruston 1998, Kamondi et al. 1998). Unlike the stereotyped all-or-none form of the classical AP, however, synaptically-evoked dendritic spikes are of variable magnitude and can consist of complex superpositions of fast and slow events. As discussed below, this highly nonlinear electrical behavior of dendritic origin is likely to profoundly influence the integrative properties of nerve cells.

5. *Synaptic Inputs*

Synaptic potentials, the principal inputs signals to which a neuron responds, likewise are evoked by changes in the conductance of ion channels in the postsynaptic membrane. In this case, the conductance changes are triggered by the binding of neurotransmitter molecules to receptors exposed on the extracellular surface of the membrane. The normal sequence of events in synaptic transmission is as follows. An AP invades the presynaptic axon terminal. The voltage upswing in the presynaptic terminal opens voltage-dependent calcium (Ca^{++}) channels leading to a rapid influx of Ca^{++} ions. A small number of vesicles (usually zero or one) containing neurotransmitters fuse with the presynaptic membrane at specialized release sites and disgorge their contents into the synaptic cleft. The neurotransmitter (e.g., glutamate or GABA) diffuses the several nanometers across the synaptic cleft in a fraction of a millisecond, and binds with receptors on the postsynaptic membrane. Upon binding, the receptor-channel complex undergoes a conformational change, causing the channel to open. Alternatively, in the case of 'second messenger' transmission, the protein to which the neurotransmitter binds is not the channel itself, but instead triggers a cascade of biochemical reactions inside the postsynaptic cell which lead indirectly to the opening or closing of ion channels. In either case, the resulting conductance change(s) cause the membrane potential in the postsynaptic compartment to fluctuate toward the characteristic battery value for that synapse, often called the synaptic 'reversal potential.' The reversal potential at an AMPA- or NMDA-type excitatory synapse is typically 0 mV, 70 mV more positive than the resting potential. This value results from an approximately equal conductance to Na^+ and K^+ ions. A synapse with a reversal potential around 0 mV is termed 'excitatory' since under normal operating conditions with a negatively polarized membrane potential, activation of the synapse drives the membrane transiently to more positive potentials (i.e., towards 0 mV). This 'depolarization' increases the likelihood (or rate) of outbound AP production in the postsynaptic cell. Conversely, the reversal potential is -65 mV for $GABA_A$-type inhibitory synapses which conduct chloride ions, and -90 mV for $GABA_B$-type inhibitory synapses which conduct K^+ ions. In either case, activation of the synapse opposes synaptic excitation by transiently pulling the membrane potential to more negative values, thus reducing the likelihood (or rate) of AP generation.

6. *Synaptic Integration*

To this point we have seen how transmitter-dependent and voltage-dependent gating of ion channels can produce the fluctuations in membrane potential that underlie both synaptic responses and action potentials. But what are the rules governing synaptic integration?

The simplest 'linear' model of dendritic integration holds that the neuron sums its excitatory and inhibitory influences across the dendritic arbor, followed by a global thresholding nonlinearity at the cell body (McCullough and Pitts 1943, Rosenblatt 1962, Rumelhart and McClelland 1986). Despite its simplicity (Fig. 1), this formulation has influenced thinking about neuronal function in both the experimental and computational communities for many years. Extensions to this idea that explicitly include the 'passive cable' properties of the neuron—accounting for attenuation of signals with time and distance (Rall 1964, Koch et al. 1982, Agmon-Snir and Segev 1993)—have also been highly influential, and have led to a variety of productive interchanges between theory and experiment (Rall et al. 1992).

From the perspective of currently available neurophysiological data, however, much of it derived from direct intradendritic recordings, synaptic integration is likely to be more complex than can be captured by linear or passive models of dendritic function (Segev and Rall 1998, Mel 1999). The question is complicated for three reasons, principally relating to uncertainties as to how or to what extent synaptic inputs interact with each other through shared voltage or chemical signals within the dendrites.

First, the spatio-temporal patterns of synaptic activation driving neurons in the CNS *in vivo* are not well known, so that current best guesses as to the electrical and chemical environments in which synapses normally operate remain largely rooted in conjecture.

Second, synaptic inputs may trigger secondary events in the dendrites mediated by voltage-dependent NMDA, Na^+, K^+, and Ca^{++} channels, which could influence significantly how excitatory and inhibitory synaptic inputs are combined to produce overall neuronal responses. Synaptically-evoked dendritic spikes have been observed *in vitro* (Golding and Spruston 1998, Schiller et al. 2000) and *in vivo* (Kamondi et al. 1998), though it is not yet known whether such events occur commonly in awake, behaving animals, and if so whether they are localized

to one or many dendritic branches, and so on. In addition, the spatial distributions of these channels on the cell surface, their kinetic properties, their dependence on voltage and other modulatory factors, and their efficacies under normal *in vivo* operating conditions, are not sufficiently well known to make firm predictions regarding the contributions of these channels to the integrative processes of the cell.

Third, cable theory (Jack et al. 1975, Koch 1999) informs us that the dendritic morphology of a typical CNS neuron—typically consisting of many thin-branch subtrees radiating outward from the cell body and/or main dendritic trunks—is ideally suited to isolate voltage signals within specific subregions of the dendritic arbor. Thus, the electrical coupling between a thin branch and a main trunk (or soma) is expected to produce a significant attenuation of distally generated voltage signals when measured at the trunk, with the most profound suppression of signal transmission reserved for rapid voltage transients associated with fast synaptic potentials or dendritic spikes. Thus, if thin branches, which receive the vast bulk of the excitatory input in most neurons, routinely generate fast spike-like events in response to synaptic excitation, the cable structure of these cells appears designed to ensure that distally-generated spikes are seen only in strongly attenuated form in other parts of the cell. Clearly, this type of functional compartmentalization could complicate the issue of synaptic integration, since the overall response of the cell would depend on whether synaptic inputs were delivered to the same or different dendrites, and whether dendritic spikes were generated or not.

7. *Computational Implications*

In spite of many uncertainties, currently available experimental data, combined with insights gained from theoretical and modeling studies, seem to support the hypothesis that dendritic trees are compartmentalized functionally, and could consist of a moderately large number of, for example, up to 100, independent integrative 'subunits.' The idea that dendritic trees are compartmentalized functionally has been discussed in a variety of forms over the years (Koch et al. 1982, Shepherd et al. 1985, Rall and Segev 1987, Mel 1992a, 1992b, 1993, Mel et al. 1998, Archie and Mel 2000, Poirazi and Mel 2001). In the continued absence of empirical evidence sufficient to confirm or deny this "nonlinear subunit" hypothesis, several recent modeling studies have provided support for one simple version of the idea. According to this model, (a) each dendritic subunit, corresponding to a single branch or small thin-branch subtree, acts as a separately thresholded neuron-like summing unit, and (b) the many subunit responses are summed in the main trunks and the cell body to produce the cell's final output (Fig. 1). This formulation ignores differences in the locations of synapses within a subunit, as justified by theoretical considerations and biophysical modeling studies indicating that voltage signals are communicated far more effectively within a dendritic branch than between branches—especially for branches separated by a main trunk or the cell body. In mathematical terms, the neuron's overall output y is given by a sum of m independently-thresholded subunit responses. Each subunit response is given by a weighted sum of k inputs followed by a fixed subunit nonlinearity, g, which is typically modeled as a sigmoid or power function

$$y = \sum_{j=1}^{m} g\left(\sum_{i=1}^{k} w_{ij} x_{ij}\right) \tag{2}$$

This simple formula, which is similar to the expression describing a conventional two-layer neural network (Bishop 1995), was found to predict the time-averaged firing rates produced by a biophysically detailed compartmental model of a pyramidal cell, a model which included passive membrane properties, a spatially-extended dendritic morphology, synapses with realistic conductance models, and voltage-dependent ion channels in the dendrites (Archie and Mel 2000).

From a functional perspective, it is interesting to note that the sum-of-thresholded-subunits model captures the underlying structure of a variety of classical and extraclassical receptive field properties found in cortical neurons, suggesting that dendritic subunits could contribute in important ways to the many quotidian signal-processing functions of cortical tissue (Mel 1999). In a different vein, a recent modeling study has shown that subunit-containing neurons as described in Eqn. (2) could increase the trainable storage capacity of neural tissue by orders of magnitude (Poirazi and Mel 2001).

While it is clear that many questions remain to be answered regarding the information processing functions of single neurons, one may be confident that with the steady advance of experimental methods available to study neuronal function at the required level of detail, coupled with the enormous flexibility of hypothesis testing provided by the computer modeling enterprise, the many fascinating secrets of neural information processing will gradually yield in the coming years.

See also: Artificial Neural Networks: Neurocomputation; Associative Modifications of Individual Neurons; Cerebral Cortex: Organization and Function; Ion Channels and Molecular Events in Neuronal Activity; Memory: Synaptic Mechanisms; Neural Networks and Related Statistical Latent Variable Models; Neural Networks: Biological Models and

Applications; Neural Networks, Statistical Physics of; Neurotransmitters; Synapse Formation; Synapse Ultrastructure; Synaptic Efficacy, Regulation of; Synaptic Transmission

Bibliography

Agmon-Snir H, Segev I 1993 Signal delay and input synchronization in passive dendritic structures. *Journal of Neurophysiology* **70**: 2066–85

Amitai Y, Friedman A, Connors B, Gutnick M 1993 Regenerative electrical activity in apical dendrites of pyramidal cells in neocortex. *Cerebral Cortex* **3**: 26–38

Archie K A, Mel B W 2000 An intradendritic model for computation of binocular disparity. *Nature Neuroscience* **3**(1): 54–63

Bishop C 1995 *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK

Golding N L, Jung H-Y, Mickus T, Spruston N 1999 Dendritic calcium spike initiation and repolarization are controlled by distinct potassium channel sybtypes in CA1 pyramidal neurons. *Journal of Neuroscience* **19**: 8789–98

Golding N L, Spruston N 1998 Dendritic sodium spikes are variable triggers of axonal action potentials in hippocampal CA1 pyramidal neurons. *Neuron* **21**: 1189–200

Jack J, Noble D, Tsien R 1975 *Electric Current Flow in Excitable Cells*. Clarendon Press, Oxford, UK

Kamondi A, Acsády L, Buzsáki G 1998 Dendritic spikes are enhanced by cooperative network activity in the intact hippocampus. *Journal of Neuroscience* **18**: 3919–28

Koch C 1999 *Biophysics of Computation*. Oxford University Press, New York

Koch C, Poggio T, Torre V 1982 Retinal ganglion cells: A functional interpretation of dendritic morphology. *Philosophical Transactions of the Royal Society of London B* **298**: 227–64

Lübke J, Markram H, Frotscher M, Sakmann B 1996 Frequency and dendritic distribution of autapses established by layer 5 pyramidal neurons in the developing rat neocortex: Comparison with synaptic innervation of adjacent neurons of the same class. *Journal of Neuroscience* **16**: 3209–18

McCullough W, Pitts W 1943 A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–33

Mel B W 1992a The clusteron: Toward a simple abstraction for a complex neuron. In: Moody J, Hanson S, Lippmann R (eds.) *Advances in Neural Information Processing Systems, Vol. 4*. Morgan Kaufmann, San Mateo, CA, pp. 35–42

Mel B W 1992b NMDA-based pattern discrimination in a modeled cortical neuron. *Neural Computation* **4**: 502–16

Mel B W 1993 Synaptic integration in an excitable dendritic tree. *Journal of Neurophysiology* **70**(3): 1086–101

Mel B W 1999 Why have dendrites? A computational perspective. In: Stuart G, Spruston N, Häusser M (eds.) *Dendrites*. Oxford University Press, Oxford, UK, pp. 271–89

Mel B W, Ruderman D L, Archie K A 1998 Translation-invariant orientation tuning in visual 'complex' cells could derive from intradendritic computations. *Journal of Neuroscience* **17**: 4325–34

Poirazi P, Mel B W 2000 Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron* **29**: 779–96

Rall W 1964 Theoretical significance of dendritic trees for neuronal input-output relations. In: Reiss R (ed.) *Neural Theory and Modeling*. Standard University Press, Stanford, CA, pp. 73–97

Rall W, Burke R, Holmes W, Jack J, Redman S, Segev I 1992 Matching dendritic neuron models to experimental data. *Physiological Reviews* **72**(4): S159–86

Rall W, Segev I 1987 Functional possibilities for synapses on dendrites and on dendritic spines. In: Edelman G, Gall W, Cowan W (eds.) *Synaptic Function*. Wiley, New York, pp. 605–36

Rosenblatt F 1962 *Principles of Neurodynamics*. Spartan, Washington

Rumelhart D, McClelland J 1986 *Parallel Distributed Processing*. MIT Press, Cambridge, MA

Schiller J, Major G, Koester H J, Schiller Y 2000 NMDA spikes in basal dendrites of cortical pyramidal neurons. *Nature* **404**: 285–89

Segev I, Rall W 1998 Excitable dendrites and spines—earlier theoretical insights elucidate recent direct observations. *Trends in Neuroscience* **21**: 453–60

Shepherd G 1998 *The Synaptic Organization of the Brain*. Oxford University Press, Oxford, UK

Shepherd G M, Brayton R K, Miller J P, Segev I, Rinzel J, Rall W 1985 Signal enhancement in distal cortical dendrites by means of interactions between active dendritic spines. *Proceedings of the National Academy of Science USA* **82**: 2192–5

Stuart G, Spruston N, Häusser M 1999 *Dendrites*. Oxford University Press, Oxford, UK

B. W. Mel

Neuropsychological Functioning, Assessment of

The aim of neuropsychological assessment is the observation and measurement of cognitive function in relation to damage or dysfunction of the brain. Typical causes of brain dysfunction are stroke, head injuries, infections, neurodegenerative diseases such as Parkinson's disease or Alzheimer's disease, and psychiatric disorders such as depression and schizophrenia. The combination of neuropsychological assessment and recently developed brain imaging techniques, which measure changes in cerebral blood flow or metabolism that accompany changes in neuronal activity, has provided major advances in the understanding of the brain circuits underlying human behavior. This article addresses an overview of the basic concepts, most frequently used neuropsychological tests, and the specific problems of neuropsychological assessment.

1. Basic Concepts in Clinical Neuropsychology

The main areas of cognitive function which are affected by damage or dysfunction of brain structures are perception, orientation, motor abilities, attention, language, memory, and executive function, as well as

affect and social behavior. Perception refers to the processing of stimuli and sensations from the external environment and the body. Perceptions involve the visual, auditory, tactile, and olfactory domains but most of our knowledge has derived from studies of the visual system. Visual perception is disrupted after lesions to the posterior regions of the brain, the occipital and temporal cortex. Specific lesions can disrupt the processing and perception of color, shape, or movement of objects, or lead to specific problems in recognizing faces or the meaning of objects.

With respect to complex motor abilities apraxia is an important concept. Apraxia refers to the inability to execute purposeful acts on verbal command and the inability to immitate gestures (Benton 1994).

Orientation refers to the specific awareness of a person in relation to time, place, situation and other people. Orientation requires intact attentional and perceptual abilities. Impairments of orientation are among the most frequent symptoms of brain disease (Lezak 1995). Transient loss of orientation is a common consequence of global disruption of brain function due to traumatic brain injury. Its duration depends upon the severity of the injury.

Attention is the ability to focus a 'mental spotlight' on sensory inputs, motor programs, memories, or internal representations (Kolb and Wishaw 1996), and also involves the capacity to concentrate on a task for a given period of time. Impairments in attentional functions are often observed as a result of diffuse brain damage (such as Alzheimer's disease) and in a range of neuropsychiatric diseases such as schizophrenia or depression (Maruff and Pantelis 1999). Attentional problems are also seen in developmental disorders such as Attention Deficit Hyperactivity Disorder (see *Attention-deficit/Hyperactivity Disorder (ADHD)*). A particularly striking form of an attentional deficit is contralateral neglect. Stimulus neglect is the failure to detect or to attend to stimuli in the visual space or body side contralateral to a parietal brain lesion in the absence of sensory deficits. In most cases neglect of the left side occurs after a right hemisphere lesion.

Another basic concept in neuropsychology is memory. Memory deficits are observed after traumatic brain injury, stroke, and in most neurodegenerative and neuropsychiatric diseases. There is general agreement that memory representation in the brain involves several subsystems, which are associated with different neuroanatomical structures. The most influential contemporary memory model is the 'multiple memory systems approach' proposed by Squire (1987). Declarative or explicit memory refers to information, which has been encoded in a specific spatiotemporal context (e.g., a name, face, or story) and can be assessed by conscious recollection. Semantic memory refers to knowledge about the world, whereas episodic memory mainly involves autobiographical information. Implicit or non-declarative memory refers to the influence of prior experience on behavior, even if such experiences are not remembered or cannot be 'declared.' Typical examples are the acquisition of motor and nonmotor skills or priming effects (Daum and Ackermann 1997). The brain area which mediates declarative memory formation is the hippocampus and related structures (Squire 1987). Non-declarative memory is dependent on a range of other areas, e.g., the basal ganglia and the neocortex (Daum and Ackermann 1997).

Another important concept of memory theories is working memory. Working memory is responsible for the integration of multiple stimuli, and the short-term maintenance of an active representation of information. Thus, working memory plays an important role in complex tasks such as learning, comprehension, and reasoning. There is general agreement about the fact that the prefrontal cortex mediates working memory functions (Fuster 1989). Working memory consists of two subsystems. One system is important for processing language and is called the phonological loop whereas the visuospatial sketch pad is important for processing visuospatial information. These two subsystems are controlled by an attentional control system, the central executive (Baddeley 1998).

Executive functions refer to supervisory abilities involved in the planning, initiation, maintenance, and adjustment of goaldirected behavior. They are activated whenever the routine selection of operations is insufficient and inappropriate. They are thus important for reasoning and problem-solving. The prefrontal cortex is critically involved in the supervision or execution of cognitive processes and is therefore involved in executive control (Mayes and Daum 1997).

Disorders of language are frequently observed after damage to the left hemisphere of the brain. Two areas are of particular importance for language processing. One is Broca's area and lies in the left frontal lobe. This region is essential for the expression of language whereas Wernicke's area, which lies in the temporal lobe, is important for language comprehension. Impairments of verbal abilities, which are affected in the comprehension of language, are referred to as aphasia. Damage to Broca's area leads to difficulties in language production with a milder impairment of comprehension. In Wernicke's aphasia, the main symptom is a deficit in the comprehension of spoken language.

Changes in affect, personality and social behavior are also frequent sequelae of brain dysfunction. Emotional and personality changes can be the direct result of brain damage. Examples are apathy and loss of interests and initiative, but also loss of tact and inappropriate social behavior, which may occur after frontal lesions (Kolb and Wishaw 1996). Emotional problems may also result from insight in the cognitive problems, which occur after brain damage, and the resulting life changes. Emotional and personality changes are most frequently due to an interaction between brain damage and the reaction to disability.

2. Neuropsychological Standard Tests

The most common approach to clinical neuropsychological assessment is the use of standard test batteries such as the 'Wechsler Adult Intelligence Scales' (Lezak 1995) and the 'Luria-Nebraska Neuropsychological Battery (see Lezak 1995). In this approach, the same tests are administered, irrespective of the specific lesion or disorder in an individual patient (Benton 1994). The use of such test batteries has several advantages e.g., normative reference data and the possibility to compare profiles across various neurologic conditions (Levin 1994). A more recent procedure involves the exploration of an individual's clinical history, including detailed information about the site and extent of the lesion, or the severity of the disease and the administration of a series of specific tests. Individual problems and areas of preservation of cognitive function are assessed in relation to specific damage, which allows the development of an individual rehabilitation program.

2.1 Assessment of Global Intellectual Functions

The most widely used test of general abilities is the 'Wechsler Intelligence Scale' (Wechsler 1987). The test includes six verbal tests and five performance tests. An overall IQ score is generally derived from these subtests. In addition, IQ scores can be derived separately from the verbal tests and the performance tests. Damage to the right hemisphere leads to a reduced performance IQ, whereas lesions to the left hemisphere tend to depress the verbal score.

2.2 Visuoperceptual Capacities

Perceptual tests primarily assess visuoperceptual and visuospatial abilities. Clinical investigations have shown that visual impairments are mainly associated with diseases and dysfunction of posterior regions of the right hemisphere.

Visual abilities are frequently assessed by tasks requiring the copy of line drawings. Examples are the complex figure or the 'Benton Visual Retention Test' (see Lezak 1995). Tests of shape, object, and visual space perception are assessed by the 'Visual Object and Space Perception Test' which includes e.g., the recognition of objects from silhouettes or the appreciation of the selective position of dots in a spatial arrangement (Warrington and James 1991).

A variety of tests have been developed for face recognition. The stimulus material is often comprised of familiar and unfamiliar faces. Face recognition is typically impaired in patients with damage of the right hemisphere. Tests, such as the 'Florida Affect Battery,' not only require basic processing of faces e.g., in face matching tasks, but also the recognition of emotional facial expressions (Bowers et al. 1992).

2.3 Attention

A common test for measuring attentional capacity is 'Digit span,' a subtest of the 'Wechsler Intelligence Scale'. Random sequences of numbers are read aloud and have to be reproduced by the subject in forward or backward order. A visuo-spatial analogue is the 'Corsi Block-tapping test'. In this test, the sequence of spatial positions of wooden cubes which are fixed on a board has to be reproduced. Frequently used concentration tests are 'Digit Symbol' a subtest from the 'Wechsler Intelligence' or 'letter cancellation tests'. More recent tests have focused on the assessment of everyday attentional requirements, such as the ability to efficiently check phone books or maps. One of the most widely used ecologically valid instruments is the 'Test of Everyday Attention' by Robertson et al. (1996).

2.4 Memory

Memory profiles (e.g., differences in verbal or visual memory abilities, short-term, or long-term memory) are determined in batteries consisting of a range of different memory tests. A typical example is the 'Wechsler Memory Scale' (Wechsler 1987). Everyday memory abilities form the basis of the 'Rivermead Memory Test' (Wilson et al. 1985). Test demands are e.g., remembering a new route, a newspaper article, a hidden object, or an appointment.

An alternative approach to test batteries is to administer a range of different individual tests. To assess verbal memory, typically word lists, stories, or paired associates are presented, which have to be recalled or recognized immediately or after a delay interval of up to one hour (for a summary see Lezak 1995). Visual or spatial memory is assessed by the immediate and/or delayed recall of geometrical figures, spatial positions, or the recognition of visual patterns or faces.

2.5 Language

Tests of verbal abilities have been developed to tap the functional efficiency of the left hemisphere and to assess aphasia. One example is the vocabulary test, a subtest of the 'Wechsler Intelligence Scale', which assesses the ability to define common concepts. The 'Token Test' was developed to assess comprehension of spoken instructions. It measures the capability to carry out commands of varying complexity (e.g., 'touch the blue square and the white square'). 'Verbal Fluency' is evaluated by the quantity of words that can be produced from a range of categories, e.g., names of countries or animals. Aphasic patients almost always perform badly. In addition, verbal fluency has been shown to be a sensitive indicator of left frontal lesions or neurodegenerative disease.

Brief screening instruments, such as the 'Aphasia Screening Test' and the 'Multilingual Aphasia Examination', allow an efficient assessment of aphasia (see Lezak 1995). The 'Multilingual Aphasia Examination' provides performance profiles concerning different aspects of language such as naming, repetition, verbal fluency, understanding, and spelling. Comprehensive instruments such as the 'Boston Diagnostic Aphasia Examination' (Goodglass and Kaplan 1983) assess a variety of specific aspects of language performance and allow the differential diagnosis of different aphasic syndromes.

2.6 Executive Functions

Frequently used tests of executive functions are the 'Wisconsin Card Sorting Test,' the 'Stroop test' (see Lezak 1995), and the recently developed 'Hayling Test' (Burgess and Shallice 1997). The 'Wisconsin Card Sorting Test' requires the detection of sorting rules (color, shape, and number of elements) and the modification of such rules in response to feedback. The 'Stroop Test' and the 'Hayling Test' assess the ability to suppress habitual responses. In the Stroop test, color words that are printed in a different color, have to be read aloud; the color of the written words has to be inhibited (e.g., the word 'red' printed in green). The Hayling Test requires the completion of sentences by unrelated words, obvious semantically related completions have to be suppressed. Further tests of executive function involve planning or problem solving within the context of complex puzzles such as the 'Tower of London Task' (see Lezak 1995).

2.7 Affect and Personality

Changes in personality or emotional processing play an important role in neuropsychological assessment, since they may affect performance on cognitive tests to a considerable degree. Depression in particular, is often observed after brain damage. The 'Beck Depression Inventory' assesses the degree and severity of depression and is frequently used in neuropsychological assessment (Beck 1987). The 'Minnesota Multiphasic Inventory' measures personality and emotional adjustment and is thought to be sensitive to affective disorders (Butcher et al. 1989) *Minnesota Multiphasic Personality Inventory (MMPI)*. Changes in social behavior and social interactions are inferred on relatives' reports and behavioral observations in the testing situation.

3. Clinical Groups

Within the context of neuropsychological assessment, three main groups of clinical conditions can be differentiated. The first group includes people suffering from localized brain damage due to stroke, head trauma, or infections such as herpes simplex. Their cognitive problems vary and largely depend on the site and extent of the brain damage. The second group comprises people with neurological, mainly neurodegenerative diseases, such as Parkinson's disease, Alzheimer's disease, and multiple sclerosis (see Lezak 1995). In Parkinson's disease the most frequently reported cognitive deficits are executive dysfunction, working memory deficits, and visuo-perceptual problems. Alzheimer's disease is associated with episodic memory impairment, accompanied by language deficits, visual–spatial problems, and changes of personality. In multiple sclerosis the cognitive impairment affects a broad range of functions such as memory, learning, attention, concentration, reasoning, speed of information processing, and executive functions.

In recent research neuropsychological deficits have also been reported in psychiatric diseases, in particular in depression and schizophrenia. These findings make an important contribution to our understanding of the relation between cognitive impairments and psychiatric symptoms, and may help to identify predictors of the course of the illness (Keefe 1995). Neuropsychological impairments in schizophrenia comprise deficits in attention, executive functions, verbal memory, and learning (Keefe 1995). Attentional problems have been attributed to a failure in the selection of relevant from irrelevant emotions and cognitions.

In depression impairments of memory, executive functions, attention, and concentration have been reported. Therefore problems mainly occur on tasks which require effortful processing (Keefe 1995).

4. Specific Problems of Neuropsychological Assessment

Neuropsychological tests are often very demanding and are dependent upon a high degree of compliance, motivation, and cooperation. They require the ability to understand and to follow instructions, and to concentrate for a considerable period of time. The potential influence of nervousness and test anxiety on cognitive performance also needs to be taken into account. The financial, emotional, and social advantages of illness and injuries may encourage malingering which poses a problem for neuropsychological assessment (Lezak 1995). Malingering individuals often show impairments that are greater than would be expected by chance or by a patient's medical and social history. In addition, errors are made that are unrelated to the lesion or complaint. The analysis of inconsistencies with repeated testing may also contribute to the detection of malingering (Lezak 1995). Advances in functional imaging techniques and the combination of neuroimaging and neuropsychological assessment procedures will be of considerable help if psychological and organic aspects of cognitive dysfunction have to

be distinguished (see *Personality Theory and Psychopathology*).

See also: Alzheimer's Disease: Behavioral and Social Aspects; Alzheimer's Disease, Neural Basis of; Attention: Models; Attention, Neural Basis of; Cognitive Control (Executive Functions): Role of Prefrontal Cortex; Cognitive Functions (Normal) and Neuropsychological Deficits, Models of; Cognitive Neuropsychology, Methodology of; Declarative Memory, Neural Basis of; Implicit Learning and Memory: Psychological and Neural Aspects; Perception of Extrapersonal Space: Psychological and Neural Aspects; Prefrontal Cortex; Visual Perception, Neural Basis of; Working Memory, Neural Basis of

Bibliography

Baddeley A 1998 Recent developments in working memory. *Current Opinion in Neurobiology* **8**: 234–438

Beck A T 1987 *Beck Depresion Inventory*. The Psychological Corporation, San Antonio, TX

Benton A L 1994 Neuropsychological assessment. *Annual Review of Psychology* **45**: 1–23

Bowers D, Blonder L X, Heiman K M 1992 *The Florida Effect Battery-Manual* Unpublished manuscript at the Center for Neuropsychological Studies University of Florida, Gainesville FL

Butcher J N, Dahlstrone W G, Graham J R et al. 1989 *Manual for the Restandardized Minnesota Multiphasic Personality Inventory*: MMPI-2. University of Minnesota Press, MN

Burgess P W, Shallice T 1997 *The Hayling and Brixton Tests*. Thames valley Tests Company Limited, Bury St Edmunds, UK

Daum I, Ackermann H 1997 Neuropsychological abnormalities in cerebellar syndromes—fact or fiction? In: Schmahmann J D (ed.) *The Cerebellum and Cognition*. Academic Press, San Diego, CA

Fuster J M 1989 *The Prefrontal Cortex: Anatomy Physiology and Neuropsychology of the Frontal Lobe*. Raven Press, New York

Goodglass H, Kaplan E 1983 *Boston Diagnostic Aphasia Examination BDAE*. Lea and Febinger, Philadelphia

Keefe R S 1995 The contribution of neuropsychology to psychiatry. *American Journal of Psychiatry* **152**: 6–15

Kolb B, Wishaw I Q 1996 *Fundamentals of Human Neuropsychology*. W H Freeman and Company, New York

Levin H S 1994 A guide to clinical neuropsychological testing. *Archives of Neurology* **51**: 854–9

Lezak M D 1995 *Neuropsychological Assessment*. Oxford University Press, Oxford and New York

Maruff P, Pantelis C 1999 Attention and neuropsychiatric disorders. *Current Opinion in Psychiatry* **12**: 339–44

Mayes A R, Daum I 1997 How specific are the memory and other cognitive deficits caused by frontal lobe lesions? In: Rabbitt P (ed.) *Methodology in Frontal and Executive Functions*. Erlbaum, Hove, UK

Robertson I H, Ward T, Ridgeway V, Nimmo-Smith I 1996 The structure of normal human attention: The test of everyday attention. *Journal of the International Neuropsychological Society* **2**: 525–34

Squire L R 1987 *Memory and Brain*. University Press, Oxford, UK

Warrington E K, James M 1991 *Visual Object and Space Perception Battery*. Thames Valley Test Co, UK National Rehabilitation Service, Gaylord, MI

Wechsler D 1987 *Wechsler Memory Scale Revised Manual*. The Psychological Corporation, San Antonio, TX

Wilson B A, Cockburn J, Baddeley A 1985 *The Rivermead Behavioral Memory Test Reading*. Thames Valley Test Co, UK National Rehabilitation Service, Gaylord, MI

J. Uekermann and I. Daum

Neuroticism

Neuroticism is a general dimension of individual differences. It is widely regarded as the single most important trait of human personality. Its importance is reflected in the fact that it has been included in virtually every trait model that has been developed during the twentieth century (see *Personality Structure*). This entry examines basic characteristics of the trait and then considers its broader significance.

1. Basic Aspects of the Dimension

Personality traits can be ordered hierarchically, ranging from narrow lower-order characteristics to very broad higher-order dimensions. Neuroticism is one of a small number of general traits that comprise the highest level of this hierarchy. This trait represents individual differences in the extent to which a person experiences the world as threatening, problematic, and distressing (Watson et al. 1994). High scorers on this dimension experience various negative emotions (e.g., anxiety, depression, anger, guilt, shame) more frequently and intensely than low scorers. They report suffering from a wide variety of problems (including minor health problems) and generally blame themselves for these problems; because of this, they tend to feel inadequate and inferior. Neurotic individuals are both self-critical and excessively sensitive to criticism, and they easily feel slighted. They report elevated levels of stress in their lives (see *Stress: Psychological Perspectives*), and indicate that they cope poorly with that stress. Finally, they tend to view the world with a negativistic bias, and so see threats, problems, and crises where others do not. In contrast, low scorers on this dimension are self-satisfied and view the world as benign; they report few problems, low levels of negative emotions, and little stress.

1.1 Historical Development of the Concept

As noted earlier, a neuroticism-like dimension has been included in virtually every prominent trait model. For instance, Cattell included the higher-order

dimension of Adjustment-vs.-Anxiety in his highly influential taxonomy of personality. However, the trait was given many different names by personality researchers, and it is noteworthy that most of the early labels emphasized the relaxed, self-satisfied end of the dimension. For instance, the trait was called 'Emotional Control,' and 'Emotional Stability' by Fiske and Guilford, respectively. Similarly, the dimension was usually labeled 'Emotional Stability' in the early literature on the 'Big Five' or 'five-factor' model of personality (Wiggins and Trapnell 1997). The term 'neuroticism' itself was first popularized in the 1950s by the influential English psychologist Hans Eysenck, who included a scale with that name in his Maudsley Personality Inventory.

This diversity of labels caused confusion in the field for several decades, because it was unclear whether all of these concepts actually reflected the same basic trait. However, integrative research in the 1980s clearly established that the same basic dimension was being assessed in all of these models (McCrae and Costa 1985, Watson and Clark 1984, Watson et al. 1994). This integration, in turn, stimulated an explosion of research on the trait during the late 1980s and 1990s. During this time, 'neuroticism' emerged as the preeminent label for the dimension, due in large part to a growing awareness of its prominent clinical correlates (which are discussed below).

1.2 Origins and Course of the Trait

The genetic basis of neuroticism has been extensively investigated (Clark and Watson 1999). Heritability estimates based on twin studies typically fall in the 0.40 to 0.60 range, with a median value of approximately 0.50 (see *Behavioral Genetics: Psychological Perspectives*). It also should be noted, however, that heritability estimates decline significantly with age and are consistently lower in older respondents. The remaining variance can be attributed entirely to the unshared environment; indeed, the evidence overwhelmingly suggests that the common rearing environment (i.e., the effects of living together in the same household) exerts virtually no effect on the development of this trait. The nature of this genetic component has not yet been clearly established. The current thinking in the field, however, is that it gives rise to individual differences in sensitivity to painful, punishing, and stressful stimuli; that is, high scorers on the trait are innately disposed to show stronger adverse reactions to noxious and threatening stimuli. In support of this view, studies have shown that neurotic individuals report larger increases in negative emotion in response to unpleasant mood inductions. In light of this evidence, it is not surprising that neuroticism scores are strongly stable over time (see Clark and Watson 1999). Stability correlations typically fall in the 0.70 to 0.80 range over retest intervals of five to 10 years. Stability coefficients still exceed 0.60 over much longer time spans; Helson and Klohnen (1998), for instance, obtained a 25-year stability correlation of 0.65 in women assessed at ages 27 and 52. However, this stability is superimposed on a significant age-related decline (see *Personality Development in Adulthood*). Specifically, neuroticism scores peak in late adolescence and then begin a long, slow decline that continues at least into middle adulthood (Clark and Watson 1999, Helson and Klohnen 1998). After that, the evidence is less clear, but it appears that scores on the trait stabilize and show no further change. There also is a modest sex difference, with women scoring slightly higher on the trait than men (see *Gender Differences in Personality and Social Behavior*).

2. Broader Significance of the Trait

As stated previously, neuroticism is widely regarded to be the single most important trait of human personality. This view reflects the fact that this dimension has important correlates across a broad range of functioning. This section explores three particularly rich research areas: (a) psychopathology, (b) job and marital satisfaction, and (c) health.

2.1 Neuroticism and Psychopathology

Neuroticism has stronger, clearer, and broader associations with psychopathology than any other personality trait (Clark and Watson 1999, Mineka et al. 1998, Widiger et al. 1999). Elevated neuroticism has been linked to a wide array of clinical syndromes, including anxiety disorders, mood disorders, substance use disorders, somatoform disorders, eating disorders, personality and conduct disorders, and schizophrenia. Indeed, it is increasingly apparent that most psychological disorders are associated with elevated scores on this dimension. The breadth of these links is unsurprising, given that emotional distress is one of the defining criteria of both neuroticism and psychological disorder. Neuroticism is particularly strongly related to clinical syndromes that contain a prominent component of subjective distress. For instance, the trait has extremely strong associations with major depression, generalized anxiety disorder, and borderline personality disorder. These links are so strong, in fact, that syndromes such as generalized anxiety disorder and borderline personality disorder essentially may represent extremely high levels of neuroticism (Clark and Watson 1999, Widiger et al. 1999). In support of this argument, recent twin study evidence strongly suggests that neuroticism, generalized anxiety disorder, and major depression are genetically indistinguishable; in other words, they can all be

traced to a single genetic diathesis that apparently reflects an underlying vulnerability to subjective distress and negative emotionality (Mineka et al. 1998).

In contrast, neuroticism is more weakly related to disorders that are not centered on the experience of emotional distress. For instance, individuals with specific phobias (e.g., fear of spiders or other small animals) show only modest elevations on the trait (Mineka et al. 1998).

2.2 Neuroticism and Dissatisfaction

High scorers on this dimension report a pervasive dissatisfaction with themselves and the world around them. This dissatisfaction influences almost every aspect of their lives. For instance, numerous studies have shown that neurotic individuals report greater marital dissatisfaction and show greater instability (e.g., increased risk of divorce) in their relationships (Karney and Bradbury 1995). It is noteworthy that much of this evidence is longitudinal, demonstrating that neuroticism scores obtained early in a relationship predict its subsequent course. The most striking evidence along these lines was reported by Kelly and Conley (1987), who followed a large sample of couples from the mid-1930s through the early 1980s. Their results indicated that respondents who initially were high in neuroticism (assessed between 1936 and 1941) were more likely to become divorced over the course of the study. Moreover, neuroticism significantly predicted subsequent levels of marital dissatisfaction assessed in 1954–5 and 1981. Thus, the trait showed significant predictive power over a time span of more than 40 years (see *Personality and Marriage*).

Similarly, several studies have shown that high scorers on the trait report significantly greater job dissatisfaction (Judge et al. 1998). Again, much of the relevant evidence is longitudinal. Most notably, Staw et al. (1986) found that a neuroticism measure assessed in adolescence was a significant predictor of job satisfaction nearly 50 years later, even after controlling for objective differences in work conditions.

2.3 Neuroticism and Health

Throughout much of the twentieth century, the 'psychosomatic hypothesis'—that is, the idea that chronically elevated levels of negative emotion eventually will lead to illness—was a dominant theoretical paradigm in behavioral medicine. Given that neurotic individuals report more frequent and intense episodes of negative emotion, it was widely assumed that neuroticism was a significant vulnerability factor for physical illness. Consistent with this assumption, neuroticism scales consistently show significant, moderate correlations with measures of physical symptoms and health complaints. Moreover, this association is quite broad, with neurotic individuals reporting increased levels of headaches, chest pain, dizziness, nausea, coughing, sneezing, congestion, and numerous other problems (see Watson and Pennebaker 1989).

However, research conducted during the 1970s and 1980s indicated that neuroticism was not strongly or consistently related to objective indicators of health status, such as health-related behaviors (e.g., number of physician visits for illness), biological markers (e.g., immune system functioning), and health outcomes (e.g., mortality rates). In other words, neuroticism is a much better predictor of subjective health complaints than it is of objective health status. Because of this, it is now widely recognized that neuroticism is an important nuisance factor that must be assessed and analyzed in health research.

3. Future Directions

Neuroticism clearly has important implications for almost every aspect of human functioning. However, researchers still know relatively little about how this trait actually works. How do innate genetic/biological differences give rise to such striking individual differences in distress and dissatisfaction? What maladaptive cognitive processes and behavioral strategies contribute to the obvious misery of neurotic individuals, and can they be eliminated through psychological treatment? These questions will drive the next generation of research on neuroticism.

See also: Genetic Studies of Personality; Personality and Adaptive Behaviors; Personality Development in Adulthood; Personality Psychology; Personality Structure; Personality Theories; Personality Theory and Psychopathology

Bibliography

Clark L A, Watson D 1999 Temperament: A new paradigm for trait psychology. In: Pervin L A, John O P (eds.) *Handbook of Personality: Theory and Research*, 2nd edn. Guilford, New York, pp. 399–423

Helson R, Klohnen E C 1998 Affective coloring of personality from young adulthood to midlife. *Personality and Social Psychology Bulletin* **24**: 241–52

Judge T A, Locke E A, Durham C C, Kluger A N 1998 Dispositional effects on job and life satisfaction: The role of core evaluations. *Journal of Applied Psychology* **83**: 17–34

Karney B R, Bradbury T N 1995 The longitudinal course of marital quality and stability: A review of theory, method, and research. *Psychological Bulletin* **118**: 3–34

Kelly E L, Conley J J 1987 Personality and compatibility: A prospectius analysis of material stability and marital satis-

faction. *Journal of Personality and Social Psychology* **52**: 27–40
McCrae R R, Costa Jr P T. 1985 Comparison of EPI and psychoticism scales with measures of the five-factor model of personality. *Personality and Individual Differences* **6**: 587–97
Mineka S, Watson D, Clark L A 1998 Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology* **49**: 377–412
Staw B M, Bell N E, Clausen J A 1986 The dispositional approach to job attitudes: A lifetime longitudinal test. *Administrative Science Quarterly* **31**: 56–77
Watson D, Clark L A 1984 Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin* **96**: 465–90
Watson D, Pennebaker J W 1989 Health complaints, stress, and distress: Exploring the central role of negative affectivity. *Psychological Review* **96**: 234–54
Watson D, Clark L A, Harkness A R 1994 Structures of personality and their relevance to psychopathology. *Journal of Abnormal Psychology* **103**: 18–31
Widiger T A, Verheul R, Van den Brink W 1999 Personality and psychopathology. In: Pervin L A, John O P (eds.) *Handbook of Personality Theory and Research*, 2nd edn. Guilford, New York, pp. 347–66
Wiggins J S, Trapnell P D 1997 Personality structure: The return of the Big Five. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 737–65

D. Watson

Neurotransmitters

At one time it was believed that neurons signalled each other or their cellular targets by electrical means. However, with an accumulation of experimental evidence it became clear that although electrical transmission does occur, by far the major mode of synaptic transmission was due to the release of chemicals, referred to as neurotransmitters.

1. Chemical Neurotransmitters in the Nervous System

Although suggestions of chemical transmitters were made by DuBois-Raymond, Elliott Dixon, and Langley, it remained for Otto Loewi in 1921 to directly demonstrate the release of neurotransmitters from nerve endings. Loewi's inhibitory chemical, which was released following vagal stimulation of the heart, was called *vagusstoff* and subsequently shown to be acetylcholine. Later, acetylcholine (ACh) was identified as the neurotransmitter at a variety of peripheral synapses such as the neuromuscular junction of skeletal muscle, cardiac muscle, visceral smooth muscle, and at preganglionic synapses in the autonomic nervous system. However, to identify ACh or any other neurotransmitter in the central nervous system (CNS) was considerably more difficult since one could not stimulate a specific neuron and collect the exudate. Eventually, with the development of electrophysiological and neurochemical techniques and the designation of a set of criteria to define a neurotransmitter molecule, a number of these chemicals have now been identified. These will be individually discussed after we examine the rigid criteria for a neurotransmitter.

(a) The transmitter must be synthesized and stored in the presynaptic terminal. Experimental evidence that satisfies this criterion include cytochemical and subcellular biochemical evidence and a demonstration of the loss of the suspected transmitter following surgical or chemical lesioning of the presynaptic neuron.

(b) Stimulation of the presynaptic neuron should release the transmitter. With current sophisticated techniques, it is possible to collect and identify the released transmitter.

(c) Application of the putative transmitter to the synapse should mimic the response of electrical stimulation of presynaptic nerve tracts. Conversely, drugs that block the response of the electrically stimulated release of the transmitter should also block the response of the applied suspected transmitter.

(d) A mechanism must exist for the rapid termination of transmitter activity. This would be either enzymatic catabolism or a high affinity reuptake system for the transmitter either in the presynaptic terminal or adjacent glial cells.

Before discussing the certified neurotransmitters, it is appropriate to clarify some confusion in the literature on the definitions of neurotransmitters, neuromodulators, and neurohormones. The criteria for classifying a compound as a neurotransmitter have already been discussed. Neuromodulators, as the name implies, do not have intrinsic activity but modulate ongoing synaptic activity. Thus, if transmitter release at a synapse was prevented, the addition of a neuromodulator to the preparation would have no effect. Neuromodulators may affect synaptic transmission, by acting on presynaptic receptors, or the synthesis, release, reuptake, or metabolism of transmitters. In contrast to neuromodulators, neurohormones do have intrinsic activity but differ from neurotransmitters in that following release from either a neuronal or glial source, they travel in some circulation to produce their effects at a distance from their release site.

Having defined these terms it is necessary to point out that it is a futile exercise to classify a neuroactive agent as a transmitter, modulator, or hormone until one knows its activity and its site of action, since these agents can play different roles. Thus, for example, dopamine is a certified transmitter in the striatum, yet it can also be released from the hypothalamus, and travel in the hypophyseal–portal circulation to act on the pituitary to inhibit the release of prolactin. In this

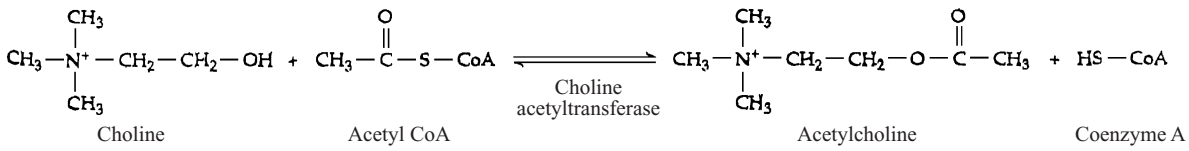

Figure 1
Synthesis of acetylcholine.

case dopamine is obviously acting as a neurohormone. Similar examples exist amongst the other transmitters to be discussed such as norepinephrine, serotonin, and acetylcholine.

Another class of agents that causes confusion in the literature are the neuroactive peptides. These compounds, often referred to as co-transmitters, can be released from nerve terminals to affect synaptic transmission. For example, vasoactive intestinal polypeptide (VIP) is released from the salivary gland along with ACh and increases secretion while simultaneously increasing the affinity of ACh for the cholinergic receptor in the salivary gland. Another peptide, substance P, is currently thought to be the actual neurotransmitter in the spinal cord rather than just being a modulatory agent.

As noted earlier, the first neurotransmitter to be identified was acetylcholine (ACh). ACh is synthesized from acetyl CoA and choline, catalyzed by the enzyme choline acetyltransferase (Fig. 1). The transmitter activity of ACh is terminated by hydrolysis of ACh to acetate and choline, catalyzed by the enzyme acetylcholinesterase. (This is the enzyme, incidentally, that is inhibited by the so-called 'nerve gases'.)

In contrast to the other neurotransmitters to be discussed, the rate-limiting step in the synthesis of ACh is the uptake of choline into the presynaptic terminal. In the terminals, ACh, like the other transmitters, is stored in small sacs, referred to as synaptic vesicles, from which it is released by an exocytotic mechanism (see below).

In the peripheral nervous system, ACh is the transmitter at autonomic ganglia, at parasympathetic postganglionic synapses, and at the neuromuscular junction. In the CNS, experimentally a more difficult area in which to localize neurotransmitter activity, major cholinergic tracts originate in the medial septal nucleus, the vertical limbus nucleus of the diagonal band, the horizontal nucleus of the diagonal band, and the nucleus basalis of the substantia inominata. These tracts respectively innervate the hippocampus, the olfactory bulb, and the cerebral cortex and amygdala. Cholinergic interneurons are found in the striatum. The first cholinergic tracts that were identified were the motoneuron collaterals to the Renshaw cells in the spinal cord.

When ACh is released from nerve terminals, it exerts its physiological effects by acting on two different types of receptor, muscarinic and nicotinic. (Each of these exist in multiple sub-types which when activated produce different responses. See XXX.)

The major disease associated with ACh is myasthenia gravis, an autoimmune disease whereby circulating antibodies affect the ACh receptor at the neuromuscular junction. Circumstantial evidence implicates ACh in cognitive functions, particularly in the early stages of Alzheimer's Disease.

2. *Biogenic Amines*

Four neurotransmitters come under the chemical classification of biogenic amines. These are epinephrine, norepinephrine, dopamine, and serotonin. Although epinephrine is the transmitter in frogs, in mammals its role has been supplanted by norepinephrine. Epinephrine's function in the mammalian brain is still unclear and may be limited to a hormonal role.

Starting with tyrosine, the catecholamines (norepinephrine and dopamine) are synthesized in a cascade of reactions beginning with the rate-limiting enzyme, tyrosine hydroxylase. Figure 2 depicts the enzymes and cofactors involved. The catecholamines are catabolized by two enzymatic pathways (Fig. 3) involving monamine oxidase, a neuronal mitochondrial enzyme, and catechol-o-methyltransferase, a cytoplasmic enzyme, found primarily in the kidney and the liver. However, as noted earlier, when norepinephrine and dopamine are released into the synapse, their activity is terminated by reuptake into the presynaptic terminal rather than by enzymatic catabolism. The reuptake is inhibited by a number of antidepressant drugs.

Noradrenergic neurons arise from the locus coeruleus, the lateral tegmental system, and a dorsal medullary group and innervate virtually all areas of the brain and spinal cord. Central effects of noradrenaline stimulation are not clear but appear to involve behavioral attention and reactivity.

Peripherally where noradrenaline is released from postganglionic sympathetic neurons of the autonomic nervous system, the major effects are to regulate blood pressure, relax bronchi, and relieve nasal congestion. These effects are mediated by the major receptors, α and β, each again with multiple subtypes.

At one time dopamine was thought to be just an intermediate in the conversion of tyrosine to noradrenaline. It is now clear, however, that dopamine is

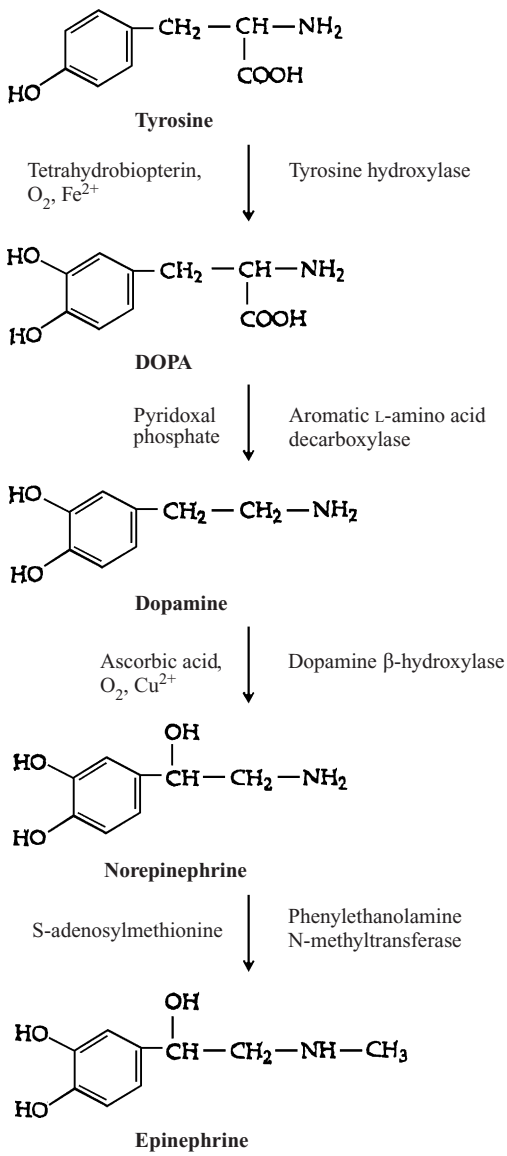

Figure 2
Synthesis of the catecholamine transmitters. The synthesizing enzymes are shown to the right of each arrow, and the enzyme cofactors to the left.

a major player in the CNS with its implication in Parkinson's disease and in schizophrenia. Dopamine cells originate in the substantia nigra, ventral tegmental area, caudal thalamus, periventricular hypothalamus, and olfactory bulb. Dopaminergic terminals are found in the basal ganglia, the nucleus acumbens, the olfactory tubercle, the amygdala, and the frontal cortex. The nigrostriatal pathway is particularly important since its degeneration is involved in Parkinson's disease. Initially, dopamine receptors were classified as D_1 or D_2. Currently the subtypes consist of D_1 through D_5 with the possibility of a D_6. All the receptors are coupled to G proteins as their second messenger. Arising from the observation that a correlation existed between therapeutic doses of antipsychotic drugs and inhibition of binding of dopamine receptor antagonists, the D_2 receptor has been fingered in the pathophysiology of schizophrenia. The atypical neuroleptic drug clozapine, however, exhibits a greater affinity for the D_4 receptor, dopaminergic transmission in the nucleus accumbens, involving both D_1 and D_2 receptors, is believed to be involved in the reward activity of abused drugs such as cocaine. The catabolism of dopamine is shown in Fig. 4.

The last of the biogenic amine neurotransmitters to be discussed is serotonin (5-hydroxytryptamine). Its synthesis and its catabolism are depicted in Figs. 5 and 6. In addition to its presence in the CNS, serotonin is found in the GI tract and in blood platelets. It is also localized in the pineal gland where it serves as a precursor to the hormone melatonin. Serotinergic neurons innervate the limbic system, the neostriatum, cerebral and cerebellar cortex and the thalamus. Currently, 18 serotonin receptor subtypes have been identified. Most are G-protein linked except for the 5-HT_3 receptor which is ligand gated. Hallucinogen drugs have been shown to act on the 5-HT_{2A} receptors. Serotonin receptor antagonists that are relatively specific have been used to treat migraine headaches, body-weight regulation and obsessive–compulsive disorders.

Decarboxylation of the amino acid histidine results in the formation of histamine, a still questionable neurotransmitter. This amine does not qualify as a transmitter according to the rigid definitions outlined earlier, since no evidence exists for either its release on stimulation of a neuronal tract, nor is there a rapid reuptake mechanism or enzymatic catabolism to terminate its activity. Histaminergic neurons are located almost exclusively in the ventral posterior hypothalamus and project throughout the entire CNS. Three histamine receptors have been described, H_1, H_2 and H_3. Antagonists of H_1 are the well-known antihistamine drugs which exhibit a sedative action. H_2 antagonists are used to block gastric acid secretion. H_3 receptors are autoreceptors which, when activated, inhibit the release of histamine.

3. *Amino Acid Transmitters*

Acetylcholine and the biogenic amines discussed above account for only a small percentage of neurotransmitter activity in the brain. By far the most prevalent transmitters are the amino acid transmitters, with glutamate in particular accounting for about one-third of synaptic activity. The amino acid transmitters

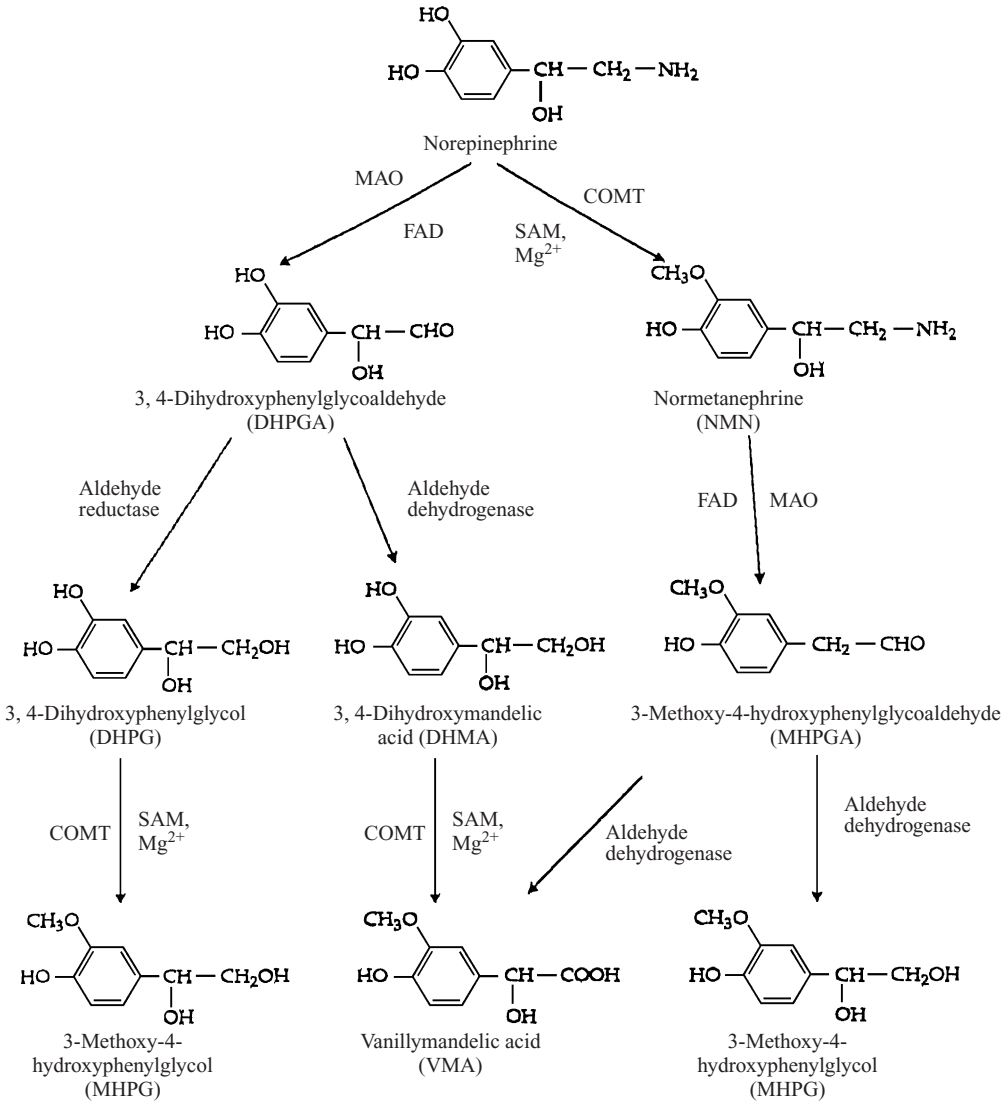

Figure 3
Catabolism of norepinephrine follows multiple pathways depending upon whether the process begins with deamination (left pathway) or O-methylation (right pathway). Aldehyde products of MOA (DHPGA and MHPGA) may be reduced to DHPG and MHPG, or oxidized to DHMA and VMA.

are glutamate (and possibly aspartate), γ-aminobutyric acid (GABA) and glycine.

Glutamate is synthesized by transamination of χ-ketoglutarate (Fig. 7) which is produced in the citric acid cycle, which in turn originates from glucose via glycolysis. Aside from its neurotransmitter activity, glutamate, with the highest concentration of any amino acid in brain, is involved in peptide and protein synthesis, in the detoxification of ammonia in the brain, and as a precursor of the major inhibitory transmitter in the brain, GABA. With so many roles to play, it is easy to understand why the amino acid is compartmentalized, with one pool involved in intermediary metabolism, and the transmitter pool sequestered in synaptic vesicles. Curiously, glutamate that is released on nerve stimulation is taken up into astrocytes where it is converted to glutamine via glutamine synthetase (Fig. 7). It then diffuses out of these glial cells to neuronal terminals where it is hydrolyzed to glutamate via glutaminase, and stored in vesicles ready for release. Glutamate and aspartate are the major fast excitatory transmitters in the CNS

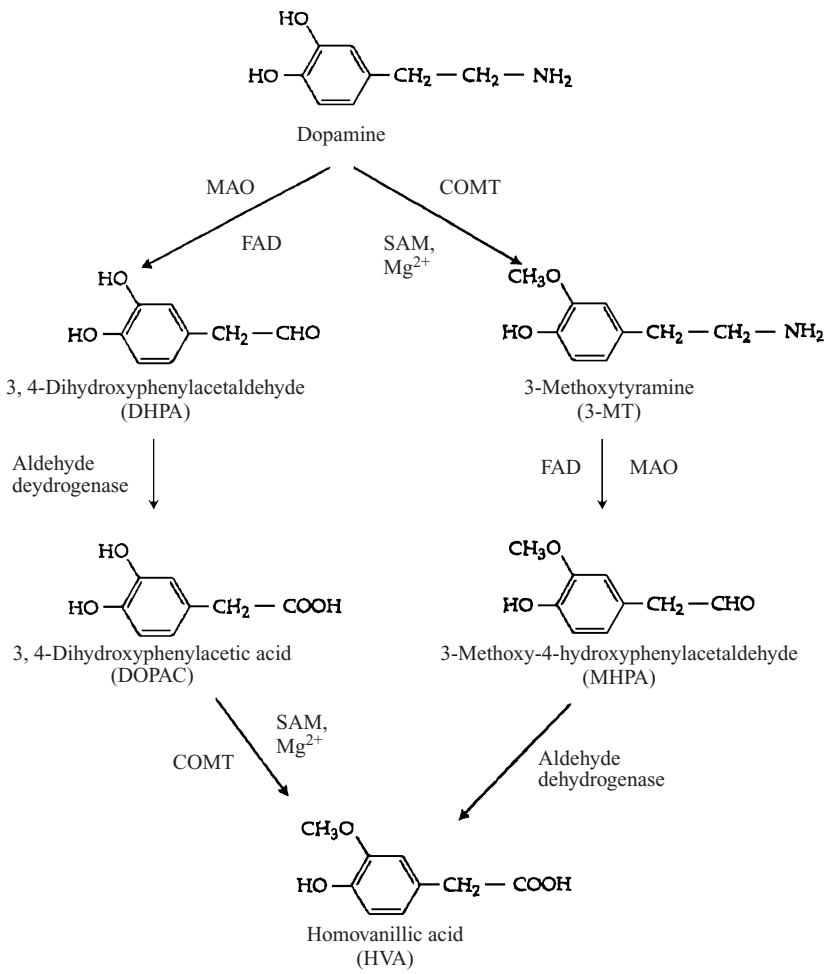

Figure 4
Catabolism of dopamine by (left pathway) oxidative deamination by MAO or (right pathway) O-methylation. DOPAC and 3-MT are indicators of intraneuronal catabolism and of catabolism of released DA, respectively.

and the spinal cord, with a widespread distribution. Currently, the case for aspartate as a neurotransmitter is not entirely clear, though it may have a distinct role in certain cortical and hippocampal pathways.

There are currently four different subtypes of excitatory amino acid receptors, referred to as AMPA, kainate, NMDA and ACPD, comprising a total of fourteen receptors. The first three are ionotropic (ligand gated) receptors and the fourth is metabotropic (G-protein linked). Glutamate as a neurotransmitter has been implicated in excitotoxicity, i.e., as a neurotoxin involved in neurological disorders. Also, glutamate has been shown to play a role in long-term potentiation that may underlie memory and learning.

Two other amino acid transmitters are GABA and glycine, both of which are inhibitory transmitters. The synthesis of GABA from glutamate and its catabolism is depicted in Fig. 8. It has been suggested that 10 to 40 percent of terminals in the cerebral cortex, hippocampus, and substantia nigra use GABA as the major inhibitory transmitter. Other GABAergic areas are the cerebellum, striatum, globus pallidus, and olfactory bulb, as well as the spinal cord.

The two main GABA receptors, each with subtypes are $GABA_A$ and $GABA_B$. $GABA_A$ is a ligand-gated chloride anion channel whereas $GABA_B$, localized primarily at presynaptic terminals, is a G-protein-linked receptor. Anxiolytic drugs such as diazepam (Valium) bind to the $GABA_A$ receptor to modulate its activity. Certain antiepieptic drugs appear to operate by increasing GABAergic activity.

Glycine, a nonessential amino acid, is the second major inhibitory transmitter in the CNS. Serine, arising from glucose, is thought to be the major source

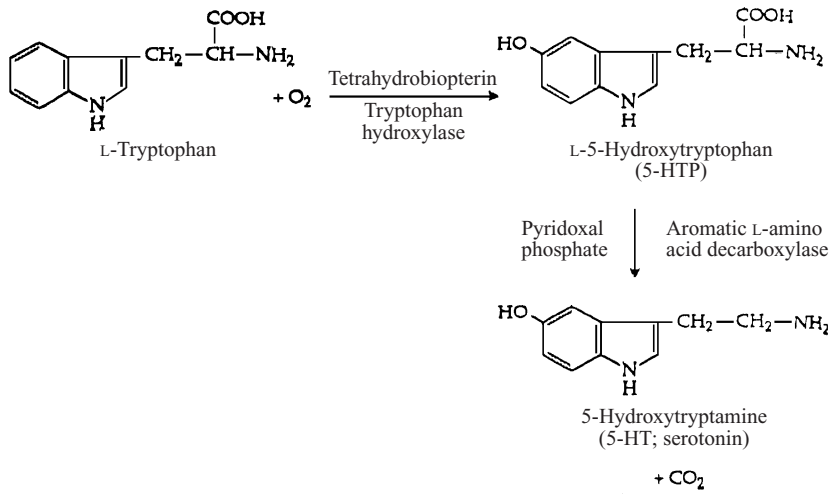

Figure 5
Serotonin (5-HT) synthesized from the amino acid tryptophan in two steps, catalyzed by the enzymes tryptophan hydroxylase and aromatic L-amino acid decarboxylase. The cofactor for each reaction is shown.

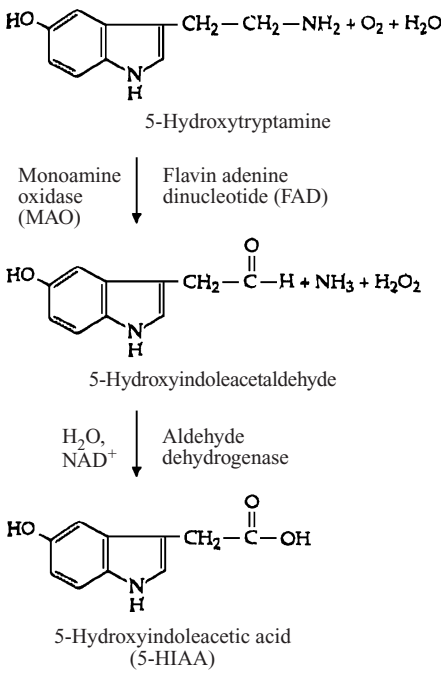

Figure 6
Serotonin is initially catabolized by the mitochondrial enzyme monoamine oxidase to yield the intermediate 5-hydroxyindoleacetaldehyde, which is rapidly converted to 5-hydroxyindoleacitic acid by the enzyme aldehyde dehydrogenase.

of glycine in neurons (Fig. 9). The conversion of serine to glycine is catalyzed by serine hydroxymethyltransferase with tetrahydrofolate as cofactor. As a neurotransmitter, glycine operates almost exclusively in the spinal cord, with a possible involvement in the lower brain stem and the retina. The glycine receptor is similar to the $GABA_A$ receptor in that it promotes hyperpolarization by increasing chloride conductance. The glycine receptor can be identified by the binding of the convulsant drug strychnine. This observation supports the presumed role of glycine in sensorimotor systems and is further supported by the production of spastic mice who have a genetic defect in the glycine receptor. Recently a new role for glycine has arisen with the finding that the amino acid regulates the NMDA (glutaminergic) receptor.

4. *Release and Reuptake of Neurotransmitters*

Although covered in other articles, two major aspects involving neurotransmiters should be mentioned. These are the mechanisms involved in the release of the transmitters and the reuptake mechanism involving particularly the biogenic amines. With regard to transmitter release, as noted earlier, the agents are sequestered in synaptic vesicles. Briefly, when the neuron is depolarized, calcium enters the terminal and the vesicles at the presynaptic release sites fuse with the plasma membrane and open to release the transmitter. A large number of polypeptides is involved in this complex process that involves both release and reconstitution of the plasma membrane and repletion and recycling of vesicles.

The second process is the role of transporters that are involved in the reuptake of transmitters into the

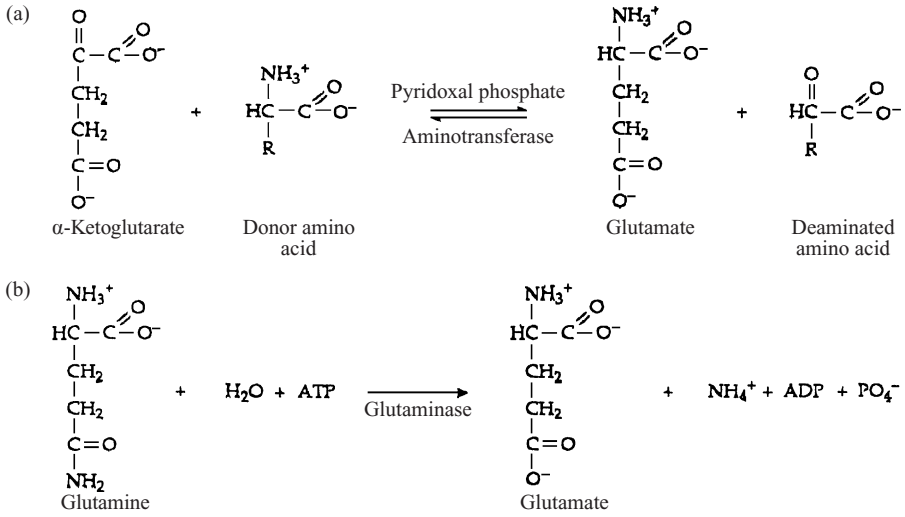

Figure 7
Glutamate is synthesized in neural tissues by two major pathways, (a) from α-ketoglutarate and donor amino acids by means of transamination, and (b) from glutamine by the action of the enzyme glutaminase.

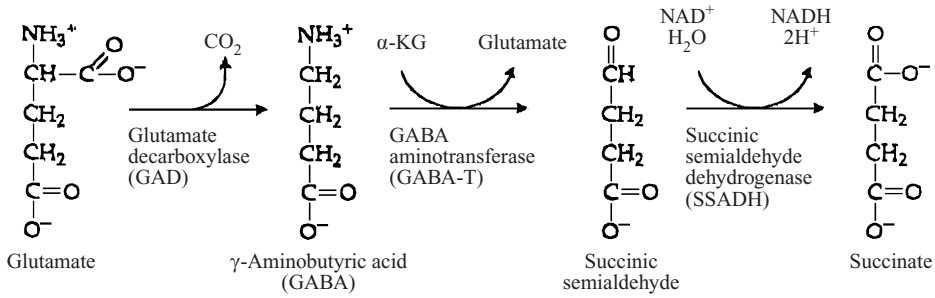

Figure 8
GABA synthesis from glutamate by the enzyme glutamate decarboxylase. Breakdown is catalyzed by GABA-aminotransferase, which forms succinic semialdehyde and regenerates the precursor glutamate from α-ketoglutarate. Most of the succinic semialdehyde is converted to succinate.

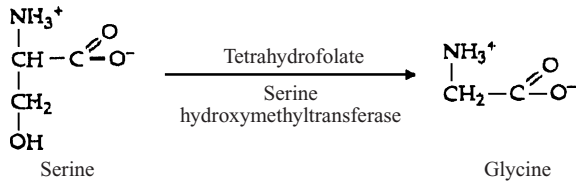

Figure 9
Glycine synthesis from the amino acid serine, catalyzed by the enzyme serine hydroxymethyltransferase. Tetrahydrofolate and pyridoxal phosphate are required as co-factors.

presynaptic terminals. Plasma membrane transporter proteins have been cloned for all the transmitters (but not for choline) and utilized an ATP-dependent process which also requires an Na^+ and Cl^- as co-transporters. These transporters are different from those that pump the transmitters into vesicles; the vesicle transporters utilize an ATP-dependent proton pump mechanism.

The agents that are covered here are the certified neurotransmitters. However, it should be kept in mind

that in addition to neuropeptides, which at certain sites appear to act as transmitters rather than as neuromodulators, a variety of compounds has also been suggested as having neurotransmitter roles. These would include adenosine, ATP, and a variety of non-peptide hormones. There is very little that is absolute in neuroscience.

See also: Synaptic Efficacy, Regulation of; Synaptic Transmission; Synapse Ultrastructure

Bibliography

Amara S G, Arriza J L 1993 Neurotransmitter transporters. *Current Opinion in Neurobiology* **3**: 337–44

Cooper J R, Bloom F E, Roth R H 1996 *The Biochemical Basis of Neuropharmacology*, 7th edn. Oxford University Press, New York

Cotman C W et al. 1994 Excitatory aminoacid transmission. In: Bloom F E, Kupfer D J (eds.) *Psychopharmacology: the Fourth Generation of Progress*. Raven Press, New York, pp. 75–85

Feldman R S, Meyer J S, Quenzer L F 1997 *Principles of Neuropsychopharmacology*. Sinauer Association, Sunderland, MA

J. R. Cooper

Neutrality: Axiological

'Axiological neutrality'—or 'ethical neutrality'—is an imperative imposed on social scientists. It requires of them, when they make policy prescriptions or recommendations, to distinguish clearly between on the one hand the purely scientific content of their presentation, i.e., the facts they are seeking to explain and the theoretical explanations themselves, and on the other hand the ethical value judgments necessary to reach these prescriptions. This distinction cannot be avoided, since social science theories as such are totally unable to express ethical value judgments—judgments about what social states are good or bad, just or unjust, desirable or undesirable. The axiological neutrality imperative would be transgressed by a social scientist who pretended to 'prove scientifically' the desirability or undesirability of a state of affairs. Such a 'proof' would necessarily rest upon the concealment—involuntary or not—of implicit ethical value judgments.

1. The Logical Foundation of Axiological Neutrality

The source of axiological neutrality is to be found in a double logical impossibility:

(a) the impossibility to deduce a value judgment from factual propositions;

(b) the impossibility to deduce a factual proposition from value judgments.

A factual proposition is true if it corresponds—and false if it does not correspond—to the empirical phenomenon that it is supposed to describe. A value judgment ('social state A is better than social state B'), on the other hand, is a *preference* between social facts, and it does not 'correspond' to any empirical phenomenon: it cannot be said 'true' or 'false.' Since a proposition logically derived from factual propositions is also a factual proposition (true or false), it cannot be a value judgment. Conversely, a proposition logically derived from a value judgment is also a value judgment (neither true nor false), and thus cannot be a factual proposition (Ayer 1936).

The consequence of this dichotomy between facts and values is that social science theories *alone* are unable to express value judgments. These theories are supposed to provide causal (deductive) explanations of social facts. They exclusively consist of factual and logical propositions: their premises cannot contain any value judgment, because it would prevent them from explaining empirical phenomena (impossibility (b) above). And it is therefore impossible to deduce from these theories any value judgment (impossibility (a)). From this point of view, social science is no different from any other empirical science. In and of itself, a theory seeking to explain the origin of an inequality cannot yield a value judgment approving of or condemning morally this inequality. For the same reason, a physical theory explaining the solidity of a building cannot yield a value judgment in favor of or against this solidity.

In order to prevent a possible misunderstanding, it must be observed that the factual propositions making up the theories of the social sciences may be propositions *describing* the moral preferences, the esthetic judgments, or the tastes of the agents of the system under study. But when the social scientist describes or explains these preferences, they are considered as mere facts (the fact that an agent has a given preference). It may of course happen that he or she approves of or condemns these preferences, but this value judgment plays no role whatsoever in the explanatory (deductive) theories themselves.

As a consequence of the double impossibility above, social scientists cannot prescribe or recommend a policy—unless they *combine* the scientific facts and theories with an ethical value judgment specifying the end(s) to be achieved. Given an end E, scientific theories can in principle determine the means M that can be used to reach it (or they can tell us whether this end is impossible to reach). Actually, the proposition 'If M, then E'—'If the means M are used, then the end E will be reached'—is a factual proposition: it can in principle be said to be 'true' or 'false.' On the other hand, propositions such as 'It is morally desirable to

attain end *E*' and 'It is morally desirable to use means *M*' are value judgments: they are not factual propositions, and are neither true nor false. In other words, a social scientist can use scientific knowledge of cause and effect in order to determine the means to an end, but not in order to advocate this end or these means from a moral point of view (Simon 1976). It goes without saying that if the theories are wrong or the relevant facts not fully known, the means in question will (generally) not actually permit one to reach the end.

In the 1940s, the founders of the 'New Welfare Economics' (economists such as John Hicks and Paul Samuelson) believed that they could overcome the dichotomy between facts and values by resorting to the unanimity principle; that is, by using only value judgments accepted by all the members of the society ('Pareto-optimality' seemed to be such a unanimous value judgment). Against this view, a number of objections have been raised. First, even if a value judgment is unanimous, it remains a value judgment and can never be ethically neutral (Sen 1970). Second, really unanimous (and thus nonconflicting) value judgments are probably less numerous than economists would like to believe. Last, such value judgments are morally questionable insofar as they tend to promote the status quo (Rothbard 1997).

2. *Axiological Neutrality as a Moral Imperative*

This dichotomy between facts and values implies that social scientists have a responsibility towards their audiences, be it their scientific colleagues, the general public (which they may reach through the media), or more especially the 'captive' audience of their students. The imperative of axiological neutrality commands social scientists not to make their audience believe that science alone—as a system of factual and logical propositions—can demonstrate the value of their own moral or political convictions. The social scientist, in other words, must not indulge in moral teaching while pretending to be engaged in pure empirical science. In his classic text about axiological neutrality, Weber (1917) emphasized that the professors of social sciences had to respect this imperative, first by being aware of it themselves, and second by making their audience and especially their students aware of the ethical value judgments founding their prescriptions. When social science theories are put into practice, value judgments are necessarily required. Such judgments are not scientific (empirical) propositions, and their existence and content should be explicitly stated—especially when these propositions may elicit interpersonal disagreement.

Weber went very far in the defense of neutrality since he recommended that professors confine themselves to purely scientific presentations, and if possible totally refrain from expressing their ethical preferences in front of their students. But on this point, his conception of axiological neutrality was too demanding. For if empirical theories are devoid of value judgments, the latter cannot be excluded from scientific *activity*. A social scientist cannot do science—try and discover the true explanations of phenomena considered as important—without first laying down value judgments. To decide to do science is to express a preference and thus to follow a value judgment. The selection of phenomena considered as important and deserving of a scientific explanation also rests on a value judgment. Social scientists may thus be guided by their ethical or political convictions in the choice of their subject matter: poverty, unemployment, violence, ethnic relations, etc. This is in part what makes the imperative of axiological neutrality necessary, since a clear distinction must therefore be made between value judgments and scientific, factual propositions. Axiological neutrality does not mean that scientific activity can get rid of all value judgments. It is obvious that if scientists had to remain perfectly neutral towards value judgments, they would not be able to choose an interesting subject matter, or to prefer truth to error. Science would simply become impossible (Gouldner 1963, Homans 1978).

Axiological neutrality is a moral duty inherent in intellectual honesty. Transgressions of this imperative may of course originate in dishonesty, but also in honest error. Here is a famous example of an involuntary confusion between facts and values chosen from the field of economics. After the law of marginal utility was discovered in the 1870s, some of the greatest economists at the time (such as Edgeworth and von Wieser) thought they could scientifically justify an egalitarian policy. The law of marginal utility, from which contemporary price theory originates, shows that when the stock of a good possessed by an agent grows, the subjective utility (satisfaction) for this agent of each additional unit of the good diminishes. The first unit is used to satisfy the most urgent need, the second one a less urgent need, the third one a still less urgent need, and so on: each additional unit is allocated to the satisfaction of less and less urgent needs. From this law is deduced the fact that, all other things being equal, the greater the quantity of money (monetary income) possessed by an agent, the smaller the subjective utility for this agent of an additional unit of money. Conversely, the smaller the monetary income of an agent, the greater the subjective utility of an additional unit of money. It would then seem as if the following reasoning could be made on purely empirical grounds: if the State decides to reduce the inequalities of income by a monetary transfer from the rich to the poor, and if this transfer does not reduce production (i.e., does not have any adverse effect on work and investment), then the total utility of society will grow, since the utility lost by the rich will be less than that gained by the poor. Robbins (1935) has shown that this deduction is not scientific: the law of

marginal utility can only be applied to *one* agent, and does not enable us to make comparisons of utility *between* persons. It is thus impossible to deduce from this law the proposition that the utility of the income taken from the rich is smaller than the utility of the income given to the poor. This proposition is an ethical value judgment—a preference given to what the poor gain over what the rich lose—that has been confused with a factual proposition. In order to deduce the egalitarian policy, this ethical value judgment is indispensable.

3. *An Objection Against Axiological Neutrality*

Strauss (1953) has raised an important objection against the weberian imperative of axiological neutrality: if value judgments are not scientific (see impossibility (a) in Sect. 1), then they are not based on reason and are thus *arbitrary*; but if value judgments are arbitrary, why should they be obeyed? Why should social scientists respect the moral imperative of axiological neutrality? Furthermore, if values are arbitrary then scientific activity is also arbitrary since it requires a series of value judgments (see Sect. 2). In other words, if the logical foundation of axiological neutrality is accepted, then the rational basis for the formulation of value judgments is undermined, and scientific activity becomes utterly arbitrary. The ideal of neutrality is supposed to protect science from the unwarranted interference of value judgments, but it results in the destruction of science since it implies that the selection of scientific problems and the practical utilization of science are purely arbitrary. According to Strauss, in order to avoid these nihilistic consequences the foundation of axiological neutrality must be questioned, and it must be admitted that a science of morals, a science of value judgments is possible.

Very briefly, a reply to this argument of Strauss is that the impossibility scientifically (empirically) to derive any value judgment does not imply that these judgments are necessarily irrational. The notion of rationality is not limited to the empirical domain. There also exists a rationality of norms, of values, an *axiological rationality* in the sense of Boudon (1995): a value can be justified by reasons, by arguments whose soundness and validity can be rationally discussed. Just because value judgments cannot be established on a purely empirical basis does not mean that they are arbitrary. It is true that among the advocates of axiological neutrality are authors believing in the pure irrationality and subjectivity of values (like Weber himself), but there are also among them authors defending the rationality of values (Boudon 1995, Rothbard 1997).

See also: Ethics and Values; Explanation: Conceptions in the Social Sciences; Knowledge, Sociology of; Positivism, History of; Positivism: Sociological; Scientific Culture; Scientific Reasoning and Discovery, Cognitive Psychology of; Sociology, Epistemology of; Sociology: Overview; Theory: Sociological; Values, Sociology of; Weber, Max (1864–1920)

Bibliography

Ayer A J 1936 *Language, Truth, and Logic*. V. Gollancz, London

Boudon R 1995 *Le Juste et le Vrai: Études sur l'Objectivité des Valeurs et de la Connaissance*. Fayard, Paris

Gouldner A W 1963 Anti-Minotaur: The myth of a value-free sociology. *Social Problems* **9**: 199–231

Homans G C 1978 What kind of a myth is the myth of a value-free social science?. In: Homans G C (ed.) *Certainties and Doubts: Collected Papers, 1962–1985*. Transaction Books, New Brunswick, NJ

Robbins L 1935 *An Essay on the Nature and Significance of Economic Science*, 2nd edn. Macmillan, London

Rothbard M N 1997 Praxeology, value judgments, and public policy. In: Rothbard M N (ed.) *The Logic of Action, I: Method, Money, and the Austrian School*. Elgar, Cheltenham, UK

Sen A K 1970 *Collective Choice and Social Welfare*. North-Holland, Amsterdam

Simon H A 1976 *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*, 3rd edn. The Free Press, New York

Strauss L 1953 *Natural Right and History*. University of Chicago Press, Chicago

Weber M 1917 The meaning of 'ethical neutrality' in sociology and economics. In: Schils E A, Finch H A (eds.) *Max Weber on the Methodology of the Social Sciences*. The Free Press, Glencoe, IL

R. Fillieule

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Missing Pages

10622 to 10624

New

'New' Eugenics, The: Role in Society

1. Introduction

With the continued progress of the Human Genome Mapping Project to identify all the genes in the body and the technologies that are developing from it, there is renewed concern about eugenic applications. Eugenics means using science for the qualitative and quantitative improvement of our genetic constitution. The subject was initiated by the Ancient Greeks and developed enthusiastically by a Victorian scientist, Francis Galton, with considerable support from his cousin Charles Darwin in the latter half of the nineteenth century (Kevles 1995). Its scope has increased enormously since the recent revolution in molecular genetics. Genetic databases can be easily obtained for individuals either before or at birth, and are being constructed for entire populations; and genetic markers are being developed to predict risks of developing cancer, diabetes, early heart attacks, and dementias. These markers can be used for diagnosis to see if the early embryo possesses deleterious alleles predisposing to disease or even carries personality 'traits' such as a tendency to become alcoholic, obese, or neurotic in later life. Such embryos can be discarded without the need to terminate pregnancy in favor of ones with 'better' genetic endowments leading to the concept of 'designer babies.' Genetic markers can also be used in childhood to predict the likelihood of developing common diseases in later adult life and this may lead to discrimination by insurance industries, employers, adoption agencies, and other state institutions.

The past history of eugenics has been appalling, with gross abuses in Germany between 1933 and 1945 when sterilization and mass extermination were performed. Revulsion generated by the laws of any state that interferes with people's reproductive choices or freedoms—whether for the theoretical purpose of improving the 'gene pool' or for other reasons such as the current demographic policy of China to restrict one child per family leading to changes in the gender ratio—has led to eugenics acquiring an evil reputation (Duster 1990).

However, the new genetic technologies are here to stay. To prevent such abuses for the future there have to be mechanisms to regulate these new techniques when applied to medical and social problems.

2. The Need for Regulation

The new eugenics requires regulating the varied and powerful social applications that are being developed, for example, 'designer' childbirth, cloning, disease prediction, insurance, industry, etc. Society will expect rules and laws to protect them from the excesses of overenthusiastic doctors and scientists, and from the greed and ambition of global corporations (Rifkin 1999).

Society's response to the new developments is not going to be easy. There is a variable and fine line that demarcates law from ethics. The line separates what can be left to individuals to choose for themselves from what requires enforcement by the law. The law and rules of professional bodies related to eugenics at present are a mass of inconsistencies. One problem is that major public interests such as education, health, and welfare, which should be the proper concern of law, are so intertwined with the realm of private and individual interests of the family. The latter could be left to individuals to work out for themselves based on their own wishes, needs, and preferences. For the future we will have to balance the practical applications arising from the new research against the need to protect both individuals and society from its dangers; and to resolve in some way the many conflicts of values that will arise.

Law is needed to negotiate between fact and value, and should be firmly based on reason and judgment (Brownsword et al. 1998). But unfortunately the facts and techniques are changing all the time from month to month. One regulatory body, the UK Human Fertilisation and Embryology Authority, developed a set of rules to cover the use of human embryos produced by test-tube fertilization. But these rules do not specifically apply to embryos produced by nuclear transfer techniques. As of 1999 there is at present no UK legislation in force specifically directed at human cloning. When human cloning was banned in the USA, laboratories there started to use enucleated egg cells from the cow for transfer of a human nucleus to make cow–human hybrids. So the law may have to be changed again; and just as quickly the scientists will find a way round it.

The problem is likely to get worse as the demand for new treatments of fertility is constantly increasing. This is partly because there is a continuing fall in the number of babies available for adoption and a growing tendency for women to delay motherhood to establish their careers first until they reach an age when

complications of childbirth are more likely to arise. For example, the number of women undergoing *in vitro* fertilization rose from approximately 12,000 in 1992 to 19,000 in 1994 in Western countries.

3. Regulatory Agencies

Legislators do not always understand the intricacies of the science they are called upon to regulate. So sets of regulatory agencies composed of a small group of scientists, lay-people, and lawyers are set up to provide recommendations and guidelines. These regulatory bodies have proliferated and range from local hospital ethical committees to national ethical committees to international bodies such as the UK Human Genetics Advisory Commission, the Advisory Committee on Genetic Modification, the UK Human Fertilisation and Embryology Authority, and the European Medicines Evaluation Agency. It is not surprising that with the numbers involved they are already starting to give confused or contradictory advice. To give one example: the UK Alzheimer's Disease Genetics Consortium have drawn up guidelines for what they consider to be the best possible practice for the genetic diagnosis of Alzheimer's dementia. They propose that before genetic testing informed consent should be obtained from the patient but the results should not usually be given back to the patient, except under the most exceptional circumstances, and then only after the appropriate counseling. However, another body, the Association of British Insurers, have demanded as a part of their genetic policy that for the purchase of insurance products of greater than £100,000 the results of genetic tests should be declared to them. A third body, the UK Human Genetics Advisory Commission, recommended a moratorium for two years on the use of genetic information for insurance purposes until one can find out what actuarial relevance it may have. But this was not adopted—it was considered by Government that the insurance industry should manage its own affairs without the need for interference (Galton and Ferns 1999).

Additional problems of leaving ethical decisions involving the genetic technology to locally interested consortia is that wider issues may be overlooked. For example, apo-E genotyping has been in use as an important diagnostic aid for blood fat disorders in hospital clinics worldwide for the past decades. This information, now available in casenotes and medical reports, can be used to predict the occurrence or diagnosis of Alzheimer's dementia in some of the patients who were tested (Galton and Krone 1991). There is no clear policy for confidentiality of these records or how they should be used. Who should have access to the information? The individuals were initially tested for a different condition, namely a cholesterol transport disorder. Should they or their first-degree relatives be now told of the expected risks of developing a dementia. Or should their primary care physician be informed of the results? Or should the results just remain undisclosed in the hospital records?

In this article some of the other questions and problems that regulation or legislation can give rise to will be presented. The answers will necessarily be evasive, and no solutions are easy to implement. But the responses can vary according to whether professional ethical codes, moratoria, referenda, or legislation are deemed appropriate.

4. Regulation for Industry

By the year 2010 ideas derived from the Human Genome Mapping Project may account for sales of $60 billion worth of products per annum; that is half the international sales of the pharmaceutical industry in 1992. New genetic knowledge translated into products will make a profit in almost any marketplace in the developed world. The profit motive is already starting to threaten the spirit of collaborative enthusiasm and cooperation in the fields of scientific research. To give just one example: there is a primitive tribe of hunter-gatherers, the Hagahai, living in a remote and inaccessible mountain range of Papua New Guinea. There was a suspicion from epidemiological studies that some of them were making antibodies to a virus that can cause a cancer of the blood, leukemia. Cells were taken from a young man of the tribe and cultured in the laboratories of the National Institutes of Health (NIH) in Bethesda, MD, USA. The cells were indeed found to be making antibodies and the NIH researchers were quick to see the profits to be made out of them. They applied for patent rights relating to the cell line. It was agreed that the Hagahai should receive some (unspecified) share of any commercial profits arising from the 'invention.' In 1996 the patent was officially disclaimed by the NIH after a heated international controversy developed about the exploitation and biocolonialism of primitive peoples (A. Pottage in *Law and Human Genetics*, Brownsword et al. 1998).

It can be seen that cells and genes of remote rural peoples are becoming the intellectual property of corporations and government institutions that hope to use the proceeds to fund their genetic and research laboratories. They claim in justification that free market incentives and the profit motive are the best and most efficient way to advance types of basic medical research which are very costly.

4.1 Databases

The Icelandic parliament passed a bill in December 1998 to make it legal for a private company, deCODE Genetics, to acquire a comprehensive database containing the very detailed medical records of almost the

entire population of Iceland, about 270,000 persons. This database is of vital interest because it contains large family trees for the vast majority of the population, and the records have been kept meticulously up-to-date. Iceland has an isolated population with very little immigration for the best part of 1,000 years. It is therefore an ideal population to investigate for the genetic basis of disease. deCODE Genetics, largely funded by American investors, has signed a contract with the Swiss pharmaceutical giant, Hoffman La Roche, worth about $200 million giving them exclusive access to the database to search for the genetic origins of 12 common diseases. It could be considered that the entire population of Iceland is being turned into a biomedical research commodity with a large commercial potential. The DNA part of the database is now being collected and will cover about 90 percent of the Icelandic population. Opposition to the bill was based on three major arguments. (a) There would be no individual signed consent for transfer of the health information to third parties and this would constitute an unwarranted invasion of privacy. (b) It may lead to loss of trust in the confidentiality of the doctor–patient relationship even if the health information were to be encrypted before the transfer. (c) Finally a monopoly was being given to one company to pursue this research. The project will cost about $150 million—more than twice the entire research budget for Iceland. So proponents argue that a private company is needed to fund the project since no governmental agency would have the marketing experience or readiness to float such a venture. The bill was finally passed and the DNA part of the population database is being assembled (Galton and O'Donovan 2000).

Such accurate and extensive medical records combined with a matching DNA bank could provide a Cornucopia to uncover the genetic origins of many diseases and be of inestimable health benefit to Icelanders as well as other populations. But if the database were to fall into the wrong hands it could provide yet another means for discrimination against an 'undesirable' or minority element in society. Passing the act also broke one of the fundamental principles of scientific research: that any research plans have to be evaluated and approved by an independent ethics committee before recruitment of patients should begin. deCode managed to convince the Icelandic government to pass the laws which avoided the necessity for review by a Bioethics or data protection committee. The government claimed that the act resulted from an informed democratic decision, but only 13 percent of the nation considered themselves to have a good enough grasp of the nature of the bill according to a Gallop poll conducted in November 1998. The act was passed by the Icelandic parliament before the public really understood all the issues. A provision in the bill ensures that the database will remain anonymous and therefore informed consent is not necessary. This then raises the problem that if some Icelanders are found to possess genetic variants that predispose them to a serious and treatable disorder they will not be told about it. Yet it is the doctors duty to inform such patients, so clearly the database is not going to be as anonymous as initially promised. The more genetic technology that is introduced into society, especially by Industry, the more important it becomes to protect the rights of individuals from becoming subservient to either state procedures or becoming exposed to market-led commercial forces (Berger 1999).

5. *Regulation of Home Genetic Testing*

'Do-it-yourself' tests for detecting pregnancies or genetic identification of paternity have been a recent commercial success. The discovery that genetic variants that predispose to some types of inherited forms of cancer has provoked widespread public interest. Demands for home tests by the public have been partly satisfied by commercial laboratories. Tests for genetic variants that cause a rare cancer of the large bowel have been put on the market and performed by a private company La Corp (Baltimore, MD, USA) after referral by a doctor. Many inadequacies and errors were found in the use of this service. In a study conducted in 1995, 177 patients were tested but only 19 percent received genetic counseling before the test to explain what was going to happen. Only 17 percent provided informed consent to have the test. In 32 percent of the cases the referring doctor misinterpreted the test results. Some patients at risk of developing the cancer were given false negative results because the tests were not done on affected family members to see which particular genetic variant occurs in that family. The scope for misapplication and misinformation of patients and creation of public anxiety in view of the other variable factors that can influence the final outcome are enormous. Yet already other commercial DNA-based tests are appearing. One developed by Oncor has received recent Food and Drugs Administration (FDA) approval in the USA to be used for risk prediction for the recurrence of breast cancer. Clearly regulation and tests of quality for the use of these measurements in the marketplace are going to be required for the future to protect the public from errors and the generation of unnecessary anxiety (Galton and Ferns 1999).

6. *Life and Health Insurance: Disease Prediction*

In its simplest form life or health insurance is a way of relieving the harmful effects of chance events such as disease or accidents. By pooling the modest premiums from a large number of people who are at the same risk, this can provide resources for making large payments to those who suffer harm or damage.

Standard insurance forms represent voluntary contracts between an individual and the insurance com-

pany and ask details about recent medical treatment, medical conditions, and family medical history. Insurance companies may require medical examination or consent for access to medical records. Assessment of risk plays a part in making the insurance contract and it is for the proposer to disclose the information required or otherwise withdraw from the contract. Failure to disclose is likely to lead to a void contract in case of a claim.

There are several ways such health or life insurance could be organized. There are basically three types of models: solidarity, mutuality, or altruism. Insurance based on solidarity takes no account of the different levels of risk that individuals bring to the pool; premiums are set at a uniform level, or based on the ability to pay. The British National Health Service is a good example of this type. Insurance based on mutuality relates the premiums to be paid to the level of risk each person brings to the common pool. The premiums will therefore vary with risks involved. A car driver with a clean driving record obtains a lower premium for car insurance than drivers with a careless record. This is perhaps fair because the bad driver could have taken more trouble to avoid the accidents. Conversely, penalizing someone with higher premiums through no fault of his or her own because they have inherited some faulty genes seems unfair—it is like financially punishing them for their skin color. Everyone from birth onwards has the chance of contracting one of a variety of afflictions that could be associated with defects in their DNA, but not all of these will come to light. Calculation of such risks can be difficult, and risk prediction for disease in life and health insurance may draw heavily on the genetic details of the person. Finally, insurance based on altruism argues that an individual should not be penalized for the chance inheritance of unfavorable genes over which he or she has no control, and that if anything health insurance as an instrument of social justice should be made cheaper not more expensive for them. This would be a noncommercial operation and would probably require that government provides some sort of a support system.

In the UK most life insurance is commercially organized and based on mutuality, whereas most health insurance in the UK is organized by the National Health Service, based on solidarity. For life insurance this means that the insurers try to adjust the premiums to match the risks involved. For single gene disorders such as the muscle wasting disease (Duchennes' muscular dystrophy) or cystic fibrosis, this is not difficult to do since if the individual possesses the bad genes he/she will almost certainly go on to develop the disease. The usual underwriting practice of insurance companies is to accept people with a family history of Huntington's disease who have not been tested but to charge them higher premiums. People known to carry the gene for Huntington's disease are usually refused cover (although they can obtain considerably better annuity rates if they retire because their life expectancy is reduced). However, for more complex disorders such as heart attacks, diabetes, or high blood pressure, the calculated risks are far more uncertain. Genes may be involved not in causing the disease, but in conferring a predisposition or susceptibility to the disease. Prediction of risks will be uncertain because of two major factors: (a) the disease will often not become apparent unless certain environmental conditions are encountered and (b) there is a variable interaction among genes that can modify the severity of the disease. A women carrying the susceptibility genes for diabetes mellitus may not manifest the disease until she becomes pregnant. The pregnancy in some way tips her into the diabetic state and her blood sugars rise and she passes sugar in her urine. After the pregnancy the diabetes disappears, but may recur at a later date as the women ages or becomes obese. As an obese person with diabetes she may sometimes see the disease disappear if strict dieting results in weight loss. How then should her premiums be adjusted if the declared aims are to take risks into account? As a general rule the risk is calculated at the time of taking out the insurance—if not diabetic at the time then no increased premium is applied.

The insurance industry in the UK has already laid out guidelines for the disclosure of genetic tests before the purchase of life insurance, disability income insurance, and critical illness cover. The Association of British Insurer's code of practice (UK) recommends that for purchase of life insurance up to a total of £100,000 which is directly linked to a new mortgage for a private dwelling, the results of any genetic tests must be reported but may not be taken into account by the insurance company. It is implied that for sums greater than £100,000 such genetic tests will be taken into account and the premiums adjusted accordingly. Companies insist that premiums paid must truly reflect the extent of risks of disability. Otherwise people at high risk will purchase higher claim policies and stand to make a financial gain out of the premiums paid by the lower risk individuals. This would be robbing the common pool of people of their money paid as premiums. The insurance industry is a commercial organization and cannot subsidize one type of policyholder who may withhold information on his/her genetic risks by charging the rest of the policyholders more than the appropriate premiums. The insurance industry calls this 'adverse selection' or ' antiselection' and although their arguments are rather bureaucratic in tone they are socially very important. There is already evidence that antiselection operates on the purchase of insurance products at rates above £100,000. However, if after genetic testing you cannot obtain a mortgage, this may discourage one from starting a family. This can result in an indirect eugenic effect.

Insurance companies in the UK cannot make applicants take genetic tests but if a test has already

Table 1
National attitudes and laws for the use of genetics in the insurance industry

The Council of Europe Convention on the protection of Human Rights and Biomedicine (1997) prohibits discrimination on the grounds of a person's genetic constitution and allows predictive testing to be carried out only for medical purposes. Unfair discrimination with respect to genetic make-up is prohibited under the Convention. Current signatories are (in alphabetical order): Czech Republic; Denmark; Estonia; Finland; France; Greece; Iceland; Italy; Latvia; Lithuania; Luxembourg; Moldova; Netherlands; Norway; Portugal; Romania; San Marino; Slovakia; Slovenia; Spain; Sweden; The former Yugoslav Republic of Macedonia and Turkey. Slovenia and Slovakia have both also ratified the Convention.

Australia
LISA's (Life, Investment and Superannuating Association of Australia Incorporated) Draft Policy on Genetic Testing (1997): calls on insurers not to initiate genetic tests. Results may be requested if they exist but should not form the basis of preferred risk underwriting. Strict confidentiality is required. LISA hopes to have its draft policy endorsed by the Australian Competition and Consumer Commission in due course.

Austria
1994 law stipulates that neither employers nor insurers are able to obtain or request the results of genetic analysis on their employees, candidates, policyholders, or insurance applicants.

Belgium
The first European country to enact legislation on the subject of genetics and insurance. The condition is that genetic data cannot be transmitted.

Canada
Canada's Human Reproductive and Genetic Technologies Act was introduced in 1996 but failed to complete its passage through Canada's Parliament. If enacted it would have prohibited 13 unacceptable uses of new human reproductive and genetic technologies although none of these applied to insurance. It is thought likely that similar legislation will now be introduced in the near future.

Denmark
The Danish Insurance Contracts Act (1997) does not allow insurers to ask for or use genetic information or use information about their relatives. Danish law is unusual in prohibiting hereditary information to be collected, while allowing genetic testing related to existing diseases or illnesses for risk evaluation purposes. Danish insurers must register personal information kept on policyholders and there are restrictions on the use of such data.

France
April 1994, The French Federation of Insurance Companies made the announcement that for a period of five years, its members would not use genetic information when determining applicants' insurability, even if favorable genetic information was brought by the applicant.

Hong Kong
The Hong Kong Federation of Insurers has set up working parties to consider the issues relating to genetic and HIV testing. Traditionally Hong Kong has followed legislative changes in the UK and, latterly, the USA; but it aims to ensure consistency with accepted international standards while maintaining its reputation as a territory with an objective and modern legal framework.

Netherlands
A five-year moratorium which expired in 1995 was extended indefinitely with the provision that it can be cancelled at any time with two years notice. Insurers have agreed to abstain from ordering genetic tests or using existing genetic information for policies below a certain level (NGL 300.000)

New Zealand
Existing genetic test results may be used to assess risk but an insurer cannot request a test.

Norway
Norwegian law prohibits requesting, receiving, possessing, or making use of genetic information.

United Kingdom
ABI Code of Practice on Genetic Testing (1997): ABI member companies will not ask applicants to undergo genetic tests or offer reduced premiums for negative tests. Ratings must be based on a relevant and reliable genetic test indicating an increased risk.

Human Genetics Advisory Commission (1997): has recommended that insurers should 'respect a moratorium on requiring disclosure of results of genetic tests' and is continuing its investigation of employment and insurance issues.

United States of America (USA)
The Human Genome Privacy Act (proposed in 1991 but never enacted): expanded the provisions of the then Privacy Act to cover genetic testing. It requires a government agency to permit an individual to inspect and have a copy of any genetic information the agency maintains about that individual. It also prohibits a government agency from disclosing genetic information except to the individual or the individual's representative, when the individual has authorised disclosure, or as otherwise authorized under the Act. The Act allows unauthorised disclosure: (a) within an agency for official use, (b) to a medical professional for use in connection with care or treatment of the individual, (c) for specified health and safety reasons, and (d) as directed by a court. Although not enacted, this proposed legislation points to society's concern with protecting the privacy of medical information.

The states of Arizona, California, Colorado, Florida, Indiana, Louisiana, Maryland, Minnesota, Montana, New Hampshire, New Jersey, New York, North Carolina, Ohio, Oregon, Tennessee, Virginia, and Wisconsin prohibit or restrict the use of genetic test information for nontherapeutic purposes, including insurance underwriting. Effectively, individuals cannot be denied coverage, have a policy canceled, or be charged a higher premium based on their genetic characteristics.

been taken the result must be given to the insurer if it asks for that information. If a genetic test is taken after the policy has been bought the insurer cannot ask for the results. For example, a woman aged 30 years with a known mutation in the breast cancer gene (*BRCA1*) may have a seven-year reduction in life expectation. If she has had the genetic test the insurers would expect her to pay a higher premium. The new genetic tests for chronic illness such as diabetes, dementias, or Parkinson's disease may profoundly affect the purchase of long term care insurance or critical illness cover by individuals and leave many people unprotected when they come to need most help. In the future the state will probably have to intervene to provide some form of protection for such people.

Insurance companies often decline to insure against pre-existing conditions. Does the possession of a liability gene for, say, Alzheimer's dementia constitute a pre-existing condition? Do such people, although healthy now, come into the category having a disease? If having a genetic defect counts as a disease, then the whole population would come into a disease category. We all possess some deleterious genetic variants, but they may never come to light as disease unless we encounter a particular set of environmental conditions. However, risk discrimination by insurers could be the slippery slope leading to a social underclass on the basis of their genetics who will find it impossible to afford any form of health or life insurance (Law et al. 1998).

To deal with this problem of 'adverse selection,' the European Council at Strasbourg in their document on Human Rights and Biomedicine (1997) stated that 'any form of discrimination against a person on the grounds of his or her genetic heritage is prohibited.' But there is already some evidence that insurance companies are dealing unfairly with relatives of families where a genetic disease occurs. The data come from a postal survey of 7,000 members of seven different British support groups for families with genetic disorders such as cystic fibrosis, Huntingdon's disease, and muscular dystrophy; and a sample of 1,033 members of the general public (Law et al. 1998). The response rate to the postal questionnaire was rather low as is usual with the number of questionnaires that are currently being circulated. About 13 percent of respondents who represented no actuarial risk on genetic grounds, either being healthy carriers or being altogether unaffected, had perceptions that they had been discriminated against on the basis of their family history. They were asked to pay higher premiums or were refused insurance outright. This may have been simply due to confusion or ignorance about genetic disease on the part of the medical officers employed by the insurance company; or possibly a way for them to try to avoid adverse selection of people who they thought were at increased risk of an inherited disease. Or it may be the respondents thought they were being victimized by their family history when in fact this was not the case. Clearly the claims for actuarial fairness by the insurance industry and the claims for social justice by the European Council will require new forms of regulation to clarify this problem. One possible solution might be to set up a type of state insurance scheme rather like the National Health Service so as to avoid penalizing individuals who have been given a bad genetic constitution through no fault of their own. In this case compensation would be paid by the more fortunate members of the rest of society who through no merit of their own inherited a good set of genes. Private insurance would still have a place alongside the State Insurance Scheme in the same way that Private Medicine operates alongside the NHS. But to many people it appears fair for the state to redistribute resources from the genetically lucky to the genetically 'underendowed.'

The responses of other European countries to the problem of adverse selection and genetic discrimination are presented in Table 1.

In the USA where health insurance is commercially organized and based on a mutuality model, many people have no health insurance at all. A great deal of recent legislation has aimed to restrict the use by health insurers of genetic tests and more broadly of any 'genetic information.' Federal legislation enacted in 1996 forbids insurance companies from using even family history to exclude people from schemes for group health insurance. The state of New Jersey has gone further and bans the use of genetic information for any insurance or employment purposes without written consent of the person involved. They have not suggested any practical measures to prevent adverse selection but appear to be making it easier for some groups to obtain health insurance. This is to prevent the trend for health care in the USA being converted from a social service to an economic commodity, sold in the marketplace and distributed on the basis of who can afford to pay for it (Pokorski 1997).

See also: Darwinism: Social; Ethical Issues in the 'New' Genetics; Eugenics as an International Movement; Eugenics as the Basis of Population Policy; Eugenics, History of; Eugenics in Asia; Eugenics in Europe; Genetic Counseling: Historical, Ethical, and Practical Aspects

Bibliography

Appleyard B 2000 *Brave New Worlds*. Harper Collins, London

Association of British Insurers 1997 *Life Insurance and Genetics*. A policy statement of the Association of British Insurers

Berger A 1999 Private company wins rights to Icelandic gene database. *British Medical Journal* **318**: 11

Brownsword R, Cornish W R, Llewlyn M (eds.) 1998 *Law and Human Genetics*. Hart Publishing, Oxford, UK

Duster T 1990 *Backdoor to Eugenics*. Routledge, New York

European Commission of Advisors on the Ethical Implications of Biotechnology 1996 *The Ethical Aspects of Prenatal Diagnosis*. European Commission of Advisors on the Ethical Implications of Biotechnology, Brussels

Galton D J 2001 *In Our Own Image: Eugenics and the Genetic Modification of People*. Little Brown & Co., Boston

Galton D J, Ferns G A A 1999 Genetic markers to predict polygenic disease: A new problem for social genetics. *Quarterly Journal of Medicine* **92**: 223–32

Galton D J, Galton C J 1998 Francis Galton: And eugenics today. *Journal of Medical Ethics* **24**: 99–105

Galton D J, O'Donovan K 2000 Legislating for the new predictive genetics. *Human Reproduction and Genetic Ethics* **6**: 39–47

Galton D J, Krone W 1991 *Hyperlipidaemia in Practice*. Gower Medical, London

Great Britain. Parliament. House of Commons, Science and Technology Committee 1995 *Human Genetics: The Science and its Consequences*. HMSO, London

Harris J 1998 *Clones, Genes and Immortality*. Oxford University Press, Oxford, UK

Kennedy I, Grubb A 1994 *Medical Law*. Butterworths, London, pp. 666–9

Kevles D J 1995 *In the Name of Eugenics*. Harvard University Press, Cambridge, MA

Kevles K J, Hood L (eds.) 1992 *Scientific and Social Issues in the Human Genome Project*. Harvard University Press, Cambridge, MA

Kitcher P 1996 *The Lives to Come*. Penguin, Harmondsworth, UK

Law L, King S, Wilkie T 1998 Genetic discrimination in life insurance: Empirical evidence from a cross sectional survey of genetic support groups in the UK. *British Medical Journal* **317**: 1632–5

Lewontin R C 1993 *The Doctrine of DNA*. Penguin, Harmondsworth, UK

Peters T 1997 *Playing God*. Routledge, New York

Pokorski R J 1997 Insurance underwriting in the genetic era. *American Journal of Human Genetics* **60**: 205–16

Powledge T 1974 Genetic screening as a political and social development. In: Bergsma D (ed.) *Ethical, Social and Legal Dimensions of Screening for Human Genetic Disease*. Stratton International Medical Books, New York

Rifkin J 1999 *Biotech Century*. Putnam, New York

Rose S, Lewontin R C, Kamin L J 1990 *Not in Our Genes*. Penguin, Harmondsworth, UK

Weir R F, Lawrence S C, Fales E (eds.) 1994 *Genes & Human Self-knowledge*. University of Iowa Press, Ames, IA

Wilson E O 1978 *On Human Nature*. Penguin, Harmondsworth, UK

D. J. Galton

New Religious Movements

1. Definition

The concepts of 'newly arising religion,' then 'new religion,' entered Japanese vocabulary after the Second World War when restrictions imposed on religions other than state Shinto were lifted (Inoue 1991, Mullins et al. 1993). Around the late 1960s, western scholars began to use the term 'new religious movement' (NRM) to describe (a) religions that were new to the West, although, like Tibetan Buddhism, they could have existed for hundreds of years in another part of the world; and (b) religious organizations referred to in the media and popular parlance as 'cults' or 'sects,' and which, had they emerged earlier, the scholars themselves would continue to classify as cults or sects.

There are a number of problems related to the use of the term. First, it does not refer to all new religious groups (the United Reformed Church, e.g., would not be called an NRM although it only came into being in 1972 as the result of a merger between English Presbyterians and Congregationalists). Second, some groups classified as NRMs protest that they are not new but can claim a heritage of hundreds of years. ISKCON (the International Society for Krishna Consciousness), e.g., traces its lineage through the sixteenth-century monk, Chaitanya Mahaprabhu. Interestingly, in Japan those movements which have appeared since 1970 are now referred to as '*new* New Religions'—and Christian groups are referred to as

new religions (Mullins et al. 1993). Third, several groups deny that they are religions; many prefer to call themselves spiritual, philosophical, or educational; some, such as the Raëlians, are avowedly atheistic. NRMs can, however, be defined loosely as new in so far as they have become visible in their present form since World War II, and religious in so far as they address at least some of the questions of ultimate concern that traditionally have been addressed by more established religions—questions such as 'What is the meaning of life?' 'Is there life after death?' 'Is there a God?'

A further problem has arisen from political battles arising out of the conflicting images of the movements that are constructed by several categories of persons (Barker 1995). The media, who tend to be interested in the more sensational aspects of a group, and anticultists (see below), who disseminate only negative accounts, both use the terms 'cult' or 'sect.' Scholars, in their desire to give as objective an account as possible, had introduced the term NRM partly to distance themselves from the blanket condemnations associated *a priori* with the label 'cult.' This, however, led to accusations from anticultists that the scholars' use of the concept was an insult to 'genuine religion' and little more than an attempt to whitewash the movements. By the late 1990s, this had resulted in a preference among some scholars and the movements themselves to use the less discriminating terms 'alternative' or 'minority religions.'

2. *Numbers*

The number of NRMs is unknown and will, of course, differ according to the precise definition that is used. It is, however, possible that in the West there are around 1,500 groups—more if all the 'self religions' are included (see *Religiosity: Modern*). Several thousand more may be found in Latin America, the former Soviet Union, and throughout Asia, including the Indian subcontinent. It has been estimated that over 10,000 new religions are to be found in Africa, although many of these are referred to more frequently as African Independent Churches.

The number of members is even more difficult to assess. There are a few large, international movements (Soka Gakkai claims around 10,000,000 members worldwide; Scientology claims over 8,000,000 have taken one or more of its courses); but there are probably less than 100 groups in the West with more than 1000 adherents; the majority are small, local groups with no more than 50 or so followers at any one time. The difficulty in obtaining accurate figures is confounded by the fact that most movements have various levels of membership, ranging from the fully committed, who may live with and work for their movement, to 'sympathizers' who might belong to more than one group (especially within the New Age or Human Potential movement). Furthermore, both the NRMs and their opponents tend to exaggerate membership figures, the former because they wish to be seen as more successful than they really are, the latter because they want to exaggerate the threat that the movements pose, and because adherence to a brainwashing thesis (which claims members are recruited by irresistible and irreversible mind-control techniques) leads them to deny the high turnover rate, with large numbers of converts deciding to leave after a few months' membership (Barker 1989).

3. *Variety*

Perhaps the most important generalization to be made about NRMs is that one cannot generalize about them. There have, of course, been new religions throughout history, including early Christianity, early Islam, and early Methodism. But while most of the movements that arose in previous times were offshoots of the prevailing traditions in their society of origin, the range of traditions, permutations, and combinations to be found within the current wave of NRMs is without precedence. This may in part be explained by the dramatically increased social and geographical mobility, the phenomenal expansion of the mass media, and various other processes subsumed under the rubric of postindustrial globalization. Thus, one can find those brought up in, say, the Judaeo–Christian tradition deciding to join (from, respectively, the Hindu, Buddhist, Shinto, or Islamic traditions) the Brahma Kumaris, Nichiren Shoshu, Mahikari, or Hizb ut-Tahrir. There are, moreover, NRMs that draw on the beliefs and practices of Wicca, Paganism, Occultism, and Esotericism; others have turned for their inspiration to modern philosophies and ideologies, psychological and psychoanalytic theories, and astrology; yet others owe much to ideas to be found in science fiction writing (Barker 1989, Melton 1992, Miller 1995, Robbins 1988).

Some NRMs, such as the Unification Church, have relatively systematic theologies; others consist of what seems to outsiders to be a jumble of uncoordinated pronouncements. Practices include chanting, meditation, ritual magick, dance, formal liturgy, and group therapy. Some members of NRMs live together in rural communes, others live separately in high-rise apartments. Life styles may be ascetic or hedonistic. Attitudes towards women, children, and the outside world vary. Some groups indulge in sexual orgies, others practice celibacy. Some are rich, others poor. The movement's organization may be democratic or authoritarian. The leader may be seen as a guru, teacher, prophet, Messiah, god, or God. He (or sometimes she) may perform miracles, may channel messages from Ascended Masters, or may be in touch with aliens from another planet. Some NRMs are destructive, others benign. Most have positive effects

some of the time for some of their members and negative effects at other times for others.

Attempts to classify NRMs have been largely unsuccessful, the sheer variety having made elegant typologies such as that which Wilson (1970) elaborated for the largely Christian-based sects of the nineteenth century well nigh impossible. The most effective attempt has been Wallis' distinction between 'world-rejecting,' 'world-affirming,' and 'world-accommodating' new religions (Wallis 1984).

4. *Characteristics*

There are, nevertheless, certain characteristics that are commonly found among NRMs merely because they are both new and religious. First, new religions, by definition, are composed of first-generation members, and converts to any religion tend to be more zealous than those born into their religion. Second, the membership is usually atypical of the general population; e.g., during the 1970s, NRMs in the West appealed disproportionately to young adults from middle-class backgrounds. The movements had, thus, an enthusiastic, healthy, well-educated, but inexperienced membership with few responsibilities. Third, the founder frequently wields charismatic authority, being unconstrained by rules or tradition, and unaccountable to anyone (except, perhaps, God). Fourth, NRMs often embrace dichotomies that allow little qualification. The world is divided unambiguously into 'them' vs. 'us' people; 'true' vs. 'false' beliefs; 'good' and 'godly' vs. 'evil' and 'satanic' morals and practices; and time is split between 'before' vs. 'after' significant events such as conversion or the coming apocalypse. Fifth, NRMs, given that they present an alternative to the status quo, commonly are regarded with suspicion and/or antagonism by the host society. Understandably, such fears have been reinforced by horrific incidents such as the mass suicides and/or murders associated with the Peoples Temple (Guyana 1978), the Branch Davidians (Waco 1993), the Solar Temple (Quebec, Switzerland, and France 1994/5), Aum Shinrikyo (Tokyo 1995), and Heaven's Gate (San Diego 1997).

5. *Changes*

A sixth characteristic of NRMs is that they are likely to change far more rapidly than older, more established religions. Taking demographic factors alone: youthful converts get older; the arrival of a second generation will mean not only that resources (time and money) have to be devoted to their care and socialization, but also that they are likely to question many of the beliefs and practices their parents had adopted. As children cannot be expelled as easily as stroppy converts, the movement could have to adapt its beliefs and practices. Furthermore, charismatic leaders die, their authority usually being replaced by a more predictable and accountable leadership. Many of the sharp boundaries that distinguished the NRM from the rest of society become permeated by qualification and accommodation, and the movements that survive tend, though not always, to merge less noticeably into the pluralistic *smorgasbord* of contemporary democracies.

6. *Cult-watching Groups*

During the 1970s, there arose a number of organizations devoted to exposing the perceived dangers of NRMs in an attempt to have them controlled or outlawed completely. These became generically known as the anticult movement (ACM) (see Bromley and Hadden 1993). Originally many of the members of the ACM were worried relatives of converts who believed that their loved ones were in danger, but they were soon joined by others, including ex-members and persons such as lawyers, private detectives, politicians, mental health professionals, and counselors who developed a specialist interest in NRMs. Anticultists range from persons who have engaged in forcible 'deprogramming' for thousands of dollars (a practice now rarely carried out in the West, although still prevalent in Japan), to those who merely attempt to educate the general public through literature, lectures, and counseling. By the 1990s, the American Family Foundation had become the largest and most professional such organization. Alongside the ACM, there also developed a counter-cult movement that criticizes the beliefs of NRMs. Among the most active of these are the Dialogue Centres in Europe, and the Watchman Fellowship in the USA.

Other groups, such as INFORM, based at the London School of Economics, RENNER at the Universities of Copenhagen and Aarhus, and CESNUR, in Turin, are more scholarly in their orientation, with the largest collection of material being housed at the University of California, Santa Barbara under the direction of Gordon Melton. Several Government Enquiries have been commissioned to look into the activities of the NRMs. The Eastern European, French, Belgian and, to a lesser extent, German reports have tended to rely either solely or predominantly on anticult resources and to recommend the movements should be restricted, while the Swedish, Dutch, and Ontario reports have paid more attention to scholarly research, and tend to share the position of the British and US governments that NRMs should be treated in law like any other religions unless they are found guilty of criminal malpractice.

See also: Charisma and Charismatic; Charisma: Social Aspects of; Millennialism; Religion and Youth; Religion: Definition and Explanation; Religion: Evolution

and Development; Religion, History of; Religion: Morality and Social Control; Religion, Sociology of; Religiosity: Modern; Weberian Social Thought, History Of

Bibliography

Barker E 1989 *New Religious Movements: A Practical Introduction*. HMSO, London

Barker E 1995 The scientific study of religion? You must be joking! *Journal for the Scientific Study of Religion* **34**(3): 287–310

Bromley D, Hadden J 1993 *The Handbook on Cults and Sects in America* (Parts A&B). JAI Press, Greenwich, CT

Inoue N (ed.) 1991 *New Religions: Contemporary Papers in Japanese Religion 2*. Institute for Japanese Culture and Classics, Kokugakuin University, Tokyo

Melton J G 1992 (original edn. 1986) *Encyclopedic Handbook of Cults in America*. Rev. edn. Garland, New York and London

Miller T (ed.) 1995 *America's Alternative Religions*. State University of New York, Albany, NY

Mullins M R, Susumu S, Swanson P L (eds.) 1993 *Religion and Society in Modern Japan*. Asian Humanities Press, Berkeley, CA

Robbins T 1988 *Cults, Converts and Charisma*. Sage, Newbury Park, CA

Wallis R 1984 *The Elementary Forms of the New Religious Life*. Routledge & Kegan Paul, London

Wilson B R 1970 *Religious Sects: A Sociological Study*. Weidenfeld & Nicholson, London

E. Barker

Newcomb's Problem

Newcomb's problem is a decision puzzle whose difficulty and interest stem from the fact that the possible outcomes are probabilistically dependent on, yet causally independent of, the agent's options. The problem is named for its inventor, the physicist William Newcomb, but first appeared in print in a 1969 paper by Robert Nozick. Closely related to, though less well known than, the Prisoners' Dilemma, it has been the subject of intense debate in the philosophical literature. After more than three decades, the issues remain unresolved. Newcomb's problem is of genuine importance because it poses a challenge to the theoretical adequacy of orthodox Bayesian decision theory. It has led both to the development of causal decision theory and to efforts aimed at defending the adequacy of the orthodox theory.

This article surveys the debate. The problem is stated in Sect. 1. Arguments for each of the opposed solutions are rehearsed in Sect. 2. Sect. 3 contains a sketch of causal decision theory. One strategy for defending orthodox Bayesianism is outlined in Sect. 4.

1. The Problem

There are two boxes on the table in front of you. One of them is transparent and can be seen to contain $1,000 dollars. The other is opaque. You know that it contains either one million dollars or nothing. You must decide whether to take (a) only the contents of the opaque box (call this the *one-box option*); or, (b) the contents of both boxes (the *two-box option*). You know that a remarkably accurate Predictor of human deliberation placed the million dollars in the opaque box yesterday if and only if it then predicted that you would choose today to take only the contents of that box. You have great confidence in the Predictor's reliability. What should you do?

2. Opposed Arguments

There is a straightforward and plausible expected value argument for the rationality of the one-box choice. If you take only the contents of the opaque box, then, almost certainly, the Predictor will have predicted this and placed the million dollars in the box. If you choose to take the contents of both boxes, then, almost certainly, the Predictor will have predicted this and left the opaque box empty. Thus the expected value of the one-box choice is very close to $1,000,000, while that of the two-box choice is approximately $1,000. The one-box choice maximizes your expected value. You should therefore choose to take only one box.

But there is also a straightforward and plausible dominance argument supporting the two-box choice. The prediction was made yesterday. The $1,000,000 is either there in the opaque box or it is not. Nothing that you do now can change the situation. If the $1,000,000 is there, then the two-box choice will yield $1,000,000 plus $1,000, while the one-box choice yields only the $1,000,000. And if the opaque box is empty, there is no million dollars to be had: the two-box choice gives you $1,000, the one-box option leaves you with no money at all. Whichever situation now obtains, you will be $1,000 better off if you take both boxes. The two-box choice is your dominant option. You should therefore choose to take both boxes.

Care must be taken in applying the second argument, for dominance reasoning is appropriate only when probability of outcome is independent of choice. Example: Nasty Nephew has insulted Rich Auntie who threatens to cut Nephew from her will. Nephew would like to receive the inheritance, but, other things being equal, would rather not apologize. Either Auntie cuts Nephew from the will or she does not. If she does, Nephew receives nothing, and in this case prefers not having apologized. If she does not, then Nephew receives the inheritance, but once again, prefers inheritance with no apology to inheritance plus apology. Does a dominance argument give Nephew reason to refrain from apologizing? Clearly that line of thought

is fallacious. It ignores the fact that whether or not Nephew is cut from the will may depend upon whether or not he apologizes.

Does this observation vitiate the two-boxer's appeal to dominance? That depends on what one means by 'dependence.' What the Predictor did yesterday is certainly probabilistically dependent on what you choose today. The conditional probability that the opaque box contains \$1,000,000 given that you choose the one-box option is close to one, while the conditional probability of the \$1,000,000 being present given that you choose to take both boxes is close to zero. But yesterday's prediction is causally independent of today's choice, for what is past is now fixed and determined and beyond your power to influence. We are used to these two senses of dependence coinciding, but Newcomb's problem illustrates the possibility of divergence when, for example, there is reliable information to be had about the future. Armed with the distinction, two-boxers maintain that dominance reasoning is valid whenever outcomes are causally independent of choice.

The dominance argument can be made even more vivid by considering the following variation on the original problem. The opaque box has a transparent back. Your dearest friend is sitting behind the boxes and can see the contents of both. Your friend is allowed to advise you what you should do. What choice will your friend tell you to make? Obviously: to take both boxes. Now it would be irrational to ignore the advice of a friend who has your best interests at heart and who you know is better informed than you are. And should not the fact that you know in advance what your friend would tell you to do in this variation of the problem determine your opinion about what to do in the original version?

One-boxers reply that the kind of agent who follows the friend's advice, and two-boxers in general, will very likely finish up with only \$1,000. 'If you're so smart', they taunt, 'why ain'cha rich?' Two-boxers may counter that Newcomb's problem is simply a situation in which one can expect irrationality to be rewarded. If 'choosing rationally' means no more than choosing so as to maximize expected reward, it may seem hard to see how such a response can be coherent, but see Gibbard and Harper (1978), Lewis (1981b).

3. *Causal Decision Theory*

Those two-boxers who are skeptical that orthodox Bayesian decision theory of the kind developed in Jeffrey (1983) can deliver the right answer to Newcomb's problem face the challenge of providing an account of rational decision making consistent with this judgment. By the early 1980s several independently formulated but essentially equivalent versions of an alternative account had appeared. (See, for example, Cartwright (1979), Gibbard and Harper (1978), Levi (1975), Lewis (1981a), Skyrms (1982), Sobel (1978). According to this alternative account, now widely referred to as causal decision theory (CDT), calculations of expected value should be sensitive to an agent's judgments about the probable causal consequences of the available options. The differences between this causal theory and orthodox Bayesian decision theory are perhaps best appreciated by proceeding from a definition of expected value that is neutral between the two accounts.

The notion of expected value relates the value assigned by an agent to an option A to the values the agent assigns to the more particular ways w in which A might turn out to be true. All parties to the debate agree that the expected value $V(A)$ of an option A is a probability weighted average of the values of these various w. That is:

$$V(A) = \sum_{w} V(w).P_A(w)$$

where P_A is the probability function that results when the agent's subjective probability function P is revised so as to accept A. Where the theories differ is on the method of probability revision deemed appropriate to this context. According to the orthodox Bayesian theory, the appropriate method is revision by conditionalization on the option A:

$$V(A) = \sum_{w} V(w).P(w/A)$$

whereas causal decision theorists maintain that for each w the weight should be given by the probability of the subjunctive conditional: if A were chosen then w would be true. This conditional is usually abbreviated $A \Box\!\!\rightarrow w$. Thus:

$$V(A) = \sum_{w} V(w).P(A \Box\!\!\rightarrow w)$$

That these two definitions of expected value do not coincide is far from obvious. But in fact they do not coincide because they cannot. David Lewis, one of the co-founders of CDT, demonstrated the impossibility of defining a propositional connective $\Box\!\!\rightarrow$ with the property that for each A, C the probability of the conditional $A \Box\!\!\rightarrow C$ equals the conditional probability of C given A. (See Lewis (1976) and *Counterfactual Reasoning, Quantitative: Philosophical Aspects* and *Counterfactual Reasoning, Qualitative: Philosophical Aspects*.)

But the divergence of these two definitions is best seen by applying them to Newcomb's problem. There are two options: the one-box choice (B_1) and the two-box choice (B_2); and four possible outcomes: receiving nothing (N), receiving \$1,000 (K), receiving \$1,000,000 (M), or receiving \$1,001,000 (i.e., 'getting the lot' L). Assume that the values you assign to these outcomes are given by the dollar amounts, and suppose finally that you have 99 percent confidence in

the Predictor's reliability, i.e., that you assign the conditional probabilities: $P(M/B_1) = 0.99$, $P(N/B_1) = 0.01$, $P(L/B_2) = 0.01$, $P(K/B_2) = 0.99$.

Then on the orthodox account:

$$V(B_1) = V(M).P(M/B_1) + V(N).P(N/B_1) = 990{,}000$$
$$V(B_2) = V(L).P(L/B_2) + V(K).P(K/B_2) = 11{,}000$$

and so according to the Bayesian theory the one-box choice maximizes expected value.

Before the causal theory can be applied, probabilities must be assigned to the relevant subjunctive conditionals: $B_1 \Box\!\!\rightarrow M$, $B_1 \Box\!\!\rightarrow N$, $B_2 \Box\!\!\rightarrow L$, and $B_2 \Box\!\!\rightarrow K$. The first of these, $B_1 \Box\!\!\rightarrow M$, is the conditional: 'If you were to take only the contents of the opaque box, then you would receive \$1,000,000.' Now this conditional is true if and only if the Predictor placed the \$1,000,000 in the opaque box yesterday. Similarly the conditional $B_2 \Box\!\!\rightarrow L$: 'If you were to take the contents of boxes, then you would receive \$1,001,000' is true just in case the Predictor put the \$1,000,000 in the opaque box yesterday. It follows that:

$$P(B_1 \Box\!\!\rightarrow M) = P(B_2 \Box\!\!\rightarrow L) = \mu$$
$$P(B_1 \Box\!\!\rightarrow N) = P(B_2 \Box\!\!\rightarrow K) = 1 - \mu$$

where μ is whatever probability you assign to the proposition that yesterday the Predictor placed \$1,000,000 in the opaque box. Thus:

$$\begin{aligned} V(B_1) &= V(M).P(B_1 \Box\!\!\rightarrow M) + V(N).P(B_1 \Box\!\!\rightarrow N) \\ &= 1{,}000{,}000.\mu \\ V(B_2) &= V(L).P(B_2 \Box\!\!\rightarrow L) + V(K).P(B_2 \Box\!\!\rightarrow K) \\ &= 1{,}000{,}000.\mu + 1{,}000 \end{aligned}$$

and so according to the causal theory the two-box choice maximizes expected value. Note that these assignments reflect the causal decision theorist's advocacy of the dominance argument in a situation in which outcomes are causally independent of choice, for to say that C is causally independent of A is just to say that $P(A \Box\!\!\rightarrow C) = P(C)$.

There are a couple of delicate issues involved here. First of all, one must resist the temptation to say such things as:

> The \$1,000,000 is there in the opaque box, but if you were to choose to take only the contents of that box, then the \$1,000,000 would not be there.

A thought to which one might be led by the consideration is that:

> If you were to make the one-box choice now, then the Predictor, being so reliable, would have to have predicted yesterday that that was what you would do today.

Of course these conditionals would be true if your choice today had the power to affect the past, if, for example, your choosing to take only one box could somehow cause the Predictor yesterday to have predicted that that was what you would do. But that is not so. Newcomb's problem is not a story about backward causation. The Predictor predicted what you will do, but that prediction was not caused by what you will do.

But of course the two conditionals above are natural enough things to say. The point is not that this sort of backtracking interpretation of the subjunctive conditional is impossible—conditionals are sometimes used in precisely this way. It is rather that backtracking conditionals are not relevant to the present discussion, for they fail to reflect causal structure. So let the causal decision theorist simply stipulate that the choice-to-outcome conditionals not be given a backtracking interpretation. The truth-value of $A \Box\!\!\rightarrow C$ is evaluated, roughly speaking, by supposing the antecedent A to be true, and then determining whether the consequent C follows from that supposition. 'No backtracking' means that in supposing A to be true, one continues to hold true what is past, fixed, and determined. (For further discussion of these issues see *Counterfactual Reasoning, Quantitative: Philosophical Aspects* and *Counterfactual Reasoning, Qualitative: Philosophical Aspects*.)

Second, it is reasonable to be suspicious of CDT's appeal to μ your present unconditional probability that the \$1,000,000 is in the opaque box. Given that you know the Predictor to be highly reliable, the best evidence you now have as to whether or not the \$1,000,000 is in the box is provided by your current beliefs about what you will choose to do. So μ will have a value close to 1 if you believe that you will take only one box, and a value close to 0 if you think that you will take both boxes. The value of μ thus helps to determine your expected values, while at the same time being determined in part by those expected values. Some opponents of CDT (e.g., Isaac Levi) have found this problematic, maintaining that agents cannot coherently assign probabilities to the outcomes of their own current deliberations, while one advocate of CDT (Brian Skyrms) exploits this very feature of the theory in giving a dynamic account of deliberation as a kind of positive feedback mechanism on the agent's probabilities that terminates only as the agent's degree of belief that a certain choice will be made approaches the value 1 (see Skyrms (1990)).

Further criticism of CDT has focussed on the partition problem. As Levi has pointed out (1975), the deliverances of CDT depend upon the particular way in which outcomes are specified. CDT may prescribe different courses of action when one and the same choice situation is described in two different ways. So unless there is some preferred way of partitioning outcomes, CDT is inconsistent. Example: if the outcomes in Newcomb's problem are taken to be (1) you

receive the contents of the one (opaque) box only, and (2) you receive the contents of the two boxes, then CDT can be made to prescribe the one-box choice, for these two outcomes are clearly under your, the agent's, causal control.

Defenders of CDT take the moral of this to be that outcomes to a well-posed decision problem must be specified to a degree of detail that is 'complete in the sense that any further specification of detail is irrelevant to the agent's concerns' (Gibbard and Harper 1978). Some opponents of CDT consider this unnecessarily limiting and a mark of the orthodox theory's superiority, for the evidential theory may be applied in a way that is robust under various different ways of partitioning outcomes (see, for example, the articles by Levi and Eells in Campbell and Sowden (1985)).

For a re-characterization of CDT that does not mention causation, see *Counterfactual Reasoning, Quantitative: Philosophical Aspects* and *Counterfactual Reasoning, Qualitative: Philosophical Aspects*. The most complete and sophisticated account of CDT to date is to be found in Joyce (1999).

4. The Tickle Defense

The relation between Newcomb's problem and the Prisoners' Dilemma has been noted and debated (see, for example, Campbell and Sowden (1985)). But whereas the Prisoners' Dilemma has been discussed widely by social and behavioral scientists in several fields, the discussion of Newcomb's problem has been confined mainly to the philosophical literature. This may be because the significance of the problem is more purely theoretical than practical, but is perhaps due also to the somewhat fantastic nature of the Newcomb problem itself (in Richard Jeffrey's (1983) phrase: 'a prisoners' dilemma for space cadets').

Less fantastic problems of the Newcomb type may arise, however, in situations in which a statistical correlation is due to a common cause rather than to a direct causal link between the correlated factors. Thus, for example, though it is well established that there is a strong statistical correlation between smoking and lung cancer, it does not follow that smoking is a cause of cancer. The statistical association might be due to a common cause (a certain genetic predisposition say) of which smoking and lung cancer are independent probable effects. (This is sometimes referred to as 'Fisher's smoking hypothesis'). If you have the bad gene you are more likely to get lung cancer than if you do not, and you are also more likely to find that you prefer smoking to abstaining. Suppose you are convinced of the truth of this common cause hypothesis and like to smoke. Does the fact that smoking increases the probability of lung cancer give you a reason not to smoke?

Surely it does not. Smoking is, for you, the dominant option, and it would be absurd for you to choose not to smoke in order to avoid lung cancer if your getting lung cancer is causally independent of your smoking. Here CDT delivers what is, pretty clearly, the correct answer. On the face of it this would seem to provide a much more clear cut counter-example to orthodox Bayesian theory than Newcomb's problem.

But this is not clearly a counter-example at all. Here the strategy of the defenders of orthodoxy has been to argue that in this more realistic kind of problem, the theory, correctly applied, prescribes the same choice as CDT. Let S be the proposition that you smoke, and let C be the proposition that you get lung cancer. Although you know that smoking and lung cancer are statistically correlated, it does not follow that C and S are probabilistically dependent by the lights of your subjective probability assignment. That is because you will typically have further information that will screen off the probabilistic dependence of C on S. In the present example, this might consist of your knowing that you have a craving to smoke. Let T be the proposition that you have this craving or 'tickle' (the strategy being discussed is customarily referred to as the 'tickle defense'. Then $P(C/T\&S) = P(C/T\&\text{not}S)$. Now since you know that you have the tickle, $P\ (T) = 1$ and hence according to your probability assignment $P(C/S) = P(C/\text{not}S)$. The non-causal probabilistic dependence of C on S has vanished (see Eells (1982) but note also the reply by Jackson and Pargetter in Campbell and Sowden (1985).

See also: Bayesian Statistics; Bounded and Costly Rationality; Counterfactual Reasoning, Qualitative: Philosophical Aspects; Counterfactual Reasoning, Quantitative: Philosophical Aspects; Decision Making: Nonrational Theories; Decision Support Systems; Decision Theory: Bayesian; Decision Theory: Classical; Game Theory; Game Theory and its Relation to Bayesian Theory; Prisoner's Dilemma, One-shot and Iterated

Bibliography

Campbell R, Sowden L 1985 *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*. The University of British Columbia Press, Vancouver, BC

Cartwright N 1979 Causal laws and effective strategies. *Nous* **4**: 419–37

Eells E 1982 *Rational Decision and Causality*. Cambridge University Press, Cambridge, UK

Gibbard A, Harper W L 1978 Counterfactuals and two kinds of expected utility. In: Hooker C A, Leach J J, McClennen E F (eds.) *Foundations and Applications of Decision Theory* **1**: 125–62. Reprinted (abridged) in Campbell and Sowden (1985)

Jeffrey R C 1983 *The Logic of Decision*, 2nd edn. University of Chicago Press, Chicago

Joyce J M 1999 *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge, UK
Levi I 1975 Newcomb's many problems. *Theory Decision* **6**: 161–75
Lewis D K 1976 Probabilities of conditionals and conditional probabilities. *Philosophical Review* **85**: 297–315
Lewis D K 1981a Causal decision theory. *Australasian Journal of Philosophy* **59**: 5–30
Lewis D K 1981b 'Why ain'cha rich?' *Noûs* **15**: 377–80
Nozick R 1969 Newcomb's problem and two principles of choice. In: Rescher (ed.) *Essays in Honor of Carl G. Hempel*, Reidel, Dordrecht, The Netherlands. Reprinted in Campbell and Sowden (1985)
Skyrms B 1982 Causal decision theory. *Journal of Philosophy* **79**: 695–711
Skyrms B 1990 *The Dynamics of Rational Deliberation*. Harvard University Press, Cambridge, MA
Sobel J H 1978 *Chance, Choice and Action: Newcomb's Problem Resolved*. University of Toronto, ON

J. Collins

News: General

The communication of news has been part of a storytelling tradition since time immemorial. Storytelling to a recurring public is a way of making common sense of the environment. Stories and news about people, friends, and relatives have always passed through the grapevine. Both news and rumor give nourishment for social discourse; something that can be adapted to local experiences in the company of a small group of familiar faces (see *Rhetorical Analysis*).

1. *Definition and Background*

In its modern and more formal usage 'news' refers to certain types of texts produced by the personnel, institutions, and methods of news media. The news media today include news agencies, newspapers, news magazines, radio and television, and Internet services. Each medium has its own form linked to some means of presentation, which is peculiar to their dominant technology of dissemination. However, news journalism also has some common features of content, irrespective of medium, based on similar methods of choosing stories, retrieving information, and of presenting facts effectively.

A discussion over what are the typical contents of 'news' has been a part of news sociology for as long as the subfield has existed, i.e., since the 1960s. There is not as yet any widespread agreement as to the number and kind of criteria. But researchers emphasize some qualities of news texts more often than not. News is supposed to tell something that has really happened recently, to be important for the nation or for the community, and to interest a substantial part of the audience. Thus Mitchell Stephens (1997, p. 9) defines news as 'new information about a subject of some public interest that is shared with some portion of the public.' Brian McNair (1998, p. 4) stresses that 'news claims to be (i.e., is presented to its audience as) truthful statement about, or record of, some hitherto unknown (new) feature of the actual, social world.'

Even so, journalists thrive on rumor but do not necessarily publish all that they hear. Different media may differ in what they consider a threshold for what can be made public; sometimes depending on their means of presentation, the risk of litigation, or on the ideals of journalism media adhere to. News in popular tabloids differs in many respects from news in television and in the quality press.

News as a social habit existed before the first newspaper appeared in Antwerp during a siege of the city in 1605, and even before the printed word. Tidings, rumor, and bits of information from near and far away could be learned at certain places: in the corners of a village square, in postal offices, in coffee houses and taverns, in front of churches, at the harbor, in parks etc. *Gazetta* was the name of a small Venetian coin for the price of which you could purchase a newssheet carrying commercial and other news at the Rialto Bridge in seventeenth century Venice. From the sixteenth century, and long into the nineteenth century, flysheets provided stories comparable to the sensational tabloid news of today: romantic, adventurous, and melodramatic events; supernatural phenomena; the royal court; executions, wars, and witchcraft etc.

Regularly published newspapers, usually weeklies, were established at the beginning of the seventeenth century in England, the American colonies, Holland, Denmark etc. They were operated by jobbing printers who also doubled as postmasters and booksellers. Since few, if any, journalists were employed, printers took their news generally from other publications, mostly from the capital, from foreign newspapers, and from letters to the editor. With regularity of publication newspapers could report continuing stories, largely as a series of minor incidents successively published. Early newspapers almost exclusively used foreign news. Censorship or the threat of censorship made local news more dangerous to publish. Over time, as censorship was eased or lifted, national news stories became a more important part of the news content (Wilke 1987).

Periods in national media histories are sometimes linked to major changes in media technologies, in economy, and to the size of the audience, which can be compared, cross-nationally to some extent. Periodization of the developing style and content of journalism is often linked to changes in the political system where the causal links are sometimes evident, sometimes unclear. Commonly used labels in media histories are

the early elite press, the party press, the mass press, national public service vs. commercial broadcasting etc.

The development of journalism was, until the mid-nineteenth century, closely bound to the book trade, in content as well as in authors. As the press expanded in terms of the number of fully employed journalists, journalism became more distinct. Divergent and seemingly contradictory trends have been observed cross-nationally. However, a general trend seems to be toward more variety in journalistic genres and methods of expression, especially after radio and television became the fastest news media. In the USA and in many countries of Europe two basic styles of journalism has been identified since the mid-nineteenth century, one discursive and another sensational. There is simultaneously a tendency toward a 'long journalism' with a slightly 'top-down' perspective, and another toward a more grassroots-oriented 'bottom-up' populist perspective.

The first trend has produced stories set within a wider frame of time and space. In a study of worklife journalism, criminal and accident reporting in *New York Times*, the *Chicago Tribune* and the *Oregonian* between 1894 and 1994, Barnhurst and Mutz (1997) demonstrated that reporting developed toward longer stories, wider contexts, and more analysis. The 'journalists identified individuals less often by name and more often by demographic group' (Barnhurst and Mutz 1997, p. 41). Many events nowadays do not become newsworthy unless they fit into 'a longer body of interpretations and themes' (Barnhurst and Mutz 1997, p. 41). This development has been possible by the mobilization of substantial journalistic resources, which again is dependent upon advertising revenues and upon some ownership concentration in the audience market.

The other trend—mainly for apolitical tabloids—has been toward spot news as entertainment: centered both on ordinary individuals and on celebrities set within dramatized events where persons, actions, and feelings are foregrounded. Benjamin H. Day, pioneer of the tabloid genre from the 1830s, said that 'News, properly so called ... must generally tell of wars and fightings, of deeds of death, and blood, of wounds and heresies, of broken heads, broken hearts, of accidents by fire or flood, a field of possessions ravaged, property purloined, wrongs inflicted, rights unavenged, reputations assailed, feelings embittered, and oppressions exercised by nations, communities, or individuals' (quoted in Huntzicker 1999, pp. 14–15). This tabloid formula seems to have been employed in urban boulevard newspapers in all kinds of political systems except the totalitarian ones.

2. *The News Paradigm*

Social scientists often use changes in technology as an index of wider social developments. Thus, the electric telegraph—invented in 1844—and the practice of news agencies have served a wide range of explanations of why and how journalism has changed since the mid-nineteenth century. This infrastructure is said to produce alternately: the 24-hour news cycle, the 'inverted pyramid form' of presentation, and 'objectivity' in journalism. Jürgen Wilke (1984) found e.g., that between 1856 and 1906 reported events that had occurred in the past 24 hours increased from 11 percent to 95 percent in the German press. The international news agencies like Reuters have been described as pioneers of neutral reporting and of the 'lead' or the inverted pyramid form (Read 1999). Likewise, Donald Shaw (1967) explained the alleged fall of the party press in the USA in the 1880s as a result of the increasing use and dependence on wired and politically neutral news by the newspapers. Many of these explanations have been challenged recently as technological determinism, ignoring journalism as a genuinely cultural phenomenon following its own trajectory.

Richard Kaplan (1997) found, for example, that the fall of the party press in Detroit did not coincide with the increased use of wired news. Michael Schudson (1978) interpreted the rise of 'objectivity' in journalism as influenced by a philosophical pragmatism in contemporary USA at the beginning of the twentieth century and by a detached professionalism in a growing number of occupations at the time. The canon of 'objectivity' became current in the late 1920s. But even by the turn of the twentieth century establishing facts was an ideal in journalism. 'Objectivity' in addition asked for balanced facts and opinions, which required new discursive practices (see *Journalism*).

The US newspaper industry had an unprecedented growth between 1880 and 1920 with widely expanding circulation, which included new layers of society in the informed public. In combination with previous breakthroughs in paper and print technology—which made cheap copy prices possible—it transformed the press into our first mass medium. The working environment and the job definition of reporters as different from writers changed rapidly (Schudson 1995), and the audience easily became strangers as the circulation area expanded. 'Journalists were thus becoming something of a separate class, a separateness indicated by increasing numbers of articles and lectures by journalists intended to explain how the press worked' (Dicken-Garcia 1989, p. 230).

This was also a time ripe with experiments in journalism. The various ideals, values, criteria, definitions, etc., which were eventually assimilated into the modern idea of proper news, became a way of thinking in journalism from the beginning of the twentieth century. Høyer (1995) constructs a 'news paradigm' out of five distinct elements emerging in this period: (a) the event, (b) the idea of neutral news criteria, (c) the news interview, (d) the inverted pyramid method of presentation, and (e) objectivity. This combination of journalistic methods and genres appeared gradually as

separate social and cultural inventions over a period of less than 50 years. The location of most innovations was New York and London, which were also centers of applied modern newspaper technology in the latter part of the nineteenth century. The five elements can be specified as follows:

(a) The event may be defined as an interaction, an accident, a meeting, a crime etc. which ideally shall begin and end between deadlines. With regular news services and the electric telegraph the time frame of a new event became the 24-hour publication cycle.

(b) News criteria are the use of nonsubjective criteria for the selection of news to be published.

(c) The news interview has now become the main method of collecting information. It is a peculiar kind of conversation where the information source feels obliged to respond to inquiries, without the journalist responding psychologically since the journalsist is talking on behalf of a third—silent but more important—part: the public. This form of interview became widespread in the US press during the 1870s and 1880s and was at first met with outright disdain from the elite, since it gave newspapers more control of what should become public news (Schudson 1994)

(d) The 'inverted pyramid' is a rhetorical form. It became a standard from the 1890s. The inverted pyramid emphasizes fast and short presentation of the event, in the headlines and especially in the opening paragraph of the story, known as the 'lead.' The lead contains the major facts and actors involved. The net result of this method was stories lacking an internal chronology with implied explanations. Stories were supplied with chronologies turned downside up, and written in a dry and down-to-earth style of language, hiding the journalist as the author responsible for the story. By the flexibility of computer technology the lead technique in presentation has recently been modified. Graphics, frames for facts and summaries of previous events etc., serve as satellites to the main story, and allow the journalist to start closer to the chronological beginning of the reported event.

(e) Objectivity—or more succinctly 'balance' of information from contending sources of information—is obtained by using many different authorities as sources. The number and variety of information sources attributed is sometimes given as the operationalized definition of 'objectivity' in journalism. This form of neutrality is different from the scientific or philosophical idea of objectivity, where the quality of data is methodically evaluated and intersubjectively controlled by repetitive tests.

The 'omnibus' newspaper became the arena for these diverse innovations; it became the answer to a growing heterogeneity of newspaper readers and to their diversity of interests. Increased advertising, cheap newsprint, etc. allowed for more space for journalism and for typographical experiments. The sum of stories is implicitly supposed to tell the most important of the latest events.

There are some apparent interconnections between the diverse elements of the news paradigm. The electric telegraph, which inspired the 24-hour news cycle, also gave a time frame for news stories to be fresh and new, which subdued long-winded histories with complex cause and effects linkages. The event by definition became something that happened in a very short time. The emphasis on actuality, variety, brevity, and speed easily inspires the inverted pyramid form of presentation. When facts, events, and drama are foregrounded without any explicit interpretation by the journalist, the trustworthiness of facts and opinions becomes essential. The main method of gaining authority or 'objectivity' in journalism became the news interview of expert and elite information sources.

3. *Diachronic and International Perspectives*

The technology and economics of news production soon spread to many countries, while the diffusion of 'the news paradigm' lagged behind. The marked expansion in circulation between 1880 and 1920 observed in the USA has e.g., been documented in Denmark, Finland, and Norway (Thomsen 1972, Tomilla 1988, Høyer 1995). Scattered data suggest that it took several more decades before the US-inspired news paradigm became widely adapted in most European countries, except for the UK. In 1905 and 1908 major newspapers in Copenhagen and Stockholm respectively adopted 'American journalism' (Marzolf 1982), while Jean Chalaby (1996) suggests the 1930s for France. Countries like Spain and Russia lagged even further behind for obvious reasons. The local press seems to lag behind the major cities everywhere. The historical geography of changing journalism is different from the changing media technology, even if the diffusion process has not yet been mapped out in more detail.

The most common reason given for this cultural lag between the Anglo-American world, Scandinavia, and Continental Europe is the strong tradition of a party press following a sustained period of censorship. In France and the Germanies, for example, the original ideal of a journalist was the mature gentleman of the press—the *redakteur*—commenting on the affairs of the world and interpreting trends in politics, arts and science from a given philosophical position (Esser 1998). With regard to Central and Eastern Europe, Owen V. Johnson (1999, p. 8) points to how journalism overlapped with politics and literature from the latter half of the nineteenth century to the Soviet period. Young writers made their living as newspaper journalists until they could make their mark as writers or politicians. In contrast, the Anglo-American ideal was the outgoing aggressive news hunter unbound by political allegiances (Köcher 1986).

These differences have been explained by the longer periods of censorship in Continental Europe, which

also produced a conception of 'state-professions' as regulated by the authorities, different from the Anglo-American system. By contrast to the US journalist the German *redakteur* demonstrated independence by giving personal and informed opinions which could deviate from more officially accepted views.

Jean Chalaby (1996) focuses more on cultural factors like the sharper hierarchy of literary genres in France as compared to Britain, fostering a literary and enigmatic journalism directed toward an elite public as the interpretive community. This tradition continued for decades after censorship was abandoned (Esser 1998, Johnson 1999). On the other hand, the introduction of an authoritarian regime in democratic Estonia 1933–40 did not prevent the breakthrough of more modern and US-inspired journalistic methods (Harro 2000). In Scandinavia the party press subsided gradually somewhere in the 1970s and 1980s. But even before journalistic objectivity was given priority over alternative journalistic values, other elements of the news paradigm were adopted. Nonseid and Høyer (1998) found in a content analysis of four Oslo newspapers 1908–30 only some traces of evidence for Donald Shaw's (1967) thesis that news agencies influenced journalism more generally. They studied occurrences of 'the inverted pyramid,' which became a more common story telling method during the period, both for wired and self-produced stories. But, except for the wired news, a clear majority of news throughout the period was still cast in different and older forms of story presentation. The influence of news agencies could be traced, but was weak. Alternative cultural and political factors were apparently stronger.

The former USSR had different news criteria and different concepts of autonomy in media. The media should first of all be concerned with what the public ought to know according to the communist credo. The news should depict the party line as the guideline for correct behavior. The staple news diet contained all kinds of stories about economic and social progress within socialist countries. Soviet news was not sensational and rarely reported accidents of any kind. TASS—the news bureau—provided between a third and a half of the contents of published Soviet news. In spite of this, events were not necessarily defined by the 24-hour news cycle. The timing of publication was mostly bound to political routines and necessities (see *Mass Communication*: *Normative Frameworks*). As a reflection of this conception of what is news Halliki Harro (2000) demonstrates in a recent content analysis of Estonian newspapers 1920–99 that the 'inverted pyramid' hardly existed during the Soviet period.

These international examples demonstrate a variety that denies the simple connections between technological and cultural innovations claimed by technological determinists. Even if there may be some links between media technology and the journalism, the connection is certainly indirect and not synchronous. Media technology seems to have diffused much faster and more widely than innovations in journalism, which is 'part of, and responding to, a deeper cultural movement' (Schudson 1991, pp. 175–189).

4. *News in Electronic Media*

Radio news started gradually in the 1920s, first by some amateur ham operators reading directly from newspapers to an unknown public. Music, and to a lesser extent lectures and humor, were the great attractions of the new medium, not news. Until the mid-1930s few journalists were assigned to radio news. Only in 1934 did news become an independent department in the British BBC (Schlesinger 1978). Typically, at the beginning, news stories consisted of short bulletins of wired news from major news agencies.

From the outset the press tried hard to repress, and later to control, radio news. The press in Europe most often controlled the main news agencies from where broadcasting corporations got most of their news. In return for access to the wired news the press in the 1920s and early 1930s demanded that broadcasters limit transmission to one news program at around 7.00 p.m., after listeners had got their evening newspapers.

Scarcity of frequencies regulated by the state, as a common good, made public service broadcasting more vulnerable to government control compared to the press. Authorities were determined to avoid controversial topics in radio. Political parties, trade unions, and special interest organizations were thus not welcome as topics or actors in the news. This made radio news rather unrealistic and broadcasters became the voice of some unidentified official establishment.

World War II broke most of these restrictions. Backgrounds and interpretations were needed to make sense of what happened and to bolster national morale. Day-round bulletins, mixed with reportage and interviews, now became the rule. Radio reporters became identifiable persons, sometimes celebrities, as they reported directly from the war theaters. Instrumental in this process was also the tape-recorder, which came into use in 1939.

Television news started out as a copy of the cinema newsreel, with presenters heard but not seen. But the audience wanted to see the faces of announcers, who eventually developed into the role of the anchorperson and moderator. The anchorperson connected the steadily differentiating parts of a televised news story: a short announcement amounting to 'headlines,' a lead-type message followed by interviews, reports from eyewitnesses or reporters on location, and expert comments back in the studio to wind up the story. A study of 17 newscasts from eight countries identified different styles of anchorpersons in Western Europe: the 'Germanic' newsreader who briefly introduces the topic and leaves the substance of information to field

reporters, as contrasted to the 'Romance' moderators who provide lengthy transitions between stories and subjective comments in the style of an editorial writer (Heinderyckx 1993).

The Internet promises a lower threshold of access to the distribution of news and views for the regular public in the industrialized world. But to be effective as a news channel it seems that the Internet so far must rely on existing networks of people with special interests. The news sites serve as points of convergence both for a like-minded public internationally and for all sorts of changing information and images within specific topic areas (see *Information Society*). Within finance and some scientific disciplines, the Internet already serves as a major publishing channel or as a marketplace for news. But for the moment it lacks the ability to reach the broader audience during a regular news cycle. For the Internet to become a general news channel, its use must expand considerably above its present level of penetration.

See also: Agenda-setting; Alternative Media; Broadcasting: General; Celebrity; Entertainment; International Communication: History; Journalism; Mass Communication: Normative Frameworks; Mass Communication: Technology; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Mass Media, Representations in; News Interview; Printing as a Medium; Public Relations in Media; Rhetorical Analysis

Bibliography

Barnhurst K G, Mutz D 1997 American journalism and the decline in event-centered reporting. *Journal of Communication* **47**(4)

Chalaby J 1996 Journalism as an Anglo-American invention. A comparison of the development of French and Anglo-American journalism 1830s–1930s. *European Journal of Communication* **11**(3)

Dicken-Garcia H 1989 *Journalistic Standards in Nineteenth-century America*. University of Wisconsin Press, Madison, WI

Esser F 1998 Editorial structures and work principles in British and German newsrooms. *European Journal of Communication* **13**(4)

Gans H J 1979 *Deciding What's News. A Study of CBS Evening News, NBC Nightly News, Newsweek & Time*. Pantheon Books, New York

Golding P, Elliot P 1977 *Making the News*. London, Longman

Harro H 2000 Changing journalistic conventions in the press. Empirical studies on daily newspapers under different political conditions in 20th century Estonia. Ph.D. dissertation, University of Oslo

Heinderyckx F 1993 Television news programs in Western Europe: A comparative study. *European Journal of Communication* **8**(4)

Høyer S 1995 *Pressen mellom teknologi og samfunn* (The press between technology and society). Universitetsforlaget, Oslo

Høyer S, Lauk E 2000 The paradoxes of the journalistic profession. An historical perspective. In: Björk U J, Nordenstreng K (eds.) *A Hundred Years of the International Journalist*. University of Luton Press, Luton, UK

Huntzicker W E 1999 *The Popular Press, 1833–1865*. Greenwood Press, Westport, CT

Johnson O V 1999 The roots of journalism in Central and Eastern Europe. In: Aumente J, Gross P, Hiebert R, Johnson O V, Mills D (eds.) *Eastern European Journalism. Before, During and After Communism*. Hampton Press, Cresskill, NJ

Kaplan R 1997 The American press and political community: Reporting in Detroit 1865–1920. *Media Culture & Society* **19**(3)

Köcher R 1986 Bloodhounds or missionaries: Role definitions of German and British journalists. *European Journal of Communication* **1**(1)

McNair B 1998 *The Sociology of Journalism*. Arnold, London

Michell S 1997 *A History of News*. Harcourt Brace, Fort Worth, TX

Nonseid J, Høyer S 1998 Diffusion of the news paradigm in Norway 1908–1930. Paper for International Association for Mass Communication Research, Glasgow, July

Read D 1999 *The Power of News. The History of Reuters*, 2nd edn. Oxford University Press, Oxford

Schlesinger P 1978 *Putting 'Reality' Together. BBC News*. Constable, London

Schudson M 1978 *Discovering the News. A Social History of American Newspapers*. Basic Books, New York

Schudson M 1991 Historical approaches to communication studies. In: Bruhn Jensen K, Janowski J W (eds.) *A Handbook of Qualitative Methodologies for Mass Communication Research*. Routledge, London

Schudson M 1994 Question authority: A history of the news interview. *Media, Culture & Society* **16**(4)

Schudson M 1995 *The Power of News*. Harvard University Press, Cambridge, MA

Schudson M 1999 The sociology of news production revisited (again). In: Curran J (ed.)

Shaw D L 1967 News bias and the telegraph: A study of historical change. *Journalism Quarterly* **44**(1)

Thomsen N 1972 *Dagbladskonkurrencen 1870–1970* (The Competition between Daily Newspapers 1870–1970). Gads, Copenhagen

Tichenor P J, Donohue G A, Olien C N 1980 *Community Conflict & The Press*. Sage, Beverly Hills, CA

Tomilla P 1988 *Suomen lehdistön historia* (The history of the Finnish Press). Kustannuskiila, Kuopio, Finland

Tuchman G 1978 *Making News. A Study in the Construction of Reality*. Free Press, New York

Wilke J 1984 The changing world of media reality. *Gazette* **39**(3)

Wilke J 1987 Foreign news coverage and international news flow over three centuries. *Gazette* **39**(2)

S. Høyer

News Interview

The news interview is a central means by which journalists gather information that will later be worked up into finished news stories. It is also employed as a finished news product in its own right,

a basic form in which news is packaged for public consumption. The latter usage is particularly important within contemporary broadcasting, as a substantial proportion of news and public affairs programming now takes the form of a journalist interviewing one or more public figures before the media audience.

1. The News Interview as a Genre

The news interview is a familiar and readily recognizable genre of broadcast programming. Its basis in comparatively spontaneous interaction distinguishes it from the fully scripted narratives that comprise most traditional news programs. The news interview also differs from other interaction-based genres of broadcast talk (e.g., talk shows, panel discussions, etc.) by its distinctive constellation of participants, subject matter, and interactional form. In a prototypical news interview, the interviewer is known as a professional journalist rather than a partisan advocate or celebrity entertainer. Interviewees are public officials, experts, or others whose actions or opinions are newsworthy. The discussion normally focuses on matters related to recent news events, is highly formal in character, and is managed primarily through questions and answers. The closest relative of the news interview is the press conference, which shares all of these features, but involves a large number of participating journalists and is held at the behest of the public figure rather than the news media (see *Talk Show in Media*; *Television: Genres*).

2. Origins and Institutionalization

The news interview prototypically involves the confluence of representatives of two important societal institutions—journalism and politics. Accordingly, the history of the news interview is deeply intertwined with the co-evolution of these institutions.

Although it now seems quite natural for journalists to interview elected officials and other prominent public figures, it has not always been so. In the US, interviewing was virtually nonexistent for the first half-century of the nation's existence. Institutions of national government only gradually became publicly accessible, and even as journalists were granted access first to the House of Representatives and later the Senate, verbatim quotations normally were prohibited (Leonard 1986). The aloofness of government officials was matched by the disinterest of most journalists. Newspapers during this period were financed by political parties and were vehicles for editorial opinion rather than reportage in the contemporary sense.

The practice of interviewing can be traced to the rise of the American penny press in the 1830s; the first papers to devote themselves primarily to 'news rather than views' and to employ reporters devoted to the task of newsgathering. But published interviews with public figures did not become common journalistic practice until the late nineteenth century. This new form of journalism first expanded rapidly in the US, and then more slowly in England and other European countries in part at the prompting of American journalists in Europe. This expansion did not occur without controversy—interviewing frequently was attacked as an artificial and unduly intrusive journalistic practice (Schudson 1994).

Although these criticisms would not disappear entirely, the news interview became increasingly accepted as normal journalistic practice in the early decades of the twentieth century. This development roughly coincides with the growing stature and professionalization of journalism, and the shift within government from backstage intragovernmental negotiations to public relations as tools of governance. Accordingly, the three US presidents most responsible for institutionalizing the presidential press conference—Theodore Roosevelt, Woodrow Wilson, and Franklin Roosevelt—were each progressive reformers with ambitious political agendas who used such conferences to build public support for their policies. The advent of television further increased the prominence of news interviews and press conferences, as weekly and eventually nightly news programs were devoted to journalistic questioning of public figures.

In short, what used to be regarded as extraordinary has become standard practice across the political spectrum. Just as journalists were once criticized for questioning public officials, now public officials are subject to criticism if they fail to make themselves sufficiently accessible to journalistic interrogations.

3. Contemporary Norms and Practices

The news interview is not merely a reflection of journalistic and political institutions; it is also a social institution in its own right. Conduct within the news interview is organized around the roles of interviewer and interviewee and is governed by a complex matrix of social norms and conventional practices.

The most fundamental and pervasive characteristic of news interview interaction is that it unfolds as a series of questions and answers. This is both an empirical regularity that typifies news interview talk (Heritage and Roth 1995), and a social norm that the participants are obliged to uphold (Greatbatch 1988). The question–answer framework may seem obvious, but its very obviousness makes it constitutive of the news interview as a recognizably distinct form of interaction. Moreover, underlying this normative framework is a far less obvious substrate of practices that are necessary to produce interaction in manifest compliance with the question–answer norm (Heritage and Greatbatch 1991). These practices include the

systematic avoidance of a wide range of acknowledgement tokens and other responsive behaviors (e.g., uh huh, yeah, right, oh, really) which are absolutely pervasive in ordinary conversation but which become incongruous in a context where the parties are supposed to restrict themselves to the actions of questioning and answering. Both the question–answer norm and the practices that underlie it usually are taken for granted by interview participants, but they may become more fully conscious of the ground rules at problematic or contentious moments, when those rules may be appealed to explicitly as a means or complaint or self-defense.

In building questions, interviewers are sensitive to two further journalistic norms which are difficult to reconcile. On the one hand, they are supposed to remain formally neutral in their conduct. While absolute neutrality is an unattainable ideal, interviewers do strive to maintain a 'neutralistic' posture by restricting themselves to just asking questions, avoiding all forms of acknowledgement, and avoiding flat assertions except as prefaces to a question (Clayman 1988) or as attributed to a third party (Clayman 1992). On the other hand, interviewers are supposed to be adversarial in their treatment of public figures and should not allow the latter to use the interview as a personal soapbox. Interviewers pursue the ideal of adversarialness in part through the content of their questions, raising matters that cut against public figures' own interests, and subjecting their previous responses to challenge. Adversarialness is also pursued through the underlying form of such questions (Clayman and Heritage in press b)—for example, by designing questions in ways that narrow the parameters of an acceptable response, by 'tilting' questions in favor of one particular response over others, and by encoding presuppositions that are difficult for the interviewee to counter or refute. An important resource for these various forms of adversarialness are the preliminary statements that interviewers often make when leading up to a question. Such statements are justifiable as providing 'background information' necessary to render the question intelligible to the audience, but they can be mobilized in ways that allow the interviewer to maximize control over the discussion agenda and exert pressure on recalcitrant interviewees. The balance that is struck between the ideals of neutrality and adversarialness is a signature that distinguishes individual interviewers, the news programs on which they appear, and historical periods characterized by dominant styles of interviewing.

Interviewees, in responding, face a different set of cross-cutting pressures. Adversarial questions create an incentive for evasive responses, encouraging interviewees to be less than forthcoming or to shift the discussion agenda in a more desirable direction. However, the normative question–answer framework obliges interviewees to answer straightforwardly, so that failure to do so can be costly. Interviewers often counter such maneuvers with probing follow-up questions and negative sanctions; audience members may infer that the interviewee has some ulterior motive for avoiding the question; and acts of evasion are often singled out in subsequent news coverage (Clayman 1990). Accordingly, interviewees almost always design their evasive responses in such a way as to minimize these undesirable consequences (Clayman and Heritage in press b). They may choose to sidestep the question in an overt or explicit manner, which allows for equally explicit forms of 'damage control'—for example, justificatory accounts and displays of deference to the interviewer. Alternatively, when evading the question covertly, they may take steps to obscure what is transpiring—for example, by giving it the surface form of an answer.

Many contemporary interviews involve multiple interviewees who represent diverse and frequently opposing viewpoints. The panel interview format is attractive to broadcasters, not only because it promotes varying degrees of dramatic conflict (Olsher in press), but also because it creates a division of labor that helps to reconcile the divergent ideals of neutrality and adversarialness. With partisan interviewees playing the role of adversary vis à vis one another, the interviewer is left free to act as an impartial catalyst.

4. *Evolving Styles of Questioning*

Comparative research on the news interview remains underdeveloped, but styles of questioning appear to have changed substantially since the advent of broadcasting. Journalists' questions to public figures have become less deferential and more adversarial during this period. In US presidential press conferences, this shift is apparent not only in the substantive issues raised within questions, but also the manner in which such questions are designed (Clayman and Heritage in press a). Simple one-sentence questions have given way to increasingly complex questions with extended prefatory remarks. Questions have also become narrower in focus, more blunt or direct, and more 'tilted' in favor of a particular answer. If this pattern reflects a more general trend in interviewing, then journalists have come to exert increasing pressure on interviewees to address inconvenient, unflattering, or incriminating topics.

The rise of adversarial questioning appears to have developed somewhat differently in America and Britain. In Britain, a robust tradition of government regulation of broadcast journalism, coupled the absence of competition prior to 1958, combined to foster a highly deferential style of questioning in BBC interviews of the 1950s. When the BBC monopoly was replaced by a duopoly in 1958, the resulting competition fueled a sudden and dramatic increase in adversarial questioning. In America, where government regulation of broadcasting has been compara-

tively minimal and where competitive pressures have been present from the outset, adversarial questioning has grown more steadily from a higher baseline. However, at least for the case of questions to the president, a sharp rise in adversarialness followed the revelations of the Watergate affair and continued to affect presidents in the post-Nixon era.

Whatever its causes, the rise of adversarial questioning has transformed the news interview into a formidable instrument of public accountability. It is now much more difficult for officials to make purely self-serving statements in the context of a news interview. However, this revolution has stimulated a counter-revolution by politicians and public officials—increasingly sophisticated strategies of evasion, aided by a burgeoning cottage industry of media advisors and consultants. Moreover, because adversarialness in news interviews is expressed to officials directly and in public, it may be contributing to a much broader development in Anglo-American culture: namely, a reduction in the social status accorded to agents of government and other institutional authorities, and a reduction in the social distance separating such elites from ordinary citizens (see *Broadcasting: General*; *Journalism*; *News: General*; *Political Communication*).

Bibliography

Clayman S E 1988 Displaying neutrality in television news interviews. *Social Problems* **35**(4): 474–92

Clayman S E 1990 From talk to text: Newspaper accounts of reporter-source interactions. *Media, Culture and Society* **12**(1): 79–103

Clayman S E 1992 Footing in the achievement of neutrality: The case of news interview discourse. In: Drew P, Heritage J (eds.) *Talk at Work*. Cambridge University Press, Cambridge, UK, pp. 163–98

Clayman S E, Heritage J in press a Questioning presidents: Journalistic deference and adversarialness in the press conferences of Eisenhower and Reagan. *Journal of Communication*

Clayman S E, Heritage J in press b The news interview: *Journalists and Public Figures on the Air*. Cambridge University Press, Cambridge

Greatbatch D 1988 A turn-taking system or British news interviews. *Language in Society* **17**(3): 401–30

Heritage J, Greatbatch D 1991 On the institutional character of institutional talk: The case of news interviews. In: Boden D, Zimmerman D H (eds.) *Talk and Social Structure*. University of California Press, Berkeley, CA, pp. 93–137

Heritage J C, Roth A L 1995 Grammar and institution: Questions and questioning in the broadcast news interview. *Research on Language and Social Interaction* **28**(1): 1–60

Leonard T C 1986 *The Power of the Press*. Oxford University Press, Oxford, UK

Olsher D in press Inviting interplay in panel format news interviews.

Schudson M 1994 Question authority: A history of the news interview in American journalism, 1830s–1930s. *Media, Culture and Society* **16**: 565–87

S. E. Clayman

Neyman, Jerzy (1894–1981)

1. Life

Jerzy Neyman was born of Polish parentage in Bendery, Russia, on April 16, 1894. He entered the University of Kharkov in 1912, studying mathematics and statistics. His instructors included the famous Russian probabilistic S. N. Bernstein, and his earliest papers were in the then relatively new subject of measure theory. Neyman also pursued teaching graduate studies in mathematics in Kharkov from 1917 to 1921, interrupted very briefly by an arrest as an enemy alien by the Russian government. In 1921, Neyman left Kharkov for Warsaw and obtained (with the help of Sierpinski) a position there that permitted him to continue his studies, culminating in his doctoral dissertation in 1923 on statistical problems in agricultural experimentation.

After teaching in Warsaw and Cracow, Neyman obtained a postdoctoral fellowship to study under Karl Pearson at University College, London in 1925, where he first met some of the leading statisticians of his day, including W. S. Gosset ('Student'), R. A. Fisher, and Egon S. Pearson (the son of Karl Pearson, and an important statistician in his own right). The next year Neyman then obtained a fellowship from the Rockefeller Foundation and studied in Paris, attending mathematical lectures by Borel, Lebesque, and Hadamard. During the summer of 1927 Neyman returned to Poland, and resumed teaching at the Universities of Warsaw and Cracow.

From 1928 to 1934 Neyman remained in Poland, working in both applied and mathematical statistics. He pursued a broad range of applied statistical interests in agriculture, biology, chemistry, and socioeconomics, ultimately leading to his appointment as head of the Statistical Laboratory of the Nencki Institute of Experimental Biology. But at the same time Neyman also continued to collaborate with Egon Pearson, work that was to result in some of their most important papers on the theory of statistical tests.

In 1933 Karl Pearson retired as Professor of Statistics from University College. Although Fisher was Pearson's obvious choice as successor, the two had for more than a decade been bitter enemies and, in Solomon fashion, it was decided to divide the Department in two: a Department of Genetics to be headed by Fisher (as Professor of Genetics), and a Department of Applied Statistics to be headed by Egon Pearson. Not unnaturally, Egon Pearson immediately invited his collaborator and friend Neyman

to join him at University College in the new Department of Statistics, initially as a Senior Lecturer and later as Reader. Neyman accepted Pearson's offer, a decision that was to have profound consequences for his career.

Despite the tensions that immediately arose between the two Departments, Neyman initially remained on cordial terms with Fisher (as evidenced both in their personal correspondence and their 1934 exchange at a meeting of the Royal Statistical Society during the discussion of Neyman's paper on confidence intervals). But their personalities and scientific philosophies were so different that conflict between the two was inevitable. Neyman's paper (1935) on the design of experiments (which implicitly criticized some of Fisher's most important work) enraged Fisher, and the clash between the two during the discussion of Neyman's paper led to a complete and permanent severing of relations between them. (Neyman's unfortunate 1961 retrospective polemic provides a useful illustration of just how bitter the dispute between the two became, although it exaggerates Fisher's role prior to the 1950s: after the first several years, Fisher largely ignored Neyman in print rather than attacking him until the middle of the 1950s.)

In 1937 Neyman toured the USA, lecturing both at the U S Department of Agriculture (at the invitation of E. Deming) and a number of major universities, resulting in the offer of a position at the University of California at Berkeley. Attracted by the opportunity to build from the ground up a statistical group, Neyman accepted the offer and permanently left London for Berkeley in 1938.

Over the next several decades Neyman used his intellectual gifts, force of personality, and international standing in the statistical community to make the University of California at Berkeley one of the leading centers of statistical research in not just the USA, but the world. In 1955 the continuing expansion of Neyman's Statistical Laboratory led to the formation at Berkeley as an independent Department of Statistics.

Throughout his career at Berkeley Neyman continued to vigorously promote statistics at the international level as well in the USA; particularly noteworthy in this regard being the institution of the Berkeley Symposia on Probability and Mathematical Statistics, held every five years from 1945 until 1970. International visitors also came to Berkeley on a regular basis, such visits often either arranged by, or due to Neyman.

During his long life Neyman received numerous honors, many of these coming after his retirement in 1961. (The retirement was in name only. In an unusual exception to Berkeley's rules regarding Emeritus status, Neyman was permitted to continue as Director of the Statistical Laboratory.) In addition to being elected a member of the National Academy of Sciences in 1964, Neyman was also elected as a foreign member of the Royal Society, and the Polish and Swedish Academies of Science. Other honors included medals (the US Medal of Science and the Guy Medal in gold of the Royal Statistical Society), and honorary doctorates (from Berkeley, the University of Chicago, the University of Warsaw, the University of Stockholm, and the Indian Statistical Institute).

Neyman remained active to the end, running a weekly seminar at which many of his visitors and guests spoke. He died in Berkeley, California on August 5, 1981, after a brief illness.

2. *Scientific Contributions*

2.1 *Hypothesis Testing*

Although Neyman wrote papers in several areas both prior to his collaboration with Egon Pearson, and after his departure for the USA, the decade from 1928 to 1938, when he collaborated with Pearson, marked the period of his most important and lasting contributions to statistics. The first of these (Neyman and Pearson 1928) introduced the concepts of two types of error (type one and type two error), and the likelihood ratio, using the latter to derive as special cases some of the standard tests then in use. Ironically, in the light of Neyman's later violent opposition to Bayesian methods, Pearson refused to be the co-author of a second paper, drafted by Neyman (1929), because it appealed to the use of prior probabilities. Despite this disagreement, however, their collaboration continued, and in the short period of six years (from 1933 to 1938) they proceeded to lay the foundations of the modern theory of hypothesis testing.

The paper of Neyman and Pearson (1933) on statistical tests, formalized the process of hypothesis testing by providing a clear mathematical framework. In the case of simple hypotheses the Neyman–Pearson lemma was stated and proved; and in the case of composite hypotheses the concepts of similar and best critical regions were introduced. Then, in a pair of papers that appeared in 1936 they further advanced the theory by introducing the concept of unbiased critical region and illustrated the use of power curves to compare tests (Neyman and Pearson 1936a). They then discussed the role of sufficient statistics and the novel concept of uniformly powerful tests (Neyman and Pearson 1936b). Finally, in a last paper (Neyman and Pearson 1938) they returned to the question of the existence of unbiased critical regions.

2.2 *Statistical Estimation*

Neyman introduced the terminology of confidence interval, in his 1934 paper on sampling (Neyman 1934). Originally Neyman deferred to Fisher's priority (Fisher 1930) in discovering the basic underlying idea

(that one could find sets having coverage probabilities independent of the parameter being estimated). But after his break with Fisher, faced both with Fisher's criticism and indeed denial that Neyman's confidence intervals were fiducial intervals at all, Neyman proceeded on his own path, developing a complete theory of confidence intervals (Neyman 1937).

2.3 Other Work

Neyman made many other important contributions to mathematical and applied statistics. Notable examples include his work on sampling theory, in particular what is currently termed Neyman allocation in stratified sampling (Neyman 1938). (Neyman notes in the introduction to the paper that the problem was posed to him during his lectures in 1937 at the US Department of Agriculture, one of the proposes being Milton Friedman) Other contributions included the so-called 'smooth' goodness of fit tests, models for contagion, BAN (best asymptotically normal) estimation, and so-called $C(\alpha)$ optimal tests of composite hypotheses. Much of his later work was applied and done in collaboration with his long-time colleague and friend, Elizabeth Scott.

3. Influence

It would be hard to overstate the impact that Neyman had on both the subject of modern statistics and the statistical profession. Although the English school of statistics, led (among others) by Galton, Karl Pearson, Weldon, Edgeworth, Gosset, Yule, and Fisher, both transformed the subject and vastly extended its scope, the process was undoubtedly hindered by concerns and doubts regarding both its conceptual underpinnings and rigorous mathematical structure. By advancing a purely frequentist setting for the subject, Neyman made the subject palatable to those suspicious of its Bayesian roots; and by presenting it as a natural branch of modern mathematics, using precise definitions and presenting many of its problems in purely mathematical guise, Neyman made it an attractive area of research for many mathematicians.

This phenomenon was particularly striking in the theory of confidence intervals. Fisher's 1930 paper, 'Inverse Probability', in which the term 'fiducial probability' first appears, presents, in all essential respects, a general theory of confidence intervals for the one parameter case. (Fisher noted that in a wide class of cases the probability integral transformation could be inverted to provide a parameter-independent probability statement concerning intervals containing the parameter; the theory presented by Fisher even contained a sample theory interpretation of the probability in question.) But, approaching the subject from the perspective of inductive inference (as opposed to what Neyman later termed 'inductive behavior'), Fisher perceived difficulties in the extension of his approach to the multi-parameter case. Neyman, in contrast, abstracting out what he believed to be the mathematical essentials of the problem (the construction of sets that would contain a parameter with a long-term frequency independent of all unknown parameters), saw no difficulties in such an extension and proceeded to propose one in his 1934 paper. Fisher's subsequent attacks and insistence that Neyman's confidence intervals were not fiducial intervals in the sense intended by Fisher were then seized on by Neyman, who was quite content from then on to portray his own theory as one clearly different from Fisher's. (It is indicative of the magnitude of the break between the two that is his 1937 memoir on confidence intervals, Neyman not only did not cite Fisher's 1930 paper, but in a final footnote asserted that he had independently discovered the confidence interval arising from the t-statistic in 1930.) The conceptual and mathematical clarity of Neyman's approach eventually triumphed in statistical practice and textbook presentations (although even today the tension between the two approaches continues in the guise of the conflict between conditional and unconditional inference).

Similarly, the mathematical elegance and clarity of the Neyman-Pearson theory of tests made it the standard mode of presentation of the subject in many statistical textbooks despite Fisher's equally vehement opposition. (Fisher believed in, and indeed played a crucial role in the development of statistical tests of significance for a single null hypothesis, but thought that when more than one hypothesis was in question, parameter estimation rather than hypothesis testing was indicated.)

Closely tied to the acceptance of Neyman's approach was his impact on the statistical profession, especially in the USA. Under Neyman's leadership, the Berkeley Statistical Laboratory attracted and trained many of the leading figures in mathematical statistics in the decades immediately after the war, including (to name only a few) Erich Lehmann, Joseph Hodges, David Blackwell, Henry Scheffé, and Lucien Lecam.

Many of Neyman's papers from 1923 to 1945 are conveniently collected in two volumes published in 1967 by the University of California Press, Berkeley and Los Angeles: A Selection of Early Statistical Papers of J. Neyman, and Joint Statistical Papers by J. Neyman and E. S. Pearson.

See also: Estimation: Point and Interval; Fisher, Ronald A (1890–1962); Significance, Tests of; Statistics, History of

Bibliography

Fisher R A 1930 Inverse probability. *Proc. Camb. Phil. Soc.* **26**: 528–35

Neyman J 1934 On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Soc. Ser. A.* **97**: 558–625
Neyman J 1935 Statistical problems in agricultural experimentation (with K. Iwaskiewicz and St. Kolodziejczyk). *J. Roy. Soc. Ser. B. Suppl.* **2**: 107–80
Neyman J 1937 Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A* **236**: 333–80
Neyman J 1938a Contributions to the theory of sampling human populations. *Journal of the American Statistical Association* **33**: 101–16
Neyman J 1952 *Lectures and Conferences on Mathematical Statistics and Probability* 2nd ed., rev. and enl. Graduate School, US Department of Agriculture, Washington, DC
Neyman J 1950 *First Course in Probability and Statistics.* Holt and Co., New York
Neyman J 1961 Silver jubilee of my dispute with Fisher. *Journal of the Operations Research Society of Japan* **3**: 145–54
Neyman J, Pearson E S 1928 On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20**: 175-240 (Part I), 263–94 (Part II)
Neyman J, Pearson E S 1933 On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A* **231**: 289–337
Neyman J, Pearson E S 1936a Contributions to the theory of testing statistical hypotheses. *Stat. Res. Mem* **1**: 1–37
Neyman J, Pearson E S 1936b Sufficient statistics and uniformly most powerful tests of statistical hypotheses. *Stat. Res. Mem.* **1**: 113–37
Neyman J, Pearson E S 1938 Contributions to the theory of testing statistical hypotheses. *Stat. Res. Mem.* **2**: 25–57
Reid C 1982 *Neyman—from Life*. Springer-Verlag, New York
Zabell S L 1992 R A Fisher and the fiducial argument. *Statistical Science* **7**: 369–87

S. L. Zabell

Nietzsche, Friedrich (1844–1900)

Nietzsche was born on October 10, 1844, in Röcken near Leipzig into a family with a long tradition of pastors. A defining traumatic experience of Nietzsche's childhood was the illness and death of his father, who died in 1849 of a brain disease; Nietzsche feared suffering a similar fate his entire life. One year later the pastor's widow moved to Naumburg on the river Saale (near Leipzig) with Nietzsche and his sister Elisabeth, born in 1846. Having attended the municipal boys' school and the Gymnasium of the diocese, the 14-year-old Nietzsche was awarded a vacancy at the renowned state school Zur Pforte. Here Nietzsche received thorough instruction and showed extraordinary talent in ancient languages and German literature. He composed, wrote poems and literary treatises. The content and problems in the philosophy of the later Nietzsche were prefigured in his writings at school, which is borne out by early evidence—even earlier than the remarkable essay on *Fate and History* (1862). Texts from his tenth year of life onwards inform us of Nietzsche's intellectual development, which took the form of a preoccupation with Christian religion and morality, markedly influenced by a study of Greek thought, and was pursued, for social and cultural reasons, in a highly furtive and brooding manner.

In the fall of 1864, Nietzsche began a course of study in theology and philology in Bonn; in 1865 he continued his study of philology in Leipzig and finally decided against the theological career which had been designated for him. Although an investigation concerning Diogenes Laertius submitted for a university competition was awarded a prize, Nietzsche had plans for writing a philosophical dissertation concerning Kant and even considered changing his field of study to chemistry. Neither plan came to fruition. Instead, following completion of his military service (1867–8) with the mounted canoneers in Naumburg, the 24-year-old student was unexpectedly named Professor of Classical Philology at the University of Basel in February 1869 with a recommendation from the influential teacher Friedrich Wilhelm Ritschl.

After a brief tour in the Franco–German War, broken off after some weeks due to severe illness, Nietzsche applied in January 1871 for a vacant chair of philosophy at the University of Basel, but without success. Nietzsche had close contacts to his colleague Jacob Burckhardt, whose course of lectures—later to be published under the title *Observations of World History*—he attended in the summer of 1872. He also maintained a friendship with Richard and Cosima Wagner from the very beginning of his tenure in Basel.

In *The Birth of Tragedy*, his first publication, Nietzsche interpreted Greek culture in the spirit of Schopenhauer as the result of a reconciliation between two drives which mutually cultivate and challenge one another: the Apollinian, standing for those concepts created and formed by art, and the Dionysian, being the deeper lying, formless foundation of the existence of its counterpart. Nietzsche believed that the theoretical approach to science beginning with Socrates and continuing into the present was in crass opposition to life, and hoped this approach might be overcome through Wagner's cultural-political plans. His ambition to act as a political author and educator of the German people was formative for the composition of five lectures *On the Future of Our Educational Institutions*, as well as for a series of shorter essays which were written between 1871 and 1873 and first published in the posthumous works: *Philosophy in the Tragic Age of the Greeks*, which influenced our view of pre-Socratic philosophy decisively. *On the Pathos of Truth* (1872) and *Truth and Lie in an Extra-moral Sense* (1873) contain Nietzsche's radical critique of metaphysics and knowledge, which seeks to prove both truth and knowledge to be measures set by

humans. The *Warning to all Germans* (1873) was conceived as a programmatic essay to accompany Wagner's opera festival in Bayreuth and raised the demand for a kind of *Machtpolitik* under the aegis of culture.

Between 1873 and 1876, Nietzsche published four critical essays on culture and contemporary topics, the *Untimely Meditations*: the first work contains a sharp polemic against David Friedrich Strauss, who is represented as the very model of cultural philistinism in the academic establishment of the time. Nietzsche came out against the historicism of the nineteenth century in the second work and demanded a revision of the value of history for human life, such that historical inquiry might be placed at the service of the living moment. In the third and fourth works, Arthur Schopenhauer and Richard Wagnerare portrayed as paradigmatic examples of the unique individual with a heroic life, to which Nietzsche attached hopes for a rebirth of culture as an aesthetic unity of all life experience.

The first symptoms of physical illness had become apparent during Nietzsche's time at Pforta. Beginning in 1873, Nietzsche was plagued by ever more frequent migraine-like attacks which forced him to interrupt his teaching. In addition, Nietzsche suffered from a chronically upset stomach and severe pain in the eyes. He took a leave of absence from the university for two semesters in the fall of 1876, after various therapies and several spa-visits showed no effect on his health. The intensity and frequency of his attacks increased so dramatically that in 1879, at only 34 years of age, Nietzsche submitted a request for his dismissal from the university. He obtained a pension settlement which was unusually generous for the time, but nevertheless extremely modest. Living on his pension from the University of Basel, Nietzsche led an unstable, lonely existence in constant search of favorable climates. He spent his winters mostly on the French and Italian Rivieras (Venice, Turin, Nice), his summers in Sils-Maria in Engadin. With the exception of occasional visits to Germany, his social contact was limited almost entirely to correspondence with Lou von Salomé, Paul Rée, Franz Overbeck, Carl von Gersdorff, and Heinrich Köselitz. Preoccupation with the history of his own illness was to influence Nietzsche's concepts of 'health' and 'illness' as philosophical and existential categories, and played a significant role in his diagnosis of culture.

With the publication of *Human, All-too-human* in 1878, and following the final break with Wagner, Nietzsche completed a transition to an independent manner of thinking in the style of an exposing critique. This committed him to a new literary style—a kind of essay writing *in nuce*. In the aphoristic works *Mixed Opinions and Sayings* (1879), *The Wanderer and his Shadow* (1880) (both published in a 2nd edition as Volume 2 of *Human, All-too-human*) and *Dawn. With Thoughts Concerning Moral Prejudices* (1881), Nietzsche achieved fundamental insights into the origin, capacity, and importance of moral acts through his genealogical and psychological method. He thus arrived at a devastating result by showing the immorality of morals. With the loss of all transcendent and historical certainties there could only be for Nietzsche an 'experimental philosophy' founded on the basis of the perspectivism of human existence, a condition of the finite worldliness of human knowledge and action. It is this 'experimental philosophy' which Nietzsche draws from physical experience in *The Gay Science* (Books 1–4, 1882).

Nietzsche created his 'ideal' of a creative combination of all human contradictions through the literary rebirth of a Persian sage in *Thus Spoke Zarathustra* (Parts 1 and 2, 1883, Part 3, 1884, Part 4 as a private publication). Here we find the conception of the *Übermensch* or superman, as well as the doctrine of the eternal return of the same. We also find a theory of the will to power as the original unity of all mental and physical strength and potential—a theory intended to lend Nietzsche's practical expectations of a revaluation of all values a metaphysical foundation.

In the years 1886 and 1887 Nietzsche began a retrospective of his previous works, some of which he edited anew, added to, and prefaced. He thereby consciously sought a connection to his previous critical and disillusioning phase, begun with the first book of aphorisms in 1878. Simultaneously, the most important critiques of morality appeared: *Beyond Good and Evil* (1886) and *On the Genealogy of Morals* (1887). In this last work Nietzsche advances the thesis that Judeo-Christian religion has not only provoked a revaluation of all values with its life-hostile morality, but that—in an erstwhile established allegiance with modern science—it threatened the future of humanity. Nietzsche saw salvation from this threat in the merciless destruction of the old values, which would thereby both overcome nihilism and make a renewed revaluation of values possible.

In 1887 came yet another serious deterioration in Nietzsche's health. He nevertheless wrote his *Hymn for Life* at the end of the year, the only musical composition he ever submitted to press, and was extraordinarily productive in the months that followed. In the spring of 1888 Nietzsche wrote *The Case of Wagner* while continuing his main work, *Will to Power*. In the course of the year 1888 he slipped into a kind of ecstatic state which was hardly burdened by sickness. This enabled him to be immensely productive. Having broken off preparations for his main work, Nietzsche concentrated with increasing haste on other projects: a summary of the main themes of the planned work, *Twilight of the Idols* (1889) and *The Antichrist* (1894), a final reckoning with Richard Wagner in *Nietzsche contra Wagner* (1894), the autobiography *Ecce Homo* (1908), and a late collection of poems, *Dionysian Dithyrambs* (1894). Signs of a growing public interest in Nietzsche's work increased,

and Georg Brandes held lectures on Nietzsche in Copenhagen.

An expression of Nietzsche's emotional state in this period are effusive letters in which he predicts immanent decisions concerning his fate. He portrays himself as playing the key role in these decisions. There is evidence of a progressive loss of a realistic view of himself by the end of 1888; Nietzsche seemed to be in a delirium of exaltation. Texts and fragments from the turn of the year document a gradual crossing over into insanity. In January 1889 Nietzsche experienced a mental breakdown, and he lived thereafter in a kind of mental derangement. His physicians' diagnosis—until this day uncertain—was progressive paralysis, that is, a continual decline of the activity of the brain, presumably as the result of a latent, long-term syphilis infection.

Following Nietzsche's brief stay in the psychiatric clinic in Basel and a year-long residence at the psychiatric clinic in Jena, his mother selflessly took over his care. Nietzsche himself was in a state of progressive dementia and lost nearly all physical and mental capacities in the course of the final 10 years of his life. Nietzsche's sister took over his care at her Villa Silberblick in 1897, following the death of their mother. Nietzsche's sister's relationship to the sick man and her connection to his work were rather ambivalent, however.

At the time of his death, on August 25, 1900, in Weimar, Nietzsche had long since become a symbolic figure in European literature. Today he is, as a philosopher, the modern classic *par excellence*.

See also: Existential Social Theory; Historicism

Bibliography

Behler E, Montinari M, Müller-Lauter W (ed.) 1972 *Nietzsche-Studien. Internationales Jahrbuch für die Nietzsche-Forschung*. Walter de Gruyter, Berlin, Germany

Danto A C 1967 *Nietzsche as Philosopher*. Macmillan, New York

Deleuze G 1967 *Nietzsche et la philosophie*. Presses Universitaires de France, Paris

Gerhardt V 1999 *Friedrich Nietzsche*, 3rd edn. Beck, Munich, Germany

Janz C P 1978 *Friedrich Nietzsche. Biographie in drei Bänden*. Carl Hanser Verlag, Munich, Germany

Jaspers K 1936 *Nietzsche. Einführung in das Verständnis seines Philosophierens*. Walter de Gruyter, Berlin

Kaufmann W 1974 *Nietzsche. Philosopher, Psychologist, Antichrist*. Princeton University Press, Princeton, N J

Kaulbach F 1980 *Nietzsches Idee einer Experimentalphilosophie*. Böhlau, Cologne, Germany

Müller-Lauter W 1971 *Nietzsche. Seine Philosophie der Gegensätze und die Gegensätze seiner Philosophie*. Walter de Gruyter, Berlin

Nehamas A 1985 *Nietzsche: Life as Literature*. Harvard University Press, Cambridge, MA

Nietzsche F 1966 *Werke in drei Bänden*, Schlechta K (ed.). Carl Hanser Verlag, Munich, Germany

Nietzsche F 1967 *Werke. Kritische Gesamtausgabe*, Colli G, Montinari M (eds.). Walter de Gruyter, Berlin

Nietzsche F 1975 *Briefwechsel. Kritische Gesamtausgabe*, Colli G, Montinari M (eds.). Walter de Gruyter, Berlin

Reichert H W, Schlechta K (eds.) 1968 *International Nietzsche Bibliography*, 2nd edn. University of North Carolina Press, Chapel Hill, NC

Ross W 1980 *Der ängstliche Adler. Friedrich Nietzsches Leben*. Deutsche Verlags-Anstalt, Stuttgart, Germany

Schmidt H J 1991 *Nietzsche absconditus oder Spurenlesen bei Nietzsche. Part I and II: Kindheit*. IBDK, Berlin

Schmidt H J 1993–4 *Nietzsche absconditus oder Spurenlesen bei Nietzsche. Part I und II: Jugend*. IBDK, Berlin

Volz P D 1990 *Nietzsche im Labyrinth seiner Krankheit. Eine medizinisch-biographische Untersuchung*. Königshausen & Neumann, Würzburg, Germany

V. Gerhardt

Nomads/Nomadism in History

1. The Term 'Nomadism'

Nomadism is a specific way of life practised in the dry belt of the Old World (Fig. 1), a socio-ecological mode of culture (*sozioökologische Kulturweise*) (Scholz 1995), whose internal processes, governing factors, and external appearance essentially obey the elementary 'law' of safeguarding survival. Within the Old World dry belt, nomadism was able to re-appear in new and original forms, at all times and places, fundamentally separate from the evolutionary stages of sedentary life. Nomadism is oriented towards living 'in' and 'with' nature, rather than dominating and exploiting it; as such, it represents a region-specific mode of subsistence that makes optimal use of ecological possibilities and sociopolitical conditions in order to survive. Thus it is an elementary, constitutive, and independent component of the cultural and social evolution of this dryland zone.

2. Debatable 'Stage Theories'?

Pohlhausen (1954, pp. 153, 167) still believed that pastoral nomadism started around AD 950 and that evidence existed of its monocentrist origin. But even the extensive ethnoarcheological studies of the past decades have failed to establish when and where nomadic peoples first appeared. The emergence of nomadism a—significant event in social history—may well have been due to 'the herder of the steppe and oasis farmer having broken away from his agriculturalist brother' (Wissmann 1961, p. 28). According to Lattimore (1940) this, or a similar process, may have

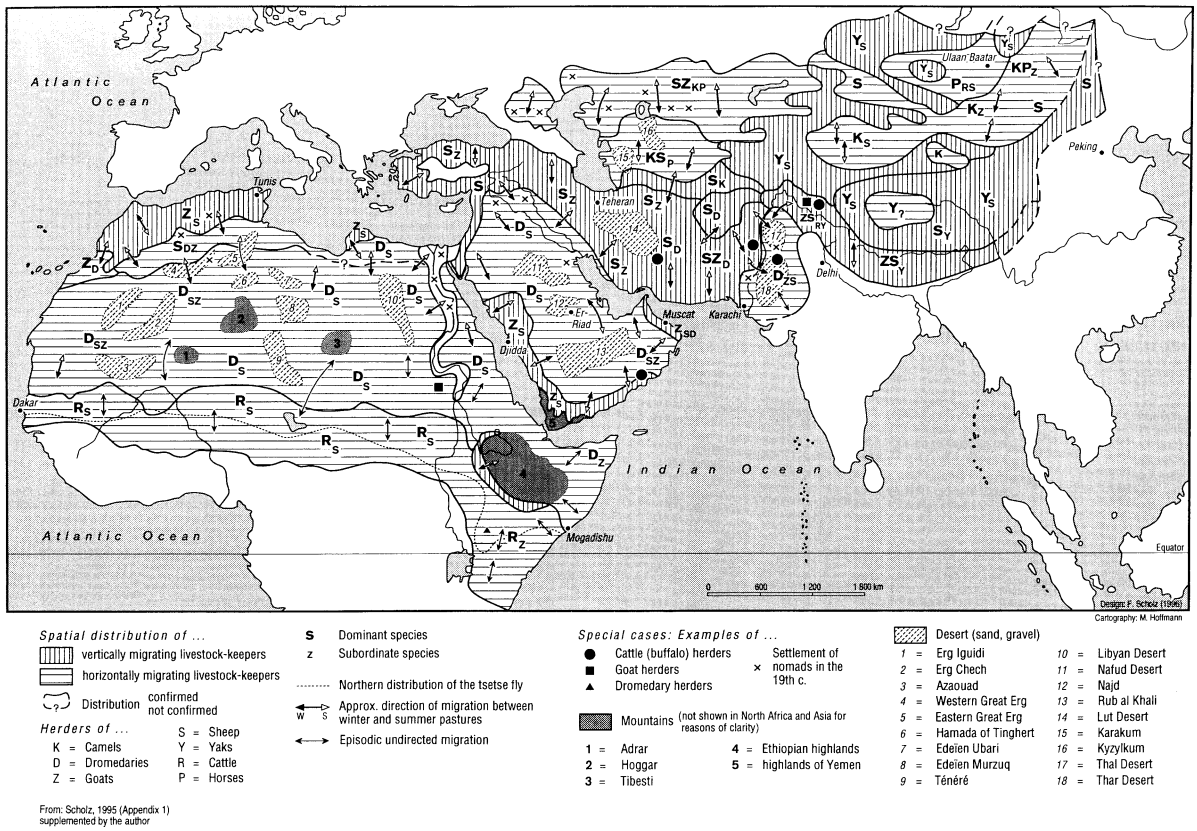

Figure 1
Nomadic Livestock-keepers in the Old World Dry Belt (circa 1900)

occurred at the Inner Asian frontier, the marginal region between settled agriculture (practised by the Chinese) and nonsedentary livestock-keeping (practised by the Mongols). Indeed, it could have occurred anywhere within the Old World dry belt (Fig. 1).

Plausible concepts of the origin of nomadism have sought to convey various (three-) stage theories of cultural and economic history (Vajda 1968). While avoiding a lengthy discussion of such explanations—which have largely outlived their relevance—mention should at least be made of Adam Smith (1723–90) and Friedrich List (1789–1846). In his comments on the evolution of 'productive forces' in agriculture, List (1961, p. 57) expressed the opinion that, in genetic terms, the nomad represents a preliminary stage before the agriculturalist. This view prevailed from the nineteenth century up to late in the twentieth century and believed to be able to locate the centre of origin of nomadism in the Altai region. This theory may seem logical at first sight, but in the end it does not fit the facts and has been considered refuted since Hahn (1891, p. 487) postulated that mixed farming was the prerequisite for the origin of nomadism.

More complex theories of origin were developed by, among others, Pohlhausen (1954) or Dittmer (1965), who believed the predecessors of pastoral nomadism (*Hirtennomadismus*) to have been alpine/mountain nomadism arising from plow agriculture and subsequently, transhumance and cowherding (*Küherwirtschaft*). By contrast, Bobek (1959, p. 272) sees nomadism as a branch of clan farming (*Sippenbauerntum*) and relates this development 'on the one hand to the availability of extensive pasturelands that are only suitable for seasonal use and not really for permanent settlement and, on the other, to the dual economic interests of mixed farmers,' who are able 'to change their mode of subsistence relatively easily and quickly by shifting the emphasis from cultivation to animal husbandry and vice versa.' Basically similar ideas are put forward by Vajda (1968), special mention being made of the conception of a repeated alternation between nomadic and farming life as a result of natural catastrophes, armed conflicts, and economic setbacks. According to de Planhol (1975) the constant demographic surplus of the nomads substantially contributes to this situation.

These different views are plausible, but by no means imperative, in terms of time and space. The processes they describe could have happened anywhere and at any time, even long before the first written references (fourth millennium) to nomads.

3. Ethnoarcheological Findings

These cultural–historical views are, in the final analysis, conclusions drawn from archeological and ethnoarcheological research. Convincing work has recently been published on nomadic societies, which are generally not amenable to research methods based on permanent material evidence (Francfort 1990). These studies have thrown light on prehistoric forms of dwellings and camps, their primary theme being the repeated alternation between sedentary agricultural and mobile livestock-keeping groups, i.e., they provide evidence of the recurring emergency of nomadic groups, of nomadism. These processes have been described for Baluchistan, for the Negev Desert and for the Eastern Desert of Egypt. The view that the nomad did not solely evolve from the steppe farmer has been represented in various models. According to Lees and Bates (1974) the initial stage of nomadism was the transition from agriculture with animal husbandry to specialized irrigation-fed agriculture. On the one hand, this involved a population increase, specialized division of labor, and higher food requirements. On the other hand, animal husbandry was driven away from irrigable land, further and further into the steppes and mountains. Seasonal wanderings over increasing distances were the result, with nomads keeping their own separate herds. The different needs of the nomads and oasis dwellers to safeguard their existence, led to a symbiotic relationship of conflict and coexistence between the groups. The course of this relationship depended on the respective military power and was conductive to the nomads' organizational independence. With reference to Lattimore (1940), Nissen (1980, p. 287) developed a similar hypothesis regarding early Babylonia whose history is marked by recurring invasions by 'so-called nomadic groups … in the fertile grounds.'

In fourth millennium Sumerian texts there are already many pointers to the repeated abandoning of agriculture in favor of mobile livestock-keeping or to the separation of livestock-keepers from agriculturalists; this process probably occurred at the start of the first millennium. Gilbert (1983) even puts forward the view that the developing urban centers the Zagros mountain region caused an increase in the demand for animal products. Simultaneously, however, more intensive agriculture was accompanied by a reduction in local animal husbandry, and niches opened up for mobile livestock-keeping and the nomadic way of life. This process is supposed to have first occurred in the seventh millennium and then again in the third and first millennia BC.

If the domestication of herd animals is accepted as the prerequisite for nomadism, then nomadism presumably did not exist in North Africa before about 5500 BC when the Sahara was a grassland niche which was increasingly occupied by pastoral people. Similarly, livestock-keeping communities would not have occurred in East Africa prior to about AD 500.

It is unlikely that it will ever be known when this transition to, or origin of, nomadism first occurred. Yet such an event is, in the final analysis, less interesting and important than the fact that, given the right natural and social conditions, nomadism could and did appear at different times and places, evidently unconnected with each other, as archeological evidence suggests. It seems to have been the interplay between nature and society, together with the repeated long-term alternation of arid and less arid conditions and short dry periods since the early Holocene, that forced peoples to adopt a peripatetic way of life in order to survive. However, it is likely that social and political changes were no less important, such as the rise and fall of empires, population pressures, and famines and/or tribal conflicts (Planhol 1975). They triggered large- and small-scale migrations, caused nomadic hordes to overrun settled land, and forced settlers to move out into the steppes in times of emergency. In the final analysis, these are the causes of the 'interplay between migratory and sedentary ways of life.'

4. Historical Interpretation and Evidence

The first written reports about nomadic groups date to the fourth and third millennia. These and later scripts were almost all written by sedentarists and reflect their opinions and prejudices. Cuneiform texts, for example, always describe nomadic territory in negative terms, as the home of spirits and demons, the same word being used to denote both the desert and the underworld. In Asia the nomad is portrayed 'as a beast of hell' and 'incapable of civilized life' according to the Chinese Confucians (Khazanov 1981, p. 142), or stereotyped as a 'barbarian,' a wolf or tiger. Little has changed since. However, now it is mainly Europeans who describe the historical role of nomads as destructive or nomadic existence as ahistorical.

The history of nomadism is voluminous, one-sided as regards its sources, doubtless complex and difficult to unravel, and its inner contradictions have never been critically analysed (Braudel 1990). Its contradictory nature consists in the dichotomy between the ephemeral duration and recurrence of nomadic patterns. This will be kept in mind in the following discussion.

Historical descriptions are based on the generally accepted experience that the nomad raided settled land suddenly, like a cavalry attack, and after a while either disappeared or was assimilated. By contrast, the

agriculturalist, natural conditions permitting, occupied nomadic country for permanent or at least long-term settlement and remained in logistic contact with the hinterland. Admittedly, nomads had a 'considerable impact on the social and political development of many sedentary societies with more diversified economies and advanced technologies' (Khazanov 1981, p. 142). Apart from climatic reasons, this process gained importance as the mobility range of nomadic groups was enhanced by more powerful pack animals and the availability of lighter metal equipment. In addition, population pressure on the nomads increased, large (presumably wealthy) sedentary empires tempted nomads to trade or raid, and the scenario of conflict developed between central rule and peripheral tribes (Jettmar 1966), Planhol 1975). Yet these waves of nomadic raids on cultivated land are only part of the story, although sources mention them first of all. Much rarer are reports of sedentarists adopting a nomadic way of life. In this context Jettmar (1966, p. 5) even speaks of a 'victory of nomadism.' Even in Inner Asia, a region of almost classical 'antagonism' between nomads (Mongolia) and sedentarists (China), the situation is more one of overlapping and reciprocity (Lattimore 1988, p. 328). Hence authors have spoken of an almost oscillation of nomadism's economic space or of a recurrent emergence of nomadism.

Indeed, empirical experience suggests that sedentism in conjunction with (oasis) agriculture offered safety and food in times of peace, but it was the mobile mode of life with herd-keeping in deserts and steppes that offered a means of survival in uncertain times. Considering the natural and socio-political settings and the short-and long-term changes occurring in the respective regions of the Old World dry belt, it may be assumed that a switch from sedentism to mobility or from agriculture to mobile livestock-keeping, and hence from agriculture to nomadism, occurred just as frequently as vice versa.

The historical view distinguishes two phases (Bobek 1959):

(a) The first phase differs only slightly from events in prehistoric times as described by archeologists. It is characterized by the keeping of small livestock, close links between agriculture and pastoralism, and probably also short, sometimes climate-induced switches from one to the other, leading to the recurrent separation of herders and cultivators within the same family or the same tribe. There is fairly reliable evidence of this from the Middle East. Similar events probably occurred in the Eurasian steppes although the actual 'nomadization' of the steppe agriculturalists presumably dates to the Scythian period (Iron Age, beginning around 800 BC) (Wissmann 1961, p. 30).

(b) This first phase ends during the second millennium BC. It was followed by First, domestication of the horse in the Eurasian steppes (in Turan about 2800 BC), by orienting entire peoples towards nomadism, by domestication of the camel in the Middle East (end of the second millennium), and resulting in greater mobility and the emergence of mounted warriors and hordes or 'empires' (e.g., the Cimmerians, Thracians, and Scythians, and later the Huns, Tatars, and Timurids). Second, the emergence of 'wealthy' centers of high culture, based on irrigation. They form the backdrop of the conflicts between nomadic and sedentary peoples that shape the subsequent history of this region. The clay tablets in the archives at Mari allowed scholars to infer these conflicts from the appearance of nomadic groups such as the Yaminites and Hanaeans (nineteenth century), the Sutaeans (eighteenth century), the Arameans (fourteenth century), or the Chaldeans (ninth century), in the irrigated parts of the drylands of Western Asia/Mesopotamia. Lasting reminders of this are the fortified towns and villages of Iran, the famous 'walls' built as a defence against the nomadic threat in China (Great Wall, third century BC against the Huns, AD fifteenth century against the Mongols), in phase with Turan or NE Mongolia at the mouth of the Gan, in the Middle East (Limes Syriae), North Africa (Limes Numidiae), Russia (Limes Daciae), or in central Europe (Limes Germaniae/Raetiae).

In the seventeenth century BC the Hyksos descended on Egypt, introducing the horse, and in the eleventh century the Midians overran Palestine and Canaan. Since the ninth century the 'Arabs' (nomadizing Arab = Bedouins), mounted on camels, repeatedly entered fertile Mesopotamia and in the third century they came to Egypt. In the sixth century, the Persian empire in the northern Middle East expanded northward (beyond the Jaxartes River) into the lands of the sedentarized steppe peoples. Like the later westward expansion of Hun rule (starting in the first century), this interrupted the sedentarization process that prevailed in the Eurasian steppes at the end of the Scythian period. Subsequently the entire region experienced the upheavals of the migration of peoples. From then on, changes occurred more rapidly, and the nomad/Bedouin continued to play a major role.

In the first two centuries AD Roman rule in the Middle East triggered a wave of nomadization; in this case the shift from a non-Bedouin to a Bedouin way of life. This process continued, as the decline of Marib Oasis (Yemen, sixth century) and the contemporaneous revival of Bedouin life show. Over the following centuries it received a new stimulus from the spread of Islam, particularly from the Islamic-Turkish nomadic armies of the Seljuqs (1048) in the mountains of Anatolia and Iran, Iranian tribes as far as Baluchistan, Sind and Punjab, and the Beni Hilal (1051) and Beni Sulaym in North Africa. In the Khanate of the nomadic Khazars, north of the Caucasus, the sedentarization process that had begun in the eighth century was succeeded by a wave of renomadization as a result of conquests by the Pechenegs and Cumans in the eleventh and twelfth centuries. The same happened to the Polovtsy (Kipchaks) after

the Mongol invasions in the thirteenth century. In the Iranian mountains, too, they triggered a real nomadization wave described as 'medieval Bedouinization' Planhol 1975, p. 238). The Mongol impact stretched as far as Western Asia, where for example, Baghdad was destroyed in 1258, the irrigation systems of Mesopotamia were devasted, arable land reverted to pasture, and the desert was the only safe refuge for any length of time Wissmann 1961, p. 47). Similar consequences followed the attacks by Timur in the fourteenth century or by the nomadic Turkmen tribes in Anatolia in the sixteenth century, where 'nomadism' disappeared in the West and began to flourish in the East (Braudel 1990, p. 139). Until the nineteenth century Egypt experienced repeated phases of movement and settlement, and the shift between sedentarization and nomadization also continued into the nineteenth century in the Eurasian steppes, where it was quite common among the Turkmen of southern Kazakhstan for 'a nomad who had lost his livestock to settle down on land, or for a rich farmer to buy a herd and to shift again to nomadism' (Khazanov 1978, p. 121). This process has been impressively described for the Arabian Peninsula, for western Afghanistan, or for the western Sahara at the start of the twentieth century, where oasis farmers aspired to the 'higher' class of nomadic herders. With some exceptions, however, the trend since the late nineteenth century has been generally towards modern agriculture and sedentary life (Scholz 1995).

5. *Conclusions*

This fragmentary outline of the 'origin' and history of nomadism as a socioecological mode of culture may be summed up in the following four conclusions:

(a) To a large extent, authors agree that stockkeeping (domestication of herd animals by farmers) forms the basis/prerequisite for nomadism.

(b) Historical sources 'report' in great detail, nomadic invasions of settled lands and nomadic switches to sedentism. This is probably why these events have enjoyed particular academic attention, wide and biased regard, and a historical interpretation of considerable influence on society.

(c) Primarily ethno-archeological studies have indicated that the shift from sedentary life to mobility, from agriculture to nomadism, occurred no less frequently. However, historical sources say little about such switches, and so there is some justification for assuming that insufficient importance has been attached to the historical and social relevance of these processes.

(d) Nomadism as a socio-ecological mode of culture (*sozioökologische Kulturweise*) and, is a human survival strategy in response to specific ecological and sociopolitical conditions within the Old World dry belt. As such, it represents a regionally immanent, logical 'idea' that can/could emerge recurrently, both locally and throughout the whole region.

The switch between sedentism and mobility, cultivator and nomad, occurred in both short- and long-term sequences, affecting both small groups and entire peoples. Its impact was local, but could also involve entire areas and regions.

In this context, it is no longer really necessary to consider Pohlhausen's (1954) hypothesis of a monocentric origin of pastoral nomadism and a multicentric origin of its various regional forms. The above archeological and historical 'facts,' and the conclusions derived from them, suggest that it is more logical to assume that the 'idea' of a nomadic mode of culture was latent everywhere in the natural ecological and sociopolitical setting/structures of the Old World dry belt. It only required a trigger, a change in prevailing conditions, for it to revive as a survival strategy or, indeed, to disappear.

See also: Gypsies and Travelers; Middle East and North Africa: Sociocultural Aspects; Migrations, Colonizations, and Diasporas in Archaeology; Pastoralism in Anthropology

Bibliography

Bobek H 1959 Die Hauptstufen der Gesellschafts- und Wirtschaftsentfaltung in geographischer Sicht. *Die Erde* **90**(3): 259–98

Braudel F 1990 *Das Mittelmeer und die mediterrane Welt in der Epoche Philips II*. Bd. I. Suhrkamp Verlag, Frankfurt am Main, Germany

Dittmer K 1965 Zur Entstehung des Rinderhirtennomadismus. *Paideuma. Mitteilungen zur Kulturkunde* **11**: 8–23

Francfort H-P (ed.) 1990 *Nomades et sédentaires en Asie centrale*. Centre National de la Recherche Scientifique, Paris

Gilbert A S 1983 *On the Origins of Specialized Nomadic Pastoralism in Western Iran. World Archaeology* **15**: 105–19

Hahn E 1891 Waren die Menschen der Urzeit zwischen der Jägerstufe und der Stufe des Ackerbaus Nomaden? Das Ausland. *Wochenschrift für Erd- und Völkerkunde* **64**(25): 481–87

Jettmar K 1966 Die Entstehung der Reiternomaden. *Saeculum* **17**(1–2): 1–11

Khazanov A M 1978 *Characteristic Features of Nomadic Communities in the Eurasian Steppes*. In: Weissleder W (ed.) *The Nomadic Alternative*. Mouton, The Hague, The Netherlands, pp. 119–26

Khazanov A M 1981 Myths and paradoxes of nomadism. *Archives Européennes de Sociologie* **22**: 121–53

Lattimore O 1940 *Inner Asian Frontiers of China*. Oxford University Press, New York

Lees S, Bates D G 1974 The origins of specialized nomadic populations: A systematic model. *American Antiquity* **39**: 187–93

List F 1961 Das natürliche System der politischen Ökonomie. Berlin Ökonomische Studientexte, 2. Akademie-Verlag, Lizenznummer: 202. 100/27/61

Nissen H J 1980 The mobility between settled and non-settled in early Babylonia: Theory and evidence. *Colloques International du CNRS* **580**: 285–90

Planhol X de 1975 *Kulturgeographische Grundlagen der islamischen Geschichte*. Artemis, München, Germany
Pohlhausen H 1954 *Das Wanderhirtentum und seine Vorstufen. Eine ethnographisch-geographische Studie zur Entwicklung der Eingeborenenwirtschaft*. Braunschweig.
Scholz F 1995 *Nomadismus—Theorie und Wandel einer sozioökologischen Kulturweise*. Erdkundliches. Wissen, Stuttgart, 118. Steiner
Scholz F 1997 Gedanken zu einer These über die Entstehung des Nomadismus als sozioökologische Kulturweise. *Petermanns Geographische Mitteilungen* **141**(3): 187–96
Vajda L 1968 *Untersuchungen zur Geschichte der Hirtenkulturen*. Wiesbaden. *(Study about the history of the pastoral culture)*, Bd. 31, O. Harrassowitz
Wissmann H von 1961 *Bauer, Nomade und Stadt im islamischen Orient*. Paret R (ed.) Die Welt des Islam und die Gegenwart. Kohlhammer, Stuttgart, Germany, pp. 22–63

F. Scholz

Nonequivalent Group Designs

In nonequivalent group designs, comparisons are drawn between two or more groups of study participants with each group receiving a different treatment. The purpose, as in all quasi-experiments, is to assess the relative effectiveness of the different treatments. What distinguishes a nonequivalent group design from other designs is the manner in which the study participants are assigned to the different treatments. In a randomized experiment (see *Random Assignment: Implementation in Complex Field Settings*), participants are assigned to treatments atrandom. In a regression-discontinuity design (see *Regression Discontinuity Design*), participants are assigned to treatments based on their value on an ordered quantitative variable. In a nonequivalent group design, participants are assigned to treatments in a nonrandom, nonquantitatively ordered manner. For example, a nonequivalent group design arises when participants self-select themselves into the different treatments, when an administrator assigns participants to different treatments, or when different treatments are implemented in different sites. In analyzing data from nonequivalent group designs, researchers must be concerned with biases that can arise because of the nonrandom assignment. The discussion describes the most widely used statistical analyzes and design features for addressing these biases.

1. *Threats to Internal Validity*

Because the participants are not assigned to treatments at random, the treatment groups in a nonequivalent group design will differ, to one degree or another, even before the treatments are administered. Such differences among the treatment groups are called selection differences. In the nonequivalent group design, selection differences are a threat to internal validity (see *Internal Validity*) because they can bias the estimate of the treatment effect, that is, selection differences can cause post-treatment differences between the treatment groups, even if there are no treatment effects.

In addition to selection differences, other threats to internal validity can also arise in nonequivalent group designs. For example, Cook and Campbell (1979) list several additional threats to internal validity, including compensatory equalization of treatments, compensatory rivalry, resentful demoralization, diffusion or imitation of the treatment, and local history. Threats to validity that involve recognition that different participants receive different treatments, such as resentful demoralization, are likely to be more severe in randomized experiments than in nonequivalent group designs (Fetterman 1982, Lam et al. 1994). On the other hand, threats such as local history will often be more severe in nonequivalent group designs than in randomized experiments. Although a variety of threats to internal validity can arise, the remainder of the present discussion will be concerned only with taking account of selection differences because it is that threat that most distinguishes nonequivalent group comparisons from other types of comparisons.

2. *The Prototypic Nonequivalent Group Design*

In the prototypic nonequivalent group design, only two treatments are compared: an experimental and a comparison treatment. In addition, the participants in each of these two treatment conditions are assessed at exactly two points in time, once before and once after the treatments are implemented. Differences between the two treatment groups on the post-treatment measure are then used to estimate the effects of the treatments and the pretreatment measure is used to take account of the effects of selection differences. If there were no pretreatment measure, the results would usually be uninterpretable because there would be no way to take account of the effects of selection differences.

Various statistical procedures have been proposed for taking account of the effects of selection differences, with the assistance of the pretreatment scores. The four most commonly used methods are described below. In general, all four of these procedures are most credible when the pretreatment measure is operationally identical to the post-treatment measure.

Each of the four analysis strategies imposes different assumptions about the nature of the selection differences. If the assumptions of a given procedure are correct, the resulting estimates of the treatment effects will be unbiased by selection differences. However, if the assumptions are incorrect the estimates of treatment effects are likely to remain biased, perhaps

severely so. Therefore, researchers must be diligent in choosing an analysis strategy whose underlying assumptions fit the circumstances at hand. Unfortunately, researchers will seldom, if ever, have enough information to know which set of assumptions is correct. As a consequence, researchers are advised to conduct multiple analyzes that impose a range of plausible assumptions about the nature of selection differences. Even then, researchers will typically need to be cautious in interpreting the results from prototypic nonequivalent group designs.

2.1 Change-score Analysis

In change-score analysis, the estimate of the treatment effect is the mean difference between the experimental and comparison groups in the amount of change that takes place from the pretreatment measure to the post-treatment measure. For this change from pretreatment to post-treatment to be meaningful, the pretreatment and post-treatment measures must be operationally identical. A change-score analysis can be implemented by calculating a pretreatment to post-treatment difference (i.e., a change score) for each individual and then performing a test of the mean difference between the treatment groups on the change scores. The same result can also be obtained using the pretreatment and post-treatment scores in a repeated-measures analysis of variance.

The underlying assumption in the change-score analysis is that the mean change from pretreatment to post-treatment will be the same in the two treatment groups, in the absence of a treatment effect. This assumption may or may not be correct. Sometimes treatment groups change at different rates so the gap between them either increases or decreases over time, even in the absence of a treatment effect. Such an outcome is called a selection-by-maturation interaction. For example, gaps in income often increase over time because the rich tend to get relatively richer. Patterns such as this can be accommodated in some cases by transforming the pretreatment and post-treatment measures to have equal variances (Judd and Kenny 1981). In other circumstances, treatment groups change so that the gap between them decreases over time. For example, the gap would tend to decrease if individuals in one or both groups were selected for their extreme scores on the pretreatment measure and subsequent performances regressed back toward the mean. This is called a regression artifact (Campbell and Kenny 1999).

Whether alternative explanations can account for the observed results sometimes depends on the pattern of the outcomes (Cook and Campbell 1979). For example, when the experimental group starts out below the comparison group but ends up higher, the pattern of outcomes is called a crossover interaction. Such a pattern seldom can be plausibly explained as due to either a selection-by-maturation interaction or regression toward the mean. However, patterns that can be interpreted credibly are not always easy to produce in practice.

2.2 Matching or Blocking

In matching, each participant from the comparison group is paired with a participant from the experimental group so that they have the same or similar pretreatment scores. Participants who cannot be matched are discarded from the analysis. The effect of the treatment is then estimated by comparing the post-treatment scores in each matched pair (e.g., via a matched-pair t test). In blocking, the participants are matched in groups (or blocks) based on their pretreatment scores, and then comparisons are drawn between the post-treatment scores within each of the blocks (e.g., via a 'randomized' blocks analysis of variance).

The degree to which matching or blocking removes bias due to selection differences depends on several factors. First, bias reduction depends on how similar the participants are on the pretreatment measures within each pair or block (Cochran and Rubin 1973). As the number of pretreatment measures increases, usually it is harder to form pairs or blocks in which the participants are similar on all the measures. This problem can be addressed by matching or blocking participants on the basis of propensity scores, which are single-variable aggregates of multiple pretreatment measures (Rosenbaum 1995).

Second, bias will remain to the degree the pretreatment scores contain measurement error. Under some assumptions, measurement error can be taken into account by matching or blocking participants using pretreatment scores that are corrected for their degree of unreliability.

Third, bias reduction depends on the degree to which the pretreatment measures tap all the factors that influence the post-treatment scores and on which the treatment groups differ. For example, to the extent (a) the participants within each pair or block are not matched on, say, motivation, (b) the treatment groups differ in motivation, and (c) motivation influences the post-treatment scores, then matching or blocking will fail to control for the effects of all selection differences. Unfortunately, there is seldom any way to be sure that all the relevant selection differences have been assessed and included in the analysis. Often the best, though still fallible, way to assure that most of the relevant factors are taken into account is to use pretreatment measures that are operationally identical to the post-treatment measures.

2.3 Analysis of Covariance

Analysis of covariance (ANCOVA) is analogous to matching and blocking. The difference is that

ANCOVA matches participants mathematically rather than by physically forming pairs or blocks of similar participants (Reichardt 1979). By matching mathematically, ANCOVA avoids the problem of inexact matches that can arise when using pairs or blocks, but does so at the expense of having to impose assumptions about the shape of the regression surface between the pretreatment and post-treatment scores.

An analysis of covariance is accomplished by regressing the post-treatment scores on to both pretreatment measures and a dummy variable that indicates membership in the different treatment groups. The estimate of the treatment effect is the regression coefficient for the group-membership dummy variable. Although not as problematic for ANCOVA as for matching or blocking, having a large number of pretreatment measures can make the analysis difficult to implement, and this difficulty might again be reduced by using propensity scores.

Like matching or blocking, ANCOVA is susceptible to biases due to measurement error in the pretreatment scores. Analytic adjustments can be made in ANCOVA to address unreliability in the pretreatment scores (Sorbom 1978). ANCOVA is also just as susceptible as matching or blocking to biases due to omitting pretreatment measures that influence the post-treatment scores and on which the treatment groups differ. Unless the pretreatment measures capture all of the selection differences that influence outcomes, the results of ANCOVA, like matching or blocking, are likely to remain biased.

2.4 Selection Modeling

In selection modeling (see *Selection Bias, Statistics of*), the post-treatment scores are regressed on to both pretreatment measures and a 'selection' variable that models the assignment of participants into the treatment groups. The estimate of the treatment effect is the regression coefficient for the selection variable in this regression equation (which will be called the 'outcome' regression equation).

The selection variable that appears in the outcome regression equation can be constructed in a variety of ways depending on the circumstances (Barnow et al. 1980). One way to construct the selection variable is to set it equal to the predicted scores from what will be called a 'selection' regression equation wherein a group-membership dummy variable is regressed linearly on to both the pretreatment measures that were used in the outcome regression equation and additional pretreatment measures that were not included in the outcome regression equation. The underlying assumption is that the unique pretreatment measures that are added to the selection regression equation are related to group assignment but do not themselves influence the post-treatment measures.

An alternative way to create the selection variable is to set it equal to the predicted scores when a group-membership dummy variable is regressed onto the pretreatment measures using probit or logistic regression. Using a probit or logistic, rather than a linear, regression produces a selection variable that is a nonlinear function of the pretreatment measures. The underlying assumption in this analysis is that the relationship between the pretreatment and post-treatment scores would be linear, in the absence of a treatment effect.

3. More Elaborate Nonequivalent Group Designs

A variety of design features are described below that can be added to the prototypic nonequivalent group design to improve the interpretability of the results (Shadish and Cook 1999). The logic is to create a design where the pattern of results that would arise from a treatment effect is different from the pattern of results that would arise from selection differences (Reichardt 2000) (see *Person-centered Research*).

3.1 Multiple Outcome Measures

The credibility of inferences from nonequivalent group comparisons can sometimes be increased by adding another outcome measure, often called a nonequivalent dependent variable (Cook and Campbell 1979). Causal inference will be enhanced if selection differences have the same effects on both the original and additional outcome measures, but the effect of the treatment differs on the two measures (or vice versa). For example, in a study of a program to aid the homeless (Braucht et al. 1995), the primary intent of the treatment was to decrease substance abuse and only secondarily to improve other outcomes such as family relationships. In contrast, selection differences, such as differences in motivation to change one's lifestyle, should have had just as large an effect on family relationships as on substance abuse. That a large outcome difference appeared on measures of substance abuse but not on measures of family relationships increased the researchers' confidence that the positive results on substance abuse were due to the treatment, rather than to selection differences.

3.2 Multiple Comparison Groups

The credibility of inferences from nonequivalent group designs sometimes can be increased by using two comparison groups in the design rather than just one. For example, a design with two comparison groups allows the experimental group to be compared to each separately. This could strengthen confidence in a causal inference if the treatment were known, say, to have an effect in the same direction in both comparisons while the effects of selection differences were known to operate in opposite directions. In such a

case, similar results in the two comparisons suggests the treatment, rather than selection differences, is responsible for the outcome.

Alternatively, additional comparison groups might be included in the design so that different comparison groups could be used to satisfy the assumptions of different statistical analyzes or multiple comparison groups could be used to assess the plausibility of the assumptions of a given statistical analysis. For example, a change-score analysis could benefit greatly if, relative to the experimental group, one comparison group is initially inferior and the other initially superior on the pretreatment assessment. The most interpretable pattern of results would be where the mean changes from pretreatment to post-treatment were the same in the two comparison groups but different in the experimental group. Such an outcome would rule out both a selection-by-maturation interaction and regression toward the mean more credibly than would be possible with only a single comparison group.

3.3 Multiple Time Points

It can be useful to collect both pre- and post-treatment measures at additional time points. Pretreatment measures at additional time points can be used to improve the prediction of growth trajectories in the absence of the treatment. For example, suppose pretreatment measures are collected at both times 1 and 2, different treatment protocols are introduced at time 3, and post-treatment measures are collected at time 4. Then growth trajectories estimated using the time 1 and 2 measures can help distinguish between treatment effects and selection-by-maturation interactions. With even more pretreatment measures collected over time, researchers could also assess the plausibility of regression toward the mean. In either case, growth trajectories could be estimated using either mean level analyzes such as change-score models or individual level analyzes such as latent growth curve models (McArdle 1988, Muthén and Curran 1997) (see *Growth Curve Analysis*) or hierarchical linear models (Bryk and Raudenbush 1992).

Alternatively, pretreatment measures at times 1 and 2 could be analyzed as if they were a pretreatment and a hypothetical post-treatment measure. Because the different treatments were not introduced until later, such 'dry-run' analyses should produce null results except to the extent they are biased by selection differences or other threats to validity. In this way, the results of such analyzes could be used to choose or modify statistical procedures so they are less likely to be biased by selection differences when used to analyze the actual pretreatment and post-treatment data (e.g., Wortman et al. 1978).

In a similar vein, post-treatment observations at additional time points sometimes can be used to distinguish between treatment effects and selection differences. For example, imagine that the effect of the treatment would increase over time while the effect of selection differences would remain constant or decrease over time. In such a case, a pattern of observed differences that increase over time would contribute to one's confidence that these differences are due to the treatment. For example, the conclusion that smoking causes cancer was strengthened when it was found that the rate of lung cancer in women smokers rose 20–30 years after smoking became prevalent in women, while the lung cancer rate in nonsmoking women remained relatively constant. The alternative explanation that genetic differences between smokers and nonsmokers was responsible for the observed association between cancer and smoking (because genes might produce both a liking for cigarettes and greater susceptibility to lung cancer) could not explain such an outcome, because genetic differences would be expected to be relatively stable over that 'short' a time period.

3.4 Variations in Amounts or Types of Treatments

Adding observations at additional time points or adding multiple comparison groups provides the opportunity to introduce repetitions of the same treatment, different amounts of the same treatment, or different types of treatments. For example, suppose the treatment could be introduced, removed, and then reintroduced to the experimental group and that observations could be collected at all the time points before and after these interventions for both the experimental and comparison groups. Further, suppose the effect of the treatment is short lived (so the treatment's effects are not present when the treatment is removed) and that the effects of selection differences are constant over time. Then a pattern of outcome differences that varies over time in accord with the varying interventions could be strong evidence in favor a treatment effect.

Another possibility is the switching replication design (Cook and Campbell 1979) where the treatment is administered to the experimental group at one point in time and then administered to the comparison group at a later point in time, with observations being collected on each group initially and after each of the treatment administrations. The premise of this design is that the difference between the groups in the effect of the treatment varies over the time periods because the administration of the treatment varies, but the effect of selection differences remains the same because the composition of the groups remains the same.

Another potential source of complex patterns with which to distinguish treatment from selection effects are designs in which groups of participants receive varying amounts of the treatment. For example, one set of the participants in the experimental group might receive the standard dose of a treatment while other

sets of participants receive either half or twice the standard dose. The underlying premise of such designs is that different treatment doses produce systematically different outcomes (sometimes called the dose-response curve for the treatment), while the effects of selection differences would follow a different pattern.

3.5 The Benefits of Disaggregation

The preceding discussion describes potential benefits of additional outcome measures, comparison groups, time points, and treatments. Incorporating these design features often requires collecting additional data, but sometimes can be accomplished by disaggregating the available data. For example, sometimes it is possible to add a nonequivalent dependent variable by disaggregating the data from an outcome variable that consists of a battery of measures. Or a comparison group sometimes can be added by dividing what was initially a single comparison group into two groups. Or comparisons between different dosage levels of the treatment might be obtained by partitioning the experimental group according to the amount of treatment received. Even additional time points sometimes can be added to a design by partitioning data that initially were aggregated, say, at yearly time intervals, into data aggregated only at monthly intervals.

4. Minimizing Selection Differences

Researchers should try to implement designs and analyses that are most capable of taking account of selection differences in the specific situations at hand. Applying a complex statistical procedure to data from a prototypic nonequivalent group design may be appealing in theory, but a much simpler statistical analysis combined with a more elaborate design will often be more credible in practice.

However, regardless of the simplicity or elaborateness of the research design and statistical analysis, the credibility of a nonequivalent group comparison tends to increase when selection differences are made small and the treatment effects are made large. Random assignment tends to make selection difference small. In nonequivalent group designs, selection differences can often be minimized by using an assignment procedure that is as close to random as possible. For example, Langer and Rodin (1976) assessed the effects of a treatment by drawing comparisons across the clients on different floors of a nursing home. In this case, the assignment of clients to treatment conditions appeared to be quasi-random in that there were no obvious mechanisms by which different types of clients would be assigned systematically to different floors.

Often selection differences are smaller when the experimental and comparison groups are composed of individuals recruited from the same organization or locale (called internal controls) than from different organizations or locales (called external controls). A special case of internal controls arises when the experimental and comparison groups are made up of earlier and later cohorts (Cook and Campbell 1979). For example, the effect of a treatment that is given to college sophomores in a given year could be assessed by drawing a comparison with the performance of sophomores at that college from the prior year. Such a design is worth considering when the cohort of sophomores from the prior year is more similar to the treatment group than would be a contemporaneous comparison group of freshman or juniors from the same college or of sophomores from a different college.

5. Summary

Primarily because of biases due to selection differences and because of the difficulty in taking these biases into account, nonequivalent group designs generally produce less credible results than randomized experiments. Nonetheless, nonequivalent group designs are used widely because they can be far easier to implement than randomized experiments. Researchers should try to take account of the biasing effects of selection differences by the judicious combination of analysis strategies and design features. It is also advisable to minimize selection differences *a priori* so as to minimize one's reliance on either statistical procedures or design elaborations.

See also: Experimental Design: Randomization and Social Experiments; Field Observational Research in Anthropology and Sociology; Hierarchical Models: Random and Fixed Effects; Internal Validity; Psychological Treatments: Randomized Controlled Clinical Trials; Quasi-Experimental Designs; Random Assignment: Implementation in Complex Field Settings; Regression Discontinuity Design; Sample Surveys: Survey Design Issues and Strategies; Selection Bias, Statistics of

Bibliography

Barnow B S, Cain G C, Goldberger A S 1980 Issues in the analysis of selectivity bias. In: Stromsdorfer E W, Farkas G (eds.) *Evaluation Studies Review Annual*. Sage, Newbury Park, CA, Vol. 5

Braucht G N, Reichardt C S, Geissler L J, Bormann C A, Kwiatkowski C F, Kirby Jr. M W 1995 Effective services for homeless substance abusers. *Journal of Addictive Diseases* **14**: 87–109

Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park, CA

Campbell D T, Kenny D A 1999 *A Primer on Regression Artifacts*. Guilford Press, New York

Cochran W G, Rubin D B 1973 Controlling bias in observational studies. *Sankhya* (*Series A*) **35**: 417–46

Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand McNally, Chicago
Fetterman D M 1982 Ibsen's baths: Reactivity and insensitivity. *Educational Evaluation and Policy Analysis* **4**: 261–79
Judd C M, Kenny D A 1981 *Estimating the Effects of Social Interventions*. Cambridge University Press, New York
Lam J A, Hartwell S W, Jekel J F 1994 'I prayed real hard, so I know I'll get in': Living with randomization. In: Conrad K J (ed.) *Critically Evaluating the Role of Experiments*. New Directions for Program Evaluation, No. 63. Jossey-Bass, San Francisco
Langer E J, Rodin J 1976 The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting. *Journal of Personality and Social Psychology* **34**: 191–98
McArdle J J 1988 Dynamic but structural equation modeling of repeated measures data. In: Nesselroade J R, Cattell R B (eds.) *The Handbook of Multivariate Experimental Psychology*, 2nd edn. Plenum, New York
Muthén B O, Curran P J 1997 General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods* **2**: 371–402
Reichardt C S 1979 The statistical analysis of data from nonequivalent group designs. In: Cook T D, Campbell D T (eds.) *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand McNally, Chicago
Reichardt C S 2000 A typology of strategies for ruling out threats to validity. In: Bickman L (ed.) *Research Design: Donald Campbell's Legacy*. Sage, Thousand Oaks, CA, Vol. 2
Rosenbaum P R 1995 *Observational Studies*. Springer-Verlag, New York
Shadish W R, Cook T D 1999 Comment–Design rules: More steps toward a complete theory of quasi-experimentation. *Statistical Science* **14**: 294–300
Sorbom D 1978 An alternative methodology for analysis of covariance. *Psychometrika* **43**: 381–96
Wortman P M, Reichardt C S, St. Pierre R G 1978 The first year of the educational voucher demonstration: A secondary analysis of student achievement test scores. *Evaluation Quarterly* **2**: 193–214

C. S. Reichardt and M. M. Mark

Nonparametric Statistics: Advanced Computational Methods

Standard inference procedures are based solely on data whose probability distributions have been specified in terms of a few parameters. Bayesian methods incorporate existing knowledge (historical information, expert knowledge, previous data, etc.) about the parameter into the inference process. This is done by assigning a probability distribution, called the prior distribution, to the parameter governing the distribution of the data. This prior distribution represents the information available to us concerning the parameter before gathering data. After some data have been observed, we can compute the conditional distribution of the parameter given the observed data; this is called the posterior distribution and it represents our knowledge of the parameter at this stage. If one wants to estimate a function of the parameter, then one can just find the expectation of that function under this posterior distribution; it will be the best estimate under squared error loss. Similar methods of estimation under other loss functions are possible and they all depend solely on the posterior distribution. Thus Bayesian methods concentrate first on finding the posterior distribution or on simulating from the posterior distribution.

In a nonparametric problem, the common unknown distribution F of the data is the 'parameter.' A Dirichlet measure is a distribution for F and will serve as a prior distribution in nonparametric problems. The real problems here are how to define such nonparametric prior distributions for this and other wider classes of problems, and how to carry out the necessary computations.

In Sect. 2 below we describe various ways to define the Dirichlet measure, which is a natural choice for a prior distribution in nonparametric problems. Closed form solutions for inference in Bayesian nonparametric problems are given in several cases together with some references for further study.

Since closed form solutions will in general not be available in interesting situations, advanced computational approaches are necessary to carry out inference in nonparametric Bayesian problems. Such computational methods have been used in regular nonparametric problems as well. In Sect. 3.1 and 3.2 we give examples of computational intensive methods in two standard nonparametric problems, namely testing for independence and testing for concordance of two variables. In Sect. 3.3, the estimation of a distribution function when observations have been censored by some censoring mechanism is treated as a Bayesian nonparametric problem. We describe the advanced computational approach, known as Gibbs sampling, which can be used for this problem. We present the exact Bayes estimate and the Bayes estimate obtained by Gibbs sampling to demonstrate that power of the computational method. For further comparison purposes, the estimate previously obtained by frequentist methods is also presented.

1. *Bayesian Methods and Dirichlet Measures*

Bayesian methods are quite general and apply to both parametric and nonparametric inference problems. It is easier in the parametric case since the unknown parameter usually varies in a finite dimensional space and we are familiar with probability distributions in finite dimensional spaces. In nonparametric problems the unknown 'parameter' can be as complex as the underlying distribution of the observed random variable. This is not only an infinite dimensional quantity,

it also satisfies some inequality restrictions; for instance a distribution function $F(x)$ satisfies $0 \leqslant F(x) \leqslant 1$ for all $x \in R_1$, where R_1 is the real line, and is also a nondecreasing right continuous function with limits 0 and 1 as x tends to $-\infty$ and ∞, respectively. Thus the first problem faced by Bayesians in nonparametric inference problems is an appropriate definition of a probability distribution for $F(\cdot)$. Any such prior distribution will be useful only if it is tractable and yields closed form or computable solutions to the nonparametric inference problems on hand. These problems were daunting at first, but were elegantly solved by Ferguson (1973) who introduced Dirichlet measures as prior distributions.

We will look closely at probability measures P (which are equivalent to distribution functions) on R_1, the real line. The same can be done for probability measures on more general spaces. For any partition of $\mathscr{A} = (A_1, \ldots, A_k)$, consisting of disjoint measurable sets whose union is R_1, the vector $P(\mathscr{A}) = (P(A_1), \ldots, P(A_k))$ is a k-vector of nonnegative numbers adding to 1. Notice that P is completely determined if $P(\mathscr{A})$ is known for all or most partitions $\mathscr{A}$ and satisfies obvious additivity and consistency conditions. It can be shown that a consistent family of probability measures for $P(\mathscr{A})$ as $\mathscr{A}$ varies among all finite partitions will define a unique probability measure for P. This is the approach of Ferguson (1973).

To describe Ferguson's Dirichlet measure, we will first look at finite dimensional Dirichlet distributions. These are distributions in R_k concentrated on the subset

$$\mathscr{P}_k = \{\mathbf{p} = (p_1, \ldots, p_k);\ p_1 \geqslant 0, \ldots, p_k \geqslant 0, \Sigma_i p_i = 1\}$$

A random vector $(Y_1, \ldots, Y_k) \in \mathscr{P}_k$ is said to have a finite dimensional Dirichlet measure with parameters $(\beta_1, \ldots, \beta_k)$ with $\Sigma_j \beta_j > 0$, if $Y_i = Z_i/(\Sigma_j\ Z_j)$ where $Z_1, \ldots, Z_k$ are independent Gamma random variables with scale parameter 1 and shape parameters $\beta_1, \ldots, \beta_k$, respectively. When $\beta_i > 0, i = 1, \ldots, k$, this distribution is given by the more familiar density function of $(Y_1, \ldots, Y_{k-1})$:

$$\frac{\Gamma\left(\sum_i \beta_j\right)}{\Pi_j \Gamma(\beta_j)} \prod_j y_j^{\beta_j - 1}$$

where $y_k = 1 - \sum_1^{k-1} y_j$. This finite dimensional Dirichlet measure will be denoted in symbols by $\mathscr{D}(\beta_1, \ldots, \beta_k)$.

Let $\mathscr{P}$ be the space of all probability measures on R_1. Ferguson defined a probability measure called a Dirichlet measure for a random probability measure P taking values in $\mathscr{P}$. (To be more rigorous, we should also define the σ-field in $\mathscr{P}$, but we shall avoid going into such details.) Ferguson defined the Dirichlet measure with parameters α and μ, where $\alpha > 0$ and μ is a probability measure on R_1, as follows: for every finite partition $\mathscr{A}$, $P(\mathscr{A})$ has the finite dimensional Dirichlet distribution with parameters $\mathscr{D}(\alpha\mu(A_1), \ldots, \alpha\mu(A_k))$. In particular the distribution of $P(A)$ is a Beta distribution with parameters $(\alpha\mu(A), \alpha\mu(A^c))$. It takes some more mathematical machinery to show that such a Dirichlet measure exists, but again we will not go into this here. The Dirichlet measure above will be denoted in symbols by $\mathscr{D}(\alpha, \mu)$.

Many interesting properties of Dirichlet measures are known. These are easy to derive from the representation of Dirichlet measures given by Sethuraman (1994) which is described in the next paragraph. The Dirichlet measure $\mathscr{D}(\alpha, \mu)$ gives probability one to the class of all discrete distributions in R_1 even if μ is a continuous distribution. The expected value and variance of $P(A)$ are $\mu(A)$ and $\mu(A)\mu(A^c)/(\alpha+1)$. The expected value and variance of $\int f dP$ is $\int f d\mu$ and $(\int f^2 d\mu - (\int f d\mu)^2)/(\alpha+1)$.

An easy constructive definition of a Dirichlet measure was announced in Sethuraman and Tiwari (1982) and proved in more detail in Sethuraman (1994). Consider independent random variables $Y_1, Y_2, \ldots, \theta_1, \theta_2, \ldots$ where the Y's are identically distributed (*i. i. d.*) with common distribution μ and the θ's are identically distributed with a common Beta distribution with parameters 1 and α. Define

$$p_1 = \theta_1, p_k = \theta_k \prod_{j=1}^{k-1} (1 - \theta_j) \tag{1}$$

$$P(A) = \sum_k p_k I(Y_k \in A) \tag{2}$$

where for any event B, the indicator function $I(B) = 1$ if B occurs and $= 0$ otherwise. Clearly, P is a random discrete probability measure putting random probability mass p_k at Y_k. Its distribution is the Dirichlet measure $\mathscr{D}(\alpha, \mu)$.

Let $X_1, X_2, \ldots$ be random variables such that, given P, they are i.i.d. with common distribution P. Let F_n be the empirical distribution of $X_1, X_2, \ldots, X_n$, that is, $F_n(x) = \frac{1}{n}\sum_1^n I(X_i \leqslant x)$. The conditional distribution of P, given the sample $X_1, X_2, \ldots, X_n$, was shown by Ferguson (1973) to be the Dirichlet measure with parameters $(\alpha + n, \mu_n)$, where

$$\mu_n = \frac{\alpha}{\alpha+n}\mu + \frac{n}{\alpha+n} F_n. \tag{3}$$

Blackwell (1973) and Blackwell and MacQueen (1973) give alternative ways of defining Dirichlet measures and establishing the above result on the posterior distribution.

We will now briefly describe some popular uses of the Dirichlet prior in nonparametric problems. Suppose that we have data consisting of $X_1, \ldots, X_n$ which, given P, are independent and identically distributed with common probability distribution P. Suppose that we want to estimate P. The frequentist's usual es-

timator is $F_n(X) = \frac{1}{n}\sum_1^n I(X_i \leqslant x)$ the empirical distribution of the data $X_1, \ldots, X_n$.

A Bayesian would place a Dirichlet prior $\mathscr{D}(\alpha, \mu)$ as a prior distribution for P. The measure μ would represent his average guess of the unknown P and α would represent his measure of confidence in this guess, with large values of α denoting more confidence. Then the posterior distribution of P is the Dirichlet measure $\mathscr{D}(\alpha+n, \mu_n)$ where μ_n is given by Eqn. (3). Under squared error loss, the Bayes estimate of any function $g(P)$ is simply its expectation of $g(P)$ under this posterior distribution. Thus the Bayes estimates of P, $P((-\infty, x])$ and $\int f(x)dP(dx)$ are μ_n, $\mu_n((-\infty, x])$ and $\int f(x)\mu_n(dx)$, respectively. Under absolute value of error loss function, the Bayes estimate of the median of P is the median of μ_n. The expected Bayes loss of all these estimates are obtained in Ferguson (1973). To summarize these results in simple terms, the Bayes estimate of P is μ_n which is a convex linear combination of the parameter μ of the prior distribution and the empirical distribution function F_n of the observed data, where the weight given to μ, the prior guess, goes to zero as the sample size of the data increases, a fact that agrees well with intuition. Susarla and Van Ryzin (1976) obtain Bayes estimates of P when the observations are censored by independent censoring variables, which we discuss later. For more examples of Bayes estimates of other functionals see Yamato (1975), Hollander and Korwar (1982), Zalkikar et al. (1986), Wells and Tiwari (1989), and Lu et al. (1994).

The influence of the parameter μ in the prior distribution goes to zero as n gets large and the Bayes estimate is close to F_n. In fact, if one were to formally put the parameter α of the prior distribution to be zero, then the Bayes estimate is F_n; this has led some researchers to refer to α as the prior sample size. There is a difficulty with this notion. The prior distribution is not defined if $\alpha = 0$. However, the prior distribution converges to a proper prior distribution as $\alpha \to 0$ and μ is held fixed. This limiting prior corresponds to a random degenerate distribution, and as such will not be an acceptable prior distribution for Bayesian problems (see Sethuraman and Tiwari 1982). In spite of this, many researchers check to see if they get familiar frequentist estimators when $\alpha \to 0$.

This summarizes the first results concerning applications of Dirichlet measures for Bayesian analysis in one-sample problems. Similar results which have wider applicability have been obtained for two-sample problems, Bayesian density estimation, censored samples, reliability, point processes, etc.

There has been work in finding prior distributions which are more general than Dirichlet measures, like measures that concentrate on absolutely continuous distributions, or distributions with increasing failure rate, and so on. The problem of constructing a random probability distribution requires a clear understanding of the structure of a probability measure. Consider partitioning the whole space into two sets A_0, A_1, and then partitioning each of these as A_{00}, A_{01} and A_{10}, A_{11}, respectively. At the nth stage we will have 2^n sets A_{t_n}, where t_n is a n-vector of zeros and ones. The set A_{t_n} will be partitioned as $A_{t_n,0}$, $A_{t_n,1}$, and we will have 2^{n+1} sets at the $(n+1)$th stage, and so on. Suppose the partitioning sets from all stages, namely $\{A_0, A_1, \ldots, A_{t_n}, \ldots\}$, generate the Borel σ-field $\mathscr{B}$ in R_1. Any probability measure P on R_1 gives rise to the quantities ($u_e = P(A_1)$, $u_0 = P(A_{01} \mid A_0)$, $u_1 = P(A_{11} \mid A_1), \ldots, u_{t_n} = P(A_{t_n 1} \mid A_{t_n}), \ldots$) all of which lie between zero and one and conversely any sequence on numbers ($u_e, u_0, u_1, \ldots, u_{t_n}, \ldots$) in [0, 1] will give rise to a unique probability measure P on R_1. Furthermore, the range of each u_{t_n} is unrestricted. Thus it is easy to assign probability distributions to (u_{t_n}). Any such distribution induces a probability measure on the space of probability measures on R_1 and can serve as a prior distribution in nonparametric problems. If we assume that the u's are independent with u_{t_n} having a Beta distribution with parameters $(\alpha\mu(A_{t_n 1}), \alpha\mu(A_{t_n 0}))$, then Blackwell (1973) showed that the resulting distribution for P is $\mathscr{D}(\alpha, \mu)$. By choosing other distributions for the u's one can obtain prior distributions that concentrate on absolutely continuous distributions, distributions which are symmetric, unimodal, or have increasing failure rate, etc.

Another way to introduce prior distributions for P arises from the fact that, when the distribution of $X_1, X_2, \ldots$ are i.i.d. P given P, the distribution of $X_1, X_2, \ldots$ is exchangeable. In fact, from De Finetti's theorem, if $X_1, X_2, \ldots$ is exchangeable, then there is a function P of $X_1, X_2, \ldots$, which is a random probability measure, and given this P, the $X_1, X_2, \ldots$ become i.i.d. P. The distribution of this P is a prior distribution on the space of all probability measures on R_1. Thus every sequence of exchangeable random variables corresponds to a prior distribution on the space of probability measures. Using this approach, Blackwell and MacQueen (1973) showed that the Pólyaurn scheme produces an exchangeable sequence of random variables leading to Dirichlet measures. Exchangeable distributions are usually derived as i.i.d. random variables with a common P having a prior distributions rather than vice versa, and so this method has not yielded a rich variety of prior distributions. There is, however, a mathematical beauty in this method.

2. *Computational Methods in Nonparametrics: Bootstrap and Gibbs Sampling*

Two methods, namely the bootstrap (see *Resampling Methods of Estimation*) and Gibbs sampling methods (see *Monte Carlo Methods and Bayesian Computation: Overview*), have proved very successful in handling a multitude of inference problems that could not be handled before since the various distributions needed

for inference are generally mathematically intractable. These sampling methods replace the mathematical machinery with computational methods to obtain approximations for large classes of distributions. A lot of ingenuity still has to go into each situation to make these computational methods effective. Much theory has also been developed to prove that these computational methods work when the number of iterations and sample sizes become large, but we will not describe those developments here. We will give a general introduction and describe some examples.

Let $X_1, \ldots, X_n$ be i.i.d. with common distribution P. Let $F_n = \frac{1}{n}\sum_1^n I(X_i \leqslant x)$ be the empirical distribution function which is an estimate of P. Suppose that we wish to estimate the parameter $\theta = g(P)$, where $g(P)$ is some function of P. The population mean $\theta = \int xP(dx)$ is an example of such a parameter. An obvious estimate of this θ would be $g(F_n)$. One would also like to give the precision of this estimate, namely the standard error of this estimate. This can be done using the bootstrap. The bootstrap method says that if we draw a sample $X_1^*, \ldots, X_n^*$ of size n from F_n (i.e., a simple random sample of size n with replacement from the originally observed sample) and compute its empirical distribution function $F_n{}^*$ then $g(F_n{}^*)$ bears the same relation to $g(F_n)$ as $g(F_n{}^*)$ bears to $g(P)$. If we repeat this procedure K times where K is large, then the standard deviation of the K values of $g(F_n{}^*)$ will be an estimate of the precision of the estimate $g(F_n{}^*)$. This, in a nutshell, is the bootstrap method in its simplest form, as introduced by Efron (1979). The asymptotic accuracy of the bootstrap has been established among others by Bickel and Freedman (1981) and Singh (1981). This method can apply to parametric as well as nonparametric problems. We will illustrate the bootstrap in two nonparametric problems below.

2.1 Testing for Concordance

Kendall's τ is a measure of concordance, namely the tendency of two random variables X and Y to increase or decrease together. It possesses the desirable property that it is invariant under monotone increasing transformations of X and Y and so depends only on the rankings and not the exact magnitudes of X and Y. Let (X, Y) have a joint distribution H. Let $I(\cdot)$ stand for the usual indicator function. Define the concordance between a pair of vectors (x_1, x_2) and (y_1, y_2) as

$$C((x_1, x_2), (y_1, y_2)) = I((x_2 - x_1)(y_2 - y_1) > 0) - I((x_2 - x_1)(y_2 - y_1) < 0)$$

which is equal to 1 if $(x_2 - x_1)$ and $(y_2 - y_1)$ are of the same sign, -1 if they are of opposite signs and 0 if either $x_1 = x_2$ or $y_1 = y_2$. The population value of τ is defined by

$$\tau = E[C((X_1, X_2), (Y_1, Y_2))] \tag{4}$$

Table 1
Hunter L values and consumer panel scores for nine lots of canned tuna

Lot	Hunter L value (X)	Panel score (Y)
1	44.4	2.6
2	45.9	3.1
3	41.9	2.5
4	53.3	5.0
5	44.7	3.6
6	44.1	4.0
7	50.7	5.2
8	45.2	2.8
9	60.1	3.8

Source: Rasekh et al. (1970)

Note that if X and Y are independent then $\tau = 0$. Consider a sample $(X_n, Y_n) = ((X_1, Y_1), \ldots, (X_n, Y_n))$ which is i.i.d. with common distribution function $H(x, y)$. A natural sample estimate of τ_n of τ is given by

$$\tau_n = \frac{2}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} C((X_i - X_j), (Y_i - Y_j)) \tag{5}$$

The above definition can be simplified if there are no ties in the data. The statistic τ_n can also be recognized to be a U-statistic and so one can derive its asymptotic distribution, see for instance Hoeffding (1948), under both the null and alternative hypotheses. However, the asymptotic distribution of τ_n depends on the joint distribution $H(x, y)$ and the variance should be estimated to produce confidence intervals for τ. We give below the details of a simple use of the bootstrap to solve this problem and illustrate it with a numerical example taken from Hollander and Wolfe (1999, Sect. 8.1).

Let $H_n(x, y)$ be a joint distribution that places mass $\frac{1}{n}$ at each of the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. This H_n is the empirical distribution function of the bivariate sample. Let $(X_n^*, Y_n^*) = ((X_1^*, Y_1^*), \ldots, X_n^*, Y_n^*))$ be a sample of size n from H_n and let τ_n^* be the τ-statistic computed from this sample. By repeating this K times, one gets K values of τ_n^*. Let a_n, b_n be the $B\alpha$-largest and $B(1-\alpha)$-largest values of these K values of τ_n^* arranged in increasing order. The interval $[a_n, b_n]$ is an approximate $1-2\alpha$ confidence interval for τ. Standard theorems show that the confidence level of this confidence interval converges to $1-2\alpha$ as K, $n \to \infty$.

Rasekh et al. (1970) wish to evaluate the relative importance of factors contributing to tuna quality, in particular the importance of the Hunter L value factor. A small subset of the data is given in Table 1 (see also Hollander and Wolfe 1999).

The value of τ_n from this sample is $\frac{4}{9} = 0.4444$ and this is an estimate of the population value of τ. This shows that there seems to be a positive dependence

between the Hunter L value and tuna quality. A confidence interval for τ can now be obtained by the use of the bootstrap. One thousand samples were generated by the bootstrap sampling scheme from the given data using a program written in the statistical package S-Plus. The 90 percent confidence interval for τ was found to be $[-0.0278, 0.7500]$, the interval containing 90 percent of the 1000 values of τ_n generated by the bootstrap. Since zero falls in this interval, we cannot rule out that the Hunter L value has no influence on tuna quality, if we use a 90 percent confidence interval as a two-sided test of independence with type-I error probability approximately 10 percent.

2.2 Testing Independence

Suppose that we want to test that two random variables X and Y are independent. Consider a sample $(X_n, Y_n) = ((X_1, Y_1, \ldots, (X_n, Y_n))$ which is i.i.d. with common distribution function $H(x, y)$ with marginal distribution functions $F(x)$ and $G(y)$. We wish to test the hypothesis of independence, namely that

$$H(x, y) = F(x)\, G(y) \tag{6}$$

for all x, y. One may restate this problem as asking whether the joint distribution given by taking the marginals to be independent fits the actual distribution very closely; in this sense this problem can be considered as a goodness-of-fit problem. (For other goodness-of-fit problems see *Goodness of Fit: Overview*). A standard test is to reject for large values of the statistic

$$T_n = T_n(X_n, Y_n) = \sup_{x,y} |\, H_n(x, y) - F_n(x) G_n(y) \,| \tag{7}$$

where $H_n(x, y)$, $F_n(x)$, $G_n(y)$ are the joint empirical and marginal empirical distributions. How should we choose the critical value for T_n?

The statistic T_n is distribution-free under the null hypothesis, if the distributions are continuous. The bootstrap implementation, which is valid even when the distributions are not continuous, proceeds as follows. Notice that $F_n(x)$ and $G_n(y)$ are the sample estimates of the marginal distribution functions. Under the null hypothesis, $F_n(x)G_n(y)$ will be an estimate of the joint distribution function. Thus one collects a sample $(X_n^*, Y_n^*) = ((X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*))$ from the distribution $F_n(x)G_n(y)$ and computes the statistic $T_n^* = T_n(X_n^*, Y_n^*)$. This is repeated K times. Let $c_n(\alpha)$ be the $K\alpha$-th largest value of these T_n^*s arranged in increasing order. This $c_n(\alpha)$ is taken to be the critical value at level α and the null hypothesis is rejected if T_n is larger than $c_n(\alpha)$. This implementation of the bootstrap is described and illustrated in two articles by Romano (1988, 1989), where it is shown that this bootstrap test is consistent, if $K, n \to \infty$.

Table 2
Psychological test scores of nonidentical male twins

Pair no	First twin	Second twin
1	277	256
2	169	118
3	157	137
4	139	144
5	108	146
6	213	221
7	232	184
8	229	188
9	114	97
10	232	231
11	161	114
12	149	187
13	128	230

Source : Clark et al. (1961)

Table 2 gives the data considered by Clark et al. (1961) in a study concerned with the relationship of scores on various psychological tests and certain physical characteristics of 13 nonidentical male twins (see Hollander and Wolfe 1999, Sect. 8.1).

The value of the T_n given by Eqn. (7) for the data is 0.1775. The bivariate distribution in which X and Y are independent and have the marginal distributions given by the above data was constructed. One thousand bivariate samples were generated from this distribution and the 5 percent upper cut-off point of the T_n statistic calculated from these 1,000 samples were found to be 0.1893. This shows that the scores of the two twins can be concluded to be independent at the 5 percent level.

2.3 Gibbs Sampling

Gibbs sampling is a computational method to calculate distributions which are intractable by mathematical means. Suppose that we wish to obtain samples from a joint distribution H of a bivariate random variable (X, Y) or just want to evaluate $\int g(x, y)\mathrm{d}H(x, y)$ for some function $g(x, y)$. This may be difficult because the joint distribution H is quite complicated or only indirectly defined. Suppose that the conditional distributions $Q(x \mid y) = P(X \leqslant x \mid Y = y)$ and $R(y \mid x) = P(Y \leqslant y \mid X = x)$ are known sufficiently well enough to generate observations from them. Then we can proceed as follows. Let (X_0, Y_0) be arbitrary. Generate X_n from the distribution $Q(\cdot \mid Y_{n-1})$ and Y_n from the distribution $R(\cdot \mid X_n)$, $n = 1, \ldots, N$. One can check that $(X_0, Y_0), (X_1, Y_1), \ldots,$ is a Markov chain with a transition function for which H is an invariant distribution. If this transition function satisfies some mild con-

Table 3
Exact and computational Bayes estimates of \$F(x)\$ based on Kaplan and Meier Data (\$\α = 8, \mu =exponential with scale parameter \$0.12\$)

x	0.80	1.00	2.70	3.10	5.40	7.00	9.20	12.10
Exact	0.1083	0.1190	0.2071	0.3006	0.4719	0.5256	0.6823	0.7501
Gibbs	0.1083	0.1190	0.2084	0.3011	0.4706	0.5261	0.6802	0.7476
KM	0.1250	0.1250	0.1250	0.3000	0.4750	0.4750	0.7375	0.7375

Source : Doss (1994)

ditions, then this Markov chain converges in one of several senses to H, and so, for large N, one can take

$$\frac{1}{N}\sum I((X_i,\ Y_i)\in A) \text{ and } \frac{1}{N}\sum g(X_i,\ Y_i)$$

to be approximations to $H(A)$ and $\int g(x, y)\, \mathrm{d}H(x, y)$. If the convergence is stronger, then one can take (X_N, Y_N), as approximately having distribution H. This is the essence of the Gibbs sampler of which there are many other variations. See Gelfand and Smith (1990), Tierney (1994) (for more details see *Monte Carlo Methods and Bayesian Computation: Overview*).

We will now illustrate the Gibbs sampler in a nonparametric censoring problem. Suppose that, given F and G, $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ are independent with $X_1, \ldots, X_n$ i.i.d. with common distribution F and with $Y_1, \ldots, Y_n$ i.i.d. with common distribution G. We will assume that the distribution functions F and G are continuous functions. Suppose that we observe only $Z_i = \min(X_i, Y_i)$, $\delta_i = I(X_i \leqslant Y_i)$, $i = 1, \ldots, n$; Z_i is the censored value of X_i if $\delta_i = 0$ and the actual value of X_i if $\delta_i = 1$.That is, sometimes we observe an X_i and at other times we just know that $X_i \in (Y_i, \infty)$. The problem is to estimate $F(x)$ or the survival function $\bar{F}(x) = 1 - F(x)$. A Bayesian approach would be to put a prior on (F, G) and look at the conditional distribution of F given the data. Assume that under the prior distribution, F and G are independent and that F has a Dirichlet measure $\mathscr{D}(\alpha\mu)$. The distribution of G will be irrelevant and it can be shown in this Bayesian problem that we can consider the values of $Y_1, \ldots, Y_n$ as constants (see Lo 1993 for instance). We can rearrange the sample so that $Z_1 = X_1, \ldots, Z_r = X_r$, $Z_{r+1} = Y_{r+1}, \ldots, Z_n = Y_n$ where r is the number of uncensored X's. The number of censored observations is $m = n - r$. This means the sample consists of $X_1, \ldots, X_r$ and the information $X_i \in A_i$, $i = r+1, \ldots, n$, where $A_i = (Y_i, \infty)$, $i = r+1, \ldots, n$. The Bayesian posterior distribution is the conditional distribution of F given the data. We will show now how Gibbs sampling can help us to simulate from this distribution after we describe earlier work on a closed form expression for the Bayes estimate of F.

The Bayes estimate of F is the expected value of F under this posterior distribution. This can be obtained without evaluating the complete posterior distribution as shown in Susarla and Van Ryzin (1976). Let $N_n(A) = \sum_1^n I(Z_i \in A)$ be the number of observations (censored or not) that fall in A. Let $Z_1^* < Z_2^* < \ldots < Z_m^*$ be the distinct values among the censored observations $Z_{r+1}, \ldots, Z_n$. Let λ_k be the number of censored observations equal to Z_k^*, $k = 1, \ldots, m$. Then the Bayes estimate of $\bar{F}(x)$ as obtained by Susarla and Van Ryzin (1976) is given by

$$\hat{\bar{F}}_n(x) = \frac{\alpha\mu(x, \infty) + N_n(x, \infty)}{\alpha\mu(0, \infty) + n} \prod_{j=1}^{m} \left[\frac{\alpha\mu([Z_j^*, \infty)) + N_n([Z_i^*, \infty))}{\alpha\mu([Z_j^*, \infty)) + N_n([Z_j^*, \infty)) - \lambda_j} \right] \quad (8)$$

for $Z_l^* \leqslant x < Z_{l+1}^*$, $l = 0, 1, \ldots, m$ where we set $Z_0^* = 0$, $Z_{m+1}^* = \infty$. This estimate converges to the standard Kaplan–Meier frequentist estimate (Kaplan and Meier 1958) as $\alpha \to 0$.

A full Bayesian analysis will require the posterior distribution of F. This can be accomplished in this problem through Gibbs sampling. This will also yield a way to obtain the posterior estimate of the survival function, by computational methods, which then can be compared with the exact estimate of Susarla and Van Ryzin given in (8). The conditional distribution of F given $X_1, \ldots, X_r$ (the observed X's) is $\mathscr{D}(\alpha + r, \mu_r)$ where μ_r is as defined in (3) with $n = r$. Let $F, X_{r+1}, \ldots, X_n$ be such that F has distribution $\mathscr{D}(\alpha + r, \mu_r)$ and given F, let $X_{r+1}, \ldots, X_n$ be i.i.d. F. The posterior distribution is the conditional distribution of F given $X_i \in A_i$, $i = r+1, \ldots, n$.

It is clear that we do not know the precise values of $X_{r+1}, \ldots, X_n$, but we know only that $X_i \in A_i$, $i = r+1, \ldots, n$. The standard theory of Dirichlet distributions shows that the conditional distribution $Q(F \mid X_r, \ldots, X_n, \text{data})$ of F given $X_r, \ldots, X_n$ and the data is equal to $\mathscr{D}(\alpha + n, \mu_n)$ where μ_n is given by Eqn. (3). We also know that the conditional distribution $R(X_{r+1}, \ldots, X_n \mid F, \text{data})$ of $X_{r+1}, \ldots, X_n$ given F and the data, is just that of independent random variables with distributions $F_{A_{r+1}}, \ldots, F_{A_n}$ where for any set A, F_A represents that distribution F restricted to the set A. Thus, in the joint distribution of F and $(X_{r+1}, \ldots, X_n)$ given the data, we know the two conditional distributions. This provides the necessary parts to carry out Gibbs sampling. After simulating an observation $(F, X_{r+1}, \ldots, X_n)$ from the joint distribution of F and

$(X_{r+1}, \ldots, X_n)$ given the data, we can retain just the F which then can be considered as an observation from the posterior distribution.

There is a special difficulty in this example of Gibbs sampling. How does one generate a distribution F from a Dirichlet measure? This F is not a simple as a real valued random variable. The constructive definition of a Dirichlet measure given in Sethuraman (1994) will be useful here. Suppose that at stage 0 of Gibbs sampling, we start with arbitrary or well chosen $F^0, X^0_{r+1}, \ldots, X^0_n$. Let $\mu^0_n = \frac{1}{n}[\alpha\mu + \sum_1^r \delta_{X_i} + \sum_{r+1}^n \delta_{X_i^0}]$. Then F^1 has to be generated from $\mathscr{D}(\alpha+n, \mu^0_n)$. Thus we may take $F^1(x) = \sum_k^\infty p_k I(W_i \leqslant x)$, where $W_1, W_2, \ldots$ are i.i.d. μ^0_n and the p_k's are given by (1) based on independent $\theta_1, \theta_2, \ldots$ which are i.i.d. Beta $(1, \alpha+n)$. This is an infinite sum and cannot be computed exactly. In fact, we will avoid generating this F^1 in a precise fashion and go to the next step of generating $(X^1_{r+1}, \ldots, X^1_n)$ as i.i.d. observations from this F^1 and satisfying the restrictions $X^1_i \in A_i$, $i = r+1, \ldots, n$. The constructive definition of Sethuraman (1994) states that $X^1_{r+1}, \ldots, X^1_n$ can be taken to be $W_{I_{r+1}}, \ldots, W_{I_n}$ where $I_{r+1}, \ldots, I_n$ are drawn i.i.d. from the random discrete distribution $\{p_k, k = 1, \ldots\}$. Thus one first generates a sufficiently large vector $W_1, \ldots, W_N$ which are i.i.d. μ^0_n. The random index I_{r+1} is then generated as $\min\{j: \sum_1^j p_i \geqslant U\}$, where U is a uniform random variable on [0, 1]. This means that we will need to generate only a finite (though random) number of p_is to generate I_{r+1}. If the value $X^1_{r+1} = W_{I_{r+1}}$ is not in A_{r+1} we will repeat this step until $X^1_{r+1} \in A_{r+1}$. This will give us X^1_{r+1}. We continue until the remaining $X^1_{r+2}, \ldots, X^1_n$ are generated. In a similar fashion, we generate $(F^M, X^M_{r+1}, \ldots, X^M_n)$ for $M = 1, 2, \ldots K$, noting that at each step we did not have to calculate the infinite series for F^M. The expected value of F^M, namely μ^M_n or the average of μ^M_n over several such runs, will be the computational Bayes estimate of μ. This computational method is described in Doss (1994). A result from Arthreya et al. (1994) can be used to justify that the Gibbs sampler converges.

We illustrate the above with the following example taken from Doss (1994). Kaplan and Meier (1958) presented the following data on lifetimes on eight subjects in which four were not fully observed, that is, were censored (indicated with a +):

$$0.8,\ 1.0+,\ 2.7+,\ 3.1,\ 5.4,\ 7.0+,\ 9.2,\ 12.1+.$$

The exact and computational Bayes estimates based on these data are very close to one another as can be seen from Table 3. For comparison purposes, the original standard frequentist estimate of Kaplan and Meier (KM) is also presented in the same table. As expected, the Bayesian estimate is smoother and varies over a wider range.

One can extend the above to the case where F is specified to come from a family $H(x-\theta)$, where H has a Dirichlet prior, and one can place an independent prior distribution on the real valued parameter θ. Doss (1994) extends the above Gibbs sampling method for this more general prior and obtains the posterior distribution and the Bayes estimate of θ. An illustrative numerical example can also be found in that paper.

The above constitute only a typical example of the use of Gibbs sampling in nonparametric Bayesian problems. There have been many interesting applications of Gibbs sampling; for some early references see Escobar (1994, 1995), and Escobar and West (1995).

3. Conclusion

Nonparametric methods have established themselves not just as competitors to parametric methods, but actually as being superior to them under certain circumstances, while still providing valid and robust procedures. Incorporating prior information and expert knowledge into nonparametric procedures becomes the next important advance in statistical methodology. In the first section we described various ways of defining Dirichlet measures which can serve as prior distributions in nonparametric problems. It turns out that the actual Bayesian solution to a particular nonparametric problem may not be in a simple closed form and therefore may require some intensive computing. The bootstrap method and Gibbs sampling represent two such computational procedures. We have presented two examples of the bootstrap: to test for concordance and to test for independence, and have illustrated the Gibbs sampler to estimate the survival distribution when censoring is present.

See also: Bayesian Statistics

Bibliography

Athreya K B, Doss H, Sethuraman J 1994 On the convergence of the Markov chain simulation method. *Annals of Statistics* **24**: 69–100

Bickel P J, Freedman D A 1981 Some asymptotic theory for the bootstrap. *Annals of Statistics* **9**: 1196–217

Blackwell D 1973 Discreteness of Ferguson selections. *Annals of Statistics* **1**: 356–8

Blackwell D, MacQueen J B 1973 Ferguson distributions via Pólya schemes. *Annals of Statistics* **1**: 353–5

Clark P J, Vandenberg S G, Proctor C H 1961 On the relationship of scores on certain psychological tests with a number of anthropomorphic characters and birth order twins. *Human Biology* **33**: 163–80

Dalal S R 1979 Dirichlet invariant processes and applications to nonparametric estimation of symmetric distributions functions. *Stochastic Processes and their Applications* **9**: 99–107

Doss H 1994 Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Annals of Statistics* **22**: 1763–86

Dykstra R L, Laud P 1981 A Bayesian nonparametric approach to reliability. *Annals of Statistics* **9**: 356–67

Efron B 1979 Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**: 1–26

Escobar M D 1994 Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**: 268–77

Escobar M D 1995 Nonparametric Bayesian methods in hierarchical models. *Journal of Statistical Planning and Inference* **43**: 97–106

Escobar M D, West M 1995 Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**: 577–88

Ferguson T S 1973 A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**: 209–30

Gelfand A E, Smith A F M 1990 Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409

Hoeffding W 1948 A class of statistics with asymptotically normal distribution. *Annals of Mathematic Statistics* **19**: 293–325

Hollander M, Korwar R M 1982 Nonparametric Bayesian estimation of the horizontal distance between two populations. In: Gnedenko B V, Puri M L, Vincze I (eds.) *Nonparametric Statistical Inference*. North Holland, New York, pp. 409–15

Hollander M, Wolfe D A 1999 *Nonparametric Statistical Methods*, 2nd edn. Wiley, New York

Kaplan E L, Meier P 1958 Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**: 457–81

Lo A Y 1981 Bayesian nonparametric statistical inference for shock models and wear processes. *Scandinavian Journal of Statistics* **8**: 237–42

Lo A Y 1993 A Bayesian bootstrap for censored data. *Annals of Statistics* **21**: 100–23

Lu H H S, Wells M T, Tiwari R C 1994 Inference for shift functions in the two-sample problem with right-censored data: With applications. *Journal of the American Statistical Association* **89**: 1017–26

Mauldin R D, Sudderth W D, Williams S C 1992 Pólya trees and random distributions. *Annals of Statistics* **20**: 1203–21

Meyn S P, Tweedie R L 1993 *Markov Chains and Stochastic Stability*. Springer-Verlag, New York

Rasekh J A, Kramer A, Finch R 1970 Objective evaluation of canned tuna sensory quality. *Journal of Food Science* **35**: 417–23

Nummelin E 1984 *General Irreducible Markov Chains and Nonnegative operators*. Cambridge University Press, Cambridge, UK

Revuz D 1984 *Markov Chains*, 2nd edn., North Holland, Amsterdam

Romano J P 1988 A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association* **83**: 698–708

Romano J P 1989 Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics* **17**: 141–59

Sethuraman J 1994 A constructive definition of Dirichlet priors. *Statistica Sinica* **4**: 639–50

Sethuraman J, Tiwari R C 1982 Convergence of Dirichlet measures and the interpretation of their parameter. In: Gupta S S, Berger J O (eds.) *Statistical Decision Theory and Related Topics III(2)*. Academic Press, New York, pp. 305–15

Singh K 1981 On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics* **9**: 1187–95

Susarla V, van Ryzin J 1976 Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association* **71**: 897–902

Tierney L 1994 Markov chains for exploring posterior distributions. *Annals of Statistics* **22**: 1701–62

Wells M T, Tiwari R C 1989 Bayesian quantile plots and statistical inference for nonlinear models in the two sample case with incomplete data. *Common Statistical Theory Methods* **18**: 2955–64

Yamato H 1975 A Bayesian estimation of a measure of the difference between two continuous distributions. *Report of the Faculty of Science at Kagoshima University* **8**: 29–38

Zalkikar J N, Tiwari R C, and Jammalamadaka S R 1986 Bayes and empirical Bayes estimation of the probability that $Z > X + Y$. *Common Statistical Theory Methods* **15**(10): 3079–101

M. Hollander and J. Sethuraman

Nonparametric Statistics: Asymptotics

Nonparametric methods are methods of statistical inference without precise assumptions on the shape of probability distribution of the observations. More precisely, a statistical model where the joint probability distribution is determined by a finite set of real parameters, or a finite dimensional real vector is called parametric, and methods of statistical inference without assuming a parametric model, is called nonparametric. Nonparametric methods, and especially nonparametric tests, have a long history. They can be traced back to the nineteenth century. Before 1940, rank correlation, rank sum tests, and permutation tests were known. However, they were often considered as merely 'quick and simple methods, substituting for more exact' procedures like t-tests, analysis of variance, etc., which were thought to be tedious.

The true nature of nonparametric methods has come to be understood since the 1950s and some of the nonparametric procedures have been found to have not only validity irrespective of the shape of underlying distribution, but also to have rather high efficiency under standard cases such as the normal distribution. Hence they have a high degree of robustness. Although the computation of the distribution of nonparametric statistics is now not difficult with modern computers, small sample distributions are usually complicated. Though they may not be difficult to compute for specific cases, it is not easy to understand their general features. Hence asymptotic theory is called for, and gives the result not for only null cases but also for alternative cases, leading to the definition and calculation of the relative efficiency of the procedures.

The term nonparametric is sometimes used in a different senses. Thus nonparametric regression usually means not that the distribution of the error is unspecified, but that the form of the regression

function is not specified by a finite number of parameters. However, such problems belong to a different area of theory and hence are not dealt with in this article. Basic aspects of the asymptotic nonparametric theory are expounded for some basic cases. There are two main nonparametric procedures: (a) testing hypotheses, and (b) interval estimation.

1. Two-sample Tests

Among nonparametric procedures, the most common is the nonparametric hypothesis testing, where the hypothesis to be tested is expressed in a nonparametric framework. Simplest and most often appearing in practical applications is the so-called two-sample problem where it is required to test the equality of two probability distributions. More precisely, we have two sets of sample observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ where X_is and Y_js are real random variables distributed according to continuous distributions, F and G respectively, and it is required to test the hypothesis that the two distributions are identical, that is

$$H\colon F \equiv G$$

A nonparametric test function ϕ is a measurable function of $\mathbf{X}$ and $\mathbf{Y}$, satisfying the conditions that

$$0 \leqslant \phi(\mathbf{X}, \mathbf{Y}) \leqslant 1,$$

and

$$E(\phi(\mathbf{X}, \mathbf{Y}) \mid F \equiv G) \leqslant \alpha \quad \text{for all possible } F$$

When $E(\phi \mid F \equiv G)$, which is called the size of the test, is always equal to α, the test is called a similar, or exactly nonparametric test.

When F is assumed to be continuous, there is a general method to derive similar tests. We define the order statistics $\tilde{X} = (X_{(1)} < X_{(2)} < \cdots < X_{(n)})$ and $\tilde{Y} = (Y_{(1)} < Y_{(2)} < \cdots < Y_{(m)})$ where $X_{(i)}$s are values of X_is arranged according to their magnitudes, and $Y_{(j)}$'s are ordered values of Y_js. Given $\tilde{X}$ and $\tilde{Y}$, the conditional distributions of X and Y are independent of F and G. Hence the pair of $\tilde{X}$ and $\tilde{Y}$ is a sufficient statistic, and we may consider ϕ as the function of $\tilde{X}$ and $\tilde{Y}$ (Fraser 1957, Lehmann 1959). Now we also define the combined order statistics $Z = (Z_1 < Z_2 < \cdots < Z_{n+m})$ where Z_ks are either one of the X_is or one of the Y_js, thus $\tilde{Z}$ is the set of values of $\tilde{X}$ and $\tilde{Y}$ combined and ordered. Then we have

$$P\{\tilde{X} = (Z_{i_1}, Z_{i_2}, \ldots, Z_{i_n}) \mid \tilde{Z} = (Z_1, Z_2, \ldots, Z_{n+m})\} = \frac{1}{{}_{n+m}C_n}$$

that is the conditional probability that the values of $\tilde{X}$ being equal to any combination of n values in $\tilde{Z}$ is always the same and equal to $(1/{}_{n+m}C_n)$, whatever the probability distribution $F \equiv G$ be, provided only that F is continuous so that no two values of Xs and Ys are equal.

Then if the test function satisfies the condition that

$$E(\phi(\tilde{X}, \tilde{Y}) \mid \tilde{Z}) \equiv \alpha \tag{1}$$

that is, the conditional expectation of α given the combined order statistic $\tilde{Z}$ is always equal to α under the hypothesis, it follows that

$$E(\phi(\tilde{X}, \tilde{Y}) \mid F \equiv G) = \alpha \quad \text{for all } F = G \tag{2}$$

hence ϕ is a similar test (Lehmann 1959). Conversely if ϕ is similar, it is proved that (2) must hold for almost all values of $\tilde{Z}$.

1.1 Test Statistic

Various similar test functions can be constructed in the following manner. Let t be a function of n values depending on the combined order statistics $\tilde{Z}$. Then the conditional distribution given $\tilde{Z}$ of $T = t(\tilde{x})$ can be computed from (2). Choose a constant c and also $0 \leqslant r \leqslant 1$, and define

$$\phi(\tilde{x}) = \begin{cases} 1 & \text{if} \quad t(\tilde{x}) > c \\ r & \text{if} \quad t(\tilde{x}) = c \end{cases}$$

so that $E(\phi(\hat{x}) \mid F \equiv G) = \alpha$. Thus we get a similar nonparametric test procedure.

There are two types of such functions. The first is a function whose marginal distribution is independent of the distribution F under the hypothesis. The rank of $x_i = x_i$ among $\tilde{Z} = (Z_1, Z_2, \ldots, Z_{n+m})$ $Z_1 < Z_2 < \cdots < Z_{n+m}$ is defined to be R_i if x_i is equal to Z_{R_i}. If t depends only on the ranks of X_is the distribution of $T = t(X) = t(R_1, R_2, \ldots, R_n)$ is clearly independent of F. By proper choice of t, most common nonparametric tests are obtained. The simplest one is the rank sum $T = \sum_{i=1}^{n} R_i$, and the hypothesis is rejected when T is either too large or too small.

The weighted rank or rank score sum test is a generalization of the rank sum test, defined by the criterion function $T = \sum_{i=1}^{n+m} c(R)_i$ where $c(R)$ is a function of the rank R. The score function $c(R)$ is often defined to be equal to the expectation of the R-th order statistic on the sample of size $n+m$ from some distribution. For example, one can take $c(R) = \xi_{R \mid n+m}$, the expectation of the R-th order statistics in the sample from the standerd normal distribution, or $c(R) = \frac{1}{R}$, the same from the exponental distribution. A more simple function is

$$c(R) = \begin{cases} \pm 1 & \text{for } R \lesseqgtr \frac{n+m+1}{2} \\ 0 & \text{for } R = \frac{n+m+1}{2} \end{cases}$$

which is called the median test. Sometimes a non-monotone function is used such as $c(R) = |\xi_{R(n+m)}|$.

Consider any statistic $T = t(X_1, X_2, \ldots, X_n)$. Its conditional distribution given $\tilde{Z}$ is called the permutation distribution of T. Then a nonparametric test can be obtained by using T as test criterion with the critical point calculated from the permutation distribution.

Using the idea of permutation tests every parametric test statistic can also be used as the second type of nonparametric test criterion. For example, if F and G are both normal distributions with possibly different means but with common variance, the best (in various senses) test statistic is the t-statistic defined as

$$T = \sqrt{\frac{nm}{n+m}}\frac{\tilde{X}-\tilde{Y}}{S}$$

where

$$S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{j=1}^{m}(Y_i-\bar{Y})}{(n+m-2)}$$

T is distributed according to the t-distribution with degrees of freedom $n+m-2$ under the normal hypothesis. But we can also consider the permutation distribution of T given $\tilde{Z}$. It can be shown that given $\tilde{Z}$, the statistic T is equivalent to $\tilde{X}$, the mean of the first sample.

1.2 Asymptotic Distribution Under the Hypothesis

The conditional distribution of any statistic $t(\tilde{X})$ given $\tilde{Z}$ is difficult to calculate in general, and so its asymptotic approximation plays an important role. A basic result under the hypothesis is the following (Fraser 1957).

Theorem: Under some regularity conditions, the conditional distribution of $T = (1/\sqrt{nv})(\bar{X}-\mu)$ approaches the standard normal distribution given $\tilde{Z} = (Z_1, Z_2, \ldots, Z_{n+m})$ when n and m tends to be large where

$$\mu = \frac{1}{n+m}\sum_{i=1}^{n+m} Z_i, \quad v^2 = \frac{\sum_{i=1}^{n+m}(Z_i-m)^2}{n+m-1}$$

Here the regularity condition is troublesome to express rigorously, but it requires basically that all moments of Z_i relative to v are bounded.

The theorem is readily generalized to the case when instead of $\tilde{X}$

$$\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

is used. Here g is a function which may depend on $\tilde{Z}$. Then the asymptotic normality of the null distribution of rank sum statistics and various permutation statistics can be derived, and asymptotic critical values for test statistics are obtained.

Thus for example, for the rank score sum test statistic,

$$T = \sum_{i=1}^{n} c(R_i)$$

$$E(T) = \frac{n}{n+m}\sum_{r=1}^{n+m} c(r) = \overline{nc(r)}$$

$$V(T) = \frac{nm}{(n+m)(n+m+1)}\sum_{i=1}^{n+m}(c(r)-\overline{c(r)})^2$$

and the approximate critical points can be obtained from the normal distribution. The asymptotically conditional distribution of the usual t statistic is equivalent to its permutation distribution, and the percentage point obtained from the t distribution approximately corresponds to that of the permutation distribution (Fisher 1935).

This fact may be interpreted to establish that the common t-test obtained under the assumption of normality is also asymptotically nonparametric, in the sense that it has approximately correct size when the sample size is large, if the two populations are the same, even if not normal.

1.3 Asymptotic Power

Although nonparametric tests are introduced to guarantee the level of the test under general conditions, it is also important to consider their power under the alternatives, consideration of which is the most important factor in choosing among many nonparametric tests. Before computers were developed, some nonparametric tests were preferred to the parametric tests because of computational simplicity, but this is no longer an issue.

In order to calculate the power we have to specify the alternative. Assume that the density function of $\mathbf{X}$ is expressed as $f_\theta(x)$ and that of $\mathbf{Y}$ as $f_0(x)$. Then the hypothesis to be tested is $H: \theta = 0$. Under the alternative, conditional distribution of $\tilde{X} = (X_1, X_2, \ldots, X_n)$ given $\tilde{Z}$ is expressed as

$$P(X_1 = Z_{i_1}, X_2 = Z_{i_2}, \ldots, X_n = Z_{i_n} \mid \tilde{Z}) = c_\theta(\tilde{Z})\prod_{j=1}^{n} f_\theta(Z_{i_j})$$

where $c_\theta(\tilde{Z})$ is a constant depending on $\tilde{Z}$. The conditional distribution of permutation test statistic can be derived from this formula.

When θ is small, that is the distribution of F_θ is close to F_0, the distribution under the alternative can also be approximated by the normal distribution. More precisely, consider a sequence of alternatives $\theta_N = \frac{\xi}{\sqrt{N}}$, and samples with

$$n = [\lambda N] \quad m = [(1-\lambda)N]$$

where

$$0 < \lambda < 1$$

is a constant. Then we have

Theorem: (Fraser 1957, Nikitin 1995) Under a set of regularity conditions, the conditional distribution given $\tilde{Z}$ of

$$T = \sqrt{\frac{N\lambda}{1-\lambda}}\,\frac{\tilde{X}-\mu}{v}$$

approaches the normal distribution with mean ψ and variance 1 where ψ is given by

$$\psi = \xi\sqrt{\lambda(1-\lambda)}\int x\,\frac{\partial}{\partial\theta}f(x,\theta)\,|_{\theta=0}\,\mathrm{d}x$$

Normal approximation for other types of statistics can also be derived. Under the same set up, the asymptotically best test statistic (assuming f is known) is distributed asymptotically normal with variance 1 and mean

$$\psi^* = \xi\sqrt{\lambda(1-\lambda)}\,\sqrt{I(\theta)}$$

when $I(\theta)$ is the Fisher information, that is

$$I(\theta) = \int\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)^2 f(x,\theta)\,\mathrm{d}x$$

The ratio $\psi^2/\psi^{*2} = e$ is called the relative efficiency of the nonparametric test, since the nonparametric test based on the sample of size N is nearly equal (in terms of power) to the best test based on a sample of size eN.

For the two-sample problem the most common model is that of shift parameter, that is, the two distributions are related by a real parameter θ so that

$$F(x) \equiv G(x-\theta)$$

Then the hypothesis corresponds to $H{:}\,\theta = 0$. For the general rank score test if the score function $c(R)$ can expressed as

$$c(R) = J\left(\frac{R}{n+m+1}\right)$$

where J is a continuous and smooth function with the conditions

$$\int_0^1 J(u)\,\mathrm{d}u = 0, \quad \int_0^1 J^2(u)\,\mathrm{d}u = 1$$

then the ψ above can be expressed as

$$\psi = \theta\sqrt{\lambda(1-\lambda)}\int f'(x)J(F(x))\,\mathrm{d}x$$

For example for the rank sum test,

$$J(u) = \frac{1}{\sqrt{12}}\left(u-\frac{1}{2}\right)$$

and

$$\psi = \theta\sqrt{\frac{\lambda(1-\lambda)}{12}}\int f'(x)\left(F(x)-\frac{1}{2}\right)\mathrm{d}x$$
$$= \theta\sqrt{\frac{\lambda(1-\lambda)}{12}}\int f^2(x)\,\mathrm{d}x$$

from which its relative efficiency for the normal case is calculated to be equal to $3/\pi = 95$ percent.

If J is chosen to be equal to

$$J(u) = c\,\frac{\mathrm{d}}{\mathrm{d}u}h_f(u), \quad \text{where } h_f(u) = f(F^{-1}(\mu))$$

the asymptotic efficiency of the rank score test is 100 percent. Thus for any specific distribution there corresponds a nonparametric rank score test which has 100 percent efficiency (Lehmann 1959).

It is also easily proved that the permutation test derived from the best test statistic under a specified distribution is asymptotically equivalent to the best test. Since the best nonparametric test is dependent on the unknown shape of F, there is no uniformly best nonparametric test. But some information about the shape of the distribution is available from the sample, which may be utilized to choose the nonparametric test among a class of tests.

For example, the best score function $J(u)$ can be approximately constructed from the sample information. Such a method is called an adaptive procedure. There have been proposals and theoretical results about adaptive nonparametric procedures, and there exist adaptive procedures which are 100 percent asymptotically efficient for all distributions within a rather large class (Nikitin 1995). But these are usually very complicated, and some of them have distributions notoriously slow to converge asymptotically. Hence they are not necessarily recommendable in practical situations.

2. *Common Nonparametric Tests*

There are various problems of testing statistical hypotheses, where several types of nonparametric tests are derived in similar ways, as in the two-sample case.

2.1 *Equality of Several Populations*

Suppose that we have k sets of samples, each of size n_i from the population with distribution F_i. The hy-

pothesis to be tested is $H: F_i \equiv F$. Let $X_i = (X_i, X_{i2}, \ldots, X_{in})$ be the set of the values in the sample from the i-th population, and $\tilde{Z} = (X_1, X_2, \ldots, X_k)$ conditional distribution given $\tilde{Z}$ is expressed as the total set of values of the k samples combined. Then under the hypothesis the

$$P(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k \mid \tilde{Z}) = \frac{n_1! n_2! \ldots n_k!}{N}$$

And nonparametric tests can be derived from this permutation distribution.

One class of such tests can be obtained from permutation distribution of the usual test criteria such as

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2}{(k-1)S^2}, \quad S^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2}{(N-k)}$$

where

$$N = \sum_{i=1}^{k_i} n_i, \quad x_i = \sum_{j=1}^{n_i} X_{i,j}/n_i, \quad \bar{\bar{X}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{i,j}/N$$

or in case $n_j = n$

$$R = \max_{i,i'} \sqrt{n}\,(\bar{X}_i - \bar{X}_{i'})S$$

When n_is are large, $(k-1)F$ is distributed asymptotically according to the chi-square distribution with $k-1$ degrees of freedom and R has the same asymptotic distribution as the same as the normal studentized sample range (Randles and Wolfe 1979).

Another class of criteria is obtained by substituting the rank score $c(R_{i,j})$ for $X_{i,j}$, where $R_{i,j}$ is the rank of $X_{i,j}$ in $\tilde{Z}$. Test criteria corresponding to the F test can be expressed as

$$T = \sum_{i=1}^{k} n_i(\bar{c}_i - \bar{\bar{c}})^2/v^2$$

where

$$\bar{\bar{c}} = \frac{1}{n} \sum_{r=1}^{N} c(r), \quad v^2 = \frac{1}{N-1} \sum_{j=1}^{N} (c(r) - \bar{c})^2,$$

$$\bar{c}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} c(R_{i,j})$$

The distribution of T can be approximated by the chi-square distribution. Under the alternative close to the hypothesis, the asymptotic distribution of T is expressed as a non-central chi-square distribution. More precisely, when the distribution F_i is expressed as $F_i(x) = F_{\theta i}(x)$ with real parameter and known function $F_\theta(x)$, the hypothesis expressed as $H: \theta_i \equiv \theta_0$, and with the sequence of samples of size $n_i = \bar{\lambda}_i N$, $\sum_{i=1}^{k} \lambda_i = 1$ under the sequence of alternatives $\theta_i = \theta_0 + \xi_i/\sqrt{N}$, the statistic T is distributed asymptotically as the non-central chi-square distribution with degrees of freedom $k-1$, and non-centrality $\psi = \sum_{i=1}^{k} \lambda_i \xi_i^2 \times \delta$. The constant δ depends both on the shape of the distribution and the score function $c(R)$. The relative efficiency of such tests can be defined as in the two-sample case, and with the same score function, the relative efficiency of the rank score square sum test is equal to that of the rank score test in the two-sample case (Lehmann 1975).

2.2 Location of a Symmetric Distribution

Consider the case when $X_1, X_2, \ldots, X_n$ is a sample from a symmetric distribution centered at θ, i.e., its probability density function $f(x-\theta)$ is an even function $f(-x) = f(x)$, but otherwise is not specified. It is required to test the hypothesis $H: \theta = \theta_0$. Non-parametric test procedures can be obtained in the following way.

Define $Z_i = |X_i - \theta_0|$ and $\varepsilon_i = sgn(X_i - \theta_0)$. Then it is easily shown that under the hypothesis, ε_is are independent and $P(\varepsilon_i = \pm 1) = 1/2$. Let $\tilde{Z} = (Z_1, Z_2, \ldots, Z_n)$ be the set of values of Z_i. Then given $\tilde{Z}$, the conditional distribution of the statistic

$$T = t(X_1, X_2, \ldots, X_n) = t(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n, Z_1, Z_2, \ldots, Z_n)$$

is obtained. Statistics of the form $T = \sum_{i=1}^{n} \varepsilon_i g(Z_i)$ have the mean and variance $E(T) = 0$, $V(T) = \sum_{i=1}^{n} g(Z_i)^2$. Its conditional distribution can be approximated by the normal distribution when n is large.

Let R_i be the rank of Z_i. Then the test based on $T = \sum_{i=1}^{n} \varepsilon_i R_i$ is called the signed rank sum test, and more generally $T = \sum_{i=1}^{n} \varepsilon_i c(R_i)$ is called a signed rank score test statistic. The relative efficiency of such a test is defined can calculated in a completely similar way, as in the two-sample case.

2.3 Independence of Two Variables

Let (X_i, Y_i), $i = 1, 2, \ldots, n$ be a sample from a bivariate distribution. Let $F(x, y)$ be the joint distribution function. Consider the hypothesis that X and Y are independent, i.e. $F(x, y) \equiv G(x)H(y)$ assuming G and H are absolutely continuous but without any further specification.

Let $\tilde{X} = (X_1, X_2, \ldots, X_n)$ and $\tilde{Y} = (Y_1, Y_2, \ldots, Y_n)$ be the set of X-values and Y-values. Then under the hypothesis the conditional distribution of (X_i, Y_i),

$i = 1, 2, \ldots, n$ given $\tilde{X} = (x_1, x_2, \ldots, x_n)$ and $\tilde{Y} = (y_1, y_2, \ldots, y_n)$ is expressed as

$$P = (X_1 = x_{i_1}, Y_1 = y_{i_1}, X_2 = x_{i_2}, Y_2 = y_{i_2}, \ldots, X_n = x_{i_n}, Y_n = y_{i_n}) = \frac{1}{n!^2}$$

for any permutation $(i_1, i_2, \ldots, i_n)$ and $(j_1, j_2, \ldots, j_n)$. This distribution is also called the permutation distribution.

Then we may define the generalized correlation coefficient

$$R = \frac{\sum_{i=1}^{n} \phi(X_i)\phi(Y_i)}{\sqrt{\sum_{i=1}^{n} \phi(X_i)^2 \sum \phi(Y_i)}}$$

with the condition

$$\sum_{i=1}^{n} \phi(X_i) = 0, \quad \sum_{i=1}^{n} \phi(Y_i) = 0$$

Under the hypothesis given $\tilde{X}$ and $\tilde{Z}$

$$E(R) = 0 \quad V(R) = \frac{1}{n-1}$$

As n tends to infinity the distribution of R approaches the standard normal distribution (Kendall 1948).

When $\phi(X_i) = R_i$, R is called the rank correlation coefficient (or more precisely Spearman's ρ). When $\phi(X_i) = X_i$, R is equal to the usual (moment) correlation coefficient. Hence it can also be interpreted as a nonparametric correlation coefficient if its permutation distribution is taken into consideration.

2.4 Equality of Multivariate Distributions

Multivariate two-sample problems can be treated in the same way as in the univariate case. Suppose that we want to test the equality of two bivariate distributions.

Let $X = \{(X_{1,1}, X_{1,2}), (X_{2,1}, X_{2,2}), \ldots, (X_{n,1}, X_{n,2})\}$ be the bivariate sample of size n from the first distribution, and $Y = \{(Y_{1,1}, Y_{1,2}), (Y_{2,1}, Y_{2,2}), \ldots, (Y_{m,1}, Y_{m,2})\}$ be the sample of size m from the second distribution. The hypothesis to be tested is that the two distributions are continuous and identical, but not otherwise specified.

Let $\tilde{Z}$ be the totality of the $n+m$ pairs of values of $\tilde{X}$ and $\tilde{Y}$. Then given $\tilde{Z}$, the conditional probability that the pairs in X are equal to the specific n pairs in $\tilde{Z}$ is equal to $1/{}_{n+m}C_n$ as in the univariate case. Nonparametric tests can be derived from this fact.

Define $T_1 = \sum g_1(X_{i,1})$ and $T_2 = g_2(X_{i,2})$. Then under the hypothesis the conditional distribution given $\tilde{Z}$ of (T_1, T_2) approaches a bivariate normal distribution as n and m get large (under a set of regularity conditions). Hence we can define

$$\chi^2 = s^{11}(T_1 - m_1)^2 + s^{12}(T_1 - m_1)(T_2 - m_2) + s^{22}(T_2 - m_2)^2$$

where

$$m_1 = E(T_1 \mid \tilde{Z}), \quad m_2 = E(T_2 \mid \tilde{Z})$$

and s^{11}, s^{12}, s^{22} are the elements of inverse of conditional variance and covariance matrix of T_1 and T_2. Then under the hypothesis χ^2 is asymptotically distributed as chi-square distribution of 2 degrees of freedom. By various choices of the function g_1, g_2, we can get bivariate versions of rank sum, rank score, etc., tests (Puri and Sen 1971).

3. Nonparametric Confidence Intervals

Nonparametric test procedures can be applied to construct nonparametric confidence intervals. For example, consider the two-sample location shift model i.e., the two distributions are related as $F(x) = G(x - \theta)$. Suppose we want to construct a confidence interval $(l(X, Y), u(X, Y))$, such that

$$P\{l(X, Y) < \theta < u(X, Y\} \geqslant 1 - \alpha$$

for all θ and for all continuous G.

Now consider the hypothesis $\theta = \theta_0$. It is equivalent to the hypothesis that $X_i' = X_i - \theta_0$, $i = 1, 2, \ldots, n$ has the same distribution as Y_i. Hence any two-sample nonparametric test procedure can be applied to $(X_1', X_2', \ldots, X_n')$ and $(Y_1, \ldots, Y_m)$. Let $\phi(\tilde{X}, \tilde{Y})$ be the test function. Then $E_{\theta_0}(\phi(\tilde{X}, \tilde{Y})) \leqslant \alpha$.

Now consider the class of hypotheses for any specific value of θ, and let the test functions corresponding to θ be $\phi_\theta(\tilde{X}, \tilde{Y})$. Corresponding to $\phi_\theta(\tilde{X}, \tilde{Y})$, let $S \subset R^1$ (real line) be the set of values of θ for which $\phi_\theta(\tilde{X}, \tilde{Y}) = 0$. Then for any θ we have $P_\theta(\theta \in S) = E_\theta(1 - P_\theta(\tilde{X}, \tilde{Y})) \geqslant 1 - \alpha$. Therefore S is a confidence region of θ with level $1 - \alpha$, and if S is always defined as an interval, it gives a nonparametric confidence interval with level $1 - \alpha$.

With a monotone increasing score function we can always construct such a confidence interval as follows:

Let $R_{\theta,i}$ be the rank of $X_i - \theta$, $i = 1, 2, \ldots, n$ in the combined set of values $X_i - \theta$, $i = 1, 2, \ldots, n$ and Y_j, $j = 1, 2, \ldots, m$. Let $c(R)$ be the monotone increasing score function. Then we can determine two constants a and b such that

$$P\{a \leqslant \sum_{i=1}^{n} c(R_{\theta,i}) \leqslant b\} \geqslant 1 - \alpha$$

Then define

$$\underline{\theta} = \inf \left\{ \sum_{i=1}^{n} c(R_{\theta,i}) \leqslant b \right\}$$

$$\bar{\theta} = \sup \left(\sum_{i=1}^{n} c(R_{\theta,i}) \geqslant a \right)$$

Since $\sum_{i=1}^{n} c(R_{\theta,i})$ is monotone non-increasing in θ, $\underline{\theta} < \theta < \bar{\theta}$ gives a confidence interval of level $1-\alpha$.

Also from the consideration of the permutation distribution of $T = \sum_{i=1}^{n} X_i'$, we can construct a nonparametric confidence interval of the form

$$\bar{X} - \bar{Y} - D < \bar{\theta} < \bar{X} - \bar{Y} + D$$

When n and m are large, asymptotic results can be applied, and the above interval becomes nearly equivalent to the t-interval derived from the assumption of normality.

Bibliography

Fisher R A 1935 *The Design of Experiments*. Oliver and Boyd

Fraser D A S 1957 *Nonparametric Methods in Statistics*. Wiley

Hajek J 1969 *A Course in Nonparametric Statistics*. Holden-Day

Hettmansperger T P 1984 *Statistical Inference Based on Ranks*. Wiley

Hollander M, Wolfe D A 1973 *Nonparametric Statistical Methods*. John and Sons

Kendall M G 1948 *Rank Correlation Methods*. Griffin

Lehmann E L 1959 *Testing Statistical Hypotheses*. Wiley

Lehmann E L 1975 *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day

Nikitin Y 1995 *Asymptotic Efficiency of Nonparametric Tests*. Cambridge University Press, Cambridge, UK

Pratt J W, Gibbons J D 1981 *Concepts of Nonparametric Theory*. Springer-Verlag

Puri M L, Sen P K 1971 *Nonparametric Methods in Multivariate Analysis*. John Wiley and Sons

Randles R H, Wolfe D A 1979 *Introduction to the Theory of Nonparametric Statistics*. Wiley and Sons

Walsh J E 1965 *Results for Two and Several Sample Problems, Symmetry, and Extremes (Handbook of Nonparametric Statistics)*. Van Nostrand

K. Takeuchi

Nonparametric Statistics: Rank-based Methods

A nonparametric statistical procedure is a statistical procedure, such as (a) a hypothesis test, (b) a confidence interval, (c) a multiple comparison procedure, (d) a confidence band, that has desirable properties that hold under mild assumptions concerning the underlying population from which the data are drawn. Typically, it suffices to assume the underlying population is continuous.

The term 'distribution-free' has a more precise meaning than the term 'nonparametric.' For example, in the context of hypothesis testing, a distribution-free statistic is one whose probability distribution, under the null hypothesis being tested, is independent of the underlying population distribution. In the context of interval estimates, a distribution-free $100(1-\alpha)$ percent confidence interval, say (Δ_L, Δ_U) for an unknown parameter Δ, is a random interval having the property that the probability that Δ is contained in the random interval is $1-\alpha$, independent of the underlying population distribution. There are related concepts for distribution-free multiple comparison procedures (see Sect. 1.2) and distribution-free confidence bands for distribution functions (see Sect. 11.5 of Hollander and Wolfe (1999)).

The success of nonparametric statistics over the last 70 years and its continued vigorous development today are due to the wide applicability of the nonparametric approach and the excellent efficiency properties of nonparametric procedures. Some of the main advantages can be briefly summarized as follows

(a) Nonparametric procedures provide exact P-values for tests, exact coverage probabilities for confidence intervals, exact experimentwise error rates for multiple comparison procedures, and exact coverage probabilities for confidence bands. These exact properties do not require an assumption that the underlying population have a specific form, such as a normal population.

(b) Nonparametric procedures have excellent efficiency properties with respect to their parametric competitors.

(c) Nonparametric procedures are relatively insensitive to outlying observations in the sense that a few wildly big (or small) observations that are far removed from their fellow observations have little influence on the procedure (test, confidence interval, multiple comparison procedure, etc.).

Section 2 describes some basic aspects of two- and k-sample location problems. These problems and their generalizations drove the early development of nonparametric statistics, and the procedures developed remain extremely useful for data analysis today. More modern topics in nonparametrics include Bayesian and computationally intensive procedures (cf. *Nonparametric Statistics: Advanced Computational Approaches*). Section 1.7 of Hollander and Wolfe (1999) provides a brief history of nonparametric statistics and contains source references to work in other currently active areas including regression problems and survival analysis. Other recent books covering various aspects of nonparametric statistics include Sprent (1993), Gibbons (1997), and Hettmansperger and McKean (1998). Earlier applied books featuring

nonparametric methods based on ranks include Noether (1967) and Lehmann (1975). Randles and Wolfe (1979) is an introduction to the theory of nonparametric statistics.

1. Rank-based Procedures for Location Alternatives

The primary reasons for using ranks are to obtain exact procedures that enjoy a distribution-free property and to achieve robustness by diminishing the deleterious effects of outlying observations and departures from the model. One might think that the reduction to ranks from original data values would sacrifice too much information, but power and efficiency studies have shown that is not the case. Indeed the nonparametric rank-based procedures are only slightly less efficient than their normal theory competitors when the underlying populations are normal (the home court of normal theory methods), and can be much more efficient when the underlying populations are not normal. We begin our description of rank-based methods with procedures for the two-sample location problem, then consider the k-sample location problem and the multivariate two-sample location problem. There is an analogous body of nonparametric techniques (which parallels the methods described in Sect. 2) for one-sample problems, paired-sample problems, randomized blocks and multivariate one-sample problems. That body of procedures also enjoys excellent efficiency and robustness properties. See Hollander and Wolfe (1999) and Hettmansperger and McKean (1998) for coverage of these methods.

1.1 The Two-sample Location Problem

A basic problem in statistics is the two-sample location problem. Assume that we have a random sample $X_1,\ldots, X_m$ from population 1 having distribution function $F(x) = P(X \leqslant x)$ and an independent random sample $Y_1,\ldots, Y_n$ from population 2 having distribution function $G(x) = P(Y \leqslant x)$. We wish to test the null hypothesis H_0 that the two populations are equivalent, that is,

$$H_0: F(x) = G(x), \quad \text{for all } x. \tag{1}$$

If we are interested in location alternatives, these can be modeled using the location shift model

$$G(x) = F(x-\Delta), \quad \text{for all } x \tag{2}$$

where the parameter Δ is commonly called the 'treatment' effect. Model (2) signifies that a randomly selected value Y from population 2 has the same distribution as $X+\Delta$ where X is a randomly selected value from population 1, that is $Y \stackrel{d}{=} X+\Delta$, where $\stackrel{d}{=}$ means 'has the same distribution as.' When $\Delta = 0$, H_0 is satisfied. In addition to testing H_0, it is useful to estimate Δ and give a confidence interval for Δ. An excellent way to test H_0, estimate Δ, and give a confidence interval for Δ is to use procedures based on (and derived from) the famous Wilcoxon (1945) two-sample rank sum statistic.

The Wilcoxon two-sample rank sum statistic is

$$W = \sum_{j=1}^{n} R(Y_j) \tag{3}$$

where $R(Y_j)$ is the rank of Y_j in the joint ranking from least to greatest of the $N = m+n$ X and Y. Significantly large values of W indicate $\Delta > 0$; significantly small values of W indicate $\Delta < 0$. A one-sided upper-tail distribution-free test of $H_0: \Delta = 0$ versus $H_1: \Delta > 0$, at the α level of significance is,

$$\text{reject } H_0 \text{ if } W \geqslant w_\alpha; \quad \text{otherwise accept } H_0 \tag{4}$$

where w_α is chosen to make the type-I error probability of the test equal to α and is the upper α^{th} percentile point of the null distribution of W.

Similarly, a one-sided lower-tail distribution-free test of $H_0: \Delta = 0$ versus $H_2: \Delta < 0$, at the α level of significance, is

$$\text{reject } H_0 \text{ if } W \leqslant n(m+n+1) - w_\alpha; \quad \text{otherwise accept } H_0. \tag{5}$$

A two-sided distribution-free test of $H_0: \Delta = 0$ vs. $H_3: \Delta \neq 0$, at the α level of significance, is

$$\text{reject } H_0 \text{ if } W \geqslant w_{\alpha/2} \text{ or if } W \leqslant n(m+n+1) - w_{\alpha/2}; \quad \text{otherwise accept } H_0. \tag{6}$$

This is the two-sided symmetric test with $\alpha/2$ probability in each tail of the distribution.

The tests based on W are distribution-free under H_0 (1) assuming the common value of F and G is a continuous distribution. For then, all $\binom{N}{n}$ possible assignments of the Y-ranks are equally likely, each having probability $1\backslash\binom{N}{n}$ when H_0 is true. Values of w_α can be obtained from Table A.6 of Hollander and Wolfe (1999). Exact P-values can also be obtained using the software StatXact (1999).

The large sample approximation to procedure (4) is

$$\text{reject } H_0 \text{ if } W^* \geqslant z_\alpha; \quad \text{otherwise accept } H_0 \tag{7}$$

where

$$W^* = \frac{W - E(W)}{\{\text{var}_0(W)\}^{\frac{1}{2}}} = \frac{W - \{n(m+n+1)/2\}}{\{mn(m+n+1)/12\}^{\frac{1}{2}}} \tag{8}$$

is the standardized version of W. It can be treated as an approximate $N(0, 1)$ random variable under H_0. In (7), z_α is the upper α percentile point of a $N(0, 1)$ distribution.

If ties are present in the data, assign to tied observations the average of the ranks for which those observations are competing. Then procedure (4) remains applicable but it is now approximate rather than exact. Alternatively, one can use an exact conditional test based on the exact conditional randomization distribution of W, see Hollander and Wolfe (1999, Comment 4.5). When there are ties, the null mean $E_0(W)$ is unchanged but the null variance $\mathrm{var}_0(W)$ is reduced to the right-hand-side of (9). Thus when applying the large sample approximation in the presence of ties, use average ranks and replace $\mathrm{var}_0(W)$ in (8) by

$$\mathrm{var}_0(W) = \frac{mn}{12}\left(m+n+1-\frac{\sum_{j=1}^{g}(t_j-1)t_j(t_j+1)}{(m+n)(m+n-1)} \right) \quad (9)$$

where g denotes the number of tied groups, t_j is the size of tied group j, and an untied observation is considered to be a tied group of size 1.

Thus far in this section we have considered testing $H_0{:}\Delta = 0$. To test $H_0''{:}\Delta = \Delta_0$ where Δ_0 is some specified nonzero number, subtract Δ_0 from each Y to form $Y_1' = Y_1 - \Delta_0, \ldots, Y_n' = Y_n - \Delta_0$. Then compute W as the sum of the ranks of Y' in the joint ranking from least to greatest of the m X's and the n Y''s. With W computed in this manner, the procedures described earlier in this section may be directly applied to test H''_0.

There are medical situations for comparing treatments when various restricted randomization designs are employed. In such situations, the test based on W will differ from that described here. See Hollander and Peña (1988) and Mehta et al. (1988) for details.

A point estimator of Δ, derived from the Wilcoxon rank sum statistic, is the Hodges–Lehmann (1963) estimator, namely

$$\hat{\Delta} = \mathrm{median}\{Y_j - X_i,\ i = 1, \ldots, m,\ j = 1, \ldots, n\} \quad (10)$$

That is, $\hat{\Delta}$ is the median of the mn $(Y-X)$ differences. Roughly speaking $\hat{\Delta}$ is the value $\hat{\Delta}$ that should be subtracted from each Y so that the value of W, when applied to the aligned samples $X_1, \ldots, X_m$, $Y_1 - \hat{\Delta}, \ldots, Y_n - \hat{\Delta}$, is equal to $n(m+n+1)/2$. One can think of the method as estimating Δ by the amount $\hat{\Delta}$ the Y's should be shifted so that the sample $X_1, \ldots, X_m$ and the pseudo-sample $Y_1 - \hat{\Delta}, \ldots, Y_n - \hat{\Delta}$ look like two samples from the same distribution as viewed by W.

A symmetric two-sided $100(1-\alpha)\%$ distribution-free confidence interval for Δ, obtained by inverting the test statistic, is

$$\Delta_L = U^{(C_\alpha)}, \quad \Delta_U = U^{(mn+1-C_\alpha)} \quad (11)$$

Table 1
Plasma glucose values for healthy and lead-poisoned Canadian geese (ranks in parentheses)

Healthy geese	Lead-poisoned geese
297 (8)	293 (7)
340 (12)	291 (6)
325 (10)	289 (4)
227 (1)	430 (14)
277 (3)	510 (15)
337 (11)	353 (13)
250 (2)	318 (9)
290 (5)	

Source: March et al. 1976.

where

$$C_\alpha = \frac{n(2m+n+1)}{2} + 1 - w_{\alpha/2} \quad (12)$$

$w_{\alpha/2}$ is the upper $\alpha/2$ percentile point of the null distribution of W, and $U^{(1)} \leqslant \cdots \leqslant U^{(mn)}$ denote the ordered values of the mn $(Y-X)$ differences. For large m and n, the integer C_α may be approximated by

$$C_\alpha \approx \frac{mn}{2} - z_{\alpha/2}\left(\frac{mn(m+n+1)}{12} \right)^{\frac{1}{2}}. \quad (13)$$

The value C_α given by (13) will typically not be an integer. To be conservative, when using the large sample approximation to the confidence interval, take C_α to be the largest integer that is less than or equal to the right-hand-side of (13).

Example 1. We illustrate the two-sample Wilcoxon procedures with data from March et al. (1976) (see also Hollander and Wolfe (1999)) who examined differences between healthy and lead-poisoned Canadian geese. One of the measures they considered was plasma glucose (in mg/100 ml plasma). The data for eight healthy geese and seven lead-poisoned geese are given in Table 1. From Table 1, we find $W = 68$. Referring $W = 68$ to Table A.6 of Hollander and Wolfe (1999) with $n = 7$ and $m = 8$, we find $P_0(W \geqslant 68) = 0.095$. (In Table A.6, $n \leqslant m$, therefore designate the values corresponding to the smaller sample as Y.) That is, the probability, under the null hypothesis, of getting a value as larger or larger than 68, is 0.095. This is the P-value for the one-sided test of $\Delta = 0$ versus $\Delta > 0$. (StatXact software gives the P-value 0.0946). Thus one would reject $\Delta = 0$ in favor of $\Delta > 0$ for all type-I error probabilities $\alpha \geqslant 0.095$ and accept $\Delta = 0$ for all type-I error probabilities $\alpha < 0.095$. Thus there is borderline evidence that the lead-poisoned geese have higher plasma glucose values.

To compute the Hodges–Lehmann estimate $\hat{\Delta}$, we first obtain the ordered $Y-X$ differences $U^{(1)} \leqslant \cdots \leqslant U^{(56)}$. The ordered differences are: -51,

$-49, -48, -47, -46, -44, -36, -34, -32, -22, -19, -8, -7, -6, -4, -1, 1, 3, 12, 13, 14, 16, 16, 21, 28, 28, 39, 41, 41, 43, 56, 62, 63, 64, 66, 68, 76, 90, 91, 93, 103, 105, 126, 133, 140, 153, 170, 173, 180, 185, 203, 213, 220, 233, 260, 283$. Then $\hat{\Delta}$ is found to be 41, the average of $U^{(28)}$ and $U^{(29)}$. To obtain a 96 percent confidence interval, set $\alpha/2 = 0.02$ and from Table A.6 of Hollander and Wolfe (1999) obtain $w_{0.02} = 74$. Then from (12), $C_{0.04} = 11$, and from (11) we obtain $\Delta_L = U^{(11)} = -19$, $\Delta_U = U^{(46)} = 153$.

1.2 The k-sample Problem

The two-sample problem of Sect. 1.1 is readily extended to $k \geqslant 2$ samples, one sample from each of k populations. The basic null hypothesis is that the k samples can be viewed as a single combined sample from a common population. Typically we are interested in alternatives that specify that the locations (medians, means) of the k populations are not equal. More formally, let $X_{ij}, X_{2j},\ldots, X_{n_j j}$ denote the random sample of n_j observations from the jth population, which is assumed to have an unknown continuous distribution F_j, $j = 1,\ldots, k$ where $F_j(x) = P(X_{ij} \leqslant x)$ for $i = 1,\ldots, n_j$. We also assume that all $N = \sum_{j=1}^{k} n_j$ random variables are mutually independent. The null hypothesis of equivalent populations is

$$H_0: F_1(x) = F_2(x) = \ldots = F_k(x), \quad \text{for all } x. \quad (14)$$

To study location alternatives it is natural to introduce the location shift model

$$F_j(x) = F(x-\tau_j), \quad j = 1,\ldots, k \quad (15)$$

where F is an unknown continuous distribution function and τ_j is the unknown treatment effect corresponding to the jth population. In submodel (15), the null hypothesis H_0 of (14) becomes

$$H_0: \tau_1 = \tau_2 = \ldots = \tau_k. \quad (16)$$

An excellent way to test H_0 is to use the Kruskal–Wallis test (see Kruskal (1952), Kruskal and Wallis (1952)). Let τ_{ij} denote the rank of X_{ij} in the joint ranking of all N observations from the k samples. Furthermore, let

$$R_j = \sum_{i=1}^{n_j} r_{ij}, \quad R.j = R_j/n_j, \quad j = 1,\ldots, k \quad (17)$$

denote, respectively, the sum and average of the ranks received by the jth sample. The Kruskal–Wallis statistic is

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(R.j - \frac{N+1}{2} \right)^2 = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1). \quad (18)$$

Significantly large values of H indicate the τ's are not all equal to a common value. A distribution-free test at the α level of significance, of H_0 (16), against the alternative that the τ's are not all equal, is

$$\text{reject } H_0 \text{ if } H \geqslant h_\alpha; \quad \text{otherwise accept } H_0 \quad (19)$$

where h_α is chosen to make the type-I error probability of the test equal to α and is the upper α^{th} percentile point of the null distribution of H. (For $k = 2$, this test reduces to the two-sided Wilcoxon test of (6)).

The test based on H is distribution-free under H_0 (14) assuming the common distribution of F is continuous. For then all $N!/(\prod_{j=1}^{k} n_j!)$ possible assignments of n_1 ranks to the sample 1 observations, n_2 ranks to the sample 2 observations, $\ldots n_k$ ranks to the sample k and Wolfe (1999). StatXact performs the exact test as well as the large sample approximation next described.

The large sample approximation to procedure (19) is

$$\text{reject } H_0 \text{ if } H \geqslant \chi^2_{k-1,\alpha}; \quad \text{otherwise accept } H_0 \quad (20)$$

where $\chi^2_{k-1,\alpha}$ is the upper α^{th} percentile point of a chi-squared distribution with $k-1$ degrees of freedom.

For a ties correction to the large sample approximation, compute H using average ranks. That is, assign each of the observations in a tied group the average of the ranks those observations would receive if they were not tied. Then refer to H' to a chi-squared distribution with $k-1$ degrees of freedom where

$$H' = H \Big/ \left(1 - \left(\sum_{j=1}^{g} (t_j^3 - t_j)/(N^3 - N) \right) \right) \quad (21)$$

In (21), g denotes the number of tied groups and t_j is the size of tied group j (and an untied observation is considered to be a tied group of size 1). The ties correction of (21) does not yield an exact α level test. To obtain an exact α-level test in the presence of ties, see Hollander and Wolfe (1999, Comment 6.8).

If the hypothesis H_0 (16) of equivalent treatments is rejected, it is natural to investigate which particular pairs of treatments differ from one another. This is the aim of multiple comparisons. There are various multiple comparison procedures available (cf. Miller (1981), Hettmansperger and McKean (1998), Hochberg and Tamhane (1987), Hollander and Wolfe (1999)). Here we will describe and illustrate the multiple comparison procedure proposed independently by Dwass (1960) and Steel (1960) for the equal sample size case and generalized by Critchlow and Fligner (1991) to allow unequal sample sizes. The procedure is based on $\binom{k}{2}$ separate pairwise rankings, corresponding to ranking jointly each pair of samples.

For each pair of treatments (u, v), let W_{uv} denote the Wilcoxon rank sum of the ranks attained by the n_v observations in sample v in the joint ranking of the n_u observations from sample u and the n_v observations from sample v.

Table 2
Ages (in months) of infants for walking alone (ranks in parentheses)

Active exercise	Passive exercise	No exercise	8-week
9.00 (1.5)	11.00 (10)	11.50 (12.5)	13.25 (20.5)
9.50 (3.5)	10.00 (7)	12.00 (16.5)	11.50 (12.5)
9.75 (5)	10.00 (7)	9.00 (1.5)	12.00 (16.5)
10.00 (7)	11.75 (15)	11.50 (12.5)	13.50 (22)
13.00 (18.5)	10.50 (9)	13.25 (20.5)	11.50 (12.5)
9.50 (3.5)	15.00 (23)	13.00 (18.5)	
$R_1 = 39$	$R_2 = 71$	$R_3 = 82$	$R_4 = 84$

Source: Zelazo et al. 1972.

Let

$$W^*_{uv} = \sqrt{2}\left(\frac{W_{uv} - E_0(W_{uv})}{\{\mathrm{var}_0(W_{uv})\}^{\frac{1}{2}}}\right) = \left(W_{uv} - \frac{n_v(n_u+n_v+1)}{2}\right) \Bigg/ \left(\frac{n_u n_v(n_u+n_v+1)}{24}\right)^{\frac{1}{2}}, \quad 1 \leqslant u < v \leqslant k. \tag{22}$$

A distribution-free two-sided all treatments multiple comparison procedure, at an experimentwise error rate of α, is

$$\text{decide } \tau_u \neq \tau_v \text{ if } |W^*_{uv}| \geqslant w^*_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v \tag{23}$$

The constant w^*_α is chosen to make the experimentwise error rate equal to α, that is, w^*_α is chosen so that

$$P_0(|W^*_{uv}| < w^*_\alpha,\ u = 1,\ldots,k-1;\ v = u+1,\ldots,k) = 1-\alpha \tag{24}$$

where the probability $P_0(\cdot)$ is computed under H_0. Selected values of w^*_α can be found in Table A.16 of Hollander and Wolfe (1999). Also see Steel (1960, 1961) and Critchlow and Fligner (1991). A large sample approximation to procedure (23) is

$$\text{decide } \tau_u \neq \tau_v \text{ if } |W^*_{uv}| \geqslant q_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v, \tag{25}$$

where q_α is the upper α^{th} percentile point of the distribution of range of k $N(0, 1)$ random variables. Values of q_α can be found in Table A.17 of Hollander and Wolfe (1999). For the justification of the approximation, see, for example, Hollander and Wolfe (1999 pp. 246–7) and Hochberg and Tamhane (1987). There are also various procedures available to provide simultaneous confidence intervals for contrasts (cf. Hollander and Wolfe (1999, Sect. 6.9) and Hettsmansperger and McKean (1998, Sect. 4.3). Miller (1981) and Hochberg and Tamhane (1987) are books that specialize in multiple comparisons.

Example 2. We illustrate the k-sample procedures of this section with data from [Zelazo et al. (1972)] (see also Brown and Hollander (1977)). The study concerned infants who were randomized to one of four treatments to determine if stimulation of the walking and placing reflexes in the newborn accelerates the time at which infants learn to walk by themselves. The subjects of Table 2 were about one-week-old male infants from middle-class and upper-middle-class families. Each infant was randomly assigned to one of four groups: an experimental group (active exercise) and three control groups (passive exercise, no exercise, eight-week control). We apply the Kruskal–Wallis test to test if there are significant differences among the four groups. The reader will note from the rankings in Table 2 that there are five tied groups of size 2, one tied group of size 3, and one tied group of size 4. Applying the large sample approximation with the correction for ties given by (21), we find $H' = 6.88$. This agrees with the calculation obtained using StatXact, namely $H' = 6.88$ with an approximate P-value of 0.0758 obtained by referring $H' = 6.88$ to a chi-squared distribution with three degrees of freedom. Thus there is evidence that the treatments are not equivalent. We now turn to multiple comparisons to see which treatments differ and which yield the earlier times to walking alone.

We will apply the approximate multiple comparison procedure given by (25). Performing the six two-sample rankings using the data of Table 2, we find from (22), $W^*_{12} = 2.717$, $W^*_{13} = 2.038$, $W^*_{14} = 3.098$, $W^*_{23} = 1.132$, $W^*_{24} = 2.066$ and $W^*_{34} = 1.033$. From the distribution of the range of 4 $N(0, 1)$ random variables, we find $q_{0.10} = 3.240$ and $q_{0.20} = 2.784$. Thus at an approximate experimentwise error rate of 0.10, procedure (25) makes no significant decisions but at an experimentwise error rate of 0.20, we decide, since $|W^*_{14}| = 3.098 > 2.784$, that $\tau_1 \neq \tau_4$ and there is evidence that active exercise leads to earlier walking than is obtained by the eight-week control. We note that the conservative nature of this experimentwise error rate

approach to multiple comparisons (as well as the two-sided procedure used here) makes it difficult to find significant differences at a low experimentwise error rate. If, however, one performs individual one-sided two-sample Wilcoxon rank sum tests of the form given by procedure (4), with sample sizes 6 and 5, as is appropriate for the (1, 4), (2, 4), and (3, 4) comparisons, the $\alpha = 0.063$ critical value for the upper-tail procedure (4) is 39 (see Table A.6, Hollander and Wolfe (1999)) and since $W_{14} = 42$, $W_{24} = 38$, and $W_{34} = 34$, we find that active exercise leads to earlier walking times than the eight-week control. Similarly, for the individual one-sided two-sample tests with sample sizes 6 and 6, as is appropriate for the (1, 2), (1, 3), and (2, 3) comparisons, the $\alpha = 0.066$ critical value for the upper-tail procedure (4) is 49 (see Table A.6, Hollander and Wolfe (1999)), and since $W_{12} = 51$, $W_{13} = 48$, and $W_{23} = 44$, we are able to judge active exercise as leading to earlier walking times than passive exercise.

1.3 The Multivariate Two-sample Location Problem

The Wilcoxon statistic of Sect. 1.1 can be generalized to the multivariate case (cf. Puri and Sen (1971, Chap. 5)). Let

$$\boldsymbol{X}_\alpha^j = (X_{1\alpha}^{(j)},\ldots, X_{p\alpha}^{(j)})', \quad \alpha = 1,\ldots, n_j, \quad j = 1, 2 \tag{26}$$

be independent p-dimensional random vectors with $\boldsymbol{X}_\alpha^j$ having distribution function $F_j(x)$. The hypothesis (1) is generalized to

$$H_0: F_1(\boldsymbol{x}) = F_2(x) \quad \text{for all } \boldsymbol{x}. \tag{27}$$

As with the procedures in Sects.1.1 and 2.2, we will define a test intended to detect location differences of the form

$$F_j(\boldsymbol{x}) = F(\boldsymbol{x}+\boldsymbol{\delta}_j), \quad j = 1, 2 \tag{28}$$

where for model (28), H_0 can be written

$$H_0: \boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \boldsymbol{0}, \tag{29}$$

that is, the random vectors can be viewed as coming from a common p-variate population.

To test H_0 versus the location alternatives $\boldsymbol{\delta}_1 \neq \boldsymbol{\delta}_2$, let W_i be the sum of the second sample's component i ranks in the joint ranking from least to greatest of the $N = n_1+n_2$ values in the pth component of the first and second samples. We let $R_{i1},\ldots, R_{in}$, denote the ranks obtained by the first sample and $R_{i,n_1+1},\ldots, R_{i,n_1+n_2}$ denote the ranks obtained by the second sample in this joint ranking. To compute this multivariate two-sample Wilcoxon-type test, we thus need to perform p two-sample rankings, each of size $N = n_1+n_2$. Let

$$V_i = W_i - n_2(N+1)/2, \quad i = 1,\ldots, p \tag{30}$$

and let $\boldsymbol{V} = (V_1, V_2,\ldots, V_p)'$. The test statistic is

$$V^* = \boldsymbol{V}'\boldsymbol{A}^{-1}\boldsymbol{V} \tag{31}$$

where the $p \times p$ matrix $\boldsymbol{A}$ has elements

$$a_{ii} = n_1 n_2(N+1)/12,$$

$$a_{ij} = [n_1 n_2/\{(N-1)N\}]\cdot\left[\sum_{s=1}^{N} R_{is}R_{js} - \{N(N+1)^2/4\}\right]. \tag{32}$$

The matrix $N^{-3}\boldsymbol{A}$ is a consistent estimator of the asymptotic covariance matrix of $N^{-(3/2)}\boldsymbol{V}$.

An asymptotically distribution-free test of H_0 defined by (27) against the location alternatives $\boldsymbol{\delta}_1 \neq \boldsymbol{\delta}_2$, at the approximate α level, rejects H_0 if $V^* > \chi^2_{p,\alpha}$. This test is not distribution-free for finite samples, but a conditionally distribution-free test based on V^* can be obtained using a permutation principle (Sect. 5.3, Puri and Sen (1971)).

For simplicity we have chosen to describe (and illustrate in Example 3 below) the multivariate two-sample Wilcoxon test. For the k-sample multivariate Wilcoxon test, see Puri and Sen (Chap. 5, 1971). For point estimation of $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ see Sect. 6.3 of Puri and Sen (1971) and for Bonferoni confidence bounds for $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$, see Puri and Sen (1971, pp. 258–9).

Example 3. The taste aversion data in Table 3 are from J. C. Smith and P. Smith (personal communication 1999). Taste aversion learning is a situation in which an animal (or human) drinks a taste solution and then is subjected to some uncomfortable experience (e.g., radiation exposure, injection of some toxic drug such as lithium chloride, exposure to a high intensity magnet, etc.). After the 'pairing' of the taste and the drug, the subject is subsequently tested with the taste solution to see if an aversion has occurred. There has been much interest in how and where this association of taste and drug is made in the brain. It involves: (a) Where is the taste sensation mediated in the brain? (b) Where is the drug sensation mediated in the brain? (c) Where is the association of the taste and drug mediated?

Taste is 'sensed' by taste nerves in the oral cavity and makes the first synapse in the nucleus of the solitary tract (NTS) and the second synapse in the parabrachial nucleus (PNB). R. Norgren believes that the association (point (c) above) is made in the PBN. He has shown that when he destroys the PBN, a rat is unable to learn a conditioned taste aversion. In contrast, the rat can learn an aversion that does not involve taste, but does involve texture, temperature, etc. J. C. Smith and P. Smith can take a rat that normally prefers cold water over warm water, give it a pairing of cold water with lithium chloride and turn it into a warm water 'preferrer.' Norgren sent Smith and Smith some PBN operated rats (along with some sham operated controls) for testing with the warm/cold

Table 3
Total water intake (component ranks in parentheses)

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
PBN	56.22 (10)	53.64 (9)	50.54 (7)	50.66 (7)	47.44 (6)	45.68 (5)
	13.76 (1)	18.84 (1)	18.94 (1)	18.52 (1)	20.28 (1)	21.44 (1)
	25.40 (3)	27.44 (3)	30.20 (3)	31.00 (4)	30.84 (3)	33.52 (3)
	28.22 (4)	40.92 (4)	42.04 (5)	41.64 (5)	41.92 (4)	45.64 (4)
	18.76 (2)	25.18 (2)	22.58 (2)	22.40 (3)	21.72 (2)	26.24 (2)
	57.92 (11)	59.54 (11.5)	58.88 (12)	62.48 (12)	58.62 (12)	59.32 (12)
	65.02 (13)	61.28 (13)	71.48 (13)	65.14 (13)	72.66 (13)	60.06 (13)
Control	51.12 (7)	50.72 (7)	53.20 (8)	52.22 (9)	55.74 (10)	54.58 (7)
	45.94 (6)	44.34 (5)	32.88 (4)	20.22 (2)	45.14 (5)	55.04 (8)
	53.22 (8)	59.54 (11.5)	55.82 (10)	49.80 (6)	52.32 (7)	55.42 (10)
	41.36 (5)	50.44 (6)	49.64 (6)	52.08 (8)	56.00 (11)	53.56 (6)
	53.34 (9)	53.62 (8)	57.24 (11)	53.98 (11)	54.24 (8)	55.12 (9)
	58.02 (12)	55.28 (10)	53.96 (9)	53.12 (10)	55.20 (9)	56.16 (11)
	$W_1 = 47$	$W_2 = 47.5$	$W_3 = 48$	$W_4 = 46$	$W_5 = 50$	$W_6 = 51$

Source: J C Smith and P Smith (private communication 1999).

Table 4
Values of $E(W, t)$ for various distributions

F	Normal	Uniform	Logistic	Double-exponential	Cauchy	Exponential
$E(W, t)$:	.955	1.000	1.097	1.500	∞	3.00

water procedure. Norgren predicted the PBN rat could learn the warm/cold discrimination since there is no reason to think this involves taste. The problem with making lesions in the brain is that most of the brain parts do more than one thing. Before Smith and Smith tested rats for Norgren, they felt it was a good idea to see if the PBN rats eat laboratory food and drink water like normal rats. The data in Table 3, with $n_1 = 7$ PBN-lesion rats and $n_2 = 6$ control rats, come from those observations. They represent a group of rats with PBN lesions and a control group that received surgery but the PBN was spared. Smith and Smith measured water intake over six days. They suspected *a priori* that there might be decrement in water drinking for the PBN rats. The analysis based on the two-sample multivariate Wilcoxon test leads to the conclusion that there are not significant differences between the two groups.

We used an S-plus program to calculate V^* (31) and found $V^* = 5.368$. Referring $V^* = 5.368$ to a chi-squared distribution with four degrees of freedom yields a P-value of 0.498. Thus there is no evidence of differences in total water intake between the PBN and control groups. We note, however, that the test based on V^* is two-sided. Since Smith and Smith suspected *a priori* that the PBN rats might drink less, we should also check the one-sided P-values associated with the one-sided two-sample Wilcoxon test (4) to see if there is evidence that the PBN rats did less drinking. The one-sided P-values corresponding to $W_1 = 47$, $W_2 = 47.5$, $W_3 = 48$, $W_4 = 46$, $W_5 = 50$, and $W_6 = 51$ are (from Table A.6, Hollander and Wolfe (1999)), respectively 0.267, 0.267, 0.223, 0.314, 0.147, 0.117. Thus, again, we do not find significant differences between the two groups. In other words the individual tests are consistent with the overall test based on V^*. The latter, of course, accounts of the dependencies between days.

1.4 Efficiencies of Location Procedures

The location procedures of Sects. 1.1, 1.2, and 1.3 have desirable efficiency properties with respect to their parametric competitors derived under the distributional assumption of normality. For translation alternatives, the Pitman asymptotic relative efficiency (Puri and Sen (1971, Sect. 3.8.1)) of the test based on W with respect to the normal theory two-sample t-test is

$$E(W, t) = 12\sigma_F^2\left(\int f^2(x)\,\mathrm{d}x\right)^2 \tag{33}$$

where σ_F^2 is the variance of the population with distribution F and f is the probability density corresponding to F. Expression (33) is also the asymptotic relative efficiency of the estimator $\hat{\Delta}$ (10) with respect to $\bar{\Delta} = \bar{Y} - \bar{X}$ (Hodges and Lehmann, 1963); the asymptotic relative efficiency of the confidence interval given by (11) relative to the normal theory confidence interval centered at $\bar{\Delta}$ (Lehmann, 1963); the asymp-

totic relative efficiency of the Kruskal–Wallis test relative to the normal theory one-way layout $\mathscr{F}$-test (Andrews, 1954); the asymptotic relative efficiency of the multiple comparison procedure (23) with respect to the corresponding normal theory procedure based on sample means (Sherman, 1965). Hodges and Lehmann (1956) showed that for all populations $E(W, t)$ is at least 0.864. Values of this efficiency expression for various distributions are given in Table 4. The Pitman asymptotic relative efficiency of the test based on V^* (31) with respect to a likelihood ratio criterion test is quite complicated and depends on the univariate marginal distributions and bivariate marginal distributions of F. It is given by Puri and Sen (1971, p. 212).

2. Conclusion

We have highlighted some rank-based procedures in two- and k-sample location problems. These procedures and their generalizations played a key role in the early development of nonparametric statistics. They remain attractive to data analysts today because they enjoy excellent efficiency properties, they are relatively insensitive to outliers, and they are valid under very mild assumptions concerning the underlying populations from which the data are obtained. The subject of nonparametric statistics continues to be vigorously studied. Areas of interest at the time of this writing include Bayesian and computer intensive methods (See *Nonparametric Statistics: Advanced Computational Methods*), regression, and survival analysis. Recent books on nonparametric statistics that would be useful to the interested reader include Spent (1993), Gibbons (1997), Hettmansperger and McKean (1998), and Hollander and Wolfe (1999).

See also: Nonparametric Statistics: Advanced Computational Methods; Nonparametric Statistics: Records; Statistics: The Field

Bibliography

Andrews F C 1954 Asymptotic behavior of rank tests for analysis of variance. *Annals of Mathematical Statistics*. **25**: 724–36

Brown Jr B W, Hollander M 1977 *Statistics: A Biomedical Introduction*. Wiley, New York

Critchlow D E, Fligner M A 1991 On distribution-free multiple comparisons in the one-way analysis of variance. *Communications in Statistics—Theory and Methods* **20**: 127–39

Dwass M 1960 Some k-sample rank-order tests. In: Olkin I, Ghurye S G, Hoeffding W, Madow W G, Mann H B (eds.) *Contributions to Probability and Statistics*. Stanford University Press, Stanford, CA, pp. 198–202

Gibbons J D 1997 *Nonparametric Methods for Quantitative Analysis*, 3rd edn. American Sciences Press, Syracuse, New York

Hettmansperger T P 1984 *Statistical Inference Based on Ranks*. Wiley, New York

Hettmansperger T P 1985 Multivariate location tests. In: Kotz S, Johnson N L, Read C B (eds.) *Encyclopedia of Statistical Sciences*. Wiley, New York, Vol. 6, pp. 79–83

Hettmansperger T P, McKean J W 1998 *Robust Nonparametric Statistical Methods*. Arnold, London

Hochberg Y, Tamhane A C 1987 *Multiple Comparison Procedures*. Wiley, New York

Hodges Jr J L 1967 Efficiency in normal samples and tolerance of extreme values for some estimates of location. In: Le Cam L, Neyman J (eds.) *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, Vol. 1, pp. 163–186

Hodges Jr J L, Lehmann E L 1956 The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*. **27**: 324–35

Hodges Jr J L, Lehmann E L 1963 Estimates of location based on rank tests. *Annals of Mathematical Statistics*. **34**: 598–611

Hollander M, Peña E 1988 Nonparametric tests under restricted treatment-assignment rules. *Journal of the American Statistical Association* **83**: 1144–51

Hollander M, Wolfe D A 1999 *Nonparametric Statistical Methods*, 2nd edn. Wiley, New York

Kruskal W H 1952 A nonparametric test for the several sample problem. *Annals of Mathematical Statistics*. **23**: 525–40

Kruskal W H, Wallis W A 1952 Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47**: 583–621

Lehmann E L 1963 Nonparametric confidence intervals for a shift parameter. *Annals of Mathematical Statistics*. **34**: 1507–12

Lehmann E L 1975 *Nonparametrics: Statistical Methods Based on Ranks*. Holden Day, San Francisco, CA

March G L, John T M, McKeown B A, Sileo L, George J C 1976. The effects of lead poisoning on various plasma constituents in the Canada goose. *Journal of Wildlife Diseases* **12**: 14–9

Mehta C R, Patel N R, Wei L J 1988 Computing exact significance tests with restricted randomization rules. *Biometrika*. **75**: 295–302

Miller Jr R G 1981 *Simulataneous Statistical Inference*, 2nd edn. Springer, New York

Noether G E 1967 *Elements of Nonparametric Statistics*. Wiley, New York

Puri M L, Sen P K 1971 *Nonparametric Methods in Multivariate Analysis*. Wiley, New York

Randles R H, Wolfe D A 1979 *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York

Sherman E 1965 A note on multiple comparisons using rank sums. *Technometrics* **7**: 255–6

Sprent P 1993 *Applied Nonparametric Statistical Methods*, 2nd edn. Chapman and Hall, London

StatXact 1999 Cytel Corporation Cambridge, Massachusetts

Steel R G D 1960 A rank sum test for comparing all pairs of treatments. *Technometrics* **2**: 197–207

Steel R G D 1961 Some rank sum multiple comparisons tests. *Biometrics* **17**: 539–52

Wilcoxon F 1945 Individual comparisons by ranking methods. *Biometrics* **1**: 80–3

Zelazo P R, Zelazo N A, Kolb S 1972 "Walking" in the newborn. *Science* **176**: 314–5

M. Hollander and J. Sethuraman

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Nonparametric Statistics: Records

In a sequence of observations, it is natural to designate as a 'record' any observation which exceeds in value all previous observations in the sequence. Technically this should be called an upper record, to distinguish it from record observations smaller than all preceding observations (lower records). Thus, for example, in Riverside California the (upper) record July 4th temperature was 108°, observed in 1957. The first person to investigate suitable stochastic models for records was Chandler (1952). The model that he proposed can be called the standard record value model, to distinguish it from more elaborate models which have subsequently developed.

1. *The Standard Record Model*

In the standard record value model we envision a sequence $X_1, X_2, \ldots$ of independent and identically distributed (i.i.d.) random variables each having the same continuous distribution, denoted by F. Thus $P(X_i \leqslant x) = F(x)$ for every x. The assumption of continuity of F is convenient since it eliminates the possibility of ties. An observation X_j will be called a record (or an upper record value) if its value exceeds that of all previous observations. By convention the first observation is a record, the zeroth record. A sequence of record times is naturally defined by $T_0 = 1$ and

$$T_i = \min\{j: X_j > X_{T_{i-1}}\}, \quad i = 1, 2, \ldots. \tag{1}$$

The record sequence $\{X^{(n)}\}_{n=0}^{\infty}$ is then defined by

$$X^{(n)} = X_{T_n}, \quad n = 0, 1, 2, \ldots. \tag{2}$$

In order to discuss the stochastic behavior of the record process, it is convenient to introduce several related processes:

(i) the record increment process $\{J_n\}_{n=0}^{\infty}$ defined by

$$\begin{aligned} J_0 &= X^{(0)} \\ J_n &= X^{(n)} - X^{(n-1)}, \quad n = 1, 2, \ldots \end{aligned} \tag{3}$$

(ii) the inter-record time sequence $\{\Delta_n\}_{n=0}^{\infty}$ defined by

$$\begin{aligned} \Delta_0 &= 0 \\ \Delta_n &= T_n - T_{n-1}, \quad n = 1, 2, \ldots \end{aligned} \tag{4}$$

and (iii) the record counting process $\{N_n\}_{n=1}^{\infty}$ defined by

$$N_n = \{\text{number of records among } X_1, X_2, \ldots, X_n\}, \quad n = 1, 2, \ldots \tag{5}$$

The record counting statistics T_n, Δ_n, N_n have distributions which do not depend on the nature of the common continuous distribution of X_i (i.e., F). The distribution of $X^{(n)}$ will depend on F. The determination of the distribution of the record sequence is particularly simple in the case in which X_i has a common standard exponential distribution. Because of the lack of memory property of the exponential distribution, the record-increment process will again consist of i.i.d. standard exponential variables. It is convenient to mark by an asterisk the observations and records from a standard exponential sequence. Thus $\{X_n^*\}_{n=1}^{\infty}$ are i.i.d. standard exponential random variables with a corresponding record sequence $X^{*(0)}, X^{*(1)}, \ldots$. The record increments $J_n^* = X^{*(n)} - X^{*(n-1)}$ are, as remarked, i.i.d. standard exponential variables. It is then trivial to verify that the distribution of the nth record $X^{*(n)}$ will be a gamma distribution, i.e.

$$X^{*(n)} \sim \text{Gamma}(n+1, 1), \quad n = 0, 1, 2, \ldots \tag{6}$$

If X is a random variable with continuous distribution function F it follows that

$$H(X) = -\log(1 - F(X)) \tag{7}$$

has a standard exponential distribution. Consequently

$$X \stackrel{\mathrm{d}}{=} F^{-1}(1 - \mathrm{e}^{-X^*}) \tag{8}$$

where X^* is a standard exponential variable. Here $\stackrel{\mathrm{d}}{=}$ denotes 'has the same distribution as.' In (8) X is represented as a monotone transformation of X^*. It follows that the record sequence associated with i.i.d. X_i can be related to the record sequence associated with a sequence of standard exponential variables. Thus we have

$$X^{(n)} \stackrel{\mathrm{d}}{=} F^{-1}(1 - \mathrm{e}^{-X^{*(n)}}), \quad n = 0, 1, 2, \ldots \tag{9}$$

Using the relationship (9) and the known distribution for $X^{*(n)}$ (given by (6)) we can readily write down an expression for the distribution of $X^{(n)}$, the nth record from an i.i.d. sequence with common distribution function F:

$$F_{X^{(n)}}(x) = 1 - [1 - F(x)] \sum_{k=0}^{n} [-\log(1 - F(x))]^k / k! \tag{10}$$

If F is absolutely continuous with corresponding density f, then $X^{(n)}$ will be absolutely continuous with a density of an attractively simple form (obtained by differention of (10) or by change of variable in the gamma density of $X^{*(n)}$):

$$f_{X^{(n)}}(x) = f(x)[-\log(1 - F(x))]^n / n! \tag{11}$$

The joint density of the records from a standard exponential sequence is readily written down and from it we can, in the absolutely continuous case, obtain a joint density for records corresponding to an i.i.d. sequence with common density f and distribution function F. Thus

$$f_{X^{(0)},X^{(1)},\ldots,X^{(n)}}(x_0, x_1, x_2,\ldots, x_n) = f(x_n)\prod_{i=o}^{n-1} h(x_i)$$
$$-\infty < x_0 < x_1 < \cdots < x_n < \infty \qquad (12)$$

where $h(x) = f(x)/[1-F(x)]$, the failure rate function of the distribution F. We might hope that, for large n, $X^{(n)}$ will be asymptotically normal. It is after all a nice function of a sum of $(n+1)$ i.i.d. random variables (as displayed in (9)). However, the nature of the asymptotic distribution of $X^{(n)}$ depends crucially on the upper tail behavior of the common distribution of X_i, i.e., of F. Three kinds of limit distributions can be encountered.

We introduce the symbol φ_F to denote the inverse of the function H (defined in (7)): thus

$$\varphi_F(u) = F^{-1}(1-\mathrm{e}^{-u}) \qquad (13)$$

If $X^{(n)}$ denotes the nth record in a sequence of i.i.d. X_i with common distribution F then (Resnick 1973) a necessary and sufficient condition for asymptotic normality of $X^{(n)}$ is that for every real x

$$\lim_{s\to\infty} \frac{\varphi_F(s+x\sqrt{s})-\varphi_F(s)}{\varphi_F(s+\sqrt{s})-\varphi_F(s)} = x \qquad (14)$$

If (14) holds, as it does for quite a broad class of distributions F, then the appropriate normalizing constants are simply expressed using the function φ_F. We have

$$\frac{X^{(n)}-\varphi_F(n)}{\varphi_F(n+\sqrt{n})-\varphi_F(n)} \xrightarrow{\mathrm{d}} Z \qquad (15)$$

where $Z \sim N(0, 1)$.

To illustrate this result consider the case in which the common distribution function of X_i is of the Weibull form. Thus

$$F(x) = 1-\mathrm{e}^{x^{\gamma}}, \quad 0 < x < \infty \qquad (16)$$

where $\gamma > 0$. In this case

$$\varphi_F(u) = u^{1/\gamma} \qquad (17)$$

It is not difficult to verify that condition (14) holds and consequently, using (15), we conclude, in this case, that

$$\frac{X^{(n)}-n^{1/\gamma}}{(n+\sqrt{n})^{1/\gamma}} \xrightarrow{\mathrm{d}} Z \qquad (18)$$

where $Z \sim N(0, 1)$.

It is possible to encounter non-normal limit laws for $X^{(n)}$, and it is possible to have cases in which $X^{(n)}$ cannot be normalized to have a non-trivial limiting distribution. See Resnick (1973) for details on this.

It should be remarked that the normalizing constants displayed in (15) are not necessarily the simplest ones available. The denominator in (15) can often be replaced by a simpler asymptotically equivalent quantity. An important special case in which this occurs is the case in which the original X_i are themselves standard normal variables, i.e. when $F = \Phi$ (the standard normal d.f.). In this case probably the simplest normalization for $X^{(n)}$ is as follows:

$$\sqrt{2}[X^{(n)}-\Phi^{-1}(1-\mathrm{e}^{-n})] \xrightarrow{\mathrm{d}} Z \qquad (19)$$

where $Z \sim N(0,1)$.

2. Some Special Cases

Since the nth record, $X^{(n)}$ can be represented as a function of a sum of i.i.d. exponential variables, namely

$$X^{(n)} \stackrel{\mathrm{d}}{=} \varphi_F\left(\sum_{i=0}^{n} X_i^*\right) \qquad (20)$$

it is sometimes possible to encounter cases in which the distribution of $X^{(n)}$ is easily described. This will be true if φ_F is a relatively simple function. Examples include the Weibull case (already introduced) in which $\varphi_F(u) = u^{1/\gamma}$, and the Gumbel case (in which $\varphi_F(u) = \log u$). Pareto records are representable in terms of products of i.i.d. uniform (0, 1) random variables. Thus if X_i has a Pareto (σ, α) distribution, i.e. if

$$F(x) = 1-\left(\frac{x}{\sigma}\right)^{-\alpha}, \quad x \geqslant \sigma \qquad (21)$$

where $\alpha > 0$ and $\sigma > 0$, then

$$X^{(n)} \stackrel{\mathrm{d}}{=} \sigma\left(\prod_{j=1}^{n} U_j\right)^{-1/\alpha} \qquad (22)$$

where U_i are i.i.d. uniform (0, 1) random variables.

3. Record Times and the Number of Records

The times of occurrence of the record values and the record-counting process have distributions which, as remarked on in section 1, do not involve F, provided it is continuous. It is convenient to introduce a sequence of record indicator variables $\{I_n\}_{n=0}^{\infty}$ defined by $I_n = 1$ if and only if X_n is a record value. It turns out that these indicator random variables are independent

random variables though not identically distributed. In fact

$$P(I_n = 1) = 1/n, \quad n = 1, 2, \ldots. \tag{23}$$

Recall that T_i is used to denote the time of occurrence of the ith record ($T_0 = 1$ by convention). The distribution of T_i is expressible in terms of Stirling numbers of the first kind. Thus

$$P(T_i = n) = S^i_{n-1}/n! \tag{24}$$

All the record times $\{T_i\}_{i=1}^{\infty}$ have infinite expectation, even T_1. The implied rarity of records may be grasped by considering the record-counting process $\{N_n\}_{n=1}^{\infty}$ where

$$N_n = \{\text{number of records among } X_1, X_2, \ldots, X_n\}. \tag{25}$$

Clearly

$$N_n = \sum_{i=1}^{n} I_i \tag{26}$$

and the fact that I_i are independent Bernoulli variables allows us to readily compute the moments of N_n. Thus

$$E(N_n) = \sum_{j=1}^{n} \frac{1}{j} \approx \log n + \gamma \tag{27}$$

where $\gamma = 0.5772\ldots$ (Euler's constant) and

$$\operatorname{var}(N_n) = \sum_{j=1}^{n} \frac{1}{j}\left(1 - \frac{1}{j}\right) \approx \log n + \gamma - \frac{\pi^2}{6} \tag{28}$$

Records, in the standard record model, are thus remarkably rare. In a million observations we expect to observe only about 14 records (and one of those is X_1).

The inter-record times, $\Delta_n = (T_n - T_{n-1})$ also have infinite expectations. Indeed much of the wait for the nth record, i.e. T_n, is accounted for by Δ_n, the waiting time for the last one of the n records. Neuts (1967) observed that $\log \Delta_n$ and $\log T_n$ have the same asymptotic distribution. Thus

$$(\log T_n - n)/\sqrt{n} \xrightarrow{\mathrm{d}} Z \tag{29}$$

and

$$(\log \Delta_n - n)/\sqrt{n} \xrightarrow{\mathrm{d}} Z \tag{30}$$

where $Z \sim N(0, 1)$. Using (29), or directly, one can verify that N_n also is asymptotically normal, i.e.

$$(N_n - \log n)/\sqrt{\log n} \xrightarrow{\mathrm{d}} Z \tag{31}$$

where $Z \sim N(0, 1)$.

4. *k-records*

It is sometimes of interest to keep track of other outstanding values in our sequence of observations. To this end we define a sequence of initial ranks, $\{\rho_n\}_{n=1}^{\infty}$.

$$\rho_n = \sum_{j=1}^{n} I(X_n \leqslant X_j) \tag{32}$$

For example $\rho_n = k$ if when X_n is observed it is the kth largest among $X_1, X_2, \ldots, X_n$. We say that X_n is a k-record if $\rho_n = k$. Our usual records are l-records in this framework. With this notation we can envision, for each k, a sequence of k-records, $X^{(n)}_{(k)}$. Remarkably, these sequences of random variables are i.i.d. (a consequence of Ignatov's theorem). Thus $X^{(n)}_{(k)} \overset{\mathrm{d}}{=} X^{(n)}$, for $k = 1, 2, \ldots$.

This result even holds when the common distribution of X_i, F, is not continuous with a suitably modified definition of records.

5. *Characterizations Involving Records*

The distributional properties of the record sequence $\{X^{(n)}\}_{n=0}^{\infty}$ of course are determined by F, the common distribution of X_i. The converse is trivially also true. For if we know the distribution of $X^{(0)}$, we know the common distribution of X_i. It is interesting to inquire whether partial knowledge of the distribution of the record sequence is adequate to determine F. It turns out that just the sequence of expected record times is almost enough. Kirmani and Beg (1984) show that, provided that F is continuous and that the $(1+\delta)$th moment of F exists for some $\delta > 0$, then F is completely determined by the sequence $\{E(X^{(n)})\}_{n=0}^{\infty}$.

The common distribution of X_i, F, can also be determined by the regression of one record on the preceding record. Assuming that F is absolutely continuous we can write

$$\begin{aligned} E(X^{(n+1)} | X^{(n)} = x) &= \int_x^{\infty} y f(y)\,\mathrm{d}y/[1 - F(x)] \\ &= k(x), \quad \text{say} \end{aligned} \tag{33}$$

Note that $k(x) = E(X_1 \mid X_1 > x)$ and it is well known that this conditional mean function uniquely determines the distribution of X_1.

The unusual character of the record increment process when X_i are exponentially distributed can lead to a spectrum of characterizations of the exponential distribution. Recall that if F is exponential, the record increments are themselves i.i.d. exponential variates. The converse is also true but of course much less need be assumed about the record increments. For example it is enough to assume F is continuous and that, for some n, $X^{(n)}$ and $J_{n+1} = X^{(n+1)} - X^{(n)}$ are independent.

6. Statistical Inference Based on Record Data

Suppose that a finite sequence of records $X^{(0)}, X^{(1)},\ldots, X^{(n)}$ is available based on a sequence of i.i.d. random variables with common distribution F. We wish to make inferences about F based upon the observed values of $X^{(0)}, X^{(1)},\ldots, X^{(n)}$.

Because of the previously remarked-on rarity of records, we can expect that the number of observed records, n, will not be large. Consequently we will not be able to avail ourselves of any large sample distribution theory for the records, neither can we hope to successfully employ any non-parametric approach to the estimation of F. There simply will not be enough data. In some specific parametric cases it is possible to develop suitable record-based estimates.

If X_i have common distribution $f(x;\theta)$ where θ is, say, k-dimensional, then in principle we can get a maximum likelihood estimate of θ by maximizing the likelihood of the first n records (as given in (12)). This may prove to be a challenging optimization problem in some cases. In certain cases it is however not difficult. If the common distribution of X_i is a translated exponential density, i.e. if

$$f_{X_1}(x;\theta_1,\theta_2) = \theta_2^{-1}\mathrm{e}^{(x-\theta_1)/\theta_2}I(x \geqslant \theta_1) \tag{34}$$

then the m.l.e.s (maximum likelihood estimates) based on $X^{(0)}, X^{(1)},\ldots, X^{(n)}$ are readily found to be

$$\hat{\theta}_1 = X^{(0)} \tag{35}$$

$$\hat{\theta}_2 = \frac{1}{n+1}[X^{(n)} - X^{(0)}] \tag{36}$$

Since exponential record increments are independent, it is not difficult to determine the joint distribution of these m.l.e.s and to compute means, variances and covariances. If unbiased estimates are desired, suitable candidates are

$$\hat{\theta}_{1,U} = \hat{\theta}_1 - \left(\frac{n+1}{n}\right)\hat{\theta}_2 \tag{37}$$

and

$$\hat{\theta}_{2,U} = \frac{n+1}{n}\hat{\theta}_2 \tag{38}$$

Closed form expressions for the m.l.e.s are also obtainable if the common distribution of X_i is uniform. In most other cases (normal, logistic, etc.), iterative solution is required to obtain the m.l.e.s.

If the common distribution of X_i is assumed to be a location-scale family, i.e. if

$$X_i = \mu + \sigma Z_i \tag{39}$$

where Z_i has a known distribution, then best linear unbiased estimates of μ and σ can be obtained provided that we are able to obtain expressions for the means, variances and covariances of the records associated with a sequence of i.i.d. Z_i. The translated exponential case is particularly easy to deal with in this context, and it is not difficult to verify that the estimates displayed in equations (37) and (38) are actually the best linear unbiased estimates. In some cases it is also possible to construct good point and interval predictors for future records. Again the translated exponential model is most tractable for such analyses.

Records of course will not be rare if X_i increase stochastically as i increases. Foster and Stuart (1954) recognized that record-counting statistics could be used to test for trends in a sequence of observations. They developed tests based on the number of lower and upper records in a sequence.

7. Alternative Record Models

In the standard record model, records are rare. In the Olympic games, records are broken with high frequency. Clearly, the standard record model is not appropriate for analyzing Olympic data. The population of Olympic athletes is growing and improving, so any attempt at modeling using i.i.d. X_i is bound to be inappropriate. Two simple modifications have been proposed to provide models in which records are more abundant. The first is the so-called F^α record model. In it we assume that X_i has distribution $F^{\alpha(i)}$ for some non-decreasing sequence $\{\alpha(i)\}_{i=1}^{\infty}$. In this model X_i is distributed like the maximum of $\alpha(i)$ i.i.d. random variables, each having common distribution F. Since $\alpha(i)$ is increasing as i increases, it follows that X_i are stochastically increasing and records will be encountered more frequently, as desired. This model was introduced by Yang (1975) who assumed that $\alpha(i)$ grew geometrically with i. Incidently, even this model seems to fail to fit Olympic data. A second model with potential for generating more records than the standard model is the linear drift model. In this model we assume that X_i are independent random variables but they are not identically distributed. Instead we assume that

$$X_i \stackrel{\mathrm{d}}{=} Z_i + ic \tag{40}$$

where $c > 0$ and Z_i are i.i.d. with common distribution function F. Balerini and Resnick (1995) have investigated properties of this record model. Other models have been proposed under which the distribution of X_i changes whenever a record is observed. This too can lead to an enhanced frequency of records.

In many applications it is actually unrealistic to assume that X_i arrive at equally spread intervals of time, as is tacitly assumed in the standard record model. It is more realistic to envision that arrival times

are governed by some stochastic process called the pacing process. Both homogenous and nonhomogenous Poisson processes have been considered as pacing processes. Interest is frequently focused on record arrival times in such paced processes.

8. Multivariate Records

If the individual observations, X, are k-dimensional it is not obvious how one should define a record. Perhaps the most compelling definition is that we declare an observation X_n to be a record if every coordinate of X_n exceeds in value the corresponding coordinate of every preceeding X_i. With this definition records will be extremely rare. For example if $k \geqslant 2$ and if the coordinates of X are independent, we can expect to see fewer than two records in an infinitely extended sequence of trials. Remember that X_1 is always a record, so less than one non-trivial record can be expected. With positive probability, the first record will stand for all time! It turns out that the number of records will be finite with probability one if and only if the coordinates of X are asymptotically independent. Several alternative definitions of multivariate records are described in Arnold, Balakrishnan and Nagaraja (1998). An extension to the case in which X_i take on values in an arbitrary partially ordered set is described in Goldie and Resnick (1989). In some applications interest may center in the coverage of $X_1, \ldots, X_n$ where X_i are i.i.d. k-dimensional random variables. In this case we may be interested in the growth of the k-dimensional volume of the convex hull of $X_1, X_2, \ldots, X_n$ as n increases. This can be viewed as the study of a sequence of record volumes, but the successive volumes are not independent and the study of such problems falls outside the scope of the usual record value models.

9. Further Reading

More details may be found in two complementary volumes on the topic of record values: Ahsanullah (1995) and Arnold, Balakrishnan and Nagaraja (1998). An excellent source for discussion of the asymptotic distribution theory of records is Resnick (1987). Most books on order statistics will include some discussion of record values.

Bibliography

Ahsanullah M 1995 *Record Statistics*. Nova Science, New York

Arnold B C, Balakrishnan N, Nagaraja H N 1998 *Records*. Wiley, New York

Ballerini R, Resnick S I 1995 Records from improving populations. *Journal of Applied Probability* **22**: 487–502

Chandler K N 1952 The distribution and frequency of record values. *Journal of the Royal Statistical Society, Series B* **14**: 220–8

Foster F G, Stuart A 1954 Distribution-free tests in time-series based on the breaking of records (with discussion). *Journal of the Royal Statistical Society, Series B* **16**: 2–22

Goldie C M, Resnick S I 1989 Records in partially ordered sets. *Annals of Probability* **17**: 678–89

Kirmani S N U A, Beg M I 1984 On characterization of distributions by expected records. *Sankhya, Series A* **46**: 463–5

Neuts M 1967 Waiting times between record observations. *Journal of Applied Probability* **4**: 206–8

Resnick S I 1973 Limit laws for record values. *Stochastic Processes and their Applications* **1**: 67–82

Resnick S 1987 *Extreme Values, Regular Variation and Point Processes*. Springer, New York

Yang M C K 1975 On the distribution of the inter-record times in an increasing population. *Journal of Applied Probability* **12**: 148–54

B. C. Arnold

Nonpartisanship: Political

Nonpartisanship is the notion that governmental decision making and even the selection of public officials should properly be based upon neutral principles of merit rather than the partisan interests of clashing forces.

1. Origins of Political Parties

Political parties developed in the United States at the end of the eighteenth century and, with the advent of democratic systems of governance, in Western Europe during the mid-1900s. Parties linked office holders in national capitals with political organizations in the countryside and, through these organizations, with popular electorates. By the end of the nineteenth century, well-organized party 'machines' dominated politics and government in the US, while in Europe socialist and labor parties, in particular, created cohesive networks through which to mobilize their supporters for political combat.

Through their competition for votes, parties helped bring about full electoral mobilization and the advent of popular democracy in the West. This process is described by what might be called the 'neo-classical model' of party competition. As developed by political scientists V. O. Key and E. E. Schattschneider, the neo-classical theory asserts that in a democratic polity, high levels of elite conflict will inevitably lead to increased rates of mass participation as contending forces organize and engage in competitive efforts to mobilize political support. V. O. Key credited Thomas Jefferson with setting the stage for large-scale popular mobilization in the US when he built local party organizations and 'lined up the unwashed' to defeat his Federalist foes (Key 1942). The Federalists fol-

lowed suit, albeit reluctantly, and built party machines to mobilize the 'unwashed' for their cause. Writing during the 1950s, Schattschneider referred to this phenomenon as 'expanding the scope of the conflict' and claimed that it was a central feature of democratic political processes. Schattschneider argued that popular mobilization was most likely to be initiated by the losers in inter-elite struggles who hoped to change the outcome by enlarging the universe of participants (Schattschneider 1960).

2. *Opposition to Parties*

Opposition to allowing political parties to play a major role in the political process has been an important element in the political ideologies of two types of political forces. The first of these, which political scientist Shefter calls an absolutist coalition, denies both the validity of popular representation in governmental decision making and the legitimacy of differences in political opinion, and looks to a leader or ruling elite to identify and implement the most appropriate and effective programs and policies for the entire nation (Shefter 1994). In the nineteenth century, absolutist coalitions were formed by traditional elites who asserted that the interests of the nation would best be served by a crown that stood above factional and partisan squabbling. In the twentieth century, similar arguments were made by military juntas and Fascists who propounded the idea of a government that would speak for the nation-as-a-whole, as well as Leninists who asserted that only a revolutionary vanguard divorced from partisan concerns possessed the level of consciousness needed to bring about a just society.

A second type of political coalition often identified with a doctrine of non-partisanship consists of rationalizing upper-middle-class reformers and their allies. Shefter identifies the ideological bases of this coalition as Benthamite in England, Positivist on the continent, and Progressive in the US (Shefter 1994). Unlike their absolutist counterparts, upper-middle-class reformers neither reject political deliberation nor object to the representation of a variety of interests in governmental decision making. However, generally they have contended that an objective and neutral 'national interest' should take precedence over partisan concerns. Reformers have seen this national interest, moreover, as most likely to be discovered by experts applying scientific criteria and judgments to resolve national problems. This view helped justify the enormous autonomy granted to the civil service in both Germany and Britain in the nineteenth and twentieth centuries. In similar vein, turn of the twentieth century American Progressives sought to build an autonomous and powerful national bureaucracy while weakening political parties and elected institutions. To this end, Progressives fought for such measures as civil service reform, non-partisan electoral processes, and non-partisan local government. During the 1960s and 1970s, upper-middle-class reformers continued the American Progressive tradition by organizing a congeries of groups that collectively came to be identified as the public interest or 'new politics' movement (Vogel 1989). Among other things, new politics reformers undertook to shift the locus of a great deal of decision-making power to the federal courts on the grounds that they were less partisan than the Congress and the presidency and, hence, better able to identify and serve objective national interests (Neely 1981). The Progressive view remains quite evident in the US today in the form of demands for non-partisan mechanisms of decision making. For example, a recent conference cosponsored by the Council for Excellence in Government and the prestigious Committee for Economic Development recommended the creation of a nonpartisan, federally chartered 'Center' that would offer opinions on all proposed congressional enactments prior to further congressional action. Its proponents assert that such a Center would ensure that new programs would serve national rather than mere partisan purposes (Council for Excellence 2000).

3. *Nonpartisanship and Democratic Politics*

A number of academic critics have argued that strong parties are a necessary condition for mass democracy and that, conversely, nonpartisanship impedes democratic governance. As the late V. O. Key pointed out long ago, the persistence over time between groups possessing a measure of identity and continuity is virtually a necessary condition for popular influence (Key 1949). Party identity increases the public's capacity to recognize its options. Continuity of party division facilitates organization of the electorate on the long-term basis necessary to sustain any serious popular influence in the governmental process. Organized party efforts promote high levels of popular political involvement. In the absence of parties, says Key, there exists little basis for 'effectuation of the popular will' (Key 1949). The weakness of party organizations in the US is one reason why American voter turnout has dropped to historically low levels (Burnham 1970, Aldrich 1995).

4. *Nonpartisanship and Class Politics*

Equally important are the class implications of partisan versus nonpartisan modes of political enterprise. Though presented by its advocates as an effort to develop impartial political principles, nonpartisanship, itself, is not a neutral doctrine. It has a distinct class bias. Party organization is generally an essential ingredient for effective political participation and

competition by groups lacking substantial economic or institutional resources. Thus, party building typically has been the strategy pursued by groups that must organize the collective energies of large numbers of individuals to counter their opponents' superior material means or institutional standing. In the course of European political history, for example, disciplined and coherent party organizations generally were developed first by political forces representing the aspirations of the working classes. Parties, French political scientist Duverger notes, 'are always more developed on the Left than on the Right because they are always more necessary on the Left than on the Right' (Duverger 1963). Thus, in the US, the Jeffersonians built the first mass party as a counterweight to the superior institutional and economic resources that could be deployed by the incumbent Federalists. In a subsequent period of American history, the Jacksonians sought to construct a coherent party organization to compete against the superior resources that could be brought to bear by their adversaries. This effort led to the Jacksonian era's famous controversies over the 'spoils system' for appointing loyal party workers to governmental posts.

In both the US and Europe, the political success of party organizations forced their opponents to copy them in order to meet the challenge. It was, as Duverger points out, 'contagion from the left,' that impelled politicians on the right to attempt to build party organizations (Duverger 1963). But, at the same time, groups with the means to acquire and hold political power without much in the way of party organization often sought to discourage the development of an organized partisan politics, preferring nonpartisan modes of decision making. As Huntington has observed, upper-class groups generally have opposed parties because they saw them as threats to the established social and political order (Huntington 1968). George Washington's warning against 'the baneful effects of the spirit of party' came to be echoed by representatives of many social, economic, and military elites who saw their right to rule challenged by groups able to organize the collective energies and resources of ordinary citizens.

This type of opposition to party was the basis for a number of the institutional reforms of the American political process promulgated during the Progressive era. Many Progressive reformers were undoubtedly motivated by a sincere desire to rid politics of corruption and to improve the quality and efficiency of governmental services in the US. But, simultaneously, from the perspective of middle- and upper-class Progressives and the financial, commercial, and industrial elites with whom they were often associated, the weakening or elimination of party organization would also mean that power would more easily be acquired and wielded by those with wealth, power, and education. The list of antiparty reforms of the Progressive era is a familiar one. The 'Australian' or official ballot reform took away the parties' traditional privilege of printing and distributing ballots and encouraged split-ticket voting. The introduction of nonpartisan local elections eroded grass-roots party organization. The extension of merit systems for administrative appointments stripped party organizations of much of their access to patronage and, so, reduced their ability to recruit workers. The introduction of the direct primary diminished party leaders' control over candidate nominations. These reforms by no means destroyed party organizations. They did, however, diminish the political vitality of American political parties and contributed to a decline in American voter turnout at the turn of the century.

In a similar vein, the upper-middle-class forces that organized the new politics movement of the 1970s hoped to undermine the working class political machines and trade unions that depended upon partisan politics and to exercise political influence through nonpartisan 'public interest' groups. During the late 1960s and early 1970s, hundreds of interest groups and public interest law firms were established by upper-middle class activists to further such goals as environmental quality, the elimination of nuclear weapons, consumer safety, and women's rights (Vogel 1989). The leaders of these groups, such as Common Cause, the Sierra Club, the National Organization for Women, and the various organizations formed by consumer activist Ralph Nader, sought to distinguish them from traditional interest groups by claiming to serve broad public interests rather than the more narrow interests of corporate patrons generally served by conventional interest groups and political parties (Berry 1999). The creation of public interest groups was promoted by major liberal funding agencies such as the Ford Foundation which provided tens of millions of dollars through its 'Fund for the Republic.' In addition, the elite media generally supported these groups and seldom questioned their claim to speak for some neutral and nonpartisan public interest.

5. *Nonpartisanship and the Public Interest*

Thus, nonpartisanship is often equated with political neutrality and the search for an objective public interest. A politics without parties, however, is profoundly biased. Essentially, the upper classes do not need party organization to participate effectively in political life while those further down the social hierarchy do require such organization to play a significant role. A nonpartisan politics is not likely to be a very democratic politics.

See also: Party Responsibility; Political Machines; Public Interest

Bibliography

Aldrich J H 1995 *Why Parties?* University of Chicago Press, Chicago
Berry J M 1999 *The New Liberalism: The Rising Power of Citizen Groups*. Brookings, Washington, DC
Burnham W D 1970 *Critical Elections and the Mainsprings of American 1st edn. Politics*, Norton, New York
Duverger M 1963 *Political Parties*. Wiley, New York
Council for Excellence 2000 *Excellence in Government Performance: Where Are We Headed?* Council for Excellence in Government, Washington, DC
Huntington S P 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, CT
Key V O 1942 *Politics, Parties and Pressure Groups*. Crowell, New York
Key V O Jr 1949 *Southern Politics in State and Nation*, New York
Neely R 1981 *How Courts Govern America*. Yale University Press, New Haven, CT
Schattschneider E E 1960 *The Semi Sovereign People*. Holt, Rinehart and Winston, New York
Shefter M 1994 *Political Parties and the State*. Princeton University Press, Princeton, NJ
Vogel D 1989 *Fluctuating Fortunes*. Basic Books, New York

B. Ginsberg

Nonprofit Sector, Geography of

Nonprofit organizations exist worldwide and their geographic distribution has become a recent focus of investigation. Research has shown that the scope and dimensions of nonprofit sectors vary within and between countries, regions, and cities. To explain these sometimes complex patterns, researchers have considered economic, social, and political variables.

1. A Recent Field of Study

While philanthropy and nonprofit organizations have long been considered in a number of social science disciplines, the nonprofit sector as a distinct field of study is a fairly recent phenomenon. Sparked in the early 1970s by a growing recognition of the importance of the nonprofit sector and debates about the future of the welfare state, interest in nonprofit organizations has grown rapidly in a number of disciplines, particularly sociology, history, economics, political science, public policy, and public administration. While much of the early research was done in the United States, the study of nonprofit sectors has recently also grown rapidly in other countries around the world.

Nonprofit researchers are interested in defining the nonprofit sector, delineating its scope, and understanding the factors determining its existence and activities. This has led researchers to a set of interrelated questions about the sector. Why are there nonprofit organizations? What kinds of activities are they involved in? Where are they located?

While definitions vary, it is generally agreed that the nonprofit sector is a set of social entities that are: formally organized, private (nongovernmental), nonprofit distributing, self-governing, and voluntary to some degree (Salamon and Anheier 1998). Nonprofits have generally been cast alongside for-profit organizations and government agencies as one societal option for providing material and nonmaterial goods and services in models of three-sector economies (Weisbrod 1988). The organizations in each sector are economic entities, but the rationale for their existence varies between sectors. For-profits are oriented toward the pursuit of self-interest, government organizations toward the provision of public goods using coercive authority if necessary, and nonprofits toward the mobilization of public spirit and altruism to create social and civic benefits.

2. Geographic Variation Abounds

The extent of the nonprofit sector has been shown to vary at a number of geographical levels, including: nations, regions within nations (such as states), cities, and neighborhoods. Variation can be in terms of the numbers of organizations, their activities, and their organizational characteristics.

To explain these variations, researchers have focused on factors ranging from the general to the particular. General factors include social, cultural, economic, and political realms. Much of the research has considered more specific factors, usually related to supply and demand. Supply-side factors include public and private money, labor, volunteers, and entrepreneurs. The demand side has focused on the desire or need for different types of goods or services, which is usually seen as dependent upon individual or community characteristics. Supply and demand have also been linked to general factors, for example demands due to a history of discrimination, lack of access to alternate services as a result of public policy, or demands due to the esthetic desires of a community.

Given that the study of nonprofit geography is a new research agenda carried out by investigators in a number of different disciplines, there is as yet no definitive set of findings and conclusions. We will, therefore, outline the various approaches that are being pursued, highlight representative findings, and speculate on the future.

2.1 Cross-national Variation

An early model (Smith 1973) linked formal voluntary organizations to a number of general conditions in a society that were related, in turn, to societal modernization. In this model, nonprofit prevalence is

positively related to the frequency, extent, and permanence of communication among the members of a society; the number of different goals available to them; the degree to which collective action is permissible in the society; and the degree to which there are resources and payoffs for collective action.

A variety of additional explanations have also been proposed to account for cross-national variation in the definition, prevalence, and role of the nonprofit form. One variable frequently cited is the increased demand for different types of public goods generated in more religious, ethnic, or ideologically heterogeneous populations (Weisbrod 1988). Democratic government, with its mandate to serve the majority, will not satisfy this 'heterogeneous demand.' For-profits, likewise, will not provide these public goods. It is, consequently, left to nonprofits to supply them.

In addition, a number of country-specific characteristics have been considered by researchers. These include religious, ideological, and political traditions; historical contingencies; and the institutional models that result from the interaction of these factors. An example would be historically based variations in the definitions of what is public and what is private and the state-private sector relations that result (including relations between the organizations in each sector).

An ambitious project is currently underway to assess the nonprofit sector worldwide more systematically. This project includes 22 developed and developing countries. Overall, the sector has been found to be extensive and growing. Descriptive analyses of the data acquired to date show that national nonprofit sectors vary in terms of the numbers of nonprofit organizations, the services they provide, their revenues, and their employment (Salamon et al. 1999). A variety of factors to explain this variation have been proposed and tested. In terms of general trends, the rise of a third sector on the global level has been linked to the substitution of private service provision for welfare state or socialistic government services, increased activity in the areas of development and environmentalism, the rise in middle-class expectations, and improvements in communication technology.

A more extensive analysis of this data (Salamon and Anheier 1998) used seven countries to test major theories of the nonprofit sector. Explanatory variables included: societal heterogeneity (as outlined above), the supply of social entrepreneurs, a lack of trust in the for-profit market sector, a relatively weak welfare state which has not displaced a previously existing nonprofit sector, and a relatively large welfare state which uses existing and new nonprofits to deliver government-funded services. In addition to these, a more integrated 'social origins' theory was also included. In this theory, choices about whether to rely on market, for-profit or government provision of goods and services are determined by patterns of historical development and broader social, political, and economic relationships. Of particular interest are class relationships and patterns of state-society relations. While partial support was found for a number of factors, social origins theory was found to provide the greatest overall explanation, generating four models of nonprofit sector development.

2.2 Regional and State Variation

Research on nonprofit geography within nations has been carried out primarily in the United States. Most of this work has involved case studies or comparisons among a small number of areas. Only a few studies have focused on regions or states, undoubtedly because of the difficulty of acquiring data on these areas. For example, the Internal Revenue Service (IRS) and the National Center for Charitable Statistics at the Urban Institute in Washington, DC provide information on nonprofit organizations in the United States. However, only certain data are collected, and these only from larger nonprofits, which are required to provide information to the IRS. It has been estimated that only about a third of the total number of nonprofits provide such information. Given these limitations, researchers are often required to collect their own data, a significant undertaking for areas as large as states and regions.

In terms of regional variation in the United States, a few descriptive studies have been done. These typically find that the largest number of nonprofits and the greatest nonprofit expenditures are in the northeast region and the lowest in the south. At the level of the state, a number of studies are currently underway to determine the scope of the nonprofit sectors in particular states. In addition, a comparative study of state nonprofit incorporations (Abzug and Turnheim 1998) included a number of demand variables (state fiscal health, crime, poverty, unemployment, and heterogeneity) as well as supply variables (the state legal environment and the legitimacy of the nonprofit form). While a number of these factors were significant when considered separately, when taken together, the legitimacy of the nonprofit form had the most explanatory power.

2.3 City and Metropolitan Variation

Numerous studies in the United States have examined nonprofit distributions and spatial patterns both within and between cities or metropolitan regions. Prominent within-city research includes the work of McPherson and his colleagues (McPherson and Rotolo 1996), who used ecological models of organizations to study nonprofits in a Midwest community. Nonprofits located where potential members with particular demographic characteristics (potential organizational resources) were found within the

metropolitan region. In addition, Wolch and Geiger (1983) studied the distribution of social welfare and community service nonprofits across municipalities in Los Angeles County and found it to be positively related to community needs and resources. Wolch has also analyzed the geography of children's services in Los Angeles County, and found significant variation and gaps in coverage due to zoning, land prices, and neighborhood characteristics.

In terms of cross-city research, a major descriptive study by the Urban Institute in the early 1980s gathered and compared data on the nonprofit sectors in 12 sites across the United States (Salamon 1987). The research showed significant variation in metropolitan nonprofit sectors. In a more detailed study, Wolch (1990) identified voluntary—rich and voluntary—poor metropolitan areas among 15 United States cities. A significant trend revealed in the data was that welfare state reorganization, including privatization, stimulated the growth of voluntary sectors. This effect, though, was shaped by local economic and demographic factors.

The most extensive research on metropolitan nonprofit sectors has been carried out by Julian Wolpert. Beginning with studies of Philadelphia and New York his research expanded to compare regions and has most recently been focused on 85 cities across the United States (see for example Wolpert 1993). Regional differences in nonprofit sectors have been linked to regional economic cycles. Metropolitan differences have been related to community variation in generosity (measured by both private giving and public spending), with higher rates of generosity being positively related to income and negatively related to population and distress levels. In addition, nonprofits were more prevalent and more targeted to human services in center cities than in suburbs.

Political culture has been related to nonprofit presence in a study of two metropolitan areas (Bielefeld and Corbin 1996). The research used a classification of political culture based on citizen attitudes about the proper role of government. It showed that the nonprofit sector was larger in a political culture where citizens supported government redistribution of wealth as opposed to a political culture that favored a government role restricted to promoting the market or reinforcing the status quo in the community. This was also found in a study of nine metropolitan regions (Bielefeld 2000). This study also supported Wolpert's conclusions about generosity.

2.4 Neighborhood Variation

Relatively few systematic spatial studies of nonprofit organizations have been carried out at the neighborhood level. A study of nonprofits across Dallas, Texas (Bielefeld et al. 1997) found that nonprofit organizations were more strongly influenced by demographic factors (representing resources and need) in their immediate areas, as opposed to the demographic characters of areas further removed from them. These results indicate that the nonprofit sector is closely tied to the dynamics of local communities. Given these findings, the lack of research at this level may be a serious oversight.

Most other neighborhood studies have been more descriptive and have used geographical information systems (GIS) techniques to relate providers and needs spatially, within or across neighborhoods in metropolitan areas. In a typical example, a project by the Center on Nonprofits and Philanthropy at the Urban Institute used a spatial analysis to examine the relationships between the nonprofits that serve families and children and the socioeconomic characteristics of three low-income neighborhoods in the District of Columbia. The purpose was to show how neighborhood environments, including the presence of nonprofit organizations, affected the well being of community residents (DeVita et al. 2000).

3. The Need for Convergence

With few exceptions, geographers have not carried out the research described above. In addition, the results have not found their way into the mainstream of that discipline. Consequently, the geographic study of nonprofit sectors is in its infancy and much remains to be explained. Closer ties between the nonprofit research community and geographers are likely to prove beneficial to both camps. Nonprofit research has clearly shown that significant geographical variation exists in the nonprofit sector and has suggested a number of important variables to account for it. This should convince geographers that the nonprofit sector is a relevant area of inquiry. On the other hand, geography, with its conceptualizations and analyses of location, place, relationships, movement, and region, clearly has much to offer nonprofit researchers in terms of methodology, variables, and theories. This is especially true for economic and cultural geography, which considers many of the same kinds of issues which interest the nonprofit research community (see for example Lee and Wills 1997). It is because of this convergence of interests that a more well-developed geography of the nonprofit sector should be pursued.

Bibliography

Abzug R, Turnheim J 1998 Bandwagon or band-aid? A model of nonprofit incorporation by state. *Nonprofit and Voluntary Sector Quarterly* **27**: 300–22

Bielefeld W 2000 Metropolitan nonprofit sectors: Findings from NCCS data. *Nonprofit and Voluntary Sector Quarterly* **29**: 297–314

Bielefeld W, Corbin J 1996 The institutionalization of nonprofit human service delivery: The role of political culture. *Administration and Society* **28**: 362–89

Bielefeld W, Murdoch J C, Waddell P 1997 The influence of demographics and distance on nonprofit location. *Nonprofit and Voluntary Sector Quarterly* **26**: 207–25
De Vita C J, Manjarrez C, Twombly E C 2000 *Organizations and Neighborhood Networks that Strengthen Families in the District of Columbia: Final Report to the Ammie E. Cassey Foundation*. The Urban Institute, Washington, DC
Lee R, Wills J 1997 *Geographies of Economies*. Arnold, London
McPherson J M, Rotolo T 1996 Testing a dynamic model of social composition: Diversity and change in voluntary groups. *American Sociological Review* **61**: 179–202
Salamon L M 1987 Partner in public service: The scope and theory of government-nonprofit relations. In: Powell W W (ed.) *The Nonprofit Sector: A Research Handbook*. Yale University Press, New Haven, CT
Salamon L M, Anheier H K 1998 Social origins of civil society: Explaining the nonprofit sector cross-nationally. *VOLUNTAS* **9**: 213–48
Salamon L M, Anheier H K et al. 1999 *The Emerging Sector Revisited: A Summary*. Institute for Policy Studies, Johns Hopkins University, Baltimore, MD
Smith D H 1973 Modernization and the emergence of voluntary associations. In: Smith D H (ed.) *Voluntary Action Research*, D. C. Heath & Co, Lexington, MA
Weisbrod B A 1988 *The Nonprofit Economy*. Harvard University Press, Cambridge, MA
Wolch J 1990 *The Shadow State*. The Foundation Center, New York
Wolch J, Geiger R K 1983 The urban distribution of voluntary resources: An exploratory analysis. *Environment and Planning, A* **15**: 1067–82
Wolpert J 1993 Decentralization and equity in public and nonprofit sectors. *Nonprofit and Voluntary Sector Quarterly* **22**: 281–96

W. Bielefeld

Nonsampling Errors

Data from sample surveys are subject to two types of errors: sampling and nonsampling errors. Sampling errors result from the fact that we only have information from a portion of the population of interest rather than from the whole population. Nonsampling errors include errors from all other sources: the sampling frame, the questionnaire, the interviewer, and the respondent. Data from both censuses and surveys may suffer from nonsampling errors while only survey data is subject to sampling error.

Statisticians have traditionally been most concerned with controlling and measuring sampling errors, perhaps because they are easier to handle. Sampling errors are controlled by choice of the survey design; the size of the sampling error can be estimated using the survey data itself. The emphasis of most classical survey sampling books such as Cochran (1977) and Kish (1965) is on measuring sampling error and designing surveys to reduce that error.

More recently, authors such as Lessler and Kalsbeek (1992) and Biemer et al. (1991) have focused on issues related to nonsampling error. Such errors must typically be assessed using information outside of the survey (or census) data or using a statistical model. As is noted by Mosteller (1978), sampling errors typically contribute to the variance of an estimate from survey data (and do not exist for census data), while nonsampling errors can contribute to both variance and bias of an estimate. In collecting survey or census data we try to control, reduce, or avoid nonsampling errors as much as possible because their size and effect on estimates are very hard to assess.

Frame errors, response errors, nonresponse errors, and other sources of errors are all nonsampling errors. The following describes these nonsampling errors and methods for dealing with them.

1. Frame Error

1.1 Definition of Frame Error

The sampling frame is the list from which units are drawn for the sample. The 'list' may be an actual listing of units, as in a phone book from which phone numbers will be sampled, or some other description of the population, such as a map from which areas will be sampled. Frame error results when the sampling frame is not an accurate and complete representation of the population of interest.

Frame errors arise from erroneous inclusions, erroneous exclusions, and multiple inclusions. Erroneous inclusions occur if the frame includes units not in the population of interest so that invalid units have a chance of being included in the sample. Erroneous exclusions occur when members of the population of interest are missing from the list so that they have no chance of being sampled. Multiple inclusions occur if some population members appear multiple times on the list so that they have a higher chance of being sampled.

For an easy example of these types of frame errors, consider sampling households within a city using the telephone white pages as the sampling frame. Some numbers listed in the phone book are no longer working numbers because the people in the household that had the number have moved out of the city; these are erroneous inclusions. Some households choose to have their phone numbers unlisted; these are erroneous exclusions. Some households have multiple listings in the phone book because they have additional telephone lines for a computer modem or a fax machine or for the children; these are the multiple inclusions.

Frame error is a problem because a sample drawn from such a frame may not represent the population of interest, from which you think you are sampling, and some units may have a higher probability of being included in the sample than most units. These errors may result in sample estimates that are biased for the population quantity being estimated. For example, if

wealthier households are more likely to have multiple phone lines, then those households are more likely to be included in the sample. This could lead to estimates of, say, amount spent on entertainment that are biased high.

1.2 Dealing with Frame Error

As is the case with all nonsampling errors, the best way to control frame error is to avoid or minimize it. Methods for doing this depend on the type of sampling frame that you are using. We will discuss two commonly used methods in this section: random digit dialing (RDD) and dual-frame sampling.

1.2.1 Random digit dialing. For telephone surveys of individuals in households, RDD is often used to avoid problems associated with using a telephone directory as the sampling frame. To create the sampling frame for an RDD sample, one must first list all working area codes in the area to be sampled. Within each area code, one must list all working exchanges. If all possible four-digit numbers that could complete the phone number were used to make up the sampling frame of phone numbers, there would be numerous nonworking numbers and business numbers included. Because calling such numbers to determine that they are not valid household numbers is time-consuming and expensive, other methods may be used to reduce the number of invalid phone numbers in the frame. One method takes advantage of the fact that phone numbers are typically given out in banks of 100. Thus, the frame is taken to be the list of area codes with their exchanges followed by all possible first two digits of the remainder of the number. These eight-digit numbers are clusters of 100 possible phone numbers. A random sample of these clusters is taken and then two two-digit numbers are chosen at random to give two potential phone numbers within each cluster. The two sampled numbers within each cluster are called. If both are nonworking numbers or business numbers, then the cluster is discarded because it is likely that the remaining possible numbers in that cluster are also invalid numbers.

Once a valid household is reached, one still must determine who is to be interviewed within the household. If a single adult respondent is to be selected, a method proposed by Kish (1949) may be used to randomly select that adult. Because Kish's method involves creating a roster of all individuals within the household, which can be time-consuming and may be considered intrusive by potential respondents, other methods such as asking for the person who had the last birthday may be used as a quick method of selecting an individual within a household (see Lavrakas 1993, pp. 111–13 for a description of the last birthday and next birthday methods).

There are, of course, some problems with RDD samples. For example, persons who live in households with no telephone are not included in the sampling frame. Households with multiple telephone lines have a higher probability of being selected for the sample. One must account for this by asking respondents how many phone lines they have that are used for voice correspondence and use this information to adjust the selection probability for that household. If one is searching for a very specific population, such as households with disabled persons or early voters, then many of the households reached by telephone will be ineligible for the survey and must be screened out of the sample. The higher the proportion of households that must be screened out because of ineligibility, the higher the final cost of the survey will be.

1.2.2 Dual frame sampling. In many sampling settings there are multiple lists or sources that can be used to develop a sampling frame. For example, the US Department of Agriculture uses lists of farms and area maps to select samples for their crop surveys. Samples from the lists of farms are easier to obtain but miss some farms; the area samples are difficult to handle but encompass all possible farms (see Bosecker and Ford 1976, for details.) A survey of patrons of the arts may use lists of season ticket holders to the symphony, ballet, and opera along with lists of persons who have made contributions to art museums. The challenge when multiple lists are used is to unduplicate the lists or determine how many ways a person could be selected for the sample. Unduplicating lists to create a single sampling frame is complicated by problems such as incorrect spellings of names, use of complete first names vs. initials, and spouses who may donate together or individually. Procedures for unduplicating lists must search for possible matches as well as for exact matches. These possible matches may need to be resolved by hand. Alternately, sampling may be done from the multiple lists and respondents may be asking how many of the lists they should appear on. This allows one to determine the probability that an individual was selected for the sample.

2. Response Error

2.1 Definition of Response Error

Response error occurs when a respondent intentionally or unintentionally gives an answer to a survey question that is not correct. An intentional response error occurs when the respondent purposely gives an incorrect answer, for example, to a question that asks for sensitive information about illegal activities (drug use, cheating on taxes), for personal information on sensitive topics (being a victim of rape), or about behavior that may not be perceived as socially ac-

ceptable (not voting, owning a gun). The respondent may give an intentionally incorrect answer for self-protection if he or she is concerned that the survey responses will not be held confidential and may be used against him or her. Or the respondent may not wish to appear in a bad light to the interviewer and may give the more socially acceptable answer.

Unintentional response errors can occur if questions are poorly worded and the respondent does not understand the question or terms in the question. The order in which questions and response choices are presented to a respondent can influence the responses that are given. In addition, response errors can occur if the respondent does not take time to think about a question that asks for an opinion and instead takes the easy option of agreeing with the statement or choosing the answer that is easiest to recall (called satisficing). (See Krosnick and Alwin 1987 for details on response-order effects and satisficing.)

2.2 Avoiding Response Error

Carefully designed and worded questionnaires, and, for in-person or telephone surveys, well-trained interviewers are the best tools for avoiding response errors (see *Sample Surveys: Cognitive Aspects of Survey Design*). Questionnaires must be pilot tested to ensure that respondents understand questions and that all possible responses are allowed for on closed-form questions.

2.3 Randomized Response Techniques for Sensitive Questions

Randomized response methods have been proposed by a number of researchers (see, for example, Warner 1965, and Nathan 1988) to avoid the problem of having respondents intentionally give an incorrect answer to a question. The idea is to protect the respondent by ensuring that the interviewer is unaware of which question the respondent is answering.

There are several versions of the randomized response technique. In one, the respondent randomly selects one of two types of cards containing opposite versions of the question. For example,

Version A:

I have used cocaine at least once. YES NO

Version B:

I have never used cocaine. YES NO

Whichever the respondent answers, yes or no, the interviewer does not know whether the respondent has used cocaine or not. The data from such responses, however, can be used to estimate the proportion of the population that has used cocaine as long as there are unequal numbers of Version A and Version B cards.

Let p = population proportion of interest (here the proportion that has used cocaine). Let $\theta \neq 1/2$ be the proportion of Version A cards. Suppose that Y of the n respondents answered 'yes' to whichever question they were asked. One can use conditional probability arguments to show that an unbiased estimate of p, assuming simple random sampling and ignoring the finite population correction factor, is

$$\tilde{p} = (\theta - Y/n)/(2\theta - 1)$$

The variance of this estimator is

$$\mathrm{Var}(\tilde{p}) = p(1-p)/n + \theta(1-\theta)/n(2\theta-1)^2$$

which is larger than that obtained by asking the question directly, $p(1-p)/n$, because we have introduced extra variability by asking a randomly chosen question. Notice that the variance will decrease as θ gets closer to 0 or 1, so that most of the cards are of one type (Version A or Version B). Thus, from a mathematical perspective, we prefer θ far from 1/2. Respondents, however, prefer θ close to 1/2 so that they feel confident that the interviewer does not know what question is being answered. The tradeoff is an increase in variance in order to ensure that the respondent will provide a true, unbiased response.

In practice, randomized response techniques are not as successful as one would hope. The explanation of the technique may confuse the respondent, or the respondent may not really believe that the interviewer does not know what question is being answered. Thus, such techniques must be used with great care.

3. Nonresponse Error

3.1 Definition of Nonresponse Error

Almost every data collection endeavor results in missing data or nonresponse. In telephone or in-person surveys and censuses, for example, some sampled individuals cannot be located during the interview period, some are ill or unable to communicate with the interviewer, and some refuse to respond to the survey. The best that one can hope for when there is nonresponse is that the missing data are like a random sample of the units selected for the survey. If this is the case then the data are said to be missing completely at random and the effect on estimates is that they are more variable since they are based on a smaller sample size than planned. More often, however, those who do not respond to the survey differ from those who do respond on at least some survey variables of interest. In these cases the data are said to be missing at random if the missingness can be explained by other variables that are observed; nonresponse is nonignorable if the missingness depends on the value that is not observed (see Little and Rubin 1987 for a more complete discussion of types of

missingness). For example, people who are not at home to respond to a telephone survey in the evening may watch less television than those who are at home. In this case, estimates of average hours of television viewing per week based only on data from respondents will be biased low. (Note that the above example also illustrates why it is not appropriate to replace nonresponding units in your sample with other units from the population: if there are differences between responders and nonresponders, choosing additional responders only repeats the bias on a larger scale.)

Nonresponse is a particular problem in longitudinal or multiple-wave surveys and studies where sampled individuals may drop out of the study after several interviews. If the reason for dropping out is related to the variables of interest, then nonresponse bias is introduced. For example, individuals in housing units sampled for the US National Crime Victimization Survey are interviewed seven times over 3.5 years. Individuals who move out of sampled housing units are not followed. If individuals who experience crime are more likely to move, perhaps to a safer neighborhood, then estimates of victimization rates will be biased low (see Lohr 1999 for a description of the National Crime Victimization Survey). A similar problem occurs in longitudinal medical studies if, for example, patients given the placebo see no change in their condition and thus are more likely to drop out of the study than patients receiving a treatment (see *Sample Surveys: Model-based Approaches*).

There are two types of nonresponse in surveys and censuses: unit nonresponse and item nonresponse. If no survey information is collected from a sampled unit this is termed unit nonresponse; if a sampled unit responds to some, but not all, survey questions this is termed item nonresponse. There are different methods for dealing with unit and item nonresponse.

3.2 Dealing with Nonresponse Error

The best way to deal with nonresponse error is to try to avoid it. Thus, good survey organizations use repeated callbacks on telephone and in-person surveys, and reminder post cards or multiple mailings on mailed surveys to try to reach sampled individuals. Introductory letters or statements at the beginning of questionnaires should describe the purpose of the survey, provide information on who is funding and conducting the survey, stress the importance of the person's responses to the survey, and assure the respondent that all survey responses will be kept confidential. Some surveys provide a payment or gift to respondents, or draw for prizes from among the respondents to encourage response.

In spite of the best efforts to encourage response, most surveys suffer from nonresponse. A researcher thus has the responsibility, at a minimum, to report the nonresponse rate for the survey along with any results from the survey. There can be many definitions of the response rate for a survey, depending on whether one counts all units originally sampled, only those that led to eligible respondents, only those that resulted in some contact with a respondent, and so forth. See the American Association for Public Opinion Research (1998) for standard definitions of response rates.

The first step in dealing with nonresponse is to conduct exploratory data analysis on the survey data. For unit nonresponse, after determining the overall nonresponse rate one should then look at response rates by other variables that are available for both responding and nonresponding units. For example, in a mailed survey of a larger city, one could look at response rates by zip code. If there are big differences in response rates for various areas of the city, say that there are higher response rates from the more affluent areas, then survey estimates for variables that are related to affluence, such as support of the arts, are likely to be biased. In a survey of university alumni, one can compare respondents and nonrespondents on any of the variables available in the university alumni database, such as GPA. If, for example, those with higher GPAs responded more often than those with lower GPAs, then survey estimates on variables related to GPA, say starting salary in the first job after graduation, are likely to be biased. If one does not find differences such as those suggested above, then at least one knows that there are not differences on the variables that could be explored. This does not, however, insure that there are no differences between respondents and nonrespondents on the survey variables of interest.

For item nonresponse, one should start by determining how much nonresponse there is for each variable with some missingness. For key study variables, one should look at summaries of other variables according to whether or not the key variable is missing. For example, if income is a key variable and subject to nonresponse, one could look at average age, average years in job, proportion married, proportion with high school diploma or less, and so forth for respondents who reported their income and those who did not. Obvious differences between the two groups on the other variables suggest that there are likely to be differences between the respondents and the nonrespondents on the income variable. It also suggests that the other variables might be useful in determining an adjustment for the nonrespondents, as described below.

3.3 Weight Adjustments

The sample-based weights assigned to cases in a sample survey can be interpreted as the number of people in the population of interest that the individual in the sample represents. The initial weight is usually the inverse of the selection probably for each sampled

unit. One commonly used method to adjust for unit nonresponse in a sample is to increase the weights of respondents to account for nonrespondents. This adjustment is usually done within weight-adjustment cells within which it must be assumed that nonrespondents are 'just like' nonrespondents on variables of interest. That is, within weight-adjustment cells, the nonrespondents can be thought of as a random sample of the sampled units in that cell.

As a simple example, suppose you have a simple random sample of $n = 100$ individuals: 50 men, of whom 30 respond, and 50 women, of whom 40 reply. (Since we are assuming that a simple random sample of individuals was selected, all individuals in the sample will have equal selection probabilities and thus equal initial weights.) If you believe that within gender the nonrespondents are 'just like' the respondents (that is, the nonrespondents can be thought of as a random sample of those originally sampled) on survey variables of interest, then you may decide to use male/female as the weight adjustment cells. The 30 responding males would represent themselves and the 20 nonresponding males. Thus, the original weights of the males must be multiplied by 50/30. Similarly, the original weights of the women would be multiplied by 50/40 to account for the nonresponding women.

In practice, there are usually more than the two nonresponse adjustment cells described in the simple example here. Little and Rubin (1987) describe methods for creating the nonresponse adjustment cells. Brick and Kalton (1996) provide details on weight adjustments for nonresponse.

3.4 Imputation Methods

Imputation methods are most often used to deal with item nonresponse. Imputation is defined as replacing a missing value by a plausible value. An advantage of imputation is that it creates a completed dataset that can be analyzed using standard statistical methods. A disadvantage is that most subsequent analyses ignore the fact that imputation was done and, therefore, underestimate the variance by ignoring the uncertainty due to the imputation. There are many methods for creating plausible values to impute for a missing value. The better methods make use of information in the available data or in an auxiliary data source to estimate the missing values. The following describes some common methods of imputation. More details and additional methods may be found in Little and Rubin (1987) and *Statistical Data, Missing*.

Mean imputation involves replacing a missing value of some variable by the sample mean computed from all individuals who reported a value for that variable. For example, mean income from those who reported income would be imputed for all those who refused to report income. The main disadvantages of mean imputation are that it distorts the distribution of the variable by creating a spike at the mean value, and it distorts relationships and correlations among variables. For example, the relationship between years experience on the job and income is not preserved when mean income is imputed for all those missing income regardless of years of experience. For this reason, mean imputation may be done within imputation cells that are determined by variables thought to be related to the missing values. Thus, years experience may be used to divide respondents into groups, and the mean income within each group is imputed for those missing income in the group. Mean imputation within group still created spikes in the data (at the mean for each group) and distorts the relationships among variables, but these effects should be less than with overall mean imputation.

Regression imputation involves using a regression model fit from complete cases to estimate missing values. For example, complete cases may be used to develop a model for income based on age, education, tenure in job, and other available data. This fitted model would then be used to estimate income when it is missing. Because this method may also create spikes in the data when there are few possible combinations of the predictor variables used in the regression model, a variation involves adding a random error to the estimated value to be used for imputation.

Cold deck imputation involves using auxiliary data as a source for missing values. For example, data from administrative records might be used to obtain missing survey values, or, in a panel survey, values from a previous interview for the same respondent could be used to replace missing values. Hot deck imputation replaces missing values with values taken from other similar respondents' records in the same dataset.

The imputation methods described above all result in underestimation of the variance because the imputed values are treated as 'truth' and subsequent analyses typically ignore the uncertainty in the imputations. Multiple imputation, first developed by Rubin (1987), allows one to account for this uncertainty. In multiple imputation, a stochastic imputation procedure is used to create, say, m imputations for each missing value. These m completed datasets are used in the standard analysis of the data. The variation among estimates from the m analyses is used to estimate the variance due to imputation. It is combined with the usual variance from the standard analyses to provide an overall variance of the estimates. For most larger surveys, m can be fairly modest: $m = 5$ is often adequate. For more details, see *Multiple Imputation*.

4. Other Sources of Error

There are other sources of nonsampling error that can effect survey data. Errors caused by interviewer bias, panel rotation, survey mode, coding, and data entry are all errors that should be avoided in conducting

surveys. We will briefly consider panel rotation and survey mode errors.

Panel rotation error occurs when responses to a survey change depending on the length of time the panel has been included in the survey. Bailar (1975) described the effects of time-in-sample for the US Current Population Survey. In general, the longer a person has been in the sample, the less likely the person is to report being unemployed. A similar effect is found in the US National Crime Victimization Survey; persons in the survey for a longer time are less likely to report a victimization. Possible explanations for this phenomenon include that respondents learn to answer the survey so as to minimize the amount of time it takes to complete the survey or that differential dropout leads to a biased sample over time.

Survey modes include in-person, telephone, self-administered paper-and-pencil questionnaires and, most recently, self-administered Web questionnaires. The same questions asked using different survey modes can elicit different responses (see *Sample Surveys: Cognitive Aspects of Survey Design*). For example, respondents are more concerned about their self-images in personal interviews than on paper-and-pencil surveys. Thus, they may be more likely to report personal problems on a self-administered questionnaire than in a face-to-face interview. On the other hand, security and confidentiality concerns may make respondents less likely to respond truthfully to surveys on the Web. The accuracy of responses to Web surveys is an area of ongoing research.

Additional information on these other sources of nonsampling error can be found in references such as Groves (1989) and Biemer et al. (1991). An excellent example of a survey for which all sources of error, both sampling and nonsampling, are explored in detail is the US Current Population Survey (see US Department of Commerce 2000).

Bibliography

American Association for Public Opinion Research 1998 Standard definitions: Final dispositions of case codes and outcome rates for RDD telephone surveys and in-person household surveys. AAPOR, Ann Arbor, MI

Bailar B A 1975 The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association* **70**: 23–30

Biemer P P, Groves R M, Lyberg L E, Mathiowetz N A, Sudman S 1991 *Measurement Errors in Surveys*. Wiley, New York

Bosecker R R, Ford B L 1976 Multiple frame estimation with stratified overlap domain. *Proceedings of the Social Statistics Section of the American Statistical Association*. 219–24

Brick J M, Kalton G 1996 Handling missing data in survey research. *Statistical Methods in Medical Research* **5**: 215–38

Cochran W G 1977 *Sampling Techniques*, 3rd edn. Wiley, New York

Groves R M 1989 *Survey Errors and Survey Costs*. Wiley, New York

Kish L 1949 A procedure for objective respondent selection within the household. *Journal of the American Statistical Association* **44**: 380–7

Kish L 1965 *Survey Sampling*. Wiley, New York

Krosnick J A, Alwin D F 1987 An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly* **51**(2): 201–19

Lavrakas P J 1993 *Telephone Survey Methods*, 2nd edn. Sage, Newbury Park, CA

Lessler J T, Kalsbeek W D 1992 *Nonsampling Error in Surveys*. Wiley, New York

Little R J A, Rubin D B 1987 *Statistical Analysis with Missing Data*. Wiley, New York

Lohr S L 1999 *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA

Mosteller F 1978 Errors: Nonsampling errors. In: Kruskal W H, Tanur J M (eds.) *The International Encyclopedia of Statistics*. Free Press, New York, pp. 208–29

Nathan G 1988 *A Bibliography on Randomized Response: 1965–1987*. *Survey Methodology* **14**: 331–46

Rubin D B 1987 *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York

US Department of Commerce, Census Bureau 2000 *Current Population Survey: Design and Methodology*. Government Printing Office, Washington, DC

Warner S L 1965 Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60**: 63–9

E. A. Stasny

Nonstandard Reasoning

Logic since Antiquity has been concerned with 'the laws of thought,' immutable principles of reasoning from premises to conclusions. In the twentieth century, this monolithic view changed to one of different human *reasoning styles* and matching logical systems. With the emergence of information and cognition as central scientific themes, new items keep appearing on this broader agenda. These are often called 'nonstandard logics,' different from the established regime. This label is relative. The background to current studies of reasoning are the major approaches to valid inference in standard logic (semantic or proof-theoretic), reflecting philosophical argument and mathematical proof. Over time, these have acquired extended vocabularies and deviant notions of inference with linguistic and other scientific motivations. Nonstandard logics often originate in computer science and artificial intelligence (AI) (nonmonotonic logic, abduction). The resulting landscape is diverse, but there are unifying themes such as structural rules, preferences, resources, knowledge, or architectures for combining systems. These reflect cognitive trends in modern logic: reasoning about one's own information and that of others, update processes, games, communication, or heterogeneous information.

1. Standard Logic: Valid Consequence, Proof Calculi, Completeness

Standard logic covers the achievements of the renaissance of logic, starting with Boole and Frege in the nineteenth century, and culminating in the work of Gödel, Tarski, Turing, and others in the 1930s. The resulting core of the field consists of propositional logic (codifying reasoning with the key operations 'not,' 'and,' 'or') and predicate logic of quantifiers ('for all,' 'there exists'). Both are based on the semantic conception of valid consequence, demanding 'transmission of truth' from premises P to a conclusion C:

every situation in which all the premises P are true also makes the conclusion C true.

To describe and produce valid inferences, one can sometimes use simple mechanical methods, such as the truth tables of propositional logic. More general, and more relevant to capturing actual reasoning, are proof-theoretic approaches (Hilbert), deriving conclusions from premises via rules. Proof systems of various sorts also drive mechanized reasoning in AI and computational logic. Crucially, the semantic approach based on truth and the syntactic one of proof coincide for quantificational logic: this is Gödel's celebrated Completeness Theorem, the first significant insight into logical systems *per se*. Another is Church's Theorem (based on Gödel's even more famous 'Incompleteness Theorem'), saying that no mechanical algorithm can decide validity in quantificational logic.

2. Extending the Vocabulary: from Mathematics to Natural Language

Logical systems use 'logical forms' with fixed operators for key notions (Booleans, quantifiers), plus variable parts for expressions that can be interpreted freely (a typical valid form is: 'from premises A-*or*-B, and *not*-A to conclusion B'). The above systems have been successful in analyzing mathematical proof, but gradually extensions had to be made to deal with the kind of argument found in philosophy, linguistics, and in the latter part of the twentieth century, computer science and cognitive science. For example, logical semantics of natural language deals with generalized quantifiers ('most,' 'few'), plus a spectrum of expressions for reasoning involving possibility, time, knowledge, action, etc.—for both single and multiple agents. The extra vocabulary is 'nonstandard,' the canons of validity remain classical.

3. Alternative Views of Validity and Logical Operations

There are also genuine deviations from the classical core. A recurrent historical example is the view that conclusions only follow validly from *consistent* premises. Then, nothing follows from a contradiction A-*and*-*not*-A, while on the 'standard' account, everything does.

3.1 Deviations in Natural Language and Philosophy

The idea that 'natural reasoning' in our ordinary language differs from standard logical calculi is a recurrent theme in the twentieth century—high-lighted in the 'paradoxes of material implication': natural 'if … then … ' does not behave like the truth-functional Boolean conditional. This has led to many deviant calculi for 'modal' or 'relevant' implication. Philosophical critiques of the standard account of valid consequence still appear today.

3.2 Deviations in Scientific Reasoning

Other deviations arise in the foundations of science. 'Constructivism' in mathematics views proofs as constructions of mental entities, whose properties must be established by computational means. *Intuitionistic logic* accepts only those inferences as valid which have a constructive proof. Thus, one cannot prove an assertion 'there exists an x such that $F(x)$' without producing an explicit example. This line is generalized in constructive 'type theories.' In physics, quantum mechanics has inspired *quantum logics* with failures of Boolean distributivity laws in reasoning about the location or momentum of particles. But most of all, reflecting the close analogies between 'reasoning and reckoning' (as Hobbs put it), computer science has inspired many new systems of reasoning. Examples from the 1970s are *dynamic logic* describing the effects of programs and actions generally, *logic programming* (using proofs as computation procedures) and perhaps most radically, *linear logic* in the 1980s, treating propositions as computational 'resources' that can be used only once. The result of all this is alternatives to classical logic, developed with equal rigor, but different in their valid principles, general philosophy, and mathematical properties.

4. Varieties of Inference

That 'natural reasoning' comes in different styles, with possibly different logical features, was already observed by Bolzano in the early nineteenth century. Scientific reasoning employs stricter standards than common sense. Variety also occurs in AI with problem solving or planning (deductive or heuristic), where systems are more 'optimistic' than standard logic.

4.1 Nonmonotonic Logics

Practical reasoning uses defaults going beyond standard logical conclusions. Thus, we use some rule

('Dutch trains run on time') unless counter-examples are explicitly indicated, e.g., when planning our escape from Amsterdam's futuristic Science City. Or we take an answer 'John, Mary' to the question 'Who were skating?' as conveying not just that John and Mary skated, but that *no others did*, assuming a maximally informative respondent. Such additional inferences may be withdrawn later, because of further information: the train may be carrying soccer fans, the respondent may be a government employee. Reasoning in which C may follow from P, but no longer from P + Q is called 'nonmonotonic'—while classical consequence is monotonic: once drawn, conclusions persist. A general pattern behind nonmonotonic reasoning is 'circumscription,' which, in its most abstract form, assumes that situations come ordered by some preference, and then makes C follow from P if

C holds in all most preferred situations where P holds.

In the escapist example, one 'prefers' the more likely situations where trains run according to schedule to those containing a disturbing factor. Preferences may reflect plausibility, utility, or simplicity, depending on circumstances. A classical consequence is the special case when no models are preferred to others. Thus, classical inferences are also valid in circumscription. This is a general phenomenon in AI: most logics that diverge from classical logic are richer, whereas the deviant logics of Sect. 3 were poorer, putting greater demands on inference.

4.2 Directions and Purposes in Reasoning

The preceding systems all view reasoning as a forward-oriented process of accumulating conclusions from given premises. But in addition to this forward deductive mode, there is also 'backward' *abduction* (Peirce) seeking plausible premises to support already given observations. Moreover, there is *induction* (Mill, Carnap), looking for plausible generalizations beyond the immediate content of the premises. These different directions and purposes all occur when analyzing activities such as 'refutation,' 'explanation,' or 'confirmation' in the philosophy of science. Abduction is studied also in connection with logic programming, and induction in the foundations of probability and uncertainty.

4.3 Reasoning with Uncertainty and Probability

Probabilistic reasoning seems a major mode of cognition in its own right. Its standard formalism is not logic, but probability theory. However, there is an increasing literature on nonstandard, often qualitative probabilistic reasoning, including 'fuzzy logic' and 'possibility theory,' while key probabilistic notions such as randomness and independence of events are migrating into logic. In particular, new versions of predicate logic have been proposed that admit dependencies between variables. Interestingly, these often become *decidable*, unlike the standard system.

5. Logical Dynamics: Reasoning as an Activity

Under the influence of computer science and cognitive science, many logicians view semantic interpretation or reasoning as cognitive processes. This is closer to actual argumentation, where dynamic timing and procedural conventions are crucial to outcomes. Within this trend, *dynamic semantics* views the meaning of a proposition as the update of a hearer's information state produced by it. Thus, propositions become cognitive procedures, and natural language a programming language modifying information states of other agents. One major dynamic notion of valid consequence is

in final information states reached by processing the successive premises, updating with the conclusion will not have any further effect.

A broader perspective arises in *belief revision theory*, which has updates adding information to the current state, contractions (removing information), and revisions (doing both)—doing justice to the 'zigzag character' of much of our reasoning. Information states can be those of individual agents, but interactive communication updates collective information states of groups of agents, often pictured as 'Kripke models.' Perhaps the most 'activist' perspective on reasoning is that of *game theory*. Logic games have existed since the fifties (Lorenzen, Hintikka), suggesting dialogical alternatives to proof and semantics, in terms of winning and losing argumentation games. Their take on valid consequence employs the typical game-theoretic notion of a *strategy*, a response pattern to any sequence of moves by one's opponents:

in fair rational debate, the proponent of the conclusion C has a winning strategy for upholding C against any play by an opponent granting the premises.

Logic games tie up the study of reasoning with that of rational behavior in general.

6. Systematic Theory of Reasoning

With the rapid growth of nonstandard logics, the unity of the meta-theory for classical logic gets lost. It is too early for a new synthesis, but some unifying themes are noticeable.

6.1 Structural Rules

Nonmonotonicity is a hallmark of nonstandard reasoning systems—though a symptom, not a diagnosis. However, it has turned out useful to classify styles of reasoning by their *structural rules*, such as

monotonicity, transitivity ('if P implies Q, and Q implies R, then P implies R'), and others, which move premises around, or duplicate them when needed. The structural rules of classical logic treat premises as an unordered set, whose presentation is irrelevant. By contrast, nonstandard reasoning tends to be more sensitive to the formulation of data, which shows in failures of classical structural rules. Thus, in dynamic consequence, permuting two premises may change the update achieved—reflecting the order-dependence of natural language.

6.2 New Vocabulary

Nonstandard consequence may reinterpret traditional logical operations. Thus, in logic games, a negation signals a 'role switch' between two players. Nonstandard reasoning may also involve new operations, outside the classical core. In particular, with games, there is not one conjunction, as in Boolean logic, but *three*: choosing which game to play at the start, playing one game after the other, or playing two games concurrently. The resulting calculi are much richer in vocabulary than classical ones.

6.3 Mechanisms and Structures

More important than individual 'logical laws' are general mechanisms behind nonstandard reasoning systems. Striking new notions are *preference*, *dynamics*, and *resources*, mentioned before. These provide a much richer picture of actual reasoning, and 'parameters' that can be set for different purposes. Other important trends go beyond the single-sentence habitat of standard logic—addressing larger-scale, longer-term data structures and processes. These include *cooperative architecture* (how do different styles of reasoning cooperate to achieve useful effects?), *heterogeneous information* (how do symbolic and nonsymbolic, e.g., visual data cooperate in reasoning?), and *meso-structure* (which higher-level discourse and theory structures drive reasoning?). Most strikingly, reasoning is a *social activity* involving agents that communicate, cogitate and act—and *multiagent logics* of knowledge and collective information update are conspicuous in contemporary research.

7. Conclusion

Modern logic arose out of reflection on proof in mathematics and argumentation in philosophy, and created the standard systems of the twenty-first century. Modern studies of reasoning, while retaining the modus operandi of this phase, deal with a much broader spectrum of laws, structures, and processes, and in doing so, merge with disciplines such as methodology of science, argumentation theory, linguistics, computer science, or cognitive psychology. In this perspective, 'nonstandard reasoning' signals nothing specific, except this broad horizon.

See also: Artificial Intelligence: Uncertainty; Axiomatic Theories; Deductive Reasoning Systems; Game Theory; Knowledge Representation; Logics for Knowledge Representation; Mathematical Models in Philosophy of Science; Scientific Discovery, Computational Models of

Bibliography

Abramsky S, Gabbay D, Maibaum T (eds.) 1992–2000 *Handbook of Logic in Theoretical Computer Science*. Oxford University Press, Oxford, UK

Barwise J (ed.) 1977 *Handbook of Mathematical Logic*. North-Holland, Amsterdam

Gabbay D, Guenthner F (eds.) 1983–8 *Handbook of Philosophical Logic*. Reidel, Dordrecht, The Netherlands; Revised and expanded edition to appear 2001, Kluwer, Dordrecht, The Netherlands

Gabbay D, Hogger C, Robinson J (eds.) 1993–8 *Handbook of Logic in Artificial Intelligence and Logic Programming*. Oxford University Press, Oxford, UK

Gabbay D, Woods J (eds.) in press *Handbook of Argumentation Theory*. Kluwer, Dordrecht, The Netherlands

Van Benthem J, ter Meulen A (eds.) 1997 *Handbook of Logic and Language*. Elsevier, Amsterdam

J. van Benthem

Nontraditional Families and Child Development

The process of development among children and adolescents growing up in different kinds of family environments has long been a topic of interest to social scientists. For many years, most researchers have assumed that conditions for child development are most favorable in families that include two heterosexual parents, one of each sex, who are married to one another, and who are biologically related to the child. Fathers in such families are assumed to be employed fulltime outside the home, and mothers are assumed to work only in the home, where they are responsible for childcare and upkeep of the household, but do not earn money. Even though the existence of such so-called 'traditional families' has not characterized much of human history, and even though many families today—even in Westernized countries—do not fit this pattern, it has nevertheless been widely assumed as the norm against which other family rearing environments should be measured (Lamb 1999).

There are a number of ways in which families may differ from a 'traditional family,' and hence be regarded as 'nontraditional.' For instance, in many families, mothers are participating in the labor market; indeed, in the USA today, most mothers of young children and even infants are employed outside the home. At least in part as a result, many infants and children spend some of their waking hours in childcare settings outside of the family home. Many families include only one parent (whether because of parental separation, divorce, or other reasons), and many families include parents who are not biologically related to their children (e.g., adoptive parents, stepparents). In addition, many children are growing up in families headed by lesbian or gay parents, and may thus live with two same-sex parents.

1. *Research on Nontraditional Families*

What is known about child development in nontraditional families, and about the impact of nontraditional family contexts on child development? It is widely agreed that multiple interacting factors in intersecting trajectories can influence the development of children in different kinds of family environments. Thus, no single perspective or factor is likely to account for all of the variation that can be observed. An important lesson to be learned from research on child development in nontraditional family environments, however, has been that family stability, and the qualities of relationships and interactions among family members are likely to be more important than family structure and household composition as proximal determinants of outcomes for children and adolescents. In families and households whose membership remains stable over time, it is not so much who the parents are or where the children are cared for during the hours while parents work, but what the child's interactions with parent figures are like, and the qualities of family relationships that are likely to have the greatest influences on the development of children and teenagers growing up in them. Research on nontraditional families suggests that, given reasonable stability of family membership over time, it is family process rather than family structure that provides the final common pathway through which outcomes for children and adolescents emerge (Amato 1994, Patterson 1997, Lamb 1999).

2. *Research on Parental Divorce and Child Development*

Research on parental divorce and remarriage has illustrated how stressful interparental conflict and family disruption can be in childhood and in adolescence (see *Divorce and Children's Social Development*). On scales designed to assess stressful experiences in childhood and adolescence, parental separation and divorce rank near the top. In addition, other events that are commonly associated with parental divorce (e.g., moving from one home to another, entering a new school) are also extremely stressful. As a result, it is not surprising that children of divorce show elevated behavior problem profiles, are more likely to have difficulties in school, and are more often referred for mental health treatment than are same-aged children from nondivorced homes. Although some of these problems are typically resolved within two years after the divorce, others persist. Adolescent offspring of divorced parents are also more likely than others to show conduct problems, to drop out of high school, to be unemployed, to be involved in delinquent activities, and to associate with antisocial peers (Hetherington et al. 1998). Some of these difficulties extend even into adulthood (Cherlin et al. 1991, Hetherington et al. 1998). Many of the problems experienced by children and adolescents when their parents divorce seem to have begun even before parental separation or divorce (Amato and Booth 1996, Block et al. 1986), and are associated primarily with poor quality parent–child relationships and interactions. Consistent with this view, when children's predivorce levels of functioning are taken into account, in longitudinal designs, differences in adjustment after divorce are reduced (Amato and Booth 1996, Block et al. 1986). Similarly, when contacts with noncustodial parents following divorce are studied, it is the quality rather than the frequency of contact that is related to children's adjustment (Emery 1988). Taken together, results of research on divorce seem to highlight the importance for children of parent–child relationships that buffer them from the effects of stressful life changes.

3. *Research on Child Development in Other Types of Nontraditional Families*

As difficult for children as family instability and difficulties in parent–child relations can be, other variations from the so-called 'traditional family' arrangements, especially those that do not involve disruption of the family unit, seem to be less troublesome.

3.1 *Maternal Employment and Nonparental Childcare Arrangements*

One deviation from the 'traditional family' norm that seems to have little or no negative impact on children is maternal employment. Mothers' participation in the labor force has increased dramatically in the USA since the mid-twentieth century (Gottfried et al. 1999), with the result that most mothers of infants and young children are now employed. Results of a large and

diverse body of research now show convincingly that maternal employment *per se* is not detrimental to the development of children or adolescents; far more important than the fact of maternal employment are the actual parent–child relationships within the family (Gottfried et al. 1999). Interestingly, one correlate of maternal employment in dual-earner families seems to be increased paternal involvement with children; again, the impact of such increased involvement is likely to be associated with the qualities of actual father–child interactions. Related to the increase in maternal employment during recent years is a concomitant increase in the use of out-of-home childcare arrangements. Here again, despite concerns that children and especially infants might be put at risk by out-of-home care arrangements, research has shown that high-quality out-of-home care has little or no discernible impact upon existing parent–child relationships or upon children's overall levels of adjustment (see *Child Care and Child Development*).

3.2 Adoptive Families

Another way in which families can differ from the 'traditional' normative family is through adoption. Adoptive families may be very like 'traditional families' in all ways except the adoption of a child who is biologically unrelated to the parents, or they may be different in many respects (e.g., number or sexual orientation of parents). Consistent with the notion that family stability is an important influence on children's development, one of the best predictors of adoptive children's overall adjustment is the age at which they were adopted. Those who were adopted early in life (e.g., before the age of two years) are, on average, more well adjusted than those who were adopted later, or who have had many placements (Brodzinsky 1987). Other research has suggested that adoptive parents, who are generally highly motivated for parenthood, provide home environments that are at least equal to those provided by parents whose children are biologically related to them. Although a minority of adopted adolescents experience difficulties relating to their status as adoptees, most report that being adopted is only a minor part of their identity (Bemon & Sharma 1994).

3.3 Families Headed by Lesbian and Gay Parents

Even though parents are generally expected to be heterosexual, many families also include one or more parents who are lesbian, gay, or bisexual (Patterson 1992, 1997, Patterson and Chan 1999, Tasker and Golombok 1991). In some cases, children are born to a heterosexual couple in the context of a marriage that eventually breaks up when one or both marital partners adopts a non-heterosexual identity. In these families, children may then grow up with either a custodial and/or a noncustodial lesbian, gay, or bisexual parent. In other families, lesbian, gay, and bisexual adults become parents after assuming non-heterosexual identities. In these families, children may be adopted or they may be conceived in one of a variety of different ways (e.g., using donor insemination or surrogacy arrangements).

Despite negative expectations voiced by many different observers, research by social scientists has failed to reveal any detrimental effects attributable to the experience of growing up with lesbian, gay, or bisexual parents (Patterson 1992, 1997, Tasker and Golombok 1991, 1997). In childhood, children of lesbian mothers are no more likely than other children to have conduct problems or social difficulties. Children with lesbian mothers are no more likely than those with heterosexual mothers to show disturbances in sex-role development. The overall life adjustment of the young adult offspring of divorced lesbian as compared with divorced heterosexual mothers is quite similar (Tasker and Golombok 1997). Research has yet to focus on children born to or adopted by gay fathers, or on adolescents or young adults who were born to lesbian or gay parents. Results of available research, however, suggest that in many ways development of children of lesbian and gay parents is similar to that of children in more 'traditional' family settings (Patterson and Chan 1999).

Although development of children with lesbian or gay parents is similar in many ways to that among children in 'traditional' families, there are also some notable differences. For instance, young adult offspring of divorced lesbian mothers report more positive attitudes about lesbian and gay people than do the same-aged offspring of divorced heterosexual mothers (Tasker and Golombok 1997). Another clear difference between the experiences of children parented by lesbian or gay versus heterosexual couples involves the divisions of family labor that are likely to be observed in their homes. Children raised by lesbian or gay couples are more likely than those raised by heterosexual couples to observe their parents using egalitarian divisions of labor, especially for childcare. Some research suggests that relatively equal division of childcare duties may be associated with positive outcomes among children of lesbian mothers, but other studies suggest that it is unrelated to child outcomes except insofar as it relates to parental relationship satisfaction, and the issue is currently under study (Patterson and Chan 1999).

4. Conclusion and Future Directions

One of the important findings of research on 'nontraditional families,' then, is that considerable variation in family structure and membership can occur without measurable detrimental effects on children. As long as family stability is high, and the qualities of

parent–child relationships and interactions are good, then children appear capable of developing along positive trajectories within a wide array of different family environments. Thus, even though cultural change has engendered considerable anxiety among many different observers, numerous deviations from the 'traditional family' have not proven to be detrimental to development of children. On the contrary, if family stability is maintained and if the qualities of parent–child relationships and family processes are positive, then nontraditional families of many descriptions can provide environments that foster positive child development.

As nontraditional families grow in numbers and proliferate in variety, it seems likely that research on children's development will increasingly include them. Future research on nontraditional families should focus not only on structural features of these families but also on their functional characteristics, such as qualities of interactions and relationships among family members. Comparisons of important influences on children across family types could yield significant insights into which elements of family life are necessary and which are optional in creating environments that support and enable positive development among children and youth.

See also: Divorce and Gender; Divorce, Sociology of; Gay, Lesbian, and Bisexual Youth; Lesbians: Social and Economic Situation; Parenting: Attitudes and Beliefs; Queer Theory

Bibliography

Amato P R 1994 The implications of research findings on children in stepfamilies. In: Booth A, Dunn J (eds.) *Stepfamilies: Who benefits? Who does not?* Lawrence Erlbaum, Hillsdale, NJ

Amato P R, Booth A 1996 A prospective study of divorce and parent–child relationships. *Journal of Marriage and the Family* **58**: 356–65

Bemon P, Sharma A S 1994 *Growing Up Adopted: A Portrait of Adolescents and Their Families*. Research Institute of Minnesota, Minneapolis, MN

Block J H, Block J, Gjerde P F 1986 The personality of children prior to divorce: a prospective study. *Child Development* **57**: 827–40

Brodzinsky D M 1987 Adjustment to adoption: a psychological perspective. *Clinical Psychology Review* **7**: 25–47

Cherlin A J, Furstenberg F F, Chase-Lansdale P L, Kiernan K E, Robins P K, Morrison D R, Teitler J O 1991 Longitudinal studies of effects of divorce in children in Great Britain and the United States. *Science* **252**: 1386–9

Emery R E 1988 *Marriage, Divorce and Children's Adjustment*. Sage, Newbury Park, CA

Gottfried A E, Gottfried A W, Killian C 1999 Maternal and dual-earner employment: family environment, adaptations, and the developmental impingement perspective. In: Lamb M E (ed.) *Parenting and Child Development in 'Nontraditional' Families*. Lawrence Erlbaum, Mahwah, NJ

Hetherington E M, Bridges M, Insabella G 1998 What matters? What does not? Five perspectives on the association between marital transitions and children's adjustment. *American Psychologist* **53**: 167–84

Lamb M E (ed.) 1999 *Parenting and Child Development in 'Nontraditional' Families*. Lawrence Erlbaum, Mahwah, NJ

Patterson C J 1992 Children of lesbian and gay parents. *Child Development* **63**: 1025–42

Patterson C J 1997 Children of lesbian and gay parents. In: Ollendick T, Prinz R (eds.) *Advances in Clinical Child Psychology*. Plenum, New York, Vol. 19, pp. 235–82

Patterson C J, Chan R W 1999 Families headed by lesbian and gay parents. In: Lamb M E (ed.) *Parenting and Child Development in 'Nontraditional' Families*. Lawrence Erlbaum, Mahwah, NJ

Tasker F, Golombok S 1991 Children raised by lesbian mothers: the empirical evidence. *Family Law* **21**: 184–87

Tasker F, Golombok S 1997 *Growing Up in a Lesbian Family: Effects on Child Development*. Guilford, New York

C. J. Patterson

Nonverbal Communication, Social Psychology of

Nonverbal communication refers to the ways in which human beings convey information about their emotions, needs, intentions, attitudes, and thoughts without the use of verbal language. Nonverbal cues serve important functions in human social life, including expressing emotions; conveying interpersonal attitudes such as friendliness, insult, or dominance; regulating affect; regulating turn taking between people in conversation; and facilitating one's own speech production. Nonverbal signals are important in many psychological processes, including attachment, attraction, social influence, deception, self-presentation, and interpersonal self-fulfilling prophecies. Cultural, gender, and other group differences in nonverbal behavior have been documented as well as individual differences in usage and in the accuracy of nonverbal cue transmission and reception.

1. Nonverbal Communication as a Scientific Discipline

The topic of nonverbal communication emerged as a significant subject of empirical inquiry in the 1950s, although interest in expression has a long history (e.g., Darwin 1965). The study of nonverbal communication is interdisciplinary, spanning sociology, anthropology, ethology, communication, and virtually all of the subdisciplines within psychology (social, personality, developmental, psychophysiological, clinical, and cognitive). The empirical study of nonverbal behavior has

been greatly aided by audiovisual and computer technology which permit archives of data to be collected and analyzed. The *Journal of Nonverbal Behavior* was founded in 1979.

2. Features of Nonverbal Communication

People possess a rich understanding of how nonverbal cues are used within their cultural group and can 'read' others' cues with a high degree of accuracy; however, this knowledge is mostly implicit and often out of conscious awareness. Nonverbal communication is also ever present in social interaction because, although it is possible to refrain from speaking, it is not possible to refrain from emitting nonverbal cues, actively or passively, that will be given meaning by others. Because nonverbal cues are constantly subject to interpretation, miscommunication can easily occur. Nonverbal communication can occur with conscious intention, as when a person gives a 'thumbs up' gesture to indicate approval, or spontaneously, as when a person unknowingly reveals sadness through a slumped posture and a down-turned mouth. In this article, the term 'nonverbal behavior' is used interchangeably with 'nonverbal communication,' although some have argued that the latter term should be reserved for behavior that is known to have the properties of a shared meaning code. Although it was once thought that nonverbal communication would prove to have many of the properties of verbal language, it is now clear that the popular term 'body language' is a misnomer.

Nonverbal cues do not have much syntactic structure, and they do not combine in a rule-bound manner to produce nonverbal 'sentences.' Moreover, most nonverbal behaviors take on clear meanings only with the addition of information about the situation, antecedent events, characteristics of the communicator, and other nonverbal and verbal cues, and even then the meanings of nonverbal cues can be ambiguous. Some nonverbal behaviors are discrete and have arbitrary meanings (similar to verbal language); the most notable category of this kind is the hand gestures that are called emblems, which have fixed verbal equivalents within a culture, with examples being the 'peace sign,' the 'A-okay sign,' and the 'finger' in North American culture. However, the great majority of nonverbal behaviors are continuous rather than discrete, are not hierarchically organized, and have referents that are intrinsic or iconic rather than arbitrary (Ekman and Friesen 1969). Nonverbal cues often accompany spoken words, and when they do they can augment or contradict the meanings of the words as well as combine with the words to produce unique interpretations. An example of the latter is sarcasm, which is produced when positive verbal messages are paired with negative nonverbal cues conveyed by the face or voice.

3. Types of Nonverbal Cues

For convenience, the study of nonverbal communication is often organized into channels representing different modalities or body parts. Channels that are defined in terms of individual behavior include facial expressions, head movements, posture and body movement, hand movements, self-touching, leg positions and movements, and nonlinguistic vocal behavior. Within each of these, further distinctions can be made such as brow vs. mouth regions of the face, and, within the vocal channel, variations in speed, amplitude, and contour. Physical characteristics such as height, weight, and facial physiognomy are also studied as 'nonverbal communication' insofar as people draw inferences (correct or not) about a person based on these characteristics. Channels of nonverbal behavior that are defined dyadically include interpersonal distance (proxemic behavior), other-directed touch, gaze, and synchrony (coordinated position or movement). Behaviors that are not easily classifiable as either nonverbal or verbal include verbal interruptions, pauses and hesitations, dysfluencies in speech, verbal reinforcers ('mmmm,' 'uh-huh'), and manual language systems such as American Sign Language. Finally, some uses of physical objects are also studied as nonverbal communication, for example clothing, hairstyles, and makeup; similarly, environmental features such as room and seating design are often considered 'nonverbal' insofar as they influence people's behavior. Cataloging of nonverbal behaviors into channels can obscure important similarities in function across channels. For example, both saying 'mmm-hmmm' and nodding can serve the function of encouraging the speaker to continue. Although it is commonly assumed that the main function of nonverbal behavior is to express or reveal emotion, this is only one of a number of distinct functions. Another emotional function is to regulate and modulate one's own emotional states via internal feedback processes. Nonemotional functions include conveying interpersonal attitudes (such as friendliness, insult, or dominance), coordinating turn-taking between people in conversation, and facilitating one's own speech production.

4. Basic Approaches to the Study of Nonverbal Communication

Broadly speaking, there are two main approaches to the study of nonverbal behavior. The first is normative, concerned with describing typical nonverbal behavior patterns, including the meanings that are ascribed to those patterns, between and among groups of people, such as different cultures, ages, and genders. The second approach is concerned with individual differences, that is, the investigation of variation among individuals within groups. An example of this ap-

proach would be the study of personality correlates of smiling. Within both the normative and individual-differences approaches, research can focus on emitted behavior or on accuracy of nonverbal communication. The study of emitted behavior consists of measuring and describing the specific nonverbal behaviors produced by people in different roles, relationships, situations, and groups, and what functions and impact they have. Accuracy of nonverbal communication is further divided into decoding accuracy or the accuracy with which a person can interpret the meanings of nonverbal cues expressed by others, and encoding accuracy or the accuracy with which others can interpret one's own emitted cues. Encoding accuracy can be further broken down into intentional encoding and spontaneous encoding, the former referring to success in deliberately conveying messages via nonverbal cues and the latter referring to the transparency of one's expressions even when not consciously trying to communicate. Often, people unconsciously mimic others' expressions. Because expressive mimicry occurs even in infants, it has been suggested that these vicarious processes are innate and help to establish the basis for emotion recognition and emotional self-knowledge.

5. *Methodology*

Nonverbal communication research utilizes both observational and experimental methods, in both field and laboratory settings. Research may also be either qualitative (i.e., using ethological or participation observation methodology as in Eibl-Eibesfeldt 1975 and Goffman 1959) or quantitative. A wide variety of coding methods exist for quantifying nonverbal behavior in various channels of communication. Quantification is typically done in terms of frequency of occurrence or duration. When research is performed in controlled settings, the use of audio and videotape recordings allows for complex measurement systems to be applied. An important distinction is between descriptive and inferential assessment. In a descriptive coding system, particular behaviors are quantified through coding, rating, or mechanical measurement (e.g., number of hand movements per minute, average duration of gaze at partner, rating of extent of smiling, average vocal amplitude). In an inferential system, interpretations are made by observers; examples would be ratings of anger, anxiety, deceptiveness, or dominance.

Inferential ratings can be valuable in that they capture complex patterns of behavior in an overall interpretation (provided that interobserver reliability is adequate). It is also possible to ask how descriptive behaviors combine to create a global impression. For example, speaking in a loud voice, touching the other person, and engaging in relatively less gazing at the partner while listening (compared to while speaking) all contribute to the inference of dominance. When investigators wish to establish the validity of inferences about nonverbal cues, they must establish a criterion for deciding what the cues mean. Because nonverbal cues are often emitted outside of the encoder's conscious awareness, one often cannot rely on the encoder's self-report as a criterion; moreover, in many research settings (e.g., in public), it is not possible to ask encoders about the meanings of their behaviors. A common alternative is the use of the consensus of observers as the criterion; for example, if observers agree that an expression looks sad, then 'sadness' becomes the operational definition of the expression's meaning. Another method is to identify the meaning of expressions on the basis of situational context; for example, the facial expression of a person who is confronted with a dish of rotting meat would be assumed to reflect 'disgust.' Finally, it is possible to identify emotional meanings through the analysis of facial muscle movements. In the Facial Action Coding System (Ekman and Friesen 1978), emotional expressions are identified based on the patterns of activated facial muscles. 'Happiness,' for example, is signaled by certain movements of the Zygomatic major and Orbicularis oculi muscles (which pull out the mouth and crinkle the eyes, respectively). In order to study vocal nonverbal cues separately from verbal cues, investigators use content-free speech. One technique to accomplish this involves asking encoders to recite neutral-content words or phrases (such as affectively ambiguous sentences or the alphabet) while trying to convey different nonverbal messages. Less artificial is electronic filtering, wherein the highest frequencies are removed from an audiotaped voice, rendering the words unintelligible while the expressive qualities that result from amplitude, frequency, contour, hesitations, and other acoustic properties remain audible. Often, it is advantageous for investigators to develop standardized cue sets in the form of photographic slides, videotapes, or audiotapes, which can be reused with many observers. The most common application of this approach is to measure individual differences in decoding accuracy. Once the investigator determines what is the 'correct answer' for each item in the set (for example, which emotion is being conveyed), it is a simple matter to use the items as a test on which each test-taker's accuracy score is the number of items judged correctly. Several tests of this nature have been developed and validated (e.g., the Profile of Nonverbal Sensitivity; Rosenthal et al. 1979).

6. *Some Established Findings*

6.1 *Meanings Associated with Nonverbal Cues*

As stated earlier, nonverbal cues often do not have clear meaning in isolation from other cues and information about the circumstances. However, some

generalizations can be made. The face is well studied as the source of information on emotions (Ekman 1982). Patterns of facial muscle movements have been identified for at least six basic emotions—happiness, sadness, anger, fear, surprise, and disgust. These patterns produce highly consistent emotion recognition in many cultures of the world. This, in combination with evidence from primates and infants, suggests a biological basis for emotional expression and recognition. Although facial expressions often reveal felt emotion, they are also used for social signaling. Expressions are often used as messages of affiliation, encouragement, empathy, or threat; they also help to regulate and punctuate the flow of conversation in dialogue. The social signaling function of nonverbal communication includes the use of nonverbal cues for deception. In general, decoders are not highly accurate in detecting deception from nonverbal cues. The nonverbal cues that are valid indicators of deception are not necessarily those that people believe are associated with deception. Movements of the eyes signal attention; in friendly interactions, more gaze signifies a more positive attitude. More gazing, more smiling, closer interpersonal distance, and more direct body orientation all convey psychological immediacy. Hand gestures emitted during speech are used in part as facilitators of lexical access (retrieving the words one wants to use in speech). They can also serve nonaffective interpersonal functions, for example, turn coordination and the signaling of ellipsis, help seeking, and digression, as well as informational functions (for example, when the hands are used to depict the shape or size of an object). Some movements of the hands, particularly self-touching, as well as erect or fidgety body movements and higher vocal pitch, are related to anxiety or stress.

6.2 *Group Differences*

Gender differences in nonverbal communication are well established (as summarized by J. A. Hall 1984). Some of the differences are present in children and adolescents as well as in adults, although in general few of the male–female differences are large in absolute magnitude. Women smile, nod, and gaze more than men; their faces are more expressive; they establish closer interpersonal distances and are approached closer than men; they touch each other more than men do; and they are more skilled in emotion expression and in decoding the expressions of others. There is considerable debate over the origins of these differences. Cultural differences have also been examined. (Most research reported in this article is based on North American behavior.) Debate has flourished over the question of cultural universality vs. cultural specificity, with anthropologists arguing for a high degree of cultural specificity and psychologists and ethologists arguing for more universality. In fact, both positions are correct depending on the question being asked. A given nonverbal behavior may have similar meanings in many places (e.g., smile means happiness, scowl means threat), but the degree to which that behavior is emitted may vary across cultures. The term display rules is used to describe local norms for when, and to what extent, and by whom, different nonverbal behaviors are to be used (Ekman 1982). Often observed is the difference between contact and noncontact cultures, those that encourage physical and sensory involvement and those that do not (E. T. Hall 1966). An example would be reduced frequency of interpersonal touch and greater interpersonal distances in Northern European vs. Mediterranean, Middle Eastern, or Latin American countries. Gestural emblems, mentioned earlier, which often have meanings associated with sex or insult, show marked cultural differences in meaning, producing many opportunities for embarrassment between communicators from different cultures. Differences in emitted behavior and in communication accuracy have also been documented with respect to mental illness, pain and physical illness, and age.

6.3 *Individual Differences*

With respect to individual differences, the main topics are nonverbal correlates of personality and accuracy of nonverbal communication. The personality trait of extraversion, for example, has several nonverbal correlates, including louder and more fluent speech, closer interpersonal distances, and heightened levels of gaze. Individuals who are higher on decoding accuracy have been shown to be better adjusted, less hostile and manipulating, more interpersonally democratic, less dogmatic, and more cognitively complex; they also receive higher ratings of interpersonal sensitivity from friends and higher supervisors' ratings of clinical ability.

7. *Future Directions*

Some of the important questions remaining to be answered in future research are the following. (a) Origins of nonverbal communication abilities. Although nonverbal abilities can be measured and are related to social and psychological functioning, very little is known about how and why differences in ability develop. (b) Process of nonverbal cue judgment. The judging of nonverbal cues has been called an 'automatic' or 'intuitive' process simply because it often occurs rapidly and without much conscious deliberation. Future research will address nonverbal cue judgment in terms of cognitive processes (e.g., attention, memory) in order to understand not only what inferences people make about nonverbal cues but also what rules they apply and how they do so. (c) Dynamic cues and cue configurations. Most research to date has focused on one nonverbal channel at a time, one person at a time, and temporally static cues.

In the future with the help of computer technology, research will more often address how multiple channels of communication combine to convey meaning, the nature of nonverbal behavior between interacting partners or in groups, and nonverbal behavior as it changes over time within interactions.

See also: Communication and Social Psychology; Communication: Philosophical Aspects; Emotion and Expression; Gesture in Linguistics; Personality and Marriage; Pragmatics: Linguistic; Verbal Communication, Psychology of

Bibliography

Darwin C 1965 *The Expression of the Emotions in Man and Animals*. University of Chicago Press, Chicago, (originally published in 1872)

DePaulo B M 1992 Nonverbal behavior and self-presentation. *Psychology Bulletin* **111**: 203–43

DePaulo B M, Friedman H S 1998 Nonverbal communication. In: D T Gilbert, S T Fiske, G Lindzey (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, pp. 3–40

Eibl-Eibesfeldt I 1975 *Love and Hate*. Schocken, New York

Ekman P (ed.) 1982 *Emotion in the Human Face*, 2nd edn. Cambridge University Press, Cambridge, UK

Ekman P, Friesen W V 1969 The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica* **1**: 49–98

Ekman P, Friesen W V 1978 *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday/Anchor, Garden City, NY

Hall J A 1984 *Nonverbal Sex Differences: Communication Accuracy and Expressive Style*. The Johns Hopkins University Press, Baltimore, MD

Hall E T 1966 *The Hidden Dimension*. Doubleday, Garden City, NY

Knapp M L, Hall J A 1997 *Nonverbal Communication in Human Interaction*, 4th edn. Harcourt Brace, Fort Worth, TX

Krauss R M, Chen Y, Chawla P 1996 Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? *Advances in Experimental Social Psychology* **28**: 389–449

Rosenthal R, Hall J A, DiMatteo M R, Rogers P L, Archer D 1979 *Sensitivity to Nonverbal Communication: The PONS Test*. The Johns Hopkins University Press, Baltimore, MD

Russell J A, Fernandez-Dols J M (eds.) 1997 *The Psychology of Facial Expression*. Cambridge University Press, Paris

Scherer K R, Ekman P (eds.) 1982 *Handbook of Nonverbal Behavior Research*. Cambridge University Press, New York

J. A. Hall

Nonviolence: Protective Factors

Violent behavior is the outcome of complex transactions between biological, psychological, and social factors under specific situational conditions. Although the typical manifestations of aggressiveness change from childhood to adulthood, interpersonal differences seem to remain relatively stable over time. This is particularly due to a subgroup of early starters who are life-course-persistent in their antisocial behavior and frequently engage in serious and violent offending (Moffitt 1993).

Prospective longitudinal studies have revealed well-replicated causes and risks for such a development (Loeber and Farrington 1998; Loeber and Stouthamer-Loeber 1998). However, even the best predictions explain only a moderate amount of variance. Many children who are exposed to high risk do not develop serious aggressive behavior. Others who seem to enter the early-starting pathway leave it again after some time. Although these processes of abstinence and desistance are very important for explanation and prevention, far less research has been directed toward the issue of nonviolence in high-risk youngsters than vice versa (Lösel and Bender 2001). These studies are part of the more general research on resilience, that is a process of positive adaptation in spite of significant adversity (Luthar et al. 2000). However, research in this field is subject to various conceptual and methodological problems.

1. Conceptual and Methodological Issues

Major problems are the definition and operationalization of risk, the ambiguity of protective factors, and the multidimensionality of behavioral outcomes. Resilience is not a static personality characteristic. The positive outcome of nonviolence may vary over time and developmental circumstances depending on the presence or absence of risk and protective factors. However, these risk and protective factors are not different classes of variables but often the negative vs. positive pole of the same variable (e.g., low vs. good intelligence). Risk can be defined by objective variables (e.g., low SES) or subjective indicators (e.g., violence-prone attitudes). In the latter case, protective mechanisms may be partially confounded with both risk and behavioral outcome. Similarly important is the degree of risk; without a substantial risk, a good outcome may be misinterpreted as a result of a protective mechanism. In contrast, the more risks are present, the less variance is left over for protective effects.

Protective factors are variables that differ between groups of comparable high risk who show positive vs. negative behavioral outcome (Rutter 1985). A good example of research is the Kauai Study that investigated resilient and deviant children from early childhood to adulthood (Werner and Smith 1992). Other authors suggest that a factor should be labeled protective only if it reduces the problem behavior when risk is high but has no impact when risk is low (Stattin et al. 1997). In both concepts, the search for protective mechanisms refers to interaction effects (for other methodological

concepts, see Luthar et al. 2000; Masten et al. 1990). However, the theoretical interpretation of such effects may be difficult when multiple risks or protective variables are combined. This is often the case because single variables show only small correlations with later violence.

Given a specific risk constellation, even the same pole of a variable may have both risk and protective effects depending on the kind of outcome. For example, shyness in childhood may be a protective factor against violence but be a risk for anxiety and depression. Therefore, risk and protective factors should not be conceptualized too broadly, but it must always be asked: 'risk for and protection against what type of problem behavior?' Sometimes, a positive behavioral outcome is defined not only by the absence of psychopathology but also by indicators of particularly good adaptation and competence. In the present paper, positive outcome is defined by the absence of violence or other serious offending in spite of various risk constellations. This contains both processes of general abstinence and desistance after a period of violence.

2. *Examples of Protective Mechanisms*

The development of persistently antisocial behavior can be conceptualized as the outcome of a chain reaction of multilevel, biopsychosocial influences from childhood to adolescence (see Fig. 1).

The relations do not represent a necessary causal sequence, and most of the constructs reveal only low correlations with each other. Although the cumulation of factors increases the probability of violence, even multiple risks may not lead to serious problem behavior when one or more links in the chain are 'broken' because of positive personal or social resources. The following factors may exert such a protective function.

2.1 *Biological Factors*

Biological risks for aggressive behavior have been found on various levels, for example: behavior genetics, autonomic underarousal, functional deficits in the prefrontal brain, hormonal influences such as high testosterone, neurotransmitter processes, prenatal factors such as maternal smoking, or perinatal complications (Raine et al. 1997). However, we can rarely assume simple causal relationships. Complex interactions between biological and social factors seem to be more the rule than an exception. For example, low heart rate level (HRL) is primarily a risk for aggression in children from relatively normal family backgrounds. Pregnancy and birth complications are particularly related to violence when combined with maternal rejection, low SES, or family instability.

Similarly, biological variables can contribute to protective mechanisms in the context of given social risks. For example, autonomic arousal as indicated by high HRL or enhanced skin conductance (SC) reduce the risk of criminality for males with a criminal father, for boys from a low SES background, or for boys from families with poor parent–child relationships. It was also found that boys who desisted from antisocial behavior had a higher HRL, higher SC arousal, and better SC conditioning than persistent criminals. These results can be interpreted within neuropsychological theories on imbalances between the behavior activation and inhibition system or on prefrontal dysfunctions in aggressive individuals. Perhaps, the stress hormone cortisol may have a similar inhibiting effect on both the testosterone level and aggressive behavior. Enhanced salivary cortisol accompanies shyness, anxiety, depression, and high arousal in some studies. Thus, biological protective mechanisms may reflect an inhibited temperament that is associated with a reduced risk of antisociality. However, there are also comorbidities between externalizing and internalizing problems and specific forms of overcontrolled aggression.

2.2 *Temperament Characteristics*

Impulsivity, hyperactivity, negative emotionality (neuroticism), sensation seeking, and risk taking are personality dispositions for violence. These and related characteristics describe a child with a 'difficult' temperament. In contrast, an 'easy' temperament seems to have a protective function (e.g., Werner and Smith 1982). Children who did not develop aggressive and other antisocial behavior in spite of social and individual risks had advantageous dispositions such as positive mood, low impulsivity, flexibility in adaptation, sociability, and regularity of biological functions. According to Moffitt et al. (1996), boys who behaved untypically well in both childhood and adolescence showed the least difficult temperament at ages 3 and 5 years. However, positive temperament traits should not be interpreted as innate and stable resiliency because those boys who recovered from early antisocial behavior improved in temperament.

Children with a difficult temperament are also more frequently victims of parental hostility, neglect, and physical abuse. This, in turn, increases the risk of later aggressive behavior. In contrast, an easy temperament may reduce or avoid such negative family influences. The child's personality can also help them to cope successfully with unfavorable social experiences. Both perspectives are suggested by studies on disadvantaged and maltreated children. For example, Cicchetti and Rogosch (1997) used Block's personality concepts of ego overcontrol (e.g., rigid impulse containment), undercontrol (e.g., impulsivity), and ego resiliency (dynamic capacity to modify the level of ego control

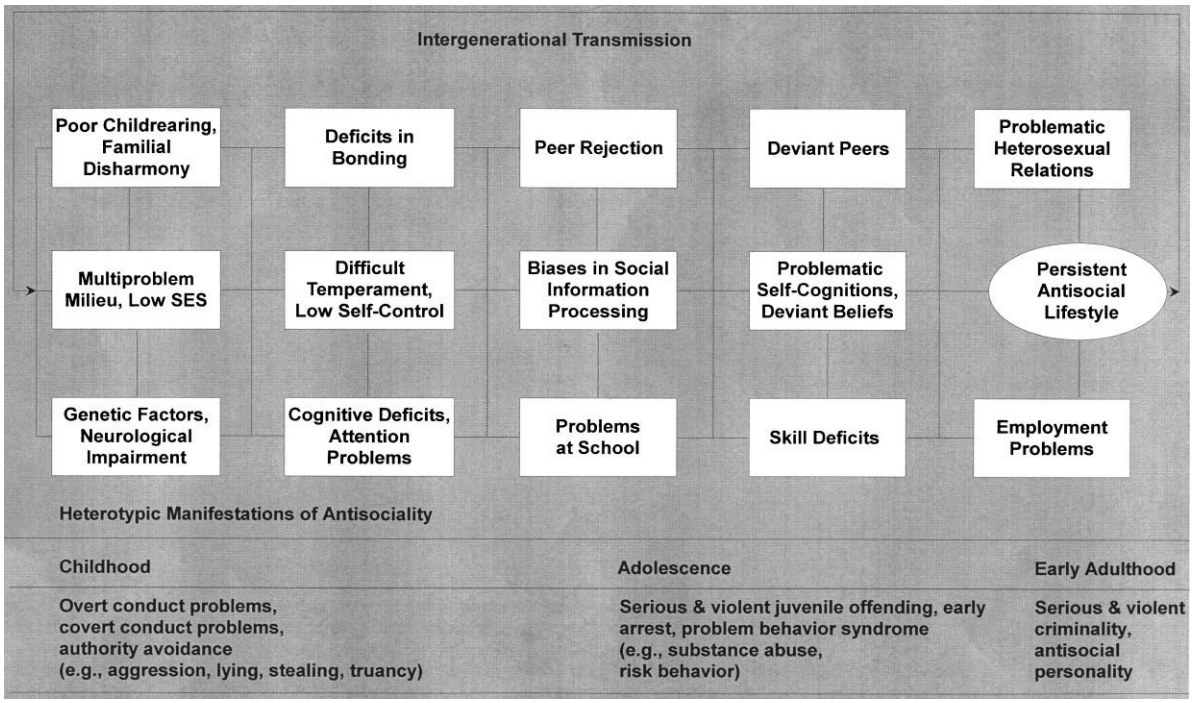

Figure 1
A simple model of cumulating risks in development of persistent antisocial behavior

according to environmental demands). These authors found that maltreated children who were more ego resilient or more overcontrolled adapted better and were less a target of continued abuse. However, inhibited children may sometimes 'pay the price' for better protection against aggressive behavior by exposure to a higher risk for internalizing problems.

2.3 Cognitive Competencies

Whereas low intelligence is a risk factor for persistently violent and delinquent behavior, above-average or high intelligence can have a protective function (Werner and Smith 1982). This is both a main effect and a buffer against risks such as a criminal father, a multiproblem milieu, or earlier problem behavior. A protective effect of cognitive competencies is in line with theories on deficits in neuropsychological functioning, social information processing, and self-control in aggressive youngsters. Intelligent children are more able to plan their behavior, anticipate consequences, solve conflicts verbally, and develop alternatives to aggressive reactions (Rutter et al. 1998). Probably, it is not primarily general intellectual capacity but more competencies in planning, communication, and interpersonal problem solving that help to cope with social risks of antisocial behavior.

Individuals with good cognitive competencies may not only avoid aggression in concrete situations but also experience more positive reactions in a variety of social contexts. Thus they are able to construct a more favorable environment. For example, high-risk individuals who planned their lives more carefully had better school experiences and chose better partners than others from similar backgrounds (Quinton et al. 1993).

2.4 Childrearing and Attachment

Parenting behaviors such as aggression, rejection, inconsistency, or laxness are well-replicated risk factors for serious and violent offending in youngsters. In contrast, parental involvement, positive emotionality, firm supervision, as well as good and reflective communication may help children to adapt relatively well in spite of risk factors for antisocial development (Stouthamer-Loeber et al. 1993, Werner and Smith 1992). Safety and stability in the family also seem to moderate negative environmental influences such as neighborhood violence (Richters and Martinez 1993). However, results on interactions between family variables and community violence are mixed.

On the one hand, protective functions of parenting behavior operate through reinforcement and the modeling of nonaggressive behavior. On the other hand,

emotionally responsive and supporting parental behavior increases the probability of secure attachment as an internal working model for the future. Attachment to a competent caregiver promotes a structured, predictable, and regulated environment that is relevant for general neuronal functioning and the regulation of aggressive behavior. Although there is some controversy about the stability and generality of attachment types, secure bonding proves to be a protective factor against various risks for behavioral problems. These include, for example, family conflicts, child abuse, multiple life burdens, a mentally ill parent, parental divorce, or a deprived lower-class milieu (Lösel and Bender 2001).

Protective functions of an emotionally positive and accepting as well as norm-oriented and adequately demanding upbringing are not restricted to the family context. Such an 'authoritative' climate in school or in residential care institutions can also have a positive effect on children at risk for aggression (Lösel and Bliesener 1994). This is important because these youngsters often come from families in which protective childrearing features are lacking.

2.5 School Achievement and Bonding

Low academic achievement and deficits in school bonding are replicated predictors of violent offending (Hawkins et al. 1998). It is not clear how far school problems reflect a cause, an effect, or merely a correlate of antisocial behavior. However, academic achievement and good school bonding can have a protective function when risks for antisociality are present. Prospective studies have shown that success at school, regular attendance, achievement orientation, parental interest in a school career, and higher education are related to nonviolence or a recovery from antisocial behavior in childhood (Moffitt et al. 1996, Werner and Smith 1992). Good school achievement and school bonding still seem to exercise a protective function against antisocial behavior beyond the effects of intelligence.

Positive experiences in school are a source of self-affirmation that helps to compensate for negative experiences in the family. A good school career is also an indicator for bonds to society and increases the chance for nondeviant opportunities in life. Similar effects are found for good work behavior, motivation for vocational improvement, and job stability. Such characteristics lead to positive turning points of previous delinquents in late adolescence or young adulthood (Sampson and Laub 1993).

2.6 Peer Groups and Social Networks

Resilience research suggests that social support and involvement in social networks have a protective effect against various emotional and behavioral problems. However, in the case of aggressive behavior, this general relation needs to be differentiated. Whereas support from nondeviant family members, relatives, teachers, educators, or friends may function in a protective way (Werner and Smith 1992), belonging to a delinquent peer group is a strong predictor of violent offending and gang membership (Thornberry 1998).

Whether the peer network has a risk or a protective effect depends on the characteristics of the individuals involved. Having nondelinquent peers protects against the onset or persistence of antisocial behavior, whereas delinquent friends promote the opposite (Stouthamer-Loeber et al. 1993, Moffitt et al. 1996). Because individuals join groups according to their own interests and attitudes, deviant peers may be both a precursor and a result of one's own antisociality. Children with early onset, undersocialized-aggressive behavior seem to be rejected by 'normal' youngsters and thus more likely to join groups with antisocial problems.

How far the peer group exerts long-term positive or negative influences on the development of antisocial behavior depends not least on the quality of relations to family and school. Satisfaction with the social network also seems to play a role. Both antisocial and nondeviant adolescents continue in their respective behavior when they are satisfied with their network (Bender and Lösel 1997). However, not being embedded in a social network has a protective effect for those with previous antisociality, but a risk effect for those who were nondeviant before. Similarly, heterosexual relations may have a risk or a protective effect depending on the individual age, gender, and social context.

2.7 Self-Related and Social Cognitions

The literature on resilience emphasizes a positive self-concept and self-efficacy as a relatively general and strong protective factor. Although this view is supported partially by results on aggressive behavior (e.g., Werner and Smith 1992), self-esteem plays a more ambiguous role in this field. Subgroups of violent individuals do not have a negative but an inflated self-esteem (Baumeister et al. 1996). This is not just the case in adult psychopaths. Already aggressive children have more idealized ratings of their own competence and relationship quality than nonaggressive children. Such youngsters may show aggression more readily in response to perceived negative feedback from others. Idealized and defensive self-schemata contribute to forms of social information processing that trigger and reinforce aggressive behavior (e.g., hostility bias, egocentric goal setting, positive evaluation of aggressive reactions (Crick and Dodge 1994).

That specific self-related and social cognitions have a protective effect against aggressive behavior is

supported by findings on offender treatment. Here, cognitive–behavioral approaches show particularly positive results. Such programs address social perception, self-critical thinking, anger management, social perspective taking, victim awareness, creative thinking, interpersonal problem solving, social skills and attitudes. Ethical and religious beliefs can also have a protective function against antisocial behavior.

2.8 Neighborhood and Community Factors

Although it is methodologically difficult to separate individual and family characteristics from neighborhood and community characteristics, the latter seem to harbor a specific risk for antisocial behavior. In analogy to this, the protective model assumes that positive features of the broader social context, such as low poverty, low violence, low unemployment, many intact families, and high cohesion, may buffer negative influences on the individual and microsocial level. For example, Kupersmidt et al. (1995) found no main effect of neighborhood factors, such as income, ethnicity, and family structure, on childhood aggression. However, there were various interactions suggesting that the same neighborhood may be protective for one family type while functioning as a risk or neutral factor for others. Middle-class neighborhoods, for example, had a protective effect only on the aggressive behavior of black children from low-income, single-parent homes. Wikström and Loeber (2000) only found a negative effect for disadvantaged neigborhoods on older children with low individual or family risks.

Neighborhoods with low violence may have a protective effect because they provide fewer negative role models and reinforcements for aggressive behavior (Thornberry 1998). However, subjective experiences of the local climate have also to be considered. For example, Seidman et al. (1998) reported more adolescent antisocial behavior in neighborhoods with moderate poverty and violence than in neighborhoods that ranked high on such structural features. This relation held primarily for middle (not young) adolescents who perceived their neighborhoods as more hassling and less cohesive. There was also a subgroup with low antisociality who experienced themselves as disconnected to their neighborhood. This may indicate a protective effect of the family's actively reduced involvement in a disadvantaged and violent neighborhood.

2.9 Conclusions and Perspectives

Research on protective factors and mechanisms is not the counterpart of traditional risk research but an integrated part of differentiated explanations of the natural history of violent and nonviolent behavior. Protective functions against violence and other serious antisocial behavior may emerge on and between biological, psychological, and social levels at various ages. Although relatively few studies are designed specifically to detect protective mechanisms against violence, the following factors seem to contribute to nonviolence in spite of serious risks:

(a) Genetic dispositions and high autonomic arousal
(b) An easy or inhibited temperament
(c) Personality dispositions of ego resiliency
(d) Above-average intelligence and good planning
(e) Emotionally secure bonds to the mother or other caregivers in or outside the family
(f) Affiliation to nonaggressive peers and (for specific risk groups) some social isolation
(g) Supportive relations to nondeviant friends or heterosexual partners
(h) Warmth, consistency, and supervision in child-rearing in the family, at school, or in other institutions
(i) Adult role models of prosocial behavior and resilience under difficult developmental circumstances
(j) Bonding to school and experiences of success at school
(k) Vocational achievement and employment stability
(l) Experiences of self-efficacy in nonaggressive activities and a good but not inflated self-esteem
(m) Cognitive schemata, beliefs, and information processing that are not aggression-prone
(n) Experiences of meaning and structure in one's life
(o) Neighborhoods that do not concentrate too many high-risk families, are nonviolent, and socially integrated in a nondeviant mode.

Probably, none of these factors are protective under all kinds and degrees of risk. Some counteract violence only under specific circumstances. The protective function of others may depend on their dosage, age periods, gender, or co-occurrence with other individual or social resources. More research is needed to understand the interaction of various levels of risk and protection.

Further levels of influence need also to be integrated. For example, low exposure to violence in the media, effective fire-arm control, and generally low violence in the society may function in a protective mode. Target hardening, victim assertiveness, informal neighborhood control, and other forms of situational crime prevention can also counteract the behavioral manifestation of dispositions to aggressiveness.

Last but not least, it is necessary to integrate experimental research on prevention and intervention more strongly with the correlational-longitudinal research on protective factors (Farrington and Coid 2001). Although successful programs seem to address similar factors as those that are protective in the natural environment, there is not yet much exchange between the two areas of research.

See also: Aggression in Adulthood, Psychology of; Attachment Theory: Psychological; Networks: Social; Self-regulation in Adulthood; Self-regulation in Childhood; Violence and Media; Violence, History of; Violence in Anthropology; Violence: Public

Bibliography

Baumeister R F, Smart L, Boden J M 1996 Relation of threatened egotism to violence and aggression: the dark side of high self-esteem. *Psychological Bulletin* **103**: 5–33

Bender D, Lösel F 1997 Protective and risk effects of peer relations and social support on antisocial behaviour in adolescents from multi-problem milieus. *Journal of Adolescence* **20**: 661–78

Cicchetti D, Rogosch F A 1997 The role of self-organization in the promotion of resilience in maltreated children. *Development and Psychopathology* **9**: 797–815

Crick N R, Dodge K A 1994 A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin* **115**: 74–101

Farrington D P, Coid J (eds.) 2001 *Prevention of Adult Antisocial Behavior*. Cambridge University Press, Cambridge, UK (in press)

Hawkins J D, Herrenkohl T, Farrington D P, Brewer D, Catalano R F, Harachi T W 1998 A review of predictors of youth violence. In: Loeber R, Farrington D P (eds.) *Serious & Violent Juvenile Offenders: Risk Factors and Successful Interventions*. Sage, Thousand Oaks, CA

Kapersmidt J B, Griesler J C, De Rosier M E, Patterson C J, Davis P W 1995 Childhood aggression and peer relations in the context of family and neighborhood factors. *Child Development* **66**: 360–75

Loeber R, Farrington D P (eds.) 1998 *Serious and Violent Juvenile Offenders: Risk Factors and Successful Interventions*. Sage, Thousand Oaks, CA

Loeber R, Stouthamer-Loeber M 1998 Development of juvenile aggression and violence: Some common misconceptions and controversies. *American Psychologist* **53**: 242–59

Lösel F, Bender D 2001 Protective factors and resilience. In: Farrington D P, Coid J (eds.) *Prevention of Adult Antisocial Behavior*. Cambridge University Press, Cambridge, UK (in press)

Lösel F, Bliesener T 1994 Some high-risk adolescents do not develop conduct problems: A study of protective factors. *International Journal of Behavioral Development* **17**: 753–77

Luthar S S, Cicchetti D, Becker B 2000 The construct of resilience: A critical evaluation and guidelines for future work. *Child Development* **71**: 543–62

Masten A S, Best K M, Garmezy N 1990 Resilience and development: Contributions from the study of children who overcome adversity. *Development and Psychopathology* **2**: 425–44

McCord J, Tremblay R (eds.) 1995 *Preventing Antisocial Behavior: Intervention From Birth Through Adolescence*. Guilford Press, London

Moffitt T E 1993 Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review* **100**: 674–701

Moffitt T E, Caspi A, Dickson N, Silva P, Stanton W 1996 Childhood-onset versus adolescent-onset antisocial conduct problems in males: Natural history from ages 3 to 18 years. *Development and Psychopathology* **8**: 399–424

Quinton D, Pickles A, Maughan B, Rutter M 1993 Partners, peers and pathways: Assortative pairing and continuities in conduct disorder. *Development and Psychopathology* **5**: 763–83

Raine A, Farrington D P, Brennan P, Mednick S A (eds.) 1997 *Biosocial Bases of Violence*. Plenum, New York

Richters J E, Martinez P E 1993 Violent communities, family choices, and children's chances: An algorithm for improving the odds. *Development and Psychopathology* **5**: 609–27

Rutter M 1985 Resilience in the face of adversity. Protective factors and resistance to psychiatric disorder. *British Journal of Psychiatry* **147**: 598–611

Rutter M, Giller H, Hagell A 1998 *Antisocial Behavior by Young People*. Cambridge University Press, Cambridge, UK

Sampson R J, Laub J H 1993 *Crime in the Making: Pathways and Turning Points through Life*. Harvard University Press, Cambridge, MA

Seidman E, Yoshikawa H, Roberts A, Chesir-Teran D, Allen L, Friedman J L, Aber J L 1998 Structural and experiental neighborhood contexts, developmental stage, and antisocial behavior among urban adolescents. *Development and Psychopathology* **10**: 259–81

Stattin H, Romelsjö A, Stenbacka M 1997 Personal resources as modifiers of the risk for future criminality. *British Journal of Criminology* **37**: 198–223

Stouthamer-Loeber M, Loeber R, Farrington D P, Zhang Q, van Kammen W, Maguin E 1993 The double edge of protective and risk factors for delinquency: Interrelations and developmental patterns. *Development and Psychopathology* **5**: 683–701

Thornberry T P 1998 Membership in youth gangs and involvement in serious and violent offending. In: Loeber R, Farrington D P (eds.) *Serious & Violent Juvenile Offenders: Risk Factors and Successful Interventions*. Sage, Thousand Oaks, CA

Werner E E, Smith R S 1982 *Vulnerable but Invincible*. McGraw-Hill, New York

Werner E E, Smith R S 1992 *Overcoming the Odds*. Cornell University Press, Ithaca, NY

Wikström P-O, Loeber R 2000 Do disadvantaged neighborhoods cause well-adjusted children to become adolescent delinquents? *Criminology* **38**: 1109–42

F. Lösel

Normative Aspects of Social and Behavioral Science

Since their very beginning, genuine concern about the scientific character of social and behavioral sciences has continuously stimulated the development of sophisticated methodologies for this group of disciplines, but also shifted interest away from the variety of normative aspects ineluctably connected with them. Most important among these widely differing aspects are (1) object and objective of the social sciences and their bearing on methods and possible results, (2) the role of value judgments in social and behavioral science, and (3) the feasibility of rational argumentation about norms. In the sequel, the concept of norm will be used in a wide sense to cover norms followed

implicitly or explicitly, norms of conventional, methodological or ethical character, and norms of different levels of generality.

1. Constitutive Norms and Ethics of Science

Unquestionably, science neither originated, nor continued to exist, as a natural event; it has always been a purposeful enterprise, established for the benefit of mankind. Soon after the social sciences had been called into being, they were considered by some of its greatest exponents to fulfill this function best if they be kept 'value-free,' which, of course, did not mean to say that their function was itself without value. Regrettably, confusion about the proper place of value in the social and behavioral sciences has accompanied theoretical reflection on them since the nineteenth century until now (see Sect. 3).

Historically, social and behavioral sciences originated from the need for scientific guidance (and, if possible, foundation) of political efforts towards improvement of the economic condition and social status of sections of the population of Western nations. Up to this day, social workers, as well as many social and behavioral scientists, are motivated in their activities by a strong commitment to the amelioration of social states-of-affairs and the discovery and correction of social disadvantages. Generally, this does not inhibit the objectivity of their scientific work, to which its motivation, despite its normative character, remains external. Likewise, the choice of fields of research and of particular problems, however dependent on practical, mostly normative preconsiderations, will generally not influence the scientific character of investigations into the fields and problems chosen.

Yet, this independence from historical roots, individual motivation, and problem-guided selection of research areas must not blind us to the fact that the object of the social and behavioral sciences is indeed a peculiar one: to restrict the study of human action and behavior to properties and relations accessible by methods effective in the natural sciences would amount to forgoing the acquisition of additional knowledge available due to the human nature shared by investigator and investigated object. The world of experience is not static, it is essentially a world of becoming; and the world investigated by social and behavioral scientists, comprising human action and behavior, exhibits throughout a becoming, events directed towards certain ends, a change from something that is, to something that, in the agent's opinion, is to be. Purposeful action, guided by norms (for the choice of particular ends and of suitable means for their realization), distinguishes man from inanimate objects (and from most living beings), enriching the equipment of human investigators of human actions by the faculty of interpretive understanding. Even though the investigator and his or her subject may belong to fundamentally different cultural contexts and may follow mutually incompatible norms, they share the orientation towards purposes and norms in general.

Needless to say, social and behavioral sciences conform to the norms constitutive for science in general, such as commitment to impeccable objectivity, unbiased interaction with other persons as objects of social inquiry, and strict observance of methodological norms for scientific investigation (e.g., collecting and sampling of data, accurate data processing, complete and unmanipulated presentation of results). The ethics of science, however, has more and more transcended these normative aspects of the scientific community and is now mainly investigating the import of general moral considerations on the production and use of scientific knowledge, particularly the (individual and general, in some cases even global) consequences of the application of technological and medical knowledge. It will be clear that, despite the novelty of many pertinent problems (e.g., those connected with analysis of the human genome), this is merely an extension of ethical reflections on prohibitive or restrictive norms concerning experimenting with human beings, manipulation of their behavior, their reproduction, their physical and mental equipment, or of the use of scientific or technological knowlegde and results for the impairment or destruction of human lives, e.g., for purposes of eugenics, of euthanasia or of warfare. Among the normative aspects of social and behavioral sciences, one of the most important is that opened up by the insight that 'a procedure may satisfy criteria of methodological adequacy and be morally unjustified' (Beauchamp 1998, p. 578) at the same time.

2. Values and Value Judgments in Social and Behavioral Science

Traditionally, problems of the kind just mentioned have been dealt with in philosophy as questions of 'values.' Individuals often prefer an object, or a state of affairs, or (performing) an action, to other objects, states of affairs, or actions; they 'valuate' them positively, considering them more desirable than the other ones and declaring them to be 'higher' values. For example, truth is a scientific value, charity a Christian moral value, reliability (of a person, or of a report, or of a tool) a value in everyday life.

Values as reasons behind valuations and thereby behind actions are omnipresent in society and are a legitimate object of social and behavioral sciences. Investigators from these fields may or may not share the values of the persons whose (valuation-guided) actions they investigate, but they must by no means let these values influence the conduction or the results of their scientific work. There is no problem in scrutinizing the compatility of a person's values and to describe

and analyze the inner conflicts of this person in case they are incompatible, or to investigate (as an historian or as a sociologist) the controversies, open hostilities, or even material combats between adherents of conflicting values. Doubt would, however, be cast on the scientific character of such research if the investigator would 'take sides,' approving or disapproving of values and actions of his subjects, and even more so if the researcher would use his scientific findings for a denouncement of, or as apologetics for, some state of affairs or situation in which his subjects are involved.

Max Weber (1864–1920) propounded his famous demand for a value-free social science explicitly with the intention of clearly separating the (fully justified) domain of scientific investigation from the (equally justified, but fundamentally different) domains of social commitment and social policy. In his opinion, it should be possible to discern, in a social scientist's report on his research, the parts that may be accepted by any one independently of his or her valuations and to argue for changes of social situations, of laws etc. without abusing for this purpose the results of scientific research as if they cogently implied practical consequences. This last-mentioned step has frequently been repudiated as a typical case of the 'naturalistic fallacy' of overlooking that a normative conclusion can never be logically inferred from descriptive premisses alone (i.e., without at least one further normative premiss). However, the position of the 'value-supporters' is not that simplistic, as is shown by the debates in the *Werturteilsstreit* (the controversy over the role of value-judgments in social science and economics, see Weippert 1939) and in its partial resumption in the *Positivismusstreit* (see Adorno 1969). The differences of opinion concerned not only the role and legitimacy of value-judgments in science, but on a deeper level the (extension of the) concept of science, the concepts of foundation and justification, and the possibility of rational argumentation for or against normative propositions.

3. *The Feasibility of Argumentation about Norms*

Max Weber and his adherents in the *Werturteilsstreit*, as well as the 'critical rationalists' in the *Positivismusstreit*, were convinced that objectivity of normative propositions could never be reached since the necessary intersubjectivity seemed to them precluded by the essential subjectivity of all valuation. At most, norms could be shown to be inconsistent, or to be incompatible with other norms. The other side, representative of the 'critical theory' of the *Frankfurter Schule*, criticized the restriction of 'foundation' and 'justification' to deductive procedures and envisaged an ultimate foundation (*Letztbegrüendung*) of norms by the exhibition of necessary presuppositions of any speech acts whatever, i.e., of human communication in general. They failed to convince their opponents since, pragmatically, proving that something is necessarily presupposed does not entail evidence that it is well founded (Gethmann 1979). More recent approaches from the perspective of 'methodical constructivism' have shown how to reach intersubjectively acceptable foundations for norms relative to the best possible consensuses among all persons concerned, i.e., exposed to possible consequences of actions that are guided by the norms in question.

Apart from this, attempts have been made to reach consensus as to fundamental (or primordial) situations which constitute ultimate purposes or ends for any human living together, i.e., community and communication. The most fundamental ones derive from the *condition humaine* including the facts that men are born, have to grow up, have to reproduce within a limited period of reproductive potency, and will die some day; that they are reliant on other human beings and on material support, etc. Institutions of society, deliberately created for the fulfillment of such needs, may be valuated as suiting their purpose to a higher or lower degree. Even though one must admit the striking fact that human beings can willfully repudiate submission to even the most fundamental exigencies, there is no reason to deny the possibility of rational argumentation for or against the acceptance of certain norms and actions (in a group, or society, or culture) if other norms or actions have already been accepted (see, e.g., Greer 1969, p. 49).

See also: Conventions and Norms: Philosophical Aspects; Critical Theory: Frankfurt School; Ethics and Values; Ethnicity, Crime, and Immigration; Language and Animal Competencies; Medical Experiments: Ethical Aspects; Norms; Research Conduct: Ethical Codes; Terrorism; Value Pluralism; Values, Anthropology of; Values, Development of; Values: Psychological Perspectives; Values, Sociology of

Bibliography

Adorno T W (ed.) 1969 *Der Positivismusstreit in der deutschen Soziologie*. Luchterhand, Berlin

Albert H 1967 Theorie und Praxis. Max Weber und das Problem der Wertfreiheit und der Rationalität. In: Oldemeyer E (ed.), *Die Philosophie und die Wissenschaften*. Hain, Meisenheim am Glan, Germany; also in: Albert H 1972 *Konstruktion und Kritik. Aufsätze zur Philosophie des kritischen Rationalismus*. Hoffmann und Campe, Hamburg, Germany

Beauchamp T L 1998 Value judgments in social science. *Routledge Encyclopedia of Philosophy*. Routledge, London

De Gré G 1955 *Science as a Social Institution. An Introduction to the Sociology of Science*. Random House, New York

Gethmann C F 1979 Zur formalen Pragmatik der Normenbegründung. In: Mittelstrass J (ed.) *Methodenprobleme der Wissenschaften vom gesellschaftlichen Handeln*. Suhrkamp, Frankfurt, pp. 46–76

Greer S 1969 *The Logic of Social Inquiry*. Aldine, Chicago

Grunwald A, Saupe S (eds.) 1999 *Ethik in der Technikgestaltung. Praktische Relevanz und Legitimation*. Springer, Berlin
Kambartel F, Mittelstrass J (eds.) *Zum normativen Fundament der Wissenschaft*. Athenäum, Frankfurt
Lorenzen P 1984 *Normative Logic and Ethics*. 2nd edn. Bibliographisches Institut, Mannheim, Germany
Lorenzen P 1985 Autonomie und empirische Sozialforschung. In: Lorenzen P (ed.) *Grundbegriffe technischer und politischer Kultur*. Suhrkamp, Frankfurt, Germany, pp. 137–57
Thiel C 1972 *Grundlagenkrise und Grundlagenstreit. Studie über das normative Fundament der Wissenschaften am Beispiel von Mathematik und Sozialwissenschaft*. Anton Hain, Meisenheim am Glan, Germany
Weippert G 1939 Vom Werturteilsstreit zur politischen Theorie, *Weltwirtschaftliches Archiv* **49**: 1–100; also in Weippert G 1966 *Sozialwissenschaft und Wirklichkeit*. Vandenhoeck & Ruprecht, Göttingen, Germany

C. Thiel

Norms

Norms are a major focus of attention of all social sciences, particularly sociology, political science, economics, anthropology, and social psychology. Whereas these social sciences are interested in the empirical study of norms, jurisprudence and philosophy (moral philosophy and deontic logic) address the question of what 'good' norms are and how normative reasoning can be formalized. The literature on norms is so vast that a short article on this subject has to be restrictive. The following sections address the most important questions of the study of norms. These are, first of all, how norms are (or should be) defined and measured. Second, the central theoretical questions are how norms originate and, if they exist, what effects they have.

1. *The Definition of Norms*

The literature abounds with definitions of norms. An overview and discussion of many norm definitions is provided by Gibbs (1965), Biddle and Thomas (1966), Rommetveit (1953), and Williams (1968). Crawford and Ostrom (1995) give a detailed analysis of the 'grammar' of norms. In this section, the dimensions of the norms concept that most definitions include are described, and which definition of norms is preferable is discussed.

George C. Homans' definition of norms includes most of the dimensions that are ingredients of other definitions as well:

> A norm is a statement specifying how a person is, or persons of a particular sort are, expected to behave in given circumstances—expected, in the first instance, by the person that utters the norm. What I expect of you is what you ought to do (1974, p. 96).

Homans refers to four dimensions. One is *oughtness*: a norm is an expectation that some behavior ought to be (or ought not to be) performed. This is a basic characteristic that almost all norms definitions have in common. Definitions that include this criterion may be called *oughtness definitions*. This kind of definition is mostly used by sociologists.

A second dimension is that people are expected to behave in a certain way 'in given circumstances.' Thus, norms may be *conditional*. A norm such as 'you shall not kill' holds only in certain circumstances or for certain kinds of actors (for example, soldiers are allowed to kill in times of war). Homans' definition further implies that norms refer to *behavior*, not to beliefs or attitudes. In contrast, Morris' definition (1956) also includes beliefs and attitudes. Thus, demands that one should believe in the existence of god (a belief) or should love his parents (an attitude) are norms. Fourth, the *actors* (i.e., the norm-senders) who hold (or utter) a norm are individuals. In other definitions, the actors are collectives such as parliaments.

Two other criteria are used in many definitions of norms. Morris (1956) defines norms as 'generally accepted, sanctioned prescriptions…' (p. 610). Thus, an expectation is, by definition, only a norm if the expectation is *shared* by the members of some group, and if conforming to or deviating from the expectation is *sanctioned* with some probability.

It is important to note that a normative expectation is not identical to a preference. For example, I may not *like* when you smoke in my company, but I may think that you are *allowed* to smoke; in a similar vein, smoking may be prohibited but I may like your smoking.

A second type of definition—which may be called a *behavioral definition* of norms—states that a norm exists if there is a *behavioral regularity* and if *sanctioning* occurs with some positive probability in the case of not performing the respective behavior. 'Sanctioning' means that others impose a cost on a person if a behavior is or is not performed. For example, a behavioral regularity may be that people keep promises. If someone does not keep a promise and if the person is sanctioned with some probability a norm exists. *Refraining* from a behavior may be a behavioral regularity as well. If people usually do not smoke (a 'behavioral' regularity) and if smoking triggers sanctions there is a nonsmoking norm. Thus, 'behavior' implies 'refraining from behavior.' This definition is typically used in economic analyses of institutions. The latter term often refers to single norms, as the following definition illustrates: 'an institution is a regularity of behaviour or a rule that is generally accepted by members of a social group, that specifies behaviour in specific situations, and that is either self-

policed or policed by external authority' (Rutherford 1996, p. 182). 'Institution' often also denotes sets of norms as well as organizations (for a discussion see Calvert 1995, pp. 216–18).

If a norm exists according to the behavioral definition, there need not be any element of oughtness. Assume that smoking is allowed (i.e., there be no nonsmoking norm in the sense of the oughtness definition) and let nonsmoking be a behavioral regularity. Some nonsmokers do not like to be exposed to smokers. If nonsmokers get into contact with smokers it happens that smokers are sanctioned: nonsmokers move to other tables in restaurants or they express their anger about being disturbed; furthermore, nonsmokers may reduce private contacts with smokers. These actions may be costly to smokers, but they have nothing to do with a norm. Thus, although there is a behavioral regularity (most people don't smoke) and although there is some probability that deviations are 'sanctioned' no norm is involved. Only if 'sanction' is defined as a reaction to the violation of a norm (in the sense of the oughtness definition) the behavioral definition includes oughtness. But most proponents of the behavioral definition do not like the oughtness definition because they prefer a pure 'behavioral' definition that allegedly allows easier measurement of norms (see below).

There are other definitions of norms in the literature. For example, Ensminger and Knight (1997, p. 2) define norms as 'informal rules that structure behavior.' Thus, only if a rule causes behavior can it be called a norm. If this definition is applied it is no longer possible to explain when norms emerge because a norm exists by definition only if it causes behavior. Since such definitions are relatively rare they will not be discussed here.

Laws (that will not be addressed in this article) are a subset of norms: they are norms that are promulgated by a public institution. There is a growing literature that explores differences between norms and laws (see, e.g., Posner and Rasmusen 1999).

What definition is to be preferred? It is striking that despite the multitude of definitions in the literature, authors usually don't give reasons for preferring a certain definition over others. One important criterion for choosing a definition is the extent to which a definition refers to the explanatory problems social scientists want to solve. With regard to norms, social scientists are interested in explaining, among other things, when people accept normative expectations and under what conditions what kinds of sanctions are attached to normative expectations. Answering these questions is not possible if several dimensions of norms are included in a single complex definition. In those definitions, certain combinations of values are contained (such as existence of a behavioral regularity *and* a positive probability of sanctioning). Thus, only this combination of values can be explained. It is further important to know how the dimensions are

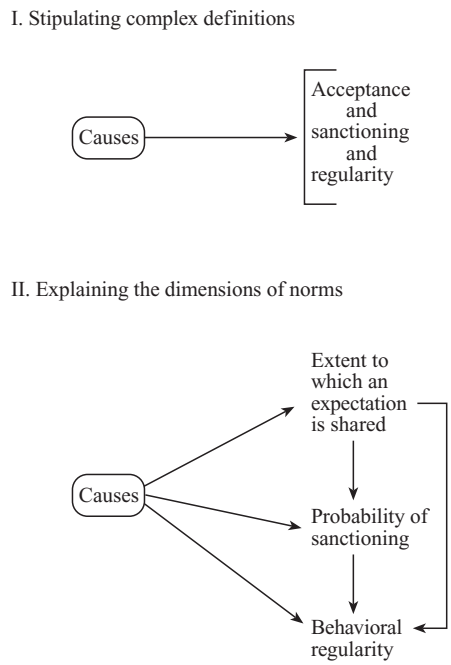

Figure 1
Different procedures for explaining norms

causally related. For example, if norms are accepted and if strong sanctions go with the norm, it is likely that the behavior that the norm prescribes is performed. These complex interdependencies cannot be addressed if several dimensions are incorporated into a single definition. The upper part of Fig. 1 shows that specifying complex definitions implies looking at certain values of the dimensions simultaneously. Treating the dimensions separately allows us to explain them and look at their interdependencies, as the lower part of Fig. 1 indicates (for a similar argument see Hechter and Opp in press).

It is further assumed that the definition used has theoretical import: it must permit the formulation of true and informative theories. The extent to which this criterion is fulfilled can only be judged if a theory with the respective concepts exists and has been tested. As long as such a theory does not exist the definition is a preliminary working definition.

2. *How Can Norms Be Measured?*

If historians want to get information about norms they will look at *documents* such as books about manners (see, e.g., Elias 1994). To learn about the norms of a cohort of young people, sociologists may analyze their diaries. The norms of voluntary associations may be taken from their statutes. One problem with these sources is to what extent the written norms are actually

in effect. Diaries may only reflect the norms of a specific group of middle class youths. Books of manners may reflect rather the ideals of the author than the manners that are effective. A second problem is the extent to which written documents state the conditions under which a norm holds. Members of a group may often not be aware of those conditions so that it is questionable whether documents state a given norm in all its complexity.

It seems that the method of *observation* is especially suitable for ascertaining norms according to the behavioral definition: behavior, in contrast to expectations, can be observed. But the *principal* possibility of observing behavior does not mean that the respective behavior can also *practically* be observed easily. Is it easy, e.g., to measure the frequency of promise keeping (or of breaking promises), of certain sexual behavior, or of smoking? It is practically impossible to observe these behaviors for a large group: numerous observers and intrusion in the private sphere would be necessary. Even if these difficulties can be overcome, recording the behavior (even on a video-tape) is not sufficient. A given behavior may have quite different meanings. For example, making a promise is a verbal behavior and the observer must know the language of a group in order to ascertain the extent to which promises are made or kept. Many kinds of sanctioning are verbal behaviors as well, and a given behavior may have different meanings. Shaking of the head may express disapproval but may also be a sign of astonishment. Thus, although behavioral regularities and sanctioning are in principle more accessible than expectations, practical measurement by observational techniques is extremely difficult. The conclusion is that the measurement argument that is often invoked for a behavioral definition is very questionable.

Surveys are probably used most frequently to measure norms. There are different measurement procedures. For example, Jackson (1966) measures the extent of approval or disapproval of a behavioral dimension (such as the extent of approval of the number of times a person speaks in an hour's session of a discussion group). If 'frequency of behavior' is plotted on the x-axis and average approval/disapproval of the members of a group on the y-axis, a curve ensues that shows what behavior is more or less approved or disapproved. This procedure also yields various statistical measures. Another measurement procedure is to formulate various items that are designed to capture, among other things, the conditionality of a norm. For example, one of the items used to measure protest norms reads: 'nobody can be expected to take part in political actions that might send her or him to jail' (Opp and Gern 1993, p. 668). A third measurement device is the factorial survey that seems best suited to ascertain the conditionality of a norm. The first step is to specify the conditions under which a norm holds. For example, people may endorse an obligation to protest if the expected repression by police action is low or if the expected success of the collective action is high. Thus, the dimensions 'expected repression' (with values no or yes) and group success (with values no or yes) may be important conditions for a felt obligation to protest. As a next step, descriptions of situations (the vignettes) are constructed from all possible combinations of the dimensions. Each of these situations is then judged by the respondents of the survey: they are asked to what extent a behavior (such as participating in a demonstration) is more or less obligatory in a given situation. The data enable the researcher to ascertain what effect each of the dimensions has on the overall normative judgment performing certain actions under a whole set of conditions (for details see Jasso and Opp (1997, Opp in press).

The well-known problems of surveys also hold for the measurement of norms. For example, people may not be able to report norms because they have never been faced with situations in which the norm holds. There may also be socially desirable answers when people hold deviant norms.

Qualitative researchers may in general object that norms are not given but are largely 'constructed' in specific situations (see, e.g., Fine in press). This implies that the measurement of norms is impossible because they do not exist. Survey research does not confirm this view. If norms do not exist the prediction is that factorial surveys yield a large number of missing values which did not turn out to be the case (Jasso and Opp 1997, Opp in press).

It is surprising that despite the centrality of norms in the social sciences, there is rather little effort to develop standardized measurement procedures of norms. A detailed comparative discussion of the different procedures is also missing.

3. *The Origins of Norms*

The major difficulty of reviewing theories about the emergence of norms seems to be the great variety of propositions, scattered in the literature, and the diversity of theoretical approaches. Recent reviews of the so-called new institutionalisms (see, e.g., Hall and Taylor 1996) testify this. Nonetheless, a review of various propositions and approaches indicates that there seems to be one basic idea that underlies all or at least most explanations of norms. This is the *instrumentality proposition* (Opp in press). The idea is that norms emerge if their emergence is in the interest of a collective of people. Norms are thus instrumental.

It is important to distinguish the instrumentality proposition from a *functionalist hypothesis* arguing that norms emerge if they contribute to the equilibrium or survival of a social system. This hypothesis is flawed because it is not possible to explain the existence of a cultural item by the effects that it would have if it

existed. Such functionalist arguments are still widely used in the literature (see the critique by Hardin (1980) of Ullmann-Margalit (1977)). In contrast, the instrumentality proposition holds that situations or events that jeopardize the realization of group goals of actors provide incentives to the actors to set up rules that are designed to make the group better off.

The instrumentality proposition is consistent with the widely held view that the emergence of norms is likely if there are *externalities*.These are actions which incur costs (negative externalities) or provide benefits (positive externalities) to other actors. Those who are affected by negative externalities have an interest in establishing norms that reduce the externalities. Those who cause positive externalities have an interest to establish a norm that compensates them. The basic reference for this theory is Demsetz (1967) (for a discussion see Eggertsson 1990). This theory underlies Coleman's explanation of norms (Coleman 1990, Chaps. 10 and 11).

Norms are public goods, i.e. goods that if they are provided are consumed (or enjoyed) by every member of a group even if the member has not contributed to their provision. For example, a non-smoking norm holds for everybody. Public goods theory which is consistent with the instrumentality proposition argues that 'groups exist in order to supply their members with some desired joint good. This good can be attained only if members comply with rules that are designed to assure its production' (Hechter 1987, p. 41; see also Oliver 1980, Heckathorn 1990, Coleman 1990). Thus, norms are instrumental to achieve group goals: norms are (second-order) public goods which are instrumental for providing (first-order) public goods. This theory suggests a certain mechanism of how norms emerge: they are created by *human design* such as laws.

Assuming that individuals may have biased knowledge about the effects of their actions on the emergence of a norm (a lawmaker may issue a law that is not accepted) or about the effects of a norm on the goals of a group (the norm has unintended negative effects), it may happen that norms are inefficient. Such norms may persist when the costs of changing a norm are high.

We find the instrumentality proposition also in the work of qualitative sociologists. Fine (in press) explains why there is a norm that restrains mushrooming. He notes that mushrooming is competitive: if someone picks a mushroom another one cannot. Mushroomers are thus exposed to a negative externality. In this instance, the norm 'first come, first serve' might apply. However, this solution would breed strain. Mushroomers were interested in good personal relations that were inconsistent with an untamed competition. Thus, the emergence of a norm about constraining the collection of mushrooms is, among other things, due to the goals of the group of mushroom collectors.

Norms are often not the result of human design but the unintended and sometimes undesired result of human action. The emergence of such 'spontaneous' norms is consistent with the instrumentality proposition as well. Assume that all of a sudden being exposed to smokers becomes a strong externality for nonsmokers. They will increasingly punish smokers. Nonsmokers may be interested in the production of a 'private good' (being free of smoke) and not in establishing a general nonsmoking norm. Due to extensive sanctioning behavior smokers will increasingly stop smoking and will get a bad conscience when they smoke in the presence of nonsmokers. A general nonsmoking norm has emerged. Various other mechanisms of the spontaneous emergence of norms are described in the literature. See particularly the literature on conventions (Schelling 1960, Lewis 1969, Sugden 1995). Other examples of the spontaneous emergence of norms are explored by Mackie (1996), Opp (1982, in press), and Horne (in press).

The instrumentality proposition assumes 'instrumental rationality': individuals have goals and they perceive that a norm x generates outcomes that attain their goals. Boudon (1997, pp. 19–22) argues that 'many moral feelings are not the product of instrumental rationality' (p. 20). For example, the act of stealing cannot be shown to have negative consequences; stealing is bad because it contradicts basic moral principles. However, the instrumentality proposition does not rule out that the goal of a group may be to preserve general social principles (or values).

The claim that the instrumentality proposition (and the theory of rational action in general) assumes 'instrumental rationality' or is 'consequentialist' may have different meanings. A behavior (or a norm) may be instrumental in the sense that it is *related* to phenomena in a noncausal manner. Assume that a person P performs action A and knows that this is in accordance with a norm N. The fact that A *corresponds to* N certainly cannot be construed as a causal effect of A. But it is a behavioral situation that may be an incentive for an individual actor to perform the respective action. Therefore, rational choice theory can be applied. In general, terms such as 'outcomes,' 'behavioral consequences,' and 'instrumentality' are used in a wide sense that does not include only (actual or perceived) *causal* effects of actions (or norms). This is confused, e.g., by Elster (1989, Chap. 3). For a critique see Opp (1997).

Are there other propositions about the emergence of norms that are not based on a 'rationalistic' process? Sociologists usually explain norms by the *socialization proposition* claiming that norms are transmitted (see, e.g., Boyd and Richerson 1985). This may occur spontaneously (a norm is imitated) or in a planned way (parents socialize their children). One question that the socialization proposition does not address is *under what conditions people transmit norms*. Whatever these conditions are, sociologists seem to argue that it is not the overall instrumentality of a

norm for a group. This is suggested by the fact that norms may be dysfunctional (Merton 1968, Chap. III). However, one would expect that parents who socialize children (or members of a group who socialize new members) are driven by incentives, i.e. by their personal goals (Coleman 1990, Chap. 11): parents teach their children good manners because they think that this is in the interest of the child or because they do not want to get a reputation as bad parents who are not able to teach their children how to behave. Thus, parents aim at providing a private good that might lead to a norm on the macrolevel.

The socialization proposition further does not say under what conditions the targets of socialization subscribe to a norm and how norms emerge before they are transmitted.

It might be argued that the previous propositions are inconsistent with a cultural approach such as symbolic interactionism. Hall and Taylor (1996) argue that a cultural approach does not reject that individuals are purposive, goal oriented, or rational (p. 949). But, they argue, the goals of individuals are conceptualized much broader: not only material well-being is at issue but individuals or organizations are assumed 'seeking to define and express their identity in socially appropriate ways' (p. 949). This is correct for a narrow neoclassical version of rational actor theory but not for a wider version (Opp 1999) which includes goals that go beyond material well-being. For example, preserving his or her 'identity' may mean that members of a group subscribe to the goal of acting in accordance with important others. Furthermore, invoking 'world views' as conditions for norm emergence is also not against a rational choice approach because actors may realize that certain norms contradict *values*, and this may be a cost that discourages the acceptance of a norm.

4. *The Effects of Norms*

Assume that a norm has come into existence. What effects does the norm have on people's actions? Role theory posits that norms are organized in roles that individuals take over. The message of the *homo sociologicus* is that individuals behave according to their roles (see, e.g., Biddle and Thomas 1966). There are several problems with this proposition. One is that behavior that is not regulated cannot be explained. For example, there is no norm prescribing that I should not go to a cinema and instead watch television. A second problem is that the proposition is wrong—people often deviate from role expectations—and that such deviations cannot be explained. A third problem exists if there are conflicting role expectations: the theory does not tell us what expectation is heeded. The basic assumption of this theoretical orientation is that individuals are socialized into their roles and follow them blindly.

An alternative to this oversocialized conception of man (Wrong 1961) is to assume that in situations in which norms hold there is choice—as in any other situation. Individuals realize that they have at least two options: to follow or not to follow a norm. Each of these options has certain costs and benefits, depending on the situation. If norms are internalized this means that breaking a norm ensues in a bad conscience which is a cost. The severity and probability of external sanctions, including informal sanctions, are costly as well (see, e.g., Epstein 1968, Hechter 1984, Heckathorn 1990, Posner and Rasmusen 1999). But other non-normative costs and benefits may instigate an individual to break a norm. What exactly the costs and benefits are that govern behavior if norms are given depends on the explanatory problem and cannot be said in advance. For example, when people consider divorce, marriage norms play a more important role in Catholic countries than in Protestant countries. In committing murder norms may be more important than in tax evasions. In both situations, the kinds of non-normative costs are very different. This argument suggests that *homo sociologicus* does not contradict *homo oeconomicus*.

Is there always a choice situation if a norm holds in a given situation? For example, if a person decides to take a subway will she or he consider every time to pay or not to pay the fare? It often happens that at some time a person makes a decision to act in a certain way and then performs the same action again without deliberating. The action then becomes a habit. But it also happens that deviating from a norm is so costly that this action is never considered. This type of situation indeed resembles the situation that role theory seems to portray.

When the effects of norms are considered, it is useful to distinguish between two types of effects. One is conformity: the emergence of a norm (and the pertinent sanctions) has the effect that more people perform the behavior that is in accordance with the norm than before. If, e.g., a government prohibits civil servants to smoke at their workplaces, smoking will decrease.

This example may illustrate a second type of *indirect* effect of norms. If fewer civil servants smoke, this may instigate others to stop smoking as well. This will reduce the number of people who get cancer and, thus, expenditures of health insurance companies will decrease. Their membership will be reduced. If civil servants have special consumer preferences, the profits of certain industries may rise. Indirect effects are often subtle and difficult to discover. For example, if companies have to pay increasing benefits to employed women (such as time off before and after pregnancy that the employer has to pay), the costs of employing women increase. The effect may be that in the long run unemployment of women increases and salaries decrease. This would certainly not have been intended by those who made the law. Increasing punishment for

some crimes may amount to a higher benefit of committing other crimes and thus may lead to a displacement effect. In general, changing norms may change the costs or benefits of various types of actions and may thus have unanticipated and undesired indirect effects. What the indirect effects of norms are cannot be said in advance. A detailed analysis of the incentives that prevail in the situation for which the effects of norms are to be explained is necessary.

See also: Coleman, James Samuel (1926–95); Collective Beliefs: Sociological Explanation; Conformity: Sociological Aspects; Control: Social; Institutionalization; Institutions; Integration: Social; Law and Everyday Life; Law: Anthropological Aspects; Law: Economics of its Public Enforcement; Law, Sociology of; Parsons, Talcott (1902–79); Policing; Sanctions in Political Science; Status and Role: Structural Aspects; Structure: Social; Values, Sociology of

Bibliography

Biddle B J, Thomas E J (eds.) 1966 *Role Theory: Concepts and Research*. Wiley, New York

Boudon R 1997 The present relevance of Max Weber's *Wertrationalität*. In: Koslowski P (eds.) *Methodology of the Social Sciences, Ethics, and Economics in the Newer Historical School From Max Weber and Rickert to Sombart and Rothacker*. Springer, Berlin

Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago

Calvert R L 1995 The rational choice theory of social institutions: cooperation, coordination, and communication. In: Banks J S, Hanushek E A (eds.) *Modern Political Economy. Old Topics, New Directions*. Cambridge University Press, Cambridge, UK

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA and London

Crawford S E S, Ostrom E 1995 A grammar of institutions. *American Political Science Review* **89**: 582–600

Demsetz H 1967 Toward a theory of property rights. *American Economic Review* **57**: 347–59

Eggertsson T 1990 *Economic Behavior and Institutions*. Cambridge University Press, Cambridge, UK

Elias N 1994 *The Civilizing Process: The History of Manners and State Formation and Civilization*. Blackwell, Cambridge, MA

Elster J 1989 *The Cement of Society. A Study of Social Order*. Cambridge University Press, Cambridge, UK

Ensminger J, Knight J 1997 Changing social norms: common property, bridewealth, and clan exogamy. *Current Anthropology* **38**: 1–24

Epstein A L 1968 Sanctions. In: Sills D L (eds.) *International Encyclopedia of the Social Sciences*. Macmillan, New York

Fine G A in press Framing norms: Mushrooming and the culture of expectations and explanations. In: Hechter M, Opp K-D (eds.) *Social Norms*. Russell Sage Foundation, New York

Gibbs J P 1965 Norms: the problem of definition and classification. *American Journal of Sociology* **70**: 586–94

Hall P A, Taylor R C R 1996 Political science and the three new institutionalisms. *Political Studies* **44**: 936–57

Hardin R 1980 The emergence of norms. *Ethics* **90**: 575–87

Hechter M 1984 When actors comply: monitoring costs and the production of social order. *Acta Sociologica* **27**: 161–83

Hechter M 1987 *Principles of Group Solidarity*. University of California Press, Berkeley, CA

Hechter M, Opp K-D in press What have we learned about the emergence of norms? In: Hechter M, Opp K-D (eds.) *Social Norms*. Russell Sage Foundation, New York

Heckathorn D D 1990 Collective sanctions and compliance norms: a formal theory of group-mediated social control. *American Sociological Review* **55**: 366–84

Homans G C 1974 *Social Behavior. Its Elementary Forms*. Harcourt, Brace & World, New York

Horne C in press Sociological perspectives on the emergence of norms. In: Hechter M, Opp K-D (eds.) *Social Norms*. Russell Sage Foundation, New York City

Jackson J M 1966 Structural characteristics of norms. In: Biddle B J, Thomas E J (eds.) *Role Theory. Concepts and Research*. Wiley, New York

Jasso G, Opp K-D 1997 Probing the character of norms: a factorial survey analysis of the norms of political action. *American Sociological Review* **62**: 947–64

Lewis D 1969 *Convention. A Philosophical Study*. Harvard University Press, Cambridge, MA

Mackie G 1996 Ending footbinding and infibulation: a convention account. *American Sociological Review* **61**: 999–1017

Merton R K 1968 *Social Theory and Social Structure*. Free Press, New York

Morris R T 1956 A typology of norms. *American Sociological Review* **21**: 610–13

Oliver P 1980 Rewards and punishments as selective incentives for collective action. *American Journal of Sociology* **85**: 1356–75

Opp K-D 1982 The evolutionary emergence of norms. *British Journal of Social Psychology* **21**: 139–49

Opp K-D 1997 Norms, rationalizations and collective political action. A rational choice perspective. *Swiss Journal of Economics and Statistics* **133**: 241–74

Opp K-D 1999 Contending conceptions of the theory of rational action. *Journal of Theoretical Politics* **11**: 171–202

Opp K-D in press Social networks and the emergence of protest norms: some hypotheses and an empirical test. In: Hechter M, Opp K-D (eds.) *Social Norms*. Russell Sage Foundation, New York

Opp K-D, Gern C 1993 Dissident groups, personal networks, and spontaneous cooperation: the East German revolution of 1989. *American Sociological Review* **58**: 659–80

Posner R A, Rasmusen E B 1999 Creating and enforcing norms, with special reference to sanctions. *International Review of Law and Economics* **19**: 369–82

Rommetveit R 1953 *Social Norms and Roles*. Universitetsforlaget, Oslo, Norway

Rutherford M 1996 *Institutions in Economics. The Old and the New Institutionalism*. Cambridge University Press, Cambridge, UK

Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA

Sugden R 1995 The coexistence of conventions. *Journal of Economic Behavior and Organizations* **28**: 241–56

Ullmann-Margalit E 1977 *The Emergence of Norms*. Clarendon Press, Oxford, UK

Williams R M 1968 The concept of norms. In: Sills D (ed.) *International Encyclopedia of the Social Sciences* Vol. II. Macmillan, New York
Wrong D H 1961 The oversocialized conception of man in modern sociology. *American Journal of Sociology* **26**: 183–93
Young H P 1998 *Individual Strategy and Social Structure. An Evolutionary Theory of Institutions*. Princeton University Press, Princeton, NJ

K.-D. Opp

Norms in Science

It is generally recognized that communication among scientists and scholars is governed by a particular scientific ethos, i.e., by a set of rules that are supposed to establish trust in, and guarantee the reliability of, the knowledge created in the process. This ethos, although rarely made explicit in science, was given its most succinct and influential formulation by the American sociologist of science Robert K. Merton (1957 (1942)), who defined it in terms of four basic norms, and with that laid the groundwork for the *Sociology of Science*. This article presents the norms, their significance in scientific and scholarly communication, and some criticisms challenging their actual existence.

1. *The Norms of Scientific Communication*

The scientific ethos refers to those patterns of behavior and implicit norms among scientists that can be traced back in history to the establishment of the academies in England and France in the seventeenth century. Regardless of many changes in detail there is a remarkable continuity in this ethos over a period of more than three centuries. Merton's formulation of the norms of science can be understood as an analytical condensation of the behavioral patterns that evolved over the duration of this period into a set of institutional imperatives or norms. The ethos, in his words, 'is that effectively toned complex of values and norms which is held to be binding on the man of science' and which, although not codified, can be inferred from, above all, the 'moral indignation directed toward contravention of the ethos' (Merton 1957, pp. 551–2). The four key norms, according to Merton, were the following.

(a) Universalism is the principle that truth claims are 'subjected to pre-established impersonal criteria' (Merton 1957, p. 553) irrespective of the social attributes of their protagonists, e.g., nationality, race, class, or religion.

(b) Communism (later, and preferably termed 'communality') refers to the norm that the findings of science 'are a product of social collaboration and are assigned to the community' (Merton 1957, p. 556). Property rights are kept to a minimum, and the scientist's claim to intellectual property may only be exchanged for recognition and esteem by the members of his community.

(c) Disinterestedness demands from the scientists to resist the temptation of using improper means (such as fraudulent manipulation of data) in the search for and communication of 'true' knowledge to their own advantage. It is a fundamental institutional element that has its basis in the public and testable nature of science, and it contributes to the integrity of scientists by the accountability to their peers.

(d) Organized skepticism is 'both a methodological and an institutional mandate.' It refers to the scrutiny of claims and beliefs by scientists' peers on the basis of empirical and logical criteria (Merton 1957, p. 560).

These norms or values are inter-related and together shape a system of communication that is uniquely geared to produce knowledge that may be considered 'true,' in the sense of being reliable, but by no means final. Universalism guarantees, at least formally, general social accessibility and at the same time prevents the intrusion of any other criteria (political, religious, ethnic) in the communication of knowledge other than those accepted as belonging to science itself. Communality subjects all knowledge to general and open communication, thereby subordinating the proprietary interests of the scientist to the less tangible rewards of gaining recognition by obtaining priority of discovery. It contributes to the social cohesion of the scientific community. Disinterestedness constitutes the self-reference of scientific communication. This means that science differs from the professions in that it has no clientele. The clients of science are scientists themselves; since all are engaged in a communally beneficial search for truth, fraud or quackery may lead to temporary advantages at best. Organized skepticism is the reverse side of the norm of disinterestedness, as it stipulates the impersonal scrutiny of any claim to truth as a general principle of scientific communication. It is institutionalized in the peer review system of journals and funding agencies.

Merton's list of the norms of science was one of several descriptions of the 'scientific ethos' formulated during the late 1930s and early 1940s in reaction to the threats against science and democracy by fascist states in Europe (Hollinger 1983). Merton's own account was written under the immediate stimulus of direct political interventions in science in National Socialist Germany (Mendelsohn 1989). Later treatments of the scientific ethos have largely ignored this peculiar historical (and ideological) context of its formulation and have focused on its sociological relevance. Bernard Barber, working in the structural-functionalist tradition of Talcott Parsons and Robert Merton, added three values to the list of norms: faith in the virtue of rationality, 'emotional neutrality as an instrumental condition for the achievement of ration-

ality,' and 'individualism' (Barber 1953, pp. 86, 88). However, partly because of their overlap with the Mertonian norms and partly because of their different theoretical status, they never attracted the same attention.

1.1 The Theoretical Status and Significance of the Norms in a Sociology of Science

In order to appreciate the significance of the norms one has first to realize that science in its modern form, as an institution which has the goal of the 'extension of certified knowledge' (Merton 1957, p. 552), is an unlikely social invention. The improbability lies in the fact that science emerged from a historical situation in which the production of knowledge was immediately linked to the glorification of God or to the pursuit of economic utility. The unprecedented progress of science began when it became a 'self-validating enterprise' (Merton 1957, p. 551), i.e., a separate functional subsystem of society (Luhmann 1990). On the institutional level, the remarkable social innovation was the separation of social status, religious affiliation, and, above all, prestige from the communication of truth claims. The rules and principles which have been poignantly dubbed 'technologies of distance' (Porter 1992) were instrumental in creating trust in scientific communication. They were gradually extended from individuals to institutions, allowing for a communication network that reached beyond generational limits in time, as well as beyond the social and geographical limits of personal acquaintances. Historically the process of the formation of modern science spans from the seventeenth to the early nineteenth century when the university emerged as the core institution and disciplines formed marking the establishment of self-validation and autonomy (Stichweh 1984). Merton's list of norms condensing the institutional patterns have emerged over this period into an analytical scheme which serves to explain the unique status of science—as a set of methods to produce certified knowledge and to accumulate the knowledge that accrues from their application, as well as a set of cultural values and customs that governs these activities. Thus, with the articulation of the norms, Merton also defined the aim for sociology of science to be concerned primarily with the moral compulsion of methods and their impact on behavior, not with their technical expediency. Merton in effect conceived of the sociology of science as an analysis of the social prerequisites and institutional configurations of science. It would in no way be concerned with the substantive findings of science or their methodological underpinnings. Consequently his sociology never challenged the philosophical assumptions that shaped the concept of science at a particular historical moment and which have changed since then, a limitation for which his views have subsequently been criticized. Under the influence of Thomas Kuhn (1962) sociologists began to look into the normative nature of scientific paradigms and thereby crossed the line into a fully-fledged 'sociology of scientific knowledge.'

The theoretical status of the scientific ethos has been a persistent issue. It has often been argued that scientists do not behave in accordance with these norms and that they are subject to historical change (Barnes and Dolby 1970). However, the existence of norms is not necessarily reflected in actual behavior. Merton's theoretical construction of the norms is a complex combination of different elements: (a) social-psychological patterns of attitudes that are expressed in internalized (but not necessarily explicit) reactions to violations of the norms, in the awareness that one's own actions or those of others are breaching a code; and (b) social-structural patterns of sanctions, i.e., mechanisms institutionalized in science that sanction, positively or negatively, certain behaviors like plagiarism (negatively) or the open exchange of information (positively). Thus, to demonstrate whether the norms are operating or not, requires a more sophisticated approach than mere observation of behavior.

Merton himself responded to the challenges to the validity of the norms by pursuing two research strategies: (a) the analysis of apparent paradoxes in scientists' behavior and the resulting psychological and social conflicts (Merton 1973, (1968)); and (b) the study of typical reasons for attacks against science or for conflicts between science and its societal environment (Merton 1957 (1937)).

On the one hand Merton observed a peculiar ambivalence among scientists toward priority disputes and their own interest in priority of discovery. (His empirical evidence consists of biographical materials and correspondence of scientists dealing with priority disputes.) He attributed this ambivalence to the institutional mandate of disinterestedness and the primacy of the expansion of knowledge. Here again, sociological elements are coupled with social-psychological ones. The need for recognition felt by the individual scientist is not an expression of idiosyncratic vanity but is initiated by the institution. 'Continued appraisal of work and recognition of work well done constitute one mechanism that unites the world of science' (Merton 1973, p. 401). Merton reacted to the changes in social-psychological patterns effected by institutional changes. Observing the increasing number of multi-authored journal articles that indicate the rising importance of cooperative forms of research, he concluded that the 'interest in priority'—by no means historically invariant—may give way to an interest in the recognizability of one's work in a research team (Merton 1973, p. 409).

Merton's second approach to documenting the validity of the scientific ethos was to analyze conflicts between science and its societal environment. He observed two sources of hostility toward science: the

conflict between methods or important results of science with important social values, and the conflict between the sentiments embodied in the scientific ethos and those of other institutions (Merton 1957, p. 537). The most obvious conflict of the latter type arises in totalitarian political orders—the kind Merton had in mind at the time of his original writing—between the scientific ethos and the political code that routinely discredited science as 'cosmopolitan,' 'liberalistic,' or 'bourgeois.' A more complex effect may be seen to accrue from utilitarian demands on science. Originally in opposition to the demands of the church, the state, or the economy, the orientation to 'pure' science may undermine the social esteem of science. Although this purist orientation may be methodologically justified, as it minimizes bias and error by separating the production of knowledge from its application, its social effect may be resentment and revolt against the possible negative consequences of science (Merton 1957, pp. 543–5). This diagnosis has proven to be farsighted. Since the 1960s public resistance to, and critique of, technological projects have steadily increased and the social contract for 'pure science' has come under pressure.

Merton's formulation of the scientific ethos as a description of the special nature of science as an institution has been fruitful in two ways. On the one hand it focused on the effects of structural changes in science on the institutional mechanisms that guarantee the production of certified knowledge, through channeling scientists' behavior. On the other hand it focused on the conflicts that emerge between science and other institutions. The question then, is what becomes of the scientific ethos as science and its institutional context change over time?

2. *The Scientific Ethos and the Changing Nature of Science*

Critics of Merton have pointed to the many changes science has undergone that are not accounted for in an ethos that is basically geared to an individualist concept of gentlemanly science. Merton himself addressed these changes and granted some plausibility to the view that the customs of scientists must have changed in view of the fact that all the 'basic demographic, social, economic, political, and organizational parameters of science have acquired dramatically new values' (Merton 1973, pp. 328–9). He enumerated certain typical phenomena such as the premature publication of results before a theory is corroborated at the cost of more careful competitors, and an intensified rivalry occasioned by specialization as well as a new kind of cynicism, amorality, and disillusionment that goes along with it. However, scientists' interest in priority as such has not changed as these actions demonstrate. Rather than indicating a change of the ethos, public reactions to apparent or real breaches of the ethos reveal an 'instructive paradox,' namely, that the 'customs governing the public demeanor of scientists and the public evaluation of contemporaries have become more exacting rather than less' (Merton 1973, p. 338). Merton's observation is supported by heightened concerns among governments and the media about spectacular cases of fraud in science in the 1980s. Although it is by no means clear that the incidence of fraudulent behavior among scientists has increased, public awareness and vigilance have grown stronger.

Other theorists have suspected that the ethos of science is threatened by fundamental changes in social values. Ezrahi sees a challenge to the scientific ethos and to liberal-democratic values by the counterculture that emerged from the radicalism of the 1960s, which favored values such as personality, authenticity, and creativity without standards. The belief that science is an enterprise governed by universalistic norms that produces objective knowledge by abstracting from its bearers is replaced, in his view, by an individualization of the perception of reality and the production of knowledge (Ezrahi 1990, p. 54). In a similar vein Forman focused on the effects of postmodernist values on science. He observed three main changes in the production of knowledge as causing the spread of these values: overproduction of knowledge and a resulting plurality of standards; instrumentalization of knowledge effecting an ethos of production as an end in itself; and an affirmation of knowledge as 'bound' and 'interested' resulting from an alleged incommensurability of different stocks of knowledge (Forman 1997).

Whether or not these diagnoses are correct will ultimately be verified by observable changes in the scientific ethos and in the emergence of radically different patterns of the production of certified knowledge. Merton's formulation of the scientific ethos has provided invaluable groundwork for an understanding of the interaction between normative, psychological, and institutional mechanisms in the production and communication of scientific knowledge.

See also: Conventions and Norms: Philosophical Aspects; Ethical Issues in the 'New' Genetics; Ethics and Values; Normative Aspects of Social and Behavioral Science; Norms; Research Conduct: Ethical Codes; Research Ethics: Research; Research Funding: Ethical Aspects; Science, Sociology of; Technology, Social Construction of; Value Pluralism

Bibliography

Barber B 1953 *Science and the Social Order*, George Allen and Unwin, London

Barnes S B, Dolby R G A 1970 The scientific ethos: A deviant viewpoint. *Archives Européennes de Sociologie* **XI**: 3–25

Ezrahi Y 1990 *The Decent of Icarus. Science and the Transformation of Contemporary Society*, Harvard University Press, Cambridge, MA
Forman P 1997 Recent science: Late-modern and post-modern. In: Söderquist T (ed.) *The Historiography of Contemporary Science and Technology*, Harwood Academic Publishers, Amsterdam
Hollinger D A 1983 The defense of democracy and Robert K. Merton's formulation of the scientific ethos. *Knowledge and Society: Studies in the Sociology of Culture Past and Present* **4**: 1–15
Kuhn T S 1962 *The Structure of Scientific Revolution*. The University of Chicago Press, Chicago
Luhmann N 1990 *Die Wissenschaft der Gesellschaft*, Suhrkamp, Frankfurt, Germany
Mendelsohn E 1989 Robert K. Merton: The celebration and defense of science. *Science in Context* **3**(1): 269–89
Merton R K 1957 Science and democratic social structure. In: Merton R K (ed.) *Social Theory and Social Structure*. The Free Press, Glencoe, IL
Merton R K 1973a Behavior patterns of scientists. In: Storer N W (ed.) Merton R K, *The Sociology of Science. Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago
Merton R K 1973b The ambivalence of scientists. In: Storer N W (ed.) Merton R K, *The Sociology of Science. Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago,
Porter T 1992 Quantification and the accounting ideal in science. *Social Studies of Science* **22**: 633–52
Stichweh R 1984 *Zur Entstehung des modernen Systems wissenschaftlicher Disziplinen. Physik in Deutschland 1740–1890*. Suhrkamp, Frankfurt, Germany

P. Weingart

North America and Native Americans: Sociocultural Aspects

The study of Native Americans of the United States and Canada was important in the development of anthropology in North America and elsewhere. In this article the Native American population is discussed in terms of its aboriginal state, the impact of European colonization upon it, and the contemporary situation of Native Americans. Also discussed is the anthropological study of Native Americans.

1. *Origins*

Scholars have for some time generally accepted the notion that *Homo sapiens sapiens*—those who became the Native Americans—first arrived in the Western Hemisphere from the Asian continent. In this respect, they acknowledge the biological affinities between Native Americans today and peoples in eastern Asia. There was also general agreement that *H. sapiens sapiens* arrived via the Bering Land bridge, which has existed at least four times in the past 70,000 years. The issue was not how *H. sapiens sapiens* arrived, but when. The conservative view, based on clearly dated human skeletal remains, is that humans arrived only around 10,000 years ago; the liberal view, based on less clearly dated human skeletal remains, as well as other types of evidence, e.g., toolmaking traditions, language, archaeological sites, is that humans have been in North America at least twice as long, if not considerably longer. Scholars have also revised their thinking as to how humans arrived in the Western Hemisphere, initially and/or subsequently. Another migration route gaining increasing acceptance is a maritime one, whereby humans migrated along the coast and thus around the edges of large glaciers which covered much of the northern part of North America at that time period. It is generally accepted now that there was not just one migration, or even just two: the first, the ancestors of American Indians; and the second, the ancestors of Inuit (Eskimos) and Aleutian Islanders. Most scholars would now accept the notion that there were at least three migrations: one, ancestors of most American Indians; two, ancestors of Athabascan American Indians; and, three, ancestors of Inuit and Aleuts. There is even acceptance of possible other migrations, as evidenced by the controversy and speculations surrounding 'Kennewick man.'

Some of these migrations may have come from areas other than Asia. Confusing the scholarly study of arrivals and their timing is the idea of continued contact between peoples of the Pacific Rim following migrations. Posed to provide clarification of migrations, dates and affinities are developing techniques of analysis, e.g., DNA. Future generations of scholars will provide future answers and, undoubtedly, future controversies. It is clear, however, that what is emerging is a complex and dynamic picture of the peopling of the North American continent.

2. *Early Native Americans*

However and whenever, and how many separate times humans arrived in the Western Hemisphere before the late fifteenth century, they eventually developed elaborate and diverse societies and cultures. They, for example, developed a remarkable variety of material cultures, ones obviously shaped by the environments in which they lived and by how they exploited these environments. Some were foragers, exploiting different facets of their environments for food, clothing, shelter, etc. during different seasons of the year. Others were hunters, of deer and buffalo, particularly, but earlier of mammoths, mastodons and giant sloths; or they were fishers, for example, of salmon in the Pacific Northwest and Alaska. Yet others were maritime hunters of whales, seals and walruses. Many American Indians became agricultural peoples, typically grow-

ing maize by 1492—which diffused northward from present-day Mexico—but also other foods (e.g., squash and beans), along with maize, in the Southwest and East, or camas in the Pacific Northwest.

Cultural life also became rich. Multiple rituals and religions developed and flourished, along with diverse patterns of kinship and marriage, political organization and governance, and practices whereby youths became adults. Europeans considered many of these to be exotic, and compared them to ones they were familiar with in Europe and elsewhere in the world. Varied patterns of political organization and governance developed as well. For example, tribal patterns of organization diverged widely, ranging from small, extended kinship systems to triblets to tribes to confederacies and chiefdoms. Probably the most spectacular Native American peoples—from the European point of view—were the chiefdoms in Florida and the southeast to the Mississippi River described by the chroniclers of the explorer Hernando de Soto, whose expedition traveled the area from 1539–1543. These were large, dense settlements controlled by rulers, one of whom was the well-known 'Lady of Cofitachequi' (from present-day South Carolina). These chiefdoms soon declined following the arrival of European explorers, probably mainly from epidemic diseases brought by the Europeans. Some, however, such as Cahokia, across the Mississippi River from present-day St. Louis and the largest city in the Western Hemisphere north of Mexico, had reached their zenith prior to European contact. Perhaps a close second from the Europeans' definition of spectacular Native American peoples were the pueblos of the southwest, described at about the same time as were the Florida chiefdoms by the explorer Hernando Coronado, and located generally along either side of the Rio Grande river. Other regions of aboriginal North America had Native American civilizations that impressed Europeans to some degree, although Europeans were decidedly unimpressed with some Native American societies and cultures they encountered.

3. *Size and Distribution Throughout North America*

Native American populations, in the form of hundreds of distinct peoples, were distributed widely throughout North America ca. 1492.

3.1 *Aboriginal Population Size*

Scholars continue to debate the size of the Native American population present in North America north of Mexico ca. 1492, and thus also the magnitude of population decline that began soon after European arrival and continued to about 1900. James Mooney (1928) of the Smithsonian Institution formulated the classic estimate of aboriginal population size at, as he expressed it, 'first extensive European contact': he asserted that there were some 1.153 million Native Americans north of the Rio Grande river. Subsequent scholars have offered both slightly lower and far larger population estimates, ranging from Kroeber's (1939) 1.025 million to Dobyns (1983) 18 million (for north of Mesoamerica, an area including northern Mexico as well as the United States, Canada, and Greenland). The estimate I formulated and continue to use is 7-plus million, composed of some 5 million for present-day United States, and some 2 million for Canada and Greenland.

3.2 *Distribution*

Populations were not distributed evenly, however. For example, Kroeber (1939) used the concept of 'cultural areas' to divide Native American groups north of Mexico into areas sharing similar cultural traits as adaptations to similar environments. According to him, the largest populations were found in the southeast and California; the largest population concentrations were found in the Pueblo and Lower Columbia regions (see Kroeber 1939, table 8). More recently, Ubelaker (1988) argues that the largest area populations were found in the southwest and the northeast.

4. *Impact of European Contact*

Contact with Europeans following the beginning of their extensive colonization of the Western Hemisphere after 1492 brought fundamental changes to native populations, and their societies and cultures. During the 1990s, 'holocaust' emerged as the metaphor to view the population collapse of Native Americans accompanying European expansion into this hemisphere (see Thornton 1987). Likewise, this 'holocaust' has emerged as crucial to understanding the full impact of colonialism upon Native Americans, and their subsequent social, cultural, biological, and even psychological changes. Native American societies and cultures, and Native Americans as biological and psychological entities were all impacted by demographic collapse after 1492.

4.1 *Tribal Decimation and Relocation and/or Concentration*

During the first four centuries following initial European contact, virtually all Native American groups experienced population decimation, and relocation and/or concentration, to some degree. Early relocations were sometimes voluntary, as American Indians moved westward to escape encroaching Euro-

peans on America's east coast. This in turn produced conflicts with other American Indians as they were pushed further west by the immigrant Indians. Later, however, relocation became the official policy of the US government.

Native American population size in the United States and Canada declined for some 400 years, reaching a nadir population around the beginning of the twentieth century of perhaps only 375,000. The causes of the population decline, of course, were epidemics and other diseases introduced from Europe and Africa, e.g., smallpox, the plague, measles, forms of influenza, whooping cough, cholera, typhoid, diphtheria, scarlet fever, yellow fever, malaria, and various sexually transmitted diseases. European, Euro-American, and American colonialism in all its forms also contributed to Native American population reduction through warfare and genocide, relocations, and destruction of economic systems as well as the association of disease with all of these. Important, too, in the equation was the inability of Native American populations to recover from population losses as their societies and cultures were undermined by colonialism.

Accompanying this decimation were further patterns of population relocation and/or concentration, typically onto reservations. The largest pattern of population relocation occurred as official US government policy in the first half of the nineteenth century, whereby large populations of Indians in the southeast—primarily the Cherokee, Chickasaw, Choctaw, Creek, and Seminole—were removed to lands west of the Mississippi River, lands which were designated Indian Territory and eventually became the eastern half of the state of Oklahoma in 1907. Other tribes were removed into what became the state of Oklahoma (named after Oklahoma Territory, the former western half of the state) as well, particularly ones from other parts of the Southern and Western Plains, the Northern Plains, and the so-called 'Old Northwest,' around the Great Lakes.

US government troops, especially following the United States Civil War which ended in 1865, were used in particular to subjugate Native Americans by forcing them onto reservations in Oklahoma Territory, Indian Territory, and elsewhere in order to contain them and reshape their ways of life. American Indian struggles against such repression included new spiritual movements such as the Ghost Dances of 1870 and 1890, and the Peyote Religion, which arose among many Indian peoples. However, struggles also culminated in infamous massacres, e.g., at Sand Creek, Colorado; Fort Robinson, Nebraska; and Wounded Knee, South Dakota.

4.2 Tribal or Individual Indian

After establishing a reservation system for American Indians, the US government soon began to undermine the integrity of many Indian lands when it saw the reservation system as 'not working' to 'civilize' Indians. From the late nineteenth century into the early twentieth century, policy was introduced to allot parcels of land on some reservations, and in Oklahoma and Indian Territory to individual Indians, thereby eliminating many reservations. A reduced reservation pattern therefore emerged, with most remaining reservations being in the west and representing only small remnants of former lands.

The idea was that Indians were to become individual Indian citizens on privately owned parcels of land. This culminated in the Citizenship Act of 1924, which made all American Indians in the United States citizens for the first time and gave them full voting privileges. 'Left over' or 'surplus' lands were then 'opened' to claims by non-Indians. Accompanying allotment were larger political and corporate desires for American Indian lands. The decades following the period were characterized by frauds whereby individual allotments ended up in the hands of land barons, speculators, and corporations, as Debo (1940) has shown for the Cherokee, Chickasaw, Choctaw, Creek, and Seminole in Oklahoma; and as Meyer (1994) has shown regarding the White Earth Reservation in Minnesota.

American Indian tribal life was undermined greatly by allotment. Tribes able to retain tribal land in the form of reservations survived as tribes and as Indians better than those that were not able to do so. This included Indians on the Plains and in the west and southwest. In 1934, the Indian Reorganization Act was enacted, acknowledging that an Indian tribe had 'rights to organize for its common welfare.' The Act served to codify tribal governments under the all too watchful eyes of the Bureau of Indian Affairs. And, in turn, it provided the basis for the operations of the more than 330 federally recognized tribes, on reservations or not, and hundreds of federally recognized Alaskan Native Villages.

The larger American society's vision of American Indians as citizens was also manifest in its system of Indian education. In the early days, America established schools to educate Indians. In fact, the American system of higher education which began in the seventeenth century—the 'Ivy League' schools, in particular—was in part to educate Native Americans, as reflected in the charters of, for example, Harvard University and Dartmouth College. With the development of the reservation system and the development of the system of elementary and secondary education in the United States, a natural alliance was born, and some treaties provided funds for reservation schools. By the beginning of the twentieth century, there were more than 200 government schools on reservations, supplemented by schools funded reluctantly funded by states for Indian youth, and some schools in soon-to-become Oklahoma, primarily under the auspices of the Five Tribes there. By the

mid-1920s, however it became apparent that these schools—particularly the boarding ones, but also the day schools—were not effective as avenues to mainstream America: rather, they were a national disgrace. In 1928, the Meriam Report (Meriam 1928, p. 348) stated that the 'Indian problem is essentially an educational one,' and called for a redirection of Indian education. Eventually, schools for American Indians became more sympathetic to Indian cultures, and began to incorporate Indian history and culture into the curriculum. Following World War II, however, boarding and day schools for Indians began to decline, and by the 1950s public school education for Indians became the norm.

4.3 Urbanization

Coinciding with the increase in the American Indian population beginning with the twentieth century, and with the dispossession of Indians from their lands, were the beginnings of a small but steady increase in the proportion of the population that lived in urban areas. Less than one-half of one percent of the American Indian population of the United States lived in urban areas in 1900; by 1950, more than 13 percent of the population lived in urban areas; at the time of writing, well over half of the American Indian population lives in urban areas, and the remainder live on reservations or in rural areas. (Only about half of the Native Americans in Canada live in cities, a substantial increase from about one-third in the 1970s, and a mere one-eighth in the early 1960s.) In part, this was a result of American Indian servicemen and women returning to cities rather than to small towns or rural areas following World War II. It was also a result of the mid-twentieth-century Bureau of Indian Affairs relocation program that encouraged and assisted American Indians to relocate to urban areas.

5. Native Americans Today

The Native American population in the United States and Canada experienced significant population recover during the twentieth century. From the nadir population of about 375,000 at the start of the twentieth century, it increased to more than 3 million by the end of the century. (See Shoemaker (1999), Ubelaker (1988), and Thornton (2000) for considerations of Native American population recovery.) This occurred because Native American mortality rates decreased, and fertility rates remained higher for Native Americans than for the total population. It also occurred because of changing census definitions of Native Americans in the United States, whereby individuals formerly classified as non-Native Americans became classified as Native Americans through self-identification in the censuses.

5.1 Tribal Membership

Most individuals self-identified in US censuses reported a tribal affiliation. The largest reported tribal affiliations are Cherokee, Navajo, Chippewa (Ojibwa), Sioux, Choctaw, Pueblo, Apache, Iroquois, Lumbee, and Creek. Not all of those in the Native American population are actual members of Native American tribes, however. Only about two-thirds of those self-identified as Native American in the US decennial censuses are actual members of one of the federally recognized tribes or Alaskan Native Villages.

In Canada, the situation is different. There one must be registered under the Indian Act of Canada to be an 'official' Indian. Categories of Indians there include either status (or registered) Indians—those recognized under the Act—and nonstatus (or nonregistered) Indians—those who never were registered or gave up their registration (and became 'enfranchised'). Status Indians, in turn, are subdivided into treaty and nontreaty Indians. Canada also has the *Metis*, individuals of Indian and white ancestry who are recognized as such but not legally recognized as Indians. Most Canadian Indians are registered, however.

5.2 Distribution

Only about one-quarter of Native Americans in the United States live on the more than 300 reservations and trust lands in the United States. About half of these live on the ten largest reservations (and trust lands): Navajo Reservation (and trust lands); Pine Ridge Reservation (and trust lands); Gila River Reservation; Papago Reservation; Rosebud Reservation (and trust lands); San Carlos Reservation; Zuni Pueblo; Hopi Pueblo (and trust lands); and Blackfeet Reservation. Conversely, about two-thirds of the registered Indians in Canada live on more than 2000 reserves there.

According to census data, the states with the largest Native American population are, in order, Oklahoma, California, Arizona, New Mexico, Alaska, Washington, North Carolina, Texas, New York, and Michigan. United States cities with the largest Native American populations are New York City, Oklahoma City, Phoenix, Tulsa, Los Angeles, Minneapolis-St. Paul, Anchorage, and Albuquerque. Canadian provinces with the largest number of Native Americans are Ontario, British Columbia, Saskatchewan, and Manitoba. And, there is now in Canada a new Native-American-controlled province of Nunavut.

6. Anthropology and the Study of American Indians

Europeans who colonized North American soon developed intellectual interests in the native peoples here. They had to explain where the natives came

from, but they were also in awe of native customs and cultures. Early explorers even sent Native Americans back to Europe, so that Europeans could see them for themselves. Eventually, prominent European scholars such as Thomas Malthus and G. W. F. Hegel developed an interest in these people and sought to integrate them into their study of human history, although Malthus considered Native Americans as 'wretched', and Hegel thought they lacked history and would disappear.

In the late nineteenth century, however, Lewis Henry Morgan wrote of the native peoples of North America, including his two classics, *League of the Ho-de-no-sau-nee, or Iroquois* (1851) and *Ancient Society* (1877). He argued that they were disappearing and called for their study—documentation—before this happened. Others, such as Emile Durkheim, Sigmund Freud, Karl Marx, and Frederick Engels, also drew upon Native American societies and cultures in their writings. Although the interest here was decidedly on social evolution—with Native Americans being at the lower stages—intellectual interests changed in the early twentieth century toward how individual elements in societies and cultures were integrated to form a whole. Anthropology in the United States was now separating itself from sociology and coming into its own, with a decided interest in the native peoples there. In California, for example, Kroeber heeded Morgan's earlier call and sought in the first decades of the twentieth century to describe California Indians before they disappeared, or changed their ways of life fundamentally. In this regard, the well-known Ishi—the so-called 'last wild Indian'—became a person of curiosity for Kroeber and others. Anthropology and history soon rejoined forces at mid-century, however, with the decided interest in studying Native Americans using the newly developed idea of 'ethnohistory.'

It was not too long before Native Americans became more involved in setting research agendas about themselves, and reacted against the study of them by anthropologists. Many rallied around Vine Deloria Jr.'s critique of anthropology in his *Custer Died for Your Sins* (1969), and his depiction of anthropologists as 'anthros.' Coinciding with this was the emergence of Native American studies in colleges and universities in the United States, and Native studies in Canadian counterparts, as a reaction toward both anthropology and the neglect of the study of Native Americans in other disciplines (see Thornton 1998a).

At the time of writing, anthropology, Native American studies, and Native Americans have achieved a truce, although an occasionally delicate one, in no small part because anthropologists have gained a new respect for Native Americans. Coinciding with an increased degree of self-determination of Native Americans in American society was the passage of the National Museum of the American Indian Act in 1989, both creating a national museum at the Smithsonian Institution devoted to Native Americans, and mandating the return of Native American human remains and funerary objects held at the Smithsonian, primarily in the National Museum of Natural History. A year later, the Native American Graves Protection and Repatriation Act was passed which mandated the repatriation not only of human remains and funerary objects from all institutions receiving federal funding but also the return of Native American objects of cultural patrimony and sacred objects. The Smithsonian, already required to repatriate human remains and funerary objects, was exempt from this legislation. However, the National Museum of the American Indian Act was amended in 1996 to include the required return of objects of cultural patrimony and sacred objects held at the Smithsonian.

Smithsonian repatriation is monitored by a seven member committee mandated by congress, and it is required that the majority of its members be selected from nominations submitted by Native American tribes. Since its inception in 1990, it has been chaired by Russell Thornton. To date, several thousand sets of human remains and numerous cultural objects have been returned to Indian tribes from the Smithsonian. Included here are human remains from the Sand Creek and Fort Robinson Massacres, Ghost Dance shirts and other objects from the Wounded Knee Massacre, and Ishi's brain. (Among the thousands of remains yet to be returned are Sitting Bull's braid, cut from his body at his autopsy and sent to the Smithsonian.)

These national repatriation laws and state laws as well, have in turn forged new relationships between museums, archaeologists, physical anthropologists and ethnographers, and Native Americans (Thornton 1998b).

The study of Native Americans, by both non-native and native scholars, in anthropology and other disciplines, will continue; however, Native Americans will have an increasing role in determining what, how, and by whom it is studied.

See also: Aboriginal Rights; Cultural Contact: Archaeological Approaches; Ethnic Identity/Ethnicity and Archaeology; Ethnicity: Anthropological Aspects; Genocide: Anthropological Aspects; Hegemony: Anthropological Aspects; Hegemony: Cultural; Holocaust, The

Bibliography

Debo A 1940 *And Still the Waters Run: The Betrayal of the Five Civilized Tribes*. Princeton University Press, Princeton, NJ

Deloria V Jr. 1969 *Custer Died for Your Sins: An Indian Manifesto*. Macmillan, New York

Dobyns H 1983 *Their Number Become Thinned: Native American Population Dynamics in Eastern North America*. University of Tennessee Press, Knoxville, TN

Kroeber A L 1939 Cultural and natural areas of native north America. *University of California Publications in American Archaeology and Ethnology* **38**: 1–242
Meriam L 1928 *The Problem of Indian Administration*. Johns Hopkins University Press, Baltimore
Meyer M 1994 *The White Earth Tragedy: Ethnicity and Dispossession at a Minnesota Anishinaabe Reservation*. University of Nebraska Press, Lincoln, NE
Mooney J 1928 The aboriginal population of America north of Mexico. *Smithsonian Miscellaneous Collections* **80**: 1–40
Morgan L H 1851 *League of the Ho-de-no-sau-nee or Iroquois*. Burt Franklin, New York
Morgan L H 1877 *Ancient Society*. Holt, New York
Shoemaker N 1999 *American Indian Population Recovery in the Twentieth Century*. University of New Mexico Press, Albuquerque, NM
Thornton R 1987 *American Indian Holocaust and Survival: A Population History since 1492*. University of Oklahoma Press, Norman, OK
Thornton R 1998a Institutional and intellectual histories of Native American studies. In: Thornton R (ed.) *Studying Native America: Problems and Prospects*. University of Wisconsin Press, Madison, WI
Thornton R 1998b Who owns our past? Repatriating Native American human remains and cultural objects. In: Thornton R (ed.) *Studying Native America: Problems and Prospects*. University of Wisconsin Press, Madison, WI
Thornton R 2000 Population history of native North Americans. In: Haines M R, Steckel R H (eds.) *A Population History of North America*. Cambridge University Press, Cambridge, UK
Ubelaker D 1988 North American Indian population size, A.D. 1500 to 1985. *American Journal of Physical Anthropology* **77**: 289–94

R. Thornton

North America, Archaeology of

1. Description of the Field

North American archaeology is defined as the study of the archaeological record of North America's past peoples. The field involves the investigation of prehistoric and historic Native American societies, the non-Indian explorers and colonists they encountered, and the development of the modern, pluralistic nations of Canada, USA, and Mexico. The region is unrivaled for the investigation of past hunter–gatherer societies whose subsistence practices involved the harvesting of wild plants and animals.

A sample of this diversity includes the hunting and fishing peoples of the far north who developed sophisticated toolkits of bone, antler, ivory, and stone for subsisting in the tundra and boreal forest (Norton, Dorset, Thule peoples); the maritime-adapted peoples of southern California and the Northwest Coast who resided in coastal villages, had access to ocean-going canoes, and harvested a bounty of shellfish, fish, and sea mammals from the ocean; the specialized hunters of the Plains who pursued bison and other game over thousands of years using an ever-changing assortment of stone projectile points and hunting techniques; and the foragers of the arid Great Basin who left behind a detailed record of their technology and life ways over many generations in rock shelters and caves (See Ames and Maschner 1999, Frison 1991, Glassow 1996, Madsen and O'Connell 1982, McGhee 1996).

North America is also noted for the study of native peoples who practiced horticulture and lived in settled communities. They include the peoples of the Eastern Woodlands who constructed an impressive range of burial mounds, temple mounds, and geometric 'effigy' mounds, participated in extensive trade networks, and cultivated both indigenous and foreign crops (Adena, Hopewell, and Mississippian peoples); the ancestors of the contemporary Iroquois who constructed palisaded, long-house settlements in the Northeast; the farming peoples of the Plains who settled the Missouri River in villages comprised of large, earth lodge houses, often protected by ditches and palisade walls; the Pueblo dwellers of the plateau and highlands Southwest who lived in masonry and adobe apartment complexes and practiced elaborate religious observances in underground ceremonial chambers; and the Hohokam peoples of the desert Southwest who lived in large towns with ball courts and platform mounds supported by an irrigation system with hundreds of miles of canals (see Fagan 1995, Knight and Steponaitis 1998, Plog 1998, Trigger 1986).

North American archaeology also involves the study of Spanish, Russian, English, and French colonial sites, as well as later mining communities, plantations, ranches, and towns. An increasing number of studies are focusing on the history of the modern State using the combination of archaeological remains and texts (Deetz 1996, Milanich and Milbrath 1989).

The immense region of North America is often subdivided into smaller geographic areas for study. The authoritative *Handbook of North American Indians* published by the Smithsonian Institution in 20 volumes subdivides North America into 10 geographic areas: Arctic, Subarctic, Northwest Coast, Plateau, Plains, Northeast, Southeast, Great Basin, California, and Southwest (see Fig. 1). Each area is characterized by its own history of research, archaeological societies, avocational groups, scientific journals and meetings, and regional textbooks. The archaeology of each area is often divided into three broad periods (prehistoric, protohistoric, historic). The prehistoric period encompasses many thousands of years of native cultures prior to evidence of early European contact. The protohistoric period refers to the temporal interval when the effects of European contact were first being felt on native populations (Euro/Asian diseases, trade), but prior to permanent European settlements. The historic period begins when European colonies

Figure 1
Geographic areas of North American archaeology (Source: Washburn 1988. *Handbook of Native American Indians, Vol. 4: History of Indian-White Relations*. Copyright by Smithsonian Institution Press)

first became established in an area, and regularized interactions between natives and Europeans took place.

2. *History of North American Archaeology*

North American archaeology did not develop into a true scholarly enterprise until about 1850. The primary catalyst was the creation of the Smithsonian Institution in 1846, as well as the founding of the Peabody Museum of Harvard University in 1866. Both institutions sponsored pioneering fieldwork and publication programs. But it was the establishment of two arms of the Smithsonian Institution (the National Museum and Bureau of Ethnology) in 1879 that made a lasting legacy to the field. A concerted effort was made to study the traditional practices of Native American peoples before they were inundated by Western technology. A four-field anthropological

approach was initiated that involved fieldworkers in the study of native languages (linguistic anthropology), the investigation of contemporary native communities (ethnography), the examination of Native American biological diversity (biological anthropology), and the documentation of past native peoples through the study of prehistoric and historic artifacts (archaeology). Classic examples of this holistic anthropological approach include the debunking of the 'mound builders' myth by Cyrus Thomas, who demonstrated that the prehistoric mounds of the Eastern Woodlands were constructed by Native Americans and not by lost 'civilized' peoples, and the innovative research of F. H. Cushing and J. W. Fewkes among the Pueblo peoples of the American Southwest. Smithsonian scholars were also notable in setting the standards of rigorous scientific research that continue today, such as those developed by William Henry Holmes and Aleš Hrdlicka to evaluate the age of early human remains.

The growth of the field really began during the first part of the twentieth century when other institutions began sponsoring archaeological programs. Many of the field and laboratory methods that distinguish North American archaeology today were first developed or refined at this time by such scholars as Max Uhle in California, Nels Nelson and A. V. Kidder in the American Southwest, James Ford in the Southeast, and H. B. Collins in the Arctic/SubArctic. Excavations were undertaken using either arbitrary levels or natural/cultural strata, and regional chronologies were generated based on the stratigraphic relationship of time-sensitive artifacts, such as projectile points and decorated ceramics. Seriation was also employed to time-order artifacts from the earliest to the latest, thus providing a relative chronology for sites and areas.

The holistic anthropological perspective distinguished the North American field from archaeological practices in Europe. Most notable was the use of direct historic analogy and the direct historical approach by North American scholars. The former employs ethnographic observations of native communities to generate interpretations of nearby archaeological remains left behind by their ancestors. The direct historical approach, originally advocated by Duncan Strong and Waldo Wedel in the Plains and Robert Heizer in California, is a more sophisticated program that examines the dynamics of culture change and persistence among native societies over many generations. Multiple lines of evidence (archaeological, historical, ethnographic, linguistic, biological data) are employed to study native cultural developments over the prehistoric, protohistoric, and historic periods.

Since 1950, North American archaeology has witnessed three theoretical perspectives that have each contributed greatly to the current state of the field. The first, culture history, developed hand-in-hand with the construction of regional chronologies in the early decades of the twentieth century. It involved cultural classifications of archaeological remains into time–space units. Temporal units outlined the normative or typical settlement layouts, house shapes, artifact types, and subsistence practices for a local region over time. With the advent of radiocarbon dating and dendrochronology (tree ring dating), calendar dates could be assigned to each cultural period. Local regions with similar cultural traits were then grouped together into broader geographic areas or culture areas. The current division of North American archaeology into 10 geographic areas is a direct consequence of cultural historical reconstructions.

The second perspective, 'processual' or 'new' archaeology in the 1960s and 1970s, continues to be a major influence in North American archaeological method and theory. Identified with Lewis Binford and his followers, this program stressed cultural evolutionary models, formal hypothesis testing, explanation of cultural processes, and the ecological adaptations of humans to the environment. Processual archaeology provided a new impetus for re-examining hunter–gatherer societies, and for examining the evolution of 'complex' political and religious organizations in many areas of North America.

The third perspective, 'postprocessual' archaeology, was inspired by British and American scholars in the 1980s and 1990s whose field research was largely outside North America. Unlike culture history and processual archaeology that were nurtured and developed in North America, the impact of postprocessual archaeology has been somewhat muted. The major contributions to date include the growing importance of historical analyses, the employment of multiscalar approaches, and the broadening of topics to include studies of gender, ideology, identity formation, and so on. (For more detailed treatments on the history of North American archaeology see Trigger 1990, Willey and Sabloff 1995.)

3. *Current State of North American Archaeology*

The field is currently in transition. The number of archaeologists has grown substantially since 1950, and employment opportunities have changed dramatically. While most archaeologists were once associated with universities or museums, the vast majority are now employed by the private sector. This was primarily the result of federal and state legislation enacted since 1966 that provides protection, management, and oversight to archaeological remains on public lands (e.g., see *Cultural Resource Management (CRM): Conservation of Cultural Heritage*). Academic (museum and university) archaeologists now comprise the smallest group of practicing professionals. The largest group are those working for cultural resource management (CRM) firms undertaking 'contract' research in accordance to federal,

state, and local laws, followed by archaeologists employed by federal and state agencies to oversee the management of public lands and cultural resources. The passage of NAGPRA (Native American Graves Protection and Repatriation Act) in 1990 has given a legal voice to Native Americans about the disposition of human skeletal remains and associated funerary artifacts in museums and the conduct of archaeological research on federal lands (e.g., see *Archaeology, Politics of*).

North American archaeology is becoming increasingly specialized, in large part because of the rapid growth of the field, the geometric increase of information resulting from CRM research, and the greater difficulties of obtaining field permits. Long gone are the days when one individual or institution could control or synthesize all the pertinent information for the entire field. Most archaeologists now focus their research in one of the ten geographic areas, and many confine their field work to considerably smaller areas of specialization.

What integrates North America archaeology as a field today are its shared historical roots, common research problems and methods, and political concerns about future funding, the protection of archaeological sites, and continued access to the archaeological record. The major research problems that crosscut the field include: (a) the study of early human antiquity, (b) documenting and understanding the diversity of hunter–gatherer societies, (c) considering the processes and implications of horticulture, (d) examining the organization of 'complex societies' characterized by powerful chiefs and religious specialists, and (e) understanding how native peoples responded to different kinds of colonial institutions (mercantile, military, religious), and how the outcomes of these encounters directed cultural developments in postcolonial contexts.

Perspectives on the early peopling of North America are rapidly changing as new discoveries are made (especially in South America). The model of Clovis 'big game' hunters crossing the Bering Land Bridge and entering North America approximately 11,500 years ago is giving way to new scenarios. Current thinking suggests that multiple migrations by land and possibly even sea may have taken place many generations earlier by peoples from diverse homelands (e.g., see Meltzer 1993).

Theoretical perspectives of evolutionary ecology and optimal foraging models are now employed in the study of hunter–gatherers, especially in the Great Basin and California, where human adaptations to different cultural and natural landscapes are considered (Madsen and O'Connell 1982). Great strides have been made in understanding the kinds of indigenous plants (e.g., sunflowers, sumpweed) domesticated in the Eastern Woodlands, as well as the spread and local adaptations of corn agriculture in the Northeast, Southeast, Southwest, and Plains (Smith 1992). Innovative models are being tested on the sociopolitical organizations, religious practices, ideological structures, and gender relationships in prehistoric Mississippian, Hohokam, Iroquoian, and Pueblo societies (Knight and Steponaitis 1998, Plog 1998; see also *Chiefdoms, Archaeology of*). Considerable work is now being devoted to the study of Spanish, French, English, Russian, and American settlements, and the development of the modern capitalistic state (Deetz 1996, Milanich and Milbrath 1989, Trigger 1986; see also *Cultural Contact: Archaeological Approaches*).

4. Future Directions

The field will continue to change rapidly in the next few years. Women will make up the majority of archaeologists, and an increasing number of underrepresented minorities will be recruited to join its ranks. One impact of NAGPRA is that more concerted efforts will be made to train Native Americans as professional archaeologists. Public education and outreach programs in archaeology will multiply, and competition for archaeological funding from government and private sources will no doubt increase. As more stringent efforts are made to protect and manage archaeological sites on public lands, noninvasive field methods will be refined and greater emphasis will be placed on the study of existing archaeological collections in museums and other repositories. Undoubtedly, new research problems will be addressed. One issue stimulated by NAGPRA is how to define legitimately cultural affiliations between archaeological remains and contemporary tribal groups (see Fagan 1995 for an excellent synthesis of North American Archaeology, Thomas 1994 also provides an excellent discussion of the field along with museums and sites that can be visited).

See also: Historical Archaeology; Horticultural Societies; Hunter–Gatherer Societies, Archaeology of

Bibliography

Ames K M, Maschner H D 1999 *Peoples of the Northwest Coast: Their Archaeology and Prehistory*. Thames and Hudson, New York

Deetz J 1996 *In Small Things Forgotten: The Archaeology of Early American Life*. Anchor Books, New York

Fagan B M 1995 *Ancient North America*, 2nd edn. Thames and Hudson, New York

Frison G C 1991 *Prehistoric Hunters of the High Plains*, 2nd edn. Academic Press, New York

Glassow M A 1996 *Purisimeño of Chumash Prehistory: Maritime Adaptations Along the Southern California Coast*. Harcourt Brace College Publishers, Fort Worth, TX

Knight V J, Steponaitis V P 1998 *Archaeology of the Moundville Chiefdom*. Smithsonian Institution Press, Washington, DC

Madsen D B, O'Connell J 1982 *Man and Environment in the Great Basin*. Society for American Archaeology, Washington, DC
McGhee R 1996 *Ancient People of the Arctic*. University of Washington Press, Seattle, WA
Meltzer D J 1993 *Search for the First Americans*. Smithsonian Institution Press, Washington, DC
Milanich J T, Milbrath S 1989 *First Encounters: Spanish Explorations in the Caribbean and the United States, 1492–1570*. University of Florida Press, Gainesville, Florida
Plog S 1998 *Ancient Peoples of the American Southwest*. Thames and Hudson, New York
Smith B D 1992 *Rivers of Change: Essays on Early Agriculture in Eastern North America*. Smithsonian Institution Press, Washington, DC
Thomas D H 1994 *Exploring Ancient North America: An Archaeological Guide*. Macmillan, New York
Trigger B G 1986 *Natives and Newcomers: Canada's 'Heroic Age' Reconsidered*. McGill-Queens University Press, Montreal, Canada
Trigger B G 1990 *A History of Archaeological Thought*. Cambridge University Press, Cambridge, UK
Washburn W E 1988 American Indians Vol. 4: *History of Indian–White Relations*. Smithsonian Institution Press, Washington, District of Columbia
Willey G R, Sabloff J A 1995 *A History of American Archaeology*, 3rd edn. W. H. Freeman, San Francisco, CA

K. G. Lightfoot

North America: Sociocultural Aspects

For five centuries peoples from Europe and Africa and those native to the New World have interacted in North America within continually shifting spatial and political arrangements. From the mid-nineteenth century onward these interactions became more complex following abolition of slavery, incorporation of Latin American populations of variously mixed ancestry, immigration from Asia, a closed settler frontier, and expanded citizenship rights for people of color and women. Sociocultural anthropologists have studied communities, subgroups, issues, and processes within this multicultural landscape through ethnographic fieldwork and by 'upstream' ethnohistorical research. They have also formulated more wide-ranging conclusions about structures of power and trajectories of change.

1. *Colonial and Early National North America*

Some five million or more speakers of hundreds of tongues (belonging to at least six language families) inhabited native North America in the 1490s when English, French, and other Western Europeans began to explore southward along the Atlantic coast from Labrador, and Spaniards northward from Florida, and later Mexico. By 1800 the more than four million whites and one million blacks living on the eastern seaboard far outnumbered the remaining 600,000 Native Americans, most of them now beyond the 'frontier' of white settlement. This demographic reversal involved free migration from Europe, enforced transportation from West and Central Africa, and massive population decline among American 'Indians.'

New World peoples bore no immunity to Old World diseases, and scores of smallpox, cholera, influenza, and other epidemics decimated them. Moreover, a cycle of change set in motion by the fur trade led to warfare over hunting and trading territories, and devastation from European-introduced guns and alcohol. In addition, permanent mixed farming in the north and middle colonies, and plantations in the south, left decreasing room for Native American corn–bean–squash swidden agriculture. The first Indian reservations were created in New England during the 1600s, and following the independence of the United States in 1783 (Canada remained a British dependency), Indian policy was placed in the Department of War. In 1830 the Indian Removal Act authorized the US president to relocate eastern groups west of the Mississippi River; overt ethnic cleansing later occurred in Texas where entire peoples were exterminated, and in California where numerous shootings and massacres followed the 1849 Gold Rush.

Unlike Indians, whose thinning numbers were welcomed by white settlers, African-Americans were induced to multiply, particularly after the overseas slave trade became illegal in 1807, and by 1850 their numbers reached 3.6 million. Of diverse languages and cultures (like the Indians), they were not deliberately divided by 'tribe' but rather experienced pressure to acculturate into a homogeneous enslaved population. African languages, and most now-unpracticed skills and conventions of African forms of life, were lost. The scattered free black populations, amounting to 400,000 by 1850, were all English speaking, though traces of Africa remained in music, folkcrafts and lore, and interpersonal behavior.

As elsewhere in the New World, numerous white males impregnated enslaved African women, but in North America a 'one drop rule' prevailed under which any known African ancestry classified a person as black, a situation that persisted after slavery and into the present. White and black North Americans became accustomed to observing a range of skin colors and facial features within the 'Negro' or 'colored' population, and to detecting African ancestry among light-skinned persons who were members of black communities or who attempted to 'pass' into the white population (Harris 1964).

White North Americans numbered 20 million by 1850. Mainly of British Isles origin, they included

other northern Europeans who by the 1800s had 'melted' into the English-speaking majority, but also a persisting French-speaking Canadian enclave. (Spain was displaced from Florida in 1819.) The northern and mid-Atlantic family-farm pattern accompanied whites expanding into the 'middle west' between the original colonies and the Mississippi River, and steadily shifted from subsistence to market production (mainly of wheat); large-scale white-owned commercial plantations worked by enslaved blacks spread across the south to Texas, interspersed with 'poor white' smallholdings on less productive land.

Along the Atlantic coast, industrial production in mill-towns and urban factories proliferated during the early nineteenth century. Some skilled artisans and inventors joined commercial investors in a solidifying capitalist class, while others, with laborers, deskilled 'mechanics,' and farm-family offspring, moved into a proletarianizing and geographically mobile working class; smaller numbers of clerks, now located in offices separated from workplaces, fell in between. White lifestyles increasingly diverged by class, especially for women—upper-class women managed household staffs, family visits, and charity efforts; working-class women performed the tasks of reproducing male labor, and often worked for wages or took in boarders as well. By the 1840s a moralistic evangelical Christianity was entrenched that extolled the rich for their successes, sanctified hierarchy, and blamed the poor for their own condition (Wallace 1978).

2. *1850s–1920s: Red, Black, and White Americas Engaged*

In 1851 Morgan published his *League of the Ho-de-no-sau-nee*, or *Iroquois*; combining fieldwork, a comparative perspective, and theoretical interests in political organization, this book was the first full-scale anthropological ethnography. Though Morgan included sections on 'intercourse with Europeans' and 'future destiny of the Indian,' the volume concerned primarily Iroquois institutions and customs, in particular the dispersed *matriclans* (or 'tribes') and the overarching council of *sachems* uniting the five Iroquois 'nations.' Morgan, however, devoted a chapter to 'The New Religion' proclaimed by Handsome Lake five decades earlier, a revitalization movement (Kehoe 1989) formed in the wake of depopulation, conquest, and reservation life.

Further ethnographies of Native American institutions and customs did not follow until the US Bureau of American Ethnology was created in 1879, and works on the Zuni by Cushing, Omaha by Dorsey, and Inuit and Kwakiutl by Boas appeared in the 1880s and 1890s. Meanwhile, the buffalo-hunting Plains culture, itself a post-1600 development following the diffusion of horses from Spanish Mexico, were being harried out of existence by settlers, soldiers, receding reservations, and nineteenth-century epidemics which reduced the US Native American population to its nadir of 250,000 in 1890. That year more than 150 Sioux were killed at Wounded Knee by US troops, Indian resistance ended, and the frontier 'closed' (Kehoe 1989). Most studies of Native Americans published through the 1920s were 'salvage ethnography' aimed at reconstructing prereservation lifeways, but an exception was Radin's *Autobiography of a Winnebago Indian* (1920). This 'life history' followed a man born about 1880 from his Wisconsin youth of hunting, canoe travel, and fast-induced vision questing, to an adulthood of craft sales to whites, farm and railroad wage labor, wild west show employment, federal annuity payments, alcoholism, attacks from whites, imprisonment, and conversion to the Christian-influenced pan-Native American peyote religion.

In 1887, attempting to transform reservation land tenure from collective to private ownership, Congress had passed the Dawes Act which defined as 'Indian' only persons who were at least 'half bloods' (later replaced by a 'one-quarter blood quantum' standard); Native Americans' own kinship and participation-based criteria of societal membership were nullified (Jaimes in Gregory and Sanjek 1994). This restrictive logic of who could be Indian contrasted with the 'one drop rule,' which was ratified in the 1896 Plessy vs. Ferguson Supreme Court ruling that Louisiana could compel a white-appearing, one-eighth black passenger to ride on 'colored' railroad cars, and which legalized racial segregation for the next six decades.

Following the abolition of slavery and bestowal of citizenship on African-Americans after the Civil War, oppression in southern states had intensified, lynching proliferated, and racial segregation expanded. The first ethnography of black Americans, *The Philadelphia Negro* (1899) by DuBois (Baker 1998), was influenced by the British and American urban survey and settlement house movements, and reflected its author's training in history as Harvard's first black Ph.D. It depicted life in an inner city neighborhood, relations between southern migrant and northern-born African-Americans, and the ever-present realities of white–black inequality; extensive interviews were combined with participant observation on 'the function of the Negro church' and 'social classes and amusements.'

Inspired by the same sources as DuBois, sociocultural research on white North Americans and European immigrants (39 million had arrived in the US and Canada between 1850 and 1924) began a dozen years later at the University of Chicago. There monographic studies continued through the 1930s of itinerant 'hobo' workers, youth gangs, dancehall employees and patrons; Poles, Jews, French Canadians (and Mexicans and Chinese); and central city 'gold coast' apartment, artist colony, and rooming house dwellers. The first participant–observer community study—of Muncie, Indiana, a town of 35,000, and 85 percent 'native white

of native parentage'—was conducted in 1924–5; crediting anthropologists Rivers and Wissler for its holistic methodology, *Middletown* (Lynd and Lynd 1929) detailed work, family life, education, leisure, religion, and government in this glass industry-dominated town.

During the 1850s–1920s decades, Mexicans incorporated into the US after the 1846–8 Mexican–American war, as well as later cross-border immigrants, were illegally denied political rights by being classified as 'Indians,' who collectively were granted US citizenship in 1924 (Menchaca 1993). Puerto Ricans, US subjects since the 1898 Spanish–American war, became citizens in 1917; and during this widening of the citizenship pool, women achieved the right to vote in 1920. Asian immigration was severely restricted after 1882, and naturalization denied; it was not until 1952 that all Latin American and Asian immigrants were permitted to become US citizens.

3. *1930s–WWII: Projects and Regions*

Coordinated teams and research schemes marked the energetic North American ethnographic fieldwork of the 1930s and early 1940s. Warner during 1930–5 led a Harvard team studying 'Yankee City' (Newburyport, Massachusetts), a town of 17,000, and authored five volumes (condensed in Warner 1963). He identified a six-tiered status system of 'social classes'—the 'upper-upper' old-family aristocracy, and 'lower-upper' newly-wealthy (together just 3 percent of Yankee City); 'upper-middle' professionals and managers (10 percent); the 'lower-middle' and 'upper-lower' 'level of the common man' (together 61 percent); and the 'lower-lower' poor (25 percent). While later criticized for presenting a top-down view (the common man admittedly made fewer distinctions), Warner and colleagues did document social class in town geography, associations and cliques, churches, aspirations of non-Yankee white immigrant groups, politics, and public ceremonies.

A more intensive, if one-person, study of an immigrant and second-generation Italian neighborhood in Boston was conducted by Whyte during 1936–40, with advice from veterans of the Yankee City team. Perhaps the richest of all North American ethnographies, it challenged the 'disorganized slum' model of Chicago sociology, and focused on unemployed 'corner boys' in their 20s, their 'college boy' rivals, and gambling racketeers with ties to police and machine politics. (For the 1955 edition of *Street Corner Society* Whyte added a revealing 80-page autobiographical appendix on research methods.) Meanwhile, Warner initiated a team study of 'Jonesville,' a mid-western prairie town; despite the absence of upper-uppers and only two ethnic groups—Norwegians and Poles—the findings were similar to Yankee City, and Warner highlighted national similarities at 'the level of the common man,' now valorized by service in World War II.

The third leg of Warner's regional comparative scheme was a study of Natchez, Mississippi, a half-white, half-black trade center with an 80 percent black rural hinterland. In 1933–5 he dispatched one white and one African-American couple to study its two 'color-castes,' as Warner termed the polarized racial structure which he contrasted with the individual and group class mobility available to white ethnics. In his view, the emerging three-tiered class structure among blacks was overshadowed by race; in the future, he proposed, black education and wealth levels would approach the more elaborate, six-tiered white class structure, and then the caste bar to intermarriage might erode (Warner in Davis et al. 1941). North America was (and is) not there yet, and his students documented how interpersonal etiquette and economic and political institutions, legally and extralegally, operated to maintain white dominance.

Warner also sponsored a study of African-American life in Chicago conducted by a predominantly black team during 1935–8. Acknowledging inspiration from DuBois, the resulting 800-page volume, *Black Metropolis*, surveyed northward migration; the 'color-line' in housing, employment, marriage, and public accommodations; black upper, middle, and lower class life; and Chicago's 'intricate web of families, cliques, churches, and voluntary associations,' concluding that 'Negroes do not usually think of their institutional life as something "inferior." In fact, they express considerable pride in it' (Drake and Cayton 1962).

A comparative project of rural community studies conducted under the US Bureau of Agricultural Economics' auspice during 1939 (Leonard et al. 1978) included four variants of the white family-farm pattern: small-scale New Hampshire dairymen (and their underemployed working-class 'floater' neighbors); Iowa commercial feed-grain and livestock growers; capital-intensive Kansas wheat producers employing a plethora of specialized machinery; and low-tech Pennsylvania Amish pietists whose agrarian innovations over two centuries had diffused widely in Europe and North America. Two other studies documented inequities faced by Spanish-speaking New Mexicans displaced from stock-grazing lands by intrusive Anglos, and a racially-segregated Georgia county where white dairy-farm owners offered 'milk-hand' employment to the black local majority of tenant cotton-farmers barely surviving in the post-boll weevil economy. Everywhere New Deal farm assistance and job programs were intrinsic to rural life, and except for the Amish, New Mexicans, and black Georgians, the same radio programs, popular music, and movies enjoyed nationwide had displaced locally-produced entertainment.

Following Mead's 1930 participant–observer study of contemporary Omaha reservation life, acculturation research on Native Americans blossomed at

Columbia University, and also at the University of Chicago where an Indian Education, Personality and Administration Project begun in 1941 resulted in several monographs with strong acculturational and psychological content, emphases evident as well in Native American life histories produced by Yale anthropologists. A number of white participants in the Warner and Native American research schemes were among the 21 anthropologists employed as administrators and researchers by the War Relocation Authority which during 1942–6 supervised 10 concentration camps for 110,000 Japanese American citizens and their Japan-born parents. (Canadians of Japanese ancestry were also incarcerated.) Coincident with the beginnings of 'applied anthropology' in US industry and government administration, this wartime episode led to no sustained anthropological focus on Asian Americans during the next two decades, but did generate persisting questions about fieldwork ethics and responsibilities (Starn 1986).

4. *Postwar to 1950s: Corporatism and Stocktaking*

While Warner had studied the takeover of a Yankee City factory by a national company, and Bureau of Agricultural Economics rural fieldworkers documented mass media influence, only after World War II did corporate power and mass culture became central topics in sociocultural fieldwork. Unfortunately, however, the volume of North American research now shrank as anthropologists had greater funds and opportunities to do research abroad.

In rural California, Goldschmidt (1978) contrasted the dampening of community life by corporate agribusiness with the greater civic participation and lesser economic and racial stratification in locales characterized by family farms. Large-scale production in fact depended on federally subsidized irrigation which by law was intended for family-size units; when Goldschmidt's findings threatened their noncompliant water use, agribusiness growers employed political influence to vilify and suppress his (and other rural) research, and to maintain control of irrigation resources. (By 1975 the number of US farms had fallen from 6.5 million in 1920 to 2.8 million, 'vertical integration' of the food industry through production contracts between corporations and family farms was widespread, and low-wage, proletarianized farmwork was on the increase.)

Powdermaker (1950) studied movie producers, writers, directors, and actors during a year of fieldwork in Hollywood, focusing on the 'big five' studios' decisions about film content. Concerned with the 'models for human relationships' and 'folk heroes' presented to the mass North American audience, and the frustration of creativity within the industry, she identified the sources of power behind the ideological messages and standardization in Hollywood's 'assembly line' films: conservative Christian censors who threatened boycotts and profit-motivated 'front office executives' dedicated to least-common-denominator formulas. Since the vertically integrated studios also controlled film distribution, their decisions determined what was seen.

During the 1950s more and more food products and films were consumed in burgeoning suburbia. (By 1970 more North Americans lived in suburbs than in cities.) Studies appeared of working and lower-middle class Levittowners, and of upper-middle class 'exurbanite' and 'organization man' commuters. The richest of these ethnographies, *Crestwood Heights* (Seeley et al. 1956), described an affluent Toronto suburb where households valued privacy and possessions, leisure and vacation activities were tightly scheduled, gender differences between career-centered husbands and home-centered wives were magnified, and parents and child-rearing experts encouraged 'responsible' adult-like child behavior, which the schools reinforced.

In 1955 the *American Anthropologist*, North America's leading anthropology journal, published a special issue on 'The USA as Anthropologists See It' (Lantis 1955). Arensberg placed the harvest of 1930s–1940s community studies within a historical and regional typology, and outlined the emerging metropolitan-area configuration of central city and suburbs. Schneider and Homans stressed the impact of class mobility on cognitive kinship maps. Several authors (most ingeniously Mason) assayed writings on American national character and values, but none tied the 'traits' of independence, competition, conformity, faith in progress, and professed equality of opportunity to their white, male, and frontier/family-farm/business-owner roots; nor did they discuss alternative values of women, workers, people of color, and pietists who were denied realization of, denigrated by, or rejected this 'dominant value profile of American culture' (see Spindler and Spindler 1983, Mullings 1997, Wallace 1978). Japanese and Mexican Americans were 'ethnic groups' like Poles and Jews; African-Americans were not; and Native Americans surfaced only in an article on language. Little in this 1955 special issue presaged what the 1960s were about to unleash.

5. *1960s–1980s: Subcultures, Ethnicity, Diversity, and Power*

The rebirth of North American ethnography in the 1960s began with a dozen books on African-Americans (Hannerz 1975); if offering narrower views of black class differences than DuBois or the 1930s studies, this work was attuned to the movement for racial equality, and answered a call for 'relevance' within anthropology. Ethnographies of Mexican Americans, Puerto

Ricans, urban Indians, schools, and crime followed in the 1970s, and of workers, health care settings, gays and lesbians, white ethnics, rural life, homeless persons, and immigrants in the 1980s. Aims and theoretical frameworks varied as political and policy interests, and in-group documentation and vindicationist goals, intersected with undercontextualized notions of 'subcultures' and 'adaptive strategies' in the 1960s, an ethnicity boom in the 1970s, and exploration of gender, sexual, generational, occupational, and religious cultural diversity in the 1980s. Expertly crafted ethnographies—new classics—appeared concerning mentally retarded adults by Edgerton, alcoholic 'urban nomads' by Spradley, a school principal's milieu by Wolcott, dying children by Bluebond-Langner, male factory workers by Halle, Japanese American kinship by Yanagisako, and female hospital employees by Sacks. Between 1980 and 1992, Moffatt estimated, the volume of research on North America (excluding Native Americans) surpassed all that had come before (Moffatt 1992, Spindler and Spindler 1983).

In addition, established anthropological interests in family and kinship, and new ones spurred by the 1970s feminist movement, received attention. Clark and Anderson (1967) pioneered ethnographic research on the growing elderly population, redefining this 'social problem' to include study of kinship and friendship networks, lifecycle transitions, counterproductive American values, medical and economic institutions, and political advocacy. Reacting to a spate of cognitive studies of American kin terms, Schneider's controversial *American Kinship* (1968) distilled the 'enduring, diffuse solidarity' of ties among relatives by blood and by law, here understood as 'a system of symbols, and not ... at any other level.' (A little-known 1975 volume usefully analyzed his empirical kinship data.) Rapp (1987) synthesized historical, ethnographic, and feminist writings to clarify work, exchange, and gender patterns among poor, stable wage-earner, affluent middle-class, and transgenerationally-wealthy households and families; an efflorescence of research expanded this terrain during the 1980s (Ginsburg and Tsing 1990), demonstrating that 'gender is not just about women; it is about the social *relationship* between men and women and the ... cultural construction of femininity and masculinity' (Mullings 1997).

Religion, a topic of passing attention in many ethnographies, was the focus of Sinha's research (1966) in mid-western 'Mapletown,' one of few studies in the US or Canada by a non-North American anthropologist. Noting how Christianity was intertwined with American patriotism as well as class-stratified by denomination, he contrasted its 'rigid' and 'attenuated' qualities with the more transcendental and pervasive Hinduism of his native India. (By 1980, Hinduism was an ethnographically documented North American religion.) Fundamentalist Christians were less than 5 percent in Mapletown, but their increasing national numbers and influence by the 1980s drew the attention of Spuhler (1985), who analyzed the beliefs and organization of Bible-based 'scientific creationism,' a viewpoint endorsed by President Ronald Reagan and accepted by 44 percent of Americans, and Harding (1987), who studied the politically activist Moral Majority, focusing on the ideology and process of 'born again' conversion, and on anti-abortion efforts.

A plea to 'study up' into government and corporate bureaucracies, the upper classes, and the organization and abuse of power in the US—a call for an anthropological political economy to anchor reborn ethnographic energies—was registered by Nader (1972); for Canada, see Hughes (1964) and Sweet (1976). A fascinating commentary on these issues was the 'fleeting glimpses of America' of Fei, a Chinese anthropologist whose 1979 visit to the US was his first since 1944. Fei noted the now dominant, and federally subsidized, automobile–highway–supermarket–suburb complex; telephone communication organizing personal life; computerization of information, and of public and private decision-making; the rise of technical and managerial personnel to comprise one-quarter of the labor force; the power of government's corporate-contractors; the one-sixth of Americans below the poverty line; persisting and disproportionate unemployment among African-Americans; and a shortage of 'scholars who are trying to understand the entire society.' As the 1980s closed, Nash (1989) came nearest to absorbing Nader's charge and Fei's observations in her historically-attuned, well-contextualized ethnography of downsizing and gentrification in a Massachusetts factory town.

6. *1990s–Twenty-first Century: Transitions and Polarizations*

By the year 2080, the historic white US majority will fall below 50 percent, and people of color will comprise the majority. Rising intermarriage rates for whites, Latin Americans, Asians, and blacks will complicate government-imposed categories of race and ethnicity, while continuing immigration will replenish ethnic and racial populations. Against the rising pace of multiracial, multiethnic encounter, wealth is more polarized than ever before, and white migration from cities and coasts to suburbs and interior states is dividing the white population (Sanjek 1998). Countering the 1950s and 1970s emphases on ethnicity, historically-grounded understandings of the racialization of blacks, Latin Americans, and Asians in North America underscore the value of both race and ethnicity as analytic concepts (Gregory and Sanjek 1994, Menchaca 1993), a position advocated by Warner in the 1930s. The situation of Native Americans remains unique: while populations are growing, intermarriage rates exceed 50 percent, and, paradoxi-

cally, as Native American languages face extinction, control of reservation land in the US, and both rights and acreage in Canada, are increasing.

Ethnography in the twenty-first century will continue to focus on rights and representations—workplace and language rights of immigrants, civil rights of sexual minorities, rights to healthcare for the elderly and poor, rights to adequate education for children, 'taxpayers' rights for the affluent, rights for Christians opposed to abortion and evolution, housing and employment rights for African-Americans, 'color-blind' rights for whites in an increasingly color-full North America—and representations by corporate mass media, by school curricula and cultural institutions, by grassroots protestors and their advocates, and by members of diverse constituencies themselves (Mullings 1997, Zavella in Gregory and Sanjek 1994).

Bibliography

Baker L 1998 *From Savage to Negro: Anthropology and the Construction of Race, 1896–1954*. University of California Press, Berkeley

Clark M, Anderson B 1967 *Culture and Aging: An Anthropological Study of Older Americans*. Charles Thomas, Springfield, IL

Davis A, Gardner B, Gardner M 1941 *Deep South: A Social Anthropological Study of Caste and Class*. University of Chicago Press, Chicago

Drake S, Cayton H 1962 [1945] *Black Metropolis: A Study of Negro Life in a Northern City*. Harper Torchbooks, New York

DuBois W E B 1967 [1899] *The Philadelphia Negro: A Social Study*. Schocken, New York

Fei X 1979–80 *Fleeting Glimpses of America*. Dagongbao (Oct.–Nov.), Hong Kong; Wen Hui Bao (Jan.), Shanghai [English trans. Whyte M]

Ginsburg F, Tsing A (eds.) 1990 *Uncertain Terms: Negotiating Gender in American Culture*. Beacon, Boston

Goldschmidt W 1978 [1946, 1947] *As You Sow: Three Studies in the Social Consequences of Agribusiness*. Allanheld, Osmun, Monclair, NJ

Gregory S, Sanjek R (eds.) 1994 *Race*. Rutgers University Press, New Brunswick, NJ

Hannerz U 1975 Research in the Black Ghetto: A review of the sixties. In: Abrahams R, Szwed J (eds.) *Discovering Afro-America*. Brill, Leiden

Harding S 1987 Convicted by the holy spirit: The rhetoric of fundamental baptist conversion. *American Ethnologist* **14**: 167–81

Harris M 1964 *Patterns of Race in the Americas*. Walker, New York

Hughes E C 1964 A sociologist's view. In: Dickey J (ed.) *The United States and Canada*. Prentice-Hall, Englewood Cliffs, NJ

Kehoe A 1989 *The Ghost Dance: Ethnohistory and Revitalization*. Holt, Rinehart & Winston, New York

Lantis M (ed.) 1955 The USA as anthropologists see it. *American Anthropologist* **57**: 1113–295 [special issue]

Leonard O, Loomis C, Bell E, MacLeish K, Young K, Kollmorgen W, Moe E, Taylor C, Wynne W 1978 [1941–3] *Culture of Contemporary Rural Communities*. Greenwood, Westport, CT

Lynd R, Lynd H 1929 *Middletown: A Study in American Culture*. Harcourt, Brace, New York

Mead M 1932 *The Changing Culture of An Indian Tribe*. Columbia University Press, New York

Menchaca M 1993 Chicano Indianism: A historical account of racial repression in the United States. *American Ethnologist* **20**: 583–603

Moffatt M 1992 Ethnographic writing about American culture. *Annual Review of Anthropology* **21**: 205–29

Morgan L H 1962 [1851] *League of the Ho-de-no-sau-nee, or Iroquois*. Cornith, New York

Mullings L 1997 *On Our Own Terms: Race, Class, and Gender in the Lives of African American Women*. Routledge, New York

Nader L 1972 Up the anthropologist—perspectives gained from studying up. In: Hymes D (ed.) *Reinventing Anthropology*. Vintage, New York

Nash J 1989 *From Tank Town to High Tech: The Clash of Community and Industrial Cycles*. State University of New York Press, Albany, NY

Powdermaker H 1950 *Hollywood the Dream Factory: An Anthropologist Looks at the Movie-makers*. Secker & Warburg, London

Radin P 1963 [1920] *The Autobiography of a Winnebago Indian*. Dover, New York

Rapp R 1987 [1978] Urban kinship in contemporary America: Families, classes, and ideology. In: Mullings L (ed.) *Cities of the United States: Studies in Urban Anthropology*. Columbia University Press, New York

Sanjek R 1998 *The Future of Us All: Race and Neighborhood Politics in New York City*. Cornell University Press, Ithaca, NY

Schneider D 1968 *American Kinship: A Cultural Account*. Prentice-Hall, Englewood Cliffs, NJ

Seeley J, Sim R A, Loosley E 1956 *Crestwood Heights: A Study of the Culture of Suburban Life*. Basic Books, New York

Sinha S 1966 Religion in an affluent society. *Current Anthropology* **7**: 189–95

Spindler G, Spindler L 1983 Anthropologists view American culture. *Annual Review of Anthropology* **12**: 49–78

Spuhler J 1985 Anthropology, evolution and 'scientific creationism.' *Annual Review of Anthropology* **14**: 103–33

Starn O 1986 Engineering internment: Anthropologists and the war relocation authority. *American Ethnologist* **13**: 700–20

Sweet L 1976 What is Canadian anthropology? *American Anthropologist* **78**: 844–50

Wallace A 1978 *Rockdale: The Growth of an American Village in the Early Industrial Revolution*. Norton, New York

Warner W L 1963 *Yankee City*. Yale University Press, New Haven, CT

Whyte W F 1955 [1943] *Street Corner Society: The Social Structure of an Italian Slum*. University of Chicago Press, Chicago

R. Sanjek

Nosology in Psychiatry

A nosology is a classification of disease (from the Greek words *nosos*, disease, and *logos*, discourse), whereas a taxonomy is a classification of animals or plants. A nomenclature is simply an approved list of

categories or titles. In order to constitute a classification its categories must be mutually exclusive and jointly exhaustive, and may have other formal relationships to one another as well, being arranged either in tiers or in a hierarchy.

1. History: From Hippocrates to the 1940s

For at least the last 2,500 years most generations of physicians, and several philosophers, have produced a nosology. Usually this was not because they were preoccupied with classification. It was because a nosology almost inevitably reflects the range and variety of diseases recognized by its authors, the extent of their clinical experience, their assumptions or knowledge about etiology, and their familiarity with or ignorance of the detailed manifestations of illness. Nearly all these nosologies have incorporated several categories of what we now regard as mental disorder, thereby demonstrating that their authors regarded these disorders as diseases to be treated by physicians, and usually with the same range of assumptions about causation and treatment as other diseases.

It is apparent from the corpus of writings attributed to him, that Hippocrates recognized at least six different forms of mental illness:

(a) Phrenitis (acute mental disturbance with fever),
(b) Mania (acute mental disturbance without fever),
(c) Melancholia (chronic mental disturbance),
(d) Epilepsy,
(e) Hysteria (convulsions and other paroxysmal symptoms), and
(f) Scythian disease (cultural transvestism).

His contemporary, Plato, on the other hand, distinguished between Natural Madness originating in physical disease and several kinds of Divine Madness (Prophetic Madness given by Apollo, Religious Madness given by Dionysus, Poetic Madness inspired by the Muses, and Erotic Madness inspired by Aphrodite and Eros).

Many of the nosologies developed over the next 2,000 years are listed by Menninger (1963). Most were fairly simple and incorporated Hippocrates' categories of mania, melancholia, and epilepsy. Others, like those of Boissier de Sauvages and William Cullen at the end of the eighteenth century, contained several hundred categories of disease arranged in tiers mimicking the orders, genera, and species of plants. Only in the nineteenth century, with the emergence of physicians specializing in the treatment of insanity (known as alienists), did separate classifications of mental illness start to appear. They quickly proliferated, however, because every self-respecting alienist, and certainly every professor, produced his own nosology, some based on behavioral differences, others on presumed differences in etiology, others on psychological differences, some on a mixture of all three. Indeed, as a medical historian subsequently observed, 'to produce a well-ordered classification almost seemed to have become the unspoken ambition of every psychiatrist of industry and promise.'

It was soon apparent that this multiplicity of classifications, with diverse meanings often attributed to the same terms, rendered communication between one hospital and another almost impossible, and with this came recognition of the need for a single nosology acceptable to everyone. At the Congress of Mental Medicine in Antwerp in 1885 a commission was appointed, chaired by a Belgian alienist, Morel, to produce this. The commission duly delivered a simple typology that was adopted formally by the International Congress of Mental Science in Paris in 1989, but as the English psychiatrist, Hack Tuke, had predicted, its nosology was generally ignored. However, as this was the first formal attempt to secure international agreement, and the forerunner of the present International Classification, its 11 categories of insanity have some historical significance. They were:

(a) Mania,
(b) Melancholia,
(c) Periodical insanity (Baillarger's *folie a double forme*),
(d) Progressive systematic insanity,
(e) Dementia,
(f) Organic and senile dementia,
(g) General paralysis,
(h) Insane neuroses (hysteria, epilepsy, hypochondriasis, etc.),
(i) Toxic insanity,
(j) Moral and impulsive insanity, and
(k) Idiocy, etc.

While Morel and his colleagues were engaged on this ultimately fruitless task several national organizations were pursuing similar but more modest aims. In Britain, the Royal Medico-Psychological Association (the forerunner of the Royal College of Psychiatrists) produced a classification in 1882 which it recommended to its members for general use. Further drafts were produced in 1904, 1905, and 1906 before the Association finally accepted that there was no prospect of persuading its members to restrict themselves to the diagnoses listed in any official nomenclature, British or international. A similar sequence of events took place in the USA. In 1886 the Association of Medical Superintendents of American Institutions for the Insane (the forerunner of the American Psychiatric Association) adopted a modified version of the British classification of 1882. This was revised in 1913 to incorporate Kraepelin's concepts of dementia praecox, involutional melancholia, and manic depressive insanity but, even so, many states and asylums still preferred to use their own nosologies, or declined to be bound by any nomenclature, national or local. Indeed, the New York State Commission in Lunacy continued to use its own private classification until 1968. In 1937 the need for an internationally agreed nosology was

discussed once more at the Second International Congress for Mental Hygiene. Although it soon became clear that there was still little hope of any real consensus emerging, the simple eight-category classification proposed by the French psychiatrist Bersot was adopted by two countries, Portugal and Switzerland, and remained in use in both for 20 years.

2. *The Origins of the World Health Organization's International Classification*

Largely at the instigation of two medical statisticians, William Farr in London and Jacques Bertillon in Paris, the International Statistical Congress had, as far back as 1853, recognized the need for '*une nomenclature uniforme des causes de décès applicable a tout les pays*' (a uniform nomenclature of causes of death applicable to all countries) and had not only produced such a nomenclature but revised it regularly for the next 50 years. Probably because of the importance of mortality statistics to public health and to governments, these successive editions of what came to be known as the Bertillon Classification of Causes of Death were used increasingly widely; and in 1900 the French Government assumed responsibility for the classification and convened a series of international meetings in Paris in 1900, 1920, 1929, and 1938, and thereby produced four successive editions of what was by then called the International List of Causes of Death.

When the World Health Organization (WHO) came into being in 1948 one of its first public actions was to produce a sixth revision of this International List. Renamed the International Statistical Classification of Diseases, Injuries and Causes of Death (*ICD*), this was for the first time a comprehensive nosology covering the whole range of disease, rather than merely causes of death, and so for the first time included a classification of mental illness. In fact, Section V of this sixth revision (*ICD-6*), entitled 'Mental, Psychoneurotic and Personality Disorders,' contained 10 categories of psychosis, nine of 'psychoneurosis,' and seven of 'disorders of character, behavior and intelligence,' most of them subdivided further. Unfortunately, although the nomenclature of *ICD-6* was adopted unanimously by the 1948 revision conference, and was duly 'recommended for use' by all the member states of the WHO, its mental disorders section failed to gain acceptance, and 11 years later it was found to be in official use only in Finland, New Zealand, Peru, Thailand, and the UK. History was repeating itself. Deeply concerned by this state of affairs, the WHO commissioned an English psychiatrist, Erwin Stengel, to investigate the situation and if possible to make recommendations.

The situation Stengel encountered he described as one of 'almost general dissatisfaction with the state of psychiatric classification, national and international' and the attitude of many psychiatrists towards the classification of mental disorders seemed to have become 'one of ambivalence, if not of cynicism' (Stengel 1959). Many countries had official classifications of their own, some of recent origin, others dating back to the 1930s, but none, new or old, was regarded as satisfactory by its users, or used conscientiously and consistently by them. In the UK, although the General Register Office published mental hospital admission rates and other data in the format of the *ICD*, the returns from individual hospitals from which these data were derived frequently showed a blatant disregard for, or ignorance of, the requirements of the nomenclature. In the USA, Section V of the *ICD* was ignored completely, despite the fact that American psychiatrists had taken a prominent part in drafting it. Instead, a nosology drawn up by the American Psychiatric Association and published in 1952 as a *Diagnostic and Statistical Manual* (*DSM-I*) was in widespread use. Even so, there were serious local problems. The New York State Department of Mental Hygiene still insisted on retaining its own classification and influential psychoanalysts like Menninger (1963) argued forcefully that classifying patients into categories of illness did more harm than good and should be abandoned.

3. *Reasons for Dissatisfaction with Psychiatric Classification*

Many factors contributed to the unwillingness of psychiatrists to commit themselves to using a common nosology and to the general dissatisfaction with all classifications. Most fundamentally, there was no rational criterion available to enable the psychiatric community, faced with the need to choose between several alternative nosologies, to decide which was the best, or most valid. In other branches of medicine the old diagnostic concepts and nosologies inherited from the eighteenth century or before had largely been replaced by new concepts and classifications based on at least a partial understanding of etiology. Phthisis had been replaced by pulmonary tuberculosis, myxoedema by hypothyroidism, and Mediterranean anaemia by thalassaemia, and no physician doubted the validity of these new concepts or the value of classification. Psychiatry, on the other hand, remained as ignorant as ever about both the pathology and the underlying causes of most of its common disorders, and so was compelled to continue defining these disorders by their syndromes, just as Sydenham had done in the seventeenth century, and went on using ancient Hippocratic terms like mania and melancholia.

A further problem was that many of these disorders appeared to merge into one another. A few conditions, like Down's syndrome and Huntington's chorea, did seem to be clearly demarcated but most were not.

Patients with the characteristic symptoms of one disorder not infrequently presented with the typical symptoms of a quite different disorder on a subsequent occasion, and it was commonplace to encounter patients with a mixture of the characteristic symptoms of what were supposed to be unrelated disorders. Such problems bedevilled the distinction between, for example, schizophrenic and manic depressive psychoses, and between depressive illnesses and anxiety states, and resulted in persistent controversies either about how best to distinguish the disorders in question or, even worse, to doubts about the validity or usefulness of the diagnoses themselves.

The combination of these two problems—the lack of understanding of etiology or of any demonstrable pathology, and the absence of obvious boundaries between several of the most common syndromes—meant that, in many instances, there was no rational means of choosing between alternative diagnostic concepts or alternative nosologies. In this situation most psychiatrists, understandably, opted to use whichever diagnostic terms and whichever classification were most familiar, particularly if they were their own.

Other considerations also contributed to the problem. Few of the diagnostic terms in common use held important therapeutic implications. Before the 1960s there were few treatments that were only effective in a single diagnostic category. Electroconvulsive therapy was effective in most acute psychoses—schizophrenic, melancholic, or manic. Chlorpromazine and the other new neuroleptic drugs appeared to have a similarly broad range of efficacy. Most influential of all, particularly in North America, was the fact that a diagnosis had ceased to be a necessary pre-condition to treatment, because psychoanalytically oriented psychotherapy was widely regarded as an effective and appropriate treatment across almost the whole range of psychiatric disorders from mild character disorders to florid schizophrenia. The implications of diagnoses for prognosis were equally tenuous. There were, to be sure, well established differences in outcome between populations of patients with different diagnoses, both in the short and the long term. On average, the course and outcome of schizophrenic and affective psychoses, or of depressions and anxiety states, were significantly different. But the overlap was very substantial, and as a result a diagnosis was of only limited value in predicting outcome in individual patients. By the early 1960s it was also becoming apparent that the reliability of psychiatric diagnoses, even those of experienced psychiatrists, was often alarmingly low, and that even under optimal conditions disagreements were commonplace—either because the patient gave different information to different interviewers, or because that information was interpreted in different ways, or because there were covert differences in the interviewers' diagnostic concepts.

Finally, many psychiatrists and psychologists were concerned about the harmful effects diagnostic labels often had on their patients. Labels like hysteria or schizophrenia sometimes encouraged hospital staff to treat the patients in their care as 'hysterics' or 'schizophrenics' rather than as distressed people. Several diagnostic terms also had pejorative overtones, lowering the patient's self-esteem even further than their illness itself had already done and encouraging relations, neighbors, colleagues, and employers to belittle them.

4. *The Response to these Problems*

The most important recommendations in Stengel's report to the WHO were that all diagnoses should be explicitly shorn of their etiological implications and regarded simply as 'operational definitions' for specified types of abnormal behavior, and that the next revision of the *ICD* should have an accompanying glossary 'available from the beginning in as many languages as possible' providing a brief description, or thumbnail sketch, of all the disorders in its nomenclature. Other, mainly American, psychiatrists advocated operational definitions as a means of improving reliability. The problem with the brief descriptions Stengel was advocating was that, like traditional textbook descriptions, they described the typical features of a clinical syndrome without making it clear which combinations of those features were adequate to establish the diagnosis in question. A typical textbook description, or the thumbnail sketch of a glossary, lists the typical symptoms of disease X as A, B, C, and D, and often goes on to say that E and F are sometimes present as well. But it does not spell out which combinations of those symptoms are adequate and which are not to establish diagnosis X, and as a result different diagnosticians make different assumptions, and so often end up making different diagnoses. An operational definition is simply a statement, or set of statements, which makes it clear for all possible combinations of A, B, C, D, E, and F which are and are not adequate to establish diagnosis X.

The 1960s also saw the introduction of multi-axial classifications. The Swedish psychiatrist Essen-Moller had long advocated a classification with separate axes for symptomatology (i.e., the syndrome) and etiology, primarily as a means of resolving the difficulties that inevitably occur in a classification in which some disorders are defined by their symptoms (e.g., mania) and others by their presumed etiology (e.g., puerperal psychosis). As he pointed out, having separate axes would make it possible to identify all manic illnesses, whether or not they occurred shortly after childbirth, and all puerperal illnesses regardless of their symptomatology. Shortly afterwards, a WHO working party, set up to produce some much needed recommendations for an improved classification of

childhood disorders in the *ICD*, made a firm recommendation that a tri-axial classification with separate axes for the clinical syndrome, the child's intellectual level and 'aetiological and associated factors' should be tried out and, if these pilot studies proved successful, incorporated into the next revision of the *ICD*.

4.1 The Eighth and Ninth Revisions of the International Classification (ICD-8 and ICD-9)

Stengel's report and the recommendations of its own expert committee were taken seriously by the WHO, and the next (eighth) revision of the International Classification, which was published in 1965 and came into use in 1969, was a considerable improvement on its predecessors. For the first time all psychiatric disorders were brought together in the Mental Disorders section of the classification, thus making it self-sufficient and easier to use. The sections dealing with personality disorders, sexual deviations, mental retardation, alcoholism and drug addiction were all expanded and recast and a new group of reactive psychoses was introduced at the request of Scandinavian psychiatrists. Even more important, the American Psychiatric Association was persuaded, despite considerable domestic opposition, to commit itself to using the nomenclature of *ICD-8* instead of that of its own *Diagnostic and Statistical Manual*. The German and Scandinavian psychiatric societies also agreed to use the new nomenclature so that, for the first time, all the major contributors to the psychiatric literature, with the notable exception of France, were committed officially to using the same classification. Unfortunately, the committee responsible for producing the companion glossary that Stengel had so strongly recommended did not start work until 1967 and the glossary was not published until 1974. The hiatus was partly filled, though, by the publication in 1968 of two national glossaries—American and British—to the nomenclature of *ICD-8*.

The ninth revision (*ICD-9*) was published in 1978, and this time the companion glossary was available from the beginning. But despite a series of eight international seminars held between 1965 and 1972, attended by psychiatrists from over 40 countries, there were no major changes. A few new categories, like nondependent abuse of drugs, were added in recognition of the broadening scope of psychiatric practice, a clearer distinction (reminiscent of Plato's distinction between natural and divine madness) was drawn between organic psychoses and other psychoses, and there was a substantial expansion of the number of categories allocated to disorders of childhood, but that was all. None of the fundamental problems, like the low reliability of diagnostic assignments, the tensions created by the incorporation of etiological assumptions into diagnostic terms, and the increasingly complex and chaotic classification of depressions, had been tackled effectively. It was also clear why this was so. The need for international consensus—the agreement of nearly 200 sovereign states to use the same nomenclature—was making it almost impossible to introduce radical changes; and an increasing number of sometimes incompatible terms were being added to the nomenclature with each new revision, because individual psychiatric societies always fought for the inclusion of their own favourite terms more persistently than they were prepared to oppose the attempts of others to do likewise. It was becoming apparent, in other words, that if fundamental improvements were ever to be achieved it was unrealistic to expect the WHO to be the driving force.

5. The Third Edition of the American Psychiatric Association's Classification (DSM-III)

For the research community the most important shortcomings of all classifications of mental disorders, national and international, were the low reliability and uncertain validity of most of their diagnostic terms. The Washington University Department of Psychiatry in St Louis in the USA reacted to this situation by publishing what came to be known as the Feighner criteria—a set of operational definitions, intended for research use, of the 15 major syndromes they regarded as having been adequately validated (Feighner et al. 1972). Other rival sets of definitions, like the Research Diagnostic Criteria, soon followed, and in this climate the American Psychiatric Association committed itself to producing a new, third, edition of its *Diagnostic and Statistical Manual*. When it was finally published in 1980 this *DSM-III* was dramatically different from any previous nosology of mental disorders. It was multi-axial with independent recording systems for five different kinds of information—the clinical syndrome or syndromes; associated personality or developmental disorders; associated physical disorders; the severity of recent psychosocial stresses; and the highest level of social functioning in the past year. The grouping of clinical syndromes was also novel, with all affective disorders brought together and the traditional distinction between psychotic and neurotic disorders abandoned. Even more provocatively, many of psychiatry's most hallowed terms, like hysteria, manic-depressive illness, and even psychosis and neurosis, were discarded and replaced by unfamiliar and rather unattractive terms like somatoform disorder, substance use disorder, and paraphilia. A brave but only partially successful attempt was also made to define the term 'mental disorder.' Finally, and most important of all, nearly all the 200 categories of disorder in the manual were provided with operational definitions, and it was demonstrated in extensive field trials that this made it possible to achieve higher reliability than had ever been attained previously.

None of the innovative features of *DSM-III* was entirely novel. The advantages of both multiple axes and operational definitions had been pointed out long before and both had already been introduced into clinical practice in some centers. The concept of hysteria had been under attack for over a generation, as had the distinction between psychotic and neurotic disorders. What was new was that a powerful professional body like the American Psychiatric Association was prepared to commit itself and its members to all these innovations simultaneously. For a time the fate of *DSM-III* hung in the balance, for in the eyes of many older and still influential psychiatrists its authors had gone too far. But before long its advantages were increasingly apparent and eventually it became a best seller, not only in the USA but worldwide.

With hindsight the adoption of *DSM-III* was a crucial step in the development of a scientific nosology of mental disorders. Certainly, all subsequent nosologies, national and international, have been deeply influenced by it. Of all its innovative features the commitment to operational definitions was the most important. Not only did this result in a major improvement in the reliability of psychiatric diagnoses and eliminate the unsuspected differences in diagnostic criteria that had previously bedevilled international comparisons, they made it explicit that diagnostic terms like schizophrenia were simply convenient designations for more or less arbitrary groupings of symptoms.

6. *Contemporary Nosologies*

6.1 *The Fourth Edition of the American Psychiatric Association's Classification (DSM-IV)*

The American Psychiatric Association published a revision of *DSM-III* in 1987 and a new fourth edition (*DSM-IV*) in 1994. Neither abandoned any of the fundamental features of *DSM-III*. For the most part they simply corrected the anomalies that had come to light, reworded several operational definitions in order to improve their reliability or to conform with changes in expert opinion, and introduced a few additional categories of disorder. A new 'clinical significance' criterion was also added to the definition of many disorders in *DSM-IV* to counter the criticism that too many people met the *DSM-III* or *DSM-IIIR* criteria for common conditions like major depression and phobic disorder and that many of them were not significantly handicapped.

The authors of *DSM-IV* also endorsed and expanded the rather cumbersome *DSM-III* definition of 'mental disorder' as 'a clinically significant behavioral or psychological syndrome or pattern that occurs in an individual and that is associated with present distress or disability or with a significantly increased risk of suffering death, pain, disability, or an important loss of freedom.' In addition, it was stipulated that 'this syndrome or pattern must not be merely an expectable and culturally sanctioned response to a particular event, for example, the loss of a loved one,' and that 'whatever its original cause, it must currently be considered a manifestation of a behavioral, psychological, or biological dysfunction in the individual,' thereby emphasizing that 'neither deviant behavior nor conflicts that are primarily between the individual and society are mental disorders unless the deviance or conflict is a symptom of a dysfunction in the individual.' The introduction to *DSM-IV* also emphasized, rather more succinctly, that 'the term mental disorder unfortunately implies a distinction between mental disorders and physical disorders that is a reductionistic anachronism of mind/body dualism. A compelling literature documents that there is much physical in mental disorders and much mental in physical disorders. The problem raised by the term mental disorders has been much clearer than its solution, and, unfortunately, the term persists in the title of *DSM-IV* because we have not found an appropriate substitute.'

6.2 *The Tenth Revision of the International Classification (ICD-10)*

The WHO published a tenth revision of the *International Classification* in 1992, after extensive consultations and field trials in several different parts of the world. Although the titles and definitions of many of the disorders listed in the Mental and Behavioural Disorders section of this *ICD-10* are different from those of either *DSM-III* or *DSM-IV* its debt to the former is obvious. The general format of the two nosologies is very similar and, crucially, *ICD-10* provides 'diagnostic criteria for research' for every category (i.e., operational definitions) as well as 'clinical descriptions and diagnostic guidelines' for ordinary clinical use (WHO 1992, 1993). This was only possible because, by the time the member states of the WHO were called upon to endorse the draft text of *ICD-10*, there had been time for the international psychiatric community to come to terms with the radical innovations of *DSM-III* and to appreciate their advantages. Significantly, though, no attempt was made to define the term 'mental or behavioral disorder.' Indeed, there is an apologetic note in the introduction explaining that 'the term disorder is used … so as to avoid even greater problems inherent in the use of terms such as disease and illness. Disorder is not an exact term, but it is used here to imply the existence of a clinically recognizable set of symptoms or behavior associated in most cases with distress and with interference of personal functions.'

At present, therefore, two nosologies of so-called 'mental disorders' are in widespread use—*DSM-IV* and *DC-10*. In their essentials they are very similar to one another. In points of detail, particularly the titles and subdivisions of individual syndromes and the

detailed definitions of these disorders, they are often different. The American classification is used mainly in North America and by research oriented psychiatrists further afield. The *International Classification* is used elsewhere, particularly in Europe and developing countries, and by all governments for their formal statistical returns to the WHO. In some ways it is regrettable that there are now, and will be for the foreseeable future, these two rival nosologies. However, the confusion generated by two rivals is much less than that produced by a dozen, the nightmare of the 1950s. There are also advantages in having some choice. The coexistence of alternative titles, subdivisions, and definitions for several of the major syndromes helps to emphasize their arbitrary nature. It also invites formal comparisons, which may help to provide the evidential base for better categorization and definition in the future. It is likely, too, that for financial reasons the American classification will be revised more frequently in the future than the *International Classification*, and some sections of the nosology may need to be revised quite quickly in the light of etiological research.

7. *Shortcomings of Contemporary Nosologies*

7.1 *Uncertain Validity*

The most obvious weakness of both *DSM-IV* and *ICD-10* is the limited evidence for the validity of most of their constituent categories. The combination of a limited understanding of the etiology of most mental disorders and the lack of evidence for natural boundaries or 'zones of rarity' separating their clinical syndromes (see *Differential Diagnosis in Psychiatry*) means that most existing diagnostic concepts and definitions are not securely based. There are exceptions, of course. The etiologies of, for example, Huntington's chorea, Down's syndrome, the Fragile X syndrome, and Creutzfeldt Jacob disease are all well established, and enough is known of the neuropathology and etiology of Alzheimer's disease and vascular dementia, and of the various syndromes associated with the use of psychoactive substances, for them to be reasonably secure. But the same cannot be said for most of the other major syndromes. For many, there is cogent evidence that genetic factors are making a major contribution to their etiology, and that life experiences and other environmental factors are also involved. It seems increasingly likely, though, that many different genes, each on its own of relatively small effect, are responsible for the genetic basis of most syndromes, and that some of these genes are risk factors for more than one syndrome—for affective as well as schizophrenic psychoses, for example, or for anxiety states as well as depressions. Some adverse life experiences appear to predispose to a similarly broad range of apparently unrelated disorders.

7.2 *The Dimensional Alternative*

An etiological framework of this kind does not easily lend itself to a categorical classification of disorders. For this and other reasons the possibility of representing variation in symptomatology by dimensions rather than categories needs to be considered. The idea is not, of course, new. Wittenborn developed an elaborate multidimensional representation of the phenomena of psychotic illness in the mid-twentieth century and other psychologists and psychiatrists have subsequently advocated and used dimensional models to portray depressive or chronic schizophrenic illnesses, or even mental disorder as a whole. Dimensional classification solves at a stroke all the problems associated with boundaries and comorbidity, and may also be a powerful means of predicting outcome. Even so, there is no realistic prospect at present of any formal national or international classification adopting a dimensional format. Medicine as a whole is too firmly committed to categories of illness, and in any case dimensions are better suited to the portrayal of variation in populations than to day-to-day decisions about the treatment of individual patients. This does not mean, though, that dimensional systems may not have an important role for research purposes, or as an experimental alternative to a traditional nosology.

8. *Future Developments*

In time, as understanding of the underlying determinants of most of the major syndromes slowly accumulates, mental disorders will cease being defined by their syndromes and be defined instead by their neuropathology, or perhaps by a characteristic combination of genetic risk factors. This may not happen quickly, and when it does the classification of mental disorders—if they are still so called—may bear little resemblance to our present nosologies. In the immediate future, however, major change seems improbable. The American Psychiatric Association has already committed itself to producing a fifth edition of its *Diagnostic and Statistical Manual* (*DSM-V*), probably some time between 2007 and 2010, and at present it seems unlikely that it will differ in any radical way from *DSM-IV*.

See also: Differential Diagnosis in Psychiatry

Bibliography

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III*), 3rd edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*), 4th edn. American Psychiatric Association, Washington, DC

Feighner J P, Robins E, Guze S B, Woodruff R A, Winokur G, Munoz R 1972 Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry* **26**: 57–63
Kendell R E 1975 *The Role of Diagnosis in Psychiatry*. Blackwell, Oxford, UK
Menninger K A 1963 *The Vital Balance: The Life Process in Mental Health and Illness*. Viking Press, New York
Stengel E 1959 Classification of mental disorders. *Bulletin of the World Health Organization* **21**: 601–63
World Health Organization 1974 *Glossary of Mental Disorders and Guide to their Classification: For Use in Conjunction With the International Classification of Diseases, 8th revision*. World Health Organization, Geneva, Switzerland
World Health Organization 1978 *Mental Disorders: Glossary and Guide to their Classification in Accordance with the Ninth Revision of the International Classification of Diseases*. World Health Organization, Geneva, Switzerland
World Health Organization 1992 *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. World Health Organization, Geneva, Switzerland
World Health Organization 1993 *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization, Geneva, Switzerland

R. E. Kendell

Nostalgia, Anthropology of

Nostalgia Isn't What it Used to Be, laments the title of actress Simone Signoret's autobiography (1978). The title reflects that sense of nostalgia that periodically accompanies feelings that life does not seem quite as vivid, or quite as good, as it did in an earlier, more golden time. Signoret's title epitomizes this nostalgic sense, by suggesting that even the present's poignant pangs of sentimental longing for the past seem neither quite so romantic, nor quite so melancholic as they once did.

The individual experience of nostalgia involves a sentimentalized longing for the past. Most often, however, this is a combination of a remembered, an imagined, and a reinterpreted past, which in memory seems more benevolent, loving, and problem-free than the actual past was. Individual nostalgia typically involves recreated memories of childhood as a time of innocence, when life seemed to follow a set course, and one felt embraced by family and community in a nexus of human and humane relationships, rather than encumbered by legalistic and bureaucratic ones. Childhood is, thus, often recaptured in nostalgic memory as unproblematic, when it was likely fraught with its own problems and tensions. The anthropology of nostalgia attempts to understand how such flushes of romanticized longing for the past that individuals periodically feel, become part of the collective imagination. It seeks to explain why nations, communities, or other collectivities of people seem to experience widespread cycles of nostalgic reminiscing for a presumed shared past.

1. *Uncertainty and the Collective Imagination*

The anthropology of nostalgia suggests that collective nostalgia seems to arise in times of transition or uncertainty. Surging nostalgic sentiment frequently underlies a collective identity crisis indicating a sense of confusion or uncertainty surrounding a proclaimed cultural or national identity. Increasing urbanization, technological changes, or globalization, all of which can cause swift changes in local lifestyles, can also give rise to collective nostalgia. Even among people seeking change, a seemingly contradictory combination of the rhetoric of progress, with a pronounced increase in the rhetoric of nostalgia is frequently noted. The anthropology of nostalgia suggests that this apparent contradiction is really a corollary response, indicating a fundamental ambivalence to sudden change and so-called modernization. Some anthropologists, such as Ivy (1988, 1995), contend that the contemporary condition of postmodernism, typified by rapidly changing lifestyles accompanied by the pronounced sense that life is shifting and uncertain, induces collective nostalgia.

The sense that life is shifting and uncertain can also be associated with the feeling that one no longer has strong emotional bonds to other humans. In his treatise on the sociology of nostalgia, Davis (1979, p. 141) contends that collective nostalgia is often a response to cultural transitions that leave masses of people with feelings of loneliness and estrangement. According to Davis, such nostalgia is part of a 'collective search for identity' which 'looks backward rather than forward, for the familiar rather than the novel, for certainty rather than discovery' (Davis 1979, pp. 107–8). Smith (1982, p. 128) notes that the 'past will always seem more stable than the present.' Therefore, images of the past are invoked to offset the perceived threat of cultural change or identity loss. These tend to be highly romanticized images of the past. Tuan points out 'the cult of the past calls for illusion rather than authenticity' (1977, p. 194).

Nostalgic projections of a sentimentalized past involve what Hobsbawm and Ranger (1983) call 'invented traditions.' This term does not necessarily mean such traditions are false, but rather recognizes that the past is constantly undergoing recreation and reinvention by those in the present, in response to present needs.

2. *Country, Community; Childlike, Calendric: The Place and Time of Nostalgia*

Anthropologists working in different areas of the world have noted that collective nostalgia frequently romanticizes rural areas, a sense of community, childhood, and calendric rituals. Romanticization of the countryside or rural life is common in collective

nostalgia. Increasing urbanization often draws people away from small towns and villages where they belonged to networks of relationships, into the anonymity of city life. This influx of population into cities can also break down existing urban communities. Concomitantly, increasing industrialization and technological development change not only the methods humans employ in productive work activities, but the relationships constructed via these activities. The *gemeinschaft* involvement thought to characterize preindustrial communities, in which life and work found meaning within a network of social relationships, is replaced by *gesellschaft* society, in which human interactions are more purely economic or business exchanges (Toennies 1957). Romanticization of rural areas can result from urban anomie, or from frustrations with a perceived *gesellschaft* lifestyle.

Romanticization of rural life frequently is noted in areas that have undergone rapid urbanization and industrialization. In Western Europe, industrialization created a romanticized image of the preindustrial countryside. Twentieth-century North America saw periodic flushes of nostalgia for small town life. Some Asian countries have experienced nostalgia for agrarian village life after periods of rapid industrial or economic growth.

Such romanticization of rural life suggests a desire to negate feelings of estrangement, by imagining a place of belonging. The particular images of rural life vary in each area noted, but all suggest a desire to return to a closer bond with nature, and to the nexus of reciprocal human relationships idealistically projected as typifying 'home town' or 'home village' life. Affirmations of a rural centrality, of the countryside, or of a 'back-to-nature' lifestyle, proffer the hope that a return to the presumed virtues of the past, can offset the negative consequences of industrialization and technological development. The nostalgic emulation of deeply embedded human relationships thought to characterize rural areas, and 'natural' lifestyles, involve a quest for community and the desire to affirm a collective identity.

'Community' is a powerful, multivocal symbol in many cultures. Despite variations in its particular ideal constructions, generally it speaks to a desire for belongingness, to an essential need for human interaction in order to lead meaningful lives. Modern nostalgia, as a response to widespread feelings of homelessness and feared cultural loss, projects symbolically mediated images of a past rural existence to address the threat of estrangement by embracing an ideal of collective belongingness that represents 'community,' the 'good life,' 'wholesomeness,' and moral virtue.

For many individuals nostalgic projections of a home town, or home village past, represent what Bellah calls 'communities of memory,' rural areas or home town communities where they actually did grow up. According to Bellah (1985, p. 154), people who grew up in 'communities of memory' know the stories of a community's past, share its ideal values, and understand what it means to participate in the ritual, aesthetic, and seasonal practices that defined the community as a way of life. However, such imagery can invoke intense nostalgia even among those who did not grow up in such communities through the construction of 'communities of imagination.'

Nostalgia for a past way of life, which rural imagery suggests, is associated with nostalgia for the personal past, or childhood. Sentimentalized projections of community and of childhood both reflect a desire to find a 'place' where one belongs and feels secure. In mass nostalgic longings, there is a tendency to collapse geographical and temporal dimensions. Rural areas are geographical places that exist contemporaneously with urban areas. However, in the collective landscape of nostalgia, rural areas become symbols of an earlier existence. Childhood is a time in people's lives. However, in the nostalgic imagination, this time becomes a place. Childhood becomes that place where one felt sheltered and secure. Reflecting this emphasis on childhood, nostalgic imagery frequently projects maternal tropes. Both 'mother' and 'community' suggest intimate places of nurturance. For the young child, mother is not just a person who provides love and warmth, but the child's primary 'place,' a 'haven of stability' where 'fundamental needs are heeded and cared for without a fuss' (Tuan 1977, pp. 137–8). The idealized past community is likewise that 'place' where one can belong, where one has obligations to others, and can depend on others.

Calendric rituals and holidays are also a profound focus of collective nostalgia. Cultural calendars punctuate experience with special days laden with symbolic meanings that have the capacity, in Zerubavel's (1981, p. 46) terms, to 'synchronize the sentiments of a large number of people' by creating 'emotional rhythms that affect large collectivities.' Special calendric events integrate community life. They also play important symbolic and affective roles in both the nostalgic remembrance of community, and the nostalgic imagination of community. For example, in an analysis of English Victorian life, Gillis (1989) contends that Christmas offered a nostalgic spiritual pilgrimage, a symbolic journey of return and reunion to home, community, and 'mother.' In Japan, *Gyoji* are the annual cycle of calendric festival events, emotionally situated in seasonal associations, that formerly orchestrated life in agrarian villages. Creighton (1998) describes how the nostalgia-driven 'retro-boom' prompted collective revivals of this traditional calendar of *gyoji* events, in rural and urban areas.

3. *Nostalgia and Nationalism*

Anthropologists have pointed out that nostalgia is used to promote nationalism, through the projection

of emotionally laden images which are then linked to a national way of life. This can involve a positive affirmation of a proud collective identity for a nation, or conversely, a negative and dangerous fostering of ultranationalism. In his famous treatise on nationalism and invented communities, Anderson says all national communities are imagined. He further stresses the importance of understanding 'the style in which they are imagined' (1985, p. 15).

The nostalgic invocation of a presumed pristine past, and sentimental reaffirmations of nostalgic holidays, are part of the stylistic imagination of national communities. Anderson claims that Christmas in modern Western nations functions as a calendric event which helps perpetuate 'that remarkable confidence of community in anonymity which is the hallmark of modern times.' Despite its appeal in different Western nations, its sentimentally loaded symbolic associations with growing up in a particular place, allow it to be associated with an imagined national community. Likewise in many Asian countries, New Year festivi-t-ies frequently are used to reaffirm national identity, even though the holiday exists and is celebrated in different forms in other places.

Nostalgic associations with childhood experience are also used to promote nationalistic sentiment. 'Motherhood' frequently becomes a dominant, multivocal symbol used in the promotion of nationalism in diverse nation states. Creighton (1991) explores the irony of motherhood as a nationalistic symbol. 'Mother' as a symbol could suggest a universal or shared humanity, since everyone presumably had a mother. However, mother symbolism is frequently used to emphasize a nationalistic identity and the exclusion of outsiders, by projections of a culturally specific mother prototype that is then linked to national identity.

In times of political, social, or economic strain, global or regional uncertainty, such uses of nostalgia in the promotion of nationalism can usurp grass roots attempts to reinvigorate bonds among people in local communities. At the other end of the spectrum, nationalistic uses of nostalgia can obviate against attempts to build associations that foster closer ties among human beings that go beyond national boundaries. Nationalistic nostalgia offers a sense of belonging to an imagined national community, but it is a form of belongingness which frequently requires the exclusion and even repudiation of nonmembers.

4. Conclusions

The anthropology of nostalgia attempts to understand why there appear to be periods of mass nostalgia. Anthropologists have noted that collective nostalgia frequently accompanies or follows periods of change, uncertainty, or alienation.

The collective landscape of nostalgia is typically rural in nature, reflecting an apparent collapse of time and place in the human imagination. Rural areas exist in present time, but are romantically envisioned as representing a past, more pristine way of life. In this collective nostalgic romanticization, place has been confused with time. Significant people of childhood, such as mother, and calendric holidays of a given culture that are emotively experienced during childhood, have strong nostalgia-inducing capacity. In this collective nostalgic romanticization time—the time of childhood—has become a place, a place where security and belongingness seemed assured.

Nostalgia can be channelled to promote nationalism through the projection of emotionally laden symbolism which is then asserted to represent 'our way of life,' making the nation–state into an 'imagined community.' Anthropologists call this an 'imagined community' because in real communities the ties that bind people are created through actual, on-going, personal interactions. The use of nostalgia to promote nationalistic sentiment can have positive consequences in creating a positive collective identity, or negative consequences in fostering ultranationalism.

Collective nostalgia can involve a melancholic longing for the past that prevents people from moving forward. However, many anthropologists suggest that collective nostalgia can also have positive effects. Nostalgic recreations of the past can inspire people to attempt to create deeper human relationships, and a stronger sense of community in their present lives. Nostalgia for the past can be part of a constructive process that allows people to deal with the uncertainty of the present in order to move forward. Collective romanticized sentiments help create a nostalgically imagined assertive linking of the present with the past, and thus with the future. Images of a presumed past in which life seems to have been more stable and loving often help people to offset the threats of insecurity and alienation in the present, allowing them to move forward into an uncertain future with renewed assurance that there is a continuity between past, present, and future, and that the human values they associate with the past still do and will continue to exist.

See also: Collective Beliefs: Sociological Explanation; Collective Memory, Psychology of; Community, Expression of; Community Studies: Anthropological; International Organization; Nationalism and Expressive Forms; Nationalism: General; Nostalgic Social Thought; Ritual; Tradition, Anthropology of; Traditions in Sociology

Bibliography

Anderson B R O'G 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Bellah R N, Madsen R, Sullivan W M, Swidler A, Tipton S M (eds.) 1985 *Habits of the Heart: Individualism and Commitment in American Life*. University of California Press, Berkeley, CA
Creighton M R 1991 Maintaining cultural boundaries in retailing: How Japanese department stores domesticate 'things foreign.' *Modern Asian Studies* **25**(4): 675–709
Creighton M R 1998 Pre-industrial dreaming in post-industrial Japan: Department stores and the commoditization of tradition. *Japan Forum* **10**(2): 1–23
Davis F 1979 *Yearning for Yesterday: A Sociology of Nostalgia*. Free Press, New York
Gillis J R 1989 Ritualization of middle-class family life in nineteenth century Britain. *International Journal of Politics, Culture, and Society* **3**(3): 213–35
Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK
Ivy M 1988 Critical texts, mass artifacts: The consumption of knowledge in postmodern Japan. *South Atlantic Quarterly* **87**(3): 419–44
Ivy M 1995 *Discourses of the Vanishing: Modernity, Phantasm, Japan*. University of Chicago Press, Chicago
Signoret S 1978 *Nostalgia Isn't What it Used to Be*. Weidenfeld and Nicolson, London
Smith M E 1982 The process of sociocultural continuity. *Current Anthropology* **23**(2): 127–42
Toennies F 1957 *Community and Society* [*Gemeinschaft und Gesellschaft*]. Michigan State University Press, Lansing, MI
Tuan Y-F 1977 *Space and Place: The Perspective of Experience*. University of Minnesota Press, Minneapolis, MN
Zerubavel E 1981 *Hidden Rhythms: Schedules and Calendars in Social Life*. University of Chicago Press, Chicago

M. R. Creighton

Nostalgic Social Thought

The word nostalgia is a modern one, but the group of sentiments that it designates is as old as human language. Johannes Hofer (1688), a young medical student at the University of Basle coined the word in his thesis *A Medical Dissertation on Nostalgia*. This neologism, made up of two classical Greek words (*nostos*, meaning return and *algos*, meaning pain) added to the clinical dictionary an illness already known in German-speaking Switzerland, and in particular in the Canton of Bern, by the name of *Heimweh*. It referred to a condition that was especially widespread among Swiss soldiers stationed at faraway garrisons and in foreign countries. Since the romantic period, the word nostalgia has been freed from its strictly medical meaning and echoed in other languages with expressions such as *mal du pays*, *rimpianto*, homesickness, *añoranza*, and *saudade*, although never quite coinciding with these terms. Nostalgia underwent a transformation analogous to that undergone by the ancient word 'melancholy.' It went from being an illness to describing a feeling.

To the doctors of the end of the seventeenth century, cases of nostalgia were demonstrated by states of high fever, hallucinations, depression, slow blood circulation, nausea, lack of appetite, insomnia, and delirium. These symptoms were explained by the young Hofer as being due to the concentration of 'vital spirits' and to moods generated in the singular location of the mind and around a single idea, that of the fantasy of returning to the homeland. Although the therapy recommended for cases of nostalgia—the patient's return home—was agreed upon almost unanimously, widely differing ideas were held about the causes of the illness. Following Hofer, observers (such as Scheuchzer and Du Bos) place emphasis on the air. Leaving behind the pure air of the mountains may have been responsible for the fall into the state of nostalgia. However, they also looked for the social causes at the basis of the illness. Rousseau, for example, highlights the link between poverty and nostalgia. Since in the state of poverty in which some mountain villages found themselves the basic domestic element is all that their inhabitants possess, the occurrence of the loss of that element provokes nostalgia. Rousseau also predicts that with the spread of affluence the illness could disappear. An illustrious Swiss doctor, Baron Albrecht von Haller, author of the article *Nostalgie* in the *Supplement* of the *Encyclopédie* (Haller 1777), locates the causes of the illness in the very structures of the Swiss cantons. The resistance of the cantons to contamination due to the arrival of foreigners, and to their being granted citizenship, was thought to be responsible for the strong nostalgia that struck these youths on abandoning their mountain villages where a closed, familiar world was the dominant or exclusive one.

While at first considered by Swiss doctors such as Scheuchzer, Zwinger and Tissot to be an illness particular to the military, later nostalgia was analyzed as an illness that can strike further afield—for example, in young women who enter into service in foreign countries. It also became an object of study in colonial medicine.

The symptoms found in cases of nostalgia are, throughout the eighteenth century and for the first few decades of the nineteenth century, those described by the doctor De Sauvages, professor at Montpellier (1763). These are *morositas*, *pervigilia*, *anorexia* and *asthenia* or, in other words, dullness and lethargy, insomnia, lack of appetite, and weakness. Often, above all in the garrisons of Swiss soldiers, the state of nostalgia was believed to be caused by a pastoral song, the '*ranz des vaches*,' the song accompanying the animals' return to the stables at sunset. The song was, in fact, later prohibited by cantonal decree. However, for many observers, including Zwinger, Rousseau, and Haller, the very fact of listening to this melody, or other voices that recall the faraway homeland, in a foreign country is what arouses the feeling of nostalgia. This movement that, from the timbre of a voice, from the motif of a song, from a rhyme, from a countryside or mountain lullaby, suddenly shifts towards another

voice or towards a face once familiar but now far away, is the very journey of nostalgia. This movement exhibits a similarity to the phenomenon described by Freud as *das Unheimliche*: the return to familiar images yet under a wave of foreign words in foreign places and times, in sum the return of the well-known from within the strange and the unknown, the perturbed and anguished creation and transformation of the domestic (*Heim*) into the threatening (*unheimlich*).

Due to the linguistic nature of this relationship between sound and imagery, and to the even more widespread experience of this turbulent remembering, nostalgia eventually became removed from medical attention. From being the subject for the clinic it moved into the domain of feelings. That is to say, it moved to a weaving of passions that literature—not only that of the romantic variety—elects as its chosen field of observation and analysis. The entry on *Nostalgie* in the *Encyclopédie méthodique* (Pinel 1821), the first part written by Pinel and the second by Boisseau, while dealing with the description of symptoms, shifts the illness toward a more general condition of sadness and anguish brought about by being far from home. All those who had experienced being far away from home could be affected by nostalgia. By the middle of the nineteenth century, the clinical framework had been enlarged. The description of the symptoms became so extensive that the positions defining nostalgia as a modern form of the ancient melancholia are justifiable. Some, such as Mutel (1849), claim that in order better to define the nature of the illness, the word *apodalgia* (*apodèin*: to be far away) would be more suitable. Others include in the diagnostic framework the numerous examples of nostalgia to be found among animals, especially of the domestic variety (although Buffon, in the *Histoire Naturelle*, had already recounted cases of nostalgia among horses). Although nostalgia disappeared from the treatises of clinical medicine, it remained a term used by psychiatry to explain, in conjunction with other causes, certain pathological behaviours, in particular among adolescents in the metropolises.

Linguistic dictionaries began to contain the term nostalgia rather late, and in many languages the French word *nostalgie* is often indicated as being its source or original root, rather than the Greek composite derived in 1688 by Johannes Hofer, a student of Johann Jacob Harder. The *Dictionnaire de l'Académie Française* added the word in 1839, and the Italian *Vocabulario degli Accademici della Crusca* in 1863. Yet, as far back as the second decade of the nineteenth century the use of the term, not only for the designation of an illness, but also for indicating an internal condition and state of mind, became diffused throughout the Europe of exiles and *émigrés*. Through this, beginning at the end of the nineteenth century and little by little throughout the twentieth, the word nostalgia has been added to dictionaries of the Romance, the Anglo-Saxon, and the Slavic languages. Curiously, however, the word entered much later into Greek language and culture despite its origins in the myth of the Odyssey, the classic hero of return, or *nostos*.

Entering into the language of narration and poetry, nostalgia opened out into a fan of meanings and was contaminated by the forms originating in romantic *reverie* or in the spleen, and from perplexing, turbulent, and undefined forms of memory. In Victor Hugo's writing, nostalgia is linked strongly to the experience of exile. However, in the case of Baudelaire who, even in his youthful journey to the tropics had felt the wound or, according to his mother's account, the disease, as an emptying of his self of all content: the distant country that one is longing for is the country never to be known ('*nostalgie du pays qu'on ignore*'). After Baudelaire, nostalgia may still preserve the reference to distant times and places, yet at the same time may make this link fade into emptiness, and into the undefined and vague sensation of distance and foreignness.

In separating itself from pathology and progressively inhabiting the forms and modes of feeling, nostalgia incorporates multiple, and often undefined, semantic areas. Its objects can at times be Eden and at others, the language that comes before a Babylon, childhood, the good old days, the frugality of customs, or the unknown. Yet, above all, the specific notion of nostalgia relates to the separation and distance from one's own country and language, and from the faces of acquaintances and friends. The feeling of nostalgia often arises from the representation of the 'birthplace' as a faraway point in space and time. But beyond the evocation of place, poets have established a relationship and link with the mother tongue, which then becomes the very root of poetry.

Nevertheless, in the representation of nostalgia, the re-rooting of a physical and delineated place hides that which is in fact distant, unreachable, and lost: the 'time' lived in that place. It was Kant who, in a passage of his work *Anthropology*, suggested that it is not a place that is sought in the desire to return but a time, the time of youth: the therapy of returning home advised in clinical cases of nostalgia leads to both delusion and recovery as, in returning to the homeland, one is faced with the realisation not only that the place has changed, but that the times passed therein are over for ever. The real object of nostalgia is the 'irreversibility of time': 'the object of nostalgia is the misery of the irreversible' wrote the French philosopher Jankélévitch in his (1974) book dedicated to the subject.

On the dark threshold of a time never to return, nostalgia tours the shadows of the interior, becoming a silent, painful conversation with the that which once was, and no longer is. Literary writing feeds on this conversation: both the 'remembering' of which Leopardi speaks, as well as the '*mémoire involontaire*' that is the theme of Proust's *Recherche* are connected

in this sense with nostalgia. In the irreversible and in oblivion memory opens passages, searching for the language for the representation of what is lost, and which only language may bring back to life. This journey taken by literature transforms nostalgia from a closed and regressive element into an open, fantasizing, and meditative one.

The ambivalence of nostalgia—obsessive and fantastic, anguished and creative—also crosses into the condition of exile and emigration. The literature of exile—from Ovidio's *Tristia* to the works of the Russian *Emigrantkaja Literatura*—entertained an inventive and profound relationship with the language of origin, and with the images arising from loss and irreversible separation. Nostalgia is able to transform the experience of disorientation in the experience of the encounter with the other.

However, against this development of language and knowledge, there is, or there can be, a mode of nostalgia that believes in the possibility of returning to that which once was, or that is imagined to have been. In reference to this regressive trend, some languages have registered the use of the adjective 'nostalgic': nostalgic people are those who mourn a regime or political system after being taken over by a new order. The view of the past can become an illusion of the return and rebirth of the past itself. This working of the imagination can feed fantasies of origin, or obsessive searching for roots, either by individuals, or by a groups. But the fervour of belonging—of ethnicity, language, custom, and 'civilization'—can lead nostalgia into the exercise of discrimination and violence, as the twentieth century has witnessed tragically. Forms of racism and persecution have looked for their bases in nostalgia and belonging, either to an epoch, or to an abstract or mythical principle. On the other hand, the emergence and development of the feeling of nostalgia in language, in the representation of fantasy and memory tell a different story, one belonging to literature, music, the arts, and linguistic forms of internal experience.

See also: Aristotelian Social Thought; Imagination: History of the Concept; Marxist Social Thought, History of; Pragmatist Social Thought, History of; Republicanism: Impact on Social Thought; Romanticism: Impact on Social Thought; Social Question: Impact on Social Thought; Totalitarianism: Impact on Social Thought; Tradition, Anthropology of; Traditions in Sociology; Utilitarian Social Thought, History of; Utopias: Social; Weberian Social Thought, History Of

Bibliography

Frigessi Castelnuovo D, Risso M 1982 *A Mezza Parete. Emigrazione, Nostalgia, Malattia Mentale*. Einaudi, Turin

Haller A von 1777 Nostalgie. In: *Supplément aux Dictionnaires des Sciences. des Arts et des Métiers*. L'imprimerie des éditeurs, Livonno, Vol. 4

Hofer J 1688 *Dissertatio medica de nostalgia oder Heimwehe* [A medical dissentation on nostalgia]. Typis Jacobi Bertschii, Basle

Jankélévitch V 1974 *L'irréversible et la Nostalgie*. Flammarion, Paris

Masciongelo P, Vecchio S (eds.) 1989 *Nostalgia Scritti Psicoanalitici*. P. Lubrina, Bergamo

Pinel Ph 1821 Nostalgie. In: *Encyclopédie Méthodique. Medicine*, Chez H. (April–September): Agasse, Paris, Vol. 12

Prete A 1992 *Nostalgia. Storia di un Sentimento*. Cortina, Milan

Starobinski J 1966 Le Concept de Nostalgie. *Diogène* **(April-September)**: 92–115

A. Prete

Notestein, Frank Wallace (1902–83)

1. Life and Career

Frank Notestein did more than any other person in the twentieth century to develop demography as a field of study in the USA, and probably the world. He was born on August 16, 1902, in Alma, Michigan to a family with academic connections. His undergraduate education was in the mid-West in Alma and Wooster colleges, but he earned his Ph.D. in Economics at Cornell University. There he came under the influence of Walter Willcox, which was important both because Willcox was interested in social and population statistics and because of his connections with the Milbank Memorial Fund (MMF), New York, which was to become an important early funder of population research. At Wooster Notestein met Daphne Limbach whom he married in 1927.

Notestein worked with Edgar Sydenstricker from 1928 to 1936 helping to establish the MMF's population program, carrying out studies of both differential fertility and contraception, and arranging for MMF to fund a demographic survey in rural China. During this time the *Milbank Memorial Fund Quarterly* became, and remained for over a decade, the primary journal where demographic research was reported. In 1936 Frederick Osborn convinced Albert Milbank to fund Princeton University to establish the Office of Population Research (OPR). This was the first program in the USA to combine demographic research and graduate education. Notestein became its first head, 1936–59, directing and synthesizing its work with profound consequences for demographic thought. Because of Princeton's departmental structure OPR's faculty were assigned to appropriate university departments (Notestein to the Department of Economic and Social Institutions), as were its students for most of their course work. OPR provided

a single course, given for the first years solely by Notestein, for which it was later to award its own certificate. This was where most of OPR's demographic discoveries and theories were first taught. OPR, because of its head start and its high standards, provided from its faculty and students an extraordinary proportion of the leading demographic teachers and researchers as the discipline expanded through the USA, and indeed the world, in the second half of the century. The two-year-old bibliographical journal *Population Index* was moved to OPR and Notestein became one of the editors.

Notestein took leave from OPR during 1946–8 to become the first Consultant Director of the new United Nations Population Division, which acted as the secretariat for the United Nations Population Commission. Ryder (1984, p. 10) believes that the decision that the United Nations should have such a commission was probably influenced by OPR's wartime research on demographic change in Europe, commissioned by the League of Nations. Notestein left Princeton in 1959 to become President of the Population Council, finally to retire from the Council back to Princeton in 1968 to the position of Visiting Research Demographer at OPR. The Population Council was established as a result of a meeting in Williamsburg, Virginia in 1952. It was held on the initiative of John D. Rockefeller III, but Notestein prepared the agenda and appears to have arrived with a clear concept of the form such an institution should take. For some years the Council was the only organization which could provide expert advice to Third World governments on contraceptive technology and the establishment of family planning programs. During Notestein's presidency its annual budget rose from US$1 million to US$11 million and its professional staff from 10 to 90, with an increasing presence in developing countries. It also offered scholarships, predominantly to students from developing countries, to pursue graduate studies elsewhere in demography or reproductive biology.

To understand Notestein's achievements it is necessary to have some knowledge of the man himself. He was plain-speaking and straightforward, as was his prose. He had a great capacity for leadership, often in a fatherly way, perhaps because the Notesteins had no children. He had a gift for creating and building institutions, which was enhanced by his closeness to some of the rich and powerful, notably John D. Rockefeller III and Frederick Osborn. He moved easily between the foundation and the university worlds, partly because he believed that the aim of both should be to improve the human lot. He had great love for Princeton University but was no believer in the ivory tower. He was strangely modest, saying that others had greater intellectual contributions to make than himself, but that his strength lay in his gift for recognizing ability and providing it with an opportunity. His successors at the various institutions he developed were men he had brought into the demographic field: Ansley Coale at OPR, John Durand in the United Nations Population Division, and Bernard Berelson at the Population Council.

Notestein's world was the institution he was building and a major source of information the seminars and papers of that institution. When first presenting demographic transition theory in its more fully developed form, he described it as 'some of the work which my colleagues and I, of Princeton's Office of Population Research, have undertaken …' (Notestein 1943, p. 165). Prominent among the colleagues contributing to demographic theory were Dudley Kirk and Kingsley Davis. The most employed of Notestein's outside sources were two works which significantly appeared in his first year of teaching: *World Population* (Carr-Saunders 1936), which documented Third World mortality declines and population growth, and *Medical History of Contraception* (Himes 1936), which showed how ancient were attempts to restrict fertility. In his two most famous papers on demographic transition, 'Population—the long view' (1945) and 'Economic problems of population change' (1953), there are only a couple of references which go beyond these two sources and work that either he or his colleagues had undertaken at MMF, OPR or the UN. All his papers drew quite early on ongoing research projects at MMF or OPR, which themselves were not often fully reported until years later. Many of his papers were given first to conferences of mixed disciplines and achieve their impact from their simplicity, common sense, and clarity. After 1936 surprisingly few of his publications were papers submitted without solicitation to demographic journals. All, as Coale (1979) notes of his writing and teaching, were marked by skepticism, respect for evidence, and insistence on rigor and technical competence.

2. *Contribution to Demography*

The greatest support for demographic studies and for their funding came between the World Wars from those in the eugenics movement. Their major concerns were that the lower socioeconomic classes were outbreeding the higher ones (and, they assumed, the genetically brighter) and, in the USA, that immigrants had higher fertility than the native-born. The contribution of Sydenstricker and Notestein (1930) was to present, virtually without comment on the implications for society, technically sophisticated analyses of a sample of previously untabulated 1910 US census returns to show that immigrants had higher fertility than the native-born in the first generation but lower in the second. They also showed there were marked rural–urban fertility differentials and somewhat smaller occupational ones. They tried to free the

debate from ideology. After 1936 Notestein ceased publishing in the area of fertility differentials, although it was an important element in the 1941 Indianapolis Survey, which was recommended by a committee of which he was a member. That study was the world's first household fertility and family planning survey and was the direct ancestor of the global programs of such surveys in the later decades of the century.

During the 1930s Notestein, together with Regine Stix, analyzed information from interviews in 1932–4 of women who had attended a Margaret Sanger family planning clinic in 1931–2. Three conclusions were to be taught to all who passed through OPR. The first was that deliberate fertility control completely explained the fertility transition and that it did not have biological origins. The second was that 'New patterns of living and new values brought growing interest in family limitation that spread the use of known methods and stimulated the development of new ones. In a real sense, modern birth control is as much the result of new interest in family limitation as its cause' Stix and Notestein (1940, p. 150). The third conclusion, deriving from the discovery that most of these women (predominantly Jewish refugees from Europe) had earlier employed *coitus interruptus* (withdrawal) to limit fertility, was that all societies at all times could have reduced family size if the socioeconomic and cultural circumstances had made this desirable.

Notestein's name is most strongly linked with demographic transition theory, which is as important an intellectual framework in modern demography as evolution is in biology. In its simplest sense, namely, as societies evolve and become richer, mortality falls, to be followed later by a decline in fertility, the concept existed all through the twentieth century. Notestein taught demographic change within such a framework from the inception of his OPR course. His Cornell supervisor, Willcox, had published a paper drawing these conclusions in 1916, and Warren Thompson had divided the process into analytical stages in 1929. Notestein's and OPR's contributions were to identify the conditions of fertility decline, the role of contraception, the great spurt of population growth which occurred before the transition was over, and the spread of the phenomenon. This was greatly facilitated by the sponsorship of four major studies of demographic change in Europe at OPR by a division of the League of Nations transferred from Geneva to Princeton for the duration of World War II, and subsequent funding by the Carnegie Foundation through MMF. This work was published in the period 1944–6, but by February 1943 Notestein was reporting to a meeting of the American Philosophical Society the nature of the 'vital revolution,' the growth of population because of the 'lagging transition in growth rates,' the spread of this revolution to Southern and Eastern Europe, and the fact that mortality was already declining in agrarian Asia portending rapid population growth there.

The globalization of demographic theory was permitted by the funding in 1943 by the Geographer of the US Department of State of studies of population change in India (now India, Pakistan, and Bangladesh) and Japan. Later studies were to cover Formosa (now Taiwan), Malaya (now Malaysia), and Lebanon. Notestein's interest in Asia had been whetted by his part in funding and analyzing the population segment of John Lossing Buck's land utilization survey carried out in China in 1929–31. His conviction that the colonial world would face mounting population growth in the postwar period had been strengthened by the Annual Conference of the Milbank Memorial Fund in Princeton in early 1944 where OPR researchers gave papers on rapid population growth in India, Japan, and Egypt, and agreed that global demographic transition was underway. This led to his paper, 'Population – the long view' (1945), presented to a 1944 conference of food specialists at the University of Chicago, and a 1952 address to a conference of agricultural economists, 'Economic problems of population change' (1953). With these two papers the theory was complete. Notestein maintained that fertility declined first in the urban populations of industrialized societies because of the development there of conditions that eroded the traditional family and ascribed status, favoring instead individualism and self-advancement. The decline was retarded because of the slow erosion of ancient institutions and beliefs which had been required to encourage the high fertility necessary to offset the high mortality of premodern times.

But Notestein was already entering a new stage in his life. He and other OPR researchers had recorded that European fertility had begun to fall centuries after mortality did so, with the result that population of European extraction increased sevenfold. This seemed to be an impossible scenario for agrarian Asia and suggested that mortality decline there might meet a Malthusian barrier strangling both the mortality transition and hopes for rising living standards. Already in 'Population—the long view' (1945, p. 52) he had raised the possibility that Asian population growth might be held to doubling by 'wise and vigorous leadership' and 'studied efforts … to induce declining fertility.' By 1952, in 'Economic problems of population change' (1953, p. 28), his position was that 'It is within the bounds of possibility that the wise use of modern methods of communication and training to promote higher age at marriage and the practice of birth control would bring a considerable reduction of the birth rate even in peasant societies.' Birth control interventions were increasingly attempted, many with the assistance of the Population Council during Notestein's presidency. Fertility in most of the developing world has since declined, earlier than the OPR researchers of the 1940s thought likely, but neither Notestein nor others at OPR incorporated the new

vital revolution into a revised demographic transition theory. Perhaps Notestein's stature and influence rested most on the extraordinary extent to which most of his colleagues and students retained the consensus which had been reached in OPR during the 1940s and which he expressed so well.

3. Revisionists

The movement of Notestein and others toward being more involved in policy formulation and less the detached intellectual observer disquieted parts of the academic demography community. Dennis Hodgson (1983, pp. 2–3) saw this change as being unfortunate, from the scientific point of view, arguing that 'There is a tendency … for crisis research to be policy research rather than to be strictly academic in nature.' He contrasted this with an earlier, perhaps happier period, when 'Understanding, not change, was the goal.' Simon Szreter (1993) went much further and attributed the changes in Notestein's and his colleagues' interpretations and attitudes to demographic change to their appreciation of the USA's strategic interests during the Cold War period.

Hodgson's criticism probably would not have worried Notestein. He had always held that the overriding purpose of research was to help the human race, writing in his last published article that he was 'both a foundation and university man,' which he did not find incompatible, because most social scientists are 'reformers who are basically interested in the interplay of evolving technology, socioeconomic institutions, and population change as major determinants of the human condition' (Notestein 1982, pp. 651–87). He had never ruled out policy interventions but regarded them as not only desirable but even to be expected, not only on the mortality side of demographic transition but also on its fertility side. He regarded them as social experiments worth undertaking and studying. His concluding chapter to Stix and Notestein (1940, pp. 150ff) referred to the need to do something about the high fertility, poverty, and human suffering of half the world, and then discussed possible family planning interventions in distressed areas of the USA. His acceptance of assistance from the League of Nations, the Department of State, the World Bank, and others shows his determination to fund OPR adequately to investigate to its fullest the demographic transition as it had occurred in Western European populations and as it moved around the world. Indeed, if he had not done so, we would have heard little of him or of OPR. Even his later papers place less emphasis on saving the world than on giving priority to demographic research and its funding. The charge that Notestein gave priority to Asia because it was the prime area of American strategic interest, especially as China became communist (Szreter 1993, pp. 675–80), overlooks the fact that Notestein regarded Asia as the main continent facing potential Malthusian pressures throughout his academic life.

Frank Notestein died at 80 years of age, of complications arising from emphysema, on 18 February, 1983 in retirement at Newton, Pennsylvania, only a short drive from Princeton.

See also: Demographic Measurement: General Issues and Measures of Fertility; Demographic Models; Demography: Twentieth-century History; Eugenics as the Basis of Population Policy; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility Theory: Embodied-capital Theory of Life History Evolution; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Global Population Trends; Human Ecology: Insights on Demographic Behavior; Immigrant Fertility; Population Cycles and Demographic Behavior; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Population, Economic Development, and Poverty

Bibliography

Caldwell J, Caldwell P 1986 *Limiting Population Growth, and the Ford Foundation Contribution*. F Pinter, London

Carr-Saunders A M 1936 *World Population: Past Growth and Present Trends*. Clarendon Press, Oxford, UK

Coale A J 1979 Notestein, Frank W. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*: *Bibliographical Supplement*. Free Press, New York

Himes N E 1936 *Medical History of Contraception*. Williams & Wilkins, Baltimore, MD

Hodgson D 1983 Demography as a social science and policy science. *Population and Development Review* **9**: 1–34

Notestein F W 1938 A demographic study of 38,256 rural families in China. *Milbank Memorial Fund Quarterly* **16**: 57–79

Notestein F W 1943 Some implications of population change for post-war Europe. *Proceedings of the American Philosophical Society* **87**: 165–74

Notestein F W 1945 Population—The long view. In: Schultz T W (ed.) *Food for the World*. University of Chicago Press, Chicago, IL, pp. 36–57

Notestein F W 1953 Economic problems of population change. In: *Proceedings of the Eighth International Conference of Agricultural Economists*. Oxford University Press, London, pp. 13–31

Notestein F W 1982 Demography in the United States: A partial account of the development of the field. *Population and Development Review* **8**: 651–87

Ryder N B 1984 Frank Wallace Notestein (1902–1983). *Population Studies* **38**: 5–20

Stix R K, Notestein F W 1940 *Controlled Fertility: An Evaluation of Clinic Service*. Williams & Wilkins, Baltimore, MD

Sydenstricker E, Notestein F W 1930 Differential fertility according to social class. *Journal of the American Statistical Association* **25** (169): 9–32

Szreter S 1993 The idea of demographic transition and the study of fertility change: A critical intellectual history. *Population and Development Review* **19**: 659–701

Thompson W S 1929 Population. *American Journal of Sociology* **34**: 959–75
Willcox W 1916 The nature and significance of the changes in the birth and death rates in recent years. *Journal of the American Statistical Association* **15** (113): 1–15

J. C. Caldwell

Number Systems, Evolution of

1. The Term 'Number System'

A 'number system' consists of a set of mental entities and operations (see *Numeral Systems*) together with systems of symbols (see *Mental Models*, *Psychology of*) such as number words, tallies, or numerical sign systems and of symbolic activities such as counting, addition, or multiplication representing them. Number systems evolved in the course of the development of most civilizations. This evolution changed not only numbers and number representations but also the meaning of the term 'number' itself.

Symbolic activities such as counting, the use of tallies for representing quantities, or performing calculations by means of tokens or systems of written symbols, are usually considered practices dealing with 'numbers.' However, such symbolic activities historically precede an explicit number concept. Explicit concepts of number are the outgrowth of reflection upon such activities.

When first explicit concepts of abstract 'numbers' emerged in history, they were restricted to the number system which is called today 'natural numbers,' that is, to the system '1, 2, 3, 4, …' The subsequent history of the term number is characterized by a series of extensions of the number concept to cover 'integers,' 'rational numbers,' 'real numbers,' 'complex numbers,' 'algebraic numbers,' and other mathematical objects designated, with lesser or greater justification, as a number system. These extensions usually included numbers of earlier stages as special kinds of numbers, but necessarily also dismissed certain properties which seemed to be essential in earlier stages. Consequently, any extension of number systems was accompanied by controversial discussions about the question whether the newly-considered mathematical objects were really numbers. These doubts are often reflected in the designation of such numbers as 'irrational,' 'negative,' 'imaginary,' etc. Finally, the variety of potential candidates for the term number became so great that the answer to the question of what is and what is not a 'number' turned into a merely conventional decision. The focus of attention shifted from the investigation of types of numbers as they were historically determined towards the study of their common structures.

2. Cultures without Number Systems

Until recently, 'prearithmetical' cultures existed without any arithmetical techniques at all such as finger counting, counting notches, counting knots, or tokens. The best known examples are the Australian aborigines (Blake 1981, Dixon 1980) and certain South American natives (Closs 1986) whose languages did possess terms for quantities, yet of an exclusively qualitative nature, terms such as 'many,' 'high,' 'big,' or 'wide.' (see *Australian Aborigines: Sociocultural Aspects*; *Justice: Philosophical Aspects*; *South America: Sociocultural Aspects*). The quantitative aspects of an object of cognition were not yet distinguished from its specific physical appearance (see *Indigenous Knowledge and Technology*). Since from periods before the Late Neolithic no objects or signs have been identified that might have had some kind of arithmetical function, it is likely that up to that time human culture remained entirely prearithmetical. While artifacts with repeated signs and regular patterns from periods as early as the Late Paleolithic (see *Modernity: Anthropological Aspects*) have occasionally been interpreted as representations of numbers (Marshack 1972), such an interpretation seems to be untenable since these sign repetitions lack the characteristic subdivision by counting levels which is present in all real counting systems (Damerow 1998).

3. Proto-arithmetic

The occurrence of the simplest genuine arithmetical activities known from recent nonliterate cultures (see, e.g., Saxe 1982) date back to the Late Neolithic and the Early Bronze Age. These activities aiming at the identification and control of quantities are based on structured and standardized systems of symbols representing objects (see *North America and Native Americans: Sociocultural Aspects*; *Ritual and Symbolism, Archaeology of*). Their emergence as counting and tallying techniques may have been a consequence of sedentariness. Symbols are the most simple tools for the construction of one-to-one correspondences in counting and tallying that can be transmitted from generation to generation. The organization of agricultural cultivation, animal husbandry and household administration led to social conditions that apparently made symbolic techniques useful, and their systematic transmission and development possible. Such techniques are 'proto-arithmetical' insofar as the symbols represent objects and not 'numbers,' and consequently are not used for symbolic transformations which correspond to such arithmetical operations as addition and multiplication.

Early explorers and travelers encountering indigenous cultures using proto-arithmetical techniques often anachronistically interpreted their activities from a modern numerical perspective and believed the limita-

tions of proto-arithmetic resulted from deficient mental abilities of such peoples. It was only in the first half of the twentieth century that anthropologists and psychologists challenged these beliefs (e.g., see *Klein, Melanie* (1882–1960); Boas, Franz (1858–1942)) and began seriously to study culturally specific mental constructions connected with proto-arithmetical activities (Levy-Bruhl 1923, Wertheimer 1925).

4. *From Proto-arithmetic to Systems of Numerical Symbols*

A further step in the evolution of number systems resulted from the rise of early civilizations and the invention of writing. This transition is particularly well documented in the case of the Ancient Near East (see *Writing, Evolution of*). A system of clay tokens possessing proto-arithmetical functions has been identified which may have been widely used already during the period from the beginning of sedentariness in the areas surrounding the Mesopotamian lowland plain and in the Nile valley around 8000 BC until the emergence of cities around 4000 BC (Schmandt-Besserat 1992).

In the fourth millennium BC this proto-arithmetic system of clay tokens became the central tool for the control of a locally-centralized economic administration. In order to serve this purpose, the proto-arithmetical capabilities of the system were exploited to the limits of their capacity. The safekeeping of tokens in closed and sealed spherical clay envelopes demonstrates that they were used for the encoding of important information. The tokens were complemented and later completely replaced by markings that were impressed on the surface of such clay envelopes or on the surface of sealed clay tablets. Thus, the system was transformed into a more suitable medium by substituting written signs for tokens. Around 3200 BC, the numerical notations were supplemented with pictograms, and the tablets achieved a more complex structure. They now displayed several quantitative notations arranged according to their function, as we find in an administrative form. The earliest attestation of addition as an operation with symbols is provided by the notation of totals of the entries on such tablets, usually inscribed on the reverse of the tablet containing these entries.

Based on their origin, the oldest written numerical notation systems exhibited, for a short period, very unusual characteristics. Since they initially still represented units of counting and measurements, and since the numerical relations between these units were dependent on the counted or measured objects, the numerical signs had no uniquely determined values. Their values resulted from their metrological context, and changed with the respective areas of application without any apparent attempt to attain unambiguous numerical values for the signs. From the viewpoint of modern arithmetic, the signs thus represented in different contexts different numbers (Damerow 1995, Damerow and Englund 1987). Some 500 years later, however, the system had developed into a sort of numerical sign system as is known from other early civilizations. In particular, the signs had now attained fixed numerical values.

5. *Systems of Numerical Symbols in Early Civilizations*

Most, if not all, advanced civilizations, in particular the Egyptian empire, the Mesopotamian city-states, the Mediterranean cultures, the Chinese empire, the Central American cultures, and the Inca culture, have, independent from each other or by adaptation from other cultures, developed systems of numerical symbols that exhibit similar semiotic characteristics. The basic symbols of these systems were the same as at the proto-arithmetical level, signs for units and not for numbers, but since they represented now complex metrologies they had necessarily also to deal with fractions. Furthermore, part of the systems were now complex symbol transformations, that is, arithmetical techniques such as Egyptian calculation using unit fractions (Chace 1927), sexagesimal arithmetical techniques of the Babylonians (Neugebauer 1934), transformations of rod numerals on the Chinese counting board (Lǐ and Dù 1987), calendrical calculations in the pre-Columbian culture of the Maya (Closs 1986, Thompson 1960), or techniques of the use of knotted cords (*quipu*) as administrative tools by the Inca (Ascher and Ascher 1971–1972, Locke 1923).

The development of these numerical techniques was historically closely related with the administrative problems that had arisen through the concentration of economic goods and services in the governmental centers of early state organization (see *State: Anthropological Aspects*). The dramatic rise in the quantities of products had to be controlled, and the immense variety of decision-making implications had to be executed administratively (Høyrup 1994, Nissen et al. 1993).

The rules dictating the use of numerical signs reflected this function; they corresponded to their meaning in the social context. The numerical sign systems thus exhibited a great variety of structures with no internal differentiation between rules representing a universal number concept and rules depending on the specific kind of symbolic representation and its function to control quantities in a specific social setting. It was only the cultural exchange between advanced civilizations with developed systems of numerical notations that created the preconditions for the differentiation between universal numbers and specific notations.

6. *The Euclidean Conceptualization of Number*

The emergence of an abstract concept of number as a consequence of cultural exchange is particularly well known from the ancient Greek culture which played a crucial role in the historical process of transforming the heritage of early Near Eastern civilizations into the intellectual achievements of the hegemonic cultures of the Hellenistic and Roman world. The reflective restructuring of divergent bodies of knowledge brought about new kinds of general concepts and, in particular, an abstract concept of number (see *Reflexivity in Anthropology*).

The oldest known examples of general propositions about abstract numbers, for instance, the statement that the number of prime numbers is infinite, are handed down to us through the definitions and theorems of Euclid's Elements (Heath 1956). The relevance of these definitions and theorems was no longer based on their practical applicability, but only on their role within a closed system of mental operations. These operations reflected certain arithmetical activities in the medium of written language. In the case of the number concept these operations were the techniques of counting and tallying. Euclid defined a number as a 'multitude composed of units' and hence restricted the abstract concept of number to natural numbers. Platonism, in particular, dogmatically excluded from theoretical arithmetic all numerical structures that did not come under the Euclidian definition. Fractions of a number, for instance, and even more so irrational numbers such as the square roots of natural numbers that are not squares were not accepted themselves as numbers.

In order to circumvent the problems of incommensurability, a theory of proportions was developed, not for numbers, but for entities designated as 'magnitudes' (Heath 1956). This theory served at the same time as a substitute for an expanded concept of number that could cover all the arithmetical activities and symbolic operations existing at that time.

At a practical level, however, operations with notations for other types of number systems had been developed and were used long before they were theoretically reflected in an extended number concept. Any exchange of goods by merchants or by administrators of a centralized economy implicitly involves fractions. The balancing of debits and credits in bookkeeping similarly involves negative numbers. Accordingly, arithmetical techniques for dealing implicitly or even explicitly with fractions or with negative and irrational numbers had thus been developed already in early civilizations long before they were reflected in an explicit number concept.

7. *The Impact of Algebra*

The problems resulting from the restrictive Greek number concept were aggravated with the renaissance of ancient mathematics in the Early Modern Era. In particular, the development of algebraic notation techniques and of analytical methods applied to continuous processes in mechanics made the distinction between numbers and magnitudes obsolete. The use of variables in order to solve equations was indifferent to this conceptual distinction. Fractions, negative numbers, irrational numbers, infinitesimals, and infinity became more and more accepted as numbers. Although infinitesimals were again expelled in the nineteenth century after their missing logical foundation resulted in serious contradictions in the conceptual system of mathematics, they were finally reintroduced in a new form by the nonstandard analysis of the twentieth century.

The introduction of algebraic notations thus contributed to the transformation of existing arithmetical techniques into conceptualized number systems, but it also initiated the creation of completely new ones. A new type of number system with no immediately obvious meaning emerged, reflecting no longer primary arithmetical techniques but rather abstract operations with variables and equations. A well-known example is provided by 'imaginary' square roots of negative numbers which occurred as solutions to higher degree equations. Such solutions, although calculated correctly according to rules for numbers, seemed to make no sense since squaring a number never results in a negative number. Nevertheless, it turned out to be possible to operate consistently with such imaginary numbers as if indeed they were, in fact, numbers. They were, thus, finally accepted as 'complex numbers.'

The justification of seemingly meaningless number systems was not the only problem posed by the development of mathematics in the nineteenth century. The term number became questionable for other reasons too. On the one hand, a productive new theory, even today called simply 'theory of numbers,' was created; it is, however, a discipline with a specific topic. The term number is in this theory restricted to (positive and negative) integers. On the other hand, a growing number of new types of mathematical systems emerged which, with greater or lesser emphasis, were called number systems, for example, 'algebraic numbers,' 'transcendental numbers,' 'higher order complex numbers,' 'quaternions,' 'biquaternions,' etc. In order to organize the hodgepodge of artificially-constructed number systems, a heuristic principle was formulated. The construction of number systems should follow the 'principle of the permanence of formal laws' of an *arithmetica universalis* (Hankel 1867), that is, they should be constructed in such a way that some of the rules of ordinary numbers are preserved. It was hoped, in particular, that 'higher order complex numbers' of a certain type might serve as the foundation for an infinite number and virtually all possible number systems, an expectation which later turned out to be misguided.

8. Formalism and the Decline of Number

The formalistic approach, which in the nineteenth century led to the flourishing number systems and ascribed to them a key role in the foundation of mathematics, reduced their importance in the course of the structuralist rebuilding of mathematics in the twentieth century. In fact, the recombination of structures and operations of mathematical objects in order to create new ones with new properties made it more and more difficult to distinguish number systems from other systems of mathematical objects. The designation of an object as a number was increasingly seen as merely an historical convention. Number systems were downgraded to instances of structures abstracted from them, in particular, of algebraic structures resembling addition and multiplication and topological structures determining distance and continuity. Not numbers, but 'sets' equipped with such structures, which were designated artificially as 'groups,' 'fields,' 'rings,' 'lattices,' 'compact spaces,' etc., became the building blocks of modern mathematics.

See also: Brain, Evolution of; Civilization, Concept and History of; Cultural Evolution: Overview; Culture in Development; Evolution of Cognition: An Adaptationist Perspective; Human Cognition, Evolution of; Indigenous Knowledge: Science and Technology Studies; Knowledge, Anthropology of; Mathematical and Logical Abilities, Neural Basis of; Mathematical Education

Bibliography

Ascher M, Ascher R 1971–1972 Numbers and relations from ancient Andean Quipus. *Archive for History of Exact Sciences* **8**: 288–320

Blake B J 1981 *Australian Aboriginal Languages*. Angus and Robertson, London

Chace A B 1927 *The Rhind Mathematical Papyrus*. Mathematical Association of America, Oberlin, OH

Closs M P (ed.) 1986 *Native American Mathematics*. University of Texas Press, Austin, TX

Damerow P 1995 *Abstraction and Representation: Essays on the Cultural Evolution of Thinking*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Damerow P 1998 Prehistory and cognitive development. In: Langer J, Killen M (eds.) *Piaget, Evolution, and Development*. Erlbaum, Mahwah, NJ

Damerow P, Englund R K 1987 *Die Zahlzeichensysteme der Archaischen Texte aus Uruk*. In: Green M W, Nissen H J (eds.) *Zeichenliste der Archaischen Texte aus Uruk*. Gebr. Mann, Berlin

Dixon R M W 1980 *The Languages of Australia*. Cambridge University Press, Cambridge, UK

Hankel H 1867 *Theorie der complexen Zahlensysteme insbesondere der gemeinen imaginären Zahlen und der Hamilton'schen Quaternionen nebst ihrer geometrischen Darstellung*. Voss, Leipzig

Heath S T L (ed.) 1956 *The Thirteen Books of Euclid's Elements*. Dover, New York

Høyrup J 1994 *In Measure, Number, and Weight: Studies in Mathematics and Culture*. State University of New York Press, Albany, NY

Levy-Bruhl L 1923 *Primitive Mentality*. Macmillan, New York

Lǐ Y, Dù S 1987 *Chinese Mathematics: A Concise History*. Clarendon, Oxford

Locke L L 1923 *The Ancient Quipu or Peruvian Knot Record*. The American Museum of Natural History, New York

Marshack A 1972 *The Roots of Civilization: The Cognitive Beginnings of Man's First Art, Symbol and Notation*. Weidenfeld and Nicolson, London

Neugebauer O 1934 *Vorgriechische Mathematik*. Springer, Berlin

Nissen H J, Damerow P, Englund R K 1993 *Archaic Bookkeeping: Early Writing and Techniques of Economic Administration in the Ancient Near East*. Chicago University Press, Chicago, IL

Saxe G B 1982 Culture and the development of numerical cognition: Studies among the Oksapmin of Papua New Guinea. In: Brainerd C J (ed.) *The Development of Logical and Mathematical Cognition*. Springer, New York

Schmandt-Besserat D 1992 *Before Writing*. University of Texas Press, Austin, TX

Thompson J E S 1960 *Maya Hieroglyphic Writing*. University of Oklahoma Press, Norman, OK

Wertheimer M 1925 *Über das Denken der Naturvölker: Zahlen und Zahlgebilde*. In: Wertheimer, M (ed.) *Drei Abhandlungen zur Gestalttheorie*. Palm and Enke, Erlangen, Germany

P. Damerow

Numeral Systems

A numeral system is a part of a natural language, primarily devoted to the expression of positive whole numbers. A number is a mathematical abstraction; a numeral is a word or phrase expressing a number. The international scientific ('Arabic') notation provides a universally known, unambiguous means of representing numbers, which gives the study of numeral systems a clear advantage over other areas of language, in which the representation of meaning is problematic. The basic semantic operations relevant to the interpretation of complex numerals (such as *five thousand six hundred and seventy eight*) are addition and multiplication. Widespread understanding of these operations also gives the study of numeral systems an especially clear compositional semantics.

The spoken and written-out forms of numerals differ in various ways from the Arabic notation. There may be a difference in base number; thus French expresses 80 as *quatre-vingts*, literally 4 20s. There may also be a difference in the order of constituent elements; thus German expresses 64 as *vierundsechzig*, literally $4+60$. The scientific notation is compact,

having no explicit representation of addition and multiplication, and relying on the place-value convention to indicate powers of the base 10. But natural language numeral systems are not as compact, having explicit representations of the addition operation (e.g., *and* in *two hundred and sixty*), and explicitly naming the powers of the base number (e.g., in English with the suffix *-ty* and words such as *hundred* and *million*). As such vernacular forms precede the scientific notation (both in the history of a language and in the order of acquisition by individuals), it is the spoken and written-out forms of numerals, and their mapping onto numbers, that concern us here.

Some languages have no numerals:

> The one obvious gap in Australian vocabularies is the lack of any system of numbers. It is usually said that there are only numbers 'one,' 'two,' 'several' and 'many'; some languages appear also to have 'three,' although this is frequently a compound form. Kenneth Hale has suggested that these forms are not numerals in the strict sense, but are rather indefinite determiners, like English *a* and *some* (Dixon 1980, pp. 107–8).

In languages with a numeral system, forms for expressing positive whole numbers are in frequent use in everyday life, whereas forms for expressing more technical concepts, such as negative numbers and fractions, either do not exist or are restricted to technical discourse. Subsidiary systems, such as systems for naming dates, clock times, amounts of money, weights and measures, or telephone numbers, are derivative of the basic numeral system of a language, and are of less interest to the linguist, often being of recent invention, in some sense artificial, and following idiosyncratic rules characteristic only of themselves.

Numeral systems, though well integrated into their host languages, are nevertheless somewhat atypical of language subsystems as a whole. Numerals are the only expressions for which a conventional sequence exists. Thus there is an agreed answer to the question, 'What comes after sixteen?' But there is no such agreed sequence, in any language, for, say, the adjectives and adjectival phrases of the language. Obviously this sequencing is due to the especially clear semantic domain of numerals, the sequence of natural numbers. Numeral systems, beyond the most rudimentary, are also atypical in that they are usually explicitly taught. Children are drilled in reciting a conventional sequence of numerals, whereas almost every other part of the core grammar of a language is acquired without explicit teaching or drilling. This might be grounds for arguing that numerals themselves are to some extent artificial and not part of the core grammar of any natural language. But if we take the notion of what belongs to a language broadly, there can be no doubt that a language's numeral system is a proper part of it; a section on the numeral system of a language is incorporated into any decent reference grammar.

1. Numerals and Grammatical Number

Grammarians use the term '(grammatical) number' for the singular/plural distinction marked on nouns in most languages, and supplemented in some languages by categories such as dual, and even trial. Grammatical number tends to be indicated in the morphology of nouns (e.g., by an affix), whereas numerals are distinct words or phrases. Grammatical number has a loose relationship with numeral expressions in a language. One might expect that in any language the numeral expressing 1 would accompany a singular noun, the numeral expressing 2 would accompany a plural noun (or a dual noun if the language has duals), and any other numeral would be accompanied by a plural noun. But such a neat pattern is not universal. In Hungarian, for example, a noun following any numeral, for whatever number, is grammatically singular; the plural form of a noun is used only in the absence of any preceding numeral expression. In (Egyptian) Arabic, the situation is more complex, with a singular noun accompanying the form for 1, a dual noun going with the form for 2, a plural noun with forms for the numbers 3–10, and a singular noun with forms for 11 upwards. Examples are given below.

ENGLISH	HUNGARIAN	ARABIC
'minute'	*perc*	*daqiiqa*
'minutes'	*percek*	*daqaayiq*
'one minute'	*egy perc*	*daqiiqa waħda*
'two minutes'	*két perc*	*(itneen) daqiiqteen*
'three minutes'	*három perc*	*talat daqaayiq*
'ten minutes'	*tiz perc*	*ʕashar daqaayiq*
'eleven minutes'	*tizenegy perc*	*ħidaashar daqiiqa*

2. Simple Lexical Numerals

Cardinal numeral words are shaped and distributed like adjectives or nouns. Across languages, the lower-valued numeral words (typically up to about the range 2–4) tend to behave clearly like adjectives, and the higher-valued numeral words (especially those used as multiplicative bases, such as *hundred* and *million*) behave like nouns. Indeed, there is an example in the table above, in which the Arabic numeral for 1, *waaħid*, follows its noun, as adjectives do in Arabic, while the other numeral words (for 2, 3, 10, 11, 20) precede their noun, unlike adjectives. (See Corbett (1978a, 1978b) for details.)

In several other respects, the lower-valued numerals, up to about 2–4, behave differently from other numeral words. In some languages, there are different gender forms for just the first few numerals. Thus in Welsh there are masculine/feminine pairs *dau/dwy* (= 2), *tri/tair* (= 3), and *pedwar/pedair* (= 4), but only one form for other numerals. In some case-marking languages, numerals for values up to about 2–4 impose a special case (e.g., Genitive, as in Russian) on a following noun (See Hurford (2000) for examples).

Ordinal numerals (e.g., *first*, *second*, *third*,… *twentieth*,…), where they exist, are adjectives derived from cardinal numerals. The lower-valued ordinals (again typically up to about the range 2–4) tend to be formed suppletively, without any obvious morphological connection to the corresponding cardinal numeral. Thus *first* is not formed from *one*, whereas *fourth* is formed by suffixing *-th* to *four* (see Hurford (in press) for tables of such examples from a range of languages).

3. From Lexicon to Syntax

In some languages, the simple lexical numerals stop at 2, and higher values are expressed by syntactic combinations. An example is given here from a Queensland, Australia language cited in Tylor (1891, p. 243)

(a) *ganar*
(b) *burla*
(c) *burla-ganar*
(d) *burla-burla*

Other languages make the break into syntax after 3, 4, and 5, and most languages have simple numeral words for values up to 10, and then resort to syntactic combinations. After 10, almost without exception in any language, the only values for which there are morphologically simple words are those used as multiplicative bases, such as 20, 100, 1,000, and so on. (English *eleven* is an exception, but it is historically derived from a more complex form; Russian *sorok* (= 40) is a genuine exception. Another source of possible exceptions is the bodypart numeration practices found in some parts of Papua New Guinea. In such practices, parts of the body are touched in sequence as a way of recording or communicating numbers, typically starting with the left little finger, moving through the left digits, up the wrist, forearm, elbow, upper arm, shoulder, neck, head, and down to the right little finger. Thus, bodypart words can be said to have secondary numeral meanings (see Saxe (1981) for details).

4. Basic Syntax of Numerals

(The material in this section is set out in much greater detail in Hurford (1975).) With remarkable uniformity, the basic form of most syntactically complex numerals in most languages can be generated from a universal schema of just two simple phrase structure rules

$$\begin{aligned} &\text{NUM} \rightarrow \left\{ \begin{array}{c} \text{DIGIT} \\ \text{NUMPHRASE (NUM)} \end{array} \right\} \\ &\text{NUMPHRASE} \rightarrow \text{NUM} \quad \text{M} \end{aligned} \qquad (1)$$

Here, 'NUM' represents the category Numeral itself, the set of possible numeral expressions in a language; 'DIGIT' represents any single numeral word up to the value of the base number (e.g., English *one*, *two*,…, *nine*); and 'M' represents a category of mainly noun-like numeral forms used as multiplicational bases (e.g., English *-ty*, *thousand*, and *billion*). The curly brackets in the rules enclose alternatives; thus a numeral may be either a DIGIT (e.g. *eight*) or a so-called NUMPHRASE (numeral phrase) followed optionally by another numeral (e.g., *eight hundred* or *eight hundred and eight*). If a numeral has two immediate constituents (i.e., is not just a single word) the value of the whole is calculated by adding the values of the constituents; thus *sixty four* means $60 + 4$. If a numeral phrase (as distinct from a numeral) has two immediate constituents the value of the whole is calculated by multiplying the values of the constituents; thus *two hundred* means 2×100.

4.1 Language-specific Idiosyncrasies

Two sorts of reservation must be made about the claim for the universality of the two basic phrase structure rules presented above. One sort of reservation is a set of language-specific modifications, made on a case-by-case basis for each separate language, essentially dealing with the ways in which the particular language's irregularities deviate from the universal pattern embodied in the two rules. These idiosyncrasies typically affect the lower-valued expressions (up to about 20, but sometimes higher) only. For instance, the rules, if given an English lexicon, besides generating the regular numeral phrases *sixty*, *seventy*, *ninety*, would also generate **twoty*, which needs to be changed to *twenty*, and **onety*, which needs to be changed to *ten*. The rules above order a numeral phrase before any accompanying numeral, which could be a digit-word, thus generating, correctly for English, for example, *seventy two*, but incorrectly for German, for example, *fünfzig drei*. Languages which place summands in a lower-before-higher order require a rule (obligatory or optional, as the case may be) reordering the constituents of a numeral. Languages also typically augment the universal rule schema above with their own idiosyncratic rules inserting particles (e.g., *and*, *et*, *und*) denoting addition and, less often, multiplication.

It may be objected that the above forces the numerals of all languages into a Procrustean bed, distorting the true picture in individual languages. Such an objection turns out, on detailed study of individual languages, to be misplaced. Even in the numeral system of just one language, there is usually an indication of a regular underlying system, captured by the rule schema above, to which most of the numeral expressions in the language, including all of the higher-valued ones, conform. Using a handful of idiosyncratic statements to capture irregularities is in fact more economical than postulating a nonuniversal underlying scheme for any particular language.

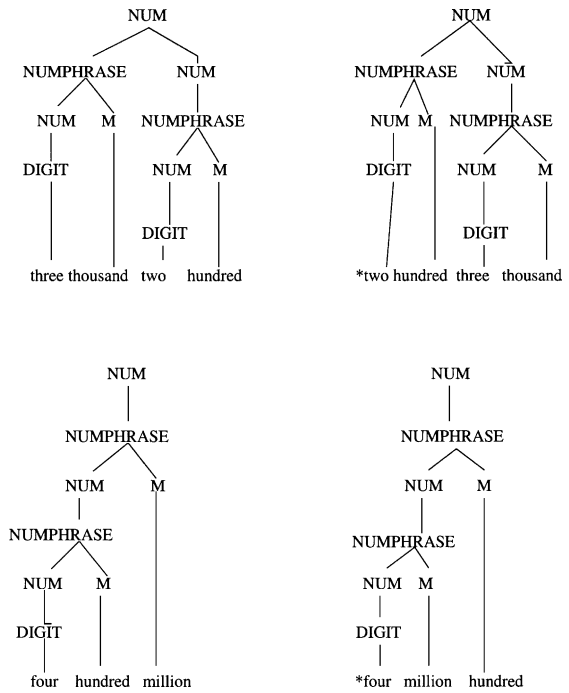

Figure 1
Structures of complex numerals

4.2 The Packing Strategy

The second necessary reservation to the universal rule schema above involves a way of constraining the massive overgeneration resulting from the recursive use of the category NUM in the rules. Without further constraint, these rules would generate (given an English lexicon) such non-existent forms as the following: **sixty ten*; **eighty hundred*; **two hundred*, *two hundred*, *two-hundred*; **three million thousand hundred*.

The universal rule schema generates hierarchically structured strings, which can be represented in tree diagrams such as in Fig. 1. These examples go in semantically equivalent pairs, as can be determined by adding the values of constituents of a NUM and multiplying the values of constituents of a NUMPHRASE. The first pair both mean 3,200, and the second pair both mean 400,000,000. But only one member of each pair, the left-hand structure, is well-formed. With strikingly few exceptions across all languages, ill-formed numeral structures generated by the universal rule schema can be eliminated by a universal principle known as the Packing Strategy, stated informally below.

Packing Strategy: within any part of a numeral structure, the sister of a NUM must have the highest possible value.

Another way of stating this is: in a numeral structure, if the sister of a NUM is not the highest valued well-formed structure of its category (equal to or lower in value than the value of the whole structure) generated by the universal schema together with the lexicon of the particular language, the whole structure is ill-formed.

Thus, **two hundred three thousand* above is characterized as ill-formed because of the existence of a well-formed NUMPHRASE *three thousand* which could be the sister of the next-to-top NUM ($3{,}000 < 3{,}200$), and which has a value greater than that of the actual NUMPHRASE sister of the next-to-top NUM ($3{,}000 > 200$). Similarly, **four million hundred* above is characterized as ill-formed because of the existence of a well-formed M *million* which could be the sister of the next-to-top NUM ($1{,}000{,}000 < 400{,}000{,}000$), and which has a value greater than that of the actual M sister of the next-to-top NUM ($1{,}000{,}000 > 100$).

The Packing Strategy works in coordination with the lexicons and other rules of particular languages. Beside the kinds of examples just given, it can be applied to explain some language-specific idiosyncrasies, such as the fact that in French 70 is expressed as *soixante-dix* ($= 60 + 10$), and not as $50 + 20$, or $40 + 30$, for example. The Packing Strategy can also be used to explain the actual sets of Ms (base words) found in numeral systems. In English, there are Ms with values of 10, 10^2 (*hundred*), 10^3 (*thousand*), $(10^3)^2$ (*million*), $(10^3)^3$ (*billion*), and $(10^3)^4$ (*trillion*). In Chinese there are Ms with values of 10, 10^2, 10^3, 10^4, $(10^4)^2$, and $(10^4)^3$. If one posits internal hierarchical structure for these lexical items, with exponentiation as the relevant arithmetical operator, the existence of such smoothly regular sequences of Ms can be predicted by means of the Packing Strategy.

The Packing Strategy is an unusually powerful generalization across languages. It is not without exceptions. The Packing Strategy predicts incorrectly, for instance, that English expressions such as *twenty three hundred* would be ill-formed, with *two thousand three hundred* being the only well-formed expression for 2,300. It is also falsified as an exceptionless generalization by the series of multiplicative bases found in Indian languages, exemplified here from Tamil: 10 *paththu*, 10^2 *nuru*, 10^3 *ayiram*, 10^5 *laksham*, 10^7 *kodi* (see Hurford (1975) for more fully argued details of the universal phrase structure rule schema for numerals and the Packing Strategy).

5. Linguistic Discontinuity

The mathematical number series, as defined by Peano, is smoothly continuous from zero to infinity, admitting of no discontinuities. The jump from 1 to 2 is of exactly the same nature as the jump from 2,001 to 2,002. But the sequences of numerals in natural languages show several discontinuities, and all at

around the same points. These discontinuity points are typically:

(a) around three, before which lexical numerals show various irregularities, for example, suppletive ordinals, agreement in gender, unusual case government, as mentioned above; psychologically, three is about the limit of subitization (instant number judgement without counting);

(b) after 10, at which point the first syntactically complex numerals typically appear, using addition;

(c) around 20, at the point when the first multiplicative expressions are used, which often coincides with a switch in the way addition is expressed (e.g., English DIGIT + -*teen* switches to DIGIT-*ty* DIGIT);

(d) (less often) at a point associated with some higher base number, such as 100 or 1,000, when different detailed rules may begin to apply (e.g., German switches from lower-plus-higher to higher-plus-lower order of summands)

Such linguistic discontinuity suggests a system growing by stages over time, with periods of stasis between each stage. The discontinuities are historical growth marks, like the rings in a tree trunk.

6. Some Idiosyncratic Systems

Linguistics is interested in discovering and explaining both the universal pattern to which all languages are cut and the range of variation within that pattern. Some linguists concentrate more on the universals, some more on variety and idiosyncrasy . The numeral systems of many languages are strikingly similar, but in this section a few examples will be given from unusual systems.

Not many numeral systems use subtraction, but the traditional Yoruba system uses it extensively. An example is *ẹ́rìn diń láàdọ́ọ̀ta*, glossed by Abraham (1958) as '4 as-to-a-reduction off 60–10.' *Áàdọ́ọ̀ta* is a contraction of a longer expression meaning literally '10 from (20 3).' Clearly the structure of this expression should reflect the bracketing [4[10[20 3]]]. In this structure 20 and 3 are combined by multiplication, yielding 60; after this 10 is subtracted from 60, giving 50; and finally 4 is subtracted from 50, giving 46. The last two operations performed are both subtraction. Note also that Yoruba here uses a highly unusual order for elements in a multiplicational relationship, in which the higher (base) number precedes the lower number. Almost universally in languages, the order low-before-high is preferred for numerals in a multiplicational relationship, as also for the combination numeral–noun; the combination noun-before-numeral is extremely rare.

The ancient Hawaiian system multiplied the usual bases by 4 before applying them, as shown by the numeral for 968 *elua lau me eha kanahaku-mamawalu*, literally 'two four-hundreds and four forties and eight' (Judd et al. 1945).

A nonstandard mathematical operation, called 'overcounting' is used extensively in Mayan languages and some others. An example is quoted in Hurford (1975)

> Thus the numeral 41, instead of being expressed by a word meaning 'twice twenty plus one,' is *huntuyoxkal*, viz. the first (*huntu*) of the third score; *oxkal* being three times twenty, or sixty.

In a few languages, there are somewhat different numeral systems, depending on what is being counted. In Kusaiean, a Micronesian language, one system is used for counting fish and things related to fishing; another system is used for counting other things.

7. Psycholinguistic and Philosophical Aspects

There is not space here to explore some interesting psychological and philosophical questions related to number and numerals. These include: is human knowledge of number handled by domain-specific parts of the brain? Do children learn the meanings of numerals from concrete examples, such as *three bricks*, *four teddy bears*? To what extent are the universal properties of numeral systems attributable to innate dispositions in the minds of language acquirers, or are they (complementarily) conventional historic-cultural constructs? Logically, are numerals proper names or predicates? In either case, what do they denote, platonic entities or sets of collections of objects? Discussion of these issues, and more, can be found in Dehaene (1997) and Hurford (1987).

See also: Number Systems, Evolution of

Bibliography

Abraham R C 1958 *Dictionary of Modern Yoruba*. University of London Press, London

Corbett G 1978a Universals in the syntax of cardinal numerals. *Lingua* **46**: 355–68

Corbett G 1978b Numerous squishes and squishy numerals in Slavonic. *International Review of Slavic Linguistics* **3**: 43–73

Dehaene S 1997 *The Number Sense*. Oxford University Press, New York

Dixon R M W 1980 *The Languages of Australia*. Cambridge University Press, Cambridge

Hurford J R 1975 *The Linguistic Theory of Numerals*. Cambridge University Press, Cambridge, UK

Hurford J R 1987 *Language and Number: The Emergence of a Cognitive System*. Blackwell, Oxford, UK

Hurford J R 2001 The interaction between numerals and nouns. In: Plank F (ed.) *Noun Phrase Structure in the Languages of Europe*. de Gruyter, Berlin

Judd H P, Pukui M K, Stokes J F G 1945 *Introduction to the Hawaiian Language*. Tongg Publishing Company, Honolulu, HI

Saxe G B 1981 Bodyparts as numerals: A developmental analysis of numeration among remote Oksapmin village populations in Papua New Guinea. *Child Development* **52**: 306–16

Tylor E B 1891 *Primitive Culture*. John Murray, London, Vol. 1

J. R. Hurford

Nursing Homes

Nursing homes are institutions that provide an age-friendly environment and care for elderly people suffering from physical, psychological, or social deficits. There are semidependent living arrangements for those capable of self-care but with some chronic limitations; these institutions offer meals, housekeeping, and transportation services. In addition, there are dependent living arrangements which are hospital-like environments that provide 24-hour supervision, intermittent services, and nursing care for severely impaired individuals. In the USA and other Western countries at similar levels of economic development, 5 percent of the total population over age 65 are institutionalized, and about 20 percent have to face institutionalization at some point in life. The probability of institutionalization increases with age. At age 85 and above, 25 percent of the population live in long-term care institutions (Kendig and Pynoos 1996). Demographic changes in the Western world will increase the number of elderly individuals—especially of very old persons—in societies. Some demographic projections report that by the year 2050, 30 percent of the elderly population will be aged 80 and above. In addition, declining birthrates and increases in women employment rates are likely to result in a decrease in potential family caregivers. These demographic developments suggest that nursing homes will remain an important living arrangement, even though social policy since the early 1990s has been focusing on community-based living and support systems (see *Age, Sociology of*).

1. *Consequences of Institutionalization*

The majority of people dread the idea of institutionalization because private homes and apartments within specific communities are strongly linked to biographies. Individual living arrangements, i.e., personal furniture, belongings, decorations, etc., can be regarded as an extension of the self and are not very easily abandoned. The move to an institution is sometimes considered as a first step towards death. Risk factors for institutionalization are high age, being female, widowhood, and childlessness. The main risk factor is impairment of functional status due to chronic, mental and/or physical diseases. Up to 50 percent of the inhabitants of nursing homes suffer from dementia. Nevertheless, people of similar vulnerability also live in communities. Therefore, the interaction between individual deficits and inadequate social support seems to be of crucial importance for the necessity of institutionalization.

Relocation is regarded as a critical life event, a major stressor which can lead to behavioral, psychological, and physical deterioration (Horowitz and Schulz 1983). Some studies, though, question this relocation-stress hypothesis and report positive outcomes of institutional living. These include better family relations, improved morale and positive alterations in lifestyles of the residents. However, there are severe methodological problems in detecting the effects of relocation. The effects of multimorbidity and various impairments are difficult to separate from effects due to institutionalization. Furthermore, it is almost impossible to realize control group designs (Kahana 1987). One important condition for positive effects seems to be that relocation is the result of the elderly person's own decision. If the selection of the environment is under the control of the elderly individual and if the institution provides compensatory support suitable for this individual, positive outcomes are most likely (M. M. Baltes et al. 1991). Unfortunately, however, many elderly people are relocated into institutions from acute hospitals or other institutional settings, which means that they have no control over the time of relocation or the choice of institution.

Personal coping strategies are an important mediator in dealing with the stress of relocation, adaptation to new living conditions, and maintenance of a sense of control. Since Goffman's (1961) classical study about the humiliation and deprivation of choice in 'total institutions', nursing homes are widely regarded as one example of violation of the basic needs for personal control. The negative impact of loss of control for elderly people has been proven in many studies (Heckhausen and Schulz 1995). It is important to note, however, that institutionalization does not necessarily lead to a total loss of control, rather to a change of certain aspects of control or differential means of expression of control. Contrary to common belief, not only active but also passive behavior can serve as having control. A resident exhibiting dependent behavior, for example, might execute control in order to secure social contact with staff members (M. M. Baltes 1996).

2. *Ecological Aspects*

Nursing homes are complex systems in which different ecological aspects interact and influence the residents' well-being and behavior. Nursing home residents who suffer from physical and cognitive restrictions are particularly vulnerable with regard to these physical and social environmental conditions. Although physical and social aspects are not independent from one another, they are described here separately for reasons of clarity.

2.1 Physical Environment

Housing and nursing homes are a major issue of theoretical and empirical research efforts in ecological and social gerontology. The importance of the physical environmental appearance in nursing homes was expressed by the classical environmental docility hypothesis, which stresses the increasing importance of environmental factors when specific losses in daily functioning occur (Lawton and Simon 1968). Deficits such as visual, mobility, or cognitive impairments are extremely stressful for elderly people, and architects have designed prothetic environments in order to compensate for these problems. The competence-press model stresses the idea that environmental demands that have a manageable difficulty level can optimize residents' performances (Lawton 1982). According to this hypothesis, high demands and low competence will result in stress and deterioration. Just manageable environmental press, however, might activate the individual's reserve capacities.

Other lines of reasoning in ecological gerontology stresses the importance of congruence between personal preferences and needs on the one hand, and physical and social characteristics of the environment on the other hand (Kahana 1982). An optimal person–environment fit with regard to necessary assistance and respect for privacy and autonomy is extremely important for the subjective well-being of residents. Due to interindividual differences in residents, the realization of a person–environment fit for every resident is an ambitious challenge for the nursing home providers.

Regnier and Pynoos (1992) suggest a conceptual framework with 12 environmental and behavioral principles in order to structure the organization of nursing homes: (a) provision of opportunities for privacy; (b) provision of opportunities for social interaction; (c) promotion of opportunities for control, choice, and autonomy; (d) fostering a sense of orientation and wayfinding; (e) provision of a safe and secure environment; (f) accessibility and functioning of basic requirements; (g) provision of stimulation and challenge; (h) consideration of sensory aspects; (i) enhancing familiarity through personal objects; (j) consideration of esthetics and appearance; (k) provision of opportunities for personalization of the environment; (l) adaptability of environmental conditions. Along these guidelines, nursing homes should be organized according to the needs of their specific residents.

2.2 Social Environment

Important predictors for subjective well-being are social relations and social activities. Therefore, an integration into the social world of a nursing home seems an important factor for positive adaptation. Any institutionalized resident is offered opportunities for social interaction with other residents and with staff members.

2.2.1 Residents as social partners. Nursing homes offer a variety of possibilities for establishing social contacts which, however, are often either not taken up or do not automatically lead to increased well-being of the residents. Research findings report that encouragement for social activities is difficult because individual preferences in activities differ. Furthermore, admission age of residents and the prevalence of demented elderly individuals are high in institutions, which increases the difficulties of setting up new contacts and satisfactory social exchange.

Socioemotional selectivity theory provides another explanation for the difficulties of enhancing social activities. Carstensen et al. (1997) argue that for elderly people, emotional criteria are crucial for the selection of social relationships. Therefore, the quality and not merely the quantity of social contacts is the decisive criterion for positive effects of relationships. Research findings demonstrating the importance of a partner or a close person as an important predictor for well-being support this view. These findings are relevant for nursing home residents because two main predictors for institutionalization are widowhood and childlessness. In other words, the important condition of having an intimate confidant does not exist for many residents, and likewise, motivation for activities with other residents may be low because these social contacts cannot substitute the missed intimate relationship. It remains an open question whether institutions could encourage emotional satisfying relationships between residents. So far, empirical results show that outer institutional contacts contribute considerably to the life satisfaction of residents.

2.2.2 Staff members as social partners. Nurses and other staff members are highly attractive and valued social partners for the residents because they represent the world outside the institution and have the means to execute emotional and instrumental power. Based on social learning theory, Margret Baltes conducted a research program which showed that cognitions and behavior patterns of staff have tremendous influence on autonomy and dependency of residents (M. M. Baltes 1996). Research findings have shown that the social reality in nursing homes can be described as overprotection and encouragement of dependent behavior. Elderly residents are regarded as incompetent and helpless by staff, who, according to the medical model, feel obliged to offer maximum assistance. Elderly residents and staff show a predictable pattern of behavior interaction: they seem to act according to a social script. This means that both staff and residents share common cog-

nitions about appropriate role behaviors. In particular, research results from cross-cultural studies have demonstrated—on a behavior level—that dependent behavior by residents is followed by staff support behavior for dependency. In addition, independent behavior by residents tends to be ignored or even punished by staff. Consequently, the display of dependent behavior by residents ensures the desired social contact with staff members. Unfortunately, this mode of control for social contact may have negative effects on the independent functioning of the residents because abilities might get lost with time without training.

3. *Intervention*

Despite a substantial amount of basic gerontological knowledge, transfer of this knowledge into practice is difficult. Therefore, a considerable number of intervention studies have been conducted in nursing homes in order to assess the outcomes. The major aim of interventions is to increase the well-being of residents. In order to enhance the quality of life of residents, different means and levels of intervention have been taken up.

On the individual level, behavior modification programs are often applied in order to change residents' behavior (Carstensen et al. 1996). First, residents' problems like sleeping disorders or urinary incontinence are the focus of intervention. Second, rehabilitation of lost functions concerning activities of daily living like eating or dressing are addressed. Third, behaviors that disturb other residents or staff like wandering or shouting might be treated with behavior modification. Especially in this latter case, ethical principles and legitimate interests of all parties have to be carefully considered before intervening.

On the group level, interventions concerning residents and interventions focusing on staff must be differentiated. The enhancement of social activities among the elderly—such as organizing games or introducing pets into the institution—are often the target of efforts. Other changes aim at increasing the sense of control, for example, through participation in decisions. Finally, cognitive functions are a major area of interventions. Reality orientation and memory trainings are group activities offered in most nursing homes. The effects of these interventions, however, remain unclear. Some authors claim that the social activity in itself is an important factor for the well-being of the residents rather than effects on cognitive functioning.

Few interventions aim at changing staff behavior in order to improve a positive impact on the residents' well-being. Maintenance or rehabilitation of everyday competence through adequate treatment by staff has been of major interest. Training programs for staff often include communicating the knowledge about normal aging, the importance of environmental conditions, and behavior modification (M. M. Baltes 1996).

On the system level, desired changes include the whole institution. Improvement of the physical environment is one key element for this kind of intervention. In particular, residents suffering from dementia need a therapeutic architecture that compensates for their deficits as much as possible (Regnier and Pynoos 1992). In order to improve orientation, different colors and noticeable symbols can be used on different floors. Low demanding elevators, simple sanitary instruments, and safe areas for physical activities like wandering are examples of matching special needs of demented patients. Moos and Lemke (1994, 1996) developed a Multiphasic Environmental Assessment Procedure (MEAP), an instrument that gathers descriptive data for different environmental dimensions. It measures physical and architectural features, institutional policy, residents' functioning, and staff characteristics. In addition, resident and staff perceptions of the social environment and social climate of an institution are measured. This instrument has been used in many studies in order to describe nursing homes or to examine interventions (Moos and Lemke 1996).

4. *Conclusions*

The theoretical and empirical research described in this article show a variety of encouraging results in that institutions are able to offer satisfying environments for their residents. In addition, many interventions have been successfully applied. One major problem of this research, however, is the fact that most studies were conducted in the 1970s or 1980s. Since 1980, the population of nursing home residents has changed noticeably. First, admission age has raised considerably. Most residents nowadays are age 85 or above when they move to an institution. Second, the proportion of residents suffering from dementia has increased tremendously—up to 50 percent in many institutions. Third, community-based services and new living arrangements are expanding; therefore healthier people are less likely to choose an institution, and the number of residents with chronic mental illnesses in nursing homes will increase further. These changes in the characteristics of institutionalized elderly people mean a major challenge for the providers and for future gerontological research.

Baltes and Baltes (1990) proposed a general model of adaptivity to the increasing biological and social losses in old age, the model of selective optimization with compensation. The three components—selection, optimization, and compensation—are regarded as fundamental processes or strategies to ensure adaptive development (P. B. Baltes 1997). The process of selection refers to a reduction or transformation of

domains of functioning in which the individual strives to perform efficiently. This involves a realistic evaluation of individual priorities, environmental demands, and biological capacity. The second component, optimization, reflects the traditional conception of development as an enrichment and augmentation of functioning with regard to quality and quantity. The third component, compensation, becomes important when losses occur (e.g., hearing loss) or behavioral capacities are reduced due to physical or mental constraints. The three processes are also operative in institutions. Nursing homes can be regarded as special worlds that alleviate adaptation to biological decline. The process of selection is realized by the provision of a less demanding environment. Compensation is offered by various means of assistance (e.g., care, prepared food) that reduce the effects of diminishing capacities, and optimization is supported by opportunities to pursue activities in those domains which are important to the individual resident (M. M. Baltes et al. 1991).

The model of selective optimization with compensation offers a valuable theoretical framework for future ecological and intervention research in institutions. It is an open question in how far the three components of positive adaptation can be supported with regard to residents suffering from dementia. Gerontological research has to generate answers concerning the balance between protection and compensation on the one hand, and stimulation and optimization on the other hand. The development of suitable offers for individuals with different levels of impairments and mental restrictions within one institution is of great importance for the well-being of the residents.

See also: Age, Sociology of; Aging and Health in Old Age

Bibliography

Baltes M M 1996 *The Many Faces of Dependency in Old Age*. Cambridge University Press, New York

Baltes M M, Wahl H-W, Reichert M 1991 Institutions and successful aging for the elderly? *Annual Review of Gerontology and Geriatrics* **11**: 311–37

Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80

Baltes P B, Baltes M M (eds.) 1990 *Successful Aging: Perspectives from the Behavioral Sciences*. Cambridge University Press, New York

Carstensen L L, Edelstein B A, Dornbrand L (eds.) 1996 *The Practical Handbook of Clinical Gerontology*. Sage Publications, Thousand Oaks, CA

Carstensen L L, Gross J, Fung H 1997 The social context of emotion. *Annual Review of Gerontology and Geriatrics* **17**: 325–52

Goffman E 1961 *Asylums: Essays on the Social Situation of Mental Patients and Other Inmates*. Anchor Books, Garden City, NY

Heckhausen J, Schulz R 1995 A life-span theory of control. *Psychological Review* **102**: 284–304

Horowitz M J, Schulz R 1983 The relocation controversy: Criticism and commentary on five recent studies. *Gerontologist* **23**: 229–34

Kahana E 1982 A congruence model of person–environment interaction. In: Lawton M P, Windley P O, Byerts T O (eds.) *Aging and the Environment: Theoretical Approaches*. Springer, New York, pp. 97–121

Kahana E 1987 Institutionalization. In: Maddox G L (ed.) *The Encyclopedia of Aging*. Springer Pub. Co., New York, pp. 355–7

Kendig H, Pynoos J 1996 Housing. In: Birren J E (ed.) *Encyclopedia of Gerontology: Age, Aging, and the Aged*. Academic Press, San Diego, CA, pp. 703–13

Lawton M P 1982 Competence, environmental press, and the adaptation of older people. In: Lawton M P, Windley P G, Byerts T O (eds.) *Aging and the Environment: Theoretical Approaches*. Springer, New York, pp. 33–59

Lawton M P, Simon B B 1968 The ecology of social relationships in housing for the elderly. *Gerontologist* **8**: 108–15

Moos R H, Lemke S 1994 *Group Residences for Older Adults: Physical Features, Policies, and Social Climate*. Oxford University Press, New York

Moos R H, Lemke S 1996 *Evaluating Residential Facilities: The Multiphasic Environmental Assessment Procedure*. Sage Publications, Thousand Oaks, CA

Regnier V, Pynoos J 1992 Environmental intervention for cognitively impaired older persons. In: Birren J E, Sloane R B, Cohen G D (eds.) *Handbook of Mental Health and Aging*, 2nd edn. Academic Press, San Diego, CA, pp. 763–92

S. Zank

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

O

Obedience: Social Psychological Perspectives

Obedience refers to the act of carrying out requests or commands emanating from a person of higher status within a social hierarchy. It is to be distinguished from compliance—which also refers to a form of acquiescence, but does not necessarily involve a situation of status inequality. Obedience is also different from conformity (see *Conformity: Sociological Aspects*). In the latter, the source of influence comes from others of equal status to the target of influence; such as, a group of peers or other members of one's society or culture (see *Social Influence, Psychology of* and *Influence: Social*). Additionally, while obedience—as well as compliance—is a result of direct influence attempts, conformity is usually the product of implicit, indirect group pressures felt by the individual to be impinging on him or her. Thus, while obedience represents the maintenance or creation of social inequality, conformity is a social leveler.

1. Distinctive Feature of the Social-psychological Approach

Social-psychological perspectives on obedience to authority see it as a necessary feature of organized social life, contributing importantly to its efficient and orderly functioning. Their most distinctive feature is the primary role accorded to situational factors. Grounded in Kurt Lewin's ahistorical approach, which emphasized contemporaneous determinants of behavior, social psychologists tend to focus on situational determinants to account for variations in obedience. This is in contrast to the personality psychologist who focuses on inner, stable dispositions to account for obedience and its variations. A prime example of the latter is the personality construct of authoritarianism, which consists of a number of interrelated elements, including a powerful propensity to obey authority (Adorno et al. 1950, Altemeyer 1988).

2. The Contribution of Stanley Milgram

The most important social-psychological contribution to understanding obedience was that of Milgram (1963, 1974). Although his aim was to study destructive obedience, his findings and insights are readily applicable to more mundane and benign forms of obedience. Milgram conducted a series of laboratory experiments at Yale University from July 1961 to May 1962 to answer the following question: If one person orders another to harm a third person, under what conditions would the second obey the first person's orders and under what conditions would he or she defy them? In his experiments, subjects who had been recruited for a study of memory and learning were instructed to give what appeared to be increasingly dangerous shocks to a learner every time he made a mistake on a word-matching task. The subjects delivered the shocks by pressing one of 30 switches arranged horizontally along the front panel of a box-shaped 'shock-generator.' Each subsequent switch was labeled with higher and higher voltages of shock, in 15-volt steps, beginning with 15 volts for the first switch and ending with 450 volts for the last one. The rule was that the subject teacher had to increase the shock level one switch at a time each time the learner made an error. In some of the conditions, as shock levels became more intense and painful, the learner responded with shrieks and complaints of increasing urgency and demanded that the experimenter terminate the experiment. After a certain point, the learner became silent, suggesting the possibility that he may have passed out.

The main finding was an unexpectedly high degree of obedience. As many as 65 percent of the subjects, ordinary residents of New Haven and Bridgeport, Connecticut, were fully obedient, that is, continued administering what they were led to believe were increasingly painful and dangerous shocks until the 450-volt maximum in response to the experimental authority's orders. In actuality, the shocks were not real and the learner was a confederate who made errors according to a pre-arranged schedule. Some of Milgram's additional findings were: Obedience declined as the distance between teacher and learner was reduced and when the distance between the experimenter and subject was increased; obedience increased when the subject was only indirectly involved in punishing the learner, that is, he read the wordlist, but someone else pressed the switch; the subject's obedience rate diminished drastically after seeing two of his peers refuse to continue and thereby providing models of disobedience; and men and women were equally obedient, although the latter reported experiencing greater tension. Milgram believed that such extreme obedience to potentially destructive com-

mands is made possible by the acceptance of the authority's definition of the situation and by the subject's entry into an 'agentic state,' a mental state whose central feature is the relinquishing of responsibility to the person in charge. According to Milgram, his experimenter represented a legitimate authority, a person who is seen as having the right to issue orders and as someone whom we feel an obligation to obey.

If the legitimacy of an authority is seen to be of prime importance in enabling him or her to be a powerful source of influence, a logical question is: How do authorities create and maintain their legitimacy? Recent research by Tyler (1997) has addressed this question. He distinguishes between two approaches that suggest why people see authorities as legitimate, and, hence, feel an obligation to obey them: an instrumental model and a relational model. According to the first approach, a prime determinant of peoples' acceptance of the legitimacy of a leader is whether they are seen as a source of benefits or desirable outcomes. According to the second approach, it is the nature of the treatment people receive from authorities—in particular the integrity and caring that mark their actions—that will determine their perceived legitimacy. Tyler finds that, although both instrumental and relational factors have an influence, the latter is the more dominant factor in judgments of an authority's legitimacy.

3. *Other Social-psychological Approaches*

According to French and Raven (1959) and Raven (1999), a person's authority or power to influence others can derive from other factors besides from his or her legitimacy. Specifically, according to these authors, social power can derive from six different bases. Besides legitimate power, they are: reward, coercive, expert, referent, and informational power. When an authority or persuader has reward power, people obey or yield because they see the former as a potential source of rewards. Coercive power is operating when people yield out of fear of punishment. Persons are said to possess expert power when others go along with their wishes because of their expertise or special knowledge. An authority or influencer is said to have referent power when people acquiesce to his or her commands or wishes because they like or identify with him or her. And finally, an authority or other source is wielding informational power when people obey or are influenced because they find the information provided by that person to be intrinsically convincing or compelling.

According to Kelman (1973), there are three social processes that help weaken the normal human reticence against inflicting harm on another and thereby facilitate people's obedience to destructive orders: authorization, routinization, and dehumanization. Through authorization, individuals are relieved of the necessity of assessing the morality of the actions they are directed to carry out. When an action is authorized, the usual moral considerations fade into the background, and the individual no longer feels responsible for his or her actions, nor any guilt, should those actions be harmful. Routinization—the absorption of the individual in the mechanical and routine details that are necessary for carrying out the authorized actions—minimizes the possibility that he will worry about its consequences. Dehumanization—viewing the target of one's harmful actions as less than human—further insures that the morality of the commands issued by an authority will not be questioned.

Going beyond obedience in authority–subordinate relationships between individuals, Kelman and Hamilton (1989) analyzed the effects of political authority—the nation–state and its various institutions—on the individual citizen. A central feature of their analysis is a threefold distinction among people's broad orientations to political authority—rule, role, and value orientations—that they see as potential mediators of the nature of citizens' responses to morally questionable policies or destructive orders, as the following quote indicates:

> For the *rule-oriented*, moral principles are largely irrelevant both to the actions of the citizen in an authority context and to the actions of the state in pursuit of its primary purposes: ensuring public order and national security. This view has a cynical overtone. The state cannot afford the luxury of concerning itself with higher principles, and individuals can be expected to do only what they must in order to stay out of trouble. For the *role-oriented*, special moral principles apply to the citizen in relation to the state as well as to the state itself. The morality of obedience—the obligation to obey and support the government—at the level of the individual citizen and the duty to pursue the national interest at the level of the state both override standard moral considerations. Finally, for the *value oriented*, principles of personal morality continue to govern the behavior of individuals as citizens, and adherence to moral principles is fundamental to the legitimacy of state action. Actions by and on behalf of the state may not always conform precisely to the expectations that govern personal relations, but they have to be chosen and evaluated with reference to standard moral principles (Kelman and Hamilton 1989, italics added).

4. *A Contemporary Problem: Harmful Obedience in the Workplace*

An important implication that Milgram (1974) had drawn from his research was about the potential dangers inherent in organizational behavior: When an individual 'merges … into an organizational structure, a new creature replaces autonomous man, unhindered by the limitations of individual morality, freed of humane inhibition, mindful only of the sanctions of authority.' This observation about the evaporation of

individual responsibility in hierarchical organizations anticipated the kinds of tragedies that later resulted from car manufacturers' employees' readiness to carry out their bosses' directives to produce unsafe automobiles, and the tobacco companies' employees' acquiescence to policies aimed at deceiving the public.

A current emphasis in research on obedience to authority is the systematic study of this 'darker side' to behavior within hierarchical organizations such as businesses and companies. Thus, for example, Hamilton and Sanders (1995), in a cross-cultural study, examined the factors that would affect judgments about a subordinate's responsibility when his company's actions could result in physical harm to the public. Brief et al. (2000) showed that, when a company president's orders not to hire minority group members was accompanied by a business justification —'Our organization attempts to match the characteristics of our representatives with the characteristics of the population to which they will be assigned'— subordinates were more likely to discriminate against Black job candidates than when no such justification was provided.

5. *The Pervasiveness of Obedience to Authority*

Obedience to authority would seem to be one of the universals of social behavior, transcending both time and place, as suggested by two quantitative analyses. One analysis, using the findings of Milgram's standard conditions and those of all the replications conducted by others, correlated the amount of obedience found in each study with its year of publication and found absolutely no relationship between the two variables (Blass 2000). Another analysis (Blass 1998), also involving Milgram's standard conditions and their replications, reveals that the average obedience rate found in US studies (61 percent) was very similar to the average rate of obedience obtained in replications carried out in other countries (66 percent). As Milgram once wrote:

> We do not observe compliance to authority merely because it is a transient cultural or historical phenomenon, but because it flows from the logical necessities of social organization. If we are to have social life in any organized form—that is to say if we are to have society—then we must have members of society amenable to organizational imperatives (letter to Professor Alan Elms, Sept 25, 1973; The Stanley Milgram Papers, Yale University Archives).

6. *Future Directions*

Recently, there has been a discernible shift in emphasis away from obedience toward defiance—both as a focus of research (e.g., Modigliani and Rochat 1995, Rochat and Modigliani 2000) and in the implementation of changes in the workplace that facilitate employees' disobedience of illegitimate or destructive orders from their superiors, such as the creation of hot lines and the placement of ombudsmen. In the future, one can expect a substantive increase in these kinds of developments.

See also: Authoritarian Personality: History of the Concept; Authoritarianism; Authority, Social Theories of; Community Power Structure; Compliance and Obedience: Legal; Conformity: Sociological Aspects; Influence: Social; Organizations: Authority and Power; Power in Society; Social Influence, Psychology of; Status and Role, Social Psychology of

Bibliography

Adorno T W, Frenkel-Brunswik E, Levinson D J, Sanford R N 1950 *The Authoritarian Personality*. Harper, New York

Altemeyer B 1988 *Enemies of Freedom: Understanding Right-wing Authoritarianism*. Jossey-Bass Publishers, San Francisco

Blass T 1998 A cross-cultural comparison of studies of obedience using the Milgram paradigm. Paper presented at the annual convention of the American Psychological Association, San Francisco

Blass T 2000 The Milgram paradigm after 35 years: Some things we now know about obedience to authority. In: Blass T (ed.) *Obedience to Authority: Current Perspectives on the Milgram Paradigm*. Erlbaum Associates, Mahwah, NJ, pp. 35–59

Brief A P, Dietz J, Cohen R R, Pugh S D, Vaslow J B 2000 Just doing business: Modern racism and obedience to authority as explanations for employment discrimination. *Organizational Behavior and Human Decision Processes* **81**: 72–97

French J R P Jr., Raven B H 1959 The bases of social power. In: Cartwright D (ed.) *Studies in Social Power*. Institute for Social Research, University of Michigan, Ann Arbor, MI, pp. 150–67

Hamilton V L, Sanders J 1995 Crimes of obedience and conformity in the workplace: Surveys of Americans, Russians, and Japanese. *Journal of Social Issues* **51**(3): 67–88

Kelman H C 1973 Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. *Journal of Social Issue* **29**(4): 25–61

Kelman H C, Hamilton V L 1989 *Crimes of Obedience: Toward a Social Psychology of Authority and Responsibility*. Yale University Press, New Haven, CT

Milgram S 1963 Behavioral study of obedience. *Journal of Abnormal and Social Psycholog* **67**: 371–8

Milgram S 1974 *Obedience to Authority: An Experimental View*. Harper & Row, New York

Modigliani A, Rochat F 1995 The role of interaction sequences and the timing of resistance in shaping obedience and defiance to authority. *Journal of Social Issues* **51**(3): 107–23

Raven B H 1999 Reflections on interpersonal influence and social power in experimental social psychology. In: Rodrigues A, Levine R V (eds.) *Reflections on 100 Years of Experimental Social Psychology*. Westview Press, Boulder, CO

Rochat F, Modigliani A 2000 Captain Paul Grueninger: The chief of police who saved Jewish refugees by refusing to do his duty. In: Blass T (ed.) *Obedience to Authority: Current Perspectives on the Milgram Paradigm*. Erlbaum, Mahwah, NJ

Tyler T R 1997 The psychology of legitimacy: A relational perspective on voluntary deference to authorities. *Personality and Social Psychology Review* **1**: 323–45

T. Blass

Obesity and Eating Disorders: Psychiatric

This article covers obesity and the eating disorders of anorexia nervosa and bulimia nervosa. People diagnosed as obese, anorexic, or bulimic eat too much or too little. As a result, they may weigh too much or too little, with consequent adverse effects on their health. Even considering each eating disorder separately, there are substantial individual differences with respect to etiology and the relative success of different types of treatment.

1. Anorexia Nervosa

People with anorexia nervosa eat less food than is necessary to sustain a healthy weight. To be diagnosed as anorexic someone must have a maximum body-mass index (BMI) of 17.5 kg/m^2 (BMI measures someone's mass per meter of height, and is calculated as weight in kilograms divided by the square of height in meters). Along with this loss of weight, females must stop menstruating to be classified as anorexic. Anorexics may occasionally go on feeding binges and then compensate for the calories ingested using such means as self-induced vomiting, laxatives, or excessive exercise. Anorexia nervosa usually occurs in females ranging in age from their teens to their 30s. The overall prevalence of the condition has been estimated at between five and ten anorexics per 1000 late-adolescent and early-adult females (American Psychiatric Association 1994).

Using the word anorexia for this disorder is misleading because, contrary to its meaning—which is lack of appetite—anorexics are completely obsessed with food and food consumption. They carefully calculate just how many calories they can and do consume. They think and dream about food constantly (Bruch 1978). Many have a different perception of their body size compared with the impression that other people have of them. Anorexics tend to see themselves as being heavier than they really are (Garfinkel and Garner 1982).

1.1 Possible Explanations and Causes

A great many explanations for the cause of anorexia nervosa have been proposed. These explanations include, first, the current fashion image of thinness, which encourages young women to feel that they must be very thin in order to be attractive (Garfinkel and Garner 1982). In addition, anorexia nervosa may be a way to gain attention, an attempt at individuality, a denial of sexuality (in addition to not menstruating, girls who are severely underweight lack secondary sexual characteristics), or an attempt to resolve various problems associated with overdemanding parents (Bruch 1973). In addition, the neurotransmitter systems of anorexics tend to exhibit a variety of abnormalities, although it is not clear if these abnormalities are a cause or an effect of the condition (Kaye 1995). Several of the behavioral characteristics of anorexia nervosa—dawdling over food, obsession with food, and hyperactivity—have also been observed in humans and nonhumans who have been deprived of food either in the laboratory or under natural conditions (Logue 1991).

It appears that no one cause will suffice to explain all cases of anorexia nervosa; in fact, it is likely that no one cause will suffice to explain any single case. There are many paths to this very complex disorder.

1.2 Treatments

Treatments for anorexia nervosa can be divided into two groups: short-term treatments designed primarily to induce an immediate weight gain in the patient, sometimes without her consent; and long-term treatments designed primarily to maintain or prolong a weight gain (Van Buskirk 1977). An example of a short-term treatment is nurturance, one or more different techniques that attempt to improve the patient's nutrition. Nurturance may take the form of bed rest accompanied by a high-calorie diet, or of force feeding the patient through a tube inserted down her throat (Garfinkel and Garner 1982). Other examples of short-term treatments are behavior therapy, in which engaging in activities known to be pleasant for the anorexic is made contingent on eating (Van Buskirk 1977), and drug therapy, in which specific types of medication are used to enhance appetite or mood (Walsh 1995). Each of these types of short-term therapy can be useful for obtaining an initial weight gain. However, given that there are potential medical problems associated with using any drug, use of drugs should be approached with caution.

Examples of long-term treatments are individual psychotherapy, family psychotherapy, and self-help groups. Within each of these types of treatment are many different approaches. Unfortunately there have been few controlled studies on the long-term effects of different treatments of anorexia nervosa, therefore we do not have a clear picture as to which treatments are successful in the long term for which types of patients.

Anorexia nervosa is a very serious disorder with an average mortality rate of 5 percent. However, most treated patients improve (Steinhausen 1995). We know that the overall treatment package used for patients, usually consisting of some form of nurturance plus psychotherapy, seems to have fairly good results.

2. Bulimia Nervosa

People suffering from bulimia nervosa repeatedly eat, in a short period of time, a very large amount of food (known as binge eating). In addition, bulimics frequently rid themselves of the excessive calories that they have consumed using self-induced vomiting, laxatives, starvation, or vigorous exercise. Bulimics are usually within normal weight ranges, and are, as with anorexia nervosa, mostly young women. The prevalence of bulimia nervosa has been estimated as between 1 and 3 bulimics per 100 adolescent and young-adult females.

Similar to anorexics, bulimics appear to have a distorted view of their own weight, and are obsessed with controlling their food intake. However, unlike anorexics, they repeatedly slip and consume huge amounts of food. Once the binge starts, it seems uncontrollable. Then the bulimic usually attempts to compensate for the effects of the excessive food consumption (American Psychiatric Association 1994).

Repeated binging and purging can have devastating physiological effects. Some of the less serious effects include sore throat and swollen glands caused by repeated vomiting; the vomitus being acidic, it eats away at the enamel of the teeth, making bulimics prone to tooth decay. More serious is that both the vomiting and the use of laxatives can cause dehydration and cardiac arrhythmias (Mitchell 1995). In addition, possibly due to their repeated cycles of binging and purging, bulimics have a significantly lower metabolic rate than do nonbulimics, making it relatively easy for bulimics to gain weight (Energy expenditure 1989; Gwirtsman et al. 1989).

2.1 Possible Explanations and Causes

Many of the possible explanations and causes that have been proposed for anorexia nervosa have also been proposed for bulimia nervosa. However, one possible cause appears to be unique to bulimia; many bulimics are depressed. In addition, bulimics tend to come from families with a history of mood disorders. Many mood disorders appear to have a genetic component, and it has been suggested that a single, genetically based chemical disorder may be at least partly responsible for both depression and bulimia (Strober 1995).

2.2 Treatments

As with *anorexia nervosa*, various types of psychotherapy have been used in treating bulimia. In addition, consistent with the hypothesis that the causes of bulimia nervosa and depression are in some way physiologically related, several researchers have reported improvement in the eating behavior of bulimics after administration of antidepressants. However, these findings are by no means uniform. Furthermore, some researchers feel it is necessary to combine psychotherapy with antidepressants to ensure both the cessation of unhealthy eating behavior and an increase in healthy eating behavior (Walsh 1995).

3. Obesity

There are many definitions of obesity. A common one is based on the BMI: people with BMIs in excess of $30\,kg/m^2$ are considered to be *obese*, and people with BMIs between 25 and $30\,kg/m^2$ are considered to be overweight. However, the BMI definitions of obesity and overweight are not perfect. The BMI does not distinguish between lean body mass, such as muscle and bone, and stored fat. Some researchers believe that the limits for overweight and obesity should be adjusted upwards as people age (Bray 1987).

Obesity is an increasing problem in the United States; approximately one-third of United States adults are obese. In general in the United States, more women than men are obese and obesity increases with age (Williamson 1995). Increasing obesity among children is a particular problem in the United States (Kennedy and Goldberg 1995). It has recently been estimated that the direct costs of health care related to obesity constitute 5.7 per cent of the total health-care costs in the United States. Medical conditions associated with obesity include high cholesterol, high blood pressure, and diabetes (Rössner 1998)

3.1 Possible Explanations and Causes

Contrary to popular belief, there are few consistent differences in the eating behavior of obese and nonobese people. Many people who are obese eat no more calories or no more sweet or fatty foods than do people who are not obese. Therefore particular aspects of eating behavior are, on average, not useful in explaining obesity (Logue 1991).

In contrast, there is no question that genes play a role in obesity; the biological children of obese parents also tend to be obese, and weight tends to be more similar between identical twins (who have identical genes) than between fraternal twins who on average share half of their genes; Gurney 1936). However, it is important to note that there are also significant correlations between the weights of adoptive mothers and their children; the environment also plays an important role (Foch and McClearn 1980).

One way in which genes may influence obesity is through the number of adipose (fat storage) cells that someone has. When these cells are not full, there is more hunger, and the number of adipose cells that someone has is influenced by heredity. In addition, if

someone gains weight, the number of adipose cells can increase, but their number can never decrease (Björntorp 1987, Sjöström 1980).

Much research has focused on energy expenditure as a factor contributing to obesity. Research has shown that individuals with identical body weights may have different metabolic rates; the lower the metabolic rate, the greater the person's tendency to become obese (Geissler 1988, Ravussin et al. 1988). Various factors can influence metabolic rate. For example, low levels of food consumption, such as occur with dieting, can decrease metabolic rate, and this decrease may last even after food consumption increases (Elliot et al. 1989). In contrast, exercise can increase metabolic rate, and the increase may last even after the exercise has stopped (Brownell 1995). Therefore lack of exercise and (paradoxically) dieting may contribute to obesity.

Many other factors have been demonstrated to contribute to eating large amounts of food. Such factors include easy access to highly palatable food, stress-induced eating, relatively great physiological responsiveness to food stimuli in the environment, and a learning history such that food comes to satisfy needs other than nutritional (Logue 1991). Nevertheless, it is important to remember that eating large amounts of food by itself does not necessarily result in obesity.

3.2 Treatments

There are very few treatments that have been reliably shown to decrease obesity for an extended period of time. Drugs such as amphetamines, diets, and behavior therapy can all result in a weight loss, but such losses are usually temporary. This is not surprising given what we know about adipose cells and metabolic rate, as described above.

One treatment that has long-term success is gastric bypass surgery, in which the lower part of the stomach is bypassed by first disconnecting it from the top part, and by then connecting the top part of the stomach directly to the intestine. Bypass surgery is extremely successful in decreasing obesity. Five years after the operation, patients have only about 40–50 percent of the excess weight that they had before the operation (Kral 1995). Patients also appear to be better adjusted and happier after surgery; their weight loss enables them to engage in many activities that were previously difficult or impossible, plus they feel more attractive. However, there can be some potentially dangerous side effects from this operation, with vomiting being the most common (Bray and Gray 1988). In addition, major surgery of any kind always involves risk. Therefore bypass surgery is not recommended for anyone unless they are severely overweight and have been unable to lose weight by other, less drastic, means.

The other treatment that has been shown to have significant long-term success in terms of weight loss is exercise. Exercise uses calories and under certain conditions can elevate metabolic rate for several hours following the exercise, thus burning additional calories (Brownell 1995). Further, exercise has more benefits than simply burning calories: cardiovascular fitness, increased muscle tone and bone strength, and improved mood (Dubbert and Martin 1988, Folkins and Sime 1981). Thus exercise has much to recommend it as a means of weight control.

4. Conclusion

It is extremely unfortunate that in a country such as the United States, where so many of life's problems have been solved or ameliorated by industrialization, other problems, such as obesity, anorexia nervosa, and bulimia nervosa have been enhanced. Our industrialized environment with its easy access to food, labor-saving devices, and fashion industry has aided and abetted situations in which people eat too much or too little. As long as our environment contains these elements, treating these disorders will remain extremely difficult. Nevertheless, scientists are making some progress in determining the best methods for treating obesity, anorexia nervosa, and bulimia nervosa.

See also: Body Image and Gender; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Eating Disorders, Determinants of: Genetic Aspects; Hunger and Eating, Neural Basis of; Obesity, Behavioral Treatment of; Obesity, Determinants of: Genetic Aspects; Weight Regulation: Psychological Aspects

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association, Washington, DC

Björntorp P 1987 Fat cell distribution and metabolism. In: Wurtman R J, Wurtman J J (eds.) *Human Obesity*. New York Academy of Sciences, New York

Bray G A 1987 Overweight is risking fate: Definition, classification, prevalence, and risks. In: Wurtman R J, Wurtman J J (eds.) *Human Obesity*. New York Academy of Sciences, New York

Bray G A, Gray D S 1988 Obesity. Part II—Treatment. *Western Journal of Medicine* **149**: 555–71

Brownell K D 1995 Exercise in the treatment of obesity. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford, New York

Bruch H 1973 *Eating Disorders*. Basic Books, New York

Bruch H 1978 *The Golden Cage*. Harvard University Press, Cambridge, MA

Dubbert P M, Martin J E 1988 Exercise. In: Blechman E A, Brownell K D (eds.) *Handbook of Behavioral Medicine for Women*. Pergamon, New York

Elliot D L, Goldberg L, Kuehl K S, Bennett W M 1989 Sustained depression of the resting metabolic rate after massive weight loss. *American Journal of Clinical Nutrition* **49**: 93–6

Energy expenditure and the control of body weight 1989 *Nutrition Reviews* **47**: 249–52

Foch T T, McClearn G E 1980 Genetics, body weight, and obesity. In: Stunkard A J (ed.) *Obesity*. Saunders, Philadelphia, PA

Folkins C H, Sime W E 1981 Physical fitness training and mental health. *American Psychologist* **36**: 373–89

Garfinkel P E, Garner D M 1982 *Anorexia Nervosa*. Brunner/Mazel, New York

Geissler C 1988 Genetic differences in metabolic rate. In: Birch G G, Lindley M G (eds.) *Low-calorie Products*. Elsevier, New York

Gurney R 1936 The hereditary factor in obesity. *Archives of Internal Medicine* **57**: 557–61

Gwirtsman H E, Kaye W H, Obarzanek E, George D T, Jimerson D C, Ebert M H 1989 Decreased caloric intake in normal-weight patients with bulimia: Comparison with female volunteers. *American Journal of Clinical Nutrition* **49**: 86–92

Kaye W H 1995 Neurotransmitters and anorexia nervosa. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford, New York

Kennedy E, Goldberg J 1995 What are American children eating? Implications for public policy. *Nutrition Reviews* **53**: 111–26

Kral J G 1995 Surgical interventions for obesity. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guildford, New York

Logue A W 1991 *The Psychology of Eating and Drinking: An Introduction*. Freeman, New York

Mitchell J E Medical complications of bulimia nervosa. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford, New York

Ravussin E, Lillioja S, Knowler W C, Christin L, Freymond D, Abbott W G H, Boyce V, Howard B V, Bogardus C 1988 Reduced rate of energy expenditure as a risk factor for body-weight gain. *New England Journal of Medicine* **318**: 467–472

Rössner S 1998 Why prevention of obesity has become necessary—A foreword. *Appetite* **31**: 407–36

Sjöström 1980 Fat cells and body weight. In: Stunkard A J (ed.) *Symposium on Obesity: Basic Mechanisms and Treatment*. Saunders, Philadelphia, PA

Steinhausen H-C 1995 The course and outcome of anorexia nervosa. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford, New York

Strober M 1995 Family-genetic perspective on anorexia nervosa and bulimia nervosa. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford, New York

Van Buskirk S S 1977 A two-phase perspective on the treatment of anorexia nervosa. *Psychology Bulletin* **84**: 529–38

Walsh B T 1995 Pharmacotherapy of eating disorders. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford, New York

Williamson D F 1995 Prevalence and demographics of obesity. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford, New York

A. W. Logue

Obesity, Behavioral Treatment of

1. Introduction

1.1 Definition

Cognitive behavior therapy is the leading treatment for obesity because it produces significant weight loss with minimal adverse side effects. Treatment components include a low-fat, balanced-deficit diet of approximately 1200–1800 kcal per day and a gradual increase in physical activity to 30 min per day. Patients keep extensive records of food intake and physical activity and use stimulus control procedures, such as limiting places and activities associated with eating. Modification of negative cognitions is increasingly used in treatment (Brownell 2000).

The mechanics of treatment are straightforward. Weekly individual or group sessions are provided for an average of 20 weeks. Bimonthly follow-up visits are provided to facilitate the maintenance of weight loss. University-based programs treat 8–20 patients in a group, but commercial programs may include as many as 100 people. Sessions are highly structured, and manuals frequently are provided to participants. Although the clinician typically introduces a new weight control skill each week, most of the session is spent reviewing homework, setting specific behavioral goals, and helping patients find solutions to barriers (Wadden and Foster 2000).

1.2 Scope of the Article

This article briefly reviews the development of the behavioral treatment of obesity. It examines short- and long-term results of treatment and considers methodological issues and future directions for research.

2. History of Obesity Treatment

2.1 Treatment in Ancient Greece through the Early Century

The treatment of obesity can be traced back to Hippocrates in Ancient Greece. His 'remedy' consisted of hard labor, sleeping on a firm bed, eating once a day, consuming fatty foods, and working naked as many hours a day as possible. From the time of Hippocrates to the twentieth century, obesity treatment remained remarkably consistent, targeting various aspects of food intake and exercise. This consistency is somewhat surprising in view of the ever-changing views of the etiology of obesity during this time (see Brownell 1995).

The greatest shift in treatment occurred in the 1940s and 1950s when psychoanalysis dominated mental health practice in America and other countries. Psychoanalysis contended that psychiatric problems were the expression of unconscious sexual or aggressive impulses. In this view, obesity was regarded as a manifestation of a basic personality problem that could involve oral fixation or efforts to compensate for feelings of sadness or inadequacy. Resolution of these conflicts was thought to be necessary for the patient to lose weight. Surprisingly, despite the lack of empirical support for this approach, psychodynamic therapy is still used to treat obesity. For a more extensive review of historical origins of obesity, see Bray (1990).

2.2 Founding and Development of Behavioral Treatment of Obesity

With the emergence of behavioral treatment in the 1960s, the focus shifted from internal, unconscious conflicts to the external behaviors associated with obesity. Ferster and co-workers were the first to suggest the application of behavioral principals to obesity treatment (Ferster et al. 1962). In this conceptualization, obesity was viewed as a learned disorder that could be 'unlearned' through specific behavioral principles. The main principles were operant conditioning (i.e., a behavior becomes more or less frequent as a function of the consequences it produces) and classical conditioning (i.e., a stimulus that is repeatedly paired with an unconditional stimulus will eventually provoke the same response as the unconditional stimulus). Thus, obesity resulted when a neutral stimulus (i.e., activities, people, places, times of day) were paired with eating and, subsequently, elicited eating in the absence of hunger. Appropriate eating habits and weight loss would be achieved with the use of stimulus control (i.e., limiting the activities, conditions, and people associated with eating) and reinforcement contingencies (i.e., pairing negative consequences with overeating and positive ones with eating appropriately).

The first empirical evaluation of behavioral treatment was published by Stuart (1967). Stuart's one-year program, which provided 26 half-hour individual sessions, produced an impressive average weight loss of 17.2 kg (about 19.2 percent of initial body weight) in eight obese individuals. The success of this program sparked a revolution in the treatment of obesity. Dozens of clinical trials followed in the next decade.

Over time, the content and process of treatment became more refined. In the 1970s, group behavioral treatment of obesity emerged, in large measure because it provided an economical and efficient means for graduate students to collect dissertation data. Regardless of its origin, group treatment caught on quickly and, today, is the standard in hospital practice and clinical trials. The 1980s also witnessed significant changes, as investigators sought new ways to induce larger weight losses and improve the maintenance of weight loss. Very low calorie diets were introduced, and exercise training became an increasingly important treatment component. Maintenance strategies, in addition to exercise, included the use of frequent (i.e., biweekly) follow-up visits and social support.

Several theorists have made a major impact on the field. Keesey and Stunkard's notion of a biological 'set point,' around which body weight was regulated, suggested that there were biological limits to weight loss. Bandura's self-efficacy theory and Marlatt and Gordon's relapse prevention model provided new methods to induce and maintain changes in eating and activity habits.

2.3 Evolution and Spread of Behavioral Treatment of Obesity

Today, behavioral treatment is considered integral to any comprehensive weight management program. It is combined with pharmacotherapy and surgical treatments for obesity and is used in commercial weight loss programs throughout the world. The principal components have remained the same, including self-monitoring, stimulus control, cognitive restructuring, and education in nutrition and exercise. However, emphasis increasingly has shifted toward viewing obesity as a chronic disorder that requires long-term treatment, to help individuals sustain changes in their eating and activity.

3. Research on Behavioral Treatment of Obesity

3.1 Short-term Results

Randomized control trials conducted in the 1970s showed that behavior therapy was consistently superior to comparison conditions, and produced an average loss of 3–4 kg in eight weeks (Table 1). As both the length and content of treatment expanded in the 1980s, weight losses nearly doubled (Table 1). However, the rate of weight loss remained unchanged at 0.4–0.5 kg per week, suggesting that longer treatment, not changes in its content, produced the greater weight loss. With longer treatment, attrition rates have increased, but still remain impressively low at, that is, 15–20 percent. A standard 20-week treatment currently produces an average loss of 8.5 kg, equal to a 9 percent reduction in initial body weight. Weight losses are associated with significant improvements in cardiovascular and psychosocial health.

3.2 Long-term Results

The overarching problem with behavioral treatment has been impaired maintenance of weight loss (Wilson 1994). In the year after treatment, patients typically

Table 1
Summary of randomized clinical trials from 1974 to 1995 examining behavioral treatment of obesity

	1974	1985–1987	1991–1995
Number of studies	15	13	5
Sample size	53.1	71.6	30.2
Initial weight (kg)	73.4	87.2	94.9
Length of treatment (week)	8.4	15.6	22.2
Weight loss (kg)	3.8	8.4	8.5
Loss per week (kg)	0.5	0.5	0.4
Attrition (percent)	11.4	13.8	18.5
Length of follow-up (week)	15.5	48.3	47.7
Loss at follow-up (kg)	4.0	5.3	5.9

Source: The data are adapted and updated with permission from Brownwell and Wadden (1986). All studies sampled were published in the following four journals: *Addictive Behaviors*, *Behavior Therapy*, *Behaviour Research and Therapy*, and *Journal and Clinical Psychology*. All values, except for number of studies, are weighted means. Thus, studies with larger sample sizes had a greater impact on the mean values than did studies with smaller sample sizes.

regain about 30–35 percent of their weight loss (Table 1). Weight regain generally increases over time. In 3–5 years, 50 percent or more of patients return to their baseline weight (Wadden and Foster 2000).

Long-term (i.e., 52–78 week) weekly or every-other-week group behavioral sessions promote the maintenance of weight loss. On average, patients maintain their full end-of-treatment loss as long as they attend maintenance sessions. Unfortunately, attendance of these sessions declines over time and, after treatment ends, patients tend to regain weight (Perri et al. 1988). Thus, maintenance sessions appear only to delay rather than to prevent weight regain.

A small minority, however, of obese individuals are able to lose weight and keep it off. The National Weight Control Registry is comprised of individuals who have lost a minimum of 13.6 kg (i.e., 30 lbs), using a variety of methods, and maintained the loss for at least one year (Klem et al. 1997). Participants have reported that they consume a low-calorie diet (1300–1800 kcal per day) and walk the equivalent of 45 kilometres per week. Based on these findings, it appears that behavioral treatment does target the essential components for long-term weight control. However, further research is needed to identify methods to facilitate patients' long-term adherence to a low-fat, low-calorie diet and high-activity lifestyle.

4. *Methodological Issues in the Conduct of Research*

4.1 *Sampling Issues*

Sampling bias is a major criticism of the behavioral treatment of obesity. The typical patient in a weight control study is 100 kg, educated, middle-aged, Caucasian, and female. Patients are likely to be highly motivated, as reflected by their willingness to complete the multiple questionnaires and medical tests required in a research study. Clearly, this subject profile does not reflect the heterogeneity of overweight and obese people in the general population. It is possible that if behavioral treatment were delivered to randomly selected members of the general population, they would achieve a less favorable outcome because of potentially lower motivation. Alternatively, they might achieve a better outcome given that participants in clinical trials are usually thought to have the most refractory obesity. The generalizability of findings from research trials needs to be determined. Similarly, studies are needed to determine whether increasing the cultural appropriateness of treatment will improve outcomes in minority populations (Kumanyika et al. 1992).

4.2 *Assessment Issues*

Behavioral scientists were first attracted to research on obesity because weight provided a 'simple' outcome measure. Thus, the 'success' of behavioral interventions typically has been assessed by the amount of weight lost. Most investigators believed that the greater the weight loss, the greater the success. However, the emphasis on achieving large weight losses, or ideal weight, has declined in the past decade with the realization the most patients cannot achieve more than a 10–15 percent reduction in initial weight. Thus, reaching ideal weight has given way to the achievement of a healthier weight. Losses of 5–15 percent of initial weight frequently are sufficient to improve weight-related health complications.

The shift from weight loss to health-related outcomes has important implications. For example, if behavioral treatment failed to produce long-term weight loss, it might succeed in modifying other outcomes associated with improved physical or mental health. Reducing fat intake, increasing physical activity, and improving coping skills may produce benefits beyond those obtained through weight loss alone. Blair and co-workers, for example, have shown that individuals who are fat but physically fit have a lower rate of all-cause mortality that do persons who are lean but unfit (i.e., sedentary). Thus, additional efforts are needed to expand the definition of 'successful' treatment to include multiple health domains. The Study of Health Outcomes of Weight Loss (SHOW) is a 6,000 person, multicenter, 12-year trial that will assess whether weight loss (and/or improved fitness) decreases mortality in obese individuals with type 2 diabetes.

Assessment of eating and activity habits has always been problematic. Obese, as well as lean, individuals underestimate their food intake while overestimating their physical activity. Difficulty in measuring these

variables has slowed progress in understanding the etiology of obesity and in identifying possible behavioral phenotypes. A new generation, however, of pedometers and accelerometers provides more accurate estimates of daily physical activity. Similarly, the doubly-labeled water method yields very accurate estimates of total energy expenditure and, thus, of food intake. These methods will allow investigators to assess more clearly both the biological and genetic factors that contribute to obesity.

4.3 Selection of Research Questions

Dozens of studies have demonstrated the benefits of the behavioral weight loss 'package.' However, it remains to be determined which of the many treatment components (e.g., cognitive techniques, relapse-prevention, self-monitoring, stimulus control) contributes to success. A few preliminary studies have attempted to identify the active ingredients in behavioral treatment. One study found that cognitive treatment did not significantly improve the effects of standard behavioral treatment (Collins et al. 1986). Another suggested that adding relapse prevention training to long-term treatment did not improve outcome compared to long-term care alone (Perri 1992). Additional controlled studies are needed to replicate these findings and determine the critical ingredients of the behavioral package. This could have important implications for treatment delivery if, for example, it were found that having patients simply 'weigh in' each week and submit a food record was as successful as having them attend 90-minute group sessions.

5. Future Directions

Investigators must continue efforts to improve the maintenance of weight loss. Combining behavioral treatment with pharmacologic agents, including sibutramine or orlistat, could well facilitate this goal. However, behavioral methods to prevent relapse must be examined, such as improving body image and/or promoting patients' acceptance of more modest weight losses.

Further research is also needed on the effects of tailoring treatments to subgroups of the obese population. Some individuals, for example, may do better with group treatment and others with individual therapy. Some patients may respond best to portion-controlled diets and others to diets of conventional foods. People with body image concerns or binge eating might benefit most from treatments focused on these disturbances. Research in this area is in its infancy and will require the identification of behavioral phenotypes, as noted earlier.

Finally, as the prevalence of obesity increases, public health campaigns will be needed to prevent the development of this disorder by attacking the 'toxic' environment that lies at the root of the epidemic. Technical advances, such as internet-based programs, may help broaden the impact of prevention and treatment efforts. Additionally, integrating treatment into community settings, primary care practices, and workplaces are needed. Ultimately, a partnership among multiple sectors of society will be required to target the many factors that have contributed to the fattening of industrialized nations (WHO 1998).

See also: Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Eating Disorders, Determinants of: Genetic Aspects; Health and Mortality, Body Stature as a Measure of; Obesity and Eating Disorders: Psychiatric; Obesity, Determinants of: Genetic Aspects; Weight Regulation: Psychological Aspects

Bibliography

Bray G A 1990 Obesity: Historical development of scientific and cultural ideas. *International Journal of Obesity* **14**: 909–26

Brownell K D 1995 History of obesity. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford Press, New York, pp. 381–5

Brownell K D 2000 *The LEARN Program for Weight Management 2000*, 8th edn. American Health Publishing Company, Dallas, TX

Brownell K D, Wadden T A 1986 Behavior therapy for obesity: Modern approaches and better results. In: Brownell K D, Foreyt J P (eds.) *The Handbook of Eating Disorders: The Physiology, Psychology, and Treatment of Obesity, Bulimia, and Anorexia*. Basic Books, New York

Collins R L, Rothblum E D, Wilson G T 1986 The comparative efficacy of cognitive and behavioral approaches to the treatment of obesity. *Cognitive Therapy Research* **10**(3): 299–318

Ferster C B, Nurnberger J I, Levitt E B 1962 The control of eating. *Journal of Mathetics* **1**: 87–109

Klem L, Wing R, McGuire T, Seagle M, Hill O 1997 A descriptive study of individuals successful at long-term maintenance of substantial weight loss. *American Journal of Clinical Nutrition* **66**: 239–46

Kumanyika S, Morssink C, Agurs T 1992 Models for dietary and weight change in African American women. *Ethnicity and Disease* **2**: 166–75

Perri M G 1992 Improving maintenance of weight loss following treatment by diet and lifestyle modification. In: Wadden T A, Van Itallie T B (eds.) *Treatment of the Seriously Obese Patient*. Guilford Press, New York, pp. 456–77

Perri M G, McAllister D A, Gange J J, Jordan R C, McAdoo W G, Nezu A M 1988 Effects of four maintenance programs on the long-term management of obesity. *Journal of Consulting and Clinical Psychology* **56**(4): 529–34

Stuart R B 1967 Behavioral control of overeating. *Behavior Therapy* **5**: 357–65

Wadden T A, Foster G D 2000 Behavioral treatment of obesity. *Medical Clinics of North America* **84**(2): 441–62

Wilson G T 1994 Behavioral treatment of obesity: Thirty years and counting. *Advances in Behavioral Research Therapy* **16**: 31–75

World Health Organization 1998 *Obesity: Preventing and Managing the Global Epidemic*. WHO, Geneva, Switzerland

T. A. Wadden

Obesity, Determinants of: Genetic Aspects

1. Background

Obesity is a worldwide epidemic. Thus it has been estimated that about 7 percent of the adults of the whole world have a body mass index (BMI; body weight in kilograms divided by height in meters squared) in the obesity range (i.e., 30 kg/m^2 or more). This translates into about 250 millions of obese adults (Seidell 2000). This amounts to about 300,000 deaths per year in the USA alone attributable to obesity (Allison et al. 1996). From a slowly increasing rate of obesity during the 1960s and 1970s, the last two decades have seen an explosion in the rate of increase. A rise of 30 to 50 percent in the prevalence rate between 1980 and 1998 has been noted in many countries (WHO 1998, NHLBI 1998). It is thought that this epidemic reflects an interplay between environmental and genetic influences. Although environmental changes must underlie the rapid progression of obesity in recent years, this explosion is occurring on a genetic background, perhaps one characterized by a vulnerability in the presence of current living conditions. Some populations such as the Polynesians and Micronesians, the Native Americans, and Africans in the diaspora seem particularly susceptible.

In this article we will examine the role that genetics has played in this epidemic. Recently, scientists involved in the study of the cause of human obesity have become optimistic about the possibility of identifying the genes associated with the predisposition to this disease. There are good reasons to believe that this new enthusiasm is justified. A growing understanding of the human genome, the high degree of homology between humans and common laboratory mammal models for a large number of genes and chromosomal regions, and the availability of a whole variety of technologies and tools to study and manipulate DNA in the laboratory are among the most important reasons for the present level of hope in the obesity research community.

2. Current Theory

The animal literature on obesity has produced a useful base for approaching the genetics of human obesity (Bray and York 1998, Kalra et al. 1999). This animal literature teaches us that we can view obesity as a disorder in a normally well-controlled system. There are four components in this system:

(a) The brain, or controller, which receives information from many sources, transduces this information into useful signals and controls food intake, the autonomic nervous system, muscular activity, and metabolism.

(b) The efferent signals by which the brain regulates food intake and metabolism involve voluntary muscle groups, the autonomic nervous system, and the hormonal system.

(c) The controlled system, which includes the organ systems involved with identifying, ingesting, digesting, absorbing, storing, or metabolizing nutrients and the excretion of gaseous, liquid, and solid waste.

(d) The afferent signals, which provide information to the brain, include the sensory organs, vagal and sympathetic nervous systems, circulating hormones like leptin, nutrient signals, and peripherally generated hormones.

This feedback approach can be used to locate the defects that produce significant obesity (Fig. 1).

Several defects in the brain produce obesity. Clinically, at the turn of the twentieth century, Frolich and Babinski described separate cases of hypothalamic obesity tumors that produced obesity (see Bray 1993). 'Hypothalamic obesity' can be produced routinely by damaging structures in the medial hypothalamus (paraventricular nucleus, PVN). This is probably the result of damaging areas that are critical to activating inhibitory systems in the medial hypothalamus.

Four genetic manipulations of this system also produce massive obesity. Leptin deficiency or leptin receptor defects impair the control of the afferent signals generated by leptin acting on the arcuate nucleus (ARC in Fig. 1). In the absence of leptin, neuropeptide Y (NPY) increases and proopiomelanocortin (POMC) decreases food intake, a setting that matches the hypothalamic lesion. Genetic overexpression of agouti-related peptide (AGRP), which can block the inhibitory effects of α-MSH produced from POMC by post-translational processing or targeted disruption at the melanocortin-4 receptor, also produces massive obesity (Bray 1998).

The other defects identified in Fig. 1 produce more moderate forms of obesity. Defects in the cholecystokinin$_A$ (CCK$_A$) receptor and in the bombesin-3 receptor produce modest obesity by interfering with afferent signaling targeted disruption of the serotonin 2C receptor. Similarly, overexpression of corticotropin releasing hormone (CRH) or disabling of the glucocorticoid receptor in the hypothalamus produces moderate obesity.

Melanin-concentrating hormone increases food intake. Targeted disruption of this gene produces leanness, suggesting a tonic involvement in feeding and body fat regulation. Two other peptides that increase food intake, NPY and orexin, do not appear to be essential since targeted disruption of peptide pro-

duction does not affect food intake or body fat. Any of the peptides listed above might serve as the basis for a new drug treatment for obesity.

3. Evidence from Genetic Epidemiology

In a report published by the Carnegie Institute of Washington in 1923, C. B. Davenport (Davenport 1923) described the first comprehensive attempt to understand the role of inheritance in human body mass for stature. His study demonstrated quite convincingly that body mass index (BMI) values were more similar among family members than among unrelated persons. However, he noted that normal weight parents sometimes have obese adult offspring. He also observed the converse: obese parents frequently have normal weight adult descendants. Thus, the genetics of human obesity is likely to be complicated.

Since Davenport's report, the literature on genetics and obesity has expanded and deepened as a variety of new approaches have come to the fore. The Danish Adoption Registry Study, the Veterans Administration Twins Study, the Quebec Family Study, and several other studies were instrumental in reawakening interest in this problem. An overview of some of these genetic epidemiologic data is presented below.

3.1 Heritability Levels

The level of heritability has been considered in a large number of twin, adoption, and family studies. The level of heritability is simply the fraction of the population variation in a trait (e.g., BMI) that can be explained by genetic transmission. Such studies use a variety of procedures to control or take into account age and sex differences. Results obtained by a good number of investigators indicate that the heritability level estimates depend on how the study was conducted and on the kinds of relatives upon which they are based (Table 1). For instance, studies conducted with identical twins and fraternal twins or identical twins reared apart have yielded the highest heritability levels with values clustering around 70 percent of the variation in BMI.

In contrast, the adoption studies have generated the lowest heritability estimates of about 30 percent or less. The family studies have generally found levels of heritability intermediate between the twin and the adoption study reports. A few investigations have included all or most of these kinds of relatives in the same analysis. Using analytical techniques developed to use all the information and maximum likelihood procedures, these studies have concluded that the true heritability estimate for BMI in large sample sizes was between 25 and 40 percent. Recent surveys undertaken in Sweden and the USA with the collaboration of severely obese and morbidly obese subjects together with information obtained on their parents, siblings, and spouses confirm that the genetic contribution to obesity may indeed be around 25 to 40 percent of the individual differences in BMI (Bouchard et al. 1998).

3.2 Familial Risk for Obesity

The risk of becoming obese when a first-degree relative is overweight or obese can be quantified using the lambda coefficient (λs) defined as the ratio of the risk of being obese when a biological relative is obese compared with the risk in the population at large, i.e., the prevalence of obesity (Risch 1990). Age- and gender-standardized risk ratios (equivalent to λs) obtained from 2349 first degree relatives of 840 obese probands and 5851 participants of the National Health and Nutrition Examination Survey III (NHANES III) revealed that the prevalence of obesity (BMI $\geqslant$ 30) is twice as high in families of obese individuals than in the population at large (Lee et al. 1997). Moreover, the risk increases with the BMI threshold used to define the obesity. Thus, the risk of extreme obesity (BMI $\geqslant$ 45) is about eight times higher in families of extremely obese subjects.

The familial clustering of morbid obesity was also investigated (Adams et al. 1993) in 221 families (n = 1560, subjects ages 18 years and older) in Utah ascertained through a single morbidly obese proband.

Table 1
Overview of the genetic epidemiology of human body fat/obesity (The table is based on the trends in about 50 different studies. In most of these studies, the BMI was the phenotype considered. In some cases, skinfolds or estimates of percentage body fat or fat mass were used. From Bouchard 1994)

	Heritability/ transmission	maternal/ paternal	familial environment
Nuclear families	30 to 50	No	Minor
Adoption studies	10 to 30	Mixed results	Minor
Twin studies	50 to 80	No	No
Combines strategies	25 to 40	No	Minor

All probands had to be at least 45.5 kg over their ideal body weight. About 48 percent of the morbidly obese probands had one or more first-degree relatives who were also morbidly obese. Using data from 20,455 families from the general Utah population, the authors estimated that the prevalence of morbid obesity reached 6 percent in the Utah population. Thus, the risk of morbid obesity was about eight times higher (6 percent vs. 48 percent) in families segregating for morbid obesity than in the general population. More recently, using data from 15,245 participants aged 7 to 69 years from the 1981 Canada Fitness Survey, it was shown that the familial risk of obesity was 5 times higher for relatives in the upper 1 percent of the distribution of BMI than in the general Canadian population (Katzmarzyk et al. 1999).

4. *Animal Models*

4.1 *Single Gene Models*

Between the first wave of genetic and family studies before World War II and the second wave beginning around 1985, there was a nearly 50-year hiatus in human studies. Filling this gap was a growing literature on the heritage of obesity in animal studies. Before World War II, only the dominantly inherited yellow-obese mouse had been identified. Following the spontaneous appearance of a recessively inherited form of obesity in mice in 1950, four other recessively inherited forms of obesity were described in mice (Table 2). Two genetic forms of obesity were also found in rats.

These animal models, through the skepticism of many, have proved to be a boon to the entire field of genetics and obesity. The genetic defect in the yellow-obese mouse unraveled in 1992, and a defect in a similar pathway is now known to produce human obesity in some individuals. The discovery of leptin, the defective product in the obese (*ob*/*ob*) mouse, has opened a whole new field of research on the control of energy stores. Leptin deficiency is the cause of obesity in three families described so far, and treatment with leptin reverses the defects. A defect in the leptin receptor causes obesity in mice, rats, and humans. Failure to convert prohormone peptides into their active peptides causes obesity in mice as well as humans. Thus, the lessons from experimental biology have paid rich rewards in the study of human obesity.

4.2 *Transgenics*

Studies in transgenic animals have added new depth and a few surprises to our understanding of the genetics of obesity (Table 3). This new technology produces two kinds of engineered animals by altering the germ-cell line. One type of transgenic mouse overexpresses a given peptide. One example of obesity produced by overexpression is the yellow-obese mouse. Overexpression of either the agouti protein itself or an agouti-related peptide will produce massive obesity.

The second transgenic strategy is to produce targeted disruption of a single gene—a so-called knockout. Knocking out leptin production or knocking out the melanocortin-4 receptor in the hypothalamus will each produce massive obesity. These engineered rodents have proven to be extremely useful for the dissection of the mechanisms regulating energy balance and for the understanding of the biological pathways leading to leanness or fatness.

5. *Human Single Gene Defects*

Although individuals affected with Prader-Willi Syndrome, Bardet-Biedl Syndrome, or other Mendelian obesity syndromes represent a very small fraction of the obese population, they reveal that single gene defects can be important in the development of human obesity. The number of cases with single gene mutations causing obesity or associated with it has increased substantially over the past few years, but still represents a very small fraction of all the obesity cases worldwide. There are now 30 cases reported in the literature involving 11 mutations in 7 genes (Pérusse

Table 2
Single-gene mutation rodent models of obesity

	Rodent			Human		
Mutations[a]	Chr	Genes	Inheritance	Chr	Genes	Gene product
Diabetes (db)[b]	4	Lepr	Recessive	1p31	LEPR	Leptin receptor
Fat (fat)	8	Cpe	Recessive	4q32	CPE	Carboxypeptidase E
Obese (ob)	6	Lep	Recessive	7q31.3	LEP	Leptin
Tubby (tub)	7	Tub	Recessive	11p15.5	TUB	Insulin signaling protein
Agouti yellow (A^y)	2	A^y	Dominant	20q11.2	ASIP	Agouti signaling protein
Fatty (fa)	5	fa	Recessive	1p31	LEPR	Leptin receptor
Koletsky (fa^k)	5	fa^k	Recessive	1p31	LEPR	Leptin receptor

Table 3
Transgenic models that affect body fat

Major Increase	Minor Increase
Leptin knock-out	Glucocorticoid receptor brain anti-sense
Leptin receptor knock-out	Glut-4 overexpression in fat
Melanocortin-4 receptor knock-out	CRH overexpression
n-HLH2 knock-out	Intracellular adhesion molecule-1 knock-out
Agouti overexpression	Brown adipose tissue Ablation
POMC knock-out	FABP II knock-out
AGRP overexpression	Bombesin-3 knock-out
Serotonin 2C receptor knock-out	CCK_A receptor deficient—Rat Glutamine:F6-P aminotransferase overexpression
Decrease	**Enhanced Susceptibility to High-Fat Diet**
Leptin overexpression	UCP-1 knock-out
MCH	MCH overexpression
Glut-4 knock-out	Block effect of high-fat diet
LPL overexpression	PTB-1B knock-out
Glucokinase over-expression	—
Dopamine D-1R knock-out	—
Proteinkinase A (RI_α) knock-out	—

et al. 1999, Chagnon et al. 2000). These cases are listed in Table 4 and are reviewed briefly below.

In the leptin receptor (LEPR) gene, a G to A base substitution in exon 16 was found in three severely obese individuals who also showed pituitary dysfunction (Clément et al. 1998). Mutations in the proopiomelanocortin (POMC) gene were found in two patients with severe early-onset obesity, adrenal insufficiency, and red hair pigmentation (Krude et al. 1998). One was a compound heterozygote for two mutations in exon 3, a G to T change at nucleotide (nt) 7013, causing a premature termination at codon 79 and a 1 bp deletion at nt 7133 which also caused a premature termination, but at codon 131. The other subject was homozygous for a C to A transversion at nt 3804 in exon 2. Mild obesity along with combined hypothyroid and hyperthyroid symptoms was reported in a heterozygous patient with a point mutation in exon 10 of the thyroid hormone receptor, beta (THRB) gene, which produced a stop codon with a 28 aminoacid deletion and abolished T3 binding (Behr et al. 1997).

Four markedly obese individuals have been shown to be carriers of a missense mutation in the peroxisome proliferator activated receptor, gamma (PPARG2) gene (Ristow et al. 1998). This Pro to Gln conversion at position 115 was associated with a higher BMI compared with an obese control population without the mutation. One case of an impaired prohormone processing associated with a G to A substitution at codon 483 in the protein convertase subtilisin/kexin type 1 (PCSK1) gene was also reported (Jackson et al. 1997). One of the earliest identified mutations causing obesity was reported by Montague (Montague et al. 1997) in the leptin (LEP) gene. Two severely obese children with a mutation in codon 133 (G398Δ) were then identified. Subsequently, three individuals with a mutation (C105T) in exon 3 of the LEP gene were reported (Strobel et al. 1998). Further investigation of the same family yielded an additional female member carrying the same mutation (Ozata et al. 1999).

Two different groups reported on nine patients who were found to have mutations in the melanocortin 4 receptor (MC4R) gene associated with a dominant form of obesity (Yeo et al. 1998, Vaisse et al. 1998). Two heterozygous males shared a CTCT deletion at codon 211 which resulted in a missing leucine and a premature stop codon. Seven other subjects had a GATT insertion at nt 732 resulting in the expression of a nonfunctional truncated receptor. More recently, Hinney et al. (1999) described six female obese subjects with mutations in the MC4R gene, two with the already identified CTCT deletion at codon 211, and four others (two probands and their respective mothers) with a novel mutation at position 35 leading to a premature stop codon generating a truncated protein product. Seven missense mutations in MC4R of unknown significance in seven other extremely obese subjects (BMI 99th percentile) were also described.

6. *Scanning the Whole Human Genome*

Several other strategies have been used in the effort to identify the genes and mutations responsible for the predisposition to obesity. Among them, one finds a

Table 4
Cases of human obesity caused by single-gene mutations (adapted from Perusse et al. 1999 and Chagnon et al. 2000)

Gene	Location	Mutation	Case no.	Sex	Age	Wt (Kg)	BMI (Kg/m²)	References
LEPR	1p31	G → A (exon 16)	1	F	19	166	65.5	Clément et al. 1998
			2	F	13	159	71.5	
			3	F	19	133	52.5	
POMC	2p23	G7013T & C deletion at nt 7133 exon3	4	F	3	30	NA	Krude et al. 1998
		C3804A exon2	5	M	7	50	NA	
THRB	3p24.1-p22	Cys434Stop C → v A exon 10	6	F	15	46	26.3	Behr et al. 1997
PPARG	3p25	Pro115Gln	7	M	65	NA	47.3	Ristow et al. 1998
			8	F	32	NA	38.5	
			9	M	54	NA	43.8	
			10	M	74	NA	37.9	
PCSK1	5q15-q21	Gly483Arg A → C^{+4} intron5	11	F	3	36	NA	Jackson et al. 1997
LEP	7q31	G398Δ (codon 133)	12	F	8	86	45.8	Montague et al. 1997
			13	M	2	29	36.6	
		C → T (codon105) (exon 3)	14	F	6	NA	32.5	Strobel et al. 1998
			15	M	22	NA	55.8	
			16	F	34	NA	46.9	
			17	F	30	130	54.9	Ozata et al. 1999
1.1.1 MC4R	18q21.3	ΔCTCT nt 631–634 (codon 211)	18	M	4	32	28	Yeo et al 1998
			19	M	30	139	41	
			20	F	20	NA	42.1	Hinney et al. 1999.
			21	F	43	NA	37.6	
		GATT insertion at nt 732 (codon 246)	22	F	58	NA	51	Vaisse et al. 1998.
			23	F	35	NA	57	
			24	F	34	NA	50	
			25	M	24	NA	33	
			26	F	11	NA	30	
		C105A (Tyr35X)	27	F	10	NA	31.3	Hinney et al. 1999
			28	F	17	NA	45.9	
			29	F	31	NA	48.2	
			30	F	36	NA	38.6	

whole series of studies on candidate genes. These have been reviewed elsewhere (Bouchard et al. 1998, Chagnon et al. 2000) and will not be discussed here. In general, however, the results of these observational studies have been disappointing, with small sample size, small effect size, and lack of replication studies or failure to replicate being the most common problems.

Another approach has been that of performing genome-wide scans with a view to identifying chromosomal regions of interest. These studies have typically been performed with pairs of sibs, sometimes whole nuclear families or pedigrees, and are based on about 300 microsatellite markers and more. The results of two such genomic scan studies pertaining to massive

obesity are summarized briefly here. They are based on families recruited through a proband with a BMI greater than 40.

The first has been conducted on a French cohort including 514 subjects (264 sib-pairs) from 158 nuclear families. Multipoint linkage analysis using affected sib-pairs revealed significant evidence of linkage with obesity (Lod score = 4.85) on chromosome 10p12 (Hager et al. 1998). The second study was based on a mixed sample of Caucasian and African-American subjects from 124 nuclear families (Lee et al. 1999). The authors tested linkage and found significant evidence of linkage ($3.02 \leqslant$ Lod score $\leqslant 3.16$) between obesity and markers located on chromosome 20q13. Other genomic scans for BMI or body fat levels measured variously have also been published from the San Antonio Heart Family Study, the Pima Indian Sibling Study, and the Quebec Family Study, but will not be reviewed here as they focus on less severe obesity.

7. Conclusion

It is quite likely that genetic susceptibility to the more severe forms of human obesity is stronger than for the milder types of obesity or for simple overweight. This is concordant with the observation that the risk of morbid obesity is higher when parental obesity is present and with an early-onset of the disease as shown by a recent study (Whitaker et al. 1997).

Obesity is a complex multifactorial trait evolving under the interactive influences of dozens of affectors from the social, behavioral, physiological, metabolic, cellular, and molecular domains (Bouchard et al. 1998). Segregation of the genes is not easily detected in familial or pedigree studies and whatever the influence of the genotype on the etiology, it is generally attenuated or exacerbated by nongenetic factors. From the research currently available, a good number of genes seem to have the capacity to cause obesity or increase the likelihood of becoming obese. This may be a reflection of how most human obesity comes about. In other words, the susceptible genotypes may result from allelic variations at a good number of genes. There are, however, cases in which a single mutation in a key gene is sufficient to cause obesity. Several such mutations have been uncovered in the past few years and many new ones are likely to be evidenced in the future. However, it must be recognized that the common forms of obesity do not arise as a result of a single mutation. Recent progress in animal genetics, transfection systems, transgenic animal models, recombinant DNA technologies applied to positional cloning, and methods to identify loci contributing to quantitative traits have given a new impetus to this field. Because the peptides and monoamines that are being targeted also have other behavioral consequences, the entire scope of information in the behavioral and social sciences will be enhanced. The stage is now set for major advances to occur in the understanding of the genetic and molecular basis of complex diseases such as human obesity.

See also: Body: Anthropological Aspects; Body, Evolution of; Body Image and Gender; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Eating Disorders, Determinants of: Genetic Aspects; Genetic Studies of Behavior: Methodology; Genetics of Complex Traits Through the Life Cycle; Hunger and Eating, Neural Basis of; Obesity and Eating Disorders: Psychiatric; Obesity, Behavioral Treatment of

Bibliography

Adams T D, Hunt S C, Mason L A, Ramirez M E, Fisher A G, Williams R R 1993 Familial aggregation of morbid obesity. *Obesity Research* **1**: 261–70

Allison D B, Faith M S, Nathan J S 1996 Risch's lambda values for human obesity. *International Journal of Obesity* **20**: 990–9

Behr M, Ramsden D B, Loos U 1997 Deoxyribonucleic acid binding and transcriptional silencing by a truncated c-erbA beta-1 thyroid hormone receptor identified in a severely retarded patient with resistance to thyroid hormone. *Journal of Clinical Endocrinology and Metabolism* **82**: 1081–7

Bouchard C (ed.) 1994 *The Genetics of Obesity*. CRC Press, Boca Raton, FL

Bouchard C, Perusse L, Rice T, Rao D C 1998 The genetics of human obesity. In: Bray G A, Bouchard C, James W P T (eds.) *Handbook of Obesity*. Marcel Dekker, New York

Bray G A 1993 Commentary on classics of obesity: 4. Hypothalamic obesity. *Obesity Research* **1**: 325–8

Bray G A 1998 *Contemporary Diagnosis and Management of Obesity*. Handbooks in Health Care Company, Newton, PA

Bray G A, York D A 1998 The MONA LISA hypothesis in the time of leptin [Review]. *Recent Progress in Hormone Research* **53**: 95–117 [discussion: 117–18]

Chagnon Y C, Perusse L, Weisnagel S J, Tankinen T, Bouchard C 2000 The human obesity gene map: The 1999 update. *Obesity Research* **8**: 89–117

Clément K, Vaisse C, Lahlou N, Cabrol S, Pelloux V, Cassuto D, Gourmelen M, Dina C, Chambaz J, Lacorte J M, Basdevant A, Bougneres P, Lebouc Y, Froguel P, Guy-Grand B 1998 A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392**: 398–401

Davenport C B 1923 *Body build and its inheritance*. Carnegie Institution of Washington, Washington, DC

Hager J, Dina C, Francke S, Dubois S, Houari M, Vatin V, Vaillant E, Lorentz N, Basdevant A, Clement K, Guy-Grand B, Froguel P 1998 A genome-wide scan for human obesity genes reveals a major susceptibility locus on chromosome 10. *Nature Genetics* **20**: 304–8

Hinney A, Schmidt A, Nottebom K, Heibult O, Becker I, Ziegler A, Gerber G, Sina M, Gorg T, Mayer H, Siegfried W, Fichter M, Remschmidt H, Hebebrand J 1999 Several mutations in the melanocortin-4 receptor gene including a nonsense and a frameshift mutation associated with dominantly inherited obesity in humans. *Journal of Clinical Endocrinology and Metabolism* **84**: 1483–6

Jackson R S, Creemers J W M, Ohagi S, Raffin-Sanson M L, Sanders L, Montague C T, Hutton J C, O'Rahilly S 1997 Obesity and impaired prohormone processing associated with

mutations in the human prohormone convertase 1 gene. *Nature Genetics* **16**: 303–36
Kalra S P, Dube M G, Pu S, Xu B, Horvath T L, Kalra P S 1999 Interacting appetite-regulating pathways in the hypothalamic regulation of body weight [Review]. *Endocrine Reviews* **20**(1): 68–100
Katzmarzyk P T, Pérusse L, Rao D C, Bouchard C 1999 Familial risk of obesity and central adipose tissue distribution in the general Canadian population. *American Journal of Epidemiology* **149**: 933–42
Krude H, Biebermann H, Luck W, Horn R, Brabant G, Gruters A 1998 Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nature Genetics* **19**: 155–7
Lee J H, Reed D R, Li W D, Xu W, Joo E J, Kilker R L, Nanthakumar E, North M, Sakul H, Bell C, Price R A 1999 Genome scan for human obesity and linkage to markers in 20q13. *American Journal of Human Genetics* **64**: 196–209
Lee J H, Reed D R, Price R A 1997 Familial risk ratios for extreme obesity: Implications for mapping human obesity genes. *International Journal of Obesity and Related Metabolic Disorders* **21**: 935–40
Montague C T, Farooqi I S, Whitehead J P, Soos M A, Rau H, Wareham N J, Sewter C P, Digby J E, Mohammed S N, Hurst J A, Cheetham C H, Earley A R, Barnett A H, Prins J B, O'Rahilly S 1997 Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387**: 903–8
National Institutes of Health and National Heart, Lung, and Blood Institute 1998 Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: The Evidence Report. *Obesity Research* **6** (Suppl. 2): 51S–209S
Ozata M, Ozdemir I C, Licinio J 1999 Human leptin deficiency caused by a missense mutation: Multiple endocrine defects, decreased sympathetic tone, and immune system dysfunction indicate new targets for leptin action, greater central than peripheral resistance to the effects of leptin, and spontaneous correction of leptin-mediated defects. *Journal of Clinical Endocrinology and Metabolism* **84**: 3686–95
Pérusse L, Chagnon Y C, Weisnagel J, Bouchard C 1999 The human obesity gene map: The 1998 update. *Obesity Research* **7**: 111–29
Risch N 1990 Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics* **46**: 222–8
Ristow M, Müller-Wieland D, Pfeiffer A, Krone W, Kahn C R 1998 Obesity associated with a mutation in a genetic regulator of adipocyte differentiation. *New England Journal of Medicine* **339**: 953–9
Seidell J C 2000 The current epidemic of obesity. In: Bouchard C (ed.) *Physical Activity and Obesity*. Human Kinetics Publishers, Champaign, IL
Strobel A, Issad T, Camoin L, Ozata M, Strosberg A D 1998 A leptin missense mutation associated with hypogonadism and morbid obesity. *Nature Genetics* **18**: 213–15
Vaisse C, Clement K, Guy-Grand B, Froguel P 1998 A frameshift mutation in human MC4R is associated with a dominant form of obesity. *Nature Genetics* **20**: 113–14
Whitaker R C, Wright J A, Pepe M S, Seidel K D, Dietz W H 1997 Predicting obesity in young adulthood from childhood and parental obesity. *New England Journal of Medicine* **337**: 869–73
World Health Organization 2000 *Obesity—Preventing and Managing the Global Epidemic. Report of a WHO consultation on obesity*. World Health Organization, Geneva
Yeo G S H, Farooqi I S, Aminian S, Halsall D J, Stanhope R G, O'Rahilly S 1998 A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nature Genetics* **20**: 111–12

G. A. Bray and C. Bouchard

Object Recognition: Theories

The goal of object recognition is to determine the identity or category of an object in a visual scene from the retinal input. In naturalistic scenes, object recognition is a computational challenge because the object may appear in various poses and contexts—i.e., in arbitrary positions, orientations, and distances with respect to the viewer and to other objects. The visual information that lands on the retina is a two-dimensional projection of a three-dimensional world. Object recognition is a challenge because many different three-dimensional objects may yield the same two-dimensional projection, and a three-dimensional object in different poses and contexts may produce grossly different two-dimensional projections. Consequently, theories of object recognition must overcome the effect of viewpoint on the appearance of an object. *Object constancy*, also known as *viewpoint-invariant recognition*, is achieved when an object is identified as being the same regardless of its pose and other objects in the scene.

Object recognition involves matching representations of objects stored in memory to representations extracted from the visual image. The debate in the literature concerns the nature of the representation extracted from the image. Broadly, theories of object recognition might be contrasted along five logically independent dimensions (aggregated from Hummel 1994, Hummel and Stankiewicz 1998, Tarr and Bülthoff, 1998, Tarr 1999):

(a) What are the primitive features or parts extracted from the visual image that form the basis of the representation? Features proposed include conjunctions of edge segments, two-dimensional contours, wavelets, local surface patches (with a specified orientation, color, and texture), generalized cylinders (cylindrical shapes whose principal axis and/or cross-sections may be stretched or otherwise distorted), or geons (primitive three-dimensional shapes, described further below).

(b) How stable is the set of features extracted across transformations of the image? The processing that underlies feature extraction might—at one extreme—produce a set of features that depends strongly on the relationship of the viewer to the object, or—at the other extreme—a set of features that is the same for nearly any two views of an object.

(c) How is the configuration of the features encoded? The pose of each feature might be repre-

sented with respect to a single point of reference, or the pose of one feature might be specified relative to the poses of other features, possibly in a hierarchy of features of features.

(d) How stable is the representation of a configuration of features across transformations of the image? The representation of the configuration might—at one extreme—depend strongly on the relationship of the viewer to the object, or—at the other extreme—be the same for nearly any two views of an object.

(e) What relationships are used to describe the possible configurations? The relationships might range from qualitative (e.g., *above*) to quantitative (e.g., 1.5 *units along the z axis*), with intermediate possibilities (e.g., *between* 2 *and* 5 *units along the z axis*).

Although theories of object recognition differ in many other regards, these five dimensions have been the primary focus of theoretical debates and behavioral studies. By situating various theories of object recognition in a space characterized by these five dimensions, contrasts among theories can be made explicit. Because the five dimensions are fairly independent of one another, experiments can be designed to efficiently explore the space of possibilities by considering one dimension at a time.

1. Frames of Reference

To understand better these five dimensions, it is necessary to introduce the concept of a *frame of reference* with respect to which visual features and their configurations are encoded. The reference frame specifies the center location, the up-down, left-right, and front-back directions, and the relative scale of axes used to form a description. Figure 1 shows two different reference frames. In terms of the reference frame centered on the telephone, the buttons would be described as being on *top* of the base and in *front* of the handle, whereas in terms of the reference frame centered on the viewer, the buttons would be described as being in the *back* of the base and to the *left* of the handle. A reference frame prescribed by the gaze, head orientation, and/or torso position of the viewer is known as *viewer-based* or *egocentric*. A reference frame prescribed by intrinsic characteristics of an object, such as axes of symmetry or elongation, or knowledge of the object's standard orientation, is known as *object-based* or *allocentric*. A frame of reference makes certain information in the stimulus explicit and therefore readily available for utilization in information processing, and hides other information, making it less available. The stability of a set of features and their configuration across transformations of the image (dimensions a and d above) depend directly on the frame of reference. If an allocentric frame is used to encode features and their configurations, the representation should be stable across transformations of the image, and object constancy should be readily achieved. If an egocentric frame is used, representations of two different views of an object may have little overlap, and object constancy is not a given. It is possible to imagine a frame of reference that is allocentric in some respects (e.g., center position and principal axis) but egocentric in other respects (e.g., left–right and up–down directions), allowing for an intermediate degree of stability.

2. Two Influential Classes of Theories

Traditionally, theories of object recognition have often been divided into two broad classes, *view-based* theories and *structural-description* theories (Hummel and Stankiewicz 1998, Tarr 1999). The key assumption of view-based theories is that the features extracted from the image are viewpoint dependent (dimension b) and that configurations of features encoded are also viewpoint dependent (dimension d). The key assumption of structural-description theories is that the pose of each feature is represented with respect to the reference frames of other features, not to a global reference frame (dimension c). Because the central

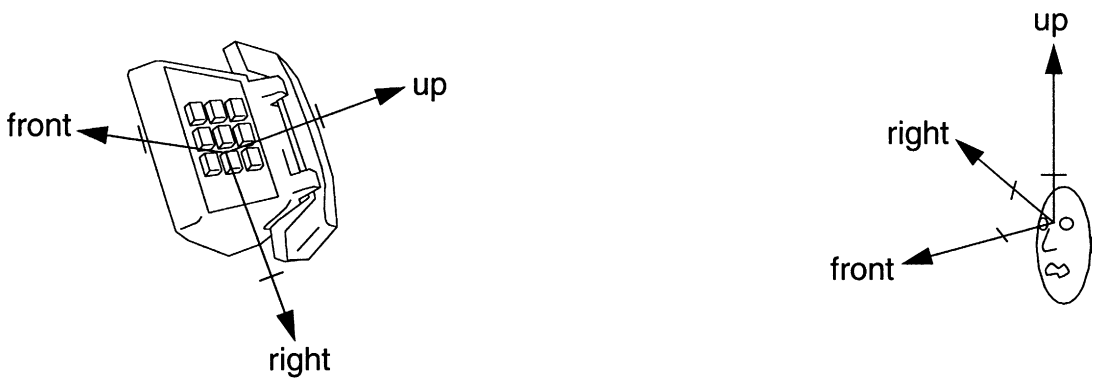

Figure 1
Two reference frames that can describe the telephone

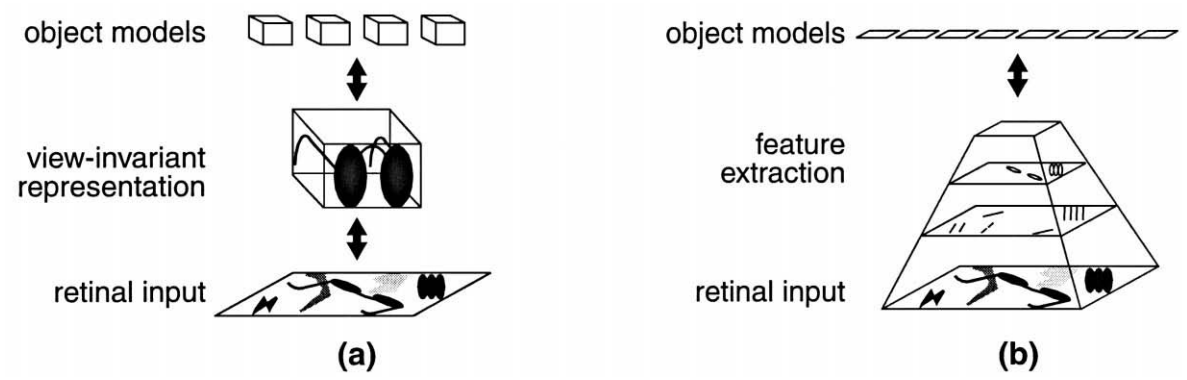

Figure 2
Two different approaches to achieving viewpoint-invariant object recognition: (a) from a structural-description theory approach; (b) from a neurobiologically motivated view-based theory approach

assumptions of the two classes focus on different dimensions, the classes are not necessarily incompatible. Historically, however, they have been viewed as antagonistic, due in part to the comparison of specific theories of each class, which make particular assumptions along all five dimensions. Several influential theories of each class are described in the following section.

2.1 Structural-description Theories

An early structural-description theory of object recognition was outlined by Marr and Nishihara (1978, see also Marr 1982, Pinker 1984). According to this theory, an internal description of an object's structure is constructed from observed visual features, essentially by transforming the viewer-based retinal input to an *object-based representation* (Fig. 2a)—a representation of the relationship of object parts to one another, independent of the viewer, using allocentric frames of reference for both the parts and configurations. This transformation solves the problem of viewpoint-invariant recognition, because every view of an object maps to the same object-based representation (ignoring the issue of occluded features).

Since Marr and Nishihara's seminal work, a variety of structural-description theories have been proposed. Some focus on rigid and two-dimensional objects—either implicitly or explicitly—in which case the object-based representation can be constructed using only image-plane transformations and recognition can be achieved in a straightforward manner by matching the object-based representation to stored templates of familiar objects (Hinton 1981, Humphreys and Heinke 1998, Olshausen et al. 1993, Zemel et al. 1988).

Other structural-description theories in psychology tackle the task of recognizing the sorts of objects encountered in the real world—complex, articulated objects with three-dimensional structure (Biederman 1987, Hummel and Biederman 1992, Hummel and Stankiewicz 1996, 1998). In the style of Marr and Nishihara, these theories operate on structural descriptions that decompose an object into its parts. Structural descriptions can be hierarchical, decomposing parts into parts of parts. The parts are described in terms of a relatively small catalog of primitives. Parts are linked by a small vocabulary of spatial relationships that describe how one part is situated with respect to another.

For example, the influential *recognition-by-components* (*RBC*) theory of Biederman (1987) invokes 36 primitives, called *geons*. Geons are detected by non-accidental properties in an image. The attributes of a geon include whether its edges are straight or curved, whether its size is constant, expanded, or contracted, and whether it is symmetric about one or more axes. The relationships among geons are captured by qualitative properties such as whether one geon is larger than another, whether one geon is above, below or to the side of another, and whether the geons are joined end-to-end or end-to-side. Because in principle any shape can be described by the finite set of primitives and relationships, the Biederman theory allows for the representation of novel shapes. Geons and their relationships can be represented by a graph structure, and recognition of an unknown shape can be performed by matching the structure of the unknown shape to the structure of a stored model. A particularly elegant aspect of this theory is that only one stored model is needed to achieve object constancy.

Hummel and Biederman (1992), Hummel (1994), and Hummel and Stankiewicz (1996, 1998) have proposed structural-description theories that differ from the Biederman theory along certain dimensions. All claim that frames of reference for describing features and their configurations are at least partially

egocentric. Additionally, the model of Hummel and Stankiewicz (1998) can accommodate quantitative as well as qualitative spatial relationships.

2.2 View-based Theories

Turning now to view-based theories, consider a naive scheme in which two-dimensional images of viewed objects are stored during learning, and these stored templates are matched pixel-by-pixel to an image containing an object to be recognized. Although this scheme seems hopelessly simplistic, minor variants of the scheme achieve a surprising degree of generalization to novel poses of three-dimensional objects by interpolation between stored poses (Poggio and Edelman 1990, Poggio and Shelton 1999, Siebert and Waxman 1990, Ullman 1989, Weinshall et al. 1990).

View-based theories that focus on explaining human vision go beyond the simple two-dimensional template matching idea by taking into account the coarse-scale anatomy of visual cortex (Fukushima and Miyake 1982, Hubel and Wiesel 1979, Le Cun et al. 1989, Mozer 1991, Perrett and Oram 1998, Reisenhuber and Poggio 1999, Sandon and Uhr 1988, Wallis and Rolls 1997). Specifically, two properties of visual cortex are generally deemed relevant. First, visual cortex is hierarchically organized, with simple, low-order, view-specific feature detectors at the earliest stage of vision, and increasingly complex, higher-order, and view-invariant detectors at subsequent stages (depicted in Fig. 2b by the pyramid structure). Second, information is processed from many locations in the visual field simultaneously and at many scales and orientations. Rather than being forced to choose a single reference frame, parallelism of the visual system allows multiple transformations of detectors at each stage in the hierarchy. In this framework, the focus of processing is on extracting features that reliably indicate the presence on an object, not on constructing a view-invariant representation. Nonetheless, a partially view-invariant representation may come to dominate later stages of processing in the service of recognition. A view-based theory may require multiple characteristic stored views to achieve object constancy.

2.3 Contrasting and Unifying the Theories

In recent years, a heated debate has ensued between proponents of view-based and structural-description theories (Biederman and Gerhardstein 1995, Biederman and Kalocsai 1997, Edelman 1997, Edelman and Duvdevani-Bar 1997, Perrett et al. 1999, Tarr 1999, Tarr and Bülthoff 1995). The debate is not between the two classes in terms of their essential claims, because the two classes address orthogonal dimensions in the space of possible theories. Rather, the debate is between a particular family of view-based theories and a particular family of structural-description theories—a view-based theory in which features are coded with respect to a single egocentric reference frame, and a structural-description theory in which structural information about features is explicitly encoded and in an allocentric reference frame. It is in principle possible to develop view-based theories that encode explicit structural information about features (Tarr and Bülthoff 1998), and structural-description theories based on egocentric frames of reference (Hummel and Stankiewicz 1998).

Each class of theories has its strengths. View-based theories are more consistent with the neurobiology of visual cortex and with evidence suggesting the neurobiological and psychological reality of view-tuned representations. Structural-description theories are more flexible in that they can represent novel and complex shapes, they can generalize from single examples, and have the ability to categorize objects as well as recognize individual members of a category. Given the growing understanding that the two classes of theories have complementary strengths and are not fundamentally incompatible, the next few years should see the development of accounts that integrate view-based and structural-description theories.

See also: Face Recognition Models; Face Recognition: Psychological and Neural Aspects; Neural Representations of Objects; Perceptrons; Perceptual Constancy: Direct versus Constructivist Theories; Perceptual Organization; Recognition Memory (in Primates), Neural Basis of; Recognition Memory, Psychology of; Statistical Pattern Recognition; Vision, High-level Theory of; Vision, Low-level Theory of

Bibliography

Biederman I 1987 Recognition-by-components: A theory of human image understanding. *Psychological Review* **94**: 115–47

Biederman I, Gerhardstein P C 1995 Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance* **21**: 1506–14

Biederman I, Kalocsai P 1997 Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London, Series B (Biological Sciences)* **352**: 1203–20

Edelman S 1997 Computational theories of object recognition. *Trends in Cognitive Science* **1**: 296–304

Edelman S, Duvdevani-Bar S 1997 A model of visual recognition and categorization. *Philosophical Transactions of the Royal Society of London, Series B (Biological Sciences)* **352**: 1191–1202

Fukushima K, Miyake S 1982 Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition* **15**: 455–69

Hinton G E 1981 A parallel computation that assigns canonical object-based frames of reference. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* Morgan Kaufmann, Los Altos, CA, pp. 683–5

Hubel D H, Wiesel T N 1979 Brain mechanisms of vision. *Scientific American* **241**: 150–62

Hummel J E 1994 Reference frames and relations in computational models of object recognition. *Current Directions in Psychological Science* **3**: 111–6
Hummel J E, Biederman I 1992 Dynamic binding in a neural network for shape recognition. *Psychological Review* **99**: 480–517
Hummel J E, Stankiewicz B J 1996 An architecture for rapid, hierarchical structural description. In: Inui T, McClelland J (eds.) *Attention and performance XVI: Information Integration in Perception and Communication*. MIT Press, Cambridge, MA, pp. 93–121
Hummel J E, Stankiewicz B J 1998 Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Cognition* **5**: 49–79
Humphreys G W, Heinke D 1998 Spatial representation and selection in the brain: Neurophychological and computational constraints. *Visual Cognition* **5**: 9–47
Le Cun Y, Boser B, Denker J S, Hendersen D, Howard R E, Hubbard W, Jackel L D 1989 Backpropagation applied to handwritten zip code recognition. *Neutral Computation* **1**: 541–51
Marr D 1982 *Vision*. Freeman, San Francisco
Marr D, Nishihara H K 1978 Representation and recognition of the spatial organization of three shapes dimensional. *Proceedings of the Royal Society of London, Series B* (*Biological Sciences*) **200**: 269–94
Mozer M C 1991 *The Perception of Multiple Objects: A Connectionist Approach*. MIT Press/Bradford Books, Cambridge, MA
Olshausen B A, Anderson C H, van Essen D C 1993 A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience* **13**: 4700–19
Perrett D I, Oram M W 1998 Visual recognition based on temporal cortex cells: Viewer-centred processing of pattern configuration. *Zeitschrift fuer Naturforschung* (*C*) **5**: 518–41
Perrett D I, Oram M W, Hietanen J K, Benson P J 1999 Issues of representation in object vision. In: Farah , Ratcliff G (eds.) *The Neuropsychology of High-level Vision: Collected Tutorial Essays*. Erlbaum, Hillsdale, NJ, pp. 33–61
Pinker S 1984 Visual cognition: A introduction. *Cognition* **18**: 1–63
Poggio T, Edelman S 1990 A network that learns to recognize 3D objects. *Nature* **343**: 263–6
Poggio T, Shelton C R 1999 Machine learning, machine vision, and the brain. *AI Magazine* **20**: 37–55
Reisenhuber M, Poggio T 1999 Hierarchal models of object recognition in cortex. *Nature Neuroscience* **2**: 1019–25
Sandon P A, Uhr L M 1988 An adaptive model for viewpoint-invariant object recognition. In: *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ, pp. 209–15
Seibert M, Waxman A M 1990 Learning aspect graph representations from view sequences. In: Touretzky D S (ed.) *Advances in Neural Information Processing Systems*. Morgan Kaufmann, San Mateo, CA, Vol. 2, pp. 258–65
Tarr M J 1999 News on views: Pandemonium revisited. *Nature Neuroscience* **2**: 932–5
Tarr M J, Bülthoff H H 1995 Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance* **21**: 494–1505
Tarr M J, Bülthoff H H 1998 Image-based object recognition in man, monkey, and machine. *Cognition* **69**: 1–20
Ullman S 1989 Aligning pictorial descriptions: An approach to object recognition. *Cognition* **32**: 193–254
Wallis G, Rolls E T 1997 Invariant face and object recognition in the visual system. *Progress in Neurobiology* **51**: 167–94
Weinshall D, Edelman S, Bülthoff H H 1990 A self-organizing multiple-view representation of 3D objects. In: Touretzky D S (ed.) *Advances in Neural Information Processing Systems*. Morgan Kaufmann, San Mateo, CA, Vol. 2, pp. 274–81
Zelmel R S, Mozer M C, Hinton G E 1988 TRAFFIC: Object recognition using hierarchical reference frame transformations. In: Touretzky D (ed.) *Advances in Neural Information Processing Systems*. Morgan Kaufmann, San Mateo, CA, Vol. 2, pp. 266–73

M. C. Mozer

Objectivity, History of

Although the term 'objective' only came to have its modern sense in the eighteenth century, issues bound up with objectivity have been rife since seventeenth-century disputes over the respective merits of accounts based on facts and those relying on deductive argument (see Daston 1992, 1994). Objectivity stands in contrast to subjectivity: an objective account is one which attempts to capture the nature of the object studied in a way that does not depend on any features of the particular subject who studies it. An objective account is, in this sense, impartial, one which could ideally be accepted by any subject, because it does not draw on any assumptions, prejudices, or values of particular subjects. This feature of objective accounts means that disputes can be confined to the object studied. But since there must be some input from the subject in terms of theories, assumptions, etc., this is possible under two conditions: neutral assumptions and shared assumptions. The first is obviously preferable to the second (because the shared assumptions might be mistaken), but there are questions as to whether the first is possible only in a limited range of cases, whether it is possible at all, and whether the second is possible in all, some, or no cases.

Much recent literature in the social studies of scientific knowledge effectively denies that either of these is possible, and argues that if we are to understand how scientific projects come to be accepted, we need to examine not their purported objectivity or truth but the rhetorical strategies employed by their proponents (see the discussions in Fuller 1993, Golinski 1998, and the essays in Megill 1994; see also *Scientific Knowledge, Sociology of*). Concern with the nature and possibility of objectivity is not new, however, and by looking at some traditional questions that have been raised about objectivity, we can attempt to identify what is genuinely problematic about objectivity and what is not.

The nature of objectivity can best be brought out by considering challenges to it. There are three traditional

kinds of challenge that bring out some of the principal issues: the claim that objectivity is impossible; the claim that objectivity is always relative; and the idea that the form objectivity takes in the natural sciences might not be generalizable.

1. Is Everything Subjective?

The ancient Pyrrhonists denied that objectivity was possible (see Annas and Barnes 1985). Everything, they argued, is relative to the observer, to the conditions under which the observation is made, and to the state the object is in when it is perceived. For example, what color we observe something to be will depend upon us—it will look different depending on whether we have jaundice or not—upon the medium in which it is observed—it will depend upon whether we observe the object through air or red-tinted glass—and it will depend upon the state of the object—such as whether the object is in a solid form or a powder (gold is yellow when solid and white when powdered, for example). Observation is always inescapably relative to a host of factors and not only can we not avoid this relativity, but, the Pyrrhonists argued, we cannot single out any factor as being objective. There is no objective fact of the matter as to what color something is, or even whether it is moving or stationary, large or small: it depends on the circumstances of viewing and we have no grounds for preferring observations made by a stationary observer to those of a moving one, those of a close observer to those of a distant one, or even those of an intoxicated observer to those of a sober one.

Now even if we were to accept that there are never any optimal viewing conditions (remembering that these kinds of argument often work by choosing examples where there are no optimal viewing conditions, and then moving illegitimately to the claim that there are *never* optimal viewing conditions), we would not thereby be forced to dispense with objectivity. An objective theory can predict how a powder will look when viewing through different colored glasses, how something will look if we have jaundice, and it may predict color changes when a solid is ground into a powder; and we can certainly correlate changes in perceived speed with our state of motion, even our state of motion relative to one another. In other words, an objective theory is not one that simply chooses one set of viewing conditions over others and then stipulates that these will be the objective ones: it is one that holds for all viewing conditions, and any choice of one set over others does not affect the objectivity of the claims of the theory.

2. Is Everything Relative?

Any perceptual judgement involves, as well as input from the object perceived, various inputs from our perceptual and cognitive faculties, and we might reasonably expect this latter type of input to take on an increasingly significant role as we move away from straightforward observation. What we see and what we judge is, therefore, dependent on various perceptual-cum-cognitive interpretations that we place on our observations: there is no such thing as a 'view from nowhere' (see Nagel 1986). This does not rob these observations of objectivity. To think this is to assume that objectivity can be secured only in the absence of interpretation and judgment, but the question of objectivity does not even arise in the absence of these. Objectivity becomes an issue when we raise the question of the standing of our judgments and interpretations. We are not seeking a view from nowhere: we are seeking interpretations and judgments that hold up against all alternatives when judged by the same criteria. The fact that judgments involve conceptual structuring and interpretation (conceptual structuring and interpretation which might have been different) in itself has no bearing at all on the objectivity of those judgments.

The trouble is that when we move beyond this, to the question of theoretically motivated or theoretically guided interpretations of phenomena, differences in these interpretations may result in differences in what counts as evidence, differences in what counts as demonstration, and in what counts as a satisfactory explanation (see Kuhn 1970 and Gaukroger 1978). Here we do face problems of objectivity. In such cases, questions may well arise of whether, in not treating a particular kind of evidence as revealing, for example, one is being objective. Now failure of objectivity may be very difficult to decide in such cases, but we must not confuse failure of objectivity with differing standards of objectivity. Kuhn's work, in particular, has sometimes been read in such a way that the move from one paradigm to another is akin to a religious conversion rather than something objective (Kuhn himself used this analogy, but arguably as a way of capturing the phenomenology of the experience of moving from one paradigm to another, not as something that was intended to throw light on how changes to the content of science are made). There may indeed be paradigm shifts that reflect no objectivity and are like religious conversions, but the ones that Kuhn described, which are very much the success stories of the physical sciences, are clearly not of this type: the move from Aristotelian to Newtonian to Einsteinian cosmologies is a progression with a clear rationale in terms of objectivity, which a move in the opposite direction, for example, would not be.

Genuine cases of different standards of objectivity are difficult, if not impossible, to find, but those who have tried to establish a relativity of objectivity to the investigators' beliefs or theories or values have sometimes invoked two cases in twentieth-century physics—relativity theory and quantum mechanics—which seem to provide examples of the relativity of observa-

tional results with respect to the observer. In fact, they do not. In the relativity case, what Einstein showed in his Special Theory was that we have to alter our assessment of what is relative and what is fixed, but no increase in relativity, as it were, results from this. In classical physics, it had been assumed that how quickly a body is moving will appear differently to different observers: if one is traveling in the same direction as the body, for example, it will appear to be moving more slowly than if one passes it traveling in the opposite direction. In other words, velocity is relative to the observer. But the speed of light in a vacuum turned out to be constant no matter in what direction, or at what speed, the measuring apparatus was moving. Einstein argued that the only explanation for this was that the speed of light was constant, but that space and time contracted or dilated depending on the direction and speed of the observer. Space and time, which were previously absolute, are now made relative to the observer (but not merely relative: they really do—objectively—dilate and contract!), but the speed of light, which was previously relative to the observer, is now made absolute.

The case of quantum mechanics is rather more complex, but again we cannot draw any general conclusions about the relativity of observations from it. The quantum mechanics case arises because we cannot measure both the position and momentum of a particle: measurement of one affects the other. One early interpretation of this phenomenon, the Copenhagen Interpretation, maintained that particles were not in a physical state when not being measured: it was the act of measurement, which involved an interaction between a measuring apparatus and a measured system, that put them in a physical state. Moreover, in the Copenhagen Interpretation, this interaction is very problematic and peculiar because the measuring system obeys classical laws, whereas the measured system obeys quantum laws. The lesson that some have drawn from this is that measurement or observation affects the system observed, so we can never know what the system would be like without the observation or measurement, which is what we seek to know.

The peculiarities of the quantum mechanics case make it difficult to understand even in its own right, but this has not prevented some from trying to take it as a general model. Making a very contentious interpretation of something which is poorly understood into a source of guidance is clearly foolish. Nevertheless, does there not seem to be an analogy between this case and one that may arise in the observation of human behavior, for example, where the fact of being observed might inevitably and perhaps unconsciously lead one to modify one's behavior? The analogy breaks down, however, because the case of modification of behavior is something which we can develop techniques for dealing with, whereas this is explicitly not possible in the Copenhagen Interpretation. Such techniques have been around since antiquity, and were developed in the sixteenth century in the context of legal and historical testimony in a particularly interesting way (see Franklin 1963). Francesco Patrizzi, in his *Dialogues on History* of 1560, had attempted to show that the historian can either be impartial, or informed, but not both. He begins by rejecting secondary sources as virtual hearsay, and he divides primary sources into the partisan and the objective. Then, replying on a number of Machiavellian assumptions about the nature of rulers, he sets up a dichotomy between the partisan observer and the objective observer. Partisan observers—in this case, those sympathetic to the ruler—in virtue of being partisan, have access to the relevant information, because the ruler can rely upon them; but because they are partisan they will not provide an objective rendering of this information. Objective observers, namely those who are prepared to be critical of the ruler if this is merited, on the other hand, will not, because they are objective, have the ruler's confidence, and so will not have access to the source of the relevant information. (If this seems far-fetched, just think of the widespread practice of government briefings to selected journalists.) Patrizzi's conclusion is that it is 'utterly and totally impossible for human actions to be known as they were actually done.' The response from his contemporaries was that there are a number of ways in which we can establish credibility and plausibility, along the lines applied in law where, for example, we have doubts about a witness' credibility. In such cases we take into account such factors as the probative value of reinforcing testimonies: the problem is not one of a skeptical challenge to the objectivity of history *per se* but rather problems about evidence and reliability which practitioners of the discipline will usually be best placed to deal with. The same range of solutions is available to the anthropologist or psychologist, for example, who is concerned that the observation of his or her subjects causes them to modify their behavior or statements in some way.

3. Does Every Type of Explanation Call for Objectivity?

Objectivity seems to require us to stand back from the phenomena we are examining, as it were, so that we can investigate and assess them from a wholly impartial standpoint, which could be used for any phenomena of this kind. It is not always clear, however, that this approach is either the only one, or the most informative one. Consider the case of an anthropologist studying a rain dance. Assume that rain dances do not actually cause it to rain, and have never done so, yet are always performed at times of drought, and assume that there is evidence that, in times of drought, social strife and uncertainty about the nature of authority increases. In this case, it seems

reasonable to answer the question why the natives perform the rain dance in functionalist terms, namely that the rain dance secures social cohesion. But there is also a sense in which this is a wholly inappropriate explanation: the natives perform the rain dance only when they want it to rain, and their reason for performing it is clearly because they believe that this will increase the chances of rain.

There are three kinds of response one can make to this issue. The first is to insist that only the first explanation is objective and impartial. It requires us to stand back from the phenomena and treat them as we would any other: after all, functional explanations can be given for phenomena in our own societies—one can imagine a functionalist explanation of regular presidential or parliamentary elections as securing social cohesion, for example—as well as in others. The theory it works with is something that has been developed in such a way as to eradicate partiality and the imposition of the values of its advocates. Generally speaking, one crucial thing that we seek in scientific explanations is algorithmic compression; that is, particularly economical statements of the principles underlying data that enable us to generate that data and predict outcomes. The functionalist criterion meets this requirement in a way that the account which sticks to what the participants themselves believe does not.

The second response is to argue that the issue hinges on the fact that the functionalist account fails to—indeed cannot but fail to—capture the thought that motivates the participants in the dance. One way in which this contrast is sometimes expressed is in terms of the distinction between reasons and causes: giving the reasons someone has for doing something (or interpreting the behavior) and giving the causes of their behavior are two different things. Some philosophers—most notably Vico and Dilthey—have argued that the natural sciences and the *human sciences* (see *Human Sciences: History and Sociology*) are quite distinct (see Berlin 1976, Dilthey 1976 and Ermarth 1978) in this respect. In particular, while it might be appropriate to 'stand back' from phenomena in the natural sciences to achieve objectivity, this is inappropriate in the case of the human sciences, where we are dealing not with an objectified realm, but with human beings who have intentional states, emotions, exercise judgment, etc., things they share with the investigator and which the investigator is therefore in a position to interpret and make sense of: something quite different from what we do in the natural sciences. In this view, to take the functionalist road would be to treat the rain dance as if it were on a par with a natural, nonintentional phenomenon and it would thus fail to appreciate the nature of the phenomenon under investigation. Proponents of such a view are not denying that objectivity in the sense of impartiality is appropriate in the human sciences, but rather that criteria of objectivity, or guides as to how objectivity might be achieved, can simply be imported from the natural sciences, because what is at stake is not the discovery of causes but the interpretation of behavior. Similarly with a third response, which denies that there is just one way of explaining the rain dance: different accounts bring to light different aspects of the phenomenon. The success of an explanation on this account is relative to what one wants explained, and there is no reason why the choices should be restricted to two mutually exclusive accounts.

These disputes raise a deep question about just how significantly the means by which objectivity is to be secured can vary. But they do not raise a question about the nature of objectivity as the dispute is over how objectivity in the human sciences is best achieved, not over whether objectivity consists of something different in the natural and human sciences. What it consists of can be spelled out in operational terms: whether one is seeking to discover the causes of behavior or to interpret it, what we seek is an impartial account, one which could ideally be accepted by any subject on the basis of evidence and/or arguments.

4. *The Limits of Objectivity*

It is a hallmark of objectivity that one compares one's theory with the evidence, but sometimes it is difficult to distinguish between the evidence for one's theory and one's expectations about what a plausible theory should or will reveal. One always judges a theory against what one already knows or believes to be the case, and while in some cases the theory may change one's beliefs and expectations, in others it may lead one to challenge the theory by trying to develop a viable alternative, for example. There may simply be no objective matter of fact, in particular cases, about which direction to go in: one direction may be no more objective than another. In other words, there is nothing objective about accepting evidence at face value: sometimes it is the right thing to do, sometimes not. In the face of potentially falsifying evidence, for example, it is often better to revise one's theory, to accommodate the theory to the evidence, than to abandon it. Some philosophers have tried to capture the conditions under which it is no longer productive to continue to revise a theory (e.g., Lakatos 1972), while others have denied that this is possible (e.g., Feyerabend 1974).

What these disputes suggest is that guidelines marking out how we are to secure objectivity are much more difficult to formulate than in the heyday of theories of scientific method, when it was assumed that comparison of theories with the evidence was all there was to the matter. But this is, of course, quite different from the claim that the securing of scientific results is not a matter of objectivity.

See also: Content Validity; External Validity; History of Science; History of Science: Constructivist Per-

spectives; Internal Validity; Laboratory Studies: Historical Perspectives; Measurement Theory: History and Philosophy; Objectivity of Research: Ethical Aspects; Objectivity: Philosophical Aspects; Quality Control, Statistical: Reliability and Life Testing; Quantification in the History of the Social Sciences; Reliability: Measurement; Scientific Instrumentation, History and Sociology of; Scientific Knowledge, Sociology of

Bibliography

Annas J, Barnes J 1985 *The Modes of Scepticism*. Cambridge University Press, Cambridge, UK

Berlin I 1976 *Vico and Herder*. Chatto and Windus, London

Daston L 1992 Objectivity and the escape from perspective. *Social Studies of Science* **22**: 597–618

Daston L 1994 Baconian facts, academic civility, and the prehistory of objectivity. In: Megill A (ed.) *Rethinking Objectivity*. Duke University Press, Durham, NC

Dilthey W 1976 *Selected Writings* [ed. and trans. Rickman H P]. Cambridge University Press, Cambridge, UK

Ermarth M 1978 *Wilhelm Dilthey: The Critique of Historical Reason*. University of Chicago Press, Chicago

Feyerabend P 1974 Zahar on Einstein. *British Journal for the Philosophy of Science* **25**: 25–8

Franklin J H 1963 *Jean Bodin and the Sixteenth Century Revolution in the Methodology of Law and History*. Columbia University Press, New York

Fuller S 1993 *Philosophy, Rhetoric, and the End of Knowledge*. University of Wisconsin Press, Madison, WI

Gaukroger S 1978 *Explanatory Structures*. Harvester Press, Brighton, UK

Golinski J 1998 *Making Natural Knowledge*. Cambridge University Press, Cambridge, UK

Kuhn T S 1970 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Lakatos I 1972 Falsification and the methodology of scientific research programmes. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK

Megill A (ed.) 1994 *Rethinking Objectivity*. Duke University Press, Durham, NC

Nagel T 1986 *The View from Nowhere*. Oxford University Press, New York

S. Gaukroger

Objectivity of Research: Ethical Aspects

1. Introduction

This article elucidates the generally accepted idea that scientists should strive to be objective. It provides a definition of 'objective research,' describes the evolving debate concerning objectivity, examines the foundations of objectivity in science, and discusses some of the principles of conduct relating to objectivity, such as honesty and openness. In exploring this subject, the article addresses several topics related to objectivity, such as error and bias, peer review, cooperation and trust in science, conflicts of interest, social responsibility, and the moral limits of objectivity.

2. What is Objective Research?

The phrase 'objective research' has several different meanings in the social sciences and humanities. We often think of objective research as research that is *unbiased or impartial*. Thus, an objective researcher (or project, report or study) is like a judge who attempts to give a fair hearing to both sides of a legal dispute. The objective judge listens to both sides of a dispute without giving undue consideration to either. We also think of objective research as *value-free* i.e., as research that is unaffected by moral, economic, social, political, or religious values. Thus, an objective researcher is like a judge who attempts to make decisions based on legal and empirical evidence without allowing his values to affect his judgment and reasoning. We also think of objective research as *reliable or trustworthy*. Thus, an objective researcher is like a thermometer that reliably reports the temperature. We can trust the thermometer's report. Finally, we also think of objective research as research that is *factual or real*. Objective theories (hypotheses or concepts) are theories that correctly depict facts or real phenomena, and they can be contrasted with fiction, propaganda, speculation, and fraud. All of these conceptions of objective research capture some important aspects of our use of this phrase; there are a variety of ways of understanding objectivity (Harre and Krausz 1996). For the purposes of this article, objective research can be defined as research that is: (a) unbiased or impartial, (b) value-free, (c) reliable or trustworthy, or (d) factual or real.

3. The Evolving Debate about Objectivity

The notion that scientists have ethical duties to be objective is hardly a modern invention. Scientists and philosophers who laid the foundations for the modern scientific method, such as Francis Bacon, Rene Descartes, Galileo Galilei, and William Harvey, all viewed science as objective inquiry. Editors of the first scientific journals instituted peer review in order to publish articles based on objective evidence and reliable methods (LaFollette 1992). Charles Babbage (1970) was perhaps the first scientist to emphasize the ethical component of scientific objectivity in print. He coined the terms 'fudging,' 'cooking,' and 'trimming' to describe some of the unscrupulous research practices he observed in England in the 1800s. In the last half of the 1900s, powerful discoveries and inventions, such as atomic power, the computer, and genetic engineering, have compelled researchers to recognize the social impacts of science and technology and the

relationship between objective inquiry and moral responsibility (Shrader-Frechette 1994).

In the last two decades, politicians, the press, and the public have raised questions about scientific integrity and ethics as a result of highly publicized cases of misconduct in science and questionable research practices (Broad and Wade 1993). Some of these episodes include allegations of fraud, plagiarism, misappropriation of funds, and exploitation of subordinates, as well as controversies concerning secrecy, conflicts of interest, patents, and the funding of science. In response to these concerns, funding agencies, professional associations, and research institutions have adopted codes of ethical conduct in science that address issues relating to objectivity and other aspects of scientific integrity and ethics (Sigma Xi 1986, American Association for the Advancement of Science 1980, Panel on Scientific Responsibility and the Conduct of Research 1992, Committee on the Conduct of Science 1994).

While this debate about ethical duties concerning objectivity was taking place, many scholars and scientists began to question the idea that science is or can be objective. This controversy about the objectivity of science is also not new—the ancient Greek philosophers developed accounts of objective knowledge in response to sophistry and relativism—but its twentieth-century incarnation has spurred an extensive inquiry into the nature of scientific method and practice. Inspired by Thomas Kuhn's monumental work on the history of science, *The Structure of Scientific Revolutions* (1962), scholars from the social sciences and humanities have challenged the traditional view of science as objective knowledge. According to these scholars, otherwise known as 'constructivists' scientific research is strongly influenced by moral, economic, historical, social, political, or religious interests or values (Shapin and Shaffer 1985, Latour and Woolgar 1986). Defenders of the traditional view have responded to these critiques with their own arguments (Kitcher 1993, Gross and Levitt 1994). These debates about the objectivity of science—the so-called 'science wars'—have migrated from the halls of academia to the outside world, and they have received considerable attention from the press.

For the purposes of this article, this debate is relevant insofar as it has some bearing on the ethical aspects of objectivity in research. With this in mind, it is important to realize that strong versions of constructivism may be incompatible with the idea that researchers have ethical obligations to be objective, since it makes little sense to claim that a person has an ethical obligation that he cannot fulfill. If it is impossible to at least strive to be objective, then the claim that researchers have ethical duties to be objective is moot (Resnik 1998). While it is important to acknowledge and understand factors that can undermine or compromise objectivity in research, this article will assume that it is at least possible for researchers to strive to be objective, even if they often fall short of this ideal.

4. *Foundations of Research Ethics*

Ethical duties in science stem from two main sources: (a) the scientific profession itself, and (b) society at large (Resnik 1998). Ethical standards (or norms) promote the aims and goals of science in several ways. Many standards promote the advancement of human knowledge. For example, honesty in science is important in obtaining truth and avoiding error and deception. Other ethical standards play a key role in promoting cooperation and trust among scientists by creating institutional structures that direct individual interests toward common goals (Merton 1973). For example, authorship practices reward researchers for their efforts and ingenuity, insure that publications meet standards of quality and originality, and hold researchers accountable for their contributions.

Scientists also should adhere to ethical standards in order to promote social values, uphold the public's trust in science, and secure the public's support for research. For example, data fabrication can lead to harmful social consequences, especially if the fabricated results pertain to applied sciences like medicine or engineering. Data fabrication can also be construed as a form of lying, which is widely regarded as immoral. Data fabrication violates the public's trust in science because the public trusts that scientists will not fake data. Finally, fabrication can undermine the public's support for science if people become concerned about unethical conduct among researchers and decide to shift government funds to other programs, such as education or healthcare.

5. *Ethical Principles and Objectivity*

There are a variety of ethical principles (or standards) that promote objectivity in research. The principles discussed below should be viewed as general guidelines instead of absolute rules. They must be elaborated, assessed, and interpreted before they can be applied to particular situations or cases.

5.1 *Honesty*

The principle of honesty implies a general prohibition against falsifying, fabricating, or misrepresenting data, results, or other types of information pertaining to scientific publication. Honesty applies to a variety of other aspects of research, such as grant proposals, peer review, personnel actions, accounting and finance, expert testimony, informed consent, media relations, and public education. As mentioned previously,

honesty plays a key role in the search for knowledge and in promoting cooperation and trust among researchers. Few scientists or scholars dispute the importance of honesty and most people understand what it means to fabricate or falsify information pertaining to research. However, a few points of clarification will be helpful.

5.2 Error, Bias, and Disagreement

First, it is important to distinguish between dishonest research and error. Dishonesty implies the intent to deceive; error does not. Making an honest mistake in recording or analyzing data should not be equated to faking data. Although errors and dishonesty often have similar results, i.e., the deception of an audience, they arise from different motives. However, scientists also have an obligation to avoid making errors, since errors can also damage the quest for knowledge and undermine cooperation and trust among researchers. Although mistakes are, and always will be, a part of science—scientists are human beings, after all—excessive or egregious errors should be denounced as a form of negligence, since these errors have detrimental effects on research or its practical applications.

Scientists have an obligation to prevent errors by carefully scrutinizing and reviewing their own work and the work of others. Although some errors result from haste or sloppiness, more subtle follies are due to systematic errors (or biases) and self-deception. For example, since the 1800s, racial and ethnic biases have plagued research on human intelligence. More recently, self-deception (or simple carelessness) was probably at work in the cold fusion research conducted by Stanley Pons and Martin Fleischmann. Scientists also have a responsibility for reducing the effects of error by publishing errata, corrections, retractions, or apologies. Although the peer review system is not perfect, it can play an important role in reducing errors. Hence, reviewers and editors have ethical responsibilities pertaining to objectivity (LaFollette 1992).

It is also important to distinguish between dishonesty and disagreement. Researchers often have fundamental disagreements about methods, concepts, and theories. Scientists may disagree about research results (or their significance) without viewing their opponents as dishonest. Before an act can be judged to be dishonest, researchers must have some agreement on what constitutes the honest application of methods, concepts.

5.3 Misrepresentation of Data

The concept of 'misrepresentation,' unlike 'fabrication' and 'falsification,' is neither clear nor uncontroversial. Most scientists will agree that fabrication is making up data and falsification is changing data. But what does it mean to misrepresent data? As a minimal answer to this question, one can define 'misrepresentation of data' as 'communicating honestly reported data in a deceptive manner.' But what is deceptive communication? The use of statistics presents researchers with numerous opportunities to misrepresent data. For example, one might use a statistical technique, such as multiple regression or the analysis of variance, to make one's results appear more significant or convincing than they really are. Or one might eliminate (or trim) outliers when 'cleaning up' raw data. Other ways of misrepresenting data include drawing unwarranted inference from data, creating deceptive graphs of figures, and using suggestive language for rhetorical effect.

However, since researchers often disagree about the proper use of statistical techniques and other means of representing data, the line between misrepresentation of data and 'disagreement about research methods' is often blurry. Since 'misrepresentation' is difficult to define, many organizations have refused to characterize misrepresenting data as a form of scientific misconduct. On the other hand, it is important to call attention to the problem of misrepresenting data, if one is concerned about promoting objectivity in research, since many of science's errors and biases result from the misrepresentation of data.

5.4 Conflict of Interest

In order to promote objectivity in research, it is important for scientists to minimize the effects of financial, personal, professional, and political interests that may potentially undermine their ability to conduct and evaluate research. Conflict of interest has emerged as an important ethical concern in science in this century as many researchers now have significant financial interests in their work (Macrina 1995). Private support for research and development (R&D) has risen steadily since 1970. Today, industry provides over $100 billion in R&D funding in the United States alone, and it is expected that the trend toward the privatization of research will continue. Scientists working in the private sector receive remuneration in the form of salary, stock options, or royalties. Even researchers funded by governments or universities can have significant financial interests in the form of patents or copyrights. Although financial, and other, interests do not invalidate research—good research can still occur in troublesome circumstances—research that is conducted when conflicts exist is more likely to be biased or unreliable. For example, a scientist whose research is sponsored by a pharmaceutical company is more likely to publish results that favor the company than a scientist who is not working for the company. Even when research is not biased, the presence of a significant conflict can create the perception of a bias. For example, if people learn that a food company has sponsored research demonstrating the health benefits

of its fat free potato chip, they may believe that the research is biased, even if it is not. Conflicts of interest can arise in experimental design, publication, peer review, personnel decisions, expert testimony, media interviews, and other aspects of science (Resnik 1998).

How should scientists deal with conflicts of interest? First, scientists should disclose actual or potential conflicts of interest to those parties who may be affected by these conflicts. Most journals and funding organizations now require researchers to disclose sources of funding and potential conflicts of interest. Sometimes these disclosures are made available to the public when articles are published. Disclosure can help buffer the impact of bias by allowing reviewers, scientists, and the lay public to scrutinize and question research that may be affected by economic, social, or political interests. Second, although it is not possible to avoid all conflicts of interest, researchers have a duty to avoid those conflicts that cannot be corrected by peer review or may have a dramatic impact on science or society. For example, researchers should not review a grant application if the applicant is a close friend or bitter enemy. The decision to avoid a potential or actual conflict of interest, also known as 'recusal,' can promote objectivity by removing a researcher from a situation that could create a bias or the perception of bias.

5.5 Openness

Since openness provides scientists with a reliable defense against error, bias, self-deception, dishonesty, and other phenomena that can threaten objectivity, it is an important ethical principle in research. Openness promotes objectivity in several ways. First, it enables researchers to criticize each other's work. Criticism can occur before, during, or after peer review. Frequently, the most important criticism takes place before an idea appears in print. Second, openness helps researchers make effective use of human and technological resources. Problems can be solved more effectively when researchers cooperate with each other than when they work in isolation (Merton 1973).

Although openness is an important ethical principle in research, it often conflicts with personal or institutional interests. Sometimes researchers refuse to share data in order to protect scientific priority or intellectual property rights. For example, a researcher who is developing a new product may decide not to share preliminary data until they have applied for a patent. A researcher who discovers a new phenomenon may not disclose the discovery until the paper has been accepted for publication. Sometimes researchers refuse to share data in order to protect their reputations. For example, a scientist may decide to not share preliminary results from a study until they have had enough time to verify and analyze those results. Researchers may refuse to share data because they have signed a contract with a corporate sponsor not to disclose data without prior authorization. Since many different parties—government, industry, colleagues, the media—may seek to obtain scientific information, it is likely that researchers will often face difficult choices when deciding how and when to share information.

6. The Moral Limits of Objectivity

While objectivity is one of science's most important principles, it should be acknowledged that this norm sometimes conflicts with other ethical duties, such as the obligation to protect human subjects and social responsibility (Shrader-Frechette 1994). For example, anthropologists may refuse to publish or disclose information in order to protect and promote the welfare of the human populations they are studying. In psychology, researchers have a duty to keep information about human subjects confidential in order to protect their privacy, dignity, and autonomy. An ecologist who is interviewed by the media must choose between their desire to advocate for the environment and their scientific duties to remain objective. Although scientists should strive to be objective, there are some situations where scientists should relax their commitment to objectivity in order to promote other moral and social values.

See also: Ethical Dilemmas: Research and Treatment Priorities; Objectivity, History of; Objectivity: Philosophical Aspects; Research Conduct: Ethical Codes; Research Ethics, Cross-cultural Dimensions of; Research Ethics: Research

Bibliography

American Association for the Advancement of Science 1980 *Principles of Scientific Freedom and Responsibility*. American Association for the Advancement of Science, Washington, DC

Babbage C 1970 *Reflections on the Decline of Science in England*. Augustus Kelley, New York

Broad W, Wade N 1993 *Betrayers of Truth*, 2nd edn. Simon and Schuster, New York

Committee on the Conduct of Science 1994 *On Being a Scientist*, 2nd edn. National Academy Press, Washington, DC

Gross P, Levitt N 1994 *Higher Superstition*. Johns Hopkins University Press, Baltimore, MD

Harre R, Krausz M 1996 *Varieties of Relativism*. Blackwell, Oxford, UK

Kitcher P 1993 *The Advancement of Science*. Oxford University Press, Oxford, UK

Kuhn T 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

LaFollette M 1992 *Stealing into Print*. University of California Press, Berkeley, CA

Latour B, Woolgar S 1986 *Laboratory Life*, 2nd edn. Princeton University Press, Princeton, NJ

Macrina F 1995 *Scientific Integrity*. American Society of Microbiology, Washington, DC

Merton R 1973 *The Sociology of Science*. University of Chicago Press, Chicago
Panel on Scientific Responsibility and the Conduct of Research 1992 *Responsible Science*. National Academy Press, Washington, DC, Vol. 1
Resnik D 1998 *The Ethics of Science*. Routledge, London
Shapin S, Shaffer S 1985 *Leviathan and the Airpump*. Princeton University Press, Princeton, NJ
Shrader-Frechette K 1994 *The Ethics of Scientific Research*. Rowman and Littlefield, Boston
Sigma Xi 1986 *Honor in Science*. Sigma Xi, Research Triangle Park, NC

D. B. Resnik

Objectivity: Philosophical Aspects

There are two main kinds of philosophical questions about objectivity: metaphysical and epistemological. Metaphysical objectivity concerns the extent to which the existence and character of some class of entities depends on the states of mind of persons (i.e., their knowledge, judgment, belief, perception, or response). Epistemological objectivity concerns the extent to which we are capable of achieving knowledge about those things that are metaphysically objective. Many philosophers working in the Anglo-American traditions also worry about semantic objectivity, that is, about whether or not the propositions in some realm of discourse (physics, psychology, ethics, etc.) can be evaluated in terms of their truth or falsity. For a discourse to be semantically objective, and for the statements in the discourse to be true, then the things referred to by the terms of that discourse (e.g., quarks, desires, justice, etc.) must be metaphysically objective.

1. *Metaphysical Objectivity*

An entity (or a class of entities) is metaphysically objective if its existence and character is independent of the human mind. This 'independence requirement' is central to metaphysical objectivity (Brower 1993, Sober 1982), though its proper interpretation raises two important questions: first, in what way must a metaphysically objective thing be 'independent' of the human mind; and second, how much independence of the relevant kind is required?

1.1 *What Kind of Independence?*

The existence and character of some entity might be independent of the human mind in three senses: causally, constitutionally, and cognitively. Only the last two will matter for metaphysical objectivity.

An entity is causally independent of the human mind as long as the causal trajectory producing it did not involve the human mind. Shoes, for example, are causally dependent on the human mind because the existence and character of any particular pair of shoes depends causally on a cobbler having had certain beliefs and desires (e.g., a desire to make a particular kind of shoe, and true beliefs about what needed to be done to produce such shoes). By contrast, the existence and character of the earth is causally independent of the human mind: no human intentions played a causal role in bringing about the existence of the earth or its specific character. Metaphysical objectivity, however, does not require causal independence. Even entities that are causally dependent on the human mind can be mind-independent in one of the other two senses (below), and thus still be metaphysically objective.

An entity is constitutionally independent of the human mind if its existence and character is not constituted by or identical with the mind. Certain historical forms of philosophical 'idealism' (such as those of Bishop Berkeley and Hegel) held that the world was constitutionally dependent on the mind (the human mind, or perhaps the mind of God). Conversely, the claim that some entity is metaphysically objective almost always involves denying its constitutional dependence on the mind. The exception is for psychological entities (e.g., beliefs, desires, emotions, etc.): such things cannot be constitutionally independent of the mind since they just are facets of the mind. Yet surely psychological facts may also be metaphysically objective. If so, it must be in the final sense of 'independence' from the mind.

An entity is cognitively independent of the human mind if its existence and character does not depend on any cognizing state of persons, for example, belief, sensory perception, judgment, response, etc. (A 'cognizing' state is one which is receptive to features of the world and thus is a potential source of knowledge of the world.) A metaphysically objective thing is, accordingly, what it is independent of what anyone believes or would be justified in believing about it (or what anyone perceives it to be or would perceive it to be under certain conditions, etc.). On this account, psychological facts about a person are metaphysically objective in virtue of not depending on what an observer of that person believes or would be justified in believing about that person's psychological state. (This assumes that mental content is 'narrow,' not 'wide,' a technical debate in the philosophy of psychology that must be set aside here.)

Any kind of metaphysically objective fact (except for psychological facts) must necessarily be constitutionally independent of the mind. All metaphysically objective facts must also be cognitively independent. The common-sense picture of the natural world presumes that its contents are metaphysically objective in this sense: ordinary people think that atoms and zebras and sulfur are not simply identical with the

mind and that they are what they are independent of what people may believe or be justified in believing about them. Science, then, aspires to epistemological objectivity by trying to accurately depict the way things (objectively) are.

1.2 How Much Independence?

There can be degrees of cognitive independence, and thus degrees of objectivity that may be distinguished; not everything that is objective may prove to be objective in the sense in which common sense understands the constituent elements of the natural world to be objective. The crucial notion for cognitive independence is independence from the cognizing states of persons: beliefs, sensory perceptions, judgments, responses, and the like. Thus, this notion of objectivity supposes that there is always a difference between what 'seems right' about some state of affairs and what actually 'is right.' For example, it may seem right to John (based on a sensory perception) that there is a table in front of him, but it may not be right that there is a table there: it could be an optical illusion. The table, then, is objective in some sense since its existence is independent of what 'seems right' to John.

It is possible, then, to distinguish four kinds of claims about objectivity. According to subjectivism, what seems right to the cognizer determines what is right. According to minimal objectivism, what seems right to the community of cognizers determines what is right. According to modest objectivism, what seems right to cognizers under appropriate or ideal conditions determines what is right. According to strong objectivism, what seems right to cognizers never determines what is right. Subjectivism and strong objectivism represent the two classical and opposed philosophical positions of antiquity: Protagoras held that 'man is the measure of all things' (subjectivism) (Plato, *Theaetetus* *152a, *166a–*8b), while Plato embraced a kind of strong objectivism (Plato, *Phaedo* *741–*75b, *Republic* *475–*80, *508d-e). The Protagorean position denies the objectivity of the world and everything in it: whatever each individual takes to be the case is the case (for that individual), and thus the existence and character of any particular thing depends (epistemically) on the (individual) human mind. By contrast, the Platonist affirms the complete and absolute objectivity of the world: What really is the case about the world is never fixed by what any person or all persons believe, has (or have) reasons to believe, or could have reasons to believe. Mistake, on a global scale, even under ideal epistemic conditions, is a possibility for the Platonist. This latter position is often described as 'realism' (or 'metaphysical realism').

Minimal objectivism and modest objectivism occupy conceptual space between these two familiar, historical positions. Minimal objectivism holds that whatever the community of cognizers takes to the case is the case. This view, like its pure Protagorean cousin, issues in a kind of relativism (what is the case is relative to a particular community of cognizers), but by abstracting away from the subjectivity of the individual cognizer, it introduces a minimum amount of objectivity. It is also a kind of objectivity with some useful domains of application. What is and is not fashionable, for example, is probably minimally objective. What seems right to John about what is fashionable can be objectively wrong: John may be out of sync with the styles of his community, and thus it would be correct to say, 'John is mistaken in thinking that a plaid shirt and striped pants go well together.' But it does not seem that the entire community can be wrong about what is fashionable: in that sense, what seems right to the community determines what really is fashionable.

In most domains, however, minimal objectivity would be viewed as too close to subjectivism for comfort. Modest objectivity thus abstracts even further away from dependence on actual cognizers, individual or communal. Something is modestly objective if its existence and character depends only on what cognizers would believe under certain idealized conditions for cognition, conditions like full information and evidence, perfect rationality, and the like. (By hypothesis, under ideal conditions, all cognizers would come to the same belief about things.) Everyone on the planet can be wrong in his or her beliefs about a modestly objective entity; beliefs formed under ideal epistemic conditions, however, can never be wrong. This latter point is what differentiates modest from strong objectivity.

Some philosophers have defended the idea that the truth about the world is at best modestly objective (e.g., the doctrine Putnam calls 'Internal Realism' in Putnam 1981): what is true in any domain is simply whatever inquirers would agree upon under epistemically ideal conditions. Versions of this idea have since been subjected to withering criticisms (e.g., Johnston 1993), and it has few adherents. Still, modest objectivity, like the other conceptions, may be particularly apt in certain domains. Consider, for example, facts about color. It seems natural to say that there are (modestly) objective facts about color even though the color of an object is not fully independent of the human mind. For example, we might say that something is red if and only if normal perceivers under normal viewing conditions would be disposed to see it as red. Color, on this account, depends on the human mind—on human response or perception—but only on human response under appropriate conditions. One important recent idea is that evaluative facts might be modestly objective in a similar way (e.g., Pettit 1991): evaluative facts would depend on human responses to morally significant situations under appropriate conditions. In both cases, however, it is important to specify the conditions under which human response fixes the reference of a concept in a

nonquestion begging way. It obviously will not do, for example, to define 'normalcy' of perceivers and of viewing conditions by reference to their getting the right result (i.e., seeing all and only the red things as red). Some philosophers doubt whether the conditions can be specified in a nonquestion-begging way (e.g., Wright 1992).

2. *Epistemological Objectivity*

The demand for epistemological objectivity is the demand to be free of bias and other factors that distort cognition, that prevent the things being cognized from presenting themselves as they really (metaphysically) are. More precisely, epistemological objectivity requires that the cognitive processes and mechanisms by which beliefs about the world are formed be constituted in such a way that they at least tend toward the production of accurate representations of how things are. Notice that epistemological objectivity does not require that cognitive processes always yield true representations: that would demand more than is attainable and more than is even expected. Epistemological objectivity obtains when either of the following is true: (a) the cognitive processes at issue reliably arrive at accurate representations, or (b) the cognitive processes are free of factors that are known to produce inaccurate representations. The obstacles to epistemological objectivity will vary with the domain under consideration. In the sciences generally, and the social sciences in particular, 'values' are often thought to present a special obstacle to epistemological objectivity insofar as they influence the choice of research topics and, most seriously, the selection and evaluation of evidence (Weber 1949). As one contemporary author explains: 'Values are thought to damage inquiry epistemically because they are held to be subjective—they come from us, not the world. Therefore, to allow values to influence inquiry into the nature of the world is to allow such inquiry to be subject to a control other than the world itself' (Railton 1985, p. 818).

Epistemic values or norms, for example, norms about when evidence warrants belief, must of course play a role in all scientific inquiry; the worry is about nonepistemic values or norms, like the political ideology of the inquirer or the political climate in which inquiry takes place. Yet even here the effect on objectivity depends on the context and the precise role values plays in inquiry. For example, where the political climate is hostile to the discovery of certain kinds of objective truths, only those inquirers motivated by a contrarian political ideology themselves are likely to pursue those objective truths (Miller 1991). Here, of course, values supply a motive for allowing 'the nature of the world' to control the course of inquiry, rather than interceding and distorting cognition.

More generally, one 'need not require freedom from all value and bias in order to have objective inquiry' since 'there may yet exist mechanisms of belief-formation that incorporate feedback from the object to the inquiring subject' (Railton 1985, p. 818). The mechanism at issue is causal: metaphysically objective things make themselves felt causally, whatever our theoretical preconceptions and values. This, however, is merely an external criterion for objectivity, and does not yet show how inquirers could determine whether or not their inquiry is epistemologically objective. Here, however, inquirers might look for certain familiar markers of epistemological objectivity, like the existence of intersubjective agreement in judgment, the publicity of evidence and standards of proof, and the reproducibility of the evidence for a judgment by different inquirers: 'when these conditions are met, subjective contributions and biases will be excluded as far as possible' (Railton 1985, p. 818). That physics constitutes a cross-cultural, global community of inquirers strongly suggests that it is epistemologically objective: if it were not, then one would expect local differences (in interests, ideology, and the like) to lead to markedly different discourses of physics. Of course, the absence of intersubjective agreement does not by itself demonstrate lack of epistemological objectivity; the question is always what the best explanation for lack of such agreement is supposed to be. In the case of the social sciences, where objective truths may conflict with entrenched interests, it should hardly be surprising that there should be no agreement about certain social scientific questions. In other cases, though, the suspicion will be strong that it is entrenched interests and values that distort cognition of the social world and warp scientific inquiry accordingly.

3. *Semantic Objectivity*

Semantic objectivity is a property of statements, rather than things or cognitive mechanisms. Philosophers in the Anglo-American traditions of the twentieth century, who approach most philosophical problems by framing them first as problems about language and its relationship to the world, have been most concerned with questions of semantic objectivity. Typically, philosophers are concerned with a class of statements characteristic of a particular branch of discourse: say, physics or psychology or ethics or aesthetics. Some branch of discourse enjoys semantic objectivity when its statements are generally apt for evaluation in terms of their truth and falsity. (Not every statement in the discourse need be determinately true or false—the property of 'bivalence'—since few discourses outside pure mathematics are bivalent.) Cognitivism is the doctrine that some branch of discourse is semantically objective. (Wright 1992 disputes whether truth-aptness is the relevant criterion for demarcating objective

discourses; most discourses satisfy the minimum syntactic requirements to make use of the truth-predicate appropriate. The issue about the objectivity of the discourse must be located elsewhere according to Wright.)

Thus, for example, the discourse of natural science is presumed to be a cognitive discourse: scientific statements about the natural world are generally either true or false. But what of statements in ethics like, 'That distribution of resources is unjust,' or 'Harming a defenseless animal is morally wrong?' Many philosophers have thought that there are, as a metaphysical matter, no facts in the world corresponding to the 'injustice' of a distribution or the 'moral wrongness' of an action (Gibbard 1990, Mackie 1977, Stevenson 1944). Most philosophers who deny the metaphysical objectivity of morality claim that its semantics is noncognitive: rather than stating facts (that are either true or false), ethical statements, according to noncognitivism, express attitudes or feelings of various kinds (Stevenson 1944, Gibbard 1990 are the most sophisticated versions of this view). Noncognitivists bear the burden, then, of explaining away the surface grammar and logical structure of ethical discourse which makes it indistinguishable from ordinary empirical discourse (compare: 'This distribution is unjust,' with 'That chair is red').

A minority of philosophers, however, agree that morality is not metaphysically objective, but nonetheless maintain that the surface grammar of ethical discourse should be taken at face value: ethical discourse purports to state facts, and is thus a cognitive discourse. It is, unfortunately, a cognitive discourse almost all of whose statements are false (since there are no metaphysically objective moral facts in the world). (The only true ethical statements, on this view, will be negations of ethical judgments: 'No, slavery is not really morally wrong.') This doctrine is known as 'error theory' (Mackie 1977). Error theories about any discourse make it puzzling, however, why the discourse should persist, let alone occupy the central role that ethical discourse does in human lives: why would people continue to engage in a putatively fact-stating discourse that never succeeds in stating any facts? Noncognitivism, as a semantic doctrine, at least identifies an important role for ethical discourse: namely, the expression of feelings and attitudes about matters of real moment to human beings and their social existence.

Most philosophers who accept that ethical discourse is cognitive do so because they also believe that morality is metaphysically objective (in some sense) (Brink 1989, Brower 1993, Pettit 1991, Railton 1986): if there are metaphysically objective moral facts, then moral statements will not be systematically false as error theory has it. How could there be such facts? For those who think moral facts are strongly objective (e.g., Brink 1989, Railton 1986), the central idea is that moral facts are simply identical with (or supervenient upon) natural facts. For example, perhaps the property of being 'morally right' is just identical with the property of 'maximizing human well-being,' where the latter may be understood in purely psychological and physiological terms. In that case, whether an action X is morally right is a strongly objective matter, since it is simply a scientific question whether action X will in fact maximize the relevant kinds of psychological and physiological states in the world. The crucial claim, plainly, is that moral facts are to be identified with certain kinds of natural facts. Again, many philosophers are skeptical that this claim can be defended (e.g., Gibbard 1990).

4. *Other Directions*

A minority of philosophers recently have disputed whether the traditional ways of conceptualizing objectivity are adequate (Dworkin 1996, McDowell 1997, Nagel 1997). In particular, these philosophers have raised two kinds of doubts about the preceding kinds of characterizations of objectivity. First, these philosophers (especially McDowell) question whether the conception of metaphysical objectivity (especially strong metaphysical objectivity) does not presuppose a vantage point on the way things 'really' are to which we can have no access. Second, these philosophers wonder whether the conception of metaphysical objectivity as mind-independent is not a paradigm too closely tied to a picture of the objectivity of the natural world, and thus either does not make sense or should not be applied with respect to the objectivity of domains like ethics or aesthetics. In these evaluative domains, it makes no sense to ask about whether there are evaluative facts 'out there' 'in the world.' The objectivity of evaluative discourse is simply a matter of its susceptibility to reasons, of our ability to subject ethical positions to rational scrutiny and discussion. These revisionary views of objectivity have not, to date, found much favor among philosophers, and they have also been extensively criticized (Wright 1992, Leiter 2001, Svavarsdóttir 2001).

See also: Aristotle (384–322 BC); Norms; Norms in Science; Objectivity, History of; Relativism, Pluralism, and Skepticism (Philosophical Perspective): Cultural Concerns; Relativism: Cognitive; Relativism: Philosophical Aspects; Truth and Credibility: Science and the Social Study of Science; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Brink D O 1989 *Moral Realism and the Foundations of Ethics*. Cambridge University Press, Cambridge, UK

Brower B W 1993 Dispositional ethical realism. *Ethics* **103**: 221–49

Dworkin R 1996 Objectivity and truth: You'd better believe it. *Philosophy and Public Affairs* **25**: 87–139
Gibbard A 1990 *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Harvard University Press, Cambridge, MA
Johnston M 1993 Objectivity refigured: Pragmatism without verificationism. In: Haldane J, Wright C (eds.) *Reality, Representation and Projection*. Oxford University Press, New York
Leiter B 2001 Objectivity, morality, and adjudication. In: Leiter B (ed.) *Objectivity in Law and Morals*. Cambridge University Press, New York
Mackie J L 1977 *Ethics: Inventing Right and Wrong*. Penguin, Harmondsworth, UK
McDowell J 1997 Projection and truth in ethics. In: Darwall S, Gibbard A, Railton P (eds.) *Moral Discourse and Practice: Some Philosophical Approaches*. Oxford University Press, New York
Miller R 1991 Fact and method in the social sciences. In: Boyd R, Gasper P, Trout J (eds.) *The Philosophy of Science*. MIT, Cambridge, MA
Nagel T 1997 *The Last Word*. Oxford University Press, New York
Pettit P 1991 Realism and response-dependence. *Mind* **100**: 587–626
Putnam H 1981 *Reason, Truth and History*. Cambridge University Press, Cambridge, UK
Railton P 1985 Marx and the objectivity of science. In: Suppe F, Asquith P (eds.) *PSA* 1984, Vol. 2. Philosophy of Science Assoc., Lansing. Also in: Boyd R, Gasper P, Trout J (eds.) *The Philosophy of Science*. MIT , Cambridge, MA
Railton P 1986 Moral realism. *Philosophical Review* **95**: 163–207
Sober E 1982 Realism and independence. *Noûs* **16**: 369–85
Stevenson C L 1944 *Ethics and Language*, 1st AMS edn. Yale University Press, New Haven, CT
Svavarsdóttir S 2001 Objective values: Does metaethics rest on a mistake? In: Leiter B (ed.) *Objectivity in Law and Morals*. Cambridge University Press, New York.
Weber M 1949 *The Methodology of the Social Sciences*. Free Press, Glencoe, IL
Wright C 1992 *Truth and Objectivity*. Harvard University Press, Cambridge, MA

B. Leiter

Objects: Material

Material objects are defined in this article as handy, portable parts of the inanimate environment (Gibson 1979) that are always cultural artifacts, either by fabrication or by appropriation of natural objects. The perception and use of objects is shaped both by the object's *affordances* and the individual's motivational state. Affordances are either primarily practical or primarily symbolic. *Instruments* are used to physically manipulate the environment, while *symbols* are used to refer to something. Following from their physical uses, instruments may also symbolize the activity or end they serve (a hammer symbolizes hammering). Finally, instruments may also have *connotations* (a car may evoke power, speed, or wealth).

Objects are thus not simply physical entities of a certain size and weight. They are products of culture. Culture endows them with instrumental and symbolic values, which are in turn elaborated on by individuals. Entire disciplines (e.g., archeology) and institutions (e.g., museums) are concerned with the collection and interpretation of objects. Theories of the meanings and functions of objects have been developed by cultural anthropology, sociology, and psychology.

1. Cultural Artifacts

1.1 Systems of Objects

In the 1960s, *semiotic* approaches (see *Semiotics*) described systems of objects that constitute *codes*, that is systems of signs, comparable to language. These codes were stipulated to be autonomous from the social and economic systems. Roland Barthes (1964) analyzed the system of clothing fashion. He differentiated three structural levels of fashion: technological (textile structure), iconological (photographic presentation), and verbal (writing about fashion). Other semiotic analyses concerned everyday classifications of objects in museums or department stores (see *Semiotics*). In a related vein, structural anthropologist Lévy-Strauss analyzed systems of objects, for example *food* or *clothing*, as binary codes that establish basic cultural categories and determine cultural practices (see *Structuralism*).

Semiotic and structural approaches describe systems of material signs or categories synchronically. Systems of objects may also be described in the same formal manner diachronically. Richardson and Kroeber (1940), for example, identified a cyclical pattern in the change of formal properties of female dress across centuries. Barthes concluded from this that semiotic systems have a temporal dimension, but not a historical dimension. However, in 1982 Lowe and Lowe continued Kroeber's analysis in time and demonstrated that historical changes were not cyclical and were influenced by particular historical and social changes.

Others constructed a diachronic perspective on systems of objects by using the metaphor of *evolution*. George Kubler (1972) constructed the history of *style* in *art* as that of successive solutions to formal problems, an idea that he extended to instruments such as ploughs. Sigfried Giedion (1948) described the history of instruments in terms of constitutive and transitory facts, alluding to the evolutionary dichotomy of geno- and phenotypes. The evolution of Mickey Mouse and of teddy bears during the twentieth century have been analyzed in terms of maximizing congruence with the innate preference for infantile forms. In the quasi-evolution of forms of cultural

artifacts, however, environmental demands such as aesthetic preferences and utility compete with each other, as demonstrated by David Norman (1988) with regard to the evolution of the telephone.

1.2 Objects Serve Traditions and Transfer Cultural Knowledge

Objects made of durable material may survive several generations. They lend themselves to passing on information to succeeding generations. Traditions are maintained by objects that explicitly symbolize continuity with the past, such as monuments, museum exhibits, old buildings, and holy objects. Graveyards, tombstones, and relics in churches mark the interface between public tradition and private remembering, for which, among others, family heirlooms, private relics, and pictures are used.

Instruments pass on cultural knowledge implicitly by inviting individuals to use them in specific ways. Similar to texts that complement individuals' declarative memory, objects serve to complement individuals' and cultures' procedural memory for how to do things (Norman 1988, see *Cognition, Distributed*). Objects also embody cultural imperatives, so that the emergence of new cultural *norms* is often accompanied by the invention and social diffusion of new instruments. Norbert Elias used the historical introduction of forks to illustrate the embodiment of norms in instruments in the 'process of *civilization*.' Forks embodied new table manners which forbid eating with the hands.

Paleontologist André Leroi-Gourhan (1964) saw the evolution of human technology as the substitution of humans' corporal and mental functions with external aids. In Lev Vygotsky's theory of the ontogenesis of higher mental functions, this phylogenetic exteriorization of physical and mental functions is complemented by the ontogenetic internalization of cultural symbols and instruments. However, also in ontogenesis the external and later mental use of symbols is followed by an increasing use of external means for thinking.

1.3 Social Status in Traditional and Modern Societies

Mary Douglas (1973) saw ritual objects such as uniforms, marital rings, or graduation gown as constituting culture by providing an intelligible universe and as serving the preservation of social structure (see *Ritual*). With the deritualization of modern societies, ritual objects have increasingly been substituted with consumer goods. They are less durable and serve consumption, and access to goods is less regulated than access to ritual objects. Conspicuous consumption was described by Theodore Veblen (1899) as a normative technique for claiming social status. The more rare, the more expensive, the more useless the good, the more luxury successfully transfers status to the individual. Pierre Bourdieu (1979) introduced the concept of habitus as a set of aesthetic preferences that defines a lifestyle and serves to create social distinction. The habitus is not under the individual's control, but is socialized early in life. Therefore parvenues and impoverished gentry can be identified by tastes and postures incongruent with their economic situation.

1.4 Individualization of Natural or Fabricated Objects

Georg Simmel (1923) hypothesized that increasingly mass-fabricated and commodified objects could no longer be appropriated by the individual or used for self-cultivation, since they were more and more determined by dynamics of production and the market. In 1948 Sigfried Giedion claimed that man was increasingly determined by machines. In 1968 Jean Baudrillard lamented that goods were no longer sold for their utilitarian value, but only for connotations produced by advertisement. These critics seem nostalgic of a traditional past in which individual objects were produced locally and were exchanged rather than traded. To these critics it appeared that in premarket societies objects had been embedded in, and were expressive of personal relationships. In contrast, modern national and global markets of mass-produced goods created anonymous, estranged relationships between individuals and between individuals and objects.

One example of the embeddedness of objects in personal relationships is the socially structured handling and shared consumption of food, drink, and drugs. Dining together (*comensality*) establishes a bond between participants (Fischler 1990). In contrast to objects that are consumed, and that therefore exist only as uncountables, gifts are usually singular, individual objects that persist. Marcel Mauss (1924) analyzed gift exchange in traditional societies as a system of reciprocal obligations that structure a group and increase its cohesion.

Since the 1980s, cultural anthropology has become less nostalgic for exchange and has focused on commodities and modern consumption as media that provide and express cultural categories. Daniel Miller (1995) even suggests that commodities and consumption patterns will substitute for kinship patterns as the key concept in cultural anthropology. To be exchanged as impersonal commodities, objects need to be comparable to other goods on a scale of value (*money*). Even modern societies do not consider all objects to be commodities. These inalienable possessions (Weiner 1992) include royal insignia, holy objects, ritual objects, gifts, public territories, and human organs. Objects may acquire and lose the status of commodity by processes of diversification

and encapsulation (Appadurai 1986). Encapsulation may concern goods that are collected, such as stamps, or foreign or old instruments displayed as exotica or antiques. Encapsulation is usually accompanied by individualization of the object, for instance by leaving traces of use or by giving the object a personal name (Rogan 1990).

Objects may also become endowed with private meanings that are shared only by some, or even only one, member of a culture. Objects may take on metaphorical and metonymical meanings. A lock of a loved one's hair, a cinema-ticket from a couple's first date, a pen that belonged to one's grandmother all acquire additional metonymical meanings by their association with specific events or persons. These meanings are not easily accessible to outsiders. Objects always have cultural meanings, to which private meanings may be added, increasing their individuality.

2. *Personal Objects*

Objects that have been individualized and appropriated often gain special significance for the individual (Csikszentmihaly and Rochberg-Halton 1981). Objects to which an individual has become attached have been termed personal objects (Boesch 1991). In a general sense, personal objects are used to increase a sense of *self* or *identity*. Studies of uses of personal objects show that most objects tend to be used either instrumentally or symbolically. The most frequent instrumental uses are those of regulating levels of activity (stimulating, helping to relax), the most frequent symbolic uses those of reminding (Dittmar 1992, Habermas 1999). Other symbolic uses are self-presentational and self-communicative.

2.1 *Possessions and Instrumental Uses*

William James defined the self's extension by all that an individual could experience as, and call, her or his own. This material self may include family, personal possessions, and any other objects or persons the individual may identify with, and for which they may show self-related emotions such as pride or fear. Although possessions may contribute to self-definition, James' proposal does not specify what an individual will identify with.

Objects may contribute to the self-dimension of a sense of mastery or self-efficiency by offering a familiar instrument that increases an individual's ability to physically manipulate the environment, or by offering stimulation through a structure that is complex and novel enough to evoke interest and curiosity, but simple and familiar enough to be explorable. Objects may contribute to identification with one's body by extending the body's power (instruments, vehicles) or by accentuating and decorating it (clothes).

2.2 *Self-presentation*

Objects used to symbolize identity and to present the self favorably are usually located in the personal space (on or near the body) or in personal places. Living rooms, offices, houses, and vehicles are preferred showrooms of identity. Low status individuals with little or no personal place, like homeless or adolescents, have to use their bodies as a stage for marking their identities. Bodies may be marked by accessories, such as clothes, jewelry, or glasses, by modifying the body through tattoos or haircuts, or by modifying body-shape through dieting, body-building or exercise. Also most instruments are closely linked to the body and used to symbolize the individual's profession (e.g., a physician's stethoscope).

The use of objects to symbolize identity allows for the rapid, economical formation of expectations which may inform the decision of whether and how to interact with an approaching other. What is symbolized ranges from very general social categories such as age, gender, or social class, to professional, ideological, or idiosyncratic identities. Identity-statements may range from taking a stance towards a given identity such as gender or ethnicity, to claiming membership, individuality, or personal worth.

Situations that invite the self-presentational use of objects are ones of heightened self-awareness such as when being observed or when in an unfamiliar context. Individuals with a disposition for public self-consciousness prefer personal objects that serve self-presentation, such as clothes. Compensatory motives to boost self-worth were proposed by Alfred Adler and were related to the use of objects by Robert Wicklund and Peter Gollwitzer (1982). When individuals perceive themselves as lacking central attributes of an aspired identity, they use the mechanism of symbolic self-completion to claim that identity, through the use of symbolic objects. An individual who aspires, for example, to be a good tennis player (but is not) may claim that identity by wearing tennis accessories suggesting good performance.

2.3 *Personification*

Objects may be personified by giving them a name and attributing human capacities to them. Objects that invite personification are ones that represent creatures such as dolls or stuffed animals, that invite imaginary dialogues, such as diaries; cozy objects; instruments that ensure survival, such as vehicles; magic or religious objects such as charms or Christian crosses; or very complex objects that suggest having a life of their own such as computers. Personified objects resemble imaginary companions and symbolize aspects of a generalized other. Communicating with a personified object externalizes the process of thinking or self-communication. Also, it provides the individual with a companion in times when significant others

may be missing. In the place of an external other, personified objects may thus serve to confirm the individual's spontaneous gestures and identity; or, by reinforcing the 'process of self,' the internal dialogue between I and Me (G. H. Mead).

2.4 Mnemonic Uses

Objects are often used as prospective reminders to do things (e.g., a knot in a handkerchief). In a less intentional way, the human environment is usually structured to remind us to do things when appropriate. Light switches, for example, are placed near the door to remind us to turn off the lights when leaving the room. Objects that explicitly serve to complement retrospective memory were termed souvenir by Pierre Janet (1928). When personal objects serve mnemonic functions, they extend not only to the past, but also to distant places and absent others, and sometimes to an aspired future as well. Souvenirs serve primarily to remind and are therefore exempted from practical uses. All personal objects, however, at least secondarily also remind of the situations and persons that have been associated with them, and, more specifically in the case of gifts, of the situation when the object was given and of the giver.

Memorabilia, mementos, or souvenirs serve autobiographical purposes by referring to the personal past. Objects may also serve to remind of significant others. Usually they refer to family members or intimate partners. Most often the emotional link to significant others is symbolized in situations of temporary or involuntary separation. The more an individual is attached to significant others, the more these links are symbolized by objects in times of separation. Transitions, such as starting university, migration, or moving into a nursing home, are bridged by reminding uses of objects. Reminding objects apparently reinforce the mental representation of attachment figures and other secure bases. Using objects for reminding serves the self by reinforcing a sense of self-continuity and by reinforcing the mental representation and experience of a link to significant others whose real or imagined response is essential to identity.

In protracted mourning, personal relics of deceased loved ones may become personified and treated as if they were the one they stand for. These linking objects (Volkan 1981) no longer refer to an absent other but stand in his or her place. Here the triadic relationship between sign, signified, and interpretant has degenerated to a dyadic relationship between interpretant and object as an imaginary other.

2.5 Transitional Objects

The precursor of personal objects is, in about half of all toddlers in Western countries, the transitional object, often a teddy bear or soft blanket. It serves as substitute or additional secure base when the attachment system is activated. Winnicott (1971) coined the term transitional object to indicate that it eases the transition from total dependence on the primary attachment figure to independence. As well, Winnicott hypothesized that toddlers would only develop attachment to a transitional object if they had developed a secure attachment pattern.

Thus material objects are part of cultural systems and serve cultural functions of tradition, social functions of regulating status, and economic functions as commodities. Also material objects serve as media for individuals' relation to culture, to other individuals, and to themselves.

See also: Action Theory: Psychological; Freud, Sigmund (1856–1939); Ritual and Symbolism, Archaeology of; Symbolic Boundaries: Overview; Symbolism in Anthropology; Symbolism (Religious) and Icon

Bibliography

Appadurai A 1986 (ed.) *The Social Life of Things. Commodities in Cultural Perspective*. Cambridge University Press, Cambridge, UK

Barthes R 1964 *Éléments de sémiologie*. Éditions de Seuil, Paris

Baudrillard J 1968 *Le système des objets*. Gallimard, Paris

Boesch E E 1991 *Symbolic Action Theory and Cultural Psychology*. Springer Verlag, Berlin

Bourdieu P 1979 *La distinction: Critique sociale du jugement*. Les éditions de Minuit, Paris

Csikszentmihaly M, Rochberg-Halton E 1981 *The Meaning Of Things: Domestic Symbols Of The Self*. Cambridge University Press, New York

Dittmar H 1992 *The Social Psychology of Material Possessions*. St.Martin's Press, New York

Douglas M 1973 *Natural Symbols: Explorations in Cosmology*. Barrie and Jenkins, London

Fischler C 1990 *L'homnivore. Le goût, la cuisine et le corps*. Éditions Odile Jacob, Paris

Gibson J J 1979 *The Ecological Approach To Visual Perception*. Houghton Mifflin Company, Boston

Giedion S 1948 *Mechanization Takes Command*. Oxford University Press, Oxford, UK

Habermas T 1999 *Geliebte Objekte: Instrumente und Symbole der Identitätsbildung*. Suhrkamp, Frankfurt, Germany

Janet P 1928 *L'Évolution de la mémoire et de la notion du temps*. Cahine, Paris

Kubler G 1972 *The Shape of Time*. Yale University Press, New Haven, CT

Leroi-Gourhan A 1964 *La geste et la parole*. Editions Albin Michel, Paris

Lowe J W G, Lowe E D 1982 Cultural patterns and process: A study of stylistic change in women's dress. *American Anthropologist* **84**: 521–44

Mauss M 1924 *Essai sur le don*. Presses Universitaires de France, Paris

Miller D 1995 Consumption and commodities. *Annual Review of Anthropology* **24**: 141–61

Norman D 1988 *The Psychology of Everyday Things*. Basic Books, New York
Richardson J, Kroeber A L 1940 Three centuries of women's dress fashion: A quantitative analysis. In: Kroeber A L (ed.) 1952 *The Nature of Culture*. University of Chicago Press, Chicago pp. 358–72
Rogan B 1990 On the custom of naming artifacts. *Ethnologia Europea* **20**: 47–60
Simmel G 1923 *Philosophische Kultur*. Gustav Kiepenheuer, Potsdam, Germany
Veblen T 1899 *The Theory Of The Leisure Class*. Houghton, Boston
Volkan V D 1981 *Linking Objects and Linking Phenomena*. International Universities Press, New York
Wicklund R A, Gollwitzer P M 1982 *Symbolic Self-Completion*. Erlbaum, Hillsdale, NJ
Weiner A B 1992 *Inalienable Possessions*. University of California Press, Berkeley, CA
Winnicott D W 1971 *Playing and Reality*. Tavistock, London

T. Habermas

Observational Methods in Human Development Research

Today it goes without saying that developmental psychology is of scientific value, but looking back in history reveals that our ancestors might have disagreed with the current understanding of human development. Which sources of information can be used to find out more about our ancestors' perceptions of the course of life? There is nothing like today's developmental research, but there is the option to have a closer look at written documents and pictures that tell us more about how the course of life formerly was divided and how stages like childhood, youth, adulthood, and older age were characterized (Ariès 2000).

The course of life was not only described in written documents. There are many pictures, prints, and paintings that show the different stages of life, either lined up one after another, or as segments of a wheel—the wheel of life—or, since 1531, as a staircase. Most staircases are divided into 10 segments of 10 years each, arranged symmetrically with its peak between 50 and 60. The steps ascend from 0 to 50 and descend from 60 to 100. The human is led to the Last Judgment after death, where the decision is made whether he or she will go to heaven or to hell. Such pictures were not meant to be realistic; they were supposed to be a reminder of human mortality. In the twentieth century pictorial representations like this became rare and the staircase of life became a caricature of human development.

Today's knowledge society is proud to be able to distinguish between facts and opinions, evidence and assumption, fantasy and reality about human life courses. We tend to forget that it has not always been that way. And the reason why it all changed is that people started to conduct diary studies to describe empirically the courses of development (Wallace et al. 1994, Jaeger 1985).

1. *Clara and William Stern's Diary Studies*

April 7, 1900 was the date of an extremely critical life event for the university lecturer William Stern and his wife Clara. At 2 a.m. their first child, Hilde, was born. The birth of Hilde was the start of a project that was to determine the Stern's lives fundamentally for the next 15 years. William Stern, an experimental psychologist and philosopher who had studied and graduated at the University of Berlin, became—along with Piaget and Vygotsky—one of the most significant developmental psychologists of the twentieth century. His gifted wife, who unfortunately was not able to study, became an amateur scientist. She did not just work for, but with, her husband. Both of them were pioneers.

If one wants to start a research project, it is highly recommended to get prepared long before it actually starts: formulating questions, applying for funding, hiring assistants. When Hilde Stern was born, the Sterns started a project without doing any of this. They were not aware of how intense and long-lasting their project would be. We do not know when and why they had the idea of describing their first child's development—and also that of their two subsequent children—in the form of diaries. All their documentations are preserved. Legible duplicates can be accessed with the help of a database (Stern and Stern 1900–18).

Clara and William Stern wrote diaries for each of their children. Looking at the extent and duration of the documentations it becomes clear that the diaries are not continual documentations of development. The older the children get, the less is documented. It is also remarkable that the extent of the diaries for the three children varies. The most detailed descriptions are the ones for the first child. What were Clara and Willam Stern's reasons for not continuing to describe the development in so much detail?

When the Sterns started the project, diary studies were common tools in developmental research. Like Preyer (1882) the Sterns tried to differentiate between observational facts and their interpretations. The scientific value of observations increases according to the ability to describe them vividly but without judgments and explanations. Balancing empathy and distance is difficult for all who conduct diary studies. In this respect Clara Stern was almost perfectly able to describe her observations—even in the eyes of her very critical husband. The Sterns changed Preyer's approach in that that they did not observe their children at certain times, but 24 hours a day they remembered or wrote down notes about everything remarkable in the development of their children. In the evening they would sit down at their desk to write everything into the diaries. Remarkable things would be more likely to be written down than things that had become part of

a routine. Those who closely observe a child's development make up their own scale when it comes to comparing the child observed with other children. This scale is used to decide whether a development is similar or different, usual or unusual, known or unknown. This way everything remarkable is mainly reduced to what is classified as individually different.

Nobody who writes a diary functions like a machine that is doing observations. The Sterns were far from turning their home into a laboratory. Being parents they acted like parents, not like scientists trying to turn their children into guinea pigs. The children did not feel watched or studied. It was by chance that they found out about the diaries at a young age, not thinking anything of it. Unlike most other diary studies, the Sterns' documentations decribe not only the first three years of their children's development; some of their documentations go far beyond those years. Writing diaries about the first three years of a child's development is quite easy because the child is housebound. If the mother is the one who is writing the diary and at the same time the one who the child relates to closely, there is almost nothing she does not know about. That changes completely once the children are able to leave the house on their own. The mother does not know what happens when the child visits other families or interacts with peers. She can only write down what the child tells her, but not her own observations.

What are the facts written down in the Sterns' diaries? It is difficult to say because they neither tried to answer certain *a priori* developmental questions, nor did they categorize their observations. Many researchers think that this kind of documentation rejects all fundamental principles of scientific work. The accusation is that their unorganized observations are just a collection of anecdotes with maybe some heuristic value. The Sterns' reason for not using standardized observations and descriptions was not only their desire closely to examine an emerging human life; arguments not to use standardized methods can also be found in William Stern's philosophical work (Behrens and Deutsch 1991). In Stern's view, every individual is determined by experience and the ability to experience. The Sterns decided not to turn the observed individuals into objects and individual biographies were the results. Their closeness to life, details, and vividness are still fascinating.

2. *From Diaries to Paradigms*

With the help of diary studies developmental psychology turned into an empirical science. Until the nineteenth century men were the ones who wrote developmental diaries. So far nobody had examined critically if the observations of those philosophers, psychologists, physicians, and educationalists really were their own, or if some of them just wrote down what mothers and nurses told them. Women entered academic psychology conducting diary studies in which they observed their own or other children of the family. By writing diaries they were able to prove that their observations were just as objective and precise as those of their male colleagues. Diary studies were most popular in the end of the nineteenth and the beginning of the twentieth century.

The differentiation between personal and scientific diaries certainly is not an easy task in the case of the Sterns' documentations. In the twentieth century diary studies became less important in developmental psychology (Wright 1960). The reasons are obvious: diary and single case studies could not be generalized and therefore were of little scientific value. Diary studies are based on observations; language is used to write down subjective impressions. The objectivity of observations can be improved if supported technically, e.g., with a tape recorder or video camera. Diary studies are documentations of events that occur under natural circumstances which makes it more difficult to compare different children, even different stages of development of the same child. Standardized conditions like tests and experimental excercises started to be used more frequently to measure courses of development.

Due to the rise of behaviorism in psychology diaries have widely lost their earlier reputation. They were replaced by less subjective methods of observation. Developmental tests make it possible to compare an individual's performance with that of a reference group, because test items are being selected according to empirically determined (age) norms. This approach can be applied to any contents of interest across the full life span. Developmental tests, however, cannot capture a person's spontaneous behavior under natural circumstances. This was made possible by advancements in modern audio and video technology allowing for online registrations of ongoing behavioral exchanges in social interactions. Recordings can be made of free interactions of mother–child dyads as well as of partially standardized interactions like in the strange situation test that became the central tool of attachment research. Diary observations suffer from memory, attention, or motivation deficits of the diarist, while technical recordings do not—but they lead to other problems. Recorded material is not self-explanatory. It has to be perceived, identified, categorized, and interpreted. Observers (raters) may disagree on what is 'in' the material. Agreement measures, for example Cohen's Kappa, reflect the consensus among observers. This is a way to introduce qualitative objectivity measures into the application of observational methods (Bakeman and Gottman 1987).

Beside developmental tests and recordings of naturally occuring behavior, a third category of observations became increasingly important in developmental research as in many other areas of psychology:

(experimental) paradigms. They require highly standardized conditions and often make use of advanced recording technology. If one wants to rank observational methods along a dimension of situational control, a bipolar dimension will appear with traditional diaries as one pole and paradigms as the other. In contemporary developmental psychology paradigms shoot up like mushrooms. They concentrate on precisely defined, sometimes diminutive aspects of behavior. Paradigms like preferential looking, head turning, referential communication, or still face do have many advantages compared to other observational methods, especially diaries. If the technical equipment works, one can use those paradigms any time and everywhere in innumerable variations. Researchers do not have to wait for certain phenomena to appear in the natural course of development; instead they are able to provoke the phenomena they are interested in at any time under controlled conditions. Therefore paradigms can reproduce themselves by generating one more time the data asked for.

There are still many good reasons to conduct diary studies nowadays (Hoppe-Graff 1989). Working with a video camera seems to be more objective than writing diaries. A videotape is not the final result though. Observers have to start interpreting what has been taped. Exactly those kind of interpretations have led to the argument that diary studies lack objectivity. Natural observations cannot be replaced by experiments or tests. The average individual does not exist, and results from experiments do not give detailed information about whole courses of development. The Sterns' diary studies showed that by comparing the development of three children, universal as well as individual trends could be identified. These trends can now be examined more closely. Maybe diary studies will become more important once again when psychology does not divide humans into structures and functions anymore, but rather starts to stress the combination of both in each individual person. An excellent example for modern diary studies is Mendelson's (1990) *Becoming a Brother*.

Standardized observations like tests and paradigms, examining representative samples if possible, have slowly replaced natural, unstandardized observations. Developmental psychology changed fundamentally because ideal scientific objectivity was what most researchers strived for—the individual case being replaced by the representative sample, the natural observations replaced by measures of artificial paradigms. Maybe the future will neutralize the contradiction between naturalness and artificiality with the help of 'natural paradigms' (Deutsch et al. 2001).

3. Conclusions

Nowadays modern studies in developmental psychology do not focus on single persons, but on psychical functions. Samples, whose individuality is a negligible quantity, are used to measure specific developmental changes. Individuals only become visible if their development is unusual, e.g., it does not progress at all, as in the case of a severe developmental disorder, or it progresses much faster or slower than average.

Since statistical measures have been introduced to developmental psychology, generalizations refer to central tendencies of samples which are supposed to be as representative as possible for a certain age group or particular time. Inductive generalizations, in which case by case one has to decide where similarities to all those that already have been examined end and differences between subgroups of individuals start, are no longer very common.

Today developmental psychology is characterized by a striving for ideal objectivity oriented according to experimental psychology. That is why extensive observations of individuals under natural circumstances have been pushed to the periphery. There they mainly survive as anecdotes which do not prove anything but—at the best—can be used as illustrations. Nevertheless exceptions exist, e.g., in the field of language development. Here the spontaneous use of language cannot be reduced to tests for language development and some standardized tests that claim to be paradigms. Other areas in developmental psychology are dominated by paradigms in which almost no spontaneous data from a single being is recorded and documented.

Artificial standardizations have to meet the scientific ideal of objectivity that is oriented according to reproductivity and generalization. The value of a scientific study is not measured by its ecological validity but by its objective and reliable realization based on a representative sample of subjects. Perhaps the future of observational methods does neither belong to natural observations, nor to artificial paradigms, but to natural paradigms in which everyday life and research are no longer opponents but partners.

See also: Case Study: Logic; Case Study: Methods and Analysis; Case-oriented Research; Developmental Psychology; Field Observational Research in Anthropology and Sociology; Lifespan Development, Theory of; Narratives and Accounts, in the Social and Behavioral Sciences; Observational Studies: Outcome-based Sampling; Observational Studies: Overview; Participant Observation

Bibliography

Ariès P 2000 *Geschichte der Kindheit*. Deutscher. Taschenbuch-Verlag, Munich, Germany

Bakeman R, Gottman J M 1987 Applying observational methods: A systematic view. In: Osofsky J D (ed.) *Handbook of Infant Development*, 2nd edn. Wiley, New York, pp. 818–54

Behrens H, Deutsch W 1991 Die Tagebücher von William und Clara Stern. In: Deutsch W (ed.) *Über die verborgene Aktualität von William Stern*. Peter Lang, Frankfurt, Germany, pp. 19–37

Deutsch W, Wagner A, Burchardt R, Schulz N, Nakath J 2001 Person in the language of singletons, siblings, and twins. In: Bowerman M, Levinson S (eds.) *Language Acquisition and Conceptual Development*. Cambridge University Press, Cambridge, UK

Hoppe-Graff S 1989 Die Tagebuchaufzeichnung: Plädoyer für eine vergessene Form der Längsschnittbeobachtung. In: Keller H (ed.) *Handbuch der Kleinkindforschung*. Springer, Berlin, pp. 233–51

Jaeger S 1985 The origin of the diary method in developmental psychology. In: Eckardt G, Bringman W G, Sprung L (eds.) *Contributions to a History of Developmental Psychology*. Mouton, Berlin, pp. 63–74

Mendelson M J 1990 *Becoming a Brother*. MIT Press, Cambridge, MA

Preyer W 1882 *Die Seele des Kindes*. Grieben, Leipzig, Germany

Stern C, Stern W 1900–18 Die Tagebücher. Elektronische Abschrift der unveröffentlichten Tagebücher aus dem Nachlass. Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands

Wallace D B, Franklin M B, Keegan R T 1994 The observing eye: A century of baby diaries. *Human Development* **37**: 1–29

Wright H F 1960 Observational child study. In: Mussen P H (ed.) *Handbook of Research Methods in Child Development*. Wiley, New York, pp. 71–139

W. Deutsch

Observational Studies: Outcome-based Sampling

1. Introduction

An outcome-based sample in an observational study is one obtained by stratification on the basis of response behavior whose explanation is the sarget of study. Observations on response and explanatory variables (covariates) are collected within each stratum. These are then used for statistical inference on the conditional distribution of the response, given the covariates. For example, a study of occupational choice may draw a sample stratified by occupation, so the first stratum is a sample of engineers, the second stratum is a sample of educators, and so forth. Data are collected on covariates such as gender and utilization of training subsidies. The observations are then used to infer the impact of training subsidies on occupational choice.

Outcome-based sampling is also termed endogenous, intercept, on-site, experience-based, response-based, or choice-based sampling, and in biostatistics, retrospective or case-control sampling. To illustrate, case-control designs sample from one stratum that has a specified outcome (the cases) and a second stratum that does not (the controls). There may be further stratification of the controls so that they resemble the cases on some covariates such as age. Outcome-based samples can be contrasted with exogenous or prospective samples that sample randomly from the target population, or stratify solely on covariates. Outcome-based sampling methods have had wide application in social science, biostatistics, and engineering; examples are Akin et al. (1998), Domowitz and Sartain (1999), Laitila (1999), and van Etten et al. (1999).

2. Design and Robustness Issues in Endogenous Samples

Outcome-based samples are useful for studying rare responses that are difficult to observe in random samples of practical size, such as risk factors for a rare disease, and for studying outcomes that randomly sampled subjects are unable or unwilling to report. An example of the second type is the study of mental illness as a risk factor for homelessness, where the only practical sample methodology may be one that surveys cases who are homeless and controls who are not (Early 1999, Lacy 1997). Outcome-based sampling may also offer economies in recruiting and interviewing subjects. For example, consumers who purchase and register a product form an outcome-based stratum that the manufacturer can survey inexpensively.

Outcome-based samples may arise by deliberate design, as in case-control studies, but can also be unintentional. General purpose surveys often use cluster sampling to facilitate enumerating and contacting subjects. The cluster stratification is exogenous for most purposes, but has the properties of an outcome-based sample for behavior that is related to the stratification. For example, cluster sampling by census tract or area code becomes outcome based in studies of residential location of migration behavior. Further, screening, self-selection, and attrition create statistical problems like those in outcome-based samples. For example, a sampling protocol that selects subjects by random-digit dialing will over-sample individuals with more than one telephone line, resulting in a sample that is endogenous for study of the demand for telephone service. Differential attrition in which, for example, individuals who live alone are more difficult to contact or more frequently refuse to be interviewed, will produce a sample that is endogenous for the study of living arrangements.

Outcome-based samples are less robust than random samples. Because stratum boundaries coincide with the behavioral categories under study, variations in sampling protocol across strata are confounded with behavioral response. For example, in case-control studies of risk factors for disease, variations in question format between cases and controls, or even the context for questions induced by the disease status, can alter recall on exposure to hazards. Unlike prospective studies where randomization of treatments and double-blind designs can minimize the risks of contamination, retrospective studies are vulnerable

to investigator bias and latent causal factors that are correlated with measured risk factors. As a result, differences in reported exposure to risk factors between cases and controls may arise from spurious sampling effects as well as from a true causal effect. The difficulty of avoiding contaminating bias in case-control settings is discussed further in Austin et al. (1994), Feinstein (1979), Lilienfeld and Lilienfeld (1979), and Taubes (1995).

3. *Statistical Inference in Endogenous Samples*

Statistical methods developed for random samples will often be inconsistent or inefficient when applied to endogenous samples. The essential problem is that the analysis is attempting to infer properties of the conditional distribution of outcomes given covariates, using observations that are drawn from conditional distribution of covariates given outcomes. The solution to the inference problem is to incorporate the mapping between the conditional distributions in the analysis, either by reweighting observations so that they behave *as if* they were drawn from a random sample, or by reweighting the probability model for a random sample so that it is consistent with the empirical sampling process. The statistical issues in analyzing outcome-based samples were treated in a seminal paper by Manski and Lerman (1977), with further results by Manski and McFadden (1981) and Cosslett (1981). A largely independent treatment of logistic regression in the case-control model appeared in biostatistics, notably in the papers of Anderson (1972) and Prentice and Pyke (1979). Recent papers by Hsieh et al. (1985) and Breslow (1996) sharpen the statistical analysis of case-control designs and show their relationship to the general problem of endogenous sampling.

Table 1 is a contingency table that depicts, schematically, the population probability law for an outcome variable y and a vector of covariates z. (This exposition treats y and z as discrete, but the discussion applies with minor modifications to the case where y and/or some components of z are continuous.) The joint probability of a (y, z) cell can be written as the product of the conditional probability of y given z times the marginal probability of z, $p(y, z) = P(y|z) \cdot p(z)$. The row sums give the marginal probability $p(z)$ of z, and the column sums give the marginal probability $q(y) = \sum_z P(y|z) \cdot p(z)$ of y. Dividing a cell probability by its column sum gives the conditional probability of z given y, $Q(z|y) = P(y|z)p(z)/q(y)$. This mapping between conditional probabilities is Bayes' Law. The target of statistical analysis is the conditional probability $P(y|z)$, sometimes termed the response probability. In applications, $P(y|z)$ is usually assumed to be invariant under treatments that alter the marginal probability of z; then knowledge of $P(y|z)$ permits the analysis to forecast y in new populations or under policy treatments where the z distribution is changed. (A conditional probability with this invariance property is sometimes said to define a causal model. It is true that a causal structure will imply this invariance property, but it is also possible for the invariance property to hold, making forecasting possible, without the presence of a deeper causal structure. Further, there are straightforward statistical tests for the invariance property, while detection of true causal structures is beyond the reach of statistics. For these reasons, it is best to avoid the language of causality and concentrate instead on invariance properties.)

Random sampling draws from the table in proportion to the cell probabilities. Exogenous stratification draws rows, with probabilities that may differ from the population marginal probabilities $p(z)$, and then within a row draws columns in proportion to their population conditional probabilities $P(y|z)$. A simple outcome-based sampling design draws columns, with probabilities that may differ from the population marginal probabilities $q(y)$, then within a column draws rows in proportion to their conditional probabilities $Q(z|y) = P(y|z)p(z)/q(y)$.

More complex endogenous sampling designs are also possible. A general framework that permits a unified analysis of many sampling schemes characterizes the sampling protocol for a stratum s in terms of a probability $R(y, z, s)$ that a member of the population in cell (y, z) will qualify for the stratum. The joint probability that a member of the population is in cell (y, z) and will qualify for stratum s is then $R(z, y, s) \cdot P(y|z, \beta_0) \cdot p(z)$. Then the proportion of the population qualifying into the stratum, or qualification factor, is $r(s) = \sum_z \sum_y R(z, y, s) \cdot P(y|z) \cdot p(z)$, and the conditional distribution of (z, y) given qualification is $R(z, y, s) \cdot P(y|z) \cdot p(z)/r(s)$. When a fraction of the sample $f(s)$ is drawn from stratum s, the

Table 1

	y_1	y_2	…	y_J	
z_1	$P(y_1\|z_1, \beta_0)p(z_1)$	$P(y_2\|z_1, \beta_0)p(z_1)$	…	$P(y_J\|z_1, \beta_0)p(z_1)$	$p(z_1)$
z_2	$P(y_1\|z_2, \beta_0)p(z_2)$	$P(y_2\|z_2, \beta_0)p(z_2)$	…	$P(y_J\|z_2, \beta_0)p(z_2)$	$p(z_2$
:	:	:		:	:
z_K	$P(y_1\|z_K, \beta_0)p(z_K)$	$P(y_2\|z_K, \beta_0)p(z_K)$	…	$P(y_J\|z_K, \beta_0)p(z_K)$	$p(z_K)$
	$q(y_1)$	$q(y_2)$	…	$q(y_1)$	1

probability law for an observation from the pooled sample is $g(y, z) = \sum_s R(z, y, s) \cdot P(y|z) \cdot p(z) f(s)/r(s)$. The conditional distribution of y given z in this pooled sample is

$$g(y|z) = \frac{P(y|z) \sum_{\Theta} R(z, y, s) \cdot f(s)/r(s)}{\sum_{\tau'} P(y'|z) \sum_{\Theta} R(z, y', s) \cdot f(s)/r(s)}$$

Note that this conditional probability depends on the marginal distribution of z only through the qualification factors.

When the sampling protocol is exogenous (i.e., $R(y, z, s)$ does not depend on y), the conditional probability $g(y|z)$ for the pooled sample equals the population conditional probability $P(y|z)$. Consequently, any statistical inference procedure designed to reveal features of the conditional probability $P(y|z)$ in random samples will apply to an exogenously stratified sample. In particular, if $P(y|z)$ is in a parametric family, then maximization of the random sample likelihood function in an exogenously stratified sample will have the same properties as in a random sample. However, in an endogenous sample in which the qualification probability $R(y, z, s)$ does depend on y, the conditional probability $g(y|z)$ for the pooled sample is *not* equal to $P(y|z)$. Consequently, statistical inference assuming that the data generation process is described by $P(y|z)$ is generally statistically inconsistent. Also, the distribution of covariates in an endogenous sample will differ from their population distribution, with

$$g(z) = p(z) \cdot \sum_s (f(s)/r(s)) \sum_y R(z, y, s) P(y|z)$$

and a corresponding correction factor must be applied to the sample empirical distribution of z to estimate population quantities consistently.

Manski and McFadden (1981) propose that statistical inference when $P(y|z)$ is parametric be based on the conditional likelihood $g(y|z)$, and term this the conditional maximum likelihood (CML) method. When the qualification factors $r(s)$ and sample frequencies $f(s)$ are known or can be estimated consistently from external samples, and the forms of $P(y|z)$ and $R(z, y, s)$ allow identification of any unknown parameters in $R(z, y, s)$, this approach is consistent. In general, the probability $g(y|z)$ is not in the same parametric family as $P(y|z)$. To illustrate, suppose a population has a binomial probit dose-response curve, $P(2|z) = \Phi(\alpha + z\beta)$, and $P(1|z) = 1 - \Phi(\alpha + z\beta)$. Suppose the sample consists of a randomly sampled stratum 1 with $R(z, y, 1) \equiv 1$, plus a stratum 2 drawn from the population with response $y = 2$, with $R(z, y, 2)$ equal to one if $y = 2$, and zero otherwise. This is called an enriched sample. The qualification factors are $r(1) = 1$ and $r(2) = q(2)$. If $q(2)$ is known, a consistent estimate of the slope parameter β in the dose-response curve can be obtained by the CML method with

$$g(1|z) = \frac{[1 - \Phi(\alpha + z\beta)] f(1)}{[1 - \Phi(\alpha + z\beta)] f(1) + \Phi(\alpha + z\beta)[f(1) + f(2)/q(2)]}$$

By contrast, likelihood maximization using $P(y|z)$ is not consistent for β.

An important simplification of the CML method occurs for what in biostatistics is termed logistic regression in case-control designs. Suppose that the vector of covariates is partitioned into components $z = (v, x)$ with v discrete. (In biostatistics, v will often include variables such as age and gender whose distributions are matched between cases and controls.) Suppose that $P(y|v, x)$ has a multinomial logit form,

$$P(y|v, x) = \exp(\alpha_y + \gamma_{yv} + x\beta_y) / \sum_y \exp(\alpha_y + \gamma_{y'v} + x\beta'_y)$$

In this model, the β_y are slope coefficients for the covariates x, and α_y and γ_{yv} are response-specific effects and interactions of response-specific and v-specific effects. Suppose that the qualification probability $R(v, x, y, s)$ does not depend on x. For identification, a normalization such as $\alpha_1 = \gamma_{1v} = \beta_1 = 0$ is imposed. The conditional probability $g(y|z)$ is again of multinomial logit form, with the same β_y parameters but with the remaining parameters shifted from their population values by sampling factors,

$$g(y|v, x) = \exp(\alpha^*_y + \gamma^*_{yv} + x\beta_y) / \sum_{y'} \exp(\alpha^*_{y'} + \gamma^*_{y'v} + x\beta_{y'})$$

with

$$\alpha^*_y + \gamma^*_{yv} = \alpha_y + \gamma_{yv} + \log\left(\sum_s R(v, y, s) \cdot f(s)/r(s)\right)$$

Note that consistent estimation of this model requires the inclusion of all the alternative specific effects and interactions that are modified by sampling factors. However, if these variables are included, then the slope parameters β_y are estimated consistently without further adjustments for endogenous sampling. (If the raising factors are estimated rather than known, there is an additional contribution to the asymptotic covariance matrix (Hsieh et al. 1985). If the model had included interactions of v and x, then the coefficients on these interactions would contain sampling factors that must be removed to obtain consistent estimates of the corresponding population interactions). The simplification above was first noted by Anderson (1972).

Another solution to inference in choice-based samples is the weighted exogenous sample maximum likelihood (WESML) method proposed by Manski and Lerman (1977) in their seminal paper on this

subject. This solution weights the observations, with a weight for (y, z) that is inversely proportional to $\sum_s R(z, y, s) \cdot f(s)/r(s)$. When these weights are finite and can be calculated or estimated from external data, consistent estimates can be obtained when the weighted observations are treated *as if* they were drawn from an exogenous sample. This is a pseudo-maximum likelihood method, and the associated covariance matrix will be of the sandwich form, with an outer product of the pseudo-score in the middle and the inverse of the pseudo-information matrix on the outside. For the case of pure outcome-based sampling, where the strata correspond to the possible values of y and $R(y, z, s)$ is one when $y = s$ and zero otherwise, the weights are $r(y)/f(y)$, the ratio of the population to the sample frequency of observing y.

A third approach to inference in outcome-based samples, developed by Cosslett (1981), achieves a semiparametric efficiency bound when $P(y|z)$ is parametric. This estimator is derived by maximum likelihood estimation that first formally concentrates out the marginal probability $p(z)$. The result is an objective function similar to $g(y|z)$, but with parametric sampling factors that satisfy side conditions. Hsieh et al. (1985) adapt Cosslett's analysis to improve the efficiency of the CML method in the presence of auxiliary information.

4. Extensions and Related Issues

Closely related to outcome-based sampling in the form described here is the problem of analyzing samples that are subject to self-selection. Several strands in the statistical and econometric literature have investigated estimators appropriate to such data: a seminal paper by Heckman (1974) on sample selection, further work on endogenous stratification by Hausman and Wise (1977), and related work on switching regression by Goldfeld and Quandt (1975).

Extensions of the basic framework for inference in outcome-based samples have been made for a variety of problems. Breslow (1996) provides methods for complex case-control sample designs. Imbens (1992) provides methods for combining outcome-based survey data with aggregate statistics. Imbens and Lancaster (1996) and McFadden (in press) have studied the problem of analysis of endogenously recruited panels.

5. Conclusions

Samples in observational studies often fail to qualify as random or exogenously stratified, and instead have the properties of endogenous or outcome-based samples. This may occur inadvertently due to attrition, self-selection, or stratification that is correlated with the response being studied, or may result from deliberate design decisions to facilitate observation, as in case-control studies. This article has described the robustness issues and inference methods that are relevant for these samples.

Bibliography

Akin J et al. 1998 Price elasticities of demand for curative health care with control for sample selectivity on endogenous illness: An analysis for Sri Lanka. *Health Economics* **7**: 509–31

Anderson J 1972 Separate sample logistic discrimination. *Biometrika* **59**: 19–35

Austin H, Hill H, Flanders W, Greenberg R 1994 Limitations in the application of case-control methodology. *Epidemiologic Reviews* **16**: 65–76

Breslow N 1996 Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* **91**: 14–28

Cosslett S 1981 Maximum likelihood estimation for choice-based samples. *Econometrica* **49**: 1289–316

Domowitz I, Sartain R 1999 Determinants of the consumer bankruptcy decision. *Journal of Finance* **54**: 403–20

Early D 1999 A microeconomic analysis of homelessness: An empirical investigation using choice-based sampling. *Journal of Housing Economics* **8**: 312–27

Feinstein A 1979 Methodological problems and standards in case-control research. *Journal of Chromic Diseases* **32**: 35–41

Goldfeld S, Quandt R 1975 Estimation in a disequilibrium model and the value of information. *Journal of Econometrics* **3**(4): 325–48

Hausman J, Wise D 1977 Social experimentation, truncated distributions, and efficient estimation. *Econometrica* **45**(4): 919–38

Heckman J 1974 Shadow prices, market wages, and the labor supply. *Econometrica* **42**: 679–94

Hsieh D, Manski C, McFadden D 1985 Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association* **80**: 651–62

Imbens G 1992 An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica* **60**: 1187–214

Imbens G, Lancaster T 1996 Efficient estimation and stratified sampling. *Journal of Econometrics* **74**: 289–318

Lacy M 1997 Efficiently studying rare events: Case-control methods for sociologists. *Sociological Perspectives* **40**: 129–154

Laitila T 1999 Estimation of combined site-choice and trip-frequency models of recreational demand using choice-based and on-site samples. *Economics Letters* **64**: 17–23

Lilienfeld A, Lilienfeld D 1979 A century of case-control studies. Progress? *Journal of Chronic Diseases* **32**: 5–13

Manski C, Lerman S 1977 The estimation of choice probabilities from choice-based samples. *Econometrica* **45**: 1977–88

Manski C, McFadden D 1981 Alternative estimators and sample designs for discrete data analysis. In: Manski C, McFadden D (eds.) *Structural Analysis of Discrete Data with Econometric Applications*. pp. 51–111

McFadden D in press On endogenously recruited panels. *Journal of Applied Econometrics*.

Prentice R, Pyke R 1979 Logistic disease incidence models and case-control studies. *Biometrica* **66**: 403–11

Taubes G 1995 Epidemiology faces its limits. *Science* **269**: 164–9

van Etten M, Neumark Y, Anthony J 1999 Male-female differences in the earliest stages of drug involvement. *Addiction* **94**: 1413–19

D. McFadden

Observational Studies: Overview

1. Introduction

1.1 Definition

Cochran (1965) defined an *observational study* as a comparison of treated and control groups in which

> the objective is to elucidate cause-and-effect relationships [… in which it] is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures.

Treatments, policies and interventions that are expected to benefit recipients can be studied in experiments that randomly allocate experimental subjects to treatment groups, so comparable subjects receive competing treatments (Boruch 1997, Piantadosi 1997, Friedman et al. 1998). Observational studies are common, nonetheless, in most fields that study the effects of treatments on people, because harmful or unwanted treatments cannot be imposed on human subjects. When treatments are not randomized, subjects assigned to competing treatments may not be comparable, so differences in outcomes may or may not be effects caused by the treatments (Rubin 1974).

1.2 Examples

Two examples of observational studies are described. Later sections refer to these examples.

1.2.1 Long-term psychological effects of the death of a close relative. Lehman et al. (1987) attempted to estimate the long term psychological effects of the sudden death of a spouse or a child in a car crash. Starting with 80 bereaved spouses and parents, they matched 80 controls drawn from 7581 individuals who came to renew a drivers license, matching for gender, age, family income before the crash, education level, number and ages of children. Contrasting their findings with the views of Bowlby and Freud they concluded:

> Contrary to what some early writers have suggested about the duration of the major symptoms of bereavement … both spouses and parents in our study showed clear evidence of depression and lack of resolution at the time of the interview, which was five to seven years after the loss occurred.

1.2.2 Effects on children of occupational exposures to lead. Morton et al. (1982) asked whether children were harmed by lead brought home in the clothes and hair of parents who were exposed to lead at work. They matched children whose parents worked in a battery factory to unexposed control children of the same age and neighborhood, comparing the level of lead found in the children's blood, finding elevated levels of lead in exposed children.

In addition, they compared exposed children whose parents had varied levels of exposure to lead at the factory, finding that parents who had higher exposures on the job in turn had children with more lead in their blood. Finally, they compared exposed children whose parents had varied hygiene upon leaving the factory at the end of the day, finding that poor hygiene of the parent predicted higher levels of lead in the blood of the child.

1.3 Central Issues

Without random assignment of treatments, treated and control groups may not have been comparable prior to treatment, so differing outcomes may reflect either biases from these pretreatment differences or else effects actually caused by the treatment. The difficulty in distinguishing treatment effects from biases is the key consequence of the absence of randomization.

A variable measured prior to treatment is not affected by the treatment and is called a covariate. A variable measured after treatment may have been affected by the treatment and is called an outcome. An analysis that does not carefully distinguish covariates and outcomes can introduce biases into the analysis where none existed previously.

A pretreatment difference between treated and control groups is called an *overt bias* if it is accurately measured in the data at hand, and it called a *hidden bias* if it is not measured. For instance, if treated subjects are observed to be somewhat older than controls, and if age is recorded, then this is an overt bias. If treated subjects consumed more illegal narcotics than controls, and if accurate records of this are not available, then this is a hidden bias. Typically, overt biases are immediately visible in the data at hand, while hidden biases are a matter of concerned speculation and investigation.

In most observational studies, an effort is made to remove overt biases using one or more analytical methods, such as matched sampling, stratification, or model-based adjustments such as covariance adjustment. Adjustments are discussed in Sect. 2. Once this is accomplished, attention turns to addressing possible

hidden biases, including efforts to detect hidden biases and to study the sensitivity of conclusions to biases of plausible magnitude. Hidden biases are discussed in Sects. 3–5.

2. *Adjusting for Overt Biases*

2.1 *Matched Sampling*

2.1.1 Selecting from a reservoir of potential controls. Matching is the most direct and intuitive method of adjustment, in which an effort is made to compare each treated individual to one or more comparable controls. Matched sampling is most common when a small treated group is available together with a large reservoir of potential controls (Rubin 1973a). For instance, in the study of bereavement by Lehman et al. (1987), there were 80 bereaved spouses and parents and 7581 potential controls, from whom 80 matched controls were selected.

The structure of the bereavement study is typical: computerized administrative records aid in identifying and matching treated and control subjects, but additional information is required from matched subjects for research purposes; in this case, psychiatric outcomes are required. Often it is not practical to obtain the needed research data from all controls in the reservoir and instead a sample of controls is used. Matched sampling selects controls who appear comparable to treated subjects.

2.1.2 Goals of matching. Matching attempts to do three things: (a) produce matched pairs or sets that appear similar in terms of observed covariates, (b) produce treated and control groups with similar distributions of observed covariates, (c) remove that part of the bias in estimated treatment effects that is due to overt biases or imbalances in observed covariates. For instance, with pairs matched for age, goal (a) would be accomplished if each treated subject were matched to a control of the same age, whereas goal (b) would be accomplished if the entire treated group had an age distribution that closely resembled the age distribution in the matched control group, for instance, the same mean age, the same quartiles of the age distribution, and so on.

It is possible to demonstrate (Rosenbaum and Rubin 1983) that goal (a) suffices but is not necessary for goal (b), and that goal (b) suffices for goal (c). Pairs that are individually matched for age will balance age, but it is also possible to balance age with matched pairs that are not individually matched for age, and any matching that balances the distribution of age removes the part of the bias in estimated treatment effects due to imbalances in age. This is fortunate, because it is difficult to match individually on many covariates at once, but it is not so difficult to balance many covariates at once. For instance, with 20 binary covariates, there are 2^{20} or about a million types of individuals, so even with thousands of potential controls, it will often be difficult to find a control who matches a treated subject on all 20 covariates. Nonetheless, balancing 20 covariates is often quite feasible. As a result, modern matching algorithms focus on balancing covariates, while comparability within pairs plays a secondary role. A tool in balanced matching is the propensity score.

2.1.3 Propensity score. The *propensity score* is the conditional probability of receiving the treatment rather than the control given the observed covariates (Rosenbaum and Rubin 1983). Note carefully that the propensity score is defined in terms of the observed covariates even if there may be hidden biases due to unobserved covariates. In the simplest randomized experiment, treatment or control is assigned by the flip of a fair coin, and the propensity score equals $\frac{1}{2}$ for all subjects no matter what covariates are observed. In an observational study, subjects with certain observed characteristics may be more likely to receive either treatment or control, so the propensity score varies with these observed characteristics.

Matching on one variable, the propensity score, tends to balance all of the observed covariates. Matching on the propensity score together with a few other observed covariates—say, matching on the propensity score together with age and gender—also tends to balance all of the observed covariates. If it suffices to adjust for the observed covariates—that is, if there is no hidden bias due to unobserved covariates—then it also suffices to adjust for the propensity score alone. These results are Theorems 1 through 4 of Rosenbaum and Rubin (1983).

In practice, the propensity score is unknown and must be estimated. For instance, one might estimate the propensity score using logit regression (Cox and Snell 1989) of assigned treatment on observed covariates, perhaps including interactions, quadratics and transformations of the covariates. Matching on a single variable, such as the estimated propensity score, is often practical whereas matching on many covariates is not; yet matching on the propensity score balances many covariates.

2.1.4 Structure of matched sets. Most commonly, each treated subject is matched to exactly one control, but other matching structures may yield either greater bias reduction or estimates with smaller standard errors or both. In particular, if the reservoir of potential controls is large, and if obtaining data from controls is not prohibitively expensive, then the stan-

dard errors of estimated treatment effects can be substantially reduced by matching each treated subject to several controls (Smith 1997).

It is not necessary to match every treated subject to the same fixed number of controls. In terms of bias reduction, the optimal form is a *full matching* in which each matched set may have one treated subject and several controls or else one control and several treated subjects (Rosenbaum 1995). In this way, when a certain pattern of observed covariates is typical of controls, then a treated subject with this pattern will have several controls, but if instead the pattern is typical of treated subjects, several treated subjects may share one control. In a simulation study, full matching was much better than matching with a fixed number of controls.

2.1.5 Matching algorithms. Matching algorithms often express the difference in covariate values between a treated subject and a potential control in terms of a distance. One then matches a treated subject to a control who is close in terms of this distance. In the reservoir of potential controls, the one control who is closest to the first treated subject may also be the closest to the second treated subject, and some rule or algorithm is needed to assign controls to treated subjects. Optimal matching minimizes the total distance within matched sets by solving a minimum cost flow problem. The use of optimal matching in observational studies is illustrated in Rosenbaum (1995) and an implementation in the statistical package SAS is discussed by Bergstralh et al. (1996).

2.2 Stratification

An alternative to matching is stratification, in which subjects are grouped into strata so that, within each stratum, treated and control subjects have similar distributions of observed covariates. For instance, one might divide subjects into strata based on age, and then compare treated and control subjects of about the same age.

Cochran (1968) showed that five strata formed from a single continuous covariate can remove about 90 percent of the bias in that covariate. It is difficult to form strata that are homogeneous in many covariates for the same reason that it is difficult to match exactly on many covariates; see Sect. 2.1.2. As a result, strata are formed using the propensity score that balance many covariates, so within a stratum, subjects may vary in terms of age, gender, race, and wealth, but the distribution of these variables would be similar for treated and control subjects in the same stratum (Rosenbaum and Rubin 1983).

The optimal stratification—that is, the stratification that makes treated and control subjects as similar as possible within strata—is a full matching as described in Sect. 2.1.4. An optimal full matching, hence also an optimal stratification, can be determined using network optimization, as discussed in Sect. 2.1.5. See Rosenbaum (1995) for proof and specifics.

2.3 Model-based Adjustments

Matched sampling and stratification compare treated subjects directly to actual controls who appear comparable in terms of observed covariates. For instance, as noted in Sect. 1.2.1, Lehman et al. (1987) compared bereaved individuals to matched controls who were comparable in terms of age, gender, income, education, and family structure. In contrast, model-based adjustments, such as covariance adjustment, use data on treated and control subjects without regard to their comparability, relying on a model, such as a linear regression model, to predict how subjects would have responded under treatments they did not receive.

If the model used for model-based adjustment is precisely correct, then model-based adjustments may be more efficient than matching and stratification, producing estimates with smaller standard errors. On the other hand, if the model is substantially incorrect, model-based adjustments may not only fail to remove overt biases, they may even increase them, whereas matching, and stratification are fairly consistent at reducing overt biases (Rubin 1973b, 1979).

Alas, in practice, one can often detect substantial errors in a model, but rarely can have much confidence that the model is precisely correct, so there is rarely a clear choice between efficiency of model-based adjustments when the model is true and the robustness of matching and stratification when it is not. This has led to the use of combined techniques.

In simulation studies, Rubin (1973b, 1979) compared matching, covariance adjustment, and combinations of these techniques, concluding that the use of covariance adjustment within matched pairs was superior to either method alone, being both robust and efficient.

3. Detecting Hidden Bias

3.1 Elaborate Theories

In an observational study, the investigator takes active steps to detect hidden biases—to collect data recording visible traces of unobserved pretreatment differences if they exist. In this, the investigator is aided by an 'elaborate theory,' defined by Sir Ronald Fisher in the following discussion from Cochran (1965, Sect. 6).

> About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: 'Make

your theories elaborate.' The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold.

Consider the study of lead exposures by Morton et al. (1982), discussed in Sect. 1.2.2. Their elaborate theory made three predictions: (a) higher lead levels in the blood of exposed children than in matched control children, (b) higher lead levels in exposed children whose parents had higher exposure on the job, and (c) higher lead levels in exposed children whose parents practiced poorer hygiene upon leaving the factory. Since each of these predictions was consistent with observed data, to attribute the observed associations to hidden bias rather than an actual effect of lead exposure, one would need to postulate biases that could produce all three associations. Additionally, these observations are consistent with other studies of lead exposure. The use of elaborate theories and further examples are discussed in Campbell and Stanley (1963), Cook and Campbell (1979), Campbell (1988), Rosenbaum (1995, Sects. 5–8), and Sobel (1995).

3.2 Two Control Groups

The simplest and perhaps the most common study design to detect hidden bias uses two control groups. Since neither control group received the treatment, systematic differences between the control groups cannot be effects of the treatment and must instead be some form of hidden bias. One seeks two control groups that would differ in their outcomes if a specific unobserved bias were present. In this way, hidden biases should produce a difference between the two control groups, whereas an actual treatment effect should produce relatively similar results in the two control groups and a different result in the treated group.

This is the principle of 'control by systematic variation,' which is due in general to Bitterman (1965), and is applied to observational studies by Campbell (1969) and Rosenbaum, 1995a Sect. 7). If one cannot ensure comparability of treated and control groups with respect to a particular variable, say by matching, because that variable cannot be observed, then one seeks two control groups that are known to differ markedly on that unobserved variable, that is, to systematically vary the unobserved variable. Control by systematic variation is an instance of 'strong inference' in the sense discussed by Platt (1964).

In an interesting study that used two control groups, Zabin et al. (1989) investigated the effects of having an abortion on the education, psychological status, and subsequent pregnancies of black teenage women in Baltimore. In the US, there is currently active debate about whether abortions should be legal, with concern expressed about both the fetus and the mother. The effects—harmful or beneficial—of abortions on young women are therefore of interest, but a moment's thought reveals some serious difficulties in devising an empirical study. A woman may have an abortion only if she is pregnant, in which case, if she does not have an abortion, then with good health she will have a child. An advocate who believed abortions should be legal might wish to compare a teenager having an abortion to a teenager having a child. An advocate who believed abortions should be illegal would probably object to this comparison, saying that it compares two people who are far from comparable, because one has a child and the other does not. If having an abortion is anywhere near as harmful as having an unwanted child, this second advocate might say, then abortions are harmful indeed. The second advocate might say that if abortion was illegal, young women would be more careful about avoiding unwanted pregnancies, so it is more appropriate to compare a young woman who had an abortion to a young woman who was not pregnant. In addition to these ambiguities about what is meant by 'the effect of abortion,' such a study would face powerful selection effects. The decision by a young pregnant woman to have an abortion or have a child may reflect many characteristics, such as the degree to which she is determined to complete high school. Sexually active teenagers who do not use adequate birth control may tend to differ in many ways from those who do not become pregnant.

Zabin et al. (1989) used two control groups derived from 360 young black women who came for pregnancy tests at two clinics in Baltimore. The 'treated' group consisted of women who were pregnant and had an abortion. One control group consisted of women who were pregnant and had a child. The second control group consisted of women whose pregnancy test revealed they were not pregnant. Because the second control group consisted of women who suspected they might be pregnant, it is likely to be less biased than a control group of women who simply were not pregnant. The second control group did not have to decide whether to have an abortion or a child, so it is not subject to the self-selection effect that formed the treated group and the first control group. If one were viewing nothing but a selection effect based on the decision to have an abortion or a child, then the second control group should be found between the treated group and the first control group. Many outcomes were studied with varied results. As one illustration, consider 'negative educational change' two years after the pregnancy test. In the abortion group, 17.8 percent had a negative educational change, whereas in the child bearing group it was 37.3 percent and in the negative pregnancy test group it was 37.4

percent. Despite the ambiguities in this study, it would be hard to argue that this pattern of results supports a claim that abortions cause negative educational change at two years. The two control groups together provide more information about hidden biases than either group provides alone, in part because the control groups systematically vary certain important unobserved differences.

4. *Appraising Sensitivity to Hidden Bias*

The analytical adjustments discussed in Sect. 2 can often remove overt biases accurately recorded in the data at hand, but there is typically a legitimate concern that treated and control groups differed prior to treatment in ways that were not recorded. A sensitivity analysis asks how such hidden biases might alter the conclusions of the study.

The first formal method of sensitivity analysis was developed to aid in appraising the effects of cigarette smoking on human health. Responding to objections that smoking might not cause lung cancer, but rather that there might be a genetic predisposition both to smoke and to develop lung cancer, Cornfield et al. (1959, p. 40) wrote:

> … if cigarette smokers have nine times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers must be at least nine times greater than among nonsmokers.

This statement was an important conceptual advance. It is, of course, commonly understood that association does not imply causation—that any observed association can be explained by a hypothetical argument, perhaps an implausible hypothetical argument, postulating an unobserved variable. Although this common understanding is always correct, it contributes little to our appraisal of an observational study, because it makes no reference to what was actually observed in this empirical investigation, and no reference to scientific knowledge about the phenomenon under study. A sensitivity analysis replaces the statement—'association does not imply causation'—by a specific statement about the magnitude of hidden bias that would need to be present to explain the associations actually observed. Strong associations in large studies can only be explained by large, perhaps implausible, biases. Aspects of the method of Cornfield et al. (1959) are discussed by Greenhouse (1982). Although important as a conceptual advance, this method of sensitivity analysis is limited in that it ignores sampling error produced by having a finite sample rather than the population as a whole, and it is applicable only to binary outcomes.

A related method addresses sampling error and may be used with outcome measures of all kinds. Observing that in a randomized experiment, subjects in the same stratum or matched set have the same probability of receiving the treatment rather than the control, the sensitivity analysis introduces a sensitivity parameter Γ which measures the degree of departure from random assignment of treatments. Specifically, two subjects who look the same in terms of observed covariates, and so are in the same matched set or stratum, may differ in their odds of receiving the treatment by at most a factor of Γ. In a randomized experiment, $\Gamma = 1$. If $\Gamma = 2$ in an observational study, then if two people look the same in terms of observed covariates, one might be twice as likely to receive the treatment as the other because they differ in terms of a covariate not observed. For each value of Γ, it is possible to place bounds on conventional statistical inferences—e.g., for $\Gamma = 2$, perhaps, the true P-value is unknown, but must be between 0.001 and 0.021. Similar bounds are available for point and interval estimates. One then asks: How large must Γ be before the conclusions of the study are qualitatively altered? If the conclusions change for $\Gamma = 1.2$ then the study is very sensitive to hidden bias—very small departures from a randomized experiment can explain the observed association between treatment and outcome—but if the conclusions do not change for $\Gamma = 9$, then only quite dramatic hidden biases can explain the observed association. A nontechnical introduction to this method of sensitivity analysis is given in Rosenbaum (1991), and a detailed development in Rosenbaum (1995, Sect. 4).

Methods of sensitivity analysis are discussed by Angrist et al. (1996), Copas and Li (1997), Cornfield et al. (1959), Gastwirth (1992), Greenhouse (1982), Lin et al. (1998), Manski (1990, 1995), Marcus (1997), and Rosenbaum (1995).

5. *Reducing Sensitivity to Hidden Bias*

5.1 *The Goal: Sharper Effects, Smaller Biases*

Given the choice, one would prefer conclusions insensitive to hidden biases, so that there is less ambiguity about the effects caused by the treatment. In general, an observational study will be less sensitive to biases of plausible size if the treatment effect size is made clearer, sharper or larger, and if the magnitude of bias that is plausible is made smaller. A diverse collection of strategies used in observational studies may be viewed as efforts to bring about this situation, that is, to observe larger, sharper treatment effects while rendering large biases less plausible. If successful, these strategies would reduce sensitivity to hidden bias.

This section discusses three such strategies, namely choice of circumstances, instrumental variables, and coherent hypotheses.

5.2 Choice of Circumstances

In the earliest stages of planning an observational study, the investigator chooses the circumstances in which the study will be conducted. These choices often affect both the size of the treatment effect and the magnitude of hidden bias that is plausible. As a result, these choices affect the sensitivity of the study to hidden bias. Some illustrations of this use of 'choice' will be briefly mentioned. See Rosenbaum (1999b) for detailed discussion of these cases and others with examples, and see Meyer (1995) and Angrist and Krueger (1998) for related discussion.

The investigator examines a broad research hypothesis, one that makes predictions about innumerable circumstances, but examines that hypothesis in particular circumstances where dramatic effects are anticipated, free of distortions from other factors that affect these same outcomes. The investigator seeks a situation in which there is a genuine control group completely shielded from the treatment, and a treated group that received the treatment at an intense dose. One prefers a treatment that is imposed suddenly, not gradually, and haphazardly, not in direct response to characteristics of the individuals under study.

For instance, in the study in Sect. 1.2.1 by Lehman et al. (1987), the effects of bereavement were studied following sudden deaths of close relatives from car crashes. Moreover, employing standardized criteria, they used only those car crashes for which the victim's car was not responsible, reasoning that alcohol or drug abuse or certain forms of psychopathology might increase the risk of a car crash and also be associated with unfavorable psychological outcomes in the family.

As another example, Angrist and Lavy 1999 studied the effects of class size on academic achievement in carefully chosen circumstances. In Israel, a rule dating back to Maimonides requires class rooms with no more than 40 students, so a class with 41 students is divided in two classes of about 20 students. Whereas in the United States, a class of size 40 is likely to be found in a different economic environment from a class of size 20, in Israel a comparatively haphazard event—the enrollment of an additional child—often separates these two situations.

Still another example is Card's (1990) use of the Mariel Boatlift to study the effects of immigration on labor markets.

5.3 Instrumental Variables

In some circumstances, the treatment assignment itself is subject to hidden biases that are plausibly quite large, but there is another variable, called an instrument, which has no effect on the outcome, but is associated with the treatment. The hope, sometimes realistic, is that instrument is free of hidden bias, or at least that it is affected by hidden biases that are much smaller. In an instrumental variable analysis, the magnitude of the treatment effect is modelled in terms of the treatment received, but the assignment variable is the instrument, not the treatment.

A nice example with a clear theoretical development is given by Angrist et al. (1996). They consider aspects of the health effects of military service, but of course, people who serve in the military differ in many ways from those who do not. Their instrument is the draft lottery in use in the United States between 1970 and 1973. Many people who were drafted did not serve, and many who served were not drafted, so the draft is not at all the same as the treatment under study, namely military service. However, the draft was close to random, so it is not plausible that it is affected by large hidden biases. The instrumental variable analysis models the treatment effect in terms of actual service, but uses the draft as the 'assignment variable' in a sense made precise in Angrist et al. (1996) and in the discussion published with that paper. See also Moffitt's (1996, p. 463) discussion of his 'prosaic economic example.'

It is possible to conduct an exact sensitivity analysis for hidden bias affecting the instrument; see Rosenbaum (1999a). An example presented there concerns the Card and Krueger (1994) study of the effects of increasing the minimum wage in New Jersey and not in neighboring Pennsylvania, where the treatment effect is modelled in terms of the change in wages and the state is used as an instrument.

5.4 Coherent Hypotheses

A coherent hypothesis makes many highly specific predictions about the effect of a treatment on several outcome variables, perhaps in several treatment groups, perhaps involving one or more dose-response relationships. Hill (1965) claimed coherence as an important criterion in appraising evidence that a treatment is the cause of the outcomes with which it is associated.

Coherence is a distinct concept. Unlike sensitivity analyses, coherence asks what the pattern of associations implies about hidden biases, not the degree of sensitivity of a single association. Efforts to detect hidden biases, say using systematic variation with two control groups, are efforts to collect data so that a treatment effect would reveal one pattern of associations while the most plausible hidden biases would reveal a different pattern. Unlike efforts to detect hidden biases, coherence concerns a complex pattern of associations anticipated by a causal theory, without specific reference to the patterns that plausible biases might produce. The everyday metaphor of fitting like a glove—a causal theory that fits data like a glove—is sometimes helpful in distinguishing coherence from other techniques. A hand and a glove are very similar but very unusual shapes—their size, the number of

digits, the pattern of the digit lengths, the single opposed digit, etc.—and these features are not mere repetitions of one another—that there are *five* digits does not entail that *one* digit will oppose the other *four*, or that the *third* digit will be longest etc. Coherence is related to 'pattern matching' as discussed by Campbell (1966), 'triangulation' as discussed by Webb (1966), and 'multiple substantive posttests' as discussed by Shadish and Cook (1999).

Informally, one feels that if numerous predictions of a causal theory are simultaneously confirmed, it should be more difficult to attribute this to hidden bias—the study should be less sensitive to bias. This is sometimes true, sometimes false, but when it is true, the reduction in sensitivity to bias can, in principle, be substantial. See Rosenbaum (1997) and the references there for discussion. The pattern of responses that does most to reduce sensitivity to bias entails weak associations among multiple outcomes each of which is strongly associated with the treatment (Rosenbaum 1997 Sect. 4.3.4).

See also: Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Panel Surveys: Uses and Applications; Quasi-Experimental Designs; Tocqueville, Alexis de (1805–59)

Bibliography

Angrist J D, Imbens G W, Rubin D B 1996 Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* **91**: 444–69

Angrist J D, Krueger A B 1998 Empirical strategies in labor economics. Working paper #401, Industrial Relations Section, Princeton University. To appear in the *Handbook of Labor Economics*

Angrist J D, Lavy V 1999 Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* **114**: 533–75

Bergstralh E J, Kosanke J L, Jacobsen S J 1996 Software for optimal matching in observational studies. *Epidemiology* **7**: 331–2

Bitterman M 1965 Phyletic differences in learning. *American Psychologist* **20**: 396–410

Boruch R E 1997 *Randomized Experiments for Planning and Evaluation: A practical guide*. Sage Publications, Thousand Oaks, CA

Campbell D R 1966 Pattern matching as an essential in distal knowing. In: Hammond K R (ed.) *The Psychology of Egon Brunswik*. Holt, Rinehart and Winston, New York, pp. 81–106

Campbell D T 1969 Prospective: artifact and control. In: Rosenthal R L, Rosnow R L (eds.) *Artifact in Behavioral Research*. Academic Press, New York

Campbell D T 1988 *Methodology and Epistemology for Social Science: Selected Papers*. University of Chicago Press, Chicago, pp. 315–33

Campbell D T, Stanley J C 1963 *Experimental and Quasi-experimental Designs for Research*. Rand McNally, Chicago

Card D 1990 The impact of the Mariel Boatlift on the Miami labor market. *Industrial and Labor Relations Review* **43**: 245–57

Card D, Krueger A B 1994 Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* **84**: 772–93

Cochran W G 1965 The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A*. **128**: 134–55

Cochran W G 1968 Effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24**: 295–313

Cook T D, Campbell D T 1979 *Quasi-experimentation*. Rand McNally College Publisher, Chicago

Copas J B, Li H G 1997 Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, Series B*. **59**: 55–96

Cornfield J, Haenszell W, Hammond E C, Lilienfeld A M, Shimkin M B, Wynder E L 1959 Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22**: 173–203

Cox D R, Snell E J 1989 *Analysis of Binary Data*. Chapman and Hall, London

Firth D, Payne C, Payne J 1999 Efficacy of programmes for the unemployed: Discrete time modelling of duration data from a matched comparison study. *Journal of the Royal Statistical Society, Series A*. **162**: 111–20

Friedman L M, Furberg C D, DeMets D L 1998 *Fundamentals of Clinical Trials*. Springer, New York

Gastwirth J L 1992 Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* **33**: 19–34

Greenhouse S W 1982 Jerome Cornfield's contributions to epidemiology. *Biometrics Suppl*. **38**: 33–45

Hill A B 1965 Environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine* **58**: 295–300

Lehman D, Wortman C, Williams A 1987 Long-term effects of losing a spouse or a child in a motor vehicle crash. *Journal of Personality and Social Psychology* **52**: 218–31

Lin D Y, Psaty B M, Kronmal R A 1998 Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54**: 948–63

Manski C F 1990 Nonparametric bounds on treatment effects. *American Economic Review* **80**: 319–23

Manski C 1995 *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA

Marcus S 1997 Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *Journal of Educational and Behavioral Statistics* **22**: 193–202

Meyer B D 1995 Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics* **13**: 151–61

Meyer M M, Fienberg S E (eds.) 1992 *Assessing Evaluation Studies: The Case of Bilingual Education Strategies*. National Academy Press, Washington, DC

Moffitt R A 1996 Comment on "Identification of causal effects using instrumental variables" by Angrist, Imbens and Rubin. *Journal of the American Statistical Association* **91**: 462–5

Morton D E, Saah A J, Silberg S L, Owens W, Roberts M, Saah M 1982 Lead absorption in children of employees in a lead-related industry. *American Journal of Epidemiology* **115**: 549–55

Piantadosi S 1997 *Clinical Trials: A Methodologic Perspective*. Wiley, New York

Platt J R 1964 Strong inference. *Science*. **146**: 347–53

Rosenbaum P R 1991 Discussing hidden bias in observational studies. *Annals of Internal Medicine* **115**: 901–5
Rosenbaum P R 1995a *Observational Studies*. Springer, New York
Rosenbaum P R 1997 Signed rank statistics for coherent predictions. *Biometrics* **53**: 556
Rosenbaum P R 1999a Using combined quantile averages in matched observational studies. *Applied Statistics* **48**: 63–78
Rosenbaum P R 1999b Choice as an alternative to control in observational studies (with discussion). *Statistical Science* **14**: 259–78
Rosenbaum P R, Rubin D B 1983 The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55
Rubin D B 1973a Matching to remove bias in observational studies. *Biometrics* **29**: 159–83 Correction, 1974 *Biometrics* **30**: 728
Rubin D B 1973b Use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29**: 185–203
Rubin D B 1974 Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**: 688–701
Rubin D B 1979 Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74**: 318–28
Shadish W R, Cook T D 1999 Design rules: More steps toward a complete theory of quasi-experimentation. *Statistical Science* **14**: 294–300
Smith H L 1997 Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* **27**: 325–53
Sobel M 1995 Causal inference in the social and behavioral sciences. In: Arminger G, Clogg C C, Sobel M E (eds.) *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Plenum Press, New York, pp. 1–38
Webb E J 1966 Unconventionality, triangulation and inference. In: *Proceedings of the Invitational Conference on Testing Problems*. Educational Testing Service, Princeton, NJ, pp. 34–43
Zabin L S, Hirsch M B, Emerson M R 1989 When urban adolescents choose abortion: effects on education, psychological status, and subsequent pregnancy. *Family Planning Perspectives* **21**: 248–55

P. R. Rosenbaum

Obsessive–Compulsive Disorder

1. Clinical Aspects

1.1 Clinical Symptomatology and Course

Obsessive–compulsive disorder (OCD) is a chronic, disabling disorder that is characterized by recurrent obsessions and uncontrolled compulsions. Obsessions are defined as persistent ideas, thoughts, impulses, or images that are experienced as intrusive and inappropriate. Usually patients attempt to ignore or suppress the obsessions, which typically deal with contamination, doubts, or aggressive imaginations. Compulsions are repetitive behaviors (e.g., handwashing, checking locked doors) or mental acts (e.g., counting, repeating words) that the person feels driven to perform in response to an obsession or according to rigid rules. Characteristically, compulsions counteract the anxiety associated with an obsession. Diagnosis by DSM IV also requires that the person recognize that the obsessions or compulsions are excessive or unreasonable; they must also cause marked distress, consume at least one hour a day, or considerably interfere with daily functioning.

OCD typically begins in late adolescence, though onset in childhood is not uncommon. The disorder tends to exhibit a waxing and waning course over the life span, with periods of relative good functioning and limited symptoms punctuated by periods of symptomatic exacerbation. A small minority of subjects exhibits either complete remission of their disorder or a progressive, deteriorating course. There is a high comorbidity with major depression, panic disorder, phobias, and eating disorders. Approximately half of all patients with Tourette's disorder meet criteria for OCD, although less than 11 percent of patients with OCD meet criteria for Tourette's disorder.

1.2 Differential Diagnosis

Obsessive–compulsive (OC) symptomatology is also found in various other psychiatric disorders. It can be difficult to differentiate OCD from major depression, which is sometimes associated with obsessive thoughts. In contrast to OCD, obsessional thinking in depression usually subsides in parallel to the remission of depression. Psychotic symptoms can also provoke obsessive thoughts and compulsive behaviors, but are accompanied by other characteristics of schizophrenia. Principally, OCD patients differ from schizophrenic patients by their knowledge that obsessions and compulsions are inappropriate, absurd, or irrational ('egodystonic'). However, when the condition is very severe the patient might experience his obsessions or compulsions as more ego-syntonic and is transiently not sure about the real risk of danger or the real necessity to perform a certain ritual in order to prevent a negative event. There is a subgroup of OCD patients who are frequently misdiagnosed as schizophrenic due to these 'psychosis-like symptoms'. OCD also displays some similarity to obsessive–compulsive personality disorder, which is associated with an obsessive concern for details, perfectionism, and rigidity. However, a genuine syndrome of obsessions and compulsions is not present in obsessive–compulsive personality disorder. Patients with OCD experience distress connected with their symptoms. By contrast, the traits of patients with obsessive–compulsive personality disorder very often disturb others more than

themselves. Finally, OCD is sometimes misdiagnosed in patients suffering from phobias (e.g., phobic avoidance of any assumed risks to be infected by the AIDS virus). However, phobic patients never develop the kind of rituals which are characteristic of OCD (see *Differential Diagnosis in Psychiatry*).

1.3 Epidemiology

The lifetime prevalence in the US, Canada, and comparable countries has been found to be 2.2 to 2.3 per 100 persons (Kessler et al. 1994); in other countries, prevalence rates were between 1 and 3 percent. Most remarkably, these rates contradict the previous traditional view of OCD as a rare disorder on the basis of published clinical reports (see *Mental Illness, Epidemiology of*).

2. Pathogenesis

There is strong evidence that neurobiological as well as psychological factors play an important role in the pathogenesis of OCD. Phenomenological similarities between focal dystonias and OCD are related to the persistent, perseverative, repetitive and involuntary nature of the symptoms. The occurrence of OC symptoms as part of the postencephalitis Parkinsonian syndrome (Schilder 1938) stimulated the search for a biological link between basal ganglia dysfunction and OC symptomatology. Compulsive behavior and obsessional thinking have been observed after basal ganglia lesions (Laplane et al. 1989, Swedo et al. 1989) and in patients with Gilles de la Tourette's syndrome (Pauls et al. 1986). Accordingly, symptomatic dystonias can be associated with structural brain lesions (Kostic et al. 1996). Patients suffering from spasmodic torticollis (Bihari et al. 1992) or from blepharospasm, conditions which are also related to basal ganglia dysfunction, showed significantly more OC symptoms than normal controls or patients with hemifacial spasms, a condition with similar symptomatology, but caused by an irritation of the facial nerve (Broocks et al. 1998b).

2.1 Neuroanatomical Hypothesis of OCD

Since around 1990 positron emission tomography (PET) and other brain imaging studies have produced further evidence for basal ganglia involvement in the pathogenesis of OCD. It has been postulated that a disturbed striatal filter function causes a disinhibition of cerebral activities generated in orbitalfrontal and cingulate structures (Baxter et al. 1992, Benkelfat et al. 1990, and Schwartz 1998). During an implicit sequence learning task, healthy subjects recruit the right striatum whereas patients with OCD fail to do so and instead exhibit aberrant medial temporal activation that may serve a compensatory function (Rauch et al. 1997). It has been assumed that the striatum may be the principal site of primary pathology in OCD and related disorders, a theory that finds support from

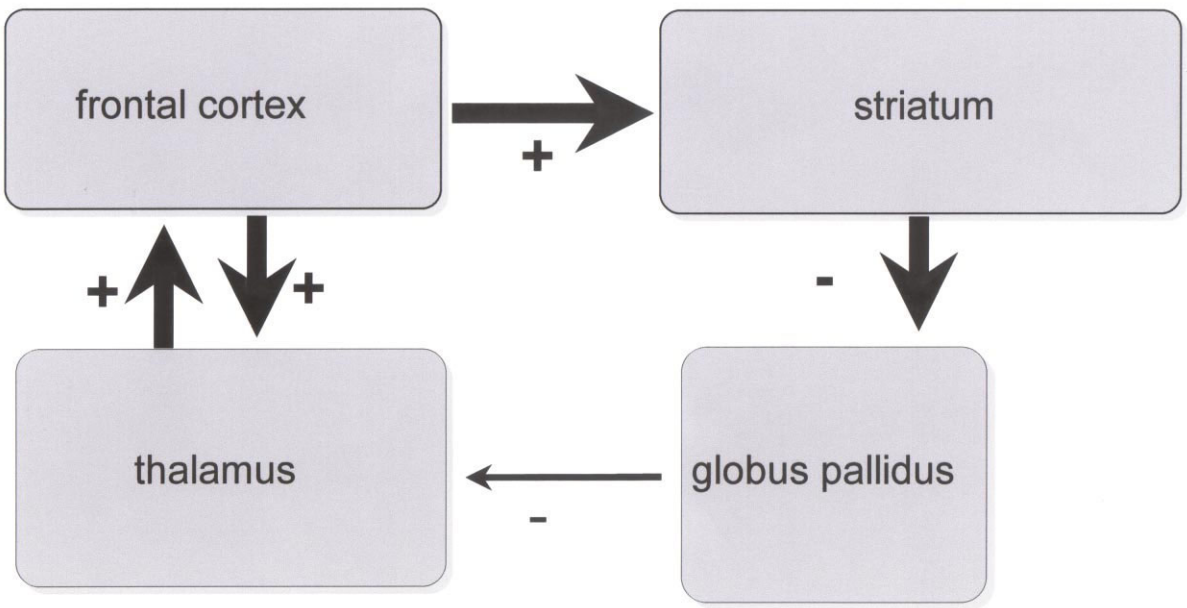

Figure 1
The proposed self-sustaining circuit in OCD.

models of pathogenesis related to autoimmune processes known to cause striatal damage (Swedo et al. 1989). Also, N-acetyl aspartate (NAA) levels, as measured by magnetic resonance spectroscopy (MRS), which are interpreted as indices of early neuronal degeneration, have been found to be reduced within the striatum (Bartha et al. 1998, Ebert et al. 1997).

The pathological increase of metabolic activity in the orbitalfrontal cortex (OCF) demonstrated in PET studies that most OCD patients experience the neurobiological expression of the sensations of anxiety and worry. The caudate nucleus and other striatal structures are stimulated by the OCF. Leading researchers in this field have postulated that in OCD a positive feedback mechanism between the OCF and the striatum is activated, which also involves thalamic structures. Physiologically, inhibitory pathways from the caudate nuclei to the thalamus (via the *globus pallidum*) have the task of keeping this neurosensory circuit in balance. One explanation of the proposed dysbalance in OCD could be that an increased inhibition of the *globus pallidum* results in a desinhibition of the thalamus creating a self-sustaining circuit (Fig. 1). Pharmacological and neurosurgical interventions are obviously able to decrease this neuronal overactivity with the consequence of decreased OC symptoms. Pretreatment metabolic correlation between orbital cortex and caudate and between orbital cortex and thalamus were abolished by successful CBT (Baxter et al. 1992, Schwartz et al. 1996). It is fascinating that psychological interventions, which are primarily mediated by various cortical structures, are also able to affect these orbitofrontal-subcortical neurocircuits.

Neuropsychological testing has revealed that patients with OCD are impaired on measures of spatial working memory, spatial recognition, and motor initiation and execution (Purcell et al. 1998, Schmidtke et al. 1998). In contrast, performance of these tasks by patients with panic disorder or depression did not differ from that of controls.

2.2 Neurochemical Hypothesis of OCD

The most compelling findings suggesting a role of serotonergic pathways in the symptomatology of OCD come from well-replicated treatment studies. Most OCD patients improve during treatment with serotonin (5-HT) reuptake inhibitors including clomipramine, fluoxetine, fluvoxamine, paroxetine, and sertraline, but not with non-5-HT selective antidepressants (Murphy et al. 1996, Pigott and Seay 1999). This observation was replicated in a recent multicenter trial enrolling 166 patients with concurrent OCD and major depression, in which the SSRI sertraline was more effective in reducing depressive and OC symptoms than the primarily norepinephrine reuptake inhibitor desipramin (Hoehn et al. 2000). Using a placebo-controlled design, discontinuation of clomipramine was followed by an increase of OCD symptoms (Pato et al. 1991). Interestingly, coadministration of the 5-HT2 receptor antagonist metergoline to OCD patients being treated with clomipramine was associated with higher OC symptoms in comparison to placebo coadministration (Benkelfat et al. 1989). An exacerbation of OC symptoms after administration of the non-selective serotonergic agonist, m-chlorophenylpiperazine (m-CPP) was observed in several, but not all studies (Broocks et al. 1998a, Charney et al. 1988, Hollander et al. 1992, Zohar et al. 1988).

When oral m-CPP was re-administered during chronic treatment with clomipramine (Zohar et al. 1988) or fluoxetine (Hollander et al. 1992, Pigott et al. 1990), OCD patients did not show an exacerbation of OC symptoms although plasma prolactin and cortisol responses to m-CPP were not attenuated. This suggests that chronic clomipramine or fluoxetine treatment in OCD patients may be associated with the development of an adaptive subsensitivity in one or more serotonergic subsystems (Murphy et al. 1996).

2.3 Genetic Findings

In a relatively small twin study concordance rates for OCD showed a non-significant trend towards higher rates in MZ versus DZ twin pairs (33 versus 7 percent). There are only few family studies of OCD employing adequate research methods (operational diagnostic criteria, controls, and blind assessment of relatives). However, two out of three studies indicate familial aggregation in relatives of patients with OCD versus relatives of normal controls. Discrepant findings might reflect differences in methodology as well as etiologic heterogeneity within the diagnostic category. That is why several researches have now begun to subtype patients by such variables as age at onset, presence of neurological symptoms, and symptom patterns of the disorder. There is a familial and perhaps genetic relationship between Tourette's disorder and chronic motor tics and some cases of OCD. As there is considerable evidence of a genetic contribution to Tourette's disorder this finding also supports a genetic role in a subset of cases of OCDs.

Recently, new evidence has emerged for possible immunological factors in the pathogenesis of OCD. It has long been known that about 10 to 30 percent of patients with rheumatic fever, caused by a streptococcal infection, develop Sydenham's chorea. Over 70 percent of these patients manifest OC symptoms in addition to the choreatiform movements. A monoclonal antibody to B lymphocytes, D8/17, developed using mice immunized with B lymphocytes from a patient with rheumatic carditis, has been used to measure B-cell D8/17 antigen in a small sample of patients with childhood onset of OCD or Tourette's

syndrome. A22 31 patients, 29 percent of whom had pure OCD, showed an elevated percentage of D8/17 antigen, contrasted with only 1 of 21 controls. This antigen might therefore serve as a marker for vulnerability to some forms of childhood OCD. (see *Mental Illness, Genetics of*).

3. Treatment

For decades, OCD has been considered as a fascinating but treatment-refractory mental condition. Conventional pharmacological approaches as well as psychodynamic therapy did not lead to satisfactory improvements of OC symptomatology. The first major advances in the treatment of OCD came from two directions. Clomipramine, which differs from other tricyclic antidepressants by its potent inhibition of serotonin reuptake, proved to exert specific anti-obsessional effects. At the same time, the introduction of special cognitive-behavioral techniques began to continuously improve the prognosis of OCD. In the following years, many studies have shown that behavior therapy, especially exposure with response prevention, and selective serotonin reuptake inhibitors (SSRIs) significantly reduce OC symptoms.

3.1 Psychological Treatment

Exposure/response prevention has been compared to various control treatments including relaxation, anxiety management training, and pill placebo (Kozak et al. 1999), leading to solid evidence for the efficacy of this procedure. Summarizing all controlled trials, average responder rates were 83 percent immediately after treatment and 76 percent after an average follow-up interval of 29 months.

Current exposure/response prevention programs usually combine *in vivo* exercises to feared stimuli with imaginative exposure. Repeated prolonged exposure leads to a psychophysiological habituation process, which also affects irrational beliefs held by the patient. It is crucial that the duration of the exposure exercise continues at least until the patient experiences a decrease of the induced distress. During this time therapists have to be alert to detect and restrain covert neutralizing thoughts (Salkovskis 1999). If possible, a friend or a family member should be taught to support the patient between sessions in order to achieve adequate exposure. Irrespective of the emotional arousal, the patient has the task to describe precisely the experienced emotions and the accompanying cognitions. This description is the basis of the cognitive restructuring which most therapists view as an essential component of successful exposure.

In recent years, cognitive approaches have been proposed for the treatment of OCD, especially obsessional ruminations (Salkovskis 1999). According to the cognitive hypothesis, obsessional problems occur when intrusive cognitions are interpreted as an indication that the person may be responsible for a certain harmful event, which has happened in the past or will happen in the future. Cognitive therapy is thought to work by altering both strongly held beliefs associated with OCD symptoms and interpretations of intrusive mental experiences (Fig. 2). It is also important to help patients to identify and to modify underlying general assumptions, which might represent the roots of the more overt dysfunctional thoughts. As in other anxiety disorders, the general style of therapy is that of guided discovery. Further details of the cognitive treatment of obsessions including the use of loop tapes are described in Salkovskis 1999.

Schwartz has proposed a cognitive-biobehavioral self-treatment for OCD (Schwartz 1996). It integrates the results of neuroimaging research into the cognitive approach based on the literature on the behavioral neurobiology of the basal ganglia. In a four-step process, the OCD patient learns to identify his or her obsessions and compulsions, to label them as such ('It's not me, it is my OCD'), and to attribute them to abnormal activity of the fronto-striatal system. While 'working around OCD symptoms,' patients learn new adaptive responses to intrusive OCD thoughts and urges. They are tutored to become increasingly aware of avoidance behaviors and to use treatment techniques to perform previously avoided behaviors (Schwartz 1996).

In addition to exposure, response prevention and cognitive techniques, the multimodal treatment of OCD also focuses on social deficits, reduced tolerance for intensive emotions, and reduced emotional awareness. In addition, it deals with deficits in problem-solving and self-initiated alternative behaviors as well as interpersonal problems. In severe cases, complex combinations of multimodal interventions are often needed, including marital or family interventions as well as individual or group treatments. Multimodal CBT has been shown to be effective with very low dropout rates in severely ill inpatients (Hohagen et al. 1998) and outpatients (Hand 1998). Furthermore, its long-term efficacy has been demonstrated in follow-up studies in unselected patients (Hand 1998). Whether multimodal behavioral therapy is superior to exposure therapy alone and whether it is appropriate to treat refusers, dropouts, and nonresponders to conventional exposure therapy with response prevention, are still open questions (see *Behavior Therapy: Psychiatric Aspects*).

3.2 Drug Treatment

In comparison to clomipramin, selective inhibitors of serotonin reuptake (SSRI) have the advantage of a very low toxicity and a more favorable side effect

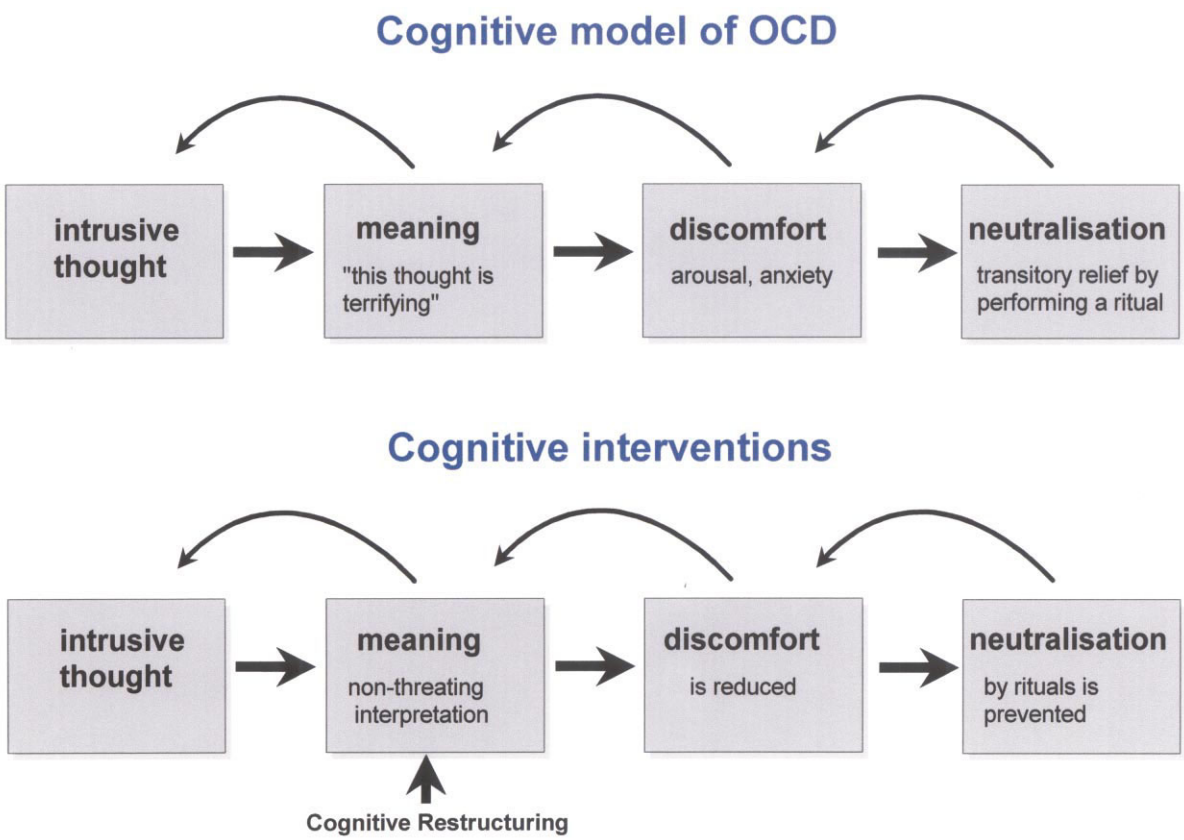

Figure 2
Cognitive therapies.

profile. Consequently, these agents have been the focus of several large-scale, placebo-controlled studies of OCD. Interestingly, very low placebo response rates were observed in most studies. Based on meta-analyses of effect sizes in placebo-controlled trials clomipramine was considered more effective than the SSRIs. However, recent direct comparisons demonstrated similar efficacy between clomipramine and different SSRIs (Pigott and Seay 1999). Pharmacological treatment of OCD usually involves relatively high daily dosages (clomipramine up to 250 mg; fluvoxamine up to 300 mg; fluoxetine, paroxetine, and citalopram up to 80 mg; and sertraline up to 200 mg a day), but some investigators believe that longer duration of treatment rather than higher dosage is essential for response. Ten weeks of treatment trial is currently the minimum recommendation, but it is not unusual to have treatment continue for several months before peak efficacy is reached.

The algorythm in Fig. 3 can be recommended for the pharmacological treatment of OCD patients.

When adequate trials with serotonin reuptake inhibitors, including augmentation strategies, have not been rewarded by a satisfying improvement, it seems justified to try irreversible inhibitors of the monoaminoxydase A such as phenelzyne or tranylcypromine, although the evidence from the literature is not sufficient yet.

When a patient does not respond to a pharmacological or psychological therapy, a repeated motivational and functional analysis might reveal inner resistance and other obstacles which have prevented the patient from therapeutic success so far. (see *Antidepressant Drugs*).

3.3 Combined Pharmacological and Cognitive-Behavioral Treatment (CBT)

Several studies examining the combination of pharmacotherapy with CBT have suggested that the combination of both treatment approaches is associated with only small additional benefits (Cottraux et al. 1993, Foa et al. 1992, Marks et al. 1988, and van Balkom et al. 1998). However, Hohagen et al. 1998 reported that the combination of multimodal CBT

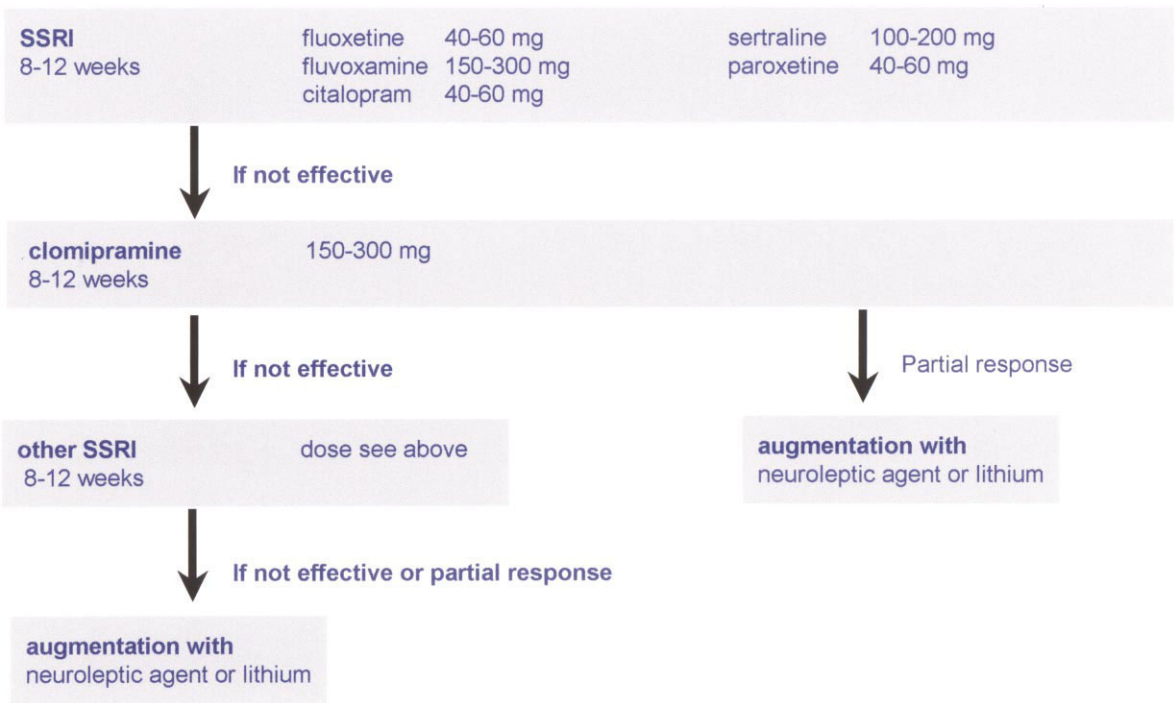

Figure 3
Further procedures in treatment of refractory cases.

with fluvoxamine is superior to the single treatment conditions in patients who suffer predominantly from obsessions or have a high degree of secondary depression. (see *Psychotherapy and Pharmacotherapy, Combined*).

3.4 Prognosis

The presence of schizotypal, borderline, and avoidant personality disorders predicts a less favorable prognosis, whereas duration of illness, sex, age, and type of symptoms had no clear predictive value. In cognitive behavioral therapy the best predictor of good outcome at the end of treatment and of follow-up was early compliance in doing exposure homework (De Araujo et al. 1996).

4. Summary and Future Tasks

Despite all progress in treatment, a considerable percentage of patients experience only a partial unsatisfactory response. At least 25 percent of OCD patients do not comply with exposure therapy/response prevention (Schwartz 1998). In addition to dropouts, 20 percent to 30 percent do not respond sufficiently. Since OCD is a chronic disorder, recurrence of the OCD symptomatology is likely when patients do not adhere to the newly adopted cognitive-behavioral habits due to motivational problems, acute life stress, or depression. With SSRI treatment alone, 40–60 percent of patients have not shown significant progress and have to be classified as non-responders.

In summary, there is still an urgent need to improve the present treatment options for OCD. The few ECT case studies in intractable OCD show considerable benefits up to one year, but these results have to be confirmed in controlled trials. There is initial evidence that repetitive transcranial magnetic stimulation affects acute OC symptoms and could represent a new treatment option in future. In severely ill patients, some people can obtain at least partial relief from OCD by neurosurgery (Jenike 1998). The most frequently performed operation, stereotactic subcaudate tractotomy or capsulotomy, intersects the connections between the frontal lobe and the thalamus and exerts its efficacy on the proposed functional imbalance in OCD between the frontal lobe and subcortical structures. Cingulotomy resulted in significant improvement in 25 to 30 percent of patients previously unresponsive to medication and behavioral treatments (Baer et al. 1995). This rate of improvement seems independent of the changes in anxiety and depression scores.

Pharmacogically, new augmentation strategies should be evaluated. The addition of buspirone or lithium has been of limited success, but larger scale studies are required. Interestingly, the addition of haloperidol in patients, who were refractory to adequate treatment with the serotonin-uptake inhibitor fluvoxamine alone, lead to a clinically and statistically significant improvement (McDougle et al. 1994). This effect was strongly related to a comorbid chronic tic disorder, which is present in up to one third of all OCD patients. Because OCD patients typically require prolonged pharmacotherapy, the risk of tardive dyskinesia has severely limited the use of additional neuroleptic treatment in OCD. The development of the new 'atypical neuroleptics' which—in clinical dosages—have not been found to induce extrapyramidal side effects (EPS) and tardive dyskinesia could lead to a new intriguing treatment option for OCD patients (Bogetto et al. 2000).

The lack of experienced therapists and the cost of multiple individual treatment sessions raise the question whether these interventions could be applied successfully in a group therapy context. The efficacy, acceptance and cost-effectiveness of short-term interventions for patients suffering from less severe forms of OCD needs further improvement. Self-directed exposure using self-help manuals (Foa and Wilson 1991) seems to be effective in a subgroup of OCD patients. More research is required to identify patients who are able to overcome their problems by pure self-help or manual-guided exposure strategies. Another intriguing research field is the question whether neurobiological abnormalities associated with OCD can be modulated by effective treatment strategies. With cognitive-behavioral treatment, significant decreases in caudate glucose metabolism were found in treatment responders compared with nonresponders. It would be of great interest if parameters, which —in comparison to PET findings—can be more easily measured in clinical settings, turn out to be predictors of treatment success or predictors of relapse at the end of a successful treatment.

See also: Antidepressant Drugs; Anxiety and Anxiety Disorders; Behavior Therapy: Psychiatric Aspects; Defense Mechanisms; Differential Diagnosis in Psychiatry; Mental Illness, Epidemiology of; Mental Illness, Genetics of; Personality Theory and Psychopathology; Psychoanalysis in Clinical Psychology; Psychotherapy and Pharmacotherapy, Combined

Bibliography

Baer L, Rauch S L, Ballantine H J, Martuza R, Cosgrove R, Cassem E, Giriunas I, Manzo P A, Dimino C, Jenike M A 1995 Cingulotomy for intractable obsessive-compulsive disorder. Prospective long-term follow-up of 18 patients. *Archives of General Psychiatry* **52**: 384–92

Bartha R, Stein M B, Williamson P C, Drost D J, Neufeld R W, Carr J J, Canaran G, Densmore M, Anderson G, Siddiqui A R 1998 A short echo 1H spectroscopy and volumetric MRI study of the corpus striatum in patients with obsessive-compulsive disorder and comparison subjects. *American Journal of Psychiatry* **155**: 1584–91

Baxter L J, Schwartz J M, Bergman K S, Szuba M P, Guze B H, Maziotta J C, Alazraki A, Selin C E, Ferng H K, Munford P 1992 Caudate glucose metabolic rate changes with both drug and behavior therapy for obsessive-compulsive disorder. *Archives of General Psychiatry* **49**: 681–9

Benkelfat C, Murphy D L, Zohar J, Hill J L, Grover G, Insel T R 1989 Clomipramine in obsessive-compulsive disorder. Further evidence for a serotonergic mechanism of action. *Archives of General Psychiatry* **46**: 23–8

Benkelfat C, Nordahl T E, Semple W E, King A C, Murphy D L, Cohen R M 1990 Local cerebral glucose metabolic rates in obsessive-compulsive disorder. Patients treated with clomipramine. *Archives of General Psychiatry* **47**: 840–8

Bihari K, Hill J L, Murphy D L 1992 Obsessive-compulsive characteristics in patients with idiopathic spasmodic torticollis. *Psychiatry Research* **42**: 267–72

Bogetto F, Bellino S, Vaschetto P, Ziero S 2000 Olanzapine augmentation of fluvoxamine-refractory obsessive-compulsive disorder (OCD): a 12-week open trial [In Process Citation]. *Psychiatry Research* **96**: 91–8

Broocks A, Pigott T A, Hill J L, Canter S, Grady T A, L'Heureux F, Murphy D L 1998a Acute intravenous administration of ondansetron and m-CPP, alone and in combination, in patients with obsessive-compulsive disorder (OCD): behavioral and biological results [In Process Citation]. *Psychiatry Research* **79**: 11–20

Broocks A, Thiel A, Angerstein D, Dressler D 1998b Higher prevalence of obsessive-compulsive symptoms in patients with blepharospasm than in patients with hemifacial spasm. *American Journal of Psychiatry* **155**: 555–7

Charney D S, Goodman W K, Price L H, Woods S W, Rasmussen S A, Heninger G R 1988 Serotonin function in obsessive-compulsive disorder. A comparison of the effects of tryptophan and m-chlorophenylpiperazine in patients and healthy subjects. *Archives of General Psychiatry* **45**: 177–85

Cottraux J, Mollard E, Bouvard M, Marks I 1993 Exposure therapy, fluvoxamine, or combination treatment in obsessive-compulsive disorder: one-year followup. *Psychiatry Research* **49**: 63–75

De Araujo L A, Ito L M, Marks I M 1996 Early compliance and other factors predicting outcome of exposure for obsessive-compulsive disorder. *British Journal of Psychiatry* **169**: 747–52

Ebert D, Speck O, Konig A, Berger M, Hennig J, Hohagen F 1997 1H-magnetic resonance spectroscopy in obsessive-compulsive disorder: evidence for neuronal loss in the cingulate gyrus and the right striatum. *Psychiatry Research* **74**: 173–6

Foa E B, Kozak M J, Steketee G S, McCarthy P R 1992 Treatment of depressive and obsessive-compulsive symptoms in OCD by imipramine and behaviour therapy. *British Journal of Clinical Psychology* **4**: 279–92

Foa E, Wilson R 1991 *Stop Obsessing*. Bantam, Philadelphia, Vol. ISBN 0553353500

Hand I 1998 Out-patient, multi-modal behaviour therapy for obsessive-compulsive disorder. *British Journal of Psychiatry,* (Suppl.) **1998**: 45–52

Hoehn S R, Ninan P, Black D W, Stahl S, Greist J H, Lydiard B, McElroy S, Zajecka J, Chapman D, Clary C, Harrison W 2000 Multicenter double-blind comparison of sertraline and desipramine for concurrent obsessive-compulsive and major

depressive disorders. *Archives of General Psychiatry* **57**: 76–82

Hohagen F, Winkelmann G, Rasche R H, Hand I, Konig A, Munchau N, Hiss H, Geiger K C, Kappler C, Schramm P, Rey E, Aldenhof J, Berger M 1998 Combination of behaviour therapy with fluvoxamine in comparison with behaviour therapy and placebo. Results of a multicentre study. *British Journal of Psychiatry*, (Suppl.) 1998: 71–8

Hollander E, DeCaria C M, Nitescu A, Gully R, Suckow R F, Cooper T B, Gorman J M, Klein D F, Liebowitz M R 1992 Serotonergic function in obsessive-compulsive disorder. Behavioral and neuroendocrine responses to oral m-chlorophenylpiperazine and fenfluramine in patients and healthy volunteers. *Archives of General Psychiatry* **49**: 21–8

Jenike M A 1998 Neurosurgical treatment of obsessive-compulsive disorder. *British Journal of Psychiatry*, (Suppl.) 1998: 79–90

Kessler R, McGonagle K, Zhao S, Nelson C B, Hughes M, Eshleman S, Wittchen H U, Kendler K S 1994 Lifetime and 12-month prevalence of DSM-II-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychology* **51**: 8–19

Kostic V S, Stojanovic-Svetel M, Kacar A 1996 Symptomatic dystonias associated with structural brain lesions. *Canadian Journal of Neurological Sciences* **23**: 53–6

Kozak M J, Liebowitz M R, Foa E B 1999 Cognitive Behavior therapy and pharmacotherapy for OCD. In: Goodman W K (ed.) *Obsessive-Compulsive Disorder: Contemporary Issues in Treatment*. Harper, Washington, DC

Laplane D, Levasseur M, Pillon B, Dubois B, Baulac M, Mazoyer B, Tran Dinh S, Sette G, Danze F, Baron J C 1989 Obsessive-compulsive and other behavioural changes with bilateral basal ganglia lesions. A neuropsychological, magnetic resonance imaging and positron tomography study. *Brain* **112**: 699–725

Marks I M, Lelliott P, Basoglu M, Noshirvani H, Monteiro W, Cohen D, Kasvikis Y 1988 Clomipramine, self-exposure and therapist-aided exposure for obsessive-compulsive rituals. *British Journal of Psychiatry* **152**: 522–34

McDougle C J, Goodman W K, Leckman J F, Lee N C, Heninger G R, Price L H 1994 Haloperidol addition in fluvoxamine-refractory obsessive-compulsive disorder. A double-blind, placebo-controlled study in patients with and without tics. *Archives of General Psychiatry* **51**: 302–8

Murphy D L, Greenberg B, Altemus M, Benjamin J, Grady T, Pigott T 1996 The Neuropharmacology and Neurobiology of Obsessive-compulsive Disorder. In: Westenberg H G M (ed.) *Advances in the Neurobiology of Anxiety Disorders*. John Wiley & Sons Ltd., pp. 279–97

Pato M T, Zohar-Kadouch R, Zohar J, Murphy D L 1991 Return of symptoms after discontinuation of clomipramine in patients with obsessive-compulsive disorder. *American Journal of Psychiatry* **145**: 1521–5

Pauls D L, Leckman J F, Towbin K E, Zahner G E, Cohen D J 1986 A possible genetic relationship exists between Tourette's syndrome and obsessive-compulsive disorder. *Psychopharmacol Bulletin* **22**: 730–3

Pigott T A, Pato M T, Bernstein S E, Grover G N, Hill J L, Tolliver T J, Murphy D L 1990 Controlled comparisons of clomipramine and fluoxetine in the treatment of obsessive-compulsive disorder. Behavioral and biological results. *Archives of General Psychiatry* **47**: 926–32

Pigott T A, Seay S M 1999 A review of the efficacy of selective serotonin reuptake inhibitors in obsessive-compulsive disorder. *Journal of Clinical Psychiatry* **60**: 101–6

Purcell R, Maruff P, Kyrios M, Pantelis C 1998 Neuropsychological deficits in obsessive-compulsive disorder: a comparison with unipolar depression, panic disorder, and normal controls. *Archives of General Psychiatry* **55**: 415–23

Rauch S L, Savage C R, Alpert N M, Dougherty D, Kendrick A, Curran T, Brown H D, Manzo P, Fischman A J, Jenike M A 1997 Probing striatal function in obsessive-compulsive disorder: a PET study of implicit sequence learning. *Journal of Neuropsychiatry and Clinical Neurosciences* **9**: 568–73

Salkovskis P M 1999 Understanding and treating obsessive-compulsive disorder. *Behavioral Research and Therapy* S29–52

Schilder P 1938 The organic background of obsessions and compulsions. *American Journal of Psychiatry* **94**: 1397–414

Schmidtke K, Schorb A, Winkelmann G, Hohagen F 1998 Cognitive frontal lobe dysfunction in obsessive-compulsive disorder. *Biological Psychiatry* **43**: 666–73

Schwartz J M 1996 *Brain Lock*. Harper Collins, New York

Schwartz J M 1998 Neuroanatomical aspects of cognitive-behavioural therapy response in obsessive-compulsive disorder. An evolving perspective on brain and behaviour. *British Journal of Psychiatry*, (Suppl.) 1998: 38–44

Schwartz J M, Stoessel P W, Baxter L J, Martin K M, Phelps M E 1996 Systematic changes in cerebral glucose metabolic rate after successful behavior modification treatment of obsessive-compulsive disorder. *Archives of General Psychiatry* **53**: 109–13

Swedo S E, Rapoport J L, Cheslow D L, Leonard H L, Ayoub E M, Hosier D M, Wald E R 1989 High prevalence of obsessive-compulsive symptoms in patients with Sydenham's chorea. *American Journal of Psychiatry* **146**: 246–9

van Balkom A J, de Haan E, van Oppen P, Spinhoven P, Hoogduin K A, van Dyck R 1998 Cognitive and behavioral therapies alone versus in combination with fluvoxamine in the treatment of obsessive compulsive disorder. *Journal of Nervous and Mental Disease* **186**: 492–9

Zohar J, Insel T R, Zohar Kadouch R C, Hill J L, Murphy D L 1988 Serotonergic responsivity in obsessive-compulsive disorder. Effects of chronic clomipramine treatment. *Archives of General Psychiatry* **45**: 167–72

A. Broocks and F. Hohagen

Occidentalism

As an inherently fluid geohistorical category that expresses varying understandings of the 'West,' 'Occidentalism' is a polysemic 'conceptual monster' (Latouche 1996, p. 7) that conjures up a sense of unity out of shifting geographical and historical referents. In light of the West's role as a colonial and imperial power, renderings of Occidentalism are inherently normative and comparative and are thus exceptionally susceptible to political positioning and ideological perspective. In ordinary use Occidentalism refers to the 'customs, institutions, characteristics, ways of thinking etc. of Western Nations' (OED). In scholarly works influenced by critiques of *Eurocentrism* and of *Orientalism*, Occidentalism refers to the West's self-representation, to other societies' representations of the expanding West, or to representations of cultural difference framed in

terms of Western political epistemologies. Occidentalism is thus an imprecise term that denotes a wide range of representations of the West as well as Occidentalist modalities of representation.

1. The Rising West and its Hegemonic Representation

1.1 Westernization, Colonialism, and European Hegemony

Occidentalism's multiple renderings have reflected the West's imperial history and sustained or contested its rise to global dominance. Hegemonic accounts locate the West's origins in ancient Greek civilization, even if some interpretations extend its roots deeper into Egyptian and Mesopotamian cultures. They present Rome's assimilation of Greek culture and the expansion of the Roman empire—its project to create 'one world'—as initiating a process of westernization or occidentalization.

Through the rise and fall of the Holy Roman Empire and until the end of the Middle Ages, westernization and the West meant Christianization and Christendom. Concentrated in Western Europe, Christendom was overshadowed by the dynamic expansion of Islam after the seventh century, and it had no influence over the empires and civilizations of Asia, Africa, and what came to be known as the Americas. Occidentalism thus designated a rather confined political and ideological formation.

The colonization of the Americas was a turning point in the making of the modern West and in the occidentalization of the world (O'Gorman 1958; Dussel 1995). When Columbus landed in what Spain called the 'Occidental Indies,' Europe was still on the margins of the world's major civilizations in the Middle East and China, and Islam was the most expansive of the major religions. The Spanish and Portuguese colonization of the Americas made it possible to think for the first time in global terms and to represent Europe as the center of the world. Supported by the labor and wealth of the Americas, the Habsburgs constructed the most extensive empire the world had ever known, 'the empire where the sun never sets,' stretching from the Americas to Eastern Europe. This period of Iberian dominance saw the rise of Occidentalism as a legitimating discourse of imperial rule that concerned at once the evangelization and assimilation of America's native populations and the making of Spain and Portugal as imperial powers.

The process of *transculturation* (Ortiz 1940) involving Europe, America, and Africa shifted power in the old world toward the Mediterranean and eventually to the Atlantic. Following the lead of Spain and Portugal, England, France, and Holland competed for control of the Americas, participating in the subjugation and conversion of its native populations as well as in the massive enslavement of African peoples and their forced relocation to the new world. Between the sixteenth and eighteenth centuries, through momentous transformations in politics (the English, US, French and Haitian revolutions), in culture (the Enlightenment), and economic relations (the expansion of global trade and production), power shifted within Europe towards its northwestern nations. By the end of the nineteenth century, their development as nation-states and their colonial expansion in Asia and Africa, conditioned by their earlier European colonial experience in the Americas, consolidated Europe's position as the world's dominant capitalist center.

During its rise as a global power, Europe fashioned itself as the apex of civilization in opposition to its overseas territories. It organized its rule both at home and abroad through taxonomies of cultures intimately linked with gender, class, and racial hierarchies. Since the nineteenth century and up to World War I, Occidentalism's core referent was northern Europe, even as it included the United States as an emerging capitalist power with imperial interests in the Caribbean, Latin America, and the Pacific. During this period of capitalist expansion, nation-state formation, and high colonialism, dominant renderings of Occidentalism were framed in terms of evolutionary schemes drawn from the natural and social sciences. The Occident, which during the early colonial period had fashioned itself in relation to 'savages' located in distant lands, was now defined in relation to 'primitives' who were projected backward into the distant past (Fabian 1983; Mignolo 1998; Trouillot 1991).

1.2 Westernization, US Hegemony, and Neoliberal Globalization

European global hegemony came to a crisis in the twentieth century as a result of two world wars, decolonization, and the consolidation of the United States as the dominant world power and new icon of the West. After World War II, the pursuit of competing visions of independent national *development* turned colonial and postcolonial areas, designated as the 'third world,' into a battlefield in a global struggle between the major socialist nations or 'second world' and the advanced capitalist powers or 'first world.' In the context of the Cold War, the West (the first world), which came to include ascendant Japan, referred centrally to the 'free (capitalist) world' led by the United States, and the East, whose basic referent had been the 'exotic' cultures of the Orient, designated the 'totalitarian (socialist) world' behind the 'iron curtain.' This 'three-world scheme' (Pletsch 1981) has ceased to reflect the power relations resulting from the fall of Soviet socialism in 1989 and the widespread acceptance of the capitalist market as the foundation of social rationality and the source of new trans-

national political and economic entities. While the fluid boundaries that had defined the West in relation to the rest of the world have been further detached from their customary geographical referents, Western economic and political power has become more concentrated and centralized through transnational networks that link dominant sectors across the globe. In this more abstract political cartography, the West has been variously identified as a cultural formation involved in a 'clash of civilizations' in an era of attenuated political conflicts and renewed ancestral cultural hostilities (Huntington 1993), or as the less visible imperial center of subtle modalities of westernization disguised as globalization (Amin 1997).

2. *After Orientalism: Occidentalism within Academia*

The West's rise to global dominance has entailed the hegemony of Western social thought. It placed 'Europe' or the 'West' not just at the center of history, but also as the source of the theories and academic disciplines in terms of which global and local histories have been narrated (Wallerstein et al. 1996; Chakrabarty 1992). A variety of oppositional Occidentalisms, reflecting different conceptions of the expanding West, have contested Western supremacy. For example, at the turn of the nineteenth century, for German romantics Occidentalism designated the soulless cultures of England and France, and for Latin American *Arielistas* it referred to the crass materialism of the United States. After WWII, for many Third World intellectuals and activists, decolonization entailed the rejection of Occidentalism as the expression of an imperial culture. At the close of the twentieth century, for Middle Eastern fundamentalists, Occidentalism means the decadent culture of a political rival, and for nationalist Japanese modernizers it denotes a civilization to be both emulated and counterposed to their own valued traditions. From different locations, struggles against Western supremacy have provided critical visions of the West and drawn attention to the connection between the West's political centrality and the Eurocentric disposition of its epistemological order. Informed by this critique and conditioned by the worldwide erosion of modernist projects in the last quarter of the twentieth century, the examination of Western epistemologies and conceptions of historical progress has been reflected in developments in social theory associated with poststructuralism, postmodernism, and the linguistic turn.

2.1 *Occidentalism and Orientalism: Inversion, Self-reflection, Critique*

Since the 1970s, in intellectual circles influenced by these transformations, critics of Eurocentrism have described in relational terms the mode of representation the West exerts on the East. Through the influence of Said's (1978) pathbreaking discussion of Orientalism as a form of discourse about the Orient rooted in Western epistemologies and political dominance, Orientalism has come to refer to eurocentric and stereotypical representations of any culture.

Reversing the gaze, scholars have generally used Occidentalism as the inverse of Orientalism to designate the study of the West (Hanafi 1991) as well as simplified or stereotypical renderings of the West by non-Western societies. In one of the earliest uses of Occidentalism understood as the opposite of Orientalism, Nader (1989) discusses how Muslim men deploy Western conceptions of women to subjugate women in their own societies. Along similar lines, Chen distinguishes 'official Occidentalism,' by which he means a negative image of the West used by the Chinese socialist government to control its own populations, from 'anti-official' Occidentalism, which refers to counter-discourses of Western superiority used by the Chinese intelligentsia to support their political opposition to the government (Chen 1992, p. 691; for a related oppositional use of Occidentalism, see Findley 1998).

Carrier coins the category 'ethno-Occidentalism' to refer to 'the essentialist rendering of the West by members of alien societies,' and reserves the term 'Occidentalism' for 'the essentialist rendering of the West by Westerners.' Regarding the 'essentialization' of concepts as inherent to 'the way Westerners and probably most people think,' he argues that relating representations of otherness to those of the West would lead to a better understanding of cultural representations of difference (1992, pp. 198–207). Armstrong (1990) uses Occidentalism to refer to the formation of specific forms of racialized and gendered Western selves as the effect of Orientalist representations upon Western societies. Reflecting on its use in scholarly works, Ning (1997, p. 62) suggests that Occidentalism, understood as the counterpart of Orientalism, is an even more imprecise and indeterminate term, a 'quasi-theoretical concept' that has not led to an independent discipline like Oriental studies. In a discussion that focuses on the genealogy of Occidentalism in the Americas, Mignolo (1998) shows how Latin American thinkers have used this term as a critical category to refer to Western imperial culture as well as to the project to transcend it, as in Fernández Retamar's concept of 'postoccidentalism' and Coronil's call to go 'beyond Occidentalism.'

2.2. *Occidentalism as an Imperial Epistemology*

Treating the partiality of imperial representations of cultural difference as the effect of historically specific unequal power relations, Coronil (1996, pp. 56–58) argues that Occidentalism is an imperial epistemology that expresses a constitutive relationship between

Western constructions of difference and Western worldwide dominance. He develops Occidentalism as an analytical category that links the production of Western representations of alterity to the implicit constructions of selfhood that underwrite them. As an ensemble of representation practices, Occidentalism consists of accounts which produce polarized and hierarchical conceptions of the West and its others and makes them central figures in narratives of global and local histories that separate the world's components into bounded units; disaggregate their relational histories; turn difference into hierarchy; naturalize these representations; and thus participate in the reproduction of asymmetrical power relations. These operations are apparent in three dominant Occidentalist modalities of representation: the dissolution of the other into the self; the incorporation of the other into the self; and the destabilization of the self by the other. The subordination of geography to history in these modes of representing modernity contributes to the narration of local histories in terms of the West's alleged universal history and thus to obscuring the participation of non-Western societies in the making of the modern world. In this usage, Occidentalism is an analytical category that refers to representations of cultural difference framed in terms of Western political epistemologies rather than a term that designates characterizations of the West; it refers to the conditions of possibility of Orientalism, not to its symmetrical counterpart.

2.3 Occidentalism and Globalization

The conditions that have enabled Western representations of cultural difference in the modern world in terms of radical alterity, subordination, and exoticism include colonialism and imperialism, a hierarchical system of nation-states, capitalism as a global mode of production of regions, persons, and things, a division of the person into private individual and public citizen, faith in secular science, and an expansive *universalism* supported by the sciences and by ideologies of progress. At the onset of a new millennium these conditions are being modified by transformations epitomized by the hegemony of neoliberal globalization. The intensified commodification of social life and the breakdown of commodities into bundles of risk spread across space and time have heightened the significance of the market as an organizing principle of geopolitical and cultural differences. Critical accounts of globalization treat it as both a unifying and divisive process that involves a dialectic of globalization and localization, universalization and particularization, spatialization and temporalization, and cultural homogenization and ethnic diversification. In contrast, hegemonic discourses of globalization emphasize unity over difference and produce an image of an increasingly integrated globe no longer radically divided between the West and its others, thus masking persisting operations of othering that continue to take place through various forms of naturalization such as the racialization of class distinctions and the essentialization of cultural differences.

Insofar as discourses of neoliberal globalization emphasize the potential equality and universality of humankind through the market, they imply a shift of focus in the representation of cultural difference from alien subjects located outside Western centers to subordinated populations dispersed across the globe. As the West seems to dissolve into the market and westernization appears as globalization, political projects to civilize alien others are being recast as self-induced processes of incorporation by willing citizens-consumers. In contrast to Occidentalist modalities of representation based upon a binary opposition between a superior Western self and its subordinate others, prevailing discourses of globalization appear as a compelling but circuitous form of Occidentalism whose power derives from its capacity to obscure the presence of the West and the operations through which it constitutes itself in opposition to others. Its mode of constructing cultural difference through subalterity rather than alterity reflects conditions that continue to contradict ideals of equality and universality, and yet open up the possibility for their fuller realization (Coronil 2000).

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Ethnocentrism; Globalization and World Culture; Hegemony: Anthropological Aspects; Imperialism, History of; Orientalism; Western European Studies: Culture; Western European Studies: Religion; Western European Studies: Society.

Bibliography

Amin S 1997 *Capitalism in the Age of Globalization. The Management of Contemporary Society*. Zed Books, London

Armstrong N 1990 Occidental Alice. *Differences. Journal of Feminist Cultural Studies* **2**(2): 3–40

Carrier J 1992 Occidentalism: The world turned upside-down. *American Ethnologist* **19**(2): 195–212

Chakrabarty D 1992 Postcoloniality and the artifice of history: Who speaks for 'Indian' pasts? *Representations* **37**: 1–37

Chen X 1992 Occidentalism as counterdiscourse: 'He Shang' in Post-Mao China. *Critical Inquiry* **18**(4): 686–712

Coronil F 1996 Beyond occidentalism: Toward non-imperial geohistorical categories. *Cultural Anthropology* **11**(1): 51–87

Coronil F 2000 Towards a critique of globalcentrism: Speculations on capitalism's nature. *Public Culture* **12**: 351–74

Dussel E 1995 *El encubrimiento del otro. Hacia el orígen del mito de la modernidad*. Nueva Utopía, Madrid

Fabian J 1983 *Time and the Other. How Anthropology Makes its Object*. Columbia University Press, New York

Findley C V 1998 An Ottoman occidentalist in Europe: Ahmed Midhat meets Madame Gulnar, 1889. *American Historical Review* **103**(1): 15–49

Hanafi H 1991 *Muqaddim ah fi ilm al-istighrab*. al-Dar al-Fanniyah, Cairo, Egypt
Huntington S 1993 The clash of civilizations? *Foreign Affairs* **72**(3)
Latouche S 1989 *L'occidentalization du monde: Essai sur la signification, la portée et les limites de l'uniformisation planétaire*. Editions la Découverte [1996. *The Westernization of the World*. Polity Press, Cambridge, UK]
Mignolo W 1998 Posoccidentalismo: el argumento desde la América Latina. In: Castro-Gómez S, Mendieta E (eds.) *Teorías sin disciplinas*. Miguel Angel Porrúa, México
Nader L 1989 Orientalism, occidentalism, and the control of women. *Cultural Dynamics* **2**: 323–55
Ning W 1997 Orientalism versus occidentalism? *New Literary History* **28**: 57–67
O'Gorman E 1958 *La invención de América: la universalización de la cultura occidental*. Universidad Nacional Autónoma de México, México
Ortiz F 1940 *Contrapunteo cubano del trabaco y el azúcar*. La Habana [1995. *Cuban Counterpoint: Tobacco and Sugar*. Duke University Press, Durham NC]
Pletsch C 1981 The three worlds, or the division of social scientific labor, circa 1950–1975. *Comparative Studies in Society and History* **23**(4): 565–90
Said E 1978 *Orientalism*. Pantheon, New York
Trouillot M-R 1991 Anthropology and the savage slot. In: Fox R G (ed.) *Recapturing Anthropology*. School of American Research Press, Santa Fe, NM
Wallerstein I et al. 1996 *Open the Social Sciences. Report of the Gulbenkian Commission on the Restructuring of the Social Sciences*. Stanford University Press, Stanford, CA

F. Coronil

Occipital Lobe (Visual Cortex): Functional Aspects

The occipital lobe, at the back of the brain, is one of the four lobes into which each cerebral hemisphere is conventionally divided. The lobes are defined purely by anatomical landmarks, not functionally. However, the occipital lobe has a certain functional unity in that, so far as we know, it is entirely concerned with processing visual information. How this processing works at the neural level, and how it is organized structurally, provides some insights about the mechanisms of visual perception (see *Visual Perception, Neural Basis of*). Much of this knowledge comes from studying the occipital lobe of the macaque, but the human occipital lobe appears generally closely similar. It maintains a basic spatial organization of visual information, but also extracts and organizes visual attributes such as local shape, motion, color, and depth. To understand this organization, it is important to appreciate that processing goes on in the cerebral cortex, a thin, folded sheet of tissue across which a set of distinct and specialized visual areas are laid out; but also that the cortex is layered, with different layers serving as the location of inputs and outputs of information for a particular area.

1. *The Retinotopic Projection to Area V1*

The fibers delivering visual information to the cerebral cortex terminate in a region around the occipital pole of each hemisphere, known as striate cortex, primary visual cortex, or area V1. Each incoming fiber, and each individual neuron in V1, carries information about a specific small region of the visual field (its receptive field). These are arranged in an orderly way across the cortex, so that neurons with neighboring receptive fields are close together and overall, area V1 in each hemisphere is laid out like a spatial map of the opposite half of the field of view (see *Topographic Maps in the Brain*). The field is split, with the right half-field represented in the left hemisphere, and vice versa. In this map, the vertical midline of the field is represented furthest back, near the occipital pole, and moving forward across the cortex corresponds to moving horizontally outwards in the field of view. This orderly 'retinotopic' mapping must be functionally important, presumably because it allows neurons serving nearby field locations to interact readily. This means that simply organized, local connections between cells can serve to link the features that form spatially connected objects.

The scale of the retinotopic map is not uniform. Much more information comes from the densely packed cells of the central retina than from an equivalent area in the peripheral field, and this is reflected in a correspondingly larger area of cortex devoted to the central region. In this way, cortical organization provides the greater processing power needed for the fine detail and spatial precision of central vision.

2. *Binocular Organization*

An important function of area V1 is that it is the first point in the visual pathway at which information from the two eyes can be integrated. The fibers from each eye are segregated into alternating bands as they arrive in layer 4 of cortical area V1 (see Sect. 4). The first real interaction between left- and right-eye signals occurs when these converge on cells outside layer 4. This interaction allows both eyes to contribute to a unified visual representation of the world. It also allows neurons in V1 and beyond to register the disparity between the two eyes' views, which serves as the basis for stereoscopic depth perception.

3. *Stimulus Selectivity of Cortical Neurons*

The number of neurons in area V1 is at least a hundred-fold greater than the number of input fibers. This enormous information-processing resource means that V1 does not simply map the visual input: it also transforms it. Each local region of the visual field is represented by an array of neurons, with each neuron responding selectively to a particular type of stimulus.

Most V1 neurons respond to contours, and are selective in several, if not all, of the following respects:

strongest response to a preferred orientation of the contour, with response falling off as the angle is further from this optimum ('orientation tuning');

a preferred type of contour: light bar, dark bar, or edge;

a preferred spatial scale, so that the cell may respond either to the broad pattern of light and dark, or to fine detail;

stronger response to one direction of motion than for the opposite direction;

differences in binocular balance: some cells respond predominantly to the right eye, some to the left, and some only give their best response when both eyes are activated together. In the latter, binocular group, cells are selective for stereoscopic depth, i.e., they respond best to objects that are either nearer, farther, or close to the distance on which the eyes are converged; and

color, with one range of wavelengths preferentially activating the cell and another inhibiting it (sometimes combined with an enhanced response to color contrast).

This array of differently tuned neurons means that the cortex can represent all these diverse properties according to which particular pattern of neurons is active (a population-based 'place code'; see *Visual Perception, Neural Basis of*).

4. *Columnar Organization*

Retinotopic mapping (Sect. 2) implies that within a small region of V1, all the cells deal with a local region of the field of view. The cortex also has a finer-grain spatial organization, reflecting the differences in tuning described in Sect. 3. Connections between neurons are predominantly vertical (i.e., perpendicular to the cortical surface). Presumably as a result of these connections, neurons in a vertical column have similar response properties. In particular, there are columns of cells responding to the same orientation, and columns within which a particular eye predominates. From column to column, there is also a spatial ordering; in a horizontal track across V1, the preferred orientation changes in an orderly sequence. As with the retinotopic organization, this order means that a regular, local pattern of connectivity between cells creates communication between cells responding to contours with the same or similar orientation. Presumably, this gives a structural basis for linking together the segments of a contour that extends across many cells' receptive fields, and so helps to establish the neural representations of shapes and objects.

Thus, at the fine scale, cortical layout is not purely a retinotopic map. However, both retinotopic and orientation-based mapping mean that orderly spatial connectivity between cells can establish communication between cells responding to similar stimuli. The columnar organization has been described here for V1, but it is believed to be a general principle that applies also to the extra-striate areas described in Sect. 6.

5. *Interleaved Pathways through V1 and V2*

The visual pathway leading to V1 is in fact two subpathways, and this division runs through processing in V1 and beyond. One group of cells, the magno cells, have large receptive fields, fast responses, and do not respond selectively to particular colors of light. The parvo cells, in contrast, respond less well to rapidly changing stimuli, but better to fine spatial detail, and are wavelength-selective (e.g., activated by one range of colors and inhibited by others). These two classes of cell appear to separate the information useful for detecting motion on the one hand, from that required for fine detail and for color vision on the other. Correspondingly, the perception of motion and of color seems to work largely independently of each other.

This division is maintained in the cortex. In V1, the fibers of magno cells terminate in a distinct layer (4B) containing many motion-selective cells. Within the parvo input, color-specific information is also anatomically separated into a regular pattern of 'blobs' that can be revealed anatomically by staining cortical tissue. Within these blobs, cortical neurons show color-related responses and are generally not orientation-selective. Thus, in the fine structure of area V1, three different kinds of processing are segregated but interleaved: motion information in layer 4B, color information in the blobs, and spatial pattern (orientation) in the 'interblob' regions.

V1 sends information to a neighboring area, V2, which forms a second retinotopic map. The three types of visual information interleaved in V1 are kept segregated in V2, into three parallel sets of stripes that send their output to different extra-striate areas (Sect. 6). Thus, apparently a major role of areas V1 and V2 is to marshal different kinds of information into separate cortical processing streams.

6. *Specialized Extra-striate Areas*

V1 and V2 are only two of what are now recognized to be a large number of distinct cortical visual areas forming an interconnected hierarchy. Most of the occipital lobe can now be divided into a mosaic of areas, distinguished by:

mapping: each area (at least at the lower levels of the hierarchy) forms a distinct, complete retinotopic map of the visual field;

connectivity: there is an orderly set of projection fibers connecting one area to another and preserving the retinotopic mapping. Connections with 'higher' and 'lower' levels in the hierarchy are made by different

cortical layers, which allow the hierarchical structure to be defined; and

function: each area has a characteristic balance in the type of stimulus selectivity found in its neurons (e.g., color selectivity common and directional motion selectivity rare, or vice versa), leading to the idea that different areas have distinct functional roles in vision. This finding from macaque single-neuron recording studies is supported by functional brain imaging in humans (see *Functional Brain Imaging*). Imaging reveals a set of areas, each of which is activated by specific visual properties (e.g., one area by color differences, a different area by coherent motion in the visual field), and is also most active when the subject performs specific visual tasks (e.g., identifying colors). This fits with findings in some patients with brain lesions, who show a deficit limited to one specific aspect of vision (e.g., a loss of just color or motion processing).

The main areas that can be identified as retinotopic maps have been labeled as V3, V3A, V4 and V5. Together they form a third tier (or possibly, third and fourth tiers) of a processing hierarchy above V1 and V2. Area V5, also known as MT, lies on the boundaries of the occipital and temporal lobes in the human brain. V5 has been especially intensively investigated because it has a very well defined specialization for processing motion information. At the same level, but complementary in its specialization, is area V4, which has the highest proportion of color-selective neurons, a type of selectivity almost absent in V5. V4 also shows processing of more complex aspects of spatial pattern than those found in V1 and V2 neurons.

Areas V4 and V5 represent a further stage in the segregated streams of processing found in V1 and V2 (see Sect. 5), with V5 as the destination of the motion information derived from layer 4B of area V1, and V4 receiving the color and form information derived from the blobs and interblob regions. This division is an aspect of a broader division between 'dorsal' and 'ventral' cortical streams. The dorsal stream consists of areas carrying motion and spatial information to the parietal lobe, where it provides the visual basis for controlling actions. The 'ventral' stream carries pattern and color information required for the identification of objects in the temporal lobe.

Functional anatomy, then, supports a view that the visual process is based on distinct modules with different functions; in particular, that there is a broad separation of the modules which tell us 'what' object we are looking at, from those that tell us 'where' or 'how' we can act upon it. In both streams, the higher levels of the hierarchy beyond the occipital lobe contain multiple, distinct areas, in which the retinotopic organization has generally been lost. It seems to be a general principle that as visual information is progressively transformed and elaborated in the brain, the less relevant is retinal location for the way in which it is represented.

While it is clear that the visual areas can be ordered into a hierarchy, the flow of information is not simply bottom-up. Almost every connection from a lower to a higher area has a corresponding connection in a reverse direction. Thus, in the neural structure, there exists the potential that low-level processing, e.g., in V1 and V2, does not run simply automatically, but is controlled and directed by top-down processes. There is also the potential for iterative loops, by which information is exchanged between lower and higher areas and progressively refined. And although the different modules have distinct functions, there are rich interconnections between them at the same level, including connections between dorsal and ventral streams. Occipital cortex is a modular system, but one in which the modules closely interact to produce integrated visual perception.

See also: Color Vision; Color Vision Theory; Motion Perception Models; Motion Perception: Psychological and Neural Aspects; Topographic Maps in the Brain; Vision, High-level Theory of; Vision, Low-level Theory of; Vision, Psychology of; Visual Memory, Psychology of; Visual Perception, Neural Basis of; Visual Space, Geometry of; Visual System in the Brain

Bibliography

Hubel D H 1988 *Eye, Brain and Vision*. Scientific American Library, New York

Rockland K S, Kaas J H, Peters A 1997 *Cerebral Cortex, Vol. 12: Extrastriate Cortex in Primates*. Plenum Press, Oxford, UK

Sereno M I 1998 Brain mapping in animals and humans. *Current Opinion in Neurobiology* **8**: 188–94

Zeki S 1993 *A Vision of the Brain*. Blackwell, Oxford, UK

O. Braddick

Occupational Health

1. Introduction: Work and Health

During historical periods preceding industrialization, most of the work that humans were involved in was unpaid. The borders between work and leisure were not clear. It was with industrialism that paid work started to play such an important role in human life. For several generations, paid work has been the activity that has generated a family's needs for survival. In this article the emphasis will be on paid work, and most of the discussion will be on the main paid work activity if a subject has several activities that are paid.

Health is not only the absence of illness. It is also a positive state in its own right. Most of the research that has been published, however, regarding work and health has been devoted to the establishment of relationships between certain adverse working conditions and illness. Accordingly most of the article will be on work–illness relationships. But we shall also discuss possible pathways between good working conditions and the promotion of health.

An important basis for our understanding of the effects of working conditions on health has been the physiological stress research which was initiated by Hans Selye (1976). Initially stress researchers were studying acute reactions to energy demanding situations. During later years, they have examined the biological consequences of long-lasting adverse life situations.

2. *The Physiological Basis of the Effects of Work on Health*

Genetic evolution and selection processes are slow. Therefore our biological reactions to stressful situations are likely to be approximately the same as they used to be several thousand years ago. It has been emphasized by many authors (Levi 1972, Frankenhaeuser 1980) that this may not be functional for man in modern societies. It was functional to become strong, courageous, and insensitive to physical pain for a human being that faced dangerous animals. Such situations are uncommon for the modern human. Instead, most of the contemperous dangers have to do with complicated social interactions, and in such situations physical energy mobilization is not functional.

3. *Energy Mobilization and Anabolism at Work*

In Selye's (1976) general adaptation syndrome, 'stress' was seen as the general reaction to a nonspecific challenge or adverse condition. The situation that induced stress was labeled a stressor (factor which induces stress). Since this has central importance to the understanding of stress at work, a short description of the energy mobilization will be given here. The most important biological process is the provision of energy—glucose and free fatty acids enter the blood, and these are used for the immediate production of energy. But there are several parallel phenomena, all of which aid the body in the physical fight or flight. Examples are lowered excretion of water and salt, decreased sensitivity to pain, and decreased inflammatory responses to infections. Since energy mobilization (resulting in elevated blood concentration of glucose and free fatty acid) has the highest priority, anabolism (restorative and regenerative activities in the cells) is down regulated. Anabolism is central to the body's central defense of all the organ systems that need constant rebuilding and restoration. If this goes on for a long time (several months) increased sensitivity to physical and psychological stress in bodily organs is the ultimate result.

Another consequence of long-lasting demands for energy mobilization is that the endocrine systems may change their regulatory patterns. This means that the ability of the body to stop energy mobilization when it is no longer needed may be disturbed or that the counter-regulation (inhibition of energy mobilization) is inhibited. Most processes in the body have a counter-regulatory mechanism that operates in order to inhibit a process that has been stimulated.

4. *Physiological Short- and Long-term Adaptation to Requirements at Work*

The hypothalamo–pituitary–adrenocortical system (HPA axis) is of central importance to the biological arousal mechanism. The steroid hormone cortisol which is released from the adrenal cortex is the most extensively studied 'agent' in this system. Cortisol facilitates biological arousal in most organ systems. Disturbed regulation of the HPA axis has been described as a possible consequence of long-lasting arousal. There are at least three kinds of such regulatory disturbances in the HPA axis, all of which are of relevance to health at work:

(a) Lack of inhibition of arousal. All reactions in the body evoke counter-regulatory reactions. This is true also of energy mobilization which should stop when the threat disappears. Normally cortisol inhibits higher centers in the HPA axis when it has been in the aroused state for some time. When this inhibition does not take place in the normal way, the levels tend to become constantly elevated.

(b) Lack of response to stimulation of the HPA axis. This means that plasma cortisol levels are low and that the normal activation does not take place when the HPA axis is stimulated artificially.

(c) Unstable levels. In this kind of disturbance the plasma cortisol levels may be low at rest, but certain situations—that are specific to the individual and have to do with the symbolic significance of the event—trigger a strong HPA axis reaction that may seem to be out of proportion to the importance of the trigger. This occurs in subjects who have been exposed to extraordinarily stressful situations, such as torture, war, disasters, and incest.

The first kind of regulatory disorder, (a), is typical of Cushing's syndrome (overproduction of corticosteroids including cortisol due to a disease in the adrenal cortex) and could be one alternative regulatory disturbance in psychiatric depression. Psychiatric depression is mostly associated with constantly high cortisol levels although in most cases, however, the circadian rhythm is preserved (Rubin et al. 1987)

The second kind, (b), has been shown to exist in a subgroup of patients with the diagnosis chronic fatigue syndrome or CFS (Demitrack et al. 1991). Since at least one other group of patients, those with fibromyalgia (a disorder characterized by easily fatigued muscles which are painful when fatigued), share some symptoms that are typical of this regulatory disorder with CFS patients, it is possible that part of the myalgia problem is that they also suffer from this kind of disturbance.

The third kind, (c), of regulatory disorder is found in post-traumatic stress disorder (PTSD) (Charney et al. 1993). The key element in the etiology of PTSD seems to be that the long-lasting extreme arousal level increases the number of cortisol receptors on the cell surfaces and also increases their sensitivity. Since the sensitivity is so high, only low concentrations of cortisol are needed to keep the system working at rest. This may be the reason why low serum cortisol levels are found at rest in patients with PTSD.

5. *Unemployment and Stress*

Unemployment is a common phenomenon in all industrialized countries. In most European countries it is presently of the order of 10 percent. Research has tried to establish to what extent unemployment contributes to poor health. At present there is agreement that poor health increases the risk of becoming unemployed in periods of financial crisis but there is also agreement that unemployment *per se* contributes to poor health. Longitudinal studies of subjects who have gone through the different stages of unemployment (Brenner and Levi 1987) or started activities again after a long period of unemployment have shown that several physiological systems are markedly affected by these processes.

For instance, there is an elevation of the secretion of cortisol (as an index of energy mobilization) and prolactin (which mirrors passiveness in a crisis situation). It is accordingly reasonable to assume that health could be affected. There is published research showing that long-lasting unemployment increases the risk of mental disorder (Hallsten 1998) and is also associated with a lifestyle that increases the risk of developing cardiovascular disease, particularly in young people (Janlert et al. 1991, Janlert and Hammarström 1991). Economic hardship may add to the health consequences (Vinokur et al. 1996). Jahoda (1979) formulated the basic reasons why work is such an important activity in modern life—the most central one being that work provides structure to life. There is extensive research showing that organized support can help the unemployed to regain employment (Vinokur et al. 1995). Evaluation of such programs, however, also shows that these programs may increase passiveness and feelings of low self-esteem if the initiative of the individual is taken away.

6. *Combined Effects of Physical and Psychosocial Adversities at Work*

Extensive research has been devoted to the disentanglement of physical from psychological stress in studies of health and disease. For instance, in the study of the importance of ergonomic load to the development of musculoskeletal disorders, it has been considered important to adjust for psychological stress. Extensive epidemiological studies of large samples of subjects have shown that ergonomic load contributes to the development of musculoskeletal disorder but that psychosocial factors at work may also contribute significantly (Kilbom et al. 1996). Recent studies have shown that combinations of physical and psychosocial factors may be of particular importance (Vingård et al. in press).

It has been pointed out by several authors (House and Cottington 1986) that combinations of physically adverse (chemicals, radiation, physical load, infections) and psychosocial conditions may be of particular importance to health in a changing world. It is not possible to monitor all new chemicals that are used, and our psychosocial conditions are changing. Accordingly, analyses of interactions will be of increasing importance (Rantanen 1999)

It has been important for the study of relationships between working conditions and health to have theoretical models which form the basis for hypothesis testing. The interplay between the individual and the environment has been an important point of departure in all theoretical modeling in this field. Some of the models have focused on environmental, for instance, the demand–control model, and some on individual aspects, for instance, the eustress–distress model. The person–environment fit (PE fit) and the effort–reward imbalance models are examples of models which take the interaction between the individual and the environment into account. These four models will be described briefly below.

7. *Theoretical Models for the Study of Work and Health—Individual or Environment*

The individual's response to demanding work situations has been the point of departure for Frankenhaeuser and her co-workers (Frankenhaeuser 1980) who introduced one of the most influential models, the eustress–distress model. According to this model, the individual responds to the work situation with different degrees of effort. The bodily reactions to a high degree of effort are influenced by the concomitant degree of joy or distress in the situation. High effort with joy is labeled eustress while high effort with marked distress corresponds to a negative stress reaction. Laboratory research indicated that joyless effort was associated with elevated catecholamine as well as elevated cortisol excretion, whereas the eustress situation was associated with increased catecholamine

excretion but no elevation of cortisol excretion. The physiological damage associated with the negative reaction was assumed to be much worse than that of the eustress situation.

The environmental counterpart of the effort–distress model is the demand–control model which has been used extensively during later years. Accordingly this will be described in more detail. The demand–control model was introduced by Karasek. It is a two-dimensional synthesis of the demand ('stress psychological') and the 'lack of control' (sociological) research traditions (Karasek and Theorell 1990, Karasek 1979). It was assumed that the employee's possibility to utilize and develop skills (skill utilization) was closely related to authority over decisions. Skill utilization has to do with the employee's control over the use and development of his/her skills, whereas authority over decisions has to do with the employee's control over decision making relevant to his work tasks. The two factors have often been added to one another to constitute decision latitude.

The other dimension in the original demand—control model, psychological demands, includes qualitative as well as quantitative demands.

The high demand–low decision latitude combination, job strain, is regarded as the most dangerous for illness development. According to the theory, this kind of situation, if prolonged and repeated for a long time, increases sympathoadrenal arousal and at the same time decreases anabolism, the body's ability to restore and repair tissues (Karasek and Theorell 1990).

The combination of high psychological demands and high decision latitude is defined as the active situation. In this situation, the worker has been given more resources to cope with high psychological demands because he/she can make relevant decisions, such as planning working hours according to his/her own biological rhythm. In addition he/she gets good possibilities to improve coping strategies—facilitating feeling of mastery and control in unforeseen situations. This situation corresponds to psychological growth.

The low demand—high decision latitude situation—the relaxed one—is theoretically the ideal one, whereas the low demand–low decision latitude situation which is labeled passive may be associated with risk of loss of skills and to some extent psychological capacity.

It soon became evident that a third component was needed in the demand–control model—social support. This dimension had already been introduced previously in job stress research (House 1981), see below. Johnson has discussed this component extensively. He has pointed out that collective support may be difficult to separate theoretically from collective control. Workers who are exposed to adverse working conditions may join forces and increase their control over the working conditions. The worst hypothesized situation is iso-strain—job strain and lack of social support (Johnson et al. 1989).

There is a striking difference between the effort–distress model and the demand–control model. Whereas the former is designed to explain differences in individual response to stressors, the latter is designed to assess the environmental conditions, the stressors. There is of course a relationship between the models since eustress reactions are likely be more common in the active job situation and negative stress reactions more common in the job strain situation.

8. *Interaction Between Individual and Environment*

One of the most influential models has been that of the Michigan school (Katz and Kahn 1966). This theory postulates that adverse reactions at work arise when the person–environment fit is poor. In this model the interaction between a person's characteristics and the environment is the central theme. If the characteristics of the job do not fit the characteristics of the person there is a high likelihood that negative stress reactions may arise. There is considerable empirical support for this hypothesis (Kahn 1981). Another important theme is the expectation of the person—those who have very positive expectations may be more disappointed than those who have negative expectations if the conditions do not fit. The Michigan school also added social support at work as a central theme (House 1981).

According to Siegrist (1996), a crucial job factor for health consequences is the degree to which the workers are rewarded for their efforts. When a high degree of effort does not meet a high degree of reward, emotional tensions arise and illness risk increases. The effort–reward imbalance model is described in another chapter by Siegrist in this book. It is basically a sociological theory that includes individual processes. The effort component includes individual (intrinsic effort, which is close to psychological demands in the demand–control model) as well as environmental aspects (extrinsic effort). The reward component includes three different aspects of reward, namely monetary reward, promotion possibilities, and self esteem.

During later years, considerable efforts have been made to explore the relationship between theoretical job stress models and 'hard endpoints' for health development. Previously the scientific literature was criticized because most of the relationships had been observed in studies which used self-reports both in descriptions of the stressors and the health outcomes (Kasl 1982). There may be a risk in such studies that a pattern of complaining may affect both explaining variables and health outcomes. Therefore, studies have included prospective designs as well as objective outcomes (such as myocardial infarction or cardiac death) and objective recordings of stressors at work (through population means for each occupation re-

garding decision latitude, for instance, or by means of expert ratings of the crucial stressors). The demand–control model has been more extensively studied than the other models.

9. Job Stress Models and Disease

The significance of the demand–control–support model to cardiovascular disease, in particular high blood pressure and myocardial infarction (Schnall et al. 1990, Theorell and Karasek 1996, Melamed et al. 1998), has been tested in many epidemiological studies. There is considerable support for the hypothesis that the combination of high demands and lack of control is associated with increased risk of developing a myocardial infarction before retirement age. There are limitations in the findings, however: the findings are stronger for blue-collar workers than for white-collar workers, more established for men than for women, and more pronounced before the age of 55 than after that age.

The control (decision latitude) dimension has had a more consistent relationship with myocardial infarction risk than the demand dimension. Loss of decision latitude has also been shown to be associated with increased risk of developing a first myocardial infarction in the near future, even after adjustment for other risk factors (Theorell et al. 1998). The support dimension has been less extensively studied than demand and control.

The findings are more convincing for myocardial infarction than for cardiovascular risk factors (Kristensen 1999). High blood pressure during activities has shown a convincing relationship with job strain, whereas high blood pressure at rest (measured in the doctor's office) has not been associated with the demand–control model in any consistent way. Serum lipids and smoking patterns have not shown consistent relationships with job strain (Theorell et al. 1998). Accelerated blood coagulation may be a mechanism that operates in the relationship between job strain and myocardial infarction risk since recent studies have shown that high plasma fibrinogen is associated with low decision latitude (Tsutsumi et al. 1999, Brunner et al. 1996).

Variations over time in job strain have been shown to be associated with changes in endocrine parameters and immune system function of relevance to stress theory. For instance, in one study of men, plasma testosterone (which is important to anabolism) decreased and immunoglobulin G increased (indicating increased immune system activity) when job strain increased and vice versa (Theorell et al. 1990a, 1990b). In another study plasma testosterone increased when the total working life situation improved (Grossi et al. 1999).

Studies of the effort–reward model in relation to risk factors have shown that in men high blood pressure and atherogenic lipids are associated with poor reward in relation to high effort. The patterns of association are different for men and women (Peter et al. 1998, Siegrist 1996).

The PE fit model and related models have been used in studies of cardiovascular risk factors and positive findings have been reported (Cooper and Payne 1978).

Sickness absence has been shown to be positively associated with high demands and low decision latitude (North et al. 1996). Musculoskeletal disorders such as low back pain and neck–shoulder pain have also been studied in relation to the demand–control model. The findings have been mixed. The relationships depend on the group and the kind of disorder studied. In some studies, social support at work is crucial whereas in others psychological demands or decision latitude may be important. When the associations between musculoskeletal disorder and psychosocial work environment are adjusted for physical load, the associations are often weakened. Among female workers, physical load and adverse psychosocial conditions are frequently correlated (Josephson et al. 1999).

10. Integration of Effort–Reward Imbalance and Demand–Control

There is evidence (Bosma et al. 1998) that the decision latitude component of the demand–control model and the effort–reward imbalance model are contributing independently of one another to the prediction of episodes of coronary heart disease. This finding confirms that the models are related to different psychosocial mechanisms linking psychosocial working conditions to health outcome. The models have psychological demands (extrinsic effort) in common but control (decision latitude) and reward are clearly different. A logical step would be to combine the models. Even if resources (control and support) are optimal for the development of good coping strategies in a highly demanding situation, the employees will need reward for high effort, and hence balance between the components is needed.

Most of the relationships that have been observed in this field would be without significance in society if it were not possible to prove that it is possible to improve health by means of changes in working conditions. Accordingly, the literature on job interventions and their possible effects on health is of considerable importance.

11. Environmental and Individual Psychosocial Interventions at Work

Quality of working life experiments have been carried out since the early 1980s. They have been evaluated in terms of productivity but not frequently of health, and

it has been shown that they have very often been associated with a lasting increase in productivity (Kopelman 1985). They have mostly been based on the theories outlined above, and two basic ideas have been to increase decision latitude and to improve social support. From an intervention perspective, decision latitude has two components, namely intellectual discretion—possibility for the worker to decide how his knowledge should be used and developed—and authority over decisions—which has to do with work democracy, how decisions are made, and how workers are informed about important decisions. Increase in intellectual discretion has been achieved by means of job enrichment or increased worker responsibility for the complete product.

Increase in decision authority has been achieved by means of job enlargement or flattened organization hierarchy. Increase in social support, finally, has been achieved by means of improved feedback and the formation of more cohesive work groups. During later years, both in North America and in Scandinavia, the starting of regular staff meetings for systematic discussion of important decisions regarding work routines and goals has been used as a tool for increased decision latitude and social support.

The theoretical models described above have also been used in efforts to prevent illness and to promote health. The occupational health care team has often been an important vehicle in Scandinavian job redesign (see for instance Wallin and Wright (1986). The occupational health care team has the possibility to work both with the structure of the company and individual workers. It has the possibility to carry out individual measurements of psychological and somatic health indicators and to monitor these through the change process. It has been the experience of health oriented job redesigners that the concomitant emphasis on structure and individuals is helpful both for individual motivation to follow health promotion advice (which is strengthened when workers discover that there is willingness in management to improve the structure) and for success in instituting structural change (which will benefit from engagement of the individuals and their health problems).

Even in Sweden with its long tradition of work redesign, very few studies have actually documented health consequences of this type of job intervention. Most of the published studies have been based on self-administered questionnaires. One study of cardiovascular risk factors (Orth-Gomér et al. 1994) showed that a similar program aiming at improved decision latitude and social support for the employees (which was evaluated in a controlled study) was followed by a reduced ratio of LDL/HDL cholesterol (the ratio between harmful and protective cholesterol) in the experimental but not in the comparison group. As in all experiments of this kind, the design was not 'clean' in the sense that all other changes had been ruled out, but the findings could be interpreted to mean that job redesign aiming at improved social relationships and decision latitude could lead to decreased cardiovascular risk.

Psychosomatic symptoms, such as gastrointestinal symptoms, tiredness, and depression, reportedly decreased among employees in part of the Volvo company after a number of psychosocially oriented job changes had been instituted, such as role change in supervisors in the direction of more facilitation, less authoritarian leadership style, promotion of personal contact between customers and employees, and increased teamwork (Wallin and Wright 1986).

The redesign process has been described in many case studies. The health consequences of improved work organization, however, have been in scientific focus in few published studies. Experiences in all EU countries have recently been summarized (Kompier and Cooper 1999). Evaluations of systematic efforts to improve working conditions by means of increased decision latitude, support, and reward in order to improve health are more common in the Northern countries than in other countries.

12. *Notes on Future Work*

In an increasingly unstable labor market in the future, employment security will be less common and global competition will increase (Aronsson and Göransson 1999). The education level of all workers in the world will rise. If we use the basic concepts of demands, decision latitude (including both intellectual discretion and authority over decisions), social support, effort, and reward in order to describe possible scenarios for the future, we might hypothesize the following to occur.

Psychological demands will continue to rise, both because there is more cognitive demand (memorizing numbers, handling computers, etc.) and more emotional demands. Numbers of collaborative partners are rising because of the increasing specialization of work, and thus there will be an increasing likelihood of complicated emotional communication for everybody.

Decision latitude is becoming polarized and this will continue in the future. Workers with a high level of education will have more and more decision latitude, whereas less favored groups will have periods of unemployment interrupted by periods of short employment. This will be a poor basis for developing decision latitude for low education workers. Studies in several countries have already shown that workers in temporary employment have a bad position in decision processes. There are indications from several countries that differences in health between favored and less favored workers will increase.

The basis for social support will change gradually since workers on all levels will rotate more frequently between worksites in different projects than pre-

viously. Employers will have a special responsibility for the organization of social support in worksites.

See also: Job Stress, Coping with; Stress at Work; Stress in Organizations, Psychology of; Well-being and Burnout in the Workplace, Psychology of; Workplace Safety and Health; Worksite Health Promotion and Wellness Programs

Bibliography

Aronsson G, Göransson S 1999 Permanent employment but not in a preferred occupation: Pyschological and medical aspects, research implications. *Journal of Occupational Health Psychology* **4**: 152–63

Bosma H, Peter R, Siegrist J, Marmot M 1998 Two alternative job stress models and the risk of coronary heart disease. *American Journal of Public Health* **88**: 68–74

Brenner S O, Levi L 1987 Long term unemployment among women in Sweden. *Social Science and Medicine* **25**: 153–61

Brunner E, Davey Smith G, Marmot M, Canner R, Beksinska M, O'Brien J 1996 Childhood social circumstances and psychosocial and behavioural factors as determinants of plasma fibrinogen. *Lancet* **13**: 1008–13

Charney D S, Deutch A Y, Krystal J H, Southwick S M, David M 1993 Psychobiological mechanisms of post-traumatic stress disorder. *Archives of General Psychiatry* **50**: 294–330

Cooper C L, Payne R (eds.) 1978 *Stress at Work*. Wiley, Chichester, UK

Demitrack M A, Dale J K, Straus S E, Laue I, Listwak S H, Krusi M J P, Chrousos G P, Gold P W 1991 Evidence for impaired activation of the hypothalamic-pituitary-adrenal axis in patients with chronic fatigue syndrome. *Journal of Clinical and Endocinological Metabolism* **73**: 1224–34

Frankenhaeuser M 1980 Psychoneuroendocrine approaches to the study of stressful person–environment transactions. In: Selye H (ed.) *Selye's Guide to Stress Research*. Van Nostrand Reinhold, New York, Vol. 1

Grossi G, Theorell T, Jurisoo M, Setterlind S 1999 Psychophysiological correlates of organizational change and threat of unemployment among police inspectors. *Integrated Physiology and Behavioral Science* **34**(1): 30–42

Hallsten I 1998 *Psykiskt välbefinnande och arbetslöshet* (*Psychological Health and Unemployment*). Arbete och Hälsa 7. National Institute for Working Life Research, Solna, Sweden

House S J 1981 *Work Stress and Social Support*. Addison-Wesley, Reading, MA

House S J, Cottington E M 1986 Health and the workplace. In: Aiken L H, Mechanic D (eds.) *Applications of Social Science: Clinical Medicine and Health Policy*. Rutgers University Press, New Brunswick, NJ

Jahoda M 1979 The impact of unemployment in the 1930's and the 70's. *Bulletin of the British Psychological Society* **32**: 309–14

Janlert U, Asplund K, Weinehall L 1991 Unemployment and cardiovascular risk indicators. Data from the Monica survey in northern Sweden. *Scandinavian Journal of Social Medicine* **1**: 14–18

Janlert U, Hammarström A 1991 Alcohol consumption among unemployed youths: Results from a prospective study. *British Journal of Addiction* **87**: 703–14

Johnson J, Hall E, Theorell T 1989 The combined effects of job strain and social isolation on the prevalence and mortality incidence of cardiovascular disease in a random sample of the Swedish male working population. *Scandinavian Journal of Work Environment and Health* **15**: 271–9

Josephson M, Pernold G, Ahlberg-Hulten G, Harenstam A, Theorell T, Vingard E, Waldenstrom M, Hjelm E W 1999 Differences in the association between psychosocial work conditions and physical work load in female- and male-dominated occupations. MUSIC-Norrtalje Study Group. *American Industrial Hygiene Association Journal* **60**(5): 673–8

Kahn R 1981 *Work and Health*. Wiley, New York

Karasek R A 1979 Job demands, job decision latitude and mental strain: Implications for jobs redesign. *Administrative Science Quarterly* **24**: 285–307

Karasek R A, Theorell T 1990 *Healthy Work*. Basic Books, New York

Kasl S V 1982 Chronic life stress and health. In: Steptoe A, Mathews A (eds.) *Health Care and Human Behavior*. Academic Press, London

Katz D, Kahn R 1966 *Social Psychology of Organizations*. Wiley, New York

Kilbom A, Armstrong T, Buckle P, Fine L, Hagberg M, Haring-Sweeney M, Martin B, Punnett L, Silverstein B, Sjogaard G, Theorell T, Viikari-Juntura E 1996 Musculoskeletal disorders: Work-related risk factors and prevention. *International Journal of Occupational and Environmental Health* **2**(3): 239–46

Kompier M, Cooper C 1999 *Preventing Stress, Improving Productivity. European Case Studies in the Workplace*. Routledge, London

Kopelman R E 1985 Job redesign and productivity: A review of the evidence. *National Productivity Review* (Summer): 237–55

Kristensen T S 1999 Challenges for research and prevention in relation to work and cardiovascular diseases. *Scandinavian Journal of Work Environment and Health* **25** (6, special issue): 550–7

Levi L 1972 Stress and distress in response to psychosocial stimuli. *Acta Medica Scandinavica Suppl.* **191**: 1–151

Melamed S, Kristal-Boneh E, Harari G, Froom P, Ribak J 1998 Variation in the ambulatory blood pressure response to daily work load—the moderating role of job control. *Scandinavian Journal of Work Environment and Health* **24**(3): 190–6

North F M, Syme S L, Feeney M, Shipley M, Marmot M 1996 Psychosocial work environment and sickness absence among British civil servants: the Whitehall II study. *American Journal of Public Health* **86**: 332–40

Orth-Gomér K, Eriksson I, Moser V, Theorell T, Fredlund P 1994 Lipid lowering through work stress reduction. *International Journal of Behavioral Medicine* **1**: 204–14

Peter R, Alfredsson L, Hammar N, Siegrist J, Theorell T, Westerholm P 1998 High effort, low reward, and cardiovascular risk factors in employed Swedish men and women: Baseline results from the WOLF study. *Journal of Epidemiology and Community Health* **52**(9): 540–7

Rantanen J 1999 Research challenges arising from changes in worklife. *Scandinavian Journal of Work Environment and Health* **25**(6, special issue): 473–83

Rubin R T, Poland R E, Lesser I M, Winston R A, Blodget A L 1987 Neuroendocrine aspects of primary endogenous depression. I. Cortisol secretory dynamics in patients and matched controls. *Archives of General Psychiatry* **44**: 328–36

Schnall P, Schwartz J E, Landsbergis P A, Warren K, Pickering T 1990 The relationship between job strain, workplace, diastolic blood pressure and left ventricular mass index. *Journal of the American Medical Association* **263**: 1929–35

Selye H 1976 *The Stress of Life*. McGraw-Hill, New York
Siegrist J 1996 Adverse health effects of high effort/low reward conditions. *Journal of Occupational and Health Psychology* **1**: 27–41
Theorell T, Karasek R A 1996 Current issues relating to psychosocial job strain and cardiovascular disease research. *Journal of Occupational and Health Psychology* **1**: 9–26
Theorell T, Karasek R A, Eneroth P 1990a Job strain variations in relation to plasma testosterone fluctuations in working men—a longitudinal study. *Journal of Internal Medicine* **227**(1): 31–6
Theorell T, Orth-Gomer K, Eneroth P 1990b Slow-reacting immunoglobulin in relation to social support and changes in job strain: A preliminary note. *Psychosomatic Medicine* **52**(5): 511–6
Theorell T, Tsutsumi A, Hallquist J, Reuterwall C, Hogstedt C, Fredlund P, Emlund N 1998 Decision latitude, job strain and myocardial infarction: a study of working men in Stockholm. *American Journal of Public Health* **88**: 382–8
Tsutsumi A, Theorell T, Hallqvist J, Reuterwall C, de Faire U 1999 Association between job characteristics and plasma fibrinogen in a normal working population: a cross sectional analysis in referents of the SHEEP Study. Stockholm Heart Epidemiology Program. *Journal of Epidemiology and Community Health* **53**(6): 348–54
Vingård E, Alfredsson L, Hagberg M, Kilbom Å, Theorell T, Waldenström M, Wigaeus Hjelm E, Wiktorin C, Hogstedt C 2000 How much do current and past occupational physical and psychosocial factors explain of low back pain in a working population? Results from the MUSIC-Norrtälje study. *Spine* **25**(4): 493–500
Vinokur A D, Price R H, Caplan R D 1996 Hard times and hurtful partners: How financial strain affects depression and relationship satisfaction of unemployed persons and their spouses. *Journal of Personality and Social Psychology* **71**: 166–79
Vinokur A D, Price R H, Schul Y 1995 Impact of the JOBS intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology* **23**: 39–74
Wallin L, Wright I 1986 Psychosocial aspects of the work environment: A group approach. *Journal of Occupational Medicine* **28**: 384–93

T. Theorell

Occupational Health and Safety, Regulation of

This article outlines developments in the legal regulation of the health and safety of workers and others at work. It does not purport to offer a comprehensive survey of international occupational health and safety (OHS) practice and theoretical perspectives. Instead, it traces historically the main developments in the field of OHS regulation, with an emphasis on Europe and North America. Within this framework, and with a special focus on the movement from command and control to self-regulatory models, the article examines three themes of particular importance in the development of OHS regulatory regimes, wherever found. These are the distinguishable features of OHS standard setting, the variable models of enforcement, and the broader ideological significance of OHS regimes, most particularly the relationship between OHS and industrial relations.

1. The Emergence of the Traditional Approach to OHS Regulation

OHS regulation is generally traced back to the early nineteenth-century Factory Acts in Great Britain. The earliest enactment was the Health and Morals of Apprentices Act 1802, essentially an extension of Elizabethan Poor Law. It regulated the working hours and conditions of apprentice pauper children in the water-based rural cotton factories of Northern England. The enforcement provisions relied on informers and visitors appointed by the magistracy from the clergy and justices of the peace, but were generally regarded as completely ineffective (Thomas 1948, pp. 11–12), largely because of the limited powers given to visitors, the lack of interest of magistrates in the factory legislation, and the fact that magistrates were too closely linked to mill owners to enforce the law (Gunningham 1984, p. 38).

The ineffectiveness of this early factory legislation led to the enactment of the Factory Regulation Act of 1833, which regulated the employment of children in a wide range of textile trades. The major historical significance of the 1833 Act was that it was the first to be enforced by an independent salaried factory inspectorate vested with broad powers, including the right of entry to factories and mills; the right to make rules, regulations, and orders; and the power to initiate and hear court proceedings to enforce the Act. Over time, however, these broad powers were reduced, at the initiative of inspectors themselves, until they conformed to a more 'legalist' regulatory form (where inspectors' powers were confined to inspecting and bringing enforcement action), which overcame allegations of class bias by presenting decision-making as a passive, neutral, nondiscretionary and technical process of implementation of the rules set out in the Act (Field 1990).

While the enactment of the 1833 Act was partly a response to working-class pressure, and agitation by middle- and upper-class philanthropists for shorter working hours for children and young people, it was undoubtedly also a result of calls for effective uniform regulation by larger urban manufacturers who had voluntarily introduced shorter working hours and who feared unscrupulous competition from smaller and rural-based manufacturers (Carson 1974, Marvel 1977). For various reasons (see Bartrip and Fenn 1983), the early factory inspectors found the 1833 Act

difficult to enforce. Prosecutions were time-consuming and costly; over time the ratio of inspectors to factory premises fell significantly; inspectors were ambivalent about the degree to which regulatory offences committed by businesses creating wealth and employment were morally reprehensible, and were deterred by hostile attitudes of magistrates towards the provisions of the Act. The inspectors also found that their assumption that offences against the Act were only being committed by a few employers had been ill-founded, and that contraventions were widespread, and intrinsic to the prevailing production system (Carson 1979). As a consequence, they developed an approach to enforcement which relied principally on advice and persuasion, and resorted to formal prosecution only in the worst cases and where there was evidence of wilful employer disobedience or obstinacy. Although officially criminal offences, contraventions of the Factory Act were accepted as customary and those responsible were rarely subjected to criminal prosecution. This 'ambiguity' of factory crime (Carson 1980), although historically contingent, has proved to be deeply entrenched ideologically in most of the world's OHS regulatory systems.

The Factories Amendment Act 1844 extended the coverage of the legislation to the working hours of women, and included the first machinery safety provisions. In compliance with inspectors' wishes, it also deleted all references to wilfulness in many of the employers' offences (Carson 1979, pp. 53–4). The 1833 and 1844 Acts marked the beginning of the traditional approach to the regulation of health and safety at work, in which the government set down highly detailed and technical requirements in strict liability 'specification' standards, and provided for the enforcement of these standards by a state inspectorate which relied principally upon informal enforcement responses. This regulatory model was reproduced in other industries in Britain, transported to the British colonies, and strongly influenced the OHS regulatory provisions of the rest of Europe (see Baldwin and Daintith 1992). In the USA, similar models of OHS regulation were established at state level from the late 1860s until the end of the nineteenth century. The Federal government relied upon state regulation until 1970, when the Occupational Safety and Health Act was enacted (Noble 1975, Ashford 1976).

2. *Criticisms of the Traditional Command and Control Approach*

The weaknesses of this traditional approach are well known (see Robens 1972). It frequently resulted in a mass of detailed and technical rules, often difficult to understand, and difficult to keep up to date. Standards were developed *ad hoc* to resolve problems as they arose, and concentrated mainly on factory-based physical hazards, resulting in uneven coverage across workplaces. Specification standards did not encourage or even enable employers to be innovative and to look for cheaper or more cost-efficient solutions. They also ignored the now well-accepted view that many hazards do not arise from the static features of the workplace, but from the way work is organized.

Critics have also pointed to the inadequacy of state approaches to enforcement, and to the ideological role of traditional factory legislation in encouraging dependency on state measures and in separating OHS from the conflictual terrain of industrial relations. This results in a lack of involvement of workers and their representatives in developing standards and measures for improving health and safety at the workplace. One exception to this trend has been some of the Scandinavian countries, most notably Sweden, which, from early in the twentieth century, has provided for trade union involvement in OHS (Navarro 1983).

Since 1970, there have been major developments in the regulation of OHS. The general trend has been away from traditional 'command and control' models, towards a more 'self-regulatory' system in which duty holders are given more flexibility and choice of methods to achieve OHS standards, there is greater scope for worker involvement in OHS, and approaches to state inspection focus as much on systems of work as on workplace hardware.

3. *The Modern Approach to OHS Regulation*

3.1 *Standard Setting*

Since the mid-1970s, the trend in Europe and North America has been to replace or supplement specification standards with a combination of general duties, performance standards, process requirements, and documentation requirements. The trend has been most marked in Europe, Canada, and Australia.

The 'general duty' provisions require the duty holder (usually the employer, but in some countries including the self-employed, occupiers of workplaces, and manufacturers, suppliers, and designers of plant and substances) to exercise reasonable care in the provision and maintenance of systems of work so as to ensure that workers are not exposed to health and safety risks. For example, the US Occupational Safety and Health Act 1970 (OSHA) Sect. 5 specifies that in all cases not covered by specific standards promulgated under the Act, the employer has a general duty to 'furnish to each of his employees, employment and a place of employment which are free from recognized hazards that are causing or are likely to cause death or serious physical harm to his employees.'

The general duty provisions tend to be supplemented by a range of standards set out in subordinate

legislation. Instead of telling duty holders exactly how they are to achieve compliance, performance standards define the duty holder's duty in terms of goals they must achieve, or problems they must solve, and leave it to the initiative of the duty holder to work out the best and most efficient method for achieving the specified standard.

Process requirements prescribe a process, or series of steps, that must be followed by a duty holder in managing specific hazards, or OHS generally. They are often used when the regulator has difficulty specifying a goal or outcome, but has confidence that the risk of illness or injury will be significantly reduced if the specified process is followed. For example, as a result of the EC 'Framework Directive' of 1989, 'On the Introduction of Measures to Encourage Improvements in the Safety and Health of Workers at Work,' all EC member states are required to introduce risk assessment processes into their OHS statutes. Process-based standards have spawned greater reliance on documentation requirements. Increasingly OHS statutes are requiring duty holders to document measures they have taken to comply with process-based standards, performance standards, and general duty standards. An example is the requirement for health and safety plans and files in the 1992 EC Directive to Implement Minimum Health and Safety Requirements at Temporary or Mobile Construction Sites.

A current debate in OHS standard setting is whether OHS regulators can take measures to require employers and other duty holders to go beyond compliance with minimum standards, and to implement comprehensive OHS management systems which involve continuous improvement and a systems-approach to OHS management (see Gunningham and Johnstone 1999, and Frick et al. 2000). The debate is most advanced in provisions for the management of major hazardous facilities, which have largely been enacted in reaction to major disasters. Many jurisdictions now require facility operators to submit a 'safety case' for each of its installations demonstrating that they have evaluated risks and set up a system of management and control to deal with the identified risks (Kaasen 1991, and Cullen 1990, and see the International Labour Organization's 1993 Convention on the Prevention of Major Industrial Accidents, and the EU Directives 82/501/EEC and 96/82/EEC on the control of major hazard facilities).

3.2 Provisions Requiring Duty Holders to Seek Expert Assistance in OHS

Occupational health services can play an important role in supporting employers and improving their understanding of OHS matters. The EC Framework Directive (1989) requires employers to enlist 'competent services or persons' to assist in the implementation of the OHS duties in the Directive and 'to organize protective and preventive measures.' For example, consistent with provisions in other continental European nations, the Swedish Work Environment Act requires employers to provide an occupational health service. Most large companies have their own occupational health services, while joint occupational health service centers provide services to small- and medium-sized organizations. This continental tradition is not followed elsewhere, although UK regulations and some Australian and Canadian statutes require the appointment of 'competent persons' or workplace health and safety officers to assist employers to implement their OHS obligations.

3.3 Worker Participation

The nineteenth-century model of OHS regulation typically did not require worker involvement in OHS. There are, however, good reasons for involving workers in OHS. Workers have a direct interest in OHS because they run the major risks of injury if the law fails to protect them. Moreover, workers often know more about the hazards associated with their workplace than anyone else, because they regularly work with them. Where process standards require hazards at work to be identified and evaluated, workers' experience and knowledge is crucially important in successfully completing both of these tasks.

European, Canadian, and Australian OHS statutes have made some statutory provision for OHS committees, and for worker health and safety representatives. Representatives generally have the right to be consulted over OHS issues and to be provided with OHS information. Some European, Australian, and Canadian statutes go further, and give elected health and safety representatives the right to require compliance with OHS regulatory provisions and to direct that dangerous work ceases. The EC Framework Directive requires employers to involve workers and their representatives in the implementation of the various duties set out in the directive. The US OSH Act to date has made no provision for worker participation in OHS, although the Occupational Safety and Health Administration's (OSHA's) recent proposed safety and health program rule requires worker involvement in the development of a safety and health program.

In contrast, in Scandinavia the worker participation elements of the OHS regulatory system derive from the general industrial relations system, and do not depend on the OHS statutory framework. For example, the Swedish Work Environment Act 1977 enshrines the concept that worker control is a critical aspect of a healthy working environment, and provides health and safety representatives with extensive powers, including the right to halt dangerous processes, and to participate in the planning of new work premises and processes. Industrial and local agree-

ments regulate much of the administration of OHS. In Sweden, Spain, Italy, and Norway, the construction industry makes provision for coverage of small workplaces by enabling unions to nominate regional safety representatives for small workplaces (normally under 50 workers) to support the OHS initiatives by employers.

3.4 Inspection Approaches Aimed at Promoting OHS Management Systems

In some countries, OHS inspectorates have developed enforcement strategies that focus as much on OHS management as on workplace hardware. For example, in the USA, some OSHA inspection programs abbreviate inspections if employers can show that they have effective safety and health programs. Under the 'focused inspection' program, piloted in the construction industry, if an OSHA inspector conducts an inspection and finds an effective safety program operating on-site, then the remainder of the inspection will be limited to the top four hazards that kill workers in the construction industry. If these hazards are well controlled, the inspector concludes the inspection. If not, the inspector conducts a complete site inspection. Such approaches to inspection are also increasingly being used in Europe and Scandinavia.

3.5 State Enforcement of OHS Obligations

The modern OHS inspectorates can call upon a range of enforcement powers and, in most countries, have a broad discretion to determine enforcement action, including negotiating compliance using informal processes such as education, advice, and persuasion. Since the 1970s, inspectorates in most countries have been given administrative sanctions, such as improvement and prohibition notices and on-the-spot fines.

In the USA, enforcement is less discretionary, and principally involves administrative sanctions (citations imposing financial penalties and requiring the abatement of hazards) (see Kaletsky 1997, Shapiro and Rabinowitz 1997), with prosecution for a limited number of intention-based offences. In other countries, the OHS inspectorates have been criticized for not adopting a sufficiently strong approach to enforcement, and for relying too heavily on advisory and persuasive approaches. Some commentators argue for a more vigorous approach to enforcement utilizing the 'self-regulatory' principles of the enforcement pyramid promoted by many regulatory theorists (see Rees 1988, Ayres and Braithwaite 1992, Chap. 2). The challenge is to develop enforcement strategies that deter the worst offenders, while at the same time encourage and help most employers to comply voluntarily. Such an approach to enforcement would require the availability of larger financial penalties, and an array of new corporate sanctions (Gunningham and Johnstone 1999, Chap. 4).

See also: Employment and Labor, Regulation of; Job Design and Evaluation: Organizational Aspects; Occupational Health; Regulation: Working Conditions; Workplace Safety and Health; Worksite Health Promotion and Wellness Programs

Bibliography

Ashford N A 1976 *Crisis in the Workplace: Occupational Disease and Injury*. MIT Press, Cambridge, MA

Ayres I, Braithwaite J 1992 *Responsive Regulation: Transcending the Deregulation Debate*. Oxford University Press, New York

Baldwin R, Daintith T (eds.) 1992 *Harmonization and Hazard: Regulating Health and Safety in the European Workplace*. Graham & Trotman, London

Bartrip P W J, Fenn P T 1983 The evolution of regulatory style in the nineteenth century British Factory Inspectorate. *Journal of Law and Society* **10**: 201

Carson W G 1974 Symbolic and instrumental dimensions of early factory legislation. In: Hood R G (ed.) *Crime, Criminology and Public Policy*. Heinemann Educational, London

Carson W G 1979 The conventionalisation of early factory crime. *International Journal of the Sociology of Law* **7**: 37–40

Carson W G 1980 The institutionalization of ambiguity: Early British Factory Acts. In: Geis G, Stotland E (eds.) *White Collar Crime: Theory and Research*. Sage Publications Beverly Hills, CA

Cullen Lord (chairman) 1990 *Piper Alpha Inquiry*. HMSO, London

European Agency for Safety and Health at Work 1997 *Priorities and Strategies in Occupational Safety and Health Policy in the Member States of the European Union*. European Agency for Safety and Health at Work, Bilbao, Spain

Field S 1990 Without the law? Arthurs and the early Factory Inspectorate. *Journal of Law and Society* **17**: 445–68

Frick K, Jensen P L, Quinlan M, Wilthagen T (eds.) 2000 *Systematic Occupational Health and Safety Management: Perspectives on an International Development*. Elsevier, Amsterdam

Gunningham N 1984 *Safeguarding the Worker: Job Hazards and the Role of the Law*. Law Book Company, Sydney, Australia

Gunningham N, Johnstone R 1999 *Regulating Workplace Safety. Systems and Sanctions*. Oxford University Press, Oxford, UK

Henriques U R Q 1979 *Before the Welfare State: Social Administration in Early Industrial Britain*, 1st edn. Longmans, London

Kaasen K 1991 Post Piper Alpha: Some reflections on offshore safety regimes from a Norwegian perspective. *Journal of Energy and Natural Resources Law* **9**: 281

Kalestsky R 1997 *OSHA Inspections. Preparation and Response*. McGraw-Hill, New York

Marvel H P 1977 Factory regulation: A reinterpretation of the early English experience. *Journal of Law and Economics* **20**: 379–402

Navarro V 1983 The determinants of social policy, a case study: Regulating health and safety at the workplace in Sweden. *International Journal of Health Services* **13**: 517–61

Noble C 1986 *Liberalism at Work: The Rise and Fall of OSHA*. Temple University Press, Philadelphia, PA

Rees J V 1988 *Reforming the Workplace: A Study of Self-regulation in Occupational Safety*. University of Pennsylvania Press, Philadelphia, PA

Robens, Lord (chairman) 1972 *Safety and Health at Work, Report of the Committee, 1970–72*, London, HMSO, Cmnd 5034 (Robens Report)

Shapiro S, Rabinowitz R S 1997 Punishment versus cooperation in regulatory enforcement: A case study of OSHA. *Administrative Law Review* **49**: 713

Thomas M W 1948 *The Early Factory Legislation: A Study in Legislative and Administrative Evolution*. Thames Bank Publishing, London

R. Johnstone and W. G. Carson

Offenses against the Laws of Humanity: International Action

This article examines and analyzes the evolution of the concept of the laws of humanity, and of the offenses associated with it in the modern era. It traces the varying interpretations on the precise contours of the concept, with particular emphasis on the Martens Clause which first appeared in the Preamble to the 1899 Hague Convention II. Although Hague Convention II dealt with the laws and customs of land warfare, the Clause's underlying notion of transcendental humanitarian principles reflected a basic commitment to human dignity and welfare; a commitment which was not conditioned on the status of the individual (combatant or noncombatant) or on the circumstances under which the individual's rights and responsibilities were to be exercised (peacetime or wartime). This is followed by a discussion of key turning points in the understanding of the term crimes against humanity which encapsulates the range of offenses associated with the laws of humanity. In this context, the article surveys the debates surrounding the modalities of the international community's response and action on the aftermath of both World Wars, and the growing intersections between international humanitarian law (laws of war) and international human rights law. In recent years, the concept is being reassessed, as a result of the evolving jurisprudence of the two *ad hoc* tribunals for the former Yugoslavia (ICTY) and Rwanda (ITR) relating to crimes against humanity and the conflict nexus requirement.

1. *The Laws of Humanity*

The term laws of humanity refers to transcendental humanitarian principles which 'are common to all human communities wherever they may be' (Pictet 1988). To be sure, the precise list of humanitarian principles shared by all communities irrespective of history, level of development, and cultural traditions is a matter of dispute. Yet the modern era has witnessed—especially since the later Enlightenment—a concerted effort to articulate the normative parameters of a fundamental baseline of welfare consistent with the preservation of human dignity. Ironically, the first substantive endeavors in that direction focused in an area of human activity whose very essence is considered the denial *par excellence* of humanity: warfare. The debates surrounding the attempts to alleviate the worst excesses of armed conflict were premised on an understanding of human suffering which transcended borders.

One of the earliest initiatives in the area of codified humanitarian controls appeared during the US Civil War. The United States Army General Order No. 100, *Instructions for the Government of Armies of the United States in the Field* (Known as the Lieber Code) expressly protected the civilian population. In addition, paragraph sixteen of the Code stipulated that 'military necessity does not admit of cruelty—that is, *the infliction of suffering for the sake of suffering or for revenge, nor of maiming or wounding except in fight*' (Hartigan 1983, Allen et al. 1996). Although the Lieber Code was a piece of domestic legislation issued in the context of a noninternational conflict, many of its provisions subsequently influenced agreements dealing with international conflicts. For example, the 1868 *St Petersburg Declaration*, the first major international agreement in which the contracting parties agreed to renounce a particular weapon in warfare, made similar references to unnecessary suffering. In particular, the *Declaration* recognized that the legitimate object of war (the weakening of the enemy's military forces) 'would be exceeded by the employment of arms which uselessly aggravate the sufferings of disabled men, or render their death inevitable'; and 'That the employment of such arms would, therefore, be contrary to the laws of humanity' (Roberts and Guelff 1989, Kalshoven 1987). Thus, causing unnecessary suffering, whether to civilians by failing to observe the combatant/noncombatant distinction, or to combatants by the use of certain types of weapons, was clearly perceived as a violation of the laws of humanity.

These concerns were not confined to the so-called Western world. Nobuhiro Sato, a Japanese strategist writing on military issues during the late eighteenth and early nineteenth centuries, argued that there was a clear rule against the killing of prisoners of war, and that its violation constituted an inhuman act. Moreover, he pointed out that 'in enemy land, coercing inhabitants and plundering their food was inhuman and that violent action was dishonourable' (Adachi 1988). Likewise, Andres Bello, a Venezuelan jurist who published in 1832 what is considered to be the first modern book on international law in Latin America, was a stronger believer in natural law principles. He

also advocated that the object of war was the weakening of the enemy, and for that purpose combatants were entitled to use the most effective methods and means 'provided they are not unlawful or contrary to natural law.' Concerning the issue of persons wounded in battle, Bello emphasized that it was the victor's responsibility to assist them, 'and that the laws of humanity and honour prohibited killing or despoiling them' (Ruda 1988). Despite these developments, it was not until the 1899 Hague Peace Conference that the Martens Clause, with its reference to the laws of humanity, reaffirmed the continuing relevance of customary law principles in guiding state practice.

1.1 The Martens Clause

The Martens Clause was inserted in the Preamble to the 1899 Hague Convention II with respect to the laws and customs of war on land and read as follows (Scott 1918):

> Until a more complete code of the laws of war is issued, the High Contracting Parties think it right to declare that in cases not included in the Regulations adopted by them, populations and belligerents remain under the protection and empire of the principles of international law, as they result from the usages establishes between civilized nations, from the laws of humanity, and the requirements of the public conscience.

The need for such a clause, drafted by the Russian delegate de Martens, arose as a result of strong disagreements among convention delegates on the issue of the legal status of resistance fighters in occupied territory. Certain delegates felt that if a very restrictive definition of the term belligerent were to be adopted, this would be tantamount to the criminalization of legitimate patriotic resistance against the invading forces. On the other hand, several delegates argued that soldiers had a right to be protected from potentially deadly action emanating from 'apparently peaceable inhabitants' (Holls 1900).

The Martens Clause broke the deadlock and has since been reaffirmed in all major codification efforts. Yet despite the fact that it constitutes the most important statement of humanitarian principles in the corpus of international humanitarian law, its exact significance remains a matter of dispute and there is no scholarly consensus on its interpretation. Most analysts seem to agree that the Clause symbolizes the continuing relevance of customary international law norms after the adoption of conventional (treaty-based) law ones (Greenwood 1995). Concerning wider interpretations, some have argued that given the incompleteness of the treaties regulating armed conflict, the Clause constitutes the ultimate humanitarian safeguard: namely, a reminder that a certain type of conduct may be proscribed even if it is not explicitly prohibited by a treaty (Ticehurst 1997). The latter interpretation is definitely consistent with the spirit of the Clause, which points to a fundamental baseline of human welfare, that is to be preserved irrespective of the contextual element, if a commitment to human dignity is to be taken seriously. In such a context, the term laws of humanity (or 'principles of humanity' in later versions of the Clause) refers to those norms, which prohibit means and methods not conducive to the attainment of a legitimate military objective. It is an attempt to balance military necessity against the requirements of humanity (Kalshoven 1987).

2. Crimes Against Humanity

Discussion on the laws of humanity resurfaced on the aftermath of World War I, when the victorious Allies created a Commission on the Responsibility of the Authors of the War and on Enforcement of Penalties. The purpose of the Commission was to collect evidence of violations of international law committed by Germany and its Allies, and to suggest a framework for the prosecution of those responsible. After an exhaustive study, the majority of the Commission concluded that the Central Powers carried out the war 'in violation of the established laws and customs of war and the elementary laws of humanity.' (The United Nations War Crimes Commission, UNWCC 1948.)

The first category of offenses could be classified as war crimes. The second category was composed of offenses 'committed on the territory of Germany and her Allies against their own nationals' (UNWCC 1948). There was an interesting precedent concerning the second category. During the War, when the Turkish Government's massacre of the Armenian population became known, the Governments of France, Great Britain, and Russia issued a Declaration denouncing these atrocities as 'crimes against humanity and civilization for which all the members of the Turkish Government will be held responsible, together with its agents implicated in the massacres' (Schwelb 1946). Thus, it was the latter set of offenses which was characterized as crimes against the laws of humanity, that constituted the precursor to the crimes against humanity provisions of the London and Tokyo Charters. The juxtaposition of these two types of offenses was a potentially major legal development, since, for the first time offenses against the laws of humanity could acquire a status under international law that was separate and distinct from war crimes.

The Commission's recommendations were challenged by the US delegation, which objected to the inclusion in the report of references to the laws of humanity. The American members argued, amongst other things, that there were important differences between the laws and customs of war and the laws and principles of humanity, and that the Commission's responsibilities lay in the investigation of the former, not the latter, because

> The laws and principles of humanity vary with the individual, which, if for no other reason, should exclude them from consideration in a court of justice, especially one charged with the administration of criminal law … . A judicial tribunal only deals with existing law and only administers existing law, leaving to another forum infractions of the moral law and actions contrary to the laws and principles of humanity. A further objection lies in the fact that the laws and principles of humanity are not certain, varying with time, place, and circumstance, and accordingly, it may be, to the conscience of the individual judge. There is not fixed and universal standard of humanity. (UNWCC 1948.)

Thus, in the US view, violations of the laws and principles of humanity were not justiciable, since they constituted moral rather than legal infractions. It was this position which eventually prevailed and, as a result, no reference to crimes against the laws and principles of humanity appeared in the text of the Treaty of Versailles (Van Schaack 1999). However, a very different perspective on this issue was to emerge as a result of the World War II experience.

2.1 The Nuremberg Era and its Legacy

The atrocities committed by the Axis Powers during World War II ensured the rapid transition of the offenses against the laws of humanity from obscurity to prominence. Contrary to its earlier position, the United States did endorse the inclusion of crimes against humanity among the offenses for which the German leaders were to be tried by the International Military Tribunal (IMT) in Nuremberg. Article 6(c) of the Tribunal's Charter defined the offense as (Roberts and Guelff 1989):

> Crimes against Humanity: namely, murder, extermination, enslavement, deportation, and other inhumane acts committed against any civilian population, before or during the war, or persecutions on political, racial or religious grounds in execution of or in connection with any crime within the jurisdiction of the Tribunal, whether or not in violation of the domestic law of the country where perpetrated.

Three elements of the definition clearly stood out: (a) 'before or during the war;' (b) 'against any civilian population', and (c) 'whether or not in violation of the domestic law of the country where perpetrated.'

The first element clearly implied that certain inhumane acts constituted international crimes not only in time of war but also during peace. This idea coupled with the reference to the irrelevance of whether such an offense constituted a crime under the law of the country where it was committed, meant that perpetrators could be held criminally responsible under an international law, regardless of domestic-law provisions to the contrary. This could pose a very serious threat to the idea of national sovereignty. However, this potentially revolutionary development was grounded by the definition's requirement that these acts had to be connected with any crime within the jurisdiction of the Tribunal, i.e., crimes against the peace, and/or war crimes. What came to be known as the war or conflict nexus requirement severely limited the scope of crimes against humanity.

The second element meant that any civilian population came under the protection of the provisions relating to crimes against humanity, regardless of its status vis-à-vis the perpetrator, including the civilian population residing in the perpetrator's country. According to the UNWCC, there are two types of crimes against humanity: 'crimes of the "murder-type" such as murder, extermination, enslavement, deportation, etc., and "persecutions." With regard to the latter the provision requires that they must have been committed on political, racial or religious grounds' (the so-called discriminatory element). Undoubtedly, the reference to any civilian population constituted a subversive of state sovereignty development.

Finally, the third element exemplified—as aforementioned—the growing belief that certain acts violative of elemental notions of human dignity could not enjoy immunity from the international prosecution even if legal under domestic law (UNWCC 1948).

Despite its subsumption under war crimes, and the debates surrounding its conceptual contours, the term crimes against humanity had an immediate effect on the quest for a postwar normative framework. The international community determined—in the words of the UN Charter—'to reaffirm faith in fundamental human rights, in the dignity and worth of the human person' specified one of the United Nations' goals as the promotion and encouragement of human rights. Any major commitment to the promotion of human rights was inherently subversive of state prerogatives: by subjecting the behavior of national governments towards their own populations to international scrutiny, the international community was challenging the basic parameters of the Westphalian order. The first major manifestations of this trend were: (a) the adoption, by the UN General Assembly, of the Universal Declaration of Human Rights (UDHR) without dissenting vote; and (b) the signature of the UN Convention on the Prevention and Punishment of the Crime of Genocide. The former, which is the foundational document of the modern era of human rights, was proclaimed by the General Assembly 'as a common standard of achievement for all peoples and all nations.' Although it provided for no enforcement mechanisms, the intrusive nature of the human rights discourse was to signal the beginning of a new era. The latter's importance went well beyond the fact that it constituted the first UN-sponsored human rights treaty. According to Article I of the Convention, the crime of genocide was the first crime against humanity to be delinked from the conflict nexus requirement ('… genocide whether committed

in time of peace or in time of war, is a crime under international law …'). Moreover, it was the first crime against humanity which could be committed by both State and non-State agents. These developments pointed to the importance of the substantive intersection between the laws of war and international human rights law.

3. *Recent Developments*

The laws of war and international human rights law have traditionally addressed separate sets of issues: the former has been concerned with the treatment of combatants and noncombatants by their opponents in wartime (including the methods and means of warfare), while the latter has been concerned with the relationship between states and their own nationals in peacetime, as well as in cases of public emergency. Yet, despite their differences, both bodies of law share a fundamental commitment to human welfare, irrespective of the circumstances under which challenges to that welfare may emerge, and of the particular identities of the potential victims.

This shared commitment is best encapsulated in the aforementioned Martens Clause, and has deepened primarily as a result of the exponential growth of human rights law in the post-1945 period. A good example is Article 3, common to the 1949 Geneva Conventions. This article refers to a minimum set of humanitarian standards applicable to noninternational conflicts. It stipulates that people who, for a variety of reasons, are not involved in the conflict 'shall in all circumstances be treated humanely, without any adverse distinction' of any kind. In some of its key provisions, the influence of the human-rights perspective is clearly in evidence. The references to unconditional humane treatment reflect Article 2 of the UDHR; the references to the prohibition of cruel treatment and torture, as well as those relating to outrages upon personal dignity, reflect Article 5 of the UDHR; and the references to all indispensable judicial guarantees reflect Article 10 and Article 11, Section 1 of the UDHR (Allen et al. 1996).

After genocide (itself an offspring of the crimes against humanity provision of the Nuremberg Charter), no other concept is more firmly anchored in both bodies of law than crimes against humanity. Recent developments have only served to strengthen these connections. The serious violations of international humanitarian law witnessed in the territories of the former Yugoslavia and Rwanda prompted the international community into action. For the first time since the Nuremberg and Tokyo Tribunals, the international community established *ad hoc* international tribunals to deal with these types of offenses. The United Nations Security Council established the ICTY and the ITR with resolutions 827 (1993) and 955 (1994) respectively. Although both Tribunals have jurisdiction over crimes against humanity, and both list the same prohibited acts, there are substantive differences in the language used. In the ICTY Statute, crimes against humanity are the enumerated acts 'when committed in armed conflict, whether international or internal in character, and directed against any civilian population'; while in the ITR Statute, they are the same acts 'when committed as part of a widespread or systematic attack against any civilian population on national, political, ethnic, racial or religious grounds.'

As these passages indicate, the concept has evolved considerably since the Nuremberg Trial. Three issues clearly stand out: first, that there is no international war nexus requirement; crimes against humanity can occur during noninternational conflicts (ICTY Statute). Second, that crimes against humanity can occur during peacetime situations (ITR Statute). Third, that the widespread or systematic attack requirement is considered as a key element of the offense, but the discriminatory aspect is considered as a requirement for all crimes against humanity, not those of the persecution type alone (ITR Statute) (Meron 1998, Fenrick 1999). Moreover, the evolving jurisprudence of the two Tribunals has clarified important rules of customary international law. For example in the Tadic case, the ICTY stated that 'It is by now a settled rule of customary international law that crimes against humanity do not require a connection to international armed conflict. Indeed … customary international law may not require a connection between crimes against humanity and any conflict at all.'

Yet not all judicial pronouncements have been well received by the scholarly community. While the delinking of crimes against humanity from the war nexus requirement was widely considered as long overdue, the ICTY has been criticized on other fronts. In particular, the Tribunal has been faulted for its view that the discriminatory element was a requirement for all crimes against humanity, instead of being confined solely to the persecution-type of offenses (Meron 1998, Fenrick 1999).

The recently adopted Rome Statute of the International Criminal Court (July 1998) constitutes the latest development in this situation. It is the first treaty text which provides a rather comprehensive treatment of the term. Article 7 of the Statute defines crimes against humanity as 'any of the following acts when committed as part of a widespread or systematic attack directed against any civilian population, with knowledge of the attack.' The list of prohibited acts includes all those listed in the ICTY/ITR Statutes (with some further refinements), and two additional offenses: the enforced disappearance of persons, and apartheid. At this stage, what is emerging from the Rome Statute's definition is a confirmation of both the severance of the term from any type of conflict, and of the discriminatory element's association with persecution-type of offenses only.

4. Concluding Remarks

The controversy surrounding the conceptual contours of crimes against humanity continues. Some analysts have argued that the way to deal with the conceptual ambiguities and the decreasing differentiation between crimes against humanity and war crimes, is to replace the latter term with the former (Green 1997/98). This view has been criticized on several grounds, the most important being that in war, death or injury of civilians is not unlawful, if occurring in the context of an attack against a legitimate military objective (Fenrick 1999). Yet, the fact that such a view is seriously proposed is indicative of the way in which the term has evolved in the modern era. It is a development which can not be fully understood or appreciated, save in the context of the growing influence of international human rights law. After all, the term laws of humanity refers to transcendental humanitarian principles not dependent on particular circumstances. It is a message reflected in the reconceptualization of crimes against humanity, the generic term encompassing the offenses against the laws of humanity. It is both promising and slightly ironic that the more the term evolves, the more it seems to reconfirm the need to transcend contextual factors, and reaffirm the international accountability of perpetrators of inhumane acts.

See also: Civil Liberties and Human Rights; Human Rights, Anthropology of; Human Rights in Intercultural Discourse: Cultural Concerns; Human Rights: Political Aspects; War: Causes and Patterns; War Crimes Tribunals

Bibliography

Adachi S 1988 The Asian concept. In: UNESCO *International Dimensions of Humanitarian Law*. Henry Dunant Institute, Geneva

Allen R, Cherniack M, Andreopoulos G 1996 Refining war: Civil wars and humanitarian controls. *Humans Rights Quarterly* **18**: 747–81

Fenrick W J 1999 Should crimes against humanity replace war crimes? *Columbia Journal of Transnational Law* **37**: 767–85

Green L C 1998 'Grave breaches' or crimes against humanity? *8 USAFA Journal of Legal Studies 19* http://web.lexis-nexis.com/universe/docu

Greenwood C 1995 Historical development and legal basis. In: Fleck D (ed.) *The Handbook of Humanitarian Law in Armed Conflicts*. Oxford University Press, New York

Hartigan R S 1983 *Lieber's Code and the Law of War*. Precedent, Chicago

Holls F W 1900 *The Peace Conference at the Hague and its Bearings on International Law and Policy*. Macmillan, New York

Kalshoven F 1987 *Constraints on the Waging of War*. International Committee of the Red Cross, Geneva

Meron T 1998 War crimes law comes of age. *92 American Journal of International Law 462* http://web.lexis-nexis.com/universe

Pictet J 1988 Introduction: Humanitarian ideas shared by different schools of thought and cultural traditions. In: UNESCO *International Dimensions of Humanitarian Law*. Henry Dunant Institute, Geneva

Roberts A, Guelff R (eds.) 1989 *Documents on the Laws of War*. Clarendon, Oxford

Ruda J M 1988 The Latin American concept of humanitarian law. In: UNESCO, *International Dimensions of Humanitarian Law*. Henry Dunant Institute, Geneva

Schwelb E 1946 Crimes against humanity. *British Year Book of International Law* **23**: 178–226

Scott J B (ed.) 1918 *The Hague Conventions and Declarations of 1899 and 1907*. Oxford University Press, New York

Ticehurst R 1997 The Martens clause and the laws of armed conflict. *International Review of the Red Cross* **317**: 125–34

United Nations War Crimes Commission 1948 *History of the United Nations War Crimes Commission and the Development of the Laws of War*. HMSO, London

Van Schaack B 1999 The definition of crimes against humanity: Resolving the incoherence. *Columbia Journal of Transnational Law* **37**: 787–850

G. J. Andreopoulos

Old Age and Centenarians

Whereas it is obvious that an individual must have lived 100 years in order to be called a centenarian, there is little consensus about the definition of the characteristics of old age or the chronological age associated with its average onset. Classification of old age is based variously on sociodemographic characteristics (e.g., retirement), biological and health factors (onset of physical decline, multimorbidity, and frailty), and subjective age identification. Depending on country, culture, birth cohort, and idiosyncratic factors, the period of old age could extend over 20 to 40 years of an individual's life (e.g., age 60 to 100+). Cultural and scientific knowledge about old age and centenarians lags behind demographic changes such as increased average life expectancy and the expansion of the population over age 60 in many countries. We know relatively little about normative and differential psychological change during old age and even less about predictors of longevity.

1. Is Old Age One or More Life Phases?

Several authors propose that it might be appropriate to consider old age as involving not one but several life phases. Some researchers distinguish the *Young Old* from the *Old* and the *Oldest Old* (e.g., Neugarten 1974, Suzman et al. 1992) whereas others refer to a *Third Age* and *Fourth Age* (e.g., Laslett 1991, Baltes 1997, Baltes and Mayer 1999).

What is the basis of these proposed classifications? Neugarten (1974) coined the term young old to identify

a group of retirees that was not so much identifiable in terms of chronological age but rather by their relatively good health status and social engagement. Based on demographic features, Suzman et al. (1992) identified 85 years of age as the criterion of membership in the oldest old category. The population over age 85, unlike any other age strata, is typically characterized by a unique excess of women over men, higher levels of comorbidity and institutionalization, and greater consumption of medical and care services. Individuals born between 1890–1915 who comprise the oldest old observed in 2000 also exemplify a set of characteristics that are cohort-specific: for example, they have a lower level of education compared with subsequent cohorts, and a higher likelihood of longterm widowhood.

The term *Third Age* was used in the title of Les Universités du Troisième Age in France in the 1970s, and later in 1981 at Cambridge University to characterize a program designed for an older (senior) group of students with the end-purpose not being professional entry but rather lifelong personal education (Laslett 1991). Laslett described the Third Age as an era of personal achievement and fulfillment and the *Fourth Age* as an era of final dependence, decrepitude, and death. In Baltes' (1997) use of the distinction, the criterion is a dynamic and changing one. He proposes that the onset of the Fourth Age be considered as the chronological age where 50 percent of the birth cohort has died. At that age, processes of biological aging can be expected to be predominant in the aging process.

At the end of the 1990s, a number of cross-sectional and longitudinal studies began to focus not only on the demographic, health, and social distinctions between the proposed phases of old age, but also on possible differences in psychological profiles. Examples include representative surveys such as Asset and Health Dynamics of the Oldest Old (AHEAD: Soldo et al. (1997), the Berlin Aging Study (BASE: Baltes and Mayer 1999), Danish Twin Study (Andersen-Ranberg et al. 1999), and the Swedish Adoption/Twin Study of Aging (SATSA: McClearn et al. 1997). Although centenarians have become a target group for biodemographers (e.g., Vaupel et al. 1998), well-designed psychological studies of representative samples of centenarians were rare in the year 2000.

2. Central Theoretical Questions about Old Age

Psychological theory and research on old age has been shaped by three general goals (see *Aging, Theories of*):

(a) to map trajectories of change in specific domains of psychological functioning during old age (e.g., cognition, personality, and social behavior) and determine processes underlying change;

(b) to determine the sources of observed heterogeneity in psychological functioning and longevity in individuals over the age of 70 years; and

(c) to understand systemic interdependencies across domains of functioning and subgroup differences (see *Differential Aging*).

Historically, the prime focus has been to determine average patterns or trajectories of within person (intraindividual) change over time (aging) in terms of level, direction, and rate. Research questions have centered on factors linked to the onset and rate of decline as well as asking whether decline goes hand in hand with structural dedifferentiation (e.g., reduced complexity) or structural disorganization (breakdown).

2.1 Proposals about Trajectories of Change

Although decline is the expected trajectory of change in old age, there are several proposals about differences in the nature, rate, and onset of decline. To begin, it is considered important to separate decline trajectories that are pathologically determined from those that are normative (i.e., usual and age-related) and to determine to what extent interventions (e.g., memory training, social contact) can modify the onset and rate of decline (e.g., Rowe and Kahn 1987). Analogous to the study of early versus late (delayed) development or maturation in early life, there is an interest in the identification of factors linked to the delayed onset of multiple negative changes in old age (e.g., compression of morbidity; see Fries 1980). Death-related trajectories (terminal decline) have been distinguished in intellectual functioning and some indicators of perceived well-being (e.g., subjective health, depressivity; see Berg 1996). These decline trajectories are larger and qualitatively different from normal age changes and occur during a window of 5–7 years before death. Birren and Cunningham (1985, p. 23) suggested that terminal decline represented the emergence of a cascade of structural change.

Trajectories reflecting a relative maintenance of functioning and delayed decline have been labeled 'successful aging.' These provide insight into what might, in principle, be possible for increasing numbers of older persons, given the evolution of social and cultural environments supportive of old age (Baltes 1997). Baltes characterized successful aging as the potential to maximize the gains of a long life and minimize the losses inherent in old age (see *Human Development, Successful: Psychological Conceptions*). Successful aging is indicated by high life quality, subjective well-being, relative good health, and longevity. Rowe and Kahn (1997) suggested that these outcomes reflect the interrelationships of three components: low risk of disease and disease-related disability; high cognitive and physical functional capacity; and an active engagement with life (especially positive interpersonal relations and productive activity).

Do trajectories of change differ among domains of psychological functioning in old age? Research pub-

lished until 2000 suggests that, at least up to the early eighties, cognitive functioning is more vulnerable to decline than is the case for other areas of psychological functioning such as personality, well-being, or social relationships (Baltes and Mayer 1999; see *Cognitive Aging* and *Personality Development in Adulthood*). A multitude of training and intervention studies have shown that although some aspects of cognitive decline can be alleviated or compensated, it is not possible to restore performance on cognitive tasks involving memory or perceptual speed to levels that might have been achieved by individuals when they were younger (see *Aging and Education*). Personality dispositions are thought to show structural and interindividual stability at least into the late seventies. Analyses of large cross-sectional and longitudinal samples of individuals aged 35–84 years suggest age-related changes in the expression of the five major personality dispositions: age gradients for Extraversion, Neuroticism, and Openness are generally negative, and those for Agreeableness and Conscientiousness are positive (e.g., McCrae et al. 1999). Proposals about changes in subjective and emotional well-being in the period of very old age are inconsistent (e.g., Diener et al. 1999). Whereas well-being may be sustained in the Third Age, the Fourth Age appears to be a phase of life that places great demands on the self-regulatory capacities of individuals.

2.2 Heterogeneity in Old Age

Gerontologists and lifespan theorists generally assume that interindividual differences are due to the cumulative effects of biogenetic, health, and personal life history factors. This assumption opens a realm of questions about the distribution of interindividual differences and whether this distribution expands or contracts in range in old age. Three positions on these questions reflect different lenses on the forces of heterogeneity across the life course.

The first position argues that early adult life is a phase when within-cohort heterogeneity expands greatly, and that the consequences of this expansion are maintained for the rest of life. Young adulthood, for example is a life phase that stimulates the initiation of a wide range of life courses in terms of education, work, family, and health behaviors (e.g., Elder 1998, Vaupel et al. 1998). The second position suggests that interindividual heterogeneity may increase with age (e.g., Dannefer 1988). The combined effects of unique experiences over many years should magnify differences between individuals. Genetically-based differences would also have had more time to be expressed and to contribute to divergence. Moreover, older adults, somewhat freer from social constraints, might be more likely to select their own courses of action. Counter to these two positions is a third proposal that, at least among the oldest old, we may observe reduced heterogeneity for those aspects of functioning that predict longevity because of selective survival (see Sect. 3, however).

2.3 Subgroup Systemic Differences in Functional Profiles

Gerontological researchers have long searched for subgroups exemplifying patterns of aging associated with differential survival, pathology, and lifestyle (e.g., Williams and Wirths 1965). Underlying this search is the concept that, over time, subgroups reflect the systemic operation of different combinations of psychological profiles. Early ideas about successful and less-successful lifestyles, and subgroups of normal versus pathological agers have been advanced in subsequent research with representative samples (e.g., Rowe and Kahn 1997). Smith and Baltes (1997), for example, reported nine subgroups extracted from the Berlin Aging Study, four reflecting different patterns of desirable functioning and five less desirable functioning. The relative risk of membership in the less desirable profile groups was 2.5 times higher for the oldest old than for those aged 70–84, suggesting a systemic transition in psychological functioning in the Fourth Age. It is an open question whether centenarians share a common profile of psychological characteristics (e.g., intellectual abilities, personality characteristics, lifestyle behaviors, belief systems) that contribute to and afford adaptation to a long life.

3. Research Issues

Researchers interested in understanding the processes underlying trajectories of change and subgroup differences in old age are faced with the difficult task of separating the effects of several explanatory sources: factors that are age-related, pathological, death-related, cohort-related (see *Life Course: Sociological Aspects*), and factors associated with sample selectivity and selective mortality.

In research design, chronological age is often used as a proxy for the time interval required for change to occur, and researchers translate their findings about change (or age differences in cross-sectional studies) into age functions or age trajectories. Some processes of change in old age, however, may instead be time-dependent (e.g., require a certain length of time to occur) or death-related (e.g., co-vary with factors associated with distance from death). Self-regulation is an example of a time-dependent process: Time is required to adjust to a life event or new health-related constraint. Some time-dependent processes may also interact with age (e.g., the young old and oldest old may differ in adaptation or recovery time).

The extent to which selective survival represents alternative explanations for observed group differ-

ences in cross-sectional and longitudinal designs is a general problem of research on advanced old age. Selective survival effects can be manifested in means, variances, and the patterns of intercorrelations between variables, and will appear for all variables that co-vary with the variables on which the selection is defined (Nesselroade 1988). For each year above average life expectancy, individuals who survive represent a proportionately more select subgroup of their birth cohort. Because data on the nonsurviving members of the cohorts are usually not accessible, we do not know whether the selection effects nullify, create, or diminish observed differences. With regard to gender, the suspicion is that the 'structural' compositions of samples of older men and older women differ because of gender differentials in longevity (Smith and Baltes 1998). Compared with older women, samples of older men are probably more positively selected on factors that contribute to longevity (especially sex-specific factors). Precisely what these factors might be is an open question (see Wachter and Finch 1997).

The selective survival component clearly sets a constraint on the interpretation of findings about age and gender differences in the period of very old age. A further constraint is introduced by sample selectivity (i.e., the question whether the younger and older men and women who agreed to participate in the study represent the heterogeneity of their respective surviving population subgroups). Sample selectivity is a particular issue in many studies of centenarians. If researchers screen for dementia, the psychological profiles that are reported from many studies only reflect cognitively well-functioning centenarians (estimated dementia prevalence at age 100 is 40–60 percent). Furthermore, time of data collection together and cohort-specific experiences make it difficult to generalize findings about predictors of longevity.

4. Future Directions

Ideally, we need a combination of experimental work and multiwave longitudinal data collected over many years, indeed until the deaths of all participants, in order to understand fully the mechanisms and processes of psychological change in old age. Longitudinal studies should be designed so that time intervals between measurement are age-appropriate: A 7-year interval might be sufficient to observe significant change in the seventies, but will contribute to sample attrition due to mortality among 90-year-olds. A new set of studies that follow individuals 90+ years of age until their death and collect extensive information about psychological functioning would provide valuable information about possible trajectories of change that distinguish centenarians from their long-lived peers who die prior to age 100. As we begin to understand more about the sources and consequences of different psychological change trajectories, we will be in a better position to intervene in the cultural and social context of old age. Revised social judgments about autonomy and dependency in old age (M. M. Baltes 1996) and the creation of supportive life environments to compensate for personal frailty may well contribute to higher life quality and a longer life in the Fourth Age of the future (e.g., Baltes 1997, Olshansky et al. 1998; see *Ecology of Aging*).

See also: Aging and Health in Old Age; Ecology of Aging; Life Course: Sociological Aspects; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development; Personality Development in Adulthood; Plasticity in Human Behavior across the Lifespan; Well-being (Subjective), Psychology of

Bibliography

Andersen-Ranberg K, Christensen K, Jeune B, Skytthe A, Vasegaard L, Vaupel J W 1999 Declining physical abilities with age: A cross-sectional study of older twins and centenarians in Denmark. *Age and Ageing* **28**: 373–7

Baltes M M 1996 *The Many Faces of Dependency in Old Age*. Cambridge University Press, New York

Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 2366–80

Baltes P B, Mayer K U (eds.) 1999 *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, New York

Berg S 1996 Aging, behavior, and terminal decline. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*. Academic Press, San Diego, CA, pp. 323–37

Birren J E, Cunningham W 1985 Research on the psychology of aging: Principles, concepts, and theory. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 2nd edn. Van Nostrand Reinhold, New York, pp. 3–34

Dannefer D 1988 Differential gerontology and the stratified life course: Conceptual and methodological issues. *Annual Review of Gerontology and Geriatrics* **8**: 3–36

Diener E, Suh E M, Lucas R E, Smith H L 1999 Subjective well-being: Three decades of progress. *Psychological Bulletin* **125**: 276–302

Elder G H 1998 The life course and human development. In: Lerner R M (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. 1, pp. 939–1028

Fries J F 1980 Aging, natural death, and the compression of morbidity. *New England Medical Journal* **303**: 130–5

Laslett P 1991 *A Fresh Map of Life: The Emergence of the Third Age*. Harvard University Press, Cambridge, MA

McClearn G E, Johansson B, Berg S, Pedersen N L, Ahern F, Petrill S A, Plomin R 1997 Substantial genetic influence of cognitive abilities in twins 80 or more years old. *Science* **276**: 1560–3

McCrae R R Jr, Costa P T Jr, de Lima M, Simoes A, Ostendorf F, Angleitner A, Marusic I, Bratko D, Caprara G V, Barbaranelli C, Chae J-H 1999 Age differences in personality across the adult life span: Parallels in five cultures. *Developmental Psychology* **35**: 466–77

Nesselroade J R 1988 Sampling and generalizability: Adult development and aging research issues examined within the

general methodological framework of selection. In: Schaie K W, Campbell R T, Meredith W, Rawlings S C (eds.) *Methodological Issues in Aging Research*. Springer, New York, pp. 13–42
Neugarten B L 1974 Age groups in American society and the rise of the young-old. *Annals of the American Academy of Politics and Social Sciences* **415**: 187–98
Olshansky S J, Carnes B A, Grahn D 1998 Confronting the boundaries of human longevity. *American Scientist* **86**: 52–61
Rowe J W, Kahn R L 1987 Human aging: Usual and successful. *Science* **237**: 143–9
Rowe J W, Kahn R L 1997 Successful aging. *Gerontologist* **37**: 433–40
Smith J, Baltes M M 1998 The role of gender in very old age: Profiles of functioning and everyday life patterns. *Psychology and Aging* **13**: 676–95
Smith J, Baltes P B 1997 Profiles of psychological functioning in the old and oldest old. *Psychology and Aging* **12**: 458–72
Soldo B J, Hurd M D, Rodgers W L, Wallace R B 1997 Asset and health dynamics among the oldest old: An overview of the AHEAD Study. *The Journals of Gerontology: Series B* **52B** (Special issue): 1–20
Suzman R M, Willis D P, Manton K G (eds.) 1992 *The Oldest Old*. Oxford University Press, New York
Vaupel J W, Carey J R, Christensen K, Johnson T E, Yashin A I, Holm N V, Iachine I A, Kannisto V, Khazaeli A A, Liedo P, Longo V D, Zeng Y, Manton K G, Curtsinger J W 1998 Biodemographic trajectories of longevity. *Science* **280**: 855–60
Wachter K, Finch C E 1997 *Between Zeus and the Salmon: The Biodemography of Aging*. National Academy of Sciences, Washington, DC
Williams R H, Wirths C G 1965 *Lives Through the Years: Styles of Life and Successful Aging*. Atherton Press, New York

J. Smith

Olfactory System

The olfactory system belongs to the chemical senses (smell, taste). We are dealing here with a sensory system whose receptive field is situated in the upper part of the nose and whose information is taken up and transmitted via a CNS protrusion, the olfactory nerves, or olfactory bulb. This sensory function essentially serves the purpose of self-preservation by enabling individuals to recognize food, assess its quality, and receive warning signals (burnt smell, poisons). The stimulus threshold is extremely low in humans, and yet humans are poor smellers compared with animals. The stimulus threshold varies considerably according to the nutritional status.

The nerve fibers of the olfactory system are spread much wider in the CNS than is the case with any other sensory system. The olfactory system exerts an essential part of its influence on subcortical centers like the limbic system. Thus, smell has a high emotional component, which accounts for its great social importance. This explains why olfactory impressions are used to express behavior patterns in all languages. In German, for example, the expression *Ich kann Dich nicht riechen* (I cannot smell you) means 'I don't like you.'

Humans have special inherited and acquired preferences (hedonics). Thus, for example, the smell of a forest elicits a positive emotion in many people, whereas negative feelings are associated with the smell of isobutyric acid (sweat).

Because of the broad spread of the secondary and tertiary olfactory fields in the CNS, an odorant can evoke its strong conscious reaction only initially; the sense of smell must then adapt rapidly to be ready for new odorants, and we lose the perception.

Apart from the olfactory sense for self-preservation, referred to henceforth as the 'classical olfactory system,' the nose has a second chemical sense that long received little attention in humans: the vomeronasal organ or system. Its purpose lies in the perception of pheromones within the sexual sphere and thus in the propagation of the species as well as in signaling and controlling aggressions. While hormones distribute messages within the body, pheromones take over communication to the outside (Jellinek). Little is known about the neuronal pathways and connections of the vomeronasal system. An intensive and exclusively subcortical process is probably involved, which means that we are not consciously aware of the stimulation of this organ. As with the classical olfactory sense, marked hedonics also exist here that are controlled by the immunological system.

1. *Anatomy of the Classical Olfactory System*

The zone of smell of the classical olfactory system in the nose lies in the upper part of the nasal cavity. It covers an area of about 5 cm on either side of the nose, specifically in the region of the superior nasal conchae, the septum and the ethmoid bone (Fig. 1). This region harbors about 100 million bipolar olfactory cells in humans (220 million in dogs). The sensory epithelium also contains stem cells with the lifelong capacity for replacing olfactory cells that have died. The peripheral processes of the bipolar sensory cells rise to the mucosal surface (Fig. 2). They bear up to 1,000 hairs (cilia) on which the primary contact takes place between aromatic molecules and receptors. The hairs differ in length among the various organisms. The longer the hairs, the more points of contact for the aromatic molecules. Organisms with long sensory hairs are therefore described as macrosmatic, those with short ones as microsmatic. The hairs must be moistened constantly so as not to be destroyed by desiccation. The olfactory zone thus has special mucous glands which, in contrast to other glands in the nose, do not form proteolytic enzymes. Odorants would otherwise be broken down into substances with a different smell prior to contact with the sensory cells.

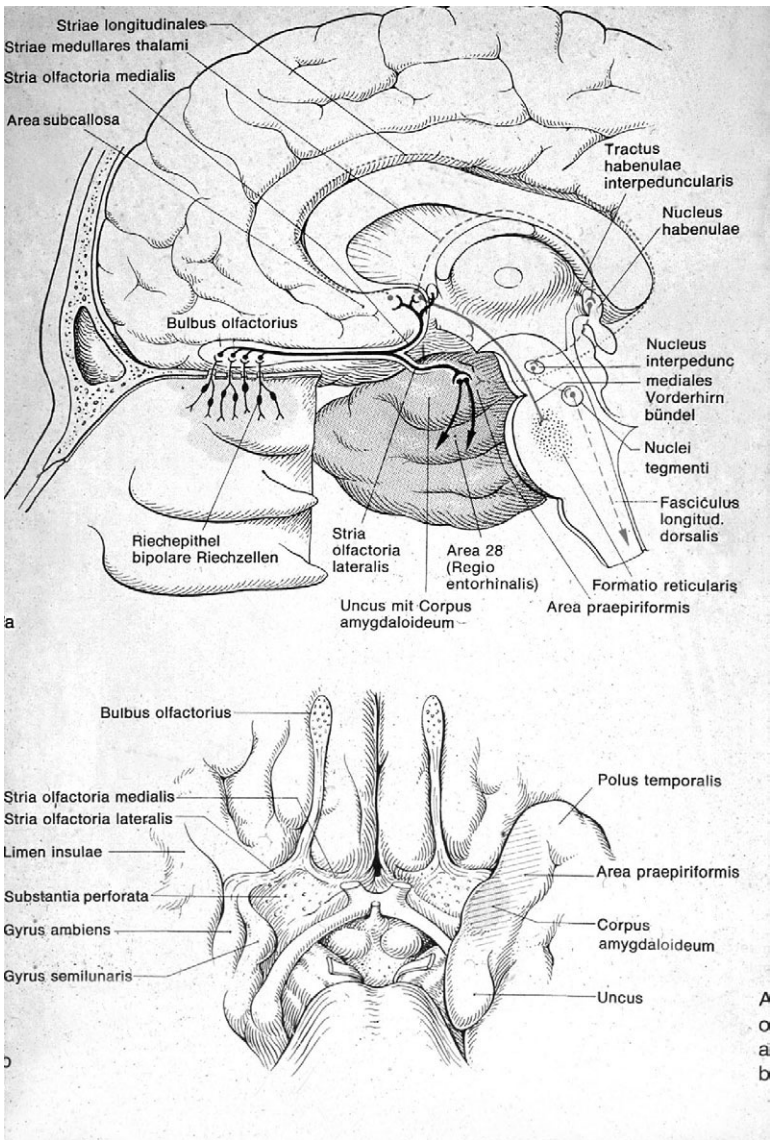

Figure 1
Anatomy of the nose with the olfactory zone

The aromatic molecules dock at special receptor proteins situated on the hairs. These proteins activate signal proteins (G proteins) which in turn trigger a cascade that sends an electric signal (action potential) along the central process of the cell to the bulb of the olfactory nerve (Reed 1992). This bulb contains cell accumulations (granules), each granule corresponding to a specific receptor molecule (Reed 1992). The expression of the specific olfactory receptor proteins is genetically determined (Buck and Axel 1991), one gene encoding each receptor molecule. Altogether, we possess about 1,000 different olfactory receptors. Approximately 1 percent of our genes are used here, which corresponds to the largest family of genes found in humans. This underscores the profound importance of the sense of smell in man (Axel 1995).

There are about 30,000 odorants. Humans can recognize about 3,000 of these but only differentiate 300. We have a poor memory for smells, a much poorer olfactory than visual memory. Thus, we can only remember about 20 odorants.

Odorants must be volatile and lipid-soluble, otherwise they cannot be perceived. However, perception does not take place solely in the area of the olfactory

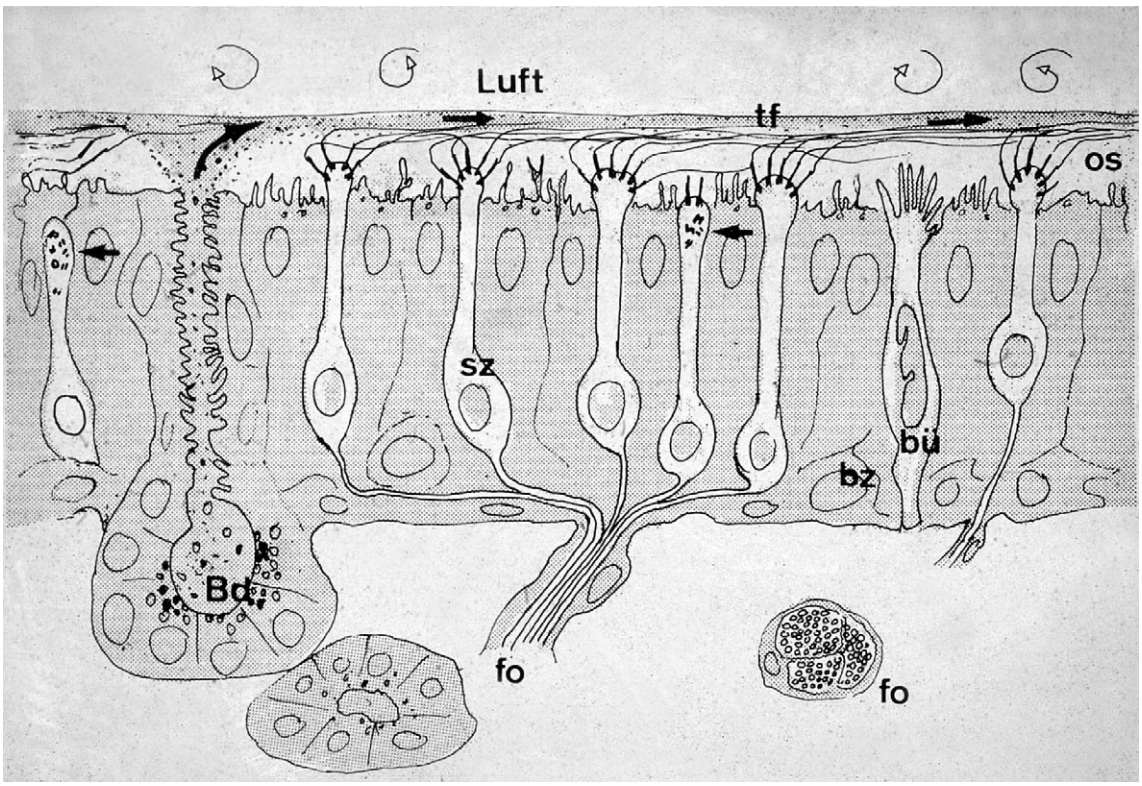

Figure 2
Detail of the olfactory mucosa with ciliated bipolar sensory cells. The cilia differ in length. (a) cilia in microsmatic organisms (for example, humans); (b) cilia in macrosmatic organisms (for example, dogs, eels, salmons)

sensory cells. Thus, it is possible to distinguish between odorants that excite the olfactory system almost exclusively, for example, coffee, cinnamon, tar, or the scent of roses, and odorants that stimulate both the olfactory system and fibers of the sensory trigeminal nerve in the nose, for example, menthol or camphor (sensation: cool) or formalin and vinegar (sensation: pungent). In addition, there are odorants that stimulate both the olfactory system and taste fibers on the base of the tongue, for example, pyridine (sensation: bitter) and chloroform (sensation: sweet).

2. *Anatomy of the Vomeronasal System*

The vomeronasal organ is situated in the anterior and lower mucosal areas of the nasal septum (Fig. 3). It comprises a 2–8 mm pouch with a diameter of 0.2–2 mm. The numerous mucous glands opening into this pouch ensure constant irrigation. At the end of the pouch there are elongated sensory cells with small mobile saccules resembling handles. Numerous blood vessels and nerve fibers are situated basal to this cell layer. The precise spread of these fibers and of the information thus transmitted to the CNS is not known. However, stimulation of the vomeronasal organ is not perceived consciously.

The vomeronasal organ develops between the 37th and 43rd prenatal day in humans (Smith and Bhatnagar 2000). At this time, it is morphologically identical to its counterpart in animals, for which chemosensory activity has been demonstrated. Its structure changes between the 12th and 14th week. The mucosal cells develop cilia and lose contact to nerves. Thus, there are discussions as to whether the organ loses its function in humans at a very early stage, that is, in infancy. However, some findings (Monti-Bloch et al. 1998, Grosser et al. 2000) have demonstrated that the vomeronasal organ can be stimulated directly and specifically by a pheromone in adults (see below).

3. *Measuring Methods*

Odorants of graded concentration (e.g., 'Sniffin' sticks, butanol test (Gudziol) are inhaled to test perception within the scope of the classical olfactory system (Hummel et al. 1997). The results are based largely on subjective data. Since this leaves open the possibility

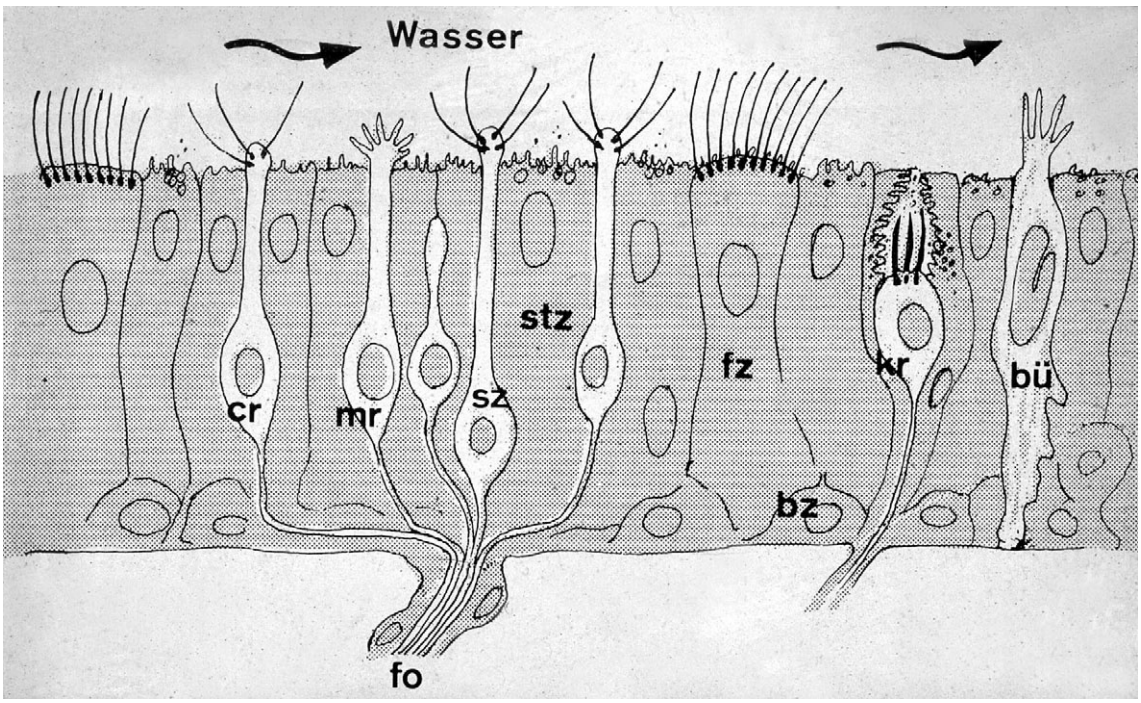

Figure 3
Anatomy of the nasal septum

of simulation and aggravation, expert opinions always require clarification by taste tests with substances that can only be smelled (e.g., peppermint solution) or by objective examination procedures. In the objective smell test, an odorant is introduced into the nose by pulsed application, and the brain waves are measured via the vertex synchronous to inhalation. The evoked test curve obtained by summation indicates the activation of the olfactory system—but not necessarily the perception of the odorant—in the cerebral cortex. Moreover, no qualitative statements can be made on the question of whether an odorant is perceived as such.

4. *Physiology of the Classical Olfactory System*

Very complex apparatuses were used to investigate the threshold of the classical olfactory sensation. It is $1–2 \times 10^{-14}$ mol/l in humans and 2×10^{-16} mol/l in eels. From the fact that only 10 molecules excite a receptor cell and 40 trigger a smell perception, it may be calculated that the olfactory system of the eel will be stimulated by even a thimbleful of rosewater poured into 16 times the amount of water in Lake Constance and well distributed (Jung 1994). This good olfactory sense of fish surely also explains why a salmon usually finds his home river. It probably smells its specific odorants, and this should give us at least some idea of the impact our water pollution has on the olfactory system of fish.

Measurements have shown that the olfactory threshold changes significantly in relation to the nutritional status. Thus, the threshold is $1–2 \times 10^{-14}$ mol/l in a fasting state and rises to $3.5–5.5 \times 10^{-14}$ mol/l after a meal. The biochemical basis for this threshold variation is the sugar content in the blood. Sugar is the nutritive substance of the CNS. It makes sense for the olfactory threshold to differ in the fasting state: hungry humans and animals require a good olfactory sense for self-preservation, since they must hunt for 'prey' or use the olfactory organ for orientation (salmon fast when returning to their home river; their olfactory sense is thus optimal).

5. *Physiologically-based Cooking*

One of the secrets of the famous art of modern French cooking (*nouvelle cuisine*) is to maintain the fasting state and thus the good olfactory sense of the guest by serving very small portions in the individual courses. It is especially important to go easy on sweet foods and to serve dry sherry as an aperitif, for example. Sweet foods should only be consumed as a dessert. Conversely, serving a sweet aperitif, for example, port wine

or a rich appetizer can mask the taste of a poorly seasoned dish.

There are individuals with a particularly keen olfactory sense. They are usually highly paid for testing wines and other luxury items like coffee or tea or may also work as perfumers. On the other hand, there are people who cannot perceive certain odorants. There are, for example, such specific, genetically determined anosmias for isobutyric acid (sweat), hydrogen cyanide (bitter almond), and the scent of freesias. More than 80 specific anosmias have been identified thus far.

Contrary to subjective reports by women, no change has been detected in the classical olfactory sense during the menstrual cycle (Hummel et al. 1997).

In 1986 *National Geographic* magazine published an article on the sense of smell in humans. For illustration, they included six encapsulated odorants to be identified after being released by rubbing. This evolved into a gigantic global smell test with 1.5 million participants. Some of the results are physiologically significant. The test clearly demonstrated, for example, that men have a poorer sense of smell than women (something women have always said). This difference is enhanced in old age. The olfactory sense is best at the age of 20 and declines continuously with age (about 40–50 percent by the age of 80). On the other hand, the ability to smell the scent of a rose hardly decreases at all. Roses are, therefore, a suitable present for older people. Smokers consider themselves poor smellers and they are quite right.

Different odorants trigger completely different physical reactions in humans. The scent of jasmine and lemon increases alertness, while that of lavender and roses have a calming effect.

A Japanese company introduces a lemon scent into their air conditioning system in the morning to increase motivation, the scent of roses at noon to calm tempers, and different wood fragrances in the afternoon to mobilize energy reserves (PM 1/1990).

Buying behavior can also be stimulated. Adding aromatics to the air lengthens customers' shopping time by 16 percent, makes them 15 percent more willing to buy, and increases sales by 6 percent (Stöhr 1995).

There are many secretory glands in the olfactory zone of the nasal mucosa. Their primary function is to moisten and thus protect the cilia of the olfactory cells, but they also rinse out odorants. Thus, the organ of smell very quickly regains its readiness for a new stimulation. The olfactory sense adapts very quickly though to odorants that are present constantly. In other words, an odorant is no longer perceived as such after a few minutes. A dog that has taken up a trail must, therefore, run in zigzags, since he would otherwise lose the scent by adaptation of his olfactory sense. This also applies to perfume users. A person smelling a perfume will no longer perceive it after a short time if the wearer remains at a constant distance.

Hedonics are understood to be the assessment of smell sensations. They can be positive or negative but differ from organism to organism and also among members of the same species. Thus, the smell of decay is negative for humans but positive for vultures. Europeans value the smell of pizza spices more highly than that of soybean oil. The opposite is true for Asians. Hedonics can be inherited (smell of decay) and acquired and are already passed on during gestation. Fetuses are known to react to olfactory stimuli in the mother after the 28th week of pregnancy. Thus, newborn rabbits already show a selective preference for the food their mothers preferred.

6. *Physiology of the Vomeronasal System*

The literature on the anatomy and physiology of the vomeronasal organ in humans is extensive and in part extremely controversial. Thus, Shoyi et al. found that the vomeronasal system has very high sensitivity to various stimulating substances, though the sensory cells in the nasal pouch do not possess distinct cilia like the olfactory cells of the classic olfactory system. In contrast, Smith and Bhatnagor (2000) already detected a transformation into ciliated cells in infancy. On the one hand, a physiological receptor function has been negated in man (Trotier et al. 2000). On the other hand, women randomized in a double-blind manner reacted specifically to the male pheromone androstadienone when applied to their vomeronasal organ in picogram quantities. Direct application of the female pheromone pregna-4,20-diene-3,6-dione (vomeropherin PDD) to the vomeronasal organ of men elicited a significant and dose-dependent parasympathomimetic reaction and also decreased serum LH and testosterone (Monti-Bloch et al. 1998). The threshold for pheromone detection is extremely low in animals. Hence, a male silk moth will find a female even at a distance of 10 km, though not if it is sitting under a glass bell. As mentioned above, the stimulating substances of the vomeronasal system are pheromones, androstenone, and androstenol being among those found in humans. They are formed in the nasolabial groove, the axilla, the mammillae and the sexual organs. We do not perceive consciously their stimulation of the vomeronasal organ. If we think we are smelling pheromones, these are bacterial breakdown products of the respective substances, for which we also develop marked hedonics. Famous in this connection is a letter Napoleon wrote to his wife, Josephine, requesting that she no longer wash herself because he was coming home from the field.

Pheromones control not only sexual activity but also aggressive behaviors among conspecifics (Dulac 2000). Dogs mark their territory, for example. Smelling pheromones is particularly important for newborn creatures, especially while still helpless, for example, blind animals. They smell their mother's mammary glands via pheromones. Kangaroo babies are very

small creatures at birth. Guided by pheromones, they find their way unaided to the mother's pouch, where the mammary glands are situated.

Hedonics for pheromones are controlled genetically (Wedekind and Penn 2000, Wedekind and Furi 1997). They are influenced by the immune system. Thus, hedonics for the pheromones of a partner are known to be all the more positive, the greater the difference between the immunocompetence of a couple, as measured by the major histocompatibility complex. This is nature's attempt to prevent marriages between brothers and sisters. Hedonics has been found to undergo a reversal in the course of pregnancy, during which time a previously preferred partner is not very attractive. This effect is biologically understandable. The reversal only lasts as long as the pregnancy. However, women taking oral contraceptives also show a reversal of hedonics, and this circumstance harbors an implicit threat to their partnership.

7. *Encyclopedia and Definition of Olfactory Disorders*

Normosmia: normal sense of smell
Anosmia: total absence of the sense of smell
Dysosmia: partial impairment of the sense of smell (hyperosmia, hyposmia)
Anosmia and Dysosmia: specific, genetic: lack of receptor proteins

Respiratory: – marked deviations of the nasal septum
– turbinate hyperplasia
– mucosal swelling (for example, in colds)
– nasal polyps
– tracheotomized patients

Epithelial: – drying of the nose (ozena)
– exogenous noxae (cadmium, chromic acid, butylene glycol, benzoic acid, hydrogen selenide (sometimes even one breath of a noxa will permanently destroy the sense of smell)
– overgrowth of the olfactory epithelium by normal nasal mucosa

Neural: – virus infections, especially those associated with the common cold (rhinoviruses, coronaviruses, adenoviruses)
– traumas in an anteroposterior (for example, boxing) or posteroanterior (falling on the back of the head) direction that shift the brain tangentially to the skull base, rupturing the olfactory nerve fibers at the point where they pass through it (Fig. 1)

Endocrine: – Cushing's syndrome, hypothyroidism, diabetes, B12 deficiency

Central: – traumas, tumors, influenza infection, meningitis

Parosmia: smelling a substance that is not there; The affected individual knows that the substance smelled is not there
Cause: often nasal (partial destruction of olfactory epithelium) but also central

Pseudosmia: misinterpretation of a real olfactory impression
Cause: psychoses

Phantosmia: olfactory hallucinations, for example, cacosmia. The affected individual does not complain about a defective sense of smell and refuses to believe he is mistaken
Cause: Central lesions
temporal lobe tumor or abscess (as an aura in uncinate fits)
schizophrenic psychoses

See also: Appetite Regulation; Emotion, Neural Basis of; Food Preference; Hunger and Eating, Neural Basis of; Limbic System; Memory Psychophysics; Psychophysics; Psychophysiology; Taste, Neural Basis of

Bibliography

Axel R 1995 The molecular logic of smell. *Science America* **273**(4): 154–9

Breitmaier E, Terpene A 1999 *Duefte, Pharmaka, Pheromone*. Teubner, Stuttgart, Leipzig, Germany

Buck L, Axel R 1991 A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* **65**(1): 175–87

Dulac C 2000 Sensory coding of pheromone signals in mammals. *Current Opinions in Neurobiology* **10**(4): 511–8

Grosser B I et al. 2000 Behavioral and electrophysiological effects of androstadienone, a human pheromone. *Psychoneuroendocrinology* **25**(3): 289–99

Gudziol H *Butanol-Schwellentest nach Cain*. Klinikum der Friedrich-Schiller-Universität Jena, Klinik und Poliklinik für Hals-, Nasen-, Ohrenkrankheiten, Germany

Hummel T et al. 1997 'Sniffin' sticks': Olfactory performance assessed by the combined testing of odor identification, odor discrimination and olfactory threshold. *Chemistry of Senses* **22**(1): 39–52

Jellinek S Menschliche Pheromone: Ein Durchbruch in der Parfuemerie? Dragoco Report 1/99, S5–S28

Jung H 1994 *HNO-Highlights 4Nr. 3*. MMW Verlag, Munich, Germany

Monti-Bloch L et al. 1998 Modulation of serum testosterone and autonomic function through stimulation of the male human vomeronasal organ (VNO) with pregna-4,20-diene-3,6-dione. *Journal of Steroid Biochemistry and Molecular Biology* **65**(1–6): 237–42

Reed R R 1992 Signaling pathways in odorant detection. *Neuron* **8**(2): 205–9

Smith T D, Bhatnagar K P 2000 The human vomeronasal organ. Part II: Prenatal development. *Journal of Anatomy* **197**(3): 421–36
Stöhr A 1995 Olfaktorische Stimulationen am Point of Sale. Arbeitspapier Nr. 16 der Reihe Konsum und Verhalten, Universität Paderborn
Trotier D et al. 2000 The vomeronasal cavity in adult humans. *Chemistry of the Senses* **25**(4): 369–80
Wedekind C, Furi S 1997 Body odour preferences in men and women: Do they aim for specific MHC combinations or simply heterozygosity? *Proceedings of the Royal Society of London Series B: Biological Sciences* **264**(1387): 1471–9
Wedekind C, Penn D 2000 MHC genes, body odours, and odour preferences. *Nephrology Dialysis Transplantation* **15**(9): 1269–71

H. Scherer and A. Quast

Oligarchy (Iron Law)

1. *Michels' Sociology of Political Leadership*

Michels' classic (1959) can be read as an extended argument that continually returns in various forms to the following syllogism:

(a) strong organization is necessary to achieve collective democratic goals,

(b) strong organization requires delegation of popular sovereignty to an autonomous and self-perpetuating leadership cadre, and,

(c) an autonomous and self-perpetuating leadership cadre is incompatible with true democracy.

The paradox in the syllogism is evident: organizations whose professed goals are democratic cannot themselves be democratic.

Michels' argument was strongly influenced by his friend and mentor Max Weber's perspective on bureaucracy (Weber 1978). Following Weber, Michels argued that the differentiation of an organizational elite from the rank-and-file was a product of the technical division of labor necessary in complex mass organizations. Weber's pessimistic analysis of the 'iron cage' of modern rationality was reflected in Michels' own pessimism about the dilemma faced by the party rank-and-file: the benefits of bureaucratic organization were manifold, but could only be realized by forsaking popular sovereignty. Weber also argued that the 'routinization of charisma' that led to modern bureaucratic organization was driven by the concerns of the administrative staff to secure and stabilize their position. In Michels' argument, the conservative tendency of organizational elites stems from their concern to maintain their status and livelihood.

As in Weber, Michels' argument about mass bureaucratic organization has two elements which coexist in some degree of tension. On the one hand, the rise of bureaucracy is driven by the demand for services provided by a stunningly efficient organizational machine; on the other hand, bureaucracy is a mechanism of elite domination (Kitschelt 1989, pp. 69–70 distinguishes between a 'weak' theory of oligarchy based on organizational efficiency and a 'strong' theory based on elite domination). Thus, Michels moves back and forth from arguing that oligarchy is explained by the masses' adoration of and gratitude towards their leaders to arguing that it is explained by elites' ability to co-opt dissent through control over patronage and the media. The tension can be highlighted as follows: despite the leaders' subversion of the revolutionary goals and democratic aspirations of the masses, the masses remain grateful and obedient.

To explain this tension, Michels turned to another source of intellectual influence: a conservative view of the masses associated with the Italian elite theorists Vilfredo Pareto and Gaetano Mosca and the French sociologist Gustave Le Bon. Pareto and Mosca drew a sharp distinction between the elites and the masses and argued that the competence and energy of the elites made it possible for them to rule the unenterprising masses. Influenced by Le Bon's study of the irrationality of crowds, Michels argued that the masses are weakly organized and hence politically immature and even incompetent. Unable to organize themselves, the masses have a strong need for leadership. The relatively strong emphasis on psychological factors in *Political Parties*—such as the argument that the masses have 'a need to pay adoring worship' to their leaders—seems to derive from Le Bon (though we cannot forget Weber's analysis of charisma). The psychological distance between elites and masses leads the rank-and-file to behave in a deferential way towards their leaders.

While elite theory is often seen as a rejection of the Marxist analysis of class struggle—and Michels certainly used his analysis to deflate the optimism of Marxism—elements of a Marxian analysis can also be seen in *Political Parties*. At numerous points, Michels suggests that the organizational asymmetry between elites and masses is akin to a class struggle. With their monopoly over the means of production (organization), the elite is essentially able to extract surplus rents from the social movement. He analyzes the organizational differentiation of elites and masses as a process of embourgeoisement. Even (or especially) leaders of proletarian origin become culturally, educationally, and financially differentiated from their original social class and consequently lose all fraternal feeling towards the working class.

2. *Parties, Social Movements, and Organizations*

Michels' articulation of the iron law of oligarchy has had a broad impact on subsequent studies of political parties, social movements, and organizations. Nevertheless, the basic thrust of this subsequent literature has been to qualify and nuance Michels' portrait of the distribution of power in organizations.

One of the first important studies to follow up on

Michels' analysis in the political party literature was McKenzie's (1963) study of the power structure of the British Conservative and Labour Parties. McKenzie found considerable evidence that a relatively small group of leaders in both parties wielded disproportionate power. But he also noted that revolts of the rank-and-file occasionally replaced unpopular leaders. 'The "law of oligarchy," he concluded, 'is certainly not an "iron" law' (McKenzie 1963, p. 644).

A major limitation of Michels' analysis is that he underestimates the factionalism of many political parties. In a historical analysis of the German Social Democratic Party (Michels' major example of the iron law), Schorske (1955) argues that the conservatism Michels attributed to bureaucratization really reflected the dominance of a faction representing a particular constituency. Indeed, as Ansell and Fish argue (1999), it may be factionalism rather than bureaucratization that is the structural basis for the indispensability of certain party leaders. In a comparison of factional politics in the German Social Democratic Party and the British Labour Party, Koelble (1991) argues that Michels also errs in not accounting for the way different internal institutional structures shape the dynamics of intra-party democracy.

Party scholars have also sought to provide alternative descriptions of the nature of power in party organizations. Eldersveld (1964), for example, contrasted oligarchy with 'stratarchy,' an organization where political power was diffused across different levels of the party. In the most general and systematic treatment of party organization in the literature, Panebianco (1988) has described political power within parties in terms of a 'dominant coalition,' a term intended to avoid the unitary image Michels gave to the party elite. Like Eldersveld, Panebianco argues that power is often distributed across different institutional locations inside and outside the party.

In a study of two Belgian ecology parties, Kitschelt and Hellemans (1990) asked party activists how they explained deficiencies in intra-party democracy. They found that party activists primarily supported a view of party oligarchy based on the technical efficiency of leadership and rejected a more conspiratorial view of elite domination. They argue that these parties do exhibit internal stratification, but

> resemble loosely coupled institutions with multiple countervailing power centers rather than tightly organized, centralist bureaucratic hierarchies (Kitschelt and Hellemans 1990, p. 164).

They also discover an interesting paradox: party activists in the party with the most plural power structure were more inclined to criticize elite domination. Kitshelt and Hellemans attribute this to the greater radicalism and participation of these activists, who are consequently more sensitive to elite domination.

Social movement theory has also been attentive to the implications of Michels' iron law. Very much in the spirit of Michels, Piven and Cloward (1977) launched an important debate in the field when they argued that institutionalized protest largely stifles successful mobilization by 'poor people's movements.' Paralleling the trend in political party studies, however, much of the social movement literature has sought to circumscribe the iron law. Zald and Garner (1987), for instance, argued that oligarchy is only one possible outcome that depends on the character of leadership, organizational structure, and membership participation. Kriesi (1996) has argued that only 'instrumental movements' (those not concerned with the collective identity of their constituency) are susceptible to oligarchy.

Finally, Michels has had an important influence in the field of organization theory. Notably, Selznick's (1984) influential critique of the 'myth of grassroots democracy' in the Tennessee Valley Authority was inspired by Michels' analysis. It is important to point out that for Selznick, Michels' critical insight was not so much the phenomenon of elite domination, but the process of 'goal displacement'—the tendency of organizations facing internal and external pressures to forget or compromise their original goals. Ideals may be displaced in the name of organizational imperatives even where elite domination does not hold true.

Otherwise, the trend in studies of organizations (and in organization theory) has also been to emphasize the contingent and plural nature of power in organization. While for the most part accepting the iron law, Lipset et al. (1956) analyzed internal democracy in the International Typographical Union (ITU) as a 'deviant case.' They found the requisites of internal democracy in the active participation of printers in organizational affairs, in the federal rather than unitary structure of organization, and in a history of factional in-fighting that led to an institutionalized two-party system within the ITU. Also stressing the importance of federal organization and factionalism as important conditions for internal democracy, Stepan-Norris and Zeitlin (1996) found internal democracy prevalent in industrial unions where there was a history of insurgency 'from below' against craft unionism.

The thrust of the literature in each of these fields has been to downgrade the 'iron law' to a 'tendency under certain conditions.' Nevertheless, oligarchy remains a 'tendency' worth attending to, and this is Michels' lasting contribution.

See also: Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Leadership: Political; Leadership, Psychology of; Organization: Overview; Organizations: Authority and Power; Organizations, Sociology of; Pareto, Vilfredo (1848–1923); Power: Political; Weber, Max (1864–1920)

Bibliography

Ansell C K, Fish M S 1999 The art of being indispensable: Noncharismatic personalism in contemporary political parties. *Comparative Political Studies* **32**(3): pp. 283–312

Eldersveld S J 1964 *Political Parties: A Behavioral Analysis*. Rand McNally, Chicago

Kitschelt H 1989 *The Logics of Party Formation: Ecological Politics in Belgium and West Germany*. Cornell University Press, Ithaca, NY

Kitschelt H, Hellemans S 1990 *Beyond the European Left: Ideology and Political Action in the Belgian Ecology Parties*. Duke University Press, Durham, NC

Koelble T A 1991 *The Left Unraveled: Social Democracy and the New Left Challenge in Britain and West Germany*. Duke University Press, Durham, NC

Kriesi H 1996 The organizational structure of new social movements in a political context. In: McAdam D, McCarthy J P, Zald M N (eds.) *Comparative Perspectives on Social Movements*. Cambridge University Press, Cambridge, UK, pp. 152–84

Lipset S M, Trow M A, Coleman J S 1956 *Union Democracy: The Internal Politics of the International Typographical Union*. The Free Press, Glencoe, NY

McKenzie R T 1963 *British Political Parties: The Distribution of Power Within the Conservative and Labour Parties*. St. Martin's Press, New York

Michels R 1959 *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracies*. Dover Publications, New York

Panebianco A 1988 *Political Parties: Organization and Power*. Cambridge University Press, Cambridge, UK

Piven F F, Cloward R A 1977 *Poor People's Movements: Why They Succeed, How They Fail*. Pantheon, New York

Schorske C E 1955 *German Social Democracy, 1905–1917: The Development of the Great Schism*. Harvard University Press, Cambridge, MA

Selznick P 1984 *TVA and the Grass Roots*. University of California Press, Berkeley, CA

Stepan-Norris J, Zeitlin M 1996 Insurgency, radicalism, and democracy in America's industrial unions. *Social Forces* **75**: 1–32

Weber M 1978 *Economy and Society*. University of California Press, Berkeley, CA

Zald M N, Garner R A 1987 Social movement organizations: Growth, decay, and change. In: Zald M N, McCarthy J D (eds.) *Social Movements in an Organizational Society*. Transaction Books, New Brunswick, NJ, pp. 121–41

C. K. Ansell

Olson, Mancur (1932–98)

1. Biography

The son of Norwegian immigrants to the US, Mancur Olson was born in Grand Forks, North Dakota, USA, on January 22, 1932, and raised near there on a farm. He died on February 19, 1998 at College Park, Maryland. He studied economics at North Dakota State University, Oxford University, and Harvard University. The greatest fraction of his career he spent as an economics professor at the University of Maryland, but before this he undertook teaching assignments at the US Air Force Academy, Princeton University, and service in the US Department of Health Education and Welfare. Influential in professional associations, he had been president of the US Southern Economic Association, US Eastern Economic Association, Public Choice Society, and Vice-President of the American Economic Association. He was a Fellow of the American Academy of Arts and Sciences, an Honorary Fellow of University College, Oxford, and recipient of many other honors and awards. At the time of his death he was principal investigator and director of a US Government sponsored development aid/research project with offices in 30 countries, and an ambitious agenda to bring the benefits of his life's work to the design and reform of the institutions of governance throughout the developing and former communist countries of the world.

2. Antecedents to Public/Collective Choice

One of the founders of 'Public or Collective Choice' as a subdiscipline and movement within economics, Olson devoted his intellectual life to the analysis of collective action. Early, as a student, he intuited that the then standard paradigm as to how people interacted economically with each other was seriously incomplete. In the traditional model it was argued explicitly or implicitly that economic interaction while widespread and diverse, was normally mediated by prices and markets between otherwise isolated and independent agents. The first founding principle of 'Public Choice' and a concern which permeates Olson's entire life work is recognition that people exchange, produce, consume, and share many goods, in many environments which are not at all mediated by markets and prices, yet the behavior and analysis thereof is unquestionably 'economic.' Typically, in these environments people belong to identifiable groups and their interaction partakes of their group identity in a manner not conveyed by Adam Smith's butcher or baker. This raises to paramount importance questions of how people and organizations who belong to a group do and should act to achieve their goals.

2.1 Oligopoly

Central to the Public Choice nexus is how the size of a group influences the behavior of its individual members and thus aggregate outcome. Other features of group membership such as homogeneity or diversity are also important but size was the first aspect to which Olson attended. With respect to industrial behavior, the question of how small numbers of firms would

behave had occupied economists' attention for a generation by the time Olson became a student. Still earlier the pricing, output, employment, and other profit maximizing behaviors of a single, monopolistic, firm had been analyzed successfully as had those of a market consisting of a large number of atomistic competitive firms (Marshall 1890). And by the time Olson arrived at Harvard in 1957, for the case of small numbers of firms in an industry, theories of oligopoly interaction (Cournot 1897, Robinson 1933, Chamberlin 1933) had emerged which demonstrated that in pursuing rational profit maximization for themselves individual suppliers would fail to achieve maximum profits for their industry—their group—as a whole. The reason for this is a lack of coordination: oligopolists fail to anticipate and include in their decision calculus the effects their narrow profit making decisions have on the profits of other firms. In fact it would be individually irrational for any single firm to include such external benefits in its decision calculus since the individual firm could gain more by free riding off the efforts of other firms. A similar coordination failure motivated the fast developing field of Game Theory that provided a rigorous framework for analysis of oligopolistic behavior. Actually Olson's mentor at Harvard (E. Chamberlin) had been one of the leaders in the development of oligopoly in the form of 'monopolistic competition.' Here individual firms in an industry resemble monopolists in that they recognize their supply decisions will influence the price they can obtain. At the same time their industry is competitive because the price actually charged by any one firm influences the price-supply opportunities of every other firm.

2.2 *Public Goods*

Parallel with economists' study of producer oligopoly, discoveries preceding Olson showed that provision of some types of goods to consumers could lead to similar failures of coordination (Samuelson 1954). 'Public Goods' or 'Collective Goods' are such that once available to any member of a specified group, the same quantity and quality is equally available to every other member at no extra cost: this is because such goods are not used up when enjoyed and if available to any consumer cannot be excluded from any other group member. Early examples included the lighthouse beacon which warns all ships (nonexcludable), and whose utilization by one ship in no way reduces its availability to other ships (nonrival). More contemporary examples include protection of pollution free environments, provision of national defense, preservation of natural species, establishment of political freedoms, provision of property right guarantees by an impartial and universal police/judiciary, and maintenance of government by rule law. Broadly speaking these all should be understood to be economic 'goods,' and all display the property that their enjoyment by one person in no way diminishes their availability to others to enjoy. But if every member of a group enjoys a good irrespective of who supplies it, no one acting voluntarily will provide the good—at least no one will provide a sufficient amount of it. Instead all will have an interest in seeing others pay the supply costs while they enjoy the benefit so produced without making any cost contribution themselves. Every individual, that is, has good reason to free ride—in whole or in part—on the provisions of others. One, therefore, should anticipate that for the provision of such public goods, free markets based on individual rationality will fail. This means that the amount or quality of collective good provision to follow from individual rational behavior will fall short of what the aggregate of individuals in a group desire. In short, notwithstanding the fact that a group is nothing more than the collection of individuals who belong to it, individual rationality, even assuming it to be universal, will not secure the rational group outcome.

3. *The Voluntary Provision Model and Olson's Logic of Collective Action*

The initial reaction of economics to this understanding was to conclude that instead of individual action, society must rely on government power to enforce individual contribution via coercive taxation/expenditure. Olson's first insights came from taking the next step: because government itself is subject to the quandary which the free rider problem creates, and may not produce an optimum outcome, it cannot be assumed that government will ordinarily close the gap between individual and group rationality. Olson, therefore, asked what to expect of individual voluntary behavior in the absence of governmental or coercive intervention. Such is the topic of *The Logic of Collective Action* (1965), Olson's Ph.D. thesis and first great contribution. Once the question was thus posed, it became clear that the range of circumstances where collective action obtained was far more vast, complex, and rich than the early oligopoly paradigm might lead one to expect. Moreover, contrary to simplistic theory sometimes arrangements arise (in the form of intragroup monitoring or selective extra rewards to those who do not free ride) which allow the free rider failure to be reduced and possibly altogether circumvented. Olson's discovery and elaboration of the 'voluntary provision' model of public goods changed, in fact redefined, the topic for economists; it introduced public goods to other social sciences, especially political science and sociology and elevated it to a unifying principle throughout the social sciences. The result was an explosion of research into the structure of voluntary provision, and into the myriad of applications of the model, ranging from strictly economic to anthropological, legal, sociological, and pol-

itical, from labor union behavior, to allocations within families, voluntary philanthropic giving for a group or social goal, and from public interest lobbying, special private interest favor seeking, and professional associations to the behavior of nations as they formed alliances.

In all of these instances Olson discerned three factors paramount in explaining the degree of success a group has in providing itself with a collective good without government coercion: these were (a) the size of the group, (b) the heterogeneity of its members, (c) and the ability of the group to offer selective incentives to individual members to induce contribution to the collective good. In fact the international alliance example served as the first instance of voluntary provision which could be verified empirically. Olson and Zeckhauser (1966) analyzed successfully the allocative behavior of the NATO Alliance as it provided its constituent members deterrence against attack and defense should attack materialize. The Olson-Zeckhauser analysis confirmed empirically prediction from Olson's theory, namely that when the membership of a group is heterogeneous, the large and/or rich members will voluntarily contribute disproportionately large shares of the total public good—an effect Olson's had named 'exploitation of the great by the small.' Thus crucially, Olson and Zeckhauser demonstrated that Olson's insights into the applicability of economic structure to political phenomena could be formulated into rigorous logic and empirically tested against reality. Later extensions of the voluntary provision model have lead to surprising insights (Bergstrom et al. 1984) into equilibrium outcomes when individuals in a group interact but do not coordinate these interactions. These include demonstrations that, up to a point, in equilibrium a group's total voluntary provision of a public good is insensitive to the distribution of wealth among its members, and beyond this point actually increases with greater inequality in the distribution of wealth. And Olson's voluntary provision model continues to be the paradigm for numerous behavioral patterns, from cooperation among countries in regulating the international economy to controlling pollution, to the economics of charitable giving.

4. *Public Choice, Political Economy, and Olson's Rise and Decline of Nations*

From the first principle of 'Public Choice' mentioned above, it follows that the extramarket behavior of the individuals who make up governments is itself governed by individual search for individual satisfaction. Always committed to individual rationality embodied in calculated decision, in the middle stages of his career Olson's elaboration of the public choice theme develops several corollaries to this proposition.

First, rather than being a utopian instrument to correct social inefficiencies in the provision of public goods stemming from free rider behavior, government behavior itself is riven by free-rider failures. The problem of correcting for free-rider deficits is itself in other words beset by free-rider effects, for example, who will contribute the resources, attention, and energy necessary to form environmental policy to protect the public from degradation of the atmosphere? The answer is: precisely not the general public which benefits from environmental control but instead those interest groups sufficiently well organized to exact or induce participation by their membership. This leads to a second corollary.

The concept of collective goods and free riders applied to the larger social mosaic implies a hierarchy or cascade of nested interest groups. First of all people will act in their own interests, not the interests of their group; but once they have overcome intragroup free riding successfully and harnessed their individual incentives to the group, if they ever do, then those groups will act in their own self interests, not in the interests of their larger society.

A third corollary applies these principles to redistribution, with introduction of the notion of a distributional coalition, a group which by collective or collusive action can gain an increased share of society's wealth not by working to create a larger societal total, but by imposing external costs on others in the society. Such groups work though various instruments, many effected and enforced by the government which they influence; these include discriminatory taxes or tariffs, barriers to mobility, or to entry as well as downright corruption and theft.

It was Olson's focus on this structure which brought the free-rider problem to the widest academic audience of sociologists, historians, and political scientists, and which showed the applicability of the idea to current world developments and practical policy. His formulation of the idea, propounded in *The Rise and Decline of Nations* (1982), assembles these corollaries into one grand principle: the notion that entire societies fail or succeed on the quality of their management of free riding. It is the success of interest groups in overcoming obstacles to their own collective action and the failure of any larger polity to overcome the obstacles to its collective need to leash protectionists that molds the destiny of nations. Moreover, there is an inherent tendency over time, Olson argued, for efficiency destroying and innovation strangling special interest groups to proliferate in stable societies, such as the US and UK with institutions which survived World War II more or less in tact. By contrast, those societies which experience the social upheavals typical of war or revolution—like Japan or Germany at the end of World War II—find that special interest groups and distributional coalitions are dismantled, thereby giving the society as a whole another start at constructing social and economic

institutions afresh which will encompass the aggregate collective interest of all the people. The idea of social and institutional sclerosis, sometimes called 'demosclerosis,' was thus advanced as a crucial determinant of societal economic success. An audacious idea, but one subject to empirical verification. The idea was noticed by social scientists the world around and resonated in the popular press.

5. *The Character of Governance: Olson's Power and Prosperity*

The final focus of Olson's energy—summarizing papers from the last decade of his life in the posthumous *Power and Prosperity* (2000)—was on the greatest of the free rider puzzles of all: how does government even come into existence, given the immense obstacles to collective rationality presented by individually rational free riding? Reflecting on this question, he was well aware of the immense rewards a society reaps from establishing a peaceful social order, and submitting itself to the institutions of governance. So there is no doubt that good governance is in the collective interest of a society; by investing in certain public goods governments can increase the productivity of the society dramatically, and this benefits everyone. Moreover, for groups larger than families or at most tribes, governments are the only institutions that can do this. These are public goods, such as establishment of a rule of law, institution and maintenance of property rights, provision of common defense, stable currency, contract enforcement, elements of infrastructure, and public education and health, and so on. Such investments in public goods benefit every one in the society by making its whole industrial, commercial, economic base much more productive.

Olson was also deeply aware that provision of public goods is far from the only or even major function of government. A second function is redistribution. Historically, almost all systems of government have emerged as vehicles of redistribution from the people to an autocracy, monarchy, or oligarchy. Contemporaneously, our generation has seen massive increases in government enforced redistributions within countries throughout the world. These redistributions, which play so crucial a role in the political economy of the modern state, are undertaken by governments which exact taxes from some groups and make transfers for the benefit of other groups. The beneficiaries in this process are in general the groups to whom governments are beholden, and the taxpayers those to whom they are not. Thus, (a) redistribution to favored interests and (b) allocation to enhance everyone's productivity are the two primary economic functions of government. Public good allocation, and group discriminatory redistribution are conflicting resource use options between which any government must choose. Irrespective of its political make up this is a choice every government must make and a supremely important one for the welfare of its members.

But how is the conflict between the two objectives reconciled? In brief, Olson argues that the redistribution that a society accepts is the price it must pay to achieve the gigantic collective benefits of peaceful governance. But a precondition for this bargain even to be struck in the first place is that free rider obstacles be overcome sufficiently for government to be enduring (Olson 1991, 1993). Governance by transient 'roving bandits' is inefficient; only when the roving bandit settles down to become a 'stationary bandit,' that is, an established monarch, can society enjoy the immense benefits of peaceful development. Assuming this much success in conquering free riding pitfalls, Olson then extrapolated a theme from earlier in his career: the idea of groups as differentiating themselves along a scale of their 'encompassing interests.' This is the notion that a group that reflects the interests of a large proportion of its society in its economic calculus will behave very differently from one with a narrow membership. The more encompassing group will internalize more of the cost it inflicts on society in pursuit of its distributional goals. And this provides the clue to his discovery (Olson and McGuire 1996) that the political logic of these invest–redistribute decisions will lie on a spectrum between two polar extremes. At one extreme is an idealized, consensual democracy where all redistribution issues have been settled unanimously and taxation only serves public good investment objectives. At the other extreme is perfect autocracy ruled by a dictator who taxes and spends solely to satisfy his own selfish desires. This work then shows that realistic societies can be analyzed as a mixture, in fact a weighted average, of these two polar cases. Thus, in making the choice between investment in the entire society and redistributive taxation from the powerless to the powerful, every polity behaves somewhat like a pure democracy and somewhat like a selfish dictatorship. Informed by this, logic McGuire and Olson derive the allocation and distribution rules which various governments across the spectrum will follow, showing, for example, that even a purely selfish dictator (providing he controls an enduring government) will not only transfer resources to himself; he will also invest in his economy. Why? Because investment will make the economy prosper and add even further to the tax take. Also intriguing is the proof that all societies and forms of rule are limited by an automatic governor on their incentives to tax. This contradicts the pessimistic expectations of some political moralists that extension of the democratic franchise would necessarily induce further voracious redistributive taxation, ending only in the ultimate destruction of the economy itself. It demonstrates, to the contrary, incentives to tax for every ruling group

reach a point well before confiscation where redistribution voluntarily ceases.

The thrust of Olson's thought toward the end of his life was to bring these insights to bear on the puzzling economic trajectory of the Soviet Union which seemed so economically cumbersome succeeded by Russia whose fortunes so declined after its markets were unleashed.

6. *Olson's Impact*

Olson's legacy extends broadly across the overlapping frontiers of social science and philosophy. His life was marked by numerous conferences devoted to examination of his ideas and books devoted to his insights such as Sandler (1992). His impact derives first of all from his ability to develop a simple idea so that it becomes powerfully applied across a grand subject area. An enthusiast for basic ideas, and ever the economist, he illuminated political processes using economics ideas and showed how political processes could hobble economic productivity. The result was to gain far wider acceptance of economic method and theory in the work of political scientists. Many before Olson could take for granted that if some outcome would be beneficial to all members of a group, then the group would act in concert to bring it about, thinking in terms of joint rather than individual action, without asking how the joint action would be enforced. After Olson this happened no longer. He had the rare ability to take highly abstract but ultimately simple ideas and to apply them to the reality of economic and political development. Because of his work awareness of the difficulties and paradoxes of achieving collective rationality, no longer confined to a cadre of specialists, has entered into the discourse of the largest universe of social science and philosophy.

Concrete witness to his impact is the IRIS (Institutional Reform and the Informal Sector) project. This US Agency for International Development vehicle was created to promulgate of Olson's ideas in more than 30 countries around the globe, many struggling with the reform and adjustment process in a transition from socialism. Although he explored and exposed the difficulties of achieving socially desirable outcomes through uncoordinated individual action, he was no naive advocate of central planning or coercion. For him the operations of governments were to be subjected to the same analytic scrutiny as markets, and government failures had to be exposed just as fearlessly as market failures. Thus, it was within the practical context of practical reform of institutions that Olson's final contributions took effect as an advocate for the idea that good governance can manage free-rider challenges and achieve prosperity in the developing world (Olson 1996). His life was marked by innovative ideas, personal energy, and an unflinching conviction that institutional reform could yield prosperity over poverty. The shift in appreciation of the links between good governance and economic success which this caused within social science is his living bequest.

Bibliography

Bergstrom T C, Blume L, Varian H R 1986 On the private provision of public goods. *Journal of Public Economics* **29**: 25–49

Chamberlin E H 1933 *The Theory of Monopolistic Competition*. Harvard University Press, Cambridge, MA

Cournot A 1897 *Researches into the Mathematical Principles of the Theory of Wealth*. Macmillan, London

Marshall A 1890 *Principles of Economics*. Macmillan, London

Olson M L Jr 1965 *The Logic of Collective Action: Public Goods and The Theory of Groups*. Harvard University Press, Cambridge, MA

Olson M L Jr 1982 *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. Yale University Press, New Haven, CT

Olson M L Jr 1991 Autocracy, democracy, and prosperity. In: Zeckhauser R (ed.) *Strategy and Choice*. MIT, Cambridge, MA

Olson M L Jr 1993 Dictatorship, democracy, and development. *American Political Science Review* **87**(3): 567–76

Olson M L Jr 1996 Big bills left on the sidewalk: Why some nations are rich, and others poor? *Journal of Economic Perspectives* **Spring**: 3–24

Olson M L Jr 2000 *Power and Prosperity: Outgrowing Communist and Capitalist Dictatorships*. Basic Books, New York

Olson M L Jr, McGuire M C 1996 The economics of autocracy and majority rule: The invisible hand and the use of force. *Journal of Economic Literature* **XXXIV** (March): 72–96

Olson M L Jr, Zeckhauser R 1966 An economic theory of alliances. *Review of Economics and Statistics* **XLVII**: 266–79

Robinson J 1933 *The Economics of Imperfect Competition*. Macmillan, London

Samuelson P A 1954 The pure theory of public expenditure. *Review of Economics and Statistics* **36**: 387–9

Sandler T 1992 *Collective Action: Theory and Applications*. University of Michigan Press, Ann Arbor, MI

M. C. McGuire

Onomastics

'Onomastics' (from Greek *ónoma*, 'name') is the intellectual endeavor which studies names of all kinds (geographical names, personal names, boat names, names in literature, etc.). As names are embedded in, and transmitted through, language their study is often regarded as a subdiscipline of linguistics, and more precisely of lexicology, the study of words. This limited approach, however, does not do them justice, for names are not simply words with additional properties, and therefore they transcend purely lexical considerations. In this broader sense, they are also the

legitimate subject of investigation by scholars working in extralinguistic disciplines, such as history, geography, sociology, psychology, philosophy, law, religion, literary criticism, and so on. It is in this wider ambience that onomastics will be treated here. The most extensive and authoritative account of names and their study is to be found in *Name Studies: An International Handbook of Onomastics* (also containing articles in French and German), Eichler et al. (1995–6).

1. Names and Words

The study of names, whether pursued as an integral part of, parallel to, or complementary to the study of words, has a long and distinguished history, although its terminology and the concepts it represents have led to a certain amount of confusion, a state of affairs which has not been resolved adequately even in our time. It is therefore necessary to pay close attention to the criteria which are being applied in each approach. Furthermore, in English the situation is complicated by the fact that, in addition to its indigenous term *name* (Old English *nama*, *noma*), inherited ultimately from Indo-European and therefore shared with the other Indo-European languages, it has, among others, also adopted the term *noun*, which goes back to Latin *nomen*, via Old French *nun*, *num*, and is now usually applied to substantives. This contrast between *names* and *nouns* is, however, not carried through systematically, and one still comes across such largely synonymous pairs as *common noun* and *proper noun* on the one hand, and *common name* and *proper name*, on the other, in which *common* and *proper* reflect classical tradition. Though the term *name* and the concept behind it are consequently at least pan-Indo-European, it is advisable to treat its descendants with care within each individual language. Such caution is, naturally, also called for in connection with name studies in non-Indo-European languages. Otherwise, any attempt at a theory of names or, more correctly, a theory about names, is destined to founder from the outset. (For further reading see Algeo 1973, Gardiner 1954, Pulgram 1954.)

1.1 Semantic Distinctions

Though names and words share many characteristics, they differ in one major respect which has fundamental consequences: their semantic properties. If words are to be used competently, they must have meaning on the lexical level; otherwise they fail to serve the purpose of linguistic communication. Such a condition is essential to words of all kinds, whether they apply to concrete (*table*, *stone*, *street*) or abstract (*kindness*, *politics*, *religion*) phenomena. Meaningless words cannot fulfill their connotative function which is inclusive in its potential reference to any kind of 'table' or any variety of 'kindness.' This basic prerequisite is not expected of names which function perfectly well without any lexical meaning in what has been called 'the speech act of identifying reference' (Searle 1969, p. 174) which is both denotative and exclusive (*Aachen*, *Glasgow*, *Paris*; *Karin*, *Maggie*, *Robert*).

Although most names started out as words or combinations of words that, in their time, had been transparent semantically on the lexical level, it is their capacity to function while being opaque semantically that makes them such fascinating and useful evidence for scholarly investigation, i.e., for what has come to be known internationally as onomastics. What they do need instead of lexical meaning, however, is onymic contents—this city (*Toronto*), this hill (*Bennachie*), this person (*David Livingstone*), this ship (*Titanic*), this dog (*Rover*), etc. In this context, it is essential for effective communication that there is a minimum of shared knowledge about that content, in order to avoid misunderstanding or complete failure. We need to know names, not understand them. The investigation of brand names, which has become very popular in recent years, can be said to straddle onymic and extra-onymic fields of research (Eichler et al. 1995–6, Sect. XIX).

1.2 Name Clusters and Fields

Despite the basic role of names as identifiers of individual referents, to the exclusion of all others, the number of names which can justifiably be claimed to refer to just one place, person, boat, cat, etc. is very small in our day and age. For that reason, true individuation can only be expected in limited environments (one first name in one generation of a nuclear family, one street name in a small town, one fishing boat in one home port, one horse in one stable). It is, however, a truism that no name can function alone but can only be useful in relation to other names, either in loose name clusters or, particularly, in structured onomastic 'fields' (a metaphor borrowed from physics), such as all names in a family; all, or most, street names in a town; names of all fishing boats in a harbor (and perhaps even beyond); all names of horses in a stable. Research into 'onomastic fields' parallels the study of 'lexical fields' (Schmidt 1973) but, because of their respective differences in semantic status, is subject to different criteria (Nicolaisen 1988).

1.3 Lexicon and Onomasticon

It is customary to think of the inventory of words in any language as a lexicon of which dictionaries are the best-known repositories. The equivalent corpus of names may be regarded as an *onomasticon* which, like a lexicon, may range from all the names embedded in

any culture or language to the personal inventory of names known to one individual. Their conventional home is the gazetteer. It is when lexical items cross the threshold from lexicon to onomasticon as part of the naming process that their semantic status changes drastically. Even if, as often happens at least in the initial stages, onymic items are still lexically transparent (*Neustadt*, *Westhill*, *Rose*, *Smith*), the recognition of that word meaning hardly ever interferes with the (newly) acquired onomastic contents, even if these two aspects apparently contradict each other, as in the case of a *Neustadt* ceasing to be a *neue Stadt* 'new city', or of a man named *Smith* becoming a baker. In fact, a *baker* called *Baker* is likely to be an amusing anomaly rather than the prevailing norm in cultures with heritable surnames. In this sense, the long-running controversy of whether a name is in essence more of a name when it is (lexically) meaningless than if it has preserved semantic transparency on the lexical level can be demonstrated to be based on false alternatives. This judgment is supported by the realization that not every name starts out as a word or word compound with lexical meaning, because any pronounceable sequence of sounds/letters can become a name, and in some instances there are intra-onomastic transfers in a personal name becoming a place name (*Jefferson*), a place name becoming a personal name (*Pittendriech*), or a place name or personal name becoming the name of a ship (*Deutschland*, *Queen Mary*). Sometimes this process can be repeated, as when the English place name *Washington* first turns into a surname and then into a place name again (*Washington*, DC *Washington* State).

2. *Onomastic Methodology*

It is clear that in such transfers from one onymic category to another, or in names with non-lexical origins, a search for the original word meaning would not only be fruitless but also inappropriate. Since they are, however, in the minority, and since much onomastic research has its origins in the investigation of linguistic history, it is not surprising that etymological approaches have been, and still are, very common, especially in the study of place names (*toponymy*). This is, in principle, not very different from the lexicographical search for the derivation of words insofar as its chief requirement is the utilization of the earliest recorded spellings or forms, but nevertheless differs from lexical etymology because its aim is to reduce the name to the word(s) it once was in the act of naming: nevertheless, etymological research must be regarded as a fundamental first step in most kinds of onomastic procedures, as long as its successful completion is not also taken to be their sole goal (see *Etymology*). The many volumes of the county survey by the English Place-Name Society, for example, speak of the necessity for, and efficacy of, this method (EPNS 1924–).

2.1 *Toponymic Research*

The study of place names (Eichler et al. 1995–6, Sects. XV–XVII) has for a considerable time been one of the foremost occupations of the scholar of names. The main reason for this has been the very fact that names can function without the aid of lexical meaning, whereas words cannot. This characteristic has allowed names to survive not only when the words which have formed them have died out at an earlier stage of the language that coined them, but also to outlive the very death of that language itself and its replacement by another language. In Scotland, for instance, only a handful of words has been adopted by the English language from Gaelic, whereas the maps contain many thousands of place names of Gaelic origin; the same can be said of the Slavic ingredients in the vocabulary and toponymy, respectively, of the eastern parts of Germany. In some parts of Europe and in European-derived cultures, this process of toponymic survival at times of language replacement has sometimes been repeated more than once. In the north-east of Scotland, for instance, Pictish place names have survived from preGaelic times, and it is also possible to detect a preCeltic, i.e., preGaelic and prePictish layer of Indo-European river names ascribable to the same 'Old European' linguistic stratum which, on the European continent, is linked to the Bronze Age. Some of the major island names in the Northern and Western Isles reflect the sometime, presence of speakers of an as yet unidentified non-Indo-European language. In southern Europe, multiple layering in linguistic stratification is also demonstrated by toponymic evidence.

The value of place names as evidence for this kind of investigation increases with their relative antiquity. Research among living languages has shown that in periods of linguistic and cultural contact the main process by which place names pass sequentially from one language to another is chiefly by phonological adaptation. However, partial translation, full translation, and complete replacement also occur, albeit more rarely; it is also natural that many names do not survive language change. What has emerged from this type of study is the realization that lexical dialects and onomastic dialects are not always congruent (Nicolaisen 1980). It has also become clear that the translations of maps showing the spatial distribution of place names into maps demonstrating chronological sequences is fraught with difficulties.

2.1.1 Name transfer. The kind of toponymic palimpsest which these processes create provides a key to a relative chronology in the settlement history of speakers of relevant languages. More often than not, individual strata, or the stratification as a whole, reflects the importation of languages from elsewhere. The settlement of the New World by Europeans since the sixteenth century is a particularly striking

example, since many of the names designating new locations were imported unanalyzed from elsewhere, usually without regard to their semantic fitness (Zelinsky 1967). It can also be shown that 'colonial' settlers take with them in their mental luggage not just a lexicon of words but also an onomasticon of name models which are applied to the new terrain in the creation of landscapes reminiscent of home (Nicolaisen 1987). One instructive instance of this is the settlement of the Northern and Western Isles of Scotland from Norway, from the ninth century onward, for without names there are no landscapes and no habitats, but only wildernesses and disorientation.

3. *Socio-onomastics*

Initially, name scholars concentrated largely on the naming process itself and on its product, the name, but in recent decades the focus of their research has shifted more to socio-onomastic concerns, and especially to the role of names in society (Eichler et al. 1995–6, Sect. XXI). Usage and process have therefore become more central to their discussions. There are various aspects to this new direction. A major facet of place names is, in this respect, their role in politics, especially in political change, of which the emergence of a 'new Africa' is an instructive example, having prompted the creation of a large number of country names in the postcolonial era (*French Soudan* becoming *Mali*, *Gold Coast* becoming *Ghana*, *Southern Rhodesia* becoming *Zimbabwe*). Of special interest is the Republic of South Africa, in which radical internal changes have necessitated the wholesale reconstruction of its place-nomenclature, a task in which scholars with extensive expertise in onomastics have become involved (Raper 1998). More generally, in the 'global village,' the United Nations has created the 'United Nations Group of Experts on Geographical Names' (UNGEGN) to oversee particularly the standardization of names in official usage, a move which has, as a byproduct, created a greater awareness of exonyms (*Copenhagen* for *København*, *Cologne* for *Köln*, *Aix-la-Chapelle* for *Aachen*) and their legitimate but circumscribed usage.

A long-standing experience of professional involvement by name scholars in the shaping forces toward a politically, as well as linguistically acceptable toponymy is to be found in Québec and its aspirations regarding independence (Dorion 1998); and the sociopolitical implications of the use of place names by speakers of minority languages in countries such as Italy, France, Switzerland, Belgium, the Netherlands, Spain, Scotland, etc. have attracted the attention of name scholars (Kühebacher 1986), as have the alternative usage of place names in informal, unofficial onomastic registers and the minutiae of microtoponymic concerns.

3.1 *Personal Names*

It goes without saying that the study of personal names (*anthroponymy*) is central to socio-onomastic investigations, as names are an essential component of human beings as unique individuals, and of their role in society. Just as there are no landscapes without place names, there are no societies without personal names. Parallel to etymological investigations regarding other categories of names, a certain amount of basic research has gone into establishing the linguistic origins and meanings of personal names, but the main thrust has been toward a systematic investigation of personal names as markers of human identity, traditional family pressures, or fashionable models, in various contexts. These include the naming of children; the historical development of heritable surnames from their four main sources—patronymics, occupational names, nicknames, and place names—anthroponymic typologies; the morphology of personal names; the increasing use of middle names; the overt and covert use of bynames; continuity and change in favorite names, etc. (see Eichler et al. 1995–6, Sects. XI–XIII).

3.2 *Names and Education*

There has been a growing tendency to introduce the discussion of many aspects of personal names into the classroom, for didactic purposes (Franz and Greule 1999). Their own names and other people's names are of interest to children from an early age, and psychologists as well as educationalists have studied their increasing awareness of names in the development of selfhood and in the process of socialization. Personal names also form an immediate basis for children's involvement in, and contribution to, such discussions, just as investigations into local place names provide convenient inroads into local history and spatial awareness.

4. *Literary Onomastics*

The study of names in literature (Eichler et al. 1995–6, pp. 533–6) attracted an increasing number of scholars in the last few decades of the twentieth century. Its double roots in onomastics and literary criticism have made it a truly interdisciplinary pursuit. Much of the research conducted in its early phases was directed toward the role of names in individual works by individual authors. This search for name inventories as texts within texts gave, and to a certain extent still does give, literary onomastics a strong bias toward intratextual questions. However, recently there has been a conscious movement toward intertextual approaches and names as devices in intertextual relationships, or as means by which the factual world

can become fictionalized, and fictive persons and places can become factualized. A special branch of this approach is concerned with the translatability of names, on both the lexical and the onomastic levels, whereas other scholars are concerned with problems of name stylistics. The role of names as devices employed in autobiographical writings in attempts to create a credible past has also attracted special attention.

5. Conclusion

Having started out as a subdiscipline of linguistics and functioned for a long time as a pursuit which was, because of the special status of names, admirably suited to throw light on puzzling concerns of other intellectual endeavors. But gradually onomastics has abandoned this handmaiden role and has developed into an independent, though amply connected, discipline which studies names as names, for their own sake, thus establishing a foundation for their effective investigations as interdisciplinary evidence. The introduction of socio-onomastic approaches has been especially productive in this respect, but onomastics has also widened its scope and diversified its methodology in other respects, in a concerted drive toward greater sophistication and a more systematic and systemic outlook. The growing scholarly interest in matters onomastic is reflected in the existence of several academic institutes devoted to name studies and usually attached to universities or academies (in Denmark, Norway, Sweden, Finland, Belgium, Scotland, England, Northern Ireland, etc.), as well as many international, national, and regional organizations in the field, and regular congresses and conferences, often dedicated to special facets of onomastics.

See also: Etymology; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Lexicology and Lexicography; Word Meaning: Psychological Aspects

Bibliography

Algeo J 1973 *On Defining the Proper Name*. University of Florida Humanities Monographs No. 41. University of Florida Press, Gainesville, FL

Dorion H 1998 La toponymie, complice involontaire de la politique? In: Nicolaisen W F H (ed.) 1998 *Proceedings of the XIXth International Congress of Onomastic Sciences, Aberdeen, August 4–11, 1996*. 3 Vols. Department of English, University of Aberdeen, UK, Vol. I, pp. 5–19

Eichler E, Hilty G, Löffler H, Steger H, Zgusta L 1995–6 *Namenforschung/Name Studies/Les noms propres: Ein internationales Handbuch zur Onomastik/An International Handbook of Onomastics/Manuel international d'onomastique* (in three parts). de Gruyter, Berlin

EPNS 1924– *The Survey of English Place-Names*. English Place-Name Society, Nottingham, UK (75 volumes published so far)

Franz K, Greule A (eds.) 1999 *Namenforschung und Didaktik*. Schneider, Hohengehren, Germany

Gardiner, Sir A H 1954 *The Theory of Proper Names*, 2nd edn. Oxford University Press, London

Kühebacher E 1986 *Amtlicher Gebrauch des geographischen Namengutes*. Südtiroler Kulturinstitut, Bozen, Germany

Nicolaisen W F H 1980 Onomastic dialects. *American Speech* **55**: 36–45

Nicolaisen W F H 1982 Lexical and onomastic fields. In: Kazimierz R (ed.) *Proceedings of the Thirteenth International Congress of Onomastic Sciences, Cracow, August 21–25, 1978*. II. Pa'nstwowe Wydawn, Warsaw, pp. 209–16

Nicolaisen W F H 1987 Imitation and innovation in the Scandinavian place-names of the Northern Isles of Scotland. *Nomina* **11**: 75–85

Nicolaisen W F H (ed.) 1998 *Proceedings of the XIXth International Congress of Onomastic Sciences, Aberdeen, August 4–11, 1996*. 3 vols. Department of English, University of Aberdeen, Aberdeen, UK

Pulgram E 1954 *Theory of Names*. American Name Society, Berkeley, CA

Raper P 1998 Standardizing geographical names in a changed society. In: Nicolaisen W F H (ed.) *Proceedings of the XIXth International Congress of Onomastic Sciences, Aberdeen, August 4–11, 1996*. 3 Vols. Department of English, University of Aberdeen, UK. Vol. I, pp. 264–74

Schmidt L 1973 *Wortfeldforschung*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany

Searle J R 1969 *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK

Zelinsky W 1967 Classical town names in the United States: The historical geography of an American idea. *The Geographical Review* **57**: 463–95

W. F. H. Nicolaisen

Open Classroom: Learning and Teaching

1. Introduction

The open classroom is a highly structured and deliberately organized learning environment in which the teacher performs a significant and crucial role. This kind of schooling is known by several of its manifestations: open school, integrated day, open corridor, informal education and British infant schooling, terms used to describe a learning approach which discards the traditional classroom setting in favor of a less formal, highly individualized atmosphere in which direct experiences are central to the learning process. The adjective 'open' recurs because this form of education is not circumscribed in classroom space, time blocks, age grouping, or even standardized curriculums. Instead, each child proceeds at his own pace and according to his own interests within a rich and carefully patterned climate in which active in-

vestigation and enjoyment of varied materials lead to the learning of skills and concepts in an integrated, rather than an isolated, manner. This departure from traditional teaching methods affects the role of the teacher as it does the behavior of the students.

Although an agglomeration of concepts from a variety of sources have contributed to open education theory, including Rousseau and Froebel, the theoretical basis of the open classroom is found in the work of Piaget who was frequently identified in the Plowden Report as being central to an understanding of the process of learning in the open classroom. This document recognized and authenticated the informal styles of teaching and learning that were developed in English primary schools during the previous four decades and investigated by Gardner at the University of London (Horwitz 1979).

As a result of the impact of the Plowden Report upon American teacher advocates, 50 percent of all schools built in the United States during the late 1960s were of open design and a similar percentage of open schools were built in Canada, New Zealand, Australia, and the United Kingdom (Bennett et al. 1980). In addition, open classrooms have appeared in Japan, Israel, Thailand, and Norway (Bennett 1981).

2. *Learning in the Open Classroom*

2.1 *The Learning Environment*

The underlying assumption of open classrooms is that in an enriched and carefully planned environment, that supports the natural drive toward learning, children learn in encounters with the materials and people around them at their own irregular and individual pace. Children learn most intensely when they are interested and understand the relevance of what they are doing. The open classroom represents a way of thinking, a way of living with children in a learning environment that meets the individual needs of the students and therefore, its curriculum is individualized.

Four operating principles govern the learning environment. First, an open flexible space divided into functional areas rather than one fixed homogenous unit. Second, the environment is rich in learning resources including concrete materials, books, and other media. Third, the students are engaged in a variety of activities and levels of operation utilizing the contract-commitment method in which each student 'contracts' to do a certain amount of work, or to complete a project within a specified time limit. In this dynamic atmosphere, where learning cuts across subject lines, students are free to talk to each other, move about the room, and they have a significant degree of responsibility for their own progress. Fourth, the teacher divides the time into small, concentrated periods, working with individual children or small groups, rather than addressing the whole class as one group all day.

Open classroom teachers are firm in their assertion that the teacher's primary responsibility is to create an environment that stimulates learning and is altered and expanded as the needs of each child change and his or her interests become more complex. The child's response to the environment becomes the starting point for learning because children learn from the things they do and understanding comes through personal experiences. In this environment, the teacher develops a uniquely close relationship with each child, as learning is dependent upon the sensitive responding of the teacher to the needs and interests of the individual student. This cultivated skill demands that the teacher become immersed in each child's activity, identifies the moment of readiness, and gently guides the child toward direct exploration and experimentation.

2.2 *Learning Events*

In the diagnosis of learning events, the role of the teacher is to understand the child's thinking process. By listening and questioning, the teacher draws the child out so that the child can be assisted in realizing their feelings, plans, and details. Teacher interviews illustrate the delicacy used to determine when to intervene, to push, or even to assist. 'She stretches her empathy for potential, she thinks with the child as he discusses. The 'tuning in', *the thinking* '*with*', the seeing with a 'child's eye' are part of the response to a child's own question. It is this question which a child must follow to follow the development of his thinking, that is prime to the teacher—not her own preconceived questions and answers' (Weber 1971, p. 229). The teacher's intention as facilitator is not to judge but to reflect back to the child the consequences of the child's behavior by guiding the child through questions about activities and discoveries relative to the learning experience (Barth 1972). Because diagnosis occurs constantly during teacher–child interactions, errors are acceptable as they provide necessary feedback for continued, guided learning.

Through frequent note-taking describing the child's feelings, attitudes, and experiences, the teacher assesses and evaluates growth and learning. From this perspective, formal evaluation procedures are ineffective because, rather than being preoccupied with standardized and normative expectations, the teacher is vitally concerned with knowing that each child is making the best possible use of his environment and that he or she is making progress. Careful observation of experiences, linked with accurate recording, constitutes the scientific study of the learning process.

This effectiveness in the diagnosis of learning events depends entirely on the teacher's knowledge of the

child's ability, individual needs, stage of development, and learning style. The teacher's focus is on the child who, in turn, is expected to accept both responsibility and internal personal control. Barth (1972) extends this idea by identifying the child as 'an agent of his own learning' and teachers as 'facilitators of learning' who encourage, help, and assist rather than prescribe, mold, coerce, or manipulate. As a facilitator of learning, which is a highly individual process in which teaching styles are as numerous, complex, and individually effective as learning styles, there is no one specific approach to teaching. In addition, in the multiaged open classroom, students of proven ability, together with parent volunteers and aides, assist other students who are having difficulty in assimilating the knowledge, as research has shown that children benefit from differentiated staffing in which contacts with more than one teacher is more beneficial than having the same teacher all day long. Peer tutoring, where students must think through a topic carefully because they are responsible for teaching it to other students, and collaborative learning, where students share responsibility for learning something and must organize themselves and the topic to do it well, are necessary learning activities in the open classroom.

The belief that children can sit passively and absorb information is not one of the assumptions on which learning occurs in the open classroom but, instead, children must be actively engaged in the process of their own learning. Perkins (1995), in investigating the practice of ability grouping and tracking in K-12 education, viewed the open classroom (with heterogeneously mixed grades) as a superior learning environment because research shows that students understand and retain what they study when instruction engages them in thinking about, and with the content of, their curriculum. Perkins' thesis is that learning is a consequence of thinking, and if students are not constructively engaged and connected with their thinking processes and their curricular studies, they will neither fully understand nor retain them. From his perspective, the teachers' basic job in an open classroom is to insure that the students do a majority of thinking about and with the content of what they are learning. This perspective implies a number of learning techniques: *problem-based learning* (where students study content by seeking out the information needed to solve the problems); *project-based learning* (where students engage in complex and often socially meaningful projects that enable them to gain content knowledge); *engagement in understanding performances* (that ask students to use what they know in order to demonstrate and build their understanding); *infusion of critical and creative thinking into subject-matter instruction* (where students learn to analyze, critique, defend, adopt alternative points of view, and ask 'what if' questions); and the use of *authentic problems* (that have real-world significance and a messy open-ended character). Although thinking-centered learning can be complex and challenging for teachers and administrators to sustain, the ways to configure thinking-centered learning are endless.

2.3 A Learning Concern

There does, however, appear to be a danger that students in the open classroom may display less involvement in their activities and, consequently, learn less. Large discrepancies in time on task appear to be widespread as the findings reported by Bennett (1981) on open education in Britain are very similar to those reported by Fisher et al. (1978) in the United States. 'It is of interest to note that the curriculum areas which are allocated the most time show the poorest involvement levels' (Bennett 1981, p. 23). In addition, transition time (from one activity to another) which was studied as part of a survey in the United Kingdom (Bennett et al. 1980), observed examples of wasting time and concluded that students in open schools spend too much time in transition from one activity to the next or from one space to another. Similarly, New Zealand elementary students who were taught at tables rather than seated in rows had increased peer interaction and transition time between activities.

3. Teaching in the Open Classroom

3.1 Important Teacher Functions

As teachers in the United Kingdom have frequently pointed out, the teacher's primary responsibility is to set the mood or climate of the learning environment as the needs of each student change. In this stage-setting role, the teacher functions as a stimulator who encourages and guides but does not direct children. The teacher sets up opportunities for learning experiences and children's questions become curriculum-starting points. In fact, the teacher's role is to not impose on children, but to closely cooperate with their native interests and drives so that whatever they are led to do is viewed as something that comes out of themselves. Weber (1971), who calls this 'opening up a child's purposes,' values respect for the child's questions and purposes as serious forces in learning, while other educational theorists maintain that demonstrating respect for children and managing the learning environment are key teacher functions. Doyle (1986) characterized the open classroom teacher as an effective manager who ensures that students are concentrating on appropriate work in an environment where they have a high degree of choice.

Common to all open classroom facilitators is an enormous trust in learners and the ability to relinquish large amounts of control. Facilitators are more effective if they themselves are involved in learning and challenge their own beliefs and practices when con-

fronted by students. Because, generally speaking, there is no standard curriculum imposed upon the student, three teacher functions have been identified as critical roles of the open-classroom teacher: individualizer of instruction, resource person and counselor, and provider of suitable materials and activities.

Walberg and Thomas (1972) identified a wide range of beliefs and behaviors about the teacher's role which were grouped into eight themes that are multidimensional and interactive in nature, yet discrete enough to serve as an organizer for an understanding of the teacher's role in the open classroom. This eight-dimensional taxonomy, around which clusters of significant research has been conducted, are: *instruction* (guidance and extensions of learning), *provisioning* (provisioning the classroom for learning), *diagnosis* (diagnosis of learning events), *evaluation* (reflective evaluation of diagnostic information), *humanness* (respect, openness, and warmth), *seeking* (seeking opportunities to promote growth), *self-perception* (the teacher's view of herself and her role), and *assumptions* (ideas about children and the process of learning).

3.2 Planning, Relationships, and Management Skills

Research into open classroom processes reveals that effective teaching requires a high degree of planning, positive personal relationships, and excellent classroom management skills. The teacher creates the environment by restructuring the dimensions of time, space, materials, grouping, and communication. The classroom reflects a diversity of homemade materials to interest the students, books geared to a wide range of reading and interest levels, and a variety of carefully planned activity areas that encourage the student to explore.

Personal relationships are positive as the teacher interacts easily with students who enjoy being around her. The teacher is a 'real' person, cognizant of and able to express feelings, needs, attitudes, and motives. Consequently, effective open classroom teachers exhibit characteristics of self-acceptance, self-respect, security, and self-confidence; they are encouraged to be themselves and to reject adopting the role of a teacher at the expense of being oneself.

Several important aspects of teacher classroom management, which are more vital in open than in traditional classrooms, focus on clarity of explanations, vigilance and mobility, and pupil independence. Clarity of explanations are crucial in open classroom teaching as students ought to have a clear idea of what the project entails and the procedure for accomplishing the task. Vigilance and mobility indicates that the effective open teacher has excellent eye contact, and is able to split attention between the individual or group the teacher is working with and the rest of the class. Periodically, the teacher scans the class and is mobile enough to intervene in any problem before it escalates. This vigilance is transmitted to the class when, at intervals, the teacher publicly reviews the progress of different students or groups of students. Another message regularly transmitted is that of pupil independence so that the teacher is not overwhelmed with six or eight students gathered around the teacher's desk or trailing the teacher around the room.

3.3 Teacher Characteristics, Attitudes, and Reactions

The redefinition of the teacher's role as that of a facilitator, which encourages short-term decisions directed toward individual students, increases cooperation and communication among faculty and staff members. This research conclusion was reinforced by Harrison and Glaubman (1990) when they studied a systematic random sample of 42 Israeli elementary schools of which 17 had been implementing open education practices for at least three years. Teachers in the open education schools were significantly younger, had less experience than teachers in traditional schools, and reported having more frequent discussions of school goals and school problems with administrators, supervisors, parents, and members of the community. They used more resources for curriculum development and, as might be expected, the Israeli teachers in open schools described their school decision making process as being more decentralized than teachers in traditional schools. Finally, Israeli teachers, who taught in schools that adopted the open classroom approach, reported significantly more skill development, as measured by the number of hours of school-base training, than did teachers in traditional schools.

Schubert and Baxter (1982) documented a close relationship between teachers' attitudes toward open education and noncognitive student outcomes. Teachers have almost uniformly rated themselves as more open than they are and the difference is especially pronounced in the traditional classroom. When open and traditional classrooms are compared, students in an open-space school or a closed-space school (where teachers have favorable attitudes toward open education principles) consistently scored higher with regard to level of curiosity and self-esteem than students of teachers with unfavorable attitudes toward open education. A significant sample of public school teachers, who taught in both open and traditionally constructed schools, had significantly more positive attitudes toward open education than teachers who taught only in open space schools. Consequently, there is a positive relationship between the number of years of teaching experience and attitudes toward open education because teachers who taught 10 years or more in the open classroom were significantly more positive in their attitudes toward open education than teachers who taught less than 10 years.

Because intrinsic rewards (i.e., to play a significant role in students' lives) are important for open teachers

and because the classroom is the major arena for the receipt of the rewards (Sederberg and Clark 1990), teachers in open schools, generally speaking, show no significant inclination to move back into the self-contained classroom. However, 10 Canadian studies detailing adverse teacher comments report increased workload and insufficient preparation time, as teaching in an open classroom requires continuous innovation, monitoring of students' performance, and well-developed skills in maintaining order without being authoritarian (Traub et al. 1976). New Zealand teachers stated that increased workload was their major problem and that demands on them were very high, citing the necessity for continuous student evaluation and careful curriculum planning. This teacher reaction to the open classroom was reinforced by Australian teachers who maintained that the open classroom workload was having a negative effect on their private lives (Angus et al. 1975).

4. Conclusion

Research which indicated that cognitive achievement levels in open classrooms were not significantly different from those in traditional classrooms initiated the decline of open education in the United States and by the late 1970s the movement was in retreat. Today, the literature is clearly characterized by its critics but, because widespread research has indicated that students in open classrooms are more independent, less conforming, act more cooperatively, are more attentive and participate more actively in class discussions, and display more creativity than students in traditional classrooms (Rothenberg 1989), open classroom learning strategies continue to be implemented in Western Europe, North and South America, and Asia as many schools continue to offer students the option of learning in open classrooms.

See also: Classroom Climate; Educational Research and School Reform; Environments for Learning; Instructional Design; Instructional Psychology; Pedagogical Reform Movement, History of; Progressive Education Internationally; Rousseau, Jean-Jacques (1712–78); School (Alternative Models): Ideas and Institutions; Self-regulated Learning; Situated Learning: Out of School and in the Classroom; Teacher Behavior and Student Outcomes; Teaching and Learning in the Classroom

Bibliography

Angus M J, Evans K, Parkin B 1975 *An Observation Study of Selected Pupil and Teacher Behaviours in Open Plan and Conventional Designed Classrooms*. Technical Report 4, Australian Open Area Schools Project, Educational Department of Western Australia, Perth

Barth R S 1972 *Open Education and the American School*. Agathon, New York

Bennett N, Andrae J, Hegarty P, Wade B 1980 *Open Plan Schools: Teaching, Curriculum, Design*. National Foundation for Educational Research, Windsor, Berkshire, UK

Bennett S N 1981 Time and space: Curriculum allocation and pupil involvement in British open-space schools. *Elementary School Journal* **82**: 18–26

Doyle W 1986 Classroom organization and management. In: Wittrock M C (ed.) *Handbook of Research on Teaching*, 3rd edn. Macmillan, New York

Fisher C, Filby N, Marliave R, Cahen L, Dishaw M, Moore J, Berliner D 1978 *Teaching Behaviors, Academic Learning Time and Student Achievement*. Far West Lab, San Francisco

Harrison J, Glaubman R 1990 Innovations and school improvements. *Journal of Research and Development in Education* **24**(1): 21–33

Horwitz R A 1979 Psychological effects of the 'Open Classroom.' *Review of Educational Research* **49**(1): 71–8

Perkins D N 1995 *Smart Schools: Better Thinking and Learning for Every Child*. Free Press, New York

Rothenberg J 1989 The open classroom reconsidered. *Elementary School Journal* **90** (1): 69–86

Schubert N A, Baxter M B 1982 The self-contained open classroom as a viable learning environment. *Journal of Education* **102** (4): 411–15

Sederberg C H, Clark S M 1990 Motivation and organizational incentives for high vitality teachers: A qualitative perspective. *Journal of Research and Development in Education* **24** (1): 6–13

Traub R, Weiss J, Fisher C 1976 *Openness in Schools: An Evaluation Study*. Ontario Institute for Studies in Education, Toronto

Walberg H J, Thomas S C 1972 Open education: classroom validation in Great Britain and the United States. *American Educational Research Journal* **9**: 197–208

Weber L 1971 *The English Infant School and Informal Education*. Prentice-Hall, Englewood Cliffs, NJ

E. F. Thibadeau

Operant Conditioning and Clinical Psychology

The application of operant conditioning to issues confronting clinical psychology has a strong foundation in the basic science of B. F. Skinner. The use of operant principles in clinical psychology focuses on altering the environment to increase useful behaviors or decrease clinically maladaptive behaviors. Modern operant theory continues to extend the original model and the scope of the clinical problems it can address, and new therapies continue to be developed.

1. The Basic Operant Model

Operant theory was preceded by the law of effect as proposed by Edward L. Thorndike. Early versions of the law of effect stated that behaviors that produce satisfying effects tend to increase in frequency while behaviors that produce discomfort tend to become less frequent. B. F. Skinner recognized that attending only

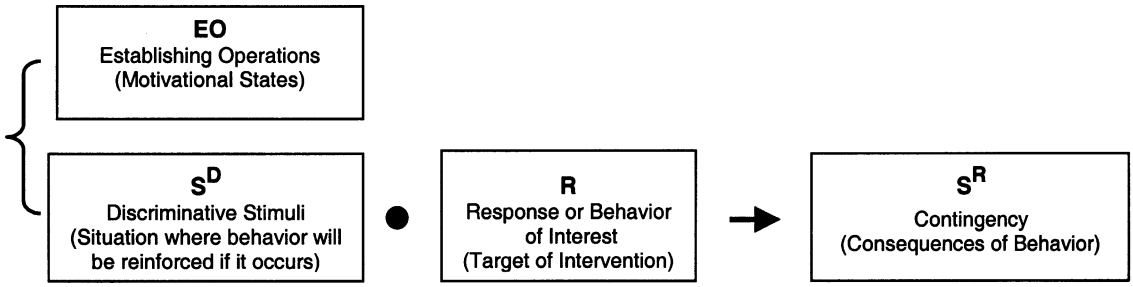

Figure 1
Representation of the multiterm contingency for the operant model of behavior

to behavior missed an important point. A behavior can produce a good effect in one situation but a very different effect in a different circumstance. In part, an appreciation of the importance of the function of behavior in the specific context in which it occurred led to the identification of the principles of operant conditioning (see *Behavior Analysis*, *Applied*).

An operant gains its name from the fact that the behavior in question *operates* on the environment or produces a certain consequence. An operant is defined by the consequence it produces and not by the physical appearance or form (also called topography) of the behavior. Operant (also called instrumental) conditioning is typically presented in what is called the 'multi-term contingency.' The analysis of a behavior takes place in the context of the circumstances in which it occurs. A common representation of the operant using three explanatory terms is presented in Fig. 1. The conventional symbols are indicated in bold above the name and descriptions.

Reading from left to right the diagram can be understood to indicate that under certain conditions a behavior will occur and, depending on what follows the behavior, the behavior will be more or less likely to occur in the future under the same circumstances. Looking in more detail at the diagram, the environmental influences that set the occasion for a behavior to occur are shown in the brackets on the left and should be considered together. The discriminative stimulus (S^D) is a stimulus in the presence of which the probability of a particular contingent consequence for a behavior or class of behaviors is greater than when it is not present. Bracketed with the discriminative stimulus is the establishing operation (EO). Establishing operations are any environmental changes that momentarily alter the effectiveness of a discriminative stimulus to evoke the behavior of interest and affect the rate of that response.

The 'R' indicates the response of interest. Responses can be overt actions or private events such as thoughts, feelings, and emotions so long as there are some observable correlates of the private events (Skinner 1945). Some critics of operant conditioning incorrectly state that only overt behavior and not private events can be studied and treated using an operant approach. In fact, operant conditioning assumes thoughts and feelings are better understood as responses to be addressed in treatment rather than causes of other behaviors (Hayes and Brownstein 1986).

The 'S^R' term indicates the contingency that follows the behavior and influences the probability of the preceding behavior occurring again under the same or similar conditions. The notation of 'S^R' is used, because typically the contingency is a stimulus that acts as a reinforcer. The term 'reinforcer' has a specific meaning in operant conditioning. A reinforcer is a stimulus that increases the probability of a given response in a particular situation when it is presented following that response and contingent upon it. The arrow indicates the temporal flow of the response being followed by the contingency. The other notation in the diagram is the dot between the discriminative stimuli and the response. The dot is interpreted as a probability function. That is, given the presence of a discriminative stimulus, there is a probability that the response will be emitted given the organism's history.

1.1 Acquisition of Behavior

Consider this simple example to clarify these terms. In the absence of intimate friendships and following social-skills training in therapy (EO), a client meets a prospective friend (S^D) and makes a social overture to them (R) that leads to the formation of a new friendship (S^R, in this case a reinforcer for approaching people as prospective friends). This sequence of behaviors leads to the strengthening of the response of initiating social contact.

Behavior is strengthened when it is followed by a reinforcing consequence. Generally, the more frequently the behavior of interest is followed by a reinforcer, the more quickly the behavior will become established. There are other well-researched factors that affect the rate of acquisition and stability of behavior (Ferster and Skinner 1957).

Distinctions are made for various types of contingencies. The most fundamental distinction is be-

Table 1
2 × 2 table defining the use of term for the operant contingency

	Positive	**Negative**
Reinforcement	A contingency is *applied* that *increases* behavior	A contingency is *removed* that *increases* behavior
Punishment	A contingency is *applied* that *decreases* behavior	A contingency is *removed* that *decreases* behavior

tween reinforcement and punishment. Any contingent stimulus that leads to an increase in the preceding behavior is called a reinforcer. Any contingency that leads to a decrease in the preceding behavior is called a punisher. The terminology is further refined by identifying whether the contingent stimulus is applied or removed to establish its effect. When a contingent stimulus is applied, it is termed positive; when a contingent stimulus is removed, it is termed negative. Table 1 clarifies the terminology of contingent stimuli.

Whether a contingency is determined to be reinforcement or punishment is determined by the effect it has on the preceding behavior, *not* the intention of the supplier of the contingency. For example, even if a parent intends to punish a child's tantrum by scolding them in order to decrease rate, if the scolding actually increases the tantrums (due to parental attention, perhaps) then the scolding has functioned as positive reinforcement, regardless of the intention of the parent. Thus, it is important to understand that exactly what type of contingency is operating is wholly determined by the observed function of the contingency on the response of interest (see *Single-case Experimental Designs in Clinical Settings*).

A common misunderstanding about contingencies is that the modifiers 'positive' and 'negative' imply some moral good or bad property of the contingency. This is not the case. The terms 'positive' and 'negative' refer only to whether the contingency (i.e., the reinforcement or punishment) is applied (positive) or removed (negative). A positive reinforcer refers to a contingency that, when applied, increases the rate of responding. A negative punisher refers to the removal of contingency that produces a reduction in the rate of responding.

1.2 Generalization

When an operant behavior has been reinforced in the presence of a particular S^D, the same behavior may still be emitted in the presence of similar but not identical stimulus conditions. When an operant is emitted in the presence of a stimulus similar to the original S^D, this is referred to as stimulus generalization. For example, if one learns to answer the door when a doorbell of a particular sound rings, one will likely answer a door when a doorbell of a somewhat different ring occurs, even if one has never heard that particular doorbell sound before. This is an example of stimulus generalization. Strictly speaking, generalization occurs when some actual physical property of the original S^D is shared with another stimulus in the presence of which a response is emitted and reinforced. A doorbell's sound may be similarly pitched and thus the operant (opening the door) may be emitted, even though the behavior has never been reinforced in the presence of that exact tone before. However, for the phenomenon to be considered generalization, there must be some formal property of the stimuli that are common. In the doorbell example, the formal property was some quality of the sound.

1.3 Extinction

Behavior that is established by operant conditioning can undergo extinction when the behavior that occurs in the presence of the S^D is no longer followed by the S^R. One of the characteristics of extinction of operant behavior is that when the S^R is removed, the organism may increase the frequency of the behavior initially and often change the topography of the behavior slightly. This is called an extinction burst and can lead to the expression of new behavior that may lead to reinforcement. If reinforcement is not forthcoming, the behavior will decrease to its pre-reinforcement baseline (referred to as the free operant level).

2. Early Applications of Operant Conditioning

There have been many applications of operant conditioning in the clinical literature. One example, in particular, demonstrates an important philosophical perspective taken by clinical operant therapists. Under the assumption that behavior is the product of environmental contingencies, two researchers, Teodoro Ayllon and Nathan Azrin (1968), studied chronically hospitalized patients at a state hospital. The patients had been in the hospital for many years and many had been diagnosed with schizophrenia and other severe forms of psychopathology. The researchers provided no direct treatment to the patients but instead altered

the behavior of the staff. They assumed that the staff were inadvertently reinforcing maladaptive behavior and failing to attend to positive patient behaviors properly. Their sole intervention was to have the staff reinforce adaptive behaviors when the patients emitted them, by giving them tokens that were redeemable for a variety of reinforcers, and ignore maladaptive behaviors. This intervention is referred to as a token economy. Remarkable improvements resulted, including a number of instances where patients who had been institutionalized for many years prior to the introduction of the token economy were functioning well enough to be discharged.

Another notable early application of operant principles to clinical problems addressed severe problems in children. O. Ivar Lovaas (1977) was among early autism researchers who used operant principles to stop self-destructive behavior and establish language. One characteristic of autistic behavior is impaired or absent language. Lovaas used reinforcement principles to establish the utterance of sounds by reinforcing sounds and then withholding the reinforcement and waiting for an extinction burst to produce sounds more like the target sound. When closer approximations to target sounds occurred, they were quickly reinforced. This use of operant conditioning to build increasingly more complex behavior is called shaping by successive approximation. While the process of teaching language in this population was slow, it can be very successful. Further work at shaping verbal and social skills has produced profound improvement in verbal and social behavior in a population that had been difficult to treat previously (Lovaas 1987).

3. *Clinical Interventions for Each Term of the Operant Model*

An operant analysis of clinical problems requires that behavior be understood as an act in context. What this means is that in order to understand a behavior, one has to understand its function in the context in which it occurs. As a simple clinical example, crying could function to express sadness, to get support, to terminate an aversive conversation, to express overwhelming happiness, or to get attention. To understand the act of crying, one would have to understand the conditions in which it occurs and the consequences that follow the crying contingently. To say that someone is crying does not demonstrate an adequate understanding of the controlling variables. If a clinician were to treat someone's crying as a reflection of sadness or depression, when its function was actually to terminate a conversation that was not going the client's way, the therapist would not achieve a good clinical outcome. When a clinician tries to understand a clinical problem, they perform a functional analysis, or an analysis of the stimulus functions of each term in the operant model (Kanfer and Saslow 1969).

It is useful to think of how problems and interventions may occur for each term of the operant model. Table 2 describes each term shown in the model followed by a description of the functional problem, a description of the process by which the problem might emerge, a clinical example, and the type of possible intervention that might be used to address the clinical problem.

One can see that in many cases, treatment involves techniques that provide training to better discriminate conditions that indicate when a response will lead to a particular desirable or undesirable outcome. An assessment of the adequacy or frequency of a particular response is also often useful. Many times an intervention focuses on shaping a better functional response for a particular set of stimulus conditions. This may take the form of social-skills training, modeling, or training by successive approximations. There are many clinical situations when clients are taught how to better apply contingencies that affect the behavior of people around them. For example, parent effectiveness training focuses on the proper and effective application contingency management to increase useful behavior in children. The same principles can be used to increase appropriate student behavior in school or increase productivity in a business environment.

Influencing establishing operations or motivational variables is the clinical intervention with the least well-developed clinical technology. In laboratory research, deprivational states are frequently used to establish the function of a discriminative stimulus for a response that will be reinforced by food. In clinical practice, inducing deprivational states is often not practical or ethical. While parents might deprive a child of their dinner, behavior therapists do not have access to such operations. Recently, motivational interview techniques have been developed to emphasize the importance of therapy and increase its reinforcing properties. These techniques have been applied with some positive effect in the area of addictions treatment (Miller et al. 1992), but the strength of the effect and the exact mechanism of action are not clear.

4. *Clinical Interventions Involving Operant and Respondent Conditioning*

Respondent (also known as classical, or Pavlovian) conditioning (see *Classical Conditioning and Clinical Psychology*) involves the pairing of unconditioned and conditioned stimuli, which ultimately leads to a conditioned stimulus that elicits a conditioned response. One of the more useful clinical heuristics has been research on how respondent and operant conditioning can combine to explain important clinical problems.

The best-known problem that has been addressed by considering both operant and respondent con-

Table 2
Clinical problems associated with each term of the operant model

Stimulus function in operant model	Functional problem	Description	Specific problem example	Possible interventions
EO Establishing Operation	Inadequate motivation	Person does not take advantage of reinforcing opportunities	Client fails to seek treatment for substance abuse	Motivational interviewing
S^D Discriminative Stimulus	Inadequate stimulus control	Response emitted under wrong conditions	Client gives inappropriate interpersonal information for a given occasion	Social-skills training
	Overly rigid rule-governance	Client is not under contingent control of environment or sensitive to environmental changes	Client's behavior not under control of actual relationship between behavior and consequences	Exposure to actual contingencies in a situation
	Ineffective arrangement of contingencies	Immediate environment not arranged to control behavior usefully	Inefficient dietary management; problems with school work	Stimulus control/problem-solving training
R Response	Deficient or ineffective behavioral repertoire	Behaviors do not lead to reinforcement	Poor social skills, parental skills deficits, social isolation, assertion deficits	Therapy provides feedback on actual function of behavior; Parent effectiveness training
	Aversive behavioral repertoire	Exhibits behavioral excesses or controlling behaviors that others find aversive	Aggressive or coercive behavior	Self-control training
	Behavioral excess	Client exhibits excess behavior that interferes with access to reinforcers	Panic/Anxiety, obsessive-compulsive behavior	Systematic desensitization; flooding; panic control therapy
S^R Contingency	Inappropriate contingent control	Reinforcers socially inappropriate or lead to unacceptable outcomes for client or others	Substance abuse; pedophilia	Self-control training, social-skills training, environmental restriction
	Non-contingent reinforcement	Person receives significant reinforcement regardless of performance	Social, school, or work difficulties	Parent training, business management training
	Overly-punitive environment	Person's behavior is under aversive control rather than positive control	Mistrust, over-caution, misses opportunity for expanding behavioral repertoire	Rearrangement of environmental contingencies
	Overly stringent self-reinforcement strategies	Person sets standards too high	Perfectionism, frequently disappointed in self or others	Self-instructional training, discrimination training

ditioning is the theory of the acquisition and maintenance of phobic behaviors. It has been suggested that phobic behaviors are acquired by classical conditioning but maintained by operant conditioning. Consider the simple example of someone bitten by a dog. In respondent conditioning terms, the dog bite is an unconditioned stimulus that produces the unconditioned response of pain and fear. Following such an incident, the next time the person approaches a dog, their fear and anxiety rises as the stimulus (the dog) gets closer. So far, the acquisition of the fearful response can be understood using a classical conditioning paradigm. If the person were to approach a variety of dogs, the fearful response would extinguish naturally, because extinction in classical conditioning is accomplished by presenting the conditioned stimulus (a dog) in the absence of the unconditioned stimulus (the dog bite). If this were the case, phobic responses would extinguish naturally over time. However, in many instances when one sees the dog and anxiety increases, a person simply turns around and walks away, thus avoiding the feared object. When that happens, the avoidance behavior is negatively reinforced (increased) by the removal of the anxiety. This increases the probability of avoiding the dog the next time such a stimulus is encountered. The avoidance of the phobic object prevents the natural extinction of phobic anxiety, because the phobic object (now a conditioned stimulus) is avoided and therefore extinction cannot occur.

Avoidance is an important issue in clinical psychology. Avoidance responses are operants that prevent the occurrence of aversive consequences before they are actually experienced. This behavior is maintained by negative reinforcement. Clinically, the liability of avoidance behavior is that the person engaging in such behavior does not experience the opportunity to test whether the anticipated aversive consequences are still in effect. Thus, the circumstances that led to the initial aversive consequences may have changed, but if the person continues to avoid the original stimulus conditions, the changes will go undetected. There may also be avoidance of other stimuli due to generalization that leads to additional restrictions in healthy functioning.

Several clinical interventions address such problems. Treatments for phobias involve therapeutic interventions that prevent or remove the instrumental benefits of avoidance. Phobia treatment involves a classical conditioning paradigm in which the behavior therapist uses exposure to the conditioned stimulus to bring about extinction (see *Behavior Therapy*: *Psychological Perspectives*). The key to successful treatment is the prevention of avoidance, which would negatively reinforce the phobic behavior (Barlow 1990).

Another clinical problem that is treated, in part, by preventing avoidance behavior is obsessive-compulsive disorder (OCD). In OCD, the client experiences intrusive thoughts or images that produce anxiety. For example, someone might be obsessed with a concern that they have failed to lock their house adequately. The thoughts are high in frequency, do not feel natural to the client, and are not under the voluntary control of the client. Obsessions are thoughts or images. They are often accompanied by compulsive behaviors that serve to reduce the obsessive thoughts. In this example, a client may go back repeatedly to check that the front door is locked, preventing them from going to work. The psychological intervention used to treat OCD is exposure to the situation that produces the obsessive behavior and response prevention so that the compulsive behaviors are not emitted (Foa et al. 1980). Eventually, the anxiety associated with the problematic stimulus extinguishes, because the function of acting to reduce the distress extinguishes.

5. *Future Directions in Operant Conditioning Research and Treatment*

In the last 15 years there have been significant advances in operant theory. Perhaps the most important advances address the limitations of the application of the operant conceptualization of stimulus generalization to verbal behavior. Humans are unique in their ability to engage in certain kinds of verbal behavior. Earlier, stimulus generalization was defined as when a stimulus with similar physical or formal properties to the original discriminative stimulus is also followed by a response that is subsequently reinforced. However, in the instance of verbal behavior, there are stimuli that have no formal properties in common that can still serve equivalent functions. A simple instance is when the written word 'g-r-e-e-n,' the sound of the word 'green' when pronounced, and the color green all serve the same stimulus function without necessarily ever having appeared together. There are no physical similarities shared by these stimuli. They can be trained to substitute for each other so that any one stimulus can entail the functions of each of the others. When each stimulus can entail the functions of all the others and demonstrate certain experimental properties (reflexivity, symmetry, transitivity), they are said to form an equivalence class.

The clinical implication for humans is that stimuli can take on the functions of other stimuli without ever having been directly learned. Though the process is debated (Barnes-Holmes et al. in press, Sidman 1986, 1994), it is clear that one can construct functionally equivalent relationships between stimuli in complex ways. What this means clinically is that a client can respond to a variety of stimuli as if they all had the same direct learning history, even if they did not. While there are experimental methods for training these relationships, they often emerge without a person knowing that an equivalent relationship is functioning

or how it came to be. A client might learn to respond to an event—for example, a relationship failure—in the same way that they might to a catastrophic event, even though a failed relationship had never previously entailed a horrible outcome. Its implications for therapy are just beginning to emerge (Hayes et al. 1999) (see *Acceptance and Change, Psychology of*). The promise of the exploration of the relational operant in the analysis of language and therapy is one of the more interesting innovations for those doing research on the cutting edge of operant theory.

In the last decade, an analysis of how to apply operant learning principles in interpersonal psychotherapy has also begun (see *Interpersonal Psychotherapy*). Kohlenberg and Tsai (1991) have detailed functional analytic psychotherapy, which describes the reinforcing functions of the therapeutic relationship that the therapist can use to affect significant behavior change in adult outpatient populations.

Advances in basic operant research are emerging at an increasing rate. Innovations in therapy are addressing new classes of problems that have not previously been the focus of an operant analysis. The role of operant conditioning will continue to maintain a prominent position in applied clinical psychology.

See also: Behaviorism; Reinforcement, Principle of

Bibliography

Allyon T, Azrin N H 1968 *The Token Economy: A Motivational System for Therapy and Rehabilitation*. Appleton-Century-Crofts, New York

Barlow D H 1990 *Anxiety and its Disorders: The Nature and Treatment of Anxiety and Panic*. Guilford, New York

Barnes-Holmes D, Hayes S C, Hegarty N, Dymond S (in press) Multiple stimulus relations, the transformation of stimulus functions, and the limits of class concept. In: Hayes S C, Barnes-Holmes D (eds.) *Relational Frame Theory: Creating a New Behavioral Agenda in Language and Cognition*. Context, Reno, NV

Ferster C B, Skinner B F 1957 *Schedules of Reinforcement*. Appleton-Century-Crofts, New York

Foa E B, Steketee G, Milby J B 1980 Differential effects of exposure and response prevention in obsessive-compulsive washers. *Journal of Consulting and Clinical Psychology* **48**: 71–9

Hayes S C, Brownstein A J 1986 Mentalism, behavior-behavior relations and a behavior analytic view of the purposes of science. *The Behavior Therapist* **9**: 175–90

Hayes S C, Strosahl K D, Wilson K D 1999 *Acceptance and Commitment Therapy: An Experiential Approach to Behavior Change*. Guilford, New York

Kanfer F H, Saslow G 1969 Behavioral diagnosis. In: Franks C M (ed.) *Behavior Therapy: Appraisal and Status*. McGraw-Hill, New York, pp. 417–44

Kohlenberg R J, Tsai M 1991 *Functional Analytic Psychotherapy*. Plenum, New York

Lovaas O I 1977 *The Autistic Child: Language Development Through Behavior Modification*. Irvington, New York

Lovaas O I 1987 Behavioral treatment and normal educational and intellectual functioning in young autistic child. *Journal of Consulting and Clinical Psychology* **55**: 3–9

Miller W R, Zweben A, DiClemente C C, Rychtarik R G 1992 *Motivational Enhancement Therapy Manual. A Clinical Research Guide for Therapists Treating Individuals with Alcohol Abuse and Dependence*. National Institute of Alcohol Abuse and Alcoholism, Rockville, MD, Vol. 2

Sidman M 1986 Functional analysis of emergent verbal classes. In: Thompson T, Zeiler M D (eds.) *Analysis and Integration of Behavioral Units*. Erlbaum, Hillsdale, NJ, pp. 213–45

Sidman M 1994 *Stimulus Equivalence: A Research Story*. Authors Cooperative, Boston

Skinner B F 1945 The operational analysis of psychological terms. *Psychological Review* **52**: 270–7

W. C. Follette

Operations Research

1. What is Operations Research?

Operations research (OR) is the name for an approach to problem solving that is intended to provide managers with a scientific basis upon which to make decisions in the interests of their organization as a whole. Modern, complex organizations are divided into functional units, each of which has its own objectives. In a private manufacturing firm, for example, the production department usually seeks to maximize production volume while minimizing costs, often calling for large inventories of relatively few product lines; the marketing department seeks to maximize sales volume while minimizing the costs per unit sale, often calling for rapid delivery of a large number of products; and the personnel department seeks to recruit the best people at the least cost, which can be achieved by a stabilized labor force. Often, these objectives are in conflict with each other. It is the aim of the firm's chief executive to harmonize these functional or component objectives so that the organization's overall objectives—profitability, return on investment, market share, stock price—are achieved. One can consider a 'system' as the interrelationship among the components or functional units of the organization. Operations research seeks to find the best solution available to achieve the system's objectives given the constraints under which the system operates. This is termed 'optimization' analysis.

The key elements of an operations research approach are the following:

(a) Formulate a problem to be studied.

(b) Construct a mathematical model that represents the system under study.

(c) Derive a solution from this model.

(d) Test the model and the solution derived from it.

(e) Establish rules for modifying the solution under specified system changes.

(f) Implement the solution.

1.1 Evolution of Operations Research

As the term implies, 'operations research' is concerned with research about operations. One of the earliest examples of OR are the Lanchester equations published during World War I. Frederick Lanchester was an aeronautical engineer who sought to translate complex military strategies into mathematical formulas. His book, *Aircraft in Warfare*, published in 1916, included the 'N square' law relating victory, numerical superiority, and weapon superiority for situations in which all units of two opposing forces engage simultaneously. Lanchester concluded that, under these assumed conditions, numerical superiority was relatively more important than weapon system accuracy, and that available combatant forces were most effective if deployed in a single force that split the enemy into two or more component units.

During World War II, the British government employed operations analysts to determine whether quantitative assessment of sensitive military data could produce recommendations that would enhance the effectiveness of British military operations. Numerous successes followed. For example, data analyzed from the Eighth Bomber Command in England in 1942 led to several operational modifications: Simultaneous release of bombs by all bombardiers instead of individualized release by each bombardier; decrease in the number of aircraft per formation from 18–36 to 12–14; and salvoing of bombs instead of presetting them to release in a string produced startling results. In 1942, less than 15 percent of bombs dropped landed within 1,000 feet of their aim point. This figure rose to more than 60 percent in 1944 after the operational modifications were adopted.

This example illustrates a number of the characteristics of OR: Systematic analysis of operational data led to the recommendation of 'counter-intuitive' recommendations that ran contrary to the experiences of the operational experts, and that, once implemented, yielded greatly improved overall system effectiveness.

The British establishment of 'operational research' units in the early 1940s were replicated by the USA during and immediately after World War II. The US Navy formed an Operations Evaluation Group, and this organization was later incorporated into the Center for Naval Analyses which performs OR and related studies for the Navy. The US Air Force, with the help of Donald Douglas of Douglas Aviation, established The RAND Corporation which has played a leading role in the creation and application of OR techniques since the 1940s. And the Weapons System Evaluation Group, reporting to the Joint Chiefs of Staff, later became a key part of the Institute for Defense Analysis. (The US Army's Operations Research Office operated independently for a number of years, but later became a 'project' within The RAND Corporation.)

In the late 1940s and early 1950s, it was realized on both sides of the Atlantic that these quantitative techniques could be applied fruitfully to industrial as well as to military operations. Initially, some of these techniques were employed by management consultants and industrial engineers to improve industrial efficiency through quality control, time-and-motion studies, and marketing analyses. But the 1950s saw the proliferation of sophisticated mathematical methods that greatly enhanced the analytical power, academic prestige, and organizational applicability of OR. These techniques included mathematical programming ('linear,' 'nonlinear,' and 'dynamic') that derived optimal allocation solutions under constraints; queuing or waiting line theory; advances in game theory, and other aspects of competitive decision-making; and the early stages of computer simulation methods.

Robert McNamara, who was a graduate of the Harvard Business School and had become familiar with these approaches as a senior executive at the Ford Motor Company, was named Secretary of Defense by President Kennedy in 1961, just after having been appointed as president of Ford. McNamara formed a 'systems analysis office' at the Pentagon, headed by Alain Enthoven, a prominent economist who had applied OR techniques at The RAND Corporation. This was the first time that OR was legitimized as an important method of analysis in US defense policy. Moreover, McNamara urged his colleagues in domestic agencies to consider the application of these techniques.

In the 1960s, the Johnson administration established the Great Society, a set of federally-funded initiatives to enhance the well-being of individuals on the bottom rungs of American society. Such domestic programs as health policy and management began to receive extensive political and then analytical attention. Simultaneously, significant numbers of military analysts active in OR development and its applications became disaffected with the escalation of American military involvement in Vietnam and sought to apply their analytical skills in nonmilitary areas. Collectively, these developments produced a third wave of OR application: domestic, government-funded programs involving social, environmental, transportation, and criminal justice policy, to name some of the more prominent areas of work.

By the 1970s, operations research groups could be found not only in the major branches of the military services and in large numbers of corporations. They were also prevalent in federal and state agencies concerned with health and human services, environmental regulation, transportation, and numerous other areas of specialization. And they were the focal

point for work among many of the nongovernmental organizations that arose during this period to support and/or critique government programs.

Another noteworthy feature of the application of OR techniques is that, historically, they were conducted largely by teams of analysts rather than by individual analysts. OR became associated widely with multidisciplinary, team problem-solving.

1.2 Academic Bases of Operations Research

The burgeoning literature on the collection of OR techniques was both a product of, and a stimulus to, academically-based graduate programs in the field. Originally, OR was represented in units within departments of industrial engineering, or groups within graduate schools of business. In some cases, separate OR departments were established on faculties of colleges of engineering, or as part of applied mathematics and statistics departments. These disciplines—mathematics and statistics, engineering, and business—became the principal bases for university advancement of the field. They were joined by economics as the applicability of game theory—formal consideration of competitive situations emphasizing the decision-making processes of the adversaries—became widely used. And, in the 1980s, as rational choice theory, a derivative of game theory, took hold in political science departments, these units also became the home for elements of operations research teaching and research.

A wide variety of US educational institutions became prominent in the teaching and research of OR. The Massachusetts Institute of Technology, the Wharton School of the University of Pennsylvania, Carnegie Mellon University, Johns Hopkins University, Case Western Reserve University, the University of California at Berkeley, and Stanford University, among others, became known internationally for the quality of their OR programs. These academic activities were matched by important programs at major universities in Europe, Israel, Japan, and elsewhere–including Lancaster in the United Kingdom, Louvain in Belgium, and Tel Aviv and Tokyo Universities.

Not unexpectedly, a corollary of these academic programs was the establishment of academic societies whose purposes were to publish the latest research in the field, to promote professional networks that would enhance interaction among specialists, and to serve as a recruiting ground for placement in academic institutions, corporations, government agencies, and nonprofit organizations.

Most noteworthy, the Operations Research Society of America (ORSA) focused on mathematical advances in the field and its many applications in the military area. A separate society, the Military Operations Research Society (MORS), was devoted exclusively to this sector. By contrast, as larger numbers began to utilize OR for industrial applications, The Institute of Management Sciences (TIMS) was organized to meet the needs of those interested in corporate environments. Eventually, in the 1990s, ORSA and TIMS merged into the Institute for Operations Research and the Management Sciences (INFORMS), reflecting the growing interaction among OR in the military, corporate, governmental, and nonprofit sectors.

2. Major Classes of Operations Research Applications

Over the decades, clusters of problems have proven to be especially amenable to treatment by operations research. They fall into the following six categories of model: inventory, allocation, waiting time, replacement, competitive, and computer simulation.

2.1 Inventory Models

Inventory models deal with the time at which orders for certain goods are to be placed, and the quantity of the order. The research problem concerns ways of optimizing these decisions, taking into account the cost of obtaining the goods, the cost of holding a unit in inventory, and the cost of shortages. More advanced inventory models deal with situations in which there are restrictions on production facilities, storage facilities, time and/or money. For several decades, more OR effort was focused on inventory models than any other area of application.

2.2 Allocation Models

This class of problems concerns a number of activities to be performed, several alternative methods for carrying them out, and a constraint that the resources or facilities available do not permit each activity to be performed in the most effective way. The analytical challenge of these problems is to combine activities and resources to optimize overall effectiveness. Of particular relevance to this work is the applicability of 'linear programming'—a class of optimization problems that address the interaction of many variables subject to a set of constraints. Subsequent advances in 'nonlinear' and 'dynamic' programming permitted the tackling of more complex assignment problems where a number of the constraints themselves vary with time.

2.3 Waiting Time Models

A third class of problems that have long been addressed by operations researchers have dealt with waiting times—when either units requiring service or facilities available for providing service wait, or stand idle. These phenomena have been the subject of a

whole class of mathematical formulations known collectively as 'queuing theory.' In these problems, the OR analyst is faced with determining the optimal number of service facilities, the optimal arrival rate or times of arrival, or both.

2.4 Replacement Models

There are a class of problems concerned with the prediction of costs for a group of items that have a probability distribution of lifespan. Decision-makers must choose between operating aging equipment or incurring the cost of replacing old equipment with new. At what age does the saving from use of new equipment compensate for its initial cost? For many years, OR analysts have developed methods for predicting costs based on the probability distribution of lifespans. These are classified as replacement models.

2.5 Competitive Models

Unlike the models cited above, which deal with issues internal to the organization, competitive models are concerned with situations in which the effectiveness of decisions by one party are dependent on decisions by another party. At the core of these models is the theory of games which assesses the strategies that players should adopt to 'win,' however that is measured. A game is competitive if there is an end-state that each player desires but not all can obtain. The players are in conflict relative to the goal. Each player has a set of possible choices which eventually bring the game to the end state.

2.6 Computer Simulation Models

With the development of advanced computer technology, OR analysts began to use these machines to build mathematical models that could simulate real-world problem situations. Then, by varying the assumptions of the model, the analysts could determine the sensitivity of the model to different parameters and conditions. The power of this technique has permitted highly complex systems to be evaluated carefully in relatively short periods of time.

3. The Intellectual Challenges

Operations research techniques have advanced enormously over the almost six decades of their development since the 1940s. The mathematical rigor of the techniques is especially impressive. Its application in both the military sectors and the commercial sectors has been extensive. Areas of government/social action and of public policy more generally have been less amenable to contributions from the operations researcher. This is because OR analysts are committed to the studying of a complex problem through the lens of a mathematical model that can be tested in quantitative terms. But a number of broader social and political issues reflect deep political struggles, and adjustments in human behavior that are exceedingly difficult to quantify. Many military operations and corporate activities have agreed widely upon quantitative measures of effectiveness.

Here, the intellectual power of OR techniques can be enormously helpful. In public policy issues, where a clash of values is often at the heart of the matter (e.g., should the US Department of Energy be twice its size, or eliminated?), quantitative models appear to be of limited value in furthering the identification of optimal choices.

A second challenge is to make OR applicable in the age of the information technology revolution. This revolution has permitted vast amounts of information to be available to people all over the globe with access to the World Wide Web. Will this revolution strengthen or marginalize the role of OR as a tool of executive decision-making? Practitioners of these techniques must tailor their models as well as their areas of application to take advantage of this revolution, or they will be left behind by new analytical approaches that are found to be more 'user friendly' for these incredibly powerful technological devices.

See also: Bounded and Costly Rationality; Decision Research: Behavioral; Dynamic Decision Making; Heuristics for Decision and Choice; Organization: Overview; Organizational Decision Making; Organizations, Sociology of; Rational Choice and Organization Theory

Bibliography

Churchman C W, Ackoff R, Arnoff E L 957 *Introduction to Operations Research*. Wiley, New York

Hillier F, Lieberman G 2000 *Introduction to Operations Research*, 7th edn. Holden-Day, San Francisco, CA

Hillier F, Hillier M, Lieberman G 2000 *Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets*. Holden-Day, San Francisco, CA

Richardson G, Pugh A 1981 *Introduction to System Dynamics Modeling with DYNAMO*. MIT Press, Cambridge, MA

Saaty T 1959 *Mathematical Methods of Operations Research*. McGraw-Hill, New York

M. Nacht

Opinion Formation

The concept of 'public opinion' has been debated across disciplines and centuries. At issue is the nature of the 'public,' the political concepts and values that constitute 'opinion,' and the process whereby the public opinion is formed. One key point of division is epistemological: whether to view political meaning as

an aggregate or a structured whole. According to the aggregative view, political meaning consists of a set of elements (representations of experience), which are related to one another either through passively learned association or through active mental calculation. The alternative view suggests that political meaning is constructed. How it is constructed delimits the nature of the representations that may be formed and the quality of the associations or calculations that can be made.

A second divide is more explicitly political. The issue here is where to locate public opinion and, by implication, where to look in order to explain it. One view assumes the public is a composite phenomenon, an aggregate function of its individual citizens. In this view, political meaning and value are attributes of individuals. Analysis therefore focuses on the qualities of cognition and, in this light, how individuals process the experiences and social influences to which they are exposed. The opposing view presumes that the public is a collective entity, a corpus of its own. Here political meaning and value are emergent properties of the collective. They are viewed as essentially cultural phenomena and are explained with reference to large-scale social forces.

1. The Mainstream Perspective: Publics and Meaning as Aggregates

Although these debates remain central in contemporary political theory, a consensus emerged in the empirical study of public opinion formation. Its basic contours were defined by Allport's (1935) seminal discussion of attitudes, the Yale persuasion studies (Hovland et al. 1953), and Converse's (1964) analysis of political belief systems. Although pursuing different topics from different disciplinary perspectives, the research is oriented by the same political and epistemological vision: the public is an aggregate of its individual citizens, and political meaning is a matter of calculated or learned associations among representations and preferences. In this light, the formation of public opinion is understood to be a process that revolves around individuals. It begins with their exposure to politically relevant experiences and information. Each individual processes this information, thereby coming to a judgment that yields an attitude. The attitudes of different individuals are then aggregated, either through informal interactions or more formal mechanisms, such as elections or opinion polls.

For the most part, this research effort has focused on the second step in the opinion formation process, the social psychology of how individuals process information (either from direct or indirect experience) when forming an attitude. The critical outcome or dependent variable is an attitude. Typically, this is defined as an individual's overall evaluation of an object with some degree of favor or disfavor. In this view, an attitude consists of three types of elements: cognitive (beliefs); affective (emotions or feelings); and behavioral (intent to act). These elements can affect one another. While they are usually consistent, they may not be. Attitudes are also related to one another.

In a way that has important consequences for attitude strength and stability, this interattitudinal structure may vary with regard to its elaboration, its coherence, and the degree to which it has a hierarchical structure. A central aim of the research is to examine how different conditions (both situational and dispositional) affect the processing of information and thus the resulting attitude. The research is designed on the presumption that all individuals operating under similar conditions will process information in the same way. Consequently subjects' responses are aggregated, and inferences are made regarding how the 'average' person will form, maintain, or change their opinions under the conditions examined.

Adopting this general framework, a wide variety of competing theories or models of public opinion formation were presented from the 1950s to the 1980s. Some posit a rational, almost logical, integration of information. A prominent example is the value-expectancy model (e.g., Fishbein and Ajzen 1975). According to this model, an individual forms or changes an attitude by reviewing their beliefs about the attitude object. A weight is assigned to each belief by multiplying its assessed value by the probability that the belief correctly pertains to the object in question. The expected values of all the beliefs are then averaged, leading to an overall attitude or judgment being formed.

Other models accept the basic features of such rational computational models but focus more on considerations of memory and its effect on what beliefs will be considered. The suggestion here is that an individual may have a diverse set of contradictory beliefs or affects associated with an attitude object. Situational and internal cues affect which of these beliefs is likely to be remembered on any given occasion. Consequently, although individuals may process beliefs rationally, the ways in which memory functions will bias which beliefs are processed and therefore what judgment is finally made. How mass media prime candidate evaluations and the apparent instability of political opinions are readily explained in these terms (e.g., Zaller 1992). On-line processing models of political candidate evaluation de-emphasize the importance of recall (e.g., Lodge and Steenbergen 1995). In this view, individuals do not judge an object by recalling all relevant beliefs. Rather, they simply remember their prior evaluation of the object and adjust this in light of the new information provided. While this information may then be forgotten, the overall assessment is carried forward when further judgment is required.

Other models of information processing suggest that judgment is more irrational. One example is the

family of cognitive consistency models which suggest that individuals relate objects to one another, not only on categorical or logical grounds, but also on the basis of evaluation or feeling. A particularly influential extension of this is Tajfel's (1981) analysis of the consequences of group identification. His research suggests that a person's desire to maintain self-esteem (positive evaluation) and their negative identification with an out-group (categorical link) will result in negatively biased evaluations and behaviors toward that group. Following a learning theory tradition, other research demonstrates how regularly pairing an unconditioned stimulus (an object already valued positively or negatively) with a conditioned stimulus (an object not valued) will shape the subsequent evaluation of the condition stimulus.

In recent years, there has been a shift to more eclectic dual processing models of attitude formation. Self-consciously integrating earlier conceptions, these models suggest that different processes will operate under different conditions. Two types of processes are distinguished. Central processes, such as the value expectancy model, are relatively demanding and involve the evaluation of relevant arguments and information. Peripheral processes are less demanding than central processes and include classical conditioned responses, reliance on peripheral situational cues (e.g., physical attractiveness of the source), or the use of simple heuristics (e.g., experts know what they are talking about). The focus of this research is to discover (a) what conditions foster what type of processing and (b) how different kinds of processes affect attitude formation under different conditions. This has led to a reconsideration of the early research on the persuasive effect of source variables (e.g., credibility, expertise), message variables (e.g., argument quality, personal relevance), recipient variables (e.g., prior attitude, personality), and context variables (e.g., distractions, audience reaction). The net result is a more complex and contingent understanding of how attitudes are formed, sustained, and changed (Petty and Wegener 1998).

Political scientists have drawn upon these models in their attempt to understand opinion formation in mass publics. Thus they have focused on who is presenting what information (the agents of socialization—family, workplace, school, social groups, and the mass media), to whom, and with what effect on individuals' judgment and knowledge of political leaders, issues, and events. The research on agent effects on opinion formation suggest that whereas the family is not as influential as initially believed, it does provide individuals with one of their most enduring characteristics: their political identity. Supporting early survey research, contemporary work indicates that group identification can influence powerfully individuals' attitudes toward issues, especially those emphasized by the group. Research also indicates that the mass media, while not the powerful persuader it was initially feared to be, nonetheless is able to affect people's sense of political priorities and to prime the specific considerations they may bring to bear when judging a particular issue or candidate.

In the case of knowledge and sophistication, the evidence suggests overwhelmingly that the vast majority of people have little accurate information about political leaders, current legislation, or foreign affairs. Nor do their judgments reflect much sophistication or consideration. What attitudes they express do not cohere in any apparent ideological fashion, nor are they particularly reliable over time. Even slight changes in how questions are posed can produce different attitudinal responses. Some studies do show correlations between apparently abstract and more concrete attitude items. This result is interpreted to suggest people do base particular judgments on more abstract core values. However evidence from open-ended interviews indicates that whereas a few people do think about politics in this deductive way, most do not. Their responses are generally ill-considered, unjustified, and readily changed by new considerations that pop to mind. Overall, the evidence suggests that citizens lack the information, understanding, or considered judgment typically called for in democratic political theory.

2. *Alternative Direction: Constructivism, Discourse, and Collective Publics*

Questions have been raised regarding the basic assumptions orienting this dominant approach to the study of public opinion. The results of open-ended interviewing have sparked an increasing interest in taking a more constructive, holistic view of political thinking. This has lead to a renewed interest in constructivist psychologies that consider basic differences in how individuals reason (e.g., Rosenberg 2002). In this view, how an individual thinks about social or political events will determine the kinds of interrelationships and the kinds of representations (and interests) they can construct. Moreover, different individuals may think in qualitatively different ways.

For example, some people tend to focus on a simple causal or categorical relationship cued by the situation when formulating an understanding, and they draw primarily on social convention or group interests to make their judgments. Other people focus on only the concrete contours of the immediate situation and draw primarily on present feelings to make a judgment. Still others infer quite abstract issues that are only implicit in the situation and judge them with reference to some notion of the general interest or abstract principle. Contrary to the common assumption of the 'average citizen,' this work suggests that people may differ significantly, both in how they are affected by the same social influences and in their

capacity to respond to different kinds of institutional regulation.

This interest in reasoning dovetails with the dissatisfaction expressed by some political theorists regarding representative democracy and its emphasis on electoral mechanisms of preference aggregation. Instead, they advocate more deliberative forms of democracy that require individuals to cooperate in the generation and resolution of political issues (e.g., Habermas 1996). In this light, considerations of the requirements of citizenship shift. Rather than personal judgment, the focus is on cooperation with others in order to come to common understanding and judgment. This involves taking the perspective of another, relating it to one's own, and then critically considering both. The latter not only entails compromise, but the self-conscious construction of new positions that incorporate and go beyond the original ones. Research motivated by these concerns seeks to determine how differing institutional arrangements impact on deliberation and with what effect on the collective judgments that ensue. It does so by eschewing a focus on individuals in favor of a direct examination of the quality of deliberations themselves and the social representations and group opinions they produce.

Finally, there is renewed interest in a more holistic, constructivist view of the public. Rejecting what is referred to as methodological individualism, social theorists have argued that consciousness and meaning making are essentially collectively structured phenomena that are best thought of as a culture or mode of discourse. Some view culture as singly structured and coherent, and others argue that it is essentially an arena of conflict. However, all agree that the individual's understanding is a generally poor reflection of a larger social process and an inappropriate object of inquiry. In this more sociological vision, changes in public opinion reflect changes in the underlying structure of collective meaning. Examples include the shift in public discourse from appeals to the sacred to reasoned argument regarding the profane, or a population coming to imagine itself as a nation (e.g., Anderson 1991). Such changes are explained with reference to collective forces such as the form of the relations of production, the mode of information, or a new geographical positioning of a population. With such changes, alternative forms of institutional regulation become possible (e.g., rule by personal power vs. rule by law) and new kinds of political values are constructed (social tradition vs. principled critique). The challenge to this line of inquiry is to develop an empirical methodology as sophisticated as the theorizing.

See also: Attitude Change: Psychological; Attitude Formation: Function and Structure; Attitude Measurement; Attitudes and Behavior; Political Culture; Public Opinion: Microsociological Aspects; Public Opinion: Political Aspects.

Bibliography

Allport G W 1935 Attitudes. In: Murchison C A (ed.) *Handbook of Social Psychology*. Clark University Press, Worcester, MA

Anderson B R O'G 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Converse P E 1964 The nature of belief systems in mass publics. In: Apter D E (ed.) *Ideology and Discontent*. Free Press, New York

Fishbein M, Ajzen I 1975 *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA

Habermas J 1996 *Between Norms and Facts*. MIT Press, Cambridge, MA

Hovland C I, Janis I L, Kelley H H 1953 *Communication and Persuasion: Psychological Studies of Opinion Change*. Yale University Press, New Haven, CT

Lodge M, Steenbergen M 1995 The responsive voter: Campaign information and the dynamics of candidate evaluation. *American Political Science Review* **83**: 309–26

Petty R E, Wegener D T 1998 Attitude change: Mutiple roles for persuasion variables. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, Vol. 1

Rosenberg S W 2002 *The Not So Common Sense: Different Ways People Judge Social and Political Events*. Yale University Press, New Haven, CT, Chap. 8, pp. 323–90

Tajfel H 1981 *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge University Press, Cambridge, UK

Zaller J R 1992 *The Nature and Origins of Mass Opinion*. Cambridge University Press, Cambridge, UK

S. W. Rosenberg

Optimal Control Theory

Optimal control theory is a mathematical theory that deals with optimizing an objective function (or functional) in a dynamic environment. In this theory, there are two basic types of variables: state variables and control variables. The state variables describe the current state of the system, for example, a state variable may represent the stock of equipment present in the economy, the size of a forest, the amount of carbon dioxide in the atmosphere, or the biomass of a certain fish species in a given lake. The control variables are what the decision maker can directly control, such as the rate of investment, or the harvesting efforts. A system of differential equations, often called the transition equations, describe how the rate of change of each state variable at a given point of time depends on the state variables and the control variables, and possibly on time itself. Given the initial values for the state variables (often called the initial stocks), the decision maker chooses the values of the control variables. This choice determines the rates of change of the state variables, and hence their future values. The objective of the decision maker is to

maximize some performance index, such as the value of the stream of discounted profits, or some index of the level of well-being of a community, which may depend on current and future values of the control variables and of the state variables. Given the initial stocks, the objective function, and the transition equations, solving an optimal control problem means determining the time paths of all the control variables so as to maximize the objective function.

It turns out that in solving an optimal control problem, it is useful to add a new set of variables, called co-state variables: for each state variable, one assigns a corresponding co-state variable. In economics, the co-state variables are often called shadow prices, because each co-state variable is an indicator of the worth of the associated state variable to the decision maker. With this additional set of variables, one can formulate an elegant set of necessary conditions, called the maximum principle, which is extremely useful in constructing the solution for the optimal control problem, and which, as a bonus, provides deep insight into the nature of the solution.

1. The Maximum Principle

We now define a simple optimal control problem and state the maximum principle for it. Let t denote time. Let $s(t) = (s_1(t), s_2(t), \ldots, s_n(t))$ be the vector of n state variables, and $c(t) = (c_1(t), c_2(t), \ldots, c_m(t))$ be the vector of m control variables. (m can be greater than, equal to, or less than n.) The transition equations are given by

$$\frac{\mathrm{d}s_i(t)}{\mathrm{d}t} = f_i(s(t), c(t), t), \quad i = 1, 2, \ldots, n.$$

and the initial values of the state variables are $s_i(0) = s_{i0}$, $i = 1, 2, \ldots, n$. The terminal time T is fixed in this simple problem. The decision maker must find the time path of $c(t)$ to maximize the sum of (a) the value of integral of the flow of net benefit, with the benefit at time t being denoted by $u(s(t), c(t), t)$, and (b) the value of the terminal stocks, denoted by $\phi(s(T), T)$:

$$\int_0^T u(s(t), c(t), t)\,\mathrm{d}t + \phi(s(T), T).$$

(The functions $u(\cdot)$, $\phi(\cdot)$, and $f_i(\cdot)$ are exogenously specified.) At each time t, the values that $c(t)$ can take on are restricted to lie in a certain set A.

For each state variables s_i, one introduces a corresponding co-state variable π_i. Let $\pi(t) = (\pi_1(t), \pi_2(t), \ldots, \pi_n(t))$. The co-state variables must satisfy certain conditions, as set out below, and their optimal values must be determined as part of the solution of the problem. A key step in solving an optimal control problem is the construction of a function called the Hamiltonian, denoted as $H(s(t), c(t), \pi(t), t)$, and defined as

$$H(s(t), c(t), \pi(t), t) = u(s(t), c(t), t) + \sum_{i=1}^{n} \pi_i(t) f_i(s(t), c(t), t)$$

The maximum principle states that an optimal solution to the above problem is a triplet $(s(t), c(t), \pi(t))$ that satisfies the following conditions.

(a) $c(t)$ maximizes $H(s(t), c(t), \pi(t), t)$ for given $s(t)$, $\pi(t)$, and t, over the set A. (This is a very attractive property, because the Hamiltonian can be interpreted as the sum of the current benefit and an indicator of future benefits arising from current actions, as reflected in the value of the rates of change of the state variables that are influenced by the control variables.)

(b) The co-state variables and the state variables are related by the equations

$$\frac{\mathrm{d}\pi_i(t)}{\mathrm{d}t} = -\frac{\partial H}{\partial s_i(t)} \tag{1}$$

$$\frac{\mathrm{d}s_i(t)}{\mathrm{d}t} = \frac{\partial H}{\partial \pi_i(t)} \tag{2}$$

Note that Eqn. (1) has the following appealing interpretation. The co-state variable π_i being the shadow price of the state variable s_i, the left-hand side of (1) may be thought of as the capital gain (the increase in the value of that asset). The optimal level of asset holding must be such that the sum of (a) the marginal contribution of that asset, $\partial H/\partial s_i(t)$, and (b) the capital gain term, is equal to zero.

(c) $2n$ boundary conditions must be met, namely, $s_i(0) = s_{i0}$, $i = 1, 2, \ldots, n$, and

$$\pi_i(T) = \frac{\partial \phi}{\partial s_i(T)} \quad \text{for } i = 1,2, \ldots, n. \tag{3}$$

The conditions given by (3) are also called the transversality conditions. They say that at the terminal time T there must be an equality between $\pi_i(T)$ (the 'internal worth' of the stock i) and its 'external worth' at the margin (recall that the function $\phi(\cdot)$ is externally specified).

The optimal time path of the control variables can be represented either in the open-loop form, i.e., each $c_i(t)$ is expressed as a function of the time, having as parameters the initial conditions s_{0i} ($i = 1, 2, \ldots, n$), or in the closed-loop form, or feedback form, i.e., an optimal rule which determines $c(t)$ using only the knowledge of the concurrent values of the state variables, and of t.

There are more complicated optimal control problems, which may involve several initial or terminal constraints on the state variables, or the optimal

choice of the terminal time T (no longer fixed), or of the initial time (for example, at what time a mining firm should abandon a mine, a researcher should drop a research project, or a city should start building an airport). These problems give rise to a complex set of transversality conditions; see Léonard and Long (1992) for a useful tabulation.

2. Connections with the Calculus of Variations and Dynamic Programming

Optimal control theory has its origin in the classical calculus of variations, which concerns essentially the same type of optimization problems over time. However, with the maximum principle, developed by Pontryagin and his associates (see Pontryagin et al. 1962), the class of problems that optimal control theory can deal with is much more general than the class solvable by the calculus of variations.

Optimal control theory is formulated in continuous time, though there is also a discrete time version of the maximum principle; see Léonard and Long (1992). In discrete time, the most commonly used optimization technique is dynamic programming, developed by Bellman (1957), using a recursive method based on the principle of optimality. Bellman's principle of optimality can also be formulated in continuous time, leading to the Hamilton–Jacobi–Bellman equation (or HJB equation for short), which is closely related to Pontryagin's maximum principle, and is often used to provide a heuristic proof of that principle (see Arrow and Kurz 1970). Such a heuristic proof reveals that the shadow price (i.e., the co-state variable) of a given state variable is in fact the derivative of the value function (obtained in solving the HJB equation) with respect to that state variable, thus justifying the name shadow price. This approach yields equations that are in agreement with (b) in the preceding section.

3. Applications of Optimal Control Theory to Economics, Finance, and Management Sciences

Economists initially made use of optimal control theory primarily to address the problem of optimal capital accumulation and optimal management of natural resources and the environment (Arrow and Kurz 1970, Hadley and Kemp 1971, Pitchford and Turnovsky 1977, Kemp and Long 1980). Optimal control theory has been used to study a variety of topics, such as macroeconomic impacts of fiscal policies in a well-functioning market economy (for example, the effect of a temporary increase in government spending on private consumption and investment), optimal consumption planning by an individual (for example, Becker and Murphy (1988) addressed the question of rational addiction: if you know that your current level of consumption of a drug will determine your degree of addiction which in turn will increase your future demand for it, how should you determine your time path of consumption of the drug, knowing the consequence of addiction?).

In management science, optimal control theory has been applied to management problems such as the optimal holding of inventories, the optimal time path of advertising (e.g., under what conditions would it be optimal to vary the intensity of advertising over the product cycle), the optimal maintenance and sale age for a machine subject to failure, and so on. In finance, theories of optimal portfolio adjustments over time, and theories on the pricing of financial options, have been developed using optimal control methods, in a stochastic setting. (See Malliaris and Brock 1982).

It is important to note that while the variable t often stands for time, and the objective function is often an integral over a time interval, it is not necessary to restrict the theory to that interpretation. In fact, in some important applications of optimal control theory to economics, t has been interpreted as an index of the type of an individual in a model where there is a continuum of types.

4. Extensions of Optimal Control Theory

There are two important extensions of optimal control theory. The first extension consists of allowing for uncertainty in the transition equations, leading to the development of stochastic optimal control theory. The second extension is to allow for strategic interactions among decision makers who pursue different objectives, or who have diametrically opposed interests. The body of literature which deals with this extension is called differential game theory.

4.1 Stochastic Optimal Control Theory

A typical transition equation in stochastic optimal control theory is

$$\mathrm{d}s = f(s(t), c(t))\,\mathrm{d}t + g(s(t), c(t))\,\mathrm{d}z$$

where $\mathrm{d}z$ represents an increment of a stochastic process z called the Wiener process. In this equation, $f(s(t), c(t))$ represents the expected rate of change in the state variable s, the term $\mathrm{d}z$ represents a random disturbance in the external environment, and $g(s(t), c(t))$ indicates that the effect of that disturbance on the evolution of the state variable may depend on the values taken by the control variable and the state variable.

In some contexts, it is more appropriate to consider random disturbances that occur only rarely, but that may have great impacts when they occur, for example, an earthquake, a currency crisis, or a discovery of an entirely new production technique. To deal with such discrete events, the Poisson process (also called jump process) is used instead of the Wiener process, which is suitable for continuous random variations.

In finding a solution of a stochastic optimal control problem, one makes use of a stochastic version of the Hamilton–Jacobi–Bellman equation. Sometimes it is possible to provide an explicit solution in the feedback form. For example, some adjustment rules for portfolios of financial assets can be derived. Another interesting application of stochastic optimal control is the problem of determining the optimal time to switch from one regime to another. This is often referred to as the optimal stopping problem. One asks, for example, when one should exercise a call option (stop waiting for a further increase in price), or when to start setting up a factory in a foreign country (stop waiting for a further favorable movement of the exchange rate).

4.2 *Differential Games*

An example of a differential game can be given involving a transboundary fishery: two nations have access to a common fish stock. Each nation knows that the other nation's harvesting strategy will have an impact on the fish stock. If the two nations do not cooperate, they are said to be players in a non-cooperative game. Each nation would be solving an optimal control problem, taking as given the strategy of the other nation. In trying to predict the outcome of such non-cooperative games, theorists have formulated various concepts of equilibrium. A Nash equilibrium of a differential game is a pair of strategies, one for each player, such that each player's strategy is the best course of action (i.e., it solves the player's optimal control problem), given the strategy of the other player.

The nature of the solution depends on the set of strategies available to the players. If one assumes that, given the initial conditions, players can only choose control variables as functions of time (and not as functions of the concurrent values of the state variables), then the equilibrium is called an open-loop Nash equilibrium. On the other hand, if players must choose strategies that specify the values to be taken by the control variables at any time as functions of the concurrent values of the state variables, and possibly of time, then the equilibrium is said to be a feedback Nash equilibrium. A refinement of the feedback Nash equilibrium requires that at any point of time, and given any situation the system may reach, the rule chosen by each player must remain optimal for that player, given that other players do not change their rules. This property is called *subgame perfectness*. A feedback Nash equilibrium that satisfies the subgame perfectness property is called a Markov Perfect Nash equilibrium.

Both deterministic and stochastic differential games have been applied to problems in economics and management science. Examples include games of transboundary fishery, advertising games, innovation games, and so on (see Clemhout and Wan 1994, and Dockner et al. 2000).

5. *Conclusions and Future Directions*

Optimal control theory has become a standard tool in economics and management science. It has facilitated the solution of many dynamic economic models, and helped to provide a great deal of insight into rational behavior and interactions in a temporal context. Several unsolved problems are currently attracting research interests. Among these are (a) finding Markov-perfect equilibria for general dynamic games with a hierarchical structure (some players have priority of moves), and (b) the search for optimal policy rules that are dynamically consistent in an environment where economic agents are rational forecasters.

See also: Decision-making Systems: Personal and Collective; Dynamic Decision Making; Game Theory; Self-organizing Dynamical Systems

Bibliography

Arrow K J, Kurz M 1970 *Public Investment, the Rate of Return, and Optimal Fiscal Policy*. Johns Hopkins University Press, Baltimore

Becker G S, Murphy K M 1988 A theory of rational addiction. *Journal of Political Economy* **96**(4): 675–700

Bellman R 1957 *Dynamic Programming*. Princeton University Press, Princeton, New Jersey

Clemhout S, Wan H Jr. 1994 Differential games: Economic applications. In: Aumann R, Hart S (eds.) *Handbook of Game Theory, with Economic Applications*. North-Holland, Amsterdam, Chap. 23

Dockner E, Jorgensen S, Long N V, Sorger G 2000 *Differential Games in Economics and Management Science*. Cambridge University Press, Cambridge, UK

Hadley G, Kemp M C 1971 *Variational Methods in Economics*. North-Holland, Amsterdam

Kemp M C, Long N V 1980 *Exhaustible Resources, Optimality, and Trade*. North-Holland, Amsterdam

Léonard D, Long N V 1992 *Optimal Control Theory and Static Optimization in Economics*. Cambridge University Press, Cambridge, UK

Malliaris A G, Brock W A 1982 *Stochastic Methods in Economics and Finance*. North-Holland, Amsterdam

Pitchford J D, Turnovsky S J 1977 *Applications of Control Theory to Economic Analysis*. North-Holland, Amsterdam

Pontryagin L S, Boltyanski V G, Gamkrelidze R V, Mishchenko E F 1962 *The Mathematical Theory of Optimal Processes*. Wiley, New York

N. Van Long

Optimization in Evolution, Limitations of

In biological evolution, natural selection is often seen as an optimizing force: an organism's adaptations are optimal in performing their function because selection has molded them over long periods of time. Indivi-

duals with better (i.e., closer to optimal) features have more offspring, which inherit these improvements. More critical scrutiny of adaptations, however, reveals that many are demonstrably suboptimal. In this article, the way natural selection works in natural populations is briefly outlined; a discussion of several reasons why the fitness-maximizing action of selection is limited then follows, and the article concludes with a short review of the consequences for evolutionary biology brought about by limitations of optimization.

1. *Natural Selection*

Some organisms, by virtue of the interaction of the environment with aspects of their morphology, behavior, physiology or other phenotypic features, leave more offspring than others with different phenotypes. Hence, on average, these fitter individuals, which are said to have a high fitness, contribute more of their genetic material to the next generation than the less fertile types. If there is some heritable component to these phenotypic differences, then the proportion of offspring exhibiting the fitter phenotype will increase. This process, continuing generation after generation, ensures that the population matches its environment and, eventually, this match can become optimal, in the sense that no better phenotype can occur.

Nevertheless, optimization may not occur, because of biological processes and properties at any of the stages described. In order to understand why, it is necessary to consider just what is being optimized. As described above, natural selection acts at the level of individual organisms. While such organismal-level selection clearly has an effect on the population, it need not maximize any property of that population. In addition, selection can (and does) act at different levels (e.g., genes, chromosomes, cells and populations), in addition to the traditional, Darwinian, individual level.

2. *Reasons for Nonoptimality*

Natural selection may fail to optimize for several reasons. Some of the causes of nonoptimality do, of course, fall under more than one of the headings.

2.1 *Organisms are Variable*

Ironically, the within-population variability, on which natural selection acts to produce a good match of an organism's phenotype to its environment, may also prevent this fit becoming perfect. Even if selection has optimized a population's mean phenotype, individuals within that population will exhibit slightly different phenotypes, some of which may be suboptimal. The phenotypic variation in this optimized population may reflect environmental differences, over either time or space, during the various individuals' development, or genetic differences among individuals, or it may be a manifestation of the interaction of environmental and genetic differences. The effect of genetic differences is known as genetic load, the difference in fitness between the best possible phenotype and the average fitness of the population. The most important form of load, segregational load, is caused by genetic variants that are found at non-negligible frequencies in the population; such variants are said to segregate in the population. The paradigmatic case is that of human sickle-cell anemia in central Africa, where two alleles, Hb^A and Hb^S, are found at the β-globin locus involved in the disease. In spite of having mild anemia, heterozygous individuals have the highest fitness because of their enhanced resistance to malaria, and selection acts to maintain both alleles in the population. An inevitable consequence of Mendelian inheritance, however, is that both sorts of homozygotes occur in each generation and these genotypes produce phenotypes with lower fitnesses, $Hb^A Hb^A$ individuals being susceptible to malaria and $Hb^S Hb^S$ individuals dying early from severe anemia.

A second form of load is mutational load, effected by recurrent deleterious mutations concurrently being eliminated by purifying selection. Because mutation is a rare event, there are usually very few highly deleterious mutants in a population and mutational load is less important as an explanation for a lack of optimality than segregational load.

In the absence of environmental or genetic differences, phenotypic variation can still arise because of random events during the development of an individual (developmental noise) that lead to differences among individual phenotypes. Hence, even though selection may act to maximize the mean fitness of a population, many individuals may be suboptimal. Of course, if the amount of phenotypic variation due to developmental noise is large and a significant number of individuals are far from the optimum, selection can be expected to act to reduce the level of noise.

2.2 *Evolution is Contingent*

Natural selection may also fail to optimize because the right sort of phenotypic variation is absent. Evolution is a contingent process; selection can only act on phenotypic variation that is already present in the population. It may be energetically more efficient to have wheels than legs, but if all individuals always have legs, the population will never optimize its energy efficiency. Of course, the phenotypic variation need not initially encompass the optimum. Selection can move a population towards it gradually, via a series of minor improvements, with new phenotypes arising each generation through mutational changes in the

gene pool and, especially, genetic recombination. The availability of small phenotypic differences on which selection can act to produce optimality can also be viewed as a developmental constraint (see Sect. 2.9).

A second aspect of contingency is that selection is blind, always acting to move a population towards the nearest local optimum, rather than towards a global optimum. In the 'legs or wheels' example above, this blindness would mean that selection would act to improve the efficiency of legs, rather than start afresh with imperfect, less efficient 'proto-wheels' and then optimizing.

Different authorities put different emphases on the importance of contingency in evolution. Some parts of evolutionary game theory (see Dugatkin and Reeve 1998, for recent work in this field), for example, stress that over time a vast range of mutations can occur and so the first aspect of contingency is less relevant. Even the second aspect can be avoided if a broad enough spectrum of mutation can occur.

2.3 *Variation may not be Heritable*

If there is no heritable component to the phenotypic variation, natural selection cannot move the population towards the optimum phenotype. Nevertheless, the vast majority of traits examined (e.g., by carrying out selection experiments on populations of a plant or animal species) do show some degree of heritability. Hence, this explanation for nonoptimality is likely to be relatively unimportant. Note, however, that a population subject to segregational load will display a lack of heritability for that trait, even though there is genetic variation underlying the phenotypic variation. Of course, if the selection pressures on that population change, then heritability will cease to be zero and the population will respond to this new selective regime. In the case of sickle-cell anemia, for instance, the absence of malaria in North America has led to a decline in the frequency of the Hb^s allele in the US Afro-American population.

Non-heritable variation can also aid optimization. Phenotypic plasticity is the ability of a particular genotype to develop differently in different environments. Crucially, the different phenotypes produced are appropriate (maybe even optimal) in the different environments. For instance, when grown in hot, sunny conditions, some plants reduce exposure to the sun, thereby lowering leaf temperature, by orienting their leaves closer to the vertical. Appropriate match to the environment is thereby achieved in a variety of habitats.

2.4 *Trade-offs Among Different Traits*

An important reason why aspects of a phenotype are not optimal is a consequence of the logical impossibility of maximizing two or more functions simultaneously. Traits may have contradictory requirements and the evolutionary consequence is a trade-off between these conflicting demands. For example, the greater a bone's weight, the stronger it is. But heavier bones are more expensive to produce and carry around, so the weight is a balance between the fitness benefits of strength and lightness. Neither the strength nor the weight of the bone is optimal. Nevertheless, this sort of trade-off can still be viewed as resulting in optimality: fitness rather than weight or strength is optimized. This view of optimality, however, reduces the applicability of Darwin's natural selection to explain why traits are the way they are because many traits must be considered simultaneously.

A second sort of trade-off arises because of the way most, if not all, genes affect several traits in an organism, a property known as pleiotropy. Traits simultaneously affected by variation at one or more loci are said to be 'genetically correlated.' If substituting an allele at a pleiotropic locus has opposite consequences for different components of fitness, the result is an evolutionary trade-off. One of the best examples of antagonistic pleiotropy (as this phenomenon is known) is that of early and late fecundity in *Drosophila* fruit flies. Flies bearing alleles that lead to greater numbers of offspring early in life have fewer offspring as they age and vice versa; no flies have the ideal maximal fecundity at all ages. But, as above, this sort of trade-off still results in fitness (in this case, lifetime fertility) being optimized.

The concept of trade-offs illustrates that it is vital to know just what is being optimized. Selection, at least that acting at the level of the individual, seeks to optimize an organism's overall fitness, not particular fitness components (e.g., early fecundity). Explanations invoking selection have been criticized for breaking up organisms, which may be treated better as integrated wholes, into separate traits, each of which is considered to be optimal (for example, Gould and Lewontin 1979). Nevertheless, some degree of 'atomization' is necessary for scientific enquiry, to avoid a stultifying holism that prevents asking any questions about aspects of an organism. We must not be prevented from studying a bird's wing simply because it is but part of a bird. Moreover, many traits must be at least 'quasi-independent' to give even the semblance of optimality seen in natural populations (Lewontin 1985).

2.5 *Selection is not Constant*

Selection pressures are temporally and spatially variable as a consequence of changes in the environment and the genetic constitution of the population itself. Seasonal variation is an obvious source of environmental change that may result in changes in the

fitnesses of different phenotypes. The annual cycling of frequencies of different chromosomal arrangements (called inversions) in some *Drosophila* populations appears to be an example of seasonal fitness differences. Hence, selection may be continually molding a population, never quite reaching optimality.

Unique events too (such as the arrival of a novel predator or disease) can also change a selective regime, although in the long term selection would be expected to move the population towards optimality. If the environment changes frequently, however, then selection will not have sufficient time to do so. The important issue is just how quickly the environment does change.

Different populations live in different local environments. In the absence of migration between populations, each can become locally adapted. Gene flow, the genetic exchange brought about by migration, inhibits this local process, however, and no population is absolutely optimal. Gene flow is invoked to explain the failure of most species to expand their range continually. Populations at the edge of the distribution cannot adapt to the conditions there (let alone expand to colonize new habitat) and should be demonstrably less than optimal.

Fitnesses may also depend on the frequency of different phenotypes, genotypes or alleles in the population. Such frequency-dependent selection often arises when unusual types have some sort of advantage by virtue of their rarity. In predator–prey situations, for instance, a rare form of the prey may be overlooked because it does not fit the predator's search image. Similarly, fitnesses may be density-dependent, affected by the population's size. These sorts of selection can drive continual genetic change in a population, with optimality never being reached. Even when frequency-dependent or density-dependent selection moves a population to a genetic equilibrium, this point need not be (and usually is not) an optimum.

2.6 Other Processes Interfere with Selection

Natural selection is not the only process effecting evolutionary change in natural populations. Each generation, the genetic recombination that occurs as part of sexual reproduction breaks up assemblages of alleles, genes and chromosomes that have survived the filtering action of selection and then re-forms less fit types. The sickle-cell anemia example demonstrates this phenomenon: the optimal heterozygosity regenerates suboptimal homozygotes through sexual reproduction every generation. A second example is the reassortment of genes that may be advantageous in certain combinations by chromosomal recombination during meiosis.

At the level of the population, genetic drift, a consequence of the finiteness of populations, randomly alters allele frequencies every generation. The size of these changes is inversely proportional to the size of the population: in large populations, the effect of drift is negligible, whereas, in small populations, it is often the major agent of change. Obviously, such random changes can move a population away from optimality, although just how far away would depend on the strength of selection for the optimum. But genetic drift can also enhance optimization by enabling a population to escape a local optimum so that selection can move it to another, possibly global, optimum.

Migration and mutation can also prevent optimization, as mentioned above. Nevertheless, mutation is also essential to optimization, providing the raw variation on which selection acts to optimize. Migration from other populations, too, can bring in novel variation on which selection can act.

Nonrandom mating, the occurrence of mating pairs at frequencies different from that expected by chance, can also interfere with the action of selection. Indeed, a population's mean fitness decreases in some simple models of one common form of nonrandom mating, selective mating, in which mates are more (or less) phenotypically or genetically similar for a particular character or gene.

2.7 Selection Need not Optimize

Perhaps the most counterintuitive reason for optimization being limited is that numerous selective regimes do not maximize fitness. While most verbal descriptions of natural selection (including Darwin's) emphasize the optimizing aspect of fitness, mathematical models show that the conditions under which fitness reaches a maximum (even a local maximum) are rather restrictive. Ignoring the effects of genetic drift, migration, mutation and nonrandom mating as described above, constant selection, arising from viability differences among different phenotypes reflecting genetic variation at a single locus, will maximize the mean fitness of the population. If selection pressures vary in any way, if selection arises because of fertility or fecundity differences or if more than one locus is involved, the population's mean fitness need not be maximized. It is even possible to construct simple mathematical models under which mean fitness is minimized. A straightforward verbal scenario (quoted in Gould and Lewontin 1979, p. 592) in which selection acts but does not improve mean fitness concerns an allele for increased fecundity, which selection rapidly fixes in the population. If the population is otherwise unchanged, no more adults are likely to survive than previously, and the size of the population is unchanged (and may even decrease if predators key into the more abundant juveniles). Neither individuals nor the population is optimized.

The non-optimizing property of fertility selection is a reflection of its formal parallels with selective mating: the number of offspring ('fitness') is a property of the mated pair, 'fertility' for fertility selection and 'the

chance of mating' for selective mating. Mean fitness can decrease in both sorts of systems. Nevertheless, it can be argued that the sorts of fitnesses that lead to such non-intuitive behavior are not often found in nature. This rarity may arise either because genetic variants bearing such fitnesses arise infrequently (a form of mutational bias) or because they are weeded out by selection in the longer term.

Systems with two or more loci do not optimize mean fitness because of genetic linkage and the non-independence of the effects of genes at the different loci (a property known as epistasis). The recombinatory effects of sexual reproduction prevent the population's mean fitness being maximized. But, again, a case can be made that alleles with strong epistatic interactions are unusual and, as a result, genes (as well as traits) are selectively quasi-independent.

2.8 Conflicts Among Different Levels of Selection

The above account describes selection acting at the level of the individual. But, at least in principle, selection can act at several levels simultaneously and, as a consequence, an optimum is not reached at one or both levels. The so-called *t* allele system (now known to involve several loci locked in a chromosomal rearrangement or inversion) in mice provides a paradigmatic example. Heterozygous males produce about 95 percent *t* sperm, rather than the 50 percent expected under standard Mendelian segregation. This difference is an example of selection at the level of the gamete, favoring the *t* allele: it is as if *t*-bearing sperm win the race to fertilize an egg more often than these males' wild-type sperm. But at the level of the organism there is strong selection against the *t*-allele, since homozygous males are sterile. Finally, there is selection at the population level reducing the frequency of *t*. Mouse populations are often small and genetic drift, together with the gametic-level selection, can occasionally result in a population having only *tt* males. Such populations then become extinct because of the sterility of these males.

Conflicts among levels of selection can be viewed in a similar way to conflicts among selection on different traits: selection is still optimizing some overall property of the system, a balance between action at the different levels. It should also be noted that the efficacy of selection at higher levels, say, populations, is less than that at lower levels, say, individuals, because population colonization and extinction rates are lower than individual birth and death rates. Nevertheless, there are clearly limits on the optimizing power of selection at a particular level.

2.9 Developmental Constraints

The way in which the developmental system of an organism is structured and operates can constrain the phenotypic response to optimizing selection. Developmental constraints can be divided into two categories (Maynard Smith et al. 1985), 'universal' and 'local,' although it is not always easy to distinguish between them. Universal constraints are those applying to all organisms because, for instance, they are a consequence of the laws of physics. The above-mentioned trade-off between the weight and strength of a bone is an example of a universal developmental constraint.

Local constraints, by contrast, pertain to a particular group of organisms: can an organism's developmental system produce a phenotype that selection will use as a step towards optimality? For example, Maynard Smith et al. (1985) argue that the failure to evolve a process of secondary thickening forces palm trees to be unbranched. Such constraints could also be described as phylogenetic or historical and are a manifestation of the contingency of evolution (see Sect. 2.2). Once evolution has started down a particular pathway, certain developments are precluded. At the origin of life, for example, the choice (for whatever reason) of left-handed amino acid isomers has restricted subsequent evolution to this option. Local developmental constraints may be overcome by selection: some species of palm trees do have branches, for instance. The importance of such constraints clearly depends on their strength, which can be difficult to gauge.

3. Consequences for Evolutionary Biology

Deciding whether optimization has been achieved is difficult. At one extreme, suboptimality is acknowledged, but explained away by invoking trade-offs, developmental constraints, or chance. Everything is considered to be as good as possible, given the circumstances, a position Gould and Lewontin (1979) pilloried as the Panglossian Paradigm. Williams (1966) has suggested that natural selection be invoked as an explanation for a trait only as a last resort. Consequently, a trait need not be considered optimal and the initial focus, at least, is on nonselective processes that may have led to its evolution. Mayr (1983), however, argued that there was nothing wrong in searching for a selective explanation for a phenotype, provided we acknowledge that evolution need not always optimize. Indeed this 'adaptationist program' has provided biology with its most powerful explanatory methodology: Krebs and Davies (1997), for instance, exemplify the productive use of this approach in the study of behavior. Moreover, Seger and Stubblefield (1996) demonstrate that using an optimization approach to study a biological phenomenon can expose sub-optimality. Nevertheless, for all the above reasons, it should be remembered that the traditional heuristic interpretation of the action of selection can be misleading. Consequently, optimization will often be limited.

See also: Adaptation, Fitness, and Evolution; Body, Evolution of; Brain, Evolution of; Darwin, Charles Robert (1809–82); Darwinism; Darwinism: Social; Evolution: Optimization; Genes and Culture, Coevolution of; Genotype and Phenotype; Intelligence, Genetics of: Heritability and Causation; Natural Selection

Bibliography

Dugatkin L A, Reeve H K (eds.) 1998 *Game Theory and Animal Behavior*. Oxford University Press, New York

Gould S J, Lewontin R C 1979 The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London Series B, Biological Sciences* **205**: 581–98

Krebs J R, Davies N B (eds.) 1997 *Behavioural Ecology: An Evolutionary Approach*, 4th edn. Blackwell Science, Oxford

Lewontin R C 1985 Adaptation. In: Levins R, Lewontin R C (eds.) *The Dialectical Biologist*. Harvard University Press, Cambridge, MA

Maynard Smith J, Burian R, Kauffman S, Alberch P, Campbell J, Goodwin B, Lande R, Raup D, Wolpert L 1985 Developmental constraints and evolution. *Quarterly Review of Biology* **60**: 265–87

Mayr E 1983 How to carry out the adaptationist program? *American Naturalist* **121**: 324–34

Seger J, Stubblefield J W 1996 Optimization and adaptation. In: Rose M R, Lauder G V (eds.) *Adaptation*. Academic Press, San Diego, CA

Williams G C 1966 *Adaptation and Natural Selection*. Princeton University Press, Princeton, NJ

H. G. Spencer

Oral and Literate Culture

The phrase 'oral and literate expression' refers to verbalized formulations communicated through speech and writing respectively. 'Oral' and 'literate' are often seen as contrasted or complementary (hence their joint treatment in this article), but the relations between them, even their juxtaposition, are matters of debate. Recent research tends to focus on their overlap, taking them as dimensions of a multiplex and socially mediated continuum. This article starts with a short overview of their main manifestations, then considers issues of definition and approach, and recent research trends.

1. Spoken and Written Expression in Human Culture

1.1 Expression in Writing

Through the ages humans have developed many forms of writing: systems for communicating relatively specific ideas through standardized and relatively permanent visible marks, conventionally ordered and interpreted. Many systems are based on close relations between written marks and speech sounds, conveyed through syllabic, consonantal or, most familiar to Western readers, alphabetic symbols (phonographic systems). Others, such as the Chinese and Aztek systems, use visual signs, sometimes pictorially based, to represent meanings rather than sounds (variously termed logographic, ideographic, and/or semasiographic systems); these allow communication across language boundaries, notably in the far-spread Chinese script. Some scholars include within writing pictographic systems such as American hobo signs or Native American pictorial records of events; others prefer narrower definitions like 'visible speech.' The classifications and boundaries of writing are thus not fully agreed, partly because most systems combine several principles and the relations between speech sounds and written signs are variable (see Boone and Mignolo 1994, Daniels and Bright 1996, see also *Writing, Evolution of*).

Writing appears in many different materials and forms, versatile enough to be turned to almost any of the multifarious purposes of human interaction. These include encapsulating, storing, and transmitting information; counting and accounting (sometimes seen as the origin of writing); formulating and extending memory; bridging space and time; and facilitating personal and artistic expression. Writing is also used in the formulation of recognized literary and/or religious works and their communication within and across human cultures. Western Europe, China, and the Islamic world are famous examples but throughout history and nowadays in every area of the globe writing has played powerful roles—for good or ill—in large-scale organizations, in personal and group expression at every level, and in the human endeavor to transcend space and time. It is inextricably embedded in human social institutions; for example, in the educational, artistic, and organizational hierarchies of the modern West.

1.2 Oral Expression

Scholars have long recognized the significance of written expression, but the rich complexities of oral forms are now increasingly understood too. Their nature and functioning are not necessarily identical to those of writing. In many disciplines, from philosophy and sociolinguistics to anthropology and education, the focus has shifted away from treating oral forms as if written texts toward active speech processes, conversations, and discourse-in-context. As with written genres, so too there are oral (or speech) genres: for example, different expected styles for political speeches, spoken prayers, or personal greetings. Oral expression uses varying permutations of human voices (speaking, singing, chanting, shouting…), singly or

combined, with generic conventions based on, for example, verbal and performance style, appropriate speaker(s) and audiences, context, or purpose. These are deployed in manifold ways in innumerable formal and informal situations, from intimate interchanges to public rituals. Many oral genres typically involve immediately present speakers and transient performances (unlike the greater permanence of writing), but some are now also enacted over far distances through electrical and electronic technologies or stored and reproduced in audio recordings. Like writing, they are both interwoven with human organization and creatively used by individuals for an infinite number of purposes (see *Linguistic Anthropology*; *Sociolinguistics*).

Certain oral genres are singled out for heightened attention. Far from 'nonliterate' formulations being crude, communal, or uncreative, as once supposed, they can be extremely sophisticated in both verbal exposition and performance. Examples include the magnificent oral artistries of twentieth-century South Slavic epics, the lofty diction of Zulu panegyrics, brilliant imagery of Somali 'miniature' lyrics, deeply felt 'wept thoughts' of some South Pacific laments, or contemporary performance poetry. Oral genres such as narratives, myths, epics, eulogies, protest poetry, laments, love poems, oratory, and song-texts (themselves a form of lyric poetry) are variously found throughout the world, as well as condensed but sometimes highly artful proverbs, jokes, anecdotes, or after-dinner tales. These are often described as 'oral literature' and although the term is not wholly satisfactory (underplaying performance dimensions, for example) it usefully highlights their literary qualities; a few scholars prefer the neologism 'orature.' Their study has enabled a radical reassessment among many scholars of the achievements of oral expression, not least in the lyrics of popular music and in cultures of Africa, Native America, or Oceania once regarded as without creative and individually shaped verbal artistry (for examples and discussion, see Feld 1990, Finnegan 1992b, Foley 1995, Okpewho 1992, Sherzer and Woodbury 1987, Tedlock 1983).

2. *The Concepts 'Oral' and 'Literate': Problems and Approaches*

The distinction between oral and literate expression seems at first self-evident, familiar from phrases like 'the spoken and written word.' However 'oral' and 'literate' are complex and multiform concepts, in both their varying cultural manifestations and the theoretical controversies surrounding them. They are all the more loaded because they invoke emotive and far-reaching issues about the nature and history of human culture.

The terms themselves are awkward in English. 'Oral' mostly refers to what is spoken, uttered by the mouth, but also sometimes to anything nonwritten: the first sense excludes deaf signing and nonverbalized material culture, the second includes them; phrases like 'oral tradition' can thus be ambiguous. The term 'literate' partly complements 'oral' but is not precisely parallel and is sometimes replaced by 'written.' While it can readily be assumed that all members of a society normally gain some capacity to speak and hear, the skills of reading and (not necessarily the same) of writing are less widely distributed.

The referents and connotations of each term have also been channeled through a series of theoretical frameworks and questions. Many of these broadly match those elsewhere in the social sciences, successively dominated by, for example, evolutionist, diffusionist, and functionalist analyses, ideas of reflection, structuralism and poststructuralism, and a variety of Marxist-inspired and cognitivist approaches. Each has it own fix on the terms and their interrelationships. Here a few are selected to illustrate the changing perspectives.

2.1 *Romantic and Evolutionary Paradigms*

Some interest in contrasts between spoken and written expression dates back at least to classical Greece but in more recent centuries the most influential model has posited a teleological progression in human history: from oral to written. Spoken and written forms are viewed as themselves distinct and as characterizing distinctive evolutionary stages, with writing destined to supersede the oral phase in the upward path to enlightened modernity. Complementing and partly overlapping this is a nostalgic attachment to what was sacrificed in the process: a romanticizing glorification of that lost 'other.' Thus, so-called 'oral cultures' are viewed ambivalently, both as the valued repository of deep-seated, communal, and emotion-sensitive qualities and as a primitive stage which modern peoples have rightly left behind.

This intersecting cluster of ideas has remained remarkably pervasive. It presents a rousing story of the proud—if dearly bought—history of humanity, culminating in the achievements of Western civilization and its deployment of alphabetic writing. The 'oral' and the 'modern' are pictured as essentially contrasting and sequential phases of human development.

2.2 *Consequences of Literacy and Orality*

The *consequences* of oral and literate forms of expression attracted much scholarly and popular interest through the mid- and late twentieth century. The British anthropologist Jack Goody argued from the 1960s that the functions and effects of literacy should be studied comparatively. Parallel questions were

raised in history, sociology, and comparative literary studies, further reinforced by interests in development, economic 'modernization,' educational applications, and the major comparative features of societal organization. Marshall McLuhan's writings on the media stirred popular imagination, intermixed with continuing evolutionary and romantic perspectives. The result was intensive discussion about the social and cognitive consequences of literacy. Among those suggested were: rationality; objectivity; the emergence of history, science, and/or philosophy; secularism (but in other accounts religious obscurantism); political and religious control (or, alternatively, freedom); 'modernization'; the 'take-off' for economic development and/or imperial rule both in early history and the present; innovation; transition from auditory to visual perception; bureaucracy; and individualism. The suggested consequences between them spanned almost every feature associated either ideally or in practice with post-Enlightenment Western society, and chimed well with narratives of the print-based Protestant Reformation and the expansionist West disseminating literacy through its civilizing mission.

Most of this presupposed some independent variable from which these consequences flowed. Usually this was literacy, with orality implicitly its more shadowy opposite. However, just what this literacy comprised was less clear. Some of the posited effects occurred in some eras but not others or were even mutually contradictory (e.g., literacy as leading to freedom *and* to oppression) so could scarcely be automatic results from some uniform entity. As in other areas of the social sciences, many analysts thus now speak more of 'implications' than unicausal 'consequences,' and of possibilism rather than determinism, recognizing that literacy takes many forms and is used differently in different situations.

2.3 '*The Orality–Literacy Debate*'

Linked to the interest in consequences was the so-called 'orality–literacy debate.' This revolved round the question of whether there is some generalizable divide between oral and literate cultures, and/or between oral and literate minds.

On one side were writers like McLuhan (1964) and Ong (1982 and earlier) who not only strongly contrasted 'literacy' and 'orality' but often envisaged human history as a series of major changes—from orality, to writing, to print. Each stage had distinctive properties, with the 'Great Divide' between 'primitive' and 'civilized' marked by literacy. This resonated with evolutionary and romantic notions, carrying connotations of orality as primeval, emotional, and homogeneous, characterized by unchanging and communal tradition rather than individual creativity, as against the rationality, individualism, science, urbanism, modernity, and economic development associated with literacy. This in turn recalled the binary division between types of society envisaged in classic nineteenth and early twentieth century social theories (e.g. see *Durkheim, Emile (1858–1917)*).

Other scholars argued that the basic question was over-general, giving a spurious identity to multiplex notions (see Boyarin 1993, Schousboe and Larsen 1989, Street 1993). They challenged the ethnocentric (Western-based) bias of the contrast and its implicit technological determinism. In their view literacy and orality are neither autonomous nor unitary. The practices of speaking and writing coexist, they are socially as much as technologically shaped, and cultural variations undermine sweeping generalizations.

A further twist comes in McLuhan's and Ong's vision of electrical and electronic technologies as ushering in a new age of 'secondary orality.' This further links to suggestions that the apparent bias towards vision in Western cultures, often associated particularly with Gutenberg's printing press, is now being redressed by new opportunities for audio-communication, hence a greater role, once again, for oral expression. The position is complicated, not least because others propound a new 'visual turn,' but computer technologies certainly open new opportunities. They also challenge older assumptions about oral forms being fleeting, evanescent, and essentially variable, in contrast to the fixed, permanent, and authoritative properties of writing. Computer 'soft' text transcends that postulated divide and erodes yet further the once-confident distinctions.

The trend is now to take more cautious account of diversity and change rather than generalizing about 'orality' or 'literacy' as supposedly uniform entities. But the question is constantly being recycled, sometimes by writers unacquainted with earlier work, and still stirs both academic and popular interest.

2.4 *Ethnographic and Historical Specificities*

Many scholars now look instead toward empirical investigation of specific cases. Here 'oral' and 'literate' turn out to be less hard-edged and autonomous than once assumed and to come in varied, culturally specific, forms. Speaking and writing intermingle in manifold genres rather than being existent in monolithic and clearly separated form. The famous Greek epics of Homer, for example, were probably first composed and performed orally, but were then transmitted in writing. Oral and written traditions are not independent streams, as once supposed, for the 'same' works sometimes appear in both modes and move between them. Examples are national anthems, carols, well-known religious texts, 'urban myths,' Hausa political poetry, and English-American ballads.

In classical and medieval Europe literary works were commonly composed for the voice and 'pub-

lished' by being read or performed aloud to an audience; comparable interplays of oral and written processes occur in sermons, plays, or conference 'papers' today. Song lyricists have often used mixes of written and oral media to compose sung words and guide their performance. Similarly, written formulations are related to oral practices in a huge variety of ways, from dictated transcriptions and performance scores to crib sheets, prompt books, or jotted notes. Radio/television broadcasts and the icons, graphics, and audio messages of modern computers further blur the once-clear divisions (for discussion and illustration see Boone and Mignolo 1994, Boyarin 1993, Cavallo and Chartier 1995, Finnegan 1992a, Foley 1995).

The notions of oral vs. literate thus no longer appear as bounded and clearly opposed, far less as carrying already known attributes when applied to cultures, genres, or individuals whether now or in the past. Rather, they emerge as broad and somewhat ill-defined terms covering a spectrum of varying practices and artifacts, resources which human beings draw on in a multitude of overlapping, merging, and intermingled ways.

3. *Recent Insights and Future Directions*

While earlier approaches still carry influence, current research could be said to be characterized by three interrelated trends.

3.1 *Multiplicity and Differentiation*

Recent analyses continue and extend the interest in historically and culturally specific contexts. Scholars now seldom speak of purely 'oral cultures,' recognizing not only the lengthy and geographically widespread interconnections between oral and graphic systems but also the complex and shifting varieties of spoken/written/electronic interactions in the present. Multiple versions of writing have existed and coexisted both in the past and in more recently developed systems, like the little-known but ingenious indigenous scripts of West Africa. A definition based on Western history and alphabetic writing is no longer accepted as universally valid.

Unlike earlier models envisaging homogeneous patterns of orality or literacy throughout a 'culture' as a whole, recent studies explore how different groups or individuals use and/or conceptualize speech and writing in diverse directions, have differential access to them, and deploy them differently in different situations.

Even within the same locality, different categories of people use writing (or, indeed, speaking) in different ways. Some can read but not write; others may be multiply literate in different languages or different genres, using them for different purposes; others again may specialize in particular forms. There are social controls on who can speak, how, and where, and on the learning, practice, and purposes of writing.

The focus is thus now on the multiplex, localized, and differentiated ways that both speaking and writing are learned, valued, or practiced according to—for example—gender, rank, religious or educational role, age, geographical area, ethnicity, or situation. Critical and postmodernist analyses emphasize that both ideologies and practices may be related to the exercise of power, and can divide as well as unite. These spill over into studies of popular culture and the contemporary media, and of the role of literacy in the ideologies and practices of imperialism. (See also *Expressive Forms as Generators, Transmitters, and Transformers of Social Power*; *Popular Culture*.)

3.2 *Practices and Processes Rather than Texts and Products*

A further trend goes beyond models of oral and literate expression as forms of *text*. This partly dates back to work on performance by linguistic anthropologists, folklorists, and others from around the 1960s (see *Linguistic Anthropology*; *Performance: Anthropological Aspects*) but has intensified through more recent social science interests in process and practice. Many scholars have turned from studying fixed products, formal typologies, or even cognitive meanings toward investigating everyday situations, intertextual processes, emotive experiences, and the diversity of socially mediated expectations about how people do or should communicate.

Such studies point to the creative and diverse involvements of individuals in oral and written interaction, revealing something of the infinite inventiveness of human beings in the composition, enactment, and reproduction of verbal expression. Even in a single communicative situation different indivi-duals may be participating differentially, actively constructing the communicative process rather than merely forming a passive or homogeneous 'audience' for an already defined message. This can be applied to written forms too. Readers approach and interpret their material in diverse ways, with marked differences between cultures, historical periods, sectional groups, and even individuals (Boyarin 1993). The reality lies there as much as in the ostensibly objective and enduring text.

The investigation of 'text' continues but the term is now less often taken as a given. Rather, it is approached as a multifaceted issue for exploration, opening up questions about, for example, the socially shaped ways in which actions and communications are or are not crystallized in verbalized forms, how these relate to one another and to other aspects of social life, and the diverse processes through which they are actively put together and experienced.

3.3 Verbal Expression and Multidimensionality

This article started by defining the key concepts in terms of *verbalized* expression. But current research is increasingly emphasizing that verbal expression consists of *more* than words. Speaking uses uttered words, certainly, but also regularly relies on concurrent *visual* communication, through participants' bodily gestures, movements, and positionings (see *Sociolinguistics*); musical, olfactory, and tactile dimensions sometimes come in too. In writing there are images and visual displays over and above signs representing speech sounds, and sometimes tangible, acoustic, and olfactory dimensions. The materiality of artifacts such as books or letters now attracts interest, both in new transdisciplinary approaches to material culture and because computer developments challenge older-established print forms and in so doing raise awareness of properties previously taken for granted. Earlier ideologies positing the neutrality, universality, and transparency of print now themselves turn out as culturally molded and implicated in material artifacts not just mental representations. The multimodality of most, perhaps all, oral and written expression is likely to form one revealing research question in the future.

There are many other perspectives on speech and writing. But one virtue of focusing on the interrelation of oral and written expression, even on partially outmoded debates, is that it brings to the surface and challenges some powerful but often unacknowledged presuppositions. In doing so it can open doors to greater appreciation for both the continuities and the multiplex powers of these primarily (but not exclusively) verbal resources of human expression.

See also: Cultural Expression and Action; Culture, Production of; Expression and Interaction: Microfoundations of Expressive Forms; Linguistic Anthropology; Literacy and Illiteracy, History of; Literacy and Orality; Literacy Education; Movement Theory and Constraints in Syntax; Popular Culture; Writing, Evolution of

Bibliography

Barton D, Hamilton M 1998 *Local Literacies. Reading and Writing in One Community*. Routledge, London

Boone E H, Mignolo W D (eds.) 1994 *Writing Without Words. Alternative Literacies in Mesoamerica and the Andes*. Duke University Press, Durham, NC

Boyarin J (ed.) 1993 *The Ethnography of Reading*. University of California Press, Berkeley, CA

Cavallo G, Chartier R (eds.) 1995 *Storia della lettura nel mondo occidentale*. Laterza, Rome [1999 *A History of Reading in the West*. Polity Press, Cambridge, UK]

Daniels P T, Bright W (eds.) 1996 *The World's Writing Systems*. Oxford University Press, New York

Feld S 1990 *Sound and Sentiment: Birds, Weeping, Poetics, and Song in Kaluli Expression*, 2nd edn. University of Pennsylvania Press, Philadelphia, PA

Finnegan R 1992a *Oral Traditions and the Verbal Arts: A Guide to Research Practices*. Routledge, London

Finnegan R 1992b *Oral Poetry: its Nature, Significance and Social Context*, 2nd edn. Indiana University Press, Bloomington, IN

Foley J M 1995 *The Singer of Tales in Performance*. Indiana University Press, Bloomington, IN

Freebody P, Gee J, Luke A, Street B 1996 *Literacy as Social Practice: An Introduction to Critical-cultural Approaches to Reading and Writing*. Falmer Press, Brighton, UK

Goody J (ed.) 1968 *Literacy in Traditional Societies*. Cambridge University Press, London

Goody J 1987 *The Interface Between the Written and the Oral*. Cambridge University Press, Cambridge, UK

McLuhan M 1964 *Understanding Media: The Extensions of Man*. Routledge and Kegan Paul, London

Okpewho I 1992 *African Oral Literature. Background, Character, and Continuity*. Indiana University Press, Bloomington, IN

Olson D R, Torrance N (eds.) 1991 *Literacy and Orality*. Cambridge University Press, Cambridge, UK

Ong W 1982 *Orality and Literacy. The Technologizing of the Word*. Methuen, London

Schousboe K, Larsen M T (eds.) 1989 *Literacy and Society*. Akademisk Forlag, Copenhagen

Sherzer J, Woodbury A C (eds.) 1987 *Native American Discourse: Poetics and Rhetoric*. Cambridge University Press, Cambridge, UK

Street B (ed.) 1993 *Cross-cultural Approaches to Literacy*. Cambridge University Press, Cambridge, UK

Tedlock D 1983 *The Spoken Word and the Work of Interpretation*. University of Pennsylvania Press, Philadelphia, PA

Wagner D A, Venezky R L, Street B V (eds.) 1999 *Literacy: An International Handbook*. Westview Press, Boulder, CO

R. Finnegan

Oral History

Oral historians include amateurs, professionals, students, family members interested in genealogy, historical association staff members, and detectives. Not all oral historians are historians—they may be anthropologists, sociologists, psychologists, medical researchers, or journalists. Some scholars, seeking to restore a human touch research subjects, have presented recorded verbal evidence as alternatives to the dehumanizing effects of data collection in survey research. Since the late nineteenth century, scholars have interviewed persons from all walks of life and analyzed their findings. When recorded wisely, and with sensitivity, oral history testimony offers the historian a richness of detail not available from any other source, and is therefore invaluable as a research tool. Oral history is a modern resource used to create documents that help us understand the past and that are stored in forms that future scholars can use in their

own research. Today the techniques of oral history rely on the availability of modern technology, principally recording devices and the computer. Oral history permits scholars to capture the experiences of persons willing to speak about themselves. This is living history, and, although it comments on the past, it is rooted in the present.

1. *Origins and Development of Oral History*

Personal testimony always has been considered an integral part of the historical record. Oral history is even older than recorded history. Many African tribes employed *griots*, trained historian–story tellers who remembered the history of their community. Others used drummers and drum language (praise poems); musicians drummed the history of the state during official ceremonies and celebrations. When they grew old, they trained younger men to take their place. The position of chief drummer was hereditary, passed on to sons.

Ancient historians often incorporated tales and legends out of the oral tradition into their writing. The first written primary sources may well have been oral testimony recorded on paper or written down from recollections. Christianity evolved a collective memory of the life of Jesus in part by imposing European notions of the Holy Land on the existing template of Jewish holy places. Over time, however, historians came to trust written documentation more, considering it more 'objective.' The advent of the printing press gave written words more legitimacy because printed text was so precise, and seemed permanent. Nineteenth-century scientists relied on published professional journals to communicate discoveries and to attest scientific proofs.

Out of all of this came the modern attitude that 'nothing has value unless it is written down.' Transcriptions of oral sources, of course, were rare, although some have survived. Ancient Egyptian scribes kept detailed written records of political administration, a tradition maintained in Christendom through the Middle Ages by monastic copyists and in Judaism by Talmudic commentators. The Inquisition kept its own transcriptions, as did tribunals assessing forms of heresy from Salem to the noxious anticommunist US Senate hearings headed by Senator Joseph McCarthy. Verbatim recording of conversations and arguments go back at least as early as the mid-seventeenth century Putney Debates in England, when soldiers and commoners were permitted to discuss publicly issues that never usually came near the surface of society.

The invention of recording devices produced revolutionary changes in oral history. Interviewers using recorders captured phrasing and stress in a person's voice. Listeners could speculate about whether an interviewee was nervously trying to mislead the interviewer. Centuries later, with the invention of the wire recorder in the 1930s, came the first scholarly efforts to interview people and collect memoirs in recorded fashion. Some police officials used wire recorders as early forms of lie-detector devices, in the same unscientific way as law enforcement officials in earlier decades had turned to phrenology (mapping of skull structure). These primitive devices eventually gave way to tape recorders, which made sound reproduction much more lifelike. More important, the appearance of hand-held recorders made it possible to leave the recording studio and seek interviews onsite.

In 1948, the first major oral history research office opened under the direction of Professor Allan Nevins at Columbia University. One of the reasons for Nevins' interest was the impact on American society of World War II, a conflict so traumatic that it created a sharp dividing line between pre- and postwar days. Radio played a major role in transmitting information during and after the War, and interviews had become an important part of news programming and, to some extent, of newsreels. Even before the War, journalism, especially in the form of such national news magazines as *Time and Life*, helped popularize interviewing. Sometimes interviews were combined with extensive use of photographs. Lack of access to transcripts, however, made it impossible to know to what extent judicious (or malicious) editing may have altered the intent or meaning of interview material. Not until oral history projects developed the means to transcribe interviews and make them available to researchers could it be said that interviews became the basis for serious scholarship.

At the end world War II in 1945, several American government agencies set out to write detailed histories of the war and of the home front. Some projects broadened this effort to include civilian leaders, especially the men who played important political roles during the New Deal in the 1930s. Many programs in the early days of television also used interviews to elicit opinions and newsworthy material for public dissemination. Interviewing by scholars fell into two major areas: folklore research, and projects carried out by archivists and historians to assemble historical documentation from statements and comments by persons involved in historical events.

Institutional oral history projects initiated in the 1950s and 1960s on university campuses and in government agencies followed the Columbia model, concentrating on people who had played key roles in historical events. This was, as the Columbia University Center's directors acknowledge today, 'a top-down history … elitist to the core.' As a result, those asked to provide oral memoirs—some life histories involving dozens of hours of taping—were overwhelmingly white males who had held positions of importance in government and in society. It was taken for granted that the ideal interviewer was someone with as little involvement in the world of the interviewee as possible.

This set off oral history as a distinctive kind of fieldwork, subject to 'scientific' rules, and reassured those who felt that interviewing was too emotional a venture to produce objective results. The emphasis clearly was on collection development, not public use. This approach was hardened to some degree by the fact that the Oral History Association (founded in 1966) at first tended to be dominated by archivists and institutional historians who saw the purposes of their craft in a very traditional manner.

Scholars such as Wula K. Baum, Paul Buhle, Michael Frisch, Sherna Gluck, and Daphne Patai led the way in popularizing and charting the path in North American universities for the use of oral sources for the construction of the African past. This approach gained influence when scholars realized the inadequacy of written sources, for the most part for the period up to the era of European colonialism in Africa. In the 1970s, according to the Oral History Association, 'the volume of oral history began to reach floodgate proportions.' New approaches emerged, often at odds with the traditional model. Debates over methodology flourished. Some leading historians registered their disapproval of using oral sources in reconstructing the past. 'Old men drooling about their youth? No!' angry A. J. P. Taylor reputedly said. Other conservatives registered skepticism about the 'activist' thrust of the new oral history. Although communities constantly retell their own stories, 'ordinary' people, after all, still appear only in fleeting glimpses in records compiled by others for their own reasons. The new generation of community and social historians that took up oral history after the 1960s in many cases chose the interview process not only to gather information but also to raise consciousness among the persons selected to be interviewed. It was acknowledged that even people considered relatively obscure create their own history, and they do so within their own conceptions of its value and use in the culture. Some of the newer recruits to oral history emphasized the fact that oral history helps people see their own place in history, giving them a sense of self-worth, and therefore performing a dual function: historical research and consciousness raising. Oral history has flourished in continental Europe, tending to concentrate on amassing extensive collections of testimony and then analyzing them in the larger historical context. Alessandro Portelli has published path-breaking studies not only providing substance but also raising methodological issues. His *The Death of Luigi Trastulli and Other Stories* (1991) has become a classic in the field. Lutz Niethammer, Professor of German history at the Friedrich-Schiller University of Jena in 1989, published a monumental three-volume study *Lebensgeschichte und Sozialstruktur im Ruhrgebeit, 1930–60*, concentrating on the impact of Word War II and the immediate postwar period on the lives of ordinary residents of the Ruhr. French oral historians have produced major studies, including the 1976 *Grande Industrie et Migration Provinciale à Paris au début du XXème siècle* (*1910–30*). *Trajets Sociaux et Modes de Vie des Migrants: une Enquête d'Histoire Orale* (en collaboration avec I. Bertaux-Wiame) In Italy, Piero Caleffi published *Un Mondo Fuori dal Mondo. Indagine Doxa fra i Reduci dai Campi Nazisti*, a book based on the anguished testimonies of 317 Italian former deportees into Nazi death camps. A Croatian immigrant to Canada has launched a 'Slavonic Oral History' web site offering online testimonies.

2. *Basic Issues in Oral History*

Perhaps as a result of the tension between the old and new approaches, practitioners began to examine the techniques employed in oral history research. Arguments surfaced over the extent to which interviews could yield 'objective' information about the past, about the relationship between interviewee and interviewer, questions of manipulation (by persons on either side of the microphone), and the impact of bias, hindsight, age differences, and memory, as well as other issues combined under what specialists called 'reflexivity.' The bottom-up group tended to select populist topics—strikes, workers' movements, the plight of minorities, and 'outsiders' in society, and allied themselves in spirit with such projects as the Federal Writers Project of the 1930s and the popular histories by the writer Studs Terkel in Chicago, all addressed at giving voice to ordinary people from the lower classes.

One of the reasons why debate over methodology and validity in oral history has become heated is the fact that people feel it is more difficult to interpret spoken statements than to analyze written accounts. There are also other considerations. Unlike traditional museum collecting, for example, where artifacts from exotic civilizations arbitrarily dubbed 'primitive' by the collecting society were purchased (or stolen outright), oral historians, concerned with the wishes of their collaborators, knew that they needed to take steps to assure that interviewees understood to what uses their depositions would be put. Oral historians must be highly sensitive to their subject's rights. Some argue that in oral history there must be a prearranged formula for determining what will be kept in the record and under what conditions statements and comments by interviewees may be used, and for what purpose. This implies in turn a need to negotiate initially an understanding about the use of the finished product by all parties.

Oral history methods, of course, have been employed even when subjects are unavailable to be interviewed. Emmanuel Le Roy Ladurie's (1979) *Montaillou: The Promised Land of Error* used Inquisition documents that essentially were life histories as retrospective biographies to reconstruct life in four-

teenth-century France. Tom Harrisson used the files of the British government's wartime Mass Observation papers, a compilation of interviews and reports on everyday life, to write *Living Through the Blitz*. Glen H. Elder Jr.'s classic *Children of the Great Depression* relied on a longitudinal archive of direct reports, not on live interviews.

Scholars have speculated on the relationship between memory, oral history, and historical narrative. These include Susan Porter Benson edited 1986 collection, *Presenting the Past*, John Bodnar's (1992) *Remaking America: Public Memory*, *Commemoration*, and *Patriotism in the Twentieth Century*, and David Lowenthal's (1985) *The Past is a Foreign Country*. Ethnographer Barbara Myerhoff explains that oral history permits practitioners to see themselves through the lenses of others. The resulting dialogical relationship between interviewer and interviewee can shape new notions of identity for both, one that is provocatively new and different.

A common theme is the desire to include in historical discourse the voices of members of the population typically slighted or excluded from formal historical scholarship. In Brazil and Argentina, for example, researchers have taped lengthy interviews with families of victims (and with surviving victims themselves) who were arrested and mistreated by military and police authorities during periods when habeas corpus protections have been suspended.

By the 1980s, in the USA the rising interest in oral history had produced impressive results. The John F Kennedy Center in Boston embarked on a major oral history project to document the Kennedy years. The *International Journal of Oral History* began publishing in l980, frequently raising questions about the theoretical issues posed by the interviewer–subject relationship and about how to utilize evidence gathered from oral testimony. British oral history projects completed in recent years include studies of bobbin and bonnet making, the Co-operative Movement, ballet, the experiences of the blind, and gun running in Northern Ireland. Kevin Brownlow published a massive study of the silent movie industry in Britain and the USA between 1916 and 1928, *The Parade's Gone By*, an effort to balance the high-culture ways of studying silent cinema by listening to the voices of those who worked in the industry. Alexander Harrison used extensive testimonies from participants in France's Algerian War to reveal full details of the murders and tortures committed by French terrorists in *Challenging De Gaulle*.

Since then, oral history has proliferated to an even richer degree. Meetings on oral history have gathered hundreds of researchers in cities from Istanbul to Anchorage, Alaska. Oral history now includes the Internet, and digitalized collections of interviews make it possible to read transcripts stored thousands of miles away and sometimes even hear the spoken voices from the original interview.

The perception of verifiability has continued to ensure the primacy of written and visual evidence. Document-oriented historians look for precision in form, precision in chronology, and evidence from multiple texts for corroboration. Oral history lacks these. Traditional historians use physical documents—letters, diaries, reports—to help them comprehend what really went on behind the scenes—sources that are as objectively unreliable as oral testimony. For these and other reasons, oral history's claim to legitimacy rose as traditional forms of historical writing lost its aura of invincibility. In the latter part of the twentieth century we have seen a substantial change in the nature of records themselves. Frequent use of telephones and email makes it more and more difficult for historians to attempt to reconstruct historical processes. With so much of importance now erased or never recorded, oral testimony, even if after the fact, takes on added importance.

It is evident that oral history implies a perception of the past that has continuity with the present and that has not 'stopped.' The past's impact on the present, in fact, is a major reason for pursuing oral history. The process of oral history, moreover, educates the interviewee to the importance of his/her own life in the larger flow of history, even if the person did not hold a politically or socially prestigious station in society. Oral history is an antidote to traditional history, that is always in danger of slipping into 'official' history, bolstering the outlook on one particular group over the interests and values of others. It is a fragile undertaking, since revelations through oral history are totally dependent on pre-existing knowledge 'recorded before death wipes the slate clean.' Oral historians stand to make great democratizing contributions to a discipline that in the past mostly listened to the voices of the powerful.

Recent poststructural scholarship cast doubt on the efficacy of not only oral history documentation but also of any interpretation not an original source. Jacques Derrida, after all, claimed that every 'translation' creates a new text rather than a copy of an original. Others have raised the question of cultural politics. A language, according to Susan F Frank 'expresses what its powerful speakers, such as men, want said. Women and other subaltern groups,' Frank adds, 'have a different relationship to their language.' Oral history practitioners may take these warnings in any manner they choose, but it would be wise to listen to them and be all the more careful in looking after accuracy. In editing transcripts we are reminded that changing an oral text into a written text involves choices of the highest degree of subjectivity.

Persons seeking to work in oral history must guard against any attitude that assumes that merely because the interviewer is well educated, or possesses sophisticated recording equipment, or is well connected, he or she somehow does not have to obey any rules. Asking others to record their thoughts and feelings for

public use involves ethical considerations, especially when the person interviewed feels at a status or psychological disadvantage.

3. Debates Over Methodology

A basic point in the debate over oral history involves the representativeness of the person being interviewed. How typical were the personal experiences of those that are described? Another side of the same question involves the numbers of persons involved in the project. Sometimes it may be necessary to interview several persons from similar backgrounds in order to yield corroborative testimony. Much of this can be resolved in advance by careful planning of the project. One properly thought out and conducted will not likely run the risk of readers ultimately questioning the validity of the evidence, although we know that eye witnesses can flatly contradict one another, even when debriefed immediately after an event, and therefore increasing the number of respondents does not necessarily guarantee objectivity.

The most fundamental task of any historical or biographical research, one might argue, is to discover our present in the past we claim to have represented. Psychologists have even disagreed among themselves about the nature of autobiographical memory. Sigmund Freud among many others, held that every experience of an individual's life was stored somewhere in the brain and could be recalled by some technique, whether therapy, hypnosis, interrogation, drugs, or meditation. One view rejects this, arguing, instead, that memory is considerably more malleable and subject to revision. Memories are constantly refashioned: by new information, by suggestions received from others, and by shifts in one's own emotional state of mind, aging, trauma, and the passage of time. The task of deciphering 'unselfconscious memory, memory that is lived out rather than recited' is formidable.

It has been noted earlier that the pitfalls of oral sources differ little from those of traditional written documents. It is the basic dialectic of the subjective that all practitioners of historical reconstruction must confront. Rather than 'oral history' as an autonomous mode of knowledge, the challenge to the oral historian is to find insightful ways to interpret them, in combination with the widest array of other sources to produce 'history.' These points should be familiar enough to serious students of history, although the academy has gone far to remove itself from the 'real world.' Oral history cannot work unless the researcher learns how to listen. This is the key to successful oral history, not fancy tape recorders or teams of interviewers.

Gabriela Rosenthal, a psychotherapist, tells us that to reconstruct a life history one must analyze the data on which it is based, to find out how the narrator has approached the project and to discover 'whether the informant is orienting toward personal or the interviewer's relevances.' Given these and other pitfalls inherent in trying to figure out the meaning of the spoken statements, the safest way to use oral evidence can only be to apply rigorously the same methods (scientific but also intuitive) that historians employ in evaluating conventional written material. Common sense is probably the most reliable tool of all. The oral history process, moreover, offers a corrective element that historians working with written documents do not have: contributors can be asked to expand on what they have said, or challenged if it seems dubious. As George Trevelyan noted, long ago, the real test is whether the evidence is handled 'in good faith.'

The fact that oral history at times approaches the realm of psychotherapy raises complex ethical issues in that sphere. Dianne W. Hart, interviewing families caught up in the struggle against the Samoza dictatorship in Nicaragua, yielded one transcript that was so disturbing to her that she showed the transcript to a clinical psychologist. She tells this in a footnote in another of her books, and describes the psychoanalyst's diagnosis. She used a pseudonym, so the identity of the interview subjects remains anonymous. Still, some may object to what may seem to be an extraordinary invasion of privacy, even cloaked in anonymity.

Oral history projects are at risk of failing if the researcher attempts to uncover the unvarnished 'truth' about past events or issues. The classical Japanese film 'Rashomon,' a study of an event viewed through the eyes of different participants, showed brilliantly how eye-witness accounts differ from one another according to the personality and the point of view of the bystander. Even flawed interviews, where memory has partially failed (or when respondents embellish or distort), can produce valuable data in other forms. Objective reality, then, is a variable phenomenon. What oral historians should be after, rather, is insight gained from interpretation and background and the experience of interview and life history subjects. This insight is a building block which the historian can use, in tandem with other forms of evidence, to craft analytic interpretation. Oral history is a remarkable tool because it elicits information from sources who otherwise would remain silent, their testimony irretrievably lost.

Whether or not the subject responds with clear and accurate memory, then, in some ways is beside the point. It is human to present a subject, especially after a long passage of time, in ways that selectively leave out certain details or embellish others. People consciously and unconsciously try to put themselves in the best light. Researchers can attempt to be nonjudgmental, but respondents fear judgment and answer accordingly. Subjects obviously lying or distorting facts should be identified as such, but even in these cases

their interviews will likely add a good deal to the researcher's final synthesis. Even flawed interviews can yield useful data, even if peripheral. Trying to figure out the motives for deception also plays a part in the detective work of the researcher and may shed light on the theme or topics of interest.

Some people will decline invitations to be interviewed because they think they were 'too unimportant,' or too peripheral to the subject of interest. Researchers should use gentle persuasion to get them to change their minds. Often people do not realize how deep their memories are once questions start to be asked. Cases where self-deprecating subjects open up not only prove useful for research but also do a world of good educating them to their own self-worth. Oral history is not therapy but it may have therapeutic effects both for the subject and for the oral historian.

Oral history cannot be done in a vacuum. To ask informed questions, oral historians must know as much as possible about the historical context in which the interviews fall. The best oral history is always done in tandem with other research activities, whether done in archives, in newspaper files, with quantitative data, or in other conventional sources of historical material. Oral history without context is merely story telling, although in rare cases (say, the oral history of Bedouins, or peoples lacking a written language) this may be the only way to collect information.

Good oral history does not draw its strength from nostalgia or from colorful statements, although there are elements of style that enhance both the content and overall effectiveness of an oral record. To be worthwhile, oral evidence has to be treated as rigorously as any other source of information. It is, in David Hey's words, 'not an easy option.' The best oral history complements other forms of historical evidence as much as it brings new testimony to old debates.

The problem is that historians have only superficially attempted to evaluate and grasp what Columbia University's Ron Grele calls 'the underlying structure of consciousness which both governs and informs oral history interviews. Persons asked to record their life history, for example, may deduce (rightly or wrongly) that the interviewer wants or expects one kind or another response—perhaps a debunking one, or one at odds with popular wisdom (what, then, is the purpose for interviewing me?, someone might ask). Oral history researchers, however, are not psychologists, and should never assume that role. Common sense is usually sufficient to guide interviews through difficult stages. Compassion—but not condescension—plays an important role. Where interviews touch on emotional or traumatic subjects (unhappy childhoods, the loss of loved ones, wartime ordeals, acts of brutality), the interviewer must never press. Frequent breaks in the interview may help. It is not uncommon for subjects to lose their composure, to become angry, or to burst into tears. Sensitivity and discretion must be present at all times. The process of oral history interviewing, because it involves the structuring of memory, is actually a process in the construction of a usable past. Writing one's diary—or speaking about one's past before a microphone—is an act of anticipatory memory. This is neither scientific nor objective. Psychotherapists know that such memories sometimes are slanted or distorted subconsciously to cover for, or screen, another memory that has been repressed. Autobiographical responses seek temporal anchors for what feels precarious in the present. Points that respondents preserve in diaries and in recalled anecdotes or memories, psychologists tell us, appear to be the ones they yearn to realize and preserve in their lives—flaws mended, limitations and isolations overcome, capacities refined.

See also: Historiography and Historical Thought: Current Trends; History and Memory; History and the Social Sciences; History: Forms of Presentation, Discourses, and Functions; History: Theories and Methods; Oral and Literate Culture

Bibliography

Adams C 1987 *Across Seven Seas and Thirteen Rivers: Life Stories of Pioneer Sylhetti Settlers in Britain*. THAP Books, London

Avila C 1996 *Historia y Testemonios Orales*. INAH, Mexico City, Mexico

Baddley A 1976 *The Psychology of Memory*. Basic Books, New York

Baum W K 1977 *Transcribing and Editing Oral History*. American Association for State and Local History, Nashville, TN

Benson S P, Brier S, Rozenzweig et al 1986 *Presenting the Past*. Temple University Press, Philadephia, PA

Bodnar J E 1992 *Remaking America*. Princeton University Press, Princeton, NJ

Brownlaw K 1968 *The Parade's Gone By*. Knopf, New York

Chinn C 1988 *They Worked All Their Lives: Women of the Urban Poor in England, 1880–1939*. Manchester University Press, Manchester, UK

Elder G H Jr 1998 *Children of the Great Depression*. West View Press, Boulder, CO

Gluck S B, Patai D 1991 *Women's Words: The Feminist Practice of Oral History*. Routledge, New York

Grele R J et al. 1975 *Envelopes of Sound: Six Practitioners Discuss the Method, Theory and Practice of Oral History and Oral Testimony*. Precedent Publishers, Chicago, reprinted (1991) as *Envelopes of Sound: the Art of Oral History*. Praeger, New York

Harrison A 1991 *Challenging De Gaulle: The O.A.S. and the Counterrevolution in Algeria*. Praeger, New York

Hart D 1990 *Thanks to God and the Revolution: The Oral History of a Nicaraguan Family*. Wisconsin Press, Madison, WI

Hartley M, Ingleby J 1976 *Life and Tradition in the Yorkshire Dales*. Dent, London

Josselson R, Lieblich A (eds.) 1993 *The Narrative Study of Lives*. Sage, Newbury Park, CA

Kodama-Nishimoto M, Nishimoto W S, Oshiro C A *Hanahana: An Oral History Anthology of Hawaii's Working People*. University of Hawaii Press at Manoa, Honolulu, HI

Ladurie E, Le R 1979 *Montalliou: the Promised Land of Error*. [Trans Bray B.] Random House, New York
Lavie S 1990 *The Poetics of Military Occupation: Mzeina Allegories of Bedouin Identity under Israeli and Egyptian Rule*. University of California Press, Berkeley, CA
Laventhal D 1985 *The Past is a Foreign Country*. Cambridge University Press, New York
Lummis T 1987 *Listening to History: the Authenticity of Oral Evidence*. Hutchinson, London
Patai D 1988 *Brazilian Women Speak*. Rutgers University Press, New Brunswick, NJ
Salazar A 1990 *No Nacimos pa' Semilla*, 1st edn. CINEP Bogotà, Colombia
Sebe Bom Meihy J C 1996 *Manual de Historia Oral*. Edições Loyola, São Paulo, Brazil
Thompson P R (ed.) 1982 *Our Common History: The Transformation of Europe*. Humanities Press, Atlantic Highlands, NJ
Thompson P 1978 *The Voice of the Past: Oral History*. Oxford University Press, Oxford, UK
Thompson P R, Perks R 1989 *Telling it How it Was: A Guide to Recording Oral History*. BBC, London
Vansina J 1965 *Oral Tradition: A Study in Historical Methodology*. Routledge and Kegan Paul, London
Vansina J 1966 *Kingdoms of the Savanna*. University of Wisconsin Press, Madison, WI
Vansina J 1978 *The Children of Woot: A History of the Kuba Peoples*. University of Wisconsin Press, Madison, WI

R. M. Levine and J. C. Sebe Bom Meihy

Order Statistics

The field of *order statistics* deals with the properties and applications of ordered random variables and of functions involving them. Simple examples, useful as descriptive statistics, are the sample maximum and minimum, median, range, and trimmed means. Major applications include the estimation of location and scale parameters in complete or censored samples, selection procedures, the treatment of outliers, robust estimation, simultaneous inference, data compression, and probability plotting. Although important in the theory of nonparametric statistics, the subject of order statistics is very largely parametric, that is, its results depend generally on the form of the population(s) sampled.

1. Introduction

Let the variates $X_1, \ldots, X_n$ be arranged in ascending order and written $X_{1:n} \leqslant \cdots \leqslant X_{r:n} \leqslant \cdots \leqslant X_{n:n}$. Then $X_{r:n}$ is the rth order statistic ($r = 1, \ldots, n$), a term also applied to the corresponding observed value, $x_{r:n}$. The *median* is $X_{(n+1)/2:n}$ for n odd and $\frac{1}{2}(X_{n/2:n}+X_{n/2+1:n})$ for n even. Helped by their simplicity, the median and the *range*, $X_{n:n}-X_{1:n}$, have long been valuable as measures of location and scale, respectively.

Consider the following illustrative data, with $i = 1, \ldots, 8$

x_i:	16	40	23	12	240	33	17	27
$x_{i:8}$:	12	16	17	23	27	33	40	240

These figures may represent the annual incomes (in multiples of \$1,000) at age 30 of eight high-school dropouts. Because of the *outlier* $x_5 = 240$, the median, $\frac{1}{2}(23+27) = 25$, clearly gives a better picture of the group than the mean, 51. The median would in fact be unchanged as long as $x_5 \geqslant 27$, which shows just how *robust* the median is.

Next, suppose that the x_i are test scores out of a total of 40. If we are unable to correct the impossible score of 240, we eliminate or *censor* it. The median is then $x_{4:7} = 23$, no longer very different from the mean, $\bar{x} = 24$. The range is not robust, dropping from 228 to 28, and is therefore useful in detecting outliers, which are seldom as obvious as in the above data (see Sect. 6.3). For 'well-behaved' data, the range provides a simple measure of variability or scale.

In time series, $X_1, X_2, \ldots$, these descriptive statistics are particularly helpful when based on small moving subsamples $X_j, X_{j+1}, \ldots, X_{j+n-1}$ ($j = 1,2, \ldots$). The resulting *moving* (or *running*) *medians* provide a current measure of location that, unlike the moving mean, is not too sensitive to outliers. The sequence of moving medians represents a smoothing of the data, with important applications in signal and image processing. Moving ranges are useful in quality control charts, where their sensitivity to outliers is a desirable feature. If more robust current estimators of dispersion are required, one can turn to the difference of the second largest and the second smallest in each moving sample, etc.

Again, the above x_i may represent response times to a stimulus. The *extremes*, 12 and 240, are then of particular interest (see Sect. 5.2). Of course, we would want more extensive data for which we may also wish to study the *upper* and *lower records*, here (16, 40, 240) and (16, 12) (see *Nonparametric Statistics: Records*).

2. Basic Distribution Theory

If $X_1, \ldots, X_n$ is a random sample from a population with cumulative distribution function (cdf) $F(x)$, then the cdf of $X_{r:n}$ is given by

$$F_{X_{r:n}}(x) = P(X_{r:n} \leqslant x) = P$$

(at least r of $X_1, \ldots, X_n$ are less than or equal to x)

$$= \sum_{i=r}^{n} \binom{n}{i} F^i(x)\,[1-F(x)]^{n-i} \qquad (1)$$

since the term in the summand is the binomial probability that exactly i of $X_1, \ldots, X_n$ are less than or equal to x.

In handling order statistics, usually it is assumed that the underlying variate X is continuous. Note that this assumption is not needed in the derivation of Equation (1). Also, if X can take only integral values, then $P(X_{r:n} = x) = F_{X_{r:n}}(x) - F_{X_{r:n}}(x-1)$. However, if X is discrete, the probability of ties among the observations is positive, which tends to complicate the distribution theory (see, e.g., David 1981, p. 13, Arnold et al. 1992, p. 41).

If the probability density function (pdf) $f(x)$ exists, as we will assume from here on, then differentiation of Eqn. (1) yields

$$f_{X_{r:n}}(x) = \frac{n!}{(r-1)!(n-r)!} F^{r-1}(x) f(x) [1-F(x)]^{n-r} \quad (2)$$

Also, the joint pdf of $X_{r:n}$ and $X_{s:n}$ $(1 \leqslant r < s \leqslant n)$ is, for $x \leqslant y$ (e.g., David 1981, p. 10, Arnold et al. 1992, p. 16)

$$f_{X_{r:n}, X_{s:n}}(x, y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \times F^{r-1}(x) f(x) [F(y)-F(x)]^{s-r-1} f(y) [1-F(y)]^{n-s} \quad (3)$$

Equations (2) and (3) are fundamental and permit the calculation, usually by numerical integration, of the means, variances, and covariances. Also, Eqn. (3) makes possible the derivation, by standard methods, of the pdf of functions of two order statistics, such as the range $W_n = X_{n:n} - X_{1:n}$. To facilitate the use of order statistics, extensive tables have been developed, especially when $F(x) = \Phi(x)$, the standard normal cdf (see Pearson and Hartley 1970/1972, Harter and Balakrishnan 1996, 1997).

3. Distribution-free Interval Estimation

Three important types of distribution-free intervals involving order statistics may be distinguished. For simplicity we assume that $F(x)$ is strictly increasing for $0 < F(x) < 1$.

3.1 Confidence Intervals for Quantiles

The (population) quantile of order p, ξ_p, is defined by $F(\xi_p) = p$, $0 < p < 1$. In particular, $\xi_{1/2}$ is the median. Since, for $r < s$ $P(X_{r:n} < \xi_p < X_{s:n}) = P(X_{r:n} < \xi_p) - P(X_{s:n} < \xi_p)$, it follows from Eqn. (1) that the interval $(X_{r:n}, X_{s:n})$ covers ξ_p with probability

$$P_1 = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}$$

3.2 Tolerance Intervals

The focus here is on the probability, P_2, with which $(X_{r:n}, X_{s:n})$ covers a proportion γ of the underlying population. Thus $P_2 = P\{F(X_{s:n}) - F(X_{r:n}) \geqslant \gamma\}$ Since $F(X)$ is a uniform (0,1) variate, U, $P_2 = P(U_{s:n} - U_{r:n} \geqslant \gamma) = 1 - I_\gamma(s-r, n-s+r+1)$, with the help of Equation (3), where $I_\gamma(a,b)$ is the incomplete beta function.

3.3 Prediction Intervals

Let $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ be independent variates with cdf $F(x)$. The Xs may represent current observations and the Ys future observations. What is the probability, P_3, that $Y_{t:n}$ $(t = 1, \ldots, n)$ falls between $X_{r:m}$ and $X_{s:m}$ $(1 \leqslant r < s \leqslant m)$? It can be shown that $P_3 = P(Y_{t:n} > X_{r:m}) - P(Y_{t:n} > X_{s:m})$, where

$$P(Y_{t:n} > X_{r:m}) = \sum_{i \geqslant r} \binom{m}{i} \binom{n}{r+t-1-i} \Big/ \binom{m+n}{r+t-1}.$$

A good account from the user's point of view of all these intervals is given in Chap. 5 of Hahn and Meeker (1991).

4. Estimation of Parameters

An important simple example is the use of range in quality control to estimate the standard deviation (sd) σ in a normal sample. If $w_{n;\alpha}$ denotes the upper α significance point of the range in a standard normal population, then from $P(w_{n;1-\alpha/2} < W_n/\sigma < w_{n;\alpha/2}) = 1-\alpha$ it follows that a $(1-\alpha)$ confidence interval (CI) for σ is

$$\frac{w_n}{w_{n;\alpha/2}} < \sigma < \frac{w_n}{w_{n;1-\alpha/2}},$$

where w_n is the observed value of the range.

An unbiased point estimator of σ is $\hat{\sigma}_W = W_n/d_n$, where $d_n = E(W_n/\sigma)$ is widely tabulated. Let $\hat{\sigma}_S$ denote the unbiased root-mean-square estimator of σ. In small samples the efficiency of $\hat{\sigma}_W$, i.e., var $\hat{\sigma}_S$/var $\hat{\sigma}_W$, is quite high (0.955 for $n = 5$).

More systematically, linear functions of the order statistics (L-statistics) can be used to estimate both location and scale parameters, μ and σ (not necessarily mean and sd), for distributions with pdf of the form

$$f(x; \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right), \quad \sigma > 0 \quad (4)$$

With $Y = (X-\mu)/\sigma$, let $\alpha_{r:n} = E(Y_{r:n})$ and $\beta_{rs:n} = \text{cov}$

$(Y_{r:n}, Y_{s:n})$ $(r, s = 1, \ldots, n)$. It follows that $X_{r:n} = \mu + \sigma Y_{r:n}$, so that

$$E(X_{r:n}) = \mu + \sigma\alpha_{r:n} \tag{5}$$

$$\mathrm{cov}(X_{r:n}, X_{s:n}) = \sigma^2\beta_{rs:n} \tag{6}$$

A major advance was made by Lloyd (1952), who recognized that the generalized Gauss–Markov theorem can be applied to the estimation of μ and σ, in view of Eqns. (5) and (6). This yields the best linear unbiased estimators (BLUEs).

$$\mu^* = \sum_{i=1}^{n} a_{in}X_{i:n} \text{ and } \sigma^* = \sum_{i-1}^{n} b_{in}X_{i:n} \tag{7}$$

where the a_{in} and b_{in} are functions of the $\alpha_{i:n}$ and $\beta_{ij:n}$ and can be evaluated once and for all. For tables in the normal case, see Sarhan and Greenberg (1962).

The statistics in Eqn. in (7) provide estimators that are unbiased and of minimum variance within the class of linear unbiased estimators. They are particularly convenient for data in which some of the smallest and/or largest observations have been censored, since maximum likelihood (ML) estimation tends to be difficult in this situation. Also, for complete and censored samples, ML estimation may produce biased estimators. However, μ^* and σ^* may not have the smallest mean squared error, even among L-statistics. Simplified approximate estimation procedures are also possible (see, e.g., David 1981, p. 128 or Arnold et al. 1992, p. 171).

5. *Asymptotic Theory*

Except for the distribution-free results of Sect. 2, the practical use of order statistics in finite samples depends heavily on the availability of tables or computer programs appropriate for the underlying population. In large samples, asymptotic theory can provide approximate results that are both simpler and more widely applicable.

It is useful to distinguish three situations as $n \to \infty$: (a) *Central order statistics* $X_{(np):n}$, where (np) denotes here the largest integer $\leqslant np+1$, with $0 < p < 1$; (b) *Extreme order statistics* $X_{r:n}$, where $r = 1, \ldots, m$ or $r = n-m+1, \ldots n$, and m is a fixed positive integer; (c) *Intermediate order statistics* $X_{r:n}$ or $X_{n-r+1:n}$, where $r \to \infty$ but $r/n \to 0$.

5.1 *Central Order Statistics*

Corresponding to the population quantile ξ_p of Sect. 2.1, it is natural to call $X_{(np):n}$ a *sample quantile*. The following important result then holds (e.g., David 1981, p. 254). Let $X_1, \ldots, X_n$ be a random sample from a population with pdf $f(x)$ and put $n_j = (np_j)$, $j = 1, \ldots, k$, where k is fixed and $0 < p_1 < p_2 < \cdots < p_k < 1$. If $0 < p(\xi_{p_j}) < \infty$, $j = 1, \ldots, k$, then the asymptotic joint distribution of $\sqrt{n}(X_{n_1:n} - \xi_{p_1}), \ldots, \sqrt{n}(X_{n_k:n} - \xi_{p_k})$ is k-dimensional normal, with zero mean vector and covariance matrix $(\sigma_{jj'})$, where

$$(\sigma_{jj'}) = \frac{p_j(1-p_j)}{f(\xi_{p_j})f(\xi_{p_{j'}})} \quad j \leqslant j'$$

As a very special case, it follows that for the median $M_n = X_{(n+1)/2:n}$ (n odd), as $n \to \infty$, $\sqrt{n}(M_n - \mu) \to N(0, \pi\sigma^2/2)$ showing incidentally that, as an estimator of μ, M_n has asymptotic efficiency (relative to $\bar{X}$) of $2/\pi = 0.637$. The same result holds for n even.

5.2 *Extreme Order Statistics*

The situation is more complicated when $p_1 = 0$ or $p_k = 1$, corresponding to the *extremes* $X_{1:n}$, $X_{n:n}$ and more generally to the mth extremes $X_{m:n}$, $X_{n+1-m:n}$, where m is a fixed positive integer. For $X_{n:n}$, suitably normalized, there are three limiting distributions, the simplest of which is the Type III extreme-value distribution with cdf $\Lambda_3(x) = \exp(-\mathrm{e}^{-x}) - \infty < x < \infty$. This distribution has been found very useful in describing the behavior of floods and other extreme meteorological phenomena. A convenient sufficient condition for $\Lambda_3(x)$ to be the appropriate limiting distribution is that the underlying distribution satisfy

$$\lim_{x\to\infty} \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{1-F(x)}{f(x)}\right] = 0 \tag{8}$$

Then

$$\lim_{x\to\infty} P\{(X_{n:n} - a_n)\,\mathrm{nf}(a_n) \leqslant x) = \Lambda_3(x)$$

where a_n is given by $F(a_n) = 1 - 1/n$. Distributions for which Eqn. (8) holds include the normal, gamma, and logistic families. For a full treatment of the extreme-value case, see Galambos (1987).

5.3 *Intermediate Order Statistics*

An example of an intermediate order statistic is $X_{\sqrt{n}:n}$. A suitably normalized form of $X_{r:n}$ may have either a normal or nonnormal asymptotic distribution (e.g., Arnold et al. 1992, p. 226).

5.4 *Asymptotic Independence Results*

A useful result is that, after appropriate normalization, upper extremes are asymptotically independent of

lower extremes. Moreover, all extremes are asymptotically independent of the sample mean (e.g., David 1981, pp. 267–70).

5.5 L-Statistics

It is clear from the above that a linear function of sample quantiles, $\sum_{j=1}^{k} c_j X_{n_j:n}$, tends to normality as $n \to \infty$ (k fixed). On the other hand, the asymptotic distribution of linear functions of the extremes, such as the range, will be more complicated. An L-statistic, $\sum_{i=1}^{n} c_{in} X_{i:n}$, involving extreme, intermediate, and central order statistics, may be asymptotically normal or nonnormal, depending on the underlying distribution and on how much weight is placed on the extreme order statistics.

6. Applications

6.1 Probability Plotting

This very useful graphical method typically is applied to data assumed to follow a distribution depending, as in Eqn. (3), only on location and scale parameters μ and σ. The ordered observations $x_{i:n}$(i = 1, …, n) are plotted against $v_i = G^{-1}(p_i)$, where G is the cdf of $Y = (X-\mu)/\sigma$ and the p_i are probability levels, such as $p_i = (\mathrm{i}-1/2)/n$. The plotting is immediate if probability paper corresponding to G is available, as it is for, for example, the normal, lognormal, and Weibull distributions. If a straight-line fit through the points (v_i, $x_{i:n}$) seems unreasonable, doubt is cast on the appropriateness of G. Often this simple procedure is adequate, but it may be followed up by a more formal test of goodness-of-fit.

If a straight line fits satisfactorily, a second use may be made of the graph: σ may be estimated by the slope of the line and μ by its intercept on the vertical axis. A detailed account of probability plotting and the related *hazard plotting* is given in Nelson (1982).

6.2 Selection Differentials and Concomitants of Order Statistics

If the top k ($< n$) individuals are chosen in a sample from a population with mean μ and sd σ, then $D_{k,n} = (X_{n-k+1:n} + \cdots + X_{n:n} - k\mu)/k\sigma$ measures the superiority, in sd units, of the selected group. The mean and variance of this *selection differential* is given immediately by Eqns. (5) and (6) as $E\,(D_{k,n}) = \sum \alpha_{i:n}/k$, var $(D_{k,n}) = \sum\sum \beta_{ij:n}/k^2$, where all summations run from $n-k+1$ to n.

Often one is interested in how this group performs on some related measurement Y. Thus, X may refer to a screening test and Y to a later test or X to a parent animal and Y to an offspring. The Y-variate paired with $X_{r:n}$ may be denoted by $Y_{[r:n]}$ and called the *concomitant of the rth order statistic*. Corresponding to $D_{k,n}$, the superiority of the selected group on the Y-measurement is, with obvious notation, $D_{[k,n]} = (\sum Y_{[i:n]} - k\mu_Y)/k\sigma_Y$. If Y is linearly related to X, such that $Y_i = \mu_Y + \rho\sigma_Y(X_i - \mu_X)/\sigma_X + Z_i$ where ρ is the correlation coefficient between X and Y, and Z_i is independent of X_i, with mean 0 and variance $\sigma_Y^2(1-\rho^2)$, then $Y_{[r:n]} = \mu_Y + \rho\sigma_Y(X_{r:n} - \mu_X)/\sigma_X + Z_{[r]}$ where $Z_{[r]}$ is the Z-variate associated with $X_{r:n}$ (e.g., David 1981, p.110). It follows that $EY_{r:n} = \mu_Y + \rho\sigma_Y\alpha_{r:n}$, $\mathrm{var}\,Y_{[r:n]} = \sigma_Y^2(\rho^2\beta_{rr:n} + 1 - \rho^2)$, $\mathrm{cov}(Y_{[r:n]}, Y_{[s:n]}) = \rho^2\sigma_Y^2\beta_{rs:n}$ $r \neq s$ so that the mean and variance of $D_{[k,n]}$, the *induced selection differential*, are $ED_{[k,n]} = \rho ED_{k:n}$, $\mathrm{var}\,D_{[k,n]} = \rho^2\,\mathrm{var}\,D_{k:n} + (1-\rho^2)/k$. The asymptotic distribution of $D_{[k:n]}$ as well as many other properties and applications of concomitants of order statistics are treated in David and Nagaraja (1998).

6.3 Outliers and Selection of the Best

Order statistics, especially $X_{n:n}$ and $X_{1:n}$, enter naturally into many outlier tests. Thus, if $X_1, \ldots, X_n$ are supposedly a random sample from a normal distribution with unknown mean and known variance σ^2, then large values of $B_n = (X_{n:n} - \bar{X})/\sigma$ make this null hypothesis, H_0, suspect. The corresponding two-sided statistic is $B_n' = \max_{i=1,\ldots,n} |X_i - \bar{X}|/\sigma$. If σ is unknown, studentized versions of B_n or B_n' can be used. The subject, with its many ramifications, is treated extensively in Barnett and Lewis (1994), where tables needed to carry out such tests are given. If the observed value of the chosen statistic is large enough, H_0 is rejected and $X_{n:n}$ (or $X_{1:n}$) becomes a candidate for further examination or rejection.

The foregoing approach is related closely to the following multiple decision procedure, illustrated on the balanced one-way classification with n 'treatment' means $\bar{X}_i$, i = 1, …, n, each based on m replications. To test the null hypothesis H_0 of equal population means $\mu_1, \ldots, \mu_n$ against the 'slippage' alternative $\mu_1 = \cdots = \mu_{i-1} = \mu_{i+1} = \cdots \mu_n$ and $\mu_i = \mu_1 + \Delta$, where i and $\Delta < 0$ are unknown, $n+1$ decisions may be taken, that is, D_0:H_0 holds or D_i: the ith treatment mean has slipped to the right. Under the usual normal-theory assumptions, an optimal procedure at significance level α is to choose D_M if (e.g., David 1981, p. 226) $C_{n,v} = m\ (\bar{X}_M - \bar{X})/(\mathrm{TSS})^{1/2} > c_{n,v;\alpha}$, where M is the subscript corresponding to the largest $\bar{X}_i$, TSS is the analysis of variance total sum of squares, and $c_{n,v;\alpha}$, with $v = n(m-1)$, is given in Table 26a of Pearson and Hartley (1970/1972). If $C_{n,v} < c_{n,v;\alpha}$ the decision D_0 is taken. Slippage in either direction can be handled similarly with Table 26b.

In the above, the focus is on the detection of a single outlier or a single superior (or inferior) treatment. The handling of multiple outliers is an important part of Barnett and Lewis (1994) (see *Statistical Analysis, Special Problems of: Outliers*).

6.4 Simultaneous Inference

Order statistics play an important supporting role in simultaneous inference procedures. A well-known example is Tukey's use of the studentized range in setting simultaneous confidence intervals for the treatment means in fixed-effects balanced designs (e.g., see *Hypothesis Tests, Multiplicity of*).

6.5 Robust Estimation

Here one gives up the classical idea of using an estimator that is optimal under tightly specified conditions in favor of an estimator that retains reasonably good properties under a variety of departures from the hypothesized situation. Since the largest and smallest observations tend to be more influenced by such departures, which include the presence of outliers, many robust procedures are based on the more central order statistics. A simple example is provided by the *trimmed means*

$$T_n(j) = \sum_{i=j+1}^{n-j} \frac{X_{i:n}}{n-2j} \quad 0 < j < \tfrac{1}{2}n$$

which include the sample median as a special case. For a very readable treatment of robust methods see Hoaglin et al. (1983) (see also *Robustness in Statistics*).

6.6 Data Compression and Optimal Spacing

Given a large random sample of size n from a population with pdf, Eqn. (4), it is possible to estimate μ or σ or both from a fixed small number k of well chosen order statistics. The loss in efficiency of estimation may be more than compensated by the storage space gained, especially as a test of the assumed underlying distribution can still be made by probability plotting. For example, in the case of a normal population, the optimal estimator of μ is, for $k = 4$,

$$\mu^* = 0.1918(X_{(0.1068n):n} + X_{(0.8932n):n}) + 0.3082\,(X_{(0.3512nn):n} + X_{(0.6488n):n})$$

where $(0.1068n)$ is the largest integer $\leqslant 0.1068n+1$, etc. The estimator μ^* has asymptotic efficiency (AE) 0.920 and, since it does not involve the more extreme order statistics, is more robust than $\bar{X}$. Also, μ^* is much more efficient than the median (AE = 0.637) (see, e.g., David 1981, Sect. 7.6).

7. Concluding Remarks

The two large multiauthored volumes (Balakrishnan and Rao 1998a, 1998b) cover many more topics in both theory and applications of order statistics than can even be mentioned here. Additional applications include *ranked set sampling* and especially the use of *median*, *order-statistic*, and other *filters* in signal and image processing.

Bibliography

Arnold B C, Balakrishnan N, Nagaraja H N 1992 *A First Course in Order Statistics*. Wiley, New York

Balakrishnan N, Rao C R (eds.) 1998a *Handbook of Statistics 16, Order Statistics: Theory & Methods*. Elsevier, Amsterdam

Balakrishnan N, Rao C R (eds.) 1998b *Handbook of Statistics 17, Order Statistics: Applications*. Elsevier, Amsterdam

Barnett V, Lewis T 1994 *Outliers in Statistical Data*, 3rd edn. Wiley, New York

David H A 1981 *Order Statistics*, 2nd edn. Wiley, New York

David H A, Nagaraja H N 1998 Concomitants of Order Statistics. In: Balakrishnan N, Rao C R 1998a, (eds.) *Handbook of Statistics 16, Order Statistics: Theory & Methods*. Elsevier, Amsterdam, pp. 487–513

Galambos J 1987 *The Asymptotic Theory of Extreme Order Statistics*, 2nd edn. Krieger, Malabar, FL

Hahn G J, Meeker W Q 1991 *Statistical Intervals: A Guide for Practitioners*. Wiley, New York

Harter H L, Balakrishnan N 1996 *CRC Handbook of Tables for the Use of Order Statistics in Estimation*. CRC Press, Boca Raton, FL

Harter H L, Balakrishnan N 1997 *Tables for the Use of Range and Studentized Range in Tests of Hypotheses*. CRC Press, Boca Raton, FL

Hoaglin D C, Mosteller F, Tukey J W (eds.) 1983 *Understanding Robust and Exploratory Data Analysis*. Wiley, New York

Lloyd E H 1952 Least-squares estimation of location and scale parameters using order statistics. *Biometrika* **38**: 88–95

Nelson W 1982 *Applied Life Data Analysis*. Wiley, New York

Pearson E S, Hartley H O (eds.) 1970/1972 *Biometrika Tables for Statisticians*, Vol. 1, 3rd edn., Vol. 2. Cambridge University Press, Cambridge, UK

Sarhan A E, Greenberg B G (eds.) 1962 *Contributions to Order Statistics*. Wiley, New York

H. A. David

Ordered Relational Structures

Ordered relational structures are the working place for representational measurement theory. If one thinks of measurement theory as a kind of interface between empirically observed phenomena and abstract structures which one can work on mathematically, a common language for both is needed. In this context the objects under scrutiny are just elements of a set and the observed interrelations (at least those which are deemed relevant by some scientific question) are modeled as abstract relations between these elements. In the investigation of such structures nobody is interested in the actual meaning attached to the object and their relations. This approach is to some extent similar to the development of geometry, where one

works on points, lines, planes, etc., without ever defining these concepts. The only thing provided is a set of axioms which these objects satisfy. (See the entry *Axiomatic Theories* for a deeper description of this approach.) Furthermore, the theory must take care of the process of mapping one structure into another while keeping its essential features intact.

Many of the phenomena for which one needs measurements exhibit some kind of dominance, preference, or ordering. Therefore, most of the relational structures which are used in measurement theory carry some kind of abstract ordering relation (linear, weak, or partial orders).

In short, an 'ordered relational structure' consists of a base-set, here always denoted by A, an order relation on A, and possibly further relations on A. It is the foundation on which representational measurement theory is built. Sometimes A is the Cartesian product of other sets, sometimes it is the union of disjoint sets and some of the relations are between these sets (see Sect. 1 for definitions). While the investigation of one such structure may be a part of mathematics, in applications in the behavioral and social sciences the common frame consists of at least two relational structures, one to describe in a qualitative way the investigated part of reality and the other a formal, mathematical, or numerical structure that is well understood or that can be investigated by means of mathematics, logic, or computation. Such a development can yield results relevant to the empirical subject under scrutiny if the investigator succeeds in faithfully mapping the qualitative structure into the formal structure.

The requirements for this process, i.e., building useful structures and mapping them 'faithfully' into one another are discussed in the sequel and in the article on the *Measurement, Representational Theory of*.

1. Relations, Functions, Operations

Let A and B be two sets. The 'Cartesian product' of these sets is denoted by $A \times B$ and consists of all ordered pairs (a, b) with $a \in A$ and $b \in B$. Any subset $R \subseteq A \times B$ is called a 'binary relation' between A and B. The case $A = B$ is of particular interest; in this situation R is called a 'relation' on A. Clearly, this construction can easily be generalized to a k-ary relation, denoting a subset of the Cartesian product of k sets, i.e., $R \subseteq A_1 \times A_2 \times \ldots \times A_k$. If in this case all sets A_i equal A, then the Cartesian product is denoted as A^k and R is a k-ary relation on A. This broad concept is often used to describe parts of mathematics, mostly with further restrictions and properties on R, but it is also possible to use it in everyday situations. For example, let A be a set of persons and let $(a, b) \in R$ denote the fact that a and b have the same birthday. Or if A is the set of participants in a tournament and $(a, b) \in R$ denotes the result 'a beats b', then R is an encoding of the outcome of this tournament.

We now classify the possible relations. Among the 'binary relations' there is the prominent class of 'order relations' (see *Partial Orders* for definitions). In this article we only consider weak and linear orders, denoted by $\precsim$. A 'weak order' satisfies by definition 'transitivity,' i.e., $a \precsim b$ and $b \precsim c$ implies $a \precsim c$, and 'completeness,' i.e., for all $a, b \in A$ holds $a \precsim b$ or $b \precsim a$. A 'linear order' is a weak order satisfying 'antisymmetry,' i.e., for all $a, b \in A$ holds, if $a \precsim b$ and $b \precsim a$ then $a = b$.

Now we are ready to define an 'ordered relational structure.' It consists of a nonempty set A, a weak order $\precsim$ on A, and possibly more relations R_i with $i \in I$ for some index set I. In the sequel an ordered relational structure is denoted by

$$\mathscr{A} = (A, \precsim, R_i, i \in I)$$

If R_i is a k_i-ary relation, then $(k_i, i \in I)$ is the *type* of the structure.

Closely related to order relations are 'equivalence relations.' Such a relation $\sim$ satisfies by definition 'reflexivity,' i.e., $a \sim a$, 'symmetry' i.e., $a \sim b$ implies $b \sim a$, and transitivity. A set on which an equivalence relation is defined can be partitioned into disjoint subsets of equivalent elements. These subsets are called 'equivalence classes.' Often one can reduce a structure by replacing the base set A with the set of its equivalence classes.

There are a few operations on relations that can be employed to generate new relations. We define them only for binary relations on a set A. Thus, in what follows $R, S \subseteq A \times A$.

$$R^*: = \{(b, a); (a, b) \in R\}$$
$$R \circ S: = \{(a, b); \exists c \in A\, (a, c) \in R, (c, b) \in S\}$$
$$R^2: = R \circ R$$
$$R^n: = R \circ R^{n-1}$$

Thus, if $(a, b) \in R$ means 'a beats b in a tournament' then $(a, b) \in R^*$ means 'a is beaten by b,' and aR^2b means 'a has beaten someone who in turn has beaten b.'

Mappings between two sets A, B are a particular kind of binary relation. A relation f between A and B is a 'mapping' (*or* 'function),' when, for all $a \in A$ there is exactly one $b \in B$ such that $(a, b) \in f$. In this case one writes $f(a) = b$ rather then $(a, b) \in f$. The comprehensive notation is

$$f: A \to B, \quad a \mapsto b = f(a)$$

A is called the 'domain' and B the 'range' of f. If $b = f(a)$ then b is the image of a (under f) and a is a preimage of b (there may be more than one). For a subset

A' of A the set of images of all $a \in A'$ is denoted by $f(A')$. In particular $f(A) = \{b \in B$; there is an $a \in A$ such that $f(a) = b\}$. A mapping is 'surjective' if $f(A) = B$ and 'injective' if $f(a_1) = f(a_2)$ implies $a_1 = a_2$. Thus, an injective function allows at most one pre-image of each element in its domain. An injective and surjective function is called 'bijective' or a 'bijection.' Bijective functions can be 'inverted;' this concept denotes the function f^{-1}: $B \to A$ such that $f^{-1}(b) = a$ if $f(a) = b$.

Sometimes bijections f: $A \to A$ are called 'permutations.' For instance, when dealing with finite sets this terminology is quite usual. Under these bijections there is one which is particularly simple, the identity mapping, denoted by id_A. It is defined by id_A: $A \to A$ and $\mathrm{id}_A(a) = a$ for all $a \in A$.

Functions can be 'concatenated,' i.e., if f: $A \to B$ and g: $f(A) \to C$ then $g \circ f$: $A \to C$ by setting $g \circ f(a) := g(f(a))$. Note that the order of f, g in $g \circ f$ is important—$f \circ g$ is not always defined and if it is, e.g., if $A = f(A) = C$, then in general $f \circ g \neq g \circ f$. One has $f \circ \mathrm{id}_A = f$ and $\mathrm{id}_{f(A)} \circ f = f$. If f is bijective then $f^{-1} \circ f = \mathrm{id}_A$ and $f \circ f^{-1} = \mathrm{id}_B$.

If the respective conditions on the domains and ranges are met then the $\circ$-operation of functions is 'associative,' i.e., $h \circ (g \circ f) = (h \circ g) \circ f$. Thus, the brackets can be omitted, and we simply write $h \circ g \circ f$.

Mathematical operations such as addition of real numbers are special functions. A '(binary) operation' $\bullet$ on A is a function.

$$\bullet: A \times A \to A \quad (a, b) \mapsto a \bullet b$$

Thus, it is a special kind of ternary (i.e., 3-ary) relation. It is a 'partial' operation if its domain is a proper subset of $A \times A$. Examples of structures with operations frequently occur in measurement theory under the name 'extensive' or 'positive concatenation' structures (see *Measurement, Representational Theory of*).

An operation is 'associative' if $a \bullet (b \bullet c) = (a \bullet b) \bullet c$. Sometimes the notion of 'weak associativity' is needed. If $a \sim b$ denotes $a \precsim b$ and $b \precsim a$ then the operation $\bullet$ is weakly associative if $a \bullet (b \bullet c) \sim (a \bullet b) \bullet c$. As already mentioned, weak association might arise in a structure $(A, \precsim, \bullet)$, where $\precsim$ is a weak order.

It should be noted in passing that the relation $\sim$ defined in the previous paragraph is an equivalence relation. It is often used if one wants to replace a given weak order by a linear order (see *Partial Orders*).

2. Examples of Ordered Relational Structures

2.1 Weights

Consider a collection of weights and let $a \precsim b$ stand for: If a is put in one pan of a balance and b in the other, then the one containing b remains in place or deflects the pointer to b's side. Let $a \bullet b$ denote putting a and b in one pan. Let A denote the set of all weights and of all combinations that can be formed by putting two or more of the weights in one pan. It can be assumed, with reasonable accuracy that $(A, \precsim, \bullet)$ is an ordered relational structure. Furthermore, the operation $\bullet$ will be approximately weakly associative and positive, i.e., $a \prec a \bullet b$ for all $b \in A$ ($x \prec y$ is short for $x \precsim y$ and not $y \precsim x$). Still more, it will satisfy 'commutativity' $a \bullet b = b \bullet a$.

2.2 Guttman Scaling

Let A be a set of testees and B a set of items to be solved by each $a \in A$. For $a \in A$ and $b \in B$ let aRb denote that person a fails to solve item b correctly. Now assume that the relation R (which is between A and B) satisfies

$$aRb,\ a'Rb,\ \text{not } a'Rb' \quad \text{implies} \quad aRb'$$

The system (A, B, R) is not an ordered relational structure in the sense of the previous section because R, being a biorder (see *Partial Orders*), is not a weak order on one of the sets A, B. However, we can easily construct an ordered relational structure from it. Let $C := A \cup B$ and on C define: for $a, a' \in A$ $a \precsim a'$ if person a' solves at least as many items as does person a, and for $b, b' \in B$ define $b \precsim b'$ if b' is solved by at most as many persons as b. Finally define $a \precsim b$ if aRb, and $b \precsim a$ if not aRb. Then it can be shown that $\precsim$ is a weak order on C; thus, $(C, \precsim)$ is an ordered relational structure.

2.3 Several Orders on One Set

Let S be a collection of different societies. Assume they can be weakly ordered with respect to social mobilization (relation K) and social frustration (relation L) of its members, and also with respect to economic development (relation M).

This is a case of a structure $\mathscr{A} := (A, K, L, M)$ where K, L, M are all weak orders on A. Let the weak orders satisfy the following compatibility relations: For all $a, b \in A$

$$\text{if } aKb,\ bLa \quad \text{then} \quad aMb \tag{1}$$

$$\text{if } aKb,\ bMa \quad \text{then} \quad aLb \tag{2}$$

$$\text{if } aLb,\ aMb \quad \text{then} \quad aKb \tag{3}$$

Suck (1990) has shown that such a structure can be transformed to an ordered relational structure on a Cartesian product which turns out to be a conjoint measurement structure satisfying independence (see *Measurement Theory: Conjoint*). As an application of such a structure Suck and Getta (1994) gave an

axiomatization and representation of Coombs's portfolio theory in which A is a set of gambles (wins and losses with certain probabilities), K is an ordering of these gambles according to preference to play such gambles by subjects, L is an ordering by expected value, and M is an ordering by perceived riskiness of the gambles. The possibility to associate this situation with conjoint measurement proved very helpful in analyzing the structure and data.

Wille and Wille (1993) considered similar relational structures. They investigated the possibility of homomorphically mapping a structure satisfying Eqns. (1), (2), and (3) into the reals without the detour via conjoint measurement. They related their investigation to the 'Huntington controversy' arising from Huntington (1968) who presented 'equations' such as:

$$\frac{\text{Social mobilization}}{\text{Economic development}} = \text{Social frustration} \quad (4)$$

This use of 'equations' was severely criticized, mainly with the argument that the variables of the equation are not measurable entities (see *Mathematical Psychology*). Formulating the structure in the above way can provide the means to solve this controversy. Either applying conjoint measurement in the sense of Suck (1990) or the results of Wille and Wille (1993) would yield a framework in which a decision on the validity of Eqn. (4) is possible.

3. *Homomorphisms, Isomorphisms*

It is often useful and even necessary to consider several relational structures at the same time. In this case it is important to describe how similar these structures are, and how one can go from one structure to the next or—more intuitively—how one can transfer knowledge from one structure to another. Examples of this technique occur in measurement theory where usually one structure has as its base set (a subset of) the reals, or more generally, when there is one structure which is thoroughly investigated and all or many of its properties are well known. In such cases it is without doubt helpful to compare such a structure to the one under empirical investigation and to point out the manner in which the two structures resemble each other.

Let $\mathscr{A} = (A, R_i, i \in I)$ and $\mathscr{B} = (B, S_j, j \in J)$ be two relational structures. Then a mapping $\varphi: A \to B$ is called a 'homomorphism' between $\mathscr{A}$ and $\mathscr{B}$ if the structures are of the same type $(k_i, i \in I)$ and for all $i \in I$ and for all $(a_1, \ldots, a_k k_i) \in A^{ki}$:

$$(a_1, \ldots, a_{k_i}) \in R_i \text{ iff } (\varphi(a_1), \ldots, \varphi(a_{k_i})) \in S_i \quad (5)$$

$\mathscr{A}$ and $\mathscr{B}$ are said to be 'homomorphic' if a homomorphism exists between A and B, If the homomorphism is a bijection, then it is an 'isomorphism.' Isomorphic structures are, in a sense, identical, although their base sets A and B can be quite different, but each property of $\mathscr{A}$ expressible in terms of the relations R_i corresponds to a property of $\mathscr{B}$ expressible in terms of the relations S_j.

If the structures $\mathscr{A}$ and $\mathscr{B}$ are both ordered, i.e., $\mathscr{A} = (A, \precsim; R_i, i \in I)$ and $\mathscr{B} = (B, \precsim'; S_i, i \in I)$ where $\precsim$ and $\precsim$ are weak orders, then the homomorphism is sometimes called 'a monotonic' function. It satisfies

$$a_1 \precsim a_2 \quad \text{iff} \quad \varphi(a_1) \precsim' \varphi(a_2)$$

If $\mathscr{A}$ is mapped homomorphically into a structure on the reals then the relation on $\mathbb{R}$ corresponding to $\precsim$ is usually, but not necessarily, the $\leqslant$ relation on $\mathbb{R}$. This situation is typical of representational measurement theory (see *Measurement, Representational Theory of*).

As an example, let $\mathbb{R}^+$ denote the positive reals and let $\mathscr{A}: = (\mathbb{R}^+, \leqslant, \cdot)$ and $\mathscr{B}: = (\mathbb{R}, \leqslant, +)$. Then $\varphi(x): = \log x$ is an isomorphism between these numerical structures.

As another example, let A be a set of players in a tournament and let $\precsim$ denote a ranking of these players with ties allowed. Here $\precsim$ is a weak order but in general it is not linear because of the possibility of ties. If $\mathscr{B}: = (\mathbb{N}, \leqslant)$ with $\mathbb{N}$ the natural numbers, then the mapping φ which associates his or her place in the ranking with each player, is a homomorphism but not an isomorphism.

Homomorphisms can be concatenated provided the range of the first and the domain of the second coincide. More precisely, let $\mathscr{A} = (A, \precsim, R_i, i \in I)$, $\mathscr{B} = (B, \precsim, S_i, i \in I)$, and $\mathscr{C} = (C, \precsim, T_i, i \in I)$ be homomorphic structures with the homomorphisms

$$\varphi: A \to B \quad \text{and} \quad \psi: B \to C$$

where φ is surjective. Then $\mathscr{A}$ and $\mathscr{C}$ are homomorphic by the homomorphism $\xi: = \psi \circ \varphi$. This fact is of particular importance when the structures are all identical; this case is considered in Sect. 4.

4. *The Automorphism Group of an Ordered Relational Structure*

An isomorphism which maps its domain structure onto itself is called an 'automorphism.' Let Aut $\mathscr{A}$ denote the set of automorphisms of a given structure $\mathscr{A}$. Clearly, Aut $\mathscr{A} \neq \emptyset$ because $\mathrm{id}_A \in$ Aut $\mathscr{A}$, i.e., the identity map is always an automorphism. If $\alpha, \beta \in$ Aut $\mathscr{A}$ then $\alpha \circ \beta \in$ Aut $\mathscr{A}$; furthermore, $\alpha^{-1} \in$ Aut $\mathscr{A}$. Thus (Aut $\mathscr{A}$, $\circ$) has the properties of a 'group,' i.e., a set with an operation on it which is associative, has an identity element (id_A in this case), and a right and left inverse (α^{-1}). Therefore (Aut $\mathscr{A}$, $\circ$) is called the 'automorphism group' of $\mathscr{A}$.

Modern measurement theory exploits the properties of the automorphism group very intensively. The main ideas on how this is done are now described.

The structure $\mathscr{A}$ is called 'homogeneous' if for each $a, b \in A$ there is an $\alpha \in \mathrm{Aut}\,\mathscr{A}$ such that $\alpha(a) = b$. The group and the structure are called 'n-homogeneous' if for any two n-tuples $a_1 \prec \ldots \prec a_n$ and $b_1 \prec \ldots \prec b_n$ there is at least one automorphism α such that $\alpha(a_i) = b_i$. Similarly, they are called 'n-unique' if for any two n-tuples $a_1 \prec \ldots \prec a_n$ and $b_1 \prec \ldots \prec b_n$ there is at most one automorphism α such that $\alpha(a_i) = b_i$. n-uniqueness can be reformulated as: if two automorphisms agree on n points then they are identical.

If M is the maximal homogeneity and N the minimal uniqueness of a structure, then the 'scale type' of $\mathscr{A}$ (or of Aut $\mathscr{A}$) is denoted by the pair (M, N). Clearly, one has $M \leqslant N$. $N = \infty$ means that there is no minimal degree of uniqueness. A structure with $N < \infty$ is called 'finitely unique.'

One of the most important results of measurement theory in the 1980s is the following theorem developed by Narens (1981) and Alper (1985, 1987).

Theorem 1 (Alper/Narens Theorem)

Let $A = \mathbb{R}$ with $\mathscr{A}$ homogeneous and finitely unique. Then it is of scale type (1, 1), (1, 2) *or* (2, 2).

Type (1, 1) is a ratio scale and type (2, 2) an interval scale (see Sect. 2 of *Measurement, Representational Theory of*)

The importance of this theorem for measurement theory is discussed in the articles on measurement theory (see *Measurement, Representational Theory of*). Suck (2000) proved a partial generalization for ordered structures with an arbitrary base set. Essentially, his result shows that for scale type (N, N), in many, though not all cases, the Alper/Narens Theorem remains true if the base set A of the structure is different from $\mathbb{R}$.

5. *Order Topology*

If one wants to exploit the power of the continuity concept for the investigation of ordered relational structures, then one must construct a 'topology' on A which is compatible with the order relation and, if possible, with the other relations as well.

A topology of A is a collection $\mathscr{T}$ of subsets of A, called the 'open sets,' such that the following three conditions are satisfied:

(T1) $\emptyset \in \mathscr{T}$ and $A \in \mathscr{T}$.

(T2) If $\mathscr{S} \subseteq \mathscr{T}$ then $\cup_{T \in \mathscr{S}} T \in \mathscr{T}$.

(T3) If $\mathscr{S}$ is a *finite* subset of $\mathscr{T}$ then also $\cap_{T \in \mathscr{S}} T \in \mathscr{T}$

The structure $(A, \mathscr{T})$ is a 'topological space.' For a development of the theory of topological space see Kelley (1955) or Dugundji (1966). The most important topological space is constructed on $\mathbb{R}$ by defining as open all subsets of $\mathbb{R}$ which are arbitrary unions of open intervals. This construction can be carried out in any ordered structure and is called the 'order topology' (see below). Thus, any $\mathscr{A}$ of the kind considered in this article can be regarded as a topological space endowed with the order topology.

In topological spaces the well-known concepts of convergence, limits, and continuity can be defined, and in this way, far reaching generalizations of familiar theorems of analysis are available in the theory of ordered relational structures.

We give the topological definition of 'continuity.' Let $(A, \mathscr{T}_A)$ and $(B, \mathscr{T}_B)$ be topological spaces. A function $f: A \to B$ is 'continuous' if for all $U \in \mathscr{T}_B$ the preimage under f, i.e., the set $f^{-1}(U) := \{a \in A; f(a) \in U\}$ is in $\mathscr{T}_A$.

Next, we are going to define the 'order topology' of an ordered structure $\mathscr{A}$. We assume that A does not contain a maximal or a minimal element. We define: for all $a, b \in A$

$$(a, b) := \{x \in A;\ a \prec x \prec b\}$$

$$[a, b] := \{x \in A;\ a \precsim x \precsim b\}$$

$$(0, a) := \{x \in A;\ x \prec a\}$$

$$(a, \infty) := \{x \in A;\ a \prec x\}$$

and call these sets intervals (with respect to $\precsim$). Note that the symbols 0, ∞ do not denote elements of A.

The intervals $[a, b)$, $(a, b]$, $(0, a]$, and $[a, \infty)$ are defined analogously. (a, b), $(0, a)$, and (a, ∞) are referred to as 'open intervals.' The open intervals are used as a 'basis' for the order topology denoted by $\mathscr{T}_{\precsim}$, i.e., $\mathscr{T}_{\precsim}$ is the collection of those subsets of A which can be generated by union from an arbitrary set of open intervals.

The construction of the order topology for an ordered relational structure is useful when the other relations of $\mathscr{A}$ are in some sense compatible with $\mathscr{T}_{\precsim}$; e.g., when there is an operation among the R_i then it should be continuous.

If $\precsim$ is a linear order, $\mathscr{T}_{\precsim}$ is easily seen to separate points, i.e., for all distinct $a, b \in A$ there are open and disjoint sets U, V such that $a \in U$ and $b \in V$. In particular, consider $a \prec b$ then, if $z \in (a, b)$ exists, $(0, z)$ and (z, ∞) are such sets; if $(a, b) = \emptyset$, then $(0, b) \cap (a, \infty) = \emptyset$ and $a \in (0, b)$ and $b \in (a, \infty)$. Technically, such a space is called a 'Hausdorff space.'

A bijective and continuous map $f: A \to B$ between topological spaces $(A, \mathscr{T}_A)$ and $(B, \mathscr{T}_B)$ is a *homeomorphism* if its inverse is also continuous. Topology is mainly concerned with properties of spaces which are shared by all spaces that are homeomorphic to a given one. These are called 'invariants' of the space.

Specializing this concept on order topologies, we observe: Isomorphic linear orderings are homeo-

morphic with respect to the order topologies and an isomorphism is a homeomorphism. Formulating this result as a theorem one gets:

Theorem 2

Let $(A, \precsim_A)$ *and* $(B, \precsim_B)$ *be two linear orderings and* $f: A \to B$ *strictly monotonic and surjective. Then* f *is a homeomorphism with respect to the order topologies induced by* $\precsim_A$ *on* A *and* $\precsim_B$ *on* B.

It follows from Theorem 2 that automorphisms of a structure $\mathscr{A} = (A, \leqslant, R_i, i \in I)$ are homeomorphisms of $(A, \mathscr{T}_{\precsim})$ onto itself.

Summarizing, we note that the order topology carries a lot of information about an ordered relational structure. Despite this fact, it has not been used very much in measurement theory until very recently. Wakker (1988) pointed out that in many cases the algebraic conditions used by Krantz et al. (1971) and others are more general than topological ones. As long as the use of topological concepts is restricted to the order topology no extra conditions need be imposed. In some modern investigations the order topology could be put to use as a fruitful tool. For instance, even the topological properties of the automorphism group can play an important role (Suck 1998, 2000).

Finally, it should be mentioned that the power of topological methods shows up in infinite structures (although the definitions are not restricted to this case). Some approaches to measurement try to avoid such structures, especially the continuum. As an example, we refer to Sommer and Suppes (1997).

6. *Completion of Structures*

The article on partial orders introduces the 'McNeille completion' of a partial order (see *Partial Orders*). This construction is also useful for ordered structures in the sense of the present article. The property of existence of infimum and supremum is equivalent to the existence of a cut element for each 'Dedekind cut.' It is called 'Dedekind completeness' (cf. Narens 1985, p. 16).

If $\mathscr{A}$ is an ordered relational structure then the McNeille completion of $(A, \precsim)$ is 'Dedekind complete.' However, a major problem arises if one wants to extend the other relations of $\mathscr{A}$, i.e., the R_i for $i \in I$, to the complete superset of A. There may exist no extension at all, or too many—more than one is already too many; Narens (1985) and Cohen (1988) investigated this problem for so-called positive concatenation structures. The results are discussed in Luce et al. (1990) in Chap. 19.4.

In the context of the order topology Dedekind completeness is equivalent to 'connectedness' in the topological sense, i.e., there are no two disjoint open sets U, V such that $A = U \cup V$. Ovchinnikov (1996) proved the surprising result that a homogeneous structure $(A, \precsim)$ with a linear order $\precsim$ has either a connected order topology or each of its maximal connected subsets consist of one element only. This result has far reaching consequences for ordered relational structures and their automorphism groups. It is at the heart of Suck's (2000) generalization of the Alper/Narens result. Thus, if a completion is feasible then many properties of the structure follow from this theory.

See also: Axiomatic Theories; Factor Analysis and Latent Structure: Overview; Luce's Choice Axiom; Mathematical Models in Philosophy of Science; Measurement Theory: Conjoint; Measurement Theory: History and Philosophy; Partial Orders

Bibliography

Alper T M 1985 A note on real measurement structures of scale type $(m, m+1)$. *Journal of Mathematical Psychology* **29**: 73–81

Alper T M 1987 A classification of all order-preserving homeomorphism groups of the reals that satisfy finite uniqueness. *Journal of Mathematical Psychology* **31**: 135–54

Cohen M 1988 Dedekind completion of positive concatenation structures: necessary and sufficient conditions. *Journal of Mathematical Psychology* **32**: 64–71

Dugundji J 1966 *Topology*. Allyn and Bacon, Boston

Huntington S P 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, CT

Kelley J L 1955 *General Topology*. Van Nostrand, New York

Krantz D H, Luce R D, Suppes P, Tversky A 1971 *Foundations of Measurement*. Academic Press, New York, Vol. 1

Luce R D, Krantz D H, Suppes P, Tversky A 1990 *Foundations of Measurement*. Academic Press, New York, Vol. 3

Narens L 1981 On the scales of measurement. *Journal of Mathematical Psychology* **24**: 249–75

Narens L 1985 *Abstract Measurement Theory*. MIT Press, Cambridge, MA

Ovchinnikov S 1996 Means on ordered sets. *Mathematical Social Sciences* **32**: 39–56

Sommer R, Suppes P 1997 Dispensing with the continuum. *Journal of Mathematical Psychology* **41**: 3–10

Suck R 1990 Conjointness as a derived property. *Journal of Mathematical Psychology* **34**: 57–80

Suck R 1998 A qualitative characterization of the exponential distribution. *Journal of Mathematical Psychology* **42**: 418–31

Suck R 2000 Scale type (N, N) and an order based topology induced on the automorphism group. *Journal of Mathematical Psychology* **44**: 582–99

Suck R, Getta A 1994 A reaxiomatization of portfolio theory. *Journal of Mathematical Psychology* **38**: 115–27

Wakker P 1988 The algebraic versus the topological approach to additive representations. *Journal of Mathematical Psychology* **32**: 421–35

Wille R, Wille U 1993 On the controversy over Huntington's equations: When are such equations meaningful. *Mathematical Social Sciences* **26**: 173–80

R. Suck

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Organization: Informal

The term 'informal organization' has been used in varying ways within organization theory, industrial sociology, and managerial writing. What the various usages have in common is reference to aspects of patterned activity in work organizations that are not part of the formally or officially designed rules, relationships, and procedures that constitute the 'formal' dimension of the organization. Informal organization is seen variously as something that works, or might be made to work, in conjunction with the formal organization, something that can exist with a degree of independence from it or something that can work against the formal aspects of the organization—especially when it takes the form of deviant behavior that resists or defies managerial authority. In spite of these differences of emphasis, all of the usages attend to the fact that one can neither understand nor intervene in what 'actually happens' in organizational settings without looking beyond what the official blueprint, organizational chart, or corporate rule book suggests happens, or should happen.

1. *The Two-sidedness of Organizational Life: The Basic Issue*

Since its first appearance in the 1930s the concept of 'informal organization' has played an important part in the history of both the primarily analytical social scientific study of work organization and more 'applied' writing intended to guide the practices of managers. The basic issue that the concept tackles remains of central importance to organizational analysis in spite of the fact that the notion of informal organization is not always used explicitly.

The basic issue that the distinction between formal and informal organization was devised to tackle is that of what might be called the fundamental 'two-sidedness' of the modern work organization. On the one hand, the work organization is a set of rules, roles, and procedures based upon the legal-rational principles of bureaucracy which is given formal and official expression through such devices as procedural manuals, rule books, and organization charts (see *Bureaucracy and Bureaucratization*). This aspect of the organization can be seen as existing separately from the human beings who are employed to operate what amounts to a scheme for an organizational machine. In this respect, the organization can be said to continue in existence regardless of the personalities who come and go as occupiers of organizational positions. However, this 'machine' can do no work until it employs human beings. Yet employees cannot simply be built into the social machine drawn up in the organizational blueprint as human cogs or injected into the system as a living fuel to power the device. Employees will inevitably bring to the organization their own wants, interests, values, assumptions, understandings, and sentiments. And they will seek to organize certain aspects of their working lives for themselves.

Individuals' acceptance of employment in the organization entails a willingness to follow instructions and fulfill tasks that are constitutive of the organizational blueprint. But this acceptance is strictly bounded and engagement in the fulfilling of corporate tasks is unlikely to satisfy all the wants and needs of every individual or to engage every aspect of their energy and humanity. A second side of organizational life thus comes into being the moment living human beings are engaged to put the first side into action. A new facet of organizational existence appears as people develop relationships, practices, pecking orders, games, languages, jokes, and stories to express or defend their 'humanness' and to pursue projects and purposes of their own—ones which sometimes fit with the priorities of the first side of organizational existence and sometimes do not.

The first academic attempt to deal systematically with this two-sidedness of work organizations, in the light of empirical social science investigation, was made by researchers at Harvard University in the USA in the 1930s. And the first attempt to do the same thing, in the light of social-science informed reflection upon executive experience, was done more or less at the same time by an individual associated with the same institution. Both attempts involved the conceptual device of informal organization. These will now be looked at.

2. *The Discovery of Informal Organization: Understanding Factory Life*

The study *Management and the Worker* written by Roethlisberger and Dickson (1939) was not the first account of the studies carried out by academics from Harvard University and members of the Western Electrical Company at that company's Hawthorne plant. But it is the most systematic report and analysis of these studies and has become a classic work in both industrial sociology and organizational analysis. It has also been drawn upon massively in management writing and continues to play a part in the education of managers. In their analysis of the Hawthorne plant, Roethlisberger and Dickson first distinguish its social organization from its technical organization. But they then make the key distinction we are concerned with here, within this notion of the factory's social organization. The social organization of the industrial plant, they say, is 'in part formally organized.'

> This formal part of the social organization includes the systems, policies, rules, and regulations of the plant which express what the relations of one person to another are

> supposed to be in order to achieve effectively the task of technical production (Roethlisberger and Dickson 1939).

'In short,' say Roethlisberger and Dickson, 'the patterns of human interactions, as defined by the systems, rules, policies, and regulations of the company, constitute the formal organization.' These are the patterns which are 'supposed to be' the ones which prevail. But their factory studies taught the researchers to challenge what they argue is a common assumption: that the 'organization of a company corresponds to a blueprint plan or organization chart.' 'Actually,' they say, 'it never does.' The study revealed what Roethlisberger later referred to as 'the emergent patterns of behavior that evolved within this formal organization' (Roethlisberger 1968). This informal organization was made up of:

> the practices, values, norms, beliefs, unofficial rules, as well as the complex network of social relations, membership patterns and centers of influence and communication that developed within and between the constituent groups of the organization under the formal arrangements but that were not specified by them (Roethlisberger 1968).

It is explicitly recognized in *Management and the Worker* that the informal organization can have positive as well as negative implications for the working of the formal organization and can be said to 'be a necessary prerequisite for effective cooperation' (Roethlisberger and Dickson 1939). The famous 'bank wiring observation room' study highlighted in the book's discussion of informal organization, however, is an example of where an informal organization developed 'in opposition to the formal organization.' The norms and practices developed by the observed group of workers effectively undermined the incentive system that was central to the formal organization of the workshop. Just what the relationship is between the formal and informal organization in any particular case, is 'an empirical question to be decided by the results of the investigation and not before it' (Roethlisberger 1968). But this is not to treat formal and informal organization as 'two separable things' within a plant's social structure. 'Formal' and 'informal' were developed in the study, Roethlisberger points out, 'as analytical concepts to refer to two different dimensions of the territory.'

3. *The Discovery of Informal Organization: Understanding the Executive Process*

Barnard's (1938) study *The Function of the Executive* included his development of a concept of informal organization influenced by the same intellectual environment which shaped the thinking of the Hawthorne researchers. The influence of the social theories of the European thinkers Durkheim and Pareto, mediated in part by the biologist Henderson and the sociologist Parsons, is apparent in both of these intellectual engagements with the practical problems of work organization and executive processes (see *Organization: Overview*). Barnard pointed to the work carried out at Hawthorne as the only work he was aware of which looked at informal organization. This, however, was 'at the production level of industrial organizations' (Barnard 1938). Barnard himself was concerned to look at its functioning at the level of the organization as a whole and, especially, to put the shaping of it at the center of the executive's responsibilities. He noted the paradox that although executives tend to speak of their work primarily with reference to the 'first side' of organizing—the formal scheme of work tasks—and, indeed, deny the importance of the second, informal side, they nevertheless intuitively recognize its vital significance. 'Learning the organizational ropes in most organizations,' he points out, 'is chiefly learning who's who, what's what, why's why, of its informal society.' It is consequently on this side of the organization that the executive needs to focus their attention.

Barnard can be read as arguing that the executive's job is primarily to shape the informal organization, in spite of the fact that the interactions of which it consists 'are based on personal rather than on joint or common purposes.' The executive shapes the informal organization so that it carries out three major functions necessary for the enterprise's continuation: communication; the maintenance of cohesiveness 'through regulating the willingness to serve and the stability of objective authority'; the retaining of feelings of 'personal integrity, of self respect, of independent choice.' In effect, then, executives work to bring the two sides of organizational existence together by attending to the 'hearts and minds' of employees. Conflict, dissent, and misunderstandings are thus overcome as they work to achieve 'synthesis in concrete action of contradictory forces, to reconcile conflicting forces, instincts, interests, conditions, positions, and ideals.'

4. *The Development of Systems Thinking and an Engagement with Weber*

As we shall see below, the type of thinking about executive or managerial work emerging from Harvard in the 1930s has immense relevance to later managerial writing. But it also opened up the possibility for more academic students of organizations to analyze work organizations in 'systems' terms. Roethlisberger was explicit, for example, in saying that the 'notion of a natural social system' should be understood as 'underlying' the 'way of thinking' which distinguishes between formal and informal organization (Roethlisberger 1968). Subsequent theorists built upon this

early foundation of a systems view of organizations. And a further influence on moves in the systems direction was the parallel recognition of the importance of the 'second side' of organizational life made by researchers concerned to counter-balance the emphasis on formalization in Weber's seminal sociological analysis of bureaucratic organization (see *Organizations: Unintended Consequences*). Merton (1940), for example, identified the emergence of the 'bureaucratic personality' whose activities tended to undermine the functioning of the formal organization. Gouldner (1954) showed in his study of an industrial plant how an 'indulgency pattern' tends to emerge whereby supervisors gain the cooperation of subordinates on certain matters by selectively ignoring the breaking of formal rules by those workers with regard to other matters. And Blau (1955) demonstrated how officials make 'procedural adjustments' to counter potentially 'dysfunctional' aspects of the functioning of the formal bureaucracy.

5. *Continuing Attention to the Two-sideness of Organization Life in Empirical Research*

The writers who set out to produce a corrective to Weber's alleged overemphasis on the first or 'formal' side of organizational life tended not to give the formal/informal distinction a central role in their analyzes, at least in an explicit way. This was not the case, however, with the British researchers who, led by Scott, were dominant in British industrial sociology in the 1950s and early 1960s. The concept of the informal structure' of the plant was used to refer mainly to patterned relations 'based on congeniality and friendship' (Scott et al. 1956). The use of the concept of informal structure to analyze how relationships between employees outside of work influenced this and could compromise the functioning of the formal structure made a significant contribution to industrial sociology.

The notion of informal organization, it would seem, was becoming vital to social science analyzes of work organizations and the activities that went on in them, whether or not it was used explicitly. As ethnographic and, especially, participant observation studies of organizational activities began to appear, increasing attention was drawn to facets of organization life which could not be accounted for in terms of formal administrative systems. Roy's (1960) participation in shopfloor work drew attention to how his colleagues used games and rituals to shape their day and overcome the monotony that the straightforward compliance with formal procedures would create. And later research by Burawoy (1979) in the same plant showed how workplace practices which might, at first sight, seem to challenge the dominant interests which the formal arrangements of the factory were designed to serve in practice accommodated workers into the dominant patterns of interest. At the managerial level, Dalton's (1959) participative observation study showed the considerable extent to which senior organizational members organized themselves and behaved in ways which would be quite unexpected from a reading of their formal roles and duties. And Mangham and Pye's (1991) interview-based study of senior executives shows how close much of the thinking of these individuals was to Barnard's conception of the executive role and its focus on informal activities. This includes their relationships with other senior managers: 'Many stressed that informal organizing, the structuring of the top team, was as important as, if not more than, the articulation of a formal organization.'

New empirical studies of work organizations continue to show the existence of patterned behaviors and relationships which have no place in the formally designed arrangements of which organizations are still frequently said to consist. These phenomena range from ones at the top of organizations to ones at the bottom and cover such diverse activities as name calling, joking, and fiddling expenses at the one end of the spectrum to coalition-building and career competition at the other. And any of these phenomena can be found, in specific circumstances, to support, undermine, or have little direct effect on the formal aspects of the organization. Such phenomena need to be conceptualized somehow. Why not therefore bring the concepts of formal and informal organization back to the center of the stage?

6. *From Informal Organization to Unofficial Aspects of Organizing?*

One of the main reasons that the concept of informal organization has been used less directly and explicitly since the 1960s is that social scientists have tended to associate it with the Harvard publications which first presented it to the world. As was pointed out earlier, this work provided much of the foundation of systems thinking—an approach which has since been rejected by many organizational researchers. It is seen as representing work enterprises in too 'unitary' a manner, as failing to recognize the multiplicity of meanings and interests existing within the organization and as underplaying the relationships between organizations and wider patterns of political economy and social inequality. Further, the engagement of these writers with the problems faced by managers of work organizations has often been seen as an inappropriate 'taking of sides' on the part of social scientists.

Part of the reluctance to use the concept of 'informal organization,' then, has been as a result of its associations. There is, however, a stronger reason for questioning the continuing usefulness of the concept. This is one which, ironically, suggests that the Harvard

work which established the importance of looking at informal organization might have so successfully influenced subsequent understanding of the two-sidedness of organizational existence that their formal/informal distinction is now too unsubtle an analytical tool. The Harvard message that the managing of work organizations must take into account both the official blueprint of the organization and nonblueprint practices has been so well received (or, equally likely, discovered through the normal processes of intuitive management learning) that managerial practices are now incorporating much of what was once 'informal' into the 'formal' aspects of their activities.

With the managers of organizations now setting out to manipulate 'corporate cultures,' publish 'mission statements,' facilitate 'teamworking,' and 'empower' workers, there is such a blurring of the formal and the informal aspects of managerial initiative that the distinction is difficult to use when setting out to analyze the organizations in which this occurs. Managers have incorporated into the 'formal' aspect of their organizations' activities that were once 'informal.' But this does not mean that the essential two-sidedness of organizational life to which the formal/informal distinction attends is something that can be forgotten about. One useful way forward would be to adopt the alternative (but closely related) distinction between official and unofficial aspects of organizing and identify the part that each of these plays in the negotiated orders of particular organizational settings (see *Organizations: Negotiated Orders*). One no longer distinguishes between what the formal documentary 'blueprint' says about the organization and observable patterns that differ from this. Instead, one distinguishes—and examines the interplay—between the managerial aspirations represented in a whole range of official documents, claims, cultural, and mission statements and the observable patterns of belief and behavior which do not accord with what is officially claimed to 'be the case.' A classic concept in organizational and managerial research is thus built upon and revised for the effective study of contemporary organizational trends.

See also: Group Processes in Organizations; Group Productivity, Social Psychology of; Human Resource Management, Psychology of; Industrial Sociology; Management: General; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture; Organizational Culture, Anthropology of; Organizations, Sociology of; Violence and Effects on Children

Bibliography

Barnard C I 1938 *The Functions of the Executive*. Harvard University Press, Cambridge, MA

Blau P M 1955 *The Dynamics of Bureaucracy*. University of Chicago Press, Chicago

Burawoy M 1979 *Manufacturing Consent*. University of Chicago Press, Chicago

Dalton M 1959 *Men Who Manage*. Wiley, New York

Gouldner A W 1954 *Patterns of Industrial Bureaucracy*. Free Press, Glencoe, IL

Mangham I L, Pye A 1991 *The Doing of Managing*. Blackwell, Oxford, UK

Merton R K 1940 Bureaucratic structure and personality. *Social Forces* **174**: 560–8

Roethlisberger F J 1968 *Man-in-organization: Essays of F J Roethlisberger*. Belknap Press of Harvard University Press, Cambridge, MA

Roethlisberger F J, Dickson W J 1939 *Management and the Worker*. Harvard University Press, Cambridge, MA

Roy D F 1960 Banana time: Job satisfaction and informal interaction. *Human Organization* **18**: 158–68

Scott W H, Banks J A, Halsey A H, Lupton T 1956 *Technical Change and Industrial Relations: A Study of the Relations Between Technical Change and the Social Structure of a Large Steelworks*. Liverpool University Press, Liverpool, UK

T. J. Watson

Organization: Overview

Theoretical approaches to organizations vary by the extent to which organizations are viewed as rational systems—instruments rationally constructed to pursue specific goals—or natural systems—valued social systems struggling to survive. In addition, organizations are variously treated as closed systems—self sufficient, relatively insulated forms—and as open systems—constituted, influenced, and penetrated by the wider environment. Conceptions of organizations also vary by level of analysis as researchers focus variously on the behavior of individual participants within organizations, the behavior of organizations as collective actors, and the role of organizations as component units within larger populations or fields of organizations. Organizations are viewed as a construction of the modernist world, but some analysts see them as products and accompaniments of industrialization while others see them as vehicles of rationalization. The proliferation of organizations and their infusion into every arena of contemporary society, results not only from their contribution to the coordination of differentiated, specialized tasks, but also their capacity to symbolize rational procedures and provide frameworks of accountability. Conceptions of organizations and theoretical approaches change over time because organizations themselves undergo change. Organizational forms have evolved from being relatively dependent and embedded in traditional societies to becoming more independent and then most recently to becoming more interdependent with and penetrated by their context. Organizations both

increasingly absorb society and are themselves being absorbed into wider, rationalized frameworks.

Organizations are a defining feature of modern societies. Modern societies are distinguished from traditional ones, in part, by the numbers and variety of organizations in existence and their utilization to support collective action in every sector. Organizations are not simply distinctive contexts within which individuals carry out coordinated activities; they are themselves collective actors, given legal recognition, and constituted so as to own and dispose of property, enter into binding contracts, and mobilize effort in the pursuit of specified goals. Organizations act to increase the reliability of behavior, but in so doing, they often lose flexibility. They are also able to provide greater accountability, maintaining records of how things are done and by whom, but in doing so, may become overly procedural and legalistic. Today's organizations, operating in a highly rationalized and dynamic environment, face different challenges than early organizations confronting a more traditional and stable context. Conceptions of organizations, theoretical perspectives, and favored levels of analysis undergo change, in part because organizations and organizing processes themselves have changed.

1. *Defining Organization*

Social theorists have long distinguished the more specific subtype, organization, from the more generic phenomenon of social organization or, even more broadly, social structure. Social life, even its most primitive forms, exhibits social organization in the sense that there exist patterned regularities in social interaction and in belief systems, each of which influences the other. Earlier theorists stressed the twin, interdependent components of normative structure, consisting of norms, values, and beliefs, and behavioral structure, consisting of recurring activities and relationships among individuals. Later theorists (Giddens 1984) emphasize schemas, rules, and procedures providing guidelines for practice, and resources, consisting of both human capacities and nonhuman tools and material things of value. The later formulations represent an advance in their recognition of the dual character of social structure, which is regarded as both medium and outcome of social action, social structure both shaping and being shaped by on-going social activities.

The most common definitional approaches to delineate organization from social organization or social structure employ substantialist formulations, stressing the priority of things or entities (Emirbayer 1997). Two subtypes are identified: self-action, conceiving of things as acting independently, under their own powers, and interaction, in which the attributes of entities are shaped by encounters with other entities. In contrast to these substantialist approaches, more recent relational perspectives have developed that stress process and dynamics as primary rather than entities. Each of these approaches is discussed and illustrated.

1.1 *Substantialist Self-action Definitions*

The most influential definitions of organizations emphasize both the ability of organizations to act independently—the self-action conception—and the features that distinguish them from other social structures. Max Weber, who provided a theoretical foundation for much subsequent work, was among the first social theorists to recognize and attempt to characterize the identifying characteristics of organizations. He developed a series of nested definitions as he endeavored to distinguish organizations from more generic forms. Weber (1922/1968, pp. 26–53) begins by defining a legitimate order as a social relation oriented to a normative framework or set of maxims viewed in some manner as obligatory. (Variations in the nature of these norms gives rise to Weber's justly famous typology of authority into traditional, charismatic, and rational-legal orders.) Social relations vary in the extent to which they are 'closed,' limiting or excluding access to outsiders. Organizations (*Verband*) are social relationships that are closed, access being limited by the actions of specific individuals, a chief or director often assisted by an administrative staff. In short, an organization is a closed group with a differentiated 'organ' to enforce its boundaries. Enterprises (*Betreib*) are organizations that have developed a system of continuous purposive activity directed to specified goals; and a formal organization (*Betriebsverband*) is an enterprise based on associative relations resting on rationally motivated adjustments of interests rather than on communal ties, based on kinship ties or the subjective feelings of the parties. The two latter definitions have been widely employed by subsequent scholars.

Chester Barnard, perhaps the most influential early American organizational theorist, was a telephone executive who drew on his own, practical managerial experiences to develop his conception of the distinctive properties of organizations. Barnard (1938) emphasized the physical, biological, and psychological limitations of individual actors, limitations that greatly constrain their ability to make intelligent choices. Cooperative systems arise to enable individuals to collectively pursue a given purpose, the purpose acting to both support and constrain individual choice and decision making. Such systems are 'effective' to the extent that they achieve their purpose; they are 'efficient' if they satisfy the motives of individual participants. To be viable, systems must induce participants to make contributions; thus, only efficient systems can survive. Organizations represent a distinctive type of cooperative system: formal organization is 'that kind of cooperation among men that is conscious, deliberate, purposeful' (p. 4), although all

organizations incorporate both formal and informal systems.

Weber's intellectual descendants include a host of sociological scholars, from Parsons to Etzioni, Blau, and Albrow. All of these theorists follow Weber in emphasizing goal specificity and formalization as distinguishing features of organizations and in recognizing the importance of power and hierarchy both in defining goals and controlling access and specifying the terms of participation. Barnard's followers encompass a diverse collection of researchers, including political scientists Simon and March, economists Arrow and Williamson, and sociologists, Selznick and Perrow. Devotees from political science and economics embrace Barnard's view of organizations as arising from the limitations of individuals—physical, but more important, cognitive constraints—organizational structures and processes acting to both inform and channel their decision making. These same theorists also stress the rationality of individuals, insisting that their participation and willingness to make contributions is contingent on the continuing receipt of appropriate inducements. Sociologists, by contrast, focus on Barnard's recognition of the motivating power of purpose and the ways in which organizational structures and procedures themselves become infused with value, so that, as both instruments and ends in themselves, survival of the organization becomes an overriding goal for participants (Selznick 1949). Organizations are observed to dispense with specific objectives when their pursuit undermines organizational viability.

These views, varied as they are, nevertheless all embrace the substantialist conception of the organization as a durable, coherent, independent entity, capable of self-action. Features distinguishing organizations from other types of social groups are identified and viewed as enduring. Perhaps the strongest statement of this position is that of March and Simon (1958, p. 4) who, early in the modern era of organization studies, declared that:

> Organizations are assemblages of interacting human beings and they are the largest assemblages in our society that have anything resembling a central coordinative system. … The high specificity of structure and coordination within organizations—as contrasted with the diffuse and variable relations among organizations and among unorganized individuals—marks off the individual organization as a sociological unit comparable in significance to the individual organism in biology.

1.2 Substantialist Interaction Definitions

Soon after the emergence of organizations as a distinctive field of academic study, more interactive versions of substantialist conceptions developed that viewed organizations as 'fixed entities having variable attributes' that interact to create diverse outcomes (Emirbayer 1997, p. 286). Among the many varieties of such approaches are conceptions of organizations as political systems. Thus, Dahrendorf, elaborating the theoretical model of Marx, views organizations as the products of conflict, stability being the result of one group achieving dominance and imposing its interests on others. A modified version, developed by Cyert and March (1963), views organizations as coalitions of interests, with diverse parties vying for dominance. Even more prominent perspectives emphasized organizations as adapting to the distinctive features of their technical and social environments. For example, contingency theorists (Lawrence and Lorsch 1967) viewed managers as adapting structural features of organizations to the technical demands of their task environments; and Pfeffer and Salancik (1978) developed a resource dependence perspective calling attention to the types of bridging strategies employed by organizations in order to avoid excessive dependence on their exchange partners. Institutional economists such as Williamson (1975) saw organizations as governance structures of varying complexity arising to economize on the transaction costs associated with the exchange of goods and services.

In all of these accounts, organizations are still seen as discrete, pregiven units but possessing attributes that vary in response to changing circumstances or external forces. Most open-system theorists, who stress the importance of environment or context on organizational structure and action, retain the substantialist assumption of the organization as entity, as thing, but recognize the vital role of context—in particular, other organizational entities—in shaping through interaction organizational characteristics.

1.3 Relational Definitions

Although still a minority perspective, a growing number of theorists have proposed adoption of a relational approach in which organizations are viewed as 'inseparable from the transactional contexts in which they are embedded,' as deriving their meaning and identity from the roles they play in these encounters (Emirbayer 1997, p. 287). Marx was a prominent general advocate of this approach to social structure, but Weick was the major early proponent of this conception of organizations. He proposes the value of shifting attention from an entity 'organization' to a set of processes, 'organizing':

> We observe either an ongoing process that appears 'frozen' and steady because it is glimpsed only briefly, or we observe that the process is continuously changing if we watch for a longer span of time. The point is that the crucial events to be explained are processes, their structuring, modification, and dissolving (Weick 1969, p. 16).

This stance has been embraced and elaborated by theorists such as Giddens (1984), who proposes that 'structure' be reconceptualized as 'structuration': the continual production and reproduction of resources

and rules or schemas, as well as by interpretive and postmodernist theorists who emphasize the social construction of reality and the reflexive and contextualized nature of knowledge. Among organizational scholars, Silverman (1970) employed an interpretive framework to argue that the dominant 'systems' based view of organizations needed to be replaced by an 'action' based account. He asserted that action analysis is particularly well suited to addressing neglected issues such as the origins of organizations and organizational change. In a more programmatic statement, Burrell and Morgan (1979) distinguished between theoretical approaches stressing stability and order versus those emphasizing radical change. Their work served to stimulate efforts to advance interpretive and radical structuralist paradigms. Among theorists developing a more interpretive or cultural approach to organizations, a number emphasize the value of appropriating and adapting the tools developed by literary and humanistic scholars and suggest that society and its component units can be viewed and analyzed as 'texts': as a structure or grammar of rules and communication acts. Some propose the use of 'dramaturgic' approaches, focusing on the expressive dimension of social relations and the role of rituals and ceremonies in expressing commitments and dealing with social uncertainties. Others assert the value of 'narrative' analysis that views organizations as ongoing conversations among interacting individuals involved in formulating and reformulating intentions, interpretations, and accounts to themselves and each other (Czarniawska 1997).

Other relational scholars have emphasized not interpretive construction but network connections. Network theorists emphasize the value of viewing organizations as shifting networks of social relations embedded in fields of wider relational systems. Location or position in the network is viewed as of greater significance in shaping organizational behavior than its internal attributes; and both the organization's powers and constraints on its action are determined by the ever-changing configuration of ties (and the lack thereof) among an organization and counteractors.

Economists have begun to take notice of the organization or firm as an important unit of analysis. After long treating it as a 'black box' whose internal structures and processes were of little analytic interest, some economists recognize that one of the most important features of an organization is its status as an independent legal entity. It is capable of entering into binding contracts in its own name, separate from those involving its individual participants. A substantialist view would simply define an organization as a contractual node; but a relational version has dloped that views the organization as a 'nexus of contracts': the firm as a 'focus for a complex process in which the conflicting objectives of individuals (some of whom may 'represent' other organizations) are brought into equilibrium within a framework of contractual relations' (Jensen and Meckling 1976, p. 311). How stable is the equilibrium reached is a function of many factors, including changes in legal systems and contracting norms as well as the composition of and changing interests of participants.

Yet another relational model has been developed by finance economists that is applicable to many types of firms, but perhaps not to the full range of organizations. These economists view the firm primarily as a collection of assets earning varying rates of returns: the firm as portfolio. Such a view favors a diversification strategy—as managers are motivated to increase assets and reduce risks by entering into many independent businesses.

Relational approaches celebrate process over structure, becoming over being. What is being processed varies greatly. In some versions it is symbols and words, in others, relationships or contracts, in still others, assets. But in relational approaches, if structures exist it is because they are continually being created and recreated, and if the world has meaning, it is because actors are constructing and reconstructing intentions and accounts and, thereby, their own and others' identities.

1.4 Definitional Variation by Perspective

In addition to the differences occasioned by underlying ontological assumptions, conceptions of organizations also vary markedly by theoretical perspective: the underlying images or metaphors that direct the attention of theorists and researchers to some rather than other aspects of organizations. Three perspectives, each embracing a number of schools, have been pervasive during the twentieth century in organizational studies (Scott 1998).

The rational system model of organizations was the dominant perspective from early in the twentieth century to the 1930s. Organizations are viewed as rationally designed instruments for the attainment of specific goals. A mechanical model provides the guiding image, and emphasis is placed on those features that differentiate organizations from other social groups, such as communities or families. Exemplars include the scientific management approach of F. W. Taylor, Weber's theory of bureaucracy (1922/1968), and Simon's model of (boundedly) rational decision making hierarchies (March and Simon 1958). The twin attributes of goal specificity and formalization—the deliberate design of roles and social structures—are privileged.

The natural system model of organizations emerged during the 1930s and became dominant in the scholarly community in the 1940s and 1950s. Organizations were viewed as social groups whose participants are joined by common purpose and informal ties, but whose primary goal is the survival of the system. Rational, formal attributes were de-emphasized as

attention was focused on interpersonal ties and leadership. An organic image supplanted that of the well-oiled machine; not distinguishing elements, but those common to all social groups were emphasized. Specific schools included the human relations group, Barnard's (1938) view of organizations as cooperative systems, and Selznick's (1949) work examining how organizations are transformed into institutional systems as participants become committed to distinctive values.

During the 1960s, general systems theory swept over the field of organization studies giving rise to the open system model of organization. This model emphasized the extent to which organizations were dependent on and penetrated by facets and forces in the wider environment. By contrast, previous models were exposed as giving insufficient attention to actors and forces external to system boundaries. But as open system models surged into the arena, they were quickly coupled with, first, rational, then natural system concerns (Scott 1998). Thus, rational-open theorists like Thompson (1967) sought to explain how organizations could attempt to behave rationally even though open to their environments; March and Simon (1958) consider how decision rules can be adjusted to accommodate greater amounts of uncertainty; and Williamson (1975) theorizes that organizations arise and develop more complex governance systems to deal with increasing amounts of uncertainly and complexity. Contingency theory, decision-making theory, and transaction cost theory represent attempts to combine rational and open systems conceptions.

These developments dominated well into the 1970s, but were then joined by a resurgence of natural system theorists attempting to incorporate and take into account open system insights. Principal approaches included socio-technical systems theory, emphasizing the need to design technologies to take into account human and social factors; resource-dependence theory, as developed by Pfeffer and Salancik (1978), stressing the ways in which economic (resource-based) dependencies generated power problems that, in turn, were addressed by political solutions. In general, organizations confronted by increasing levels of interdependence respond by creating more encompassing coordinative and governance systems. Population ecology models (Hannan and Freeman 1989) examine the ways in which similar types of organizations compete for scarce resources within the same environmental niche. And institutional theorists (Meyer and Rowan 1977, DiMaggio and Powell 1983) consider the effects of wider cultural rules and normative structures on the emergence and legitimation of organizations.

Thus, conceptions of organizations have been influenced by the theoretical stance of scholars, as debates revolved around two axes: (a) the extent to which organizations are means—disposable, deliberately designed instruments for goal attainment—or value-impregnated, ends-in-themselves; and (b) whether organizations are self-sufficient, relatively insulated forms or substantially constituted, influenced, and penetrated by the wider environment.

2. *Emergence of Organizations*

Organizations are viewed as a construction of the modernist world, but there is much debate as to whether they represent accompaniments to and products of industrialization or, more generally, vehicles of rationalization. From Adam Smith to Clark Kerr and Alfred Chandler, economists have viewed organizations as handmaidens to industrialization. The subdivision of tasks enables specialization and mechanization, but is usually accompanied by agglomeration—the gathering of workers together in a common location—and requires coordination to integrate discrete tasks, typically through a command and control structure. A more recent, but compatible conception advanced by institutional economists (Williamson 1975), views organizations as operating to reduce transactions costs occasioned by the cognitive constraints (bounded rationality) of individual actors and the need to regulate opportunism among exchange partners. In either case, organizations are viewed as emerging to enhance efficiency: to reduce costs, whether of production or transactions.

Many of the most prevalent features of organizations—rules, formalized roles, dedicated machines and established work routines—are intended to contribute to the reliability of performance. However, as often noted, these same mechanisms often prevent the organization from changing quickly enough to keep pace with environmental demands (Hannan and Freeman 1989).

Sociological institutionalists argue that organizations should be viewed as a cultural construction: a product of increasing rationalization as generalized means–ends rules are promulgated and become valued independent of their contributions to heightened effectiveness and efficiency. The proliferation of organizations, and their infusion into virtually every arena of contemporary society, is viewed as resulting not only from their contribution to the coordination of differentiated, specialized tasks but also to their capacity to symbolize rational (rule-based, means–ends) procedures (Meyer and Rowan 1977). The formal structure of organization, hierarchical controls and the compilation of records all contribute to insuring the accountability of organization performance and thereby enhance their legitimacy. However, these same structures are also associated with red tape and ritualistic practices.

Viewed historically, organizations represent new types of social forms in that they are collective or corporate actors, accorded rights, capacities, resources that are independent of those of any of its individual participants. Coleman (1974) employs changes in

the law as important indicators of the growing independence and powers of organizations as they developed during the early modern epoch. Institutionalists such as Meyer, Dobbin, and Shenhav view the same process as fueled primarily by the growth of a scientific ethos emerging during the enlightenment that created abstract and general categories to classify and enumerate first the biological and physical universe and subsequently the social world. Local knowledge and culture, over time, succumbed to a new inductive epistemology viewed as universal. In the realm of purposive social organization, engineers and managers envisioned and then gradually helped to bring about the development of a single type of social entity (the 'organization') and a generic set of skills ('managerial') for their maintenance. It is a vision that has been embraced and, hence, reinforced by, the emergence of an academic discipline and profession devoted to the pursuit of organizational studies.

3. *Levels of Analysis*

The concept of organization is employed at various levels of analysis. More micro approaches treat organizations as distinctive types of social contexts that, in turn, affect the behavior and attitudes of their participants, whether considered as individuals or as groups or teams. Thus, the first level concentrates on organization behavior. Viewed as social contexts, organizations are generally thought to be 'strong situations': much effort and various techniques and devices are employed in an attempt to influence participant behavior. Positions are typically predefined, roles are prescribed, rules formulated, authority relations specified, information selectively routed, specialized tools and equipment provided, training offered, incentives created and sanctions employed, and beliefs and values inculcated. As Pfeffer (1997, p. 100) observes: 'control is at once the essential problem of management and organization and the implicit focus of much of organization studies.' Theorists differ in the relative priority they assign to different modes of control, whether economic, social, or cultural; but organizations are apt to combine many varieties of incentives and constraints in such a manner as to maximize the likelihood of compliance.

Moving up one level from a focus on organizational behavior, many analysts focus on the structural and other features of the organization *qua* organization. Here the organizational arrangements—extent of formalization, bureaucratization, centralization, mechanization, professionalization—rather than serving primarily as hallmarks of the setting become attributes of the corporate actor to be themselves described and explained. Extensive comparative research has been conducted attempting to account for the widely varying features of organizational structure. Among the causal factors examined are the characteristics of the task and institutional (including cultural and political) environments, the demography of participants, and the life cycle of the organization. When a particular organization is singled out and its context described as it relates to this focal unit, an organization set level is being employed. This level developed in association with the early stages of the open systems perspective and serves to facilitate examination of the particular information and resource flows and types of relations affecting the organization of interest. It is instructive to replace general descriptions of environmental variables such as complexity and uncertainty with specific information on who is providing what information and resources under what terms to the focal organization. This level of analysis has been particularly supportive of analysts employing resource dependence, transaction cost, and strategic management approaches (Pfeffer and Salancik 1978, Thompson 1967, Williamson 1975).

The organizational population level is defined so as to identity and encompass all instances or a geographically bounded aggregation of all organizations of the same type. This level roughly parallels that of an industry, except that attention is limited to organizations of the same general type. The concern here is not the individual organization but the organizational form: variously defined as similar structures, a common blueprint or archetype, or distinctive constitutive information. Researchers working at this level examine the conditions under which new ways of organizing and new types of routines arise and coalesce so as to constitute a recognizable, accepted form for the pursuit of specified goals. Analysts of populations examine the conditions under which populations of organizations emerge, grow, decline, and fail (Hannan and Freeman 1989).

At the macro end of the scale, theorists have identified the organizational field as an important level linking individual organizations and organizational systems to wider, societal structures and processes. Organizational fields are defined to be 'those organizations that, in the aggregate, constitute a recognized area of institutional life: key suppliers, resource and product consumers, regulatory agencies, and other organizations that produce similar services and products' (DiMaggio and Powell 1983, p. 143). Fields are made up of multiple populations of organizations and their respective organization sets. Organizational fields vary greatly in terms of the extent of their organization, the degree of their structuration; and the emergence, development, and decline of organizational fields has become an important new focus of study.

4. *Changing Theories; Changing Organizations*

It would be odd indeed if there were no relation between our changing definitions and conceptions of organizations and changes in the phenomena itself.

Over the past few centuries, organizations have evolved from relatively embedded and dependent components of traditional society to more independent systems that attempted to insulate themselves from the surrounding society, to systems that are much more open, more flexible, and more penetrated by their context. Since the beginning of the industrial period, organizations have developed from highly centralized to more decentralized structures; and they have tended to move from more tightly to more loosely coupled systems. Rather than having their component units be fixed jobs and departments, they are increasingly made up of more diffusely defined skill categories whose participants are grouped into changing configurations of teams and project groups. Also, jobs formerly linked densely in internal labor markets providing clear career lines to employees are now more often held by temporary workers or independent contractors whose mobility routes cross organizational boundaries. Rather than attempting to incorporate all functions within the same hierarchical ownership structure, organizations increasingly find ways to contract work out to relationally-linked independent firms or to develop alliances with a shifting set of independent partners. New developments in information processing allow the collaboration of geographical remote partners and various combinations of off-site, part-time, and loosely-linked participants. And ownership is no longer viewed as necessary—and in some situations is seen to undermine—control of essential resources and services.

Some theorists, such as Perrow (1991), argue that organizations have increasingly 'absorbed society,' taking over one after another societal functions. 'Activities that once were performed by relatively autonomous and usually small informal groups (e.g., family, neighborhood) and small autonomous organizations (small businesses, local government, local church) are now performed by large bureaucracies' (p. 726). The corporation is increasingly the source of friends, marriage partners, counseling, medical care, recreation, food services, and retirement options for individual participants.

While there is much truth in this claim, Meyer (1983) stresses the obverse proposition: modern society increasingly absorbs formal organizations. Contemporary societies and international organizations generate many types of rationalized rules, occupations and experts. In order to keep abreast of the latest developments and to exhibit the requisite symbols and forms, organizations conform to the prescriptions—and incorporate representative members of—the relevant professions. And, because these rationalities are various and organized around somewhat divergent purposes, 'the institutionalization of rational organization tends to lower the formal rationality of specific organizations.' Organizations as discrete units look less like rational structures and more like 'holding companies incorporating various institutionally defined packages' (p. 262). In short, organizations are only one of many carriers of rationality in the modern world, and as the prevalence of these cultural patterns increases, organizations are less likely to determine and are more likely to absorb and reflect these diverse logics.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Networks: Social; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture; Organizational Culture, Anthropology of; Organizational Decision Making; Organizations and the Law; Organizations, Sociology of; Population Ecology; Rational Choice and Organization Theory; Weber, Max (1864–1920)

Bibliography

Barnard C I 1938 *Functions of the Executive*. Harvard University Press, Cambridge, MA

Burrell G, Morgan G 1979 *Sociological Paradigms and Organisational Analysis*. Heinemann, London

Coleman J S 1974 *Power and the Structure of Society*. Norton, New York

Cyert R M, March J G 1963 *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ

Czarniawska B 1997 *Narrating the Organization: Dramas of Institutional Identity*. University of Chicago Press, Chicago

DiMaggio P J, Powell W W 1983 The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review* **48**: 147–60

Emirbayer M 1997 Manifesto for a relational sociology. *American Journal of Sociology* **103**: 281–317

Giddens A 1984 *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, Berkeley, CA

Hannan M T, Freeman J 1989 *Organizational Ecology*. Harvard University Press, Cambridge, MA

Jensen M C, Meckling W H 1976 Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* **3**: 305–60

Lawrence P R, Lorsch J W 1967 *Organization and Environment: Managing Differentiation and Integration*. Graduate School of Business Administration, Harvard University, Boston

March J G, Simon H A 1958 *Organizations*. Wiley, New York

Meyer J W 1983 Institutionalization and the rationality of formal organizational structure. In: Meyer J W, Scott W R (eds.) *Organizational Environments: Ritual and Rationality*. Sage, Beverly Hills, CA

Meyer J W, Rowan B 1977 Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology* **83**: 340–63

Perrow C 1991 A society of organizations. *Theory and Society* **20**: 725–62

Pfeffer J 1997 *New Directions for Organization Theory: Problems and Prospects*. Oxford University Press, New York

Pfeffer J, Salancik G 1978 *The External Control of Organizations*. Harper and Row, New York

Scott W R 1998 *Organizations: Rational, Natural and Open Systems*, 4th edn. Prentice-Hall, Upper Saddle River, NJ
Selznick P 1949 *TVA and the Grass Roots*. University of California Press, Berkeley, CA
Silverman D 1970 *The Theory of Organisations: A Sociological Framework*. Heinemann Educational, London
Thompson J D *Organizations in Action*. McGraw-Hill, New York
Weber M 1922/1968 *Wirtschaft und Gesellschaft: Grundriss der verstehenden Soziologie*. J. C. B. Mohr, Tubingen, Germany. Translated by Roth G, Wittich C (eds.) *Economy and Society: An Outline of Interpretive Sociology*. Bedminister Press, New York
Weick K E 1969 *The Social Psychology of Organizing*. Addison-Wesley, Reading, MA
Williamson O E 1975 *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York

W. R. Scott

Organizational Behavior, Psychology of

Organizational behavior (OB) denotes a relatively new area of research and practice of psychology. It refers to the study of both the psychological and social antecedents and consequences of behavior of individuals and groups *in* organizations, as well as the psychological and social antecedents and consequences of the behavior *of* organizations. Stressing a psychological point of view, this article defines the field of OB and describes the main characteristics of its dominant domains, their methodological implications, and the likely future directions of OB.

1. *The Nature of the Field*

As in any new scientific field, various attempts have been made by different authors to define the field of OB. In the very first review entitled 'Organizational behavior' in the *Annual Review of Psychology*, Mitchell (1979, p. 244) considers OB as a 'distinct discipline with a focus on individual and group behavior in the organizational context,' thus making it virtually synonymous with traditional work and organizational psychology. Staw (1984, p. 628) underlines the interdisciplinary character and the two main orientations of the field, the stress on micro- and macroissues: 'organizational behavior is an interdisciplinary field that examines the behavior of individuals within organizational settings as well as the structure and behavior of organizations themselves.' According to this view, microaspects are dealt with by psychology whereas macroaspects are treated by sociology, political science, and economics, making OB appear as two distinct subdisciplines, an impression which is further fostered by different professional identifications of the students of OB. Others simply define OB operationally in terms of the organizational research topics and issues published over time in scientific psychological periodicals (Rousseau 1997) covering, for instance, leadership, motivation, organizational technology, job analysis, legal issues, and organizational performance.

A more integrative stance is taken by Schneider (1985, p. 574) in focusing on the types of studies conducted in OB: 'OB is the confluence of individual, group, and organizational studies flowing from industrial–organizational (I/O) psychology and organizational and management theory (OMT) with headwaters in psychology (social, psychometrics), sociology (organizational, work, and occupational), and management (scientific, human relations).' Given this mixed heritage of OB, its scope is extremely broad, covering both micro- and macrophenomena of organizations. McGrath (1976), in an early, even more comprehensive manner, conceptualizes OB as the intersecting area of three systems: the physical–technical/structural, the social–interpersonal, and the person system of organization members.

From a psychological point of view, OB may thus be defined as an emergent field of scientific inquiry which is constituted by the converging contributions of various social and behavioral sciences focusing on multilevel and multidimensional antecedents and consequences of psychological phenomena in organizations. The multilevel perspective of OB implies that literally *all relevant levels and dimensions* that either condition or are conditioned by psychological phenomena in organized settings fall into the scope of OB studies: the levels of the individual–personal, the group, the managerial, the technological, the organizational-structural, the interorganizational, and environmental. The notion of multiple dimensions refers here to organizational practices, processes, and outcomes of any of the relevant system levels as proper problems of OB studies.

Although the study of organizational issues has been part and parcel throughout the history of work and organizational psychology (Münsterberg 1912), it remains a moot question why it was only during the last three decades of the twentieth century that OB entered the scene as a distinct subfield of scientific inquiry. One may hypothesize that factors seem to have interacted here that are partly due to change in society and partly due to intrinsic developments of science, both of which made their effects particularly felt in the decades after World War II. Epochal shifts in the work force such as the growth of the service sector of economy, the demographic changes in terms of age and gender composition implied drastic restructuring of vocational and professional profiles. The cost-effective wholesale introduction of new information and communication technologies in work settings created dramatic challenges to a novel division of labor. Social value changes such as the changing

subjective meaning of working (MOW 1987) and the increase of international market competition may have further been contributing factors of societal change to highlight the need for the reform of work organizations and social institutions that had to go far beyond a restructuring of individual work places. Required were more holistic approaches to the study and practice of organizational change and performance that eschewed monodisciplinary psychological treatments. This is where progress in science offered itself as a promising solution: the emergence of systems theory as a meta-theoretical intellectual vehicle to facilitate multilevel studies and to link different scientific disciplines to each other in the analysis of complex scientific and practical problems appears to have been a further necessary condition for OB to acquire scientific credibility as an interdisciplinary field.

Despite its comparatively short history, OB is far from displaying a unified common theoretical framework and a corresponding set of issues under study. The most important theoretical underpinnings and the substantive problems studied in their perspective will be addressed next. These theoretical perspectives represent umbrella orientations usually encompassing diverse theoretical positions of varying scope which, depending on one's meta-theoretical penchant, may also be classified differently.

2. Theoretical Paradigms and Topical Foci

2.1 Psychology of Individual Differences

The basic tenet of differential psychology within OB is the assumption that personal characteristics constitute the core of variables to be studied by psychology either as causal (at least moderating) factors of consequent psychological/social phenomena or as the most important outcomes of individual experience in organizational settings. The microperspective dominates here. Within this theory class might also be counted the important approaches of expectancy and of goal setting theory inasmuch as individual characteristics (e.g., preferences, competencies) need to be considered by them. The former surmises that subjectively expected outcome values (subjective expected utilities (SEU)) guide individual decision-making behavior (Naylor et al. 1980), the latter postulates that goals, self-set individually or interactively by subordinates and superordinates, constitute potent predictors of subsequent behavior in organizations.

Major topical concerns within this paradigm concentrate on issues such as the role of personality and leadership style as conditioning work group and organizational performance (including absenteeism from work, turnover), work attitudes (causes and consequences of job satisfaction, meaning of working, and work commitment/motivation).

2.2 Bureaucratic Theory

An important contribution to the advancement of OB was made by the interdisciplinary team of Aston University (Birmingham, UK) under the leadership of a psychologist, Derek Pugh (Pugh and Hickson 1976). Taking Max Weber's theory of bureaucracy as the starting point, the Aston group deviated from the received single-level analytic procedure and proposed to overcome this by integrating the analysis of three levels: situational factors (size of organization, technology, environment), organizational structure (decision centralization, specialization, formalization, configuration), and member experience and behavior. A basic assumption was that cross-level effects had to be taken into account in the sense that an organization's technology would impact the division of labor, i.e., the role structure of the organization which, in turn, would influence autonomy and individual outcomes such as the evaluation of one's work setting (motivation, commitment, satisfaction). Thus the Aston approach covers both micro- and macrodimensions of OB.

As a major achievement of the Aston group, apart from introducing from its outset a theoretical multilevel and interdisciplinary perspective, must be considered its success in setting current psychometric standards for the development of instruments for an adequate analysis of organization structure and its interaction with variables on other system levels. A burgeoning body of studies followed from their lead.

2.3 Cognitive Psychology

The generally noted 'cognitive shift' in psychology during the latter part of the twentieth century soon was also to reach OB and to influence paradigmatic thinking in the field. Cognitive processes are in this perspective considered to mediate or govern behavior. Various positions fall under the umbrella category of cognitive psychology, some of which achieved prominence in OB; the three most important ones are social cognition, expectancy theory, and attribution theory.

Social cognition refers to processes among relevant actors which produce socially shared representations, views, and perceptions of organizational realities. Thus some 'social constructivist' authors claim that organizational phenomena are mainly socially constructed and articulated through negotiated goals, structures, artifacts, through myths, rituals, stories, and symbols which result in organization-specific, intersubjectively shared meanings and values (Turner 1990). Social constructivism favors an ethno-methodological and anthropological, holistic approach to organizations which lends itself to the analysis of molar organizational features such as organizational culture and climate. Under the heading of social cognition must also be mentioned the significant

contribution of equity theory, which focuses on an actor's input:output ratio in comparison with the respective ratio of similar others with the resulting energy expenditure and work satisfaction. *Expectancy theory*, as described above, addresses the cognitive processes of organization members' choices according to perceived SEU. It has shown considerable fruitfulness in the study of leadership, self-efficacy, and work motivation. *Attribution theory*, owing to its focus on perceptual processes concerning perceived causes of success and failure, has to a large extent influenced theorizing and empirical research on achievement motivation and respective performance in organizations.

2.4 Action Theory

Many different theoretical social and behavioral science approaches go under the guise of 'action theory.' Some of them play a major role in OB. A large body of German theoretical and empirical literature under the name of action control/regulation theory needs to be subsumed here (Frese and Zapf 1994). It combines elements of cybernetics and cognitive aspects of goal-oriented behavior (actions) that is assumed to be regulated by feedback loops concerning the deviations from goal achievement which necessarily includes the consideration of task and technical components of task and organization structures. The approach has proven its usefulness in a wide variety of domains, e.g., human error analysis, personality and competence development, individual task and organization design, linking socially shared meanings, and values to collective action control.

A notable French approach of sociological provenance (Crozier and Friedberg 1977) puts social actors (individuals, groups, organizations) and their various strategies of playing their power games into the center of its theoretical and empirical analysis. In this context, games are not understood in a metaphoric sense, but as concrete mechanisms to structure organizational power relations among interdependent actors negotiating degrees of freedom within given organizational and environmental constraints and opportunities. Games are here considered as the very instruments that regulate and sustain stability and integration of organized behavior based on accepted rules of the games. Thus the authors succeed in encompassing both micro- and macroaspects of OB (see *501061*).

The notion of autopoiesis, derived from biology (Maturana et al. 1980), has attracted growing interest also in OB. The concept refers to the continuous and structurally coupled adjustment of living systems to their extant environment and their continuous systemic self-reproduction. It has been applied to, among others, self-organizing capacities of institutions of public administration and their sustained regeneration of the bureaucracy and the autonomy and self-governance of organizations acting in the public area. The link of micro- and macroaspects of organizations in their given environment becomes clearly dominant in this perspective.

2.5 Systems Thinking

A lasting paradigmatic contribution to OB was made by the so-called socio-technical systems approach (STSA), developed soon after World War II by a group of social scientists of the London-based Tavistock Institute of Human Relations (Emery and Trist 1960). The approach grew out of general systems theory and is predicated on important assumptions concerning the interaction of an organization's *social* subsystem (its members, role structure) and its *technical* subsystem (task structure, technology). In linking task and technological elements to the social patterns of organization, STSA paved the ground for systematic analysis and design of work places and work organizations. Work and organizational design was postulated to proceed always through joint optimization of both subsystems in order to guarantee efficiency. The increased participation of employees in organizational change programs became a leading principle. Thus it was possible to take into account both the characteristics of organization members (their values, aspirations, expectations, social and hierarchical relations, role ascriptions, etc.) and the organization's technical dimensions.

The theoretical innovations of the STSA proved to be of great practical usefulness in providing guidelines for the design of work and work organizations which met, for instance, the demands of differential levels of human competencies and also of technical and task demands. The large majority of work humanization programs of the 1970s and 1980s profited from them in many industrialized countries. It facilitated improvements in the adequate introduction of new information and communication technologies, the creation of flexible production units, the formation of semi-autonomous work groups, and the creation of large-scale industrial democracy schemes.

3. Unresolved Issues and Future Directions of OB

The multilevel perspective of OB poses an intriguing *theoretical problem*: how can and should the relationships of several system levels be conceptualized? One case in point is problems such as the structural interrelations of organization-specific features (organization strategy and structure, membership resources, organizational practices, output) and characteristics of their institutional and societal environment (dominant culture, education and legal system, economic and technological status). While there is growing evidence that these features are interrelated and influence each other, theoretical integration remains

lacking. A possibly promising theoretical solution might be found in Giddens' theory of structuration. Giddens (1984) postulates a dual nature of institutions recursively reconstituting and maintaining each other. On the one hand, institutions, by virtue of their existence, facilitate action of individuals and of organizations, and on the other hand, by virtue of the acting actors, the institutions are recursively constituted and sustained. The theory still being on a rather abstract level requires further development in order to make it fruitful in OB.

A *methodological problem* of considerable difficulty is the need to account for effects of time. Change and continuity of organizations may be likened to quasi-stationary equilibria (Lewin 1952). Like rivers they are in continual flux while preserving their identity and basic characteristics for some time. One methodological response of OB to study change and continuity was to repeat several cross-sectional studies at various points in time, thus approaching some sort of longitudinal evaluation. Another response was to trace organizational developments historically through document analysis and interviewing. Yet another approach seems to be to employ more 'dense' descriptions of organizational processes over time by way of ethnographic participant observation. This is a choice that is increasingly advocated and implemented by many researchers in spite of its cost, simply because of its superior information yield.

Finally, it may be that at the turn of the millennium, *societal changes* as alluded to above will gather further momentum to induce also a new organizational era impacting on OB. According to Rousseau (1997), organizations seem to be faced with the challenge of increasing upheaval and transition of depth and magnitude that were hitherto unknown. She refers to governmental measures of deregulation, global competition, the explosion of information technology in manufacturing and service sectors, the uncoupling of traditional education from work demands, the emergence of small-firm employment, new and more differentiated employment relations, outsourcing among firms and inter-organizational networking, shifts from managerial prerogatives to self-management, transition from rule-regulated work to self-reliance, and continuous adaptation of workers. In short, traditional internal organizational structures are becoming more fluid and the boundaries of organizations blurred. The perceived meaning of organization as a clear-cut entity seems in need of change to adaptive *organizing*. A reorientation of OB with reference to thematic foci in addition to theoretical underpinnings would have to be the consequence of such developments. Theoretically, a renewed emphasis on the conceptualization of the dynamics of organizational and interorganizational processes is called for. New topical issues are likely and desirable to emerge as concern for OB, such as the changing nature of employment and equitable forms of remuneration, emergence of multiple loyalties, individual learning for self-management and career development, organizational learning, and intervention methods to introduce and to sustain change. OB seems to be facing new and wide areas to be examined, a task for which its intrinsic multilevel and interdisciplinary outlook will be a definite asset.

Comprehensive treatments of the field can be found in the following publications: *Handbook of Industrial and Organizational Psychology*, Vols. 1–4 (see Frese and Zapf 1994); *Annual Review of Psychology*; *International Reviews of I/O Psychology*; *Administrative Science Quarterly*; *Organizational Studies*; *Human Relations*; and *Journal of Occupational and Organizational Psychology*.

See also: Dynamical Systems Approaches to Neural Systems and Behavior; Job Analysis and Work Roles, Psychology of; Industrial and Organizational Psychology, Science and Practice

Bibliography

Crozier M, Friedberg E 1977 *l'Acteur et le Système. Les Contraintes de l'Action Collective*. Editions du Seuil, Paris

Emery F E, Trist E C 1960 Socio-technical systems. In: Churchman C W, Verhulst M (eds.) *Management Science, Models and Techniques*. Pergamon, New York, pp. 83–97

Frese M, Zapf D 1994 Action as the core of work psychology. A German approach. In: Dunnette M D, Hough L M, Triandis H (eds.) *Handbook of Industrial and Organizational Psychology*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA, Vol. 4, pp. 271–340

Giddens A 1984 *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, Berkeley, CA

Lewin K 1952 Group decision and social change. In: Swanson G E, Newcomb T M, Hartley E L (eds.) *Readings in Social Psychology*. Holt, New York, pp. 459–73

Maturana H R, Varela F J, Cohen R S 1980 *Autopoiesis and Cognition: The Realization of the Living*. Reidel, Boston

McGrath J 1976 Stress and behavior in organizations. In: Dunette M (ed.) *Handbook of Industrial and Organizational Psychology*. Rand McNally, Chicago, pp. 1351–96

Mitchell T R 1979 Organizational behavior. *Annual Review of Psychology* **30**: 243–81

MOW—Meaning of working international research team 1987 *The Meaning of Working*. Academic Press, London

Münsterberg H 1912 *Psychologie und Wirtschaftsleben*. Barth, Leipzig, Germany

Naylor J C, Pritchard R D, Ilgen D 1980 *A Theory of Behavior in Organizations*. Academic Press, New York

Pugh D S, Hickson D J 1976 *Organizational Structure and its Context. The Aston Programme I*. Saxon House, Farnborough, UK

Rousseau D M 1997 Organizational behavior in the new organizational era. *Annual Review of Psychology* **48**: 515–46

Schneider B 1985 Organizational behavior. *Annual Review of Psychology* **36**: 573–611

Staw B M 1984 Organizational behavior: A review and reformulation of the field's outcome variables. *Annual Review of Psychology* **35**: 627–66

Turner B R (ed.) 1990 *Organizational Symbolism*. de Gruyter, Berlin

B. Wilpert

Organizational Change

A simple definition of organizational change is new ways of organizing and working. It involves the movement over time from a present state of organization to some future state and may range from fine-tuning to strategic redirection. The deeper and more pervasive forms of organizational change and the development and use of new forms of work organization (for example, lean manufacturing, virtual work, and network organizations) have become a major focus of research. Across this growing and rather diverse range of studies, there remains an underlying debate on whether change can be characterized as a logical sequence of events which can be planned, or whether large-scale change unfolds in unpredictable ways.

1. The Changing Historical World of Organizational Change

Programs aimed at bringing about a fundamental change in the organization and control of work are not peculiar to the twenty-first century. For example, the development of the steam engine in 1785 and the mechanization of the textile industry in the late eighteenth century marked a significant societal transformation in changing from a mainly agricultural society into an industrial economy. By the turn of the century, Chicago meat-packing industries were experimenting in the use of continuous production lines which finally reached maturity under the new fixed-speed moving assembly lines of Henry Ford in the twentieth century.

Although there has been a long history of research into organizations and change, many commentators advocate that the current climate of change is quantitatively and qualitatively different from those experienced in the 1960s and early 1970s. The breakup of the old Soviet Union and the transformation of Eastern Europe, the launching of a single European currency, the growth in trade in the Asia Pacific region, wars, technological advances, world recessions, and changes in international trade agreements are all part of the current world in which many corporations find themselves. As an organization's ability to manage change is perceived to be increasingly central to its competitive position and ultimate survival, so the demand has increased for simple change recipes to solve complex organizational problems. However, academics who have sought to explain organizational change have been quick to show how there can be significant variation in the adoption of managerial ideas, how any substantive change is likely to engender employee distrust and resistance, and how different strategies and structures may be appropriate under different circumstances. These and other competing views have resulted in a large body of empirical research used in the formulation of 'new' approaches to understanding and managing change (Burnes 2000).

2. Why Do Organizations Change?

Many commentators have sought to isolate the factors which support and promote change and a range of 'triggers' to organizational change have been identified. Some of the main external factors are seen to comprise:

(a) Organizational growth and expansion (as an organization increases in size so may the complexity of the organization, requiring the development of appropriate coordinating mechanisms).

(b) Globalization of markets and the internationalization of business.

(c) Government laws and regulations (for example, international agreements on tariffs and trade).

(d) Advances in technology.

(e) Major political and social events (for example, the changing relationships and tensions between countries).

(f) Fluctuations in business cycles (for example, changes in the level of economic activity both within the national economy and within major trading blocks can significantly influence change strategies).

Four internal triggers to change are generally characterized in this field of study (Leavitt 1965). First, a change in an organization's technology which may involve the installation of a single piece of equipment or the complete redesign of an operating process. Second, a change in the administrative system such as restructuring lines of communication, establishing new working procedures, and introducing reward systems. Third, changing the human element of work, for example, through cultural change programs which seek to change employee behavior at work. Fourth, in changing the business focus (primary task) of the company, for example, from the development and marketing of computer software to the provision of a computer support service.

In practice, these various internal and external influences may all interlink and overlap in determining the speed, direction, and outcomes of change within an organization. Furthermore, with the growing expectation of the need to change, the managerial careers and political aspirations of business leaders can also promote visions and strategies for change. In seeking to explain this complex issue, a number of competing change perspectives have been developed.

3. Understanding Change: Models and Frameworks

There are numerous models and frameworks for understanding change (Beckhard and Harris 1987, Buchanan and Badham 1999, Kanter 1983, Mohrman et al. 1989). There are debates over whether change can best be explained with reference to the external environment (population ecology theories), whether it is largely a political and social process (institutional theory), or whether it needs to be understood in terms of the embeddedness of organizations within particular social, political, and economic structures (critical theory). Postmodernist writers have turned their attention to questioning the degree to which organizations are radically changing in shape and form to meet a new postmodern epoch and/or the degree to which we need to overturn assumptions and deconstruct modernist accounts. Within these broad spectra of studies, two views have proven particularly influential. First, those that seek to develop models which can be employed in *planning* the successful management of change. Second, those that aim to explain the dynamics of the change *process* which is seen to be messy, emergent, and iterative.

3.1 Planned Change Models

Kurt Lewin (1947) stands as an early pioneer in developing a model to service the successful management of organizational change. As a founding father of Organizational Development (OD), Lewin has had a lasting influence on planned change management models. From his work on group dynamics, he maintained that within any social system there are driving and restraining forces which serve to maintain the status quo, and that organizations generally exist in a state of equilibrium which is not itself conducive to change. As such, there has to be an unfreezing of this equilibrium through creating an imbalance between the driving and restraining forces. In the management of change, these planned models tend to focus on methods which reduce the restraining forces rather than on those that increase the driving forces. Once an imbalance has been created then the system can be altered and a new set of driving and restraining forces put into place. A planned change program is implemented and only when the desired state has been achieved, will the change agent set about 'refreezing' the organization. The new state of balance is then appraised and where appropriate, methods of positive reinforcement are used to ensure employees 'internalize' attitudes and behaviors consistent with new work regimes.

The distinct field of OD has further refined and developed this approach in applying behavioral science knowledge to the planning and design of organizational change strategies. In evolving from a human relations approach, attention is given to employee involvement in change programs and to overcoming potential resistance to change (French and Bell 1995). In building on this work, a number of 'recipes' for successful change have been developed which stress the importance of involving people and open communication. In many of the new change initiatives such as business process re-engineering and quality management, this tendency towards a recipe approach to change has been maintained. However, these rather linear approaches which forward universal prescriptions have been criticized for failing to account for the more fluid, dynamic, and processual nature of change (Dawson 1994).

3.2 Change Process Frameworks

Processual studies (such as the pioneering work of Pettigrew 1985) draw attention to the importance of context in examining unfolding processes of change, yet, unlike previous theorists, they are not drawn towards unidirectional rational models of change. They are concerned with understanding the dynamic relationship between: the political arenas in which decisions are made, histories recreated and strategies rationalized—*the politics of change*; the enabling and constraining characteristics of change programs and the scale and type of change—*the substance or content of change*; and the conditions under which change is taking place in relation to external elements such as the business market environment and internal elements, including the history and culture of an organization—*the inner and outer context of change* (Dawson 1994).

This shift in research emphasis towards studying the processes which shape change outcomes was highlighted in the *Academy of Management Chicago* 1999 *Conference* where *process* was identified as the central theme and a key concept to understanding the unfolding nature of change. A growing number of these process studies are multidisciplinary and draw on a range of perspectives and methods. Considerable theoretical importance is placed on the temporal dimension of change (the need to study change 'as-it-happens'), the substance or content of the specific change strategy, and the contextual dynamics within which change takes place. The political process is important and while actors influence change processes, the legacy of these decisions (which may become institutionalized) may in turn shape future action and decision-making activities. There is a complex dynamic between agency and structure within the unfolding sociopolitical context within which strategic choices are made (Child 1997). These processual accounts seek to accommodate the muddied and sometimes contradictory nature of change processes and reject both linear stage models and simple prescriptive 'recipe' guides. Although theoretically more

sophisticated, some of these processual approaches have been criticized for being unable to deliver any practical advice on the management of change.

4. Organizational Change in the Twenty-first Century

The perception of change as a central notion is to be expected at the start of a new century. Academic concerns are likely to continue along the route of understanding the unfolding and dynamic process of change within increasingly diverse and pluralistic settings. However, the greater interest in process studies may refuel debates about the degree to which change processes are determined by the need for organizations to adapt to new economic circumstances in order to compete and survive and the extent to which individuals and groups shape the political process of organizational change. Furthermore, as more academics get involved from different methodological backgrounds, the push to use a broader selection of data techniques is likely to gain momentum with a concomitant increase in the systematic codification of approaches (Hinnings 1997). Attention may be given to improving methods for evaluating change and drawing more substantive links between change and organizational learning. Whatever eventuates, *process* is likely to be a key concept to studies of organizational change over the next decade.

See also: Bureaucracy and Bureaucratization; Competitive Strategies: Organizational; Conflict: Organizational; Corporate Culture; Information and Knowledge: Organizational; Innovation: Organizational; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Learning: Organizational; Organization: Overview; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations, Metaphors and Paradigms in; Organizations, Sociology of; Organizations: Unintended Consequences; Rational Choice and Organization Theory; Strategic Intelligence; Strategy: Organizational; Technology and Organization

Bibliography

Beckhard R, Harris R T 1987 *Organizational Transitions: Managing Complex Change*, 2nd edn. Addison-Wesley, Reading, MA

Buchanan D, Badham R 1999 *Power, Politics, and Organizational Change. Winning the Turf Game*. Sage Publications, London

Burnes B 2000 *Managing Change: A Strategic Approach to Organisational Dynamics*, 3rd edn. Financial Times/Prentice Hall, Harlow, UK

Child J 1997 Strategic choice in the analysis of action, structure, organizations and environment: Retrospect and prospect. *Organization Studies*. **18(1)**: 43–76

Dawson P 1994 *Organizational Change: A Processual Approach*. P. Chapman, London

French W, Bell C 1995 *Organizational Development and Change*. West Publishing Company, Minneapolis, MN

Hinnings C 1997 Reflections on processual research. *Scandinavian Journal of Management*. **13(4)**: 493–503

Kanter R 1983 *The Change Masters*. Simon and Schuster, New York

Leavitt H J 1965 Applied organizational change in industry: Structural, technical and human approaches. In: March J G (ed.) *Handbook of Organizations*. Rand McNally, Chicago, pp. 1144–70

Lewin K 1947 Frontiers in group dynamics. *Human Relations*. **1**: 5–42

Mohrman A M Jr., Mohram S A, Ledford G E, Cummings T G, Lawler E E (eds.) 1989 *Large-scale Organizational Change*. 1st edn. Jossey-Bass, San Francisco

Pettigrew A 1985 *Awakening Giant: Continuity and Change in Imperial Chemical Industries*. Blackwell, Oxford, UK

P. Dawson

Organizational Climate

1. What is Organizational Climate?

Climate has been variously described as 'the way things are around here' (Schneider 1990) to 'a molar concept reflecting the content and strengths of the prevalent norms, attitudes, behaviors, and feelings of the members of the social system which can be operationally measured through the perceptions of system members, or observational means' (Payne 1990, p. 156). The *Oxford English Dictionary* defines meteorological climate as 'the prevailing atmospheric phenomena and conditions of temperature, humidity, wind, etc., of a country or region.' When applied to communities, climate is 'the mental, moral, etc., environment prevailing in a body of people in respect of opinion, some aspect of life, etc.' Thus, just as we describe climates in terms of sunshine, prevailing winds, levels of rainfall, etc., organizations are described in terms of dimensions such as leadership style (autocratic versus democratic), concern with the well-being of employees, emphasis on productivity, formalization or bureaucracy, innovation versus traditionalism, and commitment to quality in service or production.

2. Research into Organizational Climate

One of the earliest uses of the concept of climate was in research conducted by Lewin, et al. (1939), who examined the effect of the leadership style of boys upon their work groups. Leadership style was cat-

egorized as autocratic, democratic or laissez-faire and the work-group climates were described, respectively, as hostile, unproductive, and unenjoyable; busy, cooperative, productive and enjoyable; and anarchic, unproductive, not cohesive and unsatisfying.

Subsequent research has examined climate at different levels of analysis including group, department, and organizational levels. Most researchers have concentrated on climate at work group, department or organizational levels. These supraindividual climates have been operationally constructed by aggregating individual perceptions of psychological climate to the appropriate level and using the mean to represent the climate at that level of aggregation. These organizational climate measures are now widely used in public and private sector organizations to determine the prevailing climate, often being called employee attitude surveys or employee opinion surveys.

One of the reasons why organizational climate surveys have become so widely used is that climate has been shown to predict organizational productivity and profitability (Denison 1990, Patterson et al. 1997, West and Patterson 1999), employee motivation, job satisfaction, job performance, organizational commitment, and stress. For example, Patterson et al. (1997) found that an organizational climate oriented towards good human relations predicted 29 percent of the variation between companies in productivity and profitability in the UK manufacturing sector. Because of these strong associations between climate and performance, both researchers and practitioners continue to be intensely interested in the concept.

3. *Climate or Climates?*

There is uncertainty and debate about whether climate should be conceived of as a multidimensional concept or a unitary concept (Campbell et al. 1970, Payne and Mansfield 1973, James and Jones 1974). Campbell et al. (1970) argue for a small number of common dimensions that constitute climate, including individual autonomy at work, the degree of structure imposed upon people's work roles, reward orientation (positive or negative), and warmth and support. Schneider and Reichers (1983) suggest that work settings actually have many climates and they argue for the study of facet specific climates. To their way of thinking, the choice of climate dimensions should be closely linked to the criteria of interests—climates are 'for something,' e.g., customer service or safety, and there is no generic organizational climate.

However, most researchers believe that organizational climate is most usefully applied as a multidimensional construct and that some dimensions are generic and applicable to most organizations. Of course, there will be some which are unique or at least applicable to only a small range of organizations (e.g., climates for secrecy in the defense industry or for caring in health care organizations). One orientation to the mapping of climate that has proved powerful is the competing values model of organizations (Quinn and Rohrbaugh 1981) which distinguishes between human relations, rational goal, open systems, and internal process constellations of organizational values.

4. *The Competing Values Model*

The competing values approach incorporates a range of fundamental dimensions of values into a single model. It calls attention to how opposing values exist in organizations and how organizations have different mixtures of values that influence their goals and how they go about achieving those goals. These values will most fundamentally manifest themselves in whether the organizational orientation varies more towards flexibility or control, and whether the focus of the organization varies more towards the external or internal environment. A major theoretical strength of this model is its derivation from four orientations to the study of organizational effectiveness reflecting long traditions in work and organizational psychology. Thus the rational goal approach (external focus, but with tight control within the organization) reflects a rational economic model of organizational functioning, in which the emphasis is upon productivity and goal achievement. The open systems model (external focus and flexible relationships with the environment) emphasizes the interaction and adaptation of the organization in its environment, with managers seeking resources and innovating in response to environmental (or market) demands. The internal process approach reflects a Tayloristic concern with formalization and internal control of the system in order that resources are efficiently used. Finally, the human relations approach reflects the tradition derived from socio-technical theory, emphasizing the well-being, growth, and commitment of the community of workers within the organization. By combining these orientations into one model, Quinn and Rohrbaugh aimed to provide a broad conceptual map of the domains of organizational theory over the twentieth century. Such a map is useful in identifying the required topography of a climate measure applicable to a wide range of organizations. The model is also useful in reflecting the means for implementing those values in terms of managerial practices and the ends or outcomes which are emphasized or which compete in each domain. The model does not propose that organizations will locate predominantly in one quadrant, but, reflecting the rich mix of competing views and perspectives in organizations, proposes that organizations will be active and give emphasis to each domain but with differing strengths.

5. Climate Only Exists if We Agree About it

The research on organizational climate, however, is hampered by a particular methodological debate. It concerns the practice of aggregating individual employees' scores on climate dimensions and taking the mean as a representation of organizational climate. Clearly, simply using the mean may mask profound differences between organizations and the variation of responses given by employees. Indeed, research within organizations has shown clear differences between hierarchical levels and departments in ratings of climate, suggesting a number of subclimates within each organization. When the level of analysis changes from the individual to the organization, the validity of the data is uncertain. The validity of aggregate climate used to describe organizations depends upon the demonstration of agreement between individual members and their perceptions. Conceptual agreement implies a shared assignment of psychological meaning under which circumstances and individual perceptions may be aggregated and treated as higher-order constructs. In such circumstances, the use of aggregated individual data to measure organizational climate is appropriate.

The main methodological problem with this latter approach, a demonstration of shared conceptions for aggregate climates, has provoked much debate in the research literature. However, writers have not specified what is an appropriate level of agreement. Some suggest that dichotomous questionnaire items should be used to measure climate and that the only items to be treated as descriptive of organizational climate should be those in which the frequency of endorsement is not significantly different from 0 percent or 100 percent. Others have suggested that 66 percent was reasonable, but few researchers have followed either prescription.

The validity of aggregating climate perceptions to the organizational level is further called into question by the demonstration that there are significant differences between members of different organizational subgroups in their perceptions of climate. Such differences have been found between groups within the same organization, based on factors such as hierarchical level, departments, divisions, regions, and work groups (Patterson et al. 1996).

Another approach to overcoming the aggregation problem is to examine the level of inter-rater agreement between individuals within an organization, in relation to their perceptions of climate. One approach is to use the intraclass correlation coefficient. This coefficient is based on a one-way analysis of variance and assesses the ratio of variation within organizations to variation among organizations. This coefficient has been criticized as inadequate and led to the development of the within-group agreement index of multiple item scales developed by James et al. (1993). This is a technique for assessing agreement among the judgements made on multiple item scales. There has been continuing debate about the advantages and disadvantages of this orientation, although the measure continues to be widely used. It seems likely that organizational climate measures will continue to be used, despite there being no simple or widely agreed criteria for aggregation. Part of the problem is that the level of agreement within organizations will depend partly on the measures employed, the context, and the domains tapped by the measures.

6. Climate vs. Culture

A further debate in this area is over the distinction between climate and culture. Both of these concepts are root metaphors, which are very different in orientation. Organizational culture is taken to refer to the underlying patterns or configurations of interpretations of formal practices, such as pay levels, hierarchy, and job descriptions; and informal practices such as norms, espoused values, rituals, organizational stories, jargon, humor, and physical environment. 'The underlying patterns or configurations of these interpretations and the way they are enacted, constitute culture' (Martin 1995). In general, researchers agree that climate is a measure of the surface manifestations of culture and is not entirely distinct from culture. Most researchers argue that culture can only be measured by qualitative methodologies, whereas climate as a more superficial characteristic of organizations can be assessed using quantitative questionnaire measures. However, these distinctions are largely artificial and it is true that both concepts of climate and culture are the subject of continuing debate because of their lack of theoretical and conceptual underpinnings.

7. Conclusions

Because organizational climate is a phenomenon that researchers experience in practice, in working in different organizations, and because the measures of organizational climate relate to important outcome measures, it is likely that the concept will continue to be used. It does however, remain a poorly defined and weakly theorized concept in the organizational sciences.

See also: Ethics in Organizations, Psychology of; Individual Identities in Organizations; Organizational Culture; Organizations: Authority and Power; Psychological Climate in the Work Setting

Bibliography

Campbell J P, Dunnette M D, Lawler E F, Weick K E 1970 *Managerial Behavior, Performance and Effectiveness*. McGraw-Hill, New York

Denison D R 1990 *Corporate Culture and Organizational Effectiveness*. Wiley, New York
James L R, Demarée R G, Wolf G 1993 r_{wg}: An assessment of within-group interrater agreement. *Journal of Applied Psychology* **78**: 306–9
James L R, Jones A P 1974 Organizational climate: A review of theory and research. *Psychological Bulletin* **81**: 1096–112
Lewin K, Lippitt R, White R K 1939 Patterns of aggressive behavior in experimentally created social climates. *Journal of Social Psychology* **10**: 271–99
Martin J 1995 Organizational culture. In: Nicholson N (ed.) *Blackwell Encyclopaedic Dictionary of Organizational Behavior*. Blackwell, Oxford, UK, pp. 376–82
Patterson M, Payne R-L, West M 1996 Collective climates: A test of their sociopsychological significance. *Academy of Management Journal* **39**: 1675–91
Patterson M G, West M A, Lawthom R L, Nickell S 1997 *People Management and Business Performance*. Institute for Personnel and Development, London
Payne R 1990 Madness in our method. *Journal of Organizational Behavior* **11**: 77–80
Payne R, Mansfield R 1973 Relationship of perceptions of organizational climate to organizational structure, context and hierarchical position. *Administrative Science Quarterly* **18**: 515–26
Quinn R E, Rohrbaugh J 1981 A competing values approach to organisational effectiveness. *Public Productivity Review* **5**: 122–40
Schneider B 1990 The climate for service: An application of the climate construct. In: Schneider B (ed.) *Organizational Climate and Culture*, 1st edn. Jossey-Bass, San Francisco, pp. 383–412
Schneider B, Reichers A E 1983 On the etiology of climates. *Personnel Psychology* **36**: 19–39
West M A, Patterson M G 1999 The workforce and productivity: people management is the key to closing the productivity gap. *New Economy* **6**: 22–7

M. A. West

Organizational Control

1. Mainstream Approaches: Ambiguities and Absences

As Pfeffer notes, 'control is at once the essential problem of management and organization and the implicit focus of much of organization studies' (1997, p. 100). Yet there are wide disagreements as to its significance and substance. In mainstream models of organization behavior control is, at best, treated ambiguously. Some standard textbooks simply don't have chapters on the issue. Systems theory is based on equilibrium and interdependence, therefore tends to assume goal consensus and downplays systematic conflicts of interest or identity in the employment relationship.

When control is discussed it is often alongside coordination. Any complex division of labor requires mechanisms to set goals, allocate responsibilities, and evaluate the effectiveness of performance. Coordination is a more neutral term than control and more compatible with an assumption of goal consensus. In this context, control tends to be treated as a *monitoring* device, with management's role to check progress, and ensure that actions occur as planned and to correct any deviation. The emphasis is on technical inputs and outputs in a self-adjusting system, performance standards and feedback mechanisms.

Such emphases can, however, be misleading. Control often *is* present in mainstream approaches, it's just hiding in the shadows of something else. The key term used by Pfeffer is *implicit*. Texts that do have chapters with control in the title often pass over quickly into other issues such as job design, organizational structure, or leadership. In the first instance, debate is focused on designing structures which facilitate levels of control and coordination appropriate to types of work that require different levels of discretion and standardization. In the second, discussion of leadership styles such as the classic polarity of authoritarian vs. democratic, are ways of discussing control, but through the language of influence and motivation.

1.1 A Break from Command and Control?

Reinforcing this ambiguity about whether control is necessary, the more practical management theorists have continually tried to abolish it or proclaim its imminent disappearance. From the human relation's writers of the 1950s, with their distinctions between (bad) theory X and (good) theory Y, to the advocates of 1970s-style job enrichment and contemporary empowerment, control has been presented as destructive and out of date. The idea that a high degree of autonomy over their own work was a precondition for job satisfaction and high productivity has been a cornerstone of organizational psychology. From the mid-1980s onwards *necessity* was added to desirability. Influential articles such as Walton (1985) argued that new competitive pressures and the enhanced demands of a service-oriented, knowledge-based economy meant that 'command and control' was no longer an option for successful business. We were moving from control to commitment and from rules to values as means of coordination. The future was empowered employees in self-directing teams, participating in and sharing responsibility for a wider range of decisions—more voice and choice at work.

A problem with such approaches is their aspirational nature and the subsequent tendency to conflate prescription and description. It is accepted across different perspectives that changes have taken place with respect to controls. In particular there has been, through team working, increased involvement of employees in microlevel decision-making about how to do their work—a form of democratized Taylorism

(Adler 1993). However, studies are consistent in showing that most organizations, at least in Britain and North America, remain traditionally managed and continue to take a low trust, low skill path to competitiveness (Milkman 1997, Thompson and Warhurst 1998). In the context of generally tougher times, marked by downsizing and delayering, many employees experience empowerment largely as a form of work intensification. Firms are finding it harder to deliver on the kind of 'package' envisaged by Walton, where increased commitment and responsibilities would be offset by greater job security and other enhancements to the employment relationship.

While companies are indeed seeking to enhance service quality and efficiency, they are also investing in new control systems to monitor the consistency and characteristics of those outputs. In services generally there is ample evidence that organizations are scripting employee–customer interactions (Ritzer 1993), or specifying feelings rules for the mobilization of emotional labor (Hochschild 1983).

What the above discussion suggests is that mainstream conceptual frameworks are not robust enough to deal with changes in the *types* and *levels* of control. For example, it is perfectly possible to have increased operational autonomy for 'empowered' managers, existing alongside much tighter financial controls through targets and budgets.

2. *Control: The Basic Building Block of Radical Organizational Theory*

As Pfeffer (1997, p. 135) notes, those ambiguities towards the idea of control are in part responsible for the development of more critical perspectives on organizations. For a variety of radical theories, control is the core concept. Whereas mainstream perspectives treat control and coordination together, radical theorists argue that management performs a *dual* function in the enterprise. Managerial practices are a necessary means of *coordinating* diverse activities, however, they also bear the imprint of conflicting interests in the labor process, a conflict which reflects the unique nature of labor as a commodity. When capital purchases labor it has only a potential or capacity to work. To ensure profitable production, capital must organize the conditions under which labor operates to its own advantage. But workers pursue their own interests for job security, higher rewards, and satisfying work, developing their own counterorganization through informal job controls and output norms.

To resolve this problem and because they are under competitive pressure from other firms to cut costs and raise productivity, employers seek to control the conditions under which work takes place. Control is not an end in itself, but a means to transform the capacity to work established by the wage relation into profitable production. This idea of the transformation of labor power into labor is taken from Marx and is the starting point of what has become known as *labor process theory* (LPT). Its adherents describe control as a set of mechanisms and practices that regulate the labor process. Richard Edwards (1979) distinguishes three elements in any system of control:

(a) direction and specification of work tasks,

(b) evaluation, monitoring, and assessment of performance,

(c) the apparatus of discipline and reward to elicit cooperation and compliance.

Such elements may, however, be best described as *detailed* control, in that they are normally connected to immediate work processes, whereas *general* control refers to management's capacity to subordinate labor to their direction of the production process as a whole. This allows for recognition of tremendous variations in how detailed control is exercised. Such a model can even allow for employers giving workers significant discretion over tasks, as in semiautonomous work groups, *if* it maintains their overall control. Control is also not absolute but at least at the microlevel, a contested relationship.

2.1 *Strategies of Control: Debates and Disagreements*

Radical perspectives have been conditioned by Braverman's (1974) argument that the twentieth century has seen the tightening of managerial control, primarily through the application of Taylorist and scientific management strategies. He provided an important corrective to the widespread view that Taylorism was a failed system, superseded by more sophisticated behavioral theories to be used for motivational and job design tools. But it is widely recognized that Braverman overestimated the dominance of Taylorist strategies and practices, and underestimated the varied and uneven implementation.

Nevertheless, because Braverman confused a particular system of control with management control in general, the question of *alternative strategies* was put firmly on the agenda. The best known is Richard Edward's (1979) model that is based on historically successive dominant modes of control, which reflect worker resistance and changing socioeconomic conditions. A nineteenth century system of *simple* or *personal* control by employers exercising direct authority gave way to more complex *structural* forms with the transition from small business, competitive capitalism, to corporate monopolies. The first of these forms was *technical* control, typified by the use of the assembly line, which can pace and direct the labor process. The contradiction for management is that it created a common work experience and basis for unified shop floor opposition. In contrast, a system of *bureaucratic* control embedded in the social and

organizational structure of the firm rather than in personal authority offers management a means of redividing the workforce and tying it to impersonal rules and regulations.

In contrast, the other most influential model from Friedman (1977) eschews the notion of stages, preferring to set out ideal types or strategic poles of responsible autonomy and direct control which run parallel throughout the history of capitalism. Paralleling mainstream notions of mechanistic and organic structures, or low and high trust employment relations, each strategy is said to generates its own inflexibilities in areas such as hiring and firing and task specification. The choice of strategy is governed by variations in the stability of labor and product markets, mediated by the interplay of worker resistance and managerial pressure.

While the emphasis on alternative strategies was welcome, they were criticized for the 'panacea fallacy': the idea that capital always seeks and finds definitive and comprehensive modes of control as the solution to its problems. What has happened in practice is that researchers influenced by LPT have largely ignored such overarching frameworks and used the particular concepts such as technical or bureaucratic control to explain particular trends in new technology, recruitment practices, or internal labor markets. Alongside such usage, other researchers have opened up the empirical and theoretical terrain by demonstrating how control processes inside work or shaped by the broader social divisions of labor are connected to gender and family relations.

3. *Challenges, Critiques, and Responses*

While a degree of common ground emerged both about the varied influences on control strategies and the variations in combination and context through which control is exercised, critics developed deeper objections to the whole approach.

3.1 *Questioning the Centrality of Control*

One strand concerns modern versions of long-standing issues about the relations between ownership and management. Capital's interests are not given and management practices cannot be 'read-off' from them. Assumptions of a 'tight-coupling' underestimate the diversity and complexity of such practices, and the significance for decision-making processes within the enterprise. In addition, there are conflicts within management reflecting contending interest groups and the difficulty of carrying out integrative functions.

A second strand of critique questions whether the centrality given to control of labor is actually reflective of managerial behavior. It is argued that we cannot view management strategies and tactics primarily from the vantage point of the labor process. If anything, managers are dominated by problems of the outcomes of that process, including sales, marketing, supply, and cash flow. The concept of *the full circuit of capital* is used to argue that we must be concerned not only with the *extraction* of surplus value through controlling the labor process, but its realization through the sale of commodities, as well as the prior purchase of labor.

It was widely accepted within LPT that some of these objections were a necessary corrective to a debate that had often lost sight of the role of the broader political economy because of the emphasis on *management* control. However, it was argued that the management of workers and work remains at the heart of the enterprise and indeed of economic reproduction as a whole. In addition, while management strategies and practices are not always developed with labor's role in mind, it is ultimately difficult to separate a concern with 'outcomes' such as product quality or financial targets from acting on labor in some way.

LPT has also attempted to develop a more complex framework that can deal with the tensions between different management levels, functions, and ideologies in framing and exercising control. The most innovative is associated with Armstrong's work on *inter-professional competition* (1989). Management functions for capital are mediated by competition between occupational groups to become the trusted agency to carry out control functions. Armstrong charts the historical and contemporary forms of competition between engineers, personnel, accounting, and other professions within the business enterprise. Who comes out on top is a function not just of their power resources, but the societally specific constitution of managerial expertise.

3.2 *Beyond External Control: The Final Frontier?*

The other significant challenge to established control theories wishes less to deny the concept and more to shift its locus. What was discussed earlier as a move from control *to* commitment can be seen instead as a form of *internal control* that does not rely on external rewards and sanctions, or rule-following. Outside the largely consensual corporate culture literature, theorists within a labor process tradition had already gone some way down this route. Burawoy (1979) had shown how employee participation in production games and workplace rituals generated consent to existing social relations, constituting workers as individuals rather than as members of class. In other words they 'helped' to control themselves. Later, Ray (1986) argued that control by corporate culture was the last frontier, in that it had enabled organizations to generate sentiment and emotion, simultaneously

internalizing control and linking personal with corporate identity.

A further twist to such arguments has come through more recent influences from Foucault and post-structuralism. In an influential article, Sewell (1998) claims to have identified a new model of labor process control that counters the optimistic gloss of the empowerment and teamwork literatures, while moving beyond the confines of traditional LPT. 'Panoptic' forms of control augment existing practices through the interaction of 'vertical' electronic surveillance and the 'horizontal' dimension of peer group scrutiny within teams. Sewell makes much of the enhanced capacity of management to collect, display, and attribute performance data. With respect to teams, he relies on the well-known case studies of Barker (1993) and to a lesser extent McKinlay and Taylor (1996) to make a case for horizontal control. Here, team members create the intersubjective conditions for compliance by creating disciplinary rules within delegated authority, thus collaborating with management to identify and reward the 'good worker.' The implication of Sewell's argument is that such a combination solves the 'control-engage' dilemma identified by LPT and suffered in different ways by generations of managers. The new techniques allow management to rationalize knowledge and work, but not through obtrusive Taylorist and bureaucratic means, while mobilizing and manipulating the subjectivity of employees.

A range of objections can and have been made to this latest twist to accounts of organizational control (Thompson and Ackroyd 1995). Changes in the terrain and frontier of control take place continually and it would be difficult to deny that *some* shift towards electronic and self-surveillance has taken place in *some* industries. However, few of its enthusiasts adequately consider issues of time, place, and people—in other words, the generalizability of such controls outside of the high-tech, strong culture firms from which the evidence largely comes. The central problem, however, is confusion of the formal capacities of technological and managerial systems with their actual usage and effectiveness. This is compounded by a tendency to draw evidence primarily from managerial sources, underestimating the capacity of labor to find individual and collective spaces to dissent or resist. Where these problems are considered more fully, as in the McKinlay and Taylor case study of an electronics plant in Scotland, a more rounded and complex picture of the gap between intent and outcome is revealed. Finally, it is not only debatable how *internalized* the new controls are, but also how *internal*. Normative rules in strong culture companies may be different from traditional task-based control, but they are still rules. Management's sources and use of information may have increased, but there is little evidence that is unobtrusive or regarded by employees as objective and unconnected to visible authority relations. In other words, it may be too early to proclaim the death of traditional controls in a post-bureaucratic era.

4. Conclusion

Management may always be looking for the final frontier, but like the crew of Star Trek, they never really find it. This is hardly surprising. Conflicts of interest and constraints of circumstance run too deep. A close examination of contemporary organizational trends and their analysis reveals considerable continuity in theory and practice. Electronic surveillance can look remarkably like Edward's concept of technical control; the exaggerated claims for empowerment strongly resemble those made for work humanization in the 1970s. Sometimes theorists are describing the same practices in a different conceptual language. Willing internalization in one framework can be imposed normative controls in another. It would be foolish, of course, to claim that nothing changes. For example, changes in the labor force and broader cultural norms can bring new problems such as the rising tide of sexual misbehavior at work. In turn this creates the conditions for new controls such as codes of conduct that specify appropriate boundaries of behavior and what sanctions might follow if they are transgressed. While work remains a contested terrain there will be a frontier of control. It is the task of organizational analysis to explain the character, combination, and circumstances of that changing frontier.

See also: Administration in Organizations; Authority: Delegation; Authority, Social Theories of; Conflict: Organizational; Control: Social; Human Resource Management, Psychology of; Management: General; Organization: Informal; Organization: Overview; Organizational Behavior, Psychology of; Organizational Culture; Organizational Decision Making; Organizations: Authority and Power; Organizations, Sociology of

Bibliography

Adler P S 1993 Time-and-motion regained. *Harvard Business Review* January–February **71**: 97–108

Armstrong P 1989 Management, labor process and agency. *Work, Employment and Society* **3(3)**: 307–22

Barker J R 1993 Tightening the iron cage: Concertive control in self-ranging teams. *Administrative Science Quarterly* **38**: 408–37

Braverman H 1974 *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. Monthly Review Press, New York

Burawoy M 1979 *Manufacturing Consent: Changes in the Labor Process Under Monopoly Capitalism*. University of Chicago Press, Chicago

Edwards R 1979 *Contested Terrain: The Transformation of the Workplace in the Twentieth Century*. Heinemann, London

Friedman A L 1977 *Industry and Labour: Class Struggle at Work Monopoly Capitalism*. Macmillan, London
Hochschild A R 1983 *The Managed Heart: Commercialization of Human Feeling*. University of California Press, Berkeley, CA
McKinlay A, Taylor P 1996 Power, surveillance and resistance: Inside the factory of the future. In: Ackers P, Smith C, Smith P (eds.) *The New Workplace and Trade Unionism*. Routledge, London
Milkman R 1997 *Farewell to the Factory: Auto Workers in the Late Twentieth Century*. University of California Press, Berkeley, CA
Pfeffer J 1997 *New Directions for Organizational Theory: Problems and Practices*. Oxford University Press, Oxford, UK
Ray C A 1986 Corporate culture: The last frontier of control? *Journal of Management Studies* **23**(3): 287–97
Ritzer G 1993 *The McDonaldization of Society*. Pine Forge Press, Newbury Park, CA
Sewell G 1998 The discipline of teams: The control of team-based industrial work through electronic and peer surveillance. *Administrative Science Quarterly* **43**(2): 397–428
Thompson P, Ackroyd S 1995 All quiet on the workplace front? *Sociology* **29**(4): 615–33
Thompson P, Warhurst C (eds.) 1998 *Workplaces of the Future*. Macmillan, London
Walton R E 1985 Towards a strategy of eliciting employee commitment based on policies of mutuality. In: Walton R E, Lawrence P R (eds.) *Human Resource Management, Trends and Challenges*. Harvard University School Press, Boston

P. Thompson

Organizational Culture

The concept of organizational culture typically is taken to refer to the beliefs, values, behavior patterns, and understandings that are shared by members of an organization and which are distinctive of it. These may be associated with organizational symbols, stories, and myths that are expressive of the culture and help to socialize people into the organization; or images, products, clothing, buildings, and other artefacts that materially embody the culture. The concept is often distinguished from that of corporate culture (see *Corporate Culture*) as being more naturally occurring and less consciously constructed by managers. However, there are considerable differences between scholars on how culture should be defined and how it should be studied.

1. Defining Organizational Culture

One of the problems of culture as a concept is that it must always be seen in relation to some other concept—such as structure or strategy (see *Structural Contingency Theory*; *Strategy: Organizational*). In organization studies concepts of culture were borrowed, with few exceptions, from only one anthropological tradition—that of structural functionalism—when the concept of organizational culture became popular in the late 1970s and early 1980s. Much of the history of the concept since then has been of a process of appreciating its complexity and subtlety by incorporating perspectives from other disciplines, such as sociology and social psychology, as well as more recent developments in anthropology and social theory, such as postmodernism, which challenge the functionalist paradigm (see *Organizations, Metaphors and Paradigms in*).

1.1 Levels of Culture

Additional problems with defining the concept are associated with what have been called 'levels' of culture. Culture in advanced societies can be identified at the *national* level (usually but not always coincident with geopolitical boundaries), at the *regional* level (most meaningfully within geopolitical boundaries but again not exclusively) at the *organizational/institutional* level, and at the level of *subcultures* within organizations based around formal (e.g., departmental, site) or informal social groupings (see *Organization: Informal*). However, subcultures may also exist at national and regional levels, and may cut across organizational boundaries (see *Boundaries and New Organization Forms*); they may exist among industries rather than companies (see Turner 1971), and even among professions. Subcultures may be oppositional to dominant cultures (see *Conflict: Organizational*), at all levels considered, may become sources of opposition in relation to particular issues (see *Organizations: Negotiated Orders*), or they may constitute a leading edge *avant-garde* from which the rest of the organization takes its inspiration (see *Innovation: Organizational*). Diversified organizations in particular may intersect with several regional and national cultures such that it becomes difficult to identify the culture of the organization as separate from the local and national cultures in which it is embedded. Some writers distinguish between the culture of top management, or the official corporate culture, and the organizational culture, or culture of the workplace. Organizational or workplace culture is that which 'grows or emerges within the organisation and which emphasises the creativity of organisational members as culture-makers, perhaps resisting … the dominant culture' (Linstead and Grafton-Small 1992).

1.2 Definitional Elements and Problems

One division between commentators relates to the extent to which they emphasize the more obvious, concrete, and conscious aspects of culture as distinct from its tacit, submerged, and unconscious qualities. Early definitions of culture tended to emphasize the visible and behavioral, where more recent sociological definitions emphasize ideological assumptions more than the behavioral norms.

Schein (1985) defines culture from the perspective of a social psychologist as 'the deeper level of basic assumptions and beliefs that are shared by members of an organisation, that operate unconsciously, and that define in a basic, 'taken-for-granted' fashion, an organisation's view of itself and its environment.' Schein developed an influential model, grounded in small group psychology rather than anthropology or sociology, which identifies three levels of culture: *artefacts and creations* (objects, buildings, uniforms, technology, including behaviors, etc.); underpinned by *values* which are not visible but of which we are or can be made aware; and *basic assumptions* which are taken for granted, invisible, preconscious and hard to access. Furthermore, he argues that the culture reveals itself when it is most stressed, when presented with problems, rather than in its daily routine.

Because of the complex and difficult-to-access nature of human understanding, researchers will give attention only to certain dimensions of culture in any investigation, and neglect others. However, the elements that they will need to address, whether or not they decide to leave some of them out of consideration will include the extent to which a culture is distinctive or unique; the extent to which it is a guide to action, a prescription or landmark for members; the extent to which meanings are shared or unshared, explicit, or implicit; the content of common-sense or local knowledges; the typicality of attitudes and the consistency of emotional climate; the extent to which culture acts as a means of problem-solving, resolution, or containment of conflict or dilemmas; the importance of behavioral norms and their transgression; the character of patterned responses and actions (rites and rituals); the development and stability of the culture over time; forms of symbolic expression and communication in myths, rhetoric, and storytelling (see *Organizations, Metaphors and Paradigms in*; *Myths and Symbols: Organizational*); the significance of material objects, artefacts, and concrete phenomena; the significance of the immaterial, values, beliefs, and ideology; the effects of power, knowledge, and interests (see *Organizations: Authority and Power*); to what extent culture functions as a means of control (see *Organizational Control*); the nature of social motivation and commitment to the culture; and the recognition and importance of subcultures.

2. *The Importance of the Organizational Culture Concept*

The organizational culture concept became important both because of what it drew attention to and what it enabled. It underlined the fact that organizations were multidimensional, and that how individuals made sense of and felt about them could often be more important than what figures revealed about them. It placed emphasis on the hidden creative potential of human beings to transform organizations, and the importance of what happens in everyday relations rather than solely in the annual company report. It released a surge of interest in the human dimensions and 'softer' elements of organizing, providing a platform for human resources specialists to argue for the greater alignment of leadership styles, vision, recruitment methods, reward structures, and informal methods of management in relation to strategic objectives. The downside to this was that it became associated with faddish 'quick-fixes,' which were the antithesis of what culture was supposed to represent, and dramatic company turnarounds were attributed to changes in the 'culture.' For more careful observers, it drew attention to deeper elements, often obscured and unconscious, that obstructed or held back development and change, and which were in themselves hard to change. It underlined the need to move carefully, sensitively, and at a measured pace in transforming organizations in ways which would be sustained. Conceptually, despite the fact that some scholars immediately looked for what to measure in order to change culture, the concept delivered an awareness of the increasingly great complexity of organizations and organizing processes, and the need to develop richer analyses to do this complexity justice. Despite its unfortunate association with faddishness, culture became and remains a core concept in organization theory having transformed our understanding of human resource management, organizational change management, and the implementation of strategy.

3. *The Historical Development of the Organizational Culture Concept*

The idea of culture in relation to organizations has a long but tortuous history. From the 1920s, at least, it was recognized that the social dimensions of work are important elements of effectiveness through the Hawthorne studies. Jaques (1952) first coined the term culture in relation to work organization in *The Changing Culture of a Factory* which was part of a series of accounts of participatory management in the Glacier Metal company in the UK, although structural elements, reward systems, and the use of hierarchy were also important to the success of the project.

In the 1950s Gouldner, a sociologist, also identified the importance of things that were taken for granted, implicit dimensions in *Wildcat Strike* (1955). Gouldner noted the importance of an implicit system of concessions, tolerated transgressions and obligations called the *indulgency pattern* that was necessary to maintain the formal order of the organization. Similarly, organizational psychologists such as Chris Argyris noted the importance of subconscious dimensions of organization and its psychological health, noticing in particular the difference between what is

espoused by organizational members, and what they actually put into *practice*.

Another related development in the 1960s was the discovery of *negotiated order theory* which was based on work done in psychiatric hospitals by Anselm Strauss and co-workers. From the perspective of culture this work is important because it emphasizes that culture can cut across professional and nonprofessional boundaries; it is associated with typical organizational practices (e.g., care regimes); it has practical, moral, and ethical elements; it has implicit and explicit aspects; it is emergent and relationships shift over time; and it is related to organizational power and resource allocation (see *Organizations: Negotiated Orders* for a full discussion of this topic).

Around the same time, Harold Garfinkel was developing ethnomethodology which was a form of sociology which concentrated on the ways in which people make sense of their social situations, which stressed the importance of unspoken rules, talk, common sense, and the taken-for-granted aspects of social life. The idea of *membership* was also important to Garfinkel, and particularly the things people had to learn to become a 'member' of a social group. Much of Garfinkel's work overlapped with the work of anthropologists, who customarily studied exotic societies. Turner (1971), who was influenced by the work of Garfinkel, published the first book to bring the two disciplines together in looking at the way stories, rites, rituals, and humor shaped behavior in organizations.

Concerned with the nonobvious, and the importance of the implicit and taken-for-granted in forming our experience of organizations, early *Organizational Development* (*OD*) specialists began to recognize the significance of the unsaid as a barrier to transformation. They often argued that their work was to bring out the unconscious obstacles to organizational change, as a form of *cultural intervention*. In contrast to more recent literature, early antecedents emphasized the implicit and unconscious elements of experience and the *processes* of sense-making and meaning-making rather than the *content* of communication and the explicitly expressed values.

The one exception to this rule were the *organizational climate* studies which concentrated on the experiences and emotional responses which culture produces in its members. Climate is accessible and close to the surface, associated with attitudes, and there was a pronounced tendency for researchers in this tradition to want to develop instruments to measure it. As Alvesson and Berg (1992, pp. 88–9) note, much of the more positivistic work in the area of corporate culture in recent years would be better characterized as work on organizational climate.

Conceptual development during the 1970s began to take on two aspects—the development of the understanding of symbolic processes, albeit in a predominantly functionalist framework on how symbols work and the types and functions of stories, including images, rites and rituals, storytelling, humor, and what was coming to be understood as *organizational symbolism*; and the application of the culture concept to the process of changing organizations, sometimes referred to as *cultural or value engineering*. The development of the first aspect began somewhat earlier but was slower to mature; the second aspect had a dramatic impact on both the scholarly and corporate landscapes (Peters and Waterman 1982, Deal and Kennedy 1988).

3.1 Changes in Focus or Emphasis over Time

In the late 1970s culture took on a dimorphic emphasis on either organizational symbolism or corporate culture, its management and change. Culture was often seen, especially in the latter approach, as a *property* of an organization, something it has, and research questions tended to be framed according to whether culture was treated as an independent or a dependent variable. Positivistic approaches concerned themselves, with isolating the determinants of culture and measuring their effects, or identifying the properties of culture and measuring their strength. Some studies attempted to determine whether there was a relationship between the possession of a 'strong' or easily identifiable culture and success. Organizations themselves however, could be seen as types of culture, and some researchers attempted to typologize cultures, and infer their appropriateness for particular activities (see *Structural Contingency Theory*). Popular formulations of *cultural typologies* had little research base, and often confused cultural characteristics with those of the organization's structure or its task. All of these approaches are still in evidence in some of the current work.

An alternative approach regarded organizations as the outcome of cultural processes, and some researchers regarded this process as being two-way—that is organizations were also producers of cultural forms in a constant interaction. Scholarly conferences produced several seminal collections in the US, and these tended to contain contributions from both the *culture-as-property* school and the *culture-as-process* school, although the latter tended to lack much of a micropolitical perspective (see Frost et al. 1991). Trice and Beyer (1993) produced the definitive functionalist work, while Martin (1992) attempted to develop a multiperspective framework which incorporated ambiguity, from slightly to the left of the theoretical mainstream (see Martin and Frost 1996 for an extension). In Europe in 1981 the *Standing Conference on Organizational Symbolism and Corporate Culture* was formed, and its members worldwide have contributed significantly to the development of the concept over the years. The European commentators, reflecting their rather different intellectual tradition, have been a consistent source of critical treatments of

culture in which power and conflict of interest are central rather than externalized, and culture as a means of control is a focus. Alvesson and Berg (1992) produced a seminal critical review of the concept which emphasized the political and ideological dimensions of culture and Linstead and Grafton-Small (1992) produced a model of a *postmodern* approach to organizational culture that argued that culture as a phenomenon was inherently paradoxical. Subsequent postmodern treatments have tended to concentrate on culture as a form of *seduction* (in which organizational members are subject to an artificially manipulated sense of reality) or as a form of *surveillance* (in which they become self-policing cultural subjects). More recently Linstead et al. (1996) collected a number of ethnographic papers that argue most strongly that culture is a product of *difference of views*, being an ambiguous way to accommodate them at a generalized level rather than an expression of specific *shared understandings*.

4. *Emphases in Current Theory and Research*

The study of culture has given rise to a number of new directions for analysis. Perhaps the most obvious of these is the study of *language processes*, talk, rhetoric, narrative, storytelling, and deconstructive approaches. Related, emphasizing language and understanding, but from a different theoretical perspective, are a range of *cognitive approaches* grounded in psychology and cognitive anthropology. These approaches often concentrate on decision-making and learning processes, and the connection between culture and strategy as mediated through knowledge practices. They are also often associated with issues of *corporate identity*, and its relation to individual identity, another established and growing research area. Yet another important area has been the study of *emotions*, emotional labor, and the management of emotions in organizations, often taking a critical view of some of the more popular treatments of 'emotional intelligence.' Via developments in the theory of symbolism and symbolic processes, including semiotics, connections with the arts and humanities have been made as evident in the new field of *organizational aesthetics*. Where visual understanding is an element here connections have been made with marketing and advertising. At a practical level, there has also been an increase in treatments of cultural industries and performing arts organizations, and links have been drawn with cultural and media studies. Finally, *gender* issues recently have emerged to importance (Gherardi 1995).

4.1 *Methodological Issues in Investigating Culture*

There is considerable debate on how best culture can be investigated, which reflects some of the turmoil within anthropology itself. The debates are ongoing and centered around whether culture is observable, and if so whether it is measurable and how to measure it. There is also debate around whether culture consciously can be recognized and recounted by members, and the appropriateness of surveys, questionnaires, or even focused interviews as a means of locating and accessing culture. On the other hand, there are debates over whether and how fragmented and/or inferred by the observer culture is, and the consequences for participant observers. These issues are not resolved easily and because of the essentially contested nature of the culture concept, are likely to continue.

5. *Some Future Directions*

Some areas of interest already mentioned currently seem robust and likely to increase in importance, such as change management, postmodernism, and the treatment of gender and sexuality in relation to culture. *Race and ethnicity* have as yet been treated by only a few pieces of research and this must be an important area in which growth should be encouraged. The popular area of *teams and team working* has only been incidentally informed by work from the perspective of culture. The massive growth of work on *knowledge management* and transfer, innovation, and technological change remains to be fully informed by its conceptualization in terms of culture. New forms of work, including *virtual organizations* will surely have cultural impact, and the cultural consequences of *downsizing and destructuring* have not been treated in any depth from this perspective. Conceptually, the incorporation of insights from the 'new science' such as *emergence*, *chaos*, *and fractal logic*, has only just begun. Many of the critiques of the culture approach accused it of ignoring power, and although more recent treatments influenced by Michel Foucault have addressed culture by considering *power*, *discourse*, *and subjectivity* much remains to be done. Indeed, the extension of the work of Foucault into considering the importance of the body and *embodiment* in the social sciences has still had little impact in organization studies. As yet only small amounts of work have been done in some promising areas which offer exciting prospects for the continuing study of organizational culture.

See also: Organizational Climate; Organizations, Sociology of

Bibliography

Alvesson M, Berg P O 1992 *Corporate Culture and Organizational Symbolism: An Overview*. Walter de Gruyter, Berlin

Deal T E, Kennedy A A 1988 *Corporate Cultures: The Rites and Rituals of Corporate Life*. Penguin, Harmondsworth, UK

Frost P J, Moore L F, Louis M R, Lundberg C C, Martin J (eds.) 1991 *Reframing Organizational Culture*. Sage, Newbury Park, CA
Gherardi S 1995 *Gender, Symbolism and Organizational Cultures*. Sage, London
Gouldner A W 1955 *Wildcat Strike: A Study of an Unofficial Strike*. Routledge and Kegan Paul, London
Jaques E 1952 *The Changing Culture of a Factory*. Dryden Press, New York
Linstead S A, Grafton-Small R 1992 On reading organizational culture. *Organization Studies* **13**(3): 331–55
Linstead S A, Grafton-Small R, Jeffcutt P S (eds.) 1996 *Understanding Management*. Sage, London
Martin J 1992 *Cultures in Organizations: Three Perspectives*. Oxford University Press, Oxford, UK
Martin J, Frost P 1996 The organizational culture war games: A struggle for intellectual dominance. In: Clegg S R, Hardy C, Nord W R (eds.) *The Handbook of Organization Studies*. Sage, London pp. 344–67
Peters T J Jr., Waterman R H 1982 *In Search of Excellence: Lessons from America's Best-run Companies*. Harper and Row, New York
Schein E H 1985 *Organizational Culture and Leadership*. Jossey Bass, San Francisco
Trice H M, Beyer J M 1993 *The Cultures of Work Organizations*. Prentice-Hall, Englewood Cliffs, NJ
Turner B A 1971 *Exploring the Industrial Subculture*. Macmillan, London

S. A. Linstead

Organizational Culture, Anthropology of

When asked what they understand by organizational culture, most employees and managers will undoubtedly reply: 'How we think and act in our organization.' If they are asked to explain further, large differences generally appear both within and between the two groups with regard not only to how they think and act but also to which behavior is viewed positively. To a large extent this is due to the fact that organizational culture is interpreted by different people in different ways. Some take it to be a coherent set of values and norms relating to how they deal with one another, and/or with their superiors, and/or with clients. For others the term organizational culture refers above all to the ritual and symbolic behavior of employees and managers, while yet others consider it to boil down to the informal organization. There is little agreement on the right definition and the kind of research to be applied to it in the academic world either. Like the mass of current definitions of culture, there are clearly divergent views of what organizational culture means and of what an ideal organizational culture is. Concomitantly, there is a wide range of research approaches to it. Ultimately, the different definitions and research approaches can be derived from differences in epistemological, methodological, and theoretical paradigms and the related coherent visions of the phenomenon of organizational change. The representatives of the different approaches are organized separately to a large extent, they often publish separately from one another, and they attack one another from their own bastions. This article presents the main differences in academic approaches to organizational culture and sets them within a wider social and historical context. First of all, however, it is necessary to provide a brief historical account of how interest in this aspect of organization emerged.

1. Interest in Organizational Culture in Historical Perspective

In academic circles in the late 1970s, and in organizational practice a few years later, people began to realize that the instrumental-rational view of organizations that had been adhered to until then had not produced the success it had promised. This view rose to prominence in the nineteenth century and was developed further by Taylor (1912) in the theory of scientific management at the beginning of the twentieth century. This approach was characterized by increased efficiency, division of functions, and control of the production process in accordance with rational principles. Questions such as motivation and staff cooperation, which had emerged as relevant aspects of organization in the course of the well-known Hawthorne researches in the 1920s, were also tackled from this angle. After a while the idea that an organization is like a machine was dropped in favor of the metaphor of an organism, giving greater prominence to mutual functional harmonization and adaptation to the environment, but this did not lead to a fundamental rejection of the principles of scientific management.

By the end of the 1970s, however, a number of factors produced doubt. The Western economy was in recession, while that of Japan, where completely different views of management and organization held sway, was growing. Publications revealed that the distinctive features of Japanese companies were a shared culture and identification with the firm, in addition to quality, customer orientation, and considerable attention to the developmental and preparatory process. Moreover, the study by Peters and Watermann (1980) of the characteristics of several dozen successful North American companies seemed to confirm the suspicion that organizational culture was decisive for success or failure. According to Peters and Watermann, the ideal culture was one characterized by customer orientation, flexibility, and unity of purpose. This emphasis on financial success, excellence, and quality was welcomed by many young people who had enough of the pessimism, sobriety, and modesty of the 1970s. This group also favored the development of an identity of one's own organization, as it was in line with the trend towards distinction and

the development of personal lifestyles in a globalizing society in which behavior was less and less guided by tradition.

Within a short time the notion of organizational culture became a priority for managers and consultants. Very many organizations initiated trajectories for organizational change at this point, and in the course of time the list of desirable cultural aspects was extended to include such values as self-directing and communicational skills. However, many of these trajectories did not yield the success that they had been expected to provide. This is partly due to the fact that the instrumental-rational notion of management has not really been abandoned. This can be seen clearly once we take a closer look at how organizational culture has been approached by the mainstream of organizational scientists to date and which preferences are displayed by the practitioners (managers and their consultants).

2. *Organizational Culture: Stability and Unity, Diversity and Dynamics, or Both?*

Many organizational experts emphasize the values, norms, rituals, and myths shared by members of the organization. They often see the culture of an organization as a variable like structure or strategy. For others an organizational culture almost always implies internal differences and contradictions, but after a while these subcultures become reasonably stable. Yet others see organizational culture, especially in the individualizing societies of the West, as a dynamic entity in which converging, diverging, and fragmenting processes operate simultaneously. Most theorists in the latter group do not regard culture as a variable like strategy or structure; they take an organization to be a culture. In other words, in every organizational process, and thus in the establishment of a structure too, meanings are constructed, produced, interpreted, and distributed. These perspectives on organizational culture, as we shall see, have important consequences for what is regarded as the object of research, as well as for the response to the question of whether and how culture can be organized and managed.

2.1 *The Integration Approach*

A prominent and influential representative of this group is the social psychologist Schein, whose 1985 publication was a major contribution to the theoretical elaboration of the concept of organizational culture. The key element in his theory is that the norms, values, heroes, myths, and symbols of a group are based on how the group deals with the question of order ('internal integration') and the problem of adjustment to the surroundings ('external adaptation'). These basic assumptions arise during the first stage in the life cycle of a group by means of a group dynamic process in which each person has received a position and leadership has been defined. After a while they are internalized and new members are socialized into them. In other words, they are transferred from generation to generation, and there is practically no awareness of those basic assumptions (and their transfer). They refer to the relation to nature, they provide an answer to the question of what is true and legitimate, on what authority is based, what human nature entails, and how people view interpersonal relations and work (incidentally, these categories were taken from the crosscultural research conducted by the anthropologists Kluckhohn et al. in 1955). The integrating function of culture is expressed in an organization in the mission, the objectives, the means, and the opportunities for monitoring and correcting. The adaptive function is reflected in the shared language, the group boundaries, the views held on influence and status, the relations of intimacy and friendship, and the sanctions and rewards. According to Schein, groups can only have an organizational culture when they display a reasonable level of stability, homogeneity, and cohesion. This emphasis on unity, stability, and continuity can be found among virtually all the researchers who apply an integration approach to organizational culture. They consider that an organization must aim for a shared culture; difference weakens, and in this case they often distinguish between strong and weak cultures.

Research conducted from an integration perspective displays three other major characteristics. First, the research method is primarily quantitative, with a preference for questionnaires and scaled questions. Second, there is an implicit—and often explicit—assumption that culture is something that can be made, and that the person at the top has a large influence on the process of forming a culture. Organizational cultures can be changed by presenting different missions, introducing new styles of leadership, modifying systems of rewards, and altering the organizational structure. Third, most of those who adopt an integration approach display a tendency to prescribe, implicitly following a 'one best way' strategy. They often present ideal organizational cultures with characteristics such as customer orientation, the spirit of enterprise, flexibility, external orientation, empowerment, self-direction, etc. Hardly any of them ask themselves whether time and place ought to be taken into account. Most models on the relation between the success and the culture of an organization that claim to be universally applicable have been elaborated in the United States, although there are positive exceptions to this rule, such as Hofstede (1980) and Trompenaars (1993). On the basis of a wealth of empirical material, they show that there are significant national differences in attitudes to work, personal relations within the organization, time, truth, authority, and nature. They both consider that differences

in national culture are more significant and influential than differences in organizational culture. National cultures are older, cover more walks of life, and are therefore more deeply rooted. Their theory is that the larger the difference between national cultures, the more problematical cooperation between (members of) the organizations from two different countries is likely to be. In all other respects, however, Hofstede and Trompenaars share the same characteristics as other researchers on culture from an integration perspective. No attention is paid to the dynamics of culture, the theory pays little, if any, heed to other organizational aspects such as power, and the methodology is quantitative, based on scaled questions and dimensions. The hypothesis that cultural difference is decisive for the success or failure of international cooperation has been falsified in a number of researches. The (perceived) difference in power between two organizations, and the difference in power between the countries in which their head offices are located, prove to be significant intervening variables (see, e.g., van Marrewijk 1999).

Many researchers who adopt an integration perspective have elaborated typologies. A well-known one, for example, is that of Deal and Kennedy (1982): 'bet your company,' 'tough-guy macho culture,' 'process culture,' and 'work-hard-play-hard culture,' based on the dimensions of high and low risk attaching to decisions called for by the surroundings and the rapidity of feedback to the quality of those decisions. Another familiar typology is that of Handy (1985), based on the dimensions distribution of power and organizational structure: Zeus or club culture; Apollo or role culture; Athena or task culture; and Dionysus or person culture.

2.2 The Differentiation Approach

Already in the late 1970s there was a small group of researchers who concentrated on the formation of subcultures within organizations (a controversial researcher is van Maanen 1991). They take the existence of subcultures to be inherent in every organization. After all, members of organizations occupy different positions (of power), are located in different buildings, and differ in terms of gender, training, ethnicity, age, political preference, and ideology. This has often led to the emergence of (informal) subcultures, which sometimes assume the character of a counterculture in opposition to the dominant culture. Those who adopt a differentiation approach to organizational culture are thus interested in relations between subculture and power. However, they show hardly any interest in the immanent dynamics of culture, nor do many of them consider the relation between (diversity in) organizational culture and success to be an important question for empirical research.

The issue of how culture can best be directed or managed in organizations is not a main one for them. Their main research aim is to describe and explain the different systems of meaning in organizations in themselves. They make much more use of qualitative, ethnographic research methods (such as participant observation, group discussions, and interviews with lists of topics) for this purpose than do the researchers who adopt an integration perspective. There are, it is true, a few researchers who promote the advantages of cultural diversity in organizations (with the slogan 'management of diversity'), but they hardly back up their claims with detailed empirical research in the field. Managers and consultants have displayed a very minimal interest in the differentiation perspective so far. Some measure of a cautious change in their attitude can be detected in recent years, but this is mainly confined to the front stage where managers operate. Back stage, most managers still cling to the idea that difference weakens and that an ideal of unity is thus preferable.

2.3 The Fragmentation and Multiple Perspective Approaches

These two approaches are primarily products of the last few years and certainly have not come to constitute a mainstream yet. Adherents to the fragmentation approach claim that every element of culture has an ambiguous character, and that no culture or subculture is consistent and stable. This perspective has its roots in postmodernism, which calls into question the possibility of attaining an objective ideal of truth and places strong emphasis on the situational construction and production of (virtual) realities by means of signs. It takes an organization to be a nexus of configurations of meaning that partially overlap and intersect, thereby presenting individuals with many problems of choice and uncertainties (see, e.g., Martin 1992). The decisions that they take eventually lead to new cultural permutations. Research within this perspective targets the question of which meanings are constructed and produced, by whom, and when. The criticism that has been leveled against this approach is that its proponents see fragmentation, complexity, chaos, and turbulence everywhere, and fail to pay attention to the fact that organizations are always reorganized and brought into a new state of integration (see, e.g., Alvesson 1993).

Moreover, what at first sight appears to be unclear or ambiguous may turn out to have much more structure after it has been investigated in greater depth. It is thus advisable to keep an eye out for processes of integration, differentiation, and fragmentation in organizations, as well as the fact that a relation can be indicated between these three processes (Koot 1993). A culture offers not only safety, security,

and shelter, but also opportunities for the classification of who belongs to the in-group and who to the out-group. In other words, culture connects and integrates, but it also distinguishes and contrasts. When there are major clashes of interests, cultural differences may sometimes be used strategically—for example, by accentuating those differences—to keep the other at a distance and to create internal homogeneity and fighting spirit. This process takes place not only between but also within organizations. The chance of its taking place is very high if the organization is large, international, stratified, and functionally differentiated. Moreover, it appears that the rapid, forced attempt from above to bring subcultures together in order to satisfy the ideal picture of unity of culture that so many cherish actually produces differentiation. The current attempts by many managers and consultants to give their organization a contemporary culture of unity, with characteristics like self-direction, empowerment, and flexibility, are paradoxical and lead to a large degree of complexity. There is a desire for unity, on the one hand, while all (or some) are supposed to be flexible and self-directing, on the other hand. If a part of the organization is made flexible and is offered only a temporary contract, for example, this may cause contradictions between permanent and temporary staff. Besides, the temps keep introducing new cultural elements, thereby increasing the likelihood of cultural hybridization, fragmentation, and confusion. Finally, because organizations have been faced more and more with the new technologies of communication during the last few years, networks (of information) emerge between individuals within and outside the organization and virtual realities arise. The fixed cultural patterns therefore intersect more and more, and there is increasing overlapping between false and genuine.

In other words, processes of integration, differentiation, and fragmentation take place in every organization, and especially in a postmodern one. This complex interaction is the main object of research for a small group of researchers on organizational culture, a relatively large percentage of who have a background in ethnography. Neither the fragmentation nor the multiple perspective approach has met with much of a response from the practitioners in organizations so far. In most cases their reaction to the complexity that they experience is to try to control and standardize the processes of signification and attribution of meaning. It is probably here that the deeply rooted basic assumptions of the West like rationalization and individualization, both deriving from the ideals of the Enlightenment, clash with one another. Furthermore, they both entail consequences that are diametrically opposed to the point of departure: increased dependency through specialization versus the goal of liberation and control. It is likely that one of the major challenges to managers and their consultants in the twenty-first century will be to escape from this dilemma. Of course, it is also an exciting challenge for researchers: to describe, analyze, and explain the cultural paradoxicality and to draw up scenarios on this basis.

3. *The Role of Anthropologists in Research on Organizational Culture*

Anthropologists were still emphatically present in the Hawthorne research of the 1920s with their method of participant observation and their interest in how people live their everyday lives. Apart from a few exceptions, their interest soon declined dramatically. There are a number of reasons for this. The traditional field of study of academic anthropology has led to a certain bias in the object of research. The emphasis has come to focus on (the study of) marginal, oppressed, and exotic groups, along with a distaste for (the study of) the power center. Moreover, there has always been more stress on description and analysis and much less on directing and controlling phenomena. Nor should one forget that the dominant research method of anthropologists, participant observation, is relatively time-consuming and expensive, as well as threatening to find too many skeletons in the cupboard for some authorities.

However, towards the end of the twentieth century a change has taken place in the attitude of some academic anthropologists and the practitioners in the world of organizations. First of all, that world is beginning, in a very cautious way, to realize that the management of a modern organization is an exceptionally complex matter because of all the processes of globalization, localization, internationalization, immigration, and virtualization. There is an ever-growing demand for theoretical and empirical knowledge about such issues as what bonding, identification, trust, and survival mean in the fragmenting and hybrid organizations of today with their high pressure of work and fast tempo of life. This calls for research into how people go about their daily lives. Some anthropologists have come to realize that this is an exceptionally interesting field of study, and that they must drop their traditional reservations with regard to companies and the issue of management.

See also: Industrial and Organizational Psychology: Cross-cultural; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Myths and Symbols: Organizational; Organizational Climate; Organizational Culture; Organizations, Sociology of; Ritual

Bibliography

Alvesson M 1993 *Cultural Perspectives on Organizations*. Cambridge University Press, Cambridge, UK

Deal T, Kennedy A 1982 *Corporate Cultures*. Addison-Wesley, Reading, MA
Handy C 1985 *Understanding Organizations*. Penguin, Harmondsworth, UK
Hofstede G 1980 *Culture's Consequences*. Sage, London
Kluckhohn F, Strodtbeck F, Roberts J 1955 *A Study of Value Orientations*. Evanston, IL
Koot W 1993 Ambiguity and changing identities, on the dynamics of organizational culture. In: Shadid W, Nas P (eds.) *Culture, Development and Communications*. SNWS Publications, Leiden, The Netherlands
Maanen J van 1991 The smile factory. In: Frost P (ed.) *Reframing Organizational Culture*. Sage, London
Marrewijk A van 1999 *Ethnicity, Power and Cooperation in the Telecommunications Sector*. PhD thesis, VU University Press, Amsterdam
Martin J 1992 *Cultures in Organizations*. Oxford University Press, Oxford, UK
Peters T, Watermann R 1980 *In Search of Excellence*. Harper, New York
Schein E 1985 *Organizational Culture and Leadership*. Jossey Bass, San Francisco
Taylor F 1912 *Principles of Scientific Management*. Harper, New York
Trompenaars F 1993 *Riding the Waves of Culture*. Brealy Publishing, London

W. Koot

Organizational Decision Making

1. The Field

The field of organizational decision making (ODM) is not easy to define. Those who engaged in a conceptual definition have usually contrasted it with individual decision making (Shapira 1997, Butler 1997), as ODM entails multiple individuals—whereby issues of communication and conflict may arise—and multiple occasions over time and matters—thereby excluding one-shot decisions among unrelated players. On theoretical (and philosophical) grounds the issue is tricky, because individuals may not be such a unitary unit of analysis as often taken for granted: there can be conflict between objectives and 'multiple selves' also within an individual (March and Simon 1958, Elster 1985). In addition, an individual may be a subject of organizational decision making if he acts out of his organizational role rather than on a private basis (Barnard 1938). Finally, a multiperson system may be conceived as a unitary actor, if on a problem at hand its behavior can be modelized 'as if' it were a homogeneous node of knowledge and preference (Grandori 1995)—for example a firm in the context of defining an alliance with other firms, or a functional unit in the context of defining a new product with other functions. On empirical grounds, in fact, organizational decision making research has encompassed both single-actor and multiple-actors situations, provided that decisions are taken in an organized context: ongoing relations among actors (not instantaneous transactions), acting for some purpose of effectiveness (not just for leisure). Although traditionally research has been focused on processes going on within the institutional boundaries of complex actors (firms or public administration bodies), the growing relevance and interest in 'interorganizational networks' has stimulated research on organizational decision making research also across those boundaries.

The main acquisitions in the field which will be exposed here distinguish models concerned with the 'logic' or with the 'politics' of decision making, and specify how they have contributed in understanding how actors can solve and do solve problems of knowledge acquisition and conflict resolution: how organizational decision processes follow different patterns according to the state of knowledge and to the configuration of preferences, and cope with uncertainty and diversity of interests.

2. The Logic of ODM

The history of thought on organizational decision making did have a 'big bang,' with the 'theory of bounded rationality' (Simon 1955). The theory was developed in contrast with utility theory, the dominant model of rationality available at the time, developed especially in economics. The concept of bounded rationality included several components and submodels. First, it pointed out that actors (either individual or composite) on most problems cannot acquire the information required for utility maximizing calculations, either because it is too costly or because it is cognitively unfeasible to do so. Simon then distinguished 'structured' from 'unstructured' problems. In a structured problem, the actor knows what the relevant alternatives and the possible 'states of the world' are, is able to foresee the consequences of each combination of the two, knows what the value (or utility) of these consequences is for him. Even if a problem is structured, calculations of which action is optimal can be suboptimal or unfeasible because of computational complexity. As in the game of chess, the number of possibilities to be taken into account may be too high for the human (and computer) information processing capacity. In unstructured problems, the lack of knowledge—hence the state of uncertainty—is more radical. The actor does not know not only what the probabilities of payoffs are but even what the relevant alternatives, or the relevant consequences, or even the relevant objectives are. There can be uncertainty on cause–effect relations as well as on preferences (Thompson and Tuden 1959). In addition there can be uncertainty even on what has been observed, on what the 'facts' are. This state of

knowledge has been called, in later research, 'ambiguity' (Cohen et al. 1972) or 'epistemic uncertainty.' Second, Simon's research program invited 'to develop a variety of models of decision processes based on bounded rationality,' which could claim to be predictive of observed behaviors in varying conditions of information complexity. Simon outlined especially one of these possible models, the 'satisficing model' of search and choice. The simplest basic version of that model states that actors will accept the first encountered alternative superior to a given aspiration level, truncating search at that point. If it is difficult to find acceptable alternatives, aspiration levels fall, if it is easy they raise. In that version satisficing would describe primarily a decision behavior capable of generating 'good' solutions while reducing the costs of search. A more ambitious heuristic version of satisficing that would be able to solve some forms of epistemic complexity can be and has been defined: an actor formulates hypotheses about levels of result which are judged acceptable according to preference and reachable according to available knowledge; search for alternatives which are supposed to generate those results is started; either the objectives, the types of alternatives, the types of consequences considered, the cause–effect theories employed, or the observation and evaluation procedures employed, can be adjusted on the basis of experience (what proved to be possible to observe and find in reality) (Grandori 1984). Various other decision models, with claims of being able to manage even stronger conditions of uncertainty and conflict, have also been formulated in other strands of ODM research.

An 'incremental model' has been identified in research on ODM in public administration (Lindblom 1959). A starting provocative empirical finding has been that the best predictor of a year's budget for a program or activity in local government is the budget of the previous year (Wildavsky 1964). In an incremental process the decision system should be able at least to discern where it wants to go: increase rather than reduce production capacity, improve levels of education, increase well-being. However, it can generate action without making use of theories on the cause–effect relationships that regulate how the action system works and changes. Rather it employs a 'linear' and automatic choice rule—try 'incremental' solutions that differ marginally from those in use and learn afterwards.

A 'cybernetic model,' built in the conceptual frame of control theory, also relies heavily on *ex post* learning, and more precisely on learning by reinforcement (Steinbrunner 1974). A cybernetic decision strategy implies only the following kinds of judgement: the capacity to recognize situations (e.g. a configuration of costs); the capacity to recognize performance gaps with respect to a standard (works/does not work; positive/negative); and possessing a repertory of possible actions that are applicable to eliminate the gap or respond to the situation, adjourned by a principle of reinforcement (repeat successful action, avoid previously failing actions). Processes that can be described and predicted fairly well as cybernetic processes range from some real-time unreflexive stimulus–response motivation processes to some innovation processes, especially the diffusion of observable and codified technological and organizational innovations.

A 'garbage can model' constitutes a limiting case of decision making behavior, in which actions are guided by the chance encounters between actors, with an opening of attention at a certain time, and the emergence of choice occasions (Cohen et al. 1972). It is supposed to be relevant especially under ambiguity: research using that model has typically been on institutions or processes with 'ambiguous' goals or highly unpredictable consequences (e.g., universities and cultural institutions, career systems). Although it has been seen as a provocative criticism of the means-ends 'determinism' of rationalist models of ODM, the actual finding of the 'garbage can model' is that when ODM is not guided by clear knowledge and preferences, it is likely to be driven by a contextual 'determinism': choices are highly dependent on and constrained by the distribution of work loads and the time distribution of events (who is available for what choice occasion at what time).

Descriptive empirical research focused on various dimensions of ODM (see Butler 1997, Grandori 1995 for reviews)—e.g., 'discontinuity,' 'cleavage,' 'complexity'—has generally supported the proposition that the higher the level of uncertainty and the conflict among interests becomes, the less comprehensive and deductively ordered the logic of ODM becomes; the less investment in *ex ante* knowledge acquisition and the more reliance on *ex post* learning is made; the less preferences are integrated and limited agreement is sought on action.

Prescriptive comparative assessments of decision strategies have also been elaborated in management sciences. Empirically, in structural contingency research, it had early been found that organization structures tend to set decision premises and routines so as to stimulate more optimizing and computational decision processes in units dealing with structured tasks and more heuristic and empirical approached in unstructured situations. Theoretically, also building on those results, prescriptive models for the selection of decision processes as 'strategies' have been formulated. In earlier applications of a contingency approach to decision behaviors (see Grandori 1984, for a review) the decision models available in ODM research have been ordered on continuums: from 'closed system' (planned/maximizing) models to 'open system' (adaptive/incremental) models; or from 'analytic' processes (based on calculation supported by formal support and explicit knowledge) to 'un-analytic' (non-calculative action driven by rules, conventions, and

habits). Consistent with behavioral decision making research conducted at the individual level, the problem of strategy selection has been casted in those studies as a trade-off between effort and accuracy, finding that less calculative and comprehensive strategies become superior as task complexity and environmental uncertainty grows. Later, less deterministic 'contingent decision processing' models have been proposed, based on 'decision failures' analysis rather than on one-to-one correspondences between information states and decision strategies. In structured tasks any decision strategies are applicable, in principle, and strategy selection—as economists have often maintained—may be mostly a matter of information cost reduction and effort–accuracy tradeoffs. However, growing levels of uncertainty and of conflict among interests rule out progressively deductive and computational value maximizing strategies; theory-testing problem solving heuristic approaches; and finally even rule-driven noncalculative action approaches, leaving increasing terrain to the effects of chance and context—for reasons of lack of requisite knowledge and logical 'impossibility' rather than for information cost (which cannot be assessed anyway) considerations.

In the ODM research described so far, and also in recent authoritative overviews (March 1994), decision modes and strategies have been mostly seen as mutually exclusive alternatives. Recently, however, a more microanalytic view of decision processes as combinations and configurations of a variety of components and subrules has led to the possibility of mixed decision strategies; even more, to assert that some elements of each mode are necessary for any other (Lindenberg 1990): the possibility of conscious and critical decision making requires an embedding frame, accepted by convention, or learned by the experience of self or others; and building a model of a decision situation always requires some heuristic hypothesis testing, even in order to 'optimize' or in order to apply 'appropriate' solutions. Among recent trends in ODM, there has also been a tendency to deepen the implications of the 'cognitive revolution.' ODM research has sought to inquire more deeply in the processes of knowledge acquisition that frame decisions. If the main innovation of bounded rationality theory had probably been to analyze ODM with a theory of search, rather than just with a theory of choice, a subsequent and still actual challenge has been to connect the logic of ODM with the theory of knowledge. In organizational studies, this interest used to be cultivated especially at the border of the ODM field with organizational learning research and is now re-emerging at its border with the 'knowledge-based theory' of governance, with a main interest in the type of cognitive processes and organizational mechanisms which can sustain innovation. Research has also gone back to the founding connections between cognition and organization. In particular, there have been efforts at connecting research in cognitive psychology on heuristics and biases in decision making and ODM and organization systems and structures (Shapira 1997). These works have cast some doubt on the original intuition of bounded rationality theory in this respect—organization is a mean of reducing the limits of individual rationality—showing that organizational dynamics, systems, and structures can amplify rather than mitigate individual cognitive biases. A direction for future research, then, can be to discriminate between the conditions and the types of activities under which the two opposite effects are more likely.

3. *The 'Politics' of ODM*

A different, although often connected, track of development of ODM studies have been concerned more with the problem of diverging interests and the relative 'weight' or power of actors holding them than on the problem of knowledge. Often the starting point has been a criticism of Simon's 'administrative man' for 'neglection' of the political and strategic dimension (Pettigrew 1973). American ODM research on the politics of ODM has often contrasted organizational actors' power with organization system's efficiency, showing empirically that often the particularistic objectives and interests of subunits have more weight on the allocation of resources than the overall effectiveness and efficiency objectives held by the central agents of the system (Pfeffer and Salancik 1974). European studies have considered actors' power in a somehow more positive light, casting the ODM problem as 'the integration of the many,' as contrasted with the 'one actor theorem.' Research conducted in that perspective has been especially concerned with the strategic use of uncertainty and the determinants of the decision weight of organizational units and groups (Crozier and Friedberg 1977, Hickson et al. 1971). It has been argued and found that the relative influence of organizational actors in ODM is a function of 'the control of critical uncertainties,' of the degree of substitutability, of the centrality in work and information flows, and of the perception and use of these 'actors' as bases of negotiation power.

As to the problem of how to govern pluralistic ODM system, negotiation has received most attention. Decisions have been seen predominantly as negotiated actions, taken within negotiated 'rules of the game.' Other coordination mechanisms are, however, clear contenders—and/or complements—to negotiation in the governance of pluralistic and have also been studied in strands of organization studies connected or contiguous with ODM research systems (for a review of the relevant literature and authors, see Grandori 1995). 'Organizational democracies', based on the core mechanism of voting, constitute an alternative—if problems are structured enough; constitutional orders, regulated by the core mechanism of rules and norms, are another, at least in stable systems of recurring

actions; 'loosely coupled systems', in which decisions are taken in a semiautonomous way and coordination is ensured by varying mixes of buffers, codified communication, or routinization of know-how may be thought of as another governance arrangement, suitable for complex but stable action. 'Organized anarchies'—'unsegmented structures' in which everyone is free to address any problem and talk with any other actor, with little horizontal or vertical specialization constraint—would represent an extreme 'governance' form, in which there is no predefined pattern or mechanism of coordination, viable under conditions of extreme ambiguity and variability. These coordination systems have been seen, again, mostly as structural alternatives, each connected to some distinctive conditions of uncertainty and conflict of interests. As for decision strategies, however, coordination mechanisms could be assessed according to a 'failure framework' allowing for significant fungibility and equifinality among mechanisms, at least under simpler conditions. In fact, in a design perspective on distributed ODM, there has been an increasing awareness of the imperfections and paradoxes of any mode of governance of large systems composed by actors with differentiated knowledge and preferences. Although the variety of forms of 'poliarchy' mentioned above provide alternatives to the demonstrated failures of both market and hierarchical mechanisms in conditions of information complexity (Miller 1992), the limits of distributed decision making systems have also been examined. These systems are more costly—as to communication and integration costs, control losses and costs of delay—and their behavior is weakly predictable or more 'indeterminate' as far as there are usually many equilibria in such organizational games. Interestingly, in these contributions on ODM, mainly rooted in economics, a frequently recommended solution to the dilemmas of poliarchic governance is the institutionalization of framing norms and rules, which reduce the 'degrees of freedom' in actors' preference variance and enhance the predictability of behaviors; i.e., the embedding or blending of rational choice (more or less behavioral) with the suspension of critical judgement and *ad hoc* calculation in certain zones of action.

Among other recent trends in the ODM field—as to the modes for 'integrating the many'—there are signs of a growing concern for the equity dimension. The use of the concept of power has declined in studies of ODM, and the use of concepts of equity and organizational justice have gained terrain as the required complements to the notions of organizational effectiveness and efficiency in decision making. This may be seen as a healthy turn, as far as equity criteria are more precise and by definition fairer than power criteria. In addition, fairness criteria are logically necessary if the problem of governing pluralistic decision systems is tackled as in a prescriptive stance—how to devise superior arrangements for multiple actors—rather than just in a descriptive stance—how do multiple actors take decisions in organized settings.

See also: Bureaucracy and Bureaucratization; Competitive Strategies: Organizational; Conflict: Organizational; Corporate Culture; Design and Form: Organizational; Group Decision Making, Social Psychology of; Group Processes in Organizations; Information and Knowledge: Organizational; Innovation: Organizational; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Learning: Organizational; Organization: Overview; Organizational Behavior, Psychology of; Organizational Change; Organizational Decision Making; Organizations, Metaphors and Paradigms in; Organizations, Sociology of; Organizations: Unintended Consequences; Rational Choice and Organization Theory; Strategic Intelligence; Strategy: Organizational; Technology and Organization

Bibliography

Barnard C I 1938 *The Functions of the Executive*. Harvard University Press, Cambridge, MA

Butler R 1997 Decision making. In: Sarge A, Warner M (eds.) *IEBM, Handbook of Organizational Behavior*. International Business Press, London

Cohen M D, March J J, Olsen J P 1972 A garbage can model of organizational choice. *Administrative Science Quarterly* **17**: 1–25

Crozier M, Friedberg E 1977 *L'Acteur et le système*. Editions du Seuil, Paris

Elster J (ed.) 1985 *The Multiple Self*. Cambridge University Press, Cambridge, UK

Grandori A 1984 A prescriptive contingency view of organizational decision making. *Administrative Science Quarterly* **29**: 192–208

Grandori A 1995 Models of pluralistic organization: the contribution of European decision making research. In: Bacharach S, Gagliardi P, Mundell B (eds.) *Research in the Sociology of Organizations 13, ' The European Tradition.'* Jay Press, San Francisco

Hickson D, Hinings C R, Lee C A, Schnek R E, Pennings J M 1971 A strategic contingencies theory of intraorganizational power. *Administrative Science Quarterly* **16**: 216–29

Lindblom C E 1959 The science of muddling through. *Public Administration Review* **19**: 78–88

Lindenberg S 1990 Homo socio-economicus: The emergence of a general model of a man in the social sciences. *Journal of Inst. & Theor. Econ* **146**: 727–48

March J 1994 *A Primer on Decision Making*. Free Press, New York

March J G, Simon H A 1958 *Organizations*. Wiley, New York

Miller G J 1992 *Managerial Dilemmas: The Political Economy of Hierarchy*. Cambridge University Press, Cambridge, UK

Pettigrew A M 1973 *The Politics of Organizational Decision Making*. Harper and Row, New York

Pfeffer J, Salancik G R 1974 Organizational decision making as a political process: The case of a university. *Administrative Science Quarterly* **19**: 135–51

Shapira Z 1997 *Organizational Decision Making*. Cambridge University Press, Cambridge, UK
Simon H A 1955 A behavioral model of rational choice. *Quarterly Journal of Economics* **69**: 99–118
Steinbrunner J D 1974 *The Cybernetic Theory of Decision*. Princeton University Press, Princeton, NJ
Thompson J D, Tuden A 1959 Strategies, structures, and processes of organizational decision. In: Thompson J D, Hammond P B, Hawkes R W, Junker B H, Tuden A (eds.) *Comparative Studies in Administration*. University of Pittsburgh Press, Pittsburgh, PA, pp. 195–216
Wildavsky A 1964 *The Politics of the Budgetary Process*. Little, Brown, Boston

A. Grandori

Organizational Errors: Catastrophic

While it has always been possible for organizations to make errors, only in recent times has it been possible for them to make errors that can have absolutely staggering outcomes for organizations and their environments. Such errors are the normal consequences of technological development and increasing organizational interdependence, both locally and globally. In the latter part of the twentieth century science and technology introduced and identified a new type of human-made error: low probability events with enormous consequences. Examples of such errors are nuclear reactor accidents at Three Mile Island and Chernobyl, the industrial tragedy at Bhopal, the *Exxon Valdez* accident, and the collapse of Baring's Bank.

Only recently have social scientists focused on organizational predilections toward making error or insuring reliability. Some researchers look for causes of major catastrophes, while others focus on situations that could go badly but which haven't. Both are looking for the kernels of organizational processes that produce disaster and contribute to highly reliable performance. The purpose of this article is to summarize their main contributions.

1. *Three Approaches to the Study of Organizational Catastrophe*

There are three approaches to identifying processes that can result in organizational catastrophe. The first is human factors research born from the man machine interface problems identified during World War II. This approach combines the methodologies of engineering and experimental psychology and is now in its third research generation. The first generation is referred to as 'knobs and dials' human factors and dealt initially with man–machine interfaces in aviation (though it spread quickly to other industries). The second generation focused on the cognitive nature of work. The shift in emphasis was driven by technological innovation in the late 1960s and 1970s, in particular the development of computers. Most of the work of the first two generations focused on the design of specific jobs, work groups, and related human–machine interfaces to obtain reliable and safe performance. Neither generation was even remotely interested in organizational maladies that might affect performance. The emerging third generation, macroergonomics, concentrates on the overall organization machine interface.

The second approach is the multidisciplinary science of risk analysis, born in the 1970s. The foundation of this science lies in the musings of the seventeenth century French monk, Blaise Pascal. He proposed a rigorous approach to thinking about the unfolding of future events called probability theory. Probability theory enables analysts to quantify the odds of two different events occurring and then compare them. Probabilistic risk assessment (PRA) models abound and are very good at predicting events that occur with some regularity in large samples or in populations. While their practitioners often use these models to predict infrequent events, they are not very good at it. Further, the scientists with engineering and statistical backgrounds who use them typically are not interested in organizational processes, and thus, fail to include organizational variables.

The third approach is that of behavioral researchers (psychologists, sociologists, and political scientists), who examine organizations, using various methods including observations, surveys, and archival data collection. The first people in this arena were accident researchers and analysts who most frequently took a 'name and blame' the operator perspective (because then they could focus on dismissal and/or training). However, the next set of researchers interested in why organizations fail took a larger perspective and attempted to identify organizational processes that contribute to catastrophic error. It soon became clear that errors are usually caused by a combination of organizational failures.

2. *Psychological Approaches to Understanding Catastrophe*

Among the first of these was Turner (1978). His conceptualizations were based on his analysis of a large number of disasters. Before Turner's work, disaster investigations focused on the impacts of disaster. The only people interested in preconditions were engineers who were solely interested in technical determinants of technical failures. Dispelling the theory that disasters are 'bolts from the blue,' Turner

introduced the notion of incubation periods leading to disaster and identified some components of incubation. These included organizational culture, the intertwining of formal and informal communication processes, hierarchy, power, and the fact that decision-making suffered from bounded rationality. Turner used catastrophe theory to explain disasters as abrupt transitions, particularly focusing on discontinuities in information. This kind of discontinuous process has been noted by a number of other researchers.

At a more microlevel, Helmreich and co-workers' early research focused on commercial aviation safety and flight deck team performance. Recently, they examined teams in operating rooms. They identify individual and team factors that influence performance, and they situate their teams within the design and equipment environment of the operating room (Helmreich and Schaefer 1994). Given their social psychological perspective, it is not surprising that these researchers find that interpersonal and communication issues are responsible for many inefficiencies, errors, and frustrations.

3. *Sociological Approaches to Understanding Catastrophe*

After the Three Mile Island tragedy in 1979 the United States President's Commission on the Accident decided to invite social science input into what threatened to be an entirely engineering oriented investigation. That invitation lead to the publication of *Normal Accidents: Living with High Risk Technologies* by Perrow (1984). Perrow added to the Three Mile Island story archival analyses of petrochemical plants, aircraft and airways, marine systems, and other systems. He introduced the term 'normal accident' to signal that, in systems like Three Mile Island's, multiple and unexpected interactions of failures are inevitable. Perrow noted that as systems grow in size and in the number of diverse functions they serve, they experience more incomprehensible or unexpected interactions among components and processes. Among these are unfamiliar sequences of activity, or unplanned and unexpected sequences that are either not visible or not immediately comprehensible. The problem is accentuated when systems are coupled tightly, that is, they incorporate numerous time-dependent processes, invariant sequences (A must follow B), and little slack. When complexity and tight coupling combine, the likelihood of a normal accident increases.

James Short's 1984 presidential address to the American Sociological Association was one of the first highly visible treatments of risk in sociology. Short and Clarke (1992) collected a set of authors who made and continue to make seminal arguments about organizations and the dangers they create. Over the years Clarke has stressed repeatedly the ambiguous nature of risk and danger, as has Karl Weick. Clarke's recent book (1999) focuses attention on the symbolic representations organizations use to justify high-risk systems.

4. *High Reliability Organizations*

Also in 1984 a group of scholars at the University of California at Berkeley formed a loosely-knit consortium with a few scholars from other universities to examine organizations that appeared susceptible to adverse error but that seemed to operate without major incidents. They called them 'High Reliability Organizations' (HROs). This group does not contend, however, that these organizations cannot get into serious trouble. They began by using a behavioral observational approach, and by examining the United States Federal Aviation Administration's air traffic control system, a well-run nuclear power plant, and the United States Navy's carrier aviation program. The purpose of the research was to identify organizational processes that result in high reliability operations.

The group's early findings (see Roberts 1993) underscored the point that operator errors can be reduced and organization contributions can be kept separate if good design and management techniques are followed. An example is the simultaneous launch and recovery of aircraft aboard an aircraft carrier. Good design and management results in these processes being separated physically from one another. This requires that '(a) political elites and organizational leaders place high priority on safety and reliability; (b) significant levels of redundancy exist, permitting back up or overlapping units to compensate for failures; (c) error rates are reduced through decentralization of authority, strong organizational culture, and continuous operations and training; and (d) organizational learning takes place through a trial-and-error process, supplemented by anticipation and simulation' (Sagan 1993).

While members of the group no longer work together many continue to develop theories about, and do further empirical work in, high reliability organizations and in organizations that have, sadly, broken apart. They have also examined many more successful and unsuccessful organization, including the commercial maritime system, medical settings, the commercial airline industry, non-US nuclear power plants, community emergency services, and banking. One stream of the work examines our gradual loss of control of information systems. Rochlin (1997) calls this the computer trap and identifies its components. Another stream (Roberts et al. 1994) looks more specifically at how decisions must migrate around organizations to the people with expertise rather than formal or position authority, and the structural processes that unfold as the organization enters

different evolutions. This work finds, contrary to Perrow, that both tightly and loosely coupled organizations are subject to massive failure if they are tightly or loosely coupled at the wrong time.

A further stream of research (Grabowski and Roberts 1999) examines systems of interdependent organizations and is interested in whether the current mania for virtual and other temporary organizations can appropriately manage in situations demanding highly reliable operations. Larger interdependent systems do not necessarily produce new problems, but they offer more opportunities for old problems to crop up simultaneously. A final stream of the high reliability research attempts to integrate its theoretical development with the main stream of organizational theory (Weick et al. 1999).

Sagan (1993) argued that what he calls 'normal accidents theory' and 'high reliability organizations theory' are separate theories, and in the case of nuclear armament finds high reliability theory wanting. That comparison seems a bit odd since 'normal accidents theory' talks about the inevitability of accidents and 'high reliability theory' talks about creating conditions to reduce the probability of accidents. The two conceptualizations should be complementary and Sagan's test ground for reaching his conclusion is an arena in which there have been no serous accidents. More recently, Sagan (1995) pits the political science-derived theory about why more nuclear arms are better against organizational theory which says if an organization can make errors it probably will. After all, nuclear weapons are controlled (and sometimes not controlled) by organizational processes, and to the extent these break down there is no assurance we can manage nuclear weapons safely.

5. *Other Research on Organizational Catastrophes*

At the same time the high reliability project was getting under way, and in 1984, Shrivastava (1987) examined the world's most severe industrial accident, to that date, Union Carbide's chemical plant accident at Bhopal. He found that the Bhopal accident was caused by a complex set of factors: human (e.g., morale and labor relations), organizational (e.g., the plant's low importance to the parent company, mistakes the parent company made in establishing the plant, and discontinuous plant top management), and technological (design, equipment, supplies, and procedures). Together he called them 'HOT' factors.

In her 1996 analysis of the *Challenger* launch decision, Vaughan (1996) uncovered an incremental descent into poor judgment supported by a culture of high-risk technology. The culture included, among other things, the fact that flying with known flaws was normal, as was deviation from industry standards. Deviations were normalized so they became acceptable and nondeviant. This occurred within a structural secrecy. Vaughan defines structural secrecy as 'the way patterns of information, organizational structure, processes and transactions, and the structure of regulatory relations systematically undermine the attempt to know and interpret situations in all organizations' (1996).

One group of behavioral researchers focuses on crisis prevention or crisis management. While they make ameliorative prescriptions they have also uncovered additional processes that contribute to crises or their prevention. In Europe this group is lead by Reason and Rasmussen. Reason (1997) begins with a model (the 'Swiss cheese' model) of defense. Organizations build layers of defense. In an ideal world the layers are all intact, allowing no penetrations by accident trajectories. In the real world each layer has weaknesses and gaps, which can be conceived of as holes in each of a number of horizontal planes of defense. These holes are in constant flux, rather than being fixed and static. Occasionally they line up with one another and the bullet that penetrates them causes an accident. The holes are created by both active and latent failure. Defenses in depth have made modern organizations relatively immune to isolated failures but such defenses are also the single feature most responsible for organizational accidents.

Reason points out that errors and rule violations committed by those at what he calls 'the sharp end' of the organization (i.e., the operational end) are common contributors to organizational accidents but they are neither necessary nor sufficient accident causes. He notes that proceduralizing jobs is one way organizations attempt to reduce risk. The degree to which jobs can be proceduralized depends on where they are in organizations. Higher level jobs are more difficult to proceduralize than lower level jobs. There is usually a trade-off between procedures and training, the more training an individual has the less likely their job is to be proceduralized.

Rasmussen et al. (1994) build on Reason's thinking by focusing on three levels of human performance he labels skill-based (SB), rule-based (RB), and knowledge based (KB). Errors can occur at any of these levels. Reason then summarizes the varieties of rule-based behaviors and how they can lead to organizational accidents. He begins by asking the question, was the task or situation covered by procedures or training? If not improvisation occurs which is knowledge based. In this situation there is a better than 50 percent chance of mistakes. If the procedures and/or training are appropriate and they were followed as intended there is a high probability of successful performance. If the procedures are inappropriate to the task at hand latent failures are likely. If they were not followed as intended mistakes are likely to occur.

Taking a different perspective, Lagadec (1993), also European, focuses on preventing chaos in crisis. He states that a crisis situation is like a kaleidoscope, if

you touch its smallest element, the entire structure changes. Crises involve large-scale, destabilizing breakdowns that degrade exponentially. Such events begin to find echoes in their contexts. They feed on all the problems and instabilities present in their contexts. Crisis situations have the following components: initial destabilization and defeat, late or poorly processed warning, massive and shattering challenge, completely unfamiliar events, insufficient or faulty warning systems, individual inertia, sluggishness of all administrative operations, saturated operators, weak warning signals, mental stumbling blocks, and desperate struggling not to admit something could be wrong.

The organizational impediments that hamper emergency action include the absence of prior reflection (thinking about what could happen), stupefying organizational gaps (no one feels qualified to take charge), no emergency plans or imaginary capacity, a wealth of defiance, and typical mistakes and behaviors that include imprudent, uncooperative, and hostile behavior. Lagadec focuses on how to manage crises and how to develop a learning process for individuals, teams, and organizations involved to prevent crises.

In the United States, Pauchant and Mitroff (1992) have investigated crises and crisis preparedness for many years. They present a four layered onion model of crisis management. The outer layer comprises an organization's strategies for crisis management. It usually does not have any. Level three evaluates how well the everyday operating structure of the organization contributes to or inhibits crisis and examines how symbolic functions, such as the organization's formal structure, reveals the perceptions of its members. Level two addresses an organization's culture, and level one the subjective experiences of those who form the organization.

6. *Summary*

In sum, there are a variety of perspectives about the antecedents to organizational crisis. Many different roads are taken and there is little move toward consensus. Of only peripheral interest to organizational researchers are the statistical and human-factors approaches. Turner's (1978) groundbreaking work in the United Kingdom provides the springboard for organizational researchers. This was followed by a number of disparate approaches by primarily US based researchers, complimentary with work done in Europe. Finally, there is a group of researchers more interested in crisis management. All of these writers have taken the organization or groups in it as their units of analyses, though many have alluded to larger interdependent systems. Only recently have researchers attended the larger system in which organizational errors occur because of things that happen within systems of organizations. An example of this is the nuclear reactor accident at Chernobyl in which reactor problems were embedded in a culture of secrecy, and a system that selected top level managers based more on their political affiliations than their expertise.

See also: Accidents, Normal; Competitive Strategies: Organizational; Conflict: Organizational; Corporate Culture; Environmental Challenges in Organizations; Information and Knowledge: Organizational; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Learning: Organizational; Organizational Behavior, Psychology of; Organizational Change; Organizational Decision Making; Organizations, Sociology of; Organizations: Unintended Consequences; Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Risk, Sociological Study of; Risk, Sociology and Politics of; Strategic Intelligence; Strategy: Organizational; Technology and Organization

Bibliography

Clarke L 1999 *Mission Improbable: Using Fantasy Documents to Tame Disaster*. University of Chicago Press, Chicago

Grabowski M, Roberts K H 1999 Risk mitigation in virtual organizations. *Organization Science* **10**: 704–21

Helmreich R L, Schaefer H G 1994 Team performance in the operating room. In: Bogner M S (ed.) *Human Error in Medicine*. Erlbaum, Hillsdale, NJ, pp. 225–54

Lagadec P 1993 *Preventing Chaos in a Crisis: Strategies for Prevention, Control and Damage Limitation*. McGraw-Hill, Berkshire, UK

Pauchant T C, Mitroff I I 1992 *Transforming the Crisis-Prone Organization: Preventing Individual, Organizational, and Environmental Tragedies*. Jossey-Bass, San Francisco

Perrow C 1984 *Normal Accidents: Living with High Risk Technologies*. Basic Books, New York

Rassmussen J, Pejtersen A M, Goodstein L P 1994 *Cognitive Systems Engineering*. Wiley, New York

Reason J 1997 *Managing the Risks of Organizational Accidents*. Ashgate, Aldershot, UK

Roberts K H (ed.) 1993 *New Challenges to Understanding Organizations*. Macmillan, New York

Roberts K H, Stout S K, Halpern J J 1994 Decision dynamics in high-reliability military organizations. *Management Science* **40**: 614–24

Rochlin G I 1997 *Trapped in the Net: The Unanticipated Consequences of Computerization*. Princeton University Press, Princeton, NJ

Sagan S D 1993 *The Limits of Safety: Organizations, Accidents and Nuclear Weapons*. Princeton University Press, Princeton, NJ

Sagan S D 1995 *The Spread of Nuclear Weapons: A Debate*. Norton, New York

Short J F, Clarke L 1992 *Organizations, Uncertainties, and Risk*. Westview Press, Boulder, CO

Shrivastava P 1987 *Bhopal: Anatomy of a Crisis*. Ballinger, Cambridge, MA

Turner B A 1978 *Man-made Disasters*. Wykeham, London

Vaughan D 1996 *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press, Chicago

Weick K E, Sutcliffe K M, Obstfeld D 1999 Organizing for high reliability: Processes of collective mindfulness. In: Staw B M, Sutton R (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, pp. 81–123

K. H. Roberts

Organizational Size

The ultimate definition of organizational size is the sum total of social resources the organization has acquired from its environment. These resources include knowledge, power, wealth, people, and legitimation. Early in the history of the science of organizations, size usually meant the number of people employed. This no longer suffices, as will be shown.

Size has been one of the most commonly used variables in organization studies. Intuitively, size is easy to grasp and it is easy to consider its influence upon organizational behavior. Debates about the role of size have contributed to the development of organization studies as a scientific discipline.

It was Weber who stimulated empirical studies of organizations, especially after 1946, when an English translation of some of his work appeared as *From Max Weber: Essays in Sociology* (Weber 1949). Early studies were case studies and focused on a single organization or a part of an organization at a time (e.g., Selznick 1949, Gouldner 1954, Blau 1955); hence, size was outside of their concerns. But size did become a major focus after comparative studies came into being, and once a common framework and a set of instruments for measuring organizational variables had been created. Studies of multiple organizations, or comparative studies, found much variation in organization structure, begging the question of why, and what, is responsible for the variation.

Three camps emerged. One emphasized size purely on empirical grounds, the second suggested technology partly for ideological preference, and the third, culture, speculation, and romanticism.

Size as a possible determinant of structure is easy to understand. Structure is plausibly a function of how many people work in an organization. Division of labor must presume the presence of a certain minimum number of persons, or there could be no division of labor. Hierarchy of authority or centralization, too, must assume there are enough people to have a hierarchy. The same is true of such other structural variables as formalization and stratification.

The emergence of the technology camp is best understood within the political context of a resurgence of Marx in the 1960s in academia in North America. One of the important notions of Marx was worker alienation enhanced by increasing technological sophistication, and bureaucratization in modern society. Blauner (1964) tested this idea with North American data and argued that worker alienation varied by the type of technology used by the industrial organization. The implication was that technology affected worker alienation through the medium of organization structure; hence, technology is an important determinant of structure. Another study, quite independent of Blauner, demonstrated that the type of technology used by the organization called for its own appropriate structure, and when it deviates from the appropriate structure, the organization proves to be unsuccessful (Woodward 1965). Fanning the popularity of the technology camp, Perrow (1967) gave it some theoretical rationale, suggesting that the variability of the raw materials to be processed by the organization dictated the type of technology to be utilized, hence the organization structure.

The culture camp had its semistart in the work of Abegglen (1958) which suggested that in non-Western societies there could be other ways of organizing business concerns than was thought appropriate in the West. Crozier (1964) gave this camp a boost when he attempted his cultural analysis of the French organizations he had studied and analyzed. He said that those who were in the organizations he studied were French, hence carriers of French culture, with its own 'permanent French culture traits.' The culture which organization members bring into the organization would shape its structure and would help in understanding its behavior. Meyer and Rowan (1977), and more directly Hofstede (1980), called for attention to culture in organization studies. The emphasis of the culture camp, however, was on values of people and did not include either organizational size or technology. Some work arising out of this camp appears to deny the possibility of the science of organizations (Osigweh 1989).

The most commonly used indicator of organizational size has been the number of employees of the organization. Early work predating the size vs. technology controversy but using size as an independent variable attempted to predict the administrative ratio, or the percentage of personnel doing administrative work. A simple model depicting a positive effect of personnel size on administrative ratio was confirmed with school district data from California (Terrien and Mills 1956). Anderson and Warkov (1961) added complexity as another independent variable when the above model was not supported by their data from Veterans Administration hospitals in the USA. Later studies confirmed a positive effect of complexity and a negative effect of size on administrative ratio. The negative effect of size on percentage administrative personnel has been shown in studies of the optimal supervisory ratio when the tasks of persons to be supervised are similar, and the number of persons

the supervisor can supervise effectively comes down when their tasks are varied and require expert skills and knowledge (Woodward 1965).

Studies on the relationship of size with structural variables have shown interesting results. While Weber was thought to have expected positive relationships among hierarchy of authority, system of rules and procedures, and division of labor (Hall 1963), and Hage (1980) depicts, in his axiomatic theory, centralization and formalization as being associated positively with each other, empirical studies have shown inconsistent results in the relationship among structural variables and size. Hage uses complexity instead of Hall's division of labor or Aston's specialization, and his complexity is qualitatively different from division of labor and specialization. Complexity includes elements of knowledge and qualifications, and is expected to be associated negatively with centralization and formalization.

Aston's studies found a positive effect of size to specialization and formalization, and a negative one to centralization. Blau and Schoenherr (1971) reported similar results. A replication of the Aston study with 50 factories in Japan found similar results (Azumi et al. 1979). Consistent with these results, centralization has been found to be associated negatively with formalization. This led Blau and Schoenherr to speculate that centralization and formalization are alternative ways of control in which more formalization permits delegation of authority to lower levels. A more recent study of a nationally representative sample of organizations (Kalleberg et al. 1996) reports negative relations of size and bureaucratization, and size and centralization, and positive relations between size, differentiation, and formalization.

Those who are most commonly identified with the size camp were the Aston group (Pugh et al. 1969, Hickson et al. 1969) and Blau (1970) and his associates, who used the number of fulltime employees as an indicator of size. For the Aston group, size is one of their 'contextual' variables, or social resources, that are drawn in from the environment to the organization. Some have used the number of participants in the organization as an indicator of size, raising the issue of organizational boundaries, where some would include students as part of organization size, while others would exclude them as clients.

Kimberly (1976) suggested that there are four components to size: physical capacity (e.g., plant capacity, number of beds in a hospital, etc.); number of personnel; organizational inputs or outputs; and discretionary resources available (e.g., assets, wealth, university endowment). Before the appearance of the Kimberly article, most empirical studies of organizations had used personnel size as indicator of size.

The Aston group were unable to replicate the Woodward study: They found that the influence of technology was limited to workflow. One of the more consistent findings after Aston has been the positive relationship between size and what they called 'structuring of activities,' especially specialization. Blau (1970) created a 'formal theory' linking structural differentiation, administrative economy, and overheads. Increasing size is the cause of differentiation in this scheme, and differentiation has two faces: On the one hand it creates subunits in each of which people tend to do similar tasks, hence requiring relatively few supervisors; and on the other hand, the subunits created tend to engage in different tasks, which requires more supervisors or coordinators. Thus, differentiation promotes intra-unit similarity and inter-unit diversity. The former reduces administrative cost, while the latter increases it and feeds back negatively on the differentiation process.

The issue of size vs. technology has declined in importance as some researchers have moved to combine elements of both to come up with new variables such as task scope (Hage 1980) and scale (Azumi et al. 1983).

Some of the major criticism of the size camp came from Argyris (1972), who questioned whether increasing size actually causes differentiation, and Aldrich (1972a and 1972b), who suggested that size is a dependent variable of structure. One lesson that has been learned from the size vs. technology controversy is that any attempt to find one determinant of structure is likely to fail.

So what has become of organizational size as a variable for organization science? The issues surrounding structure, its variety, what may determine it, and its effects on organizational processes and performances have continued to evolve. Organizational size continues to be a commonly used variable, but it no longer has the passionate focus it once had.

See also: Design and form, organizational; Organization, general

Bibliography

Abegglen J 1958 *The Japanese Factory*. Free Press, Glencoe, IL

Aldrich H E 1972a Technology and organizational structure: A reexamination of the findings of the Aston Group. *Administrative Science Quarterly* **17**: 26–43

Aldrich H E 1972b Reply to Hilton. *Administrative Science Quarterly* **17**: 55–7

Anderson T R, Warkov S 1961 Organizational size and functional complexity: A study of administration in hospitals. *American Sociological Review* **26**: 23–8

Argyris C 1972 *The Applicability of Organizational Sociology*. Cambridge University Press, London

Azumi K, Hickson D, Horvath D, McMillan C 1979 Nippon no Soshiki Kouzou. *Soshiki Kagaku* **12** (Winter): 2–12

Azumi K, Hull F, Hage J 1983 Managing high technology in an international perspective: A comparison between Japan and the United States. In: Mannari H, Befu H (eds.) *The Challenge of Japan's Internationalization*. Kodansha International, Tokyo, pp. 111–47

Blau P M 1955 *The Dynamics of Bureaucracy*. University of Chicago Press, Chicago
Blau P M 1970 A formal theory of differentiation in organizations. *American Sociological Review* **35**(April): 201–18
Blau P M, Schoenberg R A 1970 *The Structure of Organizations*. Basic Books, New York
Blauner R 1964 *Alienation and Freedom: The Factory Worker and his Industry*. University of Chicago Press, Chicago, IL
Crozier M 1964 *The Bureaucratic Phenomenon*. University of Chicago Press, Chicago
Gouldner R W 1954 *Patterns of Industrial Bureaucracy*. Free Press, Glencoe, IL
Hage J 1980 *Theories of Organizations*. Wiley, New York
Hall R H 1963 The concept of bureaucracy: An empirical assessment. *American Journal of Sociology* **69**: 32–40
Hickson D J, Pugh D S, Pheysey D C 1969 Operational technology and organizational structure: An empirical reappraisal. *Administrative Science Quarterly* **14**: 378–97
Hofstede G 1980 *Culture's Consequences: International Differences in Work-related Values*. Sage, Beverly Hills, CA
Kalleberg L, Knoke D, Marsden P V, Spaeth J L 1996 *Organizations in America: Analyzing their Structures and Human Resource Practices*. Sage, Thousand Oaks, CA
Kimberly J R 1976 Organizational size and the structuralist perspective: A review, critique, and proposal. *Administrative Science Quarterly* **21**: 577–97
Meyer J W, Rowan B 1977 Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology* **83**: 340–63
Osigweh C A B 1989 *Organizational Science Abroad: Constraints and Perspectives*. Plenum Press, New York
Pevrow C 1987 A framework for the comparative analysis of organizations. *American Sociological Review* **32**: 194–208
Pugh D S, Hickson D J, Hinnings C R, Turner C 1969 The context of organizational structures. *Administrative Science Quarterly* **14**: 91–114
Selznick P 1949 *TVA and the Grass Roots*. University of California Press, Berkeley, CA
Terrien F W, Mills D L 1955 The effect of changing size upon the internal structure of organizations. *American Sociological Review* **20**: 11–13
Weber M 1946 Bureaucracy. In: Gertz H H, Wright Mills C (eds.) *From Max Weber: Essays in Sociology*. Oxford University Press, New York, pp. 196–239
Woodward J 1965 *Industrial Organization: Theory and Practice*. Oxford University Press, London

K. Azumi

Organizations and the Law

Scholarship on organizations and the law addresses the relationship between formal organizations—such as private businesses, public agencies, and nonprofit associations—and the legal institutions that govern them. Since both organizations and law receive separate treatment elsewhere in this encyclopedia, this article concentrates solely on the connections between the two. These connections, however, are elaborate and diverse, and different research traditions have explored different territories along three cardinal dimensions. First, separate bodies of scholarship address the relationship between organizations and at least three distinct aspects of law: (a) legal venues as facilitative arenas of organizational interaction (especially disputing), (b) legal precepts as regulative codes of organizational conduct, and (c) legal definitions as constitutive typologies of organizational experience. Second, these various aspects of the legal environment each involve multiple causal orderings: Law not only shapes organizational activity, but organizational activity simultaneously shapes law—and at times both take shape together. Third, for each aspect and each causal order, the relevant mechanisms may operate either through material rewards and penalties, or through cultural morals and meanings, or through complex mixtures of the two. The following pages examine the various images of organizations and the law that lie at the intersections of these multiple dimensions.

1. Intellectual History

1.1 Classical Roots

Classical social theorists saw organizations and the law as closely linked. To Marx (e.g., 1971a, 1971b), capitalist organizational structures provided the economic premises for 'bourgeois law,' with its characteristic coupling of formal individual rights—especially rights of property and contract—and repressive police interventions against the working class. To Durkheim (1964), society's evolution from a unified to a specialized division of labor (and the corresponding shift from communal to limited-purpose organizations) generated a parallel evolution in law, from punitive demonstrations of shared outrage to restitutive compensations for individualized damages. To Weber (1978, pp. 654–58, 956–1005), both modern organizations and modern law drew legitimacy from a shared 'legal-rational' reliance on preplanned, explicit, impersonal rules; as a result, both embodied similar tensions between responsiveness and predictability, and between autonomy and domination.

1.2 Mid-century Revival

Despite these early linkages, subsequent scholarship on organizations and on law developed for the most part independently. A few theorists (e.g., Hurst 1964) attempted to bridge the gap, but most research in each tradition remained oblivious to the other. Since the mid-twentieth century, however, the separation between organizational and legal studies has become increasingly untenable. Organizations form through the legal act of incorporation, and they dissolve

through the legal act of bankruptcy. In between, they raise capital under securities law, hire employees under labor law, and exchange goods and services under contract law. Law, for its part, arises from the demands of organizations, and it topples under pressure from organizations as well. In between, it develops meaning through organizational compliance, receives enforcement through organizational litigation, and encounters limits through organizational recalcitrance. Moreover, the organizational and legal worlds have become increasingly intertwined over time, as organizations face rising levels of litigation, regulation, and certification, and as law confronts an ever more bureaucratized society.

In the light of these affinities, a reconciliation between organizational and legal studies was perhaps inevitable, and by the mid-1960s, the early signs of convergence were beginning to emerge. Three lines of scholarship contributed to this rapprochement: a political-science literature on regulation, a sociology literature on legalization, and an economics literature on transaction costs.

1.2.1 Studies of regulation. First, from political science (and to a lesser extent from economics and sociology), a number of researchers became interested in the relationship between regulatory agencies and the industries those agencies supposedly oversaw (e.g., Bernstein 1955, Stigler 1971). Although many such investigations started with the assumption that well-chosen penalties could deter organizational wrongdoing, the empirical evidence suggested a more gloomy image: Industries often appeared to 'capture' their regulators, resulting in legal rules that limited competition and preserved prevailing practices, rather than imposing reforms. Moreover, even when agencies did attempt to regulate in the public interest, regulated organizations were less likely to be deterred from wrongdoing than to be impelled toward evasion.

1.2.2 Studies of legalization. The pessimism of the regulation literature was counterbalanced by a second body of research, originating from sociology. Examining the emergence of disciplinary hearings and other law-like procedures in the workplace, Selznick (1969) suggested that administrative and political pressures could lead organizations to develop internal practices that drew upon the public legal order for models of fairness and objectivity. Over time, these introjections of 'legality' tended to become institutionalized—both within organizations as formal policies and procedures, and outside organizations as doctrines and statutes protecting employee rights. Such legalization, Selznick argued, held the potential to transform organizations from authoritarian hierarchies into normatively-constrained polities, thereby limiting official power and securing substantial 'citizenship' rights for lower-level participants (see *Legalization*).

1.2.3 Studies of transaction costs. In contrast to the literature on regulation and legalization, a third body of scholarship, emerging primarily from economics, downplayed concerns about making organizations more socially responsible and focused instead on making them more efficient. Drawing heavily on the work of Coase (1937) and Williamson (1975), this tradition linked law and organizations through a set of general arguments about the costs and benefits of 'hierarchical governance structures,' whether legal rules or organizational policies. Because human beings are 'boundedly rational' (i.e., unable to anticipate all future contingencies) and 'opportunistic' (i.e., prone to self-interested dishonesty), certain exchanges will exhibit sizable 'transaction costs' in the market: detailed contracts will be difficult to write and enforce, and consequently hierarchical authority structures may prove more efficient than open trading. Since organizations and laws represent alternative forms of hierarchical governance, this analysis speaks equally to concerns about when the legal system should intervene in market activity and to concerns about when organizations should internalize market exchanges (see *Transaction Costs and Property Rights*).

2. Current Scholarship

In recent years, these three traditions have spawned a resurgence of scholarship on organizations and the law. This resurgence spans the disciplines of political science, sociology, and economics, with offshoots extending into traditional legal and managerial studies. This section reviews the state of this emerging literature, with emphasis on the three major aspects of the legal environment introduced above. Following Edelman and Suchman (1997), these are designated the 'facilitative,' 'regulative,' and 'constitutive' legal environments, respectively.

2.1 Law as an Arena of Interaction

At its most passive, the legal system simply provides places, personnel, and procedures to accommodate organizational interactions that might otherwise occur elsewhere. The archetypal manifestation of this 'facilitative' aspect of law is the lawsuit: organizations routinely employ litigation to open or defend markets, to attack or protect reputations, to reveal or conceal information, and so on. Courthouses are not the only

relevant legal facilities, however. Parallel activities arise in other venues as well; for example, organizations use Freedom of Information requests as vehicles for corporate espionage. The unifying theme is that, for better or worse, legal forums provide a crucial public infrastructure for organizational activity, not unlike marketplaces or convention centers. Consequently, while most studies of facilitative law portray organizations as the active players and the legal system as merely the passive arena, most also acknowledge that the contours of this arena may determine the course of the game.

While contemporary legal systems provide facilities for a wide array of organizational activities, the existing literature focuses almost exclusively on disputing—primarily through traditional litigation, but also through 'alternative dispute resolution' (ADR; see *Mediation, Arbitration, and Alternative Dispute Resolution (ADR)*) techniques, such as mediation and arbitration (see *Dispute Resolution in Economics*; *Litigation*) The best available evidence suggests that, in the USA at least, litigation rates have risen dramatically since the early 1970s, largely owing to an increase in the frequency with which business organizations sue one another. Alongside this rise in litigation, however, there has also been a substantial growth in the prevalence of ADR procedures, which proponents claim can curtail the cost, delay, and adverse publicity of conventional adjudication. The resulting climate has been described as one of 'process pluralism' (Lande 1997), with an elevated level of overall disputing activity accompanied by a reduced degree of courtroom centrality.

2.1.1 Facilitative law as independent variable. Most scholarship on these trends emphasizes the role of law as an independent variable: changes in the contours of legal forums, researchers argue, produce changes in the behaviors of organizational actors. Because much of this literature emerges from policy debates over litigation reform, investigation generally centers on whether particular procedural rules elicit or suppress lawsuits, and on whether changing rates of litigation exert positive or negative influences on nonlegal organizational activities. Standing-to-sue doctrines, class-action requirements, and contingency fee arrangements have all attracted attention as determinants of litigation behavior, and rising litigation has, in turn, received blame for a wide variety of systemic side-effects, including increased insurance use, elevated bankruptcy rates, and a societal undersupply of high-risk innovation. At the same time, a more up-beat literature on ADR asserts that the growing burdens of full-dress litigation have sparked managerial interest in nonjudicial alternatives, and that the resulting shift in dispute processing has eased workplace tensions and strengthened ongoing business relations. Other commentators, however, have noted that litigation may serve the salutary purpose of promoting corporate social responsibility, and that ADR, in contrast, may cool out genuine grievances and undercut public debate.

Research on the link between forum structure and disputing behavior generally adopts a materialist orientation, examining how legal procedures affect the concrete costs and benefits of various organizational actions. If litigants enter court strategically, then the 'house rules' should determine both the attractiveness of the arena and the moves of the players. In addition to such obvious factors as expense, delay, and the magnitude of awards, litigation rates should be particularly sensitive to the predictability of trial outcomes, since rational parties will generally prefer to settle their disputes out of court, if both sides have similar expectations about the results of a trial (Priest and Klein 1984). The most efficient judicial system, in this view, would be one in which the predictability of legal procedures encouraged disputants to resolve most matters through private negotiation, leaving the courts free to decide genuinely contested points of law with as little expense and delay as possible.

Recently, however, a number of culturalist researchers have borrowed from the literature on individual-level disputing to suggest that such rational-actor analyses may provide only a partial account, at best. Even among sophisticated business organizations, legal disagreements often comprise substantial expressive and symbolic elements, and these cultural factors may play as great a role as rational profit maximization in determining organizational disputing behavior. Thus, where materialists see litigation as 'economic competition by other means,' culturalists emphasize the relationship between law and 'softer' organizational phenomena such as alliance formation, community building, and reputation management. This perspective is particularly common in studies of ADR, which highlight the ability of dispute processing institutions to preserve or disrupt relationships, to harmonize or polarize viewpoints, and to privatize or publicize disagreements.

2.1.2 Facilitative law as dependent variable. Alongside these analyses of how legal arenas shape organizational behavior, a handful of studies have reversed the causal arrow, to explore how organizational behavior shapes legal arenas. As the transaction-cost tradition suggests, law is only one of many governance mechanisms, and if litigation arises in part from the failure of other controls, then extra-legal organizational conditions may play a crucial role in shaping the work of the courts.

For the most part, the existing literature assumes that the relevant extra-legal conditions are material rather than cultural, with factors such as economic globalization, industrial turbulence, and relational discontinuity topping the list. Clearly, however, cult-

ural conditions may also exert a significant influence. Preliminary findings suggest, for example, that professional ideologies strongly affect whether an organization's culture favors ADR or litigation; and organizational cultures, in turn, affect whether employees seek to resolve workplace grievances in-house or through the courts. Further, in a mirror image of Selznick's legalization dynamic, the spread of ADR techniques in the workplace tends to inject private organizational values into public legal discourses: Several major employment laws now explicitly encourage the use of ADR to resolve legal claims, and courts themselves increasingly employ 'appended' ADR to expedite settlement negotiations. Thus, the changes that began with rising litigation rates now appear to be coming full circle: where organizations once embraced ADR to escape the perils of formal legal arenas, the legal system is now remodeling its own facilities to replicate these initially alien organizational proceedings. Even when legal institutions are at their most passive, the relationship between law and organizations proves to be both reciprocal and complex.

2.2 *Law as a Code of Conduct*

In addition to providing a procedural forum for organizational interaction, law can also act as a system of substantive edicts, invoking societal authority over various domains of organizational activity—from antitrust to labor relations to environmental protection (see *Antitrust Policy*; *Environmental Policy: Protection and Regulation*). Formal administrative oversight represents the archetypal manifestation of this 'regulative' aspect of law, but similar applications of sovereign authority can also occur through common-law precedents, street-level policing, and privately enforced policy initiatives, such as civil claims against harassment or fraud. In all its guises, regulative law operates not merely as a neutral arena, but as a structure of precepts and enforcement mechanisms designed to channel organizational behavior in particular directions. While regulative law may not always accomplish (or even sincerely intend to accomplish) its stated goals, the distinguishing features of this aspect of the legal environment are its pervasive rhetoric of active social control and its pervasive concern with organizational compliance and evasion.

Regulation is almost certainly the most widely studied aspect of the relationship between organizations and the law. Unfortunately, most investigations to date have focused narrowly on particular substantive areas, and systematic efforts to explore underlying social processes remain relatively rare. Nonetheless, this literature displays a number of recurring materialist and culturalist themes. From a materialist perspective, regulative law serves to counterbalance market imperfections, imposing supplemental incentives and deterrents that recalibrate the implicit prices of various organizational behaviors. Organizations, for their part, seek to maximize total returns in the light of these prices, whether through petition and compliance or through manipulation and evasion. Alongside such economic dynamics, however, legal regulations also implicate significant normative and cognitive beliefs, and tactical maneuvers often cannot fully extricate organizations from these broader cultural processes (see *Schemas, Frames, and Scripts in Cognitive Psychology*; *Norms*). As statements of public values, laws shape and reflect what organizational stakeholders want, and as repositories of public meaning, laws frame and recapitulate what organizational stakeholders expect. Consequently, the underlying cultural logics of the organizational and legal realms often move in tandem, even as particularly recalcitrant firms seek to amend or evade particularly burdensome strictures.

2.2.1 Regulative law as independent variable. Tracing, first, the causal pathway from regulative edicts to organizational responses, research suggests that such legal pronouncements can operate both as material threats and as cultural sermons. As material threats, laws exert their greatest influence when standards are clear, sovereignty uncontested, and sanctions substantial (see *Deterrence: Legal Perspectives*). In practice, however, such ideal conditions rarely pertain, and in their absence, self-interested organizations react not only to the incentives enunciated by law 'on the books,' but also to the real-world gaps provided by ambiguous statutes, fragmented jurisdictions, and manipulable enforcement agencies. These gaps, substantial in themselves, grow still wider in the face of organizational characteristics such as specialization and decentralization, which marginalize legal personnel and insulate managers from responsibility for corporate transgressions. As a result, the history of regulation is replete with unintended consequences: disclosure requirements foster uninformative disclaimers, tort liability shifts hazardous activities into undercapitalized subsidiaries, and air-quality standards export pollutants to downwind communities. Although few would rule out the possibility of successful deterrence, the common theme behind much of this materialist research is that regulative law may exert its largest effects by motivating evasion, not compliance.

The pessimism of this literature stems in large part from the assumption that rational organizations lack the moral sentiments that promote individual-level obedience to law (see *Compliance and Obedience: Legal*). Recent studies of law as a cultural sermon, however, call this assumption into question. Organizational behavior, after all, reflects the activity of culture-bearing individuals operating within decision

structures that incorporate cultural values and presumptions at every turn. The leading attempt to examine regulative law through this culturalist lens comes from 'neo-institutional' sociologists, working broadly in the Selznickian tradition. To these researchers, legal compliance simply represents a special case of 'institutional isomorphism'—the process by which organizations facing similar cultural rule systems become both increasingly alike and increasingly congruent with the myths of their shared environment. In addition to applying material coercion, legal institutions also homogenize organizational behavior through both normative and cognitive beliefs: law enunciates values and ethics which organizations then elaborate and internalize; and law provides models and scripts, which organizations then imitate and enact. While the resulting isomorphism often amounts to little more than superficial ceremony, the neo-institutional perspective suggests that even 'empty' symbolism re-enacts and thereby substantiates the law's cultural claims in ways that outright defiance would not.

Thus, investigations into the impact of regulative law on organizational activity suggest an ironic result: as a material deterrent, law is rarely sufficiently explicit, authoritative, and coercive to accomplish much beyond mere symbolism. As a cultural pronouncement, however, 'mere symbolism' can prove consequential—often extending a law's practical reach (for better or worse) well beyond its official targets.

2.2.2 Regulative law as dependent variable. Investigations into the reciprocal impact of organizational activity on regulative law reveal similarly complex patterns. At the most basic level, research confirms conventional wisdom: Organizations make substantial investments in electoral politics and backstage agenda-setting, and organizational actors occupy central positions on all sides of most policy networks (see *Policy Networks*). Contrary to many popular accounts, however, organizations often seem less interested in averting government 'interference' than in using regulation to stabilize markets or outmaneuver competitors. Legal incentives are easier to control than market prices, and as materialist accounts would predict, organizations invest heavily in securing this resource pool. These efforts extend well beyond formal lawmaking to encompass subsequent implementation activities as well. Legislative compromises can leave regulatory agencies with little substantive guidance and even less monetary support, and as a result, administrators often find themselves dependent on outside organizations for information, expertise, and even political cover. Far from imposing top-down social control, the resulting process of regulatory capture tends merely to institutionalize the indigenous practices of the regulated industry itself (see *Lawyers, Regulation of*).

Culturalist accounts supplement this Machiavellian power-dependence analysis in two ways. First, organizational participation in the lawmaking process may reflect ideological concerns, as well as material interests. Professional and social movement organizations, in particular, often blend moral entrepreneurship with institutional entrepreneurship, to align the regulative environment with their own ethics and worldviews (see *Entrepreneurship*). Second, organizational and industrial culture may act as a 'filtering agent' between 'law on the books' and 'law in action' (Kidder 1983). Since popular understandings of legality come largely from mundane experiences in concrete settings, living law ultimately emerges not from legislatures or regulatory agencies but from real-world communities, where organizational beliefs and practices give form and content to otherwise vague legal principles. Indeed, many legal enactments are initially so ambiguous and contradictory as to do little more than signal that *something* in the regulative environment has changed—leaving the actual meaning of compliance to be constructed by subject organizations themselves, through an iterative process of interpretation, imitation, litigation, and reinterpretation (Edelman 1992).

This culturalist perspective suggests a somewhat less sinister view of regulatory capture. Both regulators and regulateds operate within the same institutional fields, and their shared exposure to a common set of filters and experiences can produce congruent outlooks even in the absence of *quid pro quo* power relations. Moreover, this dynamic can sometimes amplify, rather than dampen law's regulative impact: The literature reveals several instances in which savvy lobbyists, lawyers, and personnel officers have actually *exaggerated* legal demands in order to increase their own symbolic and practical importance.

2.3 Law as a Typology of Experience

As well as housing organizational interactions and imposing social control, law can also enter organizational life more subtly, as a definitional framework that identifies the legally acknowledged components of the social world and that explains their natures and attributes. The archetype of this 'constitutive' aspect of law is the doctrine of corporate personhood, which defines particular social groups as unitary collective actors with coherent legal identities, interests and capacities (see *Multinational Corporations*). Beyond incorporation, however, constitutive law also underpins virtually every other feature of the organizational world, from contracts and securities to intellectual property and employment (see *Employment and Labor, Regulation of*; *Intellectual Property: Legal Aspects*). Unlike both facilitative and regulative law, constitutive law rarely catches the public eye, but its invisibility belies its agenda-setting power: By estab-

lishing fundamental background assumptions, legal definitions help to determine which types of organizations come into existence and which types of activity gain recognition. As a toolkit of basic templates and mechanisms, constitutive law often seems more placid and routine than its facilitative and regulative counterparts; in reality, though, this aspect of the legal environment provides the building blocks that underpin the other two.

2.3.1 Constitutive law as independent variable. Because the materialist perspective emphasizes concrete conditions, it tends to downplay law's constitutive aspects: the physical realities of organizational life should not depend on the presence or absence of legal descriptions—especially since economic actors can usually specify, by mutual agreement, anything that is not already defined by law. Nonetheless, hints of a materialist approach to constitutive law appear, at least embryonically, in certain writings from the transaction cost school. Masten (1990), for example, suggests that the fundamental distinction between markets and hierarchies lies in the different 'default rules' that govern these two types of economic activity. Although a carefully crafted network of market contracts could theoretically achieve the same results as a corporate charter, the ready-made template of incorporation presumably eases the cognitive burden of constructing such a relationship. Given that human beings are only boundedly rational, it follows that many organizations would never come into existence without this supportive legal definition. Constitutive law, then, can exert a substantial influence on the organizational world, simply by creating a basic framework of categories and rights (Campbell and Lindberg 1990, Dobbin and Sutton 1998).

Culturalist analyses go even further, seeing constitutive law not merely as a starting point for strategic negotiation, but as a fundamental preconscious framework for making sense of the social world. In this view, law plays a central role in the 'social construction of reality,' with legal definitions often becoming reified and institutionalized as taken-for-granted components of 'the way the world works' (see *Disputes, Social Construction and Transformation of*; *Institutionalization* Law as Constitutive; *Reification: History of the Concept*).

Researchers have explored such constitutive dynamics in several aspects of organizational life, including inter-organizational relations, organizational structures, and the general cultural logic of 'legal rationality.' In the constitution of interorganizational relations, contract law establishes rituals for forming binding commitments, property law determines the appropriability of resources and ideas, tort law assigns responsibility for social misfortunes, and so on (see *Contracts: Legal Perspectives*; *Property: Legal Aspects*; *Torts in Law*). In the constitution of organizational structures, law defines the characteristics of particular organizational features, such as affirmative action policies and 'poison pill' takeover defenses, and law also enunciates the ground-rules for entire organizational forms, such as private companies, public agencies, and nonprofit associations. Finally, in the constitution of legal rationality, law generates sympathetic vibrations in other kindred social structures, leading organizational embodiments of 'fairness,' 'citizenship,' and the like to mimic their legal counterparts.

2.3.2 Constitutive law as dependent variable. In depicting legal definitions as the basic atoms of social structure, researchers from both the materialist and the culturalist camps tend to give short shrift to the study of constitutive law as a *dependent* variable. Nonetheless, legal categories, no matter how fundamental, are hardly facts of nature, and investigations into the relationship between organizations and the law have much to gain from examining how organizational activity may shape various constitutive constructs. The existing literature offers several thought-provoking suggestions. First, in Gilson's (1984) memorable phrase, corporate lawyers may act as 'transaction cost engineers,' responding to unresolved governance challenges by assembling pre-existing constitutive building blocks into novel configurations. Over time, the resulting 'legal devices' can serve as prototypes for increasingly routine limitations —and these standardized reproductions can, in turn, provide the basis for revised legal categories (Powell 1993). Second, even without intentional innovation, organizations may inadvertently destabilize established legal definitions through imperfect implementation or unsuitable elaboration, generating legal pluralism and contradiction as an unintended consequence of everyday activity. Third, since legitimacy often stems from sheer numerical prevalence, organizations may skew the development of constitutive law just by disproportionately employing some legal alternatives rather than others. Finally, as described above, organizations may participate directly in creating and implementing new legislation, shaping the constitutive legal environment as advocates, as filtering agents, and simply as concrete 'data points' that give substance to the law's abstract labels.

3. *Future Directions*

The study of organizations and the law has accomplished much in recent years; however, much still remains to be done. At a technical level, although both empirical and conceptual analyses abound, the former tend to focus on narrow policy concerns, while the latter tend to rest on meager evidentiary foundations. The challenge for future researchers will be to build

firmer linkages between theory and data, and from those linkages, to develop a systematic body of intercomparable results. Cross-temporal and cross-cultural studies, in particular, seem essential to resolving the field's persistent debates over causal sequences and material/cultural mechanisms.

As new research takes shape, investigators are likely to encounter many phenomena that bridge the three major aspects of law and that comingle material and cultural processes. Regulative statutes, for example, acquire meaning largely through the test cases that they generate in facilitative legal arenas—and the meaning thus acquired often includes clarification of constitutive legal definitions, and certification of 'compliant' legal devices (Edelman 1992). In such dynamics, cultural and material considerations presumably overlap, since most laws set forth norms and penalties simultaneously. Indeed, under certain circumstances, the material/cultural distinction can virtually disappear, as when material punishment becomes a symbol of societal commitment, or when cultural legitimacy becomes an asset in market competition (Suchman 1997).

In exploring such intersections, future research may accelerate a trend, already emerging among neo-institutionalists, toward conceptualizing the relationship between organizations and law less as a matter of unilateral causation than as a matter of endogenous coevolution: Given the inherent ambiguity of law on the books, the practical meaning of law in action develops only through a highly interactive process of social construction—involving both symbolic sense-making and concrete dealmaking. This 'structuration' dynamic encompasses not only official agents of the legal system but also members of the focal organizational field, all tacitly collaborating to create *simultaneously* the forums and procedures of facilitative law, the demands and sanctions of regulative law, and the entities and devices of constitutive law. Eventually this collaboration may produce a working agreement on what the law 'is' and on what it 'requires,' and that agreement may then become institutionalized in formal conventions and taken-for-granted beliefs. But saying that legal rules 'cause' organizational practices (or vice versa) would be a gross oversimplification. Rather, rules and practices arise in tandem, as part of the consolidation of a single unified organizational/legal field.

See also: Law: New Institutionalism; Legal Personality; Regulation and Administration

Bibliography

Bernstein M 1955 *Regulating Business by Independent Commission*. Princeton University Press, Princeton, NJ

Campbell J, Lindberg L 1990 Property rights and the organization of economic activity by the state. *American Sociological Review* **55**: 634–47

Coase R 1937 The nature of the firm. *Economica* **4**: 386–405

Dobbin F, Sutton J 1998 The strength of a weak state: The rights revolution and the rise of human resources management divisions. *American Sociological Review* **104**: 441–76

Dunworth T, Rogers J 1996 Corporations in court: Big business litigation in U.S. federal courts, 1971–1991. *Law and Social Inquiry* **21**: 497–592

Durkheim E 1964 [1893] *The Division of Labor in Society*. Free Press, New York

Edelman L 1992 Legal ambiguity and symbolic structures: Organizational mediation of law. *American Journal of Sociology* **97**: 1531–76

Edelman L, Suchman M 1997 The legal environments of organizations. *Annual Review of Sociology* **23**: 479–515

Gilson R 1984 Value creation by business lawyers: Legal skills and asset pricing. *Yale Law Journal* **94**: 239–313

Hurst W 1964 *Law and Economic Growth: The Legal History of the Lumber Industry in Wisconsin 1836–1915*. Belknap Press, Cambridge, MA

Kidder R 1983 *Connecting Law and Society*. Prentice-Hall, Englewood Cliffs, NJ

Lande J 1997 How will lawyering and mediation practices transform each other? *Florida State University Law Review* **24**: 839–75

Marx K 1971a [1844] On the Jewish question. In: McLellan D (ed. and trans.) *Early Texts*. Blackwell, Oxford, UK, pp. 85–114

Marx K 1971b [1842] Proceedings of the Sixth Rhenish Parliament: Debates on the law on thefts of wood. In McLellan D (ed. and trans.) *Early Texts*. Blackwell, Oxford, UK, pp. 49–51

Masten S 1990 A legal basis for the firm. In: Williamson O, Winter S (eds.) *The Nature of The Firm: Origins, Evolution, and Development*. Oxford University Press, New York, pp. 196–212

Powell M 1993 Professional innovation: Corporate lawyers and private lawmaking. *Law and Social Inquiry* **18**: 423–52

Priest G, Klein B 1984 The selection of disputes for litigation. *Journal of Legal Studies* **13**: 1–55

Selznick P 1969 *Law, Society, and Industrial Justice*. Sage, New York

Stigler G 1971 The theory of economic regulation. *Bell Journal of Economics* **2**(Spring): 3–21

Suchman M 1997 On beyond interest: Rational, normative and cognitive perspectives in the social scientific study of law. *Wisconsin Law Review* 1997: 475–501

Suchman M, Edelman L 1996 Legal rational myths: The new institutionalism and the law and society tradition. *Law and Social Inquiry* **21**: 903–41

Weber M, Roth G, Wittich C (eds.) 1978 [1925] *Economy and Society*. University of California Press, Berkeley, CA

Williamson O 1975 *Markets and Hierarchies*. Free Press, New York

M. C. Suchman

Organizations: Authority and Power

There are countless different ways of defining power in organizations and its relationship to authority. In this chapter, these differences are explored in terms of three broad perspectives identified in the literature—*managerialist*, *critical*, and *discursive* ap-

proaches. Each approach is discussed in depth, and their assumptions and insights compared. (Also see Hardy and Clegg 1999, Palmer and Hardy 2000 for extended discussions of different approaches to power.)

1. Managerialist Approaches: Authority vs. Power

Managerialist research on authority and power is predicated on a rational orientation, which seeks to make organizations more efficient and effective (or to understand why they are not). Authority is defined as the recognized, legitimate right to make a decision. Power is typically employed by those opposing management, resisting change, or pursuing self-interest, and who seek to subvert the 'legitimate' power embodied in the organizational hierarchy. When writers acknowledge that managers use power, as opposed to authority, they suggest that it is only in self-defense—to pursue organizational goals in the face of resistance. Accordingly, managerialist work on power stems from an interest in why decisions which 'ought' to be rational are instead political as interest groups use their power to influence decision outcomes in their favor. Writers in this tradition thus tend to differentiate between 'illegitimate' power and 'legitimate' authority. Since authority is a 'normal' and 'inevitable' phenomenon that results from the formal design of the organization, it is largely excluded from analysis and research concentrates on 'illegitimate' power exercised outside formal hierarchical structures.

The managerialist approach to power in organization and management theory (OMT) has been heavily influenced by early community studies of power carried out in other disciplines. In this section, an overview of this work is provided, followed by a discussion of the OMT literature in more detail.

1.1 Community Studies of Power

Early studies of community power focused on how decision outcomes are produced (Dahl 1957). Researchers analyzed key decisions to identify the power relations prevailing in a particular community. The object was to determine who made these decisions: if the same groups were responsible, the community could be said to be ruled by an élite group. If, on the other hand, a variety of groups had an impact on decision outcomes, the community—and also American society in general—could be described as pluralist. These researchers found the latter.

Several assumptions underlay this research. First, all individuals were assumed to be aware of their grievances and to act upon them by participating in the decision-making process and using their influence to determine key decisions. Second, the decision-making arena was assumed to be open to everyone and, therefore, the absence of participation indicated that actors agreed with decisions. Third, it was assumed that power was exercised only in decisions where conflict was clearly observable, and that conflict was resolved only through decision-making and not in other ways (Lukes 1974).

Lukes (1974) refers to this work as a one-dimensional view of power. He identifies a second dimension that emerged as researchers challenged the assumptions underlying the work of the pluralists. Bachrach and Baratz (1962) questioned whether decision-making processes were open to everyone, and argued that nonparticipation meant that individuals could not secure access to decision-making processes, rather than reflecting genuine satisfaction. Research began to focus on how participation was constrained by the suppression of the needs of the nonparticipants. Bachrach and Baratz showed that issues were excluded from decision-making, enabling actors to confine the agenda to 'safe' questions and the arena to 'safe' participants. They called this process 'nondecision-making' because it allowed the more powerful actors to determine outcomes from behind the scenes.

While the work on nondecision-making broadened the research agenda, it continued to operate on similar assumptions as the one-dimensional view of power. Specifically, it assumed that some form of conflict was necessary for power to be exercised (Lukes 1974). Thus, from a theoretical perspective, the first and second dimensions converge: both focus on the exercise of power in or around the decision-making arena as part of a deliberate strategy to achieve intended outcomes. Although the power mobilized through nondecision-making may be less visible than decision-making power—since opponents may be sidelined rather than directly confronted—it is still premised on the existence of conflict and opposition.

1.2 Studies of Power in OMT

An examination of early studies of power in OMT shows that, although conducted separately and with a different rationale, they adopted a similar view of power as the community studies described above. Many studies focused on uncovering the sources of power that allowed actors, who did not possess legitimate authority, to influence decision outcomes. In this way, researchers began to differentiate between formal, legitimate power—authority—and informal, illegitimate power.

Other work linked informal power to the ability to control uncertainty. The strategic contingencies theory of intraorganizational power (Hickson et al. 1971) went on to argue that subunits were powerful when they were able to cope with greater amounts of uncertainty, were central and not easily substitutable—in other words when other subunits depended upon them. A similar theoretical development

came with the resource dependency view (see Pfeffer 1981). Theorists started to identify the resources that constituted the bases or sources of power, such as information, expertise, credibility, prestige, access to higher echelon members, and the control of money, rewards, and sanctions (Pettigrew 1973). They also questioned whether simply possessing scarce resources was enough to confer power, arguing that actors also had to be aware of their contextual pertinence and sufficiently skilled to deploy them. This process of mobilizing power was typically referred to as politics (Pettigrew 1973), which further reinforced the idea that the use of power was illegitimate, dysfunctional, and self-serving.

This approach to power concentrates on the link between power and the control of scarce resources that help actors to influence decision-making processes and defeat opposition. Those individuals with the greatest access to power sources and the greatest skill at putting them to use were the most likely to prevail in decision making. Decision outcomes produced in this way were described as political, rather than rational, and thought to be at odds with the interests and objectives of those in formal authority. That power was used in the first place was due to the existence of conflict:

> it is clear that political activity is activity which is undertaken to overcome some resistance or opposition. Without opposition or contest within the organization, there is neither the need nor the expectation that one would observe political activity (Pfeffer 1981, p. 7).

In this way, the use of power was inextricably linked to situations of overt conflict associated with the struggles as actors sought to protect vested interests.

This work on power converges theoretically with the study of the first and the second dimensions of power. First, both community and OMT researchers challenged existing models of decision making. Pluralists refuted élitist views of decision making predominant in sociology, while management studies of power challenged the existence of the rational model of decision-making (although they often, implicitly or explicitly, ascribed to it as the desirable mode of functioning). Second, both bodies of literature adopted a behavioral approach that focused on the overt exercise of power in the decision-making arena, and both argued that decision outcomes were influenced by actors' access to and expertise in using power. Third, both are interested in how power determines decision outcomes (albeit that power itself is dysfunctional compared to the preferred rational functioning of organizations). Finally, both bodies of work employed similar definitions of power, i.e., the ability to get others to do what you want them to, if necessary against their will (Weber 1947) or to do something they otherwise would not do (Dahl 1957).

While managerialist approaches in OMT have produced insight into some aspects of power, existing organizational arrangements—particularly in the guise of legitimate authority—are excluded from analysis. Power represents dysfunctional interruptions in the far more desirable flow of rational decisions by self-interested actors who engage in its *political* mobilization, a term whose negative connotations has helped to reinforce the view that power is illegitimate and dysfunctional. As power and authority were pulled further apart, interest in the latter was replaced by increasing fascination with the former. Consequently, as the interest in power grew, the managerialist approach increasingly took authority for granted, never wondering how it got there or whose interests it served—power was what counted.

2. *Critical Approaches: Power, Authority and Domination*

Research in the critical tradition, a term used here to refer to a diverse body of research, adopts a rather different approach to power and authority than the managerialist studies described above. Critical theorists argue that the power embedded in formal organizational structures and processes is not neutral, rational, or logical but, rather, it specifically serves managerial interests. Hence, hierarchy is not authority, but power and domination; and actions taken to challenge it constitute resistance to domination, and not the illegitimate use of power. Critical theorists examine how power is embedded in societal and organizational structures, ideologies and narratives. They assert that much of what appears to be apolitical is, in fact, both a product and medium of power. This 'radical' view (Lukes 1974) has helped to expose how power, concealed in the legitimate trappings of structures, rules, class mechanisms and cultures, produces quiescence and prevents conflict from arising. In this section, the development of this 'radical' view of power is examined and its somewhat uneven adoption in OMT explored.

2.1 *A Radical View of Power*

Having identified the first and second dimensions of power described above, Lukes (1974, p. 24) presented a third dimension of power. He argued that power can be used to shape perceptions, cognitions, and preferences to the extent that individuals 'accept their role in the existing order of things, either because they can see or imagine no alternative to it, or because they view it as natural and unchangeable, or because they value it as divinely ordained and beneficial.' Consequently, the study of power cannot be confined to observable conflict, the outcomes of decisions, or even suppressed issues. It must also consider the question of political *quiescence*: why grievances do not exist; why demands are not made; and why conflict does not arise. Furthermore, this use of power is particularly im-

portant in sustaining the dominance of élite groups by reducing the inclination of subordinates to employ their discretionary power.

Lukes' interest in the societal and class mechanisms that perpetuate the status quo relates to the ability of dominant classes to define reality to support and justify their material domination, thus preventing challenges to their position, and avoiding the necessity of using more coercive, visible forms of power. According to this view, stable organizing power is based on the ability to avoid conflict by rendering power less visible and making it appear legitimate.

2.2 *Radical Views in OMT*

The fact that organizations tended to serve managerial interests had not been lost on organizational researchers. Nor had the political advantages of legitimacy gone completely unnoticed. One writer who attempted to draw legitimation processes into the OMT fold was Andrew Pettigrew, who argued that politics involved creating legitimacy for demands through the management of meaning. Pettigrew (1979) showed how interest groups legitimized their demands to prevent conflict through symbolic activity. Other writers examined the way in which power was embedded at a 'deeper' level in organizational structures and ideologies (e.g., Clegg 1975). In both cases, political actors try to section off spheres of influence where domination is perceived as legitimate and thus unchallenged. This use of power is largely invisible and designed to avoid conflict rather than defeat it (Hardy 1985).

Another group that adopted a critical perspective was labor process theorists who focus on the dynamics of power and resistance in relation to phenomena such as gender, technology, ethnicity, managerial work, and other aspects of the structuration of work and its organizational context. More recently, such work has critically examined the empowerment and teamwork practices that characterize contemporary management practice. Far from providing employees with more power, these practices embody subtle, invisible controls. Writers argue that empowerment terminology helps to avoid conflict and resistance by emphasizing consensus and cooperation, and managerial control is reinforced through the language of the team effort and peer pressure. A critical perspective sees empowerment as an exercise in the management of meaning, designed to enhance the legitimacy of organizational goals and to influence behavior unobtrusively (Hardy and Leiba-O'Sullivan 1998).

Despite a growing body of critical work, many writers in OMT continue to cling to managerialist assumptions and to define power in terms of conflict and illegitimacy. Rather than delve into the power hidden in and mobilized through apparently neutral structures, cultures, and technologies, these researchers preferred to view these constructs as apolitical management tools. For example, mainstream writers on organizational culture have gone to considerable length to avoid any association with power and politics, presenting cultural change in neutral terms which suggest that it is to everyone's advantage.

In summary, critical approaches to power are substantively different in their assumptions than managerialist views of power. Power is domination, authority is one form of domination, and resistance is a repudiation of that domination. Power and resistance are differentiated in order to indicate 'qualitatively different contributions to the outcome of power relations made by those who exercise power over others, on the one hand, and those subject to that power, on the other' (Barbalet 1985, p. 545). According to this approach, power is used not simply to defeat conflict, but to prevent it from arising in the first place. Critical theorists focus on this subtle, repressive use of power, attempting to reveal the hidden aspects of domination, helping to facilitate resistance, and enabling groups to free themselves from the effects of power.

3. *Discursive Power: Reconstructing Authority*

In this section, two bodies of work that examine the links between power and discourse are explored. First, the influence of Michel Foucault (e.g., 1980) is examined. Second, work on organizational discourse, which attempts to combine critical and Foucauldian insights, is assessed.

3.1 *Foucault: the End of Authority*

This approach challenges the very idea of authority by denying the existence of 'sovereign' power, which underpins both managerialist and critical approaches. Gone is the idea of a supreme authority that determines organizational outcomes, whether they are 'legitimate' goals or subtle forms of domination. Instead, a network of power relations—or discourses—captures the advantaged and disadvantaged alike.

> We use the term discourse as shorthand for a whole set of power/knowledge relations which are written, spoken, communicated and embedded in social practice. These relations have power and truth effects and it is the consequences that follow which are a major concern of a discourse analysis (Knights and Morgan 1991, p. 254).

Power as discourse is no longer a convenient, manageable resource under the control of autonomous, sovereign actors.

> People know what they do; they frequently know why they do what they do; but what they don't know is what what they do does (Foucault, quoted in Dreyfus and Rabinow 1982, p. 187).

In other words, actors may manipulate resources, structures, or meanings with the intention of achieving particular outcomes but, according to this view, no matter how hard they pull on the 'strings' of power they are unlikely to attain them.

The reasons for these limits to power and authority lie in the pervasiveness and complexity of power relations. First, this network of power produces the subject (Knights and Morgan 1991). According to postmodern views, the individual is not a stable constellation of essential characteristics but an occupant of multiple, socially constituted categories. Identities are salient only insofar as they are socially recognized. Thus the subject is not a 'given' but produced historically through power and knowledge. This effect of power is not necessarily alienating or repressive, as stressed by critical theorists: it can be productive and positive. Power transforms individuals 'into subjects who secure their sense of what it is to be worthy and competent human beings' (Knights and Morgan 1991, p. 269). This productive and positive side of power leads to the continuation, rather than the dismantling, of power relations. Also, resistance can exert a high price when the individual has to reject a part of oneself in challenging existing power relations.

Second, such approaches challenge the (critical) idea that one can escape regimes of power through knowledge, arguing that knowledge can ever be separated from the effects of power. Knowledge does not strip away the effects of power to reveal the 'truth'; all it accomplishes is the embodiment of new and different forms of power. What is accepted as 'true' are those mechanisms that 'enable one to distinguish true and false statements … the techniques and procedures accorded value in the acquisition of truth; [and] the status of those who are charged with saying what counts as true' (Foucault 1980, p. 131).

Third, the embodiment of power in every aspect of one's life—including identity—makes them difficult to repudiate. They are experienced as reality, making alternatives difficult to conceive of, let alone enact. Resistance can usually only be conceived of in terms of prevailing power relations and is subsequently 'colonized' by them. It therefore tends to reinforce existing power relations rather than overthrowing them (Clegg 1989). In this way, the emancipation for which critical approaches strive becomes illusory, despite the contention of those writers who argue otherwise.

Foucault's work posits a very different conceptualization of power than both managerialist and critical views. Not only is there no autonomous subject waiting—or willing—to be liberated through critical awareness, it is also erroneous to assume that emancipation can be achieved through the knowledge gained through critical reflection. Even if resistance was successful, the resulting situation could never be free of power—it would simply represent a *different* power regime. Moreover, these power relations affect all actors—capital and labor, employers and employees, researchers and research subjects. No one has the privilege of standing outside power.

> Power must be analyzed as something which circulates, or rather something which only functions in the form of a chain. It is never localized here or there, never in anybody's hands, never appropriated as a commodity or piece of wealth. Power is employed and exercised through a net-like organization … [Individuals] are always in the position of simultaneously undergoing and exercising this power (Foucault 1980, p. 98).

This power is hard to escape. It encompasses, surrounds, and infiltrates in ways of which individuals are often unaware. It is productive yet constraining; it makes them what they are, yet prevents them from becoming anything else. As actors try to achieve and escape power, authority becomes an illusion and politics a futile struggle.

3.2 Organizational Discourse: Resurrecting Authority

Foucault's work focuses attention on how the institutionalization of practices and structures embody sets of power relations that are deeply embedded in the taken-for-granted rules and practices that constrain action (Clegg 1989). This form of power is best understood as 'systemic.' It lies in the 'power of the system in the unconscious acceptance of the values, traditions, cultures and structures of a given institution or society' (Hardy 1994, p. 230). It also emphasizes the discursive aspects of power—how actors are locked into positions by way of the linguistic practices available to them to make sense of a situation. As the literature on power has evolved, writers have sought to combine critical conceptions of power with Foucauldian insights and, in so doing, challenge and modify both approaches to power.

This work emphasizes an analysis of discursive practices to show how discourses produce subjectivities and identities, constituting members in ways that define the possibilities for interaction in a particular context. This form of power lies neither in the possession of individuals nor the relationships between them, but is an inherent aspect of organizational life, reproduced by the day-to-day communicative practices and struggles of its members. Important power struggles are not those about the allocation of material resources or decision-making capabilities, but those that form around the ability to 'frame discursive and nondiscursive practices within a system of meaning that is commensurate with that individual's or group's own interests' (Deetz and Mumby 1990, p. 32). However, unlike critical theory's conception of 'real' interests and 'false' consciousness, discursive approaches show that, through the production of identity and subjectivity, interests are also discursively produced. Further, knowledge and meaning impinge on the identity of *all* individuals as they participate in the

production of organizational life, including those privileged as dominant and authoritative. Organizational reality is characterized less by the domination of one group over another, but by complex discursive practices that 'define what it means to be an organizational member and … allow for the privileging of managerial interests over others (Deetz and Mumby 1990, pp. 32, 39). Identity formation is thus embedded in communicative practices in which all actors participate and from which none can escape.

At the same time, these theorists have challenged Foucault's agentless and deinstitutionalized conception of power that has frustrated writers with its deterministic and pessimistic view of actors as powerless influence the discourses that define them (e.g., Reed 1998). These writers assume that actors *do* engage in purposive discursive activity as they attempt to produce particular outcomes (Deetz and Mumby 1990), albeit within the meanings and understandings embedded in the larger discursive context. Power is both enabling and constraining, productive and oppressive, local and institutional. In other words, discursive resources, embedded in multiple and contradictory discursive contexts, are mobilized by agents, even though their agency is constrained by larger discursive structures.

This approach sees discourse as the medium and product of power relations. It makes certain behaviors and practices possible and, in so doing, constitutes organizational reality. Discourses reproduce, transform, and obscure power relations and are, therefore, political.

> [D]iscourse as a political practice is not only a site of power struggle, but also a stake in power struggle: discursive practice draws upon conventions which naturalize particular power relations and ideologies, and shape these conventions themselves, and the ways in which they are articulated, are a focus of struggle (Fairclough 1992, p. 67).

Thus the act of invoking discourses through the creation and dissemination of texts and talk, and through the use of particular communicative strategies (e.g., rhetoric, narrarative, metaphors, etc.) can represent political strategies. Accordingly, researchers focus on how a particular discourse positions actors, which actors secure rights to speak, and which do not. They examine the way in which actors draw upon discourses, and the language, genres, and linguistic strategies they use. They explore the meaning attached to different categories of object in the discourse, the procedures, and practices that are invoked, and the way in which particular actors stand to gain or lose from the employment of the discourse.

In this way, researchers hope to understand the interconnections of structure and agency. Foucauldian theorists tend to dismiss both: structure, because they fail to explain how it got there in the first place; agency, because they do not believe that it exists. In this way, they are unable to distinguish between 'open doors' and 'brick walls' (Reed 1998, p. 209).

> If power/knowledge discourses are open, dispersed and ambiguous as Foucault suggests, then how does any particular discursive formation—such as 'enterprise', 'excellence' or 'designer culture'—achieve the level of institutional legitimation and range of organizational enactment presumed in the notion of disciplinary or 'panopticon' power? (Reed 1998, p. 208).

This work, implicitly or explicitly, combines insights from critical *and* postmodern views, despite the assertion of some writers that critical and postmodern epistemologies and ontologies cannot be reconciled. In so doing, it resurrects the notion of authority by acknowledging that certain actors and institutions are constructed as authorities and authoritative. Such authority may be a social construction but it is not random or arbitrary—purposive discursive activity enables some actors to construct themselves or others as authoritative, albeit with subject to limits and constraints, and in doing so, these actors are able to exercise power.

4. *Conclusions*

This chapter has examined a number of approaches to power and explored their insights into the relationship between power and authority. Managerialist views of power differentiate legitimate, formal authority from illegitimate, informal power. The former is assumed to be rational and functional, the latter is typically seen as 'political' and dysfunctional. This view of power is not contained within a particular body of literature on power; rather it permeates much of the mainstream work in OMT that sees organizations as, albeit flawed, rational decision-making systems. Critical approaches directly challenge managerialist approaches. They see organizations, not as rational, neutral, and objective but as managerial tools designed to fulfil the interests of one group at the expense of another. Power is embedded in the fabric of organizational life, hidden behind the mask of authority, and designed for domination. Discursive views of power wrestle with the notion of authority in different ways. Foucauldian approaches dispel the idea that either authority or power can be a pliable means of domination. Organizational discourse theorists have sought to rescue both power and authority from Foucauldian oblivion by resurrecting notions of agency and structure.

See also: Authority: Delegation; Authority, Social Theories of; Community Power Structure; Foucault, Michel (1926–84); Management: General; Organization: Overview; Organizational Culture; Organizational Culture, Anthropology of; Organizational Decision Making; Political Discourse

Bibliography

Bachrach P, Baratz M S 1962 The two faces of power. *American Political Science Review* **56**: 947–52

Barbalet J M 1985 Power and resistance. *British Journal of Sociology* **36**(4): 531–48

Clegg S R 1975 *Power, Rule and Domination*. Routledge, London

Clegg S R 1989 *Frameworks of Power*. Sage, London

Dahl R 1957 The concept of power. *Behavioral Science* **20**: 201–15

Deetz S, Mumby D 1990 Power, discourse and the workplace: Reclaiming the critical traditions. *Communication Yearbook* **13**: 18–47

Dreyfus H L, Rabinow P 1982 *Michel Foucault: Beyond Structuralism and Hermeneutics*. Harvester, Brighton, UK

Fairclough N 1992 *Discourse and Social Change*. Polity Press, Cambridge, UK

Foucault M 1980 *Power/Knowledge: Selected Interviews and Other Writings 1972–1977*. Harvester Press, Brighton, UK

Hardy C 1985 The nature of unobtrusive power. *Journal of Management Studies* **22**(4): 384–99

Hardy C 1994 *Managing Strategic Action: Mobilizing Change*. Sage, London

Hardy C, Clegg S 1999 Some dare call it power. In: Clegg S R, Hardy C (eds.) *Studying Organization: Theory and Method*. Sage, London

Hardy C, Leiba-O'Sullivan S 1998 The power behind empowerment: Implications for research and practice. *Human Relations* **51**(4): 451–83

Hickson D J, Hinings C R A, Lee C A, Schneck R E, Pennings J M 1971 A strategic contingencies theory of intraorganizational power. *Administrative Science Quarterly* **16**(2): 216–29

Knights D, Morgan G 1991 Corporate strategy, organizations, and subjectivity. *Organizational Studies* **12**(3): 251–73

Lukes S 1974 *Power: A Radical View*. Macmillan, London

Palmer I, Hardy C 2000 *Thinking About Management: Implications of Organizational Debates for Practice*. Sage, London

Pettigrew A M 1973 *The Politics of Organizational Decision Making*. Tavistock, London

Pettigrew A M 1979 On studying organizational cultures. *Administrative Science Quarterly* **24**: 570–81

Pfeffer J 1981 *Power in Organizations*. Pitman, Marshfield, MA

Reed M 1998 Organizational analysis as discourse analysis: A critique. In: Grant D, Keenoy T, Oswick C (eds.) *Discourse and Organization*. Sage, London

Weber M 1947 *The Theory of Social and Economic Organization*. Oxford University Press, New York

C. Hardy

Organizations, Metaphors and Paradigms in

This article comprises two sections. The first commences with a definition of metaphor. It then goes on to familiarize the reader with two key areas of debate and discussion concerning metaphor's usage in organization science. One concerns the status accorded to metaphor. The other relates to attempts to place different kinds of metaphor into hierarchical and nonhierarchical typologies. The second section of the article focuses on paradigms. The section commences with a definition of paradigms and a discussion of Kuhn's seminal (1970) work. It then examines how organization theorists have developed Kuhn's work and explores its utility in the context of organization science. In so doing, an important link between metaphor and paradigm is highlighted—one which reiterates and demonstrates the pervasiveness of metaphor within the field of organization science. In short, paradigms related to the study of organizations are shown to be influenced by dominant, but often underlying, metaphors.

1. Defining Metaphor

Metaphors may be seen as the outcome of a cognitive process that is in constant use—a process in which the literal meaning of a phrase or word is applied to a new context in a figurative sense. In this respect metaphors have an important role to play where individuals attempt to make sense of their environment. Metaphors enable the transfer of information about a relatively familiar subject (often referred to as the source or base domain) to a new and relatively unknown subject (often referred to as the target domain). What is more, when people assert that 'subject A is, or is like B, the processes of comparison, substitution, and interaction between the images of A and B act as generators of new meaning' (Morgan 1980, p. 610). Metaphors therefore have a 'generative' quality (Schön 1993). More specifically, and in the context of this article, their generative quality can be used to bring into existence new perspectives on organizations.

2. Metaphor and its Status in Organization Science

The legitimacy and value of metaphor in the social sciences, especially in relation to organization science, have been the subject of considerable debate. The debate centers on the relevance and appropriateness of what a metaphor generates—whether it is anything that actually increases knowledge and understanding of the phenomenon under consideration. Thus, and as Grant and Oswick (1996) have noted, it is not so much the existence of metaphors in organization science that is called into question by the literature, rather it is whether metaphors should be accorded a positive or negative status.

2.1 Positive Status

The suggestion that metaphors ought to be accorded a positive status in relation to organization science is based on the belief that metaphors are—in line with

their generative capacity—'liberating' in orientation. They entail using a combination of both language and thought to construct a nonliteral meaning and apply it to reality in order to shape and enhance appreciation of that reality. Seen in this light metaphors can be viewed as powerful educational devices that are highly correlated with learning. As such, they are important to the advancement of knowledge and understanding.

The liberating orientation of metaphors manifests itself in several ways. Three in particular merit attention. First, the application of a metaphor may lead to a familiar situation being seen in a completely new and highly informative light. It was this that led Tsoukas (1991, p. 566) to note how metaphors generate alternative social realities since they '…encourage different ways of thinking, which enable social scientists and lay people alike to focus upon, explain, and influence different aspects of complex organizational phenomena.'

A second interpretation of the liberating role of metaphors is one that focuses on the value of metaphor where it facilitates the learning of new knowledge. The emphasis here is not about metaphor's role in the reinterpretation of already known phenomena, but on its value where something is encountered that is a completely new experience. As an example Barrett and Cooperrider (1990, p. 223) discuss the case of a science student unable to grasp the structure of the atom. Getting the student to use the metaphor 'the atom is a solar system' portrays the electrons moving in individual orbits about a central nucleus and allows new understanding to emerge.

The third way in which metaphors have been described as liberating follows on from the first two interpretations. Its proponents emphasize that the application of metaphor to either new or existing phenomena is in itself a process of experimentation. In keeping with this line of argument, metaphor can therefore be deliberately used as an investigative tool. Such an approach is highly apparent in numerous analyses of organizations and more specifically the organizational change and development literatures (Barrett and Cooperrider 1990) (see *Organizational Change* and *Development: Organizational*).

2.2 Negative Status

The positive status accorded to metaphors is not one that is universally shared. Two main criticisms are apparent and both suggest that far from being liberating, metaphors can actually act to constrain knowledge. The first of these criticisms is that if science is about exactitude, then it follows that something which is applied in a figurative sense cannot be of any use to scientific investigation. In short, metaphors do not offer testable hypotheses and do not therefore allow 'truths' to be established. Taking this argument to its logical conclusion would mean that if metaphors are of no scientific value then the apparent reliance on metaphor in the organizational literature means that the term organization *science* is a misnomer. Furthermore, the idea that metaphors are simply nonliteral devices that need to be purged from scientific language is in line with the nonconstructivist view of the world. It is a view under which science is the only route to understanding human lived reality.

Tsoukas (1993, p. 326) has argued that metaphor's scientific credentials and therefore its liberating value can be challenged in three further ways. First, metaphors are generally 'imprecise' in that they do not create a theoretical definition of whatever it is that is under scrutiny. Consequently, one cannot tell whether the characteristics of the phenomenon being studied are theoretically relevant or irrelevant. Second, there is no way of measuring the 'goodness of fit' of a metaphor. To work, it relies on recognition of how and why the source domain of the metaphor is useful when applied to the target domain. But quantifying the extent of its success in achieving this is difficult. Moreover, the application of a metaphor is in fact a personalized cognitive process. Thus, a particular metaphor may work well for one person, but not so well, if at all, for another. Third, there is the possibility that a metaphor can be pushed too far, especially where the source concept was originally developed for use in a different scientific field. The borrowing of a scientific concept needs to be 'informed,' but it is argued that most scientists have insufficient knowledge of other disciplines for this to be a realistic proposition. Metaphors are thus put to use with scant regard for their limitations and relevance and in such situations serve no useful scientific purpose.

The second criticism of metaphors is less concerned with their scientific credentials than with their ability to reify and act as ideological distortions (Tinker 1986). A number of metaphors used in organizational analysis might be construed as 'socially partisan' in that they play down structural conflicts and fail to identify inequalities of power. For example, representation of organizations as biological, organismic, or mechanical entities (see *Closed and Open Systems: Organizational*) means that they are expected to behave automatically in ways that are in keeping with such entities and that such behavior is accepted without question. The reality is that organizations are subject to the actions of management and the owners of capital. In short, metaphors create a 'false consciousness,' one that shields organizations from any critical social analysis. For those who subscribe to this view, metaphors will only be of value where they explain and recognize social inequality and domination.

3. Types of Metaphor

A second debate surrounding metaphor concerns types of metaphor (Grant and Oswick 1996). Those involved in the debate believe metaphors can be liberating and accord them positive status, but they

also believe that the extent to which metaphors achieve this is influenced by type. A number of typologies are apparent. In essence they can be broken down into two groups—hierarchical and nonhierarchical.

3.1 Hierarchical Typologies of Metaphor

Hierarchical typologies of metaphors start with those which most influence ways of thinking about and viewing the world, and work down to those which are of minor or peripheral significance. While this is helpful in differentiating each type of metaphor on the basis of its liberating tendencies, it is also a view that has its limitations and needs to be qualified. A hierarchy could imply that it is possible to measure the impact of one type of metaphor against another. This is an impossible task; the cognitive nature of metaphors combined with their reliance on figurative language means their effects are unquantifiable and unmeasurable.

Black (1993) presents a hierarchical typology of metaphors in which metaphors are either 'strong' or 'weak.' A strong metaphor incorporates 'emphasis' and 'resonance.' Emphasis is where the words used to apply the metaphor are so effective that they express the inexpressible and cannot be substituted or varied. Resonance means that once the metaphor is understood it lends itself to further elaboration and proves to have numerous further applications. In contrast, a weak metaphor is neither emphatic nor resonant and 'might be compared to an unfunny joke, or an unilluminating philosophical epigram' (Black 1993, p. 26).

An alternative example of a hierarchical typology is the distinction made between 'surface' and 'deep' metaphors (Schön 1993). A deep metaphor is one which determines the key features of the idea or object being examined. It forms the basis on which subsequent surface level metaphors are formed. An example of a deep metaphor might be the metaphor of the organization as a human entity. The subsequent development of surface metaphors such as organizational intelligence, behavior, and learning (see *Intelligence: Organizational* and *Learning: Organizational*) are all based on this deeper metaphor but in discussing them the word human is not used.

3.2 Nonhierarchical Typologies of Metaphor

Instead of imposing relative values on the different types of metaphor they identify, nonhierarchical typologies of metaphor focus on understanding how each type of metaphor works, and when and where each type is used. Two typologies merit attention. The first concerns dead, live, and dormant metaphors.

Dead metaphors are described as having 'become so familiar and so habitual that we have ceased to be aware of their metaphorical nature and use them as literal terms' (Tsoukas 1991, p. 568). An example of such a metaphor would be the 'teeth' of a saw. On their own, dead metaphors are unable to generate any useful insights about the particular phenomenon they describe. Thus, the dead metaphor of organization (taken from the Greek word *organon*, meaning tool), offers us few insights into understanding organizational phenomena such as training, development, leadership, and motivation. For this to occur we may need to use live metaphors. Live metaphors necessitate creativity on the part of those applying them—i.e., they require interpretation and need to be thought through in order to work. Their appeal is that 'they particularly lend themselves to further conceptual development' (Tsoukas 1991, p. 568). Dormant metaphors, while used as literal terms, are distinguishable from dead metaphors in that they are really semiliteral. At first sight, the metaphorical basis of the words and phrases they encompass is not obvious, but becomes quickly apparent. Consequently, they play a positive role in organization science—examples include organizational behavior or organizational structure. Dormant metaphors may develop into either dead or live metaphors.

A second nonhierarchical typology of metaphors concerns the 'deductive' vs. 'inductive' approach to their application (Palmer and Dunford 1996). The deductive approach involves imposing a metaphor on a particular organizational phenomenon and then seeing if it generates new knowledge, insight, or understanding. The inductive approach seeks to identify the underlying metaphors that are already in use and which influence people's ways of thinking and seeing.

4. Defining Paradigms

The term paradigm is derived from the ancient Greek *paradeigma* and refers to a framework, pattern, or model. Paradigms regularly feature in debates about organization theory. For example, there is a plethora of literature concerned with the idea of organizations moving from one paradigm to another in order to understand and cope better with changes in the business environment.

Although this article is concerned with the application of paradigms to an aspect of the social sciences—specifically organization studies—it is necessary to commence with a discussion of Kuhn's seminal (1970) work on paradigms in relation to the natural sciences. Kuhn was a historian of science. His basic premise was that the role of paradigms in scientific development was as 'universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners' (Kuhn 1970, p. viii). As such, paradigms determine the development of what Kuhn termed

'normal science.' Normal science requires scientists to extend existing knowledge within a known and accepted paradigm. The scientific problems they investigate are perceived as puzzles which are solvable within the pattern of rules and assumptions that exist within the paradigm. If the scientists fail to solve the puzzle, then it is their fault; the fault does not lie within the paradigm itself. Normal science is in contrast to relatively short bursts of 'revolutionary science' where the scientist encounters increasingly perplexing anomalies, 'whose characteristic feature is their stubborn refusal to be assimilated into existing paradigms' (Kuhn 1970, p. 97). This situation gives rise to a new theory or theories and with it a new paradigm such as occurred with the shift from the Ptolemaic system, where the earth was the fixed center of the universe, to that of Copernicus, with the earth and planets moving about the sun (Kuhn 1970, pp. 116–17). The theory underpinning the new paradigm permits predictions that are incompatible with those of its predecessor. In the process of being assimilated into the thinking of the scientific community, the new paradigm displaces its predecessor.

5. *Paradigms and Organization Science*

Kuhn's work on paradigms has been subject to a number of interpretations within the social sciences. In the context of organization science, probably the most authoritative and influential application is that of Burrell and Morgan (1979). These two researchers explore the role of paradigms as views of social reality. For them, social theory in general, and organization theory in particular, is best analyzed in terms of four key paradigms—functionalist, interpretive, radical-humanist, and radical-structuralist—that represent four 'mutually exclusive views of the social world' (Burrell and Morgan 1979, p. viii). When applied to the study of organizations, each paradigm generates theories and perspectives, which are in opposition to those generated in the other paradigms. Each also reflects a group of related schools of thought. Though these schools of thought may take differing approaches and perspectives, they share common assumptions about the nature of society and more specifically organizations.

Burrell and Morgan apply the term 'paradigm' in a far broader sense than Kuhn. In so doing, they argue that social theory can be understood to comprise the four distinct but *coexisting* paradigms they identify. Under these circumstances, the sort of revolutionary shift in knowledge and thinking that Kuhn talked about is unlikely to occur since all possible paradigms are believed to be already in existence. Moreover, researchers within each paradigm are unlikely to be interested in, or have a detailed knowledge of, rival paradigms and instead continue to seek to extend, justify, prove, and defend the assumptions underlying their own. This is a point disputed by, for example, Giddens (1976) who has argued that researchers are aware of other paradigms since in their efforts to understand their own paradigm they must also learn what that paradigm does not encompass. More recently, postmodernist interpretations of paradigms have encouraged the social scientist to move from one paradigm to another, to adopt a multiple paradigm perspective, and not to privilege any one paradigm at the expense of others (see for example Hassard and Parker 1993). Such an approach allows the researcher to understand and appreciate that alternative realities exist concerning particular organizational phenomena. Furthermore, it may also facilitate the generation of a new perspective on the phenomena under consideration.

6. *Paradigms as Metatheories*

Underlying Burrell and Morgan's work is the belief that each of the paradigms they identify is underpinned by 'sets of metatheoretical assumptions about the nature of social science and the nature of society' (Burrell and Morgan 1979, p. viii). In a later work, Morgan (1980, p. 607) examined this issue in more detail suggesting that: 'any metatheoretical paradigm or world view may include different schools of thought, which are often different ways of approaching and studying a shared reality or world view.' He goes on to argue that: 'schools of thought in social science, those communities of theorists subscribing to relatively coherent perspectives, are based upon the acceptance and use of different kinds of metaphor as a foundation for inquiry.' Thus, the various research activities of social scientists are influenced by what can be termed the 'deep' metaphor (see discussion above) underpinning their particular school of thought. Furthermore, any insights into the various phenomena they investigate are generated in accordance with the explicit or implicit application of the deep metaphor to which they subscribe.

Social scientists need to practice a degree of reflexivity so that they are able to recognize that their research activities are linked to favored metaphors and therefore a favored view of reality. As Morgan points out, such recognition enables researchers to appreciate their role in relation to the social construction of social science knowledge especially where it pertains to organizations. Accordingly, he identifies a number of the key organization-targeted metaphors that are utilized by exponents of particular schools of thought within each one of the four social science paradigms (Morgan 1980, pp. 613–19). These may be summarized as follows below.

In the case of the *functionalist* paradigm, society is assumed to have a concrete and systematic existence that structures and regulates human activities and behavior and with it social relationships. Moreover, it

is believed to be possible to construct a social science research methodology that is both objective and value free. The functionalist view of the organization has dominated organization theory for over a century. The various schools of thought that it encompasses have been dominated by two metaphors in particular. The first concerns the metaphor of the organization as a machine and includes theories that portray the organization as 'rational' or 'goal' oriented (see *Rational Choice and Organization Theory*). The second portrays the organization as an organism, a living entity in constant flux and change as it seeks to interact with, and survive within, its environment. The ascription of 'organismic' status to organizations can be seen within the Hawthorne Studies, sociotechnical theories, and structural functionalism (see *Structural Contingency Theory* and *Technology and Organization*). These two metaphors have also been set up as a continuum. Thus, contingency theory seeks to contrast mechanistic and organistic organizations (see *Closed and Open Systems: Organizational*).

In contrast to the functionalist paradigm, the *interpretive* paradigm is founded on the assumption that the social world is based on a social reality that is highly subjective and personalized, so that one person's reality may be different from another. The interpretivist therefore seeks to identify those shared multiple realities that give the social world some kind of underlying uniformity and order. Research methodology is underpinned by a belief that objective social science is unobtainable, and that instead it is based upon concepts and rules that are subjectively determined and followed by researchers. Organizational metaphors underpinning particular schools of thought within this paradigm include those of language games where language becomes seen not simply as communicative or descriptive but as occupying an ontological status that facilitates the construction of the organization as a pattern of social activities. Other metaphors include the metaphor of sensemaking (Weick 1995) where members of organizations construct their realities through *post hoc* rationalization of key events, and the metaphor of identity which emphasizes the importance of identity as a construct of organizational study and examines the ways in which individuals identify with organizations (see *Individual Identities in Organizations*).

The *radical-humanist* paradigm, as with the interpretive, emphasizes that reality is socially created and sustained, but goes on to suggest that human beings are imprisoned within the parameters of the reality that they have constructed for themselves. Psychic and social processes imprison individuals in their reality and alienate them from their true potential. Radical-humanist schools of thought would, for example, see capitalism not as a form of social order (as might functionalists), but as a mode of ideological domination that alienates people from important thoughts and actions. In the context of organizations, radical-humanist schools of thought seek to identify and challenge alienating features of organizational life (see *Organizational Control*). Here, the underlying metaphor is that of the psychic prison whereby members of an organization are viewed as prisoners of an ideologically formulated mode of consciousness. It is a metaphor that underpins the work of key social theorists, notably Marx, and more specifically those addressing fundamental features of organizational life such as power (see *Organizations*: *Authority and Power*).

The *radical-structuralist* view of social reality is, like the radical-humanist, based on the assumption that society is a dominating force. However, the social world is defined as comprising a variety of concrete structures and as being highly materialistic. Various elements of this world are in opposition to one another. The tensions that this creates are held in abeyance by those with power in society, for it is in their interests to keep social order as it is, rather than to allow radical social change. Thus, the radical-structuralist paradigm is based on metaphors that depict organizations as powerful instruments of domination and as part of a wider, seemingly inevitable, process of domination that exists within society overall. These include Michels' (1949) 'iron law of oligarchy' (see *Organizations: Authority and Power* and *Organizational Control*), Weber's (1946) analysis of the 'iron cage' of bureaucracy (see *Bureaucracy and Bureaucratization*) and Marxist examinations of organizational behavior such as those provided by Labor Process theorists. In addition, a number of radical-structuralist metaphors, such as Morgan's (1981) schismatic metaphor, have shown that despite management's best efforts to maintain the prevailing structure, there is a potential for the various contradictions, conflicts, and tensions within organizations to rise to the surface so that organizations implode, fragment, or disintegrate (see *Conflict: Organizational*).

7. *Conclusions: The Value of Metaphors and Paradigms in Organization Science*

As Morgan (1980, p. 612) has noted, 'no one metaphor can capture the total nature of organizational life.' In essence, each metaphor offers a partial understanding of the whole organization. This is not to underestimate the value of metaphors, nor the important metatheoretical role they play within social science paradigms and their application to organization science. Different metaphors can be used to generate powerful and illuminating insights into particular organizational phenomena. Where applied, they often lead to a furtherance of knowledge and an increased understanding about the behavior and attitudes of organizations and their members. Moreover, the application of new metaphors has been shown to generate new ways of thinking about organizations and to overcome

the weaknesses and problems of more established, traditional metaphors. In short, new metaphors often throw up contradictory and thought-provoking analyses of organizational phenomena which cause reconsideration of what has previously been taken for granted. Similarly, literature in the field of organizational change and development (see *Organizational Change* and *Development: Organizational*), along with the management education programs that it informs, often encourages managers to recognize the particular paradigm (sometimes referred to as a frame of reference) and its associated metaphors in which they are operating and to switch to other paradigms and their associated metaphors (see especially Bolman and Deal 1991, Morgan 1997). This is believed to enable managers to appreciate better the complexities of organizations and to provide alternative perspectives and solutions to organizational problems.

See also: Boundaries and New Organization Forms; Design and Form: Organizational; Metaphor and its Role in Social Thought: History of the Concept; Myths and Symbols: Organizational; Organization: Overview; Organizational Climate; Organizational Culture; Organizational Culture, Anthropology of

Bibliography

Barrett F J, Cooperrider D L 1990 Generative metaphor intervention: A new behavioral approach for working with systems divided by conflict and caught in defensive perception. *Journal of Applied Behavioural Science* **23**(4): 219–44

Black M 1993 More about metaphors. In: Ortony A (ed.) *Metaphor and Thought*. University of Chicago Press, Chicago

Bolman L, Deal T 1991 *Reframing Organizations: Artistry, Choice and Leadership*. Jossey Bass, San Francisco

Burrell G, Morgan G 1979 *Sociological Paradigms and Organisational Analysis*. Gower, Aldershot, UK

Giddens A 1976 *New Rules of Sociological Method*. Hutchinson, London

Grant D, Oswick C 1996 Getting the measure of metaphor. In: Grant D, Oswick C (eds.) *Metaphor and Organizations*. Sage, London

Hassard J, Parker M (eds.) 1993 *Postmodernism and Organizations*. Sage, London

Kuhn T 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

Michels R 1949 *Political Parties*. Free Press, Glencoe, IL

Morgan G 1980 Paradigms, metaphors, and puzzle solving in organization theory. *Administrative Science Quarterly* **25**: 605–22

Morgan G 1981 The schismatic metaphor and its implications for organizational analysis. *Organization Studies* **2**(1): 23–44

Morgan G 1983 More on metaphor: Why we cannot control tropes in administrative science. *Administrative Science Quarterly* **28**: 601–7

Morgan G 1997 *Images of Organization*, 2nd edn. Sage Publications, Beverley Hills, CA

Palmer I, Dunford R 1996 Understanding organisations through metaphor. In: Oswick C, Grant D (eds.) *Organisation Development: Metaphorical Explorations*. Pitman Publishing, London

Schön D 1993 Generative metaphor: A perspective on problem setting in social policy. In: Ortony A (ed.) *Metaphor and Thought*, 2nd edn. Cambridge University Press, Cambridge, UK

Tinker T 1986 Metaphor or reification: Are radical humanists really libertarian anarchists? *Journal of Management Studies* **23**: 363–84

Tsoukas H 1991 The missing link: A transformational view of metaphors in organizational science. *Academy of Management Review* **16**(3): 566–85

Tsoukas H 1993 Analogical reasoning and knowledge generation in organization theory. *Organization Studies* **14**: 323–46

Weick K E 1995 *Sensemaking in Organizations*. Sage, Thousand Oaks, CA

D. Grant

Organizations: Negotiated Orders

The negotiated order of an organization is the pattern of activities which has arisen or emerged over time as an outcome of the interplay of the variety of interests, understandings, reactions, and initiatives of the individuals and groups involved in the organization. To examine negotiated orders in any given organization is to turn away from the more traditional way of looking at organizations which give primacy of attention to the pattern or ordering of activities chosen (or 'designed') by those officially in charge of the organization. Instead, the influences of people other than administrative designers on structures and patterns are recognized. The influence and power of some individuals or groups will be greater than that of others, but the ordering of activities which arises in practice is always seen as resulting from the contributions of the plurality of parties to the organization as a whole.

1. Origins of the Concept

The concept of 'negotiated order' first appeared in a study of a psychiatric hospital (Strauss et al. 1963). At that stage, the researchers did not claim a generalized relevance for the concept. However, the leading researcher, Strauss (1978), later showed how it could be used more generally to understand organizational processes, including ones in industrial organizations, and its insights have subsequently been incorporated into a broad stream of thinking about organizations which focuses on processual aspects of organizational life. The approach emerged from the symbolic interactionist tradition of social psychology, a perspective that stresses the two-way relationship between human beings and their social context. People construct their realities through processes of interaction with others,

and their identities emerge as part of this process of interacting and reality making. Such a theoretical approach, with its strongly antideterminist orientation, inevitably leads to a questioning of views of organizations that treat the formal arrangement of roles, rules, and procedures as the major determinants of the activities of organization participants. Indeed, Strauss et al. explicitly identify their study with such a critical stance on orthodoxy: 'Students of formal organization tend to underplay the processes of internal change as well as overestimate the more stable features of organizations—including its rules and its hierarchical statuses' (1963).

2. *The Negotiated Order of the Hospital*

The study of the psychiatric hospital reported by Strauss et al. (1963) focuses on the interactions of the variety of individuals and groups who are involved in its activities. These include the doctors, nurses, patients, patients' families, social workers, nutritionists, administrative workers, and so on. Although the rules, policies, official procedures, and hierarchical statuses of the hospital play a part in the patterns which emerge from the interactions which take place, the patterns are shown to be emergent. They emerge from the processes whereby different parties make use of rules, procedures, and information and negotiate between them what actually happens. These negotiations are characterized in a variety of ways, as agreements, contracts, understandings, pacts, and so on. But two things are emphasized throughout. First, it is stressed that these negotiated outcomes are always 'temporal' and have to be made and remade over time. A head nurse, for example, may gain agreement to move a patient to another ward, after it is agreed that the first ward has 'tried for a little longer' with that patient. In the light of how that patient fares, however, the patient may or may not subsequently be moved. The 'deal' will have been remade, implicitly or explicitly. Second, it is stressed that there is a patterning to what might, at first sight, appear to be a series of unique encounters. The deal between the head nurse and the doctor over a particular patient in specific circumstances, for example, takes place in the light of how interactions between members of these two occupational groups generally tend to occur across the hospital.

Out of the ongoing multiplicity of interactions and negotiations a pattern of hospital life, a 'negotiated order,' emerges. Order is thus not imposed on social actors by 'the hospital' but the 'order' which, in effect, is 'the hospital' is constructed socially by the interactions of the various individuals and groups. And each of these brings to the negotiation of order their own interests, understandings, loyalties, knowledge, occupational standing, and social origins—all these elements themselves being liable to change as they are taken into processes of reality negotiation.

3. *Parallel Early Studies*

This broad approach of focusing on interactional processes in organizations was not peculiar to this group of social scientists studying one particular mental hospital. However, the Strauss et al. study, with its concept of 'negotiated order,' has produced a valuable analytical 'handle' to attach to such processes. Goffman's classic study, *Asylums* (1961), with its own interactionist framework, has many parallels with the Strauss et al. study (which was partly influenced by early accounts of Goffman's research). And the same argument can be made with regard to Dalton's participant observation study of managerial work in an industrial setting (Dalton 1959). In fact, it was this study that Strauss (1978) used to illustrate the position he later came to take: that the negotiated order concept had a wider currency than had earlier been implied. He notes that Dalton's detailing of what Strauss calls 'sub-processes of negotiation' is especially rich. Many researchers in managerial behavior would probably agree that few studies have surpassed Dalton's analysis of how managers engage in, as Strauss puts it, 'trading off, the paying off of accumulated obligations, the covert making of deals and other kinds of secret bargains, the use of additional covering negotiations for keeping hidden the results of previous covert arrangements, the bypassing of negotiations, and the mediating of negotiations; also, a very complex balancing of accumulated favors and obligations, along with the juggling of commitments within the negotiations itself' (1978).

4. *Negotiated Order: Rules and Ethnomethodology*

One particular area where Strauss et al. (1963) follow Dalton (1959) is in attending to the frequency with which formal rules are broken, stretched, or ignored, not only to suit sectional purposes, but also to 'get things done.' A similar insight is central to ethnomethodological writing on organizations. This contributes to an understanding of how 'orders' are produced in organizations by focusing closely on how individuals, in their organizational interactions, 'work,' especially through talk, to present what they are doing as 'sensible' because it accords with rules (Bittner 1965, Silverman and Jones 1974). The rule (or even 'the organization' itself) is not something preexisting or existing outside ongoing organizational interactions. It is, rather, a resource that people draw upon to legitimize their actions and further whatever projects they are pursuing in the work context.

This perspective can be illustrated with a small case study. A van driver was overheard explaining to a factory security officer who had stopped him why he and his colleagues regularly took an illegal 'short cut' across a large industrial plant. He said, 'The company says that all employees should use their initiative,

always thinking of the customer first. By driving this way we are getting the goods out to the customer as quickly as we can. The other security men know this, and don't stop us.' Ethnomethodological analysis would point to the way this driver uses a notion of 'the organization' ('the company') as one resource and utilizes a 'rule' ('use initiative and put customers first') as another. And this case equally well illustrates the broad notion of 'negotiated order.' Informal negotiations between van drivers and security officers which allowed the driving along an illegal route through the plant to become part of the 'order' of the organization had occurred. It suited both parties (a 'quiet life' for the security people and a saving of time and effort for the drivers) and it 'got the job done.' However, we must remember that negotiated order thinking sees such deals as continually liable to change. In fact such a change was happening on the day these events were observed. A new security man had started work. And he believed that the one-way rule (which made the short-cut 'illegal') was a necessary one for protecting the safety of employees working in the plant. He too utilized a notion of 'the organization' and drew on a 'rule' when he said, 'You should realize that the company puts employee safety before even customers. Don't drive down here again.' He got his way—for the time being.

5. *Ambiguity, Interpretation, and Enactment*

It might be argued that this case is not looking at 'rules' in a proper sense, and that the individuals in the case study were talking simply in terms of inferred rules. But organizational thinking that follows a broad negotiated order approach would accept that all rules, or any other organizational procedure for that matter, are ambiguous. An important work in this tradition by March and Olsen (1976) argues that all organizational choices occur in a context where there is ambiguity about everything, from organizational goals and rules, through to the technologies used and people's knowledge of the past. Weick (1979), in another important contribution to this broad style of thinking, similarly points to *equivocality* in every aspect of the work of 'organizing.' Members of organizations create the organization (or, rather, achieve 'organizing') through continuous processes of interpretation and action based on interpretations: they *enact* both the organization and the 'environment' to which it relates.

Out of a range of partly linked theoretical origins—symbolic interactionism, 'phenomenology' and Weberian interpretative sociology—a general style of organizational analysis has developed which, in one way or another, incorporates the 'negotiated order' insight. It is a processual style of analysis: it treats organizational patterns as emergent rather than as the creations of their managerial designers. And such an emphasis has been reinforced by the influence on organizational theorizing of postmodern thinkers. Chia, for example, adopting an 'ontology of becoming' encourages the analysis of the 'discursive micro-practices which generate (the) stabilized effects' which constitute 'organizing' (1996).

The original way in which the negotiated order concept was used was criticized, even by broadly sympathetic critics like Day and Day (1977), for insufficiently grounding its analysis of interactional processes in their wider political, structural and historical context. Strauss' later work responded positively to some of these points and, in particular, demonstrated how the approach could handle such allegedly neglected matters as conflicts which are 'endemic, or essential to relations between or among negotiating parties' (1978). However, a serious reservation has to remain with regard to what the analysis of negotiated orders within organizations can add to organizational theorizing generally if it focuses primarily on interactional patterns within organizations. It is important to relate the *strategic exchanges* which occur within organizations to the wider patterns of exchange between organizations and how these exchanges relate to wider patterns of culture and political economy (Watson 1994). The negotiated order perspective has had less impact than it perhaps deserved because it was deemed inadequate as a general perspective for the study of organizations and their management. However, if the significant insights which it offers are incorporated into broader perspectives which link activities at the intra-organizational interactional level to wider societal patterns and processes, then it has considerable and continuing potential as a device for understanding organizational processes.

See also: Ethnomethodology: General; Group Processes in Organizations; Human Resource Management, Psychology of; Industrial Sociology; Learning: Organizational; Organization: Informal; Organizational Culture; Organizational Culture, Anthropology of; Organizational Decision Making; Organizations, Sociology of

Bibliography

Bittner E 1965 The concept of organization. *Sociology Research* **32**: 239–55

Chia R C H 1996 *Organizational Analysis as Deconstructive Practice*. de Gruyter, Berlin

Dalton M 1959 *Men Who Massage*. Wiley, New York

Day R A, Day J V 1977 A review of the current state of negotiated order theory: An appreciation and a critique. *Sociological Quarterly* **18**: 126–42

Goffman E 1961 *Asylums*. Anchor Books, Garden City, NY

March J G, Olsen J P 1976 *Ambiguity and Choice in Organizations*. Universitetsforlagtt, Bergen, Norway

Silverman D, Jones J 1976 *Organizational Work: The Language of Grading, and the Grading of Language*. Macmillan, London

Strauss A 1978 *Negotiations*. Jossey Bass, San Francisco
Strauss A, Schatzman L, Ehrlich D, Bucher R, Sabsin M 1963 The hospital and its negotiated order. In: Friedson E (eds.) *The Hospital in Modern Society*. Free Press of Glencoe, New York
Watson T J 1994 *In Search of Management*. Routledge, London
Weick K E 1979 *The Social Psychology of Organizing*, 2nd edn. Addison-Wesley, Reading, MA

T. J. Watson

Organizations, Sociology of

The sociology of organizations, like its larger sister the theory of organizations, is not a unified field of research. In its already long history, it has seen many different approaches which have left a more or less profound imprint on our way of thinking about organizations. It is beyond the scope of this article to attempt a complete account of this diversity.

However, taking a distant view, one can distinguish two sociological templates which structure around two understandings of the word organization. The first sees organizations as structural forms, the nature, characteristics, and dynamics of which have to be explained, whereas the second insists on the processes of structuring, i.e., organizing collective human action. Within certain limits, the *explanandum* is of course similar. For both, it is the structure and the functioning of organizations but they have different starting points, take different routes, and put the emphasis differently. Thus, the first starts with organizations and will focus on the variation of their forms: as a consequence, organizations are its basic units of analysis and it will analyze the social dynamics on the inter-organizational, sectoral, or societal level in order to explain organizational form. The second, on the other hand, will start with social action and will consider the sheer existence and maintenance of organizations as the basic problem to be addressed: as a consequence, it will concentrate its analysis on interaction processes within organizations, will take individuals (their decisions and their actions) as its basic unit of analysis, and will stress the differences between individual organizations, even if they belong to the same organizational field.

Both strands have their intellectual roots in the seminal work of Max Weber. In his *Wirtschaft und Gesellschaft* (1964), Weber develops a theory of the forms of legitimate domination in society (traditional, charismatic, and rational–legal forms of domination) which roughly define the bases of the legitimacy of power and influence in a given society and condition the types of organizations to be found within each of these regimes of domination. He then argues that bureaucracy is the organizational form developing within the framework of the rational–legal form of domination and proceeds to give an ideal–typical description of bureaucracy, i.e., a highly formalized and hierarchical form of organization characterized by the predominance of written rules and procedures, formalized tasks, precise definitions of competence, clear lines of subordination, the explicit separation of ownership and management responsibility, and the merit principle as the only legitimate route of access to the different functions in the organization. Weber sees unifying dynamics at work, pushing towards the development and generalization of bureaucratic organizations as they are both legitimate and more efficient. He sees the roots of this greater efficiency essentially in the fact that the arbitrary imposition of power and the resulting interpersonal conflicts are limited by written rules and procedures which allow for more rational, foreseeable, and standardized execution of tasks.

This analysis includes all the themes which organizational sociology has developed and dwelled on through all these years. It also contains the roots for the divide which still structures two traditions in organizational sociology today. One is concerned with the study of organizational forms as they are shaped by the constraints of efficiency (this is the main concern of what has become known as 'structural contingency theory') or legitimacy (this is the theme of the 'neo-institutional school' in organizational sociology). The other tradition emphasizes the study of interaction and decision-making processes which 'produce' as well as 'reproduce' regularities and structure, i.e., organization, and which first developed out of the empirical analysis of public and private bureaucracies.

1. From the Dysfunctions of Bureaucracy to Organizations as Constructs of Collective Action

In the 1940s and 1950s, bureaucracy was a central theme of empirical study of organizations, especially in the United States. This has certainly to do with the proliferation of huge administrative bureaucracies and the seemingly unlimited growth of the large corporation which was looked upon with both fascination, and fear. Following Weber's analyses, this movement was interpreted as the proof of the greater efficiency made possible by the standardization, formalization, and depersonalisation characteristic of this form of organization. However, following R. Michels' (1914) iron law of oligarchy, this secular movement raised fears because the conditioning powers of these huge organizations and the oligarchic and technocratic tendencies inherent in their functioning were seen as a threat to democracy and the ideals of reform: organizations as means for collective action in favor of reform set free forces which were in contradiction with these ideals.

The main interest of the early studies of bureaucracy (Merton 1940, Selznick 1949, Blau 1955, Gouldner 1954, 1955, Crozier 1961, 1964) was to have considered bureaucracies as a complex social system which had to be studied empirically in order to reconstruct and understand its informal structures and dynamics. Applying the lessons and methods from industrial sociology which developed out of the Hawthorne experiments and the human relations movement sparked by Roethlisberger's and Mayo's account of them, these studies produced results which were in contradiction to both the extrapolations and fears mentioned above. They showed that bureaucracies not only were not always efficient, but also produced informal behavior and dynamics which were often dysfunctional, i.e., detrimental to the attainment of their goals and which in any case created a greater diversity in the ways organizations function than would have been expected on theoretical grounds. Thus, Merton shows that the impersonal rules on which bureaucracies are built induce rigid and ritualistic behavior on the part of their members which are dysfunctional for their capacity to respond to the needs of their clients. Blau demonstrates that the efficiency of bureaucracies is not the product of their formal characteristics but bases his argument on the fact that their members take it upon themselves to break the rules in order to enhance the quality and efficiency of their work.

These studies, however, go beyond merely showing the inefficiencies or deviations of bureaucracies. Following Robert Merton's analysis (1936) of the unanticipated consequences of purposive social action and of the latent functions of social structures, there are two complementary explanations not only for the functioning, but also for the emergence of bureaucracy.

On the one hand, these studies show that bureaucracy is the place (and also the product) of 'vicious circles of bureaucratization,' as the attempts at countering unanticipated and dysfunctional consequences of bureaucratic structures lead to an accentuation of the very characteristics which have generated these dysfunctions. This is, for instance, the mechanism analyzed by Gouldner in his study of a gypsum mine: the formal rules instituted to strengthen control over workers and to fight against their low work morale instead increase the latter's apathy, which management tries to counter by strengthening the surveillance of the workers and an increase in the rules on which this surveillance is based. This attempt, in turn, further increases workers' apathy and passivity, and so on. Crozier's analysis (1964) of the vicious circle generated by impersonal rules and the resulting parallel power relations follows a similar pattern. He shows how the proliferation of impersonal rules in view of eliminating the uncertainties which condition the satisfactory functioning of an organization increases the power of those whose contribution is needed in order to cope with the remaining sources of uncertainty. The interpersonal power and dependence relations which are generated around these uncertainties and the (unofficial and illegitimate) privileges they produce, will justify fears of personal power and dependence relations and therefore increase the pressure for further eliminating remaining uncertainties through centralization and new impersonal rules, which in turn will generate new parallel power relations, and so forth.

On the other hand, these studies explain the bureaucratization of organizations by the latent functions it fulfills for their members. Gouldner (1954) distinguishes five latent functions of bureaucracy: impersonal rules allow for distant control, thus creating a filter and a protection by reducing interpersonal relations; they legitimize sanctions but they also restrict the freedom of behavior of the members of the hierarchy by codifying the possibilities of sanction; they make possible apathy, i.e., behavior which restricts itself to the strict application of the rules; and therefore they are a resource for bargaining with members of the hierarchy if they need some extra commitment, which they always do. Moreover, Gouldner shows that bureaucratization is an answer for the problem of succession in organizations, in so far as impersonal rules can at least partially be a substitute for the personal legitimacy which the outgoing managers enjoyed, but which the incoming managers lack. Crozier's analysis (1964) follows a similar pattern but completes Gouldner's interpretations with a broader theory: the bureaucratic mode of organization based on the proliferation of written rules and procedures is a way to avoid face-to-face relations and to escape the arbitrariness and uncertainties of direct power and dependence relations. Its latent function is a general one and relates to the necessity of taming and structuring the power and dependence relations which are at the heart of collective action and without which no cooperative endeavor is feasible.

The essential merit of these seminal studies is to enable us to go beyond the sterile opposition inherited from the human relations tradition between, on the one hand, formal structure understood to be the incarnation of rationality as well as efficiency and, on the other, the actual behavior of its members which result from their irrationality, their affectivity, and their conditioning by prior processes of socialization. In their theorizing, the rationality of formal structure is no more opposed to the affectivity of human behavior. On the contrary, they interpret these structures as an answer to the cognitive limits and to the constraints stemming from human affectivity. The structure and the goals of an organization can thus no more be understood as the expression of a logic which could be independent from, and superior to, the relations between the members of an organization. They are created by these relations and draw their

significance and their justification from them. In short, they are no longer an exogenous variable. They have been put back into the dealings between the members of an organization and become the endogenous result of them: they cannot be understood independently from the interaction and bargaining processes for which they constitute only the framework. They therefore cannot escape the limits of rationality characterizing the human behavior which produces them, and have to be analyzed together.

One can clearly see the thrust of this line of reasoning. Drawing on these studies of bureaucracy as well as on the seminal work of the Carnegie group around H. Simon on a behavioral theory of decision making in organizations (March and Simon 1958, Cyert and March 1963) and on the analyses of collective action and public or private decision making in the United States and in Europe (see Allison 1971, Crozier and Friedberg 1977, Grémion 1976, Hirschmann 1967, 1970, Lindblom 1959, Olson 1965, Schelling 1961, 1978), it points towards a behavioral conceptualization of organizations which links organizational behavior to the cognitive and relational capacities of their members. An organization, in this view, can be understood as an arena (Cyert and March 1963) or as game structures (Crozier and Friedberg 1977, 1995) where participants are free to choose their behavior, but within limits. The range of their choices is restricted, because if they do not want to lose in the transactions with the other participants, they have to take into account the rules of the game which prevail and which determine the value of the resources of each participant and the appropriate ways of using them in transactions with other participants. But within these limits, which may be variable and leave more or less leeway, they can and will actually choose their behavior which can thus be understood as the way in which they adjust to the constraints of their situation as they see it, while simultaneously trying to further their interests (in whatever way they may define them) by using the resources and opportunities they perceive. Their behavior therefore cannot be considered as completely hazardous: it is the expression of a decision which makes sense to the person who is choosing and therefore is rational in the sense of H. Simon's bounded rationality (Simon 1957, March and Simon 1958). It is a reasonable adjustment by any participant to his or her situation; i.e., the network of interdependencies within which each participant is located, an adjustment which is reasonable within the limits of his or her perception of the opportunities and constraints contained in this situation and of his or her capacity to make use of these opportunities and constraints.

Such an approach to organizations, which emphasizes the constraints on organization resulting from the limited (but extendable) cognitive and relational capacities of human beings, naturally transcends the usual distinctions between different forms of organizations (hospitals, firms, administration, etc.), and its heuristic value clearly is not limited to the analysis of formal organizations. Its target is, in fact, a much broader issue which all kinds of organizations and all forms of collective action have to solve; i.e., the problem of cooperation and coordination between actors pursuing and continuing to pursue divergent interests. It is, in a way, centered around a theory of the organizational phenomenon which aims to understand how participants who continue to pursue divergent interests can nonetheless organize or accept to be organized in the pursuit of collective goals. Such an approach radically banalizes formal organizations which, in this view, are only one of many possible forms of contexts of action the characteristics of which constrain the collective action of the various participants (Friedberg 1993). Organizations thus become the artificial device which help analyze and understand the general problem of human cooperation and coordination. Organizational sociology becomes a way of theorizing about collective social action.

2. *Efficiency and Legitimacy as Unifying Forces Shaping Organizational Forms*

The other tradition, to which we will now turn, has a different starting point. It sees organizations as structural forms, the nature, characteristics, and dynamics of which have to be explained. Thus, it starts with organizations and will focus on the variation of their forms: organizations are its basic units of analysis and it will try and analyze the social dynamics on the inter-organizational, sectoral, or societal level in order to explain organizational form. These dynamics are traced to two distinct constraints: efficiency and legitimacy, each of which is stressed by a different strand of analysis.

Efficiency is the constraint stressed by structural contingency theory. This paradigm emerged in the middle of the 1960s and developed as a critical reaction to the theoretical and methodological perspective characteristic of organizational sociology of the 1950s. With regard to methodology, the then dominant qualitative case-study method was criticized because it provided merely a thick description but no grounds for generalizations or for the construction of a general theory of organizations. With regard to theory, it was argued that the over-emphasis on motivations and human relations characteristic of organizational thinking so far had had two consequences detrimental to our understanding of organizations: the role of structure and its influence on these relations had been downplayed, while, and this was seen as even more important, the context of an organization and the way its characteristics condition an organization's structure and functioning had been largely ignored.

As a consequence, structural contingency theory set out on a different program. Its focus was not on action

or behavior within organizations, but on organizations as structured entities whose characteristics and change over time have to be explained using quantitative methods for the statistical study of samples of organizations in order to list, describe, and, if possible, measure the influence which the main dimensions of an organization's context exert on its structures, its functioning, and its performance. In other words, this paradigm was concerned with two main questions: which dimensions of context affect an organization's (mainly structural) characteristics and to what extent? what is the influence of each of these characteristics on the performance of an organization?

Structural contingency theory has been the dominant paradigm in the field of organization studies from the middle of the 1960s up to the first half of the 1980s, especially in the Anglo-Saxon world. It has generated an immense effort to determine and measure the impact of the various dimensions of context. Let us mention a few particularly significant research programs. The influence of technology on the structure of organizations has been explored by J. Woodward (1958, 1965, 1970) and C. Perrow (1967, 1970); the Aston group around D. Pugh (1963) and D. Hickson et al. (1969), as well as P. Blau in the United States (Blau and Schoenherr, 1971) have explored the link in particular between size (among other, less central variables) and organizational structure. The impact of the technical, economic, and social characteristics of an organization's environment on its structure and mode of functioning have been independently studied in the seminal work of Burns and Stalker in England (1961) and Lawrence and Lorsch in the Unites States (1967). Last, but not least, the more conceptual work of J. Thompson (1967) has also been very influential, especially in regard to his conceptualization of what he called the 'task environment' of an organization. This approach is developed further by the population ecology of organizations, which builds on the seminal work of Aldrich (1979) as well as Hannan and Freeman (1977) and which aims at studying the contextual conditions which explain the emergence, the diffusion, and the disappearance of populations of organizations which share the same characteristics and which fit certain contextual conditions or ecological niches.

The main contribution of this quantitative and apparently more scientific approach to the study of organizations has been to demonstrate empirically the impossibility to define a single best way for structuring an organization. The good, i.e., the efficient structure, cannot be defined in general and beforehand. It is a function of the context and can only be defined after the different dimensions of this context have been recognized and taken into account in the organizational design. However, this very important contribution should not cover up the theoretical and empirical shortcomings of the approach. Indeed, according to the reasoning on which it is based, context becomes a constraint because organizations are viewed as driven by the constraints of efficiency. Indeed, so this reasoning goes, organizations have to adjust to their contexts because their performance depends on this fit: in order to survive, they have to be efficient and, in order to be efficient, they have to adjust to the demands of their context. Although there is certainly some truth in this hypothesis, the empirical diversity of organizations with similar contexts has shown that structural contingency theory has vastly overestimated the unifying power of the constraint of efficiency. It has enlarged our understanding of the forces which shape organizations but in the process has overstated its case and has been proven wrong by empirical analysis.

Against this reductionism, which pretends to analyze organizations from a purely technical or economic viewpoint (the pressure of the constraint of efficiency), the neo-institutional school in organizational analysis has promoted a more sociological perspective. It emphasizes the symbolic and normative dimensions of action in and between organizations, and stresses the role of their culture, i.e., a set of cognitive and normative frames, to explain their mode of functioning. In other words, it puts forward a less intentional and rational perspective on organizations. In this view, organizations are neither the simple tools of their masters nor machines to maximize efficiency. They are also institutions, i.e., social worlds with their specific identity and culture, which, once created, take on a life of their own and develop their own ends which can never be reduced to mere considerations of efficiency.

Sociological neo-institutionalism builds on, and tries to integrate, several theoretical perspectives: the work of P. Selznick (1943, 1949) on organizations as institutions, the work of H. Simon and his group at Carnegie on bounded rationality and cognitive frames, and the work of Berger and Luckmann on the processes of institutionalization understood as processes of the social construction of reality. In organizational analysis, it brings together the work of American and European sociologists, such as N. Brunsson, P. DiMaggio, N. Fligstein, J. G. March, J. Meyer, J. P. Olsen, W. W. Powell, B. Rowan, W. R. Scott, L. G. Zucker, whose analysis of organizations is based on roughly three main premises.

First, an organization is an institution because it is structured by a set of cognitive, normative, and symbolic frames, which shape the behavior of its members by providing them with the tools necessary to observe and perceive the world around them, to interpret and understand their counterparts' behavior, and to construct their own interests as well as possible ways to further them. Through their structures—formal (organizational forms, procedures, institutional symbols) as well as informal (myths, rituals, social norms)—organizations shape the perceptions, calculations, reasoning, interpretations, and actions of their members by defining acceptable and legitimate

behavior, i.e., behavior which is appropriate in the context of its culture.

Second, no organization exists independently of other organizations which share the same characteristics and which together form an organizational field: e.g., the organizational field of universities, hospitals, schools, airlines, museums, etc. Such organizational fields have their producers and consumers, cognitive and normative frames, their power structures, control mechanisms. In short, they have their own institutional structure and their own dynamics which are brought about by competition as well as interdependence between the constitutive organizations, processes of professionalization, i.e., the establishment of cognitive and normative frames for the field, by government intervention and the like. The dynamics of these organizational fields exert unifying pressures on the individual organizations which, in order to enhance their legitimacy, tend to adopt similar, if not identical, institutional forms and procedures.

Third, while recognizing the importance of the technical and economic environment which was stressed by the structural contingency theory, the neo-institutionalist perspective is interested mainly in the influence of the societal and institutional environment. The institutional environment concerns the characteristics of the organizational field of which an organization is a part, and of the rules it has to follow if it wants to obtain resources from the field and strengthen its legitimacy in it. The societal environment designates the norms and values of modern societies which, according to DiMaggio and Powell (1983) or Meyer (Meyer and Scott 1994), are characterized by processes of rationalization (to a certain extent, this can be understood as the re-edition of Max Weber's process of the disenchantment of the world) and the diffusion of standardized norms as a consequence of increased intervention by states, professions, science, and organizational fields. Both institutional and societal environments constitute a sort of unifying matrix for organizations which have to conform to their pressures if they want to be accepted and thus able to draw resources for their functioning.

In short, the neo-institutional perspective stresses the constraint of legitimacy, as opposed to the constraint of efficiency. Rational structures (formal organizations) do not dominate the modern world because they are efficient. They adopt rationalized institutional norms because these will enable them to obtain the resources necessary for their success and their survival as they increase their legitimacy in a wider cultural environment (rationalist western society and culture).

Taken together, the two strands of reasoning mentioned in the opening remarks provide a complete panorama of the forces shaping organizations: efficiency and legitimacy are the constraints, within which the games being played in and between organizations are embedded, but the games that depend upon the cognitive and relational capacities of the individuals playing them are in turn mediating these constraints. The unifying forces stressed by contingency theory and the neo-institutionalist perspective should therefore never be overestimated: they are themselves subject to differentiating pressures stemming from the cognitive and relational capacities of the humans who play games in contexts structured by constraints of efficiency and legitimacy. These relational and cognitive capacities are never determined and are never final: they are the motor for the infinite variance which is observable in organizational life, they are the motor of innovation which succeeds in destabilizing even the best established technical or institutional environments.

See also: Action, Collective; Bureaucracy and Bureaucratization; Bureaucracy, Sociology of; Bureaucratization and Bureaucracy, History of; Institutionalism; Legitimacy; Oligarchy (Iron Law); Organization: Overview; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture; Organizational Culture, Anthropology of; Organizational Decision Making; Structural Contingency Theory; Weber, Max (1864–1920)

Bibliography

Aldrich H 1979 *Organizations and Environments*. Prentice-Hall, Englewood Cliffs, NJ

Allison G T 1971 *The Essence of Decision: Explaining the Cuban Missile Crisis*. Little Brown, Boston

Blau P M 1955 *The Dynamics of Bureaucracy*. Chicago University Press, Chicago

Blau P M, Schoenherr R 1971 *The Structure of Organizations*. Basic Books, New York

Brunsson N, Olsen J P 1993 *The Reforming Organization*. Routledge, London

Burns T, Stalker G M 1961 *The Management of Innovation*. Tavistock, London

Crozier M 1961 De la bureaucratie comme système d'organisation. *Archives Européennes de Sociologie* **2**: 18–52

Crozier M 1964 *The Bureaucratic Phenomenon*. Chicago University Press, Chicago

Crozier M, Friedberg E 1977 *L'Acteur et le système*. Editions du Seuil, Paris [trans. 1981 *Actors and Systems*. Chicago University Press, Chicago]

Crozier M, Friedberg E 1995 Organizations and collective action: our contribution to organizational analysis. In: Bacharach S B, Gagliardi P, Mundell B (eds.) *Studies of Organizations in the European Tradition*, Research in the Sociology of Organizations, Jay Press. Vol. 13, pp. 71–93

Cyert R, March J G 1963 *A Behavioral Theory of the Firm*. Prentice Hall, Englewood Cliffs, NJ

DiMaggio P J, Powell W W 1983 The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. *American Sociological Review* **48**: 147–60

Friedberg E 1993 *Le Pouvoir et la règle*. Editions du Seuil, Paris [trans. 1997 *Local Orders*, Jay Press]

Gouldner A W 1954 *Patterns of Industrial Bureaucracy*. The Free Press, Glencoe, IL

Gouldner A W 1955 *Wild Cat Strike*. Free Press, Glencoe, IL
Grémion P 1976 *Le Pouvoir périphérique*. Editions du Seuil, Paris
Hannan M T, Freeman J H 1977 The population ecology of organizations. *American Journal of Sociology* **82**: 929–64
Hickson D, Pugh D S, Pheysey D C 1969 Operations technology and organization structure: An empirical reappraisal. *Administrative Science Quarterly* **16**: 378–97
Hirschman A O 1967 *Development Projects Observed*. The Brookings Institution, Washington, DC
Hirschman A O 1970 *Exit, Voice and Loyalty: Response to Decline in Firms, Organizations and States*. Harvard University Press, Cambridge, MA
Lawrence P R, Lorsch J W 1967 *Organization and Environment: Managing Differentiation and Integration*. Harvard Business School, Cambridge, MA
Lindblom C 1959 The science of muddling through. *Public Administration Review* **19**: 79–88
March J G 1988 *Decisions and Organizations*. Blackwell, New York
March J G, Olsen J P 1984 The new institutionalism: organizational factors in political life. *American Political Science Review* **78**: 734–49
March J G, Olsen J P 1989 *Rediscovering Institutions: The Organizational Basis of Politics*. Free Press, New York
March J G, Simon H A 1958 *Organizations*. Wiley, New York
Merton R K 1936 The unanticipated consequences of purposive social action. *American Sociological Review* **1**: 894–904
Merton R K 1940 Bureaucratic structure and personality. *Social Forces* **18**: 560–8
Meyer J W, Rowan B 1977 Institutionalized organizations: Formal structures as myth and ceremony. *American Journal of Sociology* **83**: 340–63
Meyer J W, Scott R W (eds.) 1994 *Institutional Environments and Organizations: Structural Complexity and Individualism*. Sage, Thousand Oaks, CA
Michels R 1913 *Les Partis politiques*. Giard et Brière. Flammarion, Paris 1971 et 1978
Olson M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
Perrow C 1967 A framework for the comparative analysis of organizations. *American Sociological Review* **32**: 194–208
Perrow C 1970 *Organizational Analysis: A Sociological View*. Tavistock, London
Powell W W, DiMaggio P J (eds.) 1991 *The New Institutionalism in Organizational Analysis*. Chicago University Press, Chicago
Pugh D S 1963 A conceptual scheme for organizational analysis. *Administrative Science Quarterly* **8**: 289–315
Schelling T C 1961 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA
Schelling T C 1978 *Micromotives and Macrobehavior*. Norton, New York
Scott R W 1987 The adolescence of institutional theory. *Administrative Science Quarterly* **32**: 93–111
Selznick P 1943 An approach to a theory of bureaucracy. *American Sociological Review* **8**: 47–54
Selznick P 1949 *TVA and the Grass-roots*. University of California Press, Berkeley, CA
Simon H A 1957 *Administrative Behavior*, 2nd edn. Macmillan, New York
Simon H A 1957 *Models of Man, Social and Rational*. Wiley, New York
Thompson J D 1967 *Organizations in Action*. McGraw-Hill, New York
Weber M 1964 *Wirtschaft und Gesellschaft*, 2nd edn. Kiepenheuer & Witsch, Cologne, Germany
Woodward J 1958 *Management and Technology*. H.M.S.O., London
Woodward J 1965 *Industrial Organization: Theory and Practice*. Oxford University Press, London
Woodward J (ed.) 1970 *Industrial Organization: Behavior and Control*. Oxford University Press, New York and London

E. Friedberg

Organizations: Unintended Consequences

Unintended consequences are the grand theme of sociology. Like all social sciences sociology assumes that people act intentionally, however, in sociology these actions are seen to be socially conditioned, and while people act upon their intentions they produce unanticipated, unintended, even perverse consequences. While the notion of social conditioning has become part of the western worldview, this has not been the case for the notion of unintended consequences.

Organizations are often constructed by people to make up, through the pooling of resources, for the unintended consequence of individuals being unable to act in their common interest. Yet organizations produce new second-order unintended consequences in and among themselves and for society at large.

1. *History of the Problem within Sociology*

Interest in unintended consequences started in the eighteenth century with the individualistic tradition of the Scottish moral philosophers. In accord with Bernard de Mandeville's words, 'private vices, public benefits,' Adam Smith analyzed the unintended benign effects of the free market: as if moved by a hidden hand free trade transforms the self-centered motives of individual actors into a greater wealth for all. The French contract tradition provided other insights. Condorcet showed that when parties, each with consistent but different preferences on various issues, come together to collectively decide upon them, preferences on the collective level might be inconsistent, resulting in voting cycles. Rousseau described *avant la lettre* what is now known as a prisoner's dilemma, a situation in which individual and collective rationality conflict. Two persons while hunting deer fail in their objective because instead of cooperating—what they agreed upon and what is necessary to catch such a big game—one of them is tempted by the immediate gain of catching a hare. The unintended

outcome is that both are collectively worse off than they could be. Thereafter, collectivist traditions prevailed. Hegel hinted at unintended consequences in his work on the long-term historical development when he spoke about the 'cunning of reason.' However, he did not analyze why they emerge. Marx adopted these ideas on laws in history and the notion that people make history without being aware, with one exception: under communism people can decide upon their common plan and escape from unintended consequences because they are unfettered by unjust power differences. The Durkheimian program of collectivist sociology did not encourage students to look for or analyze unintended consequences, although many of the phenomena Durkheim described can be reconstructed as being the unanticipated, unintended outcome of individual actions, for example, the solidarity that flows from ceremonial gatherings, or a rise in suicide rate after an economic boom, a so-called *crise heureuse*.

Merton (1949) placed the theme back on the research agenda with his analysis of manifest and latent functions, that is, consequences of individual actions that are either intended and recognized or unintended and unanticipated. He provided case-studies of self-fulfilling prophecies, such as customers beginning a run on a bank they wrongly believe to be insolvent, causing it to go bankrupt, or the Mattheus effect in science, that is, how fame begets fame.

About the same time Elias presented his general ideas on unintended consequences in an historical study on state-formation and the refinement of manners. According to him unintended consequences, for example, refined manners, will be more common in larger groups, with more complex social networks, and smaller power-differentials.

Lately, with the emergence of a rational choice sociology and organizational economics, unintended consequences become more central to research, although often under different headings (see *Rational Choice and Organization Theory*).

2. *Unintended Consequences in Organizations*

Even a cursory review of the research literature shows that unintended consequences in organizations come in many forms and with multiple origins.

In the wake of Weber's discussion as to why Marxism did not succeed in its revolutionary goals, Michels (1911) formulated the iron law of oligarchy, probably the most famous instance of an unintended effect. This 'law' states that with time and growth of the organization there will be oligarchic tendencies in every organization, also in socialist organizations with a democratic ideology and constitution. Why is it rational not to be interested in one's own interests and not use one's democratic rights? Over time and with growing size it becomes less practical that all participate, and even leaders have to specialize and delegate. As the social distance between members and leaders grows, the cost for members to stay informed and coordinate their actions increases. This gives discretionary power to the leaders also if they only have eyes for the interests of the members. There are, however, exceptions to the iron law, like the ITU, the union of typographers that was studied by Lipset et al. (1956).

Just before WWII, industrial sociology discovered the informal group in formal organizations. Informal relations sometimes go against the objectives of the firm as they lead to rate busting. Yet other research underlined their productive contribution. The work-leader's centrality within this informal network seems critical: employees will cooperate more readily and management can keep track of ties between employees (see *Network Analysis*).

Since informal networks can have negative consequences, a work leader has to keep track of and control all the possible ties between employees and, therefore, his span of control should be limited. Research showed an intricate unintended relation between span of control, levels of hierarchy within an organization, and level of pay: a smaller span of control leads to a taller hierarchy and greater wage differentials in an organization. Promotion schemes are another known source of unintended consequences. Stouffer (1949) gave the classic example of soldiers who were less satisfied and had lower morale in companies with a higher rate of promotion. To evaluate their own status as well as to estimate what actions would be in their own interest people compare themselves to similar others. They care for relative, and not for absolute status. Recent research, however, could not replicate Stouffer's findings. Even those who are not promoted interpret high promotion rates as a sign of trustworthiness of the employer.

Lately, an important instance of unintended consequences in organizations has been discovered, that is, the performance paradox (Meyer and Gupta 1994). Where performance of a department or organization can be established, somehow measurements run down over a period of time, for of a number of reasons: (a) positive learning, that is, as performance improves, indicators become less useful, (b) perverse learning, that is, performance seems to improve without actually doing so because employees or managers adapt to the measurement system and learn how to fix the indicators, (c) selection or replacement of poor performers for better performers. Often these processes occur simultaneously.

3. *A Typology of Unintended Effects*

It is not the problem as such but its manifold manifestations, their explanation, and how they can be amended that is of interest. Yet the various forms of

unintended consequences have a common denominator in that they can all be subsumed under the problem of rationalization, since they affect efficient goal attainment.

A useful typology that covers the basic forms should at least have six dimensions (Boudon 1977):

(a) Number of individual actors that realize their goals,

(b) Whether the effects are positive, negative, or both,

(c) Number of actors affected by the outcomes,

(d) Whether they were anticipated or not,

(e) The time it takes for consequences to develop, and

(f) Whether a corporate actor or individual persons are the focal actor(s).

4. *Explanation Sketches in Sociology and Economics*

Existing theoretical knowledge on unintended consequences is not well integrated. Accounts given refer to a number of recurring elements, though: lack of or faulty information and incentives to act opportunistically. Sociologists traditionally focus on the former, and recently economists have emphasized the latter.

Until Coase explained why organizations exist, economists were silent on what happened in organizations. They exist because they are more efficient in certain situations as they save not only on production costs but also on transaction costs stemming from recurrent transactions and transaction specific investments (see *Hierarchies and Markets*).

Economists made clear that there are dilemma situations in which consequences emerge that are unintended but foreseen. The employment relation, for example, is fraught with anticipated, yet often inevitable unintended effects. Employer and employee are collectively better off if they trust each other and cooperate, yet each is tempted by opportunism.

Firm specific investments in human capital made by the employees can, depending on the alternatives both have, be misused by both parties against the other. An employer may cheat upon his promises just as the employees may sometimes shirk or reverse who controls whom.

What situational characteristics lead to these deficits in information and incentives to act opportunistically? The examples of unintended consequences discussed either stem from functional, cognitive, or structural interdependencies between various actors (Lindenberg 1997). People need each other to cooperate in the attainment of a common goal, to acquire or reduce information, and to get advice, gain approval, or emotional support. Moreover, persons that are functionally, interdependent, for example, in their tasks, will interact and it will also affect their mind.

Another relevant distinction is whether the unintended effects of a transaction affect only those involved in the transaction or also other third parties. Economists call the latter 'externalities.' Sociologists would consider them third party effects. Actors do not want to be responsible for negative externalities but they would like to internalize positive ones.

If groups grow in size, Olson (1965) showed, this logic will keep people from contributing to the common good. Many unintended effects in organizations come from the outside as external effects, for example, growing time pressure and growing wealth in society at large make people want to trade time for money resulting in many voluntary organizations nowadays being 'check book writing organizations.'

Many unintended effects come about through path dependency of social processes, that is, effects of earlier choices feed back into the social situation under which individuals decide on their next move, augmenting original differences, as with, for example, oligarchic tendencies.

Why are there exceptions to unintended effects? From the relevant research literature one might take the following advice. To come up with explanations one could stick to a behavioral theory of action and think about who are relevant actors, what are their goals, what actions would further these goals at what costs given the situation, and what actions would constitute the unintended effect. Oligarchy, for example, is reflected by nonparticipation by members and leaders not acting upon members' interests. In the case of the ITU, for example, the three kinds of interdependencies came together such that the difference in knowledge and political experience was far less than usual and that leaders were also part of the informal groups. These are all conditions that promote a choice for the democratic alternative. The absence of voting chains that might complicate originally democratic choices also helps (see Wippler 1988).

What can be done about negative unintended effects? It makes a difference whether there are central agencies that try to stabilize, correct, or foreclose the emergence of unintended negative outcomes. Functional interdependencies can be changed through social technology, for example, through redefinition of rights of actors (see *Design and Form: Organizational*).

Cognitive and especially structural interdependencies are harder to manipulate. As can be learned from principal agent and transaction cost theory, many unintended effects in organizations come from the split of ownership and usage rights, or more generally the split between authority and responsibility. Organizational rules can redefine property rights and thereby align interests such that to those who have control right over a productive capacity also have an interest in employing this capacity in a productive way. A simple application of this advice is to bind salaries of managers to performance of their department or firm. Another one is to make employees not

responsible for mistakes made by others, for example, in a sequential production process a system of quality control through backward policing allows employees to refuse products delivered by others (see Milgrom and Roberts 1992).

See also: Competitive Strategies: Organizational; Conflict: Organizational; Corporate Culture; Information and Knowledge: Organizational; Intelligence, Prior Knowledge, and Learning; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Learning: Organizational; Organizational Behavior, Psychology of; Organizational Change; Organizational Decision Making; Organizational Errors: Catastrophic; Organizations, Sociology of; Rational Choice Explanation: Philosophical Aspects; Risk, Sociological Study of; Risk, Sociology and Politics of; Strategic Intelligence; Strategy: Organizational; Technology and Organization

Bibliography

Boudon R 1977 *Effets pervers et ordre social*. Presses Universitaires, Paris

Lindenberg S 1997 Grounding groups in theory: Functional, cognitive and structural interdependencies. *Advances in Group Processes* **14**: 281–331

Lipset S M, Trow M A, Coleman J S 1956 *Union Democracy*. Free Press, New York

Merton R K 1949 *Social Theory and Social Structure*. Free Press, Glencoe, IL

Meyer M W, Gupta V 1994 The performance paradox. *Research in Organizational Behavior* **16**: 309–69

Michels R 1911/1925 *Zur Soziologie des Parteiwesens in der modernen Demokratie. Untersuchungen über die oligarchischen Tendenzen des Gruppenlebens*. Alfred Kroener Verlag, Stuttgart, Germany

Milgrom P, Roberts J 1992 *Economics, Organization and Management*. Prentice-Hall, Englewood Cliffs, NJ

Olson M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA

Stouffer S 1949 *The American Soldier*. Free Press, New York

Wippler R 1986 Oligarchic tendencies in democratic organizations. *Netherlands Journal of Sociology* **22**: 1–17

H. Flap

Orientalism

The term 'Orientalism' underwent a revolutionary shift in meaning after the publication of Edward Said's *Orientalism* in 1978. Before Said, the term conveyed several meanings, all of them quite affirmative of the Orient and of those who pursued knowledge of Oriental culture. In one meaning, Orientalism denoted an academic specialization—the study of the Orient, which, in widest compass, included the Near East, the Indian subcontinent, and East Asia. The term also demarcated a Western art, music, and literary style influenced by Eastern esthetics and philosophy, as, for example, to be found in the New England Transcendentalists in the USA. Most positively, it referenced the appreciation by Europeans of the great learning of the East. This appreciation and respect was in evidence, for example, among eighteenth-century British colonial administrators in Eastern India, who celebrated the sacred philosophic writings they uncovered from Bengal's 'golden' period, that is, from the period of Hindu rule before the Muslim conquest. In the nineteenth century, European affirmation of the Orient was expressed, for example, in Theosophy's appreciation for the religious teachings of Eastern 'Mahatmas' ('great souls') (Campbell 1980). In the same century, Edward Carpenter praised the Orient for what today we would call alternative sexualities (Carpenter 1914). Nineteenth-century Orientalism also included scholarly research, such as that of Max Müller, in philology and philosophy that pinpointed the Orient, especially India, as the ultimate source of the Indo-Aryans and, subsequently, the Greek and Roman civilizations (Schwab 1984, Müller 1883). If politics adhered to the term in any way, it was in the supposition that Orientalists respected and conserved the cultural traditions they studied.

1. Orientalism After Said

Edward Said, a professor of comparative literature at Columbia University in New York and a major commentator on Near Eastern politics, especially the Palestinian question, radically altered the meaning of Orientalism. His book *Orientalism* questioned the reality of the place called the 'Orient' and turned Orientalism into a pejorative term that carried political meanings. 'Orientalism' acquired, so to speak, permanent quotation marks around it to indicate that it can never again be used in an unself-conscious, nonreflexive, and uncritical way. After Said, to be called an 'Orientalist' is very often not a compliment, it is an accusation. Drawing mainly on material from the Near East (but with implications far beyond), Said argued that 'Orientalism' is something very different from what its adherents claim it to be, and that it functions to create inequalities of which they might not have been aware. The literary, philological, and historical studies of the Orient, the travel accounts, the depictions of its religions and arts—these, Said says, pretend to reflect or represent the Orient when, in fact, they construct it—and the image of it they build up is of a place different from and inferior to the West. The Orient is thereby created as feminine (and, consequently, weak and uncourageous), childish, superstitious, traditional (and therefore backward), des-

potic, otherworldly (and therefore unchanging), and sensual (and therefore passionate, not rational). The result is a Western portrayal of the East that dominates it by claiming to truly represent it.

For Said, the Orient fashioned in Western scholarship is complicit with European colonialism. Both colonialism and 'Orientalism' exercise empire over the societies of the Near East, India, and East Asia. The exact relationship between colonialism and Orientalism is unclear in Said's volume, as Aijaz Ahmad (1992) has noted. Said can be read to say either that the West's image of the Orient preceded and facilitated colonial conquest or that colonialism and Orientalism developed in conjunction and were mutually supportive.

Said thus made the Orient disappear as a real object or place, and appear only as a fabrication of the West. At times, Said seems to allow that a real Orient exists, and at other times he seems to deny it (as Clifford 1988 points out), but the salient issue for Said is that 'Orientalism' exists—and dominates. Similarly, although Said has received criticism for seeming to deny entirely that Western scholars might represent the Orient truly, his important message is that the representations are stereotyped and conform to a customary discourse that extends over centuries.

Said's Orientalism presents a critical viewpoint and an interpretive methodology that initiates the approach often called 'postcolonial studies' in today's social science and humanities. In this approach, determining who controls knowledge and representation and who constructs representations in colonial situations is just as important as determining who owns the productive resources or who exercises the powers of government.

2. *The Postcolonial Critique*

Important criticisms, revisions, and elaborations of Said's Orientalism have developed from such postcolonial studies, especially as geographical areas beyond the Near East have come under scrutiny. One major revision has been to bring a strong sense of history and process over time into Said's basically literary and timeless approach. Said often presents Orientalism as nearly fully formed from its outset and as relatively stable over long periods of time. If 'Orientalism' is constructed, however, there must be a building process, and it needs to be chronicled. Tim Mitchell, for example, has provided such a history of how Europeans colonized Egypt in the nineteenth century by importing representations of time, language, texts, the body, and good society. Through this imported knowledge and understanding, they convinced Egyptians that their society and culture were different—as well as inferior (Mitchell 1991). Gauri Visvanathan (1989) traces the history of English education in India, and shows how Indian students learned in their textbooks and classrooms that Western culture was superior. Because studies like these show Orientalism as a historical process rather than a finished state, they break with Said's understanding of it as an unbreachable discourse (a position he takes from Foucault). They also, however, recover his concept for meaningful historical analysis.

Other critics fault Said for presenting the West as monolithic and unchanging—in effect, he is said to represent the West in just as stereotyped and politicized a manner as he chastises Orientalists for doing to the East (see Clifford 1988 and Ahmad 1992).

One problem with Said's stereotype of European Orientalism is that it does not distinguish between positive images of the East and negative ones. Although both types of images stereotype the Orient, the positive images often served Europeans as means by which to criticize their own society. These positive images also often allowed nationalists in India and elsewhere to combat the negative stereotypes of their societies (Fox 1992).

3. *Orientalism's Consequences*

The other problem with Said's stereotypic portrayal of Europe is that he does not see the consequences for Europe itself of the construction of Orientalism. Ashis Nandy (1983) has augmented Said's approach by detailing the effects of colonialism, especially the effects of the representation made of the colonized, on European culture. Nandy argues (without using Said's terminology) that the construction of images of the colonized, such as Orientalism, had profound effects on Europeans themselves. Because they stereotyped the East as feminine, passive, and cowardly, Europeans had to emphasize a violent, active, hypermasculinity as their distinctive character, and minimize what Nandy calls the 'recessive' feminine, nonviolent, and unaggressive elements. Nandy therefore adds two points to the analysis of 'Orientalism': that such constructions have consequences even for those who have the power to construct them; and that the object of Orientalism—that is, the Oriental—is involved in an intimate, if also unequal, relationship with the European perpetrator of Orientalism.

Nandy's second point highlights the greatest limitation in Said's initial presentation of Orientalism: he did not explore the effects of European Orientalism on the Orientals themselves. This European discourse had profound effects on the self-image of the people whom it referenced. They often came to think of themselves along the same Orientalist lines as did Europeans: as inferior and backward. Sometimes, therefore, they adopted a negative self-image, in keeping with Said's understanding of the power of European representations. At other times, however, Orientalism could be the source of resistance to European domination. For example, Gandhi took

many of the negative stereotypes about India incorporated into Orientalism and turned them into positive attributes (although they were still stereotypes): Oriental 'otherworldliness' became Eastern 'spirituality' and 'passivity' became 'nonviolence' (Fox 1989). Such reversals also indicate that Orientalist discourse as a vehicle of domination can be turned back on itself.

Said's *Orientalism* signals an important epistemic break in Western social science and humanities, as is evident by the further research it inspired. His overall analytic remains an important entry point for understanding the power of representations and the domination embedded in European conceptions of other places.

See also: Colonialism: Political Aspects; Colonization and Colonialism, History of; East Asian Studies: Culture; East Asian Studies: Politics; Near Middle East/North African Studies: Culture; Postcoloniality

Bibliography

Ahmad A 1992 *In Theory: Classes Nations and Literatures*. London, Verso

Campbell B F 1980 *Ancient Wisdom Revived: A History of the Theosophical Movement*. University of California Press, Berkeley, CA

Carpenter E 1914 *Intermediate Types Among Primitive Folk: A Study in Social Evolution*. Allen & Unwin, London

Clifford J 1988 On orientalism *The Predicament of Culture*. Harvard University Press, Cambridge, MA, pp. 255–76

Fox R 1989 *Gandhian Utopia*. Beacon, Boston

Fox R 1992 East of Said. In: Sprinker M (ed.) *Edward Said, A Critical Reader*. Basil Blackwell, London, pp. 144–56

Mitchell T 1991 *Colonizing Egypt*. University of California Press, Berkeley, CA

Müller M 1883 *India, What Can It Teach Us?* J.W. Lovell, New York

Nandy A 1983 *The Intimate Enemy: Loss and Recovery of Self Under Colonialism*. Oxford University Press, Delhi, India

Said E 1978 *Orientalism*, 1st edn. Pantheon Books, New York

Schwab R 1984 *The Oriental Renaissance: Europe's Rediscovery of India and the East 1680–1880* [trans. Paterson-Black G, Reinking V]. Columbia University Press, New York

Visvanathan G 1989 *Masks of Conquest: Literary Study and British Rule in India*. Columbia University Press, New York

R. G. Fox

Orienting Response

1. *The Concept of Orienting Response*

The orienting response (OR), discovered by Pavlov (1927), is a set of reactions evoked in humans and animals by a novel stimulus. The OR is manifested in an interruption of an ongoing activity (an external inhibition) that parallels turning of eyes, head, and body movements toward a novel stimulus (the targeting response) (Konorski 1967). Subjectively, the OR is experienced as an involuntary switching of attention emphasizing a novel stimulus.

2. *Components of OR*

The behavioral manifestations of OR represent only the tip of an iceberg of reactions hidden from direct observation but revealed by instrumental research. OR is characterized by a constriction of peripherical vessels and dilation of brain vessels. The regional cerebral blood flow (rCBF) is of particular interest, being related to a local activation of neurons expressed in the generation of action potentials (APs). APs are linked with passive inflow of sodium and outflow of potassium ions from nervous cells. To keep a balance of ionic content in neurons an active sodium-potassium pump is operating. It requires additional energy supplied by oxygen and glucose during the intensification of rCBF. The increase of rCBF as an index of OR-related neuronal activation is used in positron emission tomography (PET) and functional magnetic resonance imaging (fMRI).

Neuronal networks involved in OR evocation are identified also by electroencephalography (EEG) and magnetoencephalography (MEG) (Posner and Raichle 1994). Multichannel EEG and MEG recordings are used to calculate equivalent dipoles localization representing loci of neuronal activation. OR is evident as event-related desynchronization (ERD) of background brain waves in the range of 8 to 12 Hz (alpha waves) that parallels an increase of beta (12 to 25 Hz) and gamma (25 to 200 Hz) oscillations (Naatanen 1992). The ERD is accompanied by a negative shift of the cortical steady potential that is related to an increase in neuronal excitability (Caspers 1963). The other OR index is nonspecific event-related potentials (ERPs) evoked by novel stimuli (Brunia 1997).

Direct observation of OR-related neurons is done by single unit microelectrode recording. The OR is also expressed in autonomic indicators: deceleration of heart rate and an increase of skin conductance (Graham 1979). Sensory OR components refer to an elevation of sensitivity (lowering of thresholds) and to an increase of fusion frequency (rate of discrete stimuli building up a continuous percept) (Lindsley 1961).

3. *Selective Habituation of OR*

Repeated presentation of a nonsignal standard stimulus results in a gradual reduction of OR components (OR habituation). The OR habituation represents a form of negative learning. A change of the presented stimulus produces OR, again demonstrating the con-

tribution of novelty. A standard stimulus following a novel one results in OR recovery called dishabituation (Groves and Thompson 1970). The standard stimulus characterized by a particular intensity, shape, color, duration, and interstimulus intervals builds up in neuronal networks an internal representation of all these parameters—a stimulus neuronal model. Any deviation of a stimulus from the created neuronal model produces a mismatch signal that determines OR generation. The stimulus neuronal model preserves not only simple features, but also their combinations. Thus, after habituation of OR to a complex consisting of light and sound components, elimination of either light or sound evokes OR.

The stimulus neuronal model can be characterized as a multidimensional self-adjustable reject filter. A selective tuning of the filter can be revealed by repeated presentation of the standard stimulus and intermittent stimuli deviating with respect to a particular parameter from the standard one.

The procedure applied to different stimulus parameters enables the construction of a multidimensional filter with a selective characteristic that underlies stimulus selective OR habituation. Repeated presentation of a standard stimulus reveals generalized and localized ORs. Generalized OR is evident by the involvement of a wide variety of components that rapidly habituate. Localized OR is specifically related to the applied stimulus modality. Generalized and localized ORs can be demonstrated by event-related desynchronization (ERD) of brain waves. Different brain areas are represented by specific alpha, mu, sigma, and tau generators of 8 to 12 Hz oscillations. The visual area is represented by alpha rhythm, the motor area by mu rhythm, the somatosensory area by sigma rhythm, and the auditory area by tau rhythm. A novel visual stimulus evokes a generalized OR composed of generalized ERD, skin galvanic response (SGR), and heart rate deceleration. After 15 to 20 presentations of the visual stimulus only occipital alpha-desynchronization is seen being habituated in subsequent trials. Passive hand movement produces initially a generalized ERD that under repetitions is transformed into a local mu rhythm desynchronization representing a localized OR (Sokolov 1963).

4. *OR and Conditioned Reflex*

A combination of neuronal (nonsignal) stimulus with an unconditional stimulus (US) results in an intensification of OR. The stimulus signaling a US becomes significant and produces more stable OR. In the process of stabilization of the conditional reflex (CR) the OR gradually habituates. An additional stumble, that has to be differentiated from the initial CS, results in a recovery of OR. The more difficult the differentiation task, the greater is the OR. If CSs are presented at regular intervals, a time-selective CR is elaborated that precedes actual presentation of the CS. It parallels an anticipatory OR expressed in saccadic eye movements, ERD and gradual negative shift of study potential (expectancy wave). Thus nonsignal (insignificant) stimuli produce rapidly habituatable OR, while signal (significant) stimuli triggers produce more persistent OR.

5. *OR and Vector Encoding*

A stimulus acting on an ensemble of receptors generates a combination of their excitations—a receptor excitation vector. The receptors' overlapping characteristics constitute a nonorthogonal basis. At the next stage of information processing the neuronal ensemble performs orthogonalization and normalization so that input stimuli are represented in Euclidean space by excitation vectors of a constant length.

An excitation vector evoked by an input stimulus is acting in parallel on a set of feature detectors. Each feature detector is selectively tuned to a particular excitation vector and in this way maximally excited by a respective input stimulus. Thus input stimuli encoded by excitation vectors are topically represented on a detector map. The detector map is a hypersphere in a multidimensional space. The output responses are also encoded by means of excitation vectors. A variety of specific responses can be generated by a limited number of channels constituting components of an output excitation vector (Bizzi and Mussa-Ivaldi 1998). A link between input and output excitation vectors having a set of plastic synapses modified in the process of learning so that command neuron become selectively tuned to a CS (Fomin et al. 1979).

The OR habituation as a negative learning is also due to the synaptic plasticity of specific novelty-related command neurons. The process of OR selective habituation can be explained by the formation of a Hebbian link vector opposing an excitation vector of a standard stimulus (Hebb 1949).

The model of OR information processing presented above can be illustrated by color vision. Three types of cones constitute a nonorthogonal basis for receptor excitation vectors. Three types of photopic horizontal cells, red-green, blue-yellow, and luminance, are relevant for the orthogonal basis, where intensity is given by the vector length. At the level of bipolar cells the length of excitation vectors becomes constant. This is achieved by a transition from a three- to a four-dimensional space with the addition of a neuron active under darkness. At color specific visual area (V4) different colors are specified by color-selective detectors (Bartles and Zeki 1998). The color detector map is a hypersphere in the four-dimensional space (Izmailov and Sokolov 1991). The spherical surface representing colors suggests that different colors be topically separated, constituting a colortopic cortical map similar to retinatopic, tonotopic, and somatotopic

projections. An fMRI used to localize color representations reveals color-selective areas in the human occipitotemporal cortex.

Output vector code is evident in different ballistic movements (Bizzi and Mussa-Ivaldi 1998). OR vector code is expressed in saccadic eye movements, which are selectively triggered by particular combinations of horizontal, vertical, and torsion premotor neurons. Vector code refers to autonomic OR response (Sokolov and Cacioppo 1997).

6. *Neuronal Basis of OR*

PET and fMRI in combination with EEG and MEG studies in humans have shown that OR circuits differ in passive (involuntary) and active (voluntary) OR. Passive OR is addressed to the hippocampus (Tulving et al. 1994) while active OR involves the prefrontal and parietal lobes. Single unit recording in the hippocampus of a rabbit has demonstrated that neurons under passive OR are separated into two groups. Novelty-sensitive neurons are excited by a novel stimulus. Familiarity-sensitive neurons are inhibited by a novel stimulus. By repeated presentation of a novel stimulus responses of excitatory and inhibitory cells are habituated.

Selective habituation with respect to a standard stimulus is due to a selective decrease of synaptic weights connecting activated feature detectors with novelty-related neurons (Vinogradova 1987). The mismatch signal evoked by the noncoincidence of a stimulus deviating from the standard one parallels a difference of their excitation vectors. The mismatch signal is addressed to the mesencephalic reticular formation and thalamic reticular nuclei evoking a frequency shift of their pacemaker cells from alpha waves (Moruzzi and Magoun 1949) toward gamma oscillations. Single unit recording from the reticular thalamic nucleus in rabbits demonstrates randomization response—a transition from bursting spike activity to a random spiking that parallels ERD of cortical alpha-like oscillations. By repeated presentation of a novel stimulus, neuron spiking thalamic randomization response and ERD selectively habituate in accordance with the stimulus selective habituation of other OR components (Sokolov 1975).

The hippocampus plays a distinct role in the orienting reflex and exploratory behavior. The hippocampus is characterized by (a) long-term potentiation of synaptic contacts; (b) hippocampal theta waves (4 to 7 Hz) during arousal and orienting reaction; (c) novelty and familiarity neurons.

Cortical future detectors reach the hippocampus via the perforant path contacting apical dendrites of pyramidal cells. In passing, individual fibers make numerous contacts with many pyramidal cells so that excitation is spread over a wide area of the hippocampus. A parallel branch of the perforant path contacts granule cells of the dentate fascia, which sends its mossy fibers to basal dendrites of the pyramidal cells. The potentiation of the mossy fibers' synapses gradually blocks direct synaptic contacts so that the input signal on the pyramidal neurons is selectively diminished. A new stimulus acting via synapses not blocked by the mossy fibers evokes novelty responses in the pyramidal cells that trigger OR.

A critical role of the dentate fascia in the process of habituation was shown using antibodies against neurons of the dentate fascia. This procedure results in an elimination of novelty responses. The novelty neurons become continually responsive under repeated stimulus presentation (Vinogradova 1987).

To summarize: OR is a neuronal mechanism emphasizing novel and significant stimuli against familiar and insignificant ones. One might say that OR is a high order contrasting mechanism contributing to long-term memory formation.

7. *Conclusion*

Neurons demonstrate basic characteristics of OR. They selectively habituate to different Single unit recording in animals demonstrated that hippocampal novelty-sensitive parameters of the standard stimulus respond to any change of stimulation and show dishabituation. In the hippocampus, familiarity-sensitive neurons are operating in parallel with the novelty-sensitive neurons.

Recent fMRI data show that the human hippocampus is also processing both novelty and familiarity. The left anterior hippocampus responds to perceptual and semantic novelty. By contrast, the bilateral posterior hippocampus is activated by increasing semantic familiarity. Repeated presentation of semantic stimuli results in the topographical spread of hippocampal activity in an anterior–posterior direction. The specific novelty-encoding path plays a crucial role in formation of long-term memory within the posterior medial-temporal lobe (Strange et al. 1999). PET, fMRI, and ERPs show that under active attention task additionally frontal and parietal areas, the anterior cingulate gyrus on the frontal midline are activated, demonstrating that active OR to signal stimuli has a specific circuitry differing from passive OR (Mountcastle 1979).

A novelty signal is widely distributed among different brain structures including the activating reticular formation, autonomic centers, and superior colliculus, facilitating eye and body movements as OR components. At the same time, the novelty signal acts as a distracter of ongoing behavior.

See also: Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Habituation; Pavlov, Ivan Petrovich (1849–1936)

Bibliography

Bartels A, Zeki S 1998 The theory of multistage integration in the visual brain. *Proceedings of the Royal Society London B. Biological Sciences* **265**: 2327–32

Bizzi E, Mussa-Ivaldi F A 1998 The acquisition of motor behavior. *Daedalus* **127**: 217–32

Brunia C H M 1997 Gating in readiness. In: Lang P J, Simons R F, Balaban M T (eds.) *Attention and Orienting*. Erlbaum, Mahwah, NJ

Caspers H 1963 Relations of steady potential shift in the cortex to wakefulness-sleep spectrum. In: Brazier M A B (ed.) *Brain Function: Vol. 1. Cortical Excitability and Steady Potentials*. University of California Press, Berkeley, CA

Fomin S V, Sokolov E N, Vaitkyavicus G G 1979 *Artificial Sense Organs. Modeling of Sensory Systems*. Nauka, Moscow, Russia

Graham F K 1979 Distinguishing among orienting, defensive and startle reflexes. In: Kimmel H D, van Olst E H, Orlebeke J F (eds.) *The Orienting Reflex in Humans*. Erlbaum, Hillsdale, NJ

Groves P M, Thompson R F 1970 Habituation: A dual process theory. *Psychological Review* **77**: 419–50

Hebb D O 1949 *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York

Izmailov Ch A, Sokolov E N 1991 Spherical model of color and brightness discrimination. *Psychological Science* **2**: 249–59

Konorski J 1967 *Integrative Activity of the Brain: An Interdisciplinary Approach*. University of Chicago Press, Chicago, IL

Lindsley D B 1961 The reticular activating system and perceptual integration. In: Sheer D E (ed.) *Electrical Stimulation of the Brain*. University of Texas Press, Austin, TX

Moruzzi G, Magoun H W 1949 Brain stem reticular formation and activation of the EEG. *Electrophysiology and Clinical Neurophysiology* **1**: 455–73

Mountcastle V B 1978 Brain mechanisms for directed attention. *Journal of Royal Society of Medicine* **71**: 14–29

Naatanen R 1992 *Attention and Brain Function*. Erlbaum, Hillsdale, NJ

Pavlov I P 1927 *Conditioned Reflexes*. Oxford University Press, London

Posner M I 1995 Attention in cognitive neuroscience: an overview. In: Gazzaniga M S (ed.) *The Cognitive Neurosciences*. MIT press, Cambridge, MA

Posner M I, Abdullaev Y G 1999 Neuroanatomy, circuitry and plasticity of word reading. *NeuroReport* **10**: R12–23

Posner M I, Raichle M E 1994 *Images of Mind*. Scientific American Library, New York

Sokolov E N 1963 *Perception and the Conditioned Reflex*. Pergamon Press, Oxford, UK

Sokolov E N 1975 The neuronal mechanisms of the orienting reflex. In: Sokolov E N, Vinogradova O S (eds.) *The Neuronal Mechanisms of the Orienting Reflex*. Erlbaum, Hillsdale, NJ

Sokolov E N, Cacioppo J T 1997 Orienting and defense reflexes: Vector coding the cardiac response. In: Lang P J, Simons R F, Balaban T M (eds.) *Attention and Orienting*. Erlbaum, Mahwah, NJ

Sokolov E N, Vaitkyavicus G G 1989 *Ot neirona k neirokompiuteru* [*From a Neuron to a Neurocomputer*]. Nauka, Moskva, Russia

Strange B A, Fletcher P C, Henson R N A, Friston K J, Dolan R J 1999 Segregating the functions of human hippocampus. *Proceedings of the National Academy of Science, USA* **96**: 4034–39

Tulving E, Markowitsch H J, Kapur S, Habib R, Houl S 1994 Novelty encoding networks in the human brain: positron emission tomography data. *NeuroReport* **5**: 2525–8

Vinogradova O S 1976 *Gippokamp i namiat* [*The Hippocampus and Memory*]. Nauka, Moskva, Russia

Vinogradova O S 1987 The hippocampus and the orienting reflex. In: Sokolov E N, Vinogradova O S (eds.) *The Neuronal Mechanisms of the Orienting Reflex*. Erlbaum, Hillsdale, NJ

E. N. Sokolov

Orthodoxy

The term 'Orthodox Church' is the normal term to describe a body that like the Roman Catholic Church, traces its origins to apostolic times. However, it ceased to share in the life of the Roman Catholic Church from the early Middle Ages. It would be misleading to give a single date for the separation of the two traditions. Certainly, 1054 should no longer be treated as the decisive moment of division. More important was the year 1204, when the fourth Crusade sacked the city of Constantinople and thereby demonstrated the 'otherness' of Western as compared to Eastern Christians. The very use of such terms suggests that there were cultural, not merely dogmatic, differences between these areas of the Christian world.

Meanwhile, the Orthodox have spread far beyond their one-time homelands in the 'East.' This was largely the Near East, as defined from a West European point of view. From the ninth century, Central and Eastern Europe were added to this notional 'East.' The modern move of Orthodox communities to other parts of the world tends to make 'Eastern Orthodox' an imprecise if not misleading term.

Some centuries before the separation of the Orthodox and Roman Catholic worlds occurred a different separation in the 'East.' The church council of Chalcedon (451) was the occasion on which a number of church bodies failed to accept the teaching, therefore also the fellowship of the established Orthodox Church in the Byzantine empire. It was (mistakenly) asserted that these were Monophysites. In other words it was presumed that they believed in Christ's single nature as God at the expense of his humanity. The Churches that were supposed to maintain this teaching belonged to the Near and Middle East, as well as to Africa. They are often described as non-Chalcedonian. Less precisely, they are defined as 'Oriental' Orthodox. In the course of the twentieth century, they also have moved far beyond their traditional locations. Unlike the 'Eastern' Orthodox, all of whom follow the Byzantine rite, the 'Oriental' Churches have their separate liturgical traditions.

1. The 'Eastern' Orthodox at Home

1.1 The Former Ottoman Empire

World War 1 saw the disintegration of the Ottoman empire in which the Orthodox had been always granted a recognized if subordinate role as a people (*millet*).

Under Ottoman rule this *millet* had constituted a clearly segregated society within the Muslim realm. The Orthodox patriarch of Constantinople was treated as its principal administrator or ethnarch. It was possible to escape from the restraints imposed on the *millet* by conversion to Islam. Individuals who chose this way to advancement in the Muslim administration naturally lost the esteem of their former co-religionists. Yet those Orthodox who were executed after subsequent repudiation of Islam came to be venerated as neo-martyrs.

The Orthodox Church fostered ethnic or national aspirations in Ottoman times. In effect it was a people's Church. All the more was it ready for new prominence in public life once Ottoman power had come to an end.

Thus the Church of Greece came into its own after the Greek uprising against the Turks (1821). In due course it was proclaimed the established Church of the new Greek kingdom (1852). Likewise the Serbian Church was able to proclaim its independence in 1879. This was in the aftermath of Serbian emergence from the Turkish yoke. Hence the country's Orthodox were eager to assert the unity of Church and state.

The Orthodox of Bulgaria aspired to such unity as well. But the Bulgarian Church claimed its independent status (1870) some years in advance of national independence (1878). Its aspirations were therefore condemned by the patriarchate of Constantinople as untoward, if not heretical (1872–1953). By contrast, the newly established Rumanian Church was to gain the patriarchate's approval in 1885. It had already been recognized by the Rumanian constitution as the country's national ('dominant') Church in 1866. Even after all the assaults of communism the Church could claim the allegiance of 87 percent of the population. The small Albanian Church was not to gain its independence until 1937. Unlike its Balkan partners, it could not claim to represent a majority of the population (no more than 25 percent).

These Orthodox populations could now swim with the tide and use their new opportunities to consolidate their dignity and self-esteem. But in the process they tended to despise their Muslim masters of the recent (even medieval) past. This was to motivate (politically motivated) attempts at ethnic 'cleansing' in Bosnia and Kosovo.

Memories of subjugation could also inspire a revanchist ideology among the Greeks. An aspiration to restore Constantinople as the center of an Orthodox empire animated religious and secular thinkers alike. The latter saw this as their 'Great Idea.' It created no problems until the disarray of the postwar years enticed the Greek authorities to put it into practice. Their ill-fated attempt to restore the one-time Byzantine territories of Asia Minor to the Greeks (1919–22) led to a disaster for the Greek-speaking Orthodox of the region, who were subsequently hounded thence by the victorious armies of Kemal Ataturk. The convention of Lausanne (1923) seemed to legitimize an earlier form of 'ethnic cleansing.' Significantly, this treaty determined that religious criteria should decide how peoples should be moved in the compulsory exchange of populations between Turkey and Greece. Only a small minority of Orthodox Greeks remained in Asia Minor, with the status of their former ethnarch correspondingly diminished. The patriarch of Constantinople was thought to retain merely a symbolic role. Certainly no ethnarch he, since the *ethnos* was displaced.

Only in the Church of Cyprus was the role of ethnarch to retain, even to regain, its former role. Cyprus had been taken from the Turks in 1878 and ceded to the British crown. But the ruling bishop continued to be the leader of the people. This was clearly revealed when Cyprus was seeking independence from the British. The Cypriot archbishop of the day (Makarios) was not only leader in the struggle to achieve this. In the immediate aftermath he was also to gain election as president of the new state (1960–77).

The Church of Cyprus received its independence in 431. The patriarchates of Africa (Alexandria) and the Middle East (Antioch and Jerusalem) are also rooted in the early medieval past.

Despite its once modest numbers, Alexandria claims pastoral responsibility for the whole of Africa. It has consequently acquired a sizeable flock in 60 different countries. This has modified its previously well-established character as a factor of Greek mercantile life.

Greek influence is strong among the leaders of the Jerusalem patriarchate. But it is an influence that is critically assessed by a predominantly Arab flock (perhaps 100,000 strong). By contrast, Antioch (in effect Syria and Lebanon) insists that its Arabic culture is central to its concerns. It is therefore well placed to promote dialogue with its Muslim neighbors.

1.2 The Former Russian Empire

Orthodoxy was well entrenched in Russian society by the end of the Middle Ages, and it had few competitors on native soil. But expansion of the Muscovite state from the mid-sixteenth century incorporated peoples with allegiance to other faiths. Muslim peoples were to remain proof against conversion. By contrast, the various pagan peoples of Siberia were often persuaded to join the state Church. Here, as elsewhere, conversion served as a function of colonization. More easily incorporated were the Orthodox of neighboring nations, who joined the Russian empire as the result of

one treaty or another. Thus did Ukrainians merge with the Russian Orthodox Church (1686), while the ancient Church of Georgia was accommodated merely as a province (1811). After the church reforms of Peter I (1722), the Russian Orthodox Church was overtly a state church, with the emperor's lay nominee at the head of its subservient synod. It was taken for granted that the population would normally belong to this body. To this end, Orthodoxy was taught in the schools. An annual confession and communion were seen as civic duties. Until an imperial edict of toleration was published in 1905, non-Orthodox religious rivals were restricted in their public life, if not actively restrained; some kind of pluralism might have been envisaged.

But it was not to be. The Bolshevik party, which seized power in 1917, proclaimed separation of Church and state on January 23, 1918. It was soon evident that this involved control, if not suppression of the Church. Most of the Church's many institutions were abolished, and although Stalin's constitution proclaimed 'freedom of worship' (1936), hardly a hundred churches were left open by 1939. Mere membership of the Church involved risk of liberty and life. Among the millions who vanished into concentration camps were almost all the bishops. Meanwhile, anti-religious propaganda was deployed in the formation of *homo sovieticus*. 'Scientific atheism' was taught throughout the schools and universities of the USSR. Membership of the communist party required commitment to the atheist cause.

During World War 2 a degree of toleration of the Russian Orthodox Church was introduced by Stalin in an attempt to raise national morale (1943) in those parts of the USSR that had escaped German occupation. It was toleration that extended also to the Georgian Church, which had reclaimed independence in 1917. However, such toleration signified no lessening of controls by the party and the state. Rather were these to be extended to Orthodox institutions in the larger Soviet Union of the postwar years.

With equal rigor were controls imposed on Eastern Europe. This meant that the Orthodox church establishment of Rumania and Bulgaria were used to further Soviet interests, while atheism was promoted in society at large. *Mutatis mutandis* the Serbian Church was similarly affected in Tito's Yugoslavia. Yet only in Albania was official atheism to be taken to its logical conclusion: by 1967 religion was prohibited in any of its forms.

By contrast, the Greek Orthodox Church was left free of communist rule at the conclusion of a bitter civil war (1948–9). It was therefore able to continue its existence as established Church. Only one other matched it in this role, the Orthodox Church of Finland. But the two are curiously contrasted. Whereas the Church of Greece represents 98 percent of the population (hence problems of nominal membership), the Orthodox of Finland form hardly more than 1 percent of a predominantly Lutheran population.

The end of communist regimes in Eastern Europe and the former USSR has allowed the local Orthodox Church to reassert its place in secular society. But its potency and numbers are less than might have been expected in the aftermath of large-scale persecution. However, it is no longer barred from education or from social work, or from public propagation of its faith. Lack of an official ideology has encouraged the substitution of popular (often nationalistic) Orthodoxy by spokesmen of the Church. The Russian form of this nationalism involves an anti-Western stance, with antisemitism one of its ingredients.

2. *The 'Eastern' Orthodox Abroad*

2.1 *Missions*

The missionary activities of St Cyril and St Methodios in Moravia (ninth century), and the associated translation of sacred texts from Greek into Slavonic, had legitimized the use of languages other than Hebrew, Greek, and Latin in church life. Later missionaries could therefore work unconstrained in a variety of cultures. Orthodox Christianity was able to bring literacy as well as other social tools to its clients. This was the case with the Siberian and Central Asian missions of the Russian Church. These had their extension to China (1695) and Alaska (1794). In due course there were Orthodox parishes founded in Japan (1868) and Korea (1897). These were modest-sized monocultural outposts of the Orthodox world and thus free of the conflicting loyalties that were later to burden the emigres of Western Europe, Australia, and North America.

2.2 *The Orthodox Diaspora*

Economic pressures and political oppression both led to the migration of Orthodox populations in the course of the twentieth century. Communist rule in Russia and Eastern Europe was among the most potent of the forces that propelled communities into exile. The number of North American immigrants from Russia was swelled by the Russian revolution and civil war of the post-1917 period. Communist rule in Eastern Europe after 1945 also brought Ukrainians, Rumanians, Serbs, Albanians, and Bulgarians to the new world. Meanwhile, Greeks had arrived in considerable numbers after their expulsion from Asia Minor in 1922–3. There were also Palestinians after 1948. In due course Orthodoxy came to be recognized as one of the 'major' religions of the USA.

However, there was never a united church administration. The attempt to form a largely English-speaking 'Orthodox Church in America' (1970) did

not meet with universal approval. Apart from a Greek archdiocese, which retained its links with the patriarchate of Constantinople, there was a bewildering variety of national jurisdictions, and not infrequently more than one of each. Many immigrants preferred to retain their languages and customs. Even (or especially) in exile their church life was a function of national identity. Similar fragmentation was to be observed elsewhere (Western Europe or Australia, for example).

2.3 Diaspora and Mission

However nation-oriented the diaspora, it could still have an effect on its non-Orthodox host environment. The Orthodox may not have sought to proselytize their neighbours. Nevertheless, conversions took place. A common cause might be intermarriage between partners of different faiths. But other less specific causes might also prompt the conversion of individuals or groups.

Some of these have gravitated towards Orthodoxy of their own accord: here is hardly 'mission' in the usual sense of the term. Thus the self-established Orthodox of Uganda and Kenya were assimilated into the patriarchate of Alexandria (1946), and without becoming honorary Greeks in the process. Rather the black Orthodox of Africa (one million strong) have come to outnumber the Greek Orthodox founders of the patriarchate (200,000).

3. The 'Oriental' Orthodox at Home

However, there were already large numbers of indigenous African Orthodox outside the boundaries of the Alexandrian patriarchate, members of the non-Chalcedonian ('Oriental' Orthodox) Churches of Egypt and Ethiopia.

3.1 The Coptic Orthodox Church

The Coptic (= Egyptian) Orthodox Church encompasses much of the non-Arabic part of the population, something like 10 percent of the total. This Church emerged as a distinct entity in the fifth and sixth centuries. But its distinctness was confirmed when the Arabs conquered Egypt in 639–42, since this freed it from Byzantine (= Chalcedonian) control. It has retained its character as the church of its nation. Members of the Coptic Church have experienced constraints from the Muslim establishment. Some of the discrimination exercised by extremists is not far removed from persecution. All the more noteworthy is the vigor of church life among the Copts. In a low-income community this finds its expression in service of the poor.

3.2 The Ethiopian Orthodox Church

The Coptic Church supervised the life of the Ethiopian Orthodox Church from the fourth to the twentieth centuries. Not until 1957 was the formal autonomy of the Ethiopian Church proclaimed, and this despite the number of its members (20 million), which far exceeds that of any of its 'Oriental' sister Churches. This accounts for almost half the population. The state encouraged loyalty to the Church when it was an established institution. But much loyalty survived the rigors of a Marxist revolution, not only anti-clerical but anti-religious though it was (1974).

3.3 The Syrian Orthodox Church

The 'Oriental' Churches are also represented in the Middle East. Members of the Syrian Orthodox ('Jacobite') Church are distributed between Syria, Lebanon, and Iraq. In terms of culture and membership, this Church was at its strongest in the Middle Ages, but its numbers have declined to little more than 100,000. In the search of anonymity and self-protection (many have fled post-Ottoman Turkey since the 1920s for lack of adequate safeguards) its members tend to emulate their Arab neighbors' way of life.

3.4 The Syrian Orthodox Church of South India

However, the Syrian Orthodox gain extra prominence from the existence of their co-religionists in South India, who have been linked with them for much of the period since 1665 through acceptance of a common head, the ('Oriental') Syrian Orthodox patriarch of Antioch. However, the Church of South India, with its 1,500,000 adherents, has its independent history since early Christian times, and its autonomy is fiercely defended by many of its members to this day. As an indigenous church it is well integrated into its milieu (Kerala).

3.5 The Armenian Apostolic Church

Armenian home territories are difficult to define. The Armenian republic (formerly a constituent republic of the USSR) is central to them, not least since the spiritual center of all Armenians (Etchmiadzin) is located there. It was founded in 301. However, only a third of the Armenian people reside in that republic. Many find themselves in a diaspora. But others are distributed within the confines of an adjacent administration, the catholicosate of Cilicia (Lebanon) (founded 1441).

Armenia adopted Christianity as its state religion in 301. The Armenian Apostolic Church (its name in later centuries) has remained the Church of the nation since, irrespective of changing frontiers. Armenians in the Ottoman empire gained protection by virtue of

their own *millet* (1461). But they could still be subject to pressures and persecution. The last decades of the nineteenth century and the first of the twentieth witnessed massacres of the Armenian population by the Turks that reached their climax in the early years of World War 1. Here was a genocide that claimed something like one and a half million lives, a third of the nation. Hence the dispersal of Armenians throughout the world. The persecutions confirmed Armenians in their commitment to their nation, their culture, and their Church.

4. The 'Oriental' Orthodox Abroad

4.1 Missions

The 'Oriental' Orthodox have been little involved in such political imperialism as might have prompted missionary outreach to nations other than their own.

4.2 Diaspora

Yet each of the 'Oriental' Churches has branches abroad. Thus Armenians have their communities in North and South America, in Russia, Western Europe, and the Middle East. As many are in the diaspora as in Armenia itself.

See also: Catholicism; Christianity Origins: Primitive and 'Western' History; Historiography and Historical Thought: Christian Tradition; Religion: Evolution and Development; Religion, History of; Religion, Sociology of

Bibliography

Atamian S 1955 *The Armenian Community*. Philosophical Library, New York

Beeson T 1982 *Discretion and Valour: Religious Conditions in Russia and Eastern Europe*. Collins, London

Betts R B 1979 *Christians in the Arab East: A Political Study*. SPCK, London

Davis N 1995 *A Long Walk to Church: A Contemporary History of Russian Orthodoxy*. Westview Press, Boulder, CO

Horner N 1989 *A Guide to the Christian Churches in the Middle East*. Mission Focus, Eikhart, IN

Hourani A H 1947 *Minorities in the Arab World*. Oxford University Press, London

Lane C 1978 *Christian Religion in the Soviet Union; A Sociological Study*. State University of New York Press, Albany

Pospielovsky D 1984 *The Russian Church Under the Soviet Regime 1917–1982*. St Vladimir's Seminary Press, Crestwood, NY

Rinvolucri M 1966 *Anatomy of a Church: Greek Orthodoxy Today*. Burns and Oates, London

Saloutos T 1964 *The Greeks in America*. Harvard University Press, Cambridge, MA

Sarkissian K 1968 *The Witness of the Oriental Orthodox Churches*. Beirut

Ware T 1993 *The Orthodox Church*. Penguin Books, Harmondsworth, UK

S. Hackel

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

P

Pacific Islands, Archaeology of

The Pacific Islands, also known as Oceania, have been subdivided traditionally by anthropologists into three main geographic regions: Melanesia, Micronesia, and Polynesia. Following Green (1991), prehistorians now stress the division between *Near Oceania* in the west (including the Bismarck Archipelago and the Solomon Islands), and *Remote Oceania* (which includes all of island Melanesia southeast of the Solomons, along with Polynesia and Micronesia). This latter distinction recognizes the Pleistocene settlement of Near Oceania, whereas the widely dispersed islands of Remote Oceania were discovered and settled only within the past 4,000 years. Archaeological research in Oceania has a long history, but modern efforts emphasizing stratigraphic excavations did not begin until after World War II (Kirch 2000), and have revealed the main chronological sequence for human settlement. This sequence is summarized here, followed by reviews of the development of complex societies in Oceania, and of the human impact on island environments.

1. *Early Human Settlement of Near Oceania*

The oldest known occupation sites are radiocarbon dated to ca. 36,000 years ago (the late Pleistocene), on the large island of New Guinea and in the adjacent Bismarck Archipelago (Allen 1996). At several times during the Pleistocene, New Guinea was joined to Australia as a single land mass (known as Sahul), and human entry into and expansion throughout this vast Australasian region occurred rapidly. Late Pleistocene sites in the Admiralty Islands, New Ireland, and Buka (Solomons), all would have required open ocean transport, suggesting the presence of some form of watercraft (possibly rafts, bark boats, or dugouts) (Irwin 1992).

Early human colonists in Near Oceania were hunters and gatherers, who exploited tropical rainforests as well as inshore marine resources (see *Hunter–Gatherer Societies, Archaeology of*). Long-distance communication and exchange is indicated by the movement of obsidian between islands. By the early Holocene period (after 8000 BC), there is archaeobotanical evidence for domestication of tree, root, and tuber crops (such as the *Canarium* almond and various aroids) within Near Oceania. Archaeological evidence for the cultivation of swamplands at Kuk in the Highlands of New Guinea commences as early as 7000 BC (Golson 1988). These archaeological indications confirm the long-held ethnobotanical hypothesis that Near Oceania was one of several independent centers for the origins of tropical horticulture (see *Food Production, Origins of*).

2. *Austronesian Expansion and Lapita*

During the early Holocene, the southeastern Solomon Islands marked the limit of human expansion. Beginning around 2000 BC, a major expansion or diaspora of people speaking languages belonging to the Austronesian language family commenced (Blust 1995). Their immediate homeland has generally been regarded as comprising the island of Taiwan (and perhaps adjacent areas of mainland China). The ability of early Austronesians to disperse rapidly has been attributed to their invention of the outrigger sailing canoe (Pawley and Ross 1993). The Austronesians were horticulturalists who transported root, tuber, and tree crops via their canoes, along with breeding stocks of domestic pigs, dogs, and chickens.

The Austronesian diaspora rapidly encompassed the major archipelagoes of island Southeast Asia; one branch of Austronesian-speakers expanded along the north coast of New Guinea into the Bismarck Archipelago. This branch is known to linguists as Oceanic, and the Oceanic languages (numbering about 450 modern languages) include most of those spoken throughout the Pacific. The great exception is New Guinea, where roughly 750 non-Austronesian languages are spoken.

Archaeological evidence for the initial Austronesian dispersal into the Pacific comes from both western Micronesia (the Marianas and Palau archipelagoes), and from the Bismarck Archipelago. In western Micronesia, early sites contain red-slipped pottery, some of which is decorated with lime-filled, impressed designs (Rainbird 1994). These sites, along with radiocarbon-dated sediment cores exhibiting signals of human presence (e.g., high influxes of microscopic charcoal resulting from anthropogenic burning) suggest that humans settled, the Marianas and Palau no later than 1500 BC, and possibly as early as 2000 BC.

In the Bismarck Archipelago, the initial Austronesian incursion has been correlated with the appearance of a distinctive suite of sites, also containing pottery with lime-infilled decorations, but with motifs

made largely by a technique of dentate-stamping. These sites and the associated artifacts (such as *Tridacna*-shell adzes and *Trochus*-shell fishhooks, as well as ornaments) represent the earliest known phase of the Lapita cultural complex, dating to ca. 1500–1300 BC (Gosden et al., 1989, Spriggs 1997). Early Lapita sites were frequently hamlets or villages consisting of houses elevated on posts or stilts, situated over tidal reef flats or along shorelines. Excavated plant and animal remains indicated a mixed economy with horticulture and marine exploitation. Substantial quantities of obsidian, chert, pottery, shell artifacts, and other materials were exchanged between communities (Kirch 1997).

Correlations among archaeological assemblages, language groups, and human biological populations are often complex, and need not be isomorphic. However, a strong consensus is emerging among scholars in several disciplines that the initial phase of the Lapita cultural complex can be correlated with the Proto Oceanic interstage of the Austronesian language family. Moreover, genetic evidence (such as mtDNA and hemoglobin markers) supports the view that the Lapita phenomenon reflects a substantial population intrusion into the Bismarck Archipelago, deriving out of island Southeast Asia (Hill and Serjeantson 1989). At the same time, the Proto Oceanic speakers undoubtedly had considerable interaction (cultural, linguistic, and genetic) with the indigenous non-Austronesian speaking populations who already occupied the Bismarck region in the mid-Holocene. Thus the Lapita cultural complex is seen as an outcome of cultural processes of intrusion, integration, and innovation.

3. *Human Colonization of Remote Oceania*

Beginning ca. 1300 BC, the Lapita pottery makers expanded rapidly beyond the Solomons and into the southwestern archipelagoes of Remote Oceania: Vanuatu, the Loyalty Islands, New Caledonia, Fiji, Tonga, and Samoa. Numerous radiocarbon-dated archaeological sites document that Lapita sites in all of these archipelagoes are no later than 900 BC.

We have already noted that the westernmost islands of Micronesia were colonized directly out of island Southeast Asia by Austronesian speakers ca. 2000–1500 BC. Around 2,000 years ago, Oceanic speakers who made plainware pottery (a late form of Lapita) and who used shell adzes, fishhooks, and other implements, founded settlements on several volcanic islands of central Micronesia (Chuuk, Pohnpei, and Kosrae). The atolls of the Marshall Islands were also colonized at this time.

The final stage in the human settlement of the Pacific Islands began after 500 BC, with the Polynesian dispersals eastwards out of Tonga and Samoa. Ancestral Polynesian culture and Proto Polynesian language themselves had developed in this Tonga–Samoa region between ca. 900–500 BC, directly out of the founding Lapita cultural complex (Kirch 2000, Kirch and Green, in press). While archaeologists debate the exact chronology and sequence of Polynesian dispersals, most agree that the central Eastern Polynesian archipelagoes (such as the Society Islands, Cook Islands, and Marquesas Islands) were settled first, no later than AD 300 and perhaps some centuries earlier (Rolett 1998). Remote Easter Island was discovered by AD 800–900 (Van Tilburg 1994), while the Hawaiian Islands were also well settled by this date. The twin large, temperate islands of New Zealand were colonized by Polynesians around AD 1200 (Anderson 1989, Davidson 1984). Critical to the success of this unprecedented diaspora was the double-hulled sailing canoe, capable of carrying 40–60 people on voyages lasting one month or longer (Irwin 1992). That the Polynesians reached South America and returned is suggested by preserved remains of the sweet potato (*Ipomoea batatas*), a South American domesticate, in several prehistoric Polynesian sites.

Because it was the last sector of Remote Oceania to be settled, and because its populations represent a single radiation or diaspora, Polynesia constitutes a monophyletic cultural and linguistic group. Thus, Polynesia has often been regarded as an ideal region for testing models of cultural differentiation from a common ancestor (e.g., Kirch 1984, Kirch and Green, in press).

4. *Development of Complex Societies*

When the eighteenth- to nineteenth-century European voyages of discovery inspired by the Enlightenment encountered Pacific island societies, they frequently encountered large, dense populations organized into complex, hierarchical sociopolitical formations. With populations often numbering into the tens or hundreds of thousands, such societies had two to three decision-making levels, and hereditary leaders who enjoyed elite privileges and status markers. Anthropologists classify such sociopolitical formations as chiefdoms, and indeed, the Polynesian chiefdoms are often considered the archetypal model (see *Chiefdoms, Archaeology of*).

The origins, development, and elaboration of Pacific island chiefdoms have been a major topic of archaeological research (e.g., Davidson 1984, Kirch 1984, Rainbird 1994, Sand 1995, Spriggs 1997). Based on linguistic and archaeological evidence, early Austronesian societies were characterized by some degree of internal ranking (especially between senior and junior branches of a descent line), but were probably heterarchical rather than hierarchical in structure. However, heterarchic competition (in such social arenas as marriage and exchange, as well as competition for land) between social groups provided the

basis for true hierarchy (and eventually, in the largest societies, class stratification) to emerge.

Archaeologists have identified several factors and processes that were significant in the rise of Oceanic chiefdom societies. Many of these were closely linked, and should not be considered unicausal variables. For example, population growth leading to large, high-density populations can be identified as a necessary, but probably not sufficient, cause underlying socio-political complexity (Kirch 2000, Sand 1995). The human populations of the volcanic islands typically reached densities of between 100–250 persons per square kilometer prior to European contact, resulting in intense competition for arable land and other resources. Such conditions encouraged centralized, hierarchic control, as well as providing incentives for militaristic aggression.

A second process linked to population growth and to increased hierarchy was intensification of production, including agricultural systems and other forms of production or resource extraction, as well as economic specialization (e.g., in pottery production and trade). On many islands, large-scale irrigation works or dryland field systems were developed during late prehistory. Although population increases may have initially spurred intensification, once in place such intensive production systems provided a means for surplus extraction by chiefs and other leaders, thus encouraging hierarchy. Often culturally marked as tribute, such surpluses were the economic foundation of an emergent elite, including not only hereditary chiefs, but priests, warriors, craft specialists, and others.

Ideology likewise played a key role in Pacific island societies, with the elite cadres of the larger and most complex societies actively employing ideological control as a means of legitimation. The origins of Oceanic ritual systems can be traced back to common Austronesian concepts of the sacredness of ancestors; these concepts later became elaborated as cults in which the highest chiefs were seen as directly descended from powerful gods, and hence essential to the continued well-being of the society at large. Archaeologically, the rise of elite-dominated ideological systems is particularly reflected in monumental architecture, of which the most impressive examples are the giant statue-bearing temples of Easter Island, and the site of Nan Madol on Pohnpei. Other forms of monumental architecture, however, are ubiquitous throughout Pacific islands chiefdoms. Even when monumental architecture is absent, material signs of ideological control can be quite striking, as in the multiple sacrificial interments associated with the burial of Roy Mata, a chief of Vanuatu (Garanger 1972).

Finally, competition, conflict, and warfare also characterized many of the complex societies of the Pacific, especially following the rise of large and dense populations (see also *Cultural Resource Management (CRM): Conservation of Cultural Heritage*). Archaeologically, warfare is marked by a diversity of kinds of fortifications, such as the pallisaded *pa* volcanic cones and headlands of New Zealand, or the ring-ditch fortified villages of Fiji. Another, more gruesome signal of the levels that inter-societal aggression reached on some islands is cannibalism (or para-cannibalistic treatment of enemies, such as dismembering, roasting, and the nonfunerary discard of skeletal remains). Although some anthropologists have expressed skepticism regarding the accounts of cannibalism in the Pacific by European voyagers, there is now direct archaeological evidence for cannibalistic or para-cannibalistic practices in late prehistory on Easter Island, the Marquesas, New Zealand, Mangaia, and Fiji.

5. *Human Impacts to Island Ecosystems*

The islands of Remote Oceania, due to isolation and related factors, provide model conditions for studying the effects of human colonization and land use on pristine ecosystems. Interdisciplinary research among archaeologists and natural scientists (see also *Environmental Archaeology*) over the past three decades has amplified our understanding of such human–-ecosystem interactions (Kirch and Hunt 1997).

Because of the substantial open-ocean distances isolating them from continents as well as other islands, and the difficulty of dispersal to islands, prior to human arrival oceanic ecosystems were typically characterized by: high species-level endemicity, but lower diversity in higher-level (generic and family) taxa; lowered competition; survival of archaic forms; and vulnerability to disturbance from outside agents. Larger vertebrates such as marsupials (wombats, cuscus) and rats, snakes, frogs, and most lizards were restricted primarily to Near Oceania, with only a handful of species declining in numbers eastwards to Fiji and Samoa. (The reef and marine resources of Pacific islands also display a west-to-east decline in species diversity.) Throughout most of Remote Oceania, prehuman vertebrate faunas were dominated by birds (including many flightless forms which had evolved *in situ* from flighted ancestors). Prior to human arrival, these bird populations lacked large vertebrate predators, and presumably also a typical predator avoidance strategy. They must have been extremely easy prey for the first humans to step foot on these islands.

When humans first arrived in Remote Oceania, they typically found the islands to be forested, and inhabited by a range of largely endemic species, dominated by birds, along with invertebrates such as land snails and insects. Oceanic peoples possessed a successful colonization strategy that allowed them to exist on isolated islands, by: (a) transporting in their sailing canoes stocks of horticultural crop plants, along with domestic pigs, dogs, and chickens (rats came along,

presumably as 'stowaways'); (b) clearing areas of rainforest for garden land; and (c) intensively exploiting the abundant natural avifaunal and marine resources.

This colonization strategy had several consequences for island ecosystems, all of which are increasingly well documented through both archaeological and paleoenvironmental indicators. Forest clearance on many islands is signaled in changing pollen spectra from sediment cores, with tree taxa rapidly giving away to ferns and grasses; also characteristic are sharp increases in microscopic charcoal influxes, indicating human-induced burning, in most cases probably associated with shifting cultivation. On some islands, forest clearance led to increased erosion rates, along with alluviation of valley bottoms or along coastal plains. The exploitation of natural resources is particularly evident in the zooarchaeological assemblages from early settlement sites, which are characterized by high numbers of land and seabirds, many of them representing now extinct or extirpated species (Steadman 1995). A dramatic case of avifaunal extinctions on Pacific islands is that of the moa, a group of 13 species of large, flightless birds which became totally extinct in New Zealand during the brief period of Polynesian occupation (Anderson 1989).

Cumulative effects of human actions on islands led to irreversible changes, such as dramatic declines in biodiversity, and the conversion of natural rainforests to intensively managed, anthropogenic landscapes. The consequences for human populations themselves were undoubtedly mixed. The replacement of natural ecosystems with intensive food production systems enabled the growth of large and dense human populations. At the same time, reduction or depletion of natural resources, coupled with the necessity for intensive land use, encouraged highly complex sociopolitical systems which at times competed fiercely for control of land and the means of production.

See also: Australia and New Guinea, Archaeology of; Melanesia: Sociocultural Aspects; Polynesia and Micronesia: Sociocultural Aspects; Southeast Asia, Archaeology of

Bibliography

Allen J 1996 The pre-Austronesian settlement of island Melanesia: Implications for Lapita archaeology. *Transactions of the American Philosophical Society* **86**(5): 11–27

Anderson A 1989 *Prodigious Birds: Moas and Moa Hunting in Prehistoric New Zealand*. Cambridge University Press, Cambridge, UK

Blust R 1995 The prehistory of the Austronesian-speaking peoples: The view from language. *Journal of World Prehistory* **9**: 453–510

Davidson J M 1984 *The Prehistory of New Zealand*. Longman Paul, Auckland, New Zealand

Garanger J 1972 *Archéologie des Nouvelles-Hébrides*. Publication de la Société des Océanistes No. 30, Paris

Golson J 1988 The origins and development of New Guinea agriculture. In: Harris D, Hillman G (eds.) *Foraging and Farming: The Evolution of Plant Exploitation*. Unwin Hyman, London

Gosden C, Allen J, Ambrose W, Anson D, Golson J, Green R, Kirch P V, Lilley I, Specht J, Spriggs M 1989 Lapita sites of the Bismarck Archipelago. *Antiquity* **63**: 561–586

Green R C 1991 Near and Remote Oceania: Disestablishing "Melanesia" in culture history. In: Pawley A (ed.) *Man and a Half: Essays in Pacific Anthropology and Ethnobiology in Honour of Ralph Bulmer*. The Polynesian Society, Auckland, New Zealand

Hill A V S, Serjeantson S W (eds.) 1989 *The Colonization of the Pacific: A Genetic Trail*. Clarendon Press, Oxford, UK

Irwin G 1992 *The Prehistoric Exploration and Colonisation of the Pacific*. Cambridge University Press, Cambridge, UK

Kirch P V 1984 *The Evolution of the Polynesian Chiefdoms*. Cambridge University Press, Cambridge, UK

Kirch P V 1997 *The Lapita Peoples: Ancestors of the Oceanic World*. Blackwell Publishers, Oxford, UK

Kirch P V 2000 *On the Road of the Winds: An Archaeological History of the Pacific Islands Before European Contact*. University of California Press, Berkeley, CA

Kirch P V, Green R C, in press, *Hawaiki, Ancestral Polynesia: An Essay in Historical Anthropology*. Cambridge University Press, Cambridge, UK

Kirch P V, Hunt T L (eds.) 1997 *Historical Ecology in the Pacific Islands: Prehistoric Environmental and Landscape Change*. Yale University Press, New Haven, CT

Pawley A K, Ross M 1993 Austronesian historical linguistics and culture history. *Annual Review of Anthropology* **22**: 425–59

Rainbird P 1994 Prehistory in the northwest tropical Pacific: The Caroline, Mariana, and Marshall Islands. *Journal of World Prehistory* **8**: 293–349

Rolett B V 1998 *Hanamiai: Prehistoric Colonization and Cultural Change in the Marquesas Islands (East Polynesia)*. Yale University Publications in Anthropology No. 84, New Haven, CT

Sand C 1995 "Le Temps d'Avant": *La Préhistoire de la Nouvelle-Calédonie*. L'Harmattan, Paris

Spriggs M J T 1997 *The Island Melanesians*. Blackwell, Oxford, UK

Steadman D W 1995 Prehistoric extinctions of Pacific island birds: Biodiversity meets zooarchaeology. *Science* **267**: 1123–30

Van Tilburg J A 1994 *Easter Island: Archaeology, Ecology and Culture*. British Museum Press, London

P. V. Kirch

Pain, Health Psychology of

1. Basic Terms

Pain is an adaptive phenomenon that signals impending danger to the body. Pain can, however, also be maladaptive, for example in states of chronic pain, and thus can itself become a significant disorder. Epidemiological studies show that 80 percent of the population experience recurrent pain, with more than 10 percent of the population being permanently disabled by it. Although pain has for a long time been

viewed a purely sensory phenomenon or an epiphenomenon of a medical disorder, this view changed during the last 40 years of the twentieth century. It has been recognized that pain is a psychobiological experience, with emotional aspects being as important as sensory pain characteristics. Thus nociception, the physiological process of a noxious signal being transmitted from the periphery to the brain, has been differentiated from the experience of pain, which encompasses all aspects including psychological, social and cultural factors. The international Association for the Study of Pain (IASP) definition takes this shift from a biomedical to a biobehavioral or psychobiological perspective into account by characterizing pain as an 'unpleasant sensory and emotional experience associated with actual or potential tissue damage or described in terms of such damage' (Merskey 1986). This definition was prompted by the fact that pain may often be experienced in the absence of identifiable objective pathology. A major change in the traditional view of pain was the gate control theory of pain proposed in Melzack and Wall (1965). More important than its physiological aspects, which have not all been confirmed, was the conceptual model behind it. The gate control theory stated not only that pain could be modulated by both afferent and efferent factors at the level of the spinal cord, but also emphasized that pain has sensory–discriminative, motivational–emotional, and cognitive–evaluative components as well, and thus it assigns an important role to psychological factors. A core feature of this model is the integration of peripheral stimuli with psychological variables such as mood, attention, or cultural experience in the perception of pain. Thus, the gate control theory has been instrumental in abolishing the dichotomy of psychogenic versus somatogenic pain—psychological and physiological factors always interact in the production of the experience of pain and are not mutually exclusive causes of pain.

2. *Classification of Pain*

The common distinction between acute and chronic pain—referring to pain of at least three to six months duration and/or pain that exceeds the normal time for healing in the case of an acute injury—is useful, because chronic pain is often maladaptive and needs special attention. Whereas acute pain such as pain related to medical procedures or childbirth usually leads to anxiety and apprehension, chronic pain tends to be associated with helplessness, depression, and irritability as well as interference with family function, work, or social interaction (Gatchel and Turk 1999). In addition to the large number of workdays that are lost as a consequence of it, chronic pain is the leading cause of invalidity in the age under-50s group and causes enormous costs for the health care system (Melzack and Wall 1994).

The IASP classification of pain is organized in five general categories: (a) the bodily system affected, (b) time characteristics, (c) intensity of pain (d) duration of pain, and (e) its presumed etiology. This classification lacks reliability especially in respect of the last characteristic. In addition, it neglects psychological factors that are only coded with respect to the presence of a psychiatric disorder or within the category 'psychophysiological dysfunction.' *ICD-10* as well as *DSM-IV* still adhere to the distinction between psychogenic–somatogenic or physical vs. somatoform pain, which must be viewed as obsolete given the large number of research findings emphasizing the importance of psychological factors in any type of pain, both acute and chronic (Gatchel and Turk 1999).

The biopsychosocial model of pain views pain as a complex response that can be described on the verbal–subjective, the motor–behavioral and physiological levels. Pain can be based on nociceptive input, but nociception is not a prerequisite of the experience of pain, which can be an exclusively central phenomenon, but always has physiological antecedents and consequences (Flor et al. 1990). The multiaxial classification of pain as proposed by Turk and his colleagues (cf. Turk and Rudy 1988) categorizes pain patients along both the somatic and the psychosocial dimensions. This classification has yielded three subgroups of patients in the psychosocial domain characterized as adaptive copers, interpersonally distressed, and dysfunctional, with significant differences in the course of the illness and response to treatment.

3. *The Role of Learning in Chronic Pain*

Both associative and nonassociative learning processes as well as social learning have been found to be of fundamental significance for the development of chronic pain. The repeated application of painful stimuli leads to reduced responsivity, i.e., habituation to the painful stimulation. In many states of chronic pain, sensitization rather than habituation occurs due to changes both at the level of the receptor and of the central nervous system (Woolf and Mannion 1999). Sensory information accelerates habituation and reduces activation caused by surprise, insecurity, and threat. This mechanism may underlie the effects reported in a large number of studies that support the positive results of preparatory information prior to acutely painful procedures such as surgery or bone marrow aspiration.

The most influential model of psychological factors in chronic pain was Fordyce's assumption that chronic pain can develop and be maintained due to operant conditioning of pain behaviors, i.e., overt expressions of pain. Fordyce (1976) postulated that acute pain behaviors such as limping or moaning may come under the control of external contingencies of reinforcement, and thus develop into a chronic pain

problem. Positive reinforcement of pain behaviors (e.g., by attention and the expression of sympathy), negative reinforcement of pain behaviors (e.g., by attention and the expression of sympathy), negative reinforcement of pain behaviors (e.g., the reduction of pain through the intake of medication or by the cessation of activity) as well as a lack of reinforcement of healthy behaviors could provoke chronicity in the absence of nociceptive input. Thus, pain behaviors, originally elicited by nociceptive input, may, over time, occur in response to environmental contingencies. This model has generated much research, which has not only confirmed the original assumptions made by Fordyce; it has also been shown that in addition to pain behaviors the subjective experience of pain, as well as physiological responses related to pain, are subject to reinforcement learning.

A special role has been assigned to the 'significant others' of chronic pain patients, who have high reinforcement potential. When solicitous, i.e., pain-reinforcing, spouses were present, several studies found that the patients were more pain-sensitive during acute pain tests than when the spouses were absent. The patients with nonsolicitous spouses did not differ whether the spouse was present or absent. These studies suggest that spouses can serve as discriminative stimuli for the display of pain behaviors by chronic pain patients, including their reports of pain intensity. Health care providers may also serve as discriminative cues influencing patients' responses. Solicitous responses by significant others can also lead to increased physiological pain responses in the patients, whereas the focusing on healthy behaviors by significant others can have positive effects on the pain experience.

Of equal importance is the operant conditioning related to the intake of pain medication. Patients are often told by their physicians or by well-meaning family members that they should not continue to take analgesic medication unless the pain increases to a point where it becomes intolerable (referred to as *prn* from the Latin, 'take as needed'). When pain medication is taken at that stage, both pharmacological and behavioral factors can contribute to the development of the misuse of medication and, in severe cases, even drug dependence. If analgesic medication is taken only at peak pain levels, the effect of the medication is less potent, and patients cycle between high and low levels of medication that facilitate the development of dependence. In addition, medication intake is negatively reinforcing, since the intake of medication ends an aversive state (pain). Subsequently, the pain-reducing behavior (use of analgesics) increases in frequency. Thus, both pharmacotherapists and behavioral psychologists recommend that analgesic medication should be taken not in a pain-contingent but rather in a time-contingent fashion, adapted to the specific pain level of the patient and to the half-life of the drug.

The negative reinforcement of activity levels is an important process in the development of disability. A specific activity, for example walking, is performed until pain sets in, at which point the activity is interrupted and replaced by rest. Subsequently, the pain will be reduced. The reduction of an aversive state—pain—negatively reinforces the cessation of activity. As was the case with the intake of analgesic medication, the cessation of activity has to be made dependent on the amount of activity achieved—quota-based (e.g., number of stairs climbed, distance walked) rather than on the amount of pain. Thus, the pain-reinforcing quality of rest is eliminated. This formulation supports the strategy of encouraging patients to perform activities to meet a specific quota, and not until pain is perceived as overwhelming.

The respondent conditioning model postulates that numerous, formerly neutral, cues can be associated with the experience of pain and can—over time—themselves elicit responses that lead to an increased pain response, and can create the experience of pain in the absence of nociceptive input. To illustrate this process, the patient may have learned to associate increases in muscle tension with all kinds of stimuli that were originally associated with nociceptive stimulation. Thus, sitting, walking, bending, or even thoughts about these movements, may elicit anticipatory anxiety and an increase in muscle tension. This fear of movement or 'kinesophobia' has been discussed as an important factor in the maintenance and exacerbation of chronic pain. Subsequently, patients may display maladaptive responses to any number of stimuli and reduce the frequency of performance of many activities other than those that initially reduced pain. Thus, although the original association between injury and pain results in anxiety regarding movement, over time the anxiety may lead to increased muscle tension and pain even if the nociceptive stimuli are no longer present. In addition, stress situations can increase muscle tension levels and cause sympathetic activation and may, thus, reinforce this process.

Many patients report that an acute pain problem evolved into chronic pain at a time where personal stressors co-occurred with the pain. Stress situations may serve as additional unconditioned stimuli and also as conscious stimulus for muscle tension increases, increased sympathetic activation, and subsequently pain. Nonoccurrence of pain is a powerful reinforcer for reduction of movement. Thus, the original respondent conditioning may be complemented by an operant process, whereby the nociceptive stimuli need no longer be present for the avoidance behavior to occur. People who suffer from acute back pain, regardless of the cause, may adopt specific behaviors (e.g., limping) to avoid pain and they may never obtain 'corrective feedback' because they fail to perform more natural movements. Reduction in physical activity may subsequently result in muscle atrophy and increased disability. In this manner, the physical

abnormalities proposed by biomechanical models of pain may actually be secondary to changes in behavior initiated through learning. Chronic pain patients tend to focus their attention on impending pain and subsequently avoid many types of activity, thus fostering the development of disability and depression. The release of endogenous opioids—the body's own analgesic system—may also be influenced by respondent conditioning as well as brain responses related to the experience of pain. Moreover, fear of pain and subsequent avoidance of activity is one of the best predictors of subsequent chronicity (Asmundson et al. 1999).

Response acquisition through the observation of others and modeling is an essential mechanism of learning new patterns of behavior. Children acquire attitudes about health and health care, and the perception and interpretation of symptoms and physiological processes from their parents and their social environment. They also learn appropriate responses to injury and disease, and thus may be more or less likely to ignore or over-respond to normal bodily sensations they experience. The culturally-acquired perception and interpretation of symptoms determines how people deal with illness; the sight of others in pain is an event that captures attention and may have survival value, may help to avoid experiencing more pain and help to learn what to do about acute pain. Modeling probably plays a part in the phenomenon of 'pain-prone families' families with a significantly elevated occurrence of pain problems (Whitehead et al. 1994). It was, for example, reported that children show the same pain syndromes as their parents currently have, rather than the pain problems their parents had in their own childhood. The large cultural variations in pain expression are also important. In common clinical practice, the acquisition or extinction of pain-related behavior by means of modeling has received little attention. However, there are occasional indications for the role of modeling in treating pain problems in children in burn units and in the treatment of postoperative pain.

Despite the great deal of data available on the modification of experimentally induced pain behavior by means of modeling in normal (healthy) subjects, there are few experimental results concerning chronic pain patients. Nor are there any longitudinal studies of the development of pain syndromes in 'pain-prone families.' Further investigation is necessary for the evidence of modeling as a factor in the development of chronic pain disorders.

4. *Cognitive Factors and Pain*

Cognitive–behavioral models of chronic pain emphasize that the evaluation of the pain experience by the patient greatly determines the amount of pain that is experienced as well as its negative consequences (Turk et al. 1983). General assumptions that characterize the cognitive–behavioral perspective are: (a) people are active processors of information, not passive reactors; (b) thoughts (e.g., appraisals, expectancies) can elicit or modify mood, affect physiological processes, influence the environment, and serve as impetus for behavior. Conversely, mood, physiology, environmental factors, and behavior can influence thought processes; (c) behavior is reciprocally determined by the person and by environmental factors; (d) people can learn more adaptive ways of thinking, feeling, and behaving; and (e) people are capable and should be involved as active agents in change of maladaptive thoughts, feelings, and behaviors.

From the cognitive–behavioral perspective, people suffering from chronic pain are viewed as having negative expectations about their own ability to control certain motor skills such as performing specific physical activities (e.g., climbing stairs, lifting objects) that are attributed to one overwhelming factor, namely, a chronic pain syndrome. Moreover, chronic pain patients tend to believe that they have a limited ability to exert any control over their pain. Such negative, maladaptive appraisals about their situation and personal efficacy may reinforce the experience of demoralization, inactivity, and overreaction to nociceptive stimulation.

A great deal of research has been directed toward identifying cognitive factors that contribute to pain and disability. These have consistently demonstrated that patients' attitudes, beliefs, and expectancies about their plight, themselves, their coping resources, and the health care system affect their reports of pain, activity, disability, and response to treatment.

Pain, when interpreted as signifying ongoing tissue damage or a progressive disease, seems to produce considerably more suffering and behavioral dysfunction than if it is viewed as being the result of a stable problem that is expected to improve. A number of studies using experimental pain stimuli demonstrated that the conviction of personal control can ameliorate the experience of experimentally-induced nociception. Moreover, the type of thoughts employed during exposure to painful stimulation has been related to pain tolerance and pain intensity ratings. Catastrophizing thoughts have been associated with lower pain tolerance and higher ratings of pain intensity. In contrast, coping thoughts have been related to higher pain tolerance and lower pain intensity ratings.

Certain beliefs may lead to maladaptive coping, increased suffering, and greater disability. Patients who believe their pain is likely to persist may be passive in their coping efforts and fail to make use of available strategies, even when in their repertoire, to cope with pain. Patients who consider their pain to be an 'unexplainable mystery,' may negatively evaluate their own ability to control or decrease pain, and are less likely to rate their coping strategies as effective in controlling and decreasing pain.

Once beliefs and expectancies (cognitive schemata) about a disease are formed they become stable and very difficult to modify. Patients tend to avoid experiences that could invalidate their beliefs, and they adapt their behavior in accordance with these beliefs, even in situations where the belief is no longer valid (no corrective feedback is received to discredit this belief). For example, feeling muscular pain following activity may be caused by lack of muscle strength and general deconditioning and not by additional tissue damage.

Self-regulation of pain and its impact depends upon a person's specific ways of dealing with pain, adjusting to pain, and reducing or minimizing pain and distress caused by pain—their coping strategies. Coping is assumed to be manifested by spontaneously-employed purposeful and intentional acts, and it can be assessed in terms of overt and covert behaviors. Overt, behavioral coping strategies include rest, medication, and use of relaxation. Covert coping strategies include various means of distracting oneself from pain, reassuring oneself that the pain will diminish, seeking information, and problem solving. Studies have found active coping strategies (efforts to function in spite of pain, or to distract oneself from pain such as activity, or ignoring it) to be associated with adaptive functioning; and passive coping strategies (depending on others for help in pain control and restricted activities) to be related to greater pain and depression. However, beyond this there is no evidence supporting the greater effectiveness of one active coping strategy compared to any other (Fernandez and Turk 1989). Specific coping strategies need not always be adaptive or maladaptive. It seems more likely that different strategies will be more effective than others for some people at some times, but not necessarily for all people all of the time, or even the same person at different times.

5. *Pain and Affective Factors*

The affective factors associated with pain include many different emotions, but they are primarily negative in quality. Anxiety and depression have received the greatest amount of attention in chronic pain patients; however, anger has recently received considerable interest as an important emotion in chronic pain patients. Research suggests that from 40 to 50 percent of chronic pain patients suffer from depression (Romano and Turner 1985). There have been extensive and fruitless debates concerning the causal relationship between depression and pain. In the majority of cases, depression appears to be the patients' reaction to their plight. The presence of depression is closely related to the feelings of loss of control and helplessness often associated with pain.

Several investigators have also found a close association between fear of pain and dysfunctional coping. In addition, high comorbidity between anxiety disorders and pain seems to be present. Muscular hyperreactivity to stress seems to be closely associated with fear of pain.

Anger has been widely observed in individuals with chronic pain. The internalization of angry feelings seems to be strongly related to measures of pain intensity, perceived interference, and reported frequency of pain behaviors. Anger and hostility are closely associated with pain in persons with spinal cord injuries. Frustrations related to persistence of symptoms, limited information on etiology, and repeated treatment failures along with anger toward employers, the insurance, the health care system, family members, and themselves, also contributed to the general dysphoric mood of these patients. The impact of anger and frustration on exacerbation of pain and treatment acceptance has not received adequate attention. It would be reasonable to expect that the presence of anger may serve as an aggravating factor, associated with increasing autonomic arousal, and having the effect of blocking motivation and acceptance of treatment oriented toward rehabilitation and disability management rather than cure, which are often the only treatments available for chronic pain.

6. *Biobehavioral Perspective*

A biobehavioral perspective of chronic pain needs to consider the factors discussed above and their mutual interrelationships in the explanation of chronic pain. The existence of a physiological disposition or diathesis is one important component. This predisposition is related to a reduced threshold for nociceptive stimulation and can be determined by genetic factors or acquired through early learning experiences. For example, Mogil (1999) showed that large genetic variations in individual pain sensitivity exist. Very impressive evidence for the role of early traumatic experience comes from the work of Anand et al. (1999), who showed that minor noxious experience in neonate rats leads to dramatic alterations (sensitization) in nociceptive processing in the adult organism. A further component of the biobehavioral model is a response stereotype of a particular bodily system, such as exaggerated muscular responses of the lower back muscle to stress and pain that is based both on the diathesis and on aversive experiences present at the time of the development of the response. These aversive stimuli may include personal or work stress or problematic occupational conditions and will lead not only to painful responses but also to avoidance behaviors and associated maladaptive cognitive and affective processes. The cognitive evaluation of these external or internal stimuli is of great importance in the pain response, as discussed above.

7. *Memory for Pain*

An important maintaining factor in this chronicity process is the development of pain memories. These pain-related memories may be explicit or implicit and may subsequently guide the patient's experience and behaviors (Erskine et al. 1990). For example, pain patients have a tendency to remember preferentially negative and pain-related life events and show a deficit in the retrieval of positive memories. The experience of chronic pain also leads to the development of somatosensory pain memories for example, an expanded representation of the affected body part in the primary somatosensory cortex. This expanded cortical representation is accompanied by increased sensitivity to both painful and nonpainful stimuli, and may be further enhanced by learning processes or attention to painful stimulation. A dramatic example of a learned memory for pain has been found in phantom limb pain patients (Flor et al. 1995). In upper extremity amputees, the magnitude of the phantom limb pain was found to be proportional to the amount of reorganization in the primary somatosensory cortex, namely, the shift of the cortical mouth representation into the area where the amputated limb was formerly represented. The brain obviously maintains a memory of the former input to the deafferented area. Subsequently, stimulation stemming from areas adjacent to the deafferented zone elicit sensations and pain in the now absent limb. Phantom sensations and cortical reorganization are absent in congenital amputees. The focus of the biobehavioral perspective is thus on the patient and not just on the symptoms or the underlying pathology. This focus also requires that the treatment of the patients be tailored not only to medical factors, but that it incorporates psychosocial variables that may often be predominant in states of chronic pain (see *Chronic Pain: Models and Treatment Approaches*).

8. *Future Perspectives*

At the end of the twentieth century, there has been an enormous increase of research in the neurosciences, where the mutual interactions of molecular and genetic as well as neuronal and behavioral variables have begun to be examined. This integration of the psychosocial and biomedical perspective is much needed and will bring substantial advances in our understanding of the phenomenon of pain. A further important development is the recent finding in neuroscience that the adult brain is enormously plastic and that learning processes as well as training measures have great impact on these plastic changes (Recanzone 2000). This opens the door for new approaches to treatment. It will also aid in delineating and eliminating the factors that lead to resistance to the treatment of chronic pain syndromes.

See also: Chronic Illness, Psychosocial Coping with; Chronic Pain: Models and Treatment Approaches; Pain, Management of; Stress and Coping Theories; Stressful Medical Procedures, Coping with

Bibliography

Anand K J, Coskun V, Thrivikraman K V, Nemeroff C B, Plotsky P M 1999 Long-term behavioral effects of repetitive pain in neonatal rat pups. *Physiol. Behav.* **66**(4): 627–37

Asmundson G J, Norton P J, Norton G R 1999 Beyond pain: The role of fear and avoidance in chronicity. *Clinical Psychology Review* **19**: 97–119

Erskine A, Morley S, Pearce S 1990 Memory for pain: A review. *Pain* **41**(3): 255–65

Fernandez E, Turk D C 1989 The utility of cognitive coping strategies for altering pain perception: A meta-analysis. *Pain* **38**(2): 123–35

Flor H, Birbaumer N, Turk D C 1990 The psychobiology of pain. *Advances in Behav Res Ther.* **12**(2): 47–84

Flor H, Elbert T, Wienbruch C, Pantev C, Knecht S, Birbaumer N, Larbig W, Taub E 1995 Phantom limb pain as a perceptual correlate of cortical reorganization—following arm amputation. *Nature* **375**: 482–4

Fordyce, W E 1976 *Behavioral Methods for Chronic Pain and Illness*. Mosby, St. Louis, MO

Gatchel R J, Turk D C 1999 *Psychosocial Factors in Pain: Critical Perspectives*. Guilford Press, New York

Melzack R A, Wall P D 1965 Pain mechanisms: A new theory. *Science* **150**: 3699–709

Melzack R A, Wall P D (eds.) 1994 *The Textbook of Pain*, 3rd edn. Churchill Livingstone, Edinburgh, UK

Merskey H 1986 Classification of chronic pain: Descriptions of chronic pain syndromes and definitions of pain terms. *Pain* **3**: 1–225

Mogil J S 1999 The genetic mediation of individual differences in sensitivity to pain and its inhibition. *Proceedings of the National Academy of Sciences of the United States of America* **96**(14): 7744–51

Recanzone G 2000 Cerebral cortex plasticity: Perception and skill acquisition. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Boston, MA

Romano J M, Turner J A 1985 Chronic pain and depression: Does the evidence support a relationship? *Psychology Bulletin* **97**(1): 18–34

Turk D C, Meichenbaum D, Genest M 1983 *Pain and Behavioral Medicine: A Cognitive–Behavioral Perspective*. Guilford Press, New York

Turk D C, Rudy T E 1988 Toward an empirically-derived taxonomy of chronic pain patients: integration of psychological assessment data. *Journal of Consulting and Clinical Psychology* **56**: 233–8

Whitehead W E, Crowell M D, Heller B R, Robinson J C, Schuster M M, Horn S 1994 Modeling and reinforcement of the sick role during childhood predicts adult illness behavior. *Psychosomatic Medicine* **56**: 541–50

Woolf C J, Mannion R J 1999 Neuropathic pain: Aetiology, symptoms, mechanisms, and management. *Lancet* **353**: 1959–64

H. Flor

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Pain, Management of

1. The Physiology of Pain

Pain is defined as 'an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage.' It is the most common complaint voiced by patients when they seek medical attention. The ability to perceive pain serves a protective function for the individual. The sensation warns of impending injury and elicits reflex and behavioral responses that keep tissue damage to a minimum (see Livingston 1998, Melzack 1973, Rey 1995).

In many instances, noxious stimuli applied at the skin surface (blunt trauma, burns, a surgical incision) serve to activate pain receptors. These are special nerve cells, termed nociceptors, which convert (transduce) the energy delivered at the periphery to neural impulses. Transduction is the first of four processes allowing the energy that produces noxious stimuli to result eventually in the perception of pain. Once this initial energy is transduced into electrical impulses, nerves carry the stimulus to higher levels of the nervous system. This activity is termed *transmission*, and it allows the noxious stimuli to be conveyed from the skin to the spinal cord as a series of electrical messages. As the signals arrive in the dorsal horn of the spinal cord, they are altered by other impulses within the cord and by descending impulses from higher neural centers. This is termed *modulation*. These modified and processed impulses continue upward towards the brain and allow a person to experience the perception of pain. This is an individual and highly subjective phenomenon. Another person can not reliably measure it. There is no objective appraisal of the sensation that an individual calls pain.

An understanding of the four events is critical to an appreciation of the mechanisms by which pain can be managed. The physician targets therapies to prevent the electrical signals from reaching the cortex and limbic systems of the brain where pain enters a person's consciousness. Thus treatments are aimed at interrupting pain information as they are transduced to nerve impulses; transmitted towards the spinal cord; and modulated within the cord; or by modifying the neural impulses as they enter the brain where the actual perception occurs.

Minimizing the pain a patient perceives has value on a number of fronts: humanitarian, physiologic, functional, and economic. Progress made in this area of medicine was begun by anesthesiologists in the 1980s when they began to organize services to treat surgical inpatients immediately following operations. These activities were logical extensions of the work they already performed during surgery. Acute pain services are now well established divisions of most hospitals in the United States.

Workers in the field of pain management divide pain into three separate categories: acute, chronic non-malignant, and cancer pain. Management varies depending on the etiology. Acute pain is experienced following trauma, elective surgical procedures, or the exacerbation of chronic medical conditions (e.g., sickle cell crisis). Tissue damage initiates a cascade of physiological and behavioral changes. These have the potential to progress into chronically painful states if left unattended.

1.1 Acute Pain

Mechanisms to manage operative and postoperative pain focus on the delivery of medications to specific sites within the body. Local anesthetics, such as lidocaine can be injected under the skin prior to an incision. This is exemplified by the use of a nerve field block performed immediately before an inguinal hernia repair. The medication is injected at the site of the surgery and given time to work. Local anesthetics act by blocking the conduction of impulses down nerve fibers. These nerves, then, cannot transmit to the central nervous system (CNS) the information that tissue damage (the surgical incision) is occurring. Since the CNS never receives the nerve impulses, the patient does not perceive pain.

More sophisticated techniques include the deposition of local anesthetics into areas wherein numerous nerves converge. Because these areas are densely packed with passing nerve fibers, a small amount of anesthetic can effectively block pain messages from a large segment of the body. Examples of these procedures include brachial plexus (an area of convergent nerves from the arm and hand) blockade prior to hand surgery, and epidural (a space just outside the spinal cord) blockade, which is used to blunt the intensity of labor pains prior to childbirth. One problem of the use of local anesthetic is that these agents are generally fairly short-acting. The duration of the blockade can be prolonged by co-injection of agents which decrease blood flow in the immediate area of injection. Less blood flow to the area of blockade means the anesthetic agent is taken away by the bloodstream more slowly, prolonging the presence, and effect, of the local anesthetic. Another method physicians use to prolong local anesthetic action is to pump anesthetic continuously into the needed area through a plastic tube or catheter that has been inserted into that region. This technique allows pain control for very long periods of time. For example, a catheter placed in a position that blocks pain from the chest (thorax) can be used to manage pain related to chest (thoracic) surgery for five days following the operation. A catheter is placed at the appropriate level within the epidural space and taped to the patient's skin. An infusion pump is programmed to deliver local anesthetic, opioid (such as morphine), or a mixture of both agents at an hourly

rate. When the mixture of agents is used, very dilute concentrations of each are employed. Inhibiting the transmission and modulation of nerve impulses with different classes of agent allows for synergy of effect. Excellent control of very strong pain can be achieved with the use of epidural catheter placement and the constant infusion of analgesic mixtures. Sensitivity of the affected area returns quickly when the drugs are stopped.

Unfortunately, local anesthetics block not just pain sensors, but all nerves. Thus, they produce a profound numbness and, in some cases, difficulty in moving muscles. In addition, they are only active for a short time, or require the placement of a catheter near the nerves to be affected. Therefore, other means are necessary to deal with more long-duration pain. For example, systemic medications can be delivered intravenously to blunt the perception of pain. Newer technology in the form of microprocessors and physically small infusion pumps have made the management of acute, post-surgical pain more precise. Patient-controlled analgesia (PCA) pumps provide an infusion of morphine or similar drugs at a constant background rate. This is usually 1 or 2 milligrams of morphine over the course of an hour. Furthermore, a patient can activate the pump to deliver additional morphine by the push of a button. Dosage is often 2 milligrams which can be repeated every sixth minute. This machinery allows the patient substantial control of how much morphine he or she is receiving so that the dose delivered matches his or her needs. The background infusion guarantees that some analgesia is delivered to the bloodstream to compensate for normal clearance of the analgesic by the liver and kidneys. The patient-controlled aspect enables the patient to raise immediately the blood level of the morphine should the perception of pain increase in intensity. This commonly correlates with the patient's degree of physical activity and can often be anticipated. A patient who is comfortable while lying down may feel pain when getting into a chair. In this instance, self-dosing permits the level of opioid to match the need for analgesia on a minute by minute basis. Since the pump must be activated before delivering additional medication, the risk of overdosage is minimized. Patients who are sleepy from a high serum level of morphine will not push the button for additional medication. In some instances, morphine is not the appropriate agent to use as it produces significant side-effects in some patients. In these cases, other opioids such as fentanyl, demerol, or dilaudid can be used. Individual preferences and the reactions to treatment are used to guide further therapy.

Progress has been made in the development of nonsteroidal, anti-inflammatory drugs (NSAIDs). Aspirin, acetomenophin, and ibuprofin are well known examples of these drugs. NSAIDs are a class of medications that diminish swelling and decrease the perception of pain. They act both at painful sites and in the CNS by inhibiting the enzyme cyclo-oxygenase. Intramuscular and intravenous NSAIDs have been developed so that patients in the immediate postoperative period and still unable to consume pills can benefit from the agents. The newest drugs in the class are highly selective for the desired effects and show little of the side effects which have been problems with the long-established NSAIDs. Patients are often offered intravenous doses of an NSAID in the postoperative period if they experience pain breaking through despite continuous epidural infusion of local anesthetics (see above).

New approaches to delivering analgesic drugs into the body are also being developed. Intranasal (nasal spray), pulmonary (inhaler), and transbuccal (under the tongue) routes of opioid delivery are all being researched with the goal of providing immediate pain relief. A transdermal system of the opioid fentanyl delivery is currently available but is rarely used to treat acute pain. The skin directly under the sticky transdermal patch absorbs fentanyl which acts to gradually release the medication over hours or even days. Unfortunately, changing the blood level of a drug is slow when using transdermal patches, and so the risk of drug overdose in this patient population is higher. Transdermal patches are rarely used in the postoperative phase of patient care, but are more commonly used for chronic and cancer-related pains.

1.2 Chronic Pain

Chronic, noncancer related pain affects 75 million Americans. Pain decreases the quality of life of these people and impairs their ability to work productively. Chronic pain is defined as pain that persists for longer than the expected time frame for healing, or pain associated with progressive diseases. Many patients suffer from clinical syndromes for which there are no x-ray, laboratory, or other physical evidence of something wrong, and so diagnoses rely on clinical criteria alone. These syndromes include many kinds of headache, myofascial pain syndromes (for example chronic jaw pain), fibromyalgia, neuropathic pain, low back pain, and central pain syndromes. Little is known about the underlying causes of these syndromes. Patients do, however, present with the complaint of pain. Other patients have disorders that are well described, e.g., osteoarthritis, but for which we have no cure.

Irrespective of the cause, the effects of chronic pain on a patient's life tend to be more pervasive than that of acute pain. Chronic pain affects all of the patient's social relationships, personality, and mood. Typically, people with chronic pain tend to have sleep disturbances, generalized fatigue and a decrease in their overall level of functioning. In many, the perception of pain cannot be eliminated, only lessened. The goal becomes one of managing the pain as well as is

possible while improving the person's involvement and function in his life. To this end, a multidisciplinary approach that focuses on all aspects of the patient's condition seems to have the best results. The core team consists of a pain management specialist along with representatives from pharmacy, physical and occupational therapy, psychology, and internal medicine. The care team tailors the care plan according to the needs of the individual patient. An open discussion of goals is necessary to assess objectively if progress is being made. Often the goals of productive activity and a return to work are stressed.

Treatment of chronic, non-cancer related pain conditions requires the use of analgesic medications. NSAIDs are often the mainstay of therapy. Long-term use of the agents carries the risk of easy bruising, renal dysfunction, and the development of gastric ulcers. As mentioned above, the newer NSAIDs help to limit these side effects. Patients must be warned of these complications and followed closely. The agents are potent analgesics that improve the quality of life for those who suffer from many chronic pains, especially arthritis.

Long-acting opioids can effectively control pain in a subset of chronic pain patients. Patients must be assessed carefully before the start of therapy. The potential for substance abuse or addiction must be considered even though they are not commonly encountered in patients with chronic pain. Guidelines exist which help to clarify the conditions under which the prescribing of opioids is appropriate in the treatment of chronic pain states. Common adverse side effects include constipation, pruritis, sedation, and impaired cognition. Although most of the side effects diminish over time, constipation usually persists and must be treated.

Antidepressants are effective in the treatment of many painful conditions. Their mechanism of action is unclear, but they seem to act by increasing the levels of norepinephrine and serotonin at the nerve synapses of the central nervous system. In addition to their direct effect in diminishing the perception of pain, the agents also are useful as an aid to sleep. Patients take the medication in the late evening and report improvements in their ability to fall asleep and stay asleep throughout the night. The agents have the potential for bothersome side effects which include dry mouth, blood pressure changes, urinary retention, and, rarely, heartbeat irregularities. The drugs are used as a part of an integrated care plan and improve the overall quality of the patient's life.

Anticonvulsants are effective in the treatment of many chronic pain states. The mechanism of action is again unclear. They may suppress the production of nerve impulses in nerve cells that are critical in developing the perception of pain. A newer anticonvulsant, gabapentin, provides symptomatic relief with fewer side effects than older drugs of this class. Patients notice sedation and episodes of dizziness when they begin taking the drug, but acclimate to it over a matter of days.

More invasive methods for the treatment of chronic painful states have been tried with no proven benefit. Cutting the spinal cord pathways that transmit pain messages to the brain produces temporary (6–12 month) pain relief, but the pain then returns, sometimes worse than ever. The direct injection of local anesthetics into the most painful muscle areas of myofascial pain syndrome seems to provide symptomatic relief for some of the patients. The procedures have not been tested in a randomized or blinded manner, however, and the benefits may be the result of the placebo reaction. Epidural steroid and local anesthetic injections have also been used for those who complain of vague lower back pains. The steroids are known to decrease inflammation, edema, and decrease the synthesis of prostaglandins (substances known to mediate the perception of pain). Again, the patients often report benefits in their symptoms, but there is no clear mechanism to explain the improvement.

1.3 Cancer Pain

The prevalence of cancer pain ranges from 30–50 percent of patients who are undergoing active treatment for solid tumors. The number climbs to 80–90 percent for those who suffer from advanced forms of the disease. Approximately three-quarters of these chronic pain syndromes result as a direct effect of the tumors themselves. The remaining complaints are related to the therapies administered to manage the primary or metastatic aspects of the cancer.

Neoplastic (tumor) invasion of bone or connective tissue is a major cause of severe pain. The spine is the most common site of bony metastasis. Many patients with cancer have back pain because the cancer has spread to this bony area. Continued growth of the tumor causes spinal cord and nerve root compression. Cancerous growth in this area will result in pain, as well as paralysis and the loss of bowel and bladder control for the patient. Obstruction and compression of the bowel produces intense visceral complaints. Patients will present with persistent abdominal pain, constipation, nausea, and vomiting. The tumor will have to be surgically removed, sometimes resulting in nerve injury-induced (neuropathic) pain. Radiation, frequently used to shrink tumors, can cause fibrosis, which may in turn damage peripheral nerve and central nervous system tissue. These injuries frequently cause a chronic neuropathic-type pain, along with tingling and numbness in the area.

The mainstay for management of cancer-related pain is the use of oral analgesics leading up to opioids in progressively larger doses as the disease progresses. The World Health Organization has devised an algorithm for use of oral analgesics that matches the intensity of the pain with the potency of the medication

prescribed. Although the oral route is commonly used, the transdermal approach has gained popularity because of the ease of administration. Each patch is worn for approximately three days before it must be replaced. The system has the advantage of infrequent dosing, but as mentioned above, keeping the right blood level is difficult. Most practitioners will start the patient with a relatively low dose of transdermal fentanyl, but then give the patient morphine pills to use as needed in order to supplement the relief provided by the patch. Individualization is the key to successful treatment. Each patient must be evaluated and a treatment plan formulated based on that individual's needs. In some, radiation and chemotherapy will effectively reduce the tumor mass and diminish pain. In others, there will be little hope of shrinking the tumor and the only recourse is to make these patients as comfortable as possible. This 'end-of-life' strategy is termed palliative care.

The success of continuous infusions of local anesthetics in the operating room and immediately postoperatively has prompted the developments of portable and reprogrammable infusion pumps for use by cancer pain patients. In some cases, these pumps are implanted under the skin. Pumps allow the continuous infusion of strong opioids through catheters to sites directly within the spinal canal, either epidurally or directly onto the spinal cord itself (intrathecally). The opioids act directly on nerve cells in the spinal cord to inhibit pain messages, resulting in profound pain relief. The epidural or intrathecal catheters are tunneled under the skin and attached to a pump that has within it a reservoir to hold the opioid. The machinery is programmed to deliver a preset amount of opioid as needed to provide adequate pain relief. On a regular basis, the reservoir is refilled to allow for continuous pain relief as needed.

The most invasive therapies used for those with advanced stages of cancer involve the use of ablative surgical or anesthetic procedures. As mentioned above, spinal cord incisions can produce profound pain relief. Because pain recurs over the long term, however, this procedure is only used near the end of life. Pure alcohol is toxic to nerves. Thus, another fairly drastic procedure involves the injection of pure alcohol into individual nerves of the spinal cord to destroy those nerves carrying pain messages to the CNS. In these ways, patients near the end of their life can achieve a degree of pain relief that is not attainable with less invasive measures.

Cancer pain must be viewed in the context of a progressively advancing illness. The optimum management allows for the treatment to match the intensity of discomfort. The smallest dose of the most benign medication that provides relief is used first. When the intensity of pain increases, stronger medications and more invasive techniques are used. In this way, a comprehensive range of agents and techniques are used in sequence to provide optimal pain relief.

2. *Novel Approaches to the Relief of Pain*

2.1 *Transplantation*

Recent laboratory studies have shown that transplantation of adrenal medullary tissue or of isolated chromaffin cells into CNS pain modulatory regions can reduce pain sensitivity in both acute and chronic (including neuropathic) models in rats. The analgesia produced by these transplants probably results from the release of both opioid peptides and catecholamines such as adrenaline and noradrenaline since it can be blocked or attenuated by both opiate and adrenergic blockers. Studies indicate that even over long periods there is no apparent development of tolerance.

Positive results have also been obtained in preliminary clinical studies using transplants of human adrenal gland pieces to relieve intractable cancer pain in 25 patients. Pain scores and morphine intake were measured prior to transplantation and at regular intervals following the graft implantation. Biochemical assays were performed on a sample of the patients' cerebrospinal fluid (CSF) throughout the study to assess levels of catecholamine and met-enkephalin (an opiate manufactured by the body) as an indication of graft viability. Sixteen out of 20 patients reported long lasting pain relief from these transplants (more than a year in some patients). Narcotic analgesic (e.g., morphine) intake was eliminated, reduced or at least stabilized in all patients who completed the study. These studies indicate that even over long periods, there is no development of opiate tolerance. Histological data was obtained at autopsy from some patients who had transplants for over a year. Study of the spinal cord revealed intact adrenal medullary tissue fragments distributed over several levels, with no signs of inflammation or rejection. The results of these clinical trials are encouraging over all and suggest that chromaffin cell transplants can provide relief from intractable pain. With appropriate modifications to optimize graft preparations and survival, it is likely that this approach can become an alternative means of controlling pain refractory to traditional pharmacotherapies (see Michalewicz et al. 1999, Lazorthes et al. 2001).

2.2 *Gene Therapy*

Another new approach to chronic pain is that of gene therapy. This method so far is only at the animal testing stage, but appears to be very promising. Different approaches have been taken to change the genetics of nerve cells that are involved in pain so that that function of the cells is disrupted. The first of these involves the injection of 'antisense' on to the spinal cord. Genes are essentially codes that cells read, and which cause the cell to make a particular chemical, as,

for example, a protein. Some of these chemicals are critical to the transmission of pain information from one cell to another. Antisense is essentially the code for one of these chemicals, but in reverse order. This effectively disrupts the cell's ability to read the 'real' gene and so inhibits its ability to make the chemical. If antisense is injected on to surface of the spinal cord, some of it will find its way into the correct pain neurons and so, to a greater or lesser extent, inhibit transmission of pain messages (see Peric et al. 2001).

Applying antisense to spinal cord nerve cells appears to work in animals, but is somewhat nonselective as to which cells it goes into, and is not very efficient in getting it there. For this reason, other researchers have been working on ways to use viruses as 'vectors' to carry genes that may inhibit pain transmission into the nerve cells that carry that information. One method involves putting the gene for naturally-occurring opioids into either adenoviruses (the kinds that causes the common cold) or herpes viruses (the kinds that causes cold sores). These viruses are very efficient at getting these genes into spinal nerve cells and produced long-lasting analgesia in animals. The adenovirus is injected directly into the spinal cord itself, whereas the herpes virus is placed near pain receptors outside the CNS where it is taken up and carried back to the spinal cord. The herpes virus approach has also been used to carry antisense directly into the pain-conducting nerves (see Wilson et al. 1999). All of these approaches present new and exciting approaches to the treatment of chronic pain.

See also: Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Chronic Pain: Models and Treatment Approaches; Motivation, Neural Basis of; Pain, Health Psychology of; Pain, Neural Basis of; Somatosensation

Bibliography

Lazarthes Y, Bès J C, Sol C, Pappas C D 2001 Intrathecal chromaffin cell allograff for cancer pain. In: Burchiel (eds.) *Surgical Management of Pain*. Thieme Medical Publishers Inc., New York, pp. 973–9

Livingston W K 1998 *Pain and Suffering*. IASP Press, Seattle, WA

Melzack R 1973 *The Puzzle of Pain*. Basic Books, New York

Michalewicz P, Laurito C E, Pappas G D, Yeomans D C 1999 Purification of adrenal chromaffin cells increases autinociception efficiency of xenotransplants in the absence of immunosuppression. *Cell Transplant* **8**: 103–9

Peric V, Lu Y, Laurito C E, Yeomans D C 2001 Combined effects of N-type calcium channel blockers and morphine on a-delta US. C fiber mediated nociception. *Anesthesiology and Analgesia* **92**: 39–243

Rey R 1995 *The History of Pain*. Harvard University Press, Cambridge, MA

Wilson S I, Yeomans D C, Beuder M A, Lu y, Glorioso J 1999 Antihyperalgesie effects of delivery of encephains to mouse nociceptive neurons by herpes virus encoding proeucephalins. *Proceeding of the National Academy of Sciences of the USA* **96**: 3211–16

D. C. Yeomans, C. E. Laurito and G. D. Pappas

Pain, Neural Basis of

The term 'pain' has several different connotations that are a potential source of confusion in its usage. The most frequent of such confusions consists in an equivocation of first- and third-person aspects of the target phenomenon (Metzinger 2000). There is the consciously experienced subjective qualitative aspect of pain. In this sense, to feel pain indicates a state of suffering. This aspect is only experienced from the first-person perspective and, arguably, is not directly accessible for empirical research operating from an objective, third-person perspective. Third-person aspects of pain are constituted by its neural underpinnings (e.g., the minimally sufficient neural correlate of the subjectively experienced, conscious quality of pain) and the functional properties, that is, the causal role, played by them in generating overt behavior, internal reactions, the formation of memories, etc. In this sense, pain is the percept that results from the activation of a specific sensory system, the nociceptive system, which is activated by external stimuli that threaten the integrity of the body, and by internal states of tissues such as an inflammatory process. As such, pain is a sensory experience similar to the experience of hearing a sound or tasting food as a consequence of the activation of the auditory, gustatory, or olfactory system. A comprehensive theory of pain will have to do justice to both aspects, the third-person and first-person aspects, of pain.

Since the 1970s the following definition by the International Association for the Study of Pain has gained widespread acceptance (Merskey et al. 1979):

- Pain is an unpleasant sensory and emotional experience
- associated with actual or potential tissue damage
- or described in terms of such damage.

This definition emphasizes that pain has several components, including at least a sensory one and an affective one. In most instances, pain is not a generalized feeling (*Allgemeingefühl*) such as hunger and thirst, but a localized sensation. The capacity to locate pain is due to the encoding of spatial information in the nociceptive system. Likewise, fluctuations in the perceived intensity of pain can be related to intensity coding in the nociceptive system. The discriminative power and the precision of stimulus encoding in the nociceptive system have traditionally been underestimated. In addition to this sensory-discriminative component, pain almost inevitably encompasses an affective component, mostly of a negative hedonic

quality, which is not a simple function of sensory intensity or quality. While pain may be a particularly unpleasant sensory experience, affective components are present in other sensory systems as well: a person may like or dislike a certain type of music, and a meal may have a pleasant or unpleasant taste. These affective or emotional components of sensory experience provide powerful drives to direct behavior (e.g., withdrawal, guarding, or avoidance behavior). The unpleasant emotional aspects of pain perception are therefore called the affective-motivational component of pain. The aim of this article is to present the neural basis of both components of pain from the point of view of sensory physiology. In the last section, an outlook will be given as to how sensory physiology may provide hints to understanding the subjective experience of feeling pain.

1. *The Nociceptive System*

Many tissues such as the skin, mucous membranes, muscles, joints, bones, and parts of the viscera (with the notable exception of the brain itself) are innervated by free endings of thin primary afferent nerve fibers (Raja et al. 1999). The adequate stimulus to activate these nerve endings is a strong mechanical or thermal stimulus, or one of a variety of chemicals that are either applied from the outside (as with a nettle sting or the application of hot peppers to mucous membranes) or generated within the tissue (such as bradykinin, serotonin, histamine, prostaglandins, etc.). Because they respond to several stimulus modalities, these nerve endings are said to be polymodal. The common denominator of these stimuli is that they indicate either actual or impending tissue damage. Sherrington summarized this concept in the term 'noxious stimulus,' and the sensory system that processes information on noxious stimuli is called the nociceptive system. More recently it was found that polymodality does not depend on actual tissue damage but may also be a property of the ion channels present in free nerve endings. Many nociceptive nerve endings express the so-called vanilloid receptor channel VR1 (Caterina and Julius 1999), which can be activated by a physical stimulus (heat), and by both exogenous (capsaicin and other hot spices) and endogenous chemicals (anandamide and other derivatives of membrane lipids).

Some primary nociceptive afferents (A-fibers) conduct nerve impulses relatively rapidly (about $20\,ms^{-1}$), because of the saltatory impulse conduction that is enabled by the presence of a myelin sheath; most of the primary nociceptive afferents are unmyelinated C-fibers and conduct slowly (about $1\,ms^{-1}$). A-fibers are involved in rapid behavioral responses and in motor reflexes. C-fibers are involved in slower signaling processes, including the perception of a delayed 'second pain' after a single brief stimulus such as a pinprick, changes in the sensitivity of the nociceptive system (see Sect. 4), control of local blood vessels, and trophic functions (Holzer 1992).

Primary nociceptive afferents make synaptic contact with second-order neurons in the dorsal horn of the spinal cord. Some nociceptive specific neurons have no other input, whereas another class of neurons receives convergent input from touch receptors in the skin. These neurons encode a wide range of mechanical stimulus intensities and are therefore called wide-dynamic-range neurons (Willis 1985). Spinal nociceptive neurons are subject to modulation by descending pathways from the brain stem, which are mostly inhibitory and utilize endogenous opioids, noradrenaline, and serotonin as neurotransmitters. The net result of primary afferent inputs, descending modulation, and local interactions in the spinal cord dorsal horn is then projected to spinal motor nuclei in the ventral horn, spinal autonomic nuclei in the lateral horn, to the brain stem and to the contralateral somatosensory thalamus.

Via the spinoreticular and spinomesecephalic tract projection to the brain stem, noxious stimuli may activate autonomic reflex centers of the cardiovascular and respiratory system, the descending inhibition of the spinal cord dorsal horn, and the ascending reticular activating system (ARAS) that projects to the cerebral cortex. These projections are the neural basis for many autonomic reflexes elicited by pain, for the fact that one pain may inhibit another pain, and for the strong arousal reaction to painful stimuli.

Via the spinothalamic tract, the nociceptive pathways reach two groups of nuclei in the thalamus (Craig and Dostrovsky 1998). The lateral group is situated in the ventrobasal complex and is identical to the relay station of the tactile system. These thalamic nuclei project to the primary and secondary somatosensory cortex (SI and SII), and to the insula. The medial group is mostly situated within the white matter of the thalamus (intralaminar nuclei). These thalamic nuclei project to the anterior cingulate gyrus, the basal ganglia, and also nonspecifically to wide parts of the cerebral cortex. Since its original suggestion in the late 1960s (Melzack and Casey 1968), the concept that the lateral system processes the sensory discriminative component, and the medial system the affective-motivational component of pain, is still a valid concept (Treede et al. 1999).

2. *Sensory Discriminative Functions*

The sensory discriminative component of pain may be divided into at least three aspects: stimulus localization, intensity discrimination, and quality discrimination. The human capacity to perceive the location of tissue damage is used in everyday medical practice, when doctors ask the question 'where does it hurt?' Sometimes, tissue damage and perceived pain are in

different locations, as in the case of a myocardial infarction, which may cause pain in the left arm. This phenomenon (referred pain) is due to convergence of visceral and cutaneous afferents in the spinal cord. The activity of the spinal neurons does not reveal the afferent source, and higher centers are assumed to project the resulting sensation along the more common pathway from the skin. Because of the mislocalization of pain originating from damage to the viscera and the precise localization of tactile stimuli at the fingertips, localization of noxious stimuli has traditionally been ascribed to simultaneous activation of the tactile system. Tactile acuity, however, rapidly drops outside the 'foveal' areas of the fingertips and lips. On the back of the hand, sequential tactile stimuli must be about 1 cm apart to be perceived as being given to separate locations. When laser-radiant heat pulses (which do not activate tactile afferents) are given to the back of the hand, the sequential spatial discrimination threshold is also about 1 cm (Schlereth et al. 2001).

These psychophysical findings demonstrate that the nociceptive system in humans provides sufficient spatial information to account for our capacity to know where it hurts. What is the neural basis for this capacity? Receptive fields of nociceptive neurons show somatotopic organization in the spinal cord dorsal horn (Woolf and Fitzgerald 1986), lateral thalamus (Albe-Fessard et al. 1985), and the primary somatosensory cortex (Kenshalo and Willis 1991). Receptive field sizes in SI are smaller than in the spinal cord and thalamus, possibly due to lateral inhibition, and match the pain localization capacity. Thus, all existing evidence favors the view that SI is involved in stimulus localization for the nociceptive system as well as the tactile system. Indirect evidence for this concept is provided by studies in humans that use measures of brain perfusion as parameters of activation (by positron emission tomography: PET). These studies suggest that only tasks with a stimulus-localization component activate SI, whereas studies with fixed stimulus location do not (Treede et al. 1999).

The capacity to encode different intensities of noxious stimuli is one of the criteria for identifying nociceptive neurons (Willis 1985). Intensity coding has been demonstrated for nociceptive neurons in SI and to a certain extent in SII, but also in medial thalamic nuclei. Intensity coding is a poor criterion for identifying neurons involved in sensory discriminative aspects of pain, because the affective-motivational component of pain also depends on stimulus intensity. In a study that demonstrated the different relative unpleasantness of experimental painful stimuli, perceived intensity and unpleasantness were both nevertheless related to stimulus intensity (Rainville et al. 1992).

Pain may have different qualities such as burning, stinging, or aching. Neither the number of subjectively discriminable qualities nor their neural basis are known with sufficient precision. The polymodality of primary nociceptive afferents seems to contradict the capacity for quality discrimination, but the response profiles to different noxious stimuli differ between different afferents. Quality discrimination in the nociceptive system is therefore likely to be due to a population code, similar to the encoding of the taste qualities in the gustatory system.

Current evidence supports the traditional view that the lateral nociceptive system subserves the sensory discriminative component of pain. Different aspects of this pain component (detection, localization, intensity discrimination, quality) may be processed in parallel by separate pathways. In contrast to traditional views, there is no evidence that the tactile system participates in any of these functions. Thus, nociception is established as a sensory modality within the somatosensory system.

3. *Affective Motivational Functions*

The affective-motivational component of pain sensation encompasses several aspects that are closely related: the negative hedonic quality and emotional reactions, an increase in the arousal level and stimulus-related selective attention, and the drive to terminate the stimulus causing this sensation. Some of these functions can be considered second-order sensory processing, whereas the latter is a premotor function. The affective-motivational component of pain is classically associated with the medial nociceptive system, which in turn is connected to the limbic system.

One part of the medial nociceptive system, the anterior cingulate cortex, has recently gained much publicity, because virtually all PET studies of acute pain gave evidence of activation in that area, including one study in which pain was elicited as an illusion by the interaction of two nonpainful stimuli (Craig et al. 1996). Electrophysiological recordings in humans also show activity in this area (Treede et al. 1999). The anterior cingulate cortex is a functionally heterogeneous brain area that has been implicated in the integration of affect, cognition, and response selection in addition to aspects of social behavior (for review, see Devinsky et al. 1995). Passive functions (emotion, attention) are represented more frontally, whereas a premotor part of the anterior cingulate cortex is situated more posteriorly, below the supplementary motor area. It is still debated, whether the anterior cingulate cortex may contain a nociceptive-specific area, or whether painful stimuli nonspecifically recruit several parts of the large cingulate gyrus.

The affective-motivational component of pain may also be processed in the insula. The contralateral insula was activated almost as frequently in human PET studies of acute pain as the anterior cingulate cortex (Casey and Minoshima 1997). Microstimulation in a thalamic nucleus that projects to the insula elicited pain with a strong affective component only in those patients who had previously experienced such

pain either due to panic attacks or due to angina pectoris (Lenz et al. 1995). These observations suggest that the insula may be part of a sensory limbic projection pathway for pain sensation. The insula projects to the amygdala (Augustine 1996), which is a part of the limbic system that is associated with emotions.

4. *Plasticity of the Nociceptive System*

As in all other sensory systems, the repeated presentation of noxious stimuli or the exposure of the nociceptive system to stimuli of long duration may lead to a reduction in the intensity of the evoked response (habituation and adaptation). A unique property of the nociceptive system is that under many circumstances the prior presentation of the adequate stimulus may actually enhance subsequent responses (Treede et al. 1992). This enhancement is called sensitization and it may occur at both the peripheral terminals of primary nociceptive afferents and within the central nervous system. Sensitization refers to a leftward shift in the function that relates the neural response to stimulus intensity. It is characterized by a drop in threshold and an increase in the response to suprathreshold stimuli. Spontaneous activity may also result from sensitization.

The perceptual correlate of sensitization is hyperalgesia, which is characterized by a drop in pain threshold and an enhanced painfulness of suprathreshold stimuli. Even minor tissue injury elicits a transient state of hyperalgesia, both of the injured tissue itself (primary hyperalgesia) and in surrounding uninjured skin (secondary hyperalgesia). Primary hyperalgesia is characterized by a lowered pain threshold for both mechanical and heat stimuli. Secondary hyperalgesia is characterized by a lowered pain threshold for mechanical stimuli only.

In primary hyperalgesia, the heat pain threshold may drop below the normal body temperature, which then starts to act as an adequate stimulus for the

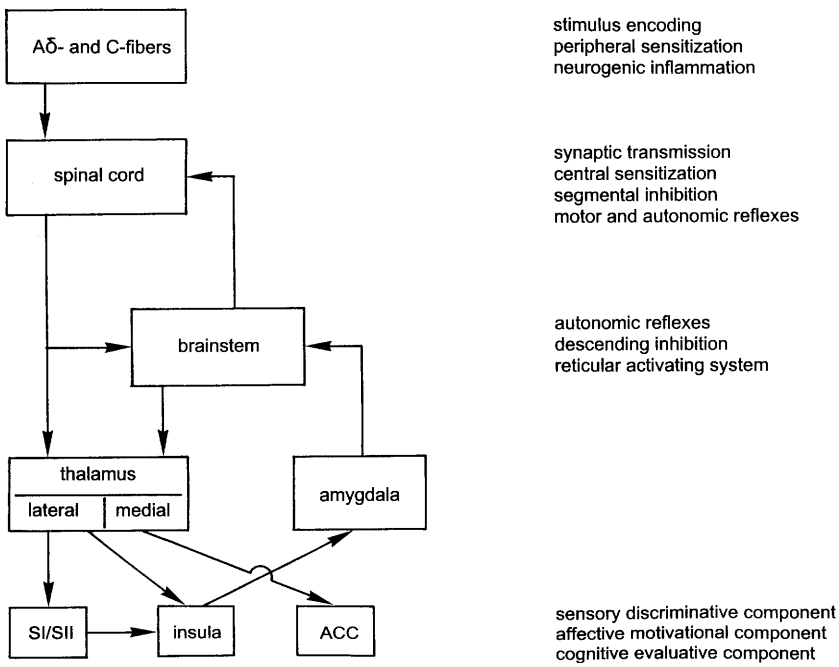

Figure 1
Major components of the nociceptive system. Free endings of Aδ- and C-fibers in skin and other tissues are activated by actual or impending tissue damage (noxious stimuli). At the first synaptic relay station in the spinal cord, their signals interact and are strongly modulated by segmental and descending connections. After another synaptic relay station in lateral thalamic nuclei, the signals reach the primary and secondary somatosensory cortex (SI, SII) as a step towards the conscious perception of a noxious stimulus as being painful. Other parts of the cerebral cortex such as the insula and the anterior cingulate cortex (ACC) are reached both in series and in parallel to SI and SII. The distributed processing of nociceptive information is thought to lead to the sensory-discriminative, affective-motivational and cognitive evaluative components of conscious pain perception. Motor and autonomic reactions to noxious stimuli are processed in lower parts of the central nervous system and are mostly independent of conscious pain perception; some of them may still occur after brain death

nociceptive system, leading to ongoing pain. Such pain may, for example, occur in inflammatory conditions and is rationally treated by cooling. The neural basis of primary hyperalgesia to heat stimuli is peripheral sensitization of primary nociceptive afferents by shifts in the temperature sensitivity of vanilloid receptors and related modifications of other ion channels (Cesare et al. 1999).

In secondary hyperalgesia, the mechanical pain threshold may drop in such a way that even gentle stimuli such as movement of cotton wool across the skin is perceived as painful (allodynia). Normal weight bearing on the sole of the feet or contact with clothing then start to act as an adequate stimulus for the nociceptive system, leading to functional deficits by avoidance behavior. Such pain may also occur chronically following lesions of the nervous system (neuropathic pain), and this pain is particularly resistant to current treatment modalities. The neural basis of secondary hyperalgesia to mechanical stimuli is central sensitization of nociceptive neurons in the spinal cord by modifications of the synaptic transmission from Aß-fiber mechanoreceptors that normally signal touch and from A-fiber nociceptors (Ziegler et al. 1999).

Central sensitization of the nociceptive system involves the activation of NMDA (N-methyl-D-aspartate) glutamate receptors in the postsynaptic neuron and several additional mechanisms. The intracellular signal pathways are similar to those found for long-term potentiation (LTP) of synaptic efficacy in the hippocampus, which is thought to be related to learning and memory (see *Neural Plasticity* and *Working Memory, Neural Basis of*). Both LTP and the opposing mechanism long-term depression (LTD) have been observed in slice preparations of the spinal cord using stimulation protocols similar to those that were found to be effective in slice preparations from the hippocampus or neocortex. In the intact organism, the triggering of LTP in the spinal cord is antagonized by the endogenous descending inhibition (Sandkühler 2000). The implications of these mechanisms for the causes of chronic pain are important but not yet fully understood. On the one hand, the intracellular signal pathways of LTP influence gene expression and may thus alter the function of nociceptive neurons for long periods of time. On the other hand, deficiencies in the balance between excitatory and inhibitory influences on the spinal cord may be a decisive factor in pain chronification.

5. *Epilogue: Pain and the Brain*

If reading the preceding sections have left the impression that the neural basis of actually feeling pain has not been addressed, this impression is correct. It is evident that pain exists only as long as it is being felt by a person. Thus, trying to understand the pain experience as a whole leads into the general mind–body problem, which is a complex philosophical issue (for a more general account of how personal-level and subpersonal-level descriptions of pain can be combined, see Bieri 1995). This article has intentionally left out the intimate relationship of the subjective pain experience and consciousness. Instead, the preceding sections have described the neural pathways that signal tissue damage and how some of their known properties explain otherwise puzzling phenomena such as the allodynia of neuropathic pain states, where lightly touching the skin may evoke unbearable pain. But how the brain ultimately synthesizes the conscious perception of pain remains a mystery. A few simple examples, however, show that the brain does synthesize this perception:

(a) Blocking the neural pathways between the damaged tissue and the brain (e.g., by a local anesthetic in the dentist's office) creates a situation where there is tissue damage but no pain.

(b) Activation of nociceptive pathways by electrical stimulation within the thalamus may elicit a vivid perception of the pain of a heart attack. In this situation there is pain but no tissue damage.

In other words, a foot cannot hurt by itself, but a brain can perceive a hurting foot. The latter is possible even in the absence of the foot, as observed in phantom limb pain. Pain due to proximal activation of the nociceptive system is called projected pain, because here the normal mechanisms of pain perception are particularly evident: pain sensation is synthesized in the brain, but is then projected into the peripheral receptive field of the active brain areas. Pain due to peripheral activation of nociceptive afferents is called nociceptive pain, but again the pain sensation is synthesized in the brain and projected into the peripheral receptive field, which happens to be the site of tissue damage. This model of pain projection into receptive fields is simplified and does not include the central representation of the body image in brain areas such as the posterior parietal cortex, but this much should be apparent: pain is always in the brain.

See also: Chronic Pain: Models and Treatment Approaches; Pain, Health Psychology of; Pain, Management of; Somatosensation

Bibliography

Albe-Fessard D, Berkley K J, Kruger L, Ralston H J, Willis W D 1985 Diencephalic mechanisms of pain sensation. *Brain Research Reviews* **9**: 217–96

Augustine J R 1996 Circuitry and functional aspects of the insular lobe in primates including humans. *Brain Research Reviews* **22**: 229–44

Bieri P 1995 Pain: A case study for the mind–body problem. In: Bromm B, Desmedt J E (eds.) *Pain and the Brain: From Nociception to Cognition*. Raven Press, New York, pp. 99–110

Casey K L, Minoshima S 1997 Can pain be imaged? In: Jensen T S, Turner J A, Wiesenfeld-Hallin Z (eds.) *Proceedings of the 8th World Congress on Pain, Progress in Pain Research and Managementi*. IASP Press, Seattle, WA, pp. 855–66

Caterina M J, Julius D 1999 Sense and specificity: A molecular identity for nociceptors. *Current Opinion in Neurobiology* **9**: 525–30

Cesare P, Moriondo A, Vellani V, McNaughton P A 1999 Ion channels gated by heat. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 7658–63

Craig A D, Dostrovsky J O 1998 Processing of nociceptive information at supraspinal levels. In: Yaksh T L (ed.) *Anesthesia: Biologic Foundations*. Lippincott-Raven, Philadelphia, PA, pp. 625–42

Craig A D, Reiman E M, Evans A, Bushnell M C 1996 Functional imaging of an illusion of pain. *Nature* **384**: 258–60

Devinsky O, Morrell M J, Vogt B A 1995 Contributions of anterior cingulate cortex to behaviour. *Brain* **118**: 279–306

Holzer P 1992 Peptidergic sensory neurons in the control of vascular functions: Mechanisms and significance in the cutaneous and splanchnic vascular beds. *Reviews in Physiology, Biochemistry and Pharmacology* **121**: 49–146

Kenshalo D R, Willis W D 1991 The role of the cerebral cortex in pain sensation. In: Peters A, Jones E G (eds.) *Cerebral Cortex*. Plenum Press, New York, Vol. 9, pp. 153–212

Lenz F A, Gracely R H, Romanoski A J, Hope E J, Rowland L H, Dougherty P M 1995 Stimulation in the human somatosensory thalamus can reproduce both the affective and sensory dimensions of previously experienced pain. *Nature Medicine* **1**: 910–13

Melzack R, Casey K L 1968 Sensory, motivational, and central control determinants of pain. A new conceptual model. In: Kenshalo D R (ed.) *The Skin Senses*. Charles C. Thomas, Springfield, MA, pp. 423–43

Merskey H, Albe-Fessard D, Bonica J-J, Carmon A, Dubner R, Kerr F-W-L, Lindblom U, Mumford J-M, Nathan P-W, Noordenbos W, Pagni C-A, Renaer M-J, Sternbach R-A, Sunderland S 1979 Pain terms: A list with definitions and notes on usage. Recommended by the IASP subcommittee on taxonomy. *Pain* **6**: 249–52

Metzinger T 2000 The subjectivity of subjective experience: A representationalist analysis of the first-person perspective. In: Metzinger T (ed.) *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. MIT Press, Cambridge, MA, pp. 285–306

Rainville P, Feine J S, Bushnell M C, Duncan G H 1992 A psychophysical comparison of sensory and affective responses to four modalities of experimental pain. *Somatosensory Motor Research* **9**: 265–77

Raja S N, Meyer R A, Ringkamp M, Campbell J N 1999 Peripheral neural mechanisms of nociception. In: Wall P D, Melzack R (eds.) *Textbook of Pain*, 4th edn. Churchill Livingstone, Edinburgh, UK, pp. 11–57

Sandkühler J 2000 Learning and memory in pain pathways. *Pain* **88**: 113–18

Schlereth T, Magerl W, Treede R-D 2001 Spatial discrimination thresholds for pain and touch in human hairy skin. *Pain* **92**: 187–94

Treede R-D, Kenshalo D R, Gracely R H, Jones A K P 1999 The cortical representation of pain. *Pain* **79**: 105–11

Treede R-D, Meyer R A, Raja S N, Campbell J N 1992 Peripheral and central mechanisms of cutaneous hyperalgesia. *Progress in Neurobiology* **38**: 397–421

Willis W D 1985 *The Pain System: The Neural Basis of Nociceptive Transmission in the Nervous System*. Karger, Basle, Switzerland

Woolf C J, Fitzgerald M 1986 Somatotopic organization of cutaneous afferent terminals and dorsal horn neuronal receptive fields in the superficial and deep laminae of the rat lumbar spinal cord. *Journal of Comparative Neurology* **251**: 517–31

Ziegler E A, Magerl W, Meyer R A, Treede R-D 1999 Secondary hyperalgesia to punctate mechanical stimuli: Central sensitization to A-fibre nociceptor input. *Brain* **122**: 2245–57

R.-D. Treede

Paleodemography: Demographic Aspects

1. Introduction

How long humans lived in the past is a question that has challenged researchers for centuries. Once the dominion of philosophers, theologians, and then historians, physical anthropologists have also struggled to answer this question over the last century. For the physical anthropologist, understanding long-term patterns of the human life span aids in placing observed changes within an evolutionary framework, linking both biological and cultural pressures and adaptations to a changing pattern of life expectancy in populations.

Exploration of this question by physical anthropologists has been primarily in the analysis of skeletal remains recovered archaeologically, and has fallen under the purview of a field developed as a subspecialty of physical anthropology—*paleodemography*.

The early days of paleodemography represented an exploration of modern demographic theory applied to ancient populations and the use of the abridged life table as a tool to aid interpretations of age-at-death profiles from cemetery samples. The 1980s marked a pivotal point for paleodemography. While there had been the occasional critique prior to the 1980s, it was not until 1982 that the great debate over the merits of paleodemography began, sparking several years of controversy within the literature (e.g., Bocquet-Appel and Masset 1982, Buikstra and Konigsberg 1985, Wittwer-Backofen 1987). Subsequent to this period, the two critical issues explored most in paleodemography have been (a) the accuracy of aging techniques, and (b) representativeness of samples.

2. Sources of Data

Paleodemographic studies have focused primarily on the reconstruction of human skeletal samples recovered archaeologically. The basic assumption has been that mortality statistics derived from the skeletal sample are sufficient to make inferences about mor-

tality in the once living population. In conjunction with and sometimes ancillary to skeletal data, archaeological evidence for settlement size and distribution have also been used to estimate population growth or to address more broadly-based questions of population structure in the past. Anthropological demography, of contemporary or recent historic hunter-gatherer and foraging populations, can also provide us with models for prehistoric populations. However, the range of fertility and mortality patterns among populations is so wide and overlapping that ethnographic analogy from subsistence base and mobility is extremely problematic. More recently, evidence from genetic studies has begun to be used to make inferences about long-term evolutionary changes in demographic structure in modern human populations over the past 250,000 years or so.

2.1 Skeletal Remains

The accuracy and reliability of age estimation techniques in particular have been central problems in paleodemography, particularly with respect to underestimation of the ages of older adults. A further issue is the necessary separation of techniques used to estimate age from the skeletal remains of adults vs. children. The latter are based on a various criteria related to known rates of development, whereas the former rely on more variable patterns of rates of development and degeneration of the bony tissues. In addition, differential rates of change between the sexes have made most techniques sex-specific. The scope of this article prevents an in-depth review of such methods, and readers are referred to such classic monographs as Acsádi and Nemeskéri (1970) or Krogman and Iscan (1986) for more detailed descriptions.

2.1.1 Estimation of Age. Determination of age from the adult skeleton can be undertaken using a variety of methods, including both quantitative and qualitative measures of continuous and discrete data. Such methods are referred to as skeletal age-indicator techniques. Adult aging techniques examine both macroscopic and microscopic changes in the morphology and structure of bone. Many macroscopic techniques focus on the pattern of age-related degeneration of bone (e.g., pubic symphysis, auricular surface, sternal end of the fourth rib). Others focus on the remodeling of bone as a biomechanical response to microfractures from everyday wear and tear. The physical anthropologist evaluates multiple criteria for each skeleton and then combines all estimates from each age-indicator method into an overall estimated age range for the individual.

Age estimation from the skeletal remains of children is based on the development of dental and skeletal tissues, including tooth formation and eruption. Estimation of age in the nonadult is much easier and more accurate than in the adult. While there are fewer techniques than are available for adult age estimation, each has a smaller range of error. Overall, estimation of age from a juvenile is accurate within a range of about plus or minus half a year.

2.1.2 Estimation of Sex. It is recognized widely in the anthropological literature that the pelvis or hipbone is the most reliable part of the skeleton for determination of sex. Both metric and nonmetric or morphological techniques have accuracy rates of better than 95 percent for correct sex, while accuracy rates based on other parts of the skeleton are usually lower. While morphological variables are often preferred because of better preservation in skeletal samples, metric variables are considered by many to be more reliable because of their greater precision.

To the regret of many researchers, however, determination of sex from the skeleton has most often been restricted to those who have survived past adolescence and who then manifest changes in the skeleton reflective of sex. While a variety of studies have investigated traits that might be sexually dimorphic in infants and juveniles, only a few have had sufficient levels of accuracy to warrant their application in osteological analyses. More promising, but still restricted by time and money, is the determination of sex by extracting ancient DNA from the bones or teeth of individuals.

2.2 Archaeological Evidence

While paleodemography has been focused primarily within the realms of skeletal biology, other forms of evidence have also been explored to answer demographic questions. In particular, archaeological demography has often shed light on issues of population structure in the distant past. Estimates of population size and growth have been attempted from settlement data by examining features such the size and area of the living site, number of dwellings, the density and distribution of artifacts and food remains, as well as from ethnohistoric estimates of population size. Even when data are available, estimates of population size must often be made through ethnographic analogy, whereby the relationship between population size and material remains observed in modern or historic groups is imposed on the archaeological site.

2.3 Genetic Evidence

Recent studies of DNA have proposed some interesting hypotheses regarding demographic changes in the past that are of interest when exploring the issue

of long-term trends in human prehistory. In particular, advances in nucleotide divergence theory have provided information on human demographic behavior since patterns of gene differences contain information about the demographic history of a species (Harpending et al. 1998). Genetic data have been used specifically to tackle questions of modern human origins, and the evidence implies human demographic expansion from a population of only several thousand about 250,000 years ago.

3. *Representativeness*

The representativeness of a skeletal series is a crucial factor in paleodemographic studies. While anthropologists do not expect samples to represent the population in absolute numbers, the fundamental assumption governing the analysis and interpretation of skeletal samples is that the *distribution and pattern of any parameter is the same in the skeletal sample as in the living population that contributed to it*. That is, the pattern observed in the skeletal sample can be used as a proxy for the real pattern within the population. However, a variety of factors serve to bias paleodemographic reconstructions relative to the true demography of the once living population. Biological factors related to the fact that skeletal samples are the nonsurvivor portion of a population, cultural factors related to the burial of the body, environmental factors related to postdepositional processes, and methodological factors related to the excavation and analysis of skeletal samples all serve to potentially bias interpretations.

Until the 1990s few attempts had been made to directly assess the validity of interpretations from paleodemographic reconstructions because of the obvious difficulty in obtaining a skeletal sample that is known to be representative of its larger population. With the more recent, detailed studies of historic cemetery skeletal samples, researchers have begun to test the representativeness of their samples by comparing the mortality data derived from the skeletal sample with the documentary mortality data associated with the cemetery from which the sample was drawn. In most cases, clear differences in demographic parameters estimated from the two sources became apparent. As a result, recognizing and controlling for these biases has become an important part of paleodemographic reconstructions.

4. *Reconstructing the Demography of Prehistoric Populations*

Traditionally, the statistical tool used by paleodemographers has been the abridged life table. Under the assumption of a stationary population, paleodemography uses mean age at death to estimate expectation of life at birth. Generally, mean age at death is considered approximately equivalent to the inverse of the birth rate in a population, but is independent of both life expectancy and the death rate. However, when the conditions of a stationary population are not met, this calculation simply represents mean age at death.

Traditionally, the abridged life table has been used by paleodemographers to estimate general mortality patterns with fertility being a by-product. However, several estimators of fertility from mortality profiles have been used. Jackes (1992) has noted that comparisons of the age structures across populations are based on the assumption that there is a relationship between juvenile and adult mortality, and that age-at-death data within very broad age categories will carry some information about the fertility rate of the population. However,

> in nonstationary populations, age-at-death distributions are extremely sensitive to changes in fertility but not to changes in mortality … . Thus, if a population is not stationary—and changing populations never are—small variations in fertility have large effects on its age-at-death distribution, while even quite large modifications of mortality have virtually none (Wood et al. 1992, p. 344).

As a result, many researchers have concluded that the age distribution of skeletal samples provides less information about mortality than it does about fertility, a position supported very early on in the demography literature. In fact, the same fertility and mortality schedules can produce different birth and death rates in populations with different age structures.

4.1 *Model Life Tables*

In the 1970s demographers expressed concern over the paucity of evidence from which to make statements regarding paleodemographic parameters, forcing investigators to extrapolate from models derived from other sources. Subsequently, model life tables from modern demographic studies formed the basis from which anthropological demographers began to develop model life tables for past populations. Skeletal samples can be compared with a model life table and the fit between model and observed mortality distributions can then be assessed statistically.

The use of model life tables in anthropological demography is twofold. First, it provides a means of assessing or compensating for biased and incomplete data, and second, it allows for the estimation of fertility rates and construction of an initial population at risk. In the early 1970s,Weiss (1973) developed a set of model fertility and mortality schedules derived from ethnographic samples of contemporary hunter-gather societies and prehistoric skeletal samples. Many investigators agree that demographic statistics derived

from contemporary non-Western societies represent an effective means of assessing skeletal age profiles of past populations. However, given the variety of conditions under which many contemporary populations live, it is difficult to be certain that ethnographic analogy will always be appropriate. Further, the application of ethnographic estimators to samples for which related sociocultural information is sparse further compounds the problem.

4.2 Hazard Analysis

Although a potentially powerful tool for paleodemographic analyses, model life table fitting techniques are still subject to potential biases resulting from the use of inappropriate model populations. As an alternative, Gage (1988) proposed the use of a hazard model of age-at-death patterns that can be fitted to survivorship, death rate, and age structure data. As Gage has noted, the Siler model of mammalian mortality is useful for studying human mortality because it reduces the number of variables while maintaining interpretive power. The model for survivorship, which can be fitted to the l_x column in a life table using nonlinear techniques to estimate the parameters, is:

$$l_t = \exp\left\{ -\frac{a_1}{b_1}[1-\exp(-b_1 t)] \right\} \exp(-a_2 t) \times \exp\left\{ \frac{a_3}{b_3}[1-\exp(b_3 t)] \right\} \quad (1)$$

where t is age, a_1 is the risk of immature mortality at the moment of birth, b_1 is the rate of decline with age of this risk, a_2 is a constant risk factor, a_3 is the risk of senescent mortality at the moment of birth, and b_3 is the rate of increase in senescent mortality with age (the rate of aging).

The instantaneous death rate or force of mortality at age t of this model is

$$h_t = a_1 e^{-b_1 t} + a_2 + a_3 e^{b_3 t} \quad (2)$$

and the proportion of individuals surviving the immature component is $\exp\{-a_1/b_1\}$ (Gage 1988). This technique provides a method of estimating age-specific mortality and fertility directly from anthropological data, and it can smooth demographic data from a variety of populations without imposing a predetermined age structure (Gage 1988).

Ultimately, the focus of paleodemography has been to refine methods to improve estimates of age at the individual level in order to get some aggregate estimate of the age structure in the population. More recently, however, researchers have begun to try and estimate the mortality distribution of a sample directly from the distribution of skeletal age-indicator stages scored. While the difference is subtle, it is important in that such techniques attempt to avoid the broad range of error associated with estimates at the individual level. If one is interested in estimating an individual's age, as in the context of forensic anthropology, then such techniques fall into the category of prediction models. However, if one is interested in shape of mortality or the force of mortality within a population then an alternative theoretical approach is required—one that *begins* with the overall age structure being estimated, and estimates of individual age being derived secondarily. This apparent paradox in paleodemography was noted by Konigsberg and Frankenberg (1994, p. 96):

> Thus, for any single skeleton from an archaeological sample, the posterior probability that the individual is a particular age conditional on the observed indicator is proportional to the prior probability that the individual is that particular age times the probability that an individual that particular age would be in the observed indicator state. However, we do not know this prior probability. If we knew the prior probability of death at particular ages, then we would know the age-at-death distribution, which is precisely what we are trying to estimate. This paradox can be solved by using a common statistical method known as maximum likelihood estimation.

While the precise statistical approach to estimating the mortality structure from skeletal age-indicator data may vary, it is clear that the emerging analytical framework on which paleodemography must move forward begins first with an estimate of the age structure, from which estimates of individual age can subsequently be calculated (cf. Hoppa and Vaupel 2001).

5. Summary

The biological basis for human survival and longevity evolved during the tens of thousands of years of human experience under premodern conditions. A deeper understanding of this basis thus depends on knowledge of mortality and fertility patterns over the long stretch of human existence. Furthermore, deeper understanding of the plasticity of survival and longevity depends on knowledge of how different mortality is today from mortality before the modern era. Paleodemography attempts to address this question by integrating evidence from a variety of sources, but focusing primarily on skeletal, archaeological, and even genetic data. Long-standing interpretations have suggested that the bulk of premodern populations had very short life spans with only a few individuals surviving to middle adulthood. However, new methodological approaches are beginning to readdress this position, and future paleodemographic studies will be able to test this hypothesis more reliably.

See also: Anthropology; Demography, History of; Evolutionary Approaches in Archaeology; Human Evolutionary Genetics; Primates, Evolution of

Bibliography

Acsádi G, Nemeskéri J 1970 *History of Human Lifespan and Mortality*. Akadémiai Kiadó, Budapest

Bocquet-Appel J-P, Masset C 1982 Farewell to paleodemography. *Journal of Human Evolution* **11**: 321–33

Buikstra J E, Konigsberg L W 1985 Paleodemography: Critiques and controversies. *American Anthropologist* **87**(2): 316–34

Gage T B 1988 Mathematical hazards models of mortality: An alternative to model life tables. *American Journal of Physical Anthropology* **86**: 429–41

Harpending H C, Batzer M A, Gurven M, Jorde L B, Rogers A R, Sherry S T 1998 Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences* **95**(4): 1961–7

Hassan F A 1981 *Demographic Archaeology*. Academic Press, New York

Hoppa R D, Vaupel J W (eds.) 2001 *Paleodemography: Age Distributions from Skeletal Samples*. Cambridge Studies in Biological and Evolutionary Anthropology. Cambridge University Press, Cambridge, UK

Jackes M 1992 Palaeodemography: Problems and techniques. In: Saunders S R, Katzenberg M A (eds.) *Skeletal Biology of Past Peoples: Research Methods*. Wiley-Liss, New York, pp. 189–224

Konigsberg L W, Frankenberg S R 1992 Estimation of age structure in anthropological demography. *American Journal of Physical Anthropology* **89**: 235–56

Konigsberg L W, Frankenberg S R 1994 Palaeodemography: 'Not quite dead.' *Evolutionary Anthropology* **3**(3): 92–105

Krogman W M, Iscan M Y 1986 *The Human Skeleton in Forensic Medicine*, 2nd edn. Charles C. Thomas, Springfield, IL

Paine R R 1997 *Integrating Archaeological Demography. Multidisciplinary Approaches to Prehistoric Population*. Center for Archeological Investigations, Occasional Paper No. 24, Southern Illinois University at Carbondale, IL

Weiss K 1973 Demographic models for anthropology. *American Antiquity* **38**(2) Part 2: Memoir 27

Wittwer-Backofen U 1987 Überblick über den aktuellen Stand paläodemographischer Forschung. *Homo* **38**: 151–60

Wood J W, Milner G R, Harpending H C, Weiss K M 1992 The osteological paradox: Problems of inferring prehistoric health from skeletal samples. *Current Anthropology* **33**(4): 343–70

R. D. Hoppa

Panel Surveys: Uses and Applications

A panel survey makes similar measurements on the same sample units (e.g., individuals, firms) at different points in time (Duncan and Kalton 1987). Panel surveys can be distinguished from: (a) cross-sectional surveys, which take measurements on sample units from a population of interest at a single point in time; and (b) repeated surveys, such as the US General Social Survey (Davis and Smith 1994), in which similar measurements are made on samples from an equivalent population (e.g., defined by identical geographical boundaries or birth cohorts) at different points of time, but without attempting to ensure that any elements are included in more than one round of data collection.

A widely used alternative term for panel survey is longitudinal survey. However, it is useful to use 'longitudinal' only in the context of longitudinal *data*; such data may be collected by a panel survey, but they could also come from retrospective reports in a cross-sectional survey.

Most panel surveys follow their sample elements for the duration of the survey. Other panel designs include a rotating panel survey, in which sample elements have a restricted panel life; as they leave the panel, new elements are added. Examples of this are the monthly US Current Population Survey (US Bureau of Labor Statistics 2000a) and the Canadian Labour Force Survey (Statistics Canada 2000), in which the sample at each month is composed of a number of rotation groups, each of which is a self-contained sample from the population of interest. One rotation group drops out and one rotation group enters each month. A split panel survey is a combination of a panel and a repeated survey (Kish 1986).

Panel surveys can differ markedly in the interval between rounds of data collection and the overall length of the survey. On the one hand, a consumer panel in market research may involve weekly contacts with respondents to collect their diary records of purchases. On the other hand, there are long-term panel surveys that may collect data only annually or even less frequently, and these surveys may continue for many years. Examples are the Panel Study of Income Dynamics (PSID), which had been collecting data on US households annually (biannually since 1997) since 1968 (Hill 1992), and the British National Child Development Study, which started in 1958 and has collected data on the sample persons on five occasions in the 33 years to 1991 (Institute of Education 2000).

Some panel surveys are concerned with population subgroups that have experienced the same event during the same time period, such as attending eighth grade in a given year, having been born in a particular week, or having been married during the same period. Since these sample universes are defined as cohorts, such studies are often called cohort studies. Examples include the various cohorts defined by age and sex in the US National Longitudinal Surveys (US Bureau of Labor Statistics 2000b). A limitation of cohort studies is that their results relate formally only to the particular cohorts from which the sample was selected. An extension of the approach is the multiple cohort design, in which several cohorts are followed in similar panel surveys. Thus, the National Longitudinal Surveys have begun three youth cohort panel surveys, in

1968 (males) and 1968 (females), 1979, and 1997, while the Monitoring the Future studies have selected samples of US high school seniors each year, and follow each sample in a panel survey (Monitoring the Future 2000).

Despite their lack of formal population representation, still other panel surveys are valuable because of their detailed interviews and often very long durations. Examples include Elder's (1999) 30-year follow-up of individuals who had attended Oakland, California elementary schools and were enrolled in the Oakland Growth Study when in fifth and sixth grades and Furstenberg's (Furstenberg et al. 1987) 30-year panel study of Baltimore teen mothers who had enrolled in a teen parenting program in 1965.

1. *Uses of Panel Surveys*

To appreciate the analytic advantages of panel surveys, it is important to realize that the time dimension complicates analyses because: (a) characteristics of population elements may change over time, for instance, an individual may be single one month and married the next; and (b) the population itself may change in composition, with both sample units entering into and exiting from the population.

These complications, especially the first, give rise to a variety of objectives for analysis (Kish 1986, Duncan and Kalton 1987), including:

(a) Estimates of population parameters (e.g., monthly rates of unemployment, public opinion regarding gun control) at distinct time points. The resulting time series of such estimates may be the input for further analysis; see, for example, Smith (1994).

(b) Measurement of net change, defined as change at the aggregate level. An example is the change in the UK unemployment rate between two consecutive months or change in US public opinion regarding gun control before and after a gun-related event.

(c) Measurement of various components of individual change, including: (i) gross change—change at the individual level between two time points (e.g., relating the poverty status of sample members in a given year to their poverty status ten years later); (ii) average change, or trend, for each individual over a period of time, with the average serving to smooth out the effects of isolated perturbations and measurement errors; or (iii) instability for each individual (e.g., the stability of locus of control over 25 consecutive weeks from a sample of elderly in Eizenman et al. 1997).

(d) Aggregation of data for individuals over time. An example is the collection of incomes on an annual basis to relate childhood-long economic well-being to children's developmental outcomes (Blau 1999).

(e) Measuring the frequency, timing, and duration of events occurring within a given time period. Examples are the proportion of persons who were ill in the past two weeks and the average duration of their illnesses, and the complete fertility histories of post-menopausal women.

If the composition of a population is sufficiently stable to be treated as static for the time period of interest, it may be possible to collect the data required to meet all of these objectives from a single cross-sectional survey conducted at the end of the period. Crucial here is that errors of measurement are acceptably small. Unfortunately, respondents' memories are notoriously inadequate when a substantial time is involved or when the characteristics or events have a low degree of salience for the respondents (Bound et al. 2000). For example, respondents are often unable to recall important events such as unemployment (Mathiowetz and Duncan 1984) or hospitalizations (Cannell et al. 1981) within periods as short as several months, and they cannot reliably recall subjective experiences (Moss and Goldstein 1979). Thus, event data often need to be collected at several points throughout the period.

Repeated surveys can meet some of these analytic goals. Their particular strength is that at each round of data collection they routinely select a sample of the population existing at that time. They thus provide a series of cross-sectional estimates that can be used for objective (a) above. They can also be used to measure overall net change (b). The major limitation of a repeated survey is that it does not yield data to satisfy objectives (c) and (d): since elements are not explicitly included in several rounds, none of the components of individual change can be measured and individual data cannot be aggregated across rounds. Data on the frequency and timing of events in a specified period, as in objective (e), can be collected in a repeated survey, but with the measurement error problems noted above.

The major advantage of a panel survey over a repeated survey is its much greater analytical potential. It enables components of individual change to be measured, objective (c), and also the summation of a variable across time, objective (d). By providing high-quality measurements of the frequency, timing, and duration of events, panel surveys provide data useful for the estimating of event-history models (Tuma and Hannan 1984, Allison 1982, Yamaguchi 1991).

A panel survey can be much more efficient than a repeated survey for measuring net change, objective (b). Let the net change be $y_2 - y_1$, where $y_1$and y_2 are the means of the variable of interest at times 1 and 2. Then the variance of the net change is $V(y_2 - y_1) = V(y_1) + V(y_2) - 2\rho[V(y_1)\ V(y_2)]^{1/2}$ where ρ is the correlation between y_1 and y_2. In a repeated survey with two independent samples, $\rho = 0$, but in a panel survey ρ is the correlation between an individual's y-values on the two occasions, which is often quite high. In this case, a panel survey will yield a much more precise estimate of net change, objective (b), than will a repeated survey of the same size. On the other hand, as

discussed below, panel nonresponse between times 1 and 2 may result in more bias in the estimation of y_2 in a panel than repeated survey.

A panel survey also permits the collection of a wider range of data on each sampled element than is possible with a repeated survey. The amount of data collected from any one respondent on one occasion is limited by the burden imposed; however, different data can be collected on different rounds of a panel survey, thus providing the possibility of accumulating an extremely wide range of variables for each sampled element. In the US Survey of Income and Program Participation (SIPP), for instance, core data on income and program participation are collected on each round, and additional data on different topics (e.g., health care utilization and financing, pension and retirement issues, housing conditions) are collected on specific rounds only (US Bureau of the Census 2000).

Some long-term panels have important intergenerational components, with data on children and their family contexts taken in early waves and measurements of the adult attainments of the children taken in later waves. Examples include population panels such as the National Longitudinal Survey youth panels, the PSID, the British Household Panel Survey (Institute for Economic and Social Research 2000), the German Socio-Economic Panel, and more specialized data collections such as the Oakland Growth Study (Elder 1999).

2. *Problems with Panel Surveys*

By attempting repeated interviews with the same sample, panel surveys have problems not found in single or repeated cross-sectional designs, the most important of which is panel nonresponse (initial-wave respondents may not respond in later waves). An additional potential problem with panel surveys is panel conditioning, where responses in a given interviewing round may be conditioned by participation in prior rounds of interviews.

Methods developed to cope with nonresponse bias include minimizing nonresponse in panel surveys and developing statistical adjustments for existing nonresponse. Existing panel surveys typically devote sizeable resources to maintaining high response rates, and sometimes are quite successful. For example, the National Longitudinal Survey of Youth conducted interviews in 1991 with 89 percent of the respondents in its initial, 1979, interview (MaCurdy et al. 1998). Losses in the British National Survey of Health and Development amounted to only 12 percent after 26 years (Atkins et al. 1981).

Incentive payments, respondent reports, persuasion letters, using administrative data for tracing, and collecting extensive contact information (e.g., on friends and relatives not living in the household who would know of address and telephone number changes) help minimize these problems (Freedman et al. 1980, Clarridge et al. 1978, Call et al. 1982).

As with any survey, sizeable nonresponse in a panel survey gives rise to concerns about nonresponse bias. The situation with the first wave of a panel survey corresponds to that with a cross-sectional survey in that very limited information is available on the nonrespondents. The situation with later wave nonresponse in a panel survey is, however, different: in this case a good deal of information is available about later wave nonrespondents from their responses on earlier waves. The earlier wave information can be used to investigate the possibility of nonresponse bias and to develop imputation and weighting nonresponse adjustments that attempt to reduce the bias (Kalton 1986, Lepkowski 1989).

With regard to conditioning, there is ample evidence from several surveys that initial responses in a panel survey differ substantially from those given in subsequent waves (Bailar 1975, 1979, Ghangurde 1982). In the case of the US Current Population Survey, estimates of unemployment from households entering the sample for the first time are almost 10 percent larger than the average over all eight monthly reporting periods. It is not clear whether there is more response bias in the initial or subsequent waves, because the repeated contact with respondents has ambiguous effects on the quality of the data. The crucial question, as yet unanswered for most phenomena reported in surveys, is whether it is merely the reporting of behavior or the behavior itself that is affected by panel membership.

It may be that data collected in subsequent panel waves is less biased, because repeated contact increases the probability that respondents understand the purposes of the study and are thus increasingly motivated to make the effort necessary to give more accurate answers. On the other hand, there is evidence from a validation study (Traugott and Katosh 1979) that extended participation in a panel study on election behavior not only increased the accuracy of responses on voting behavior but may indeed have increased the amount of voting, so that the behavior of the panel was no longer representative of the behavior of the population at large.

It seems unlikely that panel participation has pervasive behavioral effects, especially when changes in the behavior under investigation require more effort than making a trip to the polls. For example, economic behavior such as work effort, saving, commuting, and home ownership are all unlikely to be affected by responses to occasional interviews. Responses to attitudinal questions may be affected by panel membership if participation stimulates interest in the subject matter of the survey.

The limited membership in a rotating panel acts to reduce the problems of panel conditioning and panel loss in comparison with a nonrotating panel survey, and the continual introduction of new samples helps

to maintain an up-to-date sample of a changing population. Rotating panels are used primarily for the estimation of cross-sectional parameters, objective (a), for the estimation of average values of population parameters across a period of time, objective (b), and for measuring net changes, objective (c). A rotating panel survey will generally provide more precise estimates of point of time and, especially, of change parameters than a repeated survey of the same size. Moreover, a rotating panel survey will sometimes have a cost advantage over a repeated survey. This will occur when it is cheaper to conduct a reinterview than an initial interview, as for instance is the case in the US Current Population Survey where initial interviews must be conducted by personal visit whereas reinterviews on some waves may be conducted by telephone (US Bureau of the Census 1978).

The ability of a rotating panel survey to measure components of individual change, objective (c), and to aggregate data for individuals across time, objective (d), is clearly restricted. Since rotating panels are not intended to serve these objectives, they can be designed to avoid the heavy expense of following movers that occurs with nonrotating panel surveys. Thus, for instance, the Current Population Survey employs dwellings, not households or persons, as the sampled units, so that there is no need to follow households or persons moving between panel waves.

In a split panel survey, the panel survey component can be used to measure components of individual change, objective (c), and to aggregate data for individuals over time, objective (d). Its permanent overlap aids in the estimation of net change, objective (b), between *any* two waves whereas the overlap in a rotating panel survey aids only in the estimation of net change between certain prespecified waves.

Both rotating and split panel survey designs provide samples of new entrants to the population and the capacity to use their panel survey components to check on biases from panel conditioning and respondent losses.

3. *The Problem of Changing Populations*

The composition of almost every population of interest, whether of individuals, families, firms, or housing units, changes over time. Individuals enter the population when they are born, immigrate, or attain the age or other status that is used to define the population of interest, and depart when they die, emigrate, move into institutions such as jails, nursing homes, or the military, or in some other way lose that status. Families are 'born' when children leave their parents and set up their own independent households or when a divorce or separation breaks up one family into two; they 'die' when all members of the original household die or when two households are merged into a single one through marriage or less formal living arrangements. Over a survey's time span an element may both enter and leave the population, and may do so more than once, as for instance in the case of a person who has several stays in a nursing home during the course of a survey of the noninstitutionalized population. Comparable changes through construction and demolition, or incorporation, merger, and bankruptcy apply to populations of dwellings and firms.

All surveys over time must address the issues raised by these changes in population composition (Duncan and Kalton 1987). In a repeated survey the cross-sectional parameters of objective (a) relate to the population as it exists at the time the sample is drawn for the particular round of the survey. This procedure reflects both the changing values of the variables under study and the changing population over the time period involved. The estimates are thus average values for a form of average population over time.

Panel studies have adopted a variety of strategies for coping with changing populations. The simplest are panel studies of birth cohorts of individuals in the population, where as many members as possible of the original sample are followed throughout the panel period. In effect, these panels ignore all problems of population composition change except death. Disregarding nonresponse and immigration, these studies represent their respective cohorts as they age, while they gradually cease to represent individuals in the original age range. The National Longitudinal Survey cohorts of Older Men and Mature Women are examples. As discussed above, multiple-cohort panel designs avoid this problem by including in subsequent interviewing waves a sample of new entrants into the population.

Panel designs such as that of the PSID, the British Household Panel Survey, and the European Community Household Panel surveys, contain a mechanism for adding to their samples individuals and families that are 'born' into the population, so they have the potential for maintaining representative samples of nonimmigrant individuals and families throughout their duration. For example, the PSID began with a probability sample of dwellings that, in turn, provides a representative sample of subunits within those dwellings: households, families, subfamilies, transfer program recipiency units, and individuals. The selection probability of each of these units is identical to the selection probability of the dwelling itself. Other panels composed these subsamples with known selection probabilities through other means. With a properly specified set of rules regarding the definition of units, newly formed subunits of interest (including individuals) enter into the sample with known selection probabilities and reflect corresponding changes that are taking place in the population at large, for example, Hill (1992).

Because they originate from existing panel households, newly formed families and individuals are more

'clustered' by family lineage in these designs than in repeated cross-sectional samples; however, the statistical inefficiency of this clustering is heavily outweighed by the analytic advantages of being able to link these newly formed units with information on their families of origin. Thus, information on children setting up independent households can be matched to reliably reported information on their parents, information on ex-spouses can be linked together if the marital split occurred during the panel period, etc. Immigration during the panel period is usually ignored although can be a serious problem for a long-term panel survey confined to a city or a local area, but it is a less serious problem for a national panel survey.

4. *Examples of What Panel Surveys Have Found*

To help convey the analytic advantages of panel surveys, we list in this section some prominent examples of panel-survey based analyses.

4.1 *Poverty and Welfare Dynamics*

Trends in annual US poverty rates (i.e., the fraction of the population with family cash incomes below a set of 'poverty thresholds' that vary with family size) are tracked through cross-sectional estimates based on the Current Population Survey. Net change from one year to the next in these rates rarely amount to as much as a single percentage point (US Bureau of the Census 1999), producing perceptions that 'the poor' are an ever-present and little-changing group.

By applying event history methods to the time series of annual poverty measurements for PSID sample members, Bane and Ellwood (1986) characterized the nature and determinants of poverty experiences by the length of their 'spells' (i.e., continuous periods of poverty). They find that a clear majority (60 percent) of families who first begin a poverty experience do so for short (i.e., spells of 1 or 2 years) periods of time. On the other hand, a subset of poor families (14 percent) have longer-run (8 or more years) spells, and the remaining one-quarter (26 percent) have medium-length (3–7 years) spells. Thus, there is no single 'correct' characterization of poverty—transitory or persistent—since poverty experiences are a mixture of transitory and long-term.

By combining individual spells into simulated lifetime histories, Bane and Ellwood (1994) provide estimates of the likely total number of years of receipt for families just starting to receive cash assistance from the US Aid to Families With Dependent Children program. They find a roughly even distribution of first-time welfare recipients across the three time intervals; roughly one-third (36 percent) have very short welfare experiences, a third (35 percent) medium-length experiences, and the final third (29 percent) long-term receipt. With welfare, as with poverty, heterogeneity is a key feature, with many families using AFDC as short-term insurance, but a substantial fraction of recipients receiving long-term assistance.

4.2 *Intergenerational Transmission of Economic Status*

Correlations in earning between fathers and sons is sometimes treated as a useful measure of intergenerational (im)mobility. Solon (1992) argues that there are two important sources of bias in attempts to estimate such correlations: single year proxies for long-run economic status contain transitory fluctuations that, under plausible assumptions, impart a downward bias to the correlation estimates, and samples used in previous studies (e.g., Mormon brothers, white twins who served in the armed forces) tend to be peculiarly homogeneous, also resulting under plausible assumptions in a downward bias to estimates based on homogeneous than population-based samples.

Data from both the NLSY (in Zimmerman 1992) and PSID (in Solon 1992) have been used to gauge the likely biases. For example, Solon et al. (1992) find that increasing the measurement interval for economic status from one to several years causes their estimate of the father-son earning correlation to jump from 0.25, a number consistent with past studies based on single-year earnings measurement and homogeneous samples, to 0.45. Thus, intergenerational economic mobility appears to be substantially lower than had been depicted in studies based on homogeneous samples and single-year measures of income or earnings.

The more specialized panel studies of Elder and Furstenberg provide much more detail on the process of intergenerational mobility. Using 30-year follow-up data on the Oakland Growth Study cohorts, Elder (1999) found in these children of the Great Depression strong associations among economic hardship, parental psychological well-being, and adolescent well-being in intact families.

Fathers who experienced job loss and economic deprivation were more distressed psychologically and prone to explosive, rejecting, and punitive parenting. Preschool-age children in these families, especially boys, were more likely to exhibit problem behaviors, while adolescent girls were more likely to have lower feelings of self-adequacy and to be less goal-oriented. Elder (1999) speculated that the gender and age differences reflected different experiences in families during the deprived times. Adolescent boys reacted to family economic hardship by looking for economic opportunities outside of the home. This time spent taking on responsibility in extra-family contexts reduced the time they spent with their families and may have provided a buffer to the effects of negative family

interactions. Younger children and girls were unable to seek such extra-familial opportunities as readily and therefore did not have access to buffering opportunities.

Furstenberg et al. (1987) take a detailed look at intra- and intergenerational mobility in a panel study of a poor teen mothers participating in a program in a Baltimore hospital. Their 20-year follow-up revealed considerably diversity in the mothers' economic and demographic statuses, with most self-supporting but very few stably married. Most (60 percent) of the daughters of the teen mothers had not themselves become teen mothers and three-quarters enjoyed at least a modest degree of economic success. In contrast, fewer than half of the sons of the teen mothers had attained successful economic trajectories.

4.3 Unemployment and Labor Turnover

Both unemployment and job turnover are most naturally thought about in a 'spells' framework: how long does a spell of unemployment last before employment recommences? How long does a spell of employment with one firm last before the worker leaves to take (or look for) alternative employment? Given the inherently longitudinal nature of these processes, it is not surprising that panel surveys have played an important role in research on these topics. While much research on this topic has used administrative records, a panel survey has the advantages of including those who do not receive unemployment benefits, for comparison to recipients, and it continues observing spells after benefits have been exhausted (i.e., it measures total spells, not just insured spells), so one can observe changes in job-finding rates after benefits are exhausted.

An example is Katz and Meyer's (1990) study of the impact of unemployment insurance on unemployment duration. Katz and Meyer find that probability of recall declines as the duration of a spell increases, for both recipients and nonrecipients. New-job rates, however, are upward-sloping for recipients but relatively flat for nonrecipients. Moreover, there are spikes in both the recall and new-job hazard rates around 26 and 39 weeks (when benefits were typically exhausted in their sample period) for recipients but not for nonrecipients. Thus, not only is there evidence that the approaching end of benefits conditions recipients' job finding, but it also conditions firms' recall policies.

4.4 Antecedents of Life Events

There is great interest in understanding the consequences of important life events such as divorce, widowhood, unemployment, and retirement. Cross-sectional data provide comparisons of, for example, the living standards and labor supply of divorced women and 'otherwise similar' married women, or the health of unemployed or widowed men with 'otherwise similar' employed or married men. But panel surveys provide valuable post- and, interestingly, pre-event data on the same individuals.

For example, Johnson and Skinner (1986) use data from the PSID to examine work hours of women between seven years before and four years after divorce or separation. Consistent with other research, they find that the average work hours of these women rose by half (from 1024 to 1551 hours) between one year before and four years after the divorce or separation. But Johnson and Skinner (1986) also found a big jump (from 744 to 1024 hours) over the six-year period *prior* to the divorce or separation.

Zick and Smith (1991) analyze the economic antecedents and consequences of widowhood by comparing family-income trajectories of a control group of intact couples with widowed women and men between five years before and five years after the spouse's death. They too find differences prior to the event, with income changes over the five years *prior* to the death accounting for about one-quarter of the control-widow difference at $t+5$ and for fully half of the control-widower difference. A clear message from both of these analyses is that impending life events such as divorce and death produce measurable economic and social changes long before the events actually occur. An implication is that fairly long panels are needed to capture the complete dynamic.

See also: Databases, Core: Demography and Registers; Economic Panel Data; Large-scale Social Survey, Analysis of; Longitudinal Research: Panel Retention; Sample Surveys: Methods; Sample Surveys: Survey Design Issues and Strategies; Sample Surveys: The Field; Survey Research: National Centers; Surveys and Polling: Ethical Aspects

Bibliography

Allison P 1982 Discrete-time methods for the analysis of event histories. In: Leinhardt S (ed.) *Sociological Methodology 1982*. Jossey-Bass, San Francisco, pp. 61–98

Atkins E, Cherry N, Douglas J W B, Kieman K E, Wadsworth M E J 1981 The 1946 British birth cohort: an account of the origins, progress, and results of the National Survey of Health and Development. In: Mednick S A, Baert A E (eds.) *Prospective Longitudinal Research*. Oxford University Press, Oxford, UK, pp. 5–30

Bailar B A 1975 The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association* **70**: 23–30

Bailar B A 1979 Rotation sample biases and their effects on estimates of change. *Bulletin of the Statistics Institution* **48**(2): 385–407

Bane M J, Ellwood D T 1986 Slipping in and out of poverty: the dynamics of spells. *Journal of Human Resources* **21**: 1–23

Bane M J, Ellwood D T 1994 *Welfare Realities*. Harvard University Press, Cambridge, MA

Blau D 1999 The effect of income on child development. *Review of Economics and Statistics* **81**(2): 261–76
Bound J, Brown C, Mathiowetz N 2000 Measurement error in survey data. In: Heckman J, Leamer E (eds.) *Handbook of Econometrics*. North Holland, Amsterdam, Vol. 5
Call V R A, Otto L B, Spenner K I 1982 *Tracking Respondents: A Multi-Method Approach*. Lexington Books, Lexington, MA
Cannell C F, Miller P V, Oksenberg L 1981 Research on interviewing techniques. In: Leinhardt S (ed.) *Sociological Methodology, 1981*. Jossey-Bass, San Francisco pp. 389–437
Clarridge B R, Sheehy L L, Hauser T S 1978 Tracing members of a panel: A 17-year follow-up. In: Schuessler K F (ed.) *Sociological Methodology, 1978*. Jossey-Bass, San Francisco, pp. 185–203
Davis J, Smith T W 1994 *General Social Surveys, 1972–1994: Cumulative Codebook*. National Opinion Research Center, Chicago
Deutsches Institut für Wirtschaftsforchung 2000 German Socio-Economic Panel Home Page, http://www.diw-berlin.de/soep/
Duncan G, Kalton G 1987 Issues of design and analysis of surveys across time. *International Statistical Review* **55**: 97–117
Elder G 1999 *Children of the Great Depression: Social Change in Life Experience*, 25th Anniversary Edition. Westview, Boulder, CO
Eizenman D R, Nesselroade J R, Featherman D L, Rowe J W 1997 Intra-individual variability in perceived control in an older sample: The MacArthur successful aging studies. *Psychology and Aging* **12**: 489–502
Freedman D, Thornton A, Camburn D 1980 Maintaining response rates in longitudinal studies. *Sociological Methods and Research* **9**: 87–9
Furstenberg Jr F F, Brooks-Gunn J, Morgan S P 1987 *Adolescent Mothers in Later Life*. Cambridge University Press, New York
Ghangurde P D 1982 Rotation group bias in the LFS estimates. *Survey Methodology* **8**: 86–101
Hill M S 1992 *The Panel Study of Income Dynamics*. Sage, Newbury Park, CA
Institute for Economic and Social Research, University of Essex 2000 British Household Panel Survey Home Page, http://www.irc.essex.ac.uk/bhps
Institute of Education, Center for Longitudinal Studies 2000 British National Child Development Study Home Page, http://www.cls.ioe.ac.uk/Ncds/nhome.htm
Johnson W, Skinner J 1986 Labor supply and marital separation. *American Economic Review* **76**(3): 455–69
Kalton G 1986 Handling wave nonresponse in panel surveys. *Journal of Official Statistics* **2**: 303–14
Katz L, Meyer B 1990 The impact of potential duration of unemployment benefits on the duration of unemployment. *Journal of Public Economics* **41**(1): 45–72
Kish L 1986 Timing of surveys for public policy. *Australian Journal of Statistics* **28**: 1–12
Lepkowski J 1989 The treatment of wave nonresponse in panel surveys. In: Kasprzyk D, Duncan G, Kalton G, Singh M (eds.) Panel Survey Design and Analysis. Wiley, New York pp. 348–74
MaCurdy T, Mroz T, Gritz R M 1998 An evaluation of the National Longitudinal Survey on Youth. *Journal of Human Resources* **33**(2): 345–436
Mathiowetz N Duncan G 1984 Temporal patterns of response errors in retrospective reports of unemployment and occupation. In: *Proc. Survey Res. Meth. Sect.* American Statistical Association, Washington, DC, pp. 652–7
Monitoring the Future 2000 Monitoring the Future Home Page, http://monitoringthefuture.org/
Moss L, Goldstein H 1979 *The Recall Method in Social Surveys*. University of London Institute of Education, London
Solon G 1992 Intergenerational income mobility in the United States. *American Economic Review* **82**(3): 393–408
Smith T W 1994 Is there real opinion change? *International Journal of Public Opinion Research* **6**: 187–203
Statistics Canada 2000 Guide to the Labour Force Survey Statistics Canada web site, http://www.statcan.ca/
Traugott M, Katosh K 1979 Response validity in surveys of voting behavior. *Public Opinion Quarterly* **79**: 359–77
Tuma N B, Hannan M T 1984 *Social Dynamics: Models and Methods*. Academic Press, Orlando, FL
US Bureau of the Census 1978 *The Current Population Survey: Design and Methodology*. Technical Paper 40. US Government Printing Office, Washington, DC
US Bureau of the Census 1999 Poverty in the United States: 1998 (P60–207). US Government Printing Office, Washington, DC
US Bureau of the Census 2000 Survey of Income and Program Participation Home Page, http://www.sipp.census.gov/sipp/sipphome.htm
US Bureau of Labor Statistics 2000a Current Population Survey Home Page, http://www.bls.census.gov/cps/cpsmain.htm
US Bureau of Labor Statistics 2000b National Longitudinal Surveys Home Page, http://stats.bls.gov/nlshome.htm
Yamaguchi K 1991 *Event History Analysis*, Applied Social Research Methods, Vol. 28. Sage, Newbury Park, CA
Zick C, Smith K 1991 Patterns of economic change surrounding the death of a spouse. *Journal of Gerontology* **46**(6): S310–20
Zimmerman D J 1992 Regression toward mediocrity in economic stature, *American Economic Review* **82**(3): 409–29.

G. J. Duncan

Panic Disorder

Panic disorder (PD) is an anxiety disorder associated with recurrent, unexpected panic attacks, characterized by sudden surges of physical arousal and fear. In addition, most individuals with PD develop agoraphobia to varying degrees. Agoraphobia is a fear of situations in which escape might be difficult or embarrassing, or in which help might be unavailable in the event of a panic attack.

The terms 'PD' and 'panic attack' first appeared in the diagnostic nomenclature with the publication of the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III, American Psychiatric Association 1980). Initially, panic attacks were seen primarily as a feature of agoraphobia. In DSM-III, agoraphobia could be diagnosed with or without panic attacks, and individuals who experienced unexpected panic attacks in the absence of agoraphobia received a diagnosis of PD without agoraphobia.

More recently (e.g., DSM-IV, American Psychiatric Association 1994), the relationship between panic

disorder and agoraphobia has reversed, such that PD is seen as the primary problem, and agoraphobia is viewed as a complication of PD. The disorder 'agoraphobia with panic attacks' has essentially been replaced with 'panic disorder with agoraphobia' to reflect the fact that, in clinical groups, agoraphobia tends to develop as a response to having unexpected panic attacks, rather than the other way around (Craske and Barlow 1988).

1. Definitions and Description

1.1 Panic Attack

DSM-IV (American Psychiatric Association 1994) defines a panic attack as a period of intense fear of discomfort that begins quickly and peaks in ten minutes or less. A panic attack must also be accompanied by at least four out of a list of 13 symptoms, which includes eleven physical sensations (i.e., racing or pounding heart, sweating, trembling, shortness of breath, choking feelings, chest discomfort, nausea, dizziness, feelings of unreality or depersonalization, numbness or tingling sensations, hot flashes or chills) and two cognitive symptoms (i.e., fear of losing control or going crazy, fear of dying).

Panic attacks are common in the general population and are also associated with a broad range of psychological problems. For example, people who experience intense worry may have a panic attack in response to a particular worrisome thought. Individuals who are fearful of a specific situation (e.g., seeing a spider, being in a high place) may experience panic attacks when confronted with these situations. Panic attacks may also occur in the absence of any obvious cue or trigger. In fact, the hallmark symptom of PD is the presence of panic attacks that occur out of the blue.

1.2 Panic Disorder

People who suffer from PD experience recurrent unexpected or uncued panic attacks. In addition, they must experience, for at least one month, one or more of the following: (a) persistent worry about having more panic attacks; (b) frequent worry about the implications of consequences of the attacks (e.g., that an attack will lead to a heart attack, death, fainting, diarrhea, vomiting, embarrassment, or some other catastrophe); or (c) a significant change in behavior related to the attacks (e.g., avoidance of certain situations or activities). Finally, before assigning a diagnosis of PD, the clinician must establish that the panic attacks are not directly related to another psychological problem (e.g., a specific phobia), a medical condition (e.g., hyperthyroidism), or use of a substance (e.g., cocaine, caffeine). Panic disorder may be diagnosed with or without agoraphobia.

1.3 Agoraphobia

A significant proportion of individuals suffering from PD develop agoraphobia, a fear of situations in which escape might be difficult, or in which help might not be available in the event of a panic attack (American Psychiatric Association 1994). Typical situations that are often avoided by people with agoraphobia include crowded places (e.g., concerts, movies, shopping malls), driving, traveling, public transportation, flying, enclosed places, and meetings. For some individuals with agoraphobia, it is the distance from home, rather than the situation itself, that determines the level of fear and avoidance. For example, driving may be difficult, but only if the car ventures more than a few kilometers from home. For many people who suffer from agoraphobia, entering feared situations is easier if accompanied by a safe person, such as a spouse, family member, or close friend. In extreme cases, agoraphobia may be associated with a complete inability to leave the home, even when accompanied.

In addition to the obvious avoidance associated with agoraphobia, individuals with PD often engage in more subtle forms of avoidance. For example, a person might see a movie in a theater, but only if sitting in an aisle seat, near the exit. Or an individual might agree to ride a bus only while engaging in distracting behaviors such as reading or listening to a portable radio. Individuals with PD may also carry objects with them to feel safer, including medication, extra money in case of an emergency, or a portable telephone. Finally, they may engage in overprotective behaviors such as frequently checking their heart rate or blood pressure.

2. Epidemiology, Onset, and Demographics

Antony and Swinson (2000) reviewed the literature concerning age of onset and risk factors for developing PD. PD begins on average when a person is in his or her mid- to late-20s, although onset can occur any time between childhood and late in life. Risk factors include a period of life stress preceding the first attack, being female, a history of childhood separation anxiety, and a history of either substance abuse or dependence (Antony and Swinson 2000).

In the National Comorbidity Survey (a study of over 8,000 Americans in the general population), 3.5 percent of individuals reported symptoms meeting criteria for panic disorder with or without agoraphobia at some time in their life (Kessler et al. 1994). The lifetime prevalence of agoraphobia was 6.7 percent, with about two-thirds of agoraphobics having no prior history of panic attacks or panic disorder (Magee et al. 1996). This finding is in direct contrast to studies based on individuals in clinical settings, where agoraphobia almost never occurs separately from panic disorder (Pollard et al. 1989).

People with PD report impairment across a broad range of life domains including work, recreation, and social functioning (Antony et al. 1998). Furthermore, PD is associated with enormous costs to sufferers and to society, in the form of increased heath-care utilization, lost wages, and time off work (e.g., Greenberg et al. 1999). PD is also associated with a higher risk of developing other psychological problems, including other anxiety disorders, depression, certain personality disorders (e.g., dependent, avoidant), and substance-use disorders (for a review, see Antony and Swinson 2000).

3. *Biological Approaches to Understanding Panic Disorder*

3.1 *Biological Models of Panic Disorder*

Most biological approaches to understanding PD have assumed a dysfunction in one or more neurotransmitter systems, including norepinephrine (NE), serotonin (5-HT), and cholecystokinin (CCK). In addition, theorists have hypothesized that genetic and neuroanatomical factors are also involved in the development of PD.

Perhaps the most influential biological model of PD in recent years has been Donald Klein's suffocation alarm model (Klein 1993), which proposes that panic attacks are caused by a pathological misfiring of the suffocation alarm system. According to Klein, humans are equipped with two types of alarm reactions. The first is controlled by the hypothalamic–pituitary–adrenal (HPA) axis and is responsible for dealing with real or perceived emergency situations. It is this alarm system that is activated when an individual is confronted with a feared object or situation.

The second alarm system is the suffocation alarm system, which is mediated by an area of the brain known as the *locus ceruleus*. This alarm system is triggered by increased levels of carbon dioxide in the blood, a sign that individuals might be in a small enclosed area, forced to breathe their own exhaled air, which is rich in carbon dioxide. Klein hypothesized that unexpected panic attacks occur when an individual has a pathologically lowered threshold for the activation of the suffocation alarm system. Presumably, the brains of such individuals are prone to mistakenly conclude that there is not enough air to breathe. As reviewed elsewhere (Antony and Swinson 2000), evidence regarding Klein's model has been mixed.

3.2 *Evidence Regarding Biological Approaches to Panic*

Despite inconsistent evidence for Klein's model, there are many studies supporting the view that biological processes are involved in the pathophysiology of PD. PD tends to run in families and evidence from twin studies suggests that genetics plays a significant role in the transmission of PD across generations (e.g., Kendler et al. 1992). Findings from brain-imaging studies suggest that the temporal areas of the brain (particularly the right hippocampal region) are involved in the experience of panic attacks and fear (for a review, see Antony and Swinson 2000).

Evidence regarding the role of neurotransmitters in PD have included studies of altered receptor functioning, levels of particular neurotransmitters and related metabolites, and responses to substances that are known to cause changes in neurotransmitter levels. Among neurotransmitters, NE is the one most often implicated in PD and agoraphobia (Sullivan et al. 1999), although other neurotransmitter systems, such as 5-HT (Bell and Nutt 1998) and CCK (Bradwejn 1995) may also play a role. There is evidence that levels of NE metabolites predict panic-related symptoms (Garvey et al. 1990) and of altered NE receptor sensitivity in people suffering from PD (Nutt 1989). In addition, drugs that block the reuptake of NE appear to be effective for preventing panic attacks (Mavissakalian and Perel 1989).

Finally, a large number of studies have shown that certain substances that affect the NE system are more likely to trigger panic attacks in people with PD than in various comparison groups. Infusions of sodium lactate, carbon-dioxide inhalation, and yohimbine injections have all been shown to trigger panic attacks. However, it should be noted that psychological factors such as perceived control over the procedure and the presence of a safe person can influence whether an individual will experience a panic attack following exposure to these substances (for a review of panic induction studies, see Antony and Swinson 2000).

4. *Psychological Approaches to Understanding Panic Disorder*

4.1 *Psychological Models of Panic Disorder*

Attempts to explain PD using a psychological framework have come from a broad range of theoretical perspectives, including cognitive theory, learning theory, and psychodynamic theory. However, the most influential of these models have been those developed by cognitive and behavioral theorists such as David M. Clark (Clark 1986) and David H. Barlow (Barlow 1988).

From a cognitive behavioral perspective, PD is viewed as being maintained by a fear of benign physical sensations. For example, according to David Clark's cognitive model of PD, panic attacks stem from a tendency to misinterpret normal physical sensations (e.g., racing heart, dizziness) as indicating

some immediate danger. As a result of interpreting these sensations as being threatening, the symptoms escalate, leading to further catastrophic interpretations and predictions. Very quickly, the symptoms spiral into a full-blown panic attack, seemingly out of the blue. Clark and other cognitive theorists acknowledge that biological factors may also contribute to the experience of panic. The sensations that initially start the panic cycle may be caused by caffeine, intense emotions, exercise, hormonal changes, fatigue, hunger, hyperventilation, attention to bodily feelings, or other factors. Consistent with the cognitive behavioral view, 70 percent of individuals with PD report that their panic attacks are triggered by a physical sensation (Breitholtz et al. 1998).

4.2 *Evidence Regarding Psychological Approaches to Panic*

A number of studies have confirmed that people with PD are fearful of the physical sensations associated with panic attacks, compared to people without PD and people with other anxiety disorders (e.g., Taylor et al. 1992). In addition, researchers have demonstrated that people with PD are more likely than others to interpret ambiguous physical sensations as a sign of some impending physical or mental disaster and to believe these thoughts quite strongly (Clark et al. 1997).

Numerous studies have also found that people with PD have information-processing biases that are consistent with their fears. For example, a number of studies have found that PD is associated with a heightened ability mentally to track physical sensations such as heart rate (e.g., van der Does et al. 2000). In addition, results from studies of attention and memory have also been consistent with psychological models of PD. Compared to people without PD, individuals with PD attend more to threat-related information (e.g., McNally et al. 1994) and are more likely than nonanxious individuals to show threat-related biases in memory (e.g., Amir et al. 1996).

In summary, individuals with PD are typically more fearful of and more vigilant for panic-related sensations, compared to people without PD. PD is also associated with a tendency to hold fearful beliefs about panic-related sensations, and to show biases in attention and memory for threat-related information.

5. *Treatment of Panic Disorder*

In 1998, the American Psychiatric Association published practice guidelines for the treatment of PD. In addition to discussing important issues in the assessment and treatment of this condition, the guidelines reviewed the main approaches that have been shown to be useful for treating PD. Essentially, these approaches include treatment with medications, and treatment with cognitive-behavior therapy (CBT). A review of studies on the relative and combined effects of these interventions is available elsewhere (e.g., Antony and Swinson 2000). A brief summary of the major findings is provided below.

5.1 *Medication Treatments*

A large number of controlled clinical trials have established that certain medications are useful for treating PD with and without agoraphobia. The tricyclic antidepressant imipramine is the most extensively studied medication that has been shown to be effective for blocking panic attacks (e.g., Mavissakalian and Perel 1989). More recently, controlled trials have established that the selective serotonin reuptake inhibitors or SSRIs (e.g., fluoxetine, fluvoxamine, sertraline, paroxetine, citalopram) are also useful for treating PD; in fact, the SSRIs are now often considered to be the pharmacological treatment of choice for this condition (for a review, see Antony and Swinson 2000).

Finally, controlled trials have found a number of other antidepressants to be useful, including venlafaxine (a new antidepressant) (Pollack et al. 1996) and phenelzine (a monoamine oxidase inhibitor or MAOI) (Buigues and Vallejo 1987). Studies with high-potency benzodiazepine anxiolytics such as alprazolam have consistently shown these medications to be effective as well (e.g., Ballenger et al. 1988).

Despite the long list of medication options for treating PD, there is little consistent evidence that any one of these medications is more effective than the others. Choosing among medications for a particular individual usually depends on factors such as potential side effects, possible problems during discontinuation (e.g., rebound anxiety), the patient's previous history of treatment response, the cost of the drug, and potential interactions with other medications, substances, or medical illnesses.

5.2 *Cognitive-behavior Therapy*

Cognitive-behavior therapy for PD includes four main types of strategies: (a) exposure to feared situations such as crowds, public transportation, and being away from home alone (i.e., *in vivo* exposure); (b) exposure to exercises that trigger feared sensations (e.g., overbreathing or spinning to induce feelings of dizziness); (c) cognitive therapy (i.e., systematically considering evidence that contradicts fearful thinking); and (d) relaxation-based strategies (e.g., learning to slow down one's breathing). Most of the time, treatment includes various combinations of these strategies.

A number of comprehensive reviews on CBT for PD have been published (e.g., Antony and Swinson 2000,

Barlow and Brown 1996, McNally 1994). Generally, there is little evidence that it is necessary to include all four treatment components to achieve maximum results. *In vivo* exposure to feared situations appears to be particularly important for treating individuals who are agoraphobic. In addition, relaxation-based strategies are thought to be less important than the exposure-based and cognitive techniques.

Although CBT for PD lasts 10 to 15 sessions, a number of recent studies suggest that gains can be made relatively quickly, with a minimal investment of therapist time. For example, telephone-administered CBT has been used to treat people who do not have easy access to a therapist (Swinson et al. 1995) and a single session of CBT has been shown to prevent the onset of PD in a significant percentage of individuals who present to hospital emergency rooms with a panic attack (Swinson et al. 1992). Finally, a number of studies have found that brief CBT lasting only five sessions (Clark et al. 1999) and CBT delivered in a self-help format (e.g., Hecker et al. 1996) are both useful for helping people to overcome PD (see *Cognitive and Interpersonal Therapy: Psychiatric Aspects*).

5.3 Combination Treatments

Studies comparing medications, CBT, and their combinations have yielded inconsistent results (Antony and Swinson 2000). However, averaging across a large number of studies, it is reasonable to conclude that both medications and CBT are effective for treating PD and there is no clear evidence favoring either antidepressants, high-potency benzodiazepines, CBT, or their combinations for treating panic attacks (e.g., Bakker et al. 1998, van Balkom et al. 1997). However, the combination of antidepressants and *in vivo* exposure appears to be particularly useful for treating agoraphobic avoidance (van Balkom et al. 1997).

It should be noted that most studies on the treatment of PD have focused primarily on the short-term effects of treatment over a period of only two to three months. Few studies have examined the relative effects of medications and CBT after treatment is discontinued. Exceptions include three studies comparing the relative and combined effectiveness of CBT with either imipramine (Barlow et al. 2000), alprazolam (Marks et al. 1993), or fluvoxamine (Sharp et al. 1996). In all three studies, individuals who received CBT (alone or in combination with medication) were more likely to maintain their gains over the long term than individuals who received only medication or placebo treatment.

Taken together, the literature on treating PD suggests that CBT and pharmacotherapy are both effective treatments in the short term. However, there is evidence that CBT is a more effective treatment for ensuring that gains are maintained over the long term.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Fear Conditioning; Fear: Potentiation of Startle; Fear: Psychological and Neural Aspects; Mental and Behavioral Disorders, Diagnosis and Classification of; Panic, Sociology of;

Bibliography

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. (DSM-III). American Psychiatric Association, Washington, DC

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. (DSM-IV). American Psychiatric Association, Washington, DC

American Psychiatric Association 1998 Practice guideline for the treatment of patients with panic disorder. *American Journal of Psychiatry* **155**, suppl.: 1–34

Amir N, McNally R J, Riemann B C, Clements C 1996 Implicit memory bias for threat in panic disorder: Application of the 'white noise' paradigm. *Behaviour Research and Therapy* **34**: 157–62

Antony M M, Roth D, Swinson R P, Huta V, Devins G M 1998 Illness intrusiveness in individuals with panic disorder, obsessive compulsive disorder, or social phobia. *Journal of Nervous and Mental Disease* **186**: 311–15

Antony M M, Swinson R P 2000 *Phobic Disorders and Panic in Adults: A Guide to Assessment and Treatment*. American Psychological Association, Washington, DC

Bakker A, van Balkom A J L M, Spinhoven P, Blaauw B M J W, van Dyck R 1998 Follow-up on the treatment of panic disorder with or without agoraphobia: A quantitative review. *Journal of Nervous and Mental Disease* **186**: 414–19

Ballenger J C, Burrows G D, DuPont R L, Lesser I M, Noyes R, Pecknold J C, Rifkin A, Swinson R 1988 Alprazolam in panic disorder and agoraphobia: Results from a multicenter trial. *Archives of General Psychiatry* **45**: 413–22

Barlow D H 1988 *Anxiety and its Disorders: The Nature and Treatment of Anxiety and Panic*. Guilford, New York

Barlow D H, Brown T A 1996 Psychological treatments for panic disorder and panic disorder with agoraphobia. In: Mavissakalian M R, Prien R F (eds.) *Long-term Treatments of Anxiety Disorders*. American Psychiatric Press, Washington, DC, pp. 221–40

Barlow D H, Gorman J M, Shear M K, Woods S W 2000 A randomized controlled trial of cognitive-behavioral treatment vs. imipramine and their combination for panic disorder: Primary outcome results. *JAMA—Journal of the American Medical Association* **283**: 2529–36

Bell C J, Nutt D J 1998 Serotonin and panic. *British Journal of Psychiatry* **172**: 465–71

Bradwejn J 1995 Cholecystokinin and panic disorder. In: Bradwejn J, Vasar E (eds.) *Cholecystokinin and Anxiety: From Neuron to Behavior*. R. G. Landes, Austin, TX, pp. 73–86

Breitholtz E, Westling B E, Öst L G 1998 Cognitions in generalized anxiety disorder and panic disorder. *Journal of Anxiety Disorders* **12**: 567–77

Buigues J, Vallejo J 1987 Therapeutic response to phenelzine in patients with panic disorder and agoraphobia with panic attacks. *Journal of Clinical Psychiatry* **48**: 55–9

Clark D M 1986 A cognitive approach to panic. *Behaviour Research and Therapy* **24**: 461–70

Clark D M, Salkovskis P M, Hackmann A, Wells A, Ludgate J, Gelder M 1999 Brief cognitive therapy for panic disorder: A randomized controlled trial. *Journal of Consulting and Clinical Psychology* **67**: 583–89

Clark D M, Salkovskis P M, Öst L G, Breitholtz E, Koehler K A, Westling B E, Jeavons A, Gelder M 1997 Misinterpretation of body sensations in panic disorder. *Journal of Consulting and Clinical Psychology* **65**: 203–13

Craske M G, Barlow D H 1988 A review of the relationship between panic and avoidance. *Clinical Psychology Review* **8**: 667–85

Garvey M, Noyes R, Cook B 1990 Comparison of panic disordered patients with high versus low MHPG. *Journal of Affective Disorders* **20**: 7–12

Greenberg P E, Sisitsky T, Kessler R C, Finkelstein S N, Berndt E R, Davidson J R T, Ballenger J C, Fyer A J 1999 The economic burden of anxiety disorders in the 1990s. *Journal of Clinical Psychiatry* **60**: 427–35

Hecker J E, Losee M C, Fritzler B K, Fink C M 1996 Self-directed versus therapist-directed cognitive behavioral treatment for panic disorder. *Journal of Anxiety Disorders* **10**: 253–65

Kendler K S, Neale M C, Kessler R C, Heath A C, Eaves L J 1992 The genetic epidemiology of phobias in women: The interrelationship of agoraphobia, social phobia, situational phobia, and simple phobia. *Archives of General Psychiatry* **9**: 273–81

Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshleman S, Wittchen, H U, Kendler K 1994 Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**: 8–19

Klein D F 1993 False suffocation alarms, spontaneous panics, and related conditions: An integrative hypothesis. *Archives of General Psychiatry* **50**: 306–17

Magee W J, Eaton W W, Wittchen H U, McGonagle K A, Kessler R C 1996 Agoraphobia, simple phobia, and social phobia in the national comorbidity survey. *Archives of General Psychiatry* **53**: 159–68

Marks I M, Swinson R P, Basoglu M, Kuch K, Noshirvani H, O'Sullivan G, Lelliott P T, Kirby M, McNamee G, Sengun S, Wickwire K 1993 Alprazolam and exposure alone and combined in panic disorder with agoraphobia: A controlled study in London and Toronto. *British Journal of Psychiatry* **162**: 776–87

Mavissakalian M R, Perel J M 1989 Imipramine dose-response relationship in panic disorder with agoraphobia: Preliminary findings. *Archives of General Psychiatry* **46**: 127–31

McNally R J 1994 *Panic Disorder: A Critical Analysis*. Guilford, New York

McNally R J, Amir N, Louro C E, Lukach B M, Riemann B C, Calamari J E 1994 Cognitive processing of ideographic emotional information in panic disorder. *Behaviour Research and Therapy* **32**: 119–22

Nutt D J 1989 Altered central α_2-adrenoceptor sensitivity in panic disorder. *Archives of General Psychiatry* **46**: 165–9

Pollack M H, Worthington J J, Otto M W, Maki K M, Smoller J W, Manfro G G, Rudolph R, Rosenbaum J F 1996 Venlafaxine for panic disorder: Results from a double-blind placebo-controlled study. *Psychopharmacology Bulletin* **32**: 667–70

Pollard C A, Bronson S S, Kenney M R 1989 Prevalence of agoraphobia without panic in clinical settings. *American Journal of Psychiatry* **146**: 559

Sharp D M, Power K G, Simpson R J, Swanson V, Moodie E, Anstee J A, Ashford J J 1996 Fluvoxamine, placebo, and cognitive behaviour therapy used alone and in combination in the treatment of panic disorder and agoraphobia. *Journal of Anxiety Disorders* **10**: 219–42

Sullivan G M, Coplan J D, Kent J M, Gorman J M 1999 The Noradrenergic system in pathological anxiety: A focus on panic with relevance to generalized anxiety and phobias. *Biological Psychiatry* **46**: 1205–18

Swinson R P, Fergus K D, Cox B J, Wickwire K 1995 Efficacy of telephone-administered behavioral therapy for panic disorder with agoraphobia. *Behaviour Research and Therapy* **33**: 465–9

Swinson R P, Soulios C, Cox B J, Kuch K 1992 Brief treatment of emergency room patients with panic attacks. *American Journal of Psychiatry* **149**: 944–6

Taylor S, Koch W J, McNally R J 1992 How does anxiety sensitivity vary across the anxiety disorders? *Journal of Anxiety Disorders* **6**: 249–59

van Balkom A J L M, Bakker A, Spinhoven P, Blaauw B M J W, Smeenk S, Ruesink B 1997 A meta-analysis of the treatment of panic disorder with or without agoraphobia: A comparison of psychopharmacological, cognitive-behavioral, and combination treatments. *Journal of Nervous and Mental Disease* **185**: 510–16

van der Does A J W, Antony M M, Barsky A J, Ehlers A 2000 Heartbeat perception in panic disorder: A re-analysis. *Behaviour Research and Therapy* **38**: 47–62

M. M. Antony

Panic, Sociology of

The term 'panic' is widely used in everyday speech as well as in the literature of different professional areas and scientific disciplines. This leads to a very wide diversity in the attribution of both the characteristics of, and the conditions that generate, the phenomenon. However, this article confines itself primarily to how sociologists view 'panic.' Thus, it does not consider the substantial but circumscribed literature in economics that focuses solely on 'financial panics,' or another set in psychiatry that deals with 'panic states' from a mental health viewpoint. Although a few sociologists have looked at the first phenomenon under the general rubric of panic, the second has never been of any concern to them.

The focus on sociology is justified for several reasons. From the very origins of the field, sociologists have used the term as part of their professional vocabulary. The first introductory sociology textbook states that 'panic is the crowd in dissolution' (Park and Burgess 1924, p. 876) because it is the opposite of the crowd in the sense of not having any psychological unity. In particular, panic has always been considered part of the subject matter of the sociological specialization of collective behavior, which deals with nontraditional and newly emergent forms of social action (see Sociology of Collective Behavior). Most US textbooks and extended theoretical treatises on collective behavior discuss panic (Miller 1985) with some giving

extensive attention to the phenomenon. Furthermore, such empirical studies of panic as have been undertaken have been mostly by sociologists studying human and group behavior in natural and technological disasters (see *Disasters, Sociology of*). Even psychologists, who next to sociologists have been most interested in panic, primarily cite empirical and theoretical sociological sources (Schultz 1964).

1. *Past Approaches*

The first extended and systematic treatment of panic by a sociologist is a chapter by LaPiere (1938) in a little-known book on collective behavior. He sees panic as dysfunctional escape behavior generated by fortuitous, ever-varying circumstances, but involving impending danger. Panic is never formally defined except as the antithesis of regimental behavior, which is preplanned collective action for dealing with crises, such as fire drills in schools. Instead what constitutes panic is illustrated by presentations of anecdotal examples from stories of disaster behavior in journalistic and popular sources. While his vague formulations and use of popular sources are typical, LaPiere differs from many later writers in three major ways. He does not generally deal with any covert emotional or feeling state associated with the behavior, and he avoids use of the term 'irrational.' In addition, he distinguishes between collective and individual panic, which is at variance with most treatments, which imply that panic necessarily involves a multiplicity of participants.

As just implied, there is little consensus, historically or currently, on the use of the term. Nevertheless, most discussions about the nature of panic can be grouped into one of two categories. The oldest view, coming out of everyday speech, primarily equates panic with extreme and groundless fear. This is clearly related to the linguistic origins of the word which is derived from the Greek god Pan, who supposedly was able to generate sudden and overwhelming fear in the absence of any actual threat. For many writers taking this position, being seized by such a fear can lead to other irrational reactions although the essence of panic is the emotional manifestation, and not what this might otherwise overtly affect.

Another view visualizes panic as manifesting itself primarily as flight behavior. In this conception, the essence of panic is the overt behavior that is marked by the setting aside of everyday social norms, even the strongest, such as parents abandoning their young children when trying to save themselves in a life-threatening crisis. Often implicit, there is the assumption in this view that such flight behavior will occur only if there is a perception that there is a possibility of escaping the threat. Disaster researchers in particular have emphasized that hope of escape rather than hopelessness is what is involved. People who perceived themselves as totally trapped, such as in sunken submarines or collapsed coal mines, do not panic because they see no way of getting away from the threat.

Of course it is possible to put the two general formulations together. Smelser (1963, p. 131) defines panic as 'collective flight based on a hysterical belief,' which allows him to talk both about overt escape behavior and the selling behavior of investors in financial panics. But others have noted that the relationship between a terror state and rout behavior is not necessarily a direct one, given that whatever the emotion, it need not inevitably lead to flight of any kind. Apart from agreeing that panic participants are very afraid of some perceived threat, there is little agreement otherwise among students of the phenomenon on the relationship of the emotional state involved to what other, if any, behaviors will be manifested.

2. *Current Issues*

Apart from differences about the nature of panic, many discussions of panic focus on two other major themes. One such theme, and a very prominent one, assumes that panic behavior is 'irrational' (cf Le Bon [1895] 1960). This is often contrasted with the assumed rationality of most other behavior, where the means–ends relationships are in balance or where the end result is a positive one. This conception has come under sharp criticism, especially from those who have done empirical studies of the behavior of people caught in disasters and fire situations (Quarantelli 1981). These researchers note that when the behavior on such occasions is looked at from the perspective of the social actors involved, the behavior is very meaningful and far from most conceptions of irrationality (Johnson 1985). The argument is that nothing is gained by characterizing the behavior as irrational or along any similar dimension (Wenger 1980). However, it is still very common and typical for social science textbooks that discuss panic to characterize it as irrational.

A second major division among scholars is between those who argue that panic behavior is very contagious, and that human beings are easily swept up into the behavior, and those who strongly disagree with such a conception. Many of the early discussions of panic clearly assume that the participants are overwhelmed by the fearful emotion of others, and will trample over others in their path. In contrast are those researchers who take the position that whatever panic behavior is, it results from meaningful social interaction among the participants. They note that even in those rare cases of extreme instances of panic flight, very few spectators ever get caught up in the behavior.

Several specific studies have particularly influenced all of the previous arguments. One is a study by Cantril (1940) on reactions in the USA to a nationally broadcast radio show supposedly reporting as actual fact an alien invasion from Mars. The study takes the view that those who panicked upon hearing the broadcast lacked 'critical ability.' The behavior was seen as irrational. While even to this day this study is cited as scientific support for this view of panic, the research has come under sharp critical scrutiny. Analysts have noted that even taking the data reported at face value, only a small fraction (12 percent) of the radio audience ever gave even any remote credence to the idea that the broadcast was an actual news story. And the accounts of flight behavior as well as other illogical and bizarre actions reported in the book give the mistaken idea that they were obtained in the survey study done. Actually, they are taken almost unacknowledged from journalistic accounts of the time which reported that supposedly numerous people fled wildly to get away from the alien invaders. A close scrutiny of the actual news report that appeared found that most of them were sensationalized anecdotal stories that these days are typically reported in the tabloid press.

The last point is particularly important because much later Rosengren and his colleagues (1978) studied in a very systematic manner a roughly similar radio show. A Swedish radio station broadcast a fictitious news program about a nuclear plant accident that had created radioactive clouds drifting in southern Sweden. Within an hour, other media were broadcasting reports about widespread panic reactions in the population, with the print media later reporting panic reactions of various kinds on a large scale including escape flight. However, the systematic survey study found less than 10 percent of the audience gave any credence to the broadcast, with less than 1 percent showing any behavioral reaction, and no one who had engaged in flight behavior. The research showed that the supposed panic reaction of the population was almost exclusively a mass media creation. There is very little found by this research that would be supportive of any of the formulations about panic advanced to that time by social scientists.

More recently, Johnson (1988) did intensive studies of persons caught in potential panic-provoking situations such as a fire in a night club and a stampede during a rock music concert, where 160 and 11 persons respectively died. The findings are unambiguous. The great majority of involved persons did not engage in animal-like behavior, contrary to what many early writers on panic suggest occurs. Instead of ruthless competition, the social order did not break down, with cooperative rather than selfish behavior predominating. Contrary to notions of irrationality, there was much evidence for rational responses in the face of the crisis. While strong emotions were experienced, these did not lead to maladaptive behavior. These findings reinforced the ever-growing viewpoint among many researchers who have studied what seem potential panic situations, that prosocial rather than antisocial behavior predominates even in such contexts.

3. *Future Questions*

Two additional and fundamental but unresolved questions regarding the concept of panic are surfacing. Most scholars seem to agree that whatever 'panic' might mean, the phenomenon is statistically quite rare, usually involves only a handful of persons, and is of short duration. Some researchers have observed that it is very difficult to find clear-cut cases of actual panic in natural and technological disasters (and that they are also extremely rare in the other arena in which they supposedly occur, that is among soldiers in battles during wars). But the term continues to be widely used and persists, despite, the lack of empirical evidence that it happens on any significant scale; it also continues as noted by students of popular culture to be the staple of disaster movies and novels.

Wolfenstein (1957) explicitly questioned why, despite what the research evidence showed, the idea of 'panic' captures the popular imagination and continues to be evoked by scholars of human behavior. Using a psychoanalytical framework, she suggests that the fascination with the term is a psychological fantasy actually useful and functional in a variety of ways for coping with personal crises. Some disaster scholars have thought that a more sociological approach might parallel what Durkheim, a major figure in the origins of sociology, said about crime. If it did not exist, it would be necessary for human societies to create crime, at least symbolically, so as to emphasize the fact that human beings generally adhere to social norms. So perhaps the idea of the possibility of panic is necessary in society to highlight the fact that human beings in contrast react remarkably well in most stressful situations and that the social bonds between and among people usually hold. Supporting this view, as disaster researchers have noted, the mass media in reporting the *absence* of panic are behaving as if the normal expectation is that panic will occur. Whatever the merit of the parallel to crime, clearly there is a need to explain the huge discrepancy between the actual frequency of panic in any sense of the term, and the exaggerated interest in and widespread use of the word, not only in popular culture but also especially among social scientists who have not researched the phenomena.

A second basic question is whether there is still any scientific justification for the continuing use of the concept in any technical sense. Increasingly, students of the phenomenon have questioned if the term 'panic' ought to be continued to be used by students of collective behavior (Mawson 1980, Sime 1981). The logic of the argument is that such behavior as is attempted to be captured under that label can better be

dealt with by using other concepts. In other words, what is often currently described and analyzed as panic behavior can be characterized and explained by other terms. Those taking this view argue that because a word has widespread popular currency, and has been unthinkingly imported into a scientific field, this should not be seen as necessarily giving the term any legitimacy. From this perspective, it is possible that the concept of panic within collective behavior in sociology may disappear as a technical term in the future.

Bibliography

Cantril H 1940 *The Invasion From Mars: A Study in the Psychology of Panic*. Princeton University Press, Princeton, NJ

Foreman P 1953 Panic theory. *Sociology and Social Research* **37**: 295–304

Johnson N 1985 Panic and the breakdown of social order: Popular myth, social theory and empirical evidence. *Sociological Focus* **20**: 171–83

Johnson N 1988 Fire in a crowded theater: A descriptive investigation of the emergence of panic. *International Journal of Mass Emergencies and Disasters* **6**: 7–26

Keating J 1982 The myth of panic. *Fire Journal* **77**: 57–62

LaPiere R 1938 *Collective Behavior*. McGraw-Hill, New York

Le Bon G [1895] 1960 *The Crowd*. Viking, New York

Lofland J 1981 Collective behavior: The elementary forms. In: Rosenberg S, Turner R (eds.) *Social Psychology: Sociological Perspectives*. Basic Books, New York, pp. 411–46

Mawson A 1980 Is the concept of panic useful for scientific purposes? In: Leven B, Paulsen R (eds.) *Second International Seminar on Human Behavior in Fire Emergencies*. National Bureau of Standards, Washington, DC, pp. 208–13

Miller D 1985 *Introduction to Collective Behavior*. Wadsworth, Belmont, CA

Park R, Burgess E 1924 *Introduction to the Science of Sociology*. University of Chicago Press, Chicago

Quarantelli E 1954 The nature and conditions of panic. *American Journal of Sociology* **60**: 265–75

Quarantelli E 1957 The behavior of panic participants. *Sociology and Social Research* **41**: 187–94

Quarantelli E 1981 Panic behavior in fire situations: Findings and a model from the English language research literature. In: *Proceedings of the 4th Joint Panel Meeting*. The UJNRR Panel on Fire Research and Safety Building Research Institute, Tokyo

Rosengren K, Arvidsson P, Sturesson D 1978 The Barseback 'panic': A case of media deviance. In: Winick C (ed.) *Deviance and Mass Media*. Sage, Beverly Hills, CA, pp. 131–49

Schultz D (ed.) 1964 *Panic Behavior: Discussion and Readings*. Random House, New York

Sime J 1981 The concept of panic. In: Canter D (ed.) *Fires and Human Behavior*. Wiley, New York, pp. 63–81

Smelser N 1963 *Theory of Collective Behavior*. Free Press, New York

Strauss A 1944 The literature on panic. *Journal of Abnormal and Social Psychology* **29**: 317–28

Wenger D 1980 Some observations on the nature of panic behavior: A normative orientation. In: Leven B, Paulsen R (eds.) *Second International Seminar on Human Behavior in Fire Emergencies*. National Bureau of Standards, Washington, DC, pp. 214–19

Wolfenstein M 1957 *Disaster: A Psychological Essay*. Free Press, Glencoe, IL

E. L. Quarantelli

Paradigms in the Social Sciences

How is one to explain the diffusion of the concept of 'paradigm,' a word of Greek origin used by some philosophers during the eighteenth century, and which, in the writings of some contemporary scholars, has a pedantic tone? One of the reasons may be the fact that it raises what is, in the history of science, a fundamental question: is scientific progress achieved mostly by steady accretion or mostly by abrupt jumps? This concept was devised by Thomas Kuhn in a particular context, and the circumstances of its formulation poses directly the question to be treated here: Are there paradigms in the social sciences?

1. A Polysemic Concept

Thomas Kuhn explains in the preface to *The Structure of Scientific Revolutions* (Kuhn 1970) that it was during a stay at Palo Alto Center for Advanced Studies, in the company of social scientists, that he was led to formulate the concept of paradigm with the aim of making clear the essential difference between natural sciences and the social sciences. The reason given by Kuhn was the absence of a theoretical consensus in any discipline of social sciences.

> I was struck by the number and extent of overt disagreement between social scientists about the nature of scientific problems and methods…Controversies over fundamentals seem endemic among, say psychologists or sociologists… Attempting to discover the source of that difference led me to recognize the role in scientific research of what I have since called paradigms (Kuhn 1970, p. VIII).

For him, the use of the word paradigm in the social sciences is not justified.

As a preliminary it should be noted that many scholars have formulated doubts about the reality of paradigms even in the natural sciences—an issue that cannot be discussed here. Kuhn himself has in the postscript to the second edition of his book reformulated the concept of paradigm, by disentangling this concept from the notion of scientific community, and by admitting that he has used the term in two different senses. On one hand, paradigm stands for a constellation of beliefs, values, techniques and so on shared by the members of a given community; on the other, it denotes a concrete puzzle-solution. He labeled the first meaning 'disciplinary matrix': 'disciplinary, because it refers to the common possession of

the practitioners of a particular discipline; matrix, because it is composed of ordered elements of various sorts' (Kuhn 1970, p. 182). Kuhn makes a distinction between paradigms in the large sense, the disciplinary matrix, and small paradigms called 'exemplars.' This distinction is considered unclear by several sociologists of science.

In the revised version of Kuhn's thesis

> any period of scientific development is marked by a large number of overlapping and interpenetrating paradigms, which do not replace each other immediately and do not spring up full-borne, but instead emerge in a long process of intellectual competition…These concessions dilute Kuhn's original message, and in its final version the argument is difficult to distinguish from the average historian's account of the history of science (Blaug 1976, p. 154).

Among the sociologists of science, very few have asked if the concept of paradigm is, or not, applicable to social sciences. Harriet Zuckerman in her substantial *Sociology of Science* (Zuckerman 1988) does not dedicate a single line to this issue, nor do Merton, Price, Collins, Lakatos, or Mullins.

Are there in the social sciences instances of paradigmatic upheavals comparable to those created by Copernicus, Newton, Darwin, or Einstein? Can the theories of Keynes, Freud, Chomsky, or Parsons be described as paradigmatic? In the social sciences, does progress occur through paradigmatic revolutions or through cumulative processes?

For there to be a paradigm, one condition must be met: theories must refer to essential aspects of social reality. However, the more ambitious a theory is, the less it can be directly tested by the data available. In the social sciences there are no 'fundamental discoveries,' as there sometimes are in the natural sciences. Instead, unverifiable theories are constructed, partly because social reality itself changes. Kuhn held that a paradigm shift occurs when two 'incommensurable' paradigms are in competition. Then, each school rejects the premise of the others, and contests their work. Communication breakdown is the sign of an incompatibility. In the social sciences such 'incommensurability' takes the pattern of mutual ignorance.

2. *Mutual Ignorance*

There are two kinds of mutual ignorance: involuntary and deliberate. The formal disagreement may be based on personal rivalries, on contest about priority, on methodological quarrelling or on ideological hostility.

Within a formal discipline, several major theories may cohabit, but there is a paradigm only when one testable theory alone dominates all other theories and is accepted by the entire scientific community. When Pasteur discovered the microbe, the theory of spontaneous generation collapsed: contagion became the new paradigm. In the social sciences, however, we see at best a confrontation between several nontestable theories. Most of the time there is not even a confrontation but careful mutual avoidance, superb disregard on all sides. This is relatively easy, owing to the size of scientific disciplines, and their division into schools; it is true for all countries, big or small.

This mutual disregard is an old practice in the social sciences. At the turn of the twentieth century, the great scholars did not communicate, or very little. In the writings of Weber there is no reference to his contemporary Durkheim. Yet Weber was acquainted with Durkheim's jounal, *l'Année Sociologique*. For his part, Durkheim, who could read German, makes only one, fleeting reference to Weber. Yet they worked on a number of the same subjects, such as religion.

As Giddens puts it:

> Weber certainly knew of the work of Durkheim but there can be little doubt that he was barely influenced at all…The same applies in a reciprocal vein; Durkheim knew of Weber's work, and there is a minor reference to aspects of it in one passage in Durkheim's writings, but he claimed no particular kinship with those of Weber's ideas with which he was familiar (Giddens 1987, p.182).

Durkheim does no more than mention in passing Simmel and Tonnies.

Harshly criticized by Pareto, Durkheim never alluded to Pareto's work. Pareto's judgment of Durkheim's book on suicide was unfavorable. 'Unfortunately,' he wrote, 'its arguments lack rigour.' Weber seems to have been unaware of Pareto's theory on the circulation of elites, and Pareto in his turn says nothing about the Weberian theory of political leadership. There was no exchange between Weber and Freud. Ernst Bloch and Georg Lukacs met regularly with Weber in Heidelberg, but their work shows no sign of his influence. The only one of Weber's contemporaries who referred to him was Karl Jaspers, but he was a philosopher (compare Mommsen and Osterhammel 1987).

Weber and Spengler had not exerted the slightest influence on the ideas of the other. Croce said about Weber: 'I had met Weber in 1909. I had no further contact with him, nor did I follow his scholarly work, from which I had only read his early book *Roman Agrarian History*.' As was noted by Raymond Aron, each of the three great founders of sociology followed a 'solitary path.' There is no room for two theorists like Parsons and Sorokin in the same habitat; the clash between them became inevitable. The Parsonian grand theory had a short life. Robert Dahl and Arnold Rose refute the theory of C. Wright Mills; they do not adopt the stratagem of deliberate ignorance—they contest it openly. On the other hand, C. Wright Mills chooses to

ignore the books published by Berle and Means and by James Burnham, in despite of his own theory.

Mosca had accused Pareto of plagiary. The imprecision of the language of both elitist theorists, who adopted synonyms with confusing meanings, has generated in Italy in enormous literature with the participation of dozens of Italian scholars debating the priority of the authors of the 'political class' and of 'political elites' (Albertoni 1983). Pareto and Mosca are not mentioned by José Ortega Y Gasset (1883–1955). Schumpeter does not refer to Mosca, Pareto, Michels, or Burnham. Angus Campbell and Paul Lazarfeld have spent a large part of their parallel life in the study of the same political behavior; they never cite each other. Johan Galtung and Stein Rokkan born in 1930 and 1921 respectively, both citizens of the same small country (Norway), have each contributed a theory of 'center versus periphery.' They have never collaborated, avoiding each other carefully; they coexisted by ignoring each other.

Three lists of major contributions to economics were compiled in 1982 at a symposium in Berlin. The first, prepared by W. Krelle, contained 30 names, the second, by J. Timbergen, 36 names, and the third, by B. Fritsch and G. Kirchgässner, 44 names. In the first two lists, there were only two names that were the same (including Klein, Nobel prize winner, and Krelle himself?); in the first and the third, there were only nine that were the same. The second and third lists did not have a single name in common (Platt 1986, p. 350). Such widely differing views about leading figures do not say much for the coherence of economics. This is also the opinion of the economist Kenneth Boulding, who speaks of economics as being 'disorientated,' comparing it with the story of the blind man and the elephant. The Nobel prize winner Wasily Leontieff was not more indulgent: 'Year after year, economic theorists produce mathematical models without being able to get any further towards understanding the structure and functioning of the real economic system' (quoted in Platt 1986, p. 350). The monetarist George Stigler objected to six econometricians (Tinbergen, Samuelson, Malinvaud, Moore, Frisch, and Fisher) being described as the authors of 'exceptional contributions' because, he wrote, 'econometrics has no unified core or methology' and 'has not yet had a major impact on economics' (Platt 1986, p. 342). But three of these six economists have since then won the Nobel prize, as has Stigler himself. 'The existence and unity of a discipline called economics reside in the eye and mind of the beholder. Samuelson, Solow, and Tobin and others reserved their harshest words for neoclassical economists like Friedmann and members of the Chicago School, who ridicule Keynes's consumption function' (Amariglio et al. 1993, pp. 150 and 175). The same remarks can be made with reference to other disciplines, except linguistics.

If scholars belonging to the same generation and working on the same subjects ignore each other, whether involuntarily or deliberately, how could one perceive the existence of a scientific community able to adhere to a common paradigmatic Weltanschauung? Paradoxically, the closer the proximity, the more vivid is the rivalry and the more frequent the mutual ignorance. To the isolation of scholars motivated by competition should be added the weak communication between specialties within the same discipline (see *Specialization and Recombination of Specialties in the Social Sciences*).

The absence of consensus among contemporaries is compensated by the transmission of knowledge from one generation to the next, as revealed by citation patterns, particularly in handbooks and compendia. Such a generational transmission testifies that in social sciences scientific advancement occurs mostly by cumulative knowledge.

3. *Cumulative Knowledge*

Science consists largely in the adding of successive layers of sediments, which form a patrimony for each discipline and field. In all sciences 'the extent of the innovation that any individual can produce is necessarily limited, for each individual must employ in his research the tools that he acquires from a traditional education, and he cannot in his own lifetime replace them all' (Kuhn 1970, p. 183). Progress does not arise in a vacuum, but develops out of the scientific patrimony. It is difficult to find in the social sciences a virgin domain. Every decade that passes adds layer upon layer to the patrimony, the only recent exception being the 1930s in Europe. New explanations supersede older interpretations. Many examples of cumulative progress can be given. Even giants rely on the patrimony. Karl Marx refers to Adam Smith on 296 of the 1,721 pages of his *Theory of Surplus Value*. In this book he draws on the work of his predecessor in more than one page out of every six. Ralf Dahrendorf cites Marx 160 times (Dahrendorf 1957). Max Weber does not cite Karl Marx, but many of this writings were a reply to the thesis of his predecessor. He once noted that he could not have accomplished crucial parts of his work without the contributions by Marx and Nietzsche.

Darwin had recognized his debt to Malthus; Keynes would not have been able to write one of the most famous books of the twentieth century without the incremental advance achieved by several generations of economists. The 'theory of development' consists of a long chain of accumulated contributions in several disciplines.

The literature on elites is a good example, among others, of cumulative knowledge even in the case of mutual ignorance. Many contributions emphasize the sedimentation of layers of knowledge. The old Mosca–Pareto layer, impregnated with plagiarism, has become consolidated groundwork. Two dozen

important books published since 1950 have built an impressive patrimony. Obsolete theories have served as fertilizers for new complementary theories. One example of cumulative knowledge is the new influence of electoral techniques on party systems. A bibliography, even a very selective one, on this theme could easily list 200 or 300 titles in English, not to mention the many varied observations derived from the direct experience of politicians in numerous countries. From Condorcet, Bachofen, John Stuart Mill, Hare, and Hondt to Hermens, Downs, Duverger, Sartori, Lijphart, the theory is based on the contributions and successive improvements made by a very large number of scholars. (For example, the consequences of proportional representation were described by Bachofen as long ago as 1850).

For each of the dozens of important domains in the social sciences there is a substantial patrimony, the accumulated knowledge of a long chain of scholars. Imagine the following imposture-exercise. Take from the literature a classic article or chapter, for instance Max Weber's seminal chapter on 'Domination and Legitimacy,' falsify the name of the author, make a few cosmetic changes, and offer it to a major journal. Would such a manuscript be accepted today? Would not a paper that ignores the literature of the last decades preclude publication? The patrimony is common property. Although every concept or theory or method has its creators and developers, they do not need to be cited every time the term is used. Even the identity of the originator may disappear into anonymity. We do not, and cannot, remember who used for the first time terms such as 'role,' 'revolution' or 'social mobility.' Given such a patrimony, scholars today can start their research at a much higher level than did their predecessors (Dogan and Pahre 1990, p. 21). A graduate student today knows more than the founders of his or her field—even if they do not have the reasoning capacity of their forebears. That knowledge is largely acquired by cumulation is an obvious fact for all sciences; if it were otherwise, simultaneous discoveries in the natural sciences would not occur, theoretical debates in the social sciences would not be engaged. They are generated when the issue is 'in the air,' when 'time is ripe.' The dispute between Mosca and Pareto about the paternity of ideas in elite studies is similar to the dispute between Newton and Leibnitz about priorities on calculus. In elite studies, for variety of reasons 'the time was ripe' at the end of the nineteenth century. 'Discoveries are like apples on a tree,' wrote Price (Kuhn 1986, p. 60). In the history of social sciences some books have played a crucial role but they are only summits in a chain of mountains, not paradigmatic contributions in the sense given to this word by Thomas Kuhn.

Overarching theories, that is, paradigmatic frameworks, can be built on more solid ground in the natural sciences than in the social sciences, because in the former truth is universal, in the later, contextual.

4. *Contextual Diversity vs. Universal Truth*

Chemical substances keep indefinitely the same composition and are identical in all latitudes. Not so social phenomena! In contrast with the natural sciences, contextual diversity and social change are two important parameters in all social sciences. Both parameters resist ambitious paradigmatic generalizations. Extrapolating Kuhn's perception of the history of astronomy, physics, or chemistry to social sciences is like looking into a distorting mirror. Many concepts and theories do not have equal pertinence when applied to Western Europe and to tropical Africa. Dependency theory may be a pertinent theory for yesterday's Latin America and for Eastern Europe today, but not for other areas in the world. The comparative method reduces the significance of paradigmatic frameworks. Dozens of theories with universal pretentions neglecting national configurations have been invalidated.

In the natural sciences an experience can be repeated many times in identical conditions anywhere in the world (except when magnetism has an impact). On the contrary, social realities change. For instance, at the time of Malthus, overpopulation in Europe was a realistic hypothesis. Today, birth rates in this continent are not sufficient to replace the outgoing generation. Traditional values concerning religion, nation, family, erotism, or authority have changed at an accelerated rhythm in the last decades of the twentieth century. Theories reflecting the social reality one generation ago have to be replaced in order to explain contemporary trends. History never repeats itself because factors do not recombine in identical ways at all epochs. The scope of social sciences is rarely the discovery of laws with universal validity, but rather the explanation of social diversity. 'What is true in one side of the Pyrénées is not in the other side,' wrote Montaigne four centuries ago. In the social sciences truth is not universal, it is contextual. Social scientists do not make inventions, and rarely make discoveries; what they do best is to observe regularities and exceptions (anomalies). It is for this reason, that the most imaginative strategy in the social sciences is the comparative method, the best surrogate for the experimental method. Social phenomena being contextual, paradigmatic statements in political science, sociology, anthropology, social psychology or social history often appear as ethnocentric and misleading.

Most social scientists who adopt the concept of paradigm apply it to particular domains, not to their entire discipline. For some scholars, elitism is the most powerful paradigm against the Marxist paradigm. In the domain of political economy, rational choice is elevated to the rank of paradigm. Many psychiatrists are mobilized around the Freudian paradigm.

Competing theories contribute to the advancement of knowledge. To explain the fall of the Roman empire, 14 competing theories have been formulated

by great scholars, from Montesquieu to Max Weber, to which can be added a theory proposed by chemistry (saturnism). None is fully comprehensive, but each one enlightens part of the phenomenon. The history of social sciences is not a history of paradigmatic upheavals, but of competing theories, with many being invalidated, but many others constituting the foundations of the contemporary social sciences. Without competing theories, social sciences would not advance. The clash of theories leaves no room for paradigms.

A symptomatic analogy can be drawn between what Kuhn calls in the revised version of his theory, 'scientific community' (adepts of a paradigm) and what some social scientists (Almond, Merton, Smelser, Zuckerman) call 'school.' In the recent history of social sciences there are dozens of schools: monetarist, rational-choice, bio-social, individualist, postmodernist, and so on, and that means that attempts to formulate overarching disciplinary paradigms may appear as 'imperialist views.'

As Merton put it:

> Since all virtues can readily become vices merely by being carried to excess, the sociological paradigm can be abused almost as easily as it can be used. It is a temptation to mental indolence. Equipped with his paradigm, the sociologist may shut his eyes to strategic data not expressly called for in the paradigm. He may turn the paradigm from a sociological field-glass into a sociological blinker. (Merton 1957, p. 16).

In some cases a paradigm may assume the features of a dogmatic orientation. No wonder if not so many scholars adopt it: 'The notion of paradigms and paradigmatic revolutions often seem to be taken up only in order to be rejected. The beneficiary of this exercise in several cases is Lakatos, mostly because he appears as a type of moderate Kuhn' (Weingart 1986, p. 267). In the social sciences, theoretical disagreements are beneficial to the advance of knowledge. Nevertheless, the word paradigm has taken root, particularly in sociology, political science, psychology, and normative philosophy. Yet most philosophers of science reject it; most historians are reluctant to make such generalizations; most economists continue to think in terms of assumptions. It may be too late now to try to exclude this word from the lexicon, in spite of the fact that many other expressions are available (conceptual framework, assumption, dominant theory, theoretical breakthrough, grand theory, general model, axiom, and so on). It has become necessary to specify it, or to limit its use to particular domains, such as cognitive science, international relations, or hybrid demography.

See also: History of Science; History of Science: Constructivist Perspectives; Problem Selection in the Social Sciences: Methodology; Quantification in the History of the Social Sciences; Science, Sociology of; Theory: Conceptions in the Social Sciences

Bibliography

Albertoni E (ed.) 1983 *Governo e governabilità nel sistema politico e giuridico di Gaetano Mosca*. Giuffre, Milan, Italy

Amariglio J, Resnick S, Wolff R 1993 Division and difference in the 'Discipline of Economics.' In: Messer-Davidow E, Schunway D, Sylvain D (eds.) *Knowledges: Historical and Critical Studies in Disciplinarity*. University Press of Virginia, Charlottesville, VA, pp. 150–4

Blaug M 1976 Kuhn versus Lakatos or paradigms versus research programmes in the history of economics. In: Latsis S J (ed.) *Method and Appraisal in Economies*. Cambridge University Press, Cambridge, MA, pp. 149–61

Dahrendorf R 1957 *Class and Class Conflict in Industrial Society*. Stanford University Press, Stanford, CA

Dogan M, Pahre R 1990 *Creative Marginality: Innovation at the Intersection of Social Sciences*. Westview Press, Boulder, CO

Giddens A 1987 Weber and Durkheim: Coincidence and divergence. In: Mommsen W J, Osterhammel J (eds.) *Max Weber and Contemporaries*. Allen and Unwin, London, pp. 182–9

Kuhn T S 1970 [1962] *The Structure of Scientific Revolution*. University of Chicago Press, Chicago, IL

Merton R K 1957 *Social Theory and Social Structure*. The Free Press, New York

Mommsen W J, Osterhammel J 1987 *Max Weber and His Contemporaries*. Allen and Unwin, London

Platt J 1986 Is everything equal? In: Deutsch K W, Markovits A S, Platt I (eds.) *Advances in the Social Sciences*. University Press of America, New York

Price D de S 1986 [1963] *Little Science, Big Science and Beyond*. Columbia University Press, New York

Weingart P 1986 T.S. Kuhn: Revolutionary or agent provocateur? In: Deutsch K W, Markovits A S, Platt J (eds.) *Advances in the Social Sciences*. University Press of America, AS, pp. 265–75

Zuckerman H 1988 The sociology of science. In: Smelser N (ed.) *Handbook of Sociology*. Newbury Park, CA, pp. 511–99

M. Dogan

Para-lawyers: Other Legal Occupations

1. Definition

Para-lawyers are persons performing 'law work' who themselves are not lawyers. Law work in turn refers to those activities requiring at least some of the specialized legal knowledge and skill of the type associated with the work of lawyers. Thus, para-lawyers are 'law workers' who lack the full credentials and/or status associated with the legal profession but who provide services that draw on expertise in law and/or legal processes. This negative definition reflects the complexity of defining the practice of law and what constitutes membership in a legal *profession*.

Drawing the line between lawyers and para-lawyers varies from country to country because of variations in the nature of legal tasks within different legal

systems, and how those tasks are assigned to particular occupations. In many countries, the exact line is not clear. In some countries, the term most equivalent to 'lawyer' is customarily applied to an occupation engaged in a very narrow range of activities, while there are other occupations requiring substantial formal legal training that perform work which elsewhere would be the exclusive domain of the legal profession. Deciding whether to include such occupations under the label lawyers or para-lawyers requires drawing an arbitrary line. One example is the legal occupation of 'notary' as it exists in many civil law countries (Malavet 1996). Historically, the occupation of notary arose to provide a neutral person who could be trusted to draw up fair contracts to be executed by one or more persons unable to read or write; today, notaries continue to play a role in the drawing up of and execution of contracts, and apply specialized legal knowledge related to those activities. If the notary's tasks involve anything more than ministerial functions such as witnessing the execution of contracts, it is arbitrary whether or not the notary is part of the legal profession or should be categorized as a para-lawyer.

In addition to problems of drawing a line between lawyers and para-lawyers, it is difficult to draw a line between para-lawyers and 'nonlaw' occupations. For example, many nonlaw occupations require the use of some very specific legal knowledge, typically applied in a very routine way. Examples include drawing up contracts for routine property transactions (in the United States typically done by real estate sales agents), drawing up routine loan contracts (done by bank officials), settling tort-based injury claims (done by nonlawyer representatives of insurance companies), or advising on estate or tax issues (done by financial advisors and accountants). In one sense, all of these occupations could be labeled para-lawyers; for some of the occupations the para-lawyer work is only a part of their activities (e.g., real estate sales agents) while for others (e.g., insurance claims representatives) it is the vast bulk of their work.

Finally, there is no recognized standard of education or specialized training to become a para-lawyer. While there are training courses for occupations variously labeled 'paralegal,' 'legal assistant,' or 'legal executive,' these courses do not provide formal qualifications, and most people enter such occupations without any formalized training (Johnstone and Wenglinsky 1985, pp. 121–3). Other kinds of para-lawyer occupations involve the application of knowledge learned in some other role; examples include former police officers representing persons charged with driving while intoxicated in Ontario, former Internal Revenue Service (IRS) auditors representing or assisting taxpayers in disputes with the IRS in the United States, former insurance company claims inspectors or adjusters representing injury claimants in England, and legal secretaries who take on increasing responsibilities to the point where they are engaged more in law-work than in secretarial work. Thus, there is no clear way to decide whether a person offering a particular set of services, or engaged in a particular set of tasks, should be classified as a para-lawyer; nor is the term para-lawyer in common use in a way that would allow for self-identification.

2. *Empirical and Theoretical Issues*

Even with these definitional and boundary problems, the concept of para-lawyers, and para-professionals more generally, is useful theoretically and important for practical purposes. It raises the issues of how law-work is and should be structured, and what qualifications are needed to be effective in law work. In addition, there are two other lines of theoretical inquiry that are important vis-a-vis para-lawyers: 'deprofessionalization' and the gendering of occupations.

2.1 *Structure of Law Work*

Law work readily divides along at least three dimensions: institutional setting, legal advice vs. legal assistance, and routine vs. nonroutine tasks. The degree to which para-lawyers have in the past, and are today, providing legal services varies along these dimensions. As one moves among institutional settings—from office, through tribunal, to lower courts, and finally to the higher courts—the role of para-lawyers tends to decrease. Para-lawyers are most often excluded from law work in upper level trial and appeal courts, and least often from tasks that can be completed in the privacy of an office. For example, in England, lawyers enjoy a monopoly on representation in the higher courts; in contrast, office-based work generally is not limited to lawyers, nor is representation before administrative tribunals. Recent developments are bringing nonlawyers into some courts in England, particularly to handle preliminary matters in criminal cases.

Continuing with the example of England, para-lawyers direct a larger proportion of their effort toward providing advice on legal matters, broadly-defined, while lawyers are more likely to engage in active assistance. In part this is because accounting firms and advice bureaus staffed by nonlawyers are major providers of legal advice is in England. In other countries, it is likely to be the case that para-lawyers are more involved in advice-giving simply because it is less visible, and hence difficult for regulators to limit or otherwise control.

The more that routines can be established to handle a legal task, the more possible it is for someone with very specific training and/or expertise to handle the task, even if it is highly technical. For example, in the

US, appeals of certain types of government benefit claims involve highly technical issues of both regulations and medical assessment. Such cases tend to be handled by specialists, some of whom are lawyers and some of whom are para-lawyers. On the less technical side, there are many legal processes that are routine but require the completion of various kinds of forms; para-lawyers can both identify the forms needed for various processes and assist consumers in the completion of those forms.

One final issue concerning the structure of legal work and para-lawyers is whether they work as independent service providers or under the supervision of lawyers. Many lawyers employ para-lawyers to complete routine and/or specialized tasks within the lawyers' practices (Johnstone and Wenglinsky 1985); many legal departments inside organizations employ a mix of lawyers and para-lawyers all under the supervision of senior lawyers. Typically, para-lawyers working under the supervision of lawyers is an accepted delegation of responsibility, although there is substantial variation in drawing the line between what a para-lawyer may do and what requires a lawyer. One frequent boundary is the courtoom door: representation in court tends to be limited to lawyers, although that may be less so in the lower courts (the example of drunk driving cases in Ontario noted previously), and is changing under both political and economic pressures (the example of para-lawyers handling certain criminal matters in England, and the role of specialized advocates in certain types of cases such as domestic violence).

2.2 *Effectiveness in Delivery of Legal Services*

A key question with regard to the work of para-lawyers is how it compares in quality to the same activities carried out by lawyers. While lawyers claim that use of para-lawyers raises threats of poor quality work, efforts to compare systematically the work of lawyers and para-lawyers fail to support this contention (Kritzer 1998). The extant research demonstrates that para-lawyers can be as effective as lawyers in providing specific legal services. Achieving effectiveness involves a combination of specific knowledge or expertise vis-a-vis substantive law, institutional process, and other actors in the system. Para-lawyers who obtain this combination can achieve levels of effectiveness comparable to that of lawyers, and specialized para-lawyers will tend to be more effective than lawyers lacking one or more of the three core elements of knowledge.

The relationship between the service provider and the client can also influence the relative effectiveness of the provider. Specialist para-lawyers who have ongoing relationships with clients may be more effective than even specialist lawyers brought in on an *ad hoc* basis to handle a specific matter; this reflects the knowledge base created by the ongoing relationship. The payment arrangement between service provider and client may also make a difference. One study that found specialist lawyers to be more effective than specialist para-lawyers in a particular area attributed the difference in significant part to the lawyers being paid on a no-win, no-fee basis while the para-lawyers' compensation was unaffected by the outcome of specific cases.

2.3 *Deprofessionalization and Postprofessionalism*

'Deprofessionalization' refers to the shift from professional to nonprofessional status. Professional status is conferred on those occupations which maintain a monopoly over a theoretical base of knowledge, have expectation of autonomy in regulation and day to day work, and have among their clienteles a belief in the professional having a service ethos. Deprofessionalization refers to the loss of these supposedly unique occupational qualities (Haug 1973, Rothman 1984). Nonprofessionals such as para-lawyers increasingly do work once the province of professionals.

Deprofessionalization reflects several specific developments. First is the combination of the rationalization of knowledge, the rationalization of tasks employing knowledge, and the growth of tools to assist in applying knowledge. The more that tasks can be compartmentalized and the more that knowledge-based tools can be applied, the easier it is to assign the task to a person with limited, specific expertise. The result is the deskilling of knowledge-based occupations. A second development is the loss of governmental support for restrictions limiting the role of para-lawyers. In England solicitors held a monopoly on handling land transfers for many years, but these tasks are now open to licensed non-lawyer specialists; in the US, limitations have been under attack politically at both the national and state level (Commission on Nonlawyer Practice 1995). The third development is the increasing bureaucratization of legal service delivery; once organizational structures become established, there is a drive to rationalize, compartmentalize, and reduce costs. Bureaucratic structures also provide means for supervision and control, which both creates tensions for professionals expecting autonomy and provides a setting to monitor the work of para-lawyers.

A traditional claim by lawyers seeking to secure and maintain restriction on the delivery of legal services is that opening such work to para-lawyers and other nonlawyers jeopardizes the interest of clients both because of lack of competence and the absence of the stringent ethical norms of the professional. As noted in the previous section, systematic research in the United States and elsewhere increasingly challenges this claim. The loss of this argument combined with drives to contain costs through broadening options in

service-oriented markets pushes toward less and less control of 'professional' services.

The concept of deprofessionalization suggests the loss of something. An alternative concept, 'postprofessionalism' (Kritzer 1999) reflects most of the same elements but sees the opening of legal and other professional services to para-professionals such as para-lawyers as part of a natural development. Postprofessionalism builds on the growth of technological tools to access knowledge previously the exclusive realm of persons with professional-level training (Susskind 1996). These developments parallel the industrial revolution when the rationalization of tasks and the development of machines led to the shift from crafts-based production to factory-based production. Postprofessionalism captures the image of a similar shift for people who work with their heads rather than their hands: rationalization of knowledge combined with information technology tools make it possible to produce services using workers with lower levels of training and knowledge.

2.4 Gendering of Occupations

A final issue in the analysis of para-lawyers relates to the gendered nature of occupations. Historically, professions have been male-dominated while many para-professions have been female dominated. Within law-work, women were largely excluded from the professional ranks until about 1970. In the early twentieth century some of the American states formally excluded women from the bar; even where they were not excluded, women were often confined to roles such as legal secretary even if they had the formal credentials of a law degree. Since 1970, most countries have witnessed the opening of the legal profession to women, and the number of women who practice as lawyers has increased sharply.

Nonetheless, para-lawyer roles tend to be dominated by women; this is particularly true of para-lawyer positions that are specifically subordinate to lawyers (Johnstone and Wenglinsky 1985, p. 69). Even for those para-lawyer positions that function independently of lawyers, the gender composition may limit the likelihood of achieving some sort of full professional status as long as they are seen as in competition with the existing legal profession (Hearn 1982). At the same time, efforts to increase occupational status may lead to certain kinds of 'closure' strategies which bear a resemblance to those employed by lawyers: formal licensing, testing requirements, educational qualifications, etc. (Witz 1992, Johnstone and Wenglinsky 1985, pp. 165–6). Whether such efforts ultimately improve status or lock in gendered distinctions is an important theoretical and empirical issue.

A final gender-related issue concerns the professional roles assumed by many of the increasing number of women in the legal profession. A common observation in the literature on the struggles of women to achieve status and economic parity with men in the legal profession is that women lawyers are disproportionately to be found in law-work positions that are themselves in some sense gendered (e.g., family law practice, as employees rather than as principals, etc.). In some ways, the professional roles occupied by many women lawyers are closer to those of para-lawyers than to the traditional image of the work of male lawyers (Sommerlad and Sanderson 1998). This development may lead to a blurring of lines between lawyers and para-lawyers while preserving patriarchical status structures in the workplace.

3. Future Research and Theory

As theoretical and empirical concepts para-lawyers and para-professionals more generally, are underdeveloped. Current dynamics in the changing occupational structures, such as postprofessionalism, suggest that the para-lawyers should become increasingly important in the study of law work and in the delivery of legal services. The relationship of para-lawyers to the increasing role of information technology is one avenue of potential theory and research; a second is the economics of service delivery; and a third potential area is in the relationship of para-lawyer occupations to broader patterns of changing social structures, particularly the changing impact of gender on social structure.

See also: Counsel, In-house; Lawyers; Lawyers, Regulation of; Legal Professionalism; Professionalization/Professions in History; Professions, Sociology of

Bibliography

Commission on Nonlawyer Practice 1995 Nonlawyer Activity in Law-Related Situations: A Report with Recommendations. American Bar Association, Chicago

Haug M R 1973 Deprofessionalization: An alternative hypothesis for the future. *Sociological Review Monograph* **20**: 195–217

Hearn J 1982 Notes on patriarchy, professionalization and the semi-professions. *Sociology* **16**: 184–98

Johnstone Q, Wenglinsky M 1985 *Paralegals: Progress and Prospects of a Satellite Occupation*. Greenwood Press, Westport, CT

Kritzer H M 1998 *Legal Advocacy: Lawyers and Nonlawyers at Work*. University of Michigan Press, Ann Arbor, MI

Kritzer H M 1999 The professions are dead, long live the professions: Legal practice in a post-professional world. *Law & Society Review* **33**: 713–59

Malavet P A 1996 Counsel for the situation: The Latin American notary, a historical and comparative model. *Hastings International and Comparative Law Review* **19**: 389–488

Rothman R A 1984 Deprofessionalization: The Case of Law In America. *Work and Occupations* **11**: 183–206

Sommerlad H, Sanderson P 1998 *Gender, Choice and Commitment: Women Solicitors in England and Wales and the Struggle for Equal Status*. Ashgate Publishing, Aldershot, UK

Susskind R 1996 *The Future of Law: Facing the Challenges of Information Technology*. Clarendon Press, Oxford, UK
Witz A 1992 *Professions and Patriarchy*. Routledge, London

H. M. Kritzer

Parapsychology

According to *The Journal of Parapsychology* the term 'parapsychology' designates 'The branch of science that deals with *psi* communication, i.e., behaviorial or personal exchanges with the environment which are extrasensorimotor—not dependent on the senses and muscles.' The same source defines 'psi' as 'A general term to identify a person's extrasensorimotor communication with the environment.' Psi includes ESP and PK, 'ESP' (extrasensory perception) is defined as 'Experience of, or response to, a target object, state, event, or influence without sensory contact.' ESP includes telepathy (mind to mind communication without normal channels of communication), clairvoyance (extrasensory contact with the material world), and precognition (the knowledge of future events that cannot be inferred with present knowledge). The other component of psi, PK (psychokinesis) is defined as 'The extramotor aspect of psi: a direct (i.e., mental but nonmuscular) influence exerted by the subject on an external physical process, condition, or object.'

J. B. Rhine adapted the term parapsychology from the German word *Parapsychologie* in the 1930s to replace the earlier term *psychical research*. More than a change in terminology was involved. Psychical research covered a broad range of systematic investigations into spiritualistic phenomena, haunted houses, premonitory dreams, visions, and the like. Rhine's program avoided much of this material, especially that which dealt with questions of survival after death. Instead, he emphasized controlled laboratory experiments using normal individuals as subjects. Thus, the change in terminology signaled that parapsychology aspired to become an accepted scientific discipline.

1. Historical Background

1.1 Scientists and Psychic Phenomena

In 1848, the two young sisters Margaret and Katherine Fox initiated modern spiritualism with their demonstrations of rappings and other material phenomena which they claimed were caused by the spirits of dead people. Soon many other individuals, calling themselves mediums, were regularly producing phenomena which they attributed to spirits of the dead. Because science deals with material phenomena, and because material phenomena were what the spiritualists were offering as support for their paranormal claims, some major scientists became interested in investigating these claims. In 1853, Robert Hare, an important chemist from the University of Pennsylvania, began an investigation of table-tilting. He attended seances where he became convinced that the spirits of the dead were moving the tables.

In 1865, Alfred Russel Wallace, the cofounder with Darwin of the theory of evolution through natural selection, began his investigations of spiritualistic phenomena. He quickly became convinced of the reality of the paranormal and for the remaining 48 years of his life he continued his investigations and outspokenly defended the reality of spiritualistic phenomena. Sir William Crookes, the discoverer of thallium and the inventor of the cathode ray tube, began his investigations of the paranormal in 1869. He concluded, to the dismay of his scientific colleagues, that a psychic force was operating in these seances.

1.2 The Society for Psychical Research

Inspired by the investigations of such eminent scientists as Hare, Wallace, and Crookes, a group of scholars and spiritualists formed The Society for Psychical Research in London in 1882. The philosopher Henry Sidgwick, its first president, stated that the goal was to collect more evidence of the kind that had been reported by those scientists who had investigated mediums. Although Sidgwick believed that the previous investigations of mediums had already provided evidence sufficient to prove the existence of the paranormal, he was aware that the majority of the scientific community were not convinced. Therefore, the aim of the Society was to accumulate enough additional evidence of the same sort so as to force the scientific community to admit that the case for the paranormal had been proved or to accuse the proponents of insanity or gross incompetence.

In his inaugural presidential address, Sidgwick claimed that, in fact, some members of the new society had already collected just the sort of evidence that proved beyond a doubt the existence of telepathy. Sidgwick was referring to the investigations involving the Creery sisters. Unfortunately, the Creery sisters were later discovered to be signaling to each other using a code. Although an attempt was made to claim that such a code could not have accounted for some of the results of their investigations, this body of evidence was subsequently removed from the republished proceedings of the Society for Psychical Research. A similar fate befell the very next set of investigations which the Society originally offered as solid proof for the existence of telepathy. These investigations involved two young men, Smith and Blackburn, who apparently could communicate with each other tele-

pathically. Many years later, Blackburn published a confession of how he and Smith had tricked the investigators. This case was also deleted from later printings of the Society's proceedings (Hyman 1985a).

2. From Psychical Research to Parapsychology

2.1 J. B. Rhine

Joseph Banks Rhine and his wife Louisa came to Duke University in 1927 and developed a program of research based on card guessing experiments. They termed their program 'parapsychology' and focused their efforts on making this discipline into an experimental science. They replaced the regular deck of playing cards with a deck consisting of five distinct symbols—a circle, a square, three wavy lines, a cross, and a star. Each symbol occurred five times, making a deck consisting of 25 cards. The same set of cards could be used as targets for experiments in telepathy, clairvoyance, and precognition. In addition, the symbols could be presented in a variety of ways.

In a typical clairvoyance condition, for example, the deck would be shuffled and the subject would be asked to guess the symbol of each card at each location in the deck. Since there were five different symbols, the subject had one chance in five of being correct on each guess just by chance. In other words, the subject would be expected to average five correct hits on each run through a deck of 25 cards. In a precognition test, the subject would make his or her guesses before the deck was shuffled. In a telepathy test, a sender would look at the symbol on each card at the time the subject was making his or her guess.

When Rhine published his first monograph, *Extra-Sensory Perception*, in 1934 he had collected data on a total of 91,174 guesses in three years. The average number of hits per run of 25 guesses was 7.1 as compared with the chance expectation of 5. Although the results were barely 2 hits per deck higher than the expected 5, the probability of such a departure from chance, given the huge number of trials, is very low.

The scientific community, at first, took great interest in this outcome. Never before had so much scientific data been amassed in favor of a paranormal claim. After some years, however, skepticism and criticism from the scientific community prevailed. In part, this was because several attempts by independent investigators to replicate these findings failed. In addition, criticisms were aimed at both the statistical procedures as well as the experimental controls. For the most part, the statistical criticisms turned out to be unfounded or insufficient to explain away the results. The methodological critiques were more serious, and gradually parapsychologists improved their procedures for preventing sensory leakage, randomizing the targets, specifying the criteria in advance, and the like.

The paradigm that Rhine initiated under the rubric of parapsychology transformed psychical research into a serious contender for a place among the accepted sciences. It also established the basic procedures for parapsychology for a period covering the years from 1930 to around 1970. During this time the ESP cards were the predominant stimulus for experiments in telepathy, clairvoyance, and precognition. Rhine also initiated the study of psychokinesis (PK) by having subjects attempt to influence the outcomes of dice tosses. The full story of the first couple of decades of this formative period is recounted in Mauskopf and McVaugh (1980).

2.2 S. G. Soal

As already indicated, the problem of replicability haunted the early results of Rhine's card-guessing experiments. Rhine and other US parapsychologists claimed that their results, when they were successful, were equally good for telepathy and clairvoyance trials. The British parapsychologists, however, claimed that their successful experiments only involved telepathy and never clairvoyance. Perhaps the strongest critic of Rhine's early work was the British mathematician, S. G. Soal. During the years 1934–9, Soal managed to amass 128,350 guesses from 160 percipients. He reported that the number of hits was just what would be expected by chance.

A fellow parapsychologist, Whately Carington, eventually convinced Soal to review his data to see if some of his percipients might reveal a 'displacement effect.' Perhaps, he suggested, some of the subjects' guesses were systematically off by one or two targets. Soal, indeed, discovered that two of his 160 subjects showed significant hitting above chance when their guesses were matched not against the intended target but, instead, against the target for the next trial.

Soal realized that finding such displacement patterns in two out of 160 subjects after the fact could be a statistical quirk. He was able to find and persuade one of these subjects, Basil Shackleton, to participate in 11,378 more guesses during the years 1941–3. As was the case in the original experiment, Shackleton's guesses, when matched against the actual target, were consistent with chance. However, when compared with the symbol that came after the target (precognitive hitting), his guesses were well above chance. Soal reported the odds against chance for such a pattern as 10^{35} to 1. Beginning in 1945, Soal was able to gather 37,100 guess from the second subject, Gloria Stewart. Unlike her previous results or those of Shackleton, Mrs. Stewart showed no displacement effect. Instead, her guesses of the intended target were significantly above chance. Soal calculated the odds against chance of this outcome as 10^{79} to 1.

Such strikingly successful results over a period of years, especially from a strong critic of Rhine, caught

the attention of both parapsychologists and skeptics. Parapsychologists hailed the experiments as definitive proof of psi. Critics looked for possible flaws or weaknesses. The early critics managed to find some suspicious patterns, but nothing conclusive. Some proposed fanciful scenarios, involving the collusion among several investigators, to account for the results.

It wasn't until 1978 that Betty Markwick published her findings which convinced most parapsychologists and all the skeptics that Soal had indeed cheated (Hyman 1985a). Ironically, Markwick made her discovery while she was carrying out an elaborate investigation to vindicate Soal's reputation. As part of her plan, she searched tables of logarithms for matches to sequences of targets in Soal's experiments. Because Soal had claimed that he had used Chambers' tables for his random sequences, Markwick reasoned that if she could find matches between Soal's target sequences and sequences in Chambers' tables, this would show that, at least, Soal had not tampered with the target sequences. What she found were partial matches. Many of Soal's sequences had extra insertions within the sequence and these extra insertions usually corresponded with hits. The strong implication was that Soal had deliberately constructed his target sequences with blank spaces inserted systematically. In recording the percipient's guesses, he could surreptitiously insert target numbers in the blank spaces that would match the guess for that trial. Such insertions would suffice to bias the results towards a highly significant, but spurious, outcome.

2.3 *The Post-Rhine Era*

Around 1970, Rhine's specific targets and procedures began to be discarded in favor of new targets and procedures. The striking successes of the early 1930s were no longer in evidence. Results were inconsistent, replicability was still elusive, and the hit rate—even in the few successful outcomes—had become very small. Cynics attributed the declining hit rate to better controls and other methodological and statistical improvements.

Most parapsychologists, however, blamed the weak or nonexistent effects to the austereness of Rhine's procedures. The targets on the ESP cards were meaningless and abstract. The constraint on the subjects' responses to just one of a few choices and the rigidity of the experimental plans all conspired, in the view of many parapsychologists, to dampen the delicate ESP ability.

So parapsychologists began using targets that were emotionally and otherwise meaningful—actual photographs, motion picture clips, actual physical sites, etc. They also eschewed the forced-choice methods of Rhine in favor of free-responding, i.e., the percipients were encouraged to say whatever came to mind and to describe their 'mentation' in whatever words they chose. In addition, they began to take advantage of new technological breakthroughs, especially video and computer developments. These more natural targets and fewer constraints upon the subjects' responding made problems of control and scoring more complicated, but parapsychologists felt that this was a small price to pay in order to achieve what they call psi-conducive conditions.

3. *The Contemporary Scene*

Contemporary parapsychology has generated a variety of novel approaches. Computers are central to much of the research. A line of research that goes back to the late 1960s involves having subjects attempt to influence psychically the output of random number generators (RNG). Computers are also used to control many of the aspects of conducting experiments such as selecting targets, recording the guesses, reporting the outcomes, etc. Advances in neuroscience are also exploited and many experiments use physiological indices instead of, or in addition to, verbal responses. Among these new approaches, three lines of research predominate. The research on remote viewing began in 1972 at the Stanford Research Institute. Until 1995, much of that research, conducted under the auspices at first of the CIA and later the Defense Intelligence Agency was classified. The work was declassified in 1995 and evaluated by a panel consisting of the statistician, Jessica Utts, and the psychologist, Ray Hyman. Hyman and Utts disagreed about the scientific status of remote viewing. Utts concluded that the evidence clearly supported the existence of psi and Hyman that the scientific claim had not been proven. Both agreed, however, that the results of remote viewing were too unreliable to justify using them for information gathering.

The second line of research, initiated by Helmut Schmidt in 1969, involves having subjects attempt to influence or bias the outputs of random number generators (RNGs). Although many different investigators and laboratories have published research on apparent psychokinesis or PK effects on RNGs, the overwhelming number of trials in this line of research has come from the Princeton Engineering Anomalies Research (PEAR) program at Princeton University under Robert Jahn and his associates. In 1987, Radin and Nelson (Radin 1997) conducted a meta-analysis of the RNG experiments conducted from 1959 through 1987. A meta-analysis is a method for combining or pooling the results from several previously published reports involving similar procedures. They reported that the combined outcome produced an effect with odds against chance of more than a trillion to one. Radin claims that the results of such studies completed since 1987 are still consistent with that meta-analysis.

Despite these claims for the RNG research, the studies have been criticized on a number of grounds. Although the meta-analysis pools the data from all the laboratories, the pooling does not deal with the various and different patterns of outcomes from separate laboratories. For example, the PEAR researchers claim to have found a distinct pattern: The RNG outputs slightly more positive units when the subject is trying to influence it in that direction and slightly more negative units when the subject is trying to bias the output in the negative addition. In the neutral condition, the output is evenly divided but the variability is less than would be expected by a truly normal distribution. When the results from these three conditions are pooled the distribution becomes perfectly normal with a mean of zero. No other laboratory has replicated this interesting pattern. The replicability that Radin and Nelson claim is a rather weak one. In effect, any departure from the chance baseline, no matter the specific pattern, is considered equal to any other departure. The one physicist who has deliberately tried to replicate Jahn's RNG findings, Stanley Jeffers, has failed to do so.

3.1 *The Ganzfeld-psi Experiments*

The line of research that has had the most influence since the early 1970s and is at the center of parapsychology's claim of a demonstrable psychic anomaly is the ganzfeld-psi program. The March 1985 issue of *The Journal of Parapsychology* was devoted to 'The Ganzfeld Debate.' This debate consisted of a 47-page critique by Ray Hyman of the 42 ganzfeld-psi experiments known to exist at that time. This was followed by a 47-page response by Charles Honorton, one of the pioneers of this type of experiment.

Hyman (1985b) agreed that, taken as group, the overall hit rate in the ganzfeld experiments exceeded what would be expected by chance. However, he identified a number of statistical and methodological flaws characterizing the experiments. These were flaws which parapsychologists recognize as ones that should be avoided. Yet, not one of the experiments in this data base was free of all the flaws. Hyman concluded that, given the various flaws, the alleged significant levels had been vastly overinflated and the set of experiments could not be used to draw any conclusions about the existence of psi.

In his response, Honorton (1985) devised a different method for assigning flaws and concluded, contrary to Hyman, that the flaws, while there, had no impact on the outcome of the experiments. The controversy continued with parapsychologists supporting Honorton's analysis and the skeptics supporting Hyman's. Rather than continue with further responses and rebuttals, Hyman and Honorton collaborated on a joint paper to focus on areas of agreement and to suggest guidelines for future ganzfeld-psi research (Hyman and Honorton 1986).

In 1990 Honorton and his colleagues published an article in the *Journal of Parapsychology* (Honorton et al. 1990) which summarized six years of new ganzfeld-psi experiments supposedly meeting the guidelines devised by Hyman and Honorton. The new set of experiments, mainly controlled by computers, became known as the autoganzfeld experiments. The hit rate was consistent with the previous set of ganzfeld experiments and was highly significant. In 1994, a major psychological periodical, the *Psychological Bulletin*, published the article 'Does psi exist? Replicable evidence for an anomalous process of information transfer' jointly authored by the psychologist Daryl Bem and Charles Honorton (Bem and Honorton 1994) (unfortunately, Honorton died before the article was published). The same issue contained a commentary by Hyman (Hyman 1994) as well as a 'Response to Hyman' by Daryl Bem (Bem 1994).

Hyman questioned the claim by Bem and Honorton that the autoganzfeld findings were consistent with the original ganzfeld data base. He discovered two peculiar and striking patterns in the autoganzfeld data. While Hyman found these two patterns suggestive of an artifact, Bem responded that if it turned out to be a property of psi, parapsychologists should call it the Hyman Effect.

The ganzfeld-psi research continues in many laboratories around the world. Radin (1997) describes a meta-analysis of all the ganzfeld-psi studies through early 1997. He concludes that 'The overall hit rate of 33.2 percent is unlikely with odds against chance beyond a million billion to one' (p. 88). This contrasts sharply with the conclusions from a meta-analysis of 30 ganzfeld experiments by Milton and Wiseman (1999). These experiments were conducted after the autoganzfeld experiments and yielded an effect size consistent with zero. Milton and Wiseman 'conclude that the ganzfeld technique does not at present offer a replicable method for producing ESP in the laboratory.'

4. *The Case for Psi*

The strongest arguments for the reality of psi have been presented by Jessica Utts (1991) and Dean Radin (1997). Both base their case upon meta-analyses of previously published articles. Utts relies upon meta-analyses of the ganzfeld-psi experiments, early precognition card-guess experiments, RNG studies, early dice-throwing experiments, and studies relating extroversion to ESP performance. Radin includes these as well as several other meta-analyses.

The meta-analyses are used to draw two conclusions. The first is that the pooled effect size emerging from each meta-analysis, although small, is sufficiently large to reject the possibility of being due to chance. The second is that no difference in effect size can be traced to differences attributed to flaws in the ex-

perimental procedures because the effect size is essentially the same for methodologically strong and weak experiments.

It is on this basis that parapsychologists such as Utts and Radin conclude that the evidence for psi is conclusive. Radin argues that the case for psi has been made conclusive several times in the past and that the only mystery is why this is not better known. He implies that the public is unaware of the scientific case for psi because of the attempts by skeptical 'gatekeepers' to prevent the public from knowing about it.

5. *Grounds for Caution*

Despite such strong claims by Utts and Radin, the skeptics insist that there are good reasons for doubting that the evidence for psi is adequate. The meta-analysis of the most recent ganzfeld studies by Milton and Wiseman (1999) strongly suggested that the previous ganzfeld successes cannot be replicated. Worse, Milton and Wiseman found that most of the claims for correlates of psi performance also did not replicate. Although their meta-analysis is controversial, the controversy itself points to major problems with trying to base the proof of psi on meta-analysis of previously published studies. Critics have pointed out many problems with meta-analyses.

When used properly, meta-analysis can be a useful tool to help generate patterns and hypotheses about a body of data. Such a usage is consistent with what statisticians call exploratory analysis. The mistake the parapsychologists seem to make is to use the same meta-analysis both to find possible patterns and then prove the existence of such patterns. This is like having your cake and eating it too. A meta-analysis is inherently retrospective. The results are best used prospectively to make predictions about specific patterns to be found in new sets of data. So far, parapsychologists have not succeeded in doing this.

6. *Prognosis*

Although some key parapsychologists argue that their evidence strongly supports the reality of psychic phenomena, the main body of the scientific community remains unconvinced. At best the parapsychologists have demonstrated, by pooling results from previously published studies, that significant departures from the chance baseline occur more often than expected. They have yet to demonstrate, however, that these departures—if real—originate from a common cause. Nor have they demonstrated any lawfulness in these deviations. Worse, they have yet to come close to specifying a set of conditions and procedures that would enable believers and skeptics to reliably observe these so-called psi events. During the 15 years of conducting meta-analyses on parapsychological experiments, the parapsychologists claim that effect sizes have remained relatively stable. However, this cannot be taken as support for the existence of psi. In a progressive science, as the research develops and the sources of error are tamed, the effect sizes should appreciably grow larger with time. Until this happens in parapsychology, the case for psi will make little headway in the scientific community.

See also: Meta-analysis: Overview; Norms in Science; Physicalism and Alternatives

Bibliography

Bem D J 1994 Response to Hyman. *Psychological Bulletin* **115**: 25–7

Bem D J, Honorton C 1994 Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin* **115**: 4–18

Honorton C 1985 Meta-analysis of psi ganzfeld research: A response to Hyman. *The Journal of Parapsychology* **49**: 51–91

Honorton C, Berger R E, Vargolis M P, Quant M, Derr P, Schecter E I, Ferrari D C 1990 PSI communication in the ganzfeld experiments with an automated testing system and a comparison with meta-analysis of earlier studies. *The Journal of Parapsychology* **54**: 99–139

Hyman R 1985a A critical historical overview of parapsychology. In: Kurtz P (ed.) *A Skeptic's Handbook of Parapsychology*. Prometheus Books, Buffalo, NY, pp. 3–96

Hyman R 1985b The ganzfeld-psi experiment: A critical appraisal. *The Journal of Parapsychology* **49**: 3–49

Hyman R 1994 Anomaly or artifact? Comments on Bem and Honorton. *Psychological Bulletin* **115**: 19–24

Hyman R, Honorton C 1986 A joint communiqué: The psi ganzfeld controversy. *The Journal of Parapsychology* **50**: 351–64

Mauskopf S H, McVaugh M R 1980 *The Elusive Science: Origins of Experimental Psychical Research*. Johns Hopkins University Press, Baltimore, MD

Milton J, Wiseman R 1999 Does psi exist? Lack of replication of an anomalous process of information transfer. *Psychological Bulletin* **125**: 387–91

Radin D 1997 *The Conscious Universe: The Scientific Truth of Psychic Phenomena*. HarperEdge, San Francisco

Utts J 1991 Replication and meta-analysis in parapsychology. *Statistical Science* **6**: 363–403

R. Hyman

Parenthood and Adult Psychological Developments

Extensive debate since the 1950s on the question of whether and how becoming a parent alters one's life has led to three different but not mutually exclusive answers: In terms of individual and marital functioning, the transition to parenthood produces (a) pri-

marily negative changes, (b) few significant negative or positive changes, and (c) an impetus toward psychological development. We conclude that there is truth in all three alternatives.

1. Framing the Questions: 1950–80

Erik Erikson (1959) claimed that people inevitably experience a period of disequilibrium, crisis, and intrapsychic conflict when faced with new and difficult developmental tasks (see *Erikson, Erik Homburger (1902–94)*). Crises are necessary for developmental growth. If they are not resolved, fixations, regressions, and other forms of psychological dysfunction may result. Thus, normative life transitions represent both danger and opportunity.

LeMasters (1957) interviewed 57 married couples who had become parents in the previous five years. He came to what was then a startling conclusion: 83 percent of the couples had experienced an extensive or severe marital crisis within the first year after giving birth. In response, Hobbs and his colleagues conducted a series of studies (e.g., Hobbs and Cole 1976: 728) and argued that 'initiating parenthood may be slightly difficult, but not sufficiently difficult to warrant calling it a crisis.'

There were four main problems with virtually all studies of the transition to parenthood in the 1960s and 1970s: (a) They contained little description of what actually happens as partners became parents; (b) crisis was always described in negative terms and measured in arbitrary and inadequate ways; (c) couples not having babies were not included as a comparison group; (d) all the studies were retrospective and failed to measure change by following couples from before to after they had become parents.

2. Becoming a Family as a Major Life Transition: Research from 1980s to 2000

2.1 What Happens to Men, Women, and Marriage as Partners become Parents?

Longitudinal prospective studies designed to assess the impact of becoming parents emerged in the 1980s. Studies by Belsky and his colleagues (e.g., Belsky et al. 1989) and the authors of this article (e.g., Cowan et al. 1985) were guided by multidomain models of the family as a system (see also Heinicke 1984). Consistency over time and change in average level of function were examined in five family domains: each parent's (a) self-concept and sense of well-being; (b) quality of relationship with their parents; (c) quality of relationship with the baby; (d) marital quality; and (e) balance of life stress and social support. Although some measures of adult functioning and life stress remain stable from late pregnancy through the first year or two of parenthood, various studies find both quantitative and qualitative change in each of these five family domains.

New parents take on new identities and experience shifts in their self-perceptions and their role in the world. Whether or not old intrapsychic conflicts reemerge, new parents find themselves renegotiating relationships with their parents. And now, two new family relationships take center stage—between mother and child and father and child. Friendship networks and work arrangements change for both men and women. There is consistent evidence of significant qualitative and quantitative changes in the marriages of heterosexual couples. (Longitudinal studies of gay and lesbian couples making the transition to parenthood have yet to be published.) The division of family labor becomes more traditional, with women doing more housework than they had before, and far more of the care of the children than either partner expected. Levels of marital conflict increase over time, especially about 'who does what?' (Cowan and Cowan 2000). Not surprisingly, then, satisfaction with marriage declines over the early childrearing years.

2.2 Is Becoming a Parent a Major Life Transition?

We describe transitions as longterm processes that result in a qualitative reorganization of both inner life and external behavior. A major life transition involves a qualitative shift from the inside looking out (how individuals understand and feel about themselves and the world) and from the outside looking in (reorganization of the individual's or family's personal competence, role arrangements, and relationships with significant others). Because systematic longitudinal studies find both quantitative and qualitative shifts in partners' sense of self and relationships—within the family and with friends and co-workers—we have no hesitation in concluding that becoming a family constitutes a major life transition.

3. Does the Transition to Parenthood Elevate the Risk for Individual or Marital Distress?

3.1 Negative Change over Time

Is the disequilibration experienced by couples becoming parents accompanied by increased levels of individual and marital distress? The evidence supports a qualified affirmative answer. For example, women are at increased risk for depression during the transition to parenthood (e.g., Campbell et al. 1992). Approximately 10 percent of women develop clinical depression serious enough to interfere with their daily functioning in the postpartum period. There are no epidemiological data documenting the incidence of postpartum depression or other psychological distress for new fathers.

Of more than 20 longitudinal studies conducted in various parts of the USA, Germany, the UK, and Israel, only three failed to find a statistically significant decline in marital satisfaction after the birth of a first child. In two of the three studies, the couples having babies had married recently, and none of the three included first-time parents in long-term marriages. While our own study included a control group of couples not having a baby and found smaller amounts of negative change in the no-baby couples over two years, after five years the childless couples had a much higher divorce rate. Thus, in the short run, the childless couples were doing well, but over time their relationships were vulnerable, often to issues concerning whether and when to have children. Although simply adding up studies with findings supporting one side or another has drawbacks, we feel relatively assured in concluding that there is a statistically significant risk that many marriages will suffer increases in conflict and declines in satisfaction in the early years of parenthood.

It is important to note that (a) these studies report group trends, (b) not all couples show a decline in marital satisfaction (about 20 percent do not), and (c) the average decline over the first year or two of parenthood is quite small. Nevertheless, because marital satisfaction tends to decline further over time, even small increases in marital conflict or dissatisfaction during the family-making period may be important markers of potential family distress.

3.2 Consistency over Time

Despite significant shifts in average level of functioning from before to after the birth of a first child in the five domains of family life assessed in studies cited here, there is remarkable consistency over time in rank order of well-being or distress in new parents. A baby's arrival is unlikely to destroy well-functioning marriages or generate closer, more satisfying relationships between already troubled partners. The high level of cross-time consistency means that even if the transition to parenthood does not elevate the risk of marital distress, when men, women, and marriage are not doing well before the baby is born, they are likely to be distressed in the early family-making years. In terms of prevention planning, these findings indicate that it is possible to identify individuals and couples who are at heightened risk for later distress and dysfunctional behavior patterns on the basis of assessments made before their babies are born.

3.3 Single Parents

Surprisingly, the transition to parenthood has not been well studied in samples described as high risk by virtue of single parenthood, poverty, or neighborhood violence. Although preventive interventions have been offered to single mothers by home visitors (cf. Olds et al. 1998), the conception of what is needed in these interventions does not take into account that many of the high risk factors operate as additional or amplifying risks, over and above those generally associated with becoming a parent.

4. Can the Transition to Parenthood Promote Adult Development?

Given some of the negative changes and marital distress reported following the birth of a first child, it may seem surprising to ask whether becoming a parent can promote adult development. We suggest, as Erikson (1959) argued, that it is precisely this kind of stress that ultimately can lead to positive developmental change (see *Erikson, Erik Homburger (1902–94)*; *Adult Development, Psychology of*).

Development in adulthood has sometimes been defined as passing normative family life milestones (e.g., entrance to adulthood, courtship, marriage, familymaking) or entering new developmental stages. However, there are neither invariant milestones that precede parenthood nor a convincing body of theory or research to suggest that there are hierarchically integrated stage-like organizations of cognition, personality, and relationships associated with becoming a family.

Elsewhere we have advanced an alternative view (P. Cowan 1988). At the beginning of the 1960s, Allport's functional theory of personality (1961) provided a description of maturity as a general criterion of personality development in terms of six characteristics: (a) extension of the sense of self; (b) warmth in relating to others; (c) emotional security/self-acceptance; (d) realistic perceptions and skills; (e) self-insight and humor; and (f) the establishment of a unifying philosophy of life. Change in the direction of maturity occurs when a person's level of functioning can be described as having become more differentiated, integrated, and adaptive.

We are not aware of systematic longitudinal studies of parenthood as a developmental phase or stage from any of the theoretical perspectives outlined here. Anecdotal evidence from participants in our studies of several hundred couples suggests that babies do stimulate shifts in their parents in the direction of increased maturity (see P. Cowan 1988). Most new parents describe complex changes in their identities. Others describe their relationships with their baby as having 'opened them up' to deep feelings that challenge them to become more effective problem solvers. We do not know yet whether these shifts occur together in the same person, or which factors lead some new parents to become more mature.

5. Implications for Intervention

We have been disappointed to find that few service providers and researchers have created and systematically evaluated interventions for expectant parents designed to (a) reduce or prevent the negative outcomes associated with this major life transition, and (b) increase parents' ability to maintain strong and productive family relationships. Early on, Shereshefsky and Yarrow (1973) provided individual counseling for some expectant mothers. Clulow (1982) conducted monthly meetings for expectant couples, but not in a controlled study with a no-intervention comparison group. Our own intervention (Cowan and Cowan 1999) based on random assignment to a no-intervention comparison or to groups of couples meeting weekly with mental health professionals for three months prepartum and three months postpartum had initially promising results. It staved off declines in marital satisfaction from 6 to 66 months postpartum, and all intervention couples remained in intact marriages until their children were three, while 15 percent of the no-treatment controls had separated or divorced. Because of increasing evidence that marital function plays an important role in shaping parent–child relationships and affecting children's early development (Cummings and Davies 1994), we believe that preventive interventions for partners becoming parents have the potential to promote more positive trajectories of adult development and family life quality.

6. Directions for Future Research

More research is needed to understand individual differences in pathways through the transition to parenthood and why some couples show increased satisfaction and adaptation while others decline. The question of whether and how the transition to parenthood produces development in adulthood is wide open for further research. Systematic studies of interventions to facilitate individual and couple functioning as partners become parents could provide important information for strengthening new families, at the same time testing hypotheses about the mechanisms involved in producing positive change during this major life transition.

See also: Adult Development, Psychology of; Adulthood: Developmental Tasks and Critical Life Events; Fatherhood; Marriage and the Dual-career Family: Cultural Concerns

Bibliography

Allport G W 1961 *Pattern and Growth in Personality*. Holt, Rinehart and Winston, New York

Belsky J, Rovine M, Fish M 1989 The developing family system. In: Gunnar M R, Thelen E (eds.) *Systems and Development*. Erlbaum, Hillsdale, NJ, pp. 119–66

Campbell S B, Cohn J F, Flanagan C, Popper S, Myers T 1992 Course and correlates of postpartum depression during the transition to parenthood. *Development and Psychopathology* **4**: 29–47

Caplan G 1964 *Principles of Preventive Psychiatry*. Basic, New York

Clulow C F 1982 *To Have and to Hold: Marriage, the First Baby and Preparing Couples for Parenthood*. Aberdeen University Press, Aberdeen, UK

Cowan C P, Cowan P A 1999 *When Partners Become Parents: The Big Life Change for Couples*. Erlbaum, Mahwah, NJ

Cowan C P, Cowan P A, Heming G, Garrett E, Coysh W S, Curtis-Boles H, Boles A J 1985 Transitions to parenthood: His, hers, and theirs. *Journal of Family Issues* **6**: 451–81

Cowan P A 1988 Becoming a father: A time of change, an opportunity for development. In: Bronstein P, Cowan C P (eds.) *Fatherhood Today: Men's Changing Role in the Family*. Wiley, New York, pp. 13–35

Cummings E M, Davies P 1994 *Children and Marital Conflict: The Impact of Family Dispute and Resolution*. Guilford, New York

Erikson E 1959 Identity and the life cycle. *Psychological Issues* **1**: 1–171

Heinicke C M 1984 Impact of prebirth parent personality and marital functioning on family development: A framework and suggestions for further study. *Developmental Psychology* **20**: 1044–53

Hobbs D, Cole S 1976 Transition to parenthood: A decade replication. *Journal of Marriage and the Family* **38**: 723–31

LeMasters E E 1957 Parenthood as crisis. *Marriage and Family Living* **19**: 352–5

Olds D et al. 1998 The promise of home visitation: Results of two randomized trials. *Journal of Community Psychology* **26**: 5–21

Shereshefsky P M, Yarrow L J 1973 *Psychological Aspects of a First Pregnancy*. Raven, New York

P. A. Cowan and C. P. Cowan

Parenting: Attitudes and Beliefs

Parenting has been investigated extensively by developmental psychologists ever since Freud drew formal attention to its critical significance for children's social, emotional, and intellectual functioning. In studying parenting, researchers have come to believe that parents' attitudes and beliefs are centrally important in directing their treatment of children. Therefore, much effort has been focused on looking at links between how parents think about children and child rearing—both generally and in specific situations—and child outcomes. In this article, global and specific attitudes and beliefs are discussed, with an emphasis on the challenges provided by their measurement.

1. What are Attitudes and Beliefs?

Attitudes include beliefs or opinions, an evaluation of those beliefs, and an intention to act in accord with them (Ajzen and Fishbein 1980). People with a

positive attitude to exercise, for example, may reveal their positive feelings by agreeing with others who endorse the same position, by exercising more themselves, by admiring others who exercise, or by encouraging others to exercise. Thus attitudes provide a link, albeit not a perfect one, between thinking and action.

2. Global Attitudes and Beliefs

Dozens of measures of parental attitudes and beliefs have been devised, as investigators have tried to understand the relation between parents' thoughts and actions and children's resulting development. Holden and Edwards (1989) and Holden (1995) have noted the basic assumptions in this approach as well as the fact that they are sometimes violated. These include the premise that parents have pre-existing attitudes, although this is not always the case; that attitudes do not change over time, although they can certainly be modified by education and experience with children; and that parents are not ambivalent with respect to them, although they often are. When these assumptions are violated, relations between attitudes, parents' behavior, and children's development will be less robust.

2.1 Approaches and Findings

Attitudes are assessed either through interviews or questionnaires, with the latter more frequently employed. The majority of attitude questionnaires have focused on parents' views about different child-rearing practices, with a smaller number addressing views of children (Holden 1995). Questionnaires addressing ideas about child-rearing practices have focused on two major features: the quality of the parent's relationship with the child and the extent to which the parent attempts to control the child's behavior. Questions therefore deal with issues of warmth, acceptance, responsiveness, coldness, rejection, and hostility on the one hand, and permissiveness, firm control, punitiveness, and restrictiveness on the other. Typical statements, with which parents indicate their degree of agreement or disagreement, are: 'Children should always treat adults with respect' and 'I enjoy being with my child.'

Researchers have looked at relations between these attitudes and a variety of features of children's development. Positive attitudes toward warmth and reasonable firm control are moderately predictive of outcomes such as high academic achievement, high self-esteem, and a strong moral orientation, while endorsement of restrictiveness and lack of acceptance relate moderately to negative outcomes (Maccoby and Martin 1983). It should be noted that most of the research has been conducted with Anglo-European middle-class mothers and their children and that it is not clear the same pattern of relations holds for lower socioeconomic classes, in other cultural contexts, or with fathers (Parke and Buriel 1998).

2.2 Limitations of Questionnaire Measures and Needed Improvements

Although the assessment of parenting attitudes has provided useful information for developmental psychologists, dissatisfaction with modest relations between attitudes and child outcomes has fostered attempts to improve existing approaches (Holden 1995). For example, investigators have begun to show a greater concern with the actual relation between attitudes and behavior. Given the basic assumption that attitudes inform behavior, and that it is parenting behavior that determines children's development, there is, of course, little use in studying attitudes if they do not reliably reveal themselves in parenting actions. Reasonable attitude–behavior links are more likely to be found when items assessing attitudes on questionnaires match the nature of the parent's action, e.g., the response to a question about the importance of firmness in getting children to obey is related to observations of reactions when a child fails to comply with a parental request. As well, links are stronger when many examples are added together to provide an aggregated measure of a particular parenting behavior. The fact that relations between attitudes and behavior are stronger under these two conditions underlines two things: asking about general attitudes may give incomplete information when it comes to specific actions, and parenting behavior varies across situations and is determined by a number of variables so that a variety of situations needs to be assessed in order to gain an accurate impression of a particular parent's attitudes.

Holden and Edwards (1989) have raised other methodological concerns. For example, statements on questionnaires are often couched in the third person on the assumption that this may lead to more valid answers because respondents are not placing themselves in a negative light: On the other hand, this approach produces situations in which respondents are not sure whether they are to answer with respect to what might be ideal attitudes rather than what they know to be their own attitudes. Another problem is that attitude measures are based on self-report data that depends on the willingness of the respondent to be honest or even to have access to or conscious awareness of a particular attitude. Psychologists have become increasingly aware that much behavior is determined by implicit or automatic processes, and a variety of techniques have been developed to study the operation of these processes (Bugental and Goodnow 1998). Clearly, however, asking people to state their position on a given topic is not going to be effective if that position is unavailable to conscious introspection.

3. Parental Thinking in Specific Situations

One response to the problems of measuring global attitudes has been to focus on the thoughts or beliefs parents have when they are interacting with their children in specific situations. Thus, rather than ask about general and abstract beliefs and attitudes, researchers have asked about the content of parents' thinking in a particular context. For example, as noted above, inquiries about global attitudes have focused on parents' beliefs about the best child-rearing practices, e.g., authoritarianism, or the extent to which children should obey their parents. More recent inquiries focus on goals, such as obedience, that parents have in a particular situation. Thus the interest in general attitudes has been translated into an interest in goals at a particular point in time, with the recognition that goals of a parent change as a function of context (although, on average, some people may be more inclined to favor some goals more than others). Hastings and Grusec (1998) have shown that, when a child misbehaves, many parents have obedience as their goal and they are punitive. If all one knew was the extent to which those parents had endorsed a general measure of authoritarianism without knowing their specific goal, prediction of their actual behavior would be less accurate.

The study of particular goals reflects the interest of developmental researchers in attitudes toward child-rearing practices. However, interest in parenting cognitions has expanded to a wider range of content including theories about how children develop, expectations about children's abilities, and feelings of self-efficacy with respect to parenting. In the next three sections three additional areas that have received considerable attention will be described.

3.1 Causal Attributions

Parents try to find explanations for why their children have behaved in a particular way, that is, they make 'causal attributions.' Although people make causal attributions in a variety of contexts, parents are probably particularly likely to make them because they need to understand their children so that they can effectively influence them. Dix and Grusec (1985) have outlined some of the features of parents' thinking in this context. In the search for explanation, parents can make internal or dispositional attributions that find the source of action in the child's personality or character. Alternatively, they can make external attributions that locate the source of action in the external situation or environment. When a negative action is attributed to dispositional factors it is most often seen as intentional and under the child's control. In this case parents have been shown to react punitively, possibly both because of their accompanying anger and because a Western ethical system dictates that intentional misdeeds should be punished. Parents who make external attributions, believing their child was tired or provoked or did not know any better, are likely to respond in a more benign way, e.g., by reasoning in an attempt to guide their child's future actions. When the attributions are accurate, parental behavior is likely to be appropriate for modifying the child's actions. When the attribution is inaccurate, however, parenting will be ineffective, given that children who lack knowledge, for example, are merely punished and do not learn what correct actions are, or children who have knowledge do not experience the negative consequences of their actions.

3.2 Relative Control

The impact of parental thinking has also been mapped in the domain of perceptions of relative control between parents and children. In her research program, Bugental (Bugental 1992, Bugental and Goodnow 1998) has shown how parents' beliefs about such control are elicited in difficult interactions with their children. Once elicited, these beliefs have an impact on emotional reactions, behavior, and, subsequently, child outcomes. Parents who perceive their children as having more control of the difficult situation than they do experience physiological arousal that reflects feelings of threat aroused by the apparent reversal of power. Because they feel relatively powerless they resort to coercive and overly aggressive behavior. At the same time they transmit confusing and inconsistent messages to the child, both because their cognitive capacity is limited by the distracting thoughts associated with threat and because they cannot adequately inhibit the negative feelings they are experiencing. In turn, the confusing behavior leads to nonresponsiveness on the child's part that serves to reinforce the parents' negative thoughts and affect.

3.3 Attachment Relationships

Early in their development children learn that their needs for security will either be dealt with satisfactorily, rejected, or responded to inconsistently (Ainsworth et al. 1978) (e.g., see *Attachment Theory: Psychological*). These early experiences form the basis of adult working models of relationships that function throughout the life course and that manifest themselves in parenting behaviors. Thus adults who are secure with respect to the satisfaction of their emotional needs have a model of relationships that enables them to respond sensitively and supportively to their children. Adults whose model includes others as rejecting are cool, remote, and task-focused in their interactions with children, while those with models of other as inconsistent are unpredictable and confusing in their parenting actions (Crowell and Feldman 1991).

4. Automatic vs. Conscious Thinking

The earlier discussion of global attitudes noted that some of these attitudes are not open to conscious awareness. The same distinction has been made in the area of parenting cognitions, with some thoughts seen as automatic and implicit and others as more open to conscious reflection and awareness. Bugental has argued, for example, that the experience of a difficult child activates a high threat schema that operates at a preconscious level and places parents in a state of perceptual readiness to react to certain events in the environment. Thus the stress produced by a problematic interaction with a child is filtered through this schema, which in turn influences both emotional feelings and more conscious or deliberate ideation. Mental representations of childhood attachment experiences are also considered to operate outside conscious awareness (Bretherton 1985). Parents' goals and attributions about children's behavior may be more open to conscious awareness and reflection—they frequently change in response to changes in the external situation. On the other hand, some parents are not able to use situational information to change their attributions for child misbehavior (Milner and Foody 1994), an indication that even the same kinds of cognitions may be sometimes more and sometimes less automatic.

5. Future Directions

The use of global parenting attitudes and beliefs and specific cognitions to understand parent–child relationships will continue to be a significant part of the arsenal of developmental researchers. Measurement issues remain of paramount importance. Strong linkages between measured attitudes and behavior still need to be demonstrated. A fundamental question has to do with how to measure attitudes and beliefs that are automatic in nature and, as such, cannot easily be articulated by respondents even if they are perfectly willing to be open and honest in their answers. The move must be away from straightforward transparent questions to more subtle measures where 'correct' or socially acceptable answers are not so self-evident. Bugental's measure of relative control (Bugental 1992) and the Adult Attachment Interview which assesses mental representations of attachment relationships (Main and Goldwyn 1991) are examples of such measures, and they have proved to be accurate predictors of a variety of parenting behaviors and child outcomes. Another important question has to do with the origins of various parenting belief systems. Some presumably are part of the cultural context in which parents find themselves. Others come from experience with one's own children. And others emerge as a result of a parent's own experiences as a child; it is these latter, presumably, that are most automatic and most difficult to alter when alteration seems desirable. Finally, linkages among the many sets of beliefs that have taken up the attention of researchers need to be made, as well as explorations of the distinctive ways in which each contributes to the parenting process.

See also: Attitudes and Behavior; Family and Gender; Fatherhood

Bibliography

Ainsworth M D S, Blehar M C, Waters E, Wall S 1978 *Patterns of Attachment: A Psychological Study of the Strange Situation*. Lawrence Erlbaum Associates, Hillsdale, NJ

Ajzen I, Fishbein M 1980 *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall, Englewood Cliffs, NJ

Bretherton I 1985 Attachment theory: Retrospect and prospect. In: Bretherton I, Waters E (eds.) Growing points of attachment theory and research. *Monographs of the Society for Research in Child Development*. 50 Serial No. 209, pp. 3–35

Bugental D B 1992 Affective and cognitive processes within threat-oriented family systems. In: Siegel I E, McGillicuddy-DeLisi A V, Goodnow J J (eds.) *Parental Belief Systems: The Psychological Consequences for Children*, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, NJ

Bugental D B, Goodnow J J 1998 Socialization processes. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology Vol. 3 Social, Emotional, and Personality Development*. J. Wiley, New York, pp. 389–462

Crowell J A, Feldman S S 1991 Mothers' working models of attachment relationship and mother and child behavior during separation and reunion. *Developmental Psychology* **27**: 597–605

Dix T, Grusec J E 1985 Parent attribution processes in child socialization. In: Sigel I (ed.) *Parent Belief Systems: Their Psychological Consequences for Children*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 201–33

Hastings P D, Grusec J E 1998 Parenting goals as organizers of responses to parent-child disagreement. *Developmental Psychology* **34**: 465–79

Holden G W 1995 Parental attitudes toward childrearing. In: Bornstein M (ed.) *Handbook of Parenting Vol. 3 Status and Social Conditions of Parenting*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 359–92

Holden G W, Edwards L A 1989 Parental attitudes toward childrearing: Instruments, issues, and implications. *Psychological Bulletin* **106**: 29–58

Maccoby E E, Martin J A 1983 Socialization in the context of the family: Parent–child interaction. In: Hetherington E M (ed.) *Handbook of Child Psychology Vol. 4 Socialization, Personality and Social Development*. Wiley, New York, pp. 1–102

Main M, Goldwyn R 1991 Adult attachment rating and classification systems. Unpublished manuscript, University of California at Berkeley

Milner J S, Foody R 1994 Impact of mitigating information on attributions for positive and negative children's behavior by adults at low-risk and high-risk or child abusive behavior. *Journal of Social and Clinical Psychology* **13**: 335–51

Parke R D, Buriel R 1998 Socialization in the family: Ethnic and ecological perspectives. In: Eisenberg N (ed.) *Handbook of Child Psychology Vol. 3 Social, Emotional and Personality Development*. Wiley, New York, pp. 463–552

J. E. Grusec

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Parenting in Ethnic Minority Families: United States

'Parenting' refers to the training or raising of children, focusing on parents' responsibility to guide children's development toward behaviors determined as functional in the surrounding culture. Parenting practices are shaped by the contextual circumstances in which the parenting occurs. One such contextual circumstance is ethnic minority status. Ethnic minority status in the USA usually refers to people of color, those with physical characteristics differentiating them from the majority white American. Ethnic minority groups include African Americans, Asian Americans, Mexican Americans, and American Indians. The implications of ethnic minority status have been influenced greatly by the difficulties such designated groups have experienced in attempting to integrate and assimilate into the mainstream of US society.

Past conditions of involuntary immigration and slavery, legal discrimination and segregation, exclusion and restrictions, and forced removal and internment to reservations have influenced the attitudes, perceptions, and behavioral patterns of the respective ethnic minorities. Interestingly, it does not matter whether ethnic minority groups were volunteer or forced immigrants or whether they were an indigenous people, their historical struggles of assimilating into US society was qualitatively different from other noncolor ethnic Americans (Ogbu 1981, Wilson 1989). This entry considers the evolved parenting practices of the four ethnic American groups, focusing on African-American, Chinese-American, Mexican-American, and American Indian practices, illustrating their marked distinction from the dominant culture as well as variations between the ethnic groups.

1. *Parenting Behaviors and Minority Families*

Historically, past research on parenting has described practices in minority families in terms of their contrast to practices of the majority. In comparisons, dominant cultural norms and values have been used to evaluate the practices of minority parents, deeming them changeworthy. Comparing practices of ethnic minorities with those of the majority remains an important topic given the stigma and social implications such comparisons elicit for ethnic minorities in the USA. Closer examination of parenting in each minority group suggests a link between the differing historical experiences of the groups and their evolved parenting practices. Analyses of adolescent adjustment indicate the adaptive nature and functionality of parenting among ethnic minorities.

Two basic parental behaviors, warmth and control, presume to underlie most categories of parenting practices. Baumrind (1991) formed four patterns of parental authority based on differing levels of parental warmth and control. The patterns emerging from Baumrind's studies were authoritarianism, authoritativeness, permissiveness, and neglectful/rejecting. Authoritarian parents displayed high levels of parental control but low levels of warmth. Authoritative parents were high in both parental warmth and control. Permissive parents were highly responsive (i.e., warm) to their children, but did not place many demands (i.e., control) on them (Baumrind 1991). Neglectful/rejecting parents were low in both parental control and warmth.

Because authoritarian, authoritative, and permissive parenting represent a central concern of this entry, additional comments are warranted. Specifically, authoritarian parents are described as using absolute standards in order to shape and control the attitudes and behaviors of their children. They place high regard on such values as respect for authority, hard work, and maintenance of order; simply put, the parents' word is not to be questioned. Authoritative parents are best described as applying firm parental authority while engaging in and encouraging parent–child interaction with respect to socialization. Further, authoritative parents address their child's affective needs. Permissive parents do not maintain and enforce clear behavioral expectations but are highly attentive to the child's affective needs (Baumrind 1991).

Since parenting practices are assumed to be flexible and contextually sensitive to situational demands, comparative claims will not be offered regarding the advantages of any parenting style. Parents presumably incorporate their own beliefs about how the world works into their child-rearing practices. Such beliefs, developed by parental experiences, naturally have some bearing on what parents believe their children are likely to face in their own lives. Hence, a basis of understanding parenting practices is the notion that many contextual differences will shape the child-rearing practices. Each parenting style is manifested in attitudinal and behavioral attributes that vary both between different ethnic groups and within ethnic groups (see *Cultural Variations in Interpersonal Relationships*).

2. *Socialization in African-American Families*

Wilson (1989) asserts that an important social context of African-American family development involves the involuntary migration to the Americas followed by a protracted period of African-American enslavement and, after emancipation, segregation and inferior status base on race and struggle for civil rights and political freedom. The African-American family evolved into an extensive familial structure that was based on mutual support and obligation.

Two types of studies have explored the child-rearing practices common to African-Americans. Compara-

tive studies have focused on the differences between African-American child-rearing practices and those common to other ethnic groups (Rickel et al. 1988, Baumrind 1991, Bartz and Levine 1978), whereas within-group approaches focus on the variance in child-rearing practices among African-Americans only (Hill 1995, Kelley et al. 1993, Luster and McAdoo 1994). Much of the comparative research on parenting describes African-American parenting in terms of its relation to the dominant group's child-rearing practices and ideals (e.g., Baumrind 1991). In contrast, the within-group research has described the diversity of attributes and behaviors within African-American families (see *Socialization in Adolescence*).

Comparative studies on socialization have generally found African-American parenting styles to be authoritarian and more restrictive than European-American styles (Baumrind 1991, Rickel et al. 1988, Kelley et al. 1993). Also, African-American parents appear to employ gender-specific socialization practices. Specifically, African-American boys are more likely to be expected to act maturely and responsibly than are European-American boys. Parents of African-American girls have been shown to exert high levels of control. African-American fathers discourage infantile behaviors in their daughters. African-American mothers display firm enforcement and rejecting behaviors, and discourage nonconformity in their daughters (see *Poverty and Child Development*).

Bartz and Levine (1978) examined specific parenting practices in low-income African-American, European-American, and Mexican-American families. African-American parents exercised more control over their child's behavior, were more likely to press for acceleration of development, were more concerned about their children wasting time, and used supportive behaviors more frequently than European-American and Mexican-American parents. African-American parents also reported doing more to enforce their rules and expectations. Generally, African-American parents value strictness and expect their children to gain a sense of responsibility. They encouraged decision making, expressed loving concern, and closely monitored their children's behavior to assure goals such as obedience and high achievement (Bartz and Levine 1978).

Within-group ethnic studies generally focus on the specific characteristics that influence the development of a particular parenting style (Kelley et al. 1993). Generally, it has been found that mothers with a higher family income and a greater number of children report using more parental warmth and firm discipline and lower levels of control. In addition, discipline has been shown to correlate negatively to the number of adults in a family (Wilson 1989).

Inasmuch as authoritarian practices appear to be common among African-American families, there are important familial caveats that are related to familial socioeconomic levels, maternal resources, and neighborhood. Namely, older, educated, middle-class, married mothers employed more authoritative practices than do young, single mothers with less education. Conversely, young, single, less educated mothers were more likely to emphasize obedience, use physical punishment, and employ a more parent-oriented approach (Kelley et al. 1993). Also, mothers who perceived their neighborhoods as dangerous were likely to use harsh discipline and authoritarian practices.

In addition, Luster and McAdoo (1994) found that adolescents with the highest level of cognitive competence came from homes with older, more educated and high esteem mothers, a supportive environment, and a small family size. Thus, there is variance of socialization practices among African-American families.

3. *Socialization in Chinese-American Families*

Research on Chinese-American parenting practices is often described in ways similar to the characteristics of the authoritarian style. However, the child-rearing practices may be the outgrowth of the Chinese philosophy of hsaio, or filial piety, which prescribes Chinese children to fulfill the desires of their parents (Lin and Liu 1993, Kelley and Tseng 1992, Chiu 1987). Devotion to one's parents takes precedence over other relationships, including one's obligation to one's own children (Lin and Liu 1993). Parents are very involved in the training of their children, and hold a definite position of authority over them. Strict discipline is highly valued. Parental control, therefore, is apparent as children learn that their parents are 'always right' (Chiu 1987, Kelley and Tseng 1992, Lin and Liu 1993).

Kelley and Tseng (1992) examined the child-rearing practices of Chinese-American and European-American mothers. Specifically, they measured use of external threats and punishment, permissiveness, and emphasis on school-related skills, obedience, concern for others, and self-reliance. Chinese-American mothers scored higher than did European-American mothers on ratings of restrictiveness and control (Chiu 1987, Kelley and Tseng 1992).

4. *Socialization in Mexican-American Families*

Mexican Americans make up the vast majority of Latino peoples living in the USA. Other groups include Central and South Americans, Puerto Ricans and Cuban Americans. Mexican Americans have been incorporated in US society through military conquest of the Southwest and ambivalent immigration policies that at some times encourage immigration and at other times deports citizens of Mexican decent. Bartz and Levine (1978) found Mexican-American parents to exhibit more authoritative behaviors than African-

American parents but less authoritative behaviors than European-American parents. Specifically, Mexican-American parents' reported levels of firm control and warmth were intermediate between those reported by African-American and European-American parents. In examination of Mexican-American and European-American mothers from low and high socioeconomic status (SES), Cardona et al. (2000) indicated that Mexican-American and higher SES mothers reported using higher levels of discipline and lower levels of nurturing behaviors than did European-American and lower SES mothers. Moreover, higher SES Mexican-American mothers reported more frequent use of discipline than did other mothers. Further, Buriel (1997) found generational patterns in the Mexican-American parents of first-, second- and third-generational children. Parents of first- and second-generation children reported authoritative-oriented practices whereas the parents of third-generation children reported authoritarian practices. In addition parents of first- and second-generation children reported higher level of parental permissiveness than did the parents of third generation children. Taken together, it would appear that as Mexican-American parents become more acculturated and integrated into US society, they resort to more authoritarian patterns of parenting.

5. *Socialization in American Indian Families*

Although American Indians are a diverse group making up 450 distinct groups, one common experience shared by the diverse groups is their history of military conflicts with white Americans and forced removal from the ancestral homelands and placement on reservation. American Indian customs place great value on collectivism, natural harmony, humility, and respect for and importance of family. Because the family network is an important part of American Indian life, child rearing is a shared activity among the network. Through valued collectivism and the family support network, the growing child is placed in the protected environment of close relationship with familial kin (Forehand and Kotchick 1996). Several researchers have indicated that specific parenting practices among American Indians appear permissive in that American Indians report using low levels of control or punishment and high levels of warmth (Cross 1986, Red Horse 1997).

6. *Socialization of Children and Adolescents in the Context of Ethnic Differences*

Although research maintains that authoritative parenting produces the most optimal child outcomes (Baumrind 1991), numerous studies suggest that the actual positive effects of authoritative practices are correlated with some specific domains of development only (Steinberg et al. 1995) and are generally more relevant to European-American adolescents than ethnic minority adolescents (Steinberg et al. 1995, Darling and Steinberg 1993) (see *Parenting: Attitudes and Beliefs*).

The differential cultural and contextual experiences of ethnic minority groups are likely to provide an explanation for the differing effects of the authoritarian style on child outcome between ethnic minorities and the majority group. Although authoritarian parenting practices among European-American parents may be viewed as excessively harsh and punitive, and may be indicative of parental rejection and lower level of investment, similar practices among African-Americans may be interpreted as a sign of parental involvement and concern (McLeod et al. 1994). African-American mothers report using authoritarian measures to teach their children about harsh realities in a racist society (McLeod et al. 1994, Hill 1995), for teaching values such as respect, child safety, and school achievement (Kelley et al. 1993), and as a protective measure in high-risk environments (Rickel et al. 1988).

Chao (1994) suggests that Chinese-American parents' use of authoritarian and restrictive practices are more indicative of parental concern, involvement, and caring than parental hostility and aggression. Authoritarian practices are part and parcel of successful and harmonious family and community life (see *Cultural Psychology*).

7. *Concluding Remarks and Future Directions*

Several implications of the review on minority families' socialization practices are evident. First, although the concept of authoritarianism is generally analogous to parental hostility, aggression, and dominance, the common occurrence of authoritarian and restrictive parenting practices in African-American, Chinese-American, and Mexican-American populations are not necessarily associated with such negative parental attributes. Rather, the use of authoritarian procedures has demonstrated functional benefits for ethnic minority Americans. Second, although common among several of the ethnic minority American groups, the authoritarian and restrictive practices evolved from different contextual concerns. For African Americans, the use of authoritarian practices were developed and maintained in response to harsh environmental realities like low-income, inner-city living environments, and hostile discriminatory practices of US society. For Chinese Americans, traditional cultural beliefs and values of filial piety initiated the practices. For Mexican Americans, value of the strong patriarchal presence in families appears to increase the use of authoritarian practices. Interestingly, American Indian socialization practices are typically permissive, while close-kin networks maintain protectiveness. Last, different parenting practices do have different

meanings, uses, and benefits for minority and majority children (Chiu 1987, Ogbu 1981). Contending that one method embodies positive outcomes for one group does not imply that another group would or would not benefit from that practice. Parenting practices must be examined carefully from the perspective of the people who employ them and the contexts in which those practices will be employed. In order to understand ethnic minority socialization, it is important to examine applied socialization practices within their particular ecological context (see *Socialization, Sociology of*).

It is clear that the US population will continue its racial and cultural diversification. Likewise it is acceded that a diverse US population will display different kinds of socialization practices. Contextual aspects of socialization will continue to play a fundamental role in understanding and determining adaptive socialization processes. Thus, socialization practices must be appreciated for their particular situational meaning, significance, and consequence. Rather than presume peremptory negative and positive connotations of authoritative and authoritarian parenting practices, the merits of each style should be judged according to its contextual contingencies and influence on adolescent adjustment. Also, it is important that we view socialization practices from a cultural perspective. An understanding of the interacting mechanism of culture, family, and socialization may lead to an appreciation of the familial complexities experienced by various ethnic groups. Furthermore, understanding the links between ethnic minority status, socialization practices, and child outcomes may foster appropriately directed family interventions.

See also: Ethnic Identity, Psychology of; Family and Gender; Fatherhood; Gender, Class, Race, and Ethnicity, Social Construction of; Minorities; Parenting: Attitudes and Beliefs

Bibliography

Bartz K, Levine E 1978 Child-rearing by black parents: A description and comparison to Anglo and Chicano parents. *Journal of Marriage and the Family* **40**: 709–19

Baumrind D 1991 Effective parenting during the early adolescent transition. In: Cowan P A, Hetherington M (eds.) *Family Transitions*. Lawrence Earlbaum, Hillsdale, NJ, pp. 111–64

Buriel R 1997 Childrearing orientations in Mexican Amercian families: The influence of generation and sociocultural factors. *Journal of Marriage and the Family* **55**(4): 987–1000

Cardona P G, Nicholson B C, Fox R A 2000 Parenting among Hispanic and Anglo American mothers with young children. *Journal of Social Psychology* **140**(3): 357–65

Chao R K 1994 Beyond parental control and authoritarian parenting style: Understanding Chinese parenting through the cultural notion of training. *Child Development* **65**: 1111–19

Chiu L H 1987 Child-rearing attitudes of Chinese, Chinese-American, and Anglo-American mothers. *International Journal of Psychology* **22**: 409–19

Cross T 1986 Drawing on cultural tradition in Indian child welfare practice. *Social Casework* **65**(5): 283–9

Darling N, Steinberg L 1993 Parenting style as a context: An integrative model. *Psychological Bulletin* **113**(3): 487–96

Forehand R, Kotchick B 1996 Cultural diversity: A wake-up call for parent training. *Behavior Therapy* **21**: 187–206

Hill N 1995 The relationship between family environment and parenting style: A preliminary study of African American families. *Journal of Black Psychology* **21**(4): 408–23

Kelley M R, Sanchez-Hucles J, Walker R R 1993 Correlates of disciplinary practices in working- to middle-class African-American mothers. *Merrill-Palmer Quarterly* **39**(2): 252–64

Kelley M L, Tseng H M 1992 Cultural differences in child-rearing: A comparison of immigrant Chinese and Caucasian American mothers. *Journal of Cross-Cultural Psychology* **23**(4): 444–55

Lin C, Liu W 1993 Intergenerational relationships among Chinese immigrant families from Taiwan. In: McAdoo H P (ed.) *Family Ethnicity: Strength in Diversity*. Sage, Newbury Park, CA, pp. 271–86

Luster T, McAdoo H P 1994 Factors related to the achievement and adjustment of young African-American children. *Child Development* **65**: 1080–94

McLeod J, Kruttschnitt C, Dornfield M 1994 Does parenting explain the effects of structural conditions on children's antisocial behavior? A comparison of Blacks and Whites. *Social Forces* **73**(2): 575–604

Ogbu J 1981 Origins of human competence: A cultural-ecological perspective. *Child Development* **52**: 413–29

Red Horse J 1997 Traditional American Indian family systems. *Family Systems Medicine* **15**(3): 243–50

Rickel A U, Williams D L, Loigman G A 1988 Predictors of maternal child-rearing practices: Implications for intervention. *Journal of Community Psychology* **16**: 32–40

Steinberg L, Darling N, Fletcher A 1995 Authoritative parenting and adolescent adjustment: An ecological journey. In: Moen P, Elder G H, Luscher K (eds.) *Examining Lives in Context: Perspectives on the Ecology of Human Development*. American Psychological Association, Washington, DC, pp. 423–66

Wilson M 1989 Child development in the context of the black extended family. *American Psychologist* **44**(2): 380–5

M. N. Wilson and M. D'Amico

Parents and Teachers as Partners in Education

Parents have always been involved in children's preparation for adulthood, but during the twentieth century many responsibilities were taken over by schools. In Western societies a model of teachers and parents as 'partners in education' gradually evolved which currently enjoys widespread academic and political support. In the contemporary climate of globalization it may well be presented as the ideal for

all societies with mass education systems. Yet critics claim it is difficult to implement at best, and may easily serve further to disempower socially disadvantaged sectors of the population.

1. Evolution of the Partnership Model

In any society, economic, sociocultural, and political conditions, as well as pedagogical theory, help define the preferred model of home–school relations.

1.1 Historical Shifts

During the early decades of mass schooling in Western societies, as teaching methods became more specialized and when teachers were among the most highly educated members of their communities, parents were viewed as ill-equipped to understand what happened in classrooms and liable to do more harm than good if they interfered. In the more widespread affluence of the post-World War II period, parents came under much greater pressure to assist by providing a materially enriched home environment, but were only expected to become significantly engaged in discussion of their child's education if schools identified problems. Parents were not regularly invited beyond the school gate until the 1970s; opportunities were then limited initially to formal discussions about children's progress, or provision of unskilled assistance to teachers, but gradually expanded to include more responsible duties in classrooms and more freedom to engage informally with staff. By the late 1980s research proclaimed major parental involvement *as educators* to be essential to children's success in school, and parents themselves were demanding more say in educational decision-making.

1.2 Contemporary Partnership Programs

The language used to strategize home–school connections has reflected these shifts. Current policies espousing 'partnership' have implications that distinguish them in important ways from those based primarily on parental 'involvement' or 'consultation.' The partnership paradigm requires consensus between teachers and parents on the goals and processes of education, and implies a full sharing of information, skills, decision-making, and accountability.

Early partnership programs were typically the initiatives of individual school principals, but their establishment is now commonly assisted by research-based and/or organizational support. In a large-scale project in New Zealand, for example, experienced senior teachers were imported into schools as 'developers' to support and legitimate innovative practice and constructively challenge institutional resistance to collaborative ways of working (Ramsay et al. 1993). In the USA, the Center for Social Organization of Schools has theorized an extended partnership framework including other family members and the wider community (Epstein et al. 1997). Implementation involves exploring opportunities for six types of engagement: *parenting* (assisting families with childrearing and establishing home environments conducive to learning); *communicating* (developing channels for reporting student progress and school news); *volunteering* (improving involvement of families as volunteers and audiences to support students); *learning at home* (assisting all parents to help children with homework and other curriculum-linked activities); *decision-making* (including families in school governance as well as Parent–Teacher Associations and other committees); *collaborating with the community* (linking students and families with other community agencies, and providing services to the community). This model promotes 'school-like families' and 'family-like schools': parents are expected to reinforce the importance of school, and schools to recognize each child's individuality.

It is recognized that program implementation is often problematic within educational systems that have traditionally operated in a very hierarchical manner. Research suggests that projects are relatively easily established when children are most dependent (of preschool age or with special educational needs) but still difficult to operate successfully in high schools, even though it is now strongly argued that parental participation remains crucial at this level. It is widely reported that training programs continue to pay insufficient attention to preparing teachers and administrators to work in new ways with families (including dealing with attendant challenges to their professional status), and that even the best-planned initiatives rarely fulfill all criteria of partnership or manage to attract all families.

2. Critique of the Partnership Model

Despite acknowledged implementation difficulties, the literature nonetheless generally supports the view that 'partnership' remains the ideal to be pursued. However, some repeatedly articulated misgivings critique this premise.

2.1 Research Support

Some writers contend that research has *not* provided compelling support for the core claim that the impact of parental involvement is positive, that too much of the corroborative evidence comes from studies of self-selected samples of schools and parents, and that assumptions about the *direction* of causal connections between involvement and children's success have been

insufficiently interrogated. It is also argued that researchers have failed to investigate seriously some of the potentially *negative* consequences, for example, the impact of highly engaged parents whose motivation may be rooted in an antischool stance, the 'dark side' to the intense involvement of some ambitious or 'pushy' middle-class parents, and the risks associated with seeing increased parental participation as the universal panacea for schools' pedagogical or financial problems. Possibly there has been a 'conspiracy' to avoid examining the new buzzword of partnership too closely in case it fell apart (Bastiani 1993).

2.2 Partnerships and Power Relations

Whereas partnership protagonists such as Epstein argue that parents are more concerned about 'information,' 'communication,' and 'participation' than about 'power' or 'control,' others claim the language of collaboration and partnership helps to silence the naming of important power asymmetries in home–school relations. 'Participation' may therefore mean that parents play a more active role, but are given no genuine opportunity to challenge the school's authority to define the nature of that role (Fine 1993). It may deflect attention from gender disparities, such as mothers' under-representation in positions of governance or schools' reported reluctance to utilize fathers in classroom support roles.

Furthermore, parents and teachers do not necessarily mean the same thing when they use terms such as 'helping children succeed.' Discussing the enduring documentation of socioeconomic class (SES) differences in parental involvement, Lareau (Lareau 1989, Lareau and Horvat 1999) suggests that these differences have several roots, including the fact that whereas middle-class parents typically embrace opportunities to scrutinize and question school practices, low-SES parents often prefer a relationship characterized by *separateness*, expecting teachers to do their job and seeing themselves as helping best by deferring to professional expertise. Her analysis also incorporates Bourdieu's (1990) notion of *cultural capital*: high-SES children enter schools with more cultural capital potentially available for activation because schools tend to use the linguistic structures, authority patterns, and types of curricula with which they are more familiar. There may be negative consequences for parents' dignity and authority in the home if they cannot (or prefer not to) perform their designated roles as educators in accordance with these terms (see *School Achievement: Cognitive and Motivational Determinants*). Low-SES parents may feel less confident or comfortable visiting schools, and may have more reason than high-SES parents to fear that closer links with schools could result in greater family surveillance by government agencies. Middle-class domination of that aspect of 'partnership' quintessentially distinguishing it from other models—involvement of parents in all aspects of school governance—may function to sustain rather than challenge societal inequalities. While some projects work hard to minimize such impediments, or avoid them through being established in more economically homogeneous communities, the generic Western model arguably privileges the ways in which *well-educated* parents wish to, and can, engage with schools.

Most of these concerns have also been addressed with reference to multiculturalism, neocolonialism, and racism (e.g., Ramsay et al. 1993, Limerick and Nielsen 1994, Lareau and Horvat 1999) (see *Minorities*; *Multicultural Education*). Whether or not imbalances of power or influence fundamentally violate the spirit of partnership remains a key feature of current debate.

2.3 Political Intervention

In Western societies, partnership models enjoy widespread political support, provoking debate about the extent to which schools should be legally mandated to develop partnership programs and parents required to participate. Legislation might be the only way to reduce significantly existing societal inequalities *or* it could equally well serve to consolidate or amplify them. Partnership initiatives may be compromised by the effects of other political interventions in education, such as imposition of a 'market model' that positions parents as selfish 'consumers' of school services. Moreover, no less than other paradigms, the partnership model should be recognized as offering national or local governments opportunities to absolve themselves from the responsibilities of maintaining struggling public school systems (Fine 1993). While the market model seeks to lay blame primarily at the door of schools, the 'family-is-school' approach could arguably enable school failure to be more easily read as *family* failure.

3. Future Trends

To some extent, future developments will be driven by research findings that resolve the implementation problems outlined above. They will also be influenced by how communities prioritize their objectives. While the pre-eminent objective may continue to be enhancement of parents' ability to promote their own children's school success (however defined), some communities or cultures may prefer a more collective approach that encourages all adults to work on behalf of all children. Other programs may focus less exclusively on children, and place more emphasis on helping parents improve their quality of life and on the broader rejuvenation of community spirit. If educational resources are limited, families may play a critical role in the development of curriculum

materials; in nations where the legacy of colonialism still positions schools as somewhat alien institutions, schemes may be focused on integrating schools more fully into the society. Whatever the objectives, a partnership ideology suggests that programs are most likely to succeed when goals are clearly articulated and informed consensus has been reached regarding the strategies employed to achieve them. As always, however, developments will be directly or indirectly influenced by external factors such as funding support and other aspects of government policy.

'Partnership' is clearly the current model of choice for white Western middle-class communities, and has many features with the potential to inform good educational practice in a wide variety of cultural settings. Nevertheless, certain conditions must be met in order that partnership programs offer all families an equitable chance of reaping the rewards of participation. Furthermore, inasmuch as technological advances may mean that children in future spend much less time either physically located in schools or pedagogically locked into the programs of individual education systems, current ideas about 'consumers,' 'providers,' and 'partnerships' may possibly soon require radical re-examination.

See also: Families as Educational Settings; Family and Schooling; Family as Institution; Family Systems in Europe; Parenting: Attitudes and Beliefs; School Effectiveness Research; School Management; Teacher Expertise; Teaching as a Profession

Bibliography

Bastiani J 1993 Parents as partners: Genuine progress or empty rhetoric? In: Munn P (ed.) *Parents and Schools: Customers, Managers or Partners?* Routledge, London, pp. 101–16

Berger E H 1987 *Parents as Partners in Education: The School and Home Working Together*, 2nd edn. Merrrill, Columbus, OH

Bourdieu P 1990 *In Other Words: Essays Towards a Reflexive Sociology*. Polity Press, Cambridge, UK

Epstein J L, Coates L, Salinas K C, Sanders M G, Simon B S 1997 *School, Family, and Community Partnerships: Your Handbook for Action*. Corwin Press, Thousand Oaks, CA

Fine M 1993 [Ap]parent involvement: Reflections on parents, power, and urban public schools. *Teachers College Record* **94**: 682–710

Lareau A 1989 *Home Advantage: Social Class and Parental Intervention in Elementary Education*. Falmer Press, London

Lareau A, Horvat E McN 1999 Moments of social inclusion and exclusion: Race, class, and cultural capital in family–school relationships. *Sociology of Education* **72**: 37–53

Limerick B, Nielsen H (eds.) 1994 *School and Community Relations: Participation, Policy and Practice*. Harcourt Brace, Sydney, Australia

Ramsay P, Hawk K, Harold B, Marriott R, Poskitt J 1993 *Developing Partnerships: Collaboration Between Teachers and Parents*. Learning Media, Ministry of Education, Wellington, New Zealand

M. A. Payne

Pareto, Vilfredo (1848–1923)

1. Life

Fritz Wilfried Pareto was born on July 15, 1848 in Paris, France, the son of an Italian political émigré, Raffaele Pareto (1812–82). Raffaele's father, Raffaele Benedetto Pareto (1768–1831), had been made a marchese di Parigi—or, as some have claimed, even a pair de France—by Napoleon, in recognition of his collaboration in erecting the Cisalpine Republic. The Pareto family was rooted in the trade bourgeoisie of Genoa, Italy, where it had been registered in the *Libro d'oro*.

In addition to her son, whose given name was Italianized and shortened to Vilfredo, Pareto's mother, Marie Métenier (1816–89), also had two daughters. Due to a political amnesty granted by the House of Savoy, probably as early as 1858, the Pareto family was eventually able to move back to Italy. Pareto's father started to work as a French instructor for Genoa's *Scuola di Marina*. Vilfredo attended the same school, the beginning of his Italian education. When, later, Pareto's father was offered a better position in Turin, Vilfredo followed him to the capital of the kingdom of Savoy. Here, in 1867, Vilfredo received his *Licenza in scienze matematiche e fisiche*.

He then enrolled at the *Scuola di Applicazione per Ingeneri*, where, in 1870, he received his doctorate. Much later, Pareto still considered his dissertation—on 'Principi fondamentali della teoria di elasticità dei corpi solidi'—important enough to include it in his *Collected Essays on Economic Theory*. He began his professional career in Florence, working for the *Centrale della S. A. delle Strade Ferrate Romane*. From there, he went on to a management position with the *Societá dell'Industria di Ferro*, traveling widely in Germany, Austria, France, England, and Switzerland.

His career left him with a sound knowledge of practical business matters. In 1882, the year his father died, Pareto chose professional retirement and settled in Florence. At around this time, he made an unsuccessful attempt at running for political office on a platform promoting his ideal of a liberal market economy. Pareto became resolved to dedicate his remaining life to his scientific pursuits. He joined Italy's oldest learned society, the highly respected *R. Accademia Economico-Agraria*. He also became a member of a second society, *Adamo Smith*, where he gave several lectures on scholarly topics and economic policy, which he also published. Moving in the upper-class circles of Florence—which then exerted a leading cultural influence throughout Italy—Pareto made a number of influential friends: the Lord Mayor Peruzzi and his wife; the politician S. Sonnino; P. Villari, who wrote a famous biography of Machiavelli; Colli, the author of *Pinocchio*; the anthropologist Mantegazza; such renowned philologists as D. Comparetti, A.

Franchetti, and A. Linaker, as well as several others. It was at this time that he married Alessandra (Dina) Bakounine (1860–1940), a Russian (no relation to the anarchist of the same name).

When Pareto met Italy's leading economist, Maffeo Pantaleoni, they soon became friends. Pantaleoni supported Pareto, as a successor had to be appointed for the Lausanne University chair held by Léon Walras, who had had to vacate his position due to his worsening mental illness. As an academic teacher, Pareto had considerable success. However, he soon asked to be relieved of his teaching duties. An unexpected inheritance set him free to concentrate on his preference for research and scholarship. Pareto's *Feste Giubilari* was published in 1917 to great critical acclaim. While this represented a personal triumph for him, Pareto continued to view himself as a victim of Italy's politicians and the *Camorra universitaria italiana*. When the Fascists took over Italy's government, various honors were offered to him; he either declined them explicitly or did not respond to them at all.

Only once did Pareto return to Italy and then only because he was forced to visit the Rome family record registration office in person. To get a divorce and be able to remarry, he needed his birth certificate. Dina, his first wife, had left him. In the summer of 1923, he married Jeanne (Jane) Régis (1879–1948), his second wife. He had met her through a newspaper advertisement in Paris. When Pareto died two months later, on August 23, 1923, he no longer possessed an Italian passport. To be able to marry Jane, he had been forced to turn in his passport at the Republic of Fiume. Conforming to his last wishes, he was buried in Céligny, his last place of residence, in a modest grave listing only the years of his birth and death and his *nom de guerre*.

2. Works

From his beginnings as a researcher, Pareto placed immense importance on the question of method. He had named his approach the logical-empirical method. In his *Manuale d'economia politica* (1994), first published in 1906, he devoted no less than 55 paragraphs to it. In his *Trattato di sociologia generale* (1983), that number rose to 144 paragraphs. 'My only intention,' Pareto wrote, 'is to describe facts and to draw logical conclusions from them' (Pareto 1983, 55). Facts were to be weighted according to relevance, not quantity. Facts based on experience were considered valid only 'as far as experience does not contradict observation' (Pareto 1983, 6).

Pareto considered this method applicable not only to the field of economics but to the social sciences as well and—implicitly or explicitly—to every field of scholarship (Pareto 1994, Vol. 1, 14). However, every inductive approach will always contain an implicit assumption. In the sciences, this would be the supposedly homogeneous character underlying all of nature. Thus, Pareto's 'survey' is strongly similar to J. A. Schumpeter's concept of 'vision.' According to Pareto, it is quite feasible to differentiate between economic and social phenomena initially. To regard them separately does not violate the dictates of empiricism. But 'both parts then have to be integrated with each other. Only a full view of the whole thing possesses any empirical reality' (Pareto 1994).

At the heart of Pareto's *Cours d'économie politique* lies his theory of economic equilibrium. He refers to two sets of mathematical equations to describe two 'economic systems.' One system applies to barter and exchange, the other to production. If combined, both sets of equations attain validity only in a state of equilibrium (Pareto 1994, App. 131–3). 'The most interesting case,' according to Pareto, is described as a situation where the number of equations is less than the unknowns contained in the equations: 'The system is then flexible and can balance itself' (Pareto 1994, App. 131–3). 'Mathematics,' Pareto insisted, 'should be used only if it allows for an expression of certain relations between facts which otherwise could not have been expressed or which could not have been formulated using standard language' (Pareto, 559, n. 4, cf. Pareto 1983, 1732).

To describe different states of economic equilibrium, Pareto coined two terms: ophelemité and utilité. When Irving Fisher wrote a review of Pareto's work, Pareto tried in vain to discourage Fisher from confusing his terms with the terminology of Walras and Edgeworth. Fisher, Pareto argued, did not understand enough French and even misunderstood some of Pareto's mathematics (Pareto 1960, Vol. 3, p. 408).

In 1902–3, Pareto published the two volumes of *Les systèmes socialistes*, an incisive analysis of historical and present-day socialist theories, including a discussion of Werner Sombart (he had treated Karl Blücher in a similar way in his earlier *Cours*). He paid due reverence to Karl Marx as a sociologist. Economically, however, he did not give Marx much credit. Theoretically, a socialist economy might be able to solve its problems as well as any nonsocialist economy. Earlier, Pareto had already integrated socialist society and collective capital with his overall theory of economics (Pareto, Vol. 2, 714–24). Indeed, his followers, E. Barone and O. Lange, later based their own socialist theories of economics on this insight. Pareto argued against socialism not on economical but on social and anthropological grounds. He concluded that all socialist systems were eventually based on untenable social and anthropological contradictions.

In his *Manuale d'economia politica* (1994), Pareto presented a reworked version of his theory of economics with special emphasis on the various states of economic equilibrium. Typically, he differentiated between them in his famous dictum:

> We can say that the members of a community, in a certain situation, enjoy the optimum of ophelemité if it has become impossible to distance oneself from this situation even to a small degree without benefit or damage to all members of this community; even the smallest step then necessary leads to profit for some members of the community or loss for others (Pareto 1994, 33).

This concept has been named the Pareto optimum. Later, it served as the foundation for welfare economics. It led to several attempts at an improved theory of income distribution.

Pareto always insisted that to understand his *Manuale*, one needed to know his *Cours*. In a typical misreading, J. R. Hicks complained that in the *Manuale*—by now a classic on the theory of consumption—the problems of capital and interest were hardly noted (Hicks 1950), ignoring their extensive treatment in the *Cours*.

It was Pareto's intention to show 'by examples rather than declarations to prove the relation between economic phenomena and other phenomena of society' (Pareto 1994, Foreword). Thus, it becomes clear how deplorable was the omission, in the French edition of the *Manuale* (Pareto 1981) of Pareto's Foreword or *Proemio*. Still, improvements had been made to other parts of the French edition, including Pareto's mathematics. He was of the opinion that 'pure' theory was applicable only to the flow of goods and trade within a 'purely' competitive economy. Yet, the reality of monopolies, cartels, and trusts asked for a different description of modern markets and their mechanisms. Eventually, he discarded his entire theory of economic equilibrium as a 'dead end,' in spite of his groundbreaking anticipation of what later essentially came to be called the theory of monopolistic competition (cf. the French edition of the *Manuale*, Vol. 3, 40–8, 160–64, Vol. 5, 8–9, 71; App. 141). With reference to the effects of economical cycles, Pareto already differentiated between their objective and subjective sides. In this, he could claim no less a predecessor than Montesquieu.

Pareto's final great work, the *Trattato*, was preceded by the brief *Mythe vertuiste et la littérature immorale* (Pareto 1911), an unforgiving settling of scores with any kind of literary censorship. Taken to its logical conclusion, Pareto argued, censorship would have to call for all of world-class literature to be censored—starting with the Bible.

In the two volumes of his *Trattato di sociologia generale* (1983), Pareto undertook the immense task of integrating his economic and social analyses. It had long been clear to Pareto that for any model of economic action fully to grasp reality it had to encompass a theory of society. Next to those actions described as logical, recognizably illogical actions had to be equally accounted for. Such actions called illogical by Pareto are regularly cloaked by pseudo-logical explanations. On this insight, Pareto erected an entire sociology of knowledge (cf. Eisermann 1987 p. 170ff.). Logical actions, to Pareto, were not synonymous with rational actions, however. For any act to be described as rational, it would have been necessary actually to look into the mind of the social actor. Pareto did not think such introspection was feasible. Rather, he conceptualized such actions as residual social facts. Thus, Pareto arrived at an inventory of six residues, in which instinct of combinations and persistence of aggregates were the most important ones.

He distinguished between two groups of economic actors: 'rentiers' and 'speculators.' Rentiers in general are secretive, cautious, timid souls, mistrustful of all adventure, not only of dangerous ventures but also of any venture to have even the remotest semblance of not being altogether safe (Pareto 1983, 2313). Pareto's rentiers are ruled by aggregates of persistence and his speculators by their instinct of combination. Speculators

> are usually expansive personalities, ready to take up with anything new, eager for economic activity. They rejoice in dangerous economic ventures and are in the watch for them, but they work underground and know how to win and hold the substance of power, leaving the outward forms to others (Pareto 1983, 2313).

Accordingly, Pareto expanded his analysis of the economic equilibrium into an analysis of the social equilibrium (Pareto 1983, 2067–87). The crucial movement inside the system of society is construed as *circolazione delle classe scelte*, class circulation. Society is then ruled by fractions, or minorities, which are subject to a continuous turnover, or exchange (2025–46 and throughout).

Pareto always insisted that his 'pure' theory simply was an integral part of his overall economics, and that his economics necessarily formed part of his sociology. He claimed to have introduced the principle of relativity to the social sciences. He took equal pride in his anticipatory powers of foresight. Like his contemporary J. M. Keynes, he foresaw the catastrophic repercussions of the Treaty of Versailles—for purely economic reasons. As early as 1917, he asked, 'whether the Pax americana will turn out to be as productive as the onetime Pax romana?' (Pareto 1960, Vol. 3, p. 240).

See also: Economic Sociology; Economics, History of; Elites: Sociological Aspects; Mobility: Social; Political Sociology; Rationality in Society; Sociology, History of; Sociology: Overview; Welfare

Bibliography

Bobbio N 1969 *Saggi sulla scienza politica in Italia*. Laterza, Bari, Italy

Busino G 1989 *L'Italia di Vilfredo Pareto*, 2 Vols. Banca commerciale italiana, Milan

Convegno Internazionale V. P. 1975 Accademia Nazionale dei Lincei. Rome
Demaria G 1949 L'oeuvre économique de Vilfredo Pareto. *Revue de l'économique politique* **59**: 517–544
Eisermann G 1961 *Vilfredo Pareto als Nationalökonom und Soziologe*. Mohr, Tübingen, Germany
Eisermann G 1987 *Vilfredo Pareto*. Mohr, Tübingen, Germany
Eisermann G 1989 *Max Weber und Vilfredo Pareto*, Dialog und Konfrontation. Mohr, Tübingen, Germany
Eisermann G, Malinvaud E 1992 *Vilfredo Pareto und sein 'Manuale.'* Verlag Wiltschaft und Finanzen, Düsseldorf, Germany
Freund J 1974 *Pareto*. Seghers, Paris
Hicks 1950 *Value and Capital*, 3rd edn. Oxford University Press
Kirman A P 1987 Pareto as economist. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave* Vol. 3, Macmillan Press, London, pp. 804–9
Pareto V 1911 *Le mythe vertuiste et la littérature inmorale*. Rivière, Paris
Pareto V 1920 *I sistemi socialistici*, Turin, Italy
Pareto V 1952 *Scritti teorici*. Edizioni di storia e letteratura Milan
Pareto V 1960 *Lettere a Maffeo Pantaleoni*, 3 Vols. Edizioni di storia e letteratura Rome
Pareto V 1963 *The Mind and Society*, 2nd edn. Dover, New York
Pareto V 1964–89 *Oeuvres complétes*, 30 Vols. Librarie Droz, Geneva, Switzerland
Pareto V 1965 *Croncache italiane*. Morcelliana, Brescia, Italy
Pareto V 1966 *Scritti sociologici*. UTET Turin, Italy
Pareto V 1981 *Manuel d'économie politique*. Librarie Droz, Geneva
Pareto V 1920 *I sistemi socialistici*. Turin, Italy
Pareto V 1983 *Trattato di sociologia generale*, 4th edn. AMS Press, New York
Pareto V 1994 *Manuale dieconomia politica*, 4th edn. Tesi, Pordenone, Italy
Parsons T 1968 Vilfredo Pareto. In: Sills D I (ed.) *International Encyclopedia of the Social Sciences*. Vol. 11, Macmillan, pp. 411–16
Schneider E 1961 Vilfredo Pareto in the light of his letters to M. Pantaleoni, *Banca Naz. Lavoro Quarterly Review* **58**
Schumpeer J 1951 Vilfredo Pareto. In: *Ten Great Economists*. Oxford University Press, New York
Valade B 1990 *Pareto*. Presses Universitaires de France, Paris

G. Eisermann

Parietal Lobe

The parietal lobe is the part of the cerebral cortex lying between the occipital and frontal lobes, and above the temporal lobe (see Fig. 1). It is named for its proximity to the parietal bone, the wall of the skull (from the Latin *paries*, wall). It has five major functions: the primary analysis of somatic sensation (touch, the position of the limbs, temperature); the analysis of space using all sensory modalities; the specification of spatial targets for the motor system; the generation of attention; and the analysis of visual motion. Our first understanding of parietal function arose from the analysis of the deficits of patients with lesions of the parietal lobe (Critchley 1953). Subsequent insight into the mechanisms of parietal function came from neuroanatomical and neurophysiological studies.

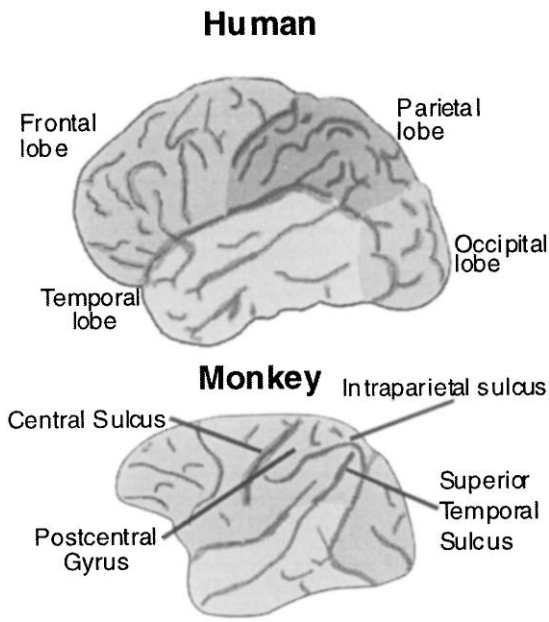

Figure 1
The four major lobes of the human and Rhesus monkey brains. The lobes are labeled in the human brain. In both the monkey and the human brain, the surface area of the cerebral cortex is so great that the cortical surface must be folded so that some tissue is on the gyral surface, for example the postcentral gyrus, and other tissue is hidden deep inside clefts called sulci. The major sulci of the parietal lobe are labeled in the monkey brain, as well as the postcentral gyrus. The lobes in the monkey brain have the same shading as the lobes in the human brain

The importance of the parietal lobe was apparent to neurologists such as Hughlings Jackson, Dejerine, Head, and Holmes early in the twentieth century. They discovered that humans with parietal lesions had difficulty with tactile perception on the side of the body opposite the lesion (the contralateral side): they were not completely unable to feel stimuli, as they might be after a lesion of the spinal cord, but they had relative deficits: they might be able to feel touch, but not pain or temperature; they might perceive touch but have difficulty localizing the stimulus; they might have difficulty knowing the position of a limb in space; they might have difficulty identifying an object placed

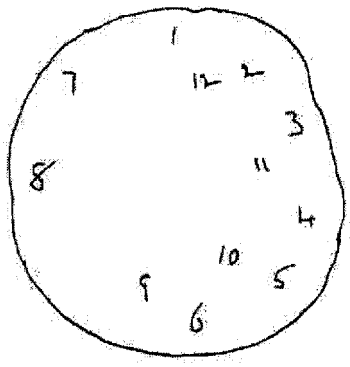

Figure 2
Drawing of clock by a patient with a right parietal lesion. Note that the numbers are displaced to the right, as if the patient is neglecting the left side of the drawing, and that the relationships of the numbers to each other are also disturbed, for example the 10 is to the right of the 9 (reproduced with permission of M. Critchley from *The Parietal Lobes*, 1953)

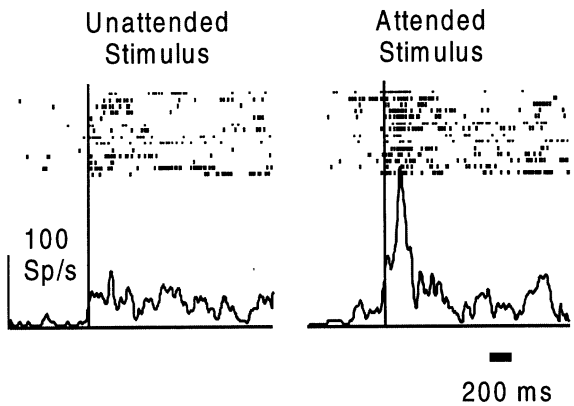

Figure 3
Enhancement of the response of a neuron in the lateral intraparietal area to an attended stimulus. The upper display is a raster diagram: in this diagram each dot represents a single action potential. Each line is a single presentation of the stimulus, which is accomplished by having the monkey make a saccade that brings the stimulus to the proper part of the retina. The graph underneath the raster display is the average firing rate of the neuron as function of time. The vertical line represents when the stimulus arrives in the part of the retina to which the cell responds. The cell responds immediately after the saccade. When the stimulus has suddenly appeared 400 ms before the eye movement, and has the attentional effect of an abruptly appearing object, the response of the cell to the appearance of the stimulus at the proper part of the retina (attended stimulus, left) is much greater than the response to the stimulus when it has been present in the environment for a long time, and is therefore not attention worthy (unattended stimulus, right) (adapted from Kusunoki et al. 2000)

in their hand without looking at it—for example, they would not be able to reach into a pocket and pull out a key, leaving coins behind. Stimuli were often perceived to be less strong on the affected side than on the normal side of the body, and comparative discriminations were coarser. Sometimes patients who can appreciate a stimulus on an affected limb when it is the only stimulus touching them cannot appreciate it when two stimuli are applied simultaneously, one on the affected limb and the other on the normal limb. The most dramatic parietal deficit is anosognosia—patients, especially those with right parietal deficits, are entirely unaware of the limbs on the contralateral side: they think that a leg in bed with them belongs to someone else.

Patients with parietal lesions do not have deficits limited to the tactile or somatosensory domain: they have striking visual deficits as well. Patients with right parietal damage tend to neglect the left visual field, the left side of space relative to the midline of their bodies, and the left side of objects (see Fig. 2). When they draw a clock they cram most of the numbers on to the right side of the clock, neglecting the left; when they attempt to bisect a line they place the midline near the right end of the line rather than the middle. After this florid neglect has recovered, right parietal patients often still neglect a stimulus in the affected visual field when another one is presented simultaneously in the normal visual field. This deficit is interpreted as a deficit in visual attention: patients with parietal lesions are less efficient at shifting their attention into the affected field than into the normal field.

Patients with parietal lesions often exhibit optic ataxia, the inability to coordinate their movements using visual cues. They have difficulty knowing where objects are relative to their body. They misreach for objects, and make inaccurate eye movements toward them. They also have difficulty transmitting visual information to the motor system. When humans reach for an object, they open their hand to the approximate width of the object before their hand arrives at the object. Thus, people reaching for a can of soda will open their hands wide to accommodate the can; people reaching for a pencil will open theirs narrowly, just to accommodate the pencil. Patients with parietal lesions cannot do this, but reach for all objects with a stereotyped hand posture. The brain mechanisms that drive limb movement have no access to the detailed visual aspects of a target, even though they might know the spatial location of the target. A remarkable aspect of parietal deficits is that, although patients do not know where an object is, they have no difficulty telling what that object is—they know, for example, that the examiner is holding a wristwatch in front of

them, but cannot find it to reach for it. They exhibit constructional apraxia—when asked to duplicate a block design they are unable to place the blocks in proper relationship to each other, but instead leave them in an unordered jumble even though they choose the proper blocks. Patients with lesions of the parietal cortex on both sides of the brain have difficulty shifting their attention away from the current object. The result of this deficit is that they only describe one in the world at a time—they may see a dog's nose but not its ears, for example. They cannot use eye movements to explore the visual world. Patients with small lesions of the parietal cortex occasionally exhibit a specific deficit for the analysis of visual motion and the transmission of this information to the motor system. They have deficits both with the motor aspects of visual motion, for example, they cannot follow a moving object with their eyes; and they have difficulty with the perception of motion—knowing the speed of an object, or judging the predominant motion of a moving random dot pattern.

Space is not represented in a unitary fashion in the parietal cortex. Instead it seems to be organized in concentric spheres, or workspaces, that correspond to the range of action of certain motor systems. Patients with small lesions of the parietal cortex may have these different workspaces affected differentially. Thus a patient who can localize a nearby stimulus by reaching out and touching it with an arm may not be able to localize a stimulus further away in the same direction by touching it with the beam of a laser pointer. These are not merely deficits in movement—the patients show perceptual distortion, such as an inability to bisect a line, in the affected workspace.

Recent studies in neuroanatomy and neurophysiology, especially those done with the Rhesus monkey, have been able to illuminate the brain mechanisms whose destruction causes the dramatic clinical symptoms of the parietal lobe. The brain is a network of information-processing cells called neurons. Each neuron serves as a node in this network, receiving information from other neurons, often in different areas of the brain, and transmitting this information to other neurons. By studying the cytology of the parietal lobe and its connections with other parts of the brain, neuroanatomists have been able to divide it into a number of functional subdivisions. Neurons transmit information by emitting a series of electrical impulses called action potentials. If a neuron responds to a stimulus it issues a burst of impulses in response to the appearance of the stimulus. By monitoring the electrical activity of neurons in the parietal lobe of monkeys that are awake and performing behavioral tasks, neurophysiologists have begun to understand the signal processing that enables the parietal lobe to perform its function in the generation of human behavior.

Physiological and anatomical studies have subdivided the parietal lobe into a number of distinct areas which perform the various different functions whose deficits appear in patients with parietal lesions. The most anterior part of the parietal lobe is the primary cortical area for somatic sensation. Most of the posterior parietal cortex, starting with the posterolateral bank of the intraparietal sulcus is concerned with space and attention. The most posterior parietal area, in the superior temporal sulcus, is important in the analysis of visual motion.

The primary somatosensory cortex consists of four subdivisions, cytoarchitectonic areas 3a and 3b, located in the posterior bank of the central sulcus, and areas 1 and 2 located on the surface of the postcentral gyrus, the brain surface between the central and intraparietal sulci. Each of these subdivisions receives projections from the somatosensory parts of the thalamus, the way station between the spinal cord and brainstem, and the cortex. Areas 3a and 2 are responsible for deep bone and muscle sensation, including the muscle spindles that measure muscle length, and joint sensors that describe limb position. Areas 3b and 1 are responsible for sensation from the skin. These areas project backward to area 5, on the postcentral gyrus and the anterior bank of the intraparietal sulcus. This area integrates the various primary sensory signals, and transmits somatosensory signals to the premotor cortex so the information can be used for not only the perception of somatic sensation but also the coordination of movement. Single neurons in the anterior parietal cortex describe the properties of tactile stimuli in a very precise manner: for example, Mountcastle showed that neurons responsive to vibration describe the intensity and frequency of that vibration so accurately that a computer could construct the nature of the stimulus by decoding the neural pulse train.

The posterior parietal cortex is subdivided into a number of areas. Those located in the intraparietal sulcus associate somatic sensation, vision, and audition for the analysis of space, and the description of this world for action. Many of the areas in the intraparietal sulcus describe particular workspaces (Colby and Goldberg 1999). The physiological properties of these areas have been studied extensively in the Rhesus monkey.

Perhaps the best-studied area of the monkey parietal cortex is the lateral intraparietal area (LIP). This area describes distant space, the space that primates explore with their eyes. It projects to the frontal eye fields and the intermediate layers of the superior colliculus, brain areas known to be important in the generation of the rapid eye movements (saccades) that human and nonhuman primates use to explore their environment. It also projects to areas in the temporal lobe, TE and TEO in the Von Bonin and Bailey nomenclature, which are important in the analysis of pattern. An important connection between the generation of saccadic eye movements and the analysis of the visual world is the process of attention, the selection of

objects in the world for special analysis. In general, human and nonhuman primates pay attention to the targets for their eye movements, and saccades and attention are closely linked, but they can be dissociated in certain cases—for example juvenile Rhesus monkeys certainly pay attention to the alpha male of their troop, but they never look directly at him. Animals attend to stimuli either because the stimulus is relevant to the current task, or because the stimulus is naturally attention worthy, stimuli that are, in the words of William James, 'big things, bright things, moving things, or … blood' (James 1890). For example, stimuli that appear abruptly in the visual field inevitably draw visual attention.

Neurons in LIP respond both to saccade targets and to attended stimuli that are not necessarily the targets for saccades. These neurons respond to visual stimuli, and their activity is modulated by whether the animal is paying attention to the stimulus. Neurons in LIP respond more to attended than to nonattended stimuli, independently of whether those stimuli are attended to because of their importance to the task or because they are in the class of stimuli that inevitably draw attention. Presumably it is the absence of this attentional modulation system that impairs the ability of parietal patients to attend to stimuli in the world contralateral to the lesion. Neurons in LIP also respond to sounds when the monkey attends to them (Anderson et al. 1997). A classic task in which to study the activity of neurons is the delayed saccade task.

In this task a monkey looks at a spot of light and another one flashes briefly in the visual field. A few seconds later the original fixation point disappears, and the monkey makes a saccade to the spatial location of the now vanished saccade target. Neurons in LIP respond to the stimulus, and then continue responding even though the stimulus has disappeared. This activity has been interpreted as either maintaining the memory of the saccade target, or focusing attention at the saccade target. It is probable that this area of the brain supports both functions: focusing attention on spatial locations, and providing the oculomotor system with information about the location of potential saccade targets. LIP serves as a model for the other, less well studied areas of the parietal lobe: neurons specify possible targets for the action appropriate for the workspace, as well as responding to attended objects in that workspace.

The most posterior part of the parietal cortex of the monkey lies in the superior temporal sulcus. This area is important for the analysis of visual motion. Neurons in the middle temporal area and the middle superior temporal area are specialized for the analysis of visual motion, and respond selectively to stimuli moving with certain speeds and direction. Electrical stimulation of these areas can affect monkeys' perception of motion, and lesions in these areas render monkeys unable to follow moving stimuli with their eyes, or to make judgments about the quality of visual motion.

See also: Navigation in Spatial Environments, Neural Basis of; Perception of Extrapersonal Space: Psychological and Neural Aspects; Somatosensation; Topographic Maps in the Brain; Visual System in the Brain

Bibliography

Andersen R A, Snyder L H, Bradley D C, Xing J 1997 Multimodel representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* **20**: 303–30

Colby C L, Goldberg M E 1999 Space and attention in parietal cortex. *Annual Review of Neuroscience* **2**: 319–49

Critchley M 1953 *The Parietal Lobes*. Hafner, New York

James W 1890 *The Principles of Psychology*. Holt, New York

Kusunoki M, Gottlieb J, Goldberg M E 2000 The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus notion and task relevance. *Vision Research* **40**: 1459–68

M. E. Goldberg

Parliamentary Government

1. Development of Parliamentary Regimes and Democracy

Parliamentary democracy is the product of the twentieth century and in many countries developed only after 1918. There were some attempts to create such a system during the European revolutions of 1848 without leading to consolidated systems. What Huntington has called the first long wave of democratization (1828–1926) in the light of research on parliamentary systems was only parliamentarization. Parliamentary government is characterized by a government *responsible to the majority in parliament*. Democracy was added to the parliamentary regimes only later. The minimal criterion of democracy is *universal suffrage*. In most countries with a regime of constitutional monarchy, such as Britain, Scandinavia, or the Benelux countries, democracy was introduced later than parliamentary regime. Only in some rather authoritarian monarchies was democratic universal suffrage introduced or preserved without granting a parliamentary regime with responsible government (France 1851 in the Second Empire, Germany 1871 in the Empire of Bismarck) (see Table 1).

Only in a few countries (France, Finland, Sweden) was the process of parliamentarization and democratization synchronized. A continuously democratic parliamentary regime existed only in France. Even Britain—the mother of 'parliamentary government'—in the nineteenth century could not be called a

Table 1
Parliamentary democracy: the merger of parliamentarization and democratization

	Continuity of a system of estates transformed into a constitutional regime							New regimes			Parliamentarization of neo-absolutist regimes					
	UK	Sweden	Norway	Finland	The Netherlands	Luxembourg	Belgium	France	Italy-Piedmont	Spain	Germany	Austria	Denmark
General male suffrage	1918	1921	1898	1906	1917	1919	1919	1848	1919	1869	1871	1907	1915
Time of creation of a government responsible to parliament	1688ff 1832–35 final conflict	1917	1884	1906	1868	1868ff	1833ff	1875 1821ff 1837 1848–51 1871ff	1860	1907 1874ff	1919 1918	1918	1901

'parliamentary democracy' (cf. Table 1). There was a divergence of *parliamentarization* (mostly implemented by liberal majorities which were reluctant to add universal suffrage to the system) and *democratization* (frequently introduced under the pressure of radicals and socialists). These groups were initially unwilling to accept the checks and balance of the liberal parliamentary regime and advocated a democratic regime with many features of plebiscitarian instead of representative democracy. Only when they came to power did they slowly accept the rules of the game which were mostly developed in former constitutional monarchies.

Democratization in some countries has ruined the parliamentary system (Italy, Austria, Germany, Spain) when counter-regime forces gained power. In other established parliamentary systems the regime in the 1930s came to the brink of a collapse (Belgium, France) under the pressure of right-wing extremism. Only after 1945 and the disastrous experiences with authoritarian or even totalitarian dictatorship, were parliamentary government and democracy combined in peaceful co-existence. Some countries (such as the Federal Republic of Germany) even then, however, showed a tendency to emphasize the legal state and *representative* parliamentary *government* much stronger than *democratic participation*. The theory of constitutional and representative government did not easily accept parliamentary democracy but interpreted the preponderance of the parliamentary majority in terms of a rather mechanistic system of checks and balances and 'mixed government.' Because responsible cabinet government developed first in Britain and Belgium, the jurists for quite a time used the term '*Coburg government style*.' Only since the 1830s was the term 'parliamentary government' accepted in Britain and France and the regime was recognized not only as a mode of governance but as a *regime sui generis* which made the old typologies of regime types according to the number of rulers rather obsolete.

2. *Types and Criteria of Parliamentary Democracy*

There are many types of parliamentary government, but most of them have developed common *institutional* criteria.

(a) *Compatibility of parliamentary mandate and ministerial office* (exceptions are the French Fifth Republic, Luxembourg and The Netherlands) in order to guarantee close cooperation between the parliamentary majority and the executive.

(b) Prime ministers are normally *members of parliament*.

(c) The government has to have the confidence of the parliamentary majority. *Ministerial responsibility* is not defined in terms of legal responsibility as in

dualistic constitutional monarchies. The government has to resign in the event of a vote of no-confidence unless it chooses to ask the head of the state to dissolve parliament in order to give the electorate the opportunity to resolve the conflict.

(d) Parliament controls the *government by raising* questions, exercises the right to interpellate and set up committees of enquiry, which facilitates the decision about whether the strongest sanction or a vote of no-confidence should be used. This right is shared by dualistic presidential systems, apart from the possibility of toppling the head of government and his or her cabinet.

(e) Some analysts have postulated as an essential feature a formal vote of the government at its first meeting with parliament (*investiture*), as in the French system under the Third and Fourth Republic and in Italy, or even demanded the *formal election of the prime minister*, which was developed in the German Basic Law of 1949. Most parliamentary regimes do not accept this condition.

(f) More widespread is acceptance that *parliament* should be *dissolved* if the prime minister has lost the confidence of the parliamentary majority. Sometimes this has not been instituted, for example in Norway and Israel. In other cases the legal possibility of dissolution has become inapplicable because it has been abused by unparliamentary presidents, as in the Third French Republic (when President MacMahon dissolved the Chamber of Deputies in 1877).

Besides the minimal institutional criteria, certain social structural features are essential for the consolidation of parliamentary government:

(a) *Organized parties* to facilitate the building of parliamentary majorities.

(b) Party-building to facilitate the development of *cabinet solidarity*.

(c) Development of the *office of prime minister*. In early systems where the estates dominated, and even in certain absolutist monarchies, the power centers—the estates or the crown—impeded the activities of first minister. A certain *hierarchization of the ministerial council* also strengthens cabinet solidarity.

(d) Development of a *political culture* favorable to appropriate parliamentary behavior and alternating government.

Not all these institutional criteria can be found in every type of parliamentary system. Therefore subtypes have been proposed. Most ideological is the distinction between *authentic British parliamentarism* and the allegedly *inauthentic parliamentary system of the French type*—close to revolutionary government. Latecomers to parliamentarization, such as Germany (1918) and Sweden (1917), have often denounced the dangers of the French type, especially the German school of Redslob.

Oddly enough, France itself has reacted strongly against the shortcomings of the parliamentary system of the Third and Fourth Republics. De Gaulle's constitutional ideas, close to those of Carré de Malberg and later developed by Michel Debré were first pronounced in 1946 in the famous speech of Bayeux but failed to obtain majority when the Fourth French Republic was developed in 1946/47. But De Gaulle got his revenge when the Fourth Republic collapsed and he was able to shape the institutions of the Fifth Republic according to his ideas. A new system was born that since Duverger has been called the *semipresidential system*.

In spite of its name this was clearly a subtype of parliamentary government though the double responsibility before parliament and the head of the state—which existed in constitutional monarchies such as that of Louis Philippe—re-emerged. But in the long run the dominance of parliament in executive-legislative relations is re-established. De Gaulle's new system was not new at all: it had developed as a Republican type of constitutional regime from 1848 (the Second French Republic, 1848–51) and re-emerged in the Weimar Republic (1919) and Finland. Semipresidential systems are not a mixed type, not even the 'three-quarter presidential system' that Yeltsin tried to develop. The backlash of a humiliated legislature was felt in Russia in 1998 when Yeltsin was forced to renounce Chernomyrdin as prime minister and accept the parliamentary candidate Primakov.

3. *Functions and Performance in Parliamentary Democracy*

Parliaments were more powerful in times of parliamentary sovereignty without universal manhood suffrage than in modern parliamentary democracies. The main actor in parliamentary democracy is no longer parliament or government, but the two powers are linked together. This linkage has strengthened the prime minister in most parliamentary regimes, unless high party fragmentation and consociational practices of conflict resolution have weakened the head of the government. Parliamentary democracy seems to be misnamed because of much discussion on the 'decline of parliament.' Therefore, *cabinet government*, *prime ministerial government*, *or chancellor's democracy* became more popular to characterize the bigger parliamentary regimes.

Parliamentary systems have to perform various functions. There is no general decline of parliament but rather a change of the main functions of the system. *Parliamentary control* is weakened in systems with close relations of parliament and government, interlaced by party government. The *controlling functions* declined or shifted to other institutions from constitutional courts and courts of account to the mass media. *Representation* and *articulation of interests* in the systems has changed over time. The *legislative function* is still important because more

social problems are regulated in modern democracies than in old constitutional regimes. The *recruiting function* of the system, however, is strengthened after democratization of the parliamentary regimes.

Parliamentarians have become less removed from their voters in terms of *social structure*. What has been called *virtual representation* in early parliaments for those not eligible to vote in *postparliamentarism* is true of the social ties between deputies and voters. But people have adapted to this development and demand less strongly to be represented by members of the same region, the same sex, the same profession, or the same social class. The influence of voters on the selection of representatives is still mediated by the parties and their formal members. Primary elections can undermine the selecting function of the parties among other disadvantages, such as a lower voter turnout in the general election. Modern democracies are not ruled according to President Lincoln's Gettysburg formula of government *of* the people, *by* the people, and *for* the people. The distance of the representatives is compensated for by increasing *responsiveness* and sensibility to the wishes of the voters. Government *for* the people in the modern democracy of fluctuation of mass public moods has increased rather than decreased.

Nonetheless the departliamentarization of democracy is lamented. This has a *domestic political* side. At the stage of policy formulation, the practice of expert committees working together with the government is growing stronger. In the decision-making arena, the government is the most important initiator of legislation. Informal initiatives behind the formal initiatives of advisory boards, constitutional courts, or international organizations are further undermining the part played by parliamentarians in legislation. The penetration of interest groups into the work of ministers and parliament has shifted the preparatary work to a round tables of extraparliamentary committees and working teams. But parliament is not just a rubber stamp to ratify the decisions of networks outside parliament. Parliaments, formally the institutional seat of the people's sovereignty, represent the framework for the coordination of networks of actors from parliamentarians, party strategists, interest groups, ministry officials and in federate states the federate units. The cosy triangles of early American parliamentary studies, composed of parliamentarians, bureaucrats, and interest groups, have become uncosy squares, since party steering was rediscovered in the 'legislative Leviathan.' In federal states such as Germany, deparliamentarization at the Länder level has led to greater participation of the Länder in the national decision-making process via intergovernmental decision-making. This further actor has been added, making an uncosy pentagon. The Maastricht Treaty has increased the number of areas that are subject to EU rules and legislation. With growing competence at the EU level, lobbying at that level is likewise growing. Meanwhile, national parliaments—more than the executive—have been downgraded to the position of lobbyists. Paradoxically the *democratization* of the European Union has led to *deparliamentarization* at the national level. Until 1979, Members of the European Parliament were tied to the national parliaments, but since the introduction of direct elections this has no longer been the case. However, the European Parliament does not yet pose a real threat to the national parliaments as it can only draw up laws when asked to do so by the Commission. Formally, EU legal acts are passed by the Council of Ministers and not by the European Parliament.

There are three threats to parliamentary democracy:

(a) In the Process of *regionalization*, regional assemblies will try to strengthen themselves via regional politics at the cost of national parliaments.

(b) In the process of *Europeanization*, more and more domains of the national parliaments will be regulated by the EU. This process will be supported by the activities of the European Court and the gradual standardization of the national legal systems.

(c) In the process of *globalization*, a consolation at the national level is the fact that worldwide organizations such as the World Trade Organisation and GATT are starting to retaliate to the EU for what it did to the national decision-makers because the room for maneuver of the European organizations is increasingly limited by the global level, although still strongly limited in some arenas.

See also: Democracy; Democracy, History of; Democratic Theory; European Integration; Globalization: Political Aspects; Parliaments, History of; Political Culture; Political Representation; Regional Government; Representation: History of the Problem

Bibliography

Blondel J, Müller-Rommel F (eds.) 1988 *Cabinets in Western Europe*. Macmillan, Basingstoke, UK

Budge I, Keman H J 1995 *Parties and Democracies*. Oxford University Press, Oxford, UK

Döring H (ed.) 1995 *Parliaments and Majority Rule in Western Europe*. St. Martin's Press, New York

Laver M, Shepsle K A (eds.) 1994 *Cabinet Ministers and Parliamentary Government*. Cambridge University Press, Cambridge, UK

Norton P 1993 *Does Parliament Matter?* Harvester Wheatsheaf, New York

Schütt-Wetschky E 1984 *Grundtypen Parlamentarischer Demokratie*. K. Alber, Freiburg, Germany

Shugart M S, Carey J M 1992 *Presidents and Assemblies*. Cambridge University Press, Cambridge, UK

Strom K 1990 *Minority Government and Majority Rule*. Cambridge University Press, Cambridge, UK

von Beyme K 2000 *Parliamentary Democracy. Democratization, Destabilization, Reconsolidation 1789–1999*. Macmillan, Basingstoke, UK

K. von Beyme

Parliaments, History of

Nowadays, a Parliament is a central institution of the democratic state under the rule of law. The term parliament comes from French and describes a forum in which social understandings are reached by verbal negotiation, through the act of speaking. A Parliament is a representative institution, elected by direct universal suffrage. It plays an indispensable role in the workings of the State, as an active participant in the coalescing of political will, in public management, and essentially through conduct of the legislative process and concentration of the ruling elite. The decisive function of policy legitimization performed by parliaments is central to representative democracy which is the predominant system of political organization in contemporary states. Experience of direct democracy, which dispenses with the representative parliamentary system or acts in parallel to it, is generally limited to referendums, Switzerland being among the few countries that practice direct democracy on a regular basis. A Parliament is an assembly of public delegates who receive a mandate for a limited period, are voted to office through elections based upon a pluralistic political-party system, and serve as a channel for the expression of the opinions of society.

The approaches of political scientists and historians to the study of parliaments are closely related. Most contemporary researchers apply the insights of both disciplines in their efforts to analyze and explain the institutional structures and workings of parliaments. Comparative studies of the trajectory of societies, of the formation of states, and of the composition of their social and economic groups are the components of research on representation, participation, and decision-making processes conducted through and by means of parliaments. The achievement that parliamentary institutions represent, in terms of the development of democracy within society throughout the nineteenth and twentieth centuries, provides the social sciences with ample scope for research. Political science, sociology, and history continually delve into the origins, composition, work methods, social relevance, political prominence, and social roles of parliaments.

1. *Functions and Types*

The functions of parliaments have differed according to the system of government. Various terms may designate a parliamentary institution according to the different systems, such as: parliament, court, assembly, council. Under a parliamentary system, executive power emanates from the legislative branch invested in parliament, and governments hold office and remain in power so long as they enjoy the confidence of the parliament. Under a presidential system, parliament serves as a counterpoint to the political legitimacy bestowed upon the president through direct presidential elections. Parliaments also exist in nondemocratic regimes, such as the fascist dictatorships and 'democratic peoples' republics' of the twentieth century. Under such systems they tend to be mere puppets or facades since they exercise no real power but rather serve as rubber stamps to endorse the power of the government, devoid of the legitimacy bestowed by a pluralistic party system and in the absence of representative electoral processes.

Depending on the process leading to their formation, parliaments can be classified according to three criteria: composition, action, and structure. Composition is subject to two variants: the first of these relates to criteria based on noble birth or propertied status, whereas the second is based on universal suffrage. Functionally speaking, contemporary parliaments can also be classified as bicameral or unicameral, depending on the constitutional framework of the State, which may have opted for a single chamber or for two chambers. In any case, the chamber (or at least one of the chambers) ought to be elected directly by the people in free elections.

In functional terms, parliaments may be classified as 'active' or 'deliberative.' Active parliaments are those that, within a parliamentary system, have the power to appoint and support the government, as is the case in most modern parliamentary democracies. The so-called deliberative parliaments are those that have no direct responsibility for the business of government, and include the pro-forma parliaments tolerated by authoritarian regimes. Under a presidential system, as a counterpoint to the legitimacy of the 'working' executive branch, the parliament assumes a predominantly deliberative role. These characteristics apply equally to unicameral and bicameral parliaments, in both federative and centralized states.

It is the decisive role performed by parliaments in the process of forming governments and providing them with support that has given currency to the term Parliamentary System to describe the form of government in which empowerment and support of the Executive Branch are dependent on a vote of confidence from the parliament. This definition applies to the form of government a term that expresses the legitimacy that stems from the legislative branch in its executive form. Under a presidential system in which the Executive is the predominant Branch of Government, as is the case in the United States, the Head of State is voted to office by direct elections and consequently his legitimacy competes with that of the

parliament. In both parliamentary and presidential systems, the parliament is said to embody the Legislative Branch and its political legitimacy stems from the ballot.

2. *Origins*

Modern parliaments stem from two roots. The first of these is the corporative assemblies of the Estates that were summoned under the *Ancien Regime* in Europe, and the second, the broad-based popular representation originating from revolutionary movements. The first model of corporative representation brought together the three Estates—namely, the nobility, the clergy, and the 'third estate' which evolved gradually into a form of parliamentary representation as sovereign rulers required assent to levy taxes and make war. This first version of the dependence of the sovereign ruler on an institution of a parliamentary nature did not lead directly to democracy nor to recognition of popular sovereignty. The origin of these lies in the second root of modern parliaments—the popular revolutionary root—from which stem: the origins of the process of definition of popular sovereignty, the concept of citizenship, the legitimacy of power based on representation and the delegation of sovereign power, the rule of impersonality of the law within a context of absolute civil equality among citizens, the extension of such rights to all adult citizens within society and the adoption and consolidation of liberal democracy, the expansion of suffrage and political parties, the constitution of government and of its control (under both the parliamentary and presidential systems), and the expression of popular will through the ballot. Both roots reflect anti-absolutist sentiments and it is from them that a representative democratic model, with elections at predetermined intervals, developed.

Both these origins of parliamentary power are clearly preserved and apparent in the structure and functioning of the British parliament. The House of Commons, based upon the weight it amassed in the civil war of 1642–9 and the upon manifestation of its social powers in the Glorious Revolution of 1688–9, assumed its role of expressing the collective will as a source of power detrimental to that of the Crown. The result of the revolution, which established the right of parliament to chose a new sovereign, is reflected in the jusnaturalist and contractual arguments presented in Locke's *Two Treaties on Civil Government* (1690)—one of the principal theoretical works in support of the legitimacy and controlling powers of parliament.

The House of Lords, with its elitist composition, preserves the corporativist tradition of the Estates. Modern French tradition, established with the Jeu de Paume Oath of June 20,1789 at the advent of the French Revolution, signals a break with the system of Estates General of the *Ancien Regime* and use of the term Parliament to designate the judicial and notarial functions, as it had since the twelfth century. The parliamentary assembly installed in the very earliest days of the revolution enshrined the principal of popular sovereignty. In both the British and the French cases, affirmation of the principle of popular sovereignty arose from revolutionary movements, whereby the political absolutism of European states was substituted by constitutionalism. The role performed by a representative system in the formation of the United States is unique. The affirmation of the will of the Americans against unconditional rule by the British Crown led to the establishment of a community-based version of the decision-making process, similar to the community practices of the Protestant congregations, which served as a collegiate political model for the provincial and continental congresses, in which no form of corporative representation was present.

3. *Evolution*

At the onset of the nineteenth century there were three types of parliaments: mixed–popular and aristocratic–bicameral under the English model; the French unicameral, popular model; and the American bicameral, popular and federalist model. In Western Europe and in the Latin-American republics, an era of constitutionalization began. Both monarchies and republics adopted political systems dependent on representative legitimacy. Under parliamentary systems (in which the Government is dependent on investiture by the Legislative Branch), and under presidential systems (in which the Executive is chosen by direct elections), parliaments assumed increasing importance in political life. Parliament became the forum for representation and expression of social and economic forces, where the repercussions of rights claimed took on resonance and the extension of political and civil rights was played out.

The role of parliaments evolved during the course of the nineteenth century from being initially consultative to being deliberative. The legitimacy of parliaments became independent of the legitimacy of the sovereign upon emerging from the electoral process. As a consequence of electoral and census systems which, throughout the 1800s, maintained cultural and economic distinctions among citizens, despite their all being held (theoretically) equal before the law, parliaments maintained an indirect corporative dimension until the institution of universal suffrage. Such distinctions were based on the systematic exclusion of women and the preference for the rich, the well-born, and the literate.

The itinerary of consolidation of the power of the Legislative Branch was implemented in different ways under the North American presidential, the British

parliamentary, and the French and German parliamentary models. The United States Congress reflects the federative structure of the State and the fact that the United States spent the nineteenth century engaged in a process of expanding its frontiers and acquiring new territory. Furthermore, the political model adopted by the American Congress implies that each state of the federation, through actions taken by its own political spheres, has joined the Union. The principle of legally joining a federation of states, together with that of popular sovereignty, is enshrined in the bicameral parliamentary systems predominant in federative states (such as the United States, Canada, Australia, Brazil, India, Austria, Mexico, Argentina, Germany, Japan, and Indonesia). In centralized states, bicameralism is the result of the continuity of mixed forms of representation, with roots stemming back to the Estates and to popular sovereignty. Both the British and the French parliamentary systems provide examples of such evolution. In both cases, tradition has been partially preserved through the upper house (the House of Lords and the Senate), whereas popular representation is ensured by the Lower House (the House of Commons and the National Assembly). This model was also adopted by the Scandinavian monarchies and also, at specific times, by such states as the German Empire (1871–1918) and the multinational Austrian–Hungarian Empire (1848–1918).

Parliaments emerged upon the European scene as a counterpoint to monarchic autocracy. In their contemporary form as 'parliaments of citizens,' they emerged from an extensive line of development going back to the principle of egalitarian representation first put forward in the eighteenth century. The parliamentary model of the late twentieth century is the fruit of the maturing of philosophical concepts with respect to human rights and the separation of powers. As the twenty-first century dawns, there is widespread social awareness that there can be no exercise of power without political representation, and that this resides in the legislative bodies. The transformation of political regimes into fully-fledged democracies has led to an expansion and diversification of the roles of parliaments. To the role of authorizing tax collection has been added deliberation of public budgets, supervision of government actions, and full participation in decisions regarding public policies.

The reform of electoral systems and extension of suffrage have placed on the stage of public and social life the professional politician, a party member who represents specific interests. The principal interest that became enshrined over the length of the nineteenth century—from the French Revolution to the end of World War I—was to force the acceptance of, respect for, and practice of popular sovereignty as the source of the legitimacy of all power.

Central to an understanding of the evolution of parliaments is the concept of popular sovereignty. It was the new element that entered the political scene in the eighteenth century, whereby Monarchic authority—though still enjoying wide support—was no longer accepted as being absolute, nor independent of social support. It found expression in parliamentary democracy which essentially signifies a system in which government is made up of a nationally-elected as sembly to which it is accountable. Acceptance of this concept was consolidated during the course of the nineteenth century when the ideal of democracy became associated with the parliamentary system of government.

'Active' parliaments—the importance of which has increased remarkably since 1945 and is sill increasing at the onset of the twenty-first century—have taken on a direct responsibility for furthering the public interest. This is in line with a historic trend, initiated in the 1790s, of delegating sovereignty to elected representative bodies. In such parliaments, the house assumes not only the role of deliberating themes of interest to society but also of making decisions relating to public polices stemming from them. Under presidential systems, as is the case of the United States, the powers of the Executive are legitimized directly by popular vote, independently of parliament. Although this is true in terms of the origin of legitimacy (the ballot), it does not mean that the concrete exercise of power is free of controls. On the other hand, the extension of suffrage and the continuous rate of development of political participation since the 1830s have brought about changes in the exercise of power which have not excluded the presidential systems. Though it may even be possible to govern without parliament—against parliament it is not possible to govern. Though a strong presidential government might tend to ignore collective assemblies, nonetheless, the democratic framework of the State serves to counterbalance the relationship. Parliamentary democracy is an achievement of the Western European States and, increasingly of a number of Latin American and former communist Eastern European States. It is also the result achieved by certain nations of the British Commonwealth which, since becoming independent of Britain, have incorporated the parliamentary system of government. Such is the case in Canada, Australia, India, and New Zealand. Democratic and parliamentary systems are rarer and less stable in Africa, the Middle East, and the Far East.

The expansion of suffrage, resulting in mass democracy and recourse to regular popular participation (formalized through political parties or indirectly through lobbies or special interest groups), have caused parliaments to assume the roles of formulation and assessment of public policies (and of decision-making under parliamentary systems of government), particularly in the financial area (through the preparation of the budget and the exercise of control over its execution). The rate at which parliamentary institutions have assumed a leading role in the political process varies from one region to another. In periods

of economic expansion or of war, the active role of parliaments may go into retreat, as occurred in the *Belle Epoque* and during World Wars I and II. Under authoritarian regimes, of the left or of the right, it is not uncommon for parliaments to be merely irrelevant decorative trimmings to the institutional architecture (as was the case in Fascist Italy, Nazi Germany, in the communist countries, and under the military dictatorships of Latin America in the twentieth century). Beginning in the 1990s, successful experiences of systems of government centered upon parliaments have been adopted in various regions of the world. In Eastern European countries, the West European parliamentary model of political organization has been widely adopted, though it runs up against the tendency to fall back on a powerful Head of State with a presidential profile that is a legacy of the centralizing tendency of a communist past which lasted from 1945 to 1990. In Latin America, with the restoration of democracy that began in the 1980s, parliaments have taken on importance in a context clearly dominated by the presidential models of the North American type. In Asia, Japan, and India active parliaments have been the rule since the late 1940s. In Africa, although practically all countries have parliamentary institutions, the parliamentary model, whether active or deliberative, remains incipient and the role of parliaments largely irrelevant.

4. *Future Prospects*

Issues discussed when examining the history of parliaments fall largely into three groups: the external environment; domestic characteristics; and the political characteristics of parliaments. The external environment includes efforts to articulate four trends: the constitutional structure of parliaments, the administrative structure, the (political or electoral) system upon which they are based, and the interest groups with which they must interact. Their domestic characteristics comprise five factors that are decisive for the process of constituting parliaments: their members (their social and professional background, political experience and aspirations, and their degree of commitment to the legislative branch); parliamentary parties, committees, and houses as operational institutions, relations between members of parliament and their constituents as a reflection of social, economic, and cultural structures. The characteristics of their politics open up the possibility of studying parliaments as forums where the repercussions of short-, medium-, and long-term structures are felt and the way in which they reflect the dynamics of political decisions.

There are four principal trends in the historical research on modern parliaments: recovery and systemization of sources; sociological analysis of the composition of parliamentary representation; a physiological examination of parliamentary speeches; analysis of the role of parliaments in public administration (the formation and control of governments, monitoring and assessment of public policies, and taking of measurements among the voters and public institutions). A great portion of these trends is typically European. In Great Britain and in Germany, for example, recovery of sources is an important approach when placing parliaments within a historical context. The same is true in Brazil, Canada, and Australia. In all these cases, this monumental work has been linked directly to the consolidation of their respective identities over time and to the legitimization of the democratic structures of the State. However, as also occurs in other walks of life, historical research trends evolve.

Not so long ago it was possible to write on the history of parliaments by concentrating solely upon formal texts. Causes and developments would be carefully taken into account, accounts taken down in writing in accordance with the level of detail required, and the consequences for political parties and other significant players in the political processes analyzed. In many of the older texts, an implicit (and sometimes explicit) premise recurring throughout the explanation of the causes and consequences of parliamentary life was that it was possible to chart their progress. History reveals the emergence of parliamentary democracy, as legislation expands the electorate (in accordance with the results of the census, especially in the nineteenth century), to a certain degree making the vote more equal and elections more 'honest.' It was perfectly possible to explore democracy through reference to certain key political dates in each country. Thus the British system came to be regarded as a paradigm by others, as it ensures mechanisms which guarantee the peaceful transfer of power.

Such emphases have been radically altered. First, historical research, in a more skeptical and less self-confident age, no longer seeks to chart the progress from a smaller to a larger state, not least because society at the end of the twentieth century has become much more ambivalent about the building of progress. Second, because much more emphasis had been devoted to continuity than to change (especially with respect to representative democracy). Third, research efforts (polling) at the constituency level have greatly increased. It is hardly surprising that such work produces revelations that do not support the progressive view. It could be argued that the nature of the political process has undergone changes for the worse, from the viewpoint of the lower classes. The way in which governments manipulate electoral laws and the rules defining constituencies and how majorities are defined (for example, proportional systems, simple majorities or mixed systems) have become relevant research themes in the field of historical research, the results of which are only now coming to light.

Furthermore, historians have begun to question the importance of the vote for participation in the political process. Indeed, a well-substantiated demand addressed to the administration is likely to be more effective than heated political speeches supported by a few votes in a legislative body. The parallel with nineteenth-century Brazilian politics (and the process of overthrowing dictatorships in the twentieth century, particularly in Latin America) are cases in point. As O'Gorman and Vernon showed for Great Britain, lack of voter participation has not been an obstacle to political participation. Members of Parliament must take a broad spectrum of interests into account if they wish to remain on good terms with their constituents. The right to vote does not necessarily bring with it access to political power. The suffragettes of the early twentieth century produced brilliant political propaganda, showing that educated, intelligent women with an interest in public affairs were denied access to the vote, whereas ignorant, vulgar, and drunken men were free to participate in it. In reality, such campaigns had a reach far beyond the mere exercise of the ballot: they were an affirmation of the importance of women in a stable political system.

Interest in electoral policies has been reinforced by detailed studies of elections, supported by long-term statistics. Such statistics have used materials that historians immediately recognize as being prime raw material for such studies. Voter registration records prior to the introduction of the secret ballot in Great Britain in 1872, for example, show who voted for whom. Such surveys make it possible to monitor changes in voting habits and to identify which candidates or parties lost favor with different social groups. The 'social geography' of elections and the 'sociological premises' of a given society thus become important instruments for historical analysis. Historical research brings together both the quantitative and qualitative techniques of the social sciences to obtain relevant results for the analysis of parliaments.

Political studies relating directly to social and occupational structures also touch upon two other lines of work carried out over the past 40 years. 'History from the bottom up' was in vogue in the 1960s and 1970s, when many researchers began to devote themselves increasingly to social structures. Attempts to discover the origins of the 'working class' and, somewhat later, of the 'middle class' involved deep social analyses. Historians more sensitive to issues of social class admit that political awareness plays a part in the issue. Indeed, class consciousness is related directly to the exercise and pursuit of power and in turn requires the existence of political structures that need to be changed. Nonetheless, in Europe, as of the early nineteenth century, debate on parliamentary reform became enmeshed in discussions on social class and on changes in the 'political culture.'

Another important intellectual development has been emergence of the so-called 'linguistic turn' which began to influence historians in the 1980s. Detailed studies of language as a vehicle for conveying ideas, conducted by historians, have been inspired by radically new ideas borrowed from literary criticism. The relevance, and even the validity, of such an approach has engendered much debate among historians versed in the empirical tradition. Many are reluctant to accept that language can build its own realities, rather than having been invented to serve economic, social, or cultural needs. Unquestionably, extreme versions of poststructuralism, which claim that a text has no validity except in its own context, undermine any type of historical investigation in that they do not accept the notion of relative importance and suggest that change cannot be explained by intentional 'privilege' attributed to particular parts of historical documents. The 'linguistic turn' causes historians to be even more strict in the control of their premises on parliamentary discourse as a historical source. Language may effectively be used to mold ideas. Studies on demagoguery, populism, and '*caudilismo*,' and their influence upon the formation of parliaments in such countries as Brazil, Mexico, and Argentina, are pioneers in this field. The study of the language used by the Chartist reformers of parliamentary democracy in Great Britain, for example, shows that their real goal was political change.

A third theme relating to the expansion of the democratic process is treated under the term 'progress.' Research shows clearly that, when it comes to political change, progress is neither simple nor linear. Parliamentary history is related to, but not defined by, the history of democratization. 'Democratization' of itself is a highly problematic and contested concept. Nonvoters are not necessarily excluded from politics. No parliamentary reform ever 'grants power' to another social group—at any rate, not directly. Women do not all of a sudden hold power simply as a consequence of a scrap of paper called a 'law.' Current research trends show that the law is more likely to ratify a political order already in place than to introduce sweeping change. The recent development of supranational parliaments (the European Union, Mercosur) provides an entirely new field, into which investigative research is about to begin.

Historians have become more involved with continuities. Much of the historical research carried out in the 1980s and 1990s was accompanied by warnings about the dangers of economic reductionism, or of any other type of reductionism. However, the determination of historians to attempt to see things 'within their context' has contributed to the clarification of a series of relevant conclusions on the subject of parliamentary reform. They are now more aware that neither the 'top-down' explanations based upon the higher policies of central parliaments, nor the 'bottom-up' approaches that concentrate on extraparliamentary forms of pressure, are of themselves sufficient. The emphasis has swung to the need to understand the

intricate interrelations of the complex factors which created the political culture of the twentieth century and, above all, World War II. The changing methods and interpretations nonetheless confirm the central role of parliaments in this culture.

See also: Citizenship, Historical Development of; Democracy, History of; Parliamentary Government; Political History (History of Politics); Political Parties, History of; Public Sphere: Nineteenth- and Twentieth-century History

Bibliography

Adamietz H 1978 *Parlamentsgeschichte*. Hanschild, Location

Bélanger R, Jones R, Vallières M 1994 *Les grands débats parlementaires*. Publisher, Laval, Quebec

Bisset A 1882 *A Short History of English Parliament*. Publisher, London

Brooks C 1999 *Governors & Government. Political and Public History of Early Modern England 1550–1850*. Arnold, London

Chacon V 1997 *História institucional do Senado*. Publisher, Brasília

Copeland G W, Patterson S C (eds.) 1994 *Parliaments in the Modern World: Changing Institutions*. University of Michigan Press, Ann Arbor, MI

Das Österreichische Parliament 1984 Vienna

Franks C E 1987 *The Parliament of Canada*. University of Toronto Press, Toronto, ON

Fraser A, Mason R H P, Mitchell P 1995 *Japan's Early Parliaments 1890–1905*. Routledge, New York

International Commission for the History of Representation and Parliamentary Institutions 1970 *Parliaments, Estates and Representation*. International Commission for the History of Representation and Parliamentary Institutions, Ashgate, London

Interparliamentary Union 1983 *Les Parlements dans le monde*. Interparliamentary Union, Paris

Kluxen K 1983 *Geschichte und Problematik des Parlamentarismus*. Frankfurt

Kühne T 1998a Parlamentsgeschichte in Deutschland: Probleme, Erträge, Perspektiven. (History of Parliamentarianism in Germany) *Geschichte und Gesellschaft* **24**(2): 323–38

Kühne T 1998b Parlamentsgeschichte in Deutschland: Probleme, Erträge, Perspektiven einer Gesamtdarstellung. *Geschichte und Gesellschaft* **24**(2): 323–38

Kurian G Th (ed.) 1998 *World Encyclopedia of Parliaments and Legislatures*. Congressional Quarterly, Washington, DC

Laundy P 1989 *Parliaments in the Modern World*. Gower, Dartmouth, UK

Loewenberg G, Patterson S C 1988 *Comparing Legislatures*. University Press of America, Lanham, MD

Olivier-Martin F 1997 *L'absolutisme français; suivi de Les parlements contre l'absolutisme traditionnel*. Librarie générale de droit et de jurisprudence, Paris

Ritter G A (ed.) 1974 *Gesellschaft, Parlament und Regierung: zur Geschichte des Parlamentarismus in Deutschland*. Kommission für Geschichte des Parlamentarismus und der politischen Parteien, Düsseldorf

The History of Parliament 1998 (CD-Rom Edition: House of Commons). Cambridge, UK

The Journal of Legislative Studies 1995 Frank Cass, London

VV. AA. 1963–1984 *Storia del Parlamento Italiano*. S. F. Flaccovio, Palermo

E. C. de Rezende Martins

Parsons, Talcott (1902–79)

From the late 1940s to the mid-1960s, Talcott Parsons was the single most influential social theorist in the world. A developer of what is now popularly called 'action theory,' 'functionalism,' and 'structural-functionalism,' Parsons spent his entire career at Harvard University, which helped considerably in institutionalizing his ideas and also in providing him access to talented graduate students. By the 1950s, his publications became a major part of what was literally a 'common curriculum' for graduate students in all major sociology departments—as well as in many of the best political science departments. By being institutionalized in this way, Parsons' publications elevated the rigor of American graduate training in sociology in particular. More than anyone else, he also defined the 'classic' theoretical tradition for sociology as a discipline and then added a more abstract, arguably more conceptually sophisticated contemporary rival to it. With this, he raised the bar for social theory worldwide, from Germany, France, and Great Britain to Japan and even the former Soviet Union. A strong case can be made today that every major social theory since the mid-1960s has been developed in direct or indirect dialogue with Parsons' functionalism (see Habermas 1981 for a related statement).

Having become literally an icon in the discipline in the first three decades of his career, in the last two decades he attracted a considerable share of the iconoclasm that more generally characterized the student movement and academia of the late 1960s and early 1970s. Parsons was subject across two decades to far greater criticism than that directed to any other theorist of his generation or since. Given the tenor of the times, most criticisms were rhetorical, often personal. Today, they leave a strong impression that few critics had bothered to read his works with the care typically afforded to basic empirical studies, let alone with the rigor and dispassion that any complicated social theory demands. Yet, the cumulative effect of the broadsides he received in the late 1960s and early 1970s was to leave the collective memory of the discipline with an understanding of Parsonian functionalism that is difficult to reconcile with often-repeated, explicitly-stated positions in his publications. What Parsons endeavored to accomplish in his day, providing the social sciences with a common language, and how he went about accomplishing it,

developing a 'conceptual framework' comprising analytical distinctions, now eludes the discipline's collective memory.

Talcott Parsons was born on December 13, 1902 in Colorado Springs, the youngest of six children of Edward S. Parsons, a Congregational minister and college administrator, and Mary Augusta Ingersoll Parsons, a housewife and active suffragist (on biographical details, I rely primarily on Nielsen 1991). Both parents were New England WASPs who traced their ancestry to mid-eighteenth century settlers, Mary to theologian Jonathan Edwards. A graduate of Yale Divinity School, Edward Parsons holds fundamentalist views of Christianity but leftist views in politics, particularly regarding the issue of economic inequality. He is an active participant in the Social Gospel movement that supported organized labor. Both parents inculcate into Talcott and his older siblings the importance of finding a mission in life and then pursuing it assiduously. Many observers comment later on the mature Talcott's remarkable energy and perseverance—including on the final two days of his life in Munich in 1979. He published over 150 sole-authored articles and 14 books or collections.

In 1917 the Trustees of Colorado College force Edward to resign as Dean after he supports two women who charged the College's President with sexual harassment. The family then moves to New York City where Talcott enters Horace Mann High School, a laboratory school for boys operated by Teacher's College of Columbia University. Talcott earns an undergraduate degree at Amherst College (as had his father and two older brothers) from 1920–4. At the time Amherst was also an experimental, progressive school. Parsons at first focuses on biology and chemistry but in his junior year becomes 'converted' to the social sciences under the influence of an unorthodox institutional economist, Walton Hamilton.

Parsons' graduate training was remarkably brief: one academic year (1924–5) at the London School of Economics as a nondegree candidate, then another (1925–6) as a degree candidate at Heidelberg University. At LSE, Parsons was influenced by Morris Ginsberg, L. T. Hobhouse, and the methodological functionalism of anthropologist Bronislaw Malinowski—as well as by fellow-student E. E. Evans-Pritchard. Parsons also met his future wife and mother of three children, Helen Banerott Walker, a Bryn Mawr College student studying banking. Her family of conservative white- and blue-collar workers consider Talcott a leftist, a 'pinky.' At Heidelberg, Parsons first learnt of the recently deceased German social theorist Max Weber and discussed his work with his wife Marianne, his younger brother Alfred, and Alexander von Schelting. Taking coursework from Edgar Salin and Emil Lederer (on economics), Karl Mannheim (a friend of Georg Lukacs) and Karl Jaspers (on Kant), Parsons selected how sociology and economics differ in portraying 'capitalism' as his dissertation topic.

During his year at Heidelberg Parsons was offered a one-year teaching appointment at Amherst in the Economics Department which also allowed him to teach an independent course in sociology. In 1927 Parsons was offered an instructorship in economics at Harvard, thus beginning a life-long career there. Harvard's Department of Sociology was a relative late-comer in the discipline, beginning in the fall 1931 under the chairmanship of Pitrim Sorokin, a 42-year old emigre from Russia recruited from University of Minnesota. Parsons, then 29, was appointed as an instructor. Disagreements and misunderstanding between the two theorists became common knowledge in the Department and across the University. Yet, throughout Parsons' career, both critics and proponents often commented on his low-key demeanor when grappling with ideas and interacting with colleagues and students. It is difficult to find accounts of him being particularly effusive, let alone dominating. A short, stocky, already balding young man who sports a moustache and always smokes cigarettes, critics also comment often on his charm when interacting with other major national and international academicians at Harvard, in the American government, and overseas.

During his earliest years at Harvard, Parsons was impressed with philosopher Alfred North Whitehead and his notion of the 'fallacy of misplaced concreteness.' Consistent with Kant, Whitehead held that scientists apprehend 'reality' or scientific 'truth' only through analytical distinctions, not more immediately—whether experientially or directly through empirical findings. A few years later in correspondence with phenomenologist Alfred Schutz from 1940–1, Parsons defended the notion of 'analytical realism'—as opposed to ontological realism or empirical realism—against Schutz's position that social scientists can somehow gain more direct access to the 'life-world' of their subjects of study (Grathoff 1978).

By contrast to Schutz and then also to ethnographers at University of Chicago and American empirical researchers more generally, Parsons endeavored across his 50-year career to identify the most irreducible analytical distinctions unique to sociology, the scientific study of social life. In pursuing this project, Parsons arrived at three successive 'theoretical syntheses' and along the way trains four remarkably talented and productive cohorts of graduate students. All three of Parsons' theories, which he called 'frameworks of concepts,' involve a 'functionalist' approach to the study of social life. This is an approach much maligned today, and yet a basic premise of functionalism is hardly controversial. Any social scientist who poses the following question is operating within a functionalist approach: What is the relationship, if any, between the substantive area of social life I am studying and the direction of social change? Here,

then, is a common ground shared by critical theorists and Parsonian functionalists but not by more historicist researchers, those who study particular events or periods in isolation and are loathe to identify any direction of social change.

While exceedingly abstract and complex, there are two ways to grasp the significance of Parsons' publications across his career. In general, he endeavored to account for the place and purpose of norms in maintaining social order both historically and cross-nationally. More particularly, he endeavored to account for the place and larger significance of professions, a unique set of occupations, in contemporary societies. Professions are a pivotal subject for sociological inquiry because they are important nongovernmental bodies in all modern societies. Yet, their prestige or status in any society hinges as much on whether they are valued culturally, supported by generally recognized social norms, and centrally situated in a social structure as on whether practicing professionals compete effectively in self-regulating markets. In all modern societies, professionals are simultaneously driven by economic pressures to maximize profits, like other workers, and also by normative pressures to conform to certain extra-economic standards of conduct that typically do not bind most other workers. Parsons was interested in accounting at a theoretical level for the rise, evolution, and institutionalization of these extra-economic restraints on self-interested behavior, and then also to explore whether and how they contribute to social order.

In his earliest publications, from the 1920s to the mid-1940s, Parsons approached professions by first distinguishing their behavior from that of economic enterprises, on one side, and religious organizations, on the other. Corporate managers and shareholders tend to act in strictly utilitarian or market-mimicking ways as they endeavor to maximize either growth or profit, a quantitative end that can be recognized interpersonally. Clerics and religious believers, however, tend to exhibit ritualistic fidelity to norms as they—presumably—seek spiritual salvation. This is a qualitative end that is transcendental or metaphysical; as such, its attainment is a matter of faith, not something that can be recognized interpersonally. Professionals, Parsons held, also exhibit fidelity to norms, but more 'voluntaristically.' They do not act in strictly market-mimicking ways, but they also do not conform to norms ritualistically with any transcendental end in view. Rather, they exhibit fidelity to norms as a means to attain qualitative ends that are worldly or empirical, and thus capable of being recognized interpersonally—such as their patients' physical or mental health, or their clients' legal innocence, or scientific truth.

Parsons came to appreciate the significance of 'voluntaristic action' after a careful reading of two economic theorists, Alfred Marshall and Vilfredo Pareto, and two social theorists, Emile Durkheim and Max Weber. This reading forms the core of *The Structure of Social Action* (1937) which, along with Parsons's 1930 translation of Weber's *The Protestant Ethic and the Spirit of Capitalism*, established Weber and Durkheim as two of the new discipline's 'classic' theorists. During the 1930s, Parsons trained a first cohort of graduate students in the classics and his own theory of 'voluntaristic action,' including: Robert Merton, Kingsley Davis, John and Mathilda Riley, Edward Devereuw, Marion Levy, Wilbert Moore, and Florence Kluckhohn (along with Edward Hartshorne who was killed by a sniper while on active service in postwar Germany).

In spring 1944 Parsons was promoted to full professor and the chairmanship. By January 1946, he transformed the Department of Sociology into the Department of Social Relations, with the explicit aim of linking structural and social psychological approaches in the scientific study of social life. In the same year, Parsons also began formal training in psychoanalysis as a Class C candidate at the Boston Psychoanalytic Institute. He had been lecturing informally on Freud since fall 1938.

As Parsons assumed his leadership role in Harvard's Department, his theory went through a brief mid-career change from the late 1940s to the mid-1950s as he considered findings from Robert Bales's study of small groups. Parsons began thinking about professions (and how to identify norms in social life more generally) in terms of six basic decisions that people performing any role or task either make explicitly or have imposed on them by cultural values, institutionalized norms, or social structures. For example, certain roles (including those performed by professionals) mandate treating others (clients or patients) in universalistic ways. By contrast, other roles (including those performed by parents) encourage more particularistic treatment of others (children). Parsons called these decisions and requirements 'pattern variables.' By his account, researchers can use six such variable-pairs (a) to distinguish professions from other occupations and (b) to identify changes in the behavior of professions both historically and cross-nationally. Parsons' major publications during this period are a collection, *Essays in Sociological Theory Pure and Applied* (1949), and three monographs: *The Social System* (1951), *Towards a General Theory of Action* (1951, with Shils), and *Working Papers in the Theory of Action* (1953, with Robert Bales and Edward Shils).

During this period Parsons trained a second cohort of graduate students in the pattern variables, including: Bernard Barber, Harry Johnson, Robin Williams, Jr., Jesse Pitts, Harold Garfinkel, Francis Sutton, and Robert Bales. More generally, as part of the buoyancy and optimism that pervaded the US after the war, many sociology departments re-evaluate their curricula. The discipline is open collectively to receiving a new theoretical synthesis, and Parsons' notion of pattern variables, while preliminary and as it

turned out fleeting, is a prime candidate. Another is Robert Merton's notion of 'middle range theory,' an idea he first proposed at the annual meeting of the then American Sociological Society in 1947 in opposition to Parsons' 'general theory.'

Finally, beginning slowly in the mid-1950s and then with more rapid developments from the early 1960s to the end of his career, Parsons approached professions within what he called the 'AGIL schema.' This is Parsons' single most significant contribution to social theory. During his transition from the pattern variables to the AGIL schema, Parsons' publications, lectures, and presentations create an unusually kinetic intellectual excitement among collaborators and graduate students. Everyone working in his circle operated with a palpable sense that social theory is in the midst of a fundamental breakthrough, a sense that American sociology has not witnessed since.

This transition period to Parsons' mature social theory is marked in particular by *Economy and Society* (1956, with Neil Smelser), and Parsons' training of a third cohort of graduate students, including: Albert Cohen, Renee Fox, Joseph Berger, Norman Birnbaum, Neil Smelser, James Olds, Jackson Toby, and Miriam Johnson. In addition, many earlier graduate students whose training had been interrupted by wartime service returned to Harvard. Talented undergraduates who come into contact with Parsons during this period include Robert Bellah and Charles Tilly. Also in transition during this period is the leadership of sociology at Harvard. Parsons resigns as chair of the Department of Social Relations in 1956, and within the next two years learns that he has diabetes. Sorokin retires in 1959.

From 1956 to 1961, Lewis Coser and Dennis Wrong in the US, David Lockwood in Great Britain, and Ralf Dahrendorf in Germany challenged Parsons' functionalism in relatively measured tones, but with uneven rigor. They were also shooting at a moving target in that Parsons was developing his mature social theory. The high-point of scholarly criticism during this period comes with a collection of essays by philosophers and sociologists edited by Max Black, to which Parsons responded with a long concluding chapter. In April 1964, Parsons was a central figure at a famous session of the German Sociology Association in Heidelberg devoted to reconsidering Max Weber's contributions. He found himself at odds with Frankfurt School critical theorists, most notably the young Jürgen Habermas and the more established Max Horkeimer and Herbert Marcuse (see Stammer 1965/1972 for a collection of these exchanges).

From this point forward, criticisms of Parsons become harsher, more personal and ideological than analytical and scholarly. Parsons was attacked, particularly in the US, for: his religious background, his WASP ethnic and cultural heritage, his writing style, his amazing productivity, his putatively strategic rather than scholarly decision to make sociology as rigorous conceptually as economics, his putative mistranslation of Weber, the putative conservative nature of his own family life including his putative responsibility for his daughter's suicide (in June 1964, two weeks after the Heidelberg debate), his putative political conservatism, and his putative general obeisance to 'American capitalism' or 'the establishment.'

Parsons unveiled the AGIL schema explicitly in 1963, with major articles on the concepts of power and influence, and then in 1964, with his first statement about 'evolutionary universals' (see the Turner 1999 collection for the 1964 article and references to the others). As fully developed, the AGIL schema isolates analytically those aspects of behavior in any role or position that contribute, respectively, to each of four general social 'functions': economic efficiency ('adaptation'), administrative effectiveness ('goal-attainment'), fidelity to social norms ('integration'), and fidelity to cultural values or a society's most basic institutional arrangements ('latency'). Parsons also proposed that four 'media of interchange' circulate between these analytical subsystems—money, power, influence, and value commitments—thereby bringing a certain orderliness or predictability to the whole. He added notions of 'systems-theory,' 'pattern maintenance,' and 'hierarchy of control' to this basic four-function breakdown of social life.

In 1973, Parsons and Platt published what is arguably his most important single work since 1937, *The American University*. In this work Parsons rethinks his approach to professions by incorporating two major points into his social theory. First, he proposes that professions and sites of professional practice (such as universities, hospitals, and research institutes) are organized in a 'collegial form,' not in a bureaucratic form or a democratic form. Second, he proposes that professions are distinguished from other occupations by their members' willingness to bear 'fiduciary responsibilities' to clients and the larger society. These two points provide the foundations for an account of the place and purpose of professions in contemporary societies that differs radically from any effort, whether by rational choice theorists or social critics on the left, to treat the professions analytically as interchangeable with other occupations or other corporate entities. By 1973, however, the collective memory of the discipline is so dominated by criticisms of the AGIL schema that this major work, arguably Parsons' single most important substantive contribution to social theory, goes largely unread by critics and proponents alike.

Through the 1960s and 1970s, Parsons trained a fourth and final cohort of graduate students in the AGIL schema and related theoretical developments, including: Victor Lidz, Jan Loubser, Leon Mayhew, Rainer Baum, Dean Gerstein, John Akula, and Willy DeCramer. Parsons also influences undergraduate Jeffrey Alexander. Only in the mid-1970s, a few years before his death, did Parsons begin methodically to

respond to at least a few criticisms of his work, most notably the informed analyses of Bershady (1973) and Warner (1978). Bershady's *Ideology and Social Knowledge* in particular marks a major sea change in commentary on Parsons' works. Bershady is the first to draw attention to Parsons' effort to ground the social sciences against relativism with analytical concepts. More than anyone else Bershady appreciates that Parsons pitched his social theory at an unusually abstract level of analysis because he sought literally a common language for the social sciences. Prior to Bershady both proponents and critics generally failed to convey the power and sweep of Parsons' social theory in form, its aspiration to unite the social sciences. Parsons' project was essentially to give structuralists and symbolic interactionists the means, the common language, by which to demonstrate to each other the central import of their respective findings.

Parsons demonstrated his capacity to 'translate' others' findings and ideas in his own scholarship. He co-authored, team-taught, or otherwise collaborated closely with a remarkable range of theorists and researchers across disciplines, including: Edward Shils, Joseph Schumpeter, Samuel Stouffer, Florence Kluckhohn, Robert F. Bales, James Olds, Morris Zelditch, Philip Slater, Neil Smelser, Lon Fuller, Kaspar Naegele, Jesse Pitts, Winston White, S. N. Eisenstadt, Robert Bellah, Victor Lidz, Gerald Platt, Edward Laumann, Erwin Scheuch, Andrew Effrat, Mark Gould, and Dean Gerstein. In addition, he engaged in private and public correspondence and debates with Frank Knight, Crane Brinton, Robert Bierstedt, Albert Blumenthal, Alfred Schutz, Chester Barnard, Philip Selznick, Eric Voegelin, Robert Merton, Kenneth Boulding, C. Wright Mills, Robert Bellah, Bennett Berger, David Riesman, Jurgen Habermas, Herbert Marcuse, Max Horkheimer, and Ernest Mayr. Finally, an impressive set of scholars was influenced heavily by Parsons himself, his students, or his theory (some of whom also influence Parsons and his students): Niklas Luhmann, Seymour Martin Lipset, Amitai Etzioni, Robert Marsh, Daniel Bell, Joseph Ben-David, Benjamin Nelson, Gabriel Almond, James S. Coleman, Karl Deutsch, David Apter, Lucian Pye, Sidney Verba, Chalmers Johnson, S. N. Eisenstadt, Ken'ichi Tominaga (Japan's leading postwar sociologist), Lawrence Brownstein, Martin Martel, Adrian Hayes, and Frank Lechner.

Parsons and Helen travelled frequently in the 1970s, including three separate trips to Japan. In early May 1979 they returned to Heidelberg, then proceeded to Munich, on the occasion of the fiftieth anniversary of Talcott's degree from Heidelberg. After Parsons' typical full day of presentations and scholarly exchanges, including a lecture on 'The Declining Significance of Social Class' attended by host Horst Helle as well as Jurgen Habermas, Niklas Luhmann, Richard Munch, and Wolfgang Schluchter, Parsons died of a stroke in the early hours of May 8. At a memorial service 10 days later in Harvard Chapel, Victor Lidz notes in eulogy: 'No sociologist of recent times has had to endure more bitter criticism than Talcott. Although a passionate man, Talcott bore criticism with equanimity as well as courage. Not daunted by even sweeping attack, he held to his own program of research and writing. Moreover, adhering to principles of value-freedom, he declined to reciprocate ideological criticism. He dealt with criticisms, always, as technical matters within the domain of social scientific theory' (Nielsen 1991).

Parsons' critics inadvertently encouraged the fragmentation of research and theory in sociology and the social sciences that Parsons endeavored to prevent. Today, in the absence of any shared framework of analytical concepts, social scientists routinely talk past each other, as opposed to revealing to each other the general importance of their research findings and theoretical developments. Parsons's 'general theory' has given way not to a finite set of readily identifiable theories of the 'middle range,' as Robert Merton anticipated in 1949, but rather to an ever more finely grained cacophony of research specialties. This trend is then recapitulated, and accelerated, by ongoing changes in graduate training. Graduate students are trained more and more narrowly, and their required survey courses in social theory hardly encourage them to move beyond any research specialty. If there is any 'common language' in the social sciences today, it is that of research methods. But even here there is more internal division among methodologists than they often wish to acknowledge.

In support of Habermas' point that we can expect any new theoretical development in the social sciences to respond directly or indirectly to Parsons' functionalism, French social theorist Pierre Bourdieu in many regards today recapitulates Parsons' AGIL schema. He identifies two poles within what he calls 'fields of power'—the economic field and the field of cultural production—and then distinguishes four fields in between: politics, the higher civil service, the professions, and the university (Bourdieu 1989/1996). The major difference with Parsons is that the French theorist's categories are more directly descriptive, less analytically rigorous. As a result, it is not clear on what basis he is distinguishing fields, why there are six fields of power rather than four or eight, or on what basis researchers may categorize particular groups or activities as components of one field or another.

By the early 1980s, a new generation of social theorists, including members of Parsons' last cohort of graduate and undergraduate students at Harvard, initiated the effort to have Parsons' publications reconsidered on scholarly grounds. Led initially by Munch (1981) in Germany and Alexander (1983, 1985) in the US, today's 'neofunctionalism' finds support from theorists whose political positions range from radical left to liberal reformist to conservative republican (see the collections edited by Hamilton 1992

and Colomy 1990). What they share, with Parsons himself, are two general principles (which Alexander 1988 now explicitly rejects). One is that it is vitally important to the social sciences to continue Parsons' effort to ground social theory on the basic analytical distinctions of the AGIL schema rather than to rely more directly on descriptive concepts (whether empirical generalizations or ideal types). The other principle is that it is incumbent today to demonstrate the empirical potential of Parsonian functionalism, to identify lines of empirical inquiry that this social theory uniquely presents to view (Sciulli 1992). A strong case can be made that Parsons' social theory has been mined only superficially (Takagi in press). An equally strong case can be made that theorists and researchers today can draw fruitfully on Parsons' basic AGIL schema, and then the first set of subdivisions he drew within each 'function,' but that they need not follow Parsons' effort to draw further distinctions within each subdivision (Mouzelis 1995).

Social theorists outside the US today who support the Parsonian tradition, even if critical of certain parts of it, include: Munch, Hans Joas, Uta Gerhardt, Horst Helle, and Harald Wenzel in Germany; Helmut Staubmann in Austria; Jens Kaalhauge Nielsen in Denmark; Pierpaolo Donati, Andrea Maccarini, Matteo Bortolini, and Riccardo Prandini in Italy; Ken'ichi Tominaga, Kiyomitsu Yui, and Kazuyoshi Takagi in Japan; and Nicos Mouzelis, Bryan Turner, and Jeremy Tanner in Great Britain.

See also: Action Theory: Psychological; Action, Theories of Social; Control: Social; Conventions and Norms: Philosophical Aspects; Durkheim, Emile (1858–1917); Educational Institutions and Society; Emergent Properties; Functionalism, History of; Functionalism in Sociology; Influence: Social; Institutions; Norms; Political Economy, History of; Professions in Organizations; Professions, Sociology of; Realism/Neorealism; Realisms and their Opponents: Philosophical Aspects; Shared Belief; Social Change: Types; Sociology, History of; Status and Role: Structural Aspects; Structuralism, Theories of; Structure: Social; Symbolic Interaction: Methodology; System: Social; Theory: Conceptions in the Social Sciences; Theory: Sociological; Value Pluralism; Values, Sociology of; Weber, Max (1864–1920)

Bibliography

Alexander J C 1983 *Theoretical Logic in Sociology, Vol. 4: The Reconstruction of Classical Thought: Talcott Parsons*. University of California Press, Berkeley, CA

Alexander J C (ed.) 1985 *Neofunctionalism*. Sage, Beverly Hills, CA

Alexander J C 1988 *Neofunctionalism and After*. Blackwell, Malden, MA

Bershady H J 1973 *Ideology and Social Knowledge*. Wiley, New York

Black M (ed.) 1961 *The Social Theories of Talcott Parsons*. Prentice-Hall, Englewood Cliffs, NJ

Bourdieu P 1989/1996 *The State Nobility: Elite Schools in the Field of Power*. Stanford University Press, Stanford, CA

Colomy P (ed.) 1990 *Neofunctionalist Sociology*. Elgar, Aldershot, UK

Grathoff R (ed.) 1978 *The Theory of Social Action: The Correspondence of Alfred Schutz and Talcott Parsons*. Indiana University Press, Bloomington, IN

Habermas J 1981/1987 *Theory of Communicative Action, Vol. 2. Lifeworld and System: A Critique of Functionalist Reason*. Beacon Press, Boston

Hamilton P (ed.) 1992 *Talcott Parsons: Critical Assessments,* 4 vols. Routledge, London

Mouzelis N 1995 *Sociological Theory: What Went Wrong? Diagnosis and Remedies*. Routledge, London

Munch R 1981 Talcott Parsons and the theory of action, I: The Structure of the Kantian core. *American Journal of Sociology* **86**: 709–39

Nielsen J K 1991 Talcott Parsons' Life, Theory and Political Orientation. Unpublished Manuscript, 5 vols.

Parsons T 1937/1968 *The Structure of Social Action*, 2 vols. Free Press, New York

Parsons T, Platt G W 1973 *The American University*. Harvard University Press, Cambridge, MA

Sciulli D 1992 *Theory of Societal Constitutionalism: Foundations of a Non-Marxist Critical Theory*. Cambridge University Press, New York

Stammer O (ed.) 1965/1972 *Max Weber and Sociology Today*. Harper and Row, New York

Takagi K in press *Talcott Parsons and American Intellectual Society*. Cambridge University Press, Cambridge, UK

Turner B S (ed.) 1999 *The Talcott Parsons Reader*. Blackwell, Malden, MA

Warner R S 1978 Toward a redefinition of action theory: Paying the cognitive element its due. *American Journal of Sociology* **83**: 1317–49

D. Sciulli

Partial Orders

The idea of ordering captures a basic faculty of the human mind: to decide for two given elements from a set of objects which one 'dominates' the other with respect to an attribute of interest, or, alternatively, which one is 'preferred' to the other.

Formally, a partially ordered set (a poset for short) consists of a pair $(P, \leqslant)$ where P is a nonempty set and $\leqslant$ a binary relation on P satisfying for all $x, y, z \in P$:

(a) reflexivity: $x \leqslant x$,

(b) asymmetry: $x \leqslant y$ and $y \leqslant x$ implies $x = y$,

(c) transitivity: $x \leqslant y$ and $y \leqslant z$ implies $x \leqslant z$.

Transitivity is, in a sense, the most important property; it is shared by all variants of order relations, though it might be a consequence of other conditions. However, asymmetry also contributes in an essential way to the meaning of a partial order. Reflexivity, on the contrary, is merely a matter of taste or con-

venience. In some cases its opposite, irreflexivity, is postulated. Such an order is called *strict*. If the asymmetry condition is replaced by symmetry, i.e., for all $x, y \in P$ $x \leqslant y$ implies $y \leqslant x$, then the resulting structure is an equivalence relation which lacks the interpretation of 'dominance' or 'preference.' Such a structure is more akin to classification or grouping with respect to similarity.

1. Related Notions and Diagrams

Sometimes asymmetry is simply dropped from the conditions. A relation satisfying reflexivity and transitivity is called a quasiorder. The investigation of quasiorders can be split into the respective properties of a partial order and an equivalence relation in the following way: Let $(Q, \lesssim)$ denote a quasiorder. Define $x \sim y$ if $x \lesssim y$ and $y \lesssim x$ both hold true. Then it is readily shown that $\sim$ is an equivalence relation on Q. Let P be the set of equivalence classes and define for $p, q \in P$ the relation $\leqslant$ by $p \leqslant q$ if $x \lesssim y$ for some $x \in p$ and $y \in q$. Then, one demonstrates two things: first, that the definition of $\leqslant$ is independent of the particular elements $x \in p$ and $y \in q$; and second, that $(P, \leqslant)$ is a poset. Thus, the 'dominance-aspect' of a quasiorder is contained in $\leqslant$, and its 'classification-aspect' is hidden in $\sim$.

One property of an intuitive meaning of dominance is not captured in the notion of a partial order: the idea that any two elements can be compared with respect to $\leqslant$. The defining conditions of a poset do not imply connectivity, i.e.,

$$x \leqslant y \text{ or } y \leqslant x \text{ for all } x, y \in P.$$

(Note: often this property is called 'completeness,' but later in this article another notion of completeness is introduced). If connectivity is added to the properties of a partial order $(P, \leqslant)$, then $\leqslant$ is called a linear or total order. A quasiorder satisfying connectivity is called a weak order. Applying the above procedure on a weak order results in a linear order on the equivalence classes of the weak order.

There is a tendency to associate the property of connectivity automatically with an order given in some field of application, e.g., if one rank orders stimuli or subjects. Doing this amounts to representing the order numerically, i.e., to assuming a function

$$f\colon P \to \mathbb{R} \text{ such that } x \leqslant y \text{ if } f(x) \leqslant f(y). \qquad (1)$$

If such a representation is possible, the order is necessarily a weak order. However, in social science, numerical representations are used widely even for orders which cannot easily be justified to be connected. For example, if the items of an achievement test are

Figure 1
The forbidden substructures of a semiorder

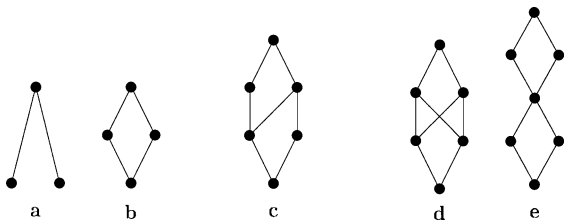

Figure 2
Lattices (b, c, e) and non-lattices (a, d)

'Rasch-scaled' (see the discussion following Eqn. (3)), a difficulty coefficient is attached to each item. This numerical scale orders the items weakly and ascribes the often made observation that subjects solve a difficult item correctly and fail at an easier one to influences of chance and error. This interpretation appears so obvious that its proponents tend to overlook the *petitio principii* which it entails, that the difficulty of an item is defined by the numerical parameter, whereas the weak ordering of the items according to a material notion of difficulty is an empirical property which will in many cases not be fulfilled.

For example, an expert judging this might well regard some items as incomparable. In such a situation a partial- (or quasi-)ordering, which allows for incomparable items, is more flexible. Thus, the assignment of numerical values, albeit ubiquitous in psychological tests, restricts drastically the range of possible models. Even partial orders can be too restrictive because transitivity is also a strong condition. The theory of knowledge spaces takes account of this aspect (see *Knowledge Spaces* for a deeper treatment of these issues).

The most convenient way of representing a poset is by its (*Hasse*) *diagram* (see Figs. 1 and 2). To construct it, one first defines $x < y$, in words 'y covers x,' to mean: $x < z \leqslant y$ implies $z = y$. Next, one draws a graph in which the vertices correspond to the elements of P, two elements are connected by an edge if one of them covers the other, and furthermore the covering element is placed above the horizontal line determined

by the covered element. Hasse diagrams are easy to draw and are of great help in theoretical considerations as well as in interpreting empirical findings.

2. Special Classes of Orders

There are numerous specializations, that is, additions of further properties, such that the resulting structure is something in between a poset and a linear order. In this section, two classes are considered which have useful numerical representations, although not the strong one of Eqn. (1). The third specialization considered here, namely lattices, is important because it tends to bridge the gap between ordering relations and algebraic operations.

2.1 Semiorders

In psychophysics one has to deal with the phenomenon of subliminal changes in intensity of physical stimuli. One stimulus may be (physically) slightly more intense than another, but a subject does not perceive the difference. For example, one might imagine a coffee-tester confronted with an extended series of cups of coffee, each with one more grain of sugar than its immediate predecessor. He/she will probably not notice the difference in taste between successive cups, but surely the difference between the first and the last cup will be clearly noticeable. In other words, indifference with respect to $\leqslant$ is not transitive. Similar observations can be made in decision theory and other branches of psychology. One way of describing such facts is to introduce sensory thresholds, thereby creating a scaling problem where one wants to scale both the subjective intensity and the magnitude of the threshold. Luce (1956) considered two conditions which seem plausible in this context (P is the set of stimuli and $x \leqslant y$ denotes the observation x is less intense than y):

(S1) If $x \leqslant y$ and $u \leqslant v$, then $x \leqslant v$ or $u \leqslant y$.

(S2) If $x \leqslant y \leqslant z$ and $w \in P$, then $w \leqslant z$ on $x \leqslant w$.

One easily shows that $(P, \leqslant)$ is a quasiorder, i.e., without losing generality, one can assume that it is a partial order. Condition (S1) says that it does not contain Fig. 1(a) as a suborder. Similarly, condition (S2) is equivalent to the fact that it does not contain Fig. 1(b) as a suborder.

From the viewpoint of measurement the most important property of semiorders is the possibility of a numerical representation with a constant threshold. More precisely, one can show that for a finite semiorder $(P, \leqslant)$ there is a function f

$$f: P \to \mathbb{R}\text{: such that } x < y \text{ if } f(x) + 1 < f(y). \quad (2)$$

The condition of finiteness is far too restrictive for Eqn. (2), but without further restrictions it does not obtain. The height of the threshold is arbitrary, that is, the number one in Eqn. (2) can be replaced by any other positive number (with a rescaled f). More on this kind of representation theorem can be found in Roberts (1979, Chap. 6.1) or Suppes et al. (1989, Chap. 16). Clearly, the converse of this theorem is true: a structure with a representation, Eqn. (2) is necessarily a semiorder.

Although the semiorder concept seems to have its origin in the psychological context of the possibility of measurement of subjective intensity described above, it is now a well-established notion in the mathematical theory of posets and several remarkable results are connected with it. Some of them are mentioned in the following text.

2.2 Interval Orders

A structure satisfying only condition (S1) but not necessarily (S2) is called an interval order. This notion derives from the numerical representation which is characteristic of this kind of order. Each element can be represented by an interval on the real line and $x < y$ or $y < x$ is tantamount to the nonoverlapping of the respective intervals. Put another way, a scale value and a threshold are attached to each element, but the height of the threshold can vary in contrast to Eqn. (2). The reader will find many interesting properties and proofs in Fishburn (1985).

Historically, it is of interest that, although the representation theorem is due to Fishburn (1970), the investigation of this structure goes back to papers of Norbert Wiener, early, in the twentieth century. What is even more remarkable is that in this work he mentioned possible applications in experimental psychology (see Fishburn and Monjardet 1992).

The theory of interval orders has applications in archeology in connection with the seriation problem. In this case P is a set of artifacts, each element of which was some time in use. From excavations in various historical sites it is known for any two artifacts whether the period of use of one follows that of another with or without overlap. If this relation satisfies (S1) then, by the interval order representation theorem, it follows that it is possible to assign a time interval to each artifact, preserving the empirical observations.

A related notion is that of a biorder. The primitives of a biorder consist of two sets P and Q and a relation ρ between P and Q, i.e., a subset $\rho \subseteq P \times Q$ satisfying for all $x, y \in P$ and $u, v \in Q$

$$x\rho u,\ y\rho v, \text{ implies } x\rho v \text{ or } y\rho u. \quad (3)$$

To understand this condition, it is helpful to think of P as a set of subjects and of Q as the items of an

achievement test. $x\rho u$ means 'subject x fails to solve item u.' The premise of Eqn. (3) says that x fails u and y fails u: if one assumes that x solves v correctly, then y has to fail u. With this interpretation, the biorder concept is the basis of Guttman scaling. Probabilistic modifications of it underly Rasch scaling, which is well-established in the theory of psychological tests. Thus, with the biorder condition Eqn. (3), a testable property is formulated which, when satisfied, justifies a numerical difficulty value for items because a numerical representation of a biorder consists of two functions f: $P \to \mathbb{R}$ and g: $Q \to \mathbb{R}$ such that:

$$x\rho u \text{ if } f(x) < g(y).$$

This, in turn, implies Eqn. (3). In the terminology of the above interpretation of a biorder $f(x)$ is the ability of subject x and $g(u)$ is the difficulty of item u. If in a biorder, the two sets coincide, $P = Q$, then Eqn. (3) reduces to (S1), i.e., the structure is an interval order.

2.3 Lattices

In a partial order $(P, \leqslant)$ subsets of P may or may not have upper or lower bounds, where upper bounds are elements which dominate all elements of the subset and lower bounds are elements which are dominated by all its elements. If each two-element subset has a unique smallest upper bound—called the supremum—and a unique greatest lower bound—called the infimum—then the structure is called a lattice. The infimum of x and y is denoted by $x \wedge y$ and the supremum by $x \vee y$. Examples of lattices and non-lattices are given in Fig. 2.

Further examples are the power set of a set ordered by set inclusion, any subset of the reals naturally ordered, and the natural numbers $\{1,2,3, \ldots\}$ ordered by divisibility, i.e., $m \leqslant n$ if m divides n. In the last example, the least common multiple and the greatest common divisor are the supremum and the infimum, respectively. It follows easily by induction that any finite subset of a lattice P has an infimum and a supremum. However, infinite subsets do not necessarily have this property. If all subsets of P have it, then the lattice is called complete. Thus, finite lattices are always complete, but the rationals with the natural ordering are not. The question of completion of lattices similar to the construction of the reals by Dedekind cuts is discussed in Sect. 5.

There is a more algebraic way of characterizing lattices, where infimum and supremum are regarded as binary operations on P:

Theorem 1 Let $(P, \leqslant)$ be a lattice. Then for all $x, y, z \in P$:

(a) $x \vee (y \vee x) = (x \vee y) \vee z$ and $x \wedge (y \wedge z) = (x \wedge y) \wedge z$

(b) $x \vee y = y \vee x$ and $x \wedge y = y \wedge x$

(c) $x \wedge (x \vee y) = x$ and $x \vee (x \wedge y) = x$.

Conversely, if a structure $(P, \vee, \wedge)$ is given with operations $\vee$, $\wedge$ satisfying the conditions (a), (b), and (c), then $\leqslant$ is defined by

$$x \leqslant y, \text{ if } x \wedge y = x$$

is a partial order on P such that $(P, \leqslant)$ is a lattice.

Lattices play an important role in pure and applied mathematics. John von Neumann, one of founders of lattice theory, developed his formulation of quantum mechanics in terms of lattices. Today they are used in many branches of mathematics—for instance, in combinatories, probability theory, logic, and functional analysis. The wide applicability of lattices in technical and social science contexts is best demonstrated by the development of 'concept analysis' by Rudolf Wille (see Ganter and Wille (1996).

3. Extension of Orders; Szpilrajn's Theorem

Is it possible to extend a partial order by adding (x, y) or (y, x) to $\leqslant$, when x and y are incomparable in $\leqslant$ and at the same time keeping the extended relation a partial order? Is it possible to extend $\leqslant$ to a linear order in this way? Both questions are answered in the affirmative by the following theorem.

Theorem 2 (Szpilrajn's Theorem) Let $(P, \leqslant)$ be a poset and $x, y \in P$ such that neither $x \leqslant y$ nor $y \leqslant x$. Then there are two linear orders $\leqslant'$ and $\leqslant''$, each of which contains $\leqslant$ as a subset, and such that $x \leqslant' y$ and $y \leqslant'' x$.

The proof of this theorem is not difficult for finite P, but entails a set theoretic twist in the infinite case. If x and y are incomparable in $\leqslant$ then one defines $x \leqslant' y$, adds all pairs to $\leqslant'$ which this choice and transitivity forces and goes ahead until no incomparable pairs are left over. The latter step of this argument is a bit of a problem for an infinite P.

Sometimes one can regard a partial order as ensuing from incomplete data collection, incomplete in so far as a linear order can be expected, but the available data give only a part of the full order information. Then the question of a linear extension is relevant, in particular, one can ask how many additional pairwise comparisons must be performed to obtain the full linear order, which is the most parsimonious way to find such an extension.

For these problems the number of linear extensions is of importance. It was shown by Fishburn and Trotter (1992) that for all partial orders with fixed $|P| = n$ and k pairs in $\leqslant$ those with a maximal number of linear extensions are semiorders. For $k \leqslant n$ these semiorders are characterized, but for larger K there seem to remain several unsolved problems.

It can be shown that a linear extension can be found with at most K questions of the form 'is x larger than

y,' where K is of the magnitude of the logarithm of the number of linear extensions (see Trotter 1995 for details). In connection with this problem it seems that a pair x, y for which many extensions with $x \leqslant y$, as well as many with $y \leqslant x$ exist is best to investigate first because, whatever the answer, it considerably reduces the number of data which remain to be collected. In this context the so called 1/3–2/3 conjecture was made. It says that each poset which is not a linear order contains a pair (x, y) for which the relative frequency of extensions with $x \leqslant y$ is between 1/3 and 2/3. This conjecture seems still open. However, bounds are known, e.g., the relative frequency is between 3/11 and 8/11. The proof of this result draws on a geometric inequality and is extremely beautiful. For semiorders the 1/3–2/3 conjecture is true. More on this topic and some of the proofs can be found in Trotter (1995).

4. Order Dimension

A corollary of Szpilrajn's Theorem is that each partial order is the intersection of its linear extensions. One can ask for the minimal number of linear extensions with this property and regard this number as a measure of complexity of the order. It is called the dimension of $(P, \leqslant)$.

All orders on three, four, and five elements are two-dimensional (provided they are not linear). There are two orders with six elements of dimension three. One can group the posets in classes of increasing dimension. The semiorders have a dimension less than or equal to three, but the interval orders can have arbitrary large dimension.

One classical result on the dimension is that it is always smaller than or equal to the maximal number of pairwise incomparable elements. For $|P| \geqslant 4$ ($|P|$ denotes the number of elements in P) the dimension is bounded above by $|P|/2$ and this is the best possible solution. It is an active area of research to find inequalities such as these which relate the dimension to other characteristics of the poset (see Trotter (1992)).

There are other variants of the dimension concept: one can replace linear orders by semiorders (and obtain the semiorder dimension) or by interval orders or by any other class of partial orders. For biorders, see Doignon et al. (1984) and Koppen (1987). These two papers attempt to link this dimension concept to techniques of extracting factors to explain data. In this way, an ordinal factor analysis is outlined. Factor analytic procedures used to be widely spread in social science research, but until now they seem to lack a sound axiomatic foundation. The order theoretic dimension concept can provide a weaker concept of a factor which is, on the other hand, much easier to justify, or empirically to test for adequacy.

5. Embeddability and Completion of Orders

Sometimes the nonexistence of suprema and infima is a nuisance; just think of the inconvenience caused when a sequence of rational numbers 'converges,' but the limit does not exist (in the rationals). One tries to avoid this situation by embedding a given poset into a complete lattice which contains the original order as a suborder. In contrast to the techniques in Sect. 3, one adds new elements to P instead of new pairs to $\leqslant$. Thus, the aim is to find a superstructure which resembles the given order as closely as possible, but which can be handled more conveniently because of the existence of suprema and infima. The prototype of this kind of extension is the time-honored construction of the reals from the rational numbers.

To begin with, some notation must be introduced. For $X \subseteq P$ let:

$$X^{\nabla} := \{y \in P\colon\ y \leqslant x \text{ for all } x \in X\}$$
$$X^{\Delta}\colon = \{y \in P;\ x \leqslant y \text{ for all } x \in X\}.$$

Thus X^{∇} is the set of lower bounds of X and X^{Δ} the set of upper bounds of X. The sets $\{x\}^{\nabla}$ and $\{x\}^{\Delta}$ for $x \in P$ are abbreviated by x^{∇} and x^{Δ}, respectively. One simple way of embedding consists of considering

$$\nabla_{(P,\leqslant)}\colon = \{x^{\nabla}; x \in P\}$$

the so called principal ideal representation. Note that $(\nabla_{(P,\leqslant)}, \subseteq)$, is a suborder of the complete lattice of the power set, isomorphic to $(P, \leqslant)$. However, this embedding is not sufficient for the purposes of this section; the superstructure carries too little information from the original one. Therefore, a more parsimonious extension must be constructed.

To this end, a pair (X, Y) is called a MacNeille cut if and only if $X = Y^{\nabla}$ and $Y = X^{\Delta}$. The set of all MacNeille cuts is a better candidate for an embedding. Indeed, one can prove:

Theorem 3 Denote by $C_{(P,\leqslant)}$ the set of all MacNeille cuts of a partial order $(P, \leqslant)$. Define for two cuts (X, Y) and (X', Y') the relation

$$(X, Y) \sqsubseteq (X', Y') \textit{ if } X \subseteq X'.$$

Then $(C_{(P,\leqslant)}, \sqsubseteq)$ is a complete lattice. Moreover, it is the unique (up to isomorphism) smallest complete lattice which contains the original poset $(P, \leqslant)$ as a supremum dense subset, i.e., each element of $C_{(P,\leqslant)}$ is supremum of elements of $(P, \leqslant)$.

A few remarks should illustrate this theorem. First, the set corresponding to $(P, \leqslant)$ in $(C_{(P,\leqslant)}, \sqsubseteq)$ is the set of cuts (x^{∇}, x^{Δ}). Second, if this theorem is applied to the rationals with the usual (linear) order, then the reals with added $\{-\infty\}$ and $\{\infty\}$ result as the MacNeille completion. The element $-\infty$ is the cut $(\emptyset, P)$ while ∞ is the cut $(P, \emptyset)$. Third, although the notion of a complete lattice only bears meaning for

infinite base sets P, the MacNeille completion makes sense also for finite posets. Figure 2(b) shows the MacNeille completion of (a) and Fig. 2(e) of (d). Finally, it is worth mentioning that the mappings $X \to X^{\nabla}$ and $X \to X^{\Delta}$ form a pair of functions that are known as a Galois connection. This construction is often used when dealing with posets; it is applicable in other contexts, for example, formal concept analysis (see Ganter and Wille (1996) is based on a Galois connection). In this respect, concept lattices are a generalization of a MacNeille completion.

See also: Majorization and Stochastic Orders; Measurement, Representational Theory of; Order Statistics; Ordered Relational Structures

Bibliography

Doignon J-C, Ducamp A, Falmagne J-C 1984 On realizable biorders and the biorder dimension of a relation. *Journal of Mathematical Psychology* **28**: 73–109

Fishburn P C 1970 Intransitive indifference with unequal indifference intervals. *Journal of Mathematical Psychology* **7**: 144–9

Fishburn P C 1985 *Interval Orders and Interval Graphs*. Wiley, New York

Fishburn P C, Monjardet B 1992 Norbert Wiener on the theory of measurement (1914, 1915, 1921). *Journal of Mathematical Psychology* **36**: 165–84

Fishburn P C, Trotter W T 1992 Linear extensions of semiorders: A maximization problem. *Discrete Mathematics* **103**: 25–40

Ganter B, Wille R 1996 *Formale Begriffsanalyse*. Springer, Berlin

Koppen M G M 1987 On finding the bidimension of a relation. *Journal of Mathematical Psychology* **31**: 155–78

Luce R D 1956 Semiorders and a theory of utility discrimination. *Econometrica* **24**: 178–91

Roberts F S 1979 *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*. Addison-Wesley, Reading, MA

Suppes P, Krantz D H, Luce R D, Tversky A 1989 *Foundations of Measurement, Vol. 2: Geometrical, Threshold, and Probabilistic Representations*. Academic Press, New York

Trotter W T 1992 *Combinatories and Partially Ordered Sets: Dimension Theory*. Johns Hopkins University Press, Baltimore, MD

Trotter W T 1995 Partially ordered sets. In: Graham R L, Grötschel M, Lovász L (eds.) *Handbook of Combinatories*. Elsevier, Amsterdam, pp. 433–80

R. Suck

Participant Observation

Participant observation has been part of the arsenal of methodological techniques employed by social scientists since around 1900, notably within the discipline of sociology. Participant observation involves the active engagement of the researcher with the members the community that he or she wishes to study, typically as an equal member of the group.

This methodology falls under the general rubric of field research. It is important to recognize that there exist alternative styles of 'field research': a range of possibilities that this chapter will not cover. However, in contrast to the manner in which observational research has traditionally been conducted in anthropology, in sociology much ethnography is based on the assumption that the researcher is at least nominally a participant in the group being studied. Yet, the line that separates participant observation and other forms of ethnographic observation is uncertain and hazy, and for some scholars the two terms have been used as synonyms.

Since the mid-1970s there has been a notable increase in the frequency and the legitimacy of the variety of field methodologies within the social sciences. These qualitative, interpretive strategies, once considered 'unscientific,' have demonstrated their utility in a number of disciplines. That this occurred during a period in which social scientific funding was being decreased is surely no coincidence. Although at the beginning of the twenty-first century qualitative field methods do not have the same standing as quantitative methods (especially in psychology and economics), their growth is astonishing.

1. Advantages

Several distinct advantages justify participant observation; among these are the benefits of rich data, validity, *Verstehen*, and economy.

1.1 Richness

This methodology, in contrast to most methods that do not involve personal witnessing, provides for rich and detailed data. To investigate action, direct observation is essential. When this is coupled with the questioning of actors about their choices, tied to some measure of active involvement, the value of participant observation is evident. The form of data collection permits an intense depiction that produces a fullness of understanding.

1.2 Validity

A second benefit is analytical validity. Because the observations are of behavior *in situ*, the researcher can rely upon the claim that the findings are close to the 'proper' depiction of the scene. If the conclusions drawn from participant observation research are not always reliable, in that other investigators might reach different conclusions, they do reflect the belief that some aspect of the world as it 'truly' is, is being described.

1.3 Interpretive Understanding

Participant observation supports the demand of Max Weber to produce research that is characterized by *Verstehen* or a personal understanding. In this way participant observation with its emphasis on both participation and observation adds to research knowledge. By directly involving the researcher in the activity, one can understand on an immediate level the dynamics and motivations of behavior. The observer becomes other than an outsider. While research projects differ significantly on this dimension, involvement has advantages for a *verstehende* analysis that other approaches can not match.

1.4 Economy

Participant observation research is typically inexpensive. While this research strategy is surely labor-intensive, it is not capital-intensive. In many cases the researcher is the only member of the project, and can set the terms of his or her own involvement. Increasingly, however, team-research projects exist as well, and in these cases wages and benefits for hired observers may involve considerable costs. Yet, in the standard participant observation project the key resources consist of a technology for inscribing observations. Perhaps for reasons of cost, connected to the time and energy involved, participant observation has proven to be a technique most typically engaged in by scholars at the early stages of their academic careers.

2. Disadvantages

Just as there are significant advantages to this methodology, disadvantages are evident. Problems relate to proof, generalizability, bias, and time commitments.

2.1 Proof

Participant observation relies upon a single case study: the examination of one place. This raises questions about the nature of proof, or, put another way, about reliability. Will two researchers examining the same or similar social scenes reach the same conclusions? Often because of different perspectives upon entering the field and different experiences within the field, findings are sharply distinct. While the observations and interpretations of those observations may be compelling, one can reasonably wonder whether any set of conclusions is definitive.

2.2 Generalizability

Even if we accept the legitimacy of analyzing one scene, on what grounds can we generalize beyond that setting? How far can our conclusions be pushed? Participant observation research has a problem in this regard because of the absence of scientific control that characterizes experimental research and produces confidence in the claim that important variables of social life have been adequately captured. As a result, the extent to which generalizability is legitimate is problematic in participant observation. Participant observers need to present a theoretical model that helps readers to judge the legitimacy of their broader claims in light of the audience's own experiences.

2.3 Bias

A strength of participant observation methodology is that the researcher's insight and perspective is taken into account, but this strength has a downside. One cannot adequately distinguish between perspective and bias. The background that the researcher brings to a scene can be distinctively different from other researchers, and, for that matter, from the perspectives of the participants in the setting. To the extent that the researcher's perspectives differ significantly from the perspectives of the participants—possible because of the generally progressive values and upper middle class status of academics—the understanding of a particular scene may be systematically biased.

2.4 Time

Just as participant observation research is relatively inexpensive, it is also highly labor intensive. This form of research requires that the researcher be present in the observed social scene. One cannot simply fly in and out, but must spend sufficient time so that the full range of activities in which participants engage are noted. Much participant observation depends upon chance—what happens to occur at the moment of observation—and, as a result, a significant investment of time is needed. While there is no definitive rule for the proper length of time necessary for observation, most projects require months, if not years, to complete. This, coupled with a modest requirement for capital equipment support means that, as noted, this methodology is particularly appropriate for younger scholars. This reality can mean that participant observation studies often do not have the depth of theoretical understanding that more likely characterizes the work of senior scholars.

3. The Growth and Development of Participant Observation

Although participant observation methodology has deep roots in ethnographic traditions (including in traveler's tales and missionary tales), participant

observation has its origins in the discipline of sociology at the turn of the twentieth century, particularly with scholars affiliated with the Department of Sociology at the University of Chicago. Notable in this early stage was the participant observation studies of Annie Marion MacLean, a student at Chicago and later professor at Adelphi University. MacLean worked and observed in department stores, sweatshops, and coalfields, in order to gain an appreciation of conditions in these scenes. By the 1920s participant observation was well established in Chicago, as evidenced in studies by Frances Donovan on the lives of waitresses and Frederic Thrasher's observations of gang life. Although members of the 'Chicago school' of sociology employed numerous methodological techniques, the group was probably best known for their observational research and life stories. By mid-century participant observation had expanded well beyond the bounds of Chicago. Works such as William Foote Whyte's *Street Corner Society* (1943), August Hollingshead's *Elmtown's Youth* (1949), Arthur Vidich and Joseph Bensman's *Small Town in Mass Society* (1958), and Herbert Gans's *The Urban Villagers* (1962), collectively demonstrated that through involvement in a community, the researcher could gain a perspective that was both empirically and theoretically powerful. The fact that many of these research projects covered life in city environments led some to describe this approach as 'urban ethnography'. Another stream of 'participant observation' research was grounded in the study of deviant groups and activities. Researchers in such domains have the dilemma of determining the degree to which participation should bolster their observation. Exemplifying such research are studies of the mentally ill, gang members, and drug dealers, notably the classic works by Erving Goffman *Asylums* (1961), and Elliott Liebow, *Tally's Corner* (1967). These two themes —urban life and social deviance—have been integral to participant observation, and despite reduced interest in the late 1980s and early 1990s, scholars, such as Elijah Anderson, Mitchell Duneier, David Snow, and Carol Stack have returned to these themes, creating a 'new urban ethnography' that focuses on issues of urban poverty and racial injustice.

Not all researchers focus on 'problem domains.' Beginning with the inspiration of the Hawthorne studies, and later the research of Donald Roy and Melville Dalton, increasing attention has been paid to observations of work places. The participant observation of organizational life has developed rapidly, creating an active investigation of organizational domains, both in traditional social science departments and within schools of management. Factories, airlines, restaurants, and high tech engineering firms have all been the focus of research. Yet, all social worlds in which individuals with a common interest meet can be analyzed through participant observation, whether clubs, sports teams, or bars.

4. *Comparison with Other Field Methods*

It is easy to assert that participant observation is synonymous with qualitative research, and, in reality, the lines are blurred. However, participant observation can be contrasted with several other techniques with which it shares some commonalties. Some scholars engage in 'ethnographic immersion.' In such investigations, the researcher becomes a full member of the group, and, in effect, studies his or her own group. This has the advantage that the researcher comes by his or her 'member's knowledge' naturally, but simultaneously the immersion may diminish the ability to examine the group life from the critical perspective that being marginal sometimes brings. Most participant observation maintains a distinction between the researcher and the full member of the group studied. Typically members are told that they are being observed—an outsider, who, for purposes of the study, may participate actively in the life of the group, even though she or he is not a full member.

A second technique, related to ethnographic immersion, has been labeled 'auto-ethnography,' akin to the methodology of introspectionism of the early years of the twentieth century. In these studies, the researcher mines his or her life experiences, hoping to demonstrate the embodied characteristics of social forces. One such example is Carolyn Ellis's emotional and analytical account of her own reaction to the lengthy dying of her significant other. This writing, while occasionally powerful, insightful, and moving, can be faulted for its reliance on subjective experience. Auto-ethnography might be seen as an extreme instance of total immersion, with nothing outside the researcher's experience; given this, such arguments become difficult for others to evaluate. The author proposes a singular account of her emotions that because it is her account leaves no room for alternative interpretations.

A third, related approach is to collect data by means of in-depth interviews. Most participant observation involves questioning of informants. However, some research relies on this questioning to the exclusion of direct observation. This method has the advantage of permitting the researcher to gain multiple perspectives of the same scene and permits the investigation of domains to which it might be impossible to gain direct access. However, in-depth interviews limit the researcher to the presentations of interested parties. The rhetoric in interviews may be systematically different from the private thoughts of the informant and may also differ from their behavior. Observations provide a check on what one is told for reasons of impression management. As always, triangulation of data provides advantages, curbing the weakness of any one methodology.

Fourth, some researchers de-emphasize participation, and are, in practice, pure observers. Some observers make precise statistical measurements of

public behaviors, but this approach is distinct from participant observation. More relevant are those scholars who examine public behavior, particularly behavior within cities. This distinguished tradition of scholarship includes such researchers as William H. Whyte, Erving Goffman, Lyn Lofland, and, more recently, Carol Brooks Gardner. A close reading of street activity, perhaps parallel to certain forms of animal ethology, permits the recognition of patterns of human activity.

Of course, within participant observation, researchers select various roles or strategies. The key dimension on which roles can be differentiated is the extent to which members are involved within group life. Adler and Adler (1987) divide participant observation methods according to the extent of participation: peripheral membership, active membership, and complete membership. In general, the peripheral member observes as an outsider to the group and its culture, the active member is directly engaged in the life of the group, while making it clear to the group that she or he does not fully belong and is conducting research. In the model of the Adlers, the complete observer essentially engages in ethnographic immersion, sometimes informing the group of the research goals.

5. *Strategies of Participant Observation*

Unlike many techniques of quantitative research and statistical analysis, there are no firm and fast rules that determine how participant observation research should be conducted. Such a reality has proven frustrating to generations of students. Traditionally a novice learned the skills of participant observation through what amounted to an apprenticeship relation with a senior researcher. However, because of the increasing numbers of field researchers, many social science departments, notably in sociology and anthropology, have instituted field methods courses in which some rudimentary 'rules' and strategies of the method are taught.

5.1 *Access*

Often the first issue that must be confronted in a research project is the question of access. How does one convince a group that they should open their space to an outsider who may criticize the group? This is surely applicable to groups that are widely perceived as deviant (criminal gangs or politically marginal movements), but it also recognizes that any group, striving to maintain boundaries that differentiate it from others, may engage in some activities that members wish to keep hidden from outsiders and which could be viewed negatively. Participant observers must, within the limits of their ethical standards, convince the group that they can be trusted. Through this process they set 'the rules of engagement.' Such rules, explicit or implicit, differ for each project. Further, this access involves an ongoing temporal process. One is rarely accepted all at once, but often the group maintains a cautious vigilance until they become satisfied that the researcher can be trusted. The participant observer must be sensitive to this personal equation.

5.2 *Socialization*

Access is only a first step. Once the researcher gains entrance to a setting he or she must determine what behavior means. In effect, the participant observer must become socialized to the environment with its norms, cultural traditions, and jargon. The ease of this process varies as a consequence of how well acquainted the researcher is with the setting prior to the observation, but unless the researcher begins as a full member, a process of acculturation must typically be mastered. At times, the group may become frustrated by the lack of competence of the observer, but more often this process of learning is valuable in methodological terms, permitting the researcher to ask questions that would be denied to others as being too obvious.

5.3 *Developing Relationships*

Participant observation is a methodology that depends upon the establishment of relationships. It relies upon sociability. One essential skill is the ability to recruit informants from the field. The most important of these are termed 'key informants,' such as Doc in William Foote Whyte's (1943) classic *Street Corner Society*. These individuals serve both a guides for the participant observer and as a means of vouching for the researcher's legitimacy within the setting, providing an insider's stamp of approval. The researcher who lacks the ability to make connections will have difficulty collecting credible research data.

5.4 *Field Notes*

The researcher does not only need to observe activity, but must subsequently inscribe that material in 'field notes,' documents that serve as the evidentiary basis of the published report. Considerable discussion has addressed what constitutes appropriate field notes and how these notes should be presented (Van Maanen 1988). These notes provide the 'facts' that are observed, including actions, statements, and the feelings of the observer. Participant observers are enjoined to write as much of what transpires as possible, particularly in the early stages of the research, when one is determining what issues are most salient. This empha-

sizes that participant observation is often an inductive methodology, with the field data revealing the important issues, rather than testing pre-established hypotheses.

5.5 Exit Strategies

Much research simply ends. Once completed, the project closes down. Exiting is not as simple in participant observation because the researcher and informants have developed social ties. Leaving the field involves the termination of a set of powerful relationships, and sometimes produces loneliness or even betrayal. As a result, the process of disengagement often involves a longer period, and some connection with the field setting continues. Researchers are often advised—both for practical and ethical reasons—to share material with their informants, either to gain additional feedback or to permit those informants to use the analysis for their own ends. The linkages that are established during the course of the observation can be a continuing source of strength, both for researchers and for informants.

6. Ethical Concerns

Although all forms of research involving human subjects raise ethical issues, the variety of ethical concerns in participant observation are particularly great because of the degree of interpersonal dynamics involved. Four ethical issues have been central to participant observation research: deception, informed consent, confidentiality, and precision of depiction. Although these do not exhaust potential ethical topics, they raise important concerns.

6.1 Deception

Most participant observers agree that it is both pragmatically and ethically necessary to announce to one's informants that one is conducting research. In part, this is because participant observation depends on the personal relations among individuals, who, for that period of time, are status peers. To claim to be something that one is not—to pretend to be a full member of the group—without revealing one's purpose could undermine the entire enterprise if it becomes discovered. Lies are often difficult to maintain for long periods. Further, deception suggests that one's interests take priority over the interests of another: can we justify such trickery? While research on groups that wish to keep their activities hidden—either deviant groups or, occasionally, elite groups—encourage deceptive practices, such strategies are troublesome.

6.2 Informed Consent

Deception involves the active misleading of others, but how should we view the absence of full disclosure? Do group members have the right to know that they are being investigated? For instance, one can easily join a leisure group without announcing that one is taking notes or planning to publish. Is this proper? As noted, good reasons may exist for announcing one's presence, permitting one's informants to serve, in effect, as research assistants. But should researchers be required to gain the 'informed consent' of their informants; further, how much information is necessary for informed consent, particularly in a research methodology that depends on inductive methods, in which the researcher may truly not be aware of what issues will eventually become central? One's topics depend on what one observes. During the past quarter century, American universities have instituted 'ethics' committees (Human Subjects Committees or Institutional Review Boards), designed to protect the 'rights' of research subjects. Frequently researchers are required to gain some measure of informed consent, but how much information is required remains a matter of debate.

6.3 Confidentiality

One element that most (although not all) researchers agree on is that informants' names and clearly identifying information not be included in any published research report, although some who critically examine elites feel that confidentiality is not essential. In general, researchers typically believe that publishing the names or identifiers of particular informants or settings is neither necessary nor warranted. At times this decision poses a problem in that field notes may describe actions or statements that can be recognized, certainly by others within the scene. As a result, participant observers must choose how much information to alter to preserve confidentiality, even when this shades the full 'facts' of the case.

6.4 Precision of Description

Although it is clearly desirable to present data that is precisely accurate—direct quotations and exact descriptions of behavior—participant observation methodology and the limits of human ability make these values unattainable goals (Fine 1993). In part, we are imperfect in inscribing reality, but in part we may recognize that the details of human activity may actually obscure our crucial analytical points. For instance, false starts, errors, and grammatical infelicities mark the speech of informants, like the speech of researchers themselves. Further, so much behavior is happening at any given moment that the researcher can easily become overloaded with trivial observa-

tions. The ethical balance that must be struck is to do justice to the events that are transpiring, while simultaneously making the account as clear as possible to permit communication with readers.

7. Summary

This brief overview has examined central features of the participant observation methodology. In doing so, it de-emphasized both the strategies of analysis (the extended case study method, grounded theory) used by practitioners and the modes of presentation (realist tales, impressionist tales). No method of data collection can exist without related techniques by which data are analyzed and presented.

Throughout the twentieth century, participant observation proved to be a powerful technique through which researchers understood a diverse set of social worlds in a way that did justice to the complexity of the activities of group members and to their own perspectives and understandings. While this methodology benefits by the triangulation with other styles of research to create adequate theoretical understanding, participant observation provides the basis by which researchers can understand the richness of the social settings that they—and others—inhabit.

See also: Field Observational Research in Anthropology and Sociology; Fieldwork: Ethical Aspects; Fieldwork in Social and Cultural Anthropology; Observational Methods in Human Development Research; Observational Studies: Outcome-based Sampling; Observational Studies: Overview

Bibliography

Adler P A, Adler P 1987 *Membership Roles in Field Research*. Sage, Thousand Oaks, CA
Fine G A 1993 Ten lies of ethnography. *Journal of Contemporary Ethnography* **22**: 267–94
Gans H 1962 *Urban Villagers*. Free Press, Glencoe, IL
Goffman E 1961 *Asylums*. Anchor, New York
Hollingshead A 1949 *Elmtown's Youth*. Wiley, New York
Liebow E 1967 *Tally's Corner*. Little Brown, Boston
Van Maanen J 1988 *Tales of the Field*. University of Chicago Press, Chicago
Vidich A, Bensman J 1958 *Small Town in Mass Society*. Princeton University Press, Princeton, NJ
Whyte W F 1943 *Street Corner Society*. University of Chicago Press, Chicago

G. A. Fine

Participation: Political

Political participation refers to the activities of the mass public in politics, including, for example, voting in elections, helping a political campaign, giving money to a candidate or cause, writing or calling officials, petitioning, boycotting, demonstrating, and working with other people on issues. Political participation figures in philosophical discussions of democracy and representative government since ancient times because it provides the means for citizens to communicate their views. Normative interest in the topic remains vigorous (see *Democratic Theory*). This article focuses upon the empirical study of political participation. Scholars look for systematic patterns that explain why some individuals are inactive and why others choose certain activities. The patterns reflect the political context and both reflect and affect the structure of political power.

Participation provides the link from the mass public to the political elite who are presumed—and sometimes shown—to respond (see *Political Representation*).

1. Differing Definitions of Participation

Scholars differ in their definition of political participation. The currently dominant view limits participation to actions that might affect others, as in the following definitions. 'By political participation we mean activity by private citizens designed to influence government decision-making' (Huntington and Nelson 1976, p. 3). 'By *political* participation we refer simply to activity that has the intent or effect of influencing government action—either directly by affecting the making or implementation of public policy or indirectly by influencing the selection of people who make those policies' (Verba et al. 1995, p. 38). The restriction to private citizens is meant to exclude from the concept activity undertaken in their official capacity by those for whom politics and governing are a vocation. A few scholars (cf. Milbrath 1965) include political involvement and activities to support the regime as participation. For these scholars, reading about politics is also political participation, while under the dominant definition it is not (since it does not have direct effects on others).

Another definitional debate in the field concerns the extent to which actions that have the effect of influencing policy without that being the actor's intent also count as political participation. Adopting the more common view, Verba et al. (1995, pp. 38–39) restrict their study to voluntary activity, which they define as follows: 'By voluntary activity we mean participation that is not obligatory—no one is forced to volunteer—and that receives no pay or only token financial compensation.' In contrast, Huntington and Nelson (1976, p. 7) explicitly include not only '*autonomous*' participation but also '*mobilized*' participation, 'defined as' 'activity that is designed by someone other than the actor to influence governmental decision-making'. Under that definition, they would include as a

participant a worker who attends a rally because his boss threatens to demote him otherwise, even if the worker has no intent of affecting government. In practice, it can be difficult empirically to distinguish mobilized from autonomous activity. For answering questions about the impact of participation upon political outcomes, the two are equivalent. The distinction is useful, however, when one focuses on the motives for participation or the impact of participation upon the actor.

2. *The Study of Participation*

2.1 *Participation as Electoral and Hierarchical: Early Studies*

The modern empirical study of political participation began developing in tandem with the modern study of electoral behavior using sample survey data. Election outcomes depend not only upon people's preferences but also upon which citizens choose to exercise their preference at the polls. Partially reflecting that link, many of these studies truncated political participation to electoral participation, that is, voter turnout and occasionally some campaign activities (see *Voting: Turnout*). Much empirical work was done in the United States, and much of that was connected to the Michigan Survey Research Center, especially under the leadership of Warren Miller and Philip Converse. These investigators also collaborated with colleagues outside the United States, especially in Europe, producing further studies addressing participation, especially voting. Under the leadership of Stein Rokkan, Scandinavian scholars produced a substantial body of work on participation as part of their more extensive studies of political behavior (see for example Rokkan 1970). These studies painted a clear sociodemographic and psychological picture of voters (and occasionally of participants in campaigns). In contrast to nonparticipants, participants in general had more money and education, were more interested in politics, more partisan, more efficacious, more likely to be men, more involved in organizations, and more rooted in their community. Some, including Rokkan, also paid close attention to the relationship between political behavior, including participation, and the citizen's position in the social cleavage structure.

Milbrath summarized the findings to date in 1965. He included political support as participation as well as electoral activities beyond voting. He promoted the influential organizing idea that participation lay on a hierarchical scale—there was a ladder of activities, with the easiest at the bottom and the most difficult at the top. 'Persons who engage in the topmost behaviors are very likely to perform those lower in rank also' (Milbrath 1965, pp. 17–18). Given the emphasis on electoral activities and given the frame created by the ladder concept, levels of participation in the United States were perceived as low.

2.2 *Large-scale Cross-national Studies*

Interest in democratic stability prompted cross-national projects that took political participation as an important variable. For Harry Eckstein, participation provided the link upwards from subordinates to superordinates (Eckstein and Gurr 1975, pp. 60–7). For Almond and Verba in *The Civic Culture* (1963), the participatory behavior of citizens in a country was one important component in characterizing their political culture, but they were just as interested in attitudes and beliefs about hypothetical behavior as in activity. Nonetheless, the data, collected in 1959 and 1960 in five countries (the United States, Mexico, Italy, West Germany, and Great Britain) provided the first major database used for cross-national comparisons of political participation, especially beyond voting turnout. Some of these analyses contributed the important conclusions that persons involved in organizations participate more in politics and that societies with denser organizational structures exhibit higher levels of participation.

Within the decade (1966 to 1971) Verba was in the field again, with an international team of collaborators, studying political activity. Surveys were completed in the United States, Japan, Austria, the Netherlands, India, Nigeria, and the former Yugoslavia. Some of the first products of this research (e.g., Verba and Nie 1972) overturned the conception of participation as a hierarchy of electoral activities. Instead, participation was expanded to include non-electoral activities and reconceptualized as modes of activity in any one or more of which individuals might specialize. Although often these are positively correlated, meaning that a participant in one is more likely to take part in others than a randomly selected person, many people nonetheless partake in one or some to the exclusion of other forms of participation. In particular, substantial numbers of people who avoided electoral politics did engage in nonpartisan participation, such as involvement with community groups.

As reported in Verba and Nie (1972) and Verba et al. (1978), as well as other publications from the project, wealth and education were reaffirmed as important predictors of participation. But the main theoretical point is that the strength of that relationship varies across societies as a function of the strength of other, 'group,' resources. Where these are strong, such as in countries with strong political cleavages tied to ascriptive characteristics, or with sociological or occupational segments with strong political leadership, the relationship of socioeconomic status (SES) to participation was relatively weak. Certain well-off segments of the population were 'inhibited' from

participation while other, possibly disadvantaged, segments were 'mobilized,' thereby together attenuating the relationship between SES and participation. While others had noted similar phenomena, the finding now had wide cross-national grounding.

The Verba and coworkers' studies deliberately ignored political protest, that is, participation in activities outside accepted conventional norms, including some that are illegal. This omission seemed unfortunate by the 1970s in light of the frequency of real-world protest. Out of the study of revolution and political violence (see *Revolution*) came renewed interest in relative deprivation. But many protest activities are not violent (see *Political Protest and Civil Disobedience*).

A group of European and American scholars launched the Political Action project, a cross-national study focused on protest. They conducted surveys in the United States, West Germany, the Netherlands, Great Britain, and Austria in 1973 to 1974, in Italy, Switzerland, and Finland two years later, and a second wave in the first three countries in 1979 to 1980. The results appeared in *Political Action* (Barnes et al. 1979), *Continuities in Political Action* (Jennings et al. 1989) and other publications. One objective of this research was to understand the relationship between protest and conventional participation. The Political Action project clearly showed that enough persons who participated in conventional politics were adding protest to their 'action repertory' to conclude that conventional and protest participation were complements rather than substitutes. The Political Action studies differed from their predecessors in showing that variations in values predict both who will be active and the form of their activity. Specifically, persons with postmaterialist values are more active and more likely to protest than are materialists. Interpretation of the results is hampered slightly by the inclusion of respondents' hypothetical willingness to protest in some of the measures of protest activity.

Other scholars investigated the effect of political institutions on participation. In particular, they showed that cross-national variations in voter turnout rates reflect differences in electoral systems, party competitiveness, registration systems, government structure, and requirements to vote (see *Voting: Compulsory*).

Attention to participation in organizations shades into studies of social movements. Where a political scientist sees a participant in a voluntary organization addressing a specific problem (and thus somehow engaged in the communal mode of participation), a sociologist sees someone active in a social movement. The latter have produced another extensive, related literature.

Approaches and/or theories from these cross-national projects have been influential in many country-specific studies of participation. These studies have contributed much that is new, but their conclusions have also tended to support the theories developed in the large projects.

2.3 Mobilization as a Predictor of Participation, Not a Type, and the Civic Voluntarism Model (CVM)

The next major theoretical development in the study of participation involved incorporation of the recognition that people are more likely to participate when they are asked to do so (i.e., when they are 'recruited' or 'mobilized') and, moreover, that mobilization follows systematic patterns. Rosenstone and Hansen (1993) developed this idea in their study of participation in the United States from 1956 to 1990, claiming that a decline in mobilization over that period accounts in large part for the observed decline in participation.

Verba et al. (1995) use the related concept of 'recruitment' cross-sectionally in their study of political participation in the United States based on a 1989–1990 survey. It is the third component of the CVM; the first two are 'resources' and 'engagement.' They argue that higher levels of any of these factors lead to greater participation—people participate in politics because they can, because they want to, and because someone asks.

'Resources' are most readily measured by income and education, but these serve as proxies for more fundamental measures. The most interesting components of resources are the 'civic skills,' operationalized by Verba et al. (1995) as writing a letter, making a speech, and attending or planning a meeting where decisions are made in one or more nonpolitical settings, such as on the job, at church, or in a voluntary association. Civic skills turn out to explain the long-observed association between organization membership and participation. 'Engagement' includes interest, efficacy, political information, and partisan attachment, and also specific issue or ideological concerns. Note the connection between this part of the model and the finding that postmaterialists are more participatory. 'Recruitment' includes both requests from individuals and those from persons in authority (e.g., on the job, at church, or in an organization). This latter route provides one path by which group resources operate.

People with greater wealth and education in general have more resources, are more engaged, and are more likely to be recruited, so overall the relationship between socioeconomic status and participation is positive. The CVM model provides a fuller understanding however of why the relationship holds. Moreover, were it applied cross-nationally, it would point to components that need measuring. For example, churches play a role in the United States not likely to be replicated elsewhere.

Cost and benefit analyses inspired by rational actor models influence much contemporary work on participation, including that of Rosenstone and Hansen

(1993) and Verba et al. (1995) (see *Rational Choice in Politics*).

Many of the factors that lead to high levels of political participation are associated with high levels of social capital (see *Social Capital*) notably including substantial organizational involvement. Since social capital is also measured with political participation, the relationship among the concepts and the phenomena need more attention.

3. Conceptual Issues in Research on Political Participation

3.1 Participation Outside Advanced Democratic Countries

Most of the research discussed above has focused on economically-advanced democratic countries, with the occasional addition of other research sites. The restriction to democratic countries reflects the particular role of political participation in a democracy. Participation provides the means by which people control their government, most decisively so in elections which determine who rules. The unique power of the vote to change rulers provides the force for most of the other modes of participation. Campaigning, expressions of opinion, community organizing, and other actions command attention in part because ultimately the opposition can win an election.

In an authoritarian regime, the electoral connection no longer exists. Nonetheless, there is still political participation. Although the people no longer select the rulers, the rulers still need to keep the population reasonably satisfied to forestall revolt, and, to do so, they need information on the people's preferences, and they need to respond to discontent. Several very interesting studies have examined participation in authoritarian regimes, such as Friedgut's study of the USSR (1979) and the work by Shi (1997) on China.

In economically less-developed countries, one frequently finds that government institutions are also less well developed, that the associational networks are weak, or that loyalties to local groups outweigh those to governmental institutions. In these circumstances, much political participation may consist of informal involvement in the local arena or in traditional institutions (see, for example, Dietz 1998).

Issues of cross-national and cross-cultural comparability loom large in all studies of participation, but they are especially salient for research in non-democratic or less-advanced countries. One must rethink what counts as 'political' in these contexts in light of the institutional situation. For example, some of these studies led to the suggestion that participation includes activity directed towards altering the implementation of policy, rather than just its formulation, or that bribery is an important mode. Certain actions that are nominally the same can serve very different functions in differing contexts. For example, 'voting' may just show regime support in one country but select a government in another. The independent variables of importance also vary. For example, because of the particular circumstances in China, 'education' does not usefully measure resources there (Shi 1997, pp. 145–8). In many countries, other institutions are likely to play the role of churches in the United States as a source of civic skills.

Matters become more complex when an authoritarian regime democratizes. Many current studies examine participation in Eastern Europe and the republics of the former Soviet Union. With the abrupt changes in regime, the process of political socialization (see *Socialization: Political*) will probably also change, providing good conditions in which to observe the effect of political attitudes on participation.

3.2 Leadership

If people participate more when they have resources, are engaged, and are recruited, then elite action can significantly alter the level of participation by affecting engagement and recruitment. Political leaders provide resources to make participation easier (such as a babysitter, a ride to the polls, or information on what to support). Leaders contact people, directly or through proxies, asking them for action. Leaders can increase engagement by framing the public debate in a way that increases the stakes of political competition by communicating this message to their listeners.

The contemporary mass media play a role very similar to leaders. They frame issues and strongly affect (at times, negatively) engagement and recruitment. However, their motivations differ from those of leaders as they seek attention rather than support, and their actions are not targeted.

The political world contains many persons vying for power, each of whom needs large numbers of supporters to achieve power. Only some are successful in recruiting others and, as noted above, theory has not yet explained the circumstances of success. But political participation will increase when many are active and successful, and it will be higher than otherwise expected for persons who are targets of their efforts.

See also: Alliances: Political; Campaigning: Political; Citizen Participation; Cross-pressures: Political; Democratic Theory; Efficacy: Political; Faction: Political; Nonpartisanship: Political; Political Protest and Civil Disobedience; Political Representation; Power: Political; Rational Choice in Politics; Revolution; Social Capital; Social Movements, Sociology of; Socialization: Political; Voluntary Organizations; Voting: Class; Voting: Compulsory; Voting, Sociology of; Voting: Turnout

Bibliography

Almond G, Verba S 1963 *The Civic Culture*. Princeton University Press, Princeton, NJ
Barnes S H, Kaase M, Allerbeck K R, Farah B G, Heunks F, Inglehart R, Jennings M K, Klingemann H D, Marsh A, Rosenmayr L 1979 *Political Action: Mass Participation in Five Western Democracies*. Sage, Beverly Hills, CA
Conway M Margaret 2000 *Political Participation in the United States*. 3rd edn. CQ Press, Washington, DC
Dietz H 1998 *Urban Poverty, Political Participation, and the State*. University of Pittsburgh Press, Pittsburgh, PA
Eckstein H, Gurr T R 1975 *Patterns of Authority: A Structural Basis for Political Inquiry*. Wiley, New York
Friedgut T H 1979 *Political Participation in the USSR*. Princeton University Press, Princeton, NJ
Huntington S P, Nelson J M 1976 *No Easy Choice*. Harvard University Press, Cambridge, MA
Jennings M K, van Deth J W, Barnes S H, Fuchs D, Heunks F J, Inglehart R, Kaase M, Klingemann H-D, Thomassen J J A 1989 *Continuities in Political Action*. de Gruyter, Berlin
Milbrath L W 1965 *Political Participation*. Rand McNally, Chicago
Rokkan S 1970 *Citizens Elections Parties*. Oslo University Press
Rosenstone S J, Hansen J M 1993 *Mobilization, Participation, and Democracy in America*, Macmillan, New York
Shi T 1997 *Political Participation in Beijing*. Harvard University Press, Cambridge, MA
Verba S, Nie N H 1972 *Participation in America*. Harper & Row, New York
Verba S, Nie N H, Kim J-O 1978 *Participation and Political Equality*. Cambridge University Press, Cambridge, UK
Verba S, Schlozman K L, Brady H E 1995 *Voice and Equality*. Harvard University Press, Cambridge, MA

C. J. Uhlaner

Parties: Litigants and Claimants

This article focuses on parties who are subject to legal proceeding in a nonprofessional capacity. Excluded are lawyers, judges, and other officials who make their living from legal work; also excluded are the individuals who serve as witnesses and jury members in trials. The focus is on actors who are the *subject* of legal decisions. They include litigants, claimants, and those subject to investigation and scrutiny by law enforcers in both criminal and administrative capacities.

It is important at the onset to differentiate between those who initiate legal proceedings and defendants. Plaintiffs in civil courts and complainants in criminal courts exercise volition in choosing to take disputes to court. While defendants in criminal and civil court proceedings may indirectly exercise volition, in the sense that the legal proceedings may be a consequence of their own prior decisions, the choice to be involved in the legal proceeding itself is typically not something they would choose on their own.

Disputes and allegations of illegal behavior that end in court are merely the tip of the iceberg (Felstiner et al. 1980–81). On the civil side, people may formally or informally negotiate a settlement 'in the shadow of law', or may 'lump' an unresolved dispute due to high costs of proceeding, inability to find an advocate who will take the case, or fear of retaliation for making legal claims. In many places, criminal cases are resolved by plea bargains instead of a formal trial. Complainants may also have their cases turned away by gatekeepers as not being sufficiently 'legal' to warrant court attention (Yngvesson 1993, Merry 1990).

1. *Dimensions of Difference in Parties*

The research literature focuses on (a) the distinction between individuals and organizations, and (b) demographic differences related to social power—primarily gender, race, and class. Both lines of research attempt to account for differences in the experiences of various types of parties with legal systems, but in somewhat different ways.

Comparisons of individuals and organizations have been made in studies testing Marc Galanter's theory of why 'haves'—parties with greater resources—may have a greater advantage in legal settings (Galanter 1974). Most empirical tests of the theory have examined success rates for different classifications of organizations compared to individuals and each other, using appellate court cases in which there is a written record. In addition, white-collar crime research has also emphasized differential treatment of organizations and individuals, and poses additional explanations for why organizations may experience more advantageous treatment than individuals when they violate the law.

Studies of gender, race, and ethnicity tend to focus more on case processing, and to examine variations among different categories of individuals in far greater detail, with minimal attention to variations among organizations. A variety of methodologies have been employed, including observational studies, analysis of legal cases, and quantitative survey studies. Some studies focus on variations in resources and skills that affect an individual's chances of negotiating successfully through legal proceedings, while others focus on the social categories and standards used by legal decision-makers in making decisions on how to proceed with cases.

2. *Comparisons of Individuals and Organizations*

Galanter's classic 1974 article 'Why the "haves" come out ahead' argues that organizations are more likely, compared to individuals, to have structural advantages that increase the likelihood of winning in court. In addition to their greater financial resources and

ability to hire legal talent, large organizations are more likely to be 'repeat players' who are better able to mobilize legally and to use law to meet strategic goals (Galanter 1974). A number of studies of state and federal appellate court decisions have found support for Galanter's argument (Songer et al. 1999). The most consistent finding is that government organizations, particularly federal and state governments, experience a substantial advantage over both business and individuals, while business organizations appear to have a modest advantage over individuals. Studies that distinguish between larger business corporations and smaller business organizations generally find a stronger advantage for big business (Farole 1999, Songer et al. 1999).

Much of the appellate court research has been conducted in the United States, but similar patterns have also been found for other countries governed by the English common law tradition. Moreover, Songer et al. (1999) found considerable stability over time of the basic pattern of results for the US Court of Appeals, suggesting some generality to the pattern of greater organizational success in appellate courts.

A study of Philippine Supreme Court decisions, however, found strikingly different results, with individuals experiencing greater success in winning cases than organizations (Haynie 1994). Haynie speculates the legitimacy of high courts in developing countries may be more precarious. This may lead courts to make more decisions with redistributive effects in a bid to enhance their legitimacy by establishing their independence from other powerful actors in society.

Research on lower courts points to mixed evidence regarding the thesis that organizations are advantaged over individuals. Anthropological studies of the lower courts suggest that organizations face fewer difficulties than individuals in getting their complaints against others into court. Yngvesson (1993) found that the majority of complaints brought by individuals were characterized as 'garbage' cases by court clerks, who often expended considerable effort to keep them out of court (see also Merry 1990). In contrast, complaints filed by police officers—representatives of a local governmental agency—were routinely admitted with minimal review as to their appropriateness for legal action. Similarly, motor vehicle complaints, code violations, and welfare fraud cases filed by government agencies, as well as debt collection cases filed by local businesses, were given a higher priority.

No doubt one of the reasons why complaints filed by the police, other government agencies, and businesses were more readily accepted and processed by the lower courts is that they are more likely to be explicitly linked to legal provisions and couched in the language of law and legal doctrine. This is in part due to the focus of such cases on disputes about money and property, while complaints brought by individuals tend to involve interpersonal disputes and problems with relationships. In addition to the more explicitly 'legal' character of property disputes, the greater training and proficiency of organizational representatives in formulating their complaints in legal terms may also assist getting borderline cases into court. As one police officer stated, while discussing a complaint he filed of 'deliberate and malicious' damage to a tree. *If you really want to, you can find a law to do what you want to do*' (Yngvesson 1993, p. 21; emphasis in original).

Organizational studies abound with examples of the strategic use of the law to achieve organizational goals. Examples in the business sector include innovative interpretations of tax laws in tax avoidance schemes (McBarnet 1991) and the strategic use of bankruptcy law (Delaney 1992). In the government sector, prosecutors in the USA have creatively used the RICO (Racketeer-Influenced and Corrupt Organizations) law to expand prosecutions beyond the original target of organized crime to include white-collar crimes such as commercial and financial fraud, tax evasion, embezzlement, and bribery (Friedrichs 1996).

While the extension of RICO laws to white-collar offenses offers an example of legal innovation that increases the advantage of large government organizations over individuals and business organizations, the research literature on white-collar crime finds that, as suspects and defendants in criminal proceedings, organizations are advantaged over individuals. As Sutherland (1983) noted, illegal behaviors by organizations are typically handled by regulatory agencies instead of criminal law enforcement agencies. Sanctions by regulatory agencies are relatively mild, typically involving fines, cease and desist orders, and other civil sanctions. While regulatory agencies may seek criminal sanctions, the criminal prosecution of organizations remains relatively rare.

The allocation of cases of organizational misconduct to civil courts and administrative enforcement, rather than criminal courts, cannot be attributed to a lesser seriousness of organizational misconduct generally. Cost comparisons typically find that illegal behaviors by organizations tend to be more deadly and expensive in the aggregate than crimes committed by individuals (Poveda 1994). Survey research generally finds that the public considers corporate crimes that results in loss of life as serious in nature (Friedrichs 1996). Some authors have also argued that greater alienation and cynicism by the public against public institutions generally is an additional cost of organizational crime (Maier and Short 1995).

Various explanations exist for why 'street crimes,' primarily committed by lower-class individuals, are relegated to the criminal courts, while of offenses of organizations are handled primarily by civil courts and administrative enforcement actions. Legal conflict theorists argue that those who have power and money are better able to influence legislators to exclude their behavior from criminal statutes (Green 1997). Even

when laws define misconduct by organizations as criminal, practical difficulties to securing criminal convictions include the diffusion of responsibility within organizations, and the greater complexity and cost of amassing evidence for organizational crimes (Friedrichs 1996).

Moreover, contemporary theories of organizational misconduct point to the influence of organizational culture and processes or organizational decision-making as causes of misconduct that do not conform easily to models of individual choice such as underlie much of the criminal law (Ermann and Rabe 1997). As Vaughan notes in her study of the *Challenger* disaster, 'repetition, seemingly small choices, and the banality of daily decisions in organizational life—indeed, in most social life—can camouflage from the participants a cumulative directionality that too often is discernible only in hindsight' (Vaughan 1996).

Finally, the pragmatics of securing cooperation from organizations to prevent future harms may be a key factor in the allocation of law enforcement activities related to organizational misconduct to regulatory agencies. While US regulatory agencies have historically been more likely to use adversarial, deterrence-oriented tactics such as the imposition of fines and penalties, regulatory agencies in most countries typically focus their energies on using educational and persuasive techniques to secure compliance (Braithwaite 1985).

The primary rationale for cooperative law enforcement strategies is that, while compliance with law for individuals often means refraining from committing illegal acts, compliance for organizations typically involves perspective behaviors, such as buying new equipment, training employees, and undertaking new routines and processes that are both expensive and time-consuming. Utilizing adversarial tactics may be counterproductive if they result in organized resistance by powerful actors (Bardach and Kagan 1982). Braithwaite (1984) frankly notes that a consequence of cooperative enforcement techniques is the introduction of distributive inequities, but argues that the payoff in increased effectiveness in saving lives and reducing human misery is worth the cost.

Counterbalancing the advantages organizations have in avoiding criminal proceedings for instances of illegal behavior, research on the civil courts typically finds that, when cases go to trial and the plaintiffs win, civil court verdicts are typically harsher for organizations than for individuals (Chin and Peterson 1985, Maccoun 1996). The 'deep pockets' hypothesis suggests that organizations are both more likely to be sued and to have higher damages assessed because they usually have more assets than individuals. While 'deep pockets' may explain decisions by both potential plaintiffs and attorneys to pursue cases, jury simulation studies have found little evidence that greater wealth explains the differential size of damage awards. Instead, it appears that juries apply different standards to organizations than individuals (see Maccoun 1996). However, the greater harshness of civil damage awards against organizations should be placed in context, in that few lawsuits go to trial.

3. *Demographic Differences Related to Social Power*

Other articles in this encyclopedia discuss race, class, gender, and the law in greater detail than is possible here. The purpose of this article section is to contrast apparent commonalities in what legal parties want from legal decision-makers with the difficulties they may experience in obtaining what they want due to barriers arising from their position in society.

Research on procedural justice suggests a high degree of commonality in what people want from legal proceedings (Lind and Tyler 1988, Tyler 1988). Studies conducted in different countries and in a variety of legal settings have consistently found that perceptions of fair process affect satisfaction with legal experiences as much or more than objective outcomes and perceptions of distributive justice. Moreover, procedural justice perceptions are generalized beyond the specific experience to affect the legitimacy of other legal and political authorities (Lind and Tyler 1988) and also expectations of fair treatment in other legal settings (Stalans 1994).

According to the group-value model of procedural justice, encounters with legal authorities convey profoundly symbolic messages to individuals about their social standing (Lind and Tyler 1988, Tyler and Lind 1992). The concept of standing emphasizes basic concerns individuals have about their membership and inclusion in social groups, about whether they are valued members of the community. Tyler (1988, 1994) argues that a substantial consensus exists on how people define procedural justice, and suggests that fair process is a means by which multicultural societies can maintain their legitimacy amid the diverse goals and interests of different groups within society. Results from other studies, however, caution that procedural justice may be a less important concern than outcomes for people who do not have a sense of identification with the larger collectivity (Huo et al. 1996).

Despite the emphasis of the group-value model of procedural justice on social standing and intergroup relations, surprisingly little procedural justice research explicitly addresses issues of social stratification. Yet many of the most pressing and controversial issues in law and society research involve the ways in which external social structures of power and domination are reproduced in legal settings during the enactment of legal procedures.

Anthropological research indicates that social stratification is intertwined with relational concerns and definitions of 'legal' issues in ways that systematically

privilege some groups over others in actual legal proceedings. To begin, the types of disputes that are often of central concern to women and lower-income individuals involve troubled relationships, including domestic violence. Yngvesson (1993) notes that clerks in New England courts were reluctant to bring cases involving relationship problems before judges. Even divorce lawyers who make their living from the dissolution of troubled relationships tend to be unresponsive and aloof to clients' tales of relational woes (Sarat and Felstiner 1995).

Merry (1990) notes that cases involving relational issues are more likely to be diverted from court into mediation—and that mediation itself and the court employees who manage mediation have a devalued social status in the court system. While mediation explicitly focuses on relationship issues, the application of rules of discourse that discourage conflict during the process of mediation may serve to systematically disadvantage women, compared to more adversarial court proceedings that focus on advocating and protecting legal rights (Conley and O'Barr 1998).

Discomfort with handling troubled relationships is not confined to Western legal systems. Griffiths (1997) reports that, in Botswana, a common ruling by *kgotla* in marital disputes is simply to order the parties to go home and 'live together in harmony and peace.' In Kenya, Swahili women who tell stories of domestic conflict and discord in Islamic Kadi courts violate social norms against revealing family secrets in public (Hirsch 1998). While some African courts encourage the airing of relationship issues, detailed analyses reveal the systematic shaming of women during the process of talk about relationships (Conley and O'Barr 1998).

The devaluation of relationship talk surfaces in Conley and O'Barr's study of US litigants in small claims court. Conley and O'Barr described two distinct types of narrative structures used by litigants. Rule-oriented accounts are ordered sequentially, describing the unfolding of events in a linear fashion while noting violations of specific rules, duties, and obligations by the other party. They articulate cause and effect and assess responsibility. Relational accounts, on the other hand, focus more on social position and general rules of social conduct. They are more likely to be full of details about the life of the speaker and the history of his/her relationship with the other party, including details that are not relevant to the legal issue at hand.

As Conley and O'Barr (1998) note, rule-oriented accounts conform to the logic of law in Western societies. Legal decision-makers readily understand rule-oriented accounts, while they have more difficulty processing relational accounts. It takes more work for judges and lawyers to translate the jumbled and rambling accounts in relational narratives into the rule-oriented framework of law. Many do not make the effort, to the disadvantage of individuals who employ relational accounts. This gives an advantage to repeat players, to those who can afford to hire legal counsel, and to better-educated, more articulate parties who bring to court greater skills in couching their narratives in forms readily comprehensible to legal decision-makers.

Conley and O'Barr also describe stylistic features of speech that can serve to reproduce structures of social domination (Conley and O'Barr 1998, Ainsworth 1993). Conley and O'Barr note that women and men of lower social status in society are more likely to evidence styles of speech that project deference and uncertainty, thereby undermining their own credibility. 'Powerless' speech includes the use of hedge words, polite forms, tag questions, exaggerated imprecision about quantities, and a rising intonation at the end of declarative statements. Although Lakoff (1975) first identified these speech styles as gendered features of language, studies in legal settings have found that some men also use these forms of speech, to their detriment (Conley and O'Barr 1998, see also Ainsworth 1993).

While dominant social orders tend to reproduce themselves in legal settings, the process is by no means immutable and unchanging (Griffiths 1997, Hirsch 1998). As Conley and O'Barr (1998) note, many individuals of subordinate social status are capable of adopting a rule-oriented discourse strategy and modifying their speech styles. In addition, the existence of a 'chivalry bias' favoring women in criminal cases (Anderson 1976) and the harsher sentences given to high-prestige individuals convicted of white-collar crimes (Weisbrud et al. 1991) highlight that the general social order is not simply reproduced in legal settings, but can be transformed in ways that create reversals of social fortune.

At the same time, it important to note that some patterns of domination and unequal treatment are more deeply engrained in the fabric of legal interactions and less amenable to modification or manipulation (e.g., Matoesian 1993, 1995). Studies of legal case processing in particular illuminate seemingly ubiquitous processes of social categorization of parties by legal decision-makers according to race, class, and gender in the course of identifying suspects, managing caseloads, evaluating credibility, and making decisions about how hard to work on cases (e.g., Waegel 1981, Frohmann 1997). While social categorization often works to further organizational goals and to increase the speed and efficiency of legal decision-making, it does so at the cost of procedural and distributive inequities that arise when individuals are judged through the prism of presumed group characteristics.

See also: Class and Law; Conflict and Conflict Resolution, Social Psychology of; Conflict Sociology; Courts and Adjudication; Critical Race Theory; Disputes, Social Construction and Transformation of;

Feminist Legal Theory; Gender and the Law; Justice and its Many Faces: Cultural Concerns; Justice, Access to: Legal Representation of the Poor; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Litigation; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Race and the Law

Bibliography

Ainsworth J E 1993 In a different register: The pragmatics of powerlessness in police interrogation. *Yale Law Journal* **103**: 259–322

Anderson E A 1976 The 'chivalrous' treatment of the female offender in the arms of the criminal justice system. *Social Problems* **23**: 350–57

Bardach E, Kagan R 1982 *Going by the Book: The Problem of Regulatory Unreasonableness*. Temple University Press, Philadelphia, PA

Braithwaite J 1984 *Corporate Crime in the Pharmaceutical Industry*. Routledge and Kegan Paul, London

Braithwaite J 1985 *To Punish or Persuade? The Enforcement of Coal Mine Safety Laws*. State University of New York Press, Albany, NY

Chin A, Peterson M 1985 *Deep Pockets, Empty Pockets, Who Wins in Cook County Jury Trials*. RAND Institute for Civil Justice, Santa Monica, CA

Conley J M, O'Barr W M 1990 *Rules Versus Relationships: The Ethnography of Legal Discourse*. University of Chicago Press, Chicago, IL

Conley J M, O'Barr W M 1998 *Just Words: Law, Language, and Power*. University of Chicago Press, Chicago, IL

Delaney K J 1992 *Strategic Bankruptcy: How Corporations and Creditors Use Chapter 11 to their Advantage*. University of California Press, Berkley, CA

Ermann M D, Rabe G A 1997 Organizational processes (not rational choices) produce most corporate crime. In: Lofquist W S, Cohen M A, Rabe G A (eds.) *Debating Corporate Crime*. Anderson, Cincinnatti, OH, pp. 53–68

Farole D J Jr. 1999 Re-examining litigant success in state supreme courts. *Law and Society Review* **33**: 1043–58

Felstiner W, Abel R, Sarat A 1980–81 The emergence and transformation of disputes: Naming, blaming, claiming. *Law and Society Review* **15**: 631–54

Friedrichs D O 1996 *Trusted Criminals: White Collar Crime in Contemporary Society*. Wadsworth, Belmont, CA

Frohmann L 1997 Convictability and discordant locales: Reproducing race, class, and gender ideologies in prosecutorial decision-making. *Law and Society Review* **31**: 531–56

Galanter M 1974 Why the 'haves' come out ahead: Speculations on the limits of legal change. *Law and Society Review* **9**: 95–160

Green G S 1997 *Occupational Crime*, 2nd edn. Nelson–Hall, Chicago, IL

Griffiths A 1997 *In the Shadow of Marriage: Gender and Justice in an African Community*. University of Chicago Press, Chicago, IL

Haynie S L 1994 Resource inequalities and litigation outcomes in the Phillipine Supreme Court. *Journal of Politics* **56**: 752–72

Hirsch S F 1998 *Pronouncing and Persevering: Gender and the Discourses of Disputing in an African Islamic Court*. University of Chicago Press, Chicago, IL

Huo Y J, Smith H J, Tyler T R, Lind E A 1996 Superordinate identity, subgroup identity, and justice concerns: Is separatism the problem? Is assimilation the answer? *Psychological Science* **7**: 40–45

Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum, New York

Lakoff R 1975 *Language and Women's Place*. Harper and Row, New York

Maccoun R J 1996 Differential treatment of corporate defendants by juries. *Law and Society Review* **30**: 121–62

Matoesian G M 1993 *Reproducing Rape: Domination Through Talk in the Courtroom*. University of Chicago Press, Chicago, IL

Matoesian G M 1995 Language, law, and society: Policy implications of the Kennedy Smith rape trial. *Law and Society Review* **29**: 669–701

McBarnet D 1991 Whiter than white-collar crime: Tax, fraud insurance, and the management of stigma. *British Journal of Sociology*

Meier R F, Short J F Jr. 1995 The consequences of white-collar crime. In: Geis G, Meier R F, Salinger L M (eds.) *White-collar Crime: Classic and Contemporary Views*, 3rd edn. The Free Press, New York, NY, pp. 80–104

Merry S E 1990 *Getting Justice and Getting Even: Legal Consciousness Among Working-Class Americans*. University of Chicago Press, Chicago, IL

Poveda T G 1994 *Re-thinking White-collar Crime*. Praeger, Westport, CT

Sarat A, Felstiner W L F 1995 *Divorce Lawyers and Their Clients*. Oxford University Press, London

Songer D R, Sheehan R S, Haire S B 1999 Do the 'haves' come out ahead over time? Applying Galanter's framework to decisions of the US Court of Appeals, 1925–1988. *Law and Society Review* **33**: 811–32

Stalans L J 1994 Formation of procedural beliefs about legal arenas: Do people generalize from loosely related past legal experiences? *Psychology, Crime, and Law* **1**: 1–19

Sutherland E H 1983 *White Collar Crime: The Uncut Version*. Yale University Press, New Haven, CT

Tonry M 1995 *Malign Neglect: Race, Crime, and Punishment in the US*. Oxford University Press, New York

Tyler T R 1988 What is procedural justice? Criteria used by citizens to assess the fairness of legal procedures. *Law and Society Review* **21**: 103–88

Tyler T R 1994 Governing amid diversity: The effect of fair decisionmaking procedures on the legitimacy of government. *Law and Society Review* **28**: 809–32

Tyler T R, Lind E A 1992 A relational model of authority in groups. In: Zanna M (ed.) *Advances in Experimental Social Psychology*. Academic Press, NY, Vol. 25, pp. 115–92

Vaughan D 1996 *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press, Chicago, IL

Waegel W B 1981 Case routinization in investigative police work. *Social Problems* **28**: 263–75

Weisbrud D, Wheeler S, Waring E, Bode N 1991 *Crimes of the Middle Classes: White-collar Offenders in the Federal Courts*. Yale University Press, New Haven, CT

Yngvesson B 1993 *Virtuous Citizens, Disruptive Subjects: Order and Complaint in a New England Court*. Routledge, New York

K. A. Kinsey

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Parties/Movements: Extreme Right

Political parties and movements that current conventional labeling by political scientists and other informed analysts would regard as of the extreme right have emerged in many countries and on many occasions, since at least the early nineteenth century. The label, as currently employed, is necessarily an encompassing one because it is applied incontrovertibly to numerous parties/movements that nonetheless differ from each other in other ways, and also, more contestably, to certain political formations onto which not every recognized specialist is willing to confer that designation. Some analysts have preferred descriptions such as 'neo-populist' or 'right-populist' for movements not universally classified as 'extreme right.'

1. *The Term 'Extreme Right'*

The term 'extreme right' can be shown to have been used in a political sense, but not with quite its contemporary meaning and connotations, since the mid-nineteenth century and before. However, the term and its cognates are surprisingly recent arrivals in political science's standard vocabulary, having been regularly employed in the discipline only since the late 1950s. Lipset, for example, used the term very occasionally in his *Political Man* (1959/1963), it became conventional usage among European (except German) commentators during the 1960s, and it achieved final recognition in American political science with Lipset and Raab's *The Politics of Unreason* (1970/1971). One reason for the slowness with which the actual term came to be adopted in the USA was the popularity of the alternative, and still used, 'radical right' in order to describe political phenomena that most, but not all, present analysts would now call 'extreme right.' Lipset had also been responsible for the introduction of the term 'radical right' into US political science, and so for its popularity during much of the following decade. It was perhaps taken from the usage in German of *rechtsradikal*. This neologism, although also postwar, can be traced back to its introduction into German in the early 1950s, probably influenced by the fact that Germany's most significant extreme right party was then the later proscribed *Sozialistische Reichspartei* (SRP). This designation, though sometimes also *extrem rechtsradikal*, was widely used by German authors in the 1960s and 1970s, and has only relatively recently been superseded among political scientists by *rechtsextrem/rechtsextremistisch*. The distinction was also a bureaucratically important one in Germany, since a purported and casuistical distinction between *rechtsradikal* and *rechtsextrem* was the basis on which the German Office for the Protection of the Constitution justified varying monitoring and control strategies toward different right-wing formations. By the 1990s extreme right activities in Germany made this distinction increasingly dubious and inoperable, and the latter term now in practice subsumes the former one.

2. *A Definition of 'Extreme Right'*

A definition of extreme right parties/movements has necessarily to be extensive and complex, for it must be appropriate for such movements as diverse in time and location as the American Anti-Masonic Movement of the 1820s; the Know-Nothing Movement of the 1850s; the French Boulangist Movement of the 1880s; interwar fascism in Italy and other European countries; Nazism in Germany; McCarthyism in 1950s USA; Poujadism in 1950s France; George Wallace's 1968 American Independent Party (AIP); more recent European examples such as the French *Front National* (FN) and the German *Die Republikaner* (REPs); Latin American paramilitaries in countries such as Nicaragua and Colombia; Vladimir Zhirinovsky's Liberal Democratic Party in Russia; and even purportedly more mainstream political parties of the present such as Austria's *Freiheitliche Partei Österreichs* (FPÖ) and Switzerland's *Schweizerische Volkspartei* (SVP). The contemporary American militias are usually regarded as extreme-right, but they are idiosyncratic examples, their most obvious European analogue perhaps being the militant neo-Nazi defense groups of Germany and some Scandinavian countries. Although not all analysts would agree on the 'extreme right' designation for every one of these, each has nonetheless attracted many such attributions from historians, political scientists, or informed journalist observers. Commentators from some countries (such as France and The Netherlands) have been more ready than those elsewhere unwaveringly and without collateral justification to label the contemporary examples as 'extreme right.' An inclusive definition of the extreme right has to be based on ideological features such as those below and, although each would not necessarily apply with equal emphasis and significance to all extreme right parties/movements, all political scientists using this concept would understand that many or most would, at leadership or mass-support level, characterize a political phenomenon so designated:

selective inclusion, especially within the nation and often based on imposed assumptions about ethnic or religious similarity (which may take the form of aggressive nationalism or ethnocentrism);

selective exclusion, directed against foreigners and/or indigenous or immigrant ethnic minorities, but also against social minorities such as homosexuals, travelers (gypsies), or in the past even the Masons;

racism, based on biological perspectives about supposed inferiority or on the fabrication of ethnic and cultural boundaries;

anti-Semitism, which may be both specific and directed at contemporary targets, but which may in some cases involve denying, or mitigating the scale of, the Holocaust and glorifying the deeds of the Third Reich;

a preference for authoritarian initiatives by a strong state;

the cult of a leader figure;

a preference for a hierarchical social order;

antisystemic and antipluralist political perspectives;

overt hostility to political opponents, often associated with the use, encouragement, or tolerance of violence;

low tolerance of social change; and

nostalgia for the past.

2.1 Extreme Right Phenomena in the Twentieth Century

It is useful to give a very general, if crude, periodization for extreme right phenomena of the twentieth century, although in each of these periods there was nonetheless a considerable variety of such parties/movements:

1920–45. The era of 'classic' anti-Semitic Nazism and fascism.

1945–70. The era of the 'old wave' extreme right, often comprising anti-Semitic neo-Nazi/neofascist movements with a yearning for the 'classic' era (e.g., the SRP or Italy's *Movimento Sociale Italiano*, MSI) and nostalgic anticommunist movements (e.g., McCarthyism).

1975–present. The era of the 'new wave' extreme right, appealing particularly on a basis of xenophobia and racism against ethnic minority immigrant groups and asylum seekers.

This periodization is not perfect, particularly because of anomalies among the US cases: the AIP, for example, anticipated the issues of the 'new wave' extreme right. However, it is a useful heuristic device for discussing theoretical approaches to the analysis of these movements.

3. Theoretical Approaches to the Extreme Right

Fashions in political science towards the analysis of the support of extreme right parties/movements have not been constant. While it is important not to exaggerate these changes, it is true that research in the 1950s and into the 1970s, conducted on the vote for Nazism (the *Nationalsozialistische Deutsche Arbeiterpartei*, NSDAP) and on 'old wave' postwar movements, was especially concerned with answering questions about the partisan origins and the social base of these examples of extreme right support. The lack of any individual-level data on voters for the NSDAP led to the use of different combinations of aggregate data in a substantial body of research about Nazism's partisan origins and social base. The most recent and comprehensive research in this tradition has argued the case for considerable social variety in the NSDAP's later support. Postwar movements for whose support there existed survey data were frequently analyzed in the first instance in terms of the social categories to which they particularly appealed. Ideological and attitudinal dispositions towards the extreme right were not ignored, of course (especially not by research drawing on perspectives in *The Authoritarian Personality*, Adorno et al. 1950), but they were not always treated as the principal explicandum; indeed, many authors (e.g., some of those writing about McCarthyism in *The Radical Right*, Bell 1963/1964) simply imputed motive. Researchers working on movements that have emerged in the 'new wave' era have been more inclined to recognize that most extreme right movements draw from a variety of social sources and have given greater attention to the attitudinal, psychological, and contextual bases of such support. The disproportionate support for many 'new wave' extreme right movements among the self-employed and among younger voters from manual-worker backgrounds (especially males) has been noted (e.g., in the cases of the FPÖ, the FN, and the German *Deutsche Volksunion*, DVU), but it has not been a central plank to theorizing about these movements.

In giving an overview of the theoretical perspectives towards extreme right parties/movements, it is essential to distinguish between the pre-World War II and postwar cases. Many of the interbellum movements played major political roles, especially so, of course, in the cases of Italy and Germany. Even in countries where such movements did not achieve power, they often had substantial impact upon national politics during the 1930s (as in France or The Netherlands). On the other hand, in the postwar period extreme right parties/movements, though not without some important effects upon the content of political agendas, have not usually achieved actual governmental office—significant exceptions being Italy briefly in 1959 and during the mid-1990s, Austria in 2000, as well as the problematic cases of Spain till 1975, and of contemporary Switzerland in the light of the recent metamorphosis of the SVP towards the extreme right.

Most contemporary theoretical analyses of the interbellum extreme right focused on German Nazism and Italian fascism. Some influential perspectives came from self-avowed Marxists and, though incorporating views about the social base of these movements, attributed to them a particular role in the balance of class forces. Perhaps the more subtle of these Marxist perspectives were those by Gramsci, who used a concept of Caesarism, and by Trotsky and Thalheimer, who drew Bonapartist analogies and regarded fascism as a resort by capitalists to perpetuate their interests in a situation of class stalemate between the proletariat and capital. Postwar Marxists have

sometimes resurrected such perspectives in looking at Italian fascism and Nazism and have even sought to apply them to postwar movements, but, to most analysts, these approaches are no longer relevant. Instead, the great majority of political scientists writing on the 'old and new wave' extreme right have sought, albeit in varying ways, to focus on the social situations that may produce extreme right sympathy and behavior.

A widespread feature of numerous theories of both the 'old and new wave' extreme right is to see them as supported by those who in some way have suffered, or see themselves to have suffered, from social change. This theorizing may take the form of predicating status loss (e.g., Hofstadter and Lipset on McCarthyism), or disorientation as a result of social strain (e.g., Parsons), or a reaction against the growth of postmaterialist values (e.g., Ignazi), material loss due to processes of modernization or globalization (e.g., Heitmeyer). Theories with a social-psychological basis, as by those who have applied perspectives derived from *The Authoritarian Personality* and the Frankfurt School to 'old and new wave' movements, also assume a resistance to, and intolerance of, certain types of social change. 'New wave' extreme right movements since the 1970s, because in many cases they have made racism and xenophobia the principal basis of their appeal, have particularly attracted theorizing seeking to explain the presence of such attitudes among individual voters and certain groups of voters.

4. *The Tendency of Extreme Right Parties to Schism*

Extreme right parties/movements frequently exhibit distinctive organizational characteristics in comparison with mainstream political parties. The former have a particular tendency to splits, and contain strong schismogenetic features. This is sometimes because of personal animosities between leaders, who are often psychologically disposed to be authoritarian and are intolerant both of different perspectives and also of rivals. Because of the variety of ideological appeals offered by extreme right parties, their leaderships and supporters often have a mixture of social origins and this too can lead to divisive differences over minutiae of ideology (a characteristic shared with extreme left movements). Extreme right parties, particularly those in the postwar 'new wave' period, have exhibited a remarkable tendency to fracture. In France, Germany, Great Britain, The Netherlands, and Switzerland, to name only some examples, the indigenous extreme right has had a history of division and internal dissension. Indeed, many of the apparent electoral successes in the 1980s and 1990s by extreme right parties foundered on the organizational collapse of the elected legislative grouping, something that has been a particular feature of the German extreme right parties, the REPs and the DVU.

5. *The Future for Extreme Right Parties/Movements*

It would be scientifically insupportable to suggest that such parties/movements will wither away. A troubled world is likely to provide just too many dynamics to underpin their emergence, especially in certain Third World contexts. In the liberal democracies, however, despite some successes in obtaining municipal control (as the FN has had since 1995) and breakthroughs by parties such as the FPÖ and the SVP (both of which emerged from mainstream political formations), there are few indications in most countries that, on a national level, extreme right parties will emerge on any scale from a political exclusion zone. Support even for the most successful examples has in most cases reached a zenith and elevated levels of one-time support frequently erode. These parties depend on a pariah status for attracting the votes of the politically alienated, and sometimes, as in France, where a party faction has wanted to seek accommodations with mainstream politics, the result has been a damaging split to the greater disadvantage of the faction arguing for such a reapprochement.

See also: Anti-Semitism; Authoritarian Personality: History of the Concept; Authoritarianism; Fundamentalism (New Christian Right); Marxism and Law; McCarthyism; National Socialism and Fascism; Nazi Law; Racism, History of; Racism, Sociology of; Radicalism; Right-wing Movements in the United States: Women and Gender; Xenophobia

Bibliography

Adorno T W, Frenkel-Brunswick E, Levinson D J, Sanford R N 1950 *The Authoritarian Personality*. Harper, New York

Bell D (ed.) 1963/1964 *The Radical Right: The New American Right Expanded and Updated*. Anchor, Garden City, NY

Betz H-G 1994 *Radical Right-wing Populism in Western Europe*. Macmillan, Basingstoke, UK

Betz H-G, Immerfall S (eds.) 1998 *The New Politics of the Right: Neo-populist Parties and Movements in Established Democracies*. 1st edn. St. Martin's Press, New York

Hainsworth P (ed.) 2000 *The Politics of the Extreme Right: From the Margins to the Mainstrem*. Pinter, London

Heitmeyer W, Bushse H, Liebe-Freund J, Möller K, Müller J, Ritz H, Siller G, Vossen J 1992 *Die Bielefelder Rechtsextremismus-Studie: Erste Langzeituntersuchung zur politischen Sozialisation männlicher Jugendlicher*. Juventa, Weinheim, Germany

Ignazi P 1994 *L'Estrema Destra in Europa*. Il Mulino, Bologna, Italy

Kaplan J, Weinberg L (eds.) 1998 *The Emergence of a Euro-American Radical Right*. Rutgers University Press, New Brunswick, NJ

Kitschelt H, McGann A J 1995 *The Radical Right in Western Europe: A Comparative Analysis*. University of Michigan Press, Ann Arbor, MI
Lipset S M 1959/1963 *Political Man: The Social Bases of Politics*. Anchor, Garden City, NY
Lipset S M, Raab E 1970/1971 *The Politics of Unreason: Right-wing Extremism in America, 1790–1970*. Heinemann, London
Merkl P H, Weinberg L (eds.) 1997 *The Revival of Right-wing Extremism in the Nineties*. Frank Cass, London
Pfahl-Traughber A 1994 *Volkes Stimme?: Rechtspopulismus in Europa*. Verlag J H W Dietz Nachfolger, Bonn, Germany
Scheuch E K, Klingemann H D 1967 Theorie des Rechtsradikalismus in westlichen Industriegesellschaften. In: Ortlieb H-D, Molitor B (eds.) *Hamburger Jahrbuch für Wirtschafts- und Gesellschaftspolitik*. J C B Mohr, Tübingen, Germany, pp. 11–29

C. T. Husbands

Partner Selection across Culture, Psychology of

Although family structures, forms, and functions vary across cultures, marriage is a central universal phenomenon in the lives of most people. Selection of a partner is a key facet of marriage, and its mechanisms vary in consonance with the social psychological construction of the concept and meaning of marriage in a given culture. Every culture has rules or normative frameworks governing the orientation towards partner search and selection.

Interestingly, a range of changes is evident in partner selection within and across cultures, some of which are a consequence of the shifts in the macro contextual conditions, such as socioeconomic development and the women's movement, which have impacted various cultural structures and processes. Importantly, these factors have introduced transformation in social psychological orientations, and created novel modes of interpretation of, and greater flexibility in the adherence to, the traditional cultural rules for partner search and selection.

1. *Theoretical Perspectives on Partner Selection*

On what basis and how do people select partners for marriage? This question has elicited considerable scientific interest, and transcends disciplinary boundaries. Different disciplines have generated theories and models to study partner selection systems and strategies. In general, however, two major theoretical perspectives are evident: the evolutionary perspective and the social psychological perspective. A recent trend is to merge disciplinary perspectives and approaches in order to comprehend the complexities involved in partner search and achieve a clearer understanding of the motives that determine the different elements of the process.

1.1 *Evolutionary and Social Psychological Perspectives: Towards a Synthesis*

Evolutionary theories enable understanding of the ultimate significance of heterosexual relationships, specifically the determinants of partner selection, whereas social psychological perspectives provide insight into the more proximate factors involved.

The evolutionary models of partner selection essentially view individuals as acting on the basis of evolved mechanisms, selected with a focus on maximizing ancestors' genetic fitness or increasing the likelihood of raising healthy offspring (Kenrick et al. 1993). These models contend that gender differences in preferences or selection criteria are cross-culturally universal, and have unfolded through sexual selection pressures over millions of years. As the physical condition of females is necessary for offspring survival, males generally tend to focus upon characteristics indicative of this aspect. On the other hand, females place more value on males' ability to contribute to resources.

A prominent theoretical model in social psychological perspectives on partner selection is the social exchange model, which posits that individuals search for partners who have resource value that is equitable with one's own 'market value.' In other words, there is an exchange of valued traits or resources (Kenrick et al. 1993). Based upon the market value of each trait, which varies cross-culturally, both women and men seek partners who will maximize gains and minimize costs. Social psychological models generally explain gender differences in terms of structural powerlessness of women and traditional sex role socialization patterns. Evolutionary theories, on the other hand, emphasize gender differences in terms of the reproductive investment hypothesis (Buss and Barnes 1986).

Evidence of the universality of certain gender differences in partner selection preferences across cultures supports the role that evolutionary forces play in shaping cultural factors that mediate the gender-based criteria for partner selection (Buss et al. 1990). Nevertheless, predictions related to psychological mechanisms in selection preferences offered by evolutionary models need to be viewed in conjunction with specific cultural-environmental contexts (see *Evolutionary Social Psychology*).

The sexual strategies theory proposes an integrated contextual-evolutionary perspective of partner selection. Based on the premise that human mating is inherently goal-directed, the theory incorporates factors from the larger context that have a bearing on partner selection and states that mating strategies are context-dependent in terms of the social milieu and the duration of the relationship. Also, the principles that govern selection by women and men are different and,

to that extent, the psychological strategies or mechanisms adopted to maximize reproductive success and offspring survival also differ (Buss and Schmitt 1993).

2. *Culture as Mediator of Psychological Orientations*

Culture is a significant mediating factor in partner selection. The cultural context provides shared schemas of meaning and knowledge systems that serve as a framework for individual experiences. The individual as a self-reflective and interpreting subject constructs and modifies the cultural schemas, thereby engendering transformation in the cultural rule systems for different domains, including partner selection. Thus, the cultural schemas on the one hand serve as a sort of template for individual and collective experiences, and on the other hand the individual constructs and modifies them in the course of cultural history (Eckensberger 1990) (see *Cultural Psychology*).

Social relationships and rule systems are at the core of culture. In cross-cultural psychology, these are conceptualized in terms of two main differentiations of self-concerned or other-concerned individual psychological orientations, related to Western and non-Western contexts, respectively (Markus and Kitayama 1991) (see *Cross-cultural Psychology*). The dichotomy of individualism and collectivism is commonly used to conceptualize the two types of cultures. Individualism emphasizes individual interests, needs, goals, and independence; collectivism refers to an orientation towards collective interests and norms with the aim to maintain social harmony (Hui and Triandis 1986) (see *Personality and Conceptions of the Self*).

The implications of these orientations for the structure and function of the primary social group, that is, the family, are of particular interest, in terms of the impact upon significant life events of individuals, including marriage and its integral process—selection of a partner.

3. *Cross-cultural Perspectives*

Multicultural studies of partner selection reveal interesting patterns of commonalties and diversities within and across cultures, which can be substantially attributed to changes in macro contexts as a result of industrialization and modernization.

Traditional and modern or collectivist and individualist cultures represent different clusters of qualities preferred in a prospective mate. Traits such as chastity, and domestic characteristics including desire for home and children, and being a good housekeeper are valued in traditional-collectivist cultures such as in China, India, Indonesia, Taiwan, Iran, the Palestinian Arab community, South Africa, and Colombia. The modern-individualist cultures (e.g., the USA, Canada, most of Western Europe) consider such characteristics as irrelevant in the partner selection criteria or place a comparatively low value on the same. Relatively uniform gender differences across cultures are observed specifically in men's emphasis on physical appearance and value for traditional characteristics such as domestic skills, and women's emphasis on resource potential. Sexual dimorphism is evident most in the collectivist Asian and African cultures (Buss et al. 1990).

In the Western world, partner selection is essentially based on one's individual criteria, largely independent of familial and societal rules. Also evident is the growing utilization of innovative formal searching services. As one focuses on non-Western societies, specifically some selected Asian and African ones, the scenario reflects an interesting melange of 'traditional' and 'modern' (including Western) patterns in the process of partner selection.

The South Asian countries of China and Japan (Jew 1994, Wang 1994) demonstrate the continuance of certain traditional patterns of partner selection, such as arranged marriages, alongside the changes being introduced by Western influences. Individuals who marry are beginning to have more decision-making power. Nevertheless, parental opinion and approval are generally sought, if only as a mark of duty, obligation, and respect for them.

The ongoing transformation across the non-Western societies in particular, reveals the dynamic interface between culture and the self-reflective intentional individual who is engaged in the complex process of cultural reconstruction (see *Action Theory: Psychological*). Traditional normative rule systems provide a common frame for individuals' experiences of partner selection. In response to the larger context of rapid economic and social influences, individuals are engaging in a process of reinterpretation and reconstruction of the existing rule systems. In this context, societies that are predominantly characterized by collectivist perspectives are beginning to incorporate elements of the individualist perspective, in indigenized forms. The contemporary scenario illustrates a 'transitional phase' with varying blends of 'traditional' and 'modern' or 'collectivist' and 'individualist' perspectives and practices of partner selection represented cross-culturally.

4. *The Indian Context*

In Hindu India, marriage or *vivaha* is an inevitable significant life ritual (*samskara*). Marriage is considered as an alliance between two families, rather than two individuals.

4.1 *Cultural Rules for Partner Selection and Emerging Social Psychological Orientations*

Caste (*jati*) compatibility is the most important factor in a Hindu marriage. The position of the family in the

caste hierarchy determines the field of selection. This rule is now applied with greater flexibility, and there are incidences of intercaste marriages. A growing emphasis on individual characteristics of the prospective partner is also evident.

Although the arranged marriage remains a largely prevalent strategy of partner selection, the system has been modified to accommodate the preferences of adult children, and the traditional 'unconsented' model has given way to a 'consented' model (Banerjee 1999). The late 1990s also witnessed the growing involvement of external intermediaries, such as matrimonial advertisements and marriage bureaus, in partner selection.

The contemporary Indian context of partner selection reflects the influences of industrialization and modernization, including the global women's movement and the ensuing changes in values and accepted practices. These changes are revealed in the greater assertion of individual choice, the parallel instances of resistance to certain normative rules, and the diluting role of family and kin networks. Although the evolving climate is more conducive to the articulation of individualistic orientations, the value of family interdependence persists.

The Indian scenario depicts the continued resilience of traditional practices, alongside its adaptive features created to 'fit' with the transforming social context. It thus reveals the two-dimensional interdependent-independent or relational-autonomous orientation characteristic of many non-Western contexts (Kagitcibasi 1997).

5. *Future Directions*

How cultures operate to 'mate' individuals is a significant issue that provides a rich context for studying the interaction among psychological phenomena, their developmental patterns, and the cultural context. The emerging trend of integrating the social psychological or cultural and evolutionary models needs to be strengthened. Furthermore, indigenous paradigms for theory and research must be developed to trace the ongoing transitions and the diversity in developmental models across societies. Importantly, intracultural heterogeneity must be highlighted, especially in multicultural societies.

The validity of cross-cultural trends in specific gender-based selection preferences supported by the evolutionary model needs to be further corroborated. A critical question in this context is whether and to what extent women's education and economic independence actually enhance gender egalitarianism in the realm of partner selection. Importantly, interdisciplinary culture-based perspectives are imperative to develop conceptual frameworks for interpreting the wide-ranging diversity and complexity in the partner selection process.

See also: Action Theory: Psychological; Cross-cultural Psychology; Cultural Psychology; Cultural Variations in Interpersonal Relationships; Evolutionary Social Psychology; Family Theory: Economics of Marriage and Divorce; Love and Intimacy, Psychology of; Marriage; Marriage: Psychological and Experimental Analyses; Personality and Conceptions of the Self; Personality and Marriage

Bibliography

Banerjee K 1999 Gender stratification and the contemporary marriage market in India. *Journal of Family Issues* **20**: 648–76

Buss D M, Schmitt D P 1993 Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review* **100**: 204–32

Buss D M, Abbott M, Angleitner A et al. 1990 International preferences in selecting mates: A study of 37 cultures. *Journal of Cross-Cultural Psychology* **21**: 5–47

Buss D M, Barnes M 1986 Preferences in human mate selection. *Journal of Personality and Social Psychology* **50**: 555–70

Eckensberger L H 1990 From cross-cultural psychology to cultural psychology. *The Quarterly Newsletter of the Laboratory of Comparative Human Cognition*, 1990*a* **12**: 37–52

Hui C H, Triandis H C 1986 Individualism–collectivism. A study of cross-cultural researchers. *Journal of Cross-Cultural Psychology* **17**: 225–48

Jew C 1994 Understanding the Japanese family structure from the Japanese perspective. *Family Perspective* **28**: 303–14

Kagitcibasi C 1997 Individualism and collectivism. In: Berry J W, Segall M H, Kagitcibasi C (eds.) *Handbook of Cross-Cultural Psychology: Social Behavior and Applications,* 2nd edn., Allyn and Bacon, Boston, Vol. 3 pp. 1–49

Kenrick D T, Groth G E, Frost M R, Sadalla E K 1993 Integrating evolutionary and social exchange perspectives on relationships: Effects of gender, self-appraisal, and involvement level on mate selection criteria. *Journal of Personality and Social Psychology* **64**: 951–69

Markus H R, Kitayama S 1991 Culture and self: Implications for cognition, emotion and motivation. *Psychological Review* **98**: 224–53

Wang G T 1994 Social development and family formation in China. *Family Perspective* **28**: 283–302

S. Kapadia

Partnership Formation and Dissolution in Western Societies

1. *Partnership Formation*

1.1 *Marriage Trends*

In the recent past, marriage heralded the start of a first union for most couples in Western societies and the great majority of marriages survived intact until one of the spouses died. There tended to be identifiable stages

in the development of a relationship: courtship, engagement, and ultimately the marriage ceremony that was followed by the couple setting up home together. Nowadays, there is more flexibility in becoming a couple and whether they co-reside.

After World War II, the general trend in marriage behavior in European and other developed countries was towards a younger and more universal marriage pattern which reached its zenith during the 1960s and early part of the 1970s (Festy 1980, Espenshade 1985, Fitch and Ruggles 2000). Since then, marriage rates in most Western countries have declined (UN Demographic Yearbooks and Council of Europe demographic publications provide collections of relevant data). Younger generations have been marrying less, and among those who marry the trend has been to do so at older ages and over a wider range of ages than was common among their recent predecessors. In broad outline in the case of Europe, the decline in marriage rates began in Sweden and Denmark in the late 1960s, and then spread through most of Western Europe in the early part of the 1970s, and became evident in the southern European countries (Spain, Italy, Portugal, and Greece) around the mid-1970s. Since the 1980s, the decline in marriage rates continued in most European countries but at a slower pace. Changes in the mean age at first marriage provide an illustration of the extent of these changes. In 1975, the average ages of first-time brides in most Western nations were clustered in the 22–4 years range, whereas in the mid-1990s they were clustered in the range 26–9 years. The pattern of change was similar for men, but they marry, on average, two years older than women do, an age difference that has been fairly invariable over time.

1.2 Cohabitation

One of the important engines behind the decline in marriage rates and a movement to a later age at marriage is the rise in cohabitation that has occurred, particularly since the beginning of the 1980s in many countries and even earlier in Scandinavia. However, it should be emphasized that men and women living together outside marriage is certainly not new. Prior to the 1970s, it was largely statistically invisible and probably socially invisible outside the local community or social milieu. For example, in some European countries there were subgroups that were probably more prone to cohabitation than others: the very poor; those whose marriages had broken up but who were unable to obtain a divorce, as there was no such legislation, or it was more stringent than nowadays, or it was very expensive to obtain a divorce; certain groups of rural dwellers; and groups ideologically opposed to marriage.

The form of cohabitation that came to the fore during the 1960s in Sweden and Denmark, and during the 1970s in other northern and Western European countries, North America, and Australia and New Zealand, is new, and has been variously termed 'premarital cohabitation,' 'nubile cohabitation,' and 'cohabitation juvenile'; in this form, young people predominantly in their 20s and early 30s live together either as a prelude to, or as an alternative to, marriage. Additionally, with the growth in divorce, 'postmarital cohabitation' has also become more prevalent, with the divorced cohabiting either in preference to, or as a prelude to, remarriage. Unfortunately, in many data sources, it is difficult to distinguish between the 'premarital' and 'postmarital' forms of cohabitation. However, the increased prevalence of cohabiting unions is likely to lie behind much of the decline in first marriage and remarriage rates that have occurred in many Western countries in the latter decades of the twentieth century.

Data on cohabitation has tended to be scarce and generally emanated from *ad hoc* surveys. This made comparative analyses problematic, as sample sizes, coverage, and definitions could vary. During the 1990s more information from standardized questionnaires became available from Eurostat (the Statistical Office of the European Communities) and from a series of Fertility and Family Surveys carried out in the main in the first half of the 1990s under the auspices of the UN Economic Commission for Europe (ECE) (United Nations 1992).

1.2.1 The incidence of cohabitation. Analysis of data from Eurobarometer Surveys carried out in the 15 member states of the EU in 1996 provides a perspective on the incidence of cohabiting and marital unions across a range of nations. These surveys are primarily opinion surveys covering a range of topics relevant to the EU, but they contain some very basic demographic information on the respondents, including information on marital status in which 'living as married' is one of the categories, the others being the more conventional ones of single, married, divorced, separated, and widowed. Such marital status distributions may not be as accurate as those obtained in dedicated family and fertility surveys, but they probably reflect the relative position of different European countries in these categories (European Commission, 1996).

Figure 1 shows the proportions of women aged 25–29 years in the 15 EU countries who were cohabiting, married, single, or separated/divorced/widowed at the time of the survey in 1996. In these data never-married and postmarital cohabitants cannot be distinguished, but it is reasonable to assume that at these younger ages the former is likely to be the most prevalent. It is clear from Fig. 1 that there is a good deal of diversity across European states in the incidence of cohabitation. Cohabitation is strikingly most common in the Nordic countries of Denmark,

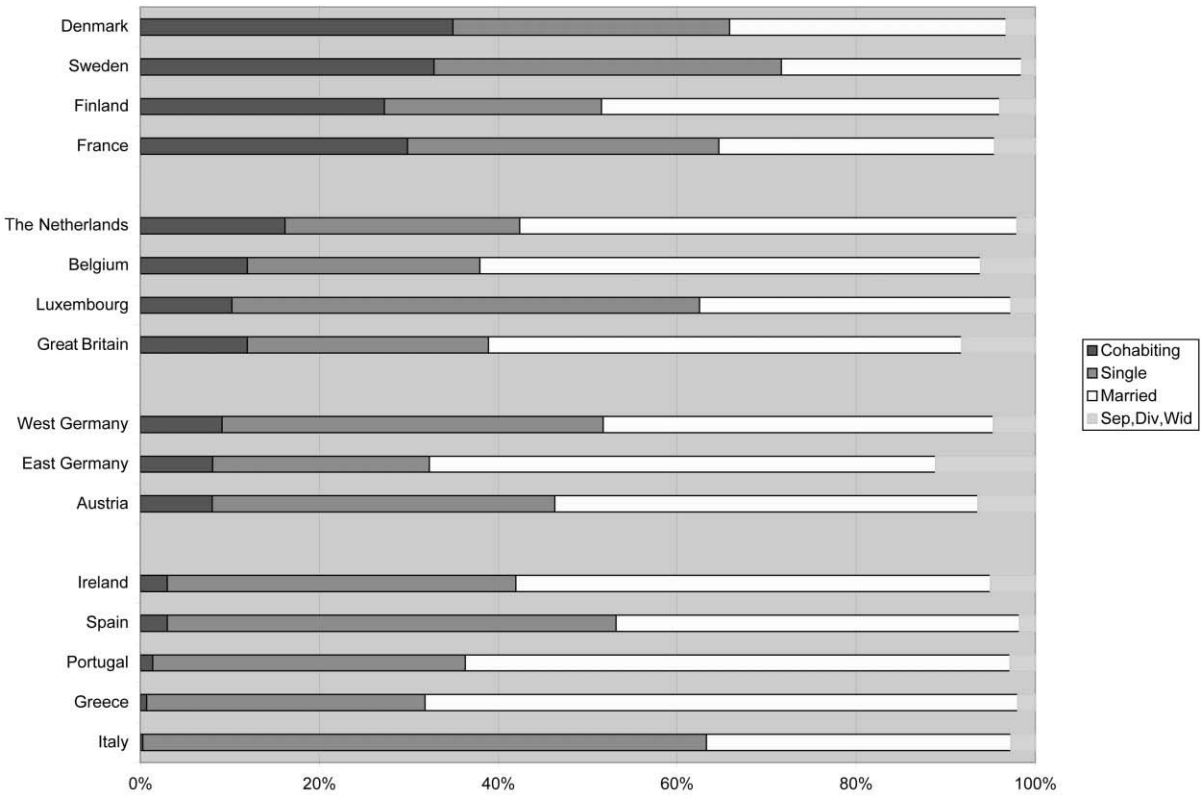

Figure 1
Marital status distribution of women aged 25–29 in 1996

Sweden, and Finland, and France also has relatively high proportions cohabiting. There is also a middle group of countries, which includes The Netherlands, Belgium, Great Britain, Germany, and Austria with mid-levels of cohabitation. Data for the US and Australia suggest that they would fall into this middle group. At the other extreme is the set of southern European countries and Ireland, where cohabitation is seemingly much rarer with only a tiny minority cohabiting.

1.2.2 Type of first partnership. The UN ECE Fertility and Family surveys carried out in the main during the first half of the 1990s included a full partnership history that incorporated dates of marriages and any other co-residential heterosexual intimate relationships. Such histories permit more in-depth examinations of partnership formation and dissolution than can be gleaned from vital registration data or cross-sectional surveys that only include current status information.

In many Western nations there have been large increases in the proportions of couples cohabiting, and nowadays cohabitation rather than marriage marks the formation of a union. Evidence on this can be seen in Table 1, which shows for two recent cohorts of women the proportions who entered their first partnership at marriage. It is clear from these data that the younger women, those aged 25–9, were much less likely to have commenced their first partnership at marriage than the older women. There are marked reductions to be seen in the proportions of women who married directly without cohabiting in most countries; for example, in France 55 percent of the older women but only 21 percent of the younger women married directly, a pattern that is repeated across many of the nations. The main exceptions are Sweden and the southern European countries. In Sweden, cohabiting rather than marrying was already well established among the older women whereas in Italy and Spain there are indications of a rise in cohabitation; but for the majority of women marriage still heralds the start of the first partnership. This is in contrast with the Scandinavian and other Western European nations where it is a minority practice.

Table 1
Percentage of women marrying directly among those who had had a first partnership according to current age group

Country	Age group	
	25–29	35–39
Sweden[a]	8	10
Norway[a]	10	42
Finland	17	31
France	21	55
Austria	21	33
Switzerland	19	31
West Germany	19	45
Great Britain	37	72
Italy	87	92
Spain	81	91

Source: Analysis UN ECE Fertility and Family Surveys and British Household Panel Study, Kiernan 1999
a Nearest equivalent cohort

It is also the case that in many European countries cohabiting unions have simply replaced the marriages of yesteryear, in that compared with the recent past there has been little change in the proportions of men and women who have formed a residential partnership by their mid-20s, whereas in other countries (most noticeably the southern European states) cohabitation is only part of the story in the decline in marriage rates (Kiernan 1999). Here, young people have been spending longer periods of time as solos than in the recent past; living with their parents (in the main), on their own, or sharing with others (European Commission 1998).

1.2.3 Duration of cohabiting unions. In many developed countries cohabitation has eclipsed marriage as the marker for entry into the first union, but subsequently many of these unions convert into marriages and others dissolve. Life-table estimates of the proportions of cohabitations that had converted into marriages or dissolved five years after union formation for a range of European countries are shown in Table 2. There is some variation in the propensity to marry across nations and age groups. Sweden exhibits the lowest conversion to marriage, with only one in three cohabitations having become marriages within five years of the start of the partnership whereas in most other countries more than one in two cohabitations had converted into marriages by the fifth anniversary of the union. In several countries there are indications of a decline in the propensity to marry over time, most noticeably in Norway and France, whereas in other countries, such as Switzerland, there is little sign of change. Turning to the extent to which cohabiting unions dissolve, we see from Table 2 that in most countries among those aged 25–9 years, around one in three unions had dissolved by the fifth anniversary of their start.

Table 2
Proportions (derived from life-table analysis) of first cohabiting unions that had converted to marriages or dissolved by five years of start of union by age of woman

	Married 5 years	Dissolved 5 years
Sweden		
1964[a]	34	37
1954	44	24
Norway		
1960[a]	56	35
1950	81	29
Finland		
25–29	60	31
35–39	66	21
France		
25–29	63	31
35–39	78	17
Great Britain		
25–29	58	36
35–39	50	41
Austria		
25–29	54	26
35–39	50	18
Switzerland		
25–29	67	38
35–39	70	26
West Germany		
25–29	57	36
35–39	51	17
East Germany		
25–29	42	27
35–39	26	15

Source: Analysis of UN ECE Fertility and Family Surveys and British Household Panel Study
a Nearest equivalent cohort

1.2.4 Who cohabits? In addition to cross-national variation in union-formation behavior, there is also variation within nations and between subgroups of the population (see Carmichael 1995 for a review). There is now robust evidence that in most nations the younger generations are more prone to cohabit than were older generations, and the more secular mem-

bers of a society and those who had experienced parental divorce during childhood are also the more likely to cohabit. There is also evidence that those residing in metropolitan areas are more prone to cohabit and that being in full-time education tends to inhibit union formation, but the association of level of educational qualifications and employment status with cohabitation is less clear-cut and tends to vary across nations.

1.2.5 A partnership transition? Many developed societies may be going through a transition in the way that men and women become couples or partners (Prinz 1995). Drawing on the experience of Sweden, which is the nation that has gone furthest in these developments, a number of stages can be identified (Hoem and Hoem 1988). To simplify, in the first stage, cohabitation emerges as a deviant or *avant-garde* phenomenon practiced by a small group of the single population, while the great majority of the population marry directly. In the second stage, cohabitation functions as either a prelude to marriage or as a probationary period where the strength of the relationship may be tested prior to committing to marriage, and this is predominantly a childless phase. In the third stage, cohabitation becomes socially acceptable as an alternative to marriage and becoming a parent is no longer restricted to marriage. Finally, in the fourth stage, cohabitation and marriage become indistinguishable with children being born and reared within both, and here the partnership transition could be said to be complete. Sweden, Denmark, and, to a lesser extent, France are countries that have made the transition to this fourth stage. These stages may vary in duration, but once a society has reached a particular stage it is unlikely that there will be a return to an earlier stage. Moreover, once a certain stage has been reached, all the previous types of cohabiting unions can co-exist. Such stages also have parallels at the level of the individual. At any given time, cohabitation may have different meanings for the men and women involved; for example, it may be viewed as an alternative to being single, or as a precursor to marriage, or as a substitute for marriage. Moreover, how a couple perceives their cohabitation may change over time and may also vary between the partners. Dissecting cohabitation in this way highlights the diversity of the phenomenon and suggests that more so than marriage it is a process rather than an event.

In sum, analyses of recently available data on union formation show there to be marked variation in the ways men and women are forming partnerships across developed nations. In the 1990s marriage was still the pre-eminent marker for entry into first union in the southern European countries, whereas in most Western European countries, and in North America, and Australia, and New Zealand, cohabitation has eclipsed marriage as the marker for first partnership. Notably, in the Nordic countries, long-term cohabitation has become more prevalent. Most Western countries appear to be experiencing changes in the ways that men and women become couples, but the question of whether most countries are on a trajectory to an ultimate destination where marriage and cohabitation are largely indistinguishable or even where cohabitation overtakes marriage as the dominant form of union awaits the future for an answer.

2. Partnership Dissolution

2.1 The Rise of Divorce

The other major development in partnership behavior is the rise of divorce. This has brought to the fore fundamental issues about the roles of men and women in society and the care and support of children. In most Western nations death still terminates the majority of marriages, but over the last few decades marriages have been increasingly dissolved by divorce before death has made any significant inroad and at a stage in the marriage when there are likely to be dependent children. In most Western countries divorce has increased since the mid-1960s, following a period of very stable divorce rates throughout the 1950s and the early part of the 1960s. Figure 2 shows trends since the early 1970s in the extent of divorce as measured by the number of divorces per 1,000 population for a range of Western countries. At the beginning of the 1970s the highest divorce rates were to be found in the US, and in Denmark and Sweden among countries in Europe. Divorce rates increased during the 1980s in most countries, and since then the rates have stabilized in many countries. Between 1960 and the mid-1980s divorce policy was either completely revised or substantially reformed in almost all the Western countries (Phillips 1988). Most countries liberalized their divorce laws, moving from fault-based divorce to no-fault divorce laws whereby fault, responsibility, or offense was no longer attributed by the law to either spouse. Following the liberalization of divorce laws, divorce rates in many countries continued their upward trend, frequently at a faster pace than in the years preceding legislative changes, followed by a period of stabilization from the mid- to late 1980s.

In the mid-1990s there was still a good deal of variation in the level of divorce across Western nations. In 1995 the total divorce rate in the EU was 0.30, whereas in 1970 it had stood at 0.11 (Eurostat 1997). This indicator is an estimate of the mean number of divorces per marriage for a notional cohort subjected to current divorce rates at each marriage duration. It can broadly be interpreted to mean that if the propensity to divorce remained unchanged over time, 30 percent of marriages would end in divorce. If

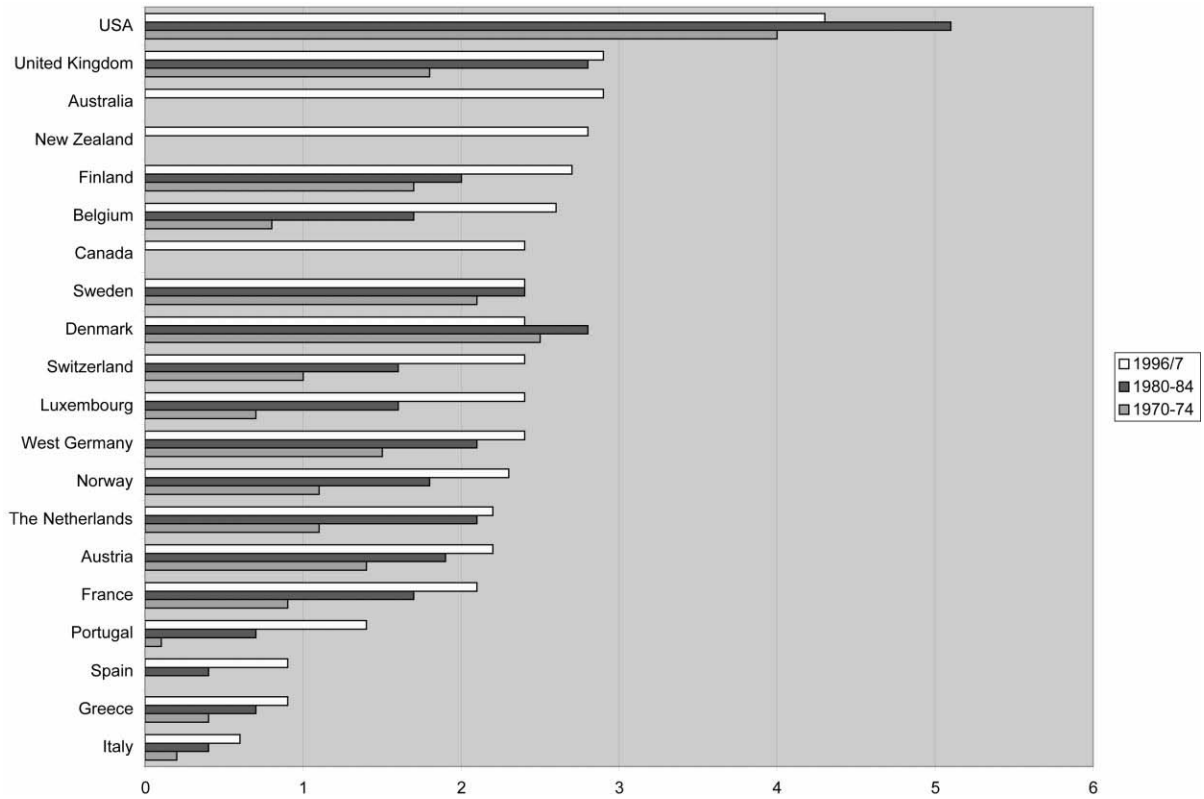

Figure 2
Crude divorce rates: divorces per 1,000 population

current rates were to prevail, nearly one in three marriages in the EU would dissolve. This is a lower level than in Canada (44 percent) and the US, where the total divorce rate has been above 50 percent since the late 1970s. Within Europe there is also a good deal of diversity in the level of divorce. During the 1990s three distinct divorce regions can be distinguished. In the Nordic countries and in England and Wales, the total divorce rate has been consistently above the 0.40 level. In Western European nations (Austria, Switzerland, Germany, France, Belgium, Netherlands, Iceland, and Luxembourg), the indicator lies between 0.30 and 0.40. In the southern European countries (Greece, Portugal, Spain, and Italy) it is below 0.20.

Divorce statistics have invariably underestimated the totality of marital dissolutions, and with the rise of cohabitation even fewer partnership dissolutions are captured by divorce registration data. In countries where divorce is more long-standing and more prevalent, there has been a tendency for marriages to break-up sooner. As a consequence, more couples are likely to be childless, and among couples with children the children are likely to be younger. Rising divorce has led to a growth in lone-parent families, the residential separation of fathers from their children, and remarried couples and stepfamilies.

2.2 Lone-parent Families

The prevalence of lone parenthood varies considerably between countries, and the proportion of families headed by a lone parent has been increasing just about everywhere, as can be seen in Table 3. As yet, no Western European country has matched the US, where more than one in four families with children are lone-parent families. Various reports made for the EU (for example, Bradshaw et al. (1996) show that the great majority of lone-parent families are headed by a woman (80–90 percent). The largest group of lone parents comprises those who have experienced a marital breakdown, the next largest group comprises widows, and lone mothers who had never been married (but not necessarily never partnered) are the smallest category. In many countries where there have been

Table 3
Lone-parent families as a percentage of all families with dependent children

Country	1980/81[a]	1990/91[a]	1996[b]
United Kingdom	14	16	23
Denmark	18	21	-
Finland	-	-	17
France	8	11	15
Belgium	9	15	15
Germany	10	15	13
Ireland	7	11	13
Portugal	12	12	12
Luxembourg	9	12	11
The Netherlands	8	12	11
Italy	7	6	11
Spain	5	8	8
Greece	4	6	7

Source: Bradshaw et al. 1996
a Eurostat LFS Social Europe 1/94—The European Union and the Family. b Eurostat Statistics in Focus: Population and Social Conditions 1998/12. Data in b from Labour Force Surveys

marked increases in childbearing outside marriage (but again not necessarily outside a partnership, see *Fertility of Single and Cohabiting Women*), this ordering may well have changed, such that never-married women with children may now constitute the second largest group of lone-mother families. Overall, the majority of lone-parent families emanate from the break-up of partnerships, either marital or cohabiting ones.

There is a good deal of evidence that children who experience the break-up of their parents' marriage or nonmarital union are more likely to experience poverty or reduced economic circumstances than children who grow up with both natural parents, but that the depth of poverty varies across nations (Bradshaw et al. 1996). The financial exigencies associated with marital breakdown arise from the precarious economic position of lone mothers, with whom most children live, and the dis-economies of scale associated with the maintenance of two households when fathers live apart from their children. The low incomes of lone-mother families are due to a combination of factors: low earnings from employment, lack of or low levels of child support from the natural father, and inadequate state support.

2.3 Remarriage

Being reared by a lone parent is frequently not a long-term arrangement, as a substantial proportion of divorced persons eventually remarry. Men are even more likely than women to remarry and are also more likely to remarry more quickly after a divorce. As well as being more likely to remarry, divorced men are more likely to cohabit than are divorced women. Remarriages are also at greater risk of dissolution than are first marriages. After an initial upsurge in remarriage rates in the early years following the enactment of more lenient divorce legislation, which occurred in most European countries, remarriage has taken a downturn due to some extent to postmarital cohabitation becoming more common.

2.4 Children and Divorce

Parental divorce and its aftermath constitute a major factor in the collective make-up of the generation of children born since the 1970s in northern European countries, as they did for the generations born since the 1960s in the US, and from whose experiences much of our knowledge on this topic emanates. This factor is increasingly being experienced by the generation born during the 1980s in many other European countries. Parental separation has been shown to affect the lives of children in both the short and long term (Amato and Keith 1991a, 1991b). Following their parents' separation, children frequently go through a crisis, during which behavior problems at home and at school are more often reported, worries become more prevalent, and anxiety levels increase. After divorce, families may have to move house through necessity rather than choice, which in turn leads to changes in schools, and neighborhood and social networks. Poverty or at least reduced economic circumstances are likely to be a prominent feature of these children's lives.

A number of studies from a range of countries have shown that children who experience the breakup of their parents' marriage subsequently have lower educational attainment, receive lower incomes, and are more likely to be unemployed and to be in less prestigious occupations in adult life than their contemporaries brought up by both parents (Jonsson and Gähler 1997, McLanahan and Sandefur 1994). In the demographic domain, young women who have experienced parental divorce are more likely than their peers to commence sexual relations early, to cohabit or marry at young ages, to bear children in their teens, and to conceive and bear children outside wedlock (Kiernan and Hobcraft 1997, McLanahan and Sandefur 1994). Men and women from disrupted families are also more likely to experience the breakup of their own marriage (Glenn and Kramer 1987, Kiernan and Cherlin 1999).

3. Conclusion

Across many Western nations there have been dramatic changes in partnership behavior, most noticeably the rise of cohabitation and divorce. The general direction of change in most Western countries is similar but there continues to be marked variations

across nations (and probably within nations) in the extent to which these developments have taken hold. Marriage once signaled the start of a lifetime partnership for the great majority of couples, but it is increasingly being replaced by more flexible and contingent partnerships that have no formal commencement and that continue only as long as both partners derive satisfaction from the relationship. A range of sociological and economic explanations has been posited to account for these changes (see *Family Theory: Complementarity of Economic and Social Explanations*).

See also: Lone Mothers in Affluent Nations

Bibliography

Amato P R, Keith B 1991a Parental divorce and the well-being of children: A meta-analysis. *Psychological Bulletin* **110**: 26–46

Amato P R, Keith B 1991b Parental divorce and adult well-being: A meta-analysis. *Journal of Marriage and the Family* **53**: 43–58

Bradshaw J, Kennedy S, Kilkey M, Hutton S, Corden A, Eardley T, Holmes H, Neale J 1996 *The Employment of Lone Parents: A Comparison of Policy in Twenty Countries*. Family Policy Studies Centre/Joseph Rowntree Foundation, London

Carmichael G 1995 Consensual partnering in the more developed countries. *Journal of the Australian Population Association* **12**: 51–86

Council of Europe 1999 *Recent Demographic Developments in Europe*. Council of Europe, Strasbourg, France

Espenshade T J 1985 The recent decline of American marriage. In: Davis K (ed.) *Contemporary Marriage*. Russell Sage Foundation, New York

European Commission 1996 *Eurobarometer No. 44*. ESRC Data Archive, Essex, UK

European Commission 1998 *Social Portrait of Europe*. Office for Official Publications of the European Communities, Luxembourg

Eurostat 1997 *Statistics in Focus: Population and Social Conditions No. 14*. Office for Official Publications of the European Communities, Luxembourg

Festy P 1980 On the new context of marriage in Western Europe. *Population and Development Review* **6**: 311–5

Fitch C A, Ruggles S 2000 Historical trends in marriage formation, United States 1850–1990. In: Waite L C, Bachrach C, Hindin M, Thomson E, Thornton A (eds.) *Ties that Bind: Perspectives on Marriage and Cohabitation*. Aldine de Gruyter, Hawthorne, IL

Glenn N D, Kramer K B 1987 The marriages and divorces of children who divorce. *Journal of Marriage and the Family* **49**: 811–25

Hoem J, Hoem B 1988 The Swedish family: Aspects of contemporary developments. *Journal of Family Issues* **9**: 397–424

Jonsson J, Gähler M 1997 Family dissolution, family reconstitution and children's educational careers: Recent evidence from Sweden. *Demography* **34**: 277–93

Kiernan K 1999 *Cohabitation in Western Europe*. Population Trends, no. 96. Stationery Office, London

Kiernan K, Cherlin A 1999 Parental divorce and partnership dissolution in adulthood: Evidence from a British cohort study. *Population Studies* **53**: 39–48

Kiernan K E, Hobcraft J N 1997 Parental divorce during childhood: Age at first intercourse, partnership and parenthood. *Population Studies* **51**: 41–55

McLanahan S, Sandefur G 1994 *Growing Up with a Single Parent*. Harvard University Press, London

Phillips R 1988 *Putting Asunder: A History of Divorce in Western Society*. Cambridge University Press, Cambridge, UK

Prinz C 1995 *Cohabiting, Married, or Single*. Ashgate Publishing, Aldershot, UK

United Nations 1992 *Questionnaire and Codebook: Fertility and Family Surveys in Countries of the ECE Region*. United Nations, New York

United Nations (annual) *Demographic YearBook*. United Nations, New York

K. Kiernan

Party Identification

Party identification is one of the central explanatory concepts used in the analysis of voting behavior in mature democracies. It refers to the enduring affective (emotional) attachment a voter feels towards a particular political party which disposes him or her to vote for that party in elections. Republican 'identifiers,' for example, are much more likely than other voters to vote Republican in a range of different elections; Democrat identifiers are similarly more likely to vote Democrat. A distinction is sometimes made between 'strong' and 'weak' identifiers, depending on the strength of the individual's attachment to the party in question. Both of these groups are contrasted with 'floating' or 'independent' voters, who exhibit no obvious attachment to one party over time and who frequently switch their votes between parties at succeeding elections. Party identification is sometimes referred to as 'partisanship.' The two terms are used interchangeably here.

1. Intellectual Context

The concept of party identification was originally developed as an integral part of the Michigan social psychological model of voting behavior. Partly through socialization experiences in childhood and in the workplace, individuals come to feel that their own political beliefs and values are 'closer' to the core ideological and policy positions of one party rather than another. Some individuals will enter adulthood as fully fledged 'identifiers' with a particular party. For others, the process may take longer. But the more frequently that an individual votes for a particular party over time, the more likely it is that s/he will come to 'feel closer to' or 'identify with' that party: s/he will

become a 'party identifier.' This, in turn, will reinforce the individual's tendency to continue to vote for the party, even when other factors might militate against it.

Although party identification is primarily a psychological concept, it also has a direct counterpart in rational choice theory in Downs' concept of the 'standing vote.' According to Downs, voters can develop a habit of voting for one party as a way of reducing the costs of continuously collecting and evaluating political information. Relatively early in their political lives, they make judgements about which party is closest to their own ideological positions and tend subsequently to continue to support that party unless and until some major political trauma causes them to reassess their early judgement. Fiorina's work on voting in the USA goes some way to combining the psychological notion of affective attachment with Down's idea of the standing vote. Fiorina argues that party identification is best seen as a running tally, or cumulation, of voters' retrospective evaluations of the major parties. In short, identification results from a voter's perception that a particular party has outperformed its rival(s) over time; it disposes the voter to support that party in the current election.

2. *Emphases in Current Theory and Research*

The concept of party identification has been deployed in four main ways. The first, and most extensive, use is as a 'nonproximate' cause of vote choice. Voters are obviously swayed by a variety of different factors in deciding how to cast their votes in elections. These include voters' perceptions of any triumphs or disasters that occur in the final weeks of a campaign, their perceptions of the candidates' character; their views of the policy stances of the parties; and their assessments of the parties' performance in office and in opposition.

For advocates of the party identification approach, a crucial additional factor is the extent to which voters are predisposed—independently of any of the above 'proximate' factors—to support one or other of the parties because they 'identify' with it. On this account, the effects of proximate causes of the vote can only be properly estimated if the underlying effects of voters' prior partisan identifications have been fully taken into account. Concomitantly, explanations of voting choices (and election outcomes) that underplay the importance of long-term party identifications risk overstating the significance of the more proximate factors that explain voting choices. In all this, there is no suggestion that party identifiers always vote for the party with which they identify. On the contrary, it is entirely possible for even strong identifiers on occasion to vote against 'their' party because other more proximate factors weigh more heavily in their vote calculations. In US presidential elections, for example, between five and twenty percent of identifiers typically vote 'against' the party they identify with. For these voters, party policies, party performance or the leadership qualities of the rival candidates (or some other proximate factors) must have weighed more heavily than traditional party loyalties. Party identification theory argues, however, that in subsequent elections 'defecting identifiers' are highly likely, other things being equal, to bring their voting behavior back into line with their partisan identity; they return to their traditional partisanship.

The second usage of the concept of party identification follows from the first. If prior identifications dispose citizens to vote in particular ways, then those identifications are also likely to color the way in which voters interpret political information. Party identification in this sense acts as a framing factor—a sort of perceptual filter—which disposes voters to view *their* party's policies, its leaders, and its record in a more favorable light than might otherwise be the case. It simultaneously disposes voters to view *other* parties' characteristics less favorably. This implies that voting choices are not only affected *directly* by party identification. These choices are also affected *indirectly* because of the 'coloring' effects of identification on voters' perceptions of the parties' competing attractions—perceptions which in turn also influence the way people vote.

A third use of the concept of party identification is made by the political parties themselves in devising electoral strategy. Party strategists are always aware of the need both to mobilize their party's own identifiers and to appeal to those voters who in the past have not been inclined to lend the party their support. In the early stages of a campaign, the main task is to trigger the support of existing partisans; in later stages it is to convert neutrals and those who identify with competing parties. Measures of party identification give strategists some idea of the size of their 'core vote.' This is the minimum level of support that parties can expect to obtain in a general election and against which, once the election outcome is known, their strategists can judge their success in recruiting new supporters. The need to provide participants in the democratic process with accurate information about their performance puts a particular premium on the accuracy of estimates of partisanship levels. For this reason debates about the measurement of party identification (which are discussed below) involve more than just matters of academic interest.

The final usage of the concept of partisanship is as a phenomenon its own right that requires analysis and explanation. To argue, as party identification theory does in relation to any given election, that a major reason why voters vote Republican or Democrat—or Conservative or Labour—is that they are already Republican or Democrat, identifiers does not, on the face of it, seem to explain very much. Indeed, it borders on the tautological. The notion that 'I vote

Republican because I am a Republican identifier' immediately begs the question as to why I identify with the Republicans in the first place. As implied above, party identification theory's answer to this question is that partisanship develops as a result of long-term exposure to a variety of socialization processes in the individual's family, school, workplace, and social milieu.

Note, however, that socialization experiences of this sort fail to provide an obvious explanation as to why some voters *switch* their identifications (from Republican to Democrat or from Labour to Conservative, for example) midway through their voting lives. Fiorina's notion of identification as a cumulative updating of retrospective assessments of party performance plays a useful explanatory role in this context. Consider a voter A who, as a result of childhood and early adult socialization experiences, identifies with—and accordingly votes for—the Conservatives at election e1. Assume further that the Conservatives appear so incompetent at the subsequent elections e2 and e3 that A switches his vote to Labour on both occasions. In these circumstances, A may also shift his identification to Labour to reflect the fact that his 'running tally' of retrospective performance now favors Labour rather than the Conservatives. Although Fiorina's analysis does not predict the precise point at which voters will switch their identification from one party to another, it offers a plausible explanation as to how, notwithstanding their early political socialization, voters can change their identifications over time.

3. *Methodological Issues: The Problem of Measurement*

A key methodological issue concerns the extent to which the party identification measures developed in the USA in the 1950s can be successfully transplanted to other political contexts. Data from successive US National Election Studies (NES) suggest that roughly three-quarters of US voters are either Republican or Democrat identifiers (the remaining quarter are 'independents') and that substantial proportions of voters retain their partisan identities over time. Given that these findings are consistent with the idea of party identification being both widespread and durable, the concept of partisanship has maintained a prominent position in analyses of electoral outcomes and electoral change. Indeed, most published accounts of US voting behavior include either controls for, or estimates of, the effects of partisanship on the vote.

Not surprisingly, when efforts have been made to extend the theories and methods of the NES to other countries, the concept of party identification has been 'exported' as a central part of the package. An early example was the inclusion of a party identification question in the 1964 British Election Study (BES) survey. The question was borrowed directly from the NES and it has been employed with only minor variations in question-wording since:

> Generally speaking, do you think of yourself as Conservative, Labour, Liberal-Democrat, (Nationalist) or what?

The phrase 'generally speaking' is intended to tap the idea that party identification—as opposed to the more immediate and ephemeral notion of current party preference—endures over time; 'think of yourself' attempts to capture the respondent's political identity.

Most analysts of British electoral politics would endorse the idea that party identification plays an important role in explaining UK voting decisions. However, considerable doubt has also been expressed as to whether the traditional partisanship question measures anything very different from voters' current party political preferences. There are two major reasons for supposing that it does not. First, responses to the standard question outlined above suggest that, since the 1960s, 90–5 percent of British voters have had some sort of party identification. This is clearly a much higher level of 'measured' partisanship than in the USA, which should in turn imply much greater stability over time in British voting patterns. Yet electoral volatility since the 1970s has not been noticeably lower in the UK than in the USA. Second, it is clear from British data that major shifts in party identification levels (in 1979, 1983, and 1997) have coincided exactly with major shifts in voting behavior —implying that both, perhaps, are indicators of current political preferences rather than separate phenomena. This conclusion is supported by monthly time-series data collected by Gallup which show that short-term fluctuations in British partisanship correlate very strongly with short-term changes in party support levels. More formal tests, based on the principles of 'convergent validity' indicate that, in Britain at least, extant measures of party identification are statistically indistinguishable from measures of voting preference. This in turn implies that models of vote choice which include party identification as an explanatory variable are tautologically mis-specified.

Although these difficulties have not been precisely replicated in all of the countries where partisanship measures have been deployed, similar problems have been encountered in a variety of different national contexts. As a result, many analysts outside the USA have expressed concern that in non-American contexts (where the concept of the 'political independent' is less well developed) the standard NES question invites respondents to admit to a partisan identity even when they do not have one. This has led a number of researchers to advocate the use of a 'filter' mechanism in measuring partisanship, such as the following:

> I am now going to ask you a question about party politics in [country]. Some people think of themselves as usually being a

supporter of one political party rather than another. Setting aside your current views about the government and opposition parties, do you usually think of yourself as being a supporter of one particular party or not?
If YES, which one is that?

The key feature of the filter is that it allows survey respondents to indicate from the outset that they do *not* usually think of themselves as being supporters of a particular political party. Survey experiments using filters of this kind have been conducted in Canada, the UK, the USA, and elsewhere. The measured levels of party identification fall considerably as a result. In Canada, switching from the traditional question to the filter approach reduces measured party identification levels from 70 percent to 40 percent of the electorate; in the UK, identification falls from around 90 percent to under 50 percent; and in the USA it falls from roughly 70 percent to 60 percent. The effects of using a filter, in short, are very considerable outside the USA, implying that levels of party identification in other countries may have been significantly overestimated in much academic research where the traditional NES question has been imported for local use. This is obviously important in purely descriptive terms but it also suggests significant measurement error problems in any statistical models that (outside the USA) have used party identification measures.

The problems of measuring party identification effectively do not stop here, however. A critical feature of partisanship is that it is, for the individual voter, an enduring feature of his or her political makeup. Partisanship should therefore be quite stable at the individual level over time, which implies that an individual should respond to the party identification question in a consistent way over time. Since 1991, the British Household Panel Survey (BHPS) has been collecting annual data on party identification in the UK using the sort of 'filter' question outlined above. In every wave of the panel, under 50 percent of respondents have indicated that they have a party identification. In the general election year of 1997, for example, 28 percent of the sample were Labour identifiers, 14 percent were Conservative identifiers, and 4 percent Liberal Democrat identifiers. As intimated earlier, these levels are clearly much lower than those recorded on the basis of the traditional NES question wording. But the measurement problem extends much further than this. Party identification is supposed to be a stable and enduring characteristic of individual voters. The BHPS reinterviews the same respondents in successive annual waves of the survey. We can therefore inspect the data to see how far individuals retain the same party identifications over time. As Table 1 shows, even if we consider only a 3-year 'rolling window' (i.e., we consider voters responses in waves 1–3, then waves 2–4, then 3–5, and so on), the levels of consistent identifiers fall even further. For example, in waves 1, 2, and 3 (1991–1993) only 9.5 percent of respondents consistently responded as Conservative identifiers and only 7.9 percent consistently as Labour identifiers. In waves 5, 6, and 7 (1995–7), only 6.9 percent were consistent Conservative identifiers and 12.5 percent Labour identifiers. The remainder (excluding the trivial number of consistent Liberal Democrats) were either consistent nonidentifiers (33.8 percent) or else switched identifications across waves (45.8 percent). In short, only around 20 percent of British voters in the 1990s were 'genuine' party identifiers in the sense that the measuring instrument used: (a) allowed them to indicate from the outset that they had no particular attachment to any party; and (b) required them to declare the same general partisan commitment on three (and only three—the figures are even lower if the length of the rolling window is extended!) successive occasions.

Table 1
Percent of BHPS respondents with consistent and inconsistent party identifications over successive three-year periods, 1991–7

	W1–3	W2–4	W3–5	W4–6	W5–7
Consistent Conservative	9.5	8.9	7.6	7.4	6.9
Consistent Labour	7.9	9.4	10.5	11.1	12.5
Consistent Lib Democrat	1.3	1.1	1.0	0.9	1.0
Inconsistent	52.1	50.2	46.8	43.2	45.8
Non-identifier	29.2	30.4	34.1	37.4	33.8

Source: British Household Panel Survey

Where does this leave us? Party identification is a theoretical construct that makes a great deal of intuitive sense. It tells us that individuals have long-term, enduring attachments to political parties and this idea in turn helps us to understand why individuals often vote for the same party in successive elections. Yet when we try to measure partisanship, things become far less clear. Generations of American scholars have happily employed the concept, confident that the NES question wording effectively captures the enduring nature of the partisan's affective commitment to his or her party. However, data from other countries—and especially from Britain—casts considerable doubt on the use of the concept in other national contexts. In the UK, large numbers of voters—much larger than are revealed by the NES question—appear to have no commitment to any political party. And of those that do, less than 20 percent retain that commitment over a three-year period. In short, if party identification denotes an *enduring* affective attachment to a particular political party, then partisanship in the UK at least is very much a minority predisposition.

Future research which seeks to employ the notion of party identification will need to address two major issues. The first, discussed above, is the need to develop operational empirical measures of the concept that are both durable and have substantive meaning in the specific national contexts in which they used. Critically, these new operational measures must distinguish effectively between voting choice and partisanship. The second issue concerns the impact of the longer-term process of 'partisan dealignment.' It is generally acknowledged that, however partisanship is measured, levels of party identification have gradually fallen in most democracies since the 1970s. Electorates have become increasingly volatile and voters increasingly prepared to behave like 'electoral consumers,' picking and choosing between parties on the basis of the rival policy and leadership packages on offer. In these circumstances, as Fiorina observed, partisan attachments acquired early in voters' political lives are likely to be less important in determining voting choices than the more temporary 'brand loyalties' associated with the perception that 'party X currently appears more competent in managerial terms than party Y.'

See also: Attitude Formation: Function and Structure; Citizen Participation; Identity: Social; Ideology: Political Aspects; Ideology, Sociology of; Participation: Political; Political Culture; Political Discourse; Political Parties, History of; Political Protest and Civil Disobedience; Polling; Public Opinion: Microsociological Aspects; Public Opinion: Political Aspects; Socialization: Political

Bibliography

Butler D, Stokes D 1974 *Political Change in Britain: The Evolution of Electoral Choice*, 2nd edn. Macmillan, London

Campbell A, Converse P, Miller W E, Stokes D 1966 *Elections and the Political Order*. Wiley, London

Fiorina M P 1981 *Retrospective Voting in American National Elections*. Yale University Press, New Haven, CT

Miller W E, Shanks M 1996 *The New American Voter*. Harvard University Press, Cambridge, MA

Sinnott R 1998 Party attachment in Europe: methodological critique and substantive implications. *British Journal of Political Science* **28**: 627–50

D. Sanders

Party Responsibility

Party responsibility is the attribution of collective responsibility to the political party in power for the performance of government. This contrasts with the attribution of individual responsibility to particular officeholders for their performance in office. Party responsibility is the core idea of a model, known as responsible party government, of how democratic government should be organized in a modern populous nation.

1. The Party Government Model

The party government model was first developed by the American writers Woodrow Wilson, Frank J. Goodnow, A. Lawrence Lowell, and Henry Jones Ford in the last two decades of the nineteenth century, and served as a base for criticizing the existing American parties and proposing party reform. It lost favor in the Progressive Era and languished until the 1940s, when it was revived by another generation of American writers led by E. E. Schattschneider in *Party Government* (1942) and *The Struggle for Party Government* (1948), with support from Herman Finer in *The Theory and Practice of Modern Government* (1932) and James MacGregor Burns *in The Deadlock of Democracy* (1963). Its visibility was maximized in 1950 with the publication of *Toward a More Responsible Two-party System*, a report of the American Political Science Association's Committee on Political Parties.

The model is based on the proposition that democratic responsibility depends on the possession and exercise of power. If an individual or group does not control the actions of government, a sovereign people cannot effectively hold that individual or group responsible for those actions. Democratic governments in modern populous nations cannot be controlled by individual officeholders, not even US presidents. A member of Congress has only a small share of the total power over government, and he or she can be held responsible only for his or her votes and speeches, not for the performance of the entire government. Even the president cannot control the actions of Congress or the Supreme Court, and so cannot meaningfully be held responsible for governmental performance.

The only entity large enough to control the entire government and public enough to be held responsible for what it does or fails to do is a political party. Therefore, only the collective responsibility of the majority party can establish effective popular control of government.

The model has three basic requirements. (a) The parties must have near-perfect cohesion on all public issues: the members of each party must act together so that they can be held collectively responsible. (b) The voters must vote for a preferred party's candidates as representatives of the collectivity, not as isolated individuals. (c) The winning party must control the entire power of government, so that it can reasonably be held responsible for the government's actions and inactions.

Ideally, the model of responsible party government should work like this: at least two (and, preferably,

only two) unified, disciplined major political parties exist. In a pre-election campaign, each sets forth its proposals for measures to deal with the nation's problems, and the differences between the two sets of proposals are great enough to provide the voters with a meaningful choice. In the election the voters vote for certain candidates for Congress and the presidency or for Parliament, *not* because of their individual eloquence and probity, but because they are members of the party the voter thinks should control the government for the next term. One party receives a majority of the votes and public offices, and takes over complete governmental power and therefore assumes complete responsibility for what government does or fails to do. At the end of its term the voters decide whether the governing party's record is mainly good or bad. If they like it, they vote to return it to power. If they dislike it, they vote to replace it with the other party, which has pointed out the governing party's errors, and stands ready to take its place.

The doctrine's advocates generally present it as a model, a theoretical construct, the merits of which are to be assessed by theoretical reasoning and discourse. They do not wish its merits to be determined by how well or badly any actual governmental system performs. However, it is clear that most of them regard the government and party system of the United Kingdom as close enough to the model to provide a basis for answering questions about how the model would actually work in a nation like the United States.

2. *The Model Used*

From its origins the model of responsible party government has been used mainly to highlight certain perceived failings of the American party system and to prescribe remedies. Three such failings have received the most attention: first, the parties do not stand for anything. As Woodrow Wilson put it,

> Provided with parties in abundance, and entertained with very nice professions of political principle, we lack party responsibility. American parties are seldom called to account for any breach of their engagements, how solemnly soever those engagements may have been entered into … 'Platforms' are built only for conventions to sit on, and fall into decay as of course when conventions adjourn. Such parties as we have, parties with worn-out principles and without definite policies, are unmitigated nuisances. (*Committee or Cabinet Government?* Wilson 1925, p. 109)

Some commentators of Wilson's persuasion said that the American major parties are like bottles bearing different labels but equally empty.

They are uncohesive and therefore irresponsible. Votes in Congress rarely fall along strict party lines, and so it is impossible for voters to tell what each party—as distinct from its individual members—has done on the issue. Hence, neither party can meaningfully be rewarded for good governmental actions or punished for bad ones. As a result, the American people can choose only between an incumbent officeholder and his or her opponent, neither of whom has much power over—and therefore cannot assume responsibility for—the actions of government. As the American Political Science Association (APSA) committee put it, the American party system consists largely of 'two loose associations of state and local organizations, with very little national machinery and very little internal cohesion ….[both of which are] ill-equipped to organize [their] members in the legislative and executive branches into a government held together and guided by the party program' (American Political Science Association, Committee on Political Parties 1950, p.v).

Some critics of the APSA report pointed out that the American constitutional system is an insurmountable barrier to the achievement of party government. Especially the institutions of separation of powers and federalism make it impossible for one party, by winning a majority of the votes and offices in any single election, to get full power over all governmental machinery and assume responsibility for what it does or fails to do.

Some advocates of party government, however, were not impressed. Schattschneider pointed out that in the past parties have had great success in making changes in the constitutional system without altering a word in the Constitution—for example, in the reshaping of the Electoral College system for electing the president—and he concluded that there are no 'grounds for excessive pessimism about the possibilities of integrating party government with the constitutional system. The greatest difficulties in the way of the development of party government in the United States have been intellectual, not legal. It is not unreasonable to suppose that once a respectable section of the public understands the issue, ways of promoting party government through the Constitution can be found' (Schattschneider 1942, pp. 209–10). The APSA Committee added that when the electorate becomes convinced that responsible parties are the best device for achieving popular control of government, the battle will be largely won, for the voters will insist that each party act together cohesively in a truly responsible manner.

3. *The Model Criticized, the Parties Defended*

From its beginnings the responsible party government model has had its critics. Perhaps the most prominent early opponent was Moisei I. Ostrogorski, a Russian writing in French. He first wrote an analysis of the British party system (especially interesting because it attacked the system often hailed as the working model

of responsible party government). His basic charge was rooted in his belief that the only meaningful form of democratic responsibility is the individual responsibility of each elected representative to his or her constituents. The worst thing about the British system, he said, is that 'the ministers … easily hide behind the collective [party] title; however incompetent or culpable they may be, it is impossible to punish one of them without punishing all … '(Ostrogorski 1902, I, p. 716).

His later (1910) volume on American parties charged them with many sins, of which the worst is that they are permanent conspiracies of political bosses interested only in winning elections so as to control government patronage. They have no interest in policy issues, and the voters cannot affect the course of public policy by voting for one party or the other. The only solution is to restore parties to what they were originally intended to be: *ad hoc* assemblies of like-minded citizens concerned with a particular policy issue, which will dissolve when that issue is settled.

The debate about what role parties should play in America and how they should be organized to play that role lay dormant until 1940, when E. Pendleton Herring published *The Politics of Democracy*. Herring argued that the unreformed parties had performed many valuable functions for American government and society, notably overcoming the potential for total gridlock inherent in separation of powers and federalism, stimulating ordinary people's interest and participation in politics, and moderating the ferocity of interest groups' demands by aiming their platform compromises and electoral appeals at the members of as many interest groups as possible rather than standing firmly and clearly for the interests of a restricted number of groups and the total rejection of those of other groups.

Herring's lead was followed by the authors of several works stimulated by *Toward a More Responsible Two-party System*. Political scientists Julius Turner and Austin Ranney argued in the *American Political Science Review* that the doctrine of responsible party government makes sense only in a democracy whose people are firmly committed to the implementation of unlimited majority rule, while in fact the American people have shown that they are more committed to the protection of minority rights. This explains why ordinary Americans have been so resistant to adopting the responsible party government model and its features of party discipline and cohesion and so admiring of individual independence in their officeholders. This same argument was developed at greater length in Ranney and Kendall's textbook *Democracy and the American Party System* (1956).

These defenders of the existing parties did not, however, silence the doctrine of responsible party government. The Schattschneider position was carried on and updated by, among others, the political scientist James MacGregor Burns (1963) and the journalist David Broder (1972). Moreover, since the 1970s both parties in Congress have become noticeably more cohesive and ideologically more distinct than they were in Schattschneider's (or Ostrogorski's) time. On the other hand, the proportion of voters voting split tickets has increased (to about 25 percent in the 1996 elections), and one of its main consequences, divided party government, with one party controlling the presidency and the other controlling one or both houses of Congress—a situation incompatible with responsible party government—has become the rule rather than the exception.

The net result is that while parties and officeholders in some parts of the national government are behaving more as the model requires, the voters' weakening party loyalties and increasing tendency to vote for individual candidates with little regard for their party labels are behaving less so. The dispute between the model's advocates and critics continues, but the party system changes only in some parts and in opposite directions.

See also: Accountability: Political; Electoral Systems; Party Systems; Political Parties; Political Parties, History of

Bibliography

American Political Science Association, Committee on Political Parties 1950 Toward a more responsible two-party system. *American Political Science Review*, **XLIV** (Suppl.)

Broder D S 1972 *The Party's Over: The Failure of Politics in America*. 1st edn. Harper & Row, New York

Burns J M 1963 *The Deadlock of Democracy: Four-party Politics in America*. Prentice-Hall, Englewood Cliffs, NJ

Croly H 1909 *The Promise of American Life*. Macmillan, New York

Croly H 1914 *Progressive Democracy*. Macmillan, New York

Finer H 1932 *The Theory and Practice of Modern Government*. Methuen and Company, London, 2 Vols.

Ford H J 1898 *The Rise and Growth of American Politics*. Macmillan, New York

Ford H J 1909 The direct primary. *North American Review* **CXC**: 1–19

Goodnow F J 1900 *Politics and Administration*. Macmillan, New York

Goodnow F J 1902 Review of Ostrogorski's *Democracy and the Organization of Political Parties*. *Political Science Quarterly* **XVIII** (June): 332–4

Herring E P 1940 *The Politics of Democracy: American Parties in Action*. W. W. Norton, New York

Lowell A L 1896a *Essays on Government*. Houghton Mifflin, Boston

Lowell A L 1896b *Governments and Parties in Continental Europe*. Houghton Mifflin Company, Boston, 2 Vols

Lowell A L 1902 The influence of party upon legislation in England and America. Annual Report of the American Historical Association for the Year 1901. Government Printing Office, Washington, DC, I, pp. 321–542

Ostrogorski M I 1902 *Democracy and the Organization of Political Parties* [trans. Clarke F]. Macmillan, New York, 2 Vols

Ostrogorski M I 1910 *Democracy and the Party System in the United States*. Macmillan, New York
Ranney A 1962 *The Doctrine of Responsible Party Government*. University of Illinois Press, Urbana, IL
Ranney A 1951 Toward a more responsible two-party system: A commentary. *American Political Science Review* **45** (June): 488–99
Ranney A, Kendall W 1956 *Democracy and the American Party System*. Harcourt Brace, New York
Schattschneider E E 1942 *Party Government*. Farrar and Rinehart, New York
Schattschneider E E 1948 *The Struggle for Party Government*. University of Maryland Press, College Park, MD
Turner J 1951 Responsible parties: A dissent from the floor. *American Political Science Review* **XLV** (March): 143–52
Wilson W 1885 *Congressional Government*. Houghton Mifflin, Boston
Wilson W 1879 Cabinet government in the United States. *International Review* **7**(August): 146–63
Wilson W 1925 Committee or cabinet government? In: Stannard Baker R, Dodd W E (eds.) *The Public Papers of Woodrow Wilson: College and State*. Harper and Brothers, New York, I, pp. 95–129
Wilson W Leaderless government. *Public Papers* **I**: 336–59

A. Ranney

Party Systems

A party system is the system of interactions between political parties that results from their mutual competition or cooperation (see Sartori 1976, p. 44). For party systems to exist there must therefore be more than one party involved. One-party states do not have party systems. Beyond the requirement for a plurality of parties, party systems can be said to exist only when these interactions between the parties are patterned, familiar, and reasonably predictable. A plurality of parties is therefore a necessary but not sufficient condition for the existence of a party system. Parties that exist alongside one another, but that do not interact in any patterned or structured fashion, do not make for a party system as such.

While it may seem difficult to conceive of situations in which a plurality of parties exists without at the same time embodying a set of patterned interactions, two types of limiting cases can be identified. The first is where each of the parties involved exists in an entirely self-sufficient world, based on a closed electoral constituency, with neither its survival nor performance being in any way effected by the other parties in the polity. Even in this limiting case, however, the absence of interactions at the electoral level might well be accompanied by the need to interact—through competition or cooperation—at the parliamentary or governmental level. Thus, while no electoral party system would exist, it might well prove possible to speak of a party system in parliament. The second type of limiting case is where a new multiparty democracy is being established more or less from scratch, such as in the wake of a democratic transition, and where neither the new parties themselves, nor their interactions, have stabilized sufficiently to allow the identification of a distinct pattern. In these situations, a party system may be in the process of being structured or institutionalized but it cannot yet be said to have definitely emerged. In their survey of party systems in Latin America, Mainwaring and Scully (1995) adopt a similar distinction in specifying party systems as either 'institutionalized' or 'inchoate.' Strictly speaking, however, an inchoate or non-institutionalized party system is a contradiction in terms: to be a system is to be institutionalized.

1. *Classifying Party Systems*

Although scholars have paid relatively scant attention to what defines a party system as such, tending instead to assume the existence of a party system in all polities where there exists a plurality of parties, they have devoted considerable effort to distinguishing between different types of party system. The most conventional and frequently adopted approach to distinguishing party systems is based simply on the number of parties in competition, and the most common distinction involved here, which goes back to Duverger (1954), is very straightforward—that between two-party systems, on the one hand, and multiparty (i.e., more than two) systems, on the other. This particular classification was also believed originally to reflect a more fundamental distinction between more or less stable and consensual democracies, such as the United Kingdom and the United States, which were seen as typical of the two-party variety, and more or less unstable or conflictual democracies, such as Fourth Republic France, Italy, or Weimar Germany, which were seen as typical of the multiparty type. Although this simple association of numerical categories of party system with levels of political stability and efficacy was later undermined by reference to a host of smaller democracies that were characterized by both a multiplicity of parties and a strong commitment to consensual government, the core distinction between two- and multi-party systems has continued to be widely employed within the literature on comparative politics, although it is sometimes modified by taking into account not only the sheer numbers of parties in competition, but also their relative size. Thus, Blondel (1968), for example, uses the relative size of the parties to distinguish four types of party system: two-party systems, two-and-a-half-party systems, multiparty systems with a dominant party, and multiparty systems without a dominant party.

The most substantial attempt to move away from a primary reliance on the simple numbers of parties in competition was that of Sartori (1976, pp. 117–323),

who combined counting the parties with a measure of the ideological distance that separated them. Sartori's typology was the first that focused directly on the interactions between the parties—the 'mechanics' of the system—and hence on the differential patterns of competition and cooperation. Following this approach, party systems could be classified according to the number of parties in the system, in which there was a distinction between systems with two parties, those with up to some five parties (limited pluralism) and those with some six parties or more (extreme pluralism); and according to the distance that separated the parties lying at either extreme of the ideological spectrum, which would either be small ('moderate') or large ('polarized'). These two criteria were not wholly independent of one another, however, in that Sartori also showed that the format of the system, that is, the number of parties, contained 'mechanical predispositions,' that is, it could help determine the ideological distance, such that extreme pluralism could lead to polarization. When combined, the two criteria resulted in three principal types of party system: two-party systems, characterized by a limited format and a small ideological distance (e.g., the UK); moderate pluralism, characterized by limited or extreme pluralism and a relatively small ideological distance (e.g., Denmark or the Netherlands); and polarized pluralism, characterized by extreme pluralism and a large ideological distance (e.g., Weimar Germany, early postwar Italy, or pre-Pinochet Chile). Sartori also noted the existence of a 'predominant-party system,' in which one particular party, such as, for example, the Congress party in India, or the old Unionist party in Northern Ireland, consistently won a winning majority of parliamentary seats.

2. *Party Systems and Structures of Competition*

At the core of any party system is the competition for executive office. It is this that structures the party system in the first place, and facilitates its institutionalization. Moreover, albeit often implicitly, the competition for executive office is also the most important criterion employed in many of the various classifications of party systems. Two-party systems, for example, are referred to as such not because there only two parties present themselves to the voters—indeed, this is rarely if ever the case—but rather because only two parties matter when it comes to forming a government, be this in a presidential or a parliamentary system. In two-and-a-half or multiparty systems, by contrast, there are more than two parties that enjoy the potential access to executive office. Even within the more complex classification developed by Sartori, it is the competition for office that proves most important. Sartori's moderate pluralism, for example, involves competition between alternative coalition governments, whereas polarized pluralism involves a patterns in which a center party or parties is more or less permanently in office, with one or more extreme parties being permanently excluded from government.

It follows from this that party systems might best be understood not only in terms of some all-embracing classification, within which the question of the numbers of parties in competition obviously plays an important role, but also in terms of contrasting patterns of government formation (Mair 1997, pp. 199–223). Three related criteria are important here. The first concerns the ways in which governments alternate with one another, where systems in which there is a regular process of wholesale alternation, in which incumbents are replaced wholly by former nonincumbents, may be separated from those in which alternation is only partially complete. Systems in which two alternative parties or coalitions alternate with one another (the USA, the UK, Fifth Republic France) may therefore be distinguished from those in which one or more parties leaves office, while another remains, albeit often with a new coalition partner (Belgium, the Netherlands). The second, relevant criterion here is familiarity, where systems in which government is always contested by the same parties or sets of parties may be contrasted with those in which patterns of government formation prove more innovative or promiscuous. In postwar Ireland, for example, voters had long become accustomed to facing a choice at election time between, on the one side, the powerful *Fianna Fáil* party, which always governed alone, or, on the other side, a coalition of all of the smaller parties. Once *Fianna Fáil* decided to opt for a coalition strategy, however, in 1989, government formation became more innovative, and the formerly familiar patterns broke down. The third criterion involves the degree of access to the government formation process, and the extent to which new parties can hope to win a place in the executive. New parties have always found it relatively easy to join innovative government coalitions in the Netherlands, for example. In the UK, by contrast, no party other than the Conservatives or Labour has gained access to government since 1945.

Putting these three criteria together enables party systems to be ranged along a notional continuum according to the degree to which the structure of competition is open or closed. At one extreme lie the wholly closed systems, such as the UK or the US, where alternation in government is always wholesale, where government formation processes are wholly familiar, and where new or third parties are always excluded. At the other extreme, it is difficult to speak of a party system at all: there is no discernible pattern in how governments alternate, the potential alternatives themselves are unfamiliar and may never have governed before, and access to office is, in principle, open to all. To travel from this latter extreme to the other is therefore to witness the progressive closure of

the structure of competition, which is simply another way of saying that systemness itself increases. Party systems as systems are strongest and are most institutionalized when the structure of competition is closed. They are weakest, and more or less nonexistent, when this structure is wholly open.

To understand party systems in this way, and hence to treat the systemness of a party system, or its degree of institutionalization, as a variable, also affords an insight into the question of the persistence and change of party systems. Strongly institutionalized party systems will inevitably enjoy a bias towards persistence, not least because the parties at the core of such systems will have a vested interest in maintaining the existing structure of competition. It is for this reason, for example, that Lipset and Rokkan (1967) could speak of the 'freezing' of west European party systems in the wake of full democratization in the 1920s. Less institutionalized systems, on the other hand, will prove more susceptible to change. This approach also affords an insight into the relationship between party systems and individual parties, for it is not just the individual parties that together create the party system, but also the party system that acts to reinforce the role and standing of the parties themselves.

See also: Democracy; Dictatorship; Electoral Systems; Participation: Political; Party Identification; Party Responsibility; Political Parties, History of

Bibliography

Blondel J 1968 Party systems and patterns of government in western democracies. *Canadian Journal of Political Science* **1**: 180–203

Duverger M 1954 *Political Parties: Their Organization and Activity in the Modern State*. Methuen, London

Lipset S M, Rokkan S 1967 Cleavage structures, party systems and voter alignments: An introduction. In: Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments*. Free Press, New York

Mainwaring S, Scully T R 1995 Introduction: Party systems in Latin America. In: Mainwaring S, Scully T R (eds.) *Building Democratic Institutions: Party Systems in Latin America*. Stanford University Press, Stanford, CA

Mair P 1997 *Party System Change: Approaches and Interpretations*. Clarendon Press, Oxford, UK

Sartori G 1976 *Parties and Party Systems: A Framework for Analysis*. Cambridge University Press, Cambridge, UK

Wolinetz S B, Brookfield U T (eds.) 1998 *Party Systems*. Ashgate, Aldershot, UK

P. Mair

Pastoralism in Anthropology

Pastoralism refers to a range of subsistence modes based on specialized herding of domestic animals on free range, and culturally dominated by values associated with such an economic adaptation. Examples are the social formations based on small-stock oriented commercial meat production and long distance trading in the Middle East, on reindeer herding and hunting in the subarctic regions or on the subsistence-oriented cattle and small-stock herding in East Africa. A common synonym to 'pastoralist' is 'nomad.' Since adjustment to animal needs necessitates as well as makes possible a mobile adaptation, the whole pastoral household or parts of the labor force may be permanently or seasonally migrating. The extent of mobility is however in fact highly variable, and dependence on herds a more significant factor than mobility in itself, so that the term pastoralism is preferable. It is useful to consider as pastoralists mainly people for which livestock have a double function as both a means for the production of utilities such as foodstuff, wool, or skins, and as a form of capital with the potential of growth, primarily used for such production (Barth 1964, p. 70). This contrasts them to 'agro-pastoralists' using their herds primarily as an agricultural capital, for example, traditional cultivators in East Africa who invest in cattle to expand the household's agricultural labor pool through bride-wealth and polygyny. Pastoralism is labor-intensive in the sense of requiring a large number of hands for different tasks, and thus also commonly contrasted with, for example, ranching, where fencing makes a much less extensive use of labor possible.

Within the limits of such definitions, 'pastoralism' refers to a wide range of very different situations, regions, and times. Milk-based economies built on cattle/small stock in East Africa and on camels/flocks in the Horn of Africa date 2,000–3,000 years back. Pastoralists exploiting arid regions use species diversification as a risk reducing strategy: different species are subject to different hazards, have supplementary production patterns and require different labor. Contemporary East African pastoralism is based on an emphasis on cow or camel milk production for subsistence purposes, supplemented with small stock rearing to produce milk or meat for market sale or household needs. In the present situation there is often a change towards commercial production of small stock for the meat market, although authorities often encourage the sale of immature cattle to the livestock industry. The basic division of labor in such societies is that young men take care of mobile adult cattle and camels out of lactation, allowing them to profit from valuable but distant pastures. Women, children, and elderly men look after lactating dams and cows, calves, and small stock. These are often managed in a less mobile fashion if conditions allow. In East African cattle cultures, a careful attitude about the reproductive potential of the herd is a prime moral value that reflects on the fundamental worth of the pastoralist as a human being and as a member of the community. Reproductive capital could traditionally

only be disposed of in socially and ritually meaningful contexts such as gift-giving, bride-wealth, or stock friendships. The typical form of social organization among such pastoralists is a confederation of intermarrying clans, led by the property-holding senior men and identifying themselves as having a common ethnicity.

Middle Eastern pastoralism has an emphasis on sheep and goat rearing for the purpose of producing milk and/or meat for domestic consumption or sale. Such pastoralism has existed in the Middle East, around the Mediterranean and in Eastern and Northern Africa for 3,000 to 4,000 years, and is today considered to have developed there as a supplement to horticulture. In the millennium before Christ, camel-saddling technology had developed enough to enable pastoral groups to involve themselves in long-distance caravaneering as well as in extensive camel-back raiding. This brought the breeding of transport and riding animals into the herding operation, and above all involved the desert dwellers in important political and economic relations with sedentary riverine or oasis polities. The spread of Islam reflected the need to find a religious basis of trust when trading out of reach of tight jural or political control. Most shepherding communities of the Middle East are today Muslims, although religious restrictions are often less tight than in urban areas. Small stock rearing provided a subsistence basis for desert or mountain-based family members who did not take part in caravaneering, which was mainly a task for select males. An interesting difference exists between Saudi Arabia and North Africa on one hand and Somalia on the other, in that Somali pastoralists exploit the camel primarily as a milk animal and do not ride it while camel milk is less important where the animal is ridden.

In central Asia other combinations of livestock have enabled other matches of raiding, trade and subsistence dairy production, such as, for example, the traditional salt trade in the Himalayas where yak and sheep are used both for transport and dairy production. Mongolian pastoralism places the strongest emphasis on horses and sheep. Although there are a number of good studies on these communities (e.g., Humphreys and Sneath 1999), they have attracted less contemporary anthropological attention than have East African or Middle Eastern pastoralists. Due to the legacy of the Golden Horde they have instead appealed to the interest of cultural historians.

Reindeer-based pastoralism in the subarctic tundra is the only contemporary form of pastoralism based on specialization in one species. It is considered the most recently developed form of specialized pastoralism and thought to have developed out of hunting only during the last millennium. Herders follow the natural movements of reindeer herds whose primary use is the production of meat. Contemporary Sami herding in Scandinavia is market-oriented, motorized, and generally a large scale operation (Ingold 1980).

Evolutionary anthropologists and philosophers of the eighteenth and nineteenth century concerned themselves with pastoralism because it was assumed to be a transitional stage between hunting and agriculture. Engels, inspired by Morgan, took a particular interest in it, because he saw livestock holdings as the origin both of private property and of male domination. Serious field research among pastoral people was undertaken only in the 1940s and 1950s when monographs on groups such as the Nuer, Dinka, and Somali came to play a decisive role in the development of British structural-functional theorizing about patrilineal descent systems. Within other scholarly traditions, attempts were made to find suitable ways of systematically classifying pastoral economies into subtypes (full nomads, seminomads, etc.) often based on the relative time spent on the move or relative dependence on agriculture. These classificatory systems for a long time dominated what was written on pastoralism but were not theoretically productive.

Substantial progress in the study of pastoralism can be traced to three different but intertwined strands of research that were prominent in the late 1960s and throughout the 1970s. One followed American traditions of cultural ecology and focussed on adaptation. The second was prominent in France and pursued Marxist analyses of modes of production. The third was inspired by Barth's theories of economic action. The conjunction of these interests, followed up by participatory fieldwork, led to a boost in understanding the technical aspects of pastoralism, the complex set of resources required, and the problems of continuity involved in sustaining herd capital over time in arid regions characterized by unpredictable climatic conditions.

As noted, pastoral production involves both a production of utilities and a continuous reproduction of livestock capital. Paine (1971) made a useful distinction between herding and husbandry. The first referred to the care for the wellbeing and growth of animals as such and husbandry is the management of the herd in its totality. The latter takes place through investments, acquisitions, or culling and sales and involves the herd owner in socially highly significant transactions that establish relations of friendship, affinity, economic insurance, spread of risks, and political alliance that are necessary for a socially and economically secure existence as a pastoralist. This implies a long-term scheming which has a very different time perspective from making herding decisions. The latter are made on a daily *ad hoc* basis, taking into consideration the latest information about the nutrition, health, and productive status of the animals as well as the quality and quantity of fodder and water available in accessible pastures, the presence of predators and human enemies, and the immediate availability of labor.

Pastoral economic power has several facets: control over the animals as food producers and as objects of

exchange, and control over the growth potential of one's livestock herd. This makes possible multifaceted property concepts. Animal wealth can be redivided, redistributed, and reorganized in a number of ways, creating diversity in rules and practices of inheritance and other forms of transgenerational property devolution. Senior men often strive to control the labor of younger men by retaining ultimate control of the latter's future herds and by summoning ritual prerogatives such as blessing power among the East African pastoralists or closeness to ancestral spirit owners of the land among the Mongols.

Property rights to grazing and water demonstrate a wide span of variation, depending on the territorial regularities of movement. East African pastoralists exploit land that can be seen as either government property or divine property, as open access or as a commons held by the ethnic group. Transhumant Middle Eastern pastoralists may have very definitive rights to seasonal routes. The Beja clans of Sudan claim tracts of land primarily for defining the political privilege of host-ship. Such differences in land right patterns reflect the degree of predictability in location and extent of seasonal rainfall or river flooding as well as the farming cycles of agricultural neighbors and concomitant fluctuations of markets. Seasonal regularity and predictability structure the possibilities for control of land, grazing, and political power. Pastoralism does not represent any single way of organizing relations of production (cf. Bonte 1981, p. 30). There is a great divergence for example between the quasi-feudal Khirgiz of Afghanistan on the one hand and the loosely organized Turkana of Kenya on the other. The Khirgiz khan distributes stock-loans to thousands of clients while among the Turkana each household strives for its own independence through cattle which are either owned by themselves or borrowed on a more egalitarian basis between stock allies. The issue of the forms of inequality that could be found within pastoral societies and the mechanisms often maintaining a relative degree of egalitarianism has interested many of the researchers referred to. If livestock rearing is considered *per se*, there are many practical problems that are common to pastoralists in different parts of the world, yet the ecological niches they occupy also allow for substantial variation. However, the insufficiency of ecological determinism for understanding pastoral cultures and social conditions has become apparent particularly in relation to such questions of equality and inequality.

Over time, anthropologists looking at pastoralism have come to place more emphasis on the various ways in which pastoral communities were integrated in larger regional economic and political systems, and on the interlinkages between pastoral production and other economic activities. An observer who considers the activities of a pastoral group in a particular place and time is easily misled to disregard the variation that a long-term perspective reveals. Later research has emphasized that pastoralism is not necessarily a static, 'pure' adaptation, but rather a stage in a flexible series of adaptations to irregular climatic circumstances. It has only existed as a pure subsistence form under particular historical conditions. What is more typical for pastoralists than a monodimensional dependence on one type of resource is flexibility in combining different resources. To survive as pastoralists in the long term, livestock-rearers have had to integrate supplementary pursuits such as venture farming, small scale-irrigation, famine-year hunting, or trade. With a longer-term perspective on some of the ethnic groups of East Africa, a gradual switch between dry-land agriculture, pastoralism, and hunting seems to be no rare occurrence. The apparent degree of specialization in and dependence on pastoralism is thus often spurious. Pastoralists whose work is dominated by care for animals still normally depend on getting parts of their diet from farm produce. Small-scale supplementary cultivation can be added to pastoralism in innumerable ways or these products can be secured from others through barter or cash trade, predation, tribute taking, or labor services. The large variation in possible solutions to this problem in itself precludes ecological determinism as a fruitful point of departure for understanding social organization. Processes of stratification and equalization within the pastoral communities can only be understood when these modes of articulation with surrounding structures are taken into account. Today, an essential issue is to what extent a pastoral group in their combination of assets have access to cash or other resources that can act as a security independent of fluctuations in herd health or climate.

In the 1970s and 1980s, the famines among African pastoralists hit by drought coincided with the international breakthrough of environmentalist discourse. Desertification became a catchy symbol of how humanity was overexploiting its brittle resource base. Hardin (1968) presented his theory of 'the tragedy of the commons,' using as an illuminating but fictive illustration a herd-owner's use of pastures. Maximizing his individual gain, the herd-owner was thought not to pay attention to the collective good. In the end all herd-owners would accordingly lose from their own faulty strategy. Hardin's theory was picked up by policy-makers dealing with pastoral development in East Africa. During the 1980s and 1990s a concern for how policies informed by Hardin's theory affect the situation of pastoralists has turned opposition to Hardin's model into a major inspiration for anthropological research. The debate has centered on the degree to which human-induced desertification is taking place. Are pastoralists to be blamed for overgrazing or to be seen essentially as conservationists? The anthropological critique of Hardin and of evolutionary ecologists is often intermixed with an ecological–technical critique, considering the long term effect of pastoral practices on species diversification

and erosion, emphasizing the resilience of desert vegetation, the relative local impact of global climatic change, and inadequacies in models relating grazing pressure to the subsequent availability of good fodder. More anthropological argumentation emphasizes the necessity to distinguish between the knowledge, values and considerations actually expressed among herdspeople and the effects of their herding practices. A methodological issue, dividing social scientists and evolutionary ecologists, is whether 'self-interest' can be analyzed as a causal variable independent of how the actor interprets his own strategies and choices. Pastoral anthropologists question the relevance of the fictive decision-making situations postulated by model-builders like Hardin to the choices real herd-owners have to make. Pastoral economies are based on constant recombination of the available resources of livestock capital and food, and in order to match productive resources and consumption needs as efficiently as possible. This makes the household a debatable unit of analysis. Anthropologists also emphasize the existence of collective institutions and norms which serve a conservationist function, and argue that external observers often think in terms of quantity and tend to neglect quality issues that pastoralists deem entirely important in relation to the composition of a herd and to fodder, minerals, and water. Neither is the parallel and intertwined rationalities of herding and husbandry appreciated. Caring for continuity involves more than a concern for water and fodder: it requires concern for securing access to territory, to docile, fertile, and reproductive stock and to reliable and knowledgeable labor—all critical resources that cannot be taken for granted but have to be continuously ensured.

While the 1970s gave fresh insights into pastoralism at a generalizable level, the field has been relatively stagnant in the 1980s and 1990s, except for the argumentation in defense against pastoralists blamed for desertification. Plentiful documentation has however been secured of threats to the viability of pastoral subsistence. Official worries over the situation of pastoral nomads are concerned with land degradation, but evidence particularly from East Africa points to problems associated with the extent of accessible pastures rather than their quality. Contemporary pastoralists suffer a successive limitation of critical grazing or water assets through the expansion of towns and of small-holder farming and large-scale mechanized agriculture, through irrigation schemes and game reserves and more brutally, through the expansion of zones of war or insecurity.

Technological changes such as motorized surveillance and transport of herds, easy fencing, mechanical watering, or stationary foddering create forms of management more profitable but less efficient in terms of providing work and food for a large number of people. A concentration of stock ownership takes place parallel to a transition to more intensive small stock rearing at the expense of cattle or camel herding, something which has been suggested to have negative ecological consequences in the long run. The diversion of farm products once provided by cultivating neighbors into other more competitive markets is another serious problem to modern pastoralists.

See also: African Studies: History; Desertification; Ecological Economics; Ecology, Cultural; Evolutionary Approaches in Archaeology; Land Degradation; Nomads/Nomadism in History

Bibliography

Barth F 1964 *Nomads of South Persia: The Basseri Tribe of the Khamseh Confederacy*. Allen and Unwin, London

Bonte P 1981 Marxist theory and anthropological analysis: The study of nomadic pastoralist societies. In: Kahn J S, Llobera J R (eds.) *The Anthropology of Pre-capitalist Societies*. MacMillan, London

Dyson-Hudson N 1972 *Introduction. Perspectives on Nomadism*. Brill, Leiden, pp. 22–9

Hardin G 1968 The tragedy of the commons. *Science* **162**

Humphreys C, Sneath D 1999 *The End of Nomadism? Society, State and the Environment in Inner Asia*. Duke University Press, Durham, NC

Ingold T 1980 *Hunters, Pastoralists and Ranchers*. Cambridge University Press, Cambridge, UK

Khazanov A M 1984 *Nomads and the Outside World*. Cambridge University Press, Cambridge, UK

Lefébure C 1979 Introduction: The specificity of nomadic pastoral societies. In: *Pastoral Production and Society*. Équipe ecologie et anthropologie des societes pastorales, Maison des Sciences de l'Homme and Cambridge University Press, Cambridge, UK

Lewis I M 1961 *A Pastoral Democracy: A Study of Pastoralism and Politics among the Northern Somali of the Horn of Africa*. Oxford University Press, London

Paine R 1971 Animals as capital: Comparisons among northern nomadic herders and hunters. *American Anthropologist* **44**: 157–72

Ruttan L M, Borgerhoff Mulder M 1999 Are East African pastoralists truly conservationists? *Current Anthropology* **40**: 621–52

Spencer P 1998 *The Pastoral Continuum: The Marginalization of Tradition in East Africa*. Clarendon Press, Oxford, UK

G. Dahl

Patient Adherence to Health Care Regimens

1. Overview

Research conducted since the 1970s has demonstrated that, on average, about 40 percent (and in some cases as many as 90 percent) of patients fail to adhere to the

recommendations they have received from their physicians for prevention or treatment of acute or chronic conditions. This is a serious problem known in the medical literature both as 'patient nonadherence' and 'patient noncompliance.' The broadest definition of the terms is that a patient fails to manifest behaviorally the health professional's (usually physician's) intended treatment regimen. It consists of such actions as the patient failing to take antibiotics correctly, forgetting or refusing to take hypertension medication, forgoing important health habits, and persisting in a dangerous and unhealthy lifestyle. As a result of failure to adhere, many patients become more and more seriously ill, they develop infections that are resistant to treatment, physicians alter patients' treatments based on misunderstanding of its initial effects, physicians are misled about the correct diagnosis, and the time and money spent on the medical visit is wasted (for reviews see DiMatteo and DiNicola 1982, Meichenbaum and Turk 1987, Myers and Midence 1998, Shumaker et al. 1998).

2. *The Conceptual and Clinical Importance of Adherence*

The issue of patient adherence has both conceptual and clinical implications. In the realm of clinical medicine, the decade of the 1990s brought amazing advances in the treatment of both acute and chronic illnesses, vastly improving quality of life for those suffering from them. Even HIV has promising treatments in the form of highly active antiretroviral therapies. Attempts at managing medical costs have enhanced the focus on preventive services. There are concerns, however, about the re-emergence of infectious diseases, as some bacterial infections that were once straightforwardly treated with antibiotics remain resistant even to the most powerful antibiotics. In HIV, viral resistance to the protease inhibitors has emerged just when control of the disease seemed possible. It is now recognized that noncompliance may be one of the major causes of these reductions in the efficacy of medical treatments, the proliferation of drug resistance, and the waste of billions of health care dollars.

In the conceptual realm, the achievement of patient adherence involves the phenomenon of social influence. A health professional essentially tries to influence a patient to change their behavior. Because medical care is delivered through the interpersonal interaction between health professional and patient, it may be best understood as a social psychological phenomenon. Thus, relevant are issues of communication, trust, persuasion, social normative influence, social reinforcement, and social support.

The understanding of compliance/adherence is complicated further by the philosophical issue of patient self-determination, often examined in detail in writings on medical ethics. It is also affected by issues of information processing, understanding, and retrieval, and the sometime interaction of these issues with that of self-determination. For example, a patient might be noncompliant simply because they never understood the regimen in the first place due to anxiety in the presence of the physician. This patient would follow the regimen had they understood it. Another patient might conveniently 'forget' to follow the treatment in an effort to exert a certain amount of self-determination.Yet another patient might choose 'alternative' medicine therapies, instead of medically recommended ones, because the former may be both easier to understand and provide the patient with a greater sense of control. Finally, by failing to adhere, another patient might be consciously or unconsciously trying to commit suicide, either through clear-headed choice or because of severe depression.

3. *Controversy Over the Term Reflects a Philosophical Issue*

Historically, the behavior of following a physician's recommendations has been termed 'compliance,' and reviewing the literature from indexed databases, particularly before 1990, has required the use of this term. Some writers (see Haug and Lavin 1983) have argued that the term compliance implies coercion, emphasizing paternalistic control on the part of the physician. Many writers have more recently used the term adherence,' which seems to imply that the patient, in a more self-determining manner, 'sticks to' the regimen if they choose to do so. Still others have argued that the appropriate term is 'cooperation,' which patient and physician may or may not achieve, depending upon their willingness and ability to negotiate effectively with one another. Studies do show that patient adherence is highly dependent upon patient involvement in medical choices and patient commitment to treatment decisions, and so cooperation, as well as participation, negotiation, and partnership are essential. The use of all these terms creates, of course, an indexing nightmare, and so the terms 'adherence' and 'compliance' have remained dominant. Although not always referred to as such in indexing, research on aspects of patient self-determination suggests it to be an essential element in adherence and health behavior change.

Whatever it is to be called, the phenomenon of a patient following through with a health professional's treatment recommendations is not a simple achievement. Given the potential for difficulties at several steps—from the physician conveying information about the regimen, to the patient remembering what was conveyed, to the patient wanting to carry out the behavior, to actually overcoming all the difficulties of doing so—one can easily understand why noncompliance rates are very high.

4. Theories Guiding the Research

Various theoretical models have driven, in some fashion or another, the research on patient adherence (see *Health Behavior: Psychosocial Theories*). They include social cognition approaches that emphasize the patient's thoughts and beliefs as primary influences on adherence behavior. Many theories use a value-expectancy paradigm, in which behavior in response to a threat to health arises from expectations that the action will be self-protective. The Health Belief Model is one such approach, and it emphasizes thoughts regarding risks and benefits of the recommended course of action (Janz and Becker 1984). The Theory of Reasoned Action takes account of beliefs and of social influences, and emphasizes the role of intentions and their tenuous connection to action. The Theory of Planned Behavior adds to this a consideration of perceived behavioral control and perceived barriers to action (Ajzen 1985). There are also attribution theories, in which people are posited to be concerned with explaining the causes of events in terms of their locus of control (internal vs. external), stability, and universality (Wallston et al. 1994).

Some models are based on efficacy beliefs regarding both the ability of the self to carry out the target behavior, and of the action to affect the intended outcome (see *Self-efficacy and Health*). There are likely to be limitations to linear rationality in decision making, and recent modifications of the social cognition approach have tried to take these limitations into account. One of these modifications is the Transtheoretical, or Stages of Change, Model that proposes that adherence occurs through progressive, nonlinear stages or steps of behavior modification and maintenance (Prochaska and DiClemente 1983). Finally, self-regulatory models view the patient as an active problem solver who strives to achieve a health goal by identifying the health threat, developing and implementing an action plan or coping procedure to deal with the threat, and appraising the outcome of the action (Leventhal and Cameron 1987) (see *Health: Self-regulation*).

No single model is universally valid, or even helpful, in the study of adherence, particularly because the complexity of treatment regimens varies widely from one disease to another. An approach that organizes the research on adherence in terms of a purely biomedical model tends to classify and analyze adherence studies by disease conditions and finds some interesting trends and patterns that might provide clinically relevant insights for the care of specific disease conditions. The difficulty with this approach, however, has been that similarities in psychosocial phenomena across disease conditions are not apparent, and findings that could be applicable to all patients may remain undiscovered. Focusing solely on psychosocial issues, without regard to the uniqueness of disease conditions, may introduce too much variation into the analysis, making trends difficult to determine. A combined 'biopsychosocial' approach, on the other hand, allows for the analysis of psychosocial factors in the context of the specifics of various diseases and treatment conditions.

5. Limitations in the Literature

Since 1968, there have been over 9,000 papers published about patient adherence (or its alternate name, patient compliance), with a ratio of more than two reviews/opinion pieces to one empirical article (Trostle 1997). Most of the empirical publications have examined only one or two predictors of adherence, either in an observational study or in an intervention to improve adherence/compliance. These studies have demonstrated, not surprisingly, that doing something to help patients comply is better than doing nothing, and that compliance can usually be found to correlate with something about the patient (e.g., motivation), the regimen (e.g., its complexity), or the interaction between patient and health professional (e.g., communication). Studies that have examined several elements of this very complex phenomenon are, on the other hand, rather few and far between. Most chronic illnesses, where adherence is very difficult (e.g., adolescent diabetes), are quite demanding psychologically and behaviorally. Designing and maintaining an effective treatment package requires understanding the entire picture of compliance, from communication of the regimen to its implementation. Unless the phenomenon is fully understood in all its complexity, it is impossible to know when, where, and how it might be best to intervene to help patients to adhere.

6. A Multidimensional Approach to Noncompliance

In examining how difficult and complex adherence to a treatment regimen might be, it is useful to consider a clinical example. Imagine a 48-year old, moderately overweight, borderline hypertensive, male patient who has been told by his physician that he should begin and maintain a program of regular exercise. The goals of this exercise are to reduce his weight, to lower his blood pressure, and improve his level of HDL cholesterol. He understands quite well what he is being asked to do, and he believes that exercise is important and worth the trouble and time it takes. His family and friends agree that the regimen is valuable, and encourage him to follow it. His family members remind and encourage him daily, and have even invested in a treadmill for him to use each evening. In this case, as in general, exercise compliance demands a multidimensional, multifactorial approach because changing the sedentary habits of a lifetime can be very

difficult and often meet with failure. Even among rather seriously at-risk patients, such as who have cardiovascular disease, diabetes, or a previous myocardial infarction, there is a 30–70 percent dropout rate from exercise. In such situations, as with our example patient here, people may have very positive attitudes toward exercise, and even strong networks of social support, but still fail. Maintenance of these very complex behavioral changes over a long period of time requires very careful attention to many factors.

7. *Factors That Affect Patient Adherence*

As noted above, there tends to be something of a disjuncture between the theoretical models of adherence, most of which are multidimensional, and the empirical findings, which are primarily unidimensional. Nevertheless, the empirical studies tell us some important facts about adherence that are useful both clinically and in terms of how they contribute, albeit in a limited fashion, to the theoretical models. Patient demographics, for example, are thought by many clinicians to be the best predictors of patient adherence. Physicians tend to use easily available information from the clinical encounter to steer their attention to those they believe to be at highest risk for nonadherence. Despite physician beliefs, however, patient age and gender seem to have very little relationship to adherence (except, perhaps, that adolescents tend to have more problems adhering than other age groups). Further, although lower income patients and those with more limited education have somewhat lower levels of adherence than more affluent and educated patients, the effect is quite small, and not nearly as important as that of other variables that can be altered.

Psychological variables, such as attributions, locus of control, and beliefs, while important in the models noted above, have not shown consistent effects on adherence. For example, believing that a disease is severe may in some cases prompt a person to take consistent action to avoid or treat that disease, and in other cases bring about denial and reckless abandonment of the regimen. On the other hand, the effect of another psychological variable—depression—on adherence is potentially noteworthy and should be examined further. Depression may bring about hopelessness and isolation, two phenomena that reduce adherence considerably; but it may be entirely treatable (see *Depression, Hopelessness, Optimism, and Health*).

Adherence has been found consistently to be dependent on the patient's social support system and family/marital environment. Further, a regimen's complexity and adverse effects must not be overlooked by physicians when making a medical recommendation; much research has found that nonadherence is related to long-term regimens with complex and confusing dosages and scheduling. Health professional-patient communication seems to affect adherence as well. Because of limitations in communication during the medical visit, patients may misunderstand what they are to do, and the anxiety of the situation tends to interfere with recall at a later time. The sensitivity and empathy in physician–patient communication often matters as well, because patients tend to be more likely to try to adhere to recommendations given by health professionals that they like and trust. Patients' health beliefs and attitudes, often influenced by their cultural patterns and expectations, may conflict with the regimen, making adherence difficult, and practical barriers, such as side effects and lack of necessary resources, may interfere with the regimen (see *Health and Illness: Mental Representations in Different Cultures*). Finally, studies show that physicians typically do not know whether or not their patients are adherent. They overestimate the degree to which their patients follow their directives, and are typically unable to identify which patients are having adherence problems. Yet, the accurate assessment of adherence is essential because achievement of adherence has been shown to make an important difference in patient outcomes and patient health status.

8. *Research Challenges*

A major research challenge in the field of adherence involves its measurement. It seems obvious that one good way to measure adherence would be simply to ask the patient. In fact, most adherence research and clinical practices have relied upon the technique of self-report, although such an approach may be biased by self-presentation and by patients' fears of reprimand by the physician. Reports of spouses or other health professionals can be useful, but these reports may vary in accuracy depending upon the opportunity to observe the patient's daily activities. Techniques such as pill counts, patient behavioral diaries, and electronic recording devices have their own drawbacks, including patient manipulation toward the goal of concealing nonadherence. Tests, such as urine or blood assay, may be useful for medication, but tend to reflect only its recent consumption. Physician reports and chart entries are unreliable and are often based on unclear criteria. Because there is no 'gold standard,' convergence of the findings of research is more difficult to achieve. Necessary to the field is the development and implementation of methods for building therapeutic trust so that patients can be forthcoming and frank about their difficulties in following treatment suggestions, and both measurement and adherence itself can be improved (Hays and DiMatteo 1987).

A second issue of concern is that most reviews of adherence have been qualitative. Many writers, in-

cluding this author, have described extensively the factors that influence patient adherence, ranging from the importance of physicians' communication skills to how simple or complex the regimen should be. Although these reviews have attempted to be exhaustive, their conclusions have been biased by several factors including limited sampling of the population of empirical studies, and entrenched, established social psychological (or other) models of behavior guiding their organization. Further, when studies are listed and the effects of different variables on adherence are described, there are invariably conflicting results that cannot be resolved qualitatively. When there is a huge amount of research available, such as on adherence, it is possible to review only a sampling of it, resulting in further bias unless that sampling is random. One good quantitative assessment of a complete set of empirical studies, however, is worth many expert opinions and reviews. Fully understanding all of the complex elements of adherence involves painstakingly organizing, reviewing, and quantitatively summarizing the entire abundant literature on each variable related to adherence employing the research technique of meta-analysis. Nothing short of this will provide the clear answers that are necessary. Such an approach would allow for a better overall understanding of the complex relationships between adherence and characteristics of patients, their lives, their diseases, their regimens, the therapeutic relationship, and the context of medical care delivery.

Finally, conceptual and empirical work is necessary to remedy the common confusion of adherence behaviors with adherence outcomes. A patient might carry out flawlessly every behavioral requirement of care, but the outcome may remain disappointing. Research must pay scrupulous attention to the accurate assessment of behavior, and recognize that physiological measures of outcome should never serve as proxies for adherence. The independent effect of adherence on achieving health outcomes, health status, functional status, psychological and social functioning, and all aspects of health related quality of life needs to be examined (see *Quality of Life: Assessment in Health Settings*).

9. *Conclusion*

As noted above, despite the complexity and multidimensionality of theoretical models of adherence, nearly all of the research has examined individual factors that correlate with adherence. These empirical investigations have provided some useful answers about adherence, but the theory that could be helpful in guiding our research questions remains disconnected from it. Fully understanding adherence has remained an elusive goal because reviews of the literature have been equivocal. Clear answers are necessary, however, as some of these questions have strong clinical and policy implications.

This review of the literature has attempted to view it as a whole, to the extent that a qualitative approach can do so, and to point out its strengths, limitations, and promising trends. At this point in time, there are many reviews, and not enough well designed, well executed, multidimensional, longitudinal empirical studies of adherence. Now that we have a compilation of studies in the univariate realm, it is necessary to work toward multifactorial explanatory approaches to adherence. These models need to examine simultaneously the effect of social, psychological, and biological variables on adherence, and to examine their unique interactional effects. Only then will a full and clear picture of this complicated phenomenon emerge.

See also: Attitudes and Behavior; Explanatory Style and Health; Health Behaviors; Health Behaviors, Assessment of

Bibliography

Ajzen I 1985 From intentions to actions: a theory of planned behavior. In: Kuhl J, Beckmann J (eds.) *Action-Control: From Cognition to Behavior*. Springer-Verlag, Berlin

DiMatteo M R, DiNicola D D 1982 *Achieving Patient Compliance: The Psychology of the Medical Practitioner's Role*. Pergamon Press, New York

Haug M R, Lavin B 1983 *Consumerism in Medicine: Challenging Physician Authority*. Sage, Beverley Hills, CA

Hays R D, DiMatteo M R 1987 Key issues and suggestions for patient compliance assessment: sources of information, focus of measures, and nature of response options. *Journal of Compliance in Health Care* **2**: 37–53

Janz N K, Becker M H 1984 The Health Belief Model: a decade later. *Health Education Quarterly* **11**: 1–47

Leventhal H, Cameron L 1987 Behavioral theories and the problem of compliance. *Patient Education and Counseling* **10**: 117–38

Meichenbaum D, Turk D C 1987 *Facilitating Treatment Adherence: A Practitioner's Guidebook*. Plenum Press, New York

Myers L B, Midence K (eds.) 1998 *Adherence to Treatment in Medical Conditions*. Harwood Academic, Amsterdam

Prochaska J O, DiClemente C C 1983 Stages and processes of self-change of smoking: toward an integrative model of change. *Journal of Consulting and Clinical Psychology* **51**: 390–95

Shumaker S A, Schron E B, Ockene J K, McBee W L (eds.) 1998 *The Handbook of Health Behavior Change*, 2nd edn. Springer, New York

Trostle J A 1997 Patient compliance as an ideology. In: Gochman D S (ed.) *Handbook of Health Behavior Research*. Plenum, New York, pp. 109–22

Wallston K A, Stein M J, Smith C A 1994 Form C of the MHLC Scales: a condition-specific measure of locus of control. *Journal of Personality Assessment* **63**: 534–53

M. R. DiMatteo

Copyright © 2001 Elsevier Science Ltd. 11115
All rights reserved.

 ISBN: 0-08-043076-7

Patriotism

'Patriotism' is about political allegiance (and, of course, loyalty), commitment, and dedication. In briefest compass, it means love of one's country or nation and is one of the oldest political virtues. It is rather more emotional than rational in its appeal and demands recognition of what is presumptively a pre-existing duty to that political order or state. One of the best exemplars of patriotism is Stephen Decatur's well-known toast in 1816:

> 'Our Country! In her intercourse with foreign nations may she always be in the right, but our country, right or wrong.'

To which John Quincy Adams replied:

> 'My toast would be, may our country always be successful, but whether successful or otherwise, always right.'

Patriotism depends upon an often unarticulated principle that is the political counterpart to the 'blood is thicker than water' adage, which reminds people that they should prefer their families before all others. Patriotism is a natural consequence of political membership or citizenship, and it is not clear where—other than membership itself—the duties of patriotism originate or what justifies them other than that membership, for patriotism is not voluntarily assumed as general political obligations are presumed to be. To deny or renounce patriotism or to act contrary to what its proponents deem appropriate is to be disloyal.

There is a Burkean quality to patriotism, both because it projects an almost organic, trans-historical unity among the member of a state, nation, or people such that each individual is inseparable from the past and because it looks to that past and to the achievements of one's political ancestors rather than to one's own accomplishments as sources of pride and holds up those achievements as standards by which the successes of the present are to be measured. In this respect, patriotism can be nurturing as well as oppressive, for it defines and constitutes the political member and, in the process, restricts that member's range of permissible options.

Patriotism is rarely invoked but is in times of stress or trouble. The call for patriots or acts of patriotism is issued when a sacrifice for the presumed good of the people or their state is needed, usually because that good is perceived to be in jeopardy or under attack. In the period since World War II, those attacks have most often been alleged to come from inside a system, from people who in an earlier day would have been labeled 'traitors' but are now more conventionally called 'disloyal' and even 'outsiders.' During World War II, in Europe, according to the *OED*, a patriot was a 'loyal inhabitant of a country overrun by the enemy, especially a member of a resistance movement.' But for nearly 150 years prior to that, also according to the *OED*, because the mantle of the patriot had been assumed by persons who were deemed not entitled to it, the term itself was somewhat discredited. Dr Johnson, amplifying his dictum that 'patriotism is the last refuge of a scoundrel,' noted that the word was used 'ironically for a factious disturber of the government.'

Etymologically, 'patriotism' springs from the Greek, via Latin, for father and father land, (*πατριώτης*, from *πάτριος* of one's fathers, *πατρίς*, one's fatherland; late Latin *patriota*, fellow-countryman [*OED*]), suggesting political membership based on kinship as well as an implicit reliance on family-like bonds to hold a state together. And in the Latinate tongues of Europe, various cognates of *patria* mean 'fatherland' or native country. Native too, as well as nation to which it is conceptually and etymologically related, suggests kinship, but these resonances have long since been buried by linguistic evolution. Their value today is primarily as reminders of the emergence of the political order from tribal, familial associations. 'Fatherland' is more obviously familial, and its roots are Germanic rather than Latinate. (Interestingly, German has adopted the Latinate forms *der Patriot* and *der Patriotismus* even though it has its own words, *der Vaterlandsfreund*, literally, *friend of the fatherland*, and *die Vaterlandliebe*, *love of the fatherland*.) But patriotism has never had this familial resonance in English, which explains why the term has an air of contentiousness and alarm about it when invoked in English and why its English-speaking history has been checkered. In that largely individualist and voluntarist world, political duty, at least since the seventeenth century, has been viewed as a consequence of intentional commitment and a subject for reason and judgment. In those terms, patriotism has often seemed like an alien concept.

Patriotism is more at home in the conceptual world of republicanism. The republican tradition has always looked upon the state as a closely-knit and relatively homogenous association. Patriotism calls upon the members of this association, when appropriate, to put aside whatever divides them and to rally in support of what they share, a practice that is integral to the communitarian predilections of republican society but stands in need of justification from the perspective of individualism. The call for patriotism—for people to be patriotic, for those who are already patriotic to come forth in the spirit of the *patria*—is an appeal to the emotions, no less than love in its ordinary sense is an invocation of affect rather than reason. And like the affections that bind friends and family members, patriotism works by reminding those at whom it is directed of their ties and of their non-voluntary relationships to other people. Thomas Paine's 'sunshine patriot' is a play on the better known 'fair weather friend,' both demeaning of the self-styled supporter who is unwilling to be inconvenienced.

Patriotism is rooted in emotions rather than reason and has to do with feelings of commitment and loyalty to one's nation or state and pride in its history and accomplishments. While it is not necessarily irrational or even unthinking, patriotism does pose as the supreme call on one's commitments, the one that trumps or overrides others with which it may compete.

With the revival of classical republican thinking in English-language political discourse, toward the latter part of the twentieth century, patriotism made an almost grudging return. Republicanism values 'community' and 'civic virtue' with which it seeks to supplant individualism, self-willed obligation, and institutional legitimacy. Further sources of the renewed appeal to patriotism are the collapse of Eastern European communism and the ensuing struggles for national identity and 'liberation.' These linked patriotism to ethnic and religious nationalism (which are akin to those flourishing in Northern Ireland, the Middle East, and on the Indian Subcontinent). Even earlier, the worker and student protest movements of the 1960s and 1970s—especially in France, Germany, the UK, and the USA—and the American civil rights movement were all surrounded by claims of 'disloyalty,' and illegitimacy, which resulted in quests for the proper bearers of the mantel of patriotism. Finally, in the USA, the period of the Cold War was marked by zealous attacks on disloyal 'communist sympathizers,' often in the name of patriotism.

In the West, contemporary proponents of patriotism attempt to harmonize it with the Enlightenment value of 'cosmopolitanism' and struggle to separate it from the virulent nationalism with which it is too easily associated and which it often resembles. These efforts seem destined to fail, for patriotism is particularistic, not universal. It tells people that what they have in common is what deeply and importantly unites them and makes them a nation or a 'people.' This unity overrides their differences. There is, in consequence, a tendency to homogenize those differences into a political blandness that could render society uninteresting and potentially stagnant. Far worse than that, however, patriotism in this homogenizing form prepares the way for the insistence that a greater, underlying good assigns places in the social order. Because of the fundamental sameness of all members, there are no remediations for deprivations. Those who are not sufficiently 'the same' are outsiders who can and should be excluded.

The modern state at the beginning of the twenty-first century is increasingly complex and heterogenous in ways that traditional patriotism cannot comprehend. Moreover, individual states cannot reject that heterogeneity and continue to exist in a world of international political and economic exchange. States today inevitably function in a world that is at odds with their claims of internal uniformity. That internal coherence—where it is more of an ideal than a fact—often leads to oppression. There are few states that do not have internal 'minority' peoples who dissent or are excluded from the presumed consensus that undergirds patriotism. While this is not to say that internal political loyalty and cultural and social diversity are incompatible, the responses of the advocates of patriotism to the circumstances that call forth their pleas are antagonistic to cultural variety. One of the hallmarks of modern politics is tolerance and forbearance by states and their members. But tolerance, by its nature, undermines both the spirit and the practice of patriotism. It is far easier—and in many respects more desirable—to give loyalty to those who are regarded as like one's self, which is among the principal reasons that the habits of obedience that patriotism fosters and on which it depends are usually rooted in the family. To bestow that same deference on an 'alien' authority or on one that appears to uphold a different set of values from one's own is often difficult and can require acts of will and judgment that are antithetical to the non-rational, emotional bases of patriotism.

At the same time, however, so long as there are territorial nations, there will be reasons for inculcating loyalty to them; nations necessarily require their members to make sacrifices from time to time, and it is certainly preferable that these sacrifices—these fulfillments of civic duties and responsibilities—be made willingly if not voluntarily and in the belief that they are justified. Individual members must have grounds for accepting the propriety of actions undertaken by and/or in the names of their states; they must feel some dedication and loyalty to their states. The inculcation of those feelings through the process generally known as 'civic education' must be done in ways that preserve the forbearance that toleration requires and do not give rise to the exclusionary and destructively ugly nationalism that often has patriotism as its partner.

See also: Identity Movements; Identity: Social; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Republicanism: Philosophical Aspects; State, Sociology of the; Xenophobia; Civil Society; Civic Culture

Bibliography

Canovan M 2000 Patriotism is not enough. *British Journal of Political Science* **30**: 413–32

Dietz M G 1989 Patriotism. In: Ball T, Farr J, Hansen R L (eds.) *Political Innovation and Conceptual Change*. Cambridge University Press, Cambridge, UK, pp. 177–93

Nussbaum M C 1996 Patriotism and cosmopolitanism. In: Cohen J (ed.) *For Love of Country: Debating the Limits of Patriotism*. Beacon Books, Boston, MA, pp. 1–16

Rorty R 1999 The unpatriotic academic. In: Rorty R (ed.) *Philosophy and Social Hope*. Reprinted from the *New York Times*, 13 February 1994. Penguin Books, New York, pp. 252–54
Schaar J H 1973 The case for patriotism. *American Review* **17** (May): 59–99
Wahlke J C (ed.) 1952 *Loyalty in a Democratic State*. D C Heath, Boston, MA
Viroli M 1995 *For Love of Country: An Essay on Patriotism a. Nationalism*. Oxford University Press, Oxford, UK

G. Schochet

Patron–Client Relationships, Anthropology of

Patron–client relations (PCR) involve asymmetric but mutually beneficial, open-ended transactions based on the differential control by individuals or groups over the access and flow of resources in stratified societies. Patrons provide selective access to resources they control, or place themselves or the clients in positions from which they can divert resources in their favor. In exchange, the clients are supposed to provide material and human resources, including their gratitude and loyalty, while boosting the patron's prestige and reputation. The study of PCR has revealed the persisting presence of interpersonal hierarchy in contemporary societies, contributing to the reevaluation of paradigms in social science. Debates continue about their institutional viability and significance in late modernity.

1. *Terminology*

The terms used originated in ancient Republican Rome, where relationships of *patrocinium* and *clientelae* proliferated during the Republic and into the Empire (later on, they were subsumed under what historians of late medieval times called *bastard feudalism*). These terms found their way into the vernaculars of Mediterranean and Latin American societies. Other terms are used for PCR elsewhere.

2. *Background*

Whereas in antiquity, PCR formed part of the normative framework of society and could hardly be thought as conflicting with legal institutions and the social ethos, in modern societies PCR are built around such a conflict. While in principle and by law the clients have access to power centers and the ability to convert resources autonomously, this potential is neutralized by the patrons' control over scarce resources, be they land, water sources, employment, loans, schooling opportunities, medical services, public security, or infrastructure. PCR may be rooted in the search for protection, in bureaucrats' providing preferential access to petitioners in return for future favors, in poor people sharing the social visibility of men or rank, while the latter use these links to consolidate their status. Union leaders, political activists, and government officials may use PCR to build a network of supporters, instrumental for gaining higher-level positions, office, and control of resources.

3. *The Study of PCR*

Expanding in the 1960s, it led (by the late 1970s and 1980s) to general and comparative analyses. In the 1990s in-depth studies renewed interest in PCR in polities under transition to market economy and democracy, re-evaluating some of the premises of earlier approaches, especially regarding the connection between PCR and underdevelopment. The growth of interest has been connected with the spread of research from anthropology to the other social sciences. From early dyadic emphases, research has increasingly revealed a complex range of PCR analyzed in terms of networks, factions and coalitions.

The study of PCR contributed to analytical shifts, as it deals with phenomena for which neither the group corporate model nor formal institutional models provide a satisfactory guide. In parallel, research revealed the complexities of studying PCR, owing to their informal character and to their cross-institutional insertion that requires interdisciplinary knowledge, bridging anthropology and sociology, political science, and economics.

4. *A Logic of Social Exchange*

Research has identified a specific logic of exchange and reciprocity beneath the wide gamut of PCR. Both the control of material and human resources and the character of trust play a crucial role in it. This logic has been defined in the literature as clientelism. A related term is patronage, used interchangeably at times and alluding to the support granted by the patron to his followers and *protégés*.

This logic implies:

Control of markets. Individuals and institutional actors are able to gather followings by dispensing selective access to valuable benefits and resources. They do so whether acting as patrons (in a strict sense), or through their influence as brokers with those who control the goods and services, or through a combination of both roles, in what has been called patron-brokerage.

Inequality and asymmetry structured though an idiom of reciprocity.

Particularism. A relationship shaped according to the particular traits of the partners rather than on the basis of entitlements or formal roles.

Favoritism. Many people are excluded from PCR or are related indirectly or intermittently. Where clientelism is highly valued, it generates 'inflationary' expectations, many of which are disappointed.

The simultaneous exchange of instrumental (e.g., economic and political) and 'sociational' or expressive resources and services (e.g., promises of loyalty). A package deal allowing for more than a restricted balanced or market exchange is built-in, identified as connected to generalized exchange. This determines expectations of entering more embracing attachments, often couched in terms of commitment, even though some PCR are very ambivalent in this respect.

PCR undermine group (i.e., class and strata) solidarity among both clients and patrons, especially among clients, and to exacerbate resentment and social antagonisms.

PCR are neither fully contractual nor legal. Most PCR are vulnerable to systemic pressures and are characterized by instability, perpetual contest and resource manipulation.

5. *Exchange Strategies*

The agent performing as client is not only expected to provide his patron with specific resources but must also accept the patron's control over access to markets and public goods, and over his ability to convert resources fully. In parallel, the patron's position is not as solid as it may seem, nor guaranteed. Never fully legitimized, it is vulnerable to attack by social forces committed to universalistic principles, by the competition of other patrons and brokers, potential and actual, and by social forces excluded from PCR. Owing to these constant threats, patrons are compelled to rely on their followers to solidify their position. The patron must also relinquish some short-term gains to project public claims and to bolster images of power and reputation. Sometimes this earns him/her the right to determine the basic rules of the social relationships. In return the client is protected from social or material insecurity, and is provided with goods, services and social advancement. Certain types of trust (focalized rather than generalized) are associated with such dynamics.

Complementary exchange strategies are built, which signal what Vincent Lemieux defined as a 'double transformation.' That is, an abdication of autonomy on the client's part and a relaxation of hierarchical controls on the patron's part, through which the client's lack of power becomes dominated power and the latter's lack of domination becomes dominating authority. These exchange strategies are not only affected by immediate, mostly technical considerations of power and instrumentality, but often encompass mutual, relatively long-term, compromises based on commitments as the prerequisite for ongoing social relationships.

6. *The Context of PCR*

Unlike societies in which hereditary ascriptive principles predominate, full-fledged clientelism flourishes where markets are no longer controlled through primordial units and allow for an open flow of resources and opportunities for mobility. (This is what distinguishes clientelism from *feudalism*, as pointed out already by Max Weber in *Economy and Society* (1968, Vol. III, Chaps. 12–13) and later on by historians of Europe). This trend, however, goes hand in hand with a strong tendency toward unequal access to markets and sociopolitical spheres. As they affect distribution and redistribution, PCR remain subject to the dynamics of political economy. Marketing economies, accelerated urbanization and the expansion of the regulatory, extractive, or even sporadic mobilizing activities of central administrations affect PCR. Research has shown that the impact of world economic trends, fluctuations in the price of commodities, the complexity of international trade, banking, and aid—all these affect the pool of patronage resources available to states and other agencies, and influence patterns of control, distribution, and redistribution. Moreover, as these arrangements are not fully legitimized, they remain vulnerable to the challenge of countervailing social forces.

7. *The Comparative Analysis of PCR*

Until the early 1980s, it was dominated by the contrast between traditional dyadic patronage and modern party-directed clientelism, a distinction that derived from the dominant paradigm of modernization. With a clear-cut developmental emphasis, this approach focused on differences in organizational complexity, deriving other aspects from it in a functionalist manner.

Since the 1980s, comparative analysis has paid attention also to other aspects. First, the way agents configure their role. Second, the styles shaping the relationship, following what Robert Paine called a 'generative transactional model of patronage.' Third, symbolic aspects such as images, discourse and trust, so important for the development of PCR. Fourth, the distinction between PCR as institutional *addenda* and clientelism as a major institutional strategy in certain societies and periods. Finally, PCR's institutional viability in processes of democratization and liberalization.

8. *Debates on Institutional Viability*

Researchers differ in assessing the institutional viability and significance of PCR in late modernity.

From one perspective, PCR neutralize the system of representation and entitlements, by placing 'friends' in strategic positions of power and control. Clienteles are depicted as inimical to the institutionalization of public accountability and contrary to a politics, open to generalization and participation, and to a discourse of rights. Other authors emphasize the pragmatics of social action, stressing that PCR are an important mechanism for obtaining transactional benefits, in resource allocation, and in providing local–regional–national mechanisms of articulation. While PCR run counter to universalistic standards, it is claimed they are sensitive to local sentiment, may solve existential problems, provide access for migrant populations, and serve political entrepreneurs. From this perspective, as long as clienteles maintain some balance of reciprocity, the participation in political and economic markets by the 'capi-clientele' (to borrow Mosca's expression)—be he or she a broker, a patron or a patron-broker—constitutes a means for individuals in their entourage to influence public decisions. In this sense, some literature alludes to PCR as reconciling public and private authority and formal and informal rules of the game.

PCR are criticized, and opposed by social forces and coalitions wishing to curtail its presence alongside bureaucratic universalism and market rationality. In parallel, sectors benefiting from patronage may see it as a pragmatic avenue of controlled freedom, useful for advancing in competitive social, economic, and political domains. This duality reflects a major tension of modern democracies, which are built on citizenship and political equality but leave the economic domain open to inequalities.

This explains the paradoxical development of clientelistic networks under macro-economic adjustment and restructuring. Liberalization, reduction of state intervention in favor of market mechanisms, privatization of services, and curtailment of union power further fragments society and heightens the need for support networks. When available, clientelism remained important throughout institutional revamping in Poland, Russia, Hungary, Brazil, Argentina, and Turkey.

In centralized polities, PCR have been depicted as a 'lubricant' to bypass the inefficient and rigid official system, constituting highly trusted informal problem-solving networks that provide material gains and support in situations of low institutional trust. In representative democracies, clientelism can be effective in power competitions, encouraging and rewarding party activists, once power is achieved, for effectively implementing policies. Patronage may become a restricted but legitimate procedure related to high office incumbents' right to appoint followers to positions of responsibility. In other respects, clientelism remains controversial and open to allegations of corruption, due to its informal particularistic use of public resources. Often, it is ignored as long as possible or is disguised as friendship, which is more acceptable in terms of the proclaimed ethos of modern equality.

Studies suggest that in post-industrial societies, patronage develops among professional and upper strata, rather than being restricted to the lower classes. Also, PCR are not confined to politics in the narrow sense, but proliferate as well in the arts, the academia, the church, the media, and even in business—whenever we deal with the power of appointment and the granting of access to benefits, goods, services, influence, and honors.

Changes in the perception of PCR are of systemic consequence if they result in the institutionalization of mechanisms through which citizens can press for entitlements without personal mediation. Such changes include: civil service reforms, non-partisan public systems, recognized charters of rights, controls over party fund raising, and non-partisan comptrollers as a prestigious and trustworthy branch of government. The functioning of these institutional mechanisms hinge on public support for a configuration of the public sphere structured around public accountability and formal responsiveness to turn more and more discrete issues into publicly negotiable and politically consequential.

See also: Exchange in Anthropology; Exchange: Social; Feudalism; Social Stratification

Bibliography

Blok A 1974 *The Mafia of a Sicilian Village*. Blackwell, Oxford, UK

Caciagli M 1996 *Clientelismo, corrupción y criminalidad organizada*. Centro de Estudios Constitucionales, Madrid

Clapham C (ed.) 1982 *Private Patronage and Public Power*. Frances Pinter, London

Dinello N 1999 *Clans for Markets or Clans for Plan: Social Networks in Hungary and Russia*. Woodrow Wilson International Center, Washington, DC

Eisenstadt S N, Roniger L 1984 *Patrons, Clients, and Friends*. Cambridge University Press, Cambridge, UK

Gellner E, Waterbury J (eds.) 1977 *Patrons and Clients in Mediterranean Societies*. Duckworth, London

Graziano L 1984 *Clientelismo e sistema politico. Il caso dell'Italia*. Angeli, Milan

Roniger L, Güne-Ayata A (eds.) 1994 *Democracy, Clientelism and Civil Society*. Lynne Rienner, Boulder, CO

Schmidt S W, Guasti L, Landé C H, Scott J C (eds.) 1977 *Friends, Followers and Factions*. University of California Press, Berkeley, CA

Willerton J P 1992 *Patronage and Politics in the USSR*. Cambridge University Press, Cambridge, UK

Weber M 1968 *Economy and Society*. Bedminster Press, New York

Wolf E 1966 Kinship, friendship, and patron–client relationships in complex societies. In: Banton M (ed.) *The Social Anthropology of Complex Societies*. Tavistock, London, pp. 1–22

L. Roniger

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Pattern Matching: Methodology

To the degree that products of the social and behavioral sciences merit the term 'knowledge,' it is because they are grounded in representations of the social world achieved by matching observations of that world with abstract concepts of it. Pattern matching is essential for attaining knowledge of everyday cultural objects such as food, clothing, and other persons. It is also indispensable for achieving knowledge of inferred entities in science—the size and brightness of distant stars, the existence and magnitude of latent psychological variables such as alienation or intelligence, and the structure of complex social processes such as social mobility, policy implementation, and economic growth. These and other inferred entities are known indirectly and vicariously, through a process of pattern matching (Campbell 1966).

1. *The Concept of Pattern Matching*

Pattern matching is a ubiquitous feature of knowledge processes in everyday life and in science. The use of pattern matching in everyday knowing, however, is different from pattern matching in mathematics, philosophy, and the social and behavioral sciences.

1.1 *Rules of Correspondence*

Rules of correspondence are prescriptions that enable the mapping of one set of objects on another. Objects in a domain, D, are related to objects in a range, R, according to a rule of correspondence such as: Given a set of n countries in domain D, if the object in D is a one-party system, assign the number '0' from range R. If it is a multiparty system, assign a '1.' Rules of correspondence perform the same matching function as truth tables in symbolic logic, pattern recognition algorithms in computer science, and the use of *modus operandi* (M.O.) methods in criminology. When a single rule of correspondence is taken to define fully and unequivocally the properties of an event or object—for example, when responses to scale items on an intelligence test are taken to define 'intelligence'—pattern matching becomes a form of definitional operationism (Campbell 1969, 1988, pp. 31–2).

1.2 *The Correspondence Theory of Truth*

Until the late 1950s, philosophy of science was dominated by the correspondence theory of truth. The correspondence theory, the core epistemological doctrine logical positivism (see Ayer 1936), asserts that propositions are true if and only if they correspond with facts. The correspondence theory also requires that factually true propositions are logically validated against formal rules of logic such as *modus ponens* ($p \supset q, p, \therefore q$) and *modus tollens* ($p \supset q, \sim q, \therefore \sim p$). To be verified, however, propositions must match facts (reality, nature). The correspondence version of pattern matching assumes a strict separation between two kinds of propositions—analytic and synthetic, logical and empirical, theoretical and observational—a separation that was abandoned after Quine (1951) and others showed that the two kinds of propositions are interdependent. Because observations are theory dependent, there is no theory-neutral observational language. Theories do not and cannot simply correspond to the 'facts.'

1.3 *Coherence Theories of Truth*

The correspondence theory has been replaced by a more complex form of pattern matching, the coherence theory of truth, which has a number of versions (see Alcoff 1996). In one version, often called the consensus theory of truth, beliefs are matched against other beliefs, with no requirement that they are tested empirically. Another version, realist coherentism (Putnam 1981), requires that two or more empirically tested beliefs are matched. William Whewell's *consilience theory of induction* is closely related to this (qualified) realist version of the coherence theory. A third version of coherence theory is methodological pragmatism (Rescher 1980). Here, beliefs must satisfy cognitive requirements including completeness, consonance, consistency, and functional efficacy, all designed to achieve optimally plausible knowledge claims. Other versions of coherence theory require the additional condition that the social circumstances under which empirically tested beliefs arise be taken into account. These other versions include 'social epistemology' (Fuller 1991) and the 'sociology of scientific validity' (Campbell 1994).

1.4 *Statistical Estimation and Curve Fitting*

Statistical principles, rules, or criteria are applied to achieve an optimal match between a curve and a set of data points or observations. An example is the *least-squares criterion* (the squared distance between observed and predicted values is a minimum or least value), where the match between the pattern supplied by a (linear or nonlinear) curve and a pattern of observations is approximate and probable, not certain as in pattern matching by rules of correspondence. The degree to which a curve and a set of observations match is summarized by coefficients of different kinds. Some of these represent the goodness-of-fit between

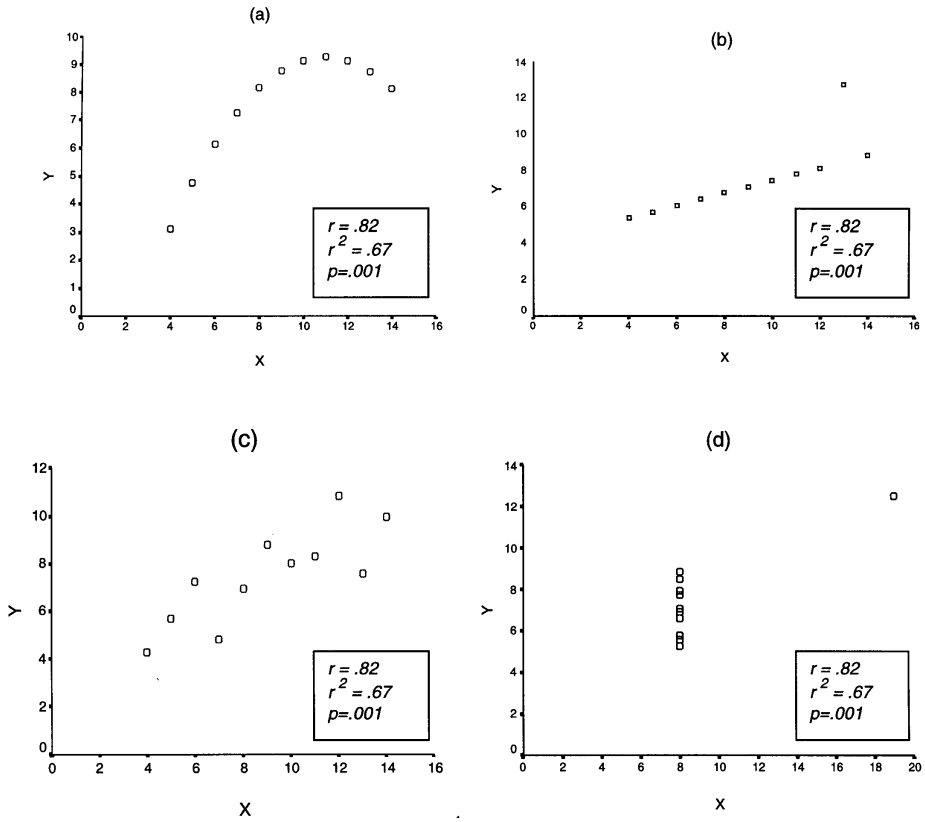

Figure 1
Statistical pattern matching: identical coefficients match different patterns (sources: Tufte 1983, pp. 13–14, Anscombe 1973)

curve and observations, while others represent the magnitude of error in pattern matching. Although coefficients are assessed according to a probability distribution, and given a *p*-value, the same observations can fit different curves; and the same curve can fit different observations. In such cases, pattern matching is as much a matter of plausible belief as statistical probability.

1.5 Context Mapping

Closely related to curve fitting is context mapping. Knowledge of contexts is essential for understanding elements within them. Context mapping, which is epitomized by the figure-ground relation in Gestalt psychology, also applies to the analysis and interpretation of statistical data. For example, observations in a scatterplot cannot be distinguished—in fact, they all look the same—when they are compared one-by-one, rather than compared as elements of the pattern of which they are elements. Context mapping is important in statistical analysis. Coefficients that summarize relations among variables are usually misleading or uninterpretable outside the context provided by scatter plots and other *visual methods in statistics* (see Tufte 1983, 1997). Figure 1 illustrates how identical sets of correlation (r), goodness-of-fit (r^2) and probability (p) statistics match different patterns. Statistics are likely to be misinterpreted outside the context provided by scatterplots, each of which has its own visual signature.

2. Theory of Pattern Matching

The first systematic effort to develop a theory and methodology of pattern matching was that of psychologist and philosopher of social science, Egon Brunswik, a professor at the University of California, Berkeley between 1935 and his death in 1955. In this period, he and his close collaborator, Edward Tolman, developed different aspects of the theory of *proba-*

bilistic functionalism. To communicate this complex theory, Brunswik developed, initially for illustrative and pedagogical purposes, a *lens model* that was based on the optics metaphor of a double convex lens. The indispensable procedural complement of probabilistic functionalism and the lens model was his methodology of *representative design*.

2.1 Probabilistic Functionalism

Probabilistic functionalism is a substantive theory and methodology (philosophy of method) that focuses on relations of adaptation and accommodation between the organism and its environment. Probabilistic functionalism investigates the ways that interdependent, intersubstitutable, and uncertain cues about external environments are used to make judgments about these environments. Environments are not fully determined and uniform in their structures and processes of causality; rather they are uncertain, unstable, and *causally textured* (Tolman and Brunswik 1935). Probabilistic functionalism also studies the manner in which knowers learn about their environments by using informational cues or indicators, continuously creating new cues and revising or abandoning old ones. An essential aspect of these cues is that they are interdependent, multiply correlated, and intersubstitutable, features that require a process of knowing that is approximate, indirect, and vicarious. Darwinian in its focus on the adaptation of the organism to its environment, probabilistic functionalism seeks to understand how these two complex systems, the organism and the environment, come to terms with one another through a process of pattern matching.

2.2 The Brunswik Lens Model

Brunswik (1952) developed his 'lens model' to illustrate the probabilistic interrelations between the organism and the environment. The lens model represents processes of perception and judgment in terms of the optics metaphor of light entering (converging on) and exiting (diverging from) a double-convex lens. The lens model, while initially metaphorical in character, was later specified in what has come to be known as the 'lens model equation' (Slovic and Lichenstein 1971, pp. 656–67):

$$r_a = R_e R_s G + C[(1 - R_e^2)(1 - R_s^2)]$$

The lens model supplies an achievement index, A, which measures the extent to which pattern matching has been achieved (see figure below). The achievement index ($A = r_{Y_e Y_s}$) is the correlation between the statistical properties of an uncertain environment (Y_e) and the statistical properties of a judge's (or knower's) response system (Y_s). A matching index, G, is the correlation between the predicted knower's response (Y_s') and the predicted criterion value for the environment (Y_e'). The lens model also estimates the degree to which indicators or cues are used in a linear or nonlinear fashion. In the equation above, the coefficient, C, is the correlation of the nonlinear residual variance in the multiple correlation coefficients R_e and R_s.

2.3 Representative Design

The methodology of *representative design* requires that experiments be conducted in realistic settings that are representative of an organism's typical ecology. Representative design is a radical departure from the classical experiment, in which one independent (treatment) variable is manipulated so as to assess its effect on a dependent (criterion) variable, with all other factors held constant through random subject sampling and statistical controls. Instead, experiments were to be conducted in natural settings, in environments that are unstable, dynamic, uncertain, and probabilistic. Such experiments require *situation sampling*, not merely the random sampling of subjects, because only the former optimizes the representative character of experiments. Representative design, in addition to its call to abandon the classical experiment, also required a rejection of the uniformity of nature presumption underlying Mill's Canons. Both ignored, and therefore failed to account for, the effects of the many interrelated contingencies that are causally relevant to experimental outcomes. Significantly, although Brunswik was an active participant in the Unity of Science Movement, and thus to some degree seems to have identified with logical positivism, he rejected the uncritical imitation of physics promoted by Vienna Circle positivists and the bulk of psychologists of his time. Psychology, and by extension other social and behavioral sciences, was practicing 'emulative physicalism' in its attempt 'to copy not only the basic methodological principles but also the specific thematic content of physics, thus nipping in the bud the establishment of (appropriate) methodological directives' (Brunswik 1952, p. 36; quoted in Hammond 1966, p. 55).

3. Methodology Development

Probabilistic functionalism and representative design have influenced the contributions of several highly influential scholars who studied with Brunswik and Tolman at Berkeley. Some of these contributions are principally theoretical, for example, social judgment theory (Hammond 1980) and evolutionary epistem-

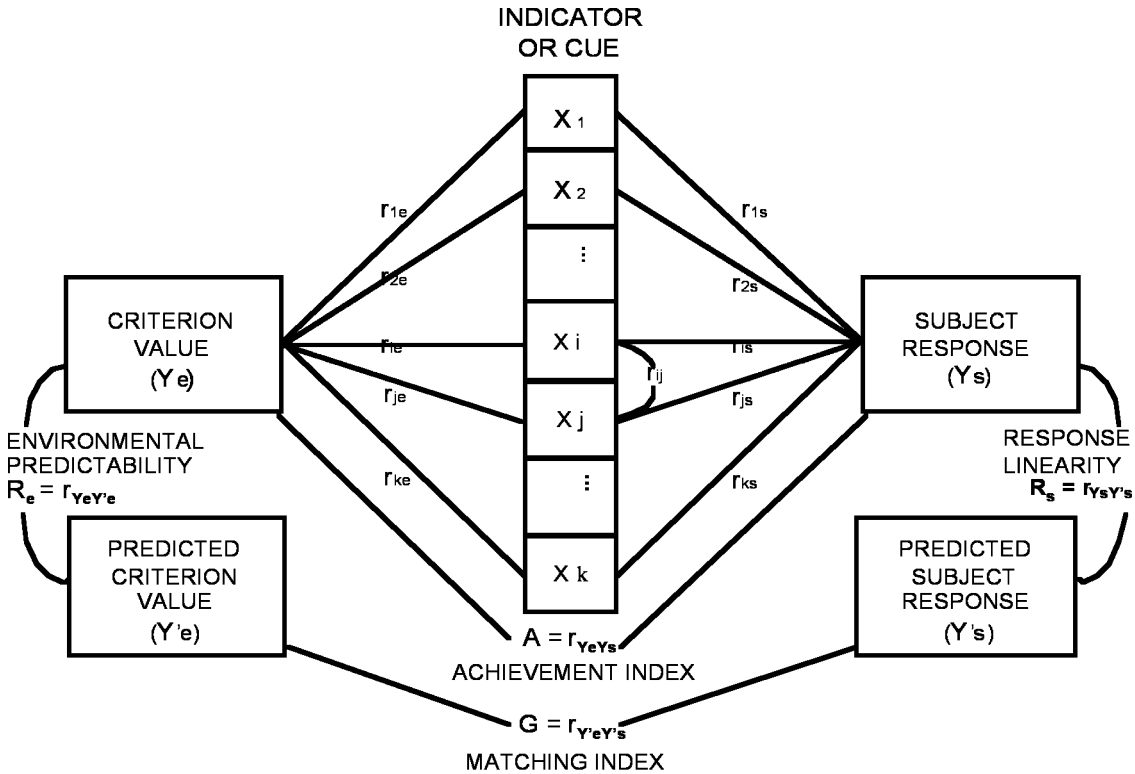

Figure 2
The Brunswik lens model (source: adapted from Slovic and Lichtenstein 1971)

ology (Campbell 1974, 1996). It is in the area of methodology development, however, where many of the most important contributions have been made.

3.1 The Lens Model and Clinical Inference

The lens model, backed by probabilistic functionalism and representative design, has been at the center of an entire research tradition on clinical inference (see Hammond 1980, 1996). Significantly, many of the most important lens-model studies have been applied studies conducted with the aim of improving learning, judgment, and behavior. In this tradition, research problems have turned on the question of how pattern matching can explain differences in achievement (predictive accuracy) among individuals with different levels of experience, including novices as well as experts. The right side of the lens model (Fig. 2) is used to predict a 'distal' environmental variable (e.g., future university enrollments) by regressing individual judgments about the distal variable on a set of interrelated and mutually substitutable informational cues (e.g., unemployment, income per capita, changes in age structure. In turn, the left side of the model focuses on predictions of that same distal variable derived from a multiple regression analysis in which the same cues (although here they are 'indicators' or 'variables') are predictors.

An important variant of this basic design is one where members of different professional groups—for example, scientists and lawyers in some area of science policy—occupy the right and left sides of the lens model. Either group may be taken as a reference class, and the calculation of the matching index, *M*, expresses the degree to which their prediction patterns match. In applied contexts, the judgments of each participant (subject) are externalized and made available to other participants. In addition, the way that each participant uses information (e.g., in a linear or nonlinear fashion) is estimated through a process of 'policy capturing.' These and other applications of the lens model have been designed to improve learning and judgments in contexts ranging from R&D planning, police work, gun control, and human factors research. The lens model has a computerized decision support program ('Policy PC') and its theoretical foundation has been redefined from social judgment theory to *cognitive continuum theory* (Hammond 1996).

3.2 Methodological Triangulation

The recognition that nature is not directly observable—that our predicament as knowers is that we must employ many intercorrelated and mutually substitutable proximal cues to infer the properties and behavior of distal objects—means that science and other forms of 'distal knowing' involve a process of pattern matching through triangulation (Campbell 1966). A number of important methodologies were developed on the basis of this recognition. One of these is the substitution of definitional operationism with multiple operationism, a substitution that involves triangulation among two or more operational definitions, each of which is seen as approximate, fallible, and independently imperfect. A second development was the expansion of multiple operationism to include theoretical constructs as well as methods for their measurement. Here the multitrait-multimethod matrix reconceptualized *construct validity* as a 'trait-method unit,' and introduced the concepts of *convergent validity* (multiple measures of the same construct should converge) and *discriminant validity* (multiple measures of different constructs should diverge) that parallel aspects of Brunswik's metaphor of a double-convex lens. The 1959 article in which the multitrait-multimethod matrix was first published (Campbell and Fiske 1959) is reputed to be one of the most highly cited in the social and behavioral sciences. A third development is multiple triangulation, including critical multiplism (Cook 1984), which involves the inclusion of multiple theories, methods, measures, observers, observations, and values. These and other forms of methodological triangulation enable 'strong inference' in the natural and social sciences (Platt 1964) and affirm that the vast bulk of what is known is based on processes of indirect, vicarious learning.

3.3 Quasi-Experimental Design

The methodology of representative design, as we have seen, rejected the classical experiment on grounds that it is unrepresentative of the usual ecology of in which knowers function. Representative design was carried forward into the applied social sciences by Donald T. Campbell and associates (Campbell and Stanley 1963, Cook and Campbell 1979). Their quasi-experimental designs were contrasted with the classical laboratory experiments in which: an outcome variable is explained by a single independent (treatment) variable (the so-called 'rule of one'); other possible explanations are ruled out through random selection of subjects; and the experimenter has virtually complete control over all contingencies. Quasi-experimentation, although it may use some of the features of classical experiments (e.g., repeated measures and control groups) should be contrasted with experiments in the analysis of variance tradition of Ronald Fisher, who envisioned experimenters who 'having complete mastery can schedule treatments and measurements for optimal statistical efficiency, with the complexity of design emerging only from that goal of efficiency. Insofar as the designs discussed in the present chapter become complex, it is because of the intransigency of the environment: because, that is, of the experimenter's lack of complete control' (Campbell and Stanley 1963, p. 1).

Because quasi-experimental designs are intended for research in settings in which numerous contingencies are beyond the control of the experimenter, many rival hypotheses (alternative explanations of the same outcome) can threaten the validity of causal claims. These rival hypotheses are organized in four sets labeled threats to statistical conclusion, internal, external, and construct validity (Cook and Campbell 1979, Chap. 2). Plausible rival hypotheses must be tested and, where possible, eliminated. This process of eliminative induction is a qualified form of Mill's *joint method of agreement and difference* and Karl Popper's *falsificationist* program. Quasi-experimentation is part of a wider evolutionary critical-realist epistemology (see Campbell 1974, Cook and Campbell 1979, Shadish et al. 2000) according to which knowers adapt to real-world environments by using overlapping and mutually substitutable informational sources to test and improve their knowledge of indirectly observable (distal) objects and behaviors. Quasi-experimentation is a form of pattern matching.

3.4 Pattern-Matching Case Studies

When quasi-experimental designs are unfeasible or undesirable, several forms of case study analysis are available. Each of these involves pattern matching.

(a) *Theory-Directed Case Study Analysis*. When a well-specified theory is available, a researcher can construct a pattern of testable implications of the theory and match it to a pattern of observations in a single case (Campbell 1975). Using statistical 'degrees of freedom' as a metaphor, the theory-directed case study is based on the concept of 'implications space,' which is similar to 'sampling space.' Testable implications are functional equivalents of degrees of freedom, such that the more implications (like a larger sample) the more confident we are in the validity of the conclusions drawn. But because theories are almost inevitably affected by the culturally acquired frames of reference of researchers, the process of testing implications should be done by at least two 'ethnographers' who are foreign to and native to the culture in which the case occurs. The process of triangulation among observers (ethnographers) can be expanded to include two (or more) cases.

(b) *Qualitative Comparative Case Study Analysis*. Two or more cases are compared by first creating a list of conditions that are believed to affect a common

outcome of interest (see Ragin 1999, 2000). The multiplication rule, r^m, is used to calculate the number of possible ordered configurations of r categories, given m conditions. When $r = 2$ and $m = 4$, there are $2^4 = 16$ configurations, each of which may involve causal order. These configurations become the rows in a 'truth table,' and each row configuration is sequentially applied to an outcome with r categories (e.g., successful vs. unsuccessful outcome). Because this method examines possible configurations, and two or more different configurations may explain the same outcome in different cases, the qualitative comparative method should be contrasted with traditional (tabular) multivariate analysis. The qualitative comparative method matches configurable patterns that have been formally structured by means of set theory and Boolean algebra against patterns of observations in case materials.

(c) *Modus Operandi Analysis*. When quasi-experimental research is not possible, *modus operandi* methods (see Scriven 1975) may be appropriate for making causal inferences in specific contexts. *Modus operandi* methods are based on the analogy of a coroner who must distinguish symptoms and properties of causes from the causes themselves. The first step is to assemble a list of probable causes, preferably one that is quasi-exhaustive. The second is to recognize the pattern of causes that constitutes a *modus operandi*—modus refers to the pattern, while operandi refers to specific and 'real' causes. The *modus operandi* of a particular cause is its characteristic causal chain, which represents a configuration of events, properties, and processes. *Modus operandi* analysis has been formalized, partially axiomatized, and advanced as a way to change the orientation of the social and behavioral sciences away from abstract, quantitative, predictive theories toward specific, qualitative, explanatory analyses of causal patterns (Scriven 1974, p. 108).

4. Conclusion

To date, the only thorough and systematic theory of pattern matching is that of probabilistic functionalism and its methodological complement, representative design. Pattern matching is essential for achieving knowledge of external objects and events, of causal regularities that have been repeatedly recognized as the same, of the character of discrete observations formed by viewing them in context, of the fit between a curve and a set of data points, of the extent to which theories cohere, and of the degree to which theory and data correspond. Underlying all pattern matching methodologies is a shared recognition that knowing in science is profoundly indirect, vicarious, and distal.

See also: Case Study: Logic; Case Study: Methods and Analysis; Causation (Theories and Models): Conceptions in the Social Sciences; Constructivism/Constructionism: Methodology; Explanation: Conceptions in the Social Sciences; Laboratory Experiment: Methodology; Problem Selection in the Social Sciences: Methodology; Qualitative Methods, History of; Triangulation: Methodology

Bibliography

Alcoff L M 1996 *Real Knowing: New Versions of the Coherence Theory of Truth*. Cornell University Press, Ithaca, NY

Anscombe F J 1973 Graphs in statistical analysis. *American Statistician* **27**: 17–21

Ayer A J 1936 *Language, Truth, and Logic*. Golanc, London

Brunswik E 1956 *Perception and the Representative Design of Psychological Experiments*, 2nd edn. University of California Press, Berkeley, CA

Brunswik E 1952 *The Conceptual Framework of Psychology*. University of Chicago Press, Chicago

Campbell D T 1959 Methodological suggestions for a comparative psychology of knowledge processes. *Inquiry* **2**: 152–82

Campbell D T 1966 Pattern matching as an essential in distal knowing. In: Hammond K R (ed.) *The Psychology of Egon Brunswik*. Holt, Rinehart, and Winston, New York

Campbell D T 1974 Evolutionary epistemology. In: Schilpp P A (ed.) *The Philosophy of Karl Popper*. Open Court Press, La Salle, IL

Campbell D T 1975 "Degrees of Freedom" and the case study. *Comparative Political Studies* **8**: 178–93

Campbell D T 1986 Science's social system of validity-enhancing collective belief change and the problems of the social sciences. In: Fiske D W (ed.)

Campbell D T 1988 *Methodology and Epistemology for Social Science: Selected Papers*. University of Chicago Press, Chicago

Campbell D T 1996 From evolutionary epistemology via selection theory to a sociology of scientific validity. *Evolution and Cognition*

Cook T D, Campbell D T 1979 *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Houghton Mifflin, Boston

Hammond K R (ed.) 1966 *The Psychology of Egon Brunswik*. Holt, Rinehart, and Winston, New York

Hammond K R 1980 *Human Judgment and Decision Making: Theories, Methods, and Procedures*. Praeger, New York

Hammond K R 1996 *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. Oxford University Press, New York

Putnam H 1981 *Reason, Truth, and History*. Cambridge University Press, Cambridge, UK

Quine W V 1951 Two dogmas of empiricism. *Philosophical Review* **60**(1): 20–43

Ragin C C 1999 Using comparative causal analysis to study causal complexity. *HSR: Health Services Research* **34**(5), Part II

Ragin C 2000 *Fuzzy-Set Social Science*. University of Chicago Press, Chicago

Rescher N 1980 *Induction: An Essay on the Justification of Inductive Reasoning*. University of Pittsburgh Press, Pittsburgh, PA

Schweder (eds.) *Metatheory in Social Science: Pluralisms and Subjectivities*. University of Chicago Press, Chicago
Scriven M 1975 Maximizing the power of causal investigations: The modus operandi method. In: Glass G V (ed.) *Evaluation Studies Review Annual*. Sage Publications, Beverly Hills, CA, Vol. 1
Shadish W, Cook T, Campbell D T 2000 *Quasi-Experimentation*. Houghton Mifflin, Boston
Slovic P, Lichtenstein S 1971 Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance* **6**(6): 649–744
Trochim W 1990 Pattern matching and psychological theory. In: Chen R (ed.) *Theory-Driven Evaluation*. Jossey Bass, San Francisco

W. N. Dunn

Pavlov, Ivan Petrovich (1849–1936)

Ivan Petrovich Pavlov is recognized as one of the most renowned of Russian scientists and is clearly the best known Russian physiologist of his era. His work on the physiology of the digestive system was well respected by his colleagues worldwide during his lifetime. This respect earned him the first Nobel Prize awarded to a physiologist and the first to a Russian scientist in 1905. Although the excellence of his work on the physiology of the digestive system is well recognized, Pavlov is even better known for his investigations into what he termed 'conditioned reflexes.' This work has had a worldwide impact on the fields of learning and memory, psychology, neuroscience, etc. Pavlov's life can be divided roughly into three phases. The first, including his early years, was a phase in which Pavlov received his early training in physiology and medicine. A second investigational phase was one in which his work on the physiology of the digestive system was done. The third includes his work, for which he became most famous, on conditioned reflexes.

1. *Pavlov's Early Life*

Ivan Petrovich Pavlov was born on September 14, 1849, in the small town of Ryazan, about 100 miles southeast of Moscow. He was the eldest of five children. His father was a priest in one of the poorer parishes. Thus, during Pavlov's early life he lived in extreme poverty. Nevertheless, the life of Pavlov and his siblings was apparently intellectually rich and their father instilled a love of learning in each of his children. Pavlov entered the ecclesiastical high school of Ryazan, and from there to the local seminary, obstentially to pursue an ecclesiastical career, as his father and other members of his family had done. Pavlov thus entered school during the Russian Cultural Revolution of the 1860s and therefore was exposed to the leaders of this revolution, such as Turgeanyev, Doskovsky, Tolstoy, etc. However, he was apparently most heavily influenced by Pisarev, from whom he learned of Darwin's theory of natural selection and the famous Russian physiologist Sechenov, who had just published *Reflexes of the Brain* in 1896. It is clear that these kinds of influences on Pavlov steered him in the direction of science and consequently he left the ecclesiastical seminary and entered the Mathematics-Physics Department at St. Petersburg State University.

In 1875 Pavlov graduated from the University with a degree in natural sciences. However, due to his interest in physiology he started work as an assistant in the Department of Physiology in the Medical Surgery Academy (now the Military Medical Academy) and in 1879 received a medical diploma. He was then recommended for an appointment with Andre Botkin, who was the most renowned clinician in St. Petersburg at the time. Botkin assigned Pavlov the task of heading his recently established experimental research laboratory. Since Botkin had little interest in experimental work, Pavlov was able to work pretty much independently on problems of his choice. He worked here until 1890. He was heavily influenced by Botkin's theories on the role that the nervous system plays in the regulation of physiological function, and it was here that Pavlov began his studies of the peripheral nerves of the heart and where in 1883 he successfully defended his doctoral dissertation in this area.

During this period of time he also became acquainted with a small circle of friends including another university student, Seraphima Vasilievna, who was in her last year at the Pedagogical Institute. They were married in 1881. During the first years of their married life they lived in extreme poverty. After a first miscarriage, Seraphima gave birth to a son but he also died the next summer. They did, however, later have four children (three boys and a girl) and, based on Seraphima's memoirs published after Pavlov's death, apparently enjoyed a happy family life. In 1883 Pavlov spent two years working abroad in the laboratories of Heidenhain and Ludwig in Germany. On his return to St. Petersburg, he resumed his work in Botkin's laboratory and in 1890 was appointed Professor in the Pharmacology Department at the Military Medical Academy. Five years later he joined the Physiology Department there, which he headed for the rest of his career. He was also later appointed the Head of the Department of Physiology in the newly established Institute of Experimental Medicine. Pavlov worked in his laboratory there until his death at the age of 87 in Leningrad on February 27, 1936.

2. *Pavlov's Work on the Digestive System*

Pavlov did most of his research on the physiology of digestion during his early years at the Institute for

Experimental Medicine. Here he was able to develop his surgical skills with chronic experiments, which he advocated over the acute experiments that were typically done in physiology at that time. Pavlov believed that the study of the entire organism in as normal a state as possible was crucial to understanding physiological systems. Moreover, he was an excellent surgeon and became quite experienced in developing surgical procedures that allowed his dogs to fully recover. In such animals the digestive processes could be studied under normal conditions. As a consequence of this approach to physiology he developed a 'stomach pouch,' which involves isolating a portion of the dog's stomach, so that it can be directly observed by the experimenter. Unlike previous researchers, the nervous input to the isolated portion of the stomach was intact in the 'Pavlov Pouch.' Using this preparation Pavlov could observe the digestive juices and enzymes that occurred as a result of normal food intake, or after direct placement of food into the stomach. It was the comparison of these two states which led to Pavlov's interesting discovery that food placed in the mouth as a normal consequence of eating, produced a significantly larger amount of gastric secretion than when food was placed directly in the stomach. Moreover, when animals were sham fed, viz. when the normal intake of food through the mouth was allowed to take place but the food was diverted externally, a significant amount of gastric secretion occurred, even though the food never reached the stomach.

It was these types of experiments, which led to Pavlov's pioneering discoveries regarding the control of the digestive system by its autonomic nervous input, primarily the vagus nerve. He showed for example that a major consequence of this neural control involves both gastric and pancreatic secretions before any food actually reaches the stomach, by the sight of the food bowl for example, or other stimuli previously associated with eating. These 'psychic secretions' turned out to be of great importance for digestion and led to Pavlov's major conclusions regarding what came to be known as the doctrine of nervism, in which the major tenet states that the physiological control of bodily systems is primarily through its nervous input. This work culminated in the publication of *Lectures on the Work of the Digestive Glands* in 1897, later translated into English in 1910 (Pavlov 1910). However, it was during this same period of time that two British physiologists, W. Bayliss and E. J. Starling, demonstrated that the secretin of the pancreatic enzymes was primarily due to release of a substance by the intestines during eating, which came to be called secretion. These investigators thus claimed that the doctrine of nervism was invalidated by their discoveries. However, as is typically the case with two opposing theories of biological function, both experimental outcomes were eventually confirmed. The new science of neuroendocrinology thus came to be established, and it is now widely accepted that nervous control over hormonal output is a ubiquitous aspect of physiological function.

3. *Pavlov's Work on Conditioned Reflexes*

As a result of his work on the digestive processes, as noted above, Pavlov also observed that non-food stimuli that became associated with food were able to elicit salivary secretions, and to a lesser extent digestive secretions, even though no food was actually present in the mouth. These 'psychic secretions' came to form the basis for the remainder of Pavlov's work and resulted in his experiments moving in a new direction, which caused an almost complete cessation of work on the digestive system proper. However, Pavlov believed that by studying 'psychic secretions,' which he referred to as 'conditioned reflexes,' he would be able to demonstrate in an objective manner how the brain controls adaptive behaviors. Pavlov developed a salivary fistula, similar to the stomach pouch, in which the release of salivation in the mouth was directed through a tube to a container outside the mouth. Through this technique he was able to discover the amount and kinds of salivary secretions that were produced by different kinds of sensory stimuli.

Using these techniques Pavlov began the study of a new kind of learning, which has come to be known as 'classical' or 'Pavlovian' conditioning. Classical conditioning occurs whenever a neutral stimulus acts as a signal for a forthcoming significant event. Thus, for example, in one of Pavlov's original experiments a pure tone, a metronome, or some other equally neutral stimulus signaled to a dog that it was about to receive an appetitive stimulus, i.e., meat powder (see Pavlov 1927). In Pavlov's experiments this neutral auditory stimulus was termed the conditioned stimulus (CS) and the meat powder the unconditioned stimulus (UnS). The UnS always elicits what Pavlov referred to as an unconditioned response (UR), viz. salivation in response to the meat powder. The behavior in which Pavlov became most interested, however, was not increased salivation in response to the meat powder, which he had previously studied intensively, but the new learned response to the neutral CS. This response also consisted of salivation, but in this case salivation in response to the initial neutral auditory stimulus, which resulted from its consistently preceding the meat powder over many CS/UnS presentations. This new learned response was referred to as a conditioned response (CR) and was thought to be only temporary and required reinforcement by the UnS for its maintenance. The UnS is thus often referred to as a reinforcer, since it reinforces the new response to the CS.

It should be noted that the CR does not, however, always resemble the UR as is the case with the original salivation experiments, which Pavlov reported. Indeed

often the CR appears to be opposite to the UR. For example, the autonomic changes associated with the contextual cues that signal drug administration, an often-studied type of classical conditioning, are opposite in direction to those produced by the drug itself (Siegel 1979) and the heart rate CR to CSs that signal aversive UnSs consist of bradycardia, whereas the UR to these same aversive stimuli alone consists of tachycardia. There are many other experimental operations that have been studied since the time of Pavlov, which produce similar new responses to an original neutral stimulus that is always followed by either a noxious or appetitive event. Pavlov thus developed a new experimental methodology for studying the role of the brain as it initiates new behaviors enabling animals to adapt to their environmental circumstances.

One of Pavlov's greatest contributions to physiology was to emphasize that the nature of the 'psychic secretions,' which were previously thought to be in the realm of psychology, could be objectively studied by physiologists through the conditioned reflex method. He thus believed that the true road to understanding brain function and therefore human behavior was through the objective physiological techniques, which he had developed. In fact Pavlov went on to consider a range of higher level functions such as thinking, reading, emotional reactivity, etc., in terms of conditioned reflexes. Using the ideas of cortical excitation and inhibition, which he studied extensively using the conditioned reflex methodology, he was thus able to explain many complicated behavioral phenomena, such as the neuroses and other psychiatric disturbances. Although Pavlov's influence on experimental psychology has been extensive and dramatic, as detailed in Sect. 4, many of the details of his theoretical interpretations have since been found to be lacking. Nevertheless the basic operational procedures for determining the effects of a signal on brain processing has been extremely influential in studying brain function.

4. *Pavlov's Influence on Modern Psychology*

During his lifetime and immediately thereafter, Pavlov had a tremendous influence on physiology and the study of brain function. Many students were drawn to his laboratory during the early part of the twentieth century to study the basic laws governing the activity of the brain, and during this period of time Pavlov received worldwide acclaim and recognition. Pavlov's work was little affected by the Russian Revolution, which was also ongoing during this time. He maintained a skeptical attitude regarding politics and government, but nevertheless his worldwide recognition led the Communists to continue to fund his research at high levels. Thus, the Soviet Union became known for its support of the study of physiology, and it was during this time that a great center for the study of physiology with many distinguished workers was developed in the Soviet Union, primarily under Pavlov's leadership.

It was his influence outside the Soviet Union, however, that resulted in Pavlov's most notable successes in science, primarily in the field of psychology, which Pavlov had previously rejected. Thus the conditioned reflex methodology was instrumental to the development of the behavioristic movement in psychology in the early part of the twentieth century. John Broadus Watson, one of the major pioneers in the development of behaviorism, utilized the conditioned reflex methodology to explain the entire subject field of psychology, which consisted, according to behaviorists, in the study of overt behavior without reference to subjective phenomena such as cognitions, learning, feelings, etc. The publication of *Psychology as the Behaviorist Views It* (1913) by Watson thus became instrumental in informing American and English scientists of Pavlov's work. The translation of Pavlov's *Conditioned Reflexes* (1927) by Anrep made his work available to English speaking scientists. As a result, experimental psychologists began to explore the new conditioned reflex methods for studying behavior at several academic centers in the USA. As was noted by Babkin (1949), however, much of the work done by the American researchers focused on skeletal reflexes, as opposed to the visceral reflexes, to which Pavlov and his students had previously devoted most of their work.

One of the basic experiments developed by American psychologists was the classical eyeblink-conditioning paradigm. Using this methodology human or animal subjects are presented with a corneal airpuff, which causes reflexive closure of the eyelids. However, when this unconditioned stimulus is preceded by a conditioned stimulus such as a light or pure tone, eventually subjects began to show anticipatory eyeblinks during the conditioned stimulus, which occur even though the corneal airpuff is not presented. This paradigm came to be used extensively in the early 1940s and 1950s as behaviorism began to gather momentum. Its success in the hands of the American psychologists led to the de-emphasis on visceral changes as conditioned responses in the USA. However, Horsley Gantt, who had previously studied with Pavlov, demonstrated in his Johns Hopkins laboratory that both visceral and skeletal responses could be studied in the same organisms, and that the two responses differed greatly in their acquisition and the conditioning parameters required to elicit learning (Gantt 1960). Much of this early work on classical conditioning was, however, overshadowed by the emphasis given to operant conditioning by other behaviorists, such as B. F. Skinner and Clark Hull. Thus, it was only during the early 1960s that classical conditioning became a popular technique for studying behavior. This was due to the development by I.

Gormezano of the classically conditioned nictitating membrane response in the rabbit (Gormezano 1966). The nictitating membrane is a third laterally moving eyelid found in the rabbit and some other mammals (e.g., cat). Using this animal preparation Gormezano and his students were able to demonstrate the parametric circumstances under which this kind of learning takes place and what kinds of visceral changes accompany them, e.g., conditioned changes in heart rate, blood pressure, etc. The popularity of this technique became even greater when integrated with the simultaneously developing field of behavioral neuroscience. Classical conditioning offers several advantages for studying concomitant brain function as well as new learned behaviors. For example, concomitant electrophysiological recording from single neurons in specific parts of the brain have led to the discovery that different brain structures are involved in visceral versus skeletal learning (Thompson 1991). Moreover, more recent brain scanning techniques in humans have demonstrated that even during simple classical eyeblink conditioning activation of several specific areas of the brain, which are known to be involved in learning and memory processes, occurs (e.g., Blaxton et al. 1996).

Thus the use of the conditioned reflex techniques originally developed by Pavlov has come to be one of the major methods used in studying brain-behavior relationships in modern psychology. Pavlov's contribution to this new technology was at a basic level and much of his earlier conclusions regarding the results of his manipulations were, of course, erroneous. However, his contributions we now know were instrumental to the development of modern psychology.

See also: Autonomic Classical and Operant Conditioning; Conditioning and Habit Formation, Psychology of; Experimentation in Psychology, History of; Psychological: Historical and Cultural Perspectives

Bibliography

Babkin B P 1949 *Pavlov: A Biography*. University of Chicago Press, Chicago, IL

Blaxton T A, Zeffiro T A, Gabrieli J D E, Bookheimer S Y, Carrillo M C, Theodore W H, Disterhoft J F 1996 Functional mapping of human learning: A positron emission tomography activation study of eyeblink conditioning. *Journal of Neuroscience* **16**: 4032–40

Gantt W H 1960 Cardiovascular component of the conditional reflex to pain, food, and other stimuli. *Physiological Review* **40**: 266–91

Gormezano I 1966 Classical conditioning. In: Sidowski J B (ed.) *Experimental Methods and Instrumentation in Psychology*. McGraw Hill, New York

Pavlov I P 1910 *Lectures on the Work of the Digestive Glands*, 2nd edn. Translated by W H Thompson. Charles Griffin and Co., London

Pavlov I P 1927 *Conditioned Reflexes*, translated by G V Anrep. Oxford University Press, London

Sechenov I M 1866 *Refleksy Golovnogo Mozga* (*Reflexes of the Brain* 1965. MIT Press, Cambridge, MA)

Siegel S 1979 The role of conditioning in drug tolerance and addiction. In: Keehn J D (ed.) *Psychopathology in Animals*. Academic Press, New York

Thompson R F 1991 Are memory traces localized or distributed? *Neuropsychologia* **29**: 571–82

Watson J B 1913 Psychology as the behaviorist views it. *Psychological Review* **20**: 158–77

D. A. Powell

Pay and Compensation, Psychology of

1. The Design of Compensation

In industrialized countries the compensation of managers and employees usually consists of three or four component parts. One part is rooted in the value of the job which is held, providing for the fixed base pay (wage or salary). The second part reflects an assessment of the effort exerted or the performance result, often called the 'performance bonus.' The third part deals with 'secondary labor conditions,' such as health insurance, retirement provisions, etc. Firms in several countries offer still additional benefits—perquisites—to their personnel (or merely their managers), such as car lease, lower interest rates on mortgage loans, etc. However, there are sizable differences between (and often also within) countries regarding the proportional contribution of each of these parts to an individual's total compensation.

Yet the common design in the composition of compensation in most countries is remarkable. This cuts across considerable differences between countries in industrial relations, wealth and welfare, and culture. Why? Probably partly because traditions in trades and industrial sectors, laws and regulations, and the ideas of labor unions and employers' federations are playing a role. And also, in all likelihood, partly because of some widely shared beliefs and attitudes about the differential impact of the various compensation components upon the work behaviors of organization members. This borders on the psychological perspective on pay, which addresses the impact of compensation upon the attitudes and work behaviors of organization members, both individually and as part of teams or larger work units. More particularly, the psychological study of pay and compensation aims to identify determining variables at different levels of analysis (such as pay system, or strength of individual expectations), moderating aspects (such as the nature of the task), and individual, group, or company-wide results (such as performance or productivity).

This perspective determines the design of this article. First, the nature of the four compensation components

is further described, in relation to some characteristic beliefs and attitudes on how they operate. Next, a few psychological theories on pay and compensation will be discussed. Finally, some implications for psychological research on pay and applications in practice are discussed.

2. Four Components and Concomitant Beliefs

Although the background of the compensation components is discussed in relation to what is characteristically expected regarding each of them, I do not imply that a company's management is always engaged in decision making along these lines. Indeed, pay systems and compensation provisions may have been introduced in a firm for reasons of tradition, imitation of other companies, or merely because alternative choices were not considered.

2.1 Base Pay

In most countries base pay constitutes the main part of an organization member's take-home pay. Yet there may be sizable differences in the proportion of base pay between trades and job categories. Traditionally, the pay of sales representatives is determined to a large degree by actually achieved sales figures. Managerial pay is characterized by a proportionally lower base-pay component than nonmanagerial pay, in particular for senior managers (Rynes and Bono 2000). Base or job pay is often set through a system of job evaluation (cf. Thierry and De Jong 1998). Increasingly, international wage and salary surveys are used, as well as more subjective impressions of the going rate on the market. Base-pay levels are usually fixed as long as the term of individual contracts or collective agreements applies.

Often the fixed base-pay level is a reflection of the belief that stability in pay contributes to individual feelings of security. It is also supposed that without the need to worry about their pay, working people will concentrate more upon the content of their tasks. Moreover, a high base-pay level may imply the notion that work motivation is considerably more affected by nonmaterial needs and goals than by material ones. Possibly this results in more intrinsic motivation (cf. Deci and Ryan 1985, Deci et al. 1999). Characteristically, levels of base pay increase each year, usually through an additional 'step' on the wage or salary scale. Although the latter increase may have been awarded to make up for increased costs of living, it often also mirrors the idea that more maturity (i.e., getting older and becoming more experienced) goes along with better performance.

2.2 Results-oriented Pay

This category, also called 'pay for performance,' is host to a variety of systems and forms. Results may be derived from individual performance appraisal sessions, in which rather abstract characteristics (such as initiative or creativity) are interpreted and considered. Results may also refer to quantitative outcome scores, for instance the number of new bank clients secured and the volume of their investments. Results may relate to individual data, but also to those of teams, units, and whole organizations (such as sales volume or net profit). These data may be provided weekly, monthly, or over a much longer period. The 'pay' is often allocated in the form of a bonus (paid only once), but it may also be given as an extra, 'structural' salary increase. The amount of results-oriented pay may be rather modest (e.g., 3 percent), but it may also exceed the base pay's proportional part of total compensation (see, more extensively, Rynes and Bono 2000, Bartol and Locke 2000, Thierry in press).

What is results-oriented pay supposed to effectuate? A widespread belief holds that it stimulates an individual's work motivation, effort, or performance. Research evidence (Thierry in press) shows that this can be frequently supported, although there are also many instances of failure. Another assumption is that results-oriented pay is perceived by many employees (and managers) as more fair, since it compensates for perceived individual and group differences in effort and/or performance (Van Silfhout 2000). Organizations in which this assumption is endorsed usually assign a particular budget to covering these merit-pay expenses. Making pay dependent upon performance results may be used to cut costs as well. Moreover, it may be used to feed performance data back to employees, and, more generally, for communicating with them.

More than is the case for other compensation components, the phase of preparing and introducing results-oriented pay is rather complicated and time consuming. It may imply that a part of the work flow is being organized more efficiently, that norms and targets are being established, that results are being fed back, that employees and managers are being consulted about the new system, and so forth. In particular, when better performance results have been reached, a point of concern is that these activities may have contributed to these results rather than the linkage between performance and pay.

2.3 Secondary Labor Conditions

There are huge differences between countries in the handling of secondary conditions. Traditionally, companies in the USA offered a very modest package to their members—leaving to them the decision whether or not further to insure themselves—whereas companies in northwestern Europe provide encompassing social security within the framework of collective labor agreements. In the latter case, conditions relate,

for example, to health insurance, disability coverage, retirement benefits, or unemployment provisions. These are intended to help working people to compensate for risks they run in their work and lives. Yet, such social security packages are not very well tailored to the individual organization member's conditions and wishes. Moreover, because of health research and technological innovations, there is a continuing trend to expand the number and the nature of social security provisions.

It is therefore an interesting development that the 'cafeteria plan' is gaining popularity on both sides of the Atlantic (Barber and Bretz 2000). According to this plan (also called 'à la carte' or 'flex-pay') organization members may exchange one or more standard labor conditions for optional ones. In US companies this plan was initiated some 40 years ago, in order to gradually expand social security coverage. Consequently, managers and employees may tailor their social security package to their particular conditions, which may lead to more utility against similar or lower costs for the company. More recent European experience, for instance in the UK and The Netherlands (cf. Thierry 1998), indicates that organizations think the plan equips them better as competitors in the labor market for 'flexible, employable' applicants. They may also express their particular culture and climate in their choice of alternative provisions. Consequently, their members are expected to be more motivated, to work better, to be more satisfied with their pay package, and to have more knowledge of their compensation components. Research evidence shows that only the latter two expectations are frequently met (Thierry 1998, in press).

2.4 Perquisites

Organizations may offer their personnel (or selected categories, such as senior managers) still further labor conditions. Early examples came in the manufacturing sector, in which an allowance for occupational clothing or cloth cleaning was sometimes provided. Current practices include (mobile) phone cost coverage, car leasing, lower mortgage interest, shares on the stock market, options on company shares, and so forth. Some of these options may be available through a cafeteria plan (see Sect. 2.3). Perquisites were provided to make up for particular inconvenient conditions and occupational costs, but current objectives relate to attracting potential applicants on the labor market and, in particular, to keeping qualified personnel within an organization.

2.5 Beliefs in Perspective

The preceding overview of compensation components shows that each component is characterized by rather distinct beliefs and expectations about how compensation may affect the attitudes and work behaviors of managers and employees. Yet an important issue is whether these beliefs are founded in concepts relevant to psychological theories on pay and compensation. Some of these will be reviewed here (more extensively in Thierry in press), but the various beliefs are to be recognized in almost all objectives of pay (cf. Thierry 1998):

to *attract* applicants from the labor market;
to *retain* qualified personnel within an organization;
to *motivate* managers and employees to perform well;
to *alert m*anagers and employees to engage in different performance behaviors in times of change;
to *balance* inconvenient working conditions with an attractive award;
to *prevent* industrial conflicts (or to solve these where they have occurred);
to *cut down* labor costs; and
to further a distinct, recognizable business unit or group *culture*.

3. Some Psychological Theories on Pay

Most psychological theories on pay have not been explicitly designed for understanding and predicting people's behaviors and attitudes regarding pay. Rather, they have been derived from more general theories, for instance regarding motivation. The theories selected here stem from both 'categories.'

3.1 Expectancy Theory

This theory concerns matters of work motivation (Vroom 1964, Van Eerde and Thierry 1996), in particular the choice process of an employee being faced with the necessity or the opportunity to mark a course of action among alternative activity patterns. Suppose that an employee considers whether to spend all their energy and resources in making the current job a great success, hoping for a promotion in the not too distant future, or whether to apply for a job in another organization, which involves taking several courses in order to become more employable. Expectancy theory holds that three cognitions determine the eventual choice:

the probability that a particular level of effort leads to a specific level of performance;
the probability that this performance level causes the attainment of particular outcomes; and
the attractiveness of those outcomes.

The alternative with the highest subjective utility for the employee will be chosen. Thus, assuming that a job elsewhere is slightly more attractive than a promotion, the employee may nevertheless opt for the internal

promotion, since they estimate the probability of getting access to courses that will increase their employability as much lower than turning the current job into a success.

Lawler (1971) has applied the last two cognitions to the importance of pay. Accordingly, its importance to an employee is a function of:

the extent to which pay is perceived as offering resources for satisfying particular needs or goals; and

the importance of these needs and goals to that employee.

In other words, the more pay is considered to be instrumental in getting more security in life (e.g., through purchasing an insurance policy), in being recognized by other people, and in gaining more independence—to mention some examples—and the more these needs are important to the individual employee, the more important pay has become. Pay is not important in itself: it may be a stronger or weaker vehicle for reaching other outcomes.

One implication bears upon the system of results-oriented pay: in order to be effective, it is necessary that an organization member perceives a clear 'instrumental' link between one or more particular performance levels and specified amounts of pay (e.g., Van Silfhout 2000).

3.2 *Equity Theory*

Expectancy theory is focused upon the individual human being, engaged in means-to-end relationships. Is that individual affected by one or more others, by smaller or larger social networks in the choice-making process? Yes, according to Ajzen and Fishbein's theory of reasoned action (1980), in which the individual's beliefs about what others think he or she should do are supposed to affect the choices he or she make. Equity theory (Adams 1963, 1965) includes 'social others' indirectly in the balance between contributions and inducements. Adams holds that an individual strives for such a balance *relative* to the balance that individual perceives between contributions and inducements as applying to a referent. A referent may be one or more human beings, a particular group, an organization's policy, and that individual at an earlier or later stage of their career. Contributions relate to long-term investments and short-term costs, such as education, abilities, and effort. Among inducements are task performance, recognition, and compensation. Equity is achieved when the person's perceived contributions–inducements ratio is equal to their perception of the referent's ratio. An unequal balance results in dissonance and inequity.

A typical setting would be an experiment in which the ratio between a person's effort (to perform a task) and pay is compared with the effort–pay ratio as perceived to apply to the referent. A high level of inequity would be the result when the person perceives their own situation in terms of high effort for low pay, and the referent's ratio as low effort for high pay. Obviously, the person would then be motivated to restore an equitable balance. But how? That is difficult to predict. Available strategies are: effecting change in contributions and/or inducements; changing the referent; cognitively reinterpreting the inequitable situation; or leaving the organization. Harder (1992) showed in a study of baseball and basketball players that under-rewarded players tended to opt for utilitarian, egotistic behaviors that were thought to lead to future higher rewards.

3.3 *Reflection Theory*

This theory specifies the meanings of pay or compensation. A basic notion is, again, that pay has no significance in itself, but that it acquires meaning to the extent that it conveys information that is vital to an individual's self-identity. Accordingly, pay 'reflects' information about events in other fields. Four categories of meaning are distinguished:

(a) *Motivational properties*. This category is derived from expectancy theory (Sect. 3.1). Pay is meaningful to the extent that it is considered to be instrumental in satisfying a person's motives.

(b) *Relative position*. This bears upon two aspects. One reflects feedback about task performance relative to the targets set. Pay may signal whether particular corrective actions are necessary. The other aspect is social comparative: pay reflects how well a person is doing in comparison with others inside and/or outside the company.

(c) *Control*. Pay may reflect the amount of control the individual exerts on their environment (both the social network and material means) as well as the extent to which they are controlled by the environment.

(d) *Spending*. Here pay reflects the utility of goods and services purchased by the individual, and the ease or hardship with which this occurred.

Scales are available for measuring these meanings. The core of reflection theory is that the more an employee or manager reads meanings into their pay, the more their performance and satisfaction with pay is affected (Thierry 1998, 2001). Obviously, the design of a pay system and the manner in which it is administered are among the variables that influence the meanings pay may have.

4. *Conclusion*

An important question raised in the introductory section is whether the beliefs and expectations which seem to be basic to the four compensation components are grounded in psychological theories on pay. Just

three theories could be discussed; nonetheless, their tenets are rather representative. Expectancy and reflection theory provide some ground for results-oriented pay, in particular regarding the importance of unambiguous performance leading to pay perceptions. Yet it is not clear how these perceptions should be structured in the case of group performance; possibly equity theory's notion on contribution—inducement ratios should also be considered. The belief embedded in the concept of base pay—that stability and predictability of pay may cause employees and managers to be focused more upon the content of their work—does not take full account of the meanings of pay (put forward in reflection theory) relative to the meanings of nonmaterial work characteristics. This is clearly in need of more research. Continued research on cafeteria plans, including perquisites, is vital, in order to learn whether this compensation component meets its expectations. All in all, however, beliefs and expectations about compensation in practice are not so much missing the point as more outspoken than psychological theory would require them to be.

See also: Money and Finances, Psychology of; Money: Anthropological Aspects; Psychological Climate in the Work Setting

Bibliography

Adams J C 1963 Toward an understanding of inequity. *Journal of Abnormal and Social Psychology* **67**: 422–36

Adams J C 1965 Inequity in social exchange. *Advances in Experimental Social Psychology* **2**

Ajzen I, Fishbein M 1980 *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall, Englewood Cliffs, NJ

Barber A E, Bretz R D 2000 Compensation, attraction, and retention. In: Rynes S, Gerhart B (eds.) *Compensation in Organizations: Current Research and Practice*. Jossey-Bass, San Francisco, pp. 32–60

Bartol K M, Locke E A 2000 Incentives and motivation. In: Rynes S, Gerhart B A (eds.) *Compensation in Organizations: Current Research and Practice*. Jossey-Bass, San Francisco, pp. 104–47

Deci E L, Ryan R M 1985 *Intrinsic Motivation and Self-determination in Human Behavior*. Plenum Press, New York

Deci E L, Koestner R, Ryan R M 1999 A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* **125**: 627–68

Harder J W 1992 Play for pay: Effects of inequity in a pay-for-performance context. *Administrative Science Quarterly* **37**: 321–35

Lawler E E 1971 *Pay and Organizational Effectiveness*. McGraw-Hill, New York

Rynes S L, Bono J E 2000 Psychological research on determinants of pay. In: Rynes S L, Gerhart B A (eds.) *Compensation in Organizations: Current Research and Practice*. Jossey-Bass, San Francisco, pp. 3–31

Thierry H 1998 Compensating work. In: Drenth P J D, Thierry H, De Wolff C J (eds.) *Handbook of Work and Organizational Psychology*, 2nd edn. Psychology Press, East Sussex, UK, pp. 291–319

Thierry H 2001 The reflection theory on motivation. In: Erez M, Kleinbeck U, Thierry H (eds.) *Work Motivation in the Context of a Globalizing Economy*. L. Erlbaum Associates, Mahwah, NJ, pp. 141–58

Thierry H in press Enhancing performance through pay and reward systems. In: Sonnentag S (ed.) *Psychological Management of Individual Performance*. John Wiley, Chichester, UK

Thierry H, De Jong J R 1998 Job evaluation. In: Drenth P J D, Thierry H, De Wolff C J (eds.) *Handbook of Work and Organizational Psychology*, 2nd edn. Psychology Press, East Sussex, UK,Vol. 3, pp. 165–83

Van Eerde W, Thierry H 1996 Vroom's expectancy models and work-related criteria: A meta-analysis. *Journal of Applied Psychology* **81**: 575–86

Van Silfhout R K 2000 *Inequality in Pay Within Organizations: Normative and Instrumental Perspectives*. PhD thesis, Tilburg University, The Netherlands

Vroom V H 1964 *Work Motivation*. John Wiley, New York

H. Thierry

Peace

The major cases of violence and peace relate to the way the human condition is cut through by fault lines, dividing humans and nature, normal people from deviants, different genders, generations, races, classes, nations, states. Each border defines at least two categories, self and other. The violence can be direct violence intended by the self to attack the basic needs of the other; or structural violence, also preventable, not intended, usually slower, but at least equally devastating. If the other is dehumanized to a category, we end up with genocide—massive category killing. With two types of violence and eight fault lines we get a 16-fold discourse, as shown in Table 1.

Most of the time and in most places there is 'narrow peace' or an absence of direct violence across most fault lines, but not 'broad peace' or an absence of direct and structural violence. Different cultures view peace differently as descriptive (how is peace possible) or prescriptive (how could peace be possible).

The following is a review of the theories of peace of six macrocultures, or civilizations. Of the six, three are from the Occident, defined here as the cultural space spanned by the Abrahamitic religions of the Book, the *kitab*, or Old Testament: Judaism, Christianity, Islam. We start with two Europes, secular Europe—the Europe of antiquity—and then Christian Europe. We then move East to West Asia (I have not designated this as the 'Middle East,' which is a Western European perspective) with Judaism, Islam, and the cradle of Christianity. Continuing eastward we come to South Asia, with what is conveniently called 'Hinduism' and its offspring, Jainism and Buddhism. And further east, in East Asia there are the Chinese and Japanese cultural amalgams with a Confucian core, Buddhism, and then Taoism in China and Shinto in Japan. We

Table 1
Classification of violence

No.	Fault lines	Direct violence	Structural violence
[1]	Nature	slaughter of 'beasts, savages, wilderness'	depletion pollution
[2]	Gender	killing women: abortion, infanticide, witch-burning	patriarchy as prison of women, putting them 'in their place'
[3]	Generation	*Privileging* abortion euthanasia	*middle-aged* schools as ghetto 'homes' as ghetto
[4]	Deviance	*Controlling dangerous persons*	
	—criminal	capital punishment	institution
	—mental	euthanasia	institution
	—somatic	euthanasia	institution
[5]	Race	*Controlling dangerous races*	
		eradication	colonialism
		slavery	slavery
[6]	Class	*Controlling dangerous classes*	
	—military	elimination	exploitation-body
	—economic		repression-mind
	—political		alienation-spirit
	—cultural		
[7]	Nation	'genocide' as narrowly defined	the state as a prison of nations run by 'majority'
	Culture		
	Ideology		
[8]	State	war (killing for food, sacrifice, conquest)	imperialism isolating 'pariah' states
	Country		
	Territory		

then move into Pacific American African spaces, picking up a Polynesian, an Amerindian, and an African tradition, ending south of Europe.

1. Europe: Pax and Eirene—The Roman/Greek and Modern Traditions

Peace can be interpreted narrowly as *absentia belli*, the absence of organized violence between groups defined by the fault lines. International or external peace is the absence of external wars: intercountry, interstate, or international (in the sense of intercultural). Social or internal peace is the absence of internal wars: national, racial, class, or ideological groups challenging central governance or each other. This concept is carried by the Roman *pax*, related to pact, as in *pacta sunt servanda*, 'treaties must be observed.' Peace as a contractual, mutually agreed relation is the source of Western international law.

Another Roman legacy, *si vis pacem*, *para bellum*, 'if you want peace prepare for war,' is the source of Western military peace theory. Peace is obtained by the balance of power, deterring the aggressor at home with defensive defense, and abroad with offensive defense. Offensive offense (attack/aggression) and defensive offense (pre-emptive warfare) are not peace concepts.

Aggression is controlled from above by pacts—the Roman Empire was capable of enforcing such pacts—and/or by the balance of power. A better word for this kind of peace is security. And there is a basic dilemma: the military capability used to deter can also be used to attack, even if the motivation is defensive. The result can be an arms race, a cold war, or even an actual war.

A basic problem with the Roman *pax* is its insensitivity to flagrant injustice and inequality, the Roman Empire itself being an example. The Greek *eirene* is closer to 'peace with justice': absence of direct and structural violence, among Greeks. But which part has priority? What if they think justice can only be obtained through war? That idea was picked up by the Marxist tradition as class war, national and international, legitimized as necessary and sufficient to obtain a just, socialist society. The Roman thinking led to the liberal tradition, tolerant of enormous inequalities, nationally and internationally, but strongly against war, both internally and externally. And Roman warfare was picked up by the conservative tradition, extolling still more inequality through

wars, provided they could be won, and even through empire-building. In the year 2000 these views still prevail.

2. *Europe: Agape and Bellum Iustum: The Christian Traditions*

The three main divisions of Christianity (Orthodoxy/Roman Catholicism/Protestantism) and the many smaller sects had Paradise and Hell as archetypes that became models for peace and war, making peace remote and static, and war a highly dynamic hell on earth. Peace is seen as ordained by a God whose law is the only valid law, and valid for all of humanity. But who is God, and how does he relate to human beings in search of peace?

A theological distinction is very useful here, between God as immanent, inside us, making us godlike, sacred, and God as transcendent, above us, saving, choosing some persons and peoples, rejecting and condemning others. We may even talk about soft and hard Christianity, depending on which God-concept is picked up, in what proportion. They do not exclude each other.

Agape (Greek for 'love') can be used as a name for the peace of a soft Christianity based on an immanent conception of God. There is God's love for humankind, through Jesus Christ; the human love for God; and the love of one's fellow beings as being God-loved. The Lord's Supper and the Eucharist are close to this concept: a community of humans, God-enlightened, *in* God. The face-to-face *corpus mysticum* is based on an identity with others so strong that there is a limit to the number of members. This may be a reason why very egalitarian, interactive, direct peace concepts spring out of smaller Christian sects, like the Quakers and the Mennonites.

Then there is the transcendent God, conceived as some male person residing above our planet, and his relation to nonbelievers, pagans, and worse still, to the heretics who have rejected God. This is where hard Christianity enters, administering Hell and torture (the Inquisition) and Holy War to the heretics.

As some kind of peace concept it takes the form of *bellum iustum*, in the Augustine-Aquinas just war tradition:

(a) Wars must be waged by a lawful authority.

(b) Wars must be waged for a just cause, to correct injustice.

(c) Wars must be waged with the right intention, not vengefully.

(d) Wars must be waged as a last resort and with prospects of success.

(e) Wars must be waged with proportionality and minimum force.

(f) Wars must be waged only against combatants.

The first five are *ius ad bellum*; the last two are *ius in bello*.

This can be seen as an effort to limit war even if accepting war as the last resort. But there is no nonviolent alternative, and it may be used to attack any kind of injustice however defined. In addition, it is not biblical but derives from St. Augustine and St. Thomas Aquinas. But this cost-benefit thinking is still with us.

3. *West Asia: Shalom/Sala'am—The Judaic and Islamic Traditions*

Looking at some famous quotes from the basic texts (see Table 2, Christianity is included as a family member for comparison) one conclusion may be that ambiguity is the message. It is neither unconditional peace, nor unconditional war. It is peace under certain conditions and war under certain conditions. The problem is to spell out those conditions. One reading of the Judaic *shalom* and the Arab *sala'am* is peace with justice. Without justice, no peace; hence war for justice is legitimate. The contradiction is mirrored in the quotes shown in Table 2.

If we define justice as absence of structural violence whereby people, nations, and states are repressed, exploited, alienated, but not by an actor, then this may lead to *bellum iustum* as injustice abounds. *Jihad*, however, translates as 'exertion' for the faith. Defending the faith by violence, against Crusades, Zionism, communism, is the fourth stage of *Jihad*.

4. *South Asia: Shanti and Ahimsa—The Hindu/Jainist/Buddhist Traditions*

There is a trend in all three faiths toward unconditional peace by *ahimsa*, nonviolence. The major carrier of this message of all times, Mohandas K. Gandhi (1869–1948) used the formula 'There is no way to peace, peace is the way.' This is a strong stand, ruling out violence as immoral and unproductive. The struggle against violence, direct or structural, is by nonviolence. And, as becomes clear from Gandhi's adulation of the small social unit, the village, peace in the sense of absence of direct violence, cannot be built top-down by heavy international and national hierarchies, in the *pax* and Hobbesian traditions.

This, in turn, is related to *shanti*, inner peace. *Ahimsa* has *shanti* as a necessary condition: with no inner peace, no nonviolence. 'Unprocessed traumas' will be acted out aggressively. If nonviolence does not lead to change of heart in the other, it is for lack of change of heart in self. Nonviolence then turns into self-purification, practiced in the little community of believers, the *sangha*, which is like a monastery.

Gandhi left behind a theory and practice of *satyagraha*, clinging to truth as he called it. Look at the list of nonviolent action: to play a major role in delivering the nonwhite world from white colonialism,

Table 2
Hard and soft peace: quotations from Judaism, Christianity, and Islam

	Judaism	Christianity	Islam
Soft peace	He will decide the disputes of the nations and settle many a people's case… till swords are beaten into plough-shares and spears into pruninghooks. No nation draws the sword against another. (Isaiah 2:4)	How blest are the peacemakers. God shall call them his sons. (Matt. 5:9)	And whoever saved a human life, should be regarded as though he had saved all mankind. (Qur'an 5:32)
Hard peace	For in the cities within the boundaries of the Promised Land you are to save no one; destroy every living thing. (Deut. 21:16)	Don't imagine that I came to bring peace to the earth. No, rather, a sword. (Matt. 10:34)	Fight for the sake of Allah those that fight against you but do not attack them first. (Qur'an 2:190)

and the white world from its own Cold War, leaving no thirst for bloody revenge behind, is no minor achievement. That gift comes out of South Asia, not the West. The contribution of the West seems to a large extent to have been to create the problems. The burden of proof is on those who teach peace by violence. To assert 'nonviolence does not work' is uninformed, given the amazing successes in the second half of the twentieth century:

(a) the liberation of arrested Jews in Berlin, February 1943

(b) Gandhi's *swaraj* campaign in India; Independence from 1947

(c) Martin Luther King Jr.'s campaign in the US South from 1956

(d) the anti-Vietnam war movement, inside and outside Vietnam

(e) the Buenos Aires Plaza de Mayo mothers against the military

(f) the 'People's Power' movement in the Philippines, 1986

(g) the Children's Power movement in South Africa, from 1986

(h) the *intifada* movement in Occupied Palestine, from 1987, in the beginning mostly nonviolent

(i) the democracy movement Beijing, spring 1989

(j) the Solidarity/DDR movements which ended the Cold War.

5. *East Asia: Ho p'ing/Heiwa—The Chinese/Japanese Traditions*

The standard translation of *ho p'ing* (Chinese) and *heiwa* (Japanese) is 'harmony.' And one reading of harmony would be not only 'absence of violence,' but 'absence of conflict.' The task of conflict resolution has already been carried out. Indications of conflict are swept under the carpet, and the person articulating conflict is frozen out of harmonious society, or prevented from expressing such views.

Take the metaphor of Chinese boxes, or in Russia the *matrushka* dolls, with one box or doll inside the other and so on, till they become extremely small. They all look alike. In modern fractal/chaos theory they talk about 'self-similarity' as something stabilizing. The basic East Asian point is that harmony is produced not by a particular structure, be that a pyramid or wheel or whatever, but by the same structure repeated at the personal, social, and world levels, within and between.

In *Tao Te Ching* we find an example: small, detached countries: 'Though they have armor and weapons nobody displays them—they leave each other in peace while they grow old and die' (No. 80). A modern version would be domestic and global democracy.

6. *Pacific, America, Africa—Ho'o ponopono, Peace Pipe, and Shir*

The following are three examples of peace as nonviolent conflict resolution.

6.2 *Ho'o ponopono—The Polynesian Conflict Resolution Circle*

Ho'o ponopono (setting straight) is a method of confict resolution that brings together around a table per-

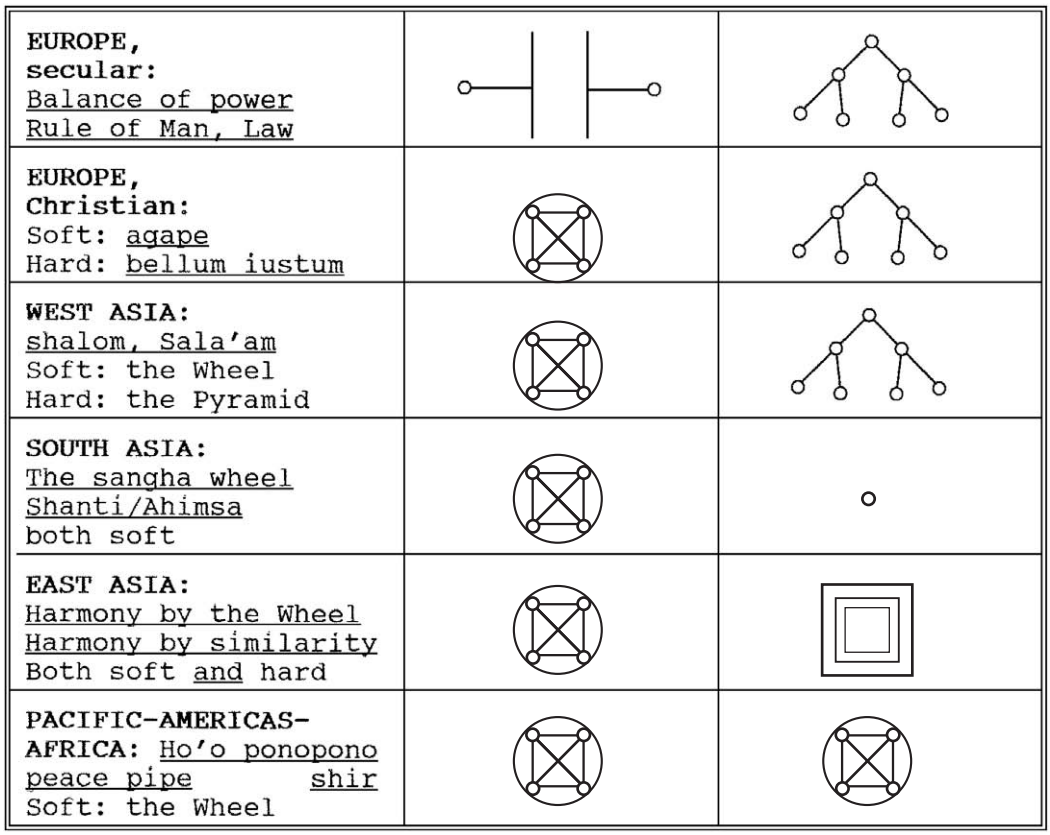

Figure 1
Geometric images of peace

petrator, victim, family members, neighbors, friends, and others, with a moderator, the 'wise man,' not from families or neighbors. The process of resolution has four stages.

Everyone is encouraged sincerely to present their version of why it happened, how, and what would be the appropriate reaction, and everybody accepts some part of the responsibility. Sincere apologies are then offered and accepted; forgiveness is demanded and offered. An action program is defined for everybody according to the principle of shared responsibility. And in the end the record of what happened (but not of what has to be done) is then burnt, symbolizing the end of the incident, and the construction of a new relation among all of them.

6.3 The Peace Pipe—The Cheyenne Conflict Resolution Symbol

With the Cheyenne, a zone of peace is created at the top of society, with the peace chiefs being models of nonviolence to be emulated by the rest. They were not to engage in any quarrels within the tribe regardless of whether their families or children were involved. They were not to engage themselves in any force or violence, even if their son was killed right in front of their tepee.You are to do nothing but take your pipe and smoke. Being a chief becomes a way of life, and the chief's home a sanctuary where others can be safe.The chiefs mediate disputes and don't take sides. Ritual, like smoking the peace pipe together, makes it possible to think with one mind, one heart. Everyone has a right to talk, nobody is interrupted, the talk continues till the end.

6.4 Shir—The Somalian Conflict Resolution Market

A traditional conflict resolution structure that brings together all the mature men in the clans involved in a conflict. Women, children and young hot-blooded warriors are excluded. Men lounge under the thorn trees during the hot, dry day. They chat and drink tea.

At some point, things will jell. The various pieces that make up the main issue for which the *shir* was called will fall into place because a social climate conducive to a solution will have slowly emerged. The result will be proper peace—a peace felt from the inside.

Conflict resolution has here been hitched on to one of the oldest institutions in the history of human beings: the market. The market is based on exchange for mutual benefit: I give, you give (*do ut des*, I give so that you give), and some principle of equal exchange, (*quid pro quo*).

7. Peace as Geometry

Five geometric archetypes for peace are shown in Figure 1: the vertical, big pyramid for the hierarchic rule from above, of human beings or of law; the wheel for the smaller, tight, equal exchange; the symmetry—like the garden of Versailles—for the balance of power; the sun for the nonviolence radiating from inner peace; and the Chinese boxes of harmony based on self-similarity, from micro to macro levels.

Different kinds of peace can also be expressed in terms better known from the history of diplomacy, focusing on interstate and internation peace: the Napoleonic peace forged by a prince/lawmaker; the silent peace brought about by groups of ordinary humans, or countries; the Nixon–Brezhnev peace negotiated at a bilateral summit; the Gandhian peace as process, not only outcome; the anonymous peace as built-in harmony. All of these are used for conflict resolution. How conflict is handled is the best peace indicator.

8. Approaches To Peace

Each society, community, family, and person has ways of being peaceful by peaceful means, but they also often have ways of being violent, summarized in their cultures. A major approach to peace concerns is to tilt the cultural balance in favor of peace, letting peace cultures be the dominant cultures. Then, using the survey of peace cultures and the peace as geometry approach we have five ideas in addition to peace culture.

The two Europes and West Asia have one idea in common: the hierarchy, the pyramid. Essentially this is peace from above, whether from a big power, a combination of big powers, or from some central authority. The current term is 'global governance,' based on global architecture. Regional versions may produce regional peace among members, excluding nonmembers. The counterpoint to the pyramid, the wheel, found in five cultural spaces, expresses the ideas of peace from below, small is beautiful, inspiring the search for alternative peace structures. Then comes the *shanti–ahimsa* combination from South Asia: the source of peace is inside you, its expression is nonviolence, which makes war abolition a realistic utopia. And we pick up the idea of conflict transformation, with East Asian harmony as outcome, *ho'o ponopono*, peace pipe, and *shir* as examples of processes.

The six approaches divide into two groups: global governance, war abolition, and conflict transformation being more familiar; nonviolence, peace structures, and peace cultures unfamiliar and challenging. Peace in the complex, highly interconnected world of the twenty-first century, bringing conflict parties very close to each other, will depend on the ability to draw upon all six—not easy given that they hail from different parts of the world.

See also: Alliances: Political; Balance of Power, History of; Balance of Power: Political; Conflict Sociology; Diplomacy; International Relations, History of; National Security Studies and War Potential of Nations; Peace and Nonviolence: Anthropological Aspects; Peace Movements; Peace Movements, History of; Religion: Peace, War, and Violence; Violence, History of; Violence: Public; War, Sociology of; Warfare in History

Bibliography

Bouquet A C, Murty K S 1960 *Studies in the Problems of Peace*. Asia Publishing House, Bombay, India

Ehrenreich B 1997 *Blood Rites: Origins and History of the Passions of War*. Metropolitan Books, New York

Galtung J 1980 Social cosmology and the concept of peace. In: Galtung J (ed.) *Essays in Peace Research*. Ejlers, Copenhagen, Vol. 5, pp. 415–36

Galtung J 1996 *Peace By Peaceful Means*. Sage, London

Galtung J 1998 *Conflict Transformation by Peaceful Means*. United Nations, Geneva

Giorgi P P 1999 *The Origins of Violence by Cultural Evolution*. Minerva E&S, Brisbane, Qld, Australia

Howell S, Willis R (eds.) 1989 *Societies at Peace*. Routledge, London

Ishida T 1969 Beyond the traditional concepts of peace in different cultures. *Journal of Peace Research* **6**: 133–45

Keeley L H 1996 *War Before Civilization*. Oxford University Press, Oxford

Prunier G 1998 Somaliland goes it alone. *Current History* May: 225–8

Shook E V 1985 *Ho'o ponopono*. East-West Center, Honolulu

Smith-Christopher D L 1998 Indigenous traditions of peace: An interview with Lawrence Hart, Cheyenne Peace Chief. In: Smith-Christopher D (ed.) *Subverting Hatred: The Challenge of Nonviolence in Religious Traditions*. Boston Research Centre for the 21st Century, Cambridge, MA pp. 85–94

Thompson W S, Jensen K M (eds.) 1991 *Approaches to Peace: An Intellectual Map*. United States Institute of Peace, Washington DC

J. Galtung

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Peace and Nonviolence: Anthropological Aspects

1. Definitions

The phrase 'peace studies' refers generically to studies of the dynamics of 'peace,' and specifically to the natural history of 'peaceable' or 'nonviolent' peoples. The two main foci, which overlap significantly, are (a) the analysis of particular conflict-resolution techniques, and (b) the holistic study of peaceable peoples.

1.1 Problems of Definition

The terms 'peace' and 'nonviolence' are relative, not absolute, referring to adaptations which are often transitory rather than to essential natures which persist indefinitely. Although researchers use the words 'nonviolence' and 'peace' as rough synonyms, the former creates methodological problems. How to observe and analyse the absence of violence? It might be useful to reserve this term for ideals such as Gandhi's *ahimsa* or related techniques such as the nonviolent protest of the 1970s 'Peace Movement' of Europe and America, from which peace studies arose and borrowed the term.

Likewise, in most Indo-European languages 'peace' is a null category. When you have no violence, you have peace. Linguistically, violence is the norm. Peaceable peoples are 'nonviolent' (Ashley Montagu) or 'low-conflict' (Marc Ross). It is difficult for students of peace to avoid drifting into the null construction of peace which pervades the cultures within which they live their lives.

By contrast, Semai, a peaceable Malaysian people, use the term *slamad* to refer to a condition of security and serenity which they nurture with the same zeal that Indo-Europeans pursue wealth and happiness. In Semai construals, *slamad* is normal, 'non-*slamad*' the null category. Recent peace studies stress the distinction between 'negative peace' (refraining from violence out of fear) and 'positive peace' (maintaining *slamad* out of love of peace). In this construction, both forms of peace are activities rather than absences of activity, involving conscious and unconscious 'peacemaking' rather than voids. Defeat involves an agent's act of surrender as well as another agent's act of conquest. 'Peacemaking' is thus the category of behaviors and social patterns on which peace studies focus.

2. Why Not Study Peace?

2.1 The Ubiquity of Peace

Any survey of the literature will uncover hundreds of articles on violence for every one on peace. Yet the most violent people spend most of their time in humdrum peaceful activities: Eating, sleeping, talking, daydreaming, scratching, and otherwise doing no harm. Most people in the most violent societies, even in wartime, spend their days in peaceful pursuits (Nordstrom 1997). The life of a soldier in wartime involves more hours of boredom than moments of terror.

No wonder. Violence hurts, and individuals dislike getting hurt. The possibility of counter-violence makes violence self-limiting in a way that peace is not. People prefer hurting those who, for whatever reason, cannot retaliate. Domestic violence is probably the commonest form of violence cross-culturally, simply because perpetrators are relatively safe from retaliation (Gelles and Straus 1988). Even then, when local values justify violence, the way human perpetrators obfuscate their actions by appealing to peace ideology reveals their uneasiness with violence: they are 'maintaining law-and-order,' 'keeping the peace,' 'disciplining students,' or 'teaching them a lesson.'

In evolutionary terms, that makes sense. Other things being equal, organisms that avoid serious fighting with peers are more likely to survive long enough to produce fertile offspring than those that usually win fights but risk counter-violence; i.e., peaceful organisms are fitter. Moreover, violence disrupts cooperative activities that enhance the fitness of social organisms like humans. Such disruption is a common theme in nonWestern peoples' reconciliation ceremonies: if we fight, who will help to harvest the crops? Fear of disruptive violence rationalizes human rulers' stamping out or severely limiting freelance violence, and monopolizing violence for themselves and their state. As a result, throughout human history peace prevails, in most places, most of the time. Peace is normal and normative. Violence is abnormal and disruptive.

2.2 Reasons Not to Study Peace

So why does peace get so little attention, even from social scientists? One reason is the definition of peace as an absence, discussed above. There seem to be several others. One probable reason is evolutionary: Violence threatens fitness. Successful organisms attend to threats. Violence is a problem; peace is unproblematic. So peace remains in the background, and violence is dramatic. Unsurprisingly, the most popular account of paleoanthropic violence is by a dramatist, not a paleoanthropologist, primatologist or prehistorian (Ardrey 1961). The very prevalence of peace makes violence salient: The exception, 'man bites dog,' makes news. Normal conditions do not.

Answers to the question of how to live in peace have potential political consequences, which might affect the lives not only of this generation but of generations to come. In the 1990s, over 2 million children died in wars, 6 million were seriously injured or permanently

disabled, a million orphaned, and 12 million made homeless. Children are the main victims of enslavement and domestic violence.

Awareness that research may be consequential can make it teleological, arriving at conclusions which reflect researchers' political predilections. Hobbes, an early student of violence and peace, deliberately excluded empirical considerations from his work, relying instead on an a priori method which he felt derived from Euclidian geometry, and concluding that the political despotism of his time was a counter-violence needed to control the freelance violence which would otherwise ruin human life. Similarly, peace studies evolved from New Left political activism in the 1970s. A large number of peace theorists are pacificists or members of quietist religious groups like Quakers or Mennonites. Many are explicitly political, arguing that, since they feel peace requires respect for human rights and social justice, anthropology should espouse such causes. Pedagogy is a central concern.

The resulting atmosphere may alienate students who hold traditional social science notions of 'objectivity.' Moreover, this sort of partisanship generates equally political responses by students of violence and others for whom, as Max Weber warned might happen, social science is the handmaiden of the Establishment and the social controls that maintain it in power. Social justice, say the reformers, will produce (positive) peace. No, say the establishmentarians, sometimes marching under the banner of objective social science, you need the hangman and the Bomb (negative peace).

Such politicization makes researching these topics difficult: For example, anthropologists have been reluctant to undertake (and funding agencies to finance) studies on domestic abuse among indigenous peoples, at least partly because the hegemonic liberal ideology frames indigenous peoples as simple and peaceful, and representing them otherwise might exacerbate the difficulties they already face in finding allies among more powerful peoples. In rebuttal, opponents seem to be reviving the nineteenth-century hegemonic representation of indigenous peoples as savage and violent, an equally political representation (e.g., Keeley 1996).

It may be impossible and even undesirable to eliminate political concerns from research that is so 'sensitive.' It would be helpful to make them explicit, as students of peace traditionally do. But scholars may be unaware of the teleological concerns that bias their results. Even researchers, after all, swim in their own culture like fish, mostly unaware of how the medium that supports them also affects how they perceive the world.

A cognate difficulty is that discussions of peaceability often degenerate into speculations about 'human nature,' a Platonic essentialist concept which serves more to terminate research than to stimulate it (but cf. Boehm 2000). If people are 'naturally' violent, violence needs no explanation; if 'naturally' peaceable, peace requires no further examination. Recent studies such as Keeley's (1996) suggest that violence has been part of human societies since early in prehistory. There may be no society in which violence has never occurred. By the same token, there is no society in which peace never occurs, in which the general rule is a Hobbesian war of each against all; and no period in prehistory or history at which no one lived in peace. One of the most interesting studies of war describes how, in the midst of all-pervasive violence, people continue doggedly to construct peace (Nordstrom 1997). The half-empty glass is half-full (Boehm 2000), and the decision on the question of human nature is more likely to reflect the political sensibilities of the researcher than the empirical data.

3. *Styles in the Study of Peace*

The two main foci of peace studies reflect the dichotomy between positive and negative notions of peace. Although the sketch given in the previous section of 'establishmentarians' versus 'reformers' is a heuristic oversimplification of impulses which overlap and may coexist within the same person, nevertheless it is not completely inaccurate to see the establishmentarian impulse in the first focus and the reformist in the second.

3.1 *Conflict Resolution Studies*

The first subfield is, at least implicitly, instrumental or 'applied' anthropology, reflecting governmental interests in international diplomacy and administrative measures to reduce freelance violence, e.g., in the school system. It comprises 'conflict-resolution' or 'mediation' studies of the sort emphasized by the journal *Peace and Conflict Studies*—www.gmu.edu/academic/pcs. In the United States, George Mason University is a center for such studies. Since World War II, conflict resolution studies in anthropology have grown in tandem with their growth in psychology, sociology, and political science. This upsurge seems to be connected with the rise of UN 'peacekeeping' efforts and interventions by the USA and Western Europe in overseas conflicts. More recently, the perceived rise of violence in the USA has led to funding of 'violence prevention programs,' for example, by the Justice Department. Most such programs stress peaceful conflict resolution, sometimes purportedly modeled on 'primitive customs.'

Most conflict resolution theory is a subset of game theory models, in which participants try to maximize benefits to themselves, and minimize their losses. The ideal solution is 'win–win,' in which for each participant the sum of the benefits outweighs the sum of the losses. Ideally, participants arrive at a 'rational'

resolution through reasonable (ideally numerate) discussion, perhaps with the help of mediators from outside, trusted by both parties, experienced in mediation, and dedicated to resolving conflicts peacefully. The 'anthropological' character of conflict resolution studies comes mostly from the attempt, in the face of skepticism from other disciplines, to deploy the concept of culture, and from the occasional choice of 'tribal' peoples as subjects. In general, however, the concerns and analytical techniques (e.g., cost–benefit analysis) are indistinguishable from those of the other social sciences.

Of course, as 'conflict theorists' have insisted since Georg Simmel, conflict is ubiquitous and need not produce violence. Conversely, conflict resolution is neither necessary nor sufficient to produce peace (Fry and Bjorkqvist 1997). Many nonWestern peoples traditionally held interminable meetings explicitly to resolve conflict. During these meetings, everyone got to speak at length until participants were physically and emotionally exhausted, and no one had any unvoiced opinions. The conflict might remain, but angry emotions dissipated and the conferees could reconcile or agree to differ.

Another practical limitation is that most conflict resolution techniques require that participants be equals or equally subject to coercion by a power greater than they are. But, as cost–benefit analysts of domestic violence point out, violence is most likely to occur when unpleasant consequences for the perpetrator are negligible, i.e., when the parties are unequal (Gelles and Straus 1988). Thucydides records a conflict resolution conference between Athenians and Melians. Since the more numerous and better-armed Athenians could (and later did) overwhelm the Melians, the Melian conference 'failed.'

The corollary is that egalitarian societies tend to be peaceable, and peaceable ones tend to be egalitarian (Fry 1999); between nations, 'mutual deterrence' depends on perceived equality. Perhaps social justice in the form of felt equality is also important. Thus, although conflict resolution theorists tend to dismiss other tactics for preventing violence as unreliable (Fry and Bjorkqvist 1997), conflict resolution techniques are most likely to succeed when least likely to be necessary.

3.2 Peace Studies

Peace studies are more ethnographic, less technical and more value-laden, finding expression in, for example, *Human Peace and Human Rights*, journal of the Commission on the Study of Peace and Human Rights of the International Union of Anthropological and Ethnological Sciences—rar@syr.edu. Syracuse University is a center for such studies as well as for conflict resolution studies in the USA. This research rests on the empirical observation that conflicts among some aggregations of people are less likely to erupt into violence than among others. One reason for such peace is a value system that promotes negative peace by abhorring violence (e.g., as stupid, scary, or self-destructive) and promotes positive peace or *slamad*, a value system somewhat like that of the 1970s 'Peace Movement' which inspired peace studies.

But that movement itself grew from the Vietnam War. Peace tends to be valued particularly by people who know violence firsthand. Peaceable values may maintain peace, in the short run, but they gain strength from the threat of violence (DeBenedetti 1978, Dentan in Silverberg and Gray 1992, in Sponsel and Gregor 1994, Nordstrom 1997). And one reason for war is that outsiders threaten *slamad*. In short, attributing peace only to peaceful values is as simplistic as attributing violence to 'innate aggression.' The stipulated variables are insufficient to produce the observed results.

The ethnology of peaceable peoples reveals how complicated is 'peace.' Documentation exists on a fairly large number of relatively peaceable societies, including traditional ethnic groups (e.g., Dentan in Silverberg and Gray 1992, Fry 1999, Howell and Willis 1989), cenobitic pacifists like the Amish or Hutterites, and other voluntary associations such as the Friends, the Rainbow Family, or Alcoholics Anonymous (AA) (1994). Most have formal and informal conflict-resolution techniques, the latter derived from those characteristic of the species as a whole (Boehm 2000, de Waal 1989). Most lack the in-group economic or political hierarchies which impede consensual conflict-resolution, although many are enclaves in hierarchical societies against whom they must maintain identity barriers.

Peaceable values are conducive to peace. Many peaceable peoples like peace better and violence less than more violent peoples do. But many, for example, the Amish and Mbuti, beat children severely. Others, such as the Semai, fantasize warfare against their oppressors, or boast, like the AA, that members can stand up to the challenges of warfare when their country calls. Generally, in-group peace seems more valued than peace with outsiders. And, as the ubiquity of peace may indicate, most people prefer peace to violence except under exceptional circumstances.

'Negative peace' based on fear of consequences seems more salient than 'positive peace' based on idealist values. Many peaceable peoples, particularly cenobitic pacifists, are patriarchal and physically punish disobedience by children. Oppressive regimes can also diminish freelance violence by retaliatory violence such as incarceration, mutilation, or capital punishment.

Avoiding others is yet another way of peacemaking. Cross-culturally, inevitable structural conflicts, such as that between a man and his in-laws over who benefits from his wife's services, may result in formalized respect/avoidance relationships in which one or

the other party expresses extreme deference to, or even avoids social contact with, the other. Avoiding a person with whom one is angry is an effective passive-aggressive communication of the anger and dramatizes the threat of withdrawing cooperation. Many peaceable societies of Southeast Asia arose in the context of a poltical economy based on slavery and coerced trade, in which the best way to avoid enslavement was to flee. Catholics and Protestants in Northern Ireland, like 'blacks' and 'whites' in the USA, tried less successfully to minimize violence by minimizing mutual contact.

Another way of dealing with inherent conflict is to transform physical violence into 'harmless' symbolic violence. To some extent, sports in the Western world do provide the 'moral equivalent of war' that William James imagined. The 'joking relationships' that occur in many societies allow people to express anxiety and frustration about possible conflicts, especially involving sex, by hostile words and deeds which local ideology defines as 'not serious.' This technique is unreliable: fans battle each other and 'locker room humor' can erupt into violence.

3.3 The Future of Peace Studies

What is needed is meticulous field studies of peoples who seem better at peacemaking than others. Such studies require avoiding essentialist notions of peaceability as a psychosocial deficiency, the inability to be violent under any circumstances. Rethinking definitions of peace and violence seems appropriate (e.g., Dentan 2000). If peaceability is an adaptation rather than an essence or changeless value, then changing circumstances should affect peaceability, so that such studies must take account of history in order to discover what variables are important. Perhaps the most promising future studies are those which make detailed comparisons between (a) peacemaking among specific human groups and among specific other primates (e.g., Cords and Killen 1998), or (b) 'controlled comparisons' between particular peaceable peoples and otherwise similar but more violent peoples, e.g. the studies of peaceable and violent Zapotec towns by O'Nell and Fry (see Fry 1999); the Robarcheks' planned book-length expansion of their chapter on peaceable and violent tropical rainforest swiddeners, the Semai and Waorani (Robarchek and Robarchek 1992); the comparison of peaceable and violent Chicago neighborhoods by Sampson et al. (1997).

3.4 The Use of Peace Studies

Improving conflict resolution techniques is of obvious importance. Are any lessons from peaceable societies of possible use to other peoples? The great student of primate conflict resolution, Frans de Waal dismisses peaceable human groups as a 'few gentle, nonmartial human societies that have managed to survive in remote corners of the world' (1989, p. 4). But many cruel martial human societies have also perished, because their cruelty and violence was as unsustainable as is peaceability as de Waal implies. The British poet Shelley's poem *Ozymandias* says it best: There is an immense broken statue in the poem, on whose base are the words: 'My name is Ozymandias, King of Kings. Look on my works, ye mighty, and despair!' All around the statue, as far as the eye can see, the desert lies unbroken. Scientists must not let the drama of violent conquest obscure the fact that organisms and societies often survive better by eschewing, preventing, or avoiding violent conflict than by engaging in it. As a long-term adaptation, neither peaceability nor violence is necessarily or essentially 'good' for survival; in the short term, either may work, depending on the circumstances. That is why peaceable people like the Semai may stop being peaceable, and violent people like the Waorani may turn to peace. Both fight and flight are active adaptations, which couple with caring for children and forming alliances ('tend and befriend') as human responses to environmental stress. The dynamics of these adaptations need the same careful study as any other ecological adaptation.

See also: Domestic Violence: Sociological Perspectives; Peace; Peace Movements; Peacemaking in History; Violence, History of; Violence in Anthropology; Violence: Public; War: Causes and Patterns; War, Sociology of

Bibliography

Ardrey R 1961 *African Genesis: A Personal Investigation into the Animal Origins and Nature of Man*. Dell, New York

Boehm C 2000 Conflict and the evolution of social control. *Journal of Consciousness Studies* **7**: 79–101, 149–183

Cords M, Killen M 1998 Conflict resolution in human and nonhuman primates. In: Langer J, Killen M (eds.) *Piaget, Evolution and Development*. Erlbaum, Mahwah, NJ

DeBenedetti C 1978 *Origins of the Modern American Peace Movement, 1915–1929*. KTO Press, Millwood, NY

Dentan R K 2000 Ceremonies of innocence and the lineaments of unsatisfied desire. *Bijdragen tot de Taal-, Land- en Volkenkunde* **156**: 193–232

de Waal F 1989 *Peacemaking Among Primates*. Harvard University Press, Cambridge, MA

Fry D P 1999 Peaceful societies. In: Kurtz L R (ed.) *Encyclopedia of Violence, Peace, and Conflict*. Academic Press, San Diego, CA

Fry D P, Bjorkqvist K (eds.) 1997 *Cultural Variation in Conflict Resolution: Exploring Alternatives to Violence*. Erlbaum, Mahwah, NJ

Gelles R, Straus M 1988 *Intimate Violence*. Simon & Schuster, New York

Howell S, Willis R (eds.) 1989 *Societies at Peace: Anthropological Perspectives*. Routledge & Kegan Paul, London

Keeley L H 1996 *War Before Civilization: The Myth of the Peaceful Savage*. Oxford University Press, New York
Montagu A (ed.) 1978 *Learning Nonaggression: The Experience of Non-literate Societies*. Oxford University Press, London
Nordstrom C 1997 *A Different Kind of War Story*. University of Pennsylvania Press, Philadelphia, PA
Robarchek C A, Robarchek C J 1992 Cultures of war and peace. In: Silverberg J, Gray J P (eds.) *Aggression and Peacefulness in Humans and Other Primates*. Oxford University Press, New York
Sampson R J, Raudenbush S W, Earls F 1997 Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* **227**: 918–24
Silverberg J, Gray J P (eds.) 1992 *Aggression and Peacefulness in Humans and Other Primates*. Oxford University Press, New York
Sponsel L E, Gregor T (eds.) 1994 *The Anthropology of Peace and Nonviolence*. Lynne Rienner, Boulder, CO

R. K. Dentan

Peace Movements

'Peace Movements' may be used in two ways. On the one hand, a peace movement is a specific coalition of peace organizations that, together with elements of the public, seek to remove a threat of war or to create institutions and cultures that obviate recourse to violence. On the other, it is the organizational infrastructure to do so. Usage is usually clarified by context. Peace organization constituencies are people with shared commitment to common values and traditions, like religious pacifism, or to a program such as world federalism. Such groups form coalitions in order to enlist public support in response to salient issues. If the issue is war or a specific war threat, peace coalitions take the form of antiwar movements. In nearly 200 years of organized peace effort, specific peace movements have affected national policies, international institutions, and popular attitudes. Taken as a whole, they can be viewed as a single, evolving, and increasingly transnational social movement that has interacted with formal analyses of war, peace, and social movements. The conceptualization of peace movements has resulted from the dialectical interaction of the movement's self-reflection on its experience and subsequent scholarly analysis.

1. Growth and Self-definition, 1815–1939

In western civilization peace has been understood mainly as the absence of war (Chatfield and Ilukhina 1994). Given war's existence, there were two main ethical alternatives: a dominant 'just war' tradition that legitimated organized violence under specific conditions, and a minority tradition of individual Christian nonresistance that rejected violence altogether. There was no sustained tradition of organized popular effort to check interstate warfare.

That was the situation when small peace societies were formed in the United States and the United Kingdom (1815) by a few local nonconformist Protestants who denounced warfare. In the next 100 years civic associations for societal change became common in western societies, providing the context in which peace organizations grew and defined themselves.

1.1 Early Internationalism

Peace constituencies broadened within an educated elite. British and American leaders solicited support on the Continent, where peace advocates tended to be secular intellectuals and where Richard Cobden's program of peace through free trade became very influential (Cooper 1991). During the 1840s leaders on both sides of the Atlantic promoted international arbitration. A few went further, a congress of nations, an international court, and even a united Europe. Then, in mid-century, peace advocates were shaken by wars in the Crimea, Germany, Italy, and America.

Gradually the movement was rebuilt in the latter third of the century. American and European leaders promoted treaties of arbitration, forming a program-oriented peace movement propelled by a bias for *internationalism*—the conviction that warfare is irredeemable and that statesmen could render it obsolete by breaking down barriers, building up international law, and seeking practical mutual interests beyond conflicts. The American Peace Society (1928) and the *Ligue internationale et permanante de la paix* (Paris 1863) popularized this approach. The *Ligue internationale de la paix et de la liberté* (Geneva, 1867), however, pressed the view that peace could be secured only by political justice among self-determining, democratic peoples. Concurrently, the First and Second (socialist) Internationals held that peace was contingent upon economic justice—the overthrow of capitalism. In practice, though, as socialists entered the political mainstream they supported programs of arbitration, arms limitation, and anticolonialism—peace as liberal internationalism, which was christened '*pacifism*' (Cooper 1991, p. 237, n. 1).

1.2 An Internationalist Peace Establishment

The movement for liberal internationalism obtained strong leverage from the First Hague Conference in 1899, when the great powers endorsed pacific alternatives to war. By then approximately 100 national and

regional societies were linked by an office in Berne, Switzerland (International Peace Bureau, 1891). Within another decade there emerged (mainly in America) several well-funded associations, notably the Carnegie Endowment for International Peace. They constituted a kind of *peace establishment*, an elite of business, legal, and governmental leaders who promoted the arbitration process. In World War I most of these internationalists proved to be patriots first. The word 'pacifism' became narrowed, especially in Anglo-American usage, to mean the absolute, individual rejection of war. Wartime peace advocates were repressed.

1.3 Peace and Internationalism Redefined

The war that shaped the twentieth-century world reshaped peace movements, which emerged more transnational and politically activist than before.

A few internationalists concluded that peace required some constraints on national sovereignty, that war is irrational, and peace requires 'a systematic effort to institutionalize managerial controls over the fragile interdependence of modern industrial civilization' (DeBenedetti 1978, p. 8). Such *liberal internationalism* was institutionalized in the League of Nations and its efforts for arms control, but also in an international network of civic League of Nations Associations. From wartime pacifism there emerged other transnational associations: the Women's International League for Peace and Freedom (1919), the pacifist International Fellowship of Reconciliation (1919), the linked British and US Quaker service committees (by 1917), and the International War Resisters League (1921).

At the same time, a new constituency of political reformers carried their campaign against corporate evil and systemic injustice into the peace movement. Women fresh from suffrage campaigns, for example, revitalized earlier analyses of war as a problem of social structure. Especially in England and America peace movements acquired a constituency of middle-class women and men who linked peace and justice and for whom peace movements were social change agents.

Following the war these activists informed and mobilized public opinion, engaging in the politics of pressure groups to challenge military spending and influence foreign policy. Peace advocates were vocal also in Germany and France, where they had been suppressed in wartime, and socialists throughout Europe took up the peace cause. Unfortunately, differences over priorities and policy hobbled peace efforts throughout the 1920s. The American movement was a sad case in point: groups promoting the League of Nations, the outlawry of war, disarmament, peace education, and international law failed to coalesce. Early in the next decade peace advocates organized internationally to support the League's 1932 disarmament conference, but in the face of aggression they soon realigned as rival supporters or opponents of collective security. No cohesive peace movement could be discerned among the bitter contests waged for public opinion.

2. External Assessment: Early Histories

The self-reflection of peace advocates on their experience affected the formal studies that began in the early 1930s with a few general histories. British scholar A. C. F. Beales, for example, interpreted *The History of Peace* (1931) optimistically in light of alternative programs broached when he wrote. A half decade later, Merle Curti's *Peace or War: The American Struggle 1636–1936* (1936) more soberly emphasized the gap between pacific ideals and the realities on which warfare is grounded (notably economic factors). Pacifist Devere Allen's earlier *The Fight for Peace* (1930) interpreted the historic American movement against his own norm of nonviolent activism and war resistance.

3. Nonviolence

Allen's book itself was part of the historic development of secular pacifism into an ideology of proactive, nonviolent change. That shift made the early *Satyagraha* campaigns of Mohandas Gandhi in India important to western contemporaries, and later to American civil rights leaders experimenting with nonviolent direct action and civil disobedience. Early sociological interpretations of *Satyagraha* were absorbed into the pacifist literature on nonviolence by the 1930s. Conversely, later experiments with the method in civil rights and antiwar campaigns expanded scholarly analyses of it. Thus nonviolent direct action developed dialectically along practical and theoretical lines (see Powers and Vogele 1997).

Moreover, the use of nonviolence in the struggle for justice applied the concept of peace as change without violence. Understanding peace as nonviolent social change for justice, rather than only as the absence of war, underlies the now common analytical distinction between positive and negative peace.

4. Cold War Mobilizations

During World War II, a fresh peace coalition formed around the United Nations (UN) ideal. It was grounded in liberal internationalism and it generated public support for a UN organization. Movement-

related nongovernmental organizations (NGOs) were responsible for including human rights under UN purview and for gaining consultative status in the organization. Even world federalism elicited much public support until it, along with initial efforts to contain nuclear arms, was undercut by the Cold War. By 1950 peace and liberal internationalism were victims of global polarization. The older peace organizations remained intact, but they lacked a salient issue around which to build any coalition.

4.1 The Nuclear Test Ban Movement

In the mid-1950s the issue of testing thermonuclear weapons on a massive scale aroused a fresh, transnational peace movement. It was grassroots based and was well organized on national levels. It made use of print and film media, mass demonstrations like the annual Aldermaston marches in England, political lobbying, and nonviolent civil disobedience such as sailing into test zones. The movement contributed significantly to a temporary moratorium on testing and to the 1963 Partial Test Ban Treaty.

4.2 The Antiwar Movement of the Vietnam War Era

Then the global movement lost coherence. The nuclear threat seemed to diminish and activists faced other issues, most dramatically the US war in Vietnam, which provoked worldwide opposition. The center of the storm was in the United States, where political radicals, pacifists, and liberal internationalists aligned in a loose, tenuous coalition. The whole repertoire of activism was employed: popular education, political pressure and lobbying, electoral politics, and street demonstrations. Although identified with radical counterculture in public opinion, opposition to the war increasingly enlisted mainstream activists and infused congressional politics. Extrication from Vietnam coincided with an exhausted peace movement and apparent détente in the Cold War.

Détente collapsed at the end of the 1970s. In response to a heightened nuclear arms threat, a new peace coalition arose that was transnational, massive, grassroots based, and aligned with environmentalists (most fully the Greens in Germany). Internationally, movement cooperation focused on UN disarmament negotiations, but it was ineffective owing to rival political agendas. Nonetheless, the movement enlisted broad public support, which it focused on the threat of nuclear war, and it reinforced arms control negotiations. It also broadened the political base in several countries and stimulated movement contacts with counterparts in the Soviet bloc.

Meanwhile, an informal coalition of peace and human rights activists in Central and North America effectively stymied overt US intervention on behalf of repressive governments there. An international campaign was mounted against South African apartheid. A small coalition called the Neptune Group helped to broker UN negotiations that led to the 1982 Law of the Sea Treaty. Movement related human rights activists in North and South America parlayed Perez Equeval's 1980 Nobel Peace Prize into leverage for victims of oppression. Throughout the century the Peace Prize Committee had broadened both its understanding of peace and the regions from which winners came; in this it reflected the peace movement itself.

The Cold War collapse in 1989 left peace advocates to redefine their movement. On the centenary of the First Hague Conference, about 4,000 delegates representing NGOs around the world assembled in The Hague to help set a UN agenda for a decade devoted to achieving a culture of peace. They explicitly defined themselves as a transnational peace movement consisting of a global network of NGOs that together reflect a world of growing, interactive social movements.

5. Scholarly Analysis: History and Sociology

Scholarly conceptualization of peace movements has tended toward a framework of analysis parallel to the historical phenomenon.

5.1 Peace Movement History

In the 1970s historical studies of peace movements flowered into a comprehensive field that expanded for two decades more. In some respects it paralleled growth of resource mobilization theory in the sociology of social movements.

Some historians revisited periods covered earlier and researched them more thoroughly. Others updated the story of peace advocacy. New patterns emerged. Given the massive work of Peter Brock, for instance, religious pacifists could be distinguished from and related to peace activists. Continuities emerged, such the growth of international organization from 1914 to 1945, the process of conceptually linking peace to justice and social change (positive peace), and the links between nonviolent activism against war and for human rights. Drawing on early sociologist Ferdinand Tönnes, 'polity' peace advocates (who stressed institutional arrangements for the sake of order) were distinguished from 'community' ones (who valued popular involvement and the well being of diverse peoples). Charles DeBenedetti incorporated much of the new work in *The Peace Reform in American History* (1980). The title was significant: treating peace as a social reform invited attention to the relationships between organizations, ideologies, and social movements.

Much has been done in the subsequent two decades, both to correct imbalances in the accounts of national movements and to connect them. Several historians rescued the vital roles of women in the movement from

the neglect of male authors, for example (see Alonso 1993). Peace historians formed a transnational network of scholars from Europe, America, Canada, Australia, and Japan, working through the *Arbeitskreis Historische Friedensforschung* (German Working Group on Historical Peace Research) and the Peace History Society. Europeans revised Beales, while an American analyzed the nineteenth century European movement in terms of relationships between patriotism and internationalism and among varying class and political groupings (Cooper 1991). Together they studied the relationship of peace movements and political cultures. US and Russian historians produced a joint history of the concept of peace in western history (Chatfield and Ilukhina 1994).

5.2 The Sociology of Social Movements

While historians were interpreting peace advocacy as a diverse reform movement, sociologists were analyzing social movements with resource mobilization theory: what are the terms on which social movements strengthen and use their resources? Basic works in the field date from the 1970s (see *Social Movements: Resource Mobilization Theory*).

The fully developed method was applied to peace movements in the 1990s, especially to the Freeze campaign of the previous decade but also in a comparative analysis of three US peace coalitions. Most recently, modified resource mobilization analysis has been applied to the dynamics of transnational NGOs and social movements (see Smith et al. 1997). One study demonstrates the inter-relationships of developing nongovernmental and intergovernmental organizations *before* 1945 (Chatfield, Chap. 2 in Smith et al. 1997), and Wittner reifies the transnational social movement model with a longitudinal study in the period *since* then (Wittner 1993, 1997).

Besides filling out the history of specific peace movements, therefore, the future of this field is likely to involve narrative and analysis of more transnational connections among peace and related movements. Future scholarship may well explore specific peace and antiwar movements as dynamic elements in the creation of transnational infrastructures in an increasingly integrated world.

See also: Civil Rights Movement, The; Conflict and Conflict Resolution, Social Psychology of; Conflict Sociology; Peace; Peace and Nonviolence: Anthropological Aspects; Peacemaking in History; Social Change: Types; Social Justice; Social Movements: Resource Mobilization Theory; Social Movements, Sociology of; Theory: Sociological; Voluntary Associations, Sociology of; World Systems Theory; Youth Movements

Bibliography

Alonso H H 1993 *Peace as a Women's Issue: A History of the US Movement for World Peace and Women's Rights*. Syracuse University Press, Syracuse, NY

Carter A 1992 *Peace Movements: International Protest and World Politics Since 1945*. Longman, London

Chatfield C 1992 *The American Peace Movement: Ideals and Activism*. G K Hall, Boston

Chatfield C, Ilukhina R (eds.) 1994 *Peace/Mire: An Anthology of Historic Alternatives to War*. Syracuse University Press, Syracuse, NY

Cooper S 1991 *Patriotic Pacifism: Waging War on War in Europe, 1815–1914*. Oxford University Press, New York

DeBenedetti C 1978 *Origins of the Modern American Peace Movement, 1915–1929*. KTO, Millwood, NY

Howlett C 1991 *The American Peace Movement: References and Resources*. G K Hal, Boston

Josephson H (ed.) 1985 *Biographical Dictionary of Modern Peace Leaders*. Greenwood, Westport, CT

Kuehl W (ed.) 1983 *Biographical Dictionary of Internationalists*. Greenwood, Westport, CT

Peace and Change: A Journal of Peace Research. Quarterly. Blackwell, Boston

Powers R S, Vogele W B 1997 *Protest Power, and Change: An Encyclopedia of Nonviolent Action from ACT-UP to Women's Suffrage*. Garland, New York

Smith J, Chatfield C, Pagnucco R 1997 *Transnational Social Movements and Global Politics: Solidarity Beyond the State*. Syracuse University Press, Syracuse, NY

Wittner L S 1993, 1997 *The Struggle Against the Bomb: Vol. 1, One World or None: A History of the World Disarmament Movement, 1954–1970; Vol. 2, Resisting the Bomb: A History of the World Disarmament Movement, 1954–1970*. Stanford University Press, Stanford, CA

C. Chatfield

Peace Movements, History of

Peace movements are social and political movements appearing since the early nineteenth century as a result of a collective effort to organize pacifism. They can be long-term phenomena aimed at a lasting peace. Alternatively, they can be stimulated by a concrete situation ready to vanish the moment the movement has reached its objective or pass into a latency period to be reactivated only if need be.

1. Peace Movement and Pacifism

Today both terms are frequently used as synonyms. This is even more justified as continental European peace activists adopted the term 'pacifism' the moment it was coined and widely picked up by the sociopolitical language. It proved to be useful as it suggests

a triple meaning i.e., 'movement,' 'theory,' and 'radicalism.' Anglo-American usage often makes a distinction between 'internationalism' and 'pacifism' with 'internationalism' comprising the moderate demands of peace movements in their entirety whereas 'pacifism' means the idea of the greatest possible radicalism in the practice of peace including absolute refusal to serve in the military. Based on British conditions Ceadel offers an elaborate distinction between 'pacificism' and 'pacifism' (Ceadel 1980).

2. Origins

The French Revolution first created conditions that allowed peace movements to arise. On the whole, they remained restricted to the European-North-Atlantic sphere far into the twentieth century. Convictions matured within the pre-Revolutionary bourgeois society that there is no rational argument to justify war. The mass experience of war following the French Revolution of 1789 and Napoleonic rule gave the impulse for the organization of an individual peace mentality. The peace movement draw important program elements from a store of peace prospects dating back to Judaic-Graeco-Roman antiquity and early Christianity which manifested themselves in christology based, eschatological expectations of religious movements that included the vision of a nonviolent society and in the religious practice of Christian peace sects (Brock 1991a) as well as in literature, the arts, philosophy and political science. Apart from the singularity of Gandhi in India (Brock 1983, Kozhuvanal, in: Brock and Socknat 1999) no indigenous nuclei for peace movements are to be found outside the European-North Atlantic sphere with the exception of Japan where they began appearing at the end of the nineteenth century particularly in connection with the reception of Christian ideas (Powles, in Brock and Socknat 1999), while other moves to organize pacifism starting around 1900 in Australia, New Zealand, Latin America, and even in Japan had either been dependent on or inspired by the Anglo-American model (see Fried 1972).

3. The Starting Period

The time spanning from the end of the Napoleonic era to the beginning of the age of imperialism was a take-off period for peace organizations. They emerged from two sources: on the one hand, religious motivation particularly characteristic of Mennonites and Quakers in North-America and England (Brock 1990, Ceadel 1996), which led to the founding of peace societies in New York and Massachusetts in 1815 and London 1816 and on the other hand a predominantly philanthropic rationalistic motivation leading to the foundation of the first peace societies on the continent in Paris in 1821 and Geneva in 1830. The discourse of both focused on how to ban war as a means of conflict resolution and to get this idea generally accepted nationally and internationally.

3.1 Difficult Dialectics

All attempts to come to grips with this problem caused by the revolution of 1789 proved to be difficult. While the social organization to ban war had been possible, peace movements had to face a new type of war that was another by-product of that secular transition. National societies turned nation-states made a concept of war feasible completely different from every preceding one, i.e., it mobilized the nation as a whole with the total annihilation of the adversary as the final goal. The concept of peace based on rationalism arose from bourgeois society yet the same society due to its increasing economic and political dynamics carried an inherent potential for war the dialectics of which constitute the peace movements' field of activity during the nineteenth century until the World War I.

3.2 Differences and Correspondence

The general development assumes a different profile in every individual nation-state (van der Linden 1987). The US peace movement owes its origin to a very specific situation (Brock 1968). In Europe's young nation-states Italy and above all the German Empire (national peace societies founded in 1887 and 1892 respectively) the development occurred with considerable delay and had a nature different from that of the British Empire, a well established nation-state, or France a nation-state having aspirations towards Europe and overseas or a multinational state like Austria-Hungary (Austrian peace society founded in 1891). It is of significance whether a country has a stable parliamentary system including minority and nonconformist groups into the national consensus or if a country is characterized by a weak parliamentary structure and/or authoritarian leadership structure as well as a tendency to marginalize dissenting political views (see Ceadel 1987).

Notwithstanding their different national conditions all peace movements of the nineteenth century have a lot in common with regard to their homogenous appearance, their organization, and their means of communication. They use societies and associations typical of the bourgeois society of the nineteenth century. The growing participation of the radical bourgeois women's movement gains importance (see Bortolotti 1985). Communication was established via national and international congresses and a broad spectrum of literature. The result was a well-organized pacifist International, whose cohesion was strengthened by the foundation of the International Peace Bureau in Bern in 1892, and remained almost intact until World War I.

Peace activists were convinced that their concept of the need for peace could be most effectively promoted by individual powers of persuasion. Cognitive learning was believed to be the most effective method for instilling an understanding of the need for peace. That is why great importance was attached to schooling and national education as 'indirect means of peace.' Peace activists believe peace as well as war to be due to an individual act of will. This conviction led to appeals for peace addressed to the rulers and delivered at international congresses (Brussels 1848, Paris 1849, Frankfurt am Main 1850, London 1851, Manchester 1852, Edinburgh 1853). This continued on a universal level with the First Universal Peace Congress in Paris in 1889. The prerequisite for this was social homogeneity within the urban society as represented by merchants, industrialists, bankers, physicians, civil servants, professors, teachers, as Chickering shows for Germany (Chickering 1975). Protestant backgrounds show significantly greater receptiveness to pacifist recruitment than Catholic ones. In countries with a predominantly Catholic majority like Italy the willingness to embrace pacifism is characteristically present in connection with economic prosperity, liberal and/or republican convictions rooted in the Risorgimento, freemasonry, etc.

Eventually the social question vanished from the agenda of the international peace movement despite the considerable contributions of renowned socialists and anarchists in organizing pacifism from the left as early as during the inaugural congress of the International League of Peace and Liberty in Geneva in 1867. This corresponds to the increasing distance between the labor movement and the peace movements. In Germany, the distance is biggest, whereas it is substantially less in France and Italy and almost non-existent in England. The sooner the labor movement had separated from bourgeois democracy while opening up to Marxist theory the more brusquely it emphasized the distance from the peace movements. For the German peace movement cooperation with social democracy remained a top priority, that long went unrealized. In accordance with their bourgeois nature, peace movements maintained regular contact with liberal parties. Wherever political liberalism became diversified like in Germany the peace movement joined left liberalism because of its manifest inclinations towards internationalism, whereas national liberalism accused the peace movement of lacking national identity, and participated in stigmatizing pacifists (Holl 1988b).

Revolutionary wars did not present an argumentative problem to the early peace movements as long as the goal of these wars was national unity (van der Linden 1987). Behind this position was the expectation that lasting peace would set in with the elimination of reactionary governments at the hands of national movements. For the bourgeois left who were active in the early peace movements war for national unification was legitimate and the last of all wars to be fought. Having found its destiny within the nation-state, a country was expected to fit into the community of nations and to submit to the international legal system in case of conflict. Pacifist arguments also legitimized national defensive warfare necessary to secure the goal reached in the process. S. E. Cooper has coined the term 'patriotic pacifism' for this position (Cooper 1991). The position on revolutionary warfare looked quite different with regard to social revolutionary goals, all the more when socialism began to dissociate itself from peace movements. Ever since bourgeois pacifism's rejection of war has been influenced by the fear that warfare could result in a social coup.

The British peace movement viewed obstruction of international free trade as yet another legitimate reason for war. In connection with the struggle over the Corn Laws economic interests were combined with pacifist propaganda. The international peace movement joined forces with the free trade movement resulting in an organizational symbiosis (Ceadel 1996) and the development of something along the lines of 'free-trade pacifism.'

As a bourgeois reform movement, the international peace movement joined abolitionism leading pacifism, utilitarianism, Quakerism, free-trade policy, and anti-slavery agitation to come together. Faced during the American Civil War with the conflicting aims of the abolition of slavery and maintenance of national unity on the part of the northern states and the defense of slavery and free trade on the Confederate side, the international peace movement leaned in the end towards the union while Quakers on both sides objected to military service (Brock 1968).

3.3 The Methods Propagated

The methods propagated for achieving and securing peace are twofold. As the prevailing state of affairs in international relations was perceived as 'international anarchy,' peace movements proposed a fundamental reform of these relations and suggested the creation of a legal codex. Concepts were discussed for legal procedures of conflict resolution to be enforced by an international tribunal and a court of arbitration with either facultative or compulsory arbitration. The underlying conviction was the call for 'peace through justice' that stimulated the origins of pacifist international law. The inherent idea was to restrain military influence on foreign affairs.

To the extent that the volatile nature of the national movements and the possibility of their exploitation by the cabinets were recognized as threats to peace, the call was made for fraternization and concepts for a league of nations on the basis of a free federation of European nations, republican constitutions, and the autonomy of communities to thus circumvent the state as a basis for social organization. This was propagated

at the Geneva Congress in 1867. In the end, corresponding attempts to organize based on Proudhonism had no chance. Continental peace movements preferred to propagate the 'United States of Europe' instead.

4. The Age of Imperialism

From 1901 onwards the supporters of peace movements called themselves 'pacifists' and their program 'pacifism.' Having the advantage of multilingual usability, the term met the requirements of an international movement. Earlier names such as 'friends of peace' finally vanished. The new term signaled an advanced level of theory and the intention of expanding the social base. The creation of the term reflects the success of both socialism and social Darwinism as social theories, and of imperialism. The spread of social Darwinist thought in Europe and in the USA that led public opinion to be more favorably disposed to imperialism and warfare had a negative effect on pacifist campaigning. Owing to its Eurocentrism and its inclination to cultural imperialism the international peace movement itself was occasionally susceptible to the imperialistic zeitgeist. The growing crisis of the global system incited the international peace movement and eminent pacifist authors to develop new concepts and theories (Grossi 1994).

Since the end of the 1860s new peace movements sprang up emphatically propagating peace negotiations by arbitration. Tsar Nicholas II's peace manifesto (1898) and the First Peace Conference in The Hague (1899) proved to be false confirmation of the movement's demand for disarmament. Despite this seeming success the peace movement's societal position remained marginal in Austria and Germany where it faced male prejudice not least because of the participation of the radical bourgeois women's movement and due to the fact that Bertha von Suttner, in whose novel *Lay Down the Arms!* (1889) a woman's concern about world peace is expressed, took a leading position in the peace movement (Holl 1988b). In his oeuvre *War in the Future in its Technical, Economic and Political Aspects* (1898) Jan Bloch predicted, against all social Darwinist and strategic expectations of the time, the disastrous consequences which were confirmed by World War I. A. H. Fried developed a peace theory he termed 'organizational pacifism' which ruled out world war because the growing network of international relations would prevent it for utilitarian reasons. Fried's theory won support from Norman Angell's book *Europe's Optical Illusion* (1909, extended 1910, *The Great Illusion*) that held future wars to be economically futile because of the interdependence of national economic interests.

The ineffectiveness of the peace movement gave way to further pacifist concepts on the eve of World War I. Their aim was to win prominent leaders in academic circles for international understanding. Their target in Germany was the 'Verband für internationale Verständigung,' founded in 1911 (Chickering 1975). Furthermore there were efforts promoting bilateral international understanding and supported by the international peace movement: on a Protestant church-based level for instance between England and Germany, and on a parliamentary level between French and German MPs. Independent from such phenomena, communities which lived according to the principle of absolute non-violence continued to exist with particularly those groups following Tolstoy's teachings receiving the greatest public attention (Brock 1991b).

5. World War I

World War I created a deep rift within the international peace movement and cut the national peace movements' flexibility to a considerable extent. The experience of war, during which traditional peace societies held on to their patriotic pacifism led to the founding of pacifist organizations with a new style and program expressing the radical pacifists' frustration with the insufficient prewar peace societies. This ended the older pacifism's abstinence from domestic policy as the dependence of foreign policy on domestic conditions was scrutinized. In Germany, the 'Bund Neues Vaterland' demanded politics of domestic reform (Holl 1988b). The British 'Union of Democratic Control' called for a new foreign policy without secret diplomacy. In addition resolute pacifism organized in the Netherlands, 'Nederlandse Anti-Oorlog-Raad,' and in the USA, the 'League to Enforce Peace.'

The new pacifism formed a more solid base for international organizations: The international congress of pacifists in the Hague in 1915 led to the short-lived 'Central Organization for a Durable Peace.' The 1915 Hague congress of women war opponents, which was rooted in the radical bourgeois women's movement, ended with the foundation of the 'International Women's Committee for a Lasting Peace' which led to the 'International Women's League for Peace and Freedom' in Zurich in 1919 (see Schott 1997). Inspired neither by the old or new style peace movements nor by Gandhi and Tolstoy's teachings, refusal of military service surfaced for the first time as a mass phenomenon, particularly in Great Britain, where objectors joined the 'No Conscription Fellowship' during World War I.

6. The Interwar Years

During the interwar period, peace movements were marked by the coexistence of radical pacifist positions with the moderate orientation of traditional peace organizations and a widening of their social base into the working classes. While radical pacifism expressed

itself in the mass-movement 'No More War,' organizational pacifism lived on in the call for the 'League of Nations' that found outstanding support in Britain. The new peace organizations in France and England differed from the German peace movement in their judgment of the imminent threat to world peace posed by right-wing dictatorships. In France, the 'Ligue internationale des Combattants de la Paix' emerged as the mouthpiece of integral pacifism in the 1930s. It played down the threat of war emanating from Nazi Germany, attacked French defense policy, and demanded strict neutrality towards current military conflicts, a position that could well have led to collaboration during German occupation (Ingram 1991). This attitude corresponded to the British 'Peace Pledge Union,' which began to loose ground only after the obvious failure of the British appeasement policy (Ceadel 1980). With the beginning of national socialist rule, the German peace movement was forced into exile or exposed to persecution (see Holl 1988a).

7. *After World War II*

The nuclear attacks on Hiroshima and Nagasaki as well as the postwar strategy of nuclear deterrence developed in the course of the East-West confrontation gave new dimensions to the discussion on war and peace. Originating from the USA and directed against the Vietnam war, a radical peace movement staging a new type of mass protest and with the characteristics of the 'New Social Movements' came into being (see Breyman 1998). In accordance with nuclear pacifism, this peace movement argued that nuclear war of aggression, nuclear preventive warfare, and nuclear defensive warfare were virtually indistinguishable from one another. Additionally, it expressed doubt with regard to the peacekeeping effect of the nuclear deterrence strategy and played up the fear of the nuclear self-destruction of humanity. These arguments motivated the movement's opposition to all weapons of mass destruction. In a large number of Western countries in the 1980s, the new peace movement achieved a mass mobilization against the installation of medium-range NATO missiles in Western Europe that in the end proved futile.

See also: Social Movements and Gender

Bibliography

Bortolotti F D 1985 La Donna, La Pace, l'Europa. L'Associazione internazionale delle donne dalle origini alla prima guerra mondiale. Franco Angeli Libri, Milan, Italy

Breyman S 1998 *Movement Genesis. Social Movement Theory and the 1980s West German Peace Movement*. Westview Press, Boulder, CO

Brock P 1968 *Pacifism in the United States: From the Colonial Era to the First World War*. Princeton University Press, Princeton, NJ

Brock P 1983 *The Mahatma and Mother India: Essays on Gandhi's Non-violence and Nationalism*. Navajivan Publishing House, Ahmedabad, India

Brock P 1990 *The Quaker Peace Testimony 1660 to 1914*. Sessions Book Trust, York, UK

Brock P 1991a *Freedom from Violence: Sectarian Pacifism from the Middle Ages to the Great War*. University of Toronto Press, Toronto, Canada

Brock P 1991b *Freedom from War: Nonsectarian Pacifism 1814–1914*. University of Toronto Press, Toronto, Canada

Brock P, Socknat T P (eds.) 1999 *Challenge to Mars: Essays on Pacifism from 1918 to 1945*. University of Toronto Press, Toronto, Canada

Ceadel M 1980 *Pacifism in Britain 1914–1945: The Defining of a Faith*. Clarendon Press, Oxford, UK

Ceadel M 1987 *Thinking about Peace and War*. Oxford University Press, Oxford, UK

Ceadel M 1996 *The Origins of War Prevention. The British Peace Movement and International Relations, 1730–1854*. Clarendon Press, Oxford, UK

Chickering R 1975 *Imperial Germany and a World Without War. The Peace Movement and German Society, 1892–1914*. Princeton University Press, Princeton, NJ

Cooper A H 1996 *Paradoxes of Peace. German Peace Movements Since 1945*. University of Michigan Press, Ann Arbor, MI

Cooper S E 1991 *Patriotic Pacifism. Waging War on War in Europe 1815–1914*. Oxford University Press, Oxford, UK

DeBenedetti C, Chatfield C 1990 *An American Ordeal. The Antiwar Movement of the Vietnam War* 1st edn. Syracuse University Press, Syracuse, NY

Fried A H 1972 *Handbuch der Friedensbewegung, Vols. I and II*. Garland Publishing, New York

Grossi V 1994 *Le Pacifisme Européen 1889–1914*. Bruylant, Bruxelles, Belgium

Holl K 1988a German pacifists in exile. In: Chatfield C, van den Dungen P (eds.) *Peace Movements and Political Cultures*. University of Tennessee Press, Knoxville, TN, pp. 165–83

Holl K 1988b *Pazifismus in Deutschland*. Suhrkamp Verlag, Frankfurt am Main, Germany

Ingram N 1991 *The Politics of Dissent. Pacifism in France 1919–1939*. Clarendon Press, Oxford, UK

Liddington J *The Road to Greenham Common. Feminism and Anti Militarism in Britain Since 1820*. Syracuse University Press, Syracuse, NY

Schott L K 1997 *Reconstructing Women's Thoughts: The Women's International League for Peace and Freedom Before World War II*. Stanford University Press, Stanford, CA

Vaisse M (ed.) 1993 *Le Pacifisme en Europe des années 1920 aux années 1950*. Bruylant, Bruxelles, Belgium

van der Linden W H 1987 *The International Peace Movement 1815–1874*. Tilleul Publications, Amsterdam

Wittner L S 1993 *One World or None. A History of the Nuclear Disarmament Movement Through 1953*. Stanford University Press, Stanford, CA

K. Holl

Peace Promotion, Psychology of

Brutal conflicts such as those in Rwanda, Kosova, and East Timor have boosted awareness of the subjective dimensions of war, genocide, and intergroup conflict

and of the need to address hatreds, fears, and emotional wounds of war in building peace. Within a multidisciplinary perspective, this article examines the psychological origins of war and destructive intergroup conflict and psychological approaches to resolving conflict nonviolently and to building peace with social justice.

1. Historical Context

Psychology became prominent partly through helping the US develop government tests for selecting military recruits in World War I. Through World War II, many psychologists contributed to war efforts, although William James, Edward Tolman, and others worked for peace.

Following World War II, social psychologists such as Morton Deutsch, Herbert Kelman, Otto Klineberg, Gardner Murphy, Charles Osgood, Muzafer Sherif, and Ralph White analyzed sources of intergroup tension and methods of nonviolent conflict resolution, with an emphasis chiefly on the Cold War. The nuclear threat animated large numbers of psychologists worldwide, as researchers such as Robert Jay Lifton, John Mack, and Milton Schwebel examined the psychic impact of living under the threat of nuclear annihilation. Throughout the Cold War, psychological research focused mostly on negative peace, that is, on stopping and preventing war (White 1986), rather than on positive peace, which involves social justice, positive inter-group relations, human rights, and sustainable development.

Following the Cold War, psychologists examined a wider array of peace issues such as women and war, militarism, children and armed conflict, environmental destruction, and postconflict healing.

2. Psychological Sources of War and Destructive Intergroup Conflict

Armed conflicts, including the intrasocietal conflicts that are now the dominant form of war, have complex political, historical, and economic roots. Much destructive conflict originates in a divergence of objective interests, competition over scarce resources such as land and water, and worsening life conditions. The 1994 genocide in Rwanda, for example, had antecedents in land scarcity, social class, and the political and economic privileging of Tutsis under the Belgian colonial regime (Prunier 1995). Many conflicts have their origins in struggles over oil, diamonds, land, and other objective resources.

Armed conflict also has subjective sources such as social identity and ideology. Even when no objective divergence of interest exists, people tend to categorize themselves as belonging to different groups and to show in-group favoritism in judging people, thereby attaining self-esteem and positive self-identity from identification with the in-group (Tajfel 1982). Typically, they judge others relative to the standards set by the in-group, leading to ethnocentric attitudes and behavior, and they derogate members of the out-group. Identity processes alone do not cause violence, but they set the stage for other subjective processes such as hostile ideologies to generate hatred and fighting. Before the Rwandan genocide, President Habyarimana, a Hutu, used mass media to create an ideology of hatred toward Tutsis. Similarly, Hitler used an ideology of racial superiority and global Jewish conspiracy to fuel anti-Semitism and enable the Holocaust (Staub 1989).

Social identity processes play a key role in interethnic conflicts, often referred to as identity conflicts. Ethnic groups are 'imagined communities' (Anderson 1991) that share distinctive language, religion, social institutions, and origin myths. Ethnic identities exercise powerful emotional influence, as children are taught to love their people, their mother tongue, and their way of life. People learn to define themselves by reference to the larger group (e.g., 'I am Russian'), to experience unity with the group, to be patriotic, and to make sacrifices for the group. The sense of unity and solidarity, integrated with a powerful sense of homeland and a sense that the group requires protection, may fuel nationalism and separatist desire for an independent national state.

Conflict escalation often stems from the interplay of objective and subjective factors. As conflict over resources escalates, negative psychological dynamics come into play, and these become self-perpetuating parts of the conflict process (Rubin et al. 1994). Each group tends to create enemy images—demonizing, exaggerated stereotypes of the 'other' that amplify fear and motivate aggressive action (White 1984). These images bias perceptions, leading adversaries to make negative attributions or inferences about motivations regarding the 'other's' behavior. For example, building military forces could be motivated by desire for security, but an adversary who harbors enemy images tends to perceive such actions as having hostile, offensive intent. Negative perceptions and fears on both sides fuel behaviors that the 'other' regards as threatening, contributing to malignant conflict spirals that tend to become self-fulfilling prophecies. By using public media to propagate enemy images, leaders mobilize society for war and genocide.

In a destructive conflict, damage to the parties' relationship becomes part of the conflict, and the emotional and social wounds encourage fighting. Viewing the other side's actions as transgressions, each party sees itself as a victim, constructs collective beliefs or myths about the traumas inflicted on it, and believes that violence is necessary for self-preservation (Volkan 1997). Further, the parties tend to become highly polarized, and intragroup pressures discourage con-

tact with the other side, which is viewed as having incompatible goals. In protracted conflicts, each side may come to view itself as irreparably opposed to the other side, leading to an ethos of antagonism and a tendency of each group to define itself in part via opposition to the other.

Subjective influences are visible also in the biased decision-making of policy elites. Many leaders' aversion to losses encourages excessive risk-taking and departure from rational decision making. In intense conflicts, leaders may experience powerful fears, which motivate misperceptions regarding their adversaries' motives, strength, and willingness to fight. As intercommunal tensions escalate, leaders may experience a cognitive constriction evidenced in reduced complexity of views of the adversary expressed in speeches. Leaders may also use inappropriate historical analogies to guide thinking about current crises. Under pressure of crises, small leadership groups having a strong *esprit de corps* and a charismatic leader may make flawed decisions guided by premature consensus, a sense of invulnerability, failure to weigh moral concerns, internal suppression of dissent, and poor contingency planning (Janis 1982). In a variety of contexts, obedience to authority is a powerful enabler of armed conflict, human rights abuses, and even genocide (Kelman and Hamilton 1989).

3. *Nonviolent Conflict Resolution*

Conflict can be constructive when it is managed in ways that promote communication, healthy relationships, and a sense of positive interdependence (Deutsch and Coleman 2000). To handle destructive conflict nonviolently, psychologists have developed or refined numerous tools designed to resolve conflict and to transform or repair damaged relationships (Rubin et al. 1994).

Negotiation, the handling of conflict through communication and bargaining by the parties themselves, has important psychological dimensions. If, as often occurs in official diplomacy between hardened adversaries, opponents view negotiation as a win–lose affair, they may dig into entrenched positions and use damaging, coercive tactics such as threats. Psychologically, a more useful alternative is principled negotiation, which reframes negotiation as a win–win affair. In this approach, parties collaborate to solve their problems and seek not to beat the opponent but to enable all parties to meet their underlying needs, thereby contributing to improved communication, empathy, and relationship. Which approaches to negotiation are likely to succeed depends not only on the situation, the kind of issues at stake, and the amount of fear and relational damage, but also on culture. For example, US negotiators favor direct communication and emphasize what is said, while Asian negotiators often communicate indirectly, relying on context and maintaining group harmony by avoiding direct confrontation.

Strong mistrust, fear of appearing weak, and poor communication can make negotiation infeasible. In this situation, a preferred approach is mediation, in which a third party helps the parties to negotiate a settlement. By proposing alternatives that the parties may not be willing to suggest themselves, mediators enable face-saving. They may also use social influence processes such as 'carrots and sticks,' a mixture of promised rewards for conflict reducing steps and promised punishments for intransigent behavior. Skilled mediators manage the conflict process, separating the parties when they are likely to say or do damaging things.

Official diplomacy, however, is limited since hardened adversaries may be unwilling to meet publicly. Also, public agreements can evoke backlashes, and official treaties may not change the polarization, fear, and hostility present in divided societies and communities. Kelman has pioneered the use of interactive problem-solving workshops (Fisher 1997) as a tool of unofficial diplomacy and relational improvement. Regarding the Israeli–Palestinian conflict, a typical workshop brings together three to six Israelis with an equal number of Palestinians for three days of private dialogue facilitated by social scientists. In analytic, problem-solving discussion, the participants examine the main issues at stake in the conflict, explore their concerns and fears, and identify possible steps and solutions that could reduce the political and psychological barriers on both sides. Empathy, problem solving, and group exploration of 'what-if?' questions are encouraged. Afterwards, participants take their new learning about the other side back into their communities, beginning the wider process of community transformation. By building positive relationships and infusing useful ideas into conflict-torn communities, this method provides a useful complement to official negotiation.

Positive contact between members of polarized, isolated groups can also reduce destructive conflict (Pettigrew 1998). Positive contact is likely to occur in a context of equal status that enables individuals to get to know each other personally and when steps have been taken to gain leaders' support and prevent damaging behavior. To reduce perceptions that the positive individuals from the out-group are exceptions, one may use cognitive strategies that enable generalization. One useful strategy is to build superordinate group identities, as occurred in postapartheid South Africa when blacks and whites forged larger identities as 'South Africans' and championed the national rugby team.

The most powerful form of cross-group contact involves cooperation toward the accomplishment of shared goals that neither group could achieve on its own. Global problems such as nuclear proliferation, environmental destruction, and the HIV/AIDS pan-

demic provide many possibilities for cooperation on shared goals that mitigates and prevents destructive conflict.

4. Building Peace with Social Justice

Beyond war prevention and nonviolent conflict resolution, peace requires social justice, social equity, and sustainable patterns of living. Psychology has analyzed how beliefs, attitudes, and practices support social injustice and '-isms' such as sexism.

In systems of severe oppression and genocide, there is a pattern of moral exclusion in which out-groups are dehumanized and removed from the moral realm, providing a rationalization for killing them. Similarly, sexism is supported by socialization norms and practices that privilege boys, encourage girls to conform to stereotypic roles, and encourage acceptance of existing patterns of patriarchy. Environmental damage through, for example, the production of mass amounts of toxic waste, often leads to environmental racism via 'not in my back yard' attitudes that encourage storage or dumping of toxic materials near impoverished neighborhoods inhabited by minority groups that lack political power. Efforts toward poverty alleviation may be thwarted by tendencies to blame the victim, enshrined in societal beliefs that people are poor because they are lazy or deserve it.

In promoting social justice, useful psychological tools include steps to increase the sense of positive interdependence between groups, media campaigns that heighten salience of justice issues and indicate specific behaviors that can help address the problem, positive modeling of tolerance and restraint by authority figures, and methods of attitude change. In addition, commitment can often be built through the 'foot in the door' method in which agreement to a small request increases the likelihood of subsequent willingness to engage in more effortful action. To change behavior on a wider level, one may use community-based social marketing methods that strategically assess points of leverage for social change, pilot selected interventions for changing attitudes and behavior, and replicate the process on a larger scale (McKenzie-Mohr 2000).

Education for peace is essential in building social justice and peace, as most people are socialized into systems of violence (Raviv et al. 1999). In schools, psychologically informed programs have developed skills of nonviolent conflict resolution, encouraged peer mediation of conflicts, enabled cooperative learning that also reduces excessive competition, and built curriculi that undermine hatred and stereotypes. Beyond schools, education for peace has included activities to enable constructive handling of family conflict, cross-conflict dialogues in divided communities, and use of mass media to show positive handling of conflict, to counteract black–white thinking, and to reduce the excessive emphasis on violence evident in television and the war-toy industry.

5. Future Directions

Since wounds of past violence create emotional vulnerability and enable cycles of violence, building peace requires the healing of emotional wounds and reconciliation. Programs of humanitarian assistance and development increasingly include psychosocial activities aimed at healing the wounds of war through normalizing activities, expressive arts, reintegration of former child soldiers, and use of local cultural networks and resources. A key task for the future is to find the most effective, culturally appropriate ways of rebuilding torn societies. There is great need of research on reconciliation, the repair of damaged relationships, and its connection with processes such as truth-telling, forgiveness, and establishment of justice. As psychology addresses these issues, it will examine a wider array of cultures and develop theories that connect microlevel change in individuals and small groups with macrolevel changes in societal and international structures.

See also: Conflict and Conflict Resolution, Social Psychology of; Dispute Resolution in Economics; Peace; Peace and Nonviolence: Anthropological Aspects; Peace Movements; Peacemaking in History; Religion: Peace, War, and Violence; War: Anthropological Aspects; War: Causes and Patterns; War, Sociology of

Bibliography

Anderson B 1991 *Imagined Communities—Reflections on the Origin and Spread of Nationalism*. Verso, London

Christie D J, Wagner R V, Winter D D 2001 *Peace, Conflict and Violence*. Prentice-Hall, Upper Saddle River, NJ

Deutsch M, Coleman P 2000 *Handbook of Conflict Resolution: Theory and Practice*, 1st edn. Jossey-Bass, San Francisco

Fisher R 1997 *Interactive Conflict Resolution*. Syracuse University Press, Syracuse, NY

Janis I 1982 *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*, 2nd edn. Houghton Mifflin, Boston

Kelman H C, Hamilton V L 1989 *Crimes of Obedience*. Yale University Press, New Haven, CT

McKenzie-Mohr D 2000 Fostering sustainable behavior through community-based social marketing. *American Journal of Psychology* **55**: 531–7

Pettigrew T F 1998 Intergroup contact theory. *Annual Review of Psychology* **49**: 65–85

Prunier G 1995 *The Rwanda Crisis*. Columbia University Press, New York

Raviv A, Oppenheimer L, Bar-Tal D 1999 *How Children Understand War and Peace*. Jossey-Bass, San Francisco

Rubin J Z, Pruitt D G, Kim S H 1994 *Social Conflict*, 2nd edn. McGraw-Hill, New York
Staub E 1989 *The Roots of Evil*. Cambridge University Press, Cambridge, UK
Tajfel H 1982 *Social Identity and Intergroup Relations*. Cambridge University Press, New York
Volkan V 1997 *Bloodlines*, 1st edn. Farrar, Straus and Giroux, New York
White R K 1984 *Fearful Warriors*. Free Press, New York
White R K (ed.) 1986 *Psychology and the Prevention of Nuclear War*. New York University Press, New York

M. Wessells

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Peacemaking in History

1. Definitions

1.1 Peace and War

War is declared, begun or unleashed, or it breaks out; peace is concluded, settled, agreed or imposed. Ordinary language thus shows an important asymmetry: war is either the outcome of forces neither side can fully control, or the result of one-sided action, while peace has to be established in an effort involving both sides, unless it is simply the result of the dying out of fighting or of the annihilation of one side. For the purpose of this article, *peace* is defined negatively as absence of war and other forms of widespread collective fighting. This static definition does not do justice to the concept of peace as a movement, a process of improvement with stages in between just a momentary absence of war and ideal, eternal peace and harmony. But this positive, dynamic view of peace concerns a different subject; this article deals only with the crucial transition from a state of war to a state of peace. *War* is defined as a state of collective armed violence between politically organized, at least *de facto* and usually *de jure* independent communities, during which special law is valid. The main characteristic of this law is the relaxation of the rules which prohibit killing. War does not presuppose continuous fighting. But if fighting ceases altogether, one can speak of a *de facto* peace. Fighting without application of different law is not war, but anarchy or rebellion. If, later on, the law of war is recognized by both sides, the rebellion or the anarchy has developed into *internal* or *civil war*. Civil war occurs between parties originally forming one political community, while war between independent communities is *external* or *international* war.

With respect to peacemaking, one fundamental aspect of war has to be mentioned: war has inevitably the character of an ordeal, of a 'judgement of God,' as power and not justice decides its outcome (Kant 1968).

1.2 Peacemaking and Pacification

Peacemaking is defined as transition from a state of war to a state of peace by a deliberate action of the parties involved. Departing from a differentiation between three types of peace introduced by Raymond Aron: equilibrium—hegemony—empire (Aron 1962), peacemaking is here placed between two extremes which mark ideal types. *Peacemaking proper* is characterized by coordination, and hence by full legal (although not necessarily material) equality and independence of the parties. If there is a new conflict between them, it will be international war. *Pacification* is characterized by subordination. It implies the legal extinction of one party as an independent entity by its incorporation into the other. Usually, it is connected with an ex post denial of the vanquished's right to conduct war by the victor. If there is a new conflict, it will start as a rebellion and may develop into a civil war. Each concrete case of war termination can be situated between these two extremes. For pacification, an overwhelming victory of one side is a necessary condition, whereas peacemaking proper tends rather to be linked with a stalemate. But the relative strength of the parties is not the only factor deciding the outcome. There are periods in history during which pacification prevails.

2. Peacemaking Before the Twentieth Century

The roots of peacemaking lie beyond the grasp of written sources. Anthropological research as well as written evidence of transactions between literate and nonliterate political communities suggest that wherever there is fighting between politically organized groups, there are also mechanisms to end fighting on more than just a *de facto* basis, which presupposes some reciprocally recognized arrangements, and thus some reliability and stability. Special oral procedures to memorize the contents of the transactions are developed (cf., e.g., the highly elaborate methods of the North American Indians in their dealings with the Europeans in the eighteenth century, which the Europeans had to adopt as long as they could not simply dictate the terms (Franklin 1938)). While the material stipulations can vary considerably according to the way of living of the peoples involved, the most important and the most difficult arrangements concern guarantees for the keeping and carrying out of the stipulations. Whatever the intention of those who conclude the peace, its oral form tends to limit its scope in time and in space. Frequent renewal, especially after the death of the head of one of the contracting parties, is the most important method to maintain the force of the agreements. Wherever writing develops, peace treaties are among the first instruments for which it is used, and they are frequently written on the most durable materials, like

stone or metal. The oldest treaties preserved date from the third millennium BC in the Ancient Near East. They are to be seen in the context of the pacification of vassals, while the first known treaty between great powers is from the thirteenth century BC, between the Hittite and the Egyptian empires. The problem of guarantees is dealt with by oaths and by the invocation of the intervention of deities, while in other cases hostages are given, either unilaterally or reciprocally, or dynastic matrimonial unions link the destinies of the two parties.

In the course of time, the main topics of peacemaking, besides peace itself and its guarantees, turned out to be territorial, financial and economic. Sometimes they were supplemented by clauses dealing with the destiny of certain populations. Peacemaking proper was almost invariably connected with an amnesty for those who had supported the enemy during the war, based on a more general promise of reciprocally forgiving and forgetting what had happened during the war. The contracting parties were to behave as if peace had never been interrupted, thus marking a sharp contrast between war and peace. There was also a strong tendency in all parts of the world to stress the perpetual character of the peace concluded, not so much as an expression of a realistic expectation but rather as a reminder of a general characteristic of peace (Fisch 1979).

Whether peacemaking proper or pacification prevailed in a certain period and in a certain area, depended to a great extent on the general nature of the relationships between political entities. Well-documented genuinely pluralistic systems of a balance of power (which tend to foster peacemaking proper) are rare. The cases which probably came nearest to it were relationships between the Greek city states in the fifth and fourth centuries BC and between the Hellenistic states before the Roman conquest. Even here, there were tendencies for hegemonies to develop as with Athens or Sparta in Greece. Later the Greek system was absorbed by an empire, first Macedonia and then Rome. It is however to be supposed that in areas and periods with many smaller political entities, of which we have few written sources, peacemaking proper was also the rule. In international systems built around powerful empires, on the other hand, there was a strong tendency to pacification, certainly in the formal, although less in the material sense. Great empires, among them the Roman, the Chinese, the Islamic Khalifate, and the Ottoman and the Mogul Empires, frequently showed a reluctance towards peacemaking proper, claiming some kind of overlordship over their neighbors and rivals. They tended not to conclude formally reciprocal treaties with them but to impose unilateral edicts on them, or else they limited the duration of treaties. The Islamic doctrine, for example, allowed no peace but only armistices of up to 10 years with non-Muslims. Peace was not agreed, but granted. However, reality often belied these pretensions; *de facto* the solution was frequently one between equals, and sometimes the empire was even in the weaker position.

In European history there is a somewhat different interplay between peacemaking proper and pacification. In the Middle Ages there were two trends towards pacification. On the one hand, the Emperor had a claim to supremacy over all the rulers of Christendom. In practice, this claim had hardly any importance; peacemaking among the great monarchs was legally and formally reciprocal. Thus, the basic structure of peacemaking in Europe after major wars was peacemaking proper long before 1648. Much more important with respect to pacification were, on the other hand, feudal relations between lords and vassals on various levels. Vassals tried to acquire full independence, which included the right to be a party in war and thus to become a subject of peacemaking, while lords tried to gain full control over their vassals, which led to attempts at pacification. The result of the century-old struggle was a balance of power-system with fully sovereign entities which, among themselves, practiced only peacemaking proper. In some countries, the ruler (usually a king) had been able to subjugate his vassals by methods of pacification, while in others, especially in the Empire, the vassals had gained full or almost full sovereignty.

Modern European history from the sixteenth century, and even more from the seventeenth century onwards, is thus characterized by a system of legally strictly equal relations. In its framework, particular methods of peacemaking were further developed. During the war, various kinds of peace feelers would be extended. Frequently, mediators played an important part, either as honest brokers or as great powers which tried to influence the settlement according to their own interests (on mediation in general, see Touval and Zartman 1985). The process of peacemaking proper tended to be broken up into three stages: armistice, with cessation of hostilities and, frequently, important material settlements—preliminary treaty, containing the core of the settlement—definitive treaty. This was, however, never a necessary pattern. Preliminary peace treaties played a really significant part only in the 19th century. Frequently, the most important decisions were taken even before hostilities ended (Pillar 1983). One of the most salient features were the peace congresses after general European wars, especially those in 1648, 1713, 1814–15, and 1919. The first of them, in Münster and Osnabrück, which led to the Peace of Westphalia, was an important stage in the development of a European system of legally equal, sovereign states. But it was not, as is frequently claimed, the point of departure for this system, which had much older roots. Such congresses developed general, often multilateral settlements which were likely to be more stable than mere bilateral peace treaties. Their success was remarkable, especially after the Congress of Vienna in 1814/15,

which constituted a kind of joint hegemony, combined with a joint responsibility for the peace, of the great powers. But after all, general peace settlements had as their corollary general wars.

This was, however, but one side of modern European peacemaking. The European states conquered, from the sixteenth to the twentieth century, vast parts of the world. In many areas of Asia and Africa, peacemaking proper between Europeans and non-Europeans was the rule up to the nineteenth century. Especially in America, there were widespread attempts at pacification. They were frequently successful: the extra-European rulers were brought into a legally subordinate position. This happened even in areas where, at first, there had been peacemaking proper, especially in India and Indonesia. The difference was felt by the contemporaries. In Spain, for example, *pacificación* officially replaced *conquista* in 1573, showing the origin of the concept in conquest and not in peace. Yet pacification was often merely a fiction, because the colonial power was too weak to impose its will. Nevertheless, in the long run, pacification tended to become the rule in most colonial dealings.

Thus, at the beginning of the twentieth century, the European experience was one of peacemaking proper in Europe and of frequent pacification outside Europe. For both purposes elaborate and successful techniques had been developed. But this also meant that war was a well-known feature of international life. One tried to avoid it as far as possible; nevertheless it belonged to the accepted experience. This acceptance was to be challenged in the twentieth century.

3. *Peacemaking in the Twentieth Century*

The First World War began as a traditional European war. The fierceness of its fighting and the destructiveness of modern weapons increased the number of combat casualties beyond that of all previous wars. This led to increased demands for the outlawry of war, that war should no longer be considered as a legitimate means of politics but as a crime, and propaganda on both sides tried to depict the enemy as a criminal.

The Second World War further strengthened the view that war was a crime, while the postwar period brought, with thermonuclear weapons, for the first time in history the possibility of a war destroying mankind altogether, so that peacemaking would be neither necessary nor possible. This did not mean, however, that smaller wars had become impossible.

These developments had consequences for peacemaking. The Covenant of the League of Nations, which was part of the peace treaties after the First World War, prohibited war at least to a certain extent (articles 10; 12–15). The Briand–Kellogg Pact of 1928 brought the renunciation of war as a means of national politics, while the UN Charter of 1945 prohibited the use of force—and thus war—altogether, with the exception of legitimate self-defense (articles 2, 4; 51), which was no real exception, as self-defense is the answer to an aggression which in turn is prohibited. Combined with the outlawry was the attempt to prevent war by a system of collective security which would replace war by police-actions of the totality of the member-states against a peacebreaker. Ideally war would have been abolished altogether by this system. But it was weak, because neither the League nor the UN had enough coercive means of its own and even less the power monopoly which would have been needed in order to reduce warfare to an efficient system of policing.

At least in theory, the outlawry of war excluded the possibility of peacemaking proper. If one side was guilty of a criminal act there could be no peace on the basis of legal equality but only punishment. Yet, as there was no efficient system of policing, war still maintained its character of a judgment of God, and there was no guarantee that the guilty party could be punished. Peacemaking reflected these facts. It tended to join elements of both peacemaking proper and pacification, the mixture being dependent on the outcome of the war. After the First World War this meant a combination of the intra-European and of the colonial tradition. There was a peace congress, but with the losers excluded, and there were peace treaties, but with the vanquished formally admitting their responsibility for the war, thus being placed into a legally inferior position. This tendency was increased after the Second World War, when the vanquished were not only deemed, and had to confess themselves, guilty of the crime of aggression, but their leaders were brought to trial for their crimes and their states were occupied and administered by the victors. Especially after the World Wars there was a tendency by the victors to devise definitive solutions at a very early stage of the negotiations if they did not simply impose them by instruments of surrender or armistices.

Peacemaking in the twentieth century outside the context of the World Wars followed this pattern to a great extent. Theoretically, or rather dogmatically speaking, there was no longer any room left for peacemaking proper. But wars could still lead to a stalemate and render traditional peacemaking necessary, although there might be no traditional final peace treaty. A particularly elaborate example was the termination of the Vietnam War 1968–73/5. Thus, peacemaking proper has maintained a certain importance in the context of the termination of wars, and it is likely to keep it as long as there is no efficient global system of policing, although the relevant activities of the UN have increased, especially after 1989, in the shape of peacekeeping and peace enforcement. Peacekeeping often takes the place of a provisional arrangement instead of formal, definitive peacemaking, while peace-enforcement has the character of an imposed mediation. The proliferation of

civil wars in recent years has even increased the importance of some traditional techniques of peacemaking, as the coexistence of former civil-war parties makes reconciliation more important than it is between independent, separate states.

4. *The Conditions of Peace: Research into Peacemaking*

It is an almost stereotyped complaint in the literature on peacemaking that, compared with research into the causes of war, very little research into the causes of peace has been carried out. This is, however, not so much the consequence of a predilection for war or even a warlike disposition of the researchers, but rather of a basic difference in the public appreciation of war and peace. In a prescientific view, war has causes, while peace has not, in the sense that peace is considered as the ordinary and war as the exceptional state of things, even though in reality periods of peace may be shorter than periods of war. The wish for peace usually does not have to be justified in the same way as the wish for war. A state of war sooner or later ends in some kind of peace, while a state of peace does not necessarily lead to war.

The scant research on peacemaking seems to be one of the consequences of these facts. Moreover it is difficult to conceive of causes of peace in a scientific sense. Peace has been successfully concluded in extremely diverse political situations, and there are no methods to calculate the probability of peace in a specific constellation.

Traditionally, peacemaking is a central object of historiography in general and of diplomatic and military history in particular. The focus is usually on individual wars and their termination. Favorite objects of research have been the plans for permanent or perpetual peace through the ages (as an introduction cf., e.g., Hinsley 1963) and the main general peace congresses from 1648 to 1919 (cf. summaries of Holsti 1991 and Osiander 1994), while a comprehensive account of the termination of World War II has yet to be written. Such research helps to understand particular cases of peacemaking, but it usually is little concerned with general causes. More systematic, comparative research has occasionally been conducted during great wars, with a view to facilitate peacemaking. This led neither to more sophisticated systematic accounts nor to the search of causes of peace, but rather to surveys of the range of material regulations of earlier peace settlements (cf., e.g., Phillimore 1917 and Phillipson 1916). The only and very important exception is Quincy Wright's influential *A Study of War* (1965), first published during the Second World War, which, however, was the result of a project initiated long before the war. Wright does not only deal with causes of war but also with causes of peace.

After 1945 there was an apparently paradoxical development. In the social sciences, the threat of an all-destructive thermonuclear war gave, especially from the 1960s, a boost to peace studies and peace research, with the *Journal of Conflict Resolution* (since 1957) and the *Journal of Peace Research* (since 1964) as leading organs. But as peacemaking was no feasible possibility in the context of a thermonuclear war, the central object was the prevention of war and the preservation and improvement of peace. Nevertheless, as experience showed the possibility and the reality of many smaller wars under the umbrella of the thermonuclear stalemate, war termination and peacemaking became the object of at least some studies in the social sciences, especially under the influence of the extremely difficult termination of the Vietnam War (cf., e.g., Fox 1970). In addition, there was the older tradition of more historically oriented research. Hentig (1952) dealt with the general modalities of peacemaking, while Fisch (1979) investigated the concepts of war and peace as they were developed in the peace treaties themselves, including the distinction between peacemaking proper and pacification. The most important basis for research in the social sciences became Wright (1965). After him, Raymond Aron's great survey of peace and war (1962) located peacemaking in the context of international studies and of the Cold War. Since the late 1960s there have been a number of studies dealing systematically with possible causes of peace, on the basis of the comparison of a greater or smaller number of war terminations, e.g., Randle (1973), Pillar (1983) and Holsti (1991) or, more formalized, Carroll (1969). On recent research see Massoud (1996). There exists, however, no theory of war termination or peacemaking, and it is unlikely that it can be developed, considering the extremely varied factors which may combine to bring about peace. The same holds for research into the conditions of maintaining a peace once concluded. Research in international law has particularly dealt with the question of the outlawry of war in the twentieth century, but less with its consequences for peacemaking. There is a challenge for new research in the growing importance of international peacekeeping, peace-enforcement and peacebuilding during recent years. For an introduction, see Otunnu and Doyle (1998). But the main problem remains the contradiction between the outlawry of war and situations of stalemate which will go on requiring traditional peacemaking as long as there is no world-wide monopoly of power. In this context, it will be important to compare the present methods of peacemaking and peace-enforcement with the traditional forms of peacemaking and pacification.

See also: Cold War, The; Conflict and Conflict Resolution, Social Psychology of; Conflict/Consensus; First World War, The; Military and Politics; Military History; Peace; Peace and Nonviolence:

Anthropological Aspects; Peace Movements; Revolution; Revolutions, Theories of; War, Sociology of; Warfare in History

Bibliography

Aron R 1962 *Paix et Guerre Entre Les Nations*, 3rd edn. Calmann-Lévy, Paris [Engl. transl. 1966 *Peace and War. A Theory of International Relations*. Weidenfeld and Nicolson, London]

Carroll B A 1969 How wars end: An analysis of some current hypotheses. *Journal of Peace Research* **6**: 295–321

Fisch J 1979 *Krieg und Frieden im Friedensvertrag*. Klett-Cotta, Stuttgart, Germany [War and Peace in the Treaty of Peace]

Fox T R 1970 How wars end. *Annals of the American Academy of Political and Social Science* **392**

Franklin B 1938 *Indian Treaties Printed by Benjamin Franklin* (with an introduction by van Doren C and bibliographical notes by Boyd J P). The Historical Society of Pennsylvania, Philadelphia, PA

Hentig H von 1952 *Der Friedensschluss. Geist und Technik einer verlorenen Kunst*. Deutsche Verlags-Anstalt, Stuttgart, Germany [The Conclusion of Peace. Spirit and Technique of Lost Art]

Hinsley F H 1963 *Power and the Pursuit of Peace. Theory and Practice in the History of Relations Between States*. Cambridge University Press, Cambridge, UK

Holsti K J 1991 *Peace and War: Armed Conflicts and International Order 1648–1989*. Cambridge University Press, Cambridge, UK

Kant I 1968 Zum ewigen Frieden. In: Weischedel W (ed.) *Id., Werke in Zehn Bänden*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany Vol. 9, pp. 191–251 (original work published in 1795) Engl. transl. 1991 Perpetual Peace. In: Reiss, H (ed.)*Id., Political Writings*. Cambridge University Press, Cambridge, UK pp. 93–130.]

Laszlo E, Yoo J Y (eds.) 1986 *World Encyclopedia of Peace*, 1st ed. (4 Vols). Pergamon, Oxford, UK

Massoud T G 1996 War termination. Review essay. *Journal of Peace Research* **33**: 491–6

Osiander A 1994 *The States System of Europe, 1640–1990. Peacemaking and the Conditions of International Stability*. Oxford University Press, Oxford, UK

Otunnu O A, Doyle M W (eds.) 1998 *Peacemaking and Peacekeeping for the New Century*. Rowman & Littlefield, Lanham

Phillimore W G F 1917 *Three Centuries of Treaties of Peace and Their Teaching*. J. Murray, London

Phillipson C 1916 *Termination of War and Treaties of Peace*. T Fisher Unwin, London

Pillar P R 1983 *Negotiating Peace. War Termination as a Bargaining Process*. Princeton University Press, Princeton, NJ

Randle R F 1973 *The Origins of Peace. A Study of Peacemaking and the Structure of Peace Settlements*. Free Press, New York

Touval S, Zartman I W (eds.) 1985 *International Mediation in Theory and Practice*. Westview Press, Boulder, CO

Wright Q 1965 *A Study of War. With a Commentary on War Since 1942*, 2nd edn. University of Chicago Press, Chicago

J. Fisch

Pearson, Karl (1857–1936)

Most of Karl Pearson's scientific work concerned human beings—he studied inheritance, physical anthropology, and disease—but his significance for the social and behavioral sciences is as a founder of modern statistics. Through the nineteenth century there had been a sense that the astronomers' methods for treating observations—the theory of errors—could be applied more generally. However, the modern conception of the scope and organization of statistics dates from the turn of the twentieth century when Pearson developed methods of wide applicability and put together the components of a discipline—the university department, courses, laboratories, journals, tables. His contemporaries may have done some of these things better but none had so much total effect.

1. Career

Karl Pearson was proud that his ancestors were of 'yeoman stock' though his father was a barrister. At Cambridge Karl read mathematics. Third place in the examinations brought him a college fellowship and the financial freedom to pursue his very wide interests. Pearson qualified as a barrister and studied social, philosophical, and historical issues. He developed his own view of man's place in a post-Christian world, expressing this in a novel and a play as well as through essays; some of those were published in the *Ethic of Freethought* (1888).

In 1884 the essayist and freelance lecturer became professor of Applied Mathematics and Mechanics at University College London. Besides carrying out the usual duties, Pearson took on two books left unfinished when their authors died. W. K. Clifford's *Common Sense of the Exact Sciences* explained the nature of mathematics and it would link Pearson's early general essays to the *Grammar of Science* (1892). The *Grammar of Science* presented the scientific method as 'the orderly classification of facts followed by the recognition of their relationship and recurring sequences.' In later editions the positivist ideas about space and force were joined by an exposition of statistical ideas. The other book, Isaac Todhunter's *History of the Theory of Elasticity*, was the history of Pearson's own specialism. Pearson actually wrote much more of this huge work than Todhunter himself. Pearson's later historical researches—a biography of Galton and lectures on the history of statistics—combined Todhunter's thoroughness with Pearson's own very different awareness of broad intellectual currents.

Pearson moved from elasticity and philosophy to biometry and statistics under the influence of W. F. R. Weldon, appointed to the college in 1890 as professor

of Zoology. Weldon was applying Galton's statistical methods to species formation and Pearson helped him extend the methods. However, Pearson was not just a mathematical consultant, he became a contributor to the biological literature in his own right. The partnership with Weldon was central to Pearson's development and he made his most important contributions to statistical thinking in these early years with Weldon. Pearson wrote over 300 pieces after 1900 but they were on lines he had already established. Of course these works demonstrated what statistical methods could achieve and were part of making statistics a recognized discipline.

In 1901 Pearson, Weldon and Galton founded *Biometrika*, a 'Journal for the Statistical Study of Biological Problems' which also published tables and statistical theory. From 1903 a grant from the Worshipful Company of Drapers funded the Biometric Laboratory. The laboratory drew visitors from all over the world and from many disciplines. Pearson wrote no textbook or treatise but the visitors passed on what they had learnt when they returned home. In 1907 Pearson took over a research unit founded by Galton and reconstituted it as the Eugenics Laboratory. This laboratory researched human pedigrees but it also produced controversial reports on the role of inherited and environmental factors in tuberculosis, alcoholism, and insanity—great topics of the day. In 1911 a bequest from Galton enabled Pearson finally to give up responsibility for applied mathematics and become Professor of Eugenics and head of the Department of Applied Statistics. During World War One, normal activities virtually ceased but afterwards expansion was resumed. However, Pearson was no longer producing important new ideas and Ronald Fisher was supplanting him as the leading figure in statistics. Pearson retired in 1933 but he continued to write and, with his son Egon. S. Pearson, to edit *Biometrika*.

Pearson made strong friendships, notably with Weldon and Galton. He also thrived on controversy. Some of the encounters produced lasting bitterness, those with William Bateson and Ronald Fisher being particularly unforgiving. Following the 'rediscovery' of Mendel's work in 1900, Bateson questioned the point of investigating phenomenological regularities such as Galton's law of ancestral heredity. The biometricians replied that such regularities were established and that any theory must account for them. After Weldon's death in 1906, the biometricians were often miscast as mathematicians who knew no biology. The quarrel about the value of biometric research faded with Bateson's death and Pearson's withdrawal from research into inheritance. Fisher, like Bateson, was a Mendelian but there was no quarrel about the value of statistics. Fisher criticized the execution of Pearson's statistical work but he also considered himself a victim of Pearson's abuse of his commanding position in the subject.

2. *Contributions to Statistics*

Pearson was a tireless worker for whom empirical research and theoretical research went hand in hand. His empirical work was always serious. He wanted to make a contribution to knowledge in the field to which he was applying his methods not just to show that they 'worked,' but so he would set out to master the field. However, his contribution to statistical methodology overshadowed his contribution to any substantive field. His statistical contributions can be divided into ways of arranging numerical facts—univariate and multivariate—and methods of inference—estimation and testing.

3. *Univariate Statistics: The Pearson Curves*

Pearson's first major paper (1894) was on extracting normal components from 'heterogeneous material.' However, Pearson considered 'skew variation in homogeneous material' a more serious problem for he held that the normal distribution underlying the theory of errors was seldom appropriate. Pearson (1895) presented a system of frequency curves; these would be of use in 'many physical, economic and biological investigations.' Systems were also developed by Edgeworth, Charlier, Kapteyn, and others, but Pearson's was the most widely used. Other systems have been devised, inspired by his, but none has been as central to the statistical thought of its time.

From probabilistic considerations—which now seem contrived—Pearson obtained a basic differential equation governing the probability density function, or frequency curve, $y = y(x)$:

$$\frac{dy}{dx} = \frac{y(x+a)}{b_0 + b_1 x + b_2 x^2}$$

The constants determine the form of the curve. Pearson developed methods for choosing an appropriate curve and for estimating the constants. His laboratory produced tables to ease the burden of calculation.

While Pearson used these curves as data distributions, he assumed that the estimators of the constants would be approximately normal, certainly in large samples. In the work of 'Student' (W. S. Gosset) and then Fisher the data distributions are normal but the exact distributions of the sample statistics are usually not, though *they* may follow the Pearson curves. 'Student' did his distribution work in Pearson's laboratory but for many years Pearson thought that such 'small sample' work was off the main track and useful only to 'naughty brewers.'

In the Pearson system the normal distribution is just one type. In Fisher's analysis of variance which appeared in the 1920s it is the *only* type. People

schooled in the Pearson system did the earliest robustness studies and one of the continuing roles of the system has been in this area. While Pearson's skepticism towards normality pervades much modern work, the modern approach to inference is so different from his that he is not a direct inspiration.

4. *Multivariate Statistics: Correlation and Contingency*

In 1888 Galton introduced correlation and soon Edgeworth was developing the theory and Weldon was applying the method in his research on crustaceans. Pearson, however, produced the definitive formulation of normal correlation and had the vision that correlation would transform medical, psychological, and sociological research.

Pearson (1896) set out to frame hypotheses about inheritance in terms of the multivariate normal distribution but these were of less enduring interest than his account of correlation and multiple regression. The familiar partial and multiple correlation and regression coefficients appear at this time either in Pearson's work or that his assistant, G. Udny Yule. Soon Pearson and those around him were applying correlation in meteorology, sociology and demography as well as in biometry.

The 1896 ideas formed the basis of a definitive theory, especially when completed by the exact distribution theory of 'Student' and Fisher. But Pearson soon moved on to qualitative, non-normal, and time series data, though not with the same success. Pearson's scheme for contingency, the qualitative analogue of correlation, has latent continuous variables—surprisingly jointly normal ones. Yule, no longer working with him, rejected the scheme. The controversy that ensued had no clear resolution. Pearson toiled unsuccessfully to create a theory of skew surfaces, uniting correlation and skew curves. Time series correlation analysis was another disputed field. Pearson thought of the true correlation between series as obscured by time trends superimposed upon them. The scheme was widely used but in the 1920s Yule criticized it effectively, yet putting nothing in its place.

From the beginning Pearson encountered problems with correlation which have lived in the underworld of statistics occasionally surfacing in fields such as path analysis and structural modeling. Pearson formalized Galton's notion of correlation as measuring the extent to which variables are governed by 'common causes.' He was soon aware of cases of 'spurious correlation' where the natural interpretation is not sustainable. He and Yule—who developed the other standard interpretation of a direct causal relationship between the correlated variables—found numerous pathological cases. They devised treatments but developed no systematic way of thinking about the problem.

Pearson—following Galton—established multivariate analysis. Yet by a curious twist the multivariate data for which Pearson struggled to create a theory of frequency surfaces is often treated by Fisher's 'regression analysis' which was a development from Gauss's univariate theory of errors rather Pearson's correlation theory.

5. *Estimation: The Method of Moments and Inverse Probability*

Pearson usually found the constants of frequency curves by the method of moments. He introduced this in his first major paper (1894) for he could find no other method for estimating a mixture of normals. Pearson went on to apply the method quite generally. It seemed particularly appropriate for the Pearson curves where the constants determine the first four moments. However, the method was put on the defensive after Fisher criticized it for not being 'efficient'—even when applied to the Pearson curves.

Probable errors (scaled standard errors corresponding to the 50 percent point) for method of moments estimators were derived by Pearson and Filon (1898) using a modification of an argument in the 1896 correlation paper. Pearson had obtained the product-moment estimate from the posterior distribution (based on a uniform prior) and its probable error from a large sample normal approximation to the posterior. In 1898 he assumed that the same technique would always yield probable errors for the method of moments. He seems to have dropped the technique when he realized that the formulae could be wrong but it was reshaped by Fisher into the large sample theory of maximum likelihood.

In 1896/8 Pearson used Bayesian techniques in a formalistic way but he could be a thoughtful Bayesian. Over a period of 30 years he wrote about the 'law of succession,' the standard Bayesian treatment of predicting the number of successes in future Bernoulli trials given the record of past outcomes—'the fundamental problem of practical statistics' he called it. Pearson held that a prior must be based on experience; a uniform prior *cannot* just be assumed. Pearson pressed these views in estimation only once when he criticized Fisher's maximum likelihood as a Bayesian method based on an unsuitable prior—essentially Pearson's own method of 1896/8. Fisher was offended by this misinterpretation of his Bayes-free analysis.

There was plenty of improvisation in Pearson's estimation work; it was eclectic, even inconsistent. He was not the consistent Bayesian that Jeffreys would have liked. Pearson used sampling theory probable errors but he did not follow up the results on the properties of estimators in the old theory of errors. Pearson was not unusual in his generation; the theory of errors he found to hand-mixed frequentist and Bayesian arguments. Some of his contemporaries were

more acute but foundational clarity and integrity only became real values towards the end of his career when Fisher, Jeffreys, and Neyman set about eliminating the *wrong* kind of argument.

6. Testing: Chi-squared

Pearson's earliest concerns were with estimation—curve fitting and correlation—though there testing could be done, using the estimates and their probable errors and making the assumption of large sample normality. However, his chi-squared paper of 1900 introduced a new technique and gave testing a much bigger role.

The chi-squared test made curve fitting less subjective. One could now say: 'In 56 cases out of a hundred such trials we should on a random selection get more improbable results than we have done. Thus we may consider the fit remarkably good.' The tail area principle was adopted without any examination. Naturally Pearson enjoyed showing that the fit of the normal distribution to errors in astronomical measurements was *not* good.

Pearson brought chi-squared into his treatment of contingency and by 1916 he had established the main applications of the chi-squared distribution—a test for goodness of fit of distributions, a test for independence in contingency tables, a test for homogeneity of samples, a goodness of fit test for regression equations.

Pearson's distribution theory involved the exact distribution of the exponent in the multinormal density, this density appearing as a large sample approximation to the multinomial. Pearson slipped in the mathematics and used the same approximation whether the constants of the frequency curves are given *a priori* or estimated from the data. Fisher realized the importance of the difference and reconstructed the theory, finding that the chi-squared approximation was valid but the number of degrees of freedom had to be altered.

Fisher continued Pearson's chi-squared work and gave significance testing an even bigger role in applied statistics. Pearson's methods were more or less vindicated by the later work of Fisher and then of Neyman and E. S. Pearson but, as in estimation, his reasons did not pass critical examination.

7. Summary

Of the many statistical techniques Pearson devised, only a few remain in use today and though his ideas sometimes find re-expression in more sophisticated form, such as the correlation curve or the generalized method of moments, there is little to suggest that Pearson continues to directly inspire work in statistics. Pearson broke with the theory of errors but in the next generation through the analysis of variance and regression the theory was restored to stand beside, even to overshadow, Pearsonian statistics. From a modern perspective Pearson's theory seems desperately superficial. Yet the problems he posed have retained their importance and the ground he claimed for the discipline of statistics has not been given up. Pearson was an extraordinarily prolific author and there is also a considerable secondary literature. There is a brief guide to this literature on the website http://www.soton.ac.uk/~jcol

See also: Multivariate Analysis: Overview; Probability: Formal; Significance, Tests of; Statistics, History of

Bibliography

Eisenhart C 1974 Karl Pearson. *Dictionary of Scientific Biography* **10**: 447–73

Hald A 1998 *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York

MacKenzie D A 1981 *Statistics in Britain 1865–1930: The Social Construction of Scientific Knowledge*. Edinburgh University Press, Edinburgh, UK

Magnello M E 1999 The non-correlation of biometrics and eugenics: Rival forms of laboratory work in Karl Pearson's career at University College London, (in two parts). *History of Science* **37**: 79–106, 123–150

Morant G M 1939 *A Bibliography of the Statistical and other Writings of Karl Pearson*. Cambridge University Press, Cambridge, UK

Pearson E S 1936/8 Karl Pearson: An appreciation of some aspects of his life and work, in two parts. *Biometrika* **28**: 193–257; **29**: 161–247

Pearson K 1888 *The Ethic of Freethought*. Fisher Unwin, London

Pearson K 1892 *The Grammar of Science* [further editions in 1900 and 1911]. Walter Scott, London [A.& C Black, London]

Pearson K 1894 Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*. **185**: 71–110

Pearson K 1896 Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society A*. **187**: 253–318

Pearson K, Filon L M G 1898 Mathematical contributions to the theory of evolution IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society A*. **191**: 229–311

Pearson K, Filon L M G 1900 On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*. **50**: 157–75

Pearson K, Filon L M G 1914/24/30 *The Life, Letters and Labours of Francis Galton*. Cambridge University Press, Cambridge, UK, Vols. I, II, IIIA, IIIB

Provine W B 1971 *The Origins of Theoretical Population Genetics*. Chicago University Press, Chicago

Semmel B 1960 *Imperialism and Social Reform: English Social-Imperial Thought 1895–1914*. George Allen & Unwin, London
Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press of Harvard University Press, Cambridge, MA

J. Aldrich

Peasants and Rural Societies in History (Agricultural History)

The concept 'rural societies' is based on the fundamental division in our society between rural and urban, but usually refers to societies dominated by rural areas. It thus has a historical dimension, and is connected with the concept of peasant societies.

1. Terminology

An intense discussion about the concept of peasant studies began in the 1950s. A large body of literature was published in the decades that followed. The discussion centered on the Anglo Saxon world since, unlike English, most other languages do not contain the same clear distinction between peasant and farmer. In German, for example, *Bauer* can mean both peasant and farmer (Rösener 1993). The need of a distinguishing term in English is also connected with the fact that peasant in English (like *paysans* in French) has a derogatory meaning which the German word *Bauer* (or the Scandinavian *bonde*) does not have.

Several contributions in the discussion about peasants and farmers were made from the 1950s until the beginning of the 1970s (important monographs: Redfield 1956, Wolf 1966, and readers: Potter et al. 1967, Shanin 1971). In recent decades, the concept of peasant has been questioned even in the Anglo Saxon world. For example, Tom Scott writes in an overview that the category of peasant 'is so broad as to become not so much an archetype as a stereotype' (Scott 1998, p. 1). However, the discussion which began during the post-war period, still forms a platform for the research on peasant and rural societies. Peasant society was defined on two levels; partly chronologically in relation to earlier and later types of social systems, and partly socially where the peasant society was defined in relation to the overall social structure of which it was a part.

In the case of the last-mentioned level of definition, the peasant society was seen as a part-society with a part-culture, i.e., a part of the large social structure. This formed the basis of the chronological separation of peasants from what was called the 'tribal society' (or even, but not any longer, the 'primitive society'). Such a society was by definition an earlier system, not being a part of a larger social unit. The emergence of the town was identified as a decisive change and resulted in peasants either choosing or being forced to send some of their produce to the town. The establishment of the State and the state authorities has also been identified as a distinguishing factor. With the advent of the State, the peasant society became clearly subordinate to, but also alienated from, a larger social system as such. To some extent, both these criteria can be said to represent the same change. The peasant society has also been described as a part of an overall socioeconomic feudal system, but every peasant society can hardly be seen as part of a feudal system (see *Feudalism*).

At the beginning of the 1970s, Teodor Shanin (in his *Introduction* to Shanin 1971) summarized the discussion and pointed at some of the characteristics that distinguished a typical peasant society: the family farm as the basic unit of multidimensional social organization; land husbandry as the main means of livelihood directly providing for the major part of the consumption needs; specific traditional culture related to the way of life of small communities; and the underdog position, specified as the domination of peasantry by outsiders. Shanin adds that, like every social entity, the peasantry must be seen as undergoing change as a part of a social process, which influences the definition of the concept peasant.

The chronological border, which separates peasants from the modern farmer has also been discussed. The modern farmer not only produces for a market, but all his actions are profit driven. He acts in the same way as other businessmen. The business itself has become increasingly industrial in nature. Different aspects were emphasised in the discussion. Wolf points to the relationship to the market as being decisive. Redfield emphasizes peasant tradition. The Marxists, who later entered the discussion, wrote that when the peasants acted as a group with class consciousness, they became a class in themselves. The Russian peasant researcher Chayanov, had been purged in 1930 by Stalin but his writings from the 1920s, which got a renaissance long after his death, emphasized the preservation of the family as being of central importance to the peasant (Chayanov 1966).

The discussion illustrates the complexity of the concept, and much of the discussion focused on the problem of definition. Could the specific African cultivator be categorized as peasant, or the Latin-American pastoralist? The fruitfulness of discussing such detailed classifications is questionable. When the core of the term has been identified, the borders drawn must be flexible to allow for interpretations, which in themselves include a dynamics. During the discussion, it also became increasingly clear that the specific historical situation must be taken into account. Rodney Hilton pointed at the essential elements in defining the English medieval peasant, but also made

clear that this definition was specific for a part of Europe, where the peasantry must be seen in the context of the medieval feudal society (Hilton 1975).

The concept of peasant will be used in this article in a very comprehensive sense, which includes those who raise crops or cattle, on a partly nonprofitable base, and are integrated in larger social structures. The definition includes peasantries in all societies where the production of foodstuffs requires the major part of the society's labor. It is debatable whether there could be peasants in societies where cultivators comprise a diminishing share of the population. Today in most countries outside the Western World, peasants continue to be tied to the soil and to the local community, with family, relatives and traditions, in a way that characterizes them as peasants. Also in a large part of Europe, there are still cultivators who live as traditional peasants. They live surrounded by their ancestors' history in houses, land, and tales and they are bound to their own farm by family-traditions. This should not be underestimated as a fact that for instance decides the course of contemporary talks on agriculture and world trade between Europe and the US.

Accordingly, the term peasant also has some relevance in today's society, and this is the explanation of why many languages have not developed the terminological difference found in English between peasant and farmer. However, the difference between peasant and farmer identifies an important historical change, and the peasant society is a purely historical concept for the Western World. Rural society should be regarded as a broader unit comprising, in addition to cultivators, craftsmen and rural industry. Rural society thus contains the seeds of proto-industrialism, and the concept is more relevant to today's conditions.

2. *Historiography*

To understand the discussion about the peasant and rural societies, it must be placed in a historiography, which both precedes and follows the postwar discussion. Rural society as a historical concept must be related to agricultural history, which mainly concerns agricultural production, but also includes other elements related to agriculture, such as politics and village communities. In rural history, the emphasis, formally speaking, is placed more on the societal aspect and in the case of agricultural history, more on the history of production and distribution. But since agricultural production is the basic activity of all peasants, in practice, rural history and agricultural history are parts of the same historiography. Rural history and agricultural history are an important part of general history, as the farming population for centuries was in the majority, but despite this, rural history and agricultural history emerged as specific disciplines relatively late.

2.1 *The First Folk-life Studies*

Scientific research into peasants and rural history in Europe began at the end of the nineteenth century. In many parts of Europe, i.e., Eastern Europe and Scandinavia, interest focused on the countries' own peasants. In the UK, research primarily focused on peasants in the colonies. In countries where agriculture played a dominating role, agricultural history was linked to the creation of national identity, demonstrated in national 'agricultural museums.' In Hungary, what is still one of the world's most important agricultural museums was founded in conjunction with the country's millennial celebration in 1898. National agricultural museums were also established in the Czech Republic and Denmark at the end of the nineteenth century.

In Eastern Europe, Scandinavia, and Germany, folk-life studies focused on the peasants who were regarded as the real heart of the nation. Later, these studies also played a prominent role under the name of ethnology or ethnography in France, Portugal, Ireland, Scotland, and other countries, but never in England. Folk-life research studied both spiritual and material culture, and described tools and houses together with folk-songs and riddles in peasant societies. A large knowledge-base was built up.

The scientific question that crystallized concerned regionality. Even if the original question was couched in terms of national identity, the answer increasingly proved the opposite. The cultural phenomena moved freely across borders, there were no obvious and given national borders in Europe. The Swede Sigurd Erixon played an important role, and a number of national ethnographic atlases, where the focus was on peasant society and its manifestations, were published. There were plans to publish a common European ethnographic atlas, but today these plans have been put on ice.

In the countries with colonies, above all the UK, the study of peasant societies outside Europe became increasingly popular at the end of the nineteenth century. Originally, these studies were part of the colonial administration, e.g., in the genre of studies of the Indian countryside, which the British Empire created as early as in the middle of the nineteenth century. Eventually social anthropology became a discipline in its own right, less and less connected with the colonial administration. In the twenty-first century it has a strong position in the Anglo Saxon world, not least in the US. Like European ethnography, anthropology was linked to the museums and the decades around the turn of the century were spent gathering data.

The goal of these studies was to understand the local community, in all its details but also its internal structure. The focus was on the individual village or region as an isolated entity. The level of detail could also be very high in the case of the material culture, but

the focus was often on the relationships between the people, e.g., family relationships. But even though the research concentrated on clarifying the internal structure of the local society, the studies gradually pointed to a different result. It was increasingly realized that the local society was integrated with, and subordinate to, larger social structures. This was one of the most important reasons why parts of the Anglo Saxon world of anthropology started using the term 'peasant' instead of 'people,' and it lead to the above mentioned debate on peasant societies.

Early on, the combination of ethnology and anthropology resulted in impressive surveys. The German researcher Paul Leser, for example, published an authoritative survey in 1931 of different types of ploughs in different parts of the world (*Entstehung und Verbreitung des Pfluges*, Munster). Not long afterwards, Leser, who was Jewish, was forced to emigrate. Folk-life studies in Germany increasingly became a part of the nationalistic ideology in the 1930s, and was concentrated on studies about the spirit and 'soul' of the people.

2.2 Agricultural History in the Interwar Period

The US was still regarded by its inhabitants as an agricultural nation into the twentieth century. In the democratic and popular tradition that was this nation's identity, the history of agriculture played a role in the explanation of its progress and success, and agricultural history crystallized early into a discipline in its own right. The oldest national journal on the subject, *Agricultural History* (1927–), was founded in the US. From the very beginning, this journal contained articles about conditions outside the US, but it has focused mainly on American agricultural history. Initially, many of the articles discussed progress in agriculture at the end of the nineteenth century and the beginning of the twentieth century. A special genre consisted of articles about diaries kept by peasants.

In Europe, the first wave of interest in the history of agriculture, in ethnography and anthropology, focused on the immediate past. There were few attempts made to create a broader time perspective, as the discipline of history was still oriented towards the history of the State and the elite. Rural history developed in France under the influence of geography, where a dominating group of scholars favored all-embracing studies where every aspect of a region should be described, including historical aspects. With geography's starting-points in the land, these studies also touched on rural history.

The historian who succeeded in giving rural history a scientific basis was the Frenchman Marc Bloch in his *Les charactères originaux de l'histoire rurale francaise*, which was published in 1931 (Bloch 1966). This was based on a series of lectures he had held in Oslo in 1929 at a research institute, which had been established to promote and carry on Scandinavian folk-life studies. Here the different threads meet. In the words of the American agrarian historian Richard Herr, Bloch was able to provide 'the kind of insights that would spark the imagination of other historians' (Herr 1993, pp. 4–5). In a subsequent study of European feudalism, Bloch explained how a rural society is part of a large social structure (Bloch 1961, in French 1939–40). He thus anticipated the discussion about peasant societies during the postwar period.

Bloch's work still provides inspiration and guidelines in a number of areas: the role of the village community, peasant uprisings, the connections between technology and social change, etc. But years passed before his books were considered to be important. General acceptance did not come until after World War II. In part, this was a result of the myths created about Bloch after he had joined the French Resistance to fight the Germans and was executed in 1944.

2.3 Agricultural History's Breakthrough after World War II

The reason for the rise in appreciation of Bloch's work should, however, be explained in terms of the large changes that occurred in the study of rural societies and agricultural history. During the 10-year period following the end of World War II agricultural history became a discipline in its own right in Western Europe. Departments of Agricultural History were established in several European countries: in the Netherlands at the agricultural university in *Wageningen*, in Germany at the agricultural university in *Stuttgart* and in the UK at the Faculty of Agriculture at *Reading* University.

National journals were started up in Europe: *The Agricultural History Review* (1953–) in the UK, *Historia Agriculturae* (1953–) in the Netherlands, *Zeitschrift für Agrargeschichte und Agrarsoziologie* (*ZAA*) (1953–) in Germany, *Bol og by* (1956–) in Denmark, *Rivista di storia dell' agricoltura* (1960–) in Italy and *Études rurales* (1961–) in France. In time, the breakthrough for agricultural history resulted in national synthesis covering several volumes. In the UK, a huge project in eight volumes was started in 1967, *The Agrarian History of England and Wales*. This work has yet to be completed. The German synthesis, *Deutsche Agrargeschichte*, consists of six volumes and was published between 1962 and 1984. In France, *Histoire de la France rurale* was published in four volumes between 1975 and 1978. Although the French volumes emphasise rural history in the title, there is no major difference between this and the other national works. In the German, two volumes are devoted to purely social questions, i.e., legal questions and peasants as a social category.

This wave of research resulted in a focus on the long-term change, and not least on the Middle Ages. At the beginning of the 1960s, two important works

were published. One of these was Georges Duby's *L'économie rurale et la vie des campagnes dans l'Occident médiéval*, published in 1962, which was a study in the tradition of Bloch (in English 1968). The following year, 1963, saw the publication of an English translation of a study published in Dutch some years earlier. Slicher van Bath's *The Agrarian History of Western Europe AD 500–1850*.

During the postwar period the agricultural crisis in the late Middle Ages became a central field of research. This economic decline had been observed before World War II (i.e., for Germany, Abel 1980). The discovery was theoretically significant since it showed that the development of production had not always been positive. Another important theoretical change was that agriculture was seen as a dynamic sector in medieval Europe, which partly rejected earlier theories about the countryside as having a restraining effect on economic development. The aftermath of World War II continued to leave its mark to some extent. Slicher van Bath, who had built up the Department of Agrarian History at the University of Agricultural Sciences in Wageningen in the Netherlands, mainly collaborated with Anglo Saxon researchers and not with the German, e.g., at the agricultural university in Stuttgart-Hohenheim. The reason for this was that Gunter Franz, who had built up the Department of Agricultural History in Stuttgart, was not only a leading expert on the German Peasant War, but also had been an active Nazi. However, the journal founded by Franz, *ZAA*, published early on historiographical surveys of agricultural history in Europe. These surveys were written by Duby, Postan, and other leading European agrarian historians. It must also be stated that several leading agrarian historians in Germany during the 1930s never became engaged in the Nazi movement, for instance, Wilhelm Abel and Friedrich Lütge.

2.4 The Transformation of Folk-life Studies after World War II

The discussion about the concept of 'peasant' resulted in a number of studies of peasant and rural societies. These include the classic studies made by Oscar Lewis in Mexico and India. Daniel Thorner's studies of peasants in India are another example (a bibliography in: *Peasants in History: Essays in Honour of Daniel Thorner*, Hobsbawm (ed.) 1980). Several studies were also carried out in Europe, often by American researchers who used anthropological methods to study villages in Southern Europe. They were somewhat surprised to meet an entirely different type of cultivator from those they usually met in the US. It was still peasants and not farmers who dominate the European villages. The classic studies here include those by Julian A. Pitt-Rivers, who studied a Spanish village (1954), and by Joel Halpern, who studied a Serb village (1958).

In anthropological research, the large databases, e.g., the Human Relations Area Field (HRAF) in Yale has been gathering information on several hundred cultures since the 1930s. But the previous focus in anthropological research on material cultural products has weakened. True, encyclopedic studies were still being made, for example, by the Frenchman Leroi-Gourhan (1943–5), but this was more and more pushed into the background. Instead, it was the social structure that attracted increasing attention, and this is noticeable not least in the discussion about peasant societies. It was still the hidden patterns that research wanted to be clarified, but now as part-cultures subordinate to the principal culture. Clifford Geertz, had made an important description of the history of agrarian technology in Indonesia (*Agricultural Involution* 1964), became even more influential as a father of the concept 'thick description,' which became a leading method in anthropology (see articles in Geertz 1973).

In Eastern Europe, folk-life studies retained and, to some extent, strengthened their, role when the Socialist people's republics were established. Research into 'the people' was emphasized as one of the main tasks of historical research. At the same time, contacts with the West were broken off. An example is how seldom the German researchers in the East and West referred to each other, even though they often made research about the same subjects.

Some contacts were however upheld. *Association International des Musées d'Agriculture (AIMA)* gradually became the meeting place between East and West for the study of rural society. AIMA was an association of agricultural museums all over Europe as well as in the US, which was formally under the jurisdiction of UNESCO. The association was formed in 1974 and its meetings were held alternately in Eastern and Western Europe (or the US). Many of the questions discussed were classic in European ethnography, e.g., the history of the plow. After the Communist collapse, this association has had a difficult time.

2.5 Reasons for the Breakthrough after World War II

There are many underlying reasons for the breakthrough of rural history in the 1950s. In the Netherlands, one of the reasons given for the establishment of a Department of Agricultural History was that the severe food shortage at the end of World War II illustrated the importance of agriculture. In all social sciences the American influence increased after World War II, not the least in sociology of agriculture and in studies of non-European peasants. This influence may also have had an impact on agricultural history.

A factor, which can hardly be underestimated, was the increased political importance of peasants on a worldwide scale and in conjunction with colonies

attaining independence, but the immediate conclusion was not that scholars regarded the peasants as decisive actors in history.

Peasants and rural societies were identified as essential elements of the social structure. Initially, the new research focused on concepts and ideas from the earliest period of folk-life studies in Europe, and emphasized that peasant culture had filtered down to the peasants from the elite. Villagers had taken over and simplified traditions from the towns or from the elite. The goal of the peasants was to imitate the leaders of society. The ideological basis for this interpretation was that society's leaders were considered to represent a higher form of civilization. The interpretation was made in more or less explicit forms by almost all the scholars participating in the discussion after World War II and in particular by George Foster (1967). Robert Redfield (1956) distinguishes between a great tradition and a little tradition, between the classical, learned elite culture, and the peasant tradition. He emphasizes that a mutual exchange takes place, where also the elite draw inspiration from below.

The view of the conservative peasant society which was unwilling to accept changes that upset traditions is similar to what is sometimes called 'the dumb peasant theory.' This implied that peasants had to be forced to develop, and this assumption guided much of the Western World's relations with the Third World throughout the 1950s and 1960s.

2.6 New Interpretations in the 1970s and 1980s

Beginning in the 1970s, studies of peasants and rural societies increased in number and became more diversified. Two partly connected phenomena began to change the picture. One was the Marxist challenge in the academic world, the other was the successful peasant uprisings in different parts of the world, not least the civil war in Vietnam. These two phenomena challenged the old thought structures, but did not lead to a new hegemony. Instead, after the major theoretical battles in the 1970s, the field was left open to a large number of different theories. This can be compared with the change undergone by rural sociology at the same time. A survey from 1990 shows that what the authors call 'the new rural sociology' from the 1970s and 1980s was characterized by a large number of interpretations (Buttel et al. 1990).

The common question with which many of the new interpretation struggled was the relationship between actors and structure, where the peasants were increasingly regarded as real actors in the historical process. One expression of this new and deeper interest in peasant studies was the *Journal of Peasant Studies* (1974–). In its editorial policy, it was emphasized that the journal would be concerned with political economy and that it would attempt to understand the role of the peasants, especially outside Europe and North America.

At the same time we also see a growing interest for rural history in the third world, and for instance, a journal on the subject was founded in India: *History of Agriculture* (1973–).

The role of class conflicts became more important, which at times resulted in very detailed analyses of the social structure of the countryside. Studies of peasant revolts and peasant uprisings were developed into a special genre with a series of studies. Searches in international databases show that there is an interesting geographical distribution of the studies. Most of the studies of peasant revolts outside Europe concern Asia (India, China, Japan, South East Asia, and Indonesia) or Latin America (Mexico, El Salvador, Bolivia, and Brazil). Studies concerning Africa, on the other hand, began early on to be more focused on the environment, while studies of peasant revolts are less common.

An example of this new orientation is the rapidly growing interest in 'Maroon societies.' The fact that slaves in Latin America as well as in the American South were able to break free and establish societies, which in some cases were able to survive for hundreds of years partly outside the control of the authorities, added a new dimension to peasant society. The first comprehensive study of Maroon societies as a specific form of peasant societies in 1970s was followed by many more (Prize 1996). This history from below was also applied in other areas.

In the European research about long-term changes new interpretations emerged. The dominating researchers had initially been supporters of Malthusian interpretations, but other interpretations gradually began to be given greater emphasis. One example of this is the class perspective introduced by Guy Bois (1976) into the interpretation of the agrarian crisis in the late Middle Ages.

The older type of studies of material culture continued, not least in the exchange between East European and West European researchers in classic fold-life research. In Denmark, the journal *Tools & Tillage* was established in 1971 and became the leading international journal dealing with the development of agriculture technology in peasant societies.

2.7 Diversification at the End of the Century

The late twentieth century has seen a gradually growing interest in rural and agricultural history. New journals have been started up, including *Rural History* (1990–) in the UK, *Noticiaro de historia agraria* (1991–) in Spain, *Skrifter om skogs- och lantbruks-historia* (1991–) in Sweden, and *Histoire et sociétés rurales* (1994–) in France.

There are new elements in this growing interest. One that is very important is the interest in environmental questions. This has had an impact on agricultural history not least in the US in the form of studies comparing how Indians and Europeans used the environment in a social context (i.e., Merchant 1989). Specialist journals have also been established, the most important of which is *Environmental History* (1989–) in the US, formed by merging two journals with the same orientation, which had been published since the 1970s. The issues dealt with have concerned sustainability in a long-term perspective in relation to current problems such as soil erosion, salinization, etc.

Interest in popular culture and social structure is increasing, and one example is the research group around Jan Peters (Peters 1997). This group also represents a successful wedding between East and West in Germany in the late twentieth century. Another strong orientation, although not only in the study of rural and peasant societies, is gender studies. Technology research has also been revived although here, more specific questions are being raised concerning the connection between technological and social change.

3. Theories and Methods—a Summary

When the question of the definition of peasants and peasant societies was raised after World War II, it proved to be both fruitful as well as inspirational to further research. The connection between peasants and social structure was brought out more clearly. However, this was only one step in a long development process involving research, and rural and agricultural history. At the end of the nineteenth century and beginning of the twentieth century, researchers concentrated on gathering material. In Europe, individual cultural elements were registered, and outside Europe whole peoples and their characteristics were dealt with. Theories about isolated nations and communities were proposed, and abandoned.

In the 1950s larger structures were identified. In anthropology, peasant societies were regarded as being a part of larger society, and in history the long waves came into focus of the research. During a following phase, which began in the 1970s, research became increasingly diversified. Much of the research emphasized the capacity of peasants to take independent action, and new fields of research, such as environmental questions and gender research, were opened up. This eventually resulted in a new wave of interest about the subject at the turn of the twenty-first century.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Economic History; Feudalism; Food in Anthropology; Food Production, Origins of; Food Security; History of Technology; Peasants and Rural Societies in History (Agricultural History); Peasants in Anthropology; Rural Geography; Rural Sociology; Social History

Bibliography

Abel W E 1980 *Agricultural Fluctuations in Europe: From the Thirteenth to the Twentieth Centuries*. Methuen, London

Bois G 1976 *Crise du féodalisme*. Presses de la Fondation National des Sciences Politiques, Paris

Bloch M 1961 *Feudal Society*. Routledge & Kegan Paul, London

Bloch M 1966 *French Rural History*. Routledge & Kegan Paul, London

Buttel F H, Larson O F, Gillespie G W 1990 *The Sociology of Agriculture*. Greenwood Press, New York

Chayanov A V 1966 The theory of the peasant economy. In: Thorner D, Kerblay B, Smith R E F (eds.). American Economic Association Translation Series, Homewood, IL

Duby G 1962 *L'économie rurale et la vie des campagnes dans l'occident médiéval*. Aubier, Editions Montaigne, Paris

Foster G 1967 What is a peasant. In: Potter J M, Diaz M N, Foster G M (eds.) *Peasant Society, A Reader*. Little, Brown, Boston, MA

Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York

Halpern J M 1958 *A Serbian Village*. Harper Colophon Books, New York

Herr R 1993 The nature of rural history. In: Herr R (ed.) *Themes in Rural History of the Western World*. Iowa State University Press, Ames, IA

Hilton R H 1975 *The English Peasantry in the Later Middle Ages*. Clarendon Press, Oxford, UK

Hobsbawn E J (ed.) 1980 *Peasants in History: Essays in Honour of Daniel Thorner*. Oxford University Press, Calcutta

Leroi-Gourhan A 1943–45 *Évolution et techniques* 1–2. Paris

Leser P 1931 *Enstehung und Verbreitung des Pfluges*. Anschendorffsche Verlags buchhandlung, Münster

Merchant C 1989 *Ecological Revolutions: Nature, Gender, and Science in New England*. University of North Carolina Press, Chapel Hill, NC

Peters J (ed.) 1997 *Gutsherrschaftsgesellschaften im europäischen Vergleich*. Akademie, Berlin

Pitt-Rivers J A 1954 *People of the Sierra*. University of Chicago Press, Chicago

Potter J M, Diaz M N, Foster G M (eds.) 1967 *Peasant Society, A Reader*. Little, Brown, Boston, MA

Prize R 1996 Preface. In: Prize R (ed.) *Maroon Societies: Rebel Slave Communities in the Americas*. Johns Hopkins University Press, Baltimore, MD

Redfield R 1956 *Peasant Society and Culture: An Anthropological Approach to Civilization*. The University of Chicago Press, Chicago

Rösener W 1993 *Die Bauern in der europäischen Geschichte*. C H Beck, Munich, Germany

Scott T 1998 Introduction. In: Scott J (ed.) *The Peasantries of Europe from the Fourteenth to the Eighteenth Centuries*. Longman, London

Shanin T (ed.) 1971 *Peasants and Peasant Societies*. Penguin, Harmondsworth, UK

Slicher van Bath B H 1963 *The Agrarian History of Western Europe AD 500–1850*. Edward Arnold Limited, London

Wolf E 1966 *Peasants*. Prentice-Hall, Englewood Cliffs, NJ

J. Myrdal

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Peasants in Anthropology

Peasants are small-scale agriculturalists who use simple technology such as hand tools, human labor, animal power, and possibly light machinery. The household is the basic unit of production and consumption. While some peasants grow crops or raise animals for markets, they usually consume most of what they produce and thus differ from farmers who produce primarily cash crops. Also, peasants typically exist in subordinate political and economic relationships with other sectors of society who receive surplus products and labor from them. From the rise of civilizations based on non-mechanized agriculture until the widespread use of modern industrial farming, peasants in rural areas have provided much of the food and other basic material resources consumed in urban society. Furthermore, peasants have provided recruits for armies throughout history, and until fairly recent times in developed nations, cities typically had high death rates relative to birth rates and thus had to be replenished by peasants migrating from rural areas. Indeed, most of the rapid growth in world population in the nineteenth and twentieth centuries was due to continued high birth rates among peasants, and correspondingly lowered death rates due to the spread of modern medicine and public health.

1. *Peasants in the History of Anthropology*

From its modern beginnings in the 1800s until the early to mid-twentieth century, anthropology focused largely on foraging, tribal, and chiefdom societies. It was not until after World War II that peasant communities became a major object of anthropological research. This shift of interest by anthropologists toward peasant communities was due to changing global conditions. Foraging and tribal societies were disappearing at ever increasing rates, and in addition to this, following the War, peasants emerged on the world stage as major political actors in the context of the Cold War. Within this global struggle between the so-called 'First World' of capitalist societies and the 'Second World' dominated by the communist Soviet Union and China, many underdeveloped 'Third World' countries were inhabited mainly by rapidly growing peasant populations. The political significance of these nations grew as concern mounted as to whether their peasant populations would veer toward communism or capitalism. Since the poverty of peasant communities was seen as a major inducement for them to turn to communism, promotion of economic development in rural areas of the 'Third World' became a priority of US foreign policy. In this geopolitical context, the anthropology of peasant communities flourished from the 1950s until approaching the end of the Cold War in the early 1990s (Kearney 1996). Much of this work was applied anthropology concerned with the improvement of agriculture, health, and education in peasant communities.

2. *Anthropological Approaches to Peasants*

A distinctive feature of anthropology is that it seeks to study and comprehend all of the major components of a community and how they are integrated. Ideally, this comprehensive method begins with examination of the physical environment of the community and how its inhabitants adapt to that environment with technology and a distinctive division of labor with which they produce food, shelter, and amenities. These features of the environment—technology, labor, and the material products of labor—are parts of a community's infrastructure, upon which are based an immaterial superstructure that comprises a corresponding patterning of social organization, politics, class positions and relations, gender identities, religion, world view, etc. Ideally, a robust anthropological treatment of a peasant community examines all of these material and immaterial components of a community and how they are interrelated. Furthermore, since peasants, by definition, typically exist within and are shaped by relations of unequal economic exchange with non-peasant communities, a complete anthropological study of a peasant community would also examine this dimension. In practice, however, few anthropologists attain such comprehensiveness. Instead, most anthropologists who study peasant societies tend to concentrate on either superstructural or infrastructural features of a local community, or seek a more comprehensive treatment by focusing on selected features of each. Similarly, anthropological studies vary considerably in the degrees to which and the ways in which they consider how peasant communities are related to non-peasants. These differing approaches and foci can be illustrated in the work of four contrasting anthropologists.

2.1 *Redfield and the 'Folk–Urban Continuum'*

According to Robert Redfield (1941), peasant communities are a kind of folk society that exists on a 'folk–urban continuum,' which has both geographic and historic dimensions. Cities represent the modern urban end of the continuum and small, isolated non-agrarian indigenous societies are the extreme traditional folk end, with peasant communities near the traditional end. Redfield saw the history of traditional societies as shaped mainly by the spread of modern features of technology, social organization, family, kinship, values, and world view outward from cities at

the urban end of the continuum toward the folk end, in a process of modernization or development. This diffusion of the traits of modernity, especially modern values and world view, would proceed faster were it not for barriers to their acceptance in the traditional culture. This model became important in programs of applied anthropology that sought to identify and overcome cultural barriers to modernization in peasant communities, which were defined as underdeveloped, that is, as waiting to shed their traditional cultures by becoming fully incorporated into the modern national culture, economy, and political system of their nation. This approach to economic, social, and political development in peasant communities became important in American applied anthropology during the Cold War as an alternative to socialist and communist paths of development.

2.2 *Foster and the 'Image of Limited Good'*

Based on long-term intensive fieldwork on the economics, social organization, and culture of a Mexican peasant community, George Foster (1979) developed a model of world view to explain peasant economic and social behavior. This model, 'The Image of Limited Good', in the superstructure of peasant society, is based on economic realities in the infrastructure. Foster notes that land, markets, employment, and other sources of income exist in limited amounts in local peasant economies and therefore there is great competition for such scarce resources and economic opportunities. There are rarely enough of the basic material resources to satisfy everyone's needs and wants. In this situation, peasant ethics and morality are based on the idea that the best accommodation to this situation of Limited Good is for all members of the community to have an equal share of different forms of Good, e.g., food, wealth, affection. Therefore, if someone gets more than their fair share of some form of Good, then, according to the logic of Limited Good thinking, someone else must get less than their fair share. Foster says that this realistic perception of material realities shapes peasant world view in general, which is expressed as envy, fatalism, individualism, fear of witchcraft, and also principles of folk medicine, proverbs, and basic features of social organization and economic behavior, all of which are barriers to development.

2.3 *Wolf and the 'Closed Corporate Peasant Community'*

Whereas anthropologists working with a modernization perspective, such as Redfield and Foster, focused on barriers in traditional peasant society to the acceptance of modern cultural traits coming from developed urban areas, other anthropologists working with Marxist concepts examined the opposite process, that is, how unequal market relations, cheaply remunerated labor, interest payments, taxes, and tribute tend to drain economic value from peasant communities. Working with this perspective, Eric Wolf (1966) developed the structural model of the 'closed corporate peasant community' to demonstrate how, contrary to Redfield, peasants were not isolated from urban society, but instead formed much of its economic base. He explored mechanisms of value extraction and how peasants attempted to minimize it by seeking to 'close themselves off' and defend themselves from exploitive outsiders. Wolf notes that the household organization of production, and the culture of envy and suspicion that Foster and others describe, tend to promote individualism and a lack of community solidarity. These traits led Karl Marx to characterize peasants as like 'potatoes in a sack,' that is, unable to organize in their own self interests. But anthropologists like Wolf who were working with a political economy orientation have also been interested in the revolutionary potential of peasants. Indeed, in the twentieth century, some of the major armed conflicts such as the Mexican, Russian, and Chinese Revolutions, and the Vietnam War, have been characterized as peasant wars (Wolf 1969) in which peasant communities sought to gain political and economic independence from non-peasant exploiters.

2.4 *Meillassoux and 'The Articulation of Modes of Production'*

Since the 1960s anthropologists have become increasingly concerned with migration from peasant communities. Claude Meillassoux (1981), who, like Wolf, also worked in the Marxist tradition, examined how, through circular migration, rural people who were unable to make a complete living as peasants could supplement their income by migrating to seek temporary work as proletarians, thus articulating the peasant and the capitalist modes of production. This situation is seen as a benefit to the modern sector because it shifts costs for the biological and social reproduction of workers to peasant communities. Also, because such peasant workers produce part of their own food and other necessities on their own land, they can accept lower wages than fully proletarianized workers, who need more cash to buy their basic necessities. Thus, whereas the modernization approach of Redfield, Foster, and others predicted that migration would hasten the development of peasant communities, the articulation perspective revealed how such partial proletarianization in some ways perpetuates peasant ways of life in the countryside. Nagengast (1991) shows how this was true even in a socialist country, such as communist Poland, that had strong national policies for full proletarianization and

modernization of the countryside. Such conditions of partial proletarianization raise questions about the basic class nature of such peasant workers and worker peasants, and what are the most appropriate political projects to defend their class interests.

3. *Economic Development, Global Change, and The Future of Peasants*

Applied anthropology in the Redfield and Foster tradition sees peasant social organization, conservative culture, and world view as barriers to the acceptance of the social and cultural traits of modernity that are essential to economic development. Foster argues that the economic realities that peasants face dispose them to be skeptical and fatalistic about possibilities for personal and especially cooperative efforts to overcome their poverty. According to this analysis, the role of applied anthropology is to understand these social and cultural dynamics of peasant communities and demonstrate alternatives to them. In contrast, applied anthropology in the Marxist tradition, as exemplified by the work of Wolf and Meillassoux, and also by dependency theory and world system theory, pays more attention to structural conditions that keep peasants in politically and economically subordinate positions so that surplus can be extracted from them and transferred to other sectors of the national and world economy, thus maintaining peasants in conditions of *de*-development. Accordingly, applied anthropology for peasant communities in this tradition is concerned with ending such unequal exchange so that de-developed nations and their peasant communities can retain more of their wealth for their own development.

Currently, increased migration between rural communities and cities and across national borders, and the social networks and complex livelihoods that result from it, have largely obliterated the cultural, social, and economic distinctions between rural and urban areas upon which the persistence of peasants depend. Also, as the rates increase at which supposed peasants migrate in and out of a variety of economic niches—ranging from subsistence farming, to wage labor, to the informal economy, etc.—so do the volume, velocity, and diversity of commodities and information that they consume also increase. These demographic, occupational, and cultural trends towards increased mobility and differentiation thus call into question the geographic, economic, social, and cultural basis of contemporary peasant society and culture (see Kearney 1996). Indeed, peasant societies that, until recently, were the most populous type in world history seem to be rapidly disappearing.

See also: Agriculture, Economics of; Development and Urbanization; Development: Rural Development Strategies; Development: Sustainable Agriculture; Feudalism; Food in Anthropology; Food Production, Origins of; Hunting and Gathering Societies in Anthropology; Indigenous Knowledge and Technology; Industrial Society/Post-industrial Society: History of the Concept; Industrialization, Typologies and History of; Land Tenure; Peasants and Rural Societies in History (Agricultural History); Rural Geography; Rural Sociology

Bibliography

Foster G M. 1979 *Tzintzuntzan: Mexican Peasants in a Changing World*, rev. edn. Elsevier, New York

Kearney M 1996 *Reconceptualizing the Peasantry: Anthropology in Global Perspective*. Westview Press, Boulder, CO

Meillassoux C 1981 *Maidens, Meal and Money: Capitalism and the Domestic Economy*. Cambridge University Press, Cambridge, UK

Nagengast C 1991 *Reluctant Socialists, Rural Entrepreneurs: Class, Culture, and the Polish State*. Westview Press, Boulder, CO

Redfield R 1941 *The Folk Culture of Yucatan*. University of Chicago Press, Chicago

Wolf E 1966 *Peasants*. Prentice-Hall, Englewood Cliffs, NJ

Wolf E 1969 *Peasant Wars of the Twentieth Century*. Harper and Row, New York

M. Kearney

Pecuniary Issues in Medical Services: Ethical Aspects

The medical services a person receives depend not only on the patients' health status, the benefits covered by public or private health plans, and the latest developments in medical knowledge but also on the financial incentives offered to doctors that encourage them to use or withhold such services. This may compromise the physicians' loyalty to patients or the exercise of independent judgment on their patients' behalf (Rodwin 1993). Ethical problems arising from financial incentives in the health care system are as old as modern medicine. But their nature is changing along with the transformation that medical care is undergoing in the industrialized countries of the west.

1. *Patient–Physician Relationship*

At the core of a health care system there is still the patient–physician relationship. In their relationship with physicians patients are people seeking help. That results in this relationship's asymmetry, one that is

unavoidable to a certain extent. Patients can have neither the technical expertise needed to provide specific medical services nor the knowledge and skill to determine what care is needed, nor are they able to coordinate and negotiate the complexities of a highly specialized medical system to make sure that the services needed are actually provided. Patients can be very inhibited by disease, by too high or too low an age, by pain, anxieties, and worries; that means they are in special need of protection. Their vulnerability is intensified even more when they expose their bodies and reveal their most intimate histories, physical and otherwise. Even if patients participate in medical decision-making, they still depend on the physician's advice and need to trust him or her. All the concepts designed to guide medical care, whose ability to function depends on patients successfully playing the role of a 'critical customer,' ignore this asymmetry and the ill person's resulting need for protection.

In modern medicine, the physician–patient relationship is assuming ever clearer forms of a production process based on a division of labor. The patient, as a suffering subject, is at the same time a 'workpiece object' and—to different degrees—a co-producer. That means he or she is directly impacted by rationalization and bureaucratization processes in the health system, and that any changes can set the course for the patient's fears and hopes, pain and comfort, life and death.

Medical services are so complex that legal norms alone cannot protect the sick from unnecessarily suffering or even dying due to mistakes arising from the physician's lack of loyalty. Or to put it another way: a large part of successful physician–patient relationships is based not on (undoubtedly necessary but inadequate) statutory norms but on ethical ones that are internalized in the physician's professional habitus.

2. *Ethics and Institutional Structures*

Most philosophers, including bioethicists, have so far overlooked or at least underestimated the nexus between acts and the institutions increasingly in control of these acts. As a result, they overlook the fact that the 'production' of medical services is not only a work process but also a socialization or learning process, the ethical values of which are learned, reproduced and also modified and repressed. The moral-molding character of the institutions must be assessed as much stronger and durable than that of professional ethicists. This nexus can be approached with simple behavioral assumptions of the kind proposed by Freidson (1970):

> First, (...) whatever motives, values, or knowledge people have come into contact with and have 'internalized,' they do not guide the behavior of most individuals unless they are continually reinforced by their social environment.
> Second, (...) the environment can, by reinforcement, lead people to forsake one set of motives, values, or knowledge in favor of another.
> And third, given the first two, the average behavior of an aggregate of individuals can be predicted more successfully by reference to the pressures of the environment than by reference to the motives, values and knowledge they had before entering the social environment. The basis of prediction is from the requirements for social 'survival' posed by the social environment and refers to the functional adaptations of the aggregate of individuals who survive. Prediction is, of course, statistically inapplicable to any individual except as a probability value.

The more the power of institutional structures to determine behavior increases, the more significant ethical 'predecisions' embedded in the structure of medical institutions become.

3. *Financial 'Conflicts of Interest' (CoI)*

It is in this sense that the notion of '*conflicts of interest*' has come into use for discussion of the ethical aspects of pecuniary issues in medical services (Spece et al. 1996). It does not mean a subjective, 'inner' conflict that arises when interests of equal standing are weighed but 'a set of conditions in which professional judgment concerning a *primary interest* (such as a patient's welfare …) tends to be unduly influenced by a *secondary interest* (such as financial gain)' (Thompson 1993). That is the case if the loyalty to patients and/or the independence of professional judgment is compromised (Rodwin 1993).

Primary interests we can define as the ethics patients expect of loyal physicians in western industrialized countries:

(a) to be loyal to patients,

(b) to act solely in the patient's interest,

(c) to make their patients' welfare their first consideration, even when their own financial interests are opposed,

(d) to keep patient information confidential.

When the ill turn to physicians, they are confident that the same physician will act in trust as their agent and not primarily in order to enhance their income or to perform as agents for third parties. Secondary interests such as a physician's financial gain and his or her professional career are not illegitimate but should be prevented from dominating. Unlike the situation with ethical dilemmas, only one side of the conflicting interests has a claim to priority. This asymmetry between interests is a distinctive characteristic of the concept (Thompson 1993). Conflicts of interests are not an expression of misguided behavior but of *latent* misguided behavior that is built into organizations' systems of incentives and sanctions. They can influence action, but they are not acts and they do not ensure disloyalty (Rodwin 1993).

4. *Transformation of the Health Systems*

In most western industrialized countries the changing structural context is distinguished by shifts in power from the providers to the payers of medical care. This is expressed in the enforcement of new, prospective forms of financing with which expenditures can be limited. Therefore, the financial risk of treatment is increasingly shifting to the providers of medical services. That means their financial success depends on the extent to which the resources consumed by diagnosis and treatment are lower than the fixed capitation fees, budgets, DRGs (diagnosis-related groups: coding in which hospital procedures are rated in terms of cost and the intensity of services delivered, regardless of the cost to the hospital to provide that service) or flat rates. Since every service was paid for or all costs reimbursed with the traditional, retrospective methods of financing, that led for the most part to incentives to increase services. With prospective payment systems, *incentives to decrease* services now predominate, and that creates specific conflicts of interests.

The pressure of limited possibilities to expand is now intensifying competition among the providers of medical services. The main parameter in this respect is minimization of operational costs. Since they are largely determined by physicians' decisions, the ability to control and steer the physician–patient relationship becomes the pivotal point of economic success.

The problem has two elements: first, the financial incentive that rewards certain medical decisions and penalizes others and, second, the degree of inevitability that competition and/or management can compel. That gives rise to different combinations of direct influence on physician–patient relationships via 'micromanagement' and indirect influence via financial incentives. As a result, the push–pull effects of financial incentives for physicians become ever harder to avoid.

5. *Micromanaged Care—Decisions Made at a Distance*

Providers exposed to the pressures of prospective financing have developed an incalculable number of management tools to control the physician–patient relationship, e.g. the use of medical protocols to assess clinical decisions, retrospective utilization reviews, systems of prospective authorization of services, the introduction of primary care physicians as gatekeepers to control referrals to specialists and hospitals, and many other devices. The crucial factor is that important medical decisions tend to be removed from the sphere of direct contact between the patient and physician and transferred to a management.

Here we are no longer dealing only with a conflict between the precept of loyalty to the patient and the interests of the organization. Rather the fact that a management makes decisions in the absence of the sick person is already a contributory factor that changes the substance of the decisions. The moral nature of such 'decisions made at a distance' is different from decisions made face to face with the other person (Bauman 1989). If a management is to make relevant medical decisions, then questions of suffering and pain, misfortune and hope, life and death must first be turned into indicators, measurable values, and formalized norms, and thus rendered bureaucratically 'banal.' The sick individual's chances of finding empathetic loyalty are thus minimized.

6. *Monetary Incentives*

Monetary incentives are demands on the physician to individually maximize utility: 'if you choose option A instead of B in your clinical decisions, you will be rewarded/penalized with a higher/lower income.' Over and beyond the direct impact on income, the incentives also assume a normative character (in terms of both medicine and ethics): if certain decision-related options are financially rewarded, they are often considered to be medically appropriate and ethically sound.

The forms of prospective payment are multifarious. In the view of their advocates they lead:

(a) to careful handling of tight resources,

(b) to less overuse of medical services,

(c) to a better quality of services, since, after all, part of the financial morbidity risk is imposed on the physician,

(d) to structural changes and integration in a way that maintains the health of the insured population through disease prevention and health promotion,

(e) to innovations resulting in measurable improvements in patient satisfaction, and

(f) they permit this without interfering with the physicians' freedom to decide, i.e. without creating red tape, simply by changing the conditions in which decisions are made.

Critics counter that conflicts of interests resulting from incentives to decrease services (a) could lead to clinically inappropriate decisions and (b) potentially harm the ethical precept of a physician's loyalty to his or her patients, who expect, after all, an independent clinical appraisal in their interest.

(a) The physicians' clinical judgement will be biased and loyalty to the sick person may be at stake when certain options involving clinical decisions are financially rewarded or sanctioned depending on their cost. For example, in the case of capitation and fund holding, the income of the primary physician drops if he or she refers patients to a specialist, prescribes costly drugs, admits them to hospital, or calls for expensive diagnostic procedures. When individual physicians or small group practices assume the insurance risk in this way, without their number of patients being large enough to compensate for the risk,

patients in need of extensive treatment become a threat to their vital interests.

Incentives for underservice should also be avoided since they cannot be an appropriate response to the wrong turns taken by care in the western industrialized countries. For one, there may be overuse in countries like the USA, Germany, or Canada, but at the same time there is also underuse and misuse. For another, such incentives may reduce the extent of certain services, but they cannot substantially alter the ratio of necessary services to unnecessary and harmful ones (Brook and Kosecoff 1988). Since, namely, the pecuniary incentive is blind with respect to what is medically proper, beneficial services have, in principle, the same chance of being reduced or expanded as superfluous and harmful ones. Only with respect to one type of monetary incentives does there appear to be acquiescent agreement: when extra earnings are linked to clearly defined and transparent, positive goals (e.g. increasing participation in screenings, house calls, better quality indicators), there will be hardly any possibility for a conflict of interests and therefore no risk to the patient.

(b) From an ethical point of view such monetary incentives are an implicit rejection of the sick persons' expectation that the physician should be their agent. If, namely, medical decisions were only made with a view to the patient's needs and in keeping with the state of the medical art (as every patient expects and professional ethics prescribe), the type of financing would be of absolutely no importance for the physician's services-related conduct. Regardless of whether a physician receives a fixed salary or payment based on capitation, the services profile would remain the same in every case. The effectiveness of monetary incentives thus stands or falls with the physicians' readiness to violate their clinical judgment and the ethical precept of loyalty to the patient and instead pursue income-related interests (or those of their organization). If they do not do so, all attempts at monetary steering are senseless.

(c) Indirect steering of the patient–physician relationship via monetary incentives to decrease services is not an alternative to the so-called 'bureaucratic model' but tends to bring about such a model. Producers of highly complex services like physicians always have more information at their disposal than the controlling management of the insurance company or the provider organization. So they can use it to manipulate the system ('gaming the system'). Ironically, strong incentives at the level of the physician–patient relationship, with its intimidating implications, produce a spiral of attempts by the physicians to outwit them and by efforts on the part of the other side to increasingly restrict the opportunities to do so, with the result of more regulation and less freedom for doctors and patients alike.

Under the pressure of economic competition, physician–patient relationships thus potentially become part of an organizational machine. The ethical consequence thereof has been formulated by an American physician as follows, 'I imagine that every doctor in private practice can on occasion think of a patient as a dollar amount couched within a symptom, but I also imagine that doctors regularly pull back from such thinking by virtue of their education and morals. The dangers here are the institutionalization of the impulse to greed by the corporation and the individual physician's subsequent removal from personal responsibility for carrying out the actions that follow from that impulse' (Bock 1988).

The conflict between the 'two cultures' (McArthur and Moore 1997) or between the ethical codes of the profession and those of business is decided in this context in favor of the latter's dominance. With the help of a comprehensive set of tools, including financial incentives, the physician–patient relationship is squeezed into a system whose goal—profitability—is a purely quantitative one. No corporation can earn too much money, its costs can never be 'too low.' The tools of monetary incentives can signal 'more or less,' but not an objective or human criterion for 'enough'; nor can they signal the point at which the physician has to stop pursuing his or the health plan's income interests. So far it has not been possible to counter such boundless systems with any effective controls oriented to protection of the individual patient. In spite of efforts to establish monitoring systems like the HEDIS quality indicators in the US there is no practical system for externally monitoring (e.g. by public agencies) all 'the subtle ways in which providers, struggling to survive market competition, may stint on services' (Relman 1993).

7. *The Trust Problem*

Rodwin (1993) has written: 'In asking physicians to consider their own interest in deciding how to act, we alter the attitude we want physicians ideally to have. For if physicians act intuitively to promote their patients' interests, we will worry less that they will behave inappropriately. But if their motivation is primarily self-interest, we will want their behavior to be monitored more carefully' (p. 153). As already shown, financial incentives can only modify physicians' behavior in the intended sense when they act in their own income-related interest. That has to collide with the patient's expectation that the physician should place the patient's welfare over his own interests.

Not only is trust at stake but also a precondition for effective medicine. Interpersonal trust is a prerequisite, for example, for the patient's willingness: to reveal potentially stigmatizing information about health-related behavior (substance use, sexual practices, etc.), to describe personal feelings and thoughts necessary to differentiate mental from physical disorders, and to

accept treatment practices or prescribed changes in personal behavior that are difficult, painful, or risky (Mechanic and Schlesinger 1996).

It is not only the trust of the individual patient that is at risk. The trust of any and all in society to have a good chance of finding a loyal physician in case of need is a 'public good' and an element of everybody's standard of living. Public goods are never realized in ideal fashion. But there are respective critical limits. For instance, despite everyday violations we still think of the public good 'trust in traffic safety' as being assured, for otherwise we would never get into a car again. It would undoubtedly be destroyed if we had to expect, say, that 10 percent of road users would no longer stop when the light turns red. The public good known as trust would be invalidated as a result. Today competing business enterprises enforce their corporate strategies at the sick bed through the micromanagement of care, financial incentives to decrease services and sanctions, that increases the risk of 'ethical externalities' and thus of the 'social costs of business enterprises' in medicine.

8. *Remedies*

When making economic decisions, people will presumably choose the alternative that best serves their interests. But, as economist Amartya Sen points out, before they can do this, they have to make another choice. They must choose to frame their decisions as economic ones that are based on self-interest vs. moral ones that are based on concern for 'what is right.' Such decisions are what most people expect from family, friends and—specifically modified according to the ethical codes of the profession—their trustees like doctors and nurses (Schwartz 1994).

As shown, institutional developments are subjecting physicians to an economic framework of decision-making with growing effectiveness. Not only that, as socialization agencies institutions are seeing to it that the professional code of ethics is already being usurped by superimposed economic considerations. That is why, at the turn of the twenty-first century, people are starting to realize that the patient's welfare must no longer depend on whether a physician is able to resist financial incentives and to ignore or submerge his or her economic interest in order to act in the patients' interest. Society has to find an answer to the principle ethical question of whether the health and life of a patient seeking help is to be decided primarily within an economic framework, and the answer does not have to be expressed solely in the individual physician's conscience but also in the structure of the institutions. There are no patent solutions here. While in some west European countries there is hardly any consciousness of the problem posed by conflicts of interests yet, a public debate about the steps to be taken toward a solution has begun in the United States (cf. biography). Consideration is being given to a broad range of possibilities. One group of proposals has in common that they would like to cap expenditures at the middle or overall social level by way of budgets, while individual physicians are to be paid a fixed salary (possibly supplemented by financial incentives tied to positive targets like outcomes and patient satisfaction). Overuse, underuse, and misuse of medical services should be prevented by supportive management as opposed to controlling and sanctioning management. Others would like to defuse the prospective payment systems, e.g. financial risk or gain should be limited, financial incentives should not directly reward decreased use of services unless there is specific evidence thereof and monitoring for underuse. Other types of safeguards proposed to protect patients are the disclosure of financial incentives, public review boards, creating independent organizations to certify quality measures, appeals boards to review cases of denied care for reasons of costs, the expansion of individual and collective patients' rights, and the strengthening of consumer organizations.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Discrimination; Discrimination, Economics of; Ethical Dilemmas: Research and Treatment Priorities; Health Care Delivery Services; Health Economics; Medical Experiments: Ethical Aspects; Medical Profession, The; Medical Sociology; Minority Access to Health Services: United States

Bibliography

Bauman Z 1989 *Modernity and the Holocaust*. Basil Blackwell, Oxford, UK

Bock R S 1988 The pressure to keep prices high at a walk-in clinic. *New England Journal of Medicine* **319**(12): 785–7

Brook R H, Kosecoff J B 1988 Competition and quality. *Health Affairs* **7**(3): 150–7

Daniels N, Light D W, Caplan R L 1996 *Benchmarks of Fairness for Health Care Reform*. Oxford University Press, New York

Emanuel E J, Goldman L 1998 Protecting patient welfare in managed care: six safeguards. *Journal of Health Politics, Policy, and Law* **23**(4): 635–59

Freidson E 1970 *Professional Dominance*, 1st edn. Aldine, New York

McArthur J H, Moore F D 1997 The two cultures and the health care revolution: Commerce and professionalism in medical care. *Journal of the American Medical Association* **277**(12): 985–9

Mechanic D, Schlesinger M 1996 The impact of managed care on patients' trust in medical care and their physicians. *Journal of the American Medical Association* **275**: 1693–7

Relman A S 1993 Controlling costs by 'managed competition'—would it work? *New England Journal of Medicine* **328**(2): 133–5

Rice T 1998 *The Economics of Health Reconsidered*. Health Administration Press, Chicago

Rodwin M A 1993 *Medicine, Money, and Morals: Physicians' Conflicts of Interest*. Oxford University Press, New York & Oxford

Salmon J W (ed.) 1994 *The Corporate Transformation of Health Care: Perspectives and Implications*. Baywood Publishing Comp., Amityville, NY
Schwartz B 1994 *The Costs of Living: How Market Freedom Erodes the Best Things of Life*. W.W. Norton, New York/London
Spece R G, Shimm D S, Buchanan A E (eds.) 1996 *Conflicts of Interest in Clinical Practice and Research*. Oxford University Press, New York & Oxford
Thompson D F 1993 Understanding financial conflicts of interest. *New England Journal of Medicine* **329**(8): 573–6

H. Kühn

Pedagogical Reform Movement, History of

1. *History of the Concept*

Theories on educational reform have been current since Greco–Roman times, in all three directions of reforms related to school, life, and society. Reform movements in the Middle Ages initially occurred on the basis of religious motives, above all where deviations from church doctrines were associated with an alternative practice. Heresy always suggested new education, because the divergences from the main doctrine picked out alternative forms of salvation as a central theme, to which the life of an own group had to be adapted. Education was not centered on children and youngsters, but it focused on religious and social renewal. It was envisaged that the 'new' education would bring about 'new people,' a central theme which was developed in Baroque didactics into a program of holistic renewal of mankind. Proofs of this theme can be found in many contexts right up to the pedagogical reform of the twentieth century, in this respect ever more closely related to the mythology and cult of 'the child' (Boas 1990). The original sacred theme was continuously secularized, without weakening the appertaining religious power. The division of education and religion had shifted Messianic expectations onto education which, since the seventeenth century, was supposed to ensure perfection in this world.

'Reform,' therefore, was never just pragmatic; it was, at the same time, a utopia that did not lose track of the salvation concept and could also understand education as a power and cause of the radical renewal. The success of this was considered to be all the better the more education could appear to be methodically mastered—a concept that was promoted by the 'inwardness' of belief within the European Reformation. The 'inner belief' should be granted by new education, and this required means of influence. The Baroque programs convinced by way of a mixture of 'wholeness' and 'method.' Similarly, sensualistic psychology in the eighteenth century was understood as a strengthening of the effectiveness of education, and pedagogical doctrines were developed with the promise of methodical innovation. At the same time a literary image of the child was created between Rousseau and the Romantic era which particularly excluded technical expectations. The reform movements of the nineteenth and twentieth centuries developed in this field of tension of 'natural education' and 'method.'

2. *Reform Movements in the Nineteenth Century up to World War I*

'Education nouvelle,' 'progressive education,' '*Reformpädagogik*' or '*nuova educazione*' are subsequent historiographic descriptions of widely divergent reform groups and reform concepts which, from the final third of the nineteenth century, opposed the foreseeable damage and weaknesses of the developing educational system. The 'grammar of schooling' (Tyack and Cuban 1995), i.e., the institutional form of the modern school, developed during the nineteenth century as a single comparable system in the individual national states. Compulsory schooling was created everywhere, teachers were professionally trained for undertaking tasks of the state, and schools developed specific institutional procedures and forms such as the grading of pupils in age groups, the dominance of subjects, largely frontal forms of lessons, the assessment of pupils' performances using grades or marks, and so on.

The weaknesses of this system were the initial theme of the reform movements, which already had international links by 1914, but in a practical sense were largely geared towards their respective national systems. Influential authors of the nineteenth century were guided by parameters laid down by Jean-Jacques Rousseau or J. H. Pestalozzi. Rousseau was considered a pioneer of natural education, while Pestalozzi was an authority on education according to moral ideals; both concepts were employed more and more against the institutional development of the school. Friedrich Fröbel's Kindergarten was of particular importance to the development of the reform movements. Central motives of the reform movements at the end of the nineteenth century, such as the focus on children, the playful manner of learning, or active methods of teaching, were disseminated in the kindergarten media, even if the practice did not always comply with the theory.

Compared with the state school development, these attempts were marginal during the nineteenth century, albeit much discussed and widely published. Significant school-reform movements were only established at the end of the nineteenth century, in a manner which covered three issues. In England and continental

Europe private alternative schools in the area of higher education ('*Landerziehungsheime*') were established which replaced or supplemented the former private school institutions. In the area of elementary school education, reform initiatives were established, above all, where social democracy gained an influence. And third, life reform movements were linked in a very different manner to the school initiatives. One of them was the Theosophical Society, which can be seen as the most influential international individual grouping of educational reform at the end of nineteenth century. In Germany it resulted in the establishment of Rudolf Steiner's Anthroposophical movement (founded in 1912). The establishment of the 'Waldorf School' (1919) in Stuttgart was the start of the largest private school reform movement to date. The new '*écoles nouvelles*,' which were established in England, France, Switzerland, Germany, and various other nations, represented the successful utilization of a pedagogical niche. In 1910 some 100 schools were registered; these operated on the basis of full-day boarding schools as 'educational communities' with a particular curriculum. These schools created a demand because they were able to respond to the dissatisfaction, above all of middle-class parents, with the rigid forms and outdated contents of the grammar schools. Furthermore, country boarding schools often provided the only opportunity to actually achieve university entrance qualifications.

School establishments on the basis of political and social change have been undertaken time and time again since the Owenite Movements in England and the USA. In 1825 the first socialistic school community was established in New Harmony, Indiana. However, most of these did not enjoy lasting success. Only a few alternatives outside the nineteenth-century association of home, school, and church were successful. The pedagogical reform movements were not only grounded in spiritual and/or political world views. Theories of child psychology were to permanently strengthen the intuitions of 'new education': the physiological evidence of the independent development of the senses and feelings, the description of the development of the child according to natural 'phases' or 'stages,' and the recording of the milieu and behavior of children in research facilities.

Together with the exemplary models of the school reform, this knowledge was authoritative for the development of teacher training. The focus of attention of the reform movement in the USA was Columbia University's Teachers College, established in 1897. Its Kindergarten and Elementary School became the experimental focus of child-centered education in the 1920s (Rugg and Shumaker 1928, Ravitch 1986). However, the Teachers College was, above all, a research center. Thus, it was not only activism, but psychological and educational research that was responsible for the success of 'new education.' In 1912 the Institut Jean-Jacques Rousseau was established in Geneva. The year 1915 saw the establishment in Berlin of the Zentralinstitut für Erziehung und Unterricht, while in 1920 Ovide Decroly established his institute of psychological research, which had previously been private, at the University of Brussels. In 1922 the London Institute of Education was set up on the lines of pedagogical reform.

These developments were favored by the cultural and social *cesura* of World War I. After 1918, political educational reforms were called for in all continental European societies, which were aimed at reacting to the devastating effects of the war. In England a radical education sector came into being; it was, however, intended to, or was forced to, operate outside the system of public education. In the USA, April 1919 saw the establishment of the Progressive Education Association (PEA), which was aimed at changing nothing less than the entire American school system (Cremin 1964, p. 241). Its public influence was considerable, also because after 1930 the academic world was opened up for 'new education' mostly via Teachers College (Cremin 1964, p. 250). The French *éducation nouvelle* gained considerable influence at the same time in the entire francophone educational world, to the extent that at the end of the 1930s the term '*expansion mondiale*' was being applied to the new education.

3. *Pedagogical Reform Between the Wars*

Three developments are of considerable importance to the period between World War I and World War II: the establishment of canonical convictions of what 'new education' is, including the appertaining major figures of eminent reformers, the development of international structures, and the very divergent politicization of the 'new education.' The core of the doctrines was child-centeredness. Education should follow the spiritual, physical, and mental growth of the child. The educational institutions must adapt their actions in line with the natural development. The mode of education dispenses, as far as is possible, with external determination, and thus dispenses with authority.

The reception of Freud's psychoanalysis and the beginnings of child therapy were of great public importance, alongside the liberal concepts of sexual education, which were discussed and practiced in all European major cities up to 1933. It was envisaged that there should be great liberty in early childhood sexuality, and considerable credit was given to Freud for this discovery. The pedagogical reception of Freud, however, is contradictory and selective—not only because of Freud's late theory of drive and aggression but also because its related cultural criticism does not fit in with the optimistic picture of a child developing in accordance with its good nature. Most reformers subscribed to the picture of the good and independent

child that cannot become neurotic or destructive unless pedagogical authority causes it to be so. This picture was simultaneously that of a gender-neutral, culturally independent and socially free child which, precisely as Rousseau had stipulated, exists for itself and needs only to be protected from damage.

The Montessori pedagogy, which became influential in England, the Netherlands, Germany and Italy, was probably the largest international education movement of the 1920s and 1930s. It saw the establishment, in each case, of independent, usually female sponsor groups, with their own communication organs and an international network. The case in Italy is particularly interesting from a political point of view, because Montessori here was not considered the actual pioneer but the theorist of the '*nouva educazione*.' Montessori cooperated with the Fascist regime between 1923 and 1993 to promote her method.'New' education's focus on the child was the subject of a dispute between Dewey and the Progressive Education Association; they were, however, never able to dissolve the dualisms 'child' and 'curriculum' or 'child' and 'society.'

The international meetings of the New Education Fellowship, which began in 1921 in Calais (topic of conference: 'The Creative Self-expression of the Child') and which ended in 1941 in Ann Arbor (topic of conference: 'Education in a World of Nations') demonstrate the scope of tension between pedagogical reformers' concentration on the child, the creative powers within the child, and the drama of world politics. It is not by chance, therefore, that the pedagogical reform appeared to be torn between the fronts of the Cold War. The Progressive Education Association was dissolved 1955 without much ado under the impression of massive criticism of the liberal and child-centered education which was held responsible for the backwardness of the American education system.

4. *Radical Years*

In 1960 Paul Goodman recollected, with reference to Bertrand Russell and Alexander Neill, the courageous 1920s and the 'revolutionary' attempts of the progressive education movement (Goodman 1960, Sect. IV/6). These attempts had, in fact, failed (Goodman 1960, Sect. IV/7), but they had not been in vain, because the education of the 'beat generation' had to link up with the movement. In the same year, 1960, the New York publisher Harold Hart put out a compilation of texts by Alexander Neill from the 1920s and 1930s under the title *Summerhill: A Radical Approach to Child Rearing* which, by 1970, had reached a circulation of more than two million copies in the USA alone.

Neill, Holt, and Goodman were initially underground authors, at least in America. The international student movement of the 1960s ensured that the concepts of the 1920s could be once again understood as for avant-garde society's innovations. John Holt's *How Children Learn* (1970) was a strong confirmation of the child-centered pedagogy without these themes being particularly historicized. They were rewritten and adopted for a new generation. Goodman's 'beat generation,' which is actually a negative description that fits with the deficit formula of educational reform, could be related to an experience of emancipation which applied, in particular, to the educational institutions. The cultural revolution was directed against the pedagogical authorities, while projects focussing on self-experience were promoted. Again, formal school education came under attack; the 'new education' (Postman and Weingartner 1969, chapter 8) was aimed at binding school knowledge to subjective meaning (Postman and Weingartner 1969, chapters 5 and 6) without further following a state curriculum. Compulsory education was called into question as 'compulsory mis-education' (Goodman 1964).

The radical approach of the educational reformers in the 1960s was not new. What is new in the USA is the link between progressive education and society reform movements which extend beyond the New Deal positions. This includes the civil rights movement, the women's movement and minorities, as well as third world movements. Paulo Freire's Pedagogy of the Oppressed (1970) is perceived by the leftwing American pedagogy as a change in the traditional European-centered reform discourse. The same applies to Ivan Illich's radical criticism of schooling, which goes back to ideas of the Centro Intercultural de Documentacion in Cuernavaca, Mexico, established in 1964. Subsequent to this criticism, the matter in question was no longer school reform, or applying the avant-garde position in the life reforms movements or cultural criticism; the focus now was on education and training.

The 'new education' was understood on one hand as a political strategy of liberation in view of third world problems. On the other hand it was seen as a withdrawal of all demands which were not compatible with the ideal of human liberty. Therefore, it was not by chance that John Holt's *Escape from Childhood* (1975) brought the radical years to an end. Even free and alternative schools such as Summerhill, Holt said, can fail, and, in fact, in greater numbers the more they have tied the happiness of children to their masts (Holt 1975, chapter 14). The problem can only be solved if the pedagogic differentiation between 'adults' and 'children' disappears, and children are granted the same rights as adults (Holt 1975, chapters 15 and 16).

In Europe and England there were comparable radicalizations which allowed for the fundamental negation of education from the linking of education with emancipation in little more than 10 years. This 'anti-pedagogy' is often justified with theories of depth psychology which also demonstrate how strong the effect of educational reform was in the 1920s; only

what was previously understood as a radical start of the 'new education' is now seen as a negative development. The brief recurrent link between movements of societal reform and 'new education' could not be renewed, among other reasons because it became evident that the 'new education' had been merely a historical project that did not allow for general applications such as change of society.

5. *Conclusion*

The real history of education of the twentieth century has confirmed rather than changed the national systems in place, irrespective of the political catastrophes and the affiliated radical breaks in culture and society. Evidently schools are robust systems which renew themselves from within and in this respect adapt those aspects of progressive education that are acceptable. These innovations never put the entire system at risk. During the 1990s this structural conservatism produced entirely different reform movements which linked up neoliberal economic theories with the old positions of freedom of learning or self-determination. Looked at from the point of view of privatization, efficiency, and commercialization (Chubb and Moe 1990), the criticism of school bureaucracy of the nineteenth century has reappeared, along the renewal of the theory of 'death of the school,' through the revolution of the interactive learning media (Perelman 1993). The critique of schooling substantiates the home schooling movement, accompanied with an internet service for freedom from schooling.

The new movements founded in the 1990s are not turning their backs on reform pedagogical motives. Criticism is now leveled at ineffective and inactive institutions whose organizational form and sponsorship should be radically changed, but there is little consensus concerning the direction in which this change should proceed. The motives are based on political change as much as on maintaining family values, and they have liberal as well as community justifications on which agreement is ultimately unlikely to be reached. As ever, the discourse on how to reform education is ongoing.

See also: Education and Gender: Historical Perspectives; Educational Philosophy: Historical Perspectives; Educational Research and School Reform; Pestalozzi, Johann Heinrich (1746–1827); School (Alternative Models): Ideas and Institutions.

Bibliography

Boas G 1966/1990 *The Cult of Childhood*. Spring Publications, Dallas, TX

Chubb J E, Moe T M 1990 *Politics, Markets, and America's Schools*. The Brookings Institution, Washington, DC

Cremin L A 1961/1964 *The Transformation of the School. Progressivism in American Education, 1876–1957*. Vintage Books, New York

Freire P 1970 *Pedagogy of the Oppressed*. Seabury Press, New York

Goodman P 1960 *Growing Up Absurd. Problems of Youth in the Organized Society*. Random House, New York

Goodman P 1964 *Compulsory Mis-education*. Random House, New York

Holt J 1964 *How Children Fail*. Pitman Publishing, London

Holt J 1970 *How Children Learn*. Dell Publishing, New York

Holt J 1975 *Escape from Childhood*. Dell Publishing, New York

Perelman L J 1993 *School's Out*. Hyperlearning, the new technology, and the end of education. Avon Books, New York

Neill A S 1960 *Summerhill: A Radical Approach to Child Rearing*. Hark, New York

Postman N, Weingartner C 1969 *Teaching as a Subversive Activity*. Delacorte Press, New York

Ravitch D 1974/1986 *The Great School Wars, New York City, 1805–1973. A History of the Public Schools as Battlefield of Social Change*. Basic Books, New York

Rugg H, Shumaker A 1928 *The Child-centered School. An Appraisal of the New Education*. World Book Company, Yonkers-on-Hudson/Chicago

Tyack D, Cuban L 1995 *Tinkering Toward Utopia. A Century of Public School Reform*. Harvard University Press, Cambridge, MA/London

J. Oelkers

Peer Review and Quality Control

Peer review is a mechanism for quality control in science, including the assessment of proposed projects and also of completed work. Its roots go back to the emerging science of seventeenth century, when novelty in natural knowledge became distinguished from technical invention and the ideals of reproducibility and publicity became established. Peer review was acknowledged by the mid-twentieth century as the unquestioned norm in 'academic science.' By then it was assumed that quality was unproblematic, owing to some special ethical disposition of scientists and their methods. With subsequent transformations in the social practice and context of science, which now extends into technology and policy, corresponding changes in quality assurance are needed. The crucial requirement is for a more encompassing community of those evaluating the products and processes of science (see, e.g., Stampa 1997).

In most other spheres of institutional activity, the formalization of quality assurance has become the norm, as for example through the wide-ranging standard-setting procedures of the International Standards Organization (ISO). In academic science, however, along with cultural pursuits like the arts, the methods are still largely informal. Science has been almost unique in having self-assessment performed by practitioners rather than by external 'critics.' To what

extent and in what ways this must change to keep pace with science's expanding role in public life has become an urgent question in the governance of science.

The assurance of quality is not a straightforward task. This has been known since the time of the Romans, as indicated by the Latin motto *quis custodiet ipsos custodes?* (Who will guard the guardians themselves?). This motto implies an indefinite iteration. It is a reminder that, however, routine may be the tasks of quality control, full quality assurance demands yet higher levels of supervision at which informality and explicit value judgments are necessary.

As long as science remained mainly academic, problems of quality were assumed to be resolved by the very nature of the scientific endeavor. The informal systems of checking by peers seemed a rational response to the problem, rather than a culturally contingent mechanism characteristic of a particular epoch. Scientific facts were believed to be discovered by an infallible method, and scientists themselves were viewed as being endowed with certain superior moral qualities that protected them and their work from ordinary human failure or error. This self-correcting property of science could be explained in sociological terms, as in the 'four norms' of scientific practice expounded by Robert K. Merton in 1942 (Merton 1973), or philosophically, as in the committed attempts at self-refutation supposed by Karl Popper to be normal scientific practice (Popper 1959).

With the onset of the industrialization of science after World War II, the self-conscious study of science as a social activity, including the methods of quality assurance, became inevitable. Growth in size, capital investment, scale, and social differentiation within science created divisions between managers and research workers, as well as between researchers and teachers in universities. A *Gemeinschaft* (community) of scholars could no longer realistically be assumed. The earliest disciplined analyses of the quality of production in science were quantitative. Derek J. de Solla Price, who devised some measures of quality and provided analyses of its distribution, did the pioneering work. He noticed that at the leading British science reference library only a minority of journals was ever requested. The contents of the others could be inferred to have no interest, and hence to be of very low scientific quality (Price 1963). This phenomenon is a reminder that 'quality' is a relational attribute. 'Fitness for purpose' depends on whose purposes are dominant; not always perhaps those of a community devoted to the advancement of learning, but possibly only of those scientists working under constraints of 'publish or perish.'

Price's studies were continued in two directions. At the Institute for Scientific Information, Eugene Garfield produced more searching and sophisticated measures of quality, using citations rather than mere number of publications. Such attempts at quantification were bound to become controversial (Brooks 1982, Garfield 1970, 1972). It was impossible to avoid bias in the selection of the relatively small set of journals used for citations; those in conventional mainstream English-language research science were inevitably privileged at the expense of all others. Further, when quantitative measures of citations came to be used as indicators of academic merit, manipulative practices, including reciprocal citations, inevitably developed. The deep problems of a quantitative proxy for quality suddenly became acute.

In a more reflective vein, Jerome R. Ravetz applied the *quis custodiet* principle to analyze the vulnerability of the quality assurance system in science. He observed that the processes of peer review are conducted largely informally and (unlike research) are not themselves normally submitted to open scrutiny and review. They require a different sort of competence, which is not part of the formal training of scientists; and they are also more open to a variety of abuses, ranging from bias to plagiarism. One can understand the phenomena of low quality, both in scientific research and in technological development in these terms. Thus, while denying that the practice of science automatically produces a higher morality, Ravetz agrees that moral standards are necessary for the successful practice of science. On this basis he stresses the importance of morale and morality (and ultimately idealism and leadership) in science (Ravetz 1996).

This analysis provides a background for the increasing interest in 'trust' as an essential element of practice in science, in society, and in their interactions. The broader society has provided resources to the esoteric activities of science because it trusts the scientific community to make good use of them. There has always been an undercurrent of distrust, based on evidence either of pointless research or of malign applications. Once science became deeply involved in technology and related policy problems that crucially affect public health and welfare, the traditional relations of trust could no longer be assumed. It appeared to be necessary for the principles and practices of accountability to be extended from the institutions of political governance (as, e.g., representative democracy) to those institutions, which govern science and its applications.

Quality control in research science has become more difficult as the relatively inflexible technical requirements of the traditional printing process have been relaxed. There is no longer a well-defined 'gateway' to publication through the institutions that control reproduction of, and hence access to documents. First through inexpensive photocopying and now through the Internet, it has become easy for anyone to distribute scientific wares to an unrestricted audience. In addition, the presence of the global media tends to bypass the traditional processes of evaluation, which were conducted personally among colleagues. Isolated scientific results can become media events (Close 1991). All those with an interest in the report, such as

consumers, politicians, regulators, and the stock markets, become potential stakeholders in the evaluation of the result. Thus, science arguably becomes accountable to a drastically extended peer community in the quality-assurance process. The criteria of quality applied by these heterogeneous actors need not be identical to those of 'public knowledge' generated within tightly knit scientific networks.

These developments may be judged in different ways. While they may seriously disrupt the procedures of quality assurance in normal science, they can also bring needed public scrutiny to bear on controversies and scandals. The demystification of scientific practice both enables such events to become news, and is fostered by their being exposed. Top scientists become like celebrities—needing the media for advertising themselves yet simultaneously hating it for its unwanted intrusions. The 'Baltimore affair,' centering on the US Nobel laureate David Baltimore's laboratory at MIT, is perhaps the most notorious case in which a dispute about scientific misconduct was blown up into a lengthy, visible, political saga that damaged all the individuals and institutions involved (Kevles 1998). The episode was symptomatic of an increasingly recognized problem of 'deviance' in science, which carries the unspoken danger that, without timely correctives, misconduct might become the norm.

All these developments affect the maintenance of trust, which is necessary for ordinary scientific practice and even more for quality assurance. As in other professional domains, the normal tendency in science has been for misconduct to be covered up by the responsible institution (not necessarily by the community of scientists). In such situations, ultimate exposure does even more damage and further erodes the basis for mutual trust. Attempts to circumvent the need for trust by increasing bureaucratic surveillance are likely to be counterproductive in their own way, by erecting impediments to free inquiry and communication among colleagues.

The relations between social science and natural science have also been transformed during the last decades, with implications for quality control. Starting with the acceptance of natural science as the ideal of knowledge, essentially independent of social forces, there has been a gradual but accelerating shift toward recognizing all sciences as incorporating social constraints and biases. An early critical interaction was in connection with the astronomical community's management of the eccentric Velikovsky (de Grazia 1966). Later, the social science community embraced Thomas Kuhn's disenchanted picture of 'normal' science (Kuhn 1970). Finally, post-Feyerabend studies of science re-examined the whole institution of scientific inquiry without presupposing any privileged status in relation to either virtue or natural knowledge (Bloor 1991, Bloor et al. 1995, Collins and Pinch 1993, Fuller 1993).

When natural scientists, led by physicists, eventually confronted the emerging socialized picture of their discipline, the reaction was so strident that 'science wars' became an appropriate label (Gross et al. 1997, Nelkin 1996, Ross 1996). Sociologists of science and post modernists were indiscriminately blamed for all the ills of science, including decline of public trust, budget cuts, resurgent Creationism, and even poor teaching of science. A physicist whose hoax article (Sokal 1996) was accepted by a leading cultural studies journal, *Social Text*, helped to crystallize the attack (Bricmont and Sokal 1998). The implication was that the critics of science had no real quality control of their productions. The science warriors' assumption was that within real science, such problems are prevented from occurring because of the verifiable empirical content of scientific research. In the ensuing debate, there was little mention of the ease of publication of erroneous or vacuous research in the standard scientific literature. Historical episodes, like Millikan's manipulation of his oil-drop results in the course of a controversy on the charge of the electron, were discounted as mere embarrassments (Segerstale 1995).

It has been presupposed thus far that 'science' refers primarily to traditional basic research. But among contemporary forms of scientific practice, curiosity-driven research with no regard for applications has been increasingly marginalized. A diversification has occurred, so that quality assurance must also be considered in such areas as mission-oriented and issue-driven research, forensic science (Foster and Huber 1997, Jasanoff 1995), and the provision of scientific advice for policy (Jasanoff 1990, Salter 1988). In addition, the products themselves and the media through which they are diffused increasingly are diversified. For example, patents are now a common outcome of a research process, and this form of intellectual property is radically different from traditional published papers (Myers 1995). Also, results are reported in unpublished consultancy advice and unpublished 'gray literature' or kept confidential within institutions or even totally sealed under 'lawyer-client confidentiality' and legal settlement agreements. With traditional peer review as the norm, the challenges of quality assurance for these new products and processes are nearly unrecognizable. A genre of critical literature has developed, with some authors directing anger at the new contexts of scientific production (Huber 1991), and others more clearly appreciating the problems they present (Crossen 1994, Jasanoff 1990, 1995).

A parallel diversification has occurred in the types of knowledge production that are accepted as legitimate. The democratization of knowledge now extends beyond the juries who assess the quality of technical evidence in courts (Jasanoff 1998) to include those who master previously esoteric aspects of their predicament (e.g., illness, contamination, pollution, oppression, discrimination, exploitation) through special-interest groups or the Internet. In addition,

claims of specialized or local knowledge are present in even more diverse contexts, as among indigenous peoples, and in systems of complementary or 'traditional' medicine. These claims are commanding increasing commercial and political support among various publics, as well as gaining explicit recognition in numerous international treaty regimes. As a result, a new philosophy of knowledge appears to be emerging, based on a new disciplined awareness of complexity, in which a plurality of legitimate perspectives is taken for granted (Funtowicz and Ravetz 1991). Modern science, with its characteristic methodology and social location, is part of this enriched whole, but not coextensive with it. The criteria and tasks of quality assurance must explicitly involve additional values and interests, incorporating even the ontological commitments of groups other than scientists. This new configuration has been termed postnormal science.

Quality assurance can, thus, be seen as a core commitment of postnormal science, replacing 'truth' as science's ultimate regulative principle (Funtowicz and Ravetz 1992). Defined in terms of uncertainties and decision-stakes, quality assurance encompasses 'public interest,' 'citizen,' and 'vernacular' sciences. In a period of domination by globalized corporate science (Gibbons et al. 1994), this effort to make scientists accountable to interested groups presents a coherent conceptual alternative for the survival of the 'public knowledge' tradition of science. Collegial peer review is thereby transformed into review by an 'extended peer community.' This new form of quality assurance will be given its formal structure and routines by those heterogeneous actors who put it into practice.

See also: Academy and Society in the United States: Cultural Concerns; Constructivism in Cognitive Psychology; History of Science; Research Conduct: Ethical Codes; Research Ethics: Research; Research Publication: Ethical Aspects; Scientific Academies, History of; Social Constructivism; Truth and Credibility: Science and the Social Study of Science; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Bloor D 1991 *Knowledge and Social Imagery*. University of Chicago Press, Chicago

Bloor D, Edge D, Henry J 1995 *Scientific Knowledge*. Athlone, Chicago

Bricmont J, Sokal A D 1998 *Fashionable Nonsense: Post-modern Intellectuals' Abuse of Science*. Picador, New York

Brooks H 1982 Science indicators and science priorities. In: La Follette M C (ed.) *Quality in Science*. MIT Press, Cambridge, MA, pp. 1–32

Close F H 1991 *Too Hot to Handle: The Race for Cold Fusion*. Princeton University Press, Princeton, NJ

Collins H, Pinch T 1993 *The Golem: What Everyone Should Know about Science*. Cambridge University Press, Cambridge, UK

Crossen C 1994 *Tainted Truth: The Manipulation of Fact in America*. Simon & Schuster, New York

Foster K R, Huber P W 1997 *Judging Science—Scientific Knowledge and the Federal Courts*. MIT Press, Cambridge, MA

Fuller S 1993 *Philosophy, Rhetoric and the End of Knowledge: The Coming of Science and Technology Studies*. University of Wisconsin Press, Madison, WI

Funtowicz S O, Ravetz J R 1991 A new scientific methodology for global environmental issues. In: Costanza R (ed.) *Ecological Economics*. Columbia University Press, New York, pp. 137–52

Funtowicz S O, Ravetz J R 1992 Three types of risk assessment and the emergence of post-normal science. In: Krimsky S, Golding (eds.) *Social Theories of Risk*. Greenwood Press, Westport, CT, pp. 251–73

Garfield E 1970 Citation indexing for studying science. *Nature* **227**: 669–71

Garfield E 1972 Citation analysis as a tool in journal evaluation. *Science* **178**: 471–9

Gibbons M C, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M 1994 *The New Production of Knowledge*. Sage, Beverley Hills, CA

de Grazia A (ed.) 1966 *The Velikovsky Affair: The Warfare of Science and Scientism*. University Books, New York

Gross P R, Levitt N, Lewis M W (eds.) 1997 *The Flight from Science and Reason*. Johns Hopkins University Press, Baltimore, MD

Huber P W 1991 *Galileo's Revenge: Junk Science in the Courtroom*. Basic Books, New York

Jasanoff S 1990 *The Fifth Branch: Science Advisors as Policy as Policymakers*. Harvard University Press, Cambridge, MA

Jasanoff S 1995 *Science at the Bar: Law, Science and Technology in America*. Harvard University Press, Cambridge, MA

Jasanoff S 1998 The eye of everyman: Witnessing DNA in the Simpson trial. *Social Studies of Science* **28**(5–6): 713–40

Kevles D J 1998 *The Baltimore Case: A Trial of Politics, Science and Character*. Norton, New York

Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

Merton R K 1973 *The Normative Structure of Science*. University of Chicago Press, Chicago

Myers G 1995 From discovery to invention: the writing and rewriting of two patents. *Social Studies of Science* **25**(1): 57–105

Nelkin D 1996 What are the science wars really about. *The Chronicle of Higher Education July* **26**: A52

Popper K 1959 *The Logic of Scientific Discovery*. Basic Books, New York

Price de Solla D J 1963 *Little Science, Big Science*. Cambridge University Press, Cambridge, UK

Ravetz J R 1996 (1971) *Scientific Knowledge and its Social Problems*. Transaction Publishers, New Brunswick, NJ

Ross A (ed.) 1996 *Science Wars*. Duke University Press, Durham, NC

Salter L 1988 *Mandated Science*. Kluwer, Dordrecht, The Netherlands

Segerstale U 1995 Good to the last drop? Millikan stories as 'canned' pedagogy. *Science and Engineering Ethics* **1**: 197–214

Sokal A D 1996 Transgressing the boundaries. *Social Text* **14**: 217–52

Stampa A E III 1997 Advances in peer review research. *Science and Engineering Ethics* **3**(1): 1–104

S. Funtowicz

People in Organizations

Organizational researchers study how individuals and organizations interact with their environment to accomplish both individual and collective goals. People are able to accomplish goals in organizations that they would be unable to accomplish alone, and they spend large portions of their lives interacting in organizational contexts. The complexities that emerge from bringing people together in organizational settings, therefore, provide a fascinating lens into the human experience. Organizational research brings together psychology, sociology, economics, and anthropology to examine both individual-level processes at work, as well as the organizational-level processes that dictate how firms relate to each other and their environment.

Reflecting this focus on both individuals and organizations, the field is split between 'macro' organizational behavior, drawing upon organizational sociology, and 'micro' organizational behavior, drawing upon social, personality, and industrial-organizational psychology. Macro-organizational behavior is concerned with the various strategies organizations use to adjust to their environment, such as the development of networks between firms. In contrast, micro-organizational behavior is concerned with the way personal characteristics combine with contextual features such as task, group, and organizational characteristics to predict work outcomes. Though the strength and uniqueness of organizational behavior rests on the equal emphasis of both individual and organizational processes and the utilization of a variety of research methods that cross disparate levels of analysis, our focus in this article is on micro-theory and research. Excellent reviews of macro-organizational theory can be found elsewhere in this collection (see *Organization: Overview*; *Administration in Organizations*; *Organizational Control*; *Organizational Culture*; *Organizations: Authority and Power*) and in Carroll and Hannan (2000) and Scott (1998).

1. *Micro-theory: Person–Situation Interactions*

Micro-organizational behavior examines both personal and situational characteristics and, as in the field of psychology, researchers debate the relative utility of each in explaining behavior. Some have emphasized the stability of attitudes and behaviors over time. For example, a person's satisfaction with his or her job remains relatively stable over years and even decades (Staw and Ross 1985). From this perspective, individual characteristics are the best predictors of behavior since they derive from personal dispositions that remain stable over time and across situations. Others have criticized this view and posited that organizations should be conceptualized as 'strong' situations that are powerful enough to shape individual behavior (e.g., Davis-Blake and Pfeffer 1989). In strong situations individual differences are unlikely to be expressed. Instead, people learn appropriate attitudes and behaviors from their co-workers, established norms, and organizational practices and procedures; these social influence processes are presumed to predict individual behavior better than are personal characteristics.

Researchers have typically considered personal and situational factors in isolation from one another, but a complete understanding of organizational behavior requires their simultaneous consideration. An interactional approach is more complex than a mere additive melding of personal and situational characteristics because it attempts to represent both personal and situational factors and their reciprocal influence.

Interactions between personal and situational characteristics may take at least four forms. First, as specified above, some situations are stronger than others, leading to different levels of behavioral uniformity. Second, work situations do not affect everyone in the same way; some people's behavior is more consistent across varying situations. Third, certain people, such as those exhibiting 'charismatic' leadership, can influence situations more than others. Finally, people do not select into situations randomly, but rather, into situations in which they think their attitudes and behaviors will be appreciated. Developing a complete theory of behavior in organizations, then, requires moving from considering personal and situational factors in isolation to considering the complexity and diversity of possible person–situation interactions and their effects on work outcomes. We illustrate this by identifying the types of person–situation interactions that are relevant to a set of vibrant research domains within organizational behavior: organizational culture, demography, leadership, and creativity.

2. *Organizational Culture: Selection, Socialization, and Person–Culture Fit*

Research on organizational culture has demonstrated that norms and values can shape individual behavior. Recent research has also demonstrated that behavior can be predicted by understanding how personal characteristics interact with cultural context. Below we discuss the four types of person–situation interactions as they relate to organizational culture.

2.1 *Organizational Culture as Strong Situation*

Though researchers agree that organizational culture is meaningful and important, they do not agree about how to define and measure it. Organizational culture

research typically draws upon theories in anthropology or cross-cultural psychology and uses ethnographic or quantitative methodologies. Some researchers have emphasized shared values and meaning (O'Reilly et al. 1991) while others have emphasized the ambiguity of cultural values and the existence of subcultures (Martin 1992). Most agree that the existence of different perceptions among members does not preclude the existence of shared assumptions. Organizational culture can be understood as a system of shared values and norms that define what is important and how organizational members ought to feel and behave. Through members' clarity about organizational objectives and their willingness to work toward these objectives, culture influences the attainment of valued organizational goals by enhancing an organization's ability to execute its strategy. This conceptualization focuses on two primary aspects of culture: first, the extent to which agreement and intensity exist about values and norms (culture strength), and, second, the extent to which these norms and values differ across settings in important ways (culture content).

Culture content refers to shared values and norms that define desirable behavior in the organization. For example, some cultures may stress the value of being a 'team player' while others may emphasize independence. For a culture to be strong, members must agree about which values are important and be willing to enforce them strenuously. Members of an organization may agree that they value, for example, being cooperative, but unless unequivocal and salient consequences result from compliance (e.g., cooperating with co-workers) and non-compliance (e.g., competing with or undercutting co-workers), the culture cannot be characterized as strong. A strong culture is characterized by social control in that members both agree about and are willing to enforce values and norms, even if such enforcement violates hierarchical lines of authority. Further, greater behavioral homogeneity among members, for better or worse, should be observed in stronger organizational cultures.

Cultural values are conveyed to members through the socialization processes that new recruits experience when they enter an organization. Though socialization often takes the form of training and contributes to increased task knowledge, information about the norms of the organization is also transmitted through training and other types of socialization (Morrison 1993). Interestingly, such normative information appears to be transmitted very early in new recruits' membership since cultural fit changes little after members' first year (e.g., Rynes and Gerhart 1990).

2.2 Behavioral Consistency Across Various Organizational Cultures

Although culture influences members' behavior, they are not merely passive recipients of social control. Individual characteristics may interact with the organization's culture to predict important behavioral and work outcomes. Research on person–organization fit, defined broadly as the compatibility between people and the organizations for which they work (Kristof 1996) has focused primarily on congruence between patterns of organizational values and patterns of individual values. New employees whose values more closely match those of the organization view it as more attractive, are more likely to join when made an offer, are able to adjust to it more rapidly, perform better, are more satisfied, and remain with it longer (e.g., Chatman 1991).

But, just as organizational cultures may differentially affect behavior, people may also differ in the extent to which their behavior is shaped by an organization's culture. For instance, compared to individualists, cooperative people were more responsive to the individualistic or collectivistic norms characterizing their organization's culture, and exhibited greater behavioral adaptation to each across the two types of organizational cultures (Chatman and Barsade 1995). Thus, congruence models, which presume an additive equivalence of person and situation factors and assume that greater congruence is always better, cannot fully explain behavior in organizations. Instead, a focus on mismatches between person and organization characteristics that challenge people to either act in accordance with the culture and thereby contradict enduring personal tendencies, or vice versa, might generate insight into such interactions.

Future research might focus on the set of characteristics, such as cooperation, self-monitoring, and self-esteem, that contribute to people's flexibility across situations. Identifying such characteristics could improve predictions of the behavioral expression of person characteristics both across time and across situations, and in particular, the extent to which an organization's culture and processes will influence member behavior. Research might investigate other organizationally relevant matches and mismatches to understand how different situations influence person–situation interactions. For example, examining mismatches between honest people and dishonest organizations may help to identify if and when good people 'turn bad.'

2.3 Individuals' Influence Over Organizational Culture

Founders and senior executives, who have legitimacy and authority, may be the most influential individuals in an organization. The person who creates the organization has significant impact on the strategies that the group develops to survive, and these are often preserved in the organization's values, practices, and promotion patterns well past the individuals' actual presence. Ironically, newcomers, who are at the other end of the spectrum in terms of authority and

legitimacy from founders, may also exert a great deal of influence on culture. This influence may be indirect; research has shown that the process of recruiting new members, including emphasizing an organization's attractions, strengthens current members' connection to their organization (Sutton and Louis 1987), and can exert mutual learning of the organizational culture (March 1991). An organization's culture may also be transformed by the entrance and exit of its members. Thus, the strength of an organization's culture may depend on the individual mobility of its members. Using simulation methods, organizations' length of service distributions have been examined as indicators of the extent to which members have been socialized into the culture's norms and values. This research has shown that variations in service distributions are positively associated with heterogeneity in organizational culture due to three distinct cultural processes: socialization, group cohesiveness, and common historical experiences (Carroll and Harrison 1998).

2.4 Situation Selection Based on Organizational Culture

Schneider (1987) developed the Attraction–Selection–Attrition (ASA) model, which posits that the key factor influencing the relationship between people and organizations is the fit between individual personality and the modal personality represented in the organization. People are differentially attracted to organizations on the basis of the anticipated fit between their own characteristics, such as personality traits, and organizational characteristics, such as culture (e.g., Cable and Judge 1997). Job seekers take an active role in the recruitment process and are attracted to careers and organizations reflecting their personal values and interests. Further, organizations have developed formal and informal strategies to identify and select recruits who will be compatible with their goals and working environment, elements that are strongly influenced by an organization's cultural values. For instance, rather than focusing on job-related criteria, selection appears to be based on such socially based criteria as 'personal chemistry,' and fit with the organization's values.

3. Organizational Demography

People may differ from each other in many ways that are both observable and unobservable. Members' observable demographic characteristics constitute a context for every other individual in the organization. For instance, a lone newcomer in an organization of established members may have a very different experience than a newcomer entering an organization characterized by members with varying tenure. Given the dramatic changes in US labor force demography over the past decade, relational demography, or the distribution of demographic characteristics within organizations, has become an active area of research.

Research on organizational demography improves upon past research on employee turnover. Turnover typically indicates a discrete departure event while demography focuses on length of service and rate of departure by taking into account the distribution of people by the length of time they have spent in the organization. This distribution can be influenced by a host of factors including personnel policies, technology, and the degree of unionization in the workforce (Pfeffer 1983). Most research on observable differences has examined the effect of demographic heterogeneity versus homogeneity on performance and has yielded mixed results (see Williams and O'Reilly 1998 for a comprehensive review). Some studies have demonstrated the positive effects of demographic heterogeneity for increasing the number of novel perspectives that can be used to solve difficult problems, increasing an organization's network of contacts, and facilitating organizational change. However, demographic heterogeneity may also lead to communication problems, less social integration in workgroups and greater turnover (e.g., Tsui et al. 1992). Demographic distribution among members across various attributes is an important situational factor that deserves further research since it can influence behavior differently depending on an individual's own demographic profile.

3.1 Behavioral Consistency Across Demographically Heterogeneous Workgroups

Research suggests that an organization's culture may influence the relationship between demographic diversity and work outcomes. For example, the purported benefits of a demographically diverse workforce are more likely to emerge in collectivistic organizations that make organizational membership more salient than membership in a demographic category (Chatman et al. 1998). An organization's culture may dictate the extent to which members view certain demographic characteristics as valuable and others as associated with lower status within an organization's informal social system (Spataro 2000). Furthermore, each attribute, such as tenure, race, or sex heterogeneity within a group, may differentially influence individual behavior and the combinations of various attributes can result in 'fault lines' which become stronger as more attributes align themselves in the same way (Lau and Murnighan 1998).

One explanation for the lack of clarity about the benefits and detriments of diversity is that researchers have neglected to consider key mediating processes between demographic composition and performance.

As shown in one recent study, heterogeneous groups initially created norms fostering independence rather than cooperation among members, but cooperative norms subsequently mediated the relationship between group demographics and performance (Chatman and Flynn, 2001). Similarly, a group's level of conflict influenced the impact of demographic heterogeneity on performance (Jehn et al. 1999).

4. Leadership

4.1 Some Individuals Can Effect Change More Than Others

Early leadership research focused on the physiological and psychological traits thought to be associated with exceptional leaders. These 'great man' theories of leadership examined the effects of personal characteristics such as height, physical appearance, and intelligence on leaders' emergence and effectiveness. This stream of research has its counterpart in more current studies examining the effects of self-confidence, extraversion, and energy level (e.g., House 1988). The aim of this approach has been to identify a leadership personality. However, it leaves many crucial questions unanswered, such as whether certain personal characteristics become more important than others depending on the organizational context, and why, regardless of formal authority, followers perceive some people as leaders and not others.

Contingency theories of leadership were advanced to explain how certain personal characteristics made a leader effective in certain situations (e.g., House and Baetz 1979). For example, leaders who initiated structure raised the productivity and satisfaction of a group working on a boring or simple task but lowered the productivity and satisfaction of a group working on a complex task, while a considerate leader raised the satisfaction and productivity of a group engaged in a boring task but had little effect on a group engaged in a task they found intrinsically interesting. Additionally, research showed that allowing members to participate in decision making increased commitment but depended on the amount of trust the leader had in his or her subordinates as well as the urgency of task completion (Vroom and Jago 1978). Thus, contingency theories of leadership were more comprehensive than trait theories; however, they still did not account for the interactive effects of leader characteristics and their situational contexts.

Recent research has focused on charismatic and transformational leadership, demonstrating that some individuals influence situations more than others. This research takes an interactional approach by conceptualizing leadership as a personal relationship between the leader and his or her followers. A leader must have certain interpersonal skills in order to inspire followers to set aside their goals and to pursue a common vision. Charismatic leaders are thought to have the ability to change their circumstances by increasing followers' motivation and commitment and, sometimes, to change the direction of the entire organization (e.g., Meindl et al. 1985). However, a leader is only charismatic if followers recognize him or her as such; followers must identify with the vision articulated by the leader. In one particularly exhaustive laboratory study of charismatic leadership (Howell and Frost 1989), confederates were trained to display qualities of a charismatic leader, such as projecting a dominant presence, articulating a large overarching goal, and displaying extreme confidence in followers' ability to accomplish this goal. In addition, norms were created in each group for either high or low productivity. In contrast to participants working under a considerate or structuring leader, participants working under the charismatic leader displayed higher task performance regardless of the group productivity norm. This finding suggests that leaders mold their styles in response to the situation. Moreover, some leaders are capable of changing the situation itself by changing followers' perceptions and motivation.

4.2 Leadership as a Function of the Strength of the Situation

Some researchers have been skeptical of a leader's ability to change situations, and have suggested that leadership is far more situationally determined than might have been assumed. The attributional theory of leadership suggests that because people tend to associate certain behaviors with those of a leader, leadership qualities will be attributed to a person displaying these behaviors (Calder 1977). Various biases emerge from this attribution, however. For instance, individuals tend to overestimate the amount of control a leader has over events that are, in fact, random or uncontrollable (Pfeffer 1977). Furthermore, a leader will be given more credit when situations are unfavorable (Meindl et al. 1985). Individuals' lay conceptions of leadership can be used or misused for the purposes of organizational impression management (Ginzel et al. 1993). In sum, leadership research has focused more on contexts and followers rather than on the characteristics of a focal leader. Future research might examine whether leaders reflect the personal characteristics of their followers or complement their weaknesses, if some followers have a greater psychological need for leadership than others, and the various substitutes for leadership, or how people can be compelled to lead themselves.

5. Creativity and Innovation

Research on creativity, like leadership, has moved from emphasizing traits, to considering the organizational context as well as the interaction between the

two. Creativity is generally viewed as distinct from innovation. Creativity occurs at the individual level, and refers to the generation of ideas that are both novel and useful. Innovation refers to the process of implementing these ideas at the organizational level.

5.1 Some People Are More Creative Across Situations Than Others

Early creativity research focused on the personality traits associated with individuals who had made creative contributions in their respective fields. People who have a wide variety of interests, are attracted to complex tasks, tolerant of ambiguity, and self-confident perform more creatively (e.g., Barron and Harrington 1981). However, trait research ignores how organizational contexts influence people's ability to perform creatively.

5.2 Some Situations Foster More Creativity Than Others

The degree to which one's job encourages intrinsic versus extrinsic motivation affects one's creative performance. Early studies suggested that when people worked on tasks that they found intrinsically interesting, adding an extrinsic reward lowered their interest in performing the task for its own sake (Deci 1971). More recent studies in organizational settings showed that individuals were most creative when they were intrinsically motivated. This intrinsic interest led them to stay focused, persist longer to complete difficult tasks, to 'play' with ideas, and suggest novel solutions (Amabile 1988). This suggests that the situational factors that are associated with decreases in intrinsic motivation, such as a controlling supervisory style, and an emphasis on external rewards, may indirectly diminish people's creative potential.

Organizational culture also influences creativity and innovation. Organizations that have mechanisms to express confidence in members, and communicate this confidence through the culture's core values, increase creativity among members (Kanter 1988). These findings are supported by a recent ethnography of a product design firm, IDEO, which creates new products by taking technologies from one industry and applying them in other industries where these technologies are unknown (Hargadon and Sutton 1997). At IDEO, employees are encouraged to create analogies between past technological solutions and current problems and to share them in brainstorming sessions. Further, employees are selected who have unique hobbies or skills that can be used to solve design problems.

Employees who have traits associated with creativity are more likely to thrive in organizations such as IDEO, which place an emphasis on creative performance. For example, employees produced the most creative work when they possessed the appropriate personal characteristics, were working on complex assignments, and were supervised in a supportive, noncontrolling fashion (Oldham and Cummings 1996). While the possibility that organizations can manage creativity through the use of a strong culture appears promising (Flynn and Chatman 2001), some worry that mechanisms of social control will stifle, not encourage creativity (Nemeth and Staw 1989). Future researchers might examine how organizations can achieve harmony and cohesion without sacrificing the flexibility and constructive conflict necessary for creativity and innovation. Many believe that creativity and innovation are the last sustainable competitive advantages and as such these issues will continue to generate a great deal of interest.

6. Conclusion

As theorists endeavor to develop a complete understanding of behavior in organizations, the analysis of both personal and situational factors, as conjoined units of behavior, will become increasingly fundamental to organizational studies. Person–situation interactions are much more complex than the simple addition of personal and situational characteristics, and these interactions may take a variety of forms. Some people are more responsive to situations than others, some situations can shape behavior to a greater degree than others, and some people have the unique capability to shape situations to their advantage or that of the organization. Furthermore, group members' personal characteristics may constitute the situational context as every individual responds to the personal or demographic characteristics of every other individual in the organization (Carroll and Harrison 1998, Tsui et al. 1992). These interactions are both complex and diverse: hence the field of organizational behavior must necessarily reflect this diversity. It will become increasingly important to observe these interactions over time. Since people and situations adjust to each other, cross-sectional research will not capture the dynamic interplay between the two. By pursuing interactional research over time, organizational behavior will increasingly evolve into a field that is as vibrant as the organizations it seeks to understand.

See also: Authority: Delegation; Innovation: Organizational; Leadership in Organizations, Sociology of; Leadership, Psychology of; Marx, Karl (1818–89); Organization: Informal; Organization: Overview; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture; Organ-

izational Culture, Anthropology of; Organizational Decision Making; Organizations: Authority and Power; Organizations, Sociology of; Schumpeter, Joseph A (1883–1950); Weber, Max (1864–1920)

Bibliography

Amabile T 1988 A model of creativity and innovation in organizations. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 10, pp. 169–212

Barron F, Harrington D M 1981 Creativity, intelligence and personality. *Annual Review of Psychology* **32**: 439–476

Cable D M, Judge T A 1997 Interviewers' perceptions of person–organization fit and organizational selection decisions. *Journal of Applied Psychology* **82**: 546–61

Calder B J 1977 An attribution theory of leadership. In: Staw B, Salancik G (eds.) *New Directions in Organizational Behavior*. St. Clair Press, Chicago

Carroll G, Hannan M 2000 *The Demography of Corporations and Industries*. Princeton Press, Princeton, NJ

Carroll G, Harrison J R 1998 Organizational demography and culture: Insights from a formal model and simulation. *Administrative Science Quarterly* **43**: 637–67

Chatman J A 1991 Matching people and organizations: Selection and socialization in public accounting firms. *Administrative Science Quarterly* **36**: 459–84

Chatman J A, Barsade S G 1995 Personality, organizational culture, and cooperation: Evidence from a business simulation. *Administrative Science Quarterly* **40**: 423–43

Chatman J, Flynn F 2001 The influence of demographic heterogeneity on the emergence and consequences of cooperative norms in work teams. *Academy of Management Journal*, **44**(5)

Chatman J A, Polzer J T, Barsade S G, Neale M A 1998 Being different yet feeling similar: the influence of demographic composition and organizational culture on work processes and outcomes. *Administrative Science Quarterly* **43**: 749–80

Davis-Blake A, Pfeffer J 1989 Just a mirage: The search for dispositional effects in organizational research. *Academy of Management Review* **14**: 385–400

Deci E L 1971 The effects of externally mediated reward on intrinsic motivation. *Journal of Personality and Social Psychology* **18**: 105–15

Flynn F, Chatman J 2001 Strong cultures and innovation: Oxymoron or opportunity. In: Cartwright S, Chatman J, Cummings T, Earley P C, Holden N, Sparrow P, Starbuck W (eds.) *International Handbook of Organizational Culture and Climate*. Wiley, Chichester, UK, pp. 263–88

Ginzell L, Kramr R, Sutton R 1993 Organizational impression management as a reciprocal influence process. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 15, pp. 227–66

Hargadon A, Sutton R I 1997 Technology brokering and innovation in a product design firm. *Administrative Science Quarterly* **42**: 716–49

House R J 1988 Power and personality in complex organizations. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 10, pp. 305–57

House R, Baetz M 1979 Leadership: Some empirical generalizations and new research directions. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 1, pp. 341–423

Howell J, Frost P 1989 A laboratory study of charismatic leadership. *Organizational Behavior and Human Decision Processes* **43**: 243–69

Jehn K A, Northcraft G B, Neale M A 1999 Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative Science Quarterly* **44**: 741–63

Kanter R M 1988 When a thousand flowers bloom: Structural, collective and social conditions for innovation in organizations. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 10, pp. 169–211

Kristof A L 1996 Person–organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personality Psychology* **49**: 1–49

Lau D, Murnighan J K 1998 Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review* **23**: 325–40

March J G 1991 Exploration and exploitation in organizational learning. *Organization Science* **2**: 71–87

Martin J 1992 *Culture in Organizations: Three perspectives*. Oxford University Press, New York

Meindl J, Ehlrich S, Dukerich J 1985 The romance of leadership. *Administrative Science Quarterly* **30**: 78–102

Morrison E W 1993 Longitudinal study of the effects of information seeking on newcomer socialization. *Journal of Applied Psychology* **78**: 173–83

Nemeth C J, Staw B M 1989 The tradeoffs of social control and innovation in small groups and organizations. *Advances in Experimental Social Psychology*. **22**: 175–210

Oldham G R, Cummings A 1996 Employee creativity: Personal and contextual factors at work. *Academy of Management Journal* **39**: 607–34

O'Reilly C A, Chatman J A, Caldwell D M 1991 People and organizational culture: A Q-Sort approach to assessing person–organization fit. *Academy of Management Journal* **34**: 487–516

Pfeffer J 1977 The ambiguity of leadership. *Academy of Management Review* **2**: 104–12

Pfeffer J 1983 Organizational demography. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 5, pp. 299–357

Rynes S, Gerhart B 1990 Interviewer assessments of applicant 'fit': an exploratory investigation. *Personnel Psychology* **43**: 13–35

Schneider B 1987 The people make the place. *Personnel Psychology* **40**: 437–53

Scott R W 1998 *Organizations: Rational, Natural and Open Systems*, 4th edn. Prentice Hall, Englewood Cliffs, NJ

Snyder M 1987 *Public Appearances, Private Realities: The Psychology of Self-monitoring*. WH Freeman, New York

Spataro S 2000 Not all differences are the same: The role of status in predicting reactions to demographic diversity in organizations. Unpublished doctoral dissertation, University of California, Berkeley, CA

Staw B M, Ross J 1985 Stability in the midst of change: The dispositional approach to job attitudes. *Journal of Applied Psychology* **70**: 469–80

Sutton R I, Louis M R 1987 How selecting and socializing newcomers influences insiders. *Human Resource Management* **26**: 347–61

Tsui A S, Egan T D, O'Reilly C A 1992 Being different:

Relational demography and organizational attachment. *Administrative Science Quarterly* **37**: 549–79
Vroom V H, Jago A G 1978 On the validity of the Vroom–Yetton model. *Journal of Applied Psychology* **63**: 151–62
Williams K Y, O'Reilly C A 1998 Demography and diversity in organizations: a review of 40 years of research. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*. JAI Press, Greenwich, CT, Vol. 20, pp. 77–140

J. A. Chatman and J. A. Goncalo

Peptides and Psychiatry

Since the 1970s, much has been learned about the role of peptides in the central nervous system (neuropeptides) and behavior. Originally thought to be rare and relatively unimportant compared to the classical neurotransmitters, they are now known to be nearly ubiquitous and extremely important in brain function. Like the 'classical' small molecule neurotransmitters, neuropeptides function as chemical mediators of neuron to neuron communication. However, unlike such classical neurotransmitters, the neuropeptides have often been evolutionarily conserved to act both as local transmitter modulators and as endocrine hormones, thus mediating complex patterns of integrated behavior.

The role of neuropeptides in facilitating complex aspects of behavior makes them ideal candidates in understanding the neurobiological bases of psychiatric disorders. Whereas the classical neurotransmitter systems are involved in the neuronal circuitry mediating all behavior and pathology, the neuropeptide systems appear anatomically distributed, but functionally more limited. Thus these systems may allow an understanding of the physiology and pathophysiology of behavioral repertoires along with the ability to treat psychiatric disorders with more specific treatment modalities. This article will briefly review neuropeptide biology and function in general. Several specific examples of neuropeptides with known behavioral significance are discussed, allowing some generalizations to be made connecting physiological behavior to pathologic disease states.

1. *Neuropeptide Biology*

Like the classical small neurotransmitters, neuropeptides also function as chemical mediators of neuron to neuron communication via presynaptic release onto postsynaptic receptors. Some neuropeptides serve their function primarily within the central nervous system (CNS), e.g., galanin and enkephalin. However, other neuropeptides serve as both neurotransmitters and endocrine hormones via pituitary release to act on peripheral sites, e.g., oxytocin and vasopressin. Other neuropeptides serve as local neuromodulators within the brain but also as hormone-releasing factors in the hypothalamus–pituitary system, e.g., corticotropin-releasing factor and thyrotropin-releasing hormone. Finally, some neuropeptides appear to have distinct and separate roles in CNS and periphery, e.g., neurotensin and cholecystokinin.

1.1 *Neuropeptide Production*

The classical neurotransmitters (i.e., glutamate, GABA, dopamine, serotonin, norepinephrine, and acetylcholine) are formed from small molecule precursors in the cytoplasm, often stored at the terminals where the neurotransmitter is packaged into vesicle pools. Control of pool size is a function of enzyme concentration and precursor availability. In contrast, peptides are the direct products of mRNA translation, essentially small protein products. Most neuropeptides are between 2 and 40 amino acids in length. They are initially formed as larger precursor proteins (preprohormones) by translation of mRNA into polypeptides that are then cleaved into various active smaller peptides. Within the cell body, vesicles of neuropeptides are packaged in the Golgi apparatus and then transported to the distal regions (axons and dendrites) of the neuron where they are released with neuronal activity. Control of neuropeptide availability is therefore largely a direct function of gene transcription and translation. Thus, change of neuropeptide expression occurs as a function of multiple hormonal and other modulatory influences on neuronal function.

1.2 *Neuropeptide Release and Inactivation*

Discovered as early as 1940 by the Scharrers, a husband and wife team, peptides are located in secretory granules (vesicles) at the neuronal terminal. Depolarization of the neuronal membrane leads to calcium influx locally, resulting in vesicle fusion with the membrane and release of peptide into the extracellular space. After release from presynaptic nerve terminals, the peptides diffuse across the synaptic cleft, binding to high affinity receptors on the postsynaptic membrane.

Termination of neuropeptide activity occurs when peptidase enzymes cleave the peptides into smaller fragments, disrupting their biologic activity. This is in contrast to small neurotransmitters that are removed from the synaptic cleft primarily by reuptake into the presynaptic terminal, with only modest breakdown by metabolic enzymes in the extracellular space. These

differences in inactivation generally lead to substantially longer periods of activity by the neuropeptides. Numerous peptidases have now been identified from endopeptidases to carboxypeptidases with differential specificities and different solubility vs. membrane-bound characteristics. These may differ among the different neuropeptide systems.

1.3 Neuropeptide Receptors

Neuropeptide receptors are generally of the G-protein coupled class of seven transmembrane receptors. Peptides released into the synaptic cleft activate these receptors via high-affinity binding and mediate internal signaling events via the cytoplasmic G-protein coupled effector. Activation of these receptors activates and inhibits a variety of second messenger systems and can result in cell firing, modulation of membrane threshold, phosphorylation and dephosphorylation events, and alteration of gene expression. These changes can thus have direct effects on the target cells or can modulate effects of other neurotransmitters. Furthermore, as in the hypothalamic–pituitary system, activation of hormone-releasing factor receptors causes release of other neuropeptides into the periphery, e.g., thyrotropin-releasing hormone activates release of thyroid-stimulating hormone from the thyrotrophs in the pituitary.

1.4 Colocalization of Neuropeptides and Classical Neurotransmitters

Neuropeptides are often found in neurons that also contain one of the small classic neurotransmitters. However, the neuropeptides and small neurotransmitters often have different dynamics of release and subserve different functions. For example, in some neurons with slow firing, release is limited to small neurotransmitter vesicles. With rapid burst firing and prolonged depolarization, however, the calcium concentration in the presynaptic terminal is significantly elevated, leading to release of neuropeptide vesicles in addition to neurotransmitter. Thus the different dynamics of release along with differential receptor activation allow these colocalized mediators to carry different functional signals in neuronal circuitry.

2. Behavioral and Psychiatric Aspects of Neuropeptides

The impetus for understanding neuropeptide biology since the 1970s has led to great advances in endocrinology and has had a significant impact on psychiatry. It is becoming increasingly evident that many psychiatric disorders are associated with neuropeptide abnormalities. Alterations in specific neuropeptide-containing neurons have been shown in Alzheimer's disease, Parkinson's disease, Huntington's disease, mood disorders, anxiety disorders, eating disorders, and schizophrenia.

There are over 100 neuropeptides now identified and currently being scrutinized. This rapid expanse of knowledge cannot be summarized here. Therefore, a few examples of neuropeptide systems are discussed below which have known behavioral, physiological, and pathological significance. As with those discussed below, many neuropeptides have multiple roles in somatic sites, the central nervous system, and as hormonal mediators linking them. Understanding the physiologic role of these molecules will likely guide the understanding of pathophysiology and treatment of the psychiatric disorders in which they are involved.

2.1 Corticotropin-releasing Factor—Model of Integrated Stress Response

Corticotropin-releasing factor (CRF) was originally identified in crude form in 1955 and finally structurally identified in 1981 as a 41 amino acid peptide. It is the primary stimulant of adrenocorticotropic hormone (ACTH) release from the anterior pituitary, initiating the hypothalamic–pituitary–adrenal (HPA) peripheral stress response. It clearly also functions as a putative neurotransmitter in the brain coordinating global responses to stressors. In higher organisms, CRF appears to mediate a complex behavioral program integrating the endocrine, immunologic, autonomic, and behavioral responses to stress.

CRF appears to be of crucial importance in many psychiatric disorders. Dysregulation of the HPA axis in major depression has been one of the most consistent findings in biological psychiatry since these abnormalities were first identified. As the field continues to understand the nature vs. nurture debate of psychopathology, the role of the endogenous stress response system in the stress–diathesis model of affective and anxiety disorders becomes paramount.

2.1.1 CRF, the physiologic mediator of stress. There is now abundant evidence that CRF and related peptides are the primary central nervous system (CNS) mediators of stress. Stress, *per se*, is generally defined as extraordinary demand on the organism or alterations in psychological homeostatic processes. Initially, the importance of CRF was thought to be primarily endocrine via activation of ACTH from the pituitary, with subsequent release of cortisol and other glucocorticoids from the adrenal gland. Since the 1980s, however, its role in the brain independent of the

HPA axis has been firmly established. Consistent with its role as a neurotransmitter mediating complex repertoires, CNS administration of CRF mimics many of the behaviors and autonomic responses seen with physiologic stress.

Neurons expressing and releasing CRF are located in neuroanatomic locations which are thought to be critical for the stress response, including areas of neocortex, limbic system, and midbrain. In the limbic system, the central amygdala and bed nucleus of the *stria terminalis* are regions of the so-called 'extended amygdala' believed to be involved in stress and anxiety. CRF neurons from these regions project to many midbrain and brainstem sites mediating internal stress responses. The *locus coeruleus* and the *raphe nucleus* in the midbrain receive these CRF projections, influencing the activation of norepinephrine and serotonin circuits, respectively. Thus stress activation of the CRF system can disrupt attention and vigilance, concentration and memory, sleep and wakefulness via these transmitter systems, and it is likely involved in their abnormal functioning in mood and anxiety disorders.

There are multiple experimental paradigms further implicating the central CRF system as the prime mediator of the acute and prolonged stress response. Rodents administered CRF directly into the CNS show decreased reproductive behaviors, altered sleep, increased grooming, increased signs of despair, increased neophobia, and altered appetite, all of which mimic the behavioral stress response. Furthermore, rodents raised with early life stress (primarily maternal separation) hypersecrete CRF from the hypothalamus and amygdala, both during the stress and later as adults.

Administration of CRF into the CNS of nonhuman primates elicits many signs of behavioral despair including increased vocalization, decreased exploration, and increased huddling. These effects are reversed by CRF receptor antagonists, as are the behavioral consequences of environmental stress. Consistent with these experiments are early studies of maternal deprivation in young Bonnet macaque monkeys. These experimental animals exhibited sustained behavioral despair with decreased locomotion, decreased exploration, and altered food intake, similar to those animals with centrally administered CRF. These animals persistently exhibited a prolonged activation of the HPA axis with elevated cortisol and ACTH peripherally.

Stress exposure to maternal–infant dyads without separation appears to have similar long-term effects. Young Bonnet macaques raised in naturalistic settings by mothers with an unpredictable food supply showed increased signs of anxiety compared to those with a predictable food supply. These symptoms were accompanied by elevated CSF CRF concentrations. Remarkably, the grown youngsters consistently showed continued signs of anxiety and affective disturbance after removal from these stressful living constraints, and their CNS CRF and HPA axis activity remained hypersensitive to other stressors later in life.

2.1.2 CRF, the pathologic mediator of depression and anxiety. The preceding paragraphs suggest that the CRF system must play a critical role in the development of stress-related disorders. It is now becoming increasingly clear that depressive and anxiety disorders have significant inheritable and environmental components. The likelihood of an individual to develop an affective disorder is related to genetic vulnerability in addition to his or her biologic capacity to respond to stress. The experiments above suggest that this capacity is somewhat plastic, influenced by extent of early life stress. Multiple lines of evidence have shown that CNS CRF systems are altered in mood and anxiety disorders.

Many studies have reproduced the early findings that CRF CSF concentrations are elevated in depressed individuals compared to healthy comparison subjects, and that the levels resolve with resolution of the depressed state. Dysregulation of the HPA axis also continues to be validated. The CRF stimulation test (depressed patients show a blunted ACTH response) along with the Dexamethasone Suppression Test (depressed patients do not suppress cortisol) as well as the combined Dex/CRF test developed by Holsboer remain very sensitive tests of major depressive disorder. Furthermore, MRI and CT evidence has revealed increased pituitary and adrenal gland size in depression, consistent with hypersecretion of CRF and ACTH, respectively.

The role of CRF in anxiety disorders is also likely critical, though less well established than in depression. In animal models, direct CNS CRF administration is markedly anxiogenic, and conversely CRF receptor antagonists are anxiolytic. In experimental models, alprazolam, a short acting benzodiazepine, leads to decreased *locus coeruleus* CRF concentrations after acute administration, an effect which may mediate its anxiolytic action. Patients with post-traumatic stress disorder (PTSD) also exhibit elevated CRF concentrations in CSF and a blunted ACTH response to CRF challenge. However, they are hypocortisolemic and exhibit supersuppression to dexamethasone, suggesting that they differ from patients with depression.

In summary, the effects of CRF administration mimic acute anxiety and chronic depression, CRF_1 receptor antagonists block some of these symptoms in model systems, and patients with depressive and anxiety disorders show alterations of the central and peripheral CRF–HPA system. These data are nicely linked to a growing understanding of the developmental role of stress on the CRF system and its possible *sequellae* later in life. This allows a reformulation of the stress–diathesis model: early untoward life

events associated with development of depression and anxiety in adulthood give rise to long-lasting changes in CRF neurons, thus increasing the individuals' vulnerability to affective and anxiety disorders. See *Depression*; *Anxiety and Anxiety Disorders*.

2.2 Oxytocin and Vasopressin—Models for Complex Behavioral Regulation

Oxytocin (OT) and vasopressin (AVP) are members of the same nine amino acid class of peptides and have similar behavioral effects. Due to their interactions and similar effect on behavior, they will be considered together here. These peptides are best characterized in terms of organizing a complex behavior, that of social affiliation. This includes reproductive behavior, parental behavior, parent–child bonding, and grooming–sitting together behaviors in nonhuman primates. They are recently evolved and mediate many mammalian-specific reproductive and social behaviors. However, there are profound species differences in receptor expression, which in some cases mediates species-specific sets of behavior.

2.2.1 Oxytocin and vasopressin mediate social interaction. The best known peripheral effects of OT are facilitation of lactation–milk ejection, and uterine contraction during labor. AVP serves in the periphery primarily to regulate blood pressure and plasma volume. These functions are mediated via the magnocellular neurons projecting from the paraventricular nucleus (PVN) of the hypothalamus to the posterior pituitary where the peptides are released into the general circulation. Central effects are mediated via the PVN projection of parvocellular neurons to areas of the limbic system, neocortex, and autonomic areas of the brainstem. These central effects include modulation of reproductive behavior, parental behavior, infant attachment, and other prosocial effects.

One of the best studied models of affiliation is the formation of pair bonds. Multiple lines of evidence have demonstrated that in several rodent model systems, OT and AVP are critical for these behaviors. The prairie vole has been an excellent model for these studies because these animals are highly affiliative, show strong monogamous behavior, frequent physical contact, nest building, and shared parental care. OT in females and AVP in males appears to be responsible for formation of partner preference. In this species, mating facilitates pair bond formation, and OT and AVP are released centrally with mating. OT antagonists in females and AVP antagonists in males block formation of the partner preference, whereas the respective agonists administered centrally facilitate partner preference in the absence of mating. These results suggest that these peptides are both necessary and sufficient for this aspect of pair bonding. Furthermore, a closely related species, the montane vole, shares many similar nonsocial behaviors with the prairie vole, but they are generally isolative, are clearly not monogamous, and males show little parental care. The critical difference in species appears to be the regulation of the OT and AVP receptor distribution in the brain. For example, the prairie vole has OT receptors in regions important for reward (*nucleus accumbens* and prelimbic cortex), whereas the montane vole shows primary distribution in the lateral septum, possibly responsible for the peptide's effect on self-grooming, but poor socialization. This is supported by evidence that centrally administered AVP increases affiliative behavior in the monogamous prairie vole, but not the montane vole. Furthermore, transgenic mice expressing the AVP receptor in the prairie vole brain distribution pattern results in a prosocial response to centrally administered AVP.

Parental behavior and infant attachment also appear to be dependent on OT and AVP systems. Parturition is associated with significant shifts in maternal behavior in some species, including relentless nest building, licking, grooming, and protection of the pups. These behaviors are blocked by central administration of OT receptor antagonists. Neonatal voles crave social contact with 'distress calls' vocalized in as early as five-day old pups with maternal separation. Centrally administered OT in eight-day old pups results in significantly reduced distress calls, but no evidence of sedation or other behavioral change, suggesting that OT influences the maternal separation response.

Evidence for other social behaviors mediated by these systems include increased levels of social grooming by centrally administered OT or AVP, and the correlation of decreased plasma OT with increased social stressors in primates. Finally, there is an interesting cognitive response to these peptides. Oxytocin appears to facilitate extinction and attenuates passive avoidance. It is hypothesized that these effects may allow the relinquishing of normal social avoidance to establish critical social bonds.

In summary, oxytocin and vasopressin have important roles in the initial affiliative and territorial stages of reproductive behaviors that are dependent on gender and species, along with influences on parental, attachment, and group social behavior. These neuropeptides appear to alter the affective processing of social stimuli via regionally specific regulation of neuropeptide receptors.

2.2.2 Possible roles for oxytocin and vasopressin in psychopathology. The data for these neuropeptides in animal affiliation suggest that they might also be important in human psychopathology in which relationships or social attachments are abnormal. Measures

of CSF OT and AVP concentrations have yielded inconsistent results in patients with schizophrenia or major depressive illness, though only a handful of studies have been performed. However, post mortem studies have suggested significant changes in several pathological processes. Significant increases in the number of hypothalamus AVP and OT cells have been found in post mortem studies of depressed subjects. Post mortem studies also revealed a > 40 percent decrease in OT cells in the Prader–Willi syndrome, a genetic disorder notable for obesity, mental retardation, hyposexuality, and inappropriate behavior. Several studies have consistently found elevated OT concentrations in OCD subjects, along with normal to elevated concentrations of AVP. One possible explanation for a role in OCD would be the relationship between OT and AVP and grooming behavior and their role in extinguishing avoidant behavior.

Most interesting would be the human disorders along the schizoid spectrum to autism, in which limited interpersonal skills and social impairment are paramount. With autism, the onset is clearly developmental prior to three years old, with most families noting decreased social interest in the first months of life. It is clearly genetically transmitted with monozygotic twins having approximately 36–91 percent concordance, compared to < 0.1 percent in the general population. In one study of autistic children, plasma OT was about half that of age-matched controls. Furthermore, autistic children failed to show the normal developmental increase in OT compared to controls. Although many clinical studies remain to be done (no study of CSF samples is yet reported), the preclinical data suggest that the OT and AVP systems are critical for normal social behavior and memory. Studies are also yet to be reported on the schizoid and avoidant spectrum disorders, but given the specificity of the OT and AVP systems, one suspects that there may be some abnormalities. Given the complexity of OT and AVP receptor gene regulation with known polymorphisms in humans, one could hypothesize that some aspects of social temperament are genetically influenced by the precise expression patterns and regulation of these behaviorally relevant peptides and receptors.

2.3 Cholecystokinin and Neurotensin—Models of Complex Neuromodulators

Cholecystokinin (CCK) and neurotensin (NT) are unrelated, evolutionarily conserved peptides of 8 and 13 amino acids, respectively. They are discussed in this section as examples of the large group of peptides that less clearly integrate a specific behavioral repertoire. They are found in numerous tissues and subserve many different roles including local paracrine, endocrine, and neurotransmitter functions. However, their similarities are intriguing. They share endocrine functions regulating feeding behavior and induction of satiety in the periphery along with neuromodulatory functions interacting with the primary reward and appetitive circuitry in the central nervous system. Furthermore, there is mounting evidence that they each may play significant roles in the pathophysiology of schizophrenia, and in the case of NT in the mechanism of action of antipsychotic drugs.

2.3.1 Cholecystokinin and neurotensin involvement in feeding, autonomic regulation, and nociception. CCK was one of the first of the gastrointestinal hormones discovered and one of the most abundant neuropeptides in the brain. In contrast, NT was initially discovered from bovine hypothalamic extracts and later found to be an important gastrointestinal hormone. In the periphery, CCK and NT are each released by the small intestine shortly after a meal, and remain elevated in the plasma for several hours. They both stimulate pancreatic and biliary secretion along with modulating small and large intestinal motility. Although they do not cross the blood–brain barrier, they have an apparently conserved function in the CNS. They are both thought to have a role in mediating satiety and inhibiting feeding within the CNS.

They share the property of regulating autonomic and nociceptive information from the midbrain to the cortex. CCK and NT released in the midbrain have been shown to directly modulate excitatory visceral transmission through the parabrachial nucleus to the thalamus. Furthermore, both have been shown to exert potent analgesic effects when injected directly into the CNS. Thus they appear to be involved in regulating sensory information representing the internal state of the organism. Modulation of autonomic outflow is evident by their role in regulating vascular tone and thermoregulation.

2.3.2 Cholecystokinin and neurotensin modulation of dopaminergic systems. In addition to the above functions, CCK and NT appear to be integrally involved in modulating the dopaminergic systems within the brain. Although they are co-localized with other neurotransmitters and modulate serotonin, acetylcholine, and other neuropeptides in some areas, there appears to be a consistent role for these peptides in modulating dopamine circuits. Dopamine (DA) is found in three principle pathways: the mesolimbic/mesocortical pathways involved in motivation, reward, and cortical processing; the nigroneostriatal pathways involved in mediation of locomotor movements; and the tuberoinfundibular pathways controlling pituitary release of prolactin. The role of CCK

and NT in modulation of the mesolimbic circuitry originating in the ventral tegmental area (VTA) is of particular interest. Within this mesencephalic nucleus, NT and CCK are co-localized in many dopaminergic neurons that project to the *nucleus accumbens*. NT and CCK are also found in cortical neurons, likely modulating the terminal projections of dopaminergic axons.

Both NT and CCK induce an initial increase in firing rate of dopaminergic neurons when released into VTA, along with an increase in DA release in the *nucleus accumbens*. This increased firing appears to be due to inhibition of D2 autoreceptors by decreasing affinity for DA. This change is thought to occur via intracellular transduction mechanisms and allosteric receptor–receptor interactions between the D2 receptor and the CCK and NT receptors. In contrast, at high doses and with chronic treatment, these peptides significantly decrease spontaneous activity of dopaminergic firing via depolarization inactivation. Thus, these neuropeptides have the ability, *in vivo*, of modulating dopaminergic function at the level of the midbrain and in the projection areas.

Whether the varied roles of feeding behavior, autonomic and nociceptive regulation, and dopaminergic modulation are functionally related is unclear. Some have speculated that these peptides were evolutionarily involved in the primary functions of feeding and feeding regulation. The CNS must be able to organize homeostatic mechanisms around this critical behavior including cardiovascular tone and temperature regulation because much of the blood volume goes to the gut postprandially. Relative analgesia postprandially might be crucial for maintaining the primary function of digestion. Finally, as complex behavior evolved, these neuropeptide systems may have served to couple the primary reward of feeding to the central motivation/reward circuitry.

2.3.3 Cholecystokinin, neurotensin, and schizophrenia. The underlying pathology in schizophrenia remains unknown. The empirical observation that all effective antipsychotics were also dopaminergic antagonists contributed to the original DA hypothesis. However, more recent research has failed to find consistent abnormalities within the dopaminergic system, *per se*, in schizophrenic patients. Additionally, evidence for disturbances in the glutamatergic, GABA-ergic, serotonergic, and cholinergic systems have accrued. One way to reconcile the vast data is that the activity of multiple neurotransmitter systems may be altered via a dysregulation of normal activity. Because neuropeptides are known to play an important role in neurotransmitter modulation, these systems have received increasing attention as both possible mediators of the pathophysiology in schizophrenia as well as potential targets for novel therapeutics.

Although several neuropeptide systems may contribute to the abnormal circuitry in schizophrenia, the majority of data point to NT and CCK as being the most likely candidates of the known neuropeptides. Their role as modulators of the dopaminergic system was outlined above. Additionally, both the serotonergic *raphe nuclei* and brainstem cholinergic nuclei are innervated by peptidergic neurons, especially NT. Within the DA system, these neuropeptides mimic the effects of antipsychotic medication in their acute activation of the VTA dopaminergic neurons and in their chronic inactivation of these neurons. Furthermore, treatment with an NT receptor antagonist appears to mimic atypical antipsychotics in that chronic administration leads to decreased VTA firing, with no effect on the nigroneostriatal system. This effect is thought to be mediated via feedback onto the VTA from the prefrontal cortex where local injection of the NT antagonist has the same effect. This would also be consistent with newer theories of schizophrenia which invoke disruption of large mesolimbic–cortical activation loops as mediating the pathophysiologic events.

Behavioral studies with centrally administered NT and CCK also mimic effects of typical and atypical antipsychotic medication. They both decrease stimulant-induced locomotion and spontaneous locomotion. They also decrease avoidance, but not escape, responding in a conditioned avoidance paradigm. Perhaps most importantly, they both appear to effect models of sensorimotor gating, which is becoming increasingly accepted as a critical objective symptom of schizophrenia. The hypothesized decreased ability of these patients to screen for appropriate sensory data may lead to 'involuntary flooding' of indifferent sensory input. This might lead to cognitive abnormalities, thought disorganization, and the positive symptoms (hallucinations and delusions) which are hallmarks of schizophrenia. Prepulse inhibition (PPI) of the startle reflex (defined by decreased startle to sound if preceded by weaker sound), and latent inhibition (LI) (reduced associative learning if subject is pre-exposed to stimulus), are the two most common measures of sensorimotor gating. In humans, PPI and LI are disrupted in schizophrenic patients, and are normalized in some studies of schizophrenics treated with antipsychotic medication. In animal models, dopaminergic agonists and other psychomimetic compounds disrupt PPI and LI, and these are returned to normal with antipsychotic treatment. NT and to a lesser extent, CCK, also clearly modulate PPI and LI in animal models. These results provide substantial evidence for the role of NT and possibly CCK in the circuits that are affected by antipsychotic medication.

Finally, there is significant evidence of abnormal regulation of CCK and NT in the CNS of schizophrenic patients. Although no consistent changes have been found post mortem with NT levels, NT receptor binding has reproducibly been found to be decreased

in some cerebrocortical regions in schizophrenia. Concentrations of NT in the CSF are reproducibly decreased in nonmedicated schizophrenics. These levels return to normal with effective treatment, and lower NT concentrations have been correlated with more severe psychopathology, particularly negative symptoms. Similar findings of NT abnormalities have not been found in affective disorders, anorexia, or Alzheimer's disease, suggesting some specificity of the findings in schizophrenia. Post mortem CCK levels have been consistently decreased in cerebrocortical and limbic regions of schizophrenic subjects, but CSF CCK changes have not been as reproducible. See *Schizophrenia, Treatment of*.

3. Summary

Since the 1970s, the biology and behavioral roles of many neuropeptides have been elucidated. They have moved from a position of relative unimportance in behavioral neuroscience to pre-eminence. In their roles as neurotransmitter and neuromodulator, paracrine and endocrine hormone, individual neuropeptides may at times subserve different functions in the CNS and periphery. However, many neuropeptides share a conserved function organizing different neural systems with the periphery in behaviors important for coordinated activity of the organism.

Corticotropin-releasing factor is essential in the physiologic mediation of stress, and likely critical in the pathophysiology of depression and anxiety. Oxytocin and AVP subserve many social roles from bonding to parental behavior, and they may underly some pathologic processes involving socialization such as autism. Finally, the peptides NT and CCK have similar roles in feeding, autonomic and analgesic regulation, and DA modulation in the brain. Their dysfunction may contribute to the pathophysiology of schizophrenia.

The knowledge of these neuropeptide systems in psychopathology provides wonderful opportunities for future rational therapeutic approaches. CRF_1 receptor antagonists show great promise in preclinical and early clinical trials for the treatment of depression and anxiety. Oxytocin has been found to improve socialization in some experiments of schizophrenic patients and may provide future hope for autistic disorders. Neurotensin and CCK receptor agonists provide important targets for future system-directed treatment options in schizophrenia. In summary, via their role in the organization of behavioral repertoires, neuropeptide systems may ultimately elucidate mechanisms and provide novel treatment options for many psychiatric diseases.

See also: Endocrinology and Psychiatry; Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of; Neurotransmitters; Women and Psychiatry

Bibliography

Arborelius L, Owens M J, Plotsky P M, Nemeroff C B 1999 The role of corticotropin-releasing factor in depression and anxiety disorders. *Journal of Endocrinology* **160**: 1–12

Argiolas A, Gessa G L 1991 Central functions of oxytocin. *Neuroscience Biobehavavioral Reviews* **15**: 217–31

Bissette G, Nemeroff C B 1995 The neurobiology of neurotensin. In: Bloom F E, Kupfer D J (eds.) *Psychopharmacology: The Fourth Generation of Progress*. Raven Press, New York

Brownstein M G 1993 Neuropeptides. In: Siegal G, Agranoff B, Albers W, Molinoff P (eds.) *Basic Neurochemistry*, 5th edn. Raven Press, New York

Burbach J P, de Wied D (eds.) 1993 *Brain Functions of Neuropeptides: A Current View*. Parthenon, Carnforth, NY

Fink H, Rex A, Voits M, Voigt J P 1998 Major biological actions of CCK—a critical evaluation of research findings. *Experimental Brain Research* **123**: 77–83

Gariano R F, Groves P M 1989 A mechanism for the involvement of colocalized neuropeptides in the actions of antipsychotic drugs. *Biological Psychiatry* **26**: 303–14

Hokfelt T G, Castel M-N, Morino P, Zhang X, Dagerlind A 1995 General overview of neuropeptides. In: Bloom F E, Kupfer D J (eds.) *Psychopharmacology: The Fourth Generation of Progress*. Raven Press, New York

Insel T R, O'Brien D J, Leckman J F 1999 Oxytocin, vasopressin, and autism: Is there a connection? *Biological Psychiatry* **45**: 145–57

Kinkead B, Binder E B, Nemeroff C B 1999 Does neurotensin mediate the effects of antipsychotic drugs? *Biological Psychiatry* **46**: 340–51

Koob G F 1999 Corticotropin-releasing factor, norepinephrine and stress. *Biological Psychiatry* **46**: 1167–80

Modahl C, Green L A, Fein D, Morris M, Waterhouse L, Feinstein C, Levin H 1998 Plasma oxytocin levels in autistic children. *Biological Psychiatry* **43**: 270–7

Nemeroff C B (ed.) 1991 *Neuropeptides and Psychiatric Disorders*. American Psychiatric Press, Washington, DC

Owens M J, Nemeroff C B 1992 The physiology and pharmacology of corticotropin-releasing factor. *Pharmacology Review* **43**: 425–73

Plotsky P M, Owens M J, Nemeroff C B 1995 Neuropeptide alterations in mood disorders. In: Bloom F E, Kupfer D J (eds.) *Psychopharmacology: The Fourth Generation of Progress*. Raven Press, New York

Rostene W H, Alexander M J 1997 Neurotensin and neuroendocrine regulation. *Frontiers in Neuroendocrinology* **18**(2): 115–73

Smock T, Albeck D, Stark P 1998 A peptidergic basis for sexual behavior in mammals. *Progress in Brain Research* **119**: 467–81

Wang Z X, Young L J, DeVries G J, Insel T R 1998 Voles and vasopressin: A review of molecular, cellular, and behavioral studies of pair bonding and paternal behaviors. *Progress in Brain Research* **119**: 483–99

K. J. Ressler and C. B. Nemeroff

Perception and Action

Perceiving without acting is hardly possible: scrutinizing an object visually presupposes directing the eyes at it, which sometimes involves moving the head or even

the whole body; a tactile investigation of an interesting object requires moving the fingers across its surface; and localizing a sound source is much easier when moving the ears and head (Dewey 1896). Likewise, acting without perceiving makes no sense; after all, actions, defined as goal-directed behavior, aim at producing some perceivable event—the goal. Performing an appropriate action requires perceptual information about suitable starting and context conditions and, in the case of complex actions, about the current progress in the action sequence. Thus, perception and action are interdependent. However, in the course of evolution humans have developed several ways to relate action to perception, ranging from simple and rigid stimulus–response (S–R) reflexes shared with many species to flexible and adaptive behavioral rules that can be tailored on the spot to the situation at hand.

1. *Reflexes*

In most species, behavior is triggered by the present situation and, thus, directly reflects the animal's immediate environmental conditions. Such reflexes can also be observed in humans, especially in infants, but here they constitute a negligible part of the behavioral repertoire. Interestingly, however, even reflexes already show the close mutual relationship between perception and action. Perhaps the best demonstration of this relationship is provided by the orientation reflex, which we experience when encountering a novel and unexpected event. On the one hand, this reflex inhibits ongoing actions and tends to freeze the body—a prime example of a stimulus–triggered response. At the same time, however, it draws attention towards the stimulus source by increasing arousal and facilitating stimulus-directed body movements. That is, the novel stimulus triggers actions that lead to a better perception of itself, thus producing a full S–R–S cycle.

Even though reflexes themselves represent a relatively inflexible way to coordinate perception and action, some researchers have suspected them to provide the basis for voluntary action (Easton 1972). For instance, the tonic neck reflex, an asymmetric pose observed in newborns with head and arm extended to one side and arm and leg flexed on the other, might facilitate the development of eye–hand coordination. Likewise, the stepping reflex, in which babies move their feet in succession when coming in contact with a solid surface, might underlie our ability to walk.

2. *Associations*

Human behavior (and that of other higher species) is surely much more flexible than exclusive control by reflexes would allow. Not only can we learn to react to particular environmental conditions and situations in a certain way, we also can unlearn what we have acquired and learn new relationships between situations and actions. Our ability to associate actions with stimulus conditions was the major topic of American behaviorism around 1890–1940, when most of the basic principles of S–R learning were empirically established. In particular, stimuli and responses become associated only if they co-occur in time, if there is at least some degree of contingency between them, and if the response is judged to be appropriate (Thorndike 1927). Although the empirical approach and the theoretical language of behaviorism has long been abandoned, its results still have a major impact on modern connectionism, the attempt to model psychological phenomena by means of artificial neural or neurally inspired networks on computers or in robots.

A major objection against the behavioristic approach to S–R learning relates to the assumed role of action outcomes. In behavioristic theories, the outcome of a given action is only judged regarding its hedonic value: a positive evaluation results in strengthening the association between the action and its antecedents whereas a negative evaluation weakens the association. However, whether it feels good or bad, the outcome of an action also informs the actor about its consequences, that is, about what he or she can achieve by performing it (Tolman 1932). And as actions aim at producing intended outcomes, perception–action learning should not be restricted to forming stimulus–response associations but comprise response–effect associations also. Indeed, studies reveal that humans (and also rats and pigeons) do acquire very stable associations between their actions and the consequences that these actions produce. Moreover, there is evidence that these associations play a major role in the control of voluntary action. That is, people control their overt behavior by forming or reactivating perceptual representations of intended goal events, which through learning have become associated with the motor patterns that have been—and must be—carried out to reach them.

3. *Rules*

The ability to learn new relations between environmental conditions and appropriate behavior provides an enormous gain in flexibility for an individual in allowing it to adapt to environmental change. Yet, learning and relearning take time and require at least some experience with a given new condition—it is thus necessarily reactive and, in a way, conservative in reflecting what one has become used to. Indeed, many forms of human behavior show these characteristics, as witnessed by the difficulties of introducing new behavioral patterns regarding, say, women, ethnic minorities, or the use of environmental resources.

Nevertheless, there are many instances where people can switch between different reactions to the same stimulus conditions on the spot, and hence more or less independently of the amount of experience with the situation. For example, even though one may have used one route to go to work 1,000 times already, it is easy to take an alternative route from one day to another, without having to unlearn the old habit or to learn the new one, and although the fact that in the case of absent-mindedness one might again go by the old route shows that the overlearned associations still exist. Therefore, people must be able to choose deliberately among alternative ways to relate environmental situations to their actions, that is, to select one out of many possible S–R rules, and behave accordingly.

The interplay between overlearned S–R associations and the voluntary implementation of intentionally selected S–R rules ('habit' vs. 'will') was a major issue in the psychological literature between 1870 and 1935, and Narziss Ach (1935) was the first to study this interplay empirically in a systematic way. In his 'combined method' he first had subjects build up new, strong associations between nonsense syllables and one type of action, and then asked them to perform another than the practiced action to the same stimuli. The resulting increase in reaction time and errors as compared with actions with neutral stimuli was taken to represent the individual 'will power' needed to overcome the previously acquired habits. After many years of neglect, the issue of how people implement and switch between alternative S–R rules received renewed interest during the 1980s and 1990s, especially in experimental and neuropsychology (Monsell and Driver 2000). The (still preliminary) results of this research provide new insights into the relationship between, and the control of, perception and action. As one would expect, the implementation of S–R rules takes effort and time. If people are to apply several sets of S–R rules concurrently, or in short succession, their performance is impaired and more error prone. In particular, they will sometimes apply the wrong rule or apply the right rule at the wrong time, especially if effort and attention deteriorate. However, even under ideal conditions, intentional control is not absolute. That is, voluntarily implementing a particular rule does not exclude or prevent contributions from overlearned S–R associations (habits). Therefore, although any S–R rule can be implemented in principle, those rules are easier to implement and to apply if they are in accordance with natural S–R relations, acquired S–R associations, and previously implemented S–R rules than if not. For instance, performance is better if stimuli and responses have features in common than with arbitrary S–R relations, so that pressing a left or right key, or turning to the left or right, is easier if signaled by the location of a stimulus that also appears on the left or right side than by the color, shape, or meaning of a stimulus.

The ability to implement and switch between the most arbitrary S–R rules allows for the highest degrees of behavioral flexibility but it is at the same time costly in terms of attentional resources. Therefore, it makes sense to see intentional rule implementation and the more automatic guidance by overlearned S–R associations as mechanisms that complement each other: implementing and acting out the same rules repeatedly and with sufficient success provides the basis for forming increasingly stronger associations between the corresponding environmental conditions and actions, in this way transforming rules into habits.

4. *Plans*

In one sense, perception and action are interdependent to a degree that makes it difficult to say where action ends and perception begins—just think of the eye-movement example. In another sense, however, evolution has provided humans with the important ability to temporally decouple our actions from perception. That is, actions can be planned, prepared, and scheduled long before their environmental trigger conditions occur. Although we are only beginning to understand how action planning works, there is evidence that it can be likened to self-automatization.

At least three phases of planning can be distinguished. First, the features of the intended action effect (i.e., the goal) need to be specified, such as the direction or end location of a reach, and the effector to be used. Next, the features belonging to one plan are integrated (i.e., their cognitive representations are temporarily coupled), to avoid confusing features belonging to different, concurrently maintained action plans. This integration process will often involve linking the plan to its anticipated trigger conditions, and hence to representations of the event that is intended to trigger the planned action. That is, planning comprises anticipating both perceptual and action events. Accordingly, planning an action changes our perception. For instance, after having planned a particular action, action-related objects and events are more salient and more easily processed, action-related features of objects are more conspicuous than other features, and objects appearing at action-related locations catch more attention than others.

The final step of planning consists of initiating the planned action. Initiation may be triggered internally, such as when a plan is carried out immediately after construction, or externally, such as when the anticipated trigger event is perceived. Interestingly, this last step does not seem to underlie overly strict cognitive control. For instance, internal initiation is more or less indifferent to the content of the plan or the progress of planning, and externally triggered plans can be seen to be called up even under inappropriate conditions or at the wrong point in time. Thus,

planning is like automatizing oneself by means of delegating control to future, often environmental, events.

See also: Attention and Action; Perceptual Constancy: Direct versus Constructivist Theories; Vision for Action: Neural Mechanisms

Bibliography

Ach N 1935 *Analyse des Willens*. Urban and Schwarzenberg, Berlin

Dewey J 1896 The reflex arc concept in psychology. *Psychological Review* **3**: 357–70

Easton T A 1972 On the normal use of reflexes. *American Scientist* **60**: 591–9

Gibson J J 1979 *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston

Jordan J S *Systems Theory and A Priori Aspects of Perception*. North-Holland, Amsterdam

Monsell S, Driver J 2000 *Control of Cognitive Processes*. MIT Press, Cambridge, MA

Piaget J 1946 *La Formation du Symbole chez l'Enfant*. Delachaux and Niestlé, Paris

Prinz W 1990 A common coding approach to perception and action. In: Neumann O, Prinz W (eds.) *Relationships Between Perception and Action*. Springer, Berlin, pp. 167–201

Thorndike E L 1927 The law of effect. *American Journal of Psychology* **39**: 212–22

Tolman E C 1932 *Purposive Behavior in Animals and Man*. Century, New York

Turvey M T 1977 Preliminaries to a theory of action with reference to vision. In: Shaw R, Bransford J (eds.) *Perceiving, Acting, and Knowing*. Erlbaum, Hillsdale, NJ, pp. 211–65

B. Hommel

Perception and Cognition, Single-/Multi-neuronal Recording Studies of

Recording of isolated single neurons from brain provides a means of estimating the relationship between an animal's behavior and the activity of the brain region from which the neuron was recorded. In this manner, neurophysiologists seek to explain all behaviors in terms of neural activity, and pair all neural activity with some perceptual, behavioral or cognitive event. Thus, the ultimate goal of neural recording and analysis is to determine how activity within a brain area relates to behavioral and/or cognitive events, in essence, seeking to answer the question: 'What do neurons encode?' This question is answered in part by recording and analyzing one neuron at a time, but there are additional characteristics of encoding which can only be answered by examining the interactions between neurons which likely underlie the overall activity of a brain region. The purpose of both single and multineuron recording, thus, is to obtain a more representative view of the neural activity in a particular brain region (Singer 2000).

Single and multiple neuron recordings are particularly powerful when applied to experiments in which the brain must form some neural 'calculation' in order to make the appropriate response. This can be manifested on many levels, from the purely mechanical, such as integration of multiple inputs in a simple neuronal circuit, to complex behavioral responses dependent on an animal's ability to 'remember' prior stimuli. Inasmuch as the recorded neural activity reveals a pattern that correlates to the cognitive process, the neural activity can therefore be considered to 'encode' information related to the process. This does not necessarily mean that the neural pattern is essential to cognition, but that the neural activity suggests information, which is available to the brain during cognition.

1. *Single vs. Multiple Neuron Recording*

Single neuron recording has revealed the function of many brain areas in terms of stimulus and response. For example, Hubel and Wiesel (1962) mapped columns of neurons in the occipital cortex that responded to specific visual stimuli. The identification of these 'ocular dominance columns'—in which successive columns responded to incremental rotations of a bar of light, and alternating regions mapped the left vs. right visual fields—were critical to understanding how visual information was encoded by the brain. Similarly, the identification of 'place cells' in hippocampus—neurons that fire only when the subject is in fairly limited regions of its environment (O'Keefe and Dostrovsky 1971), provided a correlation between neural activity and behavior, even if the exact purpose of that activity is still being debated. Single neuron recording has thus been used to identify the type of information processed by different brain areas, not just the five senses, but also control of muscles, attention, and memory. To this end, the brain has been mapped anatomically and, to a certain extent, neurophysiologically.

Broader questions of function and cognitive processing are difficult to answer with only single neuron recordings. In essence, all neurophysiological experiments use multiple neuron recording, since the activity of a single neuron is not necessarily representative of that brain area. Instead, many single neurons are recorded, then analyzed to see if a pattern is detected which is representative of neural activity in that region. As long as single neurons are recorded under con-

ditions that control variability between recording sessions, it is possible (within limits) to treat the collection of neural recordings as if they had been recorded simultaneously. Thus, single neurons are assembled in an 'ensemble' that represents the activity of multiple neurons from a single brain area. In this manner, Georgopoulos et al. (1986) observed that the activity of ensembles of neurons recorded from primate motor cortex predicted the direction of limb movement prior to actual movement. In this study, many neurons were recorded that exhibited different 'preferred' directions for limb movement. When the pattern of firing across neurons was analyzed, it was found that the direction of movement was represented by specific firing patterns prior to the actual movement. However, since the conditions under which two single neurons are recorded cannot be identical, the effect of those subtle variations on neural firing cannot be controlled. In addition, the activity of single neurons may be controlled by multiple stimuli, in which case a behavioral correlate may be shared or distributed across many neurons.

Recordings of multiple simultaneous neurons have played an important role in many exciting recent discoveries. Deadwyler and Hampson (1997) showed that patterns of neural activity in hippocampus predicted behavioral error in a short-term memory task. Laubach et al. (2000) similarly demonstrated that activity patterns in neocortical neurons predicted behavior as animals learned a motor task. Skaggs and McNaughton (1996) showed that during sleep, hippocampal ensembles 'played back' temporal firing patterns between neurons. The replayed patterns were identical to patterns generated by the animal's exploration of its environment during the waking state. In addition, Chapin et al. (1999) recently demonstrated direct neural control of a robotic arm by neural signals recorded from rats. The researchers detected patterns of neural activity in motor cortex that preceded limb movements required to operate the robotic arm. By programming a computer to detect the pattern in real-time, the computer was able to operate the arm without requiring the animal to move its limbs. These recent results indicate rapid progress toward goals of better understanding the relationship between the activity of neurons and the cognitive functions of the brain.

2. *Recording Many Neurons*

Multiple neuron recordings not only rapidly increase the total number of neurons characterized, but also allow examination of interactions between neurons that may be critical to cognitive processing. Information may simply be encoded as the mean amplitude of firing across neurons at a given point in time, or as correlation between neurons (von der Malsburg 1994), or as part of coherent oscillations between neurons (Singer and Gray 1995). No matter how the conditions are controlled during single neuron recording, interactions between neurons cannot be determined unless neurons are recorded simultaneously. The major advantage, therefore, of recording from multiple neurons in any brain area, is the potential for deciphering the codes that a brain area generates during cognitive or behavioral responding. The techniques used to record and analyze the multineuron data are critical, and not all types of multiple neuron analyses are appropriate in all cases. Misapplication of an ensemble technique can at best misrepresent the data, or at the worst lead to unworkable models of brain action. Several common uses of multiple neuron or ensemble analyses are examined below.

2.1 *Techniques for Multineuron Recording*

The ionic currents and action potentials generated by neurons produce changes in the electrical potential of the environment surrounding the neuron. These extracellular potentials are volume-conducted through the electrically-conductive extracellular fluid, and can be detected by electrodes that are either in contact with tissue or the cerebrospinal fluid. Skull surface or brain surface electrodes detect summed electrical activity resulting from the action potentials produced by millions of neurons throughout the brain, with no differentiation into single neurons. A glass pipette electrode will record neural activity from a region hundreds of microns in diameter; however, as above, the recorded neural activity is difficult to resolve into single neuron action potentials. Isolation of single neuron activity is possible with metal microelectrodes made from electrically conductive, but nonreactive material, such as tungsten, platinum-iridium, or stainless steel. Such electrodes are constructed with tip diameters approximate to the soma diameter of the neurons to be recorded and are insulated everywhere except at that tip. Sharp electrode tips ($1\,\mu$m) allow the electrode to be inserted with minimal tissue damage and allow the tip to be insinuated close to one neuron. In addition, the high impedance (resistance) of the tip ($> 1\,\mathrm{M}\Sigma$) ensures that electrical potentials from more distal neurons are attenuated compared to potentials generated by proximal neurons. More recent developments have shown that stainless-steel wires with tips up to $25\,\mu$m and tip impedances of $100\,\mathrm{k}\Sigma$ are capable of isolating single neurons if left implanted for days or weeks (McNaughton et al. 1983, Deadwyler and Hampson 1997). Single neuron action potentials are identified by biphasic waveform (typically ± 10–$100\,\mu$V amplitude, 500–$1000\,\mu$s duration). As long as the electrode tip is not moving relative to the neuron being recorded, each action potential generated by a single neuron should be the same. Therefore, the activity of a single neuron is isolated by identifying and recording all action potential 'spikes' with the

same characteristic duration and amplitude. Multiple single neurons can be identified from a single microelectrode by further identifying similar spikes from all those rejected, or by using multiple microelectrodes.

2.2 Types of Multiple Neuron Recording Electrodes

The simplest multiple electrode is a bundle of metal microelectrodes, of similar size and shape, with tips positioned to record from within a restricted area. These electrodes can contact each other without compromising recording ability, since they are insulated everywhere except at the tip. The advantage of microelectrode bundles is the large number of recording sites in a small area, the disadvantage is that there is no means of identifying where any one electrode tip is located. When electrode location is desired, arrays of microelectrodes can be used. The electrode tips are spaced in a pattern, so that each tip records from a separate, identified site. Another form of bundle electrode is the stereotrode or tetrode consisting of two or four wires twisted to keep all electrode tips in close proximity. The close proximity of recording tips on stereotrodes and tetrodes allows the recording of neurons on more than one wire, improving the ability to isolate single neurons. Finally, as stereotrodes and tetrodes can be constructed as arrays, to provide all of the benefits of both electrode types.

2.3 Discrimination and Identification of Multiple Neurons

As stated above, the essential principle of single neuron isolation and identification is the fact that the action potential from a given neuron should always be the same amplitude and duration. Thus, action potentials with the same waveform characteristics, likely were generated by the same neuron. The first technique for identifying single neurons recorded from microelectrodes is to sort the extracellular action potentials by waveform. Under some circumstances, notably when recorded from dense monolayers of neurons, two neurons may be equidistant from the electrode tip, and hence produce action potentials of similar sizes and shapes. Stereotrodes and tetrodes were developed to identify single neurons under these conditions (McNaughton et al. 1983), since it is unlikely that both neurons would be equidistant from both electrodes in a stereotrode, or from all four electrodes in a tetrode. By comparing the action potential waveforms simultaneously recorded at all electrode tips, it is possible to 'triangulate' and separate neurons that may have appeared the same to a single microelectrode.

Statistical techniques are also used to ensure that the action potentials recorded on single or multiple microelectrodes originate from a single cell. The intervals between action potentials of a randomly firing single neuron (the neuron's 'spike train') should be a *Poisson* function. A plot of the interspike-interval (i.e., between action potentials) should be similar to a graph of the *Poisson* probability distribution. In addition, autocorrelation of the spike train should reveal a symmetric distribution, with no firing within the time interval equal to the action potential refractory period for that neuron. Uniform interspike-interval distributions, or spikes occurring during the refractory period indicate a failure to isolate recordings into single neuron data (see Nicolelis 1999). These analyses, plus consistency in behavioral correlation, cross-correlation between neurons, and stability of action potential waveforms over an extended time (or multiple days) all indicate satisfactory identification of the single neuron (whether recorded with single wires, stereotrodes or tetrodes) and thus support the recording of multiple single neurons from bundled or arrayed microelectrodes.

3. Analysis of Multiple Neuron Recordings

The basic method of multiple neuron recordings is to analyze each neuron's spike train independently. This typically involves constructing a plot of average neural activity correlated to a specific cognitive event. For instance, perievent histograms depict mean firing rate synchronized to (peri-) a behavioral stimulus (event); similarly, a 'place field' recorded for hippocampal neurons consists of the correlation of neural firing rate with the animals instantaneous location in it's environment. Multiple neuron activity is then used to examine to overall average or range of responses for neurons in a given brain area. More complex analyses of multiple spike trains were first described by Gerstein and Perkel (1969) prior to development of many of the above multiple neuron recording techniques. These techniques concentrated on determining cross-correlations between neurons, and statistically extracting repeated patterns of activity across many neurons with a view toward determining the connections between neurons. Many later analyses derived from these earlier techniques.

3.1 Population Vectors

A popular type of analysis is a direct examination of ensemble firing rate patterns corresponding to specific motor movements or behavioral events. This analysis has been termed the 'population vector' and is potentially quite powerful, especially in terms of encoding directional and velocity parameters. Population vectors consist of a set of measurements of the mean firing rates for each neuron in the ensemble corresponding to discrete times when unique events, or stimuli occur. The utility of a population vector analysis depends on identifying discrete conditions or 'states' (such as tracking visual images, or degrees of rotation of a monkey forearm), which correspond to a unique set of mean firing rates across the ensemble for each state. As applied by Georgopoulos et al. (1986),

the mean firing of each neuron in an ensemble recorded in non-human primate motor cortex was computed in correlation to forearm movements. A set of population vectors was obtained which corresponded to specific angles of arm rotation, then successfully compared to new recordings to test the accuracy of the population vector which 'predicted' the forearm angle prior to movement.

3.2 Cross-correlations

Multiple neuron recordings can also be analyzed simply by looking for repeating patterns of firing across populations of recorded neurons. These patterns may consist of spatial combinations of two or more neurons within the ensemble, a temporal pattern within single neurons, a particular sequence of firing between several neurons, or combinations of spatial and temporal oscillations. Ensemble spike trains can therefore be scrutinized using cross-correlation techniques and the construction of cross-correlation histograms (CCHs; Gerstein and Perkel 1969) for this purpose. Detection of specific spatiotemporal firing patterns (i.e., correlations between neurons) may suggest ways in which those neurons may be connected, or may simply result from common 'driving' influences on several different cells. The discovery that hippocampal neurons 'replay' recent experiences during sleep (Skaggs and McNaughton 1996) relied on the observation that temporal offsets in CCHs between pairs of hippocampal place cells corresponded to the sequence in which those place cells were activated. During the waking state, the temporal offset reflected the sequential movement of an animal through different places in the environment. During the following sleep cycle, the hippocampal neural pairs were again active with the same temporal offsets, suggesting a memory of the exploring the environment. One drawback in cross-correlation studies, however, is that while it is appealing to use cross-correlations to map temporal connectivity and patterned firing between neurons within an ensemble, such procedures are not well suited to circumstances where nonstationary processes (i.e., behavior) are likely to alter temporal firing patterns while the data is being recorded (see Hampson and Deadwyler 1999). On the other hand, the demonstration that cross-correlations between hippocampal neurons were altered provided important confirmation of the perturbations in place cell firing and spatial navigation observed in transgenic mice with deficient NMDA-receptors on CA1 hippocampal neurons (McHugh et al. 1996).

3.3 Multivariate Analyses

More elaborate analyses of ensemble spike trains include multivariate statistical treatments such as linear discriminant analysis and principal components analysis (Nicolelis 1999). The differences and advantages of these analyses compared to those cited above are many. For instance, multivariate analyses simultaneously take account of the activity of all neurons in the ensemble, compared to cross-correlations which operate on pairs of neurons. Multivariate analyses incorporate both spatial (i.e., relationship between neurons within the ensemble) and temporal (time before and after events) information in the analyses, unlike population vectors which incorporate the firing of all neurons, but within an 'instantaneous' average of ensemble activity at discrete points in time. Finally, instead of producing one measure of correlation with behavioral and cognitive events for a given neuron, the multivariate components derived independent sources of variance within and between neurons comprising an ensemble.

Multivariate analyses provide the greatest potential for detecting neural patterns that correlate with perceptual, behavioral or cognitive event. Recordings of inferior temporal cortex (Miller et al. 1993), revealed many different patterns of encoding, including stimulus-specific and task (i.e., memory)-specific neural patterns. Likewise, analysis of ensembles of neurons from hippocampal (Deadwyler and Hampson 1997), parahippocampal regions (Young et al. 1997), and orbitofrontal cortex (Schoenbaum and Eichenbaum 1995) have demonstrated a broad spectrum of encoding that encompasses. Stimulus representation, memory storage, and formation of decision appropriate to the behavioral task.

4. Applications

Perhaps the best example of the utility of multiple neuron recording and analysis described above is the control of a robotic arm cited earlier (Chapin et al. 1999). In the study, ensembles of motor cortex neurons were recorded while a rat pressed a lever to cause a robotic arm to deliver water. A search for patterns of ensemble firing corresponding to the behavioral response would have required a massive search algorithm and intensive computation (as originally described Gerstein and Perkel 1969). However, principal components analysis of the ensemble spike trains revealed a pattern of neural firing that closely correlated to (and preceded) the limb movements required to press the lever and move the arm. Detection of the ensemble firing pattern while the animal performed the task allowed a computer to control the robotic arm, eventually allowing the rat to obtain water on the basis of the multiple neuron activity, without the requisite limb movement. Hence the principal components analysis allowed derivation of the brain function (behavioral response) from multiple neuron recording.

Single and multiple neuron recording have been applied to many purposes. Examination of the firing characteristics of single neurons provides a means of

correlating the activity of each neuron with the behavioral or cognitive activity of the animal. Multiple neuron recording is used simply to increase the yield of single neuron experiments, or for a more in-depth approach to correlating brain activity with external events. Analyses of the connectivity between neurons provides a framework for modeling neural circuits. Finally, extracting patterns of neural activity correlating to specific behavioral responses can result in a more detailed understanding of the role of specific brain areas in the process of cognition.

See also: Binding Problem, Neural Basis of; In Vitro Neurophysiology; Neural Synchrony as a Binding Mechanism; Neurons and Dendrites: Integration of Information

Bibliography

Chapin J K, Moxon K A, Markowitz R S, Nicolelis M A 1999 Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neuroscience* **2**: 664–70

Deadwyler S A, Hampson R E 1997 The significance of neural ensemble codes during behavior and cognition. In: Cowan W M, Shooter E M, Stevens C F, Thompson R F (eds.) *Annual Review of Neuroscience*. Annual Reviews Inc., Palo Alto, CA, Vol. 20, pp. 217–44

Georgopoulos A P, Schwartz A B, Kettner R E 1986 Neuronal population encoding of movement direction. *Science* **233**: 1416–19

Gerstein G L, Perkel D H 1969 Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science* **164**: 828–30

Hampson R E, Deadwyler S A 1999 Pitfalls and problems in the analysis of neuronal ensemble recordings during performance of a behavioral task. In: Nicolelis M (ed.) *Methods for Simultaneous Neuronal Ensemble Recordings*. Academic Press, New York, pp. 229–48

Hubel D H, Wiesel T N 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* **160**: 106–54

Laubach M, Wessberg J, Nicolelis M A 2000 Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task. *Nature* **405**: 567–71

McHugh T J, Blum K I, Tsien J Z, Tonegawa S, Wilson M A 1996 Impaired hippocampal representation of space in CA1-specific NMDAR1 knockout mice. *Cell* **87**: 1339–49

McNaughton B L, O'Keefe J, Barnes C A 1983 The stereotrode: A new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records. *Journal of Neuroscience Methods* **8**: 391–7

Miller E K, Li L, Desimone R 1993 Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *Journal of Neuroscience* **13**: 1460–78

Nicolelis M 1999 *Methods for Simultaneous Neuronal Ensemble Recordings*. Academic Press, New York

O'Keefe J, Dostrovsky J 1971 The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research* **34**: 171–5

Schoenbaum G, Eichenbaum H 1995 Information coding in the rodent prefrontal cortex. II. Ensemble activity in orbitofrontal cortex. *Journal of Neurophysiology* **74**: 751–62

Singer W 2000 Why use more than one electrode at a time? In: *New Technologies for the Life Sciences—A Trends Guide*. Elsevier, pp. 12–17

Singer W, Gray C M 1995 Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* **18**: 555–86

Skaggs W E, McNaughton B L 1996 Replay of neuronal firing sequences in rat hippocampus during sleep following spatial exposure. *Science* **271**: 1870–3

Von der Malsburg C 1994 The correlation theory of brain function. In: Domani E et al (ed.) *Models of Neural Networks II*. Springer, Berlin

Young B J, Otto T, Fox G D, Eichenbaum H 1997 Memory representation within the parahippocampal region. *Journal of Neuroscience* **17**: 5183–95

R. E. Hampson and S. A. Deadwyler

Perception: History of the Concept

'Perception' means an awareness of something, whether one's own thoughts and feelings, one's social surroundings, a business opportunity, the way to solve a math problem, or the current spatial layout. The term's meaning is often narrowed to basic aspects of perception with the senses. This article examines the history of the concept of sensory perception in philosophy and psychology. Perception has been studied as a source of knowledge and as a mental phenomenon in its own right. Visual perception was the first cognitive ability to be investigated with the methods of mathematical science, by the ancient Greeks, and was a central topic for the new philosophy and science of the seventeenth century. Quantitative studies of sensation and perception spurred the growth of experimental psychology in the nineteenth century. The twentieth century saw new perceptual theories and new techniques for examining the role of brain processes. Full understanding of how conscious awareness occurs in perception remains a goal for further research.

1. Perception in Ancient and Medieval Philosophy and Science

Human beings (as other animals) rely on their senses for survival. Although everyone uses the senses, it is not obvious how they work. Initial questions concerned their basic operation and their role in gaining knowledge. In vision, does something go out from the eye to meet the object, or does the object send

something to the eye? What do the senses reveal? Do they inform the observer of the very nature of things or are they fundamentally illusory? Or is their function primarily pragmatic?

Aristotle, working in fourth-century (BCE) Athens, wrote the first general treatise on psychology, described as *logon peri tes psyches* and entitled in Latin translation *De anima*, 'on the soul.' He considered the soul to be the principle of life or vitality, including reproduction and growth, sensory perception and purposeful activity, and intellectual theorizing. He offered a general theory of how the senses work, whereby something travels through a medium to each sense organ: light and color through air in the case of sight, vibrations in the case of sound, odors for smell, and so on. He also held that all knowledge is based on the senses. For 'proper sensibles' such as light and color, he postulated that a sample of the quality is received by the sensory soul 'without the matter' of the object. His medieval interpreters described this as the reception of a 'similitude' of the quality, or a 'real quality.' Aristotle held that sensory qualities alone do not reveal the essences of things, which are perceived by a higher cognitive faculty, the intellect, from observing changes over time. He differed from his teacher Plato, who held that in vision something goes from the eye into the air, and who taught that sensory perception is illusory, true knowledge arising from intellectual perception independent of the senses.

In this period 'optics' was about vision, not merely the physics of light (Lindberg 1976). In second-century Alexandria, Ptolemy developed a geometrical theory of vision, an extramission theory stating that something goes from eye to object. He modeled the relation between eye and object by a visual pyramid composed of a sheath of rays, with the object serving as base and the apex residing inside the cornea of the eye. The eye senses the direction and length of each ray, thereby perceiving the location of each point in the field of view, and so gaining full knowledge of the shapes, sizes, and places of things.

For centuries no theorist could formulate a mathematically rigorous intromission theory, with causation proceeding from object to eye. If one believes (correctly) that light is scattered from each point of an ordinary object in all directions, a problem arises that extramissionists did not face. It must be explained how the eye senses only one point for each point on the object, despite the fact that each point on the cornea receives light rays from all points in the field of view. Ibn al-Haytham (1030), working in Fatimid Cairo, produced the first accepted solution. He argued that the eye receives only the unrefracted rays (those that meet the cornea at right angles), and hence that it receives a cross-section of the visual pyramid preserving point-for-point the order of the scene. He also elaborated the theory that an object's true size is perceived through an unnoticed judgment that combines the visual angle under which an object is seen with the perceived distance to the object. This explanation guided psychological theories of size perception for nearly a millennium.

2. *Perception in the Modern Age*

The Scientific Revolution of the seventeenth century brought changes in perceptual theory. Kepler's discovery of the retinal image revealed the internal optics of the eye and so produced a new solution to Ibn al-Haytham's problem, though without altering the latter's geometrical analysis of size perception or his theory of it, which was further developed by Descartes and Berkeley (Hatfield and Epstein 1979). Descartes postulated physiological systems to produce distance perception directly through the ocular-muscle processes of accommodation and convergence, and Berkeley sought to replace the psychology of unnoticed judgments with associative processes. Aristotle's theory that samples of qualities are absorbed by the senses was challenged by the new physics, according to which qualities in objects must be reduced to particles having only size, shape, motion, and position (primary qualities). Secondary qualities in objects, such as color, were now to be equated with such particles and their effects on perceivers. Descartes argued that sensory perception does not reveal essences, which are perceived by the intellect independent of the senses. The empiricist Locke denied intellectual perception of that sort, and held that the new physical theories were hypotheses to be tested against sensory observation. Both Descartes and Locke considered experienced qualities such as color to be products of the human sensory system. Contrary to later interpreters (e.g., Burtt 1927), they did not regard them as illusory; if a subjectively produced color sensation is regularly or lawfully caused by qualities in objects, it can be used objectively to sort and classify such objects.

Eighteenth-century European theories of knowledge grappled with the new scientific image of the material world. Berkeley and Hume argued that the new theories devolve into paradox, as when 'red' is said both to be an experienced quality and a way light is reflected by surfaces. They sought to avoid skeptical conclusions by reducing knowledge of objects to the merely phenomenal. Reid (see Yolton 1984) countered with a 'common sense' affirmation of the reality and knowability of material objects (and of a deity, whose knowability was a primary motivation for him). Kant argued that analysis of the basis of actual knowledge in mathematics and natural science reveals that human knowledge cannot extend beyond sensory experience, so that metaphysical knowledge of God, an immaterial human soul, and the material world as it is in itself is not possible. Hegel countered that Hume, Kant, and others were wrong to think of the world in itself as lying beyond the phenomenally available world. The

natures of things are open to human knowledge, but can be grasped only through concepts that develop in a historical process of dialectic (see *Knowledge Representation*).

3. The Nineteenth Century

The nineteenth century was a golden age of sensory physiology and psychology in Germany (Turner 1994). G. Fechner, wanting to measure states of the soul, took psychophysical measurements for pressure on the skin, weights lifted by hand, line lengths perceived by sight, and the pitches of tones. Precise, empirical theories of color and spatial perception were developed by H. Helmholtz and E. Hering. In color theory, Helmholtz postulated three types of light-sensitive elements in the eye, with peak sensitivities at different wavelengths. Hering posited opponent-process mechanisms yielding four primary colors, red, green, yellow, and blue, along with a neutral white–black pairing. In spatial perception, Helmholtz argued that all ability for spatial localization is derived from experience, whereas Hering argued that the visual system has innate mechanisms for localizing sensations.

Nearly all major theorists accepted the assumption that perception is based on nonspatial, punctiform sensations varying only in quality and intensity. This assumption was shared by British associationists (e.g., J. Mill and J. S. Mill), nativists (H. Lotze), and empirists (W. Wundt, Helmholtz). It was questioned by Hering and C. Stumpf in Germany and W. James in the US. Dewey (1896) promulgated a functionalist approach to sensory perception, in which mind functions to adapt organisms to their environments.

4. Twentieth Century Trends

The idea that perception is built up from punctiform sensations was challenged by the Gestalt psychologists. Rather than punctiform sensations or two-dimensional perspective images, the Gestaltists regarded experience of the phenomenally given, meaningful world in three dimensions as the fundamental perceptual state. The American psychologist Gibson (1966) challenged the traditional theory that sensory stimulation provides greatly impoverished information about distant objects. He argued that, in vision, if one grants that organisms move and integrate information over brief intervals of time, then the information available at the eyes specifies fully the spatial layout of the scene. The British psychologist D. Marr developed computational models on the theory that Gibson-like assumptions about objects and stimulation are engineered into sensory systems. I. Rock at Rutgers University espoused the constructivist theory that perception results from inferential processes which combine impoverished sources of stimulation, an approach shared by many information processing and computer vision theorists. Some theorists preferred to view perception as sensory information processing rather than the production of conscious states. New techniques for neurophysiological recording in living animals and brain imaging in humans permitted study of neural activity in perception, though behavioral and phenomenal data guided this research. At the end of the twentieth century all these currents remained visible in perceptual theorizing (Palmer 1999).

Philosophical approaches early in the twentieth century drew inspiration from Hume's phenomenalism and James' 'neutral monism,' the latter attempting to avoid the mind–body problem by arguing that both the psychological and the physical are composed from neutral existents found in perception. This led to theories of 'sense data' that have (in vision) the phenomenal properties of two-dimensional perspective projections. Such theories were challenged by critical and representational realism, and their phenomenological claims faded as philosophers took account of the Gestaltists and Gibson (Crane 1992). Philosophical interest in perception shifted from the problem of knowledge (now allotted to the philosophies of the sciences and other cognitive practices) to perception as an instance of the mind–body problem. As the neural correlates of sensory perception became better known, there remained a theoretical gap between neural or informational accounts and the conscious states manifest in perceptual experience. It is not clear how changes in the electrochemical state of neurons in the back of the brain could *be* the experience of red, despite it being known which neurophysiological patterns yield an experience of red, and which yellow. Much has been learned about how the senses work, but the characteristic feature of perception, conscious awareness, remains unexplained.

See also: Aristotle (384–322 BC); Fechnerian Psychophysics; Helmholtz, Hermann Ludwig Ferdinand von (1821–94); Hume, David (1711–76); Kant, Immanuel (1724–1804); Locke, John (1632–1704); Mind–Body Dualism; Perception and Action; Perception: Philosophical Aspects; Perceptual Constancy: Direct versus Constructivist Theories; Psychology: Historical and Cultural Perspectives; Psychophysical Theory and Laws, History of; Psychophysics; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Burtt E A 1927 *Metaphysical Foundations of Modern Physical Science*. K. Paul, London

Crane T (ed.) 1992 *Contents of Experience*. Cambridge University Press, Cambridge, UK

Dewey J 1896 The reflex arc concept in psychology. *Psychological Review* **3**: 357–70 [In: Dewey J, *Early Works*, 1882–1896, Vol. 5. Southern Illinois University Press, Carbondale, IL]
Gibson J J 1966 *Senses Considered As Perceptual Systems*. Houghton Mifflin, Boston
Hatfield G C, Epstein W 1979 The sensory core and the medieval foundations of early modern perceptual theory. *Isis* **70**: 363–84
Ibn al-Haytham A 1030 *Kitāb al-Manāzir* [1989 *The Optics of Ibn al-Haytham*, Books I–III. Warburg Institute, London
Lindberg D C 1976 *Theories of Vision from al-Kindi to Kepler*. University of Chicago Press, Chicago
Palmer S E 1999 Vision science: Photons to phenomenology. MIT Press, Cambridge and London
Ptolemy 1989 *L'Optique de Claude Ptolémée, dans la version latine d'après l'arabe de l'émir Eugène de Sicile*. E. J. Brill, Leiden and New York [1996 *Ptolemy's Theory of Visual Perception: An English Translation of the Optics*. American Philosophical Society, Philadelphia]
Turner R S 1994 *In the Eye's Mind: Vision and the Helmholtz–Hering Controversy*. Princeton University Press, Princeton, NJ
Yolton J W 1984 *Perceptual Acquaintance: From Descartes to Reid*. University of Minnesota Press, Minneapolis, MN

G. Hatfield

Perception of Extrapersonal Space: Psychological and Neural Aspects

We take for granted our ability to know where things are and to navigate through the environment. The effortless way in which we can look at or reach for an object belies the complex computations that go into determining its location. The psychological and neural processes that underlie our spatial abilities have been extensively explored in humans and animals. In this article, selected psychological aspects of spatial perception will be reviewed, followed by a consideration of some cortical mechanisms of spatial perception and representation.

1. Psychological Aspects

1.1 Basic Processes

Spatial perception involves not one but many specific abilities. Within the visual domain these include locating points in space, determining the orientation of lines and objects, assessing location in depth, appreciating geometric relations between objects, and processing motion, including motion in depth. These spatial skills can be applied to imagined objects as well as to external stimuli, as in the classic experiments of Shepard and Cooper (1986) on mental rotation. These experiments showed that observers take longer to determine whether two objects are identical when the degree of mental rotation needed to align them increases. Similar sets of basic processes contribute to spatial representation in the auditory and somatosensory domains.

1.2 Space is Represented in Multiple Frameworks

Our subjective experience strongly suggests that we have direct access to a single coherent and overarching representation of space. Whether we localize a stimulus by sight, smell, hearing, or touch, we can respond to it with equal ease and with any motor system at our command. This introspection is misleading. There is no evidence for the existence of a single, explicit, topographic representation of space suitable for incorporating every kind of sensory input and generating every kind of motor output. On the contrary, the evidence points to multiple representations of space, in a variety of coordinate frames, and linked to separate output systems designed to guide specific motor effectors.

The particular representation of space in use at any time depends on the task the subject is trying to perform. For example, if you were to draw out a route for a hiking trip on a map, the route would be in the coordinates of the map, the piece of paper. If you were then going to walk along that route, you would have to begin by locating your current position within that coordinate frame and constructing a representation of the route with respect to your starting point. The first, map-based representation is an example of an allocentric representation, in which locations are represented in reference frames extrinsic to the observer. Allocentric representations include those centered on an object of interest (object-centered) and those in environmental (room-centered or world-centered) coordinates. The second representation, the one in the coordinates of the hiker's current position, is an example of an egocentric representation, in which locations are represented relative to the observer. Egocentric representations include those in eye-centered, head-centered, hand-centered, and body-centered coordinates. Experimental work in humans indicates that multiple reference frames can be activated simultaneously (Carlson-Radvansky and Irwin 1994).

2. Neural Aspects

2.1 Impairments of Spatial Perception in Humans

Much of our knowledge about the neural basis of spatial perception comes from observations in patients with spatial deficits following brain damage. These

include a wide range of perceptual and motor deficits, such as poor localization of visual, auditory, or tactile stimuli; inability to determine visual or tactile line orientation; impaired performance on mazes; impairment on tests of mental spatial transformations; right—left confusion; poor drawing; impaired eye movements to points in space; misreaching; defective locomotion in space; and amnesia for routes and locations. As can be seen from this partial list, spatial behavior involves many kinds of skills and, not surprisingly, a number of brain regions have been implicated in spatial perception and performance. Broadly speaking, the parietal lobe is responsible for spatial perception and representation of immediate extrapersonal space, while temporal and parahippocampal cortices are more involved in topographic memory and navigation (see *Navigation in Virtual Space: Psychological and Neural Aspects*). The frontal lobe receives input from both parietal and temporal cortex and is responsible for generating actions.

An important point about deficits in spatial perception is that they are far more common after right hemisphere damage than left. While patients with left hemisphere damage may also exhibit spatial deficits, it is clear that the right hemisphere in humans has a superordinate role in spatial processing and behavior (Heilman et al. 1985). Two kinds of spatial deficits following brain damage are particularly illuminating. First, a common sensorimotor deficit following parietal lobe damage is difficulty in using visual information to guide arm movements, referred to as optic ataxia. Patients with optic ataxia have difficulty with everyday tasks that require accurate reaching under visual guidance, such as using a knife and fork. They both misdirect the hand and misorient it with respect to the object, and are most impaired when using the contralesional hand to reach for an object in the contralesional half of space (see *Vision for Action: Neural Mechanisms*).

A second, classic disorder of spatial perception in humans is the tendency to ignore one half of space, called hemispatial neglect (see *Neglect, Neural Basis of*). The most common form of neglect arises from damage to the right parietal lobe and is manifested as a failure to detect objects in the left half of space. Neglect is more than just a visual deficit, however. It can occur separately or jointly across many sensory modalities (Barbieri and De Renzi 1989). Moreover, neglect occurs with respect to many different spatial reference frames. A patient with right parietal lobe damage typically neglects objects on the left but left may be defined with respect to the body, or the line of sight, or with respect to an attended object. Further, this neglect is dynamic, changing from moment to moment with changes in body posture and task demands (Behrmann 2000). Neglect is apparent even in the purely conceptual realm of internal images. Patients exhibit neglect when asked to imagine a familiar scene, such as a city square, and describe the buildings in it. The portion of space that is neglected changes when they are asked to imagine the scene from a different viewpoint (Bisiach and Luzzatti 1978). As this example illustrates, neglect can occur with respect to an internal image constructed by the individual. Patients with neglect show evidence of using multiple reference frames, just as intact individuals do.

2.2 Spatial Perception in Animals is Impaired by Parietal Cortex Lesions

The role of parietal cortex in spatial perception has been explicitly tested in animal studies. Monkeys with posterior parietal lobe lesions exhibit many of the same deficits seen in humans, including deficits in the appreciation of spatial relations between objects and impairments in eye movements and reaching. They perform normally on tests of object discrimination but are selectively impaired on a spatial task that requires them to judge which of two identical objects is closer to a visual landmark. In contrast, monkeys with temporal lobe lesions are unimpaired on the spatial task but fail to discriminate between or recognize objects (Ungerleider and Mishkin 1982). This double dissociation between the effects of lesions, in combination with the discovery of distinctive cortical inputs to the parietal and temporal lobes, led to the concept of the dorsal and ventral visual processing streams (see *Visual Systems: Dorsal and Ventral*).

Neurons in the dorsal stream encode the types of visual information necessary for spatial perception. Neurons in specific dorsal stream areas are selective for orientation, depth, direction and speed of motion, rotation, and many other stimulus qualities appropriate for perceiving spatial information. The dorsal visual processing stream leads to posterior parietal cortex, where many kinds of visual information converge, including information about stimulus shape. An equivalent set of somatosensory processing areas send tactile information to anterior parietal cortex. Visual and somatosensory signals converge on single neurons within the intraparietal sulcus, which divides anterior and posterior parietal cortex. Auditory signals have also been demonstrated to contribute to spatial processing in monkey parietal cortex.

2.3 Parietal Neurons Encode Spatial Information in Multiple Reference Frames

The standard approach for investigating the role of parietal neurons in spatial perception is to record electrical activity from individual neurons while the monkey performs a spatial task. Because brain tissue itself has no sensory receptors, fine microelectrodes can be introduced into the brain without disturbing the animal's performance. By recording neural responses during carefully designed tasks, neural activity

can be related directly to the sensory and representational processes that underlie spatial behavior. The general conclusion from these studies is that the function of parietal cortex is to transform spatial information from sensory coordinates into the motor coordinates that are necessary for the guidance of action (Stein 1992, Andersen et al. 1997, Colby and Goldberg 1999).

2.3.1 Head-centered spatial representation in the ventral intraparietal area (VIP). Area VIP is located in the floor of the intraparietal sulcus, where inputs from high-order visual and somatosensory cortex converge. In the visual domain, VIP neurons are characterized by direction and speed selectivity, and thus resemble neurons in other dorsal stream visual areas that process stimulus motion (see *Motion Perception: Psychological and Neural Aspects*). In the somatosensory domain, these same neurons respond to light touch on the head and face. The somatosensory and visual receptive fields of individual neurons exhibit strong spatial correspondence: they match in location, in size, and even in their preferred direction of motion. The existence of spatially matched receptive fields raises an interesting question: what happens when the eyes move away from primary position? If the visual receptive fields were simply retinotopic, they would have to move in space when the eyes do and so would no longer match the location of the somatosensory receptive field. Instead, for some VIP neurons, the visual receptive field moves to a new location on the retina when the eyes move away from the straightahead position. For example, a neuron that has a somatosensory receptive field near the mouth and responds best to a visual stimulus moving toward the mouth will continue to respond to that trajectory of motion regardless of where the monkey is looking (Colby et al. 1993). These neurons have head-centered receptive fields: they respond to stimulation of a certain portion of the skin surface and to the visual stimulus aligned with it, no matter what part of the retina is activated. Neurons in area VIP send projections to the region of premotor cortex (see Sect. 2.4.1) that generates head movements. Area VIP neurons thus contribute to the visual guidance of head movements and may play a special role in hand, eye, and mouth coordination. They operate in a head-centered reference frame in order to generate appropriate signals for a particular motor effector system, namely that which controls head movements.

2.3.2 Eye-centered spatial representation in the lateral intraparietal area (LIP). In contrast to area VIP, neurons in area LIP construct an eye-centered spatial representation. Individual neurons become active when a salient event occurs at the location of the receptive field. This can be a sensory event, such as the onset of a visual or auditory stimulus, or a motor event, such as a saccade towards the receptive field, or even a cognitive event, such as the expectation that a stimulus is about to appear, or the memory that one has recently appeared. The level of response reflects the degree to which attention has been allocated to the site of the receptive field (Colby et al. 1995).

Again, as we saw for area VIP, the animal's ability to make eye movements raises an interesting question about spatial representation. Neural representations of space are maintained over time, and the brain must solve the problem of how to update these representations when a receptor surface moves. With each eye movement, every object in our surroundings activates a new set of retinal neurons. Despite this constant change, we perceive the world as stable. Area LIP neurons contribute to this perceptual stability by using information about the metrics of the eye movement to update their spatial representation of salient locations (Duhamel et al. 1992). For example, LIP neurons are activated when the monkey makes an eye movement that brings a previously illuminated screen location into their receptive field. These neurons respond to the memory trace of the earlier stimulus: no stimulus is ever physically present in the receptive field, either before or after the saccade. The proposed explanation for this surprising finding is that the memory trace of the stimulus is updated at the time of the saccade. Before the saccade, while the monkey is looking straight ahead, the onset of the stimulus activates a set of LIP neurons whose receptive fields encompass the stimulated screen location, and they continue to respond after the stimulus is extinguished, maintaining a memory trace of the stimulus. At the time of the saccade, a corollary discharge, or copy of the eye movement command, containing information about the metrics of the saccade, arrives in parietal cortex. This corollary discharge causes the active LIP neurons to transmit their signals to the new set of LIP neurons whose receptive fields will encompass the stimulated screen location after the saccade. The representation of the stimulated location is thus dynamically updated from the coordinates of the initial eye position to the coordinates of the final eye position.

The significance of this observation is in what it reveals about spatial representation in area LIP. It demonstrates that the representation is dynamic and is always centered on the current position of the fovea. Instead of creating a spatial representation that is in purely retinotopic (sensory) coordinates, tied exclusively to the specific neurons initially activated by the stimulus, area LIP constructs a representation in eye-centered (motor) coordinates. The distinction is a subtle one but critical for the ability to generate accurate spatial behavior. By representing visual

information in eye-centered coordinates, area LIP neurons tell the monkey not just where the stimulus was on the retina when it first appeared but also where it would be now if it were still visible. The result is that the monkey always has accurate spatial information with which to program an eye movement toward a real or remembered target. The transformation from sensory to motor coordinates puts the visual information in its most immediately useful form. Compared to a head-centered or world-centered representation, an eye-centered representation has the significant advantage that it is already in the coordinates of the effector system that will be used to acquire the target. Humans with unilateral parietal lobe damage fail on an eye movement task that requires an eye-centered representation of a remembered target position. This failure presumably reflects an impairment of the updating mechanism in parietal cortex.

In sum, posterior parietal cortex plays a critical role in spatial perception and representation (Colby and Goldberg 1999). Physiological studies in monkeys show that parietal neurons represent spatial locations relative to multiple reference frames, including those centered on the head and the eye. Individual neurons (in area VIP) combine spatial information across different sensory modalities, and specific spatial reference frames are constructed by combining sensory and motor signals (in area LIP). In accord with the physiology, human neuropsychological studies show that neglect can be expressed with respect to several different reference frames.

2.4 Frontal Lobe Mechanisms of Spatial Representation

The parietal lobe transforms sensory representations of attended objects into the motor coordinate frames most appropriate for action. They do not actually generate those actions. Dorsal stream outputs to frontal cortex provide the sensory basis, in the correct spatial framework, for producing specific motor outputs. The following sections describe specific spatial representations found in three regions of frontal cortex.

2.4.1 Neurons in premotor cortex have head-centered and hand-centered visual receptive fields. Two different forms of spatially organized visual responsiveness have been described in ventral premotor cortex. First, in subdivisions representing facial movements, neurons respond to visual stimuli at specific locations relative to the head, much like those described above in area VIP. These neurons have been characterized by recording activity while objects approach the monkey's head along various trajectories. The preferred trajectory is constant with respect to the head, and is not affected by changes in eye position (Fogassi et al. 1996). This specific zone of premotor cortex receives inputs from area VIP, and uses the same form of spatial representation. In both cases, stimuli are represented in the motor coordinate frame that would be most useful for acquiring or avoiding stimuli near the face by means of a head movement.

A different kind of spatial representation has been observed in subdivisions of premotor cortex involved in generating arm movements. Here, neurons respond to visual stimuli presented in the vicinity of the hand. When the hand moves to a new location, the visual receptive field moves with it. Moreover, the visual receptive field remains fixed to the hand regardless of where the monkey is looking, suggesting the existence of a hand-centered representation (Graziano et al. 1994). A fascinating observation indicates that some neurons in this area are capable of even more abstract forms of representation. Rizzolatti and coworkers have described neurons in premotor cortex that are activated both when the monkey grasps an object, such as a raisin, and when the monkey observes the experimenter performing the same action (Gallese et al. 1996). These 'mirror' neurons encode not just the monkey's own motor actions but also the meaning of actions performed by others (see *Pre-motor Cortex*).

2.4.2 Object-centered spatial representation in the supplementary eye field (SEF). Actions are directed toward objects in the environment and toward specific locations on an object. Picking up your coffee cup requires that you locate both the cup in egocentric space and the handle in relation to the cup. The spatial reference frame that guides such movements is not limited to the egocentric representations described above. Evidence from the supplementary eye field (SEF) demonstrates that single neurons can encode movement direction relative to the object itself (Olson and Gettner 1995). The SEF is a division of premotor cortex with attentional and oculomotor functions. Neurons here fire before and during saccades. In monkeys trained to make eye movements to particular locations on an object, SEF neurons exhibit a unique form of spatial selectivity: they encode the direction of the impending eye movement as defined relative to an object-centered reference frame. For example, a given neuron may fire when the monkey looks toward the right end of a bar placed at any of several different locations on the screen, regardless of whether the eye movement itself is a rightward or leftward saccade. Moreover, the same neuron will fail to respond when the monkey makes a physically identical eye movement toward the left end of the bar stimulus. This striking result indicates that single neurons can make use of quite abstract spatial reference frames. Object-centered spatial infor-

mation could potentially guide arm movements as well as eye movements. Moreover, neuropsychological evidence indicates that an object-centered reference frame can be used to direct attention: some patients exhibit object-centered neglect after parietal lobe damage (Behrmann 2000). Parietal lobe lesions in monkeys likewise produce impairments on tasks that require an object-centered spatial representation.

In sum, behavioral and neuropsychological studies indicate that we make use of multiple spatial representations in the perception of extrapersonal space. Neurophysiological studies are beginning to uncover the neural mechanisms underlying the construction of these egocentric and allocentric spatial representations.

See also: Hippocampus and Related Structures; Motion Perception Models; Motion Perception: Psychological and Neural Aspects; Navigation in Spatial Environments, Neural Basis of; Navigation in Virtual Space: Psychological and Neural Aspects; Spatial Cognition; Wayfinding and Navigation Behavior

Bibliography

Andersen R A, Snyder L H, Bradley D C, Xing J 1997 Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* **20**: 303–30

Barbieri C, De Renzi E 1989 Patterns of neglect dissociation. *Behavioural Neurology* **2**: 13–24

Behrmann M 2000 Spatial reference frames and hemispatial neglect. In: Gazzaniga M (ed.) *The Cognitive Neurosciences*. MIT Press, Cambridge, MA

Bisiach E, Luzzatti C 1978 Unilateral neglect of representational space. *Cortex* **14**: 129–33

Carlson-Radvansky L A, Irwin D E 1994 Reference frame activation during spatial term assignment. *Journal of Memory and Language* **37**: 411–37

Colby C L, Duhamel J R, Goldberg M E 1993 Ventral intraparietal area of the macaque: Anatomic location and visual response properties. *Journal of Neurophysiology* **69**: 902–14

Colby C L, Duhamel J R, Goldberg M E 1995 Oculocentric spatial representation in parietal cortex. *Cerebral Cortex* **5**: 470–81

Colby C L, Goldberg M E 1999 Space and attention in parietal cortex. *Annual Review of Neuroscience* **22**: 319–49

Duhamel J R, Colby C L, Goldberg M E 1992 The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**: 90–2

Fogassi L, Gallese V, Fadiga L, Luppino G, Matelli M, Rizzolatti G 1996 Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology* **76**: 141–57

Gallese V, Fadiga L, Fogassi L, Rizzolatti G 1996 Action recognition in the premotor cortex. *Brain* **119**: 593–609

Graziano M S A, Yap G S, Gross C G 1994 Coding of visual space by premotor neurons. *Science* **266**: 1054–6

Heilman K M, Watson R T, Valenstein E 1985 Neglect and related disorders. In: Heilman K M, Valenstein E (eds.) *Clinical Neuropsychology*. Oxford University Press, Oxford, UK

Olson C R, Gettner S N 1995 Object-centered direction selectivity in the macaque supplementary eye field. *Science* **269**: 985–8

Shepard R, Cooper L 1986 *Mental Images and their Transformations*. MIT Press, Cambridge, MA

Stein J F 1992 The representation of egocentric space in the posterior parietal cortex. *Behavioral and Brain Sciences* **15**: 691–700

Ungerleider L, Mishkin M 1982 Two cortical visual systems. In: Ingle D J, Fouckle M A, Mansfield M A (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA

C. L. Colby

Perception: Philosophical Aspects

Perception has been of perennial interest in philosophy because it is the basis of our knowledge of the external world. Because perception is a transaction between the world and the perceiver or their mind, investigations of the nature of the world perceived, the perceiver or their mind, or the relation between the two are possible. The philosophy of perception has focused on the latter two issues; although the study of perception has led to important views about the nature of the world perceived—particularly the idealism of Bishop Berkeley (see below)—such views are rare. Philosophers have asked instead: What is the nature of the relation between the mind of the perceiver and the world perceived? and What is the nature of the perceptual states of that mind? Contemporary philosophy of perception continues to debate these traditional questions, although it places more emphasis on the nature of the perceiving mind. It also tends to focus on visual perception rather than on the other senses, assuming—controversially—that what is true of vision is likely to be true of other modes of perception. Philosophers have also investigated the justification of perceptual belief, but because this is a branch of traditional epistemology, it is not dealt with here. This article considers some of the specific questions that have been raised in the investigation of the two general questions above as well as some of the answers given to them.

1. The Perceiver–World Relation

1.1 Direct Realism, Indirect Realism, and Phenomenalism

As I look out of the window, I see objects, such as buildings, and I see their properties, such as the green

color of their roofs. Presumably, the objects and their properties continue to exist even when I am not looking at them. (I can also be said to see events, such as the passing of pedestrians along the sidewalk, and facts, such as that there is traffic on the street, but we will ignore these matters here.) It seems to me that my acquaintance with the objects I perceive is an intimate one. I have no sense that anything stands between me and the buildings I am looking at. I do not feel that I am looking at a representation of the buildings as I might if I were looking at a photograph of them, nor that I am receiving information about them that has to be decoded. I merely open my eyes, and there they are.

This common sense view of visual perception is called *direct realism*. Direct realism holds that the objects one perceives are the physical objects of the external world and that these objects are perceived directly, that is, without the aid of any intermediary object or representation. A question arises about the status of the properties of objects when they are unperceived. *Naïve direct realism* holds that all of an object's perceived properties persist when it is unperceived; *scientific direct realism* holds that only the properties countenanced by science—such as shape—persist when an object is unperceived (see Dancy 1981, Martin 1998). Object color, for example, is thought by some to be a mind-dependent property and does not, on this view, persist when an object is unperceived. Scientific direct realism holds, therefore, that color is not perceived directly.

Despite its apparent obviousness, philosophers have challenged direct realism. An important objection to the view is expressed by the *argument from illusion*. Suppose I am having a hallucination of Beaver Lake. What is the object of my experience? Since I am not perceiving a physical object, I must be experiencing some other kind of object, perhaps a mental object. Further, since the experience of hallucinating Beaver Lake is, we may suppose, identical to the experience of actually perceiving Beaver Lake, the object in each case must be identical. In all true perception, therefore, there must be some (mental) object that acts as an intermediary or representation in the perception of the external world. The view that one is directly aware of intermediaries has been expressed as the claim that there is a *given* in perception—a nonconceptual state that forms the basis for full-blooded perception which requires the possession of concepts by the perceiver.

Various strategies have been adopted to resist the argument from illusion. First, the *adverbial theory of perception* denies the inference from the fact of hallucination to the claim that there is a mental entity that is the object of that hallucination. Instead, the theory holds that when I hallucinate a red triangle, for example, I am 'experiencing redly and triangularly', where 'redly' and 'triangularly' qualify the manner of experiencing, and does not presuppose the existence of a mental object. Because no inner mental object has to be posited even in cases of hallucination, it is possible to reclaim the common sense view that in true perception it is physical objects that are directly perceived.

A second strategy is to deny the claim that the experience of true perception is indistinguishable from that of hallucination. The *disjunctive conception of perceptual experience* (see Snowdon 1990) holds that these experiences are constituted by different mental states. Thus, even if hallucinatory experience involves a mental representation, true perception need not, and could be direct.

A third strategy is to treat perception as a special kind of belief (or inclination to believe) rather than a sensory state (Armstrong 1961). On this view, to see a red triangle is to believe that there is a red triangle before one's eyes. A hallucination of a red triangle is no more than a false belief and does not, therefore, call for the introduction of a mental representation any more than the false belief that the local hockey team will win the Stanley Cup requires a mental object to be the thing believed. Some support for this view derives from the phenomenon of *blindsight* (Weiskrantz 1986), a state in which subjects with particular forms of brain damage are able to make crude perceptual judgements about the properties of stimuli in the absence of any conscious experience. If one is willing to accept the claim that blindsighted subjects are perceiving, then blindsight undermines the common sense view that all perception involves conscious experience. Since the perception-as-belief view does not make conscious experience central to perception, it is thus strengthened.

A fourth strategy for resisting the argument from illusion has been developed by information-theoretic approaches to perception which explicate the transaction between the world and the perceiver as an exchange of information (see Dretske 1981). For example, the *computational approach* to vision (Marr 1982) investigates how the information about properties of the environment contained in patterns of retinal stimulation can be extracted by the visual system.

Whatever the success of these strategies, direct realism also has trouble accounting for a different problem—that of perceptual error. If one's experience of the external world is mediated by a mental representation, then error can be explained as a mismatch between the representation and what it represents. But if there is no such representation but only a direct perception of the physical objects themselves, how is misperception possible?

These arguments support *indirect* or *representative realism* according to which external physical objects are perceived, but only indirectly, by means of the direct perception of intermediary entities (see Dancy 1981, Jackson 1977). These intermediaries are usually thought to be mental representations and have traditionally been called *ideas*, *appearances*, *sensibilia*, and *sensa*, among others. In the twentieth century they

have been referred to, rather notoriously, as *sense-data*. Much of the investigation into the nature of the relation between the perceiver and the world has focussed on two distinct questions: first, whether perception is in fact mediated in this way; and, second, whether one is consciously aware of the intermediaries. This investigation has been made more difficult by the fact that these quite different questions have often been conflated (Crane 1992).

Indirect realism also faces important objections, however; in particular, that it seems to lead to scepticism about the external world. If I am aware only of mental representations, then for all I know there is no external world at all. One response to this objection is to claim that it misconstrues the claims of indirect realism. The theory does not say that we have no access to external objects but rather that our access is by way of mental intermediaries. A second response embraces the sceptical possibility. Bishop Berkeley (1685–1753) famously argued that there is no external world but only perceptions, or, at any rate, that the external world is no more than collections of actual or possible perceptions, a view known as *phenomenalism* or *idealism*. Despite the importance of Berkeley's arguments for the theory of perception, however, the radical nature of his claims about the world have prevented them from being widely held.

Most philosophers of the first half of the twentieth century were indirect realists of the sense-datum variety. The early part of the second half of the twentieth century saw a complete rejection of sense-datum theory and indirect realism, and many contemporary philosophers are direct realists. Nevertheless, the tendency to think that there is some aspect of experience that involves a perceptual intermediary or that there are features of perceptual experience that cannot be construed as the direct perception of the external world has proved very strong. Some aspects of sense-datum theory have emerged again in the debate about qualia (see below), and even outright sense-datum theories are no longer anathema.

1.2 The Causal Theory

The *causal theory of perception* takes up a second important issue in the study of the relation between perceiver and world. The causal theory holds that the transaction between the perceiver and the world should be analyzed primarily in terms of the causal relation underlying that transaction (Grice 1961). One version of the causal theory claims that a perceiver sees an object only if the object is a cause of the perceiver's seeing it. Traditionally, indirect realism was associated with the causal theory, but not all recent versions of the causal theory are indirect theories (see Vision 1997). Indeed, the causal principle seems relevant to any view of perception, whether direct or indirect. This is illustrated by the case of a desert traveler who hallucinates an oasis. His eyes are closed, so nothing outside of his own mind is causing the experience. However, it happens that an oasis is in fact in front of him, and, coincidentally, the hallucination matches the oasis perfectly. Despite the match between the experience and the environment, however, the absence of a causal relation moves us to deny that the case is one of perception (Pears 1976).

It seems clear, nevertheless, that an object's being a cause of a perceptual state is not sufficient for perception. For example, the state of perceiving an object typically will be the effect of a long causal chain, but typically only one of the links in the chain will be the object perceived. Suppose that a cup is illuminated by a flashlight. The flashlight is a cause of the perceptual state just as the cup is, but it is not the object perceived, and it is hard to see how appeal to causation alone will capture this obvious fact (Pears 1976).

1.3 Color Perception and the Primary–Secondary Quality Distinction

A different question about the relation between the perceiver and the world is the question of how successful the process of perception is at informing the perceiver about her environment. This question has been explored primarily in the debate about color perception, the most actively investigated topic in the area since the mid-1980s. Although philosophical interest in the perception of color is as old as philosophy itself, the contemporary debate has been invigorated by significant developments in color science and the introduction of these developments into philosophy (see Byrne and Hilbert 1997).

The perception of color is related to the success of visual perception because there is some reason to doubt whether objects actually possess colors. If it could be shown that objects do in fact fail to have colors, then visual perception would be seen to be mistaken in its representation of objects as color-bearers. It is generally assumed that the only way in which colors could be properties of external objects would be if they were physical properties—properties countenanced by science. The epistemological question whether color vision is reliable, therefore, is often recast as the metaphysical question whether there is a scientific property that is sufficiently close to our conception of color to deserve the name. If there is no such property, then doubt is cast on the idea that color is indeed an ordinary property of objects, and this in turn raises doubt about the accuracy of vision quite generally.

Two facts suggest that colors are not physical. First, color does not appear in the basic scientific description of the world. Second, colors seem to be variable in a way that ordinary physical properties are not. For example, a yellow object against a red background will appear greenish, but the same object against a green

background will look reddish, a phenomenon known as *simultaneous contrast*. Because of these facts, some philosophers have concluded that color is a *secondary quality*, that is, a property the existence of which is dependent on a perceiving mind. Secondary qualities are contrasted with *primary qualities*, such as shape or mass the existence of which is not dependent on the perceiving mind.

The philosophical investigation of color has led to two opposing camps. *Color realists* accept as an obvious fact that colors are properties of external objects. The task is to identify the physical property that is identical to color and show that the variability of color, such as that due to simultaneous contrast, does not undermine the idea that colors are properties of external objects. *Color antirealists* take the variability of color to be central to its nature and infer that color, as we understand that concept, does not exist. They hold that while some of the physical properties of the environment cause viewers to have color experience, there is nothing in the environment that is correctly represented by those experiences.

A special sort of realism attempts to accommodate both camps. The *dispositional theory of color* (usually associated with John Locke (1632–1704)) acknowledges both the variability of color and its external reality. An object is navy blue, on this view, if it has the disposition—a real property of the object itself—to cause normal perceivers to have experiences of navy blue in standard conditions. Colors are identical to these dispositions and, therefore, are properties of external objects. Color is variable because objects have at least as many color dispositions, and therefore colors, as there are color experiences.

2. The Perceiving Mind

2.1 Qualia

The debates concerning the nature of the mental states involved in perception recapitulate some of those dealing with the mind-world relation. Closely related to the color debate is the question about the existence of *qualia* (singular: *quale*, pronounced kwä′lē). The term is sometimes used to refer to the properties of the external environment, such as object color, that are experienced in perception. It is also used in more philosophically radical ways to refer to the properties of experiences themselves that give experiences their 'felt' qualities. This radical understanding of qualia thus takes them to be mental entities. Think, for example, about orgasms. There is a distinct feel to the experience of an orgasm that is absent in the mere belief that one is having an orgasm. After all, others can also believe that you are experiencing an orgasm without experiencing its characteristic feel. There is, as it is sometimes put, something it is like having certain kinds of experience. According to some, this is because these kinds of experience possess qualia and beliefs do not.

Because qualia in the radical sense are mental items, the qualia debate has focused in part on whether the existence of qualia shows that *physicalism*—the view that the only entities, properties, and relations that exist are those countenanced by the physical sciences—is inadequate. (Philosophers who believe that it does have come to be called qualia freaks.) A central argument here is the knowledge argument (Jackson 1982): Mary is born and raised in an entirely black and white room. She becomes a scientist and, we suppose, learns everything a complete science can teach about perception and the world perceived. She then emerges from her room and, seeing a tomato for the first time, learns something new, namely, what it is like to see something red; that is, she experiences a red quale. But if Mary knows everything that a complete science can teach about the world and the perception of it, then she has learnt something that lies outside of the realm of the scientific, and physicalism is false.

The existence of qualia has also been taken to show that the dominant theory of mental states, called *functionalism*, is false. Functionalism holds that mental states are to be analyzed solely in terms of their relations to other mental states, relevant sensory input, and the behavioral output to which they lead. How these states are embodied in an animal, or indeed in a computer, is taken to be irrelevant to the nature of the states. Being in pain, for example, is to be analyzed as the state that is typically caused by damage to one's body; that causes one to withdraw from the source of the damage; that is related to feelings of anger or regret; and so on. But now consider which is made possible by the existence of qualia *spectrum inversion*. Inverted perceivers have systematically different color experience from normal perceivers because the qualia associated with their perceptual experience is inverted: what normal perceivers call 'red' they see as the color normal perceivers call 'green'; what normal perceivers call 'yellow' they see as the color normal perceivers call 'blue'; and so on. Despite their differences from normal perceivers, however, spectrally inverted perceivers could never be identified. After all, inverted perceivers would learn the same language as normal perceivers: they would call grass 'green' and tomatoes 'red' even if their experience of these objects was different from that of normal perceivers. If there were inverted perceivers, therefore, nothing in their behavior would reveal that they were inverted. In particular, nothing about the functional relations among their mental states revealed in their behavior would distinguish inverted from normal perceivers. By hypothesis, however, their experience is different. This argument—the *argument from spectrum inversion*—seems to show, therefore, that there has to be more to perceptual states than their functional relations.

It has been argued, however, that the very idea of qualia as properties of experience is based on a

confusion between the properties experience represents the environment as having and the properties experience itself possesses (Harman 1990). When one introspects to examine these putative properties of experience, one finds only the properties of the objects one is experiencing and not the properties of experiences. However, the former are not mysterious in the least; they are just ordinary properties of the physical world.

2.2 Perceptual Content

The qualia debate has also provided a way of investigating the nature of the *representational content* of perceptual states (see Villanueva 1996). Philosophers have asked whether there is something more to perceptual experience than what is represented by that experience—its representational content (or, simply, 'content'). The view that denies that there *is* anything more is sometimes called the *intentional theory of perception* (Harman 1990) (where 'intentional' refers to what a mental state is about). On the intentional view, once the objects and properties in the scene perceived are specified, the content of the perceptual state has been exhausted. The qualia freak denies the intentional theory. He believes that there is more to perceptual states than their content, namely, the felt qualities of perceptual experiences. Consider orgasms again. Whatever the representational content of that state—perhaps that a particular physiological process is occurring in one's body—the felt quality of the experience does not seem to be exhausted by that content.

In the realm of visual experience, intuitions are less clear. Suppose I am standing on a road on which there are two trees of the same height, one of which is at a distance of 100 meters and the second of which is at a distance of 200 meters. While I might judge the two trees to be of the same height, it is also apparent to me that the nearer tree takes up more space in my visual field. Since experience cannot correctly represent the trees both as equal in size and as different in size, it has been suggested that there is an entity, the visual field itself, that is the bearer of the latter aspect of the experience: the trees are represented as being the same size but appear in the visual field to have different sizes. There are thus two orders of properties: those of the visual field and their sister properties in the external world. On this view, the properties of the visual field are not part of the representational content of the perceptual state (Peacocke 1983).

A related question about perceptual content is whether it is conceptual or nonconceptual. The mainstream position is that all perceptual content is conceptual; the perceiver must possess the concepts for the entities and properties perceptually represented by her. For example, the ability to see a dog as a dog requires that one possess the concept 'dog.' This mainstream position evolved in part as a rejection of sense-datum theory and its commitment to a nonconceptual given in perception. Nevertheless, some of the considerations that led to sense-datum theory have emerged again as relevant, and it has been argued that some perceptual representation does not require the possession of concepts. One such consideration is the epistemological view that conceptual thought has to be grounded in something that is nonconceptual (see Sedivy 1996). A second consideration is that perceptual states seem to represent the environment in a much more fine-grained fashion than a perceiver has the concepts for. For example, it is sometimes thought that the colors we see are more varied and subtly different from each other than we have concepts for. A third consideration is the desire to link the conceptual nature of thought with the contents of the computational states posited by the computational theory of vision. The computational theory of vision describes some of the content of a visual state in terms of various mathematical concepts, but perceivers do not have to possess those concepts in order to see.

One way of fleshing out nonconceptual content is in informational terms. A second way of doing so is by means of *scenarios*. A scenario is a representation of the way the physical space around the perceiver is filled out. It is specified by identifying within a perceiver-relative frame of reference the locations of object surfaces and their visible properties. The scenario itself, rather than any mental representation, is part of the content of the perceptual state, and it captures how the world looks to the perceiver independently of the perceiver's conceptual resources (Peacocke 1992).

A third question about perceptual content has focussed on the computational theory of vision, in particular as it has been presented by David Marr (1945–80). Theorists have asked whether the contents of computational states have to make essential reference to the external world or whether they can be fully specified in terms of the perceiver's states alone. The former view is one (but not the only) version of *externalism*; the latter is a version of *internalism* or *individualism*. (In philosophical jargon, externalism is committed to *wide content* whereas individualism is committed only to *narrow content*.) Externalism and internalism are positions that have been important in recent philosophy of mind, and the debate about computational vision has been pursued largely as an instance of the wider debate about what cognitive theory says about mental content quite generally (Burge 1996).

2.3 Spatial Perception

The scenarios mentioned above in connection with perceptual content are essentially spatial and reflect a concern in the philosophy of perception about space.

The importance of this topic has been recognized since Immanuel Kant's (1724–1804) *Critique of Pure Reason* tied spatial perception to the concept of an objective world. One reason for thinking that spatial perception is necessary for objectivity is that the perception of space seems essential to the formation of a concept of something that persists over time. The notion of persistence, in turn, is necessary to form a concept of an external object and, thereby, an objective world. Imagine a space-less world constituted entirely of sound. In such a world it seems impossible to articulate the distinction between, for example, two different C# sounds occurring at different times and the very same occurrence of C# being heard, disappearing, and then reappearing. It is argued, therefore, that without the perception of space, it is impossible to distinguish between the occurrence of multiple fleeting objects of the same type and a single persistent object (Strawson 1959). And without a notion of persistence, it is impossible to develop a concept of an objective world. A closely related idea is that the perception of space—or possibly the ability to attend selectively to different parts of space—is required in order to entertain thoughts about particular physical objects.

A second issue regarding space perception is that of the concept of a surface and its role in ordinary language and theories of perception. Under typical conditions, I only see part of the surface of an object. Does this mean that I see only part of an object? Or do I see a whole object but only part of it directly? And does that imply that I see the whole object indirectly? Investigation of the nature of surfaces thus leads to a distinction between direct and indirect perception that is rather different from the traditional one and does not depend on a distinction between the physical and the mental (Stroll 1988). The philosophical investigation of surfaces is important also because the computational theory of vision is in large part about how visual perception computes the surface properties of objects.

A third issue has come to be known as *Molyneux's question* or *Molyneux's problem* (see Degenaar 1996). It was formulated by the Irish scientist William Molyneux (1656–98) and discussed by John Locke with whom Molyneux corresponded. In a letter to Locke on July 7, 1688, Molyneux formulates the question: 'A Man, being born blind, and having a Globe and a Cube, nigh of the same bigness, committed into his Hands, and being Told, which is Called the Globe, and which the Cube, so as easily to distinguish them by his Touch or Feeling; Then both being taken from Him, and Laid on a Table, let us suppose his Sight Restored to Him; Whether he Could, by his sight, and before he touch them, know which is the Globe and which the Cube?' This question was addressed by almost every major theorist of perception following Molyneux and has been of sufficient historical interest to have led Ernst Cassirer to claim that it was the focus of eighteenth century epistemology and psychology as well as the stimulus for all of Berkeley's philosophy. Molyneux's question continues to receive some attention from philosophers as a way of exploring a number of issues including the relation between different perceptual modalities, especially sight and touch; the existence and nature of the concept of space; and whether or not certain perceptual abilities are learned or innate.

Finally, the question of whether some quasiperceptual representations have significant spatial properties has been examined in the debate about mental imagery, the quasivisual experiences of imagination (see Block 1981). Philosophers and psychologists have debated the claim that mental images are essentially spatial as against the view that they are no more than the language-like representations thought by some to constitute the mental representations underlying all of cognition. Because there is now evidence that mental images are produced by the same neural machinery that produces visual experience, further investigation into imagery may shed light on the nature of visual representations, and research into vision may produce a better understanding of imagery.

See also: Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Knowledge (Explicit and Implicit): Philosophical Aspects; Memory Psychophysics; Perception and Action; Perception: History of the Concept; Perception without Awareness, Psychology of; Perceptual Learning; Perceptual Organization; Reference and Representation: Philosophical Aspects; Vision, High-level Theory of; Vision, Low-level Theory of

Bibliography

Armstrong D M 1961 *Perception and the Physical World*. Routledge & Kegan Paul, London

Block N (ed.) 1981 *Imagery*. MIT Press, Cambridge, MA

Block N J, Fodor J A 1972 What mental states are not. *Philosophical Review* **81**: 159–81

Burge T 1986 Individualism and psychology. *Philosophical Review* **95**: 3–45

Byrne A, Hilbert D R (eds.) 1997 *Readings on Color*. MIT Press, Cambridge, MA, Vol. 1

Crane T 1992 Introduction. In: Crane T (ed.) *The Contents of Experiences: Essays on Perception*. Cambridge University Press, Cambridge, UK

Dancy J 1985 *An Introduction to Contemporary Epistemology*. Blackwell, Oxford, UK

Degenaar M 1996 *Molyneux's Problem* [trans. Collins M J]. Kluwer, Dordrecht, Holland

Dretske F I 1981 *Knowledge and the Flow of Information*. MIT Press, Cambridge, MA

Grice H P 1961 The causal theory of perception. *Proceedings of the Aristotelian Society* **35**: 121–52

Harman G 1990 The intrinsic quality of experience. In: Tomberlin J E (ed.) *Philosophical Perspectives 4*. Ridgeview, Atascadero, USA
Jackson F 1977 *Perception*. Cambridge University Press, Cambridge, UK
Jackson F 1982 Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–36
Marr D 1982 *Vision*. Freeman, New York
Martin M G F 1998 Perception. In: Craig E (ed.) *Routledge Encyclopedia of Philosophy*. Routledge, London
Peacocke C 1983 *Sense and Content*. Clarendon Press, Oxford, UK
Peacocke C 1992 *A Study of Concepts*. MIT Press, Cambridge, MA
Pears D 1976 The causal conditions of perception. *Synthese* **33**: 25–40
Sedivy S 1996 Must conceptually informed perceptual experience involve non-conceptual content? *Canadian Journal of Philosophy* **26**: 413–31
Snowdon P 1990 The objects of perceptual experience. *Proceedings of the Aristotelian Society* **64**: 121–150
Strawson P F 1959 *Individuals*. Methuen, London
Stroll A 1988 *Surfaces*. University of Minnesota, Minneapolis, MN
Villanueva E (ed.) 1996 *Perception*. Ridgeview, Atascadero, Holland
Vision G 1997 *Problems of Vision*. Oxford University Press, Oxford, UK
Weiskrantz L 1986 *Blindsight: A Case Study and Implications*. Clarendon Press, Oxford, UK

I. Gold

Perception without Awareness, Psychology of

Perception without awareness (or, synonymously, unconscious perception) denotes the fact that a stimulus exerts causal influence on mental processes or behavior without being phenomenally experienced.

Research into unconscious perception began early in the history of experimental psychology (Peirce and Jastrow 1884, Sidis 1898). It is related to other domains working on unconscious processes, such as implicit memory (see *Implicit Memory, Cognitive Psychology of*) or implicit learning (Reber 1993). From the very beginning, experimenters attempted to prove the mere existence of unconscious perception. Researchers had to set up experimental demonstrations of a stimulus's causal influence on mental processes without its concomitant phenomenal experience. More recently, theorists began to ask how perception without awareness works.

The proof of perception without awareness usually needs two critical tests: (a) an indirect test that assesses the impact of a stimulus on mental processes or behavior, and (b) a direct test that establishes the absence of phenomenal experience of the stimulus in question. There was rarely doubt about the demonstration of causal influence of a stimulus, but there was—and still is—much skepticism about the proof of absence of phenomenal experience. Therefore, experimenters aimed at enhancing the validity of thresholds in order to ensure that perception was indeed subliminal. Alternatively, some researchers were looking for procedures that circumvented the problem of thresholds. Finally, there is now some evidence for unconscious perception from research in neuroscience.

1. Early Studies

Early studies attempted to show that people were able to guess at above chance level without subjectively perceiving the stimulus (see Adams 1957, for a review). Much of the classical research has been criticized for its lack of proper control conditions and of proper measurements of consciousness (e.g., Eriksen 1960, Holender 1986). Holender, for example, criticized experiments using dichotic listening tasks for their lack of evidence that processing of the stimuli was indeed unconscious. In dichotic listening tasks, participants hear different information in both ears. They are instructed to attend to the information in one ear. It has been shown that information in the unattended ear may exert some influence on ongoing processing without subjective awareness of this information. Of course, if participants turned their attention to the unattended ear, they would be able to identify the content. Holender concluded that one can not be sure that unattended stimuli are in fact unconscious. Therefore, he proposed that manipulations should not limit mental resources by distracting attention, but should limit the available data by rendering perception of stimuli impossible.

2. Perception Below Threshold

At the start of the twenty-first century, several masking techniques are used to render the material—most of them visual stimuli—undetectable or unidentifiable. For example, to measure individual thresholds of detection, masked stimuli are presented on half of the trials. On the other half of the trials, only the mask is shown. Participants have to decide whether a stimulus is present or absent. If accuracy does not exceed chance level (mostly at 50 percent), absence of direct phenomenal experience is assumed.

In a well-known study, Marcel (1983) presented masked words below the subjective threshold of detection. Then he presented two words, one similar and one dissimilar to the masked word, in respect to either graphemic information or semantic content. Participants had to decide which word was similar to the masked word, either graphemically or semantically. They were able to make above-chance

guesses for each kind of judgment. As presentation time was decreased, judgments about graphemic similarity dropped to chance level first, followed by judgments about semantic similarity.

The logic underlying this kind of study is to show a significant effect in an indirect perceptual task (e.g., semantic discrimination) when there is a null effect in a direct task (e.g., detection). However, null sensitivity in direct tasks has always been a matter of controversy. If the criteria for null sensitivity in the direct test are too lenient, one may erroneously claim to have demonstrated unconscious perception. Conversely, if these criteria are too strong, it may be virtually impossible to demonstrate perception without awareness.

Therefore, some scholars have tried to circumvent the issue of thresholds and their measurements and taken different routes to demonstrate unconscious perception.

3. *Alternative Approaches to Demonstrate Unconscious Perception*

Reingold and Merikle (1988) designed an experimental setup that allows measuring unconscious perception without assessment of thresholds. Specifically, they proposed to use identical materials and experimental conditions for direct and indirect tests. The only difference between the two tests lies in the instruction: Conscious perception of stimuli is assessed in a direct test. In contrast, a task without reference to phenomenal experience of the stimulus is given in an indirect test. Normally, one would expect that performance in the direct test would show greater sensitivity than performance in the indirect task because conscious perception optimizes extraction and use of stimulus information. If the indirect measure shows greater sensitivity than the direct measure, however, one may conclude that unconscious processes mediate task performance. This was the case in a study on implicit memory effects by Kunst-Wilson and Zajonc (1980). Participants were exposed to random shapes for 1 ms during a learning session. At test, participants were shown 10 pairs of octagons, one old and one new. They had to decide either which of the two octagons has been shown before (direct test) or which of the octagons they liked better (indirect test). In the direct test, participants scored 48 percent, which was at chance level, but they preferred the old octagon in 60 percent of the cases, indicating greater sensitivity of the indirect than of the direct test. In studies on unconscious perception, however, it proved difficult to find the proposed indirect-greater-than-direct effect pattern.

More recently, Greenwald and his colleagues proposed to use performance in a direct test as independent variable and performance in an indirect test as dependent variable in a linear regression analysis (Greenwald et al. 1995). These authors used a semantic priming paradigm as indirect test to show semantic activation without conscious perceptibility of the prime. It seems plausible that better performance in the direct test results in higher accuracy in the indirect test, yielding a linear regression function with positive slope. $X = 0$ denotes the point where conscious perception is completely absent. If there existed no effect of unconscious perception, the regression line would cross the zero point: $Y = 0$ if $X = 0$. If, however, there were an effect of unconscious perception even under complete absence of conscious identification, the intercept would be positive: $Y > 0$ if $X = 0$. In several experiments employing this kind of analysis, Greenwald and his colleagues (Draine and Greenwald 1998, Greenwald et al. 1995, 1996) were
able to show reliable indirect effects under absence of direct effects.

4. *Evidence from the Neurosciences*

Other evidence of perception without awareness comes from patients with damage to the visual cortex. They typically do not see anything in the hemifield of vision that is opposite to the side of the damage (e.g., Pöppel et al. 1973, Weiskrantz et al. 1974). However, some of these patients are able to make visual discriminations in the blind hemifield although they reportedly do not have any phenomenal experience of the stimuli. This phenomenon is called 'blindsight.' Some authors suggested that blindsight may be caused by small islands of intact tissue in the primary visual cortex (Fendrich et al. 1992). Although correct visual discriminations may be caused by small, intact islands of accurate vision, this hypothesis cannot account for all phenomena observed in both monkeys and human patients (see Weiskrantz 1995).

Brain imaging techniques led to new indirect measures. Using an unconscious priming paradigm, Dehaene et al. (1998) have shown that masked primes influenced subsequent processing, accompanied by measurable modifications of electrical brain activity and cerebral blood flow.

We have discussed studies aimed at demonstrating that unconscious perception exists. We now turn to the question of how unconscious perception works.

5. *Differences Between Conscious and Unconscious Perceptual Processes*

One may think that unconscious perception is the same as conscious perception, except for the lack of phenomenal experience. In this view, unconscious perception is just a faint shadow of conscious perception. This view has been questioned, and there is good evidence that unconscious perception is driven by a limited-capacity system, unconscious perception

by a system that has much more capacity and allows the use of information from multiple sources (see MacLeod 1998). The main evidence comes from demonstrations of qualitative dissociations between conscious and unconscious perception. This means that conscious perception affects a variable in one direction, unconscious perception in the opposite way, suggesting different underlying mechanisms (see Merikle and Daneman 1998). An elegant study showing such a dissociation has been reported by Jacoby and Whitehouse (1989). In a recognition test, a target word was presented and participants had to decide whether or not it had appeared in a list presented before. Shortly before the target, a prime was shown that was either identical to or different from the target. If the primes were shown subliminally, identical primes resulted in more 'old' responses than different primes. If the primes were shown supraliminally, however, identical primes resulted in less 'old' responses than different primes. This qualitative dissociation suggests that participants who saw supraliminal primes were able to correct their judgments for the supposed influence of the primes. If exposed to subliminal primes, however, participants were unable to avoid the impact of the prime and were biased toward 'old' responses when prime and target were identical. Several other studies yielded similar dissociations (e.g., Debner and Jacoby 1994, Marcel 1980, Merikle and Joordens 1997), suggesting that conscious and unconscious perception are governed by different mechanisms.

6. *Outlook*

There is fair evidence that unconscious perception exists. However, there is no evidence that effects of unconscious perception, caused by exposure to subliminal messages, are of any practical relevance (Greenwald et al. 1991, Vokey and Read 1985).

Beyond the proof of existence, interesting questions about unconscious perception remain open. One question pertains to the durability of unconscious perception (see Merikle and Daneman 1998). It has been shown that implicit memory effects may last for several days (see *Implicit Memory, Cognitive Psychology of*). In contrast, most studies on unconscious perception dealt with effects that lasted a few seconds. In implicit memory tasks, encoding is conscious, but retrieval is assumed to be unconscious. In perception without awareness, encoding is unconscious. Do long-term effects exist only if encoding is conscious, or is it possible to find long-term effects of unconscious encoding? Poetzl ([1917] 1960) has shown that briefly presented materials may be manifest in dreams during the following night although people were unable to remember the materials after the initial presentation. The Poetzl phenomenon, together with more recent findings on memory for information encoded during anesthesia (see Merikle and Daneman 1996 for a meta-analysis) suggest that unconscious influences may last for hours.

As discussed above, another major question is how perception without awareness works. This research has only recently begun and will intensify, stimulated by recently developed methods in the neurosciences and cognitive psychology (e.g., the process dissociation procedure, see Debner and Jacoby 1994) that open up opportunities for research into neural correlates and cognitive mechanisms of unconscious perception.

Brain imaging methods may bring new opportunities to look into neural pathways related to unconscious perception. Specifically, these methods may bring new insights into qualitative dissociations between neural substrates connected to conscious perception and neural substrates connected to unconscious perception, as demonstrated in a study by Morris et al. (1998). These authors paired an angry face with (uncomfortable) white noise. Subliminal presentation of the conditioned angry face resulted in increased neural activity in the right, but not left amygdala. In contrast, supraliminal presentation of the same face resulted in increased neural activity in the left, but not right amygdala. Moreover, it has been shown that simple stimulus features may proceed to the amygdala and trigger an emotion without involvement of cortical structures (see LeDoux 1995). Although any conclusion about presence or absence of awareness in perceiving such features is premature, it enables theorists to conceive ways of how information may influence mental processes without reaching consciousness.

See also: Consciousness, Cognitive Psychology of; Consciousness, Neural Basis of

Bibliography

Adams J K 1957 Laboratory studies of behavior without awareness. *Psychological Bulletin* **54**: 383–405

Debner J A, Jacoby L L 1994 Unconscious perception: Attention, awareness, and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**: 304–17

Dehaene S, Naccache L, Le Clec' H G, Koechlin E, Mueller M, Dehaene-Lambertz G, van de Moortele P-F, Le Bihan D 1998 Imaging unconscious semantic priming. *Nature* **395**: 597–600

Draine S C, Greenwald A G 1998 Replicable unconscious semantic priming. *Journal of Experimental Psychology: General* **127**: 286–303

Eriksen C W 1960 Discrimination and learning without awareness: A methodological review. *Psychological Review* **67**: 279–300

Fendrich R, Wessinger C M, Gazzaniga M S 1992 Residual vision in a scotoma: Implications for blindsight. *Science* **258**: 1489–91

Greenwald A G, Draine S C, Abrams R H 1996 Three cognitive markers of unconscious semantic priming. *Science* **273**: 1699–702

Greenwald A G, Klinger M R, Schuh E S 1995 Activation by marginally perceptible ('subliminal') stimuli: Dissociation of

unconscious from conscious cognition. *Journal of Experimental Psychology: General* **124**: 22–42
Greenwald A G, Spangenberg E R, Pratkanis A R, Eskanazi J 1991 Double-blind tests of subliminal self-help audiotapes. *Psychological Science* **2**: 119–22
Holender D 1986 Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences* **9**: 1–66
Jacoby L L, Whitehouse K 1989 An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General* **118**: 126–35
Kunst-Wilson W R, Zajonc R B 1980 Affective discrimination of stimuli that cannot be recognized. *Science* **207**: 557–58
LeDoux J E 1995 Emotion: Clues from the brain. *Annual Review of Psychology* **46**: 209–35
MacLeod C 1998 Implicit perception: Perceptual processing without awareness. In: Kirsner K, Speelman C, Maybery M, O'Brien-Malone A, Anderson M, MacLeod C (eds.) *Implicit and Explicit Mental Processes*. Erlbaum, Mahwah, NJ, pp. 57–78
Marcel A J 1980 Conscious and preconscious recognition of polysemous words: Locating the selective effects of prior verbal context. In: Nickerson R S (ed.) *Attention and Performance*. Erlbaum, Hillsdale, NJ, Vol. VII, pp. 435–57
Marcel A J 1983 Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology* **15**: 197–237
Merikle P M, Daneman M 1996 Memory for unconsciously perceived events: Evidence from anesthetized patients. *Consciousness and Cognition* **5**: 525–41
Merikle P M, Daneman M 1998 Psychological investigations of unconscious perception. *Journal of Consciousness Studies* **5**: 5–18
Merikle P M, Joordens S 1997 Measuring unconscious influences. In: Cohen J D, Schooler J W (eds.) *Scientific Approaches to Consciousness*. Erlbaum, Mahwah, NJ, pp. 109–123
Morris J S, Öhman A, Dolan R J 1998 Conscious and unconscious emotional learning in the human amygdala. *Nature* **393**: 467–70
Peirce C S, Jastrow J 1884 On small differences of sensation. *Memoirs of the National Academy of Sciences* **3**: 73–83
Poetzl O [1917] 1960 The relationships between experimentally induced dream images and indirect vision. *Monograph 7. Psychological Issues* **2**: 46–106
Pöppel E, Held R, Frost D 1973 Residual visual function after brain wounds involving the central visual pathways in man. *Nature* **243**: 295–6
Reber A S 1993 *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. Oxford University Press, New York
Reingold E M, Merikle P M 1988 Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics* **44**: 563–75
Sidis B 1898 *The Psychology of Suggestion*. Appleton, New York
Vokey J R, Read J D 1985 Subliminal messages. Between the devil and the media. *American Psychologist* **40**: 1231–9
Weiskrantz L 1995 Blindsight—not an island unto itself. *Current Directions in Psychological Science* **4**: 146–51
Weiskrantz L, Warrington E K, Sanders M D, Marshall J 1974 Visual capacity in the hemianoptic field following a restricted occipital ablation. *Brain* **97**: 709–28

R. Reber and W. Perrig

Perceptrons

Perceptrons was the generic name given by the psychologist Frank Rosenblatt to a family of theoretical and experimental artificial neural net models which he proposed in the period 1957–1962. Rosenblatt's work created much excitement, controversy, and interest in neural net models for pattern classification in that period and led to important models abstracted from his work in later years. Currently the names (single-layer) Perceptron and Multilayer Perceptron are used to refer to specific artificial neural network structures based on Rosenblatt's perceptrons. This article references the intellectual context preceding Rosenblatt's work and summarizes the basic operations of a simple version of Rosenblatt's perceptrons. It also comments briefly on the developments in this topic since Rosenblatt.

Rosenblatt's first (partial) report in January 1957 was titled *The Perceptron: A Perceiving and Recognizing Automaton*. A subsequent report, in January 1958, titled *The Perceptron: A Theory of Statistical Separability in Cognitive Systems* was adapted for publication (Rosenblatt 1958). A compendium of research on perceptrons by Rosenblatt and his group is *Principles of Neurodynamics* (Rosenblatt 1962).

Rosenblatt's research was done in the context of: the neobehaviorism of Hull (1952), the general principles of neuromodeling of behavior proposed by Hebb (1949), the Threshold Logic Unit (TLU) neuron model of McCulloch and Pitts (1943), proposals for brain modeling based on neurophysiology (Eccles 1953) and symbolic logic, switching theory and digital computers (Shannon and McCarthy 1956), the stimJEulus sampling models of Estes and Burke (1953), the linear operator probabilistic learning models of Bush and Mosteller (1955), and statistical classification procedures (e.g., Rao 1955).

The basic building block of a perceptron is an element that accepts a number of inputs x_i, $i = 1 \ldots N$, and computes a weighted sum of these inputs where, for each input, its fixed weights β can be only $+1$ or -1. The sum is then compared with a threshold θ, and an output y is produced that is either 0 or 1, depending on whether or not the sum exceeds the threshold. Thus

$$y = \begin{cases} 1 & \text{if} \quad \left(\sum_{i=1}^{N} \beta_i x_i \right) \geqslant \theta \\ 0 & \text{if} \quad \left(\sum_{i=1}^{N} \beta_i x_i \right) < \theta \end{cases}$$

A perceptron is a signal transmission network consisting of sensory units (S units), association units (A units), and output or response units (R units). The 'retina' of the perceptron is an array of sensory elements (photocells). An S-unit produces a binary

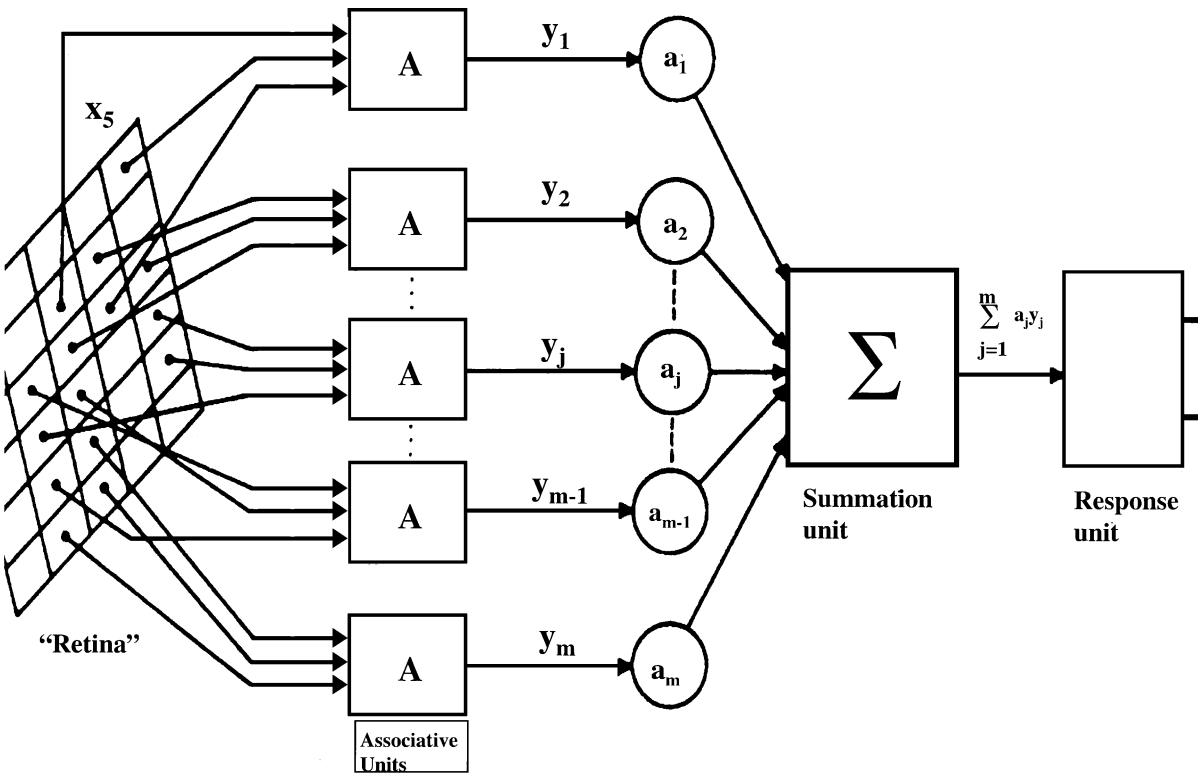

Figure 1
The mark 1 perceptron

output depending on whether or not it is excited. A randomly selected set of retinal cells is connected to the next level of the network, the A units. As originally proposed there were extensive connections among the A units, the R units, and feedback between the R units and the A units. A simpler version omitting the lateral and feed-back connections is shown in Fig. 1 This simplification is the perceptron considered here.

Each A unit behaves like the basic building block discussed above, where the $+1$, -1 weights for the inputs to each A unit are assigned randomly. The threshold θ for all A units is the same. The binary output of the kth A unit ($k = 1, \ldots, m$) is multiplied by a weight a_k, and a sum of all m weighted outputs is formed in a summation unit that is the same as the basic building blocks with all weights equal to $+1$. Each weight a_k is allowed to be positive, zero, or negative, and may change independently of other weights. The output of this block is again binary, depending on a threshold, t, that is normally set at 0. The binary values of the output are used to distinguish two classes of patterns that may be presented to the retina of a perceptron. The design of this perceptron to distinguish between two given sets of patterns involves adjusting the weights a_k, $k = 1, \ldots, m$, and setting the threshold θ.

Rosenblatt (1962) proposed a number of variations of the following procedure for 'training' perceptrons. The set of 'training' patterns, that is, patterns of known classification, are presented sequentially to the retina, with the complete set being repeated as often as needed. The output of the perceptron is monitored to determine whether a pattern is correctly classified. If not, the weights are adjusted according to the following 'error correction' procedure: If the n-th pattern was misclassified, the new value $a_k(n+1)$ for the k-th weight is set to

$$a_k(n+1) = a_k(n) + y_k(n) * \delta(n)$$

where $\delta(n)$ is 1 if the n-th pattern is from class 1 and $\delta(n)$ is -1 if the n-th pattern is from class 2. No adjustment to the weight is made if a pattern is classified correctly.

Rosenblatt conjectured that, when the pattern classes are 'linearly separable' the error correction 'learning' procedure will converge to a set of weights defining the separating hyperplane that correctly classifies all the patterns. The shortest proof of this perceptron convergence theorem was given by A. J. Novikoff. Subsequent contributions related the simple perceptron to statistical linear discriminant functions

and related the error-correction learning algorithm to gradient-descent procedures and to stochastic approximation methods that originally were developed for finding the zeros and extremes of unknown regression functions (e.g., Kanal 1962).

The simple perceptron described is a series-coupled perceptron with feed-forward connections only from S units to A units and A units to the single R unit. The weights a_k, the only adaptive elements in this network, are evaluated directly in terms of the output error.

Minsky and Papert (1969) further reduced the simple perceptron to a structure with sampled connections from the 'retina' directly to the adjustable weights. This reduced structure is referred to as a *single-layer perceptron*. There is no layer of 'hidden' elements, that is, additional adjustable elements for which the adjustment is only indirectly related to the output error. A feed-forward perceptron with one or more layers of hidden elements is termed a *multilayer perceptron*. Rosenblatt investigated *cross-coupled perceptrons* in which connections join units of the same type, and also investigated *multilayer back-coupled perceptrons*, which have feedback paths from units located near the output. For series-coupled perceptrons with multiple R units, Rosenblatt proposed a 'back-propagating error correction' procedure that used error from the R units to propagate correction back to the sensory end. But neither he nor others at that time were able to demonstrate a convergent procedure for training multilayer perceptrons.

Minsky and Papert (1969) proved various theorems about single-layer perceptrons, some of which indicated their limited pattern-classification and function approximating capabilities. For example, they proved that the single-layer perceptron could not implement the Exclusive OR logical function and several other such predicates. Later, many writing on Artificial Neural Networks (ANN) blamed this book for greatly dampening interest and leading to a demise of funding in the USA for research on ANN's. The section on 'Alternate Realities' in Kanal (1992) details why the blame is misplaced. As noted there, by 1962 (see, e.g., Kanal 1962) many researchers had moved on from perceptron type learning machines to statistical and syntactic procedures for pattern recognition.

Minsky and Papert's results did not apply to multilayer perceptrons. Research on ANN's, biologically motivated automata, and adaptive systems continued in the 1970s in Europe, Japan, the Soviet Union, and the USA, but without the frenzied excitement of previous years, which also came back starting in the early 1980s. In a 1974 Harvard University dissertation Paul Werbos presented a general convergent procedure for adaptively adjusting the weights of a differentiable nonlinear system so as to learn a functional relationship between the inputs and outputs of the system. The procedure calculates the derivatives of some function of the outputs, with respect to all inputs and weights or parameters of the system, working backwards from outputs to inputs. However, this work, published later later in a book by Werbos (1994), went essentially unnoticed, until a few years after Rumelhart et al. (1986), independently popularized a special case of the general method. This algorithm, known as error backpropagation or just backpropagation, adaptively adjusts the weights to perform correct pattern classification using gradient descent and training samples. It propagates derivatives from the output layer through each intermediate layer of the multilayer perceptron network. The resurgence of work on multilayer perceptrons and their applications in the decades of the 1980s and 1990s is directly attributable to this convergent backpropagation algorithm.

It has been shown that multilayer feedforward networks with a sufficient number of intermediate or 'hidden' units between the input and output units have a 'universal approximation' property: they can approximate 'virtually any function of interest to any desired degree of accuracy' (Hornik et al. 1989). Several modifications of the basic perceptron learning procedure that make perceptron learning well behaved with inseparable training data, even when the training data are noisy and not error free, have been proposed in recent years and various ANN learning procedures have been shown to related to known statistical techniques. The reference list includes some recent books with tutorial material covering Perceptrons and related ANN's.

See also: Artificial Neural Networks: Neurocomputation; Connectionist Models of Concept Learning; Connectionist Models of Development; Connectionist Models of Language Processing; Neural Networks and Related Statistical Latent Variable Models; Neural Networks: Biological Models and Applications; Perception: History of the Concept

Bibliography

Anderson J A 1995 *An Introduction to Neural Networks*. MIT Press, Cambridge, MA

Bush R R, Mosteller F 1955 *Stochastic Models for Learning*. Wiley, New York

Chauvin Y, Rumelhart D E (eds.) 1995 *Backpropagation: Theory, Architectures, and Applications*. Erlbaum, Mahwah, NJ

Eccles J C 1953 *The Neurophysiological Basis of Mind*. Clarendon Press, Oxford, UK

Estes W K, Burke C J 1953 A theory of stimulus variability in learning. *Psychological Review* **6**: 276–86

Haykin S 1999 *Neural Networks. A Comprehensive Foundation*, 2nd edn. Prentice Hall, Upper Saddle River, NJ

Hebb D O 1949 *The Organization of Behaviour*. Wiley, New York

Hornik K, Stinchcome M, White H 1989 Multilayer feedforward networks are universal approximators. *Neural Networks* **2**: 359–66

Hull C L 1952 *A Behaviour System*. Yale University Press, New Haven, CT
Kanal L 1962 Evaluation of a class of pattern-recognition networks. In: Bernard E E, Kare M R (eds.) *Biological Prototypes and Synthetic Systems*. Plenum Press, New York (reprinted in Sethi I, Jain A 1995 *Neural Networks and Pattern Recognition*. Elsevier, Amsterdam)
Kanal L N 1992 On pattern, categories, and alternate realities, 1992 KS Fu award talk at IAPR, The Hague. *Pattern Recognition Letters* **14**: 241–55
McCulloch W S, Pitts W A 1943 A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–33
Minsky M, Papert S 1969 *Perceptrons*. MIT Press, Cambridge, MA
Rao C R 1955 *Advanced Statistical Methods in Biometric Research*. Wiley, New York
Rosenblatt F 1958 The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**: 386–408
Rosenblatt F 1962 *Principles of Neurodynamics*. Spartan, New York
Rumelhart D E, Hinton G E, Williams R J 1986 Learning internal representations by error propagation. In: Rumelhart D E, McClelland J L et al. (eds.) *Parallel Distributed Processing*: *Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA, Vol. 1, pp. 318–62
Shannon C E, McCarthy J M (eds.) 1956 *Automata Studies*. Princeton University Press, Princeton, NJ
Werbos P 1994 *The Roots of Backpropogation: From Ordered Derivatives to Neural Networks and Political Forecasting*. Wiley, New York

L. N. Kanal

Perceptual Constancy: Direct versus Constructivist Theories

Fortunately, our perception of the environment surrounding us is usually quite veridical, accurate, and stable. While laypersons take this ability of ours for granted, students of perception have pointed out that it is actually quite a feat, since the stimulation reaching our senses changes continuously. If we take the visual perception of size as an example, we find that the size of the retinal image of, say, another person changes with that person's distance from us. In spite of these changes we perceive that person as being of constant size and not shrinking or expanding with changes in distance. This 'feat' of perception has aroused a great deal of interest and theories have been proposed to account for it. Prominent among these are two contrasting views of how it is achieved. One view, the 'constructivist,' maintains that the perceptual system is 'intelligent.' In order to attain veridical perception it utilizes high-level mental processes to correct for the changes in stimulation reaching our senses. In contrast, the 'direct' approach claims that there is no need to posit such mental processes to explain veridical perception, maintaining that the sensory information reaching our senses suffices to explain perceptual veridicality and that what is needed is a much deeper analysis of that sensory information.

1. *Stability and Constancy in Perception*

The ability of our perceptual system to overcome the effects of changes in stimulation and maintain a veridical and stable percept is called *perceptual constancy* (Epstein 1977, Walsh and Kulikowski 1998). In discussing the constancies it is common to make a distinction between the *distal* stimulus and *proximal* stimulus. The distal stimulus is the actual physical stimulus, the physically objective dimensions of the viewed object. The proximal stimulus, in the case of vision, is the very image that falls on the retina. It changes with changes in position or lighting of the physical stimulus. Constancy can be defined as the ability to correctly perceive the distal stimulus; that is, the stable properties of objects and scenes, in spite of changes in the proximal stimulus.

One example of constancy is *shape constancy*, which refers to the ability of the visual system to ascertain the true shape of an object even when it is not viewed face on but slanted with respect to the observer. In an example of shape constancy the distal stimulus might be a door half open, while the proximal stimulus will be the trapezoid shape it projects on the retinas. Shape constancy in this case would be the perception of the distal rectangular shape of the door in spite of the fact that the proximal shape is far from rectangular. Another example of constancy is *lightness constancy*. Lightness refers to the perceived reflectance of an object; high reflectance is perceived as white and very low reflectance as black, with intermediate reflectances as various shades of gray. We are capable of perceiving the distal lightness of a surface in spite of the fact that the proximal amount of light reaching our eyes changes with changes in the amount of light illuminating that surface. For example, if we hold a piece of chalk in one hand and a piece of charcoal in the other, the chalk will be perceived as white and the charcoal as black. This will be true if we observe the two in a dimly lit room or in bright sunshine, in spite of the fact that the amount of light reaching the eyes from the charcoal in the sunshine might be greater than that from the chalk in the dim room lighting.

The perceptual constancy that has received the greatest amount of research attention is *size constancy* (Ross and Plug 1998). If the distance is not very great we usually perceive the distal size of objects in spite of the changes in their proximal size as their distance from us changes. The two theoretical approaches mentioned above, the constructivist and the direct, deal with size constancy differently. According to the constructivists, size constancy is achieved through some process whereby the perceptual system perceives the object's distance and then takes the distance

into account and 'corrects' the proximal image to yield a true distal percept. The direct approach asserts that there is no need to involve the perception of distance in the perception of size. Instead, it claims that enough information exists in the visual environment to allow direct perception of size without the need for aid from additional 'higher' mental processes.

2. Two Theoretical Approaches to Perception

The two approaches, the constructivist and the direct, differ in many respects, but the main differences between them revolve around two interrelated issues, the richness of the stimulation reaching our sensory apparatus, and the involvement of 'higher' mental processes in the apprehension of our environment. The constructivists see the stimulation reaching our senses as inherently insufficient for veridical perception, necessitating an 'intelligent' mechanism that utilizes inferential processes to supplement the information available to the senses and resolve its inherent ambiguity. The direct theorists, in contrast, argue that the information in the environment is sufficiently rich to afford veridical perception with no need for sophisticated algorithms. The constructivists see perception as multistage with mediational processes intervening between stimulation and percept; i.e., perception is indirect. The direct theorists see perception as a single-stage process; i.e., it is direct and immediate. For the constructivists, memory and schemata based on past experience play an important role in perception. The direct approach sees no role for memory in perception. Finally, the two approaches differ in the aspects of perception they emphasize: the constructivists excel at analyzing the processes and mechanisms underlying perception, whereas the direct approach excels at the analysis of the stimulation reaching the observer.

2.1 The Constructivist Approach

This approach has also been called Helmholtzian, indirect, cognitive, algorithmic, and mediational, among other labels. It is the older of the two, ascribing some of its roots to Helmholtz's (1867) notion of 'unconscious inference,' which maintains that veridical perception is achieved by an inference-like process which transpires unconsciously. A somewhat similar approach was adopted and elaborated more recently by Rock (1983)in his *The Logic of Perception*. This book contains the most detailed account of the constructivist view and the evidence in its favor. The first sentence in the book is, 'The thesis of this book is that perception is intelligent in that it is based on operations similar to those that characterize thought.' (Rock 1983, p.1). Rock makes it clear that this thought-like process occurs unconsciously. He also writes, '… a summary statement of the kind of theory I propose to advance in the remainder of the book. My view follows Helmholtz's (1867) that perceptual processing is guided by the effort or search to interpret the proximal stimulus, i.e., the stimulus impinging on the sense organ, in terms of what object or event in the world it represents, what others have referred to as the "effort after meaning".' (Rock 1983, p. 16). In other words, Rock is conceiving of perception as an effortful, but unconscious, attempt at identifying an object or event.

In his discussion of size perception, Rock specifically invokes a syllogistic inferential mechanism: 'I will argue that the process of achieving constancy is one of deductive inference where the relevant "premises" are immediately known. That is to say, in the case of a specific constancy such as that of size, two aspects of the proximal stimulus are most relevant, one being the visual angle subtended by the object and the other being information about the object's distance.' (Rock 1983, p. 240). Rock sees size perception as depending on two perceptions, that of proximal size and that of distance, together leading through a syllogism to the veridical distal percept. Like other constructivists, this approach posits some sort of combination of proximal size information and distance information in the achievement of size constancy. In a similar manner, the same combinatorial process holds for all the constancies, according to the constructivist view.

The equation of perceptual and thought processes is difficult to verify empirically, but there are implications from this approach that are verifiable. As can be seen above, the constructivist approach implies that at least two perceptions are involved in perceiving size, that of the proximal size of the object and that of its distance. In a later book, *Indirect Perception*, Rock (1997) collected a large group of studies showing 'that perception is based on prior perception, implying a perception–perception chain of causality.' This interdependence of perceptual processes is something that can be examined empirically and, indeed, the studies reprinted in that book clearly evidence that chain of causality. However, it should also be noted that all of these studies utilize somewhat artificial situations where a manipulation is used in order to create some sort of illusory perception or misperception. A simple example is studies where subjects are presented with monocular views of the stimuli causing misperception of the three-dimensional spatial arrangement. Here, for example, misperceived distance is accompanied by the misperception of other perceptual qualities.

2.2 The Direct Approach

This approach is also referred to as ecological, Gibsonian, sensory, and immediate, among other labels. It is in the main an approach developed by James J. Gibson who expounded it in his *The Ecological Approach to Visual Perception* (Gibson

1979). In that book, Gibson proffered an exciting new approach to the study of visual perception that included many new concepts and new ways of looking at perception. The entire first half of the book is devoted to a novel analysis of the ambient, or surrounding, environment and the information it provides the observer. Gibson finds the classical approach of describing the stimuli for perception in terms of stimulus energies impinging upon the receptors completely unsatisfactory. He points to the differences between these energies and the optical information available in the ambient optic array. That information is *picked up* by a stationary or moving observer. Perception consists of perceiving events, i.e., perceiving changes over time and space in the optic array.

What sort of information is picked up in direct perception? Gibson suggests that there exist invariants in the optic array that serve to supply the observer with unequivocal information. These invariant aspects of the information remain constant in spite of changes in the observer's viewpoint. He musters a great deal of evidence to prove this point. Among the items of evidence he presents is a study of size perception of his where observers were asked to match the height of stakes planted at various distances in a very large plowed field with a set of stakes of varying size nearby. His finding was that size perception remained veridical no matter how far away the stake was planted. Unlike the constructivists, Gibson does not ascribe this size constancy to the taking into account of distance, but rather to the pickup of invariant ratios in the ambient optical array surrounding the observer. In the case of perception of size, he proposes two such invariant ratios. One is the amount of surface texture intercepted by the object's base as it sits on that surface. The second is the horizon ratio, the ratio between the object's height and its distance from the horizon, both measured in the two-dimensional projection of the optic array. There is no need, according to his view, for perceived distance to be involved here, or for the inferential mental processes that the constructivists purport to underlie size perception.

One of Gibson's most important contributions is the concept of *affordances*. He writes, 'The *affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*, either for good or for ill' (Gibson 1979, p. 127). In other words, affordances refer to the possible actions that can be taken by the perceiver vis-à-vis objects in the environment. Thus, for example, some surfaces in the environment are 'stand-on-able,' or 'climb-on-able,' or 'sit-on-able.' 'The psychologists assume that objects are *composed* of their qualities. But I now suggest that what we perceive when we look at objects are their affordances, not their qualities.' (Gibson 1979, p. 134), and, '… the basic affordances of the environment are perceivable and usually perceivable directly, without an excessive amount of learning.' (Gibson 1979, p. 143). Gibson is suggesting that the direct perception of, say, a chair does not constitute recognizing it as a type of furniture labeled 'chair,' i.e., categorizing it, but rather the perception of the chair's affordance, i.e., that it contains a surface that is sit-on-able.

As may be clear from the foregoing description, Gibson's conception is one of an active perceiver exploring his environment. Eye, head, and body movements are part and parcel of the perceptual process. Perception transpires continuously over both time and space. 'Space' here refers not to an empty space but to the many surfaces that make up the environment, the most important being the terrain that at times reaches the horizon. The horizon is of importance as it serves as a useful reference standard, and when it is occluded Gibson speaks in terms of an implicit horizon, presumably similar to what architects and others have called the eye-level plane. With such a conception, Gibson is totally averse to the reductionist experimental paradigms used by the constructivists. Brief exposures or looks through monocular 'peepholes' do not represent true perception, in his view.

In his book, Gibson almost totally refrains from discussing the processes underlying perception. Perception is simply the pickup of information from invariants in the ambient environment. His only allusions to underlying processes are in terms of resonance: 'In the case of the persisting thing, I suggest, the perceptual system simply extracts the invariants from the flowing array; it *resonates* to the invariant structure or is *attuned* to it. In the case of substantially distinct things, I venture, the perceptual system must *abstract* the invariants. The former process seems to be simpler than the latter, more nearly automatic' (Gibson 1979, p. 249).

3. *Summing Up and a Possible Reconciliation*

The two theoretical approaches, the constructivist and the direct, obviously differ in many respects. The former endows the perceiver with 'higher' mental capacities that allow him/her to compensate for the purported insufficiency of the stimulation reaching the sensory apparatus. In contrast, the direct approach attempts to delineate the richness of information available in the ambient environment as picked up by an active perceiver, allowing direct perception without the need for additional mental processes. What is more, the proponents of each of these two approaches utilize research methods commensurate with the tenets of the respective approach. Thus, for example, constructivist researchers utilize impoverished or illusory stimulus conditions (e.g., brief and/or monocular presentations), calling for the intervention of 'higher' processes to compensate for the lack of adequate stimulus information. The adherents of the direct theory, on the other hand, utilize rich stimulus conditions without time limitations.

In spite of the seemingly irreconcilable differences between the two theoretical approaches, a means of bridging them does exist. They can co-exist within a broader theory of perception (Norman 2001). This broader theory is based on the accumulating neurophysiological, neuropsychological, and psychophysical research findings that point to the existence of two distinct visual systems, the dorsal and the ventral (see, e.g., Milner and Goodale 1995). The central idea is that the direct approach broadly parallels the functions of the dorsal system, and the constructivist approach broadly parallels those of the ventral system. These two visual systems are found in different parts of the brain, the dorsal mainly in the parietal lobe, the ventral mainly in the temporal lobe. The two deal with different aspects of perception. The dorsal system deals mainly with the utilization of visual information for the guidance of motor behavior in one's environment. The ventral system deals mainly with the utilization of visual information for 'knowing' one's environment, i.e., identifying and recognizing items previously encountered and storing new visual information for later encounters. However, it should be stressed that both systems process similar information for somewhat different purposes. Thus, for example, both systems are involved in the perception of size, albeit with somewhat different specializations. The dorsal system picks up size information for motor tasks, such as grasping an object. The ventral system utilizes size information for distinguishing between objects, say, between a real car and an accurate toy model.

Emerging from this attempt at integrating the two approaches into a single theory is a dual-process approach to visual perception. According to this approach, much of our day-to-day pickup of visual information is carried out by the dorsal-direct system without involving much awareness. In the main the information picked up is that which allows the organism to function within its environment, i.e., Gibson's affordances. The ventral-constructivist system, on the other hand, deals with the interface between the visual input and cognition, and we are normally aware of its output. Only it possesses a long-term memory and therefore any type of identification or recognition must transpire within it. As the dorsal system is mainly concerned with directing motor behavior in one's environment it must rely on body-centered information, in absolute units, about the environment and the objects in it. In contrast, the ventral system in its attempt to recognize objects can suffice with relative, object-centered, information. It is probably for this reason that the dorsal system yields more accurate information about the visual environment than does the ventral system. Thus, it is suggested that perceptual constancy is usually attained by the dorsal system. However, there are cases where the ventral system is also involved, as when the visual information is sparse or ambiguous, or when the observer is required not only to act upon some aspect of the visual environment, but also to make a judgment about some object in that environment.

See also: Constructivism in Cognitive Psychology; Helmholtz, Hermann Ludwig Ferdinand von (1821–94); Object Recognition: Theories; Perception: History of the Concept; Perceptual Organization; Sensation and Perception: Direct Scaling; Vision, High-level Theory of

Bibliography

Epstein W (ed.) 1977 *Stability and Constancy in Visual Perception*. Wiley, New York

Gibson J J 1979 *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston

Helmholtz H von 1867/1962 *Treatise on Physiological Optics* (Southall J P C, ed.), translation of the third German edition. Dover, New York

Milner A D, Goodale M A 1995 *The Visual Brain in Action*. Oxford University Press, Oxford, UK

Norman J 2001 Two visual systems and two theories of perception: an attempt to reconcile the constructivist and ecological approaches. *Behavioral and Brain Sciences* **24**(6) in press

Rock I 1983 *The Logic of Perception*. MIT Press, Cambridge, MA

Rock I 1997 *Indirect Perception*. MIT Press, Cambridge, MA

Ross H E, Plug C 1998 The history of size constancy and size illusions. In: Walsh V, Kulikowski J (eds.) *Perceptual Constancy: Why Things Look as They Do*. Cambridge University Press, Cambridge, UK, pp. 499–528

Walsh V, Kulikowski J (eds.) 1998 *Perceptual Constancy: Why Things Look as They Do*. Cambridge University Press, Cambridge, UK

J. Norman

Perceptual Learning

1. Improvement of Perception Through Training

It has been known for a long time that visual recognition of objects improves through training. For beginners in the study of histology, all specimens, such as the liver, lung, or kidney, look quite similar. Sooner or later, however, the advanced student wonders how one could possibly miss the difference. This type of visual classification is a relatively complex visual task, but it has also been known for a long time that performance in much simpler perceptual tasks improves through practice, as for example in the cases of vernier discrimination (McKee and Westheimer 1978), stereoscopic depth perception (Ramachandran and Braddick 1973), and discrimination between line orientations (Vogels and Orban 1985). This type of learning is usually referred to as 'perceptual learning.' Compared to other domains of research on learning,

perceptual learning is a relatively new topic, so its definition is even more important than those of topics that were established earlier. Gibson (1963) proposed the following definition for perceptual learning: 'Any relatively permanent and consistent change in the perception of a stimulus array following practice or experience with this array will be considered perceptual learning.' In recent years, one would, in addition, stress the fact that the neuronal mechanisms underlying perceptual learning involve rather 'early' stages of cortical information processing. Learning usually leads to improved performance, but there are exceptions, at least under laboratory conditions with manipulated feedback.

2. *Terms Related to Perceptual Learning*

Quite a number of processes in the nervous system are able to change their responses to sensory stimulation and there is quite a number of terms related to the term 'learning.' To understand the concept of perceptual learning clearly these terms have to be defined in order to clarify differences and similarities. These related terms include 'plasticity,' 'adaptation,' 'habituation,' 'after-effects,' 'priming,' 'development,' and 'maturation' as well as 'improvement through insight.'

'Plasticity' is defined here as modifiability of functional and anatomical organization of the central nervous system leading to more appropriate function as a result of sensory experience, or to overcome limitations caused by lesions. The term 'adaptation' is most often used in relation to adjustments of information processing within a predefined working range, as a result of stimulation. A classic example is luminance adaptation—adjusting the working range of the visual system to ambient light levels, with no long-term consequences. 'Habituation,' or satiation' in the context of rewards, seems to be a special case of adaptation, namely a shift of working range towards *lower* sensitivity, as in the case of decreased reflex response after repeated stimulation. 'After-effects' can be considered, in many cases, as the result of selective short-term adaptation in cases where perception is the result of antagonistic neuronal channels, as with many after-images. 'Priming' describes the effect of a (sometimes subliminal) stimulus on a subsequent stimulus–reaction pair, or more generally, the effect of initiating a certain type of behavior. The effect of priming is usually short. 'Development' and 'maturation,' unlike learning, ascribe the thrust of the changes in behavior to genetics, not the environment; hence the mechanisms underlying changes of behavior differ radically between these phenomena on one side and perceptual learning on the other. 'Improvement through insight,' as the name indicates, is a term that should be reserved for positive changes of information processing based on cognitive processes, such as one-shot learning.

So what is so special about perceptual learning that earns it a separate entry in this encyclopedia as opposed to a passing mention under the heading 'Learning'? I would like to stress the following difference between 'ordinary' vs. 'perceptual' learning. Most forms of learning to collect better information about the outside world are generally considered to rely on relatively intelligent or complex or cognitive levels of information processing. Hence, learning leading to improved discrimination between sensory stimuli would have been considered only a decade ago to be more of the declarative type of learning (as in learning a poem) than of the procedural type of learning (as in learning to play the piano). Perceptual learning, on the other hand, is a form of learning leading to better use of sensory information which is relatively independent of conscious or declarative forms of learning but relies partly on rather low-level modifications in the central nervous system. Perceptual learning hence resembles, in many respects, procedural forms of learning that are common in motor learning, for example learning to ride a bicycle.

3. *Perceptual Learning and Plasticity of Primary Sensory Cortices*

This type of low-level modification would have been considered impossible in the 1980s. At this time, the primary visual cortex of adults was considered to be a rather 'hard-wired' first stage of visual information processing, which served to detect and extract certain 'elementary features' from the complex scenes surrounding us (cf. Marr 1982). It was considered to lack plasticity since changing preprocessing during the training of one task might have disadvantageous consequences for solving other perceptual tasks. Indeed, there is direct experimental evidence for a decrease in plasticity of the primary visual cortex during maturation. Hubel and Wiesel (1965) found that covering one eye in young cats leads to changes in their cortical wiring patterns that are reversible during their kittenhood but not thereafter. Similarly, children suffering from a squint may achieve only low visual acuity in the deviating eye; therapy can reverse this loss in small children only. Therefore it seems that a critical development phase exists during which the primary visual cortex is still plastic, but that it loses this plasticity at a later stage of development. Improvement through training in perceptual tasks by adults was therefore supposed to take place on higher levels of cortical information processing.

However, during the last decade of the twentieth century, a number of electrophysiological and psychophysical experiments cast some doubts on this interpretation and suggested that even adult primary sensory cortices showed much more plasticity than was hitherto believed. Some of this evidence will be

presented here. A first example of this evidence was the orientation specificity of improvement through learning. For example, Fiorentini and Berardi (1981) found that discrimination between complex types of grating improved through practice, but that the improvement did not transfer to stimuli rotated by 90 degrees. The same was true for a vernier discrimination task where observers had to indicate whether the lower segment of a vertically oriented vernier stimulus consisting of two almost collinear bars was offset to the right or to the left relative to the upper segment, or whether the right segment of a horizontal vernier was above or below the left segment. Observers improved their performance very significantly during training, but their performance returned to base level when the stimulus was rotated by 90 degrees. Hence, the improvement obtained through training was specific for stimulus orientation.

In a control experiment, the pushbuttons by which observers had to indicate the direction of offset were switched between hands. Observers had to push the left button when a vernier was offset to the right, and vice versa. Results did not deteriorate, indicating that not every change of experimental procedure leads to a drop in performance. Another group of observers were trained in vernier discriminations at eight different positions in the periphery of the visual field for one hour at each position. During that time, they improved performance, but returned to baseline at the transition from one visual field position to the next. The same was true for training vernier discriminations with one eye covered: observers improved their performance through training, but when they performed the same discriminations with the cover moved to the other eye, performance dropped to pretraining levels. The same specificity of improvement for the eye used during training had been observed earlier in a texture discrimination task in which observers had to discriminate a figure from its surround based on the orientation of the stimulus elements (Karni and Sagi 1991). Moreover, improvement did not transfer between a three-dot vernier and a three-dot bisection task, although the stimuli of both tasks differed by approximately one photoreceptor diameter.

4. *Neuronal Basis of Perceptual Learning*

Hence, the changes of the nervous system underlying these forms of perceptual learning should occur on a quite early level of cortical information processing, in which the neurons are already selective for different stimulus orientations—unlike in the retina—but still partly monocularly activated—unlike in all cortical areas beyond the primary visual cortex. In particular, the eye specificity of improvement points to the primary visual cortex as the most probable site for at least a large part of the changes underlying these forms of perceptual learning.

Recent electrophysiological evidence supports this conclusion of a relatively early site of parts of the neuronal changes underlying perceptual learning. Single-cell recordings in adult monkeys have demonstrated that receptive fields of neurons in the primary visual cortex of adult primates can change position after the parts of the retina that were supplying the input to these cells have been destroyed (Gilbert and Wiesel 1992), that the microcircuitry of the adult visual cortex can change (Eysel 2001), and that the distribution of mass potentials evoked by visual stimulation in humans changes as a result of training, especially pronounced for short latencies and over the primary visual cortex.

5. *Different Cortical Levels of Perceptual Learning*

Marr's insight is still true—changing the very front end of information processing as a result of learning one perceptual task would necessarily change the processing of many, if not all, other stimuli presented to the same sensors. The speed and amount of learning depend strongly on attentional control, that is on top-down influences within the brain. Hence, present models of perceptual learning increasingly emphasize that learning will occur at quite a number of different levels of information processing, and that top-down influences from 'higher' levels will play a crucial role in adjusting, in a task-dependent way, processing on the 'lower' levels. Recent experiments have indeed provided direct evidence for strong top-down influences on perceptual learning as a result of error feedback and attentional processes (Ahissar and Hochstein 1993, Herzog and Fahle 1998).

It seems that the high specificity of perceptual learning is partly lost if relatively easy tasks are learned, while specificity of improvement is highest for very difficult tasks. A possible explanation is that 'easy' tasks are learned on a relatively 'higher' level of information processing, at which the information extracted from the visual scene is better used after training than was possible before. Difficult tasks, on the other hand, may require changes at lower levels of processing that are specific, for example, to exact visual field position as well as stimulus orientation.

6. *Conclusions*

To conclude, perceptual learning differs from other forms of learning, especially declarative forms, in that it can be very task- and stimulus-specific, and probably involves functional and anatomical changes even in primary sensory cortices. Though at first sight perceptual learning seems to resemble declarative forms of learning, and to rely on relatively complex cognitive processes, the specificity of improvement for quite low-level attributes such as visual field position,

stimulus orientation, and the eye used for training a visual task indicates a strong and crucial involvement of the primary visual cortex, where neurons are still partly monocularly activated. Dependence on attention, error feedback, and 'insight,' on the other hand, demonstrates that strong top-down influences play a major role in perceptual learning and that perceptual learning also involves more cognitive levels of the brain. Hence, the study of perceptual learning processes not only shows us the amazing amount of plasticity even in adult sensory information processing at a relatively peripheral level, but also leads to a view of cortical information processing not as a feedforward system of subsequent neuronal layers but as a complex and plastic feedback system with strong and important top-down influences that shape 'lower' or 'early' parts of information processing.

Bibliography

Ahissar M, Hochstein S 1993 Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the USA* **90**: 5718–22

Eysel U 2001 Plasticity of reactive fields on early stages of the adult visual system. In: Fahle M, Poggio T (eds.) *Perceptual Learning*. MIT Press, Cambridge, MA

Fahle M, Poggio T (eds.) 2001 *Perceptual Learning*. MIT Press, Cambridge, MA

Fiorentini A, Berardi N 1981 Perceptual learning specific for orientation and spatial frequency. *Nature* **287**: 43–4

Gibson E J 1963 Perceptual learning. *Annual Review of Psychology* **14**: 29–56

Gilbert C D, Wiesel T N 1992 Receptive field dynamics in adult primary visual cortex. *Nature* **356**: 150–2

Herzog M H, Fahle M 1998 Modeling perceptual learning: Difficulties and how they can be overcome. *Biological Cybernetics* **78**: 107–17

Hubel D H, Wiesel T N 1965 Binocular interaction in striate cortex of kittens reared with artificial squint. *Journal of Neurophysiology* **28**: 1041–59

Karni A, Sagi D 1991 Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the USA* **88**: 4966–70

Marr D 1982 *Vision*. Freeman, San Francisco

McKee S P, Westheimer G 1978 Improvement in vernier acuity with practice. *Perception & Psychophysics* **24**: 258–62

Ramachandran V S, Braddick O 1973 Orientation-specific learning in stereopsis. *Perception* **2**: 371–6

Vogels R, Orban G A 1985 The effect of practice on the oblique effect in line orientation judgments. *Vision Research* **25**: 1679–87

M. Fahle

Perceptual Organization

Unless we adopt a rather artificial analytical attitude, we perceive an environment neatly partitioned into objects rather than a mere collection of sensations. Here is a (red) cherry lying on a (yellow) plate, not merely a blob of red within a yellow surround. The concept of perceptual organization, the grouping of parts into larger units, originated with Gestalt psychology in its attack on the atomistic assumption of structuralism around the beginning of the twentieth century. A then widely held view was that complex percepts are simply additive combinations of sensory elements. Each sensory 'atom' was thought to be independent of all other atoms, tied together to larger complexes just by mechanisms of associative learning. The Gestaltists, on the contrary, emphasized that what we perceive is not 'atomistic' elements, but structured forms and shapes which interact within a perceptual 'field' of dynamic 'forces.' Perception was considered to rely on interactive processes within the nervous system determined by dynamic field forces.

1. Grouping Factors in Vision

Rules for how parts of visual stimuli are grouped together, 'Gestalt laws,' were first formulated by Max Wertheimer (1923) who essentially identified four factors: nearness (proximity), similarity, common fate, and closure.

1.1 Nearness or Proximity

Other factors being equal, the nearer objects are to one another, the more likely they are to be organized into unified percepts. An example is shown in Fig. 1a, where the dots are easily organized into pairs of dots on the basis that each two neighboring dots form a unit. It is almost impossible to see the spaced dots as belonging together or, in other words, to break up the units based on nearness.

1.2 Similarity

The greater the similarity among objects, the more likely they are to be organized into perceptual units. For example, dots of the same color are readily seen as a group in distinction from dots of another color (Fig. 1b).

1.3 Common Fate

Dots which move simultaneously at the same rate and in the same direction, share a common fate, are readily seen as a group. The factor of common fate belongs to the strongest grouping factors; it ties together objects that are quite distant and different in form, size or color, given that they share a sort of event similarity in their sameness of motion. Where grouping would not otherwise occur in stationary displays, it will occur as soon as parts of the display move. A well-camouflaged animal will remain hidden only as long as it is standing still; it becomes visible ('pops out') as soon as it moves.

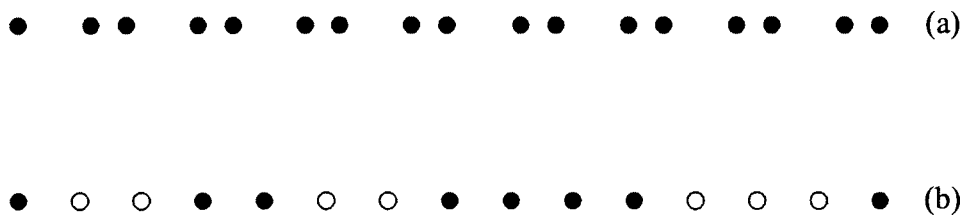

Figure 1
Demonstration of proximity (a) and similarity (b) as grouping factors

Figure 2
Demonstration of closure as grouping factor

Figure 3
Demonstration of *Einstellung* and past experience as factors in amodal completion

1.4 Closure

Components that constitute a closed entity rather than an open one are more readily organized into a single percept. This factor of closure predicts that a closed line has an advantage over an open one. For example, the two arrangements of four dots in Fig. 2a are seen as a rectangle or diamond rather than as crosses (×) or (+), because the former are closed. Closure also applies to the tendency to complete otherwise incomplete, e.g., partly occluded, objects. The configuration shown in Fig. 2b appears as a trapezoid overlapping a triangle rather than as a complicated eleven-sided figure.

1.5 Subjective Factors

The study of natural camouflage and concealment shows that those grouping factors that refer to objective or stimulus variables fare well in accounting for the utility of various surface markings of animals. In addition, there are subjective or central factors at work, such as an observer's momentary set, or *Einstellung*, and past experience, or custom, in determining how the perceptual field is organized. For example, the three separate lines shown in Fig. 3a may appear, at first glance, as a meaningless geometric arrangement of lines. However, the same configuration, just rotated by 180° (Fig. 3b), will be easily

recognized as a capital E with the missing contours being perceptually completed. Now, if the reader with the *Einstellung* of an 'E' looks back at Fig. 3a, he or she will see a rotated E rather than a mere assembly of three lines. The perceptual formation of an E implies amodal completion of contours in the absence of stimulus (modal) information.

2. *Figure–Ground Differentiation*

Why do we see things rather than the holes between them? This question addresses another important aspect of organization, the differentiation of the perceptual field into figure and ground. This aspect was first brought out clearly by Edgar Rubin (1915). For example, if we look at a pencil on a desk, the pencil appears as a well-marked part of the perceptual field, the figure, while the desk appears as a relatively formless, featureless mass which we call the ground. The figure has the character of a thing, whereas the ground appears like unformed material. The ground seems to extend continuously behind or, sometimes, in front of the figure, i.e., although the figure usually appears in front, the ground is not always identical with background. For example, in looking through a window at a tree, the window screen appears as ground, but is clearly seen in front of the figure, the tree.

2.1 *Dynamics of Figure–Ground Perception*

Rubin pointed out that even if all parts of a configuration are grouped together properly, it can appear either as object (figure) or surface (ground) under certain conditions. These so-called ambiguous or reversible figures 'flip' when, for example, the attention is directed to a specific point on the stimulus pattern. The dynamics of figure–ground perception due to the multivalence of the stimulus field are impressively demonstrated in a pattern shown in Fig. 4 (from Spillmann and Ehrenstein 1996)

New 'flowers' (figures) rise from the 'meadow' (ground) and disappear again, illustrating the vain attempt to establish a tendency toward a better and more stable percept. Apparently, the pattern allows for several groupings of equal probability and hence perception fails ever to reach a steady state. This example nicely illustrates that the visual space, as Kanizsa (1979, p. 181) artfully put it, is not a static geometrical scheme, nor is it a simple transposition on the perceptual level of the topographical arrangement of the retinal stimulus, but rather should be considered the result of an extremely dynamic event.

2.2 *Tendency Towards Prägnanz*

The Gestaltists assumed that the process of figural segregation and differentiation would follow a general tendency towards figural goodness or *Prägnanz* (Wertheimer), which denotes a processing principle for percepts to become structured in the most regular, symmetrical, and simple manner possible under the prevailing stimulus conditions. This process was understood by Wolfgang Köhler as a kind of self-organization according to electrodynamic field forces inherent in the sensory system. 'In all processes which terminate in time-independent states the distribution

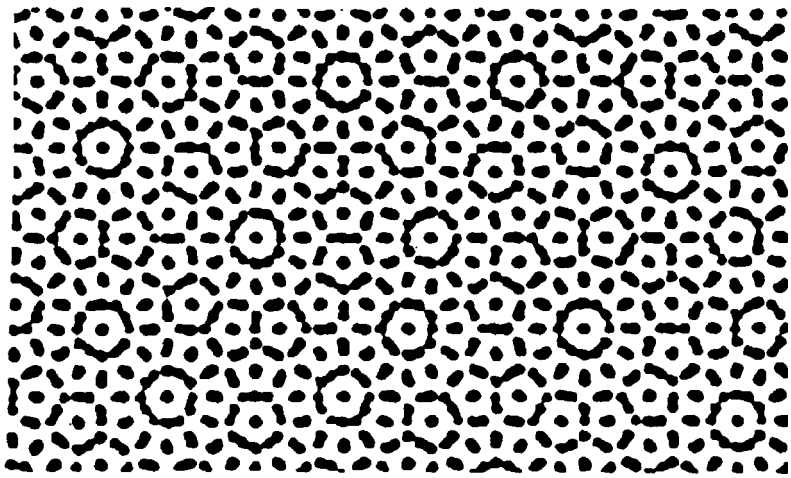

Figure 4
A pattern which affords a never-ending fluctuation in grouping (after Spillmann and Ehrenstein 1996)

shifts towards a minimum energy' (Köhler 1920, quoted from Koffka 1935, p. 108). More than any other principle, the tendency towards *Prägnanz* emphasizes the global aspects of perceptual organization, since even minor local changes in the stimulus or in the observing process (such as eye movements or changes in attention) can result in a reorganization of the entire percept (see Fig. 4). A modern approach to the problem of self-organization and *Prägnanz* can be found in the theory of neural synergetics with its tendency towards ordered steady states in dynamic vector fields (Haken and Stadler 1990).

3. Perceptual Organization in Other Sense Modalities

Although the principles of perceptual organization have been studied most thoroughly for the visual modality, they also apply to other sense modalities. For auditory perception, Bregman (1990) has provided many examples demonstrating the resemblance between audition and vision. Temporal proximity of tones is more important than spatial proximity: Tones that follow each other close together in time are perceived as belonging together. Similarity is determined primarily by the pitch of tones: tones similar in pitch tend to be grouped together in perception. When a sequence of alternating high- and low-frequency tones is played at a certain rate, listeners are able to follow the entire sequence of tone, provided that the frequency difference between tones is small and the tones are played at a slow rate. With greater frequency differences or higher rates, however, the sequence splits into two streams, one high and one low in pitch (auditory stream segregation).

Other examples—studies in tactile perception, for instance—confirm the notion that the above-mentioned factors of perceptual grouping play a similar role in different modalities. This suggests that the perceptual system can utilize organizing principles from one sensory modality to the other, i.e., in a cross-modal way. Moreover, cross-modal interactions can change the percept. A recent example concerns the perception of emotions. Listeners having to judge the emotion in the voice are influenced by whether the (visually displayed) face expresses the same emotion or a different one (De Gelder and Vroomen 2000).

4. Neural Correlates of Perceptual Organization

More than 70 years after Wertheimer's pioneering formulation of perceptual grouping factors, terms that were originally introduced to describe qualitatively properties of figural grouping and segregation are being used as a much-needed vocabulary for interpreting spatiotemporal processing of sensory neurons and neuronal systems. The Gestaltists themselves, notably Wolfgang Köhler, tried to link perceptual organization to brain processes. Inspired by the laws of thermodynamics, the brain was thought to be a homogeneous conductor of bioelectric forces rather than a complex network of interacting neurons (e.g., Köhler and Held 1949). However, the field forces postulated by Köhler as factors in perceptual organization could not be substantiated by physiological evidence. Neural correlates of perceptual grouping are today assumed to consist of receptive field organization and synchronized coupling of neuronal activity within and across cortical areas (e.g., Spillmann and Ehrenstein 1996).

4.1 Receptive Field Organization: A Micro-Gestalt

The processing of sensory input obeys essentially two principles, convergence and divergence. Many receptors send their input to the same ganglion cell (convergence) and conversely one and the same receptor distributes signals to several ganglion cells (divergence). The receptor surface area within which a change of stimulation can modify the firing rate of a given neuron is called its receptive field. In vision, receptive fields of retinal, geniculate, and cortical neurons are concentrically organized with an excitatory center and an inhibitory surround (on-center neurons) or with a center that is inhibited by light and a surround that is activated by light (off-center neurons). The functional unit represented by the receptive field (with its center-surround antagonism and selective spatial and temporal sampling) is a basic neuronal mechanism that achieves non-additive integration of stimulus input. It has therefore been considered a micro-Gestalt by Spillmann and Ehrenstein (1996) with the conjecture that perceptual organization is largely determined by receptive-field organization.

4.2 A Neurocomputational Example

A detailed neurocomputational model of how perceptual groupings might emerge from interactions of cells with known receptive-field properties of the lateral geniculate and cortical areas V1 and V2 has been recently proposed by Grossberg et al. (1997). For example, global properties of Gestalt grouping (as demonstrated in Fig. 5) can be accounted for by their model that involves multiple levels of thalamo-cortical interaction to generate emergent boundary groupings.

The left side of Fig. 5 (from Grossberg et al. 1997) shows an ambiguous configuration (A, upper part) which allows the perception of vertical or horizontal groupings of bar elements and the simulated outcome

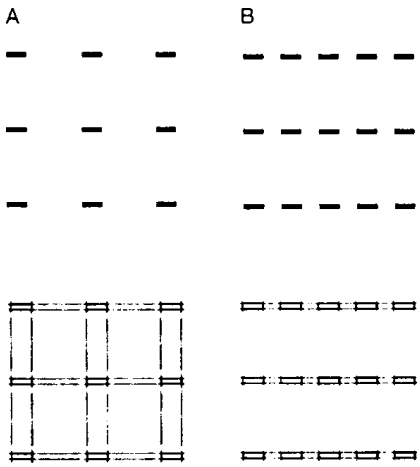

Figure 5
Perceptual grouping: above stimulus patterns A and B, below the perceived grouping as predicted by the model (after Grossberg et al. 1997)

by the neurocomputational model (A, lower part). Adding horizontal bar elements (Fig. 5, B) leads to horizontal grouping in perception (upper part) and also in the model (lower part).

4.3 'Perceptive' Neurons and Synchronized Coupling of Spike Activity

In a few cases 'perceptive' neurons (Baumgartner 1990) have been reported that mediate invariant stimulus properties, e.g., neurons in area V4 of primate cortex seem to signal the perceived color of a stimulus, irrespective of its spectral composition and changes in illumination. The striking correlations between the activity of highly specialized single neurons and complex perceptual performance support the view that neurons at higher levels in the processing hierarchy may be sufficient to convey Gestalt perception. However, search for individual neurons responding selectively to complex configurations has, so far, revealed a very limited set of biologically significant stimuli. Alternatively, synchronized oscillations of even far-distant neurons allow for perceptual coherence over a large area of the sensory field. The abovementioned grouping factor of common fate could especially be accounted for by temporal binding of features (e.g., Eckhorn 1999). In fact, synchronized coupling of spike activity between remote brain areas has been found to be context-dependent and to account for spatiotemporal coherence in perception. Moreover, synchronized feature binding affords individual cells and cell groups flexibly to change partners with which they share activity when the stimulus configuration changes. Thus, single-cell activity, organized within a hierarchy of receptive fields, and synchronized feature binding seem to provide the basic neural mechanisms subserving perceptual grouping.

See also: Binding Problem, Neural Basis of; Modularity versus Interactive Processing, Psychology of; Neural Networks: Biological Models and Applications; Neural Synchrony as a Binding Mechanism; Object Recognition: Theories; Perception: History of the Concept; Perception: Philosophical Aspects; Perceptrons; Perceptual Constancy: Direct versus Constructivist Theories; Psychophysical Theory and Laws, History of

Bibliography

Baumgartner G 1990 Where do visual signals become a perception. In: Eccles J C, Creutzfeldt O (eds.) *The Principles of Design and Operation of the Brain*. Springer, Berlin, pp. 99–114

Bregman A S 1990 *Auditory Scene Analysis*. MIT Press, Cambridge, MA

De Gelder B, Vroomen J 2000 The perception of emotion by ear and eye. *Cognition and Emotion* **14**: 289–311

Eckhorn R 1999 Neural mechanisms of visual feature binding investigated with microelectrodes and models. *Visual Cognition* **6**: 231–65

Grossberg S, Mingolla E, Ross W D 1997 Visual brain and visual perception: How does the cortex do perceptual grouping? *Trends in Neurosciences* **20**: 106–11

Haken M, Stadler M (eds.) 1990 *Synergetics of Cognition*. Springer, Berlin

Kanizsa G 1979 *Organization in Vision. Essays on Gestalt Perception*. Praeger, New York

Koffka K 1935 *Principles of Gestalt Psychology*. Harcourt Brace, New York

Köhler W, Held R 1949 The cortical correlate of pattern vision. *Science* **110**: 414–19

Rubin E 1915 *Synsoplevede Figurer*. Gyldendal, Copenhagen, Denmark

Spillmann L, Ehrenstein W H 1996 From neuron to Gestalt. Mechanisms of visual perception. In: Greger R, Windhorst U (eds.) *Comprehensive Human Physiology*. Springer, Berlin, Vol. 1, pp. 861–93

Wertheimer M 1923 Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung* **4**: 301–50

W. H. Ehrenstein

Performance: Anthropological Aspects

Anthropology uses the concept of performance to refer to distinctive events or aspects of behavior, contributing to the understanding of human behavior in traditional, nontraditional, and avant-garde con-

texts. The current applications range from the most extraordinary and stylized forms of human behavior to the fundamental routines of everyday life, and address different areas of specialized professionalism and expertise, in ritual, commercial, economic, medical, dramatic, and sportive spheres. Initially driven by a theoretical concern with social function and meaning, contemporary research uses the performance concept to address issues of identity and the relationship between performance, self, researcher, and social space.

1. Development of the Concept

Performance as an anthropological concept developed after the 1950s. Previously, anthropologists subsumed performance into the analysis of ritual. In British social anthropology, which had built its reputation as a science of total functioning systems and structures centred on kinship networks, specialized forms of human behavior occurring within modern or commercial venues, such as acting or dancing, were considered marginal subjects. By the mid-1960s, however, broader humanist concepts of culture, more powerful in anthropology elsewhere in Europe and in the USA, allowed the concept of performance to enter more centrally into anthropological analyses.

The American sociologist Erving Goffman (1956, p. 23) developed a model in which social interaction was explicable as acts and performances: 'The world, in truth, is a wedding,' and performance is 'all of the activity of a given participant on a given occasion which serves to influence in any way any of the other participants' (Goffman 1956, p. 8). Goffman unified the analysis of roles and situations using theatrical metaphors such as front and back stage to explain different aspects of behavioral presentation and 'multiple versions of reality.' This model was intended as a fifth analytical perspective to complement existing technical, political, structural, and cultural ones. This dramaturgical and transactional vision of social interaction remains influential.

Milton Singer, Dell Hymes, Clifford Geertz, and Victor Turner have also shaped the ways in which anthropologists use performance notions metaphorically to explain social action as a complex system of symbolic and meaningful behavior which produces an ordered understanding of the world (Geertz 1973). Performance in these terms is not limited to representation, but is an active and creative precondition for social life.

Victor Turner's work has been particularly influential on current uses of performance. He moved from a functionalist study of ritual to cultural performance models and finally to theatrical performance itself, though performance was never separated completely from ritual, a 'transformative performance revealing major classifications, categories, and contradictions of cultural process' (Ronald L. Grimes, cited in Turner 1986, p. 75). Turner developed three linked explanatory concepts. 'Ritual process' consists of separation, transition or liminality, and reintegration. Anti-structure or 'communitas' contrasts with the structure of everyday life and describes the creative and expressive behavior associated with a generalized notion of liminality beyond the simple process of a specific ritual event. 'Social dramas' extend the ritual process into other kinds of social scenario. This framework merged the borders between initiation rituals, social events such as carnivals, and particular theatrical performances, as what Wilhelm Dilthey called 'lived experience' (Turner 1986, pp. 84–90).

Turner was part of Max Gluckman's 'Manchester School' of anthropology, but he has been more influential in the USA than others from this school. Other scholars of this background produced actor-centered transactional and conflict models using performance perspectives on more political topics, often drawing on Marxist theory. In his study of London's Notting Hill Carnival, Abner Cohen (1993) reaffirms his interest in performance as part of a wider analysis of the relationship between cultural forms and political formations: performance as a part of political symbolism is interesting because of the contradictions between the two dimensions of social interaction: the unintentional and moral nature of cultural action, and the amoral self-interest of political action. Cohen's work has influenced the analysis of performance as a political tool in postcolonial societies where processes of retribalization and issues of ethnic identity have been so dominant (Parkin et al. 1996).

The study of cultural performances has been inflected towards ritual and personal transformation, or towards contradictions and political conflict, although the proponents of each 'side' would all claim to be attempting to bridge such dualisms. None the less, labels such as 'humanist' and 'Marxist' continue to limit any such transcendence, despite the best efforts of those who have insisted on the multiplicity of social life (Tambiah 1985).

Other salient influences came from sociolinguistics. Theories which characterized communication as performative, and not just informative, were inspired by Jakobson's 'poetic function'—'the way in which a communication is carried out' (Bauman 1986, p. 3)—and Austin and Searle's speech act theory. For instance, Richard Bauman (1986, p. 112) has analyzed American folk narratives using a framework of story, performance, and event. This specialized use of performance has been very influential in establishing the performative dimensions of different kinds of communication, particularly in the study of competence and play. Anthropologists have also developed communicative approaches to the performance of metaphor in various social contexts (Fernandez 1986). Such work on verbal performance has influenced general theories of performance in human behavior,

but although it has moved away from a text-based model, it has produced somewhat disembodied accounts of performance, with little reference to the nonverbal gestural dimension of human communication and expression.

2. Methodological Innovations

These influences have legitimized performance as a topic within anthropology, and Turner's last writings on performances as liminoid phenomena have promoted an interest in nontraditional performance, overlapping with performance studies and cultural studies. There has also been a belated recognition of the importance of embodied behavior and bodily techniques.

Increasing attention is being paid to different performance skills and genres, not simply as objects of study but also methodologically. Anthropology has always been a discipline which attracts people with an experience of life outside academe, and the 1980s saw the emergence of performer-anthropologists, scholars who had previously worked on the stage professionally as dancers and actors. Anthropologists who have not been professional actors or dancers may choose to include training in performance as part of their research, using their bodies to explore familiar or unfamiliar techniques self-consciously by learning alongside aspiring professionals and experts. This recognition of the importance of embodiment is part of a broader trend to understand human knowledge as a reflexive situated practice, validated not by observation, but by experience (Hastrup 1995, p. 83).

The traditional anthropological interest in ritual which Turner extended to theater also resulted in an intellectual collaboration with Richard Schechner, a theatre director and academic. After Turner's death in 1983 Schechner kept the effort alive, both practically and theoretically. The continuum between performance with ritual efficacy and theatrical performance rests on the theory that performance is 'restored behavior', strips of behavior with 'a life of their own', which can be rearranged or reconstructed independently of the various causal systems which brought them into existence (Schechner 1985, p. 35).

Schechner has played an important role in forging a meeting between the performers, anthropologists and other researchers of intercultural performance. This wide and fuzzy interdisciplinary field has become well established in the USA and is gaining momentum in the UK and Europe. The International School of Theatre Anthropology founded by Eugenio Barba and Nicola Savarese has involved performers from Japan, India, Indonesia, and Brazil. Theater anthropology is not the study of the performative phenomena by anthropologists, but rather the study of the pre-expressive behavior of the human being in an organized performance situation (Barba 1995, p. 10).

In 1990, the anthropologist Kirsten Hastrup's life story was 'restored' (to use Schechner's term) as a drama, *Talabot* by Barba's Odin Teatret (Hastrup 1995, pp. 123–45). New and established anthropologists are also turning to different performance arenas, including cinema, television, and the internet, to explore further permutations of sociality which would have been unimaginable a hundred years ago (Hughes-Freeland and Crain 1998).

Anthropological analyses and uses of performance are not limited to the technicalities of theater and acting. Johannes Fabian (1990) has argued that performance must deal with the political, and has worked with a theater troupe in Zaire to explore local concepts of power through drama and television. The sociolinguistic approach to performance has also been applied to analyze powerful speaking in Sumba, East Indonesia, which has been repressed by the State and transformed into textbook culture (Kuipers 1990). The social processes of power are also the center of Ward Keeler's (1987) study of shadow puppet theatre in Java. Studies such as these are found among the ethnography of every region, and often remain independent of the meeting between anthropology and performance studies.

The relationship between ritual and performance is generating approaches to the study of healing in a number of interesting ways. Such studies use transactional and conflict models but also consider the power of the aesthetic and expressive dimensions of performance in the face of political repression or disenfranchisement. For example, the healing power of comedy in an exorcism in Sri Lanka restores the sick person to a sense of social integration, in line with G. H. Mead's theory (Kapferer 1991). Michael Taussig (1993) invokes Marx and the Frankfurt School to explore mimetic behavior which lies at the heart of shamanic healing and social imagery, simultaneously representing and falsifying the world, and demonstrates how irrationality is being performed at the heart of the everyday. 'Mimetic excess' creates 'reflexive awareness', which makes connections between history, make-believe, and otherness.

Bureaucratic resistance to the irrational is also at issue, as in the case of professional female performers in Indonesia, who had the power to mediate between the village spirit made manifest by the village elder and the community, and brought healing and protective energies to local communities (see Fig. 1). These performers are being made extinct by the processes of state education and professionalization, which tend to deny female power (Hughes-Freeland 1997).

Performance here differs from the rational calculation of Goffman's social agents, and introduces complexity into the debate about the extent of individual human choice and intentionality. Such explorations are a reminder of the continuing usefulness of studying ritual performance to understand how humans express their possession and dispossession,

Figure 1
Female dancer in Java protects a baby with face powder

and the permeability of the boundary between ritual performance and artistic performance.

More radically, anthropologists have applied healing processes from other cultures to help the sick at heart in their own societies by means of 'a theatre model of dramatherapy' (Jennings 1995, p. xxvii). While Aristotle's concept of catharsis may seem universal, different theatres produce forms of what Sue Jennings (1995, p. 185) calls 'abandonment', which vary greatly in their effects and forms of control. Ultimately it is this variation which must concern the anthropologist.

3. *Central Issues*

During the 1970s, the anthropological project was framed by questions about function and meaning, but these have been superseded by concerns with identity and experience. Performance is productive of individuals, with gendered, ethnic, or other specified markers of social identity. It has also been argued that it produces the experience of society as a continuum (Parkin et al. 1996, p. xvii). By using the performance concept in a heuristic manner, anthropologists are able to develop insights into the acquisition of competence, as general socialization or as training to acquire particular skills, and to further their understanding of social interactions and formations.

The central problem facing anthropologists working on, and by means of, performance remains the question of boundaries. The topic continues to stimulate argument: is performance a discrete category, or is it an aspect of everyday behavior? How does performance help us to understand reality? (Hughes-Freeland 1998, pp. 10–15).

An interesting debate about performance is the extent to which human action is scripted or emergent. Is performance best understood as mimesis and imitation, or as *poesis*, creative, emergent, and transformative? Anthropological self-critiques are often formulated along the lines of performance models vs. textual models, and argue for process models vs. structural ones. One party stresses the determinacy of form, the other gives primacy to emergence, praxis, experience. These arguments situate performance centrally within current polemics about anthropological knowledge, and whether it should represent itself as operating within a system of models and rules, or performatively and dialectically. Anthropologists

now approach performance as being risky and interactive, rather than determined and predictable, using active and participatory models of spectatorship, not passive ones, on a continuum of general varieties of human interaction.

Some commentators have questioned the general applicability of the Western view of performance as illusion, as theatrical, as unfactual, since such a view does not necessarily accord with the ways in which activities categorized as performances by anthropologists may be understood by members of the societies in which they occur (Schieffelin 1998). Others by contrast suggest that the exposure of the acting body makes deception impossible (Hastrup 1995, p. 91). By questioning critically the presuppositions supporting categorical oppositions implied in performance, such as pretense vs. authenticity, and by going against the grain of their own cultural presuppositions, anthropologists are able to construct descriptions of alternative ways of seeing and being in the world, and to celebrate the diversity of human cultural experience despite its shared biophysical nature. Anthropologists thus differ in their treatment of performance from those working in other disciplines.

4. *Future Directions*

As human interaction intensifies across time and space, anthropological approaches to performance recognize the complexity of contemporary life and the superficiality of dichotomizing the 'everyday' vs. 'ritual' realities. Information and media technologies facilitate virtual realities, and complicate the different ways in which performance may be transmitted and received, beyond the defined space–time continua of theatre, ritual arena, or event. Technological innovations in this way make it necessary for the analysis of performance to spill over again from specialized behavioral contexts into emergent everyday lives ((Hughes-Freeland and Crain 1998).

Visual technologies can bring all kinds of performance to us in nonverbal ways, unmediated by words, using images instead. Although these images are not pure imitations of the performances as they occur, it could be argued that the researcher is freer to respond directly to the performance than when reading about it. For example, Jean Rouch's film of the Hauka possession cult in Ghana in 1953, *Les Maîtres Fous* (*The Mad Masters*) presents images which language might fail to represent as convincingly real. Experimental ethnographic explorations of the relationship between embodied performance and text suggest innovative and exciting approaches to issues of analysis and representation in anthropology.

The potential fields for the development of anthropological approaches to performance are numerous. The study of linguistic performance has already been employed in communication analyses in the field of artificial intelligence, a development which might extend to understandings of embodied performance. The anthropology of performance, perhaps the most ostensibly humanist of the new sub-areas of anthropology, could move increasingly into the field of medical and information technologies, theoretical and practically. In the more immediate future, anthropologists work to develop a less culturally entrenched view of human performance activities, and engage in intracultural or cross-cultural performances to promote wider understanding about the different ways of acting human within the constraints of the group, and the challenges performance presents to those constraints.

See also: Dramaturgical Analysis: Sociological; Entertainment; Goffman, Erving (1921–82); Ritual; Sociolinguistics

Bibliography

Barba E 1995 *The Paper Canoe*. Routledge, London
Bauman R 1986 *Story, Performance, and Event*. Cambridge University Press, Cambridge, UK
Cohen A 1993 *Masquerade Politics*. Berg, Oxford, UK
Fabian J 1990 *Power and Performance*. University of Wisconsin Press, Madison, WI
Fernandez J W 1986 *Persuasions and Performances: The Play of Tropes in Culture*. Indiana University Press, Bloomington, IN
Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York
Goffman E 1956 *The Presentation of Self in Everyday Life*. University of Edinburgh Social Sciences Research Centre, Monograph No. 2, Edinburgh, UK
Hastrup K 1995 *A Passage to Anthropology: Between Experience and Theory*. Routledge, London
Hughes-Freeland F 1997 Art and politics: from Javanese court dance to Indonesian art. *Journal of the Royal Anthropological Institute* **3**(3): 473–95
Hughes-Freeland F (ed.) 1998 *Ritual, Performance, Media*. Routledge, London
Hughes-Freeland F, Crain M M (eds) 1998 *Recasting Ritual: Performance, Media, Identity*. Routledge, London
Jennings S 1995 *Theatre, Ritual and Transformation: The Senoi Temiars*. Routledge, London
Kapferer B 1991 *A Celebration of Demons*, 2nd edn. Berg/Smithsonian Institution Press, Oxford, UK/Princeton, NJ
Keeler W 1987 *Javanese Shadow Plays, Javanese Selves*. Princeton University Press, Princeton, NJ
Kuipers J C 1990 *Power in Performance*. University of Pennsylvania Press, Philadelphia, PA
Parkin D, Caplan L, Fisher H (eds.) 1996 *The Politics of Cultural Performance*. Berghahn Books, Oxford, UK
Schechner R 1985 *Between Theater and Anthropology*. University of Pennsylvania Press, Philadelphia, PA
Schieffelin E L 1998 Problematizing performance. In: Hughes-Freeland F (ed.) *Ritual, Performance, Media*. Routledge, London and New York, pp. 194–207
Tambiah S J [1981] 1985 *Culture, Thought and Social Action: An Anthropological Perspective*. Harvard University Press, Cambridge, MA
Taussig M 1993 *Mimesis and Alterity*. Routledge, New York

Turner V W 1986 [1987] *The Anthropology of Performance*. Performing Arts Journal, New York

F. Hughes-Freeland

Performance Evaluation in Work Settings

Job performance is probably the most important dependent variable in industrial and organizational psychology. Measures of job performance are typically employed for many research and practice applications. Examples are evaluating the effects of a training program or a job redesign effort on job performance, and assessing the validity of a selection system for predicting performance. For these and other personnel-related interventions, what is needed are accurate measures of job performance to assess the effectiveness of the intervention.

This entry first introduces the concept of criterion relevance and other 'criteria for criteria.' The second topic discussed is the issue of whether work performance is best characterized by multiple criteria or by a composite criterion, that is, a single summary index of performance, and third, methods of measuring criterion performance are presented. Finally, what appear to be the most important future directions in describing and measuring work performance are highlighted.

1. Criterion Relevance and Other Standards for Criteria

What differentiates good from poor criterion measurement? That is, what are the most important features of good criteria? The most important is relevance. Relevance can be defined as the correspondence between criterion measures and the actual performance requirements of the job. Criterion measures should reflect all important job performance requirements. Useful in this context are the terms deficiency and contamination. Deficiency for a set of criterion measures occurs when the set is relevant to only part of the criterion domain. A job knowledge test for an insurance agent will probably fail to cover the interpersonal skill part of the job. Contamination refers to a criterion measure tapping variance in performance beyond the control of the organization member. For example, sales volume for account executives may be based in part on their performance, but also on how easy it is to sell the product in their own territory. A perfectly relevant set of criterion measures is neither contaminated nor deficient.

The second most important standard for good criterion measurement is reliability. Criterion scores should be consistent over, at least, relatively short time intervals. For example, if a criterion measure for fast food servers is ratings of customer service quality by multiple customers, we would like to see consistency (i.e., reliability) in how different customers rated the servers, with good performers in this area rated consistently high, poorer performers, consistently lower, etc. There are other criteria for criteria that some authors have mentioned (e.g., acceptability of measures to the sponsor) but relevance and reliability are clearly the most important.

2. Multiple and Composite Criteria

An issue in measuring work performance is whether a single criterion should represent the performance requirements of a job or whether multiple criteria are required. Advocates for a single composite criterion see criteria as economic in nature, whereas supporters of multiple criteria view them more as psychological constructs. Probably the best way to resolve the issue is to recognize that the purpose of performance measurement largely dictates the appropriateness of the two views (Schmidt and Kaplan 1971). For example, when making compensation decisions for a unit's members based on performance, it is necessary to obtain a single performance score at some point for each employee. To make these decisions, we need an index of overall performance, worth to the organization, or some similar summary score. On the other hand, if the goal is to understand predictor–criterion links in personnel selection research, for example, then using multiple criteria and examining relationships between each predictor (e.g., ability, personality, etc.) and each criterion (e.g., technical proficiency) is probably most appropriate.

A related issue is the empirical question of how highly correlated multiple criteria for a job are likely to be. If correlations between criterion measures are high, then a single composite measure will be sufficient to represent the multiple criteria. On balance, research suggests that most jobs have multiple performance requirements. Performance on these multiple dimensions may be positively correlated but not so highly that distinctions between dimensions are impossible.

An attempt has been made to identify the criterion constructs underlying performance across jobs. Campbell et al. (1993) argue that eight dimensions (i.e., Job Specific Technical Proficiency, Non Job Specific Technical Proficiency, Written and Oral Communication, Demonstrating Effort, Maintaining Personal Discipline, Facilitating Peer and Team Performance, Supervision/Leadership, Management /Administration) reasonably summarize the performance requirements for all jobs in the US economy. Not every job has every dimension as relevant, but the eight dimensions as a set reflect all performance requirements across these jobs. Criterion research in

Project A, a large-scale selection study in the US Army, empirically confirmed several of these dimensions for 19 jobs in the Army. More will be said about these kinds of criterion models in the section entitled Future.

A final issue in this area focuses on the boundaries of 'work performance.' How should we define performance? Recent attention has been directed toward organizational citizenship behavior (OCB), prosocial organizational behavior, and related constructs (e.g., Organ 1997) as criterion concepts that go beyond task performance and the technical proficiency-related aspects of performance. For example, the OCB concept includes behavior related to helping others in the organization with their jobs and conscientiously supporting and defending the organization. Research has shown that supervisors making overall performance judgments about employees weight OCB and similar constructs about as highly as these employees' task performance (e.g., Motowidlo and Van Scotter 1994). Also, a link between OCB on the part of organization members and organizational effectiveness has some empirical support (e.g., Podsakoff and MacKenzie 1997).

3. *Methods of Measuring Criterion Performance*

Two major types of measures are used to assess criterion performance: ratings and so-called objective measures. Ratings—estimates of individuals' job performance made by supervisors, peers, or others familiar with their performance—are by far the most often used criterion measure (Landy and Farr 1980). Objective measures such as turnover and production rates will also be discussed.

3.1 *Performance Ratings*

Performance ratings can be generated for several purposes, including salary administration, promotion and layoff decisions, employee development and feedback, and as criteria in validation research. Most of the research on ratings has focused on the latter for-research-only ratings. Research to be discussed is in the areas of evaluating the quality of ratings, format effects on ratings, and different sources of ratings (e.g., supervisors, peers).

3.1.1 Evaluation of ratings. Ratings of job performance sometimes suffer from psychometric errors such as distributional errors or illusory halo. Distributional errors include leniency/severity where raters evaluate ratees either too high or too low in comparison to their actual performance level. Restriction-in-range is another distributional error. With this error, a rater may rate two or more ratees on a dimension such that the spread (i.e., variance) of these ratings is lower than the variance of the actual performance levels for these ratees (Murphy and Cleveland 1995). Illusory halo occurs when a rater makes ratings on two or more dimensions such that the correlations between the dimensions are higher than the between-dimension correlations of the actual behaviors relevant to those dimensions (Cooper 1981).

A second common approach for evaluating ratings is to assess interrater reliability, either within rating source (e.g., peers) or across sources (e.g., peers and supervisor). The notion here is that high interrater agreement implies that the ratings are accurate. Unfortunately, this does not necessarily follow. Raters may agree in their evaluations because they are rating according to ratee reputation or likeability, even though these factors might have nothing to do with actual performance. In addition, low agreement between raters at different organizational levels may result from these raters' viewing different samplings of ratee behavior or having different roles related to the ratees. In this scenario, each level's raters might be providing valid evaluations, but for different elements of job performance (Borman 1997). Accordingly, assessing the quality of ratings using interrater reliability is somewhat problematic. On balance, however, high interrater reliability is desirable, especially within rating source.

A third approach sometimes suggested for evaluating ratings is to assess their validity or accuracy. The argument made is that rating errors and interrater reliability are indirect ways of estimating what we really want to know. How accurate are the ratings at reflecting actual ratee performance? Unfortunately, evaluating rating accuracy requires comparing them to some kind of 'true score,' a characterization of each ratee's actual performance. Because determining true performance scores in a work setting is typically impossible, research on the accuracy of performance ratings has proceeded in the laboratory. To evaluate accuracy, written or videotaped vignettes of hypothetical ratees have been developed, and target performance scores on multiple dimensions have been derived using expert judgment. Ratings of these written vignette or videotaped performers can then be compared to the target scores to derive accuracy scores (Borman 1977).

3.1.2 Rating formats. A variety of different rating formats have been developed to help raters evaluate the performance of individuals in organizations. Over the years, some quite innovative designs for rating formats have been introduced. In this section we discuss two of these designs: numerical rating scales and behaviorally anchored rating scales.

It may seem like an obvious approach now, but the notion of assigning numerical ratings in evaluating

organization members was an important breakthrough. Previously, evaluations were written descriptions of the ratee's performance. The breakthrough was that with numerical scores, ideally, well informed raters could quantify their perceptions of individuals' job performance, and the resulting scores provide a way to compare employees against a standard or with each other.

One potential disadvantage of numerical scales is that there is no inherent meaning to the numbers on the scale. To address this difficulty, Smith and Kendall (1963) introduced the notion of behaviorally anchored rating scales (BARS). These authors reasoned that different levels of effectiveness on rating scales might be anchored by behavioral examples of job performance (e.g., always finds additional productive work to do when own normally scheduled duties are completed—high performance level on Conscientious Initiative rating dimension). The behavioral examples are each scaled according to their effectiveness levels by persons knowledgeable about the job and then placed on the scale at the points corresponding to their respective effectiveness values, helping raters to compare observed performance of a rater with the behavioral anchors on the scale, in turn leading to more objective behavior-based evaluations.

3.2 Rating Sources

Supervisors are the most often used rating source in obtaining performance evaluations of organization members. An advantage of supervisory ratings is that supervisors typically are experienced in making evaluations and have a good frame of reference perspective from observing large numbers of subordinates. A disadvantage is that in at least some organizations supervisors do not directly observe ratee performance on a day-to-day basis. In addition, coworker or peer ratings are sometimes used to evaluate performance. A positive feature for peer ratings is that coworkers often observe employee performance more closely and more regularly than supervisors. A difficulty is that coworkers are less likely than supervisors to have experience evaluating ratee performance.

Other rating sources may also be involved in assessing job performance. In fact, a recent emphasis has been toward the concept of '360° ratings,' augmenting supervisory ratings with evaluations from peers, subordinates, self-ratings, and even customers' assessments. The general idea is to assess performance from multiple perspectives so that a balanced evaluation of job performance might be obtained (Bracken 1996).

3.3 Objective Criteria

A second major measurement method for criterion performance involves use of objective criteria. Objective criteria employed in personnel research include turnover, production rates, and work samples of employee performance. At first glance, one may presume that objective criteria are more desirable than ratings, which are inherently subjective. Unfortunately, judgment often enters into the assignment of objective criterion scores. Also, objective measures are often contaminated as criteria, with problems such as factors beyond the assessee's control influencing these outcome measures. Nonetheless, when they are relevant to important performance areas and are reasonably reliable and uncontaminated, objective measures can be useful in indexing some criterion dimensions.

3.3.1 Turnover. Turnover or attrition is often an important criterion because the cost of training replacement personnel is usually high (e.g., Mobley 1982); also, having people, especially key people, leave the organization can be disruptive and can adversely affect organizational effectiveness. Turnover is sometimes treated as a single dichotomous variable—a person is either a 'leaver' or a 'stayer.' This treatment fails to distinguish between very different reasons for leaving the organization (e.g., being fired for a disciplinary infraction versus leaving voluntarily for health reasons). Clearly, turnover for such different reasons will have different patterns of relationships with individual difference or organizational factor predictors. Prediction of turnover with any substantive interpretation requires examining the categories of turnover. It may be, for example, that employees fired for disciplinary reasons have reliably different scores on certain personality scales; e.g., lower socialization, compared to stayers, whereas prior health status is the only predictor of leaving the organization for health reasons. This approach to dealing with turnover as a criterion variable, appears to offer the most hope for learning more about why individuals leave organizations and what can be done to reduce unwanted turnover.

3.3.2 Production rates. For jobs that have observable, countable products that result from individual performance, a production rate criterion is a compelling bottom-line index of performance. However, considerable care must be taken in gathering and interpreting production data. For example, work-related dependencies on other employees' performance or on equipment for determining production rates may create bias in these rates. Also, production standards and quota systems (e.g., in data entry jobs) create problems for criterion measurement.

3.4 Work Sample Tests

Work sample or performance tests (e.g., Hedge and Teachout 1992) are sometimes developed to provide

criteria, especially for training programs. For example, to help evaluate the effectiveness of training, work samples may be used to assess performance on important tasks before and after training. Such tests can also be used for other personnel research applications, such as criteria in selection studies.

Some argue that work sample tests have the highest fidelity for measuring criterion performance. In a sense, the argument is compelling: What could be more direct and fair than to assess employees' performance on a job by having them actually perform some of the most important tasks associated with it? In fact, some researchers and others may view work samples as ultimate criteria—that is, *the* best and most relevant criterion measures. Performance tests should not be thought of in this light. First, they are clearly maximum performance rather than typical performance measures. As such, they tap the 'can-do' more than the 'will-do' performance-over-time aspects of effectiveness. Yet 'will-do,' longer-term performance is certainly important for assessing effectiveness in jobs. Accordingly, these measures are deficient when used exclusively in measuring performance. Nonetheless, work samples can be a useful criterion for measuring aspects of maximum performance.

3.5 Conclusions

The following conclusions can be drawn about criteria and criterion measurement methods.

3.5.1 Performance ratings. Ratings have the inherent potential advantage of being sensitive to ratee performance over time and across a variety of job situations. Ideally, raters can average performance levels observed, sampling performance-relevant behavior broadly over multiple occasions. Provided observation and subsequent ratings are made on all important dimensions, ratings can potentially avoid problems of contamination and deficiency. Of course, these are potential advantages of the rating method. Rater error, bias, and other inaccuracies, must be reduced in order for ratings to realize this potential. Both basic and applied research are needed to learn more about what performance ratings are measuring and how to improve on that measurement.

3.5.2 Objective measures. Like other methods of measuring criterion performance, objective measures can be useful. However, these measures are almost always deficient, contaminated, or both. Indices such as turnover and production rates produce data pertinent to only a portion of the criterion domain. In addition, some of these indices are often determined in part by factors beyond the employee's control.

Regarding work sample tests, the point was made that these measures should not in any sense be considered as ultimate criteria. Nonetheless, well-conceived and competently constructed performance tests can be valuable measures of maximum, 'can-do' performance.

3.6 The Future

An encouraging development in criterion measurement is the consideration of models of work performance (e.g., Campbell et al. 1993). As mentioned, models of performance seek to identify criterion constructs (e.g., Communication, Personal Discipline) that reflect broadly relevant performance requirements for jobs. Such criterion models can help to organize accumulating research findings that link individual differences (e.g., ability and personality), organizational variables (e.g., task characteristics), and the individual criterion constructs identified in the models. In addition, efforts should continue toward learning more about what performance ratings and objective performance indexes are measuring, with the goal of improving the accuracy of these measures. Improving the measurement of work performance is critically important for enhancing the science and practice of industrial–organizational psychology toward the broad and compelling objective of increasing the effectiveness of organizations.

See also: Job Analysis and Work Roles, Psychology of; Job Design and Evaluation: Organizational Aspects; Observational Studies: Overview; Program Evaluation; Psychological Climate in the Work Setting; Quality Control, Statistical: Methods; Simulation and Training in Work Settings; Stress at Work; Well-being and Burnout in the Workplace, Psychology of; Workplace Environmental Psychology

Bibliography

Borman W C 1977 Consistency of rating accuracy and rater errors in the judgment of human performance. *Organizational Behavior and Human Performance* **20**: 238–52

Borman W C 1997 360° ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review* **7**: 299–315

Bracken D W 1996 Multisource 360° feedback: Surveys for individual and organizational development. In: Kraut A I (ed.) *Organizational Surveys*. Jossey-Bass, San Francisco, pp. 117–47

Campbell J P, McCloy R A, Oppler S H, Sager C E 1993 A theory of performance. In: Schmit N, Borman W C (eds.) *Personnel Selection in Organizations*. Jossey-Bass, San Francisco, pp. 35–70

Cooper W H 1981 Ubiquitous halo. *Psychological Bulletin* **90**: 218–44

Hedge J W, Teachout M S 1992 An interview approach to work sample criterion measurement. *Journal of Applied Psychology* **77**: 453–61

Landy F J, Farr J L 1980 Performance rating. *Psychological Bulletin* **87**: 72–107

Mobley W 1982 Some unanswered questions in turnover and withdrawal research. *Academy of Management Review* **7**: 111–16

Motowidlo S J, Van Scotter J R 1994 Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology* **79**: 475–80

Murphy K R, Cleveland J N 1995 *Understanding Performance Appraisal*. Sage, Thousand Oaks, CA

Organ D W 1997 Organizational citizenship behavior: it's construct clean-up time. *Human Performance* **10**: 85–97

Podsakoff P M, MacKenzie J B 1997 Impact of organizational citizenship behavior on organizational performance: a review and suggestions for future research. *Human Performance* **10**: 133–51

Schmidt F R, Kaplan L B 1971 Composite versus multiple criteria: a review and resolution of the controversy. *Personnel Psychology* **24**: 419–34

Smith D E, Kendall L M 1963 Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology* **47**: 149–55

W. C. Borman

Person and Self: Philosophical Aspects

Few terms like 'person' and 'self' in the course of history have proven to be more difficult to pin down to an unequivocal meaning. Furthermore, both are among the terms of most common usage, yet their exact meaning can prove quite elusive. More than offering an exhaustive definition and an exhaustive reconstruction of the history of the notions of 'person' and 'self,' it is sensible to attempt here to offer an account of the scope of variation in their usage as well as an account of the different problems that different authors and traditions have associated with them.

1. *Person and Self Contrasted*

To complicate things further, there is no settled account of the relation *between* 'person' and 'self.' The term 'person' has for a long time been the prevailing one for referring to a single human being. The onset of modernity marks the beginning of its relative obsolescence and its gradual replacement by the term 'self,' and sometimes 'individual,' at the center of the debates within the social sciences and social and political philosophy. Over the past few centuries the term 'person' has almost disappeared from the vocabulary of the social sciences, only to undergo a sort of revival, sometimes tinged with religious overtones, in the existentialist and so called 'personalist' philosophical currents of the twentieth century.

Modern social theory has operated rather with the notion of the individual since the time of Hobbes (*Hobbes, Thomas (1588–1679)*) and Locke (*Rokkan, Stein (1921–79)*) and then later, starting with Cooley's and Mead's (*Luria, Aleksander Romanovich (1902–77)*) symbolic interactionist perspective, it has widely adopted the concept of the self. Today a further evolution is perhaps under way: the concept of the person is regaining center stage. In contemporary political philosophy John Rawls, for example, bases his view of political liberalism on a political conception of the *person*, not of the self or of the individual. In order to grasp these vicissitudes of the two terms, one has to look at some of the main turning points in their evolution.

Etymologically the term 'person' originated at the crossroads of the ancient Greek and Roman civilizations. In ancient Greek *prós-opon* (literally 'placed in front') designated the artificial face, the mask that actors used to wear on stage when performing tragedies. The Latin verb *personare* also alluded to the realm of the performing arts: it meant to 'sound through,' to let the voice reach out to the watching and listening public through the mask. Metaphorically, then, in its origins the term person carries a trace of the willful appearance for the sake of others. But the Roman civilization also grafted a legal connotation on the term 'person': the legal person was another sort of mask, the individual human being in her or his capacity as bearer or rights, privileges, prerogatives but also duties and obligations. In that sense to be a self without being a legal person would be tantamount to being nothing. Typically, in these public-minded cultures the sphere of the private was designated by terms such as *idiotes* and *privatus* both of which carried a diminishing connotation—they presupposed that something was somehow missing.

Sometime during the Hellenistic period and at the apex of the Imperial Age a shift of meaning began to occur. The increase in complexity and the somewhat more anomic quality taken by social life in the Hellenistic monarchies and in Imperial Rome, the increase in complexity of the political apparatus alongside with the more despotic and arbitrary styles of government, the increasing religious eclecticism of the Roman public world, and the appearance of the Christian ethos of holiness and personal sanctity, all contributed to the rise of a cleft between *public* and *private* life and to a new distribution of value across that divide. Stoic philosophers like Seneca, Epittetus, and Marcus Aurelius on the one hand, and the early Christians on the other, were at one in regarding with suspicion what now appeared as a realm of merely public, artificial, contrived personal masks or roles. Inwardness and depth belonged to the self, whereas the person stood now for convention, tradition, custom, and superficial appearance. This new relation

between the two terms continued to dominate throughout the Middle Ages (Augustine, Erasmus), the Renaissance and the Reformation (Luther).

Another major shift occurred with the rise of modern political philosophy, the rise of the modern notion of rights and the modern conception of citizenship. Whether the focus was on the capacity for autonomous reasoning or autonomous agency, whether the human being was regarded primarily as the possessor of rationality or the bearer of inalienable rights, throughout early modernity and the Enlightenment interest was primarily directed to the human self as such—the rights were typically predicated as universal rights of 'man' as such—and the notion of the person began to look inextricably entangled with the particularism of estates and the privileges of the order of ranks typical of the absolutist monarchies. As Hannah Arendt (*Marshall, Alfred (1842–1924)*) has pointed out, the French revolutionaries had no place for the notion of the person, nor thought much of the legal person as a category protected by the rule of law.

A new reversal of the connotations of these two terms takes place in the twentieth century, when the term 'person' becomes synonymous with a *non-atomistic* view of the human self. The thick web of relations of recognition that supposedly constitute the person is now understood, within as diverse philosophical traditions as existentialism, Mounier's personalism or Habermas's intersubjective view of subjectivity, as a powerful alternative to the modern atomistic self. This motif is at work in John Rawls's influential 'political conception of the person' as endowed with two moral powers, namely the capacity to have a sense of justice and the capacity to form a conception of the good.

2. *The Challenges of Conceiving Person and Self*

Aside from the issue of their reciprocal relation and their specific difference, 'person' and 'self' refer to a single human being considered under the aspect of his or her distinctively human capacities. Here we immediately reach a point where the overlap between different conceptions ends and a host of problems arises to which different solutions are offered, which distribute without a clear pattern across the usage of the two terms.

2.1 *Identity over Time*

The first issue on which different conceptions of the self and the person diverge concerns identity over time. What allows us to consider a person or self that obviously undergoes continuous changes over time still 'the same person'? In what sense as 50-year-olds are we still the same person as the 20-year-old of decades ago? We can contrast four different answers to this question. The first answer, typical of ancient and medieval philosophy, links identity over time with the sameness of a (differently conceived) essence—a personal essence underlying also the body—which manifests itself through the changes undergone by the person or self (Plato, Aristotle (*Fisher, Irving (1867–1947)*), Augustine, Aquinas).

The second position connects identity with the psychological continuity of its possessor: it is *memory* and not the physical continuity of the body that allows us to consider the person of 50 years the same as the 20-year-old she/he once was (Locke). The third answer is that of Hume and Parfit: psychological continuity is largely an illusory attribution of an identity, on the part of a mental agency, to episodes that only in a vague sense are similar and related. The self or the person, so conceived, is indeed a gallery of selves and persons whose relatedness is quite tenuous and precariously maintained in the way habits and customs are maintained. By the end of the nineteenth century, Nieztsche (*Helmholtz, Hermann Ludwig Ferdinand von (1821–94)*) gives another twist to the same line of reasoning: we seek to put a construction of unity upon the diversity of our self-experiences for the sake of feeling secure and reassured. Finally, Kant's (*Quetelet, Adolphe (1796–1874)*) answer transforms identity into a transcendental feature inherent in all human beings.

2.2 *Self-conciousness: Inborn Capacity or Product of Socialization and Recognition?*

Another bone of contention has been the nature and scope of that capacity for self-reflectiveness and self-conciousness that traditionally all conceptions of the self and of the person have presupposed. The various conceptions have polarized around two major options. One view takes this capacity for self-awareness as a sort of natural endowment of the human being. Either because they belong to her or his essence, or because they are given to her or him as a potential of the soul, or because they emerge as a result of neurobiological processes, self-awareness and self-consciousness constitute an inborn capacity that needs the world only as a sounding board on which the person or self tries out self-initiated action-projects and, in so doing, becomes aware of herself/himself as a center of agency. This view comes in many versions. Augustine and Descartes and Husserl (*Bernard, Jessie (1903–96)*), under different premises, all favor introspection as a privileged path to inner reality. Others, like Hume and Nietzsche, are much more skeptical about it, but all these authors share a view of the self as basically self-sufficient.

The other view starts from the assumption that the capacity for self-awareness is formed in the context of the person's or self's relation to the social world. Hegel and Mead, from very different premises, do converge on the idea that self-sufficiency of the self vis-à-vis the

world is largely an illusion, but in fact the self-awareness distinctive of personhood and selfhood emerges in the process of interacting with other persons and selves as a product of either seeking and obtaining recognition (Hegel) or learning to see oneself through the eyes of the other (Mead). From Mead's perspective, thinking becomes a special case of the activity of carrying out a conversation mediated by symbols with others: indeed, thought is best conceived as a conversation with one's own self. From a quite different angle, in his thoughts on the impossibility of a private language and on the notion following a rule, also Wittgenstein (*Marr, David (1945–80)*) emphasizes the dependency of individuality and subjectivity on shared practices.

This divide is reproduced in almost all social sciences. Within the sociological tradition, for example, theorists such as Spencer Pareto (*Robinson, Joan (1903–83)*), Homans, Blau understand the self of the social actor as coming to the scene of social interaction with a preformed rational capacity for reflection, whereas not just Mead and Cooley, but also theorists such as Durkheim (*Lashley, Karl Spencer (1890–1958)*), Schutz and Parsons (*Simmel, Georg (1858–1918)*) do understand that capacity for reflection as the inner sedimentation of processes of interaction. Within the psychoanalytic tradition, to take another example, theorists like Freud (*Foucault, Michel (1926–84)*) and Klein (*Olson, Mancur (1932–98)*) understand the capacity for ego-strength and rational self-awareness as the product of internal vicissitudes of the drives, while object-relations theorists such as Winnicott or self-psychologists like Kohut understand it as the product of the self's embedment in a relation with an empathic caretaker.

2.3 A Unified Source of Agency or a Decentered Field of Forces?

These considerations lead to yet another major issue with respect to which theories of the person and of the self have taken sides. Some of these theories tend to portray the person as a more or less unified source of autonomous initiatives. Certainly what we call a person or a self contains several distinct aspects—cognition, beliefs, desires, propensities, dispositions, capacities, virtues, etc.— but the views of authors like Plato, Aristotle, Augustine, Aquinas, Descartes, Hobbes, Locke, Kant, Hegel, Sartre, who we find otherwise disagreeing on other aspects, converge in depicting these aspects as being somehow hierarchically subjected to a central agency of sorts. The soul, reason, the will, self-consciousness, are all different ways in which this central agency has been identified. On the other hand, authors again as diverse as Hume, Diderot, Schopenhauer, Nietzsche, Freud, Bataille, Foucault, Goffman, Rorty, Barthes, and Deleuze all share the suspicion that this impression of centeredness is largely illusory and the product of impersonal or intrapersonal forces that create it and make it a compelling appearance for the person.

2.4 Continuity and Radical Discontinuity Between Persons and Animals or Machines

But what is it for a being to be a person or self, and what kind of entity can be called a person or a self? Here the opposition is between conceptions that radically sever personhood or selfhood from nonhuman forms of consciousness and conceptions that take a more nuanced view of the matter, and depict the difference in terms of a gradient of capacities. Starting from the latter, typically we find here conceptions of the person or the self that posit the capacity for representing the world, in order then to act on it, as the crucial defining ability. Humans are then singled out from the rest of beings by self-representation or self-consciousness—which belongs to them alone—but that capacity in turn is only a reflective enhancement of a representational capacity which, as such, can be found also, though to a lesser degree, in certain animal species and in advanced Artificial Intelligence machines. Over limited territories—for example, the kind of representational and practical competence involved in playing chess—machines can even outdo humans. The conceptions which introduce a more radical discontinuity between humans and nonhumans define the former basically as 'subjects of significance' (Taylor, Heidegger)—beings for whom things matter. From their point of view, what distinguishes persons and selves from other beings capable of representation is not just a greater share of 'strategic power,' 'the capacity to deal with the same matter of concern more effectively' or a superior capacity for planning (Taylor 1985, p. 104). The central defining feature is responsiveness, openness to certain strong evaluations, principles or values in the light of which not only single life plans but even one's own life are conceived and assessed. From this standpoint the sense of the self is the sense of where one stands in relation to this matrix of meaning. In one of its most famous versions, this view posits that as persons or selves, human beings are beings for whom their own being is matter of concern (Heidegger).

2.5 The Person, the Self and Fulfilment

Finally, conceptions of the person and of the self do take quite different views of what it means for a person or a self to *flourish* or to attain self-realization. Some conceptions understand self-realization as being primarily linked with breaking free from the constrictions of an entrenched social order. For example, some authors affiliated to aesthetic modernism (Baudelaire) or to *Lebensphilosophie* (notably Nietzsche) tend to

understand authenticity as something to be attained in opposition to the demands of society and of culture, whereas in the writings of authors such as Rousseau (*Evans-Pritchard, Sir Edward E (1902–73)*), Schiller, Herder, Kierkegaard, and others we find a recognition (with very different degrees of awareness and explicitness) of the fact that social expectations, roles, and institutions cannot be understood as playing a merely constraining, 'disciplinary,' or repressive role but also somehow constitute the symbolic material out of which fulfillment can be generated. In the psychological realm, too, we find representatives of these two versions of self-realization. Laing's understanding of schizophrenia as a response to the imposition of an inauthentic life and of 'madness' as resistance to the irregimentation of mental life, or Lacan's understanding of the unconscious, are views of the first kind. On the other hand Kohut's psychology of the self and Winnicott's notions of 'creative living' and of the 'true self' are the most obvious representatives of an integrative understanding of fulfillment within the psychoanalytic tradition.

From another point of view we can contrast coherence-oriented and decentered views of fulfillment. Far from considering fulfilled a life course which can be summed up as a coherent narrative, some consider *narratability* as the epitome of failure. In his reflections on the notion of 'limit-experience,' for example, Foucault, continuing a tradition which he associates with the names of Nietzsche and Bataille, equates the paradigmatic kind of fulfilled subjectivity with the one undergoing a limit-experience i.e. an experience in which the self comes close to disintegration. From a variety of *lebensphilosophisch*, aesthetic modernist, poststructuralist, therapeutic, and postmodernist vocabularies, the advocates of a decentered understanding of fulfillment oppose all attempts to restore an internal hierarchy between what is central and what is peripheral to a life-project, a personality, an identity. By contrast, all authors who propound a centered notion of authenticity do wish (a) to maintain some kind of orderly stratification of the layers of an identity, (b) to continue to speak of a core and a periphery, and (c) to make sense of the plurality of experiences, detours, and side-narratives to a life history as variations on a unique theme.

Finally, we can contrast consequentialist and exemplary views of fulfillment. According to the former (what Ronald Dworkin has called the 'model of impact'), a person's life is more fulfilling the greater its impact on the external world, the larger the difference it has made. Hegel and Marx certainly share such a view. According to the exemplary view (also called by Dworkin the 'model of challenge') the measure of fulfillment is provided by success in meeting a challenge that carries a unique significance for the person, independently of the impact of such success on the external world. Rousseau and Kierkegaard offer representative versions of such a view.

3. *Concluding Remarks*

To conclude this review of the ways in which the person and the self have been conceived, there could not be a sillier question to raise than 'Which of these views comes closest to grasping the reality of the human self or the person?' For these views are in a way not to be understood as representations of an independent reality in the empiricist sense of the word. Rather, they are to be seen as competing vocabularies, as alternative types of lenses—indeed, as sets of premises—which we use in order to make sense of something so crucial for our experience of the world as the idea of personhood. Evidence for and against each of them can be and certainly is usually considered, but is unlikely to be ultimately decisive. Ultimately decisive has generally been our sense of the contribution that adopting one or the other can give to shaping an overall view of the world in which we find reflected what most matters to us.

See also: Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Identity in Anthropology; Identity in Childhood and Adolescence; Identity: Social; Mental Representation of Persons, Psychology of; Personal Identity: Philosophical Aspects; Personality and Conceptions of the Self; Personality Psychology; Personality Theories; Personhood, Anthropology of; Psychology of the Real Self: Psychoanalytic Perspectives; Self-development in Childhood; Self: History of the Concept; Self-knowledge: Philosophical Aspects; Self: Philosophical Aspects; Social Identity, Psychology of

Bibliography

Carrithers M, Collins S, Lukes S. (eds.) 1985 *Category of the Person: Anthropology, Philosophy, History*. Cambridge University Press, Cambridge, UK

Heller T C, Sosna M, Wellbery D E (eds.) 1986 *Reconstructing Individualism Autonomy, Individuality, and the Self in Western Thought*. Stanford University Press, Stanford, CA

Ricoeur P 1992 *Oneself as Another*. Blamey K (trans.) University of Chicago Press, Chicago

Taylor C 1985 *The Concept of a Person*. Vol. 1, *Human Agency and Language*, Cambridge University Press, Cambridge, UK, pp. 97–114

A. Ferrara

Person Perception, Accuracy of

The *accuracy* of person perception (also called *personality judgment*) refers to the connection between judgments people make of the psychological attributes of others, and the actual status of those attributes in

the persons who are judged. The psychological attributes may be another person's personality traits, or his or her thoughts and feelings at a given moment. The ability to judge personality has concerned psychological researchers, off and on, for more than a half-century (see Allport 1937, Funder 1999, Kenny 1994). The ability to guess what someone is thinking and feeling, called *empathic accuracy*, is a newer topic that has begun to receive increasing attention (Ickes 1997). Research on either topic entails special difficulties because the notion of 'accuracy' raises thorny questions concerning how accuracy is to be conceptualized and evaluated. Despite these difficulties, research since 1980 can report important empirical and theoretical progress.

1. *The Importance of Accuracy*

The accuracy of person perception has obvious practical importance. Employers and other people in everyday life rely heavily on their perceptions of other persons' attributes such conscientiousness, intelligence, friendliness, dominance, and many others. It is an everyday occurrence to make decisions concerning other people as a function of beliefs about what those people are like. Moreover, clinical assessment typically relies heavily on the impressions of a psychiatrist or psychologist concerning a person of interest. Interpersonal judgments and sociometric ratings are also often used as data for psychological research. The accuracy of these perceptions is critical for the quality of the data and the research that is based on them.

The accuracy of personality judgment is also important for theoretical reasons. Accurate person perception implies knowledge about the connections between observable behaviors, on the one hand, and personality traits, on the other hand. Thus, to understand the accuracy of person perception would be to understand the ways in which personality affects behavior. Person perception is a traditional topic of social psychology (see *Social psychology*) and the connections between personality and behavior are traditionally studied by personality psychology (see *Personality psychology*). The topic of accuracy in person perception transcends the boundaries between these traditionally separate sub-fields of research.

Finally, the oldest topic in philosophy, and one of humankind's most ancient concerns, is the relationship between perception and reality. Research on the accuracy of person perception is driven by curiosity concerning just this issue.

2. *Difficulties in Accuracy Research*

In the light of the importance of the accuracy of person perception, it may be surprising to learn that researchers have neglected the topic for long periods of time (e.g., the years 1955–85), and some psychologists with relevant interests (e.g., in person perception) continue to shy away. This avoidance has been caused by several considerations. The most important consideration is the topic's inherent difficulty. At a general level, the topic of accuracy is difficult to conceptualize or to operationalize, because it requires the development of some criterion by which interpersonal judgments can be evaluated as right or wrong. Some psychologists have considered the development of such criteria to be either impossibly difficult or in principle impossible, and therefore turned their attention to other topics. At a more specific level, an important series of articles by Lee Cronbach (e.g., 1955) revealed complex and hitherto-unsuspected technical difficulties inherent in evaluating accuracy in terms of agreement among different judges. Rather than leading to improved methods, this article and related critiques managed to shut down a then-lively field of research, and appeared to intimidate a subsequent generation of researchers from addressing the topic (see Funder 1999). A further obstacle to accuracy research was the widespread influence of a set of attitudes that seemed to discourage interest in the topic. The 'person situation debate' in personality psychology (Kenrick and Funder 1988) led many researchers to believe that the connections between personality and behavior were few and weak. If this were true, it would make little sense to attempt to investigate the accuracy of person perception. At about the same time, the increasing popularity of research on 'errors' in person perception (e.g., Nisbett and Ross 1980) led to a widespread belief that human judgment is characteristically, even fundamentally mistaken. If this were true, investigations of accuracy would seem to hold little hope.

3. *Resolving the Difficulties in Accuracy Research*

Research on accuracy was able to move forward only after these difficulties began to be resolved.

First, views concerning the existence and importance of personality traits, and the abilities of human judgment, began to change during the 1980s. The person situation debate was resolved as reiterating the importance of stable attributes of personality (Kenrick and Funder 1988). Research on the errors of human judgment was reinterpreted in a broader context in which it was less often seen as implying that human judgment is always or fundamentally wrong (e.g., Swann 1984). Studies of error identify important shortcuts ('heuristics') and other processes of human cognition. However, they do not imply that judgment is inept, because the processes that produce errors in the laboratory often lead to correct judgments in real life. Moreover, research on error is fundamentally limited because it can illuminate only how judgments fail to achieve perfection, rather than how correct judgments are ever achieved. The methodological

issues identified by Cronbach and others also received a reexamination during the 1980s (e.g., Funder 1980, Kenny 1994). This reexamination yielded more sophisticated analytic techniques that allowed the methodological complications to be bypassed or specifically incorporated into analytic designs. David Kenny's 'Social Relations Model,' a data-analytic technique that accounts for effects of judge, target, and their interaction on interjudge agreement, was an important breakthrough in this regard.

Notwithstanding the issues just summarized, the thorniest issue in accuracy research is the criterion problem. By what standard can a judgment of personality be evaluated as right or wrong? The answer to this question depends on exactly how accuracy is conceptualized.

4. *Conceptualizations of Accuracy*

The accuracy of person perception is conceptualized somewhat differently by three major approaches that developed during the 1980s and 1990s.

The *pragmatic* approach, based on some early ideas by William James, has been espoused by William Swann (1984). This approach defines accuracy as the ability to make judgments that allow for success in life. The index of whether a judgment is accurate, in this view, is whether it is adaptively useful.

The *constructivist* approach, based on postmodernist viewpoints that have become widespread in academia, has been espoused by Arie Kruglanski (1989) and others. This approach either states or implies that no criterion for a person's judgmental accuracy exists beyond other judgments by other people. For example, Kenny (1994) has sometimes conceptualized accuracy as the mean judgment of all possible observers of a person. The index of whether a judgment is accurate, in this view, is whether it agrees with the judgments of others.

The *realistic* approach is consistent with the writings of Egon Brunswik (1956) and has been espoused more recently by Funder (1995). This approach assumes that psychological attributes of people really exist, though they are only probabilistically knowable through cues of uncertain validity. The real nature of a person must then be ascertained, as well as possible but always with less-than-certain validity, through multiple sources of evidence including the person's self-judgment, judgments by others, and his or her relevant behavior. The index of whether a judgment is accurate, in this view, is the degree to which it converges with these multiple criteria.

5. *The Degree of Accuracy*

When research on accuracy in person perception revived during the early 1980s, one of its first orders of business was to dispel the impression left by research on error, that human social judgment is always or fundamentally mistaken.

A large number of studies documented the accuracy of personality judgment in two ways. The most common method was to show that different judgments of the same person tend to agree. Research examined self-other agreement and agreement among judges and found that significant agreement was the rule rather than the exception across a wide band of personality traits, even after brief periods of acquaintance (e.g., Borkenau and Liebler 1992). Evidence concerning the ability of personality judgments to predict behavior is more difficult to gather and therefore was slower in coming, but eventually showed that layperson's personality judgments have an impressive ability to predict independent observations of an individual's behavior (Funder 1999).

The overall degree of accuracy in person perception is difficult to characterize, because it depends upon a number of factors including the ability of the judge, the judgability of the target, the trait that is judged, and the amount and kind of acquaintance between target and judge (see next section). But while it is for these reasons an oversimplification, it is also a fair characterization to state that the usual correlation among criteria for accuracy is in the range 0.30–0.40. This translates to the statement that one person's perception of the personality of another is correct roughly two times out of three.

6. *Moderators of Accuracy*

Research since 1980 has found that the degree of accuracy in person perception depends upon four moderator variables.

6.1 *The Good Judge*

The search for the 'good judge of personality' is one of the oldest concerns in the accuracy literature. This search motivated many investigators during the 1930s to 1950s and it was their failure to find consistent results that contributed to the three-decade hiatus that followed (Schneider et al. 1979). Part of the reason for this failure may be the presence of methodological flaws in many early studies, and another part may be that judgmental ability may not be robust. That is, a good judge of one kind of person or trait may not be a good judge of other people, other traits, or in other contexts. Research continues to address this topic but at present the 'good judge' has yet to be conclusively found.

6.2 *The Good Target*

The search for the good target of judgment, or the 'judgable' person, has had more success. As Gordon

Allport noted years ago, some people are relatively open books whereas others are difficult to understand even after years of acquaintance. According to research by Colvin (1993), the judgable person is one who has an integrated, well-adjusted personality. The consistency among such a person's actions, thoughts, and personality leads them to be coherent overall, and relatively easy to judge.

6.3 The Good Trait

An even more robust phenomenon is the difference among traits in their judgability. Relatively observable traits, like sociability and talkativeness, are much easier to judge accurately than more hidden traits like ruminativeness and anxiety. Some evidence also suggests that traits that are extremely desirable or undesirable to possess are more difficult to judge accurately than traits more neutral in evaluative connotation.

6.4 Information

The accuracy of a judgment also depends critically upon the information upon which it is based. The information variable has two facets, quantity and quality.

The quantity variable pertains to the *acquaintance effect*, the tendency to judge someone more accurately the longer one has known him or her. The straightforward expectation would be that more information is likely to produce more accurate judgment. Perhaps surprisingly, this has turned out to be a controversial proposition, with some investigators finding evidence for the acquaintance effect, and others doubting its existence on theoretical and empirical grounds.

The quality variable pertains to the kind of context in which one person observes another. Current research is pursuing the idea that, even holding amount of acquaintance constant, different contexts may produce different degrees of accuracy.

7. Theoretical Development

Historically, research on accuracy has been long on empiricism and short on theory. This imbalance is beginning to be remedied with the development of theories of how accurate judgments come about. One theory, Kenny's 'PERSON' model, conceptualizes personality judgments as the result of multiple influences including the judge's response proclivities, his or her knowledge of people in general, the actual personality of the target, and other factors. Another theory is Funder's Realistic Accuracy Model (RAM, Funder 1995). RAM conceptualizes an accurate personality judgment as the result of a process that begins with the target emitting relevant information (e.g., performing a relevant behavior), which then must become available to a judge, who then detects it and utilizes it correctly in reaching an accurate judgment. Both the PERSON and RAM models can be used to rationalize the moderators of accuracy that have been found to date, and suggest new ones for investigation in future research.

8. Conclusion

The accuracy of person perception is an important topic that, after periods of neglect, is becoming a mature field of research and a mainstream concern of personality and social psychology. The combination of empirical and theoretical progress in this area makes it a good candidate for becoming even more central within psychology in the twenty-first century.

Bibliography

Allport G W 1937 *Personality: A Psychological Interpretation*. Holt, New York

Borkenau P, Liebler A 1992 Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology* **62**: 645–57

Brunswik E 1956 *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley, CA

Colvin C R 1993 'Judgable' people: Personality, behavior, and competing explanations. *Journal of Personality and Social Psychology* **64**: 861–73

Cronbach L J 1955 Processes affecting scores on 'understanding of others' and 'assumed similarity'. *Psychological Bulletin* **52**: 177–93

Funder D C 1980 On seeing ourselves as others see us: Self-other agreement and discrepancy in personality ratings. *Journal of Personality* **48**: 473–93

Funder D C 1995 On the accuracy of personality judgment: A realistic approach. *Psychological Review* **102**: 652–70

Funder D C 1999 *Personality Judgment: A Realistic Approach to Person Perception*. Academic Press, San Diego, CA

Ickes W J (ed.) 1997 *Empathic Accuracy*. Guilford Press, New York

Kenny D A 1994 *Interpersonal Perception: A Social Relations Analysis*. Guilford Press, New York

Kenrick D T, Funder D C 1988 Profiting from controversy: Lessons from the person-situation debate. *American Psychologist* **43**: 23–34

Kruglanski A W 1989 The psychology of being 'right': The problem of accuracy in social perception and cognition. *Psychological Bulletin* **106**: 395–409

Nisbett R E, Ross L 1980 *Human Inference: Strategies and Shortcomings Of Social Judgment*. Prentice-Hall, Englewood Cliffs, NJ

Schneider D J, Hastorf A H, Ellsworth P C 1979 *Person Perception*, 2nd edn., Addison-Wesley, Reading, MA

Swann W B 1984 Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review* **91**: 457–77

D. C. Funder

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Personal Identity: Philosophical Aspects

1. Personal Identity: a Modern Problem

Identity has two meanings: it denotes category membership (e.g., X is a politician); or it identifies a specific individual (e.g., X is Mr. A). Goffman (1963) speaks of social and personal identity. Different criteria are used to determine both types of identity: physical criteria (e.g., sex, race, a thumbprint, a conspicuous scar); social criteria (e.g., general roles such as teacher, or judge, and unique roles such as queen of England or pope); and psychological criteria (competencies or personality characteristics as general variables, and unique values in any of these variables, such as the world's best tennis player).

In traditional societies, all of these criteria converge when they are used to classify individuals. They converge across contexts, across the life cycle, and across perspectives. To illustrate, the identification of X as 'tenant of the Z farm' simultaneously ascribes a social identity (e.g. X is a man, a farmer) and denotes personal identity (i.e., X is the stout, pockmarked man living at the south end of the village, who keeps initiating political debates in the pub). The identity markers used are valid in all contexts: in his family, in the village, in the pub everybody will know him as the pockmarked man running farm Z. They are valid across his life-span: From birth to death he is linked to the Z farm as its future, present, or former tenant. And descriptions from first- and third-person perspectives agree: his farmhands, his wife, and the other villagers know him as the tenant of Z, and he also identifies himself with this attribution—in fact so much so that he will have written on his tombstone: 'Here lies the tenant of Z farm.'

With modernization these dimensions diverge. As a result of urbanization, social and personal identities separate: town people can only vaguely guess each other's social identity, let alone know their personal identity. Social differentiation and cultural pluralization produce consistency problems: individuals increasingly partake in several subsystems or subcultures, each operating according to their own code or claiming specific value commitments. Thus, in different contexts they find themselves exposed to specific role expectations which quite often make conflicting demands. Rapid social change impedes the experience of personal continuity over the life course (e.g., elderly Germans have been subjects in an empire, citizens in the Weimar Republic, followers in a totalitarian regime, and citizens in a democracy again).

In general, modern sociostructural arrangements contribute less to the establishment and stabilization of identity than those of the past. From a third-person or a societal perspective, this is unproblematic. Members of modern societies are used to acknowledging or ignoring each other as strangers. When someone's identity needs to be known there are social mechanisms for identification: titles, uniforms, or the office reception area inform on status or social identity; a passport or genetic analysis enable us to determine personal identity; the curriculum vitae and certificates give evidence of personal continuity; institutionalized sanctions guarantee at least context-specific consistency. A problem does, however, arise from a first-person or internal perspective. The question 'Who am I?' is not satisfactorily answered by referring to unparalleled fingerprints or to social positions held. Rather, individuals are in need of a sense of uniqueness, and a feeling of inner consistency and continuity over and beyond matchless physical markers or straight careers.

2. Theoretical Solutions to the Identity Problem

Various solutions to this modern problem of a subjective sense of identity have been proposed. In some of his writings Erikson (1958) assumed identity to be grounded in those lifelong substantive commitments (to a partner, an occupation, a worldview) a person enters upon, having successfully solved the adolescence crisis. Nowadays, however, people keep experiencing ever new crises (e.g., midlife crises, retirement crises, etc.) in which such commitments are revoked (e.g., by divorce, religious or ideological conversions, loss of employment, or occupational changes).

Elsewhere, Erikson (1956) defined ego identity as a formal competence: the competence to uphold internal unity and continuity. This competence arises from successful resolutions of previous crises in which basic trust, autonomy, initiative, and a sense of achievement (rather than distrust, shame, doubt, and guilt, and a feeling of inferiority) have been developed. In this definition, however, people's need for substantive commitments remains unfocused.

Krappmann (1969) sees identity arising from the social recognition found in those interactions in which an individual succeeds in balancing out contradictory expectations, while at the same time signaling that they never will be merged totally in any given context. Krappmann, however, fails to specify any criteria that might enable us to distinguish opportunistic forms of balancing from the principled type of flexibility that is a prerequisite for warranting an inner sense of consistency and continuity.

Postmodern approaches deny that identity may need consistency and continuity. Instead, they propose the concept of patchwork identity to reflect the fact that there are constant changes in preferences, attitudes, roles, and relationships. However, even the external criteria that might enable us to distinguish one patchwork from the other, or to recognize sameness over time in case individual patches have been exchanged or colored, remain unclear. And from

an internal perspective, body boundaries and physical continuity are not enough to confer a sense of identity. In fact—as is well documented by research in various fields such as dissonance theory, attribution theory, and studies on dogmatism and split personality—at least some consistency seems indispensable for psychological well-being and mental health.

Cohen and Taylor (1976) focus on sense of uniqueness. Individuals seek to develop eccentric hobbies, interests, or life-styles, go in quest of adventure or unusual experiences so as to resist the fear of exchangeability that keeps creeping up on them in modern mass societies. Nevertheless, over and over again, they find that any newly invented identity marker will be imitated by others, socially co-opted, and commercialized. Thus, all attempts to secure personal identity by appropriating singular activities are—empirically—doomed to failure. Moreover, the very idea of basing identity on unique markers is misconceived because it reduces the dialectic inherent in distinctiveness to just one pole—to positively valued exceptionality—and overlooks the risk of being seen as deviant and becoming stigmatized. More importantly still, it neglects the core feature of a sense of uniqueness—its being a 'necessary byproduct' (Elster 1979). There are states such as sleeping, forgetting, or happiness that will inevitably be missed by those who directly aim for them. In the same way, a sense of uniqueness can only be experienced concomitantly; i.e., as a consequence of doing something one is intrinsically committed to doing.

Giddens (1991) conceptualizes identity as a 'consistent narrative': enveloped in the 'protective cocoon' of basic trust, individuals keep rewriting their autobiography so as to create a consistent meaning structure from the sequence of decisions made, the paths chosen, and the actions taken. This definition integrates motivational aspects (trust), substantive commitments (life decisions), and a need for consistency. The meaning of consistency, however, remains ambiguous: by insisting that identity be based on a 'morality of authenticity that skirts any universal moral criteria' (Giddens 1991, p. 79), Giddens cannot distinguish between consistency based on principles or consistency based on defensive rationalizations ('sour grapes').

3. *Aspects of Identity: Theoretical Assumptions*

From this brief overview and critical comments, several assumptions concerning constitutive aspects of identity can be derived:

(a) Commitment is constitutive of identity.

(b) Given that commitment needs an object (e.g., a person, values, projects), content is requisite for identity.

(c) What matters, however, is not the type of content chosen, but the mode of appropriating it.

(d) A subjective sense of uniqueness is a necessary by-product of an autonomous commitment.

In the following, these assumptions will first be theoretically justified. Second, I will try to substantiate them empirically.

Assumption (a) follows Frankfurt's (1988, 1993) analysis of the concept of a person. If an individual had nothing he/she would never betray, he/she would be at the helpless mercy of contingent conditions, of inner drives, of fleeting whims. He/she would not be a person but a 'wanton'. Personhood (i.e., identity) is acquired by committing oneself to ideals. The kind of ideals chosen—be they moral, aesthetic, or truth-oriented ones—is of little import. What counts is that the person—in view of his/her commitment to an ideal—is taking a stance toward spontaneous impulses and desires, and is suppressing those that are incompatible.

Frankfurt describes the way in which identity is constituted; Nozick (1981) analyzes the way in which it may be stabilized. He proposes the following thought experiments. Assume that Theseus' ship is anchored in port. As time passes, one plank after another is beginning to rot and is replaced. Finally, all the planks have been replaced by new ones. What is Theseus' ship? The renovated ship in port or the original planks piled up in the shipyard? We would consider the ship in port to be Theseus' ship, because, despite a change in its components, it has maintained its structure and its transformation is of some continuity underneath the seties. Analogously, we consider an organism to continue to be itself despite the fact that its cells are continuously replaced. Similarly, we consider a person to remain the same even if they have changed some of their views. We do so only, however, if we assume that they have changed their views autonomously, i.e., voluntarily, and with good reason. Nozick explicates this using counter-examples. Assume that, due to grave brain damage, a man has been suddenly reduced to the state of an infant, or that as a result of brainwashing, an ardent democrat has been transformed into a fanatic communist. We would say of both that they are no longer the people they once were. In contrast, we have no problem in acknowledging the identity of a former opponent to atomic energy who is now arguing *for* the use of atomic energy, if we can understand this change of mind as being motivated by reasons (e.g., a fear of global warming), and not effected by causes. In other words, we do not consider identity to be endangered if changes in mind result from learning processes. The concept 'learning process' implies that the new state is better than the former; i.e., it presupposes intersubjectively validated standards to judge the adequacy of views held. Unavoidably, everyday life is based on such collectively shared understandings of rationality—otherwise the very idea of, for example, questioning the identity of the brainwashed communist would be meaningless.

4. Aspects of Identity: Some Empirical Support

There are empirical findings that agree with these philosophical reflections. In *The Uncommitted*, Keniston (1965) describes the feeling of inner void and meaninglessness, and the sense of exchangeability some young Harvard students experienced despite the fact that from an external perspective they were considered as being highly successful; i.e., ascribed a positively valued identity. In contrast, by committing themselves to something they deem worthy and important, the Civil Rights Movement, the *Young Radicals* (Keniston 1968) increasingly came to develop a sense of uniqueness. Similarly, Helen John, a leading pioneer in protests against nuclear missiles, came to feel irreplaceable once she began to accept personal responsibility. As she put it: 'Nobody can do exactly what I am doing, in the way I am doing it. Only I can do that.' (Haste 1993, p. 339) The sense of uniqueness is not dependent on 'difference,' on singular contents; rather, it is dependent on personal commitment. Only in truly caring for something can a person feel inexchangeable.

Individuals tend to realize that caring for specific ideals is constitutive of their own identity. In the context of a longitudinal study I asked 170 17-year-olds: 'What would most make you a different person—different looks, parents, hobbies, grades, a different understanding of right and wrong; another nationality, sex; a lot of money? Why?' The answers given most often were: different parents, different sex, different moral understanding. The importance of 'ideals' to identity could be seen in the frequent references to 'moral understanding,' and in the justifications given for the response 'different parents.' Most subjects argued that growing up in a different family would make them a different person because parents so decisively influence and shape one's value orientations. There is a more indirect confirmation of the claim that ideals are constitutive of identity: many longitudinal studies find that individuals tend to hold on to the sociopolitical value orientations developed in late adolescence—even if they have experienced social and geographic mobility or fluctuations in public opinion (cf. Sears 1981). This finding may indicate that people seek to stabilize their identity during their life course by remaining true to their value commitments—irrespective of changes in context or the ideological climate.

As noted above, however, this mode of upholding identity is rendered increasingly difficult given rapid and far-reaching changes in knowledge systems and collective interpretation. In the following, I maintain that with modernization two new mechanisms that enable us to constitute and stabilize identity are evolving: value generalization and an egosyntonic motive structure. I will illustrate both using data from a cohort comparison involving 100 65- to 80-year-old, 100 40- to 50-year-old, and 100 20- to 30-year-old subjects representative of the West German population. The study focused on the understanding of moral rules and moral motivation (cf. Nunner-Winkler 2000a, 2000b). In making use of these data in the present context I do not, however, wish to imply that the ideals constitutive for identity need to be moral ones.

5. Mechanisms for Stabilizing Identity in Modernity

5.1 Value Generalization

Commitment to more abstract values increases the flexibility required if varying context conditions are to be taken into account. To give an example: when asked to judge a young mother's desire to work full-time without real economic need, older subjects express clear condemnation; they see it as a neglect of duty. Younger subjects, in contrast, no longer assume ascriptive female duties; they focus on the problem of good child care. On these more abstract terms several functionally equivalent solutions are possible. Similarly, older subjects tend to condemn a sexual relationship between two men as unnatural, deviant, or sick. Younger subjects instead tend to discuss it in terms of the quality of the relationship. Thus, increasingly concrete rules tying behavioral implications to specific context conditions (e.g., child care to being a mother; sexual relations to being of the opposite sex) are replaced by more abstract values (e.g., welfare of the child; quality of the relationship) that allow flexibility in realization without a betrayal of basic commitments.

5.2 Egosyntonic Motive Structure

Norm conformity can be motivated by various concerns: by fear of physical or institutional sanctions, social contempt (Luhmann), or pangs of conscience (Freud); by compassion (Schopenhauer); by respect for the law (Kant). In a longitudinal study on moral development it was found that children see moral behavior as being motivated not by the consequences anticipated either for the wrongdoer or for the victim (e.g., by fear of sanctions or by compassion), but by an intrinsic desire to do what is right (Nunner-Winkler 1999). The cohort comparison shows that between generations there is a change in the understanding of moral motivation indicated by changes in the emotional reactions expected to follow transgressions. Older subjects mostly refer to guilt and shame; that is, to emotions that focus on some higher internal or an external entity (e.g., the superego or a social audience); younger subjects, in contrast, are more likely to express sorrow or regret; that is, emotions that focus on the wrong committed or the harm done. Thus, norm conformity is increasingly coming to be seen not

in terms of bowing to dictates authoritatively imposed on people but rather in terms of not wanting to wrong others. This change in moral motivation in the direction of a more egosyntonic structure corresponds to a change in the cognitive understanding of morality, in as much as norms are no longer seen to be set by God or church authorities, but rather as derived from rational consensus among equals.

To conclude, with modernization—i.e., with social differentiation, cultural pluralization, and rapidly changing social conditions—new mechanisms have evolved that allow us to constitute and stabilize identity. By egosyntonic commitment to generalized values individuals may develop a sense of uniqueness, and experience consistency and continuity in their life course. In this way ego identity becomes a substitute for social role identity.

See also: Discourse and Identity; Identity and Identification: Philosophical Aspects; Identity in Anthropology; Identity in Childhood and Adolescence; Identity: Social; Multiculturalism and Identity Politics: Cultural Concerns; Religion: Nationalism and Identity

Bibliography

Cohen S, Taylor L 1976 *Escape Attempts. The Theory in Practice of Resistance to Everyday Life*. Allen Lane, London

Elster J 1979 *Ulysses and the Sirens*. Cambridge University Press, Cambridge, UK

Erikson E H 1956 The problem of ego identity. *Journal of the American Psychoanalytic Association* **4**: 56–121

Erikson E H 1958 *Identity and Uprootedness in Our Time*. Paper presented at the Annual Meeting of the World Federation of Mental Health, Vienna

Frankfurt H G 1988 *The Importance of What We Care About*. Cambridge University Press, New York

Frankfurt H G 1993 On the necessity of ideals. In: Noam G G, Wren T E (eds.) *The Moral Self*. MIT Press, Cambridge, MA, pp. 16–27

Giddens A 1991 *Modernity and Self-identity. Self and Society in the Late Modern Age*. Stanford University Press, Stanford, CA

Goffman E 1963 *Stigma. Notes on the Management of Spoiled Identity*. Prentice-Hall, Englewood Cliffs, NJ

Haste H 1993 Morality, self, and sociohistorical context: The role of lay social theory. In: Noam G G, Wren T E (eds.) *The Moral Self*. MIT Press, Cambridge, MA, pp. 175–208

Keniston K 1965 *The Uncommitted: Eliminated Youth in American Society*. Harcourt and Brace, New York

Keniston K 1968 *Young Radicals – Notes on Committed Youth*. Harcourt, Brace and World, New York

Krappmann L 1969 *Soziologische Dimensionen der Identität*. Klett, Stuttgart

Nozick R 1981 *The Identity of the Self. Philosophical Explanations*. Clarendon Press, Oxford, UK, pp. 27–114

Nunner-Winkler G 1999 Development of moral understanding and moral motivation. In: Weinert F E, Schneider W (eds.) *Individual Development from 3 to 12. Findings from the Munich Longitudinal Study*. Cambridge University Press, New York, pp. 253–290

Nunner-Winkler G 2000a Wandel in den Moralvorstellungen. In: Edelstein W, Nunner-Winkler G (eds.) *Moral im sozialen Kontext*. Suhrkamp, Frankfurt am Main

Nunner-Winkler G 2000b Von Selbstzwangen zur Selbstbinding (und Nutzenkalkülen). In: Endreß M, Roughley N (eds.) *Anthropologie und Moral*. Konigshausen Neumann, Würzburg, Germany

Rawls J 1985 Justice as fairness: Political not metaphysical. *Philosophy and Public Affairs* **3**: 223–51

Sears D O 1981 Life-stage effects on attitude change, especially among the elderly. In: Kiesler S B, Morgan J N, Kincade Oppenheimer V (eds.) *Aging. Social change*. Academic Press, New York, pp. 183–204

G. Nunner-Winkler

Personal Privacy: Cultural Concerns

Personal privacy is a vexing and invaluable concept: vexing because it is difficult to describe; invaluable because privacy allows people the possibility of freedom, and thus autonomy and personal choice. In this sense, privacy is a precondition of psychological concepts of identity: the modern ideal of an individual requires—and cannot be fully realized without—substantial privacy. Personal privacy can be pictured as the atmosphere generating part of the ecosystem in which this individual can draw breath.

Perhaps most succinctly defined by the US social philosopher, Sissela Bok (1982), privacy is 'the condition of being protected from unwanted access by others—either physical access, personal information or attention.' This protection, whether physical, political, or psychological provides freedom because it shelters people from intrusion; it fends off predators that might pre-empt personal choice; it grants control. Privacy also emphasizes interiority: the subject–while protected from undesired surveillance and intrusive access—is freed to focus more fully on self-observation, and to develop efficacious forms of self-expression.

1. Historical Background

An early template of this phenomenon can be found in Jane Austen's novel, *Pride and Prejudice*, published in 1813. Midway through the novel, its heroine Elizabeth is sitting alone in a room reading a letter when the romantic hero, Darcy, comes upon her and is startled by her solitude. Modern readers may not recognize what an unusual spectacle such a tableau offered in Austen's era. Women were rarely alone, and many were only recently literate enough to read and write letters. The book also describes Elizabeth's startling penchant for solitary walks. Her search for solitude underscores her characterization as a new, independent woman with a more self-conscious and thus a sturdier sensibility. Alone in a room with time and

peace to peruse an intimate letter, Elizabeth can contemplate her reactions to what she is reading. She can notice what she thinks and feels without immediately being distracted or having her impressions overrun or overruled by others around her. Given a pen and paper, she can respond from her own perspective; she can express an opinion, and in a larger sense, a point of view.

These behaviors, made possible by privacy, and the resulting increased capacities for autonomous voice and self-expression, distinguish Elizabeth from her less able peers, and define her as a substantial individual in a modern sense. This view is furthered by her refusal to marry Darcy when he first approaches. Although he is wealthy and upperclass, and she is from a more middle-class family with money worries, and although marriage is the only avenue into society for her, she refuses his offer because she does not love him. Her behavior is radical as it suggests love—and the sensibility of the subject—as a value of more significance than social standing or economic well-being. While in contemporary Western culture, many take for granted love as the superior foundation for marital commitment, this notion, this priviledging of private feeling, is also—historically speaking—new, having developed a popular following gradually over the past few centuries. And, as Austen illustrates, in the era of *Pride and Prejudice*, it was easily displaced by other imperatives.

Austen's heroine is only satisfied when she can choose to marry Darcy, based on her private feelings—those she discovers within her heart. Elizabeth's insistence on such a course illustrates well the confluence between forms of privacy, like solitude, and the heightened capacity for personal choice based on feelings. One emphatic version of this idea of individual integrity rooted in private experience was asserted by the US transcendentalist Bronson Alcott (1799–1888) when he wrote, 'Individuals are sacred. The world, the state, the church, and school, all are felons whenever they violate the sanctity of the private heart' (in Seipp 1981, p. 99).

While all cultures have had some notions of privacy (Barrington Moore Jr. 1984 traces its rudimentary presence in Ancient Greece and the Old Testament), the modern concept—illustrated so well in *Pride and Prejudice*—has been developing gradually since around the seventeenth century. The cultural historian Philip Aries (1962) has pointed out that until the early seventeenth century, almost no one in Europe spent time alone. But privacy is more than solitude. Alan Westin defined its four substates as solitude, anonymity, reserve, and intimacy. While these states are diverse, they all describe methods of purposefully shielding oneself from the physical and/or psychological 'gaze' of others.

The ideal of personal privacy is partly an outgrowth of a gradual social transition accompanying the mass urban migration of the nineteenth and twentieth centuries. People left communal situations—small towns and villages where their lives, monitored by church and neighbor, had been witnessed closely and controlled largely by people who knew them. Finding themselves more alone, separate, and increasingly anonymous, people, first in cities then in suburbs, realized that they were at liberty to choose their activities and behaviors with less reference to communal imperatives. Although in part following an economic infrastructure defined by industrialization and capitalism, this increased autonomy drew its cultural meaning from philosophies of individual rights and freedom. While the sources are too many to site, one would have to acknowledge the influence of John Locke, Edmund Burke, Jean Jacque Rousseau, Johann Goethe, William Wordsworth, Thomas Jefferson, Ralph Waldo Emerson, Mary Wolstencraft, Margaret Fuller, and Henry David Thoreau among many others.

Within this tradition, privacy can be defined as the right to make one's own choices relatively unobserved, unchastised, and unimpinged upon by others—whether individuals or institutions. While there is no legal 'right to privacy' in the USA, the concept began to be articulated in the late nineteenth century. In 1872 the feminist Victoria Woodhull referred to 'the right to privacy' in a newspaper piece. The famous expression, 'the right to be let alone,' was then coined by Judge Thomas Cooley in 1879. In 1890, Samuel Warren and Louis Brandeis published a landmark article, 'The Right to Privacy' in the *Harvard Law Review*. No important discussion of privacy in the USA since then has failed to refer to Warren and Brandeis.

Privacy had been increasingly understood as a precondition for human dignity. Shielded from intrusive observation, people take control over the matter of 'showing' themselves to others or not based upon their own sense of readiness. Few feel dignified when exposed to hostile or secret surveillance. Privacy functions as a practice room in which people can process and master experience before displaying themselves to others. Having this option, people feel less shame. Shame is in essence an internal overseer—employed by those parts of the mind whose job is to keep the individual in line with the communal norm. Privacy provides the space in which the communal demand can be weighed against the internal individual imperative, and a choice of behavior made through consideration rather than reflex. Since shame is also a psychic experience where the sense of self temporarily disintegrates, the shielding of the individual with privacy, the minimizing of shame, and the valuing of dignity, all enforce conditions which enhance people's sense of individual identity and worth.

2. *Theories of Privacy*

Alan Westin's landmark book, *Privacy and Freedom* (1968), defined four states of privacy. Because these

states are to some degree counterintuitive, it is useful to describe them. The first—solitude—is perhaps the most self-evident and the one most popularly associated with privacy. When privacy is mentioned, the mind often jumps to an image of a person alone, whether in nature or in a room. In US literary culture, one might think of Henry David Thoreau living at Walden Pond, or—half a century earlier—of James Fenimore Cooper's fictional frontiersman, Natty Bumpo, independently making his way through the American wilderness. Solitude is the most complete state of privacy; in it one is physically separate from others. Historically, solitude has been associated with men and linked to images of independence, moral awareness, and strength. Until recently in Western Europe and North America, women were only granted solitude for prayer. Unchosen, solitude becomes isolation and, like the extraordinarily punitive condition of solitary confinement, can be a psychically destructive experience; people, with inadequate access to others become more inclined to psychic turmoil.

Anonymity is Westin's second state of privacy. When one is with others but unknown to them, as in a crowded train or a city or a suburban shopping mall, one is anonymous. Anonymity offers privacy by allowing human proximity in the absence of knowledgeable witnesses. One is freed in this situation (for better and worse) to experience a subjective sense of oneself or of current experience unchallenged by the proximity of knowing and potentially contradicting others. The degree of anonymity that urban and suburban dwellers experience today is historically atypical and sometimes shades into the alienating sense of anomie.

The third state of privacy is reserve. Reserve is perhaps the oldest and most universally available form of privacy; the privacy one can access even when surrounded by intimates. It is the default state of privacy, the great mediator of psychic survival and human freedom. One has, thanks to reserve, the freedom not to say openly or fully what one thinks or feels, not to reveal oneself in situations that feel unsafe. Reserve is a basic element in civility and civilization, but more than that, it negotiates the distances between people, facilitating all varieties of encounter by allowing exchanges among people where disruptive observations go unspoken and social intercourse is expedited. It is also one of the equalizers in social power. The powerful, denied critical response or information by the reserve of the 'weak,' can have their intentions impeded.

Intimacy is Westin's fourth state of privacy, and perhaps the most surprising one to find in association with the others. Intimacy is a state of privacy where two people, or a small group or a family, shielded from the larger group, are able to reveal themselves more fully, physically, psychologically, sexually, and so forth. As individuals have come to live more separately and less communally, there has been an increased demand for intimacy in primary relationships. At the same time, continuing our delineation of the relationship of privacy to the ideal of the more fully developed individual, intimacy is considered a well-spring for psychic development. Within intimate relationships, individuals hope to find validation of their nascent self-expression which then ideally strengthens and reinforces their capacity to carry that private self into the public world. Intimacy also provides release in the form of freer emotional and sexual expression.

The 'possibility' within privacy also includes individual artistic expression, which in turn dovetails with the romantic emphasis on the artist as the highest realization of modern sensibility. According to this view, well articulated in Charles Taylor's (1989) compendium on the sources of modern identity, the artist, creating in private and using privacy to appropriately shield and nurture his or her creative process, expresses otherwise obscured, often internally uncovered, truths about human experience, and by so doing transforms that experience.

And while the artist may be the prototype of this style of individual, it would be a mistake to minimize privacy's broader but related function in everyday life. Take two contemporary flashpoints as illustrative: One can see the critical role of personal privacy both in the abortion debate and the gay rights movement. In both these situations, the wish of individuals to choose behavior around sexuality and procreation is placed in conflict with traditional social imperatives. Specifically, these imperatives have dictated the primacy of procreation and heterosexual procreative sex. In both cases, the progressive position has been framed as valuing individual choice above procreation specifically because this choice highly values subjective experience—the primacy of the individual heart. Thus the individual's private feelings—about readiness to have children, or the number of children to have, about whom one wants to love and how—are held to be of worth precisely because they enforce people's capacity to act according to their own lights, and to transform themselves by these actions: to become 'more fully' human through freedom.

Sigmund Freud's notions of a complex human psychology that requires and transforms itself through self-knowledge, and particularly through self-knowledge inculcated in the intimate process of psychoanalysis, fits within this modern template. In the various psychotherapies which have proliferated in the USA and Europe during the twentieth century, the therapist assists patients to become themselves more fully by verbalizing their private (even secret) heretofore inexpressible thoughts and feelings. The process of naming, the intimate yet contractual relationship, and the resulting restructuring of aspects of the patient's mind are meant to create a transformative and relatively liberating self-knowledge. The privacy of the consulting room and the confidentiality of the process are considered central to its efficacy in that

they secure a free space between social demand and individual subjectivity.

3. Threats to Privacy

Privacy is a fragile state, and one—as the twentieth century has demonstrated—that is exploited readily. No discussion of the merits of personal privacy can be complete without some exploration of its accompanying liabilities. While many of the most eloquent voices of the twentieth century—among them George Orwell, Nadia Mandelstam, and Primo Levi—have delineated the damage to liberty and individual authority brought about by hostile surveillance and its companion, terror, less has been said until recently of the equal threat to privacy caused by secrecy and social isolation. (Perhaps the one historically illuminated aspect has been in those instances when it has been tied with totalitarian power wherein the masses are placed under surveillance while the motives and behaviors of leaders are kept secret.)

But domestic secrecy turns out to be a potent destructive force with regard to personal privacy. Under the guise of privacy, more powerful members of families have had reign to exploit or abuse less powerful members—whether emotionally, sexually, or physically. These acts are significant not only because of their exploitative dimensions, but because by their very nature they stunt the freedom of the oppressed to use the privacy for the aims elaborated above. Any historical survey of social conditions would suggest that such abuse has always been amply present. Yet what has altered currently may be an ideology which pictures family life as unrealistically benign—'a haven in a heartless world' (Lasch 1977)—an idealization that sometimes denies, sometime mitigates, the inherent violence.

A corollary concern is the degree to which many of the benefits made possible by personal privacy accrue only to a privileged or powerful few. Does such a scatter imply that privacy reinforces other social hierarchies, or can it still be construed as a valuable humanistic ideal which will only find full expression if social justice increases other aspects of social equality?

So too, US social philosophers like Christopher Lasch, and the 'communitarians' Robert Bellah and Amatai Etzioni have attacked aspects of US individualism which they see both as excessive, and as indulging an unchallenged pursuit of private purpose and private gain at the expense of the welfare of a larger group. According to this critique, the goals of the individual and the community may be enough at odds to warrant more vigorous interventions on the part of the state to protect communities. Privacy becomes implicated in this debate for its functions shielding individuals and honoring their imperatives.

The technological revolution of the present age also threatens privacy. Computer databanks, video cameras, audio taping, sophisticated surveillance capabilities, and medical technologies like DNA testing make it extremely easy for people to learn private information about other people which was more difficult to access earlier in the twentieth century. Reacting to excessive social isolation and anonymity, and to the enormous profit in selling private information, there is strong interest from many quarters in developing new ways of observing others. The technological gains have come so quickly that inadequate attention is being paid to their pros and cons with regard to the value of preserving privacy. So, for example, the DNA sampling in prison of a petty offender implies a lifelong increased access to knowledge about that individual by the state. And civil liberties issues have tended to get suppressed in the enthusiasm for the new capabilities.

4. Conclusion

In summary, personal privacy protects a wide variety of modern social ideals. Privacy advocates argue that people have a right to be 'let alone'; and that, shielded from excessive authoritarian intrusions, people will make choices based on some more personal sense of what is desirable, that thus constitutes more authentic self-expression. As a corollary, these ideals justify and affirm other cultural notions which support individual liberty and dignity, the value of self-knowledge, and they privilege relationships based in love.

See also: Civil Liberties and Human Rights; Human Rights, History of; Individualism versus Collectivism: Philosophical Aspects; Love and Intimacy, Psychology of; Privacy: Legal Aspects; Privacy of Individuals in Social Research: Confidentiality; Rights

Bibliography

Aries P 1962 *Centuries of Childhood: A Social History of Family Life*. Vintage Books, New York

Bellah R, Madsen R, Sullivan W M, Swidler A, Tipton S M 1986 *Habits of the Heart: Individualism and Commitment in American Life*. Perennial Library, New York

Bok S 1982 *Secrets: On the Ethics of Concealment and Revelation*. Pantheon Books, New York

Etzioni A 1993 *The Spirit of Community*. Crown, New York

Garrow D J 1994 *Liberty and Sexuality: The Right of Privacy and the Making of Roe vs. Wade*. Macmillan, New York

Lasch C 1977 *Haven in a Heartless World*. Basic Books, New York

Moore B Jr 1984 *Privacy: Studies in Social and Cultural History*. M. E. Sharpe, Armonk, NY

Rieff P 1966 *The Triumph of the Therapeutic*. Harper and Row, New York

Seipp D J 1981 *The Right to Privacy in American History*. Harvard University Program on Information Resources Policy, Cambridge, MA
Smith J M 1997 *Private Matters: In Defense of the Personal Life*. Addison Wesley, Reading, MA
Taylor C 1989 *Sources of the Self: The Making of Modern Identity*. Harvard University Press, Cambridge, MA
Warren S D, Brandeis L D 1890 The right to privacy. *Harvard Law Review* **4/5**
Westin A 1968 *Privacy and Freedom*. Atheneum, New York

J. M. Smith

Personality and Adaptive Behaviors

Personality is the psychological system that mediates the relationship of the individual with the environment in the service of adaptation. It is the complexity of structures, processes and patterns of affect, cognition and behavior that we call 'personality' that conveys to others the sense of unity and distinctiveness that distinguish individuals from one another and that grants to each individual a sense of identity, of wholeness and uniqueness.

Because the term encompasses a great variety of phenomena it is not surprising that different ideas appear in various combinations in the traditional and common usage of the word personality: human being, person, the quality of being a person, the quality or fact of being a particular person, patterns of habits and qualities of any individual as expressed through physical and mental activities, the sum of such properties as impressing or likely to impress others, and what makes an individual conspicuously different from others.

Nor is it surprising that, given the complex nature of personality, different assumptions, foci, and strategies of research have marked the study of personality since the beginning of the discipline (Allport 1937, Murray 1938, Stagner 1937, Stern 1935).

The resolution of these different definitions and assumptions has considerable implications for the conceptualization and study of adaptation and adaptive behaviors.

1. *Common Assumptions Regarding Personality*

Before outlining these different assumptions, let us first consider some areas of commonality in the contemporary field. Most investigators in personality psychology share an 'interactionist view'. In other words, they view personality as an open system that develops and functions through continuous and reciprocal interactions with the environment (Bandura 1986, Hettema 1979, Magnusson and Stattin 1998).

Personality, then, refers to a complexity of psychological structures and processes that operate in concert in these person-situation interactions. The personality system underlies the distinctive patterns of affect, cognition, and behavior that foster coherence in individual conduct and experience over time and across settings. Molar psychological mechanisms, such as the goals, beliefs, and skills through which individuals regulate their experiences and actions, result from the synergism of underlying biological and psychological subsystems. Multiple biological systems, cognitive and affective structures, behavior, and the environment all operate as interacting determinants of what personality is at any moment within a network of reciprocal causation. The relative influence of one or another personality determinant varies in different individuals, and across activities and circumstances. Individual development and functioning are processes of continuous restructuring of the whole system and its subsystems within the boundaries set by biological and social constraints (Caprara and Cervone 2000).

2. *Competing Arguments Regarding Personality*

Despite the widespread adherence to an interactionist viewpoint, continuing debates in the field attest to the fact that many fundamental issues are still pending, with different investigators adhering to different theoretical and metatheoretical principles and assigning different importance to nature, nurture, or the individual in the shaping of their personalities (Cervone 1991). Throughout much of its history, personality psychology has been concerned with individual differences in observable variations in styles of behavior, affect, and cognition. The variations have been organized according to, and traced back, to simple systems of dispositional constructs.

Dispositional constructs are designed to encompass the variety of phenotypic expressions of individuality and to capture the consistent individual differences that are observed. Trait constructs have been posited to account for stable patterns of experience and action that people exhibit, and that distinguish them from one another. As different traits are differentially relevant to different settings and their expression is differentially sensitive to social contexts, different traits attest to the consistency of personality in various degrees depending upon their generalizability (across individuals), pervasiveness (across situations), and stability (across time). This has led dispositional theorists to conceptualize personality as a hierarchical organization and to focus on high level traits (e.g., extroversion) that organize lower level tendencies (e.g., sociability) which, in turn, supervise lower level behavioral habits (e.g., talkative) (Eysenck 1970).

In the past alternative taxonomies have been in conflict with each other. However significant con-

sensus recently has been reached on the importance of five factors (so called Big Five) that are obtained robustly: Extroversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. This five-factor model has become a point of convergence between the psycholexical tradition, mostly concerned with personality description (Goldberg 1993), and the tradition, mostly concerned with the identification of the primary structures of personality (McCrae and Costa 1996).

Significant differences are found among five-factors investigators, who may disagree on the exact nature and conceptual status of the trait variables (see Wiggins 1996). However, far more profound differences separate those who view traits as the ultimate determinants of personality from those who conclude that clustering descriptors related to habitual behaviors (as resulting from factor analysis) cannot lead too far in understanding either the regularities that explain personality functioning or the uniqueness which distinguishes persons from one another (Bandura 1999). The issue in contention is not whether people have personal dispositions, nor their determinative role in personality functioning, but rather how dispositions are conceptualized and operationalized. Dispositions may correspond to habitual styles of behavior rooted in genotypes, or to self-regulatory structures (as internal standards, aspirations and goals, and efficacy beliefs) resulting from the organization of affect, cognition, and behavior which emerges over the course of life.

In conceiving of personality as a hierarchical architecture in which basic traits are at the origins of any other individual disposition or behavior, most convinced advocates of traits as genotypes of personality derive their arguments from genetic, longitudinal, and cross-cultural studies. Genetic studies have found a consistent component of heritability for each of the Big Five (Lohelin 1992), longitudinal studies have shown a significant stability of the same traits over the life span (Costa and McCrae 1997), and cross-cultural studies have shown that the multiple descriptors used to describe personality in different cultures can be traced back to few common dimensions broadly overlapping with the Big Five (De Raad 1998). Trait psychologist converge with evolutionary psychologist in viewing adaptation as largely predetermined by the basic dispositional equipment of the species selected through generations. According to this view adaptive behaviors (in mating, fighting, solving problems, and coping with adversities) largely reflect the influence of evolved mechanisms whose nature and functioning has been determined over the eons through processes of natural selection (Buss 1997).

On the other hand, in conceiving of personality as an open, dynamic, unifying and integrating system, social cognitive and motivational theorists (Cervone and Shoda 1999) point to the emerging properties of the mind and focus on the processes and mechanisms conducive to knowledge structures which enable personality to function as a proactive self-regulatory system. Thus, they contest the generalizability to individuals of correlational findings derived from studying populations, and thus leaving large portions of variability unexplained. Rather, they focus on the construction of personality as an integrative and coherent system as it takes place over the course of life, on the processes that enable the system to function proactively with the environment and on the structures that orchestrate these processes. Ultimately they emphasize the flow of reciprocal determination among biological, psychological and social factors which sustain the whole construction, and from which the agentic properties of personality derive. In this view of personality, 'adaptation' is eminently relational and contextual. Adaptive behaviors are not ascribed to any preformed, endogenous dispositions, but to psychological systems which themselves develop through dynamic interactions with the social environment. These systems enable the person to cope in a flexible, innovative manner with the diverse challenges of life.

3. *Personality as a Self-regulatory System*

Personality can be viewed as both a construction and an agentic system. Over time, people gradually construct a sense of self through their interactions with the social world. People's capacity for forethought, performance control, and self-reflection contributes to their ability to regulate their own experiences and actions. Self-regulatory capacities enable people to contribute proactively to the course of their development. Personality, then, reveals itself as a construction endowed with agentic properties. These properties enable people to exert a broad influence on their environment and their own psychological and physical selves (Caprara and Cervone 2000).

Agentic personality functioning involves a number of distinct processes. People select the environments they encounter and thereby create opportunities for themselves, guide their behavior purposively, appraise their circumstances, and chose levels of mental and physical challenge that match their perceived capabilities. Person-situation transactions underscore the fact that personality functioning is not reactive but proactive. Individuals do not only react to the environments, but actively contribute to their transformation and creation. Individuals' distinctive ways of functioning reflect more than the unfolding of a genetic blueprint or the interactions which have taken place between the organism and the environment over the course of development. They also reflect individuals' active role in shaping their own personality.

Thus, one should look at adaptation as a dynamic, proactive process. People do not 'adapt' merely by altering themselves to fit environmental constraints and pressures. They select and transform environ-

ments in order to actualize their full potentials and maximize their own well being. Adaptation relies more on individuals' capacities to devise strategies and to implement tactics able to match successfully the requirements and affordances of multiple situations, than on fixed pattern of responses. The capacity to expand knowledge of oneselves and of situations is no less important than the capacity to master such knowledge with the flexibility that is required by the variety and continuous change of person-situation transactions.

Over the course of life, people acquire competencies, knowledge about themselves and the social world, and standards for evaluating the worth of own actions. Self-reflective and self-directive processes enable individuals to meet environmental opportunities and constraints as well as to maintain personally valued courses of action in the relative absence of tangible rewards or social support. While people show a striking ability to adapt their behavior to physical, interpersonal, and social circumstances, and also to select and create the circumstances that best fit with their purposes and abilities, successful adjustment across the life span corresponds to the best allocation of personal and social resources within the set of temporally bound opportunities and constraints.

Thus selection, optimization, and compensation have been regarded as main strategies which attest the property of personality to capitalize upon experience in the achievement of higher levels of functioning and well-being. People must select life goals, activities, and pathways that are managable; they must optimize the allocation of internal and external resources that are made available to them; they must develop strategies to compensate for lacks or losses of goal relevant means (Baltes 1997).

As people are causal agents who actively contribute to the course of their life in selecting and transforming the environment they encounters, adaptive behaviors are those behaviors that promote the conditions for the full expression of individual capacities and potentials. As individuals course of life is closely interdependent with the course of life of others, ultimately adaptive behaviors are those behaviors that serve for both the betterment of individual and collective welfare.

Future research should focus on psychological processes and social capacities that can strengthen people's capacity to adapt to a rapidly changing world. To this aim theories and findings are mostly needed to enable people to expand their knowledge and skills, to achieve greater emotional adjustment, and thus to fully realize their potentials.

See also: Interactionism and Personality; Personality Psychology; Personality Structure; Personality Theories; Self-efficacy; Self-regulation in Adulthood; Self-regulation in Childhood; Well-being (Subjective), Psychology of

Bibliography

Allport G W 1937 *Personality: A Psychological Interpretation*. Holt, Rinehart & Winston, New York

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ

Bandura A 1999 Social cognitive theory of personality. In: Cervone D, Shoda Y (eds.) *The Coherence of Personality: Social-Cognitive Bases of Consistency, Variability, and Organization*. Guilford Press, New York, pp. 185–241

Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80

Buss D M 1997 Evolutionary foundations of personality. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 318–44

Caprara G V, Cervone D 2000 *Personality: Determinants Dynamics, and Potentials*. Cambridge University Press, New York

Cervone D 1991 The two discipline of personality psychology. *Psychological Science* **2**: 371–77

Cervone D, Shoda Y 1999 (eds.) *The Coherence of Personality*. Guilford Press, New York

Costa P T, McCrae R R 1997 Longitudinal stability of adult personality. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 269–90

De Raad B 1998 Five big five issues: rationale, structure, status and cross-cultural assessment. *European Psychologist* **3**: 113–24

Eysenck H 1970 The Structure of Personality, 3rd edn, Methuen, London

Goldberg L 1993 The structure of phenotypic personality traits. *American Psychologist* **48**: 26–34

Hettema J 1979 *Personality and Adaptation*. North-Holland, Amsterdam

Lohelin J C 1992 *Genes and Environment in Personality Development*. Sage, Newbury Park, CA

Magnusson D, Stattin H 1998 Person-context interaction theories. In: Damon W, Lerner R R M (eds.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1. pp. 685–760

McCrae R R, Costa P T 1996 Toward a new generation of personality theories: theoretical contexts for the five-factor model. In: Wiggins J S (ed.) *The Five-Factor Model of Personality. Theoretical Perspectives*. Guilford Press, New York, pp. 51–87

Murray H A 1938 *Explorations in Personality*. Oxford University Press, New York

Stagner R 1937 *Psychology of Personality*. McGraw Hill, New York

Stern W 1935 *Allgemeine Psychologie auf Personalischer Grundlage*. Nijhoff, Dordrecht

Wiggins J S (ed.) 1996 *The Five-Factor Model of Personality. Theoretical Perspectives*. Guilford Press, New York

G. V. Caprara

Personality and Conceptions of the Self

A central aspect of personality involves the ways in which people perceive and think about themselves.

Unlike most nonhuman animals, human beings possess a capacity for self-reflexive thinking that allows them to think consciously about themselves in complex and abstract ways. People differ in the degree to which they self-reflect, the content and organization of their self-conceptions, the positivity of their self-evaluations, and the selves they desire and fear to become in the future. These differences in how people think about, characterize, and evaluate themselves underlie many of the characteristic patterns of thought, emotion, and behavior that comprise the human personality.

1. *Self-attention*

People differ in the amount of time that they spend attending to and thinking about themselves. Individual differences in self-attention have been studied most extensively in terms of the trait of self-consciousness. (As used in this context, self-consciousness refers only to the tendency to attend to oneself and does not involve the anxiety and awkwardness that everyday use of the term connotes.) People who are high in private self-consciousness—the tendency to think about inner, private aspects of oneself (such as one's motives, goals, attitudes, and feelings)—tend to behave more consistently with their attitudes, values, and goals than people who are low in private self-consciousness, presumably because self-attention increases people's adherence to their personal standards. Because their behavior is strongly influenced by their personal standards, people high in private self-consciousness display greater consistency in their behavior across different situations than those who are low in private self-consciousness (Carver and Scheier 1981).

Public self-consciousness—the tendency to think about the public, observable aspects of oneself (such as one's appearance, mannerisms, and image in other people's eyes)—is associated with the degree to which individuals try to control how they are perceived by other people (e.g., see *Impression Management, Psychology of*). Because they are more concerned with how they are perceived and evaluated by others, publicly self-conscious people conform more to group pressure, show greater concern for their physical appearance, and tend to be more socially anxious and embarrassable than people who are low in public self-consciousness (Carver and Scheier 1985).

Importantly, whether people are high in private or public self-consciousness has implications for how they perceive themselves. People high in private self-consciousness not only attend more to their private selves but also regard those private aspects of self as more important to who they are. In contrast, people high in public self-consciousness regard the public aspects of self as more salient and important.

2. *Self-concept*

The self-concept involves people's beliefs about their personal characteristics, not only beliefs about their traits, abilities, and physical attributes but also about their values, roles, and personal goals. People differ both in the content of their self-concepts (e.g., who and what they think they are) and in how the various elements of their self-concepts are organized. Again, these differences contribute to variations in personality.

Self-concepts develop through the individual's direct experience, as well as from the reflected appraisals of other people (e.g., see *Self-development in Childhood*; *Identity in Childhood and Adolescence*). Once formed, people's self-concepts exert a strong influence on their behavior. Perceiving oneself as a particular kind of person prescribes and proscribes certain actions. Seeing oneself as polite, for example, has implications for the behaviors that a person will and will not perform. As a result, people generally behave in ways that are consistent with the kind of person they think they are.

In addition, people's self-concepts provide them with information about the likely consequences of certain behaviors. Believing themselves to possess particular attributes leads people to expect certain outcomes to result from their actions. For example, a person whose self-concept included the trait 'intelligent' would have different expectations about performance in an academic setting than a person whose self-concept included the trait 'stupid.' Thus, the self-concept has implications for a person's sense of self-efficacy (e.g., see *Personality Development in Childhood*). Believing that one possesses particular attributes and abilities affects people's expectations regarding whether they ought to be capable of particular tasks, thereby influencing their motivation to engage in certain behaviors.

Research has supported James's (1890) notion that the self-concept is composed of multiple components. People's self-identities are composed both of personal aspects that characterize them without respect to other people (e.g., personal abilities, values, and goals) as well as social aspects that identify them with respect to their interpersonal roles and relationships (e.g., particular family relationships, friendships, and group memberships). People differ in the importance that they place on the personal versus social aspects of their self-identities (Cheek 1989), and these differences predict their behavioral choices and reactions to events. For example, people who place greater importance on the personal aspects of their identity seek jobs that allow them to fulfill personal goals (e.g., self-fulfillment, personal growth, expression of their values), whereas people who place greater emphasis on the social aspects of their identities seek jobs that provide desired social outcomes (e.g., recognition, an active social life, interpersonal relationships). People

with different prevailing aspects of identity also prefer different kinds of recreational activities—solitary activities that promote self-improvement for people who emphasize personal identity, and group activities that promote interpersonal relationships for people who emphasize social identity.

People generally try to maintain consistency among the various aspects of their self-concepts, as well as between their self-concepts and reality. People tend to be troubled by contradictions between two aspects of their self-concepts and when objective information discredits their self-perceptions, possibly because such inconsistencies undermine people's certainty in their beliefs about themselves and their worlds (e.g., see *Cognitive Dissonance*). In fact, receiving information that is inconsistent with one's self-concept may be distressing even if it otherwise reflects positively on the person. As a result, people are generally motivated to behave in ways that are consistent with their self-concepts and prefer interacting with those who see them as they see themselves. Thus, people's self-concepts tend to steer their behavior in directions that help to maintain and verify their existing self-views (Swann et al. 1987).

Beyond the fact that people hold different ideas of who they are and what they are like, people's self-concepts also differ in their clarity, complexity, and differentiation. First, people differ in the clarity and certainty of their self-concepts. People whose views of themselves are clear, confident, and certain tend to score higher in self-esteem and positive affect, and lower in depression and anxiety, than people who have unclear, unstable views of themselves (Campbell et al. 1996). People whose self-concepts are unclear or uncertain are also more easily persuaded than those with clearer and more certain self-concepts, and uncertain self-concepts are more likely to change due to experience or social feedback.

Second, people differ in self-complexity—the number of discrete ways in which they characterize themselves and in the interrelatedness of the various aspects of their self-concepts. Although initial research suggested that people with greater self-complexity are less strongly affected by negative life experiences (Linville 1987), refinements of the construct show that self-complexity bears a more complex relationship with emotional and behavioral outcomes than was first assumed (Gramzow et al. 2000, Woolfolk et al. 1995).

Third, people differ in self-concept differentiation—the degree to which they see themselves as having different personality characteristics when they are in different social roles (Donahue et al. 1993). People low in self-concept differentiation tend to see themselves as relatively constant across various social roles, whereas highly differentiated people see themselves quite differently in different roles. Although differentiation might seem to be beneficial because it promotes greater behavioral flexibility, research suggests that high differentiation is generally undesirable, reflecting lack of an integrated core sense of self. Compared to people low in self-concept differentiation, highly differentiated people tend to be more depressed and anxious, lower in self-esteem, less agreeable, less satisfied with how they perform in their roles, and lower in self-control.

3. *Self-evaluations*

People not only see themselves as being a certain kind of person with particular sorts of attributes, but they also evaluate themselves in positive and negative ways, and experience corresponding positive and negative feelings when they think about who and what they are (e.g., see *Self-esteem in Adulthood*; *Self-evaluative Process, Psychology of*). Whereas self-concept refers to the *content* of people's beliefs about themselves, self-esteem refers to people's *evaluations* of those beliefs. Although people's evaluative feelings about themselves change across contexts and over time (i.e., state self-esteem), most research has focused on stable patterns of self-evaluation across situations (i.e., trait self-esteem). Researchers often refer to people having 'low' versus 'high' trait self-esteem, but, in reality, most people's trait self-esteem falls in the neutral to positive range, and relatively few individuals evaluate themselves negatively at a global level (Baumeister et al. 1989).

Trait self-esteem has proven to be an important psychological variable that predicts a great variety of emotions and behaviors. Self-esteem is generally correlated, albeit weakly, with indices of positive mental health such as low anxiety, neuroticism, and depression, and with a high sense of personal control and self-efficacy. However, self-esteem is also associated with maladaptive behaviors such as nonproductive persistence, risk-taking (particularly when the person's ego is threatened), and a tendency to make unrealistically self-serving attributions. Controversy exists regarding why self-esteem is related to behavior, thought, and emotion as it is. Various theories suggest that self-esteem is a marker of social dominance (dominance theory), a gauge of social acceptance (sociometer theory), a buffer against existential anxiety (terror management theory), and a reflection of personal autonomy (self-determination theory) (Hoyle et al. 1999).

People differ in the degree to which their positive and negative self-evaluations are confined to particular aspects of their self-concepts vs. spread across many aspects of the self. For people with highly compartmentalized self-concepts, positive and negative information about the self is confined to different aspects of the self; for people with self-concepts low in compartmentalization, positive and negative information is mixed across many aspects of self. Thus, two people may have an equal number of aspects of self

that they evaluate positively versus negatively yet differ in the degree to which these aspects are compartmentalized. Compartmentalization is associated with positive emotions and high self-esteem for people whose positive attributes are more important to them than their negative attributes. However, compartmentalization is associated with negative affect and low self-esteem for people whose negative attributes are more important (Showers 1992).

People differ also in the stability of their self-esteem. For some people, feelings of self-esteem change very little as they move from one situation to another, whereas the self-esteem of other people is quite labile, even over short spans of time. Thus, two individuals with the same general level of self-esteem may differ markedly in the degree to which their self-feelings vacillate. People with unstable self-esteem appear to possess fragile and vulnerable feelings of self-worth that are easily influenced by their experiences. They also tend to be more prone to anger and aggression when their self-esteem is threatened (Kernis and Waschull 1995).

4. *Desired and Undesired Selves*

People differ not only in how they perceive themselves at the present time but also in how they wish to be in the future. The human capacity for self-reflection allows people to imagine themselves in the future and to adjust their present behavior in ways that move them toward desired selves and away from undesired selves. These mental representations of desired and undesired future selves serve as incentives that motivate action, and as guides that channel behavior in particular directions (Markus and Nurius 1986). Interestingly, the future selves that people fear appear to exert greater power over their behavior than the selves that they desire (Ogilvie 1987).

A great deal of human emotion is a reaction to the fortunes of people's desired and undesired selves (Leary in press). People compare themselves and their outcomes in life to their desired and undesired selves, and experience various emotions depending on whether they are moving toward or away from their desired selves. Events that have positive implications for future desired selves evoke positive reactions, whereas events that have negative implications for desired selves (or, worse, indicate movement toward undesired future selves) cause emotional upset.

Self-discrepancy theory proposes that people's emotional reactions are strongly influenced by two particular sets of concepts about the desired self (Higgins 1987). According to the theory, people compare themselves to both an ideal self (how they would like to be) and an ought self (how they think they should be). A discrepancy between one's actual self and ideal self leads to dejection-related emotions such as sadness and disappointment, whereas a discrepancy between actual and ought selves leads to agitated emotions such as anxiety and guilt. Research has generally supported the notion that self-discrepancies produce emotion, although not always precisely in the manner predicted by self-discrepancy theory.

5. *Future Directions*

Theorists have long recognized that people's real, desired, and feared self-conceptions underlie aspects of their personalities, and a great deal of research has investigated various relationships between the self and personality. However, the field lacks an overriding conceptual framework for understanding how different aspects of the self interrelate and how these aspects relate to personality structures and processes. An impending task for theorists and researchers is to integrate what is known about self-relevant aspects of personality within a single theory.

See also: James, William (1842–1910); Personality Development in Adulthood; Personality Psychology; Self-development in Childhood; Self-monitoring, Psychology of

Bibliography

Baumeister R F 1998 The self. In: Gilbert D, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*. Oxford University Press, New York, pp. 680–740

Baumeister R F, Tice D M, Hutton D G 1989 Self-presentational motivations and personality differences in self-esteem. *Journal of Personality* **57**: 547–79

Campbell J D, Trapnell P D, Heine S J, Katz I M, Lavallee L F, Lehman D R 1996 Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology* **70**: 141–56

Carver C S, Scheier M F 1981 *Attention and Self-regulation: A Control Theory Approach to Human Behavior*. Springer-Verlag, New York

Cheek J M 1989 Identity orientations and self-interpretation. In: Buss D M, Cantor N (eds.) *Personality Psychology*. Springer-Verlag, New York, pp. 275–85

Donahue E M, Robins R W, Roberts B W, John O P 1993 The divided self: Concurrent and longitudinal effects of psychological adjustment and social roles on self-concept differentiation. *Journal of Personality and Social Psychology* **64**: 834–46

Gramzow R H, Sedikides C, Panter A T, Insko C A 2000 Aspects of self-regulation and self-structure as predictors of perceived emotional distress. *Personality and Social Psychology Bulletin* **26**: 188–205

Higgins E T 1987 Self-discrepancy: A theory relating self and affect. *Psychological Review* **94**: 319–40

Hoyle R H, Kernis M H, Leary M R, Baldwin M W 1999 *Selfhood: Identity, Esteem, Regulation*. Westview Press, Boulder, CO

James W E 1890 *The Principles of Psychology*. Henry Holt, New York

Kernis M H, Waschull S B 1995 The interactive roles of stability and level of self-esteem: Research and theory. In: Zanna M (ed.) *Advances in Experimental Social Psychology*. Academic Press, San Diego, CA, pp. 93–141

Leary M R in press The self and emotion: The role of self-reflection in the generation and regulation of affective experience. In: Salovey P (ed.) *Handbook of Affective Science*. Wiley, New York

Linville P 1987 Self-complexity as a cognitive buffer against stress-related depression and illness. *Journal of Personality and Social Psychology* **52**: 663–76

Markus H, Nurius P 1986 Possible selves. *American Psychologist* **41**: 954–69

Ogilvie D M 1987 The undesired self: A neglected variable in personality research. *Journal of Personality and Social Psychology* **52**: 379–85

Showers C 1992 Compartmentalization of positive and negative self-knowledge: Keeping bad apples out of the bunch. *Journal of Personality and Social Psychology* **62**: 1036–49

Swann W B Jr., Griffin J J, Predmore S, Gaines E 1987 The cognitive-affective crossfire: When self-consistency confronts self-enhancement. *Journal of Personality and Social Psychology* **52**: 881–89

Woolfolk R L, Novalany J, Gara M A, Allen L A, Polino M 1995 Self-complexity, self-evaluation, and depression: An examination of form and content within the self-schema. *Journal of Personality and Social Psychology* **68**: 1108–20

M. R. Leary

Personality and Crime

Psychology assumes that behavior arises from the interaction between the individual and the environment. Studies show that behavior is remarkably consistent over time; or, to be more precise, the relative ordering of individuals is remarkably consistent over time (Roberts and Del Vecchio 2000). It is assumed that behavioral consistency depends primarily on the persistence of underlying tendencies to behave in particular ways in particular situations. These tendencies are termed personality traits, such as impulsiveness, excitement seeking, assertiveness, modesty, and dutifulness. Larger personality dimensions such as extraversion refer to clusters of personality traits.

Many constructs assumed to underlie behavior are not included under the heading of personality. This is true, for example, of intelligence and moral reasoning abilities, and of cognitive processes (thinking, reasoning, and decision making). However, the boundaries between personality and all these constructs are not well defined. This review will focus on the topics conventionally included under the heading of personality, and there is not space here to discuss underlying biological processes in any detail.

The main types of offending studied are: burglary, theft, violence, vandalism, fraud, and drug abuse. Since crimes are a subset of antisocial acts, any potential to commit crimes is probably part of a broader potential to commit antisocial acts, but the focus here is on crimes. The focus is also on normal personality, and as far as possible, discussions of related psychiatric conditions such as antisocial personality disorder, psychopathy, and attention deficit–hyperactivity disorder will be avoided. However, the boundary between normal and pathological behavior is not always clear, depending on judgments about how far normal functioning is impaired.

In studying the relationship between personality and crime, it is important to avoid tautological statements and tautological results. Any construct that is presumed to cause crime must measure something different from crime. Thus, it is useful to investigate how far low empathy causes crime but less useful to investigate how far aggressiveness causes violent crime or how far an antisocial personality causes crime. Early research using the psychopathic deviate scale of the Minnesota Multiphasic Personality Inventory and the socialization scale of the California Psychological Inventory was essentially tautological in showing that antisocial behavior was related to crime.

A bewildering number of personality traits and dimensions have been postulated, measured, and related to offending. In some cases, different terms are probably used to describe the same underlying construct. Only the most important topics can be reviewed here: temperament, impulsiveness, empathy, the Eysenck theory, and the 'Big Five' personality dimensions.

1. *Temperament*

Temperament is basically the childhood equivalent of personality, although there is more emphasis in the temperament literature on constitutional predisposition and genetic and biological factors. The modern study of child temperament began with the New York Longitudinal Study of Chess and Thomas (1984). Children in their first 5 years of life were rated on temperamental dimensions by their parents, and these dimensions were combined into three broad categories of easy, difficult and 'slow to warm up' temperament. Having a difficult temperament at age 3–4 years (frequent irritability, low amenability and adaptability, irregular habits) predicted poor adult psychiatric adjustment at age 17–24.

Remarkably, Bates (1989) found that mothers' ratings of difficult temperament as early as age 6 months (defined primarily as frequent, intense expressions of negative emotions), predicted mothers' ratings of child conduct problems between the ages of 3 and 6 years. Similar results were obtained in the Australian Temperament Project, which found that children who were rated as irritable, not amenable, or showing behavior problems at age 4–8 months tended to be rated as aggressive at age 7—8 years (Sanson et al.

1993). However, when information at each age comes from the same source, it is possible that the continuity lies in the rater, rather than the child. Fortunately, other studies (e.g., Guerin et al. 1997) show that difficult temperament in infancy, rated by mothers, also predicts antisocial behavior in childhood rated by teachers.

Because it was not very clear exactly what a 'difficult' temperament meant in practice, other researchers have investigated more specific dimensions of temperament. For example, Kagan (1988) and his colleagues in Boston classified children as inhibited (shy or fearful) or uninhibited at age 21 months on the basis of their observed reactions to a strange situation, and found that they remained significantly stable on this classification up to age 7 years. Furthermore, the uninhibited children at age 21 months significantly tended to be identified as aggressive at age 13 years, according to self and parent reports (Schwartz et al. 1996).

The most important results on the link between childhood temperament and later offending have been obtained in the Dunedin longitudinal study in New Zealand, which has followed up over 1000 children from age 3 years and into their 20s (Caspi 2000). Temperament at age 3 years was rated by observing the child's behavior during a testing session involving cognitive and motor tasks. The most important dimension of temperament was being undercontrolled (restless, impulsive, with poor attention), which predicted aggression, self-reported delinquency, and convictions at age 18–21 years.

2. *Impulsiveness*

The most important study of childhood temperament in relation to offending essentially identified impulsiveness as the key dimension, and it is generally true that impulsiveness is the most crucial personality dimension that predicts offending. Unfortunately, there are a bewildering number of constructs referring to a poor ability to control behavior. These include impulsiveness, hyperactivity, restlessness, clumsiness, not considering consequences before acting, a poor ability to plan ahead, short time horizons, low self-control, sensation-seeking, risk-taking, and a poor ability to delay gratification. Virtually all these constructs, measured in different ways, are consistently related to measures of offending (Blackburn 1993, pp. 191–196).

In a longitudinal study of over 400 London males from age 8 years to age 40 years, boys nominated by teachers as lacking in concentration or restless; those nominated by parents, peers, or teachers as the most daring or risk-taking; and those who were the most impulsive on psychomotor tests at age 8–10 years, all tended to become offenders later in life. Later self-report measures of impulsiveness were also related to offending. Daring, poor concentration, and restlessness all predicted both official convictions, and self-reported delinquency, and daring was consistently one of the best independent predictors (Farrington 1992).

The most extensive research on different measures of impulsiveness was carried out in another longitudinal study of males (the Pittsburgh Youth Study) by White et al. (1994). The measures that were most strongly related to self-reported delinquency at ages 10 and 13 were teacher-rated impulsiveness (e.g., 'acts without thinking'); self-reported impulsivity; self-reported under-control (e.g., 'unable to delay gratification'); motor restlessness (from videotaped observations), and psychomotor impulsivity. Generally, the verbal behavior rating tests produced stronger relationships with offending than the psychomotor performance tests, suggesting that cognitive impulsiveness (based on thinking processes) was more relevant than behavioral impulsiveness (based on test performance).

There have been many theories put forward to explain the link between impulsiveness and offending. One of the most popular suggests that impulsiveness reflects deficits in the executive functions of the brain, located in the frontal lobes (Moffitt 1990). Persons with these neuropsychological deficits will tend to commit offenses because they have poor control over their behavior, a poor ability to consider the possible consequences of their acts, and a tendency to focus on immediate gratification. There may also be an indirect link between neuropsychological deficits and offending, which is mediated by hyperactivity and inattention in school, and the resulting school failure. In discussing links between executive functions and offending, impulsiveness may be difficult to disentangle from intelligence, although Lynam and Moffitt (1995) argued that they were different constructs. Various biological theories of impulsiveness have also been proposed, focussing on either behavioral activation or behavioral inhibition (e.g., Newman and Wallace 1993).

3. *Empathy*

There is a widespread belief that low empathy is an important personality trait that is related to offending, on the assumption that people who can appreciate and/or experience a victim's feelings are less likely to victimize someone. This belief also underlies cognitive–behavioral-skills training programs that aim to increase empathy. However, its empirical basis is not very impressive. There are inconsistent results in this field, and measures of empathy are not well validated or widely accepted. There appears to be no systematic comparison of different empathy measures analogous to the research of White et al. (1994) with impulsiveness, and no longitudinal study relating early measures of empathy to later offending.

A distinction has often been made between cognitive empathy (understanding or appreciating other people's feelings) and emotional empathy (actually experiencing other people's feelings). Eisenberg et al. (1998) further distinguished between sympathy (feeling concern for another) and experiencing personal distress. Empathy can be measured in many different ways, including self-completed questionnaires, peer ratings and systematic observation. Unfortunately, results vary according to the method used (Miller and Eisenberg 1988).

The best studies of the 1990s that have related empathy to offending in relatively large samples are by Mak (1991), Kaukiainen et al. (1999) and especially Luengo et al. (1994). In Australia, Mak (1991) found that delinquent females had lower emotional empathy than nondelinquent females, but that there were no significant differences for males. In Finland, Kaukiainen et al. (1999) reported that empathy (cognitive and emotional combined) was negatively correlated with aggression (both measured by peer ratings). In Spain, Luengo et al. (1994) carried out the first project that related cognitive and emotional empathy separately to (self-reported) offending, and found that both were negatively correlated.

Some results obtained in research on psychopathy are relevant. The Psychopathy Checklist includes two correlated factors, the first measuring personality and the second measuring an antisocial life-style (Harpur et al. 1988). The first factor measures a cold, callous personality, lacking empathy, affect, guilt and remorse, and engaging in habitual lying and manipulation. This constellation of personality traits is correlated with conduct problems and delinquency. Further research on it is warranted, especially in prospective longitudinal studies.

4. *The Eysenck Theory*

Before 1990, the best-known research on personality and crime was probably that inspired by Eysenck's theory and personality questionnaires (Eysenck 1996). He viewed offending as natural and even rational, on the assumption that human beings were hedonistic, sought pleasure, and avoided pain. He assumed that delinquent acts, such as theft, violence, and vandalism were essentially pleasurable or beneficial to the offender. In order to explain why everyone was not a criminal, Eysenck suggested that the hedonistic tendency to commit crimes was opposed by the conscience, which was viewed as a fear response built up from childhood in a conditioning process.

On the Eysenck theory, the people who commit offenses are those who have not built up strong consciences, mainly because they have inherently poor conditionability. Poor conditionability is linked to Eysenck's three dimensions of personality, Extraversion (E), Neuroticism (N), and Psychoticism (P). People who are high on E build up conditioned responses less well, because they have low levels of cortical arousal. People who are high on N also condition less well, because their high resting level of anxiety interferes with their conditioning. Also, since N acts as a drive, reinforcing existing behavioral tendencies, neurotic extraverts should be particularly criminal. Eysenck also predicted that people who are high on P would tend to be offenders, because the traits included in his definition of psychoticism (emotional coldness, low empathy, high hostility, and inhumanity) were typical of criminals. However, the meaning of the P scale is unclear, and it might perhaps be more accurately labeled as psychopathy. Zuckerman (1989) suggested that it should be termed 'impulsive unsocialized sensation-seeking.'

A review of studies relating Eysenck's personality dimensions to official and self-reported offending concluded that high N (but not E) was related to official offending, while high E (but not N) was related to self-reported offending (Farrington et al. 1982). High P was related to both, but this could have been a tautological result, since many of the items on the P scale are connected with antisocial behavior, or were selected in light of their ability to discriminate between prisoners and nonprisoners. In the prospective longitudinal study of London boys, those high on both E and N tended to be juvenile self-reported offenders, adult official offenders and adult self-reported offenders, but not juvenile official offenders. These relationships held independently of other criminogenic risk factors, such as low family income, low intelligence, and poor parental child-rearing behavior. However, when individual items of the personality questionnaire were studied, it was clear that the significant relationships were caused by the items measuring impulsiveness (e.g., doing things quickly without stopping to think). Hence, it seems likely that research inspired by the Eysenck theory essentially confirms the link between impulsiveness and offending.

5. *The 'Big Five'*

Since 1990, the most widely accepted personality system has been the 'Big Five' or five-factor model (McCrae and Costa 1997). This suggests that there are five key dimensions of personality: Neuroticism (N), Extraversion (E), Openness (O), Agreeableness (A) and Conscientiousness (C). Openness means originality and openness to new ideas, Agreeableness includes nurturance and altruism, and Conscientiousness includes planning and the will to achieve. Openness and Conscientiousness seem related to intelligence, or at least to social or emotional intelligence. These dimensions are measured using a personality inventory called the NEO-PI. Controversially, McCrae et al. (2000) argued that these personality dimensions are biologically based tendencies that follow intrinsic

developmental pathways independently of environmental influences.

Because of its newness, the 'Big Five' personality model has rarely been related to offending. In Canada, Hart and Hare (1994) found that psychopathy was most strongly (negatively) correlated with Agreeableness and Conscientiousness. Similarly, in an Australian study Heaven (1996) showed that Agreeableness and Conscientiousness were most strongly (negatively) correlated with self-reported delinquency. Much the same results were obtained in the Pittsburgh Youth Study when the five dimensions were measured using mothers' ratings (John et al. 1994). It seems likely, therefore, that this pattern of results is replicable.

6. *Future Directions*

More research is needed on the theoretical linkages between personality constructs and offending and on their biological bases. Prospective longitudinal studies are especially needed in which personality traits are measured early in life and related to later offending, to investigate the relative importance of biological, individual, family, peer, and school influences on personality and crime. Different studies should be compared to investigate how far results are replicable across genders, races, and cultures (Caspi et al. 1994).

A key issue that has rarely been addressed is the relationship between personality dimensions and different types of offenders. It might be expected, for example, that violent offenders would be more impulsive than non-violent offenders. It would be useful to use personality traits to define types of people and hence move from a variable-based to a person-based approach (Krueger et al. 1994).

The main policy implication of research on personality and crime is the use of cognitive–behavioral skills training techniques to try to counteract high impulsivity and low empathy. Research on childhood temperament suggests that potential offenders can be identified early in life and that efforts should be made to change children with uninhibited and uncontrolled temperaments. Research on personality and crime should also be useful in matching types of people to types of interventions. It seems likely that different personality types would interact differently with different criminological risk factors, and hence that interventions should be tailored to personalities. This is the new frontier.

See also: Adulthood: Prosocial Behavior and Empathy; Aggression in Adulthood, Psychology of; Antisocial Behavior in Childhood and Adolescence; Crime and Gender; Crime: Sociological Aspects; Crime, Sociology of; Neuroticism; Personality Theory and Psychopathology; Prosocial Behavior and Empathy: Developmental Processes; Temperament and Human Development; Violence and Effects on Children

Bibliography

Bates J E 1989 Applications of temperament concepts. In: Kohnstamm G A, Bates J E, Rothbart M K (eds.) *Temperament in Childhood*. Wiley, Chichester, UK, pp. 321–55

Blackburn R 1993 *The Psychology of Criminal Conduct*. Wiley, Chichester, UK

Caspi A 2000 The child is father of the man: Personality continuities from childhood to adulthood. *Journal of Personality and Social Psychology* **78**: 158–72

Caspi A, Moffitt T E, Silva P A, Stouthamer-Loeber M, Krueger R F, Schmutte P S 1994 Are some people crime-prone? Replications of the personality–crime relationship across countries, genders, races, and methods. *Criminology* **32**: 163–95

Chess S, Thomas A 1984 *Origins and Evolution of Behavior Disorders: From Infancy to Early Adult Life*. Brunner/Mazel, New York

Eisenberg N, Fabes R A, Shepard S A, Murphy B C, Jones S, Guthrie I K 1998 Contemporaneous and longitudinal prediction of children's sympathy from dispositional regulation and emotionality. *Developmental Psychology* **34**: 910–24

Eysenck H J 1996 Personality and crime: Where do we stand? *Psychology, Crime and Law* **2**: 143–52

Farrington D P 1992 Juvenile delinquency. In: Coleman J C (ed.) *The School Years*, 2nd edn. Routledge, London, pp. 123–63

Farrington D P, Biron L, LeBlanc M 1982 Personality and delinquency in London and Montreal. In: Gunn J, Farrington D P (eds.) *Abnormal Offenders, Delinquency, and the Criminal Justice System*. Wiley, Chichester, UK, pp. 153–201

Guerin D W, Gottfried A W, Thomas C W 1997 Difficult temperament and behavior problems: A longitudinal study from 1.5 to 12 years. *International Journal of Behavioral Development* **21**: 71–90

Harpur T J, Hakstian A R, Hare R D 1988 Factor structure of the psychopathy checklist. *Journal of Consulting and Clinical Psychology* **56**: 741–7

Hart S D, Hare R D 1994 Psychopathy and the Big Five: Correlations between observers' ratings of normal and pathological personality. *Journal of Personality Disorders* **8**: 32–40

Heaven P C L 1996 Personality and self-reported delinquency: Analysis of the 'Big Five' personality dimensions. *Personality and Individual Differences* **20**: 47–54

John O P, Caspi A, Robins R W, Moffitt T E, Stouthamer-Loeber M 1994 The 'Little Five': Exploring the nomological network of the five-factor model of personality in adolescent boys. *Child Development* **65**: 160–78

Kagan J, Reznick J S, Snidman N 1988 Biological bases of childhood shyness. *Science* **240**: 167–71

Kaukiainen A, Bjorkvist K, Lagerspetz K, Osterman K, Salmivalli C, Rothberg S, Ahlbom A 1999 The relationships between social intelligence, empathy, and three types of aggression. *Aggressive Behavior* **25**: 81–9

Krueger R F, Schmutte P S, Caspi A, Moffitt T E, Campbell K, Silva P A 1994 Personality traits are linked to crime among men and women: Evidence from a birth cohort. *Journal of Abnormal Psychology* **103**: 328–38

Luengo M A, Otero J M, Carrillo-de-la-Pena M T, Miron L 1994 Dimensions of antisocial behavior in juvenile delinquency: A study of personality variables. *Psychology, Crime and Law* **1**: 27–37
Lynam D R, Moffitt T E 1995 Delinquency and impulsivity and IQ: A reply to Block. *Journal of Abnormal Psychology* **104**: 399–401
Mak A S 1991 Psychosocial control characteristics of delinquents and nondelinquents. *Criminal Justice and Behavior* **18**: 287–303
McCrae R R, Costa P T 1997 Personality trait structure as a human universal. *American Psychologist* **52**: 509–16
McCrae R R, Ostendorf F, Angleitner A, Hrebickova M, Avia M D, Sanz J, Sanchez-Bernardos M L, Kusdil M E, Woodfield R, Saunders P R, Smith P B 2000 Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology* **78**: 173–86
Miller P A, Eisenberg N 1988 The relation of empathy to aggressive and externalizing/antisocial behavior. *Psychological Bulletin* **103**: 324–44
Moffitt T E 1990 The neuropsychology of juvenile delinquency: A critical review. In: Tonry M, Morris N (eds.) *Crime and Justice*. University of Chicago Press, Chicago, Vol. 12, pp. 99–169
Newman J P, Wallace J F 1993 Diverse pathways to deficient self-regulation: Implications for disinhibitory psychopathology in children. *Clinical Psychology Review* **13**: 699–720
Roberts B W, Del Vecchio W F 2000 The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin* **126**: 3–25
Sanson A, Smart D, Prior M, Oberklaid F 1993 Precursors of hyperactivity and aggression. *Journal of the American Academy of Child and Adolescent Psychiatry* **32**: 1207–16
Schwartz C E, Snidman N, Kagan J 1996 Early childhood temperament as a determinant of externalizing behavior in adolescence. *Development and Psychopathology* **8**: 527–37
White J L, Moffitt T E, Caspi A, Bartusch D J, Needles D J, Stouthamer-Loeber M 1994 Measuring impulsivity and examining its relationship to delinquency. *Journal of Abnormal Psychology* **103**: 192–205
Zuckerman M 1989 Personality in the third dimension: A psychobiological approach. *Personality and Individual Differences* **10**: 391–418

D. P. Farrington and D. Jolliffe

Personality and Health

1. Associations between Personality and Health: Background

It is common folk wisdom that people who worry too much are likely to develop headaches, that nervous people are prone to ulcers, that excitable people are likely to develop skin rashes, and that overly sensitive people may be prone to nausea or fainting. Popular wisdom also has it that workers who strive compulsively for long hours at frustrating job tasks are more likely to have heart attacks, and that people who bottle up all their feelings inside them have an increased risk of cancer. Yet the scientific evidence suggests a different and more complex picture. Further, although there are indeed associations between personality and health, the causal links are not always what they first appear to be. This is important because the causal linkages tell us which interventions are helpful and which are worthless or even possibly harmful.

A significant ethical danger in studying the relations of personality and health is that individuals may be unfairly blamed for their health problems. Health and illness are complex states, usually influenced by a wide variety of biological, environmental, and psychosocial factors. Personality—the traits, motives, emotions, and behavioral patterns unique to each individual—is sometimes associated with illness, but it would be a grave oversimplification to think that the individual has direct responsibility for his or her health. On the other hand, individual behavior patterns sometimes do affect health-relevant processes, and so the person, the family, the environment, and the culture can sometimes promote the likelihood of better health.

1.1 Roots in Ancient Greece

The idea that personality and health might be related dates back thousands of years and clearly appears in the writings of Hippocrates, Galen, and their followers. The ancient Greeks, keen observers of the human condition, saw four elements, the so-called bodily humors, as key to both individuality and health. People with a healthy supply of blood might turn out to be sanguine and ruddy. Black bile (or melancholy) might lead to depression and degenerative diseases or cancer. Yellow bile (or choler) could produce a bitter, angry personality and associated diseases, and phlegm is characteristic of a phlegmatic, cold apathy. It is interesting that this scheme of four emotional aspects of personality and health was sufficiently perceptive (or ambiguous) to influence the practice of medicine for 2,000 years. Of course, notions of bodily humors have been discarded. Yet the idea that individuals can be categorized as sanguine, depressed, hostile, and repressed continues to stay with us. Modern personality and health psychology takes these chronic psycho-emotional states and combines them with a more sophisticated understanding of traits to study the associations among personality and health.

1.2 Type A Behavior

Much of the current research interest in personality and health began with theory and research on the Type A Behavior Pattern (Houston and Snyder 1988). Early in the twentieth century, a number of prominent physicians including William Osler began noting that

certain men seemed especially prone to angina and coronary heart disease. The at-risk characteristics supposedly included being ambitious, hard-working, dominant, and aggressive. In the 1930s and 1940s, the influence of psychoanalysis on psychosomatic medicine led to numerous speculations about the role of expression and repression of hidden conflicts and anger on the development of heart disease. But it was not until the mid-century writings on Type A behavior that empirical research on this topic began in earnest.

The term 'Type A' was chosen by its inventors Meyer Friedman and Ray Rosenman (e.g., Rosenman et al. 1975) because it was a nonconnotative, 'objective' designation for a so-called coronary-prone medical syndrome. Type A people were seen as struggling rapidly to accomplish many things in a short period of time, always with a deadline to meet. They were viewed as very competitive, active, and quite hostile or aggressive. Although any person can act this way from time to time, Type A people were seen as constantly driven. Individuals who did not fit this profile were labeled, by default, as Type B. That is, there was thus no independent definition of what a healthy personality style might be.

Although many hundreds of studies have been conducted on Type A behavior, the results have often been mixed and confusing. Although the initial aim was to simplify the concept by choosing a neutral term ('Type A') and avoiding related psychological concepts and theories, the characteristics of individuals cannot be adequately explained in such a sterile manner. That is, psychologists soon turned to trying to understand the trait correlates of Type A behavior, the components of Type A behavior, the basis of Type A behavior, the consequences of Type A behavior, and the health aspects of Type A behavior, even beyond coronary disease (Friedman 1992, Houston and Snyder 1988, Miller et al. 1996).

It is now clear that the fast speed and the job-involvement aspects are not important predictors of coronary disease. However, excessive competitiveness and constant hostility seem, for some people in some circumstances, to increase the likelihood of coronary heart disease. It is not yet clear the extent to which these characteristics also raise the risk of all-cause mortality, since there has been insufficient attention to construct validity.

The idea of Type A behavior proved to be an important heuristic one, leading to much empirical research since the 1970s about whether, when, why, and how personality plays a role in health. There are, however, a variety of possible mediating mechanisms.

2. *Ways that Personality is Related to Health: Causal Links*

There are many different reasons why personality is associated with health. Understanding the causal pathways is very important, because an interest in this subject usually derives from a desire to improve or promote health. However, it turns out that personality can affect health, that health can affect personality, or that the association between them can be the result of several sorts of artifacts.

2.1 *Symptom Reporting and Medical Usage*

It is well documented that personality is related to symptom reporting, symptom sensation, and entering the medical care system. Certain people, especially people who are neurotic, with anxiety and stress, are more likely to report symptoms including headache, gastrointestinal distress, and other assorted aches and pains. In other words, if questioned closely, other, non-neurotic people would be experiencing the same symptoms but would not be as likely to talk about them or to report them to their physicians.

It is also the case that people with anxiety and stress are more likely to feel pain and other symptoms. That is, they have a lower threshold for experiencing pain and a lower tolerance for pain. They also may be especially vigilant about bodily sensations (Friedman 1991/2000, Pennebaker 1982).

People with certain psychological traits are also more likely to enter the health care system. These people are generally characterized by such traits as anxiety, depression, dependency, hostility, and moodiness. Women are also generally more likely to seek health care than men. In the extreme, there are people with factitious disorders who seek health care for the rewards that they receive there, including attention, drugs, and shelter (Costa and McCrae 1987, Friedman 1990).

In all of these cases, there are associations between personality and ill health, but the association results from a greater tendency to report and perceive symptoms and seek treatment rather than from a greater incidence of organic disease. In other words, in many cases, certain people are more likely to report being 'ill,' but they do not differ much from other people in their incidence of clear organic diseases like heart disease or cancer.

2.2 *Disease-caused Personality Changes*

Many associations between personality and health are the result of disease processes that affect health. Certainly any disease with a significant impact on the brain, including Alzheimer's, Parkinson's, strokes, syphilis, AIDS, and brain tumors can produce associations between personality and health. Since the changes in personality often become apparent long before the changes in organic brain function are examined and diagnosed, it may appear that the personality is predicting or even causing the disease

when the reality is that the disease is changing the personality (Woodward 1998).

In many people, the processes are likely even more subtle and complex. Various illnesses affect the blood supply to the brain, the neurotransmitters in the brain, and communication from other parts of the body to the brain (Woodward 1998). This may result in various associations between personality and health, especially involving depression or other mood changes. For example, low levels of thyroid hormones or lessened oxygen in the blood can cause depression. Importantly, many if not most prescription and non-prescription drugs, being used to treat illness, can affect the brain and thus affect mood and personality, sometimes dramatically, sometimes subtly.

2.3 Temperamental Underlying Third Variables

Some of the associations between personality and health result from a constitutional predisposition underlying certain aspects of personality and certain threats to health. For example, some people's cardiovascular systems are hyper-reactive, meaning that their blood pressure and heart rate may show marked increases when the individual is under stress. As part of this biology, such people are more likely to be introverted and neurotic (aspects of personality) and are also more likely to suffer diseases such as hypertension. In such cases of an underlying third variable, it is not necessarily the case that efforts to change personality (such as helping an introvert become more sociable) will have any effect on health risk if the underlying physiology is not altered (Eysenck 1991, Friedman 1990, Zuckerman 1991).

Similarly, a hostile temperament, which may result from a combination of genetics, early hormone exposures, and early experiences, could lead to an association between a choleric personality and health risks. Hans Eysenck has argued that such psychobiological systems produce many associations between individual characteristics and disease, but also that the psychobiological characteristics are modifiable, thus producing changes in both personality and disease risk.

2.4 Personality-caused Illness

Probably the most interesting link between personality and health arises when personality has a fairly direct, causal influence on health. For example, chronic worrying and distress somehow may increase the likelihood of headaches. Since we cannot randomly assign people to personality and do conclusive experiments, inferences about how personality affects health must necessarily be inexact. It is important to employ a wide variety of studies ranging from epidemiological studies of psychological patterns and health down to biological studies of the physiological correlates of personality. The findings must then be pieced together.

3. Explanations of Personality-caused Illness

In those cases in which personality affects health, it is useful to distinguish among several primary causal pathways. However, given that these influences occur across time, it is often the case that these mechanisms do not function independently; rather they feed back on and interact with each other (Antonovsky 1979, Bolger and Zuckerman 1995, Friedman 1991/2000).

3.1 Poor Coping

Personality is related to coping, the ways in which an individual contends with stress. New thoughts and behaviors are constantly necessary to adapt to new challenges. With successful coping, the individual interprets and understands the challenge, draws on helpers to address it, and takes appropriate actions. Depending on the challenge, individuals who are depressed, introverted, repressed, unconscientious, vain, or conflicted may be less successful in bringing to bear necessary psychological, social, and behavioral resources (Aspinwall and Taylor 1997, Friedman 1991/2000).

A common issue involves relying on unproductive psycho-emotional ways of coping when a behavioral action could be used to solve the problem or source of stress. For example, a student might go party rather than study to deal with the challenge of an upcoming major exam. On the other hand, an individual might attempt to change the unchangeable. For example, after injury or loss, a person may need to find psychological ways of dealing with his or her new situation in order to cope effectively. Poor coping can lead either to unhealthy behaviors or unhealthy physiological reactions, or both (Maddi and Kobasa 1984, Pennebaker 1995).

3.2 Unhealthy Behaviors

It is well documented that certain personalities in certain situations are more likely to engage in unhealthy behaviors such as smoking, drinking, drug abuse, fast driving, fighting, and so on. Certain people are also less likely to engage in prophylactic measures such as using sunscreen, wearing seat belts, using condoms, brushing teeth, regularly washing hands, and so on. Some individuals are more or less likely to visit physicians, keep follow-up appointments, follow medical regimens, or have screening tests. Since these associations usually depend on at least several factors about the individual, the situation, and the unhealthy

activity, it is not usually possible to offer simple, broad generalizations.

3.3 Stress and Physiological Reactivity

Psychophysiological mechanisms often come to mind when one first considers links between personality and health. This is the idea that personality and stress directly impact internal physiological processes, thus causing disease. A more general view here is that the disruption upsets bodily homeostasis, thus allowing disease to take hold.

Much of this work dates back to Walter Cannon's writings on the 'fight-or-flight' response in which catecholamines are released and affect organs and systems throughout the body. Catecholamines (such as epinephrine) can affect processes from heart rate and blood pressure to the metabolism of fats. More recently, it has been established that the immune system is likewise affected by stress hormones. It is well documented that certain people are more likely than others to react to challenge with significant psychophysiological stress reactions (Denollet 1998, Friedman 1992, Kiecolt-Glaser et al. 1998, Zuckerman 1991).

3.4 Tropisms: Seeking Unhealthy Situations

One of the least understood but most intriguing sources of the link between personality and disease involves unhealthy situations. Some individuals gravitate towards healthier and health-promoting psychosocial environments, while other individuals find themselves pulled towards unhealthy environments. In these instances, it is not the case that personality leads directly to unhealthy behaviors or to unhealthy psychophysiological reactions. Rather, the personality facilitates entering the unhealthy situation.

Some of this motivation is derived from genetics, temperament, and early experiences. Other motivation arises from family experiences and consistent peer pressure. People who are impulsive, hostile, psychologically unstable, and have related difficulties in self-regulation are more likely to find themselves in environments that promote smoking, alcoholism, drug abuse, interpersonal violence, promiscuity, and risk-taking hobbies. Furthermore, neuroticism tends to predict to negative life events; in other words, neurotic people are more likely to experience objectively more stressful events (Taylor et al. 1997). Others, who are well socialized, conscientious, and agreeable, are more likely to wind up well educated and surrounded by mature adults. Note that someone who was genetically predisposed to alcoholism or other addiction is very unlikely to develop the health-impairing problem if raised in and exposed to very healthy psychosocial environments.

4. Disease-prone Traits

Although studies rarely assess a variety of personality traits, a variety of likely mediating mechanisms, and a variety of health outcomes, efforts to piece together the results of different types of research yield a sense of which characteristics are more or less likely to lead to good health.

4.1 Extraversion

Various theories suggest that extraversion and its components of dominance, sociability, and talkativeness should be related to good health, but the evidence is sparse. It is certainly the case that people who are well integrated into their communities, have friends, and have good relations with others are more likely to be healthier. Such social support is not, however, necessarily derived from being highly extroverted. Closer analysis reveals many people who are not especially extroverted but who have important, stable social relationships.

4.2 Optimism

Derived from popular notions of the importance of 'positive thinking,' a good deal of research has examined whether optimism is related to health. Optimism is the general expectation that good things will happen. To the extent that optimists may try harder when challenged, take better care of themselves, and cope better with the adversity of illness, optimism may sometimes be a characteristic of good health. For example, dispositional optimists tend to have a faster rate of recovery from coronary bypass surgery (Aspinwall and Taylor 1997, Scheier et al. 1989). Faster recovery with a positive outlook may also be a result of improved immune system functioning as stress is reduced (Kiecolt-Glaser et al. 1998).

However, there are also many reasons why optimism is not health protective and may be health damaging. The other side of the coin of optimism is 'optimistic bias' or unrealistic optimism (Weinstein 1984). This bias involves the faulty belief that negative events are less likely to happen to the optimist than to others. For example, one might agree that 'people' who smoke and who do not eat fruits and vegetables are more likely to get cancer, but not worry at all whether the relationship applies to oneself. Importantly, someone who is too optimistic may not take precautions or may not cooperate fully with medical treatment (Peterson et al. 1998, Tennen and Affleck 1987).

4.3 Neuroticism, Hostility, Depression

It is very well established that individuals who are more anxious, hostile, or depressed are more likely to

report being ill and seek medical care. Much research controversy has arisen over the extent to which and the circumstances under which these are causal mechanisms, with these personality traits causing disease. As noted, such individuals are more likely to feel, report, and treat symptoms. It is also often the case, however, that these symptoms result from true organic disease, promoted by psychological causes.

The best evidence concerns hostility, which was first investigated in an attempt to better understand coronary proneness. Chronic hostility (and chronic arousal of the sympathetic nervous system) raises the risk of disease and interferes with recovery (Miller et al. 1996).

There is also mounting evidence of the role of chronic depression in increasing the likelihood of development of various illnesses. Depression is associated with cortisol dysregulation, which is also linked to disruptions in lipid metabolism. It is not yet clear whether this is closely tied to hostility or is a different mechanism.

The aspect of neuroticism that is characterized by alienation, anxiety, and pessimism can produce various unhealthy effects. Such people can turn to substance abuse or overeating in an attempt to regulate their anxious states. They may develop insomnia. They may avoid interpersonal assistance and medical assistance. This may lead to a cycle of increasing anxiety, isolation, and health problems. On the other hand, someone whose neuroticism leads to a vigilance about avoiding threats, attending to symptoms, and cooperating with treatment may gain some health benefits.

Some researchers focus more generally on chronic negative affect, which includes both emotional distress like anxiety and depression as well as social interaction such as introversion, alienation, and the inability to express oneself to others. This distressed personality is sometimes referred to as 'Type D' (Denollet 1998).

4.4 *Conscientiousness and Impulsivity*

With significant attention to variables such as chronic hostility and Type A competition, it is perhaps not surprising that certain key aspects of a healthy personality have not been uncovered until recently. There is, however, rapidly growing evidence that conscientiousness and a lack of impulsivity is a key characteristic of healthy individuals. Conscientious people, who are prudent, planning, persistent, and dependable, tend to succeed in various life tasks, including staying healthy.

One major study of individuals who have been followed from childhood until their deaths in adulthood, conducted by Howard Friedman and his colleagues, has repeatedly found evidence that conscientiousness is the key personality predictor of longevity. Interestingly, the protective effect of conscientiousness does lessen the likelihood of death from accident and injury, but it is also protective against death from diseases such as cardiovascular disease and cancer. Conscientiousness does not seem to have its effects simply through one or two behavioral pathways, but rather it seems to affect a whole host of threats to health (Friedman et al. 1995). This finding of the importance of conscientiousness has been confirmed by subsequent studies by others which show that conscientiousness is associated with healthy behaviors, better self-care, better cooperation with medical treatment, and risk avoidance.

4.5 *Repression, Lack of Verbalization, and Lack of Disclosure*

It has long been recognized that among the many people who seem disease-prone, not all are hostile, impulsive, or neurotic. There are some who are too stoic, apathetic, and unemotional. These phlegmatic people, who look cool, calm, and sluggish, may be consciously or unconsciously suppressing or repressing inner conflict or feelings. This phenomenon is not well understood, but such people may appear tense or anxious upon close observation, although they assert friendliness and cooperation. Such people are often alienated or feel powerless.

A behavior pattern involving a repressive or denying coping style, lack of emotional expression, stress and anxiety, and helplessness/hopelessness is sometimes referred to as 'Type C.' A similar pattern in which individuals are unable to express emotion and seem at higher risk of disease progression is sometimes studied under the rubric of 'alexithymia.' There is speculation that such conditions promote the development or the progression of some cancers, but the relationship of this constellation of traits to other illnesses has not been much studied.

James Pennebaker has proposed that inhibition is a form of physiological work that increases the risk of illnesses. On the other hand, people who can talk about emotional experiences seem to set in motion processes to better health. Verbalizing may contribute to cognitive assimilation of stress and/or help people understand and organize their experiences. It may also often lead to increases in levels of social support.

5. *Self-healing Personality*

Although most theory and research has focused on disease-prone personalities, it is equally important to examine the characteristics that are associated with good health and longevity. These characteristics are not necessarily simply the absence of the disease-prone traits.

The 'Self-healing personality' refers to a healing emotional style involving a match between the in-

dividual and the environment, which maintains a physiological and psychosocial homeostasis, and through which good mental health promotes good physical health. The term was proposed by psychologist Howard S. Friedman in his 1991 book of the same name (Friedman 1991/2000). Self-healing, emotionally balanced people may be calm and conscientious but are responsive, energetic, curious, secure, and constructive. They are also people one likes to be around. People with self-healing personalities have better odds of a longer, healthier life.

Self-healing individuals often appear enthusiastic, and they may move in smooth, synchronized styles. Their relative lack of emotional conflict and anxiety may result in fewer speech disturbances, modulated voice tones, and an emotional consistency. Contrary to some common wisdom, a self-healing style does not involve a lack of challenge. In fact, self-healing individuals are often found in challenging, socially valued careers. They may enjoy stimulating travel because they are curious and enjoy exploration.

Interestingly, self-healing personalities are often similar to the mentally healthy orientations that have been described by humanistic psychologists (Antonovsky 1979, Csikszentmihalyi 1991, Maddi and Kobasa 1984). Humanistic psychologists such as Abraham Maslow thought that they were describing mental health when they wrote about self-actualized people who are concerned with issues of beauty, ethics, and justice. But it turns out that this spontaneous, positive, growth orientation is also associated with physical health. Indeed, it is a false dichotomy to divide physical and mental health in these ways. This positive process is sometimes also referred to by Mihaly Csikszentmihalyi as 'flow.'

5.1 *Efficacy and Hardiness*

Work by Salvador Maddi and Suzanne Ouellette Kobasa helped provide a framework for thinking about staying healthy in the face of challenge. They call these personality attributes hardiness. First, they suggest that a healthy personality maintains a sense of control. This is not necessarily a wild sense of optimism, but rather a sense that one can control one's own behaviors. Second, there is a commitment to something that is important and meaningful in their lives. This may also involve values and goals. Third, hardy people welcome challenge. For example, they may view change as an exciting challenge to their growth and development. Efficacy and hardiness generally involve a productive orientation, a sense of mastery, and a zest for life (Bandura 1997).

5.2 *Salutogenesis*

The medical sociologist Aron Antonovsky (1979) proposed an approach termed 'salutogenesis,' a theory of how and why certain people stay healthy. What is interesting about this approach is that it does not postulate a sense of control as a primary or even necessary characteristic of a healthy personality. Rather, Antonovsky refers to a 'sense of coherence' which is central to successful coping with challenge. Coherence means that the world is understandable, manageable, and meaningful. In this scheme, someone who believes that nature or a divine force is in charge of the world could have a very healthy orientation, even though the individual does not have a personal sense of control. There is so far only scattered evidence to support this promising approach.

6. *Conclusion*

In sum, certain individuals are predisposed to be vulnerable to physiological, psychosocial, and interpersonal disturbances and stresses due to a combination of genetics, emotional temperament, and early experiences. When such individuals seek out or are placed in environments that bring out, create, or exacerbate these disturbances, chronic stress patterns may result. These may be accompanied by unhealthy behaviors such as substance abuse. When these individuals then encounter microbes, toxins, or excesses such as high-fat diets, the likelihood of illness increases significantly. On the other hand, other individuals have developed processes of self-healing, which build on themselves as healthier emotional styles, physiological reactions, behaviors, and social interactions become mutually reinforcing.

Overall, there is good evidence that individual characteristics are sometimes tied to health through the biological, psychological, and social aspects of personality, and there is some evidence that focused interventions can have short-term salutary effects on health. But relatively little is understood about the precise causal inter-relations of physiology, coping, social influence, and health behavior, and little is known about the extent to which a focus on personality alone can have substantial effects on various significant health outcomes. It will be important in the future to unravel how these key processes come together and affect each other, both across short time periods and across the lifespan.

See also: Biopsychology and Health; Cancer-prone Personality, Type C; Control Beliefs: Health Perspectives; Depression, Hopelessness, Optimism, and Health; Emotions and Health; Explanatory Style and Health; Health: Self-regulation; Personality Assessment; Personality Psychology; Personality Theory and Psychopathology; Self-efficacy and Health

Bibliography

Antonovsky A 1979 *Health, Stress and Coping*. Jossey-Bass, San Francisco

Aspinwall L G, Taylor S E 1997 A stitch in time: Self-regulation and proactive coping. *Psychological Bulletin* **121**: 417–36
Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman and Co., New York
Bolger N, Zuckerman A 1995 A framework for studying personality in the stress process. *Journal of Personality and Social Psychology* **69**(5): 890–902
Costa P T, McCrae R R 1987 Neuroticism, somatic complaints, and disease: Is the bark worse than the bite? *Journal of Personality* **55**: 299–316
Csikszentmihalyi M 1990 *Flow: The Psychology of Optimal Experience*. HarperCollins, New York
Denollet J 1998 Personality and coronary heart disease: The Type-D Scale-16 (DS16). *Annals of Behavioral Medicine* **20**(3): 209–15
Eysenck H J 1991 *Smoking, Personality, and Stress: Psychosocial Factors in the Prevention of Cancer and Coronary Heart Disease*. Springer-Verlag, New York
Friedman H S (ed.) 1990 *Personality and Disease*. Wiley and Sons, New York
Friedman H S 1991/2000 *Self-healing Personality: Why Some People Achieve Health and Others Succumb to Illness*. Henry Holt, New York/(www.Iuniverse.com), 2000
Friedman H S (ed.) 1992 *Hostility, Coping, and Health*. American Psychological Association, Washington, DC
Friedman H S, Booth-Kewley S 1987 The 'disease-prone personality': A meta-analytic view of the construct. *American Psychologist* **42**: 539–55
Friedman H S, Tucker J S, Schwartz J E, Tomlinson-Keasey C, Martin L R, Wingard D L, Criqui M H 1995 Psychosocial and behavioral predictors of longevity: The aging and death of the 'Termites.' *American Psychologist* **50**: 69–78
Houston B K, Snyder C R 1988 *Type A Behavior Pattern: Research, Theory and Intervention*. Wiley, New York
Kiecolt-Glaser J K, Page G G, Marucha P T, MacCallum R C, Glaser R 1998 Psychological influences on surgical recovery. *American Psychologist* **53**: 1209–18
Maddi Salvatore R, Kobasa S C 1984 *The Hardy Executive: Health Under Stress*. Dow Jones-Irwin, Homewood, IL
Miller T Q, Smith T W, Turner C W, Guijarro M L, Hallet A J 1996 Meta-analytic review of research on hostility and physical health. *Psychological Bulletin* **119**: 322–48
Pennebaker J W (ed.) 1995 *Emotion, Disclosure, and Healing*. American Psychological Association, Washington, DC
Pennebaker J W 1982 *The Psychology of Physical Symptoms*. Springer-Verlag, New York
Peterson C, Seligman M E P, Yurko K H, Martin L R, Friedman H S 1998 Catastrophizing and untimely death. *Psychological Science* **9**: 127–30
Rosenman R H, Brand R J, Jenkins C D, Friedman M, Straus R, Wurm M 1975 Coronary heart disease in the Western Collaborative Group Study: Final follow-up experience of 81/2 years. *Journal of the American Medical Association* **233**: 872–7
Scheier M F, Magovern G J, Abbott R A., et al 1989 Dispositional optimism and recovery from coronary artery bypass surgery: The beneficial effects on physical and psychological well-being. *Journal of Personality and Social Psychology* **57**: 1024–40
Taylor S E, Repetti R L, Seeman T 1997 Health psychology: What is an unhealthy environment and how does it get under the skin? *Annual Review of Psychology* **48**: 411–47
Tennen H, Affleck G 1987 The costs and benefits of optimistic explanations and dispositional optimism. *Journal of Personality* **55**: 377–93
Weinstein N D 1984 Why it won't happen to me: Perceptions of risk factors and susceptibility. *Health Psychology* **3**: 431–57
Wiebe D J, Smith T W 1997 Personality and health: Progress and problems in psychosomatics. In: Hogan R, Johnson J, Briggs S (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, CA, pp. 891–918
Woodward J L 1998 Dementia. In: Friedman H S (ed.) *Encyclopedia of Mental Health*. Academic Press, San Diego, Vol. 1, pp. 693–713
Zuckerman M 1991 *Psychobiology of Personality*. Cambridge University Press, New York

H. S. Friedman

Personality and Marriage

The success or failure of a marriage is likely to depend, at least in part, on the personality of the individual spouses, broadly defined as the set of enduring traits and characteristics that each spouse brings to the relationship. Three questions have motivated most research on the role of personality in marriage. First, what personality traits are most relevant to predicting marital outcomes? Second, does similarity or dissimilarity between partners' personalities affect marital outcomes? Third, through what mechanisms does the personality of each spouse affect the relationship? This article reviews each of these areas of research in turn.

1. Which Personality Traits Affect Marital Outcomes?

The earliest efforts to predict marital outcomes were guided by the belief that 'a large proportion of incompatible marriages are so because of a predisposition to unhappiness in one or both of the spouses' (Terman 1938, p. 110). To explore the support for this belief, the first marital researchers conducted two kinds of studies: cross-sectional studies, in which lengthy personality inventories were administered to samples that included happily married and distressed couples, and longitudinal studies, in which personality inventories were administered to engaged couples whose marital quality was assessed several years later. Early marital researchers were more interested in determining whether spouses' personalities accounted for marital outcomes than in identifying which aspects of personality were associated with marital outcomes. Thus, the results of these initial studies were frequently reported in terms of a global personality score rather than divided into specific personality traits. Nevertheless, when Burgess and Wallin (1953), in the first review of this literature, concluded that success and failure in marriage could be predicted by the person-

alities of each spouse, they noted that personality traits associated with neuroticism (e.g., irritability, moodiness, ill-temper) appeared to be especially predictive. Spouses who reported an enduring tendency to experience negative affect across a variety of contexts seemed to be at greater risk of developing marital distress.

As these results accumulated, the same researchers noticed that other enduring characteristics of the spouses (e.g., childhood family environment, sexual history, demographic variables), were also significant predictors of marital outcomes. Once statistical techniques for conducting multiple regression became widely available, a second wave of research examined the ability of personality traits to account for marital outcomes relative to these other sources of influence. Perhaps the most impressive of these studies is that of Kelly and Conley (1987), who followed a sample of 300 couples for over 40 years. At the initial phase of data collection, engaged partners provided self-reports of their early family environment, attitudes towards marriage, sexual history, and stressful life events. In addition, the researchers contacted five acquaintances of each partner and asked these individuals to rate the personality of the subject, a substantial improvement over the self-report inventories relied on by most prior (and subsequent) studies of personality. When all of these variables were examined at the same time, only the personality ratings emerged as significant predictors of marital stability. Specifically, compared to marriages that remained intact, couples whose marriages dissolved over the course of the study were characterized by higher levels of neuroticism in both spouses and higher levels of impulsivity in husbands.

In the late 1980s, the development of the five-factor model of personality briefly promised to expand the focus of marital research beyond the negative aspects of spouses' personalities. According to this model, the many traits that had been studied in prior research on personality could be adequately summarized by five relatively independent dimensions: neuroticism, impulsivity, extraversion, agreeableness, and conscientiousness. Emerging support for this model lead marital researchers to ask whether dimensions of the 'Big Five' besides neuroticism and impulsivity may also be associated with marital outcomes. Thus, for example, Kurdek (1993), in a five-year study of newlywed couples, examined the ability of each of the five personality dimensions to predict the quality and stability of marriages over time. This study did find that wives' conscientiousness was weakly but positively associated with the stability of the marriage over five years. However, echoing the results of previous research, far larger effects were obtained for the neuroticism of both spouses. Karney and Bradbury (1995), in a meta-analysis of longitudinal research on personality and marriage, reported that of all of the Big Five personality dimensions, only neuroticism had demonstrated consistent effects on marriage across multiple studies. As the first marital researchers suggested, a spouse's 'predisposition to unhappiness' appears to predict decline and instability in marriage more reliably than other personality traits.

2. *Similarity and Complementarity*

As research on personality in marriage has developed, some have argued that the content of each spouse's personality may be less important to understanding marital outcomes than the degree of similarity between spouses' personalities. The assumption of this perspective is that spouses who share similar personalities should experience fewer opportunities for conflict and thus decreased risk of marital dissolution. Bentler and Newcomb (1978) offered some support for this view in a longitudinal study of newlywed couples. Although each spouse's level of extraversion had no independent association with their marital outcomes, the correlation between husbands' and wives' extraversion scores was significantly higher among couples who remained married than among couples who divorced.

Despite the intuitive appeal of this perspective, the accumulation of further support has been limited by a number of difficulties in assessing the effects of similarity in couples. First, the effects of similarity on a given trait are rarely independent of the effects of the trait itself. For example, Kurdek (1993) found that the difference between husbands' and wives' levels of neuroticism was greater in marriages that dissolved than in marriages that remained intact. This finding may support the role of personality similarity in marriage, but it may also result from the fact that differences between spouses are likely to be larger in couples where levels of neuroticism in either partner are higher overall. Estimating the unique effects of similarity between spouses requires that the effects of each spouse's personality be accounted for first. This is a control that has rarely been implemented.

A second difficulty with this kind of research is the fact that, as Kenny and Acitelli (1994) pointed out, 'A couple may appear to be very similar in their responses, yet they may be no more similar than a man and a woman who are totally unacquainted' (p. 419). Because there may be a typical way that most people tend to respond to personality inventories, the personalities of randomly matched people are likely to be similar to some degree. Estimating the unique degree of similarity in a given marriage requires that the 'typical cultural response' be controlled, but again this is a control that has rarely been implemented.

When similarity between spouses has been estimated accurately, support for beneficial effects of similar personalities has been sparse. For example, Russell and Wells (1991) reported that after controlling for the independent effects of each spouse's responses to a personality inventory, the association between marital satisfaction and the differences between spouses'

responses was rarely significant. Furthermore, the overall association between spouses' personalities tends to be weak (Buss 1984), even though spouses tend to be highly similar to each other on many other dimensions. In general, this research suggests that it may be spouses' similarity on those other dimensions, and in particular the degree to which spouses are similar in their values and attitudes, that may be more relevant to account for marital outcomes. With respect to personality, the nature of each spouse's personality appears to be more important than the degree of similarity between them.

In contrast to the similarity perspective, a competing view suggests that it is the degree of complementarity in spouses' personalities that is key to the successful marriage. For example, Winch (1958) argued that spouses select each other to resolve inadequacies in their own characters. Thus, extraverted people should be happiest with introverted partners and domineering people should be happiest with submissive partners. Although this is also an intuitively appealing idea, to date all attempts to obtain empirical support for the complementarity hypothesis have failed.

3. *How Does Personality Affect Marital Outcomes?*

Having established that personality traits, especially those associated with neuroticism, predict marital outcomes, the current focus of research on personality and marriage is to explain how this association comes about. Caspi et al. (1989), in their work on the effects of personality throughout the life course, proposed a broad framework that may be applied to understanding this issue. These researchers described two kinds of processes through which personality may affect development throughout the life course. 'Cumulative continuity' refers to all of the ways that individuals' personalities channel them towards certain environments and opportunities. 'Interactional continuity' refers to all of the ways that individuals' personalities affect their behavior towards and cognitions about other people. There is substantial evidence that both of these processes may account for the effects of spouses' personalities on their marriages.

3.1 *Cumulative Continuity in Marriage*

From early childhood, the personality of an individual affects not only the environments selected by the individual but also the opportunities available to the individual. Thus, for example, children who are prone to negative moods may choose career paths that are compatible with their temperament, but their choices may also be restricted by the behavioral consequences of their temperament (e.g., poorer educational attainment, lower professional accomplishment). In this way, individuals' personalities may have cumulative consequences that greatly shape the context of their adult lives. The nature of their adult environments may in turn affect the quality of individuals' marital relationships. For example, women who experience chronic depression as adolescents tend to marry at a younger age compared to women who do not experience chronic depression (Gotlib et al. 1998). Age at marriage is associated with risks of divorce, such that people who marry at a younger age are at higher risk for divorce than people who marry later in life. Thus, the stable tendency to experience negative affect in childhood appears to be associated, at least for women, with a life course that is detrimental to marriage.

In a similar vein, the personality characteristics of each spouse may give rise to stressful events and circumstances that impact the relationship. For example, individuals who score higher on measures of neuroticism describe events in their lives as more stressful than individuals scoring lower on such measures (Marco and Suls 1993). Furthermore, controlling for these reporting effects, higher levels of neuroticism appear to be associated with greater numbers of stressful life events (Poulton and Andrews 1992). To explain such findings, some researchers have suggested that individuals who are prone to experiencing negative affect frequently generate stressful circumstances that then serve to maintain their negative orientations. As a result of these processes, the marriages of individuals scoring higher in neuroticism may develop in more challenging circumstances than the marriages of less neurotic individuals. Because challenging circumstances, such as unemployment and financial difficulty, have been shown to predict divorce, this may be another route through which the cumulative consequences of each spouse's personality contribute to marital success and failure.

3.2 *Interactional Continuity in Marriage*

In addition to its role in shaping the circumstances of the marriage, personality may also affect interpersonal processes occurring within the marriage. Spouses' expectations for each other, the behaviors they exchange during their interactions, and the way they interpret those behaviors, have all been associated with the stable dispositions of each spouse. Through these mechanisms, the personality of each member of the couple may give rise to self-fulfilling prophecies, such that stable personality traits predict spouses' feelings towards each other and thus lead spouses to act in ways that confirm or enhance those feelings.

Support for these ideas has been especially strong with regard to negative affectivity, that is, neuroticism. For example, spouses who are higher in negative affectivity tend to make less charitable attributions for

their partners' negative behaviors. That is, these spouses are more likely to hold their partners responsible for negative behaviors, whereas spouses who are lower in negative affectivity are more likely to excuse their partners for such behaviors (Karney et al. 1994). Perhaps as a consequence, among newlywed couples, negative affectivity is associated with marital satisfaction, such that spouses who are higher in negative affectivity are likely to evaluate the marriage as a whole more poorly than spouses who are lower in negative affectivity (Karney and Bradbury 1997). The stable tendency to experience negative affect thus appears to affect spouses' interpretations of the marriage at multiple levels, leading to partners that are perceived as less sensitive and marriages that are perceived as less rewarding.

The effects of each spouse's personality are not merely intrapsychic, however. In established marriages, the negative affectivity of husbands is associated with the marital satisfaction of their wives, such that the wives of husbands who are higher in negative affectivity tend to be less satisfied with the marriage, controlling for their own level of negative affectivity (Karney et al. 1994). One way that the personality of one spouse may affect the satisfaction of the other is through the effects of personality on the behaviors that spouses exchange during their interactions with each other. Indeed, in established marriages, neuroticism is associated with couples' problem-solving behavior such that more neurotic spouses report feeling more negative than less neurotic spouses immediately after an attempt to resolve a marital problem (Geist and Gilbert 1996). Furthermore, observational coding of spouses' behavior indicates that more neurotic wives express more negative emotions during conflict discussions. Given that the quality of couples' problem-solving behaviors predicts stability and change in the quality of the marriage over time (Karney and Bradbury 1997), the link between neuroticism and marital interactions may be one way that the personality of an individual may influence the development of marital quality for a couple.

4. *Future Directions*

Although research on the interactional consequences of personality has focused almost exclusively on behavior during marital problem-solving interactions, personality may play an even more prominent role in the way spouses solicit and provide each other with social support. To the extent that personality, and in particular negative affectivity, affects the way spouses treat each other and the nature of the stressful events they encounter, then research on the provision of social support in response to stressful events would be a direction for future research that promises to integrate the cumulative and interactional consequences of personality in marriage.

See also: Adult Psychological Development: Attachment; Attributional Processes: Psychological; Lifespan Development, Theory of; Marriage; Neuroticism; Partner Selection across Culture, Psychology of; Personality and Adaptive Behaviors; Personality and Social Behavior; Personality Theories; Self-fulfilling Prophecies; Temperament and Human Development

Bibliography

Bentler P M, Newcomb M D 1978 Longitudinal study of marital success and failure. *Journal of Consulting and Clinical Psychology* **46**: 1053–70

Burgess E W, Wallin P 1953 *Engagement and Marriage*. Lippincott, Chicago

Buss D M 1984 Marital assortment for personality dispositions: Assessment with three different data sources. *Behavior Genetics* **14**: 111–23

Caspi A, Bem D J, Elder G H 1989 Continuities and consequences of interactional styles across the life course. *Journal of Personality* **57**: 376–406

Geist R L, Gilbert D G 1996 Correlates of expressed and felt emotion during marital conflict: satisfaction, personality, process, and outcome. *Personality and Individual Differences* **21**: 49–60

Gotlib I H, Lewinsohn P M, Seeley J R 1998 Consequences of depression during adolescence: Marital status and marital functioning in early adulthood. *Journal of Abnormal Psychology* **107**: 686–90

Karney B R, Bradbury T N 1995 The longitudinal course of marital quality and stability: A review of theory, method, and research. *Psychological Bulletin* **118**: 3–34

Karney B R, Bradbury T N 1997 Neuroticism, marital interaction, and the trajectory of marital satisfaction. *Journal of Personality and Social Psychology* **72**: 1075–92

Karney B R, Bradbury T N, Fincham F D, Sullivan K T 1994 The role of negative affectivity in the association between attributions and marital satisfaction. *Journal of Personality and Social Psychology* **66**: 413–24

Kelly E L, Conley J J 1987 Personality and compatibility: A prospective analysis of marital stability and marital satisfaction. *Journal of Personality and Social Psychology* **52**: 27–40

Kenny D A, Acitelli L K 1994 Measuring similarity in couples. *Journal of Family Psychology* **8**: 417–31

Kurdek L A 1993 Predicting marital dissolution: A five-year prospective longitudinal study of newlywed couples. *Journal of Personality and Social Psychology* **64**: 221–42

Marco C A, Suls J 1993 Daily stress and the trajectory of mood: Spillover, response assimilation contrast, and chronic negative affectivity. *Journal of Personality and Social Psychology* **64**: 1053–63

Poulton R G, Andrews G 1992 Personality as a cause of adverse life events. *Acta Psychiatrica Scandanavica* **85**: 35–8

Russell R J H, Wells P A 1991 Personality similarity and quality of marriage. *Personality and Individual Differences* **12**: 407–12

Terman L M 1938 *Psychological Factors in Marital Happiness*. McGraw-Hill, New York
Winch R F 1958 *Mate Selection: A Study of Complementary Needs*. Harper and Brothers, New York

B. R. Karney

Personality and Risk Taking

There is a long-standing and persistent belief that risk taking is a stable personality trait, often referred to as risk attitude. The belief implies that a given individual will take similar risks across a range of situations and that some people will be more risk-averse (or more risk-seeking) across situations than others. The article reviews different definitions of risk attitude that show cross-situational consistency to varying degrees. Sect. 1 shows that risk attitude defined within the expected utility (EU) framework varies greatly across situations as a function of decision content and outcome framing. Sect. 2 describes a more promising conceptualization of risk taking, within a risk–value framework. It models risk taking as a function of (a) decision makers' perception of the riskiness and value or return of different courses of action, and (b) their attitude towards perceived risk, i.e., their willingness to trade off (perceived) risk for return. Two individuals might differ in their recreational pursuits, for example, either because they assess the relative risks of skydiving, bungee jumping, and playing poker very differently (based on past experience, person A may perceive playing poker to be riskier than skydiving and thus choose to go skydiving out of risk aversion—a negative attitude toward risk, while person B may perceive playing poker as the less risky option and engage in it, also out of risk aversion) or, whether their risk perceptions agree or not, because they have different attitudes toward risk as they see it (with person C and D agreeing on the greater risk posed by skydiving, but person C being attracted by this risk and thus taking it on, and person D being repelled by it and thus choosing to play poker instead).

When modeled within this framework (as described in Sect. 3), situational differences in risk taking turn out to result from differences in the perception of risk in different situations rather than differences in willingness to take on (perceived) risk, thus restoring credibility to the notion of attitude towards perceived risk (PRA) as a stable trait. Individual differences in PRA exist, but are smaller and less systematic than individual and group differences in risk perception. While the determinants of risk perception are relatively well known at this point (see *Risk: Empirical Studies on Decision and Choice*; *Risk: Theories of Decision and Choice*) not much is known about the determinants of PRA. Personality differences in variables known to be related to risk taking seem to have their effect via differences in risk perception, as described in Sect. 4.

1. Domain and Framing Effects on Risk Taking

In the EU framework, the dominant normative model of risky decision making, the term 'risk taking' is used to characterize choice patterns. Choice of a sure outcome over a lottery with equal expected value is modeled by a concave utility function and described as risk-averse; choice of the lottery is modeled by a convex utility function and described as risk-seeking. Despite the fact that risk taking simply describes the curvature of the utility function that is derived from a series of choices, 'those who coined the term *risk aversion* had in mind the psychological interpretation that someone who prefers the expected value of a gamble over playing the gamble *does not like to take risks*' (von Winterfeldt and Edwards 1986, p. 256). Popular as well as managerial folklore tends to interpret risk taking as a personality trait.

Bromiley and Curley (1992) review the evidence for risk taking as a personality trait, i.e., as a preference for risk that is stable across situations, and find it lacking. Risk taking seems to be influenced jointly by the situation and characteristics of the decision maker. Decision domains in which the same person often shows different degrees of risk taking include games of chance/gambling, financial investing, business decisions, health decisions, recreational choices, social choices, and ethical decisions (MacCrimmon and Wehrung 1986, Weber et al. 2000).

Modeling risk taking within EU theory and defining risk attitude as the curvature of a utility function thus is clearly problematic for the notion of risk attitude as a personality trait. Attempts to restore cross-situational consistency to the construct of risk attitude by factoring differences in marginal value (e.g., the incremental value of an additional dollar or an additional life saved) out of the utility function (see *Risk: Theories of Decision and Choice*) were not successful (Keller 1985, Weber and Milliman 1997).

Prospect theory (Kahneman and Tversky 1979) generalizes EU by postulating different utility functions for outcomes that are framed as gains as opposed to losses. When outcomes are framed as gains, choices tend to be risk-averse; when the same outcomes are framed as losses (relative to a higher reference point), choices tend to be risk-seeking, further complicating the interpretation of risk taking in the EU sense as a stable trait.

2. Risk Taking and Risk Perception

In the risk–value framework (see *Risk: Theories of Decision and Choice*), risk taking is a compromise between greed (value) and fear (risk). Risk–value

models in finance equate 'value' with the expected value of a risky option and 'risk' with its variance. Generalized risk–value models allow for a broader range of risk measures. *Risk: Empirical Studies on Decision and Choice* reviews evidence that risk is perceived differently by different individuals, cultures, or subcultures. Situational differences such as outcome framing also result in different risk perception (Mellers et al. 1997). As a result, apparent differences in risk taking may be the result of differences in the perception of the riskiness of the choice options, and not of differences in attitude towards (perceived) risk. Cooper et al. (1988) report, for example, that—contrary to managerial folklore—the characteristic that differentiates entrepreneurs from other managers is not a more positive attitude towards risk, but instead an overly optimistic perception of the risks involved. For an outside observer who perceives risk more realistically, entrepreneurs will thus appear to take great risks. However, when differences in risk perception are factored out, entrepreneurs—just as other managers—demonstrate a preference for tasks that they see as only moderate in risk (Brockhaus 1982).

3. *Perceived-risk Attitude as a Stable Trait*

PRA is a measure of the degree to which individuals find perceived risk attractive (or unattractive) and therefore will choose alternatives that carry greater (or less) risk, all other things being equal. Weber and Milliman (1997) examined its cross-situational consistency by asking commuters to choose between pairs of trains that had risky arrival times (that depended on making a connection that had a stated probability) and to judge which of the two trains was the riskier one. The two trains in each pair had arrival times with equal expected value but different variance. Some pairs of trains had only positive arrival times (faster or equal to current travel times), others had only negative arrival times (slower or equal to the status quo). There was little consistency in people's risk taking across the gain and the loss domain when risk taking was defined in the EU sense. Few commuters had preferences that resulted in utility functions that were either both risk-seeking (convex) or both risk-averse (concave). However, consistency across the two domains was very high when PRAs were compared. The majority of commuters were risk-averse in both domains, i.e., consistently chose the train in a given pair that they had judged to be the less risky of the two.

In another study, MBA students participated in two sessions of an investment game where they had to pick one of six stocks (described by standard financial indicators) in each of 10 investment periods, and had to rate the riskiness of the stocks at different points throughout each session (Weber and Milliman 1997). In one session, participants lost money in most of the 10 periods, whereas in the other session they mostly made money. Choices were very different across sessions (with more switching between stocks in the failure session), as were the ratings of the riskiness of the six stocks. However, over 80 percent of investors had the same PRA in both sessions, with three-quarters consistently investing in stocks that they perceived to be less risky and one-quarter consistently investing in stocks that they perceived to be more risky.

In a cross-national study, Weber and Hsee (1998) obtained risk judgments as well as minimum buying prices for risky financial investment options from respondents in the USA, Germany, the People's Republic of China, and Poland. Both risk judgments and buying prices showed significant cross-national differences, with Americans perceiving the most risk and Chinese paying the highest prices. However, after differences in risk perception were taken into consideration, the proportion of individuals who were perceived risk-averse or perceived risk-seeking were not significantly different in the four countries, with the majority again being perceived risk-averse, and only a small percentage in each country being perceived risk-seeking.

4. *Personality, Risk Perception, and Perceived-risk Attitude*

Some psychologists have questioned the assumption of finance models that people will and should strive to minimize risk, arguing instead that people's ideal point for risk or uncertainty could differ, either as a personality difference (Lopes 1987) or as a situational difference (Weber and Kirsner 1997). Ideal-point models (Coombs 1975) assume a person will perceive the riskiness of an alternative as the deviation between the alternative's level of uncertainty or unpredictability and the person's ideal point on the uncertainty continuum. Perceived risk of an alternative with a high objective level of uncertainty would be high for a person with a low ideal point, but low for a person with a high ideal point. Individual differences in ideal points for risk and uncertainty have been measured by the construct of sensation seeking (Zuckerman 1979), which seem to have some biological basis (Zuckerman et al. 1988) and vary with age and gender (see *Sensation Seeking: Behavioral Expressions and Biosocial Bases*). Bromiley and Curley (1992) report evidence linking sensation seeking to behavioral correlates that include greater risk taking, especially in the health/safety and recreational domain. Weber et al. (2000) also report high positive correlations between sensation seeking and its subscales in several content domains, with especially high correlations between the thrill-and-adventure-seeking subscale and recreational risk

taking and the disinhibition subscale and ethical risk taking. Consistent with the predictions of ideal-point models, the path by which differences in sensation seeking seem to affect risk taking appears to be differences in the perceptions of risk, rather than differences in attitude towards perceived risk. In other words, groups known for high levels of sensation seeking (e.g., teenage boys) seem to take large risks because they perceive the levels of risk to be smaller than other groups, and not because they cherish (perceived) risk to a greater extent.

5. *Summary, Caveats, and Future Directions*

The current research consensus suggests an interactional model of risk taking (e.g., Sitkin and Weingart 1995) in which situational characteristics as well as person-centered characteristics jointly influence risk taking. Situational constraints include the content domain of the risky decision as well as contextual variables such as outcome framing and aspiration levels (Lopes 1987). Person-centered characteristics include age, gender, culture, and personality. These variables influence risk taking mostly by changing people's perception of the riskiness of decision alternatives, rather than by affecting their willingness to take on more or less risk.

Because of the domain specificity of risk taking, measures of risk attitude that employ choice situations across a range of content domains (e.g., the Choice Dilemmas Questionnaire of Kogan and Wallach 1964) have little predictive validity. Domain-specific scales of risk taking, that help to diagnose apparent differences in risk taking into differences in either risk perception and/or PRA have recently developed (Weber et al. 2000). Future research will provide additional insights into the complex interactions between personality and situation that have been explored for a range of other traits (Mischel 1999) with respect to risk taking. A combination of task analysis and theory about the reasons for risk taking and its cognitive and emotional constraints should lead to the development of gender-, culture-, and domain-specific risk taking profiles, that predict level of risk taking in a situation- and person-contingent fashion.

See also: Personality Psychology; Risk: Empirical Studies on Decision and Choice; Risk: Theories of Decision and Choice; Sensation Seeking: Behavioral Expressions and Biosocial Bases

Bibliography

Brockhaus R H 1982 The psychology of the entrepreneur. In: Kent C A, Sexton D L, Vesper K G (eds.) *The Encyclopedia of Entrepreneurship*. Prentice Hall, Englewood Cliffs, NJ, pp. 92–4

Bromiley P, Curley S P 1992 Individual differences in risk-taking. In: Yates J F (ed.) *Risk-taking Behavior*. Wiley, New York, pp. 87–132

Coombs C H 1975 Portfolio theory and the measurement of risk. In: Kaplan M F, Schwartz S (eds.) *Human Judgment and Decision*. Academic Press, New York, pp. 63–8

Cooper A C, Woo C Y, Dunkelberg W C 1988 Entrepreneurs' perceived chances for success. *Journal of Business Venturing* **3**: 97–108

Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263–91

Keller L R 1985 An empirical investigation of relative risk aversion. *IEEE Transactions on Systems, Man, and Cybernetics* **15**: 475–82

Kogan N, Wallach M A 1964 *Risk-taking: A Study in Cognition and Personality*. Holt, New York

Lopes L L 1987 Between hope and fear: The psychology of risk. *Advances in Experimental Social Psychology* **20**: 255–95

MacCrimmon K R, Wehrung D A 1986 *Taking Risks: The Management of Uncertainty*. Free Press, New York

March J G, Shapira Z 1992 Variable risk preferences and focus of attention. *Psychological Review* **99**: 172–83

Mellers B A, Schwartz A, Weber E U 1997 Do risk attitudes reflect in the eye of the beholder? In: Marley A A J (ed.) *Choice, Decision, and Measurement*. Erlbaum, Hillsdale, NJ, pp. 57–71

Mischel W 1999 Implications of person–situation interaction: Getting over the field's borderline personality disorder. *European Journal of Personality* **5**: 455–61

Sitkin S B, Weingart L R 1995 Determinants of risky decision making behavior: A test of the mediating role of risk perceptions and risk propensity. *Academy of Management Journal* **38**: 1573–92

von Winterfeldt D, Edwards W 1986 *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, UK

Weber E U, Blais A R, Betz N 2000 A domain-specific risk-attitude scale: Measuring risk perceptions and risk behavior. Working Paper, Center for the Decision Sciences. Columbia University, New York

Weber E U, Hsee C K 1998 Cross-cultural differences in risk perception but cross-cultural similarities in attitudes towards risk. *Management Science* **44**: 1205–17

Weber E U, Kirsner B 1997 Reasons for rank-dependent utility evaluation. *Journal of Risk and Uncertainty* **14**: 41–61

Weber E U, Milliman R A 1997 Perceived risk attitudes: Relating risk perception to risky choice. *Management Science* **43**: 122–44

Zuckerman M 1979 *Sensation Seeking: Beyond the Optimal Level of Arousal*. Erlbaum, Hillsdale, NJ

Zuckerman M, Simons R F, Como P G 1988 Sensation seeking and stimulus intensity as modulators of cortical, cardiovascular, and electrodermal response: A cross-modality study. *Personality and Individual Differences* **9**: 361–72

E. U. Weber

Personality and Social Behavior

The link between personality and social behavior can be approached in two ways. The first question is to what extent social behavior is caused by personality

factors. The second question is what effects social behavior has on personality. Both questions have proven surprisingly difficult for social scientists to answer.

Personality refers to the stable traits and dispositions that characterize individual people. In simplest terms, social behavior has two possible causes: personality and the situation. Although it is tempting to assign everything inside the person to personality and everything outside to the situation, the reality is somewhat more complex, because external factors (such as being insulted by others, an emergency, or an opportunity) may create temporary internal states. Personality is thus usually reserved to the relatively stable aspects of the person, as opposed to temporary states. One handy test is what aspects of the person remain the same when he or she moves from one situation to another. Personality consists of those aspects.

1. *Mischel's Challenge to Personality*

The view that social behavior derives from personality is deeply rooted in common sense and traditional views. People are widely believed to have stable personality traits that account for readily discernible patterns in their behavior. Indeed, the very concept of personality is usually understood in terms of consistency. That is, people are supposedly consistent in their motivations and actions.

Walter Mischel (1968) challenged this alleged consistency of personality in an influential book. Mischel proposed that if personality were composed of a stable, consistent set of traits, then psychological measures of traits should effectively predict social behavior quite strongly. From a review of published research studies, however, he concluded that personality traits have not been shown to have strong effects on behavior. Instead, he found that the correlations between traits and behaviors were usually small: Typically the correlation was found to be 0.20 to 0.30 out of a maximum possible 1.00. He used the term 'personality coefficient' to refer to the small correlations that were typically found.

Mischel therefore concluded that the situation is far more powerful than personality as a cause of social behavior. Following the standard statistical practice, he pointed out that a correlation of 0.30 accounts for only 9 percent of the variation in behavior, leaving the other 91 percent to the situation. Put simply, the situation was 10 times as powerful as personality in determining behavior.

This view was immediately recognized as challenging the view that personality is an important cause of social behavior. Psychologists were not the only ones to have believed in personality, for the general public has also had a long and firm belief in personality. Mischel's colleague Lee Ross (1977) coined the term 'fundamental attribution error' to refer to people's tendency to interpret someone's behavior as caused by that person's personality traits and to underestimate the importance of the situation in causing behavior. For example, people who observe an angry outburst will tend to conclude that the angry person has a hostile or aggressive personality while in fact the true cause of the angry outburst is likely to be a bad day, a provocation, a frustration, or some other cause rooted in the immediate situation. A well-known work by Jones and Nisbett (1971) pointed out that people tend to attribute their own behavior as caused by their situation (e.g., 'I was only following orders' or 'I was just reacting to what you did'), whereas they interpret other people's behavior in terms of personality traits.

Mischel's book and his striking conclusions had a major impact on psychology. One implication was that psychologists should focus on analyzing the situation rather than the individual personality in order to be able to understand and predict behavior. Social psychologists, who specialized in studying the power of situations, embraced this conclusion to the detriment of personality psychologists. This attack on personality coincided with a period of rapid growth in US psychology departments, and as a result many universities hired social psychologists rather than personality psychologists, with the result that social psychology continues to be a much larger field than personality even several decades later.

A recent analysis of the controversy by Funder (1997) summarized Mischel's argument as having three essential parts. First, a review of the literature shows that there is an upper limit to how well one can predict someone's behavior from one situation to the next, and this upper limit is small (a correlation of around 0.20 to 0.30; Nisbett 1980 later revised this upward to 0.40). Second, situations are more important than traits in explaining behavior. Third, it is no use to measure personality, and the common and traditional view of people as consistent, personality-driven beings is wrong.

2. *Defending Personality*

Personality psychologists gradually came forward to defend their field of study and the value of their work against Mischel's critique. One important point was that the correlations of 0.20 to 0.40 are not trivial and can in fact contribute useful knowledge. A correlation of 0.40 means that a prediction of behavior would be right 70 percent of the time (compared to 50 percent by chance).

Moreover, the conclusion that situations are much more powerful than personality traits does not stand up under scrutiny. Funder and Ozer (1983) analyzed the power of situational variables to predict behavior,

using several classic social psychology experiments on cognitive dissonance and bystander reactions to emergencies. Funder and Ozer (1983) found that the well-known effects of these situational factors were about the same size (between 0.30 and 0.40, when converted to correlations) as Mischel's personality coefficient. In other words, trait and situation have about equally powerful effects; it was simply not true that the situation was 10 times as powerful as personality.

Another important point was that the seeming weakness of personality traits for predicting behavior was partly due to the way psychologists did their research. Typically they would use a standard questionnaire to assess some general trait, such as helpfulness, and then they would correlate it with how helpful the person was in a specific situation (such as how the person responded to a request to donate blood). The problem with such designs is that only a single, context-specific behavior was measured, which might not match up well with a broad trait. In contrast to that procedure, some researchers began aggregating multiple behaviors. Thus, instead of only measuring whether the person acceded to one request to donate blood, one could combine many different helpful behaviors, such as taking care of sick friends, giving directions to strangers, volunteering to help on a hotline, donating money, befriending disadvantaged children, helping friends with chores, and assisting someone with schoolwork. Epstein (1979a, 1979b) found that aggregating behaviors in that way enabled traits to predict behavior at a level of over 0.80 in some cases, thus far more successfully than the 0.3 coefficient derided by Mischel.

The success of aggregation is revealing about how personality is linked to social behavior. It shows that personality traits can predict behavior in general with good success. They are, however, much less successful at predicting single behaviors. To know someone's personality is therefore to be able to predict how that person will behave in many situations, but it remains much harder to say how a given person will react on a particular occasion.

3. *Traits and Metatraits*

Another possible reason for the seeming weakness of the link between personality and behavior is that not all personalities are composed of the same kinds of traits. Bem and Allen (1974) proposed that some people are more consistent than others with respect to certain traits. Including inconsistent people in a research project will inevitably weaken the results. For example, Bem and Allen (1974) asked people whether they were consistent or inconsistent on the dimension of sociability. One person might be consistently high in sociability, or consistently medium, or consistently low. Someone else might be inconsistent, in the sense that the person acts in a highly sociable manner one day but in an unsociable or even antisocial way the next. Bem and Allen computed the correlation between trait sociability score and a behavioral measure of sociability twice: once for the people who said they were consistent, and once for the inconsistent ones. The correlation was much higher for the people who were consistent.

The implications of Bem and Allen's idea are potentially profound. For some people, the trait of being sociable versus unsociable is a stable, consistent aspect of personality, but for others it is not. One theme in personality psychology is the attempt to measure traits, but perhaps some people should not be measured on some traits. The question 'How sociable are you?' may be relatively meaningless for some people, because the answer would fluctuate from day to day and from one situation to another. For the latter, apparently, their personality lacks anything that could be described as a trait level of sociability. Baumeister and Tice (1988) used the term 'metatrait' to describe this pattern. A metatrait is literally the trait of having (or not having) a trait. For traited people, behavior will tend to be consistent. For untraited people, behavior may vary widely even within seemingly similar situations.

The idea of metatraits poses a serious challenge to both the measurement and the conceptualization of personality. Several methods have been proposed to ascertain whether people are traited on a given trait or not. There is ample evidence that the very same measures yield much stronger results for traited as opposed to untraited people (Baumeister and Tice (1988, see also Britt and Shepperd 1999). Yet other experts have proposed that these measures could potentially have statistical problems (e.g., Tellegen 1988), and there is not yet a general consensus on the best way to deal with these.

Still, the idea that people differ in their level of consistency is promising, and eventually it may turn out to be a crucial aspect of the link between personality and social behavior. The implication is that personality contains something that makes the person's behavior consistent, and that some personalities do not have whatever that is. Although Bem and Allen (1974) and most subsequent researchers have assumed that different people are traited on different traits, Snyder (1974) proposed that some people are simply more consistent than others across all traits. He developed a measure of self-monitoring that sought to assess how much each person changes his or her behavior from one situation to another. The term 'self-monitoring' refers to the monitoring of situational cues and altering one's own behavior accordingly. People who are high in self-monitoring are defined as being alert to circumstances and quick to change their behavior so as to do what might work best in any given situation. In contrast, people who score low in self-monitoring tend to act based on their inner feelings

and values regardless of the situation. Hence low self-monitors will be more consistent across different situations, whereas high self-monitors will tailor their behavior to each situation and exhibit less cross-situational consistency.

4. *Different Situations*

Two other ways of understanding the link between personality and social behavior focus on differences between situations. One of these emphasizes simply that different situations vary in their power to dictate behavior. A strong situation can override personality traits, so that nearly everyone will act the same way. In contrast, a relatively weak situation will allow behavior to be driven by personality.

Research on behavior in emergencies has provided a good illustration of this approach. Darley and Latané (1968) investigated the question of whether bystanders will come to the aid of a victim in an emergency. They found that when bystanders are alone and believe themselves to be the only ones who can help the victim, they are quite likely to act. In contrast, when there are many bystanders, each is likely to think that it is someone else's responsibility to take action, and so each will tend to do nothing. Across a series of studies, they measured an assortment of personality traits, but none of these traits predicted behavior. They concluded from their research that personality is largely irrelevant to how people respond in emergencies (see Latané and Darley 1970). They suggested that emergency situations are both powerful and unusual, and so few people develop habits or traits about how to deal with them.

The conclusion that personality is irrelevant to emergencies came under criticism from personality researchers. They proposed that the original researchers had simply used extremely powerful versions of the emergency situation, with either no one else available to help or a great many people. What about when there are only one or two other people who might help? When the emergencies were set up with just one or two other bystanders, the response rate moved to the intermediate level of 50 percent, and personality traits were quite effective at predicting which people would respond (Siem and Spence 1986, Tice and Baumeister 1985).

In short, situations do vary in how powerfully they exert pressure on people to respond in a particular way. When this power is high, there is little room for personality to make a difference. In contrast, when the power of the situation is low, personality can be decisive. This conclusion complements the metatrait view that some personalities are more strongly geared toward consistent behavior than others. An untraited person will generally respond based on the situation, whereas a traited person will act based on his or her personality. Conversely, a powerful situation will dictate behavior, whereas a weak one will leave behavior to be determined by inner traits.

The second situation-based approach was put forward by Snyder and Cantor (1998). In this theory, situations are the primary determinant of behavior — but personality dictates which situations the person enters. For example, introverted people may avoid loud parties, and so their behavior does not come under the influence of the party atmosphere. Thus, personality and situations may operate in sequence rather than in direct competition for determining social behavior. Personality involves the grander role of how people set up their lives and how they make broad choices, whereas situations exert the immediate and short-term influence over how the person acts.

5. *Behavior Shapes Personality*

Although most research has focused on the question of whether personality determines social behavior, there has been some work on the complementary question of whether behavior shapes personality. To be sure, personality is defined as stable and consistent, so such change will be slower, smaller, and more gradual than changes in behavior. Yet few people believe that personality is utterly immune to change. A compilation of works by many experts on the topic of 'Can personality change?' yielded an emphatic answer of 'yes,' even though change was acknowledged to be difficult and slow in many cases (Heatherton and Weinberger 1994).

One of the possible models of personality change involves internalization. For example, young women do not magically develop maternal personalities, nor are these traits always visible during their youth prior to motherhood, but upon assuming the role of mother a woman may begin to change her behavior, and over time these new patterns of maternal behavior can become firmly established to the point that the woman's personality can be said to have changed to fit the role.

How does the internalization process operate? One important theory was proposed by Jones et al. (1981). These authors suggested that people's self-concepts contain a great deal of information that is not necessarily consistent or well-organized. When people are induced to act in a certain way, they begin to recall other behaviors that are consistent (and to ignore previous acts that would not be consistent). Eventually the self-concept shifts to accommodate these patterns of behavior, and then the personality itself follows suit. Fazio et al. (1981) asked people a series of loaded questions that were designed to get them to think about themselves as either extraverted or introverted (e.g., 'What things would you do to liven up a dull party?' versus 'What do you dislike about loud,

crowded parties?'). Their self-concepts shifted accordingly. Later, when they were left alone with a stranger, their behavior followed from the newly induced self-concept, which implies that there was at least some lasting change in their personality. Specifically, the people who had been led to think of themselves as introverts sat quietly and did not talk with the stranger, whereas the people who had been led to think of themselves as extraverts were more likely to strike up a conversation with the stranger.

The importance of the social dimension of behavior emerged in subsequent work. Tice (1992) showed that people only internalize behaviors that are witnessed by other people. That is, when people gave their answers to another person in an interpersonal setting where another person heard them, they showed evidence of personality change. In contrast, when they gave their answers confidentially and anonymously by speaking into a tape recorder while alone, they showed no sign of internalization. The implication is that secret and private behaviors do not lead to internalization as much as do public, interpersonal behaviors. The social context—specifically, whether other people are there to witness the behavior—is decisive as to whether behaviors lead to inner change.

The practical side of this can be seen in the practices used by cults and other organizations that seek to 'brainwash' or otherwise bring about fundamental changes in people's values, motives, and other aspects of personality. Attempts to effect such change by bombarding the person with messages or influence while the person sits there passively have not proven effective. Nor is it apparently enough to get the person to comply in a private or anonymous fashion. Rather, success at brainwashing depends on inducing the person to comply actively and publicly. New recruits to a religious cult may, for example, be induced to speak out in front of the group about their new commitment to the cult or even to approach strangers to tell them about the cult, ask for donations, distribute literature, and the like. Such public actions help cement the person's commitment to the cult and bring about the inner change that the group seeks.

6. *Interactionism*

Although the intellectual debate has been cast in terms of whether the trait or the situation is the main cause of behavior, this is merely a matter of emphasis: Most experts believe in interactionism, which is the doctrine that behavior is an interactive product of the person and the situation (see *Interactionism and Personality*). Put another way, hardly anyone believes that traits are so powerful that people will always behave the same way regardless of situation, nor does anyone really believe that there are no meaningful differences among individuals.

The interaction between person and situation is implicit in several of the formulations already noted, such as the idea that personality dictates the selection of particular situations, and then the situations guide actual behavior. An explicit interactionism would assert that any approach is doomed if it focuses on merely the person or merely the situation. A crowded, noisy party will make some people mingle happily with the crowd and enjoy shouted discussions with groups of people, while others retreat to search for a quiet place where they can talk at length with a single person. Likewise, a major upcoming test will strike some people as a threat, making them so anxious that they avoid studying, whereas others will embrace the test as a challenge and prepare to do their best. Thus, seemingly identical situations elicit very different behaviors from different people, depending on their personalities.

Modern statistical techniques allow the observations of human behavior to be divided into three parts. Trait and situation each claim a share, and the interaction (between trait and situation) gets the third share. The trait share refers to how much behavior is affected directly by personality, independent of the situation (and thus consistently across different situations). The situation share involves how much all different people will give essentially the same response to the situation. Last, the interaction encompasses ways that the same situation affects different people differently.

There is unfortunately no general principle or agreement as to which of these three shares is generally the largest. It appears that some situations are quite powerful in affecting everyone the same way. Other situations are relatively weak (such as unstructured situations), and traits exert considerable influence there. Interactions do not always occur, but they are common and often generate the most interest, insofar as different people react in reliably different ways.

7. *Future Directions*

At present, it appears that personality psychology has successfully survived the crisis provoked by Mischel's (1968) critique, although some effects linger (such as the smaller size of personality psychology faculty). Instead of looking at personality and situation as competing against each other to explain behavior, the current emphasis is to understand how they interact and complement each other. A priority for the next generation of research is to assess the several models (described above) for possible interactions, such as differential traitedness, personality dictating entry into situations and situations being the proximal causes of behavior, and differential power of situations.

Another reason to expect increasing rapprochement between research programs focusing on personality

and programs focusing on situational causes is contained in the logic of scientific development. As in most fields, broad generalizations and sweeping general principles are likely to be identified first, whereas later generations have to study more fine-grained and subtle effects. The behavioral principles that are generally true for everyone were likely identified relatively early in psychology's scientific development. Subsequent generations of researchers will therefore have to focus more on principles that apply only to certain people, and these attempts to understand how situations affect particular groups will require a greater sensitivity to individual differences and a greater integration between personality and situational factors.

The study of how situations and specific experiences produce lasting changes in personality remains wide open for new ideas and creative research. The question of how personality changes is likely to be of increasing importance in the coming decades for many reasons, including the wish to facilitate child development toward desired personality outcomes, the increasingly recognized need to offset the consequences of harmful parenting or other nonoptimal upbringing patterns, and in particular the wish to enable clinical psychology to bring about the desired changes in a reliable and lasting manner.

See also: Interactionism and Personality; Personality and Adaptive Behaviors; Personality Psychology; Personality Theories

Bibliography

Baumeister R F, Tice D M 1988 Metatraits. *Journal of Personality* **56**: 571–98

Bem D J, Allen A 1974 On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review* **81**: 506–20

Britt T A, Shepperd J A 1999 Trait relevance and trait assessment. *Personality and Social Psychology Review* **3**: 108–22

Darley J M, Latané B 1968 Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology* **8**: 377–83

Epstein S 1979a The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology* **37**: 1097–126

Epstein S 1979b The stability of behavior: II. Implications for psychological research. *American Psychologist*. **35**: 790–806

Fazio R H, Effrein E A, Falender V J 1981 Self-perceptions following social interactions. *Journal of Personality and Social Psychology* **41**: 232–42

Funder D C, Ozer D J 1983 Behavior as a function of the situation. *Journal of Personality and Social Psychology* **44**: 107–12

Funder D C 1997 *The Personality Puzzle*. W W Norton, New York

Heatherton T F, Weinberger J (eds.) 1994 *Can Personality Change?* American Psychological Association, Washington, DC

Jones E E, Nisbett R E 1971 *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*. General Learning Press, New York

Jones E E, Rhodewalt F, Berglas S C, Skelton A 1981 Effects of strategic self-presentation on subsequent self-esteem. *Journal of Personality and Social Psychology* **41**: 407–21

Latané B, Darley J M 1970 *The Unresponsive Bystander: Why Doesn't He Help?* Appleton-Century Crofts, New York

Mischel W 1968 *Personality and Assessment*. Wiley, New York

Nisbett R E 1980 The trait construct in lay and professional psychology. In: Festinger L (ed.) *Retrospections on Social Psychology*. Oxford University Press, New York, pp. 109–30

Ross L 1977 The intuitive psychologist and his shortcomings: Distortions in the attribution process. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology* (Vol. 10, pp. 174–221). Academic Press, New York

Siem F M, Spence J T 1986 Gender-related traits and helping behaviors. *Journal of Personality and Social Psychology* **51**: 615–21

Snyder M 1974 The self-monitoring of expressive behavior. *Journal of Personality and Social Psychology* **30**: 526–37

Snyder M, Cantor N 1998 Understanding personality and social behavior: A functionalist strategy. In: Gilbert D, Fiske S, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, pp. 639–79

Tellegen A 1988 The analysis of consistency in personality assessment. *Journal of Personality* **56**: 621–63

Tice D M 1992 Self-presentation and self-concept change: The looking glass self as magnifying glass. *Journal of Personality and Social Psychology* **63**: 435–51

Tice D M, Baumeister R F 1985 Masculinity inhibits helping in emergencies: Personality does predict the bystander effect. *Journal of Personality and Social Psychology* **49**: 420–8

R. F. Baumeister and J. M. Twenge

Personality Assessment

To assess personality is to characterize the enduring psychological qualities that contribute to the coherence and uniqueness of the individual person and to the differences among individuals. The task of personality assessment commonly is construed in terms of the measurement of quantifiable personality variables. However, personality assessment may include both quantitative and qualitative techniques.

The set of methods that psychologists have employed to assess personality is highly diverse. There are two main reasons for this. One is that assessors have held varying beliefs about the nature of personality itself. The twentieth century witnessed a variety of formal theories of personality. These theoretical positions have guided assessors' decisions about the aspects of personality to assess and the procedures to use to assess them. Diversity in personality theory thus has fostered a corresponding diversity in assessment techniques. The second reason is that investi-

gators may assess personality with different goals in mind. One may aim, for example, to describe normal variations in personality in the population at large; to diagnose psychopathology suffered by a small subset of the population; or to obtain a detailed portrait of the psychological structures and dynamics that characterize a particular individual. Different goals commonly require different assessment methods. The methods that contemporary investigators employ are reviewed below, following a history of the field.

1. History

In a general sense, to assess personality is to evaluate the individual in a formal manner. Personality assessment, then, is as old as is the development of formal schemes for characterizing persons. Two early schemes are found in ancient Greece. The *Corpus Hippocraticum* introduced the notion that variations in four bodily humors (blood, phlegm, yellow bile, and black bile) determine one's predominant psychological temperament (sanguine, melancholic, choleric, and phlegmatic). Theophrastus, a disciple of Aristotle, provided a series of character sketches of common personality types, each defined according to a moral attribute (the liar, the adulterer, etc.). The differences between these approaches—one examining biological constitution and providing a simple system for characterizing individual differences, and the other focusing on social behavior in the moral domain and providing a more differentiated set of descriptors—interestingly foreshadows variations within the contemporary field.

The more recent centuries witnessed attempts to link personality to biological anatomy. Investigators measured anatomical features in an effort to identify underlying causes of overt psychological qualities. These efforts included the eighteenth to nineteenth century work of Lavater, who interpreted facial features; the 'organology' of Gall and 'phrenology' of Spurzheim, both of whom assessed variations in the anatomy of the skull; and the work of Lombroso, who assessed physical features that purportedly were markers of criminality. These efforts, of course, proved to be of little lasting value.

A more sophisticated conception of assessment was found in the efforts of Charcot, Janet, Freud, and Jung to diagnose the causes of psychopathology. Freud's free association technique, for example, was not only a therapeutic treatment but an assessment tool. Whatever the merits of these efforts, however, they failed to meet applied psychologists' need for efficient assessments that could be administered easily to relatively large populations.

Modern approaches to personality assessment can be traced to the late-nineteenth and early-twentieth century efforts of Galton and Binet to assess intelligence. Their standardized, paper-and-pencil assessments of individual differences in intellectual abilities provided a model for the standardized assessment of individual differences in personal styles, preferences, and behavioral tendencies. The first paper-and-pencil standardized personality inventory was the Woodworth Personal Data Sheet, which was developed during World War I to screen from the US armed forces individuals suffering from psychological disorders. Thanks heavily to their ease of administration, standarized self-report tests proved to be the most commonly employed personality assessment method throughout the remainder of the twentieth century.

A noteworthy alternative to paper-and-pencil assessment was developed during World War II by psychologists at the US Office of Strategic Services, working under the direction of Harvard psychologist Henry Murray. To screen candidates for participation in risky missions, this group assessed candidates' behavior in artificially constructed situations that contained the same types of challenges (e.g., solving practical engineering problems, displaying leadership in a group) that one might face on an actual mission. Such *in vivo* behavioral assessments can be of much predictive value. However, they are much more cumbersome and costly than paper-and-pencil self-reports, and thus are used much less frequently.

Finally, two theoretical developments in the 1950s and 1960s proved to be of historical importance to the field. One was the development of the concept of *construct validity* (Cronbach and Meehl 1955). Construct validity is concerned with the degree to which an assessment tool adequately measures a hypothesized psychological construct, as that construct is understood within a given theoretical framework. Determining the validity of a test generally involves gauging the degree to which that test predicts external criteria to which the associated construct should be related. A valid self-report test of conscientiousness, for example, would be one that adequately samples the domain of conscientious behaviors and predicts external criteria involving conscientious actions such as adhering to social norms and controlling impulses.

The second development was a paradigm crisis in the field of personality assessment in the late 1960s. Although the crisis had many sources, a critique by Mischel (1968) was particularly influential. Mischel questioned whether assessments of decontextualized or 'global' personality constructs predict specific behaviors to which they are conceptually related. His review indicated that global personality measures commonly either fail to predict behavioral criteria or predict so little of the variance that they are of little practical use. His critique prompted divergent reactions. Some tried to improve the measurement of global psychological characteristics. Others developed personality theories and associated assessment tools that were not based on global trait variables. These theoretical alternatives were, instead, grounded in the analysis of basic cognitive and affective processes, and the social contexts that activate these psycholo-

gical mechanisms. These 'social–cognitive' theories aimed to account for both trait-like consistency and situation-to-situation variability in behavior (see *Personality Theory and Psychopathology*).

2. Distinctions among Contemporary Approaches to Personality Assessment

The diversity of contemporary assessment tools can be organized according to a set of distinctions that differentiate alternative assessment efforts.

2.1 Targets of Assessment

One distinction concerns the target of personality assessment; that is, what it is that one is trying to assess. The most basic differentiation separates (a) the assessment of overt, surface-level tendencies in experience and action from (b) the assessment of internal personality structures and dynamics. Phrased simply, one may assess either phenotypic tendencies or genotypic psychological systems that causally contribute to those tendencies.

Whether one is assessing surface-level tendencies or internal psychological systems, one still must choose the precise tendencies or systems to target. These decisions generally are guided by broad theoretical principles about the nature of personality. With respect to surface-level phenotypes, the most common choice is to assess mean-level behavioral tendencies. Assessors commonly compute people's average tendency to display a particular type of behavior by aggregating together behavioral reports from diverse social contexts. This choice reflects the thinking of trait theories of personality, which view average tendencies in action as the defining expression of personality. In contrast, others caution that an assessment of mean-level tendencies is a necessary but insufficient step. It is insufficient because other features, such as variation in action across contexts (Mischel and Shoda 1995) or time (Larsen 1989), also distinguish individuals from one another. The assessment of contextual variation in action is consistent with social–cognitive theories of personality (Bandura 1986, Cervone and Shoda 1999), as detailed in Section 3.2.

Regarding the underlying genotypes, different theories again suggest different targets of assessment. Psychodynamic views explain behavior in terms of internal psychic conflicts that involve material of which persons are not consciously aware. Assessment thus must target these nonconscious psychological systems; projective tests such as the Rorschach inkblot text and the Thematic Apperception Test are designed for this purpose (see *Projective Methods in Psychology*). Social–cognitive theories contend that the core features of personality are cognitive capabilities through which people acquire social skills, reflect upon themselves, evaluate their actions, and regulate their behavior and emotional experience in accord with personal goals and standards for performance (Bandura 1986) (see *Social Cognitive Theory and Clinical Psychology*). This view dictates that assessments target an interacting system of cognitive and affective mechanisms (Cervone et al. 2001). Both psychodynamic and social–cognitive theories anticipate that there may be complex, nonlinear relations between underlying psychological systems and overt behavioral patterns. In contrast, some trait theories of personality propose that internal personality structures and overt behavior are related in a relatively direct, linear manner. Personality is construed as a small set of inherited trait structures (e.g., conscientiousness, extraversion), each of which fosters a characteristic mean level of phenotypic thoughts, feelings, and actions (e.g., conscientious acts, extraverted acts; McCrae and Costa 1996). Finally, some theorists study the role in personality functioning of personal narratives that individuals construct over the course of life (McAdams 1996). Their work suggests that personality assessment should include procedures (e.g., oral interviews or written narratives) to assess individuals' life stories.

2.2 Nomothetic vs. Idiographic Assessment

A second distinction involves overarching strategies of assessment. One strategy is to focus on an individual person and determine the potentially unique constellation of psychological variables that best characterizes that individual. Another is to focus first on the variables, and to seek a set of personality variables and associated assessment procedures that can be used to characterize the personality of any and all individuals. The former approach is an 'idiographic' strategy (from the Greek *idios*, referring to personal, private, and distinct characteristics), whereas the latter approach is 'nomothetic' (from the Greek for law, *nomos*, used here to refer to the search for universal scientific laws). A typical nomothetic strategy might use standardized tests to determine individuals' relative standing on one or more individual-difference variables. Idiographic assessment might involve interviews, unstructured self-descriptions, or biographical analyses of an individual's life.

Personality assessment may blend nomothetic and idiographic procedures. For example, one may have the same assessment goals and employ the same general testing procedures with all individuals. However, the exact content of the test items one employs may vary from person to person. Kelly's (1955) Role Construct Reperatory test is a classic example of a test that combines nomothetic procedures with idiographic content. The assessor's goal is always to identify the ideas, or constructs, that people use to understand their world. The testing procedure is

always one in which test takers are asked to indicate how a set of target persons is similar or different from one another. The exact content of the test items, however, varies idiographically. Each test taker provides a personal list of individuals who are of importance to him or her. This personalized list then comprises the list of target persons employed in the test. Contemporary assessments of belief systems and social-cognitive processes commonly mix idiographic and nomothetic methods in this manner (Caprara and Cervone 2000).

In practice, personality assessment has been dominated by the use of nomothetic procedures. The most commonly employed assessment instruments, such as the Cattell 16 P.F. Inventory, the Eysenck Personality Questionnaire (EPQ), the Minnesota Multiphasic Personality Inventory (MMPI), or the revised NEO personality inventory (NEO-PI-R), are nomothetic devices (see *Minnesota Multiphasic Personality Inventory (MMPI)*). Assessment consists of the administration of a standardized test that is used to rank individuals on a set of individual-difference dimensions.

2.3 Sources of Data

Whether one is employing a nomothetic or idiographic strategy, another question faced by the personality assessor is the source of data to draw upon. Block (1968) has drawn useful distinctions among alternative data sources; the following section modifies and elaborates upon his distinctions in light of contemporary research developments.

One can delineate at least seven sources of data that may be used in personality assessment

(a) *Behavioral Observations in Naturalistic Settings* are direct observations of a person's everyday behaviors, or the analysis of biographical records (e.g., school attendance records, records of legal infractions) that provide an objective index of such behaviors.

(b) *Behavioral Observations in Experimentally Constructed Settings* are observations of a person's behavior in situations designed by a researcher to constitute a context that is relevant to a personality construct or constructs under study. The test taker's reactions in that context are interpreted as indicators of the construct. Although the use of experimentally constructed settings has the advantage of greater experimental control, it also has a significant disadvantage. Personality characteristics often reveal themselves in the way people select environments; that is, in their choices to enter some settings and avoid others. Assessment procedures that confront all individuals with a fixed set of circumstances are insensitive to the influence of personality factors on the selection of environments.

(c) *Explicit Self-reports* ask people to describe their typical behaviors, thoughts, feelings, or personal characteristics. This is most commonly done through standardized multi-item tests, although explicit self-reports also may involve less structured techniques such as the writing of self-descriptive narratives.

(d) *Implicit Self-reports* are designed to tap beliefs or self-evaluations of which individuals may not be explicitly aware. For example, reaction-time techniques can reveal the degree to which individuals associate concepts with one another; an implicit test of self-esteem might use reaction-time measures to determine how strongly the concept of 'self' is linked to concepts of positive vs. negative emotional tone (Greenwald and Banaji 1995).

(e) *Observer Reports* are assessment procedures in which peers, family members, or other persons familiar with an individual are asked to evaluate their typical behaviors, feelings, or personal characteristics.

(f) *Psychophysiological Indices* include any of a wide variety of indicators of neural or biochemical systems that directly bear upon personality functioning and psychological differences among individuals. For example, electrophysiological measures are used to assess individual differences in cortical brain activity that may be a biological basis of overt individual differences in extraversion (Geen 1997).

(g) *Cognitive Indices* are assessments of the content or organization of a person's knowledge about themselves and the world. Many of the most important differences among individuals involve differences in knowledge and beliefs systems. A diverse set of tools, many deriving from the study of social cognition, have been employed to assess enduring individual differences in the content, complexity, and organization of knowledge structures that underlie personality functioning and individual differences (Caprara and Cervone 2000).

2.4 Comprehensiveness of Assessment

A final distinction concerns the comprehensiveness of personality assessment. Assessors sometimes wish to obtain a relatively complete portrait of the dispositional tendencies or underlying psychological dynamics that characterize an individual. Alternatively, one may desire merely to tap one or two personality variables of interest. The term 'personality assessment' commonly is applied to both comprehensive assessment efforts and measures of individual personality variables.

3. Illustrative Assessment Strategies: Five-factor Theory and Social–cognitive Theory

The conceptual distinctions drawn above are somewhat abstract. A more concrete understanding of alternative personality assessment strategies can be obtained by considering some examples of theory-

driven assessment. Two examples are considered here; namely, the personality assessment procedures that derive from five-factor theory (McCrae and Costa 1996) and social–cognitive theory (e.g., Bandura 1986, Mischel and Shoda 1995, reviewed in Cervone and Shoda 1999).

3.1 Assessing Individual Differences in Global Dispositional Tendencies; The 'Big Five' or 'Five-factor' Model

As noted above, the most common approach to personality assessment is to rank individuals on nomothetic individual-difference dimensions. The dimensions commonly represent global surface-level tendencies; that is, average tendencies to display behaviors that are representative of a given dispositional category. In a comprehensive assessment effort, individuals are ranked on a system of *n* individual-difference dimensions. The individual's personality is then represented as a point in *n*-dimensional space.

A primary question that arises is: What are the *n* dimensions? What set of dimensions is necessary and reasonably sufficient to capture individual differences in the population? A corollary question is methodological: How can one identify these dimensions? On the latter question, there has long been consensus. Investigators conduct factor analyses of the dispositional tendencies of large populations of individuals. The resulting factors then constitute the dimensions that form the core of personality assessment. Despite consensus on methods, on the former question there historically has been disagreement. Investigators in the mid-twentieth century proposed factor-analytic structures containing as many as 16 and as few as two dimensions. A major development in the 1980s and 1990s is that investigators achieved substantial consensus on the number and the nature of the factors required to assess phenotypic individual differences. Consensus centered on a five-dimensional trait taxonomy known as the big five (John and Srivastava 1999) or five-factor (McCrae and Costa 1996) model. The factors represent global individual differences in (a) extraversion or energy, (b) emotional stability or neuroticism, (c) conscientiousness, (d) agreeableness, and (e) openness to experience, particularly novel intellectual or cultural experiences. (Some interpret the fifth factor as being closer to intellect than to openness.) This five-dimensional structure has been identified in both self-reports and observer-reports, in assessment using both trait adjectives and items from professional personality inventories, and in both English (the model's language of origin) and other Indo–European and non-Indo–European languages.

The five-factor model provides an appealingly simple solution to the problem of assessing phenotypic tendencies. One merely plots people's average tendencies to exhibit actions indicative of each of five dispositional categories. Assessors also may seek more detailed information, such as people's tendency to display behaviors that are representative of narrower trait dimensions that are hierarchically related to the primary factors (in the way that sociability, for example, is hierarchically related to extraversion). Nonetheless, measuring the big five remains the core assessment task.

To some, the model also provides an appealingly simple solution to the problem of assessing causal genotypic structures. Although many proponents of the big five structure treat it merely as a descriptive model (Goldberg 1993), some suggest that it also is explanatory. In McCrae and Costa's (1996) five-factor theory, the factors are construed as 'dimensions of individual differences in tendencies to show consistent patterns of thoughts, feelings, and actions … (and) also … a property of an individual that accounts for his or her placement along this trait dimension' (McCrae and Costa 1996, p. 235). 'Agreeableness,' for example, refers simultaneously to a person's tendency to exhibit the behaviors one calls 'agreeable' and to an internal structure that corresponds to, and purportedly explains, this overt tendency.

This theoretical position has an interesting implication for assessment. It dissolves the distinction between the assessment of overt tendencies and internal psychological systems (Sect. 2.1, above). Five-factor theorists assess an individual's average behavioral tendency and interpret the measure as an index of both what the person does (an overt behavioral disposition) and what the person has (an internal psychological structure).

Although the simplicity of this approach may be appealing, its underlying logic can be questioned. Two concerns are of particular note. First, the five factors are identified in analyses of populations, yet are assumed to exist in the psychological make-up of each and every individual in the population (McCrae and Costa 1996). On purely statistical grounds, there is no reason to expect that group-level statistical parameters will replicate themselves at the level of each member of the group. (As a simple example, a group may average 40 years of age, yet no one in the group may be 40 years old.) Second, five-factor theory explains overt dispositional tendencies (e.g., a person's consistently agreeable acts) by inferring hypothetical causal constructs that contain the very properties that one is supposed to explain (e.g., agreeableness). This approach violates a standard principle of scientific explanation. As Hanson (1961, pp. 120–1) explained, 'What requires explanation cannot itself figure in the explanation,' for example, 'if the colors of objects are to be explained by atoms, then atoms cannot be colored' (also see Nozick 1981). This principle of explanation suggests that one should not explain dispositional tendencies by inferring psychological

constructs that directly embody those tendencies. It follows, then, that the assessment of surface-level tendencies and of internal psychological systems that explain those tendencies should be construed as separate tasks (Cervone et al. 2001).

3.2 Assessing Cognitive and Affective Systems and Persons-in-context: Social–cognitive Theory

An approach to personality and assessment that differs markedly from five-factor theory is that of social–cognitive theory (reviewed in Cervone and Shoda 1999). Social–cognitive theory rests on three main principles. Persons and sociocultural environments are viewed as reciprocally interacting systems. The core person variables are cognitive and affective mechanisms through which people interpret the world, reflect upon themselves, and self-regulate their behavior. Finally, personality is viewed as a complex, dynamic system; social–cognitive and affective mechanisms function in concert as interacting systems that underlie the coherence of personality functioning (Cervone and Shoda 1999).

This theoretical perspective has significant implications for assessment. One implication is that assessments of surface-level behavioral tendencies and of underlying psychological systems must be kept distinct. If overt tendencies reflect complex interactions among multiple underlying social–cognitive mechanisms, then there may be no simple relation between an overt behavioral pattern and any individual social-cognitive variable. The assessment of overt tendencies then cannot substitute for the assessment of internal structures. A second implication is that personality assessment must be contextualized. People's self-regulatory abilities enable them to vary their behavior strategically in accord with the perceived opportunities and demands of environmental settings. One must attend to these contextual variations to appreciate fully the distinctive features of an individual's personality (Mischel and Shoda 1995). A third implication is that personality assessment should include more than just an assessment of current dispositional tendencies. People possess potentials that may not be reflected in their typical daily activities. A person may, for example, possess a potential for leadership or for parenthood that only manifests itself when relevant circumstances arise (Caprara and Cervone 2000). Social–cognitive assessors commonly target skills and belief systems through which individuals contribute to their own development over the course of time (Bandura 1986). Finally, social–cognitive theory and research indicates that the content of personal beliefs, and the interconnections among elements of personal and social knowledge, may vary idiosyncratically from one person to another (Cantor and Kihlstrom 1987). This implies that assessments must be highly sensitive to individual idiosyncracy. Social–cognitive investigators commonly assess personal belief systems through idiographic methods (Cervone et al. 2001).

4. General Comments

The divergent conceptions of personality assessment that derive from five-factor theory as compared to social–cognitive theory underscore the more general point made earlier. To assess personality, one must rely upon a theory of personality. Questions of personality assessment and personality theory are inevitably intertwined.

In recent years, basic research on biological, affective, and cognitive processes increasingly has informed the study of molar behavioral patterns and differences among individuals. These advances have begun to yield an integrative psychology of personality in which diverse sources of knowledge are brought to bear upon questions of personality functioning and individual differences (Caprara and Cervone 2000). An ongoing challenge for personality assessors, then, is further to develop assessment techniques that do not merely describe surface-level behavioral patterns, but also shed light on specific underlying psychological processes that contribute to the consistency and coherence of personality.

See also: Interactionism and Personality; Minnesota Multiphasic Personality Inventory (MMPI); Personality and Conceptions of the Self; Personality Development and Temperament; Personality Development in Adulthood; Personality Development in Childhood; Personality Psychology; Personality Psychology: Methods; Personality Structure; Personality Theories; Personality Theory and Psychopathology; Projective Methods in Psychology; Social Cognitive Theory and Clinical Psychology; Social Learning, Cognition, and Personality Development

Bibliography

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall, Englewood Cliffs, NJ

Block J 1968 Personality measurement: Overview. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*, Vol. 12. Macmillan, New York, pp. 30–7

Cantor N, Kihlstrom J F 1987 *Personality and Social Intelligence*. Prentice-Hall, Englewood Cliffs, NJ

Caprara G V, Cervone D 2000 *Personality: Determinants, Dynamics, and Potentials*. Cambridge University Press, New York

Cervone D, Shadel W G, Jencius S 2001 Social-cognitive theory of personality assessment. *Personality and Social Psychology Review* **5**: 33–51

Cervone D, Shoda Y (eds.) 1999 *The Coherence of Personality: Social-cognitive Bases of Consistency, Variability, and Organization*. Guilford, New York

Cronbach L J, Meehl P E 1955 Construct validity in psychological tests. *Psychological Bulletin* **52**: 281–302
Geen R G 1997 Psychophysiological approaches to personality. In: Hogan R, Johnson J A, Briggs S R (eds.) *Handbook of Personality Psychology*. Academic Press, San Diego, pp. 387–414
Goldberg L 1993 The structure of phenotypic personality traits. *American Psychologist* **48**: 26–34
Greenwald A G, Banaji M R 1995 Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* **102**: 4–27
Hanson N R 1961 *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press, Cambridge, UK
John O P, Srivastava S 1999 The big-five factor taxonomy: History, measurement, and theoretical perspectives. In: Pervin L A, John O P (eds.) *Handbook of Personality: Theory and Research*. Guilford, New York, pp. 102–138
Kelly G 1955 *The Psychology of Personal Constructs*. Norton, New York
Larsen R J 1989 A process approach to personality psychology: Utilizing time as a facet of data. In: Buss D M, Cantor N (eds.) *Personality Psychology: Recent Trends and Emerging Directions*. Springer-Verlag, New York, pp. 177–93
McAdams D P 1996 Personality, modernity, and the storied self: A contemporary framework for studying persons. *Psychological Inquiry* **7**: 295–321
McCrae R R, Costa P T 1996 Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In: Wiggins J S (ed.) *The Five-factor Model of Personality. Theoretical Perspectives*. Guilford, New York, pp. 51–87
McCrae R R, Costa P T 1995 Trait explanations in personality psychology. *European Journal of Personality* **9**: 231–252
Mischel W 1968 *Personality and Assessment*. Wiley, New York
Mischel W, Shoda Y 1995 A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review* **102**: 246–86
Nozick R 1981 *Philosophical Explanations*. Belknap Press of Harvard University Press, Cambridge, MA

D. Cervone and G. V. Caprara

Personality Development and Temperament

Personality refers to the profiles of stable beliefs, moods, and behaviors that differentiate among individuals in a particular society. The exact form of expression and the number of profiles will vary with culture, as the number and variety of animal species vary with the ecological setting. New England Puritans would have regarded variation in piety as a major psychological trait, and citizens in Confucian China would have treated loyalty to the father as a central personality characteristic. Neither piety nor filial loyalty are regarded as primary personality traits by contemporary psychologists.

The personality types believed to be most prevalent in contemporary industrialized societies refer to usual styles of social interaction; degree of adherence to the ethical standards of the community; vulnerability to the emotions of anxiety, guilt, and anger; and receptivity to new ideas, beliefs, and attitudes. However, among individuals living in ancient agricultural societies, composed of small, stable geographically separate villages, variation in sociability would be less critical for adaptation than this trait is in the large cities of contemporary, industrialized nations.

The term temperament, on the other hand, refers to the profiles of mood, behavior, and physiology, believed to be under some genetic control, that appear relatively early in life and, together with experience, become the major personality traits of individuals in a particular society (Bates 1987). Temperamental qualities in children believed to be important at the start of the twenty-first century refer to ease of arousal to stimulation; the form and efficiency with which that arousal is regulated; activity level; irritability; form of reaction to unfamiliar people, objects, and situations; and capacity for sustained attention.

Scholars who believe that temperamental categories are applicable to adults have nominated avoidance of danger, seeking of novelty, dependence on social rewards, emotionality, activity, and sociability as primary temperamental characteristics (Buss and Plomin 1984).

1. *Temperamental Contributions to Personality*

The assumption that a young child's temperament makes a contribution to the older individual's personality has a long history. Hippocrates and Galen, two physicians living in ancient societies who were the founders of the temperamental theory, believed that the relative concentrations of four body humors—blood, phlegm, and yellow and black bile—created within each person a combination of the opposed dimensional qualities of warm versus cool and dry versus moist. The four temperamental types, called melancholic, sanguine, choleric, and phlegmatic, were the products of a particular combination of these two dimensions which were influenced, in part, by local climate and the person's diet. Melancholics were high on the qualities of cool and dry because they possessed an excess of black bile and that is why they were tense, anxious, and depressed. Sanguine individuals were high on the warm and moist qualities because of an excess of blood and were outgoing and optimistic. The choleric was high on the qualities warm and dry because of an excess of yellow bile, and, as a result, was easily angered. The phlegmatic was high on the qualities of cool and moist because of an excess of phlegm and was low in energy and initiative. These concepts remained popular among Europeans and Americans until the end of the nineteenth century.

Interest in the contributions of temperament to

personality receded during the first half of the twentieth century when political forces led to a rejection of the idea that biological processes contributed to personality. However, the contribution of temperament returned when the psychiatrists Alexander Thomas and Stella Chess (1977) nominated nine primary temperamental dimensions in children, along with three synthetic types that they called difficult child, easy child, and the child who is slow to warm up to unfamiliar situations. The difficult child, which was the least frequent category and comprised about 10 percent of the study group, was rather more likely than the other two to develop psychiatric symptoms in later childhood.

2. *Reactivity to Stimulation in Infants*

Ease of behavioral and physiological arousal to varied forms of stimulation, and the form of regulation of that arousal, are primary temperamental qualities in infants and young children (Rothbart 1989, Kagan 1994). Some infants become easily aroused by sight, sounds, and smells and thrash, cry, babble, or smile at these events. Other infants show minimal signs of arousal to the same stimulation. It is believed that about 20 percent of healthy, four-month-old infants react to varied forms of stimulation with frequent distress and vigorous motor activity. One-fourth of these infants, who are called high reactive, become very subdued, shy, fearful children in the second and third years and are more likely than others to develop anxious symptoms by school age. By contrast, about 40 percent of healthy infants show the opposite profile of a relaxed and minimally distressed reaction to stimulation. About one-fourth of these infants, who are called low reactive, become sociable, bold, and affectively spontaneous preschool children and minimally anxious seven year olds (Kagan 1994). A temperamental vulnerability to anxiety helps to explain why only a minority of children react to a particular stressful event with an anxious symptom, whether the traumatic event is an earthquake, kidnapping, divorce, or abuse. Only a fairly small proportion of children, usually less than 40 percent, react to the traumatic event with a fearful symptom or chronic anxiety.

Social experiences, especially experiences in the family and with peers, determine whether high reactive infants will develop a fearful personality and whether low reactive infants become bold extraverts. Hence, only about 15 percent of children who are extremely shy and anxious in the preschool years become introverted adolescents or adults (Caspi et al. 1988), and both the anxious-shy and the bold-sociable types are under modest genetic control.

There is disagreement as to whether temperament should be viewed as a continuous dimension or as a qualitative category. For example, some regard the tendency to approach or to withdraw from unfamiliar people and situations as a continuous dimension, while others believe that the child who usually approaches unfamiliar people and events and the child who usually avoids these events belong to separate, discrete categories. Because only about 15 percent of young children are consistently shy, avoidant, and timid in a large number of unfamiliar situations, while most children are subdued in only one type of situation, it is likely that the former child belongs to a qualitatively discrete group.

Most scientists believe that each of the many temperamental types possesses a unique set of inherited physiological features. Although most of these biological features are not yet discovered, it is known that fearful-avoidant children show greater cortical activation of the right frontal, compared with the left frontal, area as measured by absence of alpha activity in the EEG on the right side (Davidson 1994, Fox et al. 1994).

The brain contains over 150 different chemical substances, each of which influences the excitability of particular ensembles of neurons and brain circuits. It is assumed that children inherit different concentrations of these molecules, as well as different densities of the relevant receptors present on neuronal surfaces. As a result children are disposed to different mood and behavioral reactions. One example involves an enzyme called dopamine beta-hydroxylase. This enzyme, which is under genetic control, is required for the final step of the synthesis of norepinephrine. Children who have low levels of this enzyme will necessarily have less norepinephrine in their brain. Because brain norepinephrine renders a child vigilant and may increase the probability of a state of fear to novelty, children with low levels of the enzyme would be expected to be less fearful. Boys who are very aggressive or antisocial, often called conduct disorder, seem to be unafraid of punishment. These children have low levels of this enzyme (Rogeness et al. 1988). Dopamine, another essential brain neurotransmitter, affects motor behavior and cognitive processes mediated by the frontal cortex. Because the frontal cortex mediates planning and control of impulsivity, it is possible that children who inherit low levels of dopamine in the frontal area will be unusually impulsive. The rationale for administering the drug Ritalin to children who have been diagnosed as impulsive or hyperactive is that this drug increases the level of dopamine in the frontal cortex and, therefore, helps children control impulsive and potentially inappropriate behavior.

3. *The Role of Environment*

Children are exposed to different environments over the course of childhood and these environments influence temperament in different ways. Therefore, there will be a large number of different personality

profiles emerging from the combination of early inherited temperament and experience. Some of the important experiences include the socialization practices of the parents, the relationship to brothers and sisters, success or failure in school, and the quality of relationships with peers. The social class of the child's family always exerts continuous influence on the personality that develops. A large birth cohort of children born on the island of Kauai followed from birth to the fourth decade of life revealed that the social class of the child's family and early temperament were the two most important influences on the adult personality. Minimally irritable infants who were born to highly educated families were least likely to encounter school failure, display delinquent behavior, or to develop symptoms requiring referral to a physician or psychologist (Werner 1993).

3.1 Types of Challenges

The most frequent personality profiles that develop in a particular culture are a function of the challenges to which the children in that culture must accommodate. Children in most settings deal with three classes of challenge: (a) unfamiliarity, especially unfamiliar people, tasks, and situations; (b) requests by legitimate authority for conformity to and acceptance of their standards, especially standards for competent task performance; and (c) domination by and/or attack from peers. In addition, all children must learn to control two families of emotions: anxiety, fear and guilt, on the one hand, and, on the other, anger, jealousy, and resentment.

4. Temperament and Moral Emotions

Variation in the intensity of the moral emotions of shame, guilt, and anxiety can be influenced by the child's temperament and family experience. Shy, fearful children raised by mothers who used reasoning to socialize the child showed clearer signs of a strict conscience than most other children (Kochanska 1993). Sensory information from the body ascends in the spinal column and projects to the ventromedial surface of the prefrontal lobe. Children who experience consciously more intense bodily stimulation might be more vulnerable to the moral emotions than those who experience less stimulation. School-age boys who were low reactive infants and fearless children often have low sympathetic tone in the cardiovascular system. Boys who are members of this temperamental group growing up in homes with nurturant parents who socialize achievement and the control of aggression are likely to become popular group leaders. The same type of child raised by parents who did not socialize asocial behavior and who played with peers who committed asocial acts are at slightly higher risk for becoming delinquents (Caspi et al. 1995). It is believed that a small number of criminals who commit violent crimes (probably fewer than 5 percent of all criminals) were born with a special temperament characterized by impulsivity and difficulty in controlling behavior.

5. Ethnicity and Temperament

Temperamental variation among ethnic groups is a delicate issue because of the ethnic strife that exists in many parts of the world. The average genetic distances for 100 different alleles monitoring physiological function measured in geographically separate populations were largest when Asians, Africans, and European-Caucasians were compared with each other. It is likely that some of these alleles make a modest contribution to variation in temperamental qualities. Asian-American infants, compared with European-Americans, are calmer, less labile, less vocal, and more easily consoled when distressed (Kagan et al. 1994). It is relevant that Asian-American adult patients with a psychiatric diagnosis require a lower dose of therapeutic drug than European-American patients (Lin et al. 1986).

6. Personal Control

The renewed interest in temperament has generated discussion over the degree of responsibility each person should have for his or her behavior. Western society has, in the past, been Puritan in its commitment to the belief that all individuals have sufficient will power to control their behavior in most situations. This view has been eroded since the 1950s as more citizens have become willing to excuse some asocial actions as due to temperamental factors that were not within the agent's sphere of control. It is not obvious that this permissiveness is more adaptive for a society than the traditional assumption that one of the significant products of human evolution is the capacity to monitor one's behavior.

See also: Personality Assessment; Personality Psychology: Methods; Temperament and Human Development

Bibliography

Bates J E 1987 Temperament and infancy. In: Osofsky J E (ed.) *Handbook of Infant Development*, 2nd edn. Wiley, New York, pp. 1101–49

Buss A H, Plomin R 1984 *Temperament—Early Developing Traits*. Erlbaum, Hillsdale, NJ

Caspi A, Elder G H, Bem D J 1988 Moving away from the world. *Developmental Psychology* **24**: 824–31

Caspi A, Henry B, McGee R O, Moffitt T, Silva P A 1995 Temperamental origins of child and adolescent behavior problems from age 3 to 15. *Child Development* **66**: 55–68

Davidson R J 1994 Asymmetric brain function and affective style in psychopathology. *Development and Psychopathology* **6**: 741–58
Fox N A, Calkins S D, Bell M A 1994 Neuroplasticity and development in the first two years of life. *Development and Psychopathology* **6**: 677–96
Kagan J 1994 *Galen's Prophecy*. Basic Books, New York
Kagan J, Arcus D, Snidman N, Yu-feng W, Hendler J, Greene S 1994 Reactivity in infancy. *Developmental Psychology* **30**: 342–5
Kochanska G 1993 Toward a synthesis of parental socialization and child temperament in early development of conscience. *Child Development* **64**: 325–47
Lin K M, Poland R E, Lesser I N 1986 Ethnicity and psychopharmacology. *Culture, Medicine, and Psychiatry* **10**: 151–65
Rogeness G A, Maas J W, Javors M A, Masedo C A, Harris W R, Hoppe S K 1988 Diagnoses, catecholamines, and plasma dopamine-beta-hydroxylase. *Journal of the American Academy of Child and Adolescent Psychiatry* **27**: 121–5
Rothbart M K 1989 Temperament in childhood: A framework. In: Kohnstamm G A, Bates J E, Rothbart M K (eds.) *Temperament in Childhood*. Wiley, New York, pp. 59–76
Thomas A, Chess S 1977 *Temperament and Development*. Brunner-Mazel, New York
Werner E E 1993 Risk, resilience and recovery. *Development and Psychopathology* **5**: 503–15

J. S. Kagan

Personality Development in Adulthood

The study of personality is arguably the broadest subdiscipline in psychology in that it aims to understand how the whole person functions in the world (see *Personality Psychology*). Historically, personality psychology has concerned itself with grand theories about human behavior, questions about character, identity, and morality. Most of the empirical research on personality published over the past two decades fails to consider potential development in adulthood, reflecting, in part, an assumption that personality changes little in adulthood. Nevertheless there has been a long-standing interest in whether—and, if so, how—people change in systematic ways in the later years of life. The following sections summarize the research traditions and some of the central findings generated in the field of adult personality development.

1. *Defining Personality Development*

In part because personality draws so broadly from so many areas of psychology, including cognition, emotion, psychopathology, and motivation, consensus in the field over basic definitions is difficult to obtain (see *Personality Theories*). Even the preferred units of study (e.g., traits, behaviors, or psychological processes) continue to be debated. Most recently, questions are being raised about the degree to which personality is bounded within the individual or is better represented in transactions between individuals and broader collective units. Importantly, different conceptual approaches to the study of personality lead to different predictions about and evidence for personality change in adulthood. Whereas personality development is considered by some to unfold naturally and unidirectionally from temperaments inherited at birth, other personality psychologists emphasize the need to consider a complex interplay of factors that contribute to personality development, including temperamental inheritance but more importantly exposure to different types of environments, acquired beliefs and expectations and the capacity for self-regulation (Bandura 1989).

In part, evidence for stability or change in personality in adulthood reflects which facet of human functioning is studied. To be clear, there is no dispute that people change in adulthood. They do. Adults are inevitably changed in idiosyncratic ways by the life experiences they encounter, including major life events, such as becoming a parent, or less dramatic but persistent experiences associated with, for example, the pursuit of a particular career and the consequent development of a particular type of expertise. However, change in adulthood is not automatically considered *personality development*. Rather, changes must be enduring, systematic (i.e., nonrandom), and predictable by age or life stage.

2. *Major Approaches to Adult Personality Development*

Conceptions of adult personality development have evolved out of two very different traditions in psychology; clinical psychology and life-span developmental psychology. The approaches and the findings about systematic change in adulthood reflect these different paradigmatic approaches to the study of lives. Due to considerable overlap between studies of individual differences and clinical psychology, many of the oldest and most influential theories of personality, most notably psychoanalytic psychology but also ego psychology and interpersonal psychology, were developed based on clinical observations of patients. Although strikingly different in its basic tenets, social cognitive theory of personality also evolved out of close connections between clinical and personality psychology, essentially addressing differences between normal and abnormal processing involved in basic psychological functioning. Thus, the oldest approaches to personality were tied closely to understanding psychopathology.

In contrast to the traditional individual difference approach, life-span psychology was born only in about

the 1970s and reflects the view that human development is a continuous adaptive process (Baltes and Goulet 1970). Life-span psychology aims to identify and illuminate normal developmental changes in all areas of psychological functioning, including but not limited to personality, from birth until death. Perhaps most notably, life-span psychology is distinguished by the presumption that human growth is at no time during the life course complete. Consequently, the different presumptions inherent in the two approaches direct attention to different research foci. Whereas traditional adult personality psychologists ask whether traits acquired in childhood persist across adulthood, whether particular personality crises present themselves at particular stages of life, or how personality disorders are transformed in later life, life-span psychologists are more likely to target specific age-related issues, like whether people grow wiser with age, whether conceptions of the self grow more or less complex over time, and whether self-regulatory processes change in systematic ways over time. As noted above, whether one finds evidence for stability or change depends importantly on where one looks and the particular questions one poses.

The next sections include a birdseye view of the earliest approaches to personality development and brief synopses of research on personality development deriving from the trait approach to personality development and from life-span developmental psychology.

3. *Early Stage Approaches to Adult Personality Development*

Following the tradition established by child developmentalists, early thinking about adult personality development was rooted in stage theories. Whereas Freud's (see *Freud, Sigmund (1856–1939)*) psychosexual stage model of personality suggested that personality was formed in early childhood and that, barring long-term psychotherapy, was highly resistant to change, his skeptical follower, Carl Jung (see *Jung, Carl Gustav (1875–1961)*), argued that the most interesting aspects of personality do not develop fully until middle age. Jung believed that only after basic biological and reproductive issues are resolved in early adulthood, are people freed to engage in more psychic pursuits. Jung posited that whereas in early adulthood biological imperatives predominate and demand adherence to gender roles, in mid-life feminine and masculine sides of people grow more integrated and spiritual concerns grow more salient. As people age, feelings and intuitions come to dominate thoughts and sensations. Thus, albeit in rather unspecified ways, Jung advanced the position that people developed in adulthood and that spirtuality played an increasingly central role. Jung wrote far less about advanced old age, but suggested that people predictably turned inward and deteriorated psychologically as they approached the end of life and wrestled with fears about death.

In the 1930s and 1940s, in Europe and the US, several stage stages theories, such as that offered by Charlotte Bühler, concretized thinking about adult personality development and allowed for empirical tests of theoretical predictions. Of these, Erik Erikson's (see *Erikson, Erik Homburger (1902–94)*) stage theory had the most enduring influence. Like Freud and Jung, Erikson was a classic ego psychologist and his theory was grounded in the psychoanalytic tradition. However, rather than focus on psychosexual needs, Erikson argued that human needs for psychological intimacy fueled systematic development that continued into old age. According to this epigenetic theory, people pass through a fixed sequence of stages during life each of which requires the successful resolution of a central psychic crisis. In early adulthood, the central issue involves the establishment of intimacy. In middle age, generativity, namely the passing on of power and knowledge to younger generations, is achieved or failed. In old age ego-integrity vs. despair (self-acceptance of a life lived or regret and dismay) is the focal crisis in life.

In the 1960s and 1970s, major research projects aimed at profiling adult development were undertaken at Yale (Levinson 1978) and Harvard (Vaillant 1977), which continued the stage theory tradition in the US and longitudinal studies undertaken in the 1930s began to come of age. As research participants in the Stanford Terman study of 'gifted' children, for example, entered adulthood, researchers began to examine connections between childhood and adult personality. The Child Guidance and Oakland Growth Studies undertaken at the University of California at Berkeley offered resources by which to examine predictable change. At the University of Chicago, a group of social scientists, including Bernice Neugarten, Morton Lieberman, and David Guttman, formalized the study of life-span personality development.

In the end, however, empirical support for stage theories failed to withstand empirical tests. Although interesting patterns were displayed in some very homogenous samples, identified developmental patterns failed to generalize broadly. Critics of stage theories claimed that the patterns that did emerge reflected the influence of consistent surrounding social structures around highly selected groups of people, not human development. Virtually all of the longitudinal studies included predominately (if not exclusively) white, middle-class individuals, often only males. In the 1980s, Costa and McCrae (1990) essentially waged a war against stage theories declaring that personality does not change systematically in adulthood. In a decade dominated by the *trait* approach to personality, the central and reliable finding

of the 1980s was that personality changes little after the age of 30 years.

4. *The Trait Approach to Adult Personality Development*

Traits are continuous variables represented by broadly encompassing lexical terms that account for individual differences (John 1990). Traits—such as shy, lively, outgoing, anxious, and intelligent—are conceptualized as predispositions within individuals to behave in certain ways manifest across a wide variety of situations. Gordon Allport (see *Allport, Gordon W (1897–1967)*) argued that cardinal traits are those around which a person organizes life (self-sacrifice). Central traits (e.g., honesty) represent major features and secondary traits are specific traits that help to predict behavior more than underlying personality (e.g., dress type, food preferences). Allport's definition is compatible with modern trait and temperament approaches to personality which attempt to describe people in terms of one or more central features of the person.

Personality psychologists in the trait tradition seek to identify the traits along which people differ and to explore the degree to which these traits predict behavior. Many taxonomies of traits have been offered over the years, but unquestionably the five-factor model is most widely accepted today (see also *Personality Structure*). Based on factor analysis of self-descriptions, the five traits that emerge reliably across many studies of Europeans and Americans are: (a) openness to experience, (b) conscientiousness, (c) extraversion, (d) agreeableness, and (e) neuroticism.

Traits and temperaments appear to be relatively stable through the second half of life (Costa and McRae 1990). It appears that beyond the age of 30, extraverts remain extraverts and neurotics remain neurotics. Trait theorists have found reliable evidence for stability in personality well into old age. This finding emerges whether researchers ask individuals to describe themselves repeatedly over time or, alternatively, ask significant others, like spouses, to describe those same individuals repeatedly (Costa and McRae 1990). It should be noted that even though persistent rank-order differences remain the same, there is some recent evidence that modest mean level changes may appear, with older adults scoring slightly higher than younger adults on agreeableness and conscientiousness and slightly lower on neuroticism, extraversion, and openness to experience (McCrae et al. 1999). Importantly, similar findings come from studies sampling Asian and European populations. However, identified changes are quite small. Overall, there is remarkable consistency in the characteristics that distinguish individuals from one another over time. There is some evidence that the core set of traits that differentiate people are genetically based and exert their influence throughout the life course (Gatz 1992). Genetic influence is as strong in old age as early adulthood.

In summary, researchers adopting a trait approach to personality development find that along at least some of the important dimensions of personality, there is little change in personality well into old age. Critics of a trait approach, however, argue that traits communicate little about how people manage their lives in day-to-day life and because of the broadband focus exaggerate the consistency of behavior across time and situations. They criticize the trait approach for failing to better predict behavior and redirect focus to specific strategies (e.g., how an individual cognitively appraises a situation; expectancies, subjective values, self-regulatory systems, and competencies). Life-span approaches—influenced strongly by the social cognitive theory of personality (Bandura 1989)—view individuals as agentic creatures who shape their own environments (see also *Interactionism and Personality*; *Self-regulation in Adulthood*).

5. *Life-span Approaches to Adult Personality Development*

Rather than focus on taxonomies of personality, life-span developmental psychologists view development as a dynamic process aimed at adaptation (Baltes 1987). Two principal stays of life-span theory speak directly to personality. The first states that adaption is always time and space bound. In other words, behavioral adjustment must occur within a particular environmental and social niche. In life cycle context, what is adaptive in infancy and early childhood may not be adaptive in adolescence. Stranger anxiety, for example, may serve a highly adaptive function in infancy because it motivates dependent creatures to stay in close proximity to caregivers. It may also facilitate attachment to primary adult figures, a key developmental task of early life. Yet, stranger anxiety among adults is clearly maladaptive. Similarly, it can be argued that pursuing multiple prospective mates is adaptive in adolescence and early adulthood as people 'practice' intimate relationships but less so in middle and old age at which point emotional investment in a select few may hold greater gains than the continual exploration of all possible mates.

The second stay of life-span theory is that development inevitably demands selection (Baltes and Baltes 1990). In order for specialized (i.e., effective) adaptation to occur within a particular social, historical and physical niche, active and passive selections must be made. As people age, they come to have greater choice in the selection of environments and select environments that support their self-views. Throughout adulthood, people actively construct skills and hone environments to meet selected goals. There is good evidence that people narrow their social

spheres with age, for example, forming increasingly well contoured social convoys that accompany them throughout life (Carstensen et al. 1999). Caspi and Herbener (1990) found that people tend to choose spouses similar to themselves, and further show that people who have spouses similar to themselves are less likely than people with dissimilar spouses to display personality change in adulthood. Thus, it may be that stability is maintained across the life course because people actively create environments that maintain stability.

Finally, life-span theory holds that development is never fully adaptive because adaptation to one set of circumstances inevitably reduces flexibility to adapt to another. In this way development always entails gains and losses. Subsequently, life-span theory obviates the presumption that antecedent losses are the only or even the primary reasons for changes that occur with age and examines how people's relative strengths and weaknesses at different points in the life cycle influence successful adaptation.

5.1 Personality Development from a Motivational Perspective

Life-span developmental approaches, because they are rooted in adaptation, lead naturally to consideration of the ways that goals and goal attainment may change throughout the life course (Baltes and Baltes 1990, Brandtstädter et al. 1999, Carstensen et al. 1999) (see also *Adulthood: Developmental Tasks and Critical Life Events*). Motivational psychologists presume that there is continuity in basic human needs for competence, relatedness and autonomy across the life course. There is every reason to expect, for example, that regardless of age, people seek to control their worlds (see also *Control Behavior: Psychological Perspectives*).

Researchers who take a goal-focused approach have brought a different set of findings to bear on discussions of personality and aging, showing that goals and preferences do change with age and influence behavior. Carstensen and co-workers, for example, have shown that the perception of time left in life importantly influences social goals. Because aging is inextricably and positively associated with limitations on future time, older people and younger people differ in the goals they pursue (Carstensen et al. 1999). Older people are more likely to pursue emotionally meaningful goals whereas younger people are more likely to pursue goals that expand their horizons or generate new social contacts. Brandtstädter et al. (1999) argue that people adjust goal strivings to accommodate external and internal constraints placed on goal achievement at different points in the life cycle; a central finding coming out of this line of work, for example, is that older people respond to the loss of resources in advanced age by downgrading the importance of some previously desirable goals.

5.2 Emotion and Personality

Another way of conceptualizing personality, which is particularly conducive to life-span approaches, places emotional experience and emotional regulation at the core. In this view, emotions are not related simply to personality, they are the essence of personality (Carstensen et al. in press, Rothbart and Derryberry 1981) (see also *Adulthood: Emotional Development*). The emotions people feel when they face challenges, and the effectiveness with which they learn to regulate their emotions, are the cardinal components of personality development, forming the basis of individual differences in persistent tendencies to behave, think and feel in day-to-day life. Individual differences in the propensity to experience specific emotions influence not only the psychological and biological reactivity of the person in the moment, but come to influence conscious choices about preferred environments, behavioral styles, and also determine the social partners to which people are drawn.

Ryff (1995) has taken a differentiated approach to understanding emotions and well-being across the life-span. Rather than calculating global positive and negative affect as indicators of psychological well-being, Ryff conceptualizes well-being in terms of self-acceptance, environmental mastery, purpose in life, personal growth, positives relations with others, and autonomy. The dimensions appear to have differential relationships with age, with older adults scoring higher than younger adults on Environmental Mastery and Autonomy, but lower on Purpose in Life and Personal Growth. There also appear to be life-span developmental trajectories concerning the relationship between people's conception of their present status and their ideal selves along these dimensions. Older people tend to have less distance between their actual and ideal selves than do younger adults.

An emotion approach to personality has particularly intriguing implications for adult development because in adulthood emotions appear to grow more complex (Carstensen et al. in press) and emotion regulation appears to improve (Gross et al. 1997). With advancing age, the emotion-cognition-personality system also appears to become more-differentiated, with emotions becoming linked to an ever-increasing array of cognitions. To the extent that such changes influence motivation (e.g., Izard and Ackerman 1998), modify thinking and reasoning (Labouvie-Vief et al. 1989), or result in qualitative changes in subjective well-being (Ryff 1995), personality is importantly influenced.

5.3 Wisdom and Resilience

As noted above, the focus on adaptative development inherent in life-span approaches generates questions about the ways in which aging people adjust to changing resources and changing contexts. There is a

delicate balance between gains and losses that occurs in the second half of life that have important implications for personality. As people enter advanced ages, nearly inevitably they encounter increasingly difficult challenges, including the deaths of friends and loved ones, assaults on physical health, and threats to social status (see also *Coping across the Lifespan*). At the same time as experience in life increases, perspectives change. In some ironic way, the familiarity of losses may even make losses easier to take. Considerable focus in lifespan psychology, thus, has been on the ways that people effectively adjust in later adulthood. Resilience (Staudinger et al. 1995) and wisdom (Staudinger 1999) have been particular targets of interest because they involve the use of age-based experience to compensate for losses in circumscribed domains. Studies of wisdom, for example, show that contrary to popular lore, wisdom is unrelated to age in adulthood (Staudinger 1999). Even though experience-based knowledge does increase with age, wisdom requires a complex array of abilities that draw on multiple functions, some of which decline. Under optimal conditions, old age may be the time in life for wisdom to best emerge, but it does not do so normatively.

6. *Integration of Approaches and Findings about Adult Personality Development*

Does continuity or change in personality best characterize adulthood? The answer is a qualified 'yes' to both continuity and change. Along some basic descriptive dimensions, such as openness to experience and extraversion, people remain remarkably consistent in adulthood. However, in other domains just as central to personality, such as motivation and adaption, there is evidence for systematic change across adulthood. Goals change predictably with age, emotion regulation appears to improve, and well-being takes on different qualities.

7. *Future Directions*

The fundamental challenge that confronted personality researchers a century ago remains largely the same today: predicting and understanding individual differences in behavior. Students of adult personality development face the additional challenge of understanding how the differences that distinguish one person from another may change systematically over time. Although considerable progress has been made, the bulk of empirical findings generated simply show that prevalent assumptions about adult personality development in the twentieth century were wrong. For example, people do not appear to pass normatively through a fixed series of stages; and along broadband dimensions characterized as basic traits, people change very little in the second half of life. Approaches that focus on motivation and emotion are newer, but initial findings suggest that they may shed considerable light on ways that individuals change in middle and old age. Finally, at the time of this writing, the human genome project very recently was declared complete. Few, if any, scientists expect that genetic findings alone will shed much light on personality. However, they may well help to do away with customary language and algorithms (such as 'heritability coefficients') that have given credence tacitly to the idea that environmental and biological influences can be cleanly separated. Whereas in past decades, substantial discussion has centered around whether biology or environment was most influential in personality development, personality researchers will now begin to address the more interesting and more important puzzle which lies in the interaction between the two.

See also: Adult Development, Psychology of; Ego Development in Adulthood; Personality Development and Temperament; Personality Development in Childhood; Personality Psychology; Personality Theories

Bibliography

Baltes P B 1987 Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* **23**: 611–26

Baltes P B, Baltes M M 1990 Psychological perspectives on successful aging: The model of selective optimization with compensation. In: Baltes P B, Baltes M M (eds.) *Successful Aging: Perspectives from the Behavioral Sciences*. Cambridge University Press, New York, pp. 1–34

Baltes P B, Goulet L R 1970 Status and issues of life-span developmental psychology. In: Goulet L R, Baltes P B (eds.) *Life-span Developmental Psychology: Research and Theory*. Academic Press, New York, pp. 4–21

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall, Englewood Cliffs, NJ

Bandura A 1989 Human agency in social cognitive theory. *American Psychologist* **44**: 1175–84

Brandtstädter J, Wentura D, Rothermond K 1999 Intentional self-development through adulthood and later life: Tenacious pursuit and flexible adjustment of goals. In: Brandtstädter J, Lerner R (eds.) *Action and Self Development*. Sage, Thousand Oaks, CA

Carstensen L L, Charles S T, Isaacowitz D, Kennedy Q (in press) Emotion and life-span personality development. In: Davidson R, Scherer K (eds.) *Handbook of Affective Science*. Oxford University Press, Oxford, UK

Carstensen L L, Isaacowitz D M, Charles S T 1999 Taking time seriously: A theory of socioemotional selectivity. *American Psychologist* **54**: 165–81

Carstensen L L, Pasupathi M, Mayr U, Nesselroade J 2000 Emotional experience in everyday life across the adult life span. *Journal of Personality and Social Psychology* **79**: 644–55

Caspi A, Herbener E S 1990 Continuity and change: Assortative marriage and the consistency of personality in adulthood. *Journal of Personality and Social Psychology* **58**: 250–8
Costa P T Jr., McCrae R R 1990 *Personality in Adulthood*. Guilford Press, New York
Gatz M, Pederson N, Plomin R, Nesselroade J 1992 Importance of shared genes and shared environments for symptoms of depression in older adults. *Journal of Abnormal Psychology* **101**: 701–8
Gross J, Carstensen L L, Pasupathi M, Tsai J, Götestam Skorpen C, Hsu A 1997 Emotion and aging: Experience, expression and control. *Psychology and Aging* **12**: 590–9
Izard C, Ackerman B P 1998 Emotions and self-concept across the life span. In: Schaie K W, Lawton M P (eds.) *Annual Review of Gerontology and Geriatrics: Focus on Emotion and Adult Development*. Springer, New York, Vol. 17, pp. 1–26
John O 1990 The big-five factor taxonomy: Dimensions of personality in the natural language and questionnaires. In: Pervin L A (ed.) *Handbook of Personality: Theory and Research*. Guilford Press, New York, pp. 66–100
Labouvie-Vief G, DeVoe M, Bulka D 1989 Speaking about feelings: Conceptions of emotion across the life span. *Psychology and Aging* **4**: 425–37
Levinson D 1978 *The Seasons of a Man's Life*. Knopf, New York
McCrae R R, Costa P T Jr., de Lima M P, Simoes A, Ostendorf F, Angleitner A, Marusic I, Bratko D, Caprara G V, Barbaranelli C, Chae J-H, Piedmont R L 1999 Age differences in personality across the adult life-span. *Developmental Psychology* **35**: 466–77
Rothbart M, Derryberry D 1981 Development of individual differences in temperament. In: Lamb M E, Brown A L (eds.) *Advances in Developmental Psychology* Vol. 1, Erlbaum, Hillsdale, NJ, pp. 37–86
Ryff C 1995 Psychological well-being in adult life. *Current Directions in Psychological Science* **4**: 99–104
Staudinger U M 1999 Older and wiser? Integrating results from a psychological approach to the study of wisdom. *International Journal of Behavioral Development* **23**: 641–4
Staudinger U M, Marsiske M, Baltes P B 1995 Resilience and reserve capacity in later adulthood: Potentials and limits of development across the life span. In: Cicchetti D, Cohen D (eds.) *Developmental Psychopathology. Vol. 2: Risk, Disorder, and Adaptation*. Wiley, New York, pp. 801–47
Vaillant G E 1977 *Adaptation to Life* 1st edn. Little, Brown, Boston

L. L. Carstensen

Personality Development in Childhood

To understand personality development in childhood, we first define personality, and how it is structured. Then we discuss how the definition is typically translated into research, noting some curious omissions and quirks in the literature. Next, we consider the raw materials of personality and how they are changed over time. Once this is completed, we discuss outcomes of research on personality development in children.

1. Personality Development as the Organization of Individual Differences

First, perhaps the most basic question is what is personality? Common personality characteristics can be shared by groups of people—say, women, Russians, three-year-old boys—but usually we think of personality in terms of defining aspects of individual persons. In modern theory, personality is concerned with the organization of each person's unique configuration of individual differences. When personality is defined this way, rather than one difference at a time (e.g., introversion), it is possible to recognize that over time conflict within a person may occur due to competing and sometimes incompatible needs (see *Personality Structure*). The self-organization of the diverse elements that live under a common skin is motivated by the need to resolve intrapersonal conflicts and to adapt, or adjust, to the world beyond the individual. For example, we would expect an intelligent, introverted child to make compromises in the service of social adjustment different from those made by a similar child who is intelligent and extraverted. Personality is concerned with the accommodations and compromises each individual must make among the competing demands of these differences. The adjustment processes occur over time, and involve genetics, maturation, and learning. In this light, personality development in childhood can be seen as one segment in a life-long process that leads to the formation of a stable adult personality structure (see *Personality Development in Adulthood*). Common but untested assumptions are that processes of personality development are more dynamic in childhood than in adulthood, and that personality structure is more fluid and open to change in the former than in the latter. Personality development does not necessarily stop after childhood, but that period is regarded as a time of special sensitivity for the formation of the broad structure.

Second, we need to recognize how the definition of personality is typically translated into research. There are historical traditions and precedents in the personality development literature that influence the way professionals engage in research in the area. For example, the personality literature does not usually regard individual differences in ability in general, and intelligence in particular, as part of its domain, despite the illustration given in the preceding paragraph and the obvious importance of intelligence to adaptation. Similarly, differences in attachment are not well integrated into the rest of the personality development literature, despite the importance of attachment processes for adaptation. For another (but related) example, readers will sometimes encounter a distinction between 'personality development' and 'emotional development.' Some early writers used the term emotional development to focus on attachment and the psychodynamic approach to personality devel-

opment, and to separate it from other kinds of personality development. Detailed historical analyses make some of these oddities understandable, but such analyses are beyond the scope of this review (see *Infancy and Childhood: Emotional Development*). The point here is that definitions must be translated into measures and procedures, and historical precedents can lead to potential biases in the translation.

Third, developmental processes for each individual occur within a specific historical time, kinship system, social learning environment, and cultural context. Some aspects of personality are tied directly to such contextual factors (e.g., parental divorce), and leave residues. These life experience-based residues represent a legitimate aspect of personality. In the past, residues of experience relevant to personality have been conceptualized mainly in terms of 'learning,' or 'socialization' (see *Social Learning, Cognition, and Personality Development*). More recently, another approach has emerged. Residues of experience relevant to personality can be conceptualized in cognitive and information-processing terms. For example, cultures differ in the ways they teach children skills. In some cultures, mixed-age peer groups engage in competitions, with older peers teaching younger children. In those cultures, substantive knowledge acquired during peer interactions in certain domains is fused to specific beliefs about peer relations and social interaction that is missing in other cultures.

Fourth, given the potentially powerful influence of social learning and sociohistorical contexts, it is a major challenge for personality development theory to identify panhuman substantive dispositions. That is, in the past the majority of researchers have assumed that (a) personality development in all human children can be characterized by individual differences in a small set of enduring characteristics; and that (b) these characteristics induce the children to respond in consistent ways over time and across settings. These two assumptions may not be justified, and at the least they need to be evaluated empirically. Phrased more positively, personality development theory needs to integrate findings of cultural diversity and variation in children's personality with the assumptions about the nature and assumed roles for panhuman dispositions. To accomplish this task, researchers must select a small number of dispositions and enduring characteristics from the vast array of human differences that will prove to be important for theory or for life outcomes. Ideally, the ones chosen for attention in childhood will be broad in scope, related to competence in adulthood, and likely to become part of later personality structure. On an *a priori* basis, we might expect all cultures to be concerned with the development characteristics associated with success in the two major domains of 'love' (sex roles, reproduction, kin relations) and 'work' (reliability, rule compliance). It is possible, of course, that each culture will encourage the development of somewhat different configurations of characteristics within these two major domains, based on the ecological and social structures within which they live (see *Gender Differences in Personality and Social Behavior*).

2. *Four Perspectives on the Substance of Personality and Development*

To investigate the substantive nature of personality development, some further 'simplifying' assumptions are required. There have been four major perspectives on the identification of substantive characteristics in personality development (see Caspi 1998). The first major perspective is the most direct approach to childhood personality because it accepts it as a biological 'given.' It involves temperament, and focuses on childhood and infancy. Temperament, an important concept in developmental psychology, refers to early appearing, stable individual differences derived from the constitutional or biologically determined makeup of individuals (see *Temperament and Human Development*). It represents the presumed biological, emotional core around which later personality coalesces. For example, children show important individual differences in fearfulness or motor activity, and these may have heritable components. These are the kinds of differences that warrant attention developmentally because in theory they could constrain ontogenetic processes, and the kinds of experiences individuals accumulate as they move through life. Thomas and Chess (1977) stimulated modern interest in childhood temperament when they identified nine categories of differences based on clinical experiences with very young children. Examples are 'activity level' and 'persistence.' Second-order constructs were proposed to reduce the nine dimensions to three ('easy,' 'difficult,' and 'slow to warm up') based on factor analyses. Subsequent research showed that the original nine dimensions could not be recovered from item-level factor analyses. Instead, five to seven factors appeared. We will return to this perspective later.

Second is the social–cognitive process perspective of Mischel (1999). Researchers could identify individual differences in an enduring skill or tendency in children that could shape the ways children interact with their physical and social environments. These skills need not necessarily be conceptualized as traits (see *Personality and Social Behavior*). The cumulating effect of the skill on the child's outcomes could in theory have important long-range consequences (e.g., Kochanska et al. 1998). For example, the skills associated with delay of gratification are potentially relevant to both 'love' and 'work,' so they represent an especially good substantive characteristic to investigate as an aspect of personality development. Consistent with this perspective, Mischel et al., (1988) found that preschool children who are skilled in delay of gratification appear to interact systematically with the environment in

ways that allow them to achieve tasks more efficiently than their peers. Over time and settings, these skilled children accumulate accomplishments that allow them as adolescents to have higher social competencies and educational levels, and eventually as adults to achieve higher socioeconomic standing than other adults.

Third is the 'developmental targets' perspective (Caspi 1998), which is perhaps a meta-perspective because it can include both dimensional and typological theories under its umbrella. The goal here is to identify developmental antecedents of dispositions that are parts of stable adult personality structure. For example, extraversion and neuroticism are two major dimensions in adult personality structure, and represent developmental targets. It is important to know how these dimensions are formed developmentally. High levels of fearfulness or motor activity in childhood may be antecedents of adult neuroticism or extraversion, respectively. Once an empirical relation is demonstrated, then theoretical process mechanisms could be offered to explain how motor activity may be transformed or channeled into the targeted difference (e.g., extraversion). In effect, adult personality structure provides the target end states, and research is directed to finding its precursors. This approach requires knowledge of major, reliably appearing, and enduring adult personality structures (e.g., extraversion, neuroticism), but also empirically established antecedents that are not associated uniquely with configurations of sociohistorical events or cultural practices. (Of course, researchers (e.g., Elder 1999, McCrae et al. 1999) are interested in these sociohistorical and cultural patterns as a special class of personality phenomena.) Until recently, there was no consensus on the nature of adult personality structure, much less antecedents, that could meet these criteria, so personality development based on developmental targets was inhibited.

The fourth perspective could be regarded as a specialized typological version of the case outlined in the preceding paragraph. It is the person-centered, categorical approach of Block (1971). Unlike the researchers who assume personality is structured in terms of continuous underlying dimensions, the typological approach assumes that persons are better characterized as belonging in discrete categories depending on the configuration of attributes. Based on extensive empirical data on personality descriptions of persons as they moved from early adolescence to adulthood, Block had trained clinical judges sort descriptions into piles (Q-sorts), using a continuum from least characteristic to most characteristic. This Q-sort technique is person centered because each attribute is compared with other attributes in the same person. Statistical analyses can then be used to identify clusters of persons with similar Q-sort configurations. At least three adult types were found, and can be described as (a) ego-resilients, who are well functioning in all major domains; (b) overcontrollers, who were shy and lacked many interpersonal skills; and (c) undercontrollers, who were hostile, disagreeable, and showed little concern for others. These types could be regarded as developmental targets for developmental research. Longitudinal research with children (Hart et al. 1997) showed that three personality types could be replicated, suggesting avenues for theories of personality development. The largest category (ego-resilients) represents persons who are well adjusted; the other two smaller categories represent persons whose adjustment suffers from either overcontrol or undercontrol. At the least, such outcomes suggests that personality development should focus substantively on resilience and control. What is less clear is whether these types represent good developmental targets, in the sense of discrete natural classes, or are unique configurations of persons along some other underlying dimensions. It also is not clear whether these types represent the personality development of females as well as they do males (Caspi 1998).

3. *Personality Development and the 'Big Five'*

Let us move our discussion from perspectives to further research. In a series of pioneering papers in personality development, Digman and his colleagues in Hawaii began investigating the structure of child personality in teachers' ratings. Digman selected his variables to cast a wide net, and in some studies included assessments of intelligence, peer ratings, and various aptitude and personality batteries. He appeared to have no overriding theoretical agenda. At the time much of his data were collected, Digman (1994) suggested that he was an advocate for a model of child personality that included no less than ten factors. When his data were analyzed, however, a five-factor solution was reliably found (e.g., Digman and Inouye 1986). Digman recognized that the pattern he found in the child personality data was consistent with a similar pattern found in adults (e.g., Digman 1990). Perhaps the developmental targets approach could be useful after all. The time was ripe for a union between personality development researchers and researchers working with adult personality structure.

The number five seems to have a certain magical quality for personality researchers. The five-factor model, or the more descriptively accurate label 'Big Five' (BF) approach, currently occupies center stage in adult personality research (see *Personality Theories*). When people are given enough time and freedom to express their evaluations of others, five broad dimensions reappear reliably. These five dimensions are abstract 'supertraits' that summarize lower level patterns of individual differences. A useful mnemonic for remembering the dimensions of the BF is the acronym OCEAN: openness, conscientiousness, extraversion, agreeableness, and neuroticism. (The

memory aid does not convey the relative size or importance of these dimensions: extraversion and agreeableness are probably larger factors than the others.)

Two leading temperament researchers, Halverson and Kohnstamm recognized the need to bring child personality development researchers together with leading researchers working on the structure of adult personality, and they organized a conference. Papers presented at this meeting, and a few others, were published in a subsequent milestone volume entitled *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. The chapters in Halverson et al. (1994) suggested possible links among temperament differences in infants and toddlers and later personality structure in children and adults. Contributors to the Halverson et al. (1994) volume showed how aspects of temperament might be linked to BF personality structure. For the personality development researchers, temperament was the focal concept because in developmental psychology, it refers to early-appearing, stable individual differences derived from the constitutional or biologically determined makeup of individuals.

Prior to the Halverson et al. (1994) syntheses, the temperament researchers and the BF researchers were like two separate, isolated tribes. Temperament research was conducted largely as a kind of developmental psychology. Temperament was conceptualized as biobehavioral individual differences in young children, and not as personality structure, *per se* (see *Personality Development and Temperament*). Because young children cannot provide verbal self-reports, temperament researchers were usually forced to rely on objective measures and ratings by expert observers. In contrast, BF research was conducted largely as personality psychology, conceptualized as the structure of individual differences in adults. Because adults can provide verbal self-reports, BF researchers rely much more on self-reports than on objective measures or expert ratings for their data. Some BF researchers seemed mildly interested in the temperamental antecedents of the structure, but others most emphatically were not (see Hofstee 1991, p. 185). Many of the contributors to the Halverson et al. (1994) volume were internationally recognized authorities on temperament, and their suggestions about links between early-appearing, emotional dispositions and later BF personality structure gave credibility to the general line of thought.

One idea reinforced at this conference was that some aspects of temperament provide the developmental substrates for later, major dimensions of personality. One route is to link temperamental characteristics in infants and young children to each of the major BF dimensions. In theory, temperamental differences in activity at infancy could be linked positively to extraversion, and negatively to conscientiousness (e.g., Caspi 1998). Another route is to suggest ways that temperaments lay a functional foundation for later personality dimensions. For example, infants show temperament-related differences in reactions to frustration. Some infants become so upset when frustrated that they appear to be unable to respond adaptively to the frustrating event, whereas other infants seem be modulate their emotional reactions and deploy attention in adaptive ways. Ahadi and Rothbart (1994) suggested that a kind of superordinate temperament called 'effortful control' (EC) might regulate reactions to frustration. Furthermore, processes underlying EC might lay a functional foundation for both conscientiousness and agreeableness. This provocative idea suggests that (a) temperament-based processes like EC are related to perceptual processes of attention; (b) seemingly different dimensions of adult personality may be related through their common functional connection to the control of frustration; and (c) two of the major dimensions in the BF may have a common developmental substrate, and may differentiate over time into functionally separate self-regulatory systems. The implications for the socialization of emotions, achievement, and social skills are great (e.g., Kochanska et al. 1998).

In 1998, Kohnstamm et al. delivered some hard evidence in *Parental Descriptions of Child Personality: Developmental Antecedents of the Big Five?* (Kohnstamm et al. 1998). They offered a seven-culture study of parents' description of their children's personality. The seven countries involved were Belgium, China, Germany, Greece, The Netherlands, USA, and Poland. The book's title expresses the continuity beginning with early temperament research, on to Digman's pioneering work, through the Halverson et al. (1994) meeting of the minds, to the massive cross-cultural data set presented in this book. Rather than relying on top-down, theorist-imposed descriptors in existing temperament or personality measures, contributors to the Kohnstamm et al. volume asked parents to use their own words to characterize their children. The explicit assumption was that people in all seven countries frequently mention those characteristics that they think are most important or basic. In all seven countries, parents described children between the ages of two and 13 years. The descriptions were collected during face-to-face interviews by teams native to each country's language and culture. Using elaborate coding manuals, descriptions were categorized into units, which were inspired by the BF, with subcategories or facets derived mostly inductively. In addition, responses were coded as positive or negative. For example, in coding the extraversion descriptions, Kohnstamm et al. used 1A (sociability), 1B (dominance, leadership, assertiveness) and 1C (activity, pace, tempo). In this system the positive–negative poles are added so that 1A+ would include 'enthusiastic,' whereas 1A− would include 'tendency to shut self off.' After training, coders' agreement over the main categories was between 80 and 90 percent.

In theory, most cultures have a stake in supporting positive relations among people and controlling conflicts. It is reasonable then to expect persons in all cultures to recognize differences associated with harmonious relations, and to socialize children accordingly (see *Self-regulation in Childhood*). It is telling, perhaps, that the B5 dimensions that focus on interpersonal relations, extraversion and agreeableness, together account for approximately half of all parental descriptions of children. None of the other three dimensions rival either of the two interpersonal dimensions in terms of proportions of descriptions. Extraversion is one of the core concepts of trait psychology, and some form of it is measured in all widely used personality measures. In every country except Greece, extraversion was the single most frequently mentioned category at 28 percent of the descriptors (In Greece, agreeableness and extraversion were equal in importance.)

As a distinct aspect of personality, agreeableness is less well understood than extraversion. In part this may be due to an undifferentiated understanding of social behavior, which includes both social impact (extraversion) and motives to maintain positive relations with others (agreeableness). Eysenck (1994) emphasized the former aspect of social behavior in his trait model. As a consequence, personality researchers interested in the dispositional foundations of social behavior focused on extraversion at the expense of agreeableness. If agreeableness is related to 'love,' and to motives to maintain positive relations, then it is not difficult to see potential developmental continuities, particularly links to interpersonal conflict, and to childhood problems like difficult or disruptive behavior and antisocial tendencies. Data from the Kohnstamm et al. book show that in parental description, agreeableness is a large dimension, accounting for from 17 percent (China) to 26 percent (Greece) of all descriptions.

Because the Kohnstamm et al. work is massive, it is possible to give only a flavor of the findings for some of the other 'lesser' personality dimensions. Of the remaining dimensions, openness seems to be the largest category in free descriptions, but with major cultural differences. Data from the Kohnstamm et al. book show that parents in the USA make use of openness descriptors almost twice as often as do parents in all other countries except Germany. Further data focus on emotional stability (positive pole) and neuroticism (negative pole). Like extraversion, neuroticism is easily linked conceptually to early appearing temperaments (e.g., 'adaptability,' 'quality of mood'). Given the long history of research and theory on the dimension, and the applied importance of this dimension for mental health and child guidance clinics, many professional psychologists will be surprised to learn that across all seven countries, fewer than 10 percent of all parental descriptors fell into this category. One conclusion is that on a worldwide basis most parents are not concerned about negative emotionality, anxiety, or self-confidence in their children. These qualities may be taken for granted, which is not the same as being unimportant.

Previously, we suggested that all cultures should be concerned with the development of characteristics associated with success in the domain of 'work,' and this concern would be related to evaluations on conscientiousness. Nevertheless, if there is an enigmatic domain of personality development (at least as it appears in parental descriptions of children), it is conscientiousness. As a total category, proportions of parental descriptions were low but highly variable by culture, ranging from 7 percent (USA) to an outlier of 19 percent (China). Chinese parents used almost three times as many descriptors of diligence, and twice as many descriptors of carefulness, as did US parents. Closer inspection shows that Chinese parents' descriptors were drawn from the low end of the dimension (e.g., not diligent), after the child was older than three years of age. What is enigmatic, perhaps, is the low proportion of conscientiousness descriptions relative to the increasing importance of the dimension with age of child in five of the seven countries studied. The outcomes for conscientiousness may reflect cultural differences in the value of education, and in the values of modesty and humility. Chinese parents seem to stress achievement motivation more than parents in the USA, and as a result may have been more critical. Describing one's own children in very positive terms may be especially unseemly for Chinese parents.

4. *Future Directions*

In the broad-brush overview presented here, it is possible to see the recent advances in theory and research on personality development in childhood. The advances are due in part to (a) efforts to find points of convergence, and to synthesize different perspectives on personality development; (b) the discovery of certain empirical regularities, both within cultures and across them; and (c) more sophisticated research techniques that permit the consideration of several variables at a time. In particular, we see increasing awareness that both culture-based, social learning environments and dispositional characteristics contribute to the personality development process. Important questions and issues remain, however, for the future.

We began this entry with a modern definition of personality as a structural integration of individual differences. Personality development theory and research have only begun to deal with the issue of organizational structures. Theoretical accounts of the development of personality structure are rare, but rarer still as systematic empirical evaluations of such theoretical accounts. So far, the best available research has demonstrated empirical links of variables taken

one at a time, which is no small achievement. This is some distance, however, from the theoretical description of personality as a structured integration. Furthermore, it could be argued that much of the personality development literature has yet to move from the assessment of reliability to the assessment of validity. Still missing from the personality development literature are systematic demonstrations of predictive validity to theoretically important external criteria. Mean differences in frequencies in personality descriptions across cultural groups or sex do not guarantee corresponding differences in overt behavior, or the prediction of overt behavior by personality. For example, Greek parents may use more agreeableness words to describe their children than do US parents, but do agreeableness differences predict choice of conflict strategies in a more differentiated way for Greek children than for US children? In a related vein, linking infant and toddler characteristics to later childhood personality is also promising, but it is not clear how (or even whether) early characteristics are foundational developmentally for later personality structure (but see Kochanska et al. 1998, Rothbart and Bates 1998). That is, early appearing temperaments and dispositions, and life experiences like parental divorce, may precede later personality, and may even predict later personality patterns. Such connections may not be causal or developmental in nature (e.g., Lewis et al. 2000). Is childhood in fact an especially formative period for personality development, or just another way station for the unfolding of a maturation-based personality system? Alternatively, is childhood just another social ecology to which each child must adapt, but with no long-term consequences for personality development? These questions go to the heart of personality development because they address the question of the origins and maintenance of continuity in persons across time. On an optimistic note, these limitations are less obstacles than incentives for future work on personality development in childhood.

See also: Gender Differences in Personality and Social Behavior; Infancy and Childhood: Emotional Development; Personality and Conceptions of the Self; Personality and Social Behavior; Personality Assessment; Personality Development and Temperament; Personality Development in Adulthood; Personality Psychology; Personality Psychology: Methods; Personality Structure; Personality Theories; Self-regulation in Childhood; Social Learning, Cognition, and Personality Development; Temperament and Human Development

Bibliography

Ahadi S A, Rothbart M K 1994 Temperament, development, and the Big Five. In: Halverson C F Jr, Kohnstamm G A, Martin R P (eds.) *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 189–207

Block J 1971 *Lives Through Time*. Bancroft, Berkeley, CA

Caspi A 1998 Personality development across the life course. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, Vol. 3, pp. 311–88

Digman J M 1989 Five robust trait dimensions: Development, stability, and utility. *Journal of Personality*. **57**: 195–214

Digman J M 1990 Personality structure: Emergence of the five-factor model. *Annual Review of Psychology* **41**: 417–40

Digman J M 1994 Historical antecedents of the five-factor model. In: Costa P T, Widiger T A et al. (eds.) *Personality Disorders and the Five-factor Model of Personality*. American Psychological Association, Washington, DC, pp. 13–18

Elder G H Jr 1999 *Children of the Great Depression: Social Change in Life Experience*. Westview Press, Boulder, CO

Eysenck H J 1994 The Big Five or giant three: Criteria for a paradigm. In: Halverson C F Jr, Kohnstamm G A, Martin R P (eds.) *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 37–51

Halverson C F Jr, Kohnstamm G A, Martin R P 1994 *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. Lawrence Erlbaum Associates, Hillsdale, NJ

Hart D, Hofman V, Edelstein W, Keller M 1997 The relation of childhood personality types to adolescent behavior and development: A longitudinal study of Icelandic children. *Developmental Psychology* **33**: 195–205

Hofstee W K B 1991 The concept of personality and temperament. In: Strelau J, Angleitner A (eds.) *Explorations in Temperament: International Perspectives on Theory and Measurement*. Plenum, London, pp. 177–88

Kochanska G, Tjebkes T L, Forman D R 1998 Children's emerging regulation of conduct: Restraint, compliance, and internalization from infancy to the second year. *Child Development* **69**: 1378–89

Kohnstamm G A, Halverson C F Jr, Mervielde I, Havill V L (eds.) 1998 *Parental Descriptions of Child Personality: Developmental Antecedents of the Big Five?* Erlbaum, Mahwah, NJ

Lewis M, Feiring C, Rosenthal S 2000 Attachment Over Time. *Child Development* **71**: 707–20

McCrae R R, Costa P T Jr, de Lima M P, Simoes A, Ostendorf F, Angleitner A, Marusic I, Bratko D, Caprara G V, Barbaranelli C, Chae J-H, Piedmont R L 1999 Age differences in personality across the adult life span: Parallels in five cultures. *Developmental Psychology* **35**: 466–77

Mischel W 1999 Personality coherence and dispositions in a cognitive–affective personality (CAPS) approach. In: Cervone D, Shoda Y (eds.) *The Coherence of Personality: Social–cognitive Bases of Consistency, Variability, and Organization*. The Guilford Press, New York, pp. 37–60

Mischel W, Shoda Y, Peake P K 1988 The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of Personality and Social Psychology* **54**: 687–99

Rothbart M K, Bates J 1998 Temperament. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, Vol. 3, pp. 105–76

Thomas A, Chess S 1977 *Temperament and Development*. Brunner/Mazel, New York

W. Graziano

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Personality Disorders

1. General Considerations

Until the 1980s, both the scientific and the therapeutic aspect of personality disorders constituted a rather underdeveloped field despite their great significance for clinic and practice. Since the introduction of separate Axis II personality registration used in DSM-III (APA 1980) and the subsequent editions, DSM-III-R and DSM-IV (APA 1987, 1994), research and publication worldwide have been considerably stepped up, leading to the inception of professional journals and scientific societies devoted solely to the field of personality disorders. As of the beginning of the twenty-first century there is a veritable deluge of theoretical concepts and therapeutic methods.

More than in any other area of psychiatry, one still sees many inaccurate terms and concepts, and hard data on nosology, etiology, and pathogenesis of the various personality disorders remain rare. To be sure, there is a wealth of isolated pieces of etiological and pathogenetic knowledge on factors influencing the formation and progress of accentuated personality traits, but a comprehensive theoretical or empirical model remains as elusive as ever. We can therefore only address certain points and make some qualified references, as the exceedingly complex nature of the subject still does not allow any more definite statements.

'Personality disorder' is a value-free term which encompasses all those deviations from a normal personality development which require treatment. There are, necessarily, fluid borders to mental health on the one hand and mental disorder on the other.

Concepts with meaning similar or identical to personality disorder are psychopathy, abnormal personality, psychopathic development, psychopathic or dissocial personality, sociopathy, etc. In Sect. 3, the author will give an overview of the historical roots of these concepts while also addressing recent diagnostic and therapeutic developments.

2. Terms and Definitions

The most comprehensive term in this field must surely be 'personality.' One of the most complex terms in our language, it can be defined in a variety of ways.

Popularly, personality is often used as an ethical-pedagogical value judgment to mean character or temper, as when referring to someone as a strong personality or a person of character. Moral judgments are likewise made, as when we say that all human beings are persons, but we do not mean that everyone is a personality, which is characterized by a conscious and sustained exercise of thought and will.

The current psychological and psychiatric definition of personality is: the sum of all mental and behavioural characteristics which make each of us a unique individual.

On the other hand, a personality disorder occurs if, due to degree and/or a particular combination of psychopathologically relevant features in any of these areas, there is considerable suffering and/or lasting impairment of social adaptiveness (Saß 1987). According to DSM-IV, only when personality traits are inflexible and maladaptive and cause either significant functional impairment or subjective distress do they constitute personality disorders (Table 1).

3. A History of the Concepts of Abnormal Personality

The concept of psychopathy arises from a confluence of views entertained by the French, the German, and the Anglo-American psychiatric traditions. Sociocultural factors caused these conceptions of psychopathy to develop more or less independently well into the twentieth century. The following section deals with all three traditions.

Table 1
General diagnostic criteria for personality disorder

A. A lasting pattern of inner experiences and behavior which deviates markedly from the expectations of the culture the individual lives in. This pattern is manifested in two or more of the following areas:

1. cognition (i.e., ways of perceiving and interpreting oneself, others, and events)
2. emotiveness (i.e., range, intensity, stability, and appropriateness of emotional response)
3. interpersonal functioning
4. impulse control

B. The pattern is inflexible and pervasive across a broad range of personal and social situations.
C. The pattern leads to clinically significant distress or impairment in social, occupational, or other important areas.
D. The pattern is stable and longlasting, and its onset can be traced back to adolescence or early adulthood.
E. The pattern is not better accounted for as a manifestation or consequence of another mental disorder.
F. The pattern is not due to direct physiological effects of a substance (e.g., drug or medication) or a general medical condition (e.g., head trauma).

3.1 French Concepts

Pinel's (1809) concept of *manie sans délire* includes instances of 'deranged personality' and can be looked upon as the beginning of the scientific study of personality disorder as a nosological entity. In the eighteenth century, all mental diseases were regarded as fundamental disturbances of the intellect. Pinel was one of the first to point out that in some disorders, the emotions were primarily involved. Nevertheless, the early nineteenth-century definitions of madness remained centered mostly around the intellect. Indeed, to this day, psychiatric phenomenology neglects the disorders of affect.

Esquirol (1838) developed the concept of monomania, a diagnostic category based on the partial, primary–and independent–involvement of the intellectual, emotional, and/or volitional functions of the mind, including states where a single behavioral disturbance became the sole diagnostic criterion for the condition (e.g., pyromania or kleptomania). Not surprisingly, monomania was one of the sources of the concept of psychopathy. Since its inception, the concept of monomania has been criticized on clinical and medicolegal grounds both in France and abroad.

Morel's (1857) idea of degeneration was a pseudobiological account strongly tinged with the idea of original sin. He proposed that: (a) degenerative alterations are pathological deviations from the normal; (b) mental illness is mostly hereditary; and (c) degeneration is both quantitative and qualitative, i.e., new disorders may appear. According to Morel's model, all types of mental and neurological disorders can be traced back to one common hereditary origin. His idea of progressive and polymorphic degeneration was accepted as an explanation of mental illness.

3.2 Anglo-American Concepts

Prichard (1835) defined moral insanity as 'madness consisting of a morbid perversion of the natural feelings, affections, inclinations, temper, habits, moral dispositions and natural impulses without any remarkable disorder or defect of the interest or the reasoning faculties, and particularly without any illusion or hallucinations.' During the early nineteenth century, the word 'moral' had many uses, in the psychological sense chiefly to denote the affective and the conative (rather than the purely intellectual) functions.

The British concept of psychopathy was shaped by D. K. Henderson (1939), who saw 'psychopathic states' as a condition of 'constitutional abnormality.' In contrast to other (especially German) psychiatrists, he thought of constitution as deriving from both heredity and environment. There were three psychopathic states, the predominantly aggressive, the inadequate, and the creative. The former two were characterized by antisocial traits and soon became part of the Anglo-American concept of personality disorder.

Rush (1812) was the first Anglo-American psychiatrist to study individuals whose disturbances were primarily characterized by irresponsibility and aggressiveness; who showed, as he put it, a 'moral alienation of the mind.' He believed that these reprehensible acts were the manifestations of mental disease, that they were unmotivated and driven by 'a kind of involuntary power.'

Partridge (1930) defined the concept of psychopathic personality as a persistent maladjustment which cannot be corrected by ordinary methods of education or by punishment. From Partridge on, the emphasis has been on description, and etiological speculation has taken the back seat. To this day, a view of psychopathy as sociopathy dominates Anglo-American psychiatry. A parallel concept, that of the antisocial personality disorder, appeared in DSM-III (APA 1980) and was kept in DSM-III-R (APA 1987) and DSM-IV (APA 1994).

3.3 German Concepts

In Germany, the concept of psychopathy embraces most forms of abnormal personality. The current German definition can be traced back to Koch, whose monograph *Psychopathische Minderwertigkeiten* [Psychopathic Inferiorities] (1891–93) was as decisive in shaping the concept of abnormal personality in Germany as Pinel's work was in France, Rush's in the USA, and Prichard's in Great Britain. German ideas of psychopathy also influenced French and Anglo-American views, especially after the 1930s, when many German-speaking psychiatrists and psychoanalysts emigrated to these countries.

Kraepelin's concept of psychopathy, influenced by the French theory of degeneration, formed the basis of Kurt Schneider's typology, and via the latter, of today's German doctrine of psychopathy. In successive editions of his textbook, Kraepelin elaborated on his concept of 'psychopathic states' as abnormal personality. He employed the term 'psychopathic personalities' in a predominantly social sense, including also innate delinquents, unstable liars, swindlers, and pseudo-querulous individuals.

During this period, there also appeared various 'systematic' typologies to draw 'psychopathic' categories from prototypic personality theories. Foremost amongst these is Kretschmer's *konstitutionstypologisches Modell* (1919). He suggested a specific correlation between body type and personality. Kretschmer (1921) and Ewald (1924) also introduced the concept of *Reaktionstypologien*to account for specific styles of dealing with experience. Systematic typologies lost influence after the publication of Schneider's monograph in 1923.

Kurt Schneider is known especially for his famous monograph *Die psychopathischen Persönlichkeiten* [The Psychopathic Personalities] (1923). He used a 'typological approach' to personality types and tried to avoid value judgments by not including antisocial forms of behavior. Schneider defines abnormal personalities as statistical deviations from an estimated average norm, although the concept of norm is poorly formulated in his work. In his model, eminently creative and intelligent individuals are also abnormal; hence, not all abnormal personalities could be said to have psychiatric implications. Schneider defined 'psychopathic personalities [as] those abnormal personalities that suffer from their abnormality or whose abnormality causes society to suffer.' It is very important to stress that Schneider did not consider psychopathy a form of mental illness, which by definition must be associated with a brain lesion or a recognized disease process. In this, he opposed Kretschmer and Bleuler, who believed that psychosis and psychopathy were just different degrees on the same spectrum of derangement. Schneider's doctrine influenced all subsequent typologies, and current classification systems include essential parts of his concept of psychopathy. The appendix to DSM-IV even includes the 'depressive type' to encourage further research on this subject.

4. *Epidemiology, Course of Illness, and Prognosis*

According to German and American studies, 3–10 percent of the general population meet the diagnostic criteria of a personality disorder. Compared to numbers in earlier manuals, these are rather high. However, simply meeting these criteria need not imply that the individuals in question are so dysfunctional and impaired as to require treatment.

Prevalence rates are much higher among unselected psychiatric patients. After the first classification systems appeared, clinical studies found surprisingly high frequencies for personality disorders—50–80 percent—while more recent studies reported prevalence rates of 40–60 percent on the average. Forensic-psychiatric random samples yielded prevalence rates of up to 80 percent. In a large-scale international WHO study (Loranger et al. 1994), 39.5 percent of 716 psychiatric patients examined showed at least one personality disorder according to ICD-10, individual prevalence rates falling between 15.2 percent (anxious personality disorder) and 1.8 percent (schizoid personality disorder).

Clinical experience has shown that increasing age and decreasing vitality tend to attenuate 'sharp' personality traits, especially those which seriously impair social functioning, such as inconstancy, antisocial behavior, and impulsiveness. Other traits can become sharper with advancing age, above all obstinacy and rigidity.

The prognosis depends on the particular type of personality disorder, on any eventual comorbidity, and on the degree of severity. Further prognostic factors are psychostructural maturity as well as the level of psychological and social functioning. Prognostically favorable characteristics are motivation, trust in others, flexibility, and insight into one's own role in difficulties with interpersonal contact. Cases are complicated by concomitant illness, especially addiction and affective disorders. Thus, the mortality rate for patients with personality disorders and substance abuse is three times that of patients with a single personality disorder (Bohus et al. 1999).

Generally speaking, the risk of suicide is three times higher for individuals with a personality disorder than for the general population, with borderline, narcissistic, and antisocial personality disorders showing the highest incidence. These groups also show the highest degree of psychosocial impairment, with deviant actions, decreased capacity for work, and deficient skills at establishing dependable interpersonal relations.

5. *Diagnostic Problems*

In view of the complexity of the concept of personality, it becomes clear that difficulties exist in distinguishing between a normal range of varying personality traits (as many as 1800 features have been identified which might aid in characterizing possible personality traits) and personality disorders which might be of psychiatric relevance. Towards the end of the twentieth century there has been a trend towards reducing the myriad personality traits to 3, 5, or 7 essential dimensions from which the individual variations can be derived (Table 2).

Interestingly, the factor structure of behavior and its abnormalities seems to be essentially the same in the general population and in clinical groups of behaviourally abnormal patients. The differences lie mainly in the degree and the particular combination of the various dimensions, not in a fundamental difference of personality traits.

Categorical classifications are based on somatic models where the various diseases can be clearly separated from each other; personality disorders, however, show fluid borders, both between the various types and towards normality. Dimensional personality models, developed especially for scientific research, measure degree—or severity—by means of trait factors and thus assign it a position corresponding to its extent along various dimensional axes. It must be kept in mind that there is no inherent qualitative difference in personality traits between the general population and clinical groups of patients with personality disorders, only a difference in degree or severity, or eventually in a particular combination of traits. By means of factor analyses, these personality models

Table 2
Factor models of dimensional personality description

3-factor personality models	5-factor personality models	7-factor personality model
Eysenck (1977) neuroticism extroversion psychoticism	Costa and McCrae (1990) extroversion neuroticism agreeability conscientiousness	Cloninger (1994) harm avoidance novelty seeking reward dependence persistence
Millon (1981) joy/pain self/other-centeredness activity/passivity	openness to new experiences	self-directedness cooperativeness self-transcendence
Widiger et al. (1994) free-floating anxiety/eccentric behavior assertiveness/dominance social integration	von Zerssen (1988) extroversion neuroticism aggressivity conscientiousness openness (devoutness)	
DSM-IV (1994) eccentricity dramatic affect fear		

Table 3
Classification of personality disorders

DSM-IV	ICD-10
paranoid	paranoid
schizoid	schizoid
schizotypical	—
antisocial	dissocial
borderline	impulsive
histrionic	histrionic
narcissistic	—
avoidant	anxious
obsessive-compulsive	obsessive-compulsive
not otherwise specified	not otherwise specified and others

reduce the myriad personality traits to a few essential personality dimensions, independent of culture, from which the individual variants may then be derived. However, given the clinical usefulness of the categorical approach, and hence the widespread hesitancy in abandoning it, recent efforts have tended more towards a synthesis of the categorical and dimensional elements.

The classification systems used today, which increasingly determine personality diagnosis, are the *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition* (DSM-IV), published by the American Psychiatric Association (APA 1994), and the *International Classification of Diseases, 10th Edition* (ICD-10), published by the World Health Organization (WHO 1992). Both are used for making diagnoses, thus forcing clinicians, practitioners, and researchers alike to opt for one of the two systems. While the two systems resemble each other in many aspects, there are others where they completely disagree. Table 3 gives an overview of the most important discrepancies (see *Differential Diagnosis in Psychiatry*).

6. *Therapy of Personality Disorders*

Following a thorough anamnesis, there are two main options: psychotherapy and therapy with psychopharmaceuticals. In detail:

6.1 *Thorough Initial Examination*

The general procedure in treating personality disorders follows the usual rules of psychiatry. It is important that a thorough initial examination be done, with an exact medical and psychiatric anamnesis. Due to the lacking self-dystonia, the patient's self-perception may be lower than in cases showing acute psychopathological symptoms. Hence, anamnesis by significant others can also be helpful in assessing social conflict when dealing with personality disorders.

6.2 *Psychotherapy of Personality Disorders*

In the psychotherapy of personality disorders, the choice of the therapeutic procedure will depend on the particular form of personality disorder, its severity, as well as any concomitant mental disorders. Currently accepted forms of therapy include: cognitive/

behavioral therapeutic and supportive techniques, depth psychology, disorder-oriented and interdisciplinary therapies with treatments for specific problem areas of a personality, pedagogic or psychogogic therapies, sociotherapy, and dialogue sessions for couples or families. An important element of every therapy is that disorder-related behavior and interaction abnormalities also be dealt with in the patient–therapist relation, so that a big step in the therapy will consist in opening up new self-awareness and possibilities of change to the patient via the therapeutic interaction itself.

Psychotherapy of personality disorders usually requires a long time, since it involves a gradual reshaping of long-lasting characteristics in the areas of experiencing, feeling, and social behavior. Short-term crisis intervention can also be helpful if the problematic personality traits bring situations to a head and cause social conflicts. Building up a regular, trusting relation is of great importance. The frequency usually chosen is one to two hours per week. Aside from discussing current life conflicts, exploring past issues is also important as it can outline the development of the personality traits which are at the root of the trouble.

Group therapy is well suited for many personalities (provided they are not seriously deranged) since observing other patients allows learning by example. The positive feedback or criticism from the other group members can reinforce desirable behavior and put undesirable behavior in perspective (see *Cognitive and Interpersonal Therapy: Psychiatric Aspects*).

6.3 Pharmacotherapy

The second important therapy for personality disorders, one which in recent years has been increasingly developed, consists in the use of psychopharmaceuticals. It rests on the assumption that personality disorders can also have a biological cause which may be either constitutional or due to functional anomalies which were acquired later.

The aim is to reduce the vulnerability to affective and cognitive dysfunctions neurochemically and to modify certain behavioral reaction patterns, the target symptoms or syndromes being:

(a) the characteristics of the personality disorder (e.g., cognitive deficits, impulsive behavior, strong mood swings);

(b) complications due to the personality disorder (e.g., suicidal tendencies, aggressivity towards others, deficient social skills); and

(c) the axis I disorders associated with these (e.g., depressive or anxiety syndromes, compulsion, eating disorders).

The pharmacological treatment of personality disorders is by no means an alternative to psychotherapy. Rather, it is used in support and preparation for psychotherapy, as well as for crisis intervention (especially with suicidal tendencies).

Due to the dangers of side effects and habituation or dependence associated with it, long-term pharmaceutical therapy of personality disorders must always be viewed critically. Likewise, the therapist must be aware of the danger that the patient may start believing in external control or attribute therapeutic success solely to the action of the psychopharmaceuticals.

Moreover, there are special counterindications: alcohol and medicine abuse, social unreliability, bad compliance (incoming to therapy or in taking the medication). Due to the danger of late dyskinesias following long-term use, neuroleptics should be employed judiciously and substances of low or medium potency be considered. With benzodiazepines there is a particularly high risk of addiction, which is why they are generally not indicated for the treatent of personality disorders. Medication can also be tricky with some patients, who may harbor unrealistical expectations or fear to lose control. Due to these difficulties, such individuals may have to be admitted as inpatients before psychopharmaceutical therapy can be initiated (see *Psychotherapy and Pharmacotherapy, Combined*).

7. The Most Important Personality Disorders According to DSM-IV

This section offers classification of personality disorders, followed by an overview of the most important forms of abnormal personalities in DSM-IV. Please consult the manual for a complete list of the diagnostic criteria for the following personality disorders. The differential diagnosis refers solely to the Axis-I disorders (because of limited space).

7.1 Paranoid Personality Disorder

According to DSM-IV, persons with a paranoid personality disorder show a pattern of distrust and suspicion, reading malevolence into the motives of others without sufficient basis for that supposition. They often doubt the loyalty of their friends, partners, or associates and are reluctant to confide in others for fear that the information might be used against them. Persons with a paranoid personality disorder quickly react to perceived insults with excessive force, counterattacks, and/or long-lasting enmity. Mistakes are usually blamed on others. Persons with this disorder usually have recurrent suspicions regarding the fidelity of their spouse or partner. These persons are easily insulted, emotionally rigid, and persevering while at the same time appearing humorless and restrictive in their expression of warm emotions.

7.2 Schizoid Personality Disorder

The chief characteristics of the schizoid personality disorder are an inability to develop close emotional

attachments to others, seclusiveness, and a reduced capacity for emotional experience and expression in interpersonal settings. Schizoid personalities neither desire nor enjoy close relationships, have little, if any, interest in sexual experiences, and do not have friends or confidants. Persons with a schizoid personality disorder are shy, show emotional coldness and detachment, and are seemingly indifferent to praise or criticism. They take pleasure in few activities and distinctly prefer solitary ones. They may function adequately at work, especially if their job requires little social contact.

7.3 Schizotypal Personality Disorder

The schizotypal personality disorder is characterized by strong deficits in establishing interpersonal contact. Persons with schizotypal personality disorder show a pronounced fear of social situations and actively avoid them. They do not have close friends or confidants. Their interpersonal deficits are marked by acute discomfort with close relationships, constricted affect, cognitive and perceptual distortions, and eccentricities of behavior, as in their choice of dress and movement. Persons with this disorder often develop magical thinking and a belief in the occult, and sometimes ideas of reference or paranoid ideation may influence their behavior. Regarding language, there are unclear, strange, or stereotyped expressions and incorrect use of words, though not to the point of associative loosening and incoherence.

Phenomenological, biological, genetic, and outcome data (e.g., the Danish adoption studies) show a relation between schizotypal personality disorder and schizophrenia. Therefore, it is often considered a schizophrenia-spectrum disorder.

7.4 Antisocial Personality Disorder

The main characteristic of the antisocial personality disorder is a permanent and deep-seated tendency to violate and abuse the rights of others occuring from the age 15 years on. These persons show little introspection and self-criticism, lack empathy, show coldness, egotism, an exaggerated sense of entitlement, a paradoxical idea of adaptation, and weak or faulty social norms. Their behavior is marked by impulsiveness, unreliability, weak commitment, and absence of guilt feelings. They are practically beyond therapy and prognosis is generally unfavorable.

7.5 Borderline Personality Disorder

The borderline personality disorder shows a pattern of interpersonal and affective instability which often leads to sudden, short-lived bursts of extreme mood swings. The affective outbursts usually take place under conditions which are perceived as threatening, such as real or imagined abandonment or rejection.

Especially characteristic for borderline personality disorder is an alternating lifelong pattern of impulsively inflicting harm on oneself or others, including self-inflicted wounds, bulimic binge-and-purge attacks, periods of excessive alcohol consumption, or fights. Repetitive suicide threats and attempts are common.

Unlike antisocial personalities, most patients try to restrain or suppress their impulses, although these attempts often fail. This leads to unpredictable swings between a tense holding back of affective impulses on the one hand and sudden outbursts on the other. Predominant emotions are dysphoria, anxiety, anger, and chronic feelings of emptiness.

Borderline personality disorder patients are furthermore characterized by a highly unstable image and perception of themselves, which can also include aspects of gender identity, deficient orientation and plans for the future, as well as indiscriminate choice of social groups or partners. In their unstable and intense interpersonal relationships they often alternate between extremes of idealization and devaluation (splitting). A last and important area are the transient, stress-related dissociative or (pseudo)psychotic symptoms or paranoid ideas (see *Borderline Personality Disorder*).

7.6 Histrionic Personality Disorder

The main characteristics of this personality disorder are a strong need to be at the center of attention and to gain recognition. Persons with histrionic personality disorder show a pattern of excessive emotionality with self-dramatization, theatricality, and coquetry that may be reflected in their style of speech, which is impressionistic and lacking in detail.

There is a tendency for affective instability and superficiality, displaying rapidly shifting emotions. Most patients are largely suggestible and unable to maintain a steadfast pursuit of goals or value orientation. They are thus inconstant, especially in relationships. Persons with histrionic personality disorder often interact with others in an inappropriate, sexually seductive or provocative manner, using their physical appearance to draw attention to themselves. Relationships are often considered to be more intimate than they actually are.

Also clinically relevant, though uncommon, are the sometimes drastic 'pseudohysterical' cases showing aggravation, conversion, dramaticism, and improper behavior.

Manic states may be accompanied by exaggerated expression of emotion and impressionistic behavior,

but can be distinguished from histrionic personality disorder by the temporary nature of these symptoms and, of course, the presence of other specific/typical symptoms of mania.

7.7 Narcissistic Personality Disorder

According to DSM-IV, individuals with a narcissistic personality disorder have an exaggerated sense of self-importance, are preoccupied with fantasies of grandiosity or superiority, and require excessive admiration. They believe that they are 'special' and can only be understood by other special people.

Patients with narcissistic personality disorder tend to exploit others involuntarily, thinking that their own qualities and abilities entitle them to special treatment. There is a lack of empathy which becomes obvious in the unwillingness to identify with the feelings and needs of others. Further symptoms are a basically fragile self-esteem and feelings of envy and distrust towards others. With a clearly increased self-awareness and egotism, social discomfort and fear of negative opinions predominate. A particular problem is a tendency for depressive crises and resolute suicidality following an imagined insult.

7.8 Avoidant Personality Disorder

The avoidant-insecure personality disorder is characterized by a pervasive pattern of low self-esteem and hypersensivity to negative evaluation. Despite their strong wish for affection, persons with avoidant-insecure personality disorder avoid social relations, being insecure, shy, tense, and anxious. Their feelings of inferiority and inadequacy in social contact lead to a severe restriction of their social skills and roles, causing them to be reluctant to take personal risks or to engage in any new activities because they may prove embarrassing. Patients with avoidant personality disorder often show restraint with intimate relationships out of a strong fear of shame.

7.9 Dependent Personality Disorder

The dependent personality disorder is characterized by an overpowering feeling of not being able to conduct one's own life. With a self-image of weakness and helplessness, patients will seek support from others in all situations, especially from their partners. Persons with a dependent personality disorder need others to assume responsibility in most major areas of their life. They have difficulties doing things on their own or making everyday decisions without advice from others, and fear that expressing disagreement might result in a loss of support or approval. This leads to submissive behavior, sometimes to the point of being self-effacing and obsequious. In a relationship, these patients experience a constant fear of loss and abandonment, and urgently seek out somebody for support and care when a close relationship ends. They further show a cognitive distortion known as catastrophizing, which is a fearful and exaggerated estimate of the worst possible consequences of the relation ending.

7.10 Obsessive-compulsive Personality Disorder

The main characteristics of the obsessive-compulsive personality disorder are conscientiousness, perfectionism, conformism, and devotedness to work, which can be overvalued to the point of adversely affecting professional productivity and interpersonal relationships. They dislike teamwork or delegating tasks unless the others conform exactly to their way of thinking or acting. These persons show a severity and rigidity both with themselves and with others which often interferes with social functioning. They are overconscientious, inflexible, and rule-minded about matters of morality or ethics.

Obsessive-compulsive personalities are unable to separate the important from the unimportant (e.g., they cannot bring themselves to throw away old, worthless objects even when these have no sentimental value) and they have an inability to make decisions. Where money is concerned, they are miserly both with themselves and with others.

Of great importance are the interactions between the obsessive-compulsive personality disorder and depression. On the one hand, obsessive-compulsive personality traits can intensify during depression or first become disturbing, as in the form of depressive insecurity or difficulties in making decisions; on the other hand, obsessive-compulsive behavior can lead to difficulties, and hence to reactive depression, where obsessive-compulsive personality traits and depressive symptoms are closely interwoven.

7.11 Personality Disorders not Otherwise Specified

This category applies to personality disorders which are not designated by a DSM-IV diagnosis, but which cause clinically significant distress or impairment. It may also be used when the features of more than one Axis II disorder are present but the full criteria for any one disorder are not met. It also includes the depressive and passive-aggressive (negativistic) personality disorder, which are actually found in the appendix to DSM-IV, being currently under research to determine whether they should be included in DSM-V or not.

8. Concluding Remarks

Concluding, it should be noted that that all typologically arranged classifications of personality disorders are merely conventions. By giving a good description of the reality of life, they may acquire a certain plausibility, self-evidence, and clinical usefulness. A nosological diagnosis and therapy of personality disorders can only be expected for special forms, and even then only as long as it can be shown that they were caused by illnesses.

Most personality disorders, however, lie at the fringe of mental health. Instead of pathological processes, one has to deal with difficult developments in life due to a strained constitution and biography. First and foremost, these people need psychotherapeutic and psychogogic support. Far more than in mental illness, the manifestations in personality disorders are in great part shaped by such phenomena as will, decision-making, self-control, temper, intentionality, bearing, and character. Even more than mental illness, personality disorders require the biological and the psychopathological planes to be complemented by an anthropological dimension.

See also: Borderline Personality Disorder; Obsessive–Compulsive Disorder; Personality and Health; Personality Assessment; Personality Psychology; Personality Theory and Psychopathology

Bibliography

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. (DSM-III). American Psychiatric Association, Washington, DC

American Psychiatric Association 1987 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. revised (DSM-III-R). American Psychiatric Association, Washington, DC

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. (DSM-IV). American Psychiatric Association, Washington DC

Beck A T, Freeman A, Pretzer J, David D D, Fleming B, Ottaviori R, Beck J, Simon K M, Padesky C, Meyer J, Trexter L 1993 *Kognitive Therapie bei Persönlichkeitsstörungen*. Psychologie Verlags Union, Weinheim, Germany

Bohus M, Stieglitz R D, Fiedler P, Berger M 1999 Persönlich-keitsstörungen. In: Berger M (ed.) *Psychiatrie und Psychotherapie*. Urban und Schwarzenberg, Munich, pp. 771–846

Cloninger C R, Pryzbeck T R, Surakic D M et al. 1994 *The Temperament and Character Inventory (TCI): A Guide to Its Development and Use*. Washington University, St Louis, MO

Costa P T, McCrae R 1990 Personality disorders and the five-factor model of personality. *Journal of Personality Disorders* **4**: 362–71

Eysenck S B, Eysenck H J 1977 The place of impulsiveness in a dimensional system of personality description. *British Journal of Social and Clinical Psychology* **16**(1): 57–68

Esquirol E 1838 *Die Geisteskrankheiten in Beziehung zu Medizin und Staatsarzneikunde*. Voss, Berlin

Ewald G 1924 *Temperament und Charakter*. Springer, Berlin

Frances A J, Hales R E (eds.) 1986 *Review of Psychiatry, Vol 5*. American Psychiatric Press, Washington, DC

Henderson D 1939 *Psychopathic States*. Norton, New York

Jang K L, Livesley W J, Vernon P A, Jackson D N 1996 Heritability of personality disorder traits: A twin study. *Acta Psychiatrica Scandinavica* **94**: 438–44

Koch J L A 1891–3 *Die psychopathischen Minderwertigkeiten*. Maier, Ravensburg, Germany

Kraepelin E 1903–4 *Psychiatrie. Ein Lehrbuch für Studierende und Ärzte*. Barth, Leipzig, Germany

Kretschmer E 1921 *Körperbau und Charakter*. Springer, Berlin

Linehan M 1993 *Cognitive-Behavioral Treatment of Borderline-Personality Disorder*. Guilford, New York

Loranger A W, Sartorius N, Andreoli A, Berger P, Buchheim P, Channabasauanna S M, Coid B, Dahl A, Diekstra R F, Ferguson B 1994 The international personality disorders examination. *Archives of General Psychiatry* **51**: 215–24

Millon T 1981 *Disorders of Personality; DSM-III, Axis II*. Wiley, New York

Morel B A 1857 *Traité des dégénérescences physiques, intellectuelles et morales de l'espèce humaine et des causes qui produisent ces variétés maladives*. Baillière, Paris

Partridge G E 1930 Current conceptions of psychopathic personality. *American Journal of Psychiatry* **10**: 53–99

Pinel P 1809 *Traité médico-philosophique sur l'aliénation mentale*. 2nd edn. Brosson, Paris

Prichard J C 1835 *A Treatise on Insanity and Other Disorders Affecting the Mind*. Sherwood, Gilbert & Piper, London

Rush B 1812 *Medical Inquiries and Observations Upon the Diseases of the Mind*. Kimber & Richardson, Philadelphia

Saß H 1987 *Psychopathie—Soziopathie—Dissozialität. Zur Differentia/typologie der Persönlichkeitsstörungen*. Springer, Berlin

Saß H 1988 Persönlichkeit und Persönlichkeitsstörungen. In: Janzarik W (ed.) *Persönlichkeit und Psychose*. Enke, Stuttgart, Germany

Saß H 2000 Persönlichkeitsstörungen. In: Helmchen H, Henn F, Lauter M, Sartorius N (eds.) *Psychiatrie der Gegenwart. Bd.6*. Springer, Heidelberg, Germany

Schneider K 1923 *Die psychopathischen Persönlichkeiten*. Deuticke, Vienna, Austria (1. Aufl. 1923, Thieme, Leipzig Germany 1950)

Spitzer R L, Williams J B 1985 *Structured Clinical Interview for DSM-III, Patient Version*. Biometrics Research Department, New York State Psychiatric Institute, New York

Widiger T A, Costa P T 1994 Personality and personality disorders. *Journal of Abnormal Psychology* **103**: 78–91

World Health Organization (WHO) 1992 *International Statistical Classification of Diseases and Related Health Problems, 10th revision* (ICD-10). World Health Organization, Geneva

H. Saß

Personality Psychology

Personality psychology is the scientific study of the whole person. The goal of personality inquiry is to provide a scientifically credible account of human individuality. Such an account considers aspects of human nature (how a person is like all other persons), individual differences between people (how a person is like some other persons), and individual uniqueness

(how a person is like no other person). A full account of human individuality, furthermore, must situate a person's life in the biological, social, cultural, and historical contexts that give that life its meaning. This article will present a brief history of the field of personality psychology, a review of recurrent issues and controversies that have preoccupied personality psychologists, and a consideration of current and projected future research and theorizing in the field.

1. History

The roots of personality psychology can be traced back to the ancient Greeks. A pupil of Aristotle, Theophrastus (c. 300 BC) composed character sketches such as 'the flatterer' and 'the penurious man,' each of which personified (often in a humorous way) a particular personality trait. The Greek physician Galen (AD 130–200) is often credited with proposing the first taxonomy of personality traits, distinguishing among the sanguine, choleric, phlegmatic, and melancholic temperaments. It was not until the late nineteenth and early twentieth centuries, however, that scientists developed reliable methods for assessing individual differences in traits. Building on the pioneering work of Francis Galton and Alfred Binet on mental testing and spurred by the mobilization of large military forces in World War I, psychologists began to invent self-report, multi-item tests to assess individual differences in personality functioning. At about the same time, Sigmund Freud, Carl Jung, Alfred Adler, and other psychoanalytic theorists offered a strong alternative perspective on human individuality, underscoring the unconscious motivational dynamics that structure human lives and make for neurosis and conflict in human affairs.

The psychometricians and the psychoanalysts represented two very different intellectual traditions for personality psychology, both rooted in the nineteenth century. In the spirit of natural science inquiry (*Naturwissenschaften*), the former emphasized the precise quantification of common traits, readily assessed through conscious self-reports and manifest in observable behavior across many different persons. By contrast, Freud and the psychoanalysts offered a more holistic, clinically informed, and intuitive approach, consistent with the *Geistenwissenschaften* (human sciences) movement in the nineteenth century, proposing that the person is a complex and uniquely patterned whole. In the first authoritative textbook for personality psychology, Gordon Allport (1937) proposed that the two intellectual traditions might be reconciled within a single science of persons. Laying out an ambitious agenda for personality inquiry, Henry Murray (1938) published a landmark volume that drew heavily on the psychoanalytic tradition while proposing new methods for measuring personality traits, needs, and other features of human individuality. The writings of Allport and Murray helped to establish personality psychology as an identifiable discipline in the social and behavioral sciences in the 1930s.

1.1 The Development of Personality Theories

By the mid-1950s, the field of personality psychology was dominated by a number of competing grand theories. Each theory offered a comprehensive portrayal of human individuality, specifying fundamental motivations, a structure for understanding individual differences, principles of personality development, and in some cases a corresponding clinical technique for personality change (e.g., a brand of psychotherapy). Beginning with Hall and Lindzey (1957), personality textbooks typically organized the field in terms of these grand theoretical rubrics, each of which offered its own agenda for personality research.

The largest number of theories came out of the psychoanalytic tradition. Freud's seminal theory, articulated mainly between the years 1895 and 1930, proposed that all human behavior is determined by unconscious forces over which the individual has little control, forces that typically stem from sexual and aggressive drives. Rooted in instinct and early childhood experience, unconscious wishes, urges, and drives seek expression in a social environment that is predicated on the suppression of human urgings. The conflicts between internal urges and external constraints make for anxiety, defense, and a panoply of personal adaptations that define human individuality. Variations on these themes were played out in a number of other psychoanalytic theories, beginning with Jung's and Adler's rival approaches. While Jung proposed that individuality stems from each person's exploration of an evolutionarily rooted collective unconscious, Adler emphasized conscious aspects of human functioning and the prime motives of power and social interest. Around the time of World War II, neo-Freudian theorists like Karen Horney and Eric Fromm sought to link individual development to cultural factors, while Erik Erikson broadened Freud's psychosexual focus to include psychosocial issues that confront the person across the human lifespan. Moving in a different direction, postwar object-relations theorists such as W.R.D.Fairbairn and psychoanalytic self-theorists such as Heinz Kohut redoubled efforts to explore the most primitive intrapsychic factors laid down in the first two years of life and played out most importantly in the mother–infant bond.

Outside the psychoanalytic tradition, grand personality theorists came in behaviorist, trait, and humanistic/phenomenological varieties. Seeking to operationalize Freudian concepts in terms amenable to midcentury behaviorism, Neal Miller and James Dollard developed an influential social learning theory

of personality, which was followed by more cognitively based theories offered by Julian Rotter and Albert Bandura. Employing the method of factor analysis of personality trait scales, Raymond B. Cattell and Hans Eysenck conceived of personality in terms of a finite set of objectively defined traits. Eysenck's three-trait scheme—highlighting introversion–extraversion, neuroticism, and psychoticism—proved especially influential. Finally, a wide range of grand theories were built around a humanistic vision of the person as a more-or-less rational being, endowed with rich conscious experience and motivated to actualize his or her own potential. Included here were Allport's own theory, the humanistic approaches offered by Carl Rogers and Abraham Maslow, George Kelly's personal construct theory, and existential personality theories offered by Rollo May, Victor Frankl, and others.

1.2 The Elaboration of Constructs

Postwar personality psychologists borrowed liberally from the grand theories to identify individual-difference variables—or 'constructs'—that could be reliably measured through self-report scales, ratings, behavioral observations, and other techniques, and then used in hypothesis-testing research. Popular constructs of the 1950s and 1960s included anxiety, achievement motivation, authoritarianism, extraversion, ego resiliency, field independence, and locus of control. Investigators designed programs of study to examine the antecedents, consequences, and correlates of these individual differences in personality. Around each construct, then, developed a nomological network of findings—a body of research that elaborated the construct and defined its role and significance in personality functioning. With increasing specialization and the emphasis on discrete constructs, the field of personality psychology began to turn away from the grand theories to focus on problems in the measurement of individual differences. Omnibus personality inventories, such as the Minnesota Multiphasic Personality Inventory (MMPI) and the California Psychological Inventory (CPI), measured many different variables at once. But many postwar researchers preferred to focus on one construct at a time, refining measures and articulating in great detail that construct's nomological net.

The field of personality psychology experienced a crisis in confidence in the late 1960s and 1970s. Some critics lamented the decay of interest in integrative theory and the obsessive focus on constructs over persons. Others, most notably Walter Mischel (1968), argued that personality constructs themselves were not good predictors of behavior. In Mischel's view, social situations account for vastly more variance in human behavior than do individual differences in personality traits. What was often called the 'trait versus situation debate' preoccupied the field in the 1970s and early 1980s. As the debate died down, proponents of both sides seemed to agree that personality theories and research need to explicitly embody interactionism—the idea that behavior is always a function of the interaction of the person (e.g., traits) and the environment (e.g., situations). In the 1980s and 1990s, furthermore, strong evidence was adduced for the efficacy of personality constructs. First, studies increasingly showed that traits do predict general trends in behavior over time and across different situations. Second, longitudinal studies showed that individual differences in traits are often remarkably stable over long periods of time. Third, evidence from behavior genetics, especially studies with twins, showed that most personality traits exhibit at least moderate heritability. And fourth, some consensus in the field began to emerge concerning the number and kind of basic personality traits. Based on extensive factor-analytic studies, the popular Big Five conception proposes that traits cluster into five different factors, which may be labeled extraversion, neuroticism, conscientiousness, agreeableness, and openness to experience.

2. Perennial Issues and Controversies

In his first textbook, Allport (1937) foresaw a number of issues that were destined to stimulate recurrent debate in the field of personality psychology. The one that most preoccupied Allport himself was the tension between nomothetic and idiographic approaches to personality inquiry. While nomothetic approaches seek to establish general laws of behavior that apply across persons, idiographic approaches, as embodied in the case study, focus on the unique or characteristic patterning of an individual person. The vast majority of published research in personality psychology is nomothetic, typically involving the testing of hypotheses about personality constructs and processes. But if the field itself is supposed to be concerned with human individuality, Allport argued, then some form of idiographic inquiry must be included. Skeptics have countered that the results of single-case studies cannot be generalized, and thus have little scientific value. But proponents of idiographic approaches maintain that case studies are often excellent arenas for hypothesis discovery, for applying general theories, and for illuminating complex personality organization. Along with Allport and Murray, Robert White (1952) championed the intensive study of individual lives. Recent years have witnessed a resurgence of interest in idiographic approaches and considerable optimism about integrating them with more conventional nomothetic methods.

A forerunner to the 'trait versus situation debate' of the 1970s was Allport's identification of the problem of generality versus specificity in behavior. To what extent is a person's behavior generally consistent across situations, as opposed to being specific to the

vagaries of particular situations themselves? Mischel (1968) argued that Allport and most other personality psychologists overplayed the generality idea, expecting their constructs to predict general trends in behavior across many different situations. In Mischel's (1968) view, the empirical data were much more supportive of a specificity position. Although trait constructs have regained their currency in recent years, many personality psychologists have retained a healthy skepticism about cross-situational generality, and some have proposed that some personality constructs themselves need to be defined in contingent, situational terms.

A third issue concerns measurement. The most popular personality measures have always been self-report questionnaires. But many critics have argued that such measures are unable to assess especially subtle, implicit, or unconscious aspects of human individuality. As an alternative, some have championed projective techniques, wherein the person responds freely to ambiguous cues (e.g., inkblots, story scenes). For example, David McClelland (1961) built a highly successful research program around the assessment of achievement motivation in imaginative stories told to picture cues (the Thematic Apperception Test, or TAT). Others, most notably Jack Block (1971), refined Q-sort rating procedures that bypassed self-report for the evaluations of expert judges. While a plethora of measurement techniques may be seen in the field today, the self-report questionnaire, nonetheless, remains the coin of the realm.

A fourth controversy is the often-observed disconnect between grand personality theories and construct-based personality research. While some argue that a good deal of personality research has been directly or indirectly inspired by the grand theories, others contend that the grand theories should be dismissed as historical artifacts. The controversy is especially acute with respect to psychoanalytic theories. Many Freudian ideas, for example, have proven resistant to empirical scrutiny (e.g., the Oedipus complex) or have been jettisoned as outdated or just plain wrong (e.g., the death instinct). By contrast, some ideas that have traditionally been associated with psychoanalytic approaches have become incorporated into mainstream psychological research. Of most importance in this regard is the now generally accepted notion that a good deal of human information processing occurs in an automatic, implicit, and nonconscious manner. Thus, while psychoanalytic theories have exerted a strong impact on Western thinking more generally, their current and future status in personality psychology appears ambiguous at best.

3. *Current and Future Trends*

The wide range of research and theorizing that currently characterizes the field of personality psychology may be organized into three different levels or arenas of human individuality: dispositional traits, characteristic adaptations, and integrative stories. The first domain considers individuality from the standpoint of general, cross-situational tendencies; the second examines more situationally and contextually grounded aspects of persons; and the third encompasses efforts to understand individuality in the more idiographic sense of how the person understands his or her life in toto.

3.1 *Dispositional Traits*

Dispositional traits are those relatively general and nonconditional aspects of human individuality that go by such names as 'extraversion,' 'friendliness,' 'depressiveness,' and 'conscientiousness.' Each trait is viewed as a bipolar, linear dimension upon which individuals can be said to differ. For example, research has shown that people scoring at the high end of the extraversion trait are viewed by others as especially sociable and fun-loving, typically attend more social events, are more sexually active, and report higher levels of positive affect in everyday life, compared to people scoring at the low (introverted) end of the trait (e.g., Eysenck 1967). Evidence for the predictive efficacy of trait scores is growing, with the strongest data coming for aspects of extraversion (sometimes reconceptualized as positive affectivity) and neuroticism (sometimes reconceptualized as negative affectivity).

Psychophysiological experiments have begun to document possible linkages between neural systems and individual differences in traits. For example, Jeffrey Gray and others have suggested that individual differences in extraversion/positive-affectivity may be linked to a behavioral activation system in the brain, responsible for motivating behavior aimed at achieving goals and obtaining positive emotional rewards. These kinds of advances, along with growing evidence for longitudinal consistency and heritability of broad traits, help to underscore the field's conviction that dispositional traits are fundamental to a scientific account of human individuality (see McCrae and Costa 1990, Wiggins 1996).

Increasingly, research on personality traits has become subsumed within the Big Five framework. While most of this research focuses on individual differences among well-functioning adults, recent efforts have also been made to reconfigure psychopathology in terms of (extreme) scores on personality traits. Individual differences in infant and childhood temperament, assumed to be strongly driven by genetic differences, have also been conceptually linked to the five-factor taxonomy. And some theorists have speculated about how individual differences in Big Five traits reflect varying strategies for solving fundamental social tasks that humans have perennially faced over the course of evolution. The integration of trait psychology with evolutionary theory would appear to be a promising direction for the future.

3.2 Characteristic Adaptations

Despite substantial progress in trait research in recent years, there would appear to be considerably more to personality than traits. While traits are useful for predicting and understanding consistent, cross-situational differences in people, they are less effective in the explanation of more particular, contextually anchored behavior. Accordingly, personality psychologists have proposed a host of motivational, developmental, and strategic constructs that are contextualized in time, place, or social role. These include particular motives and goals (e.g., the power motive), life tasks (e.g., the intimacy life task), relational styles (e.g., secure attachment style), defense mechanisms (e.g., projection), coping strategies (e.g., emotion-focused coping), values (e.g., dogmatism), developmental issues (e.g., generativity), self-schemas (e.g., possible selves, self-with-other representations), and the like, all of which may be grouped under the generic rubric of characteristic adaptations. If traits sketch an outline of human individuality, characteristic adaptations begin to fill in some of the details. More contingent and delimited than dispositional traits, characteristic adaptations speak to what a person wants or fears, often during particular periods in life or within particular domains of action, and what life methods the person may use to get what is wanted or to avoid getting what is not wanted at a particular time, in particular places, and/or with respect to particular roles.

While trait approaches have their historical roots in the grand factor theories offered by Eysenck and Cattell, studies of characteristic adaptations connect more readily to theories emphasizing human cognition, motivation, and development. In that they address especially concrete issues that can be observed in everyday adaptation to stress and challenge, moreover, characteristic adaptations speak a language that is compatible with the concerns of many clinicians. They also appear to shed light on proximal personality processes by specifying what cognitions or emotional states may give rise to particular behaviors. Recent years have witnessed considerable progress in the study of a wide range of characteristic adaptations (Pervin and John 1999), but the field has yet to offer a comprehensive framework, akin to the Big Five, for organizing these far-flung constructs and research programs. In a general sense, though, most adaptations link directly or indirectly to the two great domains of social life—love and work—or what David Bakan termed 'communion' and 'agency.'

3.3 Integrative Stories

Since the time of Allport, some personality psychologists have maintained that while conventional personality constructs such as traits and characteristic adaptations are useful for predicting behavior and accounting for individual differences, they are unable to account for the integrative nature of personality functioning and the sense of a person's wholeness. Broader concepts such as self, ego, and character structure have often been invoked to refer to this integrative, holistic quality. For example, Jane Loevinger (1976) proposed that the ego is the master synthesizer of subjective experience, and she developed a rigorous program of research to operationalize stages of ego development through the coding of sentence-completion tests. More recently, a number of theorists and researchers have proposed that people synthesize their own experience into personal stories and that such internalized and evolving narratives of the self provide human lives with unity, purpose, and meaning. Silvan Tomkins, Dan McAdams (2001), and Hubert Hermans have all developed narrative theories of personality that attempt to account for human individuality in terms of the different kinds of life stories that people construct. This movement recalls Allport's emphasis on the idiographic, and the study of lives tradition championed by White and others. But it adds a more nuanced understanding of the psycholiterary aspects of social life, especially under conditions of cultural modernity and postmodernity, and it proposes agendas for nomothetic research whereby individual differences in life narrative form and content can be operationalized and subjected to hypothesis-testing research.

In sum, current research and theorizing in personality psychology suggest that personality itself may be viewed as a patterning of dispositional traits, characteristic adaptations, and integrative life stories, evolving over time and complexly situated in a cultural and historical milieu.

See also: Gender Differences in Personality and Social Behavior; Genetic Studies of Personality; Interactionism and Personality; Personality and Adaptive Behaviors; Personality and Conceptions of the Self; Personality and Crime; Personality and Health; Personality and Marriage; Personality and Risk Taking; Personality and Social Behavior; Personality Development and Temperament; Personality Development in Adulthood; Personality Development in Childhood; Personality Disorders; Personality Psychology; Personality Psychology: Methods; Personality Structure; Personality Theories

Bibliography

Allport G W 1937 *Personality: A Psychological Interpretation*. H. Holt, Rinehart & Winston, New York

Block J 1971 *Lives Through Time*. Bancroft Books, Berkeley, CA

Eysenck H J 1967 *The Biological Basis of Personality*. Thomas, Springfield, IL

FreudS [1916] 1961 Introductory lectures on psychoanalysis. In: J. Strachey (ed.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth, London, Vols. 15–16
Hall C S, Lindzey G 1957 *Theories of Personality*. Wiley, New York
Hogan R, Johnson J, Briggs S 1997 *Handbook of Personality Psychology*. Academic Press, San Diego, CA
Loevinger J 1976 *Ego Development: Conceptions and Theories*. Jossey-Bass, San Francisco
McAdams D P 2001 *The Person: An Integrated Introduction to Personality Psychology*, 3rd edn. Harcourt College Publishers, Fort Worth, TX
McClelland D C 1961 *The Achieving Society*. D. Van Nostrand, Princeton, NJ
McCrae R R, Costa Jr. P T 1990 *Personality and Adulthood*. Guilford Press, New York
Mischel W 1968 *Personality and Assessment*. Wiley, New York
Murray H A 1938 *Explorations in Personality: A Clinical and Experimental Study of Fifty Men of College Age*. Oxford University Press, New York
Pervin L A, John O P 1999 (eds.) *Handbook of Personality Theory and Research*, 2nd edn. Guilford Press, New York
White R W 1952 *Lives in Progress: A Study of the Natural Growth of Personality*. Drydren Press, New York
Wiggins J S 1996 (ed.) *The Five-factor Model in Personality: Theoretical Perspectives*. Guilford, New York

D. P. McAdams

Personality Psychology: Methods

All empirical research methods in psychology are concerned with the measurement of variation and covariation. Three methods for studying three types of variation and covariation can be identified. Experimental methods discern how behavior and experience vary across different environments. Developmental methods describe how behavior and experience vary over time. Finally, differential methods measure relatively enduring differences in characteristics of persons and the covariation among these differences. Although personality psychology occasionally employs experimental and developmental methods, its primary use of differential methods defines its distinctive position in psychology (Cronbach 1957). The specific methods used by personality psychologists can be further defined by (a) the type of individual differences under study, (b) the source of information about these differences, and (c) the purpose of assessing individuals' distinguishing attributes.

1. *Types of Differences Measured*

Gordon Allport's (1937) definitive etymological analysis of the term 'personality' finds that *persona*, the Latin root of 'personality,' originally referred to a mask worn by actors, and thus connoted a false or superficial appearance. Before long, however, *persona* also came to designate more substantial attributes that qualify a person to perform particular roles. Today, 'external' and 'internal' aspects of personality continue to be distinguished. Ordinary-language descriptions of personality often refer to external social impressions. Trait terms representing these impressions tend to be broad and evaluative (e.g., nice, charming, nasty, irritating), and sometimes metaphorical (cold, prickly, warm, slick). Social impressions are explained by observable behavioral traits. For example, to impress others as 'warm,' a person must consistently smile, make eye-contact, talk tenderly, express affection, and show kindness and sympathy. External (social and behavioral) aspects of personality are typically measured by observer judgment (see Sect. 2.2).

Behavioral consistencies, in turn, are explained by 'inner' traits that guide and motivate behavior. Inner traits that refer to persistent thoughts, expectations, attitudes, thinking styles, and mental abilities are called 'cognitive traits.' Inner traits involving recurring feelings and emotional dispositions are called 'emotional' or 'motivational traits.' Personality psychologists often rely on the self-report method (see Sect. 2.3) to assess inner traits. Cognitive styles and competencies (e.g., creativity), are typically assessed by performance tasks (see Sect. 2.4).

2. *Types of Individual Differences Data*

Differential methods can be organized according to the types of the data one can use to study individual differences. The acronym LOST conveniently summarizes the four major types of individual differences data: life events, observer judgments, self-reports, and tests, or L-data, O-data, S-data, and T-data (see Block 1977).

2.1 *L-data: Life Events*

'Life events' refer to relatively objective facts about people that are often a matter of public record. Some examples are birth date, number of years of education and degrees obtained, marital status and number of children, occupational record, church attendance, hospitalizations, membership in societies, criminal convictions, leadership positions, and property ownership. Although life event information is often actually obtained by self-report methods (Sect. 2.3) its factuality can be independently verified. Research indicates a very high correlation between self-reported and independently verified life events, even under highly evaluative conditions such as applying for employment.

In both applied (see Sects. 3.1 and 3.2) and basic (see Sect. 3.3) personality research, life events are sometimes used as criteria to be predicted by other types of personality data and sometimes as predictors themselves. When gathered by self-report, L-data is interpreted according to self-report methods (Sect. 2.3).

2.2 O-data: Observer Judgments

Observer judgments are a fundamental source of information about external traits for two reasons (Block 1978, Hofstee 1994). First, an observer's judgment of external traits is direct and noninferential. Second, idiosyncratic biases and errors of judgment tend to cancel out when judgments are averaged across multiple observers. This means that even inner traits may be more validly assessed by observer judgments than self-reports, despite an individual's privileged access to his or her inner thoughts and feelings (Hofstee 1994).

2.2.1 Retrospective, integrative judgments versus one-time, direct observation. Observer judgments are often made by acquaintances who use their past knowledge to summarize their perceptions of someone's personality. Concern about potential distortions and memory limitations involved in making retrospective judgments lead some psychologists to favor direct behavioral observation. Despite potential distortions, retrospective judgments are more likely than single encounters to produce the representative sampling necessary for accurately assessing relatively enduring characteristics of individuals (Kenrick and Funder 1988).

2.2.2 Normative and ipsative frames of reference for observer judgments. For observer judgments to indicate what is distinctive about someone's personality, the judgments must be compared to some reference point. Take, for example, a study in which observers record from behind one-way glass instances of different types of behavior in a nursery school classroom over a period of one year. At the end of the year, a particular behavior count by itself—say 42 acts of aggression—is insufficient for describing a child as relatively aggressive or nonaggressive. A normative frame of reference would compare the number of aggressive acts for that child to the average number of aggressive acts across all children. An ipsative frame of reference compares the number of aggressive acts for the child to all other types of acts recorded for that child.

In retrospective observer judgments, judges are often instructed to compare a person to people in general with a rating scale. The middle point (e.g., '3' on a 1–5 rating scale) represents a theoretical normative reference point (i.e., people in general), while greater numbers represent higher levels of the trait and lesser numbers lower levels. Unipolar rating scales are anchored by a single trait word at the high end, while bipolar scales are also anchored with the psychological opposite at the low end (e.g., 'thoughtful' vs. 'inconsiderate'). Defining the anchors with concrete, descriptive phrases instead of abstract adjectives improves measurement validity. Often scores on several related rating scales will be averaged to produce an overall score on a trait. For example, broad extraversion versus introversion might be assessed by the average of ratings on more specific scales such as talkative vs. silent, outgoing vs. reserved, and so forth.

When interpreting rating scores, psychologists need to consider whether to take the numerical score at face value. For example, a rating of 4 on a 1–5 scale of thoughtfulness is above the theoretical norm of 3, but if the computed average rating for a large group of people is 4.5, a rating of 4 could be interpreted as a relatively low level of thoughtfulness. This problem is further complicated when some judges restrict their ratings to one portion of the scale. A 5 might actually indicate a much higher value from a judge who assigns mostly 3s than from a judge who assigns mostly 4s and 5s.

Psychologists unwilling to accept scores at face value will recalibrate all of a judge's ratings with respect to the mean of all ratings made by that judge. This process is called 'ipsatizing' scores. Some rating procedures expressly call for ipsative assessment in the act of judgment itself. For example, judges using the California Q-set (Block 1978) are required to sort 100 personality descriptions into nine categories following a specified distribution (five descriptions in each of the extremely characteristic and uncharacteristic categories, eight descriptions in each of the quite characteristic and uncharacteristic categories, and so forth, to 18 characteristics in the relative neutral or unimportant category).

Formats other than rating scales for observer judgment of personality include questionnaires and adjective checklists. Items in some personality questionnaires are phrased in ways that allow either self-report (Sect. 2.3) or observer judgment. Adjective check lists contain a set of adjectives that judges check if they believe them to apply to the person being judged. Scores on adjective check lists are computed by counting the number of judges who check a particular adjective and/or by summing the checks for a group of adjectives considered to measure the same trait.

Factors that potentially limit the validity of observer judgments include misleading communications, stereotypes, and gossip about those being judged; unfairly positive or negative attitudes from judges who like or dislike the targets; and insufficient knowledge due to judges knowing the target only in certain roles and settings. These limitations and methods for overcoming them are discussed by Block (1978), Hofstee (1994) and Kenrick and Funder (1988).

2.3 S-data: Self-Reports

The two basic types of self-report instruments are projective tests and objective questionnaires. In projective testing, respondents are asked to finish incomplete phrases or construct stories about intentionally ambi-

guous images. Following guidelines developed by the test author and community of test users, psychologists score the respondent's protocol for psychological themes. Proponents of projective tests claim that these instruments are able to tap deep, unconscious needs and motives; critics insist that scoring projective protocols is too subjective and unreliable. Research indicates that carefully designed projective tests can be as reliable and valid as objective measures.

Personality questionnaires (Angleitner and Wiggins 1986) rarely consist of questions anymore. Instead, questionnaire items are statements about one's self and other people. Respondents express how much they agree with each statement or the degree to which they think the statement applies to them. The most comprehensive personality questionnaires contain several hundred items. Any subset of items within a personality questionnaire that is scored for a particular trait is called a 'personality scale.' The major personality questionnaires contain as many as several dozen different scales. Items are collected into scales and responses to items are scored according to one of four strategies outlined below: empirical, rational-intuitive, theoretical, or factor-analytic.

2.3.1 Empirical scales. Paul Meehl (1945) argued that psychologists would be naively optimistic to take questionnaire item responses at face value or to attempt to judge their hidden psychological meanings expertly. More prudent, he suggested, would be to treat each item response as a bit of behavior whose meaning must be determined by its empirical correlates. Empirical scales are constructed by locating all items on a questionnaire that tends to be answered differently by two groups known by other methods to differ on a particular personality trait. A person's score on an empirical scale is defined by the number of responses that match the responses given by one of the groups used in original scale construction. That is, if a person answers many items the same way as a group of people known to be aggressive, that person is considered likely to be aggressive also.

2.3.2 Rational-intuitive scales. The rational-intuitive approach to personality scales suggests personality traits can be assessed straightforwardly by items whose content, according to common sense, seems relevant to the trait (Wolfe 1993). Thus, a person who endorses items such as 'I am an aggressive person' and disagrees with items such as 'I never get in fights' would receive points on a rational-intuitive personality scale of aggressiveness. The obviousness of rational-intuitive scales perennially raises concerns about self-enhancement (exaggerating socially desirable traits and denying undesirable traits), but research indicates that respondents self-enhance on personality questionnaires no more than they do in everyday life. Furthermore, research indicates that rational-intuitive scales validly predict relevant criteria as well as scales constructed by any other method.

2.3.3 Theoretical scales. Like rational-intuitive scales, theoretical scales are comprised of items whose content is judged to be relevant to the personality characteristic being assessed. The difference is that the relevance is not apparent to common sense and can only be seen by professionals versed in a particular theory. For example, the theoretical items 'I am fascinated by fire,' 'I would do anything on a dare,' 'My father and I always fought,' 'I own a gun,' and 'Women find me charming' seem unrelated to common sense, but, to a Freudian, these are items likely to be endorsed by a man with a phallic personality character resulting from an unresolved Oedipal conflict.

2.3.4 Factor-analytic scales. Factor-analysis is a statistical method for identifying clusters of items that tend to be answered the same way. This method, like the empirical method, begins with a large set of items that are administered to a group of respondents. If respondents who agree with item 'A' also tend to agree with items 'B,' 'C,' 'D,' and so forth, these items are deemed to measure the same psychological trait (Briggs and Cheek 1986). The nature of the trait is normally determined by rational-intuitive inspection of the content of the items. Factor analysis can be applied to scales as well as items, and factor analytic research has repeatedly indicated that much of the content of personality falls into five broad factor domains: extraversion, agreeableness, conscientiousness, emotional stability, and a fifth factor variously called intellect, imagination, or openness to experience. Many psychologists regard the 'Big Five' or 'Five-factor model' (Wiggins 1996) as a major integrating focus for future research.

2.4 T-data: Laboratory Tests

Assessing personality by testing respondents in laboratories or other controlled conditions is motivated by two interests. The first is establishing objective, replicable procedures that cannot be affected by biases and errors potentially found in observer judgments or self-reports. In particular, the measurement of involuntary reactions such as changes in electrodermal conductivity or heart rate to assess anxiety, or pupil dilation to assess interest, is seen as a way of circumventing dissembling that can occur with self-reports. Unfortunately, as Block (1977, 1978) points out, T-data are always indirect measures of personality, and laboratory tests that might seem to be reasonable measures of personality have a record of failing unpredictably.

The second motivation for using laboratory tests is the particular suitability of such procedures for measuring certain traits. Laboratory tests are particularly apt for assessing cognitive performance variables. For example, a personality theory built around the mental accessibility of different concepts would naturally be tested by measuring reaction time to words representing different concepts. Cognitive styles are almost invariably measured by performance tasks.

Personality theories that attempt to explain mental or behavioral differences in terms of underlying biological differences also require laboratory facilities to assess those differences. In addition to traditional psychophysiological recordings, new laboratory tests for measuring the physical basis of personality include biochemical assaying, positron emission tomography and functional magnetic resonance imaging. Details of these methods can be found in sources such as Davidson (1999), and Pickering and Gray (1999).

3. *Three Purposes of Measurement*

The three major purposes for measuring individual differences are: (a) making decisions about people, (b) helping people make decisions, and (c) conducting basic research. All three cases involve prediction (for further information see Wiggins 1973). In decision-making, personality assessments are used to predict how an individual person will think, feel, behave, or be perceived by others if various courses of action are pursued. In research, predictions are made about how different forms of individual differences are related to one another.

3.1 *Making Decisions About People*

In clinical and counseling psychology, personality scores are used to predict what course of therapy will best help persons with psychological problems. In personnel psychology, personality scores are used to predict which individuals will perform best if hired and placed in particular jobs. Some applied psychologists endorse what they call an 'assessment' approach to these decisions, in which the decision maker intuitively weighs all the information gathered about the person. In contrast, a 'statistical' approach inserts personality scores into a mathematical equation developed from past empirical research. Studies indicate that statistical predictions are almost invariably more accurate than assessment predictions.

3.2 *Helping People Make Decisions*

Methods for helping people make decisions with personality measures differ from making decisions about people only in terms of who is making the decisions. Individuals seeking greater satisfaction in personal relationships, for example, may complete a personality questionnaire to increase self-insight, much as they might read a self-help psychology book. Likewise, individuals uncertain about career choice can complete questionnaires that predict which careers would be most satisfying. Traditionally, psychologists interpret and discuss personality scores with clients, but some self-help personality measures are completely self-administered.

3.3 *Basic Research: Uncovering Covariation Among Variables*

The usefulness of all applied personality methods depends upon the ability of researchers to construct valid personality measures and to ascertain reliable covariation among these measures. This process, called 'construct validation,' is identical with any other type of scientific hypothesis testing (Hogan and Nicholson 1988). In the typical case, construct validation takes the following form. A researcher's theory predicts that individual differences in a particular personality trait (say, conscientiousness) will covary with differences in some L-data (say, job performance). A method is devised to measure the personality trait with O-, S-, or T-data and the life event with L-data. Successful prediction supports both the validity of the measures and the theory that led to the prediction. Predictive failure means either the hypothesis was incorrect, a procedural error occurred (inappropriate research sampling, administration, or scoring), or one or both of the measures lack validity. Progress in personality research occurs when many successful predictions leads to the acceptance of a measure as 'well-validated.' Careful research with well-validated measures always advances knowledge because even predictive failures indicate the need to revise hypotheses.

Because valid measurement is crucial to the entire personality enterprise, a significant amount of personality research is directed at improving measurement methods. Some of this research aims to clarify the dynamics of the measurement process, that is, the psychological processes that occur during observer judgments and self-reports. A second line of research employs computers to administer, score, and interpret personality tests. When computer programs for analyzing personality data are combined with artificial intelligence, 'observers' with artificially constructed personalities will some day make 'observer judgments' of personality worldwide over the Internet.

See also: Personality Assessment; Personality Psychology; Personality Structure; Personality Theories; Projective Methods in Psychology

Bibliography

Allport G W 1937 *Personality: A Psychological Interpretation*. H Holt, New York

Angleitner A, Wiggins J S 1986 *Personality Assessment via Questionnaires*. Springer-Verlag, Berlin

Block J 1977 Advancing the psychology of personality: Paradigmatic shift or improving the quality of research? In: Magnusson D, Endler N S (eds.) *Personality at the Crossroads: Current Issues in Interactional Psychology*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 37–64

Block J 1978 *The Q-Sort Method in Personality Assessment and Psychiatric Research*. Consulting Psychologists Press, Palo Alto, CA

Briggs S R, Cheek J M 1986 The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality* **54**: 106–48

Cronbach L J 1957 The two disciplines of scientific psychology. *American Psychologist* **12**: 671–84

Davidson R J 1999 Biological bases of personality. In: Derlega V J, Winstead B A, Jones W H (eds.) *Personality: Contemporary Theory and Research*, 2nd edn. Nelson Hall, Chicago, pp. 101–25

Hofstee W K B 1994 Who should own the definition of personality? *European Journal of Personality* **8**: 149–62

Hogan R, Nicholson R A 1988 The meaning of personality test scores. *American Psychologist* **43**: 621–6

Kenrick D T, Funder D C 1988 Profiting from controversy: Lessons from the person-situation debate. *American Psychologist* **43**: 23–34

Meehl P E 1945 The 'dynamics' of structured personality tests. *Journal of Clinical Psychology* **1**: 296–303

Pickering A D, Gray A D 1999 The neuroscience of personality. In: Pervin L A, John O P (eds.) *Handbook of Personality: Theory and Research*, 2nd edn. Guilford Press, New York, pp. 277–99

Wiggins J S 1973 *Personality and Prediction: Principles of Personality Assessment*. Addison-Wesley Pub Co, Reading, MA

Wiggins J S (ed.) 1996 *The Five-Factor Model of Personality*. Guilford Press, New York

Wolfe R N 1993 A commonsense approach to personality measurement. In: Craik K H, Hogan R, Wolfe R N (eds.) *Fifty Years of Personality Psychology*. Plenum Press, New York, pp. 269–90

J. A. Johnson

Personality Structure

Personality structure is the branch of personality psychology that is concerned with quantitative representations of the ways in which persons consistently differ from one another in their characteristic adaptations to social environments (see *Personality and Adaptive Behaviors*). The basic unit of study is the 'personality trait' which is considered to be a moderately heritable disposition to behave consistently across different situations. Such traits are thought to be manifest in the individual differences in thinking, feeling, and behaving reported in responses to personality questionnaires or to lists of trait-descriptive adjectives. Current research in the field of personality structure emphasizes two quite different structural representations of personality traits: the Five Factor Model (FFM) of personality traits and the Interpersonal Circumplex (IPC) model of interpersonal dispositions. These two models originated in different historical contexts, are based on different assumptions concerning the nature of traits, and employ different statistical models for representing dimensions of personality. Nevertheless, the FFM and the IPC models of personality structure may be viewed as complementary rather than competitive, and as being applicable to different but equally important realms of knowledge.

1. *Origins of the Five Factor Model of Personality Structure*

The idea that variations in human characteristics have an evolutionary origin was suggested at the end of the eighteenth century by Erasmus Darwin (1731–1801) and it later bore fruit in the empirical researches of two of his grandsons, Charles Darwin (1809–82) and Francis Galton (1822–1911). That Galton was uniquely qualified to conduct research on 'hereditary genius' is suggested by the accomplishments and eminence of the Darwin family and by the fact that Galton was, by any criteria, a genius himself. Galton coined the term 'eugenics' (Greek *eugenes* 'well-born') for the field which studied the inheritance of desirable characteristics, especially intelligence. In 1904, Galton endowed a permanent chair in eugenics at University College, London and appointed the mathematician, Karl Pearson (1857–1936) as its first professor. Pearson developed the correlation coefficient to express the relation between the intelligence of parents and the intelligence of their offspring. His colleague, Charles Spearman (1863–1945) developed the method of factor analysis (see *Personality Psychology: Methods*) which, when applied to mental ability data, suggested a general factor of intelligence ('g') in addition to more specific factors. Raymond B. Cattell (1905–98), the founder of modern personality structure research, received his doctorate under Spearman and was the first to suggest extending the applications of factor analysis to the study of individual differences in personality characteristics.

Cattell (1967) had a master plan for the study of personality structure, which he and his associates implemented over a period of 65 years: (a) determine the totality of trait-descriptive terms in a given language that denote individual differences in thinking, feeling, and behaving ('surface traits'); (b) reduce correlated clusters of these surface trait terms by

factor analysis to their underlying determinants ('source traits'); and (c) confirm, by factor analysis and experimentation, the extent to which these source traits, as measured by personality questionnaires, may also be found in different media of observation such as ratings by observers, self-ratings, and behavioral responses to controlled test stimuli. Cattell's master plan set the stage for two distinct, but conceptually related, approaches to the study of personality structure that have only recently converged.

2. Two Traditions of Five Factor Model Research

2.1 The Lexical Tradition

The totality of surface traits in a given language may be defined with reference to the words that are used to mark individual differences in thinking, feeling, and behaving found in an unabridged dictionary of that language. Within the lexical tradition, representative words (typically adjectives) that have been selected from dictionaries are administered in self-report and observer-report format to groups of respondents who indicate the extent to which each word is descriptive of themselves or of someone they know well. Such data are then analyzed by first calculating the extent of co-occurrences (correlations) among words, and then determining the underlying 'structure' of the entire set of words by factor-analytic techniques.

Lewis R. Goldberg (1981) maintained that: (a) the frequency with which an attribute is represented in a language corresponds to the general importance of that attribute in a given culture and (b) the most important dimensions of personality are those 'universal dimensions' that are common to all languages. Goldberg and his colleagues have studied personality trait-descriptors in English and 12 other languages (e.g., German, Italian, and Turkish) and have concluded that there are *five factors* (which they called the 'Big Five') that may be considered universal dimensions of personality surface trait structure (see *Cross-cultural Psychology*). The Big Five dimensions of personality that have been identified in many different language groups are: (a) Extraversion (talkative, assertive, verbal); (b) Agreeableness (kind, cooperative, warm); (c) Conscientiousness (organized, systematic, practical); (d) Neuroticism (anxious, moody, emotional); and (e) Intellect/Openness (intellectual, imaginative, complex).

2.2 The Multivariate-trait Tradition

The second part of Cattell's master plan involved the development of personality questionnaires that would capture the basic dimensions of personality inherent in language. A personality questionnaire consists of items such as 'I like to attend parties and other social gatherings where I can meet new people' to which the test-taker is required to respond 'true' or 'false'; or more typically, to rate the degree to which the statement is self-descriptive on a scale ranging from 'not at all' to 'highly descriptive.' Most questionnaires are designed to measure many dimensions or variables, and hence are called 'multivariate.' Because responses to questions ('true vs. 'false'; 'descriptive' vs. 'non-descriptive') can be scored with complete reliability, the tests are referred to as 'objective.'

Cattell's suggestion that Spearman's statistical method of factor analysis could be applied to personality data was indeed a fruitful one, but the new field of personality structure also inherited some of the problems that had been associated with the use of factor analysis to determine the number and nature of mental abilities. Foremost among these problems was the issue of determining the *number* of factors 'underlying' the intercorrelations among personality scales. For example, the two most prominent figures in the field of personality structure during the twentieth century, R. B. Cattell in the United States and H. J. Eysenck in England, strongly disagreed about the number of factors that were required for the comprehensive measurement of personality structure. Cattell argued that at least 16 factors were necessary and he included these factors in his Sixteen Personality Factor Questionnaire (16 PF). Eysenck held that there were, at most, the three factors included in the Eysenck Personality Questionnaire (EPQ). A 1953 survey of the literature by psychologists at the Educational Testing Service identified and categorized the approximately 450 factors that had been reported up to that date. By the 1960s, the field of personality structure was being severely criticized, not only for its failure to agree on a common set of factors, but for its failure to demonstrate any substantial 'real world' correlates of such factors. At the same time, however, it was argued that the 'Big Two' dimensions of personality, Extraversion (see *Self-esteem in Adulthood*) and Neuroticism (see *Neuroticism*) were large, ubiquitous, and almost unavoidable dimensions of personality for which substantial and important correlates had been reported (Wiggins 1968). These dimensions had figured prominently not only in the research programs of Cattell and Eysenck, but in many other research contexts as well.

2.3 Reconciliation of the Two Traditions

P. T. Costa, Jr., and R. R. McCrae (1994) developed a personality questionnaire that included scales for the dimensions of Neuroticism and Extraversion suggested by Cattell and Eysenck, as well as a measure of Openness to Experience that was suggested by an analysis of Cattell's questionnaire. Their instrument, the NEO-Personality Inventory (NEO-PI), was later expanded to include the additional dimensions of Agreeableness and Conscientiousness that had been

identified in the lexical work of Goldberg and others. During the 1980s and 1990s, Costa and McCrae conducted an unprecedented number of empirical studies with the NEO-PI that established the utility of this instrument, as well as the comprehensiveness of what came to be known as the Five Factor Model (FFM) of personality structure. In addition to revitalizing the field of personality structure, the work of Costa and McCrae reaffirmed the basic logic of Cattell's master plan which called for an integrative framework for research that included both the lexical and multivariate-trait traditions.

3. *Research with the NEO Personality Inventory*

The NEO-PI has been an especially useful measure in the 'applied' contexts of personnel selection (Barrick and Mount 1991) and psychiatric diagnosis (Costa and Widiger 1992). In addition, the validity of the NEO-PI as a conceptually-justified measure of personality traits has been demonstrated with reference to the following criteria that would be required of such a measure.

3.1 *Heritability and Stability*

All five of the NEO-PI dimensions are moderately heritable (see *Behavioral Genetics: Psychological Perspectives*) and remarkably stable over the adult life of an individual. After the age of 30, these traits, as William James put it, appear to be 'set like plaster.' Costa and McCrae (1994) reported retest correlations ranging from 0.6 to 0.8 even after an interval of 30 years.

3.2 *Consensual Validation and Cross-cultural Invariance*

There are reasons to be skeptical of the 'truthfulness' of individuals' self-reports to personality inventories and for that reason, it is desirable to obtain ratings by knowledgeable observers of the respondent's personality characteristics on the same items (see *Person Perception, Accuracy of*). In many of the empirical studies conducted by Costa, McCrae, and associates, an Observer Form of the NEO-PI was completed by spouses, peers, or other knowledgeable informants. In general, substantial agreement has been obtained between self- and observer-ratings, indicating a satisfactory degree of 'consensual validity.' The NEO-PI has been translated into more than 30 languages, including non-Indo-European languages such as Thai and Shona. The success of these translations in producing the expected five-factor structure in other cultures attests to the cultural generalizability of the Five-Factor Model.

4. *Origins of the Interpersonal Circumplex Model*

Circular representations of human nature can be traced from ancient astrology to the drawings of Leonardo Da Vinci to the modern interpersonal circumplex tradition, which began with Harry Stack Sullivan. Sullivan defined psychiatry as the study of *interpersonal relations* and considered 'personality' to reside in an individual's patterns of relationships with significant others over time. Timothy Leary (1957) and his associates attempted to operationalize Sullivanian concepts in terms of concrete measurement procedures. Their analysis of clinicians' observations of the things that patients did to each other (and to themselves) in group psychotherapy led to a taxonomy of interpersonal variables that was empirically well captured by a *circular* arrangement of 16 (later 8) variables organized around the two coordinates of dominance (vs. submission) and affiliation (vs. hostility).

4.1 *The Structure of Interpersonal Behavior*

The Five Factor Model and Interpersonal Circumplex are based on different assumptions about the nature of personality and employ quite different structural models. The tradition that led to the current FFM focused on the *individual* and attempted to provide a comprehensive 'list' of the factors necessary for describing the individual's most important personality traits. The tradition that led to the current IPC focused on *dyadic relationships* and attempted a detailed description of a two-dimensional 'interpersonal space' within which interpersonal actions and reactions could be classified. The IPC tradition assumes that a person's characteristic interpersonal behavior may be expressed as a *combination* of scores on the two bipolar coordinates of Dominance (vs. Submission) and Affiliation (vs. Disaffiliation).

Interpersonal space is interpreted, quite literally, in terms of the geometry of the circle. The two underlying coordinates of this circle are Dominance (90°) vs. Submission (270°) and Affiliation (0°) vs. Disaffiliation (180°). A number of theoretical rationales have been given for the assumption that these coordinates are the 'basic' reference points for understanding interpersonal behavior. For example, it has been suggested that David Bakan's (1966) distinction between agency (being a differentiated individual, striving for mastery and power) and communion (being part of a larger entity, striving for intimacy and union) provides a metatheoretical basis for interpreting Dominance and Affiliation as the basic axes of social life. Within this conceptual framework, Dominance vs. Submission and Affiliation vs. Disaffiliation are the nodal reference points on the IPC and all other behaviors may be interpreted as 'blends' of these two coordinates. For example, Extraversion (45°) is an

equal blend of Dominance and Affiliation; Introversion (225°) is a combination of submissive and disaffiliative tendencies.

4.2 The Nature of Interpersonal Space

Unlike the competing inventories of personality structure found within the multivariate-trait tradition (e.g., Cattell, Eysenck, Costa, and McCrae), different IPC measures share the same structure, but differ in their domains of application. For example, IPC models of personality and emotions have coexisted comfortably for many years and are likely to continue to do so (Plutchik and Conte 1997). Four contemporary IPC measures have quite different areas of application: The Interpersonal Adjective Scales provide a measure of relatively normal interpersonal *traits* or dispositions. The Inventory of Interpersonal Problems measures *problems* of living as perceived by respondents or others. The Impact Message Inventory measures the *covert reactions* of respondents to different target persons corresponding to different locations on the IPC. The Supportive Actions Scale measures tendencies to provide various kinds of *social support* to others in need of assistance. It is clear from the approximately 1,000 references that appeared in Donald Kiesler's (1996) survey of the IPC literature that both the utility and generalizability of the IPC structure are well established.

5. Current Status

Examination of the relevant literature of the past decade reveals that the two major models of personality structure today are the Five-Factor Model or 'Big Five' and the Interpersonal Circumplex model. An early rapprochement between these two models occurred when proponents of each of the models agreed that the Extraversion and Agreeableness dimensions of the FFM were rotational variants of the interpersonal dimensions of Dominance and Affiliation and that a full description of personality structure cannot ignore the additional dimensions of Conscientiousness, Neuroticism, and Openness/Intellect. It has also become increasingly apparent that the FFM, which had been characterized by some as an 'atheoretical' model, lends itself to interpretations from a variety of different theoretical perspectives (Wiggins 1996).

See also: Extraversion; Genetic Studies of Personality; Neuroticism; Personality and Conceptions of the Self; Personality Assessment; Personality Development and Temperament; Personality Development in Adulthood; Personality Development in Childhood; Personality Psychology; Personality Theories

Bibliography

Bakan D 1966 *The Duality of Human Existence: An Essay on Psychology and Religion*. Rand McNally, Chicago

Barrick M R, Mount M K 1991 The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology* **44**: 1–26

Cattell R B 1967 The *Scientific Analysis of Personality*. Penguin, Harmondsworth, UK

Costa P T, Jr., McCrae R R 1994 'Set like plaster'? Evidence for the stability of adult personality. In: Heatherton T, Weinberger J L (eds.) *Can Personality Change?*. American Psychological Association, Washington, DC, pp. 21–40

Costa P T, Widiger T A Jr. (eds.) 1994 *Personality Disorders and the Five-Factor Model of Personality*. American Psychological Association, Washington, DC

Goldberg L R 1981 Language and individual differences: The search for universals in personality lexicons. In: Wheeler L (ed.) *Review of Personality and Social Psychology*. Sage, Beverly Hills, CA, Vol. 2, pp. 141–65

Kiesler D J 1996 *Contemporary Interpersonal Theory and Research: Personality, Psychopathology, and Psychotherapy*. Wiley, New York

Leary T F 1957 *Interpersonal Diagnosis of Personality*. Ronald Press Co., New York

Plutchik R, Conte H R (eds.) 1997 *Circumplex Models of Personality and Emotions*. American Psychological Association, Washington, DC

Wiggins J S 1968 Personality structure. *Annual Review of Psychology* **19**: 293–350

Wiggins J S (ed.) 1996 *The Five-factor Model of Personality: Theoretical Perspectives*. Guilford Press, New York

J. S. Wiggins

Personality Theories

Within psychology, personality refers to the individual's distinctive and enduring characteristics, including stable patterns of thoughts, feelings, emotions, and behavioral tendencies. No other area of psychology covers as much territory as the field of personality does; it is at the intersection of the study of human development, abnormality and deviance, temperament, emotions and thought, learning, social relations, and even the biological foundations of human behavior. Although the traditional focus is on stable individual differences in basic tendencies or dispositions (e.g., extraversion, conscientiousness), recent approaches also emphasize the psychological processes (such as learning, motivation, emotion, and thinking) that underlie them. Increasing attention is also given to the possible biological-genetic roots of these tendencies and processes, and their interactions with social-environmental influences.

Personality theorists try to answer such questions as: How stable and enduring are particular psychological qualities? What is the role of inheritance in the expression of personality, and what is acquired through experience with the environment? Does what

we do and think and feel characteristically depend mostly on the individual or on the situation? How do different types of people interact with different types of situations? Historically, there have been five main approaches to personality that address these questions: the psychodynamic, the behavioral, the phenomenological, the trait and biological, and the social cognitive. All are concerned with uncovering basic, general principles that can account for the wide range of behavior that people are capable of engaging in, and each proposes different units for conceptualizing and studying people. Each approach is considered in turn.

1. *Psychodynamic Approaches*

The psychodynamic approach to personality was founded by the Viennese physician Sigmund Freud; (see *Freud, Sigmund (1856–1939)*). Rejecting the earlier belief that people's behavior is under their conscious and rational control, Freud believed that behavior is psychologically determined by underlying unconscious causes and motives. Freud saw patients who displayed strange symptoms that had no discernible physical cause, such as a young woman who appeared to be blind although tests of her vision showed that her eyes and visual system were undamaged. To explain such puzzling cases, Freud proposed an 'anatomy' of the mind consisting of three institutions or agencies of mental operation—the id, the ego, and the superego. The id—the primary, instinctual core—obeys the 'pleasure principle,' seeking immediate gratification of basic biological sexual and aggressive impulses, regardless of reality considerations. The superego, on the other hand, represents the moral standards of a person's society, obtained through the internalization of parental values, rules, and characteristics in the course of socialization. The ego mediates between the instinctual demands of the id and the outer world of reality, localizing the appropriate objects for gratification in the environment so that the impulses of the id can be satisfied. It operates by means of logical thinking and rational planning. Its tasks include: (a) the control of unacceptable impulses or desires from the id; (b) the avoidance of pain produced by internal conflict in the efforts to control and master those unacceptable impulses; and (c) the attainment of a harmonious integration among the needs of both the id and the superego.

According to Freud ([1915] 1957), the id, the ego, and the superego are always in dynamic conflict (thus the term psychodynamics). Freud believed that in this conflict, a person's basic impulses persist and press for discharge, but the people or 'objects' at which they are directed, and the manner in which they are expressed, are transformed and disguised unconsciously to make them more acceptable and to reduce conflict and anxiety. When the young child becomes anxious about his or her own impulses (e.g., because of fear that they will lead to parental punishment), attempts are made to repress them or otherwise disguise and redirect them. The main defense mechanism that Freud proposed—repression—massively inhibits threatening impulses or events, making them unconscious. In projection, one's unacceptable impulses or aspects are attributed to other people.

Modern psychodynamic theories tend to place much less emphasis on biological drives and highlight the individual's social circumstances and relationships with significant others. Attachment theorists, for example, emphasize the quality and varieties of early mother–child attachment relations and their consequences in the course of development (Ainsworth and Bowlby 1991). Based on experiences in this relationship, the child develops internal working models (mental representations) of others, of the self, and of relationships, which guide subsequent experience and behavior. Children who have positive, gratifying experiences with significant others tend to develop internal working models of others as responsive and giving, and of themselves as worthy of attention; those who have had painful or unsatisfying experiences are likely to develop internal models that reflect those troubled relationships. Early attachment styles may enduringly influence relationships throughout life, including the way one parents one's own children (Hazan and Shaver 1987; see *Attachment Theory: Psychological*; *Bowlby, John (1907–90)*).

Many of the ideas introduced by the psychodynamic approach, notably that much mental activity occurs outside of mental awareness, also are being reconceptualized in light of current theory and findings on how thought and memory work. For example, when cognitive (mental) representations of significant others are activated by a newly encountered person, the feelings and attributes associated with this representation may be applied to that new person in making inferences about him or her (Andersen and Berk 1998). This is consistent with Freud's concept of transference, said to occur when the patient responds to the therapist as if he or she were the patient's father, mother, or some other childhood figure. This modern approach, however, views transference in information-processing terms rather than as a reflection of the psychosexual impulses and conflicts hypothesized in classical psychodynamic theory (Westen 1998; see *Mental Representation of Persons, Psychology of*).

2. *Behavioral Approaches*

Although many psychologists were intrigued by the insights of Freud and his followers, they were primarily dedicated to developing a more scientific, rigorous approach to personality that would be objectively testable, preferably by laboratory tech-

niques. One of the pioneers in the behavioral approach, the US psychologist B. F. Skinner (1904–90), criticized many motivational concepts as being no more than labels, and even fictions, that were unhelpful for explaining what people actually do. Skinner and other behaviorally oriented psychologists analyzed problematic behaviors (e.g., aggressiveness, bizarre speech patterns, smoking, fear responses) in terms of the observable events and conditions that seem to vary with them. They then tried to discover the external events that strengthened their future likelihood and that maintained or changed the behavior of concern. Behavioral approaches have led to innovations for therapy by attempting to modify problematic behavior not through insight, awareness, or the uncovering of unconscious motivations, but rather by addressing the behavior itself and modifying it directly. Such behavior therapies emphasize the individual's current maladaptive behaviors (rather than their historical origins) and assume that they can be understood and changed by learning principles that deal with the acquisition and extinction of behavior patterns.

Although the systematic rigor of behavioral approaches was widely appreciated, the relevance of the approach for understanding the complex lives of people under the often unpredictable social conditions of life remained in question. Thus, a number of theorists began to make learning and behavior theories more 'social,' not limiting themselves to simple reinforcement principles and increasingly relying on mental or 'cognitive' processes in their account of the development of complex social behavior. In Albert Bandura's (1986) social learning approach, for example, personality is seen as the outcome of social learning processes through which distinctive beliefs, self-efficacy expectations, as well as personal standards that guide characteristic behavior patterns, are acquired. He emphasized that much social learning, from table manners and interpersonal relations to working habits to coping patterns, occurs through observation of the behavior of social models without any direct rewards or reinforcement administered to the learner (see *Social Learning, Cognition, and Personality Development*).

In another direction, the assumption that the 'laws' of conditioning are universal and apply broadly has been challenged by convincing evidence that organisms (including humans) seem biologically prepared to learn some associations or pairings more readily than others. Prewired dispositions in the brain seem to make humans distinctively prepared to learn diverse high-level mental activities, from language acquisition to mathematical skills to space perception (Pinker 1997). In yet another direction, technological advances in brain imaging now allow researchers to use methods like Magnetic Resonance Imaging (MRI) to observe areas within the brain that become activated in the course of mental activity, such as the emotional centers that may be especially important bases for individual differences (Ledoux 1996). Advances in technology have therefore made it possible to objectively study mental events, such as emotions, memories, and attention, going far beyond the early behaviorism that confined itself to overt behavior.

3. *Phenomenological Approaches*

In the middle of the twentieth century, phenomenological approaches arose, in part, as a humanistic protest against the earlier psychodynamic and behavioristic views. Phenomenologically oriented theorists argued that personality is not merely passively molded by internal motivational or external situational forces that 'shape' what the person becomes. Instead, people are active agents in the world and have a measure of control over their environment and their own lives. In this view, people are considered capable of knowing themselves and of being their own best experts. Self-knowledge and self-awareness become the route to discovering one's personality and genuine self.

Phenomenological approaches to personality (sometimes called self theories, construct theories, and humanistic theories), tend to reject many of the motivational concepts of psychodynamic theories and most of the environmental determinism of behavioral theories. Instead, their focus is on the development of an active 'self': People develop self-concepts and goals that guide their choices and their life course. Understanding personality, as well as the person's goals and choices, requires attention to how the individual characteristically perceives, thinks, interprets, and experiences or even 'constructs' the personal world.

George Kelly's (1905–67) theory of personal constructs, for example, emphasized people's subjective perceptions as the determinants of their behavior. Kelly believed that, just like scientists, people generate constructs and hypotheses both about themselves and about how the world works; they use these constructs to anticipate, understand, and control events in their lives. Therefore to understand people, one has to understand their constructs, or personal theories. Problems develop when the constructs people generate don't work well for them, when they are 'bad scientists' and fail to 'test' their constructs or hypotheses against the realities of the environment, or when they see themselves as helpless victims of their own personalities or life situations. Kelly's principle of 'constructive alternativism' held that all events in the world, including one's own behavior and characteristics, can be construed in multiple, alternative ways. While it is not always possible to change these events, one can always construe them differently, thus influencing how one is affected by them and how one reacts to them.

Carl Rogers (1902–87), another pioneer of the phenomenological approach, proposed two systems:

the organism and the self (or self-concept). The organism is the locus of all experience, which includes everything potentially available for awareness. The self is that portion of the perceptual field that is composed of perceptions of characteristics of the 'I' or the 'me.' It develops from experiences and interactions with the environment, and also shows a tendency towards actualization. Rogers maintained that the central force in the human organism is the tendency to actualize itself—to move constructively in the direction of fulfillment and enhancement. The self may be in opposition or in harmony with the organism. When the self is in opposition or incongruence with the experiences of the organism (e.g., when the self tries to be what others want it to be instead of what it really is), the person may become anxious, defensive, and rigid. However, when the self is open and accepting of all of the organism's experiences without threat or anxiety, the person is genuinely psychologically adjusted, for the self and the organism are one.

In contemporary work, the 'self' is seen as multifaceted and dynamic, consisting of multiple self-concepts that encode different aspects of the person (e.g., self as lover, self as father, the 'ideal' self, the 'actual' self) and become differentially salient depending on context (Markus and Nurius 1986). According to Higgins (1987) for example, a perceived discrepancy between the mental representation of the person one would ideally like to be (the ideal self) and the representation of who one actually is (the actual self) makes one more vulnerable to feelings of dejection, such as disappointment or dissatisfaction. In contrast, a discrepancy between one's representation of who one ought to be (the ought self) and the actual self can lead to feelings of agitation such as fear and worry. Motivation for behavior change arises from the conflicts each individual feels among his or her various representations of the self. For instance, upon receiving a low grade on an exam, an undergraduate may subsequently study very hard to relieve the guilt of not living up to what she herself perceives to be her responsibility as an exemplary student. Alternatively, she may re-evaluate her negative interpretation of past events, thinking about all of the good grades she has got in other classes and the myriad of other activities she is involved in (see *Personality and Conceptions of the Self*).

4. *Trait and Biological Approaches*

In everyday life, people readily characterize each other in terms of personality characteristics: he or she is friendly, assertive, submissive, conscientious, and so on. The essence of the trait approach, whose fundamental premises date back to the ancient Greeks, is the assumption that behavior is primarily determined by a number of stable, generalized personality traits that express themselves in many contexts. Guided by this assumption, advocates of this approach try to identify and measure individuals' traits and to discover the most fundamental traits on which people can be compared.

A principal focus of research on traits is on measurement—that is, the development of quantitative ways of finding and describing important stable individual differences. Traits are inferred from questionnaires, ratings, and other reports about the person's dispositions. Usually, the person's self-reports (or someone else's reports about the person) are taken as direct signs of the relevant traits. For example, the more one rates oneself as aggressive, the more one is assumed to have an aggressive disposition. The trait approach recognizes that behavior can vary depending on the situation but has focused on individual differences in the overall response tendency averaged across many situations.

Some consensus has grown among many researchers to focus on five large factors or dimensions of personality that have emerged from statistical analyses of traits. These factors, often referred to as the 'Big Five,' comprise openness to new experience, conscientiousness, extraversion (or outgoingness), agreeableness, and neuroticism. Considerable stability has been demonstrated on trait ratings and questionnaires related to the Big Five, particularly during the adult years (McCrae and Costa 1990; see *Extraversion*; *Neuroticism*).

In a different direction, the British psychologist Hans Eysenck (1916–97) and his associates have led the way in connecting psychological dispositions to their biological foundations. According to Eysenck, introverts need only a small amount of stimulation to overstimulate their central nervous system (CNS) which then leads them to become withdrawn in their behavior. In extraverts, by contrast, the CNS is not easily stimulated, leading them to seek activities that will increase stimulation levels, for example, by socializing more actively and by seeking activities such as parties. In support of his theory, Eysenck (1971) found that extraverts reported earlier, more frequent, and more varied sexual experiences. In another study, introverts showed greater changes in their brain wave activity in response to low frequency tones (Stelmack and Michmaud-Achron 1985), indicating their lower threshold for stimulation to the CNS.

4.1 *Genetic Roots of Personality*

The rapidly developing field of behavioral genetics studies the role of inheritance in personality, both in terms of dimensional traits (such as extraversion–introversion) and temperaments (such as general levels of emotionality, sociability, and activity; see *Temperament and Human Development*). Behavioral genetics most often uses the 'twin method' to assess genetic influence, comparing the degree of similarity on trait measures obtained for genetically identical (mono-

zygotic) twins as opposed to twins that are fraternal (dyzygotic) and are only 50 percent similar genetically. To the degree that genetic factors affect a trait, it follows that identical twins must be more similar than fraternal twins with regard to that trait. Estimates of genetic influence vary across studies and measurement strategies. For example, they tend to be higher for self-reports of personality than for observational studies (Miles and Carey 1997). Nevertheless, the general conclusion that identical twins are more similar than fraternal twins has received considerable support, and indicates an important role of heritability in personality (Plomin et al. 1997).

Even highly heritable dispositions, however, can be constrained and limited in their full expression, as when a person's growth and ultimate height are affected by nutrition, disease, or development. Environmental influences also can change the hard wiring of the brain—the neuronal structures themselves—and thus produce stable changes within the person at the organic level. For example, stress can shrink the size of the hippocampus, a brain structure basic for higher order mental functions (Sapolsky 1996). Thus, although the social environment cannot affect the structure of DNA (barring biochemical or radiation exposure), it can influence their expression, the brain, and the person's personality (see *Behavioral Genetics: Psychological Perspectives*; *Stress: Psychological Perspectives*).

Unquestionably, one's genetic endowment has extensive influence on one's life and personality development. Just as noteworthy, however, the same findings also point to the importance of experiences and the environment for personality (Plomin et al. 1997). The expressions of genetic influences and the situations and events the person experiences are in continuous interaction, and it may be this interplay that is most important in personality development.

5. *Social Cognitive Approaches*

In the 1960s, a 'cognitive revolution' took place in psychology, as attention turned from behavior to the person's mental (cognitive) processes and structures. Although behavioral approaches had asserted that stimuli control behavior, evidence accumulated to suggest that the perceivers' mental representations and cognitive transformations of stimuli can determine and even reverse their impact. Such transformations were illustrated in research on the determinants of people's ability to forgo gratifying their immediate impulses for the sake of more highly desired but temporally distant goals. This kind of 'willpower' has been studied extensively in the delay of gratification paradigm, in which preschoolers are faced with the dilemma of being able to have a small, less desired reward now (e.g., one little pretzel or marshmallow) or having to wait instead for a larger, more desired reward (e.g., two little pretzels or two marshmallows). The number of seconds that preschoolers are able to delay predicts a variety of positive developmental outcomes, including their SAT scores and their social competence in adolescence (Mischel et al. 1989; see *Self-regulation in Childhood*; *Self-regulation in Adulthood*). The duration of delay time itself depends importantly on how the children represent rewards mentally and selectively attend to different features of the situation. For example, if the young child focuses cognitively on the consummatory qualities of the reward objects, such as the pretzel's salty, crunchy taste, he or she tends to be able to wait only a short time. By thinking about the stick pretzels as little logs or about the marshmallows as puffy clouds, however, the child may be able to wait much longer for the reward. These results indicate that what is in the children's heads— not what is physically in front of them—determines their ability to delay.

Concurrent with the cognitive revolution, questions arose about cross-situational traits and psychodynamics as basic units of personality. Walter Mischel (1968) in particular showed that what people do characteristically depends importantly on the situation and context. For example, the person who seems conscientious about work may show a very different pattern with family. These findings indicated that broad trait scores and projective psychodynamic methods do not accurately predict what a given person will do in different kinds of situations and fail to explain the seeming inconsistency or variability that was observed within individuals across those situations.

The social cognitive approach to personality emerged in the 1960s and 1970s as an attempt to understand both the stable ways in which the person cognitively processes social information and the ways in which the person's behavior varies with regard to the situation. In this approach, the basic 'units' of personality are conceptualized as a set of five relatively stable person variables (Mischel and Shoda 1995): the individual's encodings or construals (of self, other people, situations); expectancies and beliefs (about outcomes and one's own self-efficacy); subjective goals and values; competencies (for the construction and generation of social behavior) and self-regulatory strategies and plans in the pursuit of goals; and their affects (feelings, emotions, and affective responses). Furthermore, clues about the person's underlying qualities may be seen in when and where a type of behavior is manifested, not only in its overall frequency. If so, the patterns of situation–behavior relationships shown by a person might be a possible key to individuality and personality coherence, rather than a source of error to be eliminated systematically (see *Interactionism and Personality*).

Evidence for the existence and meaningfulness of the stable situation–behavior relations predicted by the social cognitive approach came from an extensive

observational study conducted in a residential summer camp setting for children (Shoda et al. 1994). In this study, children's behavior was recorded specifically in relation to five camp-relevant situations (e.g., being teased, threatened, or provoked by peers; being praised by an adult). The children's social behavior was observed on selected dimensions (e.g., aggression, withdrawal, friendly behavior) as it occurred in relation to each of the intepersonal situations. The individuals' situation-behavior, 'if … then …' personality 'signatures' (e.g., if situation X, then they do A, but if situation Y, then they do B) were found to be both distinctive and stable. For example, one child was consistently more verbally aggressive than others when warned by an adult, but showed less aggression than others on average when approached positively by a peer. In contrast, another child was most verbally aggressive when approached positively by peers, but not particularly aggressive when warned by an adult.

To account for such findings, a cognitive-affective personality system (CAPS) theory has been proposed (Mischel and Shoda 1995). In this theory, the individual is characterized not only by the particular subset of goals, ways of encoding the world, and self-regulatory competencies that may be potentially accessible to him or her, but also by the distinctive and stable organization of relationships among the person variables available in the personality system— i.e., the person's distinctive 'network.' When a person encounters a particular situation, the CAPS network is sensitive to particular features of situations, which become encoded and activate situation-relevant cognitions and affects (thoughts and feelings) within the system. These units, in turn, make other cognitions and affects accessible while inhibiting others. The organization of relations within the person's network remains relatively stable and invariant across situations. However, as the individual moves across situations that contain different psychological features, different mediating units and their characteristic interrelationships become activated in relation to these psychological conditions.

Thus the person's unique 'network'—which is distinctive both in the types of mediating units available within the system as well as in the relationships among these units—guides and constrains the activation of the specific cognitions, affects, and potential behaviors when an individual processes situational features. It constitutes the basic structure of personality and reflects and underlies the individual's uniqueness. When the *ifs* posed by the situation change, so do the *thens* generated by the personality system, but the relationship between them is stable. This type of system is intrinsically interactionist, and has been shown to generate both overall mean differences in a given behavior as well as the specific *if … then …* profiles that are a person's behavioral 'signature.'

To illustrate such a system in action, suppose a person is especially sensitive to rejection and is disposed to expect it, to look for it, and to react intensely to it (Downey and Feldman 1996). Such 'rejection sensitive' people may see even innocent or ambiguous behavior from a significant other as intentional rejection, triggering such thoughts as 'she doesn't love me,' which activate further thoughts and feelings of rejection and potential betrayal and abandonment. In turn, the person's scripts for coercive or controlling behaviors may become activated, leading to angry or hostile and even abusive reactions. Over time such hostility is likely to lead to actual rejection even when there was none before, further strengthening the cycle that characterizes this personality type.

Thus, the defining *if … then …* profile of rejection sensitivity—its behavioral signature—may include both being more prone than others to anger, disapproval, and coerciveness in certain types of situations in intimate relationships, as well as being more supportive, caring, and romantic than most people in other situations (e.g., in initial encounters with potential partners who are not yet committed to them). This example illustrates how the personality system remains stable, although the surface behaviors it generates change in relation to the situation. As in a musical piece, the notes played at any moment change, but they do so in an organized pattern that reflects the structure of the composition.

6. *Future Directions: Toward a Cumulative Science of Personality*

Historically, the field of personality has been characterized by many alternative approaches and conceptions of personality which competed against each other. Different approaches and theorists claimed to offer a comprehensive, unitary view of personality, to the exclusion of alternative viewpoints or approaches. However, the different approaches at least in part asked different questions, usually dealing only with selected aspects of the diverse phenomena subsumed under the construct of personality.

Personality psychologists are increasingly recognizing the importance of multiple approaches to understanding the person and many are trying to find ways to integrate them within a broader theoretical framework. Current psychodynamic approaches are beginning to incorporate insights from research on memory to test the fundamental tenets of psychodynamic theory. Similarly, some behavioral approaches are integrating findings on social learning and cognitive processes, aided by current technology to assess interactions between biological variables and learning. The phenomenological approach, with its emphasis on the person's subjective internal reality, continues to inform work on self-conceptions and how these influence the individual's choices and life course. Trait approaches are beginning to specify the boundary conditions within which traits will be selectively

activated and visible in behavior. Finally, cognitive-social approaches are incorporating the contributions of cognitive science and social psychology into the same framework used to understand the individual's characteristic cognitive-affective processes and dynamics. If these trends continue, personality psychology may be evolving into a more unified field that addresses the distinctive characteristics that people have and the psychological processes that underlie them, conceptualized within one overarching theoretical framework.

See also: Freud, Sigmund (1856–1939); Genetic Studies of Personality; Infant and Child Development, Theories of; Interactionism and Personality; Personality and Adaptive Behaviors; Personality and Conceptions of the Self; Personality and Crime; Personality and Marriage; Personality and Risk Taking; Personality and Social Behavior; Personality Assessment; Personality Development and Temperament; Personality Development in Adulthood; Personality Development in Childhood; Personality Psychology; Personality Psychology: Methods; Personality Structure; Psychological Development: Ethological and Evolutionary Approaches; Social Learning, Cognition, and Personality Development

Bibliography

Ainsworth M D S, Bowlby J 1991 An ethological approach to personality development. *American Psychologist* **46**: 331–41

Andersen S M, Berk M S 1998 Transference in everyday experience. Implications of experimental research for relevant clinical phenomena. *Review of General Psychology* **2**: 81–120

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall, Englewood Cliffs, NJ

Downey G, Feldman S I 1996 Implications of rejection sensitivity for intimate relationships. *Journal of Personality and Social Psychology* **70**: 1327–43

Eysenck H J 1971 Introverts, extroverts and sex. *Psychology Today* **4**: 48–51

Freud S [1915] 1957 *Instincts and their Vicissitudes. Standard edition*, Vol. 14. Hogarth, London

Freud S 1959 *Collected Papers*. Basic Books, New York, Vols. 1–5

Hazan C, Shaver P 1987 Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology* **52**: 511–24

Higgins E T 1987 Self-discrepancy: A theory relating self and affect. *Psychological Review* **94**: 319–40

Kernberg O 1976 *Object Relations Theory and Clinical Psychoanalysis*. Jason Aronson, New York

Kohut H 1980 *Advances in Self Psychology*. International Universities Press, New York

Ledoux J 1996 *The Emotional Brain*. Simon and Schuster, New York

Markus H, Nurius P 1986 Possible selves. *American Psychologist* **41**: 954–69

McCrae R R, Costa P T Jr. 1990 *Personality in Adulthood*. Guilford Press, New York

Miles D R, Carey G 1997 Genetic and environmental architecture of human aggression. *Journal of Personality and Social Psychology* **72**: 207–17

Mischel W 1968 *Personality and Assessment*. Wiley, New York

Mischel W, Shoda Y 1995 A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review* **102**(2): 246–68

Mischel W, Shoda Y, Rodriguez M L 1989 Delay of gratification in children. *Science* **244**: 933–8

Pinker S 1997 *How the Mind Works*. Norton, New York

Plomin R, DeFries J C, McClearn G E, Rutter M 1997 *Behavioral Genetics*, 3rd edn. W. H. Freeman, New York

Rogers C R 1980 *A Way of Being*. Houghton Mifflin, Boston

Sapolsky R M 1996 Why stress is bad for your brain. *Science* **273**: 749–50

Skinner B F 1974 *About Behaviorism*. Knopf, New York

Shoda Y, Mischel W, Wright J C 1994 Intra-individual stability in the organization and patterning of behavior. Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology* **65**: 1023–35

Stelmack R M, Michaud-Achorn A 1985 Extraversion, attention, and habituation of the auditory evoked response. *Journal of Research in Personality* **19**: 416–28

Westen D 1998 The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science. *Psychological Bulletin* **124**: 333–71

W. Mischel and R. Mendoza-Denton

Personality Theory and Psychopathology

The idea that enduring personality traits are closely related to various disease states or behavioral disorders can be traced back to initial speculation by the early Greeks. Hippocrates believed that all disease stemmed from imbalances in four bodily humors: yellow bile, black bile, blood, and phlegm. While Hippocrates' early ideas bear a rudimentary resemblance to some recent approaches to personality, most contemporary theories of the relationship between personality and psychopathology are considerably more complex, and sensitive to both endogenous and environmental variables. Before the personality–psychopathology relationship can be understood clearly, the concept of personality must be defined.

1. The Concept of Personality

The word personality is derived from the Greek *persona*, which referred to masks used in the early theater. Over time, the term has come to refer not only

to the exterior presentation of an individual, but the integrated internal systems which allow individuals to adapt to their environment. For example, Allport (1937) defined personality as 'the dynamic organization within the individual of those psychophysical systems that determine his unique adjustments to his environment.' This definition highlights the idea that personality reflects both psychological and biological processes that are used in coping with the everyday world. These points are made more clearly by Millon (1996, p. 4), who defines personality as 'a complex pattern of deeply embedded psychological characteristics that are largely nonconscious and not easily altered, expressing themselves automatically in almost every facet of functioning. Intrinsic and pervasive, these traits emerge from a completed matrix of biological dispositions and experiential learnings and ultimately comprise the individuals distinctive pattern of perceiving, feeling, thinking, coping, and behaving.'

Another important characteristic of personality is that it pertains to long-standing patterns of behavior which are relatively stable over time and are typically referred to as traits. The American Psychiatric Association has defined personality in the *Diagnostic and Statistical Manual of Mental Disorders* (4th edn.) as 'enduring patterns of perceiving, relating to, and thinking about the environment and oneself.' Taken together, the many definitions of personality which have emerged over the years highlight several consistent features: the integration of biological and psychological systems, which results in trait-like behavioral stability and function to promote environmental adaptation.

2. *Conceptual Debates Related to Personality*

Numerous conceptual debates have characterized the evolution of the study of personality. Pervin (1990) outlines several of these issues: 'Is personality primarily governed by internal factors of the person or environmental effects?'; 'Is personality determined primarily by nature or nurture?'; 'Is personality influenced by past experiences, present factors, or future expectations?'; and 'To what degree is personality modifiable?'

While each of these topics is significant in and of itself, this article will highlight several other key issues regarding the nature of personality which have generated considerable discussion.

3. *Trait-like Stability vs. Situational Determinism*

As noted previously, personality has historically referred to trait-like behavioral patterns which are stable across situations and time. However, a series of empirical studies in the 1920s and 1930s raised serious doubts about the trait-like stability of behavior and highlighted the significance of situational factors in determining variability in behavior. Classic studies by Columbia University Teachers College (1928) suggested that children's level of honesty did not follow a trait-like pattern of predictability, but was significantly influenced by the child's immediate environmental situation. Furthermore, numerous other studies emerged suggesting that specific behaviors measured across situations and time were quite variable, further calling into question the idea of trait like stability of behavior (Mischel 1968). Consequently, this era was characterized by a polemic atmosphere in which person vs. situation debates were common.

Ultimately, this criticism of trait constructs was followed by a period of greater integration of situational and trait theories of personality. For example, work by Epstein highlighted that personality traits may not be accurate predictors of specific individual acts, but may be excellent predictors of aggregated behavior across time and situations (Epstein 1979, Epstein and O'Brien 1985). In other words, traits may be good predictors of a person's behavior 'over the long haul.' Furthermore, theories emphasizing person–situation interactionism evolved and brought into greater focus the ways in which personality traits may interact with particular situational variables. The interactionist solution to the person–situation debate was to consider which personality and situational factors interact to produce consistent behaviors vs. those which interact to produce variable behavior (Ekehammer 1974). Behavior within this approach is considered to flow from transactions between a person and the environment, each of which influences the other.

4. *Should Disturbances in Personality be Conceptualized as Categories or Dimensions?*

Personality researchers frequently debate about whether personality can best be understood in categorical or dimensional terms. That is, is it more useful to determine what 'type' of personality an individual displays rather than what profile of various personality 'traits' best describes the individual? This becomes particularly significant for the study of disorders of personality, given that the primary classification schemes for psychiatric disorders (e.g., *Diagnostic and Statistical Manual of Mental Disorders, International Classification of Diseases*) rely on categorical models (see *Syndromal Diagnosis vs. Dimensional Assessment, Clinical Psychology of*). Categorical models parallel traditional, medical disease-based practices in which a person is considered either to have a given disease entity or not. Such an approach has a long history in psychiatry and psychology and offers several advantages: (a) clinicians are generally familiar with

the descriptive and clinical terminology and can apply it easily; (b) predictions can be made about a person based on the description of the category; for example, if a person is categorized as sociopathic, we may assume that rule violation or criminal activity is likely; (c) personality types have clear and vivid descriptions which facilitate communication; (d) a disease based approach is appropriate to the extent that personality disturbances are discrete and homogeneous diagnostic entities with minimal overlap between categories.

Unfortunately, the empirical literature has not supported this idea. People generally do not have a particular personality type that can be easily categorized. Rather, personality appears to be better described as an amalgam of traits, and most empirical studies suggest that personality is best conceptualized in such terms. Categorical models of personality have also been criticized because different measures of the same personality category show relatively low concordance or agreement (i.e., poor convergent validity) and also individuals who meet diagnostic criteria for a given personality type often meet criteria for several other types (i.e., poor discriminant validity). Indeed, if personality could be best understood in terms of unique and nonoverlapping categories, such problems with construct validity would not be so prominent. Furthermore, setting the minimum threshold for determining the presence of a personality disorder is a fairly arbitrary procedure and confusion regarding normality vs. abnormality is common. Finally, many personality disorder categories have not shown substantial stability over time, which would be anticipated with a true trait oriented personality disorder.

In contrast, the dimensional personality trait perspective has received considerable support in empirical studies. For example, when individuals with personality disorders are compared with individuals with normal personality, the distribution of personality scores does not show an 'either–or' pattern, suggesting that personality disordered individuals differ from normals only in the *magnitude* of certain key personality traits. Furthermore, a large number of factor analytic studies of personality traits in both normal samples and personality disordered samples yield the same basic factor structure for each population, most often implying that four or five broad personality traits, or factors, may account for variations in both normal and disordered personality. These five factors are frequently referred to as the Five Factor Model (FFM) and they are traditionally labeled in the following manner: Factor 1, surgency or extroversion; Factor 2, agreeableness; Factor 3, conscientiousness; Factor 4, emotional stability or neuroticism; and Factor 5, culture or openness to experience.

From a clinical perspective, the lack of empirical support for categorical distinctions between normal and abnormal personality and different personality types has led some researchers to suggest that personality disorders do not fit the medical model, and consequently personality 'diagnoses' should not be made. Rather, personality should be 'assessed' along core dimensions, such as the FFM or other clinical dimensions, (e.g., Minnesota Multiphasic Personality Inventory, see Helmes and Reddon (1993) for a review). Livesley et al. (1994) suggest that the simple presence of a personality disorder be determined on the basis of a behavioral pattern that results in personal or interpersonal *dysfunction*, but is then assessed, or described, on the basis of a dimensional system (e.g., FFM). Such an approach acknowledges that certain personality styles indeed cause distress (i.e., disorder), but retains the comprehensive assessment afforded by the dimensional approach. Whether or not such a system will ever be employed in psychiatric or psychological practice is likely to be determined more by political than by scientific forces.

5. *Relationship Between Personality and Psychopathology*

The idea that personality displays an important relationship to various forms of psychopathology has a long-standing history in psychopathology research. Conceptual understandings of the nature of this relationship have varied and generally represent one of several perspectives (Klein et al. 1993, Lyons et al. 1997, see Fig. 1). First, the *predispositional hypothesis* implies that personality variables precede and increase the risk of developing psychopathology. Also, personality and psychopathology are considered to be independent entities in terms of underlying causes or pathophysiology. An example of this perspective can be seen in early speculation that particular personality traits such as compliance, perfectionism, and dependence increase the risk of developing anorexia nervosa. Similarly, one may speculate that in some people, dependency may precede and increase the chances of developing depression. A second perspective, frequently called the *complication* or *scar hypothesis*, implies the opposite relationship between personality and psychopathology. From this view, psychopathology precedes and influences the development of particular personality traits which are considered the residue of the more active psychopathological syndrome. For example, research suggests that personality variables such as interpersonal dependence or obsessionality may increase after the onset of a major psychiatric syndrome such as depression, but then decrease as the syndrome resolves. The 'personality trait' may not be a true trait, but instead a complication of the depression.

A third perspective posits that both personality and psychopathology arise from the same underlying causal process. From such a *spectrum* or *common cause hypothesis*, a third variable is thought to increase the risk of both entities. From a spectrum perspective,

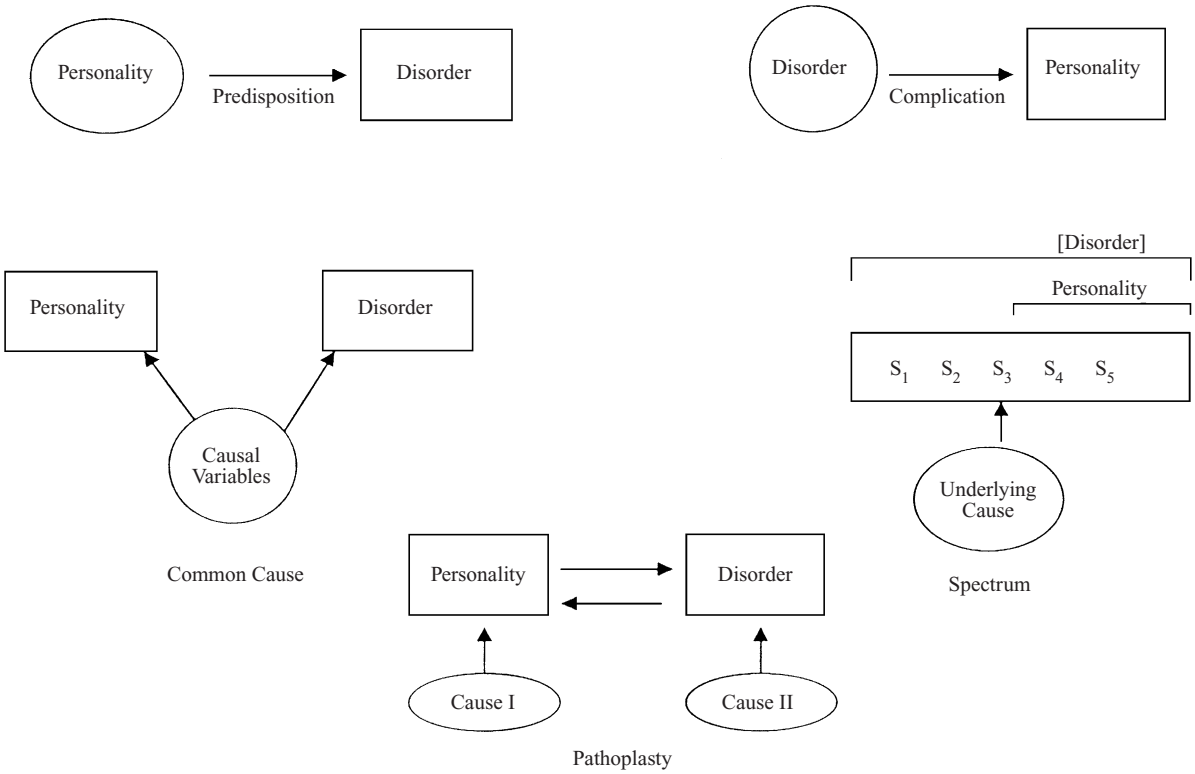

Figure 1
Models of relationship of personality and psychopathological disorders; S_1–S_5 = symptoms

personality and psychopathology are considered to be quantitative variations of the same underlying process. For example, it could be hypothesized that schizotypal personality traits and schizophrenia each are 'caused,' in part, by a particular genetic pattern. Here, schizotypal personality is not thought to cause, or be caused, by schizophrenia. Rather, it reflects variations of the same fundamental disturbance. This 'spectrum disorder' may fluctuate from periods of schizotypal personality to more active schizophrenic episodes. Common cause models are similar to spectrum approaches in specifying a common causal variable, but differ because they suggest that personality traits and specific forms of psychopathology are independent entities and not quantitative variations of one particular disorder or psychopathology. For example, obsessive personality traits and certain forms of depression may be posited to be independent entities which both arise from the same early experiences in a family, but they would not be considered variations of the same disorder. Finally, the *pathoplasty hypothesis* refers to the effect that personality may have on the clinical presentation and course of a particular form of psychopathology. This model does not stipulate any specific causal significance to personality, but simply implies that variations in personality will affect the psychopathology *after* it emerges. For example, an individual with high degrees of the personality trait of extroversion may behave differently when depressed than an individual with very low degrees of extroversion (i.e., introversion). Here personality 'colors' the way the disorder is manifest.

Unfortunately, there is a general lack of empirical data to clarify which of these models best characterizes the complex relationship between various forms of psychopathology and a myriad of personality traits. However, it is likely that specific personality traits will show different types of relationships to different forms of psychopathology and furthermore that a given personality trait may relate to psychopathology in a variety of ways. For example, in the study of neuroticism and depression there is evidence to suggest that neuroticism predisposes an individual to develop depression (predisposition hypothesis), predicts the course of depression (pathoplasty hypothesis), and may be exacerbated by the presence of depression (complication hypothesis). Clearly, the study of the relationship between personality and psychopathology is in its infancy and additional empirical studies are needed.

6. Methodological Considerations

Disentangling the relationship between specific personality traits and various forms of psychopathology will require complex research designs and statistical procedures. Clearly, longitudinal studies of individuals will be necessary to examine which facets of personality increase the risk of which disorders. Such research will also clarify which personality dimensions will reflect complications of specific psychopathologies. Additionally, family history studies and behavior genetic designs employing twin studies or adoption paradigms may help to determine more precisely how various forms of personality and psychopathology are transmitted in families and the extent to which they represent variations of the same process.

7. Summary

The study of personality has revealed that in spite of the power of environmental situations to determine behavior, people tend to be relatively consistent across time and situations in terms of their general patterns of behavior. Furthermore, it seems that people tend to differ along a few primary personality dimensions, or factors, which underlie the concept of personality. While personality traits may increase the risk of developing certain disorders or illnesses, the experience of an illness or disorder may also modify personality. Furthermore, people's personality styles will influence the way in which their disorders develop, change over time, and also respond to treatment. Powerful scientific studies which follow individuals over time will ultimately help us to clarify how personality increases the risk of disorder and furthermore how disorder may modify personality.

See also: Personality Assessment; Personality Disorders; Personality Theories

Bibliography

Allport G W 1937 *Personality: A Psychological Interpretation*. Holt, New York

Columbia University Teachers College 1928 *Studies in the Nature of Character, Studies in Deceit*. Macmillan, New York, Vol. 1

Ekehammer B 1974 Interactionism in personality from a historical perspective. *Psychological Bulletin* **81**: 1026–48

Epstein S 1979 The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology* **37**: 1097–126

Epstein S, O'Brien E J 1985 The person–situation debate in historical and current perspective. *Psychological Bulletin* **98**: 513–37

Helmes E, Reddon J R 1993 A perspective on developments in assessing psychopathology: A critical review of the MMPI and MMPI-2. *Psychological Bulletin* **113**: 453–71

Klein M H, Wonderlich S A, Shea T 1993 Models of relationships between personality and depression: toward a framework for theory and research. In: Klein M H, Kupfer D J, Shea M T (eds.) *Personality and Depression*. Guilford Press, New York pp. 1–54

Livesley W J, Schroeder M L, Jackson D N, Jang K L 1994 Categorical distinctions in the study of personality disorder: implications for classification. *Journal of Abnormal Psychology* **103**: 6–17

Lyons M J, Tyrer P, Gunderson J, Tohen M 1997 Heuristic models of comorbidity of axis I and axis II disorders. *Journal of Personality Disorders* **11**: 260–9

Millon T 1996 *Disorders of Personality: DSM-IV and Beyond*, 2nd edn. Wiley, New York

Mischel W 1968 *Learning Theory and Personality Dynamics*. Ronald Press, New York

Pervin L A 1990 A brief history of modern personality theory. In: Pervin L A (ed.) *Handbook of Personality Theory and Research*. Guilford Press, New York, pp. 3–18

S. Wonderlich

Person-centered Psychotherapy

Client-centered therapy is an approach to psychotherapy based on trust in the self-directive capacities of the individual. In this respect, it contrasts with other therapeutic orientations where the therapist characteristically acts as an expert.

1. Historical Overview

Psychologist Carl R. Rogers first described this new approach to psychotherapy in a talk in 1940, not realizing the impact it was going to have. This was followed by a full-length book on psychotherapy containing a clearly stated theory of therapy together with a verbatim account of an eight-interview case. This made up approximately two-fifths of the book and was a pioneering method of case presentation at the time.

As client-centered theory and practice developed in the 1940s and 1950s at Ohio State University and the University of Chicago, a far-reaching body of research on a new hypothesis grew up: that if the therapist offered, and the client experienced, a particular kind of relationship characterized by genuineness, unconditional positive regard, and empathy, a self-directed process of growth would follow. Moving to the University of Wisconsin from 1957 to 1963, Rogers and his associates undertook a major research project which tested the client-centered hypothesis with schizophrenic patients. In 1964, he moved to La Jolla, California, using the approach in the United States and abroad, in small and large groups, school systems, workshops, and conflict resolution. The broader application of the principles of client-centered therapy

became known as the person-centered approach. Rogers died in February 1987. The movement he fathered but did not wish to dominate is carried on by a diverse and dedicated international community.

2. Basic Therapeutic Concepts

The following are the fundamental concepts of person-centered psychotherapy:

(a) An actualizing tendency which is present in every living organism, expressed in human beings as movement toward the realization of an individual's full potential.

(b) A formative tendency of movement toward greater order, complexity, and inter-relatedness that can be observed in stars, crystals, and microorganisms, as well as human beings.

(c) Trust that individuals and groups can set their own goals and monitor their progress toward these goals. Individuals are seen as capable of choosing their therapists and deciding on the frequency and length of therapy. Groups are trusted to develop processes that are right for them and to resolve conflicts within the group.

(d) Trust in the therapist's inner, intuitive self.

(e) The therapist-offered conditions of congruence, unconditional positive regard, and empathy:

(i) Congruence has to with the correspondence between the thoughts and behavior of the therapist, who is genuine and does not put up a professional front.

(ii) Unconditional positive regard, also identified as 'caring,' prizing,' and 'nonpossessive warmth,' is not dependent on specific attributes or behaviors of the client.

(iii) Empathy reflects an attitude of profound interest in the client's world of feelings and meanings, conveying appreciation and understanding of whatever the client wishes to share with the therapist.

(f) Self-concept, locus-of-evaluation, and experiencing are basic constructs which emerge from the client's own interaction with the world.

(i) The self-concept is made up of the person's perceptions and feelings about self. Self-regard or self-esteem is a major component of the self-concept.

(ii) Locus-of-evaluation refers to whether the person's values and standards depend on the judgments and expectations of others, or rely upon his or her own experience.

(iii) Experiencing has to do with whether the person, in interacting with the world, is open and flexible or rigid and guarded.

(g) The Internal Frame of Reference (IFR) is the perceptual field of the individual, the way the world appears, and the meanings attached to experience and feeling. It is the belief of person-centered therapists that the IFR provides the fullest understanding of why people behave as they do, superior to external judgments of behavior, attitudes, and personality.

3. Theory of Psychotherapy

The basic theory of person-centered therapy is that if therapists offer, and clients experience, a particular kind of relationship characterized by genuineness, unconditional positive regard, and empathy, they will respond with greater self-direction and self-expression, an increased openness to inner and outer experiencing, more mature behavior and ability to deal with stress, and a concept of self that is more positive and more realistic.

4. Evolution of Theory and Practice

After Rogers assumed a professorship at Ohio State University at the beginning of 1940, he received an invitation to address Psi Chi, the psychological honor society, at the University of Minnesota, on December 11, 1940, the date often cited as the time when client-centered therapy was born. Rogers described a newer therapy which had the aim of helping individuals not only to solve their present problems but to grow in the capacity to solve future problems as well in a more integrated way, that took advantage of a general drive toward health, growth, and adjustment, that stressed emotional elements more than intellectual aspects, that emphasized the immediate situation rather than the past, and that viewed the therapeutic relationship itself as a growth experience.

The talk generated an intense reaction, both favorable and unfavorable. Rogers expanded his thinking into the book, *Counseling and Psychotherapy* (Rogers 1942), almost two-fifths of which was made up of 'The Case of Herbert Bryan,' consisting of the typescripts of eight phonographically recorded verbatim interviews. This kind of presentation was revolutionary, standing in sharp contrast to the subjective accounts of therapy being published at the time, and provided objective research data for the study of the therapeutic process.

Rogers and his students at Ohio State developed methods of classifying client statements and counselor responses, and of measuring self-regarding attitudes. The concept of self emerged as a central construct of personality organization, and great progress was made in the objective study of personality change and the behavior in therapy of therapist and client.

Research on person-centered psychotherapy advanced in a major way at the University of Chicago Counseling Center during Rogers' tenure there from 1945 to 1957. An entire issue of the *Journal of Consulting Psychology* in 1949 was devoted to a report of the 'parallel studies' project, comprising six investigations of the same group of 10 completely recorded and transcribed cases, with pre- and post-tests. The studies included measures of feelings regarding self and others, the pattern of client content (e.g., statement of problem, understanding or insight, discussion of plans), defensiveness, and maturity of behavior.

There was an evaluation of outcome using the Rorschach test, an analysis of the relationships among all these, and counselor ratings of improvement on a 1 to 9 scale. Two key findings were that (a) the measures applied to each interview provided a meaningful picture of the client's adjustment at the beginning, during, and end of therapy, and (b) there was a significant relationship between counselor estimates of success and ratings based on the interview-analysis measures. Significant relationships were not found between Rorschach results and those of the five interview measures. It was also concluded that the research methodology was applicable to therapeutic approaches generally, and that such efforts would help put psychotherapy on a scientific basis.

The 'parallel studies' project was succeeded five years later by a larger study (Rogers and Dymond 1954), which contained complete data on 29 clients seen by 16 therapists, as well as on a matched control group. It measured changes in self using the Q-technique of British psychologist William Stephenson, a sophisticated method of quantifying the way people viewed themselves at present and ideally, leading to the findings that the self-concept of clients in therapy improved significantly, and to a significantly greater degree than control group subjects, and that therapy achieved a significant increase in congruence between self and ideal. Qualitative analysis showed clients in therapy growing in their feelings of self-confidence, self-reliance, self-understanding, inner comfort, and comfortable relationships with others, with a decrease in negative feelings about themselves.

Person-centered and general psychotherapeutic theory advanced substantially with the introduction of the concept of 'necessary and sufficient conditions of therapeutic personality change' (Rogers 1957). Empathy, congruence, and unconditional positive regard were clearly defined. Capable of quantification, they stimulated hundreds of research projects. The Relationship Inventory (Barrett-Lennard 1998), an instrument measuring these conditions, has also been used in a vast number of research projects in psychotherapy and other human relations applications such as parent–child, student–teacher, and worker–employer relationships (see *Therapist–Patient Relationship*).

The most rigorous exposition of person-centered thinking, 'A theory of therapy, personality, and interpersonal relationships,' was published in Sigmund Koch's *Psychology*: *A Study of a Science* (Rogers 1959). A pioneering research project on therapy with schizophrenic patients and normal adults at the University of Wisconsin broke new ground in working with and doing research on this population (Rogers et al. 1967).

On Becoming a Person (Rogers 1961) included a summary of the personality and behavioral changes in therapy supported by research, an objective description of the fully-functioning person, the empirical evidence for the conditions which facilitate psychological growth, and a description of the author's struggle to reconcile the role of scientific investigator with that of therapist operating at the height of personal subjectivity. A unique, decades-long, research effort in person-centered psychotherapy which supports Rogers' description of the fully-functioning person focuses on the understanding of personality integration, which involves physiological, perceptual, cognitive, and interpersonal subsystems (Seeman 1983). Seeman has gone outside of the usual person-centered sources, drawing on the work of analysts Erik Erikson and Heinz Hartmann, and ego development specialist Jane Loevinger, among others. His conclusions are similar to Rogers, that high-functioning persons are healthier, more efficient in their perception of reality, have superior environmental contact, high self-esteem, confidence and trust in themselves, and possess a sense of autonomy that facilitates the development of caring and generative relationships.

5. *Evaluation*

Orlinsky and Howard (1978) concluded that the results of relevant studies were too variable to support the Rogerian hypothesis. But they were impressed by the evidence of 13 outcome studies of patients' perceptions of therapist attributes such as nonpossessive warmth, positive regard, and acceptance; all 13 investigations yielded a significant positive correlation between such perceptions and good therapeutic outcome.

Additionally, in 1978, Orlinsky and Howard reviewed 15 studies of the relationship between outcome and patients' perceptions of their therapists as empathically understanding, and found such an association preponderantly positive, with only two or three presenting evidence to the contrary. Fourteen of 20 investigations yielded a significant positive association between therapist self-congruence and outcome, with six showing null and marginally mixed results. Nine of 10 studies focusing on the 'process' or 'experiencing' levels conceptualized by Rogers in 1957 found significant positive correlations between experiencing and good therapeutic outcome.

Orlinsky and Howard (1986) came to these conclusions about the relation of outcome in psychotherapy to aspects of therapeutic process which bear upon the conditions offered by person-centered therapists. (a) Role-investment, empathic resonance, mutual affirmation, and the overall quality of the relationship were consistently related to patient outcome in 50–80 percent of the large number of findings surveyed in this area. (b) Therapist genuineness, particularly as perceived by patients, was often but not consistently associated with better outcomes. (c) Patients' perceptions of therapist empathy were very consistently related to good outcome. (d) Therapists' warmth or acceptance toward their patients, especially

but not only as observed by patients, was quite consistently related to good outcome in therapy.

Lambert et al. (1986), analyzing scores of studies on therapy outcome, concluded that therapist personal factors such as trust, warmth, acceptance, and wisdom were crucial ingredients, even in the more technical therapies. Similar conclusions were reached by Patterson (1984) and more recently in an examination of psychotherapy outcome research studies by Bozarth (1998) with these conclusions: successful psychotherapy depends primarily on the therapist–client relationship and the client's internal and external resources; the type of therapy is largely unrelated to outcome; the training and experience of therapists are irrelevant to successful therapy; clients who receive therapy improve more than those who do not; the evidence is weak that there are specific treatments for particular disabilities; empathy, genuineness, and unconditional positive regard are the relationship variables that correlate most consistently with effective therapy.

Watson (1984), following a careful review of the relevant research, concludes that because of methodological inadequacies, there is insufficient evidence either to support or refute the effectiveness of the 'necessary and sufficient conditions.' (see *Psychological Treatment, Effectiveness of*).

6. Concluding Comments

Person-centered psychotherapy has been evaluated assiduously, internally and externally, since its formulation by Carl Rogers in 1940. Its concepts and methods continue to be controversial in a field that is dominated by orientations that advocate guidance by experts. Its uniqueness in the depth of belief in the self-directive capacity of individuals and groups assures its future as a vital alternative. It continues to thrive internationally as part of a broader person-centered approach with implications for group process, education, and conflict resolution. Contributing to the health of the movement is the diversity of interpretation and implementation of Rogerian principles; it includes 'purists' and those who believe that a directive dimension is sometimes required. Different organizations and conferences afford the opportunity to explore these differences.

See also: Clinical Psychology in Europe, History of; Clinical Psychology in North America, History of; Interpersonal Psychotherapy; Multimodal Therapy in Clinical Psychology; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy, History of: Psychiatric Aspects; Psychotherapy Integration; Therapist–Patient Relationship

Bibliography

Barrett-Lennard G T 1998 *Carl Rogers' Helping System*. Sage Publications, London

Bozarth J 1998 *Person-centered Therapy: A Revolutionary Paradigm*. PCCS Books, Ross-on-Wye, UK

Kirschenbaum H, Henderson V L 1989 *The Carl Rogers Reader*. Houghton Mifflin, Boston

Lambert M L, Shapiro D A, Bergin A E 1986 The effectiveness of psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York

Orlinsky D E, Howard K L 1978 The relation of process to outcome in psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 2nd edn. Wiley, New York

Orlinsky D E, Howard K L 1986 Process and outcome in psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York

Patterson C H 1984 Empathy, warmth, and genuineness in psychotherapy: A review of reviews. *Psychotherapy* **21**: 431–8

Rogers C R 1942 *Counseling and Psychotherapy*. Houghton Mifflin, Boston

Rogers C R 1957 The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology* **21**(2): 95–103

Rogers C R 1959 A theory of therapy, personality, and interpersonal relationships, as developed in the client-centered framework. In: Koch S (ed.) *Psychology. A Study of Science. Vol. III: Formulations of the Person and the Social Context*. McGraw-Hill, New York, pp. 184–256

Rogers C R 1961 *On Becoming a Person*. Houghton Mifflin, Boston

Rogers C R 1980 *A Way of Being*. Houghton Mifflin, Boston

Rogers C R, Dymond R F (eds.) 1954 *Psychotherapy and Personality Change*. University of Chicago Press, Chicago

Rogers C R, Gendlin E T, Kiesler D J, Truax C B (eds.) 1967 *The Therapeutic Relationship and its Impact: A Study of Psychotherapy with Schizophrenics*. University of Wisconsin Press, Madison, WI

Seeman J 1983 *Personality Integration*. Human Sciences Press, New York

Watson N 1984 The empirical status of Rogers' hypotheses of the necessary and sufficient conditions for effective psychotherapy. In: Levant R E, Shlien J M (eds.) *Client-centered Therapy and the Person-centered Approach*. Praeger, New York

N. J. Raskin

Person-centered Research

Person-centered research here means research that focuses on the person as a functioning totality within the domain studied. The person then becomes the main conceptual unit and also often the main analytical unit. In many methodological realizations of this approach, individual patterns of values in the

variables under study become the main analytical units and subjected to, for instance, classification analysis or other types of pattern analyses. Person-centered research can be contrasted to variable-centered research where the focus is on the variable as the main conceptual and analytical unit. Variable-centered research is far more common but is not the topic of this article.

Person-centered research does not need to be quantitative and can in certain situations be carried out by case-oriented research and by using a qualitative approach. Obviously, the study of the single individual, studied quantitatively using the p-technique, is in one way person-centered, but is not normally so according to the definition given above (since the focus is usually then on relationships between variables within the individual). P-technique and time-series analysis is also discussed elsewhere in the encyclopedia. The perspective given in this article is of carrying out quantitative person-centered research on a sample of persons aiming at explaining and understanding inter-individual and/or intra-individual differences.

A very short historical overview is given of the emergence of person-centered research, followed by a presentation of the theoretical foundation of the modern person-approach. Finally an overview of common methods for carrying out person-centered research is presented.

1. *The Emergence of Person-centered Research*

A major type of person-centered research is the typological approach by which individuals are categorized into different types. It has ancient roots and is already reflected in the classical categorization of individuals into four basic temperaments: sanguine, phlegmatic, melancholic, and choleric. The typological view can be regarded as a reflection of a view that people are functioning wholes and that there are only a limited number of typical ways in which a functioning system can be organized. From this perspective, finding and understanding these types are important scientific goals. Creating typologies is also a reflection of a basic tendency in man of categorizing encountered phenomena.

The typological approach has been, and is still, stronger in the life sciences (especially taxonomy in biology and diagnosis in medicine) than it is in the social and behavioral sciences (for an historical overview, see Misiak and Sexton 1966). Gangestad and Snyder (1985) discussed the importance of a categorical approach in personality research and pointed to the emergence of types, each sharing a common influence (like an infectious disease caused by a specific germ). The typologies presented by Jung, Kretschmer, Sheldon, and others are still of interest today and psychiatric diagnoses in the Kraepelin tradition are influential in clinical practice. There is also a concern with typological research in sociology, including, for instance, the search for 'ideal types' (Bailey 1994).

The word 'typology' has many meanings—see Cattell 1957 for an overview of 45 different meanings of the word. It has lead to much confusion and even to resentment. This resentment was often directed against the connotation of 'innate' implied in many earlier typologies and against their subjectivity. Waller and Meehl (1998) make a number of thoughtful distinctions with regard to type and typology, preferring the use of the terms taxon for, roughly speaking, a meaningful type, and taxometrics as a generic name for (their) procedures for finding taxa. It should be pointed out that neither of these limitations apply to modern typological research within the person-approach described.

One of the first proponents of a comparatively modern person-centered research strategy was William Stern who already in the beginning of the twentieth century discussed an approach in which individual patterns in many traits were the units of analysis. Other early proponents of the importance of considering persons as functioning wholes were Kurt Lewin and Gordon Allport. A systematic approach to the person-centered study of personality development was undertaken by Jack Block based on longitudinal data. He presented an empirically based typology of longitudinal personality types and was, as far as we know, the first one to use the term 'person-approach' (Block 1971). Examples of research looking for basic categories are given by the work of Lars Bergman and David Magnusson studying stability and change in patterns of extrinsic adjustment problems, by Sigrid Gustafson studying a psychopathy-linked pattern called aberrant self-promotion, by Lea Pulkkinen, studying personality styles in a developmental perspective, by Richard Robins, Oliver Johan and Avshalom Caspi. searching for personality types, and by Ed Seidman relating typical experiential neighborhood profiles to antisocial behavior.

Perhaps the most forceful modern development of person-centered research has taken place within the new developmental science (Cairns et al. 1996). There, a holistic-interactionistic paradigm has grown strong and a new type of person-centered research has emerged (Magnusson 1999). It is focused on the study of individual development and has been called 'person-approach' (for an overview, see Magnusson 1998). Its theoretical and research strategic fundaments are described in Sect. 3.

2. *Theoretical and Research Strategic Fundaments of the Person-approach*

The person-approach refers to a holistic view on individual functioning and development. Magnusson and Allen (1983) summarized the essence of a person-

approach in the following way: 'The person oriented approach to research (in contrast to the variable centered approach) takes a holistic and dynamic view; the person is conceptualized as an integrated totality rather than as a summation of variables' (p. 372). A basic proposition of a holistic perspective, with consequences for the application of a person-approach, is that individual functioning and development can be described in terms of dynamic, complex, and adaptive processes in which mental, biological, and behavioral factors in the individual, and social, cultural, and physical factors in the environment are involved.

The person-approach is equally applicable to research on the current functioning of individuals and to research on individual development. Thus, it has implications for research in any field of psychological inquiry. Hitherto it has mainly been used in developmental and personality research.

For a long time the holistic view was considered to be too vague and lacking of specific content to serve as the basis for strong theories for understanding and explaining individual functioning and development. However, since 1970 the holistic model has been filled with content to the extent that it now forms a scientific basis for planning, implementation, and interpretation of empirical studies on specific aspects of individual functioning and development.

Substantively, contributions come from psychological research on mental factors and behavior and from biological and medical research on the functioning of the brain and the role of physiological factors in the total functioning and development of individuals. New findings in these areas have helped to enrich the 'black box' in S-R models with substantive contents and have helped in closing the gap between different explanations of behavior in terms of mental, biological, and environmental factors.

Contributions to the effectiveness of a holistic model as a theoretical framework for empirical research on specific phenomena, also derive from the presentation of modern models for dynamic, complex processes, particularly the general systems theory. Three aspects of these models are important for the discussion here of a person-approach.

(a) From a holistic perspective, mental, behavioral and biological aspects of individual functioning, and social, cultural, and physical factors of the environment are incorporated into one integrated theoretical framework. Thus, the role of a single variable cannot be finally investigated and understood in isolation from its context.

(b) A central principle in dynamic, complex processes is the principle of dynamic interaction (continuously ongoing reciprocal influences) as contrasted to statistical interactions in data. The new models for dynamic, complex processes provide a theoretical framework for investigating and understanding the dynamic processes of interaction of operating factors within the individual, and the continuous reciprocal interaction between the individual and the environment in the person-environment system.

(c) For the relations among operating factors at the level of the individual, mutual dependencies characterized by dynamic interactions and non-linearities can be a characteristic feature. The same holds true for the interaction of a single individual with his/her environment. For instance, individuals' psychological and physiological stress reactions to increasing stimulation from the environment are often nonlinear. In fact, most common variable-oriented research methods are not suited for handling nonlinear relations and dynamic interactions. This is exemplified by the fact that many such methods use the correlation matrix as the data to be analyzed, a matrix which mainly reflects linear relations, not nonlinear relations and interactions.

The process of organization of current and developmental processes which takes its start at conception is guided by the principle of self-organization. Self-organizing ability is a characteristic of open systems and refers to a process by which structures emerge without 'prescriptions from the outside.' Within subsystems, the operating components organize themselves to maximize the functioning of that subsystem with respect to its purpose in the total system. At a higher level subsystems organize themselves in order to fulfill their role in the functioning of the totality. Within a specific system, say the cardiovascular system, each of the operating factors (e.g., systolic blood pressure, diastolic blood pressure, and heart rate) do not function and develop independently of the others. The specific role of each operating factor is determined by the role it plays in the system. The operating factors are organized and function in terms of functional configurations, in what will be referred to in the following text as patterns. From this perspective, the important individual differences are not to be found in differences in any single variable taken out of its context of other, simultaneously operating variables. Instead, they are to be found in differences in the patterning of operating variables in the system under investigation. This applies to all levels of organization.

For the discussion of the implications of the person-approach for an effective measurement model and methodological approach, two aspects of the organization of mental, biological, and behavioral structures and processes are fundamental:

(a) Within subsystems individuals differ to some extent in the way in which operational factors are organized and function.

(b) Only a limited number of states are functional for each subsystem and for the totality.

The number of ways is restricted in which operating factors in a certain subsystem can be organized in patterns, in order to allow the subsystem to play its functional role in the totality, and the number of ways is also restricted in which subsystems can be organized

to form the total pattern for the total organism (cf. Bergman and Magnusson 1997). This view implies, among other things, that the studied complex systems have inherent restrictions which lead to (a) certain states being frequent or typical and (b) other states not occurring. The first aspect has been much studied and provides a motivation for the search for types. The second aspect has been much less studied but can also be of importance: What does not occur but, in principle, could occur contains information about how the system under study functions (cf. the concept of 'white spots' as discussed by Bergman and Magnusson 1997).

As emphasized above, the person-approach is a theoretical perspective. As such it forms the theoretical framework for the planning, implementation, and interpretation of specific substantive issues. This implies, among other things, that it has to be distinguished from the methods applied for treatment of data within the perspective of a person-approach. However, in most cases, pattern-based methods emerge as the most natural method choices. A brief overview of a selection of such methods is given in Sect. 4.

3. *A Selection of Methods for Carrying out Person-centered Research*

3.1 *Classification and Cluster Analysis*

It was mentioned in the introductory section that classification is a basic approach for carrying out person-centered research. Before discussing different types of classification analysis a few basic properties of such an analysis should be mentioned. In most forms of classification analysis the basic data are contained in the similarity or dissimilarity matrix between all pairs of subjects. Similar subjects are then grouped together in the classification analysis. The (dis)similarity between each pair of subjects could be a subjective rating but is most commonly calculated according to a formula that takes into account the (dis)similarity of the profiles of values in the variables for the two subjects. One common formula that takes into account differences in both level and form of the value profiles is the averaged squared Euclidean distance. It should be pointed out that the choice of (dis)similarity coefficient usually has important consequences for the results obtained in the classification analysis and that this choice should always be made on the basis of considerations in the specific case. The measurement characteristics of the different variables must be considered since the results of many methods are not invariant under linear transformations of the involved variables.

A classification can be achieved in diferent ways:

(a) It can be strictly theoretically derived as in the construction of ideal types.

(b) It can be totally empirically driven as in cluster analysis where subjects are classified together in clusters on the basis of their profile similarity. This type of classification approach is the most common one and is described.

(c) Some kind of model-based quantitative approach can also be used.

Examples of (c) are latent structure analysis and latent class analysis, originally proposed by Paul Lazarsfeld, where a number of latent classes is assumed and then the fit to the empirical data is tested and parameters are estimated (Goodman 1974). If the model holds, all relationships between variables within a latent class should disappear (the assumption of local independence).

3.1.1 Cluster analysis. In cluster analysis, a large number of methods are available for classifying objects on the basis of their (dis)similarities. Major types of cluster analysis are hierarchical methods (agglomerative or divisive), partitioning methods, and methods that allow overlapping clusters. Within each type of methods a variety of specific methods and algorithms exist. Perhaps the most common form of analysis is the agglomerative hierarchical cluster analysis. This group of methods starts with each of the n subjects being its own cluster. In Step 1 the two most similar subjects are joined to form one cluster giving in all n-1 clusters. In Step 2 the two most similar clusters are joined to form one cluster, giving in all n-2 clusters. The process is repeated until every subject is in one and the same cluster that occurs at Step n-1. The result is a hierarchical classification tree. Often the researcher concentrates on one cluster solution with a fairly small number of clusters and each cluster is described by its profile of means in the variables studied (called the cluster centroid). All members in that cluster should have profiles similar to the centroid. Different specific methods of hierarchical agglomerative cluster analysis have different rules for how to decide which two clusters are most similar. For instance, in the single linkage (= nearest neighbor) method the similarity between two clusters is given by the (dis)similarity of the two subjects, one from each of the two clusters, that are most similar. What method of cluster analysis is most appropriate of course, depends on the specific situation. Evaluations of the sensitivity of different clustering algorithms to the effects of errors of measurement and of the ability to recover a known cluster structure indicate that, as expected, no method appears to be generally superior to the others. Methods that often see to perform well include Ward's minimum variance method and average linkage cluster analysis (two hierarchical methods), and k-means relocation analysis based on a reasonable start classification (Morey et al. 1983). More recently, methods based on so called Beta-

flexible clustering have been suggested. For an overview of methods for cluster analysis see Gordon 1981.

3.1.2 Should all subjects be classified? It has long been recognized that multivariate outliers may disturb the results of cluster analysis and it has been suggested that in some situations the coverage must be less than 100%, i.e., not everybody can be classified. Bergman (1988) drew attention to the fact that there often exist a small number of 'unique' individuals, not similar to any other subjects that should not be forced into a cluster. He presented a procedure, RESIDAN, for *a priori* identifying and analyzing separately what he calls a residue of unclassified subjects. In a subsequent residue analysis, rare or non-existent patterns are studied which may be of theoretical significance (cf. the discussion about 'white spots' in Sect. 3).

3.1.3 Classification in the study of individual development. Important and intriguing methodological problems for person-centered research occur when studying individual development. A variety of approaches are available.

(a) Directly analyzing longitudinal patterns of values in variable profiles have often proved disappointing. If only profile form is studied, a direct longitudinal classification might work better than otherwise, since the amount of information that has to be summarized by the cluster membership variable is drastically reduced. An example of the usefulness of a longitudinal classification strategy in this situation is given in Block's (1971) study of longitudinal personality types.

(b) The analysis of cross-sectional patterns followed by linking over time is simpler and also often more robust than most other methods. One standard method of this type is called Linking of Clusters after Removal of a Residue (LICUR, Bergman 1998). LICUR is suitable in situations where both form and level of the profile is considered to be of importance. It includes procedures for removing a residue before the cluster analyses that are performed at each age separately and procedures for deciding the number of clusters. The results of the cluster analyses are linked between adjoining ages by cross-tabulating the classifications obtained at the two ages and it is tested for over frequented and under frequented cells (i.e., cluster membership combinations occurring more or less frequently than expected by chance). LICUR is applicable in a large number of settings, for instance when different variables are measured at the different measurement occasions and during periods of dramatic developmental shifts. Model-based alternatives to LICUR are provided by longitudinal extensions of latent class analysis, for instance latent transition analysis developed by Collins and Wugalter (1992).

3.2 Some other Approaches for Person-centered Analyses

3.2.1 Identifying a set of typical classes. Sometimes the interest is not in a complete classification but rather in identifying homogenous subgroups of subjects ('dense points' in a multivariate space) believed to indicate system states that are in some way 'important.' It is then believed that many subjects have entered a stable 'functioning' system state with a characteristic value pattern and that there are only a limited number of such states. Frequent typical patterns are regarded as indicators of such optimal states. One may also be interested in recognizing patterns that *a priori* are considered as 'important.' For overviews of various techniques relating to these issues see Young and Fu (1986).

3.2.2 Studying all possible patterns using configural frequency analysis. Configural Frequency Analysis (CFA) is a set of methods for studying all possible value patterns for a set of studied variables. The involved variables have to be discrete and are often dichotomized to make the number of value patterns to be examined manageable. CFA was originally suggested by Gustav Lienert. Lienert and his coworkers have developed CFA in a number of ways and in Germany it has become a research tradition. For a basic introduction see Krauth and Lienert (1982). A more recent overview is given by von Eye (1990) who also has contributed to various newer developments of CFA. The idea of types in a psychiatric setting was elaborated by Joachim Krauth who pointed to the importance of identifying combinations of variable values that emerge above chance level. His line of reasoning relates to the discussion about optimal systems states and designs as being more frequent than other states and designs. For further information on CFA, see von Eye (1990).

3.2.3 Analyzing specific properties of patterns. Instead of considering the complete value patterns in a pattern analysis, specific aspects of the patterns can be highlighted. Two such analyses are the following:

(a) In some cases the maximum and minimum scores of the subject's profile of scores are at focus. The variable taking the maximum value and the variable taking the minimum value may be seen as the essential features of the profile with the other scores providing the background. Of course, this implies that the different variables are scaled in a way that allows for comparisons between variables.

(b) The degree to which a subject's profile is even or uneven can be studied in different ways and an obvious measure of profile scatter is its variance around the mean of the scores in all variables constituting the profile.

3.2.4 Abstracting group memberships from qualitative information. A good example of this approach is the work done by Singer et al. in linking life histories to mental health outcomes. Richly detailed descriptions of individual lives form the input data for a procedure for discerning generalizable features of aggregates of multiple lives.

4. Some Final Comments

We believe that the theoretical perspective included in the person-approach presented above provides person-centered research with a useful meta-theoretical framework. But to reiterate: a person-approach should not be confused with a methodological approach using pattern analysis although such methods often are natural within a holistic orientation.

Sometimes reality may not be continuous but may rather operate to produce more or less discrete types. Or put differently: only certain configurations of system states are in some way optimal and become stable and often observed. A discussion of this issue was given in Sect. 3. It can then be argued that in person-centered research, methods for pattern-based analysis are often more useful than standard variable-based methods. Many common variable-based methods do not handle interactions well and in many situations pattern-based methods are more naturally used for this purpose. In fact, taking interactions seriously, such an approach, focusing on variables, tends to become very complicated. Lee Cronbach even made the metaphor of entering a hall of mirrors when one pursues such a goal.

It has sometimes been claimed that results from many methods used in person-centered research, for instance cluster analysis, are untrustworthy. In order to evaluate the validity of results from such studies two aspects should be discussed separately: (a) the technical aspects of the methods applied and (b) the appropriateness of the application of these methods in different settings. There is nothing wrong technically with any of the major methods used to carry out person-centered analysis. The problem arises when a method is inappropriately applied. Inconsistent results are often caused by the use of a clustering procedure that does not match the combined requirements of the problem under study and the data available. Important considerations for obtaining useful results from a cluster analysis are: (a) That the variable profile under study adequately summarizes the information *Gestalt* of interest; (b) that the values in the different variables are comparable and an appropriate (dis)-similarity coefficient has been chosen; (c) that a sound clustering algorithm has been chosen; (d) that only variables are included in the value profile which have a reasonably high reliability and, if finding homogenous clusters is the focus of interest, the profile is constituted by only a limited number of variables; and (e) that in many cases not all subjects should be classified. A number of validation procedures are available for studying the results of a classification analysis.

Two issues that are sometimes confused in person-centered research are the question about identifying generic classes ('natural clusters') and the question about ascribing the subjects in the sample to the appropriate class. It is a general experience that the second purpose tends to be more difficult to achieve. If the purpose is to identify typical value profiles that frequently emerge in different settings, one way of validating these typical profiles is to compare centroids between different samples/split-halves. Those that replicate might be regarded as types. These typical profiles need not, of course, together comprise a total classificatory system that encompasses everybody in a specific sample. On the contrary, the usefulness of a partial typology should be stressed since it can often be a more realistic goal. For instance, methods for determining how well the (dis)similarity matrix is represented by the classification solution are sometimes erroneously used for evaluating to what extent the first purpose has been achieved. Such methods are mainly relevant when the study of individual class membership is at focus.

Applying variable-oriented methods it is often assumed that (the same) linear relations approximately hold over individuals and that the essential features of a multivariate data set are captured by, for instance, the correlation matrix. This enables the researcher to use modern powerful statistical methods to construct models of data that are testable within the confinements of these basic assumptions. However, as was pointed out in Sect. 3, in settings when these assumptions cannot be assumed to hold a person-centered approach emerges as an alternative. It is natural that the recognition of multivariate complexity and higher-order interactions that follows with this perspective also makes it extremely difficult to formulate a coherent testable model of the data. Of course, this does not mean that person-centered research needs to be explorative in its methods. Theoretical considerations will lead to expectations about, for instance, types and antitypes and about typical developmental streams which are testable using various methods.

Finally, during the last decade person-centered research has received increased attention. For this we believe there are sound motives, as explicated in the section about the person-approach. In some situations one can trace a disappointment in the meager outcome in the understanding of a studied process given by even sophisticated variable-oriented methods. Applied with good judgment, person-centered research may then offer a deeper insight in how a system works. This applies to systems at the level of the individual as well as at other levels.

See also: Classification: Conceptions in the Social Sciences; Configurational Analysis; Single-subject Designs: Methodology; Statistical Clustering; Time Series: General

Bibliography

Bailey K D 1994 *Typologies and Taxonomies*. Sage, New York

Bergman L R 1988 You can't classify all of the people all of the time. *Multivariate Behavioral Research* **23**: 425–41

Bergman L R 1998 A pattern-oriented approach to studying individual development: Snapshots and processes. In: Cairns R B, Bergman L R, Kagan J (eds.) *Methods and Models for Studying the Individual*. Sage, Thousand Oaks, CA, pp. 83–121

Bergman L R, Magnusson D 1997 A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology* **9**: 291–319

Block J 1971 *Lives Through Time*. Bancroft Books, Berkeley, CA

Cairns R B, Elder G H Jr, Costello E J 1996 *Developmental Science*. Cambridge University Press, Cambridge

Cattell R B 1957 *Personality and Motivation Structure and Measurement*. World Book, New York

Collins L M, Wugalter S E 1992 Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research* **27**: 131–57

Eye A von 1990 *Introduction to Configural Frequency Analysis. The Search for Types and Antitypes in Cross-Classifications*. Cambridge University Press, Cambridge, UK

Gangestad S, Snyder M 1985 To carve nature at its joints: On the existence of discrete classes in personality. *Psychological Review* **92**: 317–49

Goodman L A 1974 Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**: 215–31

Gordon A D 1981 *Classification: Methods for the Exploratory Analysis of Multivariate Data*. Chapman and Hall, London

Krauth J, Lienert G A 1982 Fundamentals and modifications of configural frequency analysis (CFA). *Interdisciplinaria* **3**, Issue 1

Magnusson D 1998 The logic and implications of a person approach. In: Cairns R B, Bergman L R, Kagan J (eds.) *Methods and Models for Studying the Individual*. Sage, Thousand Oaks, CA, pp. 33–63

Magnusson D 1999 Holistic interactionism—a perspective for research on personality development. In: Pervin L A, John O P (eds.) *Handbook of Personality*, 2nd edn. Guilford Press, New York, pp. 219–47

Magnusson D, Allen V L 1983 Implications and applications of an interactional perspective for human development. In: Magnusson D, Allen V L (eds.) *Human Development: An Interactional Perspective*. Academic Press, New York, pp. 369–87

Misiak H, Sexton V 1966 *History of Psychology*. Grune & Stratton, New York

Morey L C, Blashfield R K, Skinner H A 1983 A comparison of cluster analysis techniques within a sequential validation framework. *Multivariate Behavioral Research* **18**: 309–29

Waller N G, Meehl P E 1998 *Multivariate Taxometric Procedures. Distinguishing Types from Continua*. Sage, Thousand Oaks, CA

Young T Y, Fu K S 1986 *Handbook of Pattern Recognition and Image Processing*. Academic Press, Orlando, FL

L. R. Bergman and D. Magnusson

Personhood, Anthropology of

'The Anthropology of Personhood' compasses the definition and study of three conceptual terms: 'person,' 'self,' and 'individual.' It is the exploration of the nature of the identity of the individual actor and the relationship between that identity and the symbolic forms and material practices of different sociocultural milieux.

On neither the meaning nor the implication of the above conceptual terms, however, can significant disciplinary consensus be found; a number of writers have addressed definitional issues directly (Harris 1989), but large disagreements remain. The dissension is perhaps understandable when one considers that what is fundamentally at issue is the nature of individual consciousness, and its manifestation—emergence, development, and construal—in different times and places. What is the relationship between 'nature' and 'nurture' in this context: between consciousness as a biological given and as sociocultural construct; between consciousness as privately experienced and as publicly validated; between consciousness as uniquely embodied and as collectively elicited and shared?

The position taken in this article is that an anthropological study of personhood must be careful to distinguish between what may be termed 'individualism' and 'individuality': roughly, between the conventional conceptualization of individual actors in particular settings on the one hand and their personal experience on the other. Much of the conceptual difficulty which the discipline has experienced arises from a conflating of these two, a confusion which goes back to Durkheim.

1. Social Structure and System

1.1 The Durkheimian Individual

Durkheim conceived of human beings as *homo duplex*; on one side there was the biological and personal (comprising the individual body with its material senses and appetites) and on the other the social and moral (the conceptual and conscientious). The individual thus led a double existence: one rooted in the physical organism and one (morally, intellectually, spiritually superior) in a social organism of collectively and uniformly held ideas and practices. Between the two there was ongoing antagonism and tension, but through inculcation into a public language and culture, humankind was capable of rising above mean (animal) 'individuality' and becoming part of a collective conscience in which the (sacred) traditions of a society were enshrined. If individuals were indeed conscious of themselves as individuals, then this was

equally a product of their socialization in a collective conscience; 'individualism' was a social product like all moralities and all religions.

From Durkheim's (structuralist) descendants a collectivist narrative was elaborated which conceptually subsumed the individual actor within sociocultural contexts. Mauss (1985) took it upon himself to detail how experience was constituted by cultural categories, and thus how society exerted its force on the physiological individual. He outlined a purported evolution in individual consciousnesses which could be tied to the particular forms of social structuration. First, comes the tribal stage of *personnage*, where individuals are conceived of as ephemeral bearers of a fixed stock of names, roles, and souls in clan possession; having no existence independently of the clan, individuals possess no inner conscience. Next comes the Classical stage of *persona* where individuals are conceived of as independent and autonomous citizens of a state; they are responsible, legal persons but still they possess no inner life or identity beyond the civic. With the rise of Christianity comes the stage of *personne*; conceived of as indivisible and rational, possessing a conscience, indeed, a unique sacred soul, the individual now serves as the foundation of all political, economic, and legal institutions. Finally comes the peculiar Western stage of *moi*: the individual as a 'self,' with self-interestedness and self-knowledge, as validated by modern institutions of psychological science.

Beginning with the assumption that sociocultural holism represents the paradigmatic form of human consciousness and experience, Dumont (1986) set out to plot more precisely the origination and progression of the singular Western idea of the autonomous individual as bearer of supreme moral value by comparing it with the Oriental 'archetype.' The Christian *personne*, he suggests, is reminiscent of the Hindu figure of the 'world-renouncer,' someone who seeks ultimate truth by forgoing the world in favour of his own individual spiritual progress. In the world-renouncer one finds a Western-like individualist for whom society has lost its absolute reality, and who, despite the constraining interdependence Indian society ubiquitously imposes on its members, has become self-conscious and self-sufficient. The evolution of the Western individual and the 'substitution' of self for society as absolute value marks the triumph of a (Christian) religious ideology in which worldly renunciation becomes the collective norm. Christ's teaching that man possesses a soul which absolutely transcends the world of social institutions and powers (which is absolutely individual in relation to God and meets others' only in God) engenders a community of 'outworldly' individuals who meet on earth but have their hearts in heaven. Nonetheless, Dumont concludes, the evolution need not end there; in the wake of such movements as multiculturalism, nationalism, fascism, and Islamic fundamentalism, the cultural future of individualism is, to say the least, unpredictable.

Characterizing the above, 'Durkheimian,' line of thought, in short, is the idea that the individual actor of Western society is the result of a recent and exceptional historico-cultural development. The concept of the 'individual' and its moral and social significance is absent elsewhere, with no *a priori* differentiation being made between individual and role, between self and society. Learned in this paradigm, moreover, it is not surprising that in much anthropological reportage on the person one finds a denial of individualism and little discussion of individuality. Inasmuch as 'the individual' exists, it is as a particular manifestation of 'the person'; it is not individuals who are seen to be diverse so much as the working parts of the complex social systems of which they are components and conduits. Individuals, in short, become collective constructs within the contexts of specific cultural ideologies and social practices, their psyches emanations of certain pregiven and prestructured life-worlds of socialization.

For instance, Strathern (1988) describes Melanesian personhood not in terms of individuals—distinct actors with discrete emotions, awareness and agency—but 'dividuals': beings constituted by properties, goods, and substances as these are ongoingly exchanged with others. The Gahuku-Gama, Read (1955) elaborates, are conscious of themselves only as members of common categories of relations; rather than, say, 'friendships' between unique individuals there are only relationships between socially defined positions. Similarly, Myers (1986) records Australian Aboriginal 'emotions' not as pertaining to deep-psychological experiences of the individual but as cultural evaluations of people and circumstance which persons must make in order to live in fluxional social settings. Geertz (1973, pp. 360), finally, relativizes Western conceptions of the autonomous individual by emphasizing the collective role-playing fundamental to the Balinese *habitus* of the person. Culture, Geertz suggests, translates into symbolic 'control mechanisms' by which the breadth and indeterminateness of actors' potential social lives are programmatically reduced to the specificity and narrowness of their actual ones.

In sum, a Durkheim-inspired anthropology of personhood has eventuated in comparative enquiries into the diversity of sociocultural conceptions of the human individual, how these conceptions connect with other sociocultural 'institutions' (family, law, exchange), and how these different ways of conceptualizing determine differences in experience. 'Personhood,' on this view, amounts to 'those culturally constituted and socially conferred attributes, capacities and signs that mark a moral career and its jural entitlements in a particular society' (Poole 1998, p. 842). 'Selfhood' refers to a human being as a locus of experience, as this experience is culturally allocated

and defined—usually as relational, socially embedded, and sociocentric. 'Individual' compasses an actor's uniqueness to the extent that this latter is realized in certain collectively structured life-worlds.

1.2 The Non-Durkheimian Individual

There have been exceptions to the Durkheimian line, nevertheless: ethnographies and analyses which deny the priority (ontological, developmental, historical) of the societal in its causing and conditioning of individual experience. Tracing the 'origins of English individualism,' Macfarlane (1978) sets out specifically to refute those theorists who would see individualism as a recent socio-cultural development—whether epiphenomenal upon religious or Renaissance or Enlightenment or capitalistic or industrialistic thinking. For:

> the majority of ordinary people in England from at least the thirteenth century were rampant individualists, highly mobile both geographically and socially, economically 'rational,' market-oriented and acquisitive, ego-centred in kinship and social life (1978, p. 163).

In New Guinea, meanwhile, Burridge (1979) describes how most people are 'individuals' and 'persons' in different respects and at different times, where 'persons' may be understood as those who realise given sociocultural categories, prescribed by traditional morality and the social order, while 'individuals' are those who use their perception and intuition to create anew. If persons are products of current sociocultural conditions, then individuals exist in spite of them. Each 'spatially bounded organism' is able to switch between these two modalities, Burridge concludes, such individuality representing an existential imperative which pre-exists culture. This also accords with Wikan's (1990) reappraisal of the Balinese, where the institution of cultural norms of personal composure, grace and placidity do not obviate the individual emotional effort and will be recognized as necessary to effect and sustain their appearance.

An emphasis on individual actors, their personal experience and agency, similarly found expression in early anthropological theorists of consciousness (such as Gregory Bateson and Anthony Wallace), in the transactionalism with which Fredrik Barth (and also F. G. Bailey and Robert Paine) developed Leachian insights into instinctually 'interested' individuals, in the work of network analysts (such as John Barnes and Ulf Hannerz), and in the burgeoning of studies within interpretive anthropology more generally which focus on the constructions of experience made by particular individuals in interaction. These accord more with Weberian sentiments that collectivities must be approached solely as the resultants and modes of organization of the particular acts of individual persons, since it is they who remain the upper limit and the sole carrier of meaningful conduct.

Notwithstanding, structuralist and poststructuralist discourses still denigrate an anthropological sensitivity of this kind as 'methodological individualism': as an erroneous couching of explanation in terms of individuals' characteristics, their behaviors, and interests, such that the 'foundational' conditions of sociocultural reality are obscured. The center ground of anthropology, it may be true to say, remains the preserve of 'methodological collectivism,' positing social phenomena as determined by factors which bypass individual rationality, and hence envisaging cultural development quite independently of individual consciousness. There is a continuing insistence, in other words, that the distinction between the individual and the societal is specific to the West and must be collapsed in favor of the latter—or at least of 'social relations'—for anthropology persuasively to encounter cultural others.

On the other hand, there is a growing insistence that it is a peculiar ethnocentrism for anthropologists to fail to allow in the 'others' we study the personal complexity which we recognize in ourselves: as both individuals and persons, role-players and rebels, conventional and creative. The anthropology which has sought most deliberately to keep this truth in view may be termed 'existential' or 'phenomenological,' after the philosophical traditions associated with such writers as Kierkegaard, Nietzsche, Husserl, and Sartre (Rapport 1997). It sets out with very different tenets and intentions from the 'sociological' orientation towards social structures and systems predominant above, privileging instead individual existence and experience.

2. Existence and Experience

To become human is to become individual, Geertz (1973, p. 52) has admitted, adding that we become individual in an environment of sociocultural forms and patterns in whose terms we give form, order, direction, and point to our lives. The important question which this raises is the precise relationship between sociocultural forms and the individual lives lived by them. Becoming human and individual in an environment of common sociocultural forms need not eventuate in becoming 'the same' as others, or even necessarily comparable, nor becoming after a deterministic or directly influenced fashion. For in this becoming, energy, agency, intention, and interpretation can be seen to remain properties of self-conscious subjects; indeed, it is individual agents who are responsible for the creation, animation, deployment, and development of systems of forms in their usage of them.

An intrinsic dichotomy between individual and world is often eschewed as a peculiarity of 'Western' sociocultural milieux, as we have heard, and hence as methodologically and analytically inapplicable. This is hardly defensible. Human beings may be

socioculturally situated, but they are also, always, interpretively autonomous—responsible for a 'personalization' of the world—and inexorably and inevitably (by) themselves. It is not good enough simply to say that only Western culture valorizes the concept of the individual—as 'individualism'—and therefore only in Western society do individuals act distinctively—with 'individuality.' For, while the former might be said to describe a particular sociocultural form of behavior (the pursuit of self-distinguishment), the latter concerns a condition which is an human universal: by virtue of a unique consciousness, each of us perforce engages with the world as other and is possessed of distinct perspectives upon it.

Individualism, more precisely, pertains to a specific historico-cultural conceptualization of the person which includes such notions as: the ultimate value and dignity of human individuals, their moral and intellectual autonomy, their rationality and self-knowledge, spirituality, right to privacy, self-sovereignty, and self-development, and their voluntary contracting into a society, market, and polity. Individuality, by contrast, refers to that unique awareness, and awareness of awareness, which is the mark of human embodiment. It is the universal nature of existence that human beings engage with others by virtue of discrete sense-making apparatuses (nervous systems and brains); they possess discrete centers of perceptual activity in discrete bodies. Furthermore, the human body (and in particular the brain) gives rise to knowledge of the world, to a perspective upon the world, which is inherently individual: human beings come to know themselves within the world by way of cognitions and perceptions, thoughts, feelings, and imaginings, which are unique to them (Edelman 1992).

Not only is there an individuality intrinsic to human consciousness but also to human agency. For, it is in terms of his perceptions that an individual comes to act, each individual becoming a discrete centre of motivation and intention. Not only is the individual's being-in-the-world universally mediated by very particular interpretative prisms which distance him from it, then, but while intrinsically 'of the world,' the individual inexorably comes to know the world, and act towards it, as 'other.' Individuals experience and interpret (and experience themselves interpreting) and therefore they are.

An individuality of consciousness and agency is current whatever the provenance of individualism as a cultural norm. Individuality is the human *a priori*, the physical–psychical basis on which all human knowledge of the world, all creativity within it and all activity upon it, rests (including the creation and representation of sociocultural milieux). Individuality remains consequential, that is, whether or not individual consciousness is an item of collective discourse, whether or not individual reflection is publicly eschewed, and whether or not individual distinctiveness is disparaged through the institutionalization of a fund of common behaviours (concerning, for instance, notions of the person). The experience of self is essentially distinct from sociocultural representations of self (cf. Spiro 1993).

Finally, an existential anthropology of personhood can hope to 'decolonize' (Cohen 1994) the individual human subject both from overdetermining cultural conditions and overweening social institutions (discourse, language-game, collective representation, social relationship, *habitus*, praxis), and from their holistic and hegemonically minded social-scientific commentators. It can take into account both the individual agency which brings sociocultural milieux to life and also the common sociocultural forms and practices by which individuals coordinate their activities and world-views within these milieux. Retaining respect for individual cognitive processes it apprehends the ambiguous interface between convention and individuality, and portrays sociocultural milieux as comprising, constituted by, individual difference: human individuals making a diversity of meaningful worlds by way of a commonality of cultural forms. In this way, an existential appreciation is promised of the 'individuals' behind the 'persons,' and of that consciousness of 'self' by which the disjunction is mediated.

See also: Individual/Society: History of the Concept; Knowledge, Anthropology of; Person and Self: Philosophical Aspects; Phenomenology: Philosophical Aspects; Self: History of the Concept; Self-knowledge: Philosophical Aspects; Self: Philosophical Aspects

Bibliography

Burridge K 1979 *Someone, No One*. Princeton University Press, Princeton, NJ

Cohen A P 1994 *Self Consciousness*. Routledge, London

Dumont L 1986 *Essays on Individualism*. University of Chicago Press, Chicago

Edelman G M 1992 *Bright Air, Brilliant Fire*. Basic Books, New York

Geertz 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York

Harris G G 1989 Concepts of individual, self and person in description and analysis. *American Anthropologist* **91**: 599–612

Macfarlane A 1978 *The Origins of English Individualism*. Blackwell, Oxford, UK

Mauss M 1985 A category of the human mind: The notion of person, the notion of self. In: Carrithers M, Collins S, Lukes S (eds.) *The Category of the Person*. Cambridge University Press, Cambridge, UK

Myers F R 1986 *Pintupi Country, Pintupi Self*. Smithsonian Institution Press, Washington, DC

Poole F 1998 Socialization, enculturation and the development of personal identity. In: Ingold T (ed.) *Companion Encyclopedia of Anthropology*. Routledge, London

Rapport N 1997 *Transcendent Individual: Towards a Literary and LiberalAnthropology*. Routledge, London

Read K E 1955 Morality and the concept of the person among the Gahuku Gama. *Oceania* **25**(4): 233–82
Spiro M E 1993 Is the Western conception of the self peculiar within the context of the world cultures? *Ethos* **21**: 107–53
Strathern M 1988 *The Gender of the Gift*. University of California Press, Berkeley, CA
Wikan U 1990 *Managing Turbulent Hearts*. University of Chicago Press, Chicago

N. J. Rapport

Personnel Selection, Psychology of

The term personnel selection has a scientific as well as a practical meaning. Scientifically, personnel selection indicates the methodology of developing and evaluating psychological methods to assess task-related qualifications which permit predictions of occupational success, job-related classification, and application of means of personnel development, including models and tools of personnel decision making. Thus, it should be referred to as personnel selection research. In organizational practice, personnel selection means the endeavor of choosing those individuals from the group of applicants who best meet the demands of the job and the organization. Although all kinds of job-related consulting for individual occupational and organizational choice are served as well by the same psychological methods, in English terminology the term personnel selection generally is used for the whole bundle of research, development, and application for all purposes in this context. Following this, this article will discuss the most important aspects relevant to these issues.

1. Comparison of Person and Job

Occupational or job-related qualification, understood as the probability of success, basically means 'who is suitable and what for?' That is, the target groups or individuals are to be defined just as well as the target positions. The requirements of the tasks or positions are to be established, and the necessary characteristics of the job-holders have to be derived on that basis. For very specific requirements, diagnostic methods are available or will be developed, or adapted, in order to measure and to compare job-relevant traits or abilities. Besides performance criteria, other indicators of occupational success can be formulated as objectives of the diagnosis, namely job and work satisfaction, health and well-being, or the tenure in the organization. Consequently, in addition to task analysis, the potential for satisfaction of the job in question has to be determined and compared with the interests and needs of the individuals.

Aside from determining the present requirements, an attempt can be made to estimate what kinds of changes in the requirements are to be anticipated; furthermore, an undeterminable amount of change has to be expected. The required development potential of a person should correspond to the foreseeable changes and offer the prospect to be up to future developments of an uncertain nature. Predictions are difficult in this area; however, some characteristics turned out to be relevant to success in most occupational contexts. Based on these general findings, it is necessary to compare person and job on three different levels (Fig. 1).

Different methods of task and job analysis are available for the investigation of occupational requirements. For our context, the most important kinds of requirements—not always clearly separable—are formulated as trait requirements (e.g., abilities and interests), behavioral requirements (e.g., skills and habits), qualification requirements (e.g., knowledge and skills), or as outcome requirements (e.g., problem solving and quality standards). In most cases the assessment of requirements is conducted by observation or by means of interviews or questionnaires.

2. Job-relevant Characteristics

The required characteristics of the job candidate are investigated, as far as feasible and appropriate, as they apply to the job-specific requirements. However,

Occupation/job	**Individual**
Requirements	Abilities, skills, and knowledge
Satisfaction potential	Interests, needs, and values
Changes	Development potential and general success-relevant characteristics

Figure 1
Comparison of person and job on three different levels

several characteristics also proved to be relevant predictors of occupational success. Preceding all is general intelligence, for which meta-analytical methods (Hunter and Schmidt 1990) provided evidence which showed that it is a valid predictor of achievement for practically all occupational fields (the effect of intelligence on work performance is mainly due to the ability to acquire job knowledge). As opposed to this, global personality traits such as extraversion and openness to experience have not shown to possess general validity. However, restricted variance in applicant samples may be responsible for this finding, adding to the problem that personality inventories usually lack job orientation. Among the global personality characteristics, the trait 'conscientiousness' seems to be of general relevance. In a meta-analysis using only European samples, Salgado (1997) found emotional stability (neuroticism) also to be a valid predictor for all jobs and criteria. While validity coefficients for these general traits turned out to be rather low, some narrower and less well-researched personality constructs showed higher validities at least for several occupational groups. Among these traits are: achievement motivation, self-confidence, and dominance. The highest predictive validity for a noncognitive trait has been found for 'integrity' (Ones et al. 1993), while its construct validity is still a controversial issue. Approaches to determine the interaction of several characteristics—especially those belonging to different diagnostic taxonomies—are only in the beginning. The application of structural equation modeling and related statistical methods should lead to substantial progress (Schmitt and Chan 1998).

3. The Trimodal Approach to Personnel Selection

The distinction of three general methodical approaches is helpful for the classification of diagnostic instruments: the *trait approach*, the *simulation approach*, and the *biographical approach* (Schuler 2001). Each one of these approaches pursues a partially independent validation logic and corresponds to specific methods of personnel assessment. With the trait or construct approach, characteristics are measured which are assumed to be relatively stable, e.g., conscientiousness and verbal intelligence. In order to assess these traits as homogeneous constructs, psychological tests are being used as typical instruments. In this case, construct validity is at the core of successful measurement. The objective of the simulation approach is the assessment of behavior similar to that required at the workplace; the appropriate validation strategy is the one which aims at representativeness or *content validity*. A typical form of measurement is the *work sample*. The third diagnostic principle is the biographical approach, although it cannot always be sharply distinguished from the other approaches. Typical assessment methods are biographical questions which can be presented in paper-and-pencil instruments (questionnaires), or in verbal form during an interview. Currently not in use, but a potentially fruitful method is the ideographical analysis of biographies (Fig. 2). For practical personnel selection, this means that in most cases multiple procedures are superior to cover complex requirements situations.

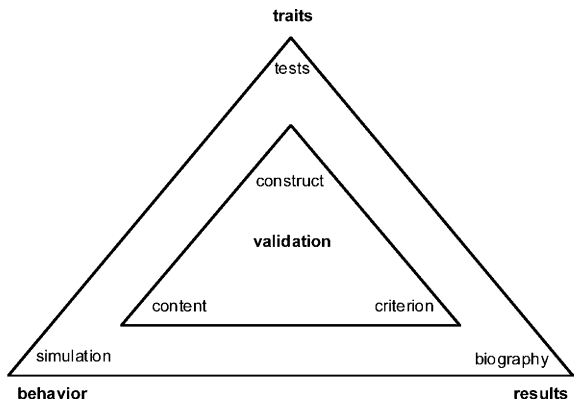

Figure 2
The trimodal approach to personnel selection

4. Instruments for Personnel Selection

First, the section discusses different types of single procedures, subsequently more complex procedures are presented (for a more detailed report, refer to Schmitt and Chan 1998, Schuler 2001).

4.1 Tests

Psychological tests are standardized routinely used methods for the measurement of individual behavioral characteristics, from which conclusions can be drawn pertaining to the individuals' traits or their behavior in other situations.

In scientifically controlled aptitude measurement, tests are the most commonly used instruments. They are applied to practically all occupational fields and job-relevant abilities. The most important kinds of tests used in personnel selection are:

(a) Tests of general cognitive ability (intelligence);
(b) Tests of specific cognitive abilities;
(c) Tests of attention and concentration;
(d) Tests of the sensory and motor abilities;
(e) Other achievement tests;
(f) General personality tests;
(g) Specific personality tests;
(h) Tests of attitudes, motivation, and interests.

Numerous tests have been published and are available from the test publishers. However, large organizations (such as the Federal Labor Agency, industrial companies, the military) often employ methods specifically designed for their purposes. The number of applications of these methods depend on the relevant occupational group. Figure 3 (data from 1990) shows that the extensiveness with which large companies use tests differs substantially between European countries.

4.2 Work Sample Tests

Work sample tests, being the prototype of simulation-oriented diagnostic tools, are standardized tasks which require assessees to carry out success-relevant occupational behavior. A high similarity of predictors and criteria is intended. The construction of work samples generally follows the same principles as that of psychological tests. A substantial difference is that it is largely done without 'translating' the occupational tasks into trait requirements. Rather than inferring a predisposition (trait) from a 'sign' (test behavior), conclusions are drawn from a sample of behavior for similar future behavior (Robertson and Kandola 1982). Examples are the preparation of a workpiece by a mechanic or a 'test lecture' by a teacher. A new class of work samples has been created by using video technology and virtual reality.

The use of computer-based procedures for psychological testing is rapidly growing (Booth 1998). The simulation of complex and dynamic problem solving tasks in scenarios requires the application of computers. These types of problems, originally developed in cognitive psychology, confront the applicant with a virtual environment where he/she has to control a system so that certain objectives will be reached. Hereby, the outcome variables (e.g., sales) as well as process variables (e.g., decisions), which are characteristic of a person's working behavior, can be observed.

There is also a special type of work samples that can be described as a kind of hidden intelligence tests, e.g., *job knowledge tests*. As construct validation research showed, cognitive ability is the primary determinant of knowledge acquisition. An extensive meta-analysis by Dye et al. (1993) proved job knowledge tests to be among the most powerful predictors of occupational success, especially in cases of high similarity between jobs and tests, and high job complexity.

4.3 Biographical Information

The basic principle of the biographical approach in personnel selection is the direct prediction of future behavior from past behavior. Application documents are being evaluated primarily with respect to that sort of information, although other (e.g., formal) aspects also require attention. Essential biographical elements are the curriculum vitae and references that contain information about job experience and achievements, as well as school and university grades. Methods of analysis to gain a maximum of information out of application documents have not been developed yet.

Prototypical for the biographical approach is the biographical questionnaire. These instruments are often difficult to distinguish from personality tests. In comparison to personality items, biographical items are generally related to the past, and they typically concern behavior and verifiable events. Classical principle of questionnaire construction is the validation of each item or even of each response alternative within a multiple-choice item with respect to external criteria. This results in a task-specific and often also organization-specific item selection—and as a consequence in limited generalization. For this reason, the 'blind empirical' approach is often combined with a rational approach which is requirement- and construct-related (Stokes et al. 1994). In most cases the most straightforward sources of performance-related biographical information are the level of education, school, and college grades. While the level of education is of high predictive value, school grades provide highly valid information for later educational performance, but the validity for predicting job performance is only moderate (Roth et al. 1996). Yet, in the 1980s and 1990s, predictive validity has even decreased for High School grades (Schuler 2001).

4.4 Interviews

Typical modes of employment interviews vary from a totally free conversation over partly structured to fully structured variants using standardized procedures, questions, and response evaluation. Ensuring interview validity can best be achieved by constructing the interview tool closely adapted to job requirements, standardizing questions, providing examples of response evaluation on behaviorally anchored rating scales, and combining several types of questions. The construction of interview questions can follow the same psychometric principles as test construction. Among the more elaborated types of interviews are the *Situational Interview* (Latham et al. 1980) and the *Multimodal Interview* (Schuler and Funke 1989). By means of the latter, it could be demonstrated that both construct- and simulation-oriented principles of assessment can be realized in an interview, as well as the biographical principle. Moreover, a sequence of structured and unstructured interview components takes account of a candidate's preferences for selection instruments that allow behavioral control, i.e., situations of low structure.

4.5 Multiple Procedures

The purpose of multimodal assessment procedures is to take into account a heterogeneous requirement

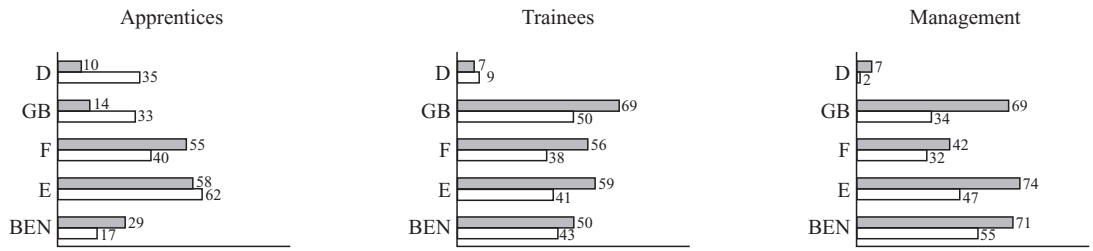

Note: Germany (D), Great Britain (GB), France (F), Spain (E), and the Benelux (BEN). Contrasted are the use of personality tests ▭ vs. ability and intelligence tests ▭ in percent of companies.

Figure 3
Application of psychological tests in several European countries

constellation by combining different methods and diagnostic principles. Basically, all types of single procedures can be combined; however, meta-analytical calculations show that noncognitive predictors can provide substantial supplements only in a few cases, i.e., incremental validity, in addition to intelligence tests. The most important of these predictors are work samples, specific personality tests, and structured interviews. As multiple selection procedures for supervisors and trainees, *assessment centers* are often used (Thornton 1992). They usually consist of work-samples such as group discussions, organizational tasks, presentations, and dyadic role-plays. Characteristic for an assessment center is that several applicants participate at the same time, and that the evaluation is carried out by several independent assessors. As far as the procedure succeeds to grasp the entirety of a person's potential for future development, we speak of *potential analysis*. Where reliable criteria are available, single components of multimodal procedures can be weighted for the calculation of a total score (e.g., by multiple regression).

5. *Evaluation of Selection Instruments and Procedures*

The evaluation of personnel selection procedures has to examine test theoretical criteria, especially objectivity, reliability, and validity. In addition to these, instruments have to be assessed according to organizational efficiency or advisory efficiency, to social validity as well as to further ethical and legal aspects.

5.1 *Validity*

Among the criteria of psychometric quality, validity is the most important, the others are necessary, but not sufficient conditions for validity. Among the variants of validity (or strategies of validation), *predictive validity* plays a crucial role, since the objective usually is the prediction of future occupational success. The following diagnostic procedures demonstrated good or sufficient validity (Salgado 1999, Schmidt and Hunter 1998): tests of general intelligence, work samples, job-related structured interviews, biographical questionnaires, some specific personality tests (integrity, conscientiousness, achievement motivation), and multimodal procedures (assessment center, potential analysis); tests of job knowledge and assessment of job performance are in a similar range. It must be noted that there are some moderating conditions (e.g., the validity of biographical questionnaires for young persons is low) and that validation leads to different coefficients for different measures of success—i.e., supervisory assessment, position level, income, assessment of potential. The aspect of *content validity* is important especially during the steps of instrument construction, that of *construct validity* in the examination of the psychological meaning of test scores.

5.2 *Organizational Efficiency*

Organizational efficiency is operationalized by aspects of practicability; these are, above all, the *expenditure* of the procedure, the required competence for its application, and its availability. Above all, *economic* utility is a core element of organizational efficiency. It depends on several parameters, especially validity, selection rate, and base rate as well as the variance in the performance criteria (that means, selection utility grows with the size of subsequent achievement differences). Supplemented by different economic indicators, the expected benefit can be calculated. Such calculations frequently lead to high utility estimates for the application of personnel selection instruments. *Advisory efficiency* may be understood as usefulness of diagnostic procedures to support individual occupational and organizational choice. Due to the relatively small general relationship between interests and abilities, such benefit is usually assumed. Advisory efficiency is important primarily in job counseling at labor agencies, but also in the context of organizational personnel development.

5.3 *Applicants' Reactions*

The applicants' *acceptance* of the selection process and their reactions to personnel selection methods is considered an independent quality criterion. Procedures which are better accepted are distinguished by information about job requirements as well as by transparency, feedback, and the possibility to control the situation. These requirements are better fulfilled in interactive procedures (i.e., interview) than in paper-and-pencil-tests or in biographical documentation (i.e., school grades).

5.4 *Ethical and Legal Aspects*

Ethical, professional, as well as *legal* aspects of personnel selection have received increasing attention. Examples for concrete ethical problems are: use of undue diagnostic methods, negligence of job requirements, application of intended distress during selection procedures, unjust invasion of privacy, lack of confidentiality of personal data, conflict of interest in relation to the employers and the clients. In this context, endeavors have been intensified to organize selection processes as fair and socially acceptable procedures (Schuler et al. 1993).

The appropriate legal basis—in Germany it is the general individual law as well as the labor law—are subject to substantial international differences. In Europe, an adaptation process takes place at present.

6. *Concluding Comments*

The adequacy of the application of personnel selection instruments in specific cases depends on a number of factors. If we disregard all required differentiations, the value of diagnostic information for the most relevant occupational groups can be judged as is presented in Fig. 4, which gives a summary of all is discussed above, that is, validity, efficiency, social validity (acceptability and fairness) as well as ethical and legal aspects. We have to get used to the idea that all these aspects will be important for the use of personnel selection methods in the future.

As a result of innovative instruments of personnel selection being vividly required by practitioners, the chance for the development and application of new approaches to assessment has increased. However, our insight into the incremental validity of additional methods or the value of combined procedures has not improved in the same way. Once science has moved towards that direction, we could approach a project that at this point of time still has the character of a vision—to draw diagnostic information not only from single, specifically constructed procedures, but from a variety of human characteristics. Performance results may be combined with self-appraisals and other assessments, evidence on abilities may be drawn from preferences and personal values, and biographical incidents, in connection with physiological data and even characteristics of the physique or other tabooed indicators, may have their place in highly valid predictor combinations.

However, one prerequisite for such a multi-indicator diagnosis are models that go beyond the available multivariate statistics, although current structural equation models may represent one step into that direction. One possible attempt is to apply the principles of fuzzy logic to validity estimation of such combined conclusions. On the other hand, the perspective of the employee or, as we call it, the *social validity* of these procedures must not be overlooked by

Types of Employment / **Selection method**	**Apprentices**	**Blue Collar**	**Employees w/o supervisory functions**	**Trainees**	**Management**		
					Lower	Middle	Top
Application documents	+	±	+	+	+	+	+
Additional references	−	±	±	±	±	+	+
Structured, job-related interviews	+	±	+	+	+	+	+
Academic grades	+	±	±	+	±	−	−
Biographical questionnaire	−	−	±	±	+	+	−
Specific, job-related personality tests	+	±	+	+	+	+	±
Achievement and knowledge tests	+	+	+	+	+	±	±
General and specific intelligence tests	+	±	±	±	±	−	−
Work samples	+	+	+	+	+	+	+
Assessment center/analysis of potential	±	−	±	+	+	+	±

Note: + appropriate; ± relatively appropriate; − inappropriate or not required

Figure 4
Adequacy of selection instruments for groups of employees

concentrating on such 'technical' innovations (Schuler et al. 1993). To keep the further development of diagnostic techniques in balance with these requirements will be an important challenge for personnel selection research in the future.

See also: Job Analysis and Work Roles, Psychology of; Personality Assessment

Bibliography

Booth J F 1998 Uses of PC technology in selection and assessment. *International Journal of Selection and Assessment* **6**: 57–60

Dye D A, Reck M, McDaniel M A 1993 The validity of job knowledge measures. *International Journal of Selection and Assessment* **1**: 153–7

Hunter J E, Schmidt F L 1990 *Methods of Meta-analysis*. Sage, Newbury Park, CA

Latham G P, Saari L M, Pursell E D, Campion M A 1980 The situational interview. *Journal of Applied Psychology* **65**: 422–7

Ones D S, Viswesvaran C, Schmidt F L 1993 Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance [Monograph]. *Journal of Applied Psychology* **78**: 679–704

Robertson I T, Kandola R S 1982 Work sample tests: Validity, adverse impact and applicant reaction. *Journal of Occupational Psychology* **55**: 171–83

Roth P L, BeVier C A, Switzer II F S, Schippmann J S 1996 Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology* **81**: 548–56

Salgado J F 1997 The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology* **82**: 30–43

Salgado J F 1999 Personnel selection methods. In: Cooper C L, Robertson I (eds.) *International Review of Industrial and Organizational Psychology*. Wiley, London, UK pp. 1–54

Schmidt F L, Hunter J E 1998 The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin* **2**: 262–74

Schmitt N, Chan D 1998 *Personnel Selection, A Theoretical Approach*. Sage, Thousand Oaks, CA

Schuler H 2001 *Lehrbuch der Personalpsychologie* [Personnel Psychology]. Hogrefe, Goettingen Country

Schuler H, Farr J L, Smith M (eds.) 1993 *Personnel Selection and Assessment: Individual and Organizational Perspectives*. Erlbaum, Hillsdale, NJ

Schuler H, Funke U 1989 The interview as a multimodal procedure. In: Eder R W, Ferris G R (eds.) *The Employment Interview—Theory, Research, and Practice*. Sage, Newbury Park, CA pp. 183–192

Stokes G, Mumford M, Owens E (eds.) 1994 *Biodata Handbook: Theory, Research and Use of Biographical Information in Selection and Performance Prediction*. Consulting Psychologists Press, Palo Alto, CA

Thornton G C 1992 *Assessment Centers in Human Resource Management*. Addison-Wesley, Reading, MA

H. Schuler

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Pes

Pestalozzi, Johann Heinrich (1746–1827)

1. *Introduction*

Johann Heinrich Pestalozzi was one of the most distinctive writers of his time, and was already regarded as one of the most famous pedagogs in Europe in his own lifetime. He owed his reputation as a writer essentially to his best-selling novel *Lienhard and Gertrud* (1781), and owed his fame as a pedagog to his work in Stans (1799), Burgdorf (1800–1803) and in Yverdon (1804–1825) where for 21 years he directed an institute in which children from all over Europe were educated.

Johann Heinrich Pestalozzi's work deals with a wealth of subjects in the fields of general politics, social policy, constitutional law, jurisprudence, literature, anthropology, philosophy, pedagogy, and social pedagogy, and it is documented in 30 volumes in the Critical Edition *Sämtliche Werke und Briefe*, 15 of which have hitherto appeared (also available on CD-ROM).

Pestalozzi was, in both his personal and literary life, actively opposed to social injustice, oppression, exploitation, dependence, and lack of political freedom. His ambition was to free the poor and their children from powerlessness, to help them to be able to secure their own individual subsistence, as well as to abolish the very political conditions which caused this hardship. He constantly gave central importance in his work to social justice, the satisfaction of needs, ownership and work, domestic order, and basic needs for all, including children. The subjects he addressed were rooted in social reality. Pestalozzi reflected upon and analyzed the everyday life of the poor masses, the everyday life of children, and the everyday political life of his time.

The special form of this relation to reality Michel Soetard (1981) has convincingly called the 'Principle Reality.'

2. *Life and Work*

2.1 *1746–1799: Childhood, Youth, and Early Career*

2.1.1 Family origins and childhood. Johann Heinrich Pestalozzi was born in Zurich on the 12 January 1746, the son of the surgeon Johann Baptist Pestalozzi and his wife Susanna, née Hotz. At that time, Zurich was a center of European culture, a city state with an exclusive merchant and entrepreneurial aristocracy. The town was regarded as rich, with a small number of old families holding power in the city council. Neither his father's family nor his mother's—who originally came from the Zurich area where her father was a parish priest—belonged to this group. The married couple produced seven children in 9 years of marriage from which only three grew to maturity. When Pestalozzi—one of seven children—was 5 years old, his father died at the age of 33, leaving his mother to take care of four small children. In order to lead a bourgeois life in spite of the financial predicament, economizing and making-do became a staple of their dismal everyday life. Pestalozzi had therefore to endure poverty. In the country at his grandfather's house in Höng, where his grandfather was a parish priest, he became acquainted early on with the miserable social condition of the provincial population which was to occupy him for a lifetime.

2.1.2 School and studies. In Zurich, Pestalozzi—who had the civic rights—grew up with the sons of the patrician families whose carefree material situation was permanently laid out before his eyes. He enjoyed the schooling of a city-dweller, and went to the city elementary school from 1751 to 1754; he was then a pupil from 1754 to 1763 at the Schola Abbatissana in Zurich, the Schola Carolina, and the Collegium humanitatis (also in Zurich), the latter being a Latin school. After some initial difficulties he was a good pupil, though because of his financial situation he was always an outsider because of his 'extraordinary carelessness and uncleanliness.' From 1763 to 1765 he studied at Zurich's theological college, the Carolinum. The appearance of an anonymous pamphlet he had co-written caused the abrupt end to his studies. It was clear to him in any case that he did not wish to become a priest.

In Zurich there were—following the example of other European countries—patriotic and physiocratic societies. From 1764, Pestalozzi belonged to the circle of patriots of the Helvetic Society. The Zurich scholars Bodmer and Breitinger in particular had their circle of influence here. The much-admired model was Rousseau. Republican works were read and written. It

was in this context that Pestalozzi's first publications *Agis* (1766) and *Wünsche* (1766) appeared. In 1767, Pestalozzi began a short agricultural apprenticeship with the Bern physiocrat Tschiffeli.

2.1.3 Livelihood. In February 1769, Pestalozzi became a farmer in Müllingen and ran into heavy debt. On 30 September 1769, he married *Anna Schulthess* (1738–1815), and on the 13 August 1770 their only child Hans Jakob, called Jaqueli (1770–1801), was born. Pestalozzi built a manor house, the 'Neuhof auf dem Birrfeld' which he moved into with his young family in 1771. Crop failure, bad purchases of land, and other mistakes led to failure and almost to financial ruin. In 1794, he founded an institution for the poor at the Neuhof. His initial interest was more entrepreneurial than pedagogical; in his *Neuhof-Schriften* (published in 1777), Pestalozzi accompanied his essays with a journalistic advertising campaign for support for his project. In his *Bitte an Menschenfreunde*, he successfully enlisted the support of, amongst others, Niklaus Emanuel Tscharner (1727–1794), the chairman of the Bern Society for Economics. Pestalozzi discovered his interest in all things pedagogical during his time at the Neuhof, motivated by his interaction with the poor children of the Neuhof: 'I saw how the overwhelming cruelty of self-interest constantly destroys the body and soul of almost—and I mean almost—all of these children.' Pestalozzi's institution failed financially, but he was later to make use of his experiences in the field of national education.

In the middle of the Neuhof crisis, the political work about Zurich *Freyheit meiner Vaterstatt* (1799)—which remained unpublished—was written in which he lamented the wealth of the city and its devastating effects on the constitution and political culture of Zurich.

2.2 From 1780 to 1799: Livelihood as an Independent Writer

Out of the necessity of having to finance himself, Pestalozzi became a writer. In 1780 he published *Abendstunde eines Einsiedlers* whose theme is cultivating people: 'Cultivating people towards the truth means cultivating their being and their nature towards a reassuring wisdom.' Then he wrote the first part of *Lienhard und Gertrud*, which appeared in 1781. *Lienhard und Gertrud* is a pedagogical novel which its author broadened into a novelistic pedagogical text. The theme of the novel is the story of the education and cultivation of an entire village. With *Lienhard und Gertrud*, Pestalozzi wrote about a fictitious village which he called Bonnal in a novel showing both the development and the education of the characters. The subject of the novel is bound together by a kaleidoscope of multifaceted and fictitious stories of education, cultivation, and socialization.

In the second (1783) and third (1785) books, it becomes clear how many individual fates have been embroiled in the deplorable state of affairs in Bonnal for many years, and that it is thereby not only a question of individual error, but at the same time of failings which lie more deeply in the political structure. Only in the fourth book (1787) does it become possible to continue telling the story of Bonnal and to organize it structurally as a story of reform, with, of course, further novelistic twists and turns and subplots. The village as the scene of the action is increasingly abandoned. Paragraph by paragraph Pestalozzi outlines his ideas for new legislation, the distribution of land, jurisdiction, and the culture of debate and justice. The novel is therefore increasingly interrupted on the level of action and plot, becomes more difficult to read, and thus loses its original novelistic character. The literary code transforms into a politically discursive text. *Lienhard und Gertrud* became Pestalozzi's greatest literary success. Pestalozzi was soon to become known far beyond the Swiss borders, first of all not as a pedagog, but as a writer. The novel brought him the title of Honorary Citizen during the French Revolution.

In *Gesetzgebung und Kindermord* (written in 1780 and printed in 1783), Pestalozzi declared the prevailing legal practice throughout Europe of executing child-murderesses—mostly after extreme humiliation and torture—as inhuman. He mounted an emotive defence of a large number of child-murderesses, whom he viewed as innocent. To him, poor girls and women had the fundamental right to sexuality and marriage, to their children, to protection and support, and to recognition and help. The work was a plea against capital punishment.

In 1782, Pestalozzi published *Christoph und Else*, a novel of instruction, even a 'novel' for the novel *Lienhard und Gertrud*. It was to be used as a guide to how adults should read *Lienhard und Gertrud*. A peasant family talk together over 30 evenings about the plot, the theme, and the constellation of characters in *Lienhard und Gertrud*. On 26 August 1792, Pestalozzi was appointed an Honorary French Citizen and never renounced the honorary title in spite of all the criticism of the Terror, an action which aroused suspicion in conservative circles. The outcome of Pestalozzi's examination of the revolution was the work *Ja oder Nein?* (1793) in which he made the Ancien Régime responsible for the devastating outrages in Europe and thus justified the revolution, though at the same time he sharply denounced the revolutionary Terror.

In 1797, Pestalozzi's main philosophical work *Meine Nachforschungen über den Gang der Natur in der Entwicklung des Menschengeschlechts* appeared. The initial question was 'What am I, and what is mankind?' Pestalozzi expounded a natural state, a social state,

and a moral state of humanity, and went on to discuss the individual human being as a 'work of himself,' a 'work of nature,' and a 'work of morality.'

In 1798, the revolution also reached Switzerland. For a short time, a Helvetic Republic was established. Pestalozzi became the editor of the 'Helvetisches Volksblatt'—the publication organ of the new Helvetic government—and at first worked on behalf of the government. He wrote texts, appeals, and pamphlets, as well as *Abhandlung über die Natur des Helvetischen Zehnden*. Whereas the typical pamphlets on both sides at the time of the revolution constituted obvious works of political agitation in favor of one side or the other, Pestalozzi was a mediator who urged the often very hostile opposing groups towards agreement. He became a soother of passions in his pamphlets. Pestalozzi wanted to promote the unity of the Swiss in the crisis year of 1798, the unity between the poor and rich population, the unity of town and country, and the unity of the Inner-Swiss (Catholic) cantons and the majority Protestant cantons.

2.3 1799–1827 The Pedagog and Institute Director: From Stans to Yverdon

2.3.1 Stans. At the age of 53, Pestalozzi received a government commission to run an orphanage in Stans. From 7 December 1798 to 8 June 1799, he lived there together with a housekeeper and as many as 80 children by the end. His pedagogical work and his pedagogical writing were now supported by two pillars. The first was a love of children and a passion for his work. Pestalozzi made himself into the 'father' and 'mother' of all the children; he tried to transpose the notion of the great importance of satisfying basic needs. The second was that the fundamental principle of the 'method'—which was on the one hand an almost technocratic syllabic reading method—was at the same time elevated by the idea of an elementary method of moral education. From then on, Pestalozzi's elementary education and elementary method were central themes. The institute, however, came to a sudden end. Pestalozzi had to hand over the building to French soldiers returning from the Russian campaign for use as a hospital and leave the institute, an action which made him fall into deep depression and despair. During this time, he wrote his strongly influential work, the Stanser Brief (1799), which contained a passionate plea for the malleability of human beings, including, and indeed especially, for the malleability of poor, neglected children.

2.3.2 Burgdorf. Pestalozzi became a teacher in Burgdorf on 23 July 1799. The announcement of a schoolmasters' conference in Burgdorf Palace followed on the 24 October 1800. The appearance of *Wie Gertrud ihre Kinder lehrt* followed in 1801. In the first part, Pestalozzi reported on Stans in autobiographical form, on his first endeavors in Burgdorf, on his young colleagues, some of whom accompanied him for a considerably long time, and on how he conceived method, the unity of knowledge, ability, and feeling. Pestalozzi developed a research program which he continued in Yverdon and which aimed at developing suitable learning aids in the three directions which he soon named 'HeadHeartHand.' The principles of the new teaching method consisted of the ability to visualize, simplification, logical progression, and individualization. The method was aimed at developing the autonomy of the child, at creating an independent personality which could secure its own subsistence. From November 1802 to February 1803, Pestalozzi traveled to Paris as a member of the Consulta. He was, however, unable to push through his concept for national education there. In June 1804, the institute at Burgdorf was moved to Münchbuchsee, and at the end of August the opening of a subsidiary institute followed in Yverdon, a town in French-speaking Switzerland, and on the 6 July 1805, Münchbuchsee was closed.

From 1804 to 1825, Pestalozzi ran one of the largest European educational institutes of the time in the rooms of the palace at Yverdon in French Switzerland. Yverdon was an experimental school on the level of its teaching, education, and teacher training. In order to be able to disseminate the new knowledge acquired from the everyday pedagogical research, a printing works was set up specifically for this purpose. Pestalozzi, who until that time had always written his works alone, now published journals, teaching books, teaching material, discourses on teaching methods, appeals, and speeches with his colleagues. At the center of these works stood Pestalozzi's concept of an elementary education of people which could be applied to different teaching disciplines.

At the height of its impact in 1809, the institute had 150 pupils from 6 to 16 years old, 31 teachers and assistant teachers, and 40 young people as observers from all over Europe who wished to study Pestalozzi's method. Pestalozzi's concept of elementary teaching was already being followed ever more closely in Prussia, Spain, and many other countries. Pestalozzi's most important work from the Yverdon period is *An die Unschuld, den Ernst und den Edelmut meines Zeitalters und meines Vaterlands* (1815), a treatise in the form of an address in which Pestalozzi set out his program for life: 'The beginning and the end of my policy is education.' Pestalozzi analyzed the approach of the industrial age with its phenomena of concentration, mechanization, and standardization. Only in education did he see the means to prepare people for the Industrial Age. The work caused a great sensation and was a literary success.

The size of the staff and the differences of interest and opinion within the teaching body increasingly led to endless quarrels which finally caused the downfall of the institute. Pestalozzi left Yverdon embittered in 1825, and returned to the Neuhof.

In the last 2 years of his life, he wrote the works *Meine Lebensschicksale als Vorsteher meiner Erziehungsinstitute in Burgdorf und Iferten* (1826), the *Schwanengesang* (1826), and finally the *Langenthaler Rede*. During the crisis which industrialization and the reorganization of the states into large administrative states gave rise to, Pestalozzi saw a great opportunity for man to return to himself, and for his inner renewal by means of a comprehensive elementary education.

After a short illness, Pestalozzi died on 17 February 1827. On 19 February, he was buried near the Neuhof in the small village of Birr.

3. *Critical Reception and Research Developments*

The research literature on Johann Heinrich Pestalozzi's life and work, amply documented by the research papers and synopses of the Pestalozzi literature of the past 200 years, at first glance appears immensely extensive. This lies in the fact that Pestalozzi, in the history of (German-speaking) pedagogy and educational sciences, has without doubt experienced a considerable *impact* history which has admittedly only to a certain degree coincided with the *reception* history of his numerous written works. Above all, Pestalozzi's teaching method is prescribed in many countries around the world and his best-known works have been translated into many languages. Should one follow the numerous publications up to the present, it becomes apparent that, whereas in the first half of the twentieth century the 'whole Pestalozzi' became the subject, individual or partial aspects began, however, to dominate more and more markedly within the framework of the differentiation of educational science and its adjacent sciences. In this way, the paradigm exchanges in educational science are mirrored in the respective cognition interests of analyses. If the Pestalozzi from the history of ideas took center stage in scientific literature in the middle of the twentieth century, the research controversy about the *political Pestalozzi* (Rang 1967) was sparked off in the 1960s and 1970s. In the course of the 1970s with its far-reaching education and school reform, interest was then concentrated rather more on the practitioner and reformer. Within the research history, it furthermore became clear that, after a very intensive, bibliographically broad and verifiable acceptance of Pestalozzi the politically reforming pedagog, the research in the 1980s stagnated; indeed the number of papers and studies about Pestalozzi fell off considerably. That only changed in 1996. Since then, the trend has been moving in the direction of an examination of the research into context and reception (the impact history) alongside biographical research using new scientific methods, aimed at producing a differentiated situating of the work as regards history and content. An important impetus comes from the work of the Pestalozzi Research Center which, dividing the work between the Pestalozzianum and the University of Zurich, is pushing forward with the Critical Edition, and publishes the *Neue Pestalozzi-Studien* and the *Neue Pestalozzi-Blätter*.

Pestalozzi's journalistic work is so diverse that it not only touches on areas of social and cultural practice as regards content, or even focuses on them explicitly as a central theme, but also represents groups of different discourse methods. In order to understand the work, interdisciplinary research is required e.g., into rhetoric, eloquence and discourse, the socioaesthetic culture of sentimentalism, social history, the relationship between the written and the spoken, the problematic nature of text selection, the relationship between truth and fiction, the theory of femininity and the ensuing mother cult of the eighteenth century.

See also: Education: Phenomena, Concepts, and Theories; Educational Institutions, History of; Educational Learning Theory; Educational Research and School Reform; Pedagogical Reform Movement, History of

Bibliography

Hager F P, Tröhler D 1996 *Pestalozzi—wirkungsgeschichtliche Aspekte. Dokumentationsband zum Pestalozzi-Symposium 1996*. Haupt, Bern/Stuttgart/Wien

Korte P 2001 *Pädagogisches Schreiben um 1800*. Der Status von Schriftkultur, Rhetorik und Poetik bei Johann Heinrich Pestalozzi

Kraft V 1966 *Pestalozzi oder das Pädagogische Selbst*. Eine Studie zur Psychoanalyse pädagogischen Denkens. Julius Klinkhardt, Bad Heilbrunn

Liedtke M 1992 *Johann Heinrich Pestalozzi mit Selbstzeugnissen und Bilddokumenten*. Rowohlt, Reinbek bei Hamburg

Pestalozzi J H 1946ff *Sämtliche Briefe*. Kritische Ausgabe. NZZ, Berlin/Zurich

Pestalozzi J H 1994 Sämtliche Werke und Briefe auf CD-ROM. Pestalozzianum, Zurich

Oelkers J, Osterwalder F 1995 *Pestalozzi—Umfeld und Rezeption. Studien zur Historisierung einer Legend*e. Beltz, Weinheim und Basel

Osterwalder F 1996 *Pestalozzi—ein pädagogischer Kult. Pestalozzis Wirkungsgeschichte in der Herausbildung der modernen Pädagogik*. Beltz, Weinheim und Basel

Rang A 1967 *Der politische Pestalozzi*. Frankfurt/Main

Soetard M 1981 *Pestalozzi ou la naissance de l'éducateur* (1746–1827). Peter Lang, Berne/Frankfurt/Las Vegas

Stadler P 1988 *Pestalozzi. Geschichtliche Biographie. Band 1. Von der alten Ordnung zur Revolution*. Verlag Neue Zürcher Zeitung, Zürich

Stadler P 1993 *Pestalozzi. Geschichtliche Biographie. Band 2. Von der Umwältzung zur Restauration*. Verlag Neue Zürcher Zeitung, Zürich

Tröhler D 1997ff *Neue Pestalozzi-Studien*. Haupt, Bern/Stuttgart/Wien

P. Korte

Phantom Limbs

Many patients awake from the anesthetic after an amputation feeling certain that the operation has not been performed. They feel the lost limb so vividly that only when they reach out to touch it, or peer under the bedsheets to see it, do they realize it has been cut off. This startling realization does nothing to subdue the reality of the limb they experience, and may even intensify the sensations that define it. The first detailed description of this phenomenon was provided by the American Civil War surgeon, Silas Weir Mitchell (1871), who dubbed it the 'phantom limb.' The term so accurately captures the nature and subjective experience of the phenomenon that it has been adopted ever since. This entry reviews the nature of phantom limb pain and nonpainful phantom limb sensations which arise after limb amputation.

1. Phantom Limb Pain

For many amputees, a distressing problem is phantom limb pain. While the distinction between a painful and nonpainful phantom is not always clear-cut, it is useful to separate them for purposes of description, classification, and treatment. Painful phantom experiences vary in a number of dimensions. These include intensity, quality, location, frequency, and duration. Many amputees report a painful intensification of the tingling, 'pins and needles' or *paresthetic* quality (i.e., dysesthesias) that defines the nonpainful phantom limb. Some sufferers describe bouts of paroxysmal shooting pain that travel up and down the limb. Others report the phantom to be in a cramped or otherwise unnatural posture that gives rise to excruciating pain. Many amputees describe the pain in the phantom limb as indistinguishable from the pain they experienced in the limb before amputation. In still others, the phantom may be immobile or paralyzed so that attempts to move it generate pain. Finally, the phantom is often the seat of an intense burning pain, as if the hand or foot were held too close to an open flame. Frequently amputees suffer from several types of pain. Phantom limb pain, it appears, is equally prevalent among men and women, and is not related to side, level, age, or cause of amputation (Jensen and Nikolajsen 1999).

Surveys based on several thousand amputees reveal that between 78 percent and 85 percent of patients continue to experience significant amounts of phantom limb pain more than 25 years after amputation (Sherman et al. 1997). Prospective studies of the incidence of phantom limb pain are consistent with these estimates (Jensen et al. 1985, Nikolajsen et al. 1997). Between 50 percent and 71 percent of amputees report phantom limb pain two years after amputation. There is a reduction in the frequency and duration of attacks over time. Equally striking is the low success rate of treatments for phantom limb pain: in the long term only 7 percent of patients are helped by the more than 50 types of therapy used to treat phantom limb pain (Sherman et al. 1997).

The different qualities of phantom limb pain, its high prevalence and incidence, and its low treatment success rate point to the complexity of the problem. Phantom limb pain is not a unitary syndrome, but a symptom class, with each class subserved by different etiologic mechanisms. There is not just one cause of phantom limb pain. The development and severity of this disorder are determined by the interaction of multiple factors that vary across individuals. These include ectopic activity from peripheral nerves that were transected during amputation and have formed a neuroma, contributions from the sympathetic nervous system, sensitization of central nervous system structures due to the effects of amputation, inflammation and ectopic activity, alterations in the functional organization of sensory nucleii, past experience with pain, and the immediate physiological, motivational, and psychological states of the individual.

1.1 Pain Memories in Phantom Limbs

A striking property of phantom limb pain is the presence of a pain that existed in a limb prior to its amputation (Katz and Melzack 1990). This class of phantom limb pain is characterized by the persistence or recurrence of a previous pain, has the same qualities of sensation, and is experienced in the same region of the limb as the pre-amputation pain. Cases studies of amputees have revealed pain 'memories' of painful diabetic foot ulcers, bedsores, gangrene, corns, blisters, ingrown toenails, cuts and deep tissue injuries, and damage to joints and bony structures. In addition, the phantom limb may assume the same painful posture as that of the real limb prior to amputation, especially if the arm or leg has been immobilized for a prolonged period.

The proportion of amputees who report similar pain before and after amputation may be as high as 79 percent (Katz and Melzack 1990), although according to prospective studies, the incidence is approximately 10 percent at one year after amputation (Jensen et al. 1985, Nikolajsen et al. 1997). Pain also persists in patients with deafferentation that does not involve amputation. In these conditions, the involved body part is still present but it is devoid of sensibility due to an interruption in the supply of sensory (afferent) information (i.e., deafferentation). Brachial plexus avulsions, in which the sensory nerve roots supplying the arm and hand are torn from the spinal cord, often

produce pain that is felt in the deafferented and anesthetic region. Similarly, patients with spinal cord injuries may complain of pain referred to body parts below the level of the transection, including, for example, a patient who continued to feel the pain of an ingrown toenail after a complete spinal cord break.

Painful and nonpainful sensations also persist or recur after surgical removal or deafferentation of body structures other than the limbs, such as breasts, teeth, and internal and special sense organs. Ulcer pain has been reported to persist after subtotal gastrectomy with removal of the ulcer. Patients have reported labor pain and menstrual cramps after total hysterectomy, rectal pain and hemorrhoids after removal of the rectum and anus, the burning pain of cystitis after complete removal of the bladder, and the pain of a severely ulcerated cornea after enucleation of an eye.

As noted above, not all phantom limb memories are of painful experiences. Nonpainful preamputation sensations do recur, but they are not as common and tend to include the sensation of objects that were once worn on the limb. These *superadded* sensations vary in complexity from such simple somatosensory qualities as the sensation of bandages that once wrapped a wound, a plaster cast, finger rings, and wristwatches to the multimodal, perceptually integrated phenomenon of a phantom foot clad in a sock and a shoe of specific type and color (Katz and Melzack 1990).

Taken together, these case reports and studies of amputees reveal that pain memories are not merely images or cognitive recollections; they are direct experiences of pain that resemble an earlier pain in location and quality. They are perceptually complex experiences that may even involve information from multiple sensory modalities including visual, olfactory, tactile, and motor components that had accompanied the original experience. The precise details of the experiences of pain involve localization, discrimination, affect, and evaluation—that is, all the dimensions of perceptual experience—and these properties are a function of integrated brain activity. It is likely that the outputs of sensitized spinal cells activate the neural structures in the brain that subserve memories of earlier events.

2. *The Nonpainful Phantom Limb*

The most salient property of the nonpainful phantom is the tingling, paresthetic quality, but sensations of temperature, posture, length, volume, and movement are also very common (Jensen and Nikolajsen 1999). The incidence of the nonpainful phantom is estimated to be between 80 percent and 100 percent. In fact, the presence of a nonpainful phantom is so common that it is considered a 'normal' sequela of amputation. It is standard practice to prepare the patient preoperatively for its subsequent occurrence.

The frequency with which a painless phantom occurs does not appear to be related to the extremity, side, level of amputation, age, gender, or reason for amputation (Jensen and Nikolajsen 1999). There is considerably more variability in the estimates of the duration of the nonpainful phantom. In some amputees, the phantom may disappear within days or weeks of the amputation. In others, it remains unchanged for up to 60 years.

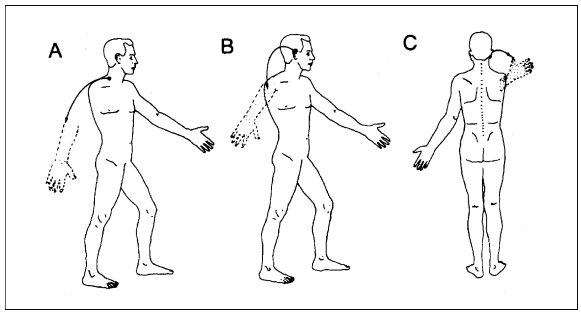

Figure 1
Illustrations of a patient who sustained a complete forequarter amputation of his right shoulder in a work accident and later developed phantom limb pain. The figurines depict the process of telescoping following a series of treatments with transcutaneous elecrical nerve stimulation. Dots represent points where stimulation was applied. Arrows indicate the sites of referred phantom limb sensations
Source: Adapted from Katz and Melzack 1987, with permission
A The patient reported a phantom limb of normal length that consisted of only the lower arm and hand, with a gap between the shoulder stump and phantom elbow. During stimulation applied at a point above the clavicle, the patient reported that his phantom hand begun to swell and become warm.
B In response to stimulation applied at a point on his right ear, he remarked that his phantom arm had 'telescoped' into the stump so that the elbow protruded from the stump.
C Later, stimulation of the same point resulted in a further retraction of the phantom, leaving only the wrist and hand attached to the stump

2.1 *Fading, Telescoping, and Shrinking of the Phantom Limb*

In many cases, the normal phantom undergoes a number of predictable changes within several years of amputation (Katz 2000). These include fading, telescoping, and shrinking. Immediately after amputation the phantom limb usually feels of normal length and size. Over time, the proximal portions begin to fade and soon disappear (Fig. 1). At this point, the phantom limb consists of only the extremity and major joints. These are felt at a normal distance from the body. The intervening parts have disappeared so that the phan-

tom hand or foot feels as if it were hanging, unattached, in empty space. In the case of an amputation performed at the shoulder, the phantom limb may consist of only the lower arm and hand, with a gap between the shoulder stump and the phantom elbow. In approximately one-third of amputees, a process of 'telescoping' begins. Among upper extremity amputees, the phantom is perceived gradually to approach the stump so that the hand is located in phenomenal space on a level with the elbow of the other arm. As the process continues, the gap between the stump and phantom continues to close; the amputee may find that the hand is protruding from, or attached to, the end of the stump. Later, the phantom hand may retract into the stump so that only the tips of the fingers jut out. In some cases, eventually even these may disappear completely and permanently into the stump.

There are times when a completely telescoped phantom will temporarily lengthen so that the hand or foot is once again felt at a normal distance from the stump. Less commonly, the converse also occurs; a phantom which is already telescoped may temporarily retract completely into the stump, or one which is of normal, or shortened length may become even shorter. The circumstances and stimuli which bring about these changes are extremely varied and range from peripheral stimulation of the stump and other regions of the body to more central influences.

The process of telescoping is sometimes accompanied by a marked shrinking of the phantom to the dimensions of a child's limb. This gradual diminution in perceived size or volume occurs in both upper- and lower-limb amputees. The extent of shrinking that occurs is variable. Adult amputees have reported the phantom to be the size of a doll's or baby's hand, a silver dollar, or even as small as a postage stamp (Katz 2000).

2.2 Cortical Maps and Phantom Limbs

Factors governing the fading, telescoping, and shrinking of the phantom are not well understood. The gradual fading of the phantom is thought to represent a perceptual correlate of the re-establishment of inhibitory control over hyperactive or spontaneously active cells which subserve the phantom. The gradual telescoping and shrinking have been hypothesized to depend upon input from intact peripheral fibers located in the stump. These experiences appear to be perceptual markers of the short- and long-term reorganizational changes that occur after amputation in central nervous system structures including the somatosensory cortex (Katz 2000, Merzenich et al. 1984). The perceived distance between the phantom and the stump may be a function of the distance separating their respective representations in the cortical map.

2.3 Psychophysical Correlates of Phantom Limb Sensations

Although a normal phantom occurs whenever nerve impulses from the periphery are blocked or otherwise removed, it is also true that direct stimulation of the amputation stump frequently exaggerates the tingling or paresthetic quality of sensation typical of the painless phantom limb. Careful questioning of amputees reveals that the nonpainful phantom limb is not perceived as a static phenomenon. The paresthetic quality of sensation, which defines the phantom limb percept, is in a constant state of flux, with changes occurring in intensity, body part, or both.

One mechanism that accounts for the paresthetic component of the phantom limb is a cycle of sympathetic-efferent somatic-afferent activity (Katz 2000). Changes in the intensity of phantom limb paresthesias reflect the joint activity of cholinergic (sudomotor) and noradrenergic (vasomotor) postganglionic sympathetic fibers on primary afferents located in the stump and stump neuromas. Release of acetylcholine and norepinephrine from postganglionic sympathetic fibers produces transient vasoconstriction and heightened skin conductance responses. In addition, neurotransmitter release onto apposing peripheral fibers trapped in stump neuromas increases primary afferent discharge. This information is transmitted rostrally, where it gives rise to referred phantom sensations upon reaching central structures subserving the amputated parts of the limb. The moment-to-moment fluctuations in the intensity of phantom limb paresthesias reported by many amputees may, in part, reflect a cycle of sympathetic-efferent somatic-afferent activity. Increases in the intensity of phantom limb paresthesias would follow bursts of sympathetic activity, and decreases would correspond to periods of relative sympathetic inactivity. If central sensitization has also developed, through prior injury, trauma during amputation, or peripheral inflammation, or if the sympathetic-sensory coupling involves nociceptors, the sensation may be one of dysesthesia.

2.4 Psychological and Emotional Processes Influence the Phantom Limb

The idea that emotional and psychological processes can cause pain traditionally has been tied to the notion of psychopathology. However, it is becoming increasingly clear that under certain circumstances pain may be triggered by these processes in psychologically healthy individuals as well. The aggravation or alleviation of pain referred to phantom body parts may be mediated in part by psychological processes that alter anxiety levels. Phantom breast pain after mastectomy is provoked by emotional distress in 6 percent of women three weeks after surgery, and in 29 percent one year later. Of lower-extremity amputees, 50 percent report that attacks of phantom limb pain are

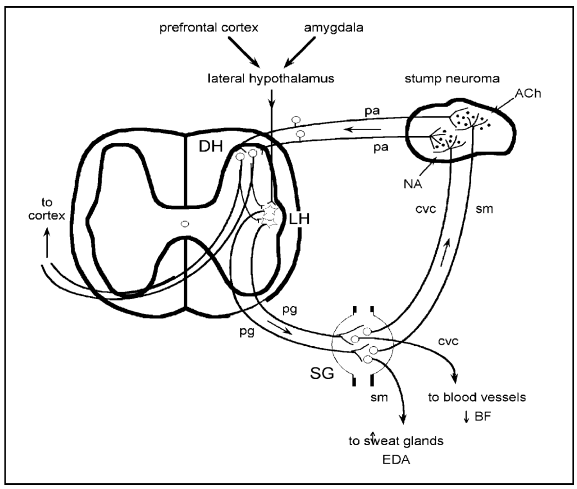

Figure 2
Schematic diagram illustrating a mechanism of sympathetically generated phantom limb paresthesias. Spontaneous sympathetic activity or excitatory inputs descending from cortex (e.g., due to the perception of a salient event, loud noise, thought, feeling, etc.) increase the discharge rate of preganglionic (pg) sympathetic neurons with cell bodies in the lateral horn (LH) of the spinal cord and terminals in the sympathetic ganglion (SG). These neurons excite postganglionic noradrenergic (NA) cutaneous vasoconstrictor (cvc) and cholinergic (ACh) sudomotor (sm) fibers that impinge on effector organs (vascular smooth muscle and sweat glands) in the stump and on sprouts from large diameter primary afferent (pa) fibers that have been trapped in a neuroma. The release of ACh and NA on effector organs results in increased electrodermal activity (EDA) and decreased blood flow (BF) to the stump. Release of these chemicals in the neuroma activates primary afferents that project to spinal cord dorsal horn (DH) cells subserving the amputated parts of the limb. These neurons, in turn, feed back to the preganglionic sympathetic neurons and project rostrally where the impulses contribute to the perception of phantom limb paresthesias. If DH cells have been sensitized due to injury, or nociceptive primary afferents are activated, then the perception may be one of dysesthesias
Source: Adapted from Fields 1987, with permission

triggered by emotional distress as long as seven years after amputation. A combination of progressive relaxation training and Electromygraphic (EMG) biofeedback of stump and forehead muscles produces significant reductions of phantom limb pain and anxiety, which are sustained for up to three years.

Psychological or emotional processes can bring about transient but profound alterations in the quality and intensity of phantom limb sensations. These processes include concentration, distraction, relaxation, fright, forceful reminders of the events that led to amputation, and witnessing cruel and violent acts. One amputee, interviewed by the present writer, described his reaction to an accident involving his wife by reporting 'goose bumps and cold shivering down the phantom [leg]. It went through me. Everything emotional will get you that.' Another amputee stated, 'It's like everything I feel goes there—the good and the bad' (Katz 2000).

2.5 A Centrally-triggered Sympathetic-efferent Somatic-afferent Mechanism

The model schematically represented in Fig. 2 outlines a mechanism through which cognitive and affective processes associated with higher cortical and limbic centers may alter phantom limb sensations. The reciprocal connections between cortical, limbic, and lateral hypothalamic structures are well documented. The lateral hypothalamus is involved in the control and integration of neural activity associated with affectively charged behavior, and has direct projections to the lateral horn of the spinal cord. The intensity of phantom limb paresthesias and dysesthesias may thus be modulated by higher brain centers involved in cognitive and affective processes via a multisynaptic network of descending inputs, which impinges on preganglionic sympathetic neurons, producing diffuse peripheral autonomic discharge and activation of primary afferent fibers located in stump neuromas.

3. Summary and Conclusions

The phantom limb is not perceived as a static entity but as a frequently changing perceptual experience dependent upon the current sensory input, emotional state, and past experience of the individual amputee. Phantom limb phenomena range from simple, diffuse sensations of tingling to perceptually complex experiences of pains and lesions that originally were felt in the limb prior to amputation. While phantom pains and other sensations frequently are triggered by the perception of salient events, thoughts, and feelings, there is no evidence that the painful or painless phantom limb is a symptom of a psychological disorder. The sympathetic nervous system may provide an important link between higher brain centers involved in cognitive and affective processes and phantom limb sensations through its peripheral actions on primary afferents located in stump neuromas.

See also: Chronic Pain: Models and Treatment Approaches; Pain, Neural Basis of; Parietal Lobe; Somatosensation

Bibliography

Fields H L 1987 *Pain*. McGraw-Hill, New York

Jensen T S, Nikolajsen L 1999 Phantom pain and other phenomena after amputation. In: Wall P D, Melzack R (eds.) *Textbook of Pain*, 4th edn. Churchill Livingstone, Edinburgh, UK, pp. 799–814

Jensen T S, Krebs B, Nielsen J, Rasmussen P 1985 Immediate and long-term phantom pain in amputees: Incidence, clinical characteristics and relationship to pre-amputation limb pain. *Pain* **21**: 268–78

Katz J 2000 Individual differences in the consciousness of phantom limbs. In: Kunzendorf R G, Wallace B (eds.) *Individual Differences in Conscious Experience*. John Benjamins, Amsterdam, pp. 45–97

Katz J, Melzack R 1987 Referred sensations in chronic pain patients. *Pain* **28**: 51–9

Katz J, Melzack R 1990 Pain 'memories' in phantom limbs: Review and clinical observations. *Pain* **43**: 319–36

Merzenich M M, Nelson R J, Stryker M P, Cynader M S, Schoppmann A, Zook J M 1984 Somatosensory cortical map changes following digit amputation in adult monkeys. *Journal of Comparative Neurology* **224**: 591–605

Mitchell S W 1871 Phantom limbs. *Lippincott's Magazine of Popular Literature and Science* **8**: 563–9

Nikolajsen L, Ilkjaer S, Krøner K, Christensen J H, Jensen T S 1997 The influence of preamputation pain on postamputation stump and phantom pain. *Pain* **72**: 393–405

Sherman R A, Devor M, Jones D E C, Katz J, Marbach J J 1997 *Phantom Pain*. Plenum Press, New York

J. Katz

Phenomenology in Human Science

Conceiving of phenomenology as a methodology of the human sciences imposes two types of constraint on the treatment of phenomenology. Since both types have formal as well as thematic implications they deserve a few introductory words of clarification.

1. The Asynchronism of Phenomenology and the Human Sciences

Relating phenomenology with the social and behavioral sciences involves a many-sided problem of *asynchrony*. When phenomenology emerged as a philosophical enterprise at the end of the nineteenth and the beginning of the twentieth century there were no social and behavioral sciences. At best there were currents at thought, hardly disciplines, that only very much later, after significant modifications and paradigmatic shifts, were categorized as social or behavioral sciences. There was, however, for instance psychology, known, criticized, adapted, or rejected by early phenomenologists, and those attitudes were reciprocated (Spiegelberg 1960). Above all, psychology was not (and never became) a unitary science, but a set of competing (and changing) schools of thought. A similar variety and changeability holds for early sociology. Also phenomenology changed and has kept changing, sometimes with one and the same author, Husserl himself being the most notorious example (Embree et al. 1997).

Hence it happened that the psychology to which the 'early Husserl' referred was quite different from the ones the 'middle' or even the 'late' Husserl was confronted with. This variety increases considerably if different phenomenologists are considered, such as Husserl, Pfänder, Scheler, Heidegger, Merleau-Ponty, Sartre, and Schütz. Excepting Husserl, their names indicate that these philosophers were not only, at least not always, phenomenologists. If one defines, as it is done within the scope of this article, phenomenology as a methodological attitude the latter can be adopted for a specific purpose, but does not necessarily make its user a phenomenologist for good. Historically seen, 'doing phenomenology' (Spiegelberg 1960) is—for philosophers as well as human scientists—a much more appropriate characteristic than 'being a phenomenologist.' After all phenomenology has, according to Spiegelberg (1960), been a movement rather than a school. Variety has always been greater than unity.

Also in the sciences with which phenomenology got related there has been considerable variance. The early psychology, for example, of associationism and elementism had so little in common with gestalt psychology which, in turn, was so different from behaviorism or from modern cognitivism that for each of them the relationship with phenomenology has carefully to be redefined, in fact from both ends.

In sociology, whose relationship with phenomenology has been more intensive and durable, it is important to identify carefully the theoretical position with which phenomenology is related. More than is the case in other social sciences sociology provides approaches and specialties engendered or influenced by phenomenological thinking, such as ethno-methodology and conversation analysis and, above all, there is an academically established phenomenological sociology which, mainly thanks to the oeuvre of Schütz (1962/66), has been considered the prototype of phenomenology in the context of social sciences. 'Phenomenological Psychology,' however, as it was introduced by Husserl (1925/1977) will have to be disregarded in this article, as far as it is the name for a branch of transcendental philosophy. The prominent exception remains Gurwitsch's (1964, 1966) creative synthesis of gestalt theoretical psychology and phenomenology.

At the end of the twentieth century we register many unilateral as well as mutual influences between phenomenology and the human sciences—the latter term, by the way, popularized by phenomenologists like Merleau-Ponty, Strasser, and Ricoeur. Among these sciences there are many of very recent origin. The

phenomenology with which they relate may be of any vintage (idealistic, realistic, existential, hermeneutic), sometimes treated in an eclectic manner. The responsibility, however, for the resulting phenomenological eclecticism rests with the human scientists.

To summarize the references to the first type of constraint: whoever is interested in the relationship between phenomenology and the human sciences must keep in mind its asynchronous historicity and multiplicity.

2. *Phenomenology as Methodology*

The second type of constraint is the limitation to *methodology*. On the one hand, this should be self-evident if one considers the number and variety of the sciences involved (Natanson 1973). An encyclopedia of phenomenology (Embree et al. 1997) lists phenomenological approaches and tendencies in about two dozen nonphilosophical disciplines, most of which belong to the human sciences. It is their inequality in terms of age, formalization, homogeneity, empiricism, and affinity with natural or cultural sciences that renders their comparability with phenomenology so complicated.

On the other hand, the limitation to methodology should not be too constraining; phenomenology itself is widely considered a specific methodological approach. Even if occasionally phenomenology is ontologically demarcated and referred to as the *systematic study of human consciousness or experience*, this is not much of a constraint if one accounts for *intentionality* with which phenomenologically (almost) all mental processes (perceiving, thinking, judging, remembering, feeling, desiring, willing) are endowed, namely, as meaning-something-as-something in being aware of it. Both this perspectival meaning-relatedness and the corresponding 'horizon of all horizons,' the *lifeworld*, in which 'things appear in terms of their experiential qualities, values, and uses' (Welton 1997), have become criteria of a phenomenological approach or orientation in the human sciences. The terms of 'approach' and 'orientation' are given preference to the totalizing (and hence overdrawn) designation of a phenomenological psychology, sociology, etc.

3. *Performing the* Epoché

Before these criteria can be exemplified a brief look into the central (Husserlian) method of *epoché* may be permitted although its proper field of application, transcendental phenomenology, is beyond the scope of this article. Yet the basic idea of *epoché* and of the related method of *reduction* provides the master-key to phenomenological methodology.

As agents in their everyday life human beings experience the world, including their own acting, perceiving, feeling selves, as really existing. Husserl (1977) considered this universal belief in the actuality (*Wirklichkeit*) of the world the general thesis of the *natural attitude*. For the purpose of a phenomenological analysis, this belief has to be suspended (as far as posssible). Suspension means neither negating nor doubting the existence of world and self. The belief, normally pervading all acts and states of consciousness, is merely 'bracketed,' that is, put out of operation. These metaphors are to signify the transition from the natural to the phenomenological attitude. By means of this transition world and self are transformed into, that is, reduced to, mere phenomena.

4. *Criteria of a Phenomenological Approach*

While the transcendental-phenomenological purpose of performing the *epoché* is of little or no interest in the human sciences and even phenomenologists like Merleau-Ponty (1962) have done their phenomenological analyses without performing *epoché*, the suspension of beliefs is an important methodological tool which is operative in what traditionally have been called the functions of a phenomenological approach in human science: (a) the critical, (b) the descriptive, and (c) the interpretive function. Although critique, description, and interpretation are distinct methods which for many purposes must be kept apart, under phenomenological premises they turn out to be interrelated.

4.1 *The Critical Function*

Given the fundamental postulate 'to the things themselves,' the primordial method is some kind of *epoché*. Critical is here the practice of disbelief with respect to all explicit and silent assumptions one has held so far about the matter to be studied, the theories, hypotheses, and methods, provided by the pertinent science, but also a self-critical avoidance of personal habits and predilections of thinking that may bias or distort the required 'fresh look' at the phenomenon in question. Equally critical the phenomenological observer has to be with respect to the language in which the unbiased observation has to be described. The theories that are to be suspended are often implied in the concepts that are intended to be descriptive, but actually are interpretive. A special difficulty arises here for the psychologist since the language for mental states and processes is predominantly metaphorical.

4.2 *The Descriptive Function*

As anticipated these critical caveats are also an introduction into the problem of phenomenological description in the human sciences. Although as old as

phenomenological methodology one caution has to be reiterated, since its disregard has turned out to be one of the most obstinate prejudices, namely that a phenomenological (-psychological) approach amounts to a method of introspection. The opposite is closer to the truth: phenomenological or 'intentional(ist)' description focuses on persons, selves, or others, as they are encountered in everyday situations. To clarify what this means a synthesis will be presented of conceptual-methodological recommendations that can be gathered from the writings of Husserl, Gurwitsch, Löwith, Merleau-Ponty, and from the phenomenological approaches in human science, mainly in psychology, psychiatry, and sociology.

The methodological foundation of phenomenological description was laid by Husserl who, in his 'principle of all principles,' demanded to describe things as they are experienced, strictly within the limits, but also to the full extent in which they are experienced. A merely consequent, but historically influential specification of this postulate was given by Gurwitsch who, in his doctoral dissertation of 1929 (a study of relations between phenomenology and gestalt theory) had adopted the *field concept* (see *Lewin, Kurt (1890–1947)*) and applied it to Husserl's concept of noema: a 'theme' in the center of what one is aware of is always experienced within a 'thematic field' and further 'marginal' objects of consciousness (Gurwitsch 1964). His study of the relationship between theme and thematic field explicates what Husserl since his 'Ideas' has developed as the *horizonal structure of consciousness*.

Closely connected was Gurwitsch's thematization of *perspectivity* (*avant la lettre*): an object, perceived from different angles and distances, that is, from varying viewpoints, presents itself as identical in a system of noemata. Both conceptions, perspective, and horizon, have become essential guidelines for phenomenological descrption:

(a) Whatever is experienced (perceived, cognized, remembered, imagined, anticipated), is experienced from a given viewpoint (an epistemological insight going back to Leibniz) and as the same from varying (i.e., successive) viewpoints.

(b) Whatever is experienced from a given viewpoint is experienced in an horizonal structure that corresponds to and changes with the point of view or perspective. Horizon means, as the everyday usage suggests, both limitation and openness of experience. Moving one's 'standpoint' leads to a shift of horizons, which holds for physical, bodily, as well as mental locomotion.

(The fact that human beings are inevitably *en route* to the horizon without ever reaching it, only the places where it was, has been made a topic of phenomenological anthropology by van Peursen 1954.)

Both perspective and horizon have become interrelated structural elements of the phenomenological conception of *situation*. Whatever or whomever a person experiences is encountered in a perspectivally structured situation. The essential *situatedness of experience* entails that in everyday experience, there is no 'decontextualized' object (thema): human beings coexist with nonhuman objects and, inversely, objects, as utensils, refer to their users, as Heidegger (1962) has shown.

Karl Löwith, a student of Heidegger, was one of the first who, in his study of *The Individual in the Role of Fellow Human Being* (1928), developed a phenomenological (i.e., nonbehavioristic) concept of human behavior or comportment (*Verhalten*). He distinguishes between three ways of relating in behavior: (a) with others (fellow human beings), (b) with nonhuman objects (e.g., utensils), and (c) with oneself. He demonstrates that and how each of these intentional directions involves the other two. For example, relating with a fellow being involves not only objects belonging to him or her or to oneself, but also a modification, although not necessarily conscious, of one's self-concept.

Decades later, in a simplified version, the social psychologist T. M. Newcomb, who had developed a model of communicative acts, in which two persons communicate about something, argued that for social psychology there is no (purely) interpersonal relation in an environmental vacuum, nor any object-relation in a social vacuum.

Löwith's conception of the interdependence of the world of fellow human beings (*Mitwelt*), the world of things (*Umwelt*), and self, was further elucidated and made relevant for phenomenological description in Gurwitsch's study of the 'encounter of fellow human beings within the horizon of the natural surrounding world,' in which the world of utensils (*Zeug-Umwelt*) is a prototype of what Max Scheler—in critical opposition to Husserl's 'platonizing' phenomenology—had introduced as 'milieux' (from which Gurwitsch derived his (German) title of the *Milieuwelt*). Not unintentionally, Scheler (1973) had made use of the term 'milieu.' After its misuse in deterministic theories of descendence, Scheler reinstated milieu as a relational term: milieu is what has effects on an organism. For humans, only those objects constitute the milieu which correspond to the values of human attitudes.

'Milieu' came to be replaced by *Umwelt*. This term was introduced into biology by von Uexküll (1909) to signify the subjective environment that corresponds to the structure and state of an organism's 'inner world.' As sensory *Merkwelt* and as motor *Wirkwelt* Umwelt is species-specific. Uexküll himself generalized the concept to the meaningful ambient of persons. As the environment-as-experienced-and-acted-upon by human (and animal) beings *Umwelt* is quasisynonymous with the phenomenologically conceived situation. Beyond phenomenology, it has been adopted for *The Study of Relations in Public* by Erving Goffman (1971) and, within the 'dramaturgical model'

of social life, by Rom Harré (1979). In the context of gestalt theory K. Koffka (1935), for whom phenomenology meant 'as naive and full a description of direct experience as possible,' introduced a distinction between the 'geographical' and the 'behavioral' environment. The famous exemplification is the German legend of the man who during a snowstorm rode over the frozen Lake Constance but dropped dead when he learnt what he had just done. The 'geographical' environment was a lake, the 'behavioral' environment a snow-covered plain. In other words and phenomenologically, the horseman's behavior was 'a riding-over-a-plain, but not a riding-over-a-lake.' Koffka's example, although fictitious, once more illustrates the validity of the 'postulate of all postulates' and the full applicability of intentionality to behavior and environment.

Without regarding his own approach as phenomenological, Charles Taylor (1964) refers to Koffka (as well as to Merleau-Ponty), when he proposes the method of 'intentional description' for the analysis of 'intentional environments.'

The globalization of the environment as it is experienced and acted upon is to be seen in one of the major meanings of *lifeworld*, the world as the ensemble of what is taken for granted. Its structures have been analysed by Schütz and Luckmann (1973/1989).

4.3 From Description to Interpretation

Since the taken-for-grantedness is a significant, but normally latent feature of everyday situations its analysis goes beyond mere description. In a phenomenologically oriented psychology and psychiatry the *situated person*, conceived as body-subject in meaningful and spatiotemporally articulated environments, has since Merleau-Ponty (1962) and the Utrecht School (Kockelmans 1987) been in the focus of interest (Herzog 1992, Herzog and Graumann 1991). The situated person's intentional relatedness in its various manifestations of perceiving, acting, desiring, feeling, remembering, communicating, and sharing their situation with others, has become the topic and task of phenomenological studies in the human sciences. Phenomena, however, are, as Heidegger (1962, §7) argued, only partly and certainly not initially apparent, but hidden. Also what is taken for granted is hidden from attention. To uncover the implicit efficacy of the partly hidden, *heuristic categories* have been developed, some of which will be briefly indicated.

A subject in intentional relation to objects and/or other persons is the minimal constitutive structure of a situation. As has been argued before, the subject has a place from which the situation is perceived or acted upon which, in turn, implies that the subject is or has a body occupying that place (as viewpoint) at a given time and for a certain duration and at a distance from whatever or whomever the body-subject experiences.

Following the implications of these few situational features will reveal a host of possible conditions or determinants of the experiences to be studied. To name a few: the body-subject *per se* is of a given gender, age, state of health, mood, civil or social status, group or class or ethnic membership, plus any of a set of psychological, for example, personality characteristics, etc. Some of these features will be relevant for the experience or behavior to be studied, some will not. But none of them, however inconspicuous or hidden, must *a priori* be considered irrelevant or even 'taken for granted.'

The same careful attention is to be paid to the other elements of the situation: its spatiotemporal articulation, its historicity, and last, but not least, its sociality and communicative structure. Since the latter term covers a wide range from spoken and written language to nonverbal communication and silence, the categories developed in philosophies and the techniques provided by the various sciences of language and signs may help to identify and to elucidate the variety and the intricacies of situated discourse and communication.

5. Concluding Comments

While it is true that many researchers, mainly those from an experimental tradition, shun the time, the effort, and the expense necessarily connected with phenomenological description and hermeneutic interpretation, it is equally evident that phenomenological methodology has been established firmly in social and behavioral sciences with a long-standing exchange with phenomenology (like psychiatry). But this methodology is also gaining a foothold in sciences without this tradition (like behavioral geography, ecology, political science, and even artificial intelligence). One may speculate whether the increasing interest in phenomenological methodology is a consequence of the generally growing realization that many of the traditional natural sciences, at least, to the extent that human beings are involved in natural and technical processes, are also human sciences.

See also: Husserl, Edmund (1859–1938); Phenomenology in Sociology; Phenomenology: Philosophical Aspects

Bibliography

Embree L, Behnke E A, Carr D, Evans J C, Huertas-Jourda J, Kockelmans J J, McKenna W R, Mickunas A, Mohanty J N, Seebohm T M, Zaner R M (eds.) 1997 *Encyclopedia of Phenomenology*. Kluwer, Dordrecht, The Netherlands

Gurwitsch A 1964 *The Field of Consciousness*. Duquesne University Press, Pittsburgh, PA

Gurwitsch A 1966 *Studies in Phenomenology and Psychology*. Northwestern University Press, Evanston, IL
Gurwitsch A 1979 *Human Encounters in the Social World* (ed. Métraux A, trans. Kersten F). Duquesne University Press, Pittsburgh, PA
Heidegger M 1962 *Being and Time* (trans. Macquarrie J, Robinson E). Harper & Row, New York
Herzog M 1992 *Phänomenologische Psychologie—Grundlagen und Entwicklung* (Phenomenological Psychology—Foundations and Development). Asanger, Heidelberg, Germany
Herzog M, Graumann C F (eds.) 1991 *Sinn und Erfahrung—Phänomenologische Methoden in den Humanwissenschaften (Meaning and Experience—Phenomenological Methods in the Human Sciences)*. Asanger, Heidelberg, Germany
Husserl E 1977 *Phenomenological Psychology: Lectures Summer Semester 1925* (trans. Sconlon J). Nijhoff, The Hague
Kockelmans J J (ed.) 1987 *Phenomenological Psychology: The Dutch School*. Nijhoff, Dordrecht, The Netherlands
Merleau-Ponty M 1962 *The Phenomenology of Perception* (trans. Smith C). Humanistic Press, New York
Merleau-Ponty M 1964 Phenomenology and the sciences of man. In: Edie J M (ed.) (trans. Wild J) *The Primacy of Perception*. Northwestern University Press, Evanston, IL, pp. 43–96
Natanson M (ed.) 1973 *Phenomenology and the Social Sciences*, 2 vols. Northwestern University Press, Evanston, IL
Ricoeur P 1981 *Hermeneutics and the Human Sciences* (trans. Thompson J B). Cambridge University Press, Cambridge, UK
Scheler M 1973 *Formalism in Ethics and Non-Formal Ethics of Values: A New Attempt Toward the Foundation of an Ethic Personalism*. (Trans. Frings M S, Funk R L) Northwestern University Press, Evanston, IL
Schütz A 1962/1966 *Collected Papers*, 3 vols. Nijhoff, The Hague
Schütz A, Luckmann T 1973/1989 *The Structure of the Life-world*, 2 vol. (trans. Zaner R, Engelhardt T Jr). Northwestern University Press, Evanston, IL
Spiegelberg H 1960 *The Phenomenological Movement*, 2 vols. Nijhoff, The Hague
Strasser S 1963 *Phenomenology and the Human Sciences*. Duquesne University Press, Pittsburgh, PA
van Peursen C A 1954 L'horizon. *Situation* **1**: 204–34
von Uexküll J 1909 *Umwelt und Innenwelt der Tiere*. [*Environment and the Inner Worlds of Animals*.] Springer, Berlin
Welton D 1997 World. In: Embree L et al (eds.) *Encyclopedia of Phenomenology*. Kluwer, Dordrecht, The Netherlands, pp. 736–43

C. F. Graumann

Phenomenology in Sociology

1. The Origins and Scope of Phenomenological Sociology

Phenomenological sociology is the prescientific study of social life and the process by which humans interpret, experience, and understand their individual and collective realities. The work of the social philosopher and sociologist Alfred Schütz (1899–1959) provides the most important foundation for phenomenological sociology. Framed in the general atmosphere of the debate between scientific and antiscientific movements that arose in the late nineteenth century, phenomenology places the social sciences in the context of everyday life (Thomason 1982). Strongly influenced by Henri Bergson, Edmond Husserl and William James, Schütz argues that a prescientific understanding of everyday life provides the only means by which a science of society is possible. 'The sciences that would interpret and explain human action must begin with a description of the foundational structures of what is prescientific, the reality which seems self-evident to men remaining within the natural attitude. This reality is the everyday life-world' (Schütz and Luckmann 1973, p. 3).

Just as the natural scientist must first understand the composition and interactions of subatomic particles in order to understand the nature of chemical reactions, the sociologist, Schütz argues, must understand the common-sense world. Unlike the subatomic world, however, the world of everyday life is permeated with the understandings given by those who constitute it. The task of the phenomenological sociologist is to understand how people make sense of their own lives. According to the phenomenologist, these subjective sense-making activities in everyday life are based on taken-for-granted knowledge passed down from generation to generation. In order to understand scientifically these subjective social actions the scientist must replace common sense explanations with objective scientific constructs derived from social theory (see *Theory: Sociological*). The phenomenologist seeks to understand social action in its own subjective terms, yet to describe this world scientifically using the tools of an objective science of society. The foundation of all social science, according to the phenomenologist must be the life-world.

2. The Life-world

Schütz states that the life-world (*Lebenswelt*) can be understood as 'that province of reality which the wide-awake and normal adult simply takes for granted as common sense' (Schütz and Luckmann 1973, p. 3). The everyday life-world provides us with a sense of the 'real.' It is through our position in, and experience of the life-world that we are social beings engaged and affected by the social and natural worlds. The most important characteristic of the life-world, according to Schütz, is that it is taken-for-granted. By this he means that individuals apprehend their worlds and its problems as self-evidently real—'that's just the way it is.' As a result, most individuals, most of the time, give little thought to the 'true' nature of the world around them. 'It is the unquestioned givenness of the life-world for its denizens, including those whose business

it is, in professional terms, to analyze problems of the social world, that phenomenology finds the ground of the social scientist's activity' (Natanson 1973, p. 40).

Taken-for-grantedness arises out of the typification of the phenomenal world. That is, our perceptions are categorized from a shared stock of knowledge as 'this' or 'that' type of thing. Our typifications, however, are only of an approximate nature. Such categories are held only until further notice. If contravened by future experiences, typifications must either be abandoned or reformulated. Taken-for-grantedness is further enabled through the use of time tested recipes for social action. Of the unlimited realm of potential social action open to individuals, most potential actions are circumscribed by a taken-for-granted sense of what is possible and not possible in such typical situations.

The universe of potential recipes for social action is also part of the social stock of knowledge from which typifications are drawn. An important point to be made is that the social stock of knowledge complete with its typifications and recipes for social action is pragmatic in nature. We simply tend to do what works and to avoid what does not work. Through typification and the use of recipes for social action the world becomes unproblematic and a matter of common sense.

3. *Phenomenology and Science*

A phenomenological study of the social world is one that addresses how humans experience the life-world. In contemporary sociology, phenomenological research has largely been identified with any research addressing the subjective perspectives of social actors. This, however, is a misunderstanding. Properly understood, phenomenology is prescientific, an attempt to ground the social sciences in human experience, a subjective appreciation of the human condition (Embree 1988, p. 270). Schütz (1962a) argues that if we are to understand social reality all social scientific constructs must not only be linked to what people experience, but how they experience it. Scientific understandings must be connected to the experiential process through which people in their everyday lives actually experience the world. He states, 'correctly understood, the postulate of subjective interpretation as applied to economics as well as to all other social sciences means merely that we always can—and for certain purposes must—refer to the activities of the subjects within the social world and their interpretation by the actors in terms of systems of projects, available means, motives, relevances, and so on' (Schütz 1962a, p. 35).

Phenomenological sociology has sometimes been inappropriately labeled as 'anti-scientific.' This misconception stems from the anti-scientific intellectual climate from which phenomenology arose (Thomason 1982). Sensitive to these movements, Schütz indeed understood the dehumanizing possibilities of science. His answer to this problem, however, was not to abandon science, but rather to ground the social science in the motives and realities of everyday life. Phenomenology does not necessarily attempt to replace or debunk science, but rather to provide a foundation for empirical investigations.

4. *The Contribution of Phenomenology in Sociology*

The possibility of a phenomenological sociology has been partially realized in two current sociological traditions: social constructionism and ethnomethodology. While both approaches address the life-world and its subjective realities, neither has fully realized the potential Schütz saw for a truly phenomenological sociology. That is, both have addressed the subjective nature of the social world, but neither has been the foundation of an objective science of society.

4.1 *Social Constructionism*

Phenomenological interest in the mediated and negotiated nature of all knowledge gave rise to social constructionism in sociology. Social constructionism, simply stated, is the study of the way in which people agree upon and define what reality is. Perhaps the most important work in this field is *The Social Construction of Reality: A Treatise in the Sociology of Knowledge* (Berger and Luckmann 1966) (for example, see *Sociology, Epistemology of*).

At the heart of social constructionism can be found one of phenomenology's most important concepts, the epoch. The epoch amounts to a radical skepticism not unlike Descartes' method, where belief in the 'realness' of the world is suspended. Schütz (1962c, p. 102) however, claims that Descartes was 'not radical enough,' that Descartes conclusion '*cogito*, *ergo sum*' ['I think therefore I am'] failed to doubt and analyze the thought itself. By methodologically setting aside or 'bracketing' the everyday world and even the nature of thought itself, the phenomenologist seeks to 'go beyond the natural attitude of man living within the world he accepts, be it reality or mere appearance' (Schütz 1962c, p. 104). The purpose of the phenomenological reduction (epoch) is to strip through the world of belief to the realm of consciousness and to examine the resulting contents.

Thomason (1982) points out that social constructionism in sociology stemming from the epoch can best be described as a methodological constructionism. Unlike ontological constructionism, methodological constructionism does not question the existence of a real world, but rather suggests that what we know of the world is always mediated and indirect. Methodological constructionism most closely resembles

Schützian phenomenology. Schütz states 'I am afraid I do not exactly know what reality is, and my only comfort in this unpleasant situation is that I share my ignorance with the greatest philosophers of all time' (Schütz 1964, p. 88, Thomason 1982, p. 4). That is, the methodological constructionist examines the agreements people make about the world, but does not question that a real world exists. Social constructionism remains an important field of inquiry in the natural and social sciences, but is of particular importance in sociology.

4.2 Ethnomethodology

Ethnomethodology is the second intellectual tradition linked to phenomenology. Ethnomethodology is also connected to the epoch, but most importantly to Schütz's commitment to the importance of the everyday life-world. However, unlike the prescientific and proscientific work of phenomenology, ethnomethodology as formulated by Garfinkel (1963) represents a radical break from the traditional models of social science with which Schütz had once tried to reconcile (Lynch 1988, p. 93). Ethnomethodology seeks to understand the method by which individuals construct, negotiate, and agree upon reality, but questions the possibility of an objective science of the subjective human condition. As a radically subjective pursuit, ethnomethodology falls short of the objective science of the life-world Schütz envisioned. Concerning such radically subjective endeavors Schütz (1962b, p. 52) maintains 'a method which would require that the individual scientific observer identify himself with the social agent observed in order to understand the motives of the later, or a method which would refer the selection of the facts observed and their interpretation to the private and subjective image in the mind of this particular observer, would merely lead to an uncontrollable private and subjective image in the mind of this particular student of human affairs, but never to a scientific theory.' While ethnomethodology remains an important influence in sociology, as currently formulated it falls short of the phenomenological sociology Schütz envisioned.

Without question, phenomenology has had a major impact upon modern sociology. Social constructionism and ethnomethodology each display a commitment to the epoch and the fundamental importance of the life-world, and therefore can directly be traced to phenomenological thinking. Both methods of analysis remain viable sociological traditions, and will no doubt continue to inform social research.

See also: Constructivism/Constructionism: Methodology; Ethnomethodology: General; Hermeneutics, History of; Hermeneutics, Including Critical Theory; Interactionism: Symbolic; Macrosociology–Microsociology; Phenomenology: Philosophical Aspects; Schütz, Alfred (1899–1959); Social Constructivism; Social Psychology: Sociological; Sociology, History of; Sociology: Overview; Weber, Max (1864–1920)

Bibliography

Berger P L, Luckmann T 1966 *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Doubleday, Garden City, NY

Embree L 1988 Schutz on science. In: Embree L (ed.) *Worldly Phenomenology: The Continuing Influence of Alfred Schutz on North American Social Science*. University Press of America, Washington, DC, pp. 251–74

Garfinkel H 1963 *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ

Lynch M 1988 Alfred Schutz and the sociology of science. In: Embree L (ed.) *Worldly Phenomenology: The Continuing Influence of Alfred Schutz on North American Social Science*. University Press of America, Washington, DC, pp. 71–100

Natanson M 1973 Introduction. In: Natanson M (ed.) *Phenomenology and the Social Sciences*. Northwestern University Press, Evanston, IL, Vol. 1, pp. 3–46

Schütz A 1962a Common-sense and scientific interpretation of human action. In: Natanson M (ed.) *The Problem of Social Reality: Collected Papers Volume One*. Martinus Nijhoff, Boston, pp. 3–47

Schütz A 1962b Concept and theory formation in the social sciences. In: Natanson M (ed.) *The Problem of Social Reality: Collected Papers Volume One*. Martinus Nijhoff, Boston, pp. 48–66

Schütz A 1962c Some leading concepts of phenomenology. In: Natanson M (ed.) *The Problem of Social Reality: Collected Papers Volume One*. Martinus Nijhoff, Boston, pp. 99–117

Schütz A 1964 The problem of rationality in the social world. In: Brodersen A (ed.) *Collected Papers II: Studies in Social Theory*. Martinus Nijhoff, The Hague, The Netherlands, pp. 64–90

Schütz A, Luckmann T 1973 *The Structures of the Life-World, Vol. II*. Northwestern University Press, Evanston, IL

Thomason B C 1982 *Making Sense of Reification: Alfred Schutz and Constructionist Theory*. Humanities Press, London

J. Williams

Phenomenology: Philosophical Aspects

Phenomenology has been one of the most influential twentieth century philosophical traditions. It began with the appearance of Husserl's *Logical Investigations* in 1900–01 and of Pfänder's *Phenomenology of Willing: A Psychological Analysis* in 1900, and continued with the work of philosophers such as Johannes Daubert, Adolph Reinach, Moritz Geiger, Max Scheler, Edith Stein, Gerda Walther, Roman Ingarden, and Wilhelm Schapp. These phenomenologists were active in Munich and Göttingen and the movement to which they belonged is often called 'realist phenomenology.' Throughout the century

phenomenology has been the object of a variety of increasingly far-reaching transformations: by Husserl himself, whose ideas in 1913 marked a turn towards an idealist and transcendental form of phenomenology and then by Heidegger, Sartre, and Derrida, among many others. Among phenomenologists who remained faithful to realist phenomenology were Ortega y Gasset, in Spain, and Felix Kaufmann and Alfred Schütz in Vienna, where Brentano's teaching at the end of the nineteenth century had launched the project of an exact descriptive psychology and caught the imagination of Austrian philosophers such as Meinong, Ehrenfels, Twardowski, and Husserl. In all its forms phenomenology has continued to influence sociology and the philosophy of the social world.

Phenomenology was initially conceived of as the careful description of the mind—act phenomenology—and its objects—object phenomenology. Successful description, it was thought, would lead to an account of the main types of mental phenomena, the main types of their objects, and of the necessary and often *a priori* interconnections between these. In Husserl's *Logical Investigations* knowledge, verification and logical structures are described as though they could in principle be instantiated by a solitary subject or even by solitary mental acts. But phenomenologists quickly began to describe the social aspects of mental phenomena and to analyze the nature of social objects.

Among the most important results of the early phenomenologists are their accounts of awareness and understanding of other minds, motives, social acts, collective intentionality, the main forms of social life, and the nature of social objects. These contributions to the phenomenology of social phenomena were also employed in discussions of social and political philosophy and in the creation by Scheler of the sociology of knowledge (for partial overviews see Toulemont 1962, Theunissen 1977).

1. *Empathy, Understanding, and Motivation*

In what does Mary's awareness of Sam's sadness consist? Phenomenologists of nearly all stripes attempted to defend two claims about such awareness—sometimes called 'Einfühlung' or 'empathy'—which also, they thought, hold of awareness of physical objects. Awareness of other minds can be, and often is, direct and perceptual. Together the two theories constitute direct realism about other minds. Critical realism about our awareness of physical objects is the claim that we do not directly see things but infer from sensory data to their existence. Similarly, critical realism about other minds is the theory that we are not directly aware of other people or their mental states but infer to these on the basis of sensory data. On the phenomenological view, Mary can directly see Sam and his sadness and her perception need involve no judgement or conceptualization and thus no inference. This view rests on the arguments given by early phenomenologists to show—against Kant and Helmholtz—that visual perception of things and events need involve neither inference nor concepts.

Traditionally the most popular variant of critical realism about other minds has appealed to reasoning from analogy. On the basis of observed analogies between her own behavior and mental states and observation of Sam, Mary is capable of attributing sadness to Sam. The locus classicus for criticisms of accounts of this kind is Scheler's 1913 appendix to his *On the Phenomenology and Theory of Feelings of Sympathy*, '*On the other I*.?' Scheler's own view contains the claim that Mary's awareness of Sam's sadness is as direct and as subject to error as her awareness of her own mental states (see on this Schütz's papers on Scheler in Schütz 1970, 1973). Less ambitious and more sober defenses of direct realism were provided by Edith Stein (1917/1980), who gives a subtle account of the relation of expression between mental states, on the one hand, and behavior and gestures, on the other hand; by Husserl (1973), who argues that Mary sees Sam's body as similar to her own body without judging this and thus without any reasoning from analogy; and by Schütz (1932). Phenomenological accounts of empathy also discuss the claim that some types of awareness of the feelings of another person involve make-believe feelings. Of the Gestalt psychologists influenced by phenomenology, Köhler (1947, Chap. 8) and Koffka (1946, Chap. 1) retain the claim that awareness of other minds need involve no judgement or inference but endorse critical realism. Bühler (1927, §9) defends both parts of the direct realism of the phenomenologists but within a naturalist and biological perspective that takes into account the dynamics of understanding other minds and the role of both steering mechanisms and criteria in bringing about 'mental contact.'

No account of empathy is complete without an account of what it is to understand other people. From the beginning, phenomenology was concerned with the analysis of the nature of motives and motivation and understanding thereof. In particular, the task was seen to be to provide an account of the relation between the internal causation of mental states and behavior, on the one hand, and motivation, on the other hand. The relation of motivation or justification was held to involve a number of distinct possible moments: the optimal case, where a mental state or disposition such as an intention, does indeed justify an action; conflict among motivations; the case where a mental state, for example, an emotion, is not justified by the subject's beliefs or perceptions. Many views were discussed and endorsed by different phenomenologists: what motivates and what is motivated are never the possible terms of a causal relation, for example, because they are ideal propositional contents; what motivates and what is motivated can be the

terms of a causal relation because they are concrete instantiations of ideal contents; the relations of causality and of motivation are always distinct but may have the same terms; there is a relation of superposition between motivation and causal relations; relations of motivation are causal relations but mental causality differs from physical causality. Finally, it was argued that motivational justification differs from both inductive and deductive justification, in particular because it often connects terms that cannot stand in the last two relations. Husserl himself thought that it was a 'basic law' of all psychological phenomena that they could stand in relations of motivation (Pfänder 1900, Husserl 1913/1969, Stein 1917, 1980).

In his 1932 *Der sinnhafte Ausbau der sozialen Welt: Eine Einleitung in die verstehende Soziologie* [*The Phenomenology of the Social World. An Introduction to 'verstehende' Sociology*], Alfred Schütz made use of the phenomenology of temporal experience in his analysis of motivation and understanding. 'In-order-to motives' explain how actions come about, genuine 'because-motives' explain by reference to preceding experiences how a practical project comes about. The present in which Sam is aware of his own motives, projects, and the states of affairs which are the objects of projects, provides him with a temporal perspective which is not available to Mary. This does not, Schütz thinks, amount to a denial that Sam's motivation can be grasped by Mary, only to the claim that her temporal perspective on his motivation and behavior must be segmented and discontinuous (Schütz 1932/1974, see also Grathoff 1995).

A further step towards accounting for the connection between understanding and temporal perspectives was taken by Wilhelm Schapp in his 1953/1976 book *In Geschichten verstrickt* (*Entangled in Stories*) in which such traditional phenomenological categories as action, its horizons, intentionality, and states of affairs are reworked in narratological terms.

2. *Social Acts*

Promises, Hume said, are the cement of society. But what is a promise? Reinach's 1913/1989 reply to this question is the most impressive piece of philosophical analysis ever penned by a phenomenologist (see Mulligan 1987). Like orders, requests, questions, and answers, promises are social acts. Such acts are typically addressed to someone, grasped by another person and are performed in the very act of speaking. They bring about changes. Thus a promise creates both an obligation and a claim. Actions of different sorts bring about states of affairs that satisfy or fulfil social acts; execution of the behavior commanded satisfies a command. Reinach warns against possible misunderstandings of the category of social acts due to dragging in the ideas to which we are accustomed. A command is neither a purely external action, nor is it a purely inner experience, nor is it the announcing to others of such an experience (Reinach 1983, p. 19).

Reinach's account of the variety of social acts and of their effects is only the first step in his demonstration of their role in social intercourse. The second step is provided by his analysis of six possible modifications of social acts. In his analyses of such solitary mental acts as seeing or judging, Husserl had already noted that such acts could be modified. Thus one modification of seeing is visual imagining, one modification of judging is supposing. Similarly, argues Reinach, each type of social act can be modified in various ways. A promise may be a sham or pseudo promise, the promisor does not really intend to do what he promises to do. A command may be a conditional command: 'In the event that such and such occurs, I order you to do such and such.' The social act of informing someone that something is the case, however, is not subject to this sort of modification. A conditional social act is to be distinguished from a social act that has conditional content, for example a promise to do one thing if some other thing occurs. A social act may be performed by more than one person.

> [We] have here to do with the case where each of the persons performs the act 'in union' with the others, where each knows of the participation of the others, lets the others participate and participates himself, we have one single act which is performed by two or more persons together (Reinach 1983, p. 24).

Similarly, a social act may be addressed to more than one person. Finally, a command or a promise, for example, may be performed in the name of another person; there are representative social acts.

3. *Collective Intentionality*

The intentionality of acts and states, their relation to an object or state of affairs, is always the intentionality of one subject. But whereas the intentionality of solitary acts such as judging, inferrings, or seeings typically involves only a more or less explicit awareness of the self—if I judge that it is raining, I am aware that I am judging, although this is not part of what I judge—there are acts, states, and activities which involve a more or less explicit awareness of others whether these are presented as 'you,' 'she,' 'they,' 'one' or 'us.' This phenomenon, nicely expressed in German by conjoining 'Mit' (with) and many a psychological verb, noun, or adjective (e.g., 'Mitschuld' collective guilt), is now called 'collective intentionality.' One example of the phenomenon, as we have seen, is furnished by joint social acts. Reinach also points out that in the case of actions that are not social acts it is possible to speak of several performing subjects of one and the same action. There is a way of acting 'in union.' The criminal law's concept of 'complicity,' as it seems to us, has to base itself on this, and such collective actions are also important for public law,

administrative law, and international law (Reinach 1983, pp. 24–5).

But the starting point for an account of collective intentionality is not 'Mithandeln' (joint action, co-action) but the analysis of the mental acts and states involving affective participation such as 'Mitgefühl,' fellow-feeling or sympathy, those involving shared beliefs and those involving trust, confidence, or belief in someone or something (see Scheler 1966, IV B 4, Reiner 1934, pts. 1 and 2).

Scheler's 1913 account of sympathy distinguishes between identification of emotional contagion, vicarious feeling, feeling something together with someone ('*Mit-einanderfühlen*'), fellow-feeling ('*Mitgefühl*'), benevolence or love of humanity, and personal love, and argues that these stand in different sorts of relation of dependence or foundation to each other (see also Walther 1922).

The phenomenologists made three important claims about collective intentionality. First, although for every type of collective intentionality implicit awareness of feeling, believing, willing, or acting together with others is to be distinguished from thoughts containing the concepts expressed by the different personal pronouns, such concepts depend on the relevant types of implicit awareness. Representations employing 'we' and 'they' presuppose presentations which help to fix the reference of our uses of the personal pronouns and cognate forms as in 'Cosa Nostra,' 'Sinn Fein,' 'Mare Nostrum,' and 'Pater Noster.' Second, collective intentionality cannot be reduced to mutual belief or knowledge. There is, Scheler claims, a type of 'co-experiencing something' which cannot be understood by saying that A experiences something that is experienced by B, and that both, in addition, know of their experiencing it (Scheler 1913/1973, p. 526).

Finally, the sense or feeling of togetherness may be actual or habitual, in which case it constitutes a type of habitual background (see Walther 1922). The notion of background, like that of horizon, was first explored within phenomenology in connection with perceptual experience.

4. *Emotions and Explanation: Ressentiment*

Of all the psychological phenomena analyzed by phenomenologists, emotions and related phenomena enjoyed by far the most attention. One reason for this is the conviction that a philosophy of mind which concentrates exclusively on beliefs, inferences, and desires can only scratch the mind's surface and that explanation and understanding within the human sciences will suffer from any such limitation. This conviction in turn relies on the claim that emotional dispositions, tendencies, long-standing states and sentiments, both those directed towards the past and those directed towards the future, determine desires and preferences. In order to mark the importance of emotions Scheler emphasizes that mind or spirit ('Geist') comprehends not only cognitive phenomena but also the will and affective phenomena, an emphasis also prominent in the fiction and essays of the Austrian thinker Robert Musil.

One important illustration of these claims is provided by Scheler's analysis of ressentiment and subsequent interest in the topic. Building on suggestions of Nietzsche and the historical sociologist Sombart, Scheler takes ressentiment to be a long-lasting psychological attitude that comes into being when emotions and sentiments such as envy, rancor, revenge, resentment, and spite are repeatedly bottled up and lead to 'value-illusions' and false evaluations. In particular, it is a felt impotence or incapacity to realize or live up to certain values that leads the woman of ressentiment to false evaluations. In the (slightly modified) example from La Fontaine's fable, the fox initially sets a high value on the grapes and changes its mind about their value solely as a result of its inability to obtain them. Scheler's account relies on his view that value judgments are true or false—'cognitivism.' It is indeed essential to the mechanism described that it involves a re-evaluation, and not merely one or another type of change of preference. But it is not necessary to assume cognitivism. A weaker version of the account would simply have it that ressentiment involves a transition from an attribution of value to an attribution of disvalue (or vice versa) that is caused in the way described by Scheler and is thus, in the absence of any other justification, unjustified, unmotivated, or irrational. The central historical example discussed by Scheler is Nietzsche's claim that early Christianity was characterized by pervasive ressentiment. One of the many other examples he discusses is the claim that the disinterested disposition to assist in the punishment of criminals and a certain sort of moral indignation together with the relevant evaluations emerge from lower-middle class envy. This claim was the object of extensive empirical investigation by the Danish sociologist Svend Ranulf (1930, 1938), who castigates Scheler's almost complete disregard of the canons of empirical inquiry.

5. *Types of Coexistence*

Scheler's taxonomy of forms of social coexistence draws on distinctions to be found in the classical writings of Tönnies, Weber, and Simmel and on his own philosophy of mind and ethical psychology. In particular, four types of coexistence are distinguished with reference to different types of collective intentionality: masses, communities, societies, and superordinate entities such as nation-states, state-nations, empires such as the Belgian Empire, and the Catholic Church. A mass is characterized by affective contagion, its members do not behave as autonomous

persons. The community—families, tribes, clans, peoples—displays such forms of collective intentionality as sympathy, trust, piety, loyalty, and collective responsibility rather than individual responsibility. A community forms a milieu or niche within which certain shared values are prominent. Empathy and understanding are there, typically automatic and unreflective. Philosophers have often assumed that the choice between the two accounts of empathy distinguished above is exclusive: either awareness of other minds is always immediate or it is always a process involving inference, for example analogical reasoning. Scheler rejects the assumption and argues that although the former account fits communities, the latter is typical of societal relations. Within societies self-interest, mistrust, and a lack of solidarity go together with the predominance of the artificial relations engendered by social acts such as promising and contracts which connect autonomous, individually responsible persons. The relations that constitute a community, unlike those created by promises and contracts, have no determinate temporal duration or terminus. In his account of such 'collective persons' as the nation–state—an account that he only sporadically attempts to make consistent with the requirements of ontological individualism—these are said combine certain traits of communities and societies. They resemble communities in that they are not typically held to enjoy a determinate duration and in the affective importance their members attach to the past. But like the autonomous persons standing in societal relations they enjoy personal responsibility. Finally, each form of coexistence is characterized by different values: the vital values of the community—its life, survival, flourishing, and death—contrast both with the values of civilization—utility, agreeableness, sociability—of society and with the spiritual values of nation–states and state–nations—culture, law, language. (On the relation between the account of values given by Scheler and by the founders of sociology, see Boudon 2000.)

These forms of coexistence are related, Scheler thinks, in the following ways. Societies are founded on communities, states on communities, and societal unities. Thus the societal relations created by promises and contracts rely on the fact that the persons standing in such relations have some familiarity with the phenomenon of trust characteristic of communities (see Scheler 1966, VI, B 4, Scheler 1963, 334ff., Stein 1970).

6. *Sociology of Knowledge*

Scheler's extensive writings on sociology in the 1920's, in particular his *Problems of a Sociology of Knowledge* and other papers collected in his 1925 *The Forms of Knowledge and Society*, display a growing appreciation for American Pragmatism. Together with the work of Karl Mannheim they established the sociology of knowledge. Scheler employs his accounts of collective intentionality and of affective phenomena in order to set out a comprehensive formulation of the view that scientific knowledge and technological knowledge are inseparably bound up with the 'instinctual structure' of the bourgeoisie, the desire to dominate nature, and with the predominance of vital values. Scientific knowledge and its sociology is opposed to philosophical and cultural knowledge ('Bildungswissen') and to religious knowledge ('Heilswissen') and the sociologies of the spiritual values essential to these, all of which rest on the natural world view and natural language. His account of the link between science and mastery of nature and his distinction between scientific and a superior, non-scientific knowledge have proved remarkably influential throughout the century (see Leiss 1994). But on two points Scheler differs from the majority of his many successors. He retains a realist view of scientific knowledge. The values he distinguishes from those of science, in his philosophy as in his sociology, are spiritual: salvation is not emancipation and philosophy does not involve hearkening to the voice of being.

7. *Social Objects*

What sort of an entity is a social object? There are, it seems, social entities of many different kinds—enduring 'substances' such as judges and the unemployed, processes such as electioneering, events such as voting, states such as ownership. But, as Reinach pointed out, neither of the two venerable philosophical distinctions, physical vs. psychological, temporal vs. ideal, is of much help if we want to understand the status of, for example, the claims and obligations created by promises.

Through the act of promising something new enters the world. A claim arises in the one party and an obligation in the other. What are these curious entities? They are surely not nothing … but they cannot be brought under any of the categories with which we are otherwise familiar. They are nothing physical; that is certain. One might rather be tempted to designate them as something psychical or mental, that is, as the experiences of the one who has the claim or the obligation. But cannot a claim or an obligation last for years without any change? Are there any such experiences? Recently one has begun to recognize again, in addition to the physical and psychical, the distinct character of ideal objects. But the essential mark of these objects, such as numbers, concepts, propositions, is their timelessness. Claims and obligations, by contrast, arise, last a definite length of time, and then disappear again. Thus they seem to be temporal objects of a special kind of which one has not yet taken notice (Reinach 1983, p. 9).

Husserl baptized such entities 'bound idealities.' Unlike 'free' idealities such as numbers, bound idealities such as word types, word meanings, and other

cultural entities such as institutions and tools come into being at a particular time and are 'earth bound … bound to particular territories'; if there were men on Mars there would be 'Mars bound' idealities (Husserl 1973, §65). Like many free idealities, bound idealities such as word meanings, rules, and types of ceremonies are multiply instantiable. The most thorough account of such entities is Ingarden's (1931) ontology of word-meanings and of works of art.

If social entities are brought into being then it is plausible to say that they are mind-dependent entities. But since mental activities belong to the category of episodes it might seem that social entities must all have the reality of episodes. This 'actuality theory,' from the German psychologist Wilhelm Wundt, was rejected in the two most substantive phenomenological ontologies of the social world, *A Contribution to the Ontology of Social Communities*, by Gerda Walther (1922), and *Foundation of the Theory of the Social Bond* (1932) by Husserl's Japanese pupil, Tomoo Odaka. Social entities are, they argue, brought into being and sustained by the mind, but do not all belong to the category of episodes. A thorough criticism of holism about social entities within the framework of Husserl's theory of objects is given by Felix Kaufmann (1929/30, Kaufmann's (1936, 1944) philosophy of the social sciences is, however, closer to the ideas of the logical positivists than to phenomenology).

What type of mental activity is responsible for social entities? According to Walther and Otaka, collective intentionality, implicit and explicit, and in particular we-intentionality, is a necessary condition for the emergence and maintenance of social entities. What, runs Walther's influential question, makes a group of laborers working for a constructor, who interact, know this, pass bricks along and receive such social acts as orders, into a community? What are the marks of a common form of life? Among the thirteen features she puts forward is the sense of togetherness or collective intentionality, a feature also necessary for the existence of all products of a community.

8. Sociology and History of Phenomenology

A notable sociological feature of phenomenology is the role played by female philosophers. A later parallel is the role of women within analytic philosophy, phenomenology's most influential rival.

Many of the distinctions, approaches, and theses of the phenomenologists were rediscovered within analytic philosophy. Reinach's analysis of social acts anticipates John Austin's account of speech acts. The phenomenological discovery of collective intentionality anticipates the work of John Searle and Raimo Tuomela. Similarly, the claim that social entities depend for their existence on collective intentionality has been defended by Searle. Jon Elster's work on such phenomena as ressentiment and on emotions in general (see Elster 1985, 1999) confirms the phenomenologists' conviction that a desire–belief–inference psychology and sociology are inadequate. A variety of theories of motivation and of its relation to causality have been canvassed since Wittgenstein. Scheler's account of communities and their relations to other types of social formation anticipates many of the contemporary varieties of communitarianism. Thus phenomenology is now by and large of interest only to the historian.

See also: Atomism and Holism: Philosophical Aspects; Collective Behavior, Sociology of; Collective Beliefs: Sociological Explanation; Collective Identity and Expressive Forms; Collective Memory, Anthropology of; Husserl, Edmund (1859–1938); Individualism versus Collectivism: Philosophical Aspects; Joint Action; Phenomenology in Human Science; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Schtz, Alfred (1899–1959); Shared Belief; Theory of Mind

Bibliography

Boudon R 2000 'La théorie des valeurs de Scheler vue depuis la théorie des valeurs de la sociologie classique'. Forthcoming.

Elster J 1983 *Sour Grapes. Studies in the Subversion of Rationality*. Cambridge University Press, Cambridge, UK

Elster J 1999 *Alchemies of the Mind. Rationality and the Emotions*. Cambridge University Press, Cambridge, UK

Grathoff R 1995 *Milieu und Lebenswelt. Einführung in die phänomenologische Soziologie und die sozialphänomenologische Forschung*, Suhrkamp, Frankfurt, Germany.

Husserl E (1900–1901, 1913) 1975 *Logische Untersuchungen*, ed. Holstein E. Martinus Nijhoff, The Hague, The Netherlands (1970 *Logical Investigations* [Trans. Findlay J N]. Humanities Press, New York)

Husserl E (1913) 1969 *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Erstes Buch: Allgemeine Einführung in die reine Phänomenologie*, ed. Biemel W. Martinus Nijhoff, The Hague, The Netherlands (1931 *Ideas: General Introduction to Pure Phenomenology* [Trans. Boyce Gibson W R]. Collier-Macmillan, London and New York)

Husserl E (1931) 1973 *Cartesianische Meditationen und Pariser Vorträge*, ed. Strasser S. Martinus Nijhoff, The Hague, The Netherlands (1960 *Cartesian Meditations*. [Trans. Cairns D. Martinus Nijhoff, The Hague, The Netherlands)

Husserl E 1939 *Erfahrung und Urteil. Untersuchungen zur Genealogie der Logik*, ed. Landrebe L. Academia, Prague (1973 *Experience and Judgment*. Routledge, London)

Ingarden R 1931 *Das literarische Kunstwerk* (1973 *The Literary Work of Art. An Investigation on the Borderlines of Ontology, Logic and Theory of Literature*. Northwestern University Press, Evanston, IL)

Kaufmann F 1929/30 Soziale Kollektiva. *Zeitschrift für Nationalökonomie* **1**: 294–308

Kaufmann F 1936 *Methodenlehre der Sozialwissenschaften*. Springer, Vienna, Austria

Kaufmann F. 1944 *Methodology of the Social Sciences*. Oxford (reprinted New York 1958, this is not a translation of Kaufmann 1936)

Köhler W 1947 *Gestalt Psychology*. Liveright, New York

Koffka K 1946 *The Growth of the Mind. An Introduction to Child-Psychology*. Kegan Paul, London

Leiss W 1994 *The Domination of Nature*. McGill-Queen's University Press, Montreal & Kingston
Mulligan K 1987 Promisings and other social acts: Their constituents and structures. In: Mulligan (ed.). *Speech Act and Sachverhalt: Reinach and the Foundations of Realist Phenomenology*. Hijhoff, Dordrecht, The Netherlands, pp. 29–90
Odaka T 1932 *Grundlegung der Lehre vom sozialen Verband*. Springer, Vienna, Austria
Pfänder A 1900 *Phänomenologie des Wollens. Eine psychologische Analyse*. Johann Ambrosius Barth, Leipzig, Germany
Ranulf S 1930 *Gudernes Misundelse og Strafferetens Oprindelse i Athen*, Det Kgl. Danske Videnskabernes Selskabs Historisk-filologiske Meddelelser XVIII, I, Copenhagen, Denmark, (1933/4 *The Jealousy of the Gods and Criminal Law at Athens*, Vols. I–II, Williams and Norgate, London)
Ranulf S 1938 *Moral Indignation and Middle Class Psychology. A Sociological Study*. Levin & Munksgaard. Ejnar Munksgaard, Copenhagen, Denmark
Reinach A (1913) 1989 Die apriorischen Grundlagen des bürgerlichen Rechtes. In: Schuhmann K, Smith B (eds.) *Sämtliche Werke* (1983 'The A Priori Foundations of the Civil Law,' Aletheia, Vol. III, Philosophy of Law, 1983, 1–142)
Reiner H 1934 *Das Phänomen des Glaubens. Dargestellt im Hinblick auf das Problem seines metaphysischen Gehalts*. Max Niemeyer Verlag, Halle (Saale), Germany
Schapp W (1953) 1976 *In Geschichten verstrickt: Zum Sein von Mensch und Ding*. B. Heymann, Wiesbaden, Germany
Scheler M 1955 Das Ressentiment im Aufbau der Moralen, Vom Umsturz der Werte, *Gesammelte Werke* 3, pp. 33–148 (1961 Ressentiment. Trans. Holdheim W W. First edition ed. Coser L A. Free Press of Glencoe, New York, 1961; reprint: Schocken Books, New York, 1972.
Scheler M (1924) 1960 Probleme einer Soziologie des Wissens. Die Wissensformen und die Gesellschaft, *Gesammelte Werke*, 8 (1980 *Problems of a Sociology of Knowledge* [Trans. Frings M S, ed. Stikkers K W]. Routledge & Kegan Paul, London)
Scheler M 1963 Schriften zur Soziologie und Weltanschauungslehre. *Gesammelte Werke* 6
Scheler M (1913/1916) 1966 Der Formalismus in der Ethik und die materiale Wertethik, Gesammelte Werke, 2, ed. Scheler M, Francke Verlag, Bern and Munich (1973 *Formalism in Ethics and Non-Formal Ethics of Values. A New Attempt Toward A Foundation of An Ethical Personalism* [Trans. Frings M S, Funk R L]. Northwestern University Press, Evanston, IL)
Scheler M (1913) 1973 Wesen und Formen der Sympathie. *Gesammelte Werke* 7 (1954 *The Nature of Sympathy* [Trans. Heath P Routledge & Kegan Paul], London. Reprinted, Archon Books, Hamden, CT, 1970)
Schütz A 1970 *Studies in Phenomenological Philosophy*, *Collected Papers*, *III*, Martinus Nijhoff, The Hague, The Netherlands
Schütz A 1973 *The Problem of Social Reality, Collected Papers I*. Martinus Nijhoff, The Hague, The Netherlands
Schütz A (1932) 1974 *Der sinnhafte Ausbau der sozialen Welt. Eine Einleitung in die verstehende Soziologie*. Suhrkamp, Frankfurt, Germany (1967 *The Phenomenology of the Social World*. Northwestern University press, trans. Walsh G, Lehnert F)
Stein E S (1917) 1980 *Zum Problem der Einfühlung*. Kaffke, Munich (1989 *On the Problem of Empathy* [Trans. Stein W] 3rd rev edn. *The Collected Works of Edith Stein*. Institute of Carmelite Studies, Washington, DC)
Stein E S 1970 *Beiträge zur philosophischen Begründung der Psychologie und der Geisteswissenschaften. Eine Untersuchung über den Staat*, Niemeyer, Tübingen, Germany. This contains (1) (1919, 1922) "Psychische Kausalität" (Psychic causality); (2) (1920, 1922) "Individuum und Gemeinschaft" (Individual and Community); (3) (1921, 1925) "Eine Untersuchung über den Staat" (An Investigation of the State)
Theunissen M 1977 *Der Andere. Studien zur Sozialontologie der Gegenwart*. Walter de Gruyter, Berlin & New York
Toulemont 1962 *L'Essence de la Société selon Husserl*. Presses Universitaires de France, Paris
Walther G 1922 Ein Beitrag zur Ontologie der sozialen Gemeinschaften mit einem Anhang zur Phänomenologie der sozialen Gemeinschaften, *Jahrbuch für Philosophie und phänomenologische Forschung*, *VI* (Sonderdruck). M. Niemeyer, Halle, Germany

K. Mulligan

Philanthropy, Economics of

Philanthropy is defined as benevolent behavior, usually in the form of charitable gifts, toward others in society. Charitable giving accounts for a significant fraction of income in the USA and other nations, and is often directly encouraged through government tax policy. This entry discusses the motives, causes, and influences of philanthropic behavior. Special emphasis is placed on how government policy, through the charitable deduction on income taxes and through direct grants to charities, can affect the amount of philanthropy in the economy.

1. Introduction

In the late 1990s the media tycoon Ted Turner announced a gift of one billion dollars to the United Nations, software titan Bill Gates donated five billion dollars toward health and education in the developing world, and Oseola McCarthy, an 89-year-old laundrywoman from Mississippi, gave her life savings of $150,000 to establish scholarships for black college students in her state. These are examples of philanthropic, or charitable, behavior—individuals freely giving money to help others. Most people are, in fact, philanthropists, although not on the scale of the three mentioned above. Through the last few decades of the twentieth century, over two-thirds of Americans donated money to charity in any given year, with households giving over two percent of income on average. In 1997, those that gave to charity averaged donations of over $1,000. In fact, people spend more on charitable giving than they do on electricity, telephones, or car insurance.

In addition, the government is an active partner with private donors in funding the charitable sector of the economy. It does this in two ways. First, are direct grants to charities. In 1994, for instance, charitable

organizations reported receiving $51.9 billion in government grants. The second way the government spends money on charity is by excluding income spent on charitable donations from income tax. The US tax code was modified in 1917 to allow taxpayers to deduct their charitable donations from their taxable incomes, and by 1998 individuals were claiming over $86 billion in charitable deductions.

This entry will describe how economists think about and do research on philanthropy, what economists know about charitable giving, and how public policy influences it. The focus will be primarily on gifts of money and assets by individuals, although corporations and other institutions also make charitable gifts (see Boatsman and Gupta 1996), and people also give by volunteering time (see Brown and Lankford 1992). In addition, the discussion will center on philanthropy in the USA, since a large majority of the academic research has centered on the US experience (for information on charitable organizations internationally, see Anheier and Salamon 1996). Finally, the entry will highlight what challenges remain ahead for the study of philanthropy.

2. *An Economist's View of Charitable Behavior*

Economics is founded on the view that people are self-interested. Economists assume that since individuals have a choice in how they behave, they must always make the choices that they think are the best at the time. This assumption, called the 'axiom of rationality,' provides the bedrock from which economic models of behavior are formed.

Applying this framework to philanthropic behavior, we can ask why would people who work hard for their money simply turn around and give it away? Does this not contradict the assumption that individuals are self-interested? Maybe. But by the axiom of rationality, we should first look for ways in which an individual believes that giving to charity is the best thing to do with that money at that time.

One possibility is that people desire more of the service provided by the charity. National Public Radio (NPR), for instance, is a charitable organization that survives largely on donations. Over a million people a year give to NPR, with most giving under $100. Are these givers getting their money's worth? Certainly not. Anyone can listen to NPR broadcasts, even without donating, and anyone who gives $100 is unlikely to notice any increase in broadcast quality as a result. It follows that a self-interested person is better off giving nothing and taking a 'free ride' on the donations of others. Hence, this alone cannot be an adequate explanation for why people give to charity (Andreoni 1988).

A second reason is that individuals may be getting something directly from the charity in exchange for their contributions. For instance, big donors to the opera may get better seats, or donors to a university may get buildings named for them. While this must surely matter for large donors, most individuals only receive tokens (a coffee mug?), if anything, in exchange for donations.

A third reason could be that individuals get some internal satisfaction—a 'warm-glow'—from giving to their favored charity, and the more they give the better they feel (Andreoni 1989). This would mean that giving to charity is like buying any other good, such as ice cream. We easily accept that people have a natural taste for ice cream, so why not a natural taste for warm glow? If we accept this, then we can analyze charitable giving just like any other consumer good—when income goes up people should want more, and when the price goes up people should want less.

Scholars have examined preferences for giving from a number of different approaches, and it seems that warm glow is in fact a core economic motivation for giving (see Andreoni 1993). That is not to say it is the only reason for giving. People may also give from a sense of religious duty, from pressure at work or from friends, or as a signal of social status (Rose-Ackerman 1996). At the heart of all of these processes, however, seems to be a basic human interest in helping others or doing one's share. This, of course, raises a deeper question that may require a broader social science perspective to answer: why do people enjoy acting unselfishly? This is a question that, fortunately, we do not have to answer to continue with this article. We can take as our launching point that people get joy from giving, and we can apply our economist's tools for analyzing consumer behavior to begin studying charitable giving.

3. *The Facts about Philanthropy in the USA*

How do we learn about charitable giving? One source is surveys of consumers. The Independent Sector for instance, surveyed about 2,500 households by telephone. Surveys are valuable since they can obtain information on age, education levels, and other personal characteristics of respondents. A disadvantage is that individuals must rely on imprecise memories when answering questions, or may be reluctant to give true information about their incomes or donations.

Table 1
Sources of private philanthropy, 1998

Source of gifts	Billions of dollars	Percent of total
Individuals	134.8	77
Foundations	17.1	10
Bequests	13.6	8
Corporations	9.0	5
Total for all sources	174.5	100

Source: Giving 1999

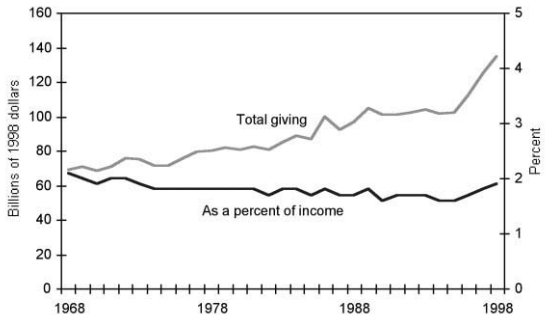

Figure 1
Giving by individuals, 1968–98[a]. Dollars are inflation-adjusted to 1998 values.
Source: Giving 1999

A second important source is samples of tax returns. Since individuals who itemize their tax returns in the USA can take a charitable deduction, we can learn about donations for this sector of the economy. The advantage of tax returns is that the information on them is precise, as long as people do not cheat on their taxes. (Slemrod 1989 explored this potential problem and found that, while there is some evidence of cheating by overstating charitable deductions, the effects are small and do not appreciably affect the analysis.) The disadvantage to this data is that tax returns contain very little information about the personal characteristics of the filers that would be helpful in explaining giving, such as education levels or religious affiliation, nor can we learn about the giving habits of those who do not itemize their tax returns. Since no data source is perfect, economists must conduct many studies on many data sources in order to uncover the 'facts' on charitable giving.

3.1 Sources, Totals, and Trends in Giving

Charitable donations can come from individuals, charitable foundations, corporations, or through bequests. While all are significant, by far the dominant source of giving is from individuals. Table 1 shows that in 1998 individuals gave over 134 billion dollars to charity, or 77 percent of the total dollars donated. The second biggest source, foundations, was responsible for only 10 percent of all donations.

The trends in giving over the last 30 years can be seen in Fig. 1. Total giving has been on a steady rise,

Table 2
Private philanthropy by income, age, and education of the giver, 1995

	Percent of households who give	Average amount given by those who give	Percent of household income
All contributing households	68.5	1,081	2.2
Household income			
Under $10,000	47.3	324	4.8
$10,000–$19,999	51.1	439	2.9
$20,000–$29,999	64.9	594	2.3
$30,000–$39,999	71.8	755	2.2
$40,000–$49,999	75.3	573	1.3
$50,000–$59,999	85.5	1,040	1.9
$60,000–$74,999	78.5	1,360	2.0
$75,000–$99,999	79.7	1,688	2.0
$100,000 and above	88.6	3,558	3.0
Age of giver			
18–24 years	57.1	266	0.6
25–34 years	66.9	793	1.7
35–44 years	68.5	1,398	2.6
45–54 years	78.5	979	1.8
55–64 years	71.7	2,015	3.6
65–74 years	73.0	1,023	2.9
75 years and above	58.6	902	3.1
Highest education of giver			
Not a high school graduate	46.6	318	1.2
High school graduate	67.2	800	1.9
Some college	74.1	1,037	2.1
College graduate or more	82.3	1,830	2.9

Source: Author's calculations, data from Independent Sector 1995

Table 3
Private philanthropy by type of charitable organization, 1995

	Percent of households who give	Average amount given by those who give	Percent of total household contributions
Type of charity			
Arts, culture, humanities	9.4	221	2.6
Education	20.3	335	9.0
Environment	11.5	110	1.6
Health	27.3	218	8.1
Human services	25.1	285	9.5
International	3.1	293	1.1
Private and community foundations	6.1	196	1.4
Public and societal benefit	10.3	127	1.7
Recreation	7.0	161	1.4
Religion	48.0	946	59.4
Youth Development	20.9	140	3.8
Other	2.1	160	0.3

Source: Author's calculations, data from Independent Sector 1995

with especially big jumps coming in 1996, 1997, and 1998. When measured as a percentage of income, however, giving seems much more stable. Since 1968 giving has varied from 1.6 percent to 2.1 percent of income. In the most recent year shown, 1998, giving ticked up to 2.1 percent of income, the highest in 30 years.

3.2 Who Gives?

Surveys reveal that 68.5 percent of all households gave to charity in 1995 and that the average gift among those giving was $1,081. Table 2 shows that the more income a household has, the more likely the household is to give to charity, and the more it gives when it does donate. This table also reveals an interesting pattern typically found in charitable statistics. Those with the lowest incomes give over four percent of income to charity. As incomes grow to about $50,000, gifts fall to 1.3 percent of income, but then rise again to 3.4 percent for the highest incomes.

What could cause this 'U-shaped' giving pattern? One explanation is that those with low incomes may be young people who know their wages will be rising, hence they feel they can afford more giving now. It may also be due to the composition of the types of charities to which people give, since lower-income people tend to give significantly more to religious causes. Hence, it will be important to account for all the factors that may explain giving before offering explanations for the averages seen in these tables.

Table 2 also illustrates that giving varies significantly with the age and educational attainment of the givers. As people get older they are typically more likely to give to charity and to give a greater fraction of their incomes. Likewise, those with more education give more often, give more dollars, and generally give a higher fraction of income. Note that the table does not show a smooth acceleration of giving with age. Again, age, education, and income all vary with each grouping in the table and will have to be considered jointly.

3.3 What Do They Give To?

In 1997 over 45,000 charitable, religious, and other nonprofit organizations filed with the US government. Table 3 attempts to categorize these charities by the types of service they provide. This reveals that, among all types, households are most likely to give to religious organizations and to give them the most money—48 percent of all households give to religion and 59 percent of all charitable dollars go to religion.

4. Taxes and Giving

How can taxes encourage philanthropic behavior? Taxpayers in the USA who itemize their deductions can deduct their charitable giving from their taxable income (and from 1982 to 1986 the law also allowed nonitemizers a limited charitable deduction). If someone in the 15 percent marginal tax bracket gives $100 to charity and takes a charitable deduction, this person will save $15 in taxes, making the net cost of the gift $85. If the person was in the 31 percent tax bracket then the $100 gift would have a net cost of $69. Thus, the higher one's tax rate, the lower the net cost of giving. In this way the tax rate acts as a subsidy rate to giving, and those in higher tax brackets get bigger subsidies. (Note that state income taxes often add additional subsidies. See Feenberg 1987.)

Since the tax deduction subsidizes giving, we should expect that those with higher subsidies will make larger contributions. It is also natural to expect that those with higher incomes will make larger contri-

butions. However, those with higher incomes will typically also have higher subsidies, so when we see richer people giving more how can we know whether their income or the subsidy is causing the change? If we can compare people with similar subsidy rates but different incomes, and similar incomes but different subsidy rates, we can apply statistical techniques to separate the influences of income and the subsidy.

Disentangling these is, of course, very important for policy makers. If, for instance, the subsidy does not increase giving at all, then the policy is costing the government tax dollars without benefiting the charities. If, on the other hand, the subsidy lifts giving significantly, it may be an effective tool to increase social well-being. Generally, policy makers look for the increase in contributions due to the subsidy (the benefit) to be larger than the loss in tax dollars (the cost). The next section discusses how economists explore this issue.

4.1 The Effect of the Subsidy on Giving

Economists began conducting detailed empirical studies of giving in the 1970s in an attempt to isolate the effect of the subsidy on giving (see Clotfelter 1985 for a review). You can imagine asking the counterfactual policy question this way: consider a proposal that would allow a taxpayer to deduct only 70 percent (or 50 percent or 20 percent, etc.) of his or her charitable contributions for tax purposes, rather than the 100 percent allowed now. This means that the net cost of a $100 contribution would rise. For someone in the 28 percent marginal tax bracket, for instance, the cost would rise from $72 to $80. We would expect that this higher cost would discourage giving, but by how much?

The first generation of studies to explore this counterfactual relied on cross-sectional data, primarily from tax returns. Observing income and deductions for thousands of people in a given tax year, the researchers used the variations in income and cost to estimate how, on average, individuals would respond to such a change in policy. There were several limiting features of the data used in these studies. Of particular importance is that the samples, for reasons of confidentiality, typically did not include extremely wealthy people.

Depending on the source of the data, the years studied, the size of the sample, and the statistical approach, the results naturally varied. It has been generally agreed, however, that the prediction that best characterizes the results is that a policy that would increase the cost by 10 percent for all taxpayers—without affecting the other income taxes paid by the individual—would reduce contributions by about 13 percent (Clotfelter 1990). This suggests that behavior is quite responsive to the incentives offered by the tax deduction, and that the response is large enough to satisfy the policy criterion noted above. In particular, if the government pays an additional 10 percent of the cost of giving and if people respond by giving 13 percent more, then the policy generates more new charity than it costs the government in tax revenue.

These studies also confirmed that giving increases with income. If after-tax incomes were to rise by 10 percent, estimates are that giving would rise by about 8 percent (Clotfelter 1990). Notice that this quantifies the left-hand part of the U-shaped giving curve discussed in Sect. 2—since giving rises by a lesser percentage than income, giving as a percentage of income (all else equal) must be declining.

Later studies were able to supplement these findings using data from surveys rather than tax returns (see Clotfelter 1990, Steinberg 1990). Although the survey results varied, they generally confirmed the policy predictions stated above. In addition, they revealed the importance of other individual characteristics in explaining giving. Most strikingly, they consistently found that giving increases with the age and the education of the giver, again as suggested in the tables above. It could be that those who are educated have more interest in charity or that interest in charity changes with age. A more plausible explanation, however, is that researchers have no information on individual wealth. Since wealth is correlated with both age and education, it is likely that the influence of wealth is being filtered through these other variables.

A second generation of studies, conducted mostly in the 1990s, has begun to shift the consensus view noted above. Unlike the earlier studies that relied on cross-sectional data, these newer studies used panel data, that is, data with observations on the same set of individuals over several years. Panel data is generally seen to be superior to cross-sectional data. The reason is that by observing the same people at different points in time the researcher is able to get a more precise measure of how behavior responds to changes in the environment.

The most influential study with panel data is by Randolph (1995). Using tax returns on about 12,000 people from 1979 to 1988, and using statistical techniques designed explicitly for panel data, he identified a much weaker effect of the subsidy and a much stronger effect of income. His results indicate that a policy that would increase the cost of giving by 10 percent would reduce giving by only about 5 percent, whereas if income were to rise by 10 percent, giving would rise by 11 percent. Others using panel data have found similar effects. Hence, at the end of the 1990s the literature on the effect of the subsidy to giving was in a state of flux. More studies with panel data are needed to reconcile the early and later findings, and to reach a new consensus on the effects of the subsidy.

Finally, these studies cannot reveal all of the subtle and complex influences on giving. Other things outside of these studies clearly matter, such as the social circles

to which people relate, the religious organizations to which they belong, and the activities of fundraisers. While many of these influences lie beyond the scope of economic analysis, when economists have examined social factors the effects seem to be small and do not appreciably change the predictions for policy (Andreoni and Scholz 1998). Still, studies that examine noneconomic influences could lead to new insights on policies toward giving.

4.2 The Effects of Tax Reforms on Giving

During the 1980s there were two major tax reforms that dramatically changed the two important variables discussed above: the cost of giving and after-tax income. Prior to 1980 there were 25 tax brackets, with the top bracket at 70 percent. This meant that a wealthy giver, after the deduction, would pay only 30 cents of each dollar donated. The first tax reform came in 1981. This created 14 tax brackets, with the top rate falling to 50 percent. The second major tax reform, the Tax Reform Act of 1986 (TRA86), was more dramatic. It reduced taxes to two rates, 15 percent and 28 percent. (Actually, it also included a 'bubble bracket' of 33 percent for those between $71,900 and $192,930 of annual income, before returning to 28 percent for incomes above this. A 31 percent bracket was added in 1991. Subsequent changes in the law mean that in 1999 there are five brackets: 15 percent, 28 percent, 31 percent, 36 percent, and 39.6 percent.) Hence, most people, but especially those in high tax brackets, had the net cost of their charitable contributions rise.

With the costs of giving going up, many feared that there would be calamitous decline in giving. Auten et al. (1992) and Randolph (1995) looked at the experience from the 1980s and made several interesting discoveries. First, people were sensitive to the anticipation of an increase in their cost. In 1981, for instance, wealthy people saw that the cost of a $100 donation would rise to $50 in 1982 from $30 in 1981, and thus moved up a sizable share of their giving to 1981. For those earning over one million dollars (in 1991 inflation-adjusted dollars), average charitable deductions rose by $32,000 from 1980 to 1981, but fell by $64,000 in 1982. This same sector also anticipated the TRA86, raising deductions by $37,500 in 1986, then reducing them by $52,500 in 1987. Similar, although less pronounced, patterns can be seen in most other income groupings as well. In fact, Randolph (1995) shows econometrically that individuals are far more responsive to short-term (or temporary) changes in the cost of giving than they are to long-term (or permanent) changes.

How did giving change overall in response to the tax reforms? As seen in Fig. 1, dollars given tended to rise over the 1980s, despite this increase in the cost for most individuals. But this does not tell the whole story. First, even though the cost of giving went up, after-tax income also rose for middle-income people, and this should counter the effect of raising the costs. In addition, incomes in general were rising throughout the 1980s. Also, this analysis cannot account for the behavioral response of charities that may have increased their fundraising efforts in response to the tax changes.

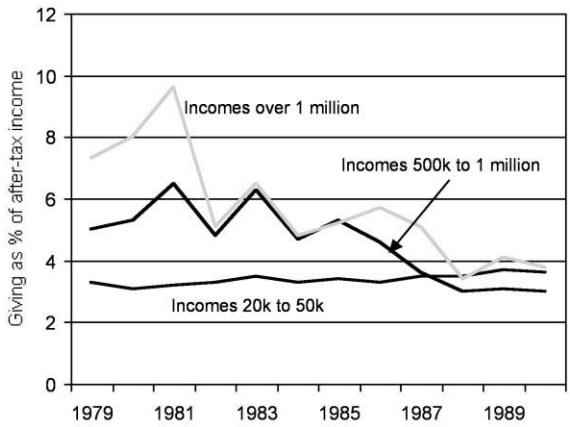

Figure 2
Changes in the Distribution of Giving: 1979–90
Source: Auten et al. 1992

Fig. 2 shows one more piece of information that is masked by the averages; that is, that the distribution of giving across income groups also changed. The figure shows that giving by the middle-income group, which is by far the most populous, actually increased as a percentage of income over the 1980s. However, the giving by the wealthy fell quite substantially. To balance this, however, far more people were entering the ranks of the super-rich—the number of tax filers with incomes over one million dollars more than tripled from 1980 to 1989.

4.3 Giving by the Very Wealthy

Auten et al. (2000), in a study of giving by the wealthy, report that in 1994 the 1 percent of US households with the highest income was responsible for over 16 percent of all dollars given. As one might imagine, describing giving by the wealthy in more detail can be complicated, in part because many give through trusts and carryover deductions through several tax years. The most distinguishing characteristic of wealthy givers is perhaps that a great deal of their giving is of appreciated assets, such as art, real estate, or stocks. Donations of appreciated assets are often a heavily tax-favored form of giving (O'Neil et al. 1996). First, the giver can take a deduction on income taxes for the current market value of the asset, not the purchase price, in the year the gift is made. In addition to this, the giver does not have to pay any tax on the capital

gains made on the asset. This means gifts of appreciated property often generate two forms of tax savings. Those who had taxable incomes over $2.5 million in 1995 had average charitable deductions of $248,069, 44 percent of which was appreciated property. People with annual incomes of $50,000–$100,000, by comparison, deducted $1,823 on average, of which only 15 percent was appreciated property.

Another distinguishing characteristic of wealthy donors is that their giving is highly variable from year to year. In any given year the median contribution of those making over $2.5 million is only 0.7 percent of income—half the rate of those in the $50,000–$100,000 range. But the 95th percentile gift of the rich is almost 21 percent of income—twice the rate of those in the $50,000–$100,000 range. When averaged over a longer period, however, these numbers become a lot more similar. This indicates that perhaps the super-rich save up their donations over a number of years in order to give bigger gifts when they do give. Why? One reason might be to gain recognition from the charity, such as a brass plate, for being a major donor. Another reason is that large donors are often asked to serve as board members of charities, hence larger gifts can lead to greater control over the future of the organization.

5. *Do Government Grants Displace Private Philanthropy?*

There are several reasons to suspect that government grants to charities might depress private donations. First, if givers are aware of the grants, then they may perceive the charity as less in need of their donations, leading them to keep the money or give it to a different cause. This is known as the 'crowding out' hypothesis (Warr 1982, Roberts 1984, Bergstrom et al. 1986). A second reason giving may decline is that the charity may be less aggressive in fundraising after receiving a government grant. Both effects would lead to government grants to offset private philanthropy.

There are also reasons to think that government grants could have the opposite effect. Citizens who are unsure about the quality of a charity could see a grant as a 'stamp of approval.' Alternatively, the grant could be used as 'seed money' that will allow the charity to expand into a bigger operation by, for instance, constructing new buildings or opening branches in different cities (Andreoni 1998). In both of these cases, government grants could actually encourage private philanthropy.

From an empirical standpoint, the question of crowding out has been a challenge, and no study to date has been definitive. Three recent studies illustrate the difficulties of this area, and the wide array of findings. Kingma (1989) studied a special sample of givers to National Public Radio (NPR) in 1986. This data is unique in that it ties each giver directly to his or her local NPR station and the grants it receives. He finds that if a typical station were to receive a $10,000 grant, private giving would fall by about $1,350, that is 13.5 percent crowding out.

Two studies using panel data on donations received by charities came to different conclusions. Khanna et al. (1995) used observations on 159 charities in the UK and found that government grants actually encouraged rather than displaced private giving. Payne (1998), using a 10-year panel of 430 US charities, applied special statistical techniques to account for the fact that government grants to charities may be caused by the same needs and interests as private giving. Without accounting for this 'simultaneity,' researchers may underestimate crowding out. Payne's analysis predicts about 50 percent crowding out.

This is clearly a difficult area to study, and all of the researchers must do the best they can with incomplete data. Moreover, researchers have thus far been unable to account for how charitable fundraising, apart from givers's willingness to donate, responds to government grants. These analyses await future study.

6. *Conclusions and Future Research*

Philanthropy is a significant factor in the US economy, accounting for about 2 percent of income over the past 30 years. Government and private givers are in a partnership to fund the services of over 45,000 charitable organizations. This entry has discussed the economic influences on private philanthropy and how it interacts with government policy.

The most important finding is that individuals are indeed sensitive to the charitable deduction in the US tax system. By subsidizing giving, the tax deduction has clearly increased giving by the private sector. Because policies and preferences change with time, new studies are always needed to assess the impact of the deduction. In the 1970s and 1980s, dozens of studies led to a consensus view that the subsidy was highly effective—a 10 percent reduction in the cost of giving led to a 13 percent increase in giving. In the 1990s, new data and new econometric analysis have challenged this view, suggesting that a 10 percent reduction in the cost leads to a 5 percent increase in giving. More studies are needed to clarify the picture about how responsive givers are to the charitable deduction.

The literature on the crowding out of private giving by government grants is also far from settled—some say the crowding out is substantial, at 50 percent, while others find government grants encourage rather than displace private giving. This research makes clear that the study of philanthropy needs to be expanded to consider the role of the charitable organizations more actively in the analysis. Not only do individuals respond to changes in tax policy, but fundraisers do too. How do charities adjust to changes in tax rates or government grants, and how responsive are indivi-

duals to fundraising efforts? Exploring these questions will deepen our understanding of philanthropy and help sharpen government policy.

Of course, to accomplish these important research goals, new and better data will be needed. There has not been a large national survey of individual philanthropic behavior since 1973, and another survey could yield many new insights. In addition, data is needed that ties together the actions of fundraisers, private donors, and the government. Such data would allow researchers to learn about the whole market for charitable giving—both supply and demand—and could help form policies to strengthen the partnership between the public and private sector for providing much needed charitable services.

See also: Altruism and Self-interest; Arts Funding; Income Distribution; Wealth Distribution; Welfare Programs, Economics of

Bibliography

Andreoni J 1988 Privately provided public-goods in a large economy: The limits of altruism. *Journal of Public Economics* **35**: 57–73

Andreoni J 1989 Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy* **97**: 1447–58

Andreoni J 1993 An experimental test of the public-goods crowding-out hypothesis. *American Economic Review* **83**: 1317–27

Andreoni J 1998 Toward a theory of charitable fund-raising. *Journal of Political Economy* **106**: 1186–213

Andreoni J, Scholz J K 1998 An econometric analysis of charitable giving with interdependent preferences. *Economic Inquiry* **36**: 410–28

Anheier H K, Salamon L M 1996 *The Emerging Nonprofit Sector: An Overview*. St. Martin's Press, New York

Auten G E, Cilke J M, Randolph W C 1992 The effects of tax-reform on charitable contributions. *National Tax Journal* **45**: 267–90

Auten G E, Clotfelter C T, Schmalbeck R L 2000 Taxes and philanthropy among the wealthy. In: Slemrod J (ed.) *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*. Harvard University Press, Cambridge, MA, pp. 392–424

Bergstrom T, Blume L, Varian H 1986 On the private provision of public goods. *Journal of Public Economics* **29**: 25–49

Boatsman J R, Gupta S 1996 Taxes and corporate charity: Empirical evidence from microlevel panel data. *National Tax Journal* **49**: 193–213

Brown E, Lankford H 1992 Gifts of money and gifts of time: Estimating the effects of tax prices and available time. *Journal of Public Economics* **47**: 321–41

Clotfelter C T 1985 *Federal Tax Policy and Charitable Giving*. The University of Chicago Press, Chicago

Clotfelter C T 1990 The impact of tax reform on charitable giving: A 1989 perspective. In: Slemrod J (ed.) *Do Taxes Matter? The Impact of the Tax Reform Act of 1986*. MIT Press, Cambridge, MA, pp. 203–42

Feenberg D 1987 Are tax price models really identified: The case of charitable giving. *National Tax Journal* **40**: 629–33

Giving USA 1999 *Annual Report*. American Association of Fund-Raising Counsel, New York

Independent Sector 1995 *Giving and Volunteering in the United States*. Independent Sector, Washington, DC

Khanna J, Posnett J, Sandler T 1995 Charity donations in the UK: New evidence based on panel-data. *Journal of Public Economics* **56**: 257–72

Kingma B R 1989 An accurate measurement of the crowd-out effect, income effect, and price effect for charitable contributions. *Journal of Political Economy* **97**: 1197–207

O'Neil C J, Steinberg R S, Thompson G R 1996 Reassessing the tax-favored status of the charitable deduction for gifts of appreciated assets. *National Tax Journal* **49**: 215–33

Payne A A 1998 Does the government crowd-out private donations? New evidence from a sample of non-profit firms. *Journal of Public Economics* **63**: 323–45

Randolph W C 1995 Dynamic income, progressive taxes, and the timing of charitable contributions. *Journal of Political Economy* **103**: 709–38

Roberts R D 1984 A positive model of private charity and public transfers. *Journal of Political Economy* **92**: 136–48

Rose-Ackerman S 1996 Altruism, nonprofits, and economic theory. *Journal of Economic Literature* **34**: 701–28

Slemrod J 1989 Are estimated tax elasticities really just tax evasion elasticities? The case of charitable contributions. *Review of Economics and Statistics* **71**: 517–22

Steinberg R 1990 Taxes and giving: New findings. *Voluntas* **1**: 61–79

Warr P G 1982 Pareto optimal redistribution and private charity. *Journal of Public Economics* **19**: 131–8

J. Andreoni

Philosophy, Psychiatry, and Psychology

From the very beginning, scientific psychiatry has raised fundamental questions within the fields of philosophy and psychology, namely as regards the nature of the 'psyche.' The question of what it means to say that something is 'psychical' leads to speculation in psychiatry insofar as special heuristic aspects are involved.

1. Philosophical Aspects of Psychiatry

In accordance with Brentano's basic idea that the psychical can be identified with the 'intentional' (Brentano 1874), one might argue that psychiatry is so important to philosophers because the disturbances of intentionality in various domains of the mental world of humans are what might be termed 'negative images' of the psychic functions which normally operate in the human mind. The special method by which Karl Jaspers (Jaspers 1913) elaborated this view of the human mind is the 'phenomenological approach' to various kinds of disturbances of mental functioning in psychiatric patients (deficiency model of psychic dimensions in Jaspers's *General Psychopathology*).

However, this is not the only way in which the relationship between psychiatry and philosophy can

be construed. The other aspect is the issue of the potential role of psychic 'normalcy' in social life. As a fluctuating criterion for assigning people to a psychiatric context, normalcy cannot be regarded as a phenomenon of 'average' and represents merely the concept of 'ideal typology' (Blankenburg 1971). From a philosophical point of view, the subject of psychiatric 'normalcy' constitutes a major problem, since it raises the issue of the existence of an 'individual norm' in the sense of the 'alterity of the other' (Lévinas 1987). Seen in this way, psychiatric disorders are not simply 'defects' but represent variants of normal psychic life which—when exaggerated—are interpreted as disorders and treated by psychiatrists.

This concept fits in with the approach of evolutionary psychiatry which considers that psychiatric disorders have evolved because they relate to psychic functions which—in their submaximal expression—yield advantages where natural selection is concerned. Viewed in this way, psychopathology does not constitute a 'statistical, average norm' but an 'individual norm,' as has been pointed out by Blankenburg (1971). In a similar sense, Spaemann and Löw (1981) have argued that health itself contains a teleological notion: 'If 99 percent of the population suffered from headaches, headaches still would not be ''normal''.' Philosophical concepts in psychiatry therefore have to address the problem of 'normalcy'; they face the difficult challenge of acknowledging deviant cognitive and mental events merely as expressions of the alterity of the other.

2. *Aspects of Greek Philosophy*

The focus of Greek philosophy is on the individual, thinking and acting intentionally, and self-reflection is thus the cornerstone of ancient Greek thought. On the basis of the hypothesis of quasi-determinism, the Athenian philosopher Socrates (470–399 BC) asserted that the will responds to the dictates of reason. He developed the Socratic—or 'maieutic'—method which elicits truths that lie hidden in an individual's unconscious by means of a process of question and answer. This method—in a completely different context—later became the basis of Sigmund Freud's psychoanalysis and the nondirective psychotherapy of Carl R. Rogers. Socrates' disciple Plato (427–347 BC) differentiated—for example, in his *Phaedo*—between the spiritual and material worlds and applied this dichotomy to man by postulating that the psyche (soul), together with its logical functions, possesses a pre-existence before the birth of the individual and survives into immortality (Plato 1966). In *Timaios*, Plato defines madness (*manteia*) as a disturbance of the vegetative components of the soul (which are located in the liver) but in the later *Phaidros* he tells us that it can also be sent by the gods as an elevating spiritual force. The aim of Plato's philosophical psychology is to create harmony between the different psychic activities. Whereas Plato's *idealism* defines ideas as the only true reality, his pupil Aristotle (384–322 BC) based his psychology partly on empirical investigation, thus representing a type of *realism*. In his systematic discussion *De anima*(*On the Soul*), Aristotle describes the soul as the *entelechy*, the self-contained purpose of the body. This purpose must be fulfilled in order to achieve *eudaimonia* (bliss). In Aristotelian thinking, philosophical self-reflection leads to psychological self-actualisation, and in his *Ethica Nikomacheia* Aristotle states that the purpose of ethics is not knowledge but action (Aristotle 1984).

Although Aristotle was the most systematic Greek philosopher and had an enormous influence on the development of modern science, the treatment of psychological matters in post-Aristotelian Hellenistic philosophy is often described as eclectic. Stoic and Epicurean philosophy, in particular, must be considered as practical psychology. Given that the increasingly urbanized Hellenistic world had a greater resemblance—in some ways—to modern society than the middle ages, this practical psychology is still of interest. Epicureans and Stoics tried to find ways to alleviate anxiety, but their anxiety is less fear of the gods than fear and disturbance caused by a complex but fragile society. Epicurean and especially Stoic thinking played an important role in Hellenistic culture and can thus be viewed as the first widespread popular psychological fashions.

After the fall of the Roman Empire, Christian Europe lost sight of some of Aristotle's major works. The philosophical psychology of the dominant neo-Platonism referred back to the static aspects of Platonism, shifting away from rationalism to metaphysical speculation on the relationship between a tangible and an intangible world, between the soul and God. St. Augustine (354–430) attempted to synthesize neo-Platonism with Christianity. Following on from Seneca and Plotinus, he leaned towards introspection as his philosophical and psychological method, considering that—alongside objective observation—self-observation of inner mental processes could function as the basis of knowledge. He therefore chose self-reflection as the basis of his epistemology, arguing that doubt presupposes the existence of the doubter (Augustine 1984), an idea that later became extremely influential as Descartes famous *Cogito ergo sum*.

3. *Metaphysics of Middle Ages and the Mind–Body Problem*

The philosophical psychology of the Middle Ages was metaphysical and—at the same time—deductive. The changeability of the soul and of world were interpreted philosophically as an inevitable but inferior property of the unchangeable, and consequently Christian dogmatism and neo-Platonism showed no interest in

the experience of psychological interaction. With the foundation of the first universities from the thirteenth century onwards, the process of philosophical debate increased and a more dynamic philosophical psychology developed as Aristotelian thinking was resuscitated through the medium of Arab philosophy. St. Thomas Aquinas (1225–74) attempted to integrate Christian dogma with elements of Greek and Arab philosophy. Aristotle differentiated between an active immortal intellect and a merely passive understanding one. Aquinas accepted this view, describing the intellect as a passive power, a *tabula rasa* which had to be filled with experience by the five exterior senses. In addition, there exists a rational power—which makes free will possible—and the soul also has vegetative components. In contrast to the neo-Platonic dichotomy of soul and body, Aquinas attempted to define man as a psycho-physical unity, as a body inspired by a soul. Taking his lead from Aristotle, Aquinas defined the soul as the form of the body and as an immortal spiritual entity. According to this theory, the body of a dead person consists of the same material or substance as that of a living person. As a result of this dichotomy, it is impossible to describe the impact of mental acts on real action without becoming entangled in serious logical difficulties. This is the *mind–body problem*: the issue of whether or not the soul or mind is an incorporeal substance, a nonmaterial entity obeying laws other than those which govern the material world. Discussion of this problem has continued until the present day.

4. *Modern Philosophical Psychology*

Modern philosophical discussions related to psychiatric problems already at its beginning. René Descartes (1596–1650) reflected on the fundamental nature of illusions and the constitution of reality. He posed the questions of certainty of reality and compared the reality of dreams with the consciousness in wakefulness. In his *Meditations on first Philosophy* Descartes arrived at the irreducibility of the cogito, putting an end to the infinite regress of doubt about possible fictional realities (Descartes 1996). As a consequence of this concept the modern subjectivistic understanding of consciousness has evolved, leading to the split between *cogitatio* and *extensio*, between soul and body, between the mental and the physical world, a split which represents one of the unresoluble fundamental problems in psychiatry. In this sense the dualistic understanding of the mind–body problem represents one of the central challenges, difficult to overcome in philosophical psychiatry.

Refuting earlier deductive speculation, the empirical philosophy of John Locke (1632–1704) again emphasized the Aristotelian view that the human mind is like a blank sheet of paper and that all the material of reason and knowledge comes from observation, experience, and reflection. David Hume (1711–76) described ideas as the copies of sensory impressions. Most of our beliefs go beyond current impressions, but all that we can be certain of is perception, meaning that it is impossible, for example, to comprehend the *self* as a separate identity. Radicalizing this point of view, the positivism of Auguste Comte (1798–1857) postulated that the sensible comprises the whole range of knowledge and that any knowledge that cannot be comprehended by the senses must be rejected.

Immanuel Kant (1724–1804) in his precritical anthropology speculated about constitutional aspects of psychiatric disorders. In his C*ritique of Pure Reason* he tried to deduce the *a priori* conditions of the possibility of experience (Kant 1974). This type of categorical analysis of mental life has been decisive for all subsequent philosophical considerations about the formation and constitution of subjectivity and consciousness.

In this sense idealism represents one of the possible forms of reactions to Kantian philosophy. For example, Georg Friedrich Hegel (1770–1831) developed an ontological approach to intellectually reconcile all elements of existence with one another within a systematic structure in which all contradictory moments of reality appear as dialectic mediations of the universal mind (Hegel 1973). The philosophical psychologist Franz Brentano (1837–1917) elaborated a comparatively modest conceptualization of the difference between the mental and the physical, thereby developing a fundamental concept of philosophy of the psyche, namely the concept of 'intentionality.' In his work *Psychology from an Empirical Standpoint*, he argued: 'Each mental phenomenon is characterised by what the scholars of the middle ages called the intentional (also probably mental) inner existence … of an object, and which we … would call the reference to content, direction toward an object' (Brentano 1874).

One may argue that the main characteristics of psychiatrically ill patients are disturbances of intentionality in Brentano's understanding of this term, that is, psychic functions are 'intentional' insofar as they refer to an ('internal') object; they can 'mean' something, can refer to something. If this fundamental function of mental life is disturbed, i.e., the *intentio recta* is disturbed, a psychiatric disorder can be anticipated.

The Danish philosopher Sören Kierkegaard (1813–55), deeply aware of Hegel's philosophical concept of the absolute and at the same time vigorously writing against it, developed a concept of subjective reality, prestructuring the later existentialistic conceptualizations of subjectivity, defining it as synthesis of soul and body, carried by spirit. His pioneer work about personal freedom formulated the questions as to *The Concept of Anxiety* (Kierkegaard 1980). As has been pointed out by Hersch (1989) these basic ideas are represented by 'existence' and 'moment,' in which the

subject works through to the moment of the 'reality of freedom,' the condition of 'anxiety.' That means that 'anxiety' cannot be understood primarily as an emotional state but rather an existential condition, an intensified form of existence, leading to an autonomous subject ('choosing the self'). Kierkegaard influenced mainly the French school of philosophical psychology (Sartre, Merleau-Ponty, Camus), whereas Brentano founded, together with his student Husserl, the German school of 'Phenomenology,' essential to the history of psychiatry (Husserl 1950). In this context 'Phenomenological reduction' has to be understood as the attention to the 'nature of the condition of things' and thus the 'things as phenomena.' A great and far-reaching concept in philosophical psychology, the philosophy of the 'language game,' has furthermore been elaborated within the philosophy of language by Ludwig Wittgenstein (1898–1951). Jaspers has pointed out the importance of philosophy of language, emphasizing that language is the 'work of man.' The universals of language also play a role in the philosophy of Arnold Gehlen (1904–76) who refered to man as a 'non-determined animal,' whereby Gehlen interprets the human nature as creating its own world as an 'open world' (Gehlen 1987). Of special importance for psychiatric philosophical psychology is the concept of Helmuth Plessner (1892–1985) who developed the idea of 'positionality' of human subjects. Whereas animals show a 'centred positionality,' human beings live, according to Plessner, in an 'eccentric positionality' (Plessner 1970). This concept plays a relevant part in psychiatric psychopathology, since disorders like delusions of reference and ideas of persecution may be interpreted as a disordered eccentric positionality (Emrich 1992). In Jean-Paul Sartre's (1905–80) *Being and Nothingness*, the 'decentering of the world' plays a fundamental role in the phenomenological analysis of the 'glance' (Sartre 1996). Sartre's analysis of the interpersonal relationship ends up in an intersubject fight of 'glances.' Later Emmanuel Lévinas (1905–95) acknowledged the 'otherness of the other' in the dimension of 'face': 'The epiphany of the face is itself a word of honour in a certain way' (Lévinas 1987). This view of responsibility in the interpersonal encounter is of great importance in psychiatry; it conceptualizes an understanding of empathy in therapeutic relationships.

As a new academic discipline, psychology was obliged to assert itself against the established materialist sciences, especially medicine and biology. Positivist thinking consequently led to strongly anti-mentalist and anti-introspectionist tendencies in psychology. The theory that it is only via *objective* observation that scientific method can formulate hypothetical laws from which it can then derive deductive arguments led to the radical behaviorist position that nothing that is the product of *subjective* introspection can be accepted. Positivism thus reduced psychology to a science that deals with behavior and which aims to predict and control that behavior. As a result of behaviorism as represented by John Watson (1878–1958) and B. F. Skinner (1904–90), academic psychology drifted away from philosophy, becoming more practical than theoretical and more concerned with predicting and controlling behavior than with understanding it (Watson 1998, Skinner 1999). The contemporary philosophy of science, however, would claim that it is an illusion to believe that science can be reduced to a simple, *objective*, and entirely rule-bound methodology. Thomas Kuhn argued that such opinions are based on science textbooks that simply fail to tell the truth about the history of science (Kuhn 1996). In contrast to academic psychology, psychiatry has retained a more heterogeneous—even if contradictory—theoretical background. The main question at the start of the twenty-first century is whether psychology and psychiatry can only be governed by certain particular scientific methodologies or if it is possible to accept the existence of different approaches to the acquisition of knowledge. Paul Feyerabend pointed out that 'the idea that science can, and should, be run according to fixed and universal rules, is both unrealistic and pernicious' (Feyerabend 1980). In actual fact, therapeutic practitioners—in particular—find it unsatisfactory to consider introspective experience and objective behavior as mutually exclusive, as if they were distant continents with no contact between them. Modern therapeutic theories are thus becoming increasingly eclectic, and the gulf which used to exist between introspective psychoanalysis and behavior therapy, for example, is in the process of disappearing.

5. *Self and Identity*

Philosophical concepts of the 'self' have led to psychological consideration of the possible nature of 'identity.' Disturbances of the inner equilibrium of the components of the self which shape identity, brought about by processes of re-presentating by remembering, therefore make clear the potential for explanation inherent in a theory of the processes of identity formation based on a psychology of the self. As Lévinas in particular has shown (Lévinas 1987), however, this cannot be restricted to a theory of the structure of the 'self'; rather it must take account of interpersonality, the existence of others (the 'trace of the other' referred to in the title of Lévinas' book). Any attempt to describe the danger to processes of identity formation posed by the existence of the other from the perspective of the psychology of the self must take account of the 'ontology of the in-between' already developed by Martin Buber (1994). For, as Theunissen (1965) demonstrates in his major work on intersubjectivity, taking the other seriously in its own being involves transcending the intentionality mode.

This philosophical/psychological concept is based

on the assumption that components of a person's self can manifest themselves in others and can exist further in them.

The basis of this primary interpersonal constitution of the psyche is the mother–child dyad of early childhood. In a way, people unconsciously always retain parts of their childhood dyad 'within them,' although they nevertheless consider themselves to be ego-centered 'individuals.' In a system of ideas in which the *cogito* is identified with the empirical ego, the 'interpersonal self' cannot be conceived. This primary (but extremely repressed or internally never realized) interpersonal constitution of the self has consequences for how people deal with what is familiar and unfamiliar, both in the other and within themselves, where aggression is concerned. Christian Scharfetter, referring to Graumann's theory of subjective perspectivity, rightly refers to the fact that precisely the most intimate personal area of a person, his own unconscious, constitutes an alien entity, an unfamiliar area of sinister, threatening mystery. Scharfetter (1994) puts this as follows: 'The other is not simply an alien entity, however. The other may very well be familiar; it can be close, it can be experienced as analogous (not homologous). Indeed, the alien is able to strengthen and reinforce the self. It is through the 'You' that the 'I' recognizes itself. This is an important factor where psychotherapy is concerned. However, confrontation with the alien within the self is painful; it requires not only cognitive efforts (Hegel's 'strenuous toil of conceptual reflection'), but also psychological effort in the form of crises in identity development.

An extremely effective defense mechanism here is that of 'projective identification.' This means that the unknown, the sinister, within the subject is discovered in others by projection and combated there. Streeck-Fischer says, for example, that 'the confrontation with the alien has become a shock experience, one which hampers successful interaction with the alien (Streek-Fischer 1994). The alien is threatened as a means of stabilizing one's own threatened self.' As Streeck-Fischer shows, this mechanism becomes particularly effective, is in a sense multiplied enormously, when one's own position is not only under pressure from the danger posed by crises in the formation of identity but is also reinforced because of social pressure. Social devaluation of specific features of the self can significantly increase projective identification in the sense of the development of aggression. Hans-Joachim Maaz, for example, shows how in East Germany 'social constraints under a repressive authoritarian system led to the self becoming alien and the alien becoming one's own.' Maaz describes the resulting aggression as follows (Maaz 1994): 'violence and hatred of other people are always evidence of internal tensions within someone, of his no longer being able to find and develop the self and the essential but of having to live a life too much of which is imposed by others.'

One of the great figures in the theory of mimesis is René Girard. In various works, notably *La violence et le sacre*, Girard developed a theory of mimetic occurrences (Girard 1977). If one accepts his thinking, the main point of the Oedipal drama, which in Freudian theory is said to mold early childhood, is not the Oedipal aspect as such—the desire for the mother in the Freudian sense—but the mimetic rivalry with the father. In an early work of 1961, *Mensonge romantique et vérité romanesque*, Girard elaborated on a discovery central to the entire development of his theory. He was able to identify a phenomenon of 'mimetic desire' in nineteenth and twentieth century novels from Stendahl to Flaubert and Proust. The novels he examined shared a peculiar basic structure: the main characters desire someone else, but their feelings are directed toward a third person, a '*mediateur*.' Analyzing Stendhal, Girard says: 'in most cases of Stendhalian desire, the *mediateur* himself desires the object, or could desire it. This desire, whether real or assumed, is precisely what makes the object infinitely desirable in the eyes of the subject.'

Desires, needs, preferences, and yearnings are thus described as being something not genuinely given, something brought forth internally by drives or make-up, but are 'read,' taken on from a third party because of mimetic dependency. The basic events in the process of identification can thus be described as follows: if we can understand the other in a recreative sense only by putting ourselves into the other's intentions, into that person's world of desires, this means that we have to identify with that other. The amazing consequence of this is that processes of identification are necessary for a proper interpretation of reality. In this sense, one can say that a rational explanation of the world is possible only if there is a fundamental willingness to identify with others.

As a result, one can argue that in psychiatric patients this rational world access is hampered by their impaired interpersonal relations, that is, due to disturbed mimetic processes (Emrich 1995).

See also: Anthropology, History of; Cognitive Functions (Normal) and Neuropsychological Deficits, Models of; Mental Health and Normality; Mind–Body Dualism; Personality Theory and Psychopathology; Phenomenology in Human Science; Phenomenology: Philosophical Aspects; Psychiatry, History of; Psychology: Historical and Cultural Perspectives; Self: History of the Concept; Theory of Mind

Bibligography

Aristotle 1984 *The Complete Works of Aristotle: The Revised Oxford Translation*. Princeton University Press, Princeton, NJ

Augustine 1984 *City of God*. Penguin, Harmondsworth, UK

Blankenburg W 1971 *Der Verlust der natürlichen Selbstver-*

ständlichkeit. Ein Beitrag zur Psychopathologie symptomarmer Schizophrenien. Ferdinand Enke, Stuttgart, Germany
Brentano F C 1874 *Psychologie vom empirischen Standpunkt*, Vol. 1. Duncker & Humblot, Leipzig [1995 *Psychology from an Empirical Standpoint*. Routledge]
Buber M 1994 *Ich und Du*. Lambert Schneider, Gerlingen [1978 *I and Thou*. Scribner, New York]
Descartes R 1996 *Meditations on First Philosophy: With Selections from the Objections and Replies* (trans. Cottingham J). Cambridge University Press, Cambridge, UK
Emrich H M 1992 Systems theory of psychosis: 'Filtering,' comparison, error correction, and its defects. In: Emrich H M, Wiegand M (eds.) *Integrative Psychiatry*. Springer, Heidelberg, Germany pp. 81–89
Emrich H M 1995 Physiognomy of the psychological—toward a theory of mimesis. *Paragrana* **4**(2): 126–43
Feyerabend P K 1980 *Against Method*. Verso, London
Gehlen A 1987 *Der Mensch*. Aula Verlag, Wiesbaden [1988 *Man*. Columbia University Press]
Girard R 1961 *Mensonge romantique et vérité romanesque*. Grassett, Paris
Girard R 1977 *Violence and the Sacred*. Johns Hopkins University Press, London
Hegel G W F 1973 *Phänomenologie des Geistes*. Suhrkamp, Frankfurt [1979 *Phenomenology of Spirit*. Galaxy Books, New York]
Hersch J 1989 *Das philosophische Staunen*. Piper, München, Germany
Husserl E 1950 *Gesammelte Werke*. Kluwer, Dordrecht, The Netherlands
Jaspers K 1913 *Allgemeine Psychopathologie*. Springer, Berlin [1998 *General Psychopathology*. Johns Hopkins University Press, Princeton, NJ]
Kant I 1974 *Kritik der reinen Vernunft*. Suhrkamp, Frankfurt [1996 *Critique of Pure Reason*. Hackett, Indianapolis, IN]
Kierkegaard S 1980 *The Concept of Anxiety: A Simple Psychologically Orienting Deliberation on the Dogmatic Issue of Hereditary Sin*. Princeton University Press, Princeton, NJ
Kuhn T S 1996 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
Lévinas E 1987 *Totalität und Unendlichkeit*. Versuch über die Exteriorität. Alber, München [1969 *Totality and Infinity: An Essay on Exteriority*. Duquesne University Press, Pittsburgh, PA]
Maaz H J 1994 Das Eigene und das Fremde im deutschen Vereinigungsprozeß. In: Egner H (ed.) *Das Eigene und das Fremde: Angst und Faszination*. Walter Verlag, Hamburg, pp. 13–27
Plato 1966 *The Collected Dialogues of Plato*. Pantheon, New York
Plessner H 1970 *Philosophische Anthropologie*. Fischer, Frankfurt
Sartre J P 1996 *Being and Nothingness*. Pocket Books, New York
Scharfetter C 1994 Im Fremden das Eigene erkennen. In: Egner H (ed.) *Das Eigene und das Fremde*. Walter Verlag, Hamburg, pp. 13–27
Skinner B F 1976 *About Behaviorism*. Pimlico, London: 1999; Random House, New York: 1976
Spaemann R, Löw R 1981 *Die Frage Wozu? Geschichte und Wiederentdeckung des teleologischen Denkens*. Piper Verlag, München, Germany
Streek-Fischer A 1994 'Haßt du was, dann bist du was'. Über Fremdenhaß und seine selbstreparative Funktion am Beispiel jugendlicher rechtsextremer Skinheads. In: Egner H (ed.) *Das Eigene und das Fremde*. Walter Verlag, Hamburg, pp. 13–27
Theunissen M 1965 *Der Andere*. Walter de Gruyter, Berlin
Watson J B 1998 *Behaviorism*. Transaction, New Brunswick, NJ

H. M. Emrich and W. Paetzold

Phonetics: Articulatory

Articulatory phonetics is the branch of phonetics concerned with describing the speech sounds of the world's languages in terms of their articulations; that is, the movements and/or positions of the vocal organs (articulators). There were independent early Sanskrit, Arabic, and Western European traditions of articulatory description, often motivated by a desire to preserve pronunciations of an earlier stage of a language, e.g., for ritual purposes. More recently, it flourished in the nineteenth century, motivated by such goals as spelling reforms, dictionary making, language teaching, and language and dialect description, and continues in this tradition to the present.

1. *Phonetic Transcription*

A system for articulatory phonetic description comprises the set of vocal organs deemed to be important for speech, and the set of their movements or positions capable of characterizing the various speech sounds. A. M. Bell's *Visible Speech* (1867), a speech teaching aid for the deaf, used quasi-iconic phonetic symbols to notate positions of the lips, tongue, and soft palate. However, the most influential system of articulator-based notation has been that of the International Phonetic Association (IPA), founded in 1886. In the IPA system, the symbols are not articulatorily iconic, being instead based on Roman letters, but they encode a well-developed system of articulatory description.

1.1 *International Phonetic Alphabet*

The articulators specified in the IPA system are the lungs, the larynx, the two lips, the upper surface of the oral cavity from the teeth back to the uvula (divided into alveolar ridge, hard palate, and soft palate), the uvula, the pharynx and epiglottis, the tongue (divided into the tip, blade, front, back, root, and sides), and the vocal folds (or cords). These are divided conveniently according to their functions in speech as follows.

Respiratory (or initiatory): The respiratory system provides an outward pulmonic airstream with a roughly constant pressure; however, since other articulators, such as the larynx, can also be used to initiate airflow, this is sometimes also called, more generally,

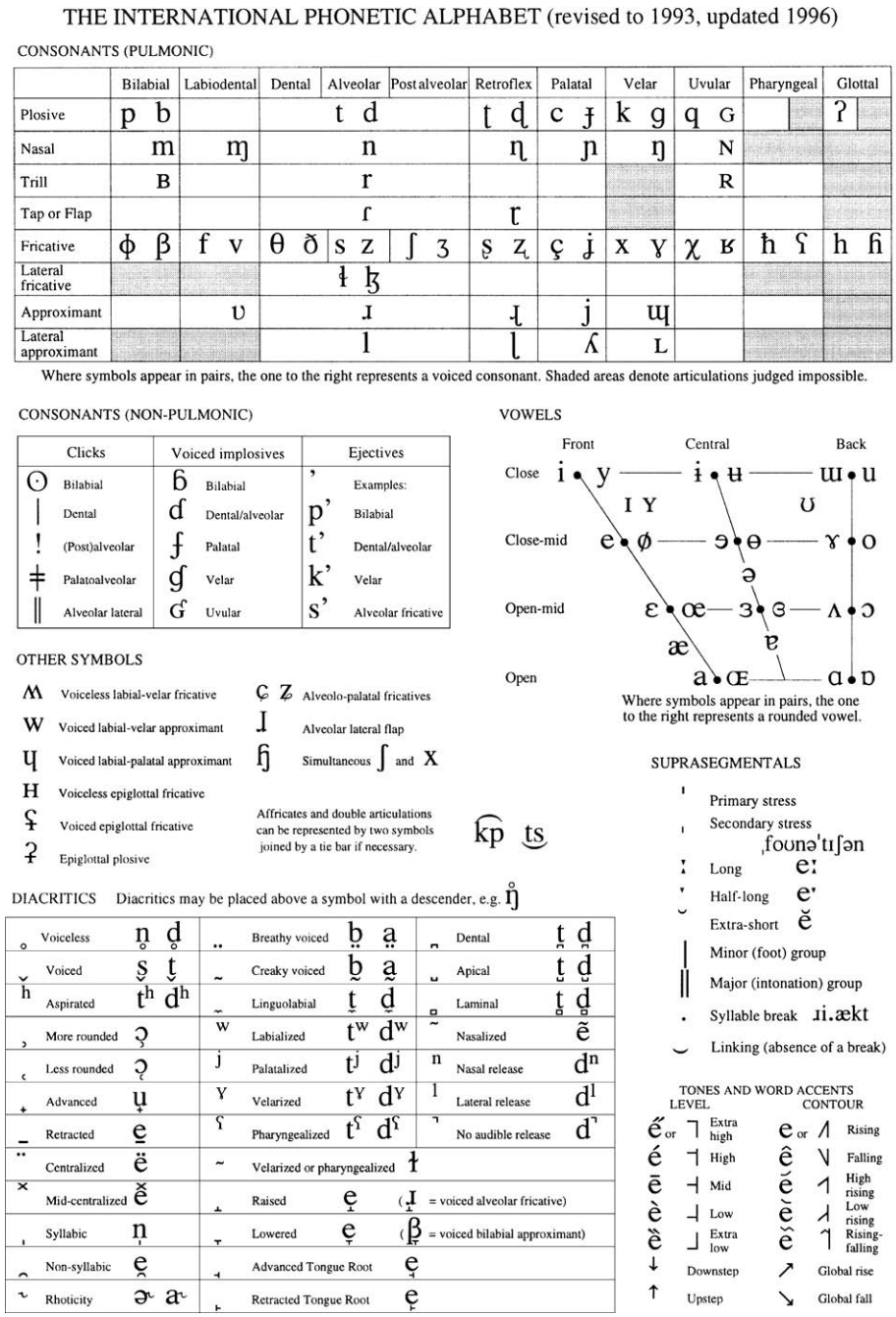

Figure 1
The 1996 IPA Chart (Source: International Phonetic Association)

the airstream process. Sounds are described in terms of the source and direction of movement of their airstream.

Phonatory: Vibration of the vocal folds (which, with their control structures, lie inside the larynx) modulates the pulmonic airstream to produce a sound source called voicing; the rate of this vibration is perceived as the pitch of the voice, and the mode of vibration is perceived as voice quality. Sounds are described in terms of the presence/absence, and if

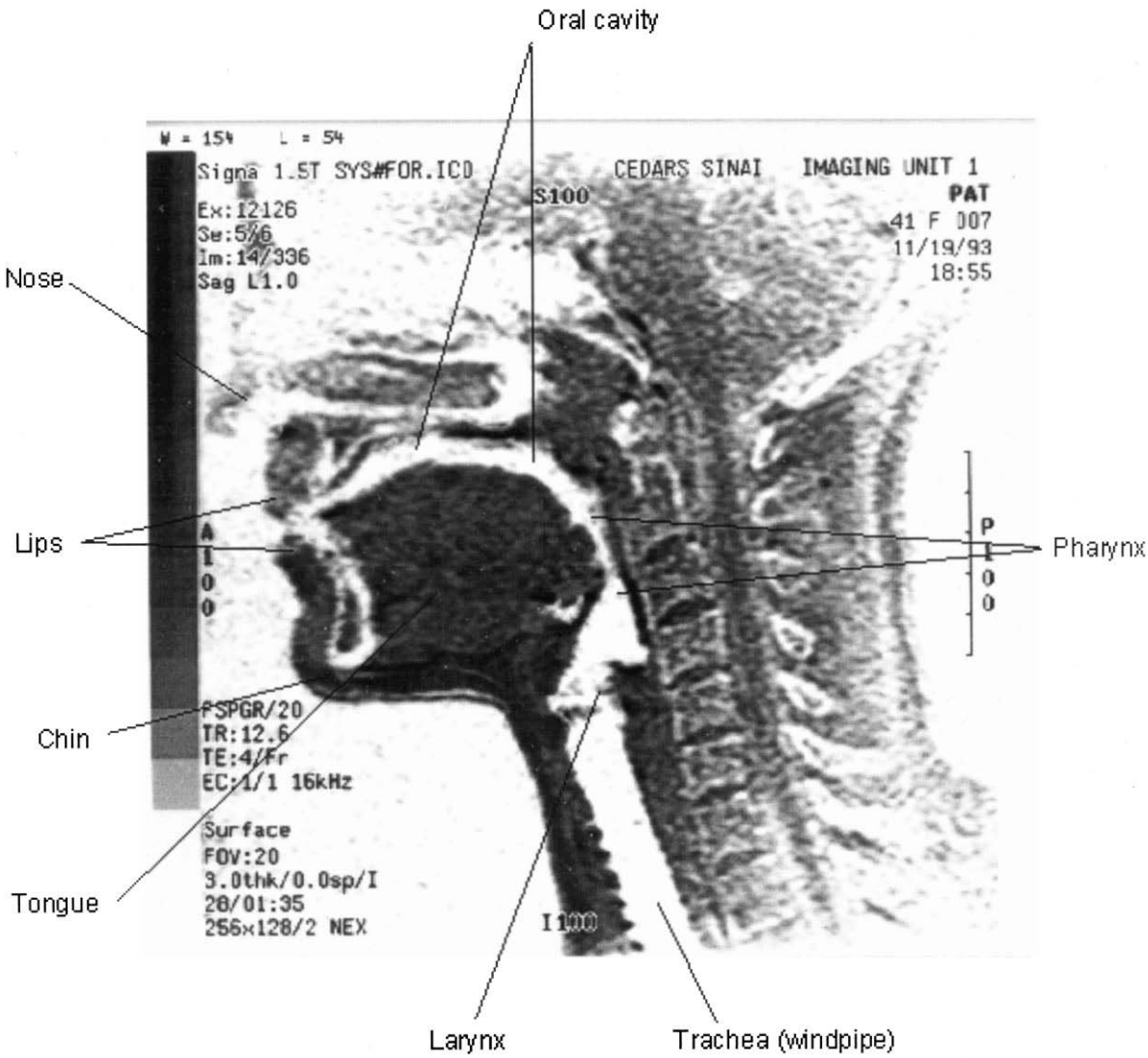

Figure 2
MR Image of midline of vocal tract during American English [s]. White portions are airspaces, black portions are structures, some of which are labeled. (Source: Fromkin 2000. Reproduced with permission)

present, the quality, of voicing; larger spans of speech are described in terms of the pitch of voicing (e.g., tones, intonation—see *Suprasegmentals*).

Articulatory (in a narrower sense): Structures above the larynx, including the pharynx walls, the tongue (and its subparts), the jaw, the velum (or soft palate—sometimes classified separately as the oro-nasal process), the uvula, and the lips, which move within the vocal tract to modify the shape, and therefore the resonances, of the airway. In addition, stationary articulators—upper teeth and hard palate (and its subparts)—provide contact sites for these movable articulators. Within a given sound, the articulators may be classified as active vs. passive, with the active articulators moving toward the passive and often forming a constriction or narrowing along the vocal tract. (While stationary articulators are necessarily passive, movable articulators may be active or passive.) For most sounds, the extreme or target positions of the active articulators relative to the passive are taken to suffice to characterize the sounds. For example, labiodental means active lower lip to passive

upper teeth; velar means active tongue body to (presumably) passive soft palate (the velum); a front vowel has the tongue as a whole to the front, relative to the surface of the palate. These locations must be further qualified according to how close the articulators come together. For example, in a stop consonant, they touch, whereas in a vowel they are still relatively far apart. These two dimensions of the location of the active articulator are sometimes called the location or place, vs. the degree, stricture, or manner, of the constriction.

The IPA uses these articulatory dimensions (and others not presented here) both to describe speech sounds, and as a basis for defining phonetic symbols to represent those sounds. These dimensions serve as a row and column labels on the well-known IPA chart, the most recent (1996) version of which is reproduced here as Fig. 1. This chart comprises two consonant charts, a vowel chart, and lists of other symbols. Each symbol represents the combination of articulatory properties expressed by the row and column labels of its chart. Square brackets are usually used to show that a symbol is part of a phonetic alphabet. Thus the first symbol on the IPA chart, [p], represents a *bilabial* (column label—made with the two lips approaching each other) *plosive* (row label—an oral stop; i.e., made with complete closure between the active and passive articulators, and the soft palate raised to prevent nasal airflow) which is *voiceless* (leftmost symbol in its cell—made without vocal fold vibration).

The consonant chart can be seen as a grid representing the articulatory space, with sounds made in the front of the mouth located toward the left of the chart, and sounds made in the throat (pharynx) located toward the right of the chart, and the degree of openness of the vocal tract given from top to bottom of the chart. The vowel chart's layout also gives a similar articulatory grid. However, rather different articulatory descriptions are given for consonants vs. vowels, even when the same articulators are involved, as can be seen from the different row and column labels. The basic descriptive dimensions for vowels are the overall vertical and horizontal position of the active articulator (the tongue), and the position of the lips. While the position of the tongue is judged implicitly relative to passive structures, these are not referred to overtly; only the active articulator, the tongue, is discussed. Thus what might be called a 'close palatal vowel with bilabial approximation' is instead nowadays generally called a 'high (close) front rounded vowel.' Not all phonetic systems make this kind of functional differentiation of vowel vs. consonant articulations, and the issue remains controversial.

The IPA system aims to provide a symbol for every phoneme of every language. However, it is a crucial property of this system that the descriptive framework is independent of the particular symbols chosen to fill the cells; other symbols could be substituted for these as long as they were defined appropriately, and the charts provide places for sounds that, while possible, have not been assigned symbols.

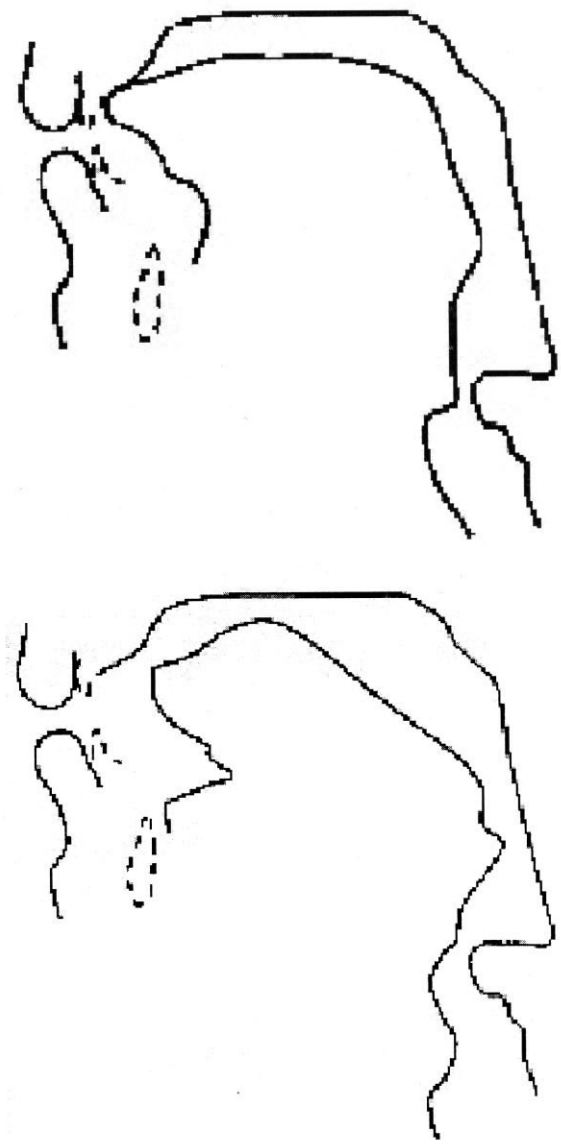

Figure 3
Tracings from MR images of midline of vocal tract during American English [1] and American English sound spelled 'r.' (Sources: Narayanan and Alwan 1997, Alwan and Narayanan 1997. Reproduced with permission)

For further explication of this chart and the IPA system, including its resources for description of suprasegmental phonetics (see *Suprasegmentals*) and of the range of speech sounds found in the world's

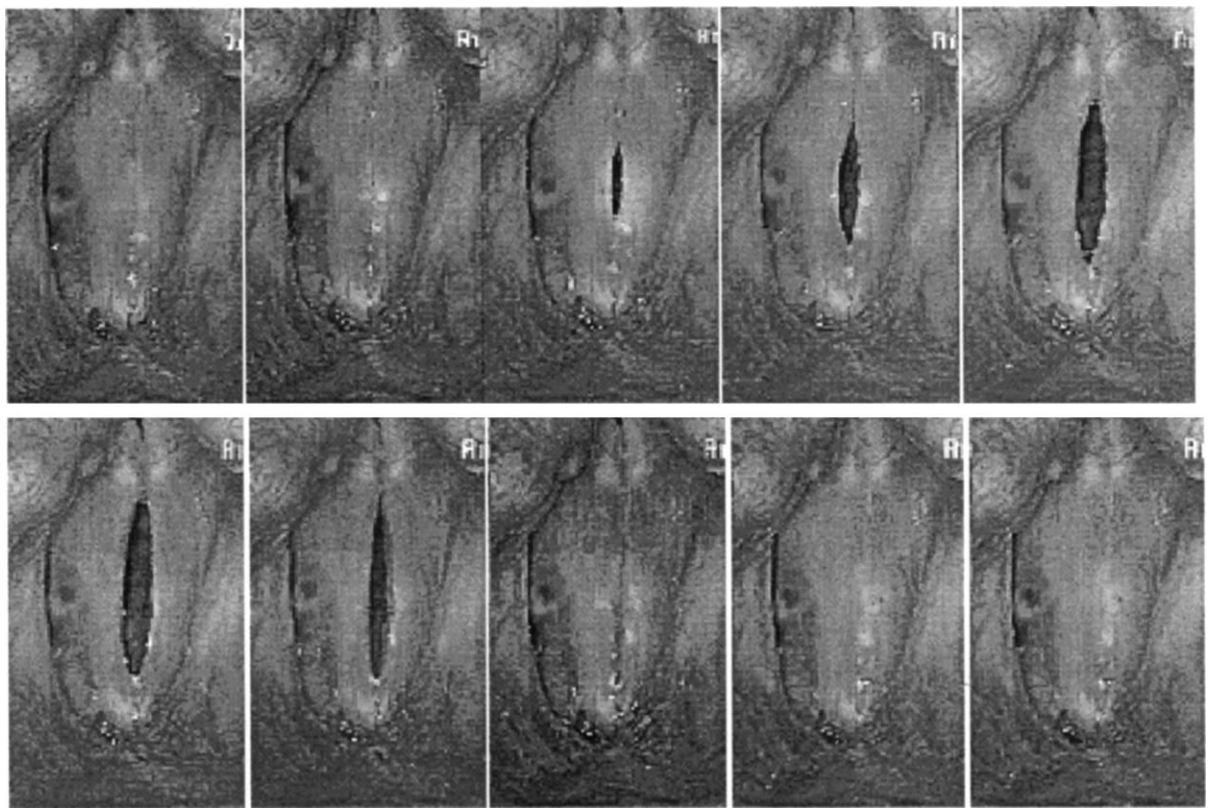

Figure 4
Frames from a video stroboscopy recording of vibrating vocal folds. (Source: Ladefoged 2001. Reproduced with permission)

languages, see the *Handbook of the International Phonetic Association* (IPA 1999), a phonetic textbook such as Ladefoged (2000), or a phonetics chapter such as that in Fromkin (2000). For more on how these and other articulatory dimensions can be used to classify speech sounds, see *Phonology*.

1.2 Nonsegmental Transcription

Speech tends to consist of alternating closings (significant constrictions) and openings of the vocal tract; these are called consonants and vowels respectively. It is a fundamental observation that the actions of the various articulators are somewhat asynchronous. That is, it is generally not the case that all the articulators move into and hold their positions for one sound, then move into and hold their positions for the next sound, and so on. The dominant traditions of phonetic description, including the IPA, assume that speech consists of sequences of separate speech segments whose diverse articulations can overlap in time. It is possible to represent much of this overlap with a phonetic alphabet (e.g., Kelly and Local 1989) but it requires cumbersome use of diacritical markings around the alphabet symbols. In contrast, competing articulatory phonetic systems do not bundle the articulatory properties into separable but overlapping segments, nor do they assign phonetic symbols to such segments. Instead, they try to record the temporal articulatory patterns of speech by using more complex representations (see, e.g., Anderson 1985; also *Phonology, Nonlinear*).

2. Study of Articulation

In most systems for articulatory phonetic description, both the identities of the articulators in general, and their positions in particular sounds are often presented in terms of static diagrams based on side-view X-rays or other images. Figure 2 gives a Magnetic Resonance image (MRI) of the vocal tract (the part of the body used in sound production), along its midline, of a speaker of American English making the fricative consonant[s]. This scan shows the larynx, the pharynx, the tongue, and the oral cavity, along with some other landmarks. Figure 3 shows tracings of the tongue and

other relevant structures made from other MRI images of the same speaker, but in this case during the English lateral approximant consonant [l] (which looks similar to [s]) and the English central approximant consonant spelled 'r' (a sound with an unusual tongue shape found in few of the world's languages). Note that, though such pictures are static, that does not mean that the articulations are necessarily held for any appreciable length of time in real speech—think of this as the frame, from a movie, which shows the most extreme displacement. In some sounds, such as stop consonants, articulations are held; in others, such as trills, movements of articulators are crucial to sound production. Direct evidence about other aspects of speech articulation can be gained using other laboratory techniques, from those dating to the nineteenth century and earlier, to the latest medical advances. For example, Fig. 4 shows successive frames from a video recording of vocal fold vibration taken with a video stroboscope.

Much of the current effort in articulatory data collection is focused on model building. Models vary in their physicality (whether they model the movements of realistic structures controlled by muscles, or whether they perform more abstract tasks), but all seek to gain insight into the kinematic and dynamical behavior of speech, either as an instance of refined motor control, or as an important part of language.

Articulatory activity ultimately serves the acoustic ends of sound production and propagation. Each speech sound has its defining acoustic properties that result from its defining articulatory properties. The different configuration of articulators define different resonance patterns of sounds, and interruptions and modifications of the airstream differentiate sounds in other ways. Static articulatory data, especially 3-D imaging of the entire vocal tract, is used in building models of the relation of articulation of acoustics—how the articulatory positions determine the acoustic properties. Models such as those of Fant (1960), and now Stevens (1998), can be applied in articulatory synthesis of speech. There is some effort to recover articulatory phonetic information from acoustic signals so that it can be used in speech recognition systems.

See also: Linguistics: Overview; Phonology; Speech Errors, Psychology of; Speech Production, Neural Basis of; Speech Production, Psychology of; Speech Recognition and Production by Machines; Suprasegmentals

Bibliography

Anderson S R 1985 *Phonology in the Twentieth Century*. University of Chicago Press, Chicago, IL

Bell A M 1867 *Visible Speech*. Simpkin, Marshall and Co., London

Fant G 1960 *Acoustic Theory of Speech Production*. Mouton, The Hague, The Netherlands

Fromkin V A 2000 Phonetics. In: Fromkin V A (ed.) *Linguistics: An Introduction to Linguistic Theory*. Blackwell, Oxford, UK, Chap. 11

IPA (International Phonetic Association) 1999 *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge, UK

Kelly J, Local J 1989 *Doing Phonology: Observing, Recording, Interpreting*. Manchester University Press, Manchester, UK

Ladefoged P 2000 *A Course in Phonetics*, 4th edn. Harcourt Brace College Publishers, Fort Worth, TX

Stevens K N 1998 *Acoustic Phonetics*. MIT Press, Cambridge, MA

P. A. Keating

Phonology

Phonology is concerned with the sound structure of words and utterances within individual languages, the way distinctions in sound are used to differentiate linguistic items within a language, and the ways in which the sound structure of the 'same' element varies as a function of the other sounds in its context. While both phonology and phonetics are concerned with the role of sound in natural language, they differ in that phonetics deals with the articulatory, acoustic, and perceptual properties of sounds from a general, language-independent point of view (see *Phonetics: Articulatory*), while phonology studies the ways in which these properties are distributed and deployed within particular languages.

1. Goals of Phonology

The study of a language's particular sound properties has a number of aspects. First, it must be able to characterize the language's *inventory*: which phonetically possible sound types occur in utterances in the language? Second, it must characterize matters of *contrast*: which of the phonetic differences that occur in the language can serve to distinguish utterances (words, sentences, etc.) from one another? Third is the matter of *contextual limitation*: even though some property P occurs in language L, are there some environments from which P is excluded? And when P is apparently a property of (some part of) some element but that element occurs in a position from which P is excluded, what other property—if any—appears in its place? And finally, there is the description of *alternation*: when the 'same' linguistic element appears in different overt forms in different environments, what systematic differences occur? What conditions govern the range of phonetically

distinct forms that can count as the 'same' word, morpheme, etc.? Different answers to these questions yield different phonological theories.

It should be noted that the present article is limited to the sound structure of spoken languages, and ignores the expression systems of manual or signed languages (see *Sign Language*). This is misleading in important respects; research has shown that most of the basic principles of spoken language phonology are also characteristic of the organization of the expression systems of signed languages as well (Coulter 1993). Just as words are composed of sounds, and sounds of component properties, signs are also composed from structured, language-particular systems of more basic constituent elements. Units such as the syllable have close parallels in signed languages. While there are clear differences that depend on modality, these appear on further examination to be relatively superficial. A comprehensive theory of phonology as a part of the structure of natural language ought to take these broader issues into account. Until quite recently, however, the possibility of deep structural parallels between speaking and signing has not been raised, and the discussion below reflects this (undoubtedly unfortunate) limitation as well.

1.1 Some History

Prior to the early twentieth century, studies of sound in language concentrated on the ways in which sounds are made (articulatory phonetics), often confusing the letters of a language's writing system with its sounds. Toward the end of the nineteenth century, however, increasing sophistication of measurement techniques made it possible to explore a much wider range of differences among sounds, and to lay out the structure of speech in vastly greater detail. Somewhat ironically, perhaps, the explosion of data which resulted from the ability of phoneticians to measure more and more things in greater and greater detail began to convince them that they were on the wrong track, at least as far as increasing their understanding of particular languages was concerned.

Much of what was found, for example, involved the observation that speech is continuous, such that whatever is going on at any particular moment is at least a little different from what has gone on just before and what will go on immediately afterwards. A full characterization of an utterance as a physical event requires the recognition of a potentially unlimited number of distinct points in time, but it is clear that our understanding of an utterance as a *linguistic* event is hindered, rather than helped, by the recognition of this continuous character of speech. Speech normally is represented as a sequence of a small number of discrete segments, strung out in consecutive fashion like beads on a string; such a *segmental* representation vastly facilitates the discovery of regularity and coherence, but it must be emphasized that there is no direct warrant for it in either the acoustic or the articulatory observable data of speech, and it constitutes a fairly abstract (though undoubtedly appropriate) theory of how sound is organized for linguistic purposes.

It is clear that the role of particular sound differences varies considerably from one language to another. Thus, in English, the vowel sound in the word *bad* is much longer than that in *bat* (more than half again as long), but such a difference in length is always predictable as a function of the following sound, and never serves by itself to distinguish one word from another. In Tahitian, in contrast, essentially the same difference in length is the only property distinguishing, for example, *paato* 'to pick, pluck' from *pato* 'to break out.' A theory of sound that attends only to physical properties has no way of clarifying the quite different functions these properties may have across various languages. This is not to suggest that phonetics is wrong, but rather that there is more to be said.

1.1.1 De Saussure and the 'phonemic principle.'

The great Swiss linguist Saussure (1916) was the first to stress that in order to understand the role of sound in language it is necessary to focus not (just) on the positive properties of sounds, but on their differences. He suggested that in the study of individual languages, as opposed to general phonetics, utterances should be characterized in such a way that two such representations might differ only in ways that could potentially correspond to a difference between two distinct messages in the language in question. Thus, since long and short vowels never (by themselves) distinguish two distinct utterances in English, the difference should not be indicated in that language; while for Tahitian, it must be. A representation with this property will be called *phonological*; it will obviously be specific to a particular language, and the distinctive elements that appear in it can be called the *phonemes* of that language.

While de Saussure enunciated this principle quite forcefully and persuasively, he provided few specific details of just what a phonological representation should look like. There are in fact a variety of ways in which his insight could potentially be realized, and much subsequent discussion in phonology hinges on these differences of interpretation.

1.1.2 The development of phonology as a theory.

Various individual investigators arrived at conclusions similar to de Saussure's about the importance of attention to language-particular sound contrasts. One of these was the Polish linguist Baudouin de Courtenay (1972), whose work actually antedated de Saussure's, but attracted little attention due to his isolation in Kazan. He developed a sophisticated view of the relation between phonetics and phonology both in individual grammars and in linguistic change. As

transmitted by his later students, Baudouin's views on the nature of the phoneme constituted an important strand in thinking about language as this developed in Russian linguistics in the early years of the twentieth century. This, in turn, provided the background from which the work associated with the Linguistic Circle of Prague grew in the 1920s and 1930s.

Two of the most prominent members of the Prague Circle were Trubetzkoy (1939) and Jakobson (1941). In their studies of Slavic languages and their histories, they stressed the notion that the collection of potentially contrastive sound types in a language was not simply an inventory, but a highly structured system. This system is organized in terms of a small number of mutually orthogonal dimensions (such as voicing, stop vs. continuant, nasality, etc.), each of which serves in parallel fashion as the basis of multiple contrasts. The notion that the fundamental terms of sound structure in language are these properties themselves and not (or at least not only) the complete sounds they characterize has remained an important component of most subsequent theorizing. The analysis of phonological structure in terms of its constituent basic contrasts, in turn, has served as a model for a variety of other disciplines in the Humanities and the Social Sciences apart from linguistics.

1.1.3 American structuralist phonology. Early thinking about sound structure in America was dominated by the anthropological interests of Franz Boas, and focused on an accurate rendering of the sound contrasts in the comparatively 'exotic' indigenous languages of the new world. Boas's student Edward Sapir, however, was concerned to place the study of language in the broader context of an understanding of the human mind and society. As such, he stressed (Sapir 1925) the notion that the elements of sound contrast in a language should be regarded as having a primarily mental reality, part of the speaker/hearer's cognitive organization rather than as external, physical events.

The rise of positivist views of science in the 1930s, and especially of behaviorist psychology, made Sapir's sort of mentalism quite marginal, and replaced it with more rigorous operational procedures for investigating notions of contrast. Especially associated with the ideas of Bloomfield (1933) and later structuralists such as Bloch and Harris (1951), the result was a theory of the phoneme based exclusively (at least in principle) on a set of mechanical manipulations of corpora of observed linguistic data, from which a set of contrasting minimal elements was to be derived. The central notion of this theory was a phonemic representation related to surface phonetic form in a way that would later be formulated explicitly as a condition of *Biuniqueness*: the requirement that given either a phonetic or a phonemic representation of an utterance in a given language, that could be converted uniquely into the other (disregarding free variation) without additional information. The perceived rigor of this notion led to its being widely taken as a model not only for the study of other areas of linguistic structure, but for other sciences as well.

1.1.4 Generative phonology. The phonemic theories of American Structuralists provided a way to characterize linguistic contrasts, the inventories of sound types used in a given language, and the ways sounds can be combined into larger structures, but other aspects of sound structure were less satisfactorily accommodated within those views. In particular, questions of the ways in which unitary meaningful elements change in shape according to their sound context (or 'allomorphy': see *Morphology in Linguistics*) failed to receive systematic treatment. Since any difference among sounds that could serve to contrast linguistic elements was *ipso facto* a difference between irreducibly basic terms, there was really no way to express the notion that a single item could take a variety of forms (as in the prefixes in *inefficient*, *imprecise*, *irregular*, *illegal*, etc.) except by simply listing the variants. Such a list is undoubtedly appropriate for cases such as the forms of English to be (*am*, *are*, *is*, *was*, *were*, etc.) which are unrelated to one another in form; but in many other cases, the variation is transparently systematic and a function of the sounds in the element's environment. This sort of variation was recognized by structuralist phonologists, but relegated to marginal status.

Beginning with work of Morris Halle, a student of Jakobson, linguists began to question the centrality of surface contrasts in sound structure. The result was a new view that allowed morphophonemic regularities as well as more superficial ones to be accommodated within a phonological description. The success of this more abstract notion of sound structure in dealing with hitherto irresolvable problems in the description of stress (see *Suprasegmentals*) contributed greatly to its success, and the resulting theory of Generative Phonology as developed in the work of Halle together with Noam Chomsky rapidly became the dominant view in the field by the middle of the 1960s.

1.2 Phonology as a System of Rules and Representations

A basic insight in the development of Generative Phonology was the proposal that it is not only the representation of linguistic elements in terms of basic contrasts that matters: an adequate theory must characterize what a speaker knows about the sound system of the language, and that includes regularities of variation and alternation as well as inventories of basic elements. Combining these two aspects of phonological knowledge required the explicit recognition of

a system of rules (expressions of regular patterns of contrast and variation in sound shape) in addition to the theory of representations. Developing an adequate theory of phonological rules, in turn, necessitated a notion of phonological representation that was related to surface phonetic reality in a much more complex and indirect way than the phonemic representations of structuralist linguistics. The central problems of phonological theory came to be articulated in terms of the theory of rules, their nature, and their interaction, and the underlying phonological representations that need to be posited in order to allow them to be expressed in their full generality.

2. *Issues in Phonological Theory*

As reviewed in Sect. 1.1 above, the problem of phonological description was originally conceived as a matter of discerning the range of contrastive sound units in a language, and arranging them in a system that brings out the dimensions of their contrast. When phonology is seen as a form of knowledge, however, as generative phonologists have stressed, the sound structure of natural language takes on quite a different form and presents quite different problems. Among these are (a) the nature of underlying ('phonological' or 'phonemic') representations as well as surface ('phonetic') representations; (b) the ways in which phonological regularities serve to relate phonological to phonetic representation, including the interactions that may obtain among regularities; and (c) the relation between phonological form and other aspects of linguistic knowledge, such as word structure (see *Morphology in Linguistics*) and sentence structure (see *Syntax–Phonology Interface*). Various aspects of these questions are addressed below.

2.1 *The Abstractness of Phonological Representation*

Initial reaction to the proposals associated with generative phonology centered on its abandonment of a phonemic representation based on the condition of bi-uniqueness. Relaxing this defining characteristic of Structuralist phonemics led to phonological representations that were considerably more abstract in their relation to phonetically observable properties. The proposals of Chomsky and Halle (1968) concerning the analysis of English, for example, involved positing final 'silent *e*' in words like *ellipse*, *burlesque* (phonologically /Ellipse/, /bVrleske/), geminate consonants in words like *confetti* (/kVnfetti/), a distinction among /k, s, c/ such that *acquiesce* is phonologically /æckwiesce/, etc. Indeed, there were no constraints whatsoever on the relation between phonological and phonetic form, apart from the desire to set up underlying forms from which as much variation as possible in the shapes of particular morphological units could be predicted. Since much of this variation is the residue of earlier sound changes that have affected the same element in different ways in different environment, many early generative analyses resembled historical accounts of the language more than they did the knowledge of current speakers.

The perception of such apparently excessive abstractness led to proposals for constraining the operation of phonological rules in grammars, and the kinds of representation that should be posited. Kiparsky (1973), in particular, suggested a variety of conditions that would have the effect of prohibiting rules of 'absolute neutralization' (by which some posited phonological distinction is eliminated uniformly in surface forms, such that it never corresponds to a phonetic distinction) and other abuses of the theory's representational freedom. This led phonologists to seek external evidence for the reality of the representations they assumed, in the form of data from poetic systems, historical change, language games, behavioral experiments, and other sources outside of the phonology itself. Such a notion would have been incoherent on the assumptions of Structuralist phonology, for which the properties of the system are a matter of its internal logic, but became important with the shift to a view of phonological form as a matter of speakers' knowledge.

It is now apparent that phonological representations, while more abstract than assumed by structuralists, are still rather constrained in the extent to which they can deviate from phonetic form. There is no real consensus in the field, however, as to what constraints are appropriate on the operation or form of grammars so as to ensure an appropriate result. In the absence of some motivated grounds for decision in general, the status of specific conditions remains somewhat arbitrary. Replacing the conception of a grammar as a set of rules by that of a set of simultaneously evaluated partially violable constraints, as discussed in a later section, severely limits the potential abstractness of underlying representations in a way that may prove to resolve these issues in a principled way.

2.2 *The Interaction of Phonological Regularities*

In the nature of a Structuralist phonemic representation, all of the regularities expressed in the grammar are mutually independent. That is, it is only necessary to know the phonemic environment of a phoneme to predict its phonetic realization; and it is only necessary to know the phonetic environment of a phonetic segment to establish its phonemic correlate. When the bi-unique relation between these levels of representation is relaxed, however, more complex possibilities arise for regularities to interact with one another.

Consider the formation of English regular plurals, for example. For the three words *cut*, *dog*, *horse*

(roughly [kæt], [dog], [hors] respectively, phonetically) and others like them, the corresponding plurals are *cats*, *dogs*, *horses* ([kæts], [dogz], [horsɪz). Assume that the phonological representations of the nouns are essentially the same as their phonetic forms, and that the regular plural ending has a single constant phonological form: /z/ for concreteness' sake, though the point to be made is independent of this choice. Now there are two distinct regularities that are involved in determining the pronunciation of the plural ending:

(a) If the noun ends in a sibilant ([s, z, š, ž, č, ǰ]), a vowel [ɪ] appears between the stem and the ending; and

(b) If the stem-final sound is voiceless, the ending is voiceless ([s]) as well.

Now consider how these principles interact in determining that the pronunciation of *horses* (phonologically /hors + z/) should be [horsɪz]. In this case, the conditions for both of the rules above, *b* as well as *a*, are met, and we might expect both to apply, yielding (incorrect) [horsɪs]. The relevant observation is the following: the vowel inserted as a result of rule *a* has the effect that the stem and the ending are no longer adjacent, and so rule *b* is inapplicable. That is, the regularity represented by rule *b* in this case presupposes that the set of clusters of consonants which will remain adjacent in surface forms (a matter which is potentially affected by applications of rule *a*) has been determined. This logical relation of presupposition between rules is generally expressed by saying that rule *a* applies 'before' rule *b*, though it is important to note that no claim of temporal priority is thereby asserted, but rather a purely logical notion.

Much of the abstractness of early generative phonological analyses was made possible precisely by the availability of rule ordering as a descriptive device: the possibility of specifying rule interactions such as the above as an independent parameter within grammars. Although arguments for the necessity of such ordering formed an important part of the literature, and particular accounts sometimes involved apparently crucial depths of ordering as great as twelve to fifteen, many linguists felt that (as part of the general campaign to reduce abstractness in phonology) stipulated ordering relations ought not to be permitted as a descriptive device. Despite the existence of cases apparently requiring irreducible specification of relative order (e.g., dialects with the same sets of rules and underlying forms, differing in their surface forms only as a consequence of different interactions among the rules), and the fact that all proposed formulations of principles from which observed rule interactions could supposedly be predicted had obvious and well established counter-examples, a consensus developed in the field that such 'extrinsic' rule ordering statements ought to be prohibited. This has left a number of kinds of empirically observed interactions unaccounted for, a problem which has persisted as phonologists have turned from rules to constraints as descriptive devices.

2.3 *The Structure of Representations in Phonology*

As already noted, the description of speech at either the phonetic or the phonological level as composed of a sequence of discrete segment-sized units is a significant abstraction from physical reality. Its justification comes not from the observable facts, but from the extent to which it allows the analyst to uncover what is orderly and coherent in linguistic structure. By the mid 1970s, however, it had become apparent that a purely segmental organization of representations impeded the description of linguistic regularity in several respects (see *Phonology, Nonlinear*).

One of the first of these problems to be discussed arose from the analysis of systems of tonal contrasts, common in the languages of China, Africa, and in fact much of the world. Careful analysis revealed two ways in which tonal properties were problematic for strictly segmental models: on the one hand, what appeared to be a single tonal specification might take as its scope more than a single segment (perhaps the vowels of several consecutive syllables, or an entire word); and on the other hand, what appeared to be a single segment (a unitary short vowel, for example) might have a tonal specification involving two or even more consecutive tonal levels. If it is assumed that each of the consecutive segmental units of which a representation is composed bears exactly one specification for each potentially distinctive property, and the specifications of distinct segments are independent of one another, both of these situations are anomalous.

As a result of these observations, the segmental view came to be replaced with an *autosegmental* notion of representation. On this picture, specifications for each potentially distinctive property (or feature) succeed one another discretely, but an additional dimension of representation is the specification of the way these are synchronized or aligned with one another. In the limiting case, where each specification for a given feature is temporally aligned with one and only one specification for each of the other features, the standard segmental picture arises, but more generally, one or more instances of many-to-one alignment may be found. Once this view was articulated, it became apparent that many properties other than those of tone were also most appropriately described in this way. Nasality, in particular, behaves similarly in many languages and the very common phenomenon of assimilation is often best treated as an alteration in the scope of some features rather than a change in their values.

On the other hand, the full flexibility of autosegmental representations predicts much more independence of specification than is actually found. Features of place of articulation typically behave as a unit, despite their logical independence from one another, as do the features characterizing laryngeal activity. This observation gave rise to the notion that phonological features are not just an unordered set, but

instead are organized into a sort of hierarchical arrangement, such that for example features of consonantal place can be referred to as a unit. There is considerable consensus on the overall form such a *feature geometry* should take, but many details and relations among properties remain unsettled.

Another representational issue concerns the existence of structural units larger than the single segment, such as the syllable. Classical generative phonology, as represented by, for instance, Chomsky and Halle (1968), makes no appeal whatever to syllables (or any other unit above the segment). This was not simply an omission on their part, but rather a systematic claim to the effect that segmental specification could always suffice to express any phonological generalization: that is, that any observation involving syllables (of which there are many in Chomsky and Halle 1968) could be satisfactorily reformulated in a way that dispensed with any units other than segments. The boldness of this claim invited immediate attack, and indeed it soon became clear that there were indeed aspects of phonological structure that required reference to syllables and other supra-segmental units. The resulting theory of the *prosodic hierarchy*, by which segments are grouped into syllables, themselves grouped together into *metrical feet*, which in turn form constituents of *phonological words* (perhaps with some additional categories, according to some writers) has become another standard assumption about phonological structure.

2.4 Interactions between Phonology and Other Areas of Grammar

The original conception of a generative phonology was as a component of the grammar that served to map lexically specified syntactic structures onto phonetic form, a representation appropriate to serve as the instructions to the language-independent mechanisms of speech production. This picture logically entails the notion that words, phrases, and sentences are already fully assembled in the input to the phonology. This assumption, however, has proven to be problematic.

The notion that phonological properties interact with the syntactic environment in which forms occur is quite a familiar one, especially in the description of accent and intonation (see *Syntax–Phonology Interface*). Word formation, too, may depend in some instances on the assignment by phonological rule of phonological properties (see *Morphology in Linguistics*) in a way that contradicts the assumption that phonology applies strictly to the output of this part of the grammar.

The most extensively elaborated picture of how phonology interacts with the rest of grammar is presented by the theory of Lexical Phonology (see papers in Hargus and Kaisse 1993 for introduction and explication). Briefly, this picture distinguishes between lexical and post-lexical regularities, with the former applying in a way that interacts with word-formation, and the latter applying to a representation in which fully formed words appear in their syntactic context. The lexical rules, in turn, may (depending on the language) be divided into two or more *strata*, for example, rules applying to *roots*, *stems*, or *words*. A process of word formation takes some lexical unit as its input and yields an output (affixed or otherwise altered in form), with this output then subject to appropriate phonological modification. The resultant representation may serve as the input to further *cycles* of word formation plus phonology, until a complete surface word is arrived at.

These units, assembled into larger syntactic constructions, are then subject to the post-lexical rules. While various forms of lexical phonology generally constitute the background assumption of much phonological discussion, the correctness of such a view of the overall architecture of grammar has come into question. In particular, the rise of constraint based theories implying a single stage in the conversion of underlying to surface form has appeared incompatible with the sequential, derivational character of lexical phonological description. Some researchers have sought to reconcile these two trends in phonological theory, while others have attempted to augment one or the other sub-theory in such a way as to include the results of the other.

2.5 Rules vs. Constraints in Phonological Description

Formulations of phonological regularities since the 1950s have drawn on the mechanisms of automata theory: standardly, these have been expressed as rewriting rules that map an input representation onto an output, one step at a time. Developments in the theory of computation, however, have given rise to an alternative possibility. Instead of converting inputs to outputs in several stages, with each step involving the imposition of a single regularity in the computation of a new representation, one might formulate all of the regularities as elements of a single system of constraints. If these constraints are ranked with respect to one another, it is possible to say that any particular constraint may be violated in the output, provided that such a violation is motivated by the need to avoid violating a more important (higher ranking) constraint. The resulting theory provides an architecture of grammar that is (at least apparently) very different from that which has dominated most of the literature in phonology. For example, the kinds of interaction among regularities treated above as matters of rule ordering can no longer be described in this way, since all of the constraints in the grammar apply (in principle) in a single, simultaneous block. Constraints can have a language-particular ranking among themselves, but this is a matter of priority, not lgical

presupposition. Other differences have been alluded to in the discussion above.

Constraint based theories along the lines of Optimality Theory (Prince and Smolensky 1993) have been shown to have a number of significant advantages over rule-based accounts, and currently constitute the most active area of investigation in phonological theory. It remains to be seen, however, what character a comprehensive theory will have once the substantial body of results achieved within earlier, apparently incompatible frameworks has been incorporated comprehensively.

See also: Phonetics: Articulatory; Phonology, Nonlinear; Saussure, Ferdinand de (1857–1913); Speech Perception; Speech Production, Neural Basis of; Speech Production, Psychology of; Syntax–Phonology Interface

Bibliography

Anderson S R 1985 *Phonology in the Twentieth Century: Theories of Rules and Theories of Representations*. University of Chicago Press, Chicago

Baudouin de Courtenay J 1972 *Selected Writings of Baudouin de Courtenay*. Indiana University Press, Bloomington, IN

Bloomfield L 1933 *Language*. Holt, New York

Chomsky N, Halle M 1968 *The Sound Pattern of English*. Harper and Row, New York

Coulter G R 1993 *Current Issues in ASL Phonology (Phonetics and Phonology 3)*. Academic Press, San Diego

Fischer-Jørgensen E 1975 *Trends in Phonological Theory*. Akademisk Forlag, Copenhagen, Denmark

Hargus S, Kaisse E 1993 *Studies in Lexical Phonology (Phonetics and Phonology 4)*. Academic Press, San Diego

Harris Z 1951 *Methods in Structural Linguistics*. University of Chicago Press, Chicago

Jakobson R 1941 Kindersprache, aphasie und allgemeine lautgesetze. In: Jakobson R (ed.) *Selected Writings, Vol. I* (English trans. 1968 Child Language, Aphasia and Phonological Universals). Mouton, The Hague, pp. 328–40

Joos M 1957 *Readings in Linguistics*, Vol. 1. American Council of Learned Societies, Washington, DC

Kenstowicz M 1994 *Phonology in Generative Grammar*. Blackwell, Oxford, UK

Kiparsky P 1973 Phonological representations. In: Fulimura O (ed.) *Three Dimensions of Linguistic Theory*. The TEC Corporation, Tokyo

Prince A, Smolensky P 1993 *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University, New Brunswick, NJ

Sapir E 1925 Sound patterns in language. *Language* **1**: 37–51

de Saussure F 1916 *Cours de linguistique générale*. Payot, Paris

Trubetzkoy N S 1939 *Grundzüge der Phonologie. Travaux du Cercle linguistique de Prague 7*

S. R. Anderson

Phonology, Nonlinear

Nonlinear phonology is a loose collection of relatively recent theories that address various aspects of the sound structure of human languages. Traditional approaches to phonology are linear, which means that they represent each word as nothing more than a string of speech sounds (called segments). Each segment is a bundle of its phonetic attributes, called distinctive features, but no further structure is assumed. Nonlinear phonology posits a richer architecture: segments are organized into a hierarchy of constituents, such as the syllable, the stress foot, and so on; and distinctive features are liberated from the segmental feature-bundle, allowing them to have domains that are larger or smaller than a single segment. This article examines some of the consequences of this hypothesis, focusing on two particular implementations of it.

1. Historical Background

Before the mid-1970s, most research in phonological theory started from the assumption that segments are linked together like beads on a string (see *Phonology*). In this view, phonological representations are very similar to alphabetic writing, with exactly one symbol standing for each segment and with little or no structure beside the linear order of the segments in a word.

The best known theory of this general type is the one in Chomsky and Halle's (1968) *The Sound Pattern of English* (SPE). The SPE theory is rigorously and consistently linear: for example, syllables play no role at all, and constituents such as the word have only a derivative status. The only departure from the alphabetic model is that each segment is assumed to be a bundle of distinctive features standing for its various attributes (such as nasal, voiced, rounded lips, and so on.

During the 1940s, two important precursors to contemporary nonlinear phonology emerged independently in the work of J. R. Firth and Zellig Harris. They shared the idea that phonological properties could have domains larger than a single segment. For example, in the South American language Terena, the meaning of 'I' or 'my' is marked by nasalizing (indicated by a tilde) a sequence of segments beginning at the left edge of the word: *owoku* 'house', *õw̃õŋgu* 'my house'; *arine* 'sickness', *ãr̃ĩnẽ* 'my sickness'. Firth or Harris would say that the domain of nasality in Terena is (approximately) the word. Their theories did not have much influence, though, and their work was largely neglected for several decades.

2. Nonlinear Phonology

In reaction to SPE's strongly segmental orientation, the mid-1970s saw a profusion of nonlinear ap-

proaches to phonological phenomena, supplying ideas and insights that continue to be studied in the present day. These models were developed to address empirical problems that had been disregarded or treated unsatisfactorily in SPE: tone, long distance assimilation processes, syllables, the typology of stress systems, and the interaction of phonology with morphology and syntax.

Two leading ideas have emerged. One is *autosegmentalism*, the notion that distinctive features, rather than segments, are the atoms of phonological representation. (This term was coined by John Goldsmith.) From an autosegmental perspective, a distinctive feature is not an attribute of a segment, as in SPE, but a free standing entity. This status allows the feature to act independently of its segmental host, approximately as in Firth's and Harris's theories. The other leading idea is the *prosodic hierarchy* (a phrase coined by Elisabeth Selkirk), which organizes segments into successively larger constituents such as the syllable or word. The prosodic constituents are analogous to the more familiar syntactic constituents such as sentence or noun phrase (see *Constituent Structure: Syntactic*), and, as in syntax, structural requirements imposed on the constituents are an important analytic tool.

Autosegmentalism and the prosodic hierarchy are complementary, not competing, hypotheses. As a matter of logic, they are not mutually exclusive, since autosegmentalism speaks to the nature of subsegmental units (distinctive features), while the prosodic hierarchy is a claim about how segments themselves are organized into larger groupings. They have also been applied to factual domains that are mostly distinct. Autosegmentalism helps to solve problems in tone and featural phonology, such as assimilation. The prosodic hierarchy is important in understanding the phonology of syllables, stress, words, and phrases.

Autosegmentalism and the prosodic hierarchy are, ultimately, claims about how language is represented in the human mind. Though they allow for a considerable range of variation between languages, they also have a core that is claimed to be universal. Hence, the main evidence for these theories comes from observations about phenomena that recur in many languages.

The rest of this article discusses the implementation of these two aspects of nonlinear phonology, the results and controversies that emerge from them, and their present and future prospects.

3. *Autosegmental Phonology*

The following example illustrates one of the key arguments for autosegmentalism. Some speakers of the Bantu language Sanga use a secret language (analogous to pig Latin in English) called Kinshingelo. Words in this secret language are formed by systematically permuting the segments of Sanga words, as in the following example:

Sanga

báá.kó.lwèé bàà.dyáá má.tá.bá àá.ŋkàá.mbò

Kinshingelo

báá.lwé.kòó dyàà.báá má.bá.tá àá.mbòó.ŋkà

Gloss

'The birds are eating my grandfather's corn.'

The accents mark tones: acute *ó* is high and grave *ò* is low. Double vowels are long, so *àá* stands for a long *a* vowel with a rising (i.e., low–high) tone. For expository convenience, syllable divisions have been indicated with a full stop.

A Kinshingelo word is formed by swapping the last two syllables of each Sanga word. But not precisely: the segments themselves are swapped, but their tones and length are not. Speakers of Sanga unconsciously regard *àá* as somehow divisible into three components: a part that describes the action of the tongue and lips in forming consonants and vowels, a part that describes the action of the larynx in controlling tone, and a part that describes the duration of the vowel. A useful analogy is an orchestral score, which contains distinct but coordinated instructions for many, different instruments. This is the central insight of autosegmental phonology: speech sounds consist of independent components which are coordinated in time.

Tone is the premier example of an independent component of speech, and in fact autosegmental phonology was first developed to deal with tone. Some other consequences of autosegmentalism for the analysis of tone include:

(a) *Persistence under deletion*. When a vowel deletes, its tone may remain and attach itself to some nearby vowel.

(b) *Floating tones*. When a vowel deletes, its tone may remain and express itself indirectly, by affecting following tones.

(c) *Toneless syllables*. Some syllables have no tonal specification at all. Their pitch is determined by linear interpolation between syllables that have tones.

(d) *Tone shift*. In the Bantu language Kikuyu, every tone shifts exactly one syllable to the right, domino fashion, so /*tòmátómàyà*/ becomes *tòmàtómáyà*

(e) *Tone melodies*. In some languages, words are limited to a small set of fixed tone melodies, or tone melodies may mark certain morphological distinctions.

In sum, tones and segments may behave independently of one another. This is the expected result of the autosegmental hypothesis (see also *Suprasegmentals*).

Not long after it was developed, autosegmental phonology was applied to nontonal phenomena as well. The idea is that every distinctive feature—[nasal], [voiced], and so on—occupies its own separate autosegmental 'tier,' allowing it to function independently of segments or other features. The main focus of

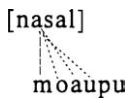

Figure 1
Autosegmental representation of Warao *mõãũpu*

research in this area has been on two related problems: assimilation and harmony.

In assimilation, one segment (the target) takes on some of the features of another, nearby segment (the trigger). Harmony is a kind of long distance assimilation, where the triggering segment and its target are not necessarily adjacent. Here are some examples from the South American language Warao, and the North American language Chumash:

(a) Warao /moaupu/ → *mõãũpu* 'give them to him'
(b) Chumash /k-sunon-š/ → *kšunonš* 'I am obedient'

In Warao, the feature [nasal] assimilates, affecting a potentially unbounded sequence of vowels after a nasal consonant. In Chumash, the feature [anterior] assimilates, changing *s* to š (the initial consonant in *ship*) if followed anywhere in the word, at whatever, distance, by another š.

Autosegmentalism contributes to the understanding of these processes in two ways. First, it provides a natural interpretation of assimilation. Because features and segments are independent, the domain of a feature can be larger than a single segment. Assimilation expands the domain of a feature from one segment to many. Second, autosegmentalism allows long distance harmony processes to be analyzed as local on the tier of the assimilating feature. These two key insights are usually expressed graphically, as in Fig. 1. The feature [nasal] is shown on its own autosegmental tier, separate from the rest of the segment. Temporal coordination between [nasal] and the rest of the segment is indicated by association lines. The solid association line is original, while the dotted association lines are the result of the process of assimilation. The effect of assimilation here is to expand the domain or scope of [nasal] to include the vowels *oau*.

The most radical autosegmental theory is one in which every distinctive feature is on a separate tier, with complete freedom to act independently of all other features. This view cannot be right, because there are strong functional associations among various subsets of the features. For example, the features that characterize states of the larynx have a considerable degree of cohesion in phonological systems, assimilating or neutralizing together.

Functional groupings of features are addressed in theories of feature geometry. (This term was proposed by G. N. Clements.) The idea is to posit a layer of representation between the segment and the features.

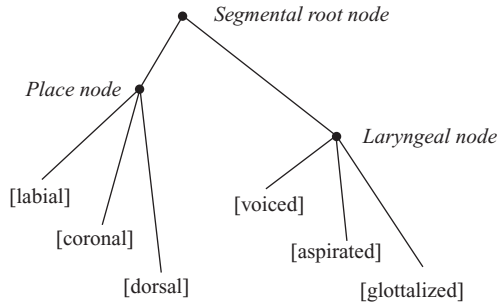

Figure 2
A model of feature geometry

Elements of this intermediate layer, called class nodes, are constituents corresponding to the observed functional groupings. A typical implementation, abstracting away from controversial details, appears in Fig. 2. The laryngeal features often act together in phonological processes, and so there is a constituent, the Laryngeal node, that unites them. Similarly, a single process will often affect the various features that distinguish *p* from *t* from *k*, and so there is a constituent, the Place node, to unite them as well. Feature geometry, like autosegmental phonology itself, is claimed to be universal. Research in this framework is oriented toward discovering a set of class nodes that will account for all (or at least all common) phonological processes in the world's languages.

4. *The Prosodic Hierarchy*

The term prosody is used by phonologists to refer to the durational, rhythmic, and phrasing aspects of speech. The study of prosody includes syllables, stress, restrictions on the size or shape of words, and many of the ways in which morphology or syntax can affect pronunciation (see *Word, Linguistics of*; *Linguistic Typology*).

A key insight into the nature of prosody is the idea of a small, universal set of prosodic categories arranged into a hierarchy (see Fig. 3). Certain additional assumptions give the hierarchy predictive value:

(a) *Grouping and layering*. Level *n* of the hierarchy imposes a grouping structure on the layer just below it, level $n-1$.

(b) *Headedness*. Every constituent at level *n* designates a specific, most prominent constituent at level $n-1$ as its head.

(c) *Structural requirements*. There are global and parochial conditions on the well-formedness of constituents. For example, languages frequently require that feet be binary, and there is also evidence of binarity conditions on other levels of the prosodic hierarchy.

So, proceeding from the bottom of the hierarchy, a syllable is a grouping of segments; the head of the

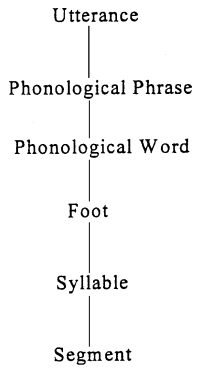

Figure 3
A prosodic hierarchy model

syllable is the peak of acoustic energy, usually a vowel. A foot is a grouping of (usually two) syllables, one of which is more prominent than the other. The more prominent syllable is the locus of linguistic stress. A phonological word is a grouping of feet, one of which is also the most prominent (the locus of main stress). A phonological phrase is a grouping of phonological words that also includes a most prominent member, and so on.

Results of surprising subtlety and generality can be derived from these modest premises. For example, many languages set a lower bound on the size of the phonological word. The Australian language Diyari is typical in this respect: it has no monosyllabic words at all. This so-called minimal word requirement follows from joint action by three different constraints: layering (the phonological word dominates the foot in the prosodic hierarchy); headedness (every word must contain at least one foot, to serve as its head); and a structural requirement that feet be binary, since they are the basic units of stress rhythm. Comparable results can be found at other levels of the prosodic hierarchy.

5. *Present and Future Prospects*

Nonlinear phonology has been an active area of research since the late 1970s. The basic autosegmental hypothesis enjoys wide and perhaps universal acceptance. Various details are matters of debate, however:

- Do all features have the same autosegmental status as tone?
- Is feature geometry an appropriate model of featural organization?
- If so, exactly how are the features organized?
- Are (binary) features the appropriate units of subsegmental structure?

The prosodic hierarchy is somewhat more controversial. Although notions of constituency do play a role in most theories of prosody, there are also models that are partly or entirely constituent-free.

In recent years, much research has focused on the ways in which Optimality Theory can illuminate topics in nonlinear phonology. A particularly productive line of investigation has been the study of constraints that align the edges of prosodic constituents with morphological or syntactic ones. These constraints relate the prosodic hierarchy to the hierarchy of grammatical constituents, offering a new perspective on the interface between phonology and grammar.

See also: Phonology

Bibliography

Chomsky N, Halle M 1968 *The Sound Pattern of English*. Harper & Row, New York
Clements G N 1985 The geometry of phonological features. *Phonology Yearbook* **2**: 225–52
Clements G N, Keyser S J 1983 *CV Phonology: A Generative Theory of the Syllable*. MIT Press, Cambridge, MA
Goldsmith J A 1990 *Autosegmental and Metrical Phonology*. Blackwell, Oxford, UK
Hayes B 1995 *Metrical Stress Theory: Principles and Case Studies*. University of Chicago Press, Chicago
van der Hulst H, Smith N 1985 The framework of nonlinear generative phonology. In: van der Hulst H, Smith N (eds.) 1985 *Advances in Nonlinear Phonology*. Foris Publications, Dordrecht, The Netherlands
Kenstowicz M 1994 *Phonology in Generative Grammar*. Blackwell, Cambridge, MA
Liberman M, Prince A 1977 On stress and linguistic rhythm. *Linguistic Inquiry* **8**: 249–336
McCarthy J, Prince A 1993 Generalized alignment. In: Booij G, van Marle J (eds.) 1993 *Yearbook of Morphology*. Kluwer, Dordrecht, The Ntherlands
McCarthy J, Prince A 1995 Prosodic morphology. In: Goldsmith J A (ed.) 1995 *The Handbook of Phonological Theory*. Blackwell, Oxford, UK
Nespor M, Vogel I 1986 *Prosodic Phonology*. Foris, Dordrecht, The Netherlands
Odden D 1995 African tone languages. In: Goldsmith J A (ed.) 1995 *The Handbook of Phonological Theory*. Blackwell, Oxford, UK
Pierrehumbert J, Beckman M 1988 *Japanese Tone Structure*. MIT Press, Cambridge, MA
Selkirk E 1981 On the nature of phonological representation. In: Laver J, Meyers T, Anderson J (eds.) *The Cognitive Representation of Speech*. North-Holland, Amsterdam
Selkirk E O 1984 *Phonology and Syntax: The Relation Between Sound and Structure*. MIT Press, Cambridge, MA

J. J. McCarthy

Photography as a Medium

For over a century and a half, photography has been the most popular and widespread means of making images. The first and primary tool of photography is the camera, and changes in the medium and its

applications are often attributed to technical developments in making, reproducing, and distributing photographs. Equally if not more significant, however, have been changes in the ways people think about photography, including ideas about what and how photographs represent, and the roles they have played in social, political, and economic life. This article examines developments and conflicts in the idea of photography within a historical framework of changing cultural practice.

1. Photography's Origins

The invention of perspective in the fifteenth century laid the foundation for the system of visual representation that became photographic. The camera obscura, used by scientists, artists, and magicians since the Renaissance, is the most important precursor to the camera. By the eighteenth century this device had evolved from a large clumsy apparatus for projecting an image onto a plane surface, to a more portable instrument, used for the curiosity and pleasure of looking at images of all sorts of objects and scenes projected by the lens.

A knowledge of light-sensitive materials was popularly available in Europe by the 1720s, and one of the riddles of photography is why nearly a hundred years passed before this knowledge was applied to fixing the camera obscura's image. Despite a broad fascination with display forms, such as the diorama, that created illusions of realistic representations of nature, the idea of the photograph as a permanent copy apparently did not exist. Suddenly, in the 1820s a range of experimenters began simultaneously to compete to be the first to successfully fix the photographic image; the desire to photograph had emerged.

Photography's invention is generally dated to 1839, when breakthroughs were announced nearly simultaneously in France and England. Louis Jacques Mandé Daguerre's induplicable monochrome on a silver-coated copper plate, and William Henry Fox Talbot's reversed image—the negative—were the first successfully publicized and patented processes. The daguerreotype enjoyed an immediate popularity, setting off a craze which by the end of the year the French press had named '*daguerréotypomanie*.' Talbot's calotype system, on the other hand, based on a negative image that could be used to print a potentially infinite number of positives in as many different sizes, laid the foundation for all subsequent developments in photography. The refinement and diffusion of these two systems, alongside many competitors, had an immediate impact on European social life and spread quickly to North America and colonial outposts.

Photographs appeared in many display forms, as cartes-de-visites and stereographs, in albums and as lantern slides, and photography surfaced as a theme in popular music, on the stage, and in artifacts ranging from home furnishings to fashion accessories. Photography as an *idea* permeated mid-century European popular culture.

2. Early Applications of the Medium

At the same time that photographic tools and processes were being simplified, an agenda was being established for how the new medium was to be used. This included the people, places, and events represented in photographs and, equally important, who was taking the pictures and who were the intended viewers.

2.1 Photographs of People

Within months of Daguerre's announcement, professional portrait studios had sprung up in Paris, and soon studios were established in other major European cities, in New York, and in Washington. From 1860 onward, portraiture expanded to provincial towns, to less industrialized European countries, and to Australia, China, Japan, Mexico, and South America.

The photographic portrait represented a democratization of the painted portrait, and was commissioned first for private use within a family. The rising status of certain photographers led to commissions from men in public life whose positions required an official portrait. Most portraits of individuals followed the strict visual conventions of the painted portrait, a head-and-shoulders composition, with the sitter's head slightly turned to avoid a direct gaze into the camera. Portraits of family groups were common, and a custom quickly developed of photographically commemorating important events such as weddings and graduations. Photographs of deceased relatives, lying peacefully in their caskets, were also highly valued.

In the 1860s, studios began to turn out millions of inexpensive *carte-de-visites*, full- or bust-length portraits of workers, tradespeople, members of the bourgeoisie and aristocracy, native people of the colonies. These mass-produced images were often instrumental in establishing the status of a politician or celebrity, and both the French and British royal families permitted the sales of their images as *cartes*. Through its broad distribution, the *carte-de-visite* also contributed to forming Western ideals of feminine beauty and fashion, as well as ideas about people from the nonindustrialized parts of the world as exotic and other. The *carte* craze further generated a surrounding industry of albums and holders for cataloguing and exhibiting the growing image collections.

In 1888 George Eastman launched the first Kodak camera, with the slogan 'You push the button, we do the rest,' and immediately opened the practice of photography to a much broader segment of the

population. Studio portraiture declined as amateurs began to photograph family and friends. The repertoire of photographic events was also expanded as picture-taking became associated with holidays, family celebrations, and vacation trips. Photographic practice had entered the domestic domain, forming a bridge between personal memory and public history. A hundred years after the first Kodak camera, photography was fully integrated into family social life. In the mid-1980s, for example, over 90 percent of US families owned a still camera and amateurs were taking more than 11 billion photographs a year (Chalfen 1987). The majority of these pictures were of family members and friends and were rarely seen outside that social group. Nonprofessionals' private photography today accounts for the largest single application of the medium. Studies have shown that people attach enormous value to these images, and losing photographs of family and friends, for example, through fire or war, is often experienced as a greater tragedy than the loss of other material possessions.

2.2 *Distant Worlds*

The early years of photography coincided with the period of colonial expansion and exploration of newly won territory for the industrializing nations of Europe and North America. Wonders of nature, the dynamic new cityscape and its architecture, and the exotic cultures of distant lands held a fascination and attraction for many. The camera was an ideal tool for showing what was new or previously unknown to an audience eager to marvel over sights they could not see with their own eyes. Photography also opened up new territories for tourism by people who, having seen a place in pictures, now had the desire and means to experience it themselves. A new aesthetic of the landscape also stimulated tourism, as for example when photographs of the rocky Scandinavian coasts brought British and German tourists north in the 1870s (Rosenblum 1984, p. 112).

Photographs of popular European tourist sites, such as the city of Rome and its ruins, were marketed as stereographs. Distant lands of North Africa and the Near East became accessible to the armchair tourists of Europe as early as the 1850s through Francis Frith's photographs. Frith's *Egypt and Palestine Photographed and Described* (1858) was the first of 11 volumes and three sets of stereographs (Rosenblum 1984, p. 122). His work was motivated in part by a concern over growing French influence in the region, and must be seen in the broader context of British expansionist goals of the period. Local Egyptian photographers also opened studios and marketed their work to the growing European fascination with views of the ruins and landscape of the Near East. Similar patterns developed in India and parts of the Far East. By 1884, for example, several thousand native photographers were in business in China, serving both local clients and the Western trade.

In North America, photography was linked to Western expansion. Photographers accompanied survey teams in order to document areas for future mineral exploitation, rail routes, and civilian settlement, at the same time that they were contributing to the visual discourse of the American frontier and its native peoples. The natural wonders of the wilderness, including Niagara Falls, were immensely popular stereographic themes, used in lantern slide shows and collected in albums. This iconography and the photographers who produced it, such as Carleton Watkins and William Henry Jackson, established the grandeur of the mountain landscape as a symbol of transcendent idealism. Their work also contributed to lobbying efforts to establish a national park system in order to save the pristine wilderness from commercial exploitation by private interests.

Nor did the city escape the camera's gaze. In contrast to images of untouched wilderness and ancient unchanging ways of life, here the focus was on physical transformation. New transport systems, bridges, and buildings made possible by new technologies and materials were transforming the city into a modern metropolis. Urban expansion was a source of national pride, and both governments and private interests supported photographic documentation of the major cities of Europe and North America. Cityscapes in panorama format were especially popular, a parallel to the painted panoramas and dioramas that had emerged as mid-nineteenth century forms of entertainment.

Photography's prominent and various positions in the emergent discourse of modernity was never more evident than in the world expositions, the late-nineteenth century spectacles of science and marketing, combining knowledge and entertainment. Beginning with the Crystal Palace in London in 1851 and followed every few years in Britain, France, and the United States, these major events both symbolized and displayed industrial expansion, made possible by new technologies and materials. Displayed side by side with the symbols of modernity were the exotic products and peoples of the underdeveloped nations, the colonies. At this point in the medium's development, photography provided an unproblematized means of expanding knowledge of the visible world, and served 'the general urge in all industrialized societies to measure, describe, and picture the physical substances of all things on earth and in the heavens' (Rosenblum 1984, p. 140).

3. *Concepts of the Image*

Photography's popularity as a democratic form of self-documentation, intertwined with colonial interests and industrial expansion, point to the camera as a

tool that linked seeing and knowledge in new ways. Within the sciences, in the arts, and in social reform movements, distinct and often conflicting visual discourses have developed around the photograph as a source of insight and knowledge. We turn now to the ways of seeing with the camera that were generated within each of these broad fields.

3.1 Photography and Science

Within the discourses of science, photography was to be a witness to scientific observation with greater accuracy than the human eye was capable of. In medicine, photography provided close visual description of the human body, and was an aid in documentation and diagnosis of medical problems and also psychological reactions and mental aberrations. Examples are portrait photographer Professor John William Draper's book *Human Physiology* (1856), illustrated with woodcuts based on photographs and Dr. Hugh Welch Diamond, also known for his pictorial art photography, who documented women inmates of the Surrey County Asylum. Charles Darwin's *The Expression of Emotions in Man and Animals* (1872) was illustrated with photographs from a variety of sources. Microscope lenses were developed and the discovery of X-rays by Conrad Wilhelm Roentgen (the 1901 Nobel Prize recipient) led to their immediate use as camera images in medical diagnosis. Motion studies and stop-motion experimentation depended heavily on photographic documentation. Eadward Muybridge's 100,000 images analyzing movements involved in animal locomotion and a range of human activities (including walking, curtseying, and bricklaying) provided a new if contested way of seeing that also influenced the arts, in particular Cubist and Futurist painters.

Taxonomy was a unifying concept of the nineteenth century scientific paradigm, and studies of the human species were formed by the delineation and elaboration of physical and cultural distinctions. Anthropology became a dominant discourse of explanation, based on historical and biological notions of culture, and relying heavily on comparative taxonomic methods in order to systematically catalogue the world's peoples (Edwards 1997). Photography was central to the anthropometric systems developed to produce comparable data for analysis and classification. The system used by the distinguished British biologist Thomas H. Huxley for the Colonial Office, for example, involved posing individuals naked with a measuring rod in order to produce 'a systematic series of photographs of the various races of men comprehended within the British Empire' (cited in Edwards 1997, p. 56). These studies of human physiogomy, used to expand evolutionary theory, contained a strong moral dimension, concerned with both creating and sustaining the complex system of social knowledge in which the images operated. Variations on the method were applied in hospitals, prisons, and other institutions where photographs of inmates were used to establish and classify social types.

The 1870s and 1880s also brought about an awareness of the contributions photography could make to anthropological fieldwork, by Edward B. Tylor and others. Yet there are few examples of systematic photographic documentation in ethnographic field studies. Early photography of the ways of life of diverse folk groups was most often carried out by missionaries and educators or by artists working in the pictorial tradition. Edward Curtis, Adam Clark Vroman, Robert Flaherty, and Frances Flaherty documented native American tribal life in its positive aspects, romanticizing in particular the close relationship to nature. Nor was it uncommon for ethnographic photographs from the turn of the century of peasant customs made in a pictorialist style to be used to advance the cause of nationalism, as in several Eastern European countries. The most common reference to an ethnographic photography worth emulating is the classic study of *Balinese Character* (1942) by Margaret Mead and Gregory Bateson, as the 'first saturated photographic research in another culture' (Collier and Collier 1986, p. 12). Visual anthropology at mid-century was more closely associated with documentary film than with photography. Ethnographic photography survived as an *ideal*, however, despite the general absence of large-scale systematic studies. The British 'Mass-Observation' project in the mill towns of northern England in the 1930s, for example, was designed as an 'objective documentation' of working class life, in the manner of an anthropological study.

The archive has worked hand in hand with the scientific endeavors of accumulating and systematizing evidence. Seen as a repository of visual knowledge, the archive is related to the ideal of a comprehensive collection of images that together represent the whole (other examples being the family album or collection of *carte-de-visites*). The structure of the archive and its classification system was intended to mirror the physical and social world that the collection represented. Following Michel Foucault's groundbreaking archeology of the human sciences (1970), museums of natural and cultural history, as well as police, medical, and other archives, can no longer be seen as disinterested institutions that sustain and perpetuate value-free systems of knowledge, but reflect instead the structures of political and economic interests that gave rise to them. The photograph's evidentiary authority is thus inseparable from the archive as a repository of knowledge and power.

3.2 Photography and Art

Contrary to the idea of the photograph as an objective visual record, it has been argued that photography is inseparable from the Western pictorial tradition. Particularly in its relationship to landscape and nature,

photography employed a set of conventions and aesthetic standards theorized by the European concept of the 'picturesque.' Emulating themes from the fine arts, many early photographers developed techniques that undermined the sharp literalness of the lens in their aspirations to make images that were at once truthful, beautiful, and inspirational. Exhibiting esthetic photographs in an appropriate context was a paramount goal of the turn-of-the-century Secessionist movements in Europe and the United States, and was important to establishing photography's status as art. Pictorialism in photography was a synthesis of the realism of the camera image, with frequently formulaic romantic compositions of peasant or working class life.

A complex and uneasy relationship between photography and the fine arts has continued, even as the repertoire of art photography has expanded. This is generally beyond the scope of this article, apart from several developments relevant to photography's relationship to the social sciences. On the one hand, themes drawn from peasant and working class life as well as of non-Western people integrated scientific inquiry with romantic images of an exotic or primitive Other as legitimate subjects of photographic art. On the other hand, attempts in the 1920s to define and explore the characteristics which distinguish photography from other visual arts laid the foundation for a more radical 'straight' photography, called New Realism or *Neue Sachlichkeit* in Europe and North America. The nonmanipulated or spontaneous photograph became a means of exploring and exposing familiar subjects in new ways. These different schools and genres of art photography became more or less systematized as museums began to add photographs to their collections. The Museum of Modern Art in New York is a noteworthy example of a leading museum where a series of influential curators established a vocabulary of photographic practice based on the premises of modernism. By the 1940s, a photographic canon was in place, including a history of art photography with specific artists, works, and genres represented in the museum's collections (Newhall 1949). Of particular importance to the social sciences was the respect this canon established for the genre of documentary photography as a legitimate form of esthetic practice.

3.3 The Photographer as Engaged Observer

Parallel to their idealized representations of European peasants, from the 1850s onward artists began to be concerned with the social and ethical consequences of industrialization, preparing the way for documentary photography's role in social reform campaigns. 'Before' and 'after' images became a staple of reform photography in the 1890s, as evidence of the positive effect the reform school or institution had on its residents. Photographs from the slums of New York (by Jacob Riis) and of child labor conditions (by Lewis Hine) were used to raise the fears and sway the consciences of politicians and a middle-class public. The photography of social reform was closely linked to new means of dissemination: the half-tone process for mass-producing photographs had been developed in the 1890s and a range of new periodicals arose in its wake.

Alongside the popular press' increased use of photographs after the turn of the century, social reform journals were among the first to recognize photography's ability to advocate a cause. Here the American *Charities and the Commons* is noteworthy. The *American Journal of Sociology* was also using photographs to illustrate articles, for example, on housing conditions. This visual material, often by women who held administrative positions in social action agencies, later declined under Albion Small's editorship, as the journal aligned itself with sociology's move to establish itself as a scientific discipline (Stasz 1979).

By the 1920s, documentary photography, as the genre came to be known, was recognized as an effective means of drawing attention to social issues. The best known example of such an effort was the photographers who worked for the US Farm Security Administration (FSA) under the leadership of Roy E. Stryker in the late 1930s, documenting the effects of the depression on rural Americans and the need for government aid. In other cases, photographers used the documentary form to draw attention to marginalized social groups, occasionally drawing on their own experience as members of the culture they documented. One of the clearest ideological expressions of photographers' identity was the Worker Photography movement, most active in Germany between World Wars I and II. The International Communist Party, convinced that photographs unavoidably reflect the class of the person holding the camera, mobilized workers in a variety of documentary projects and publications. In direct contradiction to a claim of impartiality, this photography based its authority on the photographer's membership in the group. It can thus be seen as a precursor to anthropology's interest in indigenous photographers' images as expressions of their specific cultural perspectives.

The authenticity of the documentary style was based on a complex set of factors. Seldom was the work based on a single photograph, but instead was built up through a series of images witnessing the photographer's proximity to the subject over an extended time. Despite many exceptions, the images also tend to follow a specific visual code. Straightforward photographs of people, taken in their everyday environment under available light conditions, are the most typical. If the subject appears to be aware of the camera, he or she is usually gazing directly into the lens. The photograph's claim to truth is based therefore on the

seemingly contradictory appeals of the photographer's simultaneous engagement with and objective distance from the subject.

4. Photography and Mass Society

Closely related to the ways photography has portrayed its subjects is the effect of images on those who see them, the audience. Photography was a mass medium almost from its inception, as the idea of the photograph and the ways it was marketed had a broad and immediate social and cultural impact. The 'very insatiability of the photographing eye' and the 'omnipresence of photographed images' position the medium in a vortex of complex moral, ethical, political, and esthetic issues. Sontag's influential book of essays *On Photography* (1973) brought many of these issues up for debate. Yet the exact nature of the photograph's appeal and its effect on its viewers has remained elusive. Despite extensive research on photography's impact on attitudes and behavior, results have been either inconclusive or too context-specific (regarding the viewer's age and background or the viewing context) to support any general conclusions. What is evident, instead, is a widespread belief in the photograph's influence which has in turn led to specific applications of the medium based on conceptions of the nature of photographic communication as a highly unstable synthesis of factual information and seductive attraction.

From the 1850s photographs began to be used alongside other illustrations in some magazines, and the widespread application of the half-tone process in the 1890s coincided with the emergence of new audiences in Europe and the United States. Photography was thought to appeal in particular to people who were either nonliterate or had not yet mastered the language of their new country. A new inexpensive illustrated press arose that used photography in often sensational ways, and photographs began to appear in advertising. Political campaigns made extensive use of photography from the 1920s onward, in documentary forms described above, but also in photomontage and posters. The reticence of many newspapers to make extensive use of photography (for example, the morning newspapers 'of record' in New York and London) is an exception which nevertheless rests on the same notion of the photograph's communicative power. The photograph, if seen as an unmediated copy of nature, does not require language or reason to be understood, but appeals instead directly to the emotions. Newspapers and other media that wished to base their appeal on rationality excluded photographs from their pages. Even if an image were manipulated through technical or other means, the message that the audience saw or 'received' was accepted as representing reality.

The model of mass communication that emerged first in the United States and dominated research for several decades was based on the notion of communication as a successful 'transmission' of a message from an active and powerful source to a passive receiver. The origins of the model and the research it generated can also be traced to growing concern over the power of the media, particularly in the service of political propaganda and the rise of fascism in Europe after World War I. Although only sporadically applied in research on photography, the transmission model of communication served as the underlying framework of explanation of photography's influence on its audience. The power of images is assumed to be greater, for example on the relatively unformed psyches of children and youth. 'Moral panics' continue to pursue the image and its influences, giving rise to research on Hollywood films in the 1930s, on comic books in the 1940s, on television's effects on children in the 1950s and 1960s, and video games at the turn of the century.

Countering the transmission model of photography's influence on its audience is the perspective that considers photography as a cultural practice and phenomenon which participates in the construction of power and meaning through the mutual and intertextual establishment of social and political institutions and forms of visual expression. The cultural perspective views the audience not as passive receivers, but as members of a (sub)culture who employ cultural knowledge as they produce, consume, and interpret photographs. The cultural perspective does not deny the power of the media, but rather sets it in a historical, social, or political context, and examines how the media use photographs to address or position the viewer in specific relationships to media content. The perspective shifts the emphasis from an inquiry into photography's impact on individual or group attitudes and behavior to an examination of the ways photography is used in the construction of meaning, often focusing on social values and the use of stereotypes. The medium has been decisive, for example, in the production of visual representations of gender and gender relations, as revealed in Erving Goffman's study of advertising (1979). Other studies have demonstrated how media images of women not only establish standards of feminine beauty, but also link the image of the female body to products and services in a process of commodification that objectifies all women (cf. Williamson 1978, Irigaray 1985, Bordo 1993, Solomon-Godeau 1991). Other analyses have provided insights into the formation and perpetuation of stereotypes of race, in advertising but also in the news media, with groundbreaking work carried out at the Centre for Contemporary Cultural Studies in Britain in the 1970s.

Issues of media power and influence are also linked to photography through a growing concern over the image's role in surveillance. Cameras, in addition to their continually expanding position in private life, are used increasingly in both public and commercial

spaces for the ostensible purpose of increasing security and preventing crime. Yet the images taken by these cameras also raise issues of the photograph's potential for restricting, documenting, and controlling individual behavior. As technology extends the capacity to send images over global networks, the specter is raised of expanding spheres of media influence and control that transcend the nation state. At the same time, the interactive aspects of many contemporary media raise the possibility of individual and group autonomy from transnational media content, and the ability of the individual to form and reform images that are sent and received. The cultural impact of these relatively new image technologies is a current topic of debate, particularly with regard to the moral and ethical implications of their distribution and use.

5. *Technology and Revisions of Photographic Meaning*

A central question throughout photography's history is whether the meaning of the photograph is a function of the medium *per se* and its technical, chemical, and optical relationship to the subject in front of the camera, or is formed instead through the culture which surrounds it. Benjamin described how photographic reproduction and mass distribution irrevocably altered the individual's encounter with the original, following a more general argument developed within the Frankfort School over the hegemonic power of cultural production in mass society. Yet it was not until the 1970s that a number of theorists began to address the specific problem of photographic meaning, particularly issues related to what and how the photograph represents. Roland Barthes (1977) used a semiotic analysis to unlock the problem of 'the photographic paradox': that is, the photograph represents an analogic reproduction of reality or a message without a code, and simultaneously a connotation or coded message, representing what society thinks about the subject. The major task of the denotative aspect of the image, according to Barthes, is to naturalize the cultural, that is, to transform cultural values into a message that can be 'read' as self-evident, as nature.

The deconstruction of the camera as a cultural artifact and of photography as deeply informed by culture continued on several parallel fronts. The model of the observer that dominates Western science, defined by the Cartesian perspectivalism and the direct gaze, is encoded in the optical system of the camera. The model thus defines a particular relationship between seeing and knowledge that privileges the authority of this gaze and its representations. This 'way of seeing' was critiqued most forcefully in relation to documentary photography for its dehistoricized view of marginalized social groups, a view that in turn established a visual code for representing the Other (Bolton 1989). Because photography has been constructed as a predominantly male practice, reinforced by the observer's gaze as male, female also becomes Other. Extensions of this critique to the uses of photography in the social sciences have implicated anthropology in particular for constructing its subjects as visually exotic. Related to this phenomenon, as mentioned above, is the institutional gaze, embodied and perpetuated through the archive, the museum, or the press (Tagg 1988).

In the 1970s, anthropologists also began to criticize the use of film and photography in fieldwork. Instead of being informed by anthropological concepts, the Western anthropologist's visual records reflected 'his' own culture, coded through the practices of family photography. Within the subfields of visual anthropology and sociology in the United States and Britain and their influential journals, this insight was variously developed into a broader perspective on photography, film, and media in general as sources for studying the cultural values and perspectives of those who create the images. This in turn supported a shift from image analysis to the study of the social and cultural practices in which photography is embedded. This complemented the critique being directed at the field of anthropology as a whole for perpetuating its colonial heritage through techniques of 'writing culture' that represented its subjects as Other (Ruby 1982). Anthropology's continuing interest in the photography of indigenous peoples and other subcultural groups reflects on the one hand a recognition of photography as a significant source for cultural analysis and on the other, attempts by the field to integrate the anthropological subject's own understandings into research and reports. Current ethnographic studies of media practices frequently include investigations of how images are interpreted and used by ordinary people. The photographic practices of professionals has been largely neglected, however, apart from the many biographies of well-known photographers and analyses of their work.

Digital techniques of making, reproducing, and transmitting photographs have contributed to undermining the authority of the photograph as analog or direct copy of reality. In the mid-1980s, the digitized image was seen by many, particularly within the press, as a direct challenge to established beliefs about how the photograph represents. Gradually the technique has been integrated into professional and private practices of making and distributing images, stimulating the rise of new media networks. The immediacy of image transfer and the portability of the technology have vastly increased the number of photographs in circulation, including those depicting war and human suffering. Some analysts argue that this phenomenon results in 'compassion fatigue,' reducing the moral impact of the image (Moeller 1999). In the meantime, new applications of interactive image technologies are emerging in the arts, in youth culture, and in diasporic

communities. The digital transfer of images globally has given rise in some quarters to a utopian vision with the photograph as a cornerstone of a global visual culture. More realistic analyses point to the diversities of culture worldwide and the continued vast inequalities in access to these technologies as critical impediments to a hegemonic visual culture.

See also: Mass Communication: Normative Frameworks; Mass Communication: Technology; Mass Media: Introduction and Schools of Thought; Mass Media, Representations in; Photojournalism; Scientific Instrumentation, History and Sociology of; Vision, Psychology of; Visual Anthropology; Visual Images in the Media; Visualization: Representation in Science

Bibliography

Barthes R 1977 *Image Music Text*. Hill and Wang, New York

Bateson G, Mead M 1942 *Balinese Character: A Photographic Analysis*. Special Publications of the New York Academy of Sciences 2, New York

Benjamin W 1968 The work of art in the age of mechanical reproduction. In: Arendt H (ed.) *Illuminations. Walter Benjamin*. Schocken, New York

Bolton R (ed.) 1989 *The Contest of Meaning: Critical Histories of Photography*. MIT Press, Cambridge, MA

Bordo S 1993 *Unbearable Weight: Feminism, Western Culture, and the Body*. University of California Press, Berkeley, CA

Chalfen R 1987 *Snapshot Versions of Life*. Bowling Green State University Popular Press, Bowling Green, OH

Collier J Jr, Collier M 1986 *Visual Anthropology. Photography as a Research Method*. University of New Mexico Press, Albuquerque, NM

Edwards E (ed.) 1992 *Anthropology and Photography, 1860–1920*. Yale University Press, New Haven, CT

Edwards E 1997 Ordering others: photography, anthropologies and taxonomies. In: Roberts R (ed.) *In Visible Light*. Museum of Modern Art, Oxford, UK

Goffman E 1979 *Gender Advertisements*. Harper & Row, New York

Irigaray L 1985 *This Sex which is not One*. Cornell University Press, Ithaca, NY

Moeller S 1999 *Compassion Fatigue. How the Media Sell Disease, Famine, War and Death*. Routledge, London

Newhall B 1949/1982 *The History of Photography: from 1839 to the Present*. Museum of Modern Art, New York

Rosenblum N 1984 *A World History of Photography*. Abbeville Press, New York

Ruby J (ed.) 1982 *A Crack in the Mirror. Reflexive Perspectives in Anthropology*. University of Pennsylvania Press, Philadelphia, PA

Solomon-Godeau S 1991 *Photography at the Dock: Essays on Photographic History, Institutions, and Practices*. University of Minnesota Press, Minneapolis, MN

Sontag S 1973 *On Photography*. Farrar, Straus and Giroux, London

Stasz C 1979 The Early History of Visual Sociology. In: Wagner J (ed.) *Images of Information*. Sage, Beverly Hills

Tagg J 1988 *The Burden of Representation. Essays on Photographies and Histories*. University of Massachusetts Press, Amherst, MA

Williamson J 1978 *Decoding Advertisements: Ideology and Meaning in Advertising*. Marion Boyars, London

K. Becker

Photojournalism

Photojournalism emerged in two distinct phases: during the late nineteenth century with the use of photographic images in newspapers and magazines as a new form of covering the news, and during the 1920s when photojournalism matured with the aid of technological advancements and was recognized as a form of documentary expression in most parts of the world. Unlike the spread of printing—or television later—photojournalism as a professional practice appeared almost simultaneously throughout Europe, the United States, and elsewhere. Thus, by the mid-1920s photographic expansion in Japan included the use of Leica cameras and access to the visions of *Neue Sachlichkeit* and Bauhaus practices in Germany, while similar developments in China were also directed by Western influences, although (between 1966 and the 1980s) the Cultural Revolution slowed the spread of photojournalism.

The beginnings of photojournalism occurred in the middle of the nineteenth century: photographers documented the US American-Mexican war in the 1840s; James Robertson and Felice Beato (England) photographed uprisings in India and China in the 1850s and 1860s; while Roger Fenton (England) covered the Crimean war in 1855 and Francis Frith (England) visited the Nile in 1856. Mathew Brady (USA)—with the aid of Alexander Gardner and George N. Barnard, among others—produced Civil War images in the United States between 1862 and 1864, and Jimmy Hare (USA) began working as a staff photographer for the *Illustrated American* in 1889. He later traveled to Cuba and also photographed the Russo-Japanese War, the Mexican Revolution, the Balkan War, and World War I before his retirement in 1922. Frequently, travel photographers and those accompanying expeditions, also worked as photojournalists.

The rise of photojournalism parallels the technological, social, and economic developments of the times. Technological innovations included the perfection of printing techniques to reproduce photographic images (halftones): beginning with the *New York Daily Graphic* (USA), 1873, or *Le Journal Illustre* (France), 1886, although the US press did not use halftones regularly until 1897. Also important was the development of flexible, transparent film (George Eastman, 1889), and the invention of smaller cameras (Ermanox, Leica/Germany, 1925), faster films (35mm, Kodak/USA, 1923, reversible emulsion (AGFA/Germany, 1923), and the introduction of

electronic flash equipment (Vacu-Blitz/Germany, 1929; Sashalite/England, Photoflash Lamp/USA, 1930).

The growth of photojournalism was advanced by the expansion of the reading public—with the addition of the working class—accompanied by an increased availability of inexpensive newspapers and picture magazines. Also, the public desire to 'see' the world—fuelled by the popularity of the movies (and newsreels, in particular) especially during the 1920s—created new demands on print media to legitimize photojournalism as a professional practice. In addition, the role of photography in education and propaganda, particularly in the Soviet Union, and its use in authoritarian societies, like Hitler Germany, energized photojournalism and strengthened its commitment to the ideals of social or political revolutions while providing increased picture coverage of the respective societies. The rise of worker-photographers in the Soviet Union and Germany during the 1920s also extended the notion of photojournalism to include the contributions of the working class to a photographic surveillance of the social environment. However, in post-1949 China, the use of photographs by the press remained restricted until much later (e.g., the first national photojournalism contest was held in 1981).

Photojournalists of the modern era found increasing numbers of professional outlets with the rise of mass circulation picture magazines, among them: *Arbeiter-Illustrierte-Zeitung*, *Berliner Illustrirte Zeitung*, *Vu*, *Paris Match* (France), *La Vanguardia* (Spain), *Picture Post* (England), *Look*, *LIFE* (USA), *Ogonyok* and *USSR in Construction* (USSR), or *China Pictorial* (China). Although photo agencies had existed since the end of the nineteenth century (e.g., Bain News Service, 1895, Underwood & Underwood, 1901, International News Photos, 1909 all in the USA) and photographers had been employed in considerable numbers by the press since the beginning of the twentieth century, increasing demands for photographic coverage by a rising number of media clients since the 1920s resulted in the establishment of new picture agencies. Photographs became commodities, and photographic agencies with their own staffs and national or international visibility produced and disseminated single images and photo stories. Among these photo agencies were: Mauritius, Weltrundschau and Dephot (Germany), Keystone, Rapho (France), Soyuzfoto, Novosti Picture Agency (Soviet Union), Acme Newspictures, Black Star, Wide-World Photos (USA), Scala (Italy), Nippon-Kobo (Japan) or—after World War II—agencies like Magnum, Viva, Gama, and Sygma (France), Contact Press Images (USA), and—since 1949 in China—the Bureau of Photojournalism, a government agency, reorganized as the Department of Photojournalism in the New China News Agency in 1952.

Together with traditional news agencies like the Associated Press, USA, or Reuters, England, these agencies became wholesale suppliers of visual material for the world press, working as commercial sources or cooperatives while setting the professional standards of modern-day picture coverage. They were aided by the development of wire transmission (invented in 1905 by Alfred Korn in Munich, Germany) in the early 1920s (London-New York over the Bartlane Cable system). The Associated Press New Photo Service began in 1927 and delivered photographs by train and plane; its Wirephoto network (1935) and the United Press Telephoto service (1936) helped increase the traffic in photographs, but made local editorial staffs less dependent on local photo coverage. At the end of World War II, the (New York) Times Photo service offered a worldwide news photographs service. The first satellite transmission of photographs in the United States occurred in 1960, followed by the launching in 1962 of the Telstar satellite for live transmission of pictures between Europe and the United States. Laserphoto was introduced by the Associated Press (USA) in 1977 and replaced by Photostream in 1991, an innovative transmission system that uses digital technology. The Associated Press (USA) had patented a portable digital image scanner in 1983 for the transmission of photographs from remote locations. Digital cameras appeared in 1988, and digital darkrooms in 1989 (*San Francisco Examiner*). The Associated Press (USA) began distributing news photographs via the World Wide Web in 1995.

Photo agencies also became depositories of visual representations of public life during most of the twentieth century (the Hulton Deutsch Collection (England) is one of the largest picture libraries in the world with over 15 million images; the Associated Press (USA) maintains an active picture archive with over 500,000 images, accessible over the Internet). The combined organizational network of photo agencies worldwide and their archival holdings constitutes a formidable, if not dominating cultural and commercial force in the process of visual representation. To a significant extent, journalism everywhere relies on the images that photo agencies supply.

By the end of the 1920s photojournalism was firmly established worldwide in the metropolitan press with the hiring of picture editors and the use of staff photographers in the editorial process. Smaller newspapers were considerably slower in creating staff positions for photographers. Over the years, the work of photojournalists has been preserved in the public and private archives of the press around the world. Among the most prolific photojournalists of modern times are Erich Salomon, Alfred Eisenstaedt, and Wolfgang Weber (Germany), Max Alpert, Arkady Shaikhet (Soviet Union), Eugene Smith, Margaret Bourke-White, Weegee, David Douglas Duncan, Mary Ellen Mark, Susan Meiselas (USA), Werner Bischof, René Burri (Switzerland), Robert Capa (Hungary), Ihei Kimura (Japan), Wu Yinbo, Zhang

Suicheng (China), Shomei Tomatsu (Japan), Peter Magubane (South Africa), Henri Cartier-Bresson, Martine Franck, Gilles Peress (France), Sebastiåo Salgado (Brazil), Bert Hardy, Don McCullin (England).

Photojournalism succeeded in making the world a knowable, visual experience and in creating a sense of familiarity with people and places. The presence of photojournalism in the public sphere has a profound effect on the public staging of politics. Photojournalism, by its very nature, generates attention, and because of its impact, raises issues of privacy, sensationalism, and manipulation in a marketplace that thrives on the exploitation of the image. Unlike others, photojournalists—to be effective—must seek the proximity of unfolding events, including the hazards of social unrest, revolution, or war. Consequently, they often risk their lives and sometimes pay the ultimate price in their pursuit of the perfect picture (by 1971, for instance, 25 of 35 non-Vietnamese journalists killed during the Vietnam war were photographers or cameramen).

The practice of photojournalism, beginning in the 1920s, led to an exploration of the documentary style and produced single and multiple images to convey the news. Photographs became a regular staple of newspapers and affected their make-up and design. When photographers thought editorially, the result was a new vocabulary of journalism. But, as Evans (1978/1997, p. 255) suggests, 'photography is not photojournalism'; instead, words are needed to supply more information than are conveyed by a photograph. Words help explain, confirm, and reinforce, they also place events in time and space. According to Hicks (1952, p. 5), 'the basic unit of photojournalism is one picture with words.' Photojournalism acknowledges the importance of words, either provided in the form of outlines or stories authored by the photographer or journalist-colleagues.

The picture story (or photo essay) constitutes the long form of photojournalism; its visual narrative typically focuses on a single subject and is the outcome of careful planning and extended professional commitment to a project. Ideally, the resulting presentation—not unlike an essay—consists of a flow of images, characterized by a strong opening, a descriptive middle passage, and a decisive ending. Pioneered first in the Soviet Union and Germany during the 1920s, the photo essay became a determining characteristic of *LIFE* magazine in the United States; it challenged the imagination of photographers and editors to produce a durable visual narrative that could rely on the popularity of story telling to satisfy reader interests. There is also an affinity between photojournalism and photographic fiction; for instance, photo-novelas as social reportage constitute a form of visual narrative, which has been explored since the 1970s (as literary photography or photographic writing) by contemporary photographers.

With the advent of television in the 1950s and the demise of the mass circulation picture magazine (particularly in the United States) during the 1960s, photojournalism survives in news magazines and the daily press, in particular, where it continues to help strengthen the visual culture of the late twentieth century (aided by the ease of offset printing which improved the reproduction quality of images). The introduction (since the 1930s) of color positive and negative films (with distinctive processes developed by Agfa, Cibachrome, Kodak, and Polaroid) and their increasing use by the press (since the 1960s), added a new dimension to photojournalism; it enhanced the documentary quality of images and contributed to their visual impact (*LIFE* magazine published the first color photographs in 1952). Although black-and-white photographs maintain their abstract, documentary quality (and are used for this purpose in modern advertising), the work of contemporary photojournalism relies almost exclusively on the production and circulation of color photographs.

A distinctively different form of photojournalism emerged with the demise of picture magazines in the United States and the rising popularity of the book as an alternative medium for the dissemination of visual expressions. The book celebrates authorial prerogative and individuality of form and style; it also offers opportunities for exercising creative freedom and extended personal constructions of particular aestheticized visions of the world that favor a humanistic approach to photography and avoid claims of objectivity. The resulting emergence of a new kind of photojournalism at the end of the twentieth century borrows freely from the tradition of street photography and photographers ranging from Robert Frank, William Klein, and Garry Winogrand (USA), to Henri Cartier-Bresson and Robert Doisneau (France), or Josef Koudelka (Czech Republic). The sustained, book-length visual argument can be powerful, but when photojournalism catered more exclusively to a privileged class of bourgeois readers, experts, and connoisseurs of fine photography, it crossed into the realm of art: single news photographs appeared on museum walls and became commodified objects in the commercialized sphere of art collection (especially since the mid-1970s when photographs began to be featured regularly in art galleries).

In an image conscious world photojournalism survives and its tradition is preserved and celebrated by institutions such as the International Center of Photography (USA), or the World Press Photo Foundation, organizations such as the National News Photographers Association (USA) and educational programs, such as the University of Missouri School of Journalism (USA), all of which stand for similar organizations and activities in a number of countries.

See also: Journalism; Journalism and Journalists; News: General; Photography as a Medium

Bibliography

Alabiso V 1998 *Flash! The Associated Press Covers the World.* Abrams, New York

Barret A 1978 *Die ersten Photoreporter, 1848–1914.* Krüger, Frankfurt, Germany

Carlebach M L 1997 *American Photojournalism Comes of Age.* Smithsonian Institution Press, Washington, DC

Chapnick H 1994 *Truth Needs No Ally. Inside Photojournalism.* University of Missouri Press, Columbia, MO

Evans H, Taylor E 1978/1997 *Pictures on a Page. Photojournalism, Graphics, and Picture Editing.* Pimlico, London

Gidal T 1973 *Modern Photojournalism: Origin and Evolution, 1910–1933.* Macmillan, New York

Harrison M 1998 *Young Meteors. British Photojournalism, 1957–1965.* J Cape, London

Hicks W 1952 *Words and Pictures.* Harper and Row, New York

Hopkinson A (ed.) 1995 *150 Years of Photojournalism.* Könemannn, Cologne, Germany, Vol. 2

Lewis G 1991 *Photojournalism. Content and Technique.* Brown, Dubuque, IA

Miller R 1998 *Magnum: Fifty Years at the Front Line of History: The Story of the Legendary Photo Agency.* Grove/Atlantic, New York

Neubauer H 1998 *Black Star. 60 Years of Photo Journalism.* Könemann, Cologne, Germany

Tausk P 1988 *A Short History of Press Photography.* International Organization of Journalists, Prague, Czech Republic

Yapp N (ed.) 1995 *150 Years of Photo Journalism.* Könemann, Cologne, Germany, Vol. 1

H. Hardt

Phylogeny and Systematics

Systematics is the study of the relationships among organisms, whereas phylogenetics is the study of the evolutionary relationships among organisms. In that sense, phylogenetics is a subset of systematics. Historically, before evolution was accepted as the unifying principle in biology, phylogenetics did not exist *per se*. Instead, organisms were classified on the basis of their phenotypic similarity.

1. Defining the Questions

One of the main problems with using phenotypic similarity as a basis for classification is that organisms evolve to become adapted to the environments in which they live (their niches) so that phenotypic convergence is a fairly common occurrence. For the most part, the longer ago in time that two organisms shared their most recent common ancestor, the more phenotypically dissimilar they will have become. However, phenotypic convergence as a consequence of adaptive evolution, along with rapid divergence as organisms adapt to totally novel environments, means that any classification based on phenotypes inevitably confounds phylogenetic relationships and adaptive evolution. Given the evolutionary lability of behavioral characteristics, analysis of the adaptive or functional significance of particular traits requires that the phylogenetic relationship is properly defined so that evolutionary changes in behavior can be mapped on to those in niche use.

The advent of cladism as a technique aimed at identifying phylogenetic groupings through phenotypic analysis was a notable first attempt to isolate phylogenetic groupings. Cladism, as initiated by Hennig (1966), is, however, based on a model of evolutionary change which excludes the possibility of phenotypic convergence because it assumes that the same character cannot evolve on more than one occasion and that once a character has evolved it cannot revert to its former state. More recently, the field of phylogenetic analysis has been revolutionized by the rapid production of gene sequence data. Realistic models of molecular evolution are incorporated into phylogenetic analyses of various regions of the genome to produce phylogenies (evolutionary trees) of ever-increasing accuracy. How those trees will be used to produce classifications is not clear.

At the moment, recognized groupings (genus, tribe, family, and so on) are being revised so that they describe monophyletic groups (containing all the contemporary descendants of a particular common ancestor). However, the nominal groupings involved differ according to historical precedence: systematists who originally produced the hierarchical classifications in use today ranged along an axis from lumpers to splitters. The former were less keen to classify species with a particular degree of morphological similarity into different higher level taxa than were the latter. As a consequence, even when species within the hierarchical groupings are rearranged to represent monophyletic groupings, the most recent common ancestor for all the species in a particular genus would have been earlier in some orders than the most recent common ancestor for two species separated at the family level in other orders.

The development of this article introduces natural groupings, explains how phylogenies are determined, and then shows how phylogenies can be used to understand the functional significance of behavioral differences among taxa.

2. Natural Groupings

It has always been clear to modern humans that there are natural groups of organisms: trees, birds, flowering plants, spiders, and tortoises would be cases in point. It has also been clear, at least in more recent times, that there is a natural hierarchy of groups: there are animals versus plants, but within animals there are vertebrates and invertebrates, and within vertebrates there are birds, reptiles, mammals, and amphibians. The embodiment of the hierarchical natural classification was produced by Linnaeus around 1800. No particular

mechanism was envisaged for why this natural hierarchical classification should exist, but one frequently stated reason for studying it was to provide insight into the mind of the creator. With the benefit of hindsight, insofar as classifications were based on functional similarity, what were recognized to be the same character could in fact have separate evolutionary origins.

The major sea change occurred in 1859 with the publication of Charles Darwin's *Origin of Species* in which he argued that life was created on one or few occasions and that subsequent speciation resulted in the natural hierarchy. Indeed, it is worth remembering that the only figure in the book is of an evolutionary tree. From that time on, with a few notable lapses, it has been accepted generally by working scientists that natural groupings are a consequence of evolution and that they should be monophyletic. In practice, however, it did not always turn out that way. For example, it is accepted that birds share a more recent common ancestor with crocodiles than either does with lizards. As a consequence, crocodiles should be classified with birds rather than with lizards. This means that the reptiles do not constitute a monophyletic group, but what is known as a paraphyletic group: a group made up of only those descendants of a common ancestor that resemble it (or each other) phenotypically.

Despite the acknowledged fact that the reptiles constitute a paraphyletic group which, if it were to be made monophyletic, should include the birds, there has been little serious effort on the part of biologists to execute this particular taxonomic revision. To do so, would in some sense apparently cost more effort than it is worth. Instead, reptiles without the birds are accepted as a paraphyletic natural group, while palaeontological and other evolutionary analyses should include birds within the group.

3. *Identifying Phylogenetic Relationships*

There have been many schools of systematics, and most are of merely historical rather than contemporary importance. The major division is into methods that analyze phenetic relationships vs. those that attempt to reconstruct phylogenetic relationships.

The development of phenetic methods became particularly popular in the 1960s and 1970s following the publication of Sokal and Sneath's (1963) book entitled *Principles of Numerical Taxonomy*. The book outlined apparently objective methods of character scoring and the production of phenograms (trees that summarized degrees of phenetic similarity), using the then newly and widely available computers. The fatal flaw to the whole enterprise was the absence of a model of phenetic evolution. There is an infinity of so-called clustering algorithms and distance methods, each of which is appropriate for a particular model of phenetic evolution. Unfortunately, there was no objective way to decide which to use. It might be argued that the aim of numerical taxonomy or phenetics is not to determine evolutionary relationships but rather to describe phenetic similarity, but even then there is an infinite number of distance methods and clustering algorithms, with no principled method of deciding on appropriate ones for all circumstances. The exercise foundered.

The use of phylogenetic methods was given a major boost by Hennig's (1966) book *Phylogenetic Systematics*. Hennig realised that if the phylogeneticist was willing to make certain assumptions about character evolution, then phylogenetic trees could be constructed from data based on the distribution of character states among extant taxa. If every character evolved only once and evolutionary change was irreversible, then monophyletic groups could be defined on the basis of shared derived characters. Problems arose when different characters produced different answers so that one character placed a particular species in one group while another character placed it in a different group, which meant that at least one character did not evolve according to Hennigian principles. Common sense often prevailed among cladists to distinguish between, for example, characters that might have evolved once and been lost versus those that evolved more than once. However, once cladists began to entertain the possibility that there are different models of evolution for different characters, then the advent of a model-based statistical approach to phylogenetic reconstruction became inevitable. The problem recurs, just as it did with numerical taxonomy, that for phenetic evolution there is frequently no principled way to distinguish among alternative models of evolution.

In the years immediately before genes were sequenced at such low cost and with such speed, an informative stepping stone between phenetic and phylogenetic analysis occurred with the advent of DNA–DNA hybridization technology. The method was developed and widely applied by C.G. Sibley and co-workers. Radioactively labeled DNA strands about 500 nucleotides long from a focal species were hybridized with similar strands from a number of phylogenetic relatives. The rate of dissociation of these duplexes provided a measure of genetic similarity between the sequences, thereby excluding purely morphological convergence as a factor producing errors in phylogenetic tree reconstruction. Several surprises emerged. For example, among the birds, Australian passerines seemed to be more closely related to each other than they are to passerines from other continents, thus suggesting that previous classifications based on morphological comparisons had failed to detect considerable convergent morphological evolution. Sibley and Ahlquist obtained representative samples of many thousands of bird species, representing the large majority of tribes. They then subjected the data to DNA–DNA hybridization analysis

to produce an estimated phylogeny of the birds based simply on the rate of dissociation of DNA–DNA hybrid duplexes. That phylogeny was used by Sibley and Monroe (1990) to produce a classification of birds: categories of rates of duplex dissociation were arbitrarily defined that were associated with each particular taxonomic status. Despite the various analytical difficulties of their method, and the arbitrary nature of their classification, their procedure was claimed to be both objective and repeatable. As a first attempt at a full phylogenetic classification for a major taxon, it has probably yet to be improved upon. However, while most of the major revised phylogenetic relationships that Sibley and co-workers revealed have remained intact, more appropriate methods for phylogenetic reconstruction from gene sequence data are now available.

DNA–DNA hybridization methods provide just one measure of the distance separating a pair of sampled individuals. That distance depends on the genetic similarity between those individuals in terms of base sequences. However, the same distance measure can be achieved in many different ways: if there are 20 base differences between individuals 1 and 2, and there are 20 base differences between individuals 2 and 3, that does not tell us how many differences there are between 1 and 3—just that there are between 0 and 40. All pairwise DNA–DNA hybridizations would need to be completed, even if we were accepting that DNA diverged at a metronomic rate through evolutionary time.

The production of vast amounts of gene sequence data means that every base pair becomes a character. Once we have a base sequence from a particular gene of a particular individual, it can be compared with that for every other individual for which we have equivalent data. This allows phylogenetic reconstruction using discrete rather than distance measures. What is more, we can describe models of molecular evolution with reasonable accuracy: we may know the relative substitution rates of the first, second and third base pairs, and we often have a good idea of transition to transversion ratios (the frequency with which a substitution is between purine and purine or between pyrimidines, vs. between purine and pyrimidine). We also know that some base substitutions are effectively neutral because they result in the same amino acid being coded for at the base-substituted triplet.

Before mentioning the most commonly used methods for discrete data, it should be pointed out that distance measures can be used and are used with gene sequence data. The methods differ in whether they assume branch lengths from the root to the tips of the tree are all equal and adjust the data accordingly (e.g., unweighted group pair method with arithmetic means: UPGMA) or whether they allow different rates of evolution in different parts of the tree (e.g., the neighbor joining method which often gives a good approximation to the minimum evolution tree). Despite the fact that information is lost (the evolution of individual base sites cannot be traced down the tree), there is also the problem that approximations are made which often result in biologically impossible branch lengths being reconstructed (see Page and Holmes 1998).

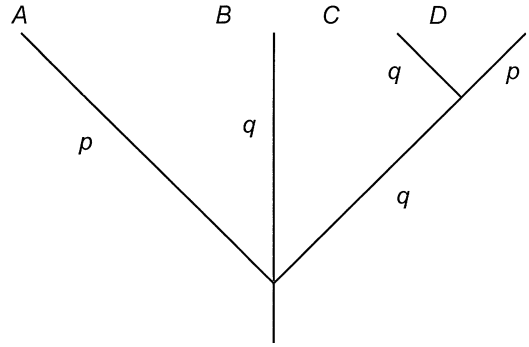

Figure 1
A phylogenetic tree that is likely to produce data that will cause parsimony or compatibility methods to give misleading results. The correct tree is shown, with p and q being probabilities of character state change along the different branches of the tree. The wrong tree is likely to be determined when p is much larger than q because then B and C are likely to be more similar than are C and D (after Felsenstein 1978, 1983)

The two most common model-based discrete methods of model based phylogenetic tree reconstruction from gene sequence data involve parsimony and maximum likelihood.

Parsimony was introduced as the so-called minimum evolution distance method by Cavalli-Sforza and Edwards (1967) as a computationally rapid route for approximating maximum likelihood trees. However, as a discrete method for finding the shortest tree, it is possible to adapt strict parsimony and weight the changes in appropriate ways (e.g., a transition might be four times as frequent as a transversion) and determine that tree with the smallest total of weighted steps. There are several problems with parsimony as a procedure. The first is philosophical—parsimony is a minimum criterion by definition and it may be argued that there is no minimum principle to evolution. The second major problem is that of 'long branch attraction' (see Fig. 1). Despite these disadvantages, parsimony remains a commonly-used procedure because of its computational speed.

Maximum likelihood is a very powerful tool in phylogeny reconstruction. The method asks which phylogenetic tree is the most likely one to have given rise to the set of gene sequences that we observe in our sample from the present day, given our specified model of gene sequence evolution. Indeed, it can go further and be used to estimate the most likely parameter estimates of our evolutionary model. The likelihood

approach is increasingly commonly used for the reconstruction of molecular phylogenies, and as computers get ever faster and more powerful, it will probably become the approach of first choice.

4. Using Phylogenetic Relationships to Analyze Behavior

When behavior is mapped on to phylogenetic tree structure, at least four types of informative analysis become possible. First, behavior itself may be seen to be responsible for phylogenetic tree structure. For example, species with particular modes of behavior may be more likely to divide into new species than others. Second, by mapping behavioral characteristics on to phylogenies, it is possible to determine which traits are ancestral and which are derived, and to determine which traits have changed most frequently over evolutionary time. Third, behavioral characteristics do not evolve independently of each other or of habitat use. If we are to understand why character covariation occurs, then phylogenetic tree structure must usually be an integral part of the necessary analysis. Fourth, adaptive radiations, such as Darwin's Galapagos finches, are well known but when we map behavior on to tree structure, it sometimes becomes apparent that independent adaptive radiations have remarkably similar structures—when the tape of life is replayed, similar arrays of diversity result. In this section, examples of each of these four types of phylogenetically-based behavioral analysis are described.

4.1 Behavior and Speciation

When behavioral traits are mapped on to phylogenies, it is sometimes possible to identify so-called 'key innovations' which appear to have resulted in high rates of speciation. For example, cichlid fishes are very speciose and their radiations seem to have coincided with a change in position of a particular muscle attachment which allowed the pharyngeal bones to manipulate individual prey items while still holding them. This released other areas of the jaw, which had been used for manipulating prey, to evolve for other functions. The cichlid radiations in African lakes are indeed characterized by a remarkable variety of feeding mechanisms. However, a problem with this explanation is that there may be other unrecognized characteristics unique to cichlid fishes that may also contribute to their explosive radiations.

A more convincing demonstration of key innovations resulting in increased rates of speciation is when the same innovation appears independently in different parts of the phylogenetic tree, and is statistically associated across a number of cases with an increased rate of speciation. One of the best examples of a repeatedly evolved such innovation is the ability of insects to feed on vascular plants, known as phytophagy. There is a wide variety of vascular plants, and whenever phytophagy evolved, there are likely to have been many vacant niches to speciate into. Mitter et al. (1988) identified 13 nodes in a phylogenetic tree of insects where all the descendant species from one branch are phytophagous, whereas all those down the other branch are not. In 11 of 13 cases there were more extant species descended from the phytophagous branch than the nonphytophagous branch, and in each of those 11 cases there were more than twice as many phytophagous than nonphytophagous species.

A final example of behavioral changes apparently causing differences in rates of speciation comes from an original suggestion made by Darwin (1871). It is known that the females of some bird species actively choose their mate on the basis of phenotype. Darwin suggested that, when populations separate, the precise criteria for mate choice might alter so that if those separated populations came back together, mate choice would have promoted speciation. The most obvious criterion for determining whether females choose their mates by phenotype is whether males differ from females in physical appearance. When passerine birds are scored as sexually dichromatic or not, then it turns out that the repeated evolution of sexual dichromatism tends to be associated with subsequently higher rates of speciation (Barraclough et al. 1995). It had taken more than 100 years before the phylogeny became available so that Darwin's theory could be tested.

4.2 Character Evolution

Mapping character evolution on to phylogenetic trees can be used to test ideas about how and why characters evolve. In order to carry out this process, it is necessary to have a phylogenetic tree showing the relationships among contemporary species, character states for each species, and a model for character evolution. Since the appropriate model for the evolution of most characters is poorly understood, ancestral character states usually are inferred from either a parsimony or other appropriate minimum evolution criterion.

Gittleman (1981) set out to test the idea that when patterns of parental care evolved among fishes, transitions were gradual and did not involve change by more than one parent. Twenty-one transitions were identified using parsimony on the phylogenetic tree, all of which were between biparental and maternal care, biparental and paternal care, no care and paternal care, and no care and maternal care. No cases involved the double transitions between no care and biparental care or paternal care and maternal care, a statistically significant deficiency if all transitions were equally likely.

A more thorough and quantitative analysis of historical patterns in parental care, this time among shorebirds, was reported by Székely and Reynolds

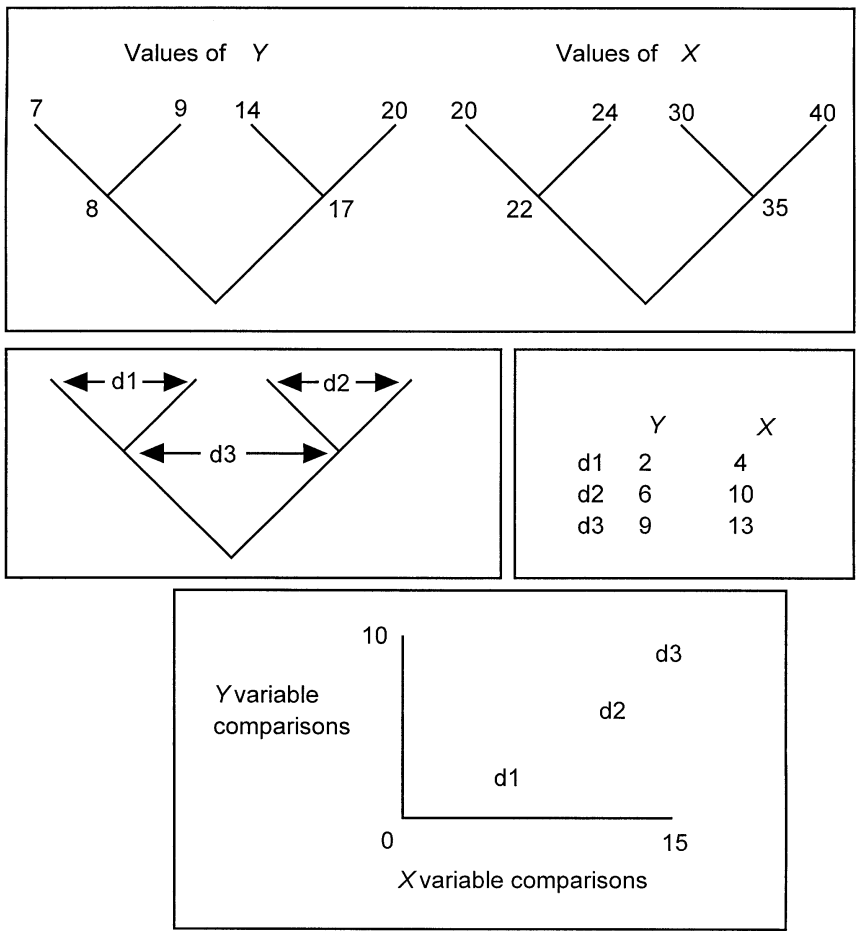

Figure 2
Comparative method of independent contrasts. Comparisons d1, d2, and d3 are independent of each other under a Brownian motion model of evolution for characters X and Y. If the two characters have shown uncorrelated evolution, values of the contrasts for one character will not be correlated with those for the other character

(1995). Again they used parsimony for ancestral character state reconstruction, but in this taxon they found that there had been an overall evolutionary reduction in paternal care, which was sometimes but not always compensated for by an increase in maternal care. The authors argued that this is a response to selection for males to partake in extra-pair matings or even to maintain multiple pair bonds.

How rapidly does behavior evolve in comparison with, say morphology and life history? Haldane (1949) suggested that one appropriate measure of evolution change was the 'darwin': a change of one logarithmic unit per million years. Gittleman et al. (1996) mapped the evolutionary change of two behaviors (group size, home range size), two life history characters (gestation length, birth weight), and two morphological characters (adult brain weight, adult body weight) using data for extant mammal species from eight taxa for which there were reasonable phlylogenies. The behaviors were more labile than the morphological characters, which were more labile than the life history characters. Such analyses may well identify characters which, for whatever reason, are the most evolutionarily constrained, leaving others more free to change.

4.3 Correlated Character Evolution

Bringing order to organic diversity is made simpler by two organizing principles: close relatives are more similar to each other than are distant relatives, and organisms share adaptations to common environments. As we have seen when attempting to reconstruct phylogenetic trees, it can be important to distinguish between these two reasons for similarity. However, the two principles cannot necessarily be

separated: because close relatives are often adapted to very similar ecological niches, they also share many adaptations in common, whether those adaptations were inherited from a common ancestor or not. If we wish to understand the extent to which similar adaptations are likely to arise in similar circumstances, it is nevertheless important to factor out identity by common descent. What is more, because close relatives share characters which may not be functionally related to each other (e.g., mammals have hair and give birth to live young, whereas birds have feathers and lay eggs), treating species values as independent points in a statistical analysis can result in functionally meaningless significant associations between characters as a consequence of pseudo-replication.

4.3.1 Independent contrast analysis. In order to tackle this problem, Felsenstein (1985) devised a method which traced the change in two or more characters on a phylogenetic tree, and defined for each character equivalent evolutionary independent contrasts which could be used for data analysis. Given character states for contemporary species, ancestral character states had to be inferred from a specified model of evolutionary change. Felsenstein used a Brownian motion model of change to estimate ancestral character states. In essence, all change that occurred between sister taxa evolved independently of change between other sister taxa. The consequence is that differences between the actual or estimated character states for sister taxa are used as the independent contrasts, which are standardized by branch lengths (variances of contrasts are larger when the time separating taxa has been longer). The direction of comparison is arbitrary, though it is conventional to order the taxa so that all contrasts for the independent variable are positive (or zero). The contrasts for the dependent variable(s) then use the same order of taxa and can take both positive and negative values. The contrasts for the dependent variable(s) are then regressed on those for the independent variables, and conventional statistical techniques (correlation or regression) are used to test for correlated evolutionary change. The null hypothesis is that change is uncorrelated, and significant departures from the null hypothesis can indicate either positive or negatively correlated character change. A hypothetical example is worked through in Fig. 2.

Although independent contrast analysis is widely used, statistical conclusions can, in fact, change when different models of character change are entertained. For example, Harvey and Purvis (1991) report a study of the evolutionary relationship between a measure of running activity and relative foreleg length (foreleg length corrected for body size) in a group of Anolis lizards. Three models of evolutionary change were considered (Brownian motion, punctuation, and minimum evolution), only the first of which produced a significant relationship between contrasts for the two variables being studied ($p < 0.025$ compared with $p > 0.10$ for the other two models). Fortunately, diagnostics are available to test whether the assumptions of Brownian motion character evolution have been approximately met, given data for a set of contemporary species and a well-resolved phylogeny (Garland et al. 1992). However, when the hypothesis of Brownian motion character evolution is rejected, there remains a paucity of statistically powerful comparative methods.

4.3.2 Sister taxon comparisons. When a phylogeny is available, but there is no obvious model of character evolution, then the technique of sister taxon comparison can prove very useful. The idea is to compare pairs of contemporary species, chosen such that each 'sister pair' shares a more recent common ancestor than either member does with any member of any other pair. The result is that the difference between each pair of sister species has evolved independently of the difference between any other pair, and ancestral character states do not need to be estimated. The loss of assumptions results in fewer possible comparisons: if there are n contemporary species in a sample, so long as the phylogeny is bifurcating, then there are $n-1$ possible independent comparisons, but at most only $n/2$ sister comparisons.

Møller and Birkhead (1992) have championed the use of sister taxon comparisons, and illustrated their use by testing the idea that high copulation rates among birds are a 'paternity insurance' mechanism. First, they asked whether rates of extra-pair copulation increased with local population density, which they did in all eight possible comparisons, which is a highly significant result (n.b. 15 potential comparisons are possible with 16 taxa when ancestral states are considered). Second, they asked whether birds living in colonies had higher rates of intra-pair copulations, which they did in 12 of 13 comparisons.

4.4 Parallel Structure of Independent Radiations

When a species invades a completely new habitat with unoccupied niches, there is the potential for an adaptive radiation to ensue. But how deterministic is evolution? Stephen Jay Gould suggested that contingency was such an important factor that, if we replayed the tape of life, we should invariably get different outcomes. That is, of course, true *sensu stricto*. However, early comparisons such as that between the placental and marsupial mammal radiations, and the very fact of evolutionary convergence, means that it is not the whole story. Recently, a number of case studies suggest a remarkable degree of similarity when the opportunity for similar radiations has been presented.

Anolis lizards in the Caribbean have invaded many small and a few large islands. On the very smallest islands, there tends to be one intermediate-sized species, but as island size increases large and small forms can coexist. Each of the largest islands supports an independently evolved *Anolis* radiation, with about seven different species occupying a variety of well-defined ecological niches. To the casual human observer there has been remarkable convergence, for example, the species living on twigs from different islands are well-nigh indistinguishable. Natural selection has led to the evolution of the same so-called ecomorphs evolving on different islands. The twig ecomorphs have a morphology that is adapted to the niche in which they live. Other ecomorphs live in the crowns of trees, on the trunks, on the ground, and so on.

While it may be easy to fool a casual human observer, would the same independently evolved ecomorphs be able to fool each other? That experiment has not been performed on *Anolis* lizards, but it has in a simpler system involving stickleback ecomorphs. At the end of the Pleistocene, after the glaciers had retreated from British Columbia, a marine stickleback invaded a number of freshwater lakes. In those lakes, parallel speciation occurred: a large benthic ecomorph evolved which feeds on invertebrates from the bottom of the lake, together with a smaller limnetic ecomorph which darts around in the open water living off plankton. There is strong assortative mating such that the benthic and limnetic ecomorphs will only mate with their own type in the wild, and have become true species. But would opposite sex individuals of the same independently evolved ecomorph from different lakes accept each other as mates? The answer is yes, in the laboratory. Rundle et al. (2000) took limnetic and benthic species from three separate lakes. Over 750 mating trials were set up in which individual females were presented with single males of the same or the alternative ecomorph, from the same or a different lake. Females were far more likely to spawn with a male of the same ecomorph as themselves, in which case it mattered not at all whether the male was from the same or a different lake.

See also: Adaptation, Fitness, and Evolution; Comparative Method, in Evolutionary Studies; Evolution, History of; Evolution: Optimization; Evolution: Self-organization Theory; Evolutionary Approaches in Archaeology; Extinction of Species; Species and Speciation

Bibliography

Barraclough T G, Harvey P H, Nee S 1995 Sexual selection and taxonomic diversity in passerine birds. *Proceedings of the Royal Society London, Series B* **259**: 211–15

Cavalli-Sforza L L, Edwards A W F 1967 Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics* **19**: 233–57

Darwin C 1859 *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life*. John Murray, London

Darwin C 1871 *The Descent of Man and Selection in Relation to Sex*. John Murray, London

Felsenstein J 1978 Cases in which parsimony or compatibility methods will be positively misleading. *Systems in Zoology* **28**: 49–62

Felsenstein J 1983 Parsimony in systematics: Biological and statistical issues. *Annual Review of Ecology and Systematics* **14**: 313–33

Felsenstein J 1985 Phylogenies and the comparative method. *American Nature* **125**: 1–15

Garland T J 1992 Rate tests for phenotypic evolution using phylogenetically independent contrasts. *American Nature* **140**: 509–19

Gittleman J L 1981 The phylogeny of parental care in fishes. *Animal Behavior* **29**: 936–41

Gittleman J L, Anderson C G, Kot M, Luh H-K 1996 Comparative tests of evolutionary lability and rates using molecular phylogenies. In: Harvey P H, Leigh Brown A J, Maynard Smith J, Nee S (eds.) *New Uses for New Phylogenies*. Oxford University Press, Oxford, UK

Haldane J B S 1949 Suggestions as to quantitative measurement of rates of evolution. *Evolution* **3**: 51–6

Harvey P H, Purvis A 1991 Comparative methods for explaining adaptation. *Nature* **351**: 619–24

Hennig W 1966 *Phylogenetic Systematics*. University of Illinois Press, Urbana, IL

Mitter C, Farrell B, Wiegmann B 1988 The phylogenetic study of adaptive zones: Has phytophagy promoted diversification. *American Nature* **132**: 107–28

Møller A P, Birkhead T R 1992 A pairwise comparative method as illustrated by copulation frequency in birds. *American Nature* **139**: 644–56

Page R M, Holmes E C 1998 *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford, UK

Rundle H D, Nagel L, Boughman J W, Schluter D 2000 Natural selection and parallel speciation in sympatric sticklebacks. *Science* **287**: 306–8

Sibley C G, Monroe B L 1990 *Distribution and Taxonomy of Birds of the World*. Yale University Press, New Haven, CT

Sokal R R, Sneath P H A 1963 *Principles of Numerical Taxonomy*. Freeman, San Francisco

Székely T, Reynolds J D 1995 Evolutionary transitions of parental care in shorebirds. *Proceedings of the Royal Society of London, Series B* **262**: 57–64

P. H. Harvey

Physical Activity and Health

The past three decades have witnessed a burgeoning of research examining the impact of physical activity on physical and mental health. This research has conceptualized physical activity as both a physiological and a psychological phenomenon. From the physiological perspective, physical activity is usually defined

as any force exerted by skeletal muscles that results in energy expenditure above the resting level. From a psychological perspective, physical activity is viewed as a behavior performed in specific social settings and controlled by motivational and volitional factors. Both perspectives complement one another, and they apply to different research questions. This article summarizes findings that refer to physiological as well as psychological aspects of the relationship between physical activity and health.

1. *The Effects of Physical Activity on Physical Health*

1.1 *Overall Mortality*

Epidemiological studies suggest that persons with moderate to high levels of regular physical activity or cardiorespiratory fitness have lower mortality rates than persons who are inactive or have low fitness levels (e.g., Sherman et al. 1999). This is supported by longitudinal research examining how changes in the level of cardiorespiratory fitness may affect subsequent mortality risks. For example, in a study by Blair et al. (1995), healthy men with an initially low level of cardiorespiratory fitness participated in two exercise tests about five years apart. Five years after the second exercise test, men in the bottom fifth of the initial fitness distribution who improved their fitness to at least a moderate level had a 44 percent lower mortality rate than men who remained in the bottom fifth. Similar results had already been found by Paffenbarger et al. (1993) in a cohort of initially sedentary Harvard alumni. Those men who started and maintained moderately intense activities extended their longevity on average by 0.72 years.

1.2 *Cardiovascular Diseases* (*CVD*)

Cardiovascular diseases, in particular coronary heart disease and stroke, are major causes of death, disability, and health care expenditures in Western societies. There is clear evidence that, within populations, the incidence of CVD is inversely related to habitual physical activity levels (Bauman and Owen 1999). Some of the studies in this field demonstrated a dose–response gradient between level of physical activity and risk of CVD outcome. Dose–response relationships were also found in those investigations in which measures of cardiovascular fitness, instead of physical activity estimates, were related to CVD mortality.

1.2.1 Coronary heart disease (*CHD*). Since 1953, about 40 studies analyzing the association between physical activity and risk of (CHD) have been published. The first study in this field was conducted by Morris et al. (1953), who examined 31,000 employees of the London Transport Executive. A comparison between bus drivers and bus conductors revealed that physically active conductors clearly had a lower risk of a 'first coronary episode' than sedentary drivers. Classifying physically active and inactive people according to job title or occupational activity is typical for studies published before 1978. In the following years, level of physical activity was defined mostly by self-report measures of leisure-time activity or by indices that combined levels of occupational and leisure-time activities. More recent investigations have also emphasized the necessity to control for potentially confounding variables (e.g., age, sex, functional status) and to adjust for factors that are considered causal links between physical activity and CHD, such as serum cholesterol, blood pressure, diabetes, or body mass index.

Reviews of relevant literature have concluded consistently that physical activity is strongly related to a reduced risk of CHD. Furthermore, there is evidence in recent studies that this relationship holds true for both men and women. For example, findings from a prospective study among 72,488 female nurses who were 40–65 years old at the onset of the study suggest that sedentary women who became active in middle adulthood or later had a lower risk of coronary diseases than their counterparts who remained sedentary. Relative to women who walked infrequently, those who walked at a brisk pace three or more hours per week had a multivariate relative risk of 0.65 for coronary events (Manson et al. 1999). Physical exertion may temporarily increase the likelihood of an acute coronary event among persons with advanced coronary atherosclerosis who are not used to regular exercise. Although this risk needs to be taken seriously, it does not affect the general fact that physical activity is associated with a substantially lower overall risk for CHD.

1.2.2 Stroke. Contrary to the results for CHD, epidemiological data do not unequivocally confirm a relationship between stroke and level of physical activity. Most investigations in this field were planned as population-based cohort studies with follow-up ranges of 5–26 years. Among those studies that were designed to detect possible dose–response effects, many did not find a clear linear association. Some studies even suggested a U-shaped relationship between physical activity level and risk of stroke. By trying to explain these contradictory findings, it was assumed that the two different types of stroke, namely ischemic stroke and hemorrhagic stroke, need to be regarded separately. Since both types have specific pathophysiologies, it may be possible that physical activity affects them in different ways. In a study by Abbott et al. (1994), inactive men had a

higher risk of hemorrhagic stroke than active men, but there was no main effect on ischemic stroke. More research is needed to further clarify whether physical activity plays a protective role against stroke.

1.2.3 Biological mechanisms. Several biological mechanisms may be responsible for the beneficial effects of physical activity on cardiovascular risks. These mechanisms involve the impact on blood pressure, atherosclerosis, plasma lipid/lipoprotein profile, availability of oxygenated blood for heart muscle needs (ischemia), thrombosis, heart rhythm disturbances, overweight, and healthier distribution of body fat.

1.3 Cancer

Cancer is one of the leading causes of death in industrialized countries. It is predicted that this proportion will increase in the next few years. Numerous studies have examined the role of physical activity in the prevention of cancer.

1.3.1 Colon cancer. Systematic literature analyses reveal that colon cancer is probably the most often-studied cancer in the field of physical exercise epidemiology. In 1996, a total of 29 studies on colon cancer were published (DHHS 1996). Of these, 18 studies considered only occupational physical activity, and 11 examined leisure-time or total physical activity. The results of the 18 studies using a measure of occupational physical activity draw a rather consistent picture: 14 studies found a significant inverse relationship between level of activity and risk of colon cancer; in four studies no such associations were observed. The 11 studies that referred to measures of leisure-time or total physical activity provided a similar pattern: eight reported an inverse association between level of physical activity and risk of colon cancer, while in three studies this was not the case.

In these studies, the level of physical activity during early adulthood was used as an effector variable. It is possible that this life period is too early to find an effect on colon cancer incidence in late adulthood. In most studies that controlled for potentially confounding dietary factors, the relationship between physical activity and risk of colon cancer remained significant. Taken together, the findings strongly suggest a significant association between level of physical activity and the occurrence of colon cancer.

1.3.2 Breast cancer. The report of the Surgeon General of 1996 reviews the results of 13 studies that examined the effects of habitual physical activity on risk of breast cancer (DHHS 1996). Of the 13 studies, four referred only to occupational physical activity; the remaining nine studies used a leisure-time or total physical activity estimate. Of the four studies that focused on occupational physical activity, two found a significant inverse relationship between physical activity and breast cancer, but the other two did not. The nine studies considering leisure-time or total activity provide a contradictory picture: two studies found a significant inverse association, two reported a nonsignificant inverse association, three others could not observe any relationship, and the remaining two identified direct associations (i.e., the more active the person was, the higher the risk of cancer). Thus, there is only limited evidence for a protective effect of physical activity against the risk of breast cancer.

1.3.3 Conclusion. Epidemiological studies have also examined the relationship between physical activity and other cancers, such as prostate, testicular, ovarian, and endometrial cancer. Firm conclusions regarding these diseases cannot be drawn because data are often very limited or are inconsistent. The only clear association in this area seems to be the one between level of physical activity and the likelihood of developing colon cancer. To explain this finding, it has been assumed that physical activity reduces gastrointestinal transit time and thus diminishes the contact duration between the colon mucosa and potential carcinogens in the fecal stream.

1.4 Further Chronic Diseases

Level of physical activity has been related to further chronic diseases and conditions, in particular to diabetes, arthritis, osteoporosis, and obesity. Several prospective cohort studies lend strong support to the hypothesis that regular physical activity offers protection against the risk of noninsulin-dependent diabetes mellitus. Whether or not physical activity may prevent the development of osteoarthritis is still unknown. In postmenopausal women, physical exercise may impede the rapid decline in bone mass, thus protecting against osteoporosis, but this effect has not yet been unequivocally established. Reducing obesity by increased physical activity alone is a relatively slow process. Physical activity programs in combination with strategies of dietary management seem to be more effective in this regard.

2. The Effects of Physical Activity on Mental Health

There is a large amount of literature dealing with the psychological effects of physical activity. Numerous reviews and meta-analyses have dealt with this literature (Fox 1999, Morgan 1997). In studies examining the effects of physical activity on mental health, the most frequently observed outcomes include depres-

sion, anxiety, mood, self-esteem, self-efficacy, perceived stress, and cognitive functioning. On the other hand, the role of physical activity in the development and treatment of psychiatric disorders such as schizophrenia, dementia, personality disorders, or eating disorders has only seldom been investigated. The following section focuses on the association of physical activity with depression and anxiety (see *Depression, Hopelessness, Optimism, and Health*).

2.1 Depression

Research using correlational designs suggests a substantial relationship between physical activity and reduced symptoms of depression in men and women (Morgan 1997). In summarizing the existing literature, it was concluded that physically inactive persons are twice as likely to have symptoms of depression than their physically active counterparts. The association between level of physical activity and depression has been found in the general population as well as in clinical subpopulations (Fox 1999).

Correlational studies are often based on cross-sectional data, leaving the question unanswered whether physical activity has an effect on depression, or whether persons with good mental health are simply more likely to be active. In an attempt to resolve this problem of causality, cohort studies have been conducted in which the relationship between physical activity and depressive symptoms were examined in a longitudinal perspective. One of these cohort studies was conducted by Paffenbarger et al. (1994), who revealed that level of physical exercise reported at an initial interview was predictive of physician-diagnosed depression about 25 years later. In this investigation with 10,201 participants, the relative risk of depressive symptoms was 27 percent lower for men who had reported spending more than 3 hours of sports activities each week than for men who had reported no such activities. However, in other population studies, the level of physical exercise at baseline was not a risk factor for incidence of depressive symptoms years later (Weyerer 1992). Thus, the empirical evidence from cohort studies is still limited and inconsistent.

Another approach to resolve the problem of causality is intervention studies conducted in community or laboratory settings (Morgan 1997). A classical and often-cited work in this field is the experimental study by McCann and Holmes (1984). Altogether, 43 women students who scored higher than the cut-off point on the Beck Depression Inventory were randomly allocated to one of three groups: (a) aerobic exercise group, (b) placebo group, and (c) no treatment group. After five weeks, the aerobic exercise group showed a significant reduction in depressive symptoms compared with the two other groups, supporting the hypothesis that increased physical activity resulted in a reduction in depression. A more recent experimental study by King et al. (1993) demonstrated that greater exercise participation was significantly related to fewer depressive symptoms, independent of changes in physical fitness or body weight. Other intervention studies have shown that the antidepressant effect of physical activity may be greatest in those with higher levels of initial depressive symptoms (Brown et al. 1995), suggesting that the psychological benefits of exercise are likely to be different for different people.

Several biological and psychosocial mechanisms have been proposed in order to explain the antidepressant effects of physical activity. In particular, it has been hypothesized that physical activity may induce changes in brain neuroreceptor concentrations of monoamines (norepinephrine, dopamine, or serotonin) or endogenous opiates (endorphins and enkephalins), which in turn are responsible for changes in the affective level (Morgan 1997). However, the observed depression-reducing effects of physical exercise may also be mediated by psychological factors such as exercise-induced feelings of mastery and self-efficacy, or by positive social and emotional experiences in the sport context.

2.2 Anxiety

Research has also suggested that physical activity may be linked to a reduction in anxiety. The psychological assessment of anxiety usually includes state and trait measures. State measures refer to the actual feeling of anxiety, whereas trait measures evaluate the stable personality characteristic of anxiety. In physical exercise research, state measures have been used mainly in pre-post-designs of intervention studies to explore whether a specific activity episode may have triggered changes in currently experienced anxiety. Results are relatively unequivocal in this respect. Immediately after a session of physical activity, most people report an improvement in mood, including a reduction in level of state anxiety (Berger and Owen 1998). However, this decrease in anxiety may persist for only 2–6 hours following the activity episode (Morgan 1997). In order to continuously profit from the anxiety-reducing effects of physical activity, it is obviously necessary to engage in physical activity on a daily basis. There is no clear evidence that regular physical exercise may have a general impact on trait anxiety (Schlicht 1994).

2.3 Further Psychological Effects

Physical activity has been considered to be a stress buffer in critical life situations (Fuchs and Hahn 1992). This has been explained by two different processes: physical exercise may affect the transactional stress-coping process, either by influencing an individual's appraisal of the potentially stressful situation, or acting as an emotion-focused coping strategy. However, empirical evidence supporting the stress-buffering role of physical activity is still very limited.

Another psychological construct often linked to physical exercise is self-esteem. In his review, McAuley (1994) concluded that the positive association between physical activity and self-esteem is observed both as an enduring consequence of habitual exercise training and as an immediate, short-term effect of single exercise episodes. Finally, several studies suggest that there is a direct relationship between level of physical activity and measures of cognitive performance, such as reaction time, nonverbal reasoning, and memory.

3. Conclusions

There is clear evidence that physical activity is significantly associated with reduced risk of all-cause mortality, all cardiovascular diseases combined, CHD, hypertension, colon cancer, and non-insulin dependent diabetes mellitus. Findings also strongly suggest that regular physical activity may reduce the risk of developing obesity, osteoporosis, depression, and state anxiety. However, effects on psychological factors such as cognitive functioning, self-esteem, body image, self-efficacy, control beliefs, or optimism are less clearly documented. Future research should focus on these emotional and cognitive factors that may be mediators in the relationship between physical activity and mental health. Based on data from epidemiological studies, it has been concluded that physical activity that increases the weekly energy expenditure by about 1,000 kcal (equivalent to about 90 minutes of running or 120 minutes of jogging) may be associated with substantial physical health benefits (DHHS 1996). But how much physical activity is needed to improve or maintain psychological well-being? Probably the mental health benefits of physical exercise are not only a matter of expended kilocalories (if at all), but also of exercise-related social interactions and emotional experiences. Clarifying the mechanisms by which physical activity may affect mental health dimensions is a challenge to sport and health psychology in the coming years.

See also: Depression, Hopelessness, Optimism, and Health; Health Behaviors; Physical Activity and Health Promotion; Sport and Exercise Psychology: Overview

Bibliography

Abbott R , Rodriguez B, Burchfiel C, Curb J 1994 Physical activity in older middle-aged men and reduced risk of stroke: The Honolulu Heart Program. *American Journal of Epidemiology* **139**: 881–93

Bauman A, Owen N 1999 Physical activity of adult Australians: Epidemiological evidence and potential strategies for health gain. *Journal of Science and Medicine in Sport* **2**: 30–41

Berger B, Owen D 1998 Relation of low and moderate intensity exercise with acute mood change in college joggers. *Perceptual and Motor Skills* **87**: 611–21

Blair S N, Kohl H, Barlow C, Paffenbarger R, Gibbons L, Macera C 1995 Changes in physical fitness and all-cause mortality: A prospective study of healthy and unhealthy men. *Journal of the American Medical Association* **273**: 1093–8

Brown D R, Wang Y, Ward A, Ebbeling C, Fortlage L, Puleo E et al. 1995 Chronic psychological effects of exercise and exercise plus cognitive strategies. *Medicine and Science in Sports and Exercise* **27**: 765–75

DHHS US Department of Health and Human Services 1996 *Physical Activity and Health: A Report of the Surgeon General*. US Department of Health and Human Services, Atlanta, GA

Fox K 1999 The influence of physical activity on mental well-being. *Public Health and Nutrition* **2**: 411–18

Fuchs R, Hahn A 1992 Physical exercise and anxiety as moderators of the stress–illness relationship. *Anxiety, Stress, and Coping: An International Journal* **5**: 139–49

King A C, Taylor C, Haskel W 1993 Effects of differing intensities and formats of 12 months of exercise training on psychological outcomes in older adults. *Health Psychology* **12**: 292–300

Manson J, Hu F, Rich-Edward J, Coldit G, Stampfer M, Willett W et al. 1999 A prospective study of walking as compared with vigorous exercise in the prevention of coronary heart disease in women. *New England Journal of Medicine* **341**: 650–58

McAuley E 1994 Physical activity and psychosocial outcomes. In: Bouchard C et al. (eds.) *Physical Activity, Fitness, and Health: International Proceedings and Consensus Statement*. Human Kinetics, Champaign, IL pp. 551–68

McCann L, Holmes D S 1984 Influence of aerobic exercise on depression. *Journal of Personality and Social Psychology* **46**: 1142–7

Morgan W P (ed.) 1997 *Physical Activity and Mental Health*. Taylor & Francis, Washington, DC

Morris J N, Heady J, Raffle P, Roberts C, Parks J 1953 Coronary heart disease and physical activity of work. *Lancet* **2**: 1111–20

Paffenbarger R, Hyde R, Wing A, Lee I, Jung D, Kampert J 1993 The association of changes in physical activity level and other lifestyle characteristics with mortality among men. *New England Journal of Medicine* **328**: 538–45

Paffenbarger R S, Lee I, Leung R 1994 Physical activity and personal characteristics associated with depression and suicide in American college men. *Acta Psychiatrica Scandinavica Supplementum* **377**: 16–22

Schlicht W 1994 Does physical exercise reduce anxious emotions? A meta-analysis. *Anxiety, Stress, and Coping* **6**: 275–88

Sherman S, D'Agostino R, Silbershatz H, Kanne W 1999 Comparison of past versus recent physical activity in the prevention of premature death and coronary artery disease. *American Heart Journal* **138**: 900–7

Weyerer S 1992 Physical activity and depression in the community: Evidence from the Upper Bavarian Field Study. *International Journal of Sports Medicine* **13**: 492–6

R. Fuchs

Physical Activity and Health Promotion

Physical as well as mental health are stabilized and promoted by regular physical activity. The shortness of this article forces us to examine its health-promoting effects only in brief. The main reflections will conc-

entrate on two subjects: health promotion as a science-based theory to change health damaging behavior and the phenomenon that people of Western societies lead a sedentary lifestyle despite the evident benefits of activeness.

1. Physical Activity and Health

Disease prevention by regular physical activity has become more and more important. Its significance has been shown in various studies. Epidemiological and naturalistic as well as experimental studies demonstrate the risk-reducing and health-enhancing effects of regular physical activity. In particular, aerobic training seems to be effective in preventing cardiovascular diseases. The psychological benefits of physical activities are decreasing negative emotional states (e.g., anxiety) and increasing positive emotional states (e.g., vigor).

Whereas the effect of exercise on physical health can be explained by adaptation processes of the metabolic and cardiovascular system, the effects on mental health remain unexplained. Different kinds of hypotheses assume physiological (e.g., thermoregulation, endorphins), psychological (e.g., attention distraction, states of *flow*), or rather unspecific mechanisms (e.g., effort justification). Further research is needed to enlighten these mechanisms. However, it can be stated that physical exercise is an appropriate means for health promotion in most cases.

2. Physical Activity and Health Promotion: Two Complementary Strategies

Health promotion is a scientific conception to help people change their lifestyles towards a state of optimal health. Recently published theories conceive health and disease as independent dimensions. In view of such a concept, reducing cardiovascular risk by regular exercise does not lead intentionally, at the same time, to optimizal physical, mental, and social wellbeing. The strategy of health promotion therefore has been extended in the past decades. Whereas in the 1980s health promotion was defined as a disease preventing strategy, it has been supplemented by a setting-based lifestyle strategy in the 1990s. Health promotion today focuses on disease prevention and health enhancement, while also aiming at changing behavior and improving the living conditions of social groups and individuals.

2.1 Physical Activity and Health Promotion: The Behavior-oriented Strategy

When focusing on personal behavior change without considering the environmental conditions, health promotion seems to have rather limited success. About 50 percent of the population in the USA do not adopt regular exercise and 25 percent are not active at all (USCDCP 1994). Those who begin an exercise program often fail to adhere to it. Most of them relapse to inactivity after the first six months (Marcus et al. 2000). Previous preventive interventions have often used threatening contents to convince people to give up inactivity (e.g., threat of coronary artery disease).

When reviewing the literature on fear communication Leventhal (1970) described how threatening intervention practices are often provoking a negative relationship between fear and persuasion. So there is no linear relationship between threat and motivation to quit health-damaging behavior. One cause of this paradoxical effect is well documented in social psychology literature as the phenomenon of reactance (a reaction to feel a restricted liberty of behavior).

In the last decade of the twentieth century Health Psychology applies theoretical models (see *Health Behavior: Psychosocial Theories*) which define the cognitive determinants of health behavior (continuum theories) and construe behavior change as a stepwise process (stage theories). Current continuum theories are the Health–Belief–Model, the Protection–Motivation–Theory, the Theory–of–Planned–Behavior and the Subjective–Expected–Utility–Theory (see *Health Behavior: Psychosocial Theories*). Empirical tests verify at least four variables as predictive for health behavior in general (see *Health Behaviors*). A behavioral change is probable on condition that a person perceives health hazards as being likely (vulnerability) and severe (severity), and on the condition that this person is supplied with disease preventing alternatives (outcome expectancies), and has enough personal efficacy (self-efficacy) to change a risky lifestyle.

With regard to physical exercise, perceived behavior control, positive outcome expectancies, social support and severity, as well as vulnerability, define valid and reliable sociocognitive predictors determining adoption to exercise for men and women in different ways (Sallis et al. 1992). In contrast, knowledge about intervention strategies that try to foster exercise adherence or maintenance to physical activity, is at an early stage (Marcus et al. 2000).

Valid social and demographic predictors of health-related activities are mainly age, gender, neighborhood environment, school attainment, ethnicity, and income aside from sociocognitive variables. This means that women, elderly people, people with a lower education, low income, ethnic minorities, and persons living in an area with low recreational potential, are little active in sports.

Moreover health-related behavior (including health damaging activities) has proved to be quite stable (Sutton 1994). It is obvious that inactive people are not aware of the risks of leading a sedentary lifestyle. Continuum theories are scarcely suited to explain

changes from the sedentary-oriented stage to that of readiness for physical activity.

Stage theories (Transtheoretical–Model, Precaution–Adoption–Process) seem rather adequate to deal with this phenomenon (Weinstein et al. 1998). They postulate distinguishable stages in which varying processes pass, leading to different levels of change: for example, precontemplation, contemplation, active change, maintenance, and perhaps relapse. Adoption of exercise has been investigated by various authors, most of them using a cross-sectional comparison of people believed to be in different stages (e.g., Prochaska 1994). The results indicate an increase in the perceived benefits of exercise from a pre- to a postdecisional stage. More experimental and longitudinal research is needed to clarify the conditions under which people change from a sedentary to an active lifestyle.

2.2 *Physical Activity and Health Promotion: The New-Public-Health Approach*

Many health problems and severe diseases in modern societies are surely caused by overeating, alcohol consumption, drug abuse, and a sedentary lifestyle. Therefore, the question of how to change health-damaging lifestyles still remains important. The models and theories mentioned above help to develop efficient intervention strategies.

Physical activity is part of a health-enhancing lifestyle. The achievement of physical activity goes beyond reducing the risk of cardiovascular diseases. Contributing to social integration gives the feeling of internal control and—especially with young people—provides the conditions for learning adequate, rule-based social behavior (see Danish et al. 1996). Exercising in groups seems to promote the learning of suitable skills to mobilize social networks.

British studies especially, however (see Blaxter 1990), substantiate that particularly middle- and upper-class members of modern societies cling to a health-enhancing lifestyle. Body image and its relevance to physical activity is now a subject of growing importance for those with a modern, middle-class lifestyle. As can be suspected, inequalities in exercise engagement are not only caused by personal income, though this is a premise for taking part in physical activities such as aerobics, calisthenics, and power training in fitness studios. Causes may be found more likely in the 'social environment' of persons. Social environment construes informal as well as formal rules and norms and it provokes attitudes and values, feelings, thoughts, and actions, which then determine the actual position of a person in a given setting (e.g., workplace, school, family). From this point-of-view a certain health-damaging behavior (e.g., inactivity) may be functional in so far as it demonstrates identity with a defined social group or class. Empirical facts show that lower-class people share a mechanistic view of the body. In general they do not expect to have personal influence on health and wellbeing. The consistent epidemiological results from USCDCP (1994) that are mentioned above, demonstrate that less physical activity participation correlates with lower educational attainment, lower income, minority status, old age, and being overweight. These factors seem to be significant barriers to exercise.

Interventions merely focusing on health-damaging behavior run the risk of ignoring the constraints and barriers of behavior change. They are liable to lead towards 'healthism.' They are likely to provoke moralistic attitudes such as blaming the victim (see Diekstra and Jansen 1990). Health promotion in the sense of New Public Health is a crossdisciplinary science that avoids such misinterpretation by preferring an ecological perspective. The *Ottawa Charter for Health Promotion* of the WHO pleads for an enlargement in the paradigm of the intervention process. Health promotion in this sense considers both personal behavior and personal environment. It includes disease prevention as well as health enhancement. Thus health promotion has the chance to look for daily life resources which are health enhancing. Those resources are, among other things, sufficient income, adequate command of work, ability to cope with daily hassles and daily uplifts, social integration and social support, self-esteem, and the feeling of internal control (Diekstra and Jansen 1990). As already mentioned, exercising in groups can contribute to the strengthening of these resources; therefore all social groups should have the opportunity to profit from these resources as well.

King et al. (1997) show that participation in an exercise program depends on its demands and type. They compared two groups of women—those who had less than 12 years of education with those who had more than 12 years. The Stanford-Sunnyvale Health-Improvement Project (SSHIP) found that the less-educated group, which was given an adequate home-based exercise program, had relatively high rates of long-term exercise adherence (two years). So, if matched to the right program, subgroups of lower classes have the chance to do quite well. Studies like the SSHIP will help to identify those variables, which can provide information on when and why people become more or less active over the course of time.

3. *Conclusions*

Physical activity is an appropriate means of enhancing physical as well as psychological wellbeing. Despite the growing evidence of the health benefits of regular physical activity, about 50 percent of adult people remain sedentary, especially those in lower socio-economic status groups. Research in Health Psychology as well as in New Public Health is challenged to clarify those conditions and program parameters which are appropriate to enhance the motivation to

exercise and which bind people together in long-term regular exercise programs.

See also: Health Behavior: Psychosocial Theories; Health Behaviors; Health Behaviors, Assessment of; Physical Activity and Health; Public Health; Sport and Exercise Psychology: Overview; Sport, Anthropology of

Bibliography

Blaxter M 1990 *Health and Life Styles*. Routledge, London

Danish S J, Nellen V C, Owens S 1996 Teaching life skills through sport: Community based programs for adolescents. In: van Raalte J L, Brewer B W (eds.) *Exploring Sport and Exercise Psychology*. APA, Washington, DC

Diekstra R, Jansen M A 1990 The importance of psychosocial interventions in primary health care. In: Schmidt L R, Schwenkmezger P, Maes S (eds.) *Theoretical and Applied Aspects of Health Psychology*. Harwood Academic Publishers, Chur, Switzerland, pp. 87–101

King A C, Kiernan M, Oman R F, Kraemer H C, Hul M, Ahn D 1997 Can we identify who will adhere to long-term physical activity? Signal detection methodology as a potential aid to clinical decision making. *Health Psychology* **16**: 380–9

Leventhal H 1970 Findings in theory in the study of fear communications. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, pp. 119–86

Marcus B H, Dubbert P M, Forsyth L A, McKenzie T L, Stone E J, Dunn A L, Blair S N 2000 Physical activity behavior change: Issues in adoption and maintenance. *Health Psycholgy* **19**: 32–41

Prochaska J O 1994 Strong and weak principles for progressing from precontemplation to action on the basis of twelve problem behaviors. *Health Psychology* **13**: 47–51

Sallis J F, Hovell M F, Hofstetter C R 1992 Predictors of adoption and maintenance of vigorous physical activity in men and women. *Preventive Medicine* **21**: 237–51

Sutton S 1994 The past predicts the future: Interpreting behaviour-relationships in social-psychological models of health behaviours. In: Rutter D R, Quine L (eds.) *Social Psychology and Health: European Perspectives*. Avebury, Aldershot, UK, pp. 71–88

USCDCP 1994 Prevalence of sedentary lifestyle—behavioral risk factor for chronic disease by education level in racial/ethnic population—United States, 1991–1992. *Morbidity and Mortality Weekly Report* **43**: 894–9

Weinstein N D, Rothman A J, Sutton S R 1998 Stage theories of health behavior: Conceptual and methodological issues. *Health Psychology* **17**: 290–9

W. Schlicht

Physical Sciences: History and Sociology

The twentieth century was the century of the physical sciences. Immensely successful at making sense of the material world, they are also deeply embedded in the social fabric, and have become associated with technological development, economic well-being, and national security and prestige. The physical sciences have also provided the knowledge required to produce a vast array of weapons and weapons systems. Their history and sociology are contested terrains, in which the issues at stake are commensurate with the ambiguity of their social role.

1. Object

The physical sciences as a coherent body of practices emerged in Western Europe in the seventeenth century (Dijksterhuis 1986/1950, Shapin 1996). At the time they encompassed the fields of physics, astronomy, and mechanics. They were at the core of a radical epistemological and social project which defined nature in mechanistic terms and as subject to law-like and predictable behavior. Their aim has been to understand but also to exploit and to manipulate their objects. Their means has been a combination of 'reason' and 'experience,' both of which had to respect locally defined and shared canons of good practice. Their practitioners mobilized mathematical tools and capitalized on the rigor of mathematical demonstrations to consolidate the logical force of their truth claims. They extended the reach of the human senses by building and using increasingly complex and powerful instruments which allowed for the controlled exploration of the properties and processes of matter.

2. Scope

The scope of the physical sciences has constantly evolved to include fields as diverse as chemistry, electricity, optics, and the study of solids. Beginning in the late nineteenth century, their potential for not simply understanding but also manipulating the material world led to their institutionalization, and the formalization of their links with the state, philanthropical foundations, and large corporations. Patrons and philosophers alike were impressed by their intellectual and pragmatic success, as revealed by the panoply of technologies which they spawned and on which millions of people depended for their well-being. Characterized as a set of techniques and practices, they were seen as a model to be emulated and reproduced. Traditional forms of knowledge were appropriated and subjected to their methods. Parts of the life sciences and the human and social sciences were colonized. Existing disciplinary boundaries were dissolved and new specialisms and fields emerged in the spaces thus created (Krige and Pestre 1997).

Consider the relationship between mathematics and the physical sciences. For Galileo and those who shared his intellectual project in the seventeenth and eighteenth centuries, the 'book of nature' was written in mathematical symbols and the laws governing its behavior could be expressed in mathematical form. This local unity between the physical sciences and mathematics persisted in the majestic constructions of

a Newton or a Laplace (Koyré 1985). Later, however, some mathematicians began to detach their objects from the empirical and the experimental (Bottazini and Dahan 2001). Particularly in Germany they reverted to a 'pure' mathematics in opposition to, and to defend themselves against, the rising intellectual and institutional importance of a mathematically informed theoretical physics which was based on, and guided, experimental work. It was this theoretical physics which transformed our conception of the structure and properties of matter in the late nineteenth and early twentieth centuries, combining theoretical analysis with philosophical reflection.

The exigencies of World War II swept aside these cultural norms and intellectual methods. A new kind of physics emerged, motivated by a pragmatic demand to get numbers out and to solve problems efficiently using an evolving toolkit of mathematical techniques. This problem-oriented rather than discipline-based approach persisted in the postwar era. It was entrenched through summer schools and conferences, and by interdisciplinary teamwork in universities and corporations, frequently undertaken with the needs of the military and the Cold War in mind. Pure mathematics continued with a life of its own, dedicated to abstraction, and often pursued with idiosyncratic brilliance in schools like the Bourbaki group in France. Its members saw themselves as the priests of high culture and disdained the vulgar materialism of those 'mathematicians' whose work was enmeshed with the empirical and the experimental.

Proponents of the physical sciences, actively encouraged by patrons, have repeatedly sought to colonize neighboring fields, redefining their objects so as to impose their methods and practices on them. The molecularization of biology is a case in point (Abir-Am 1982, Kay 1996). In the 1930s, officers of the Rockefeller Foundation, believing that the study of life needed to be purged of what they saw as its metaphysical content, actively promoted the migration of concepts and of instruments and techniques developed in the physical sciences, along with their practitioners, into biological research. The study of macromolecules of biological origin in the 1930s, and the unraveling of the molecular structure of DNA in the 1950s consolidated this reductionist project and contributed to the emergence of a new discipline, molecular biology. It was articulated around the concept of information, made use of instruments like x-ray crystallographs and electron microscopes, and benefited from an interdisciplinary symbiosis between physicists and researchers in the life sciences.

3. *Practices*

The physical sciences have always been enmeshed in complex matrices with other sciences, technology, and various social actors. This process was accelerated and consolidated in World War II. The development of radar, nuclear bombs, and long-range rockets mobilized physicists at the core of interdisciplinary teams which brought together, in addition, material scientists, chemists, mathematicians, and engineers in vast techno-scientific systems. These systems were shaped by the military and demanded unquestioning political and financial support from the state and the resources of major corporations. They also required the collaboration of hundreds and thousands of 'non-scientific' men and women who were mobilized for the war effort, either voluntarily (as in the case of women who worked around the clock in radar stations equipped to detect enemy submarines off the South African coast) or under conditions of appalling cruelty (as in the forced labor camps which built V2 rockets for the Nazi regime in the 1940s).

World War II institutionalized the hierarchical and planned organization of research in large interdisciplinary teams centered around a project or a problem (Galison 1997, Galison and Hevly 1992). The practice of some of the physical sciences after the war, notably in high-energy physics, reproduced and exemplified many of these features. It has come to be taken as the paradigmatic 'big science.' Particle physics brings together hundreds of researchers from many different institutions around major accelerators and the detectors used to explore the collision processes produced in them. It is patronized by the state as an icon of prestige and a training ground for some of the most brilliant members of a strategic elite. It relies on advanced sectors of industry for the development and supply of sophisticated technologies (in areas like high vacuum, electronics, and superconductivity).

This change of scale has demanded that physicists become adept at a variety of skills and bow to constraints imposed by various regimes. The size of the equipment built and used and the numbers who work with it, the danger of explosion with the cryogenic detectors (bubble chambers) deployed above all in the late 1950s and 1960s, and the competing demands for opportunities to perform experiments have all transformed the material culture of physics. For leaders in the field, being a physicist does not simply mean pursuing one's specialty in a multidisciplinary team. It means being adept at fundraising, at project management, and at dealing with industry. It also requires that one accepts the bureaucratization of experimental work. Committees set research priorities and allocate beam time and floor space. The laboratory directorate imposes policies intended to protect the interests of the research community which depends on the facility it manages, and of the government or organization which funds it.

4. *Instruments*

Instruments lie at the core of the physical sciences. They are the essential means to the production and certification of knowledge. They are also complex and

capricious (Collins 1992, Latour 1987, Shapin and Shaffer 1985). One has to be trained to use them, learning how to control their irregularities and to make sense of their output. The results obtained, and the truth-claims made, are impregnated with the vagaries of the instruments and the tacit, hands-on knowledge and experience of those who used them. Replication, preferably by others elsewhere with an instrument which is 'similar in the relevant respects,' is the ideal and idealized strategy used to transform these local truth-claims into universal truths which tell us something about the structure of the world. Such similarity is contestable, and often contested. It also presupposes not only agreed protocols of experimental manipulation but also a level of technological development congenial to the construction of standardized, precision instruments. This level was not reached in Europe until the nineteenth century. Instrument-makers, be they gifted craftsmen with tiny workshops in the streets of London or Paris a century or two ago, specialized firms developing and marketing standardized laboratory tools, or giant industries supplying heavy equipment, thus provide the essential technical and technological infrastructure demanded by the physical sciences.

The production of standardized instruments requires the integration of theories of how the instrument works into its design. Simple theories of optics were embedded in the telescopes, microscopes, and binoculars produced by the vast eyeglass industry built up by Zeiss at the end of the nineteenth century. Scientists themselves have played an active role in developing theoretically based instruments dedicated to their research agendas. Marie Curie not only amassed radium in order to pursue her research, she also promoted and financially backed the construction of measuring instruments which she could use to establish and impose standards for radioactivity (Boudia and Roqué 1997). Ernest Lawrence built his laboratory and his reputation on the construction of ever-more powerful particle accelerators based on a theoretical understanding of the behavior of charged particles in electromagnetic fields (Heilbron and Seidel 1989). Standardized, these instruments were promoted as tools to cure cancer, were used to manufacture enriched uranium for the Manhattan project during World War II, and were advocated as particle beam weapons during the Cold War. Felix Bloch's recondite research into the magnetic properties of the nucleus, for which he shared a Nobel prize for physics in 1952, suggested a new and sensitive technique for identifying the components of compound mixtures. Working along with his graduate students at Stanford University and the local firm Varian Associates, which specialized in resonance techniques, his nuclear magnetic resonance instrument became a standard tool for chemical analysis in the 1960s.

Instruments can inspire awe, and nowhere more than in the physical sciences where their complexity and size have often been exploited to legitimate the practices in which they are embedded and to increase the social weight of those who build them and demonstrate their powers. The sixteenth-century clock in Strasbourg cathedral was a masterpiece of craftsmanship. It was also a material illustration that the mechanical view of nature, congenial to and promoted by the partisans of the 'new philosophy,' was a feasible alternative to the dominant conception which took the natural world to be throbbing with life, intelligence, and purpose. The spectacular discharges sparked by Ruhmkorff coils in public arenas in the late nineteenth century thrilled a populace fascinated by an electricity which was replete with associations to medicine, spiritualism, and technological application. Giant particle accelerators and detectors (and telescopes), regularly put on display to impress the general public, promote and seek to justify the enormous cost of these recondite fields of research. Their size and power are limited only by technological feasibility, financial constraints and, today, political will. Each successive generation of instruments is reputedly bigger and better than its predecessors, as measured in easily understood, readily quantified units of energy (or resolving power). Each demonstrates scientific and technological leadership, inflates national pride, and fuels international rivalry.

5. *Patronage*

Patronage has made possible the emergence, consolidation, and then vast expansion of the sphere of influence of the physical sciences, notably that which has occurred over the past 150 years. Patronage by kings, emperors, the military, national governments, and private foundations has created the social niches and institutions and supplied the financial resources the practitioners of the physical sciences have needed to promote their projects. In return they have held out the promise that the knowledge they produced would improve not only understanding, but also, through that, the prestige and power of those who supported them.

Accommodation and compromise has usually characterized the relationship between practitioner and patron (Mukerji 1990, Price 1965). Both partners have fashioned identities and ideologies, and negotiated boundaries of power and autonomy, compatible with their mutual dependence on each other. Exceptionally that dependence has been denied, or subjugated to other dominant ideologies, as in the grotesque persecution of Jewish scientists by the Nazi regime in Germany in the 1930s. More usually pragmatism prevails, as in the unseemly race between the allies to secure the services of the scientists and engineers who had developed missiles used by that selfsame regime to bombard major European cities. The pragmatism of

patrons is matched by that of those whom they support; practitioners of the physical sciences have shown themselves remarkably adept at establishing a *modus vivendi* with whatever political regime provides them with the institutional and financial resources they need to pursue their work. The 'universalism' and internationalism of science so actively promoted in the twentieth century has gone along with the depoliticization of the community and the propagation of an ideology which has decoupled scientists as producers from their patrons as users of the knowledge which they have, in fact, co-produced.

The increasingly costly and complex equipment demanded by the physical sciences has led governments, above all in Western Europe, to pool resources and to fund collectively international laboratories and organizations (Krige and Guzzetti 1997). These collaborative initiatives have generally been restricted to basic research in domains of little direct strategic importance (like high-energy physics, nuclear fusion and space science, and also molecular biology). Even here the loss of sovereignty involved can be a disincentive to the participating governments. They drain resources from national programs and shift the locus of policy to an arena in which practitioners and patrons are locked into scientific, technical, financial, and political decision-making processes over which they have limited control. Indeed, the stability of intergovernmental schemes in the physical sciences depends on practitioners successfully persuading national authorities of the scientific and technical need for the laboratory and its equipment and on patrons defining the collaborative venture as an instrument of foreign policy. This has been particularly valuable for some domains of the physical sciences in the aftermath of the Cold War. With their field no longer perceived as an icon of national prestige and global power, nuclear physicists in the United States and even more dramatically in the ex-Soviet Union have seen support for their research falter, and have had to put their skills at the disposal of new and sometimes exotic patrons (like Wall Street brokers and investment banks). Their colleagues in Western Europe, by contrast, have continued to benefit from governments' use of their science as a tool to extend the 'European family' of nations to include many from the former Soviet bloc.

6. *Historiography*

Currently there are, broadly speaking, two ways of writing the history of the physical sciences (Golinski 1998). One approach focuses on product. It takes the output of research as an unproblematic corpus of established truths about the world. It is progressivist, and draws attention to scientific achievements and the logical, unidirectional connections between them. It is coupled with a sociology which defines scientific knowledge as comprising verified truths which are independent of social determination, and a corresponding view of the scientist as the member of a self-regulating community searching for objective knowledge.

The other approach focuses on process and practice. Inspired by sociological and anthropological arguments which insist that truth-claims are underdetermined by empirical evidence, it problematizes the products of research. It explores the social practices whereby consensus is forged among scientists, the processes whereby knowledge is legitimated in different times and places, and how it is reconfigured as it circulates among various groups of social actors. If for the former approach truth is simply a relationship between statements and the world, for the latter it is also a social accomplishment.

Each of these approaches emerged in a specific sociohistorical context and was shaped by a corresponding concept of science and the social role of its history. The focus on product was intended to celebrate science and scientists as the bearers of Reason and Truth, progressive forces opposing 'irrational' political ideologies which sought to impede their development and impose constraints on their field of action. The focus on process is a *fin de siècle* reaction to the success of that project. It seeks to puncture the pretensions of a science and of scientists whose cognitive authority and social power is immense, at least as far the physical sciences are concerned in some advanced industrialized countries. By insisting that social processes are central to the production of knowledge, it seeks to tie truth claims to the local contexts in which they emerge and are contested, and to draw attention to the uncertainty, ambiguity, and messiness which pervades scientific practice at the research frontier.

Each of these approaches has its associated political agendas and neither can lay claim to universal applicability. The former serves to defend the scientific approach wherever the appeal to empirical evidence and rational argument is threatened by institutionalized intolerance, and where science and technology have not yet been bent to alleviate poverty and suffering. The latter seeks to demythologize science and scientists wherever they have become entrenched and empowered in social structures, questioning the uncritical acceptance of science as a force for progress and of scientists as experts whose truth-claims and whose power lie beyond democratic debate and control.

See also: Biomedical Sciences and Technology: History and Sociology; History of Science; History of Science: Constructivist Perspectives; History of Technology; Human Sciences: History and Sociology; Scientific Disciplines, History of; Scientific Instrumentation, History and Sociology of; Scientific Knowledge, Sociology of

Bibliography

Abir-Am P G 1982 The discourse on physical power and biological knowledge in the 1930s: A reappraisal of the Rockefeller Foundations 'policy' in molecular biology. *Social Studies of Science* **12**: 341–82

Bottazini U, Dahan A 2001 *Changing Images of Mathematics. From the French Revolution to the New Millenium*. Harwood Academic Publishers, Chur

Boudia S, Roqué X 1997 Science, medicine and industry: the Curie and Joliot–Curie laboratories. *History and Technology* **13**(4): 241–354

Collins H 1992 *Changing Order. Replication and Induction in Scientific Practice*, 2nd edn. University of Chicago Press, Chicago

Dijksterhuis E J 1986/1950 *The Mechanization of the World Picture: Pythagoras to Newton*. Princeton University Press, Princeton, NJ

Galison P 1997 *Image and Logic. A Material Culture of Physics*. University of Chicago Press, Chicago

Galison P, Hevly B 1992 *Big Science. The Growth of Large-scale Research*. Stanford University Press, Stanford, CT

Golinski J 1998 *Making Natural Knowledge. Constructivism and the History of Science*. Cambridge University Press, Cambridge, UK

Heilbron J L, Seidel R W 1989 *Lawrence and his Laboratory. A History of the Lawrence Berkeley Laboratory*. University of California Press, Berkeley, CA

Kay L E 1996 *The Molecular Vision of Life: Caltech, the Rockefeller Foundation and the Rise of the New Biology*. Oxford University Press, London

Koyré A 1985 *From the Closed World to the Infinite Universe*. Chapman and Hall, London

Krige J, Guzzetti L 1997 *History of European Scientific and Technological Cooperation*. European Commission: Science, Research, Development, Luxembourg

Krige J, Pestre D 1997 *Science in the Twentieth Century*. Harwood Academic Press, Chur

Latour B 1987 *Science in Action. How to Follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge, MA

Mukerji C 1990 *A Fragile Power. Scientists and the State*. Princeton University Press, Princeton, NJ

Price D K 1965 *The Scientific Estate*. Harvard University Press, Cambridge, MA

Shapin S 1996 *The Scientific Revolution*. University of Chicago Press, Chicago

Shapin S, Schaffer S 1985 *Leviathian and the Air Pump: Hobbes, Boyle and the Experimental Life*. Princeton University Press, Princeton, NJ

J. Krige

Physicalism and Alternatives

Physicalism is the thesis that all there really is is whatever there has to be given the way reality is physically. Thus, if physicalism is true and there really are people, shopping malls, cities, and countries, then, given the way reality is physically, there have to be people, shopping malls, cities, and countries. The alternatives to physicalism imply that there are actual phenomena in addition to whatever phenomena there have to be given the way reality is physically. Substance dualism is an alternative that maintains that there are actual substances—for example, immaterial souls, or entelechies, or disembodied spirits—that need not exist given the way reality is physically. Property dualism, another alternative to physicalism, maintains that there are properties—for example, irreducible mental properties or irreducible moral properties—whose actual pattern of distribution need not occur given the way reality is physically. Physicalism is a controversial thesis. Even if it is true, it might have been false; one or more of the alternative views might have been true instead. The debate over whether physicalism or one of its alternatives is true is thus an empirical debate.

1. Physicalism

1.1 Formulating Physicalism

So-called type physicalism is the doctrine that every type or kind of thing is physical. Physicalists need not be committed to type physicalism. Many physicalists allow that some kinds of things are such that is logically possible for them to have instances in, for instance, a spiritual realm that is wholly devoid of matter. For example, it seems logically possible for certain kinds of social institutions and social contracts to have instances in such a spiritual realm. Physicalists insist only that in reality, that is, in the world as it actually is (i.e., in the actual world), any instance of any kind of thing, including any kind of social institution or social contract, is at some level of composition, wholly composed of physical entities (Hellman and Thompson 1975).

Given that being wholly composed of physical entities suffices for being physical, physicalism implies token physicalism for objects, the thesis that every object that is physical. Physicalism is not, however, equivalent to this thesis. One reason is that even if every object is physical, it might be an accident that this is so, for there might be laws of nature according to which, were certain physical conditions to obtain, wholly disembodied spirits would come into existence. That is, the laws of nature might be such that a certain brew of physical ingredients would bring into existence wholly disembodied spirits, even though, by chance, the physical ingredients in question have never come together into that brew (Witmer 1997). Since physicalists deny that there are such laws it is too weak to formulate physicalism as token physicalism for objects. Another reason it is too weak is that token physicalism for objects is compatible with there being properties whose actual patterns of distribution over objects fail to be determined by physical laws and the pattern of distribution of physical properties, and,

according to physicalists, the pattern of distribution of every property is so determined.

It would be too strong, however, to formulate physicalism in a way that implies that all properties are physical properties, for physicalists typically allow that there are functional properties. A functional property is a second-order property: it is the property of having some property or other that occupies a certain lawful or counterfactual role (Putnam 1967, Block 1994). Functional properties are realized by the first-order properties that occupy the roles in question. (A functional property may be multiply realizable in that there may be more than one first-order property that can occupy the relevant role.) Many functional properties are such that it is logically possible for them to be realized by nonphysical properties, which is why they themselves do not count as physical properties. Physicalists claim, however, that physical laws and physical conditions determine what functional properties there actually are, and that, in actuality, functional properties are ultimately realizable only by physical properties. And they claim further that physical laws and physical states and events determine the occurrence of every state or event that is an instance of a functional property. Since nonphysical properties and kinds may be cited in laws of the special sciences, it would also be too strong to formulate physicalism in a way that implies that all laws are physical laws. Physicalists maintain only that all fundamental laws are physical laws, that nonphysical laws are determined by physical laws and physical conditions.

Still, physicalism is first and foremost a thesis about what there actually is. It is the thesis that all there actually is is whatever there has to be given the way the actual world is in every physical respect—in respect of what physical objects there are, what physical properties they have, what physical relations they bear to other physical objects, what physical laws govern their behavior, etc. Physicalism can allow that there are actual nonphysical kinds of objects, nonphysical properties, relations, laws, etc., provided that they must exist, given the way the actual world is physically. What physicalism disallows is that the actual world contains anything that is unnecessary for it to be exactly as it is physically.

Call any possible world (i.e., any way the world might be) that is exactly like the actual world in every physical respect, a physical duplicate of the actual world. It is, of course, trivial that the actual world is a physical duplicate of itself. Call any possible world that is a physical duplicate of the actual world and that contains nothing other than what is required to be a physical duplicate of the actual world, a minimal physical duplicate of it. It is by no means trivial that the actual world is a minimal physical duplicate of itself. For this claim has, for instance, the consequence that there are actually people, shopping malls, cities, and countries only if there have to be people, shopping malls, cities, and countries in order for the actual world to be a physical duplicate of itself. Physicalism is equivalent to the controversial thesis that the actual world is a minimal physical duplicate of itself.

Call any possible world that is exactly like the actual world in every respect whatsoever, physical or otherwise, a duplicate simpliciter of it. A possible world could be a physical duplicate of the actual world without being a duplicate simpliciter of it. For there is nothing about the way the actual world is in physical respects that excludes there being, say, a realm of disembodied spirits, provided that this spiritual realm exerts no causal influence on the physical. Physicalism, however, disallows that there are such 'extra entities' in the actual world. Physicalism is equivalent to the following thesis: any minimal physical duplicate of the actual world is a duplicate simpliciter of the actual world (Jackson 1998). That is to say, all that there actually is is whatever there has to be given the way the actual world is physically if and only if any minimal physical duplicate of the actual world is a duplicate simpliciter of it.

The idea that any minimal physical duplicate of the actual world is a duplicate simpliciter of it can be vividly illustrated by a creation myth. Suppose that Mother Nature wanted to create an exact duplicate of the actual world. According to physicalism, then, all that She would have to do is create an exact physical duplicate of the actual world—a world exactly like it in every physical respect. Her work would then be done. She would have thereby created a world exactly like the actual world in every respect whatsoever.

1.2 The Term 'Physical'

What is meant by 'physical' in the phrase 'the way the actual world is in every physical respect'? The relevant notion of physicality is tied to physics. There is, however, a dilemma that any attempt to state the tie must face (Hempel 1969, Crane and Mellor 1990). The dilemma can be formulated thus: if by 'the way the actual world is in every physical respect' is meant the way current physics says the world is (in respect of what objects there are, what properties they have, etc.), then physicalism is no doubt false, for current physics is no doubt false. On the other hand, if by 'the way the actual world is in every physical respect' is meant the way the true, complete physics would say the world is, then physicalism is either trivially true or too obscure to be a subject of rational assessment.

While current physics is a vast improvement over any previous physics, it is, nonetheless, no doubt false. The first horn of the dilemma should thus be avoided. What, then, of the second horn? If the second horn were embraced, then it would be trivially true that the way the world is in every physical respect is the way the true, completed physics would say that it is, for the claim would be true by definition. It would not follow, however, that physicalism is true by definition,

or even that it is trivially true. For were 'the way the actual world is in every physical respect' to be understood in this way, then physicalism would be the thesis that all that there actually is is whatever there has to be for the actual world to be the way the true, complete physics says that it is. That thesis is by no means trivial. The true, complete physics would not, for instance, postulate a spiritual realm that is causally isolated from the realm governed by its laws. (Even if, contra physicalism, the true, complete theory of the actual world would postulate such realm, the part that counts as physics would not.) Yet it is by no means trivially true that there is no such realm.

Still physicalists do not maintain that 'the way the actual world is in every physical respect' means the way the true, complete physics would say the world is. For the true, complete physics might postulate phenomena that physicalists deny actually exist. Mechanics, the theory of motion, is a central branch of physics. A true, complete physics would thus include a true, complete mechanics. A true, complete mechanics would take into account all of the fundamental causes of motion. Suppose, then, that the actual world contains Cartesian minds, substances that are not extended in space, but that occupy intrinsic, conscious states that exert a fundamental causal influence on the motions of particles in the brains of human beings, and, thereby, a fundamental causal influence on the behavior of human beings. If so, then the true, complete physics would be a Cartesian-like physics in that it would postulate Cartesian minds. But were this the case, physicalist would regard physicalism as false. It may well seem preposterous that the true, complete physics is a Cartesian-like physics. Current physics is surely not that badly wrong. The point, however, is that it is not true by definition that the true, complete physics is not a Cartesian-like physics.

Physicalists indeed claim that the way the actual world is in every physical respect is the way the true, complete physics would say that it is. But this is not intended to be true by definition. Physicalists make the claim because they are betting that the true, complete physics will share certain features with current physics (Pettit 1993). Call any theory of physics that is an improvement over current physics and that retains the features in question (whatever they are), a successor physics to current physics. The physicalist is betting that the true, complete physics will be a successor physics to current physics. By 'the way the actual world is in every physical respect' physicalists mean the way that the true, complete successor physics to current physics says the world is. And thus physicalism is the thesis that all there actually is is whatever that has to be given the way the true, complete successor physics to current physics says that the actual world is. To steer a course between the horns of the dilemma, physicalists must identify the features that a theory of physics must share with current physics to count as a successor to it.

The place to look to discern the features in question is at the advances in physics and the special sciences of chemistry and biology that physicalists maintain lend support to physicalism. At the start of the twentieth century, not all chemical theorists accepted atomism, and even some atomists speculated that the explanation of chemical bonding might have to invoke fundamental chemical forces. By the middle of the first decade, however, Einstein's work on Brownian motion convinced the entire community of chemical theorists of the truth of atomism. Then, after two decades of work on the structure of the atom, quantum mechanics was proposed. Quantum mechanics provides a reductive explanation of chemical bonding that invokes no forces or energies other than those at work below the level of composition of the atom. For a time after the advent of quantum mechanics, there remained speculation that there may be a fundamental, emergent force at work in living organism, a vital élan; but with the advances of molecular biology, such as the discovery of DNA, this idea was abandoned. On the evidence, living things are wholly composed of subatomic entities (such as electrons, quarks, etc.), entities whose behavior is determined by forces and energies exerted below the level of the atom. Biological processes are wholly implemented by chemical processes, which, in turn, are wholly implemented by quantum mechanical ones (McLaughlin 1999a).

Physicalists are not committed to the truth of the whole of current physics, or even to the truth of quantum mechanics. They are, however, betting that the true, complete physics will share the following features with current physics: (a) every entity postulated by the theory will either be subatomic or, at some level of composition, wholly made up of subatomic entities; and (b) the theory will postulate no fundamental forces or energies that are not exerted below the level of the atom. Many physicalists hold that to count as a successor to current physics a physics must retain features (a) and (b). These physicalists allow for the discovery of new objects, properties, forces, fields, and laws, provided that these discoveries are consistent with (a) and (b). They maintain only that the true, complete physics will retain features (a) and (b).

Some physicalists, however, hold a more relaxed view of what features a theory of physics would have to have to be a successor to current physics. Some allow that a fundamentally new kind of macroscopic matter could be discovered that is not composed of subatomic entities, so long as all living things and everything with mental capacities is wholly composed of subatomic entities. Moreover, some physicalists allow for the discovery of emergent forces at work in living things and beings with mental capacities, so long as the forces in question are generated by spatio-temporal configurations of subatomic entities. It is logically possible that when atoms becomes so configured as to compose a cell or so configured as to compose a human brain in a certain of neural state, a

fundamental force comes into play that is not exerted by matter at lower levels of complexity. Such emergent forces would exert a fundamental 'downward' influence on the behavior of particles (Broad 1925, McLaughlin 1999a). They are ruled out by condition (b). But some physicalists allow that a successor physics to current physics could countenance such emergent forces since they would be generated by spatio-temporal configurations of atoms, and so not, for instance, by states of an entelechy or by ontologically fundamental acts of will.

Stronger and weaker notions of what counts as a successor physics to current physics can thus be distinguished, and with them stronger and weaker versions of physicalism. But even the weakest notion(s) of what counts as a successor physics to current physics would, for instance, exclude a Cartesian-like physics from so counting. The notions of a successor physics to current physics that are invoked to characterize versions of physicalism are vague since it is, for instance, somewhat vague what counts as physics. But they are not so obscure as to render any of the varieties of physicalism outside the realm of rational assessment. (The differences between the stronger and weaker versions of physicalism will not matter for what remains, and so will be ignored.)

Physicalism implies that there is a true, complete successor physics to current physics, but it is not equivalent to that claim. Nonphysicalists can accept that there is a true, complete successor physics to current physics, in even the strongest sense. Indeed, nonphysicalists can even allow that current physics may be true. For physicalism says more than that the actual world is the way that the true, complete successor physics to current physics says that it is. Physicalism is the thesis that all there actually is is whatever there has to be given the way that the true, complete successor physics to current physics says that the world is. It is this claim that nonphysicalists deny.

2. *Whether Physicalism or One of Its Alternatives*

2.1 *The Causal Priority of the Physical*

Physicalism is defended both by appeal to empirical evidence and by appeal to supraempirical considerations of overall coherence and theoretical simplicity.

Physicalists claim that physical determination is true: the objective probability of any physical event or state is wholly determined by prior physical events and states. Thus, the objective probability of the physical movements of any living organism or sentient creature, and any movement of any of its physical parts, will be wholly determined by prior physical events (perhaps those occurring on some cross-section of the backward light-cone of the physical movements in question). Thus, given physical determination, if there are actual nonphysical states or events, then either (a) they do not causally interact with physical occurrences at all, or (b) they causally interact with physical occurrences, but not by means of any underlying causal chain of physical occurrences, or else (c) they causally interact with physical occurrences derivatively, that is, by means of some underlying chain of physical occurrences (McLaughlin 1994).

Physicalists appeal to supraempirical considerations of overall coherence and theoretical simplicity to rule out alternatives (a) and (b). In response to (a), they maintain that there is no need to postulate states or events that have no physical effects in order to explain anything that requires explanation. In response to (b), they maintain that since physical determination is true, this possibility would involve a bizarre kind of autonomous causal overdetermination of the physical by the nonphysical that can be ruled out on grounds of theoretical simplicity, and not at the expense of overall coherence (Witmer 1997). Finally, in response to (c), some physicalists argue that only physical events causally interact with physical events. However, some physicalists allow that there are nonphysical events that causally interact with physical events in a derivative way, by means of underlying causal chains of physical occurrences. For they maintain that the nature of the underlying relationship is such that in any minimal physical duplicate of the actual world, the physical occurrences in that world will underlie exactly the same sorts of nonphysical occurrences that their physical counterparts underlie in the actual world, and, so, that (c) is compatible with physicalism.

Few philosophers now challenge physical causal determination, most accept the thesis on the evidence. To answer supraempirical considerations for physicalism, nonphysicalists typically argue that there are reasons to accept that there are certain actual phenomena not countenanced by physicalism, and that those reasons outweigh the consideration that the phenomena would be either epiphenomena or autonomous overdeterminers of certain physical occurrences.

2.2 *Elimination or Entailment?*

According to physicalists, all that there actually is is whatever there has to be given the way the actual world is physically. When it is claimed that a certain phenomenon actually exists that is not required for the actual world to be exactly as it is physically, physicalists have two options. They can argue that there is no phenomenon of the sort in question, and thus be eliminativists or irrealists about it. (All physicalists are eliminativists about disembodied spirits, for instance.) Or they can argue that, despite first appearances, the phenomenon in question must exist given the way the actual world is physically.

The debate between physicalists and nonphysicalists is mainly centered on the mental and on phenomena

logically dependent on the mental. Some physicalists are eliminativists about the mental, or at least about the mental as conceived by folk or commonsense psychology (Churchland 1978, Stich 1999). But most physicalists are realists about the mental, and many are realists about the moral also, and attempt to explain moral properties reductively, in part, by appeal to mental properties.

Physicalists argue that a phenomenon must exist given the way the actual world is physically by arguing that it is plausible that its existence is entailed by the totality of physical facts (Jackson 1998). The totality of physical facts includes all particular physical facts and all general ones, including all nomological ones, and includes as well a 'that's all' physical fact, a fact to the effect that the physical facts in question are all the physical facts.

Given physicalism, the totality of physical facts, in conjunction with the contingent thesis of physicalism, entails all of the facts. Thus, if, for example, it is a fact that a certain property has a certain worldwide pattern of distribution, then the totality of physical facts in conjunction with the thesis of physicalism will imply that the property has that worldwide pattern of distribution. The totality of physical facts will, of course, imply facts about the actual worldwide pattern of distribution of physical properties. The totality of physical facts, in conjunction with the assumption of physicalism, will imply facts about the actual worldwide pattern of distribution of functional properties and any properties metaphysically necessitated by physical properties, and will imply as well that no other sorts of properties are actually possessed by anything.

One way to establish that mental facts are entailed by the totality of physical facts and the truth of physicalism alone is by showing that the relevant mental concepts are definable in terms of physical and topic-neutral concepts (Smart 1959). (Topic-neutral concepts are neither physical nor mental concepts; they include the concept of a part, the concept of causation, and logical concepts such as the concept of conjunction, the concept of negation, and the concept of universal quantification.) Both analytical behaviorism and analytical functionalism entail that there are such definitions. According to analytical behaviorism, mental concepts can be defined in terms of concepts of dispositions to peripheral behavior (Byrne 1994). According to analytical functionalism, mental concepts can be defined more broadly in terms of macrophysical and topic-neutral concepts (Armstrong 1968, Block 1994, Lycan 1994).

Analytical behaviorism is widely regarded as a failed project. It is now generally acknowledged that no mental concept can be defined in terms of concepts of dispositions to peripheral behavior. Analytical functionalism continues to have proponents (Lewis 1994), but it is controversial whether it is true of any mental concept. Analytical functionalism is most plausible for intentional concepts such as the concepts of belief, desire, intention, and the like, since these concepts are, arguably, dispositonal concepts. But it is especially controversial whether analytical functionalism is true of sensory (or phenomenal) concepts such as the concept of the felt quality of pain (Block 1995, Chalmers 1996). Many philosophers, both physicalists and nonphysicalists, claim that sensory concepts fail to have *a priori* functional analyses of any sort. They maintain that there is an explanatory gap between sensory concepts and physical and topic neutral concepts in that sensory concepts are neither definable in terms of such concepts nor are their references fixed by associated physical and topic neutral concepts (Levine 1983).

Scientific functionalism and type materialism are physicalist alternatives to these *a priori* approaches to explaining how mental facts are entailed by the totality of physical facts and the assumption of physicalism. To focus for illustration on the case of pain, scientific functionalism maintains that pain is identical with a type of functional state (and, so, a type of second-order state), albeit the identity is knowable only *a posteriori* (Loar 1997). Type materialism maintains that pain is identical with a type of abstract neurobiological state, and that the identity is knowable only *a posteriori* (Hill 1991). As Saul Kripke (1980) has demonstrated, identity statements in which the identity sign is flanked by rigid designators (terms that have the same reference in any world in which they refer) are either necessarily true or necessarily false, even if only *a posteriori*. 'Pain' is a rigid designator; it is the name of a certain kind of sensation. Statements asserting that pain is identical with a certain type of functional state or ones asserting that pain is identical with a certain type of neurobiolgical state will be necessarily true, if true at all. Given the necessity of identity, if pain is identical with either a type of functional state or a type of macrophysical state, then facts about the pattern of realization of the functional state or macrophysical state in question will imply facts about the pattern of realization of pain. Thus if, for instance, pain = neurobiological state N, then facts about the worldwide distribution of N-states will entail facts about the worldwide distribution of pain (and conversely). According to type materialists, the entailments will be only *a posteriori*. Scientific functionalists make an analogous claim of *a posteriori* entailment.

Some philosophers argue that physicalism must be committed to the *a priori* entailment of all the facts (save indexical ones such as, e.g., I = Brian McLaughlin) from the totality of physical facts and the assumption of physicalism, on pain of being committed to an unacceptable view of necessity and/or an unacceptable view of conceptual reference (Chalmers 1996, Jackson 1998). Also, some of these philosophers argue that since facts about sensory consciousness will not be *a priori* entailed by the

totality of physical facts and the assumption of physicalism, physicalism is false. Scientific functionalists and type materialists agree that facts about sensory consciousness are not so *a priori* entailed, but they reject the claim that the entailments must be *a priori*. They argue that while sensory concepts are not *a priori* linked to physical or functional concepts, and so do not admit of *a priori* functional analyses of any sort, sensory concepts nonetheless apply solely in virtue of the physical or functional properties to which they refer. Suffice it to note that this dispute turns on unresolved issues about the nature of necessity and the nature of conceptual reference.

See also: Causation: Physical, Mental, and Social; Physical Sciences: History and Sociology

Bibliography

Armstrong D M 1968 *A Materialist Theory of Mind*. Routledge and Kegan Paul, London

Block N 1994 Functionalism (2). In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 323–32

Block N 1995 'Qualia'-based objections to functionalism. In: Lycan W G (ed.) 1995 *Mind and Cognition: an Anthology*. Blackwell, Malden, MA pp. 444–68

Broad C D 1925 *The Mind and Its Place in Nature*. Harcourt Brace, New York

Byrne A 1994 Behaviorism. In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 132–40

Chalmers D 1996 *The Conscious Mind: in Search of a Fundamental Theory*. Oxford University Press, New York

Churchland P M 1978 Eliminative materialism and the propositional attitudes. *The Journal of Philosophy* **78**: 67–90. Reprinted in Lycan W (ed.) 1995 *Mind and Cognition*. Blackwell, Oxford, UK, pp. 206–23

Crane T, Mellor D H 1990 There is no question of physicalism. *Mind* **99**: 185–200

Hempel C 1969 Reduction: ontological and linguistics facets. In: Morgenbesser, Suppes P, White M (eds.) *Philosophy, Science and Method*. St. Martin, New York, pp. 179–99

Hellman G P, Thompson F W 1975 Physicalism: ontology, determination, and reduction. *Journal of Philosophy* **72**: 551–64

Hill C S 1991 *Sensations: A Defense of Type Materialism*. Cambridge University Press, Cambridge, UK

Horgan T 1994 Physicalism (1). In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 471–9

Jackson F 1982 Epiphenomenal qualia. *Philosophical Quarterly* **32**: 127–36. Reprinted in Lycan W (ed.) 1995 *Mind and Cognition*. Blackwell, Oxford, UK, pp. 69–77

Jackson F 1998 *From Metaphysics to Ethics: a Defense of Conceptual Analysis*. Clarendon Press, Oxford, UK

Kim J 1999 Physicalism. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 645–7

Kripke S 1980 *Naming and Necessity*. Harvard University Press, Cambridge, MA

Levine J 1983 Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* **64**: 354–61

Lewis D 1994 Reduction of mind. In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 412–31

Loar B 1997 Phenomenal properties. In: Block N, Flanagan O, Guzeldere G (eds.) *The Nature of Consciousness: Philosophical Debates*. MIT Press, Cambridge, MA, pp. 597–616

Lycan W 1994 Functionalism (1). In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 317–23

Lycan W G (ed.) 1995 *Mind and Cognition: An Anthology*. Blackwell, Malden, MA

McLaughlin B P 1994 Epiphenomenalism. In: Guttenplan S (ed.) *A Companion to the Philosophy of Mind*. Blackwell Reference, Oxford, UK, pp. 277–88

McLaughlin B P 1999a Emergence. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 267–9

McLaughlin B P 1999b The philosophy of mind. In: Audi R (ed.) *The Cambridge Dictionary of Philosophy*. Cambridge University Press, Cambridge, UK, pp. 684–94

Melnyk A 1997 How to keep the 'physical' in physicalism. *Journal of Philosophy* **94**: 622–63

Papineau D 1993 *Philosophical Naturalism*. Blackwell, Oxford, UK

Pettit P 1993 A definition of physicalism. *Analysis* **53**: 213–23

Poland J S 1994 *Physicalism: The Philosophical Foundations*. Oxford University Press, Oxford, UK

Putnam H 1967 Psychological predicates. In: Capitan W H, Merrill D D (eds.) *Art, Mind, and Religion: Proceedings*. University of Pittsburgh Press, Pittsburgh, PA, pp. 37–48

Rosenthal D M (ed.) 1991 *The Nature of Mind*. Oxford University Press, Oxford, UK

Smart J J C 1959 Sensations and brain processes. *Philosophical Review* **68**: 141–6

Stich S 1999 Eliminative materialism. In: Wilson R A, Keil F C (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA, pp. 265–7

Witmer G 1997 *Demanding Physicalism*. Ph.D. dissertation, Rutgers University, New Brunswick, NJ

B. P. McLaughlin

Piaget, Jean (1896–1980)

A most productive scholar in the area of children's intellectual development, Jean Piaget was primarily interested in epistemology or theory of knowledge. His genetic psychology (psychology of the genesis of thought) was intended to provide a scientific basis for epistemology, with the main concern to explain how knowledge can become more rigorous and more objective. Piaget's psychological explanations were inspired by several scientific disciplines and beside his studies in psychology and epistemology he published in other fields such as biology, logic, and sociology. His work derived from evolutionist approaches rooted in the nineteenth century yet it is still in line with scientific trends at the end of the twentieth century.

The core of Piaget's theory is the conception of mind as an active system which, like an organism,

assimilates external data into its structures and can adjust to these data. Special emphasis, in the knowledge process, is placed on the subject's activity and the structures through which he or she perceives and understands things. Thus, the development of knowledge is viewed as the construction of new structures (rather than the accumulation of new information) triggered and guided by internal processes during the subject's encounter with reality. Each new stage in this process opens up new possibilities. In this interactionist conception, structures are neither predetermined nor found in the environment and there are striking similarities between biological evolution, cognitive development in children, and the development of scientific thought.

1. Seven Decades of Scientific Productivity

1.1 Childhood and Adolescence (1896–1918)

Jean Piaget was born in Neuchâtel, a small city of Switzerland, on August 9, 1896. He was the son of an historian and professor of medieval literature at the university. Of his mother, he wrote in his autobiography that she was an intelligent and energetic woman, whose psychological imbalance created problems in the family. At the age of 11 years, Jean Piaget started classifying and collecting mollusks with the help of the director of the natural history museum of his hometown. In his teens, he was already an acknowledged specialist in the field and started publishing papers in scientific journals (as of 1911). This activity led him to decide to study natural sciences at the university between 1915 and 1918. At the same time, as of the age of 16 he became passionately interested in philosophy, and particularly in the theory of knowledge or epistemology. He soon decided to devote his life to the elaboration of a biological theory of knowledge, whose goal would be to explain the evolution of ideas both in individuals and in the scientific community. This project was influenced by evolutionary approaches in the field of biology (Lamarck, Darwin) and in philosophical works on issues of intelligence (Spencer, Le Dantec, Bergson). In the following years, Piaget's professor of philosophy A. Reymond convinced him of the importance of a rationalist approach making use of logic and mathematics.

In 1918 Piaget published a philosophical and autobiographical novel (*Recherche*) describing the crisis he had gone through because of the conflict he felt between religion and science, or, in other words, between value and rationality. Eventually, the search for truth, which is the goal of science, became the most important value for Piaget. *Recherche* also contained the outline of a theory of the evolution of living and cognitive organizations based on the idea of improving forms of equilibrium between the whole and its parts. Piaget's twofold interest on the one hand in biology and the associated problems of the organization and evolution of living beings, and on the other hand in epistemology and related issues concerning the validity of knowledge and its development, forms the basis of all his work.

1.2 Complementary Studies (1918–1921)

Piaget was convinced that a theory of knowledge should be validated by psychological research, and decided to get training in scientific psychology as soon as he graduated from university with his Ph.D. thesis on mollusks in 1918. He then spent a term in Zurich, attending courses in psychopathology—where he encountered psychoanalysis—and in experimental psychology. He then studied in Paris from 1919 to 1921, where he attended lectures in psychology (specifically by Janet), philosophy (with Lalande), and history of science (by Brunschvicg who had great influence on Piaget's ideas). Courses in psychopathology provided him with a clinical interview method, which he soon used in his discussions with children. In effect, in parallel with his studies, he worked in the laboratory of Binet and Simon, standardizing intelligence tests for children. Piaget was fascinated by the reasoning of the children who did not yet answer 'correctly,' and he discovered that the study of children's thinking could be a means to validate a theory of knowledge. This program, which was to be the main focus of his work for 60 years, was very similar to the program designed by James Mark Baldwin (1861–1934), from whom Piaget borrowed several concepts and the expression 'genetic psychology.'

1.3 The Early Psychological Research (1921–1929)

In 1921, Claparède offered a position in his Institute for educational sciences (known as the Jean-Jacques Rousseau Institute) in Geneva to Piaget, who was to conduct most of his professional life in this institution. What he found there were stimulating intellectual atmosphere and students willing to assist him in his work. Together with them the young researcher started a series of studies of children's reasoning and soon published his first books: *The Language and Thought of the Child* (1923); *Judgment and Reasoning* (1924); *The Child's Conception of the World* (1926); *The Child's Conception of Physical Causality* (1927) (the titles given here and in the remainder of this text are the English translations of the French originals; the years are those of the original publications). The striking facts he discovered about the use of language, the non-differentiation of concepts and the explanations of physical phenomena in children, as well as the detailed

analyses and the synthetic categorization of these behaviors, brought Piaget almost immediate international fame. These studies showed that young children's thinking is structurally and functionally different from that of older children and adults. Young children lack a true logic of classes and relations, they have difficulties viewing reality from an objective viewpoint, and they lack moral norms, which command compliance on the basis of a feeling of necessity. Children's thinking tends to be egocentric, that is, they have difficulties in looking at things in perspective. In these early works, the passage from egocentric thought to a logical and more objective form of thinking was explained by a social factor: cooperation, or exchange between individuals who consider themselves as equals.

During this period, Piaget was involved in many activities beside his research. He taught child psychology at the J.-J. Rousseau Institute, philosophy and history of science at the University of Neuchâtel (1925–1929), was for a few months psychoanalyzed by Sabina Spielrein, a pupil of Freud, and continued to study the evolution of mollusks. In 1923 he married a student at the Geneva Institute. Piaget and his wife Valentine spent much time minutely observing their three children who were born between 1925 and 1931. These observations were the basis of several books published in the following years.

1.4 Administrative Responsibilities, Professorship and New Directions of Research (the 1930s)

In 1929, Piaget was appointed professor of history of science at the University of Geneva, where he subsequently became professor of sociology (1939) and of experimental psychology (after the death of Claparède, 1940). He also played an important administrative role at the J.-J. Rousseau Institute as associate director (1929), then as co-director (since 1932), reorganizing the Institute and establishing links between the Institute and the University. At the same time, and until the 1960s, he was director of the *Bureau International de l'éducation* and managed to convince the governments of many countries to join that organization.

Piaget's most important activities continued to be research and writing. During these years, he accumulated data on children's reasoning. However, apart from his book on moral judgment (1932)—simultaneously in French and in English—which was the sequel to the psychology books published in the 1920s, his major publications dealt with infant cognitive development. The choice of this topic stemmed from Piaget's conviction that logic has its roots in actions and also from his deep interest in the issue of the origins of knowledge. *The Origins of Intelligence* (1936) and *The Child's Construction of Reality* (1937) contained extremely original observations and explanations of infant behaviors, and departed radically from Piaget's earlier writings as far as the explanatory model was concerned. Piaget discovered behaviors that had previously been overlooked, for example, the failure of babies younger than 9 months of age to look for an object hidden in front of them. Replications of these observations and experiments many years later around the world have confirmed these findings that had resulted from the observation of his three children.

In view of theory building, Piaget analyzed the elements of knowledge, the 'schemes' of actions, which become more differentiated and coordinated as development progresses. They evolve into increasingly complex and better-adapted forms of behavior, from automatic 'reflexes' to intelligent activities. This development is a special case of the relationship between the organism and the environment, in other words of biological adaptation. It results from the interaction between the subject of knowledge, who assimilates reality to his or her structures, and the object of knowledge, which may elicit an accommodation of these structures. Cognitive development is viewed as an evolving equilibrium between assimilation and accommodation. In *The Child's Construction of Reality* (1937), Piaget studied the growth of the practical knowledge of object permanence, space, time, and causality in infants. He introduced a new and abstract explanatory notion: the mathematical concept of 'group,' which several authors in the social sciences considered as an interesting model at the time. Piaget applied it to the organization of spatial knowledge in infants and was convinced that it could be used to explain the logic of school-age children. To this end, from 1937 onwards, he worked for three years on this model, expressing it in algebraic formulae.

Although Piaget was now a famous author who had received an honorary degree from Harvard University in 1936, his two books on infant cognition were not translated in English until the 1950s. They contained too many abstract considerations and microanalyses and were definitely ahead of their time. A third book, also based on the observation of Piaget's own children, was written during this period and published later: *Play, Dream and Imitation* (1945). It dealt with imitation in infants and the origins of symbolic thought in young children. Intellectual activity as well as play and imitation were defined in terms of equilibrium or disequilibrium between assimilation and accommodation.

1.5 The Mature Years: The Advent of Structural Analysis (1940s)

During this and the following period, Jean Piaget was prodigiously productive, thanks to a strict and very

regular work regime. In the mornings he would teach, discuss research with his large team of co-workers, and perform his administrative duties. Piaget would then spend the rest of the day mostly by himself. In the early afternoon, while thinking about the topics of his current writing, he would stroll or ride a bicycle in the countryside near Geneva. During the rest of the afternoon he would write, which was the best way for him to elaborate his ideas. Even when traveling or attending conferences he would not let a day go by without writing a few pages. Although he loved classical music he almost never went out in the evenings to movies, theaters or concerts. Both in summer and in winter, his love of nature led him to repeated sojourns in the mountains of the Valais, where he both hiked and worked.

The logical model of 'grouping' of mental operations, which he had recently developed, and which accounted for the organization of inferences in overall structures, inspired much research and writing. It permitted Piaget not only to describe the limitations of young children's reasoning as he had done in his first books, but also to explain the sudden, stable, and rather generalized intellectual progress due to the advent of the elementary logic, of 'concrete operations'. Between 1941 and 1948, his work resulted in six books: *The Child's Construction of Quantities* (with B. Inhelder, 1941), *The Child's Conception of Number* (with A. Szeminska, 1941), *The Child's Conception of Time* (1946), *The Child's Conception of Movement and Speed* (1946), *The Child's Conception of Geometry* (with B. Inhelder and A. Szeminska, 1948), *The Child's Conception of Space* (with B. Inhelder, 1948). These books study the elementary mastery of basic scientific concepts. This mastery, which usually emerges between the ages of 6 and 8 or 9 years, was explained in terms of the underlying 'grouping' structures that organize mental operations. This type of structure entails the reversibility of thinking, for Piaget an essential feature of natural logic. In contrast, younger children think that the number of a set of objects or the quantity of a substance change when spatial transformations occur, and that the duration of a movement corresponds to the distance covered, without taking into account the relative speed of the moving objects. These young children have difficulties comparing the number of elements in a subclass with the number in the superordinate class. As to space, geometrical capacities like inverting a spatial order, constructing a straight line with vertical sticks using a sighting method, and drawing the horizontal line of the water level in a tilted glass also usually emerge after the age of 7 years. The basic concept studied in each book was analyzed and defined in the light of children's answers. For example, number was analyzed in terms of a synthesis between logical classification (class inclusion) and the seriation of asymmetrical relations (ordering by increasing or decreasing size); time was defined as the coordination of movements, in which the ability to consider speed played a fundamental role, and so forth.

Piaget also published a *Treaty of Logic* (1949) and with his co-workers he udied perception in children in the laboratory of experimental psychology. These studies, which attracted Piaget's attention to the phenomenon of regulation, were to provide the subject matter of a much later book. The same holds true for the papers in sociology which were published during this decade and collected in a book called *Sociological Studies* in 1965.

The second World War rendered contact with foreign colleagues difficult, but Piaget was invited in 1942 to give a series of lectures in Paris at the '*Collège de France*' (published in *The Psychology of Intelligence*, 1946). The abstract nature of his explanations struck many listeners as much as the brilliant analyses inspired by various scientific disciplines. After the war, UNESCO entrusted Piaget with several missions.

1.6 Achieving the Ultimate Goal of his Work: The Elaboration of a Genetic Epistemology (1950s)

In the 1950s, while continuing his work in genetic psychology with the help of his co-workers, Piaget devoted much of his effort to the foundation of a genetic epistemology, i.e., the study of the basic concepts and processes of the development of scientific knowledge. This used to be a topic of philosophical study but now became an object of scientific analysis. Actually, Piaget had indeed pursued epistemological issues throughout his entire previous work while ostentatiously focusing on the explanation of children's thinking. Now he wished to discuss questions of knowledge *per se* as they were raised in the sciences (but continued to use the study of children's cognitive development for this purpose). In 1950, Piaget published three volumes of an *Introduction to Genetic Epistemology*. The book begins with a general introduction defining genetic epistemology and its method and stating the main issues. Piaget develops an idea he had already expressed in earlier writings, that of a 'circle of sciences': physical reality is explained by mathematics, mathematics is the product of mind, which is a result of biological evolution, and organisms are closely connected with physical reality. The book deals in great detail with the epistemology of mathematics and physics, and in lesser detail with the epistemology of biology and the social sciences.

To make progress in the domain of genetic epistemology, Piaget founded the International Center for Genetic Epistemology, which was a working group comprising researchers in psychology and scientists from various disciplines (mathematics, physics, logic, etc.). The Center was launched in 1955 after obtaining, with difficulty, the necessary funding. A new series of books, which contained the results of the studies

conducted at the Center for genetic epistemology, appeared from 1957 onwards. About 40 volumes were published in the series until the death of Piaget in 1980.

During the 1950s, when Piaget held a position as professor at the University of Paris (1952–1963), his psychological research in Geneva continued to progress and resulted in three books with B. Inhelder: *The Development of Elementary Logical Structures* (1959), a sequel to the studies of concrete operations; *The Origin of the Idea of Chance* (1951) and *The Growth of Logical Thinking from Childhood to Adolescence* (1955). The latter book provided a new description and definition of the highest stage of logical development: the formal operational stage achieved between the ages of 12 and 15 years. The formal model of the INRC group, derived from Klein's mathematical group, provided an account of the new possibilities of reasoning which were available to adolescents. It involved a greater number of combinations of operations than the grouping structures (in particular a 'double reversibility' of thought). Moreover, the experiments by Inhelder showed how the understanding of physical phenomena developed in children and how, through the use of an adequate interview procedure, adolescents could rediscover fundamental scientific laws on their own.

1.7 The Last Two Decades (1960 To 1980): The Pinnacle of Fame

Only in the late 1950s was Piaget's work rediscovered by English and American psychologists, although he already had been a most respected professor and researcher in the French-speaking countries, and a well-known author in many other countries. During the following two decades he was among the most cited psychologists and best known scientists in the world. Most of his books were translated from French into foreign languages (a total of 24 different languages). From the end of the 1950s onwards, he received an increasing number of honorary degrees each year. When he died, he had received over 30 honorary degrees and several prestigious awards. However, these honors and the travels they involved did not prevent Piaget from continuing his work.

Until the beginning of the 1970s, studies of genetic epistemology mainly focused on problems related to logic, language, space, time, and causality. The latter topic—strangely neglected for several decades—proved difficult to study and remained on the program of the Center for Genetic Epistemology for several years in succession. On the whole, these books, while discussing issues specific to each science (e.g., the relative importance of time and speed in judgments about duration, or the possibility of geometrical intuition) validated the constructivist conception of Piaget's epistemology. Effective forms of knowledge are neither innate nor drawn from the observation of the environment. They derive from the interaction of active individuals with their environment. A consequence of this position (summarized in the short book *Genetic Epistemology*, 1970) is that the rate of intellectual development can be only moderately accelerated by education. The goal of education, according to the few texts devoted by Piaget to this topic (for example a long article in 1965) is to favor children's intellectual activity and their discovery of relationships and abstract concepts.

During the 1960s Piaget returned to his earliest field of interest: biology. He published a paper on the evolution of mollusks and about the reproduction of the 'sedum' plants, which he grew in his garden and observed carefully. In his book *Biology and Knowledge* (1967) he developed his ideas on biology and the parallel between biological and cognitive development. Simultaneously, he published three books on psychology. They dealt with the 'figurative' aspects of knowledge, whose function is to provide a kind of copy of things: *The Mechanisms of Perception* (1961), *The Child's Mental Imagery* (1966), and *Memory and Intelligence* (1968). These books showed that figurative aspects are progressively subordinated to the subject's operations: some optical illusions decrease with perceptual activity, and with the advent of logical thinking (concrete operations), mental images can adequately represent transformations, while memory of organized material improves. Intelligence is not captured by representation, i.e., a system of images or linguistic symbols; rather intelligence consists in operating on elements of reality or symbols and thus transforming them.

In the 1970s, after Piaget's retirement as professor at the University of Geneva (1971), he continued his work with a large team of researchers at the Center for Genetic Epistemology. Most of his publications dealt with processes of cognitive development. For over 20 years, his main aim had been to demonstrate that efficient forms of knowledge result from structures that organize mental operations, and that these structures, which become increasingly complex over time, are rooted in biological organization. The time had come to deal with the processes involved in the construction of these structures, a topic that Piaget had neglected since the 1930s, except for a few epistemological papers written in the 1950s. During the last decade of his life, between the ages of 75 and 84 years, Piaget developed concepts that had been introduced earlier but needed fuller consideration and a precise definition: equilibration, reflexive abstraction, generalization, the opening up of new possibilities, and so on. These concepts are highly interesting and very original ideas, difficult to operationalize, but with deep and general explanatory value. Unfortunately, Piaget was so eager to continue his work that he changed the focus of his research and his writing nearly every year, although these complex concepts would deserve more sustained treatment.

Reflexive abstraction is the process through which something new is abstracted from experience. What is abstracted is not a property of the external world, but derives from the organization of the subject's actions or operations. Equilibration is a self-regulation process, which is specific to living organizations. At the cognitive level, it is a reaction to an intellectual perturbation (such as a contradiction) which regulates the cognitive system and, in an attempt to compensate the perturbation, creates improved forms of knowledge. For his eightieth birthday in 1976, a kind of dissertation defense was organized: Piaget defended the theses developed in *The Equilibration of Cognitive Structures* (1975) before a 'jury' composed of famous scientists who represented different disciplines. In this last period, Piaget's work also dealt with the grasp of consciousness, contradiction, and dialectics. He explored the relevance to the explanation of cognition of the mathematical concept of morphism and of the idea of logic of meaning. The last work was conducted in collaboration with R. Garcia who was also co-author of *Psychogenesis and History of Science* (published posthumously in 1983). Piaget also dealt with problems of biology. His post-Lamarckian ideas attributed some importance to aspects of experience in the evolution of species (for example in *Behavior, a Motor of Evolution*, 1976).

1.8 Factors of Scientific Creativity

Piaget died on September 16, 1980, leaving behind a life's work of unusual scope, in terms of depth, originality, interdisciplinary relevance, and volume. The explanation of his exceptional scientific productivity is not only the conjunction of a brilliant mind with a great capacity for work, but also the stimulation of reflection by the results of constantly renewed research. It was also due to Piaget's sociability, which enabled him to easily convince others to work with him; and to a strong sense of a mission to which everything else was subordinate. According to Piaget, having a theory to contradict was a powerful stimulus. The dominant empiricist conceptions in psychology—neo-associationism—and logical positivism in philosophy played this role well into the 1960s. Since the mid-1960s, Piaget was busy developing arguments against the relatively widespread idea, prevalent since Chomsky published his work, that important aspects of knowledge are innate.

2. The Importance of Piaget's Work for the Social Sciences

Piaget's theory is exceptionally comprehensive, as it gives a detailed account of cognition from birth to adolescence, and also deals with scientific knowledge in adults. It contains an incredible wealth of experimental data covering these periods in life, especially the age range from 5 to 9 years. It deals with fundamental problems often ignored in other psychological theories, such as the relationship between biological activity and cognition, the relative role of the subject and the environment, and the connections between different aspects of cognition such as perception, reasoning, and memory. In some respects Piaget's work was ahead of its time and it could therefore easily establish connections with emerging disciplines such as cybernetics (since the mid-1940s) and, later, artificial intelligence. The concepts of self-regulated systems and of feedback specific to cybernetics were involved, less explicitly, in the Piagetian ideas of reversible structures (present in his work at least since 1937) and of the regulation of cognitive activities (frequently referred to by him in the early 1940s). The artificial intelligence programs which expressed in symbolic language the expertise involved in the mastery of a task bear some resemblance to the Piagetian concept of repeatable action schemes and recall Piaget's attempts to formalize mental structures. Moreover, the multidisciplinary approach known as cognitive science, which emerged in the 1980s, is completely in line with Piaget's approach, with its links between biology, psychology, logic, and physics.

One aspect of Piaget's work, genetic epistemology, which is focused on scientific knowledge as investigated through studies of children, has had little impact outside of Geneva. By contrast, Piaget's psychological work has definitely influenced subsequent orientations and research in the social sciences.

Since the 1920s, Piaget's findings about the reasoning of children had considerable influence on child psychology. The influence was strong in the 1960s and 1970s, in particular regarding the four successive stages of cognitive development, (sensorimotor, preoperational, concrete operational, and formal operational). In fact, Piaget's idea was not to define general stages of child development, but necessary steps in the integration of new forms of knowledge. His detailed findings concerning the emergence of new capacities, such as the search for hidden objects in babies, conservation of quantity in school-age children, and combinatory operations in adolescents remain useful landmarks of psychological development although they cannot be used as psychometric tests. During the same two decades, various experiments in education were directly inspired by aspects of the theory, namely the role of the child's activity and the importance attributed to general structuring abilities rather than specific learning. However, mainstream psychology and education, especially in English-speaking countries, often remained impervious to Piaget's influence, except for the growth of cognitive developmental psychology.

Cognitive developmental psychology became a productive subdiscipline of psychology, especially since

the beginning of the 1970s. Practically every developmental psychologist in the world at that time had at least some knowledge of Piaget's work and the very idea of studying cognitive development using a cross-sectional method was inspired by the work previously undertaken in Geneva. Piaget's writing often played the same role of a 'theory to contradict' that associationism or nativist conceptions had played for Piaget. Many authors in developmental psychology launched their careers by conducting experiments that were designed to prove Piaget wrong. Because his constructivist conceptions grant a limited role to social influence, and in particular to education, they were unacceptable to many. Much research has been performed especially in the USA to show that cognitive competencies existed in children well before the ages indicated by Piaget. In order to demonstrate this, the problems presented to the children were modified and compared to those presented by Piaget himself. The race for early competencies in the 1980s and 1990s gave rise to the idea that many cognitive capacities were innate. In the end, Piaget's conceptions were not followed, but thanks to his work a whole field of psychology developed.

Piaget's work paved the way for the cognitivist revolution in psychology, from the mid-1960s onwards. Psychologists had become interested in mental states, a taboo for many decades due to the predominant behaviorist conception. The new generation of psychologists could read in Piaget's work detailed studies of mental states that had been conducted over a period of several decades.

At the end of the twentieth century, Piaget's theory had indirect descendants in the fields of child psychology, cognitive science, and developmental psychology. Ideas of specialists in educational science are more directly inspired by Piaget's work, especially in French-speaking countries and in Latin America. Studies in moral development also derive directly from Piaget's research in this field. More generally, constructivist approaches which attribute importance to the activity of the subject, in applied psychology as well as in other disciplines, usually refer, albeit vaguely at times, to the work of the Swiss psychologist and epistemologist.

See also: Binet, Alfred (1857–1911); Cognitive Development: Child Education; Constructivism in Cognitive Psychology; Developmental Psychology; Infant and Child Development, Theories of; Janet, Pierre (1859–1947); Kant, Immanuel (1724–1804); Kohlberg, Lawrence (1927–87); Khler, Wolfgang (1887–1967); Lifespan Theories of Cognitive Development; Piaget's Theory of Child Development; Piaget's Theory of Human Development and Education; Psychology: Historical and Cultural Perspectives; Scientific Concepts: Development in Children; Vygotskij, Lev Semenovic (1896–1934)

Bibliography

Bringuier J-C 1989 *Conversations with Jean Piaget*. University of Chicago Press, Chicago

Ducret J-J 1990 *Jean Piaget: biographie et parcours intellectuel*. Delachaux Nestlé, Neuchâtel, Switzerland

Gruber H, Vonèche J (eds.) 1995 *The Essential Piaget*. Aronson, London

Inhelder B, Piaget J 1955 *De la logique de l'enfant à la logique de l'adolescent*. P.U.F., Paris. [English translation, 1972 *The Growth of Logical Thinking From Childhood to Adolescence*. Routledge, London]

Montangero J, Maurice-Naville D 1992 *Piaget ou l'intelligence en marche. Vocabulaire et aperçu chronologique*. Mardaga, Liège, Belgium [English translation, 1997. *Piaget or the Advance of Knowledge. Glossary and Chronological Overview*. Erlbaum, Mahwah, NJ]

Piaget J 1926 *La représentation du monde chez l'enfant*. Alcan, Paris [English translation 1972 *The Child's Conception of the World*. Littlefield Adams, Totowa, NJ]

Piaget J 1932 *Le jugement moral chez l'enfant*. Alcan, Paris [English translation 1977 *The Moral Judgment of the Child*. Penguin, Harmondsworth, UK]

Piaget J 1936 *La naissance de l'intelligence chez l'enfant*. Delachaux and Nestlé, Neuchâtel, Switzerland [English translation 1970 *The Origins of Intelligence*. Routledge, London]

Piaget J 1945 *La formation du symbole chez l'enfant*. Delachaux and Nestlé, Neuchâtel, Switzerland [English translation 1962 *Play, Dreams and Imitation*. Norton, New York]

Piaget J 1967 *Biologie et connaissance*. Gallimard, Paris [English translation 1971 *Biology and Knowledge*. The University of Chicago Press, Chicago]

Piaget J 1970 *L'épistémologie génétique*. P.U.F., Paris [English translation 1972 *The Principles of Genetic Epistemology*. Routledge, London]

Piaget J 1970 reedited *Psychologie et épistémologie*. Denoël, Paris [English translation 1971 *Psychology and Epistemology*. Grossman, New York]

Piaget J 1972 reedited *Problèmes de psychologie génétique*. Denoël, Paris [English translation 1973 *The Child and Reality: Problems of Genetic Psychology*. Grossman, New York]

Piaget J 1972 *Où va l'éducation*. Denoël-Gonthier, Paris [English translation 1973 *To Understand is to Invent: the Future of Education*. Grossman, New York]

Piaget J 1975 *L'équilibration des structures cognitives*. P.U.F., Paris [English translation 1985 *The Equilibration of Cognitive Structures* The University of Chicago Press, Chicago]

Piaget J 1981 *Le possible et le nécessaire* 2 Vols. P.U.F., Paris [English translation 1987 *Possibility and Necessity*. University of Minneapolis Press, Minneapolis, MN

Piaget J, Inhelder B 1941 *Le développement des quantités chez l'enfant*. Delachaux and Nestlé, Neuchâtel, Switzerland [English translation 1974 *The Child's Construction of Quantities*. Routledge, London]

Piaget J, Inhelder B 1948 *La représentation de l'espace*. P.U.F., Paris [English translation 1967 *The Child's Conception of Space*. Norton, New York]

Piaget J, Inhelder B 1968 *Mémoire et intelligence*. P.U.F., Paris [English translation 1973 *Memory and Intelligence*. Basic Books, New York]

Vidal F 1994 *Piaget before Piaget*. Harvard University Press, Cambridge, MA

J. Montangero

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Piaget's Theory of Child Development

Jean Piaget's (1896–1980) theory counts as one of the most influential and enduring theories of cognitive development, and is firmly grounded in biology and epistemology. Although the focus of this entry is on the child developmental aspect of his work, it is therefore important to review some of the key epistemological concepts that Piaget adopted. Unsolved issues as well as present and future developments of Piaget's theory are also briefly reviewed.

1. *The Epistemology of Piaget*

Piaget, who trained as a biologist, had as a primary objective to understand the development of knowledge in the human species, rather than to understand why and how children develop. His first originality was to address philosophical questions (such as the origin or the development of knowledge) by empirical means, in particular by relying on the study of child development (supposed to be, at the beginning of his career, a mere 'detour' of some years). He established the discipline of *genetic epistemology* (the term 'genetic' referring to the concept of genesis or evolution, as proposed by Baldwin), with the aim 'to study the roots of the various sorts of knowledge from their most elementary forms on and to follow their development in ulterior levels including scientific thinking,' and by grounding it in both historico-critical and psychogenetic approaches. Piaget's epistemology is constructivist, in the sense that knowledge is neither a mere reflection of the external world (realism), nor a projection of pre-existing structures of the mind (nativism), and the origins of knowledge are to be found in the practical and cognitive activity of the subject. Piaget's epistemology is also fundamentally interdisciplinary.

His primary hypothesis was that intelligence is a form of biological adaptation (e.g., Piaget 1967), that is, intellectual or cognitive behaviors are the products of an organism, in interaction with its environment, and the end point of biological evolution. This is not to say that psychological phenomena can be reduced to biological ones, as Piaget always opposed biological reductionism; intelligence is an organizing activity whose functioning transcends biological organization, by elaborating new structures. The main characteristic common to knowledge and to living organisms is their adaptive character. *Adaptation* itself, whether biological or cognitive, relies on two mechanisms, defined very early in Piaget's work, namely *assimilation* (incorporation of new information into an existing system) and *accommodation* (modification of existing schemes or structures by newly assimilated elements, so as to respond to the demands of the environment). By so doing, actions and operations become coordinated with one another; such coordinations define a new entity, which in turn constitutes a new object for thought and action. This is the process that Piaget labeled *equilibration*. Contrarily to frequent, but too simplistic interpretations, equilibration does not correspond to a stable (even if only temporary) state of equilibrium, but is the process of coordination in itself.

As a consequence of the functional continuity that Piaget postulated between biological and psychological entities, a given level of knowledge is always considered to result from a reorganization of the preceding one. Reconstruction means more than a mere addition of elements of a lower type in order to attain a higher level, but implies a total reorganization and a change in scale. One direct consequence of a constructivist perspective is that an epistemologist or a psychologist studying any level of development will ask what characterized the preceding level and will stress developmental progression towards higher levels. An essential mechanism for this reconstruction is *reflecting abstraction*, that is, a process by which the subject extracts characteristics or coordinations from actions or operations; it is called 'reflecting' both because it consists of a mental reorganization and because it implies a reflection or a transfer onto a higher level of what was established at a lower level. Reflecting abstraction is contrasted with empirical or simple abstraction, which bears mainly on observable characteristics of objects or classes of objects.

2. *The Psychology of Jean Piaget*

Piaget's psychology of intelligence is the most universally known part of his work, even though he himself did not consider it the main part (e.g., Piaget 1947, 1970, Piaget and Inhelder 1966). His theory transformed the field of developmental psychology by providing it with a new vision of the development of the child, to the extent that almost every developmental study has some connection with a question Piaget raised. Intelligence is defined as the most general form of coordination of the actions and operations that characterize the various developmental levels, and not as a mental faculty or an entity in itself; it develops through a succession of general stages, defined by overall structures ('*structures d'ensemble*'). Piaget was not interested in individual or task-specific performances, but in the generic behavior of subjects in classes of situations.

The successive stages or levels are general structures, which appear in a fixed order of succession, but are not necessarily tied to a given age, and which are such that a given structure n integrates, as a substructure, the preceding structure $n-1$. The number of stages defined in Piaget's publications varied from three to five. Most often, he referred to four stages: (i) the *sensorimotor stage* (approximately the first two years), itself consisting of six substages, or three periods, in which practical *schemes* develop, among which is the scheme of object permanence (operationalized by the

study of the search for hidden objects), one of the most studied domains in present day infancy research; (ii) the *preoperational stage* from approximately 2 to 6 years of age, also labeled representative (prerelational and intuitive, or prelogical) period, made possible by the emergence and the development of the *symbolic function*; (iii) the *concrete operational stage* (7 to 12 years) during which invariants (conservation), logical classes and relations, the concept of number, as well as different concepts of space develop, all of which are marked by reversibility; and (iv) the *formal operational* stage, during which adolescents develop hypothetico-deductive reasoning (i.e., beyond concrete objects and actions), and experimental schemata. Piaget used abstract algebra to formalize the underlying structures: a group of displacements for the sensorimotor stage, groupings for the concrete operational stage, and a group of four operations (INRC group) for the formal operational stage. He also attempted to formalize the pre-operational stage by means of the logic of functions. Other cognitive activities such as mental imagery, perception, memory, and language were always dealt with by Piaget in relation to the development of operational structures. They were termed 'figurative' because they attempt to represent reality as it appears. Figurative activities are contrasted with (and subordinated to) 'operative' activities which attempt to transform reality (e.g., Piaget and Inhelder 1966b).

Although Piaget grew less and less interested in the description of these general stages, this is probably the feature of his theory that has been most often retained and criticized. In particular, the generality of the stages was questioned by the existence of '*decalages*,' that is, temporal lags between acquisition of concepts supposed to pertain to the same level. Readers of Piaget have often overestimated the necessity of synchronism in acquisitions across situations as implied by the theory. It should be stressed that Piaget did not postulate that the generality of structures meant strict synchronism, even though some of his writings were admittedly ambiguous with respect to this issue (e.g., Chapman 1988).

3. *Factors of Development*

Piaget described four factors of development: (i) organic growth, and maturation of the nervous and endocrine systems; (ii) experience, which Piaget further decomposed into a factor of physical experience (actions exerted upon objects in order to extract their properties by simple abstraction) and a factor of logico-mathematical experience (extraction of knowledge by reflecting upon actions, whether effective or mental); (iii) social interactions and transmission; and (iv) equilibration. This last factor was undoubtedly the most important for Piaget; whereas he often paid little more than lip-service to the first three factors, he spent considerable time in demonstrating the necessity of equilibration, which he considered to account for the 'vection' or direction of development (Piaget 1975). Majoring equilibration ('*équilibration majorante*') is seen as the causal mechanism underlying the progression of cognitive development. It is a dynamic sequence of states of disequilibrium–equilibrium and embodies the organism's strive toward a more general equilibrium. Equilibration is the only mechanism, according to Piaget, that can account for the emergence of novel behavior, even though more recent work has disputed the idea that Piaget's theory can really account for the emergence of novel behaviors. Cognitive structures are put into disequilibrium by perturbations which resist the assimilatory process and lead to errors or failures of action. Re-equilibration is made possible by regulations, which entail the modification of an action as a function of feedback regarding its previous outcomes, ultimately leading to compensatory reconstructions. Equilibration is thus necessary when contradictions and cognitive conflicts arise, hence the importance given by Piaget to cognitive conflict as a source of cognitive development.

4. *Unsolved Issues in the Piagetian Theory*

Piagetian theory has been subjected to numerous criticisms (probably proportional to his production) on the part of developmental psychologists, only some of which are summarized here; most critics addressed, however, only isolated elements of the theory or failed to take into account the fact that Piaget's project was essentially epistemological. First, numerous decalages, which could not be predicted by the theory, have been observed between concepts supposed to belong to the same structural level; between-task correlations are relatively low, implying that all tasks do not measure the same construct. This led a number of researchers to claim that development is domain specific. Moreover, the concept of overall logico-mathematical structure proved very difficult to operationalize. Second, Piaget's portrait of the child appeared too universal and focused on the subject, independently of his/her cultural environment. Third, the theory placed too much emphasis on the development of logico-mathematical thinking, as contrasted with other forms of thinking. Fourth, Piagetian theory did not provide sufficient information on learning mechanisms and on the effects of training. Consistent with his epistemological project, Piaget was indeed interested in demonstrating the commonalities, in terms of general mechanisms, between learning and development rather than in focusing on the specificities of learning. Linked to this is the difficulty for the Piagetian theory to account for the appearance of novel behaviors. Finally, like many other, general psychological

theories, Piaget did not take into consideration individual differences; however, it has been shown that the variability in performance, whether intra- or inter-individual, is much larger than could at first be suspected and constitutes a fundamental phenomenon rather than mere noise.

5. *The Heritage of Piaget*

Despite the criticisms addressed to the theory, Piaget's theory leaves an enormous heritage, some of which is no longer even acknowledged, because it has been incorporated into developmental psychology, almost as truisms. First, the constructivist view of development that Piaget defended is presently widely held by cognitive developmentalists, as well as by educational and cognitive psychologists. Likewise, the concepts of assimilation and accommodation, even though probably too general, are adopted by a large number of theoreticians. The view of children as active thinkers has also been generally adopted. Second, although there was a period in the 1960s and 1970s during which a large number of developmental researchers argued for domain specificity in development and against the existence of general stages, there is a growing consensus that the characterization of human cognitive development needs something more general than a long list of specific concepts, even if less general than the cognitive structures defined by Piaget. In any case, most developmentalists see the need for structural analyses that transcend any particular cognitive task. Similarly, the concept of scheme (defined as an invariant abstracted across several actions, to which new actions or objects can be assimilated and which allows for generalizations) has been retained and is close to other concepts such as schemas or scripts proposed in cognitive psychology. Third, although the construct of equilibration has been criticized for being too fuzzy, similar conceptualizations are now being proposed, whether by dynamic systems theoreticians, or by neo-Piagetians.

Last, but not least, Piaget contributed enormously to the research methods used in developmental psychology, even though he was criticized for not using a standard, quantifiable methodology. He pioneered the use of a clinical method (later to be called critical method), in which the researcher probes for the child's understanding through repeated questioning. Piaget insisted that the child's incorrect responses reflect a complex, underlying cognitive system, and should not be merely recorded as failures. Also, Piaget's work was and still is a source of very ingenious empirical techniques, in many different fields; Piagetian measures such as conservation, class inclusion, spatial representation, perspective taking, object permanence, moral judgment, transitive inference, and many others, have each produced countless studies by researchers around the world. Most of these tasks were very productive, because they produced data that were both counterintuitive and generative of hypotheses about the way children think.

6. *Beyond Piaget's Theory*

As mentioned in the previous section, the Piagetian theory is still very active, not only in developmental psychology, but in neighboring fields too; many researchers around the world continue working with Piagetian tasks, even if they do not always acknowledge it, or even know it. Also from an empirical perspective, a flourishing new direction, which is very much in the spirit of Piaget's theory, is that of the 'theory of mind.' From a theoretical point of view, the core part of the theory has been retained while giving rise to new developments, as should be the case with any sound theory. Only some concepts were really abandoned, the best example of which is the logicization of overall structures. Without pretending to be exhaustive, two directions, at present, appear to be very close to Piaget's proposals, and promising for the future, namely the neo-Piagetian models, and the application of dynamic systems modeling to developmental psychology.

The objective of the *neo-Piagetian* models, initiated by Pascaul-Leone and exemplified by scholars such as Robbie Case, Kurt Fischer, Graeme Halford, and Robert Siegler (e.g., Case 1992, Case and Edelstein 1993, Pascual-Leone 1987) was to provide a response to some of the questions left open in Piagetian theory, while preserving most of its core postulates. These models have in common the search for isomorphisms or invariants across different classes of situations. As a consequence, they have maintained or rehabilitated (after the emphasis on domain specificity mentioned above) the notion of general stage; however, the structuralist features are quite different from those postulated by Piaget. First, the hypothesis of a unique, general, logical structure underlying the organization of all behaviors at a given moment of development was abandoned, and replaced by more local ones, thereby granting more importance to context. Second, isomorphisms across domains are considered to result from common constraints or 'strictures' rather than from a common structure: general limits in processing resources or attentional capacity impose a ceiling on cognitive performances while allowing for situational and individual variability. Third, the importance given to context has as corollary a much stronger emphasis on task analysis, that is, on an analysis of the characteristics of external situations together with an analysis of the strategies that children adopt when confronted with these situations and of the processing load that these strategies entail. The challenge now is to understand under which conditions characteristic forms of reasoning are most likely to emerge, and why they do not despite the presence of sufficient general

resources. Fourth, some neo-Piagetians also assign a more fundamental status to individual differences; consequently, they claim that there may exist different developmental paths for different types of subjects. This implies that development is no longer led by a single, dominant general mechanism such as equilibration, but that several general mechanisms may interact to produce development, leading to a postulate of multidimensionality and multidirectionality in development (Reuchlin et al. 1990; Rieben et al. 1990). Such a postulate has long been adopted by lifespan developmentalists (e.g., Baltes et al. 1998), but took longer to be envisioned by child developmentalists.

A new, promising way of theorization is provided by nonlinear dynamic systems theories. Such models can be used to detect, and then model, developmental transitions, that is, changes from one stable mode to another one. The emphasis is on self-organization, which can be linked to Piaget's concept of equilibration. Although some scholars have begun to use nonlinear dynamics to build specific models of processes of development, particularly in the field of motor development or in conservation tasks (e.g., Thelen and Smith 1994, Van Geert 1994), these are only the preliminaries, and are still to be taken as metaphors; moreover, they require heavy methodological designs. Nevertheless, such attempts, which may help uncover discontinuities, i.e. stages, under what appears to be continuous change, could help bring explicit modeling into developmental psychology.

See also: Cognitive Development in Childhood and Adolescence; Developmental Psychology; Infant and Child Development, Theories of; Infant Development: Physical and Social Cognition; Lifespan Theories of Cognitive Development; Piaget, Jean (1896–1980); Piaget's Theory of Human Development and Education; Theory of Mind

Bibliography

Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Damon W (ed.) *Handbook of Child Psychology*, 5th edn., Vol. 1, Theoretical Models of Human Development. J. Wiley, New York, pp. 1027–143

Case R 1992 Neo-Piagetian theories of intellectual development. In: Beilin H, Pufall P (eds.) *Piaget's Theory: Prospects and Possibilities*. L. Erlbaum Associates, Hillsdale, NJ, pp. 61–104

Case R, Edelstein W 1993 *The New Structuralism in Cognitive Development: Theory and Research on Individual Pathways*. Karger, Basel

Chapman M 1988 *Constructive Evolution: Origins and Development of Piaget's Thought*. Cambridge University Press, New York

Pascual-Leone J 1987 Organismic processes for neo-Piagetian theories: A dialectical causal account of cognitive development. *International Journal of Psychology* **22**: 531–70

Piaget J 1947 *La Psychologie de L'intelligence*. Colin, Paris. [1950 *The Psychology of Intelligence*. Routledge & Paul, London]

Piaget J 1967 *Biologie et Connaissance*. Gallimard, Paris. [1971 *Biology and Knowledge*. University of Chicago Press, Chicago]

Piaget J 1970 Piaget's theory. In: Mussen P H (ed.) *Carmichael's Manual of Child Psychology*, 3rd edn., Vol. 1. Wiley, New York [Reprinted in Mussen P H 1983 *Handbook of Child Psychology*, Vol. I. Wiley, New York]

Piaget J 1975 *L'équilibration des Structures Cognitives*. Presses universitaires de France, Paris. [1977 *The Development of Thought: Equilibration of Cognitive Structures*. Viking Press, New York]

Piaget J, Inhelder B 1966 *La Psychologie de L'enfant* 2nd edn. Presses universitaires de France, Paris. [1969 *The Psychology of the Child*. Basic Books, New York]

Piaget J, Inhelder B 1966b *L'image Mentale Chez L'enfant*. Presses universitaires de France, Paris. [1971 *Mental Imagery in the Child*. Routledge & Paul, London]

Reuchlin M, Lautrey J, Marendaz C, Ohlmann T (Eds.) 1990 *Cognition: l'individuel et l'universe*. Presses universitaires de France, Paris

Rieben L, de Ribaupierre A, Lautrey J 1990 Structural invariants and individual modes of processing: On the necessity of a minimally structuralist approach of development for education. *Archives de Psychologie* **58**: 29–53

Thelen E, Smith L B 1994 *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA

Van Geert P 1994 *Dynamic Systems of Development*. Harvester Wheatsheaf, New York

A. de Ribaupierre

Piaget's Theory of Human Development and Education

1. Introduction

The life and work of Jean Piaget are one: his work was his life and his life inspired his work (Vonèche 2001). His theory of genetic epistemology (or experimental developmental theory of knowledge and knowledge acquisition) resulted in an immense *œuvre* of several thousand pages. All stemmed out of the encounter of his philosophical preoccupations as an adolescent about the validity of knowledge (scientific and religious essentially) when compared with his experience as a young natural historian.

As an amateur philosopher, in his teens, he had some deep religious preoccupations and wanted to reconcile science and faith into one unique form of knowledge. Religious knowledge was, in his mind, essentially subjective and emotional, whereas scientific

knowledge was objective and strictly cognitive. The obvious solution for the young Piaget was double. On the one hand, 'science cannot prescribe its premisses to value judgment ... [they] are given in immediate awareness and are undemonstrable' (Piaget 1915, p. 80). On the other hand, thinking is the essence of human existence in a very Cartesian (*cogito ergo sum*) way and in opposition both to, Kant's *a priorism* on the one hand, and to, on the other hand, Brunschvig's idea of a radically contingent spiritual development over time in the history of human thought. Therefore, the solution of the problem of the origin and value of knowledge lies in a sort of experimental study of its evolution and metamorphoses over that special sort of time called ontogenesis. He considered this study as a sort of embryology of knowledge yelding the same results as biological embryology for the succession of living organisms and their hierarchical order. The stages of knowledge are equivalent to the biological scale of organisms.

2. *Basic Theoretical Assumptions*

Thus, knowledge is biological adaptation continued by different means. Whereas biological adaptation is made possible by specific organs performing specific functions and by specific species occupying specific ecological niches—thus essentially a mechanism of differentiation and speciation—things are quite different in the case of knowledge. In intelligent adaptation there is an ever increasing separation of form and content and cognitive processes exploit the organism as a whole and no longer specific parts of it.

What remains similar in biological and cognitive adaptation is the double process of assimilation and accommodation. Biological equilibrium is assured by metabolism, cognitive equilibrium by self-regulation. Cognitive self-regulation takes three different forms: rythms, self-regulations, and mental operations. These three structures go from the simplest to the more complex and from the most external interface between organism and milieu to the most internal.

If rhythm is self-explanatory, the two forms of self-regulations: homeostasis and homeorhesis, are not. Homeostasis is the process by which any variation in one direction from the point of equilibrium is counterbalanced by a movement in the opposite direction. These movements existed in rhythm, but were successive. Now, they are simultaneous in homeostasis and form the components of a system. They limit each other by a reciprocal process of alternances. It is simply and purely a feed-back mechanism.

Homeostasis assures the stability of a given system or organism by keeping it in a stationary state in spite of perturbations. Homeorhesis is the same for evolutionary systems: instead of keeping the system stationary, it keeps the direction of evolution going, in spite of perturbations, on its path. It is a feed-forward mechanism. What is common thus to regulations and rhythms is the central role of action. Every action in one direction is compensated by an action in the opposite direction, either successively or simultaneously.

An operation, for Piaget, is nothing more than an internalized action. For instance, an arithmetic operation is the action of reuniting two units: $2+2=4$. But an operation is a very remarkable action, because it is strictly reversible. Rhythmic actions are not reversible, they always take place in the oriented arrow of time. Regulations do not compensate each other totally; they always function by excess or default. They are not strictly additive and totally reversible. Operations are perfectly reversible: $4-2=2$ exactly. Such a perfection is mental, never physical.

The smart thing with Piaget's theory is that the same basic process covers the full range of adaptation from the simple metabolic assimilation of food to the most refined, abstract logico-mathematical operation. The process of transformation of rhythm into regulation and of regulation into operation is double: internalization, on the one hand, and reflexive abstraction, on the other.

But, if one thinks it over, internalization is a reflexive process and thus reflexive abstraction is enough. It is a reflection of a form on a higher level.

3. *Cognitive Development*

The process of knowledge is the same in the child and in the species. It goes, in both, from a state of adualism between the subject and the object to successive stages of differentiation and integration. The newborn baby does not distinguish self from world at first. Then the distinction is made by the baby only in the course of action. Piaget calls this period sensory-motor.

Its main features are the recognition that objects continue to exist when out of sight because they are part of a larger container called space, that they appear in time and present some relationship of causality, as far as reality is concerned, and that means and ends can be coordinated and differentiated, as far as intelligence is concerned. By reflexive abstraction, on its external immediate action, another level of equilibrium is reached by the growing child with the emergence of the semiotic function (language, symbolic play, imitation, representational drawing, dream). This new way of functioning makes present what is absent in reality. It paves the way for concepts, classes, relations, and number that will be attained in the next stage called concrete operational in which representational self-regulations become fully equilibrated into mental operations. In turn, concrete operations will be transformed into formal ones that are independent from their material support by means of a progressive dissociation between form and content until it is total and forms a complete logical system.

Thus the young adolescent is, by now, capable of hypothetico-deductive reasoning, of examining all the possibilities as well as the internal necessity and coherence of every single thought.

4. Piaget's Constructivism

The data collected by Piaget and the School of Geneva are world-famous. They have been criticized on many grounds, but always from a reductionist epistemological point of view, either nativism or environmentalism, whereas Piaget was constructivist. This constructivistic perspective has been constantly misunderstood. It should be clear that such a structure as an operation with its internal logical necessity does not correspond either to a physiological apparatus or to a societal given. No mother, no teacher teaches a baby object permanence. Nevertheless all toddlers attain this concept-in-action. Conservation of matter, weight or volume is not taught in any school or home and nobody remembers having discovered it one morning. Nevertheless, it is reached without critical period. Nor drill.

Constructivism does not deny reality, contingency or stability of scientific knowledge as it is occasionally assumed. This assumption stems out of the confusion of Piagetian constructivism with a fairly recent trend in sociology of knowledge called the social construction of science.

Unlike sociological constructivism, Piaget recognizes the remarkable stability of scientific knowledge over time. He never denied, for instance, that the successive physical theories that have appeared since Newton or even Galileo were so stable that they encompassed superseded theories as special cases of their own theory. Piaget specifically showed the striking case for theory inclusion that Galileo's system formed for Newton's and Newton's for Einstein's. He made the same remark about modern theories of force interactions in contemporary physics.

As far as contingency is concerned, he accepted without any doubt that the advance of molecular biology was made possible by the concommittant and congruent development of computer technology allowing for long and complex computations required by molecular biological methods. Brunschvig's influence had been strong enough in that sense on the young Piaget as to remain with the old man.

Reality represents a greater difficulty, because there are different ways of approaching this concept. One can consider reality as the information given by the external world to which the human mind has to conform willy-nilly by registering it or dying. Or one can consider reality as what resists the *modus operandi* of the mind, in which case an object is what objects to the mind's demands and not a mere thing, outside there, imposing upon the mind.

Reality, for a constructivist, is a constant negotiation between the subject and the object in the course of their interaction. But, once again, the term interaction has a precise meaning for Piaget. Basically, interaction can be understood in three different ways. The first meaning is a mere relationship between two separate elements. For instance, in the case of a heated iron bar that is longer than a cold one, there is an interaction between length and temperature. The second meaning implies circularity in the relationship. For instance, the circularity between the scarcity of public transportation in the countryside which is due to the excess of private cars which, in turn, is caused by the lack of public transportation. The third and last meaning is Piagetian: it supposes the unseparability between reality and mental 'schèmes.' This means that a scientific fact does not exist independently from a scientific theory.

This is true for naive theories too. Before object permanence, it is pointless for a child to search for a disappeared object, since, in the infant's naive theory, the object has been reabsorbed by the environment until it comes back by magic and not by means of a complex logico-mathematical structure called the practical group of displacements.

Thus, Piagetian constructivism should not be confused with the social construction of reality. It is a theory that drives a middle course between the extremes of empiricism and rationalism in philosophy, Darwinism and creationism in biology, Durkheim and Tarde in sociology, behaviorism and Gestalt in psychology. In addition, Piaget thought he had empirically tested the validity of his position in the course of his studies of cognitive developmental as we shall see now.

5. Cognitive Development and Education

Essentially what Piaget observed is that intelligence evolves from states of lesser knowledge to higher ones and that cognition goes from simple to complex. But, if knowledge were learned, it would go at random from some patches of learning to others according to chancy encounters with the environment. On the other hand, if knowledge were the mere unfolding of innate potentialities, why should it go from simple to complex? There is no evolutionary nor social need for such a contrived process. Thus knowledge must be constructed step by step from direct action to internalized mental operations.

At every level of this construction corresponds a mode of equilibrium between self and reality that forms the structure of a stage that Piaget defines in logico-mathematical terms. The very existence of stages is, for Piaget, the evidence that knowledge is neither the mere unfolding of innate potentialities as in nativism nor simple registration of the features of the world as in environmentalistic and social learning theories, since the notion of stage implies a total

reorganization on a higher plane of what has been attained earlier at a lower level. This is mental construction.

Such a position involves also a choice of explanatory models that has been often misunderstood. Instead of choosing a physiological model of explanation for cognitive development such as maturation or brain states and transformation as do most psychologists, Piaget selected abstract models such as mathematical groups or equilibration theory to account for psychobiological processes such as adaptation. By so doing he conformed his explanation to his observation that intelligence is biological adaptation pursued by new and more potent means.

This carries an educational lesson. Teaching and learning which are essential to human development should be adjusted to the child's level of understanding so that any pedagogical incentive speaks to the child's condition at any given moment but also appeals to the next higher structure by calling attention on the contradictions and lacunae of the existing developmental level. This is no spontaneism. This is no drill learning nor socialization. This is cooperation (in the strong sense of a Piagetian operation) with the growing mind from the part of educators whose noble task is to co-construct knowledge with every new baby who comes to this world.

See also: Cognitive Development: Child Education; Piaget, Jean (1896–1980); Piaget's Theory of Child Development; Vygotskij, Lev Semenovic (1896–1934); Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Boden M A 1979 *Piaget*. The Harvester Press, Brighton, UK

Chapman M 1988 *Constructive Evolution: Origins and Development of Piaget's Thought*. Cambridge University Press, New York

Furth H G 1969 *Piaget and Knowledge: Theoretical Foundations*. Prentice-Hall, Englewood Cliffs, NJ

Furth H G 1970 *Piaget for Teachers* 2nd edn. Prentice-Hall, Englewood Cliffs, NJ

Gruber H, Vonèche J (eds.) 1995 *The Essential Piaget*. Aronson, Northvale, NJ

Modgil S, Modgil C (eds.) 1983 *Jean Piaget: An Interdisciplinary Critique*. Routledge & Kegan Paul, London

Piaget J 1915 *La Mission de l'Idée*. La Concorde, Lausanne, Switzerland

Piaget J 1970 *Genetic Epistemology*. Columbia University Press, New York

Piaget J 1971 *Biology and Knowledge: An Essay on the Relations Between Organic Regulations and Cognitive Processes*. University of Chicago Press, Chicago (Original work published 1967)

Piaget J, Inhelder B 1969 *The Psychology of the Child*. Routledge & Kegan Paul, London (Original work published 1966)

Smith L (ed.) 1992 *Jean Piaget: Critical Assessments*, Vols 1–4. Routledge, London

Vonèche J 2001 Identity and narrative in Piaget's autobiographies. In: Brockmeier J, Carbaugh D (eds.) *Narrative and Identity: Studies in Autobiography, Self and Culture*. John Benjamins, Amsterdam

J. J. Vonèche

Pidgin and Creole Languages

1. *Introduction*

Most studies of pidgins and Creoles (PC) have focused on their origins, despite an undeniable increase during the 1990s in the number of works on structural features. Recently, some creolists have also addressed the question of whether, as a group, Creoles can be singled out as a structural type of languages. Space limitations make it impossible to discuss structural features in this essay, aside from the fact that there are no features that are peculiar to PCs, as explained in Sect. 2.

2. *What are Pidgins and Creoles?*

Strictly speaking, PCs are new language varieties, which developed out of contacts between colonial nonstandard varieties of a European language and several non-European languages around the Atlantic and in the Indian and Pacific Oceans during the seventeenth to nineteenth centuries. Pidgins typically emerged in trade colonies which developed around trade forts or along trade routes, such as on the coast of West Africa. They are reduced in structures and specialized in functions (typically trade), and initially they served as non-native *lingua francas* to users who preserved their native vernaculars for their day-to-day interactions. Some pidgins have expanded into regular vernaculars, especially in urban settings, and are called 'expanded pidgins.' Examples include *Bislama* and *Tok Pisin* (in Melanesia) and Nigerian and Cameroon Pidgin English. Structurally, they are as complex as Creoles. The latter vernaculars developed in settlement colonies whose primary industry consisted of sugar cane plantations or rice fields, which employed massive non-European slave labor. Examples include Cape *Verdian Criolou* (lexified by Portuguese) and *Papiamentu* in the Netherlands Antilles (apparently—Portuguese-based but influenced by Spanish); Haitian, Mauritian, and Seychellois (lexified by French); Jamaican, Guyanese, and Hawaiian Creole, as well as Gullah in the USA (all lexified by English); and

Saramaccan and *Sranan* in Surinam (lexified by English, with the former heavily influenced by Portuguese and the latter by Dutch). Note that although Melanesian pidgins are associated with sugar cane plantations, they apparently originated in trade settings and were adopted on the plantations (Keesing 1988).

The terms *Creole* and *pidgin* have also been extended to some other varieties that developed during the same period out of contacts among primarily non-European languages. Examples include Delaware Pidgin, Chinook Jargon, and Mobilian in North America; Sango, (Kikongo-)Kituba, and Lingala in Central Africa, Kinubi in Southern Sudan and in Uganda; and Hiri Motu in Papua New Guinea (Holm 1989, Smith 1995). In the original, lay people's naming practice, the term *jargon* was an alternate to pidgin. However, some creolists claim that pidgins are more stable and jargons are an earlier stage in the 'life-cycle' that putatively progresses from Jargon, to Pidgin, to Creole, to Post-Creole by progressive structural expansion, stabilization, and closer approximations of the *lexifier*—the language which contributed the largest part of a Creole's lexicon.

Chaudenson (1992, p. 21) argues that Creoles developed by *basilectalizing* away from the lexifier, i.e., acquiring a *basilect*, which is the variety the most different from the *acrolect*, the educated variety of the lexifier. Mufwene (2001) emphasizes that Creoles and pidgins developed in separate places, in which Europeans and non-Europeans interacted differently—sporadically in trade colonies (which produced pidgins) but regularly in the initial stages of settlement colonies (where Creoles developed). Moreover, the term *pidgin* was coined in 1807 (Baker and Mühlhäusler 1990), over two years after the term *Creole* was used in reference to a language variety. '*Pidgin English*,' apparently a distortion of '*business English*' developed in Canton, a geographical area where no large colonial plantation industry developed and no Creoles have been identified.

The term '*Creole*' was originally coined in Iberian colonies, apparently in the sixteenth century, in reference to nonindigenous people born in the American colonies. It was adopted in metropolitan Spanish, then in French and later in English by the early seventeenth century. By the second half of the same century, it was generalized to descendants of Africans or Europeans born in Romance colonies. Usage varied from one colony to another. The term was also used as an adjective to characterize plants, animals, and customs typical of the same colonies.

Creole may not have applied widely to language varieties until the late eighteenth century. Such usage may have been initiated by metropolitan Europeans to disfranchise particular colonial varieties of their languages. It is not clear how the term became associated only with vernaculars spoken primarily by descendants of non-Europeans. Nonetheless, several speakers of Creoles (or pidgins) actually believe they speak dialects of their lexifiers.

Among the earliest claims that Creoles developed from pidgins is the following statement in Bloomfield (1933, p. 474): 'when the jargon [i.e., pidgin] has become the only language of the subject group, it is a creolized language.' Hall (1962) reinterpreted this, associating the vernacular function of Creoles with nativization. Thus, Creoles have been defined inaccurately as 'nativized pidgins,' i.e., pidgins that have acquired native speakers and have therefore expanded both their structures and functions and have stabilized. Hall then also introduced the pidgin-Creole 'life-cycle' to which DeCamp (1971) added a 'post-Creole' stage.

The first creolist to dispute this connection was Alleyne (1971). He argued that fossilized inflectional morphology in Haitian Creole (HC) and the like is evidence that Europeans did not communicate with the Africans in foreigner or baby talk, which would have fostered pidgins on the plantations. Chaudenson (1992, p. 21) argues that plantation communities were preceded by homesteads on which mesolectal approximations of European lexifiers, rather than pidgins, were spoken by earlier slaves. Like some economic historians, the socioeconomic history literature suggests that in North American colonies Creole Blacks spoke the lexifier fluently. In ads on runaway slaves in British North American colonies, bad English is typically associated with slaves imported as adults from Africa. Diachronic evidence of Creoles suggests that the basilects developed during the peak growth of plantations (in the eighteenth century for most colonies!), when infant mortality was high, life expectancy was short, the plantation populations increased primarily by massive importation of slave labor, and the proportion of fluent speakers of the earlier colonial varieties kept decreasing (Chaudenson 1992, p. 21, Mufwene 2001).

According to the life-cycle model, as a Creole continues to coexist with its lexifier, the latter exerted pressure on it to shed some of its 'Creole features.' This developmental hypothesis can be traced back to Schuchardt's (1914) explanation of why African-American English (AAE) is structurally closer to North American English than Saramaccan is to its lexifier: coexistence with it in North America and absence of such continued contact in Suriname. DeCamp (1971) and several creolists who followed suit (e.g., Rickford 1987) resurrected the position by invoking 'decreolization' ('loss of "Creole" features') to account for speech continua in Creole communities.

It is in this context that DeCamp (1971) coined the term 'post-Creole continuum,' which must be interpreted charitably. If a variety is Creole because of the particular sociohistorical ecology of its development, rather than because of its structural peculiarities, it cannot stop being a Creole even after some of the features have changed. Besides, basilectal and mesolectal features continue to coexist in these communi-

ties, suggesting that Creole has not died yet. Lalla and D'Costa (1990) present copious data against decreolization in Caribbean English Creoles.

Closely related to this issue is the common assumption that Creoles are separate languages from their lexifiers and ex-colonial varieties thereof spoken by descendants of Europeans. Thus, the nonstandard French varieties spoken in Quebec and Louisiana, as well as on the Caribbean islands of St. Barths and St. Thomas, are considered dialects of French rather than Creoles. Likewise New World nonstandard varieties of Spanish and Portuguese are not considered Creoles, despite structural similarities which they display with Creoles of the same lexifiers. Although not admitted in linguistics, the classification seems to have been associated generally with whether the majority in the largely proletarian communities speaking the new, colonial vernaculars is of European or of non-European descent. Otherwise, all colonial varieties of European languages are restructured and contact-based; the current classification of vernaculars into Creoles and non-Creoles remains question begging. The embarrassment stems largely from the absence of a yardstick for measuring structural divergence from the lexifier, especially when it is dubious that it was the same in every contact setting.

It has also been claimed that Creoles have more or less the same structural design (Bickerton 1984). This position is as disputable as the counterclaim that they are more similar in the sociohistorical ecologies of their developments, or even the more recent claim that there are Creole prototypes from which others deviate in various ways (McWhorter 1998). The very fact of resorting to a handful of prototypes for the general Creole structural category suggests that the vast majority of them do not share the putative set of defining features, hence that the features cannot be used to single them out as a unique type of language. On the other hand, variation in the structural features of Creoles (lexified by the same language) is correlated with variation in the linguistic and sociohistorical ecologies of their developments (Mufwene 2001). The notion of 'ecology' includes, among other things, the nature of the lexifier, structural features of the substrate languages, changes in the ethnolinguistic make-up of the populations that came in contact, the kinds of interactions between speakers of the lexifier and those of other languages, and rates and modes of population growth.

To date the best known Creoles have been lexified by English and French. Those of the Atlantic and Indian Ocean are, along with Hawaiian Creole, those that have informed most theorizing on the development of Creoles. While the terms 'Creole' and 'creolization' have often been applied uncritically to various contact-induced language varieties, several distinctions, which are not clearly articulated have also been proposed, for instance, between pidgin, Creole, koiné, semi-Creole, intertwined varieties, foreign workers' varieties of European languages (e.g., *Gastarbeiter Deutsch*), and 'indigenized varieties' of European languages (e.g., Nigerian and Singaporean English). The denotations and importance of these terms deserve re-examining.

3. *The Development of Creoles*

The central question here is: how did Creoles develop? The following hypotheses are the major ones competing today: the substrate, the superstrate, and the universalist hypotheses.

Substratist positions are historically related to the 'baby talk hypothesis,' which I have traced back to nineteenth-century French creolists (e.g., Adam 1883). Putatively, the languages previously spoken by the Africans enslaved on New World and Indian Ocean plantations were the primary reason why the European lexifiers which they appropriated were restructured into Creoles. These French creolists assumed African languages to be 'primitive,' 'instinctive,' in 'natural' state, and simpler than the 'cultivated' European languages with which they came in contact. Creoles' systems were considered to be reflections of those non-European languages. The baby-talk connection is that, in order to be understood, the Europeans supposedly had to speak to the Africans like to babies, their interpretation of foreigner talk.

The revival of the substrate hypothesis (without its racist component) has been attributed to Sylvain (1936). Although she recognizes influence from French dialects, she argues that African linguistic influence, especially from the Ewe group of languages, is very significant in HC. Unfortunately, she states in the last sentence of her conclusions that this Creole is Ewe spoken with a French vocabulary. Over two decades later, Turner (1949) disputed American dialectologists' claim that there was virtually no trace of African languages in AAE and showed phonological and morphosyntactic similarities between Gullah and some West-African (especially Kwa) languages. He concluded that 'Gullah is indebted to African sources' (p. 254).

There are three main schools of the substrate hypothesis at the beginning of the twenty-first century. The first, led by Alleyne (1980) and Holm (1988) is closer to Turner's approach and is marked by what is also its main weakness: invocation of influence from diverse African languages without explaining what kinds of selection principles account for this seemingly random invocation of sources. This criticism is not *ipso facto* an invalidation of substrate of substrate influence; it is both a call for a more principled account and a reminder that the nature of such influence must be reassessed (Mufwene 2001).

The second school has been identified as the 'relexification hypothesis.' Its leading proponent Lefebvre (1998) argues that HC is a French relexifi-

cation of languages of the Ewe-Fon (or Fongbe) group. This account of the development of Creoles has been criticized for several basic shortcomings, including the following: (a) its 'comparative' approach has not taken into account several features that HC (also) shares with nonstandard varieties of French; (b) it downplays features which HC shares also with several other African languages which were represented in Haiti during the critical stages of its development; (c) it has not shown that the language appropriation strategies associated with relexification are typically used in naturalistic second language acquisition; and (d) it does not account for those cases where HC has selected structural options which are not consistent with those of Ewe-Fon. Moreover, relexificationists assume, disputably, that languages of the Ewe-Fon group are structurally identical and that no competition of influence was involved among them.

The least disputed version of the substrate hypothesis is Keesing's (1988), which shows that substrate languages may impose their structural features on a new, contact-induced variety if they are typologically homogeneous, with most of them sharing the relevant features. Thus Melanesian pidgins are like (most of) their substrates in having DUAL/PLURAL and INCLUSIVE/EXCLUSIVE distinctions and in having a transitive marker on the verb. Sankoff and Brown (1976) had shown similar influence with the bracketing of relative clauses with *ia*. However, the pidgins have not inherited all the peculiarities of Melanesian languages. For instance, they do not have their VSO major constituent order, nor do they have much of a numeral classifying system in the combination of *pela* with quantifiers. For an extensive discussion of substrate influence in Atlantic and Indian Ocean Creoles, see Muysken and Smith (1986) and Mufwene (1993).

Competing with the above genetic views has been the 'dialectologist,' or superstrate, hypothesis, according to which the primary, if not the exclusive, sources of Creoles' structural features are nonstandard varieties of their lexifiers. Speaking of AAE, Krapp (1924), for example, claimed that this variety was an archaic retention of the nonstandard speech of low-class Whites with whom the African slaves had been in contact. According to him, African substrate influence was limited to some isolated lexical items such as *goober* 'peanut,' *gumbo*, and *okra*. Since the late 1980s, Shana Poplack and her associates have shown that AAE shares many features with white nonstandard vernaculars in North America and England, thus it has not developed from an erstwhile Creole (see Poplack 1999 for a synthesis). Because some of the same features are also attested in Creoles, we come back to the question of whether most features of Creoles did not after all originate in their lexifiers.

Regarding French Creoles, the dialectologist position was first defended by Faine (1937), according to whom HC was essentially Norman French. This position was later espoused by Hall (1958), who argues that 'the 'basic' relationship of Creole is with seventeenth-century French, with heavy carry-overs or survivals of African linguistic structure (on a more superficial structural level) from the previous language(s) of the earliest speakers of Negro Pidgin French; its 'lexical' relationship is with nineteenth- and twentieth-century French' (p. 372). Chaudenson (1992, p. 21) is more accommodating to substrate influence as a factor that accounts for the more extensive structural divergence of Creoles from their lexifiers compared with their non-Creole colonial kin.

The 'universalist hypotheses,' which stood as strong contenders in the 1980s and 1990s, have forerunners in the nineteenth century. For instance, Adolfo Coelho (1880–1886) partly anticipated Bickerton's (1984) 'language bioprogram hypothesis' in stating that Creoles' owe their origin to the operation of psychological or physiological laws that are the same everywhere, and not to the influence of the former languages of the people among whom these dialects are found. Bickerton pushed things further in claiming that children made Creoles by fixing the parameters of these new language varieties in the their unmarked, or default, settings as specified in Universal Grammar. To account for cross-Creole structural differences, Bickerton (1984, pp. 176–7) invokes a 'Pidginization Index' (PI) that includes the following factors: the proportion of the native to non-native speakers during the initial stages of colonization, the duration of the early stage, the rate of increase of the slave population after that initial stage, the kind of social contacts between the native speakers of the lexifier and the learners, and whether or not the contact between the two groups continued after the formation of the new language variety. Some nagging questions with Bickerton's position include the following: Is his intuitively sound PI consistent with his creolization qua abrupt pidgin-nativization hypothesis? Is the abrupt creolization hypothesis consistent with the social histories of the territories where classic Creoles developed? How can we explain similarities between abrupt Creoles and expanded pidgins when the stabilization and structural expansion of the latter is not necessarily associated with restructuring by children? Is there convincing evidence for assuming that adult speech is less controlled by Universal Grammar than child language is? How can we account for similarities between abrupt creolization and naturalistic second-language acquisition?

Not all creolists who have invoked universalist explanations have made children critical to the emergence of Creoles. For instance, Sankoff (1979) makes allowance for Universal Grammar to operate in adults, too.

Few creolists subscribe nowadays to one exclusive genetic account, as evidenced by the contributions to Mufwene (1993). The 'complementary hypothesis' (Corne 1999, Mufwene 2001) seems to be an adequate

alternative, provided we can articulate the ecological conditions under which the competing influences (between the substrate and superstrate languages, and within each group) may converge or prevail upon each other. This position was well anticipated by Schuchardt (1909, 1914) in his accounts of the geneses of *Lingua Franca* and of Saramaccan. More and more research is now underway uncovering the sociohistorical conditions under which different Creoles have developed, for instance, Arends (1995), Chaudenson (1992, p. 21), Corne (1999), and Mufwene (2001).

Still, the future of research on the development of Creoles has some problems to overcome. So far knowledge of the nonstandard varieties of the lexifiers spoken by the European colonists remains limited. There are few comprehensive descriptions of Creoles' structures, which makes it difficult to determine globally how the competing influences interacted among them and how the features selected from diverse sources became integrated into new systems. Few structural facts have been correlated with the conclusions suggested by the sociohistorical backgrounds of individual Creoles. Other issues remain up in the air, for instance, regarding the markedness model that is the most adequate to account for the selection of features into Creoles' systems. For developmental issues on PCs, the following recently edited collections are good starting points: Muysken and Smith (1986), Mufwene (1993), and Arends et al. (1995). More specific issues may be checked in volumes of the Creole Language Library (John Benjamins) and of Amsterdam Creole Studies, in the *Journal of Pidgin and Creole Languages*, and in *Etudes Créoles*. Several issues of *Pacific Linguistics* also include publications on Melanesian Creoles.

4. Creolistics and General Linguistics

There is much more literature on the genesis, sociology, and morphosyntax of PCs than on their phonologies, semantics, and pragmatics. With the exception of time reference and nominal number, studies in semantics and pragmatics are scant. On the other hand, the development of quantitative sociolinguistics owes a lot to research on AAE since the mid-1960s (see, e.g., Labov 1972) and Caribbean English Creoles (e.g., Rickford 1987). Numerous publications in *American Speech*, *Language in Society*, and *Language Variation and Change* reflect this. There are also several surveys of creolistics. Holm (1988, 1989), Arends et al. (1995), and Mühlhäusler (1986, 1997), cited here, are just some of the references, which vary both in adequacy and in geographical areas of focus. It is hoped that, Corne (1999) is the beginning of a new trend of comparative studies of Creoles lexified by the same language.

Studies of structural aspects of Creoles have yet to inform general linguistics beyond the subject matters of time reference and serial verb constructions. For instance, studies of *lectal continua* have had this potential, but little has been done by creolists to show how their findings may apply to other languages. The mixed nature of *mesolects*, those intermediate varieties combining features of both the *acrolect* and the *basilect* should have informed general linguistics against the fallacy of assuming monolithic grammatical systems. However, little has been done on the subject matter. Likewise, the debate on Creole genesis could have informed historical linguistics on the importance of varying external conditions to language change.

Although lack of consensus among creolists may be invoked as a general reason for this failure to influence general linguistics, alarming indifference from theoretical linguists, especially those engaged in theories of typology and universals, is a more important reason. Consensus cannot be expected of creolistics any more than of other subfields of linguistics or any other scientific discipline. However, in the broader context of language contact (including second-language acquisition), studies of especially Creole genesis have been inspiring. For instance, Thomason and Kaufman (1988) are cited widely in studies of indigenized English. Andersen (1983) was an important step to consolidate common interests between second-language acquisition and Creole genesis. More cross-fertilization might be expected between studies of Creole genesis and those of (child) language development (DeGraff 1999), as among diverse subfields of linguistics.

See also: Archaeology and the History of Languages; Creolization: Sociocultural Aspects; Dialectology; Evolution and Language: Overview; Language and Ethnicity; Language and Society: Cultural Concerns; Language Endangerment; Language Policy: Linguistic Perspectives; Language Policy: Public Policy Perspectives; Melanesia: Sociocultural Aspects

Bibliography

Adam L 1883 *Les idiomes négro-aryens et malayo-aryens: essai d'hybridologielinguistique*. Maisonneuve, Paris

Alleyne M C 1971 Acculturation and the cultural matrix of Creolization. In: *Pidginization and Creolization of Languages*. In: Hymes D (ed.) Cambridge University Press, Cambridge, UK, pp. 169–86

Alleyne M C 1980 *Comparative Afro-American: An Historical-Comparative Study of English-based Afro-American Dialects of the New World*. Karoma, Ann Arbor, MI

Andersen R (ed.) 1983 *Pidginization and Creolization as Language Acquisition*. Newbury House Publishers, Inc., Rowley, MA

Arends J (ed.) 1995 *The Early Stages of Creolization*. John Benjamins, Amsterdam

Arends J, Muysken P, Smith N (eds.) 1995 *Pidgins and Creoles: An Introduction*. John Benjamins, Amsterdam

Baker P, Mühlhäusler P 1990 From business to pidgin. *Journal of Asian Pacific Communication* **1**: 87–115

Bickerton D 1984 The language bioprogram hypothesis. *Behavioral and Brain Sciences* **7**: 173–221
Bloomfield L 1933 *Language*. Holt, Rinehart and Winston, New York
Chaudenson R 1992 *Des îles, des hommes, des langues: essais sur la créolisation linguistique et culturelle*. L'Harmattan, Paris. Revised, English edition, *Creolization of Language and Culture*, Routledge, London
Coelho F A 1880–1886 Os dialectos românicos ou neolatinos na Africa, -Asia, ae America. *Bolletim da Sociedade de Geografia de Lisboa* **2**: 129–96 (1880–1881), **3**: 451–78 (1882), **6**: 705–55 (1886)
Corne C 1999 *From French to Creole: The Development of New Vernaculars in the French Colonial World*. University of Westminster Press, London
DeCamp D 1971 Toward a generative analysis of a post-Creole speech continuum. In: Hymes D (ed.) *Pidginization and Creolization of Languages*. Cambridge University Press, Cambridge, UK, pp. 349–70
DeGraff M (ed.) 1999 *Language Creation and Language Change: Creolization, Diachrony, and Development*. MIT Press, Cambridge, MA
Faine J 1937 *Philologie créole: études historiques et étymologiques sur la langue créole d'Haïti*. Imprimerie de l'Etat, Port-au-Prince, Haiti
Hall R A Jr. 1958 Creole languages and genetic relationships. *Word* **14**: 367–73
Hall R A Jr. 1962 The life cycle of pidgin languages. *Lingua* **11**: 151–6
Holm J 1988 *Pidgins and Creoles: Theory and Structure*. Cambridge University Press, Cambridge, UK, Vol. 1
Holm J 1989 *Pidgins and Creoles: Reference Survey*. Cambridge University Press, Cambridge, UK, Vol. 2
Keesing R M 1988 Melanesian *Pidgin and the Oceanic Substrate*. Stanford University Press, Stanford
Krapp G P 1924 The English of the Negro. *The American Mercury* **2**: 190–5
Labov W 1972 *Language in the Inner City: Studies in Black English Vernacular*. University of Pennsylvania Press, Philadelphia
Lalla B, D'Costa J 1990 *Language in Exile: Three Hundred Years of Jamaican Creole*. University of Alabama Press, Tuscaloosa, AL
Lefebvre C 1998 *Creole Genesis and the Acquisition of Grammar: The Case of Haitian Creole*. Cambridge University Press, Cambridge, UK
McWhorter J H 1998 Identifying the Creole prototype: Vindicating a typological class. *Language* **74**: 788–818
Mufwene S S (ed.) 1993 *Africanisms in Afro-American Language Varieties*. University of Georgia Press, Athens, GA
Mufwene S S 2001 *The Ecology of Language Evolution*. Cambridge University Press, Cambridge, UK
Mühlhäusler P 1986 *Pidgin and Creole Linguistics*. Basil, New York
Muysken P, Smith N (eds.) 1986 *Substrata Versus Universals in Creolegenesis*. John Benjamins, Amsterdam
Poplack S (ed.) 1999 *The English History of African-American English*. Blackwell, Oxford, UK
Rickford J R 1987 *Dimensions of a Creole Continuum: History, Texts, and Linguistic Analysis of Guyanese Creole*. Stanford University Press, Stanford
Sankoff G 1979 The genesis of a language. In: Hill K C (ed.) *The Genesis of Language*. Karoma, Ann Arbor, MI, pp. 23–47
Sankoff G, Brown P 1976 The origins of syntax in discourse: A case study of Tok Pisin relatives. *Language* **52**: 631–66
Schuchardt H 1909 Die Lingua Franca. *Zeitschrift fur Romanische Philologie* **33**: 441–61
Schuchardt H 1914 *Die Sprache der Saramakkaneger in Surinam*. Johannes Muller, Amsterdam
Smith N 1995 An annotated list of pidgins, Creoles, and mixed languages. In: Arends J, Muysken P, Smith N (eds.), pp. 331–74
Sylvain S 1936 *Le Créole Haïtien: Morphologie et Syntaxe*. Wettern, Imprimerie De Meester, Belgium
Thomason S G, Kaufman T 1988 *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley
Turner L D 1949 *Africanisms in the Gullah Dialect*. University of Chicago Press, Chicago

S. Mufwene

Pilgrimage

Pilgrimage involves the movement of a traveler or group of travelers away from home, usually with a specific, sacred goal in mind. In most cases, the term implies a journey to a place that has been sanctified by the presence or actions of a divinity or holy figure. Certain variations on this pattern are possible. Pilgrims sometimes wander continuously without a fixed destination, or the journey may be considered to occur in a purely metaphorical sense, involving spiritual development and transformation within the person. This article focuses, however, on the most common understanding of pilgrimage: literal movement across space that may also result in the crossing of conventional cultural, social, and religious boundaries.

1. *Themes in the Study of Pilgrimage*

Social scientific studies of pilgrimage were relatively sparse until the 1970s, although in the 1950s Eric Wolf had published a short but influential article on pilgrimage to the Virgin of Guadalupe in Mexico (Wolf 1958). Since the early 1970s, research has expanded and covered all of the world religions. Such academic engagement reflects a continuing and even increasing interest in pilgrimage that has become evident among participants who have better access than ever before to safe and effective modes of travel.

Many earlier studies adopted Durkheimian, functionalist perspectives in analyzing the significance and effects of pilgrimage. In other words, they interpreted the process of traveling to and gathering at a sacred place as providing ritualized opportunities for social unification and moral regeneration within populations that were normally widely dispersed. For instance,

Wolf (1958) argued that the shrine of the Virgin of Guadalupe in Mexico City represented a dominant symbol for the whole nation, condensing various historical, mythical, and cultural themes and focusing them on the activity of pilgrimage to a single place.

The most influential social scientific approach to the subject has come from the anthropologist Victor Turner (see, for instance, Turner and Turner 1978). He notes that, after an original founding event such as a vision, journey, or martyrdom, pilgrimages most frequently involve the gradual organization and routinization of the sacred journey. More speculatively, Turner argues that pilgrimage in the world religions shares certain characteristics with rites of passage in tribal societies. In both forms of ritual, the person is removed from the space and time of the everyday social environment and undergoes an emotionally and physically testing experience that can have lasting effects on perceptions of self-identity. Pilgrims thus enter the realm of what Turner calls social antistructure or *communitas*. He means by these terms that pilgrimage can involve the stripping away of everyday social statuses among participants, resulting in liberation from conformity to conventional norms and hierarchies. Pilgrims ideally cultivate direct and egalitarian bonds of comradeship during the temporary, set-apart time of their journey and dwelling at a shrine. *Communitas* therefore has the potential to contain an element of social critique since it puts the taken-for-granted quality of conventional institutions in question. Representatives of orthodoxy in many of the salvation religions often feel ambivalent towards pilgrimage and attempt either to suppress or control its tendencies to foment lay frustration with religious hierarchies.

Turner notes further that pilgrimage is frequently associated with images of death and rebirth, mirroring the symbolic destruction and reconstruction of the religious traveler's identity. For instance, many who go on the major pilgrimage (or *hajj*) to Mecca will choose to be buried in the white robes they wore on the pilgrimage. It is believed that these people will thus present themselves at the Last Judgment in the garb appropriate to a rejection of worldly, profane actions.

Turner's work has proved to be so significant partly because it appears to provide a vocabulary and theoretical framework that can be generalized and applied to varied cultural contexts. However, numerous subsequent works have criticized aspects of his thesis that pilgrimage is inherently antistructural in form and significance. Authors have pointed out that such a paradigm is possibly more reflective of a Christian ideal of human fellowship than of social and political realities in religious systems in general. Many examples of pilgrimage appear actually to reinforce situations of conflict between participants. For instance, Messerschmidt and Sharma (1981) examine Hindu pilgrimage in Nepal and find that caste and class divisions are not transcended through common engagement in the ritual. They argue further that conventional Christian ideals of egalitarianism and homogeneity contradict certain Hindu assumptions concerning the inevitable persistence of religious hierarchy.

Powerful and widely cited criticisms of the Turnerian paradigm are contained in a book that is ostensibly devoted to studies of Christian pilgrimage but presents arguments that have much wider application. Eade and Sallnow (1991) characterize pilgrimage sites as ideal locations for contesting, rather than consolidating, the sacred among religious participants. As centers of power that can provide religious legitimation for political aspirations, shrines often present powerful incentives for a particular interest group to attempt to control the behavior and beliefs of religiously motivated visitors (the point is illustrated clearly by the politically charged and often violent history of Jerusalem). Pilgrimages, according to this view, provide the material and ideological focus for competing discourses and agendas, not only between sacred and secular authorities, but also between lay pilgrims and clerics, locals and incomers, members of different denominations, and so on. Sacred sites are described by Eade and Sallnow as the equivalent of empty symbolic vessels waiting to be filled by the possibly mutually contradictory assumptions of the various people who visit them. The authors also suggest that to regard pilgrimage as a unitary institution is highly problematic, even within a single religious system, since it can contain so many historically specific behaviors and meanings. They add, however, that Christian and perhaps other forms of pilgrimage can be conceptualized as involving a basic triad of 'person,' 'place,' and 'text.'

In fact, Turner did anticipate at least some of the later criticisms that have been leveled at his model of pilgrimage. He was aware of the potential for pilgrimage sites to be co-opted by political elites, the fragility of states of *communitas*, and the capacity of pilgrimage sites to encourage widely differing interpretations from pilgrims. However, the notion that pilgrimage should be seen as an arena for contestation as much as cooperation has proved popular in recent years. Its adoption by many scholars is not surprising since it resonates with contemporary social scientific perspectives that have been influenced by postmodern understandings of culture as unstable, fragmentary, and open to multiple appropriation.

A less controversial set of issues, discussed by Turner and others, is raised by the contention that pilgrimage can encourage participants to consider themselves part of translocal religious and cultural systems. Among Hindus, for instance, a sense of participation in a sacred geography that covers the whole of the Indian subcontinent can be cultivated by travel to shrines. Some devotees choose to circumambulate the entire country, visiting sacred spots along the way. Holy places themselves may be officially

or unofficially ranked in importance, thus constituting an internally heterogeneous but widely dispersed framework of sites. Among Christians, Jerusalem (located within 'the Holy Land' itself) has often been regarded as the pre-eminent location for pilgrimage, containing as it does so many sites associated with biblical events. It is followed in importance by other locations of transnational importance, such as Rome or Compostela, and then by pilgrimage locations that have significance in national or more narrowly defined regional contexts.

2. *Methods of Study*

Pilgrimage raises unusual problems in relation to conventional social scientific methodologies. Any given site is located within a much wider political, economic, and religious framework that must be taken into account in analysis. As a religious practice, pilgrimage has often complemented (and incorporated) other translocal activities like trade, tourism, and exploration. Pilgrims themselves may move across geographical, social, and religious boundaries, so that the scholar is faced with a diffuse and literally shifting object of study.

Various methods can be used to encapsulate the most important aspects of the phenomenon. Pilgrims can be followed as they journey to a shrine (Gold 1988, Frey 1998). More commonly, a single pilgrimage center or shrine is focused upon, and the activities of various pilgrim groups observed as they pass through the sacred center (Dubisch 1995). At a wider, macro level of analysis, whole systems of pilgrimage within a religious tradition may be studied (Turner and Turner 1978), or comparisons made between pilgrimage systems in the various world religions (Coleman and Elsner 1995).

The fact that pilgrimage sites tend to persist over time provides a good opportunity to combine historical with sociological interests (Turner and Turner 1978), and much work in this area demonstrates the ways in which a given site can retain its identity as a sacred place despite the fact that it is appropriated by different and sometimes competing religious systems over perhaps hundreds of years. Guadalupe, for instance, provides a good example of a shrine that has probable pre-Christian as well as Christian roots. The mixture of religious associations increases its significance as a site whose symbolism and mythology can be invoked in current debates concerning the relationship between precolonial, colonial, and postcolonial identities in Mexico.

3. *Future Themes of Research*

Pilgrimage has the potential to be at the center of much future research that could be of importance to the social sciences in general. As a social phenomenon, it encourages the researcher to engage with some key theoretical debates in a variety of disciplines. The current popularity of pilgrimage as an activity raises the question of its significance for predictions of secularization and/or religious revitalization in various parts of the world. Some scholars have also speculated over the extent to which apparently secular activities such as going on holiday or visiting the graves of pop stars share the same structural features as pilgrimage. Certainly, it seems likely that tourism and pilgrimage will become increasingly intertwined in many cultural contexts in the future, given the popularity and flexibility of both activities. In addition, pilgrimage studies are relevant to ongoing discussions of the significance of processes of globalization in the modern world. Studies of sacralized forms of travel, involving the transference of people, ideas, and objects across national borders, can complement other research that is concerned to understand how social, cultural, and political units are being transformed by developments in technologies of mass travel and communication. Furthermore, in a world that is characterized by the displacement of populations from traditional homelands, pilgrimage may emerge as an increasingly important institutional means through which to unite diasporic communities. Tweed (1997) has recently argued, for instance, that pilgrimage allows Cuban exiles in Florida to map the landscape and history of their country of birth on to a new environment. Finally, pilgrimage provides many possible avenues of research in the light of increased social scientific interest in the body and embodiment, ranging from a focus on beliefs in miraculous healing to analysis of forms of movement associated with travel to and around sacred shrines.

See also: Migration, Theory of; Monasticism: West and East; Traditions: Social

Bibliography

Coleman S, Elsner J 1995 *Pilgrimage: Past and Present in the World Religions*. Harvard University Press, Cambridge, MA

Dubisch J 1995 *In a Different Place: Pilgrimage, Gender, and Politics of a Greek Island Shrine*. Princeton University Press, Princeton, NJ

Eade J, Sallnow M J (eds.) 1991 *Contesting the Sacred: The Anthropology of Christian Pilgrimage*. Routledge, London

Frey N L 1998 *Pilgrim Stories: On and Off the Road to Santiago*. University of California Press, Berkeley, CA

Gold A G 1988 *Fruitful Journeys: The Ways of Rajasthani Pilgrims*. University of California Press, Berkeley, CA

Morinis A (ed.) 1992 *Sacred Journeys: The Anthropology of Pilgrimage*. Greenwood Press, Westport, CT

Messerschmidt D A, Sharma J 1981 Hindu pilgrimage in the Nepal Himalayas. *Current Anthropology* **22**: 571–2

Turner V, Turner E 1978 *Image and Pilgrimage in Christian Culture*. Columbia University Press, New York

Tweed T A 1997 *Our Lady of the Exile: Diasporic Religion at a Cuban Catholic Shrine in Miami*. Oxford University Press, Oxford, UK

Wolf E 1958 The Virgin of Guadalupe: A Mexican national symbol. *Journal of American Folklore* **7**(1): 34–9

S. Coleman

Place in Geography

'Place' is vivid in experience, yet elusive as a concept. As a general term it embraces an aspect of the diversity of the world that is necessarily experienced though countless unique places. In spite of this ambiguity, 'place' in geography was considered for a long time to be more less synonymous with 'region' or 'settlement' and so obvious that it needed no clarification. In the last half of the twentieth century this apparent obviousness has been brought into question as ideas about 'place' have been subjected to critical theoretical interpretation, and as changes in the character of actual 'places' have suppressed or distorted diversity. This article describes this sequence of change and the relationships between them.

1. *'Place' and 'Places'*

Aristotle, in his *Physics*, noted that 'The question, What is Place?, presents many difficulties for analysis,' then concluded that it refers to the precise dimensions of the space that contains something—'place' is a neutral container (McKeon 1941). This is a flexible but not very edifying definition since a container can range in scale from the space on my desk where I keep my dictionary, to Canary Wharf, to Paris, to Australia, to the earth as the place of humankind.

In contrast, Strabo (1917) wrote about 5 BC in the introduction to his *Geography* that 'A knowledge of places is conducive of virtue.' He believed that geographers, with their habits of travel and knowledge of the specific parts of the world, were qualified to distinguish good settings for human activity from bad ones, and by implication that content was more important than the container. The difficult question of 'place' faded into the background for almost 2,000 years, and, while it became acceptable to describe geography as 'the study of places' (Hartshorne 1959), this was left unexplained because 'a place' was apparently an unproblematic concept. The issue of scale was left, and remains, unresolved.

So 'place' has an enigmatic character, partly general and partly specific, midway between objective fact and subjective feeling. Entrikin (1991) calls this 'the betweenness of place' and proposes that it is best viewed from a point in the middle. From this perspective, 'a place' appears as a combination of objects and meanings that distinguishes somewhere from its surroundings, regardless of scale. For Walter (1988) it is a 'location of experience; the container of shapes, powers, feelings and meanings,' and for Cresswell (1996) it is 'a meaningful segment of space' and a 'container of social power' (see *Geographic Thought, History of*; *Region: Geographical*; *Regional Identity*; *Scale in Geography*.

2. *Experience and Sense of Place*

In the 1960s a new theoretical and scientific approach to the spatial geometry of location, known as central place theory, served as a catalyst for more rigorous thinking about 'place' (see *Location Theory*). This approach considered towns and villages in terms of measurable properties and relative locations, and thus implicitly adopted an Aristotelian perspective that emptied places of content.

Geographers interested in specific places were dismayed by the aridity of central place theory but recognized that it was impossible to continue to use the concept of 'place' uncritically. Phenomenology was identified as a philosophical method that allows rigorous interpretations of place without denying diversity and meaning. The writings of M. Heidegger and M. Merleau-Ponty were used to argue that ideas of place are rooted in experiences of dwelling in specific places, and that these precede all abstractions about location, environment, or geography. Places are aspects of everyday experience, territories of symbols, memories, and associations; they are centers of being and belonging that connect people with their worlds.

Phenomenology discloses several elements of 'place' and place experience, including identity and sense of place, topophilia, and home. These insights have been absorbed into most subsequent discussions of place in geography, first in humanistic approaches and more recently in arguments which view place as a social construction (see *Humanistic Geography*).

2.1 *Identity and Spirit of Place*

Identity is what makes somewhere distinctive; it also suggests similarities, as when it is claimed that a particular fast-food restaurant is identical with one elsewhere. The term 'identity of place,' therefore, acknowledges that every place simultaneously differs from and has similarities with other places. This apparent contradiction requires clarification.

From an abstract perspective, it is possible to distinguish four essential components of place identity: location; an appearance of hills, rivers, buildings, and other objects; a community engaged in various social and economic activities; and a set of meanings and symbols. These are distinct and are not reducible one to the other (one cannot discuss location in terms of architecture, and so on), but these components are, in real places, always closely interwoven. Oaxaca in

southern Mexico is a colonial city that shares its architectural tradition and town plan with numerous cities in Latin America, and there is nothing unusual about its role as a center of regional government or its passion for fiestas. However, its location is unique, as are its elegant streets and buildings, and it seems unlikely that anywhere else has a annual festival—La Noche de los Rabanos—dedicated to carved radishes. In all places, even unremarkable ones such as new suburbs, some aspects are shared and others are distinctive, though the balance of these can vary widely. Even if appearance and activities are ordinary, meanings and symbols are always in some way distinctive.

Somewhere with exceptional place identity, like Oaxaca, is sometimes said to possess 'spirit of place.' This expression comes from the Latin *genius loci*, but the idea is derived from the ancient and widespread belief that certain places are inhabited by their own gods or spirits. In ancient Greece almost every hill-top, spring, grove, and outcrop of rock had its own guardian spirit; the Acropolis was the home of the goddess Athena. In this more secular age 'spirit of place' has lost its association with gods, and now refers to the immediately apparent yet rather elusive aesthetic quality of striking distinctiveness. Spirit of place is what makes particular places attractive and memorable.

2.2 Sense of Place

Sense of place is the capacity to recognize and respond to diverse identities of places. A place affects us directly through sight, smell, sound, and touch and sense of place is the synthetic faculty that combines these impressions with memory, imagination, and reason. (Note, however, that 'sense of place' is popularly and confusingly used to be synonymous with 'spirit of place'; thus, it might be said that 'Provence has a wonderful sense of place').

Sense of place is not a uniformly shared and undifferentiated faculty. Many people are uninterested in places, some have an instrumental view of the world that regards places as resources to be exploited, while others are finely tuned to *genius loci* and place diversity. These variations have been clarified by making a distinction between an outsider's sense of place, which looks at the world mostly in terms of objective qualities, and an insider's sense of place, which seeks to grasp the distinctive properties of particular places in terms of their meanings for those who live and work in them. An outsider's perspective pays scant attention to distinctiveness and is inclined to impose judgments on the basis of previously formed opinions; it is manifest, for example, in colonialism and in the propagation of placelessness (see Sect. 3.1). In contrast, an insider's sense of place is responsive to the idiosyncracies of environment and culture. While it comes with knowing and being known somewhere, and is most apparent in home places, it is also an ability that can be learned or enhanced by developing skills of careful observation and imagination.

2.3 Topophilia and Home

'Topophilia' lies at the intersection of spirit and sense of place. Literally it means 'love of place,' though Tuan (1974), who has given the term currency, defines it as 'the affective bond between people and place or setting.' There are, he suggests, 'landscapes of persistent appeal,' such as mountains and seashores, islands, carefully tended farms, and small towns. These are perceived through lenses of culture, familiarity, memories, even personal moods. Topophilia blends the inherent qualities of landscapes and places with the attitudes that different individuals bring to it. There are no fixed relationships in this. Sometimes a beautiful setting can be soured by ill-temper, yet at other times that same beauty may displace ill-temper.

Underlying an insider's sense of place and topophilia is the notion and experience of 'home.' It is in 'home' where the significance of individual lives is most concentrated, where people feel they belong, where love of place is strongest. Home is a powerful emotional bond that can, for example, draw back refugees and those who have been uprooted in spite of the most adverse circumstances. Because of the intensity of these associations, in phenomenological interpretations 'home' is identified as the archetypal place.

2.4 Corrupted Sense of Place

Phenomenological interpretations mostly have been concerned with clarifying the subjectivity of place experience. In emphasizing roots and home they imply that 'place' and 'sense of place' are positive and provide the foundation for a deeply rooted yet balanced outlook on the world. This is not always the case. Harvey (1989) argues that consequences of a strong sense of place include sectarian politics and parochialism. This is apparent in local and neighborhood practices that resist anything which might upset racial homogeneity or promote social equity (see *Locational Conflict (NIMBY)*). There is a fine and unclear line between such intolerance and efforts to protect a place from massive redevelopment and equivalent changes that might destroy its identity, but when that line is crossed sense of place can be said to have become corrupted. A corrupted or poisoned sense of place selectively stresses local identity and home to the virtual exclusion of shared values (Relph 1997). It is always unpleasant, and at its worst, in xenophobia and ethnic cleansing, it can be brutally violent.

3. Modernist and Postmodernist Changes in 'Places'

At the beginning of the twentieth century most peoples' lives were, as they always had been, largely confined to within about one day's walk of where they lived. Nowhere was entirely isolated from outside contact, but with lack of mobility social changes dispersed slowly and usually adapted to specific contexts, so the local distinctiveness and geographical diversity were continually reinforced. This diversity was created under social and technological conditions dramatically different from those which now prevail. Since the mid-nineteenth century technologies of electronic communications and mass travel, coupled with rapid cultural and economic changes, have collapsed distance. The ways places are made, and the ways they are experienced, have both changed. Old places may be regarded with great affection, but they are historical fragments; geographies that are being made now consist mostly of placeless or arbitrarily fabricated settings.

3.1 Placelessness

Placelessness is the condition in which different places look much the same, and, more importantly, offer the same opportunities for experience. It is the erosion of geographical distinctiveness and diversity. Examples are legion: shopping centers, apartment buildings, airport terminals, expressways. Augé (1995) calls these 'nonplaces.' The new high-rise Okura Hotel in Kobe in Japan could equally well be in Houston or Mexico City and the adjacent shopping center is so dislocated from its context that much of it only has signs in English.

In part, placelessness is the expression of modernist styles of architecture and planning that were conceived in the 1920s to take advantage of new building materials of glass, steel, and concrete. These styles are functional, standardized, and international. They respond to the essential sameness of people everywhere and can be applied anywhere, regardless of climate or culture. Placelessness is also the child of transnational corporations and a global economy that strive to provide familiar goods and services to consumers in the most profitable and efficient way possible. Modernist design and corporate capitalism fit together well. The result is that the same products, the same companies, similar buildings, and similar landscapes can be found almost everywhere.

3.2 Place Exploitation and Postmodernism

Since about 1980 placelessness has developed a more subtle and complex character. As uniformity replaces diversity, distinctiveness becomes increasingly valuable and worth maintaining. Sometimes, as in heritage preservation, the aim may be to protect intrinsic local identity, but the larger consequence has been to turn place identity into a commodity. Distinctiveness sells. In tourism, historical districts and unique landscape features are obviously attractive place qualities. Less obviously, in a global economy in which capital is mobile and businesses can locate anywhere with educated workers and reasonable land costs, a distinctive place identity can be decisive factor for a city trying to attract investment.

Once place identity is regarded as a commodity rather than an intrinsic property of somewhere, it can be exploited in many ways to promote tourism and attract other business. It can be selectively enhanced either by reconstructing destroyed buildings or historical districts, as in Place Royale in Quebec City, or by converting former industrial areas into festival markets of restaurants and galleries, as in Covent Garden in London. Links with celebrities or events, real or fictional, can promote marketability, so Atlanta is associated with *Gone with the Wind*; and Wigan has a Wigan Pier shopping center with an Orwell restaurant. If local identity is too bland elements of more distinctive places can be copied. Las Vegas, which excels in place plagiarism, has design fragments of Egypt, Italy, the tropics, New York, Ireland, and Japan in hotels and casinos on its main strip. More restrained copying is to be found in the restaurants, shopping centers, and entertainment districts of most cities, so gondolas ply the otherwise placeless waterfront in Toronto.

The exploitation and fabrication of place identity erodes distinctions between local history and contrived heritage, between fiction and reality, in order to allow capital to pursue its relentless aims. Harvey (1989) argues that these are processes of postmodernity. Their consequence is a placelessness achieved not through uniformity, but through what Sorkin (1992) calls 'ageographical variety.' Place identities have been uprooted, turned into transportable goods, and rearranged without regard for context.

4. A Progressive or Critical Sense of Place

These changes in the character of actual places raise questions about the notion of 'sense of place.' Phenomenological accounts imply that sense of place is constant across space and time and unaffected by electronic communications and increased travel. A different interpretation, based in critical social theory and political economy, proposes what Massey (1993) calls 'a progressive sense of place.' Instead of viewing places as bounded settings rooted in home and local traditions, a progressive sense of place looks outward from a specific geographical context to connections with wider processes. It sees each place as a unique point of interaction within networks of social and economic relations. From this perspective places appear as distinctive social spaces of inclusion or

exclusion, within a broad and changeful geographical context (Cresswell 1996).

A progressive or critical sense of place offers a possible foundation, that is neither nostalgic nor exclusionary, for understanding recent social and urban changes (Harvey 1996). It recognizes the global processes at work in place exploitation and, therefore, can provide insights into cultural 'displacements' and unanticipated juxtapositions of mosques, Chinatowns, and protestant churches that are the consequence of the intercontinental population movements of the last 50 years and which have turned many cities and towns into multicultural and multicentred places (see *Place, Power of*).

5. Place-making

As a concept, 'place' has evolved rapidly from a background, neutral term to one that is an important theoretical and empirical subject not only in geography, but for many of the social sciences (Walter 1988, Augé 1995, Casey 1997). A simultaneous, possibly coincidental shift from distinctiveness to placelessness and ageographical variety has occurred in the identity of actual 'places.' A repeated theme in recent discussions of these changes in 'place' and 'places,' regardless of discipline and regardless of whether the approach is theoretical, empirical, or phenomenological, is the need for individuals and local communities to be involved in making and maintaining their own places. Places have ceased to be objects for description and have come to be regarded as complex theoretical phenomena, as foundations for social resistance against remotely generated economic forces, and as centers of stability in world of change.

See also: Geography; Place, Power of; Postmodernism in Geography; Urban Geography

Bibliography

Augé M 1995 *Non-places: Introduction to the Anthropology of Super-modernity*. (trans. Howe J) Verso, London

Casey E S 1997 *The Fate of Place: A Philosophical History*. University of California Press, Berkeley, CA

Cresswell T 1996 *In Place/Out of Place: Geography, Ideology and Transgression*. University of Minnesota Press, Minneapolis, MN

Entrikin J N 1991 *The Betweenness of Place: Towards a Geography of Modernity*. The Johns Hopkins University Press, Baltimore, MD

Hartshorne R 1959 *Perspectives on the Nature of Geography*. Rand McNally, Chicago

Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK

Harvey D 1996 *Justice, Nature and the Geography of Difference*. Blackwell, Oxford, UK

Massey D 1993 Power-geometry and a progressive sense of place. In: Bird J, Curtis B, Putnam T, Robertson G, Tickner L (eds.) *Mapping the Futures: Local Cultures, Global Change*. Routledge, London and New York

McKeon R (ed.) 1941 *Aristotle: Basic Works*. Random House, New York

Relph E 1976 *Place and Placelessness*. Pion, London

Relph E 1997 Sense of place. In: Hanson S (ed.) *Ten Geographical Ideas that Changed the World*. Rutgers University Press, New Brunswick, NJ

Sorkin M (ed.) 1992 *Variations on a Theme Park: The New American City and the End of Public Space*. Hill and Wang, New York

Strabo 1917 *The Geography of Strabo*, 8 Vols. (trans. Jones H L). Heinemann, London

Tuan Y-F 1974 *Topophilia*. Prentice-Hall, Englewood Cliffs, NJ

Walter E V 1988 *Placeways: A Theory of the Human Environment*. University of North Carolina Press, Chapel Hill, NC

E. Relph

Place, Power of

The 'power of place' refers to the spatial form of landscapes, towns, and buildings as an important aspect of social and political life. The phrase is used in various contexts: analyzing social life in existing places, advocating the preservation of historic places, and formulating goals for the design of new places. As a phrase, 'power of place' combines the meanings of location, locale, and 'sense of place' (Agnew and Duncan 1989). It involves an approach to aesthetics drawn from work dealing with 'local distinctiveness' (Clifford and King 1990) in architecture, landscape architecture, cultural geography, and environmental psychology, with an approach to politics drawn from work on space in the social sciences (Hayden 1995). Its practical applications in urban planning include preservation and urban design.

1. Definitions of Place and Power

'Place' is one of the trickiest words in the English language. It carries the resonance of homestead, location, and open space in the city as well as a position in a social hierarchy. The authors of books on architecture, photography, cultural geography, poetry, and travel rely on 'sense of place' as an aesthetic concept but often settle for 'town character,' 'local distinctiveness,' or 'the personality of a location' as a way of defining it. In the nineteenth century and earlier, place also carried a sense of the right of a person to own a piece of land or to be a part of a social world, and in this older sense, place carries more political history with it. Phrases like 'knowing one's place' or 'a woman's place' still carry some territorial meanings and imply political power relationships (see *Community Power Structure*).

1.1 Place and Social Life

People experience places in ways that are critical to their well-being or distress. An individual's sense of place is both a biological response to the surrounding physical environment and a cultural creation. From childhood, humans come to know places through engaging all five senses: sight, sound, smell, taste, and touch. Research on perception shows the simultaneous engagement of several senses in orientation and wayfinding. Children show an interest in landmarks at three or earlier, and by age five or six can read aerial maps with great accuracy and confidence, as an example of the human ability to perceive and remember places (Tuan 1977).

As social relationships are intertwined with spatial perception, the subject of human attachments to places, as well as the ability to conceptualize places, is a fascinating one that defies simplistic research models in any field. Environmental psychologists define 'place attachment' as having social, material, and ideological dimensions (Altman and Low 1992). Some sociological studies of urban renewal, especially those which study the process of mourning for a lost neighborhood, have utilized attachment theory as well (Marris 1986). Many sociologists speak of 'locale' as part of fieldwork; anthropologists deal with the design of settlements and living space as part of the material culture of any group. Studies in the cognitive mapping of places often reveal the social dimensions of knowledge of the city. If place does provide an overload of possible meanings and methods of inquiry for the researcher, philosophers have argued that place's very same assault on all ways of knowing (sight, sound, smell, touch, and taste) make it powerful as a way of structuring social attachments and as a container of memory. Thus the concept of *lieux de memoire* (sites of memory, or places of memory) popularized by Pierre Nora in France, and expanded as processes of remembrance (Hayden 1998).

1.2 Place as Cultural Landscape

Cultural landscape studies, as geographer Carl Sauer and the 'Berkeley School' developed them from the 1940s on, focused on the evolution of places and included the 'combination of natural and man-made elements that comprises, at any given time, the essential character of a place' (see *Landscape Architecture*). Cultural landscape studies emphasized the connection of natural environments and built interventions. Yet cultural landscape methods for studying places, and people's shaping of them, tended to stress the physical and not the political dimensions of places, leaning to the study of rural, pre-industrial landscapes, vernacular house types, and patterns of cultivation, considering ecology but avoiding issues of political contestation. Among the noted essayists of this type was John Brinckerhoff Jackson (1997).

1.3 Places and Power Struggles

As the landscape is more densely inhabited, economic and social forces affecting it are more complex, change is rapid, layers proliferate, and often abrupt spatial discontinuities result. The shaping of urban places involves gaining political access to territory as well as owning, planning, designing, constructing, using, and demolishing various kinds of built space. One can't simply turn to work on location in economic geography (or any other kind of quantitative analysis) because there the human experience of place is an essential topic that is often lost. Rather, the cultural landscape model has been more firmly connected to urban sociology, anthropology, and political science, retaining the physical analysis necessary to convey the shape of urban places, while adding more focused analysis of social and economic contestation (Duncan and Ley 1993, Groth and Bressi 1997). Historians are also claiming this subject matter: 'The story of how different peoples have lived and used the natural world will become one of the most basic and fundamental narratives in all of history, without which no understanding of the past could be complete' (Cronon 1992).

It has often proved easier to study either the natural or the built components of an urban cultural landscape than to wrestle with the combination of the two, fusing analyses of place and power. For many environmental historians, the deployment of land and natural resources has been a central preoccupation. Urban historians have focused more on power struggles; architectural historians on styles and patterns of building. The phrase, 'power of place,' suggests that all three areas are intertwined in some of the spatial struggles of urban development. The aesthetic qualities of architecture or of the built environment cannot be divorced from those of the natural environment without strange dislocations of experience: expensive houses swathed in smog; skyscrapers built on unstable soils, freeways blocking river fronts.

1.4 Avoidance of 'Place'

In the last decades of the twentieth century, social scientists frequently avoided 'place' as a concept, and thus sidetracked the sensory, aesthetic, and environmental components of the urbanized world, in favor of more quantifiable research with fewer epistemological problems (Agnew and Duncan 1989). Melvin Webber used 'non-place urban realm' to refer to the decline of face-to-face activities and the rise of telephone and television. Some analysts have extended this by arguing for the importance of an increasingly 'placeless world' (Relph 1976) or 'non-places' (Augé 1995) dominated by 'organization space' (Easterling 1999). Others prefer to retain the concept of place, and critique unpleasant, unlivable places (Hayden 1995). A paved-over neighborhood, restructured as freeway

lanes, parking lots, and mall, should still be considered a place, if only to explain why the new space lacks the pedestrian scale, lively texture and local meanings of the older place.

2. Connections between Place and Space

Sociologists' analyses of the 'production of space' provide a framework that can be used to relate the 'power of place' to the political economy. Lefebvre (1991) argues that every society in history has shaped a distinctive social space that meets its intertwined requirements for economic production and social reproduction. In terms of production, this would be close to cultural geography in identifying spaces or landscapes shaped for mining, manufacturing, commerce, or real estate speculation. Most original is his analysis of the space of social reproduction, which ranges over different scales, including the space in and around the body (biological reproduction), the space of housing (the reproduction of the labor force), and the public space of the city (the reproduction of social relations). Here he links the physical to the social in decisive ways. Cultural critics praised Lefebvre: he 'called for a new kind of spatial imagination capable of confronting the past in a new way and reading its less tangible secrets off the template of its spatial structures—body, cosmos, city' (Jameson 1991). His analysis also fueled the 'non-place' debate: a quality of late capitalist space is the creation of many identical units—similar but not 'placeless' places—by the large commercial real estate market which has become, in itself, a distinguishing feature of the global economy.

The cultural costs in terms of identity, history, and meaning have also been weighed, and here the phrase 'power of place' is used again. In a world increasingly given over to global corporations and their large, box-like buildings unrelated to local landscapes, the importance of local places as icons of resistance increases. Place-based social movements may resist demolition, preserve historic places, or demand place-sensitive urban design and architecture.

3. Place-based Social Movements

One of the consistent ways to limit the economic and political rights of groups has been to constrain social life through limiting access to places. The interplay between the social and the spatial is constant. The women's rights movement of the late nineteenth and early twentieth centuries was a struggle against spatial segregation based on gender. Women were not allowed to walk freely in public, obtain education, or vote. So too, the civil rights movement of the 1960s emphasized overcoming spatial segregation. Jim Crow laws restricted the African American community: streets, neighborhoods, schools, colleges, hotels, stores, fire stations, swimming pools, and cemeteries were all segregated. To understand the intersecting struggles of race, class, and gender, the spatial dimensions of traditional 'woman's sphere' would have to be studied in combination with the spatial limits imposed by race or class. In the twentieth century, urban renewal often attempted to target poor neighborhoods and those inhabited by people of color. Resistance to demolition was a place-based campaign.

3.1 The Power of Place, Los Angeles

In Los Angeles, The Power of Place, a non-profit corporation, was founded by urban historian and architect Dolores Hayden. Operating from 1984–91, it documented and interpreted the political dimensions of place as historical experience, with a special concern for gender, class, and ethnicity. One premise of this project was that architects, planners, and artists, as well as specialists in preservation, have an important role to play in making the entire city more comprehensible, and its history more visible, in the minds of citizens. The group traced an itinerary of work places in downtown connected to the growth of the city: citrus farms, flower farms, oil wells, garment factories, and other economic activities between 1780 and 1940. On this itinerary, buildings were preserved and interpreted, and new public art and open space designs were created (Hayden 1995). Similar projects were organized in London by a group called Common Ground (Clifford and King 1990) and continue to be developed in many other American cities.

3.2 Civil Rights Museums

It is possible to identify historic urban places which have special significance to certain populations fighting spatial segregation of different kinds: a church where major civil rights meetings were held; a local newspaper that crusaded for fair housing; the first place in a city where women tried to vote; the first place where new immigrants were allowed to own land; the first integrated primary school. The Women's Rights National Park in Seneca Falls, New York is one such museum. There are also sites of assassinations, lynchings, massacres, and riots. The motel where Martin Luther King was shot, the Lorraine Motel in Memphis, Tennessee, is now a national civil rights museum, but every city and town has similar landmarks where territorial struggles have been waged.

3.3 The Lower East Side Tenement Museum

Another kind of museum to exploit the power of place is the Lower East Side Tenement Museum in New York. Ordinary dwellings contain the physical evi-

dence necessary to analyze the material conditions of ordinary workers' lives in earlier times. Standing inside a tenement apartment—perhaps four hundred square feet of living space for an entire family, minimal plumbing, only one or two exterior windows—leaves a visitor gasping for air and looking for light. The claustrophobic experiences of immigrants living for decades in crowded, unhealthy space are conveyed by the building in a way that a text or a chart can never match. Because a tenement was typical of hundreds of thousands of tenements all over the city, its implications for the historic urban landscape are broad. One building tells the bitter story of thousands of developers and millions of tenants. When this museum was acquired by the National Trust for Historic Preservation in the late 1990s, it signaled new prestige for social history museums everywhere.

3.4 Parades and the Power of Place

Festivals and parades also help to define the power of place by staking out routes in the urban cultural landscape. They intermix vernacular arts traditions (in their costumes, floats, music, dance, and performance) with spatial history (sites where parades begin, march, and end). African American jazz funerals marching through the streets of New Orleans, Chinese New Year parades, saint's day processions in Irish American or Italian American Catholic communities, and graveyard ceremonies for the Day of the Dead in Mexican American communities are just a few examples. Political parades have been representing communities and their causes for a long time, from workers with the tools of their trades to suffragettes. Since the 1960s, civil rights marches in Southern cities, women's marches to 'Take Back the Night,' and Gay and Lesbian Pride Parades in major cities have also been part of campaigns to achieve greater political representation.

4. The Power of Place—Applications in Urban Design

Some recent books proclaim the imminent end of urban public places. Advances in computer technology have affected many parts of daily life and often dispersed neighborhood activities. 'Disneyfication' fills theme parks and casinos with entertainments built to resemble older towns. Shopping malls take up historicist built forms as well. The global migration of capital, and the resultant production of 'organization space' emphasizes freeway interchanges or airports as locational determinants, eroding the importance of older central cities. Virtual reality promises to simulate more and more spatial experiences.

Nevertheless, human beings still situate their lives and work in material spaces of different scales. There are many signs that places are more important, and more contested, than ever. Numerous older cities and towns have been protected by a growing movement for historic preservation. New town design is frequently done in accordance with the architectural movement promoting pedestrian scale and known as the 'New Urbanism.' Similarly, planners involved with campaigns for 'smart growth' have been heartened to hear that most Americans ranked urban and suburban sprawl ('placelessness') as an important political issue in 2000.

Studies exploring the 'power of place' are rapidly increasing in academic research, although the authors are scattered in half a dozen fields—history, geography, architecture, urban planning, landscape architecture, art history, anthropology, sociology, and others. In the 1980s, 'place' was not quantitative enough for the mainstream of social science, or narrowly aesthetic enough for the mainstream of art history and design. At the start of the twenty-first century, scholarly work on the broad cultural and political importance of places is part of many interdisciplinary programs such as cultural studies or American studies, as well as all of the separate disciplines. Work on the 'power of place' also has much to offer citizens and local public officials. Scholars and practitioners take discussion of 'the power of place' and other public arts to public forums in cities and towns (Imagining America 2000). Understanding the creation of places as a social and political process can lead to new projects to extend public meaning in the urban cultural landscape.

See also: Cultural Landscape in Environmental Studies; Cultural Landscape in Geography; Land Use and Cover Change; Location Theory; Locational Conflict (NIMBY); Political Geography; Power in Society; Social Movements, Geography of; Spatial Equity; Spatial Interaction; Spatial Interaction Models; Spatial Optimization Models; Spatial Pattern, Analysis of

Bibliography

Agnew J A, Duncan J S 1989 *The Power of Place: Bringing Together Geographical and Sociological Imaginations*. Unwin Hyman, Boston, MA

Altman I, Low S (eds.) 1992 *Place Attachment*. Plenum Publishing, New York

Augé M 1995 *Non-places: Introduction to an Anthropology of Supermodernity*. Verso, London

Conzen M (ed.) 1990 *The Making of the American Landscape*. Unwin Hyman, Boston, MA

Clifford S, King A 1990 *Local Distinctiveness*. Common Ground, London

Cronon W 1992 A place for stories: Nature, history, and narrative. *Journal of American History* **78**: 1347–76

Duncan J, Ley D (eds.) 1993 *Place/Culture/Representation*. Routledge, London

Easterling K 1999 *Organization Space*. MIT Press, Cambridge, MA
Groth P, Bressi T (eds.) 1997 *Understanding Ordinary Landscapes*. Yale University Press, New Haven, CT
Hayden D 1995 *The Power of Place: Urban Landscapes as Public History*. MIT Press, Cambridge, MA
Hayden D 1998 Landscapes of loss and remembrance: The case of Little Tokyo in Los Angeles. In: Winter J, Sivan E (eds.) *War and Remembrance*. Cambridge University Press, Cambridge, UK
Hayden D, Dubrow G, Flynn C 1984 *The Power of Place: Los Angeles*. The Power of Place, Los Angeles
Imagining America: Artists and Scholars in Public Life 2000 http://www.ia.umich.edu/
Jackson J B 1997 *Landscape in Sight: Looking at America*. Yale University Press, New Haven, CT
Jameson F 1991 *Postmodernism, or, The Cultural Logic of Late Capitalism*. Duke University Press, Durham, NC
Lefebvre H 1991 *The Production of Space*. Basil Blackwell, Oxford, UK
Marris P H 1986 *Loss and Change*. 2nd edn. Routledge & Kegan Paul, London
Relph E 1976 *Place and Placelessness*. Pion, London
Tuan Y-F 1977 *Space and Place: The Perspective of Experience*. University of Minnesota Press, Minneapolis, MN

D. Hayden

Placebo Studies (Double-blind Studies)

The story of placebo and of the double-blind method is part of the greater saga of the randomized clinical trial. When regulatory agencies mandated in the 1960s and 1970s that clinical trials of new treatments conform to research methodology, including randomization, double-blind design, and placebo controls, they affirmed a shift that had taken place in American medicine. For four millennia, and well into the 1950s, medical practice had been governed by expert opinion, and treatment was mainly placebo, at worst 'filling the cemeteries' (Shapiro and Shapiro 1997).

In clinical trials, patients are the subjects of all treatment modes, including drugs, radiation, surgery, psychotherapy. This article describes how trials have been conducted and the results interpreted. Clinical trials are termed 'open label' when patients and researchers know that the active drug is being given; 'single-blind' when the patient does not know what he or she is receiving; and 'double-blind' when neither the patient nor the research team knows who is receiving what. 'Randomization' denotes the assignment of experimental subjects through a table of random numbers, dice, or other statistical tools. 'Placebo' was first defined for medical use in 1785 as a substance that was pharmacologically inert, given to reinforce a patient's expectation to get well. Glantz's definition of placebo terms the placebo effect 'a response attributable to therapy *per se*, as opposed to the therapy's specific effects' (Glantz 1997). Current usage differentiates active placebo that has pharmacologic effect from inert placebo.

Thomas Kuhn, the historian of science, defines normal science as the prevalent system of beliefs and practice. He likens practitioners to chess players, who test alternative moves but do not challenge the rules of the game. The game is challenged through discovery or a new hypothesis. Practitioners of normal science combat the skeptics, solve the problem, postpone solution, or attempt to reconstruct the discussion from new fundamentals. Kuhn quotes Max Planck: 'A new scientific truth can often prevail only when its opponents die and the new generation grows up.'

Kuhn underlines that resistance to change evolves not from wrong-headedness but from the impossibility of measuring or comparing different standards. Scientists and laymen could not accept Einstein's theory of relativity that proposed that space was curved, because Newtonian physics had proven that space was flat, homogeneous, and unaffected by the presence of matter. In the 1960s and 1970s depression was thought to be caused by psychological factors until challenged by evidence that a functional deficit of neurotransmitters, norepinephrine, dopamine, or serotonin or abnormality of their receptors is the cause. While opposition to this remains, most researchers now are refining particulars of neurotransmitter metabolism.

The acceptance of the double-blind method and placebo controls are among the most important developments in science and illustrate how new discoveries drive new theories that are eventually accepted in place of incompatible, older theories.

1. *History of Placebo*

Placebo concept and the word derives from the Hebrew 'I will walk before,' interpreted to mean 'to please.' It gained a secular connotation in the fourteenth century, to become derisive meaning one who 'servily echoes another' (Shapiro and Shapiro 1997).

The first medical definition of placebo appeared in a medical dictionary in 1785. In the late eighteenth and early nineteenth century, it was defined as medication given to please. Eventually, it meant 'a common method or medicine' and referred to treatments as well as medicaments not yet classified as inert. References to placebo as medications were retained in the medical dictionaries until the 1950s.

The use of placebos as controls in clinical research emerged from their long history as therapeutic agents. For four millennia, medical practitioners utilized thousands of substances, 0.03 to 0.04 percent of which are pharmacologically active. These physicians did not discriminate among active and inactive concoctions, and it is not known whether active drugs were used in

effective doses. In the nineteenth and twentieth centuries, some physicians' rationalization for the use of placebo was mobilizing hope.

Efforts to introduce placebo controls into research designs had been confounded in the 1930s by the growing low esteem of placebo as treatment on the part of the medical establishment. But because of self-delusion or hypocrisy while placebos were considered quack treatments prescribed by quacks, most community physicians also used placebos. They gave vitamins without acknowledging their lack of activity for the conditions for which they were prescribed, for example. Jerome Frank, the American psychiatrist, has said that psychoanalysis might be the major placebo of the twentieth century.

Reformers from different generations concerned with the practices of their mainstream clinical colleagues advocated altertness to self-deception as to the efficacy of treatments. Roueché said that 'results from 10 well-controlled cases were superior to haphazard impressions from 1,000 cases' (Marks 1997).

2. *Medical Theories*

In primitive supernaturalism, illness was attributed to actions of the gods and treated with magic administered by shamans. Greek medicine combined rational and irrational explanations and treatments, using miracles, divine healing, and natural remedies. Dream interpretation was a precursor of the confession and psychotherapy. Galenic medicine of ancient Rome was predicated on Galen's principle of opposites: if a drug caused fever, it would cure chills. Galenic theories had an answer for every question and maintained a tenacious hold on medical practice well into the nineteenth century. For example, Galen had observed pus in wounds and erroneously concluded that it was essential to healing rather than a symptom and made no attempts to relieve it. Because of this belief, thousands died of infection.

Theriac, a concoction that was the mainstay of Galenic medicine, had between 33 to 100+ substances in it, eventually being prescribed as a remedy for almost every human ailment. Theriac contained opium, which alleviated pain but was addictive. The rest of the ingredients were mainly placebos; many were harmful.

The laying on of hands was a hallmark of Christianity's return to mystical explanations of illnesses. Charles II of England used therapeutic touch on almost 100,000 persons to cure scrofula, glandular tuberculosis. Skeptics attributed cures to effects of the imagination, incorrect diagnoses to start with, and spontaneous remissions. In the nineteenth century, Claude Bernard asserted that science had to be 'founded on observation and proved by experience.' He influenced the development of studies on pathology, physiology, and bacteriology, but pharmacology continued to lag behind. However, criticisms of 'normal' medicine and its treatments were voiced, notably by Oliver Wendel Holmes, who urged the retention only of quinine for malaria, mercury for syphilis, colchinicum for gout, iodine for goiter, and ipecac for dysentery. If all medications 'could be sunk to the bottom of the sea, it would be all the better for mankind and all the worse for the fishes,' he said in 1860 in a much-quoted statement. Often paradigm shifts are spurred by knowledge acquired elsewhere.

Laboratory science was making major inroads by the end of the nineteenth and early twentieth century, when therapeutic chemical compounds began to be synthesized and replaced some of the naturally occurring medicaments. This was made possible by progress in laboratory sciences, biochemistry, physiology, genetics, and immunology, which are in turn based on chemistry and physics. Animal experiments where variables could be well controlled became standard.

While the concept of the scientific method was beginning to impact medicine in the nineteenth century, theories of therapeutics based on *a priori* reasoning, were still entertained. Homeopathy was a success because of placebo effects, but also due to its theoretical base that less is more. Miniscule dosing probably protected patients from ill affects of noxious substances dispensed in more robust doses by non-homeopathic physicians.

3. *Beginnings of the Scientific Method*

From antiquity, 'open label' experiments were conducted but results were explained by 'medical ideas (that) were speculative… and medical hypotheses that relied on a priori reasoning (Shapiro and Shapiro 1997). In the seventeenth century, Sir Kenneth Digby, an English Renaissance man, carried out a single blind experiment on a single patient. He attributed the positive results to the popular concept of sympathetic powder (Shapiro and Shapiro 1997).

In 1872, John Tyndall, an English physician, suggested that the community pray for from 3–5 years for patients ill with fatal diseases to see if differences emerged between them and others not prayed for. If or how the suggestion was carried out was not clear.

Towards the end of the seventeenth century through the early eighteenth century, some British physicians called for comparative trials using 'simple arithmetic,' i.e., 'open label' trials. These reformers were generally Scots, provincials, without access to established centers of medical influence and power, and thus, more open to new ideas. Harry M. Marks, the historian, warns that the historical record of the clinical trial is incomplete and informs that Western European and American scientists influenced each other. He points out that among medical reformers, two parallel viewpoints were emerging: medicine can be a science if it rigorously applies research that is

subsequently channeled into clinical practice, and clinical medicine can be as scientific as laboratory sciences if scientific methods are applied to judging medical treatment.

Marks further subdivides the history of the clinical trials in the US into organizational and methodological reform. The first Food and Drug Act was adopted in 1906. The American Medical Association Council on Pharmacy and Chemistry was established in 1912. The standards of the AMA Council on Pharmacy and Chemistry set by expert opinion were adopted in the Food and Drug Act of 1938.

In 1958, the AMA Council on drugs authorized the publication of a paper by Walter Modell and Raymond Houde and 'approved the use of the double-blind method, controlled comparisons, randomization, (and) statistics…' (Shapiro and Shapiro 1997). By 1970 the Food and Drug Administration was empowered to monitor the safety and effectiveness of new drugs based on adequate, well-controlled investigations using appropriate statistical methods, the result of the reforms in methodology that influenced organizational reform.

In the US, Marks credits successive generations of challengers to the methodological status quo, nineteenth-century therapeutic nihilists, and also followers of Claude Bernard's rational therapeutics that stressed knowledge of etiology before treatment. The therapeutic reformers of the twentieth century included Torald Sollmann, George H. Simmons, Thomas C. Chalmers, and Morris Fishbein. Simmons, editor of *Journal of the American Medical Association*, attacked manufacturers for flooding the market with exaggerated claims for unproven nostrums. Fishbein, another editor of *JAMA*, 'unrelentingly described the fads and follies of healing cults and medical follies (Shapiro and Shapiro 1997). Chalmers urged in 1975 that the first patient put on any new drug be part of a clinical trial.

Charles S. Pierce and Joseph Jastrow are credited with the first use of randomization in 1885. Rivers in 1908 used the double-blind design and placebo (without using the terminology) in an experiment testing the influence of alcohol and other drugs on fatigue. He noted, 'The action of suggestion can never be excluded from any medical treatment' (Shapiro and Shapiro 1997). In 1917, John E. Coover first used the terms 'controlled experiment' and 'randomization' (Shapiro and Shapiro 1997), which only became standard through Sir Ronald Fisher's book, *The Design of Experiments* in 1935. The number of experimental studies increased steadily through the work of, among others, Louis Pasteur, Robert Koch, and Rudolph Virchow. Torald Sollmann, professor and chair of the Department of Pharmacology at the Medical School of Western Reserve University, headed for 55 years the AMA Council on Pharmacy and Chemistry. Sollmann urged that it was 'essential that clinical experimentation should follow the canons of other scientific experimentation' (Shapiro and Shapiro 1997). He coined the term 'comparative method.' Publications of the Council made reference to 'the blind test' and published what was probably the first study of the 'double-blind procedure.' In 1930, Sollmann recommended the use of the 'blind test,' used the term 'placebo' for the first time in research, and dropped the term 'comparative method.' But his news had little impact on practice, and the AMA Council on Pharmacy and Chemistry continued to rely on reviews by its member experts rather than on controlled studies.

The experts were researchers of unquestionable skill and integrity, but they rarely had experimental data and focused on confronting the tactics of manufacturers, their aggressiveness and exaggeration of claims, and on manufacturer-supplied laboratory evidence. Their influence led to a 'seemingly deluded faith in experts,' from 1900–1930 i.e., themselves and colleagues of the same ilk (Marks 1997).

A major milestone was a single-blind study in 1932, when Harry Gold tested the use of xanthine against placebo (lactose) for cardiac pain. He realized that physicians were asking the patients leading questions and prejudicing answers, and thereafter, tried to blind the physicians, as well. In 1935, Hediger and Gold compared two forms of ether in a 'blind test' and legitimized the use of placebos. Kwit summarized their insights: 'Towards the end of the study we realized that the doctor must not know what he gives, the patient…what he receives, and the questioner must not know what was given' (Shapiro and Shapiro 1997).

Like Sollmann earlier, Gold in the 1930s failed to have an immediate effect on medical practice. However, throughout the next two decades, researchers flocked to work with him and the new methodology, and he supervised many double-blind, placebo controlled studies. In a 1950s interview with Shapiro, Gold said, 'The placebo and the double-blind were companions after our 1937 study.'

A 1948 joint study by the US Public Health Service and the British Research Council is generally considered the first clinical trial. It tested the effectiveness of streptomycin in the treatment of tuberculosis. Patient selection was randomized. Evaluators of the patient X-rays did not know which patients had received streptomycin and which placebo. The treatment was not blind, because streptomycin-treated patients received up to four injections per day. Since the controls were bed-rest only patients, it would have been unethical to subject them to injections of placebo.

4. *New Paradigm I*

It was not until the 1950s, when antibiotics were discovered and mechanisms of several metabolic diseases understood that modern medicine began, and

clinical research could continue to make inroads. Through the 1960s and 1970s, the scientific method superseded authority and tradition. Treatments had to show sensitivity, specificity, and predictability through statistically sound techniques, randomization, the double-blind method, and placebo controls.

Current clinical standards for research are reflected in the US Preventive Services Task Force recommendations for treatment. Guidelines of 1989 (revised 1996), rank in descending order of reliability and ascending order of bias and misinterpretation randomized controlled trials, nonrandomized controlled trials, cohort studies, case control studies, comparisons between times and places, uncontrolled experiments, descriptive studies, and expert opinion.

5. *New Paradigm II*

The new scientific canon of clinical research encountered challenges and doubts almost as soon as it became established. Questions of method were raised about blindability in double-blind studies, as well as the precise definition, activity, and validity of placebo. Critical voices questioned ethical aspects of clinical trials that deprived half the patients of a medication that might be effective. The ensuing debate polarized ethicists and physicians.

6. *Blindability*

Extensive reviews of blindability indicated that studies took its validity for granted. It was suspect in the few double-blind trials that tried to carry out such tests of validity. Basoghu, in review of the literature found that only three of 23 double-blind psychopharmacology studies maintained blindability in patients, physicians, and assessors. Side effects, explanations in informed consent forms, past experience with similar drugs, and improvement with the active drug but not placebo sabotaged blindability. In his own comparison study of blindability of alprazolam, an anti-anxiety agent, with behavior therapy for panic disorder, Basoghu found that patients, treating therapists, and assessors were able to guess correctly, to a statistically significant extent, the patient's category. Patients treated with the active drug were more accurate than those on placebo, and active drug accuracy was highest at late stage of therapy, when optimal dosing and maximal efficacy had presumably been achieved.

Morin et al. (1995), recognizing that judgments carried out at the end of a study as to what was drug, what placebo could be influenced by treatment response to active drug, compared Temazepam, a hypnotic, and an inactive placebo. They carried out three assessments over 8 weeks. All patients received cognitive or behavioral therapy. Both patient and physician groups showed a statistically significant high degree of accuracy in guessing at 1 and 8 weeks with no better than chance guessing at week 4.

From these and other studies, it appeared that certainty ratings increased with time, that experienced patients showed greater accuracy in penetrating blinding, and that patients on active drugs guessed better than those on placebo. In summary, neither the patients nor the administering physicians were blinded to any significant degree.

7. *Placebo*

Harry Gold had accepted earlier twentieth-century definitions of placebo, such as Rivers' that placebo gives an inactive imitation of the substance being tested and Sollmann's that placebo furnishes the comparative check of the natural course of the disease, and did not explore the nature of placebo and placebo effects. Current theories on the nature of placebo and placebo effects include spontaneous remission due to biological diversity, conditioning, mobilization of endogenous endorphins, and the statistical phenomenon of regression to the mean. In 1955, H. K. Beecher calculated that 35.2 percent improvement in 15 clinical trials involving different diseases was due to placebo.

McDonald and Mazzuca (1983) concede the evidence that patients improve during placebo treatment, but point out that that is not proof that placebo treatment is the reason. The percentage of improvement observed in placebo treated patients is identical to that predicted from a statistical phenomenon known as regression to the mean where both objective and subjective measures of abnormalcy tend to improve and move towards the mean with successive measurements. McDonald and Mazzuca (1983) illustrate this in a review of 30 randomly selected placebo controlled clinical trials. Placebo effects and regression to the mean are both 'not uniform, constant, or predictable.' (Regression is by definition a random phenomenon.) McDonald and Mazzuca do not negate the usefulness of placebo. In comparisons of known drug with unknown, placebo controls can clarify ambiguous results. Placebo controls are also useful in conditions that fluctuate.

McDonald and Mazzuca note that placebo effects can be created through conditioning, citing the work of Stewart Wolf on the effects of conditioning and suggestion on measurable physiologic events in human beings.

8. *Placebo Amplification*

Richard Thomson believed that fewer studies would show a significant difference between placebo and drug if an active placebo were used as control. He reviewed 75 studies of tricyclic antidepressants com-

pared with active placebo (atropine) or with an inert placebo (lactose). Indeed, as postulated, antidepressant efficacy was ranked as superior in only one of the seven atropine/placebo studies available, while the tricyclic antidepressant was ranked superior in 43 of 68 studies controlled by an inert placebo. One explanation might be that atropine might have an antidepressant effect of its own. However, Thomson concluded that placebo amplification, a situation where physiological effects of taking a drug convince subjects already committed to treatment that the treatment is, in fact, effective, was the phenomenon being observed.

Roger P. Greenberg compared a tricyclic antidepressant with a new drug being tested, as well as an inactive placebo. He postulated that in a study focused on a new drug's effectiveness, there would be less interest in proving the efficacy of the older drug. Indeed, compared with claims in earlier literature, the older drug was much less effective.

Shapiro and Shapiro (1997) reviewed 27 studies of guessing of placebo and drug allocation and found that overall, both patients and physicians made correct identifications in 60 to 80 percent of tests. Eighty percent of drug guesses, but only 50 percent of placebo guesses were correct in their own study.

Fisher and Greenberg (1995) found in a review of world literature of studies of psychotropic drugs prior to 1993 that 'double-blind is vulnerable to penetration and hence, to bias,' placing all previous psychopharmacological research in question.

To try to rectify previous research on the efficacy of psychotropic drugs, Fisher and Greenberg suggest greater complexity in research design, the use of active placebo, various crossover designs, 'conventional drug' and 'drug disguised' groups, dose ranging, and the testing of drugs in opposing contexts, i.e., in an atmosphere of optimism or pessimism. They concede that current regulations preclude a drug disguise scheme for example, and agree that a completely foolproof design is not likely ever to be contrived. Personalities of patients and researchers were a factor in correct guessing of placebo or drug. While overall studies showed agreement among evaluators, rater differences were noted. Patient noncompliance was another confounding variable.

9. *Ethics*

As early as in the 1960s, ethical considerations about clinical trials involving randomization, double-blind methodology, and placebo controls were raised. Benevolent paternalism had been a prevalent ethical guideline since antiquity. At worst, it led to egregious exploitation of patients. In the late 1950s and 1960s, H. K. Beecher, in several watershed papers, cited instances of unethical and unscientific proceedings by researchers. He recommended that patients must give 'informed consent.' The US Public Health Service responded by proposing the establishment of Institutional Review Boards. The National Institutes of Health mandated Internal Review Boards, and the Food and Drug Administration, published in 1970, rules for 'adequate and well controlled clinical evaluation' (Shapiro and Shapiro 1997).

Beecher's papers helped move medicine from its tradition of paternalism to informed consent and patient autonomy. However, many in the medical establishment complained that the patient–physician dyad was being fractured by a mass of lawyers, judges, journalists, philosophers, and legislators. And that inevitably, bureaucracy was stifling the practice of medicine.

Ethicists had entered discussions in the 1970s, questioning placebos, the double-blind procedure, and the randomized clinical trial. Sissela Bok and Beth Simmons took strong stands opposing the use of placebos without informed consent. Others considered the use of placebos to be unethical in practice but permissible in research, to the point of not informing subjects of the research design. Some proposed that it was unethical not to carry out a placebo controlled, double-blind study. Still others concluded that since placebos have shown positive therapeutic effects using a placebo as control does not deprive patients of all treatment.

Questions have also been raised about how informed is informed consent in general, and especially if the subjects are cognitively or psychologically impaired? Attempts to resolve some of the issues are also being made. Where an effective treatment exists but a new one is being tried, the potency of the established treatment, as well as the severity of the patient's condition, are considered. As part of the experimental design, patients are moved from experimental or placebo cohorts to conventional treatment if their conditions deteriorate.

In 1978, while the rest of the Western scientific community moved towards research criteria of randomization, double-blind, and placebo controls in clinical trials, Germany passed laws substituting proof of efficacy through clinical open trials or expert opinion as qualifications for acceptance and registration of drugs. Controlled studies were no longer considered as essential to proof of efficacy.

In the introduction to the Physician's Desk Reference for Herbal Medicines, the editors say that in the absence of the 'rigors of FDA review' of these preparations, the editors are basing their guidelines on 'the closest available analog to FDA approved labeling—the findings of the German Regulatory Authority's watchdog agency commonly called Commission E'…its conclusions represent the best expert consensus…'

Statisticians contend that uncontrolled studies tend to favor a positive outcome.

10. Summary

Scientific thinking, defined by uncertainty that drives observation, hypothesis formation, and further search, was a late arrival to medicine, which was still laboring under few restrictions, guided by elaborate theories, mysticism, superstition, tradition, and authority.

Scientific thinking met much resistance in the medical establishment, but open label clinical trials, clinical trials using placebo controls, and double-blind methods began and continued into the twentieth century. In the 1950s when it became possible for the first time to cure people, statisticians became essential partners in the research team (Glantz 1997). Methodological and ethical questions began to be raised in subsequent decades about contemporary scientific criteria. Patient activism had also contributed to the erosion of paternalism that afflicted research design, as well as clinical practice.

The future may render clinical trials obsolete. We may well have a database atlas of the brain and body that will enable the researchers to calculate and theorize about effects without the use of human subjects.

The reader is referred to Marks (1997) and Shapiro and Shapiro (1997) for more complete historical surveys, detail, and documentation.

See also: Alzheimer's Disease: Antidementive Drugs; Antidepressant Drugs; Schizophrenia, Treatment of

Bibliography

Basogliu M 1997 Double-blindness procedures, rater blindness, and ratings outcome: Observations from a controlled trial. *Archives of General Psychiatry* **54**: 744–8

Beecher H K 1955 The powerful placebo. *Journal of American Medical Association* **159**: 1602–6

Bystritsky A, Waikar S V 1994 Inert placebo vs. active medication: Patient blindability in clinical pharmacological trials. *Journal of Nervous and Mental Disease* **182**: 485–7

Cohen A, Posner J 1995 *A Guide to Clinical Drug Research*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Fisher S, Greenberg R P, Riter J A 1995 Placebo washout is not a meaningful part of antidepressant drug trials. *Perpetual and Motor Skills* **81**: 688–90

Frank J D, Frank J B 1993 *Persuasion and Healing*. Johns Hopkins paperback edn. The Johns Hopkins University Press, Baltimore and London

Freston J W 1986 Dose ranging in clinic trials: Rationale and proposed use with placebo positive controls. *American Journal of Gastroenterology* **81**: 307–11

Glantz S A 1997 *Primer of Biostatistics*, 4th edn. McGraw Hill, New York

Kuhn T S 1996 *The Structure of Scientific Revolutions*, 3rd edn. The University of Chicago Press, Chicago

Marks H M 1997 *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990*. Cambridge University Press, Cambridge, MA

McDonald C J, Mazzuco S A 1983 How much of the placebo effect is really statistical regression? *Statistics in Medicine* **2**: 417–27

Morin C, Calecchi C, Brink D 1995 How 'blind' are double-blind placebo-controlled trials of benzodiazepine hypnotics? *Sleep* **18**: 240–5

Niklson I A 1997 Factors that influence the outcome of placebo-controlled antidepressant clinical trials. *Psychopharmacology Bulletin* **33**: 41–51

Report of the US Preventive Services Task Force 1996 *Guide to Clinical Preventive Services*, 2nd edn. Williams and Wilkins, Baltimore, MD

Shapiro A K, Shapiro E 1997 *The Powerful Placebo: From Ancient Priest to Modern Physician*. The Johns Hopkins University Press, Baltimore, MD

Thomson R 1982 Side effects and placebo amplification. *British Journal of Psychiatry* **140**: 64–8

E. Shapiro

Placebos, Use of: Ethical Aspects

A placebo is a therapy, or an intervention designed to simulate therapy, that has unspecified activity for the condition being treated. Placebo components exist in all treatments that produce an expectation of good effects, including competent and caring professional behavior of those delivering the therapy, and status or attractiveness of the institution. Placebos effectively relieve symptoms of most medical and psychological conditions. The underlying mechanisms are increasingly a topic of theory, speculation, and research. The ethics of using placebos depends on context, theory, and purpose. This article examines evolving theories and uses of placebos and associated ethical issues.

1. History

Before modern medicine, many substances thought to have curative powers, e.g., arsenic, produced feelings of wellness, before perhaps killing the patient. Such placebo effects, due to ignorance of the physician and patient, were primary active ingredients of ancient medical practice. By the mid-1800s, Koch, Pasteur, and others demonstrated that specific bacteria could cause specific diseases, giving rise to specific pharmacologic and surgical interventions replacing nonspecific healing therapies. Placebo effects became 'nuisance variables' to be controlled for, and ethics considered when it was acceptable to use placebos instead of accepted treatment modes. Finally, placebo effects are being studied as interesting phenomena in their own right.

Their therapeutic value is increasingly recognized, and the ethics of failing to use placebos to produce beneficial results is being questioned. Today's profes-

sionals produce placebo effects unwittingly, incidentally, or deliberately in research and practice in many fields including medicine, psychotherapy, and teaching. Today's ideal professional would appropriately and deliberately combine the power of modern science and technology with the healing power of placebos. Unfortunately, impersonal cost-cutting programs may inadvertently eliminate such effective strategies.

2. *Definitions and Uses*

The ethics of placebo use depends on which definition or theory of placebos is employed.

2.1 *The Pharmacological Paradigm*

The term placebo, from Latin *placere*, 'to please,' was coined to mean a pharmacologically inert substance, e.g., a sugar tablet, deliberately disguised as medication. Thus defined, the placebo is used to satisfy complaining or malingering patients; placebo effectiveness was until recently considered proof of patient psychopathology or feigned illness. Today's physician still employs placebos when the requested medication is inappropriate, but to satisfy concerns about lying to patients, may use very weak medications rather than inert substances.

Research on new treatments may yield 'success' simply as placebo effects. To control for this 'noise' researchers randomly assign both subjects and researchers to placebo and therapy groups. Double blind designs are used because researcher belief also influences results. The therapy is deemed successful if it yields better results than the placebo. (Without a placebo group, recovery due to treatment could not be distinguished from recovery due to spontaneous natural healing, as well as from expectancy effects.) To satisfy ethical requirements, researchers:

Inform subjects that they may receive placebos;

Specify the chance of receiving a placebo;

Use placebos only when genuinely uncertain about treatment effectiveness;

When inert substances might do harm, compare experimental treatment to baseline clinical data or current effective treatment;

Monitor subjects regularly for adverse effects and remove immediately from placebos if indicated;

Provide the placebo group with effective therapy later;

In double-blind studies, ensure that the code can be broken quickly to discover whether a placebo is causing a worsening condition;

Do not use placebos with highly vulnerable populations;

Use placebos resembling the active preparation so that patient and researcher cannot detect the difference.

Use of placebo control groups has, unfortunately, become such a hallmark of publishable research that placebo controls may be used when other approaches, e.g., comparison with current effective treatments, are as rigorous, often more appropriate, and put the subject's interests first. Placebo control groups in psychiatric research on such vulnerable populations as persons suffering severe depression or schizophrenia have come under considerable scrutiny. Full comprehension and consent of the patient and next-of-kin concerning the possible use of a placebo, and careful surveillance of the patient, are crucial to ethical use of placebos when suicide or other forms of harm are plausible risks.

The pharmacological paradigm ignores sociocultural and interpersonal variables, expectations, learning, neurophysiology, and real therapeutic value of placebos. The malingering patient notion implies personality or demographic characteristics of placebo responders, but such characteristics have not been found.

2.2 *Remembered Wellness*

Newer conceptions and uses of placebos are highlighted in Benson and Friedman's (1996) article. Under the right circumstances anyone might respond to a placebo. Placebos are not absence of treatment. What persons believe can diminish pain and objective symptoms of disease. The components of 'remembered wellness'—the patient's and physician's positive beliefs and expectations, and a good relationship between the two parties—are one of medicine's most powerful assets. Impersonal medical care and drug or surgical interventions are not effective in treating most diseases. Most visits to physicians stem from conditions related to stress and other psychosocial factors (Kroenke and Mangelsdorff 1989) which are responsive to placebo effects.

Such placebo effects based on 'remembered wellness,' accompanying standard therapies comprise effective care. For example, in an experiment using five standard therapies for angina pectoris, patients who believed in the therapy experienced about 80 percent relief from chest pain, reduced nitroglycerine usage, improved electrocardiogram, and increased exercise tolerance; skeptical patients experienced about half as much therapeutic benefit (Benson and McCallie 1977). In Canner (1980) among patients given placebos, less compliant patients had a 28 percent mortality rate after five years, while compliant patients had a 15 percent mortality rate.

Placebo use has evolved to entail warm, caring, sympathetic relationships that foster positive beliefs, combined with appropriate standard specific therapies. Thus defined, placebos are devoid of deception, and integral to quality of life, health, sound professional conduct, and empowerment and self-help.

Under this conceptualization, it is reductionist scientific medicine coupled with impersonal managed care that is ethically questionable.

3. *Theories of Why Placebos Work*

Three fundamental theories of placebo mechanisms are: conditioning (learning), hope and expectancy (cognition), and neurobiology (subjective experience). These three mechanisms are not mutually exclusive. Each provides ethical insights into placebo use.

3.1 *Learning*

Classical conditioning involves pairing an unconditioned stimulus (UCS), e.g., a charismatic, enthusiastic teacher who emphasizes the excitement and importance of a topic, with a neutral stimulus (a not-particularly exciting curriculum) which becomes the conditioned stimulus (CS) and evokes enthusiasm and dedication in the student. Pairing morphine (UCS) given to patients in great pain, causes the nurse and syringe to become conditioned stimuli (CS) for pain reduction. Placebo effects persist when only the CS is present and, following principles of partial reinforcement, continue undiminished if the UCS is reintroduced periodically, but eventually extinguish if the UCS is never reintroduced. For example, if a painkiller (UCS) having severe side-effects is interspersed with a placebo in an identical capsule (CS) and intermittent reintroduction of the UCS, the effect may be a continuously effective analgesic with diminished side-effects. There is a powerful relationship between magnitude of UCS and response to the CS. For example, a massive dose of morphine (UCS) will result in a massive placebo (CS) effect. These principles of conditioned analgesia have been found with animals as well as humans.

Ethics requires disclosure of use of conditioning; e.g., following explanation of learning theory, the doctor might state that the UCS would be intermittent and that the client's own relevant powers will serve to reduce pain. Conditioning produces expectancy, and memory of UCS effectiveness, as well as measurable neurological change.

3.2 *Desire and Expectancy*

Need or desire (e.g., for relief from pain) and expectation that a given therapy is effective will mediate placebo analgesia. These cognitive factors become associated with the CS and extend and support the role of learning. Authoritative professionals can create powerful hopes and expectations which can produce a placebo response without prior exposure to an objective UCS, though such exposure (combining conditioning with expectancy) increases expectancy. Similarly, expectancy can result from learning about an agent's therapeutic value, especially when one strongly desires that effect. Thus, persons desperate for relief are highly vulnerable to quackery, fraud, superstition, or cults, which may produce significant temporary relief through placebo effects. Conversely, anything expected to cause distress, sickness, or death (e.g., feared surgery, awkward physicians, institutions with poor reputations) can have a *nocebo* effect (Hahn 1997). Culturally based expectations, and good or bad relevant prior information or experiences, as well as aspects of the current therapeutic situation will create expectations and have placebo or nocebo effects. Positive expectations result in far superior outcomes of standard mediation surgery, education, and other interventions. The ethical professional uses expectancy effects to good ends, not to deceive people.

3.3 *Neurobiology*

If placebo effects are solely products of imagination, the ethics of placebo use concerns fooling people. However, as scientists discover how subjective 'mind' acts on the physical body, the ethics of placebo use changes. The functioning of neurons is involved in mental representation (e.g., the self, an image); subjective experience (e.g., pain, feeling of wellness); and bodily change (e.g., hormone secretion, healing); and the neural network, itself, is an objective physical entity. Thus, for example, injury causes pain, morphine reduces that pain, and placebos also alter neural representation of the injury (pain) in proportion to the preceding amount of morphine. Research has demonstrated neural substrates of placebo effects, especially analgesic effects, and imaging technology holds promise of rapid advancement in this area (see Fields and Price 1997).

Neurological models of placebo effects provide a framework for research on the role of language, culture, and environmental context in helping to reduce stress and pain, reversing disease processes, and healing injury. Future discoveries in this domain may lead to rewriting the ethics of placebo use.

4. *Summary*

Placebos have evolved from deception, and 'noise' that researchers control for, to phenomena studied in their own right. Research has shown that what persons think can affect objective (structurally or functionally measurable) changes in physiological functioning. Further theory and methodological developments promise the possibility of effective self-administered placebos—a new kind of self-help, and of radically new ethical considerations regarding placebo use.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Clinical Treat-

ment Outcome Research: Control and Comparison Groups; Ethics for Biomedical Research Involving Humans: International Codes; Medical Experiments: Ethical Aspects; Placebo Studies (Double-blind Studies); Placebos, Use of: Ethical Aspects

Bibliography

Benson H, Friedman R 1996 Harnessing the power of the placebo effect and renaming it 'remembered wellness.' *Annual Review of Medicine* **47**: 193–9

Benson H, McCallie D P 1977 Angina pectoris and the placebo effect. *New England Journal of Medicine* **300**: 1424–9

Canner P L 1980 Influence of adherence to treatment and response of cholesterol and mortality in the coronary drug project. *New England Journal of Medicine* **303**: 1038–41

Fields H L, Price D D 1997 Toward a neurobiology of placebo analgesia. In: Harrington A (ed.) *The Placebo Effect: An Interdisciplinary Exploration*. Harvard University Press, Cambridge, MA, pp. 93–116

Fish J M 1973 *Placebo Therapy*. Jossey-Bass, San Francisco, CA

Frankenhaeuser M, Jearpe G, Svan H, Wrangsjea B 1963 Psychophysiological reactions to two different placebo treatments. *Scandinavian Journal of Psychology* **4**: 245–50

Hahn R A 1997 The nocebo phenomenon: Scope and Foundations. In: Harrington A (ed.) *The Placebo Effect: An Interdisciplinary Exploration*. Harvard University Press, Cambridge, MA, pp. 56–76

Harrington A 1997 *The Placebo Effect: An Interdisciplinary Exploration*. Harvard University Press, Cambridge, MA, pp. 56–76, 93–250

Kroenke K, Mangelsdorff A D 1989 Common symptoms in ambulatory care: incidence, evaluation, therapy, and outcome. *American Journal of Medicine* **86**: 262–6

Peck C, Coleman G 1991 Implications of placebo theory for clinical research and practice in pain management. *Theoretical Medicine* **12**: 247–70

Rossi E L 1993 *The Psychobiology of Mind-Body Healing*. Norton, New York

Smith T 1999 *Ethics in Medical Research*. Cambridge University Press, Cambridge, UK

Spiro H M 1986 *Doctors, Patients, and Placebos*. Yale University Press, New Haven, CT

Wickramasekera I E 1988 *Clinical Behavioral Medicine*. Plenum, New York, Chap. 5

J. E. Sieber

Plagues and Diseases in History

Disease is the sting of death and death is the necessary victory of the grave. Disease is a mechanism of adaptation—that is, of attack, equilibrium, and defense—of the living being to its habitat. Human development and evolution, in its social, cultural, and ecological setting, give rise to alterations in the human organism, both in body and mind, against which it must struggle. The evolution of the species, society, and culture takes place within a given environment and this environment may alter the pathological spectrum. If according to Bichat life is resistance to death, it may be said that life is resistance to disease. The human being seeks Hippocratic and Aristotelian equilibrium but, in doing so, he or she is frequently subjected to changes and movements that endanger this quest.

These changes, quite frequently depend on the human being, on his or her customs, social groupings, or on settlements that lead him or her to avoid malaria by settling in mosquito free zones or to control smallpox by vaccination. On other occasions, such changes do not depend on the human being but arise in the environment or within the diseases themselves that may appear or disappear of their own accord, as was the case with the 'English sweate.' On yet other occasions, these changes are merely due to cultural appreciations, such as is the case where certain sexual practices such as masturbation or homosexuality are stigmatized as diseases. On the other hand, disease is not always considered harmful: thus it may be considered a distinction of the gods, as was the case of epilepsy in the time of the classic authors. This same disease was associated with the devil by the Christians, and historically, it had always been a marked distinction of great personages from Caesar to Napoleon. Disease can also be considered a way towards perfection or transformation, leading to the creation of art or the salvation of the soul. It may also have political repercussions: it is claimed that George III's madness gave rise to problems for the British crown while Roosevelt's fragile health was considered to have placed him in a weak position in the Yalta negotiations. Socially, disease has sometimes been considered the stigma of groups or races. Medicine and politics have endeavored to preempt social changes by means of engenics which began with recommendations or orders to avoid marriages entailing risks (real or imaginary) and went on to such processes as sterilization and abortion. Hereditary problems are thus intertwined with cultural, social, religious, and moral ones. In contrast with the stark determination of former times, there is today an attempt to draw new optimism from the possibilities of genetic modification. Although, in this latter respect, there is fear of a future where such a formidable force could rest in the hands of the rich and powerful. Thus, individual disease or illness is rooted in society, being tied up with the history of great endemics and epidemics.

1. Ancient and Middle Times

Paleopathology enables us to discover the high incidence of diseases in the early inhabitants of the

Earth, especially in the cases where such diseases left their mark on the bone. It is thus possible to detect remnants of tumors, infections, and osteoporosis as also traumatisms, dental alterations, rheumatism, and osteopathy. From the first settlements in the Fertile Crescent, changes have taken place in the relationship of man and animals with their environment and it is possible to establish an extensive catalogue of diseases which largely subsist down to this very day. Particularly significant are the forms of settlement, the crops, and the contact of man with cattle, pets, and parasites. Thus in the fourth and third centuries BC, malaria had already set in the early civilizations in marshy areas where the water and the crops provided the conditions for the mosquito to infect man. Evidence for the existence of seasonal and intermittent fevers is already to be found in Chinese, Indian, Sumerian, Assyrian, Babylonian, and Egyptian cultures. The link between splenomegaly and fever is mentioned in the Papyrus of Ebers, dating from 1570 BC. The connection between fever and marshy areas appears in the Hippocratic text *On Airs, Waters and Places*. Deforestation, agriculture, temperature, and a greater population density, as well as military and commercial movements favor the spread of the disease. In the third century BC in Italy, the disease becomes acute and in the first century BC Varron coincides with Lucretius in pointing to small animals as the cause of the disease and recommended that these be either abandoned or sold and that people should dwell on high ground, this latter being traditional advice and here also defense reasons counted. The spread of malaria got worse in Rome owing to negligence and flooding and it reached its zenith at the time of the fall of the empire, a cataclysm to which it may have contributed.

The bubonic plague was endemic in the Near East in pre-Classic times, the buboes being described by Hippocrates. In 430 BC, Thucydides describes the plague of Athens in the war against Sparta, but it is doubtful that this case refers to the bubonic plague. The first reliable description of this disease is that given by Rufus of Ephesus in the first to the second century AD, in the epoch of Trajanus; and in the sixth century, the plague of Justinianus initiates the descriptions of the first great epidemic cycles. Up until the Christian era, there is equilibrium between human beings, cities, and parasites but this equilibrium breaks down with the advent of the convoys that bring the two healthiest extremes into contact, namely, China and Rome. India and the Middle East continue to be dangerous and the situation at the extremes further deteriorates. Soldiers and merchants, animals and merchandise, slaves and prisoners all contribute to linking up the extremes of the known world. A consequence of this contact was the Antonine Plague in the second century. This plague was attributed to smallpox, something that had already existed as far back as the year 3000 BC, as certain mummies show.

In the Middle Ages, leprosy spread greatly. This biblical and present-day disease is accompanied by malignant connotations since dirtiness and overcrowding provided the conditions for the spread of the otherwise not extremely contagious *Mycobacterium leprae*. In this respect, the Route to Santiago de Compostella where pilgrims slept in groups and under conditions that left a lot to be desired provided a fertile ground for this disease. Another markedly religious character for the spread of the disease is to be found in the 'Sacred Fire,' 'St Anthony's Fire,' or ergotism, during which rye ergot was eaten. These diseases are linked to many others that owed their spread to squalor and poverty, for example, scabies and lice, consumption and fevers, anthrax and ophthalmia. In the Mediterranean bowl, malaria continued to be endemic with the population tending to resort to residing on the hills and mountains, far from marshy lands. The plague set in the incipient cities with the effects of rats and lice, making this disease an explicative example par excellence of epidemic history; see in this respect, Zinsser (1934). A second epidemic cycle commences with the Black Death in 1348. At this time, Venice establishes control over persons and ships. Commerce and the cities collapse and thus the hegemony of Castille over Aragon takes place, giving rise to the origin of the modern European Kingdoms and empires.

2. *Modern Times*

European expansion meant a sudden traffic of parasites towards the new worlds. The discovery of America, with its scattered population devoted to agriculture, where there were no dirty cities and where the inhabitants were devoid of immunity to the diseases of old Europe, produced an enormous demographic crisis. Certain diseases, such as smallpox ravaged the indigenous population, first in South and later in North America. Syphilis has supposedly come from America, brought back by the Spaniards and it caused havoc in Renaissance Europe. Sea voyages led to the development of scurvy, due to the lack of fresh foods. Other diseases also came from overseas, such as yellow fever, although it originated from Africa and later from the tropics, due to the commerce of slaves and merchandise. In the seventeenth century, it took root in Brazil and is still present today in the Amazon. Soon this disease spread to the Caribbean and throughout America, reaching New York and Boston and likewise Colombia, Ecuador, and Peru. It affected warm America and Europe, especially the Iberian Peninsula. Immunity and cut offs in commerce stalled the disease and today it remains endemic in tropical areas of America and Africa.

The plague arrested demographic growth in Europe since the enimity between states and trade impeded protection against the disease. The last plague of Marseille in 1720 and the relative peace of the

eighteenth century allowed all European nations to set up sea and land defenses with Austria becoming a solid bulwark in the face of the Ottoman Empire, thus sealing off all possible spread of the plague. Towards the middle of the eighteenth century, the third wave of the disease broke out and with the exception of Europe, it spread to all countries including Asia, Africa, and paradises like America and Australia, leaving remnants in those places. The eradication of the plague in Europe was a result of the advent of public hygiene, immunity to the disease, and the disappearance of rats and the old and dirty wooden buildings. The fire of London 'purified' the city, since the rebuilding was carried out in stone, eliminating animals and dirt.

The disappearance in Europe of the plague was followed by the appearance of cholera that had been endemic in India for 2000 years. This disease was described by Western travelers in the sixteenth century and its spread to Europe and America was a consequence of pilgrimages, trade, and a lack of cleanliness in water. A great pandemic broke out in 1817 and spread through Eastern Africa and Southern Asia and in a second wave it spread through Kabul and Persia reaching Russia and Poland in 1830. It then swept across Europe and had reached America by 1832. Over two centuries, terrible waves spread from the East and as late as the 1970s, the seventh outbreak of pandemic reached Europe. However, the timely and necessary cleansing of urban water supplies helped to stall its advance progressively. Nevertheless, as late as 1892, the free Hanseatic city of Hamburg suffered an outbreak of cholera epidemic while the neighboring Altona, governed by the new Prussian Reich, was able to avoid the disease, thanks to the successful filtering of its water. Cities and governments advanced in hygiene and they attempted to bring about healthier cities by means of appropriate public health measures.

Together with the major epidemics, death was continuously caused by two groups of endemic diseases. Two seasonal cycles of disease remained rooted into modern times: one in summer affecting children and due to enterocolitis, lactation problems, dysentery, and typhus, and another in winter affecting the old and due to typhus and pneumonia. The nineteenth century sees the beginning of the demographic revolution in Europe with a sharp decline in mortality, especially maternal and infant mortality. Diseases like diphtheria, chicken pox, scarlet fever, measles, and smallpox are combated with serum and later on with vaccination. Mother care also advanced both throughout pregnancy as also at childbirth and during lactation. Tertian fevers continue to be active and are combated in the Mediterranean area by clearing up the marshy areas, avoiding the mosquito by mechanical means, or by the use of cinchona bark (Peruvian bark), and later on of quinine.

The rise of great and industrial cities with low-grade outskirts and the emigration of peasants to the city favored the spread of maladies associated with poverty, squalor, lack of appropriate food, and exposure of people devoid of defenses. Some of the diseases are strictly occupational which is the case of miners and textile workers who were exposed to injury and to the industrial poisons that affected the proletariat. Other diseases are closely related to the harsh conditions of life with the spread of typhus and pulmonary diseases, especially tuberculosis, being due to this factor. Tuberculosis spread over the turn between the eighteenth and nineteenth centuries and while it was known as the disease in fashion, affecting notable and distinguished people, it struck the proletariat much more severely. Today, tuberculosis is again a serious threat. The same can be claimed regarding the enormous spread of syphilis as also with respect to drug abuse, beginning with alcohol and continuing with marihuana, cocaine, and morphine. At the middle of nineteenth century the opium trade was legalized in China, after the victory of the British army.

3. *Contemporary Times*

This last century marked the descent of the high mortality rates in countries that reached high standards of sanitary development, that is countries that invested in public health care, such as Britain or the USSR. Such advances have been due to public health services, hospitals, antibiotics, surgery, and vaccination and they have been reinforced by the developments in immunology, pharmacology, and biotechnology and with the promising future of improved genetic material. Smallpox is the first disease to be considered totally eradicated with only some samples of the virus being kept at a few laboratories for study purposes. But other viruses have taken its place. Thus the old influenza caused three heavy outbreaks of pandemic in 1918–9, 1957–8, and 1968: these were known respectively as the 'Spanish 'flu,' which originated in the USA, the 'Asian 'flu,' and the 'Hong Kong 'flu.' Poliomyelitis originated in the USA, affecting non-immune children and this disease has only been controlled in developed countries by means of vaccination. AIDS has meant a development in virus pathology and this disease has associations with the exploitation of poor countries, drug trading, and also an elite sector that has sprung rapid research in the field. The world commotion surrounding this disease has served both to reveal human altruism on the one hand, but, on the other, contempt towards those infected by the disease, since AIDS is becoming more and more the lot of the poor, such as large areas of Africa or southern Asia and of downcast or marginal groups such as the chronically ill, drug addicts, prostitutes, and homosexuals. Nevertheless, heterosexual transmission is advancing and is placing the future of all Africa in danger.

Forms of illness are changing. Development has managed to stamp out old forms of infectious morb-

idity. On the other hand, the increase in life expectation favors the development of Alzheimer and cancer. Cardiovascular accidents and traumatism, traffic and sports, factories and radiation are something that the twenty-first century will inherit. Ghettos and unemployment, and some of the old drugs such as heroin and tobacco or new synthetic ones are causing havoc even to the rich world. But in underdeveloped countries, the old morbidity continues to exist, giving rise to very low life expectation. Death is most rampant amid the least protected, especially women, children and the elderly, with the wars causing high death rates amongst men. Under such circumstances, diseases caused by deprivation and infection continue to be the most rampant.

Malaria has been eradicated from Europe thanks to drainage, improvements in crop cultivation, mechanical means, quinine, and insecticides. Today, faith is also placed on vaccination and the genetic modification of the mosquito. Yet malaria continues to be endemic in warm zones of America, Asia, and Africa. In the interwar years, the Rockefeller Foundation sponsored a campaign against the disease. In 1955, the World Health Assembly, meeting in Mexico, warned against the resistance of the mosquitoes to insecticides. Together with the WHO and the UNICEF, the US government was financing the eradication program of the Pan-American Health Organization. Today, an effort is being made to interest governments in this disease, a disease that represents a danger to 40 percent of the population. In Africa the situation is very serious due to the increase in agriculture and irrigation as well as emigration, the political and economic problems and the bad sanitary conditions following in the wake of housing expansion and deforestation. The WHO, through the 51st World Health Assembly of 1998, set up the program 'Roll Back Malaria' and, with aid from UNICEF and the World Bank, is endeavoring to bring about economic and sanitary improvements in developing countries.

Cholera continues to be endemic in Asia, between India and the Far East, and also in warm zones in America and Africa. *Yersinia* (or *Pasteurella*) *pestis*, discovered by Yersin and Kitasato, is still to be found throughout the world, except in Western Europe, carried by rodents, rabbits, and related species. Bacteria are carried by black and brown house rats, *Rattus rattus*: rat fleas are the arthropod vectors transmitting 'epizootic' plague to humans from rodent hosts, but transmission among humans is also possible. Short-term human immunity requires frequent vaccination. Almost forgotten in the civilized world, there is nevertheless always the possibility that the devil may once more send his rats to the old, rich cities.

See also: Health: Anthropological Aspects; Middle Ages, The; Mortality, Epidemiological, and Health Transitions; Mortality: The Great Historical Decline

Bibliography

Ballester R 1977 *La historia clínica pediátrica durante el siglo XIX*. Universidad Zaragoza, Spain

Biraben J N 1975–6 *Les hommes et la peste*. Mouton, Paris

Campillo C 1993–4 *Paleopatología*. Uriach, Barcelona, Spain

Corbellini G, Merzagora L 1998 *La malaria tra passato e presente*. Università La Sapienza, Università di Cassino, Rome

Crosby A W 1972 *The Columbian Exchange*. Greenwood Press, Westport, CT

Engelhardt H T 1974 Explanatory models in medicine: Facts, theories, and values. *Texas Reports on Biology and Medicine* **32**: 225–39

Fischer-Homberger E 1979 *Krankheit Frau*. Hans Huber, Bern, Switzerland

Kiple K F (ed.) 1993 *The Cambridge World History of Human Disease*. Cambridge University Press, Cambridge, UK

Laín Entralgo P 1961 *Enfermedad y pecado*. Toray, Barcelona, Spain

McNeill W H 1977 *Plagues and People*. Oxford University Press, Oxford, UK

Peset J L 1993 On the history of medical causality. In: Delkeskamp-Hayes C, Gardell Cutter M A (eds.) *Science, Technology, and the Art of Medicine*. Kluwer Academic, Dordrecht, The Netherlands, pp. 57–74

Peset M, Peset J L 1972 *Muerte en España*. Seminarios y Ediciones, Madrid, Spain

Zinsser H 1934 *Rats, Lice, and History*. Bantam Books, London

J. L. Peset

Planetary Cartography

Planetary cartography is the making or study of maps of the planets, satellites, asteroids, and comet nuclei comprising our solar system. Maps of the Sun and other stars may be included for the sake of completeness, and future astronomers may prepare maps of planets orbiting other stars. The definition of 'map' is often biased towards modern terrestrial topographic maps, so for this article a broad definition is adopted similar to that of Harley and Woodward (1987): a map is a graphic representation designed to facilitate a spatial understanding of the object or region it represents. How many worlds have been mapped as of mid-1999? Several thousand maps of the Moon and Mars have been published, but maps of Pluto are limited to simple illustrations in journals. Asteroid 29 Amphitrite is portrayed in one simple sketch map, and the only published map of asteroid 1 Ceres is an illustration in a science fiction magazine. If Amphitrite is included but Ceres is not, 60 bodies in the solar system have been mapped in a scientific sense. When should a world be said to be mapped? It is accepted for this article that composites of multiple observations or images augmented with feature names or other in-

formation can be called maps, but not primary observational records (photographs or drawings).

The history of cartography can be portrayed as a linear evolution driven by successive developments in theory, technology and exploration, with modern scientific mapping as its zenith. Planetary cartography is particularly susceptible to this approach. Recent work (e.g., Harley and Woodward 1987) recognizes maps once dismissed as irrelevant and broadens the field to consider uses or aspects of maps overlooked in purely technical histories. Here both approaches are combined, useful references identified and areas indicated that will repay further study, for no comprehensive survey of this subject has yet been undertaken.

1. Background

The International Astronomical Union (IAU) oversees body and surface feature nomenclature, the definition of north and the locations of prime meridians, and publishes its decisions in its *Transactions*. Surface feature locations are determined by establishing a control network of points measured in multiple images of a world. Greeley and Batson (1990) summarize these technical issues. Many mapped bodies are near-spherical, but where they are not their shapes are estimated from stereoscopic images and the changing outlines of the object in views from different directions.

2. Pretelescopic Planetary Cartography

Markings on the Moon have always been visible, so representations of them are the earliest planetary maps. Drawings by Leonardo da Vinci dated about 1505 were long considered the oldest, but recently other ancient candidates have been suggested. Carvings at Knowth, Ireland (about 2800 BC) may represent the lunar markings, and one at the end of an interior passage might be illuminated at times by the rising Moon. The final form of Stonehenge, a circle containing a double horseshoe, may be a schematic map of the Moon akin to those at Knowth. Its solar alignment is most celebrated, but the full Moon occasionally rises on the monument's axis. Some Egyptian depictions of the sacred eye representing the Moon resemble the lunar markings. Plutarch gives several lunar place names. Some Arab world maps around AD 1100 resemble the lunar markings in reverse, recalling the ancient belief that the Moon is a mirror reflecting an image of the Earth. Van Eyck painted the Moon with markings in 1430, though they may be merely schematic. William Gilbert drew a lunar map with place names in about 1600, just before the invention of the telescope. Of uncertain age, not ancient but probably drawing on a long tradition, are representations of the *maria* in native American art, including rock art in the Stein Valley, British Columbia and drawings of the Barasana culture, Colombia. The only other body which could be mapped without a telescope is the Sun, since sunspots had long been known in China, but no such map has ever been described.

3. Telescopic Planetary Cartography

Following the pioneering observations of Harriott and Galileo in 1609, lunar mapping advanced rapidly until the 1800s (Kopal and Carder 1974, Whitaker 1989). Successive cartographers sought mainly to improve accuracy and detail, since the Moon is unchanging. Riccioli's nomenclature became standard and was formalized by the IAU in the 1930s. Photographs eventually replaced more stylized drawings, culminating in several superb atlases produced under the direction of Gerard Kuiper (University of Arizona) at the dawn of the space age. Frequently overlooked in histories of lunar cartography are maps prepared by and for nonastronomers, such as Homann's 1720 map comparing the nomenclatures of Hevelius and Riccioli, and numerous small charts shown as insets in celestial or world maps from roughly 1650 to 1800. Some atlases of the 1800s also contain lunar maps. Though rarely more than poor copies of astronomers' map, they may have had a wider audience than their models. No systematic study of them has yet been made.

Other worlds were more difficult targets for telescopic observers (Greeley and Batson 1990). The first map of Venus, by Bianchini in 1728, misrepresented faint markings in the planet's opaque clouds as lands and seas. Similar maps were being drawn until the 1950s. Herschel first mapped Mars in 1783. Schiaparelli in 1877 measured feature locations for positional control, described the spurious 'canals' which influenced public opinion for decades, and introduced the modern feature nomenclature. The markings of Mars are variable, so maps are made whenever it is well placed for observation (Blunck 1982). Maps of Mercury, beginning with Schiaparelli's of 1889, were compromised by the assumption of an erroneous rotation period that was not corrected until the 1960s. Markings on the four largest satellites of Jupiter are at the limit of telescopic discernability, but a map was made by Douglass in 1897, and others followed in the 1900s. Ephemeral cloud features in Jupiter's atmosphere are plotted on what are essentially meteorological maps to monitor changes. Similarly, sunspots and other ephemeral features of the solar photosphere are mapped to record the Sun's changing appearance. These 10 bodies, excluding Earth, were all that had been mapped before the Space Age.

4. Space Age Planetary Cartography

Spacecraft have visited many worlds since 1959, and like terrestrial explorers these planetary missions generate many maps. Some are used for premission planning, some for operational use during the mission, and others provide a systematic record of discoveries made or activities undertaken. The subject has not yet been described in detail and will form fertile ground for future scholarship.

4.1 Cartography for Lunar Exploration

The Soviet probe Luna 3 photographed the Moon's far side for the first time in 1959. Later missions filled remaining gaps in coverage and revealed greater detail, permitting large scale mapping of some areas. Landing craft photographed their surroundings, which were mapped at large scales. The US Geological Survey (USGS), Army Map Service, and Aeronautical Chart and Information Center produced over 1,000 maps for the Apollo missions (Kopal and Carder 1974). Astronauts used maps for site selection, training, orbital navigation, and surface activities. In the Soviet Union, MIIGAiK and other agencies produced maps for their lunar missions, including robotic landers and rovers, but never undertook a systematic large-scale mapping program. Lunar mapping in both nations almost ceased in the mid-1970s, leaving a legacy of 1:250,000 and 1:1,000,000 scale maps as well as global and regional geological maps. Public interest in Apollo resulted in numerous popular maps, atlases, and globes in the 1960s, and for a time many general world atlases included a lunar chart. Most disappeared as interest in lunar exploration waned.

4.2 Cartography for Solar System Exploration

Following Apollo, robotic spacecraft travelled to other planets, all but Pluto having been visited by 1989. In the 1990s several asteroids were explored and other bodies were observed in greater detail. Almost every mission resulted in maps, from systematic geological and topographic mapping by USGS to large-scale charts of the immediate surroundings of landers on Venus or Mars. The Soviet Union attempted no systematic mapping, except at Venus, but its landers gave the only surface views of Venus yet obtained. Mapping by USGS up to about 1994 is summarized by Greeley and Batson (1990). Forty-two worlds had been mapped from planetary spacecraft data by 1999, 20 by USGS. One significant problem concerns small irregularly shaped bodies for which geodetic methods and map projections are still being developed. Many planetary maps since 1990 take the form of digital photomosaics and maps of radar reflectivity, elevation, etc. which may be distributed on CD-ROM or via the World Wide Web, reprojected to any desired map projection and used in Geographic Information Systems. USGS has taken the lead in this, and many such products are now available for free download via NASA's Planetary Data System. The number of maps published on paper is declining, which will compromise historical studies. Public interest in Mars and the possibility of human expeditions before about 2020 led to the publication of several popular Mars maps in the late 1990s. These may be said to include maps of the Cydonia region of Mars illustrating unconventional concepts regarding the artificiality of features such as the famous 'face,' showing place names, alignments, and architectural or geological interpretations.

4.3 Telescopic Mapping in the Space Age

New instruments and techniques have enabled mapping of many worlds not visited by spacecraft, including Pluto and the asteroid 4 Vesta. Radar images of Earth-approaching asteroids have been used to derive shape models and maps, and comet nuclei have been mapped very tentatively by observing changing patterns of material ejected from their surfaces. These maps usually appear in scientific literature and are rarely seen by a nontechnical audience. Numerous stars should be included among the mapped celestial bodies, since methods have been devised for interpreting features in their spectra in ways that lead to simple maps of phenomena in their photospheres. Interferometers operating at visible wavelengths may eventually permit the mapping of surface features on planets orbiting nearby stars. Even in this solar system, the number of mapped worlds will exceed 100 not far into the twenty-first century.

5. Science Fiction Mapping

Some science fiction books and fan magazines contain maps reflecting both the stories themselves and contemporary knowledge of their settings. Like other popular maps they help shape public awareness of the planets. Post (1979) illustrates some examples, but the subject would repay greater attention. Here are considered only fictional maps of real worlds, Mars being the most prominent example. Fictional maps once showed canals and ruined cities, but now often identify lowlands which become seas on a future Mars made more earthlike by colonists. Notable among works illustrated in this way are the 'Barsoom' (Mars) novels of Edgar Rice Burroughs and magazines relating to his work, and the novels of Kim Stanley Robinson. A different class of planetary map has evolved to support role-playing games. These maps frequently include a mixture of fictional and real surface features, and have appeared in atlas and sheet map form as well as illustrations in game manuals.

6. Conclusion

Planetary cartography is a broad and rapidly evolving field with a long history. It has not yet been examined in its entirety. Maps record both the history of astronomy and space exploration and a much broader cultural response to those endeavours. A systematic effort to collect and describe this material is needed, including an attempt to preserve the increasing number of ephemeral maps on the Internet.

Bibliography

Blunck J 1982 *Mars and its Satellites*, 2nd rev. edn. Exposition Press, Smithtown, NY

Greeley R, Batson R M 1990 *Planetary Mapping*. Cambridge University Press, Cambridge, UK

Harley J B, Woodward D 1987 *The History of Cartography*. University of Chicago Press, Chicago, Vol. 1

Kopal Z, Carder R W 1974 *Mapping of the Moon*. Reidel Publishing, Dordrecht, Holland

Post J B 1979 *An Atlas of Fantasy*, new rev. edn. Souvenir Press, London

Whitaker E A 1989 Selenography in the seventeenth century. In: Taton R, Wilson C (eds.) *The General History of Astronomy*. Cambridge University Press, Cambridge, UK, Vol. 2

P. Stooke

Planning, Administrative Organization of

This article is about the role of planning in organizations, with a focus on urban planning in local government. It takes the view that it is best to analyze the general before focusing on the particular. To this end, a generic model is outlined, which sets the stage for the later discussion. The article concludes with alternative strategy options for coping with some unique aspects of the urban planning situation.

1. The Nature of Organizations

1.1 Organizations and Management

A human organization is one form of an abstract concept called a system. A system is an assembly of parts that form a whole greater than the sum of the parts. Every system has some unifying principle or force that holds it together. In the case of a human organization, this unifying element is provided by the management of the organization. Management in general refers to those people in an organization who are authorized to give direction to the other members of the organization. Understanding management is the key to understanding any organization.

The field of management is large and complex, and many theories compete for attention. However, most theorists agree, that some structure of power is a necessary part of management. The wellsprings of human motivation remain somewhat mysterious, and often people will cooperate voluntarily without coercion, but some sort of power to influence behavior, either by promise of reward or threat of reprisal, seems essential to the success of any organization. Without it, managers cannot be sure their leadership will be followed.

The problem facing management is that members of organizations are self-conscious human beings with an innate desire for individual freedom and gain, and an innate dislike of restriction and loss. They naturally compete for power, and for the benefits in freedom and gain they believe come with it. Therefore, every time a human organization is formed, and a hierarchy of management power is set up, a field of competitive tension is automatically created, as surely as a magnet generates a field of electromagnetism.

1.2 Culture of Organizations

For the purpose of discussion, this field of tension will be called the 'culture' of the organization. Defined many ways by scholars, the term 'culture' is used here to refer to the degree to which the members of an organization are willing to vest power and authority in management. Three somewhat arbitrary positions along a spectrum make useful benchmarks: (a) a culture of 'consent;' (b) a culture of 'cooperation;' and (c) a culture of 'dissent.'

In (a), organization members are accustomed to accepting the centralization of power in a management hierarchy, and look to specific directions from management for guidance on a wide variety of operations. In (b), members are accustomed to working together creatively under minimum management guidance, except for specific directions on matters that most can agree require a higher vision. In (c), members are accustomed to questioning management directions of all kinds, and frequently tend to place a higher value on individual independence than on group cooperation.

Obviously, the more a culture is oriented towards dissent, the more complex is the management challenge. To understand how managers have dealt with this problem, it is useful to construct a simple model of a generic organization, and see how variations in its structure relate to differences in internal cultures and external conditions. The model presented here is based on four tasks that can be said to constitute the primary responsibilities of management.

1.3 Management Task 1: Mission

Mission is the management task of defining and communicating the basic purpose for the creation and

existence of the organization. Its role is to act as a symbol of values, an orientation for planning, and an inspiration for action.

1.4 Management Task 2: Structure

Structure is the management task of planning and implementing the form of the organization. In practice, this usually means dividing the constituent members into clusters, called administrative units, and assigning a manager to each cluster. The result is often portrayed as an 'organization chart,' a diagram showing the administrative units, with lines linking them vertically to either a single person (called here the 'executive,' or a small group (called here the 'board,' or to both, at the top of the chart. Together these two entities, plus their support staff, constitute the management 'apex.'

The vertical linkage always takes the shape of an upside down tree. This tree represents the way that all members of the organization relate to each other in terms of their authority, that is to say, their legitimate power to exercise control over subordinates. A hierarchy controlled by an apex is always necessary, because organizations need a way to reconcile disputes in a timely manner.

1.5 Management Task 3: Procedure

Procedure is the management task of planning and implementing the functions of the organization. In practice, it means controlling the behavior of individuals within the organization to ensure that plans are executed and products are produced efficiently and effectively.

Systems theory suggests that human organizations, like biological forms of life, can be described in terms of function, as linear process (with some internal feedback loops) from 'input' to 'throughput' to 'output,' and that its output can be divided into two components, one called 'product' one called 'waste.' This simple model is popular in the fields of economics and business management, where it provides a foundation for the idea of using product to measure the functional performance of organizations of all types.

1.6 Management Task 4: Product

Product is the task of designing the products that the organization creates and exports into its external environment. Product in this sense is an extension of the concept expressed in the term 'mission.' The products of an organization should represent a more focused and defined expression of the values and directions implicit in its mission statement.

The concept of product has come to play a critical role in analyzing and comparing the relative value of different organizations. Two criteria are used, one focused on the internal environment of the organization 'efficiency,' and the other on its external environment ('effectiveness'). Both criteria use the product as a surrogate by which to measure the organization as a whole. Efficiency is the ratio derived from dividing the product by the input. It is a quantitative measure. Effectiveness is the degree to which the product satisfies the mission of the organization with regard to its external environment. It is a qualitative measure.

2. Model of a Generic Organization

Summing up the key factors outlined in Sect. 1 a graphic model can be visualized. The structure of the organization is represented by a triangle, with its vertical axis reflecting the authority hierarchy of management. The procedure of the organization is represented by a horizontal line across the bottom of this triangle, reflecting the flow of work transactions through the organization from input to output. Operators comprise the bottom layer of the triangle, performing the actual transactions that create the product as it moves from input to output. They are supervised by managers in the upper layers of the triangle.

The mission and structure responsibilities of management fall along the vertical axis of the triangle, and the procedure and product tasks fall along its horizontal axis. Mission and structure, once established, are not often changed, whereas product and procedure must be changed frequently to keep the organization in balance with its changing external environment. Structure is a key factor influencing product and procedure. Three basic variations in structure design can be observed in practice, as follows.

2.1 Structure Alternative 1: The Divisional Model

In this model, the basic principle for clustering people together into administrative units is their degree of commonality with regard to technique, skill, knowledge, or activity. This creates a set of what may be called 'functional' divisions, each one centered on a different type of specialized operation. The effect of this model is to allow the apex to delegate authority over specialized operations to subordinate managers within the hierarchy, while still retaining strong direction over the design and production of the composite final product.

This model emphasizes quality of input, by setting up hierarchies of specialized knowledge and skill that can apply technology to improve the product. Power over all four management tasks is strongly concentrated at the apex. This model can be very efficient,

but depends on a culture of consent, fast internal communication linkages, and a comprehensible external environment. It can lose effectiveness whenever any of these supporting conditions break down.

2.2 Structure Alternative 2: The Decentralized Model

In this model, the basic principle for establishing administrative units is to include in each division all the special resources necessary to produce a complete particular product. This model creates a set of 'product' divisions that share all of the characteristics of the parent organization, with the exceptions that: (a) each division produces a different product; and (b) enough freedom is withheld by the Apex to ensure the health of the parent organization.

This model emphasizes quality of output, by setting up hierarchies of situational analysts who can adapt product designs to changing external situations. Power over product and its procedure is delegated by the apex, but control over mission and structure is retained. This model can very effective in responding to local and rapidly changing product needs, but it depends on a culture of consent with regard to mission and structure, a greater aptitude for situational strategy by divisional leaders, and sufficient supplies of input resources to duplicate them across every product division. It can lose effectiveness whenever any of these supporting conditions breaks down.

2.3 Management Structure 3: The Matrix Model

In this model, the organization is divided into a set of 'functional' divisions as well as a set of 'product' divisions. The apex delegates control of both product and procedure, but avoids concentrating them entirely in the product divisions, by requiring them to share control of these tasks with the functional divisions through negotiated divisional agreements.

This model tries to achieve a balance between quality of input and quality of output, by delegating power in both directions. Power over mission and structure is retained at the apex. This model can be very effective in facing situations that call for the strengths of both the divisional model and the decentralized model. However, it heavily depends on a culture of cooperation in product and its procedure, and a culture of consent in mission and structure. Without managers and operators who value these cultures, this model can be difficult to sustain over long periods of time.

3. The Place of Planning in the Generic Model

3.1 Planning vs. Implementation

Planning is a process of choosing goals and ways to achieve them, 'before initiating action.' It is one of the basic activities of the human mind. As long as the process of planning and the initiation of action constitute sequential events within a single human mind, there is no need to reflect on the relationship between the two events. However, as soon as the planning is performed by one person, and the action to approve and implement the planning by another, the question arises: How shall the two events be connected? How this question is answered is the key to the administrative organization of planning.

3.2 The Generic Planning Process

The process of planning for a client begins with 'analysis' and ends with 'synthesis.' As defined here, it involves six sequential actions, the first three of which are analytic, and second three of which are synthetic:

(a) 'Survey' the nature of the current situation;

(b) 'Forecast' trends, assuming no purposeful intervention by the client;

(c) 'Compare' the trend scenario to the client's needs and desires;

(d) 'Invent' possible alternative scenarios achievable through intervention;

(e) 'Evaluate' the scenarios for feasibility and fitness;

(f) 'Integrate' alternative scenario elements into one final version, 'the plan.'

The planner can adopt one of two modes while performing this process. All six steps can be performed by the planner alone, with only the sixth step, 'the plan,' passed on to the client. Call this the 'advocacy mode.' Or, the client, and other interested parties as well, can be invited to participate in some, or all, of the steps. Call this the 'participatory mode.'

3.3 The Base Case: Planning and Implementation Combined

The least political complications ensue when the divisional model is used in a consent culture, and planning is performed solely by the management apex. Military organizations typically use this model this way, largely because they face such threatening external environments that internal control and secrecy in planning become cardinal virtues.

Certain kinds of business organizations also use the model this way, especially those facing fierce competition analogous to the military situation. However, businesses in the modern world increasingly seem to find it more useful to foster a cooperation culture than a consent one, and, to that end, often shift to some form of the decentralized or matrix models.

Government organizations, by contrast, especially in democracies, increasingly find themselves facing cultures of dissent in the larger organizational unit (i.e., the nation or municipality) for which they function as management. Within individual govern-

ment agencies, however, management can often develop a culture consent or cooperation, because members of the agency voluntarily contract for employment knowing the management climate of the place.

In summary, evidence from the field suggests that the model to which management naturally gravitates is the divisional one, with the synthetic steps of the planning process performed directly by the apex, supported when necessary by a delegation of the analytic steps to a subsidiary unit without hierarchical authority. Some theorists suggest that this model represents the default mode to which all organizations tend to revert, especially when facing external pressures, such as strict performance evaluation, or strong technical competition at the output end, or just scarce resources at the input end.

3.4 Delegated Planning and the Administrative Planning Unit

When external conditions overburden the planning capacity of the apex in a divisional organization, the easiest approach is to expand total capacity by setting up an administrative planning unit (APU). This may be viewed as a lesser delegation of power than is involved in re-engineering the structure of the organization into a decentralized or matrix model, but it is a delegation of power nevertheless, and a power that is potentially very broad in scope. It is easy for other administrative units in the organization hierarchy to become nervous or hostile. Like management itself, every APU must work in a power struggle.

Since the apex usually wishes to retain direct control of the organization's mission and structure, the delegated planning responsibility usually only applies to product and/or its procedure. In manufacturing organizations the former is often called 'research and design,' and the latter 'quality control.' In either case, although the plans of the APU affect the entire organization, the APU's most important client, in terms of plan implementation, is always the Apex.

3.5 The APU in Decentralized and Matrix Models

In the decentralized and matrix models, the responsibility of planning for product and procedure is already delegated in large measure to the individual product divisions. If the apex wishes to exert control or influence over the product units in these divisions, it may set up an APU to fill some planning role that is intermediary between the mission and structure role of the apex and the product and procedure role of the divisions. When this happens, the APU role is typically concentrated more on the analytic steps of the planning process than on the synthetic ones, or alternatively is focused on operational techniques for supporting a cooperation culture.

4. The Administrative Organization of Urban Planning

4.1 Urban Planning

Urban planning is a field and profession that arose after 1900, primarily in industrial nations. A search began to find methods to ameliorate the negative physical and social conditions of the industrial revolution, especially in urban areas. The concept arose that a good social reality needs the support of a good physical reality, and that the latter can be achieved by adopting a plan for the physical layout of the urban place. The idea was born that one of the major products of local government should be a spatial plan for the physical environment of its geographical area of jurisdiction.

In time, it became evident that a spatial plan by itself could not adequately deal with problems in the physical, social, and environmental systems of urban areas. Theorists concluded that other 'products' of local government (such as rules to regulate traffic, pollution, housing, and employment) were needed to supplement a spatial plan. Today, a wide typology of both spatial and operational plans are in use, as governments around the world seek to improve the form and function of their urban places for a wide variety of reasons. The challenge of getting such plans approved and implemented is made greater by one peculiar characteristic of urban planning that makes it qualitatively different from the kind of planning that all other types of human organizations engage in.

Two observations make the point. First, an organization's product is generally considered to be a creation that is extruded from the interior environment of the organization into its external environment. It is aimed at the exterior, not the interior, of the organization. Second, this product is much smaller in scale than its external environment, whether it be a battle plan for execution by a military organization, a computer or a software training program for sale by a business organization, or schools and police protection provided by a local government.

By contrast, the urban planning product is typically far larger in scale than these. More significantly, it is aimed, not externally, but internally—at the organization's own members. In fact, it is a plan to govern an important part of the lifestyle of the entire human society for which the government is the management. Because this plan is so sweeping in its internal effect, getting it approved and implemented is far more difficult than is the case for the typical

product plan in other organizations. It is more like the adoption of a labor–management agreement than the approval of a new automobile design.

4.2 Government Organizations

Other challenges to urban planning also arise because of the unique conditions under which government organizations must operate. These need some explanation. The mission of government organizations is generally to maintain a healthy and stable society. Typically, they consist of at least two tiers: national and local. National governments often face turbulent, complex, and hostile external environments, and tend to maintain a strong military function. For this and other reasons, most national governments seem to be organized structurally in the divisional model.

By contrast, local governments would seem to face less hostile challenges from their external environments, but, in fact, they must cope with some other conditions that are common to both national and local governments, and that may affect local governments more. The most significant of these are: (a) a limited managerial role within their organization (i.e., their constituency or society); (b) uncoordinated apex members; and (c) a culture of dissent. Together, these three conditions create a management situation of greater complexity than is the case for most nongovernmental organizations.

Regarding limits on governmental power, local governments are usually limited by statutes, in many more ways than is common for the management of nongovernmental organizations. This limitation of authority rests on a foundation of accepted social values, the most prominent of which are the importance given to protecting individual human rights (in democratic societies), and the importance given to protecting private economic markets (in capitalist societies).

Regarding uncoordinated apex members, democratic societies often create government structures that separate to various degrees the power of the board (the legislature) from the power of the executive (the president, governor, etc.), for the purpose of avoiding an excess concentration of power at the apex.

Regarding a culture of dissent, civil societies can often drift towards this end of the cultural spectrum for a variety of reasons, including problems created by the pace of technological change, or international relations, or changing human aspirations, or other sociological factors.

4.3 Urban Planning in Local Government

It is usual to find urban planning delegated to an APU, within a divisional local government structure. This APU inherits the same situational factors common to all APU's in divisional structures, but in addition, must deal with the unique conditions outlined above. In sum, the APU must find a way to get approval and implementation of its product in the face of: (a) the enormous social impact of its product; (b) a culture of dissent; (c) limited authority of government generally; (d) a fractured apex; and (e) competition for power with other units of government. What are some alternative approaches?

4.4 Strategies for Urban Planning APU's

The most direct approach of course, is to seek the approval of the apex, followed by the issuance of an apex command protocol throughout the organization. This approach has the merits of simplicity and efficiency, and can work well in situations where the scope of social change embraced by the plan is not too big, the cultural dissent level is not too high, the apex is not too uncoordinated, and the power struggle with other governmental units can be won. It fits well with the advocacy mode outlined earlier for the planning process method. Call this the 'apex approach' and the 'stability scenario.'

A second approach has better chances of success when the social change gets bigger, the dissent level gets higher, the apex is more uncoordinated, and the internal power struggle outcome is uncertain. In this approach, the APU needs to shift its planning process to the participatory mode, and reach out early in the process to seek the participation and counsel of all the key managers and operators within the government structure, as well as key movers and shakers within the general society. Call this the 'stakeholder approach' and the 'complexity scenario.'

A third approach may be necessary when the social change is very large, the dissent level is very high, the apex is seriously fractured, and the interagency power struggle is intense. In this approach, the APU expands the outreach scope of its participatory mode planning process to engage the support of as many individuals in the wider society as it can reach, on the theory that this third tier group may be persuaded to influence the stakeholders and the apex to approve and implement the plan produced by the APU. Call this the 'constituency approach' and the 'conflict scenario.'

4.5 Relationship of Organization to Strategy for Urban Planning APU's

Where the charter of the APU does not specify the structure of the unit, the unit's executive (usually called the planning director) can choose an organizational form. In small APU's in a stability scenario, the prevailing practice seems to mirror the tendency of other organizations to adopt the divisional form and use the advocacy mode. In a small

APU, the director is personally under a lot of external pressure, and, like a beleaguered military commander, may feel the need for internal secrecy and control, especially if the situational scenario shifts from stability to complexity or conflict.

In large APU's, especially within a complexity scenario, there are benefits to adopting the decentralized form and using the stakeholder approach, but they depend on developing an internal cooperation culture. By setting up different units to plan for subsections of the society, the scale of the planning product becomes more manageable. With responsibility for product delegated, the director is under less immediate political pressure. The weakness of this approach only surfaces when there are insufficient resources to supply all the divisions with adequate technical knowledge, and/or the need arises to make the divisional plans consistent with each other from a larger perspective.

Conflict scenarios present the most difficult situations to deal with, especially if the technical or consistency component of the plans is important. It is dangerous to use the constituency approach, because the APU may lose apex support if the elected officials feel threatened by this outreach into their electorate. However, if the director can build an internal culture of cooperation sufficient to support a matrix form, it is possible to blend a set of procedures and products adequate to the challenge, but, it is no sure thing. In the final analysis, urban planning is an art, not a science—in administrative organization as well as urban form.

See also: Administration in Organizations; Decision Making, Psychology of; Leadership in Organizations, Psychology of; Leadership in Organizations, Sociology of; Management: General; Organizational Climate; Organizational Culture; Organizational Culture, Anthropology of; Organizational Decision Making; Organizations, Sociology of; Planning Theory: Interaction with Institutional Contexts; Urban Places, Planning for the Use of: Design Guide

Bibliography

Churchman C W 1979 *The Systems Approach*. Dell, New York

Drucker P F 1973 *Management: Tasks, Responsibilities, Practice*. Harper and Row, New York

Frame J D 1995 *Managing Projects in Organizations*. Jossey-Bass, San Francisco, CA

Galbraith J R 1977 *Organization Design*. Addison-Wesley, Reading, MA

Grove A S 1983 *High Output Management*. Random House, New York

Hatch M J 1997 *Organization Theory*. Oxford University Press, Oxford, UK

Kaiser E J, Godschalk D R, Chapin F S Jr 1995 *Land Use Planning*. University of Illinois Press, Urbana, IL

Keidel R W 1995 *Seeing Organizational Patterns*. Berrett-Koehler, San Francisco, CA

Miller J G 1978 *Living Systems*. McGraw-Hill, New York

Mintzberg H 1989 *Mintzberg on Management*. The Free Press, New York

So F S, Getzels J 1988 *The Practice of Local Government Planning*. International. City/Country Management Association, Washington, DC

R. E. Tustian

Planning Ethics

Planning ethics refers to the integration of ideas from the fields of planning and moral philosophy, or ethics. As such, it covers a vast terrain of issues in planning theory and practice. Contrary to terms such as 'ethics in planning' or 'ethics of planning,' planning ethics does not privilege either planning or ethics; nor does it connote only a descriptive representation of values in planning. Instead, it is intended to depict the fact that planners and planning can benefit from incorporating ideas from moral philosophy into planning thought and professional endeavors. Further, ethics can also benefit by gaining insight as to how concepts in moral philosophy actually work in practice. Both disciplines may then be more effective in prescribing courses of action in human settlements (planning) and developing theoretical approaches of how persons ought to behave (ethics).

1. Ethical Content in Planning

Planning ethics may be divided according to several different analytic schemes. Wachs (1985) developed a particularly comprehensive and useful typology in this regard. As supplemented by Hendler (1995), it represents a way of discussing planning ethics with reference to five aspects of planning theory and practice. These attributes are: everyday behavior, plan making, administrative discretion, planning techniques, and normative planning theory. In each category are examples of descriptive analyses in which researchers describe and discuss the values planners actually have (e.g., Howe 1995, Gowdy and Hendler 1999) as well as normative analyses where planners advocate particular approaches to planning ethics (e.g., Beatley 1994, Harper and Stein 1993).

1.1 Everyday Behavior

Everyday behavior pertains to the fact that planners are, as persons or moral agents, subject to the same ethical guidelines as everyone else, although their professional status may give rise to particular rights

and responsibilities. Thus, planners may be judged according to the consequences of their actions (consequentialist ethics) or with reference to the inherent characteristics of the actions themselves (nonconsequentialist or deontological ethics). Consequentialist approaches to planning ethics include most prominently utilitarianism with its focus on creating the greatest amount of utility from either individual actions (act-utilitarianism) or actions arising from the application of ethical rules (rule-utilitarianism). Nonconsequentialist approaches, on the other hand, include Kantian ethics with its interest in the ethical treatment of persons as ends and not (only) means. While Kantian and utilitarian perspectives continue to attract much attention in philosophical and planning circles (e.g., Howe 1990, 1995), other consequential and nonconsequential ethical approaches have also proven attractive to planners and philosophers alike; these include, but are not limited to, communitarianism (e.g., Blanco 1995) which emphasizes community practices and virtues, environmental ethics (e.g., Beatley 1994, Jacobs 1995) which focuses on the relationships between and among humans and nonhuman components of our environments, and feminist ethics (e.g., Hendler 1994, Ritzdorf 1995) which attempts to incorporate into our considerations of moral issues the notions of care, trust, nurturing, and eradicating all forms of oppression. Communicative or discourse ethics (e.g., Benhabib 1992), too, has become an important part of contemporary ethical thinking and refers to procedural accounts of how one should make ethical decisions, with the presumption that, if one adheres to the rules guiding such a procedure, the outcome should, by definition, be good or right. This approach may be seen as underlying models of planning processes that emphasize consensus building and/or communicative rationality (e.g., Forester 1989).

Considerations of planners' everyday actions that are based on these sorts of ethical ideas include evaluations of behaviors corresponding to planners as professionals and nonprofessionals. Professionals are often seen as having added moral obligations in contrast to nonprofessionals, given their status in society. Doctors, lawyers, engineers, architects, and planners, among others, are trusted with going beyond motivations of profit and self-interest and focusing instead on the public interest. This status is accompanied by particular responsibilities for specialized knowledge not easily accessible to nonprofessionals, and self-regulation in the face of this specialization (Bickenbach and Hendler 1994). Specific behaviors addressed in this first category of planning ethics include, among others: avoiding conflict of interest; maintaining appropriate relationships with clients, employees, and colleagues; advertising one's services in certain manners; practicing within one's area and scope of competence; providing independent professional advice; and treating information appropriately whether in terms of confidentiality or disclosure in reference to considerations of the public interest. Providing ethical guidance for planners facing decisions in these and related areas is central to this first aspect of planning ethics that focuses on everyday behavior.

1.2 Plan Making

Plan making is that part of planning having to do with the development, implementation, monitoring, and evaluation of plans, programs, and policies. Together, these sorts of normative statements have ethical implications in that certain amenities or disamenities are distributed or redistributed among citizens (and nonhuman components of our environments) in ways that benefit certain individuals or groups and impose costs on others. Assessing the implications of such decisions and actions pertains directly to the ethical theories and approaches listed above. Kantian-based ethics, for example, may be applied via a Rawlsian perspective in which resources are said to be distributed equally or, if inequity exists, so as to improve the situation of disadvantaged segments of the population in question. Examining the consequences of a planning document using a utilitarian calculus of costs and benefits is also common (see Sect. 1.4). Whichever ethical theory is used, however, is immaterial to the central idea that almost every decision made by planners potentially provides benefits and harms to individuals, groups, neighborhoods, or jurisdictions, and that this result has ethical implications associated with it. The importance of examining the (usually implicit) underlying assumptions and resulting implications of the ethical directions of plans, programs, and policies is the focus of this area of planning ethics.

1.3 Administrative Discretion

Administrative discretion refers to the ambiguity in planners' roles in their working environments. A number of roles have been proposed, ranging from advocate to technician to mediator to social learner (e.g., Gunton 1984). The fact that a planner's role is not discretely and concretely prescribed via the planner's professional organization and/or work environment means that considerable latitude may exist in the planner's determination of an appropriate or desirable way-of-being in professional planning work. Planners thus may exercise discretion in selecting for themselves a role (or roles) but this decision must be subject to ethical scrutiny. Whether, for instance, a planner may in fact explicitly advocate for a particular point of view is a controversial issue in this area of planning ethics. While, as suggested in the discussion of plan making above, any plan or program or policy does necessarily advocate for one course of action as opposed to

others, what and how such directions are advocated is sometimes seen as problematic within planning organizations.

In addition to choices among roles is the fact that administrative discretion highlights the issue of accountability in that planners are not elected officials and yet they may represent the interests of one group over another at any one point in time and in any given plan-making exercise. While such an affiliation may be couched in terms of the public interest, any value-laden connection between planning and developers on the one hand and environmentalists or social justice advocates on the other, for example, may be regarded as an improper exercise of the planner's discretion regarding that planner's role. Despite the fact that this representation of planning work is rooted in the largely outmoded conception of planning as value-free technical work, planners must acknowledge and address the fact that their discretion may be perceived as potentially troublesome.

1.4 Planning Techniques

Planning techniques are those methods or processes followed by planners in preparing or evaluating their plans, programs, and policies. Planners have a considerable array of analytic techniques at their disposal. Cost–benefit analysis, risk assessment, environmental assessment, program evaluation, participatory methods, and many others may be included in this regard. These techniques may aid a planner in deciding a course of action to follow or in assessing the effects of particular actions on different publics. Most of these techniques, and perhaps especially cost–benefit analysis, have been connected conceptually and methodologically to utilitarianism, as follows. First, utilitarianism relies upon the quantification of 'goods' and 'harms' so that both can be summed and compared. This is a starting point, too, for cost–benefit analysis in which costs and benefits are identified, measured, added, and contrasted in order to determine whether a course of action would result in more benefits than costs. Second, this sort of conclusion is clearly based on utilitarian thinking in which a maximum amount of good is sought.

The fact that planning techniques are linked to ethical theories is not particularly troubling. This kind of observation attests to the post-positivist world in which we find ourselves where the planner is not regarded as a value-free technician but as an individual who necessarily brings certain biases, values, and conceptual lenses to their work. Similar, then, to contemporary views of science, defining one's perspective and unearthing one's assumptions is an essential first step in doing research and in doing planning. While this approach to science and to planning continues to be debated, an issue arising from the mere association of certain ethical approaches to certain planning techniques remains. This is the fact that techniques based on consequential or utilitarian thinking can be seen as over-represented in planning ethics in that methods based more on nonconsequential thinking do not have the prominence of their philosophical cousins. One branch of moral philosophy thus gets short shrift in helping planners to develop and evaluate plans—a branch that is generally regarded as just as viable and important as consequential approaches. Lake (1993) discusses one aspect of this phenomenon in the context of feminist ethics and techniques based on geographical information systems (GIS). He argues that the data-aggregation focus on GIS is incompatible with the focus on the 'concrete other' (Benhabib 1992) in feminism. The same could be said of other ethical theories that attempt to highlight individual experiences and inherent value (the 'right') as opposed to formulaic assessments of consequences (the 'good'). Uncovering assumptions and seeking a broader distribution of ethical underpinnings for planning techniques is of central importance to this area of planning ethics.

1.5 Normative Planning Theory

With its focus on the nature and intent of planning as a profession, as well as its scope or mandate, normative planning theory falls clearly within the realm of ethical analysis. What the planning profession *should* profess (Haworth 1984) is a question that demands this sort of scrutiny.

There have been many disparate suggestions regarding the ethical mandate of planning. Udy (1980) proposed that the four eternal values of goodness, beauty, truth, and love lay at the heart of planning and should be recognized as the normative direction for planning initiatives. Blanco (1995) developed an alternative list that included reason, community, environment, and democracy. Krumholz and others (e.g., Krumholz and Forester 1990) have written and practiced extensively in the area of equity planning, and Davidoff (1978, pp. 69–70) forcefully asserted that,

> if a planner is not working directly for the objective of eradicating poverty and racial and sexual discrimination, then she or he is counterproductive. If the work is not specific in its redistributive aims, then it is at best inefficient. If the work is not aimed at redistribution, then a presumption stands that it is amoral. These are strong words. They must be. So long as poverty and racism exist in our society, there is an ethical imperative for a single direction in planning.

In the face of these conflicting sets of values, the field and profession of planning continue to have a normative nature that can only be called ambiguous. Each professional planning organization defines for itself its

central ethical values and then applies them in creating a regulatory structure for its practitioners. In contrast, then, to the fundamental pursuit of health in medicine and the quest for justice in law, planning cannot express its basic ethical norms as succinctly nor with as much certainty. Developing such a rubric or becoming more accepting of a diverse set of principles of moral guidance is a key part of this last aspect of planning ethics.

2. *Regulating Planning Ethics*

All of these ethical aspects of professional behavior are addressed in the codes of professions such as planning (see, e.g., Hendler 1990). Codes of conduct, especially, provide ethical guidance in the form of rules, sanctions, and encouragement towards, or away from, particular behaviors. Codes of ethics are more general and are usually seen as articulating a normative basis for the profession in the form of key values that are thought to be central to the professional community in question. In planning, these values often include environmental integrity, public participation, holistic decision making in terms of temporal, geographic, and disciplinary scope, and a respect for a social justice and a multitude of diversities. While not easily enforceable, codes of ethics provide for a unified voice of the moral intent of a profession and contribute to community building and identity formation within a profession such as planning.

Codes of conduct are used in assessing the behavior of individual professionals and, if necessary, imposing sanctions if the rules under scrutiny are found to have been violated. Suspension, financial penalty, expulsion, and education/counselling are examples of the kinds of sanctions levied by professional organizations. In that planners rarely have to be certified to practice their work, considerable debate is directed towards the licensing of planners and attendant issues regarding their discipline. While some planning institutes are content to regulate only their own members and show little concern for the behaviors of non-members, others have endeavored to expand their membership and refine the notion of planning so that only their members may refer to themselves as professional planners.

Together, codes of ethics and conduct may be seen as a bridge linking abstract moral theories and practical work such as planning in that the sorts of ethical issues discussed here are articulated precisely and resulting discussions may shed light on the implications of different ethical ideas 'on the ground.' As such, codes are growing and developing documents that have evolved significantly during the past two decades of planning ethics discussions. They themselves have also stimulated much in the way of discussion regarding the efficacy of the values they present as being good or right (e.g., Lucy 1988). Through the processes employed in formulating a code and in subsequent discussions and revisions, planning ethics has enjoyed much in the way of constructive critique (e.g., Kaufman 1990).

Some planning organizations go beyond codes *per se* in educating their members about ethics and contributing to the interpretation of their codes by publishing ethical advisories that apply given sections of a code to a particular planning issue (e.g., AICP 1996). As the sophistication of planning ethics discourse increases, codes and planning practice can only be expected to benefit from an enhanced understanding of, and perhaps consensus about, the fundamental values that underlie planning activities.

3. *Planning Ethics in Theory, Practice, and Education*

Discussions of planning ethics have been most pervasive in planning theory where philosophical analyses of planning and planners are commonplace. It is here that arguments regarding the central tenets and values of planning have arisen, and varying normative models of planning have been developed. Planning practice, however, has only recently begun to emphasize planning ethics in the form of increasingly comprehensive and thoughtful professional codes, as well as enhanced disciplinary proceedings. Empirical analyses of planners' values and accompanying ethical approaches have assisted in our constructing snapshots of the moral terrain in which planners have placed themselves. While consensus is not apparent, certain loci of values around environmental integrity, social justice, and democratic decision making often emerge. In addition to such descriptive studies are normative assessments of planning activities that include both subfields of planning such as transportation, environment, education planning, waste management, etc. (e.g., Hendler 1995, Beatley 1994), as well as specific planning behaviors (e.g., Howe 1995).

Planning education, which usually has the intent of both training future professionals as well as being part of universities, where intellectual growth and development is a priority, has begun to 'mainstream' discussions of planning ethics in its curricula. Far from inculcating particular values, teaching about ethics in planning programs usually has the goal of helping students engage in rigorous ethical analysis as well as preparing them to practice in a profession which espouses certain principles. Thus, values are presented along with the means to examine them critically. This performs the dual function of satisfying planning institute expectations regarding the education of future professionals as well as the more academic intent of teaching students how to think clearly and coherently. Having the skill of developing and critiquing logical arguments helps students and professionals work through ethical dilemmas they will certainly

encounter in their employment and contribute to the increasingly sophisticated nature of planning ethics—theory, practice, and education.

4. Challenges in Planning Ethics

While the value-laden nature of planning endeavors is generally accepted, certain challenges and questions remain. Examples of these include the following. The similarities and differences between public and private planners continue to be analyzed in terms of their moral obligations. The fast-encroaching world of computer technologies and associated planning applications remains ill examined in terms of the fit of these technologies and the moral intentions of the planners and the planning profession they serve. As planning becomes broader substantively with focuses on land use planning but also on social planning, health planning, environmental planning, etc., it remains to be seen whether a single professional organization may be appropriate to govern planners' activities. Further, addressing planning issues in ways that not only accommodate but also celebrate racial, ethnic, sexual, and other forms of diversity continues to occupy the interests of many people who write about and practice planning. Current emphases on planning processes and public participation contribute to resolving such issues but do not address questions regarding such processes as issues of identity, representation, method, and practical considerations regarding the efficient use of planning resources. Empowerment, while often discussed in the context of planning theory, is seldom found in planning codes. Economic development, though central to the 1990s climate of planning, is not well represented in the ethical guidance supplied to planners. And so on. Clearly, planning will continue to have much to consider in the way of discussions regarding its ethical mandate and the methods to be employed in addressing the issues that arise within the context of its thought and practice.

As ethics has highlighted for planners the values in their work, and the need to address critically the values that lie at the base of their theories and their practices, so, too, has planning underscored the importance of discussing ethical issues with an eye to the application of any theoretical stance based in moral philosophy. Thus, planning ethics, as an interdisciplinary field of inquiry and one which does not rest on the laurels of either 'parent' discipline (i.e., planning or ethics), planning ethics promises to contribute to an especially critical and grounded approach to posing, and answering, the sorts of questions and challenges listed here.

See also: Environmental Conservation: Ethical Concerns; Environmental Justice; Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives; Ethics and Values; Neighborhood Revitalization and Community Development; Planning Issues and Sustainable Development; Planning, Politics of

Bibliography

American Institute of Certified Planners (AICP) 1996 *AICP Code of Ethics and Professional Conduct*. AICP, Washington, DC

Beatley T 1994 *Land Use Ethics*. Johns Hopkins, University Press, Baltimore, MD

Benhabib S 1992 *Situating the Self*. Routledge, New York

Bickenbach J, Hendler S 1994 The moral mandate of the 'profession' of planning. In: Thomas H (ed.) *Values and Planning*. Avebury Technical, Aldershot, UK, pp. 162–77

Blanco H 1995 Community and the four jewels of planning. In: Hendler S (ed.) *Planning Ethics. Readings in Planning Theory, Practice and Education*. Center for Urban Policy Research, New Brunswick, NJ, pp. 66–82

Davidoff P 1978 The redistributive function in planning. Creating greater equity among citizens of communities. In: Burchell R, Sternlieb G (eds.) *Planning Theory for the 1980s*. Center for Urban Policy Research, New Brunswick, NJ, pp. 69–72

Forester J 1989 *Planning in the Face of Power*. University of California Press, Berkeley, CA

Gowdy A, Hendler S 1999 Environmental values and professionalism in planning: a case study. *Plan Canada* **39**(2): 28–32

Gunton T 1984 The role of the professional planner. *Canadian Public Administration* **27**(3): 399–417

Harper T, Stein S 1993 Normative ethical theory: is it relevant to contemporary planning practice? *Plan Canada* September: 6–12

Haworth L 1984 Orwell, the planning profession, and autonomy. *Environments* **16**(2): 10–15

Hendler S 1990 Professional codes as bridges between planning and ethics: a case study. *Plan Canada* **30**: 22–9

Hendler S 1994 Feminist planning ethics. *Journal of Planning Literature* **9**(2): 115–27

Hendler S (ed.) 1995 *Planning Ethics. A Reader in Planning Theory, Practice and Education*. Center for Urban Policy Research, New Brunswick, NJ

Howe E 1990 Normative ethics in planning. *Journal of Planning Literature* **5**(2): 123–50

Howe E 1995 *Acting on Ethics in City Planning*. Center for Urban Policy Research, New Brunswick, NJ

Jacobs H 1995 Contemporary environmental philosophy and its challenge to planning theory. In: Hendler S (ed.) *Planning Ethics. A Reader in Planning Theory, Practice and Education*. Center for Urban Policy Research, New Brunswick, NJ, pp. 83–103

Kaufman J 1990 American codes of planning ethics: content, development and after-effects. *Plan Canada* **30**: 29–34

Krumholz N, Forester J 1990 *Making Equity Planning Work*. Temple University Press, Philadelphia, PA

Lake R 1993 Planning and applied geography: positivism, ethics, and geographic information systems. *Progress in Human Geography* **17**(3): 404–13

Lucy W 1988 American Planning Association's ethical principles include simplistic planning theories. *Journal of the American Planning Association* **54**: 147–9

Ritzdorf M 1995 Feminist contributions to ethics and planning theory. In: Hendler S (ed.) *Planning Ethics. A Reader in*

Planning, Theory, Practice and Education. Center for Urban Policy Research, New Brunswick, NJ, pp. 104–19
Udy J 1980 Why plan? Planning and the eternal values. *Plan Canada* **20**(3/4): 176–83
Wachs M (ed.) 1985 *Ethics in Planning*. Center for Urban Policy Research, New Brunswick, NJ

S. Hendler

Planning Issues and Sustainable Development

'Sustainable development' or 'sustainability' for short simply means that in a global context any economic or social development should improve and not harm the environment. This concept has developed from a global political process over the last three decades of the twentieth century into one that now touches every part of society. Sustainability has cut across all disciplines and professions, and has developed many complexities in relation to how people plan for their futures (Pezzoli 1997). Its application to planning is full of interest and controversy.

Town and regional planning has its origins firmly in a desire to improve the environment and create a better future. The visions of Ebenezer Howard, Patrick Geddes, and Lewis Mumford, for example, were all rooted in the need to make a greener city. Today there are many planners who use sustainability to rationalize why they are planning the way they do. This article will try to clarify the ways that sustainability is shaping how planning is done, as well as the differences in the ways that planners relate to sustainable development.

The issues relating to sustainability that appear to be uniting planners are:

(a) How necessary it is to use holistic planning frameworks; and

(b) How planning processes must start with community/civil society perspectives.

The issues which appear to divide planners are:

(a) How planners view the size of cities; and

(b) How planners view density and car dependence.

Some resolution will be attempted on these issues.

1. *The Concept of Sustainability*

Sustainable development was defined by the Brundtland Commission in 1987 after they were given the task of resolving the apparent global conflict between development and the environment (WCED 1987). They concluded that it is possible to develop the world in a way that improves the environment if:

(a) High income nations see their responsibility to reduce their resource consumption with every new step in the development process;

(b) Low income nations direct development to those who need it most as only then will population growth stabilize and misuse of natural resources begin to ease;

(c) All nations need to see development as meeting community goals, thus development must be inherently a 'bottom up' process; and,

(d) All professions and disciplines need to adjust to a more holistic, integrated approach if sustainability is to be converted into real change.

The thrust of this article is that, within the sustainable development framework of thinking, there are clear differences in what planners are suggesting sustainability means for cities. However, the overall situation cannot be seen as just a series of raging controversies, and indeed there are substantial areas of agreement.

2. *The Agreements: Holistic Frameworks and New Planning Processes in the Age of Sustainability*

Dyck (1998) suggests that sustainability is a concept which challenges prevailing planning theory because planning has emerged primarily from a modernist social context. In this perspective, modernism has associated growth or development with increased resource use, trickle-down economics, and 'expert-oriented' planning. Very few planners are, however, espousing a modernist view in the new millennium, and even the World Bank seems to agree with the basic conceptions of how sustainable development should now be viewed (Leitmann 1999). In particular, the areas of emerging agreement are the need for new, more holistic planning frameworks and community-sensitive planning processes.

2.1 *Holistic Frameworks*

The principles of sustainability outlined above can be applied to cities, though the guidance on how this can be done was not very clear in Agenda 21 and other UN documents. It is probably true to say that the major environmental battles of the past were fought outside cities, but that awareness of the need to come back to cities is now recognized universally by environmentalists, government, and industry (United Nations Center for Human Settlements 1996).

Anders (1991), in a global review of the sustainable cities movement, pointed out that 'The sustainable cities movement seems united in its perception that the state of the environment demands action and that cities are an appropriate forum in which to act' (p. 17).

One of the strongest themes running through the literature on urban sustainability is that, to solve problems, planners need to view the city as an *ecosystem*. As Tjallingii (1993, p. 7) puts it: 'The city is

(now) conceived as a dynamic and complex ecosystem. This is not a metaphor, but a concept of a real city. The social, economic and cultural systems cannot escape the rules of abiotic and biotic nature. Guidelines for action will have to be geared to these rules.'

The city is an open ecosystem, having inputs of energy and materials. The main environmental problems (and economic costs) are related to sustaining the growth of these inputs, and managing the increased outputs. By looking at the city as a whole and by analyzing the pathways along which energy and materials (and pollution) move, it is possible to begin to conceive of management systems and technologies which allow for the reintegration of natural processes, increasing the efficiency of resource use, the recycling of wastes as valuable materials, and the conservation of (and even production of) energy.

There may be ongoing academic debate about what constitutes sustainability and how it applies to cities (Slocombe 1993, Mitlin and Satterthwaite 1994), but what is clear is that many strategies and programs around the world have begun to apply such notions both for new development and for redevelopment of existing areas. Holistic frameworks such as the Extended Metabolism Model (Newman and Kenworthy 1999), the Ecological Footprint (Wackernagel and Rees 1996), and Local Environmental Action Planning (Leitmann 1999) are all emerging as new planning frameworks.

2.2 Planning Processes

One of the strengths of the sustainability concept is that it is not a set of regulations or even highly definitive principles. Its principles need to be examined by every local community to see how they can be applied to create better long-term futures for their children. Thus planners at the local level have often found something of the reality of sustainability through the use of Local Agenda 21 Plans which have emerged as a tool for planning since the Rio conference (ICLEI 1996, 1997, Selman 1996). Over 2,000 local governments have produced such plans. Others have adopted approaches based on sustainability indicators such as Sustainable Seattle (City of Seattle 1993) or the Local Environmental Action Plan (Leitmann 1999), and many are now using ICLEI's Cities for Climate Protection system as a guiding approach.

All these new processes stress that planning has moved beyond traditional approaches such as Impact Assessment or Strategic Planning based on trend projections and functional separation. They have in common a desire to plan from a community base, integrated with the goals of the market and other levels of government, including the new global agendas generated from Rio, Kyoto, and beyond.

The stories of hope emerging at a local level are rich and diverse when examined at the grass roots level (e.g., Pathways to Sustainability Conference 1997, ICLEI 1997). The reason for this is probably that at the local level it is possible for government more easily to make the huge steps in integrating the environmental, economic, and social, in order to make policy developments that are sustainable. They are also closer to concerned people, and more distant from the single-issue powerful lobbies such as the fossil fuel and road lobbies, who are so obvious in shaping national priorities (Korten 1990, 1995).

2.3 The Disagreements: Application of Sustainability to City Size, Urban Form, and Transportation

The disagreements that are apparent among planners, who all profess to want more sustainable cities, are reducible to issues concerning urban form and transportation. They relate to the size of cities and to their density/car dependence.

3. City Size and Sustainability

Many planners subscribe to the view that, on environmental grounds (and other criteria), cities would be better if they were smaller (Troy 1996, Morris 1982). This is an old planning debate, for example, Fischer (1976) shows how optimum size cannot be found on any social or economic variable, but the debate has new currency in light of the current move to seek 'smart growth' and growth management. The desire to keep settlements small is seen by many to be a major part of what sustainability means.

The evidence from studies of settlements shows that as cities get larger they become more efficient in energy use per person and other critical sustainability variables (Newman and Kenworthy 1999, Naess 1993, Smith et al. 1998). This appears to be related to the economies of scale and density associated with bigger cities, which applies equally to technologies seen to be appropriate for sustainable development such as transit, recycling, or the new information industries. It does not mean, however, that larger cities do not reach capacity limits in terms of their airsheds and water systems. But it does mean that simple ideas of decanting people from big cities to smaller ones would not help to create a more sustainable global future.

The evidence suggests that smaller cities find the process of achieving efficiency in economic and resource terms harder than larger cities. The strong emotional appeal of 'smallness' to our age is not, however, without its basis. E. F. Schumacher's 'Small is Beautiful' approach, with its attack on modern gigantism, is still relevant as it suggests that all technology needs to be at an appropriate scale for the community that it is meant to be serving. Community-scale technology is emerging as communities begin to

assert their roles—whether the technology is in villages in the Third World or modern large cities (Korten 1990). The thrust of the New Urbanism and the 'urban village' movement is that communities need to be physically designed for and given infrastructure at the scale of the community, in cities both large and small (Calthorpe 1993).

3.1 Density and Car Dependence

In cities across the world the issue of urban sprawl is firmly on the agenda. Associated with it are the issues of density and levels of car dependence, and hence traffic. In the USA in 1998, 240 local governments had antisprawl initiatives on their ballots, and reductions in traffic had become a top priority across the country. However, debate among planners on these issues shows little sign of resolution.

The divide continues to widen between those advocating more compact cities and transit-oriented reurbanization (e.g., Cervero 1999, Whitelegg 1993, Roseland 1998, Newman and Kenworthy 1999, Calthorpe 1993, Elkin et al. 1991), and those advocating reduced densities and only technological changes to cars (e.g., Gordon and Richardson 1989, Troy 1996, Wachs and Crawford 1991, Johnson 1993). Another category of planners sit on the density/car dependence fence (e.g., Haughten and Hunter 1994), while others see the problems but tend to avoid the planning implications in low density, car-dependent planning (e.g., Leitmann 1999, World Bank 1996, Stren et al. 1992).

The data on density and car dependence is very clear. All cities appear to lie on the negative exponential curve of density and gasoline per capita (Newman and Kenworthy 1999). The same relationship is seen across cities, so that as density increases, per capita energy goes down, well beyond any income factors, for example, Manhattan uses 500 percent less energy per capita than the outer suburbs of New York. The mechanism is understood easily in terms of the relative ability to use different modes as distances between land uses spread out. It suggests that density is a powerful determinant of gasoline use, and hence if cities are to develop sustainably they should be facilitating reurbanization rather than new fringe development. Other planning studies have also shown over many decades the link between low urban density and a range of environmental problems (RERC 1974, Berry et al. 1974).

Despite this kind of evidence there is a long tradition of planners seeing density as a problem (Newman and Kenworthy 1989). Those planners advocating the benefits of sprawl who are associated with land or automobile lobbies can be discounted in debates about urban sustainability (STPP 1994). However, there are many environmental planners who also adopt the benefits of low-density planning (e.g., Gordon 1990, Todd and Tukel 1981, Coates 1981). This suggests that there may be a deeper issue in this debate; an issue that also lies behind the question of city size and sustainability.

Radberg (1995), when reviewing the form of cities required to achieve sustainability, described the rationale for a more compact city which can conserve energy. He then set out the form of a city for achieving green objectives that are 'related to the ecological perspective' for 'recycling and cultivation'; this city is 'greener, more ruralized, more spread out.' He therefore raised the questions: Is there a conflict between the low-energy city and the green city? Is there a fundamental conflict between the need to have more urban land for the green city in order to do local ecological processing, and a low-energy city with its need to minimize travel distances and hence have a frugal use of land?

The resolution of this issue as suggested by Radberg (1995) was to enable some parts of cities to be developed more intensely for reductions in energy, and other parts to be more reduced in their density so that 'greener' activities could occur there. The evidence, however, is that those planners who support the reduction of density are not able to accept that any density increases are worthwhile (e.g., Troy 1996). On the other hand, studies of urban ecology innovations by Scheurer (1999), confirmed by Beatley (2000), found that there are no increases in ecological activity in low-density areas. Indeed, the best examples of urban ecology (recycling, solar design, storm water retention, permaculture …) occurred in denser developments, mostly in inner city areas where communities were able to create innovative experiments, or in intentional rural or urban-fringe communities. In the car-dependent suburbs urban ecology innovations are rare.

Newman and Kenworthy, (1999) tried to resolve the issue of density and the green city. They suggest that green innovations can indeed occur anywhere in a city, although experiments may best be applied in rural areas, especially in places where rural population decline is occurring. However, the broader global sustainability agenda cannot be subsumed by this new rationale for density-lowering; cities need to reurbanize while conducting simultaneous improvements in urban ecology. They suggest that the difference in attitude between the two groups is as much due to confusion over where different approaches to sustainability should be employed. In simple terms, the city should be allowed to become more urban, and the countryside more rural, if the multiple goals of planning and sustainability are to be met.

A resolution of this issue may not occur while people still use the word 'sustainable' to describe extensions of low-density, car-based development (even if designed with New Urbanist principles). The strong emphasis on reductions in energy that is now

associated with sustainability will probably mean that car-based development increasingly will have to be described using other words that are less demanding of their global contribution.

See also: Environmental Economics; Environmental Vulnerability; Environmentalism: Preservation and Conservation; Sustainability Transition: Human–Environment Relationship; Sustainable Development; Sustainable Transportation; Transportation Geography; Transportation Planning; Transportation: Supply and Congestion; Urban Geography

Bibliography

Anders R 1991 The sustainable cities movement. Working Paper No. 2, Institute for Resource and Security Studies, Cambridge, MA

Beatley T 2000 *Planning for Sustainable Cities: The European Experience*. Island Press, Washington, DC

Berry B et al. 1974 Land use, urban form and environmental quality. Research Paper No. 155, University of Chicago Press, Chicago

Calthorpe P 1993 *The Next American Metropolis: Ecology and Urban Form*. Princeton Architectural Press, Princeton, NJ

Cervero R 1999 *Transit Planning*. Island Press, Washington, DC

City of Seattle 1993 Sustainable Seattle: Indicators of Sustainable Community, City of Seattle, WA

Coates G 1981 *Resettling America: Energy Ecology and Community*. Brick House, Andover, UK

Dyck R G 1998 Integrating planning and sustainability theory for local benefit. *Local Environment* **3**(1): 27–41

Elkin T, McLaren D, Hillman M 1991 *Reviving the City: Towards Sustainable Urban Development*. Friends of the Earth, London

Fischer C S 1976 *The Urban Experience*. Harcourt Brace, New York

Gordon D (ed.) 1990 *Green Cities*. Black Rose, Montreal, PQ

Gordon P, Richardson H W 1989 Gasoline consumption and cities: A reply. *Journal of American Planning Association* **55**: 342–6

Haughton G, Hunter C 1994 *Sustainable Cities*. Taylor and Francis, Bristol, PA

ICLEI (International Council on Local Environmental Initiatives) 1996 *The Local Agenda 21 Planning Guide: An Introduction to Sustainable Development Planning*. ICLEI, Toronto, ON

ICLEI (International Council on Local Environmental Initiatives) 1997 *Local Agenda 21 Survey*. ICLEI, Toronto, ON

Johnson E 1993 *Avoiding the Collision of Cities and Cars*. American Academy of Arts and Sciences with the Aspen Institute, Chicago

Korten D C 1990 *Getting to the 21st Century: Voluntary Action and the Global Agenda*. Kumarian, Bloomfield, CT

Korten D C 1995 *When Corporations Rule the World*. Kumarian, Bloomfield, CT

Leitmann J 1999 *Sustaining Cities: Environmental Planning and Management in Urban Design*. McGraw-Hill, New York

Mitlin D, Satterthwaite D 1994 *Cities and Sustainable Development*. International Institute for Environment and Development, London

Morris D 1982 *Self Reliant Cities: Energy and the Transformation of Urban America*. Sierra Club Books, San Francisco

Naess P 1993 Energy use for transport in 22 Nordic towns, NIBR Report No. 2, Norwegian Institute for Urban and Regional Research, Oslo, Norway

Newman P W G, Kenworthy J R 1989 *Cities and Automobile Dependence: An International Sourcebook*. Gower, Aldershot, UK

Newman P W G, Kenworthy 1999 *Sustainability and Cities: Overcoming Automobile Dependence*. Island Press, Washington, DC

Pathways to Sustainability Conference 1997 Proceedings of UN Conference, City of Newcastle, New South Wales, Australia

Pezzoli K 1997 Sustainable development: A transdisciplinary overview of the literature. *Journal of Environmental Planning and Management* **40**(5): 549–74

Radberg J 1995 Termites heap or rural villages? The problems of urban density and sustainability. In: *The European City—Sustaining Urban Quality*. Proceedings of Conference, Copenhagen, April. Ministry of Environment and Energy, Copenhagen, Denmark

Real Estate Research Corporation 1974 *The Costs of Sprawl*, Council of Environmental Quality, Washington, DC

Roseland M. 1998 *Towards Sustainable Communities*. New Society Press, Vancouver, BC

Scheurer J 1999 *An Evaluation of Danish Ecological Housing and Planning*. Danish Building Research Institute, Horsholm, Denmark/ISTP, Murdoch University, Perth, WA

Selman P 1996 *Local Sustainability: Managing and Planning Ecologically Sound Places*. Paul Chapman, London

Slocombe D S 1993 Environmental planning, ecosystem science and ecosystem approaches for integrating environment and development. *Environmental Management* **17**(3): 289–303

Smith M A F, Whitelegg J, Williams N 1998 *Greening the Built Environment*. Earthscan Publications, London

Stren R, White R, Whitney J (eds.) 1992 *Sustainable Cities*. Westview Press, Boulder, CO

STPP 1994 *A Citizens Guide to Transportation Planning and Livable Communities*. Surface Transportation Policy Project, New York

Tjallingii S 1991 The responsible city. Paper presented at the International Federation for Housing and Planning, International Congress, Berlin

Todd J, Tukel G 1981 *Rehabilitating Cities and Towns: Designing for Sustainability*. Planet Drum, San Francisco

Troy P 1996 *The Perils of Urban Consolidation*. The Federation Press, Sydney, NSW

United Nations Center for Human Settlements 1996 *UN Global Review of Human Settlements*. Habitat, Nairobi, Kenya

Wachs M, Crawford M 1991 *The Car and the City: The Automobile, The Built Environment and Daily Urban Life*. Michigan University Press, Ann Arbor, MI

Wackernagel M, Rees W 1996 *Our Ecological Footprint: Reducing Human Impact on the Earth*. New Society Publishers, Philadelphia, PA

Walter R, Arkin L, Crenshaw R (eds.) 1992 *Sustainable Cities: Concepts and Strategies for Eco-City Development*. Eco-Home Media, Los Angeles

World Bank 1996 *Sustainable Transport—Priorities for Policy Reform, Development in Practice*, World Bank, Washington, DC

World Commission on Environment and Development 1987 *Our Common Future*. Oxford University Press, Oxford

Whitelegg J 1993 *Transport for a Sustainable Future: The Case for Europe*. Belhaven Press, London

P. Newman

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Planning, Politics of

Planning—here defined as current efforts to affect future outcomes—is a ubiquitous human activity in both public and private life. Individuals plan when they choose their career paths, build families, and save for their retirements. More formally, both corporate managers and government officials regularly develop detailed plans to guide the future activity of their organizations. No society or time-period has a monopoly on planning. Indeed, examples of planning can be identified in virtually every culture and epoch. Some of the best-known episodes in the ancient literature (e.g., the biblical account of Joseph's grain storage) and the most celebrated public works projects in history (e.g., the building of the pyramids, the Great Wall of China and the eighteenth century reconstruction of Paris) were, fundamentally, exercises in planning.

A major aim of planning is to cope with uncertainty. By delineating precise goals, planners seek to stabilize and shape subsequent conditions. Although (nearly) everyone plans, some do so as a career. In the USA today, planning is a recognized profession with barriers to entry, distinct graduate training programs, and particular methods, modes of discourse and internal status rewards. Many planners work for government agencies or as government contractors.

1. *Planning Defined*

The term 'planning' has no fixed technical meaning in the political science or public administration literatures (Meyerson and Banfield 1955). It refers to activities ranging from socialism to program budgeting to purposive efforts to manage land use, municipal zoning, and regional development. The denials of certain planning theorists notwithstanding, planning is inescapably political because it reflects and reinforces social values (Benveniste 1989). Yet planning and politics may also be in tension with one another (Banfield 1961). At its most heroic, planning seeks to pre-commit major aspects of a society in a particular long-term direction designated by central elites. Yet knowledge is always partial and social conditions inevitably change. Heavy reliance on central planning thus risks reducing opportunities for social learning and partisan mutual adjustment, processes that are essential to a well-functioning polity. To the extent that planning accommodates such vital error-correction and conflict-resolution processes, however, it becomes hard to distinguish from ordinary forms of policymaking (Wildavsky 1979).

At least three (overlapping) dimensions of government planning may be distinguished. First is whether the objective of the plan is to shape the future operations of the overall economy (economic planning) or, alternatively, primarily to affect the government's own internal administrative operations (administrative planning). A second important dimension concerns the scope of the plan—whether the intellectual demands, and aspirations, of the planning effort are large (synoptic or comprehensive planning) or relatively modest (strategic planning). Finally, one can distinguish between plans that rely on coercion to effect change and those that seek to achieve their goals through voluntary compliance (indicative planning).

1.1 *Economic Planning*

Throughout the twentieth century, economic planning has been an important response to the perceived moral failings and practical defects of capitalism in both Europe and North America. Prior to the 1980s, the governments of the former USSR and its Eastern European satellites attempted economic planning on a grand scale, seeking directly to control all means of production (as well as consumption patterns) thorough centralized decision-making (Berliner 1957). For many, the tragic experience of central economic planning stands as a political symbol of the inherent limitations of socialism. More concretely, this experience is a legacy that the peoples of Russia and Eastern Europe are struggling to overcome.

Under Stalin and his successors, central Soviet economic planning was organized hierarchically. The top state planning agency, Gosplan, established the target rate of national economic growth and the allocation of production across industrial sectors and geographic regions. This overall national plan was then broken down by various ministries into directives assigned to local economic units. Finally, managers and engineers at the factory or farm level were required to implement particular instructions. Until the 1960s, the USSR reported strong economic growth. By the 1970s, however, it was apparent that productivity had slowed drastically and that resources were being grossly misallocated. While military facilities could generally obtain whatever they needed, the domestic sector was being starved. Living standards deteriorated. In 1987, President Mikhail Gorbachev announced a policy of Perestroika—restructuring—designed to resuscitate the moribund economy (Heilbroner 1993). But the system collapsed.

Critics argue that the Soviet planned economy failed for two main reasons. The first is inadequate information. Very early on the free market economists Ludwig von Mises and Fredrich Hayek insisted that a socialist system was doomed to fail because central economic planners, lacking the price signals of a market system, would be unable to obtain the information required to promote economic efficiency (Von Mises 1935, Hayek 1948). In fact, economic planners in the Communist world—as Polish economist Oskar Lange anticipated in an important 1938 analysis of socialism—were often able to gain much of

the information they needed by monitoring changes in plant inventory levels (Lange and Taylor 1938). The more important reason for the failure of socialist central planning, it seems, was that actors typically had little or no economic incentive to respond to their directives (Heilbroner 1993).

Although Soviet-style comprehensive economic planning was never accepted, the idea of economic planning has long enjoyed intellectual support in the West, even in the USA. The terrible human costs of the Great Depression weakened elite and popular faith in *laissez faire* capitalism and ultimately led to new forms of government intervention in the economy. Most Western governments assumed responsibility for planning for full employment. In France, 'indicative' planners sought to achieve a voluntary exchange of information among business and labor leaders, thereby allowing the French government to use its fiscal and regulatory tools to shape private investment decisions (Lindblom 1977). In the USA, a President's Council of Economic Advisers was established in 1946 to promote stability and growth. During the 1950s and 1960s, Keynesian fiscal planners actively sought to manipulate the level of consumer demand and macroeconomic output. The large tax cut of 1964 was explicitly justified in Keynesian terms. The era of activist Keynesian economic planning came to an end in the mid 1970s, however, with the simultaneous appearance of high inflation and slow growth (stagflation), which contradicted the assumptions of Keynesian models. While use of automatic stabilizers remained a fixture of fiscal policymaking, US politicians in particular lost confidence in the ability of economic planners to 'fine-tune' macroeconomic policy to the business cycle.

1.2 Administrative Planning

In the West, it is in the realm of public administration that one can identify genuine attempts at synoptic planning. Comprehensive government planning has long been central to normative theories of public administration. The influential French administrator Henri Fayol (1841–1925) argued that a 'good plan' requires clearly defined, written, comprehensive objectives together with the resources necessary to carry out these tasks. In the USA, broadly similar concepts found expression in the scientific and executive management movements of the Progressive and New Deal eras. The normative commitments and pragmatic understandings that emerged from these movements (e.g., 'one-best way' thinking; the emphasis on hierarchy and span of control; the separation of politics from administration) have had a significant impact on USA administrative practice and doctrine. In a classic statement in 1937, Luther Gulick, a member of the New York Bureau of Municipal Research who served on Franklin Roosevelt's famous Brownlow Committee, defined the executive function by the acronym POSDCORB (planning, organizing, staffing, directing, coordination, reporting, and budgeting). Effective planning was thus held to be the crucial first step to the achievement of rationality and technical efficiency in government administration. In the USA, administrative planning enjoyed a revival in the 1960s (Schick 1975). Hundreds of federal and local statutes mandated the formulation of plans for allocating intergovernmental grants, designing cities, and other purposes. The centerpiece of the new administrative planning was PPBS—Planning, Programming, Budgeting Systems—which essentially demanded that federal agencies engage in synoptic planning to promote the public interest. Administrators were to (a) articulate precise program goals, (b) comprehensively list alternative ways of achieving these goals, and (c) analyze the costs and benefits of each significant option. PPBS was introduced in the Pentagon in 1961 by Defense Secretary Robert McNamara. By 1965, President Lyndon Johnson ordered PPBS to be extended to nearly all federal, civilian departments. The promise of PPBS was considerable: Planning and policy analysis would be used to increase administrative effectiveness. Defective programs would be found wanting and, it was hoped, replaced with superior alternatives. The actual results of PPBS were disappointing to say the least, however. The reporting and monitoring requirements of the planning effort created a paperwork burden that quickly overwhelmed government officials. PPBS performed especially badly where, as was the case in many domestic agencies, bureaucratic outputs were soft and hard-to-measure. Most importantly, PPBS failed to accommodate the political and organizational incentives of budget actors. The Nixon Administration formally abolished PPBS in 1971 (Knott and Miller 1987, Kettl 1992). The failure of PPBS lends support to Aaron Wildavsky's argument that a fragmented and seemingly uncoordinated budget process generally does a better job of allocating public resources than does a system in which central planners make a synoptic, integrated evaluation of budget alternatives (Wildavsky 1964). Despite the manifest failures of PPBS, similar, comprehensive-planning exercises were adopted subsequently by US governments (e.g., zero-base budgeting and management-by-objectives).

Given the severe limits of synoptic planning, why have so many governments persisted with these sweeping reform efforts? Probable reasons include: there may be a political benefit to appearing 'rational'; planning exercises may enable elected officials to give respectability to things they wish to do for other reasons; and focusing on planning procedures is often easier than making substantive decisions (Knott and Miller 1987). The enduring interest in synoptic government planning may actually rest less on science and rationality than on a secular faith in the perfectibility of man (Wildavsky 1973). Indeed, in an important account of paradigmatic failed attempts at 'state

simplification' (e.g., the Great Leap Forward in China, collectivization in Russia, compulsory villagization in Tanzania, Mozambique, and Ethiopia), political anthropologist James Scott argues that the hegemonic planning mentality is supremely unscientific in the sense of being uncritical, unskeptical, and contemptuous of the importance of practical, local knowledge and informal processes in the face of unpredictability (Scott 1998).

2. Limits and Possibilities of Planning

Although synoptism has failed everywhere it has been tried, more limited forms of government planning clearly have succeeded. The Controlled Materials Plan of the USA in World War Two, the American Social Security program, the massive interstate highway system—each of these initiatives was bureaucratically 'planned.' Long-term goals were articulated, the stated objectives were (more or less) achieved, and the political outcomes were, if hardly uncontroversial, at least defensible. Planning works best when it is highly bounded, is focused on a specific problem, and is economical of analytical talent (Lindblom 1965, 1977, see also Devons 1950). Thus, like the formal hierarchies through which it tends to be implemented, planning has its organizational place (Alexander 1992). The challenge is to keep it there—and to render the practice and results of planning compatible with processes of social interaction.

See also: Patriotism

Bibliography

Alexander E R 1992 A transaction cost theory of planning. *Journal of the American Planning Association* **58**: 190–200

Banfield E 1961 *Political Influence*. Free Press, Glencoe, IL

Benveniste G 1989 *Mastering the Politics of Planning*. Jossey-Bass, San Francisco

Berliner J 1957 *Factory and Manager in the USSR*. Harvard University Press, Cambridge, MA

Devons E 1950 *Planning in Practice*. Cambridge University Press, Cambridge, UK

Gulick L, Urwick L 1937 *Papers on the Science of Administration*. Columbia University Press, New York

Hayek F A 1948 Socialist economic calculation: The present state of the debate. In: Hayek F A (ed.) *Individualism and Economic Order*. University of Chicago Press, Chicago

Heilbroner R 1993 Socialism. In: Henderson D R (ed.) *The Fortune Encyclopedia of Economics*. Warner, New York

Kettl D F 1992 *Deficit Politics: Public Budgeting in Its Institutional and Historical Context*. Macmillan Publishing, New York

Knott J H, Miller G J 1987 *Reforming Bureaucracy: The Politics of Institutional Choice*. Prentice-Hall, Englewood Cliffs, NJ

Lange O, Taylor F 1938 *On the Economic Theory of Socialism*. University of Minnesota Press, Minneapolis, MN

Lindblom C E 1965 *Intelligence of Democracy*. Free Press, New York

Lindblom C E 1977 *Politics and Markets*. Basic Books, New York

Meyerson M, Banfield E C 1955 *Politics, Planning and the Public Interest*. The Free Press, Glencoe, IL

Schick A 1975 The trauma of politics: Public administration in the sixties. In: Mosher F (ed.) *American Public Administration: Past, Present, Future*, University of Alabama Press, AL

Scott J 1998 *Seeing Like A State*. Yale University Press, New Haven, CT

Von Mises L 1935 Economic calculation in the Socialist Commonwealth. In: Hayek F A (ed.) *Collectivist Economic Planning*, G. Routledge and Sons, London

Wildavsky A B 1964 *The Politics of the Budgetary Process*. Little, Brown, Boston

Wildavsky A B 1973 If planning is everything, maybe it's nothing. *Policy Sciences* **4**: 127–53

Wildavsky A B 1979 Between planning and politics: Intellect vs. interaction as analysis. In: Wildavsky A B (ed.) *Speaking Truth to Power*. Little, Brown, Boston, Chap. 5, pp. 114–40

E. M. Patashnik

Planning Theory: Interaction with Institutional Contexts

During the twentieth century, the concept and practice of what is variously called town design, spatial planning, territorial management, and town and country planning developed into a complex area of public policy. The theory informing this policy field has been the product of continual interaction between ideas about how to do planning, the experience of planning in practice, the intellectual context available to those thinking about planning, and the agenda of problems and opportunities which planning has been harnessed to address. Inherently, the planning enterprise is caught up in institutional contexts, which shape its meaning, its material expression, and its consequences. This contribution illustrates the interaction between ideas about the nature, purpose, and method of planning, the context of their production and the practices at which they have been directed.

1. Cities, Governance, and Planning

Few of the ideas about planning which evolved in the twentieth century came initially out of academia. There was no significant academic infrastructure for the field until the last quarter of the twentieth century. The origins of twentieth century planning ideas lay in the experience of rapidly urbanizing societies, and the perceptions of injustice, disorganization and threats to human health and happiness which urbanization seemed to bring. These promoted political movements, pressure group politics, expert evangelism, and attempts to transform city government to cope with the complexity of issues which arise in large cities. The focus of attention ranged from the neighborhood, the city, the region, the process of governance and the

overall organization of society. Two core ideas have informed these otherwise very diffuse endeavors: firstly, that through appropriate governance action, it is possible to 'make a difference,' especially in making cities and societies more sustainable and just; secondly, that the spatiality of social processes and the qualities of place are important dimensions of well-being.

These two ideas have been developed in very different ways as illustrated below, with the aim of highlighting both the situatedness of planning ideas and how their impacts are highly dependent on the specifics of institutional milieu.

1.1 The Good City

In the early twentieth century, planning was generally understood as focused on cities, on how to manage the congested and sprawling consequences of the urbanization arising from industrialisation and new transport technologies (Mumford 1961). Throughout the century, there was concern about the qualities of cities as living places, as aesthetic products, as business locales, and as biospheric relations. 'Theories' of urban form and the arrangement of regional space dominated the discourses of city planning until at least the 1960s (Boyer 1983, Hall 1988). Protagonists, a mixture of campaigners and specialists in urban development (architects, engineers, and surveyors), battled over the relative merits of different forms. By the 1950s, most European states had some form of urban planning and land use regulation embedded in their central and local government policy frameworks and competencies.

This success in institutionalizing the idea of city planning was soon overtaken by vigorous critique of the predominant focus on urban form. Sociologists such as Herbert Gans in the USA argued that cities should be understood through the way people lived in them rather than through their physical form. Jane Jacobs, a fine observer of the way people used urban neighborhoods, argued against planners' ideas of demolishing old neighborhoods to replace them with modern buildings and the separation of living and working spaces. Regional economists and geographers challenged assumptions about the dynamics of urban growth embodied in the urban plans produced by architects, engineers, and surveyors. The emphasis of the earlier generation of town planners on achieving urban quality of life for the mass of the people was replaced by a concern with distributive justice and the 'socially just city'. This idea emphasized that the poor should not draw the short straw in the distribution of access to urban living space. Peter Hall showed that the containment of British cities by greenbelts, far from achieving gains in quality of life for ordinary citizens, had made housing generally more expensive and led to deteriorating conditions in the UK's inner cities. By the 1970s, Marxist analysts had joined the debate. Manuel Castells showed, in a study of a French city, that the aim of urban planning to improve the quality of life was a mask behind which developers and business interests could pursue strategies for realizing land value gains. This critical story had resonance in other parts of Europe where urbanization was proceeding apace, producing substantial development gains to landowners and developers. It was echoed in examples of urban renewal in the USA.

As a consequence of this critique, at least in Europe, coupled with a slowdown in the pace of urban growth, the idea of large-scale comprehensive planning fell into disrepute. Urban management practices focused on the fine grain, the neighborhood, the site, and the urban renewal area. Attention in the 1970s and 1980s turned to projects and mechanisms. The focus on the physical form of the city was displaced by ideas about appropriate planning processes (see below). Architects, engineers, and surveyors were replaced as a source of expertise for planning by regional economists, policy analysts, and those specially trained as 'planners'. The discussion of 'planning theory' turned away from ideas about what the city as such could be like. Influenced by postmodern ideas in urban design and postmodernism in philosophy, a critique developed which suggested that the whole idea of attempting to 'plan' such a complex phenomenon as a 'city' was oppressive to the multiple realities and possibilities of urban life.

By the end of the twentieth century, a new wave of interest in the quality of cities, rural areas, and urban regions was swelling up in policy arenas and among commentators on planning. The pressure for this came in part from an increasing awareness of new economic and social realities. The recognition of the openness of urban economies to the dynamics of global economic forces raised widespread concern about how to 'position' cities and urban regions in a landscape of competing regions. The appreciation from direct experience of lifestyle diversity, and its consequences for the ways different groups were using the space of an urban region, made people aware that traditional models of self-contained, hierarchically-ordered cities no longer captured the reality of urban life, especially as affected by new technologies (Sandercock 1998, Beauregard and Body-Gendrot 1999). Most influential of all were the ideas associated with the rise of a new environmental consciousness in the last quarter of the twentieth century. This brought with it conceptions of the limits which the natural environment placed on human actions, and of the need to consider transgenerational impacts not just short term ones.

This led to a wave of interest in the 'Sustainable City' (Satterthwaite 1999). For some, this meant minimizing adverse impacts on natural environments, linked to ideas about 'ecological footprints' of urban development, and conceptions of the 'carrying capacity' of environmental systems. For others, it emphasized keeping cities compact and preventing land

being used up in development. Others enlarged the concept to integrate the concern with social justice to that of 'environmental sustainability.' The dominant concept in public policy in much of western Europe by the turn of the twentieth century was 'sustainable development,' which combined the principles of minimizing ecological impacts with ways of continuing to promote economic development. However, it is clear that these two dynamics do not comfortably co-exist with each other. There are also tensions with notions of social justice, the multi-cultural city, and fostering innovative plurality. The reawakened emphasis on urban futures in planning theory thus opens up a complex set of issues that had hardly begun to be developed by the end of the millenium.

These ideas and debates about the quality of the 'good city' and the 'good environment' were shaped by the changing dynamics of urbanization processes, as well as changing expectations and values. In turn, concepts from planning theory have influenced the practices of urban management. Planners' ideas have molded the restructuring of existing urban areas, the design of new towns and settlements, and residential neighborhood design. They have affected the form and location of transport projects. They have shaped the evolution of urban street scenes through the impact of land use regulations. These regulations have in turn influenced the way land and property values are distributed and property development processes have evolved. Planners' ideas achieved this impact by framing the organizing principles and discourses used by governments in developing policy frameworks and organizing competencies for planning 'systems'. However, the organizing principles and discourses of one era become embedded in practices and can come to outlive their usefulness. By the turn of the millenium, in many parts of the world, 'planning' as practiced was identified as 'part of the problem' of governance processes that do not match the aspirations and challenges of contemporary cities and regions.

1.2 Good Government

For the protagonists of early twentieth century city planning, the everyday business of politics and government was 'part of the problem.' They sought to show a way forward by striding out ahead with their plans and designs, apart from this messy world. By mid-century, an alternative idea had taken root. Planning was a way of 'doing government' itself. In Europe, socialist movements promoted the role of planning as an alternative to allocating resources through market processes. In the USA, planning was promoted as a way of improving the quality of local government. Clientelist practices were well developed in many cities, with adverse consequences for both the efficiency and effectiveness of public policy. Bureaucratic management was no better, as 'playing by the rules' led to government actions which did not adapt to the dynamics of changing situations or local specificities. The way forward, strongly advocated in the University of Chicago School of Planning in the 1950s, was to adopt contemporary principles of business management. This meant developing strategic frameworks to articulate the goals of an organization. These could then be translated into policy principles or objectives which individual officials could use as guides as they developed responses appropriate to specific situations. This powerful idea, with its focus on 'means–ends relationships,' was strongly linked to the promotion of democratic government. All the steps justifying a particular policy could be explicitly identified and implemented by logical analysis and scientific evidence. Officials could be held to account in relation to their performance in achieving an organization's goals, while the process of strategy-formation made these goals and their justification transparent to citizens (Faludi 1973, Friedmann 1987).

Initially, it was assumed that the work of 'policy planning' would take place in a world of technical experts, reporting to politicians. As Western democracies have become increasingly educated and pluralistic, distributing power and knowledge ever more widely among business groups, pressure groups, and citizen mobilization movements, this position has been challenged. However, concepts of strategy, of evaluation of policy ideas and performance in relation to objectives, and of the systematic link between policy and implementation had become deeply embedded in expectations of government behavior by the end of the twentieth century. City planning practices have been transformed by these shifts in the way government is performed. Planning theory has both influenced the design of spatial planning processes in many parts of the world and learned from their practice.

At the core of the original idea of a 'rational planning process,' as it was often called in the 1960s and 1970s, was the view that governments—and their teams of experts—needed to be able to explain why a particular strategy was appropriate in a particular context, how it would be implemented, and whether it had been implemented. It became linked to the development of ways of making strategic and policy plans, techniques of evaluating alternative options, appraising the impacts of projects, the development of indicators to evaluate performance, and attempts at modeling situations—for example land use and transport interactions—to assess the likely impacts of different courses of action. All these are still in common use in twenty-first century urban and regional government, providing a methodological kitbag for planning office staff and consultants.

However, these early ideas attracted criticism almost as soon as they were propounded. The overall approach was challenged as deeply instrumentalist—defining the purpose of government in terms of achieving specific objectives effectively. The meth-

odologies drew on welfare economics, with its utilitarian assumptions. The approach assumed that policy ideas were translated into action in a linear way, from 'top' to 'bottom' of an organization or government system. It was seen to be over-generalized, failing to consider the differences, which might arise in different policy fields and in different institutional contexts. It also handed power to the experts, as the early city planning ideas had done. Politicians, lobbyists, and citizens were pushed to the sidelines. In the 1950s, a famous study by Meyerson and Banfield in Chicago showed that local politicians were not prepared to accept the restraints on their actions, which the rational policy-planning model implied. In the 1960s and 1970s, scholars such as Charles Lindblom, Amitai Etzioni, John Dyckman, and Richard Bolan began to show how contexts might limit how far a policy-driven planning practice might go. Others criticized the approach as promoting 'corporatist' practices, entrenching the influence of local business and political elites through their power to shape and overturn strategies. An alternative was to accept the increasingly pluralistic form of local politics, and develop a local politics around competing plans, produced by different interest groups. This approach, promoted by Paul Davidoff, became known as advocacy planning (Faludi 1973). The experience of policy practice began to look less and less like a strategy-driven process. Instead, analysts noted that the content of policy seemed to be negotiated in the flow of action, rather than being set in advance.

By the late 1970s, planning theorists turned their attention to issues of policy 'failure' and implementation 'gaps,' following landmark studies by Aaron Wildavsky and others. This work looked more closely at the institutional context in which policy frameworks were articulated and the ways planners, understood as experts and government officials, went about their tasks. It identified the problems, which arose where policy ideas were not sufficiently 'owned' or understood by politicians, or by other parts of public administration which controlled significant resources or regulatory powers. There were close parallels with work undertaken in business studies. This analysis highlighted the difficulty of promoting strategy-driven governance processes where other ways of 'doing governance' controlled the culture of local government. It also showed that 'field level' actors, officials, and experts working up the details of policies, or acting at the interface between officials and the clients of public administration, developed their own ways of understanding policies and how they might be applied.

This led to a critique of the rational planning process as a 'top down' viewpoint, challenged by a 'bottom up' viewpoint of the actors 'on the ground.' (Barrett and Fudge 1981) The linear model of the policy process was increasingly challenged by conceptions of policy development as an interactive process, continuously evolving over time and drawing in different actors at different stages. In this formulation, planning was no longer either outside government, nor imposed on governments in an attempt to transform their processes. It was caught up 'inside' politics and governance processes. In the 1980s, a significant strand of work associated with John Forester, Charles Hoch, and Howell Baum, began to address questions of the way planners committed to progressive values operated in these ethically challenging situations (Forester 1989, Campbell and Fainstein 1996).

During the 1980s, a major conceptual reorientation brought new ideas about the nature of policy processes to a central position in debates in planning theory. These built on the earlier work on 'implementation processes' in two ways. Some analysts focused on the way experts and officials made sense of the complex situations in which they found themselves. Others began to explore the way policy agendas and the relations of governance mutually constituted each other in interactive processes. Both emphasized the significance of the frames of reference and the 'social worlds' which those involved in policy arenas brought to bear on policy design and implementation. Policy development was presented less as a technical exercise in calculation and design, and more as a social process of articulating ways of thinking and ways of acting which could command support among a wide range of players in urban and regional contexts where power was likely to be diffused in complex ways rather than concentrated in the hands of a single elite. Collaborative, participative forms thus challenged top-down and 'producer-driven' approaches to public policy. Such ideas were infused with the much wider critiques of positivist science and linear models of knowledge and application developing in the 1980s in the social sciences generally. Some argued that planning was an impossible task, caught up in the discarded clothes of 'modernity.' Others, and notably John Bryson, Judith Innes, and Maarten Hajer drew on emerging interpretive approaches. These emphasized that policy meanings were actively constructed by participants through communicative processes which, if effective, could recaste policy 'discourses' and redirect resource allocation processes and regulatory practices (Healy 1997, Innes 1995).

In parallel, there was an explosion of attempts in various parts of the world to develop new, more participative and collaborative forms in planning practice. These were partly promoted by the idealism of those pursuing a democratic version of the sustainability agenda, notably in the practices of producing Environmental Audits and Appraisals for localities, as encouraged by the World Environment Congress in Rio in 1992. But they were also a manifestation of a struggle to recast the forms of governance inherited from mid-century in Europe and the US. While neoliberal strategies sought a way forward by reducing the burden of government, through strategies of 'privatization' and 'deregulation,' alternative strategies pro-

moted models of more participative democracy, which could reduce the perceived gap between the worlds of politics and government and the worlds of citizens. The objective was to refocus governance activity around the concerns of civil society and the economy, rather than those of the 'producers' of government.

By the late 1990s, ideas about consensus-building practices and collaborative strategy formation, collected and consolidated by planning theorists, were being rapidly internationally circulated to help fulfill in practical ways the ambitions of politicians and officials trying to move in new directions. In this context, the value of policy effectiveness embodied in the 'rational planning process' was supplemented by a concern with 'building democratic culture' (respectful and 'citizen-centered' styles of governance). Proponents of the new communicative planning argued that strategies that were developed interactively among affected parties (stakeholders) led to greater 'ownership' which in turn provided the social and political resources through which strategic directions could be maintained. They also argued that such processes brought civil society into a much stronger position to shape public policy agendas.

1.3 Qualities of Place and Qualities of Governance

By 2000, there were vigorous debates in the planning field about this new agenda. As in the earlier discussion of the 'rational planning model,' a key issue was the qualities of the institutional context in which such practices could flourish and how far their promotion could induce changes in the context. However, there was much less discussion of how these 'good governance' ideas might link to the qualities of a 'good city.' The two strands of thinking in planning ideas largely proceeded separately in the late twentieth century. They also potentially challenge each other, as the one stresses qualities of process and the other stresses qualities of places and their diverse relations. A governance process informed by the citizens 'of' a locality, which proceeds in a richly participative way, is likely to encounter multiple ideals about a 'good city' and a 'good living place.' What may emerge through processes of debate, struggle, and conflict resolution may be far removed from the ideals of a good city being developed in debates among professionals.

2. Planning Theories in Context

The enterprise of planning is profoundly committed to changing the ways things are and to transforming realities, whether those of cities or their governance. It is inherently normative in its value orientation, infused with the belief that deliberately designed actions today can have beneficial impacts tomorrow. Those who reject the idea that this is desirable or possible—anarchists, extreme free marketers and cultural postmodernists—will reject the enterprise of planning. Yet the idea that, in democratic states, governments should have policies and should be judged on how effectively they follow them through, is one of the great inventions of the twentieth century. It is being carried forward vigorously into the twenty-first century. What are changing, however, are the conceptions of where policies should come from, how they should be focused and who gets involved in their articulation and implementation. At the beginning of the twenty-first century it is not enough for policies to be designed by experts. The quality of government is increasingly judged by the way its policy formation attends to the multiple voices of complex societies, by its skill in recognizing the complexity of economic, social, and environmental dynamics, and its appreciation of the specific realities of the way policy programs are realized.

In this context, a policy focus on 'place quality' has a new salience. Such a focus, if grounded in the specific contingencies of particular places and the multiple perceptions and meanings of those with a stake in them, can connect the issues which concern citizens and businesses in ways which make sense to them. These tendencies recreate a policy environment in which the enterprise of planning has much to contribute, provided it can release itself from its 'entrapment' in the past and reconnect the debates about the qualities of place to those on good governance. Three recurring themes in the discussion of the 'good city' and 'good governance' represent important challenges for future spatial planning and territorial management practices.

2.1 Ideas and Actions

The first challenge focuses on the relation of theory to practice. The history of planning thought shows the many ways in which planning ideas filter into the construction of concepts, goals, and norms which guide specific actions—through advocacy and campaigning which impact on popular perception and politicians' agendas; through the design of planning systems and formal norms; through the ways these are interpreted in administrative and legal arenas; through the formulation of explicit strategic frameworks and the ways these get translated into routinely used concepts and norms; and through the training of 'planners' and the professional organizations which bring 'planners' together. Through continual interaction with these worlds, planning ideas are themselves changed. The contribution of planning thought in the twentieth century has been its preparedness, however ineffectively it is often developed, to link together the

philosophy of normative action, the social science of urban and regional dynamics and the fine grain of the practices of urban and regional governance. The challenge has been to draw out the significant strategic processes of governance which affect the qualities of places and territories, as lived, experienced, and given identity in multiple social processes (Friedmann 1987).

2.2 Substance and Process

The geographer Ed Soja has remarked that twentieth-century thought has had a peculiar difficulty in perceiving the impact of 'being in a place' and of the spatiality of social relations. The autonomous individual in a 'free society,' or the social group as a 'class' in relation to other social classes, has been much more dominant. Contemporary culture, however, notices all kinds of cleavages and fractures, and stresses the complex relations individuals may have with others. Grasping the significance of place and territory in these complex representations and relations, and developing governance processes in which to give them voice, is a major challenge for urban and regional politics and policy. In planning theory, there has long been a debate about how to relate the 'substance' of policy to the 'process' of policy development, with theories being criticized for giving too much weight to one or the other.

If the predictions made in the previous section are correct, then the second challenge for planning theory is to contribute to the co-evolution of the 'substance' and 'process' of territorial management-policy frameworks. This has major implications for institutional design—who contributes what, when, and where. It also has implications for the articulation of values. There were major debates in the mid-twentieth century about whether planning was 'value-free' or 'value-full'. If 'value-full,' were the values about process (fair, just, 'rational'), or about substance (achieving specific goals, for example, aesthetically-pleasing environments, quality of life for ordinary people, redressing the unequal distribution of resources and power etc.)? By the end of the century, process values had been extended to include qualities of inclusion, empowerment, and respect to all citizens, while substance values were infused with ideas about environmental sustainability and social justice. However, if the substance and process of governance for complex territorial problems are seen to co-evolve, then it must be expected that values too co-evolve. This implies the development of critical capacities for 'reflection-in-practice' on what is evolving, what it might lead to and what values might be promoted as a result (Schon 1983). In work in planning theory, this has led to a critique of planning strategies and the design of planning systems. It has also focused on the personal dilemmas of 'doing' planning work.

2.3 System Maintenance or Transformation

The third challenge in planning theory lies in the approach to social change. Behind any idea of the 'good city' or 'good government' is some model of the nature and dynamics of social organization. For some advocates of planning ideas, it is the model itself that needs change. For others, the model needs correction to make it work better. As John Friedmann argued in the 1960s, some planning ideas sought to maintain basic systems. Others aimed to transform them. At the time, he was contrasting the management needed to trouble-shoot in western democracies with the transformations needed in illiberal Latin-American dictatorships, newly experimenting with democracy. By the 1970s, he had become much more radical, demanding 'social transformation' in over-urbanized capitalist societies, dominated by global corporations.

Throughout the twentieth century, people have been attracted to the planning field through transformative idealism. In the second part of the century in Europe, the target of transformation was a form of capitalism which dominated life chances, structured the way urban and regional space was used and damaged environments across the globe. For many, it was a shock to discover that the 'planning systems' and 'planning practices' which actually evolved, seemed deeply committed to maintaining the very practices which they had sought to challenge. Ideas about alternative ways of organizing society, however, come up against the reality of the dynamics of social change. In some periods of tumultuous social change, any kind of planning is an impossibility. In some periods of entrenched stability, similarly there are few cracks of opportunity to mobilize to do things differently. Most of the time, most situations are somewhere in between. In the contemporary world of complex relationships and diffused power, change happens both through big events (for example, the OPEC oil crisis, the collapse of communism, the 1987 property crash, the rise to power of neo-liberal governments, etc.) and through lots of small shifts and mobilization episodes. Planning processes can help these shifts along. Instead of making plans which are assumed to be predictions of what will happen, making plans becomes more of an activity of social mobilization, encouraging new ways of understanding situations and developing new ways of working. Strategy development in such conceptions takes on more of the character of collective invention and risk-taking in trying to shape futures rather than the mid-century assumption that government could command resources and control futures.

3. The Contribution of Planning Theory

These debates in the planning field provide rich resources for imagining, developing, and critiquing governance practices as they impact on urban and regional dynamics in the twenty-first century. Formal-

ized in the planning-theory literature, they range through philosophical debate, through analysis of systems and practices, and through discussion of the situation of 'doing planning work.' If the planning field is to realize its potential to contribute, imaginatively and critically, to the 'turn' towards more place-focused policy agendas, planning theory needs to maintain a close and multi-facetted engagement with these evolving practices.

There are many 'windows' through which this engagement has developed. One is that of the 'planner,' working within specific contexts trying to make things happen differently. This has been the key contribution of John Forester, whose work speaks sympathetically to planners in complex situations around the world (Forester 1989). Another is local institutional practice. In recent years, this work has been closely linked with the disciplines of urban politics and urban and regional geography, exploring questions about the relation between politicians, political regimes, experts and citizens, and the way planning strategies and tools have been used, in particular institutional milieu. A third 'window' has focused on understanding the design of governance systems. This has examined competencies, law, resource allocation systems, and appropriate expertise, the 'top-down' levers of policy change, and how these interact with 'bottom-up' interpretive practices. Much of this work has been informed by the search for ways of changing governance practices. A fourth 'window' has concentrated on ideas about the good city and the good society, especially in relation to ideas of environmental sustainability and multicultural societies and on how these have worked out in practice. The critical challenge for the development of planning as an enterprise is to recognize that these different 'windows' do not represent separate compartments of the planning field. They merely cast a more intense lens over a particular set of relations. All the others are still there, and need to be recognized as present, even if not the focus of a particular study.

The role of planning theory, as a distinctive task in this complex and dynamic institutional terrain, is to provide a critical imagination and structuring ideas for the dynamic invention processes which are continually unfolding in political communities in cities and regions. It should offer evaluative vocabularies for the recording and critiquing of these processes, both in relation to their content and their procedures. It should interpret emergent patterns and assess their consequences. It should draw in ideas from other fields and translate them in ways understandable to the various thought-worlds caught up in city governance. It should consolidate the distinctive understanding evolved by engagement with this activity and make it available to those working in other fields. This implies that the task is always an 'in-between' one. If the 'practitioners' of planning theory withdraw too far from the worlds of those engaged in 'doing planning,' then they lose their capacity to record, shape, connect, and translate. If they stay too close to the 'doers,' they are unable to see beyond what the doers themselves see, and therefore cannot 'add value' to the work of doing. The challenge for planning theory in the twenty-first century is thus to enrich the evolution of public policy focused on place quality and to do this in a way which fosters the flow of ideas between the many arenas engaged in shaping urban and regional governance.

See also: Advocacy and Equity Planning; Development and Urbanization; Landscape Architecture; Market and Nonmarket Allocation; Metropolitan Growth and Change: International Perspectives; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Planning Theory (North American); Population Pressure, Resources, and the Environment: Industrialized World; Rural Planning: General; Strategic Planning; Sustainable Development; Urban Growth Models; Urban Places, Planning for the Use of: Design Guide; Urban Planning: Methods and Technologies; Urban Planning (Western), History of

Bibliography

Barrett S, Fudge C 1981 *Policy and Action*. Methuen, London

Beauregard R A, Body-Gendrot S 1999 *The Urban Moment*. Sage, Thousand Oaks, CA

Boyer C 1983 *Dreaming the Rational City*. MIT Press, Boston, MA

Campbell S, Fainstein S 1996 *A Reader in Urban Theory*. Blackwell, Oxford

Faludi A 1973 *A Reader in Planning Theory*. Pergamon, Oxford

Forester J 1989 *Planning in the face of power*. University of California Press, Berkeley, CA

Friedmann J 1987 *Planning in the Public Domain: From Knowledge to Action*. Princeton University Press, NJ

Hall P 1988 *Cities of Tomorrow*. Blackwell, Oxford

Healey P 1997 *Collaborative Planning: Shaping Places in Fragmented Societies*. UBC Press, Vancouver

Innes J 1995 Planning theory's emerging paradigm. *Journal of Planning Education and Research* **14**: 183–9

Mumford L 1961 *The City in History*. Harcourt, New York

Sandercock L 1998 *Towards Cosmopolis*. Wiley, London

Satterthwaite D (ed.) 1999 *Earthscan Reader in Sustainable Cities*. Earthscan, London

Schon D 1983 *The Reflective Practitioner*. Basic Books, New York

P. Healey

Planning Theory (North American)

1. The Invention of Planning Theory

In the quarter of a century prior to World War I a small international movement centered in the North

Atlantic Basin sought to establish 'city planning' (under one or another of its names) as a professional craft and an imaginative sensibility. The movement cast itself in a heroic role. The new practice of planning, speakers at the founding conferences insisted, would correct and tame (what they represented as) the 'wild' and essentially 'unplanned' urban growth of the nineteenth century. The new practice would infuse the fragmented work of professional architects, social workers, hygienists, and civil engineers with an integrative civic vision. The image of the well-ordered city—symbolized by a plan that would guide its growth 'tomorrow'—was presented as a global icon of modernity. Chicago, New York, London, Birmingham, Paris, Berlin, and Vienna would learn from one another and tutor Tokyo, Shanghai, Cairo, Delhi, Colombo, and Rio de Janeiro (Sutcliffe 1981).

Despite their expansive aspirations, the leaders of the planning movement—both before and after World War I—built their professional careers within a conception of cities as well-bounded spaces filled with physical structures that organized 'public' activities. (The founders argued among themselves: Were houses public or private, and were they appropriately part of the planning brief?) Through the first half of the twentieth century the 'theory of town planning' was articulated largely in speculative proposals for new spatial forms, in the tacit assumptions of working plans, and in textbooks legitimating the 'rational' integration of 'architecture and urbanism.' The movement did not develop a distinctive theorizing tradition assessing the connection between the design of planning practices, processes or institutions, and the character of plans; between the character of plans and of the worlds they shaped (Choay 1997).

In the second half of the twentieth century, all of that changed—though not everywhere in quite the same way, and certainly not in an instant. The demand for professional planners grew dramatically after World War II, and universities responded by creating academic programs. Where planning education remained an adjunct of engineering or architecture, procedural theories played a relatively small role in shaping the new curricula or the public statements of the professoriate. Where, however, planning was articulated as a distinctive craft, the success or failure of 'its theory' were vital elements of the new professional identity. ('Professions,' it was widely claimed, 'needed a theory.')

That 'need' did not, of course, sweep away the intellectual difficulties that made the construction of procedural planning theories so difficult. The familiar forms of social theorizing were (and continue to be) designed to explain repeated historical behaviors, or to forecast the path of systems that we cannot, or do not choose to, change. In contrast, planning processes in their varied forms attempt to control (rather than to explain) a world that does not yet—and may never—exist. No matter how sophisticated, planners find themselves torn repeatedly by the discovery that when, imaginatively, they unfreeze relations in order to change them it becomes harder to anticipate the outcomes of their actions, mistakes are legion, and the burdens of control grow heavy. The internal intellectual history of the attempt to construct and validate theories of planning in the last half of the twentieth century is centered in that dilemma.

The builders of the epistemic community devoted to planning theory created a full array of disciplinary artifacts: Journals, conferences, anthologies, review articles, 'fields' for doctoral students to master, careers, and stylized self-criticisms. Inevitably, favored failings were part of the community's identity. Insiders worried that their enterprise was too theoretical or not theoretical enough; that it ignored the spatial core of urban planning in order to run after foreign gods, or that it was too insular in its close attention to planning professionals.

Some of the builders—both early and late—imagined that they might in time satisfy the aspirations of practitioners and educators for a coherent and cogent theory that would guide and legitimate professional city and regional planning. At mid-century, the small Program for Education and Research in Planning at the University of Chicago, radiating confidence in the promise of rational planning, achieved an abiding place in the mythology of the planning Academy despite—or perhaps because of—its premature demise. By the end of the century, however, that promise seemed to most members of the community of theorists either misguided in principle or unachievable in practice. At its conferences and in its journals, the community acted as a loosely coupled network of distinctive and often contending groups. Planning theory—in the singular—appeared as an ironic misrepresentation of the varied products of those distinctly plural circles.

Individual theorists have not, of course, been immune from the arrogance of expertise. The procedural authority of the community of theorists has, however, been extremely limited in most political regimes. Unlike architects and engineers, theorists have never been certified by public or quasi-public authorities. Nowhere is the approval of a board of theorists required to validate procedural designs or to approve the sale of malpractice insurance. Theorists are not legally responsible for flawed organizational proposals, incoherent arguments, or the misreading of texts. They do, of course, hold each other accountable for simple foolishness, complex confusions, and the corruption of novices but their (loose) mutual discipline cannot be more demanding than the structure of the field will allow.

Perhaps because of that freedom from responsibility, the theorists have been intellectually ambitious. Undeterred by ordinary scholarly boundaries and without asking anyone's permission, they have extended the privileges of membership in their com-

munity to 'colleagues'—dead and alive—who had little direct connection with city and regional planning, and who rarely noticed the appropriation of their work. Sins of omission play a more important role in the stylized self-criticism of the community than sins of commission. A claim that a voice has been neglected commands the community's attention in the way that a broken tooth commands an inquiring tongue.

Whatever their initial aspirations, the theorists have not created a confident science of planning. To those who read, they have, however, suggested both alternative planning modes and subtle similarities in overtly different circumstances (Krieger 1981, Verma 1998). They have sustained an integrative bridge across diverse domains—land use, environmental and social policy, the design of policies and the associations of civil society, and local and regional economic development—that stretch far beyond the conventional scope of urbanism. Planning theory at the end of the twentieth century appeared as a counterbalance to the authority of local practices, legal mandates, urban regimes, technical expertise, and settled professional doctrines. Those were, however, powerful enemies. It is little wonder that theorists—offering at their best only criticism, questions, and suggestive intimations—lived at the margins of the professional planning enterprise with its characteristic and quite sensible respect for tacit knowledge and uniform practices. Within the main body of professional practitioners, theorists prospered only when they created formal protocols for representation, design, evaluation, and negotiation that could be replicated in many settings and invested with political authority. Geographic Information Systems (now inflated to Geographic Information Science) were the rage of planning agencies at the end of the century. GIS were, however, only the latest in a long series of tools—first among them, cost/benefit analysis—whose substantial philosophic baggage is characteristically left at the agency door (Pickles 1995).

2. *The Platform of Planning Theory*

The conception of 'architecture and urbanism' that was displaced by planning theory had integrated four assumptions:

(a) the object of planning was the physical form of a settlement;

(b) designers could specify the ideal, rational form for a settlement in a particular site;

(c) rational forms had wholly predictable and appropriate outcomes; and

(d) rational form should dominate other design preferences or proclivities.

The critical challenge to these assumptions was articulated on the one side in a direct assault on geographic determinism and the cogency of architectural conceptions of urban form; and on the other side, by an almost casual dismissal of the 'dream of a rational city.' The texts that codified theories of spatial design and the debates of the first half of the century virtually disappeared from the 'talk' of the emergent community of theorists. Nigel Taylor, in his history, *Urban Planning Theory Since 1945* (1998) lists five mid-century texts in the 'physicalist' tradition: none of them was included in the bibliography of Andreas Faludi, *Planning Theory* (1973b), Kevin Lynch, *Good City Form* (1983), or John Friedmann, *Planning in the Public Domain* (1987). Faludi (1973a, p. x) characterized his own frequent use of physical planning examples as 'coincidental, a mere reflection of the fact that is the field where I come from and in which I teach.' Friedmann did not include 'urbanism' in his graphic presentation of distinctive planning traditions since the great revolutions of the eighteenth century. New recruits into the community of theorists should (in his account) place themselves in one of those traditions as heirs of an Enlightenment belief that scientific knowledge can and ought to be applied effectively to social improvement.

The combination of a frontal and a 'casual' challenge to the integrative conception of architecture and urbanism created an intellectual space in which it was possible to rethink the practices of architecture and urban design, and to create a platform for the construction of planning theory as a field of inquiry and reflection. The major premise of that platform was that though planning processes, settings, and outcomes were entangled, it was useful to distinguish between them. Within the framework of that heuristic distinction, theories of planning should articulate and explain the attributes of each element and illuminate the relations among them: If planners—lay or professional—act in a particular way in settings of such and such attributes, what are the likely outcomes? If they seek outcomes of a particular sort in a given setting, how should they plan? What sort of outcomes should they seek, and should their choices vary from one setting to the next? The intellectual history of planning theory is usefully understood as a tale of a community learning to live on that platform—a stage of its own making. (The best way to follow this tale is probably through the series of anthologies cited in the bibliography.)

The practitioners who shaped the early-twentieth-century planning movement embraced an ambitious conception of planning as a process—and as a profession—that was uniquely rational, scientific, or self-conscious; that looked into the future and across conventional institutional domains; and that served the public interest or other comparably admirable objectives. The emergence of the new planning academy after World War II did not dim that heroic self-image. The rationality of urban form was, however, replaced by the intellectual claims of a toolkit of protocols for representing complex systems and their cybernetic controls, for evaluating projects and

programs, for modeling urban land use and transportation patterns, and for revealing the dynamics of location, spatial differentiation, and interpersonal negotiations. Those protocols—*inter alia* systems theory, experimental design, mathematical simulation, factorial ecologies, and game theory—had a profound impact on the first postwar generation of theorists, displacing the integrative conception of urbanism but retaining an essentially rationalist justification for professional planning. The mission of theory was not to describe or explain planning processes but to shape them so that they were ethically compelling and instrumentally effective (or even efficient). Normative theorizing would embed those demanding criteria in the identities of individual planners, the practices of planning agencies, and the norms of a disciplined profession. (For exemplary statements of that rationalist ethos, see Chadwick 1971 and Branch 1990.)

Those ambitious rationalist protocols and aspirations did not control the platform for very long before they evoked a critical counterattack on the 'new utopians' and the pathologies of technocracy (Boguslaw 1965). Some of these attacks came from theorists who had never been true believers in an age of uniquely rational planning. Perhaps more interesting, were subtle changes in the thought of rationalists (such as Melvin Webber or Britton Harris) responding to the difficulties of managing the indeterminate complexity of 'wicked' problems, and of bringing knowledge to bear upon even simple planning issues when ordinary words carried so many conflicting meanings (Rittel and Webber 1973, Vettoretto 1997, Harris 1989, 1997). Both internal and external voices challenged the assertion that a corps of professional planners could articulate and defend a unique social welfare function, or that the difficulties of collective action could be resolved consensually by privileged institutions or tools. Accounts of urban and regional planning activities that dealt only with the behavior of professional planners obscured the ways in which a great many agents attempted to plan, playing both with and against one another. It may be useful to represent planning as a system, but neither the governor in that system nor the feedback loop need appear as a professional planner or agency. In many settings, governance is the work of a political regime, a market, or an interorganizational field. In Charles Lindblom's account, the 'intelligence of democracy' depended on the design of interaction as well as the practice of analysis. Theories addressed only to the creation and maintenance of an expert profession of analysts could not explain adequately the dynamics of those interactive processes, institutions, and practices. Nor could they provide compelling support for their redesign (Lindblom1965, 1990).

The critique of the rationalist toolkit was extended by a reinterpretation of reason and rationality. These ideas, circa 1950, justified claims as fitting within a coherent world view; and as cogently linking a chain of assertions to core propositions without violating legitimate rules of discourse. By 1975—or even earlier—the implications of these criteria had been altered substantially by the recognition of multiple worlds, discursive formations, and pragmatic tests of truth claims.

Paradoxically, the shift was signaled most clearly by the simultaneous publication in 1973 of Andreas Faludi's anthology celebrating the achievements of postwar planning theorists, and his own treatise synthesizing these achievements in an account of three polar modes in which intelligent professional agencies sought to enhance human growth: blueprint vs. process, rational comprehensive vs. disjointed incrementalist, and normative vs. functional planning. By articulating these polarities as foundations of a positive theory of planning, Faludi revealed how little the community of theorists knew of the variability of these modes over time; and of the ways in which the field of planning organizations and institutions responded to the attributes of the settings upon which it acted. The description of three polarities virtually begged for additional options that would—and that did—enrich the planning repertoire, but that also made it impossible to cultivate a complete general theory of planning. Though Faludi's description of 'human growth' as the goal of planning had a distinguished intellectual pedigree, it was obscure and (in the event) quickly dominated by the intense interest of theorists in John Rawls' (1971) conception of 'justice as fairness.' The theorists, though deeply engaged by the philosophic disputes surrounding *A Theory of Justice* (and subsequently by Rawls' responses to his critics), did not, however, agree unanimously that justice was a superordinate goal that should guide all planning choices, that the conditions of the original position would yield a rational consensus on Rawlsian principles, or that the distinctions between rights and goods (and thick and thin) were compelling.

In practice, reason and rationality had always spoken with diverse voices. In the last quarter of the twentieth century, they were intractably pluralistic in theory. Pluralism implied that reason and rationality were neither permanent nor universal criteria. Encoded in planning doctrines, they were precious but temporary and local public orders. In different ways, both Richard Rorty and Jurgen Habermas—as they were coopted by the community of planning theorists— defined the planning 'problem' as the design of legitimate 'conversations.' Closer to the core of the community, the fascination with 'participation' as a remedy for the putative flaws of both autocratic and representative institutions privileged the voices of engaged advocates rather than heroic rationalists.

Without the taming influence of a commanding conception of rationality, the pluralist expectation encouraged an increasingly complex conception of the architecture of the intellectual field. In both positive

and normative theorizing the distinctions between process, setting, and outcome could be drawn in many different ways. A development coalition committed to promoting growth might be treated as a setting permitting a comparative study of professional planning processes in an array of urban regimes. A theory trying to explain differences in development patterns might, however, represent the coalition as the most important dimension of the urban planning process. The theorist in such a field is a designer adjusting and readjusting the three elements in order to find a process that can deal with a setting, a setting that can sustain a process, and a conception of outcomes that will not shatter a rightly cherished world. The theorist as critical reader unpacks that complex judgment to test its credibility and its delight: Is that a world worth making?

By the mid-1980s many members of the community of theorists had accepted an image of their field as a lively but not very tidy patchwork—wall *bricolage* was the fashionable descriptive image. As theorists learned to create sensitive ethnographies of practice, to test the cogency of procedural typologies, and to appreciate the disjunction of what Thomas Schelling (1979) described as 'micromotives and macrobehavior,' the harder it became to make sense of the wall. Even the most elaborate plans—symbols of the authority and heroic aspirations of practitioners—were understood by theorists as vulnerable to the (highly variable) interpretations of uncontrollable readers and the secret influences embedded in the everyday language of planners. In Judith Innes' (1990) report of her own conversion, the new 'interpretive or phenomenological view of knowledge' was 'more contextual, more evolutionary, and more complex' than the positivist scientific model in which she had cast her 1975 account of social indicators.

That was not, however, the end of the tale. Whatever their view of the political roles and possibilities of professional planning, theorists worried lest the literary forms of their arguments alienate both students and experienced practitioners; that it would encourage skepticism, and an insidious delight in uncertainty. They fantasized about a postmodern 'abyss,' in which both practical confidence and public meaning were lost.

The community struggled with those fears in the last two decades of the twentieth century. The attempt to control a pluralism that was deeply treasured took at least four different forms—some new and some part of its initial endowment. The first responsive form, as in the complementary work of Ernest R. Alexander (1996) and Andreas Faludi (1996), accepts the difficulty of articulating and defending a complete general theory of planning. It insists, however, that the patchwork of typologies, ethical claims, and empirical findings can be integrated by a robust planning doctrine into a series of contingent protocols: Within the frame of such a doctrine, if a situation changes from A to B, there is a shift from the now inappropriate mode to another that better fits the current setting. These protocols are more complex than a single theory but not so difficult (or so numerous) that they deny the possibility of achieving consensual agreements on the nature of processes, settings, and outcomes, and their relationships.

The second response addresses pluralism and the patchwork of theory in a quite different way. Surely, the worlds we plan are remarkably varied and often the devil is in the details of those differences. Nevertheless, the advocates of this response—such as David Harvey (1985)—insist, there are really only a small number of compelling ways to describe and explain planning processes, institutions, and practices; only a small number of compelling ethical systems and principled social conflicts. We misrepresent the condition of planning theory, they insist, when we fail to recognize the ways in which the great global systems—e.g., capitalism, statism, democracy, colonialism, racism, liberalism—are defended by the construction of the illusion of huge variations and the exaggerated promise of plasticity.

The third response divides the patchwork of theory into two parts. One—planning proper—is subject to all of the familiar perils of indeterminacy, complexity, and interpretive variety. The second part is the domain of planning support systems that tame those perils (even if they cannot eliminate them) by establishing consensual representations of 'fact' and 'structure,' allowing us to elaborate the implications of our own images and plans, and stabilizing public orders (Klosterman 1997). Serious support requires, of course, that the choice of representational modes and data sets address the protean requirements of planning processes. The support we seek when we are encouraging divergent thinking early in a planning process would be disruptive or inadequate as we fill in the final details of a fully elaborated scheme. Data sets or maps that cloak essentially contested meanings in a formal garb cannot transform (or support) planning debates: They only reproduce them. Strategic planning *may* not be well-served by the data that regulates operational management; or organizations in a turbulent environment by the information practices of a placid field (Emery and Trist 1965).

Finally—in what is often represented by its critics as the dominant school within the community (Healey 1997, Yiftachel 1999, Throgmorton 1999)—the communication strategies that have accentuated the pluralism of both planning and planning theory have been engaged to limit the perils of pluralism. Initially, the project of communicative enlargement served principally to encourage the voices of silent groups and to 'pay attention' to their lives, to democratize access to knowledge, and to multiple the opportunities to be heard. This project is by no means completed to the satisfaction of its partisans, but it has now taken a somewhat different turn. How is it, many theorists

(e.g., Baum 1997, Forester 1999, Mandelbaum 2000) wonder, that we sometimes manage to understand and change one another, and to resolve moral conflicts? What sorts of rhetorical practices and what sorts of settings sustain a practice of reflection, successful mediation, cooperation within and between organizations, and mutual trust? How—even more profoundly—may strangers learn to deliberate across the ordinary divides of adversarial democracies? What does 'justice' require of our speech and of the distribution of knowledge?

The current theorizing about communication in planning is capacious. It is not only about talk and texts, but about the ways in which symbolic meanings are imputed to actions and symbols are acts; about the ways in which arenas, forums, and courts (Bryson and Crosby 1992) shape communication patterns. The focus on communication accommodates ethnographies of planners at work, close studies of the dynamics of communities as moral associations, critical assessments of the networks that support the 'new information technologies,' civil societies and political regimes at every level, and the detailed protocols (Friend and Hickling 1997) that support the work of complex task forces. Inevitably, the links between these diverse intellectual forms sometimes seem very fragile, and the communication explanation and remedy tendentious. The power of the communication approach lies in the ways in which it makes sense of the diversity that theorists discover as they engage one another on the shared platform. Though theorists have not articulated a consensual set of findings or shared uncertainties, they have found a way to talk about their disagreements, to identify themselves within intellectual traditions, and to map prospective bridges across contested grounds.

See also: Architecture; Community, Social Contexts of; Development Theory in Geography; Planning Issues and Sustainable Development; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Urban Growth Models; Urban Places, Planning for the Use of: Design Guide

Bibliography

Alexander A 1996 After rationality: Towards a contingency theory for planning. In: Mandelbaum S J, Mazza L, Burchell R (eds.) *Explorations in Planning Theory*. Center for Urban Policy Research, New Brunswick, NJ, pp. 45–64

Baum H S 1997 *The Organization of Hope: Communities Planning Themselves*. State University of New York Press, Albany, NY

Boguslaw R 1965 *The New Utopians: A Study of System Design and Social Change*. Prentice Hall, Englewood Cliffs, NJ

Branch M C 1990 *Planning: Universal Process*. Praeger, New York, NY

Bryson J, Crosby B 1992 *Leadership for the Common Good: How to Tackle Public Problems in a Shared Power World*. Jossey Press, San Francisco, CA

Burchell R W, Sternlieb G 1978 *Planning Theory in the 1980s*. Center for Urban Policy Research, New Brunswick, NJ

Campbell S, Fainstein S (eds.) 1996 *Readings in Planning Theory*. Blackwell, Cambridge, MA

Chadwick G 1971 *A Systems View of Planning: Towards a Theory of the Urban and Regional Planning Process*. Pergamon Press, Oxford, UK

Choay F 1997 *The Rule and the Model: On the Theory of Architecture and Urbanism*. In: Bratton D (ed.) MIT Press, Cambridge, MA

Emery F, Trist E L 1965 The causal texture of organizational environments: An approach to macroorganizational theory. *Human Relations* **27**: 201–38

Faludi A 1973a *Planning Theory*. Pergamon Press, Oxford, UK

Faludi A (ed.) 1973b *A Reader in Planning Theory*. Pergamon Press, Oxford, UK

Faludi A 1996 Rationality, critical rationalism, and planning doctrine. In: Mandelbaum S J, Mazza L, Burchell R (eds.) *Explorations in Planning Theory*. Center for Urban Policy Research, New Brunswick, NJ, pp. 65–82

Forester J 1999 *The Deliberative Practitioner: Encouraging Participatory Planning Processes*. MIT Press, Cambridge, MA

Friedmann J 1987 *Planning in the Public Domain: From Knowledge to Action*. Princeton University Press, Princeton, NJ

Friend J, Hickling A 1997 *Planning Under Pressure: The Strategic Choice Approach*, 2nd edn. Butterworth Heinemann, Oxford, UK

Harris B 1989 Beyond geographic information systems: Computers and the planning professional. *Journal of the American Planning Association* **55**: 85–90

Harris B 1997 The theory of planning and of its profession. *Environment and Planning B*: *Planning and Design* **24**: 483–9

Harvey D 1985 *The Urbanization of Capital: Studies in the History and Theory of Capitalist Urbanization*. Johns Hopkins University Press, Baltimore, MD

Healey P (ed.) 1997 Planning theory, political economy and the interpretive turn: The debate continues! *Planning Theory* **17**: 10–85

Healey P, McDougall G, Thomas M J (eds.) 1982 *Planning Theory: Prospects for the 1980s*. Pergamon Press, Oxford, UK

Innes J E 1990 *Knowledge and Public Policy. The Search for Meaningful Indicators* (2nd (expanded) edn of 1975 title *Social Indicators and Public Policy: Interactive Processes of Design*). Transaction Publishers, New Brunswick, NJ

Klosterman R E 1997 Planning support systems: A new perspective on computer-aided planning. *Journal of Planning Education and Research* **17**: 45–54

Krieger M H 1981 *Advice and Planning*. Temple University Press, Philadelphia, PA

Lindblom C 1965 *The Intelligence of Democracy*. Free Press, New York, NY

Lindblom C 1990 *Inquiry and Change: The Troubled Attempt to Understand and Shape Society*. Yale University Press, New Haven, CT

Lynch K 1981 *Good City Form*. MIT Press, Cambridge, MA

Mandelbaum S J 2000 *Open Moral Communities*. MIT Press, Cambridge, MA

Mandelbaum S J, Mazza L, Burchell R (eds.) 1996 *Explorations in Planning Theory*. Center for Urban Policy Research, New Brunswick, NJ

Pickles J (ed.) 1995 *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford Press, New York
Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA
Rittel H W, Webber M 1973 Dilemmas in a general theory of planning. *Policy Sciences* **4**: 155–69
Schelling T C 1978 *Micromotives and Macrobehavior*. Norton, New York
Sutcliffe A 1981 *Towards the Planned City: Germany, Britain, the United States, and France, 1780–1914*. St. Martin's Press, New York
Taylor N M 1998 *Urban Planning Theory Since 1945*. Sage, Thousand Oaks, CA
Verma N 1998 *Similarities, Connections and Systems: The Search for a New Rationality for Planning and Management*. Lexington Books, Lanham, MD
Vettoretto L 1997 Planning theory and fragmented societies. Explorations in Melvin Webber's civic liberalism. *Planning Theory* **17**: 113–41
Throgmorton J A 1999 Learning through conflict at Oxford. *Journal of Planning Education and Research* **18**: 269–70
Yiftachel O 1999 Planning theory at a crossroad: The Third Oxford Conference. *Journal of Planning Education and Research* **18**: 267–9

S. J. Mandelbaum

Plasticity in Human Behavior across the Lifespan

The concept of plasticity originated in the medical sciences and refers to the capacity of organism to be modifiable. This variability and flexibility enables human beings to adapt to a great range of new and changing environmental conditions. Starting from this origin, the concept of plasticity has been applied to a broad scope of phenomena, the plasticity of nervous systems, changes in observable behavior patterns, and life trajectories.

Neuronal plasticity research, for example, addresses the biochemical and neuronal processes underlying the development and the variability of behavior over the lifespan; that is, it refers to the mechanisms of brain development that involve changes in structure and function (brain plasticity). A key concept in this field was the 'critical period' (Wiesel and Hubel 1963) which refers to a phase in the lifespan when neuronal structures are highly malleable based on the level and type of activity exposed to. For organisms at increasing maturity levels it was expected that neurostructural plasticity decreases with age and is absent in the mature cortex. However, subsequent research revealed substantial plasticity in the mature adult cortex, for example after lesions, strokes, and nerve damage or—even more important—after learning occurs (e.g., Buonomano and Merzenich 1998, Recanzone 2000). Studies showing activity-dependent plasticity on brain structures and transmitter systems in the aging brain provided convincing evidence that brain plasticity is not restricted to juvenile stages of development, but at least in the human species covers the entire lifespan (e.g., Cotman and Neeper 1996).

In contrast to neuronal plasticity research, the focus in lifespan psychological plasticity research lies on identifying the range and the conditions of behavioral modifiability as well as their age-related constraints across the entire life course (behavioral plasticity). In that context, the concept of plasticity generally refers to the ability of human beings to display rather different patterns of behavior as well as different life trajectories (Baltes and Willis 1982, Lerner 1984). So, plasticity can refer to variability of life course patterns (ontogenetic plasticity) or to the variability of a performance in a specific moment in time (concurrent plasticity). Moreover, the study of intraindividual behavioral variability is not restricted to cognitive functioning but extends to noncognitive domains such as motivation and personality. To understand changes in behavioral and ontogenetic plasticity across the lifespan, broader theoretical frameworks of human functioning have to be considered that take the dynamic and complex interplay between evolutionary/biology-based and culture-based influences on the nature of lifespan development into account (Baltes et al. 1998, Finch 1996).

1. Human Behavioral Plasticity as a Product of Evolution and the Foundation of Culture

Much of the behavioral and ontogenetic plasticity is based on the phylogeny of mammals, a biological strata characterized by open behavioral programs (Mayr 1974). With few exceptions, human behavior is not regulated by preprogrammed stimulus–response patterns, but entails the potential to adapt to a variety of ecologies and stimulus constellations. This plasticity opens behavioral patterns both in terms of concurrently available behavioral options and—even more important for ontogenetic change—in terms of acquiring new behavioral patterns. This holds too and is even accentuated when one views the evolution of the human mind from a perspective of evolved domain-specific behavioral and cognitive modules (Cosmides and Tooby 1994, Fodor 1983), which are open to ecological and cultural variations. In the course of human phylogeny such evolved mental modules have escaped domain-specific encapsulation (see Fodor's original conception of encapsulated modules, Fodor 1983) and thus provided the building blocks of human intellectual capacities and behavioral adaptiveness. This variability in behavior and development is as much a product of evolution as the more narrowly constrained behavioral systems of other species (e.g., Wilson 1980).

Behavioral plasticity opens the potential for systematic influences of experiences, most notably in-

fluences of socialization and instruction. Both on the part of the socializing agent and recipient, flexible regulation of behavior in the interaction allows the partners to move beyond preformatted parenting and attachment programs of behavior, and reach out for the social—and ultimately cultural—transmission of experience between generations, and more generally between experts and novices. The ability to instruct and to profit from instruction probably is the critical difference in behavioral regulation which sets *homo sapiens* apart from other primates (Tomasello et al. 1993). Only humans seem to intentionally instruct their children (or nonexpert adults) to modify behavior in accordance with a standard or goal. They correct, model subcomponents, and thus selectively modify error-prone behavior of a partner to fit a certain standard. This specific instructional potential of human interactions is the vehicle of fidelity in information transmission (Tomasello et al. 1993), the prerequisite of cultural heritage. In modern societies this instructional potential has led to the institutionalization of instruction in educational institutions, such as schools and universities. Instructional interactions between humans and the training in educational systems have in turn greatly amplified the range of plasticity in human behavior. The combination of evolutionary potential and cultural transmission can thus generate an immense, albeit not infinite, universe of behavioral plasticity in many potential developmental pathways. Indeed, this great potential for variability, plasticity, and adaptivity in human ontogeny is a cornerstone of lifespan developmental theories (Baltes 1997, Lerner 1984).

2. *Plasticity of Performance Across the Lifespan*

The range of plasticity in performance in diverse domains of functioning (e.g., athletic, intellectual, creative) is determined by the joint influences of either the genotype-related component of performance, that is, genetically-based interindividual differences in performance, or the phenotype-related component of performance, that is, intraindividual differences based on current performance factors (e.g., motivation, alertness), accumulated investment, and instruction. While genotype-related influences delimit the range of plasticity for any given individual, the phenotype-based influences determine whether this potential of plasticity is actually realized (Kliegl and Baltes 1987). Specifically, these four influences are the following. First, there are age-graded differences in the baseline opportunities for maximum performance. It is well known and empirically established that across the human lifespan, performance in various domains of functioning undergoes radical changes, typically reflecting an inverted U-shaped trajectory, with a maximum performance potential at a given age (Lehman 1953, Simonton 1988). Age of maximum performance is relatively early for most athletic disciplines, particularly the track-and-fields sports (Schulz and Curnow 1988), and moves further into adulthood for areas of peak performance which require extensive expertise, such as soloist musical performance (Ericsson et al. 1993), science, and poetry (Simonton 1988). Second, a great many contextual factors can influence an individual's baseline reserve capacity (Kliegl and Baltes 1987). Thus, the component of reserve capacity which influences current performance can be optimized by improving motivation, alertness, clarity of task instruction, and other contextual factors. Albeit opening up many avenues for intervention to promote improved performance, these performance factors are problematic from the perspective of basic research (Salthouse 1985). The performance factors can vary with age and thus confound age differences, so that accurate assessments of latent performance potential at different ages is compromised. Baltes, Kliegl and their colleagues have proposed the methodology of Testing-the-Limits as a research strategy to assess cognitive reserve capacity in terms of latent performance potentials. This research paradigm optimizes performance factors for all individuals and assesses their ability to profit from a training designed to construct a specific cognitive skill, for which the different age groups do not hold differential experiences (Kliegl and Baltes 1987). This latter component of instruction-related plasticity of performance is the third influencing factor, addressed next. Third, performance can be optimized by providing appropriate instruction. The influence of instruction on performance is a long-standing research topic in diverse fields of study and approaches to psychological functioning. Vygotsky conceptualized the zone of proximate development as the range of potential performance an individual can attain with the help and instruction of an expert interactive partner (Vygotsky 1978). This conception about developmental reserve capacity in socially constructive interactions has sprawned much research on social facilitation in parent–child dyads (scaffolding, Wood et al. 1976; apprenticeship, Heckhausen 1987, Kaye 1982, Rogoff and Wertsch 1984), child–child interactions (Rogoff 1990), cognitive aging (development-enhancing interventions, Kliegl and Baltes 1987), and educational psychology (see zone of potential development, Brown and French 1979). Finally fourth, plasticity in performance is a function of the individual's long-term investment in a given area of skill or expertise. Research on life-course trajectories of peak performance across a wide variety of domains has demonstrated repeatedly that world-class levels of performance require extensive and high-quality expert-guided training and practice over long periods of the lifespan, particularly in the early phases of acquisition (Ericsson and Charness 1994, Simonton 1988). This powerful effect of focuses and prolonged

practice and training is itself the best witness of the impressive plasticity of human performance across the lifespan.

3. Cognitive Functioning in Old Age: A Case in Point

An extensively studied area of human behavioral plasticity is cognitive functioning in old age. Numerous studies have demonstrated declining performance in fluid intelligence performance with increasing age among older adults (Salthouse 1985). This pattern of age related decline has been challenged to be reflective of true latent performance potential, since many factors such as motivation or lack of recent practice could produce performance differences such as those found between older and younger adults in intellectual functioning. This critique prompted researchers to optimize performance conditions for older adults for instance by using ecologically valid tasks. Moreover, specifically designed training programs were administered to promote improved performance in older adults (Verhaeghen et al. 1992, Willis 1989). Examples of early training studies are among others the Penn State Adult Development and Enrichment Project (ADEPT) and the Berlin PRO-ALT project (for review see, Baltes and Lindenberger 1988). These studies demonstrated substantial benefits of older adults' fluid intelligence performance from cognitive training. For tasks that were similar to those addressed by the cognitive training, older adults improved their performance. However, the training effect showed no transfer to other fluid intelligence tasks. Additional studies comparing the benefit of guided instruction with self-regulated practice showed that self-instruction was as effective in improving cognitive performance in older adults as was guided instruction. This suggests that a large share of the old age plasticity in cognitive functioning is due to the activation of already available cognitive skills, which have been fallen into disuse over the adult lifespan.

The findings of preserved substantial learning potentials up to old age and the possibly arising optimistic view of unchanged cognitive plasticity over the entire lifespan had been put in perspective by Testing-the-Limits studies aiming to delimit maximum performance potentials (Kliegl and Baltes 1987). In this research, cognitive skills are taught to subjects using specific training programs to construct an expertise in a certain cognitive domain, for example in the domain of memory for nouns. In this way, an expertise is cognitively engineered, so that the researchers know the component abilities, and can thus specifically compare subjects at different ages in terms of specific subskills. Subjects are trained extensively until they have reached asymptotic performance levels. Once maximum performance levels are achieved, time constraints are introduced. The findings indicate that although older adults can reach high levels of performance (e.g., in recalling nouns), their maximum (asymptotic) performance level after extensive training is much lower compared to younger adults. Moreover, their functioning breaks down when they encounter severe time constraints. Thus, at limits of functioning, age differences are magnified and reveal the potentially underlying neurobiological constraints associated with the biology of the aging brain (Baltes and Kliegl 1992). Finally, Testing-the-Limits methodologies have been successfully employed to identify elderly at risk of dementia (Baltes et al 1992) and distinguish pathological from normal cognitive aging (Reischies and Lindenberger 1996).

4. Plasticity in Personality Across the Lifespan

Processes of self-regulation, personality, and subjective well-being show much less systematic age variation than cognitive functioning. Indeed, the stability of life satisfaction, self-esteem, perceived control, and overall happiness in itself is proof of the adaptive plasticity of these processes. The age-normative shift from a predominance of perceived gains to ever increasing risks for decline (Heckhausen et al. 1989) would lead one to expect concomitant declines in indicators and components of subjective well-being. Instead, emotional balance remains resilient across the life span (Staudinger et al. 1993).

This resilience of the emotional balance to lifespan changes is one of the key achievements of the human motivational system, which has evolved to manage the challenges associated with the relative dearth of biologically-based predetermination of behavior. The relatively weak biological regulation of behavior furnishes a regulatory need to achieve and maintain selectivity in behavioral investment, and to compensate for the ever-present failure and loss of a behavioral system that is underdetermined and thus highly error-prone. Both selectivity and loss compensation, key concepts in lifespan developmental research (Baltes et al. 1998, Schulz and Heckhausen 1999, Carstensen et al. 1999), require adaptive motivational processes, which focus volitional investment into a chosen goal and buffer the potentially negative effects of aging-related losses. Typical examples of such metamotivational processes (i.e., secondary control, Heckhausen and Schulz 1995) which facilitate goal striving, are upward social comparisons with superior or advanced others, who can serve as models for self-improvement (Heckhausen and Krueger 1993), and enhanced valuation of the chosen compared to competing goals. Examples of metamotivational processes, which compensate for the negative effects of losses or failures, are self-enhancing comparisons with inferior others, self-protective attributions which take credit for gains but refuse to take blame for losses, and the adjustment of goals to fit one's real potential (for the latter example

see also Brandtstädter and Renner's conception of accommodation, Brandtstädter and Renner 1992). In general, the age-graded structure of opportunities and constraints sets the stage for the individual to use the age window of maximal opportunities to strive for a given goal. The regulation of motivational investment requires substantial plasticity in that the individual needs to enter cycles of engagement and disengagement with various developmental goals (Heckhausen 1999). Motivational engagement and disengagement with specific life goals should be contingent and flexibly adapted to changes in opportunities and constraints across the lifespan.

See also: Behaviorism; Human Behavioral Ecology; Life Course in History; Life Course: Sociological Aspects; Lifelong Learning and its Support with New Media: Cultural Concerns; Lifespan Development: Evolutionary Perspectives; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development

Bibliography

Baltes M M, Kühl K P, Sowarka D 1992 Testing for limits of cognitive reserve capacity: a promising strategy for early diagnostic of dementia? *Journal of Gerontology: Psychological Sciences* **47**: 165–7

Baltes P B 1997 On the incomplete architecture of human antogeny: selection, organization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80

Baltes P B, Kliegl R 1992 Further testing of limits of cognitive plasticity: Negative age differences in a mnemonic skill are robust. *Developmental Psychology* **28**: 121–5

Baltes P B, Lindenberger U 1988 On the range of cognitive plasticity in old age as a function of experience: 15 years of intervention research. *Behavior Therapy* **19**: 283–300

Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology, Vol. 1: Theoretical Models of Human Development*, 5th edn. J. Wiley, New York, pp. 1029–1143

Baltes P B, Willis S L 1982 Plasticity and enhancement of intellectual functioning in old age: Penn State's Adult Development and Enrichment Project (ADEPT). In: Craik F I M, Trehub S (eds.) *Aging and Cognitive Processes*. Plenum Press, New York, pp. 353–89

Brandtstädter J, Renner G 1992 Coping with discrepancies between aspirations and achievements in adult development: a dual-process model. In: Montada L, Filipp S-H, Lerner M R (eds.) *Life Crises and Experiences of Loss in Adulthood*. Erlbaum, Hillsdale, NJ, pp. 301–19

Brown A L, French L A 1979 The zone of potential development: implication for intelligence testing in the year 2000. *Intelligence* **3**: 255–77

Buonomano D V, Merzenich M M 1998 Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience* **21**: 149–86

Carstensen L L, Isaacowitz D M, Charles S T 1999 Taking time seriously: a theory of socioemotional selectivity. *American Psychologist* **54**: 165–81

Cosmides L, Tooby J 1994 Origins of domain-specificity: the evolution of functional organization. In: Hirschfeld L A, Gelman S A (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, Cambridge, UK, pp. 85–116

Cotman C W, Neeper S 1996 Activity-dependent plasticity and the aging brain. In: Schneider L, Rowe J W (eds.) *Handbook of the Biology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 283–99

Ericsson K A, Krampe R T, Tesch-Römer C 1993 The role of deliberate practice in the acquisition of expert performance. *Psychology Review* **100**(3): 363–406

Ericsson K A, Charness N 1994 Expert performance: its structure and aquisition. *American Psychologist* **49**(8): 725–47

Finch C E 1996 Biological bases for plasticity during aging of individual life histories. In: Magnusson D (ed.) *The Life-Span Development of Individuals: Behavioral, Neurobiological and Psychosocial Perspective*. Cambridge University Press, Cambridge, UK, pp. 488–511

Fodor J A 1983 *The Modularity of Mind*. MIT Press, Cambridge, MA

Heckhausen J 1987 Balancing for weaknesses and challenging developmental potential: a longitudinal study of mother-infant dyads in apprenticeship interactions. *Developmental Psychology* **23**: 762–70

Heckhausen J 1999 *Developmental Regulation in Adulthood: Age-Normative and Sociostructural Constraints as Adaptive Challenges*. Cambridge University Press, New York

Heckhausen J, Dixon R A, Baltes P B 1989 Gains and losses in development throughout adulthood as perceived by different adult age groups. *Developmental Psychology* **25**: 109–21

Heckhausen J, Krueger J 1993 Developmental expectations for the self and most other people: age grading in three functions of social comparison. *Developmental Psychology* **29**: 539–48

Heckhausen J, Schulz R 1995 A life-span theory of control. *Psychological Review* **102**: 284–304

Kaye K 1982 *The Mental and Social Life of Babies. How Parents Create Persons*. University of Chicago Press, Chicago

Kliegl R, Baltes P B 1987 Theory-guided analysis of mechanisms of development and aging through testing-the-limits and research on expertise. In: Schooler C, Schaie K W (eds.) *Cognitive Functioning and Social Structure over the Life Course*. Ablex, Norwood, NJ, pp. 95–119

Lehman H C 1953 *Age and Achievement*. Princeton University Press, Princeton, NJ

Lerner R M 1984 *On the Nature of Human Plasticity*. Cambridge University Press, New York

Mayr E 1974 Behavior programs and evolutionary strategies. *American Scientist* **62**: 650–9

Recanzone G H 2000 Cerebral cortical plasticity: Perception and skill acquisition. In: Gazzaniga S (ed.) *The New Cognitive Neurosciences*, 2nd edn. MIT Press, Cambridge, MA, pp. 237–47

Reischies F M, Lindenberger U 1996 Grenzen und Potentiale kognitiver Leistungsf-Àhigkeit im Alter [Limits and potentials of cognitive ability in old age]. In: Mayer K U, Baltes P B (eds.) *Die Berliner Altersstudie*. Akademie Verlag, Berlin, pp. 351–78

Rogoff B 1990 *Apprenticeship in Thinking: Cognitive Development in Social Context*. Oxford University Press, New York

Rogoff B, Wertsch J V (eds.) 1984 Children's learning in the "zone of proximal development". Jossey-Bass, San Francisco
Salthause T A 1985 A Theory of Cognitive Aging. North-Holland, Amsterdam
Schulz R, Cunow C 1988 Peak performance and age among super athletes: track and field, swimming, baseball, tennis, and golf. *Journal of Gerontology: Psychological Sciences* **43**: 113–20
Schulz R, Heckhausen J 1999 Aging, culture and control: setting a new research agenda. *Journal of Gerontology series B: Psychological Sciences* **54**: 139–45
Simonton D K 1988 Age and outstanding achievement: What do we know after a century of research? *Psychological Bulletin* **104**: 251–67
Staudinger U M, Marsiske M, Baltes P B 1993 Resilience and levels of reserve capacity in later adulthood: Perspectives from life-span theory. *Development and Psychopathology* **5**: 541–66
Tomasello M, Kruger A C, Ratner H H 1993 Cultural learning. *Behavioral and Brain Sciences* **16**: 495–511
Verhaeghen P, Marcoen A, Goossens L 1992 Improving memory performance in the aged through mnemonic training: A meta-analytic study. *Psychology and Aging* **7**: 242–51
Vygotsky L S 1978 *Mind in Society*. Harvard University Press, Cambridge, MA
Wiesel T N, Hubel D H 1963 Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of Neurophysiology* **26**: 1003–17
Willis S L 1989 Improvement with cognitive training: Which dogs learn what tricks? In: Poon L W, Rubin D C, Wilson B A (eds.) *Everyday Cognition in Adulthood and Late Life*. Cambridge University Press, Cambridge, UK, pp. 545–69
Wilson E O 1980 *Sociobiology*. Belknap Press of Harvard University Press, Cambridge, MA
Wood D J, Bruner J S, Ross G 1976 The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry* **17**: 89–100

J. Heckhausen & T. Singer

Play and Development in Children

1. Introduction

The study of play has held particular allure to social and behavioral scientists, at least since Groos (1898). Since that time it has been assumed that play serves an important role in children's development into adults. An important step in documenting the importance of play should be defining play in terms of the ways in which it is different from forms of behavior not considered playful.

2. What is Play?

No one definition of play is necessary or sufficient, yet child psychologists (Pellegrini and Smith 1998) and ethologists (Martin and Caro 1985) agree upon some common definitional criteria: play does not seem to serve any apparent immediate purpose. This assumes that children are less concerned with the outcome of their behavior than with the behavioral processes *per se*.

Play has also been defined according to phenomena that precede (e.g., an atmosphere that is familiar, safe, and friendly, a minimally intrusive adult, and children who are free from stress) and succeed play behaviors (e.g., behavior is categorized as play fighting, not aggression, if children stay together after the conclusion of the bout; if they separate, it is defined as aggression). Caution is needed as in this case as antecedents and consequences of behaviors can be considered as elicitors and outcomes, respectively, of those behaviors, rather than as components of the behavior *per se*.

3. Forms of Play

Different forms of play occur from infancy through adolescence and they have distinct antecedents and possible functions.

3.1 Exploration

Exploration is an information-gathering venture (e.g., mouthing and simple manipulation of objects) (Belsky and Most 1983). Through exploration children come to know their environments.

Exploration dominates behavior for the first 9 months of life. By 12 months, play and exploration co-occur. By 18 months play accounts for more of the child's interactions with the environment than does exploration (Belsky and Most 1983). Exploration also precedes play to the extent that children of all ages must explore an object, or know its properties, before they can play with it. The knowledge resulting from exploration is basis for play.

3.2 Fantasy Play

Fantasy, or pretense, is the paradigm case of play during childhood. Fantasy involves an 'as if…' orientation to the world and involves actions, use of objects, and verbalizations and nonliteral meanings; often, it involves playing a distinct pretend role such as mummy. Fantasy begins during the second year of life, peaks during the late preschool years, and then declines (Fein 1981) and there are reliable gender differences: girls engage in fantasy play both more frequently and at more sophisticated levels than do boys.

Fantasy play is influenced by familial variables as well as play props and play partners. Children who are securely, compared with insecurely, attached to their

mothers engage in more sophisticated pretend play (Roggman and Langlois 1987). Regarding props, play themes generally follow the themes inherent in the props available (Pellegrini and Perlmutter 1989). Further, children's pretense is more sustained and complex when they are playing with friends, compared with acquaintances, possibly because the mutuality and emotional commitment of friends may motivate children to sustain cooperation (Howes 1992).

3.3 Locomotor Play

Locomotor play is play that is physically vigorous, e.g., chasing and rough-and-tumble play (RandT). The less 'thematic' and nonsocial forms of vigorous play, such as running and jumping, account for close to 20 percent of the activities of 3- and 4-year-olds and then declines during primary school years (Pellegrini and Smith 1998).

Vigorous play is shaped by both hormonal and social events, resulting in reliable gender differences (Pellegrini and Smith 1998). During the course of normal fetal development, androgens predispose males to be more physically active than females. Socialization practices, such as father engaging in vigorous play with sons more frequently than with daughters, reinforce the difference.

4. Functions of Play

Paradoxically, play is defined in terms of its 'purposelessness' yet seen as serving an important function in children's development. The paradox can be reconciled by supposing that the apparent lack of immediate function actually conceals a function delayed in development.

4.1 What is Meant by Function?

Function has at least two meanings in the social and behavioral sciences. First, function can be defined in terms of an evolutionary history. A behavior is functional if it has added to the reproductive success of individuals. Functions can also be defined in terms of beneficial consequences during the lifespan of any particular individual. Benefits may be accrued immediately or during the period of childhood, or deferred until adulthood.

4.2 Immediate or Deferred Functions

Most theories of play have proposed that the benefits of play are not immediate to the period of childhood but deferred until later in development. In Bateson's (1976) terms this view presents play as a 'scaffolding' metaphor: play functions in the assembly of a skill and then is disassembled when the skill is complete. This remains an influential view, even though empirical studies have not provided support for this claim (e.g., Martin and Caro 1985, Pellegrini and Smith 1998).

Alternatively, play may not be an imperfect version of adult behavior but may be adapted and beneficial to the specific niche of childhood. Bateson's (1976) metaphor for the immediate functions of play is metamorphic (e.g., the sense of mastery associated with children's play probably relates to children experimenting with new activities).

5. Making Functional Inferences

Function can be inferred using cost–benefit analyses, consideration of design and contextual features, and experimental enhancement or deprivation of play components.

5.1 Cost–Benefit Analyses

With cost–benefit analyses we assume a correspondence between high costs and high benefits; low costs imply low or high benefits. Play can be costly in terms of time, caloric expenditure, and survivorship (e.g., danger). Basic cost data on children's play is, however, scanty. Yet there are some hints from the animal play research which suggest that play accounts for a limited portion of an animal's time and energy budgets—around 10 percent (Martin and Caro 1985, Pellegrini and Smith 1998). Consequently, we conclude that benefits are probably low and deferred as immediate benefits are more likely to be reaped given risks associated with survivorship.

5.2 Design Features

A design-features argument involves explication of similarities between aspects of play behavior and similar features in mature behavior. Design feature arguments have been made most frequently for RandT as its fighting-like features (e.g., wrestling) should relate to fighting skills and dominance. Problems associated with this design-features approach to the study of function relate to the fact that a mature behavior may develop from many different antecedents (Martin and Caro 1985).

5.3 Experimental Deprivation Studies

A few experimental play-deprivation studies have been conducted where juveniles are deprived of an aspect of play. The assumption is that deprivation of an aspect of play, if it is developmentally important, should

result in a 'rebound effect' when children are given the opportunity to play again; children compensate for the deprivation by engaging in high levels and longer durations of play. Most deprivation studies can be criticized on the grounds that more than one thing is involved when we deprive children of play (e.g., depriving children of social play simultaneously deprives them of social interaction).

The effects of depriving children of play opportunities have largely been limited to studies of physical play. In different field experiments, Smith and Hagan (1980) and Pellegrini et al. (1998) experimentally deprived children of opportunities for physical play. Across all experiments the results showed that increased deprivation led to increased levels of play when opportunities for play were resumed. We argued that children rebounded after deprivation so as to overcompensate for lost opportunities for physical exercise, which serves important training functions.

5.4 *Experimental Enrichment Studies*

With enrichment procedures, children are given opportunities to engage in play that vary according to theoretically relevant dimensions, such as conceptual conflict (Pellegrini 1984), and their effects are measured. Most of the criterion measures in these studies are taken in close temporal proximity to the training, thus assuming that play has immediate rather than deferred benefits.

There have been numerous studies examining the effects of and pretend play enrichment on various dimensions of children's social cognitive status, such as role taking. Most of these studies showed positive benefit effects of play; but they must be interpreted very cautiously, primarily because many play enrichment experiments have serious internal validity problems (Smith 1988). For example, many experiments confound play treatment with adults tutoring children in play and most effects are due to experimenter bias.

6. *Conclusion*

We have outlined some of the basic definitional and developmental tenets of play. The role of play in children's development is still a controversial and unresolved topic. Cost–benefit analyses suggest that there should be benefits; but they may be relatively small and possibly limited to the period of childhood, rather than deferred until after childhood. Suggested benefits include physical fitness (for physical play), fighting and dominance skills (for RandT), and affiliative and theory of mind skills (social and pretend play).

With this said, however, we note that much more research is needed before we can make sound conclusions regarding the developmental functions of play. Most basically, more descriptive studies of children's play, documenting costs in terms of time, caloric expenditure, and survivorship in contexts other than school are needed. This information could be used in the service of cost–benefit analytic studies of play. Correspondingly, these features could also be used to test functional hypotheses more directly using field experimental methods.

See also: Children's Play: Educational Function; Infancy and Childhood: Emotional Development; Piaget's Theory of Child Development; Play, Anthropology of

Bibliography

Bateson P P G 1976 Rules and reciprocity in behavioral development. In: Bateson P P G, Hinde R A (eds.) *Growing Points in Ethology*. Cambridge University Press, Cambridge, UK, pp. 401–21

Belsky J, Most R K 1981 From exploitation to play. *Developmental Psychology* **17**: 630–9

Fein G 1981 Pretend play: an integrative review. *Child Development* **52**: 1095–118

Groos K 1898 *The Play of Animals*. D. Appleton, New York

Howes C 1992 *The Collaborative Construction of Pretend*. State University of New York Press, Albany, NY

Martin P, Caro T M 1985 On the functions of play and its role in behavioral development. *Advances in the Study of Behavior* **15**: 59–103

Pellegrini A D 1984 Identifying causal elements in the thematic-fantasy play paradigm. *American Educational Research Journal* **21**: 691–703

Pellegrini A D, Horvat M, Huberty P 1998 The relative costs of children's physical activity play. *Animal Behavior* **55**: 1053–61

Pellegrini A D, Perlmutter J C 1989 Classroom contextual effects on children's play. *Developmental Psychology* **25**: 289–96

Pellegrini A D, Smith P K 1998 Physical activity play: the nature and function of a neglected aspect of play. *Child Development* **69**: 577–98

Roggman L A, Langlois J H, Hubbstait L 1987 Mothers, infants, and toys: social play correlates of attachment. *Infant Behavior and Development* **10**: 233–37

Smith P K 1982 Does play matter? Functional and evolutionary aspects of animal and human play. *The Behavioral and Brain Sciences* **5**: 139–55

Smith P K 1988 Children's play and its role in early development: a re-evaluation of the 'play ethos.' In: Pellegrini A D (ed.) *Psychological Bases for Early Education*. Wiley, Chichester, UK, pp. 207–26

Smith P K, Hagan T 1980 Effects of deprivation on exercise play in nursery school children. *Animal Behavior* **28**: 922–8

A. D. Pellegrini and P. K. Smith

Play, Anthropology of

Malleable, flexible, fluid, uncertain, changeable—all are adjectives that are used to describe play, whether as a medium of cognition, practice, or communication. Play refers to a complex of universal phenomena

among human beings that also are found among numerous animal species (Fagen 1981). All peoples, though they may not distinguish in semantic terms between play and other behavior, likely engage in playing as they shift into a conditional or subjunctive sense of being, one that opens toward a multiplicity of possibility. The mathematician and the theoretical physicist play during the process of shaping their ideas no less than children imagine, practice, and repeat their inventions in their play. Useful overviews of play phenomena are given in Spariosu (1989), Schwartzman (1978), and Sutton-Smith (1997).

Despite its pervasive presence, play is perhaps the least concrete of phenomena and therefore one of the most difficult to define. Any useful definition of play highlights its multiplicity of potential experiences. The anthropologist, Bateson (1972) perceives play as paradoxical, in that it is and is not what it seems to be. The classicist Spariosu (1989) terms play, amphibolous, or going in two directions at once. In a succinct manner, the anthropologist, Droogers (1996) defines play as, 'the capacity to deal simultaneously and subjunctively with two or more ways of classifying reality.'

1. *Play and Reality*

There is an intimate relationship between play and reality because play is first and foremost experiential, whether personal or social, and reality does not exist without its experiencing. In the creation and practice of realities of play, the imagination is crucial. Without the cognitive, the capacity to imagine realities, there is no play, just as, without the practice of experience there is no reality.

Play raises to the fore issues of alternative realities, the role of these and other multiplicities in social and personal life, and their consequences for human cognition and practice. The cognitive perceptions of play enable persons to enter into alternative realities, and perhaps into the simultaneity of multiplicity. Cognition is related intrinsically to how realities of play are imagined and framed differently from other realities. The practice of play entails the sensuous embodiment of these alternative realities; and, so, the subjective experiencing of play from within these imagined contexts that offer a sense of release, freedom, exploration, and perhaps exhiliration. The communicative component suggests that these specialized contexts of play are able to comment on the social orders from which they arise, and to send these messages back as the participants in play return to the reality of nonplay. Therefore contexts of play are metacommunicative in their relationships to social order: play communicates to its participants not only *how* to play but how such playing relates to social order. That play is framed as imaginative and malleable, and that through these qualities it generates and conveys metamessages informing players as to how to relate to aspects of social order, makes play perhaps the most fruitful of contexts of socialization, both for children (Briggs 1998) and for adults.

2. *Framing Play*

Given that play constitutes domains that cognize, practice, and communicate the shaping of imagination, are these domains to be conceptualized as incisively distinct from domains of serious reality, of not-play? This is the implication of Bateson's highly influential theory of the framing of play which separates play lineally from the serious, or not-play. According to Bateson, the cognitive metamessage, 'This is play,' brings into existence the play frame which cleaves between the seriousness of everyday reality and the imaginings which are practiced within the frame. One may adduce from this not only that play comes into existence as a corollary of its framing but also that play, and the practice of its imaginative multiplicities of possibility, is located within the restrictions of its framing. Viewed from outside the frame, play is make-believe activity that in fact does not exist. Experienced from within the frame, play may well be utterly engrossing in the embracing reality of its presence. These views accord well with Johann Huizinga's (1970) classic study of how civilization emerged through the plastic premises of play; although in the process, play itself became a temporary sphere of make-believe activity, secluded, limited, and bounded in time and space. Once play is framed through lineal cognition, it becomes a world unto itself, one of temporary limited perfection with its own internal teleologies (Gadamer 1988). Within itself the practice of play may well proceed in deadly earnest.

2.1 *Play and the Playful*

Sutton-Smith (1997) offers a useful modification to the binarism of a reality either of play or of not-play, of make-believe or of serious reality. Sutton-Smith distinguishes between play and the playful. To build on his argument that the playful plays with the very framing of play itself, one can add that the playful is a mood, an attitude, that may permeate both serious reality and play, or that suddenly may surge into presence within the mundane. In either instance, the playful is full of the impulse to perceive and feel in ways other than those offered by the immediacy of a given reality. The abrupt presence of playfulness disrupts the most routine of expectations. Playfulness injects speedy uncertainty into the most expected of social practices.

The playful attitude within mundane reality may be integral to the existence and organization of the latter; while play itself may be more 'distant,' as it were, from mundane reality. Playfulness and play complement

one another. The eruption of playful feeling and perception are a significant key to establishing the play frame in relationship to mundane reality. But the play frame itself directs the kinds of ludic or make-believe activities that take shape within it, relatively separated from mundane reality.

2.2 Play and Foundation-for-form

In order for a play frame to come into existence there seems to be a need for a form of reference, a foundation-for-form or organic model that can be altered in systematic ways and turned into someone or something else. Nonetheless, whatever is changed retains crucial similarities to the form of its foundation and so remains intimately related to it. For example, The medieval European Feast of Fools which mocked the traditional Christian Mass needed the latter so that this could be inverted and altered. The play-mass, performed in church by priests, would have had no significance for participants were it not derived from and contrasted with its everyday analogue, the sacred Mass.

To take a more mundane example, when one little American girl, aged seven, turns to her sister, aged five, and says: 'Let's play sisters,' she is establishing a play frame within which these girls can play the foundational form of their sisterhood, imaginatively and perhaps with novelty. Play changes the known signs of form into something else by altering the reified boundaries that define and characterize the phenomenon. These girls will still be sisters, but they will be able to explore, practice, and experience their sisterhood with a difference, with license, in ways perhaps closed or unimagined within their usual relationship. They can play with the strangeness of being sisters. In turn, they may carry the messages of their experience back into their ongoing relationship as sisters, changing this in minute but additive ways. Hypothetically, they playfully may enter into play-sisterhood whenever they desire.

3. Attributes of Play

Play is a way of organizing activity, rather than any particular set or content of activities (Miller 1973). Any activity may be altered playfully, and played at. But in order to do this the everyday, instrumental perception of how means lead to and produce ends must be loosened, stretched, or perhaps uncoupled entirely. Goals exist in play, but the means to attain them are given degrees of autonomy. This is why play behavior can be endlessly repetitive or follow a crooked line around artificial obstacles. The patterning for play may be suggested by how the foundations-for-form that are used in playing function in the everyday. Say that the two little sisters and their friends place chairs in a row, the back of each facing the seat of the next, and that these children imagine these chairs to be a train. The children practice playing train, until they decide that the train is a plane, and the row sprouts wings, carrying its passengers, but then is turned into a bird flying off with children on its back, becoming a row of chairs in school and kindergarten that is also a plane. In each instance, one foundation for form (train, plane, bird) is imagined to take the place of another. Each such organic model is voluntarily turned into a foundation for imaginative practice that plays in turn with the multiplicity of these organic models, one after the other or as hybrids of one another (a row of chairs that with near simultaneity is train, plane, bird).

The perception and practice of play brings into being something that had not existed before in quite the same way. This is done by changing the shape and positioning of boundaries that categorize phenomena, thereby altering their meaning. Each shift by sisters into playing sisters seems like a repetition, yet is instead a different beginning since it imagines anew and plays with the very possibilities of sisterhood. Even a slight change in the boundaries that give shape and significance to phenomena involves some degree of destruction and recreation. In its cognition and feeling, play is associated closely with creativity, creation, and destruction.

Play is frequently implicated in the making of change. Change is the making of difference, even if this difference seems to mimic or eventually to reproduce an earlier state. Play introduces malleability into social and cultural life, taking apart the given and perhaps offering alternative visions and practices. Play often is involved in learning something new, through open-ended repetition, variation, exaggeration, and experimentation. Play is not unusual in the liminal phases of initiation rituals (Turner 1969), healing rituals (Kapferer 1983), and mortuary rituals of tribal peoples, when initiates, the ill, or the dead undergo profound transformations of being. Playfulness, in a more postmodern idiom is central to the practice of certain neo-shamanic rituals whose goal is the transformation of consciousness (Lindquist 1997). These and many other changing conditions involve shifts of cognition, practice, and communication. More significantly, change (creative or reproductive) usually involves fragmentation, destruction, or transformation of some sort, whether temporary or permanent, and these processes are often violent. The violence of change and change through play are often closely related. Some form of destruction and violence (to cognition, feeling, body and being) is involved in the shattering suddenness of the playful and in play.

4. The Truth-values of Play

The relationship between play and violence relates to how play has been conceived, and to the truth-values accorded to play. In his survey of philosophical ideas

of play, Spariosu (1989) finds two trends, one that stresses the practice of play as prerational (with its roots in ancient Greek thought and poetry), and the other that emphasizes play as rational activity. In prerational thinking, play is full of violent agon and power, strongly and pleasurably emotional, arbitrary in its appearance, experienced perhaps as ecstatic, and opening toward unlimited freedom. In rational thinking, the naked, enjoyable forces of chaos and disorder, of irrational power and violence, are suppressed in favor of play as rationally framed activity that is set apart from serious reality. Here play is irreal in its irrationality, and therefore controlled, nonviolent, and limited in its freedom. To expand on Spariosu, one may add that the eruption of pre-rational play (akin to playfulness) into everyday life is full of its own unsuppressed truth-value. The truth-value of prerational play is integral to living within realities that overlap and that are continuous with one another—realities that relate to one another as multiplicities of possibility. By contrast, the make-believe status of rational play, framed and contained in special times and places (as Huizinga, Bateson, and others argue) denies any direct or immediate truth-value to play, enabling its practice to be treated as fictive, as 'just play,' 'mere play,' 'done only in play,' and so forth.

4.1 Play and Cosmology

The organization of cosmology demonstrates from another perspective the importance of the truth-values attributed to play. For example, in Indian cosmogenesis, play (or more accurately, playfulness) is implicated in the very creation of cosmos. Thus, varieties of Hindu cosmology and religion embed premises of play—malleability, multiplicity, ongoing change—at a very high level of cosmic organization. In consequence, these self-same premises of play resonate throughout the levels of cosmos. The entirety of cosmos is influenced by premises of play. In this kind of cosmology, play is a top-down idea, responsible in no small degree for the multiplicities of reality that constitute Hindu cosmos and religions (Handelman 1992, Handelman and Shulman 1997).

Here play has high degrees of truth-value and is rarely referred to as make-believe. By contrast, mainstream monotheistic cosmologies, if they pay any attention at all to ideas of play, embed these at a low level of cosmic organization, distant from cosmogenesis and the ordering of religion and society. In monotheistic cosmologies, play is more of a bottom-up idea, lineally framed, rationally contained, and separated from serious reality. Bottom-up play enters cosmology and religion from their conceptual and emotional undergrowth, and therefore is perceived as threatening. Here the messages of play are often positioned in opposition to or as a negation of the order of things; and play is perceived as fantasy, illusion, and ephemeral; yet also as subversive and as resisting the order of things. But here the practice of play has low truth-value. Thus, the truth-value of top-down play is openly full of power and violence, while these qualities are veiled, hidden, and denatured in the low truth-values accorded to bottom-up play.

5. Game and Play

The idea and practice of game is closely related to that of play and may be understood as a variation of the latter, in the following senses. Game, like play, is a way of organizing activity. Game, too, loosens the relationship between means and ends. But the existence of game depends on relatively enduring rules, rather than on foundations-for-form that are easily changeable. Therefore games constitute distinct, autonomous realities, virtualities, given their internal, rule-based organization which to a high degree determines how they are practiced. Playfulness has less of a role in the practice of games. Games are less malleable and harder to alter than play.

Yet it is precisely because of their ruledness that games can be organized to make directed change through their practice. Lévi-Strauss (1966) comments that game is a systematic way of creating inequality. Game begins with equality between its sides and produces differences which are perceived as significant. The logics of game, in fact, are often close to those of ritual which is intended to generate change through its operations. Traditional Mesoamerican ball-games had game-like aspects, yet they were ritual-like in their relationship between means and ends, perhaps ends of cosmic renewal that their practice was intended to accomplish. More generally, the presence of play and game within rites seem to contribute to the attainment of ritual goals. (This is powerfully exemplified in the roles of clowns in Native-American Pueblo rituals.)

Perhaps in more systemic ways than play, game may awaken interpretive, reflexive feelings about aspects of social and moral order. The highly agonistic Native-American Kwakiutl potlatch, a ritual full of conflict that both demonstrated and reordered social status, produced the 'play potlatch' (Codere 1956) in which participants laughed at their own strivings for power and hierarchy. The rigid gender divisions in traditional Iranian society produced games practiced by women in which these distinctions were called into question, mocked, and played with (Safa-Isfahani 1980).

The twentieth century has seen tremendous growth in the use of game as a metaphor for living and social order (Minnema 1998)—Wittgenstein's language games, Von Neumann's game theory, war games, educational games, the game of intelligence testing, games people play, the play of quantum mechanics, chaos theory, virtual realities, and so forth. In postmodernist terms, these visions are of life organized as interpenetrating games, none of which give any total

or complete vision of reality, and no certainty about the stability of forms of living. The game metaphor reflects more fluid and mobile social orders that are perceived, correctly or incorrectly, as the creations of individual agency. Television as spectacle has contributed greatly to this growth by telecasting existing competitive game forms and aiding in the invention of new ones. Sports are at the front of these developments. These multitudes of contest (in which violence is close to the social surface if not on it) continuously rank and re-rank participants over and again in hierarchies that are more kaleidoscopic, fragile, and ephemeral than they are enduring. Play as the multiplicities of possibility, and the playful as the eruption of these possibilities in the everyday, are competing with more unitary monotheistic ethoses, at least in the visions of the middle-class Western world.

Current thinking on play is reconceptualizing the problem of framing—the problem of relationships among realities, social and cognitive—to open alternatives to the ongoing rupture between play as rational and as irrational. Research on Internet chat groups plays into questions of framing (Danet et al. 1998). Issues of power in and through the practice of play are also related to questions of framing. Thus, ideas of 'dark play' (Schechner 1988), when players do not know that they are playing, and 'evil play,' when the manipulator of the frame positions herself outside it and plays others, yet pretends to be within the frame and controlled by it, relate to issues of play and power. They echo Bateson's still enduring thought that a more complex question to ask of a play phenomenon is whether it is indeed this, or whether it is pretending to be so. Imagining the pretense of play is even more powerful than play imagined as such.

See also: Game Theory; Leisure, Psychology of; Play and Development in Children; Sport, Anthropology of; Sport, Sociology of

Bibliography

Bateson G 1972 *Steps to an Ecology of Mind*. Chandler, San Francisco

Briggs J L 1998 *Inuit Morality Play: The Emotional Education of a Three-Year-Old*. Yale University Press, New Haven, CT

Codere H 1956 The amiable side of Kwakiutl life: The potlatch and the play potlatch. *American Anthropologist* **58**: 334–51

Danet B, Ruedenberg L, Rosenbaum-Tamari Y 1998 Hmmm... where's that smoke coming from? Writing, play and performance on internet relay chat. In: Sudweeks F, McLaughlin M, Rafaeli S (eds.) *Network and Netplay: Virtual Groups on the Internet*. AAAI Press, Menlo Park, CA and MIT Press, Cambridge, MA

Droogers A 1996 Methodological ludism: beyond religionism and reductionism. In: van Harskamp A (ed.) *Conflicts in Social Science*. Routledge, London

Fagen R 1981 *Animal Play Behavior*. Oxford University Press, New York

Gadamer H-G 1988 *Truth and Method*. Crossroad, New York

Handelman D 1992 Passages to play: Paradox and process. *Play & Culture* **5**: 1–19

Handelman D, Shulman D 1997 *God Inside Out: Siva's Game of Dice*. Oxford University Press, New York

Huizinga J 1970 *Homo Ludens: A Study of the Play Element in Culture*. Maurice Temple Smith, London

Kapferer B 1983 *A Celebration of Demons: Exorcism and the Aesthetics of Healing in Sri Lanka*. Indiana University Press, Bloomington, IN

Lévi-Strauss C 1966 *The Savage Mind*. Weidenfeld & Nicholson, London

Lindquist G 1997 *Shamanic Performances on the Urban Scene: Neo-Shamanism in Contemporary Sweden* (Stockholm Studies in Social Anthropology 39). Almqvist & Wiksell International, Stockholm

Miller S 1973 Ends, means, and galumphing: Some leitmotifs of play. *American Anthropologist* **75**: 87–98

Minnema L 1998 Play and (post)modern culture: An essay on changes in the scientific interest in the phenomenon of play. *Cultural Dynamics* **10**: 21–47

Safa-Isfahani K 1980 Female-centred world views in Iranian culture: Symbolic representations of sexuality in dramatic games. *Signs* **6**: 33–53

Schechner R 1988 Playing. *Play & Culture* **1**: 3–19

Schwartzman H B 1978 *Transformations: The Anthropology of Children's Play*. Plenum, New York

Spariosu M I 1989 *Dionysus Reborn: Play and the Aesthetic Dimension in Modern Philosophical and Scientific Discourse*. Cornell University Press, Ithaca, NY

Sutton-Smith B 1997 *The Ambiguity of Play*. Harvard University Press, Cambridge, MA

Turner V W 1969 *The Ritual Process*. Aldine, Chicago

D. Handelman

Plea Bargaining

Research into plea bargaining 'in action', not plea bargaining 'on the books,' has been conducted, for the most part, by studying US federal and state courts. Nonetheless, the central theme of this article is that the findings of this research are applicable in a comparative context and can, with some qualifications, help us to understand plea bargaining around the world. Plea bargaining may take different forms, and it might be referred to, or conceived of, as something different, but it does exist in criminal justice systems outside the USA. By examining US studies and the small amount of comparative research available, we can speculate that plea bargaining is a function of the raw material that courts process, and that if we permit an expanded definition of plea bargaining—even allowing for the particularities of various jurisdictions—we will see that this manner of case disposition is a near-universal province of courts. Space constraints preclude a systematic examination of what we know about these dynamics in a comparative sense; this article will discuss general themes and questions.

1. The Typology of Plea Bargaining

While scholars have concluded that plea bargaining was common in nineteenth century British courts (Zander 1998, pp. 323–34), most comprehensive studies have focused on the USA, and most have been conducted in the past 25 years. (For a summary of the early work on plea bargaining, as well as an agenda for future plea bargaining research at the time, see *Law and Society Review* 13 (2), 1979; for useful citations on subsequent related work see McCoy 1993, Vogel 1999.) In the early days, there was a debunking tone to this work, as scholars questioned and tested the assumption that trial courts 'try' cases. What was viewed as anomalous in an earlier period of US history—something which might have characterized the more political (and frequently 'sleazy') municipal courts—plea bargaining is now understood to be the central method of processing cases in trial courts. Plea bargaining was not an aberration of US courts; it was the norm. In what might be viewed as the epitaph for the role of trials in trial courts, a judge noted: 'I don't think that a trial necessarily represents a failure in the plea bargaining system. I really believe that the trial is a necessary adjunct of the system' (Heumann 1977, p. 150).

While most scholarly attention has been focused on plea bargaining in the USA, recent comparative endeavors have produced data that afford us the opportunity to understand plea bargaining as it exists (in many forms) around the world. While no two nations' legal systems or plea bargaining experiences are exactly alike, we are at the point where we can begin to construct typologies (explicit, policy, and implicit plea bargaining) that can and should invite future research. Notably, while different forms of plea bargaining will continue to exist, there is a kind of convergence at work, as the institutions, attitudes, and processes of nations around the world begin to resemble each other when it comes to the actual practice of plea bargaining.

1.1 Explicit Plea Bargaining

Plea bargaining in the USA in its most basic and commonly understood form, is the process of active, explicit negotiation between the defense attorney and prosecutor concerning the resolution of the defendant's case. Most frequently the 'deal' revolves around the benefits that will flow to the defendant if he or she chooses to plead guilty. These might include a reduction in a charge, a dismissal of a charge or charges, or a prosecutorial sentence recommendation (or an agreement for the prosecutor not to press a particular sentence). Also, increasingly, the prosecutor might agree not to add a mandatory sentence charge to the defendant's indictment. Any, or many, of these will be given to the defendant in return for a guilty plea. Generally the exchange is explicitly negotiated, and a judge may or may not also be part of the negotiating process. When not involved, the judge usually goes along with the negotiations; however, if it is felt that the agreement is not acceptable to the judge, the defendant is usually allowed to withdraw the plea.

This form of plea bargaining, as indicated above, is one in which a negotiation is undertaken explicitly and the parameters of the agreement are agreed upon specifically. While the USA (especially since the late 1960s following the US Supreme Court's approval of plea bargaining) represents the best example of a system that embraces explicit plea bargaining (indeed, the explicit variety is the one that is conjured up when comparisons with the USA are made), other nations which have experienced increases in case pressure have sought relief through increases in explicit plea bargaining. While it has been proven that this relationship is not a causal one in American courts (Heumann 1977), we can see a correlative relationship in systems such as South Africa's, for example, which is considering the expansion of the realm of explicit bargaining (and inviting judges into the process) as a response to its increasingly overwhelmed courts (Schwikkard, and van der Merve 1999, p. 352). It is important to note that plea bargaining (involving prosecutors and defendants) was already common in South Africa, and thus was not caused by a case overload; what we are seeing now, in other words, is a development of earlier practice. Other examples of nations that employ more explicit practices include Israel, which allows on-the-record plea bargaining (Harnon and Stein 1999, pp. 239–41), and Canada, which openly encourages pleas with regard to charge and sentence (Roach 1999, p. 75).

1.2 'Policy' Plea Bargaining

A second kind of plea bargaining, neither explicit nor implicit *per se*, is derived from, and relies upon, the policies a prosecutor's office may have adopted over time. Specifically, the prosecutor may announce that his/her office will not negotiate in specific types of cases (e.g., when a firearm is involved, or certain drug matters). But the prosecutor, without explicit negotiation will, as a matter of policy, substitute some lower charge for defendants who plead guilty to a 'policy case.' Thus defendants (or defense attorneys) can be reasonably certain about the benefits that will attach to a plea of guilty. This form of plea bargaining seems to be unique to the USA—a product, seemingly, of the mandatory sentencing guidelines that have stripped some courtroom actors increasingly of the discretion necessary to make explicit deals.

1.3 Implicit Plea Bargaining

A third kind of plea bargaining—what we might label 'implicit plea bargaining'—takes place whenever a

defendant pleads guilty, with the expectation that the plea will lead to a reward in terms of charge and/or sentence reduction. Unlike the first kind of plea bargaining there is no explicit negotiation; unlike the second type there is no specific policy reward for a plea. In this kind of plea bargaining, the defendant hopes that a plea—a gesture of cooperation or contrition—will yield a beneficial disposition of the case.

Plea bargaining in the USA generally took the 'implicit' form until the 1960s, and we can see that in Japan, for example, defendants will often expect a reward for their plea, especially as the national government continues its crackdown on organized crime. (For an interesting discussion of why 'elite' USA Federal judges preferred implicit plea bargaining, see Padgett 1990.) Implicit plea bargaining is, based on data we have on other countries, the most common form in the world, but it is also the most difficult to study.

Many nations deny that plea bargaining takes place, and some (the UK, for example) even make plea bargaining illegal. In the early 1970s, several official arms of the government actually tried to stop the publication of a social science study based on defendant interviews, a study which claimed that extensive coercive plea bargaining characterized the British system (Baldwin and McConville 1977). And though a subsequent study qualified the contention about the coerciveness of the process, it still maintained that a variant of what can fairly be called plea bargaining thrived in the UK (Seifer 1980, pp. 179–97).

Japan also takes a more negative attitude toward plea bargaining, declaring it to be 'illegal' (Johnson 2001, p. 2). Thus if one were to believe the 'law on the books,' one would have to conclude that Japan stands as a clear exception to the argument that plea bargaining is both central to criminal justice systems and quite common as a practice. Yet in one of the few empirical studies of Japanese law in action, Johnson demonstrates that, despite the formal law to the contrary, several modes of plea bargaining are common in the Japanese criminal justice system. An aversion to the 'crassness' of the American form of plea bargaining, coupled with the purported illegality of this activity, may lead to denials of plea bargaining's existence. But these denials are simply wrong—plea bargaining, in various forms, is alive and well in Japanese courts.

We can see this version of plea bargaining at work in some cases in Japan, and in Argentina and Italy (Sanders 1966, pp. 335–9, Carrio and Garro 1999, p. 43, Van Cleave 1999, pp. 272–4). In Japan, abbreviated—very abbreviated—'trials' are really uncontested presentations of evidence, which frequently are called 'slow pleas of guilty.' Running through these alternatives and others is something like the notion of implicit plea bargaining, the idea that not contesting a claim (inquisitorial systems typically do not admit a guilty plea) leads to an expectation of leniency in disposition. In France, to offer yet another example, defendants feel that cooperation clearly leads to sentence reductions (Frase 1999, p. 169).

One thing that tends to disguise plea bargaining in other nations (and to allow for official denials) is the existence of functional equivalents. These equivalents represent this form of negotiation in its most 'implicit' manner. In Germany, for example, the prosecutor can either waive cases if it seems there is no public interest in prosecuting them, or issue a 'punitive order' (more of an administrative than a criminal resolution of the case) when it appears that trial alternatives are not desirable. 'The prosecutor's offer of a conditional waiver as well as the punitive orders have been considered a German variant of plea bargaining.' (Blankenburg 1996, pp. 284–5, see also Weigand 1999, pp. 208–9).

2. *Trends in Plea Bargaining in the USA*

Because plea bargaining can take place in a variety of ways—all of them by definition subtle and discrete—this manner of bargaining presents scholars with the greatest challenge; to perceive, isolate, and understand negotiations of this kind seems to require the kind of comprehensive comparative studies that have not yet been performed. The implications for defining plea bargaining in these three ways ought to be made clear. Explicit bargaining, for a long while, was viewed as illegitimate, with judges in court often concerned with demonstrating for the record that nothing had been promised in return for a plea (see, e.g., Blumberg 1967, pp. 88–94); so for much of the twentieth century, plea bargaining in the USA assumed either the 'policy' or 'implicit' form. The recognition of the centrality of explicit plea bargaining by the US Supreme Court, however, rendered this unnecessary, and enabled plea bargaining to take place more openly in the court (*Brady* vs. *US*, 397 US 742 (1970); *North Carolina* vs. *Aflord*, 400 US 25 (1970); and especially *Santobello* vs. *New York*, 404 US 257 (1971); see also *Bordenkircher* vs. *Hayes*, 434 US 357 (1978)). With the advent of mandatory sentencing guidelines, however, many jurisdictions around the country have been forced to return to the more subtle, implicit manner of bargaining—moving the USA closer to the 'backroom' style of bargaining that takes place around the world and exemplifying the 'convergence' discussed above.

Policy plea bargaining and implicit plea bargaining, on the other hand, have more subtle implications. Both allow jurisdictions—prosecutors in particular—to claim that they do not plea bargain, or that plea bargaining in a category of cases has been proscribed. The symbolic value of such claims should not be underestimated. The 'bargaining' image of plea negotiations makes the explicit process vulnerable to accusa-

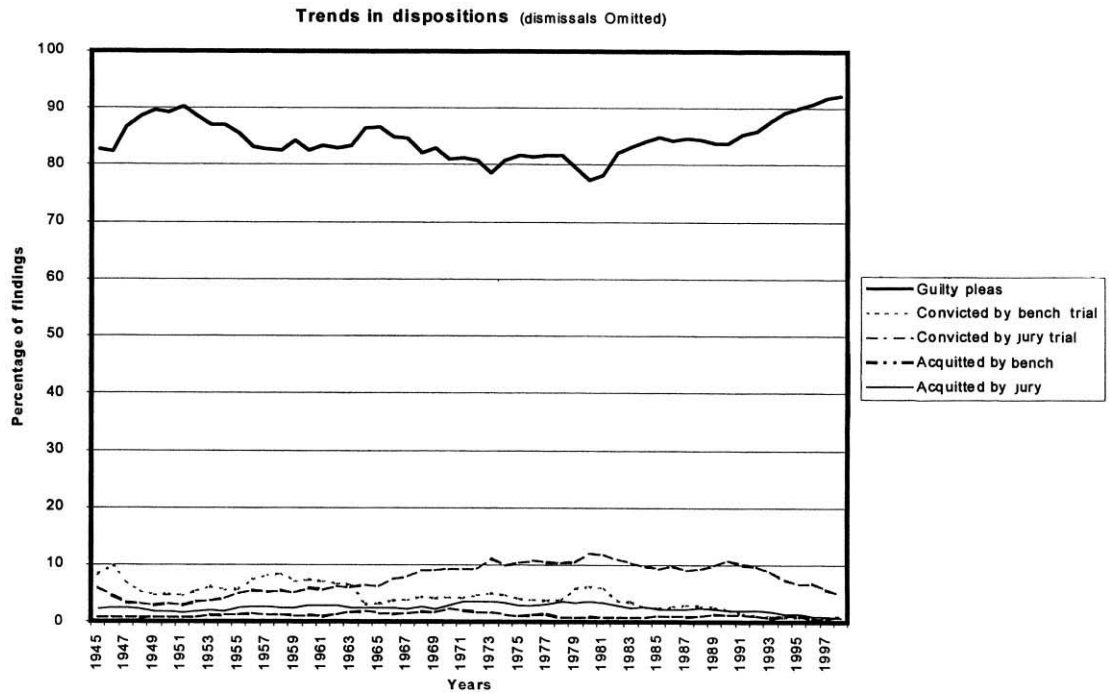

Figure 1
Guilty pleas (80–90 percent) compared with other outcomes of cases brought before the US courts 1945–97. Source: US Department of Justice, Bureau of Statistics

tions of 'bargaining with justice.' This, one might speculate, is why so few nations have engaged in the explicit form of plea bargaining. Even the USA, as we can posit from studies of the impact of sentencing guidelines, is beginning to move back toward the policy and implicit varieties of bargaining. (To appear to be in conformity with a 'no plea bargaining' policy, a court system may also make increased use of bench trials. These are often very abbreviated trials in which the defendant does not contest the case meaningfully, but nonetheless has an expectation of some benefit because he/she has waived a jury trial. See, for example, Heumann and Loftin 1979, pp. 417–21.) Yet if defendants still plead with the expectation of benefit, plea bargaining, less explicit, but plea bargaining nonetheless, can continue to flourish.

Some illustrative data on guilty plea rates are instructive. Figure 1 graphs guilty pleas, bench trials, and jury trials (convictions and acquittals) for US District Courts from 1945 through 1998 (the last year for which these data were available). The pattern is quite clear—roughly 80 percent of the defendants whose cases are not dismissed plead guilty. The very limited role of full trials in US courts is evident. Comparably, if we look at some recently reported data from three typical counties in a Midwestern state, guilty pleas(again for cases not dismissed) range from a high of 97 percent to a 'low' of 87 percent (see Hannaford and Munsterman 1999, p. 66). One cannot help but be struck by the prevalence of guilty pleas in Federal and state courts in the USA. One would be hard pressed to find an American jurisdiction in which the guilty plea—or its functional equivalent (e.g., *pro forma* bench trials, or other forms of 'slow pleas of guilty')—does not comprise at least 70 percent of dispositions, with the usual percentage considerably higher. (Regarding these pleas in their many forms, see Mather 1979, pp. 55–6, 69–70, 78–9.)

As the 1945 data suggest, and other data demonstrate, plea bargaining is not a new phenomenon in US courts. Defendants pleading guilty for a putative reward and waiving the 'Miranda rights' (rights of the citizen when faced with police interrogation, such as the right to remain silent, right to legal counsel etc.) are the norm, not the exception in these courts, and this pattern has been consistent for some time (see, e.g., Heumann 1977, pp. 29–30, Padgett 1990, pp. 413–50, Zander 1998, pp. 323–4, Vogel 1999, pp. 161–246). Indeed, at least one scholar has argued that plea bargaining increased with the adversarial nature of the courts (Feeley 1997, p. 323).

Contrary to a popular belief, and indicated by its history and its prevalence in our courts at the time of writing, plea bargaining must reflect something other

than a response to caseload pressure. The common belief that plea bargaining is caused by case pressure has been disproved by studies of American courts, and, based on the research we have thus far, we can speculate reasonably that it would be disproved in a comparative context as well. While we are aware that plea bargaining takes place (in some form) in the nations that have been studied, we are still unsure as to *why* this manner of negotiation takes different forms in different nations. That is, why is it that the UK tends to practice and to view plea bargaining in one way, while Japan proceeds in another? The results and motives are often the same—prosecutors induce (or accept) a guilty plea and deliver (or end up providing) some kind of 'reward.' So what is it, specifically, that distinguishes nations and their experiences in a comparative context? Would we find the answer in a study of institutions? Culture? History? We both invite and eagerly await scholarship that concentrates on these important questions.

3. *Why Plea Bargaining?*

To develop a rationale for plea bargaining, we must begin with the defendant's expectations/preferences. Though data may be unclear and variable depending on circumstances, most defendants believe that in most cases 'it pays to plead guilty' (Brereton and Casper 1982, pp. 45–70). Defendants often believe that their plea will be rewarded in terms of charge and/or sentence reductions. Why they think the system so rewards them varies with defendants, and seems to reflect the variety of justifications for such a reward advanced by prosecutors, judges, and defense attorneys. Some feel that the system appreciates the appearance of contrition that a plea represents—a theme frequently heard in court is that 'pleading guilty is the first step to rehabilitation.' Others feel that they are saving the state the time and expense of a trial, while others who provide information about associated crimes and criminals believe they should earn a reward for their contribution. Indeed, some Japanese scholars urge an increase in plea bargaining in order to obtain more information from defendants about organized crime (Johnson 2001, p. 35).

Finally, defendants feel that by pleading guilty they ensure the state some certainty—a certainty they themselves, in fact, enjoy more. In this sense defendants might be characterized as, almost naturally, in 'plea bargaining mode,' they seem to enter the system expecting a bargain. A study of the intentions and expectations of defendants, across nations and cultures, could shed some light on the internal dynamics of criminal justice systems around the world. Is 'bargaining mode,' for example, a product of a 'bargaining' culture—or one reputed to be interested in 'dealing'—or do defendants naturally have this inclination without state encouragement? A study that controlled for other nations' reputations and cultures might yield answers to this question.

While defendant choices and expectations drive the system, complementary calculations by the system's professional actors actually shape the defendants' views. Prosecutors, as we know from studies of American courts, are influenced by saving money and time in many of their cases and (all things being equal) value defendant pleas. Also, to the prosecutor, even in the many open-and-shut cases (often known colloquially as cases which are 'born dead') that come through the office, there is always some uncertainty about post-trial outcomes; pleas remove any uncertainty, no matter how low the probability of acquittal (Mather 1979, pp. 41–2, 65–72. For the comparable concept of cases that are 'born dead' see Heumann 1977, p. 60). Without the data to predict with certainty, it seems reasonable to suggest a kind of 'prosecutorial culture' that might transcend national borders and even legal systems. That is, while we should be careful not to generalize too much, it seems reasonable to expect significant similarities in the attitudes (and certainly headaches) of prosecutors around the world.

Defense attorneys make similar assessments about minimizing pointless work and about ensuring certainty. Finally, judges also value the ostensible finality of pleas, and the reduction of the probability of appeal after a plea. Avoiding reversal on appeal, we know from the American experience, is one factor that is said to weigh large in the support that judges give to pleas. But does the same concern apply in other settings and legal traditions? Again, these actors (and others) are an active part of legal systems around the world. What then is the explanation for the different plea bargaining patterns we see in our initial comparative assessment?

While the above discussion outlines adequately what we know of plea bargaining up to the turn of the millennium, the standard explanations for plea bargaining ought not to obscure two very important considerations. The first we might call 'all is not rosy in the plea bargaining world.' The explanations we have given read as if defendants walk arm-in-arm with prosecutors, defense attorneys, and judges into the wonderful world of pleas. To some extent, many guilty defendants do take pleas, do so without coercion, and do so aware of the advantages of their choice. But all is not rosy, particularly for defendants who want to opt for trial. There is no question that in the US plea bargaining system there is generally a very strong inclination toward a plea; defendants who insist on trials are often in conflict with this clear reversal of the 'presumption of innocence' standard. Though we do not have good evidence for this, defendants may suffer harsher (i.e., harsher than the plea offer) dispositions upon conviction. Though these dispositions can be rationalized (there is no penalty for trial; pleas were rewarded because of the possibility that a trial might

lead to an acquittal; defendant not penitent; etc.), the specter of trials being viewed as unacceptable options in a plea bargaining world is distasteful.

Second, it is important to clarify the case-pressure/plea-bargaining argument. We have said that case pressure does not cause plea bargaining; both contemporary courts and those of the past with no realistically conceivable case pressure plea bargained most of their cases; crowded or uncrowded courts plea bargain for the reasons we have indicated. However, to say that plea bargaining is not caused by case pressure is not to say that case pressure may not affect plea bargaining; for many US courts, volume may be substantial and staffing less than optimal. In many courts, case pressure may affect dispositions, but we do not have a clear sense of how this might operate. Several possibilities come quickly to mind—rewards for pleas, which may exist in a 'no pressure' system, might become even greater; the reward for a plea may stay the same, but the 'penalty' for trial may be increased; the prosecutor may choose to screen out more cases through dismissals.

Case pressure, while not the cause of plea bargaining, can still affect the nature and form of the negotiation entailed, and we need more comprehensive comparative data on the effects of this pressure in other national settings. Recognizing that plea bargaining is fundamentally not a function of case pressure helps us to understand the centrality of pleas in US courts; acknowledging that case pressure can affect court disposition patterns helps us to understand why so many assume (incorrectly) that plea bargaining was adopted to manage heavy caseloads. We can speculate as to the relationship between plea bargaining and case pressure in other nations, but without data that considers the multiple variables involved, we are limited to our conjecture.

4. *External Influences on Plea Bargaining: The USA as a Case Study*

The single most dramatic change in American criminal court processes in recent years has been the increased use of mandatory sentences in particular, and adoption of sentencing guidelines in general (see *Punishment: Social and Legal Aspects*). In an effort to reduce judicial discretion in sentencing, the US Congress and individual states have circumscribed this discretion by requiring mandatory minimum sentences in certain cases, and generally requiring sentences within a narrow sentencing grid for defendants with particular background characteristics (e.g., number of prior convictions) who are convicted of particular crimes (see Tonry 1996, pp. 6–12, DuBow 1998, pp. 231–45). However, in limiting the open exercise of judicial discretion in the interests of reducing disparity in outcomes, the effect has actually been to increase (but mask) the discretion of prosecutors, while confining these activities to the 'backroom.' By constraining judicial discretion, prosecutorial discretion has been enhanced; restricting sentencing ranges for judges has promoted less visible, but equally consequential, sentencing practices by prosecutors.

The significance of increased prosecutorial discretion in the USA, as suggested above, seems to be that—among other things—American plea bargaining is beginning to appear more like the negotiations that take place in other nations. This 'convergence' as we might call it, has tempered both the desire and capacity for actors in the American system to engage in explicit plea bargaining; 'deals,' therefore, increasingly must be of the 'implicit' or 'policy' variety. While this makes plea bargaining more difficult to study—relegated, as it is, to the literal and metaphorical 'backrooms'—it presents scholars with an interesting series of questions to consider.

5. *Conclusion*

As we move beyond the exploration of plea bargaining itself, a number of very important matters—e.g., the relationship of plea bargaining to sentencing reform and to court caseloads, the effects of culture, traditions, and institutions on plea bargaining activities, and the appearance of a convergence of plea bargaining practices, move to the fore. These are matters that should be high on our research agendas as we begin to fashion a new, better informed, and less naive set of plea bargaining questions for the twenty-first century.

See also: Courtroom Consultation: Psychology; Courts and Adjudication; Criminal Defenses; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Due Process in Law; Negotiation and Bargaining: Role of Lawyers; Punishment: Social and Legal Aspects; Trials: Cultural

Bibliography

Baldwin J, McConville M 1977 *Negotiated Justice: Pressure to Plead Guilty*. Martin Robertson, London

Blankenburg E 1996 Changes in political regimes and continuity of the rule of law in Germany. In: Jacob F, Blankenburg E, Kritzer H, Provine D M, Sanders J (eds.) *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT, pp. 249–314

Blumberg A 1967 *Criminal Justice*. Quadrangle Books, Chicago

Bradley C (ed.) 1999 *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC

Brereton D, Casper J 1982 Does it pay to plead guilty? *Law and Society Review* **16**: 45–70

Carrio A, Garro A 1999 Argentina. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 3–52

DuBow S S 1998 Introduction to models for sentencing reform in the United States and effects on the sanctioning process. *Law and Policy* **20**(3): 231–45

Feeley M 1997 Legal complexity and the transformation of the criminal process: The origins of plea bargaining. *Israel Law Review* **31**:(183)
Emmelman D 1996 Trial by plea bargain: Case settlement as a product of recursive decisionmaking. *Law and Society Review* **30**(2)
Frase R 1999 France. Bradley C (ed.) 1999 *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 143–85
Goldstein A, Marcus M 1977 The myth of judicial supervision in three inquisitorial systems: France, Italy, and Germany. *Yale Law Journal* **87**: 240–83
Hannaford P, Hans V, Munsterman T 1999 How much justice hangs in the balance. *Judicature* **83**(2): 66
Harnon E, Stein A 1999 Israel. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 217–44
Heumann M 1977 *Plea Bargaining: The Experiences of Prosecutors, Judges, and Defense Attorneys*. University of Chicago Press, Chicago
Heumann M, Loftin C 1979 Mandatory sentencing and the abolition of plea bargaining: The Michigan felony firearm statute. *Law and Society Review* **13**(2): 417–21
Jacob F et al. (eds.) 1996 *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT
Johnson D 2001 Plea bargaining in Japan. In: Feeley M, Miyazawa S (eds.) *Sho Sato Collection*. Macmillian, London
McCoy C 1993 *Politics and Plea Bargaining*. University of Pennsylvania Press, Philadelphia, PA
Mather L 1979 *Plea Bargaining or Trial?* Lexington Books, Lexington, MA
Murray J, Rau A, Sherman E (eds.) 1998 *Process of Dispute Resolution: The Role of Lawyers*, 2nd edn. Foundation Press, Westbury, NY
Padgett J 1990 Plea bargaining and prohibition in the Federal courts, 1908–1934. *Law and Society Review* **24**(2): 413–50
Provine M 1996 Courts in the political process in France. In: Jacob F et al. (eds.) 1996 *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT, pp. 177–248
Roach K 1999 Canada. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 53–80
Sanders J 1996 Courts and law in Japan. In: Jacob F, Blankenburg E, Kritzer H, Provine D M, Sanders J (eds.) *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT, pp. 315–88
Schwikkard P, van der Merve S 1999 South Africa. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Caroline University Press, Durham, North Carolina, pp. 319–60
Seifer R 1980 Plea bargaining in England. In: McDonald W, Cramer J (eds.) *Plea Bargaining*. Lexington Books, Lexington, MA, pp. 179–97
Stith K, Cabranes J 1998 *Fear of Judging*. University of Chicago Press, Chicago
Tonry M 1995 *Malign Neglect—Race, Crime, and Punishment in America*. Oxford University Press, New York
Tonry M 1996 *Sentencing Matters*. Oxford University Press, New York
US Department of Justice 2000 *Sourcebook of Criminal Justice Statistics 1998*. Bureau of Justice Statistics, Albany, NY
Van Cleave R 1999 Italy. In: Bradley C (ed.) *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 245–82
Vogel M 1999 The social origins of plea bargaining: Conflict and the law in the process of state formation, 1830–1860. *Law and Society Review* **33**(1): 237–43
Weigand T 1999 Germany. In: Bradley C (ed.) 1999 *Criminal Procedure: A Worldwide Survey*. Carolina University Press, Durham, NC, pp. 187–216
Zander M 1998 Plea bargaining goes back a hundred years. *New Law Journal* **148**(6830): 323–24

M. Heumann

Plural Societies

The idea of the 'plural society' was first coined by a British colonial administrator, Furnivall, on the basis of practical experience in Southeast Asia, in his classic volume, *Colonial Policy and Practice: A Comparative Study of Burma and Netherlands India* (1939), more as a descriptive device than as theory. It reflected the condition, common in colonial and later, newly independent postcolonial states, of deeply divided societies, usually along ascriptive lines of ethnicity or race (sometimes called 'communal'), but also of culture, religion, or language. Characteristically, such societies are deeply segmented, with weak institutional integration and little consensus over 'national' agenda, unity, or identity. These tendencies are enhanced by the fact that the core values of each of the constituent (ethnic or other) communities are often incompatible, and that they each pursue separate goals and interests, in a fashion which Furnivall described as 'mixing' without 'combining,' and 'meeting only in the market place,' for basic economic needs. In the colonial context such societies were held together largely by coercion rather than consensus. Boeke, a Dutch colonial administrator in Furnivall's tradition, developed similar ideas about segmented societies in his *Economics and Economic Policy of Dual Societies, Indonesia as Exemplified by* (1953), which he conceptualized in binary ('dual') terms, highlighting the colonial/indigenous cleavage, particularly in the domain of modern vs. traditional economic institutions.

One common source of confusion arises around two diametrically opposed meanings of the form 'pluralism' which may refer either to the *plural* society, or to its obverse, viz. a cohesive *pluralist* society integrated through a multiplicity of cross-cutting linkages, as conceptualized by such political scientists as (Dahrendorf 1959). This represents a continuation of the tradition of Alexis de Tocqueville, whose model of nineteenth-century America as knit together by webs of associations, helped to neutralize ethnic and other divisions.

Beginning in the 1960s, when most colonial states achieved independence, the term 'plural society' migrated into academic discourse, as anthropologists, sociologists, and political scientists attempted to inject

some theoretical rigor into the study of such weak states, held together as much by coercion as by the market. Among the pioneers were Smith and Kuper writing initially on (especially South) Africa and the West Indies. Together, they edited *Pluralism in Africa* (Kuper 1969), while Smith also authored *Stratification in Grenada* (Smith 1965b) and *The Plural Society in the British West Indies* (Smith 1965a). Smith and Kuper distinguished three types, or levels of pluralism. Cultural pluralism, as a form of extreme heterogeneity, sets the necessary, though not sufficient, preconditions for social pluralism, which is marked by the existence of relatively closed communities with separate core values and institutions and engaging in few cross-ethnic contacts. Finally, where constituent groups within a state are 'differentially incorporated,' or have unequal civil and legal status as citizens, there exists a situation of structural pluralism, a feature often found in racially divided societies. These writers, and to some extent Van den Berghe (1970), largely due to the particularities of the societies they studied, were most concerned with the type of pluralism known as ethnic stratification. Most common in 'white-settler' (hence, not strictly postcolonial) societies, social structure is envisaged in the form of a vertical hierarchy of ethnic/racial communities, which are simultaneously ranked as classes (occasionally called 'eth-classes'), the most egregious example of the era being apartheid South Africa.

In other world regions, such as Southeast Asia, another kind of plural society was identified, where the ethnic/cultural/communal sections, rather than being ranked hierarchically, live side-by-side as 'parallel,' 'pillared,' or 'segmental' communities (Benedict 1962, Morris 1967), each largely institutionally self-sufficient. Characteristically, individuals in such societies are primarily identified as members of their collectivity (or 'ethnicity'), which constrains them in most activities and life-chances, from religion, kinship practices, education, and occupation, to political party membership. In maximally plural situations (Despres 1968), there is little opportunity for autonomous individuals to act independently or to forge cross-community ties without reference to the 'basic givens' of their birth status. 'National' unity consists of little more than opportunistic alliances of multi-ethnic elites, while most other societal institutions and relations remain divided along ethnic lines. Societies of this type included Indonesia (Geertz 1963), Malaysia (Ratnam 1965, Nagata 1979), and Mauritius (Benedict 1962), where the presence of colonially-imported immigrant communities of Chinese and Indians exacerbated the cleavages and even conflict. In these places, and in such comparable plural states as Nigeria, Sri Lanka, Guyana, and Fiji, even national political elections are fought along communal lines, where the principal unity and co-operation emerges at the top, through cross-ethnic connections among elites for their mutual advantage. Political scientists describe such electoral arrangements as 'consociational' (Lijphart 1977), and these are probably implicated in the familiar 'crony-capitalist' style of leadership of many rapidly developing states, such as Indonesia and Malaysia.

Where the impetus for linking social, economic, and political institutions and issues to ethnic identity is state policy, there is little prospect for any form of assimilation ideology or practice, with all its implications for lack of national identity or shared 'imagined community.' In face of internally contested goals and identities, departments or ministries of 'national unity' are sometimes created to impose or direct what daily living cannot, but without some 'depluralization' (Kuper and Smith 1969), both the objective structures and subjective perceptions of separateness persist. Until this happens, in the view of some political scientists of the 1960s and 1970s, societies are not yet politically modern (Inkeles 1969). What is needed to overcome these problems, according to Geertz (1963) is an 'integrative revolution,' whereby transcendent civic loyalties come to supercede particularistic ones, or turned to the benefit of a more cohesive state. In effect this involves the creation of more cross-cutting institutional and individual networks in the public domain, based on superethnic principles and issues, of a neutral political space corresponding to a pluralist or even to a 'civil society'. While it has been claimed (Eriksen 1994) for Mauritius, that such depluralization is achievable over time, particularly under the impact of an open market economy, and arguably that even South Africa is moving in this direction, it is at best an unpredictable and oscillating process. Evidence (Eriksen 1994) does suggest that some communal cleavages can be retooled to work for, rather than against, national unity as in Mauritius and Trinidad, whereas post-Suharto Indonesia, post-Abacha Nigeria, and contemporary Sri Lanka, on the other hand, remain vulnerable to the centrifugal tendencies of their internal plural divisions. As Tambiah (1989) has suggested, the initial post-colonial euphoria of nation-building may be sufficient in the short term to paper over the plural cracks, but they do seem to re-assert themselves subsequently as a new round of internal ethnic resource and power politics, or 'retribalization' (Cohen 1974) comes into play. The language of 'ethnonationalism' covers many of the same problems, in the sense that ethnonational movements represent politically mobilized ethnic communities, which may eventually be consummated in an independent or secessionist state, in the manner of Croatia, Serbia, Macedonia, and Bosnia of the 1990s. Others are not (yet) consummated, but remain poised for secession from larger (plural) states, as in Quebec or the Tamil northeast of Sri Lanka, exemplifying the perennial problem of such polities.

In one variation of the plural society's political and socially fissiparous tendencies, the binary type is considered by some observers (Milne 1981) to be the

most fragile. In his selections of 'bipolar' plural states, notably Malaysia, Guyana, and Fiji (to which could be added Sri Lanka and Belgium), the argument is made for the presence of a zero-sum or winner/loser mentality in matters of resource and power distribution, such that A's gain is invariably construed as B's loss, and vice versa. In situations of multiple division, where three or more communities are present, there are more possibilities for cross-cutting coalitions, whereby groups make different alliances according to interest, such that one can neutralize another, and this is particularly clear in the case of political party relations. In practice, however, even in situations where three or more nominal parties or plural segments exist, the actual arrangements may operate in *de facto* binary fashion. Thus in Malaysia, where three principal ethnic-based and several other political parties play the electoral field, in most major issues the public rhetoric is framed in terms of Malays (or Muslims) vs. non-Malays (or non-Muslims), the second category consisting of a collective assortment of other ethnic and religious constituencies. In Indonesia, the plethora of indigenous ethnic groups are packaged politically and symbolically as one (*pribumi*) in face of the (relatively tiny) immigrant Chinese community. Similarly, in Canada, the array of diverse immigrant ethnic communities count as a single unit (the 'Rest of Canada/ROC') or Anglophone, in opposition to Francophone Quebec in the business of federal politics. The reduction of multi- to bipolarity significantly affects the oppositional character and abrasiveness of internal relations, at least at the maximal political level.

Plural constructions of society tend to focus exclusively on the divisiveness of ethnic, racial and cultural, or religious identities and their derivative institutions. Conspicuously missing in most of the plural literature is any reference to class stratification and its potential for the creation of cross-cutting cleavages, or, viewed from another perspective, of unity in a different direction. Where class enters the picture at all, as in the writings of Kuper and Smith, the emphasis is on the fact that each ethnic/racial selection is internally stratified by its own unique set of strata, none of which correspond with those of any other, hence their cooperative, class-action potential is neutralized. No South African black miner would find common cause with a white worker in the same sector, nor a Malay with a Chinese rubber tapper. However, the deflection of class to ethnic/racial images is often as much the product of a state worldview or agenda, whose interest it is to frame all local issues in ethnic terms and to rule by playing off one community against another. This is also intended to a avoid labor unrest or economic instability, and is reinforced by banning unions (or channeling them along ethnic lines), and other class-like activities. One consequence is that members of such societies spontaneously perceive all inequalities as based on ethnic/communal or sometimes religious statuses. It is this kind of mindset which causes the Chinese in many countries to be unilaterally classified as the 'exploiters,' while conveniently ignoring the existence both of poor Chinese and rich local elites. Furnivall's 'meeting in the market place' does not necessarily seem to create the de Tocquevillian or Dahrendorfian conditions necessary for a pluralist society.

The lack of attention to the potential or existence of class expression in plural societies may be as much ideological as a reflection of the actual situation. In contrast to Marxian convictions that the fundamental determinant of social inequality is material, and that ethnicity is merely an epiphenomenon of class, plural society theorists maintain that ethnic and racial identities are *sui generis*, the cause of most social inequality and conflict.

Since the late 1970s, the term 'plural' seems to have yielded to or been superceded by the label 'multicultural' in many places. A careful examination of so-called multicultural societies today reveals a number of parallels with the older 'plural' term, in its emphasis on ethnic, racial, or cultural identities as the primary source of diversity and division, often at the expense of class. Some commentators maintain that states which promote multicultural policies, which make a virtue of ethnic difference, and even design many social activities and programs along these lines, are a cynical official device ('myths') to obscure more basic class or race inequalities and lack of opportunities for mobility or political empowerment (Peter 1981) for Canada. Proponents of this view suggest that the choice of the term 'culture' as the operating principle conveys a bland, benign, and misleading image of tolerance and equality of opportunity, turning 'difference' away from ethnic conflict and to the benefit of national unity. Given that most societies today are in the process of becoming demographically, socially, religiously, and culturally more multicultural through immigration, even the older, more stable states of Europe, once confident of their national character and identity, are now rethinking these as they come to terms with unaccustomed internal ethnic and cultural diversity. Particularly pressing are the handling of religious and legal pluralness by the host society. The actual policy responses to *de facto* multiculturalism, however, vary by country and over time, as Germany's (1999) reconsideration of the status of its largely Turkish 'guestworkers' (*Gastarbeiter*) illustrates. In the case of the US, the long-established ideology of assimilation is challenged by ideas of multiculturalism as policy (Walzer 1992, Waldron 1992). Increasingly, the emphasis is on the acceptance of difference rather than on the promotion of homogeneity. Also at issue is the question of recognition of (ethnic/cultural/racial) collectives or group rights vs. the more established privileging of the individual and of individual rights as the underpinning of western liberal democracy (Kymlicka 1995). This has implications for

such 'corrective' policies introduced by some contemporary plural states in the direction of quotas or of affirmative action, which could be construed as benign attempts to condone or perpetuate pluralness (Porter 1975), and hark back to what could be interpreted as compromising the standards of liberal democracy and a civic nation. This dilemma, of simultaneously recognizing group identities and integrity while maintaining liberal democratic ideals, is most articulately treated in Taylor (1992) dissections of Canadian multiculturalism.

A general survey of current world trends indicates a declining emphasis on assimilation as a route to nation-building, from Israel to Australia, largely as a response to accelerated and extensive transnational movements and to the existence of deterritorialized communities. This raises new problems of internal social cohesion within states, as they face the question as to whether multiple identities (and even citizenships) among their populations, (who may simultaneously also be participating in the activities of transnational communities offshore), are necessarily mutually exclusive or to the detriment of the state. The original, rather divisive forms of pluralness may, in certain circumstances, form the foundations of a new style of pluralistic politics of diversity or difference, in dialogic as much as conflictive relations, operating between, as well as within states.

See also: Ethnicity: Anthropological Aspects; Multiculturalism, Anthropology of

Bibliography

Benedict B 1962 Stratification in plural societies. *American Anthropologist* **64**(6): 1235–46

Boeke J 1953 *Economics and Economic Policies of Dual Societies, as Exemplified by Indonesia*. Institute of Pacific Relations, New York

Cohen A 1974 *Urban Ethnicity. ASA Monographs, No. 12*. Tavistock, London

Dahrendorf R 1959 *Class and Class Conflict in Industrial Society*. Routledge & Kegan Paul, London

Despres L A 1968 Anthropological theory, cultural pluralism and the study of complex societies. *Current Anthropology* **9**(1): 3–26

Despres L A 1975 *Ethnicity and Resource Competition in Plural Societies*. Aldine, Chicago

Eriksen H 1994 Nationalism Mauritian style: Cultural unity and ethnic diversity. *Comparative Studies in Society & History* **36**(3): 549–74

Furnivall J H 1939 *Colonial Policy and Practice. A Comparative Study of Burma and the Netherlands India*. Cambridge University Press, Cambridge, UK

Geertz C (ed.) 1963 The integrative revolution. In: Geertz C (ed.)*Old Societies and New States: The Quest for Modernity in Asia and Africa*. Free Press, Glencoe, IL, pp. 105–57

Inkeles A 1969 Participant citizenship in six developing countries. *American Political Science Review* **57**(4): 1120–41

Kuper L, Smith M G (eds.) 1969 *Pluralism in Africa*. University of California Press, Berkeley, CA

Kymlicka W (ed.) 1995 *The Rights of Minority Cultures*. Oxford University Press, Oxford, UK

Lijphart A 1977 *Democracy in Plural Societies: A Comparative Explanation*. Yale University Press, New Haven, CT

Milne R S 1981 *Political in Ethnically Bipolar States: Guyana, Malaysia, Fiji*. University of British Columbia Press, Vancouver, BC, Canada

Morris H S 1967 Some aspects of the concept plural society. *Man* **2**(2): 169–84

Nagata J 1979 *Malaysian Mosaic: Perspectives from a Polyethnic Society*. University of British Columbia Press, Vancouver, BC, Canada

Peter K 1981 The myth of multiculturalism and other fables. In: Dahlie–Fernando T (ed.) *Ethnicity, Power & Politics in Canada*. Methuen Press, Toronto, Canada, pp. 56–7

Porter J 1965 *The Vertical Mosaic*. University of Toronto Press, Toronto

Porter J 1972 Dilemmas & contradictions of a multi-ethnic society. *Transactions of the Royal Society of Canada* **10**: 193–205

Ratnam K J 1965 *Communalism and the Political Process in Malaya*. University of Malaya Press, Kuala Lumpur, Malaya

Smith M G 1965a *The Plural Society in the British West Indies*. University of California Press, Berkeley, CA

Smith M G 1965b *Stratification in Grenada*. University of California Press, Berkeley, CA

Tambiah S 1989 The politics of ethnicity. *American Ethnologist* **16**(2): 335–49

Taylor C 1992 *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton, NJ

Van den Berghe P L 1970 *Race and Ethnicity: Essay in Comparative Sociology*. Basic Books, New York

Waldron J 1992 Minority cultures and the cosmopolitan alternative. *University of Michigan Journal of Law Reform* **25**: 751–992

Walzer M 1992 Comment. In: Taylor C (ed.) *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton, NJ, pp. 99–104

J. Nagata

Pluralism

The term 'pluralism' was introduced into political science by the Englishman Harold Laski in a number of essays published mostly while teaching at Harvard University between 1916 and 1920. For the term, Laski was indebted to the pragmatist philosophy of William James, whose book, *A Pluralistic Universe* (1909), Laski hailed as of 'vital significance for political theory.' In philosophy, pluralism, as distinguished from 'monism,' denoted a system of thought which recognizes more than one ultimate principle. Although Laski borrowed the word 'pluralism' from American pragmatism, the ideas he advanced under the rubric of pluralism were taken largely from a heterogeneous group of European, and in particular English, legal historians and political theorists whose work Laski encountered while a student at Oxford immediately prior to World War I. Among the most important of these were the German medievalist Otto Von Gierke;

Frederick Maitland, the preeminent legal historian in England; J. Neville Figgis; and Laski's Oxford tutor, Ernest Barker, who in 1914 coined the terms 'polyarchism' and 'the new federalism' to describe this new set of ideas.

The defining project of the English pluralists (as they would come to be known by the time Laski returned to England in 1920) was to vindicate the rights and autonomy of associations by challenging the modern doctrine of unlimited, unitary state sovereignty. Pluralists drew an invidious contrast between the 'federalistic structure of medieval society,' in which a dense network of autonomous associations flourished, and the modern notion of state sovereignty, which, Maitland lamented, reduced 'to impotence and then to nullity all that intervenes between Man and State.' But pluralism was much more than an historical description of the world we have lost; the concept of sovereignty was not just unfortunate, it was untrue to the facts of social life. 'We say,' Figgis explained, 'that the world is *not*, as a fact, composed of a few vast unities known as States, set over against crowds of isolated individuals; but is a society of societies, each and all with rights, liberty and life of their own.' Corporations were not legal fictions created by the state, but rather were natural organizations with their own autonomous existence. The idea of unlimited state sovereignty, Laski argued, was 'a fiction of logic … out of accord with the facts of life'; the 'history of societies' showed repeatedly that organized groups, even unpopular ones like militant female suffragists, striking miners, or conscientious objectors, imposed limits on the allegedly sovereign state, forcing public officials to adopt policies to which they were opposed.

Although pluralists like Laski and Figgis frequently appealed to 'the real facts of human society,' their challenge to what they termed 'the monistic theory of the state' was more normative than empirical, though typically the 'is' and the 'ought' were woven seamlessly together. In arguing that 'the state is only one among many forms of human association,' for instance, Laski meant not only to highlight the empirical fact of the social and political importance of associations and groups, but also, and more important, to advance a radically anti-statist normative vision. For the young Laski, the state was owed no more loyalty than other groups; each group, including the state, deserved only the loyalty it earned 'by virtue of its achievement.' The 'ought,' in Laski's mind, derived from the 'is': It was the 'facts themselves' that 'dissolve … the inherent claim of the State to obedience.' Not all pluralists, however, shared Laski's anarchic and syndicalist sympathies. Others, like Figgis and Barker, for instance, allowed that the state had an important role to play in regulating conflict between associations and thus in creating the political conditions necessary for associations to pursue their own ends. But while Figgis and Laski might disagree on what should replace the monistic state, they both looked to pluralism less to explain the world than to change it. Their main objective was to help the groups they cared most deeply about, which, in the case of Figgis, an Anglican priest, were religious associations, and in the case of Laski, as well as the guild socialist theoretician, G. D. H. Cole, were labor and the trade unions. And herein lies an explanation for English pluralism's precipitous decline in the 1920s. For pluralists failed to persuade others, or even eventually themselves, that pluralism was the most effective means to help the groups they wanted to assist. Laski himself gradually abandoned or watered down his pluralist ideas because, along with the rest of the left, he came to believe, particularly after the Great Crash of 1929, that labor had more to gain from an active, even paternal, state than from a minimal, pluralist state.

Ironically, at the time that English pluralism as a normative vision and intellectual force was in decline, the empirical study of organized groups was picking up steam in American political science, particularly with the publication of Peter Odegard's *Pressure Politics: The Story of the Anti-Saloon League* (1928) and Pendleton Herring's *Group Representation Before Congress* (1929). These sorts of empirical studies of the role of organized groups in the American political process never interested Laski. Indeed, they only confirmed his belief that the trouble with American political science was that 'everywhere the emphasis is on description and not principle,' a diagnosis diametrically opposed to that held by most American political scientists interested in the role of groups during this era. They believed, with Arthur Bentley, that there was too much moralizing about the evils of special interest groups and too little objective empirical description and analysis of group behavior in the political process. Over the next quarter century American political scientists worked diligently to remedy these empirical shortcomings, such that by 1950, a year before the publication of David Truman's landmark book, *The Governmental Process*, (Truman 1971/1951), Merle Fainsod could conclude that 'there are few fields which American political scientists have cultivated as intensively and profitably as the analysis of the dynamic interplay of interest groups as they give shape to public policy at almost all governmental levels.' (cited in Crick 1959)

In 1950 the discipline was relatively rich in empirical studies of interest group influence but the term 'pluralism' was rarely, if ever, used, though political scientists were beginning to speak of the 'group basis of politics' or the 'group theory of politics.' Truman as well as Bertram Gross had helped to revive interest in Bentley who, in his 1908 book, *The Process of Government* (Bentley 1967/1908), had argued that groups, not laws, institutions, or constitutions, were the essential 'raw materials' of government. Truman followed Bentley in trying to understand the government process by treating groups, specifically interest

groups, as 'a primary unit of analysis.' Over the next decade this increasingly influential mode of analysis would become known by advocates and even more frequently by detractors as 'analytical pluralism' or simply 'pluralism.'

Critics were quick to charge pluralists with believing that groups were all that mattered, often pointing to Bentley's audacious remark that 'when the groups are adequately stated, everything is stated.' But most of those tagged as pluralists, including most especially Truman, never held to anything so implausible. Pluralists looked to groups (in Truman's words) as 'a major explanatory variable,' but explicitly rejected the notion that government officials merely passively registered or mirrored the sum total of organized group preferences. The relative importance of groups and government officials in the shaping of policy, Truman emphasized, was an empirical question, not one to be settled by 'preassigning weights and roles.' Analytic pluralists, to be sure, generally avoided using the term 'the state,' but this was not because their explanations were 'society-centered' or because they believed government officials lacked autonomy or importance vis-à-vis societal groups, but rather because they doubted, for epistemological as well as empirical reasons, that the myriad institutions and officials of government, particularly in the United States, were coherent enough to warrant a locution that treated the government as a unitary actor with a single will (Ellis 1992). The same doubts fed pluralists' skepticism of concepts like 'the public interest' or 'the will of the people,' since society no less than the state was a complex mix of rival interests and objectives. In this respect, the modern American pluralists were kindred spirits of the early-twentieth-century English pluralists.

Another strain of criticism of American pluralism objected not to its focus on societal groups but to its portrayal of the distribution of power within the United States as dispersed. Pluralists, it was claimed, slighted the severe inequalities in resources and power in American society, and believed mistakenly that the system was self-correcting: that the mobilization of one group would necessarily lead to the mobilization of an equal and opposite countervailing pressure. Pluralists were also said to believe that all important or legitimate interests and opinions were represented within the political system, or even that all groups had substantially equal access to the policy-making process. As the ideological dust kicked up in the 1960s has slowly settled, scholars (e.g., Smith 1990, Hirst 1987) have begun to recognize that few, if any, of those scholars commonly described as pluralists harbored such naive views of the political process. The important question, in any event, is not whether it is possible to find someone myopic enough to believe that all groups have equal, or nearly equal, power, but whether pluralists are logically committed to these propositions. How widely dispersed power is within any given policy domain or in any given place or time is an empirical question that cannot be settled by methodological or theoretical fiat. Nor can it be answered satisfactorily without a comparative perspective: the same distribution of power may be reasonably characterized as either concentrated or dispersed depending on one's comparative referent or on one's normative benchmark. There is no inherent reason that an empirical commitment to the study of interest groups *à la* Truman need predispose the analyst to find any given distribution of political power. Nor need a normative commitment to pluralism, understood as wide dispersion of political influence and power, necessarily predispose one to find pluralism as a sociopolitical fact in any given setting.

The term 'pluralist' is a particularly confusing label because it is often applied to those who conclude that power in a given locale is dispersed rather than concentrated. Pluralism can be a useful term to describe either an empirical reality—one in which there is, for example, widespread dispersal of power, and bargaining rather than hierarchical decision-making—or a normative commitment to such dispersed power, but it does not help to call those scholars who find dispersed power 'pluralists.' When Robert Dahl (1961) finds a pattern of dispersed influence in contemporary New Haven politics, that no more makes him a pluralist than his discovery of concentrated oligarchic power in early New Haven makes him an elitist.

Dahl remains the archetypal pluralist in the minds of many, even though he is not a group theorist on the model of Bentley and Truman. Whether or not Dahl counts as a pluralist, he is certainly among the keenest student of the dilemmas of pluralism. As a graduate student at Yale in the 1930s, Dahl took a year-long seminar on recent political thought from Francis Coker, who had written the entry on pluralism for the *Encyclopedia of the Social Sciences* (1934). There Dahl encountered Laski and the English pluralists as well as Coker's ambivalent critique, in which he praised pluralists for drawing attention to the importance of autonomous groups intermediate between the individual and the state, while also criticizing pluralists for slighting the need for centralized political control. State control was necessary, Coker argued, not only to referee intergroup conflict but also to protect the 'unorganized interests of the general body of citizens' against the organized. Moreover, without active state intervention pluralist group autonomy would result in stronger interests, particularly employers abrogating the rights of weaker interests, particularly wage-workers. Dahl amplified his teachers' arguments in *Politics, Economics and Welfare* (1953). It was one thing, Dahl and his co-author Charles Lindblom said, to recognize (as did Laski) that government officials often could not 'enforce decisions by unilateral, hierarchical means' but must often bargain with

powerful and autonomous groups to negotiate 'temporary armistices.' However, it was 'quite another to turn this social fact into a prescription of the desirable, and to argue in effect that politicians should not even attempt to exercise "the last say," but should turn that power over to national organizations bargaining among themselves.' The fundamental dilemma of a pluralist democracy, in Dahl's (1982) view, is that autonomous associations, groups, or organizations are highly desirable and yet also capable of doing great harm if not controlled by a central authority.

Why is pluralism desirable? Advocates of pluralism see autonomous organizations fulfilling a variety of beneficial purposes. Some emphasize pluralism's role in facilitating individual development and self-expression, while others emphasize the importance of groups in integrating the individual into society. Associations are also seen as seedbeds of civic virtue, an enabling environment in which citizens learn the habits of self-government. A different emphasis is evident among those who stress pluralism's defensive political function, as an organizational counterweight to the coercive powers of the central government. Pluralism's vital contributions to democratic governance has long been Dahl's primary focus, a concern that has renewed resonance as scholars, politicians, and citizens attempt to understand and direct the process of democratization in the former communist countries in Eastern Europe and the Soviet Union. In the absence of vibrant organizational pluralism, the experiences of these nations suggest, it is difficult for formal democratic institutions such as elections, parliaments, and presidents to curb hierarchy and domination.

Pluralism may be essential to the functioning of large-scale democracies but that does not mean, as Dahl himself emphasizes, that pluralism is without disadvantages. To begin with, group leaders may dominate or oppress group members, particularly where the costs of exit from a group are high. Less ominously, the group's elites may fail to represent the opinions of its membership—the elites may be more conservative or more radical, more compromising or more confrontational, than the membership. The English pluralists at times obscured these problems by treating organizations, whether churches or trade unions, as if they were undifferentiated organic wholes with 'real' personalities (even while they denied such a real personality to the state), but American analytic pluralism was rarely, if ever, guilty of reifying groups. Far from treating large groups as if they were homogenous, as is sometimes alleged (Kariel 1968), Bentley and other American group theorists were keenly aware, in Odegard's words, that all groups 'are themselves almost infinitely divisible into subgroups whose interests are not always congruent or compatible with one another.'

A second problem with organizational pluralism is that all groups are not created equal: organized interests, all other things being equal, have an advantage over unorganized interests, and some interests are easier to organize than others. Autonomy for the lambs may well mean a quick dinner for the wolves. It is the weaker groups, as Schattschneider (1960) argued, who generally have the greatest interest in 'enlarging the scope of conflict' by appealing to public authority and thereby modifying the imbalance in private power relations. Recognition that group autonomy may serve the interests of the most powerful groups is a potent argument against the radically antistatist vision of the English pluralists, but not nearly so troubling to the pluralist vision of post-war American political scientists, most of whom were New Deal Democrats. The problems of group autonomy are real, but they need to be counterbalanced with an appreciation for the equally real limits of central control. Absent the articulation of group preferences and mobilization of group interests, a central political authority would define who was a group and what were legitimate interests. Without group pressures, the state would find it difficult if not impossible to gauge the intensity with which preferences are held. Group autonomy without central control may be bad, but it is no worse than central control without group autonomy.

Other critics worry that, even if all groups were equal, pluralism may, in Dahl's words, 'deform civic consciousness' by fostering particularistic or short-run demands at the expense of a broader common good or long-run social needs. Whether civic consciousness is higher or the common good is more closely attended to in those countries or policy domains where interest groups are less prevalent or less assertive is an empirical question, the answer to which is far from self-evident. A related concern is that private associations may be given essentially public functions, thereby removing or reducing democratic control over governmental decisions. One manifestation of this problem is European-style corporatism, in which bargains are worked out between peak associations representing labor, business, and farmers. In the United States, the problem of alienating democratic control may take the form of powerful groups capitalizing on the fragmentation of the system by 'capturing' governmental officials in a specialized policy domain (McConnell 1966, Lowi 1969). Dispersed power is not inconsistent with groups monopolizing power in specialized segments of the political system. Again, the extent to which this describes accurately the policy-making in any given system or policy domain is an empirical question.

Among the most common criticisms of pluralism is that it was little more than a self-satisfied apologia for American or, alternatively, Western, democracy. Whether or not those who were associated with pluralist ideas were satisfied with the state of democracy in America or Western Europe, there is no reason to think pluralism as a normative vision is inherently

conservative. Hirst (1989), for instance, uses pluralist ideas to launch a blistering critique of the British political system, particularly Parliament's claims to unlimited sovereignty. Measuring Western democracies against an exacting yardstick of pluralist principles, Hirst (1997) finds severely wanting not only the West's narrowly plebiscitarian vision of democracy, in which 'the people periodically choose who are to be their elected masters,' but also the over-centralization of the provision and delivery of public services. Taking pluralism seriously, Hirst maintains, means reshaping governance radically, not only in the nation state but at the level of international institutions like the European Community. Dahl himself draws upon pluralist ideas to argue for a dramatic devolution of power to cooperatively owned economic enterprises. And in the context of the non-Western world, even the tamest pluralist ideas can constitute a bracing challenge to entrenched state power and economic privilege. Writing in the 1968 *International Encyclopedia of the Social Sciences*, Henry Kariel wrote an obituary for political pluralism, which, he said, 'as an ideology … only lingers quietly as a submerged, inarticulate ingredient of Western liberalism' (Kariel 1968). Over three decades later, it would appear that reports of pluralism's decline, let alone death, have been greatly exaggerated. Particularly as a normative vision, as an antidote to plebiscitary democracy or to centralized state control, political pluralism has shown renewed appeal and relevance across the globe.

See also: Civic Culture; Cleavages: Political; Democracy; Democratic Theory; Interest Groups; Interest Groups, History of; Lobbying; Participation: Political; Party Systems; Political Parties; Political Representation; Sovereignty: Political; Voluntary Associations, Sociology of; Voluntary Organizations

Bibliography

Baumgartner F R, Leech B L 1998 *Basic Interests: The Importance of Groups in Politics and in Political Science*. Princeton University Press, Princeton, NJ

Bentley A F 1967 [1908] *The Process of Government*. Belknap Press, Cambridge, MA

Coker F 1934 Pluralism. In: Seligman E (ed.) *Encyclopedia of the Social Sciences*. Macmillan, New York

Crick B 1959 *The American Science of Politics*. University of California Press, Berkely, CA

Dahl R, Lindblom C 1953 *Politics, Economics and Welfare*. Harper, New York

Dahl R A 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT

Dahl R A 1982 *Dilemmas of Pluralist Democracy: Autonomy vs. Control*. Yale University Press, New Haven, CT

Ellis R J 1992 Pluralist political science and 'the state': Distinguishing between autonomy and coherence. *Polity* **24**: 569–89

Gunnell J G 1995 *The Declination of the 'State' and the Origins of American Pluralism*. In: Farr J, Dryzek J S, Leonard S T (eds.) *Political Science in History: Research Programs and Political Traditions*. Cambridge University Press, Cambridge, UK

Herring P 1929 *Group Representation Before Congress*. Johns Hopkins Press, Baltimore, MD

Hirst P Q 1987 Retrieving pluralism. In: Outhwaite M, Mulkay M (eds.) *Social Theory and Social Criticism*. Basil Blackwell, Oxford, UK

Hirst P Q (ed.) 1989 *The Pluralist Theory of the State: Selected Writings of Cole G D H, Figgis J N, and Laski H J*. Routledge, London

Hirst P Q 1997 *From Statism to Pluralism: Democracy, Civil Society and Global Politics*. UCL Press, London

James W 1909 *A Pluralistic Universe*. Longmans, Green, New York

Kariel H S 1961 *The Decline of American Pluralism*. Stanford University Press, Stanford, CA

Kariel H S 1968 *Pluralism*. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York

Laski H J 1917 *Studies in the Problem of Sovereignty*. Yale University Press, New Haven, CT

Laski H J 1919 *Authority is the Modern State*. Yale University Press, New Haven, CT

Laski H J 1921 *The Foundations of Sovereignty and Other Essays*. Harcourt, Brace and Company, New York

Lowi T J 1969 *The End of Liberalism*. W. W. Norton, New York

McConnell G 1966 *Private Power and American Democracy*. Knopf, New York

Nicholls D 1975 *The Pluralist State*. St. Martin's Press, New York

Odegard P 1928 *Pressure Politics: The Story of the Anti-saloon League*. Columbia University Press, New York

Polsby N W 1980 *Community Power and Political Theory*, 2nd edn. Yale University Press, New Haven, CT

Runciman D 1997 *Pluralism and the Personality of the State*. Cambridge University Press, Cambridge, UK

Schattschneider E E 1960 *The Semisovereign People*. Holt, Rinehart and Winston, New York

Smith M J 1990 Pluralism, reformed pluralism and neo-pluralism: The role of pressure groups in policy-making. *Political Studies* **38**: 302–22

Solomon S G 1983 'Pluralism' in political science: The odyssey of a concept. In: Solomon S G (ed.) *Pluralism in the Soviet Union*. St. Martin's Press, New York

Truman D 1971 [1951] *The Governmental Process: Political Interests and Public Opinion*, 2nd edn. Knopf, New York

R. J. Ellis

Pluralism and Toleration

Pluralism is a characteristic of the contemporary lifeworld that has developed in the course of modernity into the dominant basis for social freedom and political justice. Today, the features as well as the relevance of this basis have not, gone uncontested (Schultz 1974, Kelso 1978). We generally understand pluralism not only as a diversity, but rather as a right to an equal coexistence and confrontation of natural or social elements which lack a governing principle of unity. Tolerance is closely associated with the notion of pluralism insofar as it

refers to those individual virtues which derive from the recognition of this right to an equal coexistence of a multitude of natural and social elements. Pluralism and tolerance consider the same phenomenon from different perspectives. Whereas pluralism emphasizes the diversity of equally entitled, free groups, tolerance stresses the 'live and let live' of others' liberty.

1. *The Idea of Pluralism*

Pluralism can be understood in a metaphysical sense as designating those philosophies which allow for a diversity of entities in their explanations of the world. The pragmatist philosopher William James maintained the existence of an infinite number of limited facts, connections, and systems which are in constant flux; which always lead to new relationships and structures; and for the cognition of which, fixed categories and criteria do not exist, only perspective orientation points (James 1909). Pluralism can also refer to the methodological neutrality of the social sciences of an empiristic orientation, to the multiplicity of cultural forms, from which 'all sociology has its starting point' (Scheler 1980, p. 25). Today, the term is primarily understood in a social and political sense in which empirical and normative elements converge.

1.1 *Empirical and Normative Elements*

Empirically speaking, pluralism denotes the diversity not only of confession and religion (religious pluralism), but also of social groups together with their lifestyles (social pluralism), of basic values (pluralism of value), and of the powers which determine politics (political pluralism). For this many-dimensional diversity, our life-world has primarily benefited from the religious wars of the Reformation, the European Enlightenment, and the Bourgeois and Industrial Revolutions. Through those factors, a process of differentiation was set into motion which dissolved both the relatively homogenous and stable relationships of the so-called Old-European societies and the connection between church and state. Moreover, this process brought forth open societies, the numerous and constantly expanding groups of which had their own interests, convictions, and manners of conduct. More recently, pluralism has gained a new momentum through the process of globalization (Höffe 1999), in particular also through migration movements, many refugees, and those looking for asylum. Second, pluralism purports—and here one finds its normative content—that diversity and difference is acknowledged and approved of; all functionally different groups are equally entitled to develop themselves in a free and unconstrained manner.

The notion of pluralism is of particular relevance in modern democracy theory (Fraenkel 1979, Oberreuter 1980). It is considered here as a critical notion, namely as that central mediative category between the individual and the state, with which the liberal atomization of individuals just as the omnipotence and an all-encompassing capacity of the state are rejected. In a democracy, so runs the empirical thesis of pluralism; many-faceted powers are at work, which, on the basis of a fundamental consensus over institutionalized arbitration with and against one another, struggle for influence and importance. As far as the normative content of pluralism is concerned, various social forces, notwithstanding their differences in content and function, have the same right to engage themselves in the political process by the means of, e.g., voting, of contributions to public deliberation, or by addressing themselves to members of parliament. These forces include trade groups, unions, and employers' associations; scientific organizations, churches, and political parties; and even civil action groups. Politics is not derived from the conception of a common good which is clearly recognizable, prior to all political confrontations and which does justice to all parties concerned. Political decisions are a result of the constant dispute between competing interest groups which have to mutually monitor one another and, in constantly new coalitions, 'scuffle' together into a compromise which might be accepted by a majority.

1.2 *The Critique of Pluralism*

The criticism of the idea of pluralism is leveled partly at its empirical thesis and partly at its normative content; the normative criticism pertains to either pluralism of value or political pluralism.

Pluralism of value becomes visible in the form of the ideologically neutral state. It is reproached with relativism and even nihilism, because it does not even acknowledge certain very general values as binding and is held responsible, therefore, for the crises of meaning and orientation which threaten contemporary industrialized societies. Pushed to the fore, the ideologically neutral state will abolish itself, because it cannot legitimize the framework in the limits of which various powers are capable of freely expressing themselves.

The most significant criticism of *political pluralism* stems from Carl Schmitt and his followers, who fear the centrifugal, 'anarchistic' tendencies of pluralism. Through the openness and insecurity of pluralistic democracy, the state's power of decision making is left crippled; moreover, the compromises resulting from such a situation correspond to neither moral nor economical ideals of sound politics.

Supporters of the critical theory hold pluralism for desirable, but contest the existence of the empirical presuppositions for its realization. They point to the socially underprivileged, who are either effectively shut out of participating in the bargaining process of weighing competing interests, or are, at least, clearly at a disadvantage when up against more powerful

organizations. The normative idea of an equal right for all groups is, therefore, rather a postulate than social reality. Other critics maintain a 'standardization of subjects,' who are no longer capable of an autonomous definition of their interests and, hence, of a genuine pluralistic differentiation. Where a free concert of interest groups once existed, one now finds bureaucratic apparatuses in their place, the internal dynamics of which dominates the political scene.

1.3 Pluralism and Liberal Democracy

Political pluralism presupposes two conditions as given. On the one hand, all interests and the groups which represent them should have equal rights in public deliberation and in the process of political will-formation; on the other hand, the common good should realize itself through the struggle of interests between groups of equal rights. Both conditions are fulfilled, even in developed democracies, at best in approximation. For in pluralistic society, citizens with their interests and convictions have no direct access to politics; their access stands in need of mediation through clubs, organizations, and political parties. In this manner, interests and convictions do indeed gain political power, however at different levels of intensity. The mediation of interests and convictions very often leads to perceptible distortions of their influence.

Due to varying historical or social conditions, many interests and convictions are represented more powerfully than others. The explanation of this fact might be seen in the tighter organization of certain associations, to the better support by more potent money lenders, or to the fact that such groups have attractive media at their disposal, or better relations to the mass media. An equality of rights in modern times is not realized through the abolishment of legal privileges, characterizing hierarchical feudal or class societies. Pluralism does not simply describe our social and political reality. It is to a significant degree really a postulate, according to which societies need to be changed and modified.

A modification within pluralism, however, is possible only to a limited extent: first of all, because powerful groups do not freely relinquish their predominance; second, as can be observed sociologically (Olson 1965), many interests are much easier to organize than others, which may lead to the superiority of certain economic groups (employers as much as employees). Furthermore, mergings and the oligarchical tendencies within organizations endanger civic freedom, the essence of pluralism. Dependence on the government gives way to dependence on both the bureaucracy of associations and its functionaries. Finally, even a fair arrangement between a plurality of interest groups does not simply lead to the realization of the common good; for, as a rule, long-term interests get the worst of short-term interests, just as general interests get the worst of particular interests. It is for this reason that the protection of the environment is so easy to push aside.

In taking an overview of these arguments, one must neither adhere to a conspiracy theory, nor to an extreme criticism of capitalism, in order to preserve a residue of skepticism with regard to the affirmation of an empirically existing pluralism. Hence, the plurality of intermediary groups is without a doubt a pillar of freedom; a pluralistic society is invariably better than a nonpluralistic one; and pluralism itself an important structural element of the liberal-constitutional democracy. Nevertheless, pluralism alone cannot warrant its own significance; it is only a helpful not in any way sufficient organizational form of freedom and equality. One should not easily identify pluralism and liberal democracy because the notion of pluralism brings to light only diversity and competition, and not the complementary elements of commonality and cooperation.

A liberal democracy consists not only of the coexistence and confrontation of a diversity, but also of the regulation of coexistence and confrontation. Regulation brings about community: the mutuality of problems to be solved and, in particular, that fundamental (constitutional) consensus which defines the normative content and the rules of pluralism. In this manner, pluralism shows itself to be a one-sided, nondialectical notion that stresses plurality instead of unity, competition instead of cooperation.

1.4 A Relative Pluralism

Because empirical and normative elements converge in the notion of pluralism, the normative valuation has already begun in the empirical thesis. It is continued with a *pragmatic* consideration: the decentralizing of political power, the institutional and social separation of the process of will formation, just as the diversity of convictions and manners of conduct increase the chance of an efficient coexistence, in which superfluous frustrations and conflicts decrease.

The comprehensive *moral* evaluation begins with an ethical–anthropological insight. Because of its lack of instinct and a wanting specialization of organs, mankind is not fixed to definite manners of conduct. Due to its capacity for language and reason, mankind is much more reflective in relation to its doings and omissions; this not only enables it to act freely and responsibly, but also constitutes a particular challenge. Therefore, mankind, be it as individual, a small or a large group, has the task and the right to develop its own convictions and practices, and to determine and pursue a particular conception of a good life. Different convictions and practices are to be expected, on account of the historical factors under which this process of personal as well as social self-determination takes place. The form as well as the results of self-determination are limited by different perspectives, and a single individual cannot lay claim to the

unconditional 'truth' for himself (Rawls 1993, p. 36). In the pluralism of interests, of values, and of life-plans, the wealth of human self-realization is expressed just as well as the limitation of every concrete design.

On the other hand, different groups are already related to each other through the social division of labor; but the division of labor represents only one of the many aspects which make mediation necessary. Within one's own group, the acknowledgment of interests and convictions is commonplace. However, with regard to other groups, they stand in both competition and conflict with other groups. Because the solution of conflicts is not secured by instincts, there exists the danger of collective self-destruction of mankind. In order to check this danger and, at the same time, to avoid a crude sort of social Darwinism which makes the weak even weaker and the vital and cunning even stronger, general obligations are necessary, which guarantee a humane and just management of conflicts. For that reason, not an absolute and merely a relative pluralism that develops in the context of fundamental commonalities is really desirable.

Pluralism of this sort can encounter the suspicion of being relativistic and nihilistic. Even if modern societies are skeptical of a timelessly valid code of values, even if they do not confer infallibility upon a particular personal, political, religious, or intellectual instance, they nonetheless need at least a few generally valid obligations removed from the opposition of particular designs of faith and existence. Therefore, human rights is accrued a particular meaning: as the personal right to liberty, the political right to participation, and as social and cultural rights, they are the condition and expression of the mutual acknowledgment of fundamentally equal persons. Therefore, human rights refer to those claims which every individual may legitimately raise: they belong to a sphere beyond a legitimate pluralism and which every legal order, every state, especially pluralistic democracy owes to each individual. Through human rights even those minorities which come off badly in the pluralistic struggle of competition are protected (Rawls 1971, Höffe 1995).

2. *The Idea of Tolerance*

Tolerance is realized through the recognition of those groups which are differentiated by interests, beliefs, and lifestyles. In this manner, the idea of pluralism is, through its normative content, aligned from the outset with tolerance; and tolerance, like pluralism, has its basis of legitimation in the idea of justice and in the modern principle of liberty.

2.1 *Its Origin in Religious Tolerance*

Like the general principle of liberty, the particular principle of tolerance found its principal acknowledgment only after long intellectual and sociopolitical debate. In this context, religious tolerance plays a particular role (Püttner 1977).

Because polytheistic religions make a lesser claim to truth than monotheistic ones, they do not face many difficulties as far as the toleration of diverse religious beliefs is concerned. Pre-Christian Rome allows all conquered peoples to practice and even propagate their own cults. Yet the Christians were excluded from the rule of tolerance, but not for religious reasons; from their refusal to observe the transmitted gods, one fears a backlash within the entire commonwealth.

As long as the Christians are persecuted, they appeal to the Roman Emperors, referring to the rights which all mankind are due: that freedom of religion is granted internally and externally. But after Christianity is promoted to the status of a national religion, it soon forgets this demand and its underlying principles contained in the New Testament: even the principle of reciprocity (*Matt*. 7,12; also *Tob*. 4,16); the sermon on the mountain; the parable of the weed under the wheat (*Matt*. 13,24–30, 36–43); Jesus' conduct in general of welcoming, but not of forcing his imitation; moreover his magnanimous attitude toward sinners; finally, Paul's conception of freedom of religion as free from enforcement and his admonition to mutual tolerance (*Col*. 3,12f.; *Cor*. 1, 4,12; *Cor*. 2, 11,1; *Gal*. 5,1.13).

Whereas the *Edict of Milan* (313) leaves religious belief to its practitioners and, moreover, differentiates between state and religion and does not admit mortal or corporeal punishment for religious offences, the swift rise of schisms and heresies fills the bishops with such alarm, even in the patristic times, that they begin to bind, as in the Old Convenant, religion closely with political affairs. Especially the conversion to a foreign cult is punished to the same extent as capital crimes and the disruption of civil order.

Granted, medieval Christendom should not to be considered intolerant through and through (Patschovsky 1998). Because it regards the transmittance of faith as a matter of free conscience, the conversion of pagans by force represents a clear exception. Jews are generally and, often, truly tolerated; and for a long time the Christians live in mutual tolerance with the Muslims. But, the heretic in Christendom, accused of both arrogance and impertinence, was considered, like the Islamic apostate, as it were, an outlaw in the eyes of the law. In the eleventh and twelfth centuries, capital punishment extends to the heretic and is maintained until well after the Reformation. The foundation of medieval intolerance, the all too intimate fusion of *Sacerdotium* and *Imperium*, of spiritual and secular matters are all well preserved in Protestant countries, more likely being advantageous to the state than to the church.

In the 'confessional age' (the age of the Reformation), the idea of (religious) tolerance has to be re-established; and it is effected only after hefty intellectual debate and painful political experience. In

order to encourage the victory of tolerance in the form of the religious neutrality of the state, the mutual autonomy of the state and the church, a few fundamental insights first have to be (re)attained. At first the pragmatic consideration of the 'greater benefit,' the fact that in most cases mildness is better suited than force in order to gain back 'lost souls.' Then one also has to reach the insight that the New Testament is based on tolerance and love, moreover, that not only the conversion but also the retention of faith is both an act of freedom and of mercy, and for that reason, cannot be enforced; furthermore, that the church is a free communion, whose task is to be fulfilled in the spiritual, not in the secular realm, and that for this task every possible constraint is useless, as it would at best lead to a feigned and not to a real acknowledgment of faith.

The way for this new form of tolerance is prepared by Christian humanists like Dante, Marsilius of Padua, Nicolaus von Kues, and Sebastian Castellio, who reject all acts of violence against religious dissidents. However, they are still fixed to ideas of religious unity within Christendom, and, therefore, they search for a peaceful reconciliation by way of religious forum and dogmatic compromise between rivaling confessions. On the contrary, mystic spiritualists as well as some sects engage themselves directly for the religious liberty of each individual; they do this by appealing to the Holy Ghost, who forbids every form of slavery in the realm of the religious, including that of the institutional church.

Whereas no lasting success is allotted to either of these thoughts (peaceful reconciliation or religious individualism), the second half of the sixteenth century sees the problem shift from the religious to the 'secularized' level. Theological arguments are succeeded by considerations of politics and natural right, occasionally even of economics; unsuccessful reconciliatory talks give way to efforts of toleration, in which the denominational heterodox gains ground.

Starting in France, where the preservation of the nation is considered more important than the re-establishment of religious unity, the idea of civil tolerance, which unambiguously differentiates between the aims of the state and the aims of religion, is established against the rule of 'one confession, one law, one king.' With the idea of the confessional, later the religious neutrality of the state, the bloody experiences of confessional civil wars which, at the European level, ended in a stalemate, are finally digested. Consequentially, the relative autonomy of the state and church/religion and, with that, of confessional pluralism, is acknowledged as a necessary condition for political, even physical survival. Nevertheless, the numerous edicts of tolerance, which follow on the Peace of Augsburg (1555) and the Edict of Nantes (1598), are issued only reluctantly; Brandenburg-Prussia represents an exception for a long period. In auspicious circumstances, as in France (1685), the edicts are repealed or, as in Great Britain, where the Catholics remain second-class citizens until the nineteenth century, they are not valid for everyone from the outset.

Because intolerance endangers the free development of business and trade, Europe and North America also have economic reasons to thank for the following two phenomena: that religious liberty does indeed assert itself in the end, and that the remnants of the medieval connection between the state and the church are vanquished. The European Enlightenment, however, plays a particular role in the ultimate victory of the principle of religious tolerance. Inspired by the idea of universal human reason and liberty, such figures as Grotius and Spinoza, Milton, Locke and Voltaire, Lessing (*Nathan the Sage*), later also Schiller (*Don Carlos*) turn to the remnants of dogmatic and political intolerance within the high church and to their corresponding, amalgamated states; moreover, they lay claim to religious liberty as a fundamental personal right to liberty. But, even those propagating the Enlightenment as a rule exclude atheists and nonbelievers from the idea of tolerance, granted, as with Locke, not out of religious but rather out of political reasons; for, whoever does not believe in God cannot take a legally admissible oath.

2.2 Passive and Active Tolerance

The idea of tolerance kindled itself on the freedom of religion in Western Christianity. In our advanced, secularized world this idea no longer represents a special problem with regard to the social interaction of diverse confessions and religions. Tolerance is often enough primarily called for in relation to beliefs which are socially, politically, and culturally differentiated. Thereby, a weaker, more passive form of tolerance, the mere toleration of otherness, is to be differentiated from a stronger, more active, and creative form, the openness to differences, yes, even to their candid recognition.

As pluralistic democracy cannot arise nor continue to subsist without the mutual toleration of otherness, tolerance is, at least in the weaker form, a fundamental principle of modern states (for developments see Lecler 1955, Lutz 1977, Becker 1996). However, tolerance is not only necessary under the condition and in favor of a political pluralism. If not always immediately and distinctly perceptible, there are nonetheless differences among mankind: in need, talent, and interest; in taste and in sociohistorical origin; moreover, in the perception and interpretation of the personal or social situation in which one lives. Nor is one free of faults, prejudices, or mistakes. Therefore, at least passive tolerance is held to be a universally valid posture with regard to the civilized intercourse of mankind; it is through just such relations that the coexistence in marriage and family, in industry, in

society, in the government, even between states is made bearable as well as efficient. So, even an enlightened self-interest speaks in favor of tolerance.

However, the core dimension of tolerance is, with that aspect, not yet achieved. Its stronger form, the candid recognition of the other and otherness, is grounded in the dignity and the liberty of every human being. Therefore, tolerance consists not only of that forbearance of the peculiarities and weaknesses of one's fellow man which can be solicited from them and freely granted to them. As the coexistence of individually responsible persons with equal rights is not possible without reciprocal recognition, tolerance is a fundamental condition for the justice of mankind's coexistence in freedom; and, as a condition of justice, it is a claim that can be refused to no one.

For active tolerance, not merely the practical (moral-political) principle of liberty but also a theoretical interest in truth be adduced. Human perception being limited, the open debate of different opinions offers a better chance for establishing truth than the 'intolerant,' namely, dogmatic and of a conviction considered infallible (Mill 1989).

3. Tolerance as Both a Virtue and a Principle of the State

Religious liberty is only one example that brings to the fore the specific intellectual and political difficulties which have to be overcome in order for the principle of tolerance to become historically significant. This principle, which goes beyond religious liberty, takes primarily two different forms: tolerance as a personal attitude and as a political principle of the state and of civilization, both forms mutually supplementing and yielding one another.

Whoever affirms the self-determination (the right to life, the will to development, and the liberty) of the dissentient is realizing the attitude of tolerance. Only when one has one's own convictions, can one speak of recognizing an 'other'; and only there, where one doesn't indulge in indifference, or in the type of cynical nihilism which plainly lets everything pass as valid, including even gross injustice, can one speak of this particular attitude. Tolerance does not mean the fig leaf, behind which moral indifference and intellectual weaknesses conceal themselves. It is a critical autonomy which does not even allow certain conflicts to arise in coexistence, especially not in situations of competition; and which does not go through the inevitable debates and quarrels with the inexorable severity of a religious war. Whoever is tolerant no longer pursues a life which is founded upon self-assertion through violent conversion, or through the surmounting of adversaries. He or she endeavors to establish a communality on the basis of civil equality and communication to which the following qualities belong: the ability to listen, the capacity to take others seriously, and finally the willingness to let oneself be taught by new situations and new information. Tolerance is then accomplished by the willingness and ability to sympathize with the views and lifestyles of others.

Like pluralism, tolerance too is grounded in the following factors: the understanding that no one is entirely free from faults; the knowledge of the wealth as well as the perspectival limitations of every concrete instance of self-realization; and the insight into the limits of human knowledge. Tolerance, however, originates primarily in the recognition of the other as a free and equal person, who, on the basis of the inviolable dignity of mankind, has the right to form his or her own opinions, and to live according to them alone, or with like-minded individuals—provided that the exercise of this right does not encroach upon the same right of all others. The virtue of tolerance is one of the most important political virtues of a democratic citizen (Rawls 1993, p. 194).

As tolerance does not represent an end in itself, but rather a condition and an expression of the liberty and dignity of mankind, it terminates at the point where liberty and dignity are violated. Tolerance does not consist, as Marcuse (1965) fears, of the stabilization of existing social states, in the preservation of the status quo, and even less of the (masochistic) acceptance of injustice. Wherever the liberty and dignity of mankind needs to be protected, tolerance crosses over into accusation, criticism, and protest. The obligation to impose on no one a limitation of his or her liberty which is not imposed on all others in the same manner can be considered a criterion for the limits of tolerance. The highest principle of justice consists of this criterion of an equal liberty which is primarily spelled out in terms of human rights.

Because the justification of tolerance is found in liberty and justice, it cannot be left to the discretion of individuals and political powers. Its recognition is not only a voluntarily assumed attitude, but also a duty of justice, the implementation of which the legal order, the state, is responsible. It is therefore a political task to institutionally guarantee basic elements of tolerance, such as human rights. The first task of tolerance as a principle of the legal order and the state is therefore found in the positivation of human rights and their recognition as basic rights (personal right to liberty, political right to participation, and social and cultural rights), or, as the case may be, as fundamental aims of the state (liberty, democracy, welfare state).

As long as human rights are only recognized by coercion, tolerance remains nothing more than an acceptance accompanied by contemption toward those who think and live differently. In order to guarantee the stronger form of tolerance—coexistence based on mutual respect—tolerance as a duty of justice has to be acknowledged on its own terms; moreover, it must become through continuous practice a component of one's personal character, a genuine attitude.

Tolerance as a political principle is only accomplished when this duty of justice is freely accepted as a moral virtue of the citizen.

At the international level tolerance demands the mutual respect of different cultures and traditions, instead of valuing European culture or American lifestyle as the ideal of the development of all civilization and to devalue all other forms as primitive and barbarian. This type of respect was articulated first in Herder's (1774/1994) and later in Hegel's (1821/1970) theory of the 'spirits of nations' (*Volksgeister*), according to which different social and cultural patterns in which different peoples think and live assert themselves in the course of history. As these forms are coherent, they justify themselves to this extent.

One facilitates the recognition of other cultures' intrinsic value whenever one looks not only to that which divides but also to that which binds, that is, universal human attitudes such as readiness to aid, civil courage, and honesty. These commonalities are not only acknowledged as morally binding in the Christian West, but also in the most diverse of societies and epochs. The tolerance that began with religious liberty is accomplished only at the point, where the colorful abundance of personal and sociohistorical self-realization is freely acknowledged as both a task and a chance for creative humanity.

See also: Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Human Rights in Intercultural Discourse: Cultural Concerns; Liberalism; Pluralism; Relativism, Pluralism, and Skepticism (Philosophical Perspective): Cultural Concerns; Tolerance; Value Pluralism

Bibliography

Becker W 1996 *Anerkennung und Toleranz. Über die politischen Tugenden der Demokratie*. Palm & Enke, Erlangen

Fetscher I 1990 *Toleranz. Von der Unentbehrlichkeit einer kleinen Tugend für die Demokratie*. Radius, Stuttgart

Fraenkel E 1979 *Deutschland und die westlichen Demokratien*, 7th edn. Kohlhammer, Stuttgart

Hegel G W F 1970 *Grundlinien der Philosophie des Rechts (1821)*. Suhrkamp, Frankfurt

Herder J G 1994 *Auch eine Philosophie der Geschichte zur Bildung der Menschheit (1774)*. Deutscher Klassiker Verlag, Frankfurt

Höffe O 1995 *Political Justice*. Polity Press, Cambridge, UK

Höffe O 1999 *Demokratie in Zeiten der Globalisierung*. Beck, Munich

James W 1909 *A Pluralistic Universe*. Longmans, New York

Kelso W A 1978 *American Democratic Theory. Pluralism and its Critics*. Greenwood Press, Westport, CT

Lecler J 1955 *Histoire de la tolérance au siècle de la réforme*, 2 Vols. Aubier, Paris

Horton J, Mendus S (eds.) 1991 *John Locke: A Letter Concerning Toleration (1689)*. Routledge, London

Lutz H (ed.) 1977 *Zur Geschichte der Toleranz und Religionsfreiheit*. Wissenschaftliche Buchgesellschaft, Darmstadt

Marcuse H 1965 Repressive tolerance. In: Wolff R P, Marcuse H, Moore B (eds.) *A Critique of Pure Tolerance*. Beacon Press, Boston, MA

Mendus S (ed.) 1988 *Justifying Toleration. Conceptual and Historical Perspectives*. Cambridge University Press, Cambridge, UK

Mill J S 1989 *On Liberty (1859)*. Cambridge University Press, Cambridge, UK

Oberreuter H (ed.) 1980 *Pluralismus. Grundlegung und Diskussion*. Leske und Budrich, Opladen

Olson M Jr 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA

Patschovsky A (ed.) 1998 *Toleranz im Mittelalter*. J. Thorbecke, Sigmaringen

Püttner G 1977 *Toleranz als Verfassungsprinzip. Prolegomena zu einer rechtlichen Theorie des pluralistischen Staates*. Duncker & Humblot, Berlin

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Rawls J 1993 *Political Liberalism*. Columbia University Press, New York

Scheler M 1980 *Die Wissensformen und die Gesellschaft (1926)*, 3rd edn. Francke, Bern

Schnapper D 1998 *La rélation à l'autre. Au coeur de la pensée sociologique*. Gallimard, Paris

Schultz U (ed.) 1974 *Toleranz. Die Krise der demokratischen Tugend*. Reinbek Rowohlt

Thierrey P 1997 *La tolérance. Société démocratique, opinions, vertus*. Press Universitaires de France, Paris

Walzer M 1997 *On Toleration*. Yale University Press, New Haven, CT

O. Hoffe

Polanyi, Karl (1886–1964)

1. Biographical Notes

Karl Paul Polanyi was born in Vienna on 25 October 1886 to Michael Pollacsek and Cecile Wohl. They had married in 1881 and had six children: Laura, Adolf, Karl, Michael, Zsófia, and Paul. Polanyi's father worked for the Swiss Railways and Vienna's Stadtbahn, before setting up on his own as a railway builder in Hungary. In the 1890s the family moved to Budapest where Karl Polanyi began his secondary studies in the model school of the teachers' training institute in 1898. At the turn of the century Polanyi's fathers business went bankrupt and after indemnifying the creditors, the family was impoverished and he was forced to take a job abroad again. In spite of that, there was lively intellectual activity in the parental home. Polanyi's mother, Cecile Wohl, organized a literary salon attended, among others, by Oszkár Jászi, György Lukács and Karl Mannheim. After finishing secondary school in June 1904, Polanyi enrolled in the Faculty of Law in Budapest University. He lost his maternal grandfather and father in 1905, both of

whom had especially deeply inspired him. After attending courses in Vienna and Budapest he took part in organizing demonstrations for the progressive thinker Gyula Pikler, a professor of the Philosophy of Law in 1907. Polanyi obtained his Doctoral degree on 26 June 1909 from Bódog Somló at the Kolozsvár (Cluj) University. In 1908 he became president of the recently formed Galilei Circle, a student body set up to disseminate the scientific world view in readings and debates, without—as Polanyi later self-critically recalled—being politically involved. He qualified as a lawyer in 1912 and began to co-edit the review *Szabadgondolat* [*Free Thought*] in 1913. His writings argued for free education and were devoid of the influence of the church. In 1914, Polanyi was involved in establishing the Radical Party, of which he became the secretary.

However, the next year he was sent to the front and discharged two years later as disabled. During the Károlyi government of 1918, he was a member of the legal committee of the People's Council in Budapest and of the Students' Council. However, on the eve of the Hungarian Soviet Republic in 1919 he delivered a speech against the dictatorship of the proletariat, then immigrated to Vienna. He soon underwent a serious operation from which he took a long time to recover. He was working on a voluminous unpublished work, Behemoth, concerned with the destructive capacity of science. In this he says that the future is constantly being remade, therefore it can't give validity to our actions in the present.

In addition to Ludwig von Mises, Jakob Marschak, Felix Weil, and others, he contributed to the theoretical debate about the possibility of the planned economy in which he elaborated a proposal to solve the problem of 'Sozialistische Rechnungslegung' on the basis of the principles of guild socialism. He met a well-known figure of the revolutionary students' movement, Ilona Duczynska, in 1920. They married and had a child, Karoline Helene (Kari) in 1923. In Vienna, Polanyi first worked for Hungarian papers (Bécsi Magyar Ujság and Bécsi Kurir), and then from 1924 to 1933 he contributed to *Der Österreichische Volkswirt*. Having lost his job because of the rising danger of National Socialism, he immigrated to England where he first lived in privation, earning money by giving lectures. He was one of the masterminds behind the Christian Left movement, and concentrated on criticism of the essence of fascism.

This time he joined the Workers' Educational Association and the Extra-Mural Delegacies of Oxford and London, lecturing mainly in general economic history and social theory. His lectures focussed on the shift from *laissez-faire* to regulated capitalism, when state and industry began to interpenetrate, and interference of government with the captains of industry became regular.

In 1940 he became a British citizen, but from that year, he was resident lecturer at Bennington College in the USA and toured America with lectures. He also worked on *The Great Transformation* first published in 1944 (Polanyi [1944] 1957). After returning to the UK during the war this book earned him the post of visiting professor at Columbia University, New York, in 1947. He lectured in economic history and coordinated research into the origin of economic institutions. These efforts were summed up in the posthumously published *The Livelihood of Man* (Polanyi 1977). He continued to do research with a team after his retirement in 1953, publishing their findings in *Trade and Market in the Early Empires* (Polanyi et al. 1957). The posthumously published *Dahomey and the Slave Trade* (Polanyi 1966) was another result of the cooperation.

The family settled in Canada as US authorities had denied an entrance visa to Polanyi's wife. In 1963 they visited Hungary, where Polanyi gave a lecture about the American Economic Sociology. He died of a stroke during the night of April 23, 1964 in Toronto, Canada.

2. *The Substantive View*

The major question of Polanyi's substantive view is how to identify the place of economies in societies (Polanyi 1947, 1977). He argues that economic activity is embedded in the social and cultural setting and that the predominance of price-making markets does not apply to most primitive, archaic, or even contemporary economies. He therefore taught that the changing place of economies could be best described by combinations of comparative concepts that shed light on the alternative logic of economic reproduction. Polanyi's substantive theory was discernible in its essentials in *The Great Transformation* (Polanyi [1944] 1957) and in the lectures he wrote in preparing it. The book, a critique of the market system, is a derivative of historical knowledge and skills accumulated in political journalism. This is borne out by a later remark of his saying he had two aims with the work he wished to: transfer the more combatant Austrian workers' culture into the English labor movement; and give an up-to-date criticism of capitalism, providing intellectual support to Roosevelt's New Deal. In Polanyi's view, the imminent problem of the market system was not the existence of markets but the turning land and labor into commodities by this system. The analysis of the so-called Speenhamland law, temporarily hindering the creation of a competitive labor market as well as of the ensuing welfare acts, revealed that welfare legislation failed to efficiently remedy the mass impoverishment paradoxically entailed by growth and could only delay the emergence of a labor market from masses of wage workers.

The book was written during World War II, parallel with Hayek's *Road to Serfdom* (Hayek [1944] 1976) and after Burnham's *The Managerial Revolution*

(Burnham [1941] 1960). While Hayek argued that deserting the road of self-regulation leads to fatal social deformation, Burnham envisaged the inevitable predominance of managerial-elitist hierarchies with the growing autonomy of control functions. Polanyi focussed on the double movement of *laissez-faire* economy and society. He derived the respective distortions from the market system and from the degeneration of society's defensive mechanisms. What constitutes the continuity between these and his earlier ideas was the criticism of the market system, the rejection of faith in advancement, the doubt about interpreting class interests as material interests, and the stress on non-anonymous historical processes resulting in the market system.

These ideas can be detected in the study 'Our obsolete market mentality' (Polanyi 1947) and in the volume *Trade and Market in the Early Empires* (Polanyi et al. 1957). However, the emphasis in the former is on the more accurate clarification of the concept of embeddedness, and in the latter, on the empirical analysis of primitive and archaic economic formations.

Embeddedness, according to the substantivist view, is related to the fact that an economic actor is a social being, with a mixture of motives, serving to attain social recognition and social goods. His economic efforts are therefore usually subordinated to these goals. Hunger and gain are not universal motives—Polanyi and his colleagues argued—thus the scarcity postulate, the cornerstone of formal economics, is inadequate. As is indicated in the title of Polanyi's key study, economy can be interpreted as an 'instituted process' in a substantive sense. The term process refers to 'location and appropriate movements,' that is production and transport on the one hand and transaction and disposition on the other.

Institution lends unity and stability to these processes, which can be approached via the forms of integration. Reciprocity presupposes the symmetry of groups, while redistribution is dependent upon collecting and distributing goods or disposal over them, hence it implies centricity. Exchange, in turn, depends on the existence of price-making markets. Not every exchange has integrative significance. Polanyi argued that the simple operational exchange and decisional exchange at a set rate do not integrate the economy, whereas bargained rate exchange does. The forms of integration are not arranged in phases, in stages of development as do the types of List or Bücher (Polanyi 1968); apart from the predominant forms, others are also present, constituting varying combinations. Trade, money, and markets are not necessarily functional upon exchange in the substantive approach, unless they are fitted in a market system. The ports of trade gift trade, and administered trade usually fall outside the market system.

Similarly, only one of the money uses coincides with exchange. The rest of the uses analyzed by Polanyi—payment use, standard, ideal unit, operational device—do not necessarily tally with the market, being the typical features of the redistributive intregrative form of the economy. The market, Polanyi sums up, is not so old a phenomenon, as are trade and money, and though it may crop up in history, it only integrates into a market system in exceptional cases.

While the former book contained an implicit debate with Hayek and Burnham, this one was challenged by *Economy and Society*, the book of Parsons and Smelser, the aim of which was to integrate economic and social theory and to find an alternative explanation on the place of economy in society (Parsons and Smelser [1956] 1957).

3. *Polanyi's Influence*

If it is contended with justification that Polanyi's views were deeply influenced by the writings of Bücher, Weber, Malinowski and others (Polanyi 1977), it must also be contended that Polanyi filtered these influences and thus conveyed them to late twentieth century social science. His ideas were discussed in several volumes, selections, and bibliographies (Stanfield 1986, Polanyi-Levitt 1990, Mendel and Salée 1991). Substantivism appears to be integrated in the history of sociology as a standard chapter (Hindess 1977).

The *Trade and Market in the Early Empires* volume, as well as the posthumous works edited mainly by his disciples (Polanyi 1966, Dalton 1968, Polanyi 1977) and the disputes coming to a head in economic anthropology suggest that a school was evolving around Polanyi. In economic history, Douglass C. North (North 1977) assessed Polanyi's influence in basically appreciative terms, though criticizing his market interpretation. Polanyi also influenced modern researches centering on the interrelation between state and economy as well as studies of socialist economy and its disintegration.

According to an evaluation, in spite of his marginal academic position, Polanyi was one of the most influential thinkers of the twentieth century. As far as the sociological analysis of the economy is concerned, this statement might be proved. Polanyi's concept of embeddedness became the starting point of the new economic sociology.

See also: Capitalism; Economic History; Economic Sociology; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Economics: Overview; Hierarchies and Markets; Institutional Economic Thought; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; Market and Nonmarket Allocation; Market Areas; Market Structure and Performance; Markets and the Law; Marxian Economic Thought; Money, Sociology of; Policy

History: State and Economy; Regulation, Economic Theory of; Regulation Theory in Geography; Retail Trade

Bibliography

Burnham J [1941] 1960 *The Managerial Revolution*. Indiana University Press, Bloomington
Dalton G (ed.) 1968 *Primitive, Archaic, and Modern Economies. Essays of Karl Polanyi*. Anchor Books, Garden City, NY
Hayek F [1944] 1976 *The Road to Serfdom*. The University of Chicago Press, Chicago
Hindess B (ed.) 1977 *Sociological Theories of the Economy*. Holmes and Meier, New York
Mendell M, Salée D (eds.) 1991 *The Legacy of Karl Polanyi. Market, State and Society at the End of the Twentieth Century*. Macmillan, Basingstoke, UK
North D C 1977 Markets and Other Allocation Systems in History: The Challenge of Karl Polanyi. *Journal of European Economic History* **6**(3): 703–16
Parsons T, Smelser N [1956] 1957 *Economy and Society. A Study in the Integration of Economic and Social Theory*. Routledge and Kegan Paul, London
Polanyi K [1944] 1957 *The Great Transformation*. Rinehart, New York
Polanyi K 1947 Our obsolete market mentality. *Commentary* **3,** February: 109–17
Polanyi K, Arensberg C M, Pearson H W (eds.) 1957 *Trade and Market in the Early Empires. Economies in History and Theory*. The Free Press and the Falcon's Wing Press, Glencoe, IL
Polanyi K (in collaboration with Rotstein A) 1966 *Dahomey and the Slave Trade*. University of Washington Press, Seattle, WA
Polanyi K 1968 Bücher, Karl. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. The Macmillan Company and the Free Press, New York, Vol. 2, pp. 163–5
Polanyi K 1977 In: Pearson H W (ed.) *The Livelihood of Man*. Academic Press, New York
Polanyi-Levitt K (ed.) 1990 *The Life and Work of Karl Polanyi. A Celebration*. Black Rose Books, Montréal
Stanfield J R 1986 *The Economic Thought of Karl Polanyi. Lives and Livelihood*. Macmillan, Basingstoke, UK

G. Lengyel

Police and Politics

Sir Frederick Maitland pointed out in the late nineteenth century that the etymological relationship of the words police, policy, politic, polity, politics, and politicians is very close (Maitland 1885). All are derived from the Greek *polis*, roughly translated as city-state. The earlier English meanings of police reflected this common origin, since police could mean a constitution or a polity. In the eighteenth century, the word came to have a more specialized meaning as 'the civil force to which it is entrusted the duty of maintaining public order, enforcing regulations for the prevention and punishment of breaches of the law, and detecting crime' (OED). This purely civil meaning was broadened (1837) to include 'any body of men, officially employed to keep order, enforce regulations, or maintain a political or ecclesiastical system.' The wider definition covers gendarmeries under military discipline and security services charged with the protection of public authorities.

The origins of the term have contemporary significance since unsuccessful attempts have been made, particularly in England, to divorce police from politics, and represent the police as nonpolitical guardians of society (Reiner 1992). The police and political authority are, however, intertwined—the police are executors of the decisions and the laws of duly constituted political authorities. In order to do so, they designated as having, in Max Weber's celebrated formulation about the state, the monopoly of the legitimate use of force. As Bittner (1979) remarks, the police 'are equipped, entitled and prepared to deal with every exigency in which force might have to be used.' The legitimate use of force by the police has been confined to the territories of sovereign states. The reality of a police monopoly of legitimate force within these territories has been contested because other agencies, such as prison officers, use physical constraints (Brodeur 1983). Other arguments contend that the actual use of force weakens the authority of the police by pre-empting negotiated solutions to conflicts and provoking sometimes serious opposition to the police (Monjardet 1996). Nonetheless, the concept of police necessarily includes the availability of legitimate force.

1. Police Systems

The long neglected comparative history of the relations between the police and politics has been reinvigorated by major studies since the 1980s (Bayley 1985, Liang 1992, Emsley 1999). In the eighteenth and early nineteenth centuries, professional police, paid by public funds, appeared. In the more developed parts of continental Europe, the purposes of the police were to secure the institutions of the state, enforce the law and the decisions of the ruler, and to protect propertied classes and property, both public and private, against criminals, calamities, and disturbances. In the English experience, except in London where the Metropolitan Police was established in 1829 by Act of Parliament, the modern police emerged from the office of constable, a medieval institution with equivalents elsewhere in Europe, who was charged with public tranquility at the local or parish level (Emsley 1996). In the United States, policing developed first at the local and state levels but, in the twentieth century, there was a proliferation of law enforcement agencies at the Federal level (Federal Bureau of Investigation, Drugs Enforcement Administration, Secret Service, etc.). However, the Federal agencies, in terms of numbers of police officers, constitute only a small proportion of the whole.

A distinction is made, sometimes too sharply, between an Anglo-American and a European 'continental' tradition of police–government relations (Mawby 1990). In general terms, hierarchically organized police forces controlled directly by a central government minister were characteristic of the most powerful continental European states. Mutual suspicion has often characterized police and politicians' attitudes on either side of the Anglo-American/continental European divide. The centralized systems have been regarded in Britain and the United States as instruments, often improperly used, in the hands of unscrupulous politicians. Misuse of police forces by the Nazis and other oppressive regimes gives this view credibility. In addition, national police forces are regarded as unwieldy and inefficient organizations. In the opposing European tradition, there has often been frank disbelief that police activities in decentralized systems can be professionally efficient, properly coordinated, and democratically controlled. They are also regarded as vulnerable to the influence of private interests, and as having uneven quality. Although much attenuated, this mutual suspicion has not entirely disappeared.

Differences between police systems have diminished because of a mutual learning process and because of internationalizing influences. Tendencies are apparent towards centralization in the British and some federal systems and to decentralization among the centralized continental systems. In the 1990s the British set up the National Criminal Intelligence Service (NCIS) to counter the increasing national and international mobility and activity of criminals. Local security plans, local coordination of policing, and more municipal police forces have modified the system of uniform central control. There have been increasing international police contacts and cooperation, and increased commonality of police methods and techniques, which, in turn, have an impact on police organization. The 'projection outwards' of police systems to the rest of the world, first through the European colonial empires and then US world influence after World War II, has resulted in hybrid systems and, as in Japan, some ingenious adaptations to local circumstances. Police systems are now distributed along a spectrum of types of organization, rather than based on contrasting models (Marenin 1996, Bayley 1996).

A greater degree of consensus than in the past exists among politicians in the advanced industrial democracies about the proper role of the police. Many of the less economically advanced countries have received assistance particularly from the United States, Britain, and France in police training and methods; they have, as a consequence, attempted to integrate, with varying degrees of success, philosophies about police–political relations. The consensus among the elites of the rich world (now including most electorally important parties of the Left) is that police forces exist to protect the constitutionally public authorities, to uphold the law, and to protect liberties, rights, and property. Direct political interference in police operations is less and less acceptable. Left wing radicals, often influenced by Michel Foucault (Foucault 1995), dissent from this consensus and regard these beliefs as useful myths to uphold an obnoxious status quo.

2. *Police Accountability*

The elite consensus leaves intact the longstanding problems of how the police ought to be controlled and made accountable. These problems raise the most politically sensitive question concerning the police—'who guards the guardians?'—a question first asked in imperial Rome by Juvenal (Satires 6 1), to which there is no universally accepted answer. In recent years, public concern has been voiced about police malpractice, corruption, violence, and intrusion into privacy. The contemporary policy dilemma is how to preserve a degree of professional autonomy for the police, essential for operational efficiency, while making police forces accountable to elected authorities, accountable before the law, and professionally accountable for their competence (Walker 1998, Wilson 2000).

In liberal democratic systems, the principle of accountability to the law and judicial authorities is uncontroversial but the methods of implementing the principle vary between countries and within countries. The police, in general, are subject to the ordinary law but this has been found to provide inadequate redress for complaints about police malpractice. The result is special arrangements for investigating such complaints (Goldsmith 1991). These arrangements are always the focus of political controversy and allegations are often made that the police have operated above or outside the law (Skolnick and Fyfe 1993). The police usually have considerable operational discretion but there are wide variations in the extent to which they are directly controlled by the political authorities.

Systems of political accountability vary along two dimensions—the extent of decentralization and the closeness of control of the police by elected officials. In the continental European systems, direct hierarchical accountability to a minister who is, in turn, answerable to parliament seems to guarantee the clearest lines of democratic accountability. However, these systems tend to generate three kinds of police–political scandals. These are: allegations of improper instructions given to the police to further political or personal interests of a minister; the political covering up of improper actions by police officers, to avoid embarrassment to the government; and improper police actions encouraged by the government, such as a tough law and order policy.

In decentralized and federal systems, political accountability for different police forces is dispersed. In federations, federal, state, and local authorities control different police forces. Improper behavior by police officers or relations between politicians and the police does not, in these systems, tarnish the whole of the police institution. Police malpractice such as institutionalized racism, corruption, and excessive use of force can, however, become endemic locally. In a constabulary system like that in England where the Chief Constables have a high degree of professional independence, democratic accountability may be indirect (and ineffective) through nominated police committees. The greatest difficulties of accountability arise in societies with deep ethnic, social, or religious divisions such as Northern Ireland, Israel, and South Africa, as well as countries with large immigrant minorities with characteristics clearly distinguishable from the host society (Brogden and Shearing 1993, Brewer et al. 1988). In these circumstances, the police inevitably are regarded by the powerless as oppressive instruments of the dominant group.

3. *New Issues*

Certain policing issues, broadly connected with political accountability, emerged on the political agenda in the last third of the twentieth century, in the highly industrialized democracies. The women's movement has had considerable influence over the investigation and prosecution of certain forms of crime—rape, violence against women, and pedophilia. The movement has also accelerated the number of women police officers in all highly industrialized democracies. The great political importance accorded to drug abuse, expressed by Presidents Nixon and Reagan's 'war on drugs' and their European emulators, has resulted in considerable diversion of resources into the pursuit of drug traffickers and drug abusers, the development of new and controversial investigatory techniques (controlled delivery, money tracking, intrusive electronic surveillance), and international law agreements.

The introduction of 'value for money' financial accounting throughout government has raised new questions about police effectiveness. It has proven politically embarrassing for police agencies, as increased expenditures are rarely reflected in improved clear up rates for crime. The determination to control public expenditure has also created pressures to develop nonstate forms of policing, which had precedents in the period before the establishment of professional police forces. Modern forms were a prominent feature of United States policing since the nineteenth century; such was the fame of the private security firm, Pinkertons, that some Europeans thought it was an official American police department. But private security firms have now become a controversial feature of policing internationally. The political left has been particularly critical of this development because it seems to provide an increased level of protection for the rich, but right wing supporters of national police forces are also very reserved about it (Johnston 1992).

4. *Transnationalization of Policing*

Developments in the international system must now be taken into account in the politics of policing because the police function is no longer completely constrained by the doctrine of territorial sovereignty. The establishment of international jurisdictions is changing, and has the potential to change radically the basis of police accountability. Protection of human rights has become a major priority of 'the international community.' International judicial protection for these rights is emerging and this erodes state sovereignty. France, for example, was condemned in 1999 by the European Court of Human Rights for torture following the brutal treatment in a police station of a Dutch citizen of African origin.

'Globalization' has led to increased concern about international crime in areas such as terrorism, drug trafficking, money laundering, international fraud, traffic in human beings, and disposal of high value stolen goods. International legal instruments have been negotiated, such as the 1989 Vienna Convention on drug trafficking which has direct effects on domestic law enforcement and criminal investigation techniques, such as 'controlled delivery.' Also, a certain transnationalization of police has occurred. This commenced as early as 1923 with the setting up of the International Criminal Police Commission, the predecessor of Interpol (Anderson 1989). US Federal law enforcement agencies have posted police officials abroad since the World War II—by the 1980s, the US Department of Drug Administration had over 60 offices in 40 countries (Nadelmann 1993). In Europe, Europol and the Schengen System (the compensatory measures of police and judicial cooperation for the abolition of border controls within the European Union) have gone much further in involving international regional institutions in police cooperation (Anderson et al. 1996). Transnationalization, together with the impact of the information technology revolution, undermines the traditional hierarchies of control (including political control) of the police. Policing is becoming a more knowledge-based, networking activity. In due course, this will transform all aspects of the relationship between police and politics.

See also: Criminal Justice, Sociology of; Justice and Law; Law: Economics of its Public Enforcement; Law, Sociology of; Leadership: Political; Police, Sociology of; Policing; Political Culture; Political Discourse; Prisons and Imprisonment

Bibliography

Anderson M 1989 *Policing the World: Interpol and the Politics of International Police Co-operation*. Clarendon Press, Oxford, UK

Anderson M et al. 1995 *Policing the European Union*. Clarendon Press, Oxford, UK

Bittner E 1979 *The Functions of the Police in Modern Society*. Oelgeshlager, Gunn & Hain, Cambridge, MA

Brewer D J et al (eds.) 1988 *The Police, Public Order and the State*. St. Martin's Press, New York

Bayley D H 1985 *Patterns of Policing*. Rutgers University Press, New Brunswick, NJ

Bayley D H 1994 *Police for the Future*. Oxford University Press, New York

Brodeur J P 1983 High policing and low policing: Remarks about the policing of political activities. *Social Problems* **30**(5): 507–20

Brogden M, Shearing C 1993 *Policing for a New South Africa*. Routledge, London

Emsley C 1996 *The English Police: A Political and Social History*, 2nd edn. Longman, London

Emsley C 1999 *Gendarmes and the State in Nineteenth Century Europe*. Oxford University Press, New York

Foucault M 1995 *Discipline and Punish*. Vintage Books, New York

Goldsmith A (ed.) 1991 *Complaints against the Police*. Clarendon Press, Oxford, UK

Johnston L 1992 *The Rebirth of Private Policing*. Routledge, London

Liang H-H 1992 *The Rise of the Modern Police and the European State System from Metternich to the Second World War*. Cambridge University Press, Cambridge, UK

Maitland F W 1885 *Justice and Police*. Macmillan, London

Mawby R I 1990 *Comparative Policing Issues; the British and American System in International Perspective*. Unwin Hyman, London

Marenin O (ed.) 1996 *Policing Change, Changing Police: International Perspectives*. Garland, New York

Monjardet D 1996 *Ce que fait la Police: Sociologie de la Force Publique*. Editions de la Découverte, Paris

Nadelmann E A 1993 *Cops across Borders*. Pennsylvania State University Press, University Park, PA

Reiner R 1992 *Policing and Politics*. Harvester Wheatsheaf, New York

Skolnick J H, Fyfe J J 1993 *Above the Law: Police and the Excessive Use of Force*. Free Press, New York

Walker N 1998 Police accountability. *Encyclopedia of Applied Ethics* **3**: 541–52 Academic Press, San Diego

Wilson J Q 1989 *Bureaucracy: What Government Agencies Do and Why They Do It*. Basic Books, New York

M. Anderson

Police, Sociology of

Sociology of police as a specialized area of study can be defined as the empirical and theoretical analysis of those organizational agencies—usually but not exclusively established by state governments—that are charged with investigation and prevention of crime and/or offences against formally adopted routines and rules as well as with maintaining public order, safety, and peace. It embraces the systematic sociological enquiry about the role of police in modern societies, the general characteristics of policing bodies and their relation to other institutions, the institutional development and internal structure of police forces, the collective and individual behavior of their members, and the interactions or relations of those agencies with individuals, different social groups, institutions, and the general public.

1. *Police as a Matter of Sociological Concern*

Both police and sociology are offsprings of modernity and progressively diversified societies. Although the modern contours of both became obvious only in the nineteenth century—the former as a new formal institutionalization of governance, the latter as an accomplishment of enlightenment and a new academic discipline—there is a line of discourse which can be traced back to ancient times, when Roman emperors struggled to get or keep their rule by using armed forces and when philosophers reasoned about the art of best governance. In spite of the fact, that the modern notions of 'police' and 'sociology' were not yet discovered then, the reasoning about the necessity of establishing and maintaining a public force to control crime and to ensure public order—and the critical discussion of its actual appearance—is as old as the discourse on the *res publica*.

Later on, the issue of police was embedded into the broader philosophical topic of the state and government—as it was for the early classic sociologists: police were discussed mainly as an instrument for the sovereign's power to govern. Max Weber's often quoted notion, that the state is that human community, that successfully claims a monopoly of legitimate use of force within a certain territory, still prevails within academic reflections on the police (Bittner 1991).

With the growing urbanization of the countries in industrial transition, there was also an increasing body of literature on police matters from a pragmatic perspective in the first half of the twentieth century. Yet it took some time for the police to become an explicit and original matter in sociological thought and empirical inquiry, primarily in English-speaking countries.

2. *Determining the Subject*

At first glance and seen from the surface it seems rather trivial to determine who and what constitutes 'the police.' One would expect that the police as a public institution is represented by (sworn) officers, i.e., representatives of the state and its government,

who are (often) specially trained professionals and are invested with certain powers (like the authority to search or arrest a person). Usually this 'apparent' kind of police will appear in public as officers on the beat, a crew in a patrol car, behind a desk in a police station, or as plain-clothes detectives, doing investigations in the field. This is the common and popular image (in the Western world) of what police actually are, transmitted by the media and supported by occasional ordinary encounters. From time to time this picture is widened by the appearance of riot police in full gear, underlining the state's authorization and capability of using legitimately physical force against disobedient citizens or rioting crowds.

When one looks again at who and what constitutes police, the semblance becomes more complicated: on the level of involved actors it becomes apparent that a good portion of the workforce employed by police consists of civilian staff, like secretaries or the clerks in the forensic laboratories; some sworn officers, such as staff who are responsible for conducting statistical analysis of criminal incidents or running a computer program for matching data, are rarely out in the field, while certain field officers, like undercover agents or specialists in charge of surveillance of telecommunication, often do not act openly as police. Furthermore there are other state agents who, having similar powers of investigation or intervention, perform certain functions of policing, but are not seen or labeled as 'police.' One might think here of custom officers, secret agents, public health inspectors, or prison guards. Still others appear in a policelike guise but are clearly not officers (at least lacking their full powers): city wardens, commercial guards and patrols, hired investigators, vigilante organizations like the 'Guardian Angels,' bodyguards, stewards in a football stadium, or bouncers.

As professionals trained to use force legitimately in the name of the state's power, police share this somewhat exclusive right with the military and other law enforcement officials (prison, customs). But on the level of routine work processes, this distinguishing feature is a rather rare event compared to the overall picture of police duties. Instead, research has shown that a considerable part of the modern police workload is not at all focused on crime control and investigation of criminal cases, but consists of responding to general emergency calls, mediating conflicts, regulating motor vehicle traffic, and communication with other institutions or agencies like social services, insurance companies, etc.

Although it might be clear from a commonsense point of view, what 'police' actually are—basically that formal institution which is vested with the powers and resources to respond to criminal acts or public order disturbances and calls itself 'police'—the subject in question gets more diverse, the more we see it from a perspective of a peculiar organized social activity rather than as a matter of institutionalism. Thus a true 'sociology of policing' would cover a wider area and embrace more organizations than the initial 'police studies.'

But even from an institutional approach one has to speak of 'the police' either from a very abstract level or from a single case point of view only. Besides the common similarities in terms of historical developments, organizational models, practical strategies and tactics, and legal accountabilities, every country's police system has its own particularities and unique arrangement of forces.

It should not be overlooked that police as a subject of sociological interest is linked in many ways with other social systems, all of them carrying their own, often overlapping, bodies of literature: as an instrument of executive governance 'the police' can be seen as a segment of the *sociology of the state*. Its quality of being an important part of the criminal justice system and its more or less explicit legal bindings does make it a component of the *sociology of law*. Last but not least, on the level of empirical studies there are clear ties to the field of *organizational sociology*.

3. *Changes in Focus and Emphasis*

Although some academic work on the police can be traced back to the 1950s, the first surge of sociological studies of police work occurred as an aftermath of the increasing social and political changes which affected most countries of the Western world from the early 1960s on. In a time when social and political conflicts were brought onto the streets, and rising concerns about crime and disorder became a public issue, 'the police became increasingly visible, controversial and politicized in response to these tensions and pressures' (Reiner 2000, p. 210).

According to Cain (1979), five types of studies, following a rough chronological sequence in the shift of areas of interest, can be identified for this period: (a) a law-related concern with civil rights, when confronted with police powers and discretion; (b) examinations of the police organization by participant observations (Skolnick 1966 and Manning 1977 are well-known examples); (c) studies of the role the police take in the social construction of deviants and the fabricating of public images of deviance; (d) analysis of the interactions and relationships of the police with the social environments they are working in; and (e) studies questioning the political contexts, presumptions, and consequences of police activities. It can be said that all these topics overlap in a way and have persisted during the following decades. There are two important points to note: one is that the main characteristics of this work, coming out of universities, were motivated by a concern with issues of police deviancy and generally tended to be critical of the police as a powerful institution of the state. The other

is that this research was created overwhelmingly in the USA and the UK.

In a second wave from the early 1980s on, police research was done more often outside the academic sphere. Governmental bodies like the Home Office in the UK began actively to do their own research on police issues. Furthermore think tanks and independent research organizations began to enter the field to offer their own agendas and expertise. This development enlarged and in a way altered the agenda of police research and with it the accessible knowledge on the sociology of police. Thus the momentum of studies changed—beyond the Anglo-Saxon universe—from an outside critique of police behavior and police institutions to questions of good practice and evaluation of reform measures, generated under policy and managerialist perspectives (Reiner 2000).

4. *Current Theory and Research*

It is still the case that the bulk of police studies which are relevant for the sociology of police in one way or the other is produced in English-speaking countries, with the USA in a clear lead. This might be seen as the reason why, on an international level, the discourse on police and policing issues mainly reflects on these countries. For example debates on 'community policing' or 'zero-tolerance strategies' originate from the development of police policies in the USA. But as matters of cross-border and international police cooperation have become more important since the early 1990s, there is a growing interest in comparative perspectives (Brodeur 1995, Findlay and Zvekic 1993).

The contemporary sociology of police and policing has to take two aspects into account. First of all, there are quite a lot of significant differences in how police institutions are organized and how policing is actually achieved between different countries. This is mainly due to the diverse national historical background. The diversity lies particularly in the arrangements of the police system influenced by legal, political, and occupational cultures (Busch et al. 1985, Monjardet 1996). The other aspect is that when it comes to the level of policing, i.e., how police work is actually achieved and what kind of resources and strategies are used, some striking similarities and convergences can be noted. There seems to be an increasing transfer of issues in the international police discourse.

The sociology of police is fed in a significant way by historical research on police institutions and forces, be it on the local or the national level. These studies show that there are certain issues in reflecting on police and policing, which have persisted for a long time, most prominent among them effectiveness, deviance, and accountability.

Contemporary research on police and policing embraces also immanent occupational and organizational aspects like recruitment and education, gender and minority issues, the role of unions and professionalism and specialization. The question at stake here is what police are actually doing.

A good portion of police research literature is still concerned with issues of internal as well as external control mechanisms of individual and institutional police behavior. In this respect the institutional autonomy of the police in the face of the government and the judiciary is one side of the question, the control of the behavior of individual officers another (Marx 1988).

Increasingly matters of police management and evaluation are discussed, especially connected with institutional police reforms in the UK. Modernization seems to have been an enduring issue in police reform since the 1970s.

In theoretically oriented academic research the emphasis is clearly shifting from an institutional understanding of police towards policing as a socially structured, dynamic, and multilayered process. The background for this hidden paradigm shift is on the one hand the growing role nonstate policing bodies (security enterprises, insurance companies, vigilant citizens etc.) are gaining in many countries (Shearing and Stenning 1987) and on the other the increasing relevance of technology and information for the police (Ericson and Haggerty 1997).

5. *Methodological Issues*

The police is a specific institutional field and bears special problems for sociological research. Studies on police and policing in a sociological sense started to be done by university researchers. As long as police behavior could be observed from the outside, especially in respect to its effects on citizens, it was a matter of external description and analysis. In a way these studies could be seen as a type of 'legal journalism,' mostly case-oriented. A new field of empirical data was opened, when researchers were able to do participant observation of police work. With that, the internal perspective of the field actors were reflected in the research on police. Of course, there have always been some difficulties for police organizations granting outsiders access to internal functional mechanisms, because it carries the risk of 'occupational secrets' becoming subjects of public discussion.

Generally, from a methodological point of view three points deserve attention. First, police studies relevant for sociological interests can be roughly divided between pure empirical, pure theoretical, and those studies which combine or at least claim to combine both aspects.

Another issue is, who initiates and/or carries out the research in question. It might make a difference for methodological approaches, as well as for the broadness of the theoretical perspective, if studies are done by researchers coming from police or other govern-

ment institutions or whether they are carried out by university institutes. Furthermore there is a clear difference between research *for* and research *about* the police. Usually the former tends to be preoccupied with practical issues or management problems. Although there quite often exist interesting sociological aspects within these perspectives, a genuine sociology of police and policing is coupled with research *about* the police, i.e., its meaning in relation to the law, the state, or society in general.

Finally, the term 'police' might be a category too broad when it comes to empirical analysis and theoretical conclusions. Not only do national police systems embrace a variety of more or less specialized forces, with specific, sometimes unique ways of division of labor among units, but police organizations also refer to very different geographical ranges of jurisdiction. This situation becomes more complicated when nationally focused studies are compared on an international level: for a full comprehension of the differences and convergences national particulars have to be taken into account.

6. *Future Directions of Theory and Research*

There seem to be three major paths of future research which will contribute to knowledge about the sociology of police. One line of advancement will come from the shift that police institutions will be interpreted more and more in terms of the market. One issue will be what can be called the 'economy of police and policing.' In this regard commercialization of police functions will become more important, not only in terms of private police or privatization of policing but also in terms of adopting service orientation as a primary goal and private business structures within the framework of public police. Another research topic will be connected with the growing internationalization of police work. The emerging concept of liaison officers, the creation of multinational task forces—Schengen and Europol as well as Interpol arrangements on policing, etc.—are headed toward changes in particular in the relationship between policing and space. Finally, recently emerging approaches in maintaining order in transnational zones of disorder (i.e., the former Yugoslavia), established drug control policies, and antiterrorism policies, demonstrate the apparent convergence of internal and external security and safety concepts, and convergence between military and secret services on the one hand and police on the other.

See also: Control: Social; Crime: Sociological Aspects; Crime, Sociology of; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Law: Economics of its Public Enforcement; Law, Sociology of; Police and Politics; Policing

Bibliography

Bittner E [1980] 1991 The functions of police in modern society. In: Klockars C B, Mastrofski S D (eds.) *Thinking About Police*. McGraw-Hill, New York, pp. 35–51

Brodeur J P (ed.) 1995 *Comparisons in Policing: An International Perspective*. Avebury, Aldershot, UK

Busch H, Funk A, Kauss U, Narr W D, Werkentin F 1985 *Die Polizei in der Bundesrepublik*. Campus, Frankfurt am Main, Germany

Cain M E 1979 Trends in the sociology of police work. *International Journal of the Sociology of Law* **7**: 143–67

Ericson R V, Haggerty K D 1997 *Policing the Risk Society*. Clarendon Press, Oxford

Findlay M, Zvekic U (eds.) 1993 *Alternative Policing Styles: Cross-cultural Perspectives*. Law and Taxation Publishers, Deventer, The Netherlands

Manning P K 1977 *Police Work: The Social Organization of Policity*. MIT Press, Cambridge, MA

Marx G T 1988 *Undercover Police Surveillance in America*. University of California Press, Berkeley, CA

Monjardet D 1996 *Ce que fait la police* (*Sociologie de la force publique*). La Découverte, Paris

Reiner R 2000 Police research. In: King R D, Wincup E (eds.) *Doing Research on Crime and Justice*. Oxford University Press, Oxford, UK, pp. 205–35

Shearing C D, Stenning P C (eds.) 1987 *Private Policing*. Sage, Newbury Park, CA

Skolnick J H 1966 *Justice Without Trial*. Wiley, New York

H.-J. Albrecht and D. Nogala

Policing

Public police are a bureaucracy created by political and legislative processes to enforce the statutory or common law of crimes, and to maintain public order and safety. Police in democratic states are accountable to the law and to democratically elected legislatures and executives. A 'social science' understanding of police draws upon theory and research in sociology, urban anthropology, law, criminology, history, and political science.

Social science research on policing scarcely existed before the 1960s. Moreover, observational and survey research on policing has largely been restricted to democratic regimes. Even in those, social science research may be constrained by police officials who limit access to investigators. Since the 1960s—a decade of rising crime rates, riot, and social protest—sociolegal research on police has flourished. An it has spread to democratic states around the world, including Canada, Eastern and Western Europe, Israel, Japan, Australia, India, Brazil, and South Africa. While earlier social science primarily sought to analyze police organization, role, and culture, later social science attempted to apply these insights to policing strategies and accountability.

1. Police Organization

Police structures vary. Some nations, like France, Italy, and Israel employ centralized police systems. The police in Japan, Australia, Great Britain, and Brazil are moderately decentralized. Police in the United States are decentralized, and include federal, state, county, urban, suburban, and rural forces. Nevertheless, like other bureaucracies, the police are everywhere shaped by social and economic conditions, by the interests and values of those who exercise or have access to state power, and a concern—or lack of it—for civil liberties and rights. Thus, police can differ significantly depending upon language, politics, tradition, and ideology.

1.1 Police Organization in Totalitarian States

In the first half of the twentieth century, aside from the militarized and centralized German police of the Nazi era, Soviet policing supplied a formidable model for policing a security state. From the beginning of the Bolshevik regime, the Cheka, the Soviet secret police, were given extraordinary powers to arrest, imprison, even to execute, in the interest of defending 'the dictatorship of the proletariat.' The detection and prevention of crime to provide public safety was secondary to the task of preserving the dominance of the political system. In contemporary China, although the police are striving for some independence, they have been regarded as a coercive instrument to serve the Communist party's needs. Nations in transition from dictatorship to democracy have struggled to transform the police from their formerly authoritarian responsibilities to what would be acceptable in a democratic nation, where police are accountable to the political authority of elected officials.

2. Democratic Policing

Historians generally credit Sir Robert Peel's London Metropolitan police to be the original, remarkable, and sustaining model for democratic policing. Yet the passage of *The London Metropolitan Police Act* of 1829 followed three-quarters of a century of deep suspicion and failed attempts to create a police force. Peel's achievement was to coordinate London's need for public order with democratic concerns and values. In 1826, before submitting the police bill, he abolished the death penalty for over 100 offenses, and removed numerous criminal penalties entirely. To win the support of the Whigs for his police bill, he temporarily excluded the City of London from the authority of the new police. To quell the fears of an emerging working class, he designed blue uniforms for the police to distinguish them from the army's red coats; and, what in retrospect must be seen as a brilliant and radical move, the men in blue did not carry guns while patrolling.

Furthermore, Peel and Colonel Charles Rowan, whom Peel appointed commissioner of the new police, strongly urged moderation in the actions of the police to gain public support. Although the purpose of the police was to protect private property and to maintain public order, it was understood that police power would be restrained by the rule of law, minimum use of force, and public opinion. As time advanced, and as class conflict diminished, English police were perceived to be efficient, incorruptible, accountable, and nonpartisan.

All this was to change after 1960. Urban riots, rising crime, and police scandals combined to undermine the warm confidence the public had shown the British police for more than a century. 'From almost complete invisibility as a political issue, policing has become a babble of scandalous revelation, controversy and competing agendas for reform' (Reiner 1992, pp. 73–4).

British historians have offered revisionist, theoretically Marxist, interpretations of the origins of the English police, emphasizing their role in diminishing class conflict in the development of Great Britain's industrial capitalism. Like their American counterparts, British sociologists have observed the power of subculture and political belief to mediate the influence of law on police behavior. Others, especially McBarnet, have implicated the legal structure—judges, senior officers, vague laws, case law, including those purporting to protect civil liberty—as the basic source of police deviance (McBarnet 1983).

2.1 The Police in the United States

No police department in America was so carefully planned and organized as the London Metropolitan Police. 'The employment of police in municipal administration was governed not by theory but by convenience' (Lane 1967, p. 221). Public order and rioting posed a more serious problem than crime in early nineteenth-century Boston, when the city established its first full-time police department in 1837. The precipitating event for the creation of the Boston police department was akin to a race riot, as roving bands of Protestants destroyed nearly every Irish Catholic home on Broad Street. Similarly, the creation of New York City's police force reflected a growing intolerance for riots and disorder rather than a response to an increase in crime (Monkkennen 1981, p. 553). Nevertheless like their eighteenth-century London contemporaries, New Yorkers thought themselves engulfed in a rising tide of crime and disorder. Negro slaves made up one-fifth of New York City's population and were 'an unstable and unruly element often

treated with great brutality' (Richardson 1970, pp. 4–5).

Interracial and interethnic hostility have always posed a problem for policing in democratic nations as well as totalitarian ones. In the United States, the combination of immigration, expansion, and slavery stand behind contradictory images of America as a land of promise and of oppression. The themes of brutality against blacks, of discrimination against immigrants, of immigrant hostility toward blacks—all set against a turbulent background of social and class conflict—arise time and again in American history.

Over time, American cities witnessed a progressive exchange of functions between the citizenry and the police. Initially, the citizenry were responsible for guarding themselves against criminals through a watch system. As other agencies assumed responsibility for other areas of municipal administration, the police began to be increasingly concerned with crime. Not until the turn of the twentieth century did the American police fully resemble the blue coated forces of the present.

Twentieth-century changes in the American police reflected population growth and diversity in cities, changing social and political organization of communities and neighborhoods, and in technology available to police. During the first reform wave, from the late nineteenth century to the Prohibition-inspired 'roaring twenties,' progressive elites (such as Theodore Roosevelt and his political associates) attempted to wrest control of the police from corrupt and powerful political machines (such as Tammany Hall in New York City). They also instituted civil service examinations and other measures intended to insulate officers and commanders from political influence.

The second wave, in the 1920s and 1930s, reacted to police scandals associated with national Prohibition. The 1931 Wickersham Commission report (1931) documented shocking corruption and brutality in the criminal justice system and motivated police professionals, such as August Vollmer, to introduce reforms. The efficient, seemingly incorruptible, semi-military Los Angeles Police Department, of the 1950s and 1960s, headed by William H. Parker was hailed as a model of such 'legalistic' professionalism, although the police behavior in the Los Angeles (Watts) riots of 1965 was to suggest otherwise. That riot was not as hurtful as the 1992 riot following the Simi Valley acquittal, in a California court, of the police who beat Rodney King. But at the time it took the largest toll of any riot in American history and was followed by significant riots across the nation, most prominently in Cincinnati, Cleveland, Newark, and Detroit.

The Civil Disorder (Kerner) Commission (1968) found that, in the first nine months of 1967 alone, there were 164 disorders of varying intensity, of which 41, in 39 cities, were considered 'major.' The report singled out the police as the activating cause of the urban riots that had shocked and mystified the American public. The 1960s was also a period of anti-Vietnam war protest, generating hostility between police and protesters.

A third reform wave followed, continuing into the twenty-first century. Social science theory and research was, and remains, implicated in this project. Two main subjects dominate: sociological theories of police role, culture, and discretion, based mainly on participant observation; and theories of police administration, especially community-oriented, problem-oriented, and quality-of-life policing, sometimes grounded in participant observation, but also based on theory and experimentation.

3. *Reorienting Police*

Historically, the police emerge as a significant, and sometimes controversial, agency of social control in periods of social and economic transition accompanied by crime and social unrest. In democratic societies, policing generates issues of how police should best be organized to deal with crime, and to be accountable to the communities they serve. Police researchers who had studied the 'professional' reactive model of policing in the 1960s and 1970s became increasingly skeptical of the capacity of police to satisfy community demands for public safety and dignity. Two decades of research suggested that motorized patrol scarcely deters crime, that detectives can solve only a small percentage of crime, that rapid response doesn't solve crimes, since by the time police are told of the crime, the offender has gone. Moreover, four US Presidential Commission reports (as well as Great Britain's Scarman Report) showed how supposedly neutral and detached policing styles could generate deep mistrust of police and estrangement from the communities they policed.

Three new approaches to police organization, values and purpose were articulated and eventually studied by social scientists in the attempt to reduce crime, increase public safety, and renew some degree of confidence in urban police: community oriented policing (Skolnick and Bayley 1986, Moore 1992), problem oriented policing (Goldstein 1990), and quality of life policing (Kelling and Coles 1996, Livingston 1997). Most social scientists agree that policing is only one of a number of causes affecting crime rates, and even then demands considerable organizational rethinking and practice (Blumstein and Wallman 2000).

3.1 *Community-oriented and Problem-oriented Policing*

By the 1980s, police relations with the communities they served had become a major issue for industrial democracies in Western Europe, Australia, New Zealand, Canada, Japan, and in the United States. James Q. Wilson had demonstrated that American police departments of the 1960s had adopted 'styles'

of policing—'legalistic,' 'watchman,' 'service'—that greatly influenced their performance. 'Community-oriented' (COP) and 'problem-oriented' policing (POP) are best understood as styles of policing which, although widely discussed and praised, have not generated a consensus as to their meaning or effectiveness. Nevertheless, although there is no standardized set of precepts of COP, three elements are generally regarded as central: (a) that police and citizens are 'co-producers' of crime prevention; (b) that policing be decentralized, with focus on neighborhood groups cooperating with police; and (c) that patrol be reoriented, from vehicles to streets. In addition, COP seems especially sensitive to the police need for racial, linguistic, and gender diversity.

PO Policing is frequently associated, and sometimes confused, with COP and for good reason. Each stresses community involvement. Each is critical of traditional 'incident oriented' policing. POP suggests that police managers review 'incidents' in consultation with community members to focus on underlying causes. Furthermore, POP envisions the police executive as a skilled politician able to mobilize the larger community—the media, and organizations within and outside of government—in the service of addressing underlying problems.

3.2 *Quality of Life Policing*

In an influential essay, James Q Wilson and George Kelling (1982) maintained that disorder and crime are usually inextricably linked, in a kind of developmental sequence. Just as a broken window, if neglected, leads to more broken windows, minor disorders (such as littering, loitering, public drinking, panhandling, and prostitution), if ignored, invite more serious criminality. Applying the theory as chief of New York City's Transit Police, William Bratton and his associates scrupulously removed graffiti, and cracked down on turnstile jumpers, panhandlers, and others violating subway regulations. The result was an astonishing reduction of subway crime and a comparably successful rise in ridership.

With the election of Rudolph Giuliani as New York City's mayor, Bratton became Commissioner of the New York City Police Department (NYPD). He merged the Transit Police into the NYPD and introduced many of its precepts. In addition, he set up a sophisticated computer mapping system (COMPSTAT) to assist police commanders in allocating resources, and to hold them responsible for crime reduction. New York City crime rates subsequently declined sharply. Although proving cause and effect is difficult, since other factors might have contributed (improved economy, higher incarceration rates, changes in drug preference, demographic shifts) close observers of the police acknowledge that these police strategies reduced crime. Bratton's successor, Howard Safir, employed the strategies more aggressively. The aggressive policing, plus highly publicized instances of excessive force, led to hostile relations between the police and the communities they serve.

4. *Discretion, Culture, and Deviance*

Wherever social scientists have observed police in action, as patrol officers or detectives, in Amsterdam, London, Los Angeles, New York, Sydney, or Toronto, their orientation to the job appears to be influenced by a profusion of unwritten rules and understandings significantly influencing conduct and misconduct. Sociologists and criminal law trained observers of the 1960s were impressed by the vast amounts of discretion that police enjoyed. Police on patrol were seen ignoring minor thefts, drunkenness, disturbances, assaults, and malicious mischief in the interest of maintaining public order. These discretionary practices rendered a significant part of the criminal justice process invisible and unreviewable.

William Westley's pioneering study of the Gary, Indiana police examined how the strains of the occupation—dealing with criminals, the insane, the intoxicated, often in an adversarial relationship to the public—influenced police conduct—and misconduct—on the job (Westley 1970), Skolnick (1996) built on Westley's insights to develop a theoretical role construct, the policeman's 'working personality.' Skolnick asserted that significant features of the police role—danger, authority, and accountability to superiors—were a precursor to understanding how police perceived and acted upon the world around them.

Bittner (1970) contended that the core of the police role is their capacity to employ force. Although police often aid sick and troubled people, Bittner asserted, their main involvement is to provide coercion where needed. By now social scientists have produced a substantial literature documenting the role strains, ordinary practices, and cultural understandings of urban police.

4.1 *Code of Silence*

Although subcultural understandings vary, the legendary silence code seems to be the one powerful norm uniting police across departments and nations. Maurice Punch, who studied police deviance and corruption in New York, London, and Amsterdam, writes that police operate by a code of silence which 'dictates that you do not 'rat on your mates,' (Punch 1985). In Los Angeles, more than 20 police officers witnessed Rodney King's beating. Those who administered it must have assumed that those watching would not come forward and would back up any stories they would tell; and that departmental super-

iors would believe them. Officers in the New York City Police Department's 70th Precinct did not protest when they saw Officer Justin Volpe marching around the station house, waving a broken broomstick stained with blood and feces, the residue of his sadistic anal attack on Abner Louima.

Testifying before the 1933 Mollen Commission in New York, which was investigating a pattern of corruption that had grown into vicious criminality, Officer Bernard Cawley affirmed the silence code, and the reason behind it: 'Just say if a cop decided to tell on me, his career's ruined. He's going to be labeled as a rat.' The code of silence may be that of aspect of police culture most frustrating to the general public and to officials responsible for police oversight, including prosecutors, judges, legislators, and high ranking police officials.

4.2 *Excessive Force*

Following his study of police violence in American, Caribbean, and South American cities, Chevigny concluded that police organized on a highly military model, in socially stratified societies which have slight regard for rights, and a tradition of private vengeance are more likely to permit police to use arbitrary violence (Chevigny 1995). 'Police brutality' is an evocative media term, but it has no legal referent. To ask someone to move on, to employ physical coercion, or to shoot to kill, are all appropriate police conduct, depending on circumstances. Nevertheless, since every arrest, every handcuffing, involves an imposition of force on an essentially unwilling person, departments develop subcultural understandings about how to deal with people who challenge police authority. The widely televised beating of Rodney King by Los Angeles Police Department officers revealed a use of force so brutal it shocked viewers around the world. Yet a suburban, predominantly white jury found that the force was not 'excessive' since King was allegedly resisting arrest. (In a second, federal trial, held in Los Angeles, the officers were found guilty of violating King's civil rights.)

A blue ribbon commission (headed by Warren Christopher to investigate the complicity of the LAPD in the beating found that the LAPD reflected a counter-productive organizational culture. Officers were rewarded for the number of calls they handled and arrests they made as well as for being 'hard nosed.' Based on its time-honored notion of 'professionalism,' the department, the Christopher Commission concluded, 'emphasized crime control over crime prevention and isolated the police, from the communities and the people they serve,' and evidenced a 'seige mentality' (Christopher 1991).

In New York City, a particularly sadistic assault on a Haitian immigrant, Abner Louima, followed by two shootings of innocent African-American men (Amadou Diallo and Patrick Dorismond) led to angry protests against police use of force, especially deadly force. As in Los Angeles, critics pointed to the police culture and organizational strategies—especially 'zero tolerance' for street infractions—to explain the use of excessive force.

4.3 *Culture and Corruption*

Although researchers have distinguished among levels of police culture in management and street cops, usually, students of police culture have focused on street cops. Subcultures of police departments generate understandings about low-level deviance, such as when it is permissible to accept a favor, a free meal, a discount on purchased goods, or an outright money bribe. Corruption scandals have tainted the police in London, Amsterdam, and Brussels, as well as in the United States. Low level corruption can all too easily slide into a higher level, diluting the distinction sometimes made between 'clean' and 'dirty' money. That distinction was to break down in New York when the Knapp Commission found that officers had extorted money or narcotics from drug dealers, had failed to take enforcement action against narcotics criminals, and had given false testimony. Around a quarter of a century later, the Mollen Commission found that lying under oath was so routine the New York police had a word for it—'testilying.' A 2000 corruption scandal in the Los Angeles Police Department—which includes allegations of police shooting and killing drug dealers, beating them, stealing and selling drugs, evidence planting, false arrests, and perjury, has given the LAPD the unwelcome distinction of being the first major US police department to be monitored by the federal government.

5. *Private Policing*

With the development of mass private property in the United States—e.g., gated communities, malls, industrial and amusement parks—more than double is spent annually on commercial private security ($90 billion) than on public policing ($40 billion). Other countries, like South Africa, maintain an even higher ratio of private to public police. In light of this growth, private policing has come to have significant implications for theories of social control in democratic societies.

Although private policing sometimes resembles, overlaps, and is confused with public policing, private policing is distinctive in one major respect. Unlike public police, private police are primarily accountable to their employers, not to the citizenry. Thus, private policing is ultimately grounded in different interests, in turn influencing its understanding of its role and practices.

While public police may act as 'peacekeepers' they are nevertheless authorized and expected to enforce the criminal law, which, at least in theory, imposes punishments based on just desert. Private police, by contrast, primarily enforce norms instituted by their employer. Thus, private police usually address violations of institutional standards of comportment (dress codes, noise levels): try to prevent material damage to property (through theft, vandalism, or terrorism); and seek to avoid disruption of routines (disorderliness) of the sites for which they are responsible—such as universities, hospitals, offices, industrial parks, or airports. To achieve these goals, security firms typically adopt a *preventive* approach to misconduct—which community- and the problem-oriented police do as well—thus blurring the distinction. Indeed, Ericson and Haggerty (1997) have argued that, like private police, modern police have largely become risk managers rather than crime controllers.

In the United States, laws protecting ownership of property permit private clients to define the meaning of security and order within their property boundaries more narrowly than public police, modified only by constitutional and legislative limits protecting free speech, minorities, and women. For example, behavior acceptable in a public park, e.g., walking barefooted, may be unacceptable in a restaurant or private place of amusement, such as Disneyland. The typical sanction for violating private property rules in *exclusion* or *ejection*. Low-level criminal behavior, such as smoking a marijuana cigarette, may also result in ejection or the threat of ejection, rather than the punishments embedded in criminal statutes.

Private police in the United States are not constrained by constitutional protections (e.g., the Fourth Amendment), nor are they limited by the same kinds of civil liabilities (e.g., §1983 civil rights claims) to which public police are subjected. Private security typically derives its powers from, and to an extent is controlled by, the laws of property, contract, and labor. Nevertheless, though many states do regulate the licensing of guards and their ability to use firearms, private policing in the USA remains largely self-regulating.

6. *Controlling Police Misconduct*

Procedural law, especially under the US Constitution's Fourth (search and seizure) and Fifth (self-incrimination) Amendments, sets boundaries for police who are investigating crime by excluding evidence obtained in violation of these safeguards. Nevertheless, such protections have been narrowed by the US Supreme Court, especially in connection with drug enforcement.

Since street law enforcement involves stopping, questioning, sometimes searching and arresting, as well as ordinary peacekeeping, complaining citizens usually report one or more of the following: excessive force (hitting, pushing, tight handcuffs); abuse of authority (ordering change of location); rudeness; and ethnic slurs. Citizen Review Boards have become increasingly relied upon to address such issues and are found in Canada, Australia, Northern Ireland, and the United States. In the USA more than 80 percent of major cities have introduced such organizations. Walker (2001) concludes that policy review and change is more important than officer punishment to raise departmental performance, particularly since it is difficult to resolve cases where an officer and a complainant offer conflicting accounts.

External monitors, Walker believes, should prove more effective than citizen complaint review in elevating police performance. Civil liability suits, and most recently, federal and state oversight, have also been initiated to control police departments which appear to have deviated from their mission, either through corruption, excessive force, racial discrimination, or all three. There are no perfect models for policing or for controlling police. There is more agreement by social scientists and the public on what should not happen—corruption, brutality, racism—than on the mechanisms for controlling police. Nevertheless, certain themes related to police accountability are evident in the social science literature. First, police organizations are influenced by ideology, the most significant being a commitment to political democracy. Yet even within democracies, policing is influenced on a daily and regional basis by politics and power relationships. Second, policing is strongly influenced by the subcultural norms of policing. Third, courts possess only slight capacity to constrain police discretion on streets, neighborhoods, or management offices. Fourth, victimless crime enforcement breeds corruption, brutality, and even racism, as when minorities are racially profiled to enforce drug laws. Finally, police are more effective and accepted to the degree that they involve communities in developing their policies and practices.

See also: Criminal Justice, Sociology of; Law: Economics of its Public Enforcement; Police and Politics; Police, Sociology of; Professions, Sociology of

Bibliography

Bayley D H 1985 *Patterns of Policing*. Rutgers University Press, New Brunswick, NJ

Bayley D H 1991 *Forces of Order: Policing Modern Japan*. University of California Press, Berkeley, CA

Bayley D H 1994 *Police for the Future*. Oxford University Press, New York

Bittner E 1970 *The Functions of the Police in Modern Society*. National Institute of Mental Health, Chevy Chase, MD

Blumstein A, Wallman J (eds.) 2000 *The Crime Drop in America*. Cambridge University Press, New York

Chevigny P 1995 *Edge of the Knife: Police Violence in the Americas*. The New Press, New York
Christopher W 1991 *Report of the Independent Commission on the Los Angeles Police Department*. Los Angeles, CA
Dixon D 1997 *Law in Policing: Legal Regulation and Police Practices*. Clarendon Press, Oxford, UK
Fu H L 1994 A bird in the cage: Police and political leadership in post-Mao China. *Policing & Society* **4**: 277–92
Ericson R V, Haggerty K D 1997 *Policing the Risk Society*. University of Toronto Press, Toronto, Canada
Fogelson R M 1997 *Big City Police*. Harvard University Press, Cambridge, MA
Goldsmith A J (ed.) 1991 *Complaints Against the Police: The Trend to External Review*. Clarendon Press, Oxford, UK
Goldstein H 1990 *Problem Oriented Policing*. Temple University Press, Philadelphia, PA
Hungarian Helsinki Committee (ed.) 1988 *Police in Transition: A Survey of the Police Forces of Bulgaria, the Czech Republic, England and Wales, Germany, Hungary, Poland, Romania, Slovakia, Yugoslavia*, CD-Rom. Hungarian Helsinki Committee, Budapest
Ianni F, Reuss-Ianni E 1983 Street cops versus management cops: The two cultures of policing. In: Punch M (ed.) *Control in the Police Organization*. MIT Press, Cambridge, MA
Kelling G L, Coles C M 1996 *Fixing Broken Windows: Restoring Order and Reducing Crime in our Communities*. The Free Press, New York
Kerner H 1968 *Report of the National Advisory Commission on Civil Disorder*. US Government Printing Office, Washington, DC
Knapp W 1972 *Report on Police Corruption*. George Braziller, New York
Lane R 1967 *Policing the City: Boston, 1822–1855*. Harvard University Press, Cambridge, MA
Livingston D 1997 Police discretion and the quality of life in public places: Courts, communities, and the new policing. *Columbia Law Review* **97**: 551–672
Manning P 1977 *Police Work*. MIT Press, Cambridge, MA
McBarnet D J 1983 *Conviction*, 2nd edn. Macmillan, London
Miller W R 1977 *Cops and Bobbies*. University of Chicago Press, Chicago
Mollen M 1994 *Anatomy of Failure: A Path for Success*. Commission Report, City of New York, New York
Monkkennon E H 1981 *Police in Urban America: 1860–1920*. Cambridge University Press, Cambridge, UK
Moore M H 1992 Problem solving and community policing. In: Tonry M, Morris N (eds.) *Modern Policing*. University of Chicago Press, Chicago
Muir Jr K W 1977 *Police: Streetcorner Politicians*. University of Chicago Press, Chicago
Reiner R 1992 *The Politics of the Police*, 2nd edn. Harvester, London
Richardson J F 1970 *The New York Police: Colonial Times to 1901*. Oxford University Press, New York
Rubinstein J 1973 *City Police*. Ballantine, New York
Shearing C, Stenning P 1982 *Private Security and Private Justice: A Review of the Policy Issues*. The Institute of Policy Studies, Montreal, PQ
Shearing C, Stenning P 1984 From Panopticon to Disney World: The development of discipline. In: Doob A, Greenspan E (eds.) *Perspectives in Criminal Law*. Canada Law Books, Aurora, ON
Sherman L 1978 *Scandal and Reform: Controlling Police Corruption*. University of California Press, Berkeley, CA
Skolnick J H 1966 *Justice Without Trial: Law Enforcement in Democratic Society*. Macmillan, New York
Skolnick J H 1969 *The Politics of Protest: A Task Force Report Submitted to the National Commission on the Causes and Prevention of Violence*. Simon and Schuster, New York
Skolnick J H, Bayley D H 1986 *The New Blue Line: Police Innovation in Six American Cities*. The Free Press, New York
Skolnick J H, Fyfe J J 1993 *Above The Law: Police and the Excessive Use of Force*. The Free Press, New York
Stone C E, Ward H H 2000 Democratic policing: A framework for action. *Policing and Society* **10**: 11–46
Szikinger I 2000 The Roma in Hungary. *Policing and Society* **10**: 91–120
Tonry M, Morris N (eds.) 1992 *Modern Policing*. University of Chicago Press, Chicago
Walker S 2001 *Police Accountability: The Role of Citizen Oversight*. Wadsworth, Belmont, CA
Westley W 1970 *Violence and the Police*. MIT Press, Cambridge, MA
Wickershan G W 1931 *Report on Lawlessness in Law Enforcement*. National Commision on Law Observance and Enforcement. US Printing Office, Washington, DC
Wilson J Q 1968 *Varieties of Police Behavior*. Harvard University Press, Cambridge, MA
Wilson G Q, Kelling G L 1982 Broken windows: The police and neighborhood safety. *The Atlantic Monthly* **249** (March): 29–38

J. H. Skolnick

Policy History: Origins

1. Introduction

Public policy is as old as the history of human governance. All rulers, irrespective of the form of public rule, have sought to shape and manage the substantive order governed under their authority. Modern public policy composes a narrower domain. Co-extensive with the complex, and continuing, process of state formation (and de-formation) begun in late medieval and early modern Europe (Tilly 1975), public policy has come to refer more specifically both to what governments do as they transact with civil society, the economy, and states within a global state system and to the creation and deployment of knowledge about these sites of authoritative transactions. The semi-independent origins of policy knowledge and policy activity constitute my subject. I focus primarily on two formative moments in the West: (a) the early modern period marked by the growing concentration of state sovereignty over territory and people, and (b) the midnineteenth through to the early twentieth century when the 'social question' and unaccustomed levels and forms of international violence pressed to the fore. 'Origins' itself a word with

a composite meaning. It refers both to the beginnings of a process and to the constituent elements of what comes next. The results of decisions taken at the critical junctures I discuss, set in motion developmental pathways that further defined the contours of modern public policy.

2. *Early Modern Europe*

The very existence of public policy in the modern sense can be traced to sixteenth and seventeenth century postfeudal Europe, a period marked by the accumulation and concentration of sovereignty into national territorial states (Watkins 1934). These units claimed unique authority within their borders and recognized boundaries demarcating states from each other (in this way, international relations became distinguishable from domestic politics). Each State possessed sovereignty based on law, and with it legitimate force within a distinctive territory, an ensemble of institutions, and a vision and articulation of the common good. With these attributes, the State emerged as a calculating actor vis à vis other States, the newly separated economic sphere, and a newly distinguishable civil society (Skinner 1989).

This consolidation of sovereignty into increasingly centralized units produced the standardization of laws, coinage, taxation, language, and responsibility for security, all of which were requisites for systematic public policy. A corollary process was that of a growing set of structural separations within national boundaries distinguishing the realms of sovereignty, property, and civil society. Now, states could not simply impose their will despotically. They had to create and deploy policies to transact and coordinate with other powerholders. State–society relations now became reciprocal as policy came to be synonymous with the outputs of governments seeking to regulate an increasingly challenging environment (Mann 1988; Tilly 1990; Spruyt 1994).

The modern economy likewise was the product of this postfeudal differentiation of spheres. The key point in the development of capitalism as the dominant framework for economic development came only with the concentration of sovereignty and the establishment of an authoritative framework for property rights and economic transactions on a large scale. Capitalism prospered and urbanization (as well as rural proto-industrialization) accelerated once states emerged that were not merely extortionist but shared an interest in creating the conditions required to organize independent market transactions. Whereas early medieval face-to-face exchanges had taken place in towns whose existence depended on grants of autonomy from local lords, thus creating insecurities outside this tight embrace, by the sixteenth century, kings, in exchange for revenue, protected towns and traveling merchants on a much larger scale. With states providing a framework for capitalist development and not acting merely as rent-seekers, the linkage between the economy and the polity was utterly transformed into a new kind of strategic game. States did not ensure just any property rights, but only those consistent with state interests; capital did not give obeisance to any state, but only to those that secured its economic activities. The terms of this relationship thus became the first of two fundamental pivots of modern domestic politics in the West (North 1985).

The second was the hinge between the State and civil society. Separated from property, and concentrated in authority, the States of postfeudal Europe had to forge new, and uncertain, ties to civil society. These exchanges were marked by contestable questions about the range of activities over which the State would have authority—that is, the extent of its *hubris* in penetrating and regulating civil society—and about the ways in which members of civil society could affect the activities of the State. With the breakup of the tightly knit juridical, economic, and social units characteristic of feudalism, states could not simply compel by imposition. Instead, it was a condition of their effectiveness that they transact with society and coordinate aims with 'private' powerholders.

State–society relations thus became more reciprocal and less arbitrary; over time, the very ability of the State to utilize its authority to intervene in society produced commitments that subsequently bound the State. (Pitkin 1969; Downing 1992; Ertman 1997) The terms of these exchanges became the objects of political struggle and the subjects of expert consideration by political theorists and by political economists. Hobbes, Locke, Montesquieu, Rousseau, Millar, Steuart, the Physiocrats, and Smith, amongst a cast of many, mostly lesser, others, theorized these relationships and sought to find rules of interested engagement between the relevant parties. (Hirschman 1977) At stake was both the scope of responsibility of states for economic and moral situations and the ways in which the actions of states could be shaped by people, interests, and values outside the State itself.

The last key zone of transaction was international. Based on sovereignty understood as exclusive control over a particular territory and its people, a new international system came into being that substituted modern states with claims to unique control for formerly more fluid and overlapping jurisdictions. This change did not occur all at once, and its dimensions and decisiveness were masked for quite some time by the existence of dynastic empires, but the new global order of sovereign actors transformed the meaning of international policy about war and peace. Now, offense was geared to the extension of exclusive geographical sovereignty; defense to its protection. Contests over borders were settled by force or by diplomacy, both of which required new kinds of social knowledge (Doyle 1997).

These three sets of transaction were arrayed in a hierarchy, with the international dimension of war and peace designed to secure or extend sovereignty first amongst them because the existence and protection of such exclusive claims was the premise of the very existence of domestic policy. International policy also defined ambitions for overseas empires and for the guardianship of commerce in the globe's new Europe-centered market system (Zolberg 1981). As conflict between and amongst states now took on a global cast, so, too, did specialized policy knowledge about subjects as wide-ranging as military organization and the human qualities of 'primitive' subject populations.

Producers of modern policy knowledge were those intellectuals who came to understand and manage these new sets of transaction (Bauman 1992). Such policy intellectuals, after Machiavelli, created a modern science of government with two main ends: to point rulers towards more effective transactions among the State, economy, and civil society, and to police, educate, and civilize the population in order to improve its character and render it more rational and governable. They counseled how to expand the policy role of governments to address the security and portability of property under broadly mercantile economic arrangements and doctrines and how to govern populations subject to the authority of the State. Guided by such policy knowledge, as the territorial scope of states either expanded or became more secure, as the infrastructural capacities of governmental bureaucracies grew (including their abilities to gather and deploy statistical and other forms of information), and as a communications revolution in print and in advances to transportation took hold, governments took on increasing responsibilities for the solution of collective problems. These included new roles for both local and national authorities in such areas as poor relief, sanitation, and schooling (de Swaan 1988). Each time a new zone of policy activity was established under public aegis, either complementing or supplanting the responsibilities of private religious or secular responsibility, new specialist managers and intellectuals fashioned distinctive sets of policy ideas to organize these State-based activities. This extension of responsibility transcended even such fundamental divides as the *ancien régime* and post Revolutionary France (Woloch 1994).

In the long epoch from early modern Europe to the nineteenth century, mercantile thought produced a powerful and dominant recipe for state-building and for rules to govern international economic transactions, to shape those joining the State to the domestic economy and civil society. Its main aspects included 'regulation of the economically strong, support and direction for the economically weak, and the State's own enterprise where private initiative is wanting' (Krieger 1963, p. 557; Anderson 1974). At its early moments mercantilism was put to use for conservative purposes to buttress a social order in crisis; later, under the initiatives of Colbert, mercantilism went hand in hand with the protection of property rights and the reduction of transaction costs to provide a dynamic basis for capitalist economic development. Both facets were integral to the process of State-building in which:

> regional rulers utilized provisions for welfare as well as force to extend their control over all the inhabitants of their realms. Indeed, it was precisely the need for such a politics of welfare over a larger area when the old local welfare arrangements were being broken down that went far to establish the ruler as the head of the State. (Krieger 1963, p. 557)

3. *The Late Nineteenth and Early Twentieth Centuries*

During the first half of the nineteenth century, these formulas were pressured by the emergence of independent labor markets of more than local and regional scope (Polanyi 1944; Agnew 1986). With the introduction of such mechanisms to allocate people to jobs and set wage levels, capitalism entered a new phase. Concurrent with the development of labor markets (as well as autonomous housing markets) that helped generate new working class identities, dispositions, and patterns of collective action (Katznelson and Zolberg 1986), the period was marked by the development of compelling conceptions of popular representation and citizenship defined by the American and French Revolutions. All members of civil society now became actual or potential citizens. No longer could citizenship be restricted to those who, by holding property, demonstrated a stake in the existing commonwealth. Instead, the possibility of political (and military) participation was universalized, in the expectation that in a community of citizens people would act rationally in pursuit of individual, group, and common interests. Perhaps more so at this time than at any moment since, citizenship now possessed a radical, emancipatory edge (van Gunsteren 1985).

Within this field of tension, public policy was transformed radically in range and content. In this novel circumstance the 'social problem' emerged as a key category of public policy. This emblematic term condensed the new tensions of markets and citizenship. It put at the core of public life the question of whether, and to what extent, the political relations of citizens might modify the operation of markets, as well as the extent to which the rationale of the marketplace would dictate limits to citizenship. Prerevolutionary approaches to citizenship (whether Roman, medieval, or Enlightenment) had not ignored the connection between citizenship and minimum levels of security, but they had dealt with it by excluding those without economic means from the community of citizens. No longer was this solution possible. As citizenship

became more inclusive, states extended the scope of their policies aimed at organizing markets and mitigating their distributional inequities. Further, a modernist policy intelligentsia, nourished by Enlightenment conceptions of reason, considered citizenship, markets, and their intertwining with the confidence that they could find scientific solutions to ethical problems. Their methods became more specialized, and the professionals overtook the amateurs.

In situations characterized by the crystallization of the new institutions of markets and citizenship, practitioners of social knowledge perforce could accept or repudiate these foundations of policy. Throughout the nineteenth and early twentieth centuries, deep reactionary currents of refusal and restoration, radical renunciations of the new moral economy (Hirschman 1991), and forward-looking blueprints, the most important of which was a powerful Marxist critique and teleology, assertively rejected the new institutional and normative framework of liberalism. By contrast, the great majority of social analysts accommodated with varying degrees of enthusiasm to the new order and to its immanent power relations, and they sought to solve problems within the liberal doctrinal and institutional premises.

Even within this framework, there existed a wide array of alternative conceptualizations, ways of organizing knowledge, and contestations about policies, from right to left (at both poles straddling the liberal/non-liberal divide); but these debates, no matter how hotly contested, came to be circumscribed within a distinctive family of possibilities and they focused on a shared subject matter: the field of tension created by the simultaneous development of transactions between states and markets, especially as they concerned labor, and between states and actual or prospective citizens within civil society. The harsh realities of unequal and exploitative relations between economic actors helped create contentious class representations that intersected the radical impetus of citizenship to produce incendiary political possibilities, making both the modern State and modern capitalism precarious.

In these circumstances, producers of social knowledge—as individual scholars and theorists, and in organized settings such as universities, learned societies, political parties, organizations devoted to policy goals, and governmental bureaucracies—struggled with how to interpret and manage the new transactions linking State and market and State and citizen with very mixed and often confused motives: to preserve the social order and to reform it, and frequently both. Sometimes cast in a language of socialist transformation, sometimes in a conservative discourse, and sometimes in a vocabulary of individual rights, the center of gravity of the new social knowledge in the West was located in various attempts to find reference points and tools of administration for a form of social policy which could integrate conceptions of individual and collective goods. The policy intellectuals who undertook this quest were not utopians. Most saw no way to go back to a moment before citizenship and (labor) markets (indeed, in Britain, neither progressives nor conservatives in the late nineteenth century sought to return to the pre-1834 epoch, when social policy and economics were treated autonomously), and no way forward to a postliberal era within the ambit of normal politics (Harris 1990; Mandler 1987).

The features of policy and knowledge tended to be nationally-specific. In France, an optimistic anticlerical and republican program of the pupils of Comte and then of Durkheim articulated with the Third Republic. In Britain, New Liberalism conquered the field, displacing and reshaping virtually all its competitors; as it did in the United States, within the framework of Progressivism. In Germany, like so much else, social science and policy studies suffered the fate of the country's precarious liberalism, developing especially skittish qualities (Lepines 1988; Ringer 1969).

4. *Modern Policy Knowledge and Practice*

The place of pride in social analysis in the *ancien régime* had belonged to the moral sciences. Classical political economy had been elaborated within its embrace. In Britain, figures as diverse as Stewart, Malthus, Ricardo, the Philosophic Whigs, John Stuart Mill, and Bagehot, sought to maintain holistic portrayals of modern society while treating the specificity of the new transactions between its distinctive domains, and to do so by making the role of the State, in form and in policy, the centerpiece of their analyses. These pivotal thinkers were neither traditional political philosophers who deduced the character of the good society from first principles or moral ends, nor were they empirical social scientists who sought to comprehend the regularities governing the new transactions.

Over the course of the nineteenth century into the twentieth, this kind of relatively integrated social knowledge was replaced by professionalized, discipline-based attempts to grapple with the specific tensions inherent in the elaboration of markets and citizenship in liberal societies. At stake was the character of the extended responsibilities and jurisdiction of the State with respect both to markets and citizens. From the vantage point of the transactions between the states and markets, the central issue was how to shape the State's interventions to make markets function effectively and to make their distributional effects tolerable. From the perspective of transactions between states and citizens, at issue was the balance between freedom and domination. From the standpoint of international relations, the key issue was how to regulate the quest for power and plenty in a

manner that enhanced the prospects of peace and prosperity. Together, these questions elaborated the themes of moral philosophy under radically revised conditions (Thomas and Meyer 1984).

These were the concerns of a new genre of policy institution, the first example of which is Britain's Social Science Association founded in 1857. Created to consider the problems of social economics and to promote legislative reforms in the House of Commons with regard to education, law, penal policy, public health, and the regulation of markets, the Association self-consciously viewed itself as a conduit between politicians and an expanding citizenry with political rights, and it came to play an important interstitial role in the Victorian State by proposing ways to manage the 'social problem' attendant upon the simultaneous existence of labor markets and an expanded citizenry. Like Germany's *Verein für Sozialpolitik*, the Association fused reform and social science in its attempt to provide expert knowledge on behalf of a project its members conceived of as an endeavor of social integration. Working before the development of separate social science disciplines and before the creation of an autonomous and politically ambitious labor movement, the Association succeeded in joining under liberal auspices a wide array of intellectuals as well as Liberal MPs, trade union and cooperative movement leaders, and other reformers in a quest to bring expertise to bear on the project of creating a new liberalism capable of surmounting the now manifest inadequacies of traditional political economy (Goldman 1986).

This subsequent transformation in subject matter, style of work, and institutional organization was no simple linear development. There were two overlapping moments, each characterized by an important partition. In the first, spanning the 1860s through the 1910s, the knowledge community divided into specialists concerned with one or the other transactions of State, economy, and citizenship. Integrated knowledge separated into distinctive zones for the study of economics, politics, and society, ultimately devolving into wholly separate professional disciplines, each with distinctive intellectual tendencies (Oleson and Voss 1979). The second moment, from the 1890s through the 1920s, was characterized by a trend complementary to these new disciplinary divisions: a split between academic practitioners of the social sciences, located mainly in burgeoning institutions of higher education, whose policy concerns and impact were indirect, and scholars who emphasis was placed far more directly on policy matters and public influence.

As late as 1870, political science was hardly distinguishable from political economy and history. In the next 50 years, political science secured a disciplinary status divorced in large measure from both. It came to focus in a specialized way on constitutionalism, law, political parties, interest groups, voting and other mechanisms of representation; that is, on the rules of State–citizen transactions. During the same period, political economics shed its adjective and its claim to be an aspect, in its science, of moral philosophy. Recast as a scholarly craft in the mathematical language of post Jevons marginalism, it was not so much that economics no longer wished to concern itself with larger questions of human nature and social organization. Rather, the economists of the period sought to elaborate a science of markets on the basis of an elegant simplification about just these issues, but not for the sake of science alone. This new economics, for whom the key figure certainly was Alfred Marshall, was created in the service of the ethical goals of the period's New Liberalism. Under his remarkable influence, deductive work within economics replaced historical scholarship. This narrowing of focus was accompanied by a widening of theoretical ambition to create an economics as a timeless and institutionless body of knowledge (Soffer 1978).

Indeed, Marshall was an exceptionally important figure at two junction points: between classical political economy and disciplinary social science; and between the latter and the world of public policy. His *Principles of Economics*, published in 1890, explicitly sought to place pioneering analytical techniques at the service of human improvement. While Marshall's economics was neoclassical in the sense that all its aspects concerned the fluctuating equilibrium of supply and demand, this equilibrium depended for him on the subjective decisions of producers and consumers in pursuit of qualitative time-bound goals in pursuit of a better life. Once his new mathematical tools became central to economics, the discipline developed in ways that elaborated technique at the expense of his own value-laden preferences. The new Marshallian economics deepened its scientific capacities at the expense of its practical effects on policy issues of the day and on its ability to analyze, or even take an interest in, the rules of transaction governing the relations of State and market in particular historical settings. Just at the moment when, in the late nineteenth and early twentieth centuries, the social tensions of markets and citizenship intensified under the impact of economic depressions, the radical transformations in industrial size and class structure entailed in the second industrial revolution, international rivalries for markets, a new geopolitics characterized by imperial ventures, and, ultimately, by total war, the academic discipline of economics lost a good deal of its capacity to function as a practical policy science. In part this was a matter of a distancing from practical affairs, in part a flight to abstraction, in part a deliberate decision in favor of the ambitions of science at the expense of the ambitions of moral discourse (Church 1974; Nelson 1987).

By the end of the nineteenth century, classical political economy virtually had disappeared, as had

the gentlemen generalist historians and lawyers who had become archaic figures. These 'all-rounders' were steadily replaced by academic scholars in specialized social sciences and by a growing body of policy specialists who claimed knowledge that both drew upon and crossed over the boundaries that divided the social sciences. Labor experts, social workers, organizers of social surveys, public administration specialists, reformers concerned with social policy, and intellectuals within social movements and political parties—both inside and outside the administrative apparatuses of states—tried to come to grips with the intensification of the problems inherent in interactions between States, citizens, and markets. These policy intellectuals were a novel and distinctive breed. Their main orientation was not to the social science disciplines as such; nor were they amateur seekers after a science of legislation in the manner of their mid-nineteenth century predecessors.

In spite of its abstraction and distancing from practical affairs economics, and economists, were not left without influence. On the contrary, the new scientific economics achieved something of an encompassing contextual role with regard to the development in this epoch of a new kind of policy-relevant social knowledge that coexisted with, and drew upon, university-based social science disciplines. Economics set limits to common sense, by putting on a 'scientific' basis the assumption that market outcomes are likely to be the most optimal, at least in the sense of the most efficient. But if economics emerged secure in this role, policy analysis (unlike the policy studies of today grounded primarily in microeconomics) principally adopted the language of public administration and social engineering. This language, nestled within the assumptions of Marshallian economics, facilitated a self-image for policy intellectuals much like that of the earlier liberals of the Social Science Association which stressed the role of social knowledge in helping to transcend class conflict within a framework that accepted, at least implicitly, the divisions between State, economy, and civil society. The central concern of these policy intellectuals was how to build a State capable of dealing with the challenges of increasingly assertive working classes both as economic actors and as poorly integrated citizens. The focus of these efforts pivoted on what came, only later in the 1940s, to be called the Welfare State, to encompass the cluster of policies and institutions concerned both with the mitigation of market distributions and with the substantive content of citizenship. To one side of this new policy knowledge lay the preliberal conservatism of a Bismarck who had sought to make workers a new estate subservient to and dependent on the State; to the other, lay the strong socialist assertion that only the transcendence of markets and the integration of the domains of State, market, and citizen could resolve the contradictions of modern capitalist societies. The new liberal social knowledge decisively rejected these alternatives. Public policy today in the West is encompassed within these boundaries.

5. *Origins as Consequences: A Post World War II and Post Cold War Postscript*

Totalitarianism aside, what is striking is just how encompassing this orientation to public policy proved to be, and how dominant the policies promoted by practitioners of liberal social knowledge had, by the end of World War II, become broadly based on the model established under conditions of extreme economic and geopolitical duress by the American New Deal (Ruggie 1982). With Fascism defeated and Communism the new enemy, this expanded liberalism faced no serious competition in the normal politics of the western democracies. This is not to say that consensus replaced conflict; rather, conflicts about the contours of domestic issues took the form of debates about this or that essentially social liberal alternative, even in situations where some of the most important contestants for power presented more radical alternatives for the long haul. From quite far left to quite far right, political competitors accommodated to liberal principles and institutions. Political contests became no less vigorous, but certainly less fundamental: not capitalism versus socialism or democracy versus totalitarianism (though this antinomy defined the discourse of competition between the Soviet and American blocs), but a politics about rules of transaction between spheres within capitalism and democracy. Coherent repertoires of positions about these questions pivoting on the degree to which the State should be assertive in organizing markets and in utilizing social policy for distributional goals and on how institutions of political participation and the representation of interests should be organized distinguished 'left' from 'right' formulas about State–market and State–citizen transactions. Each position was given voice by its own identifiable policy intellectuals, but together they constituted a single, inherently liberal, knowledge community.

The benefits secured for liberal societies by the robust development of social knowledge by this kind of policy intelligentsia ought not to be gainsaid, especially in light of this century's cruel alternatives. But we should also be aware that by conceptualizing their problems at the level of economic and political market exchanges, most policy intellectuals have left unexamined a host of issues about class and State, inequality and power. By working at the level of transactions between states and markets and states and citizens and by separating these from specialized inquiry by students of international relations, they tacitly have pushed the macrostructures of capitalism and states into the background of their work. In so doing, they have underscored the great normative achievements of liberalism, but also became ac-

complices in the silences deep-seated in the liberal project.

See also: Capitalism; Citizenship and Public Policy; Civil Society, Concept and History of; Democracy; Economic History; Governments; Liberalism; Marshall, Thomas Humphrey (1893–1981); Moral Sentiments in Society; Nations and Nation-states in History; Political History (History of Politics); Political Science: Overview; Property Rights; Public Policy Schools; Rule of Law; Socialism; Sovereignty: Political; Welfare State

Bibliography

Agnew J C 1986 *World's Apart: The Market and the Theatre in Anglo-American Thought, 1550–1750*. Cambridge University Press, Cambridge, UK

Anderson P 1974 *Lineages of the Absolutist State*. New Left Books, London

Bauman Z 1992 'Love in adversity': on the State and intellectuals, and the state of the intellectuals. *Thesis Eleven* **31** (March)

Church R L 1974 Economists as experts: The rise of an academic profession in the United States. In: Stone L (ed.) *The University in Society*. Vol. 2: *Europe, Scotland, and the United States from the 16th to the 20th Century*. Princeton University Press, Princeton, NJ

de Swaan A 1988 *In Care of the State*. Oxford University Press, New York

Downing B M 1992 *The Military Revolution and Political Change: Origins of Democracy and Autocracy in Early Modern Europe*. Princeton University Press, Princeton, NJ

Doyle M 1997 *Ways of War and Peace: Realism, Liberalism, and Socialism*. W. W. Norton, New York

Ertman T 1997 *Birth of the Leviathan: Building States and Regimes in Medieval and Early Modern Europe*. Cambridge University Press, New York

Goldman L 1986 The Social Science Association, 1857–1886. *The English Historical Review* **110** (January)

Harris J 1990 Economic knowledge and British social policy. In: Furner M, Supple B (eds.) *The State and Economic Knowledge*. Cambridge University Press, New York

Hirschman A O 1977 *The Passions and the Interests: Political Arguments for Capitalism Before its Triumph*. Princeton University Press, Princeton, NJ

Hirschman A O 1991 *The Rhetoric of Reaction: Perversity, Futility, Jeopardy*. Harvard University Press, Cambridge, MA

Katznelson I, Zolberg A (eds.) 1986 *Working Class Formation: Nineteenth Century Patterns in Europe and North America*. Princeton University Press, Princeton, NJ

Krieger L 1963 The idea of the welfare state in Europe and the United States. *Journal of the History of Ideas* **24** (October–December)

Lepines W 1988 *Between Literature and Science: The Rise of Sociology*. Cambridge University Press, Cambridge, UK

Mandler P 1987 The making of the new poor law redivivus. *Past and Present* **117** (November)

Mann M 1988 *States, War, and Capitalism*. Blackwell, Oxford

Nelson R 1987 The economics profession and the making on public policy. *Journal of Economic Literature* **25** (March)

North D 1985 Transaction costs in history. *Journal of European Economic History* **14** (Winter)

Oleson A, Voss J (eds.) 1979 *The Organization of Knowledge in Modern America, 1860–1920*. Johns Hopkins University Press, Baltimore

Pitkin H (ed.) 1969 *Representation*. Atherton, New York

Polanyi K 1944 *The Great Transformation: The Political and Economic Origins of Our Time*. Rinehart & Company, New York

Ringer F 1969 *The Decline of the German Mandarins: The German Academic Community, 1890–1933*. Cambridge University Press, Cambridge, UK

Ruggie J 1982 International regimes, transactions, and change: Embedded liberalism in the postwar economic order. *International Organization* **36** (Spring)

Skinner Q 1989 The State. In: Ball T, Farr J, Hanson R L (eds.) *Political Innovation and Conceptual Change*. Cambridge University Press, Cambridge, UK

Soffer R N 1978 *Ethics and Society in England: The Revolution in the Social Sciences, 1870–1914*. University of California Press, Berkeley, CA

Spruyt H 1994 *The Sovereign State and its Competitors*. Princeton University Press, Princeton, NJ

Thomas G M, Meyer J M 1984 The expansion of the state. *Annual Review of Sociology* **10**

Tilly C (ed.) 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ

Tilly C 1990 *Coercion, Capital, and European States*. Blackwell, Oxford, UK

van Gunsteren H 1985 Notes on a theory of citizenship. In: Birnbaum P, Lively J, Parry G (eds.) *1985 Democracy, Consensus, and Social Contract*. Sage Publications, London

Watkins F M 1934 *The State as a Concept of Political Science*. Harper & Brothers, New York

Woloch I 1994 *The New Regime: Transformations of the French Civic Order, 1789–1820s*. W.W. Norton, New York

Zolberg A 1981 Origins of the modern world system: A missing link. *World Politics* **33** (January)

I. Katznelson

Policy History: Regimes

The concept of regime has multiple cognate meanings. Most broadly, it is used as a synonym for 'political system,' connoting the most general institutions and processes characterizing the exercise of political power. More narrowly, when power is so tightly centralized and controlled that the dictator seems to constitute the entire political system, regime is used to refer to the paramount political elite, as in the Franco regime of Spain or the Nasser regime of Egypt. In international relations, regime refers to the set of principles, norms, rules, and decision-making procedures that actors mutually recognize and adhere to in their interactions. The core meaning of regime shared by these usages refers to the formal and informal rules and institutions that both reflect and shape the distribution and organization of political power and that constrain to various degrees the actions of power-holders. In this article, regimes and the

major subtypes of regimes, democracies, or dictatorships, are discussed in the context of state building and with reference to the concepts of state and government. Next, research into the origins of regime types is surveyed, with special reference to the causal role of public policies. Finally, the causal role of regime types on public policy is considered.

1. Regimes and State Building

The topic of regimes assumed increasingly greater importance in the polities of the West in the seventeenth and eighteenth centuries as state building extended the infrastructural power at the disposal of central rulers. Infrastructural power (Mann 1984) is an attribute of states, and refers to the capacity of state agencies to monitor society, regulate social and economic activity, sanction certain forms of collective action while prohibiting other types, adjudicate disputes, and engage in other activities dedicated to extracting human and fiscal resources to reproduce and enhance state power. Despotic power, on the other hand, refers to varying degrees of formal and informal constraints on rulers; as an attribute of regimes, it measures the capacity of rulers to act without consulting and bargaining with other political, economic, and social groups. While in many historical empires despotic power was virtually unlimited, the practical effect of the absence of constraints was acutely mitigated because, weak in infrastructural power, the state lacked the logistic capacity to implement the ruler's whim. State formation in Western Europe threatened to combine high levels of infrastructural power with high levels of despotic power.

European state formation, however, was accompanied by the structural differentiation of the realms of politics, economics, and civil society. Over time, although operating at a variable pace in different political units, the separation between public and private became more and more evident in both theory and practice. While structural differentiation created new opportunities for rulers to extract physical and material resources, rulers increasingly had to bargain and consult not only with traditional political actors and organizations, but also with representatives of new classes and organizations. Extracting resources, moreover, soon imposed new demands for state coordination and regulation to match the increasing complexity of economy and society. Struggles over the content of public policies such as taxation and military conscription became, by the nineteenth century, intertwined with struggles over the form of the regime itself. Over the long term, state power was both massively enhanced and, via changes in the regime, the disposition of state power by rulers was made subject to new constraints and redirected to new aims.

Regimes are composed of two forms of constraints on despotic power. On the one hand, rulers may share the political stage with other collective actors possessing political power, such as parliaments, independent judiciaries, and even legally autonomous executive agencies. On the other hand, and perhaps more importantly, rulers may be compelled to negotiate with actors in civil society, such as capitalists, large landowners, religious organizations, labor unions, and diverse interest groups, who possess resources that can be converted into political power. The distribution of power (the relative strength of actors) and the organization of power (the institutionalized forms of interaction) act as constraints on despotic power and thus constitute regimes.

The concept of regime thus falls midway between those of state and government. State refers to the relatively permanent organizations of territorial-based rule used to generate power via coercion and administration; government refers to incumbents holding offices that deploy that power. Regime refers to the organization of power generated by the state and exercised by rulers and other state officials. It thus defines the routinized relations between rulers and other agents within the political apparatus, on the one hand, and between the state and civil society, on the other hand. Critical components of the regime are the formal and informal mechanisms for selecting leaders, procedures for representation and consultation, and rules by which decisions are made.

2. Types of Regimes

The semantic field of regime types is not timeless, but has instead been composed of those types considered legitimate contenders in distinct historical epochs. Thus, in ancient Greece, political theorists judged the best regime from among democracy, oligarchy, and monarchy. In post-feudal Europe, the main alternatives were absolutist versus constitutional regimes, distinguished by whether monarchs exercised both executive and legislative powers or shared legislative powers with an assembly. During the Cold War, regimes were trichotomized into democratic, authoritarian, and totalitarian regimes. With the collapse of the Soviet Union and the transition to democracy made by both authoritarian and former communist countries, current research treats almost exclusively democratic and authoritarian regimes.

Democracy is an example of what philosophers call an 'essentially contested concept,' a concept that straddles the divide between empirical and normative criteria. These normative criteria are evident in the proliferation of subtypes of democracy characterized by rival contentions about the proper goals of democracy, ranging from the defense of individual rights and freedoms of liberal democracy to the extension of rights to the economic and social realms in social democracy (Gutman 1993).

Wishing to avoid the complex debates surrounding this array of goals to be pursued by a regime in order for it to be called a democracy, most scholars engaged in empirical work on political regimes in the post-World War Two period have subscribed to a strictly institutional definition of democracy. In an early and still influential statement, Joseph Schumpeter (1942, p. 269) identified close links between democracy and capitalism, and defined the former as an 'institutional arrangement for arriving at political decisions in which individuals acquire the power to decide by means of a competitive struggle for the people's vote.' A minimal procedural definition, however, runs the risk of committing the 'electoral fallacy:' the failure to recognize situations in which civilian political leaders gain access to office via popular mandate but do not truly represent the electorate; because they win elections via corruption and manipulation of the political process; because they do not exercise *de facto* control over non-elected military officers and state officials; or because civil society does not have the informational and organizational capacity to monitor government actions and express dissatisfaction with policies. In any of these scenarios, elections do not produce genuine accountability of elected officials to the electorate. In many circumstances, then, it may be difficult to distinguish analytically between the institutional procedures of democracy and the goals democratic practices are designed to achieve. Indeed, insofar as both politicians and the electorate understand elections as an opportunity to exchange loyalty for material benefits and readily assent to these practices, then elections may be understood as opportunities to express mutuality and reciprocal solidarity, not as a means of holding governments accountable. In this scenario, votes may truly reflect genuine preferences of the electorate yet, given the cultural construction of the meaning of elections, it is not clear whether to label this regime democratic (Schaffer 1998).

One way to bridge the institutional and the goal-oriented approaches is to return to the definition of regime and consider democracy as a constraint on power. Modern democracies emerged in national political systems that were products of centuries of state formation, producing far greater infrastructural power than any political systems in the past. Barring constraints on despotic power, rulers of modern states would be capable of achieving their preferences by directly imposing them on all other actors expressing rival preferences, by firmly controlling the processes by which decisions are made to pre-empt the expression of rival preferences, or even to shape the consciousness and preferences of less powerful actors. Consequently, unconstrained rulers with infrastructural power at their disposal would be relatively certain of prevailing in conflicts with other power-holders or actors in civil society. Moreover, insofar as the structures of modern states are relatively durable, unconstrained rulers would have relatively permanent power. Finally, in the absence of constraints, power could be deployed in any and all realms of society, polity, and economy.

Movements for democracy were efforts to contain the arbitrary exercise of power and to redirect its use toward new goals by making governments accountable to and representative of an electorate whose membership was enlarged in the nineteenth century and swelled in the twentieth century. These goals were achieved by placing constitutional limitations on the actions of governments, establishing rights of citizenship, and licensing citizens to participate in the formation of governments and the formulation of policies. As a result, despotic power was sharply constrained: outcomes of conflicts over access to office or policies enacted while in office became increasingly uncertain, and any advantage enjoyed by political leaders became both increasingly temporary and subject to constitutional limitations. The basic principles of democratic regimes, in other words, are that power can be exercised in only a delimited realm of areas; the outcomes of all political conflicts are uncertain; and, therefore, governments are temporary (Przeworski 1991, Lawson 1993).

Making these principles of uncertain, temporary, and restricted exercise of power concrete requires a range of specific institutional arrangements that permit civil society to express preferences and monitor government adherence to them, and that permit an opposition to challenge governments in power with a genuine possibility of becoming the government. Eight institutional criteria are necessary for citizens to have the opportunity to formulate preferences, communicate them to governing authorities, and have those preferences considered in the conduct of government: freedom to join and form organizations; freedom of expression; the right to vote; the right to hold office; the right of political leaders to compete for support; alternative sources of information; free and fair elections; and institutions for making government policies depend on votes and other expressions of preference (Dahl 1971).

By focusing on the institutional arrangements that make the exercise of power relatively restricted, uncertain, and temporary, this definition treats all nondemocratic regimes as a homogenous residual category. In fact, via reference to more specific attributes of the distribution and organization of power, a number of subtypes of authoritarian regimes can be identified, including but not limited to autocratic, totalitarian, absolutist, traditional, monarchic, oligarchic, and sultanistic regimes. By the same token, democratic regimes can—and for many purposes should—be disaggregated into numerous subtypes. Examples of 'modified' democracy in which some features of full democracy are absent include 'controlled' democracy in which parties representing some segments of the political spectrum are banned from competition, and illiberal democracy in which civil liberties are incompletely recognized.

3. Origins of Regime Types

The major cleavage in theoretical debates over the origins of democracy and dictatorship divides structuralist accounts that assign primary causal weight to situational factors from agency-based accounts emphasizing the role of political actors who are relatively unconstrained by environmental factors. Structural explanations can in turn be subdivided into nomothetic accounts positing quasi-universal trajectories of socioeconomic change affecting regime types and historically specific explanations. Among the latter, of particular importance are theories that view regime types as functions of particular constellations of public policies.

3.1 Universal Preconditions

Building on ideas that originated in ancient Greek political thought and that can be traced through the Enlightenment, the most influential theories of regime change in the post-World War Two period have been those that understand economic development and cultural change to be important, if not absolutely necessary, influences on democratization. Studies relying on increasingly refined conceptualizations, reliable data bases, and sophisticated statistical techniques have largely replicated the findings of earlier, relatively crude empirical studies: a strong relationship exists between the level of socioeconomic development—essentially, increasing wealth accompanied by multiple correlates—and democracy. A number of scholars have suggested that this relationship verges on the status of a sociological law.

Several analytic defects, however, motivate scholars to propose alternative theoretical positions. First, questions remain about the causal mechanisms underpinning the statistical relationship between prosperity and democracy. One of the earliest statements of this hypothesis briefly alluded to a number of causal pathways by which increased wealth would promote the spread of values of moderation and toleration, thereby immunizing polities from extremist politics (Lipset 1959). The articulation of increasingly sophisticated statistical studies linking socioeconomic development to democracy has not been accompanied by commensurate analysis of the intervening causal mechanisms. There is yet no systematic evidence that economic development is causally related to democracy primarily via changes in values. Consequently, many studies of political culture and democracy reject the claim that pro-democratic changes in values are necessarily preceded by socioeconomic development. Instead, the cluster of attitudes seen as conducive to democracy has been attributed to political learning, political socialization, institutional change, international diffusion effects, and habits learned from the routine practices of politics itself (Diamond 1993). Even these sources of democratic norms of moderation and tolerance, once decoupled from economic development, have been only contingently tied to democratization.

A second problem stems from the ambiguous specification of the causal relationship between wealth and democracy: rising wealth has been hypothesized both as helping to sustain democracy, but also as helping to generate democracy. Statistical tests have frequently conflated the two issues, despite warnings by prominent scholars to maintain the distinction and treat the two issues separately. When statistical tests are designed to discover the exact nature of the empirical relationship, the best evidence is that heterogeneous factors account for the origins of democracies; once established, rising levels of wealth help to stabilize democracy without causing it (Przeworksi and Limongi 1997).

Finally, in response to democratic breakdowns in the 1960s and 1970s, scholars criticized the universality of the claim, restricting its validity to the earlier developers in Europe and Anglo-American countries, and contending that modernization in later developers would issue in diverse trajectories of political change. In one variant of this position, scholars argued that economic modernization in the twentieth century might actually require a set of public policies such as capital accumulation and investment in heavy industries that could only be implemented by authoritarian political systems; developing countries faced the terrible dilemma of a trade-off between prosperity and political liberty. Alternatively, socioeconomic modernization disrupted people from old patterns of culture and social interaction, and raised their expectations of increasing material gain. When these expectations could not be met, and in the absence of robust political institutions such as mass parties, political instability and military rule typically followed.

Thus, by the mid-1970s, the prospects for new democracies were considered highly unfavorable. Either economic development would be associated with authoritarian governments or, if one maintained loyalty to the modernization hypothesis, given the global recession of mid-decade and the debt crisis that spread through the developing world in the early 1980s, democracy-inducing development was not on the horizon. Political scientists were thus caught largely by surprise by the wave of democratization that began in Southern Europe in the mid-1970s and spread, over the next two decades, to Latin America, Eastern Europe, Asia, and Africa.

3.2 Process-based Theories

Reacting to the predictive failures of both modernization theory and its critical rivals, scholars articulated

a model of regime change that downplayed the role of sociological, cultural, and economic conditions and highlighted the importance of relatively indeterminate transition processes. These transition processes are indeterminate because imperfect information, malleable interests and identities, rapidly changing strategic calculations, and a strong role for contingent events combine in various permutations to preclude structuralist accounts of regime change. In other words, during periods of regime transitions, when previous rules and institutions have been suspended, actors typically find that their identities and interests are subject to change; consequently, it becomes extremely difficult to devise coherent strategies driving goal-oriented behavior; and events that might previously have been considered routine may take on additional significance, forcing actors to recalculate their strategies and their goals. All of these factors do not aggregate into a single model or theory of regime change; rather, they all suggest why the constraints exercised by structural conditions will be loosened, opening up considerable space for politicians to exercise initiative and craft novel outcomes that might have seemed otherwise infeasible. Just as there are no inexorable preconditions for democracy, seemingly unavoidable democratic breakdowns in the past could have been avoided by only incremental changes in actors' perceptions that would have permitted a greater degree of trust and collaboration on behalf of shared goals. Finally, to the extent that relevant actors cannot predict outcomes, normal social science methodology oriented to discovering law-like behavior will be inappropriate to the analysis of democratic transitions (O'Donnell and Schmitter 1986).

The focus on actors and their strategic interactions has undoubtedly been a useful corrective to analyses that focused exclusively on environmental factors. But actor-oriented models may have gone too far in their rejection of any meaningful role for structural conditions, prompting critics to charge this literature with excessive voluntarism. Reviewers of this literature have regularly claimed that its authors too readily dismiss the importance of a range of structural factors other than the level of economic development. Moreover, they point out, it has proven difficult for transitions scholars completely to excise structural factors that regularly appear, either in explanations of individual cases or in the acknowledgement of the 'long-run' causal impact of macro-level factors. Methodologically, these analyses have been overwhelmingly based on countries that have successfully democratized. Persistent authoritarian regimes have not been systematically incorporated into the analysis of regime change. This bias in the selection of cases and the failure to consider the full range of structural factors raises the strong possibility that those cases that democratized might all share one or more structural factors distinguishing them from those cases that have not democratized.

3.3 Macro-historical Analyses

It seems highly likely, therefore, that subsequent research will continue to draw on a third approach to regime change: macro-historical studies. These analyses occupy a middle position between the universalism of modernization theory and the particularity of transitions studies by wagering that a relatively delimited set of historically specific constellations of structural factors can be identified that will account for regime change in clusters of cases. In contrast to the emphasis on values and cultural change of modernization theory, the causal mechanisms adduced by these approaches are typically rooted in the conflictual relations of rational actors pursuing their interests.

A seminal work in this approach is Barrington Moore Jr.'s (1966) analysis of the divergent regime outcomes associated with European economic modernization. For Moore, various relations between the crown, landed aristocracies, and urban bourgeoisies at the time of the commercialization of agriculture determined whether modern regimes would be democratic or authoritarian. For democracy, a strong bourgeoisie had to be the dominant partner of landed elites in a coalition that sought to check royal power. When bourgeoisies were not available to play this leading role, divergent relations between landed elites and peasantries determined whether twentieth-century authoritarian regimes would be fascist or communist. Recent theorists follow Moore's suggestion that the relative strength of classes determines regime outcomes, but propose that the key to democracy lies in the weakening of antidemocratic landed classes and the strengthening of pro-democratic urban working classes (Rueschemeyer et al. 1992).

A second approach focuses less on relative class power and more on the consequences of divergent patterns of interclass coalitions. In this approach, groups embroiled in political-economic conflicts seek allies from among urban workers or rural cultivators, or both, as a means of vanquishing rivals. Regime outcomes are by-products of the political bargains sealing these class alliances (Luebbert 1991).

Another macro-historical approach links regime outcomes to patterns of state formation. As European rulers sought to increase tax burdens and to recruit manpower from their populations, they were forced to bargain with other power-holders who controlled access to those resources: landed elites and urban capitalists (Tilly 1990). Those bargains had a powerful influence on the structure of ensuing states and on the nature of the political regime. The availability of liquid capital in countries such as England and Holland, for example, tended to produce less bulky state apparatuses and more restrictions on despotic power, as holders of moveable assets otherwise beyond the reach of rulers were able to secure some control over public policy and more transparent and accountable administrative procedures in exchange for financing war-

making and state building. Indeed, in these instances, many of the limitations on despotic power established in the late medieval period such as parliaments, independent judiciaries, rudimentary rights, and the rule of law persisted into the early modern period and beyond to the twentieth century (Downing 1992). In contrast, in regions where urban capitalists were few in number and limited in wealth, rulers tended to bargain with landed elites. In these cases, medieval limitations on despotic power were abolished, absolutist regimes were established, and the prospects for stable democracy in the twentieth century were diminished.

These state-building patterns have not been reproduced outside of Europe. Instead, a repeated theme of non-European political development has been the deviation from European patterns due precisely to the existence of wealthier European states that financed non-European state formation, precluding rulers in the periphery from the need to bargain with their own populations. Assuming some conditional relationship between taxation and representation, the reliance of states on external sources of finance such as flows of foreign aid or sale of commodities by state-owned companies forms a potent obstacle to democracy. Moreover, the sale of advanced military hardware and the training of officer corps in both external and internal warfare gives third world states a powerful coercive edge over their civil societies and often gives military officers an edge over civilian officials. Finally, in times of crisis, foreign interventions have and undoubtedly will continue to shape regime outcomes in the non-European world.

At their best, macro-historical analyses are equally attentive to historical particularity, large-scale global processes, and the linkages between situational variables and individual action. But the search for clusters of variables explaining groups of cases can easily be replaced by purely idiographic explanations of individual cases, especially as the wealth of plausible hypotheses and of partially confirmatory evidence renders difficult the extension of any one model to multiple cases in different world-historical settings. Moreover the link between macro-structural variables and individual action is more often imputed than carefully theorized and researched. Finally, in light of efforts to capture at least some particularity of individual cases, there has perhaps been insufficient attention to hypotheses motivating more universalistic theories, such as the development of capitalism. It has been noted with increasing concern, for example, that recently created democracies in Latin America and elsewhere may be grafts of formal institutional structures on to deeply antidemocratic behavioral patterns. To what extent can these variants of democracy be explained by late development and global capitalism? Scholars seeking middle-range hypotheses have underemphasized this sort of research, the provenance of large-n statistical studies. Future research may have to give even more weight to both nomological theorems and the specification and verification of macro-micro links.

4. *Regime Types and Public Policy*

Political scientists have frequently invoked regime types as an influential cause of public policies, particularly in the study of the political determinants of economic policies. In the 1960s, scholars openly wondered whether economic development in the postcolonial world would require authoritarian governments. By the mid-1970s, the end of the postwar growth period and manifestations of domestic political instability prompted some theorists to suggest the onset of a crisis of democratic governability in the advanced industrial nations as well.

In the short term, the hypothesized implication of democratic regimes is the 'political business cycle.' According to this theory, instrumentally rational political leaders who seek above all else to win elections will use their power to alter the economy by reducing unemployment in the period leading up to elections. After being rewarded by voters with re-election, leaders will impose austerity policies to eliminate the inflation induced by pre-election policies. Under democracy, then, economic policy would always be held hostage to the goals of egoistic politicians, resulting in regular macroeconomic instability that follows a 'stop-and-go' pattern. Over the longer term, the hypothesized consequences of democratic regimes were even more alarming. In the well-known analysis of Mancur Olson (1982), special-interest groups formed, under most circumstances, distributive coalitions that sought to increase the economic welfare of their members, typically via pushing for policies that were highly detrimental to overall growth, such as protection or halting the spread of labor-replacing technology. Under conditions of liberal democracy, there were no impediments to the growth of distributive coalitions other than collective action problems within each organization, problems that were likely to be overcome with sufficient time. Consequently, the fastest growing of the advanced economies in the postwar period were those like Germany and Japan which had experienced a fascist interlude that had eliminated most special-interest groups, granting them a period of rapid growth before the inevitable decline set in. The basic logic of these arguments was applied to the developing world as well. Given a trade-off between consumption and investment, democratic governments were seen as more likely to engage in vote-buying behavior and thus less likely to impose policies that delayed consumption on behalf of capital accumulation and long-term growth.

More recent research and theorizing have cast considerable doubt on these arguments. While data

demonstrates a political business cycle in some countries, that evidence is equivocal. When larger samples of countries are studied, the evidence for a business cycle is very weak. One reason that no such cycle can be shown to be consistently evident is that voters, while certainly influenced by economic matters, are best understood as considering both retrospective and prospective components of the economy, effectively canceling out the efforts of politicians to manage the economy for electoral gains (Lewis-Beck 1988). Arguments about the long-term effects of democratic regimes on economic performance also ignored a host of institutional factors that have considerable impact on economic policy-making and performance. States with democratic regimes, for example, display considerable diversity in their structures of interest aggregation and articulation: they thus offer considerably different prospects for particularistic interests to gain access to policy makers and correspondingly different opportunities for policy makers to act autonomously. More importantly, economic performance is closely linked to the specific features of non-regime institutions. The structure of financial system intermediation and of the trade union movement, for example, have powerful effects on growth, inflation, and unemployment rates as well as on the capacity to adjust successfully to international economic change. Broadly similar arguments about the developing world attribute level of development more to specific institutions such as bureaucratic structures and financial systems than to regime types. Indeed, numerous studies have demonstrated that even authoritarian regimes frequently sacrifice long-term investment to short-term political calculations, and that democratic regimes often have the legitimacy and popular support needed to enact painful reforms.

See also: Democracy; Democratic Transitions; Development and the State; Development, Economics of; Dictatorship; Globalization: Political Aspects; Leadership: Political; Modernization, Political: Alternate Paths; Modernization, Political: Development of the Concept; Nations and Nation-states in History; Parliamentary Government; Policy History: Origins; Policy History: State and Economy; Power: Political; Presidential Government; Sovereignty: Political; State Formation; State, History of; State, Sociology of the

Bibliography

Cheibub J A 1998 Political regimes and the extractive capacity of governments: Taxation in democracies and dictatorships. *World Politics* **50**: 349–76

Dahl R 1971 *Polyarchy: Participation and Opposition*. Yale University Press, New Haven, CT

Diamond L 1992 Economic development and democracy reconsidered. *American Behavioral Scientist* **35**: 450–99

Diamond L 1993 Political culture and democracy. In: Diamond L (ed.) *Political Culture and Democracy in Developing Countries*. Lynne Rienner Publishers, Boulder, CO

Downing B 1992 *The Military Revolution and Political Change: Origins of Democracy and Autocracy in Early Modern Europe*. Princeton University Press, Princeton, NJ

Gutman A 1993 Democracy. In: Goodin R, Pettit P (eds.) *A Companion to Contemporary Political Philosophy*. Basil Blackwell, Oxford, UK

Karl T L, Schmitter P C 1993 What democracy is … and is not. In: Diamond L, Plattner M F (eds.) *The Global Resurgence of Democracy*. Johns Hopkins University Press, Baltimore, MD

Lawson S 1993 Conceptual issues in the comparative study of regime change and democratization. *Comparative Politics* **25**: 183–205

Lewis-Beck M S 1988 *Economics and Elections: The Major Western Democracies*. University of Michigan, Ann Arbor, MI

Lindenberg M, Devarajan S 1993 Prescribing strong economic medicine: Revisiting the myths about structural adjustment, democracy, and economic performance in developing countries. *Comparative Politics* **25**: 169–82

Lipset S M 1959 Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review* **53**: 69–105

Londregan J B, Poole K T 1996 Does high income promote democracy? *World Politics* **49**: 1–30

Luebbert G M 1991 *Liberalism, Fascism, or Social Democracy: Social Classes and the Political Origins of Regimes in Interwar Europe*. Oxford University Press, New York

Mann M 1984 The autonomous power of the state: Its origins, mechanisms and results. *Archives Européennes de Sociologie* **25**: 185–213

Moore B 1966 *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon, Boston, MA

O'Donnell G 1974 *Modernization and Bureaucratic-Authoritarianism*. University of California, Berkeley, CA

O'Donnell G, Schmitter P C 1986 *Transitions from Authoritarian Rule: Tentative Conclusions about Uncertain Democracies*. Johns Hopkins University Press, Baltimore, MD

Olson M 1982 *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. Yale University Press, New Haven, CT

Przeworski A 1991 *Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America*. Cambridge University Press, Cambridge, UK

Przeworski A, Limongi F 1997 Modernization: Theories and facts. *World Politics* **49**: 155–83

Rueschemeyer D, Stephens E H, Stephens J D 1992 *Capitalist Development and Democracy*. University of Chicago Press, Chicago

Schaffer F C 1998 *Democracy in Translation: Understanding Politics in an Unfamiliar Culture*. Cornell University Press, Ithaca, NY

Schumpeter J A 1942 *Capitalism, Socialism and Democracy*. Harper Torchbooks, New York

Tilly C 1990 *Coercion, Capital, and European States, AD 990–1990*. Basil Blackwell, Oxford, UK

D. Waldner

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Policy History: State and Economy

The appropriate relationship between state and economy has been one of the central faultlines of political debate, policy conflicts, and scholarly debates for centuries. Scholars are no closer to reaching consensus on these issues at the beginning of the twenty-first century than they were when Adam Smith first published *The Wealth of Nations* (1776). Yet despite the venerable nature of this faultline, there have been remarkably few efforts to provide a history of the centuries old debate on the most desirable relationship between state and economy. This article represents a preliminary effort to sketch out a brief history of these policy debates, concentrating on the nineteenth and twentieth centuries. The critical ingredients of this history are doctrines, global constraints, and theoretical confusions.

1. Doctrines

The terms of debate on the relationship between state and economy over the last two centuries is radically different from what came before. For centuries, European governments used a variety of mercantilist practices to enhance the economic growth of their territories by setting up barriers against imports, investing in education and technology, providing concessions and other support to promising enterprises, and establishing relief policies to protect the poor from economic hardship (Heckscher 1934, Reinert 1999). There is disagreement over when and how effectively these mercantilist practices were codified into a coherent body of economic doctrine, but there is consensus that for centuries, the use of governmental power to enhance economic development was seen as legitimate and desirable.

The classical economists, particularly Ricardo and Malthus, fundamentally changed the terms of discussion by conceptualizing the economy as an autonomous entity that is governed by its own principles, such as the law of supply and demand. Both Polanyi (1957) and Rothschild (1992) stress the important discontinuities between Adam Smith and the later writers of the classical school. Writing in the shadow of Newton's scientific achievements, they imagined that the workings of the economy were similar to the harmonious movements of the planets in the solar system that could now be fully explained by the laws of physics (Redman 1997). In searching for the economic equivalents of the laws of physics, the classical economists generally viewed the state as an 'outside' force whose actions were likely to interfere with the economy's capacities for self-regulation. In fact, their arguments against various mercantilist policies often emphasized that government actions were likely to have perverse consequences—opposite to those initially intended—precisely because such actions would disrupt the natural and self-regulating operation of economic laws (Hirschman 1991).

The classical economists recognized that there was a fundamental tension between their abstract model of an economy and the real world in which human beings were organized into separate nations governed by states. Just as their mercantilist predecessors, they were writing in order to shape the policies of their own government; they recognized that their vision of the proper organization of the economy could only be implemented through the use of the state. Hence, they distinguished between appropriate government policies—those that supported the full emergence of the economy as autonomous—and those government policies that were inappropriate—that were likely to retard or interfere with an autonomous economy. But while the classical economists agreed on this distinction, they were unable either to define precisely where this dividing line should be drawn or to develop a theoretically sound justification for its location. In fact, they had profound disagreements among themselves on where the line should be drawn. Despite the centrality of free trade to the classical tradition, one of its key thinkers—Malthus—argued consistently that it was appropriate for England to retain the Corn Laws that limited agricultural imports. John Stuart Mill, who played a critical role in codifying classical economics, also entertained a variety of heterodox policy ideas (Schumpeter 1954).

1.1 Purists and Pragmatists

This inability to derive the proper functions of the state from their abstract model of the economy has characterized the mainstream tradition of economic thought through the entire nineteenth and twentieth centuries (for accounts of this tradition that are less Anglo-American centered than most, see Schumpeter (1954) and Screpanti and Zamagni (1993)). The result has been a continuous tension within the tradition between 'purists' and 'pragmatists.' The purists argue that since we know that expanding the scope of markets improves economic efficiency, it follows that international free trade is always preferable to any form of protectionism. Furthermore, since governments can subvert free trade by providing subsidies to their own producers or by using regulations to discourage imports, it follows that a domestic regime of *laissez faire* in which government exercises the smallest possible role in the economy is to be preferred. Pragmatists, in contrast, starting with John Stuart Mill, have argued that precisely because the real world of actual economies diverges from the abstract model, there is an affirmative role of government that extends beyond establishing the minimal conditions for private contracting. Pragmatists have differed among themselves as to what government actions they would endorse, but some have favored tariff protection of

infant industries, extensive regulatory regimes to overcome market failures, substantial measures to redistribute income, and the provision of a wide variety of public goods, including government participation in planning of the economy's future growth path (Hutchison 1978 is particularly valuable in tracing out the pragmatism of Jevons, Marshall, and Pigou).

But this longstanding argument between purists and pragmatists within the central tradition of economics has usually been overshadowed by conflicts generated by two other important sets of ideas. The first is the challenge to the market model posed by socialist ideas, particularly those of the Marxist tradition. The second is the emergence of a counter tradition in economics that rejected both the methodology and the free trade preferences of the classical tradition.

1.2 Socialism

The socialist tradition predates Marx and includes a range of different theoretical currents, but the Marxist current has been the strongest for almost a century and a half. Moreover, it was Marx who effectively defined the intellectual achievement of the classical economists; he argued that they were the theorists and apologists of the emerging bourgeois social order whose triumph was signaled by the French Revolution. For Marx, this new social order was a coherent system that revolved around the imperatives of profit maximization. Precisely because the capitalist mode of production was a unified system, efforts to reform it were fundamentally futile. Marx repeatedly insisted that the only way for the working class to advance its collective interests was by a revolution that replaced capitalism with a radically different social order in which markets, private property, and the pursuit of profit would be eliminated.

By the end of the nineteenth century, the Marxist tradition developed its own division between purists and pragmatists. Purists continued to oppose any efforts to reform capitalism as inherently futile since they viewed it as impossible to change capitalism's fundamental logic. Pragmatists argued that meaningful reforms could be won that would strengthen the economic and political position of the working class and make possible further advances in the direction of socialism. Since the pragmatist position often resonated with the views of non-Marxist socialists and with the views of reformists in the trade union movement, it played an important part in expanding the state's role in the provision of social insurance, the regulation of industry, and the redistribution of income (Sassoon 1996, Hicks 1999).

While purists in the Marxist tradition and purist defenders of the free market have been bitter antagonists, it is worth stressing that they share certain common assumptions. They both insist that the only real alternatives for organizing societies are free market capitalism or socialism. They both see any effort to mix socialist and capitalist principles within a single society to be inherently unstable; sooner or later the instability will be resolved by movement to one of the stable poles. Moreover, they have the same view of the globe—that a world divided between socialist and capitalist societies is unstable; the world ultimately will have to be organized by only one of these principles.

1.3 The Counter Tradition

Marxism, however, was not the only oppositional doctrine that was spawned by the theoretical achievements of the classical economists. Starting as early as the 1820s, a counter tradition emerged that challenged both the classical economist's emphasis on the desirability of free trade and their methodological assumptions. But while important work has been done in this tradition in virtually every decade from 1820 to 2000, the identification of this counter tradition is largely retrospective (Reinert 1999). It does not have either the organizational or the doctrinal continuity of either mainstream economics or Marxism. Its founding figure was the German, Friedrich List, who argued that the policy proposals of the classical economists were not universally applicable; if poorer nations were to embrace free trade without qualification, they would never be able to catch up with the world's leading economic power—England. List advocated the use of governmental powers, including tariffs to protect infant industries, as a means to accelerate development of the economy's 'productive forces' (Henderson 1983). The American economist, Henry Carey, also made important early contributions to this counter tradition.

The counter tradition was further elaborated in the 1870s and 1880s by the younger generation of the German historical school, especially in the writings of Gustav von Schmoller. Schmoller rejected both the 'classical tradition' view of the economy as autonomous and their *laissez faire* policy prescriptions (Schumpeter 1954). The German historical school, in turn, inspired both the first generation of academic economists in the United States (Ross 1991) and a group of English historical economists, including William Cunningham, J. W. Ashley, and W. A. S. Hewins who revisited English economic history and found powerful evidence that English economic development had depended heavily on state policies (Koot 1987). In the first part of the twentieth century, further contributions to the counter tradition were made by the American institutionalists and the English Fabians. In the 1930s, despite his orthodox intellectual lineage, John Maynard Keynes joined this counter tradition by breaking with free trade and advocating active government management of the economy. From

the 1930s through the 1970s, this counter tradition was enriched by a broad range of scholars including Japanese economists who are little known in the West (Gao 1997), the Latin American group around Raul Prebisch that developed the idea of import substituting industrialization (McMichael 1996), and others. A particularly important contribution was the study of late-industrializing countries by Gerschenkron (1962).

Political leaders in late-industrializing countries have turned repeatedly to theorists of this counter tradition to help develop policies to accelerate their nation's development. Count Witte, the Russian modernizing minister under Tsar Nicholas II had studied List closely before taking office and followed many of List's policy prescriptions (Henderson 1983, pp. 215–17). After the Meijii Restoration in Japan, the most influential economic work was that of the German historical school. As Gao writes, 'The intellectual roots of Japanese developmentalism can be traced to the German historical school' (Gao 1997, p. 60). More recently, the spectacular economic advances of the South Korean and Taiwanese economies have been attributed to the influence of developmentalist policies inspired by the counter tradition (Wade 1990, see also the important book by Evans 1995).

The boundary lines separating these various currents of economic doctrine have sometimes been vague and indistinct. In the twentieth century, the dividing line between pragmatists within the mainstream tradition and proponents of the counter tradition has been crossed by major figures. (There are also important figures whose work spans the boundary line between Marxism and other traditions; of particular interest are those who have contributed to the theory of market socialism.) Two of the greatest economists of the twentieth century—Keynes and Schumpeter—can legitimately be claimed by both the mainstream and the counter tradition. A further complication is that the mainstream tradition of Anglo-American economics whose successive paradigms were codified by J. S. Mill, Marshall, and Keynes, entered into crisis in the early 1970s. As the influence of Keynesian ideas declined, no single theory has become dominant and the tradition has splintered into a number of conflicting views (Heilbroner and Milbert 1995). This is the theoretical context in which the political influence of the purist wing of the tradition has increased dramatically in the USA and throughout the world.

2. Global Constraints

It would be a mistake to imagine that political leaders have been free to choose among these diverse doctrines in deciding what role the state should play in their particular economy. On the contrary, these diverse doctrines have existed within a global structure of power that has constrained and limited the choices of political leaders. But this global structure of power has itself been in continuous flux, so that the severity of the constraints has varied considerably over time. Moreover, the constraints also vary within different historical periods; some countries have had more freedom to maneuver because of the specifics of their location.

The constraints are a direct result of successive national hegemonies over what Wallerstein has termed 'the modern capitalist world-system' (Wallerstein 1974, Arrighi and Silver 1999). Dutch hegemony was established initially in the seventeenth century, but it gave way, after an interregnum, to English hegemony in the early years of the nineteenth century. England's hegemony was undermined first by World War I and then by the global economic collapse of the early 1930s. After another interregnum, the US emerged as the new hegemonic power in the final years of World War II. Moreover, the geographical scope of this capitalist world-system has steadily expanded from its European core so that by the middle of the twentieth century, it encompassed the entire globe.

Hegemony represents a combination of economic, military, and cultural power. In the nineteenth century, for example, the new ideas of the classical economists were a constitutive part of the English-dominated global order. Other nations were under pressure to adopt the ideas of free trade and *laissez faire* as a way to gain access to English capital, to avoid England's military wrath, and to partake in the industrial modernity that England exemplified. The same point can be said of US hegemony in the second half of the twentieth century; it is impossible to determine whether the influence of US-sponsored economic doctrines has more to do with financial power, military power, or the powerful imagery of Hollywood and Madison Avenue.

2.1 A Rough Periodization

Even in England, there was a considerable time lag between the writings of the classical economists and the full embrace of free trade. It was only with the repeal of the Corn Laws in 1846 that England began exerting considerable pressure on other nations to embrace the doctrines of free trade and *laissez faire*. Hence, the period from 1850 to 1875 can be thought of as the first great epoch of free trade. During this period, most European nations lowered their tariff barriers and made at least some efforts to embrace *laissez faire* (Chase-Dunn et al. 2000). By 1870, many nations had embraced the gold standard by making their currencies freely convertible into gold (Eichengreen 1996). The gold standard represented the fullest expression of English hegemony, since it was designed to equilibrate automatically the global economy without governmental interference.

The dynamic shifted sharply in the next period from about 1875 to 1900. This was a period of slower

growth and declining prices that in combination with the gold standard placed severe pressures on national economies (Polanyi 1957). The result was a marked retreat from both free trade and *laissez faire*. Tariff barriers increased significantly in this era (Chase-Dunn et al. 2000), and rising protectionism was clearly a component of the rush for empire as the European powers, the US, and Japan moved to establish formal colonies (Polanyi 1957). The trade barriers meant the potential loss of access to both markets and raw materials in those regions that had been colonized by other powers. Moreover, national regulations to control market forces expanded dramatically; this was the era in which government-sponsored social insurance was first developed and began to spread.

The brief period from 1900 until the outbreak of WW I is an anomalous and contradictory era. On the one side, the pendulum swung back in the direction of lowering trade barriers (Chase-Dunn et al. 2000), and international capital movements reached unprecedented levels in this period as the global financial system became more tightly integrated. On the other side, there were few signs of a return to *laissez faire*; the state's role in managing domestic economies and providing social welfare continued to expand in many nations. Governments were under increasing pressure to make concessions to socialist movements that were growing in power and influence. Moreover, in this period, late-industrializing nations, such as Russia and Japan, did not hesitate to use governmental power to accelerate the processes of industrial development. Moreover, the intensification of the military rivalry between England and Germany in this period was very much a continuation of the imperialist conflicts that had emerged in the previous period.

Between 1914 and 1945, nineteenth-century civilization collapsed, and it appeared that the ways of thinking about state and economy developed by the classical economists had become completely obsolete. From Bolshevism to fascism to the experiments of the New Deal, it seemed that there were a multitude of divergent paths for organizing national economies (Polanyi 1957). Moreover, the imperatives of war meant that the major powers, even those that retained democratic institutions, went through extended periods of time when economic life was completely subordinated to political and military control. While much of Asia and Africa remained under colonial control, the larger Latin America nations were able to take advantage of the breakdown of global power to follow the teachings of the counter tradition (McMichael 1996). The overwhelming reality of this period was the human suffering generated by war, dislocation, depression, and unprecedented violence directed at civilians.

During the first period of US global hegemony from 1945 to 1973, the USA struggled, with the help of multilateral institutions, to limit the range of permissible variations in national economic policies. This meant an intensification of the Cold War to restrain the growth of socialism that had spread to Eastern Europe and China. The process of decolonization meant that this battle was now fought out over a far larger group of nations. But while the USA pushed strongly in this period for trade liberalization, there was no effort to restore *laissez faire*. It was taken for granted that, as with the New Deal regime in the USA, the government would play a large role in the domestic economy in managing the business cycle, regulating private firms, and producing a broad array of public goods. In this period, the pragmatists within the mainstream tradition of economics were dominant with their vision of a mixed economy. The ability of governments to pursue macroeconomic objectives was facilitated further by global rules that severely restricted international capital movements. Even poorer countries were expected to develop ambitious development plans and an elaborate system of foreign aid emerged to finance these efforts. Yet these efforts were biased strongly towards a model of development led by the private sector and countries that were too aggressive in fostering state-led development were pressured to stop (Wood 1986). A few nations, such as South Korea and Taiwan, were able to resist these pressures because of their location close to the fault lines of the Cold War, but their experience was deeply atypical (Wade 1990).

Finally, the period from 1973 to 2000 is the second great epoch of free trade. The steady progress towards reduction of trade barriers that began in 1945 was continued, and it was now joined by the elimination of barriers to international capital movement. The integration of global financial markets reached levels that had only previously been attained in the period between 1900 and 1914. Starting in the late 1970s, with Margaret Thatcher and Ronald Reagan, and intensifying after the collapse of the USSR, the old doctrines of the classical economists were now recycled under the rubric of neoliberalism. *Laissez faire* ideas were reintroduced on the grounds that government efforts to manage the business cycle were futile and excessive levels of taxation and regulation were destroying the vitality of market economies. Trade negotiations were now expanded to dismantle 'non-tariff barriers' that included domestic legislation that had the unintended consequence of discriminating against foreign products. Developing nations were told to scrap their planning processes, to privatize state-owned enterprises, and to end subsidies that made food, transportation, and housing more available for the poor (McMichael 1996).

3. *Theoretical Confusions*

In reviving *laissez faire*, neoliberals assume that they can distinguish between the legitimate, albeit limited, responsibilities of government and those illegitimate

government actions that are bound to have perverse consequences. But neoliberals have been no more successful than their classical predecessors in providing a theoretically justified demarcation line. That this problem has not been solved after two centuries of work suggests that there are deep theoretical confusions operating here.

3.1 Can the Economy be Autonomous?

The fundamental problem lies with the assumption, introduced by the classical economists, that an economy can be organized around a set of self-regulating markets. Even with the added stipulation that there will be a state and a legal system that enforces property rights and contracts, this is still not enough. This point was elaborated most forcefully by Polanyi (1957), a Hungarian refugee intellectual who developed his own theoretical synthesis drawing on insights from Marxism and the German historical school. Polanyi argued that the flaw in the belief in an autonomous economy is that land, labor, and money are not true commodities; they are not things that were produced for sale on a market. Economists assume, however, that the price mechanism will equilibrate supply and demand for these fictitious commodities just as it does for true commodities. The difficulty is that when the demand for labor suddenly falls, human beings cannot be warehoused like unsold merchandise; those who are out of work must still have income. Similarly, countries that leave it to the market to determine the supply of money and credit are likely to suffer from chronic deflation when credit is too scarce or chronic inflation when it is too abundant. With labor, money, and land, a continuous governmental role is needed to make the market for these fictitious commodities work like other markets.

In short, a market economy cannot be autonomous because the state is involved intimately in three of the most important markets. What follows from this insight is that capitalism, defined as a society organized around an autonomous market in which actors seek to maximize utility, has never existed. Polanyi went on to argue that the market societies that actually exist are constituted by two opposing tendencies. The first are the ongoing efforts to expand the scope of markets—the project first announced by the classical economists. The second are parallel efforts to protect the social order from the consequences of market competition that includes regulatory efforts, social welfare policies, and other macroeconomic policies. Polanyi emphasizes that business firms participate in both of these tendencies, since they often need to protect themselves from market situations where uncertainty is uncomfortably high.

In recent years, scholars have produced a body of work that elaborates this Polanyian insight by analyzing the existence of a number of varieties of capitalism that vary significantly in the way that key markets are organized and in the relationships between state and economy (Block 1994, Crouch and Streeck 1997, Hollingsworth and Boyer 1997). These studies represent a fundamental challenge to the assumptions of both mainstream economists and Marxists that market societies should converge towards some common pattern. At the same time, however, some commentators (Gray 1998) have expressed concern that the neoliberal rules of the global economy might have the consequence of destroying varieties of capitalism that are actually more efficient than the Anglo-American type.

3.2 Transition to a Borderless World

A second theoretical confusion has to do with the tension between a unified global market and a world of nation states. Both the classical economists and contemporary neoliberals have been motivated by a vision of a peaceful planet in which military rivalries and warfare are replaced by the competition of individuals and firms to reach higher levels of efficiency. Both have imagined a gradual withering away of the nation state as the globe becomes economically unified. While there is much that is appealing in this vision, there has been a lack of clear thinking about the actual staging of this transition to a borderless world.

Ironically, the historical periods in which this vision has been most influential were marked not by the disappearance of state power, but by the unchallenged military and political hegemony of a single power—first England and then the USA (Kindleberger 1973). The temporary dominance of a single power hardly seems a sound foundation for a transition to a world beyond nations. A closely related problem is that this transition requires that nations give up the various measures that buffer their populations from the pressures of the global economy just as those pressures are intensifying. Creating a unified global market for agricultural goods, for example, requires that governments dismantle programs designed to stabilize the income of farmers just as those farmers are becoming more vulnerable to cheaper agricultural imports. The neoliberal vision rests on the willingness of millions of ordinary people—farmers, workers, and small businesspeople to be extraordinarily flexible—to accept that years of reasonable income periodically will be interrupted by years in which their income might fall by as much as 40–60 percent. (These figures reflect income losses in Asia and Mexico in the financial crises of the 1990s.)

For the affluent, such flexibility might seem logical and reasonable; they know that years of high returns on their portfolios can alternate with periods of

negative returns. But those who lack an investment portfolio tend to resist this type of flexibility; their ability to pursue a life plan depends on some stability and predictability in their income flows. When such people are able to exercise political influence on their national leaders, they demand to be protected from extreme instability of their income. This is precisely what happened in the period from 1875 to 1900; ordinary people demanded to be protected from the costs imposed on them by global economic integration, and the result was a significant retreat from global free trade. Moreover, if governments are constrained by global power and are unable to respond to these popular demands to be protected from income instability, discontent does not disappear. It is often channeled into militant movements of the left or of the right that target both the ineffectiveness of the government in power and the foreign enemies who have imposed economic hardship on the local people. The efforts of centrist political leaders in the 1920s and 1930s to restore the gold standard and a liberal world economy were defeated by these types of movements. The deep historical irony is that the effort to create a pacified, borderless world produced exactly the opposite; extreme forms of nationalism and war on a global scale.

4. Conclusion

The period from 1973 to 2000 has seen unprecedented growth in a variety of indices of economic globalization. However, these theoretical confusions about the autonomy of the economy and the transition to a borderless world raise concerns about the sustainability of these trends. How the organization of economies and states will evolve remains an open question.

See also: Development and the State; Development, Economics of; Economics and Ethics; Economics, History of; Economics: Overview; Policy History: Regimes; Political Economy, History of

Bibliography

Arrighi G, Silver B 1999 *Chaos and Governance in the Modern World System*. University of Minnesota Press, Minneapolis, MN

Block F 1994 The roles of the state in the economy. In: Smelser N J, Swedberg R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ

Chase-Dunn C, Kawano Y, Brewer B D 2000 Trade globalization since 1795: Waves of integration in the world-system. *American Sociological Review* **65**: 77–95

Crouch C, Streeck W 1997 *Political Economy of Modern Capitalism: Mapping Convergence and Diversity*. Sage, Thousand Oaks, CA

Eichengreen B 1996 *Globalizing Capital: A History of the International Monetary System*. Princeton University Press, Princeton, NJ

Evans P 1995 *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press, Princeton, NJ

Gao B 1997 *Economic Ideology and Japanese Industrial Policy*. Cambridge University Press, Cambridge, UK

Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Harvard University Press, Cambridge, MA

Gray J 1998 *False Dawn: The Delusions of Global Capitalism*. Granta, London

Heckscher E 1934 *Mercantilism*, 2 vols. George Allen & Unwin, London

Heilbroner R, Milbert W 1995 *The Crisis of Vision in Modern Economic Thought*. Cambridge University Press, New York

Henderson W O 1983 *Friedrich List: Economist and Visionary 1789–1846*. Frank Cass, London

Hicks A 1999 *Social Democracy and Welfare Capitalism*. Cornell University Press, Ithaca, NY

Hirschman A O 1991 *The Rhetoric of Reaction: Perversity, Futility, Jeopardy*. Harvard University Press, Cambridge, MA

Hollingsworth R, Boyer R 1997 *Contemporary Capitalism: The Embeddedness of Institutions*. Cambridge University Press, Cambridge, UK

Hutchison T W 1978 *On Revolutions and Progress in Economic Knowledge*. Cambridge University Press, Cambridge, UK

Kindleberger C P 1973 *The World in Depression 1929–1939*. University of California Press, Berkeley, CA

Koot G M 1987 *English Historical Economics, 1870–1926*. Cambridge University Press, Cambridge, UK

McMichael P 1996 *Development and Social Change: A Global Perspective*. Pine Forge Press, Thousand Oaks, CA

Polanyi K 1944/1957 *The Great Transformation*. Beacon Press, Boston

Redman D A 1997 *The Rise of Political Economy as a Science*. MIT Press, Cambridge, MA

Reinert E 1999 The role of the state in economic growth. *Journal of Economic Studies* **26**: 268–326

Ross D 1991 *The Origins of American Social Science*. Cambridge University Press, Cambridge, UK

Rothschild E 1992 Adam Smith and conservative economics. *Economic History Review* **XLV**: 74–96

Sassoon D 1996 *One Hundred Years of Socialism: The West European Left in the Twentieth Century*. Tauris, London

Schumpeter J A 1954 *History of Economic Analysis*. Oxford University Press, New York

Screpanti E, Zamagni S 1993 *An Outline of the History of Economic Thought*. Clarendon Press, Oxford, UK

Smith A 1776/1976 *The Wealth of Nations*. University of Chicago Press, Chicago

Wade R 1990 *Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization*. Princeton University Press, Princeton, NJ

Wallerstein I 1974 *The Modern World-System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

Wood R 1986 *From Marshall Plan to Debt Crisis: Foreign Aid and Development Choices in the World Economy*. University of California Press, Berkeley, CA

F. Block

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Policy Information

Information is used by governments for a wide variety of policy purposes. It can be used in the formulation of public policies, by helping to identify the need for a public policy, or to select one among alternatives. It can be used in the implementation of policies, to help target or improve programs. It can be used in the evaluation of policies, such as measures of progress toward desired outcomes, or of the intended or unintended social, economic, and environmental effects. Here, however, *policy information* is taken to mean information used in the formulation, design, and selection of public policies. It may also include information from the evaluation of public policies to the extent that effects of the policies inform consideration of alternative policies.

Policy information differs from political information, such as polls of public opinion, or of support for a policy among legislators. Policy information is objective. It may be quantitative or qualitative, but it is based on evidence, rather than opinions, attitudes, or beliefs (see Citro and Hanushek 1991). Policy information comprises both data and analysis. Indeed, it is often the result of *policy analysis*, by which is meant the objective prospective, and sometimes retrospective, evaluation of the social and economic implications of changes in public policy. For example, in considering choices among social welfare policies, decision makers typically want to know what populations would be affected, in what ways, and at what cost. Policy analysis provides answers to these 'what if' questions. In order to inform choices among policy alternatives, policy analysis involves forecasts or projections into the future. As a result, they are more uncertain than evaluations of the past. Three key ingredients of policy analysis are data; analytical models, including their documentation and validation; and effective communication of the results to policy makers, in particular of the assumptions in models and uncertainty in estimates.

1. Data

Accurate and relevant data are critical for policy analysis. Government surveys are a common source (see *Government Statistics*). Since the behavior of individuals in response to decisions of employers or government agencies is often what is being modeled, data may be needed on the interactions of individuals with organizations and institutions. For example, an individual's access to health care services will depend on providers, insurers, and government health care programs and policies. Often, however, data are available only on a single unit, such as an individual, family, or firm. In this situation, analysts must meld together data from different sources.

Useful sources of data to augment survey data are administrative records, such as health care claims, tax returns, earnings reports for social security, and case files from public assistance programs. Such data may need to be linked, or matched, to the survey data. Increased concern in government statistical agencies about possible violations of the confidentiality of survey data, however, militate against such linking. Analysts in this situation must make assumptions, the validity of which the analysts may not be able to determine without the linked data.

As important as the availability of data is, it is not sufficient. Data for public policy analysis must reflect relevant and accurate measurement. For this reason, government statistical agencies need to conduct research on concepts, definitions, and measurement, and assess the quality of their data. In the highly decentralized statistical system in the United States, however, it is not uncommon for agencies to report different measures from different surveys of ostensibly the same concept, such as income, employment status, or disability. Efforts may be made to explain the differences, or a result from one survey may be adjusted, or controlled, to agree with the result from a more prominent survey. Good policy analysis, however, requires integrated databases with the best estimates that can be developed from multiple sources.

Policy analysts, as well as other sophisticated users, cope with the lack of integrated data by performing additional adjustments to make the data suitable for modeling and analysis. But not all users have the information, much less the resources, to do so. And those that do may either duplicate efforts of others or, much worse, wind up with different results.

2. Models

Models are fundamental tools of policy analysis. *Microsimulation models* are one important type of policy analysis tool. With data from individual records, a microsimulation model simulates how the records change over time and as a result of a proposed policy. To predict the effect of a change in tax law, for example, the model processes records for people as if they were filling out their tax forms. As Citro and Hanushek (1991) note:

> Models based on microsimulation techniques are conceptually highly attractive because they operate at the appropriate decision level and take into account the diverse circumstances and characteristics of the relevant population, whether it be low-income families, taxpayers, or health care providers. Such models are able to respond to important needs of the policy process for information about the effects of very fine-grained as well as broader policy changes, the effects of changes that involve complicated interactions among more than one government program, and the effects of changes on specific population groups as well as total program costs and caseloads.

Microsimulation models, however, tend to be highly complex and expensive to develop and maintain. Because of the lack of integrated data, analysts using the models must often patch together a variety of data and research results of varying, if even known, quality. As a result, often many assumptions are made; and some are unsupported. Because of these problems, some analysts may rely on simpler models, akin to accounting methods, to understand the effects of changes in a program. But the value of microsimulation models in informing a policy debate is that they can model behavioral responses to changes in policies and programs.

Many other statistical models are also used in policy analysis, from linear regression models, to simultaneous equation models, to complex macroeconomic ones. Macroeconomic models are also tools for 'what if' questions. They are often used to make unconditional forecasts, such as the rate of growth of the economy, as opposed to conditional projections. In addition to these models, a cornucopia of social science methods are also employed in policy analysis, including cost–benefit studies, risk assessments, meta-analyses, and even social experiments.

While microsimulation models seek to answer 'what if' questions, other models seek to answer 'why' questions, such as why people save, or fail to save, and to provide estimates of relationships that may be used in microsimulation models. Sometimes the research model may not be explicit, as when a database is constructed by piecing together data from different sources, in particular to determine a trend for which there is no survey-based indicator, such as the retention of children in the same grade in school. Sometimes the model may be difficult to ascertain, because political factors must also be taken into consideration; for example, in formulas used in legislation to allocate funds to school districts according to the numbers of children from families with incomes below the poverty level.

3. *Uncertainty*

Even when the model is explicitly clear, however, its uncertainty may be unclear. The estimates from models in policy analysis are conditional forecasts of the effects of a policy change. The effects may be poorly understood. They may be unknown or depend on unknown factors. The model relating cause and effect may be incorrect; or it may be uncertain as to which of many factors contribute to the effects. Moreover, the lack of data is a ubiquitous problem.

The understanding, however, of the effects of a proposed policy is important, if decisions on it are to balance the benefits against the costs. Key to that understanding are reasonable judgments about the quality of estimates of the likely effects of a proposed policy change. Policy makers need to understand the uncertainty or variability in the estimates and to judge the track record of a policy analysis model. Unfortunately, they are rarely given the information to enable them to do so.

3.1 *Model Validation*

Assessing the quality of model estimates is a common problem in statistical research, but the problem is compounded in policy analysis, because it involves conditional forecasts. Some, if not all, of the policy alternatives studied may not be adopted. And, even for alternatives that are, data may never become available on the outcome. Yet it is crucial to understand the uncertainty in model estimates, if only to evaluate the utility of different models. *Model validation* is the process for measuring the uncertainty or variability in a model's estimates and identifying the sources of that uncertainty (Citro and Hanushek 1991).

One technique of model validation is *external validation*, in which the model's estimates are compared with the outcomes the model is forecasting. Because the outcomes may never be known, however, two variants of external validation may be used. One is *ex post forecasting*, in which one takes a policy or program change adopted in the past and, applying a model to data available then, forecasts the effects in a more recent past year when outcome data are available to compare. Another variant is *backcasting*, in which the model and current data are used to simulate the effects of a reversion to program conditions in the past, and the results are compared to measures of the effects based on data in that period.

Internal validation refers to studies of the model, its components, and how it is used, as well as of the data used. Measures of internal validity derive from the data used and model estimates obtained, without any comparison to external reality. They include sampling variability and other errors in the data or other inputs, and errors of model misspecification. Estimates of variability may employ standard sample variance calculations or more modern resampling techniques. *Sensitivity analysis* looks at the effects on model estimates of changes in one or more assumptions, specifications, or components of the model. It is used to determine the factors that most affect the results, or how robust the results are to changes in assumptions.

3.2 *Documentation*

Documentation of data, models, and analyses is important, not only for validation studies but also for a better understanding of the model estimates and their quality. Documentation should include information on the policy alternatives under analysis, the data and other inputs, assumptions, interpretations,

and changes made. The information should be specified in sufficient detail so that another person could replicate the results. Otherwise, the utility of the model may be limited to only a few analysts who are experienced in its use. Moreover, as models invariably grow in complexity, even experienced users may not understand the implications of underlying assumptions, or how components of the model interact.

3.3 Communication of Uncertainty

Policy analysts should include information about the uncertainty of their estimates and the sources of that uncertainty in the presentations of results to policy makers. Sampling variability is often the most straight forward to calculate and report, but it neglects errors due to model misspecification, nonsampling errors in surveys, and other factors. Efforts should be made to convey a more complete description of uncertainty through a variety of techniques, including variance calculations, sensitivity analysis, and the proven record of a model.

It is important for policy makers to understand the uncertainty attendant in model estimates, for a number of reasons. First, the information on uncertainty allows them to distinguish differences from those that could occur solely by chance. Second, they can use the information to compare the quality of different estimates. Third, the information may help them decide how much weight to give to a policy analysis, as opposed to other, political considerations.

But there is an even more important reason for agencies that are responsible for policy analysis to measure and communicate the uncertainty in data and model estimates. It is only through such measures of quality that resources to improve the data and models can be justified. For example, in the United States, many national surveys are so sparse that they are unable to provide direct survey-based estimates for all but a handful of states. National policy makers have devolved some programs to the states, such as children's health insurance and welfare reforms, and expect estimates of outcomes to hold states accountable. The estimates can be modeled, but their much larger uncertainty than that for the national average must be communicated. Not only would policy makers know if a reported difference reflects a real difference; they could see the need to invest resources to improve data or models, or both. For example, the sample size of the Current Population Survey has been increased to improve estimates, in each state, of the numbers of children covered by health insurance.

4. Organization of Policy Analysis

Agencies conducting research in support of public policies in a department of government must anticipate policy issues if its research is to inform the policy debate when it occurs. As a result, a policy research agency may invest resources in research on policy issues for which it may never be asked for the results. When consideration of an unanticipated policy comes to the fore, the agency must be able to draw upon a synthesis of research or a review of extant data to provide the required policy information. For that reason, the policy research agency must develop data and research continually, and be cognizant of that developed elsewhere in the important subject matters of its department.

The national government may provide for unsolicited basic and applied research, in particular in academic institutions, but policy analysis rarely, if at all, is a byproduct of this research. Moreover, policy research conducted by private organizations with a stake in the outcome may not be objective. Effective policy analysis, therefore, can neither be delegated to academia nor to interested organizations. Government policy analysis agencies must either conduct or direct the research. In doing so, they may draw upon research conducted by a wide variety of institutions, such as quasi-governmental organizations, universities and academic research institutes, private foundations and other nonprofit organizations, and for-profit consulting firms.

Policy analysis in a department of government may or may not be the responsibility of a distinct, independent unit. For example, the United States has many units in the executive branch with responsibilities for policy analysis, such as the Office of Tax Analysis in the Department of the Treasury, and the Office of the Assistant Secretary for Planning and Evaluation in the Department of Health and Human Services, but no single unit in the Department of Transportation is responsible for policy analysis. Moreover, policy analysis need not be the responsibility of just the executive branch. The US Congress has developed policy analysis capabilities, through such agencies as the Congressional Budget Office, the Joint Committee on Taxation, and the Congressional Research Service, to counterbalance the policy analysis capabilities in the executive branch.

For government policy research agencies, maintaining independence from the political process itself is essential to the agency's effectiveness. 'Research in support of public policy depends for its credibility, in great part, on its conduct by those with a greater interest in the quality of the work than in the substance of the conclusion' (Geweke et al. 1999).

5. Conclusion

Developing information to inform debate on public policies requires data, models, and an understanding of uncertainty in the resulting estimates. By far the

most common problem and the bane of policy analysts is the lack of adequate data. When a panel of the National Academy of Sciences–National Research Council was asked by the US Department of Labor to advise on the development of models to assess policies for retirement income, the panel concluded instead that priority be given to funding data collection and behavioral research in preference to significant investments in large-scale projection models:

> Generally speaking, it takes more time to collect new data and analyze them than it does to build a model to use data and research results. There are more than a few instances in the history of policy analysis when complex projection models were built in a span of weeks or months. It is rare that a critical unmet data or analysis need can be supplied in that short a time. (Citro and Hanushek 1997)

Government, the panel concluded, must take the lead in data collection, so that high-quality, comprehensive databases are available for policy analysis use in all sectors—government, private, and academic.

See also: Policy Knowledge: Census; Sample Surveys, History of; Sample Surveys: Methods; Sample Surveys: The Field; Survey Research: National Centers

Bibliography

Citro C F, Hanushek E A 1991 *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Vol. I. Panel to Evaluate Microsimulation Models for Social Welfare Programs, Committee on National Statistics, National Research Council. National Academy Press, Washington, DC

Citro C F, Hanushek E A 1997 *Assessing Policies for Retirement Income: Needs for Data, Research, and Models*. Panel on Retirement Income Modeling, Committee on National Statistics, National Research Council. National Academy Press, Washington, DC

Geweke J F, Bonnen J T, White A A, Koshel J J 1999 *Sowing Seeds of Change: Informing Public Policy in the Economic Research Service of USDA*. Panel to Study the Research Program of the Economic Research Service, Committee on National Statistics, National Research Council. National Academy Press, Washington, DC

M. Straf

Policy Knowledge: Advocacy Organizations

Substantive policy knowledge is a set of beliefs that encompass perceptions of (a) the seriousness of various dimensions of a policy problem (e.g., the health impacts of air pollution, the declining quality of health care), (b) the relative importance of various causes of those problems, and (c) the expected impacts of policy options for changing current conditions. In conflicts between advocacy organizations that pursue their interests through the policy-making process—including interest groups, corporations, administrative agencies, and associations of governments—debate surrounding the various elements of policy knowledge is often technical in nature. These debates frequently employ scientific research, technical policy analyses (such as benefit/cost analyses), and policy evaluations that follow scientific norms of data acquisition and analysis.

There are four main frameworks for examining the nexus between substantive policy knowledge and the political behavior of advocacy organizations: (a) a 'civics textbook' approach; (b) an 'atheoretical skeptics' perspective; (c) the 'multiple streams' approach; and (d) the 'advocacy coalition framework.' All vary in their attention to the political process underlying public policy debates and the potential biases inherent in substantive policy information.

1. The Civics Textbook View of the Role of Science

This is basically a normative model, derived from certain fundamental principles of democratic theory. It assigns very distinct roles to three categories of actors in the policy process:

Elected officials are responsible for making basic policy decisions in a manner which reflects the distribution of values in society. They use scientific findings to help them understand trends in various problems, the factors affecting them, and the means of alleviating those problems. Their basic role is to establish clear laws and budgetary priorities for implementation by agencies.

Governmental agencies are composed largely of civil servants who should be politically neutral (i.e., faithfully implement whatever the legislature decides) and who have a special role in fostering applied research in areas of interest to the agency. Political appointees within the agency are responsible for seeing that the agency reflects the government's priorities—to the extent permitted by law.

Scientists are supposed to be neutral seekers of the truth. Their role is to understand the world and to present this information to policy-makers.

This is a normative model of how science *should* be used in making public policy, and many scientists involved in policy disputes do, in fact, view themselves as 'objective technicians' (Meltsner 1976).

According to this model, the major problem impeding communication between scientists and governmental officials is that they inhabit two quite distinct communities with different value priorities, timeframes, and methods for resolving conflict (Dunn

1980). Proposals for reform thus focus on improving communication between the two communities by exchange programs, the development of facilitating or 'translating' institutions, etc.

The textbook model has substantial limitations in practice, in large part because most actors do not behave as the model indicates they should.

Elected officials seldom pass clear laws on contentious issues with substantial technical components because very few have the expertise to obtain a decent understanding of the technical issues. Instead, they tend to pass procedural or 'framework' laws which hand over the problem to an administrative agency without clear policy priorities.

Agency officials are seldom as neutral as the civil servants in the Weberian model of bureaucracy (Knott and Miller 1987). Most agencies have a fairly clear overall mission which tells them to give priority to some values over others. Most officials who join the agency come to accept those priorities, whether out of self-selection or gradual indoctrination. Agencies are often dominated by members of a particular profession or scientific discipline who share the norms of colleagues outside the agency. Thus, most agencies can be expected to sponsor research consistent with their mission and to be skeptical of findings which cast doubt on its wisdom.

Scientists are often not neutral participants. Virtually all scientists operate within a specific 'paradigm,' i.e., a set of often implicit assumptions about basic causal assumptions and proper methods of investigation which guide research (Kuhn 1970). More importantly for our purposes, almost all scientific disciplines contain important normative assumptions which members often come to accept in an uncritical fashion. The normative assumptions behind welfare economics and benefit/cost analysis have been widely discussed (Rhoads 1985). With respect to nuclear waste disposal, Barke and Jenkins-Smith (1993) provide evidence that biologists perceive significantly greater risks than physicists, chemists, and engineers. The latter think basically in terms of dose-response curves, while biologists are more wary of the effects of *any* dose on living organism.

Moreover, scientists are often drawn to applied—rather than basic—research because they want to help solve a particular problem. Having a demonstrable effect on policy normally requires the accumulation of results over an extended period of time (Weiss 1977). The more neutral and 'apolitical' scientists are unlikely to remain interested in an issue long enough to have such an impact. Thus the most active scientists in a particular dispute are likely to be those who have been involved the longest and who are most committed to defending a particular point of view.

The end result is that scientists who have something to contribute to important policy disputes are seldom neutral (Mazur 1981). The argument does not imply that they manipulate or falsify data. Instead, disciplinary paradigms, the values underlying their discipline, and their desire to solve particular problems affect the topics they choose to research, the variables they focus on, the methods they utilize, where they place the burden of proof in cases of uncertainty, and how quickly they present various results. For example, wildlife biologists are much more likely than engineers to look for species in trouble because their disciplinary norms define species extinction as a serious problem. They are more likely to look to human technological interventions as a likely explanation because they tend to respect the beauty of natural systems. In contrast, engineers assume they can improve on nature. Members within each discipline will readily present results which are congruent with these assumptions, while incongruent results are likely to be interpreted as tentative and in need of further verification (Brown 1977).

2. *The Atheoretical Skeptics' View of the Role of Scientists in Policy Disputes*

Over the years, a number of studies have seriously questioned the textbook view that scientists participating in policy are completely neutral. First, Weinberg (1972) argued that some topics—such as risk assessments of very improbable, catastrophic events—are essentially in the realm of 'trans-science.' Second, as argued above, there is increasing recognition that the topics chosen for research, the allocation of the burden of proof in areas of uncertainty, and the presentation of results are affected by scientists' disciplinary paradigms (Stewart 1986), organizational interests (Primack and von Hippel 1974, Jasanoff 1987), and/or policy concerns (Nelkin 1971).

Although suggestive, most of these studies have suffered from two serious limitations. First, they have failed to articulate a comprehensive theory (or theories) of the role of 'advocacy science' in the overall policy process. Instead, they have usually limited themselves to identifying certain institutional/organizational features that inhibit or exacerbate the extent of advocacy (Mazur 1981). But there is very little recognition of the role of broader social norms, economic conditions, and the myriad of advocacy organizations involved in most policy subsystems. Second, almost all of this research has consisted of essays or of rather traditional case studies which (a) lack explicit theoretical grounds for case and variable selection and (b) utilize quite subjective methods of data acquisition and analysis. The result has been a plethora of interesting arguments and illustrative examples, but very little systematic, intersubjectively reliable analysis of the extent of advocacy science, the factors affecting it, and its effects on the overall policy process.

3. The Multiple Streams Framework of the Policy Process

Drawing upon the 'garbage can model' of organizational choice (Cohen et al. 1972), Kingdon (1984) developed an approach to agenda-setting and policy formulation that deals explicitly with the role of substantive policy information. In his view, policy-making can be conceptualized as three largely unrelated 'streams':

(a) A *problem stream* consisting of information about real world problems and the effects of past governmental interventions. This would include large databases of problem severity and causes gathered by governmental agencies and research organizations, as well as policy analyses of the benefits and costs of previous attempts to address those problems. This information is treated as relatively unbiased.

(b) A *policy stream* consisting of proposals to solve one or more problems. This is the province of policy analysts in government think tanks and consulting firms, as well as many legislative staff. Kingdon notes the tendency for analysts to become infatuated with a particular solution and then to seek to apply it to a wide variety of problems. For the past 30 years, for example, many resource economists have viewed tradeable permit schemes as the solution to virtually all environmental problems (Tietenberg 1984).

(c) A *political stream* consisting of elections, legislative leadership contests, and other openly political competition.

According to Kingdon, major policy reforms result when a 'window of opportunity' joins the three streams. The window may be opened by a 'focusing event,' such as an airplane crash or oil spill, that dramatically illustrates the potential seriousness of a problem, or it can be opened by a political election that brings a new leader with a new agenda to power. In response to the window, major policy change will result if the policy community develops a proposal that is financially and technically feasible and if elected officials find it advantageous to approve it.

The multiple streams approach has a number of promising features. It gives a prominent role to substantive policy information about real world problems and the impacts of past governmental interventions. It was one of the first frameworks that treats researchers and policy analysts in consulting firms, think tanks, and universities as potentially important political actors. And it acknowledges the role of serendipity—particularly in the case of focusing events—in the policy process.

On the other hand, it has several limitations. First, we are not convinced that the problem stream is nearly as neutral and apolitical as Kingdon suggests. Bureaucracies gather information on some aspects of a situation and not on others. The US Department of Agriculture gathers much more extensive information on crop production and prices than it does on nitrogen runoff from fields into streams and groundwater basins. Second, the conditions creating windows of opportunity need further analysis. Third, the multiple streams framework contains virtually no falsifiable hypotheses, and thus very little has been learned about its strengths and limitations over time (cf. Kingdon 1995, Chap. 10).

4. An Advocacy Coalition Framework of the Policy Process

The advocacy coalition framework (ACF) starts from the premise that political elites concerned with a specific problem or policy domain—such as transportation, education, health, or air pollution—will form relatively autonomous subsystems composed of interest-group leaders, administrative agency officials, researchers, and a few legislators with a specific interest in the topic. These subsystems (or policy communities) usually involve actors from multiple levels of government. Once a subsystem is formed, policy disputes become increasingly technical and specialized as actors learn more about the complexity of problems and solutions. In such a system, the role of legislatures and political parties is primarily to aggregate the decisions made by policy specialists and to provide some general guidance.

The ACF views policy change over time as primarily the result of competition among advocacy coalitions within a policy subsystem (Sabatier and Jenkins-Smith 1993, 1999). An advocacy coalition consists of interest-group leaders, legislators, agency officials, researchers, and even journalists who share a set of basic beliefs (policy goals plus critical perceptions of causal relationships) and engage in some degree of coordinated behavior in an effort to make governmental policy more consistent with those beliefs. Conflict among coalitions is mediated by 'policy brokers,' i.e., actors more concerned with fashioning an acceptable compromise than with achieving specific policy goals. While the framework focuses on competition among coalitions within the subsystem, changes external to the subsystem (such as fluctuations in socioeconomic conditions) and stable system parameters (such as constitutional rules) play an important role in major policy change.

The belief systems of advocacy coalitions are assumed to be hierarchically organized. At the highest/broadest level, the 'deep core' of a coalition's belief system consists of fundamental normative beliefs, such as the familiar Left/Right scale, that operate across virtually all policy domains. Within any given policy subsystem, however, it is the 'policy core' and the 'secondary aspects' that are most critical. The former consists of basic positions, some of them purely normative (e.g., the relative importance of different values, such as environmental protection vs. economic development), while others are a mixture of normative

and empirical (e.g., the proper scope of governmental vs. market authority for realizing those values), that operate across most or all of the policy subsystem. These policy core positions are very resistant to change, are only intermittently the subject of policy debate, and are usually changed as a result of perturbations external to the subsystem, although long-term 'enlightenment' may also play a role (Weiss 1977). Science plays a much more important role in the secondary aspects of coalitions' belief systems, as these involve disputes over the seriousness of a problem or the relative importance of various causal factors in different locales, the evaluation of various programs and institutions, and specific policy preferences.

The ACF assumes that members of a coalition will readily accept new evidence consistent with their views and seek to discount information which conflicts with their perception of the seriousness of a problem, the relative importance of various factors affecting it, or the costs and benefits of different alternatives (Lord et al. 1979). The result is 'a dialogue of the deaf?' in which members of different coalitions talk past each other. Given that policy core beliefs are resistant to change, the composition of coalitions is hypothesized to be stable over periods of a decade or more.

The ACF explicitly rejects the assumption that most bureaucrats and researchers involved in a policy area will be neutral. Some may well have no strong policy preferences, at least initially. But the framework contends that, as conflict between coalitions increases and as the interrelationships among sets of beliefs become clearer over time, initially loose groups with amorphous beliefs will coalesce into increasingly distinct coalitions with coherent belief systems; in the process, most neutral actors, particularly university scientists, will drop out. The ACF thus contends that, in well-developed subsystems, most agency officials and researchers who are active will be members of specific coalitions in terms of sharing a set of policy core beliefs and acting in concert to some degree (Sabatier and Zafonte 2000).

Policy-oriented learning involves relatively enduring alterations of thought or behavioral intentions that result from experience and/or the assessment of new information involving the precepts of belief systems (Heclo 1974). Learning *within* coalitions is relatively easy, as members share the same value-premises and are looking for the most effective means to those ends. Learning *between* coalitions is much more difficult, as people from each coalition have quite different values and distrust each other. Nevertheless, the original versions of the ACF hypothesized three conditions facilitating such learning (Sabatier and Jenkins-Smith 1988).

(a) *Issues on which there is an intermediate level of conflict*. Issues have to be important enough to generate sufficient research—usually by members of several coalitions, as well as neutrals. On the other hand, issues involving conflict between the core beliefs of different coalitions (e.g., the rights of non-human beings or the ability of humans to improve on nature) generate more heat than light. Learning across coalitions is thus most likely on issues involving important secondary aspects of the relevant belief systems.

(b) *Issues involving primarily the natural sciences*. Across-coalition learning is generally easier in the natural than in the social-behavioral sciences because the theories and accepted methods are better established and the objects of study are not themselves actors in the policy debate.

(c) *The existence of a forum* that is (i) prestigious enough to force professionals from different coalitions to participate and (ii) dominated by scientific norms. The latter assures a general consensus on the appropriate rules of evidence and a minimum of *ad hominem* attacks, as well as some attention to underlying assumptions. Possibilities for professional forums include 'blue ribbon' committees appointed by the National Research Council or a professional association, studies by organizations with a strong reputation for neutral competence (e.g., the Congressional Budget Office), or, in some cases, interagency technical advisory committees.

The following set of conditions have been developed for successful professional forums, i.e., those in which experts from the competing coalitions come to consensus on technical and/or policy issues placed before it. In general, such forums are most useful when a fair amount of scientific/technical information on various aspects of the topic exists, but the conclusions of different studies vary and the validity of much of the evidence is questioned by members of opposing coalitions. The arguments below assume that the essential task of a professional forum is to convince coalition experts that professional norms require the alteration of one or more beliefs important to a coalition.

Composition. In order for scientists with very different points of view to come to consensus and for that consensus to be accepted by the major coalitions, the technical advisory committee should include both (a) scientists clearly associated with each of the major coalitions and (b) neutral scientists. The chair should come from the latter.

The ACF assumes that the various coalitions involved in important policy disputes have scientists (or technical experts) whom they trust, presumably because they share most of the coalition's policy core beliefs and can help the coalition understand the scientific aspects of such disputes. Many interest-group leaders, legislative personnel, and agency officials who constitute the leadership core of coalitions have an intuitive awareness of the role of values in determining the agenda of technical advisory committees and how uncertainty is treated. Thus they are very unlikely to accept the recommendations of any technical advisory committee on which their point of view has not been argued by someone whom they trust.

And, if they do not trust the committee's report—particularly in a relatively decentralized political system like that of the USA—they almost always can find some decision-making venue that will enable them to circumvent or obstruct the committee's recommendations. In short, selecting a committee composed entirely of relatively 'neutral' scientists—which we perceive to have been the strategy of the National Academy of Scientists (Boffey 1975), for example—may facilitate short-term consensus but will probably fail in the long run.

On the other hand, the technical advisory committee needs to be chaired by a neutral (and probably to include a few other neutrals) whose task is to impose professional norms on the debate regarding acceptable evidence, methodologies, etc., and to indicate to coalition experts when a professional consensus is beginning to emerge. Nonprofessionals must be excluded from these deliberations in order that scientific norms can prevail in weighing the evidence for—and the assumptions behind—different points of view.

Funding and Sponsorship. Funding must come from an institution which is not perceived as being controlled by a single coalition. This will usually require funding either by (a) a neutral foundation or (b) multiple agencies representing the various coalitions. This is important because policy participants are unlikely to trust a committee funded—and therefore potentially controlled—by their opponents. Since many agencies are perceived as being part of a specific coalition (Sabatier and Zafonte 2000), single-agency technical advisory committees will lack legitimacy in the eyes of opposing coalitions.

Duration. A forum should meet at least half a dozen times over a year or so. It takes time for scientists from different coalitions to analyze their hidden assumptions, to evaluate the evidence critically, and to begin to trust each other. One-shot committees of short duration will probably not work.

Context of Policy Stalemate. A forum will be successful only in a context of policy stalemate, i.e., when each of the coalitions views a continuation of the status quo as unacceptable. In any policy dispute, arriving at a compromise acceptable to everyone requires that all the coalitions view the status quo as unacceptable. If any should view the status quo as acceptable, then it will be much less willing to compromise and may not even participate.

The assumption thus far has been that lasting policy change requires genuine learning between/across coalitions. In a relatively open and decentralized political system, such as the USA's, coalitions that do not agree with the conclusions of a professional forum have numerous points of appeal (courts, legislatures, agencies at different levels of government) which have different biases. In more centralized systems, such as UK's, where routes of appeal are severely restricted, it may be possible for a coalition to change policy simply by convincing a policy broker of the merits of its point of view without having to change the views of the other coalition(s).

5. *Summary*

This article has presented four conceptual frameworks that seek to understand the role of advocacy organizations in processing substantive policy information of a relatively technical nature in policy disputes, particularly in the USA. Both the civic textbook model and the multiple streams model view substantive policy information as being relatively unbiased, but the multiple streams has a much more sophisticated understanding of the policy process. In contrast, both the atheoretical skeptics and the ACF view substantive policy information as 'filtered' through organizational, disciplinary, and coalition lenses, but the ACF has a much more sophisticated model of the policy process in intergovernmental policy subsystems. This sophistication, however, comes at the cost of parsimony. All four perspectives in combination can be useful in generating a clearer understanding of the role of policy knowledge in the behavior of advocacy organizations.

See also: Advocacy and Equity Planning; Advocacy in Anthropology; Interest Groups; Labor Unions; Lobbying; Policy Knowledge: Nonprofit Institutions; Think Tanks

Bibliography

Barke R, Jenkins-Smith H 1993 Politics and scientific expertise: Scientists, risk perception, and nuclear waste policy. *Risk Analysis* **13**(4): 425–39

Boffey P 1975 *The Brain Bank of America*. McGraw-Hill, New York

Brown H 1977 *Perception, Theory, and Commitment*. University of Chicago Press, Chicago

Cohen M, March J, Olsen J 1972 A garbage can model of organizational choice. *Administrative Science Quarterly* **17**: 1–25

Dunn W 1980 The two-communities metaphor and models of knowledge use. *Knowledge* **1**(June): 515–36

Heclo H 1974 *Social Policy in Britain and Sweden*. Yale University Press, New Haven, CT

Jasanoff S 1987 Contested boundaries in policy-relevant science. *Social Studies of Science* **17**: 195–230

Kingdon J 1984 *Agendas, Alternatives, and Public Policies*. Little, Brown, Boston

Kingdon J 1995 *Agendas, Alternatives, and Public Policies*. Little, Brown, Boston

Knott J, Miller G 1987 *Reforming Bureaucracy*. Prentice-Hall, Englewood Cliffs, NJ

Kuhn T 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

Lord C, Ross L, Lepper M 1979 Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* **37**: 2098–109

Mazur A 1981 *The Dynamics of Technical Controversy*. Communications Press, Washington, DC
Meltsner A 1976 *Policy Analysts in the Bureaucracy*. University of California Press, Berkeley,CA
Nelkin D 1971 *Nuclear Power and Its Critics: The Cayuga Lake Controversy*. Cornell University Press, Ithaca, NY
Primack J, von Hippel F 1974 *Advice and Dissent*. Basic Books, New York
Rhoads S 1985 *The Economist's View of the World*. Cambridge University Press, London
Sabatier P, Jenkins-Smith H (eds.) 1988 Symposium on 'policy change and policy-oriented learning.' *Policy Sciences* **21**(2–3): 123–77
Sabatier P, Jenkins-Smith H (eds.) 1993 *Policy Change and Learning: An Advocacy Coalition Approach*. Westview Press, Boulder, CO
Sabatier P, Jenkins-Smith H 1999 The advocacy coalition framework: An assessment. In: Sabatier P (ed.) *Theories of the Policy Process*. Westview Press, Boulder, CO, pp. 117–68
Sabatier P, Zafonte M 2000 Are scientists and bureaucrats members of advocacy coalitions? In: Sabatier P (ed.) *An ACF Lens on Environmental Policy*. Under review
Stewart J A 1986 Drifting continents and colliding interests. *Social Studies of Science* **16**: 261–79
Tietenberg T 1984 *Environmental and Natural Resource Economics*. Scott Foresman, Glenview, IL
Weinberg A 1972 Science and trans-science. *Minerva* **10**: 209–22
Weiss C 1977 Research for policy's sake: The enlightenment function of social research. *Policy Analysis* **3** (Fall): 531–45

P. A. Sabatier and M. A. Zafonte

Policy Knowledge: Census

A census is the count of the entire population of a nation, normally tabulated by jurisdictions within the nation such as provinces, districts, cities, towns, or villages. A modern census also collects social, economic, and housing characteristics as part of the population count. Census information is heavily used in state planning and in public policy. The social sciences, especially demography, sociology, history, and economics, draw from census data across a wide range of research topics: fertility, migration, race and ethnicity, economic growth, and many others. However, with a few exceptions, the census itself has not been a research subject in social science; for example, there are no cross-national comparisons analyzing differing ways in which the census is involved in social, political, and economic processes.

1. *History/Definition*

The archaic definition of a census is 'poll tax,' indicating its early association with the state's function of revenue extraction. Although the term census dates to ancient Rome, where it referenced the registration of citizens and their property, the idea of a census predates Roman times. For example, *Numbers*, the fourth book of the Hebrew Bible, starts with the Lord instructing Moses: 'Take ye the sum of all the congregation of the children of Israel, after their families, with the number of their names, every male … from twenty years old and upward.' That is, take a census of those 'able to go forth to war.'

It is probable that taxation or military conscription or both gave rise to other census operations indicated by archeological records, such as the *Nepohualco* or 'counting place' built by the Chichimecas a millennium ago. This site, near today's Mexico City, consists of stones deposited by each inhabitant of the region, an early census databank. Other archeological records from various ancient civilizations indicate similar preoccupation by the state with carrying out systematic statistical studies. In these ancient records it is difficult to detect all of the state purposes that might have been served by a census. Certainly there was interest in economic data, as is suggested by statistical records of harvest and grain storage. Ancient Chinese records, dating from the Han dynasty (206 BC to AD 220), indicate that detailed population counts were made after each harvest. Subsequent systematic census records appear throughout Chinese history, and many scholars concur that China should be credited with the first comprehensive census (Domschke and Goyer 1986).

There were attempts to establish a periodic census in early modern times, such as the ambitious seventeenth century proposal for an annual census directed to France's Louis XIV:

> Would it not be a great satisfaction to the king to know at a designated moment every year the number of his subjects, in total and by region, with all the resources, wealth & poverty of each place?… [Would it not be] a useful and necessary pleasure for him to be able, in his own office, to review in an hour's time the present and past condition of a great realm of which he is the head, and be able himself to know with certitude in what consists his grandeur, his wealth, and his strengths? (cited in Scott 1998, p. 11).

In the modern period, however, it was the United States that initiated the first periodic census. The US Constitution (1787) requires that an enumeration of the population be taken every 10 years. This decennial census, first taken in 1790, served two state-building functions. One was periodically to reallocate seats in the House of Representatives, the legislative branch to which positions are allocated proportionate to size of population in each of the member states. Because the population might grow at different rates from one state to the next, provision was made to take a census every decade and then to calculate how many legislative seats would go to each state. The second nation-building task addressed the challenge of geographic expansion. The country was expected to expand into new territories, and these areas were to join the nation

on equal footing with the original states. The decennial census measured population growth and its geographic dispersion and thereby served as the mechanism regulating how territories were added to the state.

Though the US was early in establishing a periodic census, most European nations shortly followed in instituting some form of regular census to assist in nation-building and policy-making—the United Kingdom and France in 1801; The Netherlands, shortly after independence, in 1830; Sweden in 1855; Spain in 1857; Italy in 1861; the German Empire in 1871; and the Russian Empire in 1897. In nearly every instance, the postcolonial nations of Latin America, Asia, and Africa established a periodic census as an early step in nation building. By the end of the twentieth century only a handful of countries do not have a census.

2. *Census and Public Policy*

A wide array of public policy purposes are served by a regular official count of the population and its characteristics. Before turning to these purposes, it is necessary to distinguish between a census and a sample survey. The vast majority of state-generated statistical information comes from sample surveys, which elicit information from a part of the population in order to estimate characteristics of the entire population. A census customarily describes the effort to count everyone in the population. There is an important statistical interaction between sample surveys and the census, for the latter provides the frame necessary to draw samples for survey purposes. It also provides the basic denominator that serves as a control to determine whether a sample is demographically and geographically representative of the country. That is, does the sample have the correct proportion of age, race, ethnic, gender, and other demographic categories that have been measured in the population census, and does it reflect how the population is distributed geographically, again as measured in the census. The close relationship of a census and surveys indicates that many policy functions that rely on sample data are, in fact, drawing from a prior census.

An important distinction between the census and the survey is that the former allows estimates of population or housing characteristics to be made at comparatively small levels of geography. Policy purposes that require fine geographic detail will directly draw on census information. Policy purposes that are targeted at higher levels of aggregation tend to rely on sample survey data and thus have an indirect relation to the census.

A census supports policy-making in four broad arenas: election systems, taxation, public funding of social services, and social control.

2.1 *Election Systems*

When the basis of election to a legislature is size of the population being represented, census information is used to draw the boundaries of election units. As indicated above, this was the primary purpose for which a census is required by the US Constitution. It is now the practice in dozens of democratic countries, and not just for national legislatures. Election districts for city governments and other local and regional governments are frequently drawn to ensure approximate equal population size for each geographic area that constitutes an election unit.

Size is not the only population characteristic that can be taken into account in drawing election districts. Demographic traits such as ethnicity, language, and race have been used. This occurs when the state wants to ensure that the legislative body is 'demographically representative' of the population at large, for instance, that racial minorities have the opportunity to elect someone from their race to the legislature. Under these circumstances, census data are commonly used to monitor compliance with election laws. In the United States, for instance, census information on ancestry, gender, race, ethnicity, language spoken at home, place of birth, and year of entry into the country are critical for monitoring compliance with voting and civil rights laws.

When political party affiliations vary across different ethnic, language, or racial groups—as they often do—the use of the census to promote ethnic, linguistic, and racial balance or access will invite partisan disagreements. Census counts become the target of such disagreements, as in the events that led to the Nigerian civil war. Even census methodology can come under attack, as has happened in the United States in connection with its 1990 and 2000 census.

Basing political representation on population counts also invites jurisdictions to inflate their numbers. This incentive confronts census taking with serious methodological challenges. Unless there is a corresponding disincentive—such as basing tax rates on population size—a census can be vulnerable to strong political pressure by those with a direct stake in high counts.

2.2 *Taxation*

The most direct relation of a census to taxation policy occurs when taxes are collected from a jurisdiction proportionate to its population size. Thus a central government might calculate the tax obligation of its administrative units depending on the size and wealth of the unit, and use a census to make the determination.

A more indirect use of census (and survey) data occurs when the state measures the economic life of

the nation, using that information to help design a taxation system—whether to tax manufacturing activity, natural resources, individual income, property, at what rates and in what combinations. For example, population projections based on census data are used to design social security tax rates and to estimate probable social security payment levels well into the future.

2.3 Public Funding and Planning

Social welfare policies are generally targeted to specific population groups, such as families in substandard housing, children in poverty, the physically disabled, minorities historically disadvantaged or mistreated, victims of particular diseases. Assessing the size and geographic distribution of groups targeted for welfare payments or subsidies is the task of a nation's statistical system. Frequently public funding formulae are based on census counts of particular population groups (Nathan 1999). In the United States, for instance, in 1999 approximately $US185 billion in medical, education, housing, transportation, unemployment, and other payments depend in part on census data. Similar funding formulae using census information occur in many countries with advanced social welfare systems.

When public funds are allocated to target groups based on their size, they have a strong interest in being fully counted. But these are also population groups that are difficult to reach in standard census operations, and even if reached, will sometimes resist being counted because of concerns that census information can be used against them. This situation has given rise to advocacy organizations that work with and put pressure on census bureaus to assure as full a count as possible of their particular constituencies. Where this has occurred, census taking is no longer the domain only of professional statisticians and their government minister, but becomes of much more generalized political interest.

In addition to targeted welfare spending, national governments frequently use census information for program planning across many policy domains such as public transportation, health and education services, job training, police and fire protection, sanitation and sewage disposal, and land use.

The census also plays a role in national security policy. As indicated above, since biblical times the state has relied on census information to assess the numerical strength of its fighting age population, and also at times to design and implement actual conscription. Of course, a reliable count of the fighting strength of one nation's population will be used by another nation to calculate whether that nation poses a military threat. The first US census in 1790 was declared by the nation's leaders to have missed a sizeable portion of the population, and the Secretary of State took care to announce that fact to European powers that might otherwise have concluded that the young nation was too weak to protect its independence. Of course not just population censuses but also economic censuses help the state assess its military capacity and that of other nations.

The relative strengths and weaknesses of contending groups within a nation are also revealed by census information, and have often played a role in strategic choices that precede internal warfare. The suppression of census data in several African countries reflects a decision by the ethnic group in power to disguise the comparative numerical strength of ethnic groups challenging its rule. In the period preceding the Civil War in the United States the more rapid growth of 'free-states' compared to 'slave-states,' as revealed in the 1850 census, was a factor leading to the succession of the South from the national union.

2.4 Social Control

More generally the state uses the census as an instrument of social control, as it was routinely used during the colonial era. Great Britain, as early as the seventeenth century, introduced a census in many of its colonies in order to control more effectively its distant subjects and to assure that resource extraction proceeded smoothly. Decisions about which territories needed policing, required additional labor, could benefit from infrastructure investments, and so forth were facilitated if a reliable census could be taken. Great Britain and other colonial powers were early practitioners of what has now become commonplace. Social control policies that police borders, population movement, and even reproduction and fertility are fashioned in part on the basis of census information.

The boundary between routine social control and abusive exercise of power by the state is not easily drawn, but certainly history offers many examples of the census as an instrument of the latter. Family planning policy and fertility regulation in China and India, for instance, is often cited in this regard.

Population displacement is a further example. Census data have been used to place restrictions on freedom of movement or to force population groups to live in places they would not have chosen. An important instance of the latter is the role of US census data in the nineteenth century relocation of native American Indian tribes from their homelands to reservations. US census data were also used to facilitate the internment of Japanese-Americans at the beginning of World War Two, because they were said to pose a security risk shoul Japan invade America's west coast.

Census data have also been used in genocidal policies, the most tragic example being the Holocaust. Population registers or special censuses identifying Jews greatly assisted in the location of Jews and their transport to concentration and extermination camps. The 1939 German census, for example, was used to

identify specific individuals. In The Netherlands small-area tabulation of a census of religions taken in 1930 was used in dot-maps of Amsterdam to indicate the density of the Jewish population. Poland, France, and Norway offer additional evidence linking census data to Holocaust operations (Seltzer 1998).

These historical instances of gross misuse of census information have had one beneficial outcome. There is now much more professional and political sensitivity to issues of data confidentiality, the independence of statistical agencies, and related ethical considerations (Seltzer 1994).

3. Census Data and Democracy

Census information connects to politics, and especially to democratic politics, in ways other than its use by the government to design and implement public policy. Statistics and especially statistical series also advance key features of democratic politics. One purpose is helping the electorate to assess government performance and hold it accountable. Another is in providing a platform for social reform. A third is in the assertion of group interests (Prewitt 1999).

3.1 Democratic Accountability

The theory of electoral accountability holds that competitive elections offer to the electorate alternative portrayals of how well the government in power has performed or will perform against future challenges. Aspirants for political office present themselves in terms of their past accomplishments and their promises of future accomplishments in managing the economy, protecting the nation's security, and generally enhancing national well-being. Voters then elect, re-elect, or evict accordingly.

This simple model presumes that the voters have some way of assessing government performance. Statistical trends and social indicators play a major role. They indicate whether the economy is growing or stagnating, whether education or health or housing is improving, whether crime rates are down, whether the environment is being protected, and so forth. Arguments about who can take credit for improvements, or should be blamed for failure, are the common currency of competitive elections and are often advanced by citing statistical trends.

Consider, for example, the sharp political debate over immigration policy in many wealthier countries as they face heavy demands for entry from Central Asia, Northern Africa, and Latin America. These debates are framed in terms of the 'proper' balance between native population groups and newer immigrant groups, with census and related data providing the subtext. Governments are then judged in terms of how well they protect racial purity with anti-immigration policies; or, conversely, display humanitarian compassion or perhaps, as in the United States at present, replenish the labor pool with more open immigration policy. In either instance, as in so many areas of government programming, long-term statistical trends are invoked when assessing policy.

3.2 Social Reform

Democratic theorists have long been interested in how issues get on the political agenda, especially issues that matter most to the unorganized or disenfranchised interests of society. For instance, how do issues such as child labor or male–female wage differentials become politically salient? One route to the political agenda is though social reform action that takes as its starting point statistical profiles. Since the early days of industrialization, reform groups have used statistics to document the social conditions that they were committed to eradicating (Cohen 1982). Examples are poverty laws, prison reform, child abuse, racial discrimination in education, housing, and employment. Reform activists mobilize political participation and inform public debate by transforming previously unnoticed social conditions into highly visible social injustices. In this way resource-poor groups compensate for their lack of political organization or low levels of political participation. They use state provided statistical information as a political resource (Lacey and Furner 1993).

3.3 Group Interests

Census information can be the basis on which to establish group identity. When, for instance, the census differentiates the population into its component racial, ethnic, or tribal groups, it promotes these categories in public consciousness and in public policy. To be measured is to be politically and socially visible, and this visibility can become the basis for group representation. Public policies such as affirmative action targeted to India's lower castes or racial minorities in many countries are formulated on the principle of statistical proportionality. That is, there is pressure to ensure that racial or ethnic groups historically discriminated against are represented in high status occupations or the legislative bodies proportionate to their numbers in the general population.

This process, once started, gathers momentum. More and more groups insist on being fully counted in a census. The leaders of these groups then assert political claims on the basis of the counts. A census which is judged to undercount particular groups, such as the homeless or the linguistically isolated, is attacked by advocates for the undercounted population groups. The census and even census methodology can become a political battlefield.

The increasingly central role of statistical proportionality in democratic politics and public policy is welcomed by some but resisted by others. Proponents argue that decades or even centuries of injustice can be

reversed through social policies that correct for a history of discrimination. Because this discrimination was and even remains group-based, it requires group-based measures to overcome it. Others express concern that attaching policy benefits and penalties to persons on the base of their race or ethnicity is to abandon the liberal principle of individual rights in a democracy.

The census and what it measures is of great political interest to both sides of this debate. What is unmeasured cannot easily become the target of public policy, and certainly cannot be used in policies that employ statistical proportionality to achieve social purposes.

4. *The Census as a Subject of Public Policy*

The census, being an instrument of nation building, policy-making, public planning, resource allocation, and democratic governance, will itself be the focus of policy. The overriding policy consideration for a government considering its census is how to strike the proper balance between accountability and autonomy. On the one hand, the census is instituted and funded primarily because it serves government purposes. If it serves no useful purpose it will be underfunded and eventually discarded. On the other hand, the census shares with other scientific enterprises the requirements of professional independence. When a government does not respect the independence of census operations, it places at risk the credibility of the counts and can easily be accused of manipulating census numbers in order to protect itself or enhance its prestige.

This tension between policy accountability and scientific autonomy plays itself out in multiple ways within the policies that govern a census. At one extreme, census results and other official statistics are only very selectively released. For example, China during its Cultural Revolution (1966–76) tended to suppress all statistics or to make them available only through highly controlled government reports. In the former USSR, the 1937 Population Census was never made public, in part to keep from public attention the magnitude of the 1932–3 famine. In less extreme instances, particular statistical series are presented in a manner designed less to inform than to persuade the public. In democracies, for example, the release of inflation rates or poverty measures can be timed to benefit the incumbent party during an election period. Practices such as suppression of data or political manipulation of its presentation and release will of course undermine public confidence in census counts and related statistics.

At the other end of the policy continuum is the assumption that the census is a public good. From this perspective, national statistics serve broad social purposes that reach well beyond the immediate needs of the government in power. This perspective, of course, is compatible with granting considerable independence and autonomy to a census bureau staffed less with political appointees than with professional statisticians.

In most countries, census design and operations exist uneasily in a policy space that simultaneously includes close government oversight, on the one hand, and acceptance of the importance of scientific autonomy, on the other. The weight assigned to one or the other will vary depending on the ideology of the government in power. When the conservative government of Margaret Thatcher came to power in the United Kingdom, an explicit decision was taken to narrow the functions of the statistical service and make it serve specific, immediate policy needs. In the United States, a number of questions previously asked were eliminated from the 2000 census forms because they were not directly related to a federal program. This narrowing of the function of the US census dismayed demographers and sociologists who pointed out that the census was becoming less analytically useful, that is, was less of a public good.

The mission a census is expected to serve—narrow, immediate, and even secret government functions at one extreme or broad public good functions at the other—is the first-order policy issue that confronts a government as it establishes and funds the periodic census. There follow, then, a host of subsidiary policy issues. How are government statisticians expected to balance their duties as civil servants with their obligations to their professional peers? What criteria should be used to select the director of a census operation? How much autonomy is granted a census agency in its ongoing effort to improve methods and concepts, especially in redefining politically sensitive measures in areas such as race and ethnicity or income and poverty? What rules govern the timing and nature of data release? What is the appropriate tradeoff between data confidentiality and respondent privacy, on the one hand, and the society's demand for detailed information, quickly provided, on the other? Should a census agency only provide raw data, or should it have its own analytic capacity and thereby provide reports and interpretations? How can appropriate government oversight of a census operation guard against waste of public money or agency incompetency without itself being accused of partisan intrusion in the census?

All of these and other issues are the proper subject of political debate in the articulation of policies that govern the funding, organization, and implementation of a census. The issues are similar in nature to those that arise in the government sponsorship of science or the arts, each being an activity that requires a high degree of professional autonomy but each also being an arena of great political sensitivity.

With respect to a census, or statistical activities more broadly, thoughtful observers have articulated useful policy principles. In essence these principles

recommend that the statistical products should be designed to serve the needs of multiple political interests and not just those of the government currently in power, and should be sensitive to future policy issues rather than focused only on the present. Success in institutionalizing this principle will maintain both the relevance and the credibility of a census (Seltzer 1994).

5. *The Census as a Subject of Social Science*

The social sciences make heavy use of census data across a wide variety of topics. The quantitative study of society is severely hampered in the absence of census information. It is perhaps surprising, therefore, that studies of the census itself are so few. That is, though the census is everywhere the subject of public policy, it is seldom the subject of public policy analysis.

The exceptions to this general proposition are important, and should not be overlooked. There are historical treatments and case studies of particular censuses that are often rich in analytic detail (Anderson 1988, Anderson and Feinberg 1999, Arowolo and Daramola 1982, Medvedev 1989). There are studies of particular policy issues that emerge in the census context, such as analysis of confidentiality or of the census as an instrument of social control or the role of the census in revenue sharing schemes.

What is generally lacking, however, is social science theory about the emergence, role, function, and consequence of the census. It is largely assumed that the modern nation-state needs a census, else it would not have emerged in nineteenth century Europe and North America. This assumption, however, has not been subjected to investigation in political sociology. Theoretically rich treatments of nation building and citizenship (Bendix 1964) or of the state's effort to improve society (Scott 1998) provide a platform for studies of the political causes and consequences of the census, but with few exceptions (Alonso and Starr 1987) the work is yet to be undertaken.

The political economy literature is similarly deficient in its theoretical treatment of the census as an object of inquiry. It is easy to find assertions that the modern economy requires fine-grained demographic data for making decisions about labor market dynamics, location of retail outlets, types of products that will meet future consumer needs, and so forth. From this latter assertion, it is an easy step to claim that census data are a public good and to invoke microeconomic theory to explain the government's decision to fund a census. Using the United States as a case study, one sustained examination of this hypothesis reports that public good theory offers an insufficient explanation of the rise of the modern census (Kelman 1985).

Neither political sociology nor political economy has yet turned its analytic tools on the census as a feature of the modern state and modern economy. This will likely be corrected. If metaphors about how the 'knowledge society' will replace industrial society are taken seriously, so also is the assertion that information is the emerging infrastructure for the knowledge revolution just as railroads or canal systems were the earliest infrastructure for the industrial revolution. An accurate and credible census is basic to establishing an information infrastructure and thereby requires systematic attention by the social sciences.

See also: Policy Information; Policy Knowledge: Epistemic Communities; Policy Networks; Population Cycles and Demographic Behavior; Population Forecasts; Population Size Estimation and Capture–Recapture Methods

Bibliography

Alonso W, Starr P (eds.) 1987 *The Politics of Numbers*. Russell Sage Foundation, New York

Anderson M J 1988 *The American Census: A Social History*. Yale University Press, New Haven, CT

Anderson M J, Feinberg S E 1999 *Who Counts*. Russell Sage Foundation, New York

Arowolo O, Daramola O (eds.) 1982 *Philosophy of Population Census in Nigeria*. National Population Commission, Lagos, Nigeria

Bendix R 1964 *Nation-Building and Citizenship*. John Wiley & Sons, New York

Cohen P C 1982 *A Calculating People: The Spread of Numeracy in Early America*. University of Chicago Press, Chicago

Domschke E, Goyer D S 1986 *The Handbook of National Population Censuses: Africa and Asia*. Greenwood Press, Westport, CT

Economic Commission for Europe 1992 *Handbook of Official Statistics in ECE Member Countries*. United Nations, Geneva

Kelman S 1985 Why should government gather statistics, anyway? *Journal of Official Statistics* **1**(4): 361–79

Lacey M J, Furner M O 1993 *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge, UK

Medvedev R 1989 *Let History Judge: The Origins and Consequences of Stalinism*. Columbia University Press, New York

Nathan R 1999 The politic of printouts: The use of official numbers to allocate Federal grants-in-aid. In: Alonso W, Starr P (eds.) *Who Counts*. Russell Sage Foundation, New York

Prewitt K 1999 Public statistics and democratic politics. In: Alonso W, Starr P (eds.) *Who Counts*. Russell Sage Foundation, New York

Scott J C 1998 *Seeing Like A State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT

Seltzer W 1994 *Politics and Statistics: Independence, Dependence, or Interaction?* United Nations Department for Economic and Social Information and Policy Analysis, Working Paper Series No. 6. New York

Seltzer W 1998 Population statistics, the Holocaust, and the Nuremberg Trials. *Population and Development Review* **24**(3): 511–22

K. Prewitt

Copyright © 2001 Elsevier Science Ltd. 11573
All rights reserved.

 ISBN: 0-08-043076-7

Policy Knowledge: Contract Research

1. 'Contractual Research' Versus 'Academic Research'

'Contractual research' as a modality of the science-based generation and production of (policy-related) knowledge is basically characterized by the commissioner/producer or consumer/contractor principle: 'the consumer says what he wants, the contractor does it (if he can), and the consumer pays' (to quote Lord Rothschild's much-referred to formula, Wittrock et al. 1991, p. 47). Hence, the 'request for proposal' (RFP) through which the commissioning agency addresses the would-be contractors (in public bidding, selective bidding, or directly), as a rule, defines and specifies the questions to be answered and the time frame. In his project proposal the would-be contractor explains his research plan within the parameters set by the 'customer' and makes his financial offer which is usually calculated on a 'personnel costs plus overheads' formula.

By contrast, academic research has traditionally hinged on the claim to institutional and intellectual autonomy of scientific work. Ideally committed to the 'search for truth,' academic (social) science research is traditionally understood as selecting its subject matter, research questions, and methodology solely by the 'free' decision of the individual researcher on the basis of 'intra-scientific' criteria. The essential frame of reference is the scientific community and its reputation system centering on peer review and peer recognition. Hence, the publication of results is an indispensable vehicle of allowing peer review as well as scientific knowledge accumulation and theory building. The classical locus of academic research is the 'research university,' based on the Humboldtian idea of the 'unity of research and teaching' as well as the (public/quasi-public non-profit) research institutes operating on a 'university without students' logic (Weaver 1989). While academic research is essentially geared to 'basic' research, it may well, through an intra-scientific search for 'relevance,' take some 'applied,' policy-related orientation. The cognitive autonomy of academic research is essentially secured by its 'independent' modalities of research funding, be it through the host institution's own resources ('untied' public money, private donations, endowments, etc.), be it through funding agencies (public or philanthropic foundations) usually operating on a competitive, peer-reviewed application formula.

At this point, it might be useful, in a side remark, to recall that recently the term and concept of 'think tanks' has been occupying the discussion on social science-provided policy-advice and its institutionalization (see Stone et al. 1998). In order to avoid the present use of the inflationary use of the term, 'think tanks' might be understood and defined as research institutes which are institutionally and financially independent and see their task in generating and (most importantly) diffusing policy-relevant knowledge and therefore in exerting an influence on the political debate and agenda setting. 'Think tanks' that assume an explicitly ideological stance have been labeled 'advocacy think tanks' of which the Heritage Foundation in the USA and the Adam Smith Institute in the UK (both acting as neo-conservative/neo-liberal intellectual spearheads) are perhaps the most prominent examples.

This article is meant, within the limits of its space, to do two things. First, in its main body, it will take up a historical-institutionalist approach in 'mapping' the institutional landscape of contractual research (Sect. 2). Second, some conceptual (and ethical) key issues of contract research will be addressed (Sect. 3).

2. Across-countries Commonalities and Variance in Institutionalization of Contractual Research

2.1 Institutional and Cultural Factors Influencing the Institutional Profile ('Who is Who') of Contractual Research

The size and profile of contractual research in a country plausibly depends on institutional and cultural factors both of the political system ('on the demand side') and of the (social) science system ('on the supply side').

The political, governmental, and administrative demand for (external) science-generated policy knowledge may be contingent on the *institutional structure of government* (whether presidential or parliamentary), on the *political culture* (whether government seeks policy knowledge and advice primarily 'in house' in its own governmental apparatus, or at best 'quasi in-house' in governmental research institutes, or whether it is disposed to turn to external research capacities), and on the *administrative culture* (whether the administrative elites see themselves as professionally self-sufficient and secluded or whether they are perceptive for, and receptive to, external science-generated knowledge).

The supply of science-generated policy knowledge and advice may hinge on the *institutional setting of the social science system* (whether it traditionally revolves around 'research universities' (in the Humboldtian sense) or on non-university research institutes) as well as on *research attitudes and cultures* (whether university-based social scientists have the propensity to engage in policy-related knowledge production).

As some of these factors appear deeply, even in a 'path-dependency' manner, rooted in the countries' institutional and cultural fabric, they are likely to bring forth country-specific trajectories in the way contractual research has been institutionalized.

In Sects. 2.2 and 2.3 the attempt is made to elaborate (and explain) some distinct patterns and types of that institutionalization by dwelling on a number of countries as cases in point.

2.2 Emergence Of Contractual Research up to World War II: America's Exceptionalism

The development of the relation between university-based social science and government up to 1945 is a story of 'American exceptionalism' for a number of reasons. When the social sciences emerged at American universities, most of which were privately funded, the explicit attempt was made in the 1920s, spearheaded by the political scientist Charles Merriam at the University of Chicago, to combine the ambition to make them 'truly scientific' with an 'applied,' policy-related orientation. This applied stance of university research was fostered and supported by the early appearance of philanthropic foundations (Rockefeller, Carnegie) which made for another US American particularity and innovation. In this context the Brookings Institution was established in 1927 as a private, independently funded, policy-related institute—another American 'first'. While the federal government abstained from setting up governmental research institutes of its own, it started to draw on external, mainly university-based, social science expertise since the 1920s, opened up by President Hoover and stepped up by President Roosevelt during the New Deal (Wittrock et al. 1991, p. 39). When, during World War II, the federal government made an all-out effort to mobilize the country's entire scientific and technological potential, including the social sciences (systems analysis, psychology, etc.), 'contract research' became the decisive tool by which the federal government succeeded in catalyzing the huge technological, industrial, and intellectual resources of America's private sector which, to a large extent, include the universities (Lieske 2000, pp. 51ff.). In view of the salience of the private sector in the American institutional world it has been aptly remarked that '*contract* is symbolic of the extent that it embodies and represents the values of private enterprise and the "free" market economy that have distinguished much of the nation's history' (Owens 1996, p. 249).

In contemporary European countries the development under consideration was distinctly different. On the Continent governments showed little interest in drawing on external science-generated knowledge. Instead, for instance in Germany, governmental research institutes were put in place for the (quasi in-house) provision of policy-relevant knowledge, While the university-based social sciences, in their struggle for scientific and academic recognition, accentuated their scientific profile and shunned an applied one. Sweden was an exception to this, where since the 1930s a policy-related interaction between government and university-based social sciences has caught root. To some degree this also held true in the UK where reformist Fabians established the London School of Economics as an explicitly policy-related academic institution. It should be added, though, that in the economic field most European governments showed an early interest in eliciting and using external expertise from newly established public and private economic research institutes that were founded in Germany, France, Austria, and Sweden during the 1920s and 1930s (Gellner 1998, p. 90, Wittrock et al. 1991, p. 48).

2.3 The Upsurge of Contractual Research After 1945: Commonalities and Variance

In the following, the development in a number of countries will be highlighted as cases and types in point instead of venturing in a systematic overview.

2.3.1 USA: private for-profit institutes/firms prevail in contractual research. Responding to the mounting policy knowledge demand which came from the post-war role of the US as the hegemonic Western power and from the ever more intensive Cold War confrontation, the US government continued to heavily draw on contractual research, at first particularly in foreign and military policies. The RAND Corporation, founded in 1948, was an institutional innovation in that it was the first private, for-profit research institute which was basically financed through governmental research contracts, in that case particularly from the US Department of Defense (Abelson 1998, pp. 112ff.). During the 1960s, contractual research experienced a tremendous upsurge. The Social Action programs under the Johnson administration and the legislative mandate to have them evaluated unleashed a contractual research money flow from the federal ministries and agencies on an unprecedented scale. At the same time, typical of the US government system, Congress built up its own evaluation capacities (GAO, CBO counting some 6000 staff members) and the federal departments too set up extended 'in-house' evaluation units and personnel (at the cost of some 1.2 billion US$ in 1984). A huge amount of contractual money was put into external evaluation (amounting to some 460 million US$ in 1984, Rist 1990, p. 80); thus, evaluation research became, 'in the words of Wall Street, a growth industry' (Rossi et al. 1999, p. 11). The institutional response to the contractual money market was threefold. First, linking up with their traditional applied and policy-related orientation and experience, university-based social scientists successfully sought access to the contractual research money market, sometimes operating at self-standing (non-profit) research institutes that, while remaining affiliated with their 'mother' university, were founded for the

very purpose of doing contractual research. Second, the (quasi-public or quasi-private) non-profit institutes (such as The Urban Institute, but also the time-honored Brookings Institution) got engaged in contractual research. Third and most conspicuously, a growing number of private for-profit research institutes and consulting firms, doing 'research for business,' were founded and entered the race for research contracts (such as Abt Associates, established in 1966). Sometimes ironically dubbed 'Belt Way Bandits' (after the high way around Washington, DC, where many of them are located) they have succeeded in taking the lion's share of the contractual research money market.

2.3.2 Sweden: university-based research prevails in contractual research. Among the European countries, Sweden also makes for an exceptionalist story of contractual research. Sweden's policy-making is traditionally characterized by the practice of creating, for each major piece of legislation or policy-making, *ad hoc* commissions made up of members of parliament, administrators, representatives of interest groups, and academics; in their mandate to seek agreed upon policy solutions they epitomize Sweden's traditional consensus-seeking political style (Vedung 1992, pp. 74ff.). For preparing their recommendations to government and parliament the *ad hoc* committees dispose of contractual research money to commission policy-related information and expertise. During the 1960s when Sweden experienced a (short-lived) PPBS boom and a 'scientific revolution' in policy-making (Wittrock et al. 1991), the country became one of the European frontrunners in policy evaluation, making evaluation look 'endemic throughout the Swedish system' (Levine 1981, p. 50). Thanks to the traditional policy-related interaction between government and university-based social scientists, the latter have prevailed in the conduct of contractual research.

2.3.3 Germany: public/non-profit and private/for-profit mix in contractual research. Germany presents the case of an (at first) retarded and (then) accelerated development of contractual research. Well into the 1960s, Germany's administrative (law-trained) elite adhered to the traditional belief in its self-sufficiency and, for the rest, relied on the chain of governmental research institutes (Hohn and Schimank 1990, pp. 303ff.) for informational support. At the same time, university-based research, while still absorbed by its post-war reconstruction and expansion, largely cherished its traditional dedication to basic research and continued to be wary of applied research. In the late 1960s, this traditional distance between, if not mutual neglect of, the government and the university-based social science experienced a conspicuous rupture and reversal which was significantly the work of 'reform coalitions' formed by reformist politicians, civil servants, and university-based (social) scientists (Wagner and Wollmann 1991, pp. 74ff.). In conspicuously revamping its governmental machinery under the then dominant planning imperative and drawing, in an attempt at the 'scientification of policy making,' on social science advice on an unprecedented scale (Wollmann 1989), Germany became a European frontrunner, besides Sweden, in evaluation research (Derlien 1990, p. 148, Wagner and Wollmann 1986). The dramatic shift was evidenced by a rapid expansion of the contractual research money market with three groups of researchers competing for the research contracts. First, pursuing their new reorientation to applied and policy-related research, university-based researchers have sought and attained some access to contractual research, partly in institutional 'spin-offs' affiliated with the 'mother' university. Second, quasi-public non-profit research institutes have held a significant share in contractual research. Third, the new sector of private for-profit research institutes and consulting firms has been able to take lion's share of the contractual money market.

2.3.4 France: production of policy-related knowledge still largely in in-house or quasi-in-house manner. France can be seen as an example of the Continental European (State) tradition in which the production of policy-related knowledge has still remained largely within the government's domain in an in-house or quasi-in-house manner. France's administrative (*grands corps*) elites have retained their traditional belief in their professional self-sufficiency (Nioche 1992, p. 31). The Court of Accounts and the General Inspection of Finances constitute another barrier of traditional policy monitoring and auditing agencies. In addition, a chain of quasi-governmental research institutes, subordinated to the central ministries, has been established to provide ancillary policy information and advice. On the science system side, France has traditionally known the functional division between the teaching role of the universities and the research-focused CNRS whose close ties with the State have been seen as 'making them an integral part of the state apparatus' (Fieschi and Gaffney 1998, p. 54). Hence, France has been a latecomer to policy evaluation (Nioche 1992, p. 24); as late as in 1990 a typically centralized evaluation procedure was institutionalized, comprising, *inter alia*, a Scientific Council of Evaluation made up of academics (Duran et al. 1995, pp. 55ff.). While the contractual research money market is comparatively small, it has been dominated by CNRS institutes which quite often have teamed with university researchers. The share of private for-profit institutes and consulting firms has so far been quite modest.

2.3.5 Japan: private for-profit research institutes prevail in contractual research. While Japan's governmental bureaucracy has traditionally cherished a culture of self-sufficiency and elitist seclusion, it was also seized, during the 1960s, by a (passing) PPBS boom which resulted in the establishment of planning and analytical units in central ministries (Ide 1969) and in growing demand and contractual funding for external analyses. In 1974 the quasi-governmental National Institute for Research Advancement (NIRA) was established to initiate, coordinate, and monitor the contractual money research. On the 'supply side,' the university-based social science research has shown little propensity for applied and policy-related work (Watanuki 1991), not least because the universities have traditionally been under-equipped for research purposes. Thus, contractual research has been procured and conducted almost exclusively by private research institutes. Out of the some 225 private research institutes (as of 1992) some 40 percent are for-profit (Ueno 1998, p. 195). While the largest ones (such as Mitsubishi Research Institute and Nomura Research Institute, founded in 1970 and 1975, respectively) have been established by their 'parent' corporations, most of the smaller ones have typically been established by former (well-connected) high-ranking civil servants after their early retirement as well as by university professors who combine teaching at the university with doing paid contractual research outside.

2.3.6 European Union (EU): dynamic contractual research money market dominated by the private for-profit research sector. While EU policy-making lies largely with the European Council and the Commission, the implementation of EU policies (that is, EU regulation and EU funding) is carried out by the member states and their (for the most part regional and local) bureaucracies. Thus the EU central actors, particularly the Commission, have a 'natural' need for external policy information. With regard to EU structural funding, for instance, the procurement of such information has been secured by a technically sophisticated evaluation system consisting of *ex ante*, interim, and *ex post* evaluation and based almost entirely on external evaluation. Consequently the EU has, since the late 1980s, opened up and has massively funded a contractual research money market. The evaluation of EU policies and programs is conducted on two levels. On the EU central level the evaluation contracts (amounting to 13.5 million Euro in 1999) relate to the entire EU programs and have been won and carried out exclusively by for-profit institutes and consulting firms among which big international firms (like Anderson, Ernst and Young and the like) loom large. Another EU contractual money flow is directed at evaluating the implementation of the EU programs in the individual member countries. As these projects are smaller and require some familiarity with the national, regional, if not local, contexts, 'national' research institutes have prevailed in gaining the research contracts. The profile of the contractors differs from country to country. While, for instance, in Spain university-based research groups have the Spain-related evaluation contracts, in most other countries the majority of the contractors are private for-profit. The contractual research money flow of the EU is bound to have significant repercussions not only on the contract research scene, but also on the social science community at large. This holds true particularly in South European member countries (Greece, Spain, Portugal, and also Italy) which receive the bulk of EU funding and which, consequently, are the target area of correspondingly massive evaluation research. There is evidence that, in fact, EU contractual research funding has given strong incentives and impulses to the national research scene in these countries.

3. Contractual Research Constituting a Threat to the (Intellectual Etc.) Integrity of the Individual Researcher or Even of the Social Science Research Community at Large?

The expansion of contractual research has been accompanied, both within the contractual research camp as well as on the academic scene, by a debate about the threats which the crucial consumer/contractor constellation may pose to the intellectual (and moral) integrity of the researchers involved and beyond. It is particularly the basic commissioning and funding logic of contractual research, described in the 'Rothschild principle' as 'the consumer says what he wants, the contractor does it (if he can), and the consumer pays,' that is being suspected of jeopardizing and undermining its intellectual integrity and of translating into reality the adage 'who pays the piper calls the tune'. First of all private for-profit institutes and firms that seek and conduct research contracts as 'research for business' are seen exposed to an erosive, if not corruptive mechanism (Ham 1999, p. 276). As a spill-over from the realm of contractual research onto social science research at large, including academic research, even a 'colonization process' has been foreboded in the wake of the external setting of the research agenda, 'whereby the bureaucracies' perspective and conceptual framework (is) established as the valid epistemological framework also for scientific research' (Elzinga 1985, p. 211).

In the face of such challenges to the intellectual integrity and honesty of contractual research, initiatives have been taken, particularly among evaluators as probably the most numerous and most

professionalized group among those engaged in contractual research, to formulate and lay down standards that could guide them in their contractual work and in negotiations with their 'clients' (Rossi et al. 1999, pp. 425ff.). Exemplary are the *Guiding Principles of Evaluation* that were adopted by the American Evaluation Association in 1995. Among its five principles the maxims of integrity and honesty of research ('evaluators must ensure the honesty and integrity of their entire evaluation process') are writ large (full text in: Rossi et al. 1999, pp. 427ff.).

Bibliography

Abelson D E 1998 Think tanks in the United States. In: Stone D, Denham A, Garnett M (eds.) *Think Tanks across Nations*. Manchester University Press, Manchester/New York, pp. 107–26

Derlien H-U 1990 Genesis and structure of evaluation efforts in comparative perspective. In: Rist R (ed.) *Program Evaluation and the Management of Government*. Transaction, New Brunswick/London, pp. 147–77

Duran P, Monnier E, Smith A 1995 Evaluation à la Française. *Evaluation* **1**(1): 45–64

Elzinga A 1985 Research, bureaucracy and the drift of epistemic criteria. In: Wittrock B, Elzinga A (eds.) *The University Research System*. Almqvist and Wiksell, Stockholm, Sweden, pp. 191–220

Fieschi C, Gaffney J 1998 French think tanks in comparative perspective. In: Stone D, Denham A, Garnett M (eds.) *Think Tanks across Nations*. Manchester University Press, Manchester/New York, pp. 42–58

Gellner W 1995 *Ideenagenturen für Politik und Öffentlichkeit: Think Tanks in den USA und in Deutschland*. Westdeutscher Verlag, Opladen, Germany

Gellner W 1998 Think tanks in Germany. In: Stone D, Denham A, Garnett M (eds.) *Think Tanks across Nations*. Manchester University Press, Manchester/New York, pp. 82–106

Ham V 1999 Tracking the truth or selling one's soul? Reflections on the ethics of a piece of commissioned research. *British Journal of Educational Studies* **47**(3): 275–82

Hohn H-W, Schimank U 1990 *Konflikte und Gleichgewicht im Forschungssystem*. Campus, Frankfurt/New York

Ide Y 1969 Administrative reform and innovation: The Japanese case. *International Social Science Journal* **1**: 56–76

Leeuw F L, Toulemonde J, Brouwers A 1999 Evaluation activities in Europe: A quick scan of the market in 1998. *Evaluation* **5**(4): 487–96

Levine R A 1981 Program evaluation and policy analysis in western nations: An overview. In: Levine R A, Solomon M A, Hellstern G-M, Wollmann H (eds.) *Evaluation Research and Practice*. Sage, Beverly Hills, CA, pp. 27–60

Lieske J 2000 *Forschung als Geschäft*. Campus, Frankfurt/New York

Nioche J-P 1992 Institutionalising evaluation in France, skating on this ice? In: *OECD 1977, Review of Social Science*. OECD, Japan, Paris

Owens L 1996 MIT and the federal 'Angel': Academic R&D and federal-private co-operation before World War II. In: Numbers L, Rosenberg C E (eds.) *The Scientific Enterprise in America*. Chicago, pp. 247–72

Rist R (ed.) 1990 *Program Evaluation and the Management of Government*. Transaction, New Brunswick/London

Rossi P H, Freeman H E, Lipsey M W 1999 *Evaluation. A Systematic Approach*, 6th edn. Sage, Thousand Oaks

Stone D, Denham A, Garnett Mark (eds.) 1998 *Think Tanks across Nations. A Comparative Approach*. Manchester University Press, Manchester, UK

Ueno M 1998 Think tanks in Japan: Towards a more democratic society. In: Stone D, Denham A, Garnett M (eds.) *Think Tanks across Nations*. Manchester University Press, Manchester/ New York, pp. 188–201

Vedung E 1992 Five observations on evaluation in Sweden. In: Mayne J, Bemelmaus-Videc M L, Hudson J, Conner R (eds.) *Advancing Public Policy Evaluation*. Elsevier, Amsterdam, pp. 23–36

Wagner P, Wollmann H 1986 Fluctuations in the development of evaluation research: Do 'regime shifts' matter? *International Social Science Journal* **38**(2): 205–18

Wagner P, Wollmann H 1991 Beyond serving state and bureaucracy: Problem-oriented social science in (West Germany). *Knowledge and Policy* **4**(1,2): 56–88

Watanuki J 1991 The impact of social sciences on the process of development in Japan. In: Wagner P, Weiss C H, Wittrock B, Wollmann H (eds.) 1991 *Social Sciences and Modern States*. Cambridge University Press, Cambridge, UK, pp. 221–9

Weaver K 1989 The changing world of think tanks. *Political Science and Politics* **22**(3): 563–78

Wittrock B, Wagner P, Wollmann H 1991 Social science and the modern state. In: Wagner P, Weiss C H, Wittrock B, Wollmann H (eds.) *Social Sciences and Modern State*. Cambridge University Press, Cambridge, UK, pp. 28–85

Wollmann H 1983 *Development and State of Policy Research in Japan*. National Institute for the Advancement of Research, Tokyo

Wollmann H 1989 Policy analysis in West Germany's Federal Government Governance. *An International Journal of Policy and Administration* **2**(3): 233–66

H. Wollmann

Policy Knowledge: Epistemic Communities

Political science and policy studies have increasingly developed an interest in the role of ideas, values, and technical understanding in shaping political outcomes, particularly under conditions of perceived complexity. 'Epistemic communities' is a concept developed by 'soft' constructivist scholars of international relations concerned with agency to understand the actors associated with the formulation of ideas, and the circumstances, resources and mechanisms by which new ideas or policy doctrines get developed and are introduced to the political process.

1. Constructivism

Constructivism analyzes the social process by which actors construct meaning of the world through the application of broad ideas and reasoning patterns. Thus, constructivists specify how actors come to

recognize interests and make choices (Ruggie 1998, Haas 1992/1997). While there has been a long-standing appreciation of the importance of ideas—the fabric with which Weberian webs of significance are spun—on politics and on choice, the mechanism by which such constructions arise and persist has long remained elusive. Epistemic communities are important actors responsible for developing and circulating causal ideas and some associated normative beliefs, and thus helping to create state interests and preferences, as well as helping to identify legitimate participants in the policy process and influencing the form of negotiated outcomes by shaping how conflicts of interest will be resolved. Attention to epistemic communities provides a way to understand agency in politics and policy formation.

Assigning context and determining appropriate behavior—and thus policy—are fundamentally political processes, both in terms of actors' efforts to directly influence the process and in terms of their distributional consequences. Most IR theorists now agree that contemporary politics and international relations are characterized by conditions of complexity, as a consequence of growing interdependence and globalization. Complexity complicates decision-making. Substantive and procedural rationality are impossible under high degrees of systemic complexity (Simon 1983).

Constructivists rely on discussions of complexity to conclude that actors are unable to easily deduce preferences and understanding of options from contemporary circumstances. States and other actors are uncertain about their ultimate goals, and how to achieve them. Even whether actors choose to apply a logic of consequences or of appropriateness is itself conditioned by their prior understanding of the nature of the political circumstances in which they find themselves (March and Olson 1998). The very nature of political strategies actors deem appropriate is a consequence of the advice and framing that actors receive about the uncertain policy environment in which they find themselves, or what John Searle calls establishing that 'x counts as y under conditions c' (Searle 1995).

For international relations, more specifically, actors are uncertain about how to achieve particular interests in given circumstances (their preferences); with whom they have to interact to achieve their goals, and by what means to do so. For instance, when first confronted with new issues political actors are typically uncertain about whether to pursue their ends unilaterally or multilaterally, through which institutions, which nests to apply, which linkages are relevant, and what policies should be collectively embraced. For the policy studies literature (Ascher 1986, Thomas 1997, Social Learning Group 2001) the principal analytic focus that results from an appreciation of complexity and the role of ideas and epistemic communities is on the stages of policy making: framing, agenda setting, social learning, participation, policy choice, and, to a lesser extent, compliance or enforcement.

National and international politics and policy choice then becomes a matter of applying embedded and institutionalized beliefs about the nature of problems and the appropriate means of collective response as argued by many current theorists of IR (Ruggie 1998, Adler 2001, Haas 1990, 1992/1997), rather than the process of resolving rationally formulated state preferences. In this conception politics becomes a process of learning about the world.

Thus, ideas may play important roles in framing the policy debate, and in predisposing outcomes highlighted by each perspective. Epistemic communities are the cognitive baggage handlers of constructivist analyses of politics and ideas. They are invoked by Constructivist scholars to focus analytic attention on the process by which states formulate interests and reconcile differences of interest. Epistemic communities are a principal channel through which consensual knowledge about causal connections is applied to policy formation and policy coordination. States may learn through processes of international cooperation mediated by epistemic communities. As a consequence collective patterns of behavior reflect the dominant ideas initially brought to bear to understand the area of activity which is being governed.

Ruggie borrowed *episteme* from Foucault, to describe the overarching perspective through which political relationships are visualized and understood during historical eras (Ruggie 1975). The winter 1992 issue of *International Organization* on Knowledge, Power and International Policy Coordination refined the concept by restoring agency to the study of politics and knowledge, by focusing on the actors responsible for articulating and aggregating knowledge based understanding in areas of security, environment and international political economy (Haas 1992/1997). Epistemic communities are networks—often transnational—of knowledge based experts with an authoritative claim to policy relevant knowledge within their domain of expertise. Their members share knowledge about the causation of social or physical phenomena in an area for which they have a reputation for competence, and a common set of normative beliefs about what actions will benefit human welfare in such a domain. In particular, they are a group of professionals, often from a number of different disciplines, who share all of the following characteristics:

(a) Shared consummatory values or principled beliefs. Such beliefs provide a value-based rationale for social action by the members of the community.

(b) Shared causal beliefs or professional judgment. Such beliefs provide analytic reasons and explanations of behavior, offering causal explanations for the multiple linkages between possible policy actions and desired outcomes. While they need not agree on every element—in fact they are likely to disagree vehemently

about some elements—they do agree about the core assumptions and causal forces from which their models of the world derive and agree on the means by which such differences can potentially be reconciled.

(c) Common notions of validity: intersubjective, internally defined criteria for validating knowledge.

(d) A common policy enterprise: a set of practices associated with a central set of problems which have to be tackled, presumably out of a conviction that human welfare will be enhanced as a consequence.

This combination of factors—especially the socialized truth tests and common causal beliefs—distinguish epistemic communities from other types of policy networks and groups active in politics and policy-making. Unlike other organized interest groups active in politics and policy making, epistemic communities are bound by the truth tests to which they were socialized, and thus are more likely to provide information that is politically untainted, and thus more likely to 'work,' in the political sense that it will be embraced and followed by political authorities concerned about the need for appearing impartial, and also technically.

2. *Cognate Literatures*

The policy sciences, since Lasswell, have stressed the difficulty of applying strong rationality assumptions when conducting policy analysis, as well as advocating a neo-pluralistic approach to understanding the politics of policy making (Ascher 1986, Braybrooke and Lindblom 1963, Healy and Ascher 1995). Discursive approaches to understanding the policy sciences reinforce constructivist studies of politics and policy (Linder and Peters 1995, Stone 1989).

Similar concepts to epistemic communities were developed in other social sciences in the 1980s and 1990s to better understand the role of expertise in constructing meaning for policy makers and for developing policy usable knowledge. Knorr-Cetina, in the sociology of knowledge, coined the term 'epistemic communities' (Knorr-Cetina 1981, 1999), although this neologism was not read by the international relations research program until after the 1992 publication of the epistemic communities volume.

In comparative politics and policy studies complementary notions have been put forward to study networks of experts involved in policy making. Richardson coined the term policy communities (Richardson 1995). Reinicke has extended policy communities to include global policy networks (Reinicke and Deng 2000). Sabatier developed the notion of policy advocacy coalitions, which were an extension of earlier US work on iron triangles (Sabatier and Jenkins-Smith 1993). The concept of policy communities stresses the causal beliefs shared by policy analysts involved in policy development, but do not fully develop a causal model of their development and spread. Analysts of advocacy coalitions note the presence at different points in time of people from multiple interest groups and organizations, but do not address their beliefs systematically, nor do they look at the causal role of a kernel of individuals sharing beliefs and involved in the same policy enterprises over time. Analysts of epistemic communities, with their focus on agency and the shared beliefs of a community that shares a common policy activity over time, provide the leverage for generating social science knowledge about the development and diffusion of knowledge for policy and politics.

3. *A Constructivist Model of Institutionalized Ideas*

The development and application of an epistemic community's consensual knowledge can be best understood within a stepwise framework. Such a framework provides independent mechanisms by which new ideas are generated, selected by authoritative actors, and persist over time. It is a path-dependent model of punctuated equilibrium because many choices or decisions are largely irreversible and the array of possible choices at time (t) is causally related to decisions taken at time ($t-1$) in ways that may not be perceptible to actors at time $t+1$ (Adler and Haas 1992/1997). Ideas may exercise a causal impact later in time regardless of the reasons for their initial selection, as they continue to shape actors' understanding and guide decisions even in the absence of original conditions that contributed to their widespread acceptance.

While actors are always constructing the world, to some extent, the challenge to analysts is to identify what constructions are likely to prevail, and under what circumstances. Since current constructions are made in the shadow of past constructions, this framework provides the means to specify which ideas will be influential and the factors influencing the likelihood that they will exercise enduring influence. Unique collective outcomes involving learning are possible with the existence and involvement of an epistemic community to focus beliefs. Otherwise, political behavior remains a matter of applying prior frames to understanding political choices and options, with conflicts of interest resolved through the application of existing institutions and power resources.

4. *Constructivist Assumptions*

A number of key assumptions underlie this constructivist model. States are key juridical actors, and act willfully. States are functionally differentiated, and vary widely according to their state/society relations

and the technical capacity of the state to formulate and enforce public policies in technical domains (Risse-Kappen 1995, Katzenstein 1977, Gourevitch 1996, Evans et al. 1993). States operate in a complex policy environment, so that information about the environment is highly valued. Yet knowledge is asymmetrically distributed within and between countries. Finally, while perceptions are socially constructed, political actors and social scientists are not subject to the same set of influences. Thus social scientists may identify social forces influencing actors' choices without being equally prone to their influence.

A number of propositions about state behavior and its explanation immediately follow from these assumptions. States are uncertain about their policy environment and how their interests are to be recognized in particular instances, and thus they exercise bounded rationality. The recognition of threats and the identification of responses are subject to socialized patterns of reasoning, rather than through formal rationality in which actors have complete information and choose the most efficient alternatives. Decision-making operates according to a satisfying model in which actors express procedural rationality, at best, subject to the time constraints and structural blinders imposed by bounded rationality (Simon 1983, 1985, March 1994). Consequently, decisions and institutions are sticky, since decision-makers do not return to issues until there is a crisis.

States make choices subject to multiple sources of influence, whose organization varies by issue area (Keohane and Nye 2001). Thus, governance varies by issue area. Consequently, policy networks, organized around specific issues become the appropriate level and unit of analysis, because the array of actors, interests, institutions, and capabilities varies by issue. Even within a particular state, the design of policy-making institutions and procedures vary by issue area. Consequently, it is not surprising that international regimes in related areas can differ widely, because the articulation of interests for each regime takes a different form.

5. *Generation of New Ideas*

Epistemic communities are the agents by which policy relevant information is developed and transmitted. They are likely to be found in substantive issues where scientific disciplines have been applied to policy oriented work. Thus, epistemic communities arise in disciplines associated with natural science, engineering, ecology, and even economics. They will also be most likely to emerge in countries with well-established institutional capacities for public administration, and science and technology. They are also most likely to be present in rich (measured as a high ratio as share of GDP, and with a significant number of scientists and engineers in the country) democracies. Democracies are most likely to encourage open-ended, spirited discussions because the knowledge enterprise is independent of the state, and yet the state has the resources to access and support science.

For governments to defer to new knowledge, the presumptive knowledge claim must be regarded as authoritative. Authoritative knowledge comes from groups that command legitimacy and authority in society. Epistemic communities' reputation for authority rests on their expertise as well as their reputation for impartiality. If not seen as being impartial their advice will be widely contested and lack any independent political influence.

Such pockets of independent research are most likely to emerge when the conduct of research is seen to be far from the direct interest and influence of the sponsoring sources of the research (Dasgupta and David 1994, David 1995). Its impartiality and accuracy depends heavily on the ability to publish in peer-reviewed journals, as well as the internal commitments to truth tests by epistemic community members. A distance from research sponsors improves the actual and perceived impartiality of epistemic community members as it is less likely to be shaped by needs of the state or sponsoring agency, and will be less politically tainted. Activities most likely to be regarded as impartial are those with relatively low costs that are not directly tied to the mission of the sponsoring agency. Epistemic communities are likely to offer policy advice that will be effective, unlike the more common process of policy advice which corresponds to specific interests, because the policy proposals are developed far from the political influence of sponsoring agencies and will be more acceptable to the public when the ideas do not appear politically compromised.

6. *Selection*

New ideas will be solicited and selected only after crises, for crises will alert politicians to the need for action and will seek to gather information about their interests and options. Crises will trigger agenda setting and the search for new information.

Ideas will be selected by states and international institutions. However, not all states or international institutions are equally likely to consult members of an epistemic community or defer to their advice. Institutions with strong science and technology capabilities, and those representing pluralistic societies will be most likely to be in the first wave of those embracing new ideas. It is there that the scientific culture of epistemic communities will have a close affinity, and new groups will be able to quickly articulate new policy initiatives. Selection will probably take longer in non-democratic societies with technically strong

states that have a strong domestic scientific capacity because it takes longer for ideas to percolate up from society to the government outside of routinized channels.

Different styles of policy selection occur in different political systems, due to organizational and cultural differences between types of democracies (Weaver and Rockman 1993, Katzenstein 1977). In presidential systems, with multiple checks and balances, the flow of new ideas to the state will emerge from think tanks, universities, and governmental labs; and will be solicited directly by Congressional Committee staff, the judiciary and executive branch bureaucrats. In Parliamentary systems the sources of new ideas are more limited, and much more likely to emerge from party think tanks and be deployed through political parties to ministers and suitably socialized civil servants. At times ideas may go from universities and independent think tanks to mid level civil servants, but there are fewer universities and independent think tanks in parliamentary systems because of the predominant influence of political parties.

International institutions will also vary in their propensity to take on board new ideas. In a study of international institutions that selected new programmatic ideas from epistemic communities leading to a redirection of the core mission of the institution, Haas and Haas (1995) found a number of common institutional properties among them. Their governing bodies were not split by major political disagreements, they were led by skilled institutional heads, they had skilled professional staffs recruited on merit, and the institution maintained a porous connection with external sources of information. Few institutions have these properties. Those that do were created by their initial funders and sponsoring states to have such resources. The postwar settlement institutions had to reflect the geopolitical goals of the superpowers, and be consistent with overarching values and widely shared norms about economic management because decisions were made at the top levels of government in the wake of such system wide crises, and these policy makers were alert to consistencies between perceived state imperatives and institutional design. Institutions that were established in the interstices of the Cold War, when state leaders were paying less attention to perceptions of systemic imperatives, have greater prospects for flexibility when they reflect domestic sentiment in the superpowers, and they were created during an absence of widespread geopolitical schisms.

There are several alternatives regarding the likelihood of which ideas will be selected at a moment of crisis. If there is only one new candidate, there is little problem, and the new aspirant will be selected. If there are many, choices may be conditioned by calculations of relative political gain from each idea. Studies of trade policy and the use of ideas in formulating international trade regimes have offered a number of propositions about conditioning factors influencing the selection of competing policy ideas (Odell 1990, Goldstein 1993, Garrett and Weingast 1993). Ideas that provide a potential for building coalitions will be favored over those that do not. Simple ideas that have perceived practical utility will be chosen over more complex approaches. In either case, new ideas will only endure if they are loosely commensurate with deeper-seated beliefs and do not endanger strategic political alliances.

7. *Diffusion*

Over time ideas will be diffused. Whereas they were selected initially by deferring to epistemic communities and recruiting their members by states and international institutions that were seeking new information, the diffusion of ideas occurs by many channels that include direct persuasion, demonstration effects, and inducements and incentives.

Diffusion may occur by institutional learning, as groups apply evolving consensual knowledge to manage institutional practices. Learning occurs directly, through interpersonal persuasion, communication, exchange and reflection, that leads to the recognition or appreciation of new causal models and shared values. Leaders are socialized to accept new views and to empower their expositors.

One other diffusion mechanism is administrative recruitment, as epistemic community members or their confidantes replace officials informed with alternative perspectives. Diffusion may also occur by other mechanisms, as actors alter their behavior subject to the influence wielded by institutions that embody the ideas (those that selected an epistemic community and its ideas earlier).

International regimes informed by institutionalized ideas promulgated by strong states and by international institutions will have the effect of disseminating the ideas to other parties in which epistemic communities were not present, and through their socialization and exposure to new ideas presented during regime negotiations.

Diffusion may also occur directly as a consequence of incentives and inducements to adopt those ideas coming from states, international institutions, and international regimes. These incentives and inducements typically take the form of technology assistance and financial transfers. These will only promote diffusion to states with weak capacity, for whom the provision of such resources would be a powerful incentive to change policy. Demonstration effects by multilateral financial institutions will induce borrowing states to adopt the policy being demonstrated.

States may draw lessons from one another's experiences, and thus borrow policies and ideas from countries that they are trying to emulate (Rose 1993). This process of diffusion is not one of learning,

though, it is one of mimesis or borrowing because the lesson drawers are not adopting the policies based on the persuasive power of reason by an advocate of the policy. Rather they are imitating the actions of someone who commands respect and appears successful.

These diffusion mechanisms will operate along different time frames. While some effects may be felt immediately through new recruitment decisions, and persuasion, broader shifts in public opinion and societal effects may occur on the order of decades. As these processes overlap, the general diffusion of an epistemic community's ideas will accelerate.

8. *Persistence*

Once in place, ideas are likely to persist once they acquire a taken-for-granted element and as states and their diplomats become socialized to institutionalized regimes and practices generated by the application of new ideas. Eventually ideas get converted to domestic laws, and thus become enforced out of habit and become the basis for policy enforcement by the state. It also becomes politically costly to reverse such practices as new interest groups and policy communities mobilize around them after recognizing that material gains are possible from the application of the new ideas. Ultimately new markets may form through the application of these ideas, and thus generate economic incentives for their continued adoption once economic actors associate their profits with the application of such ideas.

9. *Empirical Examples*

9.1 *International Political Economy (IPE)*

Analysts of IPE have identified ideational dynamics in the popularization and dissemination of major economic doctrines. For instance, the repeal of the corn laws occurred after the conversion of Prime Minister Peel, in the face of likely political defeat. The adoption of liberal economic policy occurred following the creation of an influential epistemic community led by Richard Cobden able to persuade top levels of government that the national interest could be advanced by low tariffs, even in the face of economic losses to a powerful political constituency expressing particularistic protectionist interests. England's hegemonic power was then deployed to convert Europe to free trade through the 1860 Cobden Chevalier Treaty, and the US through bilateral treaties, and most European governments kept tariffs relatively low in accordance with the institutionalized ideas until the 1890s, despite a lengthy and deep economic depression.

Keynesianism spread internationally through a similar process of epistemic persuasion in a core set of countries, followed by the creation of international institutions to legitimate and create international structures based on Keynesian dictates. Keynesianism first caught on in England, then traveled to the US through academic networks of economists. After its adoption at the highest levels in the two countries, it provided the intellectual foundations for the Bretton Woods System, through which it was institutionally embedded for decades.

Structuralism and Development Economics was developed in ECLA in the 1940s, and then embedded in ECLA and UNCTAD, from where it continued to serve as development orthodoxy to a number of developing countries throughout the 1970s, as well as providing the economic justification behind the New International Economic Order reforms of the 1970s. Odell tells how a group of 'floaters' in the US treasury department influenced ending the Bretton Woods fixed exchange rates by designing its replacement (Odell 1982). The Washington Consensus on macroeconomic management has guided international public debt rescheduling during the 1990s.

9.2 *International Environmental Protection*

Although multilateral economic activities have been governed, since World War II, by the application of consensual liberal ideas about free trade, efforts to control the impact of economic activity on transboundary and environmental pollution has been based on non market and regulatory principles. Since 1972 an ecological epistemic community, articulating a new ecological management doctrine, has institutionalized its ideas in state policies and practices, in the programatic activities of international institutions, and in international regimes. As a consequence of the international institutionalization of its ideas, states were caused to undertake more comprehensive styles of environmental management for transboundary and global environmental threats, leading to selective improvements in environmental quality (Haas 2001).

A new environmental management doctrine based on ecological principles emerged in the 1960s. The ideas were developed outside the scope of most state interests and are generally regarded as being relatively uncompromised by political and institutional influence, as ecology was a relatively cheap research enterprise and its subject of study was far removed from most state interests. Members of the ecological epistemic community subscribed to holistic ecological beliefs about the need for policy coordination subject to ecosystemic laws (Haas 1990). Their ideas about

ecological management were based on a systems perspective of environmental and social systems. Ecological management proposals were based on setting comprehensive environmental standards based on conservative estimates of the ability of ecosystems to sustain stress, subject to their technical understanding of the behavior of particular ecosystems. They promoted international environmental regimes that are grounded on policies which offered coherent plans for the management of entire ecosystems, sensitive to interactions between environmental media (such as air and water), sources of pollution, and contending uses of the common property resource, rather than being limited to more traditional policies for managing discrete activities or physical resource spaces within fairly short term time horizons. They proposed treaties in which bans and emission limits were set for multiple contaminants, with environmental standards for each contaminant set according to scientific understanding about its environmental impact and its interactive effects with other contaminants.

There were few ideational competitors. Resource management bodies had traditionally been staffed by neoclassical economists and resource managers, who had been discredited by broadly publicized environmental disasters and the energy crisis of the 1970s as well as the limits to growth debate, which they had been unable to predict, and attendant popular fears of widespread resource depletion.

Following the politicization of international environmental issues in the late 1960s galvanized by widely publicized environmental disasters occurring in global commons, the United Nations convened the United Nations Conference on the Human Environment (UNCHE) in 1972. As a consequence of these new domestic demands, and in order to prepare national papers for UNCHE, governments created new environmental agencies. Most state officials were unfamiliar with environmental threats and were unable to rank order environmental threats, amounts of national emissions, what would constitute safe or dangerous concentrations of possible contaminants, and what were appropriate policies to reduce emissions. Many of these national agencies recruited epistemic community members to serve as officials or as consultants.

International organizations were created as well. The United Nations Environment Program (UNEP) was established in 1973 and was staffed principally by young epistemic community members eager to put their professional knowledge to work. The UN Economic Commission for Europe (UNECE) played a strong role in managing European air pollution. The UNECE's environmental unit was led by a former UNEP official who carried the ecological management ideas from UNEP to UNECE. After 1987 the World Bank also became active in international environmental matters. Part of the environmental reforms introduced at the World Bank was the recruitment of ecological epistemic community members and their assignment to key posts to evaluate the environmental consequences of development projects.

Diffusion occurred principally through efforts of these major environmental IOs and through environmental regimes. The IOs encouraged other IOs to internalize environmental concerns into their missions through joint projects, and encouraged governments to pursue more comprehensive environmental policies through public education campaigns, publicizing environmental monitoring findings, resource transfers, and demonstration effects.

International regimes have been even more important for the diffusion of ecological management techniques. This spread is due to the growing influence of ecological epistemic communities and the growing confidence and influence of strong international institutions informed with their ecological beliefs.

Ecological epistemic communities, often working with UNEP, helped draft international environmental regimes governing marine pollution, acid rain, stratospheric ozone protection, wetlands protection, protecting migratory species, polar bears protection, and the preservation of Antarctica. An increasing proportion of them are now based on the comprehensive ecological approach promoted by the ecological epistemic community. In 1973, three of 11 international environmental regimes were based on ecological management styles. In 1985, seven of 22 were comprehensive. In 1995, 10 of 25 were comprehensive (Haas 2001). The application of ecological management ideas to environmental regimes spanning a number of geographic areas and functional activities means that most states have accepted ecological obligations for governing a wide variety of human activities.

Ecological practices based on ecological management ideas have become locked in through a variety of mechanisms. Following ratification of international regimes, governments enforce these obligations domestically. Ecological practices get institutionalized domestically through legal precedents, bureaucratic standard operating procedures, and policy enforcement. In many countries they have acquired domestic constituencies—composed of lawyers and civil engineers who subcontract services, firms selling pollution control technologies, and environmental NGOs—that contribute political pressure for continued state enforcement of policies grounded on these ideas.

Exogenous forces have also reinforced these state commitments. A Gallup poll conducted in 1992 demonstrated new worldwide concern about the need to respond to global and transboundary environmental threats (Dunlap et al. 1992). In democratic societies public demands for environmental protection reinforce the influence of the epistemic community's ideas. International markets for pollution control technology also came into existence in the 1990s.

10. Further Research

Epistemic communities is a core concept in a progressive constructivist research program. Its application in international relations has restored agency to the study of ideas, improved the understanding of the sources of state interests under conditions of uncertainty and complexity, and expanded an understanding of regime dynamics. In policy studies it fleshes out with accounts of agency the insight that policy debates are discursive rather than rational, and thus 'best' policy analytic suggestions do not always prevail. It provides the potential for understanding the mechanisms by which knowledge change contributes to policy change.

Further work needs to be done to integrate this work into institutional studies, to better understand the character of crises that are capable of generating political upheavals and discrediting institutionalized ideas as well as how ideas become institutionalized at the national and international levels. Research opportunities beckon on the creation, mobilization and maintenance of epistemic networks. Further prospects lie in studies of variation in generation and selection of epistemic communities, and how they communicate with the public. Applications of epistemic communities' research would contribute to the development and organization of usable technical knowledge for improved policy debates on contemporary issues.

See also: Constructivism/Constructionism: Methodology; Decision Making, Psychology of; Environmental Planning; Environmentalism: Preservation and Conservation; International Organization; Organizational Behavior, Psychology of; Organizational Climate; Organizational Decision Making

Bibliography

Adler E 2001 Constructivism in international relations. In: Carlsnaess W, Risse T, Simmons B A (eds.) *Handbook of International Relations*. Sage, Beverly Hills, CA

Adler E, Haas P M 1992/1997 Conclusion. In: Haas P M (ed.) *Knowledge, Power, and International Policy Coordination*. University of South Carolina Press, Columbia, SC

Ascher W 1986 The evolution of the policy sciences. *Journal of Policy Analysis and Management* **5**(2): 365–89

Braybrooke D, Lindblom C E 1963 *A Strategy of Decision; Policy Evaluation as a Social Process*. Free Press, New York

Dasgupta P, David P A 1994 Toward a new economics of science. *Research Policy* **23**: 487–521

David P A 1995 Reputation and agency in the historical emergence of the institutions of 'open science.' Mimeograph, Stanford University, Stanford, CA

Dunlap R E, Gallup G H Jr, Gallup A M 1992 *Health of the Planet Survey*. Gallup International Institute, Princeton, NJ

Evans P B, Jacobson H K, Putnam R D (eds.) 1993 *Double-edged Diplomacy*. University of California Press, Berkeley, CA

Garrett G, Weingast B 1993 Ideas, interests and institutions. In: Goldstein J, Keohane R O (eds.) *Ideas and Foreign Policy*. Cornell University Press, Ithaca, NY

Goldstein J 1993 *Ideas, Interests, and American Trade Policy*. Cornell University Press, Ithaca, NY

Gourevitch P A 1996 Domestic sources of international cooperation. *International Organization* **50**(2): 349–73

Haas P M 1990 *Saving the Mediterranean*. Columbia University Press, New York

Haas P M (ed.) 1992/1997 *Knowledge, Power, and International Policy Coordination*. University of South Carolina Press, Columbia, SC (originally published in *International Organization* **46**(1) (Winter 1992))

Haas P M 2001 Institutionalized knowledge and international environmental politics. In: Ikenberry G J, Parsi V (eds.) *Manual di Relazione Internazionale*. Gins, Laterza & Figlie, Rome

Haas P M, Haas E B 1995 Learning to learn. *Global Governance* **1**(3): 255–85

Healy R G, Ascher W 1995 Knowledge in the policy process. *Policy Sciences* **28**: 1–19

Katzenstein P J (ed.) 1977 *Between Power and Plenty*. University of Wisconsin Press, Madison, WI

Keohane R O, Nye J P 2001 *Complex Interdependence*, 3rd edn. Longman, New York

Knorr-Cetina K 1981 *The Manufacture of Knowledge*. Pergamon, New York

Knorr-Cetina K 1999 *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press, Cambridge, MA

Linder S H, Peters B G 1995 The two traditions of institutional designing: Dialogue versus decision. In: Weimer D L (ed.) *Institutional Design*. Kluwer, Boston, pp. 133–60

March J G 1994 *A Primer on Decision Making*. The Free Press, New York

March J G, Olson J P 1988 The institutional dynamics of international political orders. *International Organization* **52**(4): 943–70

Odell J S 1982 *U.S. International Monetary Policy*. Princeton University Press, Princeton, NJ

Odell J S 1990 Understanding international trade policies. *World Politics* **43**: 139–67

Reinicke W H, Deng F 2000 *Critical Choices*. IDRC, Ottawa

Richardson J 1995 Actor based models of national and EU policy-making: Policy communities, issue networks and epistemic communities. In: Kassin H, Menon A (eds.) *State Autonomy in the European Community*. Routledge, London

Risse-Kappen T (ed.) 1995 *Bringing Transnational Relations Back In*. Cambridge University Press, Cambridge, UK

Rose R 1993 *Lesson Drawing in Public Policy*. Chatham House, Chatham, NJ

Ruggie J G 1975 International responses to technology. *International Organization* **29**(3): 557–84

Ruggie J G 1998 What makes the world hang together? *International Organization* **52**(4): 855–86

Sabatier P A, Jenkins-Smith H C (eds.) 1993 *Policy Change and Learning: An Advocacy Coalition Approach*. Westview, Boulder, CO

Searle J 1995 *The Construction of Social Reality*. The Free Press, New York

Simon H A 1983 *Reason in Human Affairs*. Stanford University Press, Stanford, CA

Simon H A 1985 Human nature in politics: The dialogue of psychology with political science. *The American Political Science Review* **79**: 293–304

Social Learning Group 2001 *Social Learning and the Management of Global and Transboundary Risks*, 2 Vols. MIT Press, Cambridge, MA
Stone D A 1989 Causal stories and the formation of policy agendas. *Political Science Quarterly* **104**(2): 281–300
Thomas C W 1997 Public management as interagency cooperation: Testing epistemic community theory at the domestic level. *Journal of Public Administration Research and Theory* **7**(2): 221–46
Weaver R K, Rockman B A (eds.) 1993 *Do Institutions Matter?* Brookings Press, Washington, DC

P. M. Haas

Policy Knowledge: Foundations

Private foundations, typically found in countries with strong market economies and comparatively weak states, directly fund public policy analysis by the support they provide to research projects, policy institutes, and individual scholars who investigate policy questions. Foundations also advance public policy agendas by supporting advocacy groups and social movements that in turn have policy agendas. In most countries foundations are prohibited from direct lobbying of legislatures, but they are generally free to support public education campaigns that can contribute to an environment within which different public policies will be favored or disfavored.

1. *Definition/History*

The private foundation is an institution that privately manages private funds in pursuit of a public good. The source of these funds is normally a wealthy individual, almost always a person whose wealth was acquired in commercial activities. Foundation funds are privately managed by self-perpetuating trustees. The requirement that foundations advance a public good derives from the law under which they are chartered. That law stipulates that the funds made available for the endowment of a foundation, being 'tax-free,' must be used for charitable or philanthropic purposes. Foundations operate primarily by giving grants from earnings on their endowments, though in rare instances will also use their endowed wealth for grant making.

The wealthy donor inclined to the philanthropic act as a way to perpetuate his or her name and social influence is not a recent practice. Plato bequeathed property so that his Academy would continue after his death, and in fact the Academy lasted nearly a millennium, from 347 BC to 529 AD, until it was deemed unchristian and dissolved by the Emperor Justinian. Popes and princes became patrons of the arts to enhance their esteem in the public eye. The royal gift, especially to provide a health or educational service, is known across history and across royal regimes from China to India to Egypt to England. More than a millennium of Islamic philanthropy built libraries, universities, and mosques. Commercial wealth has generated philanthropic endeavors across the centuries, and not just since industrial capitalism. The famous Hotel Dieu in Burgundy was established in 1443 when the benefactor donated a highly valued vineyard. The University of Uppsala was similarly endowed with a gift of productive land. Philanthropy as reflected in the royal gift, the papal patronage, and even the private endowment is the backdrop to what emerged in the advanced industrial democracies, particularly the United States, toward the end of the nineteenth century.

2. *Foundations and Industrial Wealth*

Late nineteenth century industrial wealth became the basis for a new type of philanthropy—the private foundation. Modern foundations were a self-conscious break from Victorian era charity focused on the immediate relief of suffering and poverty. The distinction between foundation grant-making and charitable giving is often captured in the 'root cause' metaphor. Rather than relieve suffering, foundations see their purpose as doing something about the deeper causes that lead to suffering in the first place. The classic example is public health such as a vaccine regimen that is designed to prevent the suffering that accompanies epidemics.

Foundations also differed from the one-time gift to endow a hospital or center of learning. The private foundation, a legally immortal institution, extends the bequest and its purposes into the distant future. That is, rather than endow an institution, the philanthropic act endowed the foundation itself. There are a number of reasons for this institutional innovation, one being the sheer magnitude of the private fortunes gained by controlling a key industry (oil, steel, banking, transport) during the unregulated and largely untaxed decades of the late nineteenth and early twentieth centuries. The private foundation, then, is a place to put wealth too substantial to give away in one's lifetime.

The twentieth century saw a vast growth in the number of private foundations, and in their aggregate asset base. There are more than 40,000 in the United States alone, and many fewer though still substantial numbers in Western Europe, Japan, Brazil, India, Mexico, and other of the wealthier nations of the world. Many foundations are comparatively small, with annual grant budgets of $1 million or less. At the other extreme are a few substantially endowed

foundations—such as the Gates and Ford Foundations in the United States and the Welcome Trust in Great Britain, whose annual grant-making approaches $1 billion. There are also corporate foundations and community foundations, which together comprise about one-fifth of all foundations in the United States and a smaller percent in other countries.

There was a surge in the number of independent private foundations in the latter decades of the twentieth century, for reasons similar to the first surge that occurred a century earlier. The late twentieth century wealth gained through such new industries as electronics, information, media, and entertainment was at a scale comparable to the nineteenth century fortunes in oil, steel, and railroads. Such growth occurs in a period when the rules of the liberal economy allow, and even encourage, the accumulation of inordinate wealth in private control and which tolerate wealth inequalities at a scale thought improper if not immoral only a few decades earlier. If wealth redistribution through public policy was not in favor in the early twentieth century, neither is it in favor in the early twenty-first. The private foundation becomes, then, a means to recycle some of this wealth, though on terms largely set by the individuals who control it.

3. *Foundations and the Political Culture*

Social Darwinian theories favored by late nineteenth century industrialists and their apologists was particularly favorable to the emergence of the private foundation (see *Darwinism: Social*). An influential statement was Andrew Carnegie's widely circulated essay, *The Gospel of Wealth* (1889a). Carnegie acquired his vast wealth in the steel industry. He argued that industrial capitalism was providing more and more benefits to ever greater proportions of the population, but acknowledged that its workings inevitably generated substantial surplus wealth under control of the few. 'We accept and welcome, therefore, as conditions to which we must accommodate ourselves, great inequality of environment; the concentration of business, industrial and commercial, in the hands of the few; and the law of competition between these, as being not only beneficial, but essential to the future progress of the race.' (p. 656). Having acquired this wealth under the benign influence of social Darwinianism, Carnegie then concluded that it was his Christian obligation to give it away.

It was Max Weber, in *The Protestant Ethic and the Spirit of Capitalism*, who gave sociological depth to the association between Protestant asceticism and responsibility—with its emphasis on discipline, secular engagement, and deferred gratification—and the thrifty, hardworking business ethic that shaped modern capitalism (see *Weber, Max (1864–1920)*). The Protestant ethic stressed the greatest possible productivity, but frowned on the luxurious enjoyment of the wealth so earned. If work is to manifest God's glory, the profits of work are to be reinvested in that which is productive and socially beneficial. To do good works in this life is a sign of grace. Not that salvation could be earned by good works, but doing good as well as being good was to be expected of wealth. This was a convenient doctrine for men, Carnegie notably among them, who were both devoutly religious and wealthy beyond easy measure. It is comforting to know that one can dedicate oneself to acquisition and yet be virtuous.

Although explicitly religious motives are not often cited in today's more secular culture, the Protestant ethic retains its hold. The notion that one is a 'steward' of surplus wealth still appears in explanations of the philanthropic act, as do echoes of the belief that good works in this life signal moral worth. Late twentieth century billionaires speak of an obligation to repay society for having been so good to them. They are in this respect reflecting the notion that the wealthy *because* they are wealthy are obliged to be philanthropic. Inherent in wealth is the responsibility to be responsible for the public good that one's wealth can promote.

The social Darwinian environment within which foundations emerge was compatible with another feature of, especially, American political culture—the weak state tradition. Understood as private funds linked to public purpose, the private foundation is a social invention consistent with this tradition. If wealth is accumulated in an amount that is too substantial to be consumed or given away in one's lifetime, and thought to be too great to bequeath to family members, there are a limited number of ways to dispose of it. Furthermore, if the political culture as well as law and taxation policy invite the wealthy to create private institutions that can function as an alternative to the state, there is much satisfaction to be gained by creating a foundation. In the twentieth century, then, the private foundation has become a major mechanism by which large-scale, privately held capital memorializes the donor, endows his or her philanthropic inclinations, and not incidentally substantially reduces tax obligations. In some instances it also serves to keep control of a business enterprise in the hands of a founder's family and descendants.

There is a long tradition of inquiry that sees in the establishment of foundations the intent by the donor to exercise lasting political–economic–social control. The very wealthy do not earn their fortunes without exercising economic control—of an industry, a marketing strategy, a scientific innovation, a new product. Nor are these individuals indifferent to how policies, regulations, tariffs, and taxes can help or hinder the acquisition of wealth. It is not much of a leap from these facts to presume that transferring private wealth to a private foundation is a way to extend political–economic influence.

For example, a Commission of the United States Congress (1915–16), observed that industry and wealth in the United States had come under the control of a few very wealthy persons. The Commission alleged that this small group intended to continue and extend this control by establishing privately managed funds for indefinite purposes. Here, then, is coupled the late nineteenth century anxiety about the concentration of economic power with the fear that private foundations would perpetuate the dominance of business interests in American political and social life.

Careful historical analysis needed to test this hypothesis is in short supply. There is, however, ample anecdotal evidence that donors who establish foundations have a strong inclination to see that their funds enhance particular political and economic doctrines, especially private enterprise and market capitalism. Perhaps there is an irony in the fact that the foundation results from the accumulation of substantial private wealth and yet in instance after instance declares that its mission is to improve the lot of the poor and powerless. It goes about that mission by helping to lower the barriers to upward mobility, by working to insure basic civil and political rights or by contributing to education and health for the poor. It does not go about that mission by calling into question the political-economic arrangements that allow for unequal wealth acquisition in the first place. This despite the oft repeated claim that foundations will find and eliminate the 'root causes' of poverty, discrimination, and illness. These root causes, however, seem not to include the economic doctrine and public policies that govern wealth acquisition.

4. *The Public Policy Agenda*

Based on statistics it collects from thousands of private foundations, the US based Foundation Center uses a ten-part subject-matter classification to report grant making: arts and culture, education, environment, health, human services, international affairs and development, public/society benefit, science and technology, social science, and religion. Although data collection is less systematic in other nations, these categories are generally serviceable across the private foundation world. Public policy is not its own discrete category, because policy analysis and policy advocacy can appear in each of the ten categories. Foundations have taken up arts policy, education policy, health policy, science policy, security policy, and much more.

Of course the policy interests of foundations alter with the times. To note obvious examples: the sizeable investment by American foundations in area and international studies, language training, and student exchanges is a product of post World War II conditions; the rise of environmental consciousness in the 1970s led to a marked increase in funding for environmental policy in Japan, Europe, and the US; the collapse of communism suddenly opened up opportunities in Eastern Europe and the former Soviet Union to which both European and American foundations were quickly responsive. Such examples could be multiplied over and over, indicating that the foundation mission and especially how it expresses itself in grant making is not frozen but rather adapts and evolves as the policy conditions change.

The foundation has a number of strategic choices available as it considers whether and how to confront the public policy agenda. Considered from the broadest possible perspective, foundations adopt strategies based on their notions about how social change can best be brought about. Does the foundation operate on the assumption that ideas drive history; or that technologies do; or social movements, market incentives, government interventions, moral exhortation? If ideas drive history, invest in intellectual efforts; if government interventions drive history, invest in specific policy analysis and advocacy; if exhortation drives history, invest in public education. Or, to use a metaphor much favored by foundations, grant making should seek out the point of 'leverage' judged to be most productive of desired social change. The point of intervention is sometimes indirectly and at other times directly linked to public policy knowledge or policy advocacy.

4.1 *New Knowledge*

Foundations have made major investments in universities and in research institutions, equipment, and personnel, as part of a broad strategy to advance basic human understanding in nearly every field of study imaginable: anthropology, astronomy, biology, chemistry, economics, history, medicine, music, physics, political science, and psychology, to name only the most obvious. These investments have, at times, been sustained over long periods, illustrated by the Rockefeller Foundation's century-long involvement with medical science. And for some foundations for certain periods of their history, the search for new, fundamental knowledge defined what they were about, what they took to be their vocation. Seldom, however, was new knowledge sought for its own sake. The foundation is focused on *public* benefits and thus even an investment in basic science will always be linked to how the new knowledge is to be applied.

4.2 *New Knowledge Applied*

Foundations have always been focused on how new knowledge can be applied in socially beneficial ways. An example is the investment in the public health profession, viewed as a means to ensure that dis-

coveries from biomedical science will reach the broadest population possible, and especially those population groups that might otherwise be ignored. There are many other examples: basic knowledge in learning theory being coupled with school reform efforts, basic knowledge in plant physiology leading to higher yielding food crops and their dissemination in poorer countries. In this application effort, the foundations make heavy use of the social sciences more generally and policy studies more specifically. (See *Policy Knowledge: Universities*.)

4.3 Direct Policy Analysis

Foundations have had a long, close relation with studies and publications that would inform public policy. This can be seen in the heavy investment in policy research institutions in the United States, and in policy reports and white papers in European countries. In the United States the independent policy institute and the independent foundation are linked because they are responsive to the same deep strand in the political culture that prefers the private to the public sector (Prewitt 1995). Foundation funding of the social sciences, which starts before the turn of the century, has generally been motivated by the quest for improved public policy. The close relationship between foundation funds and public policy analysis underscores how privately held philanthropic resources attempt to advance public purposes. This relationship has been a complicated one, attracting its share of sharp criticism from both the political left and political right. Notwithstanding this criticism, foundations have forthrightly declared their intent to fund analysis that will lead to what they view as public policy improvements. For example, the German-based Bertelsmann Foundation articulates as one of its primary goals to provide new concepts and models for the social market economy and for public administration.

4.4 Policy Advocacy

Often in combination with policy analysis, foundations have funded a seemingly endless number of organizations dedicated to fighting for or against particular public policies: in human rights, equal opportunity, school reform, nuclear disarmament, environmental protection, free speech, children's welfare, and on and on. Although a dimension of American philanthropy early in the century (Sealander 1997), such funding gained momentum in the post World War II period as the federal government became increasingly active in areas long of interest to the philanthropic sector. If foundation funds could stimulate the government to adopt the 'right' policy in civil rights or social welfare, the impact would be far greater than what could be accomplished with foundation funds alone. From this emerged the notion that philanthropic dollars could be taken to scale by leveraging government funds. That is, if government through its regulatory and taxation powers influences private expenditure patterns, philanthropy reverses the direction of influence by spending private funds on policy analysis and advocacy that will change government spending priorities. In earlier years, the policy advocacy groups supported by foundations largely focused on domestic policy. In recent years, however, attention has broadened to include the international arena and international actors; policies of the World Bank or the World Health Organization or multinational corporations are targets for internationally organized advocacy groups, especially in the fields of environment and human rights. The post-1989 preoccupation with what is often if not very precisely described as civil society funding is a combination of policy advocacy and social movement philanthropy. (See *Policy Knowledge: Advocacy Organizations*.)

Though overlapping with policy advocacy, social movement philanthropy is sufficiently different to merit independent mention. In recent decades, the major beneficiaries have been social movements organized around equality for women, environmental protection, and civil rights for racial and ethnic minorities Although each of these social movements originated outside philanthropy, foundation dollars rather quickly led to their stabilization and professionalization. Support for social movements is largely a post World War II phenomena, getting its start with the activism that characterized the 1960s. This type of funding has also been internationalized, as worldwide organizations have formed to advance environmental protection or women's rights or children's welfare. What foundations refer to as support for grass-roots organizations, or community groups, is often an element of social movement philanthropy.

Closely associated with such funding is an emphasis on *social empowerment*. This term refers to the direct empowerment of individuals, especially the traditionally powerless or disenfranchised social groups. An example is micro-credit lending, which is designed to put resources directly in the hands of the powerless poor. Social movement and social empowerment philanthropy are linked in that the former has the dual purpose of empowering its adherents and changing the social landscape. But direct empowerment need not be part of an effort to create a social movement. The beneficiaries of a micro-credit lending program or a female education strategy are not viewed as members of a movement in the same way that feminists or ecological activist are. Social empowerment, of course, is a close cousin to charity—the direct relief of suffering. But philanthropy has defined itself as not about charity so much as by its intent to remove the 'root causes' that lead in the first place to the suffering

for which relief is sought. In its effort to permanently change the condition of the disenfranchised, empowerment is distinct from direct charity.

4.5 Public Education

Foundations have often tried to influence the policy knowledge of the general public. There are numerous examples where foundation funds have been used to improve public understanding as a way to correct for what are considered 'misguided' public practices. A public inclined to permit inefficient and corrupt partisan politics, for example, might be taught to appreciate the virtues of independent expertise in government and a professional civil service—and turn of the century philanthropy took up this progressive cause. Or, a public too given to racial prejudice should be enlightened to the moral worth of racial tolerance—a major philanthropic project across the twentieth century. The examples multiply endlessly—the public is inefficient in its use of energy, unclear about the dangers of population growth, illiterate about its national history, prone to unhealthy eating habits, smokes too much, and, repeatedly, is insensitive to the plight of, depending on the times, women, workers, children, the poor, immigrants, minorities, the handicapped or others from a long list that changes its emphasis from decade to decade, but is never absent from the philanthropic portfolio.

These deficiencies in public knowledge, understanding or attitude can be corrected by the selective use of foundation resources, such as introducing new elements into the school curriculum or through targeted media campaigns. More than public education is at stake, for these efforts often have a change in public policy as the goal. The anti-smoking campaign illustrates this. Foundation supported public education intends not just to discourage smoking, but also to modify public policies on, for example, where tobacco can be sold or how it can be advertised.

What are generally described as 'liberal foundations' have been especially adept at using public education campaigns to shape public opinion, though in recent decades foundations with a more conservative inclination have supported the writings of public intellectuals on what they see as failures in public education, social welfare, criminal justice, and other projects long associated with liberal foundations.

5. Entrepreneurial Philanthropy

Through the twentieth century the attention of foundations to policy knowledge and analysis focused almost exclusively on government policies. Insofar as the behaviors of business sector firms were of interest, the strategy of choice was to affect regulation, taxation and other public policies to influence market behavior. In the latter decades of the twentieth century this exclusionary focus on government policy began to alter. Some foundations directly addressed the policies and practices of market actors. For example, in the field of public health, foundation funds were used as an incentive to induce pharmaceutical firms to invest in AIDS vaccine research and with respect to energy efficiency, foundation funds attempted to influence utility companies to give more attention to renewable energy sources such as wind and solar power. Microcredit lending (below market loans in impoverished areas for small-scale entrepreneurs) also became a favored target of foundation funds.

With the arrival on the foundation scene of young, recently wealthy entrepreneurs from the high technology industries this form of philanthropy is likely to increase. The degree to which it will supplant rather than simply add to traditional patterns of public policy funding is unknown, but it is not likely that foundations will soon if ever exit the long-established field of public policy analysis and advocacy.

6. Conclusions

Foundations are well established in advanced market economies, especially in those countries with a comparatively weak state tradition. Over the decades a receptive network of regulatory and tax policies have emerged to make it easy to establish foundations, and thus to withdraw funds from the tax stream on the assumption that these funds will serve the public good. Foundations have used these tax-free funds to support studies of alternative public policies in areas of interest such as health, environment, education, the arts and so forth. They have contributed significantly to the growth of policy knowledge. As more and more nations of the world adopt liberal market economies, it is likely that foundations will emerge in ever greater numbers. It is also likely that public policy analysis and advocacy will remain a foundation funding priority.

Bibliography

Arnove R (ed.) 1980 *Philanthropy and Cultural Imperialism: The Foundations at Home and Abroad*. G. K. Hall & Co., Boston

Carnegie A 1889a Wealth. *North American Review* **June**: 653–64

Carnegie A 1889b Wealth. *North American Review* **June**: 682–98

Cuninggim M 1972 *Private Money and Public Service: The Role of Foundations in American Society*. McGraw-Hill Book Company, New York

Foundation Giving: Yearbook of Facts and Figures on Private, Corporate and Community Foundations. 1999 The Foundation Center, New York

Fosdick R 1952 *The Story of the Rockefeller Foundation: Nineteen Thirteen to Nineteen Fifty*. Harper and Brothers, New York
Keppel F 1930 *The Foundation: Its Place in American Life*. Macmillan Company, New York
Mohn R 1997 Objectives of an operating foundation. In: (ed.) *The Work of Operating Foundations: Strategies—Instruments—Perspectives*. Bertelsmann Foundation Publishers, Gutersloh, Germany
Prewitt K 1995 *Social Sciences and Private Philanthropy: The Quest for Social Relevance*. Indiana University Center on Philanthropy
Sealander J 1997 *Private Wealth & Public Life: Foundation Philanthropy and the Reshaping of American Social Policy from the Progressive Era to the New Deal*. Johns Hopkins Press, Baltimore

J. Goldenberg

Policy Knowledge: New Liberalism

New Liberalism is at its core a contrast term, designed to distinguish it from and to set it against classical liberalism, as the latter developed in the West between the seventeenth and mid-nineteenth centuries. In use internationally between the 1880s and the 1940s, the term referred to a broad-gauged movement in social knowledge, public philosophy, state policy, and social practice intended to address the 'critique of political economy' and the 'social question' that the pernicious instabilities and distributive inequities of modern capitalism had engendered, and to provide new critical thinking about the crucial relationships between *state*, *economy*, and *society*. The movement should not be taken as a total repudiation of the older tradition, with its characteristic stress on individual autonomy, private property, and limited government. Rather, new liberalism was undertaken initially as an exploration of how and where to amend the old ideals in light of what had come to be seen as negative consequences of capitalist industrialization, and as a method for helping to shape and choose among new civic and social moralities under construction in the industrializing nations of the West.

New liberals were ultimately unable to sustain their critical assessment of the classical tradition without changing quite fundamentally their understanding of the relationship between society, economy, and state. In an intricate process of reaction and inversion, these changes shifted the meaning of political conservatism in the West, gradually severing it from its basis in quasi-feudal and aristocratic precedents and reorienting it toward a defense of personal, civil, and commercial liberty as it had earlier been defined in the 'old,' classical liberalism. The account of new liberalism provided here refers mainly to the American case, with occasional references to other (mainly the British) experiences. It identifies four phases in the development of new liberalism that aid in heightening the visibility of critical moments in its construction. These phases are: (a) market failure and the drive for new social knowledge, (b) conflicting approaches to framing the social question, (c) the 'turn' to social insurance, and (d) the social sciences, planning, and the limits of liberal statism.

1. *Market Failure and the Drive for New Social Knowledge*

The devastating market failures of the late nineteenth century, destabilizing waves of bankruptcies, mass unemployment, and chronic, desperate poverty produced an enormous disappointment of traditional liberal expectations regarding the efficacy of the market. New liberals recognized market failure and attendant social conflicts as evidence of a wider collapse, more disruptive to intellectual, social, and political order and more menacing to communities of a social contract based on unfettered competition between white males that had been implemented during the market revolution of the Jacksonian era. Rejecting the excessive individualism, localism, and states-rights obstructionism of that period as morally and intellectually bankrupt, they put themselves quite deliberately in the business of conceptual innovation, seeking a higher rationality that would place knowledge in the service of a more democratically constituted society, which would maintain itself in part by allotting intellectuals a special role in shaping a more vigorous, capable, developed, and purposeful national government.

Across the range of nations involved, the conceptual revolution that new liberals sought to promote was grounded institutionally in an international social science association movement that fostered transatlantic borrowing, and in critical journalism, settlement houses, and social surveys, all of them cranked up to record, analyse, and disseminate empirical data, propound theory, encourage the development of political leadership responsive to the 'teachings' of liberal social knowledge, improve the execution of policy, and advocate reforms intended to renovate and enable Enlightenment ideals of political equality and social justice. Within government, in investigative agencies such as improved censuses, labor, women's, and children's bureaus, trade boards, and bureaus of corporations, high-level civil servants such as Llewllyn Smith and Lord Beveridge in England, and Carroll Wright, Francis Amasa Walker, and Julia Lathrop, in the United States provided conduits for policy ideas from those burgeoning 'parastates' into the state. Within this 'great transformation' in social thought and political practice, crucial early players were the

academic social science disciplines, whose missions and boundaries were historically determined, differing from country to country. Academic professionalization proceeded rapidly in the American research universities, fueled by rapid economic growth, academic entrepreneurship, and the relative shortage of other types of creditable specialized intelligence.

But from the beginning, progressive academic economists, political scientists, and sociologists were interested in more than the provision of expertise; even as conservative university trustees (and at times conservative colleagues) disciplined their advocacy, social scientists saw themselves as crucial links in the interpretive processes needed to make the new political and social economy intelligible. Thus their specialized investigations and discourses were a critical factor in reworking the political agenda so that it placed prominently before policy makers and emerging, activist publics as problems urgently needing comprehension the range of issues that came to define the politics and policymaking of the new liberal era. Most prominent among these issues were business cycles and unemployment, rights and conditions of labor, wages and distribution, industrial organization and corporate reorganization, family welfare and gender roles, problems of the low-paid, exploited immigrant work force and—often with disabling assumptions rolled in—issues of race. They also addressed related issues of governance: the detection of wholesale corruption of law and government that inspired the clean government crusade, and the need for methods of building up the cognitive abilities and administrative capacities of public officials.

Strategic rhetorical steps were taken early that shaped the ties between American governance and intellectual life in the period. In the United States, Woodrow Wilson's early and generative 'Science of Administration' (1887) addressed republican fears of bureaucracy by placing a firewall between politics and administration. Reflecting distaste for the intense, tribal partisanship of Gilded Age politics, academics and bureaucrats (unlike many female social scientists and social reformers) adopted nonpartisanship as a standard description of their professional ethic. This pristine canon elevated everyday standards of disciplinary and agency practice, but it often hampered the effectiveness of social scientists as advocates for change, routing their activism through ostensibly apolitical venues such as foundations and think-tanks, which ceded partial control of research agendas to corporate elites.

2. *Conflicting Approaches to Framing the Social Question*

Yet despite frequent and often calculated professions of scholarly and professional neutrality, the building of new liberalism was neither a consensual nor a non-ideological process, but rather one in which newly experienced social realities disrupted expectations and categories in different and conflicting, contested ways. Deep divisions over the substance and implications of new empirical and theoretical knowledge led to a branching of new liberalism, early in its American career, into what participants saw as two distinct new liberal movements—a *statist-democratic* strand of reform thought born in the 1880s, and an opposing, *corporate-voluntarist* analysis that came together around the turn of the century. These conflicting reform discourses, worked out in dialogue with each other between the 1880s and the 1940s, opened up greater possibilities for genuine policy innovation than existed either earlier or since.

Assessing deficiencies in the mechanisms for translating an aroused social conscience into effective social action, the statist strand of new liberalism looked in new ways and on a number of fronts to a redefinition and a principled restriction of the meaning of 'private' in many of its applications, and to a corresponding expansion of what were to be taken as proper, morally necessary functions of government. Liberal statists sought more effective methods for mobilizing state agencies in behalf of a reinvigorated national interest, taken as something higher and more ethically compelling than the sum of the separate private, local, class, or sectoral interests.

The competing voluntarist vision rejected with equal conviction what—telegraphing their disapproval—its founders referred to as 'state interference.' Seeking ways to insulate and energize civil society, the anti-statists preferred to build new ordering strategies upon what they took to be the naturally evolving social institutions of a modern pluralist society, premier among them the corporation, and what they saw as organic, corporate bodies such as trade associations, trade unions, professions, and charitable associations, rather than around invigorated democracy, political mobilization, and a bureaucratic, regulatory state, which they represented as intrusive upon 'natural' evolutionary processes and prone to manipulation, most dangerously by social levelers who failed to appreciate the vulnerability of the investment system. More than merely Tocqueville revivalists, they were particularly interested in what could be achieved by way of ordering a chaotic, excessively competitive economy through innovative uses of the business contract, within a rich climate of information about the market, and through enlightened management. This effort culminated in the liberal-corporatist design fleshed out during the 1920s for peak-level business-government cooperation in achieving economic and social stabilization through energized philanthropy and local relief in hard times, but mainly preventively, through the development and legitimation of industrial self-government oriented toward controlling chronic instabilities such as overproduction.

The liberal statists placed greater confidence in deliberative democracy than this, and less in capitalist rationality or the unity of the corporate and the public interest. Drawing energy from heartland republicanism, evangelical Christianity, and European socialism, a leading theorist of the new liberal state, political economist Henry Carter Adams, marked out a path that two generations of unapologetically statist progressives would define as a democratic, progressive, collectivist or communitarian public philosophy—a distinctively American middle way between what they framed as the excessive individualism of the British model and the excessive statism of the German. The founding text for the liberal statist approach was Adams's classic paper, 'Relation of the State to Industrial Action,' which interrogated standing assumptions about the beneficence of competition and called into question the fundamental morality and the long-term serviceability to the public interest of the way the modern capitalist market was being constituted in its corporate phase. For those engaged in reflecting on issues of public philosophy, Adams, along with Richard Ely, Florence Kelley, Edmund James, Simon Patten, John R. Commons, and Jane Addams among numerous others, did more than establish a 'context of refutation' for socialists and diehard believers in *laissez faire*. More important, in relation to later developments in theoretical and policy knowledge, they directed their arguments quite explicitly to new-liberal antistatists, who were forming their own intellectual identity in the 1890s as part of a corporate-voluntarist counter-attack on liberal statism.

Reflecting back on the historical construction of the ante-bellum US liberal order, Adams pointed to three deficiencies of a regime of market competition that set the initial terms on which this new, positive vision of liberalism would be built: (a) its failure to allow society to enjoy the benefits of monopolies occurring naturally in increasing-returns industries, and thus to reap the benefits of human genius enacted in advancing technology, (b) the inevitable decline of business ethics to the lowest level, absent a generally credited method for communal determination and legal establishment of labor standards, among them a right to employment, and (c) a corruption and general hollowing out of the substance and potential of representative government in a deliberate, misguided weakening of state capacities, which Adams traced to anti-statism in the Jackson period, the lack of a respectable science of public finance in the United States, the rise of powerful corporations, and a compliant legal profession bent on shaping the protections of the law to private ends.

Naturalizing certain instances of monopoly was a critical step in the modernization of civic republicanism, designed to pull that venerable tradition out of the Jeffersonian, anti-monopoly, anti-statist discourse that had assimilated it to market liberalism. The reshaping of this core category laid an authoritative basis for state supervision of entire segments of the transportation, communication, and production systems, designed to control monopoly power. This strategic conceptual move gathered force in the regulation movements of the early progressive era, and in Theodore Roosevelt's vision of a national law of corporations, a robust national citizenship, and a muscular 'new nationalist' bureaucracy capable of disciplining the modern corporation. Classic texts such as Herbert Croly's *Promise of American Life (1909)* and the Progressive Party Platform of 1912 represented the maturing new liberal orientation to planning, in which capable, informed governance could instruct and protect a collective interest in conservation, labor standards, and consumer protection. More than merely an affirmation of the need for aggressive regulation in these regards, Rooseveltian progressivism argued for a public presence in the civic discourses and decisions regarding production, investment, and pricing. This broader vision of a deliberative process that could detect and shape the impact of American capitalism on society was later the unifying insight at work in the planning phases of the New Deal, most evident in the activities and discourses of the Public Works Administration (PWA), the National Resources Planning Board (NRPB), and the Temporary National Economic Committee (TNEC).

The Adams–Ely view was challenged immediately by procorporate academic economists Arthur Hadley and Jeremiah Jenks, close students of the sources of market failure in specific industries, and political advisers to Republican governors and presidents. Though these two extended what they saw as a natural tendency to monopoly even further than Adams did—into the whole broad range of capital-intensive industries, they considered even indirect state control of production and profit a threat to the price and investment systems that neither liberalism, capitalism, nor private morality could long endure. Conservative institutionalists such as these began constructing an alternative, more hierarchical, less democratic discourse of elite responsibility and social control, forming around the institutions of high finance and the corporation. They hoped to make the new corporate elites responsible for transclass and intercorporate cooperation, for personal growth and individual mobility along career ladders provided by corporate business, and even for community economic and moral development—to the extent that experience within a corporate culture would inculcate values of efficiency, cooperation, interdependence, and service.

This rudimentary modernization theory was the founding discourse of corporate liberalism, which down through Herbert Hoover's associative state to its full flowering in pluralist theory and liberal-corporatist public policy during the Eisenhower years equated the (properly enlightened) corporate interest with the (properly conceived and limited) public interest in sustained growth of the economy. Cor-

porate liberalism provided a systematic alternative to liberal statism, supplying a constraint on statist progressive politics in the US that was available, in anything like comparable strength, nowhere else.

More than other distinctive qualities, such as early (white manhood) suffrage, weak parties, weak labor, weak socialism, or state structure and capacities, this continuing tension between statist and corporatist discourses of the public interest, increasingly fleshed out in programmatic terms, explains—along with issues of racial exclusion that new liberals for the most part only minimally addressed—the special character of modern United States liberalism. It also helps to account for the distinctive pattern of US state formation, and particularly welfare-state development, that students of social policy, often taking the more comprehensive European welfare states as the normal path of history, have seen as tardy, insufficient, and incomplete—kept from developing by traditions of individualism and *laissez faire*.

In reality, such commitments were highly attenuated where they survived at all among the knowledge-producing classes most involved in shaping new liberalism. For democratic theorists from Adams to John Dewey, the public interest was a potent category—reclaimed from the republican past, but extensively recast to take account of concentrations of social and economic power that threatened family subsistence, community morals, responsive and responsible governance, and political, social, and economic democracy. In his seminal *The Public and Its Problems* (1927), another powerful exemplar in the new liberal 'role of the state' genre that H. C. Adams inaugurated for Americans, Dewey located the public interest as the place in social relations where effects spilled over from the parties to a contract to third parties, who bore the often devastating ancillary consequences of others' private bargains.

In the construction of democratic statism, problems raised by shoddy business ethics and new, manipulative forms of 'scientific' management rivaled concerns about monopolistic corruptions of the corporate form. Instituting another enduring theme in new liberal reflection about the role of the state, the second 'evil of competition' in Adams's treatise—pernicious insecurity, exploitative working conditions, and relentless time pressure—he proposed to remedy by labor-standards legislation based on a community moral consensus derived discursively, from detailed knowledge of the impact of industrial labor as then constituted, on the moral, mental, and civic capacities of the working class. The debilitating effect of casual labor, low wages, and chronic unemployment on working class men and women was a constant theme that women's social investigation and reform activism also did much to develop. But this emphasis on legislated labor standards should not be mistaken for an early blueprint for the welfare state. In the formative moment, the 1880s to 1910s, American statist progressives were looking not to compensatory welfare measures but to independent unions, fostered by public intellectuals and protected by the state, and to community moral pressure and democratic politics, as the means for achieving an 'economic republic' and 'industrial democracy.' Their vision included recognition of two inalienable rights of worker citizens—a 'living' wage, sufficient to support a family in reasonable comfort, and a voice in management.

3. *The 'Turn' to Social Insurance*

Underway in Europe much earlier, the new liberal 'turn' to social insurance came in the American case only in the second decade of the twentieth century. Growing up in social work, settlement, bureaucratic, and academic circles, it signaled the onset of a second phase in new liberal thinking, running parallel, but in important respects contrary, to the older movement for industrial democracy, which also continued to gather force through the New Deal and into the late 1940s. Several factors can be seen to have promoted the adoption of a social insurance perspective, among them the arrival of a broad environmentalist consensus on the causes of poverty, accumulating realism about distinctive constraints on labor's capacity to organize a positive political program, and the social tensions that compelled American cities to provide relief during cyclical downturns from the 1870s onward. Once again, academic and expert knowledge played a major role as well. European reform discourses and precedents, culminating with the British adoption in 1911 of broadly accessible unemployment insurance, were faithfully catalogued in works such as Isaac Rubinow's *Social Insurance* (1913). As class conflict accelerated, inquiries staffed by social scientists such as the famous Pittsburgh Survey and the United States Commission on Industrial Relations documented economic insecurity as the central cause, along with corporate arrogance and powerlessness to combat it in their communities, of worker unrest, suggesting an attack on inequality through aggressive labor standards reforms and state policies promoting redistribution.

Contrary to the strong government role envisaged by the USCIR and in wartime programs for improving labor standards, in the post-World War I period, with Herbert Hoover in a prominent organizing role, the Wilson and Harding administrations arranged a series of high-level conferences for planning postwar stabilization. Allied with the President's Conference on Unemployment, Wesley Mitchell's National Bureau of Economic Research drew in economists such as Leo Wolman and Wilford I. King for the vital statistical work and propagated new methods for understanding the pesky business cycle. This chain of policy thinking suggested alternative policy options, among them more responsible administration of credit by banks and government, enrichment of the informational climate for business decisions to facilitate a

microeconomic approach to steadying output and employment, the accumulation of reserve funding for countercyclical public works by (from Hoover's perspective, preferably local) government, and an actuarial approach to relieving unemployment through funds amassed by employers, unions, or government. In what must be seen as one of the most significant conceptual innovations in modern governance, these more conservative approaches transformed unemployment from a problem of the political order rooted in inequality into a 'risk,' the key policy decision becoming therefore how, and by whom, that risk—understood as a problem in the smooth management of the elastic supply of reserve labor accepted as an indispensable part of the capitalist process—could be insured. In the discourse manufactured in Hooverian circles, most of these remedies could be implemented by informed business decisions and enlightened labor-management cooperation.

The impulse for what from the perspective of liberal-statist, laborist, and social-feminist proposals for extensive redistribution of income and political voice became a critical retrenchment in social policy thinking arose not only from these liberal corporatist iniatives, but from at least three other, at first glance unlikely sources as well: the maternalist turn in the women's movement of the progressive era; a vision of the immigrant and non-white poor propagated in the social sciences, by the Chicago school of sociology most particularly; and the appropriation and aggressive expansion of carefully targeted forms of social provision in the private-sector, imagined by corporate liberals in business management, government, and the academy as a strategy for making the 'modern manors' of welfare capitalism reliable instruments for securing the loyalty and uninterrupted productive service of industrial employees.

As the central sites for the development of reform discourses that transported women's issues from the private to the public spheres, the women's movement and female social science were neither immediately nor exclusively committed to a social insurance perspective. Female social scientists such as Jane Addams, Florence Kelley, Alice Hamilton, Crystal Eastman, and Mary van Kleeck initially broadened the social question to include the appalling conditions faced by working class families in immigrant enclaves, conditions they traced as much to deliberate capitalist exploitation as to the relentless, patterned instability of the capitalist economy. Constitutive new liberal texts such as the pioneering social survey *Hull House Maps and Papers* (1895) and Margaret Byington's 1907 study for the Pittsburgh Survey of steel workers' families traced their poverty to the most discretionary of environmental causes—a calculated policy of exceedingly low wages and insufficient employment, yielding family incomes well below what was needed for subsistence, chosen by leading industrialists as a way to socialize much of the cost of maintaining a low-skilled industrial work force, including a good-sized reserve labor supply.

Female reformers such as these translated their expert knowledge of the working conditions of women and children into an ethic of public responsibility that matched and complemented the initiative of academic new liberals such as H. C. Adams for finding ways of translating communal judgments regarding the 'moral plane of competition' into policy. Often promoting and then working in state and national investigative and protective agencies, rights-thinking feminists were thus among the earliest new liberal state-builders. Yet, as conservative opposition accumulated and the suffrage crusade shifted to political mobilization, important segments of mainstream feminism took on a more elitist, exclusionist cast, and the 'female dominion' of women's institutions increasingly followed a separatist course in identifying its reform goals. Leading female reformers, sociologists, and legal realists built a maternalist case for special state-based income assistance to mothers and children, special provisions for safeguarding maternal and infant health, and special protections for republican mothers in the workplace.

Maternalists helped to shape as the American pattern what women's historians have identified as a two-channel welfare state, in which one channel provided universal coverage for male workers under reasonably generous programs such as workers' compensation, unemployment insurance, and Social Security, whereas the other offered only meager, means-tested pensions for virtuous poor widows, and little assistance of any kind for most female workers and people of color. Worthwhile in themselves as strategies for meeting real needs among the 'deserving' poor, these approaches sidestepped, or confronted only indirectly, and without formulating them as such, issues of monopoly power, proletarianized, low-wage work, and obstacles to unionization for workers of both sexes that had initially been catalysts for the development of democratic-statist new liberalism.

The social sciences provided powerful legitimation for the social insurance track as an alternative to empowering the bottom ranks of the working class. Between the 1890s and the mid-twentieth century, even as institutionalists continued to fashion economic knowledge in forms that emphasized law and capitalist agency as contexts for the political and cultural creation of markets, rival neoclassical economists cultivated statistical and mathematical models of analysis that extracted economic behavior from history and culture, naturalized the workings of the market, and turned the discipline into a potent, technocratically organized form of social engineering. As Dorothy Ross's account of the rise of scientism also documents, quantitative, behavioristic and functionalist models vied with historicism and institutionalism in other social sciences as well, perhaps most visibly in two massive collaborative efforts in Hoover-

era policy knowledge, *Recent Economic Changes* (1929) and *Recent Social Trends* (1933). The dominant sociological discourse of the 1920s, the 'Chicago School' established by Robert Park, W. I. Thomas, and their students at the University of Chicago offered an alternative to the liberal-statist explanation for poverty and urban ills that took the form of a cultural theory of urban family and community 'disorganization.' Describing a succession of discrete stages—contact, disorganization, acculturation, assimilation—that accounted for the rise of problems such as dysfunctional families, juvenile delinquency, and crime, these scholars offered a vision of the immigrant, and especially the African-American migrant poor as culturally deficient—whether from environmental, or more enduring cultural causes. They uncovered 'pathologies' in Negro personality and family patterns, an alleged inability to defer gratification rooted in the experience of slavery, and a tendency toward male improvidence and female-headed families, for example, that predicted a longer than normal acculturation process for African Americans. Culminating with Gunnar Myrdall's *American Dilemma* (1944), this new cultural analysis of poverty separated the poor from the capitalist strategies that shaped their lives, and rendered those historically constituted structures less visible.

Maternalist social knowledge and the sociology of ethnic difference merged with developments in the new sciences of management to promote the development of corporation-based programs for linking carefully constructed forms of social provision to employment with specific employers. When wartime controls on the economy were quickly abandoned, corporatism flourished in the United States (as it did in different forms during the interwar period in Europe as well). Doctrines of imperfect competition, an institutionalist analysis of the modern corporation's separation of ownership from control, arts of advertising based on a new consumer psychology that legitimized desire, and theories of an administered market allowed liberal corporatists to proffer a new economy of mass consumption as a palatable alternative to nineteenth-century producerist dreams of worker control.

Resisting this trend only formally, the American Federation of Labor, along with industrial relations experts constructing a new profession in the 1920s, recognized productivity increases as the indispensable basis for increases in wages. Where company welfare programs existed, worker loyalties based on them ran quite deep. Thus the shock of the sudden collapse of Hooverian strategies for industrial self-government in the Great Depression crisis, the failure of voluntarism more generally, the welcome proliferation of New Deal programs targeting unemployment and poverty, and the aggressive organizing efforts finally unleashed by national labor-rights legislation helped to ratchet up workers' expectations of the state, and to reinvigorate impulses, largely quiescent since the immediate postwar era, toward the development within government of planning capacities capable of heading off such emergencies in the future, and of democratizing both industrial governance and the political state.

4. *The Social Sciences, Planning, and the Limits of Liberal Statism*

From their beginnings in the Enlightenment and especially since the late nineteenth century, the social sciences were oriented toward planning, defined imprecisely enough to include methods all the way from enriching the informational climate of decision making to managing economic performance and shaping cultural expression. Most centrally a conservationists' discourse of efficient natural resource management in the 1910s and a welfare corporatists' discourse about the benefits of social cooperation in the 1920s, the American conversation about planning matured rapidly in the depression emergency. As critics correctly charged, stressing continuities in the development of a national planning capacity that bridged private and public, the National Recovery and Agricultural Adjustment Acts could seem like little more than surreptitious continuations of corporatist strategies for contriving scarcity under the cover of state compulsion. Yet a parallel, consciously statist movement also reappeared in the 1930s, fueled by FDR's intense commitment to planning, challenges on the left that promised full utilization of idle resources and full employment, and the recruitment to government planning agencies of a social scientists, including a large number of institutionally oriented economists, who energized and updated the democratic statist strand of new liberal thinking. When crisis suddenly opened the way for bureaucratic efforts to comprehend and combat depression, many of the plans necessarily came 'off the shelf,' as did the comprehensive programs for national old age, survivors, unemployment, and disability insurance. Yet by the mid-1930s institutional and policy innovations shifted the planning discourse toward methods of using resources efficiently over the long term, expanding public investment, crushing debilitating tendencies toward monopoly, and rekindling demand—policy goals institutionalists insisted were related causally. Social scientists in key nodes of conceptual innovation redefined planning and, within the constraints of opposition that also bedeviled Franklin Roosevelt's own narrower planning ambitions, attempted to give it an overt civic dimension. Thus, in the early leadership of the National Labor Relations Board, William Leiserson and others experimented with methods for linking the civic and stabilization purposes of the Wagner Act, which its framers intended to ensure the worker's right to organize independently, but also to equip trade unions

to play a role in redistributing income, supporting the workers' ability to consume, and democratizing industry.

Economists' knowledge and planners' leanings toward infusing market structures with civic purposes also blended in agencies such as the venerable Bureau of Agricultural Economics, the path-breaking, community-reconstruction oriented, ultimately disappointing Tennessee Valley Authority, and the Farm Security Administration, which addressed the issue of chronic farm surpluses, but also the economic and social marginality experienced by people at the bottom of rural hierarchies. A major study by TNEC of the concentration of economic power steered by Thurman Arnold and Leon Henderson referred the stickiness of prices, excess saving, and the slow pace of investment back to the monopoly power that allowed large corporations to administer markets and perpetuate extreme maldistribution of income and wealth.

The meaning of planning was shifted most decisively by theoretical, institutional, and practical innovations generated by the NRPB, as that New Deal agency's role was enlarged by Roosevelt from resource management to guiding government efforts to achieve recovery, and later to maintain prosperity and promote economic democracy after World War II. A body of economic knowledge produced in the early 1930s by economists such as Edward Chamberlain, Joan Robinson, Alvin Hansen, and John Maynard Keynes explained how a stubborn equilibrium, well below the level of full employment of the factors of production, could have been reached and sustained over nearly a decade. Building upon this, economists at NRPB and the Federal Reserve zeroed in on underconsumption as a factor in the slow and uneven pace of recovery and pointed to gains to be achieved by more aggressive deficit financing of public expenditures in the form of worthwhile civil works. By the definition of planning that flowed from this new thinking, the state became responsible for continuously monitoring and occasionally dramatically augmenting or shifting investment flows in socially constructive directions not attractive to the private capital markets, and for limiting entrenched corporate power over prices and redistributing income to support consumption. Among the most advanced elaborations of New Deal statism, *Security*, *Work*, *and Relief Policies* (1942) conveyed NRPB's synthesis of three related but discrete government roles: guaranteeing work for the segment of the population securely attached to the labor market, financing and extending social insurance to care for those unable to participate in the work economy, and underwriting a still highly contested version of industrial democracy through the NLRB and the self-organization of labor. Through its promotion of this triad, the NRPB enunciated a liberal statist's vision of a new social contract.

New Deal planners managed only a partial and temporary fulfillment of the moralizing role Henry Carter Adams had charted for knowledge-based civic discourse back in the 1880s, implemented in limited ways through the Fair Labor Standards Act and the Fair Employment Practices Commission. The NRPB's interest in postwar planning was one of the factors that led to the unceremonious termination of the agency in 1943, a milestone in the history of new liberalism and in the history of the social sciences as well. The commitment of New Deal planners to full employment proved untenable in the face of rising opposition to the New Deal. Though the Council of Economic Advisers authorized by the watered-down Employment Act of 1946 lacked the authority to trigger government spending at a level needed to achieve full employment, the agency continued at least through the Truman era to press for a major government role in the postwar expansion of the economy, and to jawbone against monopoly as an obstacle to growth.

Meanwhile, opponents of planning clustered around nodes represented by the resurgent free market liberalism of Frederich Hayek and by concerns among business conservatives and Southern industrialists, fanned by legal activism and the uproar created by *American Dilemma*, that federal government controls might invade local prerogatives regarding labor and race. The most influential opponents of government authority in these regards were undoubtedly corporate elites and corporate liberal thinkers who considered the associative vision of business–government cooperation a better method longterm. Herbert Stein, Allan Brinkley, and others have located a 'fiscal revolution' in economic policy thinking, underway by 1937, in which a competing body of policy knowledge suggested that the measured use of fiscal and monetary levers would enable macroeconomic stabilization without intrusion into capitalist decision-making processes or extensive interference with labor markets, which in this view could safely be left to private management in the context of 'commercial Keynesianism' and a compensatory state.

Whereas the first generation of American liberal-statist social scientists had reconstituted government as a knowledgeable, morally responsible regulator, and the second generation of planners had modeled the state as social insurer and mediator of class antagonisms, the third, New Deal generation of planners conceived a state role not well captured as state capitalism. It was closer to older Hamiltonian and Whig models of a development state responsible for liberating and energizing community efforts and for facilitating the balanced economic and civic development of the commonwealth. Policy discourses of the 1950s, developed around Cold War internal security issues, an increasingly consensual growth-oriented commercial Keynesianism, modernization theories exported from the United States as technologies for liberal internationalism, and domestic discourses of racial and gender identity, challenged the

assumptions of new liberal social science, fostered new methods for interpreting modern social relations, and inaugurated a new ideological era. But embodied in resilient structures of disciplinary knowledge formation, in durable habits that demand grounding of political argument in social empiricism, and in a chastened but resilient Enlightenment tradition of political obligation, the legacy remains—indeed, in important though little recognized ways, recurs in the dialogue between contemporary liberals and communitarians. Absent the polarizing tendencies of a world divided for half a century between capitalism and state communism, and given now the resurgence over recent decades of libertarianism and free market ideology, it becomes ever more obvious that the critical decisions in democratic governance in the West, and beyond, will henceforth be—as they were in the new liberal era, 1880s to 1940s—hard choices between variants of liberalism.

See also: Capitalism; Feminist Theory: Liberal; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Social Justice; Social Question: Impact on Social Thought

Bibliography

Alchon G 1985 *The Invisible Hand of Planning: Capitalism, Social Science, and The State in The 1920s*. Princeton University Press, Princeton, NJ

Barber W 1996 *Design Within Disorder: Franklin D. Roosevelt, the Economists, and the Shaping of American Economic Policy, 1933–1945*. Cambridge University Press, New York

Berkowitz E, McQuaid K 1992 *Creating The Welfare State: The Political Economy of Twentieth-century Reform*, rev. edn. University Press of Kansas, Lawrence, KN

Brinkley A 1995 *The End of Reform: New Deal in Recession and War*. A. A. Knopf, New York

Clawson M 1981 *New Deal Planning: The National Resources Planning Board*. Johns Hopkins University Press, Baltimore, MD

Collini S 1979 *Liberalism and Sociology: L. T. Hobhouse and Political Argument in England, 1880–1914*. Cambridge University Press, New York

Croly H 1909 *The Promise of American Life*. Macmillan Co., New York

Eisenach E 1994 *The Lost Promise of Progressivism*. University Press of Kansas, Lawrence, KN

Ethington P 1992 Hypotheses from Habermas: Notes on reconstructing American political and social history, 1890–1920. *Intellectual History Newsletter* **14**: 21–40

Farr J 1989 Understanding conceptual change politically. In: Ball T, Farr J, Hanson R (eds.) *Political Innovation and Conceptual Change*. Cambridge University Press, New York pp. 24–49

Fraser S, Gerstle G 1989 *The Rise and Fall of the New Deal Order, 1930–1980*. Princeton University Press, Princeton, NJ

Freeden M 1978 *The New Liberalism: An Ideology of Social Reform*. Clarendon Press, Oxford, UK

Furner M O 1975 *Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science, 1865–1905*. The University Press of Kentucky, Lexington, KY

Furner M O, Supple B (eds.) 1990 *The State and Economic Knowledge: The American and British Experiences*. Cambridge University Press, Cambridge, UK

Gerstle G 2001 *American Crucible: Race and Nation in the Twentieth Century*. Princeton University Press, Princeton, NJ

Hawley E 1966 *The New Deal and the Problem of Monopoly*. Princeton University Press, Princeton, NJ

Heclo H 1974 *Modern Social Politics in Britain and Sweden*. Yale University Press, New Haven, CT

Jackson W 1990 *Gunnar Myrdal and America's Conscience: Social Engineering and Racial Liberalism, 1938–1987*. University of North Carolina Press, Chapel Hill, NC

Jacoby S 1997 *Modern Manors: Welfare Capitalism Since the New Deal*. Princeton University Press, Princeton, NJ

Joas H 1993 *Pragmatism and Social Theory*. University of Chicago Press, Chicago, IL

Judis J B 2001 *The Paradox of American Democracy: Elites, Special Interests, and the Betrayal of the Public Trust*. Pantheon Books, New York

Kloppenberg J 1986 *Uncertain Victory: Social Democracy and Progressivism in European and American Thought, 1870–1920*. Oxford University Press, New York

Koven S, Michel S 1990 Womanly duties: Maternalist politics and the origins of welfare states in France, Germany, Great Britain, and the United States, 1880–1920. *American Historical Review* **95**: 1076–108

Kraditor A 1971 *Ideas of the Woman Suffrage Movement, 1890–1920*. Columbia University Press, New York

Lacey M J, Furner M O 1993 *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge, UK

Lacey M (ed.) 1989 *The Truman Presidency*. Cambridge University Press, Cambridge, UK

Milkis S M, Mileur J M (eds.) 1999 *Progressivism and the New Democracy*. University of Massachusetts Press, Amherst, MA

O'Connor A 2001 *Poverty Knowledge: Social Science, Social Policy, and the Poor in Twentieth Century US History*. Princeton University Press, Princeton, NJ

Patterson J T 2000 *America's Struggle Against Poverty in the Twentieth Century*. Harvard University Press, Cambridge, MA

Persons S 1987 *Ethnic Studies at Chicago, 1900–45*. University of Illinois Press, Urbana, IL

Reagan P 1999 *Designing a New America: The Origins of New Deal Planning, 1890–1943*. University of Massachusetts Press, Amherst, MA

Rodgers D 1998 *Atlantic Crossings: Social Politics in a Progressive Age*. Belknap Press of Harvard University Press, Cambridge, MA

Ross D 1991 *Origins of American Social Science*. Cambridge University Press, Cambridge, UK

Rueschmeyer D, Skocpol T (eds.) 1996 *States, Social Knowledge, and the Origins of Modern Social Policies*. Princeton University Press, Princeton, NJ

Silverberg H (ed.) 1998 *Gender and American Social Science: The Formative Years*. Princeton University Press, Princeton, NJ

Sklar K K 1993 The historical foundations of women's power in the creation of the American welfare state, 1830–1930. In:

Koven S, Michel S (eds.) *Mothers Of A New World: Maternalist Politics and the Origins of Welfare States*. Routledge, New York pp. 43–93
Sklar M 1988 *The Corporate Reconstruction of American Capitalism, 1890–1916: The Market, Law, and Politics*. Cambridge University Press, New York
Skocpol T 1992 *Protecting Soldiers and Mothers: The Political and Social Origins of Social Policy in the United States*. Belknap Press of Harvard University Press, Cambridge, MA
Stein H 1969 *The Fiscal Revolution in America*. University of Chicago Press, Chicago, IL
Tomlins C 1985 *The State and the Unions: Labor Relations, Law, and the Organized Labor Movement in America, 1880–1960*. Cambridge University Press, Cambridge, UK

M. O. Furner

Policy Knowledge: Nonprofit Institutions

The complexity of advanced industrial systems requires governments to acquire an enormous range of specialized knowledge and administrative expertise with which to order their affairs. Uniformity in demand, however, has not produced uniformity in the character and supply of knowledge for government purposes. There are significant variations, related to diverse cultural traditions in the means of acquiring knowledge and applying methods of expertise. For example, central government bureaucracies in Western Europe and Japan directly provide many more sources and opportunities for the development of policy expertise than they do in the USA, which in modern times has relied heavily on private research organizations operating outside government bureaucracies for functional expertise. As the federal role in economic and social policy expanded in the post-World War II era, increased federal funding did not produce a commensurate increase in the size of government. Instead, it strengthened the complicated network of relationships among the federal government, state and local governments, and private groups, including nonpartisan research institutes (Alchon 1985, Critchlow 1985, Ricci 1993, Smith 1991).

In this policy system, the federal government operates indirectly, through intermediary organizations in state and local government, and in civil society (Heclo 1980). This reliance on private, nonpartisan research organizations imparts a unique quality to the acquisition and use of knowledge by the administrative state in modern America.

The vast size, complexity, and expense of this system of acquiring knowledge by government in the USA suggests the pluralistic nature of the American polity, deeply rooted in its republican distrust of centralized government. This distrust of centralized government found expression in the nation's constitution, drafted in Philadelphia in 1787. Its famous beginning 'We the people' asserted that the source of its authority rested in the people rather than state or centralized government. This distrust of centralized government found expression in the antistatist Jeffersonian–Jacksonian tradition that emerged early in the nation's history.

1. *Social Investigation in the Nineteenth Century*

In the environment of relatively weak centralized government prior to the US Civil War, the craft of social investigation developed within a middle-class subculture of reformers, many of them women, who documented the plight of slaves, the poor, prisoners, and the insane. The affinity of antebellum social reform and feminism became evident in reports issued by American abolitionists who documented the conditions of slave life, as well in reports on the care of prisoners, the indigent, and the insane. In the 1840s, social reformer Dorothea Dix issued a series of groundbreaking reports to state legislatures calling for the humane care of the mentally ill. This tradition of empirical research into social problems conducted by private individuals associated with reform movements was extended late into the nineteenth century (Marshal 1937, Nelson 1992, Leach 1989, Daniels 1968, Dupree 1957).

By the outbreak of the Civil War, state legislatures in the most populous states regularly received detailed narrative and statistical reports on dependency, crime, public health, labor conditions, and economic development. These reports relied more on democratic and humanitarian sentiments associated with moral reform than they did on claims of science and disinterested expertise. Female reformers, as suggested by the role of Dorthea Dix, played an especially important role in this early endeavor to bring social knowledge to state legislatures and the public (Furner and Lacey 1993, Ross 1991).

While the capacity for social knowledge grew on the state level, the role of the federal government's activities in the pursuit of knowledge remained limited generally to surveying, land exploration, and modest technological development through private contracts. This relative institutional impoverishment reflected a unique antistatist tradition in America, reinforced by sectionalist politics of the period. The major institution in Washington for the gathering and accumulation of scientific and social knowledge remained the Smithsonian Institution, which opened in 1846 at the instigation of the individual for whom it was named. In the decades following the Civil War, the Smithsonian Institution became a center for people such as Francis Walker, John Eaton, John Wesley Powell, and Carroll D. Wright, who took leading roles in establishing a rudimentary infrastructure of social investigation in the nation's capital. These individuals served

as conduits for a transition from moral reform to scientific reform.

2. The Twentieth Century

By the beginning of the twentieth century a large resident class of scientific and professional civil servants had emerged in Washington. In 1904, the US Bureau of Census reported that 3,919 civil servants fell into the classification of 'professional, technical and scientific' workers. Within this group, nearly 1,100 people were described as 'special agents, experts, and commissioners,' while another 300 or so were designated simply as 'scientific experts and investigators' (Lacey 1993, pp. 127–70, Ross 1991, Blumer et al. 1991, Keller 1977, Skowronek 1982, Wilson 1909).

Many of the early organizers in the Smithsonian Institution played a critical role in the establishment of the American Social Science Association (ASSA) in 1865. The ASSA sought to serve as a clearinghouse of social knowledge in order to promote reform in education, public health jurisprudence, finance, and social economy. Directly influenced by the achievements of the Sanitary Commission during the Civil War and modeled on the British National Association for the Promotion of Social Science, the ASSA attracted commercial elite groups interested in promoting sound fiscal and monetary policy; lawyers, doctors, ministers, and teachers who sought professional affirmation; and middle-class reformers who combined humanitarian moral reform with the new scientific spirit. In this way, the ASSA, and other professional organizations that arose during this period, paved the way for a new liberalism, grounded in social empiricism (Haskell 1977, Furner 1975, 1993).

These reformers called for new measures of efficiency and economy in voluntary charity organizations and in public administration. Behind this concern with efficiency and economy lay a deep conviction that the polity was threatened by corruption of partisan political machines serving special economic interests. These political machines, run by local and state political bosses, consistently manipulated the mass of voters to maintain political power in flagrant disregard of any notion of public morality or efficiency and economy in government. Thus progressive reformers called for new social knowledge to help restore political order and representative government to American society. Beneath the catchphrase 'efficiency and economy'—the rallying cry of reformers—lay strong antimajoritarian values fused with antiparty perspectives (Critchlow 1985, Alchon 1985, Karl 1974).

This movement for efficiency found early expression in late nineteenth-century charity endeavors. Amid concern that relief was being wasted on the undeserving poor, new professionals were brought into charity organization to centralize recipient roles, provide vouchers for relief funds to be distributed to recipients, institute periodic checks on recipients, and implement budget systems within agencies to keep a close watch on funds. Essential to this reform was the use of data gathering through statistical measures and new accounting procedures. Behind this movement lay a sentiment expressed by industrialist Andrew Carnegie, who declared, 'One of the most serious obstacles to the improvement of our race is indiscriminate charity.' Carnegie's infatuation with the new social sciences would lead him to create a number of new nonprofit philanthropic institutions concerned with the gathering of public knowledge (Lagemann 1989).

At the vanguard of this movement for efficiency and economy in government were university trained and highly specialized social scientists. In the latter half of the nineteenth century, the basis of social science research was transformed from deductive reasoning derived from natural law into inductive procedure founded on historical analysis and empirical research. This new methodology allowed social scientists to mask any preoccupation with political matters in the terminology of scholarly objectivity and empirical research. The transformation of political impulses into the quest for public knowledge and administrative solutions reflected the aspirations of a new professional class anxious to gain influence in a highly industrialized society through expertise. They sought reform by placing 'the best people'—objective-minded experts—in government (Ross 1991, Haskell 1977).

The call for efficiency and economy in government, and the need to develop new public agencies for the gathering of public knowledge essential to restoration of representative government, began on the municipal level first in New York with the creation of a Bureau of Municipal Research in 1905, organized by social scientists and civic-minded philanthropists and businessmen. Municipal research bureaus would be established in several other US cities. Central to these municipal research bureaus was the creation of executive budget systems that allowed a strong executive office, public officials, and the general electorate to track how public funds were spent. Essential to the functioning of an executive budget was the creation of an agency staffed by nonpartisan experts who would prepare and analyze budgetary finances and expenditures (Critchlow 1985, Karl 1974, Smith 1991).

This demand for an executive budget system reached the national level under the administration of William H. Taft in 1910. When Taft assumed office in 1909, the federal budget stood at over a billion dollars and the federal debt had reached disturbing proportions, largely as an aftermath of the 'Panic of 1907.' In these circumstances, Taft was able to persuade Congress to appropriate $100,000 for a commission on efficiency and economy in government. To staff the commission, Taft called on the best social scientists in the country. After two years of painstaking work, the Taft Com-

mission came out with its report, *The Need for a National Budget*. Yet, because the proposal was made in an election year, a recalcitrant Congress refused to enact an executive budget system. The failure of the Taft Commission did not dismay reformers, however. In 1915 they created the Institute for Government Research—the predecessor to the Brookings Institution. Only in 1921 during the Harding administration would budget reform be enacted in the Budget and Accounting Act of 1921.

Taft's successor, Woodrow Wilson reflected in many respects the deep ambivalence that middle-class Americans felt about the state during the early twentieth century. An array of new government agencies was now involved in conservation, consumer protection, corporate regulation, trade practices, and monetary and financial policies. Yet Wilson, in his pioneering essay on public administration warned that the model of Prussian administration should be avoided in America. He declared that European bureaucratic practices would have to be adapted to fit the highly decentralized form of government in the USA. If the European model is to be employed in the USA, he told his fellow reformers, 'We must Americanize it … We must learn our constitutions by heart and we must get the bureaucratic fever out of its veins' (quoted in Cuff 1993, p. 390).

Until America's entry into the war in April 1917, neither the federal government's defense department nor civilian departments had considered mobilization tasks. As a consequence, private planning proceeded public planning. Financial and manufacturing firms such as J. P. Morgan and Du Pont Corporation, by receiving Allied war orders, were among the first to assess the financial, technical, and labor requirements of munitions production (Brownlee 1993).

War, however, forced the enormous administrative expansion of the federal government in the short period between April 1917 and November 1919. New government agencies appeared in haphazard fashion, drawing corporate and academic personnel into government management. This involvement led many to the belief that the capacity for public knowledge and for effective social action rested in the mobilization of private, public-spirited professional and business groups willing to work in cooperation with a sympathetic state. The war demonstrated for these individuals how to forge a uniquely American combination of state–society relations.

Herbert Hoover sought to put this insight into practice as Secretary of Commerce in the Harding and Coolidge administrations, and later as president, by using the techniques of private–public cooperation he had learned as a wartime administrator. Hoover sought to develop state–voluntary cooperation in the collection of specialized data required for rational decision-making in both the private and public sectors. For Hoover and others concerned with public knowledge, the war revealed just how essential government knowledge was for economic and social planning. Through the development of an administrative state he sought to develop new mechanisms for collecting public knowledge without the coercive element of state power (Hawley 1974).

In keeping with his views about the credibility of privately financed researchers, Hoover arranged for the Rockefeller Foundation, founded in 1913, to fund a comprehensive survey of social changes in American life between 1900 and 1930. The three-year study resulted in the massive 1,600-page *Recent Social Trends in the United States: Report of the President's Research Committee of Social Trends* (1933). This study presented the full panoply of available knowledge for public purposes (Karl 1974, Hawley 1974, Alchon 1985).

Although state capacity for gathering public knowledge expanded dramatically during the Great Depression era, World War II, and the postwar period, the federal government continued to rely indirectly on intermediary organizations in state and local government, and in civil society, for the accumulation of knowledge. The commensurate increase in the size of the federal government led to the creation of fluid networks of policy specialists, who operated in and between government and the private sector, identifying policy issues, generating 'solutions' to policy problems, and overseeing and evaluating the implementation and administration of policies within their specific areas of competence (Heclo 1980).

3. *Role of the Think Tanks*

Essential to these policy networks are private policy research institutes or 'think tanks,' a peculiarly American social invention. By the 1990s there were an estimated 1,200 non-government think tanks of various descriptions. World War II and the postwar continuation of a large defense program, as well as expansion of federal social programs, was a watershed in the evolution of the American think tank.

At the end of World War II, Americans made a political decision to have government train, subsidize, and regularly consult a sizable population of civilian experts on such issues as defense, welfare, health, and economic stability. Of the major think tanks, only the Brookings Institution (1916) and the Carnegie Endowment for International Peace (1910) were founded before World War II. The American Enterprise Institute was founded during the war in 1943 (Critchlow 1985, Smith 1991, Lagemann 1989, Dickson 1971).

Although think tanks are ostensibly nonpartisan, in many instances they function as extensions of state power, their influence rising and falling with changes in governments and shifts in the ideological climate of the country. In other cases, think tanks function more independently, questioning and monitoring state strat-

egies and structures. For example, the Rand Corporation of the 1940s and 1950s was designed to strengthen the programs of the US Air Force, whereas the Institute for Policy Studies (IPS) has been a major center of criticism of US foreign policy, domestic social programs, and policies in the areas of civil rights and poverty.

Often the ideological orientation of these think tanks reflected changes in the historical context and political climate, as indicated by the dominant trends in high and popular culture, changes in voting behavior, and shifts in political leadership (Ricci 1993).

The course of the Brookings Institution from the New Deal through the end of the twentieth century reflected these ideological currents. While Brookings studies in the early 1930s criticized New Deal programs, the Brookings Institution under its president Harold Moulton tried to avoid becoming directly involved in partisan politics. By 1937, however, the institution had emerged as a vociferous critic of New Deal fiscal policy.

In response to the new postwar environment and the reluctance of foundations to fund an institution they perceived as ineffective or out of touch, Robert Calkins, former head of the General Education Fund at the Rockefeller Foundation, agreed to head the institution. Calkins reorganized Brookings and recruited social scientists with liberal credentials and government experience. This new group had close ties with government, and unlike the devotees of the earlier nonpartisan ideal, aligned themselves closely with presidential administrations. In 1965, when Calkins retired, Brookings was representative of mainstream Keynesian economic thinking, and its growing influence was reflected in renewed foundation support, especially from the Ford Foundation. Under Calkins' successor, Kermit Gordon, Brookings' reputation as a liberal Democratic think tank was well entrenched. As a result, the institution became a major research center and a holding area for policy experts who were temporarily out of office. In these years, the Brookings Institution became a major center for policy innovation in welfare, healthcare, education, housing and taxation policy (Critchlow 1985, Smith 1991).

Divisions over foreign and domestic policy opened cracks in the liberal consensus in the early 1970s, however, leading many observers to see the institution as isolated from large segments of business. In 1976, the board of trustees appointed Bruce MacLaury to head the institution. A former regional Federal Reserve banker and Treasury official, MacLaury successfully courted business support, increased corporate representation on the board of trustees, and moved the institution toward a more moderate ideological stance.

By the 1970s, the Brookings Institution confronted competition from other major policy research institutions, especially the American Enterprise Institute and the Heritage Foundation—both viewed as conservative research institutions close to the Republican Party.

The American Enterprise Institute (AEI), which was founded in 1943 as the American Enterprise Association (AEA), illustrates the experience of a conservatively oriented research institution that expressed deep ambivalence about the post-World War II policy consensus (Critchlow 1985, Smith 1991). The key figure behind the establishment of the AEA was Lewis Brown, chairman of Johns-Manville Corporation. Brown realized that the New Deal had changed the political environment for business and, although he expressed anxiety about the level of government intervention, he accepted much of the New Deal program. From the start, the AEA reflected a conservative bias.

In 1954, A. D. Marshall, head of General Electric, assumed the institution's presidency and immediately hired William Baroody Sr. and W. Glenn Campbell, both staff economists at the US Chamber of Commerce, to head the research program. Under their guidance, AEA was gradually built into a modern research institute under its new name, the American Enterprise Institute. Principal support came from private foundations and major corporate sponsors. The institution's reputation was enhanced when the Nixon administration called upon a number of AEI associates to fill government positions. The AEI also emerged as a successful proponent of economic deregulation.

In 1977 Baroody retired and his son, William Baroody Jr., took over the presidency of the institution. To improve its standing in the academic community, the AEI assembled an impressive staff including Melvin Laird, William Simon, Robert Bork, Michael Novak, and Herbert Stein. The tenure of William Baroody Jr., however, ended abruptly in the summer of 1987, when an increasingly restive board of trustees forced his resignation because of cost overruns and declining revenues. Moreover, despite the large number of AEI staffers who held appointments in the Reagan administration (30 left AEI to join the administration in its first two years alone), members of the New Right, suspicious of what they considered the centrist orientation of the institution, voiced bitter criticism of the AEI.

Baroody's successor, Christopher DeMuth, bolstered the conservative orientation of the institute by bringing on board several former Reagan administration officials with strong rightist reputations.

The founding of the Heritage Foundation in 1973 revealed a new ideological climate in the analysis of public knowledge (Edwards 1997, Peschek 1987). Founded by Edwin Feulner and Paul Weyrich to provide rapid and succinct legislative analysis on issues pending before Congress, the Heritage Foundation sought to promote conservative values and demonstrate the need for a free market and a strong defense. The Heritage Foundation's articulation of conserva-

tive values in social policy, education, and government activities placed it at the forefront of New Right activity. Although sympathetic to the Reagan wing of the Republican Party, Heritage's commitment to conservative economic theories and social values allowed the institution to be critical of both Republicans and Democrats. The Heritage Foundation remained relatively small in its early years, but the election of Ronald Reagan to the presidency in 1980 enhanced the institution's prestige. By the mid-1980s the Heritage Foundation had established a solid place in the Washington world of think tanks as a well-organized, efficient, and well-financed research organization that called for the turning over of many government tasks to private enterprise, a strong defense, and a cautious approach to Russia and China.

During these years, myriad other think tanks emerged in Washington representing a range of ideological positions and specialized policy interests, including the left-oriented Institute for Policy Studies (1963) and the libertarian-oriented CATO Institute (1977). Think tanks concerned with national security included the Center for Strategic and International Studies (1962) and the Center for National Security Studies (1962) affiliated with the American Civil Liberties Union (Friedman 1997).

These private institutions served to supplement social research conducted by federal agencies and institutions. By the 1990s the number of personal and committee staff of Congress exceeded 11,000, with a further 20,000 people responding to the information needs of the legislature from the Congressional Budget Office, the Office of Technological Assessment, and the General Accounting Office (Lacey 1993). The executive branch commands far greater numbers. By the 1990s the scale of spending by the federal government for the acquisition and application of public knowledge was massive. Funding is currently routed through an estimated 180 bureaus and agencies. This research includes a mixture of 'in-house' research conducted by career civil servants, and a larger body of research conducted by outside parties via grants and contracts on a 'for profit' basis and some of it by university-based scholars on a 'not for profit' basis. An estimated third of this research activity is supported by the federal government to meet its needs, while the remaining two-thirds is authorized for state and local government, school systems, policy forces, welfare agencies, and industrial and agricultural groups. In addition, a vast array of trade and professional organization and 'public interest' groups conduct research in the areas of environmental policy, consumer affairs, civil rights, and foreign policy issues (Critchlow 1985). While the government relied on think tanks for policy innovation, program review, and policy assessment, the federal government relied on nonprofit service providers to implement social programs. In this way, the expansion of the welfare state in the 1960s and 1970s created a seeming paradox that remains apparent in the contemporary regime: the rapid growth of the nonprofit, voluntaristic sector occurred at the same time that federal social policy became increasingly centralized and active. The parallel development of the centralized welfare state and the voluntary sector was no coincidence. The growth of federal social programs mobilized the nonprofit sector, while in turn the nonprofit sector, through policy innovation and program implementation, helped legitimize the modern welfare state.

Thus, as the traditional federal system that had existed since the founding of the nation eroded in the 1960s, when the federal government extended its powers at the expense of state power, the nonprofit sector was called upon to help design, implement, and administer social programs that had once been within the jurisdiction of the public sector (Hall 1992, Kameran and Kahn 1989, Salamon 1992). As a consequence, the nonprofit sector expanded in the 1960s at the very time that the federal government enlarged its powers over the states. The result produced an unusual phenomenon—government centralization produced greater nonprofit sector voluntarism (Critchlow 1999).

4. *The 'Mixed' Welfare State*

Between 1965 and 1980 the American social welfare state underwent a significant transformation, with government spending on social welfare increasing by 263 percent in inflation-adjusted dollars (Gilbert 1983, Critchlow 1999). The expansion of the welfare state promoted the emergence of a government–nonprofit partnership. Increasingly in the 1960s, the federal government generated the funds but then turned over the delivery of services to other public state and local organizations, and private nonprofit organizations. The private, nonprofit sector became a major beneficiary of this system of services. By the 1980s, the federal government alone provided $40 billion in support to the private nonprofit sector, while state and local governments provided additional amounts from their own resources (Salamon 1992).

Federal demand enabled the nonprofit sector to grow by leaps and bounds. In 1940 there were only 12,500 charitable, tax exempt organizations. By 1990, there were over 700,000, with most of this growth taking place after the 1960s (Salamon 1992, Hall 1992, Critchlow 1999). While the federal government centralized social policy through congressional mandates, administrative and legislative regulations, and new funding arrangements, the delivery and implementation of these programs were placed in the hands of state and local governments and nonprofit organizations.

The emergence of the 'second welfare state' in the last three decades of the twentieth century—contrary

to the concerns of those who feared that government would replace the nonprofit sector altogether—created a vibrant nonprofit sector. Government, in turn, came to rely on this sector to design and implement programs, undertake demonstration projects, contract with other private agencies, and provide outside consultation and independent review agencies and activities.

As this second welfare state emerged, the nonprofit sector was called upon to fulfill a number of distinct functions. Nonprofit organizations served as *funding intermediaries* distributing funds to other parts of the nonprofit sector, *service-providing organizations* involved in healthcare, education, counseling, adoption assistance, and *political action agencies* supporting particular pieces of legislation or candidates for political office (Salamon 1992). By providing public services, these nonprofit organizations developed a symbiotic relationship with the federal government. By 1990, government grants, contracts, and reimbursements accounted for 31 percent of nonprofit service organizations' income. Just 18 percent of the total income that nonprofit organizations receive comes from private giving (Salamon 1992, Hall 1992).

The proliferation of government and nonprofit institutions involved in public policy marshals expert against expert, testing the credibility of each and enriching the information environment in which policy discourse occurs and decisions are taken. Indeed, on highly technical issues, opposing experts often directly contradict one another. Yet the growth of public and private agencies involved in collecting public knowledge allows for a more democratized policy process—much closer, in a sense, to the Jeffersonian–Jackson vision of an open, competitive democracy.

Still, two qualifications are in order: neither Jefferson nor Jackson envisioned a world of competing experts representing various interests and ideological positions; and secondly, this contemporary world of competing institutional research is not equally open to all comers. Access to financing is one limitation, and access to those in power is another.

The proliferation of expertise and the expansion of government knowledge that has arisen in response to the demand for more policy inquiry have not necessarily improved the information base on which political decisions are made. The emergence of more complicated patterns of political mobilization by a larger and more diverse array of interests and groups has necessarily politicized the nature of policy knowledge and thereby redefined the nonpartisan ideal espoused at the beginning of the twentieth century. Competition for attention to research places heavy demands on the time and attention of legislators and policy-makers who, in a value-laden universe, are left, ultimately, to decide what information counts for them on the basis of their own values. A widespread acceptance of relativism, embodied institutionally in public and private research organizations, has irrevocably altered the process of social investigation and the uses of social knowledge.

See also: Centers for Advanced Study: International/Interdisciplinary; Intellectuals, Sociology of; Nonprofit Sector, Geography of; Policy Information; Policy Knowledge: Advocacy Organizations; Policy Knowledge: Universities; Strategic Planning; Think Tanks

Bibliography

Alchon G 1985 *The Invisible Hand of Planning: Capitalism, Social Science, and the State in the 1920s*. Princeton University Press, Princeton, NJ

Bensel R F 1990 *Yankee Leviathan: The Origins of Central State Authority in America, 1859–1877*. Cambridge University Press, Cambridge, UK

Blumer M, Bales K, Kish Sklar K (eds.) 1991 *The Social Survey in Historical Perspective, 1889–1940*. Cambridge University Press, Cambridge

Brewer J 1990 *The Sinews of Power: War, Money and the English State, 1688–1783*. KNOPF, Cambridge, MA

Brock W R 1984 *Investigation and Responsibility: Public Responsibility in the United States, 1865–1900*. Cambridge University Press, Cambridge, MA

Brownlee W E 1993 Social investigations and political learning in the financing of World War II. In: Lacey M J, Furner M O (eds.) *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge, pp. 323–61

Cassedy J H 1969 *Demography in Early America: Demography in Early America: Beginnings of the Statistical Mind*. Harvard University Press, Cambridge, MA

Cohen P C 1982 *A Calculating People: The Spread of Numeracy in Early America*. Chicago University Press, Chicago

Critchlow D T 1985 *The Brookings Institution, 1916–1952: Expertise and the Public Interest in a Democratic Society*. Northern Illinois University Press, DeKalb, IL

Critchlow D T 1999 *Intended Consequences: Birth Control, Abortion and the Federal Government*. Oxford University Press, New York

Cuff R D 1993 War mobilization, institutional learning, and state building in the United States, 1917–1941. In: Lacey M J, Furner M O (eds.) *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge, pp. 388–427

Daniels G H 1968 *American Science in the Age of Jackson*. Columbia University Press, New York

Dixon P 1971 *Think Tanks*. New York

Dupree, Hunter A 1957 *Science in the Federal Government*. Harvard University Press, Cambridge, MA

Edwards L 1997 *The Power of Idea: The Heritage Foundation at 25 Years*. Jameson Books, Ottawa, IL

Friedman J (ed.) 1983 *The First Harvest: An Institute for Policy Studies Reader, 1963–1983*. Grove Press, New York

Furner M O 1975 *Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science, 1865–1905*. University of Kentucky Press, Lexington, KY

Furner M O 1993 The Republican tradition and the new liberalism: Social investigation, state building, and social learning in the gilded age. In: Lacey M J, Furner M O (eds.)

The State and Social Investigation in Britain and the United State. Cambridge University Press, Cambridge, pp. 171–242
Furner M O, Lacey M J 1993 Social investigation, social knowledge and the state: An introduction. In: Lacey M J, Furner M O (eds.) *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge
Furner M O, Supple B (eds.) 1990 *The State and Economic Knowledge: The American and British Experiences*. Cambridge University Press, Cambridge
Gilbert N 1983 *Capitalism and the Welfare State: Dilemmas of Social Benevolence*. Yale University Press, New Haven, CT
Hall P D 1992 *Inventing the Nonprofit Sector and Other Essays on Philanthropy, Voluntarism, and Nonprofit Organizations*. Johns Hopkins University Press, Baltimore
Haskell T 1977 *The Emergence of Professional Social Science: The American Social Science Association and the Nineteenth Century Crisis of Authority*. Illinois University Press, Urbana
Hawley E 1974 Herbert Hoover, the commerce secretariat, and the vision of an associative state, 1921–1929. *Journal of American History* **61**(June)
Heclo H 1980 Issue networks and the executive establishment. In: King A (ed.) *The New American Political System*. American Enterprise Institute Press, Washington, DC
Kameran S B, Kahn A J 1989 *Privatization and the Welfare State*. Princeton University Press, Princeton, NJ
Karl B 1974 *Charles Merriam and the Study of Politics*. University of Chicago Press, Chicago
Karl B 1983 *The Uneasy State: The United States from 1915 to 1945*. University of Chicago Press, Chicago
Keller M 1977 *Affairs of State: Public Life in Late Nineteenth Century America*. Harvard University Press, Cambridge, MA
Lacey M J 1993 The world of the bureaus: Government and the positivist project in the late nineteenth century. In: Lacey M J, Furner M O (eds.) *The State and Social Investigation in Britain and the United States*. Cambridge University Press, Cambridge
Lagemann E C 1989 *The Politics of Knowledge: The Carnegie Corporation, Philanthropy, and Public Policy*. University of Chicago Press
Leach W 1989 *True Love and Perfect Union: The Feminist Reform of Sex and Society*. Basic Books, New York
Marshall H E 1937 *Dorothea Dix: Forgotten Samaritan*. University of North Carolina Press, Chapel Hill, NC
National Academy of Sciences 1978 *The Federal Investment in Knowledge of Social Problems*. Washington, DC
Nelson W 1992 *The Roots of American Bureaucracy, 1830–1900*. Cambridge
Peschek J G 1987 *Policy-planning Organizations: Elite Agendas and America's Rightward Turn*. Temple University Press, Philadelphia
Ricci D 1993 *The Transformation of American Politics: The New Washington and the Rise of Think Tanks*. New Haven, CT
Ross D 1991 *The Origins of American Social Science*. Cambridge University Press, Cambridge
Salamon L M 1992 *America's Nonprofit Sector: A Primer*. Foundation Center, New York
Skowronek S 1982 *Building a New American State: The Expansion of National Administrative Capacities, 1877–1920*. Cambridge University Press, Cambridge
Smith J A 1991 *The Idea Brokers: Think Tanks and the New Policy Elite*. The Free Press, New York
Wilson W 1909 *The State: Elements of Historical and Practical Politics*, rev. edn. D C Health and Company, Boston

D. T. Critchlow

Policy Knowledge: Universities

All civilizations have needed higher education to train their ruling, priestly, military, bureaucratic, and other service elites. However it was in twelfth-century Europe that an institution recognizable as a university came into being, as a school of higher learning combining teaching and scholarship and characterized by a corporate autonomy and academic freedom. This institution of higher learning has survived in varying forms down to the present day, while evolving a flexibility which has permitted it to serve the needs of enormously different societies in every part of the world. One of those common societal needs which universities in recent years have striven to serve is that of public policy, that can be defined as national policy which is at some level an expression of the public will. Universities have come to interact with the policy process in multiple ways.

The idea of the practical application of thought to the solution of technical and social problems is virtually timeless and universal. However the direct and total involvement of the corporate universities in social and economic activity is of relatively recent origin. It is identifiable in the later stages of the Industrial Revolution and in the Europe of the nineteenth century, when the direct relevance of scientific and intellectual thought was generally discussed and debated in public. However, it was only after World War Two had dramatized the life-saving significance of knowledge production and application that in the second half of the twentieth century a clear relationship to public policy can be discerned: 'The systematic attention of universities to social reconstruction and economic innovation to all aspects of economic development—financial, industrial, entrepreneurial, organizational, managerial—its concerns for the urban infrastructure, epidemiology, family welfare and schooling, the enhancement of life chances and the encouragement of vertical and horizontal mobility, are changes occurring within living memory' (Rothblatt 1995).

1. The Foundations of the Relationship Between Universities and Public Policy

That issues of social and economic development and policy claimed the attention of the university was a combined consequence of developments within the institution itself and in its external political context. The great nineteenth-century development within the university was the emergence of a new paradigm of the institution, in which research was added to teaching as an equal part of its mission.

Twentieth-century universities have systematically developed structures that allowed them to add the function of generating knowledge to the previous ones of preserving and transmitting it. This was followed by

the rapid development of the empirical social sciences as an expanding part of university curricula. Thus research and the production of knowledge emerged as a service to society and the social sciences arrived as a science of society. The potential of both brought the university into a systematic engagement with public policy.

Outside the university in the second half of the twentieth century came the solidification of political systems that derived more fully than before from the wishes of the national public, and required technical and research-based knowledge to satisfy those wishes and to inform policies. National competition to produce new knowledge made universities the key institutions of postindustrial society.

Historically there has always been interplay between what happens inside universities and what goes on in society outside. In the twentieth century this bond became close. The kind of people who work in universities, the type of topics on which they work and the theories, methodologies, and knowledge which they produce all have relevance to the solution of policy problems and improved social wellbeing. As problems became increasingly complex and multidimensional, cross-disciplinary and multidisciplinary concentrations began to develop within universities as a way of responding to the different problems. Specific techniques, such as survey research, manpower planning, opinion polling, rate of return analysis, and systems theory are a few among the techniques and vocabularies of social analysis which universities developed and exported, to societies wishing to digest and utilize knowledge, and to policy-makers anxious for a scientific base to their choices. There was, in the second half of the twentieth century in Europe and America, a social demand from the policy arena for the kind of knowledge, ideas and techniques which trained people within universities tend to produce.

2. *Autonomy and Relevance*

The extent of this demand created a dual temptation: within the academy for faculty members to direct their work towards perceived policy issues and, outside it, for governments, with the power of the purse, to put pressure on universities to create specific policies backed by suitable academic rationales. The result was a tension between the traditional autonomous role of the university and the demands of partisan government for knowledge that would make policies competitive.

In this situation the critical issue becomes the institutional relationship between the university and the government, which has taken over the formulation of public policy, and the nature of government's claim on the ideas and people producing them, in a context where the essence of the university ethos is autonomy. In the USA and Europe, where concepts of university autonomy are relatively well established, this autonomy was protected in large measure by a *de facto* hiving off of the policy function to independent policy research institutions. Often physically close to universities, they are staffed mainly by former university academics who focus on specific policy issues in the company of like-minded scholars.

A dense network of such institutions has grown up in the USA addressing a wide range of policy issues and there is a constant two-way traffic between them and the universities, and between both these and the government and the private sector. With regard to public policy, individuals from universities who had developed strong knowledge and skills in particular policy areas would be recruited for policy positions and consultancies in government. In the USA, Henry Kissinger, Daniel Patrick Moynihan, and Joseph Stiglitz are no more than the better known among the many who took that route in the current era. The critical point here is that they were recruited because of their individual competence and interest, not because of a particular position reflected by their university. Universities as institutions in general try to avoid espousing specific public policy positions and make a virtue of eclecticism. The guiding, albeit unstated, principle by which universities diffuse the tension is that they do not make policies but rather see themselves as enlarging the knowledge base on which policies can be made. The balance is sometimes hard to hold as when, for example, a policy philosophy becomes so strong that it is associated with the institution itself, and, in this connection, the reputation and influence of the so-called 'Chicago School' in the 1960s comes to mind. Other forms of encroachment of the policy-making function on university autonomy become controversial when the policy institute and the university appear synonymous at least to the outside eye, or when the university takes government money for training for research related to debatable policy purposes.

3. *The Internationalization of the Social Sciences*

Outside the USA and Europe the postwar era of independence in Asia and Africa, and before that in Latin America, led ultimately to the internationalization of social science and hence the university's role in relation to public policy. First, it became clear that theories derived from industrialization and market behavior did not automatically apply in rural settings, imperfect markets, or centralized and state-controlled economies. Policy-making required more contextual understanding derived from history, anthropology and cultural studies and to develop a theory of problem-solving. The content, theories, and techniques available for policy consideration were enriched. Second, the issue of autonomy and relevance, as encapsulated in the relationship between government and universities, took very different expression in developing countries than in the industrialized west.

Instead of a dense array of public policy institutes, policy and research units existed as full and functioning parts of what were to begin with often single national universities. In this period the relationship between university and government was so close that either through a honeymoon sense of postindependence common purpose, i.e., to take control of the foreign-dominated state policy-making apparatus, or because a single party regime did not see much room for university autonomy in the sense of expressing views contrary to national policy, there was little call for independent policy institutes. In many such situations in Africa the universities came close to becoming part of the civil service with a responsibility to develop and present public policy, solve problems, and in so doing be relevant.

The tension between autonomy and relevance comes to the fore and becomes more acute to the extent that the university's role in producing new knowledge and a workable science of society is supplemented by additional social demand for relevance from universities, where relevance is defined not simply in terms of the institution's ability to train its students for jobs but also to define and support specific public policies. This tension was contained by a sense of common purpose between government and universities in Africa for a period but, in Latin America, autonomy was an early casualty to conformity and the independent and knowledge-producing function moved out into independent centers which in essence became the source of alternative policies that in Africa during the same period emanated from the universities.

However, with the end of the Cold War, sweeping international forces transformed the situation and created conditions permitting a new kind of relationship between universities and the policy-making process. Economic liberalization, political democratization, and administrative decentralization all created a favorable environment for genuinely autonomous universities. Universities could now redefine their relationship with the policy-making process in a context of autonomy and chose to favor coexistence with separate policy institutes in line with the Western pattern. This move was strongly promoted by the World Bank, anxious as it was to have recognizable partners for policy dialogue in an economic context increasingly dominated by debt ratios. At the end of the twentieth century the pattern worldwide is for research centers, institutes, and think tanks to multiply on the periphery of universities while university faculties and departments are becoming the place for the provision of teaching.

4. *Universities and Public Policy in the Twenty-first Century*

Universities were never the sole producers of knowledge for policy consideration but they were paramount, and they are perhaps the only institutions with the objectivity and neutrality to train and make provision for the public good dimension of public policy. However, as they enter the twenty-first century they face global forces and technological change which promise a massive transformation and significant reduction in their paramount role. They are being required to function in a larger, more complex environment than existed when they became basic knowledge producers after World War Two. Knowledge production and dissemination, research and teaching will no longer be self-contained activities but will involve integration with a far wider variety of knowledge producers than before, including industry. As always, the universities are being required to respond to external change.

The biggest single source of change in the relationship between universities and the policy process is the emergence of the Internet and the pace of change in information technology. It is influencing enormous changes in the nature of knowledge production, the corresponding role of the university and its relationship to other actors in the informing of public policy. At the same time, globalization and economic pressures of international competition are dissolving the boundaries between national institutions and disciplines and creating widely dispersed knowledge production systems in which universities are a part but have many partners.

Universities have traditionally organized their work into disciplinary science structures. Such structures will retain a privileged place in knowledge production but they will no longer be alone. The disciplinary structure of the production of knowledge is itself changing, as we witness the emergence of a new model in which knowledge is produced in the context of its application rather than in a single theoretical disciplinary structure and is becoming heterogeneous and more socially accountable. Research done in the context of application and global distribution means that 'science' cannot remain easily within the confines of a university department or academic center. The emergence of a host of new institutional arrangements linking government, industry, NGOs, and private consultancy groups will arrange knowledge production in new forms of organization that erode the traditional place of the university.

To the extent that research is cutting loose from disciplinary structures and generating knowledge which does not seem to wish to institutionalize itself in university departments and faculties, universities will be forced to pay more attention than before to links with the new institutions and centers where the research is being generated. Coalitions, partnerships, and networks are becoming more important than institutions. In terms of their relation to policy, the old division of labor in which universities focused their knowledge production on government, and industry on private application, is less relevant and the re-

lationships which count are those with the host of diverse and globally scattered institutions which now have a voice and a hand in the production of policy-related knowledge.

See also: Centers for Advanced Study: International/Interdisciplinary; Policy Knowledge: Nonprofit Institutions; Think Tanks; Universities and Science and Technology: Europe; Universities and Science and Technology: United States; Universities, in the History of the Social Sciences

Bibliography

Altbach P, McGill Peterson P *Higher Education in the 21st Century: Global Challenge and National Response*. Institute of International Education, Annapolis

Caraca J, Conceicao P, Heitor M V 2000 Towards a public policy for the research university in Portugal. *Higher Education Policy* **13**: 181–201

Court D 1979 The idea of social science in East Africa: An aspect of the development of higher education. *Minerva* **17**(2): 244–82

Dill D B, Sporn B 1995 *Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly*. Pergamon, Oxford

Gibbons M 1998 *Higher Education Relevance in the 21st Century*. The World Bank, Washington

Perkin H 2000 History of universities. *International Higher Education International* **13**: 169–204

Neave, G, van Vught F A 1994 *Government and Higher Education Relationships Across Three Continents: The Winds of Change*. Pergamon, Oxford

Rothblatt S 1995 An historical perspective on the university's role in social development. In: Dill D, Sporn B (eds.) *Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly*. Pergamon, Oxford

Scot P 1999 Globalization and the University. Unpublished manuscript

Stifel L D, Davidson R K, Coleman J S 1982 *Social Sciences and Public Policy in the Developing World*. Lexington Books, Lexington, MA

Task Force on Higher Education and Society 2000 *Higher Education in Developing Countries: Peril and Promise*. World Bank, Washington

D. Court

Policy Networks

Although the concept of policy network has taken on a variety of different definitions since its first appearance as an analytical term in the late 1970s, gradually a consensus has appeared on its meaning. A policy network refers to a set of informal and formal interactions between a variety of usually collective public (state) and private actors, who have different, but interdependent interests. Operating in a more or less institutionalized setting, these actors are engaged in horizontal, relatively nonhierarchical discussions and negotiations to define policy alternatives, or formulate policies, or implement them. This definition suggests that policy networks vary along two core dimensions: the structural pattern according to which public power is shared among the members and the degree and patterns of integration among the members. These two dimensions developed along distinct conceptual routes, only merging into a unified concept in the mid-1990s.

1. The Changing Context of Policy-making

This definition of policy networks signals a change in the way political decision-making has been conventionally understood. It contradicts the standard notion of a clear separation between state and society, with the state being seen as the highest center for social and political control. As Mayntz (1993) observes, social modernization has brought more importance to formal organizations in all areas of society and with their strengthened position comes an increasing fragmentation of power. Policy networks represent one response to the escalating functional differentiation of social subsystems. With the state no longer being the sole entity capable of organizing society, there is a dispersion of expertise and competence, a multiplication of channels for mediation and agreement, and the involvement of different levels of decision-making from the local to the supranational (Jouve 1995). Consequently, public policy is much less likely to be produced by a unitary state with adequate control over all the resources necessary for implementing decisions in the public interest.

Rather, public policy is more likely to result from strategic interaction among several or many policy actors, each with a particular understanding of the problem and its own individual or institutional self-interest (Scharpf 1997).

1.1 Historical Development

The sharing of political power between public and private actors became the first avenue of enquiry that led to the development of the policy network concept. In the mid-1970s, Schmitter and Lehmbruch (1977) suggested the concept of corporatism be reconsidered as a conceptual tool for understanding the particular relationships between public and private actors that had emerged for the management of macroeconomic policy in several European democracies. The revitalization of the corporatism concept coincided with

other suggested new concepts for capturing the changing patterns of power sharing between private and public actors: 'corporate pluralism' in Scandinavia, 'issue networks' in the US (Heclo 1978), 'policy communities' in the UK (Richardson and Jordan 1979), and 'policy networks' (Katzenstein 1978).

The early focus in each of these concepts was on broad, macro characteristics of the political system. As these concepts became more familiar, however, a number of scholars noted their potential relevance for understanding private–public relations in more restricted sectoral policy arenas. Drawing on a growing number of empirical studies, these scholars began to develop typologies of different structural patterns of public–private interaction at the sectoral or meso level (Atkinson and Coleman 1989, van Waarden 1992), prompting a certain spate of definitional disputes. By the mid-1990s, however, sectoral and state corporatism, pressure pluralism, clientelism, issue networks, and state-directed networks generally were accepted as core structural types.

In addition to patterns of power sharing between public and private actors, a second line of inquiry developed focusing on the degree of integration among the various actors involved in these horizontal policy-making settings. Rhodes (1990) and Rhodes and Marsh (1995) targeted the degree of integration as their key variable and outlined a typology of different types of 'policy networks' that ranged from fairly tightly integrated 'policy communities' to very loose and open 'issue networks.' The prevalence of large organizations in horizontal arrangements in policy-making also triggered the interest of political sociologists interested in patterns of communication between large organizations. Laumann and Knoke (1987) sought to identify network structures by examining how information was transmitted between policy actors, the nature of the resource transactions between actors, and the relative openness of boundaries in various policy domains. These studies employed quantitative social network analysis developed in sociology. By the mid-1990s, this quantitative approach had incorporated insights from other policy network writings (Pappi and Henning 1998).

1.2 Current Views on the Concept

When speaking of the degree of integration of the policy network, it is usual to employ a concept that refers to the aggregate of actors involved. The term 'policy community' is utilized most commonly for these purposes, but terms like 'policy domain' or 'policy subsystem' are also used. Research has established a number of factors that are related to integration:

(a) Openness of the policy community. The more open deliberations in the policy community tend to be, the less integrated it is likely to become.

(b) Size of the policy community. The larger the number of members of the policy community, the greater the obstacles to high integration.

(c) Frequency of interactions among community members and the degree to which these interactions become institutionalized as opposed to remaining *ad hoc* and fluid. Institutionalization normally enhances the degree of integration.

(d) Membership stability. The more changeable the membership of a policy community, the less integrated it will be.

(e) Density and structure of information exchange. The denser the ties for information exchange and the more direct relationships are, the more integrated the policy community.

Another, particularly important, factor affecting the level of integration of a policy community is the degree to which its members share basic beliefs, values, and norms. Some actors within a given policy community may come to share a set of deep normative core beliefs and from these develop basic strategies and policy positions (Sabatier and Jenkins-Smith 1993). These 'advocacy coalitions' create 'spaces of meaning' within the policy community that provide a framework for political action. The degree of integration of the policy community will be affected by how many distinct spaces of meaning it holds. Sabatier and Jenkins-Smith suggest that policy communities will normally contain two to four advocacy coalitions. They add that the number will likely be higher when the policy community is newly formed, and may dwindle to one dominant coalition as policy-making becomes routinized. The concept of an epistemic community developed in international relations has certain properties in common with Sabatier and Jenkins-Smith's advocacy coalition (see *Policy Knowledge: Epistemic Communities*). All or part of an epistemic community may act as an advocacy coalition. When this development occurs, these epistemic communities often provide important links between policy communities, whether in other nation–states or at the transnational level.

Different types of policy networks may be distinguished depending on whether state actors share political power and on the relative resource balance between private and public actors. State actors may share political power by inviting private actors to participate directly in the formulation or implementation of policy (see *Authority: Delegation*). Here private actors become policy participants. Corporatist and clientelist networks have this feature. Or, state actors may monopolize policy formulation and policy implementation, while listening to, or receiving submissions from, private actors. In this relationship, private actors perform as policy advocates. Pressure pluralism is the most common network with this feature, but it also occurs in issue networks and state-directed networks. The decision by state actors to share political power affects the overall balance of

resources between public and private actors in the policy network. Resources may be relatively balanced, they may favor the state actors, or they may advantage civil society actors. Both pressure pluralist and sectoral-corporatist networks feature a balanced relationship. Private actors have the stronger position in clientelist and issue networks while public actors dominate in the state-directed and state-corporatist types.

2. *Theoretical Status of the Concept*

Although definitional disputes surrounding the concept have gradually been resolved, the overall theoretical status of policy networks remains subject to debate. Two fairly fundamental criticisms have been directed at the concept. First, Dowding (1995) has charged that the concept is nothing more than a metaphor and will thus fail to produce theories of the policy process. Second, the concept is said to have limited relevance because of an apparent lack of suitability for explaining policy innovations and policy change.

2.1 *Policy Networks as Metaphors Only*

Dowding has argued that the attempts to provide a middle range theory of policy-making utilizing the policy network concept are bound to fail. He suggests that the driving force of the explanation, the independent variables, are not characteristics of the network as such, but of the components within the networks. Variables such as the relative openness of a policy network are cast as a mere redescription of policy phenomena. Related criticisms argue that policy network studies have failed to explain both the mutual dependence of public and private actors and the conditions under which particular policy networks arise (Thatcher 1998). Critics also lament the lack of specificity of the hypotheses that do exist, a failing bound to render any empirical testing rather unconvincing.

Although useful for spurring discussion and reflection, there are several reasons why these criticisms may be premature. First, the criticisms undervalue the descriptive utility of the policy network concept. One of the most difficult empirical problems faced in the analysis of public policies is determining specifically who the consequential actors might be in any policy domain. Although one might be able to begin to specify these actors by utilizing a good understanding of formal institutions and of the persons occupying key positions, such a specification is almost inevitably bound to be incomplete. Policy-making is replete with informal relationships and the operation of subtle, somewhat opaque norms that empower certain actors in unexpected ways. Accordingly, the methodologies developed for delimiting a policy network have a broader use in policy analysis. By identifying the central actors in policy-making, they assist the analysis of the processes by which policy ideas are translated into particular policy proposals through interchanges between an identified set of actors. This identification process becomes particularly helpful in policy areas featuring multilevel governance as Peterson (1995) has argued in referring to the European Union. This argument also has relevance to such patterns of governance as they emerge in policy areas where globalization has a decided impact.

Second, numerous empirical studies subsequent to the publication of Dowding's critique have provided convincing evidence of relationships between variations in policy networks and policy outcomes. Josselin's (1997) study of the negotiation of core directives for a single market in financial services in the EU evaluates the extent to which different network types facilitate or hinder the promotion of member-states' interests in EU-level negotiation settings. She shows that a relatively closed corporatist network at the national level is a decided asset in promoting a unified position at the EU level. Conversely, however, a more fluid issue network makes such unity more difficult, while providing additional degrees of freedom in the fluid decision-making arenas that exist at the supranational level. Similarly, Daugbjerg (1998) relates the relatively closed character of corporatist networks to patterns of political change. The more closed and cohesive the network, the more incremental the pattern of change and the more stable the utilization of policy instruments. Coleman's (1996) study of the impact of globalization on the regulation of financial services at the domestic level traces relationships between particular types of policy networks and the varying adaptation of countries to these globalizing pressures.

Finally, Dowding's critique is addressed by some scholars who are using more formal models. Pappi and Henning (1998) elaborate a complex model that suggests direct relationships between network types (pluralism, corporatism, clientelism, state-direction) and policy outcomes. Scharpf's 'actor-centered institutionalism' introduces network-based negotiation settings as one of several core types of forums for policy-making. Drawing on game theory, he sets out a series of propositions that provide a basis for modeling negotiations between actors in such networks and deriving likely differences in policy outputs based on general welfare and distributive justice criteria. Both of these theoretical approaches would appear to hold considerable promise for the development of middle level theory where policy networks will be an important independent variable.

2.2 *Policy Networks and Policy Change*

Policy networks have been conceptualized primarily using structural variables. Such structural character-

istics are helpful for building descriptions of the policy process at a particular point in time and, more recently, for positing testable explanations of variations in policy outcomes. Dowding (1995) questions, however, whether these same variables are sufficient for explaining how and why policies change and why networks themselves change. Two avenues of investigation have been pursued for addressing these criticisms. The first focuses on the relationship between policy ideas and policy networks and the second draws on an analysis of boundary rules.

Sabatier and Jenkins-Smith (1993) argue that policy networks include one or more advocacy coalitions that are composed of actors who share deep core beliefs, near core policy beliefs, and secondary beliefs about the use of policy instruments. Changes in the relative dominance of advocacy coalitions in a policy network may prompt policy change. Or members of coalitions may negotiate adjustment of secondary beliefs that, in turn, open the door to policy change. Paralleling Sabatier and Jenkins-Smith's work, Jobert (1995) argues that a 'political forum' and a 'scientific or expert' forum exist alongside a policy network. When politicians begin to experience problems that cannot be addressed well within a policy network, they begin to reach to these other fora for advice. To the extent that this advice is acted upon, the policy network may be dissolved, with new arrangements taking its place.

What both these theoretical arguments stress is the importance of looking at the interaction between policy ideas and who shares which ideas within a given policy network. Coleman et al. (1996) follow up on these insights by arguing that the degree of sharing of policy ideas in a given policy network will determine the approach to policy negotiations among actors in the network. They hypothesize that the more ideas are shared, the more policy discussions will take on aspects of what Scharpf (1997) has called 'positive coordination' or a 'problem solving' decision-making style. They suggest that positive coordination is more likely in closed networks like corporatism than in more open networks such as pressure pluralism. By coupling policy network structures to policy ideas, they then trace out different patterns for how policy change occurs. Scharpf's work stresses the importance of looking at actor preferences and cognitive orientations in studying policy processes. The richness of his theoretical work offers further opportunities for improving theories of policy change.

The second avenue for improving explanations of policy change focuses on different types of boundaries within policy communities. In their retrospective on policy network analysis in the UK, Rhodes and Marsh (1995) note that a policy community can have several policy networks. These, in turn, may operate at two levels; a central level whose members dispose of considerable resources, and a peripheral one composed of those with few or no resources. Changes in the actors at the center of a policy community and thus in the dominant policy network, or movements from the peripheral networks to the dominant one, will be crucial intervening variables in explanations of policy change. Similarly, new actors may build peripheral networks over time and then come to contest the exclusivity of the existing dominant policy network. To the extent that new actors enter an existing policy network, they may bring about a change in its structure. For example, entry of new actors into a previously closed corporatist network may foster sufficient dissent that the state refuses to involve private actors any longer in policy making. Once such actors are excluded, the network takes on pressure pluralist characteristics, altering, in turn, the policy process and perhaps policy outcomes (Coleman et al. 1996).

3. *Methodological Issues*

The empirical analysis of policy networks has centered on two variables: the degree of integration of policy communities and the division of power between state and civil society actors. Scholars have approached these empirical issues from a variety of methodological standpoints. Among these, the use of quantitative network analysis, elite interviews, and close examination of documents and archival materials are the most commonly used. The use of these methods varies depending on the particular empirical task in question.

3.1 What are the Boundary Rules of the Policy Community? Who Belongs?

Determining the boundaries of a policy network is one of the more difficult, yet more important empirical tasks in this approach. Understanding who belongs and who does not provides key information about likely political demands and patterns of political influence, factors, in turn, important to explaining policy outcomes. In addition, as noted above, observations of changes in policy network membership may be crucial to approaching the explanation of political change. Accordingly, boundary questions must be approached diachronically, with Sabatier and Jenkins-Smith (1993) suggesting a time perspective 'of a decade or more.'

Those working in the public policy field have tended to approach boundary specification by looking at relationships and then at activities and events. In the relational approach to boundary specification, they seek to identify the actors involved in the formulation and implementation of policies in selected policy domains. With these relationships in focus, they then examine variations in the organizational attributes, political ideas, and degree of participation of these actors in selected activities and events. A mix of

quantitative network analysis and qualitative techniques are used in these studies. The relationships of concern are usually determined using 'snowball-sampling' techniques to set up the right set of elite interviews. This methodology does have its weaknesses in that reliance on elite interviews is costly both in time and in money and it is difficult to replicate these interviews over time to get a dynamic perspective on boundary change.

3.2 *Does the State Share Power Directly With Policy Community Members? What is the Balance in the Distribution of Resources in the Policy Community?*

The answers to these questions provide the empirical base for determining the type of policy network in place in a given policy community. In beginning, to assess the nature of the policy network, it is essential to understand well the organizational attributes of the actors in the policy community. One of the crucial empirical legacies of the corporatist literature were some empirical guidelines for assessing such policy capacity.

The policy capacity of interest associations and of associational systems may be evaluated by looking successively at three types of phenomena. First, it is useful to assess whether associations in a policy network compete with one another for members and are part of a more comprehensive associative order. The development of an associative order is evaluated by examining the degree of horizontal and vertical integration among associations. Second, some evaluation of the internal resources of the associations is useful for assessing likely policy capacity. Frequently, however, these kinds of analysis are still insufficient, particularly for determining the degree to which associations participate in policy-making. Normally, before final determination of the structure of a policy network, it is useful again to turn to elite interviews for additional information on the extent to which civil society actors participate directly in policy formulation or implementation.

4. *Future Challenges*

The concept of a policy network was designed in the first instance for the analysis of policy making at the national and more local levels. As globalization has increased the internationalization of many policy areas, however, the likelihood rises that several policy networks will be interacting with one another across different levels of decision-making. Future research, therefore, will need to focus on improving understanding of the relationships between policy networks. Some of these networks will have the additional property of being transnational in character, that is, they will include actors from both the nation-state and the international levels.

The concept of policy community mediators offers one avenue for research on the relationships between policy networks (Coleman and Perl 1999). In many internationalized policy domains, boundary rules will create overlapping memberships between policy networks, with mediators emerging from among those actors with joint memberships. Mediation can take place at two levels. First, a mediator may operate as a pragmatic policy broker who looks for opportunities to lessen conflict within and between policy networks in order to move the policy process forward. Second, perhaps even more important, mediation takes place on an intellectual plane, involving the 'translating' of policy ideas and policy paradigms between policy networks. Drawing on Kuhn's work on scientific paradigms, Schön and Rein (1994) suggest that translation is crucial to resolving conflicts between different policy paradigms or frames. Mediators construct a new view of the world that has a place for the policy concerns of both networks and that provides a language and set of norms that enable them to speak to one another.

Success in this second role requires that the mediator be accepted as legitimate in both policy communities. Legitimacy will derive from having both adequate expertise and a socialization experience that permits straddling the boundaries of the communities concerned (Jouve 1995). Research on linkages between policy networks and on the role of policy mediators will be crucial to further theorizing on policy networks in this globalizing era.

See also: Network Analysis; Networks: Social

Bibliography

Atkinson M M, Coleman W D 1989 Strong states and weak states: sectoral policy networks in advanced capitalist economies. *British Journal of Political Science* **19**: 47–67

Coleman W D 1996 *Financial Services, Globalization and Domestic Policy Change*. St. Martin's Press, New York

Coleman W D, Perl A 1999 Internationalized policy environments and policy network analysis. *Political Studies–London* **47**: 691–709

Coleman W D, Skogstad G, Atkinson M 1996 Paradigm shifts and policy networks: cumulative change in agriculture. *Journal of Public Policy* **16**: 273–301

Daugbjerg C 1998 Linking policy networks and environmental policies: nitrate policy making in Denmark and Sweden 1970–1995. *Public Administration* **76**: 275–94

Dowding K 1995 Model or metaphor? A critical review of the policy network approach. *Political Studies–London* **43**: 136–58

Heclo H 1978 Issue networks and the executive establishment. In: Beer S H, King A (eds.) *The New American Political*

System. American Enterprise Institute for Public Policy Research, Washington, DC
Jobert B 1995 Rhétorique politique, controverses scientifiques et construction des normes institutionnelles: esquisse d'un parcours de recherche. In: Faure A, Pollet G, Warin P (eds.) *La Construction du Sens dans Les Politiques Publiques: Débats Autour de la Notion de Référentiel* [The Construction of Meaning in Public Policies: Debates surrounding the concept of the *Référentiel*]. L'Harmattan, Paris
Josselin D 1997 *Money Politics in the New Europe: Britain, France and the Single Financial Market*. Macmillan, Basingstoke, UK
Jouve B 1995 Réseaux et communautés de politique publique en action. In: Le Galès P, Thatcher M (eds.) *Les Réseaux de Politique Publique. Débat autour des Policy Networks* [Public Policy Networks: Debates on Policy Networks]. L'Harmattan, Paris
Katzenstein P 1978 Conclusion: domestic structures and strategies of foreign economic policy. In: Katzenstein P J (ed.) *Between Power and Plenty: Foreign Economic Policies of Advanced Industrial States*. University of Wisconsin Press, Madison, WI
Laumann E O, Knoke D 1987 *The Organizational State*. University of Wisconsin Press, Madison, WI
Mayntz, R 1993 Policy-netzwerke und die Logik von Verhandlungssystemen [Policy networks and the logic of negotiation systems]. In: Héritier A (ed.) *Policy-Analyse: Politische Vierteljahresschrift Sonderheft 24*. Westdeutscher Verlag, Opladen, Germany
Pappi F U, Henning C H 1998 Policy networks: more than a metaphor? *Journal of Theoretical Politics* **10**: 553–75
Peterson J 1995 Decision-making in the European Union: towards a framework for analysis. *Journal of European Public Policy* **2**: 69–93
Rhodes, R 1990 Policy networks: a British perspective. *Journal of Theoretical Politics* **2**: 293–317
Rhodes R, Marsh D 1995 Les réseaux d'action publique en Grande Bretagne. In: Le Galès P, Thatcher M (eds.) *Les Réseaux de Politique Publique. Débat autour des Policy Networks* [Public Policy Networks: Debates on Policy Networks]. L'Harmattan, Paris
Richardson J J, Jordan A G 1979 *Governing Under Pressure: The Policy Process in a Post-Parliamentary Democracy*. Robertson, Oxford, UK
Sabatier P A, Jenkins-Smith H C 1993 Policy change over a decade or more. In: Sabatier P A, Jenkins-Smith H C (eds.) *Policy Change and Learning: An Advocacy Coalition Approach*. Westview Press, Boulder, CO
Scharpf F W 1997 *Games Real Actors Play: Actor-centered Institutionalism in Policy Research*. Westview Press, Boulder, CO
Schmitter P C, Lehmbruch G 1977 *Trends Toward Corporatist Intermediation*. Sage Publications, London
Schön D A, Rein M 1994 *Frame Reflection: Toward the Resolution of Intractable Policy Controversies*. Basic Books, New York
Thatcher M 1998 The development of policy network analyses: from modest origins to overarching frameworks. *Journal of Theoretical Politics* **10**: 389–416
van Waarden F 1992 Dimensions and types of policy networks. *European Journal of Political Research* **21**: 29–52

W. D. Coleman

Policy Process: Business Participation

While business firms play an important role in policy process in all capitalist societies the extent, nature, and impact of this participation varies both over time and among countries. Business political participation has changed significantly in the USA since the 1970s, and to a lesser extent in Europe. However the degree of political power exercised by business remains an important point of contention among scholars.

1. Business Political Participation in Comparative Perspective

A critical factor underlying differences in national patterns of business–government relations is the extent and scope of government intervention in the economy (Shonfield 1965). The greater a government's impact on the direction of national economic development—measured by spending, government ownership of industry, and control over capital allocation—the closer are the political ties between business and government. In capitalist nations characterized by extensive government intervention, the government will often promote the formation of business associations in order to facilitate policy coordination both with and within the private sector. The way business is organized politically also tends to mirror the structure of the public sector: the more governmental decision-making is centralized, the greater the breadth of business political associations.

On both dimensions, the USA is distinctive: government economic intervention has historically been relatively limited and public policy making is relatively fragmented. Accordingly, there is no American counterpart to 'peak' associations such as the German Federation of Industry, Confindustria in Italy, the Comité National du Patronat in France, the Keidanren in Japan, or the Confederation of British Industry—organizations which both represent and are able to speak for business as a whole (G. Wilson 1981, 1985).

In many capitalist nations, trade associations negotiate with the government on behalf of all firms in a particular sector. Such arrangements are often characterized as corporatist (Schmitter and Lehmbruch 1979). By contrast, trade associations in the USA rarely enjoy official or quasiofficial status and they tend to be much weaker and less inclusive than their counterparts in other capitalist nations. More generally, while business firms have become much more politically active in the USA since the 1970s, business is still less politically organized than in other capitalist nations.

The nature of the interaction between business organizations and government officials also differs in

the USA. Firms in the USA typically influence public policy through 'lobbying,' a term which connotes an arms-length relationship between business and government. Corporations typically need to engage in pressure group politics to gain access to policy makers and shape the policy process. But outside the USA, business and political elites are more likely to come from similar social backgrounds and there is considerable movement of personnel between the two institutions. There are numerous opportunities for both formal and informal interaction between business and government officials and firms can expect to be closely consulted about policies which affect their interests. There is less need to 'lobby' because business organizations are themselves either a quasiofficial or official part of the public policy process.

2. *Changes in Business Political Participation in Europe*

However, business political participation has changed in Europe. In many capitalist societies a growing portion of firms are subsidiaries of multinational corporations whose headquarters are outside national borders. Such firms are less likely both to identify their interests with those of the nation in which their business is located and to lack the kind of informal access to policy makers enjoyed by domestic firms.

In addition, the role of government in managing the economy has declined in almost all capitalist polities: the importance of national industrial policies which required close business–government cooperation has diminished while a substantial number of government-owned firms have been privatized, thus creating a more independent and pluralist business community. Both of these developments have diminished the historically close ties of business and government in a number of European countries and made the structure of business political representation more fragmented.

Within Western Europe, the increasing importance of the European Union has had a significant impact on business–government relations (Mazey and Richardson 1993, Wallace and Young 1997). The Single European Act (1987) not only significantly expanded the regulatory role of the EU, but ended the unanimity rule of the Council of Ministers. The latter constitutional change meant that each member state was no longer able to exercise a veto power over EU legislation. Accordingly, firms and trade associations could no longer rely on their influence with their national governments to protect their interests in Brussels.

The result has been a substantial increase in business political representation in Brussels. Many European-wide trade associations have been established to represent business interests in Brussels. At the same time, many large multinational firms have established their own independent lobbying operations in Brussels. The interests of 80 American firms with subsidiaries in Europe are represented by the EC Committee of the American Chamber of Commerce. In addition, many national trade associations have shifted part of their political efforts to Brussels.

In some cases, the more stable corporatist arrangements which characterized business–government at the national level have been undermined and firms and industries no longer enjoy the privileged access they had when policy making took place at the national level. The result has been an EU policy process which often resembles the American pattern of 'disjointed pluralism' and fluid issue networks (Streeck and Schmitter 1991). This, however, is not true of all policy areas, many of which continue to be characterized by stable policy communities which usually privilege producer interests.

2.1 *Contemporary Changes in Business Political Participation in the USA*

The degree and forms of business participation have significantly changed in the USA since the 1970s. During the 1950s and 1960s, business devoted relatively few resources to influencing public policy in Washington. Few large firms had Washington offices and these were typically thinly staffed. An exhaustive study of business lobbying on foreign trade policies during the 1950s found most business lobbies to be poorly financed, ill-managed, and at best only marginally effective. Indeed, only a minority of either large- or medium-sized firms regularly communicated with Congress (Bauer et al. 1993).

Certainly, many firms and industries received considerable benefits from government. For example, defense contractors benefited from procurement policies, agricultural interests received substantial price support, and the oil and textile industries, and also much of American agriculture, were protected from import competition. Many industries, such as housing and energy, received important tax preferences and a substantial portion of the American economy was regulated in ways that limited price competition and stabilized markets. However, because business enjoyed considerable public support, and nonbusiness constituencies were poorly organized, business did not need to engage in extensive political activity to enjoy a wide array of economic privileges.

2.2 *The Changing Political Environment of Business*

The political environment of business began to change in the late 1960s. The social activism of the latter part of that decade reflected and contributed to a decline of

public confidence in all institutions, including the large corporation. It also spawned a new political force: a loosely allied network of consumer and environmental organizations who began to pressure for expanding government regulation of business. Led by skillful political entrepreneurs such as Ralph Nader, and enjoying extensive and for the most part favorable press coverage, the public interest movement had a major policy impact (Berry 1997). From the late-1960s through the mid-1970s business experienced a series of major political defeats in Congress. New laws were passed which expanded the regulatory role of the federal government in environmental protection, consumer protection, and occupational health and safety.

Unlike previous patterns of government intervention, which had primarily affected only specific industries, the impact of the new social regulation was much broader: now, for the first time, virtually every firm faced a large and complex array of federal rules that mandated or prohibited a wide range of business decisions. The additional costs of complying with these regulations also became more burdensome following the economic downturn of the mid-1970s, which also led to a substantial increase in public hostility toward particular industries, most notably the oil industry.

2.3 *Implications for Business Political Strategy*

Business–government relations had previously been dominated by 'insider politics': corporations had secured benefits primarily by making low-visibility appeals to key legislators and administrators. But social regulation was dominated by 'outsider politics': it was often driven by highly visible, often emotional appeals to both public and elite opinion . To counter effectively the appeals of activists and their supporters, companies had to adopt to the 'outsider' strategies of their political opponents. Moreover, within Congress power became more decentralized and party discipline weakened. This meant that business could no longer rely on a handful of influential politicians or party leaders: they had to learn how to influence the votes of 535 individual legislators.

This decade also witnessed a substantial change in the regulations governing campaign contributions by business. Since the beginning of the twentieth century, corporations had been forbidden to contribute funds to candidates for federal office. However, the investigations surrounding the Watergate break-in revealed that a large number of firms had responded to pressures from the Finance Committee to re-elect the President by making illegal contributions to President Nixon's re-election campaign. Ironically, Congress 'cleaned up' federal campaign finance by providing a legal vehicle for corporations to solicit funds from their executive employees and shareholders. The 1974 Federal Campaign Act Finance Amendments permitted corporations, and also other organizations, to establish political action committees (PACs). Corporations were allowed to pay the costs of establishing and administrating PACs, although they still could not contribute to them

2.4 *New Forms of Business Political Participation*

As a response to these developments, business political activity both increased and became more visible (G. Wilson 1981). Between 1968 and 1978, the number of corporations with public affairs offices in Washington grew from 100 to more than 500, while many other firms employed full- or part-time lobbyists. All told, by the early 1980s nearly 2,500 business firms had some form of political representation in Washington (Vogel 1989, J. Q. Wilson 1996).

The number of trade associations also grew substantially, and many moved their headquarters to Washington in order to more effectively monitor and influence national policy. Some, such as the Chemical Manufacturers Association, the Semiconductor Industry Association, and the Health Insurance Association of America, proved highly effective. But the typical industry trade association declined in importance. This was largely the result of structural changes in the American economy which made firms more diversified. With investments in a wide variety of different industries, their interests could no longer be effectively represented by an industry-based interest group. Accordingly, many large corporations developed an independent capacity to monitor political developments and to advance their political interests (Ryan et al. 1987, Harris 1989).

In order to advance its collective interests, the business community engaged in coalition building—the identification and mobilization of large numbers of companies and trade associations who shared a common interest in a particular issue. This became a highly effective strategy to maximize the political influence of substantial segments of the business community, since it enabled many firms and trade associations to pool and coordinate their political resources. Such coalitions were often difficult to establish since many firms were reluctant to commit scarce political resources to lobby on issues in which their stake was either modest or not readily apparent. But those that were establishing were often politically effective.

Many coalitions were *ad hoc* in nature: they were formed to lobby on a particular issue—and then disbanded when the issue was resolved or disappeared from the political agenda. However, some business coalitions became institutionalized (Levitan and Cooper 1984). One of the most important new 'peak' business organizations was the Business Roundtable.

It was established during the early 1970s from a merger of three *ad hoc* coalitions which had been primarily formed to lobby on issues affecting labor. Exclusively composed of the chief executive officers of approximately 200 of the nation's largest corporations, it relied on the 'unique mobilization of the talent and prestige of the nation's top corporate heads' to lobby government officials (Vogel 1989, p. 198).

Other more established business organizations representing large numbers of firms and industries also grew in importance. Of particular significance was an increase in coordinated political activity on the part of smaller firms. Between 1970 and 1979, the membership of the National Federation of Independent Business (NFIB), which represented smaller firms, grew from 300 to 600,000. The NFIB frequently took positions opposed to those of the Business Roundtable, which many small firms viewed as too willing to accommodate itself to liberal political pressures for government expansion.

Another dimension of business political participation focused on mobilizing the 'grass-roots.' By the late 1970s, corporations and trade associations were spending an estimated $850 million to help their supporters throughout the USA to communicate with, and thus influence the votes of, their representatives and Senators. Much of what Washington lobbyists now did was to identify strategies for influencing particular legislators, rather than lobby them directly. Business associations representing large numbers of smaller firms such as the National Association of Independent Businessmen and the Chamber of Commerce made particularly extensive use of grass-roots lobbying.

2.5 *The Politics of Ideas*

An important component of business participation in the policy process focused on changing public attitudes toward business in general and specific policy issues in particular (Blumenthal 1986). Corporations and trade associations made considerable use of 'advocacy advertising,' a specialized form of institutional or corporate advertising whose purpose is to influence the public's attitudes on issues of public policy. During the 1970s, more than 35 corporations and trade associations launched public advocacy campaigns and by the end of that decade, major corporate advertisers were spending about one-third of their advertising dollars on campaigns aimed at the public in their role as citizens rather than consumers.

The oil industry made extensive use of this advertising strategy during the 1970s to explain its position on antitrust policy, government regulation and industry profitability—all issues then much in the public spotlight as a result of the energy crisis. Steel firms purchased advertising space to explain their positions on import restrictions and pollution controls, the American Electric Power Company ran a multimillion dollar campaign opposing the use of scrubbers and the Caloric Control Council employed advocacy advertising to build a successful grass-roots campaign to enact legislation overruling the Food and Drug Administration's ban on saccharin. During the 1990s, the Health Insurance Association of America waged a highly effective television camping against the Clinton Administration's proposed health care plan.

The business community and its supporters also sought to influence the terms of public debate by funding research on public policy. Financial support from corporations and conservative foundations played a critical role in creating and expanding private research centers that issued and sponsored books, studies and reports that challenged liberal and left of center views on public policy. In contrast to advocacy advertising, these efforts were aimed at influencing the views of elites such as politicians, government regulators, academics, and journalists.

The Heritage Foundation, a conservative policy research institute, employed several dozen policy analysts who worked closely with the staffs of sympathetic members of Congress, tracking the progress of particular policy proposals and preparing detailed positions papers on them. The American Enterprise Institute, which became a kind of right-of-center counterpart to the more liberal and older Brookings Institution, released scores of studies, held numerous conferences, and prepared several analyses of legislative proposals—in addition to publishing seven journals and newsletters—primarily focusing on government regulation and tax policy. This research helped contribute to a more skeptical attitude about the role of government and a more positive attitude about the role of markets—first among elites and then among the public as a whole. More liberal research institutes also became more sympathetic to business perspectives.

3. *The Debate Over Business Power*

During the 1950s and 1960s, the scholarly consensus within both political science and sociology regarded business as an interest group which did not pose any distinctive problems for democratic governance (Truman 1951, Epstein 1969). There were dissenters, but they were a minority voice (Bachrach 1967). During the 1970s and 1980s the intellectual center of gravity of both disciplines began to shift: a considerable body of scholarship now judged 'the large corporation's concentration of resources and wealth (to be) an anomaly that upsets the balance between democracy and capitalism' (Menninger 1985, p. 210).

According to the critics of pluralism, business political influence has two dimensions, one structural

and the other instrumental. Structural power refers to the ability of business to shape both the terms of political debate and political outcomes by virtue of its privileged position within a capitalist system. This privileged position in turn has two roots. One is the central role of business in shaping the rate and direction of economic development. According to this analysis, if public policy does not furnish firms with sufficient inducements to invest, or alternatively if it imposes too many restrictions on their ability to accumulate capital, firms will reduce their rate of investment, thus resulting in an economic downturn. Consequently, politicians, knowing that their re-election will be adversely affected by a slowdown in economic growth, will be reluctant to support policies that might antagonize corporate investors.

Since a critical dimension of political power involves control over the political agenda and the ability to define the terms of public debate, the privileged political position of business in capitalist societies also has an ideological dimension. According to this analysis, a capitalist society is characterized by a prevailing ideological consensus that places a wide array of issues and policy alternatives off the political agenda, such as the autonomy of corporate managers, the prevailing distribution of wealth, and private ownership of the means of production—all of which protect business privileges.

While the structural dimension of business power is often exercised in subtle ways, the second, instrumental dimension of business power is more visible. It is reflected in the disproportionate resources that business has available to shape the resolution of issues that actually do come before the political process—resources which give business political advantages enjoyed by no other interest group, or the public as a whole. This dimension of business power became substantially more visible in the 1970s and 1980s as firms devoted increased resources in order to both shift the terms of public debate and affect political outcomes.

3.1 Is Business Privileged?

The claim that business enjoys a privileged political position is closely associated with Charles Lindblom, whose book *Politics and Markets* represents one of the most important and influential analyses of business political power published since the 1970s (Lindblom 1997). As he argues, capitalism in the USA enjoys considerable legitimacy: no political movement in the USA has seriously questioned private ownership or challenged the central role of markets in distributing wealth and privilege. However, while business certainly benefits from this characteristic of American society, it is not responsible for it. The hegemony of the 'liberal tradition in America' is deeply rooted in popular American culture—a culture that has long valued individual self-reliance, limited government, and private property—and which predates the rise of the large corporation (Hartz 1955, McClosky and Zaller 1984).

At the same time, American culture also contains a strong populist dimension, which, although not anti-capitalist, has periodically been expressed in considerable public hostility to the large corporation. During the twentieth century, the USA experienced four reform movements—Populism, Progressivism, The New Deal, and the public interest movement, each of which has effectively challenged a wide variety of corporate prerogatives. Moreover, in other capitalist nations, anticapitalist movements and political parties on both the left and the right have received considerable public support, which suggests that capitalism does not enjoy equal legitimacy or hegemony in all capitalist societies.

Scholars have challenged Lindblom's thesis with reference to the USA by citing the ability of non-business constituencies to challenge effectively a wide array of corporate prerogatives during the 1960s and 1970s (Vogel 1987, G. Wilson 1981, J. Q. Wilson 1996). The effectiveness of the challenges to corporate privileges that have surfaced periodically in the USA implies that corporations, like any group of interests, must engage in interest-group politics in order to defend and advance their interests. Sometimes their efforts will be successful and other times they will not. But this does not distinguish them from any other group of interests.

Vogel claims that the power of business has varied considerably over time: it was relatively substantial in the 1950s through the late 1960s, declined through the late 1970s, and then subsequently experienced a resurgence (Vogel 1989). These fluctuations in business political influence are primarily attributed to shifting public perceptions of the performance of the economy: business tends to be more influential when the economy appears to be performing relatively poorly and is less influential when it is relatively prosperous. While this analysis does help account for many of the changes in political influence from the early 1960s through the late 1980s, it cannot account for the 1990s—when economic performance was extremely strong and yet business was able to successfully resist the efforts of the Clinton Administration to increase public social spending and expand medical coverage.

Plotke (1992) argues that the slowdown in growth and the decline in profits during the 1970s was not by itself sufficient to mobilize business politically. What was also required was an alternative vision of political economic relations which both attributed the nation's economic difficulties to the expansion of government intervention and proposed an alternative model of economic growth and development. This ability to redefine the political agenda provided a referent for business interest formation that in turn made possible the mobilization of large numbers of disparate firms

and industries. An important contribution of this perspective is to emphasize the contingent nature of business political preferences, as well as the extent to which business influence both affects and is affected by the terms of public debate.

Mucciaroni (1995) analyzed the factors that affect the ability of particular sectors or firms to maintain the benefits they receive from government. Focusing on developments in four policy areas, namely trade policy, agricultural subsidies, tax expenditures, and economic regulation, over two decades, he concluded that producer-group fortunes have varied considerably both over time and among issue areas. In three of the four policy areas, producers experienced reduced benefits due to changes in both public policy norms and the structure of political decision making.

3.2 Corporate Campaign Contributions

The number of business-sponsored political action committees grew substantially in the 1970s and 1980s. While there were no corporate PACs in 1971, there were more than 1,700 in 1992 and they gave a total of $64 million to congressional candidates. Thanks to the wealth of data available due to federal reporting requirements, few dimensions of business political participation—and power—have been studied as extensively.

One focus of this research has examined the effect of campaign contributions on the voting of legislators, from which different conclusions have been drawn. A series of studies have found that business contributions affected roll-call votes on minimum wage legislation, the debt limit, the windfall profits tax, wage and price controls, trucking deregulation, and legislation affecting auto dealers (Conway 1991). However, a comprehensive study of a number of PACs and congressional voting behavior over an 8-year period found contributions to be rarely related to congressional voting patterns.

According to Wilson, 'a fair summary of the existing studies (reveals) there is no consistent and systematic evidence showing that PAC contributions, independent of other factors (such as constituency pressures, legislator ideology, and party affiliation), have a significant impact on voting in Congress. Some studies find some effects in some areas, other find no effect in other areas' (J. Q. Wilson 1996, p. 430). These findings suggest that corporations often make contributions as much to gain access as to influence votes. However, other studies reach a very different conclusion, arguing that campaign contributions do represent a critical source of business power and privilege (Clawson et al. 1992, Ferguson 1995).

The extent to which firms share similar policy preferences, and are able to cooperate to achieve them, has been an important focus on scholarly debate over the political significance of business campaign spending as well as over business political influence in general. Business campaign contributions are often coordinated. For example, in the 1984 election, business contributions went overwhelming to only one candidate in almost every race, while in three out of four congressional races, large corporate PACs have given at least nine times as much to one candidate as the other (Clawson et al. 1992, p. 160).

3.3 The Extent and Significance of Business Political Unity

This coherence, in turn, reflects not only their common interests, but also a common culture and the extent of structural ties among firms, such as loans from the same banks, sales and purchases from each other, and interlocking boards of directors. 'Business power derives in part from this ability to achieve political unity on most issues' (Clawson et al. 1992, p. 160). Another study found economic leverage between industries to be associated with similar patterns of donations (Mizuichi and Koenig 1986). Moreover, the American business community has numerous mechanisms that facilitate cooperation and coordination, especially under the leadership of large firms (Useem 1984).

Other scholars, by contrast, emphasize that while 'corporate class consciousness' does exist in the USA, it only emerges sporadically under unusual circumstances: the American business 'community' is more often a community in name only (Vogel 1978). The structure of the American economy is highly fragmented with firms having multiple sources of capital. Moreover, the American system of corporate governance encourages managers to support public policies that directly benefit their firms—even if such policies may not be in the interest of business as a whole. While companies may not typically give to opposing candidates, they nonetheless frequently do have opposing interests.

Many important public policies, such as trade policy, tax policy, health policy, environmental policy, and economic deregulation, divide the business community. A particularly important and long-standing division is between large firms and small businesses, with the latter often frequently opposing the more liberal policy preferences of 'big business.' Martin's study of the politics of human capital investment, including the bitter debate over the expansion of medical coverage in the early 1990s, reveals sharp divisions between large and small firms which precluded the enactment of reforms that arguably would have benefited business as a whole (Martin 2000).

Ironically, the politicization of business that took place during the 1970s has exacerbated the centrifugal tendencies within American business. Because firms have become more politically sophisticated and both able and willing to devote substantial resources to

influencing public policy, corporate political strategy has become more closely linked to business strategy. Firms now frequently seek to gain competitive advantages through the use of non-market strategies, which in turn increases the likelihood of policy differences among them (Baron 1995).

4. Conclusion

The latter part of the twentieth century has witnessed a substantial increase in both scholarly and popular studies of business political activity in the USA and the European Union, in large measure because this activity has become more visible. There have also been a few comparative studies of business political participation (G. Wilson 1985, 1990, Bottomore and Brym 1989). However, there remains a dearth of comparative studies of business political power and influence. Scholars who assert that business enjoys a privileged position usually assume that business is equally influential in all capitalist societies, while most empirical studies of business political influence tend to focus on only one country, or even more narrowly on specific policy areas. An important challenge of future scholarship is to compare systematically patterns of business political influence in different capitalist societies.

See also: Business History; Business Law; Community Power Structure; Corporate Law; Ethical Codes, Professional: Business Codes; Lobbying; Participation: Political; Policy History: State and Economy; Political Money and Party Finance; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulation Theory in Geography

Bibliography

Bachrach P 1967 *The Theory of Democratic Elitism: A Critique*. Little, Brown, Boston

Baron D P 1995 Integrating strategy—market and nonmarket components. *California Management Review* **37**(2): 47–65

Bauer R A, de Sola Pool I, Dexter L A 1993 *American Business and Public Policy: The Politics of Foreign Trade*. Atherton Press, New York

Berry J 1997 *Lobbying for the People: The Political Behaviour of Public Interest Groups*. Princeton University Press, Princeton, NJ

Blumenthal S 1986 *The Rise of the Counter-establishment: From Conservative Ideology to Political Power*. Times Books, New York

Bottomore T, Bryn R J (eds.) 1989 *The Capitalist Class: An International Study*. New York University Press, New York

Clawson D, Neustadtl A, Scott D 1992 *Money Talks: Corporate PACS and Political Influence*. Basic Books, New York

Conway M 1991 PACs in the political process. In: Cigler A J, Loomis B A (eds.) *Interest Group Politics*, 3rd edn. Congressional Quarterly Press, Washington, DC, pp. 199–216

Epstein E M 1969 *The Corporation in American Politics*. Prentice-Hall, Englewood Cliffs, NJ

Ferguson T 1995 *Golden Rule: The Investment Theory of Party Competition and the Logic of Money-driven Political Systems*. University of Chicago Press, Chicago

Harris R A 1989 Politicized management: the changing face of business in American politics. In: Harris R A, Milkis S M (eds.) *Remaking American Politics*. Westview Press, Boulder, CO, pp. 261–86

Hartz L 1955 *The Liberal Tradition in America*. Harcourt Brace, New York

Levitan S A, Cooper M R 1984 *Business Lobbies: The Public Good & the Bottom Line*. Johns Hopkins University Press, Baltimore

Lindblom C 1997 *Politics and Markets*. Basic Books, New York

Martin C J 2000 *Stuck in Neutral: Business and the Politics of Human Capital Investment Policy*. Princeton University Press, Princeton, NJ

Mazey S, Richardson J 1993 *Lobbying in the European Community*. Oxford University Press, Oxford

McClosky H, Zaller J 1984 *The American Ethos: Public Attitudes Toward Capitalism and Democracy*. Harvard University Press, Cambridge, MA

Menninger D 1985 Business and Politics. *Politics & Society* **xviii**, pp. 207–11

Mizruchi M S, Keonig T 1986 Economy sources of corporate political consensus. *American Sociological Review* **51**: 482–91

Mucciaroni G 1995 *Reversals of Fortune: Public Policy and Private Interests*. Brookings Institution, Washington, DC

Plotke D 1992 The political mobilization of business. In: Petracca M R (ed.) *The Politics of Interests: Interest Groups Transformed*. Westview Press, Boulder, CO, pp. 175–98

Ryan M H, Swanson C L, Buchholz R A 1987 *Corporate Strategy, Public Policy and the Fortune 500: How American's Major Corporations Influence Government*. Blackwell, Oxford, UK

Schmitter P C, Lehmbruch G 1979 *Trends Toward Corporatist Intermediation*. Sage, London

Shonfield A 1965 *Modern Capitalism*. Oxford University Press, New York

Streeck W, Schmitter P C 1991 From national corporatism to transnational pluralism: Organized interests in the single European market. *Politics and Society* **19**: 133–64

Truman D B 1951 *The Governmental Process: Political Interests and Public Opinion*. Alfred Knopf, New York

Useem M 1984 *The Inner Circle: Large Corporations and the Rise of Business Political Activity in the U.S. and U.K.* Oxford University Press, Oxford, UK

Vogel D 1978 Why businessmen distrust their state: The political consciousness of American corporate executives. *British Journal of Political Science* **8**(1): 45–78

Vogel D 1987 Political science and the study of corporate power: A dissent from the new conventional wisdom. *British Journal of Political Science* **17**(4): 385–408

Vogel D 1989 *Fluctuating Fortunes: The Political Power of Business in America*. Basic Books, New York

Wallace H, Young A R (eds.) 1997 *Participation and Policy-making in the European Union*. Clarendon Press, Oxford, UK

Wilson G K 1981 *Interest Groups in the USA*. Clarendon Press, Oxford, UK

Wilson G K 1985, 1990 *Business and Politics: A Comparative Introduction*. Chatham House, Chatham, NJ

Wilson J Q 1996 The corporation as a political actor. In: Kaysen C (ed.) *The American Corporation Today*. Oxford University Press, Oxford, UK

D. Vogel

Politeness and Language

1. What is Politeness?

If, as many have claimed, language is the trait that most radically distinguishes *homo sapiens* from other species, politeness is the feature of language use that most clearly reveals the nature of human sociality as expressed in speech. Politeness is essentially a matter of taking into account the feelings of others as to how they should be interactionally treated, including behaving in a manner that demonstrates appropriate concern for interactors' social status and their social relationship. Politeness—in this broad sense of speech oriented to an interactor's public persona or 'face'—is ubiquitous in language use. Since, on the whole, taking account of people's feelings means saying and doing things in a less straightforward or more elaborate manner than when one is not taking such feelings into consideration, ways of being polite provide probably the most pervasive source of indirectness, reasons for not saying exactly what one means, in how people frame their utterances.

There are two quite different kinds of feelings to be attended to, and therefore there are two distinct kinds of politeness. One kind arises whenever what is about to be said may be unwelcome, prompting expressions of respect, restraint, avoidance ('negative politeness'). Another arises from the fact that long-term relationships with people can be important in taking their feelings into account, prompting expressions of social closeness, caring, approval ('positive politeness'). There are many folk notions for these kinds of attention to feelings—including courtesy, tact, deference, demeanor, sensibility, poise, discernment, rapport, mannerliness, urbanity, as well as for the contrasting behavior—rudeness, gaucheness, the making of social gaffes, and for its consequences—embarrassment, humiliation. Such terms label culture-specific notions invested with social importance, and attest both to the pervasiveness of notions of politeness and to their cultural framing.

2. How Can We Account for Politeness?

Since politeness is crucial to the construction and maintenance of social relationships, politeness in communication goes to the very heart of social life and interaction; indeed it is probably a precondition for human cooperation in general. Politeness phenomena have therefore commanded interest from theorists in a wide range of social sciences. Three main classes of theoretical approach to the analysis of politeness in language can be distinguished.

(a) Politeness as social rules. To the layman, politeness is a concept designating 'proper' social conduct, rules for speech and behavior stemming generally from high-status individuals or groups. In literate societies such rules are often formulated in etiquette books. These 'emic' notions range from polite formulae like *please* and *thank you*, the forms of greetings and farewells, etc., to more elaborate routines for table manners, or the protocol for formal events. Politeness is conventionally attached to certain linguistic forms and formulaic expressions, which may be very different in different languages and cultures. This is how the 'person on the street' tends to think about politeness, as inhering in particular forms of words.

Some analytical approaches to politeness are formulated in terms of the same sorts of culture-specific rules for doing what is socially acceptable, for example, the work by Ide and others on Japanese politeness as social indexing or 'discernment' (see Watts et al. 1992). In these approaches, politeness is a matter of social norms, and inheres in particular linguistic forms when used appropriately as markers of pre-given social categories. This approach is most appropriate for fixed aspects of language use—the more or less obligatory social marking of relatively unalterable social categories.

(b) Politeness as adherence to Politeness Maxims. Another rule-based approach derives politeness as a set of social conventions coordinate with Grice's Cooperative Principle for maximally efficient information transmission ('Make your contribution such as required by the purposes of the conversation at the moment'), with its four 'Maxims' of Quality, Quantity, Relevance, and Manner (See *Pragmatics: Linguistic*). Lakoff (1973) suggested that three 'rules of rapport' underlie choice of linguistic expression, rules which can account for how speakers deviate from directly expressing meanings. Choice among these three pragmatic rules ('Don't impose,' 'Give options,' 'Be friendly') gives rise to distinct communicative styles. Leech's more detailed proposal (1983) is in the same vein. Complementary to Grice's Cooperative Principle, Leech postulates a Politeness Principle—'Minimize the expression of impolite beliefs,' with the six Maxims of Tact, Generosity, Approbation, Modesty, Agreement, Sympathy. As with Grice's Maxims, deviations from what is expected give rise to inferences. Cross-cultural differences derive from the different importance attached to particular maxims.

The conversational maxim approach shares with the social norm approach the emphasis on codified social rules for minimizing friction between inter-

actors, and the view that deviations from expected levels or forms of politeness carry a message.

(c) Politeness as face management. A more sociological perspective places 'face work' at the core of politeness. Goffman (1967) considered politeness as an aspect of interpersonal ritual, central to public order. He defined face as an individual's publicly manifest self-esteem, and proposed that social members have two kinds of face requirements: positive face, or the want for approval from others, and negative face, or the want not to offend others. Attention to these face requirements is a matter of orientation to Goffman's 'diplomatic fiction of the virtual offense, or worst possible reading' (Goffman 1971 pp.138ff.), the working assumption that face is always potentially at risk, so that any interactional act with a social–relational dimension is inherently face-threatening and needs to be modified by appropriate forms of politeness.

2.1 Universal Principles of Politeness

Brown and Levinson (1987, henceforth B&L) introduced a new perspective by drawing attention to the detailed parallels in the construction of polite utterances across widely differing languages and cultures, and argued that universal principles underlie the construction of polite utterances. The parallels they noted are of two sorts: how the polite expression of utterances is modified in relation to social characteristics of the interloculors and the situation, and how polite utterances are linguistically constructed. At least three social factors are involved in deciding how to be polite: (a) one tends to be more polite to social superiors; (b) one tends to be more polite to people one doesn't know. In the first case, politeness tends to go one way upwards (the superior is less polite to an inferior); in the second, politeness tends to be symmetrically exchanged. In addition, (c) in any culture there are norms and values affecting the degree of imposition or unwelcomeness of an utterance, and one tends to be more polite for more serious impositions. In language there are also parallels, the linguistic structures for realizing particular kinds of politeness show remarkable similarities across languages. The politeness of solidarity is characterized, for example, by use of intensifiers, in-group identity markers and address forms, exaggerated intonation patterns, forms for seeking agreement and avoiding disagreement. Avoidance-based politeness is characterized by self-effacement, formality, restraint, deference, with the use of honorifics, hedges, indirect speech acts, and impersonalizing mechanisms like pluralization of pronouns, nominalization, and passive.

Why then are these kinds of detailed parallels across languages and cultures to be found in the minutiae of linguistic expression in socially analogous contexts? Explanations in terms of social norms or rules can account for politeness in a particular social group, but not the cross-cultural patterns, which seem to require a strategic account in terms of what people generally are trying to do when they are being polite. B&L (1987) proposed an abstract model of politeness wherein human actors are endowed with two essential attributes: *face* and *rationality*. Face consists of two specific kinds of wants: positive face (i.e., the desire to be approved of, admired, liked, validated), and negative face (the desire to be unimposed upon, unimpeded in one's actions). The second ingredient in the model—rationality—provides for the ability to reason from communicative goals to linguistic means that would achieve these goals. From these two assumptions—face and rationality—and the assumption that speakers mutually know that all speakers have these attributes, B&L developed a model of how speakers construct polite utterances in different contexts on the basis of assessments of three social factors: the relative power (P) of speaker and addressee, their social distance (D), and the intrinsic ranking (R) of the face-threateningness of an imposition. P, D, and R are seen as abstract social dimensions indexing kinds of social relationship (P and D) and cultural values and definitions of impositions or threats to face (R).

B&L (1987) distinguished five general types of strategies of politeness, ranging from avoiding a face-threatening act (FTA) altogether, to carrying it out but 'off record' (indirectly). On-record realization of an FTA can be done without any redressive action at all ('baldly'). It may be carried out with positive redress, which is essentially approach based, addressing the hearer's positive face wants by emphasizing closeness and solidarity. Politeness may also be carried out with negative redress, which is essentially avoidance based, addressing negative face wants for distance, deference, freedom from unexpectable impositions. Speakers are assumed to choose the linguistic framing of their utterance from this set of strategic possibilities according to the weightiness of the FTA, which is assessed with reference to the three contextually dependent social factors P, D, and R. For low levels of FTA-threat, bald-on-record, or positive politeness is most appropriate and cost effective; for higher levels, negative politeness is required; for the highest threats, indirectness is the safe option.

The claim, then, is that, however culturally variable the kinds of social relationship and kinds of face threat might be, underlying them are pan-cultural social dimensions (relative power, social distance, ranking of face threateningness) which universally go into the reckoning—and the interpretation—of strategic language choice, and hence one can derive the cross-cultural similarities in choice of linguistic realizations of politeness strategies that empirically seem to be in evidence. B&L (1987) claimed further that this model of politeness universals could be applied in particular cultural settings as an ethnographic tool for analyzing the quality of social relationships. Stable social

relationships are characterized in part by stable patterns of language use, which may distinguish particular societies or social groups.

2.2 Criticisms of the B & L Model

The goal of B&L to formulate an 'etic' set of concepts in terms of which politeness can be analyzed in 'emic' terms for any particular society was ambitious. Critics of the B&L model reveal several major points of contention about how a theory of politeness should be formulated.

2.2.1 The universality of face. Many critics have challenged. B&L's formulation (via Goffman and Radcliffe-Brown) of positive and negative face wants, as a valid way of conceptualizing the universal underpinnings of politeness. Negative face, in particular, considered as wants for freedom from imposition, appears entirely too embedded in Western individualism to sit well with conceptions of face in some other (e.g., Asian) cultures. In part, this is due to a misconstrual: the B&L face wants are abstract, they do not necessarily correspond clearly to conscious emic notions. What B&L (1987) claimed that *underlying* very diverse folk notions is a core of two interactionally relevant wants (for ratification, and freedom from imposition) which seem to be cross-culturally applicable, as desires concerning one's public self-image in the context of the moment which are *assumptions oriented to in interaction*. Other theorists (e.g., O'Driscoll 1996, Arundale 1999) have argued for notions of positive/negative face that are even more abstract, in terms of merging/individuation or closeness/separation, as the universal heart of politeness.

Challenges to the universality of the model also extend to the proposed hierarchy of increasing politeness (from bald-on-record to positive to negative to indirectness). Assessments of the P, D, R factors are situationally and culturally very variable, it is possible to accumulate different strategies in one utterance and to balance elements of negative politeness with positive politeness in one act, and indirectness is not always seen as the most polite option. These observations have led some researchers to argue against the possibility of identifying any kind of universal basis for polite behavior; politeness is simply incommensurate across societies. Those who take this extreme relativistic line can have no explanation for the observable cross-cultural parallels in patterns of language use, for how people manage (sometimes) to understand others from culturally different backgrounds, or for cross-linguistic parallels in the diachronic sources of honorifics from politeness strategies.

2.2.2 Politeness as communicated or taken for granted. In contrast with rule-based approaches, B&L (1987) insist that politeness inheres not in words or in sentences *per se*, but in utterances uttered in a context, by virtue of the successful communication of a polite attitude or intention. Polite utterances are not necessarily communicating 'real' feelings about anothers' social persona, but expressing contextually-expected concern for face. This concern is an 'implicature,' an inference of polite intentions, not a feature inextricably attached to particular linguistic forms. Politeness is ascribed to a speech act, or to an interactional move (if you prefer), not to a strategy or its linguistic realization *per se*.

In other approaches (for example, Fraser's (1990) 'conversational contract,' Watts et al.'s (1992) 'politic behavior'), politeness is taken to be the expected *background* to interaction; it is normally not communicated but consists in following expectations as to appropriate behavior.

2.2.3 Broad vs. narrow scopes for politeness theory. A narrower view takes politeness to be strategic orientation to potential face threats; there are some situations (e.g., task-oriented ones) where politeness may be subsumed to other goals, and there are many reasons for being indirect in speech other than politeness (e.g., humor, irony, rhetorical farce). Many motivations other than politeness guide human behavior. A more inclusive view sees politeness as orientation to the social-relationship dimension of every interaction, with attention to face taken to be an omnipresent necessity. The whole continuum from extreme politeness through a quite neutral level of politeness (maintaining the status quo, 'discernment') to rudeness (outright intentional threat) then needs to be brought into the theory.

2.2.4 Politeness from the point of view of the individual, the dyad, or the social group. The B&L model takes the interacting dyad as its unit; it is about how interlocutors make inferences of politeness from one another's deviations from Gricean efficient communication, and how stable patterns of strategies characterize dyadic interaction, providing an index to the quality of the social relationship. Many politeness strategies are quintessential examples of 'intersubjective perspective-taking'—putting yourself in the others' position. A major goal of B&L (1987) was to insist on the centrality of social interaction as a significant level of social life, intermediate between the individual and society, where social/cultural facts (status, role, values, norms, rights, and obligations) are integrated with individual ones (goals, plans, strategies).

Yet the Gricean foundation of the theory and the speech-act based formulation of the B&L strategies have made many see the model as purely psychological (how a speaker calculates how to frame an utterance).

Arundale (1999), for example, argues for a theory of how face is *jointly constituted* in ongoing interaction. In fact, we need both perspectives: face is indisputably interactionally created and manipulated. Nevertheless, it can be considered from the point of view of the individual speaker or hearer (as B&L do in their production/comprehension model), or of the society or social group (as 'face constituting theory' and most sociolinguists do).

2.2.5 The indeterminacy problem. The usefulness of the B&L (1987) model as an ethnographic tool for analyzing the quality of social relationships in any society is undermined by the contextual dependency of ratings of the social factors. The problems of the mixing and ordering of strategies are compounded by the indeterminacy of context-dependent P, D, R assessments, which make it hard to code levels of politeness in any concrete situation. This is a problem with any theory in terms of actors' intentions when applied to empirical data; as both interactors and conversation analysts know, it is not always possible to be certain what interlocutors' intentions are at a particular point in natural interaction.

The B&L model of politeness as originally formulated clearly needs elaboration and revision. Nonetheless, it retains its centrality in research on politeness, largely because it provides a coherent set of concepts for analytically dissecting polite speech in different societies and contexts. It is also the only theory that begins to provide an account for the detailed cross-cultural parallels in polite speech.

3. Conclusions

Politeness has attracted an enormous amount of research attention since the 1970s, and continues to be a major focus for research in disciplines concerned with social interaction. The study of politeness phenomena can provide insight into widely differing issues; foci of interest correspondingly differ widely. They include, for example, analyses of the sequential development of politeness in natural interaction, the study of cross-cultural pragmatics and misunderstanding, the ethnography of speaking of face and politeness in different cultures and contexts, politeness strategies as providing the stylistic coherence of particular types of interaction (e.g., gender differences in speech style), politeness as a functional motivation for linguistic structure, the social psychology of face management and interpersonal perception, politeness theory applied to the analysis of formal ritual and to a view of culture as 'rhetoric,' or forms of effective expression.

Research on politeness has, however, been much weaker on the theoretical front. A major limitation is the kinds of data used in analyses. A large proportion of studies take as their data people's conscious evaluations of politeness expressed in sentences, judgements which tend to be both prescriptive and stereotypical. Far fewer studies use as data recordings of situated conversational exchanges to explore how politeness is achieved sequentially in naturally-occurring discourse, and only a handful provide the crucial kind of evidence necessary to test the universality of any theory of politeness: for a particular society, an 'ethnography of speaking' providing evidence across different contexts to show how politeness is modulated in relation to social factors (P, D, R and others) in that society.

The emphasis is research has been largely on cross-cultural *differences*, with insufficient attention addressed to the cross-linguistic/cross-cultural parallels which tend to be taken for granted when they are not disputed. But politeness has a significance far beyond the P's and Q's of appropriate behavior and speech. The wider significance of politeness is in the interactional, communicative, day-to-day basis of social life, and the conduct of social relationships. Any work in this area needs to be based in a theory of social interaction that takes account *both* of our common human nature (and ability to communicate cross-culturally) and of our cultural differences (which makes us sometimes misunderstand one another), to anchor research on politeness—this is the great weakness in butterfly-collecting approaches.

See also: Communication and Social Psychology; Communication: Philosophical Aspects; Communicative Competence: Linguistic Aspects; Expression and Interaction: Microfoundations of Expressive Forms; Interactionism and Personality; Intergroup Relations, Social Psychology of; Pragmatics: Linguistic; Social Identity, Psychology of; Social Psychology; Status and Role, Social Psychology of

Bibliography

Arundale R 1999 An alternative model and ideology of communication for an alternative to politeness theory. *Pragmatics* **9**: 119–53

Brown P 1995 Politeness strategies and the attribution of intentions: The case of Tzeltal irony. In: Goody E (ed.) *Social Intelligence and Interaction*. Cambridge University Press, Cambridge, UK, pp. 153–74

Brown P, Levinson S C 1987 [1978] *Politeness: Some Universals in Language Use*. Cambridge University Press, Cambridge, UK

Brown R, Gilman A 1968 [1960] Pronouns of power and solidarity. In: Fishman J (ed.) *Readings in the Sociology of Language*. Mouton, The Hague, Netherlands, pp. 252–76

Coulmas F (ed.) 1991 New perspectives on linguistic etiquette. *International Journal of the Sociology of Language* **92**

Dufon M A, Kasper G, Takahashi S, Yoshinaga N 1994 Bibliography on linguistic politeness. *Journal of Pragmatics* **21**: 527–8

Fraser B 1990 Perspectives on politeness *Journal of Pragmatics* **14**: 219–36

Goffman E 1967 The nature of deference and demeanor. In: Goffman E *Relations in Public: Microstudies of the Public Order*. Harper and Row, New York
Kasper G 1990 Linguistic politeness: Current research issues. *Journal of Pragmatics* **14**: 193–218
Lakoff R 1973 The logic of politeness or minding your p's and q's. In: *Papers from the Ninth Reginal Meeting of the Chicago Linguistic Society*. Chicago, pp. 292–305
Leech G 1983 *Principles of Pragmatics*. Longman, London
O'Driscoll J 1996 About face: A defence and elaboration of universal dualism. *Journal of Pragmatics* **25**: 1–32
Strecker I 1998 *The Social Practice of Symbolization*. Athlone, London
Tracy K 1990 The many faces of facework. In: Giles H, Robinson W P (eds.) *Handbook of Language and Social Psychology*. Wiley, Chichester, UK, pp. 209–26
Watts R J, Ide S, Ehlich K (eds.) 1992 *Politeness in Language*. Mouton, Berlin

P. Brown

Political Advertising

Political advertising has been the subject of intense scholarly investigation for nearly a century. Two agendas underlie this strain of social science inquiry.

The first agenda seeks to measure how political advertising, and campaigns generally, produce votes. Candidates' campaign advertisements are a major source of information about electoral choices that voters face. The intellectual origins date to Gosnell's studies of voting in Chicago in the 1920s (Gosnell 1927, 1935). How much and how do advertisements influence election outcomes? The answers to these questions have very practical uses for political practitioners, who seek information that can guide campaign strategies, and for social scientists, who seek to explain how elections are won. Many social scientists question whether political advertising has any influence on elections, because elections are readily forecasted using indicators of peace and prosperity. Key writes in *The Responsible Electorate* that 'The hullabaloo of a presidential campaign so commands our attention that we ascribe to campaigns great power to sway the multitude. Campaigning does change votes and it does bestir people to vote. Yet other influences doubtless outweigh the campaign in the determination of the vote' (1966, p. 9). On election forecasts, see Rosenstone (1983).

The second agenda concerns the extent of voter autonomy. Are votes hollow echoes of campaign messages, or does political discourse enable voters to reach reasonable decisions that reflect their own preferences and values? Electoral reform motivates much research on advertising and voter autonomy. If political advertising easily misleads voters, then restrictions on campaign advertising or some sort of 'political consumer protection' may be necessary in order to make democracy work. To the extent that there is a scholarly consensus on this matter it is that concerns about voter autonomy are overdrawn. Behavioral research in a wide variety of settings indicates that advertising has modest but important effects on election outcomes: advertising can make a difference in a close election, but it does not appear to persuade large numbers of voters to turn against a candidate or party that is well-known, well-liked, and represents the voters' interests.

1. *Advertising and Vote Production*

Political advertising covers a wide range of media that candidates use to mobilize supporters to turn out or to persuade people to vote for them. Commentary on political advertising often equates it with television advertising, as, for instance, when European commentators descry the 'Americanization' of their politics. But even in the US there is, in fact, much less television in politics than is often supposed. Campaigns for the US House of Representatives, for instance, use as much direct mail as television advertising. And in local campaigns candidates use little television but a vast range of other media, including mail, billboards, lawn signs, telephones, and door-to-door canvassing (Key 1966).

Campaigns' usage of different media appears to follow the same principles that apply to firms. Candidates choose what media to use based on the productivity of the media and their relative prices. Variation in the mix of media used at different levels of office suggests that there are economies of scale to different sorts of advertising. Personal contact is reputed to be the most effective mode of political communication and it is a staple of local elections everywhere. But it is impossible to spend 15 minutes with each of the millions voters in a national election.

Mass media are most cost-effective at reaching broad constituencies. Evidence of the rational use of media comes from US House races. The average household in the typical House district receives about two pieces of mail and sees at least two television advertisements. However, candidates choose a mix of advertising media based on the prices and productivity of those means of reaching potential voters (or consumers). Campaigns for the US House in districts with expensive broadcasting media (such as Los Angeles and New York) spend $2.50 on mail for every $1 broadcasting, or three pieces of mail and 15 seconds of advertising per person. House campaigns in districts with inexpensive broadcasting (such as Dayton, Ohio and El Paso, Texas) spend $10 on broadcasting for every $1 on mail, or one piece of mail and 8 minutes of advertising per person. (For details see Ansolabehere et al. 1999a.) To my knowledge, similar sorts of

calculations have not been conducted for other countries, though it is evident from patterns of media usage that parties and candidates in Europe, Australia, and elsewhere are similarly responsive to media prices. (See, for example, Blumler 1983 and Australian Election Commission 1997.)

Social science research has focused on how voters respond to the overall volume and different sorts of campaign advertising that they receive. The estimated effects depend in part on the methods used.

Three different approaches have been taken to the study of advertising. First, laboratory and field experiments have been used to gain precise measures of the effects of advertising exposure in specific campaign settings. Such studies control what messages individuals receive and then measure differences in opinions or behaviors in response to different sorts of messages (e.g., Gosnell 1927, 1935, Ansolabehere and Iyengar 1996, Gerber and Green 1999, Norris 1999).

Second, studies of aggregate data correlate the observed vote to levels of campaign spending. Some of these studies design their inquiries around 'quasi-experiments' which exploit factors that change levels of spending (such as personal wealth) in order to measure the effect of spending on the vote. The seminal work in this vein, for both American and European researchers, is Jacobson (1978). Subsequent research has improved on this model; see, for example, Gerber (1998). Races involving three or more candidates pose statistical problems in estimation of vote production functions, for which widely accepted statistical models have not yet been developed. Third, surveys associate reported vote or political attitudes with reported or recalled exposure to messages. Almost all research outside of the US relies on survey methods. For a discussion of the problems with survey based measures of campaign effects see Ansolabehere et al. (1999b). Analogous with the aggregate studies, instrumental variables can solve such problems.

Experiments, which offer high internal validity, show a remarkably consistent set of results about the capacity of advertising to produce votes. Researchers observe exactly what messages citizens see and can design the study so that other factors do not confound the effect of the experimental manipulation. Perhaps the earliest experimental study of political advertising and campaigning took place in the 1920s in the US. To measure the mobilizing effects of direct campaigning, Gosnell (1927) sent a cadre of graduate students out to canvass a set of blocs of the city of Chicago in an effort to manipulate the levels of voter contact. They found higher in turnout ranging from 4 to 7 percent in the areas that the students canvassed over the areas that they did not canvass. Fast forward to the 1990s. Gerber and Green (1999) replicated Gosnell's study in several towns in southern Connecticut. They found increases in participation similar to those reported in the Chicago studies 70 years earlier. They also found increases in vote shares of candidates in the neighborhood of 4 percentage points for two pieces of mail. In a series of laboratory experiments conducted in Los Angeles from 1990 to 1993, Ansolabehere and Iyengar (1996) found strikingly similar results for television advertising. Exposure to a 30-second television advertisement raised the percentage of viewers who intended to vote for a candidate by 4 to 10 points. These and many other experiments show substantial effects of political advertising on candidates vote shares. Those who see a broadcast message, receive a mailing, or talk to a campaign worker are between 4 and 10 percent more likely to support the candidate being promoted.

Aggregate studies of campaign spending yield similar, albeit lower estimates of the effectiveness of campaign advertising. Using fairly complicated, and controversial, statistical models, researchers have attempted to measure the marginal effect on the vote of spending an additional campaign dollar. Such techniques yield estimates ranging from miniscule effects of a dollar spend on the vote to estimates of approximately $10 per vote. The key papers in this subfield are Jacobson (1978), Green and Krasno (1988), and Gerber (1998).

Such estimates are in the right ballpark, but the wide range of estimates in the literature on campaign spending suggests that additional work finding valid statistical models is in order before we can put much confidence in the estimates emerging from this literature.

The problem is that it is difficult to sort out cause and effect in aggregate electoral data. Consider, for example, two similar legislators: one holds a highly competitive seat and the other holds a very safe seat. The legislator in the competitive seat must spend large amounts of money to win re-election; the legislator in the safe seat is unlikely to be challenged and if they are they do not need to spend much to retain the seat. In this example the correlation between spending and vote shares is negative! To fix this problem researchers have taken a 'quasi-experimental' approach, using factors such as candidate wealth, which affect directly how much candidates spend and in turn affect their vote shares. Improving such statistical models remains an important research challenge.

Survey research has presented the most modest estimates of the effectiveness of advertising. A long line of studies report minimal effects of campaigns in general and of advertising in particular on the vote (in the US, see Campbell et al. 1960; in the UK, see Norris 1999). Serious problems of question validity lead to serious doubts about these research findings. In the laboratory setting, Ansolabehere and Iyengar (1996) document that survey-based measures of advertising exposure miss one exposure for every one that they detect. This degree of misreporting is sufficient to explain why surveys often produce insignificant findings or even estimates in the 'wrong' direction. Quasi-experimental techniques analogous to those used in

aggregate studies can correct measurement problems (Bartels 1993).

Beyond differences in media and methodologies, a further theme in the study of political advertising is the mechanism through which messages become votes. There are three. Advertising may solidify an uncertain vote (reinforcement), it may stimulate a supporter to vote (mobilization), and it may lead a voter to change their mind (conversion). This typology is due to Berelson et al. (1954). Although political observers often talk of 'swing' voters and movement in the polls, there appears to be relatively little evidence in experiments or surveys of conversion. Few Republicans decide to vote Democrat and few Democrats decide to vote Republican in response to political advertisements. Perhaps the strongest evidence against conversion, though, comes from nonpartisans. Nonpartisans, especially low information independents, are the least responsive to political advertising.

Reinforcement appears to be the dominant mechanism in political communication. For instance, in the experimental research reinforcement produces the clearest effect on the vote. Nonpartisans and viewers of the opposite party as the sponsor of the ad showed no statistically significant response to a candidate's ads. Low information partisans of the same party as the sponsor showed by far the strongest favorable response to political advertising. That effect shrank, though it remained statistically and substantively large, with the partisan's level of information and interest in politics (Ansolabehere and Iyengar 1996). This pattern suggests that political advertising fits neatly into the more general model of information acquisition developed by Zaller, in which both predisposition to believe a piece of information and prior levels of knowledge determine responsiveness to new information (Zaller 1992).

Mobilization has proved the most controversial of the three mechanisms. Experimental research on political advertising has called into question the longstanding belief among political scientists that campaigns are inherently stimulating affairs. Berleson et al. (1954) describe the process in detail. This is essentially the model behind retrospective voting developed by Key (1966) and widely used to understand elections in the US, Europe, and elsewhere. A large number of experimental studies have demonstrated that not all viewers are stimulated to vote by political advertisements. For example, nonpartisan and low information viewers exposed to negative advertising appears express significantly lower intentions to vote than others. As a result, negative campaigning can produce lower turnout among the key group of 'swing' voters and an electorate that reflects two ideologically opposed groups.

The demobilizing effects of negative advertisements on non-partisans underscores the difference between reinforcement and mobilization. Negative advertising might win the highest vote share, mainly by reinforcing partisans, but it can lead to lower overall participation.

One important caveat about the state of social science research on political advertising is in order. The bulk of the research on the effectiveness of political advertising has been conducted in the United States. Many of the findings about mobilization and reinforcement likely depend on the electoral system, which influences the number of parties or candidates. The dynamics of campaigns and the consequences of advertising appear to be quite different in races involving more than two candidates. Johnston et al. (1992) provide excellent examples of political communications having complicated effects on competition in Canada's multi-party system. Careful experimental research on multicandidate and multiparty races, though, is in its nascency.

2. *Voter Autonomy*

The central problem for democratic theory in modern societies is that people apparently know little about the choices they face and they have little incentive to seek new information. Candidates and parties, then, have very strong incentives to provide information to voters at little or no search costs: candidates try to reach us, rather than us, them. As a result we are highly reliant on what politicians choose to say in their advertisements.

This is the problem of voter autonomy. By one telling voters have abdicated to political elites the important jobs of choosing what issues will be discussed, what elections and therefore government will be about, and even what we will think as people walk into the voting booth. From this view, people are highly susceptible to false or deceptive advertising and will vote for the candidate whom we see the most rather than the one we think is best representative. Regulation of the volume and content of political advertising may, therefore, be necessary.

The counter argument is that voters, even though they do not know much by way of details, have a clear sense of what they themselves value or prefer, and, in order to win elections, politicians must adjust their messages to what the voters want. In crafting their advertisements, a candidate selects issues on which he or she can make electoral gains, because the voters care about it, because the candidate has a strong record, or because the candidate has have taken the more popular position. This leads to majoritarian biases in what is discussed, but that is the essence of elections.

Voter autonomy presents a much more difficult set of research questions. In order to say that people are or are not misled by political communications, we must establish what it is that they prefer or value *a priori*. The effects of advertisements must then be judged against that baseline. Two questions are central.

First, do advertisements lead people to choose a candidate who is less good for their interests or values? Research showing strong reinforcement and few conversion effects suggests that the answer to this question is probably no. In both survey and experimental data it appears that people who lean toward a party become more solidly in favor of that party as the campaigns progress (Ansolabehere and Iyengar 1996, Chap. 4, Gelman and King 1990). An important variation on this arises from tactical voting in Britain, Canada, and other countries. As campaigns progress, voters supporting a candidate or party in third place or lower might vote for the second place candidate if they prefer that candidate over the front-runner. This seems to have happened in the 2000 Mexican election as support for Cardenas fell over the course of the campaign and support for Fox, who was in second place most of the race, rose.

Nonpartisans tend to move very little, and tend not to vote. The immediate political implication is that campaign discourse is polarizing: it ultimately pits Democrats against Republicans, Conservatives against Labourites. The low amount of conversion suggests, however, that people fundamentally are not misled.

Second, can advertising set the public agenda, or are people mainly responsive to what they themselves care about? While agenda setting effects have been widely documented in other areas of study, they have received surprisingly little careful study in the field of political advertising. Early evidence of agenda setting is provided by Bereleson et al. (1954, Chap. 12). Two careful studies stand out. Jamieson's (1992) *Dirty Politics* documents how the Republican and Democratic strategists use their advertising and the news media to magnify their own agendas, especially Republican's usage of crime and race. Johnston et al. (1992) demonstrate that the debate over free trade completely transformed the Canadian voters' priorities, without changing their preferences on free trade itself.

Reinforcement and agenda setting can hold conflicting implications for voter autonomy. Partisan reinforcement suggests that there is little evidence that people vote against their expressed preferences; agenda setting suggests that people are vulnerable to what the campaigns choose to discuss. There may, however, be no contradiction here. It may indeed be the case in each of the examples discussed that voters were simply very receptive to new agendas, and that elections gave voters an opportunity to change public policies on trade, affirmative action, and other issues from those set decades earlier.

3. New Directions

Political advertising poses both methodological and normative challenges for social scientists, and our understanding on both levels has advanced considerably over the last two decades. Advertising has important effects on election outcomes, but it does not appear to lead people to vote for candidates with whom they fundamentally disagree. It, therefore, does not on the whole undercut the normal workings of majority rule in democratic elections.

In terms of methods, a high priority for further inquiry is bridging the gaps between experimental, aggregate, and survey methods, and developing more powerful quasi-experimental approaches to correct measurement and simultaneity problems. To this end, research on vote production, which has relied mainly on quasi-experiments in aggregate data, should make greater use of experiments to map out features of the vote production function more fully, especially the effects of repeated exposures to messages and interactions between the messages and the viewers preferences.

In addition, there is considerable need for research on political advertising outside the US. The American focus of campaign research stems in part from existing regulations on elections and media. The US imposes few limits on how candidates and parties raise and spend campaign funds, on their access to media, and on what can be said on the various media. As a result, the US has proved a rich field to study advertising effects. Research on campaigning in European countries has placed a much greater emphasis on electoral laws, political parties, and tactical voting. There is, however, growing interest in campaign effects in European political science.

Beyond methodological differences, there is the matter of why. Reinforcement and mobilization are a sufficiently well established phenomenon that it is worth opening up the black box. Why are these the routes through which political advertising works, and why not conversion? The highly inductive approach of empirical researchers is likely to pay the biggest dividends, as political theory has yet to produce a powerful model of why people vote.

See also: Mass Communication: Normative Frameworks; Mass Media, Representations in; Political Communication; Political Discourse; Political Money and Party Finance; Political Parties, History of; Political Science: Overview; Polling; Television: Industry; Voting, Sociology of

Bibliography

Ansolabehere S, Gerber A, Snyder J M Jr 1999a *How Campaigns Respond to Media Prices*. MIT, Cambridge, MA

Ansolabehere S, Iyengar S 1996 *Going Negative: How Political Advertising Shrinks and Polarizes the Electorate*. Free Press, New York

Ansolabehere S, Iyengar S, Simon A 1999 Replicating experiments using surveys and aggregate data: The case of negative advertising. *American Political Science Review* **93**: 901–10

Australian Election Commission 1997 *Funding and Disclosure Report, 1996*. AEC, Sydney, Australia

Bartels L 1993 Message received. *American Political Science Review* **87**: 267–85
Berelson B, Lazarsfeld P, McPhee W 1954 *Voting*. University of Chicago Press, Chicago
Blumler J 1983 *Communicating to Voters: Television in the First European Parliament Elections*. Sage, London
Campbell A, Converse P, Miller W, Stokes D 1960 *The American Voter*. University of Chicago Press, Chicago
Fritz S, Morris D 1992 *The Handbook of Campaign Spending, 1990*. CQ Press, Washington, DC
Gelman A, King G 1990 Why are presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* **23**: 409–52
Gosnell H F 1927 *Getting-out-the-vote: An Experiment in the Stimulation of Voting*. University of Chicago Press, Chicago
Gosnell H F 1935 *Machine Politics: Chicago Model*. University of Chicago Press, Chicago
Green D, Krasno J 1988 Salvation for the spendthrift incumbent. *American Journal of Political Science* **32**: 884–907
Gerber A 1998 Estimating the effect of campaign spending on senate election outcomes using instrumental variables. *American Political Science Review* **92**: 401–11
Gerber A, Green D 1999 Does canvassing increase voter turnout? A field experiment. *Proceedings of the National Academy of Sciences* **96**: 10939–42
Jacobson G 1978 The effects of campaign spending in congressional elections. *American Political Science Review* **72**: 469–91
Jamieson K H 1992 *Dirty Politics*. Oxford University Press, New York
Johnston R, Blais A, Brady H E, Crete J 1992 *Letting the People Decide: Dynamics of a Canadian Election*. Stanford University Press, Palo Alto, CA
Key V O 1966 *The Responsible Electorate*. Vintage, New York
Morris D, Gamache M 1994 *The Handbook of Campaign Spending, 1992*. CQ Press, Washington, DC
Norris P 1999 Political communications in British election campaigns: Reconsidering media effects. *British Elections and Parties Review* **5**: 190–211
Rosenstone S 1983 *Forecasting Presidential Elections*. Yale University Press, New Haven, CT
Zaller J 1992 *The Nature and Origins of Mass Opinion*. Cambridge University Press, New York

S. Ansolabehere

Political Anthropology

1. Definitions

Political anthropology is a subfield of sociocultural anthropology and as such is concerned with the comparative, fieldwork-based study of politics in a broad range of social and cultural contexts. If sociocultural anthropology is above all a comparative discipline, concerned with the full variety of human social and cultural life, the existence of a coherent subfield of political anthropology presupposes a relatively clear definition of what constitutes 'the political.' Classic studies in political anthropology in the 1940s and 1950s concerned themselves with the maintenance of order and the regulation of force in societies apparently lacking the institutional forms of the modern state. In contrast, more recent studies in the anthropology of politics are as likely to be concerned with issues of citizenship and exclusion, democracy and nationalism, law and violence, but they coexist with a whole corpus of work which identifies and analyzes the workings of power in all social contexts.

There are three overlapping ways in which anthropologists might tackle the definitional issue. One would be formal: political anthropology is the study of 'political behavior,' wherever and however encountered, which in turn may be defined in various ways—for example, in terms of the pursuit of power, of strategizing, or of instrumentality. A second way would be to concentrate instead on function: the political refers to institutions and practices which perform political functions—the maintenance of social order, the succession to positions of power and authority, the settlement of disputes. The third would be substantive and based on a recognition that 'the political' is not a universal category of experience, but is a product of a certain history: it emerged in Western Europe after the Reformation, alongside the modern idea of the state, and as part of a move towards recognizably modern forms of mass politics based on the institutions of representative democracy. Insofar as such a category now also exists across the spectrum of postcolonial and postsocialist states, political anthropology is the study of whatever it is people in those places believe politics to be.

These three approaches to the anthropology of politics are further complicated by the relationship between political anthropology and other disciplines concerned with the same topics. Both the formal and the functional versions of political anthropology have obvious affinities with political science and political sociology, but less apparent overlap with political philosophy and political theory. In the early days of political anthropology, there was a reaction against philosophical accounts of human political capacities, such as social contract theories of the state, which relied on evolutionary speculation about the primitive origins of contemporary political institutions. Political anthropology's affinity for political science and hostility to political theory also stemmed from political theory's tendency to combine and confuse the ideal and the actual, accounts of how things *are* with speculations on how they *ought* to be (Fortes and Evans-Pritchard 1940, p. 4). However, in the latter half of the twentieth century, anthropological accounts of politics have become less confident about the reliance on observed behavior and the appeal to dispassionate science. In particular, since the 1970s, the influence of the work of Michel Foucault has focused attention on the alleged ubiquity of power, while opening the way for an empirical reexamination of the language and

claims of liberal political philosophy. At the same time, the recognition that the anthropological science of 'observed behavior' was almost always a colonial science has shifted attention from the comparative study of politics to a reflexive concern with the political constraints that have shaped the emerging discipline of anthropology. This double concern—with the anthropology of politics *and* with the politics of anthropology—animates Vincent's recent authoritative overview of the subject (Vincent 1990).

Translated into more conventional historical terms, there is a tale to be told of the rise, fall, and partial resurrection of political anthropology. Between the 1940s and 1960s the comparative study of political institutions became a central anthropological interest, especially for social anthropologists in the European tradition. From the early 1970s political anthropology swiftly fell out of fashion, before a number of political themes—power, nationalism, resistance, the state—returned to the anthropological agenda in the mid 1980s. Even so, many anthropologists who studied obviously political topics, but not in the style of 1960s political anthropology, were reluctant to identify themselves with the label of 'political anthropologist.'

2. *From Function to Process to Form*

A comparison of three key works from the early years of political anthropology provides a better sense of the strengths and weaknesses of canonical political anthropology. In 1940 Edward Evans-Pritchard published his monograph on *The Nuer*, subtitled 'A description of the modes of livelihood and political institutions of a Nilotic people' (Evans-Pritchard 1940). Evans-Pritchard had conducted his research among this group of people in the 1930s at the request of the colonial government. His book found the key to Nuer political order in the field of kinship, specifically in the structure of dominant lineages, and the systematic way in which segments of the population would unite in antagonism to structurally equivalent segments. The result was an extraordinarily neat and coherent account of the emergence of political order in the absence of a unitary ruler. Key issues in Evans-Pritchard's account—such as the ordered exchange of violence in the institution of the feud, or the structure of so-called 'segmentary lineage systems'—generated their own specialist literatures, and inevitable critiques, in the years that followed.

Fourteen years after Evans-Pritchard's Nuer study, Edmund Leach published an analysis of highland Burma of comparable intellectual ambition (Leach 1954). Like Evans-Pritchard, Leach concentrated on a group of people on the further fringes of colonial power. As his title suggested, Leach was also concerned with the idea of the political system, or in this case *systems*, and, like Evans-Pritchard, he rooted parts of his analysis in the everyday structures of kinship and marriage. But Leach was also concerned to historicize his account, and to insert a dynamic element. Dissatisfied with the conservative assumptions of functionalist anthropology, Leach attempted to problematize the idea of a system in equilibrium, by showing how, over a relatively wide area and a relatively long period of time, the same people could oscillate between an egalitarian and a hierarchical political structure, each of which was inherently unstable and contained within itself the germ of its opposite.

The same concern with the dynamic working through of internal contradictions could be found in Victor Turner's (1957) analysis of village micropolitics in what was then Northern Rhodesia (now Zambia). Eschewing the more geographically and historically ambitious vision of Evans-Pritchard and Leach, Turner concentrated on the meticulous analysis of specific cases of conflict within a specific village. Pattern and form were to be found not so much in a social structure, apparently frozen in time, but in the orderly sequence of events within what Turner described as his 'social dramas,' while the shared symbols in village ritual bound people together, even as political conflict drove them apart.

The studies of Evans-Pritchard, Leach, and Turner assumed canonical status in the emerging subdiscipline of political anthropology, and together they can give a fairly good idea of what kind of work political anthropologists of the 1950s and 1960s might be expected to do. First, and most obviously, political anthropologists were relatively unconcerned with the political institutions of the modern nation state, and, although all three authors were working within the political structures of late colonialism, European rule was not a central concern of their work. Their empirical focus was on relationships of kinship, marriage, and residence, and their intellectual goal was to locate the principles of local order in the details of everyday conflict and show the structural consequences of those principles as they applied across a wider social field. Yet, as we move from one author to the next, two related changes emerge, and one unresolved dilemma is left hanging. The two changes were the shift—evident in the contrast between Evans-Pritchard and Leach, and further developed in Turner's work—from a concern with *structure* to a concern with *process*; and the parallel move from relatively large-scale models of social structure to what Turner described as 'diachronic micro-sociology' (1957, p. 328). The unresolved dilemma concerned the cultural content of politics.

The 1960s were the subdisciplinary heyday of political anthropology, and many of the edited collections and monographs of this period celebrate this shift towards a processual view of politics, and evince a growing concern with the microstudy of political behavior (cf. Swartz et al. 1966, Bailey 1969). In the terms set out at the start of this article, if political

anthropology had started out as the study of political *functions* in societies apparently lacking modern political institutions, it gradually became the *formal* study of political behavior in all kinds of society. But the formal definition often rested on a hostility to the cultural meaning of politics. In *African Political Systems*, Fortes and Evans-Pritchard had insisted on the need to ignore culture for comparative purposes: 'A comparative study of political systems has to be on an abstract plane where social processes are stripped of their cultural idiom and are reduced to functional terms' (1940, p. 3). In *Political Systems of Highland Burma*, cultural differences became mere tokens in the political games analyzed by Leach (1954, pp. 16–17). Yet this antagonism to the cultural meaning of politics was never simple and complete: Turner concludes *Schism and Continuity* with a recognition of the binding force of ritual and symbols—paradoxically, a point also acknowledged by Fortes and Evans-Pritchard (1940, pp. 16–22)—and the promise, richly fulfilled in his subsequent work, of more analysis in the future of 'the cultural aspects of Ndembu ritual' (1957, p. 331).

3. *After Formalism: Domination, Resistance, and the Nation*

The emphasis on formal models of political behavior owed much to the prevailing positivism of political science, especially North American political science, but was increasingly out of joint with the anthropological *zeitgeist*. The 1960s and 1970s saw a major shift away from the formal, typologizing type of work found in political anthropology, and towards the more humanistic style once characterized by Evans-Pritchard as the 'translation of culture.' In Europe, this shift was given theoretical impetus by Lévi-Strauss' structuralism, which directed attention to the study of myth and ritual; in North America, the same role was played by the cultural analyses of Clifford Geertz and David Schneider, with their roots in Parsonian theory, and their concern with culture as a system of symbols and meanings. By the early 1970s, a leading political anthropologist described an intellectual landscape unhelpfully divided between those he characterized as the 'thought structuralists' and those interested in power relations (Cohen 1974). The task, ever since, has been to attempt some sort of *rapprochement* between anthropology's symbolic wing, concerned with the explication of cultural meaning, but often at the expense of any concern with social context and social process, and its political specialists, who had been increasingly identified with a rather narrow and culturally arid formalism.

One area of convergence that emerged in the 1980s was in the anthropology of power and resistance. This represented a development of themes that had first emerged in the feminist and Marxist work of the 1970s. To what extent are marginal or oppressed people—women, the poor—mystified by the ideological representation of their own position? Is culture a source of such mystification, or is it a resource for resistance? Is culture internally homogeneous or, as the slogan of the times put it, is it 'always contested'? Theoretical answers to these questions were found in a range of new sources—Marxist and post-Marxist theorists of culture like Raymond Williams, Antonio Gramsci, Pierre Bourdieu, and Stuart Hall; feminists like Judith Butler and Gayatri Chakravorty Spivak; the radical South Asian historians of the Subaltern Studies group; and, perhaps above all, Michel Foucault. Empirically, instances of resistance might be found in everything from changing domestic consumption patterns, to agricultural workers' self-conscious footdragging, to participation in new religious movements. From a considerable distance, it might be possible to discern a strong continuity with the 1960s analysis of local micropolitics, but the antipositivism and self-conscious poststructuralism of many writers on resistance left them reluctant to acknowledge too strong a kinship with the political anthropology of a previous generation.

A second area of convergence between anthropology's symbolic and political tendencies was in the study of nationalism. The wave of decolonization in the decades after World War II provided anthropology with an obvious area of research. But despite interesting attempts at interdisciplinary enquiry, such as Chicago's Committee for the Study of New Nations, the anthropology of nationalism failed to take off in the 1950s and 1960s. The publication of a provocative essay by a highly 'anthropological' political scientist, Benedict Anderson's (1983) *Imagined Communities*, put nationalism back on the anthropological agenda, and the emergence of ethnic and nationalist conflict in the years after the collapse of Soviet hegemony in Europe provided further impetus. However, while the new anthropology of nationalism has produced a great deal of important and imaginative empirical research, it has been somewhat less theoretically impressive. Some of the most impressive anthropological work has been on the experience of ethnic and nationalist violence, but this has been overwhelmingly studied from the point of view of the victims, rather than the perpetrators, of violence (e.g., Das 1990). The big political mystery of the twentieth century, mass mobilization in the name of the nation, remains as problematic for anthropology as for the other social sciences. In the face of the breakup of the Soviet empire and the war in former Yugoslavia, political anthropology's key contribution in the 1960s—instrumental analyses of the politics of ethnicity—now looks dangerously reductionist. The major theoretical gain from the anthropology of nationalism has been a new attentiveness to the political history of anthropology's central concept of 'culture,' and a reflexive awareness of anthropology's potential complicity in

nationalist interpretations of culture and cultural difference (Handler 1988).

4. Towards Complexity

Although the label 'political anthropology' still retains some of the stigma of 1960s formalism, the study of politics has returned to the center of sociocultural anthropology since the 1980s. There are two obvious and related reasons for this. One is an increased openness to interdisciplinary collaboration with historians, political scientists, and political theorists as anthropologists grapple with the spatial and temporal limitations of the ethnographic method. The other is a renewed critical interest in the received vocabulary of modern politics, obviously dense and complex ideas like 'the nation' and 'the state,' but also concepts of 'citizenship,' 'civil society,' and 'democracy.'

Ethnographic enquiry itself has changed, opening up new potential issues for political anthropology. A study like James Ferguson's *The Anti-Politics Machine* (1990) exemplifies what is now possible. Ferguson's book is based on an ethnographic investigation of a large and unsuccessful development project in Lesotho. It involves at once a critical examination of the 'development discourse' that informs projects like these, and a description of the systematic ways in which potential political issues, like the sources and consequences of social inequality, are swept aside in the path of the project, while all the while the reach of state power is extended into the countryside. Theoretically, Ferguson's work owes a great deal to Foucault's critique of liberal complacency, but empirically it combines local ethnographic detail with broader understandings of the translocal processes and institutions that are so characteristic of the age of development.

In some ways, the new anthropology of the state represents a return to the founding suspicion of political philosophy in *African Political Systems*. In the Preface to that volume, Radcliffe-Brown had dismissed the idea of the state as 'a fiction of the philosophers.' In a series of provocative essays, Michael Taussig has raised the issue of the productive potential of this 'fiction': the—mostly symbolic—ways in which people are subjected to the 'state effect' (Taussig 1997). In a collection of critical studies of African politics, John and Jean Comaroff (1999) subject the notion of 'civil society' to a similar process of empirical scrutiny. Thomas Hansen's (1999) study of Hindu nationalism, combines history, ethnography, and Lacanian theory, and locates the rise of politicized Hindu identity in the peculiar postcolonial trajectory of Indian democracy. These studies represent an especially interesting departure in the relationship between political anthropology and political theory. They focus on central concepts in liberal political philosophy but, rather than viewing that philosophy as a potentially universal and neutral idiom, treat it as one more folk theory of government, better viewed as an object of ethnographic enquiry than a source of wholesale explanation.

See also: Evans-Pritchard, Sir Edward E (1902–73); Nationalism: General; Political Economy in Anthropology; Political Sociology; Power in Society; Power: Political.

Bibliography

Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London
Bailey F G 1969 *Stratagems and Spoils: A Social Anthropology of Politics*. Blackwell, Oxford, UK
Cohen A 1974 *Two-Dimensional Man: An Essay on Power and Symbolism in Complex Society*. Routledge and Kegan Paul, London
Comaroff J L, Comaroff J (eds.) 1999 *Civil Society and the Political Imagination in Africa: Critical Perspectives*. University of Chicago Press, Chicago
Das V (ed.) 1990 *Mirrors of Violence: Communities, Riots and Survivors in South Asia*. Oxford University Press, Delhi, India
Evans-Pritchard E E 1940 *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People*. Oxford University Press, Oxford, UK
Ferguson J 1990 *The Anti-Politics Machine: 'Development,' Depoliticization, and Bureaucratic Power in Lesotho*. Cambridge University Press, Cambridge, UK
Fortes M, Evans-Pritchard E E (eds.) 1940 *African Political Systems*. Oxford University Press, London
Handler R 1988 *Nationalism and the Politics of Culture in Quebec*. University of Wisconsin Press, Madison, WI
Hansen T B 1999 *The Saffron Wave: Democracy and Hindu Nationalism in India*. Princeton University Press, Princeton, NJ
Leach E R 1954 *Political Systems of Highland Burma: A Study of Kachin Social Structure*. Athlone, London
Spencer J 1997 Post-colonialism and the political imagination. *Journal of the Royal Anthropological Institute* **3**: 1–19
Swartz M, Turner V W, Tuden A (eds.) 1966 *Political Anthropology*. Aldine, Chicago
Taussig M 1997 *The Magic of the State*. Routledge, London
Turner V W 1957 *Schism and Continuity in an African Society: A Study of Ndembu Village Life*. Manchester University Press, Manchester, UK
Vincent J 1990 *Anthropology and Politics: Visions, Traditions, and Trends*. University of Arizona Press, Tucson, AZ

J. Spencer

Political Communication

Political communication is an interactive process concerning the transmission of information among politicians, the news media, and the public. The process operates downward from governing institutions toward citizens, horizontally in linkages among political actors, and also upward from public opinion toward authorities. The literature in political

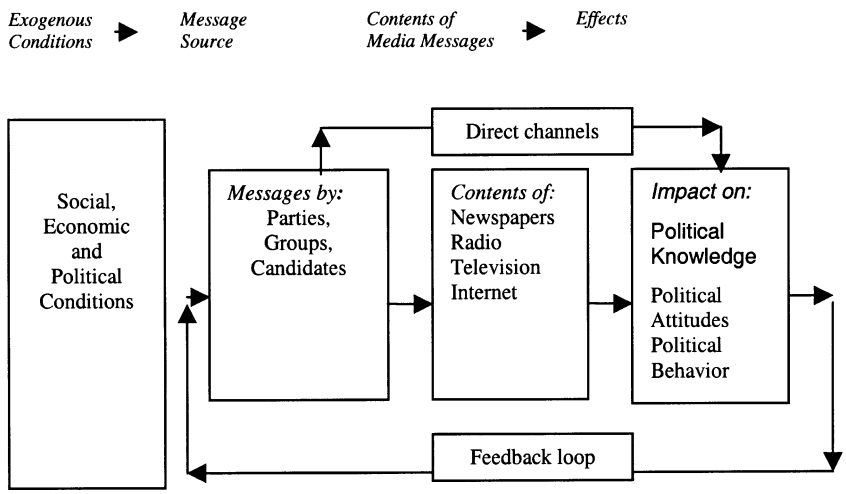

Figure 1
Systems model of political communication process

communication can be subdivided into three major categories, using a simple systems model of the process illustrated in Fig. 1 to distinguish between production, contents, and effects.

1. *Production Processes*

Work on the production process focuses on how messages are generated by political actors like parties and interest groups, and then transmitted via both direct channels like political advertisements and indirect channels including newspapers, radio, and television. Many studies have focused on the increased professionalization of political marketing campaigns in the postwar era, including the rise of the class of political consultants, pollsters, advertising executives, and their coterie, and the consequence of this process for strategic communications by political parties and interest groups. A large literature, particularly within Europe, has also studied the changing structure of the news industry, notably the economic basis of the newspaper industry and the legal structure regulating broadcasting and the press. Comparative studies have also commonly analyzed the news culture, especially the values that journalists, broadcasters, and editors employ as 'gatekeepers' in deciding 'what's news,' as well as the organizational structure of newsrooms. Recent work within this area has generated a growing body of research on the rise of new communication and information technologies, and use of the Internet by parties, new social movements, and the news media.

One of the major challenges in this subfield is to widen the scope of comparative research on a systematic basis so that we move from studies of the structure of the political communication process within particular nations to conceptual typologies, broader theories, and empirical generalizations that can be tested across different types of societies. Much of this work has traditionally focused upon postindustrial nations, particularly the USA and Western Europe, although in the 1990s increased attention has been paid to new democracies in Central and Eastern Europe, Latin America, and Asia, as well as, to a lesser extent, to the role of the news media in authoritarian regimes like Burma, China, and Cuba where the role of the free press and opposition movements remain severely restricted.

2. *Contents*

Another related mainstream research tradition has examined the content of the messages that are produced by this process, such as the amount and tone of political reporting presented in television news, the partisan balance in the press, the coverage of election campaigns and particular events, the agenda-setting reporting of policy issues, and the representation of social minorities in the news media. The most common forms of comparison are between different media, for example, differences in the messages conveyed during an election campaign in paid political advertisements, party press releases, newspaper columns, and TV news stories.

Other forms of comparison examine trends over time, such as the coverage of political scandals or social minorities in recent decades. More rarely, collaborative teams have attempted comparison

across nations, for example, concerning a specific period or event in selected major newspapers. Work in this tradition has drawn largely upon systematic forms of content analysis of a representative random sample of stories among different media, although alternative qualitative techniques for deconstructing textual and visual messages are also common. Moving beyond systematic description, which is valuable but limited, the main challenges in this area are to relate the content of the messages to either the production process (to examine their possible causes) or to their potential impact (to understand their effects).

3. *Effects*

Lastly perhaps the largest body of research, certainly in the USA, has focused at individual level on the study of the potential effects of exposure to different types of political communication messages. The most common method has been to draw upon cross-sectional or panel representative surveys or, more rarely, experimental methods. The key issues have focused upon analyzing the potential impact of exposure to different type of mediated messages (such as watching an advertisement or news story) upon either political knowledge and opinions (such as awareness about an issue, civic knowledge, or recognition of political leaders), political attitudes and values (such as support for a particular party or issue), and political behavior (such as voting turnout). Most work has focused upon the impact of the messages on the mass public or particular subgroups, like women or undecided voters, although some studies have also analyzed the effects upon middle-level elites involved in the policymaking process.

The primary challenges are threefold: (a) to expand generalizations beyond the USA, to see how far they continue to hold up within very different contexts; (b) to move beyond cross-sectional surveys, which cannot determine issues of causality, toward more dynamic designs such as panel surveys and pre–post experimental designs; and (c) to link studies of the individual-level analysis of effects to both what we know about the structure of the news industry and the contents of the messages, generating multilevel and multinational research designs. This work has made considerable progress as the study of political communications has moved increasingly into the mainstream within political science since the 1980s, but nevertheless the subfield remains predominantly US and European, and the process of internationalization in what is increasingly a global society is only starting to produce more systematic cross-national and multilevel research.

Political communication has, therefore, always been central to the electoral and policymaking process but since the 1990s certain important structural developments have fundamentally altered this process, particularly postwar trends in the mass media moving from the traditional world of newspapers, radio, and television broadcasting toward the Internet. The rest of this article outlines alternative interpretations of the nature of these trends and reflects on their consequences for the process of socioeconomic and political development around the globe.

4. *Postwar Trends in Newspapers*

Concern about traditional standards of journalism has been fuelled by major changes in the newspaper industry during the postwar era. In the USA the daily press has experienced dwindling readership and sales, especially among the younger generation, a loss of advertising market share to the electronic media, and growing concentration of ownership in larger multiple newspaper chains or a few multimedia conglomerates (McManus 1994, Baker and Dessart 1998). All these developments have had a major impact upon the profitability and economic viability of the print sector in the USA, particularly for smaller outlets (De Bens and Ostbye 1998).

Yet although the demise of newspapers has been widely predicted for decades, we should not underestimate their continued popularity and technological adaptation to new forms of production and distribution. If we compare postwar trends in circulation, controlling for population growth, the evidence shows that sales of the daily press in most postindustrial societies has not been affected by the growing availability of electronic media. The long-term trend in newspaper sales in OECD countries, per 1,000 population, has proved fairly stable. Average circulation in all OECD states was 271 newspapers per 1,000 population in 1950, rising modestly in 1980, before subsiding slightly to 263 per 1,000 in 1996 (UNESCO). In postindustrial societies, despite the massive surge in the availability of television since the 1950s, about one quarter of the population continues to buy a daily newspaper, and readership figures are even higher. The electronic media have therefore increased the choice and diversity of news outlets and formats, but at the same time they have not killed sales of the printed press. Given growing educational levels and affluence characteristic of postindustrial societies, consumption of news has not proved a zero sum game.

Moreover there are considerable variations among societies in use of the traditional mass media. Figure 2 illustrates the penetration of newspapers and TV sets per 1,000 population worldwide. The pattern confirms how far newspaper and TV penetration levels are associated with basic patterns of socioeconomic development; use of the mass media is related to a more affluent and educated population with greater leisure

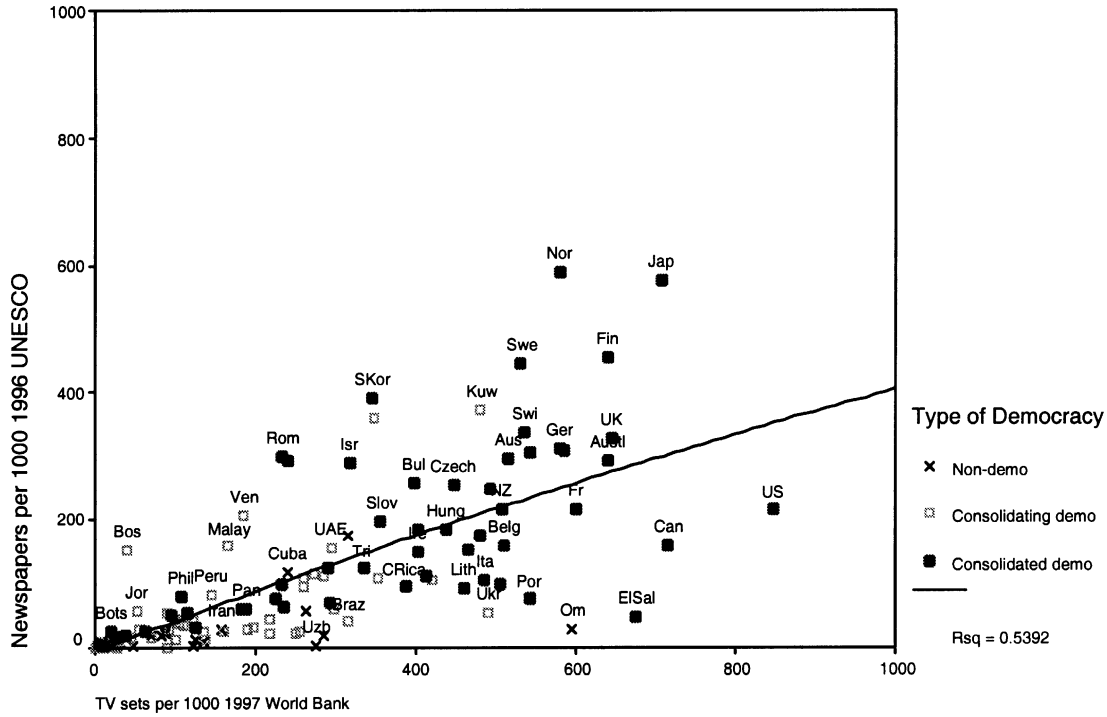

Figure 2
Worldwide distribution of newspapers and TV sets, mid-1990s. (Sources: Newspapers: UNESCO Statistical Yearbook 1997; TV sets: World Bank 2000; Type of democracy: Based on the Freedom House classification of political rights and civil liberties.)

time. Postindustrial societies tend to be the heaviest consumers of the mass media; nevertheless the pattern also shows considerable variations among these nations. Some outliers like the USA and Canada prove far more television-oriented than average (along with Southern Mediterranean Europe and much of Latin America), while others nations like Norway and Sweden (along with much of Northern, Central, and Eastern Europe) remain more newspaper-oriented.

It remains unclear whether systematic trends in the newspaper industry have changed traditional journalism, producing an increased focus on crime, sex, and entertainment, as is assumed by some critics. What seems equally plausible across OECD countries is an expansion of both lowbrow and highbrow news media in recent decades, representing a diversification of the market. A recent review of the comparative literature by Kees Brants concluded that the few available content analysis studies provide an ambiguous and sometimes contradictory picture of the growth of 'infotainment' news in different countries, rather than showing a uniform pattern: 'Where for the European countries as a whole we might see a slight tendency towards the popularization of news, there is little evidence that politicians and politics are more dramatically personalized and sensationalized than before'(Brants 1998).

Brants found that the available content analysis shows a mixed picture of the growth of 'infotainment' news in different European countries, rather than a uniform trend. Frank Esser concluded that there were marked contrasts between Germany, the UK, and the USA in the popularity of tabloid news, and that the nature and degree of competition in a particular media market is the decisive factor explaining the degree of tabloidization (Esser 1999). Moreover systematic research on long-term trends in British newspapers from 1952–97 found that the amount of political coverage in the tabloid sector had not declined over time, as many critics assume. Instead, the tabloid press in the UK has expanded its coverage of entertainment but also maintained its political news over this period (McLachlan and Golding 1999, McLachlan 1999).

Therefore across all postindustrial societies newspaper circulation levels have remained largely stable during the postwar era, yet at the same time the range of papers published in these states has contracted. The number of daily newspapers published in OECD nations fell on average by 15 percent during the postwar era, from 160 per country in 1960 to 130 in

1996, producing greater concentration of readership in fewer outlets. Many countries have introduced measures to maintain press diversity, on the assumption that we need diverse outlets for an effective civic forum. Antitrust regulations have attempted to insure competition in the ownership of the press, for example, limiting the proportion of cross-media ownership by a single company, administered by fair trade bodies like the British Monopolies and Merger Commission or the German Federal Cartel Office. Other societies like Sweden and Norway have used press subsidies as a policy instrument to protect the financial viability of the more vulnerable sectors of the press (McQuail and Siune 1998). Countries with provincial and localized newspaper markets like the USA and Germany proved particularly prone to media mergers and acquisitions, reducing pluralism and competition in many cities. Papers in smaller countries like Austria and Belgium also often experienced takeovers or closure because of a limited domestic market and imports from neighboring states with a shared common language.

Concentration of ownership in the hands of a few multinational corporations with multimedia empires has become increasingly common, notably Rupert Murdoch's News International, and the vast holdings of Bertelsmann in Germany, or Fininvest in Italy (Tunstall and Palmer 1991, Smith 1991, Sanchez-Tabernero 1993). Hence Rupert Murdoch, who started with two small Australian newspapers, built an empire in News Corp. that includes 20th Century Fox films, the Fox TV network, a number of US stations, 50 percent ownership of Sky TV, a majority interest in the STAR Asian satellite, ownership of *The Sun* and *The Times* in the UK, additional television stations in Latin America, and the book publisher HarperCollins, as well as investments in Internet companies. In the USA, Time Warner's purchase of Turner Broadcasting Systems Inc (including CNN) in 1996 created the largest media firm in the world with strong print, cable, and programming divisions. Walt Disney Company's acquisition of Capital Cities/ABC Inc for $19 billion in 1995 created the second largest media conglomerate with movie, television, cable, and Internet interests, although the purchase proved costly since ABC's profitability moved sharply into the red four years after acquisition. Conrad Black's acquisition of Southam in Canada in 1996 gave his company, Hollinger Inc, control of two-thirds of the newspapers in that country.

What will be the consequences of these developments? Many commentators like Ben Bagdikian (1997, see also Bogart 1995) fear that media mergers have concentrated excessive control in the hands of a few multinational corporations, who remain unaccountable to the public, and that only greater economic competition can change this situation. Recognition of this problem has arisen from the understanding that economic controls can constrain the media just as significantly as political controls. There is nothing new about this concern, which was often expressed during the interwar era of the press barons when proprietors like Beaverbrook and Rothermere actively intervened to further their political ambitions. Yet others like Robert Picard (1988, 1998) remain more sanguine about recent developments, arguing that we need to distinguish between concentration defined by considering the number of media outlets held by dominant firms and concentration defined by dominance in a clear geographical market. It is the latter—which can harm consumers by producing fewer choices, poorer services, and higher prices—which is important for the availability of alternative sources of political information in a democracy.

Monopolies in the local market for ideas can be harmful for pluralism. Nevertheless it must be recognized that we need to look beyond any single media sector to establish the harmful political effects of concentration, since consumers use and have access to multiple sources of news and information, from newspapers to radio, television, and the Internet. Moreover the trends toward greater concentration are not universal, as some OECD countries have seen a significant expansion in the circulation and range of daily newspapers being published in the postwar era, particularly states like Mexico and Greece where educational and literacy rates have been rising sharply, as well as more modest growth evident in newer democracies like Hungary, the Czech Republic, and South Korea.

5. *Trends in Broadcasting*

Just as there are serious concerns about the future of newspapers, so many believe that in recent decades traditional standards of television news and public affairs have come under threat from technological and economic developments. The critical factors transforming broadcasting include the proliferation of channels on terrestrial, cable, satellite, digital, and now broadband services, fragmenting the mass audience; the crisis of identity and funding facing public service television, which once enjoyed a monopoly throughout most of Europe, following the rise of myriad commercial competitors; and lastly the more recent technological convergence with the digitization and compression of images, sounds, and data which has produced a new multimedia environment breaking down the traditional boundaries between telecommunications, the audiovisual industries, and computers.

These trends have affected all OECD countries to different degrees although their impact and timing has been strongly mediated by the existing communications landscape. In the Thatcherite 1980s, deregulation and privatization had the most profound influence on public service broadcasters throughout Western

Europe (Achille and Ibanez Bueno 1994). Following the fall of the Berlin Wall, the transition to democracy in the early 1990s produced an even more radical jolt to public television in Central and Eastern Europe (O'Neill 1998). Meanwhile in the USA the long-standing rule of the three major networks experienced an equivalent coup d'etat, cut down in the 1980s by myriad competitors on cable and satellite (Kimball 1994). Despite the long-standing contrasts between the commercially dominant major networks in the USA and public television in Europe, in recent years both have faced a strikingly similar multiplication of outlets and fragmentation of television audiences, raising new concerns about the standards of programming.

Despite the deluge of commercial alternatives, the public channels remain popular; on average across all OECD states public channels maintain a 42 percent share of the television audience. This varies substantially, however, by country. Public television RAI1 and TVE1 remain market leaders in Italy and Spain, while NRK in Norway and SVT in Sweden had most of the best-rated shows in their countries. In contrast public TV has a far smaller share in some other societies such as PBS (3 percent) in the United States and NHK (18 percent) in Japan (Mediametrie 1999). OECD states can be classified into three major types: those that remain predominately *public* systems (based on an audience share of public channels of 60 percent and above), *mixed* systems (with a public share of 40–59 percent) and predominately *private* systems (with a public share of less than 40 percent). Where there is comparable data, only three OECD nations can be categorized as predominately public (Austria, Denmark and Hungary), 11 represent mixed systems, while 10 can be classified as predominately private systems. Therefore although there has been growing transatlantic convergence over the years, nevertheless broadcasting systems continue to bear the distinct imprint of their origins with radio in the 1920s, and public television continues to remain popular in many countries, rather than being swamped by commercial services.

6. *The Rise of the Internet*

Since the early 1990s the most important change to the political communication process has occurred through the rise of the Internet, particularly in postindustrial societies that are at the forefront of the information society such as the USA, Australia, and Sweden. Networked computing and computer-mediated email have existed for the scientific elite since the early 1960s but the number of users was too small to monitor through mass surveys. The key historic development transforming the Internet into the world's favorite virtual reference library, post office, and shopping mall was a series of rapid innovations: the birth of the World Wide Web (in 1990) and the launch of popular browsers to access materials including those by Mosaic (1993), Netscape Navigator (1994), and Microsoft Internet Explorer (1995) (see Berners-Lee 2000). Subsequent technological applications, like the easy transfer of .mp3 music files and video formats, and WAP-enabled digital telephony, while representing important innovations, cannot yet claim to have had an impact equal to the basic invention of point-and-click browsers.

As yet no official government statistics on the online population are collected by international agencies like UNESCO and the International Telecommunications Union (ITU), although many indirect measures of technological diffusion are available, including investment in scientific Research and Development, the spread of computer hardware, and the rate of telephone density. The most comprehensive worldwide guide estimating the size of the online population is provided by NUA. This organization regularly monitors and collates survey results conducted by different market research companies in each nation. The surveys ask a representative sample of the public in each country about use of the Internet from home, work, or elsewhere during the previous three months. NUA's database 'How Many Online' (http://www.nua.ie) currently collects data from 179 countries, covering 5.7 billion people.

The NUA evidence highlights the dramatic rise in popularity of the Internet in recent years: between 1995 and 2000 the total number of Internet users surged from about 26 to 377 million worldwide, an explosive jump within the space of a few years. The Internet became a truly global phenomenon as more and more users came online from around the world and the proportion of Americans in the online community dropped from 70 to 40 percent in 1995–2000. Despite this remarkable expansion, about one in 20 of the world's population was online in 2000, with highly uneven diffusion globally.

According to NUA estimates, in Spring 2000 Scandinavia and North America led the world in rates of Internet penetration, with one-third or more of the population online, followed by Western Europe, with about one in 10 online. Central and Eastern Europe, Asia, the Middle East, and South America fall below the world average, all with less than one in 20 online, while minimal diffusion is evident in Sub-Saharan Africa, with only 36 users per 1,000 people. In terms of levels of human development, there are stark contrasts between rich and poor nations. Most of the world's online community (87 percent) lives in highly developed nations, as measured by the UNDP index (see UNDP 1999). In comparison, the 35 societies classified by the UNDP with low levels of human development, like Nigeria, Bangladesh, and Uganda, contained only 5 percent of the online population, although home to half a billion people.

A finer-grained comparison of countries ranked by the online population reveals a pattern of widespread adoption in four clusters of societies:

(a) Throughout the smaller Nordic social democratic welfare states, especially Sweden, Norway, Iceland, and Finland;

(b) In larger Anglo-American and English-speaking nations including the USA, Canada, Australia, and the UK;

(c) In the Asian 'tiger' economies of Singapore, South Korea, and Taiwan, as well as Japan; and lastly,

(d) In a few smaller European nations with above-average Internet use such as the Netherlands, Belgium, Switzerland, Slovenia, and Estonia.

At the bottom of the national rankings, with less that 0.5 percent of the population online, few Internet users are found throughout most of the poorer countries of sub-Saharan Africa (with the exception of South Africa), as well as in many states in central Europe, the Middle East, Asia, and Latin America.

Yet at the same time if technological diffusion can be achieved in poorer societies, and it is a big 'if,' then many observers hope that the Internet will provide multiple opportunities for socioeconomic and democratic development. Digital networks have the potential to broaden and enhance access to information and communications for remote rural areas and poorer neighborhoods, to strengthen the process of democratization under transitional regimes, and to ameliorate the endemic problems of poverty in the developing world. With connectivity as the umbilical cord, enthusiasts hope that the Internet will eventually serve multiple functions as the world's favorite public library, school classroom, and medical database; post office and telephone; marketplace and shopping mall; channel for entertainment, culture, and music; daily news resource for headlines, stocks, and weather; and heterogeneous global public sphere. In the heady words of the G-8 Okinawa Charter:

> Our vision of an information society is one that better enables people to fulfill their potential and realize their aspirations. To this end we must ensure that IT serves the mutually supportive goals of creating sustainable economic growth, enhancing the public welfare, and fostering social cohesion, and work to fully realize its potential to strengthen democracy, increase transparency and accountability in governance, promote human rights, enhance cultural diversity, and to foster international peace and stability." (G-8 Okinawa Charter on Global Information Society 23 July 2000. http://www.g8kyushu-okinawa.go.jp/w/documents/it1.html)

The Internet may allow societies to leapfrog stages of technological and industrial development. On the production side, if Bangalore companies can write software code for IBM or Microsoft, and if Costa Rica can manufacture chips for Intel, then potentially entrepreneurs can offer similar services from Malaysia, Brazil, and South Africa. The Internet encourages market globalization: small craft industries and the tourism industry in Bali or the Maldives can deal directly with customers and holidaymakers in New York and London, irrespective of distance, the costs of advertising, and the intermediate distribution chains of travel agents and retail businesses (International Telecommunications Union 1999, p. 7, Dugger 2000).

The Internet also offers promise for the delivery of basic social services like education and health information across the globe, a function that may be particularly important for middle-level professionals serving their broader community (Hayward 1995, Wresch 1996). Potentially, local teachers or community officials connected to the digital world in Lagos, Beijing, or Calcutta can access the same electronic journals, books, and databases as students at the Sorbonne, Oxford, or Harvard. Distance learning can widen access to training and education, via open universities in India, Africa, and Thailand, and language websites for schools (Arunachalam 1999). Net works of hospitals and healthcare professionals in the Ukraine, Mozambique, and Stockholm can pool expertise and knowledge about the latest research on AIDS. Peasant farmers using village community centers can learn about storm warnings and market prices for their crops, along with employment opportunities in local towns. Where peripheral regions lack access to the traditional media, the convergence of communication technologies mean that potentially the Internet can deliver virtual local newspapers, streaming radio and television video, as well as other services.

Numerous examples can be cited to show the potential of digital technologies for fostering new opportunities for development in societies around the world (World Economic Forum 2000). Many South East Asian nations seek to emulate the Japanese model of development in the postwar era of reconstruction, and the knowledge-based economy in Singapore, South Korea, and Taiwan. In Malaysia, for example, the Multimedia Super Corridor has been developed to bring investment from telecommunications, multimedia, and electronics companies, and the production of silicon wafers and software. The corridor has attracted major players such as Microsoft, Sun Systems, and NTT (Japanese telecom). Under the 'Vision 2020' plan, Malaysia now boasts cellular telephone penetration rates of one in every 10 people, more and more wired schools, and 21 Internet hosts per 1,000 people. Revenue generated by the production of information and communication technology goods, like office equipment, telecommunications, and consumer audiovisuals, shows that the USA leads the world but many Asian countries are close rivals, including Japan (second), Korea (third), Singapore (fourth), Taiwan (seventh) and Malaysia (eighth) (OECD 2000, p. 24, Table 2).

Southern India is most often cited as an important area of software development, producing an estimated $3.8 billion in revenues, with this figure doubling in the

last years of the 1990s. Over one-half of India's software services are exported to the USA (OECD 2000). The Bangalore area has attracted inward investment from many major corporations, not least from the diaspora of the Asian dot.com entrepreneurs thriving in California's Silicon Valley and Cambridge's Technology Park (for a discussion see Yourdon 1996). In rural Bangladesh many isolated communities lack landline telephones. An innovative program by Grameen Telecom supplies cellular mobile phones to village women, who rent calls in their community to repay the loan and sustain thriving micro enterprises. (By Spring 2000, over 1,000 phones had been provided, serving 65,000 people, and the eventual target is 40,000 phones. See Richardson 2000, see also http://www.grameenphone.com.). With this service, local communities benefit by direct links to job, weather, and health information, as well as more efficient markets for their produce. Village telecom centers are being developed with email and fax services, along with computer literacy projects in selected schools.

In Central and Eastern Europe, Slovenia, Estonia, and Slovakia have made great strides in moving their populations online, moving well ahead of Portugal, Greece, and Austria in levels of connectivity. Hungary's ambitious Schoolnet program has allowed students in two-thirds of all secondary schools to browse the Web from their classrooms, with extensive teaching resources, interactive discussion forums, events, and competitions (see http://www.SuliNet.hu). In the Baltic, the Estonian government has provided public access points for the Internet throughout the country, using schools, post offices, community centers, libraries, police stations, and health clinics. The program has been highly successful; today more than one in 10 Estonians is online, with personal computer ownership well above average for Central and Eastern Europe (UNDP 1999, p. 64).

Progress has been slower in Africa, but nevertheless plans have been announced by Global Crossing, Lucent Technologies, and Africa One for an ambitious $1.9 billion project to link up the whole continent by 2002 through a high-speed underwater fiber optic cable, with interior countries connected through terrestrial cables, microwave, or satellite facilities, overcoming many of the current problems of the inadequate telephony infrastructure (Wired News 2000. For further details see www.AfricaOne.com). Given a high-speed backbone, and market liberalization of telecommunication services, African nations may also be able to 'leapfrog' stages of industrialization through new technology by investing in fully digitized telecommunications networks rather than outdated analog-based systems. Cellular telephony is rapidly expanding as an alternative to conventional network services; the number of subscribers in the OECD region reached almost one-quarter of the population in 1998 (OECD 2000, p. 81). This growth has had even greater impact in the developing world. In postindustrial economies there were 20 times as many mobile phones in 1998 as there were in 1990, and in developing economies there were 160 times as many, an astonishing rise (World Bank 2000, p. 299). Over a third of all telephone subscribers in Cote d'Ivoire, Cambodia, and Paraguay, for instance, are now connected via mobiles, a far higher proportion than in the USA (International Telecommunications Union 1999).

7. *The Consequences for Democracy and Democratization*

What will be the political consequences of changes in the traditional news media and the subsequent rise of the Internet? In many ways it remains far too early to say (for a detailed discussion, see Norris 2001). The Internet has generated deeply contested alternative visions about the future. Luddites fear for the worse, but technophiles hope for the better.

The most positive perspective is held by cyber-optimists who emphasize the Panglossian possibilities of the Internet for the involvement of ordinary citizens in direct democracy. In its more utopian manifestations, this view has been dubbed 'technoromanticism' (Coyne 1999), expressed in earlier eras in response to Samuel Morse's electric telegraph, Alexander Graham Bell's telephone, and Guglielmo Marconi's wireless radio (see, e.g., Rhodes 1999, Pursell 1995, Lubar 1993). In this account, digital technologies hold promise as a mechanism facilitating alternative channels of civic engagement such as political chatrooms, electronic voting in general elections and for referenda issues, and the mobilization of virtual communities, revitalizing levels of mass participation in public affairs (see Barber 1998). The use of the Internet by groups and social movements is often believed to exemplify digital politics.

The more utopian visions of the Internet suggest a future society in which virtually unlimited qualities of information become available, civic society flourishes, government decision making becomes more open and transparent, and nation-state borders are eroded as people build virtual communities for work, learning, and leisure spanning traditional boundaries of time and place. Although still in its adolescence, the core transformative capacities of the Internet include its potential for radically shrinking communications and information costs, maximizing speed, broadening reach, and eradicating distance. Compared with radio, television, and newspapers, controlled by editors and broadcasters, the World Wide Web facilitates a virtually unlimited choice of information and communication one-to-one (e.g., via email), one-to-many (e.g., via a personal home page or electronic conference), many-to-one (e.g., via an electronic poll) and, perhaps most importantly, many-to-many (e.g., an online chatroom), with a minimal role for gatekeepers or

government censors (see, e.g., Shapiro 1999, Weare 2000). Internet messages have the capacity to flow further, faster, and with fewer intermediaries.

As a result many hope that recent developments, especially the spread of new information and communication technologies, will serve to undermine authoritarian regimes, creating a 'dictators' dilemma' in countries like Burma, China, and Cuba (Drake et al.). Leaders in these nations want to facilitate economic development through the Internet and yet at the same time they seek to restrict political access. It is also believed that the Internet will have major consequences for electoral democracies, countries like Russia, Taiwan, and Mexico that are seeking to consolidate democratic political institutions.

Yet as the Internet evolves, a darker vision has been articulated among cyber-pessimists, who regard digital technology as a Pandora's box unleashing new inequalities of power and wealth, reinforcing deeper divisions between the information rich and poor, the tuned-in and the tuned-out, the activists and the disengaged. This account stresses that the global and social divides already discussed mean that Internet politics will disproportionately benefit the elite (see, e.g., Golding 1996, 1998, 2000). In this perspective, despite the potential for technological innovations, traditional interests and established authorities have the capacity to reassert their control in the virtual political sphere, just as traditional multinational corporations have the ability to reestablish their predominance in the world of e-commerce (see, e.g., McChesney 1999, pp. 182–5). In this view, so far the potential of the Internet has failed to have a dramatic impact on the practical reality of 'politics as usual,' for good or ill, even in countries at the forefront of digital technologies (Margolis and Resnick 2000). There are fears that continued inequalities in the spread of new technologies will exacerbate the traditional North–South divide evident in access to the traditional news media like newspapers, radio, and television. As such, while some developing countries may manage to leapfrog earlier technologies in the race towards the information society, other poorer societies may drop even further behind.

The debate between the cyber-optimists and pessimists continues and as the Internet continues to evolve over the first decade of the twenty-first century, then the impact of new technologies will become evident, whether for good (as some hope) or ill (as others fear). What is certainly clear is that political communications via the old world of newspapers and television is in the process of fundamental change and this process holds both threats and promises for the future of socioeconomic and political development.

See also: Advertising: Effects; Advertising: General; Broadcasting: General; Broadcasting: Regulation; Campaigning: Political; Electronic Democracy; Journalism and Journalists; Media Effects; Media, Uses of; News: General; Political Advertising; Political Culture; Political Discourse; Political Money and Party Finance

Bibliography

Achille Y, Ibanez Bueno J 1994 *Les télévisions publiques en quête d'avenir*. Presses Universitaires de Grenoble, Grenoble, France

Arunachalam S 1999 Information and knowledge in the age of electronic communication: A developing country perspective. *Journal of Information Science* **25**(6): 465–76

Bagdikian B 1997 *The Media Monopoly*, 5th edn. Beacon Press, Boston

Baker W F, Dessart G 1998 *Down the Tube*. Basic Books, New York

Barber B R 1998 Three scenarios for the future of technology and strong democracy. *Political Science Quarterly* **113**(4): 573–89

Berners-Lee T 2000 *Weaving the Web*. Harper Business, New York

Bogart L 1995 *Commercial Culture*

Brants K 1998 Who's afraid of infotainment? *European Journal of Communication* **13**(3): 315–35

Coyne R 1999 *Technoromanticism: Digital Narrative, Holism, and the Romance of the Real*. MIT Press, Cambridge, MA

De Bens E, Ostbye H 1998 The European newspaper market. In: McQuail D, Siune K (eds.) *Media Policy: Convergence, Concentration and Commerce*. Sage, London

Drake W, Kalathil S, Boas T C Dictatorships in the digital age: Some considerations on the Internet in China and Cuba. *iMP: The Magazine on Information Impacts*

Dugger C W 2000 Connecting rural India to the world. *New York Times* 28 May, http://www.nytimes.com/library/tech/yr/mo/biztech/articles/28india.html

Esser F 1999 Tabloidization of news: A comparative analysis of Anglo-American and German press journalism. *European Journal of Communication* **14**(3): 291–324

Golding P 1996 World wide wedge: Division and contradiction in the global information infrastructure. *Monthly Review* **48**(3): 70–85

Golding P 1998 Global village or cultural pillage? The unequal inheritance of the communication revolution. In: McChesney R W, Meiksins Wood E, Foster J B (eds.) *Capitalism and the Information Age: The Political Economy of the Global Communication Revolution*. Monthly Review Press, New York

Golding P 2000 Information and communications technologies and the sociology of the future. *Sociology* **34**(1): 165–84

Hayward T 1995 *Info-rich, Info-poor: Access and Exchange in the Global Information Society*. Saur, London

International Telecommunications Union (ITU) 1999 *Challenges to the Network: Internet for Development*. ITU, Geneva, Switzerland

Kimball P 1994 *Downsizing The News: Network Cutbacks in the Nation's Capital*. Woodrow Wilson Center Press, Washington, DC

Lubar S 1993 *Infoculture: The Smithsonian Book of Information Age Inventions*. Houghton Mifflin, Boston

Margolis M, Resnick D 2000 *Politics as Usual: The Cyberspace 'Revolution.'* Sage, Thousand Oaks, CA

McChesney R W 1999 *Rich Media, Poor Democracy*. University of Illinois Press, Urbana, IL

McLachlan S 1999 Who's afraid of the news bunny? The changing face of the television evening news broadcast.

Information and Democracy Research Project: Working Paper No. 3. Loughborough University, Loughborough, UK
McLachlan S, Golding P 1999 Tabloidization in the British press: A quantitative investigation into changes within British newspapers from 1952–1997. *Communication Research Centre Working Paper #9*. Loughborough University, Loughborough, UK
McManus J H 1994 *Market Driven Journalism*. Sage, London
McQuail D, Siune K 1998 *Media Policy: Convergence, Concentration and Commerce*. Sage, London
Mediametrie 1999 *One Television Year in the World. 1998*. Mediametrie, Paris
Norris P 2001 *Digital Divide: Civic Engagement, Information Poverty and the Internet Worldwide*. Cambridge University Press, Cambridge, UK
O'Neill P 1998 *Post-communism and the Media in Eastern Europe*. Frank Cass, London
Organization for Economic Cooperation and Development (OECD) 2000 *OECD Information and Technology Outlook*. OECD, Paris
Picard R G 1988 *Press Concentration and Monopoly: New Perspectives on Newspaper Ownership and Operation*. Ablex, Norwood, NJ
Picard R G 1998 Media concentration, economics, and regulation. In: Graber D, McQuail D, Norris P (eds.) *The Politics of News: The News of Politics*. CQ Press, Washington, DC
Pursell C W 1995 *The Machine in America: A Social History of Technology*. The Johns Hopkins University Press, Baltimore, MD
Rhodes R (ed.) 1999 *Visions of Technology*. Simon & Schuster, New York
Richardson D 2000 *Grameen's Telecom's Village Phone Programme in Rural Bangladesh*. Telecommons Development Group, ON, http://www.telecommons.com
Sanchez-Tabernero A 1993 *Media Concentration in Europe: Commercial Enterprises and the Public Interest*. John Libbey, London
Shapiro A L 1999 *The Control Revolution: How the Internet is Putting Individuals in Charge and Changing the World we Know*. Public Affairs, New York
Smith A 1991 *The Age of Behemoths: The Globalization of Mass Media Firms*. Priority Press, New York
Tunstall J, Palmer M 1991 *Media Moguls*. Routledge, London
United Nations Development Programme (UNDP) 1999 *Human Development Report 1999*. UNDP/Oxford, New York
UNESCO 1964 *UNESCO Statistical Yearbook*. UNESCO, Paris
Weare C 2000 Technology and politics: Linking the Internet to changes in democratic governance. Paper presented at the International Political Science Association World Congress, Quebec, August
Wired News 2000 Africa One project targets 2002. 5 June
World Bank 2000 *World Development Indicators* 2000. www.worldbank.org/data
World Economic Forum 2000 *From the Global Digital Divide to the Global Digital Opportunity: Proposals submitted to the G-8 Kyushu-Okinawa Summit 2000*. www.ceip.org
Wresch W 1996 *Disconnected: Haves and Have-nots in the Information Age*. Rutgers University Press, New Brunswick, NJ
Yourdon E 1996 *The Rise and Resurrection of the American Programmer*. Prentice-Hall, New York

P. Norris

Political Culture

At the broadest level, political culture refers to the meanings attributed to politics. All aspects of meaning, subjective and intersubjective, related to politics are included in this broad perspective—including the definition of what constitutes or determines the boundaries of the political. A central concern is the role of political culture in the transformation of power into authority, and in challenges to authority. Political culture constitutes the socially constructed and tenuously shared meanings which endow or challenge legitimacy in the political institutions, offices, and procedures of a polity. At the cognitive level, political culture defines and explains or questions the rules of the political game, and at the normative level it justifies or denies the authority of those who play it and of the institutions within which the players interact. Political culture is employed to establish or contest the legitimate boundaries of collective identity. It defines a 'we' against a 'them' and marks the centrality and marginality of various groups within the social unit. At any given historical period, a political culture will likely be characterized by a fairly limited repertoire of particularly salient themes that express the values and goals most sacred to the collectivity through key symbols, myths, rhetorical strategies, and rituals accessible through the analysis of discourse. Most contemporary approaches to political culture consider a consensus on values unlikely to occur and unnecessary for the viability of polities. The public need only resonate to common symbols to which they may ascribe radically different meanings. Even in the best of circumstances the shared meanings of political culture are highly tenuous. Like a fragile spider's web, its symbolic themes are stretched in multiple directions through contested interpretations. Political culture, for all its fragility, is essential for the survival of society and civilization.

1. The Politics of Culture and the Culture of Politics

Contemporary approaches to political culture derive from two Western intellectual traditions. One emphasizes politics—viewing it as a neutral adjudicator of conflicts including those in the realm of culture. The second tradition, descending from Montesquieu and de Toqueville, views politics and governance as cultural constructs. The first tradition is expressed by Harold Lasswell's (1960) inquiry into who gets what, when, and how. The second is illustrated by Michel Foucault's (1966) concern with multiple ways in which power is implicated in the constitution of all areas of social life. In this second tradition the definition of the political and the location of its boundaries is framed as a fundamental cultural question. Political culture

encompasses both approaches—the politics of culture, and the culture of politics. Although the term 'political culture' is used primarily by political scientists and historians (Gendzel 1997), interest in the relationship between politics and culture—including the investigation of the extent to which each constitutes an autonomous realm—is of interest in fields in which the practitioners use the term less frequently; for example, anthropology, sociology (Swidler 1984), and cultural studies.

2. *Components of Political Culture*

Symbols are the threads from which the web of culture is spun. Sherry Ortner (1973) distinguishes between elaborating and summarizing symbols. The latter, deriving their power primarily from their emotional resonance, tend to be sacred (in the broad sense)—acting as objects of reverence. The national flag is treated with such reverence as to constitute a fetish in American civil religion. The Western Wall in Jerusalem, a place of sacred centrality in Israeli political culture, and the Black Madonna of Czestochowa in Poland constitute powerful summarizing symbols.

Elaborating symbols work in the opposite direction. They provide vehicles for sorting out complex and undifferentiated feelings and ideas, thereby making them comprehensible, communicable, and translatable into orderly action. Given their cognitive nature, they play a more dominant role in ideology. Root metaphors have conceptual elaborating power. Their value derives from providing categories for comprehending order in the world. 'Exile' from and 'return' to the 'Holy Land' provide the root metaphor of Zionism—the dominant discourse (master narrative) of Israeli political culture (Aronoff 1989). Key scenarios are specific guides that provide maps for socially valued action. The Horatio Alger myth is a such a guide to achieve the idealized American success story.

The myths that endow power with legitimacy are challenged constantly by iconoclasts who call them into question and undermine them, thereby tearing the web of cultural meanings. Old truths are constantly reinterpreted as political realities change, and new cultural traditions are invented by myth-makers who repair the web and drape emergent groups with the mantle of authority. Political myths are narratives which rely on the nonrational for their emotive power. Myths present more lucid and compelling images than do more abstract principles. The more compelling the image, the more powerful the myth's appeal. Because a myth makes hidden connections between disparate ideas based on unconscious premises, public discourse in mythical form is relatively immune from rational criticism that points out its logical contradictions, or evidence which disconfirms it. When such a pattern of discourse dominates a polity, it is usually a powerful force constraining political change. Yet, political myths also play crucial roles in times of crisis and can be used—especially in rituals through which myths are enacted—to achieve even revolutionary change (Kertzer 1988). Solidarity and the Catholic Church utilized the Pope's visit to Poland effectively to orchestrate public rites and political rituals which undermined the hegemony of the Polish People's Republic (Kubik 1994). Political rhetoric, as in the case of the Pope's sermons, also constitutes an important component of political culture.

Collective memory and amnesia are molded and expressed through invented traditions which are revitalized cultural forms, such as nationalism, portrayed as continuations of more ancient customs. These elements of political culture articulate a society's self-definition and destiny. Political cultures are effectively reified when the 'reality' they define becomes taken for granted by their publics as natural or divinely ordained. This 'reality' explains and justifies (or challenges) ideologically the allocation of resources, prestige, and power in the polity. Ideologies are more concrete and rationalized versions of the general political culture articulated by groups to give legitimacy to their identities and to justify their goals. The endurance of dominant (hegemonic) political cultures at any point in time varies depending on their effectiveness in providing models for dealing with societal problems. Political cultures are dynamic because they are embedded in changing historical, social, economic, cultural and political environments that shape them, and which are in turn shaped by them. The relationship between these variables is dialectical, making the determination of causality highly problematic.

Political cultural divisions can be understood as competing temporal and spatial perceptions. Contested interpretations of myths of origin and aspects of history forgotten and remembered selectively reveal contrasting imaginings of political community. Analysis of the relationship between the cultural construction of past, present, and future, and political ideology reveals the link between contested collective memories and competing visions of nationhood. Temporal and spatial perceptions vary within and across political cultures. Social and political boundaries are related to, but not always identical to, actual geographic frontiers of the state (Lustick 1993). Alterations of the physical borders of a state, filtered through temporal cultural frames, shape citizens' senses of personal and collective security, which in turn influence their political sentiments and positions.

3. *Disciplinary and Interdisciplinary Approaches*

At the apex of the post-World War II behavioral revolution, the first systematic multinational comparative study of political culture based on the new research method of sample surveys was published. Gabriel Almond and Sidney Verba's *The Civic Culture* (1963) had a major impact on the comparative study

of political culture, leading to replications, adaptations, extensions, and criticisms of both theory and methodology. By the 1970s and early 1980s, the role of the political culture approach had diminished considerably in political science. Since the mid-1980s the concept has regained stature, during a time when the role of culture in the social sciences and humanities has undergone far-reaching reinterpretation.

At the analytical level, rational modes of thought can be attributed to the cognitive dimension of political culture, and nonrational modes relate to the normative. Most political scientists, including some students of political culture, assume the rationality of political actors and essentially ignore the nonrational dimension—with notable and important exceptions. Rational choice and formal modeling approaches have strengthened this bias, making many political scientists unreceptive to cultural approaches. Reactions against attempts to make the rational actor paradigm hegemonic in the discipline, however, and the resurgence of identity politics—most dramatically in the manifestations of xenophobic nationalism and ethnic cleansing—have contributed to the renewed interest in the role of political culture. David Laitin (1998), who shows how political culture is part of an equilibrium in a coordination game, is one of several scholars who combine rational choice and political cultural approaches. The revival of interest in political culture is most apparent in the field of comparative politics (Ross 1997), but the concept has also gained importance in international relations, particularly in the areas of conflict resolution and strategic culture (Katzenstein 1996).

The two major approaches to political culture are semiotic (via anthropology) and cognitive (via social psychology). Richard Wilson (2000) subdivides them into five approaches. Hermeneutics, influenced by the work of Clifford Geertz, is an important bridge between traditional and postmodern approaches. The interpretative semiotic approach (characterized in the introduction to this article) has been very influential in the social sciences and humanities—the boundary between these two having become increasingly blurred in recent decades. James C. Scott (1990) utilizes this approach to investigate the boundaries of the political. The culture theory approach builds upon the work of Mary Douglas. Developed by Michael Thompson, Richard Ellis and Aaron Wildavsky (1990), it postulates a universal division of polities into four ideal-type subcultures based on two criteria. Analysis focuses on the contestation within polities between competing subcultures, and explains change in terms of their different competencies and biases.

Social character is the third approach influenced by anthropology, but it also draws heavily from social psychology. It emerged from studies of culture and personality, and national character. Personality variables are analyzed in terms of types of significant value orientations, including attitudes toward and dependence upon authority, fear of disorder, and various religious and ideological positions. These studies generally are based on the analysis of the relationship between norms and behavior in specific cultural contexts. Although a few scholars define types of broad civilizations, most, like Robert Putnam's (1993) influential study of civic traditions in Italy, focus on selected case studies. The social learning approach, based more directly on psychological theories, utilizes social surveys to analyze values and beliefs to explain the role of socialization in determining values and behavior. Ronald Inglehart's (1997) work on the relationship between generational change and postmaterialist values illustrates this approach well, following the tradition of *The Civic Culture*. Other cognitive approaches incorporate Piaget's theories and conceptualize value orientations very differently from social character studies. Analysis focuses on the reciprocal relationship between individual development and social life. In some of this work moral reasoning and development is at the center of analysis.

Ironically, although political science borrowed the concept of culture from anthropology, contemporary perspectives in anthropology challenge the utility and integrity of the concept, and either avoid the term culture or bracket it with quotation marks to signal its problematic nature (Brightman 1995). Postmodern approaches deconstruct political culture by defining all cultural phenomena in terms of domination and contestation for power. Many practitioners of cultural studies use euphemisms for culture to explore the relationship between power and the imaginative expression of culture. Much innovative work on political culture crosses disciplinary boundaries.

4. *Changes in Focus and Recent Developments*

Contemporary understanding of political culture builds on the foundations of Karl Marx's notion of reification, Max Weber's analysis of legitimation, and Michel Foucault's insights into the cultural definitions of the political and his analysis of power in institutional discourses. Many scholars utilize Antonio Gramsci's conceptualization of hegemony, Eric Hobsbawm and Terence Ranger's formulation of the invention of tradition, Pierre Bourdieu's notions of habitus, doxa, and cultural capital, which reveal how power is inscribed in the scripts of everyday life, and Jurgen Habermas's concept of the public sphere. Feminist theory has also influenced contemporary formulations by calling attention to the gendered aspects of political culture. Volumes edited by Nicholas B. Dirks, Geoff Eley, and Sherry B. Ortner (1994) and by Victoria E. Bonnell and Lynn Hunt (1999) are excellent examples of such contemporary, cross-disciplinary approaches. Their appearances in two of the half dozen important interdisciplinary series on politics and culture established by major university presses are indicative of this trend.

Although there are important exceptions to these broad generalizations, traditional and contemporary approaches can be contrasted in several major respects. Traditional approaches define political culture in terms of attitudes and values, whereas more contemporary approaches view culture in terms of scenarios and discourses. The former conceive of political culture as reified, holistic, discrete, clearly bounded, coherent systems; while the latter view boundaries as fragmented, tenuous, and contested. Older approaches view traditions as objective or 'natural' and assume continuity, while newer approaches perceive traditions as 'invented' social constructions and assume they are dynamic because actors have (limited) options. Some traditional approaches were ahistorical, while most contemporary approaches ground analysis in historical contexts. Contemporary students of political culture examine the mechanisms of transmission of political culture and strategies of boundary maintenance. Traditional approaches focus on goals, while contemporary approaches focus on strategies of action and the incompatibility of strategies of various actors. The analysis of limited repertoires of cultural options available to actors seeks to conceptualize the dialectic of system and practice.

5. *Current Areas of Research and Future Directions*

An important area of research is state and society relations (Migdal 1989). The disintegration of the Soviet Union and the emergence of the new states of Central and Eastern Europe give renewed currency to interest in civil society, democracy, state and institutions, political communication and meaning, and collective action (Berezin 1997). The ongoing attempt to determine whether there may be certain cultural prerequisites for democracy has been chastened by postmodern and postcolonial critiques of the Western biases in such investigations. More culturally informed reformulations of traditional areas have resulted in the study of New Social Movements and the New Institutionalism (Powell and DiMagio 1991).

Interest in civil religion and in the relationship between religion and politics (including fundamentalism) has increased. Benedict Anderson's (1991) notion of the nation as an imagined community is one of the many publications on collective identity which, in addition to nationalism, examine ethnicity, race, gender, and sexual preference. Studies of political rituals, rhetorical strategies, strategic discourse, commemorations, museums, political funerals—including the reburial of the remains of previously buried 'heroes'—are areas of considerable innovation. The study of the cultural dimension of processes of globalization and localization (including migration) is another frontier for the study of political culture from its increasingly interdisciplinary perspective. The conceptual threads connecting the different approaches to political culture are as tenuous as the webs of meaning constituting political culture. Yet, the fragility of both does not lessen their importance.

See also: Civic Culture; Civil Society, Concept and History of; Democracy; Democratic Theory; Democratic Transitions; Globalization: Political Aspects; Ideology: Political Aspects; Legitimacy: Political; Political Protest and Civil Disobedience; Power: Political; Race Relations in the United States, Politics of; Rational Choice in Politics; Religion and Politics: United States; Socialization: Political; Zionism

Bibliography

Almond G A, Verba S 1963 *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton University Press, Princeton, NJ

Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, New York

Aronoff M J 1989 *Israeli Visions and Divisions: Cultural Change and Political Conflict*. Transaction, New Brunswick, NJ

Berezin M 1997 Politics and culture: A less fissured terrain. *Annual Review of Sociology* **23**: 361–83

Bonnel V E, Hunt L (eds.) 1999 *Beyond the Cultural Turn: New Directions in the Study of Society and Culture*. University of California Press, Berkeley, CA

Brightman R 1995 Forget culture: Replacement, transcendence, reflexification. *Current Anthropology* **10**: 509–46

Dirks N B, Eley G, Ortner S B (eds.) 1994 *Culture/Power/History*. Princeton University Press, NJ

Foucault M 1966 *Les Mots et les choses*. Editions Gallimard, Paris [1970 *The Order of Things: An Archaeology of the Human Sciences*. Random House, New York, NY]

Gendzel G 1997 Political culture: Genealogy of a concept, *Journal of Interdisciplinary History* **27**: 225–50

Inglehart R 1997 *Modernization and Postmodernization: Cultural, Economic and Political Change in Forty-three Societies*. Princeton University Press, Princeton, NJ

Katzenstein P J (ed.) 1996 *The Culture of National Security: Norms and Identity in World Politics*. Columbia University Press, New York

Kertzer D I 1988 *Ritual, Politics, and Power*. Yale University Press, New Haven, CT

Kubik J 1994 *The Power of Symbols Against the Symbols of Power: The Rise of Solidarity and the Fall of State Socialism in Poland*. Pennsylvania State University Press, University Park, PA

Laitin D 1998 *Identity in Formation: The Russian-speaking Populations in the Near Abroad*. Cornell University Press, Ithaca, NY

Lasswell H 1960 *Politics: Who Gets What, When, How*. Meridian Books, New York

Lustick I S 1993 *Unsettled States, Disputed Lands: Britain and Ireland, France and Algeria, Israel and the West Bank–Gaza*. Cornell University Press, Ithaca, NY

Migdal J 1989 *Strong Societies and Weak States: State–Society Relations and State Capabilities in the Third World*. Princeton University Press, Princeton, NJ

Ortner S B 1973 On key symbols. *American Anthropologist* **75**: 1338–46

Powell W, DiMaggio P J (eds.) 1991 *The New Institutionalism in Organizational Analysis*. University of Chicago Press, Chicago, IL
Putnam R 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ
Ross M H 1997 Culture and identity in comparative political analysis. In: Lichbach S, Zuckerman A S (eds.) *Comparative Politics: Rationality, Culture, and Structure*. Cambridge University Press, New York
Scott J C 1990 *Domination and the Arts of Resistance: Hidden Transcripts*. Yale University Press, New Haven, CT
Swidler A 1984 Culture in action: Symbols and strategies. *American Sociological Review* **51**: 273–86
Thompson M, Ellis R, Wildavsky A 1990 *Culture Theory*. Westview, Boulder, CO
Wilson R W 2000 The many voices of political culture: Assessing different approaches. *World Politics* **52**: 246–73

M. J. Aronoff

Political Discourse

1. Introduction

Discourse in general is a way of organizing human experience. It establishes frames of meaning by the recounting and interpreting of events and situations. It constructs systems of order. Political discourse applies such frames to the exercise of power—including principles of hierarchy, representation, and accountability. It deals with the narrative interpretation of events and ideas, logical and mythic, and establishes criteria and contexts for comparing and evaluating political systems. In these terms, political discourse is as old as politics, its pedigree going back at least as far as Plato for whom discourse took the form of a *dialectical logic*. Plato used this method to derive a model system composed of appropriate connections between truth, knowledge, virtue, and justice that he used to evaluate and rank different polities. Plato made explicit what might best be referred to as the relationship between discourse as rationality and rationality as discourse. But he also recognized that there were limits to the role and place of rationality in human affairs. Where the rational leaves off, he invoked a revelatory higher power, religious, mythic, magical, numerological—a Pythagorean idea of myth, (Cassirer 1946). One can say that Plato defined political discourse in terms of a *mytho-logic* that he regarded foundational for the functioning of any political community. For Plato the state, or better, the city, is a discourse community.

For Aristotle political discourse was an emergent teleological process embodying final causes, the shape, purpose, and meaning of the polity. This established a basis for a purposeful history of ideas—a narrative realizing itself in the evolution of democracy of which there are several versions. One starts with Athenian democracy as a defining and classic polity, Rome representing natural and positive law (with interpretive exegesis by Glossators and Commentators). The Schoolmen raised the issue of revelation, ecclesiastical law vis-à-vis natural law. Medieval corporatism established the principle of conciliarism (and the kings 'two bodies') (Kantorowicz 1981). Social contract theorists, from Hobbes, Locke, and Mill, defined a political spectrum, which at one end was autocratic, the other democratic. Another version uses ages: classic, ecclesiastical, enlightenment, romantic, democratic. The end of history is modern, secular, democratic government. Which makes a better story than a good history.

Although hardly a corpus or a stable body of knowledge, modern political discourse theory is composed of a variety of concepts and ideas from many disciplines. As a form of critique it tries to penetrate below the surface of the good stories that people tell themselves about politics. Discourse theories may differ but they are all concerned with what it is that people come to believe about power and its exercise, and how they come to think that way. Political science and sociology analyze different ideologies, dogmas, and belief systems in terms of their content and the extent and intensity of their support to understand threats and opposition to democracy and democratic principles.

As with Durkheim's treatment of social facts, political beliefs are no different from any political perceptions and preferences, loyalties and affiliations. Examining them requires operational techniques, quantitative analysis. Belief systems are amenable to survey methods, questionnaires, and statistical techniques (Lane 1962, Converse 1964; Putnam 1973, Tilly 1978, Sidney Tarrow 1994).

In contrast, several other disciplines break up concepts like ideology, belief systems, or culture into components on the grounds that they have become 'overkill'; that by explaining too much they explain too little. How political discourse establishes order and control and organizes and mobilizes social groups and organizations requires more differentiated categories and more interpretive knowledge. How is it that words can kill? When does ecclesiastical belief lead to murder in the cathedral? How much was Nazi ideology the cause of the Holocaust? How and when does ethnicity, religion, language, or other differentiating social criteria that establish exclusivity define 'others' in negative terms—the extreme consequence being genocide? Similarly, when do such circumstances reverse themselves? When do the same defining characteristics constitute pluralism, multiculturalism, or diversity? How does confession within the framework of legal discourse connect guilt with justice, and by so doing change the contours of the permissible? How was Apartheid in South Africa, with its biblical racism, its Boer myths of a chosen people replaced by a Truth

and Reconciliation Commission? These and similar questions have led to a virtual explosion of concepts, categories, and theories: structuralist, phenomenological, linguistic, psychological.

Despite differences among approaches, all favor the analysis of events as more than chance, i.e., as coherent and meaningfully and internally connected. In terms of political discourse theory, events are *social texts*. They can be read as *language* including words, of course, what people say about each other and their condition, but also as *signs*, *symbols*, *rhetoric*—as *semiotics*. Discourse theory addresses the problem of how best precisely to read events. It is less concerned with the content of particular ideologies than 'the content of the form' (White 1987) and the 'narrative construction of reality' (Bruner 1991).

Political discourse theory occupies a position on a continuum, the opposite end of which is political economy and rational choice. Like those approaches, it responds to a calculus of costs and benefits, losses and gains. Unlike them what is exchanged is outside the normal range of interest competition and bargaining—*symbolic* rather than economic capital. If so, then an important purpose of political discourse theory is to identify the coin by which different rationalities are measured.

While the substance of political narratives varies widely, they follow certain fairly standard trajectories, including the *recounting* of events in the form of *retrievals* and *projections*. Events serve as *signifiers* and *metaphors* in which meanings are transmitted in terms of past and similar situations, real or fancied; and as *metonymies* in which the event is a fragment or representation of some larger logical theoretical or revelatory belief system. By this means the past provides evidence for a projected and transcending state of affairs, including millennial outcomes, promised lands, states of grace, or a condition of political virtue or justice. This suggests a central question: when does political discourse lead away from a politics of give and take associated with rational economic behavior and when does it generate *symbolic capital*? (Bourdieu 1977). It is this question that will be discussed here.

First the intellectual pedigree of political discourse will be examined, indicating ways that discourse affects the way people think about politics, act on what they think, and, as a part of such action, how they form discourse communities using symbolic capital, in contrast to those predominantly organized around economic and political exchange followed by a critique and assessment of future trends.

2. *An Intellectual Pedigree*

Modern concern with political discourse might be said to begin with the *semiotic* revolution. To name only a few contributors (and leaving out forebears like Peirce, Saussure, Wittgenstein, and others), one might include linguists like Roman Jacobson, Dell Hymes, and Umberto Eco; sociologists like Basil Bernstein, Erving Goffman, and Herbert Garfinkel; philosophers like Alfred Schuetz, John Austin, Ernest Gellner, and John Searle; anthropologists such as Claude Levi-Strauss, Clifford Geertz, Victor Turner, Mary Douglas, and Jack Goody; historians such as Robert Darnton, Pierre Nora, Le Roy Ladurie and of course many others. Political scientists include Benedict Anderson, David E. Apter, and James Scott. In literary theory there is a very large cohort concerned with *narrative* and *interpretation* (Mitchell 1982): Kenneth Burke, Roland Barthes, Hayden White, Frederic Jameson, Terry Eagleton.

Which brings up the question of the range of concepts that such thinkers apply to the analysis of political discourse theory. Walter Benjamin, examining the relation between memory and the crafted story, the potentiality for what might be called the *fictive use of recounting*, implies the significance of the *imaginary real*. Levi-Strauss laid out the structural properties or better, the isomorphic ordering properties of *myth* and *theory* in the form of binaries which literally compose the order of the universe, providing rules of appropriate *intermediation*: between man and the cosmos, men and women, kin and family, earth and sky, land and water, etc. (Levi-Strauss 1955). Similarly in Umberto Eco's theories of semiotics, with their emphasis on *codes* and *codex*, *signs* and *signification*, and what he called *sign production*.

In a more literary vein Hayden White refers in his analysis of narrative to the *poetic troping of the facts* in which certain events become in effect what Eagleton called *drenched signifiers*, images and metaphors so symbolically powerful that they become incentives to action. Foucault examines the way certain *discursive codes* establish authority and power. For François Furet political discourse becomes a *nonscientific representational strategy* or a structure of time, or *time consciousness*. Ricoeur emphasizes discourse as *sin*, *confession*, and *purification*, *negative poles* and *overcoming* and the uses of metaphor. Henri Lefebvre deals with space as *terrain* and jurisdiction as *agora*—a theater of intersection between *addressor* and *addressee*. For Baudrillard an appropriate discourse produces *symbolically dense* miniaturized versions of state and society, encampments, staging areas, fortresses, neighborhoods, which form the moral center for what is yet to be. Guy Debord emphasizes political discourse as theater, drama, *spectacle*.

While each of these and many other thinkers contribute in very different ways to political discourse theory, it is possible to identify one common and underlying concern. People make stories out of events. They do so individually and collectively. Recounting individual stories makes for sociability. Collective stories have political consequences when, as myths they purport to be history, as history they are

reinterpreted as theories, and as theories they make up stories about events. Theories that become stories create *fictive truths*. Since in politics, truth-telling and story-telling are part of the same process, it becomes possible to interrogate the past in order to transform the future.

There are of course many ways to examine such matters. A useful place to begin is with the actual ways and circumstances in which individuals recount their own stories. As a process, this becomes politically interesting if such stories are *collectivized*, systematized, and formed into what have been called *master narratives*. For this purpose *agency* is required, a public figure able to play the special role of 'story-teller' whether in the form of philosopher-kings, cosmocratic figures, politicians, soothsayers, diviners, or others (Benjamin 1969).

Agency, a form of authoritative recounting, begins with some benchmark, a state of grace which then defines a *negative pole*, loss, suffering, struggle, on the basis of projecting an *overcoming project*, a preferred goal. Such recounting is intrinsically *dramatistic*—exceptional experiences punctuated by *promontory events*. Events may be tragic, including suffering, torture, betrayal, murder, and death. They can also result in superior *insight*, knowledge through *struggle*, *purification*, and *redemption*. Recounting renders the unimaginable manageable. It builds into the narrative *imaginary reals*, necessary fictions which establish boundaries and borders, affiliations and loyalties, terrains and jurisdictions, defines insiders and outsiders, separates the good citizen from the pariah. The narrative establishes *boundaries* and *clienteles*, identifying relevant *markers*, ethnic, religious, racial, kin, clan, and class, or other categories. It locates *surrogates*, *negative* (but redeeming) *others*, Genet's thief, or homosexual, Fanon's colonial African, Foucault's madman or prisoner, etc. (Foucault 1979).

It deals too with *ritual endowments*, insignia, votary paraphernalia that go with them, flags, and uniforms. It emphasizes the theater of the absurd, or the solemn occasions, the mass, the funeral, the parade, votary ceremonies, candle light parades, pistols, whips, massed guns and technology; anything which permits *symbolic intensification*. Ordinary terrain, transformed into sacred space, serves as metaphors or better *simulacra* for the national patrimony (old and new Jerusalems). Politics is the stage on which leaders perform, recapitulating historical events as theater and choreography (Plant 1992).

Time becomes crucial, especially in *disjunctive* and *transformational moments*, when leaders propose a fresh start. The French revolution tried to reinvent the calendar in its desire to start the world all over again. Time foreshortened becomes space, a *geomancy of space*, enclosure as architecture, the placement of buildings, squares, parks, transportation, homes, meeting places, amphitheaters all taking on new meaning as potential rallying ground, or as Agora's or forums, staging areas, embattled zones, etc. So too with monuments, statues, shrines, graveyards, jails, public buildings, this pub, that hideout, this square, street, neighborhood. Virtually any ordinary thing can take on significance, disrupting the daily or routine business of commerce, its taken-for-grantedness, in favor of the exceptional moment in the exceptional space (Lefebvre 1986).

Similarly with clothes, uniforms, medals, gestures (verbal and other), hair, Jacobin hats, red flags, insignia, swastikas. These emphasize the body as reference point; exaggerated posture, sexual affronts, cross dressing, racial inversions. The body politic takes on the characteristics of the individual body. It can be violated, penetrated, tortured, mobilized, transformed.

Opportunities for such *recounting* include the founding of new nations and states, revolutionary and redemptive *breaks*, a turning of the world upside down. In effect, political discourse helps produce such breaks. It also capitalizes on them when the opportunity arises. In such moments political discourse creates its own rationality, defining sharply what it is against and what it is for. It contains within itself a disordering effect in rejecting established political beliefs and institutions, and an ordering principle designed to negate and transcend the disordering condition it helped to create.

Political discourse then is bound up with history as conflict and struggle. Typical circumstances are one or other form of *inversion*; *revolutionary*, *ecclesiastical*. A classic device is some variation on Hegel's *master and slave*, or, more redemptive, the *last* becoming *first*. One can find both in the speeches and texts of early church fathers, fifteenth century mystics, and radical political revolutionaries from the Jacobins, to the Levellers and Diggers, not least of all including what Hobsbawm called *social banditry* (Hobsbawm 1959). Such situations lead to exceptional acts of daring and devotion. They involve a tension between *purity* and *danger* such that the words of the recounted narratives resonate.

Recounting is thus a way of both periodizing the past and incorporating it within a series of performances in which singular and distinctive individual events and actions stand out, the whole inspiring urgency, martyrs and heroes not to speak of loyalty, obligation, and what Bernard Yack has called political 'yearning' (Yack 1986).

Narrative is half the process. *Text* is the other. Both are required if political discourse is to become an active ingredient in political life. Required of texts in relation to symbolic capital is that they provide exegetical opportunities. *Textualization*, the transition from *orality* to *writing*, requires not only a narrator but a writer. Both are essential if those documents designated as sacral are to serve as symbolic centers of power, complete with their coterie of interpreters devoted to locating hidden meanings, logical and

magical, thus providing those esoteric nuances that make a discourse community unique and which give it its moral center. Text and coterie, with the agent at the center, suggest truth as insider knowledge. They create an *iconography of words and objects* worthy of worship and respect. So virtually all societies and social movements have certain texts which become canonical, whether they take the form of a constitution or a 'little red book.'

Whether as Marxist texts or canonical or biblical texts, the word as written embodies both logical truths and mythic or fictive ones. Some are enshrined in sanctuaries. Others as authorizing documents are on suitable display (as for example the constitution in the United States or the Magna Carta in Britain). Hence, one might say that political discourse involves the alteration of original meanings by means of mythic retrieval and logical projection, what Roland Barthes and Claude Levi-Strauss referred to as 'mytho-logics' (and what Malinowski called a mythic charter).

Which returns us to the question raised above. How can history as a representation of symbolic events become simply history—exceptional moments becoming unexceptional? Or, to put it differently, how can symbolic capital be converted into normal interest competition and bargaining within a framework of constitutional authority and law? It is a question that comes up in the context of wars, nationalist, separatist, and revolutionary movements, the rise of cults, persistent and organized protest—matters which if anything are more important today than ever before. In this sense one can say that political discourse is about how to change the world by reinterpreting it, and political discourse theories are about how to interpret the interpretations.

3. *Applications and Implications*

The twentieth century has perhaps been more witness to powerful and consequential political occasions than any previous one, not least of all because of the connection between terror and utopianism. If one includes the power of nationalism, the mobilization of its rhetoric, which accompanied the decline of colonialism and the founding of new states in Asia and Africa, one thinks too of a virtual explosion of political *agents* like Nkrumah, Senghor, Sekou Toure, and Nyerere, leaders all of whom, in the name of anticolonialism, nationalism, and independence, created their own political vernacular texts, doctrines by which they claimed exceptional and continuing political privileges (emulated by their putative descendants). In these terms too political discourse was very much a phenomenon of such populist dictatorships as Peron's Argentina not to speak of De Gaulle's myth of a glorious (and largely fictive) France.

Perhaps the exemplary defining event or historical moment, or disjunctive break, was the Russian revolution with Lenin its agent, surrounded by a band of intellectuals forming a coterie. In the texts so formed the strong elements are *orality* and *textuality*. Dramatic public spaces were carefully selected for delivery. What began in speeches was translated, that is, words became converted and modified in the form of writings many of which became canonical. As the redemptive and prophetary figure of the revolution, he wrote the bible of revolutionary principles and tactics, not simply for his own followers but subsequent generations and putative descendants. Perhaps the most significant of these disciples was Mao Zedong who 'rewrote' China's history in the form of three narratives: long, intermediate, and short; each fitted inside the other, *contextualizing* in the form of radical revolution and militant nationalism, converting time into immediacy, forming a *fictive/factual* basis for *revolutionary discourse* (Apter and Saich 1994).

Political discourse not only precipitates conflict and foreshortens time but also can extend the disjunctive moment and keep it alive. Examples of conflicts of the long *duree* include the IRA in Northern Ireland, and between the Israelis and Palestinians (where two diasporic populations remain locked together in their hostility). In both cases the parties are divided by their discourses, their retrievals of the past determining their view of the future, and so making mutual accommodation difficult as long as their territorial claims are to the same physical space. Each discourse has the same contours as the other, similar themes of loss and displacement. Each makes salvational, redemptive, and recuperative national claims. For one side to give way to the other is considered betrayal.

There are innumerable examples. The rise of fundamentalism or integralism enables the preaching of the word as *scripture*, something to be pored over and through a process of *exegesis* made to reveal hidden, deeper, and higher meanings. For some such words will be profound, for others ritualized, both appearing to offer magical depth, a common sensibility. By this means, the individual is linked to the state or sect and by multiple references and *reciprocities*. Virtually any categories will do, religion, ethnicity, clan, family, or region, village, town, or cult, political program, ideology, etc. All potentially are *condensed symbols*, complex, with multiple references and signifiers, providing opportunities for *exegetical bonding*—establishing boundaries in the mind that constitute boundaries on the ground.

There are certain recurrent themes: an original and negative condition to be transcended, a mythical past to be redeemed, a frame within which certain primary symbols are characteristic, loss, defilement, sin, stain. Grievance is converted to intentionality, time into space, and space into particular places, symbolically endowed terrains. By this means individual stories are collectivized and conveyed as master narratives to a group as a whole. Hence the narrative process itself creates political leaders acting as agents who validate

preferred outcomes embodied in texts. Such narratives contain a story within a story, the storyteller or agent figuring in his or her own story. Prototypical exilic figures are Odysseus or Moses. Forced to suffer the loss of the patrimony, wandering far from their homes, they learn to outwit enemies. By means of knowledge gained from experience, they regain the power necessary to reclaim the patrimony in the first instance or a Promised Land in the second.

The more cosmocratic the impulse, the larger than life the *principal agent*, the greater the moral density of the text, the more likely are the doctrines to intensify differences, separating believers and nonbelievers; the saved and the damned, divided by a potential millennial accomplishment. Some are privileged at the expense of others. A chosen people will claim superiority, whether by class, ethnicity, religion, nationality, race, or doctrinal persuasion. Each such group becomes both *surrogate* for higher truths and a symbol or representation of them. Loyalty is thus moral obligation, affiliation deep seated, binding groups together. Hence political discourse theory constitutes as its subject matter all those cleavages of human *association*, which especially in a context of violence acquire durable properties reinforced rather than undermined in subsequent events. In these terms, discourse communities generate their own objects, symbolic capital possessing its own *exchange value*.

There are certain serious weaknesses to political discourse theory so considered. It works best in conjunction with intensive case studies in depth. It depends on depth interviewing or political anthropology. It is not very amenable to operational techniques, survey, path analysis, and other statistical methods. Its relativism is less concerned with the content of ideologies, the conventional interest of political scientists, than in the structure of beliefs, what forms them, and how they become adopted. There is also a bias that over time rationality will triumph over nonrationality, logic over revelation, theory over myth, facts over appearances.

However one approaches it, discourse theory is not likely to disappear in the foreseeable future. It has moved in several different general directions: toward *postmodernism* in the connection between *narration and legitimation* (Lyotard 1984); *cultural studies*, with respect to the symbolization of violence and the state (Bhabha 1990); *sociolinguistics*, itself divided into three main strands, psychological, linguistic, and sociological (Giglioli 1982); towards an intermediate position between comprehensive structural theory and the concrete explanations (or justifications) people give for their actions and beliefs—what has been called 'critical hermeneutics' (Habermas 1979). Finally, for those interested in politics, power, and rational choice theory, some attention is now being paid to how symbolic capital fits into much the same framework as economic capital, using narrative to show how the two represent different forms of the same rationality (Bates et al. 1998). Indeed, if 'money' is itself the symbolic form of utility *influence* is the consequence of symbolic capital. If so, rational choice theory and discourse theory instead of occupying opposite ends of the same continuum may come to constitute their own hermeneutic circle.

See also: Discourse and Identity; Discourse, Anthropology of; Linguistic Turn and Discourse Analysis in History; Metaphor and its Role in Social Thought: History of the Concept; Narrative, Sociology of; Narratives and Accounts, in the Social and Behavioral Sciences; Political Communication; Political Culture; Semiotics

Bibliography

Anderson B 1983 *Imagined Communities*. Verso and New Left Books, London

Apter D E, Saich T 1994 *Revolutionary Discourse in Mao's Republic*. Harvard University Press, Cambridge, MA

Bates R et al. 1998 *Analytic Narratives*. Princeton University Press, Princeton, NJ

Baudrillard J 1981 *Simulacres et simulation*. Editions Galilee, Paris

Benjamin W 1969 The storyteller. In: Arendt H (ed.) *Illuminations*. Schocken Books, New York

Bhabha H K (ed.) 1990 *Nation and Narration*. Routledge, London

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK

Bruner J A 1991 The narrative construction of reality. *Critical Inquiry* **18**(1): 1–21

Cassirer E 1946 *The Myth of the State*. Yale University Press, New Haven, CT

Converse P A 1964 The nature of belief systems in mass publics. In: Apter D E (ed.) *Ideology and Discontent*. The Free Press, New York pp. 206–61

Debord G 1987 *La societe du spectacle*. Gerard Lebovici, Paris

Foucault M 1979 *Discipline and Punish*. Random House, New York

Geertz C 1973 Ideology as a cultural system. In: Geertz C (ed.) *The Interpretation of Cultures*. Basic Books, New York pp. 47–76

Gigliol P (ed.) 1982 *Language and Social Context*. Penguin, London

Habermas J 1984 *The Theory of Communicative Action*. Beacon Press, Boston

Hobsbawm E S 1959 *Primitive Rebels*. University of Manchester Press, Manchester, UK

Jameson F 1972 *The Prison-House of Language*. Princeton University Press, Princeton, NJ

Kantorowicz E H 1981 *The King's Two Bodies*. Princeton University Press, Princeton, NJ

Lane R E 1962 *Political Ideology*. Free Press, New York

Lefebvre H 1986 *La production de l'espace*. Editions Anthropos, Paris

Le Roy Ladurie E 1979 *Carnival in Romans*. George Braziller, New York

Levi-Strauss C A 1955 The structural study of myth. *American Journal of Folklore* **LXXVIII**(270): 428–44

Lyotard J-F 1984 *The Postmodern Condition: A Report on Knowledge*. The University of Minnesota Press, Minneapolis, MN
Mitchell W J T 1982 *The Politics of Interpretation*. University of Chicago Press, Chicago
Nora P (ed.) 1996 *Realms of Memory*. Columbia University Press, New York, 3 Vols.
Plant S 1992 *The Most Radical Gesture*. Routledge, London
Putnam R 1973 *The Beliefs of Politicians*. Yale University Press, New Haven, CT
Ricoeur P 1968 *The Symbolism of Evil*. Harper and Row, New York
Tarrow S T 1998 *Power in Movement*. Cambridge University Press, Cambridge, UK
Tilly C 1978 *From Mobilization to Revolution*. Addison Wesley, Reading, MA
White H 1987 *The Content of the Form*. Johns Hopkins University Press, Baltimore, MD
Yack B 1986 *The Longing for Total Revolution*. Princeton University Press, Princeton, NJ

D. E. Apter

Political Economy, History of

The term 'political economy' first appeared during the early seventeenth century in French texts that considered the contribution that trade, agriculture or manufacturing might make in augmenting the powers of the sovereign and rendering their subjects more prosperous. It is linked to the rise of the early modern state: it assumes that the interest of the monarch lies in a strong state, and that the path to a strong state leads via the augmentation of a population, and the extension of their means of subsistence. The interests of sovereign and subject therefore coincide—the strength of a state is founded upon the well-being and numerousness of its population. During the seventeenth and early eighteenth centuries an extensive literature on the contribution of trade and agriculture to the welfare of people and ruler developed in the English, French and German languages. The diverse practical propositions encountered in such writings were systematized in the course of the eighteenth century, first through the work of François Quesnay (1694–1774) and the Physiocratic 'School' with which he was associated, and subsequently through Adam Smith's (1723–1790) critique of Physiocratic and 'mercantilist' schools of 'political economy.' Smith's *Wealth of Nations* (1776) provided the foundations for a new 'principled' science of political economy, the chief architects of this new science being Thomas Robert Malthus (1766–1834) and David Ricardo (1772–1813). The 'classical political economy' which these and their contemporaries represented elaborated a limited set of economic principles turning on the proposition that, ultimately, the exchange value of a commodity was reducible to the amount of labor embodied in it. Price, therefore, inhered in the good, rising or falling according to the cost of labor. By the 1830s an independent American literature was developing on this basis, and in the final two decades of the century Dutch, French and English texts began to be read in Japan. In continental Europe Jean-Baptiste Say's (1767–1832) reception of Smithian principles was extremely influential; his writings laid the basis for the transformation of 'political economy', during the last third of the century, into a new 'economics' in which value was a product of subjective choices, and price the outcome of the degree of scarcity which these choices created.

1. *The Derivation and Evolution of the Term 'Economy'*

Although the discourse of political economy first emerged in early modern Europe, the term derives from Aristotelian beliefs concerning social and political order. Aristotle's *Politics* outlined two forms of human association: the *oikos* and the *polis*. The first, the household, was instituted for the satisfaction of private daily needs; the second, the domain of political association, structured public life. Aristotle introduced this distinction in response to Plato's argument that the art and technique of government was everywhere the same; much of *Politics* is therefore taken up with the different ways in which public life could be structured. By contrast, only one form of organization was foreseen for the household, whose head ruled over its members like a monarch over its subjects, securing the proper employment and maintenance of property and persons. Elsewhere in his *Nicomachean Ethics* Aristotle had argued that the virtuous man set no store by wealth, hence it could be no part of virtue to deliberately increase the wealth of a household. The purpose of the household was therefore the provision of an ordered material existence to its members through the allocations made by the household head. Beyond the household, in the public domain, merchants engaged in trade and money-getting, but this was activity governed by no definite conceptions of virtue or order, subordinated to the superior principles of government.

Vestiges of this way of conceiving social and political relations persisted into the eighteenth century, especially in English, French and German treatises which adumbrated the proper management of farm and field. However, in the early seventeenth century 'économie politique' adopted the central householding tenet of the Aristotelian model outlined above, projecting it into the public domain, hence creating a new 'political oeconomy'—which sought to identify the most expedient manner in which the interest of ruler and ruled might be reconciled within a state conceived as a giant household. Although this

naturally involved proposals for the augmentation of population and the means of their subsistence, the pursuit of wealth for its own ends had no definite place in this new order. In France, *commerce* was the term assigned to activity related to trade, finance, and money; literature related to this aspect of economic activity developed quite separately to that of a political oeconomy primarily focused on agricultural, subsistence activities.

2. The Sources of Power and Plenty

In the context of a European economy recovering from the ravages of the Thirty Years War, vulnerable to epidemics on the one hand and periodic poor harvests on the other, the priority given to agriculture as the foundation for any increase in numbers and welfare requires little elaboration. Overseas trade could make a small country like Holland rich, but both Britain and France were simply too large for trade to have such an impact. When in the early nineteenth century David Ricardo outlined the benefits of international trade he chose as his examples English wool against Portuguese wine—products that both countries had in abundance, of course, but where the increase in welfare consequent upon exchange had little impact upon the central issue of subsistence. And in fact the overseas trade of Britain throughout the eighteenth century facilitated the import of sugar, tea, coffee and spices, luxury goods whose significance in turn became part of a debate on the implications of the production and consumption of such goods. Not until the end of the century did grain imports exceed exports, and this reflected both an increasing population and a shift in the structure of production that formed the origins of the industrial revolution. This new pattern of international trade—in which domestic employment and a rising population increasingly depends on the export of manufactured goods in return for which food and raw materials are imported—first emerged in the nineteenth century.

States with ready access to the sea—Spain, France, Britain, and the Hanseatic cities—had the opportunity to develop overseas trade and fisheries, and an associated literature developed which both urged the extension of trade as the path to wealth, and outlined the practicalities of successful trade and commerce. Where access to the long-distance trade that the oceans offered was limited, arguments that herein lay the route to wealth and power had little force. Moreover, the profitability of merchant ventures seemed to depend upon risk and chance, not upon the wise investment of capital, the proper management of resources, and the judicious use of labor that typified a flourishing agriculture. The rhythm and progress of production in this sphere of economic activity lent itself to the conceptual framework of political oeconomy and could without serious modification be extended to manufacturing activities. German texts of the eighteenth century routinely countered *Landwirtschaft* to *Stadtwirtschaft*, the town economy to the country economy, signifying manufacture in the former and agriculture in the latter—even though much early manufacturing actually took place in rural locations. But the contrast set up in this manner simply excludes serious consideration of trade and finance as significant elements in the welfare of the state.

3. The Enlightenment Science of Political Economy

In the mid-eighteenth century a 'new science of political economy' was created, based upon a conception of the 'natural order' of society in which agricultural activity was the unique source of a 'net product.' The systematic analysis of production and circulation that was developed by the so-called 'Physiocrats' between 1760 and 1780 owed much to contemporary enlightenment rationalism, but it was also prompted by the growing problems of the French economy, and the increasingly parlous nature of the state finances.

Physiocratic political economy identified the natural foundations of economic activity, and at the same time implied that proper management of the economy required that the ruler not 'govern too much,' that excessive interference would impede the formation of wealth. The source of wealth was agriculture, whose net product circulated through the various classes of society; the greater this net product, the stronger the circulation and hence the greater the wealth. Improvements required investment, which could only take place if the net produce was not in the course of its circulation diverted into unproductive ends, such as the production of luxury goods for landed proprietors. Special taxes and duties were impediments to the free circulation of goods; and since all taxes were ultimately funded by the net product, all state revenue should be drawn directly from the sole source of this net product, agriculture.

Quesnay illustrated his ideas with a circulatory diagram, the *Tableau Économique* (1758–9), depicting the flow of net product from its agricultural source, through the hands of the landlords, thence to the 'sterile class' of manufacturers, and then, through the agricultural purchases this last group made, back again to the source. This clear diagrammatic representation of three economic classes and the flows of goods and payments between them is now seen as the original conceptualization of circular flow in an economy, but the originality of this drew little contemporary attention. Criticism of the Physiocratic system isolated two central tenets: the proposition that

non agricultural employment was unproductive, and that a single tax upon agriculture was viable, practically or intellectually. The most prominent criticism of this system can be found in Smith's *Wealth of Nations*.

4. *The Smithian Critique of the 'Mercantile' and 'Agricultural' Schools*

In Book IV of *Wealth of Nations* Adam Smith arranged his intellectual forbears into two (fatally flawed) schools of thought. The first of these was the 'commercial, or mercantile system,' which, he argued, sought to enrich one state by impoverishing others, imposing duties and a range of restraints of trade designed to attract money and gold into the land. In part a rhetorical exaggeration designed to highlight his own 'system of natural liberty,' and in part a willful misreading of an extensive and varied literature more sophisticated than he would allow, his remarks laid the foundations for a critique of 'mercantilist' policy that has survived up to the present day.

The other school to which Smith addressed critical remarks was the 'agricultural' school. He rejected the idea that agricultural labor was the sole source of wealth, and that other labor was unproductive; but he warmly supported the Physiocratic representation of the annual flow of goods as the source of wealth, and the proposal that this flow should be as little impeded as possible. These were indeed central principles of Smith's own work, together with a redefinition of what counted as 'productive' as against 'unproductive' labor.

Smith defined the 'wealth' of a nation not as an accumulation of gold, nor as simple populousness, but as the annual produce of its (productive) labor. Productive labor was that embodied in capital; it left a physical trace, unlike a service (no matter how useful or vital to the state that service might be). This could be augmented through the extension of the division of labor; labor could be made more productive by subdividing tasks and rationalizing the production process. There were two provisos attached to this: the first was that this process of the increasing division of labor had greater potential in manufacturing than in agriculture; and the second was that the degree to which it could develop was limited by the extent of the market. From this it follows that the most rapid increase of national wealth would result from the augmentation and rationalization of labor applied to manufacturing, within an economy which was both internally and externally freed from impediments to trade.

Smith was writing in the context of an economy whose structure and rhythms were dominated by agricultural production. This was not only the major economic sector; it was the immediate source of food and was prone to annual fluctuations in output and associated variations in prices. Since foodstuff represented such a major part of expenditures on the part of a majority of the population, the achievement of a low and stable price for grain was a crucial political problem throughout the eighteenth century for European governments.

Broadly speaking, there were two opposed solutions to this question. The first option was to encourage home production of grain by protecting domestic producers: high tariffs on imported grain would prevent the flooding of the market with cheap food in good years, undermining home production which would then be unable to provide adequate supplies whenever the sources of foreign grain diminished, or even dried up altogether. This option therefore secured whatever the level of home produce might be to domestic consumers, albeit at a high price. The second option also encouraged home production, but did so by permitting free import and export of grain. According to this, the availability of a market greater in extent than that offered by the national economy would stimulate domestic producers, consequently rationalizing their activities so that cheap foreign grain would not appear to be such a constant threat to domestic production. This second option proposed that the domestic population would be best served by a supply of cheap food based not upon the protection and control of the home market, but rather upon their access to an international market in foodstuffs. Smith's argument that the division of labor was a powerful engine of economic growth, and that it was limited only by the extent of the market, was therefore an argument for the benefits of free trade in all goods, augmenting the welfare of all states through the extension of the international economy.

5. *The Principles of Classical Political Economy*

Malthus' *An Essay on the Principle of Population* (1798) was conceived as a critique of contemporary conceptions of human betterment, and it certainly played a part in discrediting the conventional association of large populations with economic wealth and national progress. But his demographic arguments presupposed a closed agricultural economy, enjoying none of the benefits of commerce outlined by Smith. Some five years before the publication of the first edition of the *Essay* Britain had, however, become a net importer of grain. Henceforth, exported manufactures would pay for the imports of food required to subsist a large, increasingly manufacturing, population. Expanding (free) trade and commerce was to provide the solution for the 'Malthusian problem.'

Nonetheless, during the first two decades of the nineteenth century the elaboration of the classical theories of production and distribution presupposed

the conditions of an agricultural economy, producing a staple crop through the joint contribution of laborers, farmers and landowners. These three economic classes received wages, profits and rent, respectively, and the question which political economists sought to resolve was the relation between price formation and the distributive shares received by each class. Smith had assumed that prices were simply the sum of distributive shares; political economists sought a systematic explanation of the level of each of the elements within a framework that supposed that human labor was the sole creator of value.

Debate over Britain's Corn Laws during 1814–5 provided the occasion for the systematic linkage of the pattern of domestic income distribution to the trade regime. Malthus composed three pamphlets during this period which argued the benefits of limited protection, and suggested that rising rents were a reflection of human progress, since he regarded rent as a payment made subsequent to the settlement of wages and profits. Ricardo in his own pamphlet, known as the *Essay on Profits*, dissented from Malthus' protectionist sentiment and assumption that rising rents were part of the progress of wealth. Since he assumed that wages were fixed, he argued that there was an inverse relationship between rent and profits. High rents, therefore, meant low profits, they furthermore meant high prices for grain and hence increased the wages that manufacturers must pay their own workers. Since he also assumed that the rate of agricultural profit determined the rate of profit in general, protection for the landlord and high rents depressed the rate of profit in the economy as a whole. The solution to this prospect, he argued, was free trade in grain, which would force agricultural rationalization and lower the average price of grain, the basic foodstuff for the majority of the population. Ricardo's *Principles of Political Economy and Taxation* (1817) elaborated these points, and established a core of theoretical propositions that survived until late in the century, John Stuart Mill's *Principles of Political Economy* (1848) providing a lasting synthesis of classical principles in a manner that made the science accessible to an informed public readership.

6. The Diffusion of Political Economy

Smith's *Wealth of Nations* was by 1800 acknowledged everywhere as the canonical text of political economy, but nineteenth century English writing had relatively little impact outside English-language areas, and at this time anything written in French had a far bigger readership than texts written in English. Jean Baptiste Say had been greatly influenced by his reading of Smith, and in 1803 he produced his own *Traité d'Économie Politique*, followed by a number of new editions, a 'catechism' in 1815 and a new six-volume 'complete course of political economy' in 1828–9. Say conceived political economy as a significant part of an education for republican citizenship, and his writing was deliberately directed to a large popular audience. Since his texts were readable and in French they were quickly taken up outside France, with the result that, after Smith, Say became the most widely recognized political economist of the first half of the nineteenth century.

Say is remembered today for his 'law,' that 'supply creates its own demand.' In fact, this is a travesty of his argument on 'effective demand,' that whatever the general demand for a good, the political economist was only concerned with the demand of those with means to purchase, which means had themselves been gained from previous sales of their goods or services to others. He also argued that the process of production and distribution involved utilities, not goods, and that what the process of consumption destroyed was the utility of a good or service, and not the good or service itself. When combined with his principle of effective demand, this shifted the emphasis of his work away from the (production-oriented) English Classical Political Economy and towards the sphere of consumption. In these and other respects Say's political economy laid the path which led to the new subjective, marginal economics of the 1870s.

See also: Area and International Studies: Political Economy; Capitalism; Industrial Society/Post-industrial Society: History of the Concept; Industrialization, Typologies and History of; Labor History; Labor, Division of; Malthus, Thomas Robert (1766–1834); Marx, Karl (1818–89); Mass Media, Political Economy of; Political Economy in Anthropology; Power in Society; Smith, Adam (1723–90); Wealth Distribution

Bibliography

Faccarello G (ed.) 1998 *Studies in the History of French Political Economy. From Bodin to Walras*. Routledge, London

Griswold C L J 1999 *Adam Smith and the Virtues of Enlightenment*. Cambridge University Press, Cambridge, UK

Hilton B 1977 *Corn, Cash, Commerce. The Economic Policies of the Tory Governments 1815–1830*. Oxford University Press, Oxford, UK

Hirschmann A O 1977 *The Passions and the Interests. Political Arguments for Capitalism before its Triumph*. Princeton University Press, Princeton, NJ

Hont I, Ignatieff M 1983 (eds.) *Wealth and Virtue. The Shaping of Political Economy in the Scottish Enlightenment*. Cambridge University Press, Cambridge, UK

Letwin W 1963 *The Origins of Scientific Economics. English Economic Thought, 1660–1776*. Methuen, London

Peach T 1993 *Interpreting Ricardo*. Cambridge University Press, Cambridge, UK

Thomas W 1979 *The Philosophic Radicals: Nine Studies in Theory and Practice, 1817–1841*. Oxford University Press, Oxford, UK

Tribe K 1988 *Governing Economy. The Reformation of German Economic Discourse, 1750–1840*. Cambridge University Press, Cambridge, UK

Winch D 1996 *Riches and Poverty: An Intellectual History of Political Economy in Britain, 1750–1834*. Cambridge University Press, Cambridge, UK

K. Tribe

Political Economy in Anthropology

While in the history of economics, political economy is a widely accepted description of the general style of analysis of national wealth and international trade that emerged in the early nineteenth century in the writings of classical economists such as Adam Smith and David Ricardo, the term has come to be used in anthropology to refer to a variety of theoretical approaches that developed in the 1970s and shifted the focus of anthropological analysis away from isolatable local systems and toward 'large-scale regional political/economic systems.' In this shift, research now took 'the form of studying the effects of capitalist penetration upon these communities' (Ortner 1984, p. 141).

Although the term is sometimes applied only to developments within North American anthropology, it is misleading to see political economy in this sense as a description of a distinctively American 'school' limited to the 1970s. At least some practitioners had examined such broader links much earlier. Moreover, at around the same time, anthropologists elsewhere were also beginning to investigate more closely the links between global capitalism and the local social systems they had been trained to analyze. Doubtless American efforts in this direction were distinguishable because of their greater debt to the work of Andre Gunder Frank and the sociologist Immanuel Wallerstein on the capitalist world system, as well the influences on them of earlier materialist tendencies within American anthropology.

But, particularly in France and Britain, and often under a rather different set of theoretical influences—most notably from contemporary Marxist theorizing—anthropologists were also reexamining the relationships between capitalism and what were often described as various precapitalist modes of production or social formations in Asia, Africa, and Latin America. Any assessment of this general move on the part of anthropologists from the 1970s to engage with global economic factors and their influence on the local groups anthropologists were more used to studying must then examine both of these developments.

1. *Political Economy, Anthropology, and Global Capitalism*

The development of political economy in North American anthropology was linked to a more general intellectual and cultural challenge to America's imperial expansion, itself stimulated by opposition to the war in Southeast Asia. Consequently, political economy in the American context was on the one hand theoretically somewhat eclectic, and on the other, concerned primarily with issues of global capitalism. Given the tendency of anthropologists to engage with regions, societies, and cultures outside or on the margins of North Atlantic capitalism, much of the energy of political economists in America was devoted to analyzing the impact of capitalism on individuals and communities on what was sometimes called the 'periphery' of the capitalist system. This global emphasis in their approach to more traditional anthropological concerns, moreover, made American political economists less likely to classify these marginal individuals and groups as *pre*capitalist. They were consequently also less interested than their French counterparts in any attempt to develop a general theory of modes of production, seeing this as being of little relevance to the concerns of social and cultural anthropologists doing field studies in the present.

Materialist analyses were well established in American anthropology some time before the emergence of political economy, although it could hardly be argued that these were based on any genuine engagement with, or even sympathy for, the work of Karl Marx himself. This materialist current is commonly associated with the name Leslie White, who is assumed to have been familiar with work in the Marxist tradition, if not with Marx's own writings, although he was prevented from acknowledging the link explicitly because of a generalized American hostility to Marx. Marvin Harris was an influential advocate of materialist analyses from the 1960s, although, unlike White, he was openly hostile to Marxist theorizing, preferring a form of techno-ecological determinism that was rather nearer to that of White.

Closer to the interests of the new generation of political economists was the work of Julian Steward. Steward had formulated a more open, 'multilinear' theory of evolution. He also headed the Puerto Rico project, unusual for the ways it tackled issues to do with social change and colonial history as central concerns of anthropology.

But more than anyone else, a group of anthropologists studying in New York in the 1940s—including subsequently well-known figures such as Eric Wolf, Sydney Mintz, Elman Service, and Stanley Diamond—became the key figures in anthropology's direct engagement with colonialism and global capitalism from within a materialist perspective, an engagement that subsequently attracted the label 'political economy.' Wolf and Mintz in particular became increasingly concerned with the implications of European global expansion for the anthropological analysis of peasant communities in the Caribbean and South America. Both wrote anthropological monographs that were unusual for their discussion not only of local

systems of culture, kinship, and social organization, but of the formative links established in the long history of European colonization between local communities and the centers of the world economy in Europe and North America. Later, both also produced influential general works on the development of a world capitalist system, paying particular attention to the interconnections established in this process between Europe and North America on the one hand, and the only apparently isolated peoples on its periphery on the other (Wolf 1992, Mintz 1985).

Others, from both the New York group and outside it, contributed, if less directly, to the development of political economy. Unlike many materialists in American anthropology, Stanley Diamond made important contributions in Marxist theory, founding the journal *Dialectical Anthropology* in the 1970s, and working on a critical/phenomenological Marxism that contrasted rather sharply with contemporary Marxist theorizing in France. Eleanor Leacock made her own contribution, both to the development of a Marxist anthropology, through a pursuit of themes of kinship, family, and gender raised in Engels' *Origins of the Family*, and to political economy in some important work on the impact of the fur trade on indigenous North Americans.

Two further influences served to broaden the popularity of such analyses in American anthropology by the early 1970s. First, there was the publication of the very influential early works of the dependency theorist Andre Gunder Frank and, shortly thereafter, the sociologist Immanuel Wallerstein (see Frank 1969, Wallerstein 1974). What this work brought in particular was a more sophisticated understanding of the imperatives of the global economy, and the compatibility, indeed even functionality, within it of apparently noncapitalist economic and political forms. Anthropologists such as Jane Schneider and Peter Schneider, Katherine Verdery, Gavin Smith, Carol Smith, William Roseberry, and others associated with the new political economy 'school' did not necessarily accept wholeheartedly the 'capital-centrism' of these more global approaches, but instead entered into a sustained and fruitful debate with them over the global versus the local determinants of socioeconomic systems in what Wallerstein called the 'peripheries' of the world capitalist system.

The second formative influence, particularly on materialist anthropology in America in this period, was the growing movement against America's military intervention in Indochina. This contributed greatly to the radicalized political and intellectual atmosphere on North American campuses, where a new generation of anthropologists was being trained. But it also has specific implications for anthropology, raising as it did the possibility that the discipline itself could be implicated in the project of imperial. In 1968, in an article that generated a heated debate, one anthropologist went so far as to describe anthropology explicitly as a 'child of imperialism' (Gough 1968). Similar concerns led to significant conflicts within the professional anthropological body itself (the American Anthropological Association), which debated at its annual meetings the involvement of anthropologists in covert operations.

2. *The French Marxist Heritage*

If in North America the main parameters of political economy were laid down by the engagement of anthropologists with a predilection for materialist explanation with issues arising out of the development of a world capitalist economy, in Europe the situation was somewhat different. This is not to say there was no interest in the relationship between global capitalism and local socioeconomic organization. On the contrary, this central concern in American anthropology found important parallels, especially in France and Britain. Here, however, the debates were informed by a far more theoretically sophisticated attempt to develop a Marxist theory appropriate to the analysis of anthropological materials, a level of theoretical concern largely absent in contemporary American anthropology. And the center of this reworking of Marxism was to be found in France.

Developments in French anthropology were subject to three distinctive influences. The first was the opening up of the Marxist left, in part spurred by the developing Marxist critique of Stalinism which, among other things, produced a space for Marxist intellectuals, both inside and outside the Communist Party, to criticize the kinds of mechanical and evolutionist interpretations of Marx that up to then were the only versions permitted by Marxist parties. Second, and related to this, the Sino-Soviet conflict inspired at least some Marxist intellectuals and political activists to search for alternative models of both socialism and the ways it could be achieved. Third, unlike the case in America, French anthropology was dominated by the imposing intellectual figure of Claude Lévi-Strauss, and thus it is not surprising that when French anthropologists turned to the work of Karl Marx for inspiration, they developed a structural version of Marxism highly critical of the existential and humanist Marxism whose best-known exponent in France was Jean Paul Sartre. It is not especially surprising that the two main 'schools' in French Marxist anthropology—the first associated closely with the theories of the communist philosopher Louis Althusser, and the second with the structural Marxism of Maurice Godelier (a student of Lévi-Strauss)—while being critical of mainstream 'mechanical Marxism,' should both also have been equally if not more critical of the main alternative. While these ideas may have been influential in the wider student movement, Althusser's anthropological admirers, most notably Emmanuel Terray and Pierre Philippe

Rey, along with Godelier and his students, all sought to reformulate a Marxism that was 'scientific,' based on Marx's 'mature' work, and avoided what they took to be the pitfalls of an approach based on humanistic assumptions about a universal human subjectivity thwarted by capitalism.

As this suggests, for structural Marxists, rethinking the Marxist heritage meant going back to the work of Marx on capitalism, and what they found in doing so was the beginnings of a general theory of historical materialism within which anthropology could play a very significant role. That role was to be the formulation of a general theory of modes of production, and the laws of transition from one to another. Contra some of their critics, this was not seen as a mere exercise in antiquarianism, since such a theory would provide, it was argued, the theoretical basis for a political practice aimed at stimulating not just a transition to communism, but the shaping the nature of the communist future as well.

It is suggested that there was no single Marxist approach in French anthropology, but instead a collection of competing and often antagonistic camps. And yet, in spite of differing allegiances—to Althusser, to Godelier's version of structuralism, etc.—at least when viewed from the outside, French Marxist anthropology did seem to constitute a fairly unified discursive field for debate.

French Marxist anthropologists made important contributions to a number of areas of concern to both Marxism and anthropology (for an overview, see Kahn and Llobera 1981). A good example is provided by the debates over the nature of lineage-based societies in Africa that were sparked off by the early publications of Claude Meillassoux (see, for example, Meillassoux 1964, 1967). Scholars, including most centrally, Meillassoux himself, Emmanuel Terray, and Pierre Philippe Rey (both former students of Louis Althusser), as well as Maurice Godelier, Arienne Deluz, Pierre Bonte, and others, used Meillassoux's material on the Gouro of the Ivory Coast, as well as ethnographic data on other uncentralized African societies, to debate the nature of most of the basic concepts of historical materialism—class, mode of production, productive forces and relations of production, exploitation, articulation, social formation, patriarchy—as well as the central theoretical tenet of historical materialism—economic determinism (or what one participant called Marx's hypothesis of economic determinism in the last instance). This important debate, more than any other, serves to illustrate much of the potential theoretical benefits, as well as some of the main pitfalls, of the project of a Marxist anthropology as it was conceived by these scholars, namely a project to recover, by means of a careful reading of Marx's own work, his 'mature' theoretical project, and to examine its more general validity in situations largely unknown to its originator. Moreover, the debate added considerably to the understanding of more traditional anthropological problems to do with the nature of authority and social cohesion in kinship-based societies, and societies without centralized political systems.

3. *Marxism, Articulation, and Political Economy*

From the perspective of the study of the linkages between global capitalism and local social and economic systems, the most important effect of these debates was in contributing to the formation of a theory of what some participants called the 'articulation of modes of production.' Drawing on the theoretical work of Althusser, Etienne Balibar, and Maurice Godelier, anthropologists, and others with an interest in the concerns of political economy, produced analyses of the relations between the local and the global understood as an articulation of capitalism with different precapitalist modes of production. It is not often unappreciated that exploring this relation was always a central concern of Meillassoux's in a series of publications, including his early monograph on the Gouro. A sophisticated Althusserian analysis of articulation is to be found in the work of the brilliant Marxist anthropologist, Pierre Philippe Rey (see especially, Rey 1973). And other important contributions to this project were made in Britain by a variety of scholars influenced by French structural Marxism, including Ernesto Laclau, Barry Hindess, Paul Hirst, Talal Asad, and others, and by a group associated with the journal *Critique of Anthropology* including Felicity Edholm, Olivia Harris, Joel S. Kahn, John Gledhill, Josep Llobera, Stephen Nugent, and Maila Stivens. Mention needs also to be made of the important work of Jonathan Friedman on Marxist theory, anthropology, and globalization.

4. *Beyond Political Economy*

The central concerns of political economy have certainly not disappeared in anthropology. But it would be fair to say that in general within the discipline there have been significant shifts at the level of theory and analytic style, as well as at the level of the selection of problems and issues for research, so that very few anthropologists could now be described as political economists. The reasons for this declining popularity of political economy are varied. There has certainly been a recognition that earlier attempts to theorize the relationship between the global and the local in either a dependency/world systems mode or by means of a concept of the articulation of modes of production suffered from problematic theoretical tendencies. These include an overly functionalist understanding of the needs of capitalism, a unidimensional economism, an inattention to the spheres of meaning and agency, and so on. Political economists have, like many others,

been criticized for their eurocentrism, and for failing to appreciate that race, culture, and gender are not mere epiphenomena of social systems. Like those of many others, the truth claims of political economists have been subjected to wideranging critique by scholars influenced by poststructuralism, postmodernism, feminism, and postcolonialism.

Social and cultural anthropology has been affected profoundly by the critiques of those associated with the so-called 'writing culture school.' This has undermined claims about the reality of those other lives constructed in anthropological texts, by the demand for greater degrees of reflexivity, and by the placing of problems of cultural, religious, gender, racial, and national identity at the forefront of issues for investigation. Political economy which, despite the name, doubtless focused on economic issues more than any other, naturally was affected by these shifts. Finally, European versions of political economy, precisely because of their greater concern with issues of (Marxist) theory, have been damaged by the general intellectual and political rejection of Marxism, particularly since the late 1980s.

While political economy in the various forms that it took in Europe and North America in the 1970s and early 1980s will probably not regain the central role it once had in the discipline, for better or worse it would be misleading to pronounce it dead and buried. The tendency to see in the rise of Foucauldian and poststructuralist paradigms a complete rejection of, particularly, the basic conceptions of structural Marxism, is probably misleading. It is possible to argue that these merely take the structural Marxist project one step forward, from a critique of ideology, to a critique of discourse *tout court*.

Certainly, some of the central concerns of political economy have reemerged in more recent attempts by anthropologists to theorize the phenomenon of globalization. An anthropology seeking its roots in a variety of new social movements is very much on the agenda of anthropologists styling themselves as 'radical' and/or 'committed.' Some of the concerns of an earlier generation of political economists are evident in more recent feminist anthropology, as well as in various attempts to construct an anthropology of human rights. Recent attempts to forge a link between critical social theory and anthropology through various reworkings of the concept of modernity may be seen as an attempt to correct the economistic and unidimensionalist failings of the structural version of Marxism. There are even apparently once again calls to return to Marx. It remains to see how these will develop in anthropology in coming years.

See also: Dependency Theory; Economic Anthropology; Globalization and World Culture; Globalization, Anthropology of; Globalization: Political Aspects; Marxist Social Thought, History of; Political Anthropology; World Systems Theory

Bibliography

Frank A G 1969 *Capitalism and Underdevelopment in Latin America: Historical Studies of Chile and Brazil*. Monthly Review Press, New York

Gough K 1968 Anthropology: Child of imperialism. *Monthly Review* **19**(11): 12–27

Kahn J S, Llobera J R 1981 *The Anthropology of Precapitalist Societies*. Macmillan, London

Meillassoux C 1964 *Anthropologie Économiques des Gouro de Côte d'Ivoire*. Mouton, Paris-Le Haye

Meillassoux C 1967 Recherche d'un niveau de détermination dans la société Cynégétique. *L'Homme et la Société* **6**

Mintz S W 1985 *Sweetness and Power: The Place of Sugar in Modern History*. Viking, New York

Ortner S 1984 Theory in anthropology since the sixties. *Comparative Studies in Society and History* **26**: 126–66

Rey P P 1973 *Les Alliances de Classes*. Maspero, Paris

Wallerstein I M *The Modern World-system, Vol. 1*. Academic Press, New York

Wolf E R 1982 *Europe and the People Without History*. University of California Press, Berkeley, CA

J. S. Kahn

Political Elites: Recruitment and Careers

The question of who rules is always central to political science precisely because power is never evenly distributed. At any given time, some individuals and subgroups in society will have more power than others. While most non-Marxist theorists tend to agree that concentrations of power are inevitable, their understandings of how individuals or groups gain, maintain, and distribute influence differ widely. Further, despite the common usage of the term elite, theorists fail to agree on an appropriate definition.

Some descriptions of elites focus on the role and effectiveness of a society's leaders. Suzanne Keller, for example, defines elites as 'a minority of individuals designated to serve a collectivity in a socially valued way' (1963, p. 4). Alternatively, Eva Etzioni-Halevy (1993), following in the tradition of Lasswell (1936), Lasswell and Kaplan (1950), and Sartori (1976), grounds her definition of elites in terms of distribution of resources, explaining that 'resources are mostly (though not exclusively) accumulated in organizations, and elites are thus mostly (though not exclusively) people who control such resources through incumbency of certain top positions in organizations' (p. 221). Taking still another tack Dogan and Higley define political elites as 'the holders of strategic positions in powerful organizations including dissident ones, who are able to affect national political outcomes regularly' (1998, p. 15, Dogan 1989). In their view it is elite adjustment and compromise which creates stable political orders, even under conditions in which the population as a whole is characterized by

sharp ethnic and cultural divisions, a perspective which is shared by Luther and Deschouwer (1999).

Etzioni-Halevy's definition, which we follow in this article, allows analysts to include among elites, movements which are attempting to change the structure of leadership in the social and/or political order. Thus African Americans holding important positions in, say, the National Association for the Advancement of Colored People (NAACP) will be included among political elites.

1. *The Study of Elites*

Vilfredo Pareto (1848–1923), Gaetano Mosca (1858–1941), and to a lesser extent, Roberto Michels (1876–1936) developed the underpinnings of modern elite theory. In part they were reacting to a Marxist analysis which insisted that, in bourgeois society, basic divisions occurred between two social classes, the bourgeoisie and the industrial workers (Pareto 1935, Mosca 1939, Michels 1915). Politics and political power were perceived as effects of these arrangements. Formal or nominal political leaders of all societies were agents of the dominant class and were not free to act on their own in any but matters of relatively little consequence. By the late twentieth century, Marxists had changed their orientation somewhat, allowing the state and political leadership some freedom to act upon a few important matters, but Marxist theory, even in its present tattered form continues, basically, to urge the dominance of economic class and the mode of production as the prime moving forces (Sitton 1996).

Both Mosca and Pareto insisted that political elites play an independent role in determining the nature of regimes. They argued that all societies are hierarchical, dominated at any given time by one or more elites which, eventually, will be replaced by others. Dominant elites are political elites and will remain in power as long as they retain their unity and character.

Elite theory has changed a great deal since its initial formulation but many of the same arguments continue. Thus populists and quasi-Marxist analysts such as C. Wright Mills (1959) and William Domhoff (1990) (see also Zweigenhaft and Domhoff 1998) continue to insist that America and other capitalist societies are ruled by powerful socioeconomic groups (capitalists and the military for Mills; WASP capitalists and their dependents for Domhoff). More recently feminists and race theorists have maintained that gender or race in conjunction with class are the dominant forces in society, and consequently political elites' thoughts and behavior emerges from their economic, or racial, or gender status (McClain and Garcia 1993, Carrol and Zerilli 1993).

Alternatively, scholars like Robert Dahl (1982), Nelson Polsby (1980), Theda Skocpol (1979), Evans et al. (1985), Gabriel Almond (1978), Etzioni-Halevy (1993), Dogan (1989), and Keller (1963) argue that the political order and political elites play a significant, independent role and are not merely the captives of class, gender, or racial patterns. Indeed as noted earlier, Dogan and Higley (1998) and others go so far as to argue that it is the character of political elite conflicts and settlements which determine the nature and stability of the social order.

A number of social scientists, including Robert Putnam (1976), Samuel Eldersveld (1989, Eldersveld et al. 1995), Eric Carlton (1996), and Ronald Ingelhart (1997), have conducted empirical studies of elites around the world and through time. Their work permits some generalizations based on the evidence but does not resolve key issues concerning the structure of power. There is still little agreement as to the role of political elites as compared with other elite groups, or to what extent gender and race need to be brought into the analysis of elite organization and power.

2. *What Kinds of Elites?*

While Marx and Engels' notion of an anarchic, communal, totally egalitarian society at the dawn of history, in which women played key social and political roles, seems to have little basis in fact, the earliest groupings of human beings were members of relatively egalitarian, though male dominated, segmented societies. With little societal power available, power differentiation was, almost by definition, limited (Service 1975). Historically, the next human developmental stage involved the creation of pyramidal systems of social organization. This pattern characterized the most advanced societies down to the fourth or third millennium BCE and generally involved a number of groups or villages loosely organized under the authority of a paramount chief. Such chiefs often emerged as charismatic leaders who were successful in the hunt and/or war and whose power waned as they aged.

A new step in societal evolution, generally associated with the further development of agriculture and its consequent population growth, was the emergence of patrimonial societies, dominated by one individual (king or chief) whose authority was generally based on tradition supported by religious myths (Weber 1978). Such societies often attempted to institutionalize succession by hereditary right as a means, theoretically, of overcoming succession crises, and ensuring that the present chief or king could pass authority on to his male heirs. Historically, however, the hereditary mechanism for choosing rulers has not been very effective in maintaining social peace (Burling 1974).

Patrimonial kingdoms also took the first steps in the differentiation of political institutions with at least some authority devolving to subordinate officers whose positions were based on lineage and/or personal or contractual fealty. The expansion of power

available to the new social organizations led to greater asymmetries in the society. It is in patrimonial kingdoms and, even more so, in the great bureaucratic empires of the ancient world, i.e., China, Egypt, Rome, etc., that one can begin to analyse issues in terms of mutually supportive or competing elites (the army, the bureaucracy, religious elites) who shared at least some power with the ruler (Eisenstadt 1963). These societies, too, selected leaders on hereditary principles for the most part, though rulers were not infrequently deposed by the military or bureaucrats. Indeed military control of the succession came to characterize the Roman Empire by the third century CE. Military or bureaucratic leaders seized power to improve their own status and that of their clan or family. It was partly for this reason that Chinese and Byzantine emperors often assigned eunuchs to key bureaucratic and military roles in the empire (Balch 1985). Since they lacked direct descendents, eunuchs were considered more reliable instruments of the ruler.

3. *Political Elites in the Modern World*

In the modern world societies become more differentiated and patrimonial orders based on traditional authority gradually give way to authority derived from the beliefs in the legitimacy of 'rational laws' in the Weberian sense, agreed to by the people. Not only does the emergence of democratic structures (polyarchies to use Dahl and Lindblom's (1953) felicitous expression) increase the number of power centers and the complexity of power relations, but the issue of how power should be distributed becomes more central to political discourse. At the same time patterns of recruitment to and career development in elite institutions change dramatically.

Some groups which traditionally held political power lost it. For example at one point religious leaders had to be counted among political elites. Today religious leaders may still play a significant role on the political scene though, in the Christian world, their impact on political outcomes is far more limited than it once was. Indeed the process of secularization has proceeded in Europe, and, even the more religious United States boasts a fairly strict separation of religious and political institutions. In some Muslim countries religious leaders play a much more significant role. However, they are not, with the exception of Shi'ite Iran, members of the political elite *per se*, with the power to formulate and enforce policies (Tschirgi 1994, Zartman 1980).

There are today, also, a few patrimonial regimes dominated by particular lineages: Saudi Arabia, Jordan, and the Gulf Emirates come to mind as well as a few small African kingdoms. As traditional authority has come to be replaced by the notion that governments should be based on the rule of the people (which emerged in Europe as part of the liberal tradition), most regimes, even when not democratic, pretend to be so, basing their authority on mass parties, often dominated by charismatic leaders such as Hitler or Mao Zedong and his successors. Charisma plays an even more important role for leaders who have not institutionalized their power. One clear example is Fidel Castro, who rules directly like a traditional Latin American *caudillo* (chief), eschewing a mass party (Aguila 1990).

Military leaders, who emerge from time to time claiming to speak for the people, and/or 'to restore order' continue to play an active role in such countries as Myanmar (formerly Burma), Pakistan, Turkey, Indonesia, and a number of African and, from time to time, Latin American countries. In Turkey the military considers itself to be the inheritor of the policies of Mustafa Kemal Atatürk and has intervened in politics on a regular basis to preserve a secular state, usually withdrawing once the crisis was over. It is currently highly suspicious of a Muslim revival with important political overtones (Gole 1997). The military has played a similar role in other Muslim countries, such as Algeria, where it has been the prime competitor of fundamentalist groups. Of course the military plays an influential role in such countries as China and North Korea though not officially part of the governing structure of either regime. Military influence and indirect connection to policy making is more likely today than direct military rule over a political system which, when it occurs, typically does not last too long. In long-established and strong parliamentary regimes (as in Europe and the United States), the military role in political decision making is very limited.

4. *Elites in (Mostly) Modern Democratic Systems*

In relatively polyarchic systems with successfully functioning parliamentary or presidential regimes, legislatures elected by the populace play key roles in making binding decisions about the distribution of resources. In such societies election to office is mediated by political parties and the character of the electoral system is important. In both parliamentary and presidential regimes the key elite players tend to be lawyers and other professionals such as teachers and retired bureaucrats and politics becomes a full-time profession, replacing the old English notion of the amateur politician (King 1981). In presidential regimes the chief executive (whatever the formal title) exercises considerable power independent of the party to which he/she is attached and has, in the past, based more of his authority on charisma. Indeed, in the United States, the president, while not independent of his party, draws much of his authority from his unique relation with the mass media. Arguably the political roles of American political parties have substantially diminished in the face of the media revolution.

In parliamentary regimes, however, chief executives (prime ministers, premiers) are selected from within the party and generally are leaders of a group of almost equals and exercise less personal power than presidents. Of course this depends partly on the particular nature of the party system. Within the framework of a relatively disciplined two-party system, party members are more likely to adhere to the party line than in regimes characterized by multiple, usually weaker, parties and coalition governments. Exceptionally, during their heyday, the Communist parties of France and Italy were, despite multiparty systems with relatively weak political parties in each country, quite able to retain control over party members. Even in parliamentary systems with relatively disciplined parties, the role of the media has heightened the charisma and power of the party leader (especially when in office) above that of his closest associates. This results in greater identification of the Party with the leader (Dogan 1989, Kamarck 1995, Linz and Valenzia 1994, Mainwaring and Valenzuel 1998, Patterson and Mughan 1999, Trend 1997, Yesilada 1999, Riggs 1988).

Political parties are present even in authoritarian regimes such as the People's Republic of China. They were the preferred mode of political organization in most other Communist countries and in many African states. Generally, such parties have been instruments of mobilization and control allowing relatively little freedom. They have also been the central loci of power in the society. There have been some partial exceptions. In Mexico the PRI (The Party of the Institutionalized Revolution) effectively dominated politics but opposition was permitted and the party did not attempt to control the whole society. The Mexican system has further opened up in recent years and the one-party system is probably in a state of terminal decay. Also, the Soviet Union and its former satellites have, at least temporarily, transformed themselves into more open societies often with a substantial circulation of elites. Under the Soviet regime, and still in China, advancement in political power required working one's way up through the party with the aid of powerful patrons whose fall from favor could seriously affect one's career (MacFarquhar 1993, Li and Bachman 1989, Lane 1995).

In well-established parliamentary regimes the success of those seeking elite status depends on party position, seniority, and effectiveness in parliamentary work including the give and take of debate. Party loyalty generally remains the requisite condition of advancement in parliamentary regimes with disciplined party systems. Loyalty is less important in the American presidential regime. There power in the legislature is partly a function of seniority, good committee assignments, service to constituents, and a good relationship with the media.

In most of Europe, party organizations are relatively weak and the number of active party militants small. In less developed countries parties are even weaker and mostly quite short-lived. Parties often are the personal vehicles of individual charismatic leaders. After the leader passes from the scene, such parties usually fall apart to be replaced by still another 'personal' party. Political power in such countries is actually heavily dependent upon 'clientelist' relationships among various political families.

In modern democracies politics has increasingly become a career in which those aspiring to office start early and remain long, often beginning in local government (Eldersveld et al. 1995). With the passing of distinctly working-class socialist parties in Europe, political leaders are even more likely than they were in the first half of the twentieth century to come from middle-class backgrounds and to boast college or university degrees (Dogan 1989, Putnam 1976).

Money plays a relatively important role in elections in capitalist societies even when stringent limitations are placed upon the amounts parties are permitted to raise. However, individual wealth has not provided a passport to office, even in the United States where considerable freedom in the use of personal and contributed funds is permitted. Middle- or lower-middle-class aspirants to office can usually obtain financial support from backers who, for one reason or another are willing to bankroll them.

Among the unique elements of modern, relatively large complex societies is their heavy dependence upon highly differentiated bureaucracies. Despite important elements of bureaucratic structure in traditional empires, bureaucracies of the modern type never developed in these countries. High level bureaucratic officials, as Weber pointed out early in the twentieth century, are among key political elites (Weber 1978). They have played especially important roles in societies such as the Soviet and Chinese Communist regimes whose theoretical goal was to eliminate them. Indeed as late as the 1960s Mao Zedong's cultural revolution was designed to abolish or at least sharply reduce bureaucratic power. Mao had a point. Bureaucrats were blunting the cutting radical edge of the Chinese revolution for, as Weber argued, bureaucrats tend to routinize decision making. Mao did not succeed, for Weber was also correct in asserting, when many social theorists thought otherwise, that socialism could only increase the size and power of bureaucracy.

Weber, however, was wrong in implying the modern political order would eventually be dominated by bureaucratic elites. It is true that carrying out legislative mandates always involves administrative discretion. However, bureaucracies have never completely dominated a political system, with the possible exception of Japan in the decades after World War II (Van Wolferen 1989).

The influence of bureaucracy is heightened when the technical training of an elite corps of bureaucrats emphasizes their authority and power in the state. The

second empire in Germany (1870–1918) was characterized by the very significant role of the bureaucratic elite as was France under the Fourth Republic (1946–58). In the latter case the substantial power of the bureaucracy was partly a result of the inability of the party system to create a stable executive (Koh 1989, Peters 1978, Suleiman 1984, Farmer 1992, Thakur 1981).

For different reasons England and the United States have never created a bureaucratic elite of the continental European type. In the US, especially, the courts have added a layer of control above that created by Congressional oversight and presidential authority, and have insisted, in many cases, upon public consultation in bureaucratic decision making. Indeed, the role of courts has become so significant in many countries (though not as significant as in the United States) that some scholars now speak of the 'judicialization of politics.' Such judicialization has strongly affected bureaucratic decision making (Vallinder 1994).

In many modern political regimes, e.g., France, Japan, England, etc. higher level bureaucrats are recruited directly from the university for bureaucratic careers and often retire early to enter business or run for political office. High office in the United States federal bureaucracy is more permeable through in-house promotion and lateral recruitment. This gives American bureaucratic recruitment a more democratic cast than that of most other advanced industrial nations.

5. *Issues for the Present and the Future*

Political elites in advanced industrial societies do exercise more power in the sense of producing desired outcomes than do those in less developed countries. Elites in the poorer countries of Latin America, Africa, and parts of Asia often behave as if they are creating new public policies and yet nothing changes. This occurs because the organizational capacity of the society is limited and thus there is simply not sufficient power to carry through new initiatives. Promising beginnings frequently end in failure, too often frustrating, and eventually corrupting, those elites who begin as sincere reformers. Individuals or political groups may dominate the political orders of poorer societies but frequently they can change very little because the population lacks the capacity to respond effectively to new initiatives. Similarly efforts by richer more powerful nations to assist in or compel change are usually unsuccessful.

In general the role of women in public affairs has been very limited. In western Europe and the United States increasing numbers of women are becoming part of the political elite, though their role remains relatively small except in the Scandinavian countries. In most of the rest of the world their role is still smaller. Indeed the representation of women in official governmental organs dropped when the regimes in Communist states were overthrown. However, the offices they had held in various ineffective legislative bodies did not entail real authority so that the loss is more symbolic than real.

The distribution of power among racial and ethnic groups in increasingly multiethnic societies is very likely to be a major source of conflict in the twenty-first century. How such conflicts are to be handled is an issue which, except in a few cases, has not been dealt with effectively.

See also: Charisma and Charismatic; Charisma: Social Aspects of; Community Power Structure; Elites: Sociological Aspects; Leadership in Organizations, Sociology of; Leadership: Political; Leadership, Psychology of; Mobility: Social; Pareto, Vilfredo (1848–1923); Pluralism; Power in Society; Power: Political; Presidential Personality; Presidential Popularity; Social Mobility, History of; Social Stratification

Bibliography

Aguila J M del 1990 Development and revolution in Cuba. In: Wiarda H J, Kline H F (eds.) *Latin American Politics and Development*, 3rd edn. Westview, Boulder, CO, pp. 420–54

Almond G A, Powell G B Jr 1978 *Comparative Politics*, 2nd edn. Little Brown, Boston, MA

Balch S H 1985 The neutered civil servant: Eunuchs, celibates, abductees and the maintenance of organizational loyalty. *The Journal of Social Biological Structures* **8**: 313–28

Burling R 1974 *The Passage of Power; Studies in Political Succession*. Academic Press, New York

Carlton E 1996 *The Few and the Many: A Typology of Elites*. Scolar Press, Brookfield, VT

Carroll S J, Zerilli L M G 1993 Feminist challenges to political science. In: Finifter A W (ed.) *Political Science: The State of the Discipline II*. American Political Science Association, Washington, DC, pp. 55–76

Dahl R A 1982 *Dilemmas of Pluralist Democracy: Autonomy vs. Control*. Yale University Press, New Haven, CT

Dahl R A, Lindblom C E 1953 *Politics, Economics, and Welfare: Planning and Politico-economic Systems Resolved into Basic Social Processes*. Harper, New York

Dogan M 1989 *Pathways to Power: Selecting Rulers in Pluralist Democracies*. Westview Press, Boulder, CO

Dogan M, Higley J 1998 *Elites, Crises, and the Origins of Regimes*. Rowman & Littlefield, Lanham, MD

Domhoff G W 1990 *The Power Elite and the State: How Policy is Made in America*. De Gruyter, New York

Eisenstadt S N 1963 *The Political Systems of Empires*. Free Press of Glencoe, New York

Eldersveld S J 1989 *Political Elites in Modern Societies: Empirical Research and Democratic Theory*. University of Michigan Press, Ann Arbor, MI

Eldersveld S J, Stromberg L, Derksen W 1995 *Local Elites in Western Democracies: A Comparative Analysis of Urban*

Political Leaders in the U.S., Sweden, and the Netherlands. Westview Press, Boulder, CO
Etzioni-Halevy E 1993 *The Elite Connection: Problems and Potential of Western Democracy*. Polity Press, Cambridge, MA
Evans P B, Rueschmeyer D, Skocpol T (eds.) 1985 *Bringing the State Back In*. Cambridge University Press, New York
Farmer K C 1992 *The Soviet Administrative Elite*. Praeger Publishers, New York
Gole N 1997 Secularism and Islamism in Turkey: The making of elites and counter elites. *Middle East Journal* **51**(1): 46–58
Inglehart R 1997 *Modernization and Postmodernization: Cultural, Economic and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ
Kamarck E C 1995 Structure as strategy: Presidential nominating politics in the post-reform era. In: Pfiffner J P (ed.) *Government and American Politics, Classic and Current Perspectives*. Harcourt Brace College Publishers, Fortworth, TX
Keller S I 1963 *Beyond the Ruling Class: Strategic Elites in Modern Society*. Random House, New York
King A 1981 The rise of the career politician in Britain—and its consequences. *British Journal of Political Science* **11**(3): 249–85
Koh B C 1989 *Japan's Administrative Elite*. University of California Press, Berkeley, CA
Lane D (ed.) 1995 *Russia in Transition: Parties Privatisation and Inequality*. Longman, London
Lasswell H D 1936 *Politics: Who Gets What, When, How*. Whittlesey House, New York
Lasswell H D, Kaplan A 1950 *Power and Society: A Framework for Political Inquiry*. Yale University Press, New Haven, CT
Li C, Bachman D 1989 Localism, elitism, and immobilism: Elite formation and social change in post-Mao China. *World Politics* **42**(1): 64–94
Linz J J, Valenzia A 1994 *The Failure of Presidential Democracy*. Johns Hopkins University Press, Baltimore, MD
Luther K R, Deschouwer K 1999 *Party Elites in Divided Societies: Political Parties in Consociational Democracy*. Routledge, New York
MacFarquhar R (ed.) 1993 *The Politics of China 1949–1989*. Cambridge University Press, Cambridge, UK
McClain P D, Garcia J A 1993 Expanding disciplinary boundaries: Black, Latino and racial minority groups in political science. In: Finifter A W (ed.) *Political Science: The State of the Discipline II*. American Political Science Association, Washington, DC, pp. 247–80
Mainwaring S, Valenzuel A 1998 *Politics, Society, and Democracy: Latin America*. Westview, Boulder, CO
Michels R 1915 [trans. Paul E and Paul C]. *Political Parties; A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. Hearst's International Library Co., New York
Mills C W 1959 *The Power Elite*. Oxford University Press, New York
Mosca G 1939 [trans. Kahn H D]. *The Ruling Class*. 1st edn. McGraw-Hill, New York
Pareto V 1935 [trans. Bongiorno A, Livingston A]. (ed.) *The Mind and Society*. Harcourt, New York
Patterson S C, Mughan A (eds.) 1999 *Senates: Bicameralism in the Contemporary World*. Ohio State University Press, Columbus, OH
Peters B G 1978 *The Politics of Bureaucracy: A Comparative Perspective*. Longman, New York
Polsby N W 1980 *Community Power and Political Theory: A Further Look at Problems of Evidence and Inference*, 2nd edn. Yale University Press, New Haven, CT
Putnam R D 1976 *The Comparative Study of Political Elites*. Prentice Hall, Englewood Cliffs, NJ
Riggs F W 1988 The Survival of Presidentialism in America: Power-Constitutional Practices. *International Political Science Review* **9**(4): 247–78
Sartori G 1976 *Parties and Party Systems: A Framework for Analysis*. Cambridge University Press, Cambridge, UK
Service E R 1975 *Origins of the State and Civilization*, 1st edn. Norton, New York
Sitton J F 1996 *Recent Marxian Theory: Class Formation and Social Conflict in Contemporary Capitalism*. State University of New York Press, Albany, NY
Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge University Press, New York
Suleiman E N (ed.) 1984 *Bureaucrats and Policy Making: A Comparative Overview*. Holmes & Meier, New York
Thakur R 1981 *Elite Theory and Administrative System*, 1st edn. Sterling, New Delhi, India
Trend D 1997 *Cultural Democracy: Politics, Media, New Technology*. State University of New York Press, Albany, NY
Tschirgi D (ed.) 1994 *The Arab World Today*. Lynne Rienner, Boulder, CO
Vallinder T 1994 The judicialization of politics. *The International Political Science Review* **15**(2): 91–9
Yesilada B A (ed.) 1999 *Comparative Political Parties and Party Elites: Essays in Honor of Samuel J. Eldersveld*. University of Michigan Press, Ann Arbor, MI
Van Wolferen K 1989 *The Enigma of Japanese Power: People and Politics in a Stateless Nation*. Knopf, New York
Weber M 1978 [Roth G, Wittich C (eds.); trans. Fischoff E et al.] *Economy and Society: An Outline of Interpretive Sociology*, 2 Vols. University of California Press, Berkeley, CA
Zartman I W (ed.) 1980 *Elites in the Middle East*. Praeger, New York
Zweigenhaft R L, Domhoff W G 1998 *Diversity in the Power Elite: Have Women and Minorities Reached the Top?* Yale University Press, New Haven, CT

S. Rothman

Political Geography

Political geography is a field of inquiry concerned with the nature and implications of the evolving spatial organization of governmental regimes and formal political practices around the world. It is a widely recognized sub-discipline of geography, and it is increasingly an interdisciplinary pursuit as well. Political geography investigates why politically organized areas emerge in the places they do, and how the character and configuration of those areas both reflect and affect the social and environmental contexts in which they are situated. Whereas a political scientist or political sociologist might focus on the politics and social structure of a political entity, a political geographer would likely emphasize how and why a particular piece of the earth's surface came to be

organized as a discrete political unit, and what that unit represents and means for the peoples, institutions, and places affected by it.

1. Development of Modern Political Geography, 1890s–1970

The origins of political geography are often traced to Aristotle's effort to describe the ideal state in terms of its territory, population, and internal geographical structure (Kasperson and Minghi 1969). Its modern roots, however, lie in the writings of the German geographer, Friedrich Ratzel (1844–1904). Ratzel sought to analyze states in terms of their environmental characteristics, arguing that an understanding of the political world needed to be rooted in the human and physical character of the concrete territorial spaces occupied by states. He is most known, however, for his Darwin-inspired organic theory of the state, which treated states as competitive territorial entities vying with one another for control over parts of the earth's surface. He set forth his theory in a seminal book, *Politische Geographie* (1897), which marked the formal beginnings of the modern sub-discipline.

Ratzel's organic state theory attracted attention in both Europe and the United States (Glassner 1996). The Swedish political scientist, Rudolf Kjellén (1865–1922), used Ratzel's ideas about the territorial dynamics of state growth to develop an elaborated vision of the state as an organism, in the process giving birth to the term 'geopolitics' (see *Geopolitics*). Contemporaneously, the American naval historian Alfred Thayer Mahan (1840–1914) published an influential set of works arguing that control over sea lanes and adjacent land areas was critical for the maintenance and expansion of political power. This was soon followed by an alternative, more widely circulated geopolitical vision from the geographer-director of the London School of Economics, Sir Halford Mackinder (1861–1947). He emphasized the importance of the interior of the Eurasian landmass (the heartland) for the maintenance and spread of political power.

Given the relevance of these geopolitical contributions, and the interest they attracted, it is not surprising that the political geography of the early twentieth century was strongly identified with geostrategic thinking (Claval 1994). Ratzel and Kjellén's work became the foundation for a school of geopolitics (Geopolitik) in Germany led by Karl Haushofer (1869–1946), which eventually was appropriated by the Nazi war machine. In other parts of Europe and the United States, debates over the relative merits of Mahan and Mackinder's conceptualizations attracted considerable attention in both academic and policy circles, and led to yet other geopolitical visions, most notably those of the Dutch-born, American-educated international relations scholar, Nicholas John Spykman (1893–1943), who focused attention on the importance of the so-called rimland (coastal fringe) of Eurasia.

Despite the importance of geopolitical theorization in the early history of the sub-discipline, the first half of the twentieth century also saw expanding work on the geographical characteristics of individual states (Hartshorne 1954). The specific configuration of the world political map was undergoing substantial changes that required investigation and analysis. Geographers were being called upon to provide background on boundary disputes, the internal divisions within empires, and the impacts of new transportation technologies on political areas. World War I was a particular catalyst to such work. Indeed, the director of the American Geographical Society, Isaiah Bowman (1878–1950), was a member of the American delegation to the Paris Peace Conference where a new European political map was forged, one that reflected a combination of prevailing geopolitical visions influenced by Mackinder and understandings of the particular geographical characteristics of state territories.

During the interwar period, the study of political geography expanded modestly in colleges and universities, although relatively few geographers defined themselves primarily as political geographers. Those who did emphasized the historical territorial development and contemporary geographical characteristics of individual states. Bowman produced the first systematic political geography text of the post-World War I political order. It concentrated on the territorial problems of the new countries of the world, particularly those in Europe. Richard Hartshorne (1899–1992) saw this work as an extension of Ratzel's concern with the environmental character of state areas; his own influential work focused on the integrative and disintegrative potentials of differing state territorial characteristics and founding territorial ideologies (summarized in Hartshorne 1954).

World War II spawned increased interest in political geography offerings in some colleges and universities in the United States and parts of Europe, yet it did not lead to any fundamental reorientation of the sub-discipline. The one exception concerned the geopolitics wing of political geography, which fell into wide disfavor in some places because of its association with Nazi efforts to justify claims to territory. For decades the Soviets opposed any use of the term political geography, and in Germany itself the sub-discipline ceased to be taught altogether. In the English-speaking world, the most influential writings of the succeeding two decades dealt with geographical influences on the functional character of individual states through the study of their morphology, internal organization, social and physical homogeneity, relative location, and boundaries (e.g., Hartshorne 1954, Prescott 1965, East and Moodie 1968). Some innovative work came out of these studies, highlighting critical connections between geography and political decision making (e.g.,

Cohen 1963). For the most part, however, political geography became somewhat disconnected from conceptual and empirical developments in other parts of geography and beyond. Moreover, the positivist turn of the discipline of geography during the late 1950s and 1960s left little room for the consideration of political matters. By the end of the 1960s, these circumstances had placed political geography in a somewhat marginalized position.

2. *Political Geography in the Late Twentieth Century*

Political geography did not remain marginal for long. An important edited compilation from the late 1960s (Kasperson and Minghi 1969) provided a useful framework for teaching and research, and geographers trained in quantitative urban and economic geography began to turn their attention to the role that politics plays in the distribution of services, people, and wealth. The latter development spawned work focused at local and regional scales. Kevin Cox's (1973) seminal book on the politics of urban services gave rise to an influential set of studies in 'welfare geography' (see *Welfare State, Geography of* and *Locational Conflict (NIMBY)*) that drew on diverse theories ranging from Marxism to public choice theory. In the process, political geography came into contact with emerging 'political economy' perspectives.

A somewhat different influence on political geography came from a group of urban and economic geographers with a quantitative bent who saw the potential for the quantitative analysis of election data. Extensive work on the spatial organization of elections, voting patterns, and the geography of representation made electoral geography a widely recognized sub-specialty within political geography beginning in the 1970s (Taylor and Johnston 1979). The methods and theories of electoral geography were not closely aligned with other work in the sub-discipline (Taylor 1985), but the developing field of electoral geography widened the scope of political geography and attracted interest among a new generation of geographers. At the same time, the different scales and issues on which votes are taken encouraged a broadening of political geography's prior preoccupation with the scale of the state.

By the early 1980s, political geography was clearly emerging from its formerly marginalized status. Having shed its positivist orientation, politics was back on the agenda of geography as a whole, and political geography could therefore no longer be dismissed as peripheral to the discipline. Moreover, new work on urban, regional, and electoral themes was attracting increased attention to political geography and connecting it to issues and concerns being taken up by other social sciences. Even more traditional themes were taking on new life, with expanding interest in shifting cold-war geopolitical relations and the emergence of new regimes of control in the world's oceans (see Glassner 1996). The founding of a new journal, *Political Geography* (originally entitled *Political Geography Quarterly*), in 1982; the establishment of major research and study groups in the International Geographical Union and various national geographic organizations; and the founding of a boundary research institute at the University of Durham signaled the growing importance of the sub-discipline.

At the heart of political geography's revivification was the effort to develop a more sophisticated theoretical foundation for the sub-discipline (Reynolds and Knight 1989). Impetus for the theoretical turn came from the combination of insights being offered by the new, theoretically informed literature on urban social conflict and a growing frustration with the lack of connection between research initiatives outside political geography and more traditional work in the sub-discipline. Explicit engagement with theory took two overlapping, but sometimes conflicting forms: a concern with developing a theory of the state and a concern with theorizing the nature of the state system.

The work of Gordon Clark and Michael Dear (1984) is particularly associated with the effort to theorize the state in political geography. They promoted consideration of a variety of theories of the state and themselves advanced an approach that drew on Marxist ideas on the role of the state in the maintenance and expansion of capitalism, but which argued that the state is also the product of the institutionalized political-economic patterns and arrangements that are produced through specific state structures and functional hierarchies. Their work helped to foster wide-ranging inquires into the spatial dimensions of state policy initiatives, as well as studies of the 'local state' and its relationship to the national state. The latter was also heavily influenced by the writings of David Harvey (e.g., Harvey 1985), who focused attention on how the capitalist accumulation process was linked to the political production of spatial inequalities within cities and sub-state regions.

Although those writing on the nature of the local and national state looked to the role of large-scale structural forces, their work focused on communities existing within states (see Reynolds and Knight 1989). Some political geographers saw this as seriously limiting, arguing that intrastate processes could only be understood with reference to the dynamics and norms of the international system of states. They also argued that political geography would have little to say about international relations if it did not confront the nature and meaning of the state system. The writings of Peter Taylor—particularly his important text, *Political Geography: World-economy, Nation-state and Locality* (Taylor 1985 and later editions)—were influential in advancing this argument.

Drawing on the world-systems work of Immanuel Wallerstein, Taylor sought to situate analyses of the changing political geography of the modern world in the context of the territorial dynamics of the emerging capitalist world economy. He argued that the forms and functions of national and sub-national units are derived from the positions territorial states occupy in the global political-economic order. Taylor's particular use of world-systems approaches was not widely emulated, but his writings helped to put the very nature of the system of states at the center of the political geographic enterprise.

The new theoretical perspectives coming into political geography during the 1980s influenced both the types of issues being studied and how they were studied. The relationship between political geography and uneven development became a particularly significant focus of attention (see Smith 1984). Political geographers also turned their attention to processes of nation-building and substate nationalism, seeing these as predicated in significant ways on the nature of the state and the system of states (e.g., Johnston et al. 1988). Even some studies in electoral geography began to look at the relationship between spatial structures of power and electoral outcomes (Johnston et al. 1990).

The closing decade and a half of the twentieth century saw the emergence of new theoretical influences and a broadening of the scope of political geography (Agnew 1997). The growing influence first of structurationist, and then of post-structuralist, theories in the social sciences had clear impacts on political geography. The concern with the relationship between structure and agency that was part of the structurationist critique prompted political geographers to focus more attention on the ways in which political contexts are created through the interplay of large-scale forces and local practices. In his influential volume *Place and Politics*, John Agnew (1987) argued that political studies are generally based on the assumption that, in the modern world, local and regional identities are being replaced by national communities segmented by socio-economic differences, not by geographic sectionalism. Yet Agnew showed that this 'nationalization thesis' failed to account for the enduring impacts of local context on social and political behavior. In the process, he made a compelling case for taking the concept of place seriously in political geography (see also Johnston 1991).

The structurationist literature also brought with it growing interest in the spatial structures and arrangements that circumscribe social and political life. This offered a potential bridge between new currents of thought and longstanding efforts to analyze the development and meaning of the world political map (Gottmann 1973). An important element in building that bridge was Robert Sack's work on *Human Territoriality* (Sack 1986). Sack argued that a critical component of human social behavior was the effort to influence people and ideas through the control of specific geographical areas, and that a concern with '*human territoriality*' could therefore provide important insights into social and political processes. Sack's work directly challenged the dominant tendency to treat political territories as *a priori* spatial givens, in the process opening an important theoretical window for political geographic work focused on the nature and meaning of specific political-territorial developments.

Work on the politics of place and on territoriality has developed in conjunction with themes from the post- structuralist literature to open new vistas for political geography. The very centrality of politics and political relations to post-structuralist thought has greatly abetted this process (Painter 1995). Indeed, at some level much of human geography as currently practiced is about politics, making it difficult to establish exactly what constitutes political geography *per se*. Yet, as Agnew (1997, p. 3) has argued, 'there is nevertheless still the need for an area of study of the more formalized and institutionalized arenas of power such as the spatiality of states, interstate geopolitics, the geographies of political and social movements, and territorial conflicts involving national and ethnic groups.' These are the matters that continue to be most closely associated with the sub-discipline of political geography, and they are not particularly confining. Indeed, the recent political geography literature addresses themes that range from the role of political geographic arrangements in the constitution of gendered identities to the implications of new telecommunications technologies for the territorial integrity of states.

3. *Contemporary Approaches and Themes*

The contemporary practice of political geography echoes the diverse elements that have contributed to its development. Textbooks and research literature reveal approaches that are traceable to (a) political geography's tradition of describing the spatial configuration, geographic character, and boundaries of politically organized areas; (b) the quantitative turn of political geography in the 1960s; (c) the influence of political-economy perspectives on the sub-discipline in the 1970s and 1980s; and (d) the recent interest in post-structuralist theories of society (see Agnew 1997 for a variation on this typology). Some studies can still be neatly classified in one or another of these analytical traditions, but there is also considerable blurring among them. Nevertheless, tensions persist among proponents of these different perspectives, the most acute of which is between those who posit a world of 'real' political spaces/arrangements that we can try to comprehend through some combination of empirical and critical analysis, and those post-structuralists who

view reality as unknowable outside of the individuals and societies that construct it. The opposition between these positions has created a dualism in the political geography literature that is only beginning to be addressed (Agnew and Corbridge 1995).

The substantive topics on which political geographers are working are a product of the more general effort to make sense of the rapidly changing political, social, and economic environment of the late twentieth century. The following subsections highlight the topics that are receiving the most attention There is considerable overlap among and between these topics, but each has a core literature attached to it.

3.1 Political Geographic Assumptions of Social and Political Theory

There is a pervasive tendency in the social sciences to frame analyses of social and political developments in terms of the political territories in which they are situated. Yet the territories themselves are usually treated as little more than locational referents for whatever is being studied. A body of work in political geography seeks to expose the theoretical shortcomings of this state of affairs. Much of that work focuses particularly on the importance of understanding what is hidden and what is revealed when the modern territorial state is treated as an unproblematic spatial container for social and political processes. John Agnew in particular has argued that this creates a 'territorial trap' in which critical historical and geographical processes are ignored because the nature and meaning of territory are not problematized (see Chap. 4 of Agnew and Corbridge 1995). A number of political geographic studies support this line of reasoning by showing how the political organization of space is itself implicated in social and political processes, even shaping the very concepts of identity on which political communities are based. Work in this arena is thus tied to the larger effort to bring concepts of space into social theory.

3.2 Geopolitics and Interstate Conflict

Geopolitics is clearly back on the political geography agenda, as evidenced by the publication of a spate of recent books and the founding of a new journal, *Geopolitics*. The recent literature on the subject reveals some continuing interest in the development of formal geopolitical models, but recent work has concentrated on the changing geopolitical environment for interstate relations and on the world views and political/military priorities of key political actors. Particular attention has been paid to the emergence of the modern world economy and its impacts for the conceptualization and pursuit of geopolitics (e.g., Agnew 1998). Political geographers have also examined the locational components of conflict itself, studying where conflicts have occurred and why some areas have experienced more conflict than others (e.g., O'Loughlin and van der Wusten 1993). The interest in discourse as a window into geopolitical motives is the focus of a new literature on 'critical geopolitics' (see Ó Tuathail 1996). Influenced particularly by deconstructionist ideas, this increasingly influential literature looks at the discourse of geopolitical practice to gain insight into the ways in which spatial understandings have been developed, deployed, and manipulated in the pursuit of particular political and military objectives.

3.3 The Changing Role of the Territorial State in the Modern World

There is a lively debate in political geography and beyond about the degree to which the traditional role of the state is being undermined by the increased mobility of capital and labor, the growing power of multinational corporations, the advent of transportation and communications technologies that decrease or eliminate the friction of distance, and the rise of extra-state trading blocs and political organizations (see generally Painter 1995). Within political geography this debate finds concrete expression in studies focused on the ways in which these developments have altered the internal spatial organization and territorial character of states. Political geographers are raising questions about the changing nature and function of boundaries; the distribution of powers among and between substate, state, and superstate political-territorial entities; and the means by which governments at different scales are seeking to integrate and defend their territories. Research on such matters has been spurred by concrete changes in the world political map: the rise of the European Union; the break-up of the Soviet Union, Yugoslavia, and Czechoslovakia; and the devolution of substantial powers to substate regional authorities in countries such as the United Kingdom and Spain.

3.4 Ethno-national Challenges to the Prevailing Territorial Order

Closely related to the prior theme is research focussed on the often discordant relationship between the map of territorial states and the territorial claims of different ethno-national groups. Political geographers have long argued that the concept of the nation-state is more an aspiration than a reality. The rise of so many visible instances of ethno-national conflict over the

past few decades has moved this issue to the center of the political geography agenda. Political geographic research has focused on the competing territorial ideologies and strategies of majority and minority groups, the ways in which the political organization of territory affects intergroup interaction and intragroup identity, and the particular notions of citizenship and nationalism that have been forged in differing historical-geographical contexts (see, e.g., Johnston et al. 1988). The infusion of post-structuralist ideas into political geography has also promoted interest in the use of cultural symbols in the pursuit of territorial strategies, as well as the ways in which meanings develop around particular territories and nationalized landscapes. Work on the political geography of ethno-nationalism has highlighted the overlapping senses of territory that exist among many ethno-national groups and the need for policy approaches that recognize this state of affairs.

3.5 *The Political Geography of the Environment*

One of the clearest obstacles to the effort to confront environmental problems is the fragmentation of the planet into political spaces, few of which are meaningful ecological units. The problem is not just that ecological and political boundaries do not coincide; the very roles that territorial states play in the international state system and the world economy frequently work against collective action in the interests of the environment (Johnston 1996). Building on a tradition of examining the relationship between spaces of governance and environmental geographies, political geographers are increasingly turning their attention to environmental questions. Recent work has focused on new transnational regulation regimes and the special issues presented by environments that do not lie within the territories of individual states: the world's oceans and Antarctica. At smaller scales, political geographers have examined the circumstances giving rise to the emergence of an environmental politics in particular places, the problems and prospects of collective resource management in different political contexts, and the use of environmental strategies in the pursuit of political-territorial objectives. Some of this work is closely aligned with the emerging field of political ecology, with its focus on the role of large-scale political-economic structures in the production of local environmental outcomes.

3.6 *The Quest for Political Power and Influence*

Studies in electoral geography have long focused on the spatial dynamics of a particular type of quest for political power: that which takes place in formal elections. The geography of elections continues to attract significant attention among political geographers, but a variety of other political movements are commanding attention as well. This is a product of the increased visibility of such movements and the success of some of them in challenging long-standing political arrangements. Work in political geography has focussed on the political-territorial contexts that give rise to political movements, the spatial character of movement support, and the impacts of movement objectives on the political organization of space. A significant body of recent literature has considered how the economic, social, and political character of individual places is implicated in the rise and success of different types of political movements (Staeheli 1994). The focus on places, however, is not at the expense of larger-scale influences. Instead, many political geographers treat individual places as the product of the interaction between large-scale political and economic structures and local cultural, social, and material circumstances (Johnston 1991).

3.7 *Political Place and the Construction of Social Identity*

Closely related to the prior theme is a strand of work in political geography concerned with the ways in which social identities are shaped by the political-territorial construction of place. The plethora of new social movements that have come together around social categories such as race, gender, sexual identity, and ethnicity have attracted growing attention throughout human geography, and the recognition that such identities are rooted in territorial practices and arrangements has provided an important bridge between work in political geography and the literature of economic, social, and cultural geography (see Wolch and Dear 1989). The concept of place—as politically and socially constructed—is once again at the center of this area of research because identification with place is so often at the heart of social identity (Johnston 1991, Agnew 1997). Yet place is not necessarily presented as a benign incubator of social identity; a considerable body of literature has focused on the ways in which dominant interests create spaces/places of marginalization that reflect and reinforce existing power relations and social prejudices.

3.8 *The Emergence of New Political Spaces*

The sweeping economic, political, and technological changes of the late twentieth century are associated with the development of new political spaces that coexist, and sometimes compete, with territorial states. Many of these new spaces are derivative of the map of states in that they are either collections of states (e.g., the European Union) or sub-state regions that have recently achieved greater autonomy (e.g., Catalonia). Yet activities within and among these new state-derived political spaces are creating political spaces of

a completely new sort. These include new political spaces of cooperation along international boundaries and geographically distant political communities that are forging links outside the framework of states. Behind many of these developments is the explosion in communications and information technologies, which is fostering an increasingly deterritorialized politics that nonetheless has concrete territorial implications. The emergence of far-flung political communities united by ideology and technology poses new theoretical and empirical challenges for an area of inquiry concerned with the spatial dimensions of politics.

4. *The Way Forward*

Political geography has emerged as one of the most vibrant, active areas of inquiry within the discipline of geography, and it is likely to remain so. The political-territorial upheavals of recent years show no signs of abating. Moreover, it is likely that growing challenges to the role of the state will bring into question the tendency throughout the social sciences to treat the map of states as a static frame of reference, and will direct attention to the nature and meaning of the changing spatial organization of politics. The interdisciplinary character and impact of political geography are likely to grow with the recognition that political practices and outcomes cannot be divorced from the changing geographical spaces and contexts within which they are situated.

See also: Geopolitics; Welfare State, Geography of.

Bibliography

Agnew J A 1987 *Place and Politics: The Geographical Mediation of State and Society*. Allen and Unwin, Boston

Agnew J A (ed.) 1997 *Political Geography: A Reader*. Arnold, London

Agnew J A 1998 *Geopolitics: Re-visioning World Politics*. Routledge, London

Agnew J A, Corbridge S 1995 *Mastering Space: Hegemony, Territory, and International Political Economy*. Routledge, London

Clark G L, Dear M 1984 *State Apparatus: Structures and Language of Legitimacy*. Allen & Unwin, Boston

Claval P 1994 *Géopolitique et Géostratégie: La Pensée Politique, l'Espace et le territoire au XXe siècle*. Editions Nathan, Paris

Cohen S B 1963 *Geography and Politics in a World Divided*. Random House, New York

Cox K R 1973 *Conflict, Power and Politics in the City: A Geographic View*. McGraw Hill, New York

East W G, Moodie A E (eds.) 1968 *The Changing World: Studies in Political Geography*. Methuen, London

Glassner M I 1996 *Political Geography*, 2nd edn. John Wiley, New York

Gottmann J 1973 *The Significance of Territory*. University of Virginia Press, Charlottesville, VA

Hartshorne R E 1954 Political geography. In: James P E, Jones C F (eds.) *American Geography: Inventory & Prospect*. Syracuse University Press for the Association of American Geographers, Syracuse, NY

Harvey D 1985 The geopolitics of capitalism. In: Gregory D, Urry J (eds.) *Social Relations and Spatial Structures*. St. Martin's Press, New York

Johnston R J 1991 *A Question of Place: Exploring the Practice of Human Geography*. Blackwell, Oxford, UK

Johnston R J 1996 *Nature, State, and Economy: A Political Economy of the Environment*, 2nd edn. Wiley, New York

Johnston R J, Knight D B, Kofman E (eds.) 1988 *Nationalism, Self-Determination and Political Geography*. Croom Helm, London

Johnston R J, Shelly F M, Taylor P J (eds.) 1990 *Developments in Electoral Geography*. Routledge, London

Kasperson R E, Minghi J V (eds.) 1969 *The Structure of Political Geography*. Aldine Publishing Company, Chicago

O'Loughlin J, van der Wusten H 1993 Political geography of war and peace. In: Taylor P (ed.) *The Political Geography of the Twentieth Century: A Global Analysis*. Belhaven, London

Ó Tuathail G 1996 *Critical Geopolitics: The Politics of Writing Global Space*. University of Minnesota Press, Minneapolis, MN

Painter J 1995 *Politics, Geography, and 'Political Geography'*. E Arnold, London

Prescott J R V 1965 *The Geography of Frontiers and Boundaries*. Aldine, Chicago

Reynolds D R, Knight D B 1989 Political geography. In: Gaile G L, Willmott C J (eds.) *Geography in America*. Merrill Publishing company, Columbus, OH

Sack R D 1986 *Human Territoriality: Its Theory and History*. Cambridge University Press, Cambridge, UK

Smith N 1984 *Uneven Development*. Blackwell, Oxford, UK

Staeheli L 1994 Empowering political struggle: spaces and scales of resistance. *Political Geography*. **13**: 387–91

Taylor P J 1985 *Political Geography: World-Economy, Nation-state and Locality*. Longman, London

Taylor P J, Johnston R J (eds.) 1979 *Geography of Elections*. Penguin, London

Wolch J, Dear M (eds.) 1989 *The Power of Geography: How Territory Shapes Social Life*. Unwin Hyman, Boston

A. B. Murphy

Political History (History of Politics)

Political history is the study of how legitimate power has been defined, distributed, and justified in the past (see *Power: Political*). Two imperatives, both with ancient roots, link historiography and politics. The first is the need to affirm the identity of the political community by uncovering its roots in the past. This search for historical identity engaged both the Hebrews and Greeks at the very beginning of European culture; it moved the Romans to search for the sources of their state and prompted succeeding generations to study the history of their cultures, religions, and nations. 'No question,' wrote Jean Bodin in the sixteenth century, 'has exercised the writers of histories more than the origin of peoples' (Kelley 1998). The second link between history and politics is the desire to

find lessons in the past that might help solve the political problems of the present. 'I shall be content,' Thucydides remarked in the first chapter of *The Peloponnesian War*, 'if (this history) be judged useful by those inquirers who desire an exact knowledge of the past as an aid to the understanding of the future' (Thucydides 1996).

1. *From the Ancient World to the End of the Middle Ages*

Political history in the West begins with two masterpieces: Herodotus's history of the Persian wars, written between 445 and 425 BC and Thucydides's history of the long conflict between Athens and Sparta, which he began in 424 BC and left unfinished when he died a quarter of a century later. Products of the extraordinary burst of cultural energy that transformed fifth and fourth century BC Greece, Herodotus and Thucydides combined a deep sense of the importance of their own age—Herodotus wrote about the recent past, Thucydides about events in which he himself had taken part—with an awareness of the historian's responsibility to establish the truth about the past. Both men reflected on the problems of historical research, the reliability of eye witnesses, and the necessity of critically analyzing source material. More than any particular way of writing history, this was their most significant legacy for future historians (see *Historiography and Historical Thought: Classical Period (Especially Greece and Rome)*).

The study of history was more important in Rome than it had been in Greece, perhaps because Romans were always conscious of their complex historical relationship to the Greek culture that they had both vanquished and absorbed. Roman historians examined the problem of politics from a variety of perspectives; they wrote about leadership and institutions, military strategy and court intrigue, republican virtues, and imperial ambitions. But behind these disparate approaches was a common theme: the origins, triumph, and—eventually—the decline of the Roman state, which, as republic and empire, dominated the public life of the Mediterranean world for 700 years and then continued to haunt Europeans' historical imagination until well into the nineteenth century.

The emergence of Christianity and its adoption as the Roman Empire's official religion transformed all of western culture, including both politics and history. Christianity's most immediate contribution to political historiography was to pose the problem of church and state, which would decisively shape the theory and practice of European politics from the end of the Roman empire to the beginning of the twentieth century. Unlike the pagans whom they displaced or the Hebrew traditions upon which they built, Christians created religious institutions that were connected to, but not coincident with, the sources of political authority. 'Spiritual and secular power,' Ranke (1962) wrote, 'could approach one another and be intimately related, but they could be fused only under exceptional circumstances and for a brief period of time. Their relationships and conflicts provide one of history's most important themes.' From Eusebius's history of the church, which reflected the triumph of Constantine's conversion at the beginning of the fourth century, to Otto of Freising's twelfth-century chronicle, *The Two Cities*, which described the conflict between empire and papacy, to Ranke's own history of the popes, which bore the marks of nineteenth-century's struggles between church and state, historians searched the past for clues to the proper relationship between God and Caesar, pope and emperor, altar and throne, religious autonomy and secular authority (see *Historiography and Historical Thought: Christian Tradition*).

2. *History and the State*

Beginning in the 1490s with the French invasion of the Italian peninsula, a series of international, constitutional, and religious crises disrupted European politics and helped to create a new subject for political history. This subject was, in Quentin Skinner's (1978) concise summary, 'the concept of the state—its nature, its powers, its right to common obedience'. The concept of the state emerged from the dissolution of two closely-connected assumptions that had shaped European political thought since Augustine: first, that the secular realm of politics had to be seen as part of a sacred order; and second, that Europe's components—its free cities, ecclesiastical states, and various monarchies—were part of a single polity, governed by the universal laws of the church and sustained by the institutional legacies of imperial Rome. A good deal of medieval political theory and practice had been devoted to reconciling these assumptions with the stubborn realities of political life; in the first decades of the sixteenth century, Europeans began groping their way towards new ways of thinking about politics—and, therefore, towards a new kind of political history.

One of the first signs of this can be found in the work of Niccoló Machiavelli. As a Florentine who had been forced into exile by the collapse of his city's constitutional order, Machiavelli (see *Machiavelli, Niccolo (1469–1527)*) felt the political crisis of the early sixteenth century at first hand. In *The Prince* and *The History of Florence*, Machiavelli sketched a political world that Thucydides would have recognized immediately, a world driven by struggles for power, beset by violence, subject to forces that are often beyond human capacities to control. A similar picture emerges in the work of Machiavelli's fellow citizen and contemporary, Francesco Guicciardini, whose histories

provide an even darker picture of instability and violence. But however limited people's ability to control events, history, for both Machiavelli and Guicciardini, was a totally human enterprise, untouched by the hand of God, unfettered by a transcendent order.

Machiavelli wrote about 'the state,' but without giving it a clear institutional definition. This would be the great achievement of sixteenth- and seventeenth-century legal humanists, whose study of the law provided a new foundation for both politics and history. Of course law had always been an important part of European culture, but scholars like Guillaume Budé, Jean Bodin, François Baudouin, and François Hotman combined philological methods and historical analysis to create a new science of jurisprudence that enabled them to establish the origins, nature, and limits of the state's authority. Bodin, for example, provided the classic definition of sovereignty as the supreme source of law within a commonwealth, exercised by, but clearly independent from, the ruler's person. The work of these scholars stands at the source of one of political history's richest traditions: the study of legal institutions and constitutional development, which includes among its practitioners such giants as Montesquieu, Sauvigny, and Maine (Kelley 1970).

Law was also important to those who sought a vision of the international order to replace medieval universalism. Hugo Grotius, for example, who wrote during the Thirty Years War in the first half of the seventeenth century, argued that the international order was based on natural laws, which were divinely ordained but could be discovered by examining the past; like the French legal humanists, Grotius regarded the study of law and history as part of a single enterprise. By the end of the eighteenth century, under the influence of a new set of crises generated by the French revolution (see *French Revolution, The*) and Napoleonic conquests, a number of scholars searched for an international system that was no longer based on God-given laws but was still subject to underlying principles that the historian could discover. Heeren (1834), for instance, wrote that the 'essential property' of the European state system was its 'internal freedom, that is, the stability and mutual independence of its members. To set forth how this was formed, endangered, and preserved … is the great object of the historian.'

The historical image of a world of independent states struggling for existence within an international system was given its definitive formulation in the middle decades of the nineteenth century by Leopold von Ranke, perhaps the greatest political historian of the modern era (see *Historiography and Historical Thought: Modern History (Since the Eighteenth Century)*). Von Ranke is best known for his archival work and his insistence on the critical use of primary sources. But inseparable from Ranke the great empiricist was Ranke the moral philosopher of statehood, who regarded states as the 'ideas of God' that move through the historical order 'like celestial bodies.' His work, which combined scholarly rigor with stylistic grace, was read widely and admired throughout Europe; his methods of teaching and research were adopted everywhere (see Krieger 1977).

3. *History and the Nation*

Of all kinds of history, the history of the nation has the deepest roots and the most widely extended branches (see *Nation: Sociological Aspects*). Its first and most persistently influential expression was in the Hebrew Bible, which recounts the archtypical story of a nation's origins, defines its special historical mission, traces its survival in the face of terrible trials, and promises its eventual triumph. Their story, the Hebrews' historians insisted, is a story like no other—among the many ways in which national histories resemble one another, none is more striking than their insistence on being unique. As for so many other nations without states, this national history sustained the Jewish people's sense of collective identity through centuries of dispersal and oppression. And once the Jewish nation finally had its state, the story of the national past became—and this is also characteristic of the genre—the site of political debates about the nation's present identity and future direction (see *Historiography and Historical Thought: Jewish Tradition*).

Although national history has a genealogy that extends to the dawn of western culture, its most intellectually vigorous and politically potent versions were written in the nineteenth century when, for the first time, political history became virtually synonymous with the history of nations and nation states (see *State, History of*). The nineteenth century was, of course, the great age of historical writing, a time when historical consciousness touched virtually every aspect of European culture, from art and literature to philosophy and the social thought and even the natural sciences. For most nineteenth-century intellectuals, to think about the world was to think about historical origins, evolution, and identity. This was a time when statesmen—men like Guizot and Macaulay—wrote history and when historians— Bancroft, Droysen, Treitschke—were active in political affairs. And this was also the era when the institutional foundation for modern historical scholarship—document collections, libraries and archives, journals and reference works—was created, much of it by, and often for, national governments (see *Historicism*).

While every nation has a story to tell about itself, every national story is different, shaped by the political imperatives, cultural climate, and institutional arrangements of each national community. Nineteenth-century French historians, for instance, whose state had a unusually long and apparently uninterrupted

existence, were able to imagine a single national history. Their problem was how to find a place in this history for the trauma of revolutionary upheavals and prolonged warfare that engulfed France after 1789. For almost two centuries, French historiography was shaped by political struggles over the revolution's legacy, just as French politics was influenced by historical struggles about the revolution's character and significance. Among the works engaged by this project are such classics of political historiography as Tocqueville's *Ancient Regime and the French Revolution* and Michelet's *History of the French Revolution* (see *Tocqueville, Alexis de (1805–59)*). It was not until the 1970s that François Furet could declare that the French revolution was finally over, by which he meant that it had lost its central place in the nation's political discourse and public memory.

Each national experience poses a distinctive ensemble of questions for its historians to confront. In Britain, these are most often questions about institutional continuity rather than revolutionary ruptures. In the US, the central historical issue is the question of identity itself, which was addressed in works like Frederick Jackson Turner's classic study of the frontier in American history (see *Frontiers in History*). The problem of identity also dominated German historiography in the nineteenth century: before the creation of a German nation state, this meant a competition between different versions of the national past, after national unification, it involved demonstrating that the nation created by Bismarckian 'blood and iron' was the only authentic destination for the German people's long historical journey towards political unity. In the twentieth century German history was shaped by efforts to locate the catastrophe of National Socialism in the national past: was National Socialism, as some claimed, the destination of Germany's 'special path' (*Sonderweg*) to modernization, or was it a particularly toxic example of a broader crisis within modern culture and politics (see *National Socialism and Fascism*)?

An important element in the public life of every nineteenth century state, history was even more significant for those national communities that did not have a state of their own. For the Poles, the Irish, the Czechs, and, in the course of the twentieth century, for the peoples of the world colonized by Europeans, a shared history was the surest source of present cohesion and the brightest hope for future nationhood. After all, what else did these nations possess besides their belief in a common ancestry? Often enough, this belief had to be constructed from the scattered remnants of an imagined past—literary works discovered and transcribed by patriotic scholars, ancient sites unearthed by archaeologists, memories of lost greatness brought to life in poems and plays and popular histories. Often enough the character of the national past was bitterly contested—by Czechs and Germans in Bohemia, Protestants and Catholics in Ireland, conservatives and liberals in partitioned Poland, Muslims and Christians in the Balkans. In these parts of the world, the historical significance of John Hus, whether the town was called Derry or Londonderry, the meaning of the Polish uprising of 1863, or how to commemorate the battle of Kosovo could swiftly become matters of life or death. Every national history reflects ideological divisions and political conflict, but in nations without states, the seam between politics and history was nearly invisible, so that it was usually impossible to determine where the one began and other ended.

4. *Subversive Histories*

Although historians have often been the servants of those in power, there is also a long, rich tradition of subversive historiography, written by scholars who turned to the past in order to question the established order and suggest new paths for the future. In the nineteenth century, at the same time that historians were drawn into the service of the nation state, subversive historiography was given its most influential formulation by Karl Marx. For him, the political movements that had interested Thucydides, Machiavelli, and Ranke are surface phenomena, reflections of the deeper social conflicts from which history takes its shape and direction. Politics is merely an instrument of the ruling class, a way of justifying and defending its hold on the means of production. Traditional political historians, therefore, view the past through a distorting lens that turns the world upside down, obscuring the essential sources of change with their concern for artificial conflicts, empty slogans, and trivial events. The Marxist historian's purpose is to substitute for these ideological distortions the radical clarity of scientific materialism, which will reveal to the proletariat the roots of its oppression and thus help prepare them for their revolutionary role in history cosmic drama. From the last decades of the nineteenth century until the end of the twentieth, Marxism—in all its many forms—remained the most influential form of subversive historiography (see *Marx, Karl (1818–89)*; *Marxist Social Thought, History of*).

Well before the final collapse of international communism in 1989, Marxist historiography had begun to lose its intellectual attraction and political influence. In part this was because of the diminution of ideological energies both in the communist regimes of eastern Europe and in labor movements throughout the world. Moreover, by the 1970s, the political passions that had drawn intellectuals to Marxism during the struggle against fascism in the 1930s and against colonialism in the 1960s had cooled. The Marxist view of history was also undermined by a growing skepticism about all grand narratives, a turning away from the aspiration to grasp the meaning

of human history as a whole. And finally, Marxism declined because it could not explain very well the subjects to which more and more historians now turned, especially problems of ethnicity, race, and gender, which did not fit into the categories defined by the means of production. Not surprisingly, the form of Marxism that remained most influential in Britain, the US, and most of western Europe was associated with historians like E. P. Thompson, who tried—not always with success—to fuse class analysis with a concern for culture and politics (see *Social History*).

Since the decline of Marxism, the subversive impulse in historiography has been fed from many other sources of which the most important, at least in the last two decades of the twentieth century, was certainly the work of the French philosopher and historian, Michel Foucault (see *Boserup, Ester (1910–99)*). Like Marx, Foucault's enemies were oppression and inequality, his goals, freedom and community. But unlike Marx, he did not propose an alternative view of history or a critical science of society. Foucault's project was willfully fragmented; he called his approach to the past a genealogy rather than a narrative history. Deeply concerned about what he regarded as the repressive power of organized knowledge, he was nevertheless unwilling to abandon efforts to understand and subvert the existing order of things. Above all, this meant to understand and subvert power, which was exercised not only by the agents of the state but rather was diffused 'in the multiple forms of subjugation that have a place and function within the social organism.' Emancipation, therefore, required a struggle against all of what Foucault regarded as the 'polymorphous techniques of subjugation' (Dirks et al. 1994).

Marx (and also, if much less influentially, Freud) had sought to subvert traditional political history by showing that politics was really a reflection of deeper social (or psychological) conflicts (see *Psychohistory*). Foucault subverted the conventional image of politics from the opposite direction, that is, not by showing that politics should be reduced to something else, but rather by insisting that everything—including what was once thought of as the most intimate spheres of private life—is politically charged. Both perspectives are useful since both can illuminate new problems and neglected connections. Whether they can—and should—lead to the abandonment of the subjects that have concerned political historians since classical times is much more open to question.

5. *The Varieties of Political History in the Twentieth Century*

Scholars' interest in the past has always been influenced by contemporary crises. Thucydides was moved to write his history by the Athenians' disastrous war with Sparta, Josephus recorded the destruction of the Jewish people by the Romans, and Augustine composed his universal history under the shadow of the Vandals' sacking of Rome. In the modern era, the French revolution inspired Tocqueville, the Bolshevik revolution, Trotsky, the rise of Italian fascism, Salvemini, Nazi totalitarianism, Arendt, the world wars, Churchill—the list might easily be extended to include many other occasions when, to borrow Hegel's phrase, men's hopes were destroyed on history's slaughter bench.

The war that shattered the world order in 1914 dominated the study of international political history for a quarter of a century (see *First World War, The*). Basing their work on the documents published by all of the belligerents, historians examined both the immediate causes of the war and its deeper roots in the great power rivalries of the nineteenth century. Some of this work was obviously partisan, written to absolve the historian's own nation of responsibility; but much of it sought to strike a balance between commitment and detachment. A notable feature of this historiography was the participation of American scholars. Responding to the hopes and disappointments that attended the failure of Woodrow Wilson's vision of a conciliatory peace, these scholars—Sidney Fay, Robert Binkley, William L. Langer, Raymond J. Sontag, and many others—played an active role in controversy over the causes and consequences of the war. By the 1930s, there was a rough scholarly consensus that no one power was responsible for the war, which was thought to have come from a series of miscalculations by several different governments. In the 1960s, this consensus was shattered by the work of Fritz Fischer and his students who attempted to blame the war on Germany's reckless bid for world power. Fischer's view was bitterly attacked and then widely accepted, but no new consensus has emerged to explain what George Kennan, himself a distinguished student of international history, once called 'the great seminal catastrophe of the century' (Stevenson 1997).

In contrast to World War I, the causes of World War II did not attract much scholarly attention; virtually everyone agreed that this was Hitler's war (see *Second World War, The*). But the war's character and consequences generated an enormous historical literature, including some great works by participants such as Winston Churchill and Charles DeGaulle. Historical controversies surrounded issues such as the air war (especially the decision to use atomic weapons against Japan), wartime diplomacy, and the role of resistance and collaboration in the territories occupied by Germany and Japan. Among the historical questions about the war in Europe none are more painfully complex than those surrounding the mass murder of European Jews, conventionally known as the Holocaust. Was it, as some scholars believe, tied to a distinctive sort of German anti-Semitism? Or did it arise within the climate of violence created by the war itself? Is it best seen as a problem of German

history? Or of modernity as a whole? Could it have been prevented or at least diminished by the action of Germany's enemies? Central to the historical self-understanding of both perpetrators and victims, the Holocaust has become the century's exemplary disaster (Marrus 1987) (see *Holocaust, The*).

The origins of the Cold War, the global geopolitical rivalry between the Soviet Union and the US that dominated international politics from 1945 to 1989, were a matter of intense scholarly controversy throughout its long history. Was the Cold War, as some claimed, a defensive response by the western powers to communist aggression and expansionism? Or, as the so-called 'revisionists' argued, did the US over estimate, perhaps even manufacture, the communist threat in order to advance its own aims throughout the world? Until the collapse of the Soviet Union, both sides in this controversy depended largely upon sources from the western powers, but after 1989, the empirical basis of Cold War historiography was transformed by the opening of archives in both the former Soviet Union and its eastern European allies. As a result, historians have begun to create a new picture of international history in the second half of the twentieth century (Gaddis 1997).

Among the great themes of world history is the expansion of European power which began in the late fifteenth century and ended during the second half of the twentieth century, when the colonial states either abandoned or were expelled from their overseas possessions (see *Colonization and Colonialism, History of*; *Imperialism, History of*). European expansion was the subject of many important historical studies, including the analyses by J. A. Hobson and V. I. Lenin that inspired several generations of anti-imperialist writers. Decolonization also had a lively historiography, stimulated in part by the traumatic wars fought by France in Algeria and by the US in Indo-China. In the 1990s, historians became interested in what became known as postcolonialism, that is, the persistent and reciprocal relationship between colonized and colonizer both during and after the colonial experience itself. While many postcolonial studies take up issues that seem far distant from conventional political history—for instance, questions of gender, sexuality, and cultural identity—they are all connected to the ways in which political power is defined and distributed.

In the last decades of the twentieth century, political historians have become increasing aware of the importance of regional and loyal allegiances and institutions. A number of recent works on the revolutions of 1848, such as Jonathan Sperber's account of the Rhineland, have shown how tensions between central and local authorities affected both the origins and outcome of the revolution. Historians have also become sensitive to the historicity of the national state: Eugen Weber, for instance, has described the long and uneven process through which the French state imposed its values in the countryside, while Linda Colley, among many others, has uncovered the way in which Britain became a focus of people's interests and loyalties. The upheavals that destroyed Yugoslavia in the 1990s produced a cluster of new works on the historical sources of ethnic identity and changing nature of nationalism in the twentieth century (see *Nationalism: General*; *Nationalism, Historical Aspects of: The West*). In these studies traditional questions about state power and politics are enriched by a new concern for ethnicity, culture, and identity.

6. *Conclusion: Political History in the Twenty-first Century*

Although it no longer has the unquestioned historiographical primacy it once enjoyed, the history of politics remains a vigorous and remarkably diverse enterprise. For the general public, political history is a source of apparently endless fascination. Each year a staggering number of books, especially biographies and battle narratives, are published about subjects such as the American Civil War and World War II. Among professional historians there are signs that political history may have recovered some of the ground lost to social and cultural history in the 1970s and 1980s. An increasing number of scholars are now seeking to apply social scientific methods and concepts to the study of both domestic and international politics (see *History and the Social Sciences*). At the same time, insights from cultural anthropology (see *Anthropology and History*) and literary criticism have raised interesting new questions about the nature of political power and identity.

The history of politics is important because individuals' lives are shaped and limited by political institutions, while the very existence of their communities depends on how political power is exercised and restrained. The illusion that politics does not matter is a luxury that only those fortunate enough to live in a rich and orderly state can afford. Moreover, while the importance of national history may have declined in western Europe, for much of humanity the question of national identity remains vitally relevant; in many places people are still willing to die—and to kill—in behalf of their own version of the national story. And the emergence of a new set of problems and opportunities from the wreckage of the Cold War system (see *Cold War, The*) will certainly raise a new set of questions about the past and future of the international politics.

Does knowledge of the past help us, as Thucydides and Machiavelli and so many others believed, to master our own political problems? To be sure, few now share the confidence expressed by Lord Acton in his inaugural lecture of 1895 that 'the science of politics is the one science that is deposited by the

streams of history, like the grains of gold in the sand of a river; and the knowledge of the past, the record of truths revealed by experience, is eminently practical, as an instrument of action and a power that goes to making the future' (Acton 1957, pp. 25–26). And yet the impulse to extract lessons from the past seems irresistible; historical references and analogies are everywhere a prominent part of political discourse. If nothing else, the study of past politics should help to assay the ore left by the stream of history and to recognize the fool's gold so often on display in the rhetoric of public life.

See also: Aristotelian Social Thought; Citizen Participation; Democracy, History of; Democracy: Normative Theory; Democratic Transitions; Historicism; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Parliaments, History of; Participation: Political; Political Anthropology; Political Culture; Political Discourse; Political Economy, History of; Political Science: Overview; Republicanism: Philosophical Aspects

Bibliography

Acton J 1957 *Essays on Freedom and Power*. Meridian Books, New York

Breisach E 1983 *Historiography: Ancient, Medieval, and Modern*. University of Chicago Press, Chicago

Dirks N B, Eley G, Ortner S B (eds.) 1994 *Culture/Power/History*. Princeton University Press, Princeton, NJ

Gaddis J L 1997 *We Now Know: Rethinking Cold War History*. Clarendon Press, Oxford, UK

Gilbert F 1965 *Machiavelli and Guicciardini: Politics and History in Sixteenth Century Florence*. Princeton University Press, Princeton, NJ

Gooch G P 1959 *History and Historians in the Nineteenth Century*. Beacon, Boston

Heeren A 1834 *A Manual of the History of the Political System of Europe and its Colonies*. Clarendon Press, Oxford, UK

Iggers G G, Parker H T 1979 *International Handbook of Historical Studies*. Greenwood Press, Westport, CT

Kelley D R 1970 *Foundations of Modern Historical Scholarship: Language, Law, and History in the French Renaissance*. Columbia University Press, New York

Kelley D 1998 *Faces of History: Historical Inquiry from Herodotus to Herder*. Yale University Press, New Haven, CT

Krieger L 1977 *Ranke: The Meaning of History*. University of Chicago Press, Chicago

Laistner M L W 1963 *The Greater Roman Historians*. University of California Press, Berkeley, CA

Marrus M 1987 *The Holocaust in History*. University Press of New England, Hanover, NH

Momigliano A 1990 *The Classical Foundations of Modern Historiography*. University of California Press, Berkeley, CA

Ranke L von 1962 *Die Römischen Päpste in den letzten Vier Jahrhumderten*. Phaidon Verlag, Vienna, Austria

Stevenson D 1997 *The Outbreak of the First World War: 1914 in Perspective*. St. Martin's Press, London

Thucydides 1996 *The Landmark Thucydides*. Free Press, New York

J. J. Sheehan

Political Lawyering

Scholarship on lawyers and politics falls into two broad classes. On the one hand, as in most theories of professions—most notably functionalist and market control theories—any politics is largely absent. In a very different way, studies of lawyers' participation in the state and political institutions, where lawyers take for granted the rules of the political game, conceal the collective action of lawyers toward a common definition of the public good. There are, however, two theoretical perspectives that engage lawyers' politics: political lawyering and cause lawyering. The first approach demonstrates that lawyers have mobilized collectively, at critical points of political transition, to advance the institution-building of liberal political regimes. The second approach indicates the impact of those lawyers who pursue political agendas in direct response to perceived inequity or injustice, and in reaction to repressive or arbitrary legal and political regimes. This article briefly reviews the first class of scholarship and then analyzes in more detail political lawyering and cause lawyering.

1. The Avoidance of Politics

1.1 Functionalism and its Critics

Functionalism dominated the sociology of professions, including law, during the years 1950–70. Since functionalists sought to understand the link between knowledge and social needs—or, more precisely, to study the social control of knowledge—the logic of the profession solved the intractable problem of the balance between esoteric, dangerous knowledge in the hands of a potentially egoistic few and democratic control of those few which could in itself be destructive. Functionalists found a solution in the logic of a profession where (a) specific social structures (socialization, ethics, rewards, etc.) oriented the use and the development of knowledge in order to produce a relationship of service between professionals and society, and (b) society delegated to a 'community' a set of duties which were dominated by 'institutionalized altruism' and whose dedication to those duties was appropriately compensated by high social rewards (self-government, prestige, power, and some-

times wealth). Yet the constellation of questions and concepts that were used to study lawyers, as a specific profession, did not refer either to politics or to collective action. History was not considered as the product of social action, but rather as the change to which a profession was compelled to adapt.

Although Talcott Parsons (1964) was the central figure of functionalism, he recognized that, due to its specific 'interstitial' position in the social structure, the legal profession cannot escape the political system. Through the different choices they make, lawyers, as intermediaries, facilitate or impede the task of political authority and favor or not transformations in legal interpretation. Yet even this analysis remains rather limited, because the terms that orient lawyers' action—political authority, law, clients—are given, and because, with some room to maneuver, the legal profession is fundamentally defined by an integrative function.

In the 1970s, functionalism (among other theories) became the target of a critical sociology that leveled three charges against the sociology of professions:

(a) theory (in this case, the functionalist theory of profession) was nothing more than the reproduction of professional ideologies and a mask for gaining or keeping symbolic and material rewards;

(b) the so-called 'service relationship' was only a social mystification and, like the ideal of independence, was not justified by the quality of the service; and

(c) sociologists, therefore, were servants, unwittingly or not, of the professional elite.

Critical sociology rejected ahistorical studies, demanded realism about professional interests, and raised new questions about power, effectively anticipating the rise of market control theory.

Magali Sarfatti Larson opened up a new research program in the late 1970s. Devoted to the comparative and historical study of several professions in the United States, *The Rise of Professionalism* (1977) brought together two intellectual traditions: the European tradition, centered on Marx, Weber, and Polanyi; and the American tradition. It combined a historical framework with collective action (the mobility project), and it integrated the transformation of the market and the change of the symbolic systems through a very original analysis of the role of universities in creating legal service as a new commodity. An interpretation of Larson's book became the central resource of a new school—market control theory—whose simplified thesis posited that lawyers are rent seekers who, to maximize their profit, use two complementary strategies: to reduce the supply of legal services by excluding competitors, and to increase demand for legal services by influencing the state. All the established and distinguishing characteristics of the legal profession—self-government, professional mortality, independence, controls over entry to the profession, moderate internal competition, and so on—still exist, but only as tools to extend market control and thus to increase economic and social profit. This perspective, which received its most expansive application in Abel and Lewis's *Lawyers and Society* (1988a, 1988b, 1989), ignored politics, took for granted the capacity of professions to mobilize collectively, and assumed the ability of professions to impose their self-interested wills on legislatures and states. It directed scholarly attention away from lawyers acting collectively for reasons other than market control. Although erudite and fruitful in many other respects, Abbott's jurisdictional theory of professions has a similar result. While Abbott (1988) offers acute insights into ways that professions mobilize knowledge in their scramble for control over work, he does not classify either the political processes by which professions secure and maintain jurisdictional control, or the activities of professions beyond control of work domains.

1.2 Institutionalized Politics

Originating primarily from political science and studies of political recruitment, many studies since the 1960s have sought to understand the significance of a widely observed phenomenon—the overrepresentation of lawyers in the legislatures of many liberal democracies, e.g., the USA, India, Canada, Venezuela, Australia, and the Netherlands (see Miller 1995). What are the effects of background and networks on legislative process? Does the prominence of lawyers in legislatures make a difference to policy outcomes?

Studies of the compositional or recruitment politics of lawyers show limited results. Lawyers' actions in legislatures are rarely the expression of any collective professional orientations. Lawyers as legislators do not evince as a whole a coherent or distinctive substantive politics. They distribute themselves across political parties, so there is rarely a clear affinity with party ideology.

At least in the United States, the strong overrepresentation of lawyers in legislatures leads not to substantive outcomes but to a peculiarly lawyerly way of dealing with issues. According to Miller (1995, p. 174), who summarizes several decades of research as a context for his own findings on Congress and state legislatures, 'the numerical dominance of lawyer-politicians in US political institutions produces a government that is preoccupied with the procedurally-oriented, incrementalist myth of rights, instead of with broader substantive questions of public policy.' Thinking like a lawyer leads to reaction against radical change, it privileges procedural rationality over substance, and it produces a rights-focused approach to the solution of public problems. Policy-making alternatives are defined more narrowly and institutional parameters are largely taken for granted. Similar patterns may be observed elsewhere. Ledford's (1996) study of German lawyers in the last quarter of the nineteenth century finds that, through their partici-

pation in national political movements and parties, lawyers were actively committed to the building of the *Rechtsstaat*, a rule-of-law society that would implement the protection of civil rights. But he also found that this procedural orientation faltered when faced, a half-century later, with the substantive challenges posed by National Socialism.

2. Political Lawyering

Sociologists, historians, and legal scholars have demonstrated that, in Europe and North America, legal professions, acting collectively, have been periodic but by no means consistent agents of political liberalism at particular times and places from the seventeenth century onwards. These scholars generally define political liberalism along three dimensions:

(a) a moderate state, which is an amalgam of the rule of law, strong intermediate associations that exist between the state and the individual and will defend the latter against the former, a balance of national and local political authority, and an institutional separation of powers between the executive branch of government and the judiciary;

(b) a civil society, which combines the existence of voluntary associations, religious institutions, and publics, which do not owe their existence to the state and can stand against it; and

(c) citizenship, most notably the civil element of citizenship, in which individuals gain legal rights and personality.

These scholars identify historical episodes in several countries in which political lawyering has facilitated the construction of the moderate state, civil society, and civil citizenship. The findings of these studies may be summarized in five propositions.

2.1 The Autonomy of the Judiciary and the Autonomy of the Bar are the Principal Conditions for the Fight of Lawyers on Behalf of a Moderate State

Autonomy from the state has relevance for legal professions for two reasons. First, some independence of the judiciary from executive power is not only critical for establishment of a rule-governed state; and second, autonomy of the profession itself is the key element in this establishment, for without it no independent collective action will be possible. The fight for judicial autonomy can be found in eighteenth-century France where the *parlements* (sovereign courts with judicial and legislative functions) sought to enlarge their powers and control the person of the king and power of the royal court. Lawyers were the *parlements'* most faithful allies. They supported the independence of the *parlements* by urging effective general strikes which brought the justice system to a halt, and by becoming the political thinkers of the struggle. Lawyers articulated the foundations of the moderate state with such slogans as 'Laws are veritable contracts between those who govern and those who are governed.' Since the 1870s, the bar in the USA has strenuously exerted itself to immunize federal and state courts from control by presidents or governors, congress or state legislatures, and federal, state and municipal bureaucracies. In Germany, where the judiciary was subordinated to the state, the lawyers' campaign for a 'free bar' began in the 1830s with the aim of obtaining the institutional separation of judicial and administrative authority and the autonomy of law through adherence to judicial precedent. This movement reached its zenith in the Imperial Justice Laws of 1877–79, only to decline thereafter until the advent of National Socialism.

Seen from the vantage point of political liberalism, the autonomy of lawyers may be one of the key distinguishing features of a liberal political society. In England, the autonomy of the bar long preceded the emergence of an English state in the sixteenth and seventeenth centuries. It thus became a pillar of English constitutionalism, and was therefore able to resist effectively the attempted incursions of Charles II and James II on such 'little commonwealths' and 'little republics' that existed independently within the English body politic. In France, that autonomy grew consistently with the nurturance of the *parlements* from the seventeenth century to the present, with short interruptions during the Revolution and 1822–30. It was the absence of long-established, independently-secured autonomy in Germany, coupled with internal and external conflicts, which ultimately brought the bar under National Socialist control in 1933 and rendered it quite unable to defend judicial autonomy.

2.2 Lawyers' Struggles to Institute and Extend Citizenship are Diverse

Lawyers' orientations towards the extension of citizenship can be distinguished along two dimensions: first, between substantive and procedural rights; and, second, among civil, political, and social concepts of citizenship. French lawyers have been the champions of substantive rights, such as the eighteenth-century efforts to defend the religious rights of Jansenists and to obtain property rights for peasants in Burgundy, or the nineteenth-century efforts to transform the courtroom into a political forum to fight against emergency laws, attacks on a free press, and limits on individual freedoms. In contrast, German lawyers in nineteenth or twentieth centuries did not press directly for substantive individual rights, but rather pursued them indirectly through proceduralism—which opened them to seduction by antiliberal sentiments in the 1920s and 1930s. In the United States, fringes of the

profession variously championed substantive civil rights and rights to representation on behalf of the poor, but the central tendency of the profession has been to valorize procedural limits on state authority, as in the McCarthy period of anti-Communist purges.

2.3 Lawyers' Collective Defense of Civil Rights Does Not Extend to Struggles in Favor of Economic and Social Rights, or to Fights Against Racism, Xenophobia, or Gender Discrimination

With surprising continuity across history and nation, lawyers have been strongly defined by a 'pure' civil rights model. Lawyers are not to be found collectively on the barricades of expanded welfare rights, of national political issues such as the universal ballot, or of the restriction of racism. During the civil rights era in the United States (1950s and 1960s), a few professional associations took collective action. In nineteenth-century France, few lawyers mobilized to defend the working class. Lawyers are limited liberals, most especially in relation to citizenship.

2.4 The Forms of Lawyers' Political Action are Principally Reactive and Rely on the Authority of the Public or Civil Society

Compared to social movements, lawyers react against arbitrary or illegal actions by governments and repudiations of individual rights through trials that become '*causes célèbres*.' In Britain and France, use of the courts as forums came to constitute the political public sphere. Radical barristers in eighteenth- and nineteenth-century England used the courts to rail against corrupt government, to publicize causes, and to champion rights. French barristers used their freedom of speech in the *ancient régime* to create extended legal briefs that effectively became political tracts. Not only did they stimulate debate within the enlightened public, but they claimed for themselves a representational role as the voice of the public. By contrast, lawyers in the United States used the institutions of civil society, which were well developed, to mobilize publics. In Germany, lawyers were able neither to constitute publics as an autonomous sphere, nor to mobilize the institutions of civil society, which barely existed.

2.5 Contradictory Movements Entangle Economic Globalization with Political Globalization on Behalf of Liberal Institution-building

The market has contradictory effects on political globalization of liberal political institutions. On the one hand, the global market has homogenizing effects on nation states at the same time as it stimulates national reactions against it. The extension of the global market also introduces more contradictions within segments of legal professions and thus inhibits their capacity to act collectively. On the other hand, the conditions of financial stability and global trade require institution-building, such as the rule of law, independent courts, and private professions. Thus, differing segments of legal professions come to constitute liberal politics through institution-building that will facilitate global commerce, and through advocacy on behalf of human rights and environmental responsibility among other fields.

3. Cause Lawyering

Cause lawyering is a form of doing politics (Sarat and Scheingold 1998, 2001). Whereas legal elites separate law from politics, and legal skills from moral calling, cause lawyers seek to 'reconnect law and morality.' Unlike the classic image of the lawyer as 'hired gun,' prepared to serve a client whatever the merits of the client's case and however wide the gap between the values of lawyers and clients, cause lawyers usually identify with the values and causes of their clients. They may even join with clients in political and social movements to advance a common goal. Thus, cause lawyers have a substantive vision of the good society and they use their legal skills to produce it.

If political lawyering is defined by the defense of political rights construed narrowly, the range of causes can be enormously variegated, encompassing pay equity, animal rights, capital punishment, Bedouin land rights, struggles against military repression, women's rights, peasant movements, and human rights, among a multitude of others. (See the case studies in Sarat and Scheingold 1998, 2001.)

Scheingold (2001) shows that cause lawyering is both constant and contingent, and that cause lawyers must choose between frequently uncomfortable alternatives. There is a constancy and coherence that transcends time and place for, while 'cause lawyering is frequently directed at altering some aspect of the social, economic, and political status quo,' (Sarat and Scheingold 1998, p. 4) cause lawyers are everywhere caught between two visions of democracy—of legal and political rights versus social and economic rights. They lean towards the latter and its emphasis on 'egalitarian values and redistributive policies.' Since they face an enormous variety of political regimes (totalitarian, colonial, authoritarian, repressive, liberal democratic, neo-corporatist) with quite different sets of resources, and since their causes are highly diverse, the expression of cause lawyering and its efficacy is highly contingent.

3.1 Strategies of Cause Lawyers

Cause lawyers must usually choose between two broad strategies, although these frequently blur in practice (Scheingold 2001). To follow a *legal strategy* means operating within the very mainstream of professionalism, with which cause lawyers constantly struggle. Thus Israeli civil rights lawyers defend Palestinian 'terrorists' against Israeli security officials within Israeli courts. This may be somewhat effective on narrow grounds but leaves the larger status quo intact. To follow a *political strategy*, and to identify with the cause of clients, may take the form of a political advocacy through social and political movements that extends well beyond legal constraints, as evidenced by many cause lawyers in Latin America. This allows much greater scope of advocacy, but reduces the scope of professional collective action since it marginalizes the cause lawyers.

Similarly, cause lawyers must choose between distance from the state versus close proximity to the state. On the one hand, a close relationship with the state, through cooperation with state authorities, such as maintained by cause lawyers in post-Apartheid South Africa, may yield results while significantly constraining the scope of action, for lawyers are bound by the limits of state institutions. Yet many of the ideals advocated by cause lawyers cannot be achieved without the exercise of state authority, as in the extension and enforcement of new rights, or the complicity of state officials. On the other hand, an adversarial relationship to the state may mobilize public opinion and social movements, but it is a high-risk strategy. The closer to the state, the more the cause lawyer will be compelled to moderate the cause and compromise his or her objectives.

3.2 Resources of Cause Lawyers

Cause lawyers must often rely on resources quite different from those available to the elite or core of the profession. When cause lawyers seek to mobilize, they may be heavily resource-dependent in at least five areas.

(a) *Money*: since little funding for cause lawyers comes from third world governments, they rely heavily on western governments, development agencies and churches (see Ellman in Sarat and Scheingold 1998);

(b) *Media*: publicity, including widely distributed articles, books, print and media interviews and stories, even public demonstrations and courts, increases the impact of otherwise isolated cause lawyers (see Bisharat in Sarat and Scheingold 1998);

(c) *External validation*: in Third World countries, funding by international foundations, partnerships with international non-government organizations (INGOs), and recognition by distinguished professional bodies, cannot be underestimated;

(d) *Professional solidarity*: where cause lawyers can ally with professional legal elites or law schools, they can obtain sponsorship, protection, and legitimation at the local level; and

(e) *Churches*: in some countries, such as Brazil during the military regime of the 1970s and 1980s (see Meili in Sarat and Scheingold 1998), the church and its human rights centers offer institutional solidarity to cause lawyers (Halliday 1999).

The cause lawyering perspective also perceives an ambivalent relationship with globalization.

> 'For cause lawyers … globalization and state transformation may be a mixed blessing. In some contexts it may help to provide a political structure wherein minimal rights claims and legal defenses against abuses of state power become meaningful. In other contexts, it may erode support for the kinds of social changes which these lawyers, and the movements with which they are affiliated, seek to advance' (Scheingold 2001, p. 384).

4. Conclusion

Although they have quite different theoretical origins, the two principal approaches to lawyers' politics show some convergence. Political lawyers have been studied more in the past and cause lawyers more in the present, but in fact both can be found in every period. In both cases, three questions are central: how to explain the relationships between political orientations and lawyers' involvement; how to understand the collective construction of a defense or a cause; and how to produce collective action that will change institutional rules. The main difference between political and cause lawyers lies in the conditions which facilitate or hinder the sharing of a common political vision. Cause lawyers are usually individuals or groups defending a specific case in the context of an indifferent or hostile profession. If cause lawyering is very much a creature of, and a contributor to, legal–liberal forms of democracy, cause lawyers do so as a deviant strain within the legal profession and by making appeals for support of their cause outside the profession. By contrast, political lawyering is the expression of a common orientation of lawyers and therefore, through individual causes, it deals directly with other collective actors, especially justice and the state. The more the profession is diversified, the more its political commitment is likely to take the form of the defense of a cause; and the more it is united around some specific issues, the more it is likely to become a political actor. Thus, while the value orientations of these kinds of lawyers' politics may differ, they exist in creative tension to each other, and together may come to constitute the building and maintenance of liberal politics in the modern world.

See also: Affirmative Action, Sociology of; Citizenship and Public Policy; Citizenship, Historical Develop-

ment of; Citizenship: Political; Citizenship: Sociological Aspects; Equality and Inequality: Legal Aspects; Inequality; Justice and its Many Faces: Cultural Concerns; Law and Democracy; Lawyers; Legal Professionalism; Political Science: Overview; Public Interest Law; War Crimes Tribunals

Bibliography

Abbott A 1988 *The System of Professions*. University of Chicago Press, Chicago

Abel R L, Lewis P S C (eds.) 1988a, 1988b, 1989 *Lawyers in Society: Vols. I, II, III*. University of California Press, Berkeley, CA

Halliday T C (ed.) 1999 Review symposium: Lawyers and politics. *Law and Social Inquiry* **24**: 1007–124 (esp. essays by T. C. Halliday, S. Scheingold, M. Burrage)

Halliday T C, Karpik L (eds.) 1998 *Lawyers and the Rise of Western Political Liberalism*. Oxford University Press, Oxford, UK

Karpik L 1999 *French Lawyers: A Study in Collective Action, 1274–1994*. Oxford University Press, Oxford, UK

Larson M S 1977 *The Rise of Professionalism*. University of California Press, Berkeley, CA

Ledford K 1996 *From General Estate to Special Interest: German Lawyers, 1878–1933*. Cambridge University Press, Cambridge, UK

Miller M C 1995 *The High Priests of American Politics: The Role of Lawyers in American Political Institutions*. University of Tennessee Press, Knoxville, TN

Parsons T 1964 *The Social System*. Free Press, New York

Sarat A, Scheingold S (eds.) 1998 *Cause Lawyering: Political Commitments and Professional Responsibilities*. Oxford University Press, New York

Sarat A, Scheingold S 2001 *Cause Lawyering and the State in a Global Era*. Oxford University Press, New York

Scheingold S 2001 Cause lawyering and democracy in transnational perspective. In: Sarat A, Scheingold S (eds.) *Cause Lawyering and the State in a Global Era*. Oxford University Press, New York

T. C. Halliday and L. Karpik

Political Machines

Though the term 'political machine' is sometimes used to describe any party or electoral organization noted for its efficiency and effectiveness, its principal significance, especially in American political science, is to refer to a particular kind of party organization, that draws much of its support from patronage jobs, government contracts, and other 'spoils' of office. This second type of machine once dominated the political landscape in most large American cities, though by the late twentieth century most such machines had declined or disappeared.

1. Two Definitions and Their Application

The term 'political machine' is used in two related, but distinct ways. Particularly in the 1950s and 1960s, it was often used to refer to any party or electoral organization distinguished by its effectiveness and efficiency (see, e.g., Gottfried 1968). Usually, such an organization was headed by a leader or small leadership group and imposed some measure of discipline on its members (otherwise it would not have functioned quite so effectively), but its most important quality was its ability to acquire, maintain, and use political power on a regular, predictable basis. Hence, the comparison between a human organization and a mechanical contrivance. When used in this sense, the term has a generally neutral connotation, and the study of political machines is simply one aspect of the larger subject of political party organization.

More often, however, the term political machine has been used to refer to a particular kind of party organization: one that derives its effectiveness from the extensive use of what organizational theorists call specific, material inducements. As laid out most clearly by Banfield and Wilson (1963), a specific (as opposed to general) inducement is one 'that can be offered to one person while being withheld from others.' Material inducements include money and any other 'physical "thing" to which value attaches.' Thus, a political machine attracts support to a much greater extent than other party organizations through the use of patronage jobs, government contracts, and special favors to individual voters. Because these sorts of inducements conflict with certain widely held civic norms, and because their use frequently leads to other forms of corruption, this second use of the word machine has a decidedly pejorative connotation. Indeed, those actively involved in political combat sometimes use the term as little more than an epithet. As Theodore Roosevelt once noted, one's own organization is invariably described as a 'club' or a 'party'; it is only the opposition that has a machine.

When used in this second sense, political machines have a particularly celebrated history in American politics—and American political science. Almost every large and medium-sized city in the United States was run by one or more political machines, often for extended periods of time, during the latter half of the nineteenth century and the first half of the twentieth century. Among the cities in which machines played a particularly powerful and significant role are Chicago, New York, Philadelphia, Jersey City, Kansas City, and Memphis. Moreover, since some of the most important work in the development of political science as an academic discipline dealt with city government and politics, machines also served as the principal subject matter for what are now regarded as some of the classic early works in American political science (see especially Merriam 1929 and Gosnell 1937).

It is difficult to say with much certainty how often and in what sorts of circumstances political machines have flourished, primarily because it is difficult to get reliable information about how much various organizations used patronage or graft and whether such practices were essential to their success or just an ancillary way of enriching the leadership. There is, as has just been noted, widespread agreement about the importance of machines in American urban history. Many commentators also believe that many rural counties in the South and Midwest were once run by machines, though there seems to be little beyond anecdotal evidence to support this contention. American historians and political scientists once regularly asserted that political machines were a uniquely American institution. Over the last thirty years of the twentieth century, however, it has been recognized that machine-like organizations also exist in a number of Third World or developing countries, as well as in specific locales or regions of other Western nations, such as southern Italy (see especially Scott 1969).

2. *Explaining the Rise of Political Machines*

Why did political machines play such a dominant role in American urban politics? Three principal theories have been put forward to explain the origin and survival of machines in American cities. For shorthand purposes, these theories can be described as governmental, economic, and cultural.

According to the governmental theory, machines came into existence as a solution to the unique governing problems that beset American cities. On the one hand, cities faced an extraordinary array of difficult challenges in the late nineteenth and early twentieth centuries. Their population and geographic area grew at an explosive pace. Partly as a result, and partly as a response to new technological discoveries, cities were suddenly required to provide an array of important services—water, gas, electricity, paved streets, sewer systems, mass transit—that no American city had previously had to worry about. Unfortunately, the structure of American urban government was dramatically unsuited to these tasks. Governmental power and authority in the cities was widely dispersed among the mayor, city council, and a large number of special boards, commissions, and independently elected officials. The special function of political machines, then, was that they provided an informal way of centralizing the political power that had been fragmented among the formal organs of city government (for the classic statement of this argument, see Merton 1957).

The economic explanation for machines, by contrast, argues that machines came to power because they helped serve the needs of the large number of poor and economically insecure people who lived in American cities. When a city's population doubled in the course of just one decade, as often occurred in the late nineteenth century, the new residents necessarily consisted primarily of immigrants: sometimes from farms and rural areas within the United States, more often, especially in northeastern and midwestern cities, from foreign countries. In either case, the newcomers generally faced severe problems adjusting to their new environment and satisfying such basic human needs as food, housing, and employment. Yet, American government at this time provided little or no help at all in the way of welfare or social services. Political machines did not completely fill the gap, of course, but they were one of the few agencies that even tried, providing at least some of the urban poor with jobs, occasional gifts of food and fuel, and help when they came in conflict with the law. One early twentieth-century social worker went so far as to call the machine 'a kind of crude socialism.' In return, the lower classes, especially those of foreign background, provided a reliable fund of votes for machine candidates, and were not particularly upset if the resources the machine distributed were ultimately derived from illegal or corrupt activities.

The cultural explanation also sees foreign immigration as a key factor in the rise of urban machines, but places emphasis not on the immigrants' economic predicament, but on their political culture. Of the immigrants who came to America after 1840, few had any experience with democracy; almost none had ever cast a meaningful vote. On the contrary, most had learned to view government only as a hostile and authoritarian presence. Against that background, most immigrants naturally saw their votes less as a way of working large-scale transformations in the urban social order than as a commodity that could be exchanged for short-term personal advantages. One widely quoted formulation of this argument has been provided by historian Hofstadter (1955), who saw the clash between immigrants and natives as having been grounded in 'two thoroughly different systems of political ethics.' Where the middle-class natives 'argued that political life ought to be run … in accordance with general principles and abstract laws,' immigrants 'interpreted political and civic relations chiefly in terms of personal obligations, and placed strong personal loyalties above allegiance to abstract codes of law or morals.'

3. *The Decline of the Urban Machine*

If political machines were once an all-but-ubiquitous feature of American urban politics, they are no longer. It is difficult to be precise about exactly when the machines lost their capacity to deliver votes on a regular and reliable basis for their favored candidates. In Chicago, for example, Mayor Richard J. Daley continued to win easy re-election all the way up until

his death in 1976, but it is questionable if the machine *per se* had much to do with his last several victories. A good case can be made that he won less because of the organization's precinct captains and patronage army than because most city voters genuinely thought he did a good job of governing the city. There is, moreover, no easy way to say who is and who is not a 'machine candidate.' Many self-proclaimed 'reformers,' once in office, would begin rewarding their supporters with jobs and contracts in ways that were largely indistinguishable from the machines they defeated.

That said, it was primarily in the years between about 1945 and 1965 that the most celebrated machines in American history suffered a series of major defeats and lost their capacity to dominate politics on a citywide basis. By the late 1960s, it was frequently asserted that of all the large cities in the United States, only Chicago still had a functioning political machine.

In general, when a machine started to weaken, the first major office it would lose control of was the mayoralty. Yet, long after many machines could no longer dominate these sorts of highly visible, media-centered campaigns, they could still reliably elect a number of their supporters to posts on the city council and to many of the less publicized offices in city and county government. In these weakened circumstances, the vestiges of machine politics linger on in a number of cities.

Why did the machines die out? Each of the three theories developed earlier to explain the growth of machines also provides a ready explanation for their decline. According to the governmental theory, machines died out because, over time, many of the worst problems afflicting urban government became more tractable. Cities stopped expanding, they gradually learned from experience how to deal with such basic demands as paving streets and providing water and electricity, and, in many cases, city charters were changed in ways that brought a greater measure of order and coordination to the formal agencies of government.

Advocates of the economic theory, by contrast, argue that machines went into eclipse because the demand for their services declined. On the one hand, a series of laws passed in the 1920s effectively shut off the flow of foreign immigration; those immigrants who were already here—and their children—gradually learned to adjust to their new environment and, especially after World War II, escaped poverty and even made it into the middle class. Equally important, with the dramatic expansion of the American welfare state in the 1930s, government now provided many of the goods and services for which the machines had once held a near-monopoly.

The cultural theory, like the economic theory, sees the passage of restrictive immigration legislation as an important, long-term threat to the machines' existence. It would also point toward agencies, such as the public schools and the mass media, that were attempting to assimilate the immigrants' children and grandchildren and socialize them into a new set of political norms.

According to several of these theories, an important demographic trend of the 1950s and 1960s—the massive migration of American blacks from rural southern areas to northern cities—at least had the potential to revivify the machines. Like the European immigrants of the late nineteenth and early twentieth centuries, the black migrants faced enormous social and economic problems and had little previous experience with voting, and there is little doubt that their arrival in northern cities posed huge new problems for urban governance. Yet, for a variety of reasons, blacks proved to be an uncertain and unreliable constituency for the established machines. Racial antagonism and prejudice made many white ethnic groups unwilling to seek common cause with blacks—and made blacks unwilling to take their place as just one more group in the machine coalition. Equally important, the civil rights movement encouraged blacks to think of their votes as a means of effecting social change, and not just as a way of gaining particularistic benefits.

4. *Legacy and Enduring Significance*

How well did political machines govern American cities? To answer to this question, one must pose a further one: compared to what? It is, for example, easy to criticize the aid machines gave to the urban poor: such aid was clearly inadequate to the scope of the problem and was often given out in an arbitrary fashion. And it came at a price: most of the money that the bosses skimmed from the public treasury wound up in their own pockets, not as assistance to the less fortunate. Still, when compared to the contemporary alternatives—the almost complete absence of governmental aid and the paternalistic and demeaning nature of private charity—it is hard not to feel that the machines deserve a more favorable assessment.

At the very least, one can say that there is little compelling evidence to suggest that the machines governed worse than the 'reformers' they competed against—and some reason to think the opposite. All too often, reformers proved to be personally honest but largely ineffective, and to have strikingly little appreciation of the real problems and concerns of immigrants and the urban lower classes (see in particular (Holli 1969)

Good or bad, machines are simply not a very significant presence in contemporary American politics, national, state, or local. Nor is there any obvious reason to think they will be revived at any time in the foreseeable future. Yet, for several reasons, political machines are still worth studying. Most obvious is their historic importance: for at least a century, machines were one of the principal centers of political power and political conflict in almost every large city

in the United States. No one can hope to understand American urban history without coming to terms with the machine. Second, the struggle between machines and reformers left an enduring imprint on the structures of contemporary urban government. Many of the most common institutions in city government, including nonpartisan elections and the city-manager form of government, were originally developed as ways of undercutting the power of the machine. Long after the machines have declined or disappeared, these institutions remain.

See also: Campaigning: Political; Corruption: Political and Public Aspects; Participation: Political; Political Money and Party Finance; Urban Politics: United States; Urban Sociology

Bibliography

Banfield E C, Wilson J Q 1963 *City Politics*. Harvard University Press, Cambridge, MA

Gosnell H F 1937 *Machine Politics: Chicago Model*. University of Chicago Press, Chicago

Gottfried A 1968 Political machines. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan and The Free Press, New York, pp. 248–52

Hofstadter R 1955 *The Age of Reform*. Knopf, New York

Holli M 1969 *Reform in Detroit: Hazen S. Pingree and Urban Politics*. Oxford University Press, New York

Merriam C E 1929 *Chicago: A More Intimate View of Urban Politics*. Macmillan, New York

Merton R K 1957 *Social Theory and Social Structure*. Free Press, New York

Scott J C 1969 Corruption, machine politics, and political change. *American Political Science Review* **63:** 1142–58

W. G. Mayer

Political Money and Party Finance

Healthy democratic politics depends importantly on a free flow of information among citizens, groups, candidates for public office, and political parties. The right to speak diminishes in significance without the financial resources needed to be heard. In the legendary words of California Assembly Speaker Jesse Unruh, 'Money is the mother's milk of politics.' Yet the manner in which political money is raised and spent can undermine the legitimacy of democratic politics. Ensuring that parties and candidates have adequate resources to compete effectively in elections is problematic given the inevitable tension between the reality of economic inequality and the ideal of political equality. Efforts to prevent concentrations of wealth from undermining political equality may conflict with bedrock freedoms of speech and association. Every democracy struggles to reconcile the need for political money with the problems it begets. None have found entirely satisfactory and long-lasting solutions. The political finance tools available to policy makers—public subsidies, limits on contributions, expenditure controls, disclosure, and regulation of campaign activity—often leave reformers with unfulfilled objectives and unanticipated consequences.

1. Institutional Context

The prominence and nature of money in politics varies across institutional settings. Party-centered electoral systems differ substantially from candidate-centered systems. In the former, most often associated with the parliamentary systems of Western Europe, political money means party finance. Most, if not all, political funds are raised and spent by political parties and the primary concerns have to do with the adequacy of funds available for routine party activities, the work of party caucuses in parliament, and election campaigns. Direct and indirect public subsidies are often provided to the parties, including cash and free broadcasting time on state-owned networks. There is substantial variation in the extent of regulation of party finance, with disclosure of large private contributions and prohibitions on paid political advertising most typical and limits on contributions the least typical. Party-centered systems have not been immune from escalating campaign costs and scandals, as the Clean Hands investigation in Italy and the more recent charges against Helmut Kohl's Christian Democrats in Germany illustrate. But the problems associated with political money in party-centered systems are relatively contained and finance practices are generally viewed as in need of monitoring and repair, not radical surgery.

In candidate-centered systems, by contrast, the focus is less on political party activity and more on campaign fundraising and spending by individual candidates. In these systems, the term of choice is campaign finance rather than party finance. The United States is, of course, the most prominent example of this class of democracies and the one most atypical in its various components. While parties once dominated the recruitment of candidates and the waging of campaigns in the US, a host of developments, including the creation of a professional civil service, the spread of the direct primary, the rise of television, the reform of party rules, and the provision of substantial staff resources to public officials, have diminished their role. Party organizations in the US have become largely appendages of incumbent office-holders—more instruments for raising money within and beyond the constraints of the law than independent political actors. Political money in the US and other candidate-centered systems is widely viewed as highly problematic and much time and energy are devoted to reporting on it and debating how best to ameliorate its harmful effects.

The demand for and supply of political money are shaped by more than the extent of party or candidate-centered elections. In the US, the sheer number of contests across the federal system, the separation of executive and legislative elections, plurality rules, constitutional protections of free speech, the high cost of reaching the large American audience, the professionalism and length of campaigns, and the permeability of the policy process all contribute to the large sums raised and spent in election campaigns. The perceived stakes of elections, the incentives and opportunities for raising and spending private resources in campaigns, and the laws and norms governing its role all condition how political money flows in democracies.

2. Problems Associated with Political Money

Rising campaign costs, scandals, and equity concerns in the early and mid-twentieth century precipitated waves of regulation of political finance and subsidization of party and campaign costs in democracies around the world (Paltiel 1981). Similar concerns kept issues of the regulation and subsidy of political finance on the public agenda in more recent years. Problems arise as parties and candidates struggle to raise adequate resources to compete effectively in times when those costs have risen sharply and when groups and individuals seek to influence election outcomes and policy decisions. At one extreme lies bribery and extortion, the illegal linkage of private resources and public authority for personal gain. Short of such explicit *quid pro quo* exchanges lie such problems as the dependence of parties and candidates on large financial contributors, conflicts of interests arising when politicians receive contributions from those who seek favorable action by a legislative or administrative body, and the impact on democratic legitimacy of the appearance of special access and influence by financial contributors. Then there are the consequences of the money chase—the unending scramble for political money—for how party and public officials allocate their time, who they see and listen to, and how both affect the deliberative processes of government.

Another set of problems are associated directly with the electoral process. Rising campaign costs and inadequate political funding depress electoral competition, discourage able people from seeking public office, give a distinct advantage to incumbent office-holders and wealthy self-financed candidates, and weaken democratic accountability by muffling the voices of individuals, groups, candidates, and parties.

A final set of problems arise from the very act of trying to cope with the others. Reforms designed to alter the flow of money in elections often produce equally problematic unanticipated consequences. Political money is fungible and legal constraints can easily divert it to less accountable passageways. Ambitious politicians naturally exploit the weaknesses and push the boundaries of the law, threatening the effectiveness and credibility if not the complete collapse of the regulatory regime.

3. Political Finance Tools

Policy-makers have utilized a conventional toolbox in responding to the problems associated with political money, but the mix of tools and particular applications have varied widely among democracies.

3.1 Public Subsidies

In response to concerns about rising campaign costs and excessive dependence on financial contributors, public support of parties and campaigns has become nearly universal among democracies (Katz 1996). Subsidies take the form of direct cash assistance, the provision of goods and services (e.g., free broadcast time, mailings, voter brochures, and meeting space), and indirect support such as tax credits or deductions for contributions or reduced rates for political advertising. The subsidies are given to parties or candidates and directed toward party maintenance or campaign expenditures. Direct cash subsidies to parties are typically allocated in proportion to electoral strength while services and indirect subsidies are usually provided on a more equitable basis. In the United States, public matching funds are provided to candidates for their party's presidential nomination who raise the requisite number of private contributions and agree to limit their spending; full public grants are provided to the major party candidates for the general election as long as they agree to raise and spend no private funds. But unlike most other democracies, the US provides no free broadcasting time for parties or candidates and no public subsidies for candidates for Congress.

Public subsidies are not without risks. They can reinforce the dominance of established parties, increase the potential for control by the state, and weaken the connection of parties with their grassroots. Public financing conditioned on very low qualifying thresholds can also encourage fringe and patently offensive candidates and parties. Full public financing systems generate intense pressure for private resources to find alternate routes to influence elections and policy.

3.2 Contribution Limits

While public subsidies are the norm in political financing, limits on contributions are less common. Many democracies place no restrictions on the source or amount of contributions to parties or candidates. Among those that do, prohibitions on particular sources (e.g., civil servants, foreign donors, corpor-

ations, and unions) are most common. In the United States, corporations and unions have been precluded for many decades from making political contributions from their treasuries but they may form voluntary political action committees (PACs), through which they may contribute to parties and candidates. In the late twentieth century, even prohibitions on contributions from corporate and union treasuries have been rendered ineffectual through the emergence of 'soft money' donations to political parties not subject to federal regulation.

Some countries have placed statutory limits on the size of contributions from permissible sources, including France, Israel, Japan, and the United States. The contribution limits in the US are notably low, especially since they were not indexed for inflation when fixed by law in 1974. This decline in the real value of donation ceilings contributes to the fundraising demands on politicians and pressures for private donations to flow through other channels, problems associated more generally with contribution limits.

3.3 Expenditure Limits

A number of democracies place limits on spending by parties and candidates, although these limits often apply only to a brief period of time (e.g., the formal campaign in parliamentary systems), to candidates but not parties (or vice-versa), or they exempt entire categories of political activity that influence election outcomes. In Japan, candidates for the Diet face limits on campaign expenditures but these are largely irrelevant because they apply only to the 12-day formal campaign period. For well over a century, Britain limited the expenses individual candidates for Parliament could incur in their constituencies (and prohibited outside groups from spending on their behalf) but did not regulate political party expenditures at the national level, where most of the action is. However, a new law adopted in 2000 limits national campaign expenditures by political parties; it also limits the amount of spending by outside organizations and individuals for political activity during a general election.

Mandatory spending limits adopted in the US in 1974 were declared unconstitutional, while voluntary caps tied to the receipt of public funds in presidential elections remain intact. But presidential candidates, political parties, and interest groups have managed to work around these expenditure limits in the general election by fashioning political advertisements (known as 'issue advocacy' but indistinguishable from traditional electioneering) that the courts have ruled are not subject to regulation.

3.4 Public Disclosure

Transparency in the flow of political money is embraced by many democracies as a deterrent to corruption or influence peddling but the comprehensiveness and effectiveness of disclosure regimes vary widely. At one extreme are systems requiring only the filing of a single report on party expenditures well after the election; at the other are systems like the US with elaborate rules for reporting political contributions (as small as $200) and expenditures and for making this information publicly available on a timely basis. Yet even in the US, a significant amount of electioneering falls outside the boundaries of disclosure law. While disclosure of political money is gaining in acceptance in the democratic world—Britain, for example, broke with its longstanding practice in 2000 by requiring that large donations to political parties be publicly disclosed—some Scandinavian countries continue to consider reporting requirements as an unacceptable violation of privacy.

3.5 Regulation of Campaign Activity

One way of reducing the demand for political money is to prohibit certain costly campaign activities. A surprisingly large number of countries ban paid political advertising on television, although most couple the ban with the provision of blocks of free television time to the parties. Such prohibitions are unthinkable in the United States, where free speech guarantees preclude any significant restraint on political communication. Other countries have moved to assert that a legal protection of freedom of expression limits the extent to which campaign activity can be regulated. For example, the Australian High Court declared a 1991 law that prohibited the broadcasting of political advertising during election periods unconstitutional, on grounds that it infringed the implied right to freedom of political communication.

4. Enforcement and Effectiveness

The success of any measures designed to redirect the flow of political money depends upon the ingenuity, comprehensiveness, and timeliness of the particular mix of regulatory and subsidy measures and on the mechanisms for enforcement. Short of blatant violations of criminal statutes (e.g., bribery or extortion), departures by parties and candidates from the letter and spirit of campaign finance laws often go unpunished. Politicians are reluctant to invest the requisite authority in public agencies to enforce political finance laws, a phenomenon tellingly illustrated by the structure and operation of the US Federal Election Commission. The FEC, whose six members include three from each of the two major parties, operates under very tight statutory and budgetary reins from Congress and often deadlocks on regulatory decisions. Putting in place a reasonable and workable set of regulations, appropriate criminal and civil sanctions, and administrative and judicial mechanisms has proven a daunting task in many democracies.

For all of these reasons it is not surprising that political finance reforms often fail to reduce corruption, lessen the influence of monied interests, slow the increase in campaign expenditures, or increase electoral competition. Other institutions and policies are more effective levers for combating corruption and enhancing democratic accountability. These include broad economic policies that limit the discretionary authority of public officials; institutions and practices nurturing the rule of law; an independent judiciary; adequate pay and strong ethical requirements for politicians and civil servants; transparent budget and financial control systems; and party systems supporting stable patterns of competition, accountability, and alternation of power (Kaufmann 1997, Pope 1996).

Not all political finance measures are doomed to fail, nor are they miracle cures for the ills arising from money in politics. But since neither the abolition of all private funds from politics nor the complete deregulation of political finance are feasible or sustainable, democracies will continue to struggle not to solve these problems once and for all but to contain and manage them as best as they can.

See also: Campaigning: Political; Electoral Systems; Electronic Democracy; Lobbying; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Media Effects; Participation: Political; Parties: Litigants and Claimants; Party Systems; Policy Process: Business Participation; Political Machines; Voting, Sociology of

Bibliography

Alexander H E, Shiratori R (eds.) 1994 *Comparative Political Finance Among the Democracies*. Westview Press, Boulder, CO

Corrado A, Mann T E, Ortiz D R, Potter T 2001 *The New Campaign Finance Sourcebook*. Brookings Institution Press, Washington, DC

Gunlicks A B (ed.) 1993 *Campaign and Party Finance in North America and Western Europe*. Westview Press, Boulder, CO

Katz R S 1996 Party organizations and finance. In: Leduc L, Niemi R G, Norris P (eds.) *Comparing Democracies: Elections and Voting in Global Perspective*. Sage Publications, Thousand Oaks, CA, pp. 107–33

Kaufmann D 1997 Corruption: the facts. *Foreign Policy* **107**: 114–31

Paltiel K Z 1981 Campaign finance: Contrasting practices and reforms. In: Butler D, Penniman H R, Ranney A (eds.) *Democracy At the Polls*. American Enterprise Institute, Washington, DC, pp. 138–72

Pope J (ed.) 1996 *The Transparency International Source Book*. Transparency International, Berlin, Germany

Report for Congress 1995 *Campaign Financing of National Elections in Selected Foreign Countries*. Law Library, Library of Congress, Washington, DC

Sorauf F J 1992 *Inside Campaign Finance: Myths and Realities*. Yale University Press, New Haven, CT

T. E. Mann

Political Parties

1. Defining Characteristics

Political parties are a form of political grouping. The term is commonly applied in different political systems to widely different bodies, ranging from the loose, decentralized, and undisciplined associations of the American Democratic and Republican parties to the tightly bonded, highly centralized, and rigidly disciplined Communist and Nazi parties of recent memory. It therefore seems appropriate to state the characteristic that appear to qualify these variegated groupings as 'parties':

(a) They are groups of people to whom labels—'Republican,' 'Labour,' '*Rassemblement pour la République*,' 'Baath Socialist Arab Rebirth,' and so on—are generally applied by themselves and others.

(b) Some of the people are organized—they consciously act together to achieve common goals.

(c) The larger society recognizes as legitimate the right of one or more parties to organize and promote their causes.

(d) In some of their goal-promoting activities, parties work through the mechanisms of representative government, such as nominations and elections.

(e) A key activity of parties is thus selecting candidates for elective public office.

Most scholars agree that, taken together, these characteristics distinguish modern parties from such historic precursors as the patricians and plebeians of the Roman Republic, the Guelfs and Ghibellines of medieval Italy, the 'hats' and 'caps' of eighteenth century Sweden and the Whigs and Tories of eighteenth century England. In his classic *The Structure of Politics at the Accession of George III*, for example, Sir Lewis Namier (1929) pointed out that 'Whig' and 'Tory' were loose labels sometimes called 'parties,' but, since it was uncertain which label fitted many MPs, the groupings were more 'tendencies of opinion and personal loyalty' than fully-fledged political parties.

The last of these defining characteristics most clearly distinguishes parties from interest groups. Typical organized interest groups, such as the American Rifle Association and Handgun Control, Inc., support, finance, and sometimes organize campaigns for political candidates who take their respective positions on gun control, but they do not *nominate* candidates who will bear their labels on general election ballots. Only parties do that.

1.1 'Party'

There are three levels of meaning. Many scholars of American parties have embraced a suggestion made in the first edition of V. O. Key's *Politics, Parties and Pressure Groups* (1958 [1942]). Key observed that when most Americans talk about a particular party or

parties, they have in mind one or more of three levels of meaning of the term.

(a) *Party-in-the-electorate* refers to the degrees of attachment to particular parties held by ordinary people (a concept later developed into 'party indentification,' operationalized and analyzed by the authors of *The American Voter* (Campbell et al. 1960)). People who feel some attachment to, say the Democratic party do not necessarily work for the party, contribute money to it, or even vote for its candidates.

(b) *Party organization* denotes people who hold offices in the parties or contest elections as their nominees or otherwise work for them in election campaigns.

(c) *Party-in-government* means persons who are elected or appointed to public offices commonly associated with a particular party—for example, the president, members of the cabinet, subordinate administrators serving at the pleasure of the president, members of the party's congressional party caucuses and their leaders, and the like.

There are some occasions, Key said, when some people use the term 'Democratic party' or 'Republican party' to include all three levels of meaning, although they are rarely explicit—or perhaps aware—of what they are doing.

When considering what people say about the Democratic and Republican parties, Key advised, it is well to be as precise as we can about just what meaning of 'party' we have in mind. Most political scientists (though not most journalists) concerned with the major American parties have followed his advice.

2. Party Types

Each party in any democratic system is not the same in all respects as every other party. They differ mainly along the lines covered in Sects. 2.1–2.3.

2.1 Membership

Membership in a party carries with it the right to participate in the selection of the party's candidates for public office, and it may or may not also involve the assumption of some kind of obligation to the parties. Most of the parties in most democracies have memberships that are to some degree closed: the parties' rules in such places as Great Britain, Europe, and Scandinavia usually require aspirants to apply for membership, and party officials can accept or reject the applications (they usually accept them). To continue as a party members in good standing, the person must at least pay annual party dues and perhaps also take an oath of loyalty to the party's principles and candidates. Membership in US parties, however, is open: there is no formal application for membership, Democratic and Republican presidential and other candidates are selected in direct primaries, and the right to participate in those primaries is guaranteed by law to all registered voters. In 'closed' primaries the voters are required to register their party preferences before the election, but they and they alone control their choice of parties. In 'open' primaries they do not make any public statement of their party preferences. In 'blanket' primaries they are allowed to vote in the primaries of either or both parties, although only for one candidate for each office.

2.2 Discipline

Most human organizations have some form of discipline, in which the leaders control the members by dispensing rewards and imposing punishments for the members' behavior. The leaders of every democratic party have some disciplinary powers, but some are more effective than others.

The leaders of the British Conservative and Labour parties, for example, have strong disciplinary powers. The members of their parliamentary parties know that when 'the whip is laid on'—when the leader, through the party whips, has told the members how they should vote on a particular bill—the MPs put their political careers in great jeopardy if they disobey. The prime minister, who is leader of the party in power, controls who gets ministerial offices constituting the 'government,' and the leader of the opposition controls who gets the positions in the 'shadow cabinet.' Moreover, the leader of each party controls its national organization, which has the power to deny a rebel MP reselection as the party's official candidate in each constituency and thereby effectively end the MP's political career. Perhaps the prime minister's greatest disciplinary power stems from the fact that if his or her government is defeated on a major parliamentary vote, a new election must be held, which the party—and any of its MPs—might lose. Thus an MP's vote against the whip's instructions is, in effect, a vote to risk putting the other party in power—a powerful deterrent against defections by the majority party's MPs.

US presidents are leaders of their parties (the party that does not hold the presidency has no generally acknowledged single leader), but their disciplinary powers are much weaker than those of their British counterparts. Presidents can make campaigning for one of their party's candidates contingent upon that candidate's support of presidential programs. However, most congressional candidates' electoral success depends far more on local factors than on presidential support. Indeed, it is usually an electoral asset for a congressional candidate to be seen as independent from any 'outside' party leader, including the president. Presidential discipline in American parties thus consists of little more than asking and cajoling.

Most European and Scandinavian parties give their national organizations the power to expel from the party members of the national parliamentary parties

for bolting the party line. While such expulsion is somewhat less likely than in Britain to end a parliamentarian's political career, it is still a much greater weapon than any available to the leaders of the American Democrats and Republicans.

2.3 Cohesion

Cohesion, defined as the extent to which party members holding public office act together on policy matters, is a critical party characteristic for people concerned with party responsibility. Parties in democratic countries vary considerably in cohesion. The British Conservative and Labour parties are among the most cohesive: on almost every issue before the House of Commons the members of the parliamentary majority party vote together without exception—and usually with unanimous opposition from the other parliamentary parties.

At the other extreme, some small parties in France split in parliament on almost every issue. Other democratic parties fall between these two extremes. The larger parties in Australia, Canada, and New Zealand, as well as the leading French, German, and Scandinavian parties, are almost as cohesive as their British counterparts. The American parties in Congress fall somewhere in the middle, often mustering near-perfect cohesion on such matters as the organization of the chambers, election of presiding officers, large tax cuts, increasing social welfare expenditures, and other major issues where party positions are well entrenched. Usually one or both parties split internally on other issues such as foreign aid, abortion, and free trade.

3. Party Functions

Every democratic country has political parties at the national level, and most have them at the regional and local levels as well. Moreover, all of the nations that began their existence as democracies or replaced some form of authoritarianism with democracy have also had parties from their democratic beginnings. Evidently, then, democracy and parties are in some fashion intertwined.

Why? Most scholars would say because parties perform at least two critical functions for democratic government that no other form of political organization does or can.

3.1 Organizing Elections

Popular elections of the government's top public officeholders are traditionally the main democratic device by which a sovereign people choose their officeholders and hold them accountable. But elections, especially in democratic nations with millions of citizens, must be organized somehow. For instance, 130 million Americans fulfill the constitutional requirements for the US presidency, and theoretically the American people have the right to pick the one person they most prefer. But that right is meaningless, for no person or electorate can choose among 130 million alternatives. Some agency must aggregate the choices and reduce the 130 million to a manageable number, say, 20 or three or two. That vital function is performed in every democracy by the political parties' activity of candidate selection. Theoretically it could be performed by another agency, such as the mass media or random selection by computers. In fact, however, democratic governments have always left the task mainly to political parties.

In addition, by putting the candidates' party labels on ballots the parties provide important cues that help voters make some sense out of the 'booming, buzzing confusion' of names, offices, and issues that elections present to them—especially where there are long ballots filled with many options. Some political scientists believe that prominent party labels explain why voter turnout in partisan candidate elections is usually much higher than in nonpartisan elections or in referendums.

3.2 Organizing Government

In every modern democracy elected officials of each party in each governing body join together in some kind of party organization. In each national legislature, regional assembly, or local council there is a comprehensive organization of all the party's legislative officeholders, often called the 'party caucus' or the 'parliamentary party.' That group usually selects some of its members as a guiding committee and one of its members as the party leader.

In the two-party systems elements of the majority party's organization become the government's 'ministry,' 'cabinet,' and 'prime minister.' In multiparty democracies the leader of one of the parties negotiates with other parties' leaders to put together a multiparty coalition government with some top positions going to leaders of all the participating parties. In both kinds of democracies the parties provide the organization that not only makes possible meaningful elections but effective and accountable government as well.

Some parties and party systems may perform these functions better than others, but even the least effective makes possible the coherent government and popular accountability that democratic government must have in modern populous nations.

See also: Civic Culture; Communist Parties; Electoral Systems; Participation: Political; Party Identification;

Party Responsibility; Party Systems; Political Culture; Political Elites: Recruitment and Careers; Political Machines; Political Money and Party Finance; Political Parties, History of

Bibliography

Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. Wiley, New York

Duverger M 1954 *Political Parties, Their Organization and Activity in the Modern State* (trans. North B, North R). Methuen, London

Eldersveld S J 1982 *Political Parties in American Society*. Basic Books, New York

Epstein L D 1967 *Political Parties in Western Democracies*. Praeger, New York

Epstein L D 1986 *Political Parties in the American Mold*. University of Wisconsin Press, Madison, WI

Keefe W J 1998 [1972] *Parties, Politics, and Public Policy in America*. 8th edn., Congressional Quarterly Press, Washington, DC

Key V O 1958 [1942] *Politics, Parties and Pressure Groups*. 4th edn. Crowell Company, New York

Merriam C E, Gosnell H F 1949 [1922] *The American Party System: An Introduction to the Study of Political Parties in the United States*. 4th edn. 1949 Macmillan, New York

Michels R 1958 *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. (trans. Paul E, Paul C). Free Press, Glencoe, IL (1st edn. in French, E. Flammarion, Paris, 1914)

Namier L B, Sir 1929 *The Structure of Politics at the Accession of George III*. Macmillan, London

Ostrogorski M 1902 *Democracy and the Organization of Political Parties* (trans. Clarke F). Macmillan, New York

Ranney A, Kendall W 1956 *Democracy and the American Party System*. Harcourt, Brace, New York

Rokkan S 1970 *Citizens, Elections, Parties; Approaches to the Comparative Study of the Processes of Development*. McKay, New York

Sartori G 1976 *Parties and Party Systems: A Framework for Analysis*. Cambridge University Press, New York

Sorauf F J, Beck P A 1992 *Party Politics in America*. 7th edn., HarperCollins, New York

A. Ranney

Political Parties, History of

The definition of a political party requires a historical perspective. The idea of a party is multifaceted and developed over time, in the light of the prevailing political, social, economic, and (especially) legal realities in which it is formed. In current language—and in the specialized terminology of the social sciences—the term 'party' has been in use for longer than the term 'social class'. Etymologically, the term derives from the verb *partir*, which, in archaic French, meant 'partake,' or 'share.' The present article regards political parties as formal civil organizations that are active in public life, in which they represent specific class, professional, or ideological interests, within the framework of the modern constitutional State, and whose goal is to run for election and, upon winning, to take power as the government.

1. Approaches and Definitions

In the past the expression 'political party' was also used to describe other political phenomena. Throughout the history of societies and of their political bodies, since antiquity, the expression to 'be party to' affairs of public life has been considered a common mode of human behavior. The translators of Aristotle use the term 'party' to designate opposing social groups in Athens. Specialists on ancient Rome have used the concepts of 'Plebeian Party' and 'Patrician Party,' and the famous *De viris illustribus Romae* relates to the party disputes that dilacerated the Roman Republic. The supporters of the Crown and the defenders of the rights of the nobility were also parties. Critics of the theological hegemony of the church and of the clergy were parties within the context of the political order of Christian Europe. Noble families in opposition to the Crown were parties in the making; religious and political reformers were parties. These are broad meanings of the term that are frequently to be found in the literature. Also under the broad heading of 'parties' can be categorized all manner of social organizations that seek political ends, even if they never stand for (or win) elections, as is the case for the mini-parties or revolutionary organizations that have no constituency among the electorate and that do not accept the validity of the electoral system, or even the dominant groups in totalitarian states. Thus, the expression 'political parties' is also often applied to interest groups; propagators of ideologies, doctrines, and political positions; political model makers; and mediators between citizens and government, including a vast multiplicity of bodies that participate in public life.

There are three nuances of the term 'party' from the historical point of view: rival groups, orders, states, or dynasties; theological or ideological critics; and modern formal social organization. This third interpret-ation has, in part, inherited the historical characteristics of earlier forms, making it possible to make such distinctions as conservative and liberal parties; religious and lay groups; corporative interests and those of the masses; revolutionaries and reformers; left and right; center and radical; and progressive and reactionary. Such distinctions are not exclusive, are directly proportional to the degree of institutional solidity and of integration which the parties enjoy in political society, and always relate to the ideas that separate—or unite—individuals or groups. Such multiplicity is evident in the classification of parties, according to the ideological density of their manifestos, the form of internal structure, competitive or

noncompetitive organization, and of party systems (single-party, two-party, or multiparty; unipolar, bipolar, or multi-polar) concerning the political and legal frame they are put in. Unlike political science, historical research does not treat these issues in isolation, but rather, seeks to reconstruct how all such aspects relate to each other.

Political parties—in the contemporary sense of the term—became common in the nineteenth century, alongside the development of representative institutions and the suffrage, becoming gradually universal, especially in Western Europe and the USA. In the aftermath of the Congress of Vienna (1814–15) political parties had not yet been formally recognized as civil associations. Any association, assembly, or grouping united by specific goal or political interest was regarded as a party. Conversely, the pursuit of private interests and the application of abusive pressures was regarded as a constraint upon political life, as the goal of winning elections and of influencing government and public policy became the established criteria for identifying a group of citizens as a political party. In the USA, by the 1830s, parties as such were fully active. In the Western European states, it was in the final decades of the nineteenth century and the first decade of the twentieth century that this development took place. Modern political parties are full players in the public life of democratic societies. In effect, only constitutional states formed and consolidated over the past two hundred years have the institutional framework necessary to ensure regular political participation on the part of their citizens. Such a framework determines the legal principles for party actions, articulates the position of the parties in relation the system of government and the electoral system, and places them within the framework of civil law. Their social origin and ideological orientation notwithstanding, it was the placing of parties within a legal framework that created the party system—understood both as a juridical and political frame of the participation of citizens in political life.

Parties began to elicit attention from philosophers and the interest of modern historians from the time of the English revolutions of the seventeenth century, on through the revolutionary turmoil of 1848, and up until the death of Marx in 1883. They were observed by Bolingbroke and Madison, Hume, Burke, Benjamin Constant and Blüntschli, and the German social-democratic thinkers. At the turn of the twentieth century, party-political phenomena began to be the object of serious scientific methodological analysis. The earliest works in this field were published after the consolidation of parliamentary forms of democratic government, which had been in place since the mid-nineteenth century. The pioneering works of Ostrogorski (1902) and Michels (1989) examined the functioning of parties in the late nineteenth and early twentieth centuries in the UK, France, and the USA. These authors sought to put forward a general theory, beginning with highly particular experiences and specific historical developments, such as the establishment of federative republicanism of the USA or the burgeoning power of Parliament in the Government of the UK. There are many descriptive and historical studies on political parties in specific countries and also comparative studies which seek to define possible models. Research trends generally give greater emphasis to the western model of representative democracy, in which formalized parliamentary, electoral, and political-party systems play an essential role in public life and the practices of Government (Epstein 1993; Budge and Keman 1990, Fenske 1994).

With Max Weber, who couches his remarks on parties from a sociological standpoint, modern political parties came to be understood as a product of democracy and of universal suffrage. Jean Charlot (1971), according to Lapalombara and Weiner (1966), provides a systematically historical view of parties. In order for a group to be regarded as a political party, it should fulfill four requisites. It must (a) possess a visible and stable organization, for the articulation of its local, regional, and national components; (b) ensure continuity of organization, independently of its current managers; (c) have the explicit goal of achieving and holding decision-making power, either alone or as part of a coalition with others, at all levels; and (d) have organizational efficiency, with a view to capturing the vote and harnessing popular support. It is, however, a work by Duverger (1951) that systematizes and compiles knowledge on political parties. Two of his research elements remain valid: one puts forward a theory of the origin and of the proliferation of political parties, from an institutional perspective, relating to the party phenomenon and the dynamics of institutions, whereas the other resides in a typology of parties based upon the nature of their organization. After World War II, the discredit and mistrust with which people regarded the traditional political parties of Europe—owing to their failure to avoid the emergence of fascism and totalitarian dictatorships in Italy, Germany, Spain, and Russia—hampered the reintroduction of democracy, and parties were left fragmented and disunited. The understanding came up that healthy democratic life requires political parties within a system of government that establishes clear rules for the formation of governments to exercise the will of the majority, solely on the basis of the judgment expressed through the popular vote. The political instability of the Fourth Republic in France, and of the government of Italy, for example, provided Duverger with a rich vein of material upon which to base the most ambitious essay written on the theme of political parties in the mid-twentieth century. Duverger regards the parties of the masses as the inevitable product of universal suffrage and regards many of the residual parties as being archaic, especially in postwar France. His view is supported by the movements for ideological concentration in European

public life at the onset of the Cold War and has as its inspiration the two-party systems—which he regarded as being successful—in the USA, UK, and West Germany.

Although there is still a lack broad-based comparative studies, the contributions of Robert Dahl (1966), Lipset and Rokkan (1967), and Epstein (1993), take steps in this direction. The ideological nature of parties is given an innovative treatment by Stein Rokkan, comparable with the contribution of Duverger on party organization. The analysis of party systems was advanced by Sartori (1976) and Blondel (1978), whereas electoral sociology—beginning with the pioneering work of André Siegfried on electoral statistics and geography (1913, 1927), by Richard Rose (1984), David Butler (1981), and Alain Lancelot (1988). More recently, Daalder and Mair (1983), Pierre Avril (1990), Klaus von Beyme (1985), and Daniel-Louis Seiler (1986) have sought to formulate studies from an institutional and systemic viewpoint.

2. *Stages of Development (Phases) Change Over Time*

The first expressions of party life began with the defense of specific interests, although in different contexts and times, such as those of the aristocracy against the power of kings and governments, or of the working class against the political, economic, and social dominance of ruling groups or elites. Such defense was not passive. It was carried out by means of consistent and prolonged actions within the representative institutions, such as the estates, consultative assemblies and parliaments. It also comprised the revolutionary movements in which the respect for rights—both effectively held and merely claimed—was demanded of the political authorities, the governing establishment or the despot.

A party emerges when a political current within society finds expression through a specific (national or regional) organization. The first steps towards the organization of a political group can be observed through the 'parties' into which England polarized in the mid-seventeenth century. Although these parties were still characterized by the discipline and discourse of religion, the political effect of the change from one system of government to another imposed by those in opposition to the power of the throne, ushered in a new form of the exercise of power within society. It goes without saying that the great revolutions of the seventeenth and eighteenth centuries—in England, France, and the USA—also gave rise to factions which the majority of historians would call 'parties.'

The typical action of parties in modern societies emerged out of the political consciousness that defined and established the notion of a universal citizenship, evolves in the new space for political protagonists (both for and against the State), the emergence of organized political parties during the nineteenth century and investiture of political parties as being the only protagonists capable of adopting practices, expressing ideas, and taking power and exercising public office within the State. The parties of the bourgeois in the early nineteenth century protagonized opposition between the aristocracy and the monarchy; a similar process can be observed among the emergent revolutionary or proletarian parties of the end of the same century. In revolutionary France there were embryonic experiences of political-party organization: the Girondes, Montagnards, and the Jacobins. The greater part of these 'parties' consisted of small groups, without formal organizational structures. It can thus be stated that no formal parties existed prior to 1820. Three phases in the development of party systems can be identified in the nineteenth century: the earliest phase between 1820 and 1830, an intermediate phase in the 1860s, and a final phase around 1880.

In the 1820s and 1830s, the British Tories and Whigs emerged and consolidated long-held positions. At the same time, in many countries of Western Europe, and notably in France, there was a movement towards a more constitutional state, expanding suffrage and consolidating the system through application the principle of majority rule in the formation of governments. It was in this phase that the earliest political parties emerged in Latin America, as in the case of the *Partido Liberal* in Colombia (1815) and the *Colorado* and *Nacional* parties in Uruguay (1836), which arose as an immediate consequence of the wars of independence against Spain and of the ideological influence of the USA. In the intermediate phase, there was a clear expansion of political parties: the Conservative (1840) and the Liberal (1860) parties in Brazil; the Liberal Party in Belgium (1846); the Christian-Democrat and Radical-Democrat parties in Switzerland (1848); the Democratic (1848) and Republican (1856) parties in the USA; the Liberal Party in Italy (1848); the *Partido Radical* in Chile (1863); and the Conservative (1854) and Liberal (1867) parties in Canada.

In Latin America, where most countries had thrown off colonial rule in the first decades of the nineteenth century, only gradually did political parties begin to gain importance. As of 1850, the national aspirations of countries of South America came under the spreading influence of North American two-party system. The first expression of political life was the affirmation of a national identity. The main banner of the 'political parties' up until the mid 1800s was nationalism, defended as the means whereby these new states could affirm their independence in face of their former colonial masters. The 1880s witnessed the emergence of 'classical' political parties, generally in twos, and of either conservative or liberal persuasion—but always nationalist—in the former Spanish colonies. It should be noted that the emergence of political parties did not necessarily mean that they assumed a decisive role in political life. Under the Brazilian Empire (1822–89), a

two-party system existed as a means for the rural aristocracy to maintain its hold on power and, under the Republic proclaimed in 1889, this oligarchic system remained essentially unchanged.

In the mid-nineteenth century, growing social tensions and bitter social criticism led people with a political outlook to express their social and economic (and occasionally revolutionary) demands through formal political action, by means of the creation of political parties. Thus began the 'socialist' or 'social-democrat' phase, typified by formation of the Danish Social-Democratic Party (1871), the Spanish Socialist Party (1879), the German Social-Democratic Party (1863, 1869, 1875), the Swiss Socialist Party (1880), the Belgian Workers Party (1885), the Indian National Congress (1885), the Austrian Social-Democratic Party (1889), the Swedish Social-Democratic Party (1889), the Italian Socialist Party (1892), and many others.

World War I forced a pause for reflection and redefinition of the party systems in Europe. On the one hand, the victory (1917) of the Bolshevik party (1903), which upon seizing power set up a revolutionary and internationalist government in Russia, and, simultaneously, the conservative bias of the Treaty of Versailles (1919) on the other, mark the polarization of two separate and irreconcilable ideological currents. On one side, moderate, conservative democratic, and liberal parties, found themselves up against hard-line right or left-wing revolutionary opponents of a nationalist or internationalist persuasion, with scant respect for suffrage, democracy, or tolerance. The twentieth century opened with the founding of a series of communist parties (Russia 1903; Austria 1918; Greece 1918; USA 1919; UK 1920; Spain 1920; Italy 1921). These new protagonists on the political scene achieved varying degrees of importance in different places and their role was to become more prominent during the Cold War. Communist ideology and the proposals of revolutionary Marxist analyses were to assume the weight of a political mortgage at two moments: in Europe in the 1920s and 1930s, and throughout the world from 1945 to 1990. Social demands and economic crises deeply disturbed the political environment. The emergence of totalitarian and fascist dictatorships were to undermine political party structures throughout Europe until 1945. Nevertheless, these dictatorships organized political life by means of an imposed adherence to a single party. Only formally such a party was maintained as an element of legitimacy in the exercise of state power, despite the fact that authoritarian regimes could dispense this.

The success of the Western model of parliamentary democracy based upon a three-legged political system (representative and participative democracy, the electoral system, and the political-party system) is thus a recent phenomenon. During the Cold War, the ideals of representative democracy as being the end product of the modern liberal constitutionalism, with roots in the 1790s, appeared to be a heritage limited principally to Western Europe and the USA. A few other countries, notably former members of the British Empire—such as Canada, Australia, New Zealand, and India—have practiced representative democracy since achieving their independence. In countries such as Brazil and India, where democratic pluralism exists under specific historical and cultural conditions, many political activities take place outside the formal political-party structure.

Between the world wars, growing social awareness of rights and of the need to manage an increasingly literate populace led to structural modifications among the political parties. Economic crises in Central Europe were accompanied by increasing mistrust of parliamentary institutions and the emergence of manipulative right-wing political parties. Fascism and communism radically changed the way in which political parties were regarded. Into the parties of the masses flowed all the social dissatisfaction and frustration, making them easy prey for manipulation by dictators of every stripe. In Europe, as also in other regions, the trend towards mass parties, which in principal appeared to be a sign of the triumph of democracy, in the end generally resulted in quite the opposite effect as authoritarian rulers assumed power. In many regions, however, expectations with respect to the freedom to participate in party activities was enhanced, especially after the experience of World War II.

The triumph of the Western model of liberal democracy with parliamentary representation was one of the consequences of the new international order that emerged after 1945, except in regions under the yoke of communism (Eastern Europe, China, and Cuba). Communist regimes, however, do not dispense with mass parties or even with formal parliaments, although the role of these institutions is merely figurative. After World War II, a new phase of political life began, with the restoration of the values of liberal democracy, universal suffrage, civil liberties, and parliamentary representation. The Western tradition of recording political events—at least since the Code Napoleon (1804) and the British Reform Act (1832)—has provided a framework for the organization of political life in such a manner that the formal and legal structuring of parties has advanced in parallel to the campaign for constitutional government. This has caused parties to organize and represent themselves as both defenders of an overall social vision and as pursuers of political goals. Thus, parties assume the posture of political businesses whereby they compete within a legal framework to advance the political convictions, opinions, and beliefs—and to defend the preferences and class interests—of their constituents. Diverse and intersecting interests find expression among groups, clubs, associations, and, above all, political parties. The social economic and ideological cut-lines present in the social structures were key

factors throughout the nineteenth century. Conservatives and Liberals, labor parties and revolutionaries, opposed each other in Europe and also in Latin America. In the UK, the Conservatives and the Liberals succeeded the Tories and the Whigs and were in turn superseded by the Conservative and Labour parties of the twentieth century. In the USA in the period following the Civil War, the Democrats and the Republicans definitively consolidated their hegemony over the political scene (Katz and Mair 1992).

The history of political parties clearly demonstrates that the political opportunities in the social competition for power in the State has an effect upon the parties that, for their part, have an effect upon the structure of political opportunity, i.e., parties and institutions mutually affect each other. In democracies which work regularly, disciplined voters arise, as in the European Union, for example. These voters do respond to party activities, although they are increasingly critical of the means whereby parties pursue their ends. 'Belonging' to a party tends to be more a sociological identification with a political tendency, rather than involving formal party affiliation. This suggests a salutary clarification of the boundaries between parties and their environment. The process whereby a party is identified with its leader is an interesting historical phenomenon in the political organization of societies. For example, the Braybrooke's Manuscript (1788) refers to the parties 'of the Crown,' 'faithful to Mr. Pitt,' and 'of Mr. Fox.' Another interesting and more contemporary example is of the emergence of a 'Gaullist' party in France, after 1946. The tradition of having a strong personal leader continues very much alive in Latin America, even in the late twentieth century. This tradition has influenced the social practices of parties in various ways. In Mexico, the Institutional Revolutionary Party—PRI (1911) dominated the Mexican political scene single-handedly until 2000.

With the advance of democracy and of democratic negotiating practices, conducted with a view to making government of contemporary states feasible, all parties (including the new parties that emerged out of the former communist parties of Eastern Europe, and also certain parties in Western Europe) have had to become connected. The difficulty of winning elections based solely upon their different platforms and making their will prevail has, since the 1970s, obliged parties to emphasize their capacity to draw together to form and support coalitions. Finally, all modern parties are obliged to adopt business-like organizational structures or run the risk of failing to fulfill the demands of modern life, in terms of communication, social penetration, and funding.

Evidently, the aim of assuming the mantle of government remains the principal motivation of political parties. The historical trajectory of activities undertaken by political parties is closely associated with an awareness of the structure of political opportunities within a given political system. Revolutionary organizations that reject existing institutions, social clubs, debating societies, pressure or advocacy groups with independent objectives can be regarded as entities that discreetly play significant political roles without being formally political parties. Formal political parties must, as such, compete to control a political process whose agenda they do not establish by themselves, but which cannot function without their participation. This experience is deeply ingrained in the political parties of the late twentieth century.

3. *Unanswered Questions and Issues for Debate*

Some important questions remain unanswered. For example, the emergence of political participation—as an alternative to traditional representation in liberal democracies—is often treated both as an essential element for modernization and as its critical consequence. To assume the mantle of government and to perpetuate the values of parliamentary democracy, strong parliamentary institutions are necessary if modern political parties are to perform their roles. This is true even in countries such as the USA—and many others in Africa and Latin-America—in which the presidency is a strong and 'personalized' institution. This applies even in cases where presidents are elected by direct vote and are endowed with special powers, as occurs in France and in Portugal.

Although much is known with respect to the details of the emergence of specific party systems, the capacity to generalize as to the reasons why different types of parties emerge in the ways and forms in which they have done so, remains incipient. There are few works that examine or seek to explain the vast variety of different contexts, and comparative studies among regions with different levels of development are rare. In the 1980s, the political development of multinational regional institutions—which initially had a solely economic focus—such as the European Economic Community, ASEAN, and (in the 1990s) MERCOSUR, has unveiled a whole new field for research into the political life of such institutions, their relations with member states, and the entire theme of supranational integration. From a historical-research perspective, only now are the first comparative studies of the political parties being carried out. This applies not only to research on existing parties in specific countries or groups of countries (as is the case in the European Union), but also with respect to an examination of the historical roots of contemporary political parties.

Not only parties, but also party systems can be analyzed from the institutional standpoint. Both the parties and the party systems of the 1970s were little different from those of the 1920s. With the exception of the ecological movement, the majority of the party groupings, i.e., similar parties, can be presented

according to their similarities. Practically all of the present parties—of the left, of the right, of the center; liberal and revolutionary parties—consolidated themselves in the 1920s, and some much earlier. Political parties are present at the heart of all activities of the state in liberal democracies. However, even under communist regimes the role of the political party is performed in a well-established manner. In the 1990s, however, it could not safely be predicted that political parties would face a future with development and prestige, similar to that they enjoyed in the recent past. At the present time, a new type of political culture is emerging. Political awareness and politics—as subject for discussion—open up a new avenue of research into general participation in political life on the threshold of the twenty-first century. Culture may help a understanding as to why certain social cleavages are successfully transposed by a stable party system, whereas others are not. Political culture and the manner in which it is created is responsible for some significant differences in the traditional organization of states, or of bodies associated to them. Political culture includes not only ideals, such as values, beliefs, attitudes, but also political practices and societal behavior.

There are certain themes in the study of political parties that still require further research from a historical point of view. Politics, with the constrained use of political power, and the politics of containment have been recurrent themes in political science since the 1970s. One of the most important contentions of political activity is the institutional framework of the state. Contemporary political research has made clear the importance of institutional factors in political life. Aside from this new institutional focus, the study of political parties has demonstrated a renewed appreciation of history, of events, of the rules and of political regimes as the active forces of political life. It is perhaps commonplace to state that 'history matters.' The 'frozen cleavages' of Lipset and Rokkan, the development models of communism, fascism and the parliamentary democracy of Moore (1968), the theories of 'critical realignment' of Burnham (1970), are good examples of historical reviews and models. The legacy of history is thus one of the limitations to which institutionalism is subject.

Socio–economic constraints are generally accepted without going into the mythical and overworked concepts that have been served up since the time of Marx. Rapid changes in international finances and economic scenarios have produced intriguing works. The exercise of social power is also constrained by cognitive means, in the use of pure reason (and more specifically, of the practical one). The knowledge society and information society will have to take into consideration non-rational features of political life, and the social and ideological factors that influence the belief systems of the masses (Meny 1989). The final two decades of the twentieth century witnessed a swift retreat in the positivist distinction between fact and value. Specialists recognized that political agents were also ethical players (Rawls 1993), who have their own values and act in accordance with them, and that occasionally they may be persuaded (by political philosophers, perhaps) to adopt other, better values. Any attempt to divorce facts and values from the mental processes and from the underlying political dynamics, would be doomed to failure. Similarly, political science increasingly seeks to develop studies that take into account the complexity of the structures, processes, and results that it investigates. In order to do this, one must have a theoretical framework within which the history of the processes and the prevailing political philosophy must occupy a space which formally would have drawn opposition from previous generations.

See also: Conservatism: Historical Aspects; Democracy, History of; Interest Groups, History of; Liberalism: Historical Aspects; Parliaments, History of; Party Systems; Political History (History of Politics); Political Parties; Socialism: Historical Aspects

Bibliography

Avril P 1990 *Essais sur les Partis Politiques*. Payot, Paris

Budge I, Keman K 1990 *Parties and Democracy. Coalition, Formation and Government Functioning in Twenty States*. Oxford University Press, Oxford, UK

Butler D 1981 *Democracy at the Polls*. AEI, Washington DC

Charlot J 1971 *Les Partis Politiques*. Payot, Paris

Duverger M 1951 *Les Partis Politiques*. Colin, Paris

Epstein L D 1993 *Political Parties in Western Democracies*. Transaction Publishers, New Brunswick NJ

Fenske H 1994 *Deutsche Parteiengeschichte*. Schöningh, Munich, Germany

Galli G 1990 *Storia dei Partiti Politici Europei*. Rizzoli, Milan, Italy

Goodin R E, Klingemann H-D (eds.) 1966 *A New Handbook of Political Science*. Oxford University Press, Oxford, UK

Graham B D 1993 *Representation and Party Politics. A Comparative Perspective*. Blackwell, Cambridge, UK

Huntington S P 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, NJ

Kaase M, Newton K (eds.) 1995 *Beliefs in Government*. Oxford University Press, Oxford, UK

Katz R S, Mair P (eds.) 1992 *Party Organizations. A Data Handbook on Party Organizations in Western Democracies, 1960–90*. Sage Publications, London

Lancelot A 1988 *Les élections sous la Vème République*. PUF, Paris

LaPalombara J, Weiner M (eds.) 1966 *Political Parties and Political Development*. Princeton University Press, NJ

Lawson K (ed.) 1994 *How Political Parties Work: Perspectives from Within*. Praeger Publishing, Westport, CT

Lipset S M, Rokkan S 1967 Cleavage structures, party systems, and voter alignments: an introduction. In: S. Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments: Cross-national Perspectives*. Free Press, New York

Lloyd C (ed.) 1983 *Social Theory and Political Practice*. Wolfson College Lectures. Clarendon Press, Oxford, UK
Lloyd C 1993 *The Structures of History (Studies in Social Discontinuity)*. Blackwell, London
Lovenduski J, Pippa N (eds.) 1993 *Gender and Party Politics*. Sage Publications, London
Maisel L, Cooper J (eds.) 1978 *Political Parties: Development and Decay*. Sage Publications, London
Maisel L, Sacks P M (eds.) 1975 *The Future of Political Parties*. Sage, Beverly Hills, CA
Mayhew D R 1986 *Placing Parties in American Politics: Organization, Electoral Settings, and Government Activity in the 20th Century*. Princeton University Press, NJ
Meny Y (ed.) 1989 *Idéologies, Partis Politiques, et Groupes Sociaux*. Presses de la FNSP, Paris
Merton R K 1968 *Social Theory and Social Structure*. Free Press, New York
Michels R 1989 *Soziologie des Parteiwesens*. Kröner, Stuttgart, Germany
Moore B Jr 1968 *Injustice. The Social Bases of Obedience and Revolt*. Princeton University Press, NJ
Moore B Jr 1998 *Moral Aspects of Economic Growth and Other Essays*. Cornell University Press, Ithaca, NY
Norton P (ed.) 1996 *The Conservative Party*. Oxford University Press, Oxford, UK
Offerlé M 1997 *Les Partis Politiques*. PUF, Paris
Ostrogorski M 1903 *La démocratie et l'Organisation des Partis Politiques*. Calmann-Lévy, Paris
Pippa N (ed.) 1999 *Critical Citizens: Global Support for Democratic Government*. Oxford University Press, Oxford, UK
Pombeni P 1994 *Partiti e Sistemi Politici Nella Storia Contemporanea*. il Mulino, Bologna, Italy
Rawls J 1993 *Political Liberalism*. Columbia University Press, TN
Rohe K (ed.) 1990 *Elections, Parties and Political Traditions. Social Foundations of German Politics and Party Systems, 1867–1987*. Berg, New York
Rose R 1984 *Do Parties make a Difference?* MacMillan, London
Schumpeter J A 1942 *Capitalism, Socialism, and Democracy*. Harper and Row, New York
Seiler D L 1986 *De la Comparaison des Partis Politiques*. Economica, Paris
Skocpol T (ed.) 1999 *Democracy, Revolution, and History*. Cornell University Press, Ithaca, NY
Tufte E R 1980 *Political Control of the Economy*. Princeton University Press, NJ
Ware A 1996 *Political Parties and Party Systems*. Oxford University Press, Oxford, UK
Weber M 1919 *Politik als Beruf*. J. B. Mohr, Tübingen, Germany

E. C. de Rezende Martins

Political Protest and Civil Disobedience

Political protest refers to a multitude of methods used by individuals and groups within a political system to express their dissatisfaction with the status quo. Although protest can be registered through voting, letter writing, campaign contributions, participation in interest groups and political parties, and other traditional means, the term more commonly refers to those kinds of mass political activism that occur outside of conventional parliamentary channels, such as demonstrations and marches, sit-ins, blockades, labor stoppages, boycotts, and other methods that combine communication and persuasion, direct action, and noncooperation with political, economic, and social institutions.

Much social science research on political protest has concentrated on explaining organized, sustained forms of mass collective action such as social movements. A social movement consists of an array of organizations and individuals who share a common identity and agenda of social and political goals, and who are regarded as a collective entity by other political actors. Social movements typically pursue their goals by mixing protest actions with conventional political strategies such as lobbying, litigation, public education, and electoral mobilization. Discontented groups often resort to direct action when they are denied equal access to conventional means of participation or when such avenues are regarded to be ineffectual or inadequate for registering the intensity of their grievances. Theoretical and empirical studies of political protest have analyzed the emergence of group consciousness and solidarity, the framing of grievances, the role of social networks and community institutions in fostering cooperation, the formation of social movement organizations, the psychological make-up and motivations of individuals who engage in protest, the evolution of strategy and tactics in social movements, the influence of changing political opportunities on the course of a movement, and the efficacy of violent and nonviolent methods of protest.

Unlike protest marches, picketing, and other kinds of legal protest, civil disobedience is a particular form of political protest that is marked by the deliberate violation of the law for social purposes. Civil disobedience is not merely conscious law breaking, but rather it is law breaking with the express goal of correcting a perceived injustice. Action that would now be regarded as civil disobedience has been a regular feature of social protest throughout history even though the term itself, attributed to an 1849 essay by Henry David Thoreau, is a relatively recent creation.

The act of civil disobedience is fundamentally an act of political protest because the noncompliance is intended to convey a message of discontent to the government, media, and public. The acts of defiance are typically overt, public acts aimed at marshaling collective opposition to unjust or unconstitutional laws and policies. Civil disobedience is not restricted to protests over laws and public policies, but is also employed against the practices of private institutions such as corporations and universities, as occurred for example in the divestment campaign on American university campuses in the early 1980s, which pres-

sured universities (both private and public) to cut their financial ties to companies that conducted business in South Africa in an effort to bring an end to apartheid.

1. Forms of Civil Disobedience

Within this broad conceptualization, civil disobedience can emanate from different motivations and assume a number of forms. Civil disobedience can be active or passive, direct or indirect. An individual can either actively commit a prohibited act or passively refuse to conform to a prescribed action. For example, an anti-war protester may destroy his draft card or, alternatively, he may refuse to register for the draft. Both actions qualify as civil disobedience because the protester intentionally violates the law as an expression of his opposition to the war, knowing the legal consequences of his actions. Thoreau himself engaged in a form of passive civil disobedience when he chose to register his opposition to the institution of slavery by refusing to pay the Massachusetts state poll tax.

Civil disobedients often break laws that are unrelated to the law or policy they are protesting. For example, peace activists who disrupt operations at defense laboratories are not protesting laws against trespassing, but are using civil disobedience to dramatize their opposition to nuclear weapons. Supreme Court Justice Abe Fortas objected to the indirect use of civil disobedience on the grounds that it would give licence to people to break any law they wished in the service of a goal that they perceived to be more noble. But few scholars have subscribed to this restriction because, on many issues, there is no feasible direct method to engage in civil disobedience. There may, for example, be no direct method by which activists can use civil disobedience to protest the government's foreign policy or its failure to pass laws addressing issues of inequality and poverty.

A narrower definition of civil disobedience would exclude illegal actions taken by individuals who do not accept the constitutionality of the law they are violating. By this definition, an activist who breaks the law by setting fire to the American flag is *not* engaging in civil disobedience if he believes that his actions are protected by the First Amendment, but a flag burner who accepts the constitutionality of a law against flag burning *is* engaging in civil disobedience. For the purpose of defining civil disobedience, however, it may be unnecessary to make such a distinction based on the activist's frame of mind, because the crucial assumption underlying civil disobedience is that the law being broken is generally viewed to be valid by those in authority, so the actions of the civil disobedient are generally viewed to be illegal. Therefore, both flag burners are committing civil disobedience in the sense that they are breaking what is generally understood to be the law in an effort to change it, and the main difference between them is whether they are appealing to the courts or to the executive and legislature for remedial action.

Although Thoreau assumed that 'civil' protest would be conducted in an orderly and peaceful fashion, scholars disagree over whether civil disobedience is necessarily nonviolent by definition. Some commentators insist that civil disobedience must be nonviolent or else it cannot be distinguished from rebellion, rioting, and other types of violent direct action. The two individuals who are most responsible for the association of civil disobedience with nonviolence are Mohandas Gandhi (see e.g., Gandhi 1961) and Martin Luther King, Jr. (see e.g., King 1964). Indeed, contemporary understandings of civil disobedience have been substantially shaped by the mass campaigns for equal rights organized by Gandhi in South Africa and India and by King in the United States.

Both Gandhi and King possessed the extraordinary combination of political 'savvy' and unwavering moral and spiritual commitment to the philosophy of nonviolence. In his teachings, Gandhi insisted that civil disobedients must scrupulously obey laws that are unrelated to the target of the disobedience campaign and submit themselves to the punishment of the state for breaking a law they believed to be unjust. Civil disobedients do not flee or resist arrest, for in accepting penalties for their actions, they bring attention to the injustice they wish to correct and arouse the sense of injustice in others. Gandhi's concept of *satyagraha* or 'truth-force' refers to the moral conversion of one's adversary. In their quiet suffering, civil disobedients demonstrate their honor and integrity and their capacity to differentiate between just and unjust laws.

King regarded Gandhi as his mentor and shared his belief that the purpose of civil disobedience is to persuade the state and its citizens that an existing law violates a higher moral principle. Violence, therefore, must be avoided because it potentially undermines the communicative intent of civil disobedience. Adherence to nonviolence, on the other hand, indicates that one bears no malice or hatred for one's opponent. Like Gandhi, King believed that the ultimate goal of a campaign of nonviolent protest is to achieve a new moral understanding with one's opponents. Nonviolent resistance works, not by defeating and humiliating the opponent, but by winning the opponent's empathy and friendship.

2. Justification of Civil Disobedience

In addition to disagreements over what kinds of actions qualify as civil disobedience, there is debate over the circumstances in which civil disobedience can be justified. Civil disobedience raises issues about the nature of authority, obligation, and consent in society.

The practice originates from the idea that a person must be true to his conscience and that he is not obligated to submit to the laws or commands of the state when they violate his conception of justice. Civil disobedience therefore assumes both a right to individual freedom and the existence of a natural law that takes precedence over civil law.

Philosophical discussions of the justification for civil disobedience generally fall between legal absolutism and moral individualism. In justifying civil disobedience, it is necessary to accept that there are standards against which governments can be evaluated, and that citizens have the right to exercise such judgments. Few would argue that the law must always be obeyed; and similarly, few would contend that obeying the law is entirely a matter of individual judgment about what is morally right and wrong. One extreme supports despotic governments, while the other suggests anarchy. Whether civil disobedience undermines respect for the law is a matter for empirical inquiry. The penalties for breaking the law are usually sufficiently great that it is unlikely that people will engage in civil disobedience unless they are intensely motivated by a sense of injustice.

But do the moral criteria for civil disobedience reside only in conscience? Or are there intersubjective criteria that permit rational assessment of the morality of a law or policy? Without such objective criteria, we are in danger of every individual invoking arbitrary subjective criteria for disobeying different laws. We invite moral anarchy unless there are some agreed-upon criteria for evaluating whether a law or policy is unjust or immoral. At a minimum, the civil disobedient must be able to demonstrate that the law or policy violates principles of welfare, equality, or justice. The law or policy must be shown either to do more harm than good, to distribute benefits unequally, or to operate without the consent of certain groups. Legal methods of petition to change the law must be tried first, and the activists must grant the same rights of civil disobedience to others who believe that another law is unjust. Some would argue that the civil disobedient must also have a reasonable chance of overturning the law. According to these normative criteria, only those who believe in reciprocity and adhere to a sophisticated set of general rules of political conduct can rightfully be said to be engaged in justifiable acts of civil disobedience.

3. *Effectiveness of Civil Disobedience*

What are the mechanisms that make civil disobedience effective in practice? Gandhi recognized that civil disobedience could not be effectively employed for a global, distant goal such as national independence, but had to be used for specific, local issues on which the opposition had the power to compromise and make concessions. True to form, civil disobedience campaigns led by Gandhi were most successfully applied in local disputes concerning such issues as land revenue assessments and mill workers' wages, in which small numbers of close knit, disciplined protesters were united by their interest in the issue and capable of conforming to a norm of nonviolence.

In theory, if civil disobedience is to work through persuasion, then the protesters must invoke principles or values that are shared by a large proportion of the population. In practice, however, the preferences that are converted are more likely to be those of the spectators to the conflict than those of the adversaries of the protesters. A small number of disobedients can seldom, by their actions alone, render a law inoperative, but they may be able to disrupt the social order sufficiently to draw public attention to their grievances. Although Gandhi believed that the suffering of protesters in jail would apply pressure on the government, he also recognized that the state's use of violence would serve the interests of the protesters by influencing third parties who were witnesses to the conflict. British repression of Indian dissenters confirmed to onlookers that the British ruled by force and not through the consent of the Indian population. Similarly, in the course of the civil rights movement, it was the contrast between the nonviolent protesters and the violent response of the southern police that helped to galvanize public opinion and mobilize federal government intervention in support of racial equality.

In response to a campaign of civil disobedience, the defenders of the status quo must decide whether to offer concessions to the protesters or to escalate the conflict by using force or summoning the power of the state. It is the threat of continued disruption of the normal operations of business and government that 'persuades' the targets of the protest to negotiate with the activists and to yield to their demands. During the civil rights movement, the early success of direct action campaigns in some southern cities aimed at forcing businesses to provide equal treatment to blacks made it easier for similar campaigns to succeed elsewhere because businesses foresaw the normal course of these protests and sought to cut their losses sooner rather than later. There is little evidence that the attitudes of white southerners were converted by direct action; indeed, the attitudes of many were hardened by the confrontational tactics of the movement.

Civil disobedience thus rarely works through persuasion and attitude change alone, but also depends significantly on pressure and coercion. Actions that are aimed at transforming preferences may leave attitudes unchanged, but still be effective at promoting social change because the opponents are pragmatic enough to calculate that they are better off making concessions than dealing with the disorder created by continual protests. Civil disobedience in this respect resembles a pragmatic tactic within a repertoire of potential strategies that are available to a protest

group rather than an option that is resorted to only when the conventional channels of participation have been exhausted.

See also: Citizen Participation; Civil Rights Movement, The; Civil Society, Concept and History of; Collective Behavior, Sociology of; Nonviolence: Protective Factors; Peace and Nonviolence: Anthropological Aspects; Peace Movements, History of; Social Movements: Resource Mobilization Theory; Social Movements, Sociology of

Bibliography

Bay C 1968 Civil disobedience. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. MacMillan, New York

Chong D 1991 *Collective Action and the Civil Rights Movement*. University of Chicago Press, Chicago

Gandhi M K 1961 *Non-violent Resistance*. Kumarappa B (ed.). Schocken Books, New York

Gewirth A 1982 *Human Rights: Essays on Justification and Applications*. University of Chicago Press, Chicago

Greenawalt K 1987 *Conflicts of Law and Morality*. Oxford University Press, New York

King M L 1964 *Why We Can't Wait*, 1st edn. Harper & Row, New York

McAdam D, McCarthy J D, Zald M N 1988 Social movements. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA

Powers R S, Vogele W B 1997 *Protest, Power, and Change: An Encyclopedia of Nonviolent Action from ACT-UP to Women's Suffrage*. Garland Publishing, New York

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Sharp G 1973 *The Politics of Nonviolent Action*. Porter Sargent, Boston

Zinn H 1968 *Disobedience and Democracy: Nine Fallacies on Law and Order*. Random House, New York

D. Chong

Political Representation

Political representation is a process in which one individual or group (the representative) acts on behalf of other individuals or groups (the represented) in making or influencing authoritative decisions, policies, or laws for a polity. The *representative* may hold any of a variety of offices (executive, administrator, ambassador, judge, lobbyist, party leader) and may even be a collectivity (a governmental agency or political organization), but the modern concept most commonly refers to the role of a legislator. Similarly, the *represented* may be almost any individual or group, but is usually understood to be those individuals or groups, typically called constituents, who are entitled to select the legislators. Some important work examines the relationship between legislative and other forms of representation (Stimson et al. 1995), but most has focused on one or more of three dimensions of legislative representation: relationships between the representatives and the represented; among the represented; and among the representatives.

1. Origins

Political representation in its modern form evolved over a long period from about the fourteenth to the seventeenth century as kings and nobles convened assemblies to collect revenue and conduct wars. The bodies, usually composed of members from various estates or classes, gradually assumed independent power, and came to be thought of as representative institutions. In the debates during the Civil War period in England, defenders of Parliament connected the idea of representation clearly to the idea of consent of the governed. In 1651, the members of Parliament were for the first time referred to as 'representatives' (Pitkin 1967; also see Larsen 1966, Manin 1997).

By the eighteenth century the link between representation and popular rule had become sufficiently close that Baron de Montesquieu could praise the English constitution for establishing a republic in a large state by giving the people a voice through their representatives in the legislature. Defending the new American constitution, James Madison argued that representation not only made republics possible in large territories, but also mitigated their dangers, especially the influence of factions (Wood 1972) Representative government was not yet representative democracy, because much of the adult population could serve as neither representatives nor the (actually) represented. It was John Stuart Mill who in the late nineteenth century provided the framework for the theoretical synthesis that combined the traditional virtues of representation and the emerging ideals of democracy (Thompson 1976).

2. Legislators and Constituents

At least three aspects of the role of the representative should be distinguished. The first concerns the *mode* of representation—how constituents are represented. It expresses a continuum in which the limits are specified by the traditional contrast between trustee and delegate. The second aspect is the *locus* of representation—which citizens are represented. Among the possibilities are the nation as a whole, a political party, the state or district, and various groups and organizations. The district is more complex than is often assumed. For example, many legislators are bound to

local elites and national interest groups whose interests may diverge from those of other groups in the district, and some legislators see themselves as representatives of under-represented groups, such as blacks and women. The third aspect refers to variation in the *content* of representation—the types of issues with which a representative deals. A legislator may act, for example, as a delegate of the district majority on welfare issues, as a delegate of the national majority on civil rights, but as a trustee of both on foreign policy.

Given this complexity, it is not surprising that the most plausible normative theories refuse to privilege any one of these roles; what each representative should do depends on the nature of the particular issues, what other representatives are doing, and the biases in the political system. Similarly, when political scientists use the simpler conceptions of role, they rarely find any general correspondence between a representative's conception of role and actual behavior. The most robust accounts of legislators' conceptions posit decision trees with multiple branches and a complex set of factors (Kingdon 1981, p. 244).

On any account of the role of the representative, there is the further question of how representatives can be held accountable. The methods vary, depending on the political system, and on assumptions about what political accountability means. The form of accountability that has received the most attention is the electoral connection between representatives and constituents. Much of the empirical literature on the linkages between the policy preferences of constituents and legislators has been influenced by the classic model presented by Miller and Stokes (1963), which highlights four major variables: legislators' personal attitudes, roll call votes, constituency opinion, and legislators' perceptions of constituency attitudes. The vast literature on this linkage shows that the conditions under which each has an effect are quite variable. It is now recognized that roll call votes fail to capture much important legislative activity, and do not always reflect the actual views of legislators. A still unresolved puzzle is why individual representatives (even in safe seats) exert so much effort to please their constituents and collectively act in ways that track mass public opinion even while the evidence indicates that constituents and the mass public are not well informed about what their representatives believe and do (see Uslaner 1986).

A related trend is the growing importance of constituent service (a practice in which representatives help constituents and groups in dealing with their particular problems with government, such as immigration and state and federal grants). The electoral effect of constituency service remains a subject of some controversy among political scientists, but there is agreement that the practice is now more prevalent (Cain et al. 1987). Another effect that has provoked important work is pecuniary. Since at least the eighteenth century students of legislatures have studied and often criticized the influence of money and wealth on the behavior of representatives. Political scientists have not generally found strong correlations between contributions and roll call votes on broad policy areas even in the US system. But they have found that money may buy access—marginal time, energy, and legislative support.

3. *Constituents*

All legislative representation is group representation. Even in societies governed by individualist norms such as 'one person-one vote,' the definition of the constituency implicitly identifies a group or groups that are to be represented.

Party-based representation is associated with a modified form of proportional representation in which the share of votes that each party receives determines how many of the candidates from its lists win seats in the legislature. The system usually produces multiple parties, and thus gives voters a wide range of ways to express the interests they wish to have represented. However, it gives voters less choice about which particular candidate should serve as their representative, and in this respect differs significantly from the purer form of proportional representation, the system of the single-transferable vote, in which every successful candidate has a unanimous constituency.

Geographical representation usually relies on single-member plurality systems, and tends to produce only two or three parties. The parties represent broader interests than do the parties in modified proportional systems, but if party discipline is weak, as in the USA, the geographical basis encourages the representation of particular regional and other interests. No single type representational system has been shown to be superior in all or most circumstances, and the choice generally involves a trade-off: the wider the range of interests represented in the legislature, the less the likelihood that any particular interest will be fully represented in legislation actually adopted. The most useful scholarship on these and related questions has been comparative (Klingemann et al. 1994, Converse and Pierce 1986).

In systems such as the USA, where many of the districts that define constituencies do not follow any natural geographical boundaries, the issues of how the districts are drawn and whether they produce fair representation have stimulated important normative and empirical research. Political scientists have examined the partisan and incumbency biases in various redistricting schemes. Some of the most interesting theoretical work in this area combines empirical and normative approaches. One study has developed more analytically robust concepts of 'fair representation' in electoral systems (Beitz 1989), and another has revived the theory of descriptive representation for 'marginalized groups' (Williams 1998).

4. Legislators

Much of the research on this element of representation has dealt with the tension between individual and collective behavior in the legislature. In the USA a phenomenon that has received considerable attention is what is sometimes called legislative individualism, the tendency of legislators to act in ways that promote their own re-election without regard to the effect on collective product of the legislature. Such individualism has been shown to produce a pattern of 'particularized benefits' (dams in certain regions, tariffs for local industries, and military bases in the district) and the neglect of more general public benefits (such as education, health, and the environment. The explanations for these outcomes vary, but they are generally structural, locating the problem in the incentives of the system, rather than in the character of legislators. They generally assume that, given the structure of the political system, individually rational behavior by legislators produces collectively suboptimal outcomes in the legislature.

If all legislators are pure individualists seeking only particularized benefits for their own constituents, how (under conditions of majority rule and weak parties) can the legislature adopt any legislation at all, even the pork barrel kind? This puzzle has been extensively studied by formal theorists, who propose various solutions for achieving policy equilibria in legislative institutions. One of the most plausible relies on a committee system to set an agenda and limit the scope of conflict. But shared interests and cross-cutting cleavages would still be necessary in any actual system. These are more likely to exist or to arise in legislatures that encourage or at least permit representatives to change their positions through interaction with their colleagues. This is one of the advantages of deliberative theories of legislative representation, which allow for changes in the preferences of both representatives and the represented over time (Gutmann and Thompson 1996, Chap. 4).

See also: Accountability: Political; Apportionment: Political; Campaigning: Political; Cleavages: Political; Consociationalism; Deliberation: Political Aspects; Electoral Systems; Feminist Political Theory and Political Science; Interest Groups; Legislatures: United States; Participation: Political; Political Parties; Proportional Representation.

Bibliography

Beitz C R 1989 *Political Equality*. Princeton University Press, Princeton, NJ

Cain B, Ferejohn J, Fiorina M 1987 *The Personal Vote: Constituency Service and Electoral Independence*. Harvard University Press, Cambridge, MA

Converse P E, Pierce R 1986 *Political Representation in France*. Belknap Press of Harvard University Press, Cambridge, MA

Gutmann A, Thompson D 1996 *Democracy and Disagreement*. Harvard University Press, Cambridge, MA

Kingdon J W 1981 *Congressmen's Voting Decisions*, 2nd edn. Harper and Row, New York

Klingemann H-D, Hofferbert R I, Budge I 1994 *Parties, Policies and Democracy*. Westview Press, Boulder, CO

Larsen J A O 1966 *Representative Government in Greek and Roman History*. University of California Press, Berkeley, CA

Manin B 1997 *The Principles of Representative Government*. Cambridge University Press, Cambridge, UK

Miller W E, Stokes D 1963 Constituency influence in Congress. *American Political Science Review* **57**

Pitkin H F 1967 *The Concept of Representation*. University of California Press, Berkeley, CA

Stimson J A, Mackuen M B, Erikson R S 1995 Dynamic representation. *American Political Science Review* **89**: 543–65

Thompson D F 1976 *John Stuart Mill and Representative Government*. Princeton University Press, Princeton, NJ

Uslaner E M 1986 Legislative behavior: the study of representation. *Annual Review of Political Science* **1**: 205–37

Williams M S 1998 *Voice Trust and Memory: Marginalized Groups and the Failings of Liberal Representation*. Princeton University Press, Princeton, NJ

Wood G S 1972 *The Creation of the American Republic 1776–1787*. Norton, New York

D. F. Thompson

Political Science: Overview

Political Science is an academic discipline that has been entrenched in university curriculums for about a hundred years. It takes as its subject matter government and politics very broadly defined, meaning the study (in whatever setting) of outcomes deemed important by participants or observers. This means that there is politics, and therefore in principle political science, concerned with, *inter alia*, organized crime, organized sports, organized religion (Reese), business firms (Bauer et al. 1963), family life, and markets (Dahl and Lindblom 1953). Most commonly political scientists study social patterns and practices associated with the apparatus of the state, identified by Max Weber as an institutional complex located within a specific territory that monopolizes the legitimate use of force.

One formula describing the scope of the field is contained in the full title of one of Harold Lasswell's path-breaking books: *Politics: Who Gets What When How*. This definition, by focusing on what people 'get,' leaves uncertain the demarcation between political science and economics, and may lead to the impermissible inference that answers to the question 'who benefits' are by definition the same as answers to the question 'who governs.' Perhaps a more useful des-

cription of the scope of the discipline is David Easton's formulation of politics as 'the authoritative allocation of values' in society, entailing the study of authority and its bases, allocation and its modes, and the shaping and sharing of values. Political scientists adopt many different approaches to the study of politics, and so it may be helpful to contrast political science, which is a discipline whose practitioners are united by very little other than a focus on a population of dependent variables or effects of human intention or organization or action with, e.g., sociology or psychology, which are unified around a preferred set of independent variables or causes of varied social and personal outcomes. This is not to say that political scientists do not write about causes as well as effects; rather, when they do so, they borrow freely from other disciplines. There are no psychological or sociological mechanisms that are the monopoly of political science. What is distinctive to political science is what it is about: politics. Political science draws on and elaborates ideas from other disciplines in seeking to account for, or explain, or understand political life. This helps us understand why no clear-cut mainstream tradition exists in political science, rather a number of different traditions of inquiry, organized in part around the differing scale and character of different topics within the field, of which I shall mention several.

Political scientists draw on several branches of psychology in attempting to understand attitude formation and the preferences of citizens and individual voters, as well as the choices of leaders. The development of sophisticated survey instruments for the purpose of understanding individual attitudes and opinions has taken place almost wholly outside political science, but political scientists have been quite successful in adapting these tools to the tasks of understanding how individuals think about political outcomes, make voting choices, and form views of their political world (Converse 1987, Zaller 1992). Social psychology and the social psychological branch of cognitive science have also framed and explored principles of human behavior (some of which are called 'heuristics') that shape political scientists' explanations of political outcomes (Tversky and Kahneman 1974, Abelson et al. 1968, Simon 1957). Finally, the multifaceted Freudian and neo-Freudian traditions of inquiry have had an impact on the study of the behavior both of masses and elites (Greenstein 1969, George and George 1956).

A second focus of research for political science has been the study of social organizations, groups, and classes, in which mainstream sociology has provided intellectual inspiration. In some cases, despite the restricted domain of political scientists' interests, political scientists have entered into protracted and serious dialogue with sociological students of such topics as organizational design and reform, community power, and social capital (Wilensky 1967, Polsby 1963, Putnam et al. 1993).

On a still broader scale, political scientists study institutions as large and consequential as nation-states and their constituent elements: executives, legislatures, courts, parties, elections, interest groups, drawing on a great many specialized studies by historians, legal scholars, and journalists as well as other social scientists (Greenstein and Polsby 1975).

Relations among nation-states is a fourth domain in which political scientists work, exploring questions like the extent to which inter-state cooperation on small or less contentious matters such as the regulation of financial markets or trade increases the probability of harmony on larger matters touching national security interests and forestalls hostility (Haas 1968, Deutsch et al. 1957).

In the study of international affairs (as is also the case in other domains) we may note a significant influence on the agendas of political scientists, namely the preoccupations of politicians who are themselves the objects of study. Just as diplomats historically have sought to find a balance of power among nation-states, students of international affairs have concerned themselves with issues of international political stability, the formation of security communities, and the causes of war (Waltz 1959, Van Evera 1999). This illustrates the fact that the intellectual agendas of political scientists are made in part from political events and not only from problems posed by the ongoing literature of political science.

A fifth focus for political scientists has been the history of ideas about social and political organization, sometimes studied through the elucidation of great books in the heritage primarily of Western Civilization, sometimes through the examination of prescriptions about social organization and the textual or empirical study of prerequisites of their fulfilment (Skinner 1978, Berlin 1969, Wolin 1960, Dahl 1956). Thus a political scientist may be a historian of ideas, or a philosopher, or draw inspiration from those disciplines.

As a professional discipline organized into the curriculums of the world's universities, Political Science is, like the other social studies, only about a century old. Systematic political study, to be sure, is as old as the Greek philosophers of the fourth century BC, notably Plato and Aristotle. The oldest University chair in the subject is at the University of Uppsala which in 1622 established the Johan Skytte chair of Eloquence and Government, helpfully mentioning one of the classic disciplines (rhetoric; moral philosophy and law are two others) from which late nineteenth century political science emerged as a separate subject offered on a regular basis in the newly professionalizing universities of Europe and, especially, North America.

The editors of a new British journal of politics and international relations estimate that today about 85 percent of all the political science being studied worldwide is located in the USA. Where the discipline

flourishes elsewhere it is more likely to be limited to studies in public law or philosophy and to lack the autonomous ability to sponsor the discovery and pursuit of its own program of study as has been possible for political scientists in the US, at least since the formation of the American Political Science Association in 1903. The Association now has about 15000 members, the largest aggregation of political scientists in the world. It sponsors an annual meeting and two journals, one (*American Political Science Review*) scholarly and one (*Political Science*) professional (Somit and Tanenhaus 1964).

Like other social sciences, the pursuit of disinterested inquiry into political life requires a political atmosphere that tolerates disinterested inquiry. Without this atmosphere political science is suffocated. It can be no accident that contemporary political science is primarily to be found in those nations most advanced in their incorporation of the values of the Enlightenment, which underwrite academic freedoms.

Political science as knowledge can be distinguished from political lore or political news primarily by reference to the social organization that produces, conserves, and criticizes it. Political news is the property of a nation's political class and it is spread by journalists or chains of gossip. Political lore consists of precepts codifying and sustaining the practices of political practitioners. Political science may incorporate content from either of these bodies of knowledge but the social organization that certifies knowledge as political science at any given time and sponsors criticism and new exploration in the field is the congeries of academic departments variously called political science, political studies, politics, or government in colleges and universities worldwide. These departments promote inquiry and conserve knowledge by educating undergraduates, training graduate students, producing texts, announcing discoveries, criticizing work, and maintaining a community of mutually attentive scholars (or, more accurately, a set of sub-communities) who participate in a loosely articulated status system, recognize common intellectual ancestors, promote fads, award prizes, and otherwise establish and enforce social boundaries.

In the USA alone, some 125 departments grant the Ph.D. in political science, and roughly five times that number of colleges offer undergraduate majors in the topic. Substantial numbers of universities in Canada, Europe, India, Japan, and Australia do likewise, but graduate studies in political science in other places, where they exist, are more likely to be subsumed under law, history, or philosophy as, for example is the case at Cambridge University in the UK.

The incorporation of the term 'science' in the disciplinary label political science has from time to time provoked controversy on the ground that no inquiry into the behavior of sentient beings can possibly approximate the predictive capacities of physics or chemistry. This charge is made against all the human sciences but because of the discipline's eighteenth century—largely pre-scientific—name it is sometimes assumed that in the case of political science ambitious claims for prediction or control of human behavior are being advanced. There are several standard answers: (a) That the term 'science' in historical context makes reference not to some social analogue to atomic physics but simply to 'knowledge'; (b) that the application of political knowledge was thought to be desirable in the creation of new constitutions, notably the US Constitution of 1787; (c) that political knowledge takes many forms requiring only the recognition of patterns of social behavior. This facilitates the codification for example, of comparative information about analogous activities (voting, budget-making, social/political protest) in different political systems; (d) that the goals of prediction and control as a means of understanding social behavior are best understood as applying not to historical events but rather to variables abstracted from events which may lend themselves to historical interpretation; (e) and that social patterns and other manifestations of human behavior can be more competently observed and explained by reference to standards and canons of inquiry (especially recourse to evidence and codification of findings into propositions of some generality) associated with more scientifically advanced disciplines.

Political scientists themselves vary in the extent to which they pay explicit attention to rules and methods of scientific inquiry. As the twentieth century advanced, more and more did so, creating among other things, sub-fields devoted to methodology—mainly adapting to political materials the methods of survey research and econometrics—and to specialties sometimes called rational choice or formal or positive theory—adaptation of the economist's propensity to describe simplified versions of human activity by means of mathematical models. (Solow 1997).

Earlier, with growing influence from the 1950's onward, a so-called 'behavioral revolution' took place in which political scientists in many sub-fields of the discipline sought to introduce into their work the rigorous intellectual habits of the other social sciences with their concern for evidence and for structuring their inquiries so as to alert investigators to unexpected or generally applicable outcomes—e.g. by referring to null hypotheses, or by adhering to falsificationist standards of proof. This has facilitated storage and retrieval of information and led to the cumulation of findings in several branches of the discipline. While these practices have spread in the last half-century, they have not wholly supplanted more traditional modes of inquiry such as the close reading of classic texts or the clarification—or at least discussion—of key concepts, such as 'democracy' or 'power' or 'coalition' or 'party.' (Lasswell and Kaplan 1950, Sartori 1984).

See also: Aristotle (384–322 BC); Communitarianism: Political Theory; Governments; Political Anthropology; Political Geography; Political Sociology; State and Society; State, History of

Bibliography

Abelson R P et al. (ed.) 1968 *Theories of Cognitive Consistency: A Sourcebook*. Rand McNally, Chicago

Bauer R A, Pool I de Sola, Dexter L A 1963 *American Business and Public Policy*. Atherton, New York

Berlin I 1969 *Four Essays on Liberty*. Oxford University Press, London

British Journal of Politics and International Relations 1999 "Editorial: Studying British Politics," **1**(1): 2

Converse J M 1987 *Survey Research in the United States: Roots and Emergence 1890–1960*. University of California Press, Berkeley, CA

Dahl R A 1956 *A Preface to Democratic Theory*. University of Chicago Press, Chicago

Dahl R A, Lindblom C E 1953 *Politics Economics and Welfare: Planning and Politico-Economic Systems Resolved Into Basic Social Processes*. Harper, New York

Deutsch K W et al. 1957 *Political Community and the North Atlantic Area: International Organization in the Light of Historical Experience*. Princeton University Press, Princeton

Easton D 1953 *The Political System: An Inquiry into the State of Political Science*. 1st Edn. Knopf, New York

George A L, George J L 1956 *Woodrow Wilson and Colonel House: A Personality Study*. J. Day Co., New York

Greenstein F I 1969 *Personality and Politics: Problems of Evidence, Inference and Conceptualization*. Markham, Chicago

Greenstein F I, Polsby N W 1975 *Handbook of Political Science* (Volume I). Addison-Wesley Publishing Company, Reading, MA

Haas E B 1968 *The Uniting of Europe: Political, Social, and Economic Forces, 1950–1957*. Stanford University Press, Stanford, CA

Lasswell H D, Kaplan A 1950 *Power and Society: A Framework for Political Inquiry*. Yale University Press, New Haven

Lasswell H D 1936 *Politics; Who Gets What When How*. Whittlesey House, New York

Polsby N W 1963, 1980 *Community Power and Political Theory*. Yale University Press, New Haven, CT

Putnam R D, Leonardi R, Nanetti R Y 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, Jamaica

Reese T J 1989 *Archbishop: Inside the Power Structure of the American Catholic Church*. 1st Edn. Harper and Row, San Francisco

Reese T J 1992 *A Flock of Shepherds: The National Conference of Catholic Bishops*. Sheed and Ward, Kansas City, MO

Reese T J 1996 *Inside the Vatican: The Politics and Organization of the Catholic Church*. Harvard University Press, Cambridge, MA

Sartori G (ed.) 1984 *Social Science Concepts: A Systematic Analysis*. Sage, Beverly Hills, CA

Simon H A 1957 *Models of Man: Social and Rational*. Wiley, New York

Skinner Q 1978 *The Foundations of Modern Political Thought*. Cambridge University Press, Cambridge, UK

Solow R M 1997 How Did Economics Get That Way and What Way Did It Get? *Daedalus* **126**: 39–58

Somit A, Tanenhaus J 1964 *American Political Science: A Profile of A Discipline*. Atherton Press, New York

Tversky A, Kahneman D 1974 Judgement Under Uncertainty: Heuristics and Biases. *Science* **85**: 1124–31

Van Evera S 1999 *Causes of War: Power and the Roots of Conflict*. Cornell University Press, Ithaca

Waltz K 1959 *Man, the State, and War: A Theoretical Analysis*. Columbia University Press, New York

Weber M 1946 *From Max Weber: Essays in Sociology*. Oxford University Press, New York

Wilensky H 1967 *Organizational Intelligence: Knowledge and Policy in Government and Industry*. Basic Books, New York

Wolin S S 1960 *Politics and Vision: Continuity and Innovation in Western Political Thought*. Little, Brown, Boston, MA

Zaller J 1992 *The Nature and Origins of Mass Opinion*. Cambridge University Press, Cambridge

N. W. Polsby

Political Sociology

Political sociology is a border field between political science and sociology, and the term encompasses the overlap between these two neighboring disciplines. It denotes the analysis of the inter-relationship between the social and the political, social structures and political institutions, between the society and the state. There is no stable consensus of what counts as political sociology in contrast to sociology and political science proper. There have been great variations over time in the popularity of political sociology and in the tendencies to emphasize it as a genuine field of its own. Nevertheless a common element is that political sociology is related to the distinction between the social and the political.

1. Historical Roots of Political Sociology

The historical development of political sociology can be roughly outlined by emphasizing three periods of special importance.

(a) The birth of political sociology can be dated to the mid-nineteenth century, when the founding fathers of social science began to conceive of the social and political, the society and the state as separate entities, and to analyze their relationships (Runciman 1963, pp. 22–42, Pizzorno 1971, pp. 7–18). This type of inquiry had its intellectual roots in the writings of earlier philosophers about the state and the social life, however, as a broad and joint interest of an academic community, political sociology emerged in the schol-

arship of the nineteenth-century fathers of social science.

(b) The post-World War II studies of the conditions for and the citizen activities in political democracies with a special focus on parliamentary elections conceived them as expressions of the political competition between different social groups. The joint focus on the prerequisites of democracy and voting patterns proliferated during the 1950s and 1960s, receiving institutionalized expression in the work and activities of the Committee of Political Sociology, founded in 1959 with Seymour Martin Lipset as chairman and Stein Rokkan as secretary and functioning under the auspices of both the ISA (International Sociological Association) and the IPSA (International Political Science Association) (Rokkan 1970, pp. 1–20).

(c) The late twentieth-century studies of nation and state building placed a special emphasis on great political transformations and breakdowns of regimes. This type of inquiry received a strong impetus from the renaissance of Marxist thinking in the 1960s and 1970s, but it was not restricted to social scientists with a leftist orientation (Linz and Stepan 1978, Skocpol 1979).

1.1 The Establishment of a Distinction between Society and State

In a great number of their texts, most of the pathbreaking social science theoreticians of the second half of the nineteenth century and the early twentieth century analyzed the relationships between the social and the political. One of the first thinkers to delineate an empirical political sociology was Karl Marx in his systematic insistence on seeing politics as an expression of conflicts between different social classes. His theorizing had a factual historical background in the emergence of the bourgeoisie and the industrial working class as distinct and dominant social groups (Pizzorno 1971, pp. 7–9).

Max Weber on his part wrote many of his works as rejoinders to Marx, but even apart from his counterclaims against Marx, there is in Weber's writings continuous analyses of the social institutions and the types of authority existing in modern society (Runciman 1963, pp. 53–7).

Among the great founders of sociology Émile Durkheim is not normally thought of as a scholar of politics, but his preoccupation with order, social integration, and social solidarity had considerable implications for the analysis of the relationship between society and politics.

Vilfredo Pareto was a pioneer both in econometrics and in the sociological study of elites. His ability to relate the circulation of elites to large-scale social and political changes makes him one of the founding scholars of political sociology.

Karl Marx, Max Weber, Émile Durkheim, and Vilfredo Pareto were not primarily political sociologists, but instead more generally the great founding fathers of social science theory building. There are, however, some nineteenth-century scholars who can be labeled as specialists in political sociology. They are not unknown as scholars in history and in the study of politics, but their fame rests basically on the fact that they were important creators of the tradition of political sociology with its focus on the relationship between society and politics. Among them Lorenz von Stein, Alexis de Tocqueville, Mosei Ostrogorski, and Robert Michels deserve special mention.

Lorenz von Stein (1815–90) was a forerunner of Marxism in the sense that he explicitly emphasized how material and social conditions mould political development. His *Geschichte der sozialen Bewegung in Frankreich* (1850) contains not only a history of the French Revolution but also analyses of the importance of the property distributions and educational opportunities for political activities, and of the phenomena of class struggles and sudden outbursts of social change. It has been emphasized how some of Karl Marx's ideas were influenced by the writings of von Stein.

Alexis de Tocqueville (1805–1859) descended from an ancient and distinguished noble French family. His visit to the United States in 1831–2, which was of decisive importance for his scholarship, resulted in a study of American democracy published in two volumes *De la démocratie en Amérique* (1835, 1840). Although the work made him well known, his great scholarly fame was established only after the Second World War and to a considerable degree by the then growing political sociology. Especially the English translations of his works, *Democracy in America* (1945) and his study of the French revolution, *The Old Regime and the Revolution* (1955; originally published in French in 1856), became standard works for political sociologists. His great contribution was that he specified the conditions for pluralism and democratic systems. He argued that social cleavages and competition between social groups and strata are necessary conditions for consensus in industrially developed societies. He emphasized voluntary associations and active, partially autonomous local communities as crucial in creating and sustaining a plural democracy. In the texts of political sociology, de Tocqueville has often been contrasted to Marx. While Karl Marx envisaged conflicts and social cleavages as one-sidedly destructive, de Tocqeuville emphasized their importance for obtaining consensus in well-functioning democracies (Lipset 1960, pp. 4–9).

Although Mosei Ostrogorski (1854–1919) was born in Tsarist Russia and died in the Soviet Union, during important parts of his life he worked in Paris and London. Early in his career he published studies of the suffrage demands of the women's movement, but his main work was a critique of party democracy in

Democracy and the Organization of Political Parties (1902). Although it was an empirical study of the political party organizations in Britain and the United States, its main theme was that the mechanical and massive organization of political life into bureaucratized political parties, aimed at manipulating individuals and masses, was harmful for real democracy.

As was Ostrogorski, Robert Michels (1876–1936) was a critic of the emerging party bureacracies in modern states. Through his education and activities Michels had background liaisons with Germany, France, and Italy. Politically he started as a socialist with a syndicalist orientation but developed later profascist opinions. He took Italian citizenship in 1913, and although he was active as a professor and scholar in many countries he regarded Italy as his home country. His major work *Zur Soziologie des Parteiwesens in der modernen Demokratie* (1911) is considered a classic in political sociology. It treats many of its central themes, including the theory of elites, political movements on the left, processes of bureacratization, and the gap between democratic theory and practice. His central thesis, known under the name of *the iron law of oligarchy*, is that organization necessarily leads to oligarchy, or as Michels summarized it, 'it is organization which gives birth to the dominion of the elected over the electors, of the mandataries over the mandators, of the delegates over the delegators. Who says organization, says oligarchy.' Despite Michel's pessimism and one-sidedness, his work and the iron law of oligarchy has inspired a great number of studies of political and other civic organizations.

1.2 The Institutionalization of Political Sociology after World War II

The period between World War I and II, and especially the 1930s, were poor as regards studies in political sociology. The European political scene was turbulent, little support was given for studies in the social sciences, and in the 1930s the Nazi, Fascist, and Communist regimes ended most social science research in a number of countries. During the period between the world wars there were nevertheless some pathbreaking studies of voting and electoral behavior. Already before World War I André Siegfried had introduced the analysis of geographical variations in electoral behavior by studying voting patterns in French constituencies. A summary of his results was published in *Tableau des partis politiques en France* (1930). His central conclusion was that electoral behavior was steered by stable, deep-seated, regional traditions. After Siegfried the concept of *political traditions* became central in descriptions of voting patterns. Another innovator was the German-American Rudolf Heberle, who had started his career with an analysis of the tendencies to radicalization of the Swedish labor movement. His leading work *From Democracy to Nazism* (1945) about the appeal of Nazism to the rural population in the German province of Schleswig Holstein was published when he already had settled as professor in Baton Rouge, Louisiana, USA.

Large-scale studies of voting behavior, however, were not undertaken until after World War II, and this subject became a leading problem area of political sociology during the first postwar decades. It had a background in the postwar situation, as great hopes were attached to the possibility of building a better world with the aid of social science and research. Studies of the prerequisites of democracy became important, and parliamentary elections were seen as the major mechanism for establishing and securing democratic rule. Studies of the social background of political behavior represented an almost empty space between political science and sociology. Political scientists had been mainly preoccupied with constitutional problems and methods of steering the states, whereas sociologists had focused their interest on studies of social structure and behavior in social groups. Political sociology moved into a vacuum, and it became soon both an important field of research and an activity of interest to the general public. Although the study of electoral behavior later became mainly a field within political science, the first pathbreaking studies were made by sociologists. Theoretically fruitful conceptions and hypotheses, such as the tendency of cross-pressures to lead to political passivity, were developed at the Bureau of Applied Social Research at Columbia University under the inspiration and leadership of Paul Lazarsfeld (Lazarsfeld et al. 1944).

It was, however, through the founding of the Committee of Political Sociology at the International Sociological Association's Fourth World Congress in Stresa in September 1959 that political sociology became an established and institutionalized field of research. It led to a new kind of cooperation between political scientists and sociologists, and the committee also became a research committee of the International Political Science Association. The Committee's first chairman, Seymour Martin Lipset, had already published several innovative books in political sociology, and its first secretary, Stein Rokkan, had started to develop means for comparative, cross-national research. The year 1959 represented in many senses an important mark in the development of political sociology. In 1959 Lipset obtained the copyright for his *Political Man. The Social Bases of Politics* (1960), which became the leading text of the period of birth of an institutionalized political sociology.

The references and data sources in Lipset's book reflect the fact that at the end of the 1950s political sociology had become a small industry with young researchers all over the industrially developed world digging up data about the social bases of politics. The work of the Committee led to new developments. This is evident by a comparison of Lipset's *Political Man*

and the first comprehensive book produced by cooperation within the Committee, namely Lipset's and Rokkan's *Party Systems and Voter Alignments: Cross-National Perspectives* (1967). It contains a lengthy and theoretical introduction by Lipset and Rokkan, and chapters about individual countries, including the United States, Canada, Great Britain, Australia, New Zealand, France, Italy, Spain, West Germany, Finland, Norway, Japan, Brazil, and some West African states. Both the theoretical introduction, with its application of Talcott Parson's A-G-I-L scheme, and most of the country chapters, reflect attempts to combine a structural and institutional approach and to study both the social bases of politics and the effects of politics on social structures. One of the members of the Committee, Giovanni Sartori, described the change in orientation in a paper entitled *From the Sociology of Politics to Political Sociology* (1969). Political sociology had ceased to be a subfield of sociology and had become a genuine border field of its own.

Studies of voting patterns were by no means the only ones conducted and presented at the conferences of the Committee and by its members. The first part of Lipset's book on the *Political Man* consists of an analysis of the prerequisites of the democratic order. At the turn of 1950s and 1960s several studies analyzed the conditions of democracy. In addition to the formal criteria of democracy such as popular representation and individual freedoms, political sociologists were keen to stress some factual social conditions as prerequisites of a democratic order. Of particular importance were *effectiveness*, the government's capacity to implement policy, *efficacy*, peoples' subjective feelings of being important for the politicians, and *legitimacy*, beliefs that the existing political institutions are appropriate and lawful. Some of the studies were particularly focused on breakdowns of democracy and the rise and appeal of Fascism and Communism, such as Raymond Aron's *l'Opium des intellectuels* (1955) and William Kornhauser's *The Politics of Mass Society* (1959). Democracy was studied from many angles, such as in the study done by Lipset and his associates on the conditions for democracy in labor unions, Robert Dahl's studies of democracy on the local community level, and Ralf Dahrendorf's analyses of the conditions of conflict regulation in industrial society.

Even if studies of and data about voting and electoral behavior formed a central point of departure for the political sociologists in the 1950s and 1960s, their aim was more general than to give information about how people behave during elections. The basic interest was in studying conditions for democracy, breakdowns of regimes, nation and state building, and master processes of societal and political change. From the 1970s onward the electoral studies became methodologically more sophisticated at the same time as they increasingly became a preoccupation of specialists of political issues and mass media effects within the field of political science. Initially such studies had a center at the University of Michigan under the leadership of Angus Campbell.

1.3 The Re-entry of the State and its Transformations

The disappearance of electoral studies from the central agenda of political sociology was not an isolated phenomenon. From the beginning of the 1970s many of the original topics of political sociology lost much of their popularity. One of the reasons was that the student rebellions and leftist radicalism in the 1960s and 1970s had severely criticized the agenda of the prevailing political sociology. Another reason was that some fields of research had become routine and were incorporated into either sociology or political science.

Despite a decline of the distinctiveness of political sociology, some research themes have remained central and typical for political sociology. The problem sphere of political sociology comprises increasingly master patterns of societal change, i.e. simultaneous changes and pressures to change in the political order, the social system, and the socioeconomic composition of societies. Such studies had been initiated by members of the Committee of Political Sociology already during its preoccupation with voting studies. Thus, there were studies of the formation and fall of empires by S.N. Eisenstadt (1963), studies of the historical formation of European centers and peripheries by Stein Rokkan (Rokkan and Urwin 1983), and studies of the breakdown of democratic regimes by Juan Linz (Linz and Stepan 1978). Also many scholars who never had been associated with political sociology as such wrote influential texts. An outstanding representative in this category is Barrington Moore's often-quoted major work *The Social Origins of Dictatorship and Democracy* (1966).

In the 1980s and 1990s interest in master patterns of change, and in state and nation building proliferated. The state was reintroduced as a truly central concern. Major representatives of the new orientation were especially Theda Skocpol (1979), who penetratingly emphasized the state as the originator of social outcomes, and Charles Tilly (1975), who rather than seeking one particular important unit, has stressed how national developments were formed by a complicated web of warfare, fiscal policy, bureacratization, and state-making. Göran Therborn's (1995) studies of modern developments in the European states have also emphasized the multidimensional nature of national development and warned against beliefs in irreversible tendencies. The increased interest in state and nation building also went beyond a focus on national states by studies of worldwide tendencies of development, as seen in the work of Immanuel

Wallerstein (1974) about the origins of the European world economy in the sixteenth century.

The developments from the 1970s onward made it increasingly difficult to define the borders of political sociology. There was for instance a recrudescence of studies of ethnicity, ethnic groups, and the role of ethnic identities in the formation of nations. The studies of ethnicity are sometimes counted as part of political sociology, sometimes as a field of its own. Yet, the borderline problem is fairly clear. It is a question of where the main emphasis is. When the aim is to describe the role of ethnicity and ethnic identities in the formation of nations and political development on the macro level, as in Stein Rokkan's (Rokkan and Urwin 1983) studies of the formation of European centers and peripheries, they are of great relevance for political sociology.

It is much more difficult to evaluate the character and place of Marxist-oriented and radically critical studies which multiplied from the second half of the 1960s onward. Earlier this type of scholarship had been mainly carried out by specific individuals. In studies of the power elite, C. Wright Mills definitely treated the relationship between the social and political, but he was not generally counted among the political sociologists (Kimmerling 1996). In the 1960s there was a rise—indeed a renaissance—of whole schools of critical sociologists such as the Frankfurt School of dialectical sociology and the dependence theorists who emphasized the hopeless fate of the underdeveloped countries in a world dominated by international capitalism. It is an open question whether these and kindred groups should be included in political sociology. Usually they were not counted as such. It is again a question of how one conceives of their main emphasis. If the main intent is to present social critique, the studies are usually not counted as political sociology, however, if the studies present hypotheses and concepts useful in most studies of societal and political development, they are regarded as political sociology. In the 1980s the radical thrust in the social sciences came to an end. The studies in political sociology began to combine elements from studies of very different political shades. It is characteristic of most of the studies of nation and state building mentioned above that they had assimilated elements from both Marxist and liberal thinking and scholarship.

2. *The Consciousness-creating Functions of Political Sociology*

Whatever solution is reached as regards the labeling of radical social critiques, it is obvious that political sociology is always directed towards expanding and creating awareness of macro changes and fundamental problems of societies. Nevertheless, it is seen as important that political sociology should not basically serve political aims. There are strong claims that political sociology should be empirically oriented, aim at theory building, employ a comparative approach, and systematically consider historical factors. The combination of the two claims, the consciousness-expanding function on one hand, and the quest for systematic comparative, historically based, and theoretically oriented studies on the other, has been astutely analyzed by Birgitta Nedelmann (1997) in a critique of the failure of German postwar political sociology to account for the period of National Socialism in German history.

Even if a successful delineation of political sociology never emerges, some main problem areas can be listed with the help of the two claims and objectives listed above. Political sociology studies (a) master patterns of simultaneous social, cultural, and political change in society, (b) dominant sociopolitical and emerging new cleavages, (c) the formation of states and nations as well as breakdowns of social and political orders, (d) prerequisites of and threats to democratic regimes, (e) problematic features of modern states such as bureaucratization, centralization, tendencies to corporatism, and (f) the political consequences of globalization such as the formation of continent-based political units and institutions.

See also: Marx, Karl (1818–89); Pareto, Vilfredo (1848–1923); Political Parties, History of; Political Science: Overview; State and Society; State, Sociology of the; Tocqueville, Alexis de (1805–59); Voting: Class; Voting, Sociology of; Weber, Max (1864–1920)

Bibliography

Aron R 1955 *l'Opium des intellectuels*. Calman Levy, Paris

Eisenstadt S N 1963 *The Political System of Empires. The Rise and Fall of the Historical Bureaucratic Empires*. The Free Press, Glencoe, IL

Heberle R 1945 *From Democracy to Nazism: A Regional Case Study on Political Parties in Germany*. Louisana University Press, Baton Rouge, LA

Kimmerling B 1996 Changing meanings and boundaries of the 'political'. *Current Sociology* **44**: 152–76

Kornhauser W 1959 *The Politics of Mass Society*. The Free Press, Glencoe, IL

Lazarsfeld P F, Berelson B, Gaudet H 1944 *The People's Choice. How the Voter Makes up His Mind in a Presidential Campaign*. Columbia University Press, New York

Linz J J, Stepan A (eds.) 1978 *The Breakdown of Democratic Regimes*. Johns Hopkins University Press, Baltimore, MD

Lipset S M 1960 *Political Man. The Social Bases of Politics*. Doubleday, Garden City, New York

Lipset S M, Rokkan S (eds.) 1967 *Party Systems and Voter Alignments. Cross-National Perspectives*. The Free Press, Glencoe, IL

Michels R 1911 *Zur Soziologie des Parteiwesens in der modernen Demokratie. Untersuchungen über die oligarchischen Tendenzen des Gruppenlebens*. Klinkhardt, Leipzig, Germany

Moore B Jr 1966 *The Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon Press, Boston, MA
Nedelmann B 1997 Between national socialism and real socialism. Political sociology in the Federal Republic of Germany. *Current Sociology* **45**: 157–86
Ostrogorski M Y 1902 *Democracy and the Organization of Political Parties*. MacMillan, London
Pizzorno A (ed.) 1971 *Political Sociology. Selected Readings*. Penguin Books, Harmondsworth, UK
Rokkan S 1970 Introduction: international cooperation in political sociology. In: Allardt E, Rokkan S (eds.) *Mass Politics. Studies in Political Sociology*. The Free Press, New York
Rokkan S, Urwin D W 1983 *Economy Territory Identity. Politics of West European Peripheries*. Sage, London
Runciman W G 1963 *Social Science and Political Theory*. Cambridge University Press, Cambridge, UK
Sartori G 1969 From the sociology of politics to political sociology. *Government and Opposition* **4**: 195–214
Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge University Press, New York
Siegfried A 1930 *Tableau des partis politiques en France*. Grasset, Paris
von Stein L 1850 *Geschichte der sozialen Bewegung in Frankreich von 1789 bis auf unsere Tage*. Leipzig
Therborn G 1995 *European Modernity and Beyond. The Trajectory of European Societies 1945–2000*. Sage, London
Tilly C (ed.) 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ
Tocqueville A C de 1835, 1840 *De la démocratie en Amerique I–III*. Paris. [Published in English 1945 *Democracy in America*. Knopf, New York]
Tocqueville A C de 1955 *The Old Regime and the Revolution*. (trans. from French), Doubleday, New York
Wallerstein I 1974 *The Modern World System. Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York

E. Allardt

Political Thought, History of

Reflection about political phenomena in a wide sense of the term is as old as human community itself. It focuses on a triad of related problems that might be broadly delimited as follows. First, there is the problem of constituting a community bound together by a normative order, secondly that of establishing the legitimacy and the terms of enforcement of such an order, and, thirdly, that of finding criteria for a proper and just allocation of resources as well as of entitlements. Jointly these three sets of problems entail that political theory and the history of political theory, as distinct from, say, a theory of scientific validation or a theory of argumentation, will be characterized by an inherent duality. Thus it will not only, as these other types of theory, be an account of rules of contestation and signification. It will also have to involve an account of which such rules are and which are not upheld by a normativity that is, in the last instance, enforceable by violent means. Put in everyday language, political theory will in its core be a theory both about meaning and communication and a theory about power, enforcement, and, ultimately, violence. It is easy to see that different political theorists have come to emphasize different strands in this duality, but the duality itself is irreducible and constitutive of political thought.

1. *Political Thought in Historical Context*

Thought about such matters is found in clearly articulated form from the middle of the first millennium BCE onwards in all the major civilizations of Afro-Eurasia. Thus it is manifested in written texts in the sixth century by Confucius and two centuries later by Mencius, in China and the writings during the same period in India, as well as in the prophetical age of Jewish religion and during the era of classical philosophy in Greece. The emergence of imperial-like political orders across Afro-Eurasia in the last centuries of the first millennium BCE also gave rise to a range of texts about the nature of political order. This is equally valid for texts about the Roman and Byzantine empires in the Mediterranean region to texts highlighting features of rulership in Parthian and later also in Sassanian Persia, in Maurya India, and in the Han Empire of China.

The recording of political thought in the form of the writing of histories of such thought is however a much more recent phenomenon. At its focus it is the relationship of community-wide rules to forms of enforcement and an effort to arrive at an understanding of the historical context in which particular types of such relationships occur. The emergence of clearly articulated forms of historical thought is largely coterminous with the emergence of history and social science as forms of scholarly activity performed within reformed or newly created universities and higher education institutions from the early nineteenth century onwards. In order to understand the key features and characteristics of the writing of histories of political thought it is first of all necessary to bring out key conditions of political thought in the region in which such an early disciplinary condition first occurred, namely in Europe.

2. *Political Thought and Early Modernities*

In a broad sense political thought in early modern Europe was a way to understand the nature of ruling in human communities and to grasp the conditions of normative argumentation and enforcement in such communities. The need for new understandings of these types of phenomena was itself directly related to the gradual decline and eventual demise of the two

most prominent institutional forms of medieval Europe. Those were, first, the cultural world of Western Christendom, epitomized first by the all-embracing Christian Church and second by its secular counterpart, namely the pervasive, but elusive, political order of the Holy Roman Empire. By the twelfth and thirteenth centuries, however, city republican rule, particularly in the northern half of Italy, had emerged as a new form of political order. Originally it was often modeled on trading communities and was created out of a sense of the necessity of mutual support and protection in areas exposed to incursions and with little or no protection provided by the nominal Imperial rulers.

Similarly, incipient national monarchies were starting to assert their authority vis-à-vis both pope and emperor. In this secular process the French king played a pioneering role. In particular, he had earlier than most other kings and rulers demonstrated his ability to strengthen his own position and financial basis, e.g., by confiscating lands and goods of the order of the Templars. This example was to be followed by kings across Europe in their confiscating the lands and riches of monasteries and religious orders in centuries to come.

In this sense, the Reformation and the Counter-reformation meant that in Protestant and Catholic countries alike, national monarchies, normally of a composite nature (cf. Elliott 1992), were created or reformed. This occurred on the basis of the previous dominant ideals of church, empire, and city republics, but in a form that synthesized parts that had previously been separated. Furthermore, the new national and composite monarchies almost invariably involved efforts on the part of rulers to appropriate lands and resources that had previously been controlled by the church or religious orders. Thereby they were also able to secure a fiscal basis for the new forms of political ruling, if often at a horrendous price of protracted religious civil wars and cleansings.

Under these conditions the pre-eminent political questions became that of grasping the nature and legitimacy of a normatively binding communal order and to outline the range of legitimate means for upholding and enforcing that order. Many of the early modern classical political thinkers, from Machiavelli to Jean Bodin, provide answers to these very questions. They have, however, significantly less to say about the two other sets of questions discerned above and that were to become prominent in later political thought. These were the questions of the very constitution of a community of human beings and of the principles underlying decisions about the allocations of resources or the granting or withholding of entitlements.

All this was to change with the final breakdown of the old political and cultural order in Europe in the middle of the seventeenth century. In China, even in the embattled transition from the Ming to the Ching dynasty at roughly the same time, the general linguistic and political order was preserved. In Japan a long period of internal warfare and turmoil came to an end with the establishment of one of the most effective forms of absolutist rule, namely the Tokugawa shogunate. Across the Islamic world, the three parallel empires of the Ottomans, the Safavi, and the Mughals, constituted stable political entities. In Europe, however, a new system of states took form and was promulgated in the Westphalian peace treaty of 1648, based on an explicit recognition of religious diversity and monarchical pre-eminence. It was also based on the tacit recognition that an all-encompassing imperial polity was no longer even possible as a viable political ideal and that religious diversity was not to be overcome.

3. *Political Thought as the Constitution of a Political Community*

In this unfamiliar context, there was an urgency about finding ways of granting legitimacy and meaning to the new order that emerged in the wake of the Westphalian peace treaty and the provisional end of a century and a half of bloody civil wars in Europe. Thereby a sea change occurred in political thought, namely a deep-seated shift from a concern with the enforcement of normative political order to a concern with the nature and very emergence of political community in the first place.

The concern with the constitution of political community becomes the central one in political thought from Hobbes and Pufendorf to Locke all the way to Rousseau. Within this wide range of answers there is, however, a recurring and predominant concept, namely an inquiry into the conditions for the establishment of a social contract. The existence of such a social contract was normally seen also to entail an answer to the question of which types of enforcement of the normative order constitutive of the political community are legitimate. More specifically only those types of enforcement are sanctioned that, implicitly or explicitly, follow from the terms of the social contract itself.

Political theory from Hobbes onwards is centrally preoccupied with the nature of and the consequences of a transition from an imaginary state of nature to the contractually constituted political community. Ultimately all such forms of theorizing seem to depend on a limited set of assumptions about human beings and the nature of their joining together in societies. One such set of assumptions refer to the assignment of some kind of human rights, i.e., rights that accrue to human beings irrespective of their origins and other accidental traits and characteristics. Furthermore, there is a set of assumptions that refer to the thought-experiment of society being created through a process, normally described in terms of a contract, whereby

human beings agree about their association in political form.

Questions about enforcement, separate from an answer to the question of the constitution of a political order, based upon legitimately agreed upon contractual arrangements, were of subordinate and derivative importance relative to this overriding concern. The same is true of questions pertaining to the allocation of resources and the assignment of entitlements to the members and subjects of a political community. Thus such matters have to be dealt with and negotiated in accordance with the terms of those agreed upon in the original contractual arrangements.

Thus European political theory in this classical form emerged in a particular historical context both in European and world history. It also involved forms of theorizing that portrayed an imaginary contractual past as the point of constitution of polities. However, it was not in itself characterized by an effort to reflect on the history of political theory as a key component of developing such a theory. This form of reflexive consciousness was to become a key characteristic of yet another major transformation in intellectual and political history.

4. *The Formation of Modernity and the Transition from Political Philosophy to Social Science and History*

The late eighteenth and early nineteenth centuries are often described in terms of the coterminous occurrence of major transformations of both technological and economic practices and of political order. Often these transformations are captured with terms such as the industrial revolution and the political transformations inherent in the French and American revolutions. However, in this period there also occurs a major restructuring of ideas about nature and society and, more broadly, in the forms of knowledge and beliefs about the world. This intellectual transformation had deep repercussions for the emergence of new forms of political thought and reflection about the historical context of such forms of thought and for the decline of other forms of political thought.

In this period broad genres of thought about the world gradually gave way to more differentiated forms of knowledge. Slowly natural philosophy and political philosophy came to be replaced by a variety of different natural and social sciences. At the core of this transformation were epistemic shifts where ideas about biological and historical processes, about language and interpretation, about human beings and their capacities and bonds were created or reformulated. Some scholars have even spoken of this period as a second scientific revolution, where modern disciplines emerge in rudimentary form and where there is a shift in the institutional locations of scholarly and scientific activities toward new research laboratories and universities. In this process, society itself may even, to quote one scholar (Manent 1994, pp. 80–3), be said to have been discovered in its modern form, as distinct from the political order of the prerevolutionary period, whether it had the form of an absolutist monarchy or a constitutional one. Contrary to older forms of contractual reasoning, which postulated social contract as the hypothetical medium that had created political order, the new forms of social thought were distinctly historical and could also draw on anthropological experiences of human beings and societies in different parts of the world (Fox et al. 1995).

Broadly speaking, four distinct forms of thinking about human beings, societies, and political order emerged. These new forms of reasoning and the transition from generalized moral and political philosophy to social sciences had deep repercussions for political thought.

First, there was, with roots in the political economy and moral philosophy of the Scottish Enlightenment, a view of human beings as rational and deliberative individuals and of society as a compositional collective. This also made possible a new way of thinking about political community and its legitimacy. More precisely, it provided the possibility for thinking about political order as formed through the consent of individual human beings. However it was not, as in older contractual thinking, a consent that would have ushered in a distant and hypothetical social contract but rather a consent that could be assessed and rationally evaluated in terms of its reasonableness in its present and historical context.

Second, there emerged, often enough out of the experiences of the relative inability of the French Revolution to create stable institutions, a conception of human agency as formed and constrained by structural and systemic properties and of society as an organic totality. Such a conception would in a conservative version lead to theories of the necessity to restore an organic political order. In a radical version it could, however, also usher in demands for the overthrow of unnatural and illegitimate order. In a more broadly defined liberal or republican tradition, it might and would be compatible with ideas of the type that thinkers, often inspired by Durkheim, would propound toward the end of the nineteenth century in their combat against the ideological enemies of the republic. Thus it could support the idea that modern society was not an atomistic and distorted form of human community but could relate to an organic solidarity compatible with the life of free citizens in a free republic.

Third, and often in relation to historical and linguistic thought among representative of German philosophical idealism, a conceptualization of human beings took form in terms of the linguistic-interpretative nature of human agency and of society as an historically emergent totality. In this sense, political

community did not have to derive its legitimacy from dynastic tradition but from the historically grown experience of a community. Such communities in turn were constituted by human beings sharing a language and a history and where the community of shared life forms was seen as deeper, more genuine, and more real than the artificial boundaries of a minor princely state. It is easy to see that demands for national self-determination and for both popular rule and a nation-state easily flow from a conception along these lines.

Finally, in parallel to the dramatic development of forms of statistical analysis, there emerged a statistical-inductive way of representing the interactions of human beings and of society as a kind of systematized aggregate. Such views did not immediately entail a specific form of legitimate political community, but it would mean that any such community would have at its disposal infinitely more detailed knowledge of the polity and its members than even the most absolutist monarchy (cf. Brian 1994).

Thus the great intellectual transition entailed a deep shift in the history of political thought which affected both conceptions of the nature of a political community and the nature of legitimate and illegitimate enforcement. By virtue of this it also had profound implications for conceptions of the proper distribution of entitlements. Thus it was simply no longer possible, as observed most clearly perhaps by Tocqueville (1964, [1850]), to deny demands for popular participation with reference to tradition or dynastic legitimacy. The thought of a right to the political entitlements belonging to all members of a society, that had been manifested in the proclamations of the revolutionary upheavals, could never again, despite all efforts of the Vienna Congress and the Holy Alliance, be unthought. They were to pervade the political struggles first of European and American societies throughout the nineteenth and the early twentieth centuries and eventually in the world at large.

5. *Triumph and Crisis of Liberal Political Thought*

The end of World War I marked a sea change in world history. It spelt the end of European global pre-eminence. It also meant that the transition in Western Europe, normally from forms of constitutional monarchical rule, to parliamentary democratic rule was dramatically increased. In the Eastern half of Europe all of the four imperial political entities, the Russian, the Habsburg, the Wilhelminian, and the Ottoman, crumbled under the impact of external military defeat and internal revolutionary upheavals. However, the very triumph of liberal principles of representative democracy was coterminous with the emergence of new forms of political thought that transcended the traditions of the prewar world.

Leninist political thought, which before and during the war had been a radical version of classical Marxism, gradually came to take on the features of an official ideology of the new Soviet state. Precisely in its historical imaginary, Leninism differed from both revisionist and orthodox thought in a Marxist tradition, as epitomized by Bernstein and Kautsky respectively, of the Social Democratic mainstream of the Second International in the late nineteenth and early twentieth centuries. Thus Leninism cast its own claims to legitimacy and its right to use forms of violent enforcement, previously unthought of in late nineteenth century socialist thought, in terms of an historical account and of a projected visionary transition to a stateless society of communism.

Experiences of the mass violence of the war itself, as well as of the violent national and social struggles of the immediate postwar period, formed the environment out of which fascism emerged as a violent rejection of both Marxism and of the whole mainstream tradition of liberal political thought. Even if fascism had roots in prewar forms of radical conservative thought, it marked a form of transcendence of such traditions and indeed of the legacy of modernity itself. Thus even if fascism in power, and its radicalized form of National Socialism, embraced technological modernization, it represented a rejection of basic premises of political thought of modernity in the form it had taken since the late eighteenth century. In particular this is true of the basic presupposition of political thought of modernity of a principled equal right to participation in governance. It is also true of fascism's rejection of another basic presupposition of modernity, namely the legitimacy of a public sphere with free access for all and a principled freedom of every member of a polity to speak not only about but also to the rulers.

In a sense fascist political thought was a form of political thought that invoked history, or rather historical mythology. However, it was not in itself characterized by historical reflection. In Germany, the challenge of fascism to the political order of the Weimar republic was manifested in a range of debates, one of the most important of which concerned the nature of legal order and of the state itself. It was in this debate that the contesting positions of the radical and pro-Nazi decisionism of Carl Schmitt was confronted by the legal positivism of Hans Kelsen and the complex and sophisticated thought of Hermann Heller.

The outlines of a radical rethinking of the history of Western philosophical and political thought can be discerned in the elective affinity between fascism and some forms of philosophical thought, in particular that of Heidegger of the late 1920s and early 1930s. However, such a program was never elaborated or extended, nor was the decisive response and resistance to it on the part of a thinker such as Ernst Cassirer. Instead the military defeat of fascism in World War II

was seen to entail the definitive demise of features of political thought characteristic of fascism.

In the French context, the response to the challenges of fascism was to be found in several quarters. These would range, in political terms, from Marxism to radical Catholic political thought, and in scholarly terms, from sociology in the tradition of Durkheim to the profound renewal of historical studies of the Annales School. In the period after World War II, the French intellectual scene may have been the most important one in Europe in terms of a renewal of political thought. Thus it would comprise the nexus of contributions made by scholars elaborating a Marxian philosophical existentialist position, most importantly of course Sartre, but also thinkers such as Claude Lefort, who, starting out from such a position came to question the core of Leninist political thought. It would also comprise some of the most articulated liberal political thought as epitomized by Raymond Aron.

In the interwar years, the new challenges to liberal political thought had already given rise to a profound renewal of political thought in historical context. These were the years when new standard histories of political thought were being written. The most prominent example in the Anglo-Saxon world might have been a book that was to become a classic for generations, namely George Sabine's (1937) majestic volume *A History of Political Theory*. Refugees from continental Europe made key contributions to this debate. A classical example is the analysis of the meaning of democracy in historical context that was articulated by Joseph Schumpeter (1943) in *Capitalism, Socialism, and Democracy*. Most importantly, perhaps, Karl Popper transcended the bounds of philosophy of science and wrote one of the most quoted works in political thought during the twentieth century in his two-volume opus on *The Open Society and Its Enemies* (Popper 1966, [1945]). This was, of course, a work that was pre-eminently characterized by the links between historical analysis and political philosophy of immediate and contemporary relevance.

In parallel, leading representatives of Scandinavian political and legal science made some pathbreaking contributions both to an understanding of fascism and of the nature of democracy. In this respect the names of Herbert Tingsten and Alf Ross stand out, particularly with Tingsten's pioneering studies of fascism and his and Ross's contributions to a political theory of democracy.

One prominent characteristic of antiliberal political thought of the first half of the twentieth century was its emphasis on the right of the polity to resort to enforcement and violence against the enemies of the polity without being bound by any other constraints than those of pure convenience. Thus the constitutive boundary between members and enemies of the political community, between those chosen to fulfill the destiny of that community and those deemed to stand in the way and worthy of being removed from community, and ultimately from life, became the most crucial one. The political community itself was seen as constituted in terms of a historical account, whether of the struggle of a class for its liberation against oppression or of a people for its legitimate sphere of life and domination. Resistance to such types of political thought of a historicist nature formed the core of the argument of Popper's political philosophy. It also formed the focus for a range of contributions criticizing totalitarian political thought. Among the main contributors to this critique were thinkers as different as Jacob L. Talmon, Hannah Arendt, and Claude Lefort.

6. *Renewal and Critique in Political Thought*

The history of political thought in the first part of the latter half of the twentieth century is characterized by three parallel developments. First, political thought in this period cannot but reflect the deep tensions in a world with two superpowers in a state of cold war. At the same time large parts of the Third World were emerging out of the experiences of a colonial past and seeking a place in this new world of fundamental tensions. One manifestation of this historical context, was the absolute absence of publications in the tradition of liberal political thought in the countries under Soviet dominance, but also the relative absence of Marxist political thought at many universities in the English-speaking world. Another indirect consequence was a persistent interest in exploring the empirical basis of liberal democracy. In this sense, the contributions by scholars such as Robert Dahl and some of his colleagues at Yale university, not least Charles Lindblom, stand out as exemplary. Second, a classical mainstream tradition of historically orientated political philosophy made significant advances. Isaiah Berlin and Judith Shklar are pre-eminent representatives of this type of scholarship, and the works of Charles Taylor and Norberto Bobbio should also be included here. Third, however, philosophical and theoretical contributions were made that were to reshape the history of late twentieth century political thought. I shall highlight some of these contributions and their consequences in terms of two of the major conceptualizations of human agency and of society that were discerned in the section above about the rise of the social sciences at the turn of the eighteenth century.

In a rationalistic-deliberative tradition, a landmark was established by the publication of Kenneth Arrow's (1951) *Social Choice and Individual Values*. A few years later, economic theorizing would enter political science at large with Anthony Downs's (1957) book *An Economic Theory of Democracy*, to be followed in the early 1960s by a range of books applying game theory and social choice theory to political phenomena.

(Incidentally, as testified by Gabriel Almond (1990, p.117) the institutional role of the Center for Advanced Study in the Behavioral Sciences at Stanford was a crucial one for some of the interactions involving Arrow, Dahl, and others. These interactions were of immediate importance to the very publication of Downs's book but also more generally to the breakthrough of rationalistic forms of analysis).

A few decades later, what had once appeared as an interesting but rather marginal concern of some specialized political theorists had evolved into a broad tradition of rational choice analysis with pervasive influence on both the theory and practice of political thought, particularly in North America. Society as a compositional collective was described not just as an analytically convenient metaphor but as a representation of societal order and with consequential claims concerning the nature of legitimate political action. One thrust of theorizing in this tradition was to demonstrate the limitations, or even the impossibility, of delimiting an unambiguous common good, whether arrived at with the help of a social welfare function or of a majority decision. It is in this context that John Rawls's (1971) *A Theory of Justice* appeared as a pathbreaking book. Drawing on an individualistic and contractual form of reasoning, succinctly criticised by several scholars, most notably perhaps Michael Sandel, Rawls was able to reintroduce discussions of justice and just distributions and allocations into the core of political philosophy. Even after decades of discussions and critique, both Arrow's and Rawls's contribution retain their stature. However, both of them signal a rupture with historical and contextual reasoning rather than the opposite.

In the broad tradition of linguistic-interpretative conceptions of human action and of society as a form of emergent totality in historical context, important new developments also occurred. Thus contributions by continental European phenomenological and hermeneutic philosophers in the 1950s and 1960s came to merge with influences from Anglo-Saxon linguistic philosophy in the tradition of Wittgenstein and Austin. Three forms of such thought, in particular, emerged and have exerted a deep influence on the ways of studying the history of political thought.

In 1959, Reinhart Koselleck (1987), drawing on a tradition of hermeneutic and historical scholarship, published a book on critique and crisis. This book had the deep transformation of intellectual and political order at the turn of the eighteenth century as its focus but also contained the core of a research program on the history of key concepts of society and political order in a long-term perspective. This program was realized through the publication during the following decades of a set of nine massive volumes tracing the history of a limited number of such concepts, primarily in continental Europe.

Roughly at the same time, Jürgen Habermas elaborated another research program, drawing on the joint traditions of critical Marxist thought, hermeneutics, and speech act theory. In 1962 Habermas published a volume that had an analogous empirical focus as that of Koselleck (Habermas 1989). Thus it was a reinterpretation of the emergence of a new type of public sphere in the late eighteenth century. Habermas's subsequent long-term program has consistently linked these different traditions and may in one sense be seen as a form of historical positioning but also a radical generalization of the philosophical insights of linguistic philosophy and speech act theory.

This later tradition also inspired Quentin Skinner's (1978) two volume major opus *The Foundations of Modern Political Thought* (a work, incidentally, written largely at the Institute for Advanced Study at Princeton). Skinner came to be the major representative of a new school, often identified with his—and John Pocock's old—university, namely Cambridge. It involved the setting of new and significantly more rigorous standards for the writing of the history of political thought, particularly in its demand that the linguistic and intellectual context of the actual writing of major works be carefully brought to the fore of analysis. After decades of discussions, there can be little doubt that intellectual history has made advances in this period, inspired by the works of Skinner and several of his colleagues, that still by no means have been fully explored. At the same time, it is also clear that these works still leave room for a more extensive inquiry into the relationship between long-term institutional and intellectual change.

Finally, links between political thought and political agency are prominent in studies of mechanisms of inclusion and exclusion and on the granting and withholding of entitlements for human beings. Gender studies has clearly entailed a renaissance of the writing of the history of political thought with pioneering contributions by Carole Pateman and Seyla Benhabib, just to mention two of the most distinguished scholars on the late twentieth century.

Historically orientated scholarship has in recent years seemed to open up the possibility for situating political thought in a much more encompassing context than has hitherto been customary. Not least, scholars writing in countries outside of the North Atlantic and West European context seem now to be about to help bring about a more truly global effort to rethink the history of political thought. That would also, almost inevitably, contribute to a transcendence of some of the most severe limitations of contemporary political thought, in particular the following ones:

The relative inability of political thought in the rationalistic tradition to transcend a theory of agency that is manifestly impoverished vis-à-vis available historical research but also vis-à-vis demands that social and political thought itself cannot but make concerning a theory of agency (cf. Joas 1996, 2000).

The relative inability of the so-called communitarian critics of this tradition to supply a sustained historical account of their

preferred community and to escape a tendency of entering into circular justifications of those very societies which they themselves happened to be located or identified with.
The relative inability of research in the traditions of conceptual history and linguistic contextualism to contribute to an understanding of the institutional formation of political order in a long-term and also extra-European perspective.

In order to be able to advance beyond these limitations, it seems that political thought must become significantly more prepared to engage in studies of long-term processes of intellectual and institutional transformations. This will in all likelihood entail efforts to invigorate historical social science of a type that was characteristic of the classics of social science, not least Weber (cf. Schluchter 1996). It would then also stimulate culturally and contextually sensitive social science of a type that has been advocated most ambitiously and consistently by scholars across a range of the social and historical sciences and with S. N. Eisenstadt (1995) as one of its most prominent representatives. The history of political thought in the last three decades of the twentieth century involved impressive advances. At the turn of the millennium, however, the need for historically and comparatively orientated research programs able to transcend the limitations of contemporary political thought also stand out as possibly more urgent than at any time since the emergence of social science itself in the end of the eighteenth century.

See also: Communitarianism: Political Theory; Democratic Theory; Enlightenment; Hermeneutics, History of; Integration: Social; Intellectual History; Legitimacy: Political; Liberalism: Historical Aspects; Locke, John (1632–1704); Marxist Social Thought, History of; National Socialism and Fascism; Rational Choice in Politics; Scientific Disciplines, History of; Solidarity: History of the Concept

Bibliography

Almond G A 1990 *A Discipline Divided: Schools and Sects in Political Science*. Sage, Newbury Park, CA

Arrow K 1951 *Social Choice and Individual Values*. Wiley, New York

Benhabib S 1992 *Situating the Self: Gender, Community and Postmodernism in Contemporary Ethics*. Routledge, New York

Brian E 1994 *La Mesure de l'Etat: Administrateurs et Géometres au XVIII Siècle*. Albin Michel, Paris

Downs A 1957 *An Economic Theory of Democracy*. Harper, New York

Eisenstadt S N 1995 *Power, Trust and Meaning*. University of Chicago Press, Chicago

Elliott J H 1992 A Europe of composite monarchies. *Past and Present* **137**: 48–71

Fox C, Porter R, Wokler R (eds.) 1995 *Inventing Human Science: Eighteenth Century Domains*. University of California Press, Berkeley, CA

Habermas J 1989 *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. MIT Press, Cambridge, MA

Harré R 2000 *One Thousand Years of Philosophy: From Ramanuja to Wittgenstein*. Blackwell, Malden, MA

Heilbron J 1995 *The Rise of Social Theory*. Polity Press, Cambridge, UK

Heilbron J, Magnusson L, Wittrock B (eds.) 1998 *The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850*. Kluwer, Dordrecht, The Netherlands

Joas H 1996 *The Creativity of Action Polity*. University of Chicago Press, Chicago, IL

Joas H 2000 *The Genesis of Values Polity*. University of Chicago Press, Chicago, IL

Koselleck R 1988 [1957] *Critique and Crisis: Enlightenment and the Pathogenesis of Modern Society*. MIT Press, Cambridge, MA

Lefort C 1988 *Democracy and Political Theory*. University of Minnesota Press, Minneapolis, MN

Manent P 1994 *An Intellectual History of Liberalism*. Princeton University Press, Princeton, NJ

Pagden A (ed.) 1987 *The Languages of Political Theory in Early Modern Europe*. Cambridge University Press, Cambridge, UK

Pocock J G A 1975 *The Machiavellian Moment*. Princeton University Press, Princeton, NJ

Popper K 1966 [1945] *The Open Society and its Enemies*, 2 vols. Routledge, London

Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA

Sabine G 1937 *A History of Political Theory*. Holt, New York

Schluchter W 1996 *Paradoxes of Modernity: Culture and Conduct in the Theory of Max Weber*. Stanford University Press, Stanford, CA

Schumpeter J A 1943 *Capitalism, Socialism and Democracy*. Allen & Unwin, London

Shklar J N 1998 *Political Thought and Political Thinkers*. University of Chicago Press, Chicago

Skinner Q 1978 *The Foundations of Modern Political Thought*, 2 vols. Cambridge University Press, Cambridge, UK

Skinner Q 1996 *Reason and Rhetoric in the Philosophy of Hobbes*. Cambridge University Press, Cambridge, UK

Talmon J L 1960 *The Origins of Totalitarian Democracy*. Praeger, New York

Tocqueville A de 1964 [1850] *Oeuvres Complètes, Vol. XII, Souvenirs*. Gallimard, Paris

B. Wittrock

Political Trials

The term 'political trial' is popularly associated with 'show trials' like the Moscow Purge trials that were conducted by Stalin during the 1930s. Such trials represent an extreme example of political repression and are viewed as sham legal proceedings designed by the authorities to dramatize specific political campaigns and/or to eliminate prominent individuals. In this context the political trial is taken to be a 'show' in which politicians and not jurists are those who pull the strings and determine the results of the trial in advance. This pejorative use of the term also helps designate regimes that resort to such techniques as the antithesis

of liberal democracies. Yet, as Otto Kirchheimer observed, the most effective usage of political trials occurs precisely within liberal democracies, in societies committed to the rule of law. What can explain this gap between popular perceptions and the actual practice of political trials? Can we distinguish a political trial from a 'show trial'? Should we continue to view political trials as opposed to the very essence of liberal democracies, or should we accept them as part and parcel of this system, deserving a sustained theoretical elaboration?

In his seminal book *Political Justice* Kirchheimer defines political trials as instances in which 'court action is called upon to exert influence on the distribution of political power' (Kirchheimer 1961). This broad definition includes the use of the court by the ruling elites or by outgroups and dissenters to achieve political goals by judicial means. Many writers regard the central factor in such trials as the political motive behind them. The connecting thread between the trial of Socrates for corrupting the youth of Athens (399 BC), the trial of Jesus for blasphemy and sedition (AD 30), the trial of Joan of Arc (1431) for heresy and witchcraft, the trial of Thomas More (1534) for remaining silent when asked about Henry VIII's supremacy in religion, the trial of Galileo (1633) for heretically suggesting that the earth moved around the sun, and many other trials is that, in each case, men in power believed that the defendant was a threat to them (Becker 1971, Belknap 1981). For this reason the prototype of a political trial has been the criminal trial of a political adversary for political reasons. This broad definition encompasses trials in which the defendant has directly attacked the established order by acts such as treason or sedition, but also trials involving common crimes in which only the personality or the motive of the offender indicates their political nature. Other writers try to refine this definition by expanding its scope to legal proceedings other than criminal trials such as impeachment, congressional hearings, and truth commissions. It has also been suggested that the definition be extended to include trials involving larger group conflicts such as ethnic, race, or labor struggles. Some move the emphasis from the political motive of the authorities to the political identity of the defendant. Yet others conclude that the important factor is not the motive but the competition of stories over the identity of the polity that distinguishes such trials. Notwithstanding their differences, all writers on the subject agree that political trials are inevitable not only in authoritarian regimes but also in constitutional democracies, although this is generally denied. Critical of such denials, they set out to replace them with an honest recognition of the existence and function of political trials in both types of regimes.

Why should the term political trial cause such anxiety and widespread denial? Since law tends to stabilize the status quo it is not surprising that there are always people who feel justified in ignoring the law in order to bring about drastic and necessary changes. This is especially true for powerless groups that are prevented from participating effectively in important institutional decisions affecting their lives (Hakman 1972, Fletcher 1995). When the authorities put on trial those who challenge the basic value distribution in society, especially those who are willing to disregard existing laws to this end, the result is a political trial. Despite this regular occurrence of political trials within democratic societies the term is widely resisted because of its connotation with notions of 'show trials' in which either the charge is a 'frame-up' to conceal the political motivation of the prosecution, or the hearing itself fails to meet minimum standards of judicial impartiality and independence. For this reason, writers attempt to distinguish among different types of political trials. One classification distinguishes among the politically motivated trial, politically determined trial, and trials with substantial political consequences (Friedman 1970). Another differentiates between political 'trials' (abridgment of due process guarantees) the 'political' trial (camouflaged as a criminal trial) and the political trial (politically motivated) (Becker 1971). These efforts at classification have proved frustrating because of the immense historical variations and the widespread existence of political trials in very different regimes. Writers have identified a variety of characteristics of political trials but none has proved to be a sufficient and necessary condition. This literature has largely been descriptive but has failed to assess the compatibility of political trials with liberal values and thus to overcome the negative connotations the term arouses. In order to address this question it is necessary to expose the hidden engine of political trials—their legitimizing function—and its relation to democracy.

The legal system offers a tempting opportunity for those in power to damage enemies, tarnish their image, and isolate them from potential allies by casting them as criminals. This temptation becomes especially strong in democracies that constrain the ruling elite from resorting to more direct devices of repression. What distinguishes political trials from other political devices is their legitimizing function, i.e., their ability to turn a political adversary into a criminal defendant and thereby reinforce the legitimacy of those in power. However, this function can be served only if the public perceives the trial as providing the defendant with a 'fighting chance.' In constitutional democracies this 'fighting chance' is generally assured by the relative autonomy of the judges and the constraining effects of the rule of law (Kirchheimer 1961). Moreover, these guarantees introduce an element of uncertainty into the trial, and encourage political defendants to use it to expose the illegitimacy of the policy or value system advanced by the authorities. Indeed, from Socrates's trial (399 BC) to the Chicago Conspiracy trial (1969), there have always been defendants who have exploited

the limited forum that the political trial provided them to articulate their oppositional views. Some have adopted a defiant stance, proclaiming their right to commit the crime (Solomon Tellirian—as a protest against the Armenian massacre by the Turks in 1921; Dr Jack Kevorkian—advocating euthanasia in 1999), or refusing to respect the traditional decorum of the courtroom altogether (Bobby Seale–Black Panthers in 1969; Klaus Barbie–Nazism and Vichy France in 1987); others have played by the rules in order to win an acquittal that would confirm their claim that they were being prosecuted for their political beliefs (Dr Spock–Vietnam War in 1968). Insight into this legitimizing function and the element of uncertainty that it introduces into political trials in democratic regimes exposes the tension between politics and justice inherent in them. This has made political trials a fascinating area for research since the 1960s.

1. *Political Trials in the Transition to Democracy*

Political trials are most salient in times of transition between regimes, especially when a new democratic regime confronts crimes of the old regime. Indeed, the first attempts to seriously consider the compatibility of political trials with liberal–democratic values appear in the literature on the transition to democracy. Unlike a military revolution that sustains its authority by brute force, democratic regimes are committed to the rule of law and are inclined to address the evils of the previous regime with the help of legal devices. However, the new regime's commitment to the rule of law also makes it aware of the dangers of using *ex post facto* laws and indulging in 'victor's justice.' At such times, the various expectations from the law—to punish the guilty, to ascertain the truth about the old regime, and to enhance reconciliation in society—seem to overwhelm the legal system and to push it in opposite directions. Moreover, the forward-looking direction of the architects of democracy who are concerned with the efficacy of the transition often conflicts with the backward-looking direction of legal proceedings and their narrow focus on individual guilt (Osiel 1997, Nino 1991, Teitel 1997). For these reasons trials of transition bring to the foreground the clash between politics and justice. Two main approaches to the problem have evolved since World War II: Exemplary Criminal Trials and Truth Commissions, both of which can be considered political trials of sorts and have received extensive theoretical elaboration.

Following the end of World War II the Allies established an international military tribunal in Nuremberg (1945) to judge the leaders of the Nazi regime. This was the first time in history that the leaders of a defeated country had faced criminal prosecution for war crimes, crimes against the peace, and crimes against humanity by an international tribunal. The charge that was applied in the trial—conspiring to wage an aggressive war—allowed the prosecution to tell a general historical narrative. Could the obvious political end of using the court to teach a history lesson (through *ex post facto* laws, and victors' judges) be reconciled with the demands of liberalism? Confronted with this dilemma, Judith Shklar made the first serious effort to reconcile the legacy of liberalism with political trials occurring in the transition to democracy (Shklar 1964). A legalistic understanding of liberalism, she argued, could justify the Nuremberg trials only at the high cost of denying their political nature altogether. In her view, procedural safeguards to defendants could only guarantee that the trial would not deteriorate into a show trial, but this in itself could not justify deviating from basic liberal demands such as an independent judiciary, established court, nonretroactive laws etc. Paradoxically, Shklar came to the conclusion that the political dictate of founding a democracy justified the divergence from strict legalism in the Nuremberg trials. In other words, it was only an honest recognition of the political aspects of the trials, their educational and symbolic contribution to the building of democracy in postwar Germany, that could justify them. Later writers accepted Shklar's consequentialist approach to transitional justice but limited its applicability to trials that guarantee the due-process rights of the defendants (Osiel 1997).

A very different attempt to address the crimes of the Nazi regime was made in the Eichmann trial (1961). This time it was not the international community but an Israeli court that undertook to judge one of the central figures in the administrative massacre of the Jews. The political nature of the trial was evident, among other things, from its reliance on the particularistic law of 'crimes against the Jewish people.' Likewise the heavy reliance on victims' testimonies about their suffering (and through them the telling of the Jewish Holocaust) conflicted with the liberal demand that the trial concentrate on proving the deeds of the accused. Hannah Arendt and other writers criticized the Israeli prosecution for politicizing the trial in this way (Arendt 1963, Segev 1993, Lahav 1992). Others see the decision as stemming from the general dilemma of political trials of how to strike a balance between the need to establish the truth about an oppressive regime, the need to give voice to the victims, and the need to ascertain the guilt of the defendants (Bilsky 2001).

This dilemma is compounded when the former authoritarian regime retains enclaves of power and opposes any attempt to bring to justice the officials of the former regime. In South America, Argentina remains the only nation that has indicted and prosecuted military officers in public trials for their role in acts of repression carried out under the former regime. The political nature of these trials was evident in the decision to prosecute only the masterminds and the worst offenders, and in the greater priority given to the value of learning about the evils of the authoritarian

past than to prosecuting every individual who was involved. Again, the justification for such selective prosecution is given in consequentionalist terms—to what extent the trial serves to strengthen the emerging democracy (Nino 1991). A central dilemma that emerges in transitional trials of this kind is the breakdown of common discourse which undermines the legitimacy of the trial for some segments of society. Thus, the Argentinian court's decision to limit the evidence to proving the deeds of the accused, without allowing them to be situated within the wider historical narrative about an alleged war of self-defense against attacks on the regime by guerrilla forces, was deemed political and undermined the legitimacy of the trial in the eyes of the military (Osiel 1997).

Side by side with such efforts to bring to trial leaders of former authoritarian regimes for gross violations of human rights, there has developed the institution of truth commissions established by national or international organizations. Truth commissions reduce the tensions that arise in transitional trials by separating the various functions of learning the truth about the past, of ascertaining individual guilt, and of giving voice to victims. The Truth and Reconciliation Commission in South Africa, for example, was divided into a Committee on Human Rights Violations (victims' testimonies), a Committee on Amnesty (offender's testimonies), and a Committee on Reparation and Rehabilitation. In each committee different procedures were used (1995 Act). It further made the granting of individual amnesty conditional upon providing a full account of the offenses that were committed (Minow 1998). Truth commissions have proved to be far more effective than court proceedings in furnishing a dramatic medium for theatricalizing the new official history since the narratives of victims are rarely interrupted by lawyers and there are incentives for offenders to relate their part in the repression. Indeed, in several countries the commission's proceedings have been broadcast daily, and in others the final report, Nunca Mais (Never Again), has become a best seller (Brazil, Argentina.) Truth commissions present us with a kind of political trial in which the balance between politics and justice has been struck in ways more compatible with liberal concerns.

2. *Political Trials in Established Democracies*

The political nature of transitional justice seems from the fact that there are no overarching legal norms that are accepted as legitimate by the two successive regimes. Moreover, in these cases the trials have overwhelmingly been used to fulfill social and political functions other than ascertaining individual guilt. Liberal scholars, as we have seen, have been willing to recognize the political nature of such trials and to justify it, but have not extended their theories to trials in established democracies. The general view is that in normal times law should endeavor to keep politics out of the courtrooms so that every political trial is perceived as a corruption of the rule of law (which may be termed the pathology thesis). This distinction, however, fails to account for the striking similarities between transitional situations and periods in which there has been a significant increase in political trials in democratic societies.

In the United States political trials have occurred whenever the status quo has been challenged, generally during periods of social and political ferment unleashed by such forces as war, economic conflict, or racial discord (Belknap 1981). For example, the decade that began with the escalation of the Vietnam war in 1965 brought an epidemic of political trials culminating in the Chicago Conspiracy trial (1969–70). This case represents a microcosm of American political justice of the era, as it was directed against representatives of major antiwar groups, youth counterculture, and the Black Panther Party who all believed that the war was illegal and sought to mobilize the public against it. Although the charges were clearly political (conspiracy to cross state lines with the intent to incite riot), the trial also demonstrated the ability of the defense to politicize the trial and turn it into a forum for social protest by flouting the norms of the courtroom and ridiculing the judge (Ely 1981). In such trials as the Chicago Conspiracy trial the same breakdown of common discourse was manifested as in political trials during the transition to democracy.

The social turmoil and political trials of the late 1960s influenced the development of radical theories of law (Gordon 1982) scholars in critical legal studies, critical race theory; and feminist legal studies began to question the plausibility of the liberal ideal of separating law from politics. For these scholars the very attempt to delimit the boundaries of the category 'political trials' obscures the way in which politics enters every trial and every field of law. They argue that the liberal reliance on a distinct category of political trials often serves to legitimize the status quo. Critical theory understands politics not in the narrow sense of the term (motive of authorities, identity of defendant, etc.) but as the hegemonic ideology that shapes the interpretation of law while presenting it as neutral. The main contribution of critical writings to understanding political trials has been their rejection of the 'pathology thesis' about the relation between law and politics and their systematic efforts to uncover the ideological structures that shape different areas of law (Kelman 1987, Kairys 1982). This broad definition of political trials, however, had its drawbacks because it no longer called for a systematic investigation of the unique features of classic political trials, and shifted the attention of scholars from the court drama and social reception of the trial to the appellate court interpretation of the law.

Other scholars, still committed to the liberal framework, have identified a new form of political trial that emerged during the 1980s and 1990s. While in the old political trial the ruling authorities selected certain individuals to stand for an opposition the state wanted to eliminate, in the new political trial a section of the public turns the trial into a political trial by identifying with the victims (who are not a formal party to the trial) or with the defendant. In these cases the state does not intend a political trial and has very little control over its politicization (Fletcher 1995). The people mobilized around these trials are usually outgroups protesting their marginalization by the legal and political institutions of the state. Examples are famous rape trials and self-defense trials of battered women who killed their abusive spouses in which women's group identify with the victim and politicize the trial. The Rodney King trial and O. J. Simpson trial likewise became political trials in which African–American groups were mobilized to protest against white justice in America. The study of these trials draws particular attention to the role of the media in these cases.

The focus on group conflict and on the narrative and rhetorical aspects of political trials in on-going democracies provides scholars with the key to understanding their dynamics from a pluralistic perspective. The famous trials of Socrates, Jesus, Dreyfus, and others are examples of heroes unjustly prosecuted, but they can also be viewed as major junctions in the life of the republic where society's conflicting values are played out not only, and not even mainly, through the learned interpretation of the law, but through the human drama in and around the courtroom. In these trials the social conflict is transformed into competing narratives that capture the public's attention and offer an opportunity for collective self-reflection. Through an examination of competing values and loyalties they bring together for public consideration society's basic contradictions (Christenson 1999). From this perspective we can see how political trials, while threatening the rule of law, at the same time may contribute to the development of a more critical and democratic society.

See also: Political Lawyering; Political Protest and Civil Disobedience; Trials: Cultural

Bibliography

Arendt H 1963, 1964 *Eichmann in Jerusalem, A Report on the Banality of Evil*. Penguin Books, New York

Barrett W P 1931 *The Trial of Jeanne d'Arc*. George Routledge & Sons, London

Becker T L (ed.) 1971 *Political Trials*. Bobbs-Merrill, New York

Belknap M R 1981 *American Political Trials*. Greenwood Press, Westport, CT

Bilsky L 1999 *Political Trials: A History of Misrecognition*. Unpublished manuscript

Bilsky L 2001 Between justice and politics: The competition of storytellers in the Eichmann trial. In: Aschheim S (ed.) *Hannah Arendt in Jerusalem*. University of California Press

Boyle J (ed.) 1994 *Critical Legal Studies*. New York University Press, New York

Christenson R 1999 *Political Trials, Gordian Knots in the Law*. Transaction Publishers, New Brunswick, Canada

Conot R 1983 *Justice at Nuremberg*. Harper & Row, New York

Cover R M 1983 The Supreme Court, 1982 Term—Foreword: nomos and narrative. *Harvard Law Review* **97**: 4–70

De Brito A B 1997 *Human Rights and Democratization in Latin America: Uruguay and Chile*. Oxford University Press, Oxford, UK

Dorsen N, Friedman L (eds.) 1974 *Disorder in the Court, Report of the Association of the Bar of City of New York, Special Committee on Courtroom Conduct*. Pantheon Books, New York

Du Plessis L 1994 Amnesty and transition in South Africa. In: Boraine A, Levy J, Scheffer R (eds.) *Dealing with the Past: Truth and Reconciliation in South Africa*. IDASA, Capetown, South Africa

Ely J W 1981 The Chicago conspiracy case. In: Belknap M R (ed.) *American Political Trials*. Greenwood Press, Westort, CT

Fletcher G F 1995 *With Justice for Some, Victim's Rights in Criminal Trials*. Addison-Wesley, Reading, UK

Friedman L 1970 Political power and legal legitimacy: A short history of political trials. *Antioch Review* **30**: 157–71

Gagne P 1998 *Battered Women's Justice*. Twayne, New York

Gordon R W 1982 Some critical theories of law and their critics. In: Kairys (ed.) *The Politics of Law*. Basics Books, New York

Hakman N 1972 Political trials in the legal order: A political scientist's perspective. *Journal of Public Law* **21**: 73–126

Hariman R (ed.) 1990 *Popular Trials, Rhetoric, Mass Media and the Law*. The University of Alabama Press, Tuscaloosa, AL

Harris W R 1954 *Tyranny on Trial, The Trial of the Major German War Criminals at the End of World War 2 at Nuremberg, Germany, 1945–1946*. Southern Methodist University Press, Dallas, TX

Hesse C, Post R (eds.) 1999 *Human Rights in Political Transition*. Zone Books, New York

Kairys D (ed.) 1982 *The Politics of Law, A Progressive Critique*. Basic Books, New York

Kelman M 1987 *A Guide to Critical Legal Studies*. Harvard University Press, Cambridge

Kirchheimer O 1955 Politics and justice. *Social Research* **22**: 377–398

Kirchheimer O 1961 *Political Justice, The Use of Legal Procedure for Political Ends*. Princeton University Press, Princeton, NJ

Lahav P 1992 The Eichmann trial, the Jewish question and the American-Jewish intelligensia. *Boston University Law Review* **72**: 555–575

MacKinnon C A 1989 *Toward A Feminist Theory of the State*. Harvard University Press, Cambridge, MA

Minow M 1998 *Between Vengeance and Forgiveness, Facing History after Genocide and Mass Violence*. Beacon Press, Boston

Nino C S 1991 The duty to punish past abuses of human rights put into context: The case of Argentina. *Yale Law Journal* **100**: 2619–640

Nino C S 1996 *Radical Evil on Trial*. Yale University Press, New Haven

Orentlicher D F 1991 Settling accounts: The duty to prosecute human rights violations of a prior regime. *Yale Law Journal* **100**: 2537–615

Osiel M 1997 *Mass Atrocity, Collective Memory and the Law*. Transaction Publishers, New Brunswick, Canada
Segev T 1993 *The Seventh Million: The Israelis and the Holocaust*. Hill and Wang, New York
Shklar J N 1964 *Legalism, Law, Morals and Political Trials*. Harvard University Press, Cambridge, MA
Stone I F 1989 *The Trial of Socrates*. Doubleday, New York
Teitel R 1997 Transitional jurisprudence: The role of law in political transformation. *The Yale Law Journal* **106**: 2009–80
Zalaquett J 1988 Confronting human rights violations by former governments: Principals applicable and political constraints. In: Henkin A (ed.) *State Crimes: Punishment or Pardon*. Aspen Institute, Queenstown, MD, 23–71

L. Bilsky

Politics, Formal Modeling of

Positive theories of politics provide explanations for why political phenomena occur in the manner that they do. Examples of such phenomena include which parties or candidates are elected at certain times, what policies legislative bodies adopt, and when and how wars are fought between countries. Formal models complement these explanations by using the language and analysis of mathematics to derive the phenomena as necessary consequences of certain sets of underlying factors. As with mathematical models in other disciplines, the main advantage of such an approach is twofold: to give an unambiguous representation of the relevant moving parts of a theory (the clarity of the argument) and to enforce a logical coherence on a theory (the consistency of the argument). The main drawback is found in the required simplification of the phenomena down to a manageable number of moving parts, and the subsequent possibility of simplifying away the very essence of the problem.

1. *The Basic Structure*

Most formal models of politics fall in the category of rational choice (e.g., see *Rational Choice in Politics*), where the presumption is that observed outcomes are the result of decisions made by a certain relevant set of individuals. These could be voters and candidates in the first example above, elected representatives and appointed ministers in the second, or heads of state in the third. The general behavioral postulate is that individuals attempt to influence the outcome in a manner reflecting their preferences (tastes, values, etc.) over the set of possible outcomes. For instance, in selecting campaign platforms, a candidate who only cares about winning the current election might look to adopt policies appealing to the greatest number of voters; conversely, a candidate who cares intrinsically about the policies to be implemented might shy away from politically popular but personally unpalatable alternatives. Whatever their specific form, individuals' preferences are taken to be primitives of the model, and so are held fixed throughout the analytic determination of an outcome. However, different specifications of preferences can give rise to different outcomes. Indeed, to the extent the predicted outcome varies with changes in the underlying individual preferences, one has a potential explanation for any variation in observed outcomes as being due to variation in these preferences. The extent to which these preferences can themselves be modeled as functions of observable parameters, such as wealth or legislative seniority, permits one to generate a theoretical relationship between these parameters and observable political outcomes as well.

The influence of preferences on outcomes occurs through the choices the individuals make. The most common assumption is that each individual's choice is optimal with respect to their preferences, for example, in the above scenario the candidate selects a vote-maximizing policy. This optimality assumption may be far from a realistic representation of individual decision-making, implying as it does that individuals can effortlessly solve what are at times relatively difficult mathematical problems. However, assuming otherwise typically requires the presence of auxiliary factors (for instance, decision costs) to rationalize *non*-optimizing behavior, factors which, while reasonable, tend to create additional complications in the model. Also the optimality assumption allows one to import existing results from optimization theory to aid in the analysis.

As seen in the examples above, most formal models of politics involve more than one individual, and so this individual level of optimization often takes place within the confines of non-cooperative game theory. Such models explicitly describe the interaction among the individuals that ultimately produces a collective decision. This description begins with a 'game form' consisting of (a) the set of strategies available to each individual, and (b) the specification of which outcome occurs when a given profile of strategies is chosen (with each profile consisting of one strategy per individual). Continuing with the electoral setting, one example of a game form is where the set of strategies for a candidate is equal to the set of available policies, a voter's strategy describes which candidate to vote for based on the candidates' chosen policies, and an outcome, consisting of who wins the election and with what policy, is determined by the candidate receiving the most votes. The remaining piece of the puzzle then is the derivation of the individuals' strategies, with the fundamental concept being that of the 'Nash equilibrium.' Note that, even assuming all individuals have well-defined preferences over the set of outcomes, the notion of optimality in this multiperson environment is necessarily conditional, as an individual's best choice of strategy may depend nontrivially on the strategy

choices by the other individuals. For instance, in the electoral game above, the location of a candidate's optimal policy may depend on the voters' behavior as well as the policies endorsed by rivals. A Nash equilibrium is a profile of strategies wherein each individual's strategy is optimal, given the strategies of the others, that is, conditional on the choices by the other individuals, no one individual has an incentive to alter her choice. A Nash equilibrium outcome is thus a collective decision consistent with individually optimal behavior and constitutes the basic object of analysis in any noncooperative model. (In certain game forms, some Nash equilibria are inherently more plausible than others and so refinements of Nash equilibria, such as subgame perfection, are used.)

As with the choice of preferences, the choice of game form is in the hands of the modeler. This choice may be influenced by observable features of the political phenomenon in question, as in the above electoral example. In such situations, there will invariably be a trade-off between making the game form descriptively accurate and keeping the game form analytically tractable. Thus while the electoral rule may easily be incorporated, simplifying the campaign stage of the process to a single choice of policy, thereby avoiding the complexities found in actual elections, may be necessary to guarantee the existence of and to solve for equilibrium outcomes. At other times, there may not exist any such observable features to leverage, an example of which would be when two heads of state are bargaining over land and other resources in the shadow of war. Here one would be hard-pressed to assert that the unstructured negotiations taking place can be definitively captured in any one particular game form. Hence descriptive accuracy is not an issue, and the modeler is relatively unconstrained in the search for a 'good' game form, that is, a game form capable of generating new and valuable insights into the political phenomenon in question.

Yet to the extent that the game form incorporates observable features of the political process and that the equilibrium outcomes vary with these features, one has a second source of variation in outcomes above and beyond changes in preferences, namely, changes in the political process. For instance, the equilibrium outcomes under the above plurality electoral rule may differ from those associated with a run-off rule (where, if no candidate receives a majority of the votes, the two candidates receiving the highest number of votes are paired in a second election, in which majority rule determines the winner) even if the candidates' and voters' preferences are held fixed. This allows one to make statements about the degree to which differences in observed collective choices are due to differences in preferences, or to differences in political institutions. More ambitiously, this type of comparative institutional analysis can be used to predict the effect changes in existing political institutions will have on collective outcomes.

2. *An Alternative Approach*

While noncooperative game theory has tended to dominate the modeling landscape in recent years, much of the early formal work employed the techniques of cooperative game theory. Rather than explicitly modeling outcomes as the aggregation of individual decisions, under this approach one looks for outcomes that are stable in some sense, with this 'sense' implicitly characterizing the (unmodeled) underlying strategic interaction. Given the normative appeal and frequent use of majoritarian-based political procedures, the focus was primarily on identifying outcomes which are majority-preferred to all others in pair-wise comparisons. Such an outcome would then be immune to changes proposed by any majority coalition, whereas all other outcomes would not.

It turns out that a majority-stable outcome necessarily exists in certain environments, the most compelling being when the collective choice concerns the selection of a single number. For example, suppose the issue before a legislature is the size of the nation's military budget, the set of possible outcomes can be taken to be the set of nonnegative real numbers. Suppose further that the legislators have 'single-peaked' preferences: each has some best number, and the value to them of the outcome falls as one moves away from this best number. Then a stable outcome will exist and, if the number of individuals is odd, will be unique and equal to the median of these best numbers. This 'median voter theorem' has been at the core of numerous and varied empirical studies of collective choice, from local school budgets in the United States to European Union decision-making. Its comparative advantage as a formal model is obviously its parsimony, requiring as it does only a description of how individual preferences over outcomes are related to any underlying parameters of interest (e.g., household income in the local school budget case). Further, by imposing additional assumptions on preferences, one finds that the median voter's preferences coincide with those of the majority over the entire set of outcomes, implying that the median voter is in a literal sense representative of the collective as a whole. The benefits from this result accrue principally to the noncooperative modeler: suddenly a one-dimensional model of the interaction between the US President and the 435 members of the US House of Representatives over the size of the military budget, with majority rule dictating the decision-making among the latter, can be simplified to a two-person interaction between the President and the median member of the House.

The main shortcoming of the median voter theorem of course is its restriction on the set of outcomes to a single dimension. In certain situations this is benign, as in the above example when the policy choice in question consists of selecting the military budget.

However, a policy choice consisting of the distribution of this budget across more than two groups is inherently a multidimensional problem and the conclusions about majority rule for multidimensional problems, relative to their single-dimensional cousins, could hardly be more dramatic. In particular, for most descriptions of preferences, no majority-stable outcome exists (where 'most' has a mathematically precise definition in this setting). As an example of such nonexistence, suppose in the above scenario that each legislator cares only about how much their favorite group receives, with one-third of the legislators associated with each of three groups. Then for any distribution of the budget, the legislators for two groups (which together constitute a majority) prefer to take away any amount going to the third group and redistribute it amongst themselves, thereby guaranteeing that no distribution is majority-stable. Furthermore, when no stable outcome exists, not only is there no representative voter but in fact one can construct majority preference cycles throughout the policy space (i.e., policy *a* is majority-preferred to policy *b*, which is majority preferred to policy *c*, which … which is majority-preferred to policy *z*, which is majority-preferred to policy *a*). Thus when majority rule breaks down, it breaks down completely. Subsequent research has shown that these negative conclusions continue to hold, to a greater or lesser degree, for most nonmajoritarian rules as well.

As with the median voter theorem, it is difficult to overstate the impact these 'chaos theorems' have had on the formal modeling of politics. Clearly, in a multidimensional world, the modeler has lost the analytic shortcut of representing a collective's decision-making as being governed by a single individual (or, at least, the logical rationale for doing so). More importantly, the theorems gave rise to a new-found appreciation for political institutions, such as committee systems, restrictive voting rules, and the like as key elements in the determination of collective outcomes. Indeed, such institutions are often depicted as necessary ingredients in any multidimensional model of politics, in that, but for such institutions, collective decision-making would fall prey to the inevitable onslaught of cycling and chaos. This interpretation, while quite popular, appears inconsistent with the central tenet of formal modeling, that of equilibrium analysis, as the chaos theorems are properly seen as nonequilibrium results. Put differently, these theorems do not predict that (as commonly phrased) 'anything can happen,' for as nonequilibrium results they do not predict anything at all. Rather, they demonstrate the impossibility of any general theory of collective behavior based solely on the (majority) aggregation of preferences, and hence the necessity of being more explicit about the decision-making process so as to provide the analytical traction required to formulate a well-posed model. From this perspective, the lesson to be learned from the chaos theorems is one for the modeler of collective decision-making, rather than as one about collective decision-making itself.

See also: Action, Collective; Action, Theories of Social; Behavioralism: Political; Bounded Rationality; Game Theory: Noncooperative Games; Model Testing and Selection, Theory of; Systems Modeling

Bibliography

Austen-Smith D, Banks J S 1999 *Positive Political Theory I: Collective Preference*. University of Michigan Press, Ann Arbor, MI

Morrow J D 1994 *Game Theory for Political Scientists*. Princeton University Press, Princeton, NJ

Ordeshook P C 1986 *Game Theory and Political Theory: An Introduction*. Cambridge University Press, Cambridge, UK

Shepsle K A, Bonchek M S 1997 *Analyzing Politics: Rationality, Behavior, and Institutions*. Norton, New York

J. S. Banks

Polling

Opinion polls or surveys are used widely to study the behavior of individuals. Survey researchers select and ask questions concerning individuals' attitudes, behavior, personal characteristics, or other types of information that they can provide as respondents. These responses are counted, analyzed statistically, and evaluated or interpreted in different ways, depending on the types of questions asked.

Academicians and practitioners engaged in opinion polling have preferred to call themselves 'survey researchers,' in contrast to 'pollsters,' who carry out nonacademic polls to predict election results, or report the mass public's or others' opinions and attitudes in the news media, or who work as consultants doing proprietary research, such as market research, for political candidates, political parties (including polls for their presidents), or other clients (Converse 1987, Moore 1992, Jacobs and Shapiro 1995). Polls and surveys are synonymous, but their quality (validity and reliability) can vary widely.

1. 'Straw Polls'

The use of polling had expanded greatly worldwide by the late twentieth century. Its origin and impetus occurred in the United States beginning with election 'straw polls' in the nineteenth century, and developing into modern-day political polling and academic surveys used widely to study social, economic, and political phenomena. Polling was arguably a major part of the rise of quantification in society consistent with Max Weber's writing about the role of calculation and rationality in social life (Herbst1993). The earliest straw poll counterparts of modern opinion surveys

occurred in the 1824 presidential election, and the use of different types of these in politics and journalism increased until the emergence of more consciously scientific polling in the 1930s (Smith 1990).

These polls took many different forms, including oral counts, handraising, paper ballots, and writing preferences in 'poll books.' The participants were 'self-selected' in ways that were likely to produce a problem known as 'selection bias,' in which the results of these polls would be influenced by the characteristics of those who were able to participate, which might be different from the electorate or population to which the polls' results would be generalized. Reporters and partisans nonetheless treated these polls as indications of candidates' election chances.

2. *'Scientific Polls'*

The problems associated with the overall representativeness and accuracy of such polls became clear publicly, when the accuracy of the *Literary Digest*'s paper ballot poll, which had predicted the winners in US presidential elections from 1920 to 1932, was called into question by George Gallup and other new and more scientific-minded pollsters, statisticians, and social scientists. Gallup was the most widely known, if not the most important figure (along with Elmo Roper, Archibald Crossley, Paul F. Lazarsfeld, and Hadley Cantril) in 'scientific polling,' and approached polling with a democratic philosophy (Gallup and Rae 1940). While the first national Gallup election poll, done through what became standard in-person interviews did not predict the margin of President Franklin Roosevelt's 1936 election victory over Alfred Landon accurately, it triumphed over the *Literary Digest* in predicting the winner.

The Gallup poll represented the actual electorate far better than did the *Digest*, confirming the superiority of scientific sampling methods. In principle, initially, these methods allowed each potentially eligible adult voter (or eligible respondent, given the research objectives of a particular poll) an equal probability of selection, which was not the case in the *Digest* or other straw polls that allowed self-selection (Squire 1988). The early scientific polls involved sending interviewers to randomly-selected locations and then allowed interviewers largely to select their respondents to fill 'quotas' based on population characteristics to be represented in the poll. This led to selection biases.

These and other problems in the early polls, including ceasing to interview weeks before an election, were debated widely among pollsters and social scientists when the polls predicted incorrectly that Thomas Dewey would defeat President Harry Truman in the 1948 presidential election (Mosteller et al. 1949). Gallup and other pollsters improved their sampling methods, and the pollsters and academic researchers, who were beginning to engage in large-scale public opinion and voting studies at this time, engaged in a concerted and continuing effort to improve and defend their research (see, e.g., Sheatsley and Mitofsky 1992, Mitofsky 1998).

3. *Academic Research*

University-based survey research had its beginnings in 1939–40 with sociologist Paul F. Lazarsfeld's Office of Radio Research, later renamed the Bureau of Applied Social Research. This history is described by Rossi (1959), Natchez (1985), and especially Converse (1987). Lazarsfeld and his colleagues in *The People's Choice* (Lazarsfeld et al. 1944) did the first sophisticated survey study of presidential elections in 1940 in Erie County, Ohio. This was followed by *Voting* (Berelson et al. 1954), a study of the 1948 election in Elmira, New York. Two of the largest and most widely known research centers that developed the capacity to do long-term national surveys are the National Opinion Research Center (NORC), which was established in 1941 at the University of Denver and moved to the University of Chicago in 1947, and the University of Michigan's Survey Research Center (SRC, Institute for Social Research (ISR)), established in 1946. With the publication of *The Voter Decides* (Campbell et al. 1954), the funding and locus for national election surveys shifted to Michigan's SRC, which led to the publication of *The American Voter* (Campbell et al. 1960) and the continuing series of American National Election Studies and studies based upon them (Nie et al. 1976/1979, Miller and Shanks 1996).

4. *Further Expansion*

Political polling and market and academic survey research increased after 1948. The early and subsequent public pollsters who provided their survey results to the mass media and other subscribers (for example, Louis Harris and Daniel Yankelovich, among others) had used their political polling as advertising to generate business for their proprietary market and other types of research. Market researchers and other pollsters expanded their work into polling and consulting for political parties and candidates (Moore 1992). Polling experienced its greatest expansion beginning in the 1970s, when it was found that sufficiently representative and accurate polling could be done by telephone (based on experiments comparing polls conducted by phone with those done in-person), eventually leading all the major pollsters to do most of their polling by phone. The most important national news media, which had relied exclusively on polling by public pollsters, began sponsoring and conducting their own polls, establishing their own,

often joint, polling operations or making arrangements to contract-out polls in which they controlled the content and reporting of results (Gollin 1980, Mann and Orren 1992).

5. *Methodology and Technology*

While social scientific polling depends on an understanding and use of scientific sampling to draw inferences about a target population (e.g., national or local, the entire adult population or particular subpopulations, ordinary citizens or elites) polls can be conducted in different ways depending on the research goal, the kind of survey questionnaire to be administered, and the available resources (money and time). In-person national polling (typically randomly sampling geographic locations, then households in these locations, and then random sampling within households) is the most expensive and time-consuming because of travel costs, though these can involve longer surveys and the use of visual aids and secret ballots for responses to sensitive questions. In contrast, election 'exit polls' that are used to project election results and provide the basis for more in-depth statistical analysis of voters' behavior, are short and efficient surveys, based on random samples of voters exiting voting places. Exit polling has occasionally been controversial due to the 'early' reporting of results that might affect the behavior of those who vote later in the day.

Telephone polling (typically using Random Digit Dialing, RDD) is less costly, as it eliminates travel costs and is easier to monitor and validate than in-person interviews; however, it excludes households without telephones, and it has higher rates of nonresponse due to households that are not reached (no answers, busy signals, answering machines, and refusals to participate); it is easier to make contact with respondents and establish rapport through in-person contact. Mail surveys are the least expensive but produce the lowest response rates unless extensive follow-ups are made (Dillman 2000). Low response rates, with their increasing potential for selection or non-response bias, has emerged as the most pressing professional and methodological problem currently facing survey researchers as the result of the proliferation of telephone surveys, telemarketing, and solicitations that interrupt and impose upon people's time (Groves and Couper 1998).

Problems of survey costs and response rates have led surveyors to turn to better technology. The use of computerized RDD and computer-assisted telephone interviewing (CATI) was a major development in improving surveys (making complex surveys, including experimental designs, easier to administer, and eliminating interviewer and data processing errors) and expanding their use. The use of computer-assisted personal interviewing (CAPI) followed, including the use computers for self-administered surveys allowing for greater assurance of confidentiality for sensitive questions. Recently, there have been increasing efforts to expand sampling and polling through the Internet. While the Internet provides a way of increasing response rates through ease of contact without disruption, for potential respondents whose electronic addresses have been obtained, and through ease in administering surveys, the samples obtained may not be sufficiently representative of the target population, including those without Internet access, or who refuse to make their electronic addresses available for polling (thus echoing the limitation of the straw polls).

6. *Sources of Error*

Beyond statistical sampling error and nonresponse bias, there are other sources of error in surveys that are not easily quantified. The responses to survey questions and measurement of opinions and behavior can be affected by how questions are worded, what response categories are offered (e.g., whether a middle category or a don't know or no opinion response is offered or allowed), and whether questions should have fixed or forced choices, or whether questions should be asked and responses recorded as open-ended questions in which respondents can offer any answer they choose. Further, there may be 'context effects' that are produced by the order in which questions are asked, so that the way respondents answer questions may be influenced by what they had thought about and responded to in previous questions (Schuman and Presser 1981, Asher 1998).

A major source of error can occur depending on how research problems are formulated or specified, as when researchers assume respondents have familiarity with the object about which they are asked to offer an opinion or response. Care needs to be exercised in drawing inferences about actual behavior from survey measures of opinion, and self-reports of future—or even past—behavior. Reported voting in an election after the fact may not match actual behavior based upon data on 'validated' votes. Difficulties may also arise in how poll results are reported, such as when journalists or researchers themselves do not report enough information about their survey data and methods to allow their audiences or readers to evaluate their results (Cantril 1991, Traugott and Lavrakis 2000)

7. *Contributions to Knowledge*

The uses of surveys, how to improve polling methods, and how survey researchers should practice and report about their craft are matters that have been the

concern of the professional associations and journals established by those actively involved in the field of survey research. These include the American Association for Public Opinion Research and its journal, *Public Opinion Quarterly*; the World Association for Public Opinion Research and its journal, *International Journal of Public Opinion Research*; the Council of American Survey Research Organizations; and the National Council on Public Polls. In addition, political science, sociology, and other social science journals also publish survey studies widely.

The early voting and public opinion studies by Lazarsfeld and others were highly sensitive to methodological issues in survey research, and this led to the development and use of panel surveys (the interviewing of the same respondents over two or more points in time) and complex research designs. In addition to contributing to a large body of knowledge about the influences on voting, survey research contributed from its inception to the understanding of other forms of social and political attitudes and behavior, ranging from Samuel Stouffer's studies of *The American Soldier* (Stouffer et al. 1965 [1949]) and attitudes toward communism (Stouffer 1955), to studies of long-term trends in public opinion which utilize all the available polling data since the 1930s (Page and Shapiro 1992, Mayer 1993).

While academic researchers have made substantial use of data collected by academic survey organizations (especially the NORC General Social Surveys conducted since 1972, and the American National Election Studies covering every presidential and congressional election since 1952), increasingly they have used data made available by the public and mass media pollsters. The major archives for these data are the Roper Center for Public Opinion Research (University of Connecticut), the Inter-university Consortium for Political and Social Research (ISR, University of Michigan), and the Louis Harris survey and state poll archive at the Institute for Research in Social Science (University of North Carolina). These archives and the many polling organizations nationwide have also established and expanded access to their survey data through websites on the Internet.

See also: Democracy; Opinion Formation; Political Representation; Public Opinion: Political Aspects

Bibliography

Asher H 1998 *Polling and the Public: What Every Citizen Should Know*, 4th edn. Congressional Quarterly, Washington, DC

Berelson B R, Lazarsfeld P F, McPhee W N 1954 *Voting: A Study of Opinion Formation in a Presidential Campaign*. University of Chicago Press, Chicago, IL

Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. University of Chicago Press, Chicago, IL

Campbell A, Gurin G, Miller W E 1954 *The Voter Decides*. Row, Peterson, White Plains, NY

Cantril A H 1991 *The Opinion Connection: Polling, Politics, and the Press*. Congressional Quarterly, Washington, DC

Converse J M 1987 *Survey Research in the United States: Roots and Emergence, 1890–1960*. University of California Press, Berkeley, CA

Dillman D 2000 *Mail and Internet Surveys: The Tailored Design Method*, 2nd edn. Wiley, New York

Gallup G, Rae S F 1940 *The Pulse of Democracy*. Simon & Schuster, New York

Gollin A 1980 Exploiting the liaison between polling and the press. *Public Opinion Quarterly* **44**: 445–61

Groves R M, Couper M P 1998 *Nonresponse in Household Interview Surveys*. Wiley, New York

Herbst S 1993 *Numbered Voices: How Opinion Polling has Shaped American Politics*. University of Chicago Press, Chicago, IL

Jacobs L R, Shapiro R Y 1995 The rise of presidential polling: The Nixon White House in historical perspective. *Public Opinion Quarterly* **59**: 163–95

Lazarsfeld P F, Berelson B, Gaudet H 1944 *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press, New York

Mann T, Orren G (eds.) 1992 *Media Polls in American Politics*. Brookings Institution, Washington, DC

Mayer W G 1993 *The Changing American Mind: How and Why American Public Opinion Changed Between 1960 and 1988*. University of Michigan Press, Ann Arbor, MI

Miller W E, Shanks J M 1996 *The New American Voter*. Harvard University Press, Cambridge, MA

Mitofsky W J 1998 Was 1996 a worse year for polls than 1948? *Public Opinion Quarterly* **62**: 230–49

Moore D W 1992 *The Superpollsters: How They Measure and Manipulate Public Opinion in America*. Four Walls Eight Windows, New York

Mosteller F, Hyman H, McCarthy P J, Marks, E S, Truman D B 1949 *The Pre-election Polls of 1948*. Social Science Research Council, New York

Natchez P B 1985 *Images of Voting: Visions of Democracy*. Basic Books, New York

Nie N H, Verba S, Petrocik J R 1976/1979 *The Changing American Voter* (enlarged edn.). Harvard University Press, Cambridge, MA

Page B I, Shapiro R Y 1992 *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. University of Chicago Press, Chicago, IL

Rossi P H 1959 Four landmarks in voting research. In: Burdick E, Brodbeck A J (eds.) *American Voting Behavior*. The Free Press, Glencoe, IL

Schuman H, Presser S 1981 *Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording, and Context*. Academic Press, New York

Sheatsley P B, Mitofsky W J (eds.) 1992 *A Meeting Place: The History of The American Association for Public Opinion Research*. American Association for Public Opinion Research, Ann Arbor, MI

Smith T W 1990 The first straw: A study of the origins of election polls. *Public Opinion Quarterly* **54**: 21–36

Squire P 1988 Why the 1936 *Literary Digest* poll failed. *Public Opinion Quarterly* **52**: 125–33

Stouffer S A 1955 *Communism, Conformity, and Civil Liberties: A Cross-section of the Nation Speaks its Mind*. Doubleday, Garden City, NY

Stouffer S A, Suchman E A, DeVinney L C, Star S A, Williams R M Jr 1965 [1949] *The American Soldier: Adjustments During Army Life*. Wiley, New York, Vol. 1
Traugott M W, Lavrakas P J 2000 *The Voter's Guide to Election Polls*, 2nd edn. Chatham House, New York

R. Y. Shapiro

Polynesia and Micronesia: Sociocultural Aspects

The two geographical areas commonly referred to as 'Polynesia' and 'Micronesia' consist of a scattering of relatively small and widely disseminated islands across the central, south, and northwestern Pacific Ocean. The two areas are complementary to Melanesia, the third area customarily recognized in the insular Pacific. Together, the three areas are commonly referred to as the 'Pacific Islands' or 'Oceania,' although these categories, particularly the latter, occasionally include Australia and Tasmania, and occasionally the larger islands of Southeast Asia. While some geographical, social, cultural, linguistic, and historical generalizations can be made about each of them, Polynesia and Micronesia are meaningful entities beyond simple areal demarcations only as a result of historical contingencies, in which the history of Western colonialism in the Pacific figures prominently.

1. *General Identification*

The terms 'Polynesia' and 'Micronesia' are convenient labels for geographical areas whose social, cultural, archeological, historical, political, and linguistic significance is somewhat arbitrary. In modern-day usage, the term 'Polynesia' refers to all island and island groups falling within a large triangular area whose apexes are New Zealand to the south, Hawaii to the north, and Rapanui (Easter Island) to the west. The largest islands and island groups of the region are Tonga, Samoa, Rarotonga and the Southern Cooks, Tahiti and the Society Islands, the Marquesas, and Hawaii, in addition to the much larger New Zealand. Politically or culturally notable smaller islands include Niue, Wallis and Futuna, Tuvalu, Tokelau, the Northern Cooks, Rapanui, Pitcairn Island, and the Tuamotus, Austral, and Gambier Islands.

In addition, a geographically heterogeneous group of about 18 islands and sections of islands are commonly identified as 'Polynesian Outliers' societies because they are located outside of this triangular area: in Melanesia to the west (e.g., Takuu off Bougainville Island, Sikaiana and Rennell in the Solomon Islands, Mae and Mele in Vanuatu) and, in the case of two islands (Nukuoro and Kapingamarangi), in Micronesia to the northwest of the Polynesian Triangle.

Micronesia forms a wide arc of small islands spanning from the western region of Insular Southeast Asia to the Central Pacific, comprising the following island groups: Palau and outlying islands, Guam and the Mariana Islands, Caroline Islands, Marshall Islands, Gilbert Islands, and two isolated islands, Nauru and Banaba (Ocean Island). Contemporary political considerations may include island groups of the Central Pacific like the Phoenix and Line Islands; although geographically located in the Polynesian Triangle, these islands, which were not permanently inhabited until the modern age, are governed by Kiribati, the modern state based in the Gilbert Islands. With the exception of Guam, the islands of Micronesia are small, although some of the largest atolls in the world are found in the region (e.g., Kwajalein in the Marshalls).

Islands of many different types are found in Polynesia and Micronesia: a few continental islands (e.g., Guam); volcanic structures, some of which can be substantial in size (e.g., Hawaii); raised coral islands of different types (e.g., Tonga); atolls (e.g., Kiribati); and combinations of these.

2. *The Historical Contingency of Labels*

The societies of each of the two regions share a number of characteristics with one another, although it is impossible to arrive at a list of necessary and sufficient conditions that would determine whether a given society should be identified as Polynesian or Micronesian, or as something else. The reason for this is twofold. First, commonalities among the societies of both areas are balanced by important patterns of variation (to the extent that no feature of social organization, for example, will be found in all societies in question). Second, many societies in other areas of Oceania display sociocultural characteristics identical to those prevalent among Polynesian or Micronesian societies. Perhaps the least controversial criterion for 'Polynesian-ness' is language: all languages spoken natively in the islands are more closely related to one another than to any other language. Indeed, the reason for identifying Outlier societies as Polynesian is primarily linguistic: all Outlier communities speak languages that are most closely related to the languages spoken on islands of the Polynesian Triangle. In terms of social organization and culture, however, Outlier communities vary widely, from bearing considerable similarity to the rest of Polynesia (e.g., Tikopia) to having much more in common with their more immediate non-Polynesian neighbors (e.g., the Polynesian-speaking villages of Ouvea, Loyalty Islands).

Yet even this criterion is not devoid of caveats: on the boundary between Melanesian Fiji and Polynesian Tonga, historical linguistic evidence suggests a con-

tinuum of gradual linguistic differentiation rather than a clean break. Furthermore, linguistic factors are of little use as determiners of what Micronesia includes and what it does not: all languages spoken in the area are related to one another, but some (e.g., Chamorro of the Marianas and Palauan) are historically closer to languages spoken outside of Micronesia (e.g., in the Philippines) than to the other languages of Micronesia.

Two additional issues further complicate the problem. First, many Polynesian and Micronesian societies are increasingly diasporic, and significant communities of Samoans, Wallisians, Cook Islanders, Caroline Islanders, for example, are well established in New Zealand, Australia, the United States, and the metropolitan centers of Oceania. In some cases, emigrant communities are much more populous than the island-based communities. Second, the characterization of some areas of the region as 'Polynesian' or 'Micronesian' is historically well motivated but, from a contemporary perspective, is part and parcel of acrid political debates. Such is the case of Hawaii, New Zealand, and Guam, where the original Polynesian or Micronesian inhabitants, as well as recent migrants from other Pacific Islands, today form numerical and political minorities, albeit vocal ones in terms of identity politics. These observations highlight the problems associated with attributing human significance to the definitions of Polynesia and Micronesia.

To say that the regions are arbitrarily defined does not mean that their characterization is haphazard. Rather, the definition of sections of the globe as 'Polynesia' and 'Micronesia' is deeply embedded in a history of elaboration of certain differences and similarities and the obscuration of others to suit politically dominant agendas, of the kind that suffuses any characterization of the 'other.' In this case, this history is that of Western colonial hegemony and of the intellectual endeavors that ran alongside it. When Enlightenment-era Europeans invented it, the term 'Polynesia' ('many islands') was applied to the entire Pacific region. In the course of the nineteenth century, it became more common to restrict the term to its current referent, and to contrast it with the newly coined 'Micronesia' ('small islands') and 'Melanesia' ('black islands'). As the etymology of the latter term indicates, race and its various associations figured prominently in the newly created finer distinctions. The term 'Polynesia' came to embody (mostly positive) images of noble fair-skinned otherness already prominent in the Enlightenment and early Romantic intellect, in contrast to Melanesia in particular, which Westerners deemed to be dominated by dark-skinned and uncivilized savagery (Smith 1985, Thomas 1989).

3. *Initial Human Settlement*

The prehistory of Polynesia can only be understood in the context of human movements in the Pacific region since about 3500 BP. Around that date, a wave of people appears to have traveled gradually from west to east from Southeast Asia, settling islands of Melanesia, some of which were already inhabited while others were not. The most salient archeological evidence for this conjecture is a lowly style of decorated pottery referred to as 'Lapita,' fragments of which are found in insular Melanesia and Western Polynesia, carbon-dated to 3500–2000 BP. Lapita pottery makers and users appear to have been accomplished long-distance sailors, fisherfolk, and agriculturalists, to have organized their communities in hierarchical fashion, and to have spoken languages ancestral to most languages of insular Melanesia and Polynesia.

They reached Fiji, Tonga, and Samoa around 3000 BP, an area in which, according to some researchers, migrations may have 'paused' for a while (while maintaining active contact both externally and with the West from where they had come). Such a 'pause' would have allowed a social, cultural, and linguistic distinctiveness to emerge that would eventually become what is now recognized as Polynesian distinctiveness.

From the Fiji–Western Polynesia area, early Polynesians settled the rest of Polynesia in the course of the following two millennia, finally reaching Hawaii in about 650 AD and New Zealand around 1000 AD (Bellwood 1979, Kirch 1996). There is no convincing evidence of any subsequent prehistorical human settlement in Polynesia other than the Lapita potters and their descendants, although they intermarried with their non-Lapita-making neighbors, forming as diverse a genetic pool as is found in any other part of the world.

Micronesia's prehistory is much more heterogeneous and complex than that of Polynesia, and less well understood. Evidence of human settlement in the Mariana Islands dates back to approximately 4000 BP. The archeological record indicates a Southeast Asian connection for this early population, as does relatively more recent evidence gathered on Palau and Yap. The rest of the Caroline Islands, the Marshall Islands, and the Gilbert Islands were settled by a northwestward back-migration from eastern Melanesia. Languages spoken in this region, commonly referred to as 'Nuclear Micronesian,' exhibit greater linguistic homogeneity than other Micronesian languages, suggesting a more focused history of prehistoric settlement.

4. *Contacts and Movements*

It is difficult to issue brief general statements about the history of Polynesia and Micronesia since colonial encroachment because of the diversity of experiences and heterogeneity of influences. European travelers began sailing through Polynesia and Micronesia early in the Age of Discovery, making contacts with Islanders that were sporadic except in Guam and the Marianas, which were situated strategically on the sea

road between the Philippines and the New World. Spain claimed the islands in the mid-sixteenth century, a move which was followed by the virtual decimation of the Islanders, through disease, slaughter, and forced resettlement.

Contacts between Westerners and Islanders in both Polynesia and Micronesia began earnestly at the end of the eighteenth century, and often had equally tragic consequences on the island populations. Enlightenment-era travelers, spurred on by complex and sometimes contradictory motivations, roamed Polynesia in particular, making contact with Islanders in a variety of fashions that ranged from the hostile to the very friendly. They were soon followed by Christian missionaries (who would eventually succeed in converting all of Polynesia and Micronesia), adventurers and traders, developers seeking to establish plantations and other large-scale ventures, whalers, and of course governments seeking to establish colonies and spheres of influence, often under the pretext of protecting their citizens. In Polynesia, the key players (big and small) in the nineteenth century were predominantly British and French, to which American and German interests added competition later in the century. In Micronesia, Spain continued to hold political and economic sway, until it was completely displaced at the end of the century by the United States and Germany.

A notable consequence of Westerners' increasing visibility in nineteenth century Polynesia was the consolidation of hitherto politically fragmented chiefdoms. In Tonga, Tahiti, and Hawaii (and to a lesser extent Samoa), ambitious and astute chieftains utilized the new tools that Westerners introduced to the islands (trading goods for firearms, protection for religious conversion) to defeat their contenders and impose their rule over entire islands and island groups, establishing themselves as sovereigns under the protection of Western powers. Of these kingdoms, only Tonga remains to this day.

By the dawn of the twentieth century, European powers had divided Polynesia and Micronesia (as well as Melanesia) into sphere of colonial domination and influence. The end of World War I somewhat reshuffled the distribution of these spheres, introducing Japan as a colonial power in Micronesia (except the Gilbert Islands, which by then had become a British colony). World War II had a profound and dramatic effect on all of Micronesia, which was the stage of some of the fiercest battles, as was Melanesia. In contrast, Polynesia (other than Pearl Harbor in Hawaii) remained in the background of the conflict.

5. *Political Re-emergence*

In tune with the worldwide decolonizing trends of the second half of the twentieth century, island nations began obtaining their independence from world powers, particularly Great Britain, which by then had also delegated some of its colonial authority to Australia and New Zealand. In Polynesia, for example, Western Samoa (now Samoa) was granted independence in 1962, and Tonga ceased to be a British protected state in 1970. Decolonization in island groups held by the United States and France (Micronesia and French Polynesia, respectively, in particular) was and still is a slower, more complex, and conflictual process. A notable factor in both cases is the use by both powers of islands for nuclear testing until well into the 1990s.

Today, France maintains a colonial presence in French Polynesia and Wallis and Futuna (both Polynesian groups), which is contested in the former but receives general local approbation in the latter. The United States signed Compacts of Free Association with the Federated States of Micronesia and the Marshall Islands in 1986, and with Palau in 1994, and since then these entities have been recognized as sovereign nations despite some continuing US security interests in the islands. American Samoa and Guam are unincorporated US territories in which the federal government continues to maintain control of many functions, in spite of some local opposition in the case of Guam. The Northern Mariana Islands is a Commonwealth with more control of local affairs. Rapanui is administered by Chile, despite some political tension, and tiny Pitcairn Island (Polynesian from a geographical point of view) by Britain. New Zealand holds special political ties with three now self-governing territories, namely Niue, Tokelau, and the Cooks. Hawaii, a state of the United States, and New Zealand are both postcolonial entities in which the politics of indigeneity are tense. These varied historical connections, as well as the different configurations in which domination, resistance, and everything in between are embedded, have contributed to the sociopolitical diversity of both regions.

The states and territories of Polynesia and Micronesia include some of the tiniest political entities of the world, such as Tuvalu (10,000 inhabitants, 26 km^2). Outside of New Zealand, most are heavily dependent on world powers, relying on foreign aid, migrations, and access to preferential markets to ensure at least a semblance of economic viability and, in some cases, political stability. The notable exception is Nauru, which is one of the richest countries of the world because of decades of phosphate mining.

6. *Society and Culture*

In precontact times, the societies of Polynesia and Micronesia displayed a certain degree of sociocultural commonality, as witnessed, for example, in a tendency to have stratified political systems and in the elaboration of certain symbolic complexes such as *tapu*

(roughly, 'religiously-based prohibition') and *mana* (roughly, 'efficacy of divine origin') in Polynesia. However, the recognition of patterns of commonality must always be qualified in at least two ways. First, one must be attentive to the dynamics through which certain characteristics have come to be identified as 'typically' Polynesian or Micronesian in spite of the fact that these characteristics are absent in some societies of the region. Second, 'precontact times' is a vague characterization at best, and all island societies experienced radical changes over the centuries, some of which can be identified archeologically, in the course of which 'typical' Polynesian or Micronesian characteristics appeared in and disappeared from particular societies.

A significant number of societies in both regions appear to have been stratified political entities, the more so in Polynesia than in Micronesia. Particularly on large islands and island groups, chiefdoms were organized around a leadership that derived power through a combination of genealogical ties to both sacred and profane entities that confirmed claims to power (ascription) and personal performance that hopefully confirmed these claims (achievement). In some of more stratified societies, persons were also ranked into categories ranging from the highest ranks, which were commonly sacralized, to the lowest ranks, whose members were often not considered human. Other societies, in contrast, displayed little stratification and emphasized egalitarianism and consensual decision-making. Political organization on some islands was centered on localized kinship units, particularly in Micronesia but also on the smaller islands of Polynesia. The rough correlation between the amount of resources to be produced and organized and the degree of social stratification, at least in Polynesia, led some early researchers to propose that one engenders the other: chiefs, as resource managers, increase in importance as resources increase in volume (Sahlins 1958). Few anthropologists today would give such determinative power to resource management (and indeed the empirical evidence does not support the proposal), although it certainly plays a role in the development and elaboration of social stratification.

Prevalent patterns throughout the area centralized the identification of a common founding family ancestor, creating in some societies a pyramid-shaped kinship structure. In Polynesia, the most important branch was the senior patrilineal branch, and members of other branches reckoned their relative status according to their relative distance from the main descent branch, measured through relative patrilineality and seniority. Micronesian societies are predominantly matrilineal. There, matrilineal clans have historic claims to particular pieces of land and are the basis for social organization. Rank is based on claims of founding settlement of particular pieces of land. Probably all Polynesian and Micronesian societies also offered the possibility of bilateral affiliation, whereby kinship could be based on either patrilineal or matrilineal principles, thus providing room for maneuvering for individual advantage. In Polynesia, there emerges an east–west contrast between societies in which gender and its associations constitute a pivotal organizational principle for kinship organization (as in Tahiti and Hawaii), and societies in which seniority plays a greater role (as in Samoa and Tonga). Adoption in its various guises was and is also prevalent throughout Polynesia and Micronesia.

Many aspects of precontact (or early contact) society and culture have reproduced over time and continue to characterize contemporary circumstances. However, any statement about contemporary Polynesian and Micronesian societies must take into consideration the enormous transformations that these societies have undergone in both bygone days and more recent times, introducing dynamics and agents that have been variously resisted, localized, and transformed. Two examples will suffice. The first is the recognition that many contemporary Polynesian and Micronesian societies are vigorously diasporic. While strong sociopolitical forces continue to ground society and culture in the 'islands of origin,' these forces are increasingly subject to contestation as second- and third-generation expatriates constantly negotiate their place across different identities and allegiances. The second illustration is the fact that Christianity, which is two centuries old in some area of Polynesia and Micronesia and much more recent in others, is almost universal and highly visible in both regions. In some parts of Polynesia, it has become one of the main backbones of hierarchy, exchange systems, and kinship, and is viewed as one of the most fundamental tenets of 'tradition.' In short, like all social groupings around the world, Polynesian and Micronesian societies are dynamic and adaptable entities, whose boundaries, core cultural values, and principles of social organization are subject to change, reassessment, and potential contestation.

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Cultural Assimilation; Melanesia: Sociocultural Aspects

Bibliography

Bellwood P 1979 *Man's Conquest of the Pacific: The Prehistory of Southeast Asia and Oceania*. Oxford University Press, New York

Denoon D (ed.) 1997 *The Cambridge History of the Pacific Islanders*. Cambridge University Press, Cambridge, UK

Howard A, Borofsky R (eds.) 1989 *Developments in Polynesian Ethnology*. University of Hawaii Press, Honolulu, HI

Kirch P V 1996 *The Lapita Peoples: Ancestors of the Oceanic World*. Basil Blackwell, Oxford, UK

Kirch P V 2000 *On the Road of the Winds: An Archaeological History of the Pacific Islands Before European Contact*. University of California Press, Berkeley, CA
Kiste R C, Marshall M (eds.) 1999 *American Anthropology in Micronesia: An Assessment*. University of Hawaii Press, Honolulu, HI
Lal B V, Fortune K (eds.) 2000 *The Pacific Islands: An Encyclopedia*. University of Hawaii Press, Honolulu, HI
Sahlins M 1958 *Social Stratification in Polynesia*. American Ethnological Society, University of Washington Press, Seattle, WA
Smith B 1985/1969 *European Vision and the South Pacific*. Yale University Press, New Haven, CT
Thomas N 1989 The forces of ethnology: origins and significance of the Melanesia/Polynesia division. *Current Anthropology* **30**: 27–41

N. Besnier

Popper, Karl Raimund (1902–94)

1. *Popper's Life*

Karl Raimund Popper was born in Vienna, Austria, on July 28, 1902. His father, Simon [Siegmund Carl] Popper (1856–1932), who had come from Bohemia to Vienna, was a well-established lawyer with his offices and family residence in the center of Vienna. Karl Popper's mother, Jenny Popper (neé Schiff) (1864–1938) came from a musical family and was a gifted pianist herself. Both Simon and Jenny Popper were Jews by origin, but converted later in their lives to Protestantism. Besides their son Karl, they had two older daughters, Dora (1893–1932) and Anna Lydia (1898–1975).

Karl Popper left high school (Franz Josef Gymnasium) at the age of 16 without a degree, moved out of his parents' home and attended lectures at the University of Vienna on various subjects, in particular mathematics, philosophy, physics, and psychology. Besides occasional jobs to earn his living (e.g., coaching American students), Popper was engaged in social activities. For some time, he worked as an assistant of Alfred Adler on neglected children. Hints in Popper's autobiography indicate that this connection (through which he got acquainted with Adler's 'individual psychology') contributed to Popper's later critique of psychoanalysis as a 'pseudoscience.' In his youth, Popper had a strong favor for democratic socialism. For a few months, he even considered himself a communist, from which he refrained, in 1919, after witnessing a shooting incident in which socialist workers were killed by police forces. This experience of a useless sacrifice of lives demonstrated to him the futility of ideologies. A second, more intellectual experience in 1919 was his taking interest in Eddington's test of Einstein's eclipse predictions carried out at an expedition of the Royal Astronomical Society, whose results gained widespread publicity at the time. During a total solar eclipse which was observed from West Africa and Brazil, a shift of apparent position of stars in the presence of the sun's gravitational field could be established, as predicted by the general theory of relativity. This coined Popper's later insistence that scientific theories develop by exposing them to possibly falsifying experience. Around that time, Popper visited a lecture by Einstein in Vienna.

The young Popper also tried manual work. From 1922 to 1924 he was an apprentice to a cabinet maker. At the same time, he passed the high school exams ('Matura') as an external participant and formally enrolled as a student at the University of Vienna. He also considered becoming a musician. After being disappointed with the sort of music cultivated in the Schönberg circle, he enrolled as a student of Church Music at the Vienna Academy of Music ('Konservatorium') and stayed there for one year. He was admitted upon submission of a fugue for organ in F sharp minor. Besides his scientific studies at the University of Vienna, in 1924 he qualified as a primary school teacher. Since no appropriate teacher's post was available at the time, Popper worked as a social worker with neglected children.

In 1925 Popper became a student at the Pedagogic Institute of the City of Vienna, where he stayed until 1927. This institute, which Popper was admitted to due to his status of a social worker, had been newly founded. It was linked to the University, though formally independent. In today's terms, its purpose was to scientifically accompany and evaluate the school reform movement. This school reform movement had been politically supported by the Austrian social democrats since 1919 and was in 1925 still very strong in Vienna. It was characterized by ideas of abolishing school learning as an accumulation of facts in favor of developing a more autonomous and participating attitude. The influence of these ideas on Popper should not be underestimated. His later criticism of what he calls the 'bucket theory of knowledge,' i.e., the inductive generation of knowledge by collecting facts, is related to pedagogic proposals on how learning proceeds or should proceed. Actually, Popper's first published papers were in this field and appeared in journals belonging to the school reform movement. The very first one *Über die Stellung des Lehrers zu Schule und Schüler* (*On the teacher's attitude towards school and students*, 1925) is very significant in this respect.

At the pedagogic institute, Popper met Josefine Henninger (1906–85), who was training as a teacher and who was to become his wife in 1930. (They had no children.) He also met the psychologist Karl Bühler and the philosopher Heinrich Gomperz, both of them

professors at the University. They were of crucial importance to Popper's intellectual development. It was through discussions with Gomperz and reading his works that Popper shifted away from any psychologistic approach to the philosophy of science. From Bühler he adopted the distinction of expressive, signaling, and descriptive functions of language, which still figured prominently in Popper's late metaphysical writings.

In 1928 Popper completed his Ph.D. in Psychology with Karl Bühler and Moritz Schlick as examiners. The subject of the thesis was *Zur Methodenlehre der Denkpsychologie* (*On the Methodology of the Psychology of Thinking*, unpublished), a work which combined Popper's interest in psychology with his methodological bias that he had finally shifted to during his time at the Pedagogic Institute. In 1929 Popper qualified as a teacher for secondary schools ('Hauptschulen'), and in 1930 he became a teacher in Vienna.

Since 1922 Popper had been in contact with members of the Vienna circle, a group of philosophers around Moritz Schlick who propagated what they called 'logical positivism,' a version of empiricism which made strong use of modern mathematical logic as a tool of philosophical analysis. Mathematical logic had already been developed by Gottlob Frege (1848–1925) and Bertrand Russell (1872–1970) decades ago, but had not been appropriately appreciated within the established academic philosophy. Though never a member of the circle, Popper attended its sessions occasionally and was in contact with several of its members, in particular Rudolf Carnap, Victor Kraft, and Herbert Feigl. Although he shared certain basic attitudes of the Vienna Circle, in particular the plea for a scientifically oriented philosophy, he was highly critical of other tenets such as the theory of meaning and the form of empiricism propagated. Motivated by Feigl, he put this criticism into a large manuscript in two volumes *Die beiden Grundprobleme der Erkenntnistheorie* (*The Two Basic Problems of Epistemology*, published in 1979, the second volume only as a fragment). A shorter version of the second volume of this manuscript was published in 1934 as *Logik der Forschung* (*Logic of Scientific Discovery*), which was to become his main work.

The book was well received. Reviews appeared in many journals. Even Einstein wrote Popper a letter after the book was sent to him through intermediates. In 1935 and 1936 Popper received invitations to England, where he lectured at Bedford College and Imperial College, London, at the London School of Economics and at Cambridge. He made a lot of contacts. Of most importance for his later career were those with Ernst Gombrich and Fritz Hayek. However, attempts to find a position in England did not materialize, and in 1937 he emigrated with his wife to New Zealand, accepting an offer of a lectureship at Canterbury University College.

During the time in exile Popper intensified his interest in social and political philosophy and in the philosophy of history, besides continuing his research in logic and philosophy of science. From that time on, he published mainly in English. He wrote a series of articles *The Poverty of Historicism* (Popper 1957a originally 1944–5); and, more importantly, his two-volume classic *The Open Society and Its Enemies* (Popper 1945). For both works he had to struggle to find a publisher. Due to the *Open Society*, he acquired his initial postwar reputation and, in particular, his post, from which he carried out all subsequent work. Through the intervention of Hayek who, together with Gombrich, had made the publication of the *Open Society* with the publisher Routledge possible, he was offered a readership in 'Logic and Scientific Method' at the London School of Economics in 1945, which Popper took up early in 1946, shortly after publication of the *Open Society*. In 1949 he received a professorship of the same title at the London School of Economics, which he held until his retirement in 1969.

During the first five years of his stay in Britain, he was sort of a 'star.' The *Open Society* made him a quite popular theoretician of political liberalism. It has often been said that this book filled some intellectual vacuum in the postwar period. Popper was invited to many places and, at the same time, had a substantial audience at the London School of Economics. In the 1950s, he shifted more to working from home with less institutional involvement. At the same time, his research interests moved away from political and social philosophy and back towards the subject of the *Logik der Forschung*, i.e., the methodology of the natural sciences. However, the first English edition of this book appeared only in 1959, with a lot of material added compared with the German original. Many other books and plenty of papers followed. Up to his death, Popper produced an enormous scientific output and had a vast correspondence. From the mid-1960s on, Popper received a great number of prizes and honorary degrees in many countries, including Knighthood and Membership of the Royal Society in Britain. He died on 17 September 1994 in London. His ashes were buried in Vienna, where his wife had died in 1985.

2. *Popper's Work*

Popper's work covers nearly all areas of theoretical philosophy. His overall profile is that of a great philosophical classic. However, most important are his contributions to the methodology of science, where he established falsificationism as one of the basic orientations in the modern philosophy of science.

The basic claims of falsificationism are put forward in the *Logik der Forschung* (1934, official year given in

the book: 1935). Most of Popper's later methodological contributions can be viewed as elaborations or amendments of what was already stated in his first book. Even Popper himself treated the *Logik der Forschung* like a classical text, with all changes in the first English translation (*The Logic of Scientific Discovery* 1959) as well as in the later 10 German and 10 English editions, as a system of footnotes and appendices to the 1934 edition.

Falsificationism can be described as a 'why not' methodology in contradistinction to a 'why' methodology with respect to scientific theories. According to Popper, a scientific theory is posed without any prior justification. It gains its confirmation, or 'corroboration,' which is Popper's preferred term in later writings, not by experimental evidence demonstrating *why* it should hold, but by the fact that all attempts at showing *why* it does *not* hold have failed. The main activity of an experimental researcher is trying to refute or '*falsify*' a scientific theory, rather than to justify or '*verify*' it. Theoreticians who design theories should do this in such a way that they can be refuted most easily. In no case must a theory be immunized against possible falsification, e.g., by the introduction of *ad hoc* hypotheses. We rely on a theory only because we have not managed to falsify it (so far), not because we have any positive or direct reasons in its favor.

His own theorizing, and especially the idea of falisificationism, is considered by Popper as proposing 'methodological rules.' These methodological rules have to prove fruitful in the understanding of scientific research and, particularly, of scientific progress. The main argument in favor of his falsificationist proposal is that the converse proposal, which is called 'verificationist,' is bound to fail. Verificationism, which underlies the 'inductivist' methodology, claims that a scientific theory is supported by a collection of verifying instances. However, as has already been established by David Hume (1711–76), the inductive reasoning from a finite set of observations to a general statement is not logically valid if the domain of reasoning is infinite or at least indefinite. Furthermore, any attempt at creating a special nondeductive form of reasoning (an 'inductive logic') can only be justified by means of induction itself, thus ending up in a vicious circle. To quote Popper's famous example, no observation of white swans can verify the general statement that all swans are white.

Popper's falsificationist methodology rests on the assumption that unlike verification, falsification of a general statement by experimental or observational means is logically without any problem. The statement 'All swans are white' can be falsified by the observation of a single black swan. The logic involved here is just ordinary deductive logic, due to which the falsity of an instance derived from a general hypothesis falsifies the hypothesis itself. Seemingly singular statements such as 'The planet Pluto has a moon,' which at first glance seem to be verifiable but not falsifiable, are viewed by Popper not in isolation but as embedded in a general theory and therefore are no counterexample to what he calls the 'asymmetry between verification and falsification.'

More complicated is the example of statistical theories where definite falsification is not possible. Here Popper applied as an additional methodological rule that what is extremely unlikely can be neglected as something which cannot be reproduced deliberately. This allows Popper to cope with the standard way of rejecting statistical theories by means of a critical test (usually defined by stipulating a level of significance).

In traditional methodology, particularly that of the Vienna Circle, inductivism was tied to empiricism and guaranteed the connection between theoretical statements and their empirical basis. In falsificationism, the empirical significance of theoretical statements is kept by relating them to basic statements which possibly falsify them. However, these basic statements are no longer purely observational in some epistemological sense, but are themselves theory-laden from the very beginning. Even elementary perceptual evidence has to rely on theoretical assumptions. Thus Popper rejects any semantic division between theoretical and observational statements and, especially, the claim that a particular sort of certainty comes with the latter. According to Popper, the validity of basic statements can only be fixed in the course of the testing of a theory, and, in the end, this fixing is a matter of convention. This variant of conventionalism, which enters Popper's approach in addition to the conventionalism inherent in the choice of methodological rules, has repeatedly been the subject of criticism.

For Popper the distinction between a theory which can be falsified by means of basic statements and one which is not falsifiable serves as a demarcation criterion for empirical science vs. metaphysics. However, metaphysics, though not falsifiable, is not meaningless. Metaphysical statements may serve useful purposes, even within science. For example, the principle of causality ('any event can be causally explained'), which cannot be falsified and is therefore metaphysical, can be viewed as the useful methodological rule never to abandon the search for laws which explain observed facts. This view of metaphysics is less rigid than that of the Vienna Circle, who considered metaphysics to be meaningless throughout. It allows Popper to sharply criticize certain nonfalsifiable theories as 'pseudosciences' (such as psychoanalysis) but nevertheless to make sense of certain traditional philosophical claims or to appreciate the fruitful role of metaphysical research programs in the history of physics ('A metaphysical epilogue,' in *Quantum Theory and the Schism in Physics* 1956, published 1982b).

In his later works, Popper extended the methodological approaches of the *Logik der Forschung* towards a general account of the dynamics of scientific theories ('Truth, rationality, and the growth of scientific knowledge,' in *Conjectures and Refutations* 1963).

According to this view, theoretical progress is made by successive critique and revision of existing theories, which is governed by the idea of objective truth. Scientific progress works by replacing a falsified theory with a new one which, in the ideal case, is more encompassing than the previous one in that, while coping with the new results, it also covers the unfalsified instances of the old theory. This appeared to be the case with classical and relativistic gravitational theories in the presence of the results of the 1919 solar eclipse observations, which were so impressive to the young Popper. His later claim is that the new and more encompassing theory is nearer to the truth than the older and more special one, even if it should at some point become falsified itself. Popper has tried to characterize the distance of a theory from truth, its 'truthlikeness' or 'verisimilitude,' in numeric terms. Considerable work has been carried out in recent decades to develop improved measures.

Within the revision of scientific theories, Popper strongly distinguishes between hypotheses being tested and background knowledge which, though testable in principle, is not being questioned at the same time. This distinction makes the falsification of parts of a theory possible, and not only the falsification of a theory as a whole. Therefore Popper pleads to formulate theories in a modular way such that its parts can be tested independently of each other. He is opposed to 'holistic' views of theories, according to which, in a single test, the whole theory (Pierre Duhem) or, even more radically, all our theoretical knowledge including logic and mathematics (W. V. O. Quine) is at stake. Popper defended his model of scientific progress as a change of scientific hypotheses by means of falsification against historic approaches which denied the rational character of scientific change. So he was strongly opposed to Thomas S. Kuhn's account of scientific revolutions as irrational changes without the concept of 'progress' being applicable. Similarly, he argued with Imre Lakatos (1922–74) and Paul K. Feyerabend (1924–94), who started as Popperians, but later on made changes in falsificationist methodology by weakening and relativizing Popper's strict idea of progress by critical testing.

In *Evolution and the Tree of Knowledge* (1961, published 1972 in *Objective Knowledge*), Popper embedded the idea of theory change by falsification and revision within a general theory of the evolution of knowledge. Here 'knowledge' is viewed as a form of adaptation of an organism to the environment. Popper's schema of progress from problems via falsification to new problems is interpreted as the evolutionary transition from organisms via natural selection to organisms which are better adapted. By viewing all life as problem solving (*All Life is Problem Solving* 1994c), Popper claimed to combine theories of biological evolution with theories of scientific progress, relating his approach to evolutionary epistemology in the sense of D. Campbell. However, it must be emphasized that, unlike most evolutionary epistemologists, Popper was not a naturalist. He insisted that the cultural-scientific evolution is rational and based on conscious and deliberate criticism, not on natural selection governed by chance.

In his later work, Popper also replaced the narrow concept of 'falsification' with a broader concept of 'criticism' or 'critical examination.' This broader concept not only applies to scientific theories which are formulated in a logical language but to philosophical claims which are metaphysical. It involves tests of consistency and the investigation of how far the theory in question is able to solve particular problems. This orientation, with critical examination as its cornerstone, was called 'critical rationalism' by Popper and has since become the standard denomination of the Popper school. The main examples of Popper's dealings with metaphysical problems are his treatment of determinism and his philosophy of mind. Quite independent of problems of quantum mechanics, which often serve as arguments against determinism, in *The Open Universe*: *An Argument for Indeterminism* (1956, published 1982a), he developed an indeterministic world view. In 'Of clouds and clocks' (1966, reprinted in *Objective Knowledge* 1972), he extended this view with arguments for the openness of the future for free action. In the mind–body debate, Popper defended a mind–body dualism ('Language and the body–mind Problem,' 1953, reprinted in *Conjectures and Refutations* 1963), which he later extended to a model of three worlds of (a) physical entities, (b) mental states, and (c) objective contents of thought, which causally interact. This approach gained much attention through his discussions with J. R. Eccles which, together with their separate contributions, are presented in their joint book *The Self and its Brain* (1977).

Epistemologically, Popper was an outspoken objectivist and realist. On the Popperian view, our theories are not just methodological instruments or means of acquiring (some degree of) certainty about something, but attempts at explaining the real world. As reality is described by means of true sentences, this is another way of claiming that the goal of science is truth. Part of Popper's objectivism and realism is his rejection of any subjective interpretation of probability. In particular, he was strongly opposed to Bayesianism and its view that scientific progress is just changing the probabilistic degree of belief in certain propositions, which means that scientific dynamics is probability dynamics (*Probability Magic or Knowledge out of Ignorance* 1957b). Both the probability of hypotheses, for which Popper coined the term 'logical probability,' and statistical probability, which may be part of a hypothesis, are objective probabilities reflecting properties of what they are attributed to. Logical probabilities are inversely proportional to the content of a theory, the less probable theory having a higher content. Statistical probabilities which, in the *Logik*

der Forschung, were interpreted as relative frequencies, were later viewed as objective 'propensities' of experimental designs to develop certain outcomes (*Realism and the Aim of Science*, 1956, published 1983). This propensity interpretation of probability claims to be a solution of the problem of assigning probabilities to singular events, rather than only to events as members of a given sequence or set of related events.

Popper's attitude against probabilities as degrees of belief does not contradict the fact that he attempted to measure the degree of corroboration of a theory in probabilistic terms. However, first of all, Popper's degree of corroboration, though relying on probabilities, is not a probability itself, i.e., does not obey the laws of the calculus of probability. Second, the degree of corroboration is intended as a strictly objective measure which reflects the way the hypothesis in question was subjected to falsification, and is definitely not meant as a degree of rational belief in a hypothesis. According to Popper, epistemology is the methodological treatment of objective scientific contents and not a theory of the knowing subject ('Epistemology without a knowing subject,' 1967, reprinted in *Objective Knowledge* 1972).

Popper's social and political philosophy which was developed during his emigration in New Zealand and published in *The Poverty of Historicism* (1944–5) and *The Open Society and its Enemies* (1945) can, to some extent, be viewed as an application of falsificationism to the area of social interaction. By 'historicism' Popper denoted the claim that the methodology in the historical and social sciences differs fundamentally from those in the natural sciences but, nevertheless, allows one to formulate objective laws of historical development, which cannot be changed by subjective action. The standard example is Marxist philosophy of history (historical materialism). In contradistinction to that, Popper put forward the idea of 'piecemeal social engineering,' i.e., of a development of social institutions in small steps which can be revised and whose potential damage can be controlled. According to Popper, piecemeal social engineering can be combined with a falsificationist form of social research, which is based on trial and error as it generates models for social development. For areas of historical and sociological research which are interested in the explanation of single events rather than the testing of general laws, Popper proposed the procedure of a 'logic of the situation' which describes the acting of individuals under the hypothesis of rational purpose-oriented behavior within a certain social situation. In this context he defends the autonomy of sociology against psychologism, as this situation is described in objective terms by reference to social institutions, etc. Since 1961, Popper defended this account of social theorizing against representatives of the neo-Marxist Frankfurt school in Germany, in particular, Theodor W. Adorno and Jürgen Habermas, in a dispute which became known as the 'positivist dispute in German sociology,' and which made Popper well known to a wider audience in the German-speaking countries.

The *Open Society* extends the critique of historicism to a critique of the philosophical theories of state and society by Plato, Hegel, and Marx, which Popper considers to be philosophical precursors of political totalitarianism. According to Popper, they advocate the idea of a closed society, which is a collectivistic and organismic whole, without the distinction between natural and cultural laws. In an open society, which Popper strongly favors, individuals are personally responsible for their decisions and can therefore be critical to social regulations. Popper argues that the traditional question 'who should rule' ought to be replaced by the question of how to minimize the risk of bad rulers and how to get rid of them once they are in power. For Popper, the long-standing philosophical discussion about the best rulers belongs to justificationist thinking that has to be abandoned in favor of critical thinking, just as in general methodology. The utilitarian principle of maximization of happiness should give way to the idea of minimizing suffering.

Although there appears to be a lot of moralizing both in the *Open Society* and in many short papers and interviews on questions of morals and politics given towards the end of his life, this should not be understood as fragments of an ethical theory. For Popper, there is no such thing as ethics, because ethical sentences are not propositions in the logical sense that one could argue for or that one could reject. However, there are moral actions. These are actions which are carried out by someone who is aware of their consequences. Since the consequences of actions can be critically examined in the sense of critical rationalism, moral decisions are not irrational.

3. *Relevance and Impact of Popper's Philosophy*

Popper's methodological program was revolutionary as it reverted the established conceptual order between certain fundamental notions. He put 'negative' concepts such as 'falsification' and 'rejection' first, and made 'positive' concepts such as 'acceptance' or 'corroboration' dependent on them. Whereas according to the traditional philosophical view of science established ever since Bacon, acceptance of a theory is based on the direct generalization from observation, according to Popper, it can only rely on failed rejection. It is therefore correct to speak of Popper's 'negative methodology.' As Popper tried to show in his later writings, this is not just a program for the natural sciences, but applies to any area of scientific reasoning and even to metaphysics, where negative concepts such as that of 'criticism' come first. The credit we give to a conjecture is based on the fact that refuting it by criticism has been unsuccessful. This program is even the basis of Popper's political phil-

osophy. The negative question of how to restrain political power and to remove bad rulers from power is of primary concern compared with the positive one of whom to give power. Popper's negative methodology is a unified view of rationality under the header 'critical rationalism.'

This comprehensive perspective of theoretical reasoning and social and political action makes Popper's philosophy comparable with that of great classics like Aristotle or Kant. It was also one of the sources of the great success which Popper enjoyed still during his lifetime. As a general and unifying view, 'critical rationalism' is, of course, liable to oversimplification and, among certain adherents of Popper there has, in fact, been a considerable amount of that. On the other hand, it was only critical rationalism as a general orientation that could publicly compete with the eloquence of hermeneutic or neo-Marxist approaches. Philosophical theories like logical empirism, though with respect to intellectual clarity and responsibility by no means behind Popper's, would have been too academic to develop any major public significance.

In fact, it is not very difficult to give an outline of the basic principles of critical rationalism to somebody who is not a trained philosopher. Critical rationalism is an approach which offers something at any depth of understanding. Popper's works both cover extremely intricate discussions, e.g., of the methodological role of probability, whose significance can only be judged by highly specialist researchers, and general questions of reasoning, society, and politics at a level accessible to anybody who is really interested. With the exception of the chapters on probability and on quantum mechanics and of the appendices, the *Logik der Forschung* can easily be studied without deep expertise in academic philosophy. Even from a first reading, the average student would gain a preliminary understanding of falsificationism. Matters of presentation are very important in philosophy, and Popper was a master in that. His language is extremely clear and nonpretentious, which distinguishes him from many other philosophers, particularly in the German tradition.

Popper made a strong impression on politicians of various directions, although he was himself never involved in party politics. Both conservative and social democratic or labor politicians have referred to Popper to justify their respective programs and proposals. The term 'open society,' which Popper coined, has become widely accepted as a designation of basic features of a democratic state.

Popper's lasting impact on the shape of philosophy of science was not only due to his methodology in the narrower sense, but also due to his drawing attention to the *dynamics* of science. It is significantly due to Popper's influence that the philosophical and logical reconstruction of scientific reasoning is not just a modeling of static theories in logical terms but, to a great extent, an account of theory dynamics, i.e., of what it logically means that a theory is being given up or revised. The combination of philosophy of science with the history of science, which is reflected in many divisions of university departments, denominations of chairs, etc. emerged on this background, as philosophical theories of theory dynamics have to be tested against the historical development of theories, and, conversely, the proper understanding of historical developments requires appropriate methodological tools. Although today's theorizing differs in many respects from what Popper argued for and often contradicts his claims (i.e., Bayesian approaches in the field of belief revision), the research program of putting dynamics of theories first in the methodology of science is unthinkable without Popper's revival of the induction problem and his negative methodology as a proposed solution.

The influence of Popperian theories on the 'working scientist' should not be underestimated. If scientists have any interest in the methodology of science at all, they would know of Popper's approach. This is particularly true of the social and behavioral sciences. There, falsificationism figures prominently in textbook chapters dealing with methodological questions. One may speculate why falsificationism is so appealing to social scientists. A possible answer might be that statistical methodology, which lies at the heart of the social sciences, has a bias towards falsification. According to the most widespread statistical methodology, which is a combination of accounts by R. A. Fisher and by J. Neyman and E. Pearson, the rejection or falsification of a statistical null hypothesis at a certain level of significance is a direct and uncontroversial result, whereas the acceptance of a positively formulated statistical hypothesis is only indirect and preliminary (as it depends on features such as sample size) and, in any case, is technically more involved.

Certain results of modern computer science have challenged Popper's view on induction. His methodology comes along with the claim that we always start with theories or hypotheses which, in the course of time, are then revised to better hypotheses by trial and error. How a new hypothesis is formed was for Popper not the subject of rational investigation. In artificial intelligence and, in particular, in learning theory, algorithms have been developed to generate theories that fit given data. This does not contradict falsificationism, as these heuristically developed hypotheses would then have to be tested, just as falsificationism claims. However, the availability of such procedures destroy the myth that a theory is the extrarational product of a genius or some sort of accidental mutation, and lays the foundation for algorithmic inductive heuristics as a rational enterprise.

There are a lot of questions in the technical development of the Popperian approach, e.g., his treatment of probability and of logic, whose significance for future developments has still to be proven.

Something similar holds for Popper's metaphysical theories. However, these are less important details compared with his basic methodological claims, which will remain a standing challenge.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Causation: Physical, Mental, and Social; Explanation: Conceptions in the Social Sciences; Hayek, Friedrich A von (1899–1996); Historicism; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Positivism, History of; Positivism: Sociological; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Corri R 1996 *An Introduction to the Thought of Karl Popper*. Routledge, London

Geier M 1994 *Karl Popper*. Rowohlt, Hamburg

Hacohen M H 2000 *Karl Popper—The Formative Years, 1902–1945: Politics and Philosophy in Interwar Vienna*. Cambridge University Press, Cambridge, UK

Keuth H 2000 *Die Philosophie Karl Poppers*. Mohr Siebeck, Tübingen

Magee B 1985 *Philosophy and the Real World: An Introduction to Karl Popper*. Open Court, La Salle

O'Hear A 1995 *Karl Popper: Philosophy and Problems*. Cambridge University Press, Cambridge, UK

Popper K R 1925 Über die Stellung des Lehrers zu Schule und Schüler: Gesellschaftliche oder individualistische Erziehung? *Schulreform* **4**: 204–8

Popper K R 1934 *Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft*. Julius Springer, Vienna (10th edn. Mohr Siebeck, Tübingen 1994) [*The Logic of Scientific Discovery*. Hutchinson, London 1959 (10th edn. 1980)]

Popper K R 1945 *The Open Society and its Enemies, Vol. I: The Spell of Plato, Vol. II: The High Tide of Prophecy: Hegel Marx, and the Aftermath*. George Routledge, London (2nd edn. Routledge & Kegan Paul, London 1952, many reprints)

Popper K R 1957a The *Poverty of Historicism*. Routledge & Kegan Paul, London [originally as a series of papers in *Economica* **11** (1944) and **12** (1945)]

Popper K R 1957b Probability magic, or knowledge out of ignorance. *Dialectica* **11**: 354–72

Popper K R 1963 *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge & Kegan Paul, London

Popper K R 1972 *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford, UK (2nd edn. 1979)

Popper K R 1976 *Unended Quest: An Intellectual Autobiography*. Fontana, London (= *Intellectual Autobiography* In: Schilpp 1974)

Popper K R 1979 In: Hansen T E (ed.) *Die beiden Grundprobleme der Erkenntnistheorie*. Mohr Siebeck, Tübingen

Popper K R 1982a The open universe: An argument for indeterminism. In: Bartley W W III (ed.) *Postscript to the Logic of Scientific Discovery*. Hutchinson, London

Popper K R 1982b Quantum theory and the schism in physics. In: Bartley W W III (ed.) *Postscript to the Logic of Scientific Discovery*. Hutchinson, London

Popper K R 1983 Realism and the aim of science. In: Bartley W W III (ed.) *Postscript to the Logic of Scientific Discovery*. Hutchinson, London

Popper K R 1984 *Auf der Suche nach einer besseren Welt: Vorträge und Aufsätze aus dreißig Jahren*. Piper, München (*In Search of a Better World: Lectures and Essays from Thirty Years*. Routledge, London 1992)

Popper K R 1985 In: Miller D (ed.) *Popper Selections*. Princeton University Press, Princeton, NJ

Popper K R 1990 *A World of Propensities*. Thoemmes, Bristol, UK

Popper K R 1994a The myth of the framework: In: Notturno M A (ed.) *Defence of Science and Rationality*. Routledge, London

Popper K R 1994b Knowledge and the body–mind problem. In: Notturno M A (ed.) *Defence of Interaction*. Routledge, London

Popper K R 1994c *Alles Leben ist Problemlösen: Über Erkenntnis, Geschichte und Politik*. Piper, München (*All Life is Problem Solving*. Routledge, London 1999)

Popper K R 1996 *The Lesson of this Century: Karl Popper Interviewed by Giancarlo Bosetti*. Routledge, London

Popper K R 1998 In: Petersen A F (ed.) *The World of Parmenides. Essays on the Pre-Socratic Enlightenment*. Routledge, London

Popper K R, Eccles J C 1977 *The Self and its Brain: An Argument for Interactionism*. Springer International, Berlin

Sandra Pralong I J (ed.) 1999 *Popper's Open Society After Fifty Years*. Routledge, London

Schilpp P A (ed.) 1974 *The Philosophy of Karl Popper*. Open Court, La Salle, 2 Vols.

Shearmur J 1996 *The Political Thought of Karl Popper*. Routledge, London

Simkin C G F 1993 *Popper's Views on Natural and Social Science*. Brill, Leiden, The Netherlands

P. Schroeder-Heister

Popular Culture

For most of us, popular culture is a source of entertainment. It encapsulates mass media (such as film, television, print media), and includes sports events, advertisements, street fairs, and tourism. We engage with popular culture because we get pleasure from it; at the same time, we are also informed by it. We usually associate information with news rather than with entertainment. But other forms of popular culture that we often associate with 'mere entertainment'—for instance, tourist practices, popular music, films, sports, and television programs—also inform us in profound ways. For example, the US sitcom, the Cosby Show, may shape our ideas not just of what an ideal family is, but also about the relationship between race and upward mobility; a romance novel might instill in us expectations about intimacy as well as about gender and sexuality. Popular culture plays an important role in shaping our personal and collective identities. It also provides us with an analytic lens to understand sociohistorical processes such as class, gender and sexuality, nationalism, and transnationalism.

1. What is Popular Culture?

In the modern world, popular culture is available to what we may call 'the public' or to 'masses' of people. For this to happen, there have to be established means of production, distribution, and reception. For example, for a soap opera to exist there have to be a crew of script writers, actors, a director, a producer—in other words, a crew of people who work behind the scenes and in front of the camera. There also has to be a means to finance this production: in capitalist societies, soap operas are financed by sponsors who pay for the production and telecast of a soap in exchange for being able to advertise on it. This is how the term 'soap opera' originated: soap operas were sponsored originally by manufacturers of detergents. Even today, the manufacturers of detergents and other household goods like washing machines, kitchen appliances, and cleaning agents are key advertisers on soap operas. There also needs to be an infrastructure for its distribution. This is where television networks come in. Very often, television networks are part of larger media conglomerates or groups of companies. Television production is, therefore, part of a larger system of industrial production.

Popular culture presents us with a vantage point on the world: it is a domain that shapes our sense of self or our identity. Not only is it available generally to a large number of people, it also responds to and creates points of identification among large groups within a community or nation. Popular culture plays an important role in constituting and representing the public to itself. It is public in two senses: not only does it circulate widely and is often associated with public exhibition and display, it also configures 'the public,' 'the nation,' 'society,' 'culture,' 'family,' and other collectivities. From a sociological viewpoint, popular culture is probably the most important influence on the lives of millions. However, it is important to recognize that viewers are not just unthinking targets of popular culture. Rather, in order to understand the cultural significance of popular culture, we need to examine how its consumers respond to, reinterpret, and employ its messages.

Popular culture is a site of struggle between dominant discourses and forces of resistance. It contains 'points of resistance' as well as 'moments of supersession;' it forms a 'battlefield where no once-for-all victories are obtained but where there are always strategic positions to be won and lost' (Hall 1981, p. 233). Mass media may open up a 'zone of debate' about issues current in a given community (Gupta 1995). For instance, television shows like the US sitcom 'Ellen' enabled the rearticulation of debates around homosexuality. Popular culture, then, does not simply 'reflect' the lives and worlds of audiences; nor does it 'impact' these worlds or lives in a simple way. Instead, popular cultural texts are social formations in and of themselves.

Our subjectivities are constituted in the process of making sense of the world around us and in our interaction with others. In capitalist and postcapitalist societies, these processes are refracted by mass media. Film, television, popular music, and mass market fiction constitute and represent the public realm in cognitive, moral, and emotional terms. But these representations do not impact us as if we are blank screens—instead, 'they occupy and rework the interior contradictions of feeling and perception … they do find or clear a space of recognition in those who respond to them' (Hall 1981, p. 232). Thus, although representations are neither all-powerful nor all-inclusive, they have material effects on how people see themselves and lead their lives.

2. Scholarly Perspectives on Popular Culture

Definitions of popular culture have shifted according to the disciplinary emphases of its students. Historians of early modern Europe (e.g., Peter Burke (1978) and Robert Darnton (1984)) stress how members of different classes created and participated in popular culture and, simultaneously, were shaped and constrained by it. These historians analyze the factors influencing the popular cultures of early modern Europe, chiefly, the rise of cities and the advent of printing. They supplement thematic analyses of books with information about audiences and, thus, aim at providing the context for the meaning and uses of books.

Some of the foundational work on the relationship between capitalism and mass media is Max Horkheimer and Theodor Adorno's *The Culture Industry: Enlightenment as Mass Deception* (1944). Horkheimer and Adorno focused on how culture and the individual are 'produced' by the culture industry. They claimed that the culture industry's power lies in its ability to disseminate ideas and beliefs on a mass scale. They were concerned about how the management and control of leisure time and the socialization of the ego by the culture industry leads to a 'mass' culture that undermines the private realm.

The relationship between modern forms of popular culture and the growth of capitalism is also examined by some cultural studies scholars. Some of the most influential work in this area was produced (and has been influenced by) the Centre for Contemporary Cultural Studies at the University of Birmingham, UK. These scholars are concerned with how mass media shape dominant ideologies and discourses. Their central project is to examine the relationship between the production and consumption of popular culture and social axes of inequality, such as class, race, and gender. Although cultural studies' earliest borrowings were from literary criticism, chiefly the work of Raymond Williams and Richard Hoggart, its primary focus has been on practices of everyday life.

The greatest debts of this genre of cultural studies are to Marxist theories of culture. Its basic premise is that cultural production often takes the form of commodities. Other important propositions are: (a) popular culture is inextricably tied up with social relations; (b) popular culture can both enable social groups to define and realize their needs and reproduce asymmetries; and (c) popular culture is a terrain where social differences and struggles are articulated. These concerns reflect the influence of Gramsci on the ideological and popular struggles involved in the construction of common sense. In addition, some cultural studies scholars influenced by Lacanian analyses of film and television reconceptualize ideology in terms of how texts construct subject positions. Others have rethought the role of ideology in popular culture in terms of Foucault's analyses of discursive practices. Finally, breaking with Horkheimer and Adorno's assumption of the audience as a passive and undifferentiated mass of receivers, many cultural studies scholars conceive of audiences or consumers who actively 'decode' and interpret popular culture texts.

In addition to studying everyday practices, some cultural studies scholars also use semiotic methods of textual analysis. They treat cultural products as 'texts' to be 'read' according to interpretive strategies, including the literary analysis of narrative, semiology, and deconstruction. These analytical strategies reveal that literary conventions and forms have greater social and cultural significance than we might first suspect. For example, feminists have pointed to the place of fictional representations of romantic love in cultural constructions of femininity. Far from being 'mere' literary constructions, these representations play a crucial role in the constitution of dominant practices of gender and sexuality.

Anthropologists bring important theoretical and methodological insights to the study of popular culture because of their commitment to ethnographically analyzing the relationship between popular culture and other social institutions. Initially, mass media was neglected by anthropologists (exceptions include Bateson 1943 and Powdermaker 1950) because it was deemed 'too redolent of western modernity and cultural imperialism for a field identified with tradition, the non-western and the vitality of the local' (Ginsburg 1999, p. 297). Since the 1980s, however, many anthropologists have analyzed popular culture by studying mass media. They have examined the cultural implications of popular culture through an ethnographic attention to the everyday practices and lives of producers and/or consumers of media, and have traced how popular culture participates in the formation of identities, collectivities, and imaginaries.

Anthropological perspectives underscore that popular culture cannot be generalized across different historical and cultural contexts simply on the basis of the mass media apparatuses. Producers and consumers of mass media are, fundamentally, social and historical subjects and mass media are, primarily, cultural institutions that need to be studied in their concrete particularity. Hence, we cannot conceptualize relationships between popular culture and the social actors who engage its myriad texts in the abstract: neither can be studied outside their location in specific socio-historical contexts.

Several anthropologists have engaged in analyses of mass media texts. These scholars, together with those who study the reception of mass media, carefully situate popular culture in the discursive and historical contexts in which it is produced and consumed. These studies undermine unilinear models of 'communication flows,' and foreground the importance of examining the specific conditions in which popular culture circulates.

3. *Popular Culture as a Lens for Studying other Sociocultural Processes*

Popular culture is a terrain where class identities are articulated, negotiated, and consolidated. The production and consumption of popular culture is mediated by class. For example, class provides individuals with 'cultural constraints' and a common 'outer frame of meaning' (Darnton 1984, pp. 6–7). A mutually influencing relationship exists between elite culture and popular culture. Popular culture also provides a site for the negotiation and articulation of class identities and hegemonies. Thus, in nineteenth-century India, Bengali elites took upon themselves the task of 'purifying' popular culture produced by working-class women: while their objective was ostensibly to 'emancipate' women of their class, their efforts resulted in the consolidation of their own middle- and upper-class identities through the marginalization of working-class popular culture (Bannerjee 1989).

The role of popular culture in mediating discourses of gender and sexuality has long been acknowledged. Two pioneering works in this regard are Laura Mulvey's *Visual Pleasure and Narrative Cinema* (1975) and Janice Radway's *Reading the Romance* (1984). Laura Mulvey examined how the cinematic forms and narratives of classic Hollywood film 'suture' viewers into 'spectatorial positions' that reinscribe dominant ideas about gender and sexuality. Mulvey's work provided feminist critics of popular culture with a paradigm for examining the role of mass media in the constitution of discourses of gender and sexuality. Janice Radway's work laid the foundation for studying the role of a specific genre of texts, romance fiction, in the constitution of discourses of gender and sexuality. Radway situated women's reading practices in their social relationships and everyday lives, and studied their interpretations of different romances in order to analyze what the books mean to their readers.

Since the publication of these works by Mulvey and Radway, the relationship of popular culture to gender

and sexuality has been analyzed by scholars too numerous to cite. Some of the most noteworthy studies in this regard, however, come from scholars, chiefly anthropologists, who have examined media in non-Western contexts. These include the analysis of the relationship between popular culture, gender and sexuality in Japan (Allison 1996), and television and notions of modern womanhood in Egypt and India (Abu-Lughod 1993, Mankekar 1999, respectively). The significance of these studies lies in their decentering of Eurocentric perspectives dominant in cultural studies, media studies, and feminist studies.

Some scholars have argued that the conceptualization of popular culture was itself engendered by the emergence of nationalism (Burke 1978). Building on Benedict Anderson's insights about the role of print capitalism in the construction of nations as imagined communities (1983), several scholars have examined the role of popular culture in the formation of national identity. For instance, diverse forms of popular culture, including tourism, exhibitions, material culture and consumption, have enabled the constitution of the national imaginary in China (Anagnost 1997). In some parts of the developing world, television has been consciously deployed to construct a 'national community': in Mexico and Venezuela, indigenous soap operas called telenovelas have combined entertainment with 'social messages' about development. In India, state-owned television was harnessed to the explicitly nationalist goals of development and national integration (Mankekar 1999) and, furthermore, participated in the consolidation of an exclusionary and strident Hindu nationalism (Rajagopal 2000).

Yet, at the same time that popular culture enables the constitution of local and national communities, some of its forms are inherently transnational. Thus, the telenovelas described above circulated not only within but across national borders in Latin America, and were adopted as models for the television programs in India. In addition, the transnational circulation of popular culture has resulted in the creation of new forms of diasporic identities. The transnational circulation of newspapers, film, television programs relayed via satellite, advertisements, and mass market fiction have created, what anthropologist Arjun Appadurai terms, 'mediascapes' (1996). Mediascapes 'refer both to the distribution of the electronic capabilities to produce and disseminate information … now available to a growing number of private and public interests throughout the world, and to the images of the world created by these media' (1996, p. 35).

The boundaries of cultures and nations have thus proved to be permeable to globalized and transnational forms of popular culture. Yet, globalization has not resulted in a homogenization of culture. At different historical moments, globalization has resulted in a recasting of 'tradition' and processes of creolization (Hannerz 1996, pp. 67–77). Transnational popular culture thus forces us to revisit previously held notions of culture as bounded and systemic (Appadurai 1996, Gupta and Ferguson 1997).

See also: Adolescents: Leisure-time Activities; Critical Theory: Contemporary; Critical Theory: Frankfurt School; Cultural Assimilation; Fashion, Sociology of; Fast Food Enterprises; Film: Genres and Genre Theory; Film: History; Globalization and World Culture; Mass Society: History of the Concept; Music as Expressive Form; Television: Genres; Television: Industry

Bibliography

Abu-Lughod L 1993 Finding a place for Islam: Egyptian television serials and the national interest. *Public Culture* **5**: 493–514

Allison A 1996 *Permitted and Prohibited Desires*. Westview Press, Boulder, CO

Anagnost A 1997 *National Past-times: Narrative, Representation, and Power in Modern China*. Duke University Press, Durham, NC

Anderson B 1983 *Imagined Communities*. Verso, London

Appadurai A 1996 *Modernity at Large*. University of Minnesota Press, Minneapolis

Bannerjee S 1989 Marginalization of women's popular culture in nineteenth century Bengal. In: Kumkum S, Sudesh V (eds.) *Recasting Women: Essays in Colonial History*. Kali, New Delhi, India, pp. 127–80

Bateson G 1943 Cultural and thematic analysis of fictional films. *Transactions of the New York Academy of Sciences* **5**: 72–8

Burke P 1978 *Popular Culture in Early Modern Europe*. Harper and Row, New York

Darnton R 1984 *The Great Cat Massacre and Other Episodes in French Cultural History*. Vintage, New York

Ginsburg F 1999 Shooting back: From ethnographic film to indigenous production/ethnography of media. In: Miller T, Stam R (eds.) *A Companion to Film Theory*. Blackwell, Malden, MA, pp. 295–322

Gupta A 1995 Blurred boundaries. *American Ethnologist* **22**: 375–402

Gupta A, Ferguson J (eds.) 1997 *Culture, Power, Place*. Duke University Press, Durham, NC

Hall S 1981 Notes on deconstructing 'the popular'. In: Samuel R (ed.) *People's History and Socialist Theory*. Routledge, London, pp. 227–40

Hannerz U 1996 *Transnational Connections*. Routledge, London

Horkheimer M, Theodor A 1944 *Dialectic of Enlightenment*. Continuum Press, New York

Mankekar P 1999 *Screening Culture, Viewing Politics*. Duke University Press, Durham, NC

Mulvey L 1975 Visual pleasure and narrative cinema. *Screen* **16**(3): 6–18

Powdermaker H 1950 *Hollywood: The Dream Factory*. Houghton Mifflin, Boston

Radway J 1984 *Reading the Romance*. University of North Carolina Press, Chapel Hill, NC

Rajagopal A 2000 *Retailing Hinduism*. Cambridge University Press, Cambridge, UK

P. Mankekar

Population Aging: Economic and Social Consequences

The notion of aging is quite natural for human beings: from the age of reason, we understand that with the passage of time, every living thing ages and finally dies. Young children experience this with their goldfish or dog, adults with their parents and old people with their friends. By analogy, we apply the notion of aging even to nonliving things: for example, as a car deteriorates, we say that it is getting old and eventually we relegate it to the scrap lot. Although a good wine may improve with age, in the end it becomes undrinkable.

From the individual, the notion of aging is extended to the group formed by the individuals: thus we speak of an aging fleet of cars or an aging population, meaning that the average length of life of the entire group of these individual units changes with the passage of time. However, unlike the case for the individuals making up the group, the group itself is able not to age, or even to get younger. This is the case for human populations. For them the aging process is strictly structural, not biological. The decay generally associated with the aging of individuals should not be assumed to apply to aging populations.

1. *Population Aging: A Modern Demographic Phenomenon*

A population changes through the interaction of the fertility, mortality, and migration of its members. In recent centuries, we have witnessed remarkable changes in human fertility and mortality, which, uncontrolled since time immemorial, had always been high. In the past, there was a 'natural' fertility and mortality, and population numbers throughout the world were relatively low. Eventually, people found ways to reduce first mortality and then fertility after a certain time lag. In theoretical terms, this historical process is known as the demographic transition (Chesnais 1986). During this process, the population grows at a higher or lower rate, depending on the extent of this time lag. At the end of the process, according to the theory, both fertility and mortality are low and the population stops growing. But an inescapable law of demography says that a population either grows or it ages (Coale 1987). As all populations around the world have reduced or stopped their growth, they are all bound to age. In some countries the fertility level is well below the replacement level, and thus their populations will decrease unless they experience in-migration, a possibility not foreseen by classical transition theory.

There are several ways of measuring the aging of a population. Instinctively, one could say that a population is aging if the average—or median—age of its members as a group increases over time. Alternatively, one could look for an increase in the proportion of old people in the population. But at what age does a person become old? We do not want to enter into a debate over the exact threshold of old age or retirement, but for purposes of illustration we will use the percentage aged 65 and over in the population . To avoid delving too far into the past or future, we will restrict our discussion of demographic aging to the period from 1950 to 2050.

During the twentieth century, the growth in the number of people aged 65 and over was relatively small and there were roughly as many people old (according to this definition) in industrialized countries as in less developed countries. However, from now until 2050, this number will grow spectacularly, mainly because of the increase in less developed countries (Table 1a). But if we look at the proportion of people aged 65 and over in the population, as a measure of demographic aging between 1950 and 2050 (Table 1b), we see that the situation differs greatly from one continent to another. Africa will remain young while the more developed countries as a whole will have more than one-quarter of their populations in the 'old' category. Asia will age at an increasing rate: in 2050, it will have more than 900 million people aged 65 and over. This is two-thirds of the world's elderly population.

In the industrialized countries, the process of population aging is relatively slow, but in developing countries it will be more intense and rapid. This can be clearly shown by a comparison of Canada and China (Fig. 1). In 2050, China will be a world leader in demographic aging with almost 450 million people aged 60 and over. This is close to 30 percent of its total population.

In the face of globalization, the changes occurring over this century will give rise to a number of socioeconomic issues that would not have made themselves felt if population aging affected only the currently industrialized countries. There will be both demographic and economic and social consequences for our societies.

2. *Demographic Consequences*

As is easily demonstrated, in the early stages of the demographic transition aging is much more the result of a decline in fertility than in mortality. When mortality declines from a high level, it is primarily

Table 1a/b
Population aged 65 and over, 1950–2050

Regions	1950	1975	2000	2025	2050
(a) Population aged 65 and over, in thousands					
World Total	131 118	232 258	417 798	813 665	1 461 092
More Developed Regions*	64 202	112 177	171 069	253 912	299 249
Less Developed Regions*	66 644	118 027	248 221	561 749	1 163 054
Africa	7 070	11 068	25 102	51 932	141 287
Latin America and the Caribbean	6 179	13 841	28 034	66 183	135 897
Northern America	14 073	25 079	38 704	69 086	85 800
Asia	57 483	101 037	217 270	477 037	911 442
Europe	44 880	77 108	107 146	147 490	173 243
Oceania	933	1 608	2 820	5 788	8 636
(b) Population aged 65 and over, percentage of total population					
World Total	5.2	5.7	6.9	10.4	16.4
More Developed Regions*	7.9	10.7	14.4	20.9	25.9
Less Developed Regions*	3.9	3.9	5.1	8.5	15.0
Africa	3.2	3.1	3.2	4.0	8.0
Latin America and the Caribbean	3.7	4.3	5.4	9.5	16.8
Northern America	8.2	10.3	12.5	19.0	21.9
Asia	4.1	4.2	5.9	10.1	17.3
Europe	8.2	11.4	14.7	21.0	27.6
Oceania	7.4	7.5	9.9	14.6	18.7

* More developed regions comprise Northern America, Japan, Europe, and Australia/New Zealand. Less developed regions comprise all regions of Africa, Latin America and the Caribbean, Asia (excluding Japan), and Melanesia, Micronesia, and Polynesia.
Source: United Nations Department of Economic and Social Affairs 1999

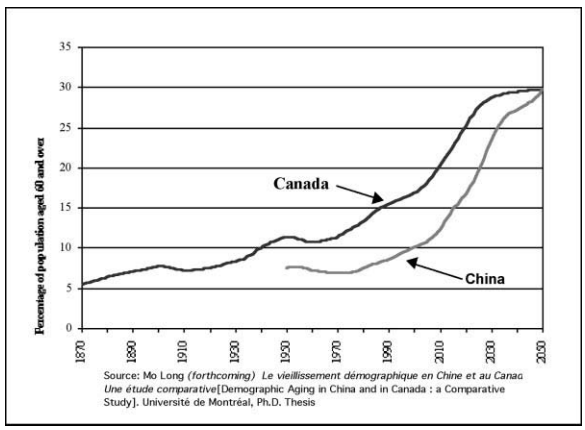

Figure 1
Population aging in Canada and China, 1870–2050

early deaths that are eliminated, in particular deaths of children and young adults. The result is a younger (not older) population, as was seen in Africa in the 1950s and 1960s (Table 1b).

But when life expectancy is already very high, as is the case now in industrialized countries, mortality becomes an important factor in demographic aging. If mortality is low and decreasing, we get a double effect. First, increasing numbers of people reach the old age threshold—because of lowered mortality before that age. Second, people live for increasing lengths of time beyond this threshold—owing to a continued decline of mortality at old age (Bourbeau et al. 1997). The main consequence of this development is that old people no longer form a homogenous group: we now have the young old, who generally are in quite good health, and the oldest old, who generally are more dependent for income security and health. Thus we have moved from having three main lifecycle stages to four. There is a new third age that extends from retirement to the oldest-old stage. This is an entity that never existed before in human history (Laslett 1989). Another important feature in mortality changes in many societies has been an increased mortality gap between men and women. Very frequently old people in these societies are and will continue to be primarily old women.

Like any other population, the elderly population is subject to a process of incessant inward and outward movement of cohorts. In most industrialized countries the incoming generations will be clearly differentiated with respect to their numbers and their characteristics. The generations now approaching retirement are the Baby Boomers born after World War II. They will be followed by the Baby Busters, who are far less numerous. These generations, particularly the women,

will have characteristics very different from those of today's old people (Marcil-Gratton and Légaré 1987), especially in terms of education.

The changes in the age structure of our societies will cause major realignments that affect not only the elderly but the entire population. In terms of health and financial security, will old people be more independent or more dependent than before? The increasing presence of very old people tends to suggest dependence, while the characteristics of the future younger elderly appear to suggest greater autonomy. Will the younger adults, who are far less numerous than the Baby Boomers, be willing to pay for a golden twilight for tomorrow's seniors? We now turn to these questions in the following sections.

3. *Economic Consequences*

3.1 *The Aging Process and the Working-age Population*

At the beginning of the population aging process, a standard age pyramid shows the population cohorts by age increasingly reduced by mortality. As the process gradually proceeds, positive impacts on the working-age population appear. The decline in mortality eliminates early deaths and increases population numbers. Then, as the effects of fertility decline come to the fore, cohort size is increasingly reduced. When mortality improvements lead to significant reductions in old-age deaths, the members of the retired cohorts become as numerous as, and sometimes more numerous than, those of working-age cohorts. The pyramid changes into a top and finally into a cylinder. This means that at the end of the aging process, all generations are of approximately the same size except for the very old. This process is reflected in the variations of what are commonly called the dependency ratios, one of which is the ratio of people aged 65 and over to those aged 19–64. Unfortunately, the use of the adjective 'dependency' may unduly give an impression of economic dependence. These indicators are merely indices of the age structure and it would be more appropriate to speak of demographic dependency ratios.

In studies of aging we frequently find the members of the working-age population equated to actual workers. While in the past this was true for men when their labor force participation rate was close to 100 percent, it is now less so, in particular for women. To assess the economic impacts of aging and our ways of dealing with them, we needed to take into account not only population numbers, but perhaps more importantly labor force participation rates and worker productivity (Fellegi 1988). If the shift from agricultural to industrial societies had great impacts on the world of work, the shift to aging post-industrial societies will have an equally great impact. This is particularly true for the change in the rate of labor force participation by women and by older workers.

The radical decline in fertility and otherwise changing family arrangements has freed women from a myriad of family duties and made them available for increased participation in the paid labor market. The women's liberation movement also contributed to changes in attitudes and women are becoming increasingly independent, not only at working age, but also, and perhaps especially, during retirement.

One of the corollaries—and achievements—of the development of industrial societies has been the right to retirement and associated pensions. In this century we have witnessed a decline in the labor force participation rate of older workers, along with a systematic decline in the age of leaving the labor market. This ever-earlier retirement from the labor market is now increasingly being questioned. As life expectancy increases, especially at old age, and as the time spent taking education is extended, the life segment in retirement is growing longer and the segment in which one works is getting shorter. When a large and growing share of the costs of social security, including pensions and health insurance, is borne by those who work, some adjustments must be considered. It is not possible to maintain contribution and benefit levels and, simultaneously, to bring down, or even maintain, the age of access to benefits.

There is a need for greater flexibility, both in wages and responsibilities for the population approaching retirement. For older workers, one should consider a gradual progression from wages to pension income, perhaps following a normal rather than S-shaped curve. This may call for a change in attitudes, not only among the workers themselves but also among employers and trade unions.

Solidarity among generations presupposes a minimum of equity between groups. When social security programs in the industrialized countries were reorganized after World War II, both the population and the economy were growing: the populations were young. Pay-as-you-go financing did not create any visible problems and it eliminated some perverse effects of funded programs, many of which had collapsed in Europe during the Depression and the war. In aging societies—particularly those which have experienced a large baby boom followed by a baby bust—a hybrid system combining pay-as-you-go and funded elements is likely to be more equitable across generations.

At present, the labor market is undergoing major changes. Workers will no longer hold full-time jobs for life the way they did previously in industrial societies. New workers will enter the labor market later in life. They will hold part-time jobs with varying employers, more often and possibly for a significant part of their working life, or be self-employed. From time to time they will have periods of unemployment or of retrain-

ing. Moreover, such patterns, which formerly applied mainly to women, will now apply to men as well. Consequently, retirement systems based on previous labor market participation and involving individual employers will be subject to in-depth revision. In addition, one should take the value of unpaid work into account. In particular, work performed to raise children or to care for aging parents needs to be recognized.

3.2 An Aging Society: A Burden?

Because population aging is equated to individual aging all too often, one tends to highlight problems for aging societies, in particular problems related to the costs of health and pension programs for the elderly. On this subject some issues need to be clarified and some myths should be dispelled.

First, the situation of old people in most countries has improved greatly largely as a result of post-war social programs. There have even been hard economic times when their situation has deteriorated less than that of younger people.

Second, health and pension costs will certainly continue to increase in aging societies, but at what rate will this happen? With regard to health Zweifel et al. (1999) have shown that increases in medical costs are more closely linked to the number of old-age deaths than to the size of the elderly population, and the number of old-age deaths is increasing much less rapidly than the size of the elderly population. The same is not true for pensions, whose costs are linked directly to the number of old people. However, one should not forget that retired people also pay taxes, both on their income and on consumption. As people age, their material needs (other than those related to health) decrease, however, and this might be reflected in their pension requirements.

Third, although incomes are on average lower at old age than at other ages, the elderly often hold a large share of the collective wealth through their assets, in particular by owning of their dwellings.

In brief, far from being an impediment to economic development, elderly people in aging societies can sometimes serve as an important actor if not the engine of such development.

4. Social Consequences

Increasingly, the group of old people is not homogeneous with regard to health status and wealth, and this has numerous consequences for their life-style. First, let us recall that to a large extent the world of the elderly is a world of old widows. This feature has become increasingly prominent over the twentieth century, as excess male mortality has grown while the difference in age at marriage has remained stable. In addition, this century has been characterized by the quest for greater independence at all ages. Old people in industrialized countries prefer to live alone if they do not live with their spouse, particularly when they are in good health. But this choice does not necessarily imply isolation. Alternatively, living together with people other than one's late spouse may become a strategy for dealing with financial or health problems, perhaps as a stage that precedes institutionalization.

Multigenerational households are much more common in less industrialized countries, mainly for cultural and economic reasons. We can only wonder what will happen in the future in these societies. In a context of rapid urbanization and scarcity of dwellings, there will be an acceleration in aging following the sharp decline in fertility. With its one-child policy, China is a particularly poignant case. Demographic constraints may become apparent as daughters and daughters-in-law become increasingly less numerous and less often available to take care of old parents.

In societies with low mortality (or equivalently, long life expectancy), early deaths are reduced to a minimum and are usually related to life-style rather than caused by exogenous diseases. This produces a great deal of interest in the population's health status, and public health policy objectives may concentrate on insuring quality of life for the aging population more than on preventing early adult deaths. The focus moves from mortality to morbidity, from life expectancy to healthy-life expectancy (Robine and Romieu 1998). By means such as surveys on seniors' functional capacities, it is possible to identify impediments to independent existence and to develop policies that will minimize the adverse effects of biological and social aging.

Such policies are centered on support networks. Assistance provided to old people may come from professionals hired by the state, the community, or private enterprise. Alternatively, they may be offered by the family, friends, or neighbors. The expanding need for services provided through such formal and informal networks is an increasing source of concern. For the informal networks, the many changes that have occurred in family structures leave future prospects unclear; for the formal networks, it is mainly the cost that worries public policy makers. In this picture one should not forget that seniors are not only receivers of assistance, but very often also providers.

5. Conclusion

Some believe that the twenty-first century will see the emergence of intergenerational conflicts similar to the gender conflicts of the twentieth century. Major transformations of the population age structure have altered the ratio between the retired and the working population and will continue to do so. In addition to solidarity among generations, any reforms to social

security programs taking account of the new demographic reality need to aim to promote intergenerational equity. For this purpose, we have seen the development of new measurement techniques, known as demographic accounting (Auerbach et al. 1999). By making use of individual-level data and by making adequate long-term projections we will be able to benefit from these new methodologies.

In conclusion, we observe that for centuries, life in all its forms was regulated by nature. Gradually, human societies became able to control their destiny by largely controlling their mortality and fertility. This resulted in great changes for the societies, and in particular it accelerated their aging. The fact that most now live to old age and that our populations progressively age can only be regarded as great cultural progress (Loriaux et al. 1990). However, like all progress, individual and societal aging must be managed skillfully. Moreover, we are in the midst of a process of shifting control from society to the citizen (Légaré and Marcil-Gratton 1990). Decisions concerning demographic behavior, like whether or not to get married, whether to have children, how many and when, whether to leave an existing marriage and contract another one, and so on, are now made on the individual level more than ever before. In a world where centenarians will become rather common, people will doubtless also want to have a say on the timing of their departure from life.

See also: Age Policy; Age, Sociology of; Age Stratification; Age Structure; Aging and Health in Old Age; Cohort Analysis; Old Age and Centenarians; Retirement and Health; Retirement, Economics of; Social Security

Bibliography

Auerbach A J, Kotlikoff L J, Leibfritz W (eds.) 1999 *Generational Accounting around the World*. University of Chicago Press, Chicago

Bourbeau R, Légaré J, Emond V 1997 *New Birth Cohort Life Tables for Canada and Quebec, 1801–1991*. Statistics Canada, Ottawa, Canada

Chesnais J C 1986 *La transition démographique: étapes, formes, implications économiques: étude de séries temporelles (1720–1984) relatives à 67 pays*. Presses universitaires de France, Paris [1992 *The Demographic Transition: Stages, Patterns, and Economic Implications*. Oxford University Press, New York]

Cliquet R, Nizamuddin M (eds.) 1999 *Population Ageing. Challenges for Policies and Programmes in Developed and Developing Countries*. United Nations Population Fund, New York

Coale A 1987 How a population ages or grows younger. In: Menard S W, Moen E W (eds.) *Perspectives on Population: An Introduction to Concepts and Issues*. Oxford University Press, New York

Fellegi I P 1988 Can we afford an aging society? *Canadian Economic Observer* **1**,**10**: 4.1–4.34

Laslett P 1989 *A Fresh Map of Life. The Emergence of the Third Age*. Weidenfeld and Nicolson, London

Légaré J, Marcil-Gratton N 1990 Individual programming of life events: A challenge for demographers in the 21st century. *Annals of the New York Academy of Sciences* **610**: 98–105

Loriaux M, Remy D, Vilquin E (eds.) 1990 *Populations âgées et révolution grise: les hommes et les sociétés face a leurs vieillissements* [Aged Populations and Grey Revolution: Humans and Societies Facing their own Ageing]. CIACO, Louvain-la-Neuve, Belgium

Marcil-Gratton N, Légaré J 1987 Being old today and tomorrow: A different proposition. *Canadian Studies in Population* **14**,**2**: 237–41

Martin L G, Preston S H (eds.) 1994 *Demography of Ageing*. National Academy Press, Washington, DC

Mo Long (forthcoming) Le vieillissement démographique en Chine et au Canada: une étude comparative [Demographic Aging in China and in Canada: a Comparative Study]. Ph.D. thesis, Université de Montréal

Murray C J L, Lopez A D (eds.) 1996 *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020* (on behalf of the World Health Organization and the World Bank). Harvard University Press, Cambridge, MA

Organisation for Economic Co-operation and Development (OECD) 1988 *Ageing Populations. The Social Policy Implications*. Organisation for Economic Co-operation and Development (OECD), Paris

Organisation for Economic Co-operation and Development (OECD) 1998 *Maintaining Prosperity in an Ageing Society*. Organisation for Economic Co-operation and Development (OECD), Paris

Organisation for Economic Co-operation and Development (OECD) 2000 *Reforms for an Ageing Society*. Organisation for Economic Co-operation and Development (OECD), Paris

Robine J M, Romieu I 1998 *Healthy Active Ageing: Health Expectancies at Age 65 in the Different Parts of the World*. REVES/INSERM, Montpellier, France

Stolnitz G J (ed.) 1992 *Demographic Causes and Economic Consequences of Population Ageing: Europe and North America*. UN Economic Commission for Europe (ECE), Geneva, Switzerland

Stolnitz G J (ed.) 1994 *Social Aspects and Country Reviews of Population Ageing*. UN Economic Commission for Europe (ECE), Geneva, Switzerland

Suzman R, Willis D P, Manton K G 1992 *The Oldest Old*. Oxford University Press, New York

United Nations Department of Economic and Social Affairs 1956 *The Ageing of Populations and Its Economic and Social Implications*. United Nations, New York

United Nations Department of Economic and Social Affairs 1988 *Economic and Social Implications of Population Ageing*. United Nations, New York

United Nations Department of Economic and Social Affairs 1999 *World Population Prospects: The 1998 Revision*. United Nations, New York

World Bank 1994 *Averting the Old Age Crisis: Policies to Protect the Old and Promote Growth*. Oxford University Press, New York

Zweifel P, Felder S, Meiers M 1999 Ageing of population and health care expenditure: A red herring? *Health Economics* **8**,**6**: 485–98

J. Légaré

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Population and Technological Change in Agriculture

Induced innovation refers to the process by which societies develop technologies that facilitate the substitution of relatively abundant (hence cheap) factors of production for relatively scarce (hence expensive) factors in the economy. For example, the choice between land saving vs. labor saving technological change has been largely determined by the relative scarcity of land and labor in a particular society (Hayami and Ruttan 1985). Population pressure has played a defining role in technology development and farmer choice of technologies. Ester Boserup (1965) argued that societies across the world have responded to increasing land scarcity in a strikingly similar and predictable manner, by increasing land productivity through agricultural intensification. The transition from low-yield, land-extensive cultivation systems to land intensive, double and triple crop systems has only been profitable in societies where the supply of uncultivated land has been exhausted (Boserup 1965, Hayami and Ruttan 1985, Pingali et al. 1987). The process of agricultural intensification has been observed in traditional as well as modern agricultural societies.

The movement from forest and bush fallow systems of cultivation to annual and multicrop cultivation systems, whereby plots of land are cultivated one or more times per year, has generally been influenced by rising population densities in traditional societies (Boserup 1965). It is no accident that the modern seed-fertilizer revolution has been most successful in densely populated areas of the world, where traditional mechanisms for enhancing yields per unit area have been exhausted.

Kikuchi and Hayami (1978) document the process of land augmentation as the cultivation frontier closes for Japan, Taiwan, Korea, and the Philippines. 'As population pressure pushes the cultivation frontier into marginal areas, we expect the marginal cost of agricultural production via expansion of cultivated area to rise relative to the marginal cost of production via intensification. Eventually the economy will reach a stage at which internal land augmentation becomes a less costly means of increasing agricultural output than external land augmentation' (Hayami and Ruttan 1985, p. 310).

In addition to high population densities, intensive cultivation will also be observed in areas with better access to markets, provided soil conditions are suitable. Intensification occurs in the less densely populated areas for two reasons: (a) higher prices and elastic demand for output imply that the marginal utility of effort increases, hence farmers in the region will begin cultivating larger areas; and (b) higher returns to labor encourage migration into well-connected areas from neighboring regions with higher transport costs. In other words, where population densities are low, the returns to intensification are high only if markets, domestic and/or export, are easily accessible, such as the Central Plains of Thailand or parts of South America's Southern Cone. If these conditions are not present, labor and other costs associated with intensive agriculture are substantially higher than its incremental economic returns.

1. Farmer-induced Technological Change in Traditional Farming Systems

Pingali et al. (1987) documented, in the case of Africa, the movement from shifting cultivation to permanent agriculture with increases in population densities and improvements in market infrastructure. As land became scarce, traditional farming communities across sub-Saharan Africa extracted higher and higher levels of output from their land through investments in land improvements and soil fertility management. Erosion control, irrigation, and drainage are the most commonly made land investments for enhancing productivity. Tiffen et al. (1994) provided an excellent case study from the Machakos District of Kenya on the process of intensification and farmer investment in land improvements.

The extent of intensification is conditional on the relative responsiveness of the soils to inputs associated with intensive production such as land improvements, manure, and fertilizers. As population densities rise, one observes the cultivation of land that requires substantially higher labor input but that is also more responsive to the extra inputs. The responsiveness to intensification is generally higher on soils with higher water and nutrient holding capacity; these are usually the heavier, clayey soils, rather than the lighter sandy soils (Pingali et al. 1987). The movement from communal access, to individual property rights, with population growth induced increases in land values, provide individuals the incentive to invest in land improvements and other technologies necessary for intensive agricultural systems (Binswanger and Deininger 1997). Formal land ownership as characterized by the possession of a title also helps the farmer in acquiring credit for making the necessary investments in land and technology (Pingali 1990).

In using population densities as a basis for predicting the level of agricultural intensification, one ought to standardize arable land by soil quality and climate. Binswanger and Pingali (1988) provide a standardized population density measure: the number of people per million kilocalories of production potential. This measure is called the agroclimatic population density. When countries are ranked by agroclimatic population densities, Niger and Kenya turn out to be more densely populated than Bangladesh, and India ranks only twenty-ninth rather than seventh as in the traditional

population rankings. This explains why one observes intensive cultivation in some sparsely populated countries with poor market infrastructure—Niger and Botswana are examples.

2. Adoption of Modern Scientific Innovations

The intensification of traditional farming systems was a process that the more densely populated regions of Asia had been through several decades, and in some cases centuries, earlier. The application of modern scientific innovations in the form of high yielding seeds and fertilizer allowed the extremely land scarce regions of Asia achieve levels of land productivity that were not possible through the exclusive reliance on traditional farmer innovation systems.

A cross-sectional comparison across Asian countries shows that the proportion of irrigated area and the percentage of area under high yielding varieties of crops (HYVs) and fertilizer use is positively related to population densities (Pingali et al. 1997). It was not a historical accident that research for enhancing the yields of the two major food crops of Asia, rice and wheat, began in the 1950s. Research investments for enhancing yields were made in response to declining opportunities for area expansion, which was the main source of output growth in the pre-Green Revolution Asia. A comparison of sources of rice output growth during the early and late Green Revolution periods shows that the overall rate of growth was similar for the two time periods. In the early period, one-third of the growth came from an increase in cropped land, while in the latter period almost all of it was due to yield increases. The adoption of the Green Revolution technologies has been spotty in other parts of the world; this reflects in part the differences in land and labor endowments relative to Asia (Pingali and Heisey 1999). Research on induced intensification has not been restricted to the agricultural economics literature alone; a substantial number of comparative studies that look at the process of intensification, over time and over space, can be found in the geography–anthropology literature; see for example, Turner and Brush (1987) and Turner and Ali (1996).

3. Labor Use and Employment and Labor Saving Technologies

In both traditional and modern agricultural systems, intensification of agriculture leads to an increase in labor use per hectare, due to an increase in the number of tasks performed, such as irrigation, drainage, etc., and due to an increase in the intensity with which some tasks have to be performed, such as more intensive manure use and weeding (Pingali et al. 1987). Intensification therefore leads to an increase in agricultural employment as well as an increase in yields per hectare. Labor productivity can only be sustained, as intensity rises, if labor saving technologies are available.

Pingali et al. (1997) have argued that the adoption of the plow in African agriculture has been closely associated with the level of intensification. Similarly, increases in labor productivity in Asia are associated with the increasing adoption of labor saving technologies, both as a means of saving on the higher labor needs due to intensification as well as a means of economizing on labor as its opportunity cost rises with economic growth. The movement from single-crop cultivation systems to a double- and triple-crop systems shifts out the demand for labor and also increases the demand for timely completion of operations. Initially, the switch to modern varieties was profitable, even with higher labor requirements per hectare, because the labor requirements per ton of paddy were lower compared with that of traditional varieties (Barker and Cordova 1978). Over time however, increasing labor demand for peak period operations led to a rise in real wages for these operations, even in densely-populated labor surplus countries such as India and Indonesia. The rise in wages was further exacerbated by the concurrent growth in the rural nonfarm sector and increasing employment opportunities in the urban sector, as has happened in Japan and Korea and now being witnessed in Southeast Asia.

Widespread adoption of labor-saving mechanical and chemical technologies has alleviated the growing labor constraints, and contributed substantially to overall productivity growth (Sidhu and Byerlee 1992). The most notable confirmation of this hypothesis comes from the now developed country, Japan, where graphs plotting the log of land productivity vs. the log of labor productivity show almost no increases in land productivity but substantial gains in labor productivity in the last 30 years (Hayami and Ruttan 1985, Craig et al. 1997).

4. Intensification and Degradation—When Does the System Break Down?

The process of population induced technological change is by no means automatic; there are several instances where societies faced with growing population pressure have not been able to achieve sustainable intensification. Failures in agricultural intensification can be attributed to one or more of the following reasons: persistence of uncertain long-term rights to land; encroachment of cultivation onto marginal lands; collective effort required for watershed-level protective investments; and an inappropriate policy environment.

Farmer incentives for regenerating soil fertility and for making erosion-prevention investments are high if secure long-term land rights exist, both to currently cultivated land and to fallow land. Where farmers do

not have long-term rights to land, the private rate of discount is higher than the social rate of discount, hence private investments in land productivity will be lower than the societal optimum. In societies with slowly and steadily growing populations, property rights to land evolved systematically, being induced by growing land scarcity. However, the evolution to secure land rights has not been successful in societies experiencing rapid population growth (through natural increase or through migration), nor in societies where this induced institutional change is circumvented by government policies (e.g., land expropriation by colonial authorities or socialist governments).

Given secure rights to land, the rate of return to land investments varies by agroclimatic zone and soil type: degradation problems are most severe where the returns to land investments are lowest. The arid fringe areas, upper slopes in the semi-arid and the humid zones, and shallow sandy soils exhibit the highest levels of erosion, other things being equal. Degradation problems are likely to be most severe in regions with relatively large endowments of marginal lands.

Degradation of marginal lands is also likely to be severe where institutional arrangements are not available for controlling access to grazing and forestry. Where privatization is infeasible or undesirable, community action or government intervention is required both for control of access and for investments in erosion control and forest replanting. Difficult dilemmas frequently arise as traditional users are threatened with a loss of their use rights. Societies are often incapable of solving these problems of ownership or controlled access before serious damage has occurred.

In several instances, inappropriate policies, both micro- and macroeconomic, have created disincentives for farmer investments in land productivity enhancing and/or resource conserving technologies (Heath and Binswanger 1996). Pingali et al. (1997) argue that even in the intensive lowland cereal production systems, where Green revolution technologies were most successfully adopted, price and trade policy distortions are leading to significant levels of environmental and ecological degradation.

The potential problems of declining labor productivity and environmental degradation are not problems of levels of population densities. Given sufficient time, it is likely that a combination of farmer innovations, savings, and the development of research facilities and institutions for dealing with soil degradation issues will be able to accommodate much more than the current population in most countries, especially in many of the less densely populated ones. However, if all these changes are required quickly and simultaneously because of high population growth rates, they may emerge at too slow a pace to prevent a decline in human welfare.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Development and the State; Development: Rural Development Strategies; Development: Sustainable Agriculture; Food Production, Origins of; Industrialization; Industrialization, Typologies and History of; Innovation and Technological Change, Economics of; Innovation, Theory of; Technology Transfer

Bibliography

Barker R, Cordova V 1978 Labor utilization in rice production. In: *Economic Consequences of the New Rice Technology*. International Rice Research Institute, Los Baños, Philippines

Binswanger H, Deininger K 1997 Explaining agricultural and agrarian policies in developing countries. *Journal of Economic Literature* **35**: 1958–2005

Binswanger H, Pingali P 1988 Technological priorities for farming in sub-Saharan Africa. *World Bank Research Observer* **3**: 81–98

Boserup E 1965 *Conditions of Agricultural Growth*. Aldine Publishing Company, Chicago

Craig B J, Pardey P G, Roseboom J 1997 International productivity patterns: accounting for input quality, infrastructure, and research. *American Journal of Agricultural Economics* **79**(4): 1064–76

Hayami Y, Ruttan V W 1985 *Agricultural Development: An International Perspective*, rev. and expanded edn., Johns Hopkins University Press, Baltimore, MD

Heath J, Binswanger H 1996 Natural resource degradation effects of poverty and population growth are largely policy-induced: the case of Colombia. *Environment and Development Economics* **1**: 65–83

Kikuchi M, Hayami Y 1978 Agricultural growth against a land resource constraint: a comparative history of Japan, Taiwan, Korea, and the Philippines. *Journal of Economic History* **38**: 839–64

Pingali P 1990 Institutional and environmental constraints to agricultural intensification. In: McNicoll G, Cain M (eds.) *Rural Development and Population (Institutions and Policy)*. Population and Development Review, a supplement to Volume 15, 1989. Oxford University Press, Oxford, UK

Pingali P, Bigot Y, Binswanger H 1987 *Agricultural Mechanization and the Evolution of Farming Systems in Sub-Saharan Africa*. Johns Hopkins University Press, Baltimore, MD

Pingali P, Heisey P W 1999 *Cereal Crop Productivity in Developing Countries*. CIMMYT Economics Paper 99-03, CIMMYT, Mexico

Pingali P, Hossain M, Gerpacio R V 1997 *Asian Rice Bowls: The Returning Crisis?* CAB International, Wallingford, UK

Sidhu D S, Byerlee D 1992 *Technical Change and Wheat Productivity in the Indian Punjab in the Post-Green Revolution Period*. CIMMYT Economics Program Working Paper 92-02, CIMMYT, Mexico

Tiffen M, Mortimore M, Gichuki F 1994 *More People, Less Erosion: Environmental Recovery in Kenya*. Wiley, Chichester, UK

Turner B L II, Ali A M S 1996 Induced intensification: agricultural change in Bangladesh with implications for Malthus and Boserup. *Proceedings of the National Academy of Sciences* **93**: 14984–91

Turner B L II, Brush S B (eds.) 1987 *Comparative Farming Systems*. Guilford Press, New York

P. Pingali

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Population Composition by Race and Ethnicity: North America

In the three nations of North America—Canada, the United States, and Mexico—ethnicity has different connotations and is officially recorded using a variety of measures. Weber's (1978) definition of ethnic groups as those that share 'a subjective belief in their common descent because of similarities of physical type or of customs' permits various attributes to be used to characterize ethnic boundaries. Not only do these traits vary from place to place, but they change over time as well. In contemporary Mexico, the primary ethnic distinction is between the indigenous and non-indigenous populations. Statistics Canada records national origin, minority and aboriginal status, and religion. Finally, the US Census Bureau solicits race and Hispanic origin, and for a subset of the population, inquires about ethnic ancestry.

1. *Historical Perspective of North America's Ethnic Composition*

The peopling of North America has been shaped by the presence of indigenous Americans, European colonization, African slavery, voluntary immigration, and for Mexico, considerable emigration. Along with differential fertility and mortality rates, these factors have dramatically transformed the continent's ethnic landscape over time.

The size of the indigenous population of North America at the time of European colonization is unknown. Estimates for the pre-Contact population of Canada and the US range from 900,000 to 18 million (Snipp 1989), while similar estimates for Mexico range from 4.5 million to 25 million (Alba 1977). Despite the lack of consensus about fifteenth-century population levels, there is widespread agreement that in the wake of European colonization, indigenous depopulation was severe. Although 200,000 Native people were estimated to inhabit Canada when the French began settlement in the early seventeenth century, less than 100,000 remained by 1867 (Anderson and Frideres 1981). In 1548, 6 million indigenous people were estimated to inhabit Mexico, but by 1605, only 1 million remained (Alba 1977). Depopulation estimates for North America as a whole have been put at 1.75 million to 4.75 million between 1500 and 1890 (Snipp 1989).

As Native American population levels dropped, the arrival of Europeans, and to a lesser extent Africans, began a repopulation process that transformed the demography of North America. Until the twentieth century, Canadian immigration would be dominated by the arrival of English and French settlers, while in the US, English colonization followed by German, French, Dutch, Irish, and Scotch–Irish immigration was accompanied by the coercive importation of African slaves. Although Africans were a feature of colonial Mexican society, their shares were extremely small and highly concentrated geographically.

Despite some similarities in their peopling processes, the three nations of North America have experienced distinct immigration trends following the European colonization. In 1870—soon after Confederation and almost 400 years after Christopher Columbus arrived in the Americas—Canada's population was overwhelmingly European-origin; 61 percent of approximately 3 million claimed roots in the British Isles and 31 percent reported French ancestry (Anderson and Frideres 1981). In contrast, the US population of over 38 million was more ethnically diverse: nearly 13 percent were 'colored' (i.e., black) (US Census Office 1883).

Three waves of immigrants diversified Canada's population during the twentieth century. Around the turn of the century, the European origins of Canadian immigrants changed, but remained overwhelmingly white. From 1896 to 1914, 3 million new immigrants responded to government incentives to settle the western prairie provinces, and the majority were German, Slavic, or Scandinavian. In a second wave from approximately 1946 to 1961, Canada absorbed 2 million newcomers, mainly Europeans who were fleeing the aftermath of World War II. Finally, in the wake of immigration reforms after 1970, the share of European immigrants dropped sharply from 76 percent in 1966 to 39 percent in 1973, while arrivals from Asia jumped from 6 percent to 23 percent (Anderson and Frideres 1981).

Like Canada, the volume and national origin composition of US immigration flows has varied over time. In the early years of the republic, the population was largely of English, German, Irish, and African origin. When the legal importation of slaves was halted in 1808, the black population stood at over 16 percent of the total (US Census Bureau 1976). Concurrent and continued immigration for Italians, Russians, Poles, and Scandinavians boosted the white population during the nineteenth and twentieth century. Following the 1965 Amendments to the Immigration and Nationalization Act, immigrants from Asia and Latin America have dominated flows. As a result, the US has become both less white and black than it was at its founding.

Mexico's demographic evolution has differed in several ways from its neighbors to the north. After the decline of the indigenous population that characterized the early colonization, and the subsequent period of Spanish settlement and African slavery, the Mexican population grew slowly by natural increase, as it was untouched by the large-scale immigration to the US and Canada. From the sixteenth through the seventeenth centuries, admixture between Europeans, Indians, and Africans produced the *mestizo* population that characterizes Mexico today. Both fertility

and mortality rates were high throughout this period, and remained so well into the twentieth century. At the beginning of the nineteenth century, when the colonial period came to a close, Mexico's total population stood at 6 million—roughly equivalent to the indigenous population prior to colonization (Alba 1977). Mexico's indigenous population rebounded following a decline between 1900 and 1920 (as did the national population in the wake of the Mexican Revolution, emigration, and an influenza epidemic).

2. *Contemporary Population Composition in North America*

Recent changes in the ethnic composition of North America derive from three processes, namely immigration, intermarriage across race and ethnic lines, and differential fertility rates of natural increase. Inconsistent reporting categories over time preclude precise estimates of the magnitude and components of change, but general inferences about the relative importance of each are possible. Accordingly, we discuss the contemporary ethnic landscape of Canada, the US, and Mexico and conclude with a broad comparison of the primary axes of ethnoracial differentiation among the three countries.

2.1 Canada

Table 1, which displays Canada's population composition in 1961 and 1996, reveals both increasing ethnoracial diversification over the 35-year period and also persisting dominance of European ancestries. Despite comparability problems over time, the two most salient differences are the absence of 'Canadian' as an ethnic category in 1960, and the allowance for multiple ancestries. In 1996 nearly 18 percent of all residents identified as 'Canadian'; this figure approaches one in five if the aboriginal population is added. Between 1961 and 1996 the aboriginal population increased its share slightly, but remains a tiny share of the total population. In 1961, virtually all Canadians (97 percent) traced their origins to a European ancestry, and of these, nearly three in four identified as either British or French origin. As recently as 1996, the majority of ethnic Canadians identify with

Table 1
Canada: population by ethnic origin, 1961–96

	1961		1996	
Ethnic origins	*N*	%	*N*	%
Single origins	*18,238,247*	*100.0*	*20,272,095*	*66.5*
British Isles	7,996,669	43.8	3,267,520	10.7
French	5,540,346	30.4	2,683,840	8.8
Other European	4,116,849	22.6	3,742,895	12.3
Western European			1,126,095	3.7
Northern European			167,285	0.5
Eastern European			867,055	2.8
Southern European			1,376,935	4.5
Other European			205,525	0.7
Arab			188,435	0.6
Asian	121,753	0.7	1,968,465	6.5
West Asian			106,870	0.4
South Asian			590,145	1.9
East and Southeast Asian			1,271,450	4.2
African	32,127	0.2	137,315	0.5
Pacific Islands			5,765	0.0
Latin, Central and South American			118,640	0.4
Caribbean			305,290	1.0
Aboriginal	220,121	1.2	477,630	1.6
Canadian			5,326,995	17.5
Other	210,382	1.2	80,840	0.3
Multiple origins	*NA*	*NA*	*10,224,495*	*33.5*
Total	18,238,247	100.0	30,496,590	100.0

Sources: Statistics Canada, *Population by ethnic origin, 1996 Census*, www.statcan.ca.
Domininon Bureau of Statistics, 1961 Census of Canada, Population: Ethnic Groups (Bulletin 1.2–5).
Notes: From the 1961 data, the category 'Negro' has been assigned to African origins.

Table 2
United States: population by race and Hispanic origin, 1960–99

Group	1960		1990		1999	
	Total (thousands)	Percentage of population	Total (thousands)	Percentage of population	Total (thousands)	Percentage of population
White	151,932	84.7	188,307	75.7	196,043	71.9
Black	18,872	10.5	29,299	11.8	33,088	12.1
Hispanic[a]	6,900	3.9	22,372	9.0	31,265	11.5
Asian and Pacific Islander[b]	878	0.5	6,992	2.8	10,219	3.7
American Indian, Eskimo, and Aleut	524	0.3	1,796	0.7	2,022	0.7
Other	218	0.1	—	—	—	—
Total	179,323	100.0	248,766	100.0	272,637	100.0

Sources: US Bureau of the Census, Census of Population 1960, Vol. 1: Characteristics of the Population, Part 1: United States Summary, Table 56. US Bureau of the Census, *Resident Population Estimates of the United States by Sex, Race, and Hispanic Origin: April 1, 1990 to June 1, 1999*.
Notes: a The estimate of the 1960 Hispanic population is from Cary Davis, Carl Haub, and JoAnne Willette, 1983. 'US Hispanics: Changing the Face of America.' *Population Bulletin*, Vol. 38, No. 3, p. 8, Table 2.
b For 1960, the Asian population comprises the categories 'Japanese,' and 'Chinese,' and 'Filipino' only. In subsequent years, Koreans, Vietnamese, Asian Indians, Samoans, Guamamians, and Hawaiians were also included.

European ancestries; nearly half (48 percent) of all single origin ancestries reported in 1996 were from Europe.

Among Canadians who report a single ancestry, the major change in ethnic composition derives from shifts in the source countries of immigrants. Prior to 1961, fewer than 1 percent of Canadian residents derived from the Middle East, Latin America, or Asia, compared to 19 percent by 1996. Canadian immigration from these regions is relatively recent, which explains their small population shares at the turn of the twenty-first century. Among the million foreign-born Canadian residents admitted before 1961, less than 5 percent were from Asia, Africa, or Latin America. But, as the volume of immigration to Canada increased, so too has the representation of these groups among new arrivals. Should current immigration trends continue, Canada's ethnic mosaic will be very different in the middle of the twenty-first century. However, the homogenization of Canadians through assimilation and intermarriage may also swell the numbers who identify in nonethnic terms as 'Canadian.'

2.2 United States

Table 2 presents the ethnoracial composition of the US in 1960 and 1990 based on the last two censuses, and for 1999 based on Current Population Survey estimates. The broad groups are the official categories approved by the Office of Management and Budget. Like Canada, the US racial classification scheme evolved to reflect the rapid growth of the Hispanic population and increasing diversification of Asian populations. Although Hispanics can be of any race, the vast majority is considered white because *mestizo* has not been an acceptable racial category for administrative reporting purposes. In 1960, the Asian and Pacific Islander category included only Japanese, Chinese, and Filipinos, but when the racial classification was revised and expanded in 1980, Vietnamese, Asian Indians, Samoans, Guamanians, and Hawaiians were added as separate racial groups. In 1990, the US Census identified nearly 1,000 unique racial and ethnic codes, which included nearly 400 American-Indian tribes, hundreds of detailed Asian and Pacific Islander groups, and many more self-identified nationalities.

Changes in the US racial classification system notwithstanding, trends in ethnic composition are clear. The white share of the total population has been falling steadily, from 85 percent in 1960 to 76 percent in 1990 and 72 percent in 1999; the black share grew slowly from 10.5 percent in 1960 and remained stable around 12 percent during the 1990s, while Hispanics and Asians increased their population shares. Census projections indicate that the Hispanic population will surpass blacks by 2003 or sooner. Less than 1 percent of the 1960 US population was Asian, but by 1999 Asians comprised almost 4 percent of all US residents—a seven-fold increase since 1960. Although Asians are the fastest growing minority population in the US, their population share remains low for historical reasons, primarily due to laws restricting their admission as immigrants until 1968. Compared to blacks, Hispanics and Asians, the indigenous population of the US remains tiny. In 1960 less than 0.3 percent of the total population consisted of indigenous groups, and by 1990 this share had increased to 0.7 percent. Part of this increase is an artifact of changed reporting criteria that increased incentives and opportunities for individuals of mixed

Table 3
Ten most frequently cited ancestries, 1990

Ancestry[a]	Percentage of reported ancestries
German	19.7
Irish	13.1
English	11.0
Afro-American	8.1
Italian	4.9
American	4.2
Mexican	4.0
French	3.5
Polish	3.2
American Indian	3.0

Source: University of Minnesota, Integrated Public Use Microdata Series, 1990 Census 1 percent Sample
a The ancestry data includes multiple-ancestry reports. A person who described herself as 'French' and 'English' would be counted twice, once under each ancestry listing.

racial origin to self-identify as Native American (Snipp 1989).

Because the US reports race and ethnic origin separately, ancestry data are instructive about the impact of historical and recent immigration trends on the ethnic landscape. Table 3 reports the 10 most frequently reported ancestries in 1990. Multiple responses are double counted; therefore the white population is more heavily weighted because they tend to report more ancestries (resulting from intermarriage) and because inter-racial marriages are typically rare. Germans, Irish, and English are the top three most frequently reported ancestries, followed by African–Americans and Italians. Only 4.2 percent of the US population identified as 'American,' but this figure would be higher if multiple ancestries were not double counted. The prominence of Mexicans among the top 10 ancestries reflects the dramatic growth of this population through immigration since 1960. Between 1970 and 1995, for example, the Mexico-born population of the US increased eight-fold, rising from 750 thousand to 6.2 million. Currently, Mexico is the largest single source of US immigrants.

Census Bureau projections using moderate assumptions about mortality, fertility and intermarriage patterns and current immigration levels indicate that by the year 2020, 65 percent of the US population will be white and 13 percent black. However, the two groups most affected by immigration, Hispanics and Asians, are projected to comprise 16 and 7 percent, respectively, of the total population. This is a marked change from 1960, when Hispanics and Asians combined made up less than 5 percent of the total. In recognition that intermarriage will continue to blur ethnic and racial boundaries, plans for the US Census 2000 will allow Americans to identify themselves using multiple races.

2.3 Mexico

Mexican ethnoracial categories are less differentiated than either Canada or the US mainly because it is a country of emigration rather than immigration. In Mexico the key racial division is between the indigenous and *mestizo* populations, and use of an indigenous language is the only measure the government now uses to draw this racial distinction. This is because designating members of indigenous groups hinges much more on sociocultural characteristics (like language) than phenotype or precise blood lineages, as practiced in the US.

Table 4
Mexico: indigenous language speakers, 1900–90 population aged 5 and above

Year	Spanish	Indigenous language*	Total population	Percentage indigenous language
1900*	11,528,345	2,078,914	13,607,259	15.3
1910**	13,200,063	1,960,306	15,160,369	12.9
1930	11,632,660	1,185,162	12,817,822	9.2
1940	18,410,534	1,237,018	19,647,552	6.3
1950	19,373,423	795,069	20,168,492	3.9
1960	33,818,174	1,104,955	34,923,129	3.2
1970	45,113,823	3,111,415	48,225,238	6.5
1980	61,665,795	5,181,038	66,846,833	7.8
1990	65,279,855	5,282,347	70,562,202	7.5
1995	74,735,782	5,483,555	80,219,337	6.8

Sources: Instituto Nacional de Estadística, Geografía e Informática, Estadísticas Históricas de México, Volume 1 (1985), Table 2.4.
INEGI, *Conteo* '95 (1995), Table 6.
INEGI, *XI Censo General 1990* (1992), Table 6.
Pick, James G. and Edgar W. Butler. 1994. *The Mexico Handbook* (1994).
Secretaría de Industria y Comercio, Dirección General de Estadisticas, *VII Censo General de Población 1950* (1952), Table 6; *VIII Censo General 1960* (1962), Tables 3, 14 and 15; *IX Censo General 1970*, Tables 4, 17; *X Censo General 1980*.
* From 1950 forward, the 'indigenous language speakers' includes both those who do and do not speak Spanish. In earlier years, it is unclear if both groups are included.
** The 1900 and 1910 figures are for the entire population, not just those aged 5 and above.

Table 4 provides a glimpse of Mexico's indigenous population and its evolution since the turn of the twentieth century. The series reported in Table 4 is not totally consistent because the criteria used to enumerate the indigenous populations at times were restricted by age (5 years and over) and at times combined speakers of indigenous languages who did and did not speak Spanish. In 1990, the majority of the indigenous population spoke over 11 different languages, but Nahutl, Maya, and Zapoteco were the three most commonly spoken. Mexico's indigenous population is concentrated in the southern and least developed states, with Quintana Roo, Yucatan, Campeche, Chiapas, Oaxaca, Guerrero, Puebla, Hidalgo, and Tlaxcala housing the vast majority of indigenous peoples.

Despite changes in enumeration criteria, the historical record shows a clear decline in the share of indigenous people in Mexico's population. The indigenous population has grown strongly in absolute terms, but this is due partly to improvements in enumeration and coverage over the period. At the turn of the century, 15 percent of Mexicans were indigenous language speakers but by 1995 this share had been halved. In relative terms, the indigenous population of Mexico is shrinking. Mexico's declining linguistic diversity is due to social and economic integration of indigenous minorities through internal migration, urbanization, and modernization processes. Further shrinkage of Mexico's indigenous population is likely in the future, but linguistic diversity is likely to persist much longer than in Canada and the US owing to the larger population base and the persisting rural isolation of many groups.

2.4 Comparative North American Perspective

Canada, the US, and Mexico have some similarities and many differences in the composition and evolution of their ethnoracial landscapes. First, despite rising immigration to Canada and the US from Asia, Africa, and Latin America, both countries have witnessed an increase in the population shares that do not self-identify in ethnic terms, but report themselves as Canadian or American. Second, by comparison to Mexico, Canada and the US have much smaller indigenous populations. Third, unlike the US, Canada and Mexico have negligible shares of residents who trace their ancestries to Africa; owing largely to the legacy of slavery, blacks comprise over 12 percent of US residents. Finally, immigration has diversified the ethnoracial landscape of Canada and the US since the 1960s, but Mexico has witnessed a negligible impact of foreign admissions. As an emigrant country, Mexico's population composition will be driven by the assimilation of indigenous populations into the national population, a process that is shaped largely by social and economic change, along with internal migration and intermarriage patterns.

See also: Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity: Anthropological Aspects; Ethnicity, Race, and Health; Ethnicity, Sociology of; Gender, Class, Race, and Ethnicity, Social Construction of; Population Cycles and Demographic Behavior; Population Ecology; Race and Urban Planning; Race: History of the Concept; Statistical Systems: Censuses of Population

Bibliography

Alba F 1977 *La Población de México: Evolución y Dilemas*. El Colegio de México, Mexico City

Anderson A B, Frideres J S 1981 *Ethnicity in Canada: Theoretical Perspectives*. Butterworth & Co, Toronto

Bean F D, Tienda M 1987 *The Hispanic Population of the United States*. Russell Sage Foundation, New York

Davis C, Haub C, Willette J 1983 US Hispanics: Changing the face of the nation. *Population Bulletin*, Vol. 38, No. 3. Washington, DC: Population Reference Bureau

Instituto Nacional de Estadística, G. e. I. *1996 Conteo de Poblaciôn y Vivienda 1995: Resultados Definitivos, Tabulados Básicos*. Aguascalientes

Smith J P, Edmonston B (eds.) 1997 *New Americans: Economic, Demographic and Fiscal Effects of Immigration*. National Academy Press, Washington, DC

Snipp M C 1989 *American Indians: The First of This Land*. Russell Sage Foundation, New York

US Census Office 1883 Compendium of the Tenth Census (June 1, 1880). Washington, DC

US Census Bureau 1976 *The Statistical History of the United States From Colonial Times to the Present*. Basic Books, New York, Tables A 1–5 and A 91–104

Weber M 1978 *Economy and Society*. University of California Press, Berkeley, CA

M. Tienda and A. Morning

Population Cycles and Demographic Behavior

1. Population Cycles in the Context of Intellectual History

Population studies and economic theory have been developed and discussed in conjunction with one another since the early nineteenth century. At this time Thomas Robert Malthus devised the ingenious theory, influenced by religion and controversial in its time, that bears his name. Ever since then it has been assumed within the social sciences almost as a matter of course that a strong linkage exists between demographic and economic factors. Malthus elaborated his classical theory around the year 1800. There were few new theoretical contributions for a long time afterwards, and those that came forth were not especially

significant. Until well into our own day, therefore, Malthus' basic ideas dominated the scene. According to his theory the causal relation between population and the economy goes chiefly in one direction—population changes play only a passive role and are determined by changes in the means of subsistence. Furthermore, in his system a single demographic variable, i.e., mortality, is mainly influenced by the economy. Although Malthus points out that moral restraint and the postponement of marriage may counteract excessively strong population growth, mortality is the dominant force adapting population size to the existing means of subsistence.

By contrast, theoretical developments have proliferated rapidly during recent decades. Entirely new models have emerged to elucidate the interaction between the population and the economy. This has given rise to population economics as a new and vigorously expanding field of scholarly endeavor. Its reasoning is usually strongly rooted in traditional economic theory and often bears the imprint of Malthusian thinking. These developments have meant that the shaping of theory has taken quite another turn and has also become more diversified, with much greater breadth and depth than before. The major characteristics of modern population economics are discussed below.

First, interest has shifted from one demographic variable to another. The strong link which Malthus believed to exist between changes in the economic environment and mortality has been called into question. Mortality variations are influenced by a number of other factors, the means of subsistence and standard of living forming only one of these, and not always the most important one. Fertility—along with the factors which determine it, of which particular mention may be made of varying age at marriage, the use of birth control, abortion, and nutrition—has been regarded instead as being determined by economic factors in a way quite different from that found in Malthus. This does not apply only to modern society. Several studies have shown that even in preindustrial society, the link between fertility and economic factors was often stronger than the relation between mortality and the economy, in both the short and the long term.

Second, population variables have been linked more closely to economic theory, especially neoclassical microtheory, and have then usually been analyzed in a more stringent manner than is the case with the Malthusian theory. Here we may point to the 'microeconomic theory of fertility' or, as it is sometimes termed, the 'new household economics' or 'new home economics.' In this theoretical construction, fertility is viewed as the result of economically rational decisions (the demand for children). Established economic methods and conceptual mechanisms are employed in order to explain variations in fertility. For example, models are devised in which fertility is regarded as a function of income and the 'price' of children in relation to the price of other goods and services. In this reasoning children are regarded as ordinary goods, exchangeable in the choice situation for other goods and services. Fluctuations in fertility are explained using only prices and incomes. The pioneering theoretical work in this area was done by Gary Becker (1960) and was then elaborated, for example by Becker and Lewis (1973); it has often been cited as one of the basic explanations of fertility variations in the Western world since World War II. Models have also been formulated along highly neoclassical lines and have been based on a traditional investment approach: cost–benefit analysis. This method regards fertility as being determined by how costs and revenues associated with children change over time. From this standpoint, accordingly, the strong decline of Western fertility which set in at the end of the nineteenth century should be regarded as a consequence of the falling revenues and rising costs which having another child would entail for the parents.

Third, Malthus' view that there is a one-way causality between the population and the economy has been abandoned. What is now emphasized is that the interaction between the population and the economy is considerably more complex than Malthus believed. In many instances, demographic events can be one of the most important factors underlying the process of economic change. In other words, the population factor may also, in many cases, be the independent variable, which in its turn influences the economy and the means of subsistence. According to this approach, numerous other elements of social change must be viewed in the context of changing population size and composition (e.g., Boserup 1965, Simon 1977, Easterlin 1980).

Fourth, the various models which have been developed to represent the interaction between the population and the economy have become considerably more dynamic in character. One important element marking this trend has been the wide-ranging research conducted into the incidence and causes of various population cycles and into their effects on social change. Intensive discussions have been conducted concerning the extent to which population cycles with varying durations and strengths can be identified; what demographic or other factors can be cited as explanations of their genesis; and what effects the various population cycles have had on demographic behavior and on the national economy as a whole.

2. *The Mechanisms of Population Cycles*

The study of population cycles goes a long way back in time. Its origins can be traced back to the northern European countries in the nineteenth century. As early as in 1855 Ejlert Sundt, a Norwegian embryonic sociologist and statistician, observed that a strong rise

and fall in a country's birthrate is repeated during the next generation, though in a less marked fashion. On reaching marriageable age an extraordinarily large birth cohort will bring into the world more children than an age group of normal size. The same law applies in reverse: after one generation an undersized cohort will produce a new undersized group. As a result, regularly recurring waves or 'echo effects' are generated in the size and age composition of the population.

Ten years later, in 1865, the Swedish demographer Fredrik Berg paid tribute to Sundt for the idea of cohort analysis, and emphasized the importance of keeping different birth cohorts apart and studying them separately. He also argued that it ought to be possible to use variations in birth cohort sizes to compute the future number of births. Berg foreshadowed modern population projection techniques.

The Swedish demographer Gustav Sundbärg was a pioneer in expressing the view that a substantial change in birthrates may have far-reaching repercussions, not only on the economy but also on institutional and social change for a long time ahead. He may also have been the first to point out that *the size of a cohort* can have an impact not only on individuals' demographic behavior, but also on their behavior in numerous other spheres. Sundbärg developed these ideas in his analysis of the strong rise in the number of births in Sweden between 1815 and 1830, and its aftereffects. Little by little the rise in births transmitted itself to higher age levels and grew in strength compared with earlier cohorts because mortality was also falling dramatically. Sundbärg believed that this rise had a far-reaching influence in many areas of social and economic development in Sweden all through the nineteenth century. He wrote in 1903:

> As they reached school age these cohorts filled up all the school rooms so that their education and upbringing became a burning issue. As they reached working age they shattered the guilds (in the 1840s) and compelled the introduction of a more liberal trade legislation, without which these masses of people would have been unable to survive. As they reached marriageable age their large numbers enforced delayed marriage, and in the end (though not until the second generation) the great emigration movement. The immense class of 'all others' who were excluded from democratic representation at the time destroyed the constitution based on the four estates (in the 1850s and 1860s). And in our own day, having reached old age they fill our poorhouses and make the problem of disability and old age a matter of topical concern.

In 1937, the German economist August Lösch noticed that sudden and dramatic changes in the population structure, caused by war for example, generated long waves in population growth, which in turn brought about far-reaching changes in the economy. Great wars are characterized by a deficit of births during the war and a surplus of births in the immediate postwar period. In a now-classical article in the *Quarterly Journal of Economics*, 'Population Cycles as a Cause of Business Cycles,' Lösch described how these features caused a population cycle which in turn gave rise to strong fluctuations in German economic growth and in the business cycle.

This discussion of the origins of population cycles and their influence on demographic and economic factors transmitted itself to the American scholarly arena after World War II. American researchers had been able to identify long cycles in US economic activity since the middle of the nineteenth century. How these cycles were generated and propagated became the object of extensive empirical and theoretical studies by American scholars during the 1950s and 1960s. They found that rather than being a passive response to changes in the economy, population cycles played an active role in determining the long cycles that had been observed in US economic activity. These long cycles could be found not only in aggregate production growth but also in most of its subcomponents. This made it possible to identify long cycles in the growth of the labor force, in capital formation, in productivity, and in utilization of capacity. In particular, there was an extensive discussion on the existence of, and mechanisms underlying, Kuznets cycles—named after Simon Kuznets—or 'long swings,' as they are often called. Models were devised to show how changes in the age structure and population growth over about 20 years generated long-term fluctuations in investment activity, mainly in building construction, and corresponding fluctuations in economic growth. In explanations of the underlying mechanisms the primary stress was laid on the interaction between demographic and economic factors. In this way long cycles in marriage frequency, birthrate, and immigration were seen to give rise to long cycles in both the growth and the utilization of resources through their effect on such factors as capital formation, housing construction, and labor supply. The causal relation in the reverse direction was also important. Through the link with unemployment and effective demand, changes in the utilization of resources intensified immigration and gave rise to long-term cycles in the pattern of immigration. In the same way, the variations in unemployment influenced job opportunities for the labor reserve (mainly married women, ethnic minorities, and older workers). In other words, there was a long-term cycle in the size of the labor supply. A side result of this wide-ranging discussion of the interaction between demographic and economic factors was that economists increasingly incorporated demographic variables into their reasoning.

After World War II the long-term cycles assumed a different character both in duration and strength. A number of the suppositions applied to earlier periods of history were changed. Some of the mechanisms which formerly generated Kuznets cycles changed character and others weakened or vanished. In par-

ticular, the great influence that immigration to the USA had exerted in the generation of Kuznets cycles after 1850 was modified. When immigration was restricted during the interwar period, its cycle-generating function vanished. Instead, this function was replaced by the marked rise in the number of births during the 1940s and 1950s and by an equally marked decline during the next two decades (e.g., Kuznets 1958, Abramovitz 1961, 1968, Easterlin 1968).

2.1 The Easterlin Hypothesis

One of the most prominent figures in this earlier research into the genesis of population cycles and their influence on demographic and economic phenomena was the American economic historian Richard Easterlin. It was also Easterlin who drew attention to the changed demographic conditions that applied to the Kuznets cycle after World War II. It therefore seems quite appropriate that it would be Easterlin who made the most trail-blazing analysis of the causes of the dramatic changes in births in the USA after World War II and their effects on the demographic and economic behavior of individuals. A number of other features of modern social change must also be viewed, he argued, in the context of demographic events.

A group of people born during a given period are marked by the same code of values and attitudes in many respects. Easterlin believes it possible to identify an important underlying reason why values and attitudes may vary from one cohort to another. Accordingly he asserts that an individual's behavior in a number of areas will be influenced by the size of the cohort to which he or she belongs, since it determines economic situation and future prospects to such a large extent. At this point Easterlin ties in with earlier thinking by Gustav Sundbärg and others, but he conducts the analysis with considerably greater precision and stringency (e.g., Easterlin 1968, 1978, 1981, Artzrouni and Easterlin 1982.)

For young individuals who belong to a large cohort, prospects for the future appear somber. Their incomes will be lower in general and the likelihood of becoming unemployed is higher than for individuals belonging to a small cohort. As a consequence they marry later and also divorce more frequently. Besides they have a greater propensity to commit suicide and to manifest higher frequency of criminality. Married women in a large cohort will participate in the labor force more often because of the family's difficult economic situation. Unmarried women have high nonmarital fertility, also resulting from social stress factors. Alcoholism is also regarded as more common in a large cohort. Values, too (political opinions, attitude to work, savings, etc.), depend on cohort size.

One key element of Easterlin's reasoning is the 'relative income hypothesis.' This holds that the individual's behavior is determined by the relation between, on the one hand, the material standard and position in society which he or she has attained, and on the other hand, what he or she had expected. The expectations are in high degree determined by family background, i.e., by the parents' economic situation, while the possibilities of translating them into reality are dependent on the prevailing labor market situation and relative wages. In other words, it is not the absolute level of living standard that determines behavior, but the relative level, in which the crucial factor is the relation between expectations and the possibilities of realizing them (material aspirations). During a period when high aims are set but the possibilities of realizing them are low, as is the case with a large cohort, the social stress factors are magnified. The stresses and strains associated with large cohort size are likely to continue throughout the whole course of a cohort's life.

In a small cohort the reverse applies. Firms have to compete for the relatively few new entrants to the labor market, which means that starting wages become high and opportunities of advancement good. Social stress is reduced considerably, and divorce, suicide and criminal frequencies become low. Moreover, in a small cohort expectations are low, because they are determined by the parents' (the preceding large cohort's) economic situation. On the other hand the possibilities of realizing expectations—and even of surpassing them—are high. Here, too, it happens that these favorable circumstances follow the small cohort all through its life.

Easterlin makes a particular point of analyzing how cohort size influences childbearing. He observes that a large cohort will bear relatively few children per woman while a small cohort will give birth to many. The linkage between birthrates and adverse economic and social effects causes the emergence of 'crowding mechanisms' operating within three major social institutions: family, school, and labor market. The result of the 'crowding mechanisms' is that the appearance of social stress factors are quite different in large and small cohorts. A large cohort adapts in order to avoid deterioration of its living standard compared with that of its parents. One such adaptation is that family formation takes place later in life. Another is that a woman in such a situation finds herself compelled to combine labor force participation with childbearing, which reduces her natality. By means of such demographic adaptation the large cohort strives to maintain its economic status compared with that of its parents. For a small cohort the situation is again reversed. Its members have no need to make any demographic adaptation in order to preserve an economic status that is favorable from the outset.

These demographic mechanisms produce what is called the 'Easterlin effect,' and they have been used as an alternative to the explanation suggested by the 'microeconomic theory of fertility' for the large fertility variations observed in the USA and Western

Europe during the postwar period. The most essential difference between the two approaches is that, according to the Easterlin model, demographic behavior is determined by incomes and consumption relative to expectations, whereas in the neoclassical theory demographic behavior is determined by absolute incomes and actual levels of consumption. Where the neoclassical economists regard the cost of children as a response to exogenous changes in female wages, Easterlin regards tastes as response to exogenous changes in relative income status. Instead of assuming a constant preference structure as in Becker's theory, Easterlin introduces a shifting one; the shift being determined by incomes and prices. The Easterlin model is regarded by many as an attempt to introduce a sociological mode of thought into economic theory.

The observation that large cohorts would give birth to relatively few children and small cohorts to many children also creates the potential for continuing fluctuations in the relative size of birth cohorts. As a result there emerges a self-generating and to a certain extent self-correcting fertility cycle of around four to five decades in length and with associated changes in a wide range of social and economic phenomena induced by fluctuations in relative cohort size. Pre-World War II Kuznets cycles of 15–25 years caused by aggregate demand fluctuations would be replaced, according to Easterlin's argument, by post-World War II swings of double the length originating in fertility changes operating with a lagged effect on the age structure of labor supply. Some scholars have included such population cycles in mathematical models (e.g., Lee 1974, Samuelson 1976, Chu and Lu 1995). With the relative income hypothesis as a starting point, Easterlin describes *inter alia* important elements of social change in the USA during the postwar era, his main objective being to explain the baby boom of the 1940s and 1950s and the baby bust of the 1960s and 1970s. As a result of low birthrates during the interwar period—and of low immigration—relatively few young people entered the labor market during the 1940s and 1950s. Labor market conditions were favorable, since firms were competing fiercely for the limited numbers on the labor market by such means as increased starting wages. The future beckoned brightly too, for the prospects of advancement appeared good in the prevailing conditions of labor shortage. As a result of the fall in the proportion of young people in the total labor force during the 1940s and 1950s, the relative incomes of the young cohorts increased during this period at the same time as fertility was rising.

Conditions for the next generation were quite the opposite. Because of the high relative incomes of their parents and the subsequent rise in fertility, the new generation was comparatively large and at the same time entertained expectations of a high material standard. Thus, when the new generation emerged on the labor market during the 1960s and 1970s its expectations were high. Because of the cohort's size, starting wages were low, the likelihood of becoming unemployed was high, and opportunities for advancement seemed rather limited. The chances of fulfilling their high expectations therefore appeared small. As a consequence, families were formed later, and childbearing was reduced. In other respects, too, as regards marriage and divorce patterns for example, female labor force participation rates, nonmarital fertility, suicide and criminal frequencies, and political alienation (students riots in the late 1960s, etc.), the 1960s and 1970s exhibited characteristics quite different from those of the previous two decades. Easterlin also predicted that a repetition of the baby boom would take place during the 1980s and early 1990s. In reality, however, fertility remained largely unchanged in these small cohorts.

2.2 *Criticism of the Easterlin Hypothesis*

Though much admired, the Easterlin hypothesis has not been exempt from criticism. Some scholars contend that the analysis is conducted at far too high a level of aggregation. They assert that the information available on the individual level concerning numbers of children and relative wages at a given time does not substantiate the Easterlin hypothesis, though there also are micro studies that report findings consistent with it. A number of scholars maintain that the theoretical ground for believing that changed relative incomes should influence fertility, labor force participation rate, and so on, are not particularly well developed either. Easterlin is also said to fail to take account of women's contribution to family income. If men's relative incomes are low at the same time as the labor force participation rate of women is high, it is fully conceivable that household incomes—which are the relevant factor at this point—are even higher than if the man had a relatively high income and the woman participated little in the labor force.

Other scholars argue that Easterlin's view of the modus operandi of the labor market is too simple and not rooted in any accepted labor market theory. In particular, one has criticized the idea that substitutability between older and younger workers is as limited as the Easterlin hypothesis holds. Easterlin is even accused of being careless with empirical data. He is said to pay too little attention to alternative explanations and to let his hypothesis rest on fragile empirical foundations, underpinning it only by *one* observed cyclical covariance between age structure relative incomes, and fertility.

Furthermore, the efforts made by Easterlin to link varying economic conditions to individual behavior in fields other than that of childbearing have not had any profound impact on current research. The studies which have been carried out give incomplete or conflicting results.

Nevertheless, the view that generation size significantly affects the attitudes and actions of individuals in a variety of areas, especially fertility behavior, has formed the basis of a number of international studies. As noted above, the relationships postulated by Easterlin seem to fit rather well for the USA. Some other countries (Australia, New Zealand, Norway, Canada) show similar patterns. In other countries there is a connection, but it is considerably weaker (Finland, France, The Netherlands, England and Wales, Sweden). For certain countries (Germany, Italy, Denmark) no such connection has been found at all. It should be observed, however, that usually it is only the reduced relation between age structure and fertility that has been studied, which is a 'second best' solution in this connection. Empirical studies have had to be based on age-specific incomes of different age groups to test the hypothesis properly. (For an overview of the 76 empirical analyses of the 'Easterlin hypothesis' that have been undertaken so far, and also of the criticism against the hypothesis, see Macunovich 1998.)

The first link of Easterlin's key relation between age structure relative incomes and fertility, has shown itself to be the more tenable and has inspired some scholars to develop the basic lines of thought both empirically and theoretically. In Easterlin's original analysis the relation is constructed on the basis of a relatively simple model of how the labor market functions. Thus, Easterlin does not pay much attention to possibilities for substitution between younger and older workers, nor does he differentiate between educational categories. An increased supply of younger labor will, therefore, result automatically in reduced relative incomes and increased unemployment for that younger element, all other things being equal. Later studies have shown with greater precision and detail how changes in the age structure have influenced relative incomes, and have established *inter alia* that cohort effects on incomes have been stronger for higher-educated categories than for lower. It has further been shown that the cohort effect on incomes has a tendency to wane over time, and that it is essentially absent for female workers. This suggests that the possibilities of substitution between younger and older women are large.

Easterlin's thesis that economic conditions can vary very widely between different demographic groups, and that a group's size influences its relative economic situation, has been clearly substantiated. But studies concerning the period after 1980 suggest that the effect applies only under special conditions. According to Easterlin, the relative income of the young ought to increase when the 'baby bust' cohorts entered the labor market in the late 1980s. Yet what one could observe was a continued decline in the relative wages of the young. It seems that changes in the supply of male workers may be able to explain the pattern of wages prior to 1980, but that a change in the structure of labor demand—in particular an increase in the demand for more educated labor—offers a better explanation of wage patterns after 1980.

Easterlin himself has noted that the fortunes of a cohort depend on a number of other factors besides cohort size alone. One would therefore expect cohort size effects to dominate a cohort's experience only when other factors are relatively constant. This was the case in the period after World War II, when governmental monetary and fiscal policies stabilized growth of aggregate demand. In addition, the relative impact of immigration on the labor force was greatly reduced as a result of restrictive legislation adopted in the 1920s. In Easterlin's view the general slowdown of economic growth after 1973 and the widespread retreat from full employment as an objective of macroeconomic policy are more important than changes in cohort size as factors explaining observed changes in fertility during the most recent decades. Furthermore, high immigration and the rise in female labor force participation also tend to limit the importance of birth cohort size, and so will government welfare programs that offer policies of full employment. According to Easterlin this is the reason why stronger cohort effects can be observed for the USA than for European nations, since the latter have more general welfare systems (see also Pampel and Peters 1995).

Just the same, it has proved possible to single out one factor—cohort size—that underlies variations in relative income and in fertility during certain periods and in some countries after World War II. This finding is moderate in scope, but solid and robust in its moderation. Only in exceptional cases can it be adduced as a dominant explanation of the variations observed. Nevertheless, the 'Easterlin hypothesis' and the many studies it has inspired, has given us new insights into the kinds of different connections and responses of economic, sociological, demographic, and psychological character which generate population cycles.

See also: Baby Booms and Baby Busts in the Twentieth Century; Cohort Analysis; Fertility Change: Quantum and Tempo; Malthus, Thomas Robert (1766–1834); Population Cycles, Formal Theory of

Bibliography

Abramovitz M 1961 The nature and significance of Kutznets cycles. *Economic Development and Cultural Change* **9**: 225–48

Abramovitz M 1968 The passing of the Kutznets cycle. *Economica* **35**

Artzrouni M A, Easterlin R A 1982 Birth history, age structure, and post World War II fertility in ten developed countries: An exploratory empirical analysis. *Genus* **38**(3–4): 81–99

Becker G S 1960 An economic analysis of fertility. In: Universities-National Bureau, Committee for Economic Research. *Demographic and Economic Change in Developed Countries*. Universities-National Bureau Conference Series No. 11. Princeton University Press, Princeton, NJ

Becker G S, Lewis H G 1973 On the interaction between the quantity and quality of children. *Journal of Political Economy* **81**(2): S279–88
Boserup E 1965 *The Conditions of Agricultural Growth: The Economics of Agrarian Change Under Population Pressure*. Aldine, Chicago
Chu C T C, Lu H C 1995 Toward a general analysis of endogenous Easterlin cycles. *Journal of Populations Economics* **8**: 35–57
Easterlin R A 1966 Economic–demographic interactions and long swings in economic growth. *American Economic Review* **16**: 1063–104
Easterlin R A 1968 *Population, Labor Force and Long Swings in Economic Growth: The American Experience*. Columbia University Press, New York
Easterlin R A 1978 What will 1984 be like? Socioeconomic implications of recent twists in age structure. *Demography* **15**(4): 397–432
Easterlin R A 1980 *Birth and Fortune*. Basic Books, New York
Easterlin R A 1987 Easterlin hypothesis. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics 2*. The Stockton Press, New York, pp. 1–4
Easterlin R A, Schaeffer C M, Macunovich D J 1993 Will baby boomers be less well off than their parents? Income, wealth and family circumstances over the life cycle in the United States. *Population and Development Review* **19**(3): 497–522
Kuznets S 1958 Long secular swings in the growth of population and in related economic variables. *Proceedings of the American Philosophical Society* **CII**: 25
Lee R 1974 The formal dynamics of controlled populations and the echo, the boom and the bust. *Demography* **11**: 563–85
Macunovich D J 1998 Fertility and the Easterlin hypothesis: an assessment of the literature. *Journal of Population Economics* **11**: 53–111
Pampel F C, Peters H E 1995 The Easterlin effect. *Annual Review of Sociology* **21**: 163–94
Samuelson P A 1976 An economist's non-linear model of self-generated fertility waves. *Population Studies* **30**: 243–7
Simon J S 1977 *The Economics of Population Growth*. Princeton University Press, Princeton, NJ

R. Ohlsson

Population Cycles, Formal Theory of

A population cycle is a roughly periodic variation over time in the size or composition of a population. Here, composition refers to the component structure of the population, for example into distinct age classes. The formal theory of population cycles consists of mathematical tools that yield qualitative and quantitative information about the characteristics of cycles, such as their amplitude or period. There are three broad types of cycle: *transient* cycles are a response to a perturbation of finite duration such as a fertility transition, *externally driven* cycles are triggered and sustained by a persistent external perturbation such as an economic cycle; *internally driven* cycles are maintained by interactions within the population such as a feedback between cohort size and cohort fertility. This article discusses the theory for these three types of cycles, dealing mainly with age-structured populations, although other aspects of population structure are considered briefly.

1. Mechanisms Driving Population Cycles

To explain the different mechanisms that drive cycles, and the kinds of formal theory that are required, it is useful to begin with Lotka's classical demographic equation which relates births $B(t)$ at the present time $t > 0$ to past births, by summing the reproductive contributions $\phi(a, t)$ made by individuals of all ages a, as

$$B(t) = \int_0^t \mathrm{d}a\phi(a, t)\, B(t-a) + H(t). \qquad (1)$$

Here $H(t)$ specifies births to individuals who were present at $t = 0$, and immigration is ignored. The theory of this classical demographic model is discussed in *Population Dynamics: Theory of Stable Populations*; *Population Dynamics: Classical Applications of Stable Population Theory*; *Population Dynamics: Theory of Nonstable Populations*.

Transient cycles typically occur in a population whose vital rates, i.e., mortality and fertility, change over a time interval, say from $t = 0$ to $t = t_1$, and are constant thereafter. A practical example of such a perturbation is a decline from high fertility to replacement fertility that starts at $t = 0$ and is completed at $t = t_1$. After time t_1, the population is described by classical demographic theory and undergoes damped oscillations towards the classical stationary state. But the dynamics during the transition are also oscillatory and require a different treatment.

Externally driven cycles result from a persistent perturbation of vital rates. One example is a periodic variation over time in fertility that is generated by an exogenous force, such as an economic cycle in wages. Another example is an irregular, stochastic variation in fertility or mortality that produces an ongoing stochastic response that has cyclical components. Formally, such perturbations are described by an appropriate time-dependence in the function $\phi(a, t)$.

Internally driven cycles result from changes in vital rates that are driven by changes in population size and composition. In such cases, the net reproduction $\phi(a, t)$ has a functional dependence on the past trajectory of births, or on past vital rates. One example is a relationship between the fertility of a cohort in its childbearing years with the size of the cohort: for a substantive discussion of such relationships, see *Ferti-*

lity Transition: Economic Explanations; *Baby Booms and Baby Busts in the Twentieth Century*; *Population Cycles and Demographic Behavior*. Another example is provided by inverse projection in historical demography (Lee 1977), in which observed counts of births and deaths over time are used in conjunction with the Lotka equation to estimate the time-variation in the level of mortality rates. Here, the constraint of matching observed counts results in a feedback between the level of mortality and population counts (Wachter 1986).

The formal theory of cycles driven by each of these mechanisms is discussed below, followed by a brief consideration of future directions.

2. Transient Cycles

A perturbation such as a fertility transition may take decades to complete, and a formal theory must provide formulas for the birth sequence $B(t)$ both during and after the perturbation. Li and Tuljapurkar (1999b) provide an explicit if cumbersome description of the decaying population cycles that result from a perturbation. This theory shows that the population undergoes damped cyclical changes during and after the transition. The characteristics of these cycles have not been explored except in particular cases; but this analysis makes possible the computation of population momentum for all gradual transitions (Li and Tuljapurkar 1999a). It is worth noting that the analytical results for this case rest on properties of the roots of Lotka's equation. The relationship between these roots and the age-pattern of reproduction has been studied by approximate (Keyfitz 1985) and exact (Wachter 1984) methods that are also useful in the analysis of externally and internally driven cycles.

3. Externally Driven Cycles

An illuminating treatment of population cycles driven by external periodic variation in reproduction was given by Coale (1972), who showed that the population responds with periodic variation in the numbers and relative sizes of its age components. The challenge for the theory lies in the nonlinearity of eqn. (1) when $\phi(a, t)$ contains a periodic component. To see this, suppose that $\phi(a, t) = \phi(a) + \psi(a, t)$ with ψ being periodic in time t, and suppose that the resulting birth sequence in eqn. (1) has the form $B(t) = B + b(t)$, with b also a periodic function. Substitution of these forms into eqn. (1) shows that the unknown $b(t)$ depends on a nonlinear product $\psi(a, t-a)\, b(t-a)$. General results for nonlinear renewal eqns. of this type are not known. There are analytical approximations which employ an expansion in the amplitude of the periodic perturbation ψ; these describe the amplitude response of the population and the phase difference between population cycles and the external driving cycle (Tuljapurkar 1985). An interesting class of periodically perturbed Lotka models has been described and analyzed exactly by Kim and Schoen (1997), Schoen and Kim (1994, 1997). The latter work demonstrates that it is sometimes possible to solve exactly the homogeneous form of equation (1), i.e., with $H = 0$, for an explicitly time-dependent $\phi(a, t)$. Such analytical examples can be valuable as a source of insight into the dynamics, and as a check on approximate theories intended for less tractable problems.

The analysis of periodic perturbations suggests that a persistent stochastic perturbation of vital rates will also generate a superposition of cycles of population change. The dynamics of a population subject to stochastic perturbation are also nonlinear, so again the theory has relied upon approximations. Lee (1975) pioneered the application of time-series methods to the linearized analysis of stochastic perturbations. To go beyond linearization, e.g., if the perturbations are large or have strong positive serial correlations, a higher order approximation may be used as discussed by Tuljapurkar (1990), Tuljapurkar and Lee (1997).

4. Internally Driven Cycles

The formal analysis of internally driven cycles is quite distinct from the analyses used in classical demography, and draws on developments in nonlinear dynamics (for a review of modern nonlinear techniques for dynamics, see Guckenheimer and Holmes (1983)). To start with, it is important to recognize that there are different forms of internal feedback. One class of feedback is exemplified by cohort size effects on fertility: large birth cohorts go on to have lower fertility than their parents and the opposite for small birth cohorts (Lee 1974, 1976, Wachter and Lee 1989). A second class of feedback is generated by social norms or expectations: for example, individual fertility is influenced by the average fertility in the population (Kohler 2000). A third class of feedback arises through an interaction between a population and external variables—for example, economic conditions, environmental conditions, or resources such as land holdings—and these external variables in turn influence fertility or mortality (Bonneuil 1990, 1992, 1997, Chu 1998, 1999, Feichtinger and Sorger 1990, Feichtinger and Dockner 1990, Lee 1987, Milik and Prskawetz 1996).

The formal theory of such feedbacks shares several elements that may be illustrated by considering a nonlinear version of Eqn. (1). Suppose that the net reproduction $\phi(a, t)$ in that equation changes with the population structure: for example, the fertility of individuals aged a at time t depends on the size of that cohort relative to the size of their parents' cohort.

Such feedbacks result in a functional relationship of the form

$$\phi(a, t) = \phi(a)h(g, a, \{B(t-s), 0 < s < S\}). \quad (2)$$

Here the function h on the right describes the effect of the other birth cohorts, born as long as $S > 0$ years ago, on the fertility of the cohort aged a. The parameter g controls the strength of the feedback, for example, by controlling the gradient of h. Lotka's equation is now a nonlinear renewal equation, which can be written in abstract form as

$$B = \mathscr{L}(\mu, B). \quad (3)$$

The nonlinear operator $\mathscr{L}$ maps past values of births into their current value. This verbal description can be formalized suitably; for an example of the mathematical details see, e.g., Cushing (1978).

In most applications the nonlinear eqn. (3) has a constant solution, $E(\mu)$. The dynamic stability of this solution is examined by studying small deviations $x(t) = \{B(t) - E(\mu)\}$: stability means that $x(t)$ should decrease over time. An expansion of the nonlinearity in eqn. (3) yields a series of terms of the form

$$x = \mathscr{H}(\mu, x) + \mathscr{G}(\mu, x, x) + \mathscr{R}. \quad (4)$$

The operator $\mathscr{H}$ is a linear operator, $\mathscr{G}$ is quadratically nonlinear, and $\mathscr{R}$ involves higher-order nonlinear terms. Local stability is determined by the eigenvalues of the linear operator $\mathscr{H}$, and these in turn depend on the elasticity of the linear response (Wachter 1991a). In typical examples, the equilibrium is locally stable when $g < g^*$, say, but becomes unstable when $g \geqslant g^*$. Local instability means that small deviations from the equilibrium births will change with time as $\exp\{(p + iq)t\}$ where $p > 0$ and q are given by eigenvalues of the linear operator $\mathscr{H}$. The local change in $x(t)$ follows a cycle of period $(2\pi/q)$. The central question now is: when such local instabilities grow and experience the nonlinear forces in equation (4), will they turn into sustained, stable cycles? If they do, increases in the parameter g past the threshold value g^* will generate a sustained population cycle, called a *limit cycle*. Since $E(\mu)$ is still a constant solution, eqn. (3) (or its solution) is said to have undergone a bifurcation at $g = g^*$.

For any value of g above but not far from the threshold g^*, the nonlinear limit cycle may be represented by a superposition of periodic solutions to the linear part of eqn. (3). Numerical illustrations of such limit cycles are given by Frauenthal and Swick (1983). A systematic procedure for computing the amplitude, period, and stability properties of this limit cycle beyond the bifurcation point is described by Tuljapurkar (1987), and substantive applications that illustrate these computations are given by Wachter (1991b) and Chu (1998). These procedures may be extended to many other models with feedback, and provide an important part of the theory of population cycles. Other feedback mechanisms may result in a different approach to bifurcation (e.g., bifurcation in fertility level, Kohler 2000).

5. New Directions

The formal theory of cycles has moved considerably beyond classical demographic theory. The nonlinear bifurcation analysis discussed above can be considerably extended. Wachter (1994) has explored global properties of feedback models, addressing the behavior of cycles far above a bifurcation threshold. Bonneuil (1992), Prskawetz and Feichtinger (1995), and Tuljapurkar et al. (1994) show that increasing feedback can lead to more complex types of population cycles including aperiodicity and chaos. Purely internal feedbacks may not be strong enough to generate demographic cycles in human populations, but feedbacks mediated by social or external interactions may well be strong enough. The challenges that face the theory now lie in two areas: confrontation with data to test and modify theories, and the development of new models that incorporate a wider range of demographic dynamics and interaction. As yet, nonlinear cycles in human populations have not been examined in the realistic and rich way that biologists have explored cycles in the structure of laboratory populations (de Roos 1997, Cushing 1997).

Bibliography

Bonneuil N 1990 Turbulent dynamics in a 17th century population. *Mathematical Population Studies* **2**: 289–311

Bonneuil N 1992 Attractors and confiners in demography. *Annals of Operations Research* **37**: 17–32

Bonneuil N 1997 *Transformation of the French Demographic Landscape, 1806–1906*. Oxford University Press, Oxford

Chu C Y C 1998 *Population Dynamics: A New Economic Approach*. Oxford University Press, Oxford

Chu C Y C 1999 Predator-prey models with endogenous decisions. *Mathematical Population Studies* **8**: 55–71

Coale A J 1972 *The Growth and Structure of Human Populations: a mathematical investigation*. Princeton University Press, Princeton, NJ

Cushing J M 1978 Nontrivial-periodic solutions of some Volterra integral equations. *Lecture Notes in Mathematics* **737**: 50–66

Cushing J M 1997 Nonlinear matrix equations and population dynamics. In: Tuljapurkar S, Caswell H (eds.) *Structured Population Models in Marine, Terrestrial and Freshwater systems*. Chapman and Hall, New York

de Roos A M 1997 A gentle introduction to physiologically structured models. In: Tuljapurkar S, Caswell H (eds.) *Structured Population Models in Marine, Terrestrial and Freshwater systems*. Chapman and Hall, New York

Feichtinger G, Dockner E J 1990 Capital accumulation, endogenous population growth, and Easterlin cycles. *Journal of Population Economics* **3**: 73–87
Feichtinger G, Sorger G 1990 Capital accumulation, aspiration adjustment, and population growth: limit cycles in an Easterlin-type model. *Mathematical Population Studies* **2**: 93–103
Frauenthal J C, Swick K E 1983 Limit cycle oscillations of the human population. *Demography* **20**: 285–98
Guckenheimer J, Holmes P 1983 *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer, New York
Keyfitz N 1985 *Applied Mathematical Demography*. Springer, New York
Kim Y J, Schoen R 1996 Populations with sinusoidal birth trajectories. *Theoretical Population Biology* **50**: 105–23
Kohler H P 2000 Social interactions and fluctuations in birth rates. *Population Studies* **54**: 223–7
Lee R D 1974 Formal dynamics of controlled populations and the echo, the boom and bust. *Demography* **11**: 563–85
Lee R D 1975 Natural fertility, population cycles, and the spectral analysis of births and marriages. *Journal of the American Statistical Association* **70**: 295–304
Lee R D 1976 Demographic forecasting and Easterlin hypothesis. *Population and Development Review* **2**: 459–68
Lee R D 1977 Methods and models for analyzing historical series of births, deaths and marriages. In: Lee R D (ed.) *Population Patterns in the Past*. Academic Press, New York
Lee R D 1987 Population cycles. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave; A Dictionary of Economic Theory and Doctrine*. MacMillan, UK
Li N, Tuljapurkar S 1999a Population momentum for gradual demographic transitions. *Population Studies* **53**: 255–62
Li N, Tuljapurkar S 1999b The solution of time-dependent population models. *Mathematical Population Studies—J. Demog*. **7**: 311–29
Milik A, Prskawetz A 1996 Slow–fast dynamics in a model of population and resource growth. *Mathematical Population Studies* **6**: 155–69
Prskawetz A, Feichtinger G 1995 Endogenous population growth may imply chaos. *Journal of Population Economics* **8**: 59–80
Schoen R, Kim Y J 1994 Cyclically stable populations. *Mathematical Population Studies* **4**: 283–95
Schoen R, Kim Y J 1997 Exploring cyclic net reproduction. *Mathematical Population Studies* **6**: 277–90
Tuljapurkar S 1985 Population dynamics in variable environments. 6 Cyclical environments. *Theoretical Population Biology* **28**: 1–17
Tuljapurkar S 1987 Cycles in nonlinear age-structured models. I. Renewal equations. *Theoretical Population Biology* **32**: 26–41
Tuljapurkar S 1990 *Population Dynamics in Variable Environments*. Springer, New York
Tuljapurkar S, Boe C, Wachter K 1994 Nonlinear feedback dynamics in fisheries: analysis of the DeRiso–Schnute model. *Canadian Journal of Fisheries and Aquatic Science* **51**: 1462–73
Tuljapurkar S, Lee R 1997 Demographic uncertainty and the stable equivalent population. *Mathematics and Computer Modelling* **26**: 39–56
Wachter K W 1984 Lotka's roots under rescalings. *Proceedings of the Natural Academy of Sciences of the USA* **81**: 3600–4
Wachter K W 1986 Ergodicity and inverse projection. *Population Studies* **40**: 275–87
Wachter K W 1991a Pre-procreative ages in population stability and cyclicity. *Mathematical Population Studies* **3**: 79–104
Wachter K W 1991b Elusive cycles: are there dynamically possible Lee–Easterlin models for U.S. births? *Population Studies* **45**: 109–35
Wachter K W 1994 The Cohort Feedback model with symmetric net maternity. *Mathematical Population Studies* **5**: 25–44
Wachter K W, Lee R D 1989 U.S. births and limit cycle models. *Demography* **26**: 99–115

S. Tuljapurkar

Population Dynamics: Classical Applications of Stable Population Theory

1. Introduction

The theory of stable populations is presented in both continuous and discrete form in the article *Population Dynamics: Theory of Stable Populations*. Virtually any real population, whether contemporary or historical, deviates from stability, but the stable model produces a standard for comparison based on simplifying assumptions that allow us to handle data limitations. Some applications of the theory appear in the introductory article just mentioned and in other articles, including *Demographic Techniques: Indirect Estimation*; *Population Dynamics: Momentum of Population Growth*. The present article will review several applications within a structure that specifies the amount of information that is required. The applications are 'classical' in the sense that they are relatively well known.

It is helpful first to distinguish between two classes of models that are widely used: survival processes and renewal processes. The prototypical example of a survival process is the single-sex, single-decrement life table, described in the article *Life Table*. A survival process describes a cohort, real or synthetic, from age 0 until some terminal age ω. The main concepts are (a) the survival function, $l(x)$, namely the proportion of the cohort that survives to each age x, and (b) the mean waiting time or remaining life expectancy, e^o_x, at any specific age x.

A renewal process, by contrast, is a survival process with reproduction. With this component, the process generates successive cohorts that are (in general) larger than necessary simply to replace the persons who have died. The cohort interpretation is supplemented with the asymptotic implications of fixed and stable overall rates (of growth, mortality, and fertility), and a fixed and stable age distribution, none of which depends on initial conditions.

If a renewal process is in equilibrium—that is, if reproduction is exactly at replacement value, then the stable population is stationary. It has an intrinsic growth rate of zero, and one does not need explicit

age-specific birth rates to analyze it, for its overall birth rate is identical to the overall death rate. There are several classical applications and generalizations of this scenario, all of which can be regarded as applications of stable population theory because a survival process is a special stable renewal process.

A stable population is typically motivated as the asymptotic consequence of projecting the population indefinitely with fixed age-specific rates. Its overall rates and age distribution converge to functions that do not depend on the initial conditions. However, unless the intrinsic growth rate is zero, there is a fundamental implausibility in this asymptotic interpretation. If the intrinsic growth rate is positive, then the overall population size will increase exponentially, an outcome that cannot be sustained indefinitely in a real population. If the intrinsic growth rate is negative, then the overall population size will *decrease* exponentially until eventual extinction. In probabilistic terms, the fixed-rates assumption implies almost certain ultimate extinction or explosion (see *Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems*). It is probably more useful to regard the stable population parameters simply as a synthesis of the underlying age-specific probabilities or rates.

2. Applications Not Requiring Schedules of Rates

Applications of stable population assumptions can be organized according to the amount of information they utilize. Probably the simplest possible applications occur when the only parameter utilized is the intrinsic growth rate, r. If we assume stability between two times t_1 and t_2, and that the population size at those two times is n_1 and n_2, then

$$n_2/n_1 = e^{r(t_2-t_1)}$$

Solving for r, we get

$$r = (\ln n_2 - \ln n_1)/(t_2 - t_1)$$

We can then estimate the population size at an intermediate time, $t_1 \leqslant t \leqslant t_2$, to be $n_t = n_1 e^{r(t-t_1)}$. This formula, or variations upon it, is often used when the population is not stable, in the absence of better information. Another simple and well-known application is to determine the doubling time associated with a given value of r. Setting $n_2 = 2n_1$ and letting T be the time required for the population to double, we obtain $2n_1 = n_1 e^{rT}$, leading to $T = (\ln 2)/r = 0.6931/r$. For example, if the annual growth rate is 0.01, or 1 percent, a stable population will double in 69.31 years, or approximately seventy years. These and related applications are discussed by Hinde (1998, pp. 154–8), for example.

Suppose that in addition to the intrinsic growth rate, r, we also know the intrinsic birth rate, b (and hence the intrinsic death rate $d = b - r$). Keyfitz (1977, pp. 12–14) uses such information to estimate the number of persons who have ever lived on Earth, which is the same as the number of persons ever born. His strategy is to piece together long intervals of time within which stability (at different levels) is assumed. If the population is stable between t_1 and t_2, then the number of person-years lived in that time interval will be

$$PY = \int_{t_1}^{t_2} n_1 e^{r(t-t_1)}\, dt = (n_2 - n_1)/r$$

PY can be converted to the corresponding number of births by multiplying by b:

$$B = bPY = b(n_2 - n_1)/r$$

(If $r = 0$, then $n_1 = n_2 = n$, $\mathrm{PY} = n(t_2 - t_1)$, and $B = bn(t_2 - t_1)$.)

3. Applications that Assume Schedules of Rates

Another class of applications arises if we know r, b, and d, and do not know the survival and fertility schedules but are willing to make some assumptions about them. Coale and Demeny (1983) blended empirical life tables into four sets of so-called regional model life tables (north, south, east, and west) for males and females, keyed to the life expectancy e_0^o at birth and e_{10}^o at age 10 (see *Demographic Models*). They also developed plausible fertility schedules, varying according to the mean age of childbearing but adding up to one child, so that the entire schedule could be multiplied by a hypothetical total fertility rate (TFR). With this information, much of it hypothetical, they were able to generate sets of stable age distributions. It is then possible to estimate a wide range of population characteristics, using only a few pieces of summary information (along with a specification of region and sex). The summary information typically includes certain combinations of the following: a measure of growth, such as r; a measure of fertility, such as b, the gross reproduction rate (GRR), or the TFR; a measure of mortality, such as d or the average age at death; or a measure of the age distribution, such as the proportion of persons below age 20. Stable age distributions generated in this way can also be used to approximate real age distributions in settings of deficient data. Some methods for the computation of population parameters using the stable population model are described in *Demographic Techniques: Indirect Estimation*.

4. Applications that Require Schedules of Rates

A number of analyses of population dynamics have resulted from examining the sensitivity of population characteristics to disturbances of a stable population. These applications require knowledge of the age-specific schedules for mortality (say m_x for each age x),

of fertility (say f_x for each x), and for any other transitions in the model. For example, Preston (1974) took partial derivatives of the basic stable population formulas to estimate the effect of various age-specific mortality changes upon the intrinsic growth rate, birth rate, and age composition. He found that the effects could be represented as simple functions of the age pattern of the cumulative changes in the mortality rates. Probably the best known of these sensitivity analyses has been on the topic of population momentum, which in its simplest form measures the eventual effect on population growth of an immediate shift to replacement fertility (see *Population Dynamics: Momentum of Population Growth*).

The classical stable population model has no in- or out-migration, but several authors have used the model in innovative ways to evaluate the impact of emigration (Keyfitz 1971) or immigration (Lesthaeghe et al. 1988). Mitra (1990) considered how a population could maintain stationarity through a combination of below-replacement fertility and a steady stream of immigration. With assumptions about the ages and fertility of the immigrants, it is possible to determine the necessary size of the stream of immigration. This model is relevant, at least as a first approximation, to the current mix of below-replacement fertility and immigration in Europe and North America.

The classical formulation of stable population theory only allows individuals to be born, procreate, and die, but later developments permit individuals to move between several statuses or states during their lifetime, as in multistate models (see *Multi-state Transition Models in Demography*). In one obvious application the states represent regions of a territory (Feeney 1970). In another, Keyfitz developed a linkage between aging and career mobility in stable populations (1973; the title of this paper is misleading because he included stable as well as stationary populations). He supposed that advancement in status across a set of discrete categories is essentially based on the steepness of the age distribution. In a setting of high growth, age (as a proxy for experience) tends to produce earlier advancement because there are relatively few people at higher levels. In a setting of low growth, seniority is less valued, because there are relatively many people at higher levels, and promotion will take longer. The approach is again a type of sensitivity analysis, in which the age at reaching a sequence of discrete gateways is expressed as a function of stable population parameters (in particular, the intrinsic growth rate).

5. *Applications that Link Individuals to One Another: Kinship*

Assume now that we have access to complete schedules of survival and fertility. In continuous form, say that $l(x)$ is the probability that a woman will survive to age x and $m(x)\,\mathrm{d}x$ is the probability that a woman will have a daughter between ages x and $x+\mathrm{d}x$ (with $m(x) > 0$ for x in the reproductive age range $\alpha < x < \beta$ within the full range of ages $0 < x < \omega$).

Single-sex kinship provides a structure by which individuals are linked through lines of descent or ancestry. Two specific models for kinship structures will be described here. In the context of kinship, reference individuals (ego) are linked to their direct descendants, such as daughters; to progenitors, such as mothers; or to their lateral kin, such as sisters. Two-sex approximations are possible. The main outputs are the probability that a specific progenitor is still alive when ego is age a, and the expected number of descendants and lateral kin (in a specific category) who are still alive, or were ever born, when ego is age a. These are obtained from various convolutions of the survival and fertility functions.

The first kinship model to be described was developed by Goodman et al. (1974, 1975). The following exposition will be limited to surviving kin when ego is age a. It can be modified to refer to kin ever born simply by deletion of the outermost survival function (that is, setting it equal to 1).

Each kinship role is expressed with a different formula, but the strategy for generating these formulas is the same for all roles. The steps to link ego with a consanguineal relative are to identify the relevant sequence of births, and the ages of the respective mothers at those dates; to incorporate the fertility rates for those ages, the probability of surviving to those ages, and, for the final link, the probability of surviving to the time when ego is age a; and integrating over all allowable combinations of ages of the linked kin. Note that the birth of ego herself, and her survival to age a, are assumed.

For example, to link ego with her daughter, the relevant birth is that of the daughter, which occurred when ego was some previous age x. If the daughter is still alive when ego is age a, then the daughter's age is $a-x$. Thus ego experienced the fertility rate for age x, and the daughter survived to age $a-x$. Multiplying in reverse order and integrating over allowable ages x, the expected number of surviving daughters for a woman age a is

$$D(a) = \int_{\alpha}^{a} l(a-x)\,m(x)\,\mathrm{d}x$$

(If $a < \alpha$ then $D(a) = 0$ and if $a > \beta$ then $D(a) = D(\beta)$.)

The stable population probability that ego's mother is alive when ego is age a was originally calculated by Lotka (1931), and is useful for estimating adult mortality when some empirical information about the incidence of orphanhood is available. If the mother was age x when ego was born, then the probability that she is still alive a years later is the probability of surviving to age $a+x$, $l(a+x)/l(x)$. Obviously this

depends on the value of x, as well as a, and these survival ratios must be weighted by the probability that the mother was age x to age $x+\mathrm{d}x$ when ego was born. In a stable population these weights will be given by $W(x)\,\mathrm{d}x = \mathrm{e}^{-rx}l(x)m(x)\,\mathrm{d}x$ (from the characteristic equation, this is the density function for mothers' ages and integrates to 1). Therefore,

$$M_1(a) = \int_\alpha^\beta [l(a+x)/l(x)]W(x)\,\mathrm{d}x$$
$$= \int_\alpha^\beta \mathrm{e}^{-rx}l(a+x)m(x)\,\mathrm{d}x$$

is the probability that ego's mother is alive when ego is age a.

The expected number of surviving sisters when ego is age a involves a similar logic but will not be given here. Ego is linked to her sisters through their mother. It is necessary to distinguish between older and younger sisters because the survivorship of the mother is problematic after, although not before, the birth of ego. The formulas can be extended to aunts, nieces, first cousins (mother's sister's daughters), and indeed any category of kin.

A noteworthy feature of this application of the basic model of a stable population is that it requires only the basic survival function and maternity function, and a structure to link individuals. As an alternative to linkage by kinship, it should be possible to substitute an organizational structure or some form of vertical and horizontal connections, but we are not aware of any analogous applications.

The formulas are deterministic, rather than stochastic, and will not produce distributions. They assume that the birth of a child is independent of the number of children a woman has already had, that survivorship is independent of biological relationship, and that the population is homogeneous. Although these limitations are severe, it would be difficult to find data corresponding to more realistic assumptions. Calculations using these multiple integrals require some form of quadrature.

In a second application of kinship (Pullum 1982), counts are not conditional upon the survival of ego or any other kin, but are defined as 'eventual expected' numbers of kin. For example, the net reproduction rate

$$N = \int_\alpha^\beta l(x)m(x)\,\mathrm{d}x$$

is the average number of female births that will be produced, a generation later, from a single female birth. This is equivalent to the eventual expected number of daughters that a woman (ego) will have, defined at the point of ego's birth and the daughters' births.

The eventual expected number of sisters, defined at the point of ego's births and the daughters' births, can be referred to with the symbol S. N is not equivalent to any of the functions defined by Goodman et al. (1974) but S does appear within their scheme as the (expected) number of sisters (older and younger added together) ever born by the time ego has reached age $\beta-\alpha$, the maximum age at which a sister can be born. This formula superfluously conditions on the survival of ego, but allows for all sisters to be born.

It can be shown that N and S are sufficient to obtain the eventual expected numbers of any other types of kin. For example, an average of N^2 granddaughters (daughters of daughters) will be born, SN first cousins (daughters of sisters of the mother), SN nieces (daughters of sisters), and so on. The sex ratio at birth will also be constant in a stable population, and by including it one can easily extend to a two-sex kinship structure. For example, define s to be the ratio of males to females at birth, so that for every female birth there will be $1+s$ births of males *and* females. This implies that for every female child there will be $1+s$ children, so N is replaced by $(1+s)N$; for every sister there will be $1+s$ siblings, so S is replaced by $(1+s)S$. (Note that instead of one mother there will be exactly two parents.) The number of grandchildren (sons and daughters of sons and daughters) will be $(1+s)^2N^2$; the number of first cousins (sons and daughters of brothers and sisters of the mother or father) will be $2(1+s)^2SN$; the number of nieces/nephews (sons and daughters of brothers and sisters) will be$(1+s)^2SN$; and so on.

Like the applications given in Sect. 3 above, this second model of kinship only requires summary measures of the stable population (in this case S and N), and does not require any schedules of rates.

6. Conclusion

These applications have been presented largely at a conceptual level, without the specific data or numerical conclusions that might help to demonstrate their value (the bibliography and other entries in this encyclopedia describe several further examples). In each case, however, the original goal was to utilize stable population theory to draw approximate conclusions about real populations in situations where data are scarce. The model approximation has certain elegant formal properties and in effect permits the simulation of data that are generally not available for real populations. Indeed, significant improvement in the availability of demographic data is a likely reason why there have been relatively few new applications of the theory in recent decades.

Bibliography

Bourgeois-Pichat J 1971 Stable, semi-stable populations and growth potential. *Population Studies* **25**: 235–54

Cerone P 1987 On stable population theory with immigration. *Demography* **24**: 431–8
Coale A J 1981 Robust estimation of fertility by the use of model stable populations. *Asian and Pacific Census Forum* **8**: 5–10
Coale A J, Demeny P 1983 *Regional Model Life Tables and Stable Populations*, 2nd edn. Academic Press, New York
Espenshade T J, Bouvier L F, Arthur B 1982 Immigration and the stable population model. *Demography* **19**: 125–33
Feeney G M 1970 Stable age by region distributions. *Demography* **7**: 341–8
Goodman L A, Keyfitz N, Pullum T W 1974 Family formation and the frequency of various kinship relationships. *Theoretical Population Biology* **5**: 1–27. Addendum 1975 *Theoretical Population Biology* **8**: 376–81
Hinde A 1998 *Demographic Methods*. Arnold, London
Keyfitz N 1971 Migration as a means of population control. *Population Studies* **25**: 63–72
Keyfitz N 1973 Individual mobility in a stationary population. *Population Studies* **27**: 335–52
Keyfitz N 1977 *Applied Mathematical Demography*. Wiley, New York
Keyfitz N 1985 The demography of unfunded pensions. *European Journal of Population* **1**: 5–30
Keyfitz N, Nagnur D, Sharma D 1967 On the interpretation of age distributions. *Journal of the American Statistical Association* **62**: 862–74
Lapkoff S F 1991 A research note on Keyfitz's 'The demographics of unfunded pensions.' *European Journal of Population* **7**: 159–69
Lesthaeghe R, Page H, Surkyn J 1988 Are immigrants substitutes for births? Unpublished paper, European University, Florence. Published in German in 1991 as Sind Einwanderer ein Ersatz für Geburten? *Zeitschrift für Bevölkerungswissenschaft* **3**: 281–314
Lotka A J 1931 Orphanhood in relation to demographic factors: A study in population analysis. *Metron* **9**: 37–109
Mitra S 1983 Generalization of the immigration and the stable population model. *Demography* **20**: 111–15
Mitra S 1990 Immigration, below-replacement fertility, and long-term national population trends. *Demography* **27**: 121–30
Preston S H 1974 Effect of mortality change on stable population parameters. *Demography* **11**: 119–30
Pullum T W 1982 The eventual frequencies of kin in a stable population. *Demography* **19**: 549–65
Schoen R 1986 The two-sex multiethnic stable population model. *Theoretical Population Biology* **29**: 343–64

T. W. Pullum

Population Dynamics: Mathematic Models of Population, Development, and Natural Resources

Although the importance of the links between population and environment is widely appreciated, there is little agreement in the literature about the nature of these links.

The dynamic link between population and natural resources—influences are never entirely unidirectional or constant over time—constitutes the main difficulty in assessing formal models and unified theories. To complicate matters further, this interrelationship depends on the level and type of economic activities, technological development, institutional settings and cultural systems. Even the variables themselves are not uniquely defined, but depend rather on the level and purpose of the analysis.

This article does not survey numerical simulation models of demographic/economic/environmental interactions (see Sanderson 1994 for a review), of which the World 3 Model (Meadows et al. 1972) is the most popular. Instead, it looks at some simple mathematical models that capture stylized facts of the link between population and natural resource dynamics. Such 'mathematical cartoons' help us to explain and understand how various population and environmental characteristics might affect each other, and they can teach us how to respond most effectively to various demographic and environmental developments.

Although it was exhaustible resources that first received theoretical attention, this article concentrates on renewable resources since these are the most interesting and least understood as regards changes in population structures and dynamics. In contrast to nonrenewable resources, the use of which is controlled largely by market prices, renewable ones are very often open-access resources. The overuse of resources if property rights are absent or poorly defined is called the 'first tragedy of the commons' (Hardin 1968) and can be corrected by restricting access to the resource. But as emphasized by Lee (1990), if the population is not fixed—as is presupposed in the models presented in this article—free access through reproduction constitutes the 'second tragedy of the commons.'

1. *Malthusian Population Dynamics and Natural Resources*

The study of interactions between population and resources has a long history. According to Malthus, population growth reduces material welfare due to diminishing returns to labor on a fixed supply of land. On the other hand the higher the level of material welfare the higher the population growth rate will be. The Malthusian model predicts that 'population will equilibrate with resources at some level mediated by technology and a conventional standard of living' (Lee 1986). Improvements in technology will be offset in the long run by increases in the size of the population, but the standard of living will not be related to the level of technology. As such, the Malthusian model provides a description of a rather primitive society with incomes not too far above the subsistence level and where local renewable resources are an important part of the economic production process.

Renewable resources (agricultural land, forests, lakes, etc.) are not in fixed supply as Malthus assumed. Renewable resources regenerate, but if the rate of

utilization (harvest) exceeds the rate of regeneration, a renewable resource will be depleted or, in the extreme case, irretrievably exhausted. By adding the dynamics of renewable resource growth to the dynamics of population growth, the Malthusian model is capable of explaining patterns of population growth and resource degradation that do not necessarily end up in a single equilibrium (Malthusian trap) (see Brander and Taylor 1998 and Prskawetz et al. 1994).

The structure of these models can best be described in terms of prey–predator dynamics, with the resources, R, being the prey and the population, P, acting as the predator. The dynamics of renewable resources (Eqn. (1)) are commonly described by the standard model of mathematical bioeconomics, where the net growth of the renewable resource, $\mathrm{d}R(t)/\mathrm{d}t$, is affected by two counteracting factors: indigenous biological growth, $g(R(t))$, and the harvest, $H(R(t), P(t))$, which depends on the stock of resources available to be harvested and the number of people who are harvesting. Indigenous resource growth is modeled by the logistic growth function $g(R(t)) = aR(t)(K - R(t))$, where the coefficient K determines the saturation level (carrying capacity) of the resource stock (i.e., K is the stationary solution of R if the resource is not degraded) and parameter a determines the speed at which the resource regenerates. The functional form of the harvest function is determined by the prevailing economic structure, and it establishes the link between the stock of resources and population. The population growth rate, $(\mathrm{d}P(t)/\mathrm{d}t)/P(t)$ (Eqn. (2)), is modeled as an increasing function of material welfare $y(t) = y(P(t), H(t))$, which is determined by the level of the harvest and will be reduced by population growth. Whenever material welfare falls below the subsistence level, the population will decline. These Malthusian population dynamics imply that population growth may well adapt to resource constraints in contrast to models with exogenous positive population growth, where the economy collapses if resources do not regenerate quickly enough.

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = g(R(t)) - H(R(t), P(t)) \tag{1}$$

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = n(y(t))P(t). \tag{2}$$

Within this class of models it is possible to investigate how the equilibrium between resources and population stock changes, dependent on the functional relations that govern the dynamics of population and resources. For instance, the efficiency of harvest technology, the degree of substitution between labor and resources, the indigenous rate of resource growth and the carrying capacity of the resource stock will have an effect on material welfare and hence on population growth. In turn, fertility and mortality will affect the resource dynamics via the input of changing labor stocks in the harvest. An empirical calibration of the model for the case of the small Pacific Easter Island characterized by pronounced population fluctuations was carried out by Brander and Taylor (1998).

Contrary to the Malthusian predictions, increasing population densities might be beneficial, as argued by Boserup (1981), and could well increase the human carrying capacity of the earth. Higher population densities will initiate technological innovations in agriculture, thereby increasing the yields so that the natural environment can support a larger population without reducing the level of welfare. Similar positive feedback mechanisms are captured in a simple mathematical cartoon of the interdependence between the growth in population $P(t)$ and carrying capacity $K(t)$ (expressed in numbers of individuals) by Cohen (1995) as

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = rP(t)[K(t) - P(t)] \tag{3}$$

$$\frac{\mathrm{d}K(t)}{\mathrm{d}t} = c\left[\frac{\mathrm{d}P(t)}{\mathrm{d}t}\right] \tag{4}$$

with $r > 0$ and c either negative, zero, or positive. The parameter c, which captures the effect of an increment of population on the carrying capacity, determines the long-term population dynamics, and it can represent technological innovation. When $c = 1$ population size grows exponentially (Euler, eighteenth century); when $c < 1$ population grows logistically (Verhulst, nineteenth century); and when $c > 1$ population grows faster than exponentially (Forester and co-workers, twentieth century).

Malthusian results (in the sense that population will grow to the point where material welfare matches subsistence demand) can be further undermined by allowing population growth to be a choice variable. Zero population growth may well be the optimal choice for individuals in the economy. Since environmental changes are often slow over the course of an individual life span, and since environmental damage may outlive its perpetrators, overlapping generation models (Eckstein et al. (1988)) provide an appropriate demographic structure. Intra- as well as intergenerational conflicts can be modeled in such a framework, taking account of the effects of increases in population and resource exploitation on the future population's quality.

The models presented so far represent traditional societies in which populations derive their living from primary occupations (agriculture, hunting, fishing, etc.) which depend on the availability of resources. But as societies become less bound to the land, more urbanized and more technological, they not only use the environment as a source of natural resources but also as a dump for waste products arising from human activity. Furthermore, an economy's production possibilities are no longer determined by the maximum

sustainable yield of renewable resources. Improvements in technology can increase the sustainable yields or reduce the resource stock required for production, and economic growth will allow for the use of man-made capital in place of natural resources. In open economies with trade, technological change, and economic growth, there is no simple and direct relationship between population growth on the one hand and the environment on the other hand.

Resources might still be at risk, albeit no longer by positive population growth alone. The risk can stem from the environmental costs of consumption and production. It is therefore of great importance to understand how these environmental impacts depend on different population structures.

2. The Impact of Population on the Environment

In a seminal paper, Ehrlich and Holdren (1971) express the impact on the environment, I, (e.g., the value of some pollutant) as the product of population, P, and per capita impact, $F(P)$:

$$I = PF(P) \tag{5}$$

Depending on whether diminishing returns or economies of scale are dominant, per capita impact can be increasing or decreasing in the total population. The most familiar form of such a demographic impact model is the so called I-PAT identity, which divides per capita impact $F(P)$ into affluence A, and technology T. Affluence can be measured as GNP per capita, and technology represents the mean impact of an extra dollar of production (impact/GNP). The latter variable encompasses the cleanliness of technology, the effects of institutions, and other factors which can alter the impact efficiency of production.

Demographic impact models (Wexler 1996) constitute multiplicative identities aimed at decomposing the effect on the environment of population on the one hand and that of consumption and technology on the other. The number of variables on the right-hand side of Eqn. (5) can be any number greater than two. Although these models are commonly used, they face several criticisms. They suppress any feedback relation between the variables on the right-hand side; their implications are highly sensitive to the choice of decomposition variables; and the method of decomposition and the role of institutions and culture are ignored. Recent research on the I-PAT identity has dealt with some of these problems. Expressing the identity in terms of variances rather than means (Preston 1996) allows one to incorporate the links between the decomposition variables. Considering households instead of individuals as the unit of consumption (MacKellar et al. 1995) highlights the effect of economies of scale as household size increases. Typical results of such analyses are that the population has a much stronger impact on the environment in developing countries than in developed countries. In the case of environmental hazards produced by industrial processes, population growth plays a minor role both in developing and in developed countries.

Milik et al. (1996) embed the I-PAT identity into a dynamic model where population P, affluence A, and technology T influence each other and vary over time. Pollution I reduces the regenerative capacity of natural resources, which in turn influences the dynamics of population and affluence. The model highlights the fact that understanding and monitoring the impact of pollution on the environment is important even when outward signs of damage (as a reduction in the resource stock) are not yet visible. Ecosystems often absorb stresses over long periods of time and then eventually reach a disruption level at which the cumulative consequences of stress appear in critical proportions. A policy of waiting until the first signs of environmental deterioration are observed before actions are taken to reduce the impact on the environment may well prove to be disastrous. When action is initiated, the impact on the environment might already be irreversible. Mathematically, these systems can best be described in terms of 'slow–fast' dynamics, where during certain periods some variables vary much more rapidly than others. While population and economic variables can be regarded as slow moving variables, the environment changes most rapidly once the system has absorbed too much stress. For instance, the emission of carbon dioxide occurred almost unobserved for several decades before any attention was paid to the impact of greenhouse gases on global warming.

It becomes difficult to quantify the impact of population on the environment if the demand for local resources is not related to local population growth and if institutional environments are not stable. In fact, 'commercialisation by converting a limited and inelastic subsistence demand to a limitless and elastic export demand (from the standpoint of the region, at least) can lead to much more rapid rates of exploitation than would be implied by population growth alone.' (Repetto and Holmes 1983).

3. Modeling Environmental Influences on Population Dynamics

Existing models of environmental influences on population exhibit two shortcomings. First, they generally do not distinguish between the separate effects the environment may have on fertility and mortality. Second, the environmental impact on population growth is commonly modeled to work only through the economic and social variables of the model. There are only two examples in the literature that model the direct effect of environmental stress on fertility and mortality.

The first example (Nerlove 1991) is based on the vicious circle of poverty, population growth, and environmental degradation as evidenced in small rural communities of sub-Saharan Africa. In these economies simple tasks like fetching water and collecting fodder and wood are mostly carried out by children. This means that children are not only valuable to parents for future income but also as a source of current income. Higher population growth leads to more resource depletion, which reduces the marginal productivity of the resource. To offset this effect families have more children, thus depleting the resource even more. In a short parable Nerlove demonstrates that increasing the birthrate when the environment deteriorates will be the optimal choice for a household that maximizes per capita harvest when 'the perceived marginal product of an additional child increases as environment deteriorates.' In addition, in a system in which land rights are acquired through cultivation, a large number of children can imply increased claim to land ownership. This positive feedback between fertility and environmental degradation is in stark contrast to Malthusian dynamics, where population equilibrates with resources. Since the marginal private and social costs of reproduction are not the same, Nerlove then shows how social interventions in the form of taxes or subsidies can be used to induce a specific birthrate corresponding to a socially desirable stationary state.

The underlying mechanism of the above-mentioned vicious circle is neither population growth nor environmental degradation but poverty, which prevents the substitution of alternative fuel sources, and the low status of women and girls, which devalues the large amount of time and effort that they must devote to daily gathering activities. To escape this vicious circle one needs to alleviate poverty and to educate women.

As suggested by Chu (1998), the increasing specialization and level of economic activity typical of advanced societies may also imply a substantial dependency of population dynamics on the state of the environment. Increasing the level of human activities destroys the earth's biodiversity, so that the ability of the environment to absorb negative shocks will decline, while increased specialization makes people less adaptive to environmental shocks. Chu suggests a simple mathematical model of the survival probability $f(\cdot,\cdot)$ to capture both effects:

$$f(m_t, x_t) < f(m_t', x_t) \quad \text{if } x_t > 0 \tag{6}$$

$$f(m_t, x_t) \geqslant f(m_t', x_t) \quad \text{if } x_t \leqslant 0 \tag{7}$$

where m_t denotes the logarithm of the number of intermediate goods (a measure of the division of labor) with $m_t < m_t'$ and x_t is a realization of the environment value X_t with the critical state $x_t = 0$. The mean and the variance of the random variable X_t will be increasing functions of m_t. By reducing the ability to adapt, increased specialization m_t' will decrease the survival probability if the environment is already in a disastrous state $x_t < 0$, while the opposite relation will hold if the environmental state is favorable $x_t > 0$.

4. Concluding Assessment of the Literature

Although there is a growing awareness of the population/environment linkage, as evidenced by the increasing number of empirical (case) studies, research has not progressed much in the Malthusian spirit of modeling the complex interaction in simple mathematical cartoons. Even worse, dynamic interactions between population and environment are often neglected—or increasingly complex simulation models are set up where anything can happen.

Further research on this topic is urgently needed. The concern should be shifted away from physical limits to growth towards the ability of social units to respond to environment/population linkages. Although they are more difficult to formalize, such models should also include organizational and institutional issues. The study by Angelsen (1999) provides a first step in this direction.

See also: Development, Economics of; Development: Sustainable Agriculture; Environmental and Resource Management; Environmental Challenges in Organizations; Environmentalism: Preservation and Conservation; Malthus, Thomas Robert (1766–1834); Sustainability Transition: Human–Environment Relationship; Sustainable Development

Bibliography

Angelsen A 1999 Agricultural expansion and deforestation: modelling the impact of population, market forces and property rights. *Journal of Development Economics* **58**: 185–218

Boserup E 1981 *Population and Technological Change: A Study of Long-Term Trends*. University of Chicago Press, Chicago

Brander J A, Taylor M S 1998 The simple economics of Easter Island: a Ricardo–Malthus model of renewable resource use. *The American Economic Review* **88**(1): 119–38

Chu C Y C 1998 *Population Dynamics: A New Economic Approach*. Oxford University Press, New York

Cohen J E 1995 Population growth and earth's human carrying capacity. *Science* **269**: 341–6

Eckstein Z, Stern S, Wolpin K I 1988 Fertility choice, land, and the Malthusian hypothesis. *International Economic Review* **29**(2): 353–61

Ehrlich P R, Holdren J P 1971 Impact of population growth. *Science* **171**: 1212–7

Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8

Lee R D 1986 Malthus and Boserup: a dynamic synthesis. In: Coleman D, Schofield R (eds.) *The State of Population Theory: Forward from Malthus*. Basil Blackwell, Oxford, UK pp. 96–130

Lee R D 1990 Comment: the second tragedy of the commons. *Population and Development Review* **16** (Suppl.): 315–22

MacKellar F L, Lutz W, Prinz C, Goujon A 1995 Population, households, and CO_2 emissions. *Population and Development Review* **21**(4): 849–65
Meadows D H, Meadows D L, Randers J, Behrens W W 1972 *The Limits to Growth*. Universe Books, New York
Milik A, Prskawetz A, Feichtinger G, Sanderson W C 1996 Slow–fast dynamics in Wonderland. *Environmental Modeling and Assessment* **1**: 3–17
Nerlove M 1991 Population and the environment: a parable of firewood and other tales. *American Journal of Agricultural Economics* **73**: 1334–47
Preston S H 1996 The effect of population growth on environmental quality. *Population Research and Policy Review* **15**: 95–108
Prskawetz A, Feichtinger G, Wirl F 1994 Endogenous population growth and the exploitation of renewable resources. *Mathematical Population Studies* **5**(1): 87–106
Repetto R, Holmes T 1983 The role of population in resource depletion in developing countries. *Population and Development Review* **9**(4): 609–32
Sanderson W C 1994 Simulation models of demographic, economic, and environmental interactions. In: Lutz W (ed.) *Population–Development–Environment: Understanding their Interactions in Mauritius*. Springer, Berlin, pp. 33–71
Wexler L 1996 Decomposing models of demographic impact on the environment. *Working Paper* WP-96–85, IIASA, Laxenburg, Austria

A. Fürnkranz-Prskawetz

Population Dynamics: Momentum of Population Growth

The momentum of population growth refers to the tendency of continued growth of a population with a history of high fertility after its fertility has fallen to replacement level. A population with sustained high fertility is growing in size and has a 'young' age distribution, that is it has a high proportion of young people. If such a population adopts replacement-level fertility, the population continues to grow before its size stabilizes. The eventual population is a stationary population, i.e., one whose size and age distribution remain constant over time. This ultimate stationary population has an older age distribution than the initial population, and its size is larger. The ratio of the eventual population to the pretransition population size is defined as the momentum of population growth.

1. *Momentum Under Specific Assumptions*

In a seminal paper, Keyfitz (1971) presented a simple mathematical expression for the population momentum under specific assumptions. Keyfitz assumed that the initial population has a stable age distribution, and that its fertility falls instantaneously to replacement-level fertility. The drop in fertility level is assumed to occur proportionately at all ages so that the age pattern of fertility remains unchanged. Mortality is assumed to be constant at all times. Under these assumptions, the population momentum M can be expressed as

$$M = \frac{be_0}{r\mu}\frac{R_0 - 1}{R_0} \tag{1}$$

where R_0 is the net reproduction rate (NRR), b is the birth rate, and r is the growth rate, all of the initial stable population, e_0 is the expectation of life at birth and μ is the mean age at childbearing in the stationary population, i.e., the mean of the net maternity function. Equation (1) provides a simple, closed-form expression that permits calculation of an exact population momentum from a few known summary measures of the initial stable population. This avoids the need to project the initial population for a long time until it stabilizes at zero growth as was done by Frejka (1973).

In observed populations, each of the four assumptions is rarely satisfied, and the resultant momentum will be larger, due mainly to delayed fertility transition and to decline in mortality. The Keyfitz momentum M in Eqn. (1) can then be considered as a lower bound of the extent of future population growth. The formula has been used widely, and it has been shown that initial populations with high growth rates produce a substantial population momentum, frequently around 1.6. That is, the initial population grows by 60 percent before its size stabilizes at a constant value even when the most optimistic fertility reduction scenario is assumed.

2. *The Anatomy of Population Momentum*

The continuation of population growth after an instantaneous drop in fertility to replacement level is due to the large number of women at childbearing ages in the initial population. To understand what 'causes' the growth of a population after a fertility drop, the population momentum can be written as a product of two ratios: (a) the ratio of the numbers of births in the ultimate to the initial population, and (b) the ratio of ultimate to initial population size when both have the same number of births. The first factor is the stable-equivalent number of births, and its value is approximately equal to $1/\sqrt{R_0}$, a value less than one for an initially growing population. The second factor is given by be_0, and is approximately equal to R_0. Since the product of the two ratios yields the population momentum, the population grows approximately by a factor of $\sqrt{R_0}$ if the pretransition population had an NRR of R_0 (Frauenthal 1975, Kim et al. 1991). Thus the growth of the population is due entirely to the eventual old age distribution relative to the initial young age distribution, but is *not* due to the larger number of births after the fertility drop. The number of births drops when fertility rates drop in spite of the large number of young people in the initial population.

Preston (1986) showed that when fertility drops to replacement level, the population segment below some age near the mean age at childbearing remains virtually constant, so that all of the growth is confined to the population segment above that age, typically around 28. Inspired by this insight, Kim et al. (1991) examined the anatomy of the population momentum in detail. It was shown that the momentum of *any* observed population is given as the ratio of the *proportion* of the observed population under the mean age at childbearing to the *proportion* of its life table population under that age. Thus, the younger the initial population, the larger is the momentum.

The age-specific momentum for broad segments of a population was also presented. Approximately, (a) the momentum for the segment of a population below age 30 is 1, that is the under-30 population size remains constant, confirming Preston's (1986) result; (b) the momentum for age group 30 to 60 is R_0; and (c) the momentum for the group above age 60 is R_0^2. The exact age below which the population size remains constant can be determined exactly (Kim et al. 1991, Kim and Schoen 1993), but age 30 works well for all practical purposes. To illustrate the age-specific momentum graphically, the trajectory of population segments in five-year age groups after the immediate fertility drop is shown in Fig. 1a (initial $R_0 = 2.9$ in this example). The number of births (here represented by the population size of ages under five) initially drops when fertility falls, then recovers somewhat before it stabilizes at a new level, which is lower than the original level. In this process, the population segment under age 30 (below the lower solid line) remains virtually constant except for a small fluctuation around time 30, which is when the smaller birth cohorts born right after the fertility drop reach that age. The total population growth of some 60 percent after the fertility drop is due entirely to the increase in the population segment above age 30.

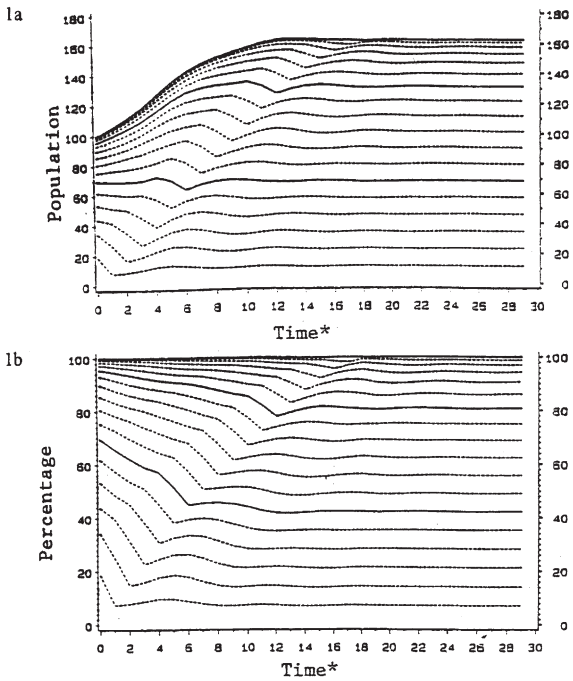

Figure 1
Age distributions and the momentum of the observed population of Singapore (1957): $R_0 = 2.9$. (a) Evolution of population segments from 0 to 90, in five-year age groups. Solid lines represent ages 30, 60, and 90. (b) Population segments from 0 to 90, in five-year age groups as percentages of total population. Solid lines represent ages 30, 60, and 90. *One time unit is five (5) years

Because the total population growth was contributed solely by the population above age 30 while the population under age 30 remains fixed in size, the population momentum and population aging occur simultaneously. The aging process of the population is shown clearly in the percent age distribution of Fig. 1b. The proportion under age 30 makes up 70 percent of the initial population but it decreases to a little over 40 percent in the eventual population, while the population above age 60 increases from a mere 4 percent to almost 20 percent, showing a substantial degree of aging of the population. The results described here hold for any initial population that is not stable, as well as for initially stable populations.

3. *Population Momentum and Population Aging*

Population momentum and population aging during the transition to stationarity are two aspects of the same phenomenon. The population momentum expresses population aging, and vice versa. The two quantities are separate aspects of the long-term change in size and age structure that results from a reduction in fertility to replacement level.

To make the relationship between the dual consequences of fertility decline—population momentum and population aging—more specific, three measures of aging were considered (Kim and Schoen 1997). They are: (a) the increase in the mean age of population (ΔA), (b) the relative size of the proportion under age 30 (F_{30}), and (c) the relative size of the proportion above age 65 (F_{65+}). The first measure, ΔA, is defined as the increase in mean age of the eventual stationary population from that of the initial population. The second measure, F_{30}, is defined as the ratio of the population under age 30 in the initial population to that in the ultimate stationary population, and the third measure, F_{65+}, is defined as the ratio of the proportion over age 65 in the initial population to that in the ultimate stationary population.

The three measures of aging are highly correlated with each other and with the population momentum, for these four quantities are linear functions of the

intrinsic growth rate, r, of the initial population. The theoretical linear relationships as well as some empirical results are presented in Kim and Schoen (1997). For empirical relationships, two data sets are used. The first is a set of Coale-Demeny West model stable populations (see *Demographic Models*) that includes 39 stable populations with the intrinsic growth rate r ranging from -0.01 to 0.05 in increments of 0.005 at three mortality levels (e_0 values of 60, 70, and 80). The second data set is comprised of 41 observed populations in 1985 (Keyfitz and Flieger 1990).

Figure 2 presents the empirical two-way scatterplots for the momentum, M, and each of the three aging measures generated by the first data set. The first plot (from left to right) shows the relationship between M and ΔA, the second M's relationship with F_{30}, and the third with F_{65+}. The momentum–mean age linear fit is quite good, and the estimated slope (as well as the theoretical value) suggests that the mean age of the population increases by a little more than two years for every 10 percent increase in population momentum. The middle plot shows that the relative size of the proportions under age 30 is not only linear, but is *identical* to the momentum. This is a direct consequence of the fact that the population size under age 30 remains constant before and after the transition, as discussed in Sect. 2. The third plot presents the relationship between the momentum and the relative size of the proportion above age 65. The fit is not as good as with the other two measures. The theoretical exposition reveals that, for this measure, the contribution from the quadratic term is not negligible, and that the coefficient of the linear term is affected by mortality levels, especially at the high end of the initial growth rate.

Thus, although the proportion over age 65 has great appeal as a conventional index of aging, it performs the poorest among the three measures of aging considered. The increase in the mean age of the population is simple and performs well, both theoretically and empirically. However, the ratio of the proportion under age 30 in the initial population to that in the ultimate population gives the population momentum *directly*.

As Fig. 2 illustrates, these linear relationships hold for *negative* momentum of population decline as well as for positive momentum of population growth. An initial stationary population with $r=0$ bridges the negative and positive momentum seamlessly.

4. Momentum Under Gradual Decline of Fertility

A population's growth potential is significantly underestimated when fertility is assumed to drop instantaneously to replacement level because no population can achieve such an abrupt fertility reduction. If fertility declines gradually over time, the resultant momentum is larger.

If an immediate fertility reduction is adopted by cohorts, the period NRR declines gradually (Frauenthal 1975). Suppose that cohorts born before time 0 maintain the old fertility level at a NRR of R_0, while all cohorts born after time 0 reproduce at replacement-level fertility. Then the time trajectory of the period NRR resembles a declining logistic curve that begins to decline at time 15 (the beginning age of childbearing), and decreases to 1 at time 45 (the maximum age at childbearing). The *extra* momentum generated by this delayed fertility transition is R_0. Thus the eventual population momentum under the assumption of a sudden drop in *cohort* fertility is given by $M_c = M \cdot R_0$.

Gradual fertility declines in period fertility have also been examined. Schoen and Kim (1998) consider a linear decline in the growth rate of births to 0 in a given period, during which the number of births grows to a higher level. In this case, NRR reaches replacement level a generation *after* the number of births stabilizes, and the time path of NRR decline is a smooth curve also resembling a declining logistic function. Li and Tuljapurkar (1999) consider a linear path of NRR decline, and derive the exact momentum when the duration of fertility transition is short (e.g., less than 15 years), then show numerically that it is a good approximation for as long as 40 years.

The two formulations give consistent results. The extra momentum increases exponentially as the duration of the fertility transition increases. More specifically, if the *period* NRR declines monotonically to replacement level over a duration, D, the *extra* momentum is larger by a factor $e^{Dr/2} \approx (\sqrt{R_0})^{D/\mu}$ (Schoen and Kim 1998). This is consistent with intuition: it can be interpreted as the growth at half the original growth rate during the given period, or as the growth at the original growth rate for half the duration. Therefore the eventual population

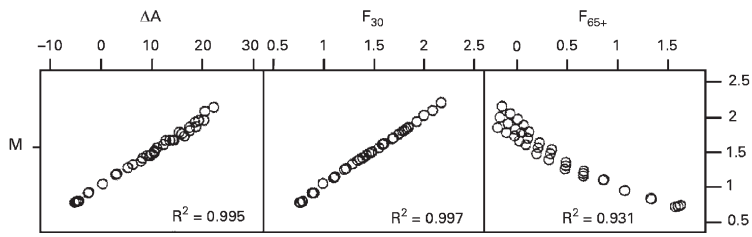

Figure 2
Scatterplots and coefficeints of determination (R^2) for M, ΔA, $F_{30,}$ and $F_{65+,}$ coale-demeny female stable population

Thus the momentum generated by a gradual period fertility transition over 30 years is approximately equal to R_0, equally contributed (multiplicatively) by the age distribution and the delayed fertility transition. The momentum under the sudden *cohort* fertility transition can also be explained in this vein. The extra population momentum, shown to be R_0, is contributed equally by the growth during the first 15 years at the original growth rate, and during the next 30 years with declining *period* fertility. The growth generated by a *sudden cohort* fertility transition that takes 45 years is equivalent to that of a *gradual period* fertility transition over 60 years.

The extra momentum generated by delayed transition is not sensitive to variations in the time path of fertility decline. A comparison of the momentum generated by a linear decline in the period NRR and that generated by the logistically declining paths (produced by the scenarios of Frauenthal 1975 and of Schoen and Kim 1998) shows that the difference is small (Li and Tuljapurkar 1999). The extra momentum is determined mainly by the number of years a population needs to complete a smooth fertility transition.

5. *Future Population Growth*

In practice, future population growth is determined by improvements in mortality and the pattern and extent of migration, as well as by the future trends in fertility decline and by the current age distribution. The population growth that results from urbanization of a highly rural population in which the rural sector has higher fertility than in the urban sector is called spatial momentum (Rogers and Willekens 1978, Schoen and Kim 1993). Modern projection/prediction studies show that as the world's populations complete demographic transitions, the momentum inherent in the current young age distribution is the major component of future population growth in most developing countries (Bongaarts and Bulatao 1999). The other components of population growth—the spatial momentum, and the growth due to changes in fertility and mortality—under realistic assumptions is shown to be less than the age momentum (Bongaarts and Bulatao 1999, Brockerhoff 1999). Thus the growth associated with the young current age distribution in developing countries will be the major contributor to the future growth of the world population.

See also: Demographic Transition, Second; Fertility Change: Quantum and Tempo; Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems; Population Dynamics: Theory of Nonstable Populations; Population Dynamics: Theory of Stable Populations

Bibliography

Bongaarts J, Bulatao R A 1999 Completing the demographic transition. *Population and Development Review* **25**: 515–29

Brockerhoff M 1999 Urban growth in developing countries: A review of projections and predictions. *Population and Development Review* **25**: 757–78

Frauenthal J 1975 Birth trajectory under changing fertility conditions. *Demography* **12**: 447–54

Frejka T 1973 *The Future of Population Growth: Alternative Paths to Equilibrium*. Population Council, New York

Keyfitz N 1971 On the momentum of population growth. *Demography* **8**: 71–80

Keyfitz N, Flieger W 1990 *World Population Growth and Aging*. University of Chicago Press, Chicago

Kim Y J, Schoen R 1993 Crossovers that link populations with the same vital rates. *Mathematical Population Studies* **4**: 1–19

Kim Y J, Schoen R 1997 Population momentum expresses population aging. *Demography* **34**: 421–7

Kim Y J, Schoen R, Sarma P 1991 Momentum and the growth-free segment of a population. *Demography* **28**: 159

Li N, Tuljapurkar S 1999 Population momentum for gradual demographic transitions. *Population Studies* **53**: 255–62

Preston S 1986 The relation between actual and intrinsic growth rates. *Population Studies* **40**: 343–51

Rogers A, Willekens F 1978 The spatial reproductive value and the spatial momentum of zero population growth. *Environment and Planning, Series A* **10**: 503–18

Schoen R, Kim Y J 1993 Two-state spatial dynamics in the absence of age. *Theoretical Population Biology* **44**: 67–79

Schoen R, Kim Y J 1998 Momentum under a gradual approach to zero growth. *Population Studies* **52**: 295–9

Y. J. Kim

Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems

1. *Extinction, Stable Size, or Explosion?*

Historically, probabilistic population theories originate from the *extinction problem*: suppose you know the probabilities of begetting 0, 1, 2, 3,… children. Now determine the probability that a population of a known number of individuals will eventually die out. Already Malthus had noted that a rapid increase of the whole population could occur together with frequent extinction of the separate families constituting the population. Rephrased in the terminology of evolutionary biology, frequent extinction of species does not contradict a rapid growth of the biomass of the world.

The mathematical explanation of this apparent paradox is due to Bienaymé (1845) and independently,

thirty years later and only in parts correctly, to Galton and Watson, who were later to lend their names to the simplest, generation-counting version of the so called branching processes that provide the mathematical structure underlying most stochastic population dynamics. Indeed, if p_k denotes the probability of giving birth to k children, $k = 0, 1, 2, 3, \ldots$ then the extinction probability of a population stemming from one ancestor will be the smallest nonnegative solution q of the equation

$$\sum_k p_k q^k = q.$$

If we disregard the trivial case $p_1 = 1$, q will equal 1 if and only if the mean number of children

$$m = \sum_k p_k k \leqslant 1.$$

If $m > 1$, then the population growth rate per generation will be m, and even though q must be < 1 it can still be pretty large: it is easy to formulate demographically or biologically plausible cases where, say, m exceeds 2 and q is still around 0.8.

If a stochastic population does not die out, it can be proved very generally that it must grow beyond all limits (see Jagers 1991). This is a consequence of stochasticity and the asymmetry between extinction and big sizes: however big the population, there remains an albeit small but still positive risk of extinction, whereas an extinct population cannot be resurrected. This is in contradiction to deterministic population theories which allow population sizes to be stable, cf. Keyfitz 1968.

A finite world can certainly not sustain infinite growth. One can therefore argue that probabilistic population dynamics show that all real populations will die out sooner or later. But before that happens there may be a long period of growth.

2. Population Growth

To describe population growth in real time rather than generation-wise, a more sophisticated description of individual life and reproduction is needed. Probabilistic population dynamics is *individual based* in the sense that it starts from such a description. Deterministic theories, on the other hand, often start directly from an assumed behavior of the population as a whole. Demographic theory is often *pseudoprobabilistic* in that it uses probabilistic notions but makes conclusions only about the expected behavior of the population as a whole. (This is the reason why there exist deterministic populations with stationary sizes: if the underlying stochastic population is critical in the sense that on average each individual is replaced by one child, then the expected total population size will be stationary, whereas the population itself will die out sooner or later.)

The most general probabilistic formulation of individual life goes as follows: at birth an individual inherits a *type* from her mother. The type determines her survival distribution as well as the distribution of age at childbearing. It also determines the types of the children she will give birth to in her turn. (Think of the type as a genotype, or a genotype combined with the social and biological conditions of life.)

In the *supercritical* case, where there is a positive chance of nonextinction, the probability laws of survival and childbearing determine a positive *Malthusian* parameter, which gives the rate of exponential growth. In the theory of branching processes this rate is usually denoted by α, so that the population will grow like $e^{\alpha t}$, as time t passes. In deterministic theory the conventional notation is often r instead.

Even though exponential growth appears to be a consequence of the independence assumed (tacitly or explicitly) between the reproduction of different individuals, it is interesting to note that such growth has recently been proved also to occur in many cases where individual reproduction depends upon the population as a whole, e.g., its size. For such results in the case of simple Galton-Watson branching processes see Klebaner (1984).

In cases where the reproduction approaches criticality as the population grows, non-Malthusian growth can occur nevertheless, e.g., in the form of linear growth (Klebaner 1994).

3. Stable Population Theory

In contrast to population size, the composition of growing populations tends to stabilize. The description of this composition, or rather the various stable compositions pertaining to different reproduction and survival laws, is the object of stable population theory.

The most well known aspect of stable population composition is the stable age distribution, first described by Euler in 1760 and since then rediscovered repeatedly: if the Malthusian parameter is $\alpha > 0$ and all individuals have the same survival function ℓ, then the age of an individual sampled at random in an old population will have the density

$$\frac{e^{-\alpha a}\ell(a)}{\int_0^\infty e^{-au}\ell(a)\,du}$$

at age a, exactly as in deterministic theory. This is, of course, a law-of-large-numbers effect; we are considering ever larger populations.

If, as in the general setup mentioned above, there are several types of individuals, each with its own life span distribution, then the situation will be more

complex. First of all, the type distribution will converge to a stable type law, determined by an eigenmeasure property. Then the survival function ℓ above will have to be replaced by an average of all the different survival functions weighted by the stable type distribution.

But the stable composition describes many other aspects of a balanced growing population besides ages. For example the probability of being first-born can be calculated as well as other aspects of family structure. In modern evolutionary biology the genealogy of individuals and, say, the mutational history plays an important role. In an old population this is described by a stable, typical ancestry, which can be used to determine time back to some event like divergence between species.

See also: Demographic Analysis: Probabilistic Approach; Population Dynamics: Classical Applications of Stable Population Theory; Population Dynamics: Momentum of Population Growth; Population Dynamics: Theory of Stable Populations

Bibliography

Athreya K B, Ney P 1972 *Branching Processes*. Springer-Verlag, Berlin

Harris T 1963 *The Theory of Branching Processes*. Springer-Verlag, Berlin

Jagers P 1975 *Branching Processes with Biological Applications*. Wiley, Chichester

Jagers P 1991 The growth and stabilization of populations. *Statistical Science* **6**: 269–83

Keyfitz N 1968 *Introduction to the Mathematics of Population*. Addison-Wesley, Reading, MA

Klebaner F C 1984 Geometric rate of growth in population size dependent branching processes. *Journal of Applied Probability* **21**: 40–9

Klebaner F C 1994 Asymptotic behaviour of Markov population processes with asymptotically linear rate of change. *Journal of Applied Probability* **31**: 614–25

P. Jagers

Population Dynamics: Theory of Nonstable Populations

1. Background

To introduce the theory of nonstable population dynamics, we start by presenting a Malthusian model studied by Lee (1974). Consider an overlapping-generation framework in which each individual lives one or two periods. The first period is childhood and the second is adulthood, and all surviving adults will be in the labor force. The Malthusian model can be characterized by the following two equations: $W_t = f(N_t)$, $b_t = g(W_t)$, where W_t is the wage rate (at time t), N is the size of the adult group, b is the crude birth rate, and $f(.)$ and $g(.)$ are two functions. The $f(.)$ function characterizes the wage/employment relationship in the labor market, and the $g(.)$ function specifies the influence of the labor market reward on fertility.

Suppose the number of children born in period t is B_t, and the child survival rate is l, then $N_t = l.B_{t-1}$. The above two equations can be combined in the following reduced form: $B_t = h(B_{t-1})$, where $h(B_t) \equiv g(f(l.B_{t-1})).l.B_{t-1}$. The above expression is a recursive equation of B_t, from which the dynamic pattern can be analyzed, and the possible existence of cyclicity can be studied. Here the Malthusian model is interpreted as a model of density dependency, where the density is characterized by the birth size of the previous period. It turns out that the Malthusian model described above may not converge to a 'dismal' steady state, as Malthus suggested. Mathematically, the B_t series may converge, diverge, or be cyclic, and the volatility of the birth series B_t crucially hinges on the feedback elasticity of $h(.)$ with respect to B_{t-1}. When this elasticity is sufficiently large, the B_t series can easily have cycles. This is the classical case of nonstable population. Modern versions of nonstable population will be discussed below, after we introduce the general population structure.

2. The General Population Structure

Let $\mathbf{N}_t \equiv (N_{t,1}, \ldots, N_{t,n})$ be an n-type population vector at period t, where $N_{t,i}$, $i = 1,\ldots,n$ indicates the size of type-i population at period t. The type here may refer to any criterion that is used to classify the population, such as age, sex, and occupation. A general formulation of population dynamics can be written as follows:

$$N_{t+1} = Q(N_t)N_t, \tag{1}$$

where $Q(N_t)$ is the transition matrix. The simplest case is when $Q(N_t) = Q$, a time-independent matrix. In this situation, the above equation reduces to a time-invariant Markov chain: $N_{t+1} = QN_t$. It is a well-known implication of the Frobenius–Perron theorem that if Q is positively regular, meaning that there is flexible cross-period mobility across types, then the population vector N_t will converge to a steady state, in which all elements of the N_t vector grow at the same exponential rate, thereby the population composition is also time-invariant. It has been shown by Chu (1998) that, for most population models with time-

independent Q, this positive regularity condition is likely to be satisfied, therefore any nonstable behavior of the population would appear only when $Q(N_t)$ is not a constant matrix, or equivalently when the relationship between N_{t+1} and N_t in Eqn. (1) has some nonlinearity. Below we shall introduce several special cases of the generic setting in Eqn. (1), and discuss their empirical relevance. The reader can see that each model corresponds to a special interpretation of the generic setting in Eqn. (1).

3. Easterlin Cycles

For demographers, the most familiar version of the dynamic models in Eqn. (1) is the Lotka–Leslie age-specific models, where the type criterion is age. Let $N_{t,i}$ be the number of population aged i at period t, B_{t-i} be the birth size at period $t-i$, and l_i be the probability that a person can survive to age i, then the following relationship must hold: $N_{t,i} = B_{t-i}\, l_i$. The transformed dynamics in terms of B_t is the well-known Lotka's equation:

$$B(t) = \sum_i B_{t-i} . l_i . m_{t,i} = \sum_i B_{t-i} . \phi_{t,i} \tag{2}$$

where $m_{t,i}$ is the average number of births per surviving member aged i at period t, and $\phi_{t,i} \equiv l_i \,.\, m_{t,i}$ is the net maternity function. Evidently, the Malthusian model presented in the previous section is a special case of Eqn. (2a) with two age types. As we mentioned, when the fertility function $m_{t,i}$ is independent of t, then the process of B_t follows a time-invariant Markov chain, and the usual convergence result applies. But if $m(t,i)$ is time dependent, then cycles or more volatile behavior may appear. One typical case of such time dependency is when the age-specific fertility rates are functions of previous birth sizes: $m_{t,i} = m_i(B_t)$, where $B_t \equiv (B_{t-1}, B_{t-2}, \ldots)$.

Suppose the birth series corresponding to Eqn. (2a) has a steady state. Taking a Taylor expansion around this steady state, Lee (1974) showed that the volatility of the birth series characterized by Eqn. (2a) hinges upon the elasticities of the net maternity function with respect to B_t. When previous birth sizes are larger, usually the density pressure is larger. Intuitively, larger feedback elasticities imply that present reproduction would be more sensitive to previous birth shocks, hence the birth series would be more volatile. Easterlin (1961) argued that the baby-boom/baby-bust cycles observed in the United States are typical examples of a feedback cycle.

How large the feedback elasticities are is an empirical question. In order to estimate such elasticities, two models have been proposed: the first is to assume that $\phi_{t,i}$ is only a function of the cohort size aged i: $\phi_{t,i} = \phi_i(B_{t-i})$, and the second is to assume that $\phi_{t,i}$ is a function of the weighted average of birth sizes of different ages: $\phi_{t,i} = \phi_i(\sum_j w_j \,.\, B_{t-j})$. The former is referred to as the cohort model, and the latter as the period model. However, empirical evidence has so far failed to back up either of these two models in producing persistent birth cycles that fit all characteristics (such as amplitude and period). Moreover, even if there does exist a cyclic solution to Eqn. (2a), the information concerning feedback elasticity is not sufficient for us to tell whether the cyclical solution in question is stable. Related analysis is rather technical, and the reader is referred to Tuljapurkar (1987) for details.

4. Predator–Prey Cycles

The predator–prey model describes a two-type population structure, where the types are prey and predator, respectively, referred to as type-1 and type-2. The simplest formulation of the predator–prey model is characterized by the following difference equations:

$$\begin{aligned} \Delta N_{t+1,1} &\equiv N_{t+1,1} - N_{t,1} \\ &= aN_{t,1} - bN_{t,1}N_{t,2} \end{aligned} \tag{3a}$$

$$\begin{aligned} \Delta N_{t+1,2} &\equiv N_{t+1,2} - N_{t,2} \\ &= -cN_{t,2} + dN_{t,1}N_{t,2}, \\ &\quad a, b, c, d > 0. \end{aligned} \tag{3b}$$

Equation (3) is obviously a special case of Eqn. (1). The prey are considered the only food resource available to the predator. Thus, if $N_{t,1} = 0$, the predator population decreases exponentially at the rate c. With the nonexistence of predators, the prey population grows exponentially at the rate a.

In the steady state of Eqn. (3), $\Delta N_{t,1} = \Delta N_{t,2} = 0$, there is only one nontrivial solution: $(N_1^*, N_2^*) = (c/d, a/b)$. Let the Jacobian matrix of Eqn. (3) around the nontrivial solution be denoted J. It turns out that the determinant of J is

$$J = \begin{pmatrix} a - bN_2^* & -bN_1^* \\ dN_2^* & -c + dN_1^* \end{pmatrix} = \begin{pmatrix} 0 & -bc/d \\ ad/b & 0 \end{pmatrix}$$

$|J| = ac > 0$ and the trace of J is zero.

Thus, we know that the eigenvalues of J are purely imaginary, meaning that besides the equilibrium point $(c/d, a/b)$ itself, the solution trajectories of Eqn. (3) are cyclic and nonstable. But it was also pointed out that when the predator–prey model embodies the concerns of diminishing or increasing returns, then the solution becomes locally stable or divergent.

5. Dynastic Cycles

Based on stylized facts found in Chinese history, Chu and Lee (1994) proposed an occupation-specific population model, in which the population is separated into

three occupations: bandits, peasants, and soldiers. Peasants grow crops and pay taxes, soldiers (rulers) collect taxes and hunt for bandits, and bandits rob food and clash with peasants and soldiers. It is assumed that soldiers are drafted by the government, and all other civilians can choose to become peasants or bandits. Their occupation choice is made depending on which occupation is expected to generate a higher utility. When the bandit/soldier ratio is relatively high, it is likely that the incumbent power regime will be overthrown by the bandit group, and the dynasty switches. It turns out that this pattern of population dynamics characterizes the dynamic evolution of dynasties in Chinese history, and the cycles were called dynastic cycles by historians.

In terms of demography, the occupation switches mentioned above are merely changes in population composition. But there is a unique feature associated with the above-mentioned occupation switching, which we shall explain below. For most animals, it is believed that when density pressure occurs, the environment becomes less favorable, and the reproduction rate of animals is reduced. One might expect that rational human decisions may be able to weaken the outside density pressure through institutional and rational regulation. But the scenario of dynastic cycles seems to be a counterexample. As noted by historians, clashes between soldiers and rebellious peasants in human history were often initiated by famine or density pressure. The higher the bandit ratio, the less likely a peasant can safely keep a good proportion of his crop, and hence the greater attraction for the peasant to join the bandit group and seize other people's crops. Thus, human beings' occupational choice between peasants and bandits has a 'demonstration effect,' which tends to magnify the originally weak density pressure and 'destablize' the dynamics of the population compositional structure.

The cyclical pattern of population composition mentioned above is not the only case we find in human history. Some researchers have considered a neoclassical economic growth model, and focused on the two economic classes in the economy: the laborers and the capitalists. Let u_t be the income share attributed to labor, and v_t be the employment rate of labor. Under some reasonable assumptions, Goodwin (1967) has shown that the dynamics of u_t and v_t mimic that of Eqn. (3). Thus, the conclusion that the capitalists' economy appears to be 'permanently oscillating' was reached. But just as the original predator–prey model is sensitive to variations in institutional specifications, Goodwin's model after some minor modification also generates qualitatively different results. Definite evidence of the nonstability of human population has yet to be found.

See also: Demographic Models; Population Cycles and Demographic Behavior; Population Dynamics: Classical Applications of Stable Population Theory; Population Dynamics: Theory of Stable Populations

Bibliography

Cyrus Chu C Y 1998 *Population Dynamics: A New Economic Approach*. Oxford University Press, New York

Cyrus Chu C Y, Lee R D 1994 Famine, revolt and the dynastic cycle: Population dynamics in historic China. *Journal of Population Economics* **7**: 351–78

Easterlin R A 1961 The American baby boom in historical perspective. *American Economic Review* **51**: 860–911

Goodwin R M 1967 A growth cycle. In: Feinstein C H (ed.) *Socialism, Capitalism and Economic Growth*. Cambridge University Press, Cambridge, UK

Lee R 1974 The formal dynamics of controlled population and the echo, the boom and the bust. *Demography* **11**: 563–85

Tuljapurkar S 1987 Cycles in nonlinear age-structured models i: Renewal dynamics. *Theoretical Population Biology* **32**: 26–41

C. Y. Cyrus Chu

Population Dynamics: Theory of Stable Populations

A population with an age distribution which remains constant over time is referred to as 'stable.' A stable population of constant size is said to be 'stationary.' These concepts are of considerable antiquity. Euler, for example, knew in 1760 the stable age distribution a population needed to have if it were to have a given growth rate. Compilers of some of the early life tables such as Graunt in 1662 appear to have assumed populations which were stationary, and nineteenth-century actuaries certainly made use of the stationary population concept (King 1902).

Subject to age-specific mortality and fertility rates which remain unchanged over time a population will eventually develop a stable age distribution which depends on those mortality and fertility rates but is independent of the initial age distribution of the population. This result, due to Sharpe and Lotka (1911) caused renewed interest in stable populations among demographers. Although demographers usually think of a stable population as one with an age distribution which is unchanging, such a definition is restrictive, because populations can also exhibit stability in respect of other characteristics as well as age and in more general situations.

1. The Population Model of Sharpe and Lotka

Sharpe and Lotka considered the male population alone and assumed that the growth of the female population would be such as to justify assumptions of

constant age-specific fertility and mortality for the males. Subsequent writers have usually applied the model to the female component of the population because of its shorter reproductive age span and the fact that the number of females in the reproductive age range plays a more significant role in determining the number of births than the number of males (Pollard 1973).

Applied to the female population, the model assumes $F(x, t)\,\mathrm{d}x$ females in the age range $(x, x+\mathrm{d}x)$ at time t and $B(t)\,\mathrm{d}t$ female births between times t and $t+\mathrm{d}t$. Then, if the average number of daughters born per female aged exactly x in time element $\mathrm{d}t$ is $\lambda(x)\,\mathrm{d}t$ and the probability that a female survives from age u to age $u+v$ is ${}_vp_u$, the following relationship is immediately apparent:

$$B(t) = \int_0^\infty F(x, t)\,\lambda(x)\,\mathrm{d}x \tag{1}$$

Noting that those alive and aged x at time t, enumerated in $F(x, t)\,\mathrm{d}x$, must either have been born at time $t-x$ $(x < t)$ or else have been aged $x-t$ at time 0 $(x > t)$, Eqn. (1) may be rewritten as

$$B(t) = \int_0^t B(t-x)\,{}_xp_0\,\lambda(x)\,\mathrm{d}x + G(t) \tag{2}$$

with

$$G(t) = \int_0^\infty F(x, 0)\,{}_tp_x\,\lambda(x+t)\,\mathrm{d}x \tag{3}$$

The solution to the integral equation (2) reveals that for large values of t, under quite general assumptions

$$B(t) \cong C\,e^{rt} \tag{4}$$

where the constant r (the *intrinsic growth rate* of the population) is the unique real solution to the integral equation

$$\int_0^\infty e^{-rx}\,{}_xp_o\,\lambda(x)\,\mathrm{d}x = 1 \tag{5}$$

and the constant C depends on the initial age distribution of the population as well as the age-specific mortality and fertility rates (Keyfitz 1968, Pollard 1973).

Since the females aged between x and $\mathbf{x}+\mathrm{d}x$ at time t are the survivors of those born x years earlier between times $t-x$ and $t-x+\mathrm{d}x$, it follows from Eqn. (4) that

$$F(x, t) \cong C\,e^{r(t-x)}\,{}_xp_0 = (C\,e^{rt})\,(e^{-rx}\,{}_xp_0) \tag{6}$$

The first factor in the right-hand side of Eqn. (6) indicates that the population grows exponentially asymptotically at rate r, while the second reveals that the age distribution is asymptotically independent of t. In other words, a *stable age distribution* is reached.

2. *The Discrete Time Model of Bernardelli, Leslie, and Lewis*

Whilst the continuous time model of Sharpe and Lotka is elegant and predicts a stable population asymptotically under conditions of constant age specific mortality and fertility, the discrete time formulation of Bernardelli (1941), Lewis (1942) and Leslie (1945), which makes similar assumptions, is more readily adapted to analyze stability in more complicated situations.

The number of females aged x last birthday at time t is represented by $n_{x,t}$. There is no migration, and no one can live beyond age k last birthday. The average number of daughters born to a female while she is aged x last birthday, those daughters surviving to be enumerated as being aged zero at the next point of time, is F_x, and the proportion of females aged x last birthday surviving to be aged $x+1$ last birthday a year later is P_x. Under these assumptions, the following matrix recurrence equation must apply to the numbers of females in successive years:

$$\begin{pmatrix} n_{0,t+1} \\ n_{1,t+1} \\ n_{2,t+1} \\ n_{3,t+1} \\ \cdots \\ n_{k,t+1} \end{pmatrix} = \begin{pmatrix} F_0 & F_1 & F_2 \cdots & & F_{k-1} & F_k \\ P_0 & & & & & \\ & P_1 & & & & \\ & & P_2 & & & \\ & & & \ddots & & \\ & & & & P_{k-1} & \end{pmatrix} \begin{pmatrix} n_{0,t} \\ n_{1,t} \\ n_{2,t} \\ n_{3,t} \\ \cdots \\ n_{k,t} \end{pmatrix}. \tag{7}$$

This equation may be written more concisely as

$$\mathbf{n}_{t+1} = \mathbf{A}\,\mathbf{n}_t \tag{8}$$

so that

$$\mathbf{n}_t = \mathbf{A}^t\,\mathbf{n}_0 \tag{9}$$

Under fairly general conditions, matrix $\mathbf{A}$ is positive regular, in which case it has a dominant eigenvalue α which is positive, of multiplicity one, and greater in absolute size than any other eigenvalue. It follows from Eqn. (9) therefore that for large t,

$$\mathbf{n}_t \cong C\,\alpha^t\,\mathbf{z} \tag{10}$$

where C is a constant which depends on the initial age distribution of the population, and $\mathbf{z}$ is the column eigenvector corresponding to the eigenvalue α. If Π_s is written instead of the product $P_0 P_1 \ldots P_S$, the elements of $\mathbf{z}$ are $\Pi_0 \alpha^{-0}$, $\Pi_1 \alpha^{-1}, \ldots, \Pi_k \alpha^{-k}$. It follows, therefore that the number of females aged x last birthday at time t is

$$n_{x,t} \cong (C\ \alpha^t)\ (\alpha^{-x}\ \Pi_x) \tag{11}$$

This is the discrete analogue of the Sharpe and Lotka formula, Eqn. (6). The second factor defines the asymptotic stable age distribution; the first shows the exponential growth of the population (Keyfitz 1968, Pollard 1973).

3. *Stability with Migration*

The most appropriate model for a population subject to significant migration must depend on the pattern that migration takes. In the case of a country accepting a constant number of immigrants each year with a constant age pattern, for example, the discrete time recurrence of Eqn. (8) needs to be modified as follows:

$$\mathbf{n}_{t+1} = \mathbf{A}\ \mathbf{n}_t + \mathbf{b} \tag{12}$$

where $\mathbf{b}$ is a constant vector of immigrants. Under this model, immigrants are assumed to adopt immediately the mortality and fertility patterns of their new homeland (see *Immigrant Fertility*). Repeated application of Eqn. (12) for $t = 0, 1, 2, \ldots$ reveals that, provided the dominant eigenvalue of $\mathbf{A}$ is not equal to one,

$$\mathbf{n}_t = \mathbf{A}^t\ \mathbf{n}_0 + (\mathbf{I} - \mathbf{A}^t)\ (\mathbf{I} - \mathbf{A})^{-1}\ \mathbf{b} \tag{13}$$

If the rate of natural increase of the population is less than zero, α will be less than one, and for very large t

$$\mathbf{n}_t \cong (\mathbf{I} - \mathbf{A})^{-1}\ \mathbf{b} \tag{14}$$

The population will become stationary with its age structure determined by the ages of the immigrants (vector $\mathbf{b}$), as well as the mortality and fertility of the host country (matrix $\mathbf{A}$). Where the fertility of the host country is greater than replacement level, α will be greater than one, and the following asymptotic formula follows from Eqn. (13) for very large t:

$$\mathbf{n}_t \cong C'\ \alpha^t\ \mathbf{z} \tag{15}$$

In this case, the population grows asymptotically with the same exponential growth rate as it did in the discrete nonimmigration case of Sect. 2. The asymptotic stable age distribution (vector $\mathbf{z}$) is also the same. The constant C', however, is different, depending on the initial age structure, the vector of immigrants $\mathbf{b}$, and the mortality and fertility of the population.

In the pathological case where the levels of mortality and fertility in the population correspond exactly to replacement, α is equal to one, and a different algebraic approach is necessary to the solution of Eqn. (12), which yields, for very large t, the following asymptotic result:

$$\mathbf{n}_t \cong C''\ t\ \mathbf{z} \tag{16}$$

The population grows linearly asymptotically and the asymptotic age distribution is the same as in the previous case. The constant C'' depends on the immigration vector, as well as the fertility and mortality rates of the population.

For a population experiencing net emigration with a proportion β_x of females aged x departing each year, the matrix model of Sect. 2 can be utilized with each of the P_x in matrix $\mathbf{A}$ replaced by $P_x - \beta_x$. The mathematical theory can then be applied in the same manner as before, and an asymptotic stable population emerges. The dominant eigenvalue and eigenvector will of course be different from those in Sect. 2. Depending on whether the dominant eigenvalue of the modified matrix $\mathbf{A}$ is less than, equal to, or greater than one, the population will ultimately decline, become stationary or grow in size.

The above models are only two of a very wide range of possible migration models appropriate in different circumstances. They demonstrate how the discrete time model of Sect. 2 may be readily modified to take account of migration and the fact that population stability will usually emerge in the presence of migration when the migration parameters, mortality, and fertility remain unchanged over time.

4. *Generalizations*

The discrete time models of the earlier sections are readily adapted to study populations partitioned according to other characteristics as well as age. Multiregional demography is the best known example (Rogers 1966, 1975). Females of a population are characterized by age and geographical region in an extended column vector $\mathbf{n}_t$ of dimension mk, where k is the maximum age last birthday attainable and m is the number of regions in the country. As well as surviving and reproducing in the manner of Sect. 2 with rates that may differ from region to region, females can move from one region to another with transition rates which depend on the regions involved and may also depend on age. The matrix recurrence equation is again Eqn. (8), but with $\mathbf{n}_t$ of dimension mk and $\mathbf{A}$ of dimension $mk \times mk$. Conditions for a unique dominant eigenvector are readily determined, and asymp-

totic stability for the population by age and geographic region is predicted.

Another partition characteristic sometimes proposed is the *parity* of woman. Females at time t are enumerated in a vector $\mathbf{n}_t$ with elements $\{n_{x,y,t}\}$ which record the numbers aged x last birthday who already have y children. The only possible transitions for a female aged x and parity y remaining within the system are to age $x+1$ and parity y or to age $x+1$ and parity $y+1$ (ignoring the small possibility of multiple births). Where there is a change in parity, a new female will be introduced into the population aged 0 last birthday and parity 0 if the mother's change in parity corresponds to a female birth. Recurrence matrix $\mathbf{A}$ contains these transition rates. The fundamental recurrence equation is again Eqn. (8). For all realistic populations, the enlarged matrix $\mathbf{A}$ will have a unique dominant eigenvector, and the population enumerated by age and parity will approach stability (Keyfitz 1968).

The discrete-time models of Sects. 2 and 3 have also been applied to workforce planning (Bartholomew 1973, Pollard 1967), membership of learned societies (where Fellows are elected on merit and there is a danger of the society becoming a rapidly aging institution), and membership of other groups including pension schemes (Sherris and Pollard 1980). A learned society, for example, might have a policy of electing five new Fellows each year. The membership of the society at time t can be summarized by a vector $\mathbf{n}_t$ with elements $\{n_{x,t}\}$ representing the numbers at the various ages at that time. The number of new members (five) and their age distribution can be summarized in a vector $\mathbf{b}$. The transition rates in the recurrence matrix are simply the proportions $\{P_x\}$ surviving from age x last birthday to age $x+1$ last birthday a year later, for all relevant ages. Assuming that the ages are listed in their natural increasing order, the transition matrix $\mathbf{A}$ is particularly simple in this case: all its elements are zero, except those immediately below the main diagonal which comprise the survival proportions $\{P_x\}$. With $\mathbf{n}_t$, $\mathbf{b}$, and $\mathbf{A}$ defined in this manner, the recurrence equation for the society is Eqn. (12). All the eigenvalues of $\mathbf{A}$ are zero, and for t greater than the number of ages involved, $\mathbf{A}^t = \mathbf{O}$. The society will approach a stationary state given by Eqn. (14).

All such linear models with realistic assumptions concerning transition between states, entry and reproduction can be shown to predict stable populations asymptotically.

5. Applications

No observed population experiences constant levels of mortality and fertility. Nevertheless the concept of a stable population finds frequent application. Keyfitz's elegant formula for the effect of population momentum (see *Population Dynamics: Momentum of Population Growth*) assumes that the population is stable immediately prior to fertility falling to replacement level. Yet the formula works remarkably well for observed populations which are not stable. Stable population theory may be helpful in deducing results in respect of populations with unreliable and/or limited demographic data. Equation (4), for example, can be used to infer an approximate value for the crude birth rate of a population where contraception is not widely practiced, using only a reliable census of the population and a more or less appropriate life table (Bourgeois-Pichat 1957). Many of the applications of model life tables rely on stable population theory (see also *Population Dynamics: Classical Applications of Stable Population Theory*).

6. Limitations

The discrete time model of Bernardelli, Leslie, and Lewis, like the continuous (female) formulation of Sharpe and Lotka, assumes that the development of the male component of the population is such as to justify the assumption of unchanging female age-specific mortality and fertility. Linear models which include the male component of the population alongside the females have also been proposed, and mathematical analysis analogous to that presented in Sect. 2 predicts a stable age–gender composition for the whole population. Nonlinear models have also been proposed to accommodate the 'two-sex problem' (see *Population Dynamics: Two-sex Demographic Models*). Many of these nonlinear models would also seem to predict populations which are asymptotically stable under conditions of unchanging rates of mortality, fertility and nuptiality.

For many developing countries, for which stable population theory had been an extremely helpful tool, mortality began to decline in the second half of the twentieth century, but fertility remained essentially unchanged from previous levels. To accommodate the improvements in mortality, the concept of a quasi-stable population was introduced: one with improving mortality but constant fertility (Coale 1963).

Bibliography

Bartholomew D J 1973 *Stochastic Models for Social Processes*, 2nd edn. Wiley, New York

Bernardelli H 1941 Population waves. *Journal of Burma Research Society* **31**: 1–18

Bourgeois-Pichat J 1957 Utilisation de la notion de population stable pour mesurer la mortalité et la fécondité des populations des pays sous-développés *Bulletin de l'Institut International de Statistique* (actes de la 30e Session)

Coale A J 1963 Estimation of various demographic measures through the quasi-stable age distribution. *Emerging Tech-*

niques in Population Research: Proceedings of the 1962 Annual Conference of the Milbank Memorial Fund
Keyfitz N 1968 *Introduction to the Mathematics of Population*. Addison-Wesley, Reading, MA
King G 1902 *Life Contingencies*, 2nd edn. Charles & Edward Layton, London
Leslie P H 1945 On the use of matrices in certain population mathematics. *Biometrika* **33**: 183–212
Lewis E G 1942 On the generation and growth of a population. *Sankhya* **6**: 93–6
Pollard J H 1967 Hierarchical population models with Poisson recruitment. *Journal of Applied Probability* **4**: 209–13
Pollard J H 1973 *Mathematical Models for the Growth of Human Populations*. Cambridge University Press, Cambridge, UK
Rogers A 1966 The multiregional matrix growth operator and the stable interregional age structure. *Demography* **3**: 537–44
Rogers A 1975 *Introduction to Multiregional Mathematical Demography*. Wiley, New York
Sharpe F R, Lotka A J 1911 A problem in age-distribution. *Philosophical Magazine*, *Ser. 6* **21**: 435–8
Sherris M, Pollard J H 1980 Application of matrix methods to pension funds. *Scandinavian Actuarial Journal* 77–95

J. H. Pollard

Population Dynamics: Two-sex Demographic Models

1. Classical Stable Population Theory

A population is in 'demographic equilibrium' if the fraction of individuals in each age-sex category remains constant over time and the population is growing (or declining) at a constant rate. Demographers call a population that is in demographic equilibrium a 'stable population.' This use of the phrase stable population blurs the distinction between questions involving the existence of a demographic equilibrium and those involving its dynamic stability; hence, I depart from standard demographic terminology and in its place use the phrase 'demographic equilibrium.' In deference to traditional usage, however, I refer to demography's standard model as 'classical stable population theory.'

Nondemographers are often surprised to learn that the classical stable population theory operates with a one-sex model. The theory postulates an age-specific fertility schedule for women and an age-specific mortality schedule for women. These two schedules, in conjunction with the very strong assumptions that both of these schedules remain constant over time, imply an equilibrium growth rate and an equilibrium age distribution for the female population.

Demography's two-sex problem is the problem of integrating men into this one-sex, female-based model. The two-sex problem is often motivated by discussing the effect on fertility of a 'marriage squeeze.' A marriage squeeze is an imbalance between the female and male populations of marriageable age brought about, for example, by wars that disproportionately reduce the number of marriageable men. A marriage squeeze can also result from a baby boom: because women typically marry older men, a baby boom increases the number of marriageable women many years before it increases the number of marriageable men. Common sense suggests that a marriage squeeze is likely to reduce age-specific fertility rates for women, but the classical stable population theory, because it assumes that age-specific fertility rates remain constant, cannot accommodate this common sense conclusion.

2. The Birth Matrix-mating Rule Model

Demography's two-sex problem is solved, in a formal sense, by the 'birth matrix-mating rule' (BMMR) model introduced in Pollak (1986). (Feeney (1972), in an unpublished doctoral dissertation, presents a two-sex model with the same basic structure, but without a satisfactory proof of the existence of equilibrium.) The BMMR model has three primitives (i.e., components not derived from more fundamental components):

(a) a birth matrix, which indicates the number of offspring born to each possible type of union (e.g., the union of a female of age i and a male of age j);

(b) a mating rule, which indicates the number of unions of each type as a function of the number of females in each age category and the number of males in each age category; and

(c) age-specific mortality schedules for males and females.

Under relatively weak assumptions, the BMMR model determines equilibrium growth rates and equilibrium age distributions for the female and male populations. It also determines who marries whom or, more precisely, the age pattern of assortative mating. The BMMR model assumes that all fertility is marital fertility and all unions are monogamous, although both of these assumptions are easy to relax. Because the BMMR model allows changes in the age structure of the population (e.g., shortages of marriageable men relative to marriageable women) to change the number of unions, it accommodates our common sense conclusion about the effect of a marriage squeeze. Throughout a marriage squeeze, the birth matrix (i.e., the number of births to each type of union) remains constant, but the number of unions changes, and these changes in the number of unions imply changes in the number of births and, hence, changes in the implied age-specific female fertility rates.

2.1 Linear and Nonlinear Models

Because the classical stable population theory utilizes a linear model, the mathematics is straightforward. The age-specific fertility and mortality schedules for the female population together define a mapping of the female population vector (by age) in period t

into the female population vector in $t+1$. Because the implied mapping is linear, demographers use the simple and powerful tools of linear algebra to investigate the existence, uniqueness, and dynamic stability of the equilibrium. The matrix corresponding to the mapping of the population vector in period t into the population vector in period $t+1$ is called the Leslie matrix. The equilibrium age distribution is an eigenvector of the Leslie matrix; the equilibrium growth rate is obtained by subtracting 1 from the corresponding eigenvalue. In linear demographic models, cases in which an equilibrium fails to exist, cases of multiple equilibria, and cases in which the dynamics do not exhibit global stability are generally regarded as pathological.

The BMMR model, like the classical stable population theory, defines a mapping of the population vector in period t into the population vector in period $t+1$. The initial population vector, together with the mating rule, determines the number of unions of each type; the number of unions of each type, together with the birth matrix, determines the number of newborns in the next period. The initial population vector, together with the mortality schedules, determines the number of individuals in each of the other age-sex categories, just as in the classical stable population theory.

Nonlinearity is a necessary feature of any two-sex demographic model. The two-sex problem arises precisely because the classical stable population theory fails to recognize that the effect on population in period $t+1$ of adding a female of age i to the population in period t depends on the number of males. In the BMMR model, this dependence is captured by the mating rule. Any mating rule for monogamous unions must satisfy two accounting requirements: a nonnegativity condition ensuring that the number of unions of each type is positive or zero, and an adding-up or balance condition ensuring that the number of mated females of each age does not exceed the total number of females of that age, with a similar requirement holding for males. This adding-up requirement is the fundamental source of nonlinearity in the BMMR model.

Because the BMMR model is inherently nonlinear, establishing the existence of equilibrium and analyzing dynamic behavior in the BMMR model is more complex than in the classical stable population theory. A 'natural' but unsuccessful strategy for establishing the existence of equilibrium is to apply a fixed-point theorem to the mapping of the population vector in period t into a suitably normalized population vector in period $t+1$; a fixed point of such a mapping—a population vector that maps into itself—is an equilibrium of the BMMR model. The difficulty with this strategy is that the mapping defined by the BMMR model carries some population vectors into the 0 vector and, hence, this strategy fails to establish the existence of a nontrivial equilibrium. An alternative proof strategy, followed in Pollak (1986), avoids this difficulty by drastically limiting the domain of the mapping and reducing the problem to a single dimension.

2.2 Uniqueness and Dynamic Stability

Multiple nontrivial equilibria are not pathological in nonlinear models, and Pollak (1990a) demonstrates by example that the BMMR model can have multiple nontrivial equilibria. An overly strong apparently innocuous sufficient condition for the BMMR model to have a unique nontrivial equilibrium is that at every population vector an increase in the number of individuals in each age-sex category does not decrease the number of newborns in the next period.

Uniqueness and dynamic stability are intimately related. In a model with multiple equilibria, initial conditions determine which, if any, of several equilibria will be realized, and small differences in initial conditions can lead to large differences in long-run behavior. Pollak (1990a) discusses four sets of demographically meaningful conditions that imply local stability. Local stability rather than global stability is the relevant issue because, inter alia, some initial population vectors must lead to the trivial equilibrium. Further work—both theoretical and empirical—is required to determine whether any of the sufficient conditions for uniqueness and dynamic stability are satisfied and whether multiple equilibria or dynamic instability are realistic possibilities in two-sex demographic models.

Simulation is a possible approach to investigating these issues, but simulation techniques are much less informative in models that lack parametric specifications. In linear models, parametric specification is a feature of the model itself, but to simulate the BMMR model, we must first specify a functional form for the mating rule, and then specify the appropriate parameter values. Explicit functional forms for mating rules are discussed in Pollak (1990b).

3. Behavioral Models

Further work is also required to transform the BMMR model from a formal into a substantive model of population age structure and growth. The transformation requires importing behavioral theories from social science into the BMMR model to explain its three primitives—the birth matrix, the mating rule, and the mortality schedules. (In the generalized version of the model in which unions can persist for more than one period, there is a fourth primitive requiring a behavioral explanation: the schedule specifying the probabilities that unions of each type will end in desertion or divorce: see Pollak 1987.) From the

standpoint of social science, however, these three primitives are different kinds of analytical constructs. The mortality schedule is often regarded as a biological datum, although there is ample precedent (from Malthus to recent concern about excess female infant mortality rates in India) for regarding mortality as endogenous. The elements of the birth matrix reflect the decisions of individuals or of families facing economic and biological constraints. The analysis of these decisions is the subject matter of economic and sociological theories of fertility.

Whether we interpret 'unions' narrowly or broadly, the mating rule of the BMMR model is more complex analytically than the mortality schedules or the birth matrix because it reflects not only individual behavior but also the interactions of individuals in the 'marriage market.' The marriage market metaphor is useful because it suggests notions of choice, competition, and equilibrium. That is, the mating rule is not a full-blown theory of marriage or union formation but a reduced form, presumably reflecting the requirements of equilibrium in the marriage market.

There are two approaches to modeling competition and equilibrium in marriage markets, one based on matching and the other based on search. Mortensen (1988) and Weiss (1997) provide accessible introductions to both matching and search approaches; Roth and Sotomayor (1990) provide the definitive treatment of matching models. Both matching and search models begin with the preferences of individuals and analyze their implications for marriage-market equilibrium. They differ, however, in their assumptions about the information available to and the alternatives faced by marriage-market participants. Matching models are nonstochastic and timeless; they implicitly assume that complete and accurate information about potential spouses is instantaneously and costlessly available to all marriage-market participants. Search models, in contrast, are stochastic and dynamic; they recognize explicitly that individuals have less than complete and accurate information about potential spouses, and that acquiring more and better information takes time and, in some search models, requires resources. In a search model, when two individuals meet in the marriage market, they must decide whether to marry or to continue searching, balancing the attractiveness of marriage now to a particular individual against the expected value of remaining unmarried and continuing to search. Because search models do not assume that individuals have complete and accurate information, they may appear to be more 'realistic' than matching models. The price that search models pay for added realism is their sensitivity to specific assumptions about the rate at which potential spouses meet in the marriage market, the impatience or time preference of individuals, and individuals' beliefs about the characteristics and beliefs of other marriage market participants.

Ecologists may find models that allow the possibility of a positive equilibrium rate of population growth unsatisfactory because such models fail to recognize density dependence and other factors that prevent population growth from continuing indefinitely.

4. Summary

To summarize: classical stable population theory is parsimonious both because it allows us to use well-known, powerful mathematical techniques to investigate existence and dynamic stability, and because it allows us to infer a population's equilibrium age structure, growth rate, and dynamic behavior from very little data. The birth matrix-mating rule model, which allows fertility rates to depend on the population's age-sex structure, is more complex analytically and more demanding in its data requirements. In return for these extravagances, the BMMR model solves demography's two-sex problem and provides a framework for addressing the marriage squeeze and other important issues in demography that require a two-sex model.

See also: Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology

Bibliography

Feeney G M 1972 Marriage rates and population growth: The two-sex problem in demography. Ph.D. thesis, University of California, Berkeley, CA

Mortensen D T 1988 Matching: Finding a partner for life or otherwise. In: Winship C, Rosen S (eds.) *Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure. American Journal of Sociology, Supplement*. University of Chicago Press, Chicago, pp. S215–40

Pollak R A 1986 A reformulation of the two-sex problem. *Demography* **23**: 247–59

Pollak R A 1987 The two-sex problem with persistent unions: A generalization of the birth matrix-mating rule model. *Theoretical Population Biology* **32**: 176–87

Pollak R A 1990a Two-sex demographic models. *Journal of Political Economy* **98**: 399–420

Pollak R A 1990b Two-sex population models and classical stable population theory. In: Adams J, Lam D A, Hermalin A I, Smouse P (eds.) *Convergent Issues in Genetics and Demography*. Oxford University Press, New York

Roth A E, Sotomayor M A O 1990 *Two-sided Matching: A Study in Game-theoretic Modeling and Analysis*. Cambridge University Press, Cambridge, UK

Weiss Y 1997 The formation and dissolution of families: Why marry? Who marries whom? And what happens upon divorce. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics, Vol. 1A*. Elsevier, Amsterdam

R. A. Pollak

Population Ecology

One part of this research program examines the demography of organizations. It finds that organizational mortality processes depend upon the age and size of the organization, as well as on characteristics of populations and environments. Moreover, there is evidence of an imprinting process, meaning that environmental conditions at certain early phases in an organization's development have long-term consequences. In particular, organizations subject to intense competition have elevated mortality hazards at all ages.

A central theme is structural inertia, the tendency for organizations to respond slowly relative to the speed of environmental change. A central argument holds that inertia derives from the very characteristics that make organizations favored kinds of actors in modern society: reliability and (formal) accountability. It follows that changes in an organization's core features are disruptive and increase mortality hazards, at least in the short-run. Research on this subject tends to support this view.

The concept of niche provides a framework for relating environmental variations and competition to population dynamics and segmentation. Much empirical work examines the niches of organizational populations in terms of dimensions of social, political, and economic environments.

Most research in this field builds on theories of resource partition and of density dependence. Resource-partitioning theory concerns the relationship between increasing market concentration and increasing proliferation of specialists in mature industries. The key implication of this theory concerns the effects of concentration on the viability of specialist organizations (those that seek to exploit a narrow range of resources). The theory of density-dependent organizational evolution synthesizes ecological and institutional processes. It holds that growth in the number of organizations in a population (density) drives processes of social legitimation and competition that, in turn, shape the vital rates. This argument implies certain nonmonotonic relationships between density and rates of founding and mortality. Both theories have been subject to empirical test in diverse populations of organizations; the preponderance of evidence supports both.

Current research emphasizes formalization (often using tools of formal logic), integration of methods of social network analysis and ecological analysis, and developing implications of organizational ecology for other aspects of social structure.

The population ecology of organizations, commonly called organizational ecology (OE) examines interactions within and between populations of organizations. This approach differs from other sociological research by focusing on the population level, relying on selection logic, and studying the life histories of all organizations in populations studied over their full histories.

This approach built initially on neo-classical population bioecology, which analyzes numerical aspects of population interactions from an evolutionary perspective. In a paraphrase of bioecologist G. E. Hutchinson, Hannan and Freeman (1977) posed the following orienting question of OE: Why are there so many kinds of organizations? They suggested that the social science literature lacked good answers to this question and that seeking answers would clarify the dynamics of the organizational world. Subsequent research has shown that organizational diversity has important consequences for individuals and social structures (Carroll and Hannan 2000).

OE sought initially to resolve a tension in the human ecology concerning what Hawley (1968) termed the principle of isomorphism, 'Units subjected to the same environmental conditions or to environmental conditions as mediated through a given key unit, acquire a similar form of organization.' Hannan and Freeman (1977) argued that Hawley's principle, like other adaptationist premises, does not apply straightforwardly in uncertain, heterogeneous environments. Extending the received wisdom about adaptation requires specification of the underlying dynamic processes and attention to processes of selection and segmentation.

1. *The Basic Demography of Organizations*

With its concentration on the vital rates of founding and mortality, OE has a strong demographic character. Key structuring processes are specified at the population level. Although mortality histories can always be associated with individual organizations, histories of foundings cannot because the absence of foundings in some period or place must be considered.

Populations must have unit character to serve as useful objects of analysis (Hawley 1968). Researchers generally regard populations as spatial–temporal instantiations of organizational forms, templates for organizing. Several conceptions of the basis of form

have guided these efforts: common resource dependence, structural equivalence in network position, and identities coded in cultural rules.

Constructing organizational demography and ecology requires attention to differences between organizations and organisms. In particular, organizations experience more varied 'birth' and 'death' events, typically lack clear parents, have multilayered, partly decomposable structures, and can transform themselves. The major methodological project has involved dealing with these complexities (Carroll and Hannan 2000).

The resulting strategy examines the full histories of organizational populations, because experience shows that early events can have lasting consequences for population dynamics. It gathers life history data on all organizations in the population(s), including the large and famous as well as the small and insignificant—this is crucial for avoiding problems of selectivity bias. It records detailed information about type of entry (new founding, entry from another industry, merger, etc.) and exit (disbanding, acquisition, transformation, etc.). Finally, it uses event-history methods to estimate the effects on vital rates of characteristics of organization, population, and environment. This research strategy has diffused widely to areas of organizational research that do not necessarily share theoretical commitments.

Ecological research finds that many more organizations have operated in well-studied industries than is usually assumed. Careful archival searches usually find information on hundreds of organizations that get overlooked in published counts. They also reveal that organizational life expectancies are generally an order of magnitude smaller than for individuals.

Much attention has been paid to organizational aging. Stinchcombe (1965) observed that organizations (and organizational populations) experience a liability of newness. Efficient organization requires trust among members, which takes time to build. Creating roles and routines, learning about the environment, and developing relationships with other organizations also take time. Once established, these patterns of relationships enhance survival chances. Much early research found this pattern. Some later research found a variation, a liability of adolescence: the hazard rises during the early lifespan—while initial stocks of endowments get exhausted—before declining. Recent research, using designs that measure (age-varying) organizational size, has produced (mixed) evidence in favor of positive age-dependence (Carroll and Hannan 2000). Two interpretations for this pattern have been proposed. If inertial forces are strong, then the possibilities of adapting to changing environments are limited and older cohorts of organizations have lower fitness—there is a liability of obsolescence. Alternatively, the accumulation of rules and routines impedes adjustment to environmental change—the liability is one of senescence. Current research seeks to clarify the conditions under which the various patterns arise—see later.

2. *Inertia and Selection*

Stinchcombe's (1965) insight that new organizations get imprinted by their environments provided inspiration for developing the ecological approach. According to Stinchcombe's argument, new organizations get tested against assumptions that are taken for granted but change over time. Hence, the characteristics of organizations that pass such tests and emerge reflect the social structure of the time. Imprinting requires both a mapping of environmental conditions onto the organization and inertia in the imprinted characteristics. Absent the second condition, later adjustments will erode the association between founding conditions and those features.

OE has built on this image of inertia (relative to the speed of environmental change). According to Hannan and Freeman (1989), inertia prevails as an inadvertent by-product of a selection process favoring the properties of reliability (low variance in the quality of performance) and accountability (the ability to construct rational accounts for actions). Achieving these properties depends on structures being reproduced faithfully over time. Yet, high reproducibility means that structures resist transformation. Therefore, this selection process inadvertently favors corporate actors with strong inertial tendencies.

Fundamental change presumably diminishes reliability and accountability. Old learning loses value, and new processes must be learned. Vestiges of the old system conflict with the new one emerging. Thus, even if fundamental change has long-term benefits (the new form might be better aligned with environments and internal processes), change might increase mortality hazards in the short-run. If so, then reorganization-prone organizations have a lower probability of representation in future populations.

Much research has tested the implications of this argument. Most well designed studies reveal that changing core structures does increase the hazard of mortality in the short-run, often substantially (Carroll and Hannan 2000). Moreover, the magnitude of this effect increases with organizational age.

A different way of viewing imprinting yielded one of OE's most robust empirical findings: density delay. Organizations subject to intense competition at founding (as evidenced by a high density of organizations in the population—see later) have elevated mortality hazards at all ages. In other words, density at founding has a (partially) delayed effect on mortality processes. This relationship has been interpreted as reflecting enduring effects either of building structures under scarcity or of positioning at the margins of resource distributions.

3. Organizational Niches

The concept of niche provides a framework for relating environmental variations and competition to population dynamics and segmentation. A population's fundamental niche consists of the combinations of environmental parameters for which its growth rate (fitness) is non-negative in the absence of competition. Much empirical work in OE examines the niches of organizational populations in terms of dimensions of social, political, and economic environments (e.g., Dobrev 1999). Even when such issues do not stand in the forefront, ecologists routinely control statistically for the effects of such dimensions in examining dynamic processes.

Niche theory has also been central to research on competition (Hannan and Freeman 1977, 1989, McPherson 1983). Populations compete if the presence of each reduces the spaces in which the others can sustain themselves (their realized niches). By this reasoning, populations compete if and only if their fundamental niches intersect. The assumed equivalence of niche overlap and competition has played a crucial role in allowing organizational ecologists to relate empirical observations on realized niches to models of population dynamics.

Niche width concerns the variance of resource utilization. Does evolution favor populations of Jacks-of-all-trades and masters of none (generalists) or specialists at a few tasks? Research shows that the organizational world tends toward segmentation into specialist populations (each highly isomorphic) when relevant environmental variations are uncertain, large in magnitude, and long in duration (Hannan and Freeman 1989, Péli 1997).

4. Resource Partitioning

Resource-partitioning theory, OE's best-developed theory of segmentation, concerns the relationship between increasing market concentration and increasing proliferation of specialists in mature industries. The theoretical imagery concerns crowding in a heterogeneous market. Organizations initially target resource segments. Specialists choose narrow, homogeneous targets; generalists choose broad, heterogeneous ones.

The theory assumes an advantage of scale (in production, marketing, or distribution). Because organizations targeting narrow segments cannot grow very large, they have higher (per unit) costs. Organizations with scale advantages expand products to broaden their appeal without losing established domains. Small organizations then face formidable invaders, and many fail. Competition is most intense in resource-dense centers, resulting in dominance by large generalists. Failures of smaller generalists cause some resources to become free. Because larger generalists can best secure these resources, the surviving generalists become larger and more general. However, generalists can rarely secure all of the newly freed resources. As the struggle among generalists proceeds to its monopoly equilibrium, the surviving generalists grow and broaden, but the set of resources they hold collectively actually declines. In other words, more resources become available for specialists as the market concentrates. Therefore, as concentration rises, the viability of specialist organizations increases as well: founding rates rise and mortality rates fall.

The empirical implications of this argument involve an interaction effect between organizational form (specialism–generalism) and concentration in affecting each vital rate. Empirical studies on diverse populations support this hypothesis (Carroll and Hannan 2000). Populations studied include American airlines, brewers, investment banks, microprocessor manufacturers, newspaper publishers, telephone companies, and wineries, as well as British automobile manufacturers and Italian cooperative banks (full citations can be found in Carroll and Hannan 2000).

5. Density Dependence

The theory of density-dependent organizational evolution synthesizes ecological and institutional processes. It holds that growth in the number of organizations in a population (density) drives processes of social legitimation and competition that, in turn, shape the vital rates (Hannan and Carroll 1992). An organizational form gains (constitutive) legitimation when it becomes taken-for-granted (Meyer and Rowan 1977). New organizational forms generally lack such legitimation, which makes organizing difficult. If the form gets promulgated, legitimation increases. Initially, when there are few organizations in a population and the form lacks legitimacy, the returns to legitimation of adding another organization are great. However, when the population contains many organizations, constitutive legitimation increases little—if at all—as density rises further. In other words, a population's legitimation increases with density at a decreasing rate and approaches a ceiling.

Competition refers to the consequences of efforts by multiple agents to acquire the same limited resources. Density-dependence theorists focus on diffuse competition because it has a clear relationship with density as the intensity of diffuse competition within an organizational population rises as a function of the number of potential bilateral competitors. As density increases linearly, the number of possible competitive relations increases geometrically. Thus diffuse competition rises with density at an increasing rate. The theory posits that founding rates are directly proportional to legitimation and inversely proportional to the intensity of diffuse competition and that these relationships are inverted for mortality hazards.

This theory (formalized appropriately) yields two theorems. First, as density increases, founding rates rise initially, reach a peak, and then decline. Second, the relationship for mortality is the mirror image. These theorems, expressed in parametric forms, can be falsified empirically. They have been tested for diverse populations, including automobile manufacturers in Europe and the USA, brewers in Germany and the USA, banks in Denmark, Japan, Singapore, and the USA, and American labor unions, life insurers, newspapers, and social movements of women and ethnic groups. These studies have supported the theory in most cases (full citations can be found in Carroll and Hannan 2000).

6. *Emerging Trends*

Current research seeks to sort out these varied arguments, learn about conditions of applicability, and (when possible) unify them. One strand begins by reconstructing the fragments in formal languages. In the case of age dependence, a partial integration has been achieved in a formulation in first-order logic (Carroll and Hannan 2000). The next steps require use of nonmonotonic logic, which allows premises to be overridden by more specific ones so that the conclusions of an argument can change as more knowledge about specifics becomes available (Pólos et al. 1999).

Efforts are underway to rebuild the foundations of the research program on an improved formal representation of organizational forms, with forms based on culturally coded and enforced identities (by Pólos et al. 2002). In this construction, density dependence plays a role in the cultural coding process, as do social movements. Processes of enforcement of identities generate inertial mechanisms as well as punishments for fundamental changes (those that violate codes).

Basing forms on identities helps clarify issues in resource partitioning. Specialization based on claims to authentic identity (as in the case of the rise of specialist producers of microbrewed beer in the USA and 'real ale' in the UK) can prevent large generalists from gaining complete control of concentrated markets. Other research seeks better representations of dynamics and segmentation. The formal model of density-dependence serves as the foundation for many of these efforts. One approach rescales the effects of density using a specification of subniches to weight density. Some research weights densities according to an organization's position within a broad technology class or by licensing restrictions, location, and price (Baum and Singh 1994). Alternatively, the weightings are set to reflect the strength of competitors (Barnett 1997). These specifications generate refined understandings of dynamics (at the cost of considerable complexity).

A second variation envisions competition as more localized than legitimation, because political authorities can better impede flows of products than ideas. For instance, research on the European automobile industry shows competition operating primarily within national markets but legitimation operating throughout Europe (Carroll and Hannan 2000). The third has shown the effects of density change systematically over a population's history, either directly as a function of population age, reflecting the growing importance of identities, statuses, and reputations in affecting life chances in mature populations or as a function of industry concentration (Barron 1999), in the spirit of resource-partitioning.

The empirical scope of OE also continues to expand. Current research focuses on a diverse set of industries. In addition to those already mentioned, these include: accountancies, bicycle manufacturers, computer manufacturers, disk-drive producers, footwear manufacturers, hotels, law firms, magazine publishers, paper manufacturers, sex shops, and television stations. A major thrust of current work is collecting data on life histories of organizations and on connections among the organizations (including flows of personnel among them). These studies highlight the value of social network analysis for understanding the ecology of organizations.

New processes are also being brought into the picture. Organizational status has been shown to affect life chances, with the impact increasing with the uncertainty in evaluations of quality (Podolny et al. 1996). Social movement research relies increasing on OE as a guide to modeling dynamics (Minkoff 1997); ecologists, in turn, increasingly view social movements as pivotal in the emergence of forms (Wade et al. 1998). Studies of careers find that organizational demography induces considerable job mobility (Haveman and Cohen 1994); at the same time, careers of managers shape strategies and thereby influence life events for organizations (Sørensen 1999).

See also: Competitive Strategies: Organizational; Darwinism: Social; Ecological Economics; Ecological Imperialism; Ecology: Organizations; Ecology, Political; Resource Geography

Bibliography

Barnett W P 1997 The dynamics of competitive intensity. *Administrative Science Quarterly* **42**: 128–60

Barron D N 1999 The structuring of organizational populations. *American Sociological Review* **64**: 421–45

Baum J A C, Singh J V 1994 Organizational niches and the dynamics of organizational mortality. *American Journal of Sociology* **100**: 346–80

Carroll G R 1985 Concentration and specialization. *American Journal of Sociology* **90**: 1262–1283

Carroll G R, Hannan M T 2000 *The Demography of Corporations and Industries*. Princeton University Press, Princeton, NJ

Dobrev S D 1999 The dynamics of the Bulgarian newspaper industry in a period of transition. *Industrial and Corporate Change* **8**: 573–605
Hannan M T, Carroll G R 1992 *Dynamics of Organizational Populations*. Oxford University Press, New York
Hannan M T, Freeman J 1977 The population ecology of organizations. *American Journal of Sociology* **82**: 929–64
Hannan M T, Freeman J 1989 *Organizational Ecology*. Harvard University Press, Cambridge, MA
Haveman H A, Cohen L E 1994 The ecological dynamics of careers. *American Journal of Sociology* **100**: 104–52
Hawley A H 1968 Human ecology. In: Sills D (ed.) *International Encyclopedia of Social Sciences*. Macmillan, New York
McPherson J M 1983 An ecology of affiliation. *American Sociological Review* **48**: 519–35
Meyer J W, Rowan B 1977 Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology* **83**: 340–63
Minkoff D 1997 The sequencing of social movements. *American Sociological Review* **62**: 779–99
Péli G 1997 The niche hikers guide to population ecology. In: Raftery A (ed.) *Sociological Methodology*. Blackwell, Oxford, UK
Podolny J P, Stuart T E, Hannan M T 1996 Networks, knowledge and niches. *American Journal of Sociology* **102**: 659–89
Pólos L, Hannan M T, Carroll G R 2002 Foundations of a theory of social forms. *Industrial and Corporate Change* Forthcoming
Pólos L, Hannan M T, Kamps J 1999 Aging by default: Building theory from fragmentary knowledge. In: Rott H, Albert C, Brewka G, Witteveen C (eds.) *Proceedings of the Fourth Dutch–German Workshop on Non-Monotonic Reasoning Techniques*. ILCC Scientific Publications, Amsterdam
Sørensen J B 1999 The ecology of organizational demography. *Industrial and Corporate Change* **8**: 713–44
Stinchcombe A L 1965 Social structure and organizations. In: March J (ed.) *Handbook of Organizations*. Rand-McNally, Chicago
Wade J B, Swaminathan A, Saxon M S 1998 Normative and resource flow consequences of local regulations in the American brewing industry. *Administrative Science Quarterly* **43**: 905–35

M. T. Hannan

Population, Economic Development, and Poverty

1. Introduction

Assessing the consequences of population on the pace and process of economic growth is one of the oldest themes in the literature on economics. These assessments have varied enormously over time, spanning the highly pessimistic to the outright optimistic. A systematic review of the major studies in this literature represents a useful way to organize a survey of the consequences of demographic change. Such an approach places the population debates in perspective, and it infuses a healthy dose of caution in appraising current debates. Specifically, how have the 'bottom-line' appraisals of the consequences of demographic change on development changed over time; why have they changed; and what are the most recent contributions to this literature?

2. The Beginnings: Malthus (1798)

The debate began with two propositions by the Reverend Thomas Malthus in 1798 (*First Essay on Population*) that population would grow at a geometric rate (e.g., 1, 2, 4, ...) due mainly to a lack of conscious restraints on fertility, and that food would grow at an arithmetic rate (e.g., 1, 2, 3, ...) due substantially to diminishing returns to increasingly scarce land. The resulting outcome would be food shortages, starvation, and deaths. In the long run, population size would be held in check by food availability and mortality. Population pressures would constrain income per capita to a low level of subsistence—a 'Malthusian trap,' as it has been termed. These images caused economics, unfairly, to be dubbed the dismal science.

Fortunately Malthus' predictions were not sustained by the preponderance of experience over the next two centuries. Couples did not breed without restraint, but rather by consciously managing fertility in response to changing conditions. Food was not unduly constrained by land availability. Instead, technology blossomed and food expanded apace in the very geographic regions where Malthus focused his empirical studies. Ironically, food *surpluses* turned out to be a 'problem' confronting many nations, and governments implemented policies designed to curtail farm production.

Clearly the Malthusian ideas regarding population–economic linkages were incomplete, and richer analytical and empirical foundations were needed. The urgency for such a framework was made apparent by demographic events. By the mid-twentieth century, it was recognized that the simultaneous occurrence of declining mortality and exceptionally high and sustained fertility in scores of developing countries was resulting in high population growth rates. A concern emerged that these rates could not be sustained over long periods of time. While, as in the past, fertility would predictably decline (*a la* the Demographic Transition), it was unclear whether such a decline would be soon or rapid enough to avoid potentially deleterious effects on welfare, economic progress, and the environment. Thus, while the 'Malthusian Problem' reappeared, approaches to assessing population consequences assumed quite different tacks. It was time for a fresh reassessment.

3. Expanded Elaborations: 1950s, 1960s, 1970s

3.1 The United Nations (1953)

The first major installment that exposed economic–demographic interactions in detail was provided in a seminal study by the United Nations in 1953, *The Determinants and Consequences of Population Trends*. This study was both balanced and comprehensive. Twenty-one economic–demographic linkages were taken up. The impact of population on some factors was judged to be positive (economies due to scale and organization), on some negative (diminishing returns), and on some neutral (technology and social progress). Overall, a net negative impact of undetermined size was considered most likely in much of the developing world, although this conclusion was guarded, and the UN predicted a wide diversity of effects according to country-specific conditions.

3.2 Coale and Hoover (1958)

A second major installment in the evolution of demographic assessments appeared in 1958 in an influential book by Ansley J. Coale and Edgar Hoover (C/H) entitled *Population Growth and Economic Development in Low-Income Countries*. Based on the simulation results of a mathematical model calibrated by Indian data, C/H concluded that India's development would be substantially enhanced by lower rates of population growth. Their analysis rested on two premises. First, household and economy-wide saving would be diminished by large families; it certainly would not be stimulated enough in response to demographic changes to provide the increased capital required by an enlarged population. Second, economy-wide investments would be skewed away from relatively productive activities since funds would be shifted toward so-called unproductive population-sensitive social expenditures (e.g., health and education). This crowding out of productive investment was linked mainly to the youthful age composition of the high-fertility population, and not population size, density, or growth, *per se*.

The C/H study had a profound impact on US population policy. It also attracted the attention of academic scholars given its focus on physical capital (as distinct from the Malthusian focus on land), then considered by many to be a (the?) key to economic development.

The C/H study was not without challenge. Over time its impact waned, in part due to the appearance of empirical studies that failed to unearth empirically strong and consistent impacts of population pressures on Third World saving. This is not surprising. At the household level, such saving impacts are fundamentally based on a life-cycle conceptualization of behavior. This requires a lengthy forward looking planning horizon. It also requires an institutional structure (e.g., developed capital markets and/or government social security options) to implement lifetime plans. These conditions do not describe the day-to-day survival mode of most residents of the Third World. Indeed, spending on children represents a form of saving (e.g., parents may expect transfers from their children in old age); and moreover, children can be quite productive both in the household and on the farm.

A second challenge to the C/H framework was the appearance of empirical studies that downplayed the role of country-wide age structures as a determinant of spending on education. Even in the face of population pressures, school enrollments and attainments advanced at historically unprecedented rates in many developing countries. This surprisingly occurred without substantial realignments in social budgets. Nations found ways to economize on education resources, and tradeoffs within education budgets were more pervasive than was assumed in the simple C/H model. In short, the empirical underpinnings of the C/H framework were questioned by an increasing body of evidence. Additionally, over time economic theory itself was in flux, highlighting both the *productivity* of human capital (considered economically unproductive in the C/H model), and the importance of technical change (also absent in C/H). While these limitations do not diminish the importance of C/H's pioneering work, they served to shift the population debates in the ensuing decades.

3.3 National Academy of Sciences (1971)

The most pessimistic assessment of population consequences since Malthus appeared in a study organized by the National Academy of Sciences. The findings, *Rapid Population Growth*: *Consequences and Policy Implications*, appeared in two parts: Volume I, a short Executive Summary, and Volume II, a large set of papers compiled by a panel of distinguished scholars. The study's findings, as summarized by the Executive Summary, are baffling and difficult to interpret. They must be handled with care.

On the one hand, the Executive Summary conveyed a highly negative, indeed an outright alarmist, assessment of the impact of rapid population growth. Twenty-five separate adverse impacts were listed and assessed; no notable positive effects of demographic change were identified. Moreover, there was a substantial disconnection between the Executive Summary and the findings of the key scholars participating in the study, as documented by the papers found in Volume II. The Executive Summary, whose pessimistic conclusions obtained wide publicity in the 1970s, was demonstrably unfaithful to the scholarly studies in

Volume II. The Summary was not vetted with the scholars who participated in the study, and indeed its authorship is unknown to this day, in spite of extensive attempts to clarify the historical record (an exhaustive historical assessment of this episode is provided in Kelley 2001). Clearly the assessment of the Executive Summary must be strongly discounted. In spite of this, however, a careful reading of the papers and supporting components of the Report do reveal an important insight that assists in illuminating the ebb and flow of population assessments over time.

Specifically, the Executive Summary was by admission based mainly on 'direct,' *shorter-run* impacts of demographic change. ('We have limited ourselves to relative short-term … issues,' p. vi). In contrast, the major research papers supporting that study in Volume II, which were much less pessimistic, were based on a *longer-run* focus. More than any other factor, the outcomes of the population debates from the 1970s to the present day would turn on this difference in time perspective. *Direct, shorter-run impacts of demographic change are almost always attenuated* (*and sometimes even offset*) *by feedbacks that occur only over longer periods of time*.

3.4 *United Nations* (1973)

In 1973 the United Nations updated its early assessment. This revision arrived at a less eclectic, and a somewhat more pessimistic (but by no means alarmist) evaluation of the various impacts of population growth. This is particularly true of anticipated difficulties of feeding the expanding populations (reverting to traditional Malthusianism), and of pressures on capital formation (reverting to the concerns of Coale and Hoover). In both cases, short-run, direct impacts were highlighted. The report downplays some indirect, longer-run feedbacks that will likely occur due to price responses, and induced innovation and institutional change.

The most important new contribution to the population debate deriving from the 1973 study was a finding by Simon Kuznets that, based on simple correlations, a net negative impact of population growth on per capita output was not obvious in the data. While his work was based on *longer-run* assessments, and while they were appropriately qualified, they were important to conditioning the bottom-line UN assessment. Given the strong priors of demographers and policy makers that the negative impacts of population growth on development were large, the inability to easily confirm this hypothesis through simple, albeit inconclusive, correlations more than any other factor kept the population debate alive during the ensuing decades. The stage was set for a new round of debate.

4. *Revisionism: The 1980s*

4.1 *Julian L. Simon* (1981)

This debate was initiated in 1981 with the publication of Julian L. Simon's *The Ultimate Resource*, which challenged the pessimistic views of economic–demographic interactions. The significant impact of Simon's book is explained by several factors. First, it concluded that population growth was likely to exert a *positive* net impact on economic development in many Third World countries *in the intermediate run*—a startling assertion that attracted extensive attention. Second, it illustrated that the outcome of population impacts on the economy are likely to hinge both on the *time* dimension of the assessments, and whether *feedbacks* are included in the analysis. Third, it reintroduced to the debates the roles of alternative demographic impacts, specifically those of density and size, to add to the popular focus on population growth and age structure. Fourth, and importantly, the book elevated the role of *technology* as a (the?) driving force of economic growth. Simon asserted that demographic pressures would favorably influence the nature and pace of technological change. Finally, it was written in a highly accessible, confrontational debating style that goaded and infuriated his skeptics and amused and emboldened his supporters.

These various attributes of Simon's framework are best illustrated by his analysis of the likelihood that demographic pressures would ultimately cause natural resource exhaustion. He illustrated with numerous graphs that over longer periods most natural-resource prices actually declined, even in the face of rising demands stimulated in part by expanding populations. Price-induced substitutions in production and consumption of natural resources, and an expansion of supply through discovery and technical advance, are offered to explain this result. Most importantly, these 'adjustments' or 'feedbacks' were in part *caused by* population pressures, a reality that could be exposed only by extending the analysis over several decades. It is this longer-run perspective, and the *feedbacks* in markets and other institutions, that account for Simon's findings.

Similarly, Simon hypothesized and attempted to document that the pace of technical change, and its bias, were actually stimulated by population pressures, particularly in agriculture where Ester Boserup's (1965) model, first expounded in *The Conditions of Agricultural Growth*, was advanced. She proposed that increasingly productive agricultural technologies are made economically feasible only at higher land densities, and therefore population pressures against increasingly scarce land could have a powerful positive effect on agricultural output growth. Simon extended this notion to observe that major social overhead projects (roads, communications, irrigation) benefited from expanded populations and scale.

While Simon's book was highly influential in its own right, arguably its greatest impact was as a catalyst in setting the stage for reassessing the basic propositions of the population debates. The 1980s witnessed a flurry of major reassessments.

4.2 *National Research Council* (1986)

The most important and influential reassessment was compiled by a Working Group on Population Growth and Economic Development under the auspices of the National Research Council (NRC). The bottom-line assessment by this group was eclectic, balanced, and nonalarmist:

> On balance, we reach the qualitative conclusion that slower population growth would be beneficial to economic development of most developing countries (p. 90).

This carefully worded statement illustrates well the central elements of all of the major population assessments in the 1980s: (a) there are both important positive and negative impacts of population growth (thus, 'on balance'); (b) the actual size of the net impact—and even whether it is strong or weak—cannot be determined given current evidence (thus 'qualitative'); (c) only the direction of the impact from high current growth rates can be discerned (thus 'slower,' and not 'slow'); and (d) the net impact varies from country to country—in most cases it will be negative, in some it will be positive, and in others it will have little impact one way or the other (thus, 'most developing countries').

It is interesting to speculate on factors that account for the guarded and qualified assessment of the NRC group. Three factors can be singled out.

First, the Summary Report highlights *individual and institutional responses* to initial impacts of population change—conservation in response to scarcity, substitution of abundant for scarce factors of production, innovation and adoption of technologies to exploit profitable opportunities, and the like.

Second, the study was compiled almost exclusively by economists, whose faith in the potential for market-induced responses to modify initial direct impacts of population change is greater than that possessed by other social and biological scientists.

Third, research which had been accumulating over the past 10 to 15 years downplayed many of the previous concerns voiced on the impacts of rapid population growth. For example, as noted by Simon and others, the concern that population growth results in the exhaustion of non-renewable natural resources is misplaced. The relationship between population growth and global resource use is not as strong as had been assumed.

Similarly, the concern about a substantial reduction of saving due to rapid population growth is not confirmed by the data. While some capital shallowing occurs, the impact of this on economic growth is not particularly strong. Finally, the concern that population growth will significantly shift resources from productive physical capital formation into alleged 'less-productive' areas was not sustained by the data. The financing of educational enrollments, which expanded significantly even in the face of population pressures, came from some combination of increases in public (sometimes deficit) spending, reduction in per pupil expenditures, and efficiency gains.

Interestingly, a new concern was elevated in importance by the 1986 Report in the area of renewable resource degradation. Unlike non-renewable resources, property rights are difficult to assign or maintain for renewable resources like fishing areas and rain forests. Overuse can result. The problem is not population growth *per se*, but rather institutional failure. Cutting population growth in half would not solve the problem; it would only postpone the deleterious outcome of resource degradation. The solution requires policies to focus on the *causes* of failure (the need for more effective property rights, market responses, and government policies to correct externalities) rather than a focus on population pressures, which mainly exacerbate the environmental responses.

4.3 *Other Studies in the 1980s*

The bottom-line assessments of five other studies in the 1980s merit citing. All reflect the balanced, longer-run methodological perspective of the 1986 National Research Council report. Specifically, all highlight direct and indirect responses to population pressures. The resulting bottom-lines, while assigning a negative net impact of rapid population growth are, on average, both qualified and guarded.

The World Bank's *World Development Report* 1984 is concerned most about countries that exhibit exceptionally rapid growth rates: '... population growth—at rates above 2 percent ... —acts as a brake on development' (p. 79), although, 'Up to a point, population growth can be accommodated ...' (p. 79).

An overall appraisal of both the *World Development Report* 1984, and the National Research Council report two years later, is offered by Dr. Nancy Birdsall, who headed the World Bank study and was also a member of the NRC Working Group. She concluded that 'Rapid population growth can slow development, but only under specific circumstances and generally with limited or weak effects' (Birdsall 1988, p. 529).

Allen C. Kelley's 1988 survey for the *Journal of Economic Literature* concludes that '... economic growth ... would have been more rapid in an environment of slower population growth, although in a number of countries the impact was probably negligible and in some it may have been positive' (p.

1715). Adverse impacts are most likely to occur where (a) water and arable land are scarce, (b) property rights are poorly defined, and (c) government policies are ineffective and biased against labor.

T. N. Srinivasan's 1988 survey parallels the conclusion of the NRC Report. He emphasizes caution in the formulation of population policies, noting that 'many of the alleged deleterious consequences [of population growth] result more from inappropriate policies and institutions than from rapid population growth. Thus policy reform and institutional change are called for, rather than policy interventions in private fertility decisions to counter these effects' (p. 7).

It appears that in the 1980s a broad consensus was struck on the net assessments of demographic change. Most analysts held that while slower population growth would indeed advance the economic progress of most developing countries, the size of the net impacts would not likely be especially remarkable by comparison with numerous other determinants of economic growth. The shift of the research emphasis during the decade to highlight the importance of (a) *time* in the analysis (modifying shorter-run direct impacts of demography with feedbacks occurring over the longer run), and (b) the need to account for the separate components of demographic change (births, deaths, age, size, density), combined to put the research programs on solid footing. While these revisionist renderings and methodologies were widely embraced, uncertainties still remained, most importantly the need to obtain more precise *quantitative* assessments, especially those that accounted both for the components of demographic change and the timing of their impacts.

5. *New Paradigms: The 1990s and Beyond*

While most of the 1990s was preoccupied with digesting the revisionist results of the 1980s, population research did advance in several areas.

First, the findings from simple correlations between the rate of population and per capita economic growth appeared to have changed. While a general *lack* of correlation was the widely obtained statistical result for the 1960s and 1970s, in the 1980s the correlation turned negative (see Kelley and Schmidt 1994). On the one hand, most analysts agreed that such simple correlations are difficult to interpret, plagued as they are by failure to adequately account for reverse causation, excessive reliance on cross-section data, sensitivity to the selection of countries, the complexity of demographic linkages that are poorly modeled, spurious correlation, econometric pitfalls, and data of dubious quality. On the other hand, the previous finding of *no* correlation for the 1960s and 1970s in the face of strongly held priors of a negative correlation literally kept the population debate alive. Now, a *change* in this relationship from one of no correlation to one of a negative correlation for the 1980s required an explanation. New questions appeared: what accounts for the changed correlations; are the new results robust; are they quantitatively important?

The ability to address these issues coincided with the emergence in the 1990s of empirical 'convergence' models of economic growth. Pioneered by Robert Barro, these empirical paradigms distinguish between factors (economic, political, social, institutional, geographic) that determine each country's *long-run* level of per capita output, and the shorter-to-intermediate-run *transition* of countries to this longer-run state. These models lent themselves to investigating the impacts of demography since they exposed both short- and long-run impacts.

Efforts to model demography using the new convergence models have varied notably. Barro (1997), for example, looked primarily on the longer-run impacts of demography, and found that reductions in the total fertility rate increased the potential for economic growth.

In yet an earlier study, Kelley and Schmidt (1995), building on the Barro core variables, distinguished between several alternative demographic influences on the economy's potential output in the long-run, (e.g., the impacts of population size and density), and timing of demographic impacts (e.g., the timing of reductions in birth and death rates) which influence both the short and long run. These timing specifications highlighted the reality that birth-rate reductions have an immediate positive impact on growth by economizing on child-rearing expenses while, in fifteen or so years, the impacts will be reversed, since there will be fewer persons entering their productive work force years. Death rate reductions, especially infant/child mortality, can have similar timing impacts. Kelley and Schmidt found that these timing/transitional features were empirically important, and the long-run impacts of both density and size added modestly to growth as well.

Bloom and Williamson (1998), also building on Barro's empirical framework (although with different core variables highlighting policy and geography), modified the demographic modeling to break out an accounting reckoning of age-compositional impacts. While explicit modeling of longer-run demographic impacts is absent in their framework, their clean accounting framework clearly exposes the impacts of changing age structures, driven by changes in fertility and mortality. These are quantitatively important impacts on the *transition* to long-run output per capita. Their results focused on East Asia where declines in fertility were rapid and shorter-run transition effects are predictably large.

Finally, Kelley and Schmidt (2001) compared the above (and other) modeling efforts in a single empirical investigation, and came up with a somewhat surprising result: demography accounts for around 20 percent of

changes in output per capita growth from 1960–95 across a wide collection of countries. While for several reasons they consider their findings qualified, it is interesting that these findings are broadly consistent with those of the 1980s. Population *does* matter: its impact is likely adverse over the period 1960–95; this impact varies from decade to decade; components of demographic change exert both positive and negative impacts; these impacts vary notably from place to place; and, as a determining variable of long-run economic prosperity, population's impact is notable, but not remarkable. In the shorter-to-intermediate run, during periods of transition (both demographic, and economic), populations impact can be elevated or diminished, depending on the pace of demographic change and especially on the country's specific institutions (government policy, efficacy of markets, definition of property rights).

See also: Boserup, Ester (1910–99); Demographic Transition, Second; Development: Socioeconomic Aspects; Economic Growth: Theory; Malthus, Thomas Robert (1766–1834); Savings Behavior: Demographic InfluencesPopulation Geography

Bibliography

Barro R J 1997 Determinants of economic growth: A cross-country empirical study. Development Discussion Paper No. 579, Harvard Institute for International Development, April

Birdsall N 1988 Economic approaches to population growth and development. In: Chenery H B, Srinivasan T N (eds.) *Handbook of Development Economics*. Elsevier Science, Amsterdam

Birdsall N Kelley A C Sinding S (eds.) 2001 *Demography Matters: Population Change, Economic Growth and Poverty in the Developing World*. Oxford University Press, Oxford, UK

Bloom D E, Williamson J G 1998 Demographic transitions and economic miracles in emerging Asia. *World Bank Economic Review* **12**: 419–55

Boserup E 1965 *The Conditions of Agricultural Growth*. Aldine, Chicago

Coale A J, Hoover E M 1958 *Population Growth and Economic Development in Low-Income Countries*. Princeton University Press, Princeton, NJ

Kelley A C 1988 Economic consequences of population change in the Third World. *Journal of Economic Literature* **26**: 1685–728

Kelley A C 2001 The population debate in historical perspective: Revisionism revisited. In: Birdsall N, Kelley A C, Sinding S (eds.) *Demography Matters: Population Change, Economic Growth and Poverty in the Developing World*. Oxford University Press, Oxford, UK

Kelley A C Schmidt R M 1994 *Population and Income Change: Recent Evidence*. World Bank Discussion Papers 249. The World Bank, Washington, DC

Kelley A C, Schmidt R M 1995 Aggregate population and economic growth correlations: The role of the components of demographic change. *Demography* **32**: 543–55

Kelley A C, Schmidt R M 2001 Economic and demographic change: A synthesis of models, findings and perspectives. In: Birdsall N, Kelley A C, Sinding S (eds.) *Demography Matters: Population Change, Economic Growth and Poverty in the Developing World*. Oxford University Press, Oxford, UK

Malthus T R 1965 *First Essay on Population, 1798, with Notes by James Bonar*. Augustus M. Kelley, Bookseller, New York

National Academy of Sciences 1971 *Rapid Population Growth: Consequences and Policy Implications*. Johns Hopkins Press for the National Academy of Sciences, Baltimore, MD, 2 Vols.

National Research Council 1986 *Population Growth and Economic Development: Policy Questions*. National Academy Press, Washington, DC

Simon J L 1981 *The Ultimate Resource*. Princeton University Press, Princeton, NJ

Srinivasan T N 1988 Modeling growth and economic development. *Journal of Policy Modeling* **10**: 7–28

United Nations 1953 *The Determinants and Consequences of Population Trends*. Department of Social Affairs, Population Division, Population Studies No. 17. United Nations, New York

United Nations 1973 *The Determinants and Consequences of Population Trends*. Department of Economic and Social Affairs, Population Studies No. 50, 2 Vols. United Nations, New York

World Bank 1984 *World Development Report 1984*. Oxford University Press, New York

A. C. Kelley

Population Forecasts

Population forecasts summarize existing knowledge about the change in the size and structure of populations. They support decision making by providing broad perspectives for future development. In other cases they may provide detailed estimates for specific planning tasks. The methods that are currently in use were mostly developed in the 1920s, 1930s, and 1940s, before modern computers were available. Although the forecasts provide us with important insight into the possible future development of the populations, a major deficiency in those methods is that they provide the forecast users with inadequate assessment of the uncertainty to be expected. The reasons for the unpredictability of demographic processes are only poorly understood. However, the development of new statistical models and the availability of fast computing allows a realistic description of forecast uncertainty in terms that are understandable to nonspecialist users.

1. Development of the Cohort-component Method

1.1 Curve Fitting Approach

Early calculations concerning the renewal of populations were based on simple mathematical models.

Define $V(t)$ as the total population at time t. If the rate of growth is constant, or $V'(t)/V(t) = r$, then the population grows exponentially, $V(t) = V(0)e^{rt}$, from the starting or *jump-off* population $V(0)$. If the rate of growth decreases with increasing population according to $V'(t)/V(t) = r(M - V(t))/M$ where M is a limiting population size, then the growth is *logistic*, $V(t) = Me^{rt}/(1+e^{rt})$. A discrete time version of the exponential model was used, for example, by T. Malthus in his *Essay on the Principle of Population* in 1798. The logistic model was introduced by F. Verhulst in 1838.

Such models were widely criticized because (a) they do not give age detail; (b) they do not allow a change from increase to decrease; (c) they are too mechanistic for human populations; (d) even though the exponential and logistic models are empirically indistinguishable when $V(t)$ is small relative to M, they give completely different long-term forecasts.

1.2 Incorporation of Age Structure

Methods for describing age-structured populations were developed around the turn of the nineteenth century in many countries, first in England by E. Cannan in 1895. By the end of the 1920s, forecasts by age and sex had also been made for the Soviet Union, The Netherlands, Sweden, Italy, Germany, France, and the United States (de Gans 1999).

Define $V(x,t)$ as the size of the female population at age x at the beginning of year t. Then, for ages $x > 0$, we have the bookkeeping equation $V(x+1, t+1) = V(x, t)p(x, t) + N(x+1, t+1)$, where $p(x, t)$ is the probability of surviving from age x to age $x+1$ during year t, and $N(x+1, t+1)$ is the net number of migrants during year t who are at age $x+1$ at the beginning of year $t+1$. For age $x = 0$ we have $V(0, t+1) = \Sigma_x f(x, t)V(x, t) + N(0, t+1)$, where the summation is over the childbearing ages (often $x = 15,\ldots, 49$), and $f(x, t)$ is the expected number of children at age 0 at the beginning of year $t+1$ that were borne by women at age x one year earlier. Corresponding formulas can be written for the male population, and they can be extended to comprise several regions. This system of calculation is called the *cohort-component* (c-c) method.

1.3 Advantages of the Cohort-component Method

The c-c method can be viewed as a response to the criticisms of Sect. 1.1 in the following sense. (a) Since the c-c method treats the population by age groups, it provides direct estimates of the future population by age and sex. That this is useful for the planning of housing, schools, hospitals etc., was understood from the beginning. (b) The c-c method makes no assumption about the growth rate. It shows that, other things being equal, the growth rate increases with the proportion of women of childbearing ages. In 1925 L. Dublin and A. Lotka introduced the concept of intrinsic rate of growth. For a closed population with unchanging fertility and mortality, it gives the asymptotic rate of growth. As such it provides important insight into the renewal of populations that cannot be detected from current growth rates. Many of the early cohort-component forecasts assumed unchanging fertility and mortality, and no migration, and showed that the size of the childbearing population, and with it the total population, would cease to grow, even though the populations had been growing fast during the early 1900s. This contradicted the forecasts by exponential and logistic models. (c) The reason why the c-c forecasts are less mechanistic than those based on exponential or logistic curves is that they utilize the available information or assumptions about the change of the *vital rates* (of fertility, mortality, and migration). The forecasters of the 1920s knew that fertility and mortality had been declining for decades. How far down would they go and how fast? Examples from other countries or from privileged social groups were used as leading indicators for future mortality. The spreading of the urban lifestyle was seen to lead to a decrease in fertility. Opinion surveys were developed around 1940 to collect data on desired family size (Whelpton et al. 1947). (d) The early forecasters knew that the exponential, logistic, and other curves often gave similar fits to past data, but led to very different forecasts. Since statistical *model choice* depended on judgment but had a drastic effect on the results, why could one not use judgment in other ways as well? The solution by Whelpton to this problem was to set up targets for future vital rates, and to connect the most recently observed rates to the targets by suitable mathematical curves. Dutch forecasters had also used similar methods. This practice has been widely followed, and most national population forecasts use judgmental methods for specifying future vital rates, often via explicit targets (e.g., Andrews and Beekman 1987).

Although the c-c method was an improvement over the curve fitting methods, it was soon understood that it is really a bookkeeping framework, and without an accurate specification of the future change in the vital rates improvements in forecast accuracy would not materialize.

As an example we may consider a forecast of Sweden that was crafted by S. Wicksell in 1926. The forecast had 1920 as the jump-off year with a population of 5.90 million. It had five fertility variants. For 1940 the most likely forecast was 6.53 million inhabitants. The high forecast was 6.87 and the low forecast was 6.28. The actual population size was 6.37 so the error was 2.5 percent. For 1970 the most likely value was given as 6.67 and the high–low interval was [6.26, 8.44]. The true value was 8.08 so the error was –17.5 percent.

Even though many of the errors were large, the high–low intervals were wide and they gave a realistic indication of the uncertainty of the forecast.

2. Foundational Aspects in the Forecasting of the Vital Rates

Most national forecasts of the vital rates are purely demographic, i.e., they do not explicitly incorporate social or economic factors. Only rarely do they take into account the fact that the forecast itself might influence the future development. Partly because of these factors, and partly because of the poor accuracy of past forecasts, there has been a lengthy debate about whether the results of the calculations would be better interpreted as mere 'projections' rather than 'forecasts.'

2.1 The Difficulty with Explanatory Models

Refinement in demographic analysis often refers to the partitioning of the population at risk into more meaningful subgroups. For example, instead of relating births and deaths to the total population (which gives the crude birth and death rates), one may relate births to age of mother and consider deaths by age and sex (which gives the age-specific fertility and mortality rates) (see *Demographic Measurement: General Issues and Measures of Fertility*; *Demographic Measurement: Nuptiality, Mortality, Migration, and Growth*; *Demographic Techniques: LEXIS Diagram*). The use of crude rates in exponential or logistic models can imply growth forever, whereas the use of the age-specific rates in a c-c model may imply a future decrease of the same population.

The analysis can be further refined. Age-specific rates may be analyzed for birth cohorts and by parity; marital and nonmarital fertility may be considered separately; or one may consider fertility by race, for example. Empirical evidence of whether this improves accuracy is mixed.

However, the more fundamental social or economic factors that determine the fertility, mortality, and migration of population subgroups are rarely used explicitly in forecasting. This is because future changes in such factors can be harder to forecast than the vital processes themselves. Keyfitz (1982) reviews many of the difficulties.

One attempt at solving the problem has been *conditioning*. By making detailed assumptions about the future course of relevant social and economic factors, forecasters have sought to formulate conditional forecasts, or *scenarios*, in which the evolution of the vital processes would follow assumed relationships with their determinants. A drawback of such an approach is that the manner in which the conditioning is done can have a major effect on the results. A *sensitivity analysis* that systematically varies some of the assumptions may then be helpful.

2.2 Self-fulfilling or Self-negating Prophecies

In some countries such as Germany and France the early population forecasts were motivated by the fear of population decline and loss of political power. In The Netherlands overpopulation was the concern. In both cases the forecasts were seen as a means of influencing social policies (de Gans 1999). Inasmuch as the forecasts were influential, they were self-negating.

In other cases such as regional forecasting, a forecast of population growth may cause the local society to develop the infrastructure of an area to allow for future growth. The favorable circumstances created may then turn the forecast into a self-fulfilling prophecy.

Although examples of both types of 'prophecies' exist, it has been difficult to demonstrate the causality in a historical context. Deliberate government attempts at lowering (or increasing) fertility, or lowering mortality, have seldom had long-lasting effects that could be attributed to the policies as opposed to other concurrent factors.

2.3 Debate on Terminology

From the very start forecasters have been aware that the future vital rates cannot be known with certainty. Evaluations of past forecasts have vividly demonstrated the unpredictability of demographic change. For example, the post-World War II baby boom was missed by all national forecasters. Mortality has similarly declined faster in the industrialized countries than the forecasters anticipated.

To safeguard their reputation some forecasters have occasionally tried to communicate the uncertainty by using the terms 'projection,' 'illustrative projection,' or even 'horoscope' for the results of their calculations. This is a reasonable practice, e.g., if the calculation consists of a c-c forecast with constant vital rates, when no one expects the rates to remain constant. However, as pointed out by H. Dorn in 1950, the producers of official forecasts typically publish calculations that are based on what they consider as the most likely future vital rates, and users also interpret the results in this way. Indeed, we may define a calculation based on what is regarded as the most likely assumptions as a *forecast*.

3. Statistical Uncertainty in Forecasting

Although probabilistic terminology is used in many branches of demography (e.g., probability of death in

life tables) (see *Life Table*), and experience has shown that forecasts of the future population have been highly uncertain, the forecasting models considered so far have all been deterministic. The recognition of the uncertainties in probabilistic terms opens up new possibilities to enhance the usability of the forecasts.

3.1 Uncertainty in the Events and in the Rates

Consider a child-bearing population of size V. Suppose the number of births during a year is B. We can think of B as a sum of V random experiments with an average probability f of a birth. The expected number of births is $E = fV$. Based on a Poisson approximation one can estimate that in a sequence of B values the relative variability (coefficient of variation) is $E^{-1/2}$. In a small regional population we may have $E = 100$, for example. Then, the births vary by about 10 percent each year due to random variation alone. However, if $E = 10{,}000$, they only vary by about 1 percent, and if $E = 1{,}000{,}000$ they vary by 0.1 percent each year. It was pointed out by J. Pollard in 1973 that much larger relative variations are observed in the numbers of births in different countries, so this type of Poisson variation does not account for the variability of the observed birth processes.

The situation changes if we allow f to change from year to year with a coefficient of variation c (or by $100c$ percent). Then the relative variation is at least c irrespective of the expectation E. The conclusion is that for forecasting large populations, the uncertainty lies in the randomness of the rates rather than in the Poisson type variation in the events themselves.

3.2 Logical Flaws in the High and Low Variants

Over the years it became customary to produce high and low variants, in addition to the most likely medium variant, to give the forecast users an indication of the plausible range of the future developments. Users have typically ignored the high and low variants. One reason has been the added complexity of having more than one number in a planning context. More fundamentally, an additional reason may have been that the high and low variants have not had a logically coherent interpretation. There are two reasons for this. First, suppose $[B_L, B_H]$ and $[D_L, D_H]$ are high–low intervals for births B and deaths D, respectively. In general, there is no guarantee that the two intervals cover an equal range, because no probability content for the intervals has traditionally been given. Any calculation based on intervals with unequal ranges is difficult to interpret. Second, suppose that the intervals are given a probabilistic interpretation, so that the probability is, e.g., 80 percent that each one of them contains the true value. This was first suggested by L. Törnqvist in 1949. Suppose there is no migration. Then the increase of the population is $I = B–D$. On can show that the usual high–low interval $[B_L–D_H, B_H–D_L]$ for I has a *higher* probability than 80 percent to cover the increase. Only if the number of births and deaths have correlation –1 does the high–low interval have probability content 80 percent. But the assumption of perfect negative correlation is known empirically to be false.

In a traditional c-c forecast a very large number of vital rates are given high–low variants, and these are combined in complex ways to produce high–low intervals for population sizes at different ages. No uniform interpretation can be given to the latter intervals. For example, a United Nations forecast of the total world population published in 1998 gives a high–low interval of [7.4, 8.8] billions for the year 2030 and [7.3, 10.7] for 2050. Since it is not known what probability contents these intervals are intended to have, it is not clear how they should be interpreted.

3.3 Expressing Uncertainty Probabilistically

Modern probability theory provides a language that can be used to describe the uncertainty of the future vital rates and vital events. This leads to a *predictive distribution* of the future population given what we have observed in the past.

The availability of fast computing allows the use of stochastic simulation to describe numerically the predictive distribution. In practice this is done so that a set of age-specific vital rates is sampled from their respective predictive distributions. A c-c calculation is then made based on the sampled values to produce a possible *sample path* of the population in each age and sex group. This is repeated thousands of times and the results are stored. For any age–sex group, or for any larger population aggregate, one can represent the uncertainty of the forecast in terms of graphics (histograms, boxplots, etc.), or in terms of statistical summaries (means, medians, standard deviations, etc.). Such techniques are available for single region populations, and they are being developed for multiregional (or multistate) populations, household forecasts, and other functional forecasts.

The key element in such calculations is the formulation of the predictive distributions of the vital rates based on statistical analysis of past data. This can be based on formal time series methods, the analysis of the accuracy of the past forecasts, and judgment.

Predictive distributions provide the forecast users with a richer view of the possible future developments than the traditional population forecasts. For users needing a point forecast only, one would typically offer the median of the predictive distribution. At the other end, a complete predictive distribution is available to a sophisticated user who wants to use formal *loss functions* in *risk analyses*. Stochastic forecasts are

well suited to the analysis of the financial stability of pension funds and other social security systems, for example.

Bibliography

Ahlburg D A, Land K C 1992 Population forecasting. *International Journal of Forecasting* **8** (Special issue)

Alho J 1997 Special section on statistical analysis of errors in population forecasting and its implications on policy. *Journal of Official Statistics* **13**: 201–304

Andrews G H, Beekman J A 1987 *Actuarial Projections for the Old-Age, Survivors, and Disability Insurance Program of Social Security in the United States of America*. Actuarial Education and Research fund, Itasca, IL

de Gans H A 1999 *Population Forecasting 1895–1945: The Transition to Modernity*. Kluwer, Dordrecht, The Netherlands

Keyfitz N 1982 Can knowledge improve forecasts?*Population and Development Review* **8**: 729–51

Land K C 1986 Methods for national population forecasts: a review. *Journal of the American Statistical Association* **81**: 888–901

Lee R D, Tuljapurkar S 1994 Stochastic population forecasts for the U.S.: Beyond high, medium and low. *Journal of the American Statistical Association*. **89**: 1175–89

Lutz W, Vaupel J W, Ahlburg D A 1998 Frontiers of population forecasting. *Population and Development Review* 24 (Supplement)

National Research Council 2000 *Beyond Six Billion. Forecasting the World's Population*. National Academy Press, Washington, DC

Smith D P 1992 *Formal Demography*. Plenum, New York

Suzman R M, Manton K G, Singer B H (eds.) 1993 *Forecasting the Health of Elderly Populations*. Springer, New York

United Nations 1998 *World Population Prospects*: *The* 1998 *Revision*. United Nations, New York

van Imhoff E, Keilman N W 1991 *Lipro 2.0: An Application of A Dynamic Demographic Projection Model to Household Structure in the Netherlands*. Swets & Zeitlinger, Amsterdam

Whelpton P K, Eldridge H T, Siegel J S 1947 *Forecasts of the Population of the United States 1945–1975*. U.S. Government Printing Office, Washington, DC

J. M. Alho

Population Policies, Politics of

Population policy has a long history of engendering political controversy that promises to persist as it is fueled by divergent views of the Good Society. The deeply held beliefs that make population policy an object of intense dispute have their roots in such issues as political and economic ideologies, religion, morality, nationalism, and ethnicity. The impetus for governmental intervention into demographic processes grows out of the realization that births, deaths, and migration are not merely a question of numbers, but are critical to the social, economic, and political character of a nation. As Kingsley Davis has pointed out, 'Never, apparently, is a demographic goal sufficient by itself to justify a population policy. People judge a population trend to be good or bad only in the light of its presumed social and economic consequences' (Davis 1971, p. 6). Although Western nations have been deeply involved in the effort to curtail fertility in the Third World beginning in the 1960s, the West's experience with population policy has been mainly to encourage higher fertility in their own countries. This article will discuss the difficulties faced by governments that adopt policies to either lower or increase birthrates.

1. *The Definition of Population Policy*

The controversial nature of population policy extends to the definition of the term itself. There has been a tendency among scholars as well as government officials to abuse and misuse the definition. Population policy has often been defined in terms designed to advance a particular social or demographic agenda rather than to provide a rigorous structure for understanding and analysis. Many writers, including population professionals and students of demography, define population policy as any action of government that directly or indirectly influences population size, distribution, or composition (Eldridge 1954). Other writers go so far as to treat population policy as any government action that directly or indirectly exerts an effect on health, well being, status, or quality of life. These definitions are broad and encompass almost everything a government does. However, this is also the weakness of these definitions as almost every government policy, from the construction of roads to the provision of health services, has demographic implications.

For those interested in the formulation, content, or outcomes of population policy, a more exacting definition is required such as that found in the United Nations volume, *The Determinants and Consequences of Population* (1973). This definition, based on statements by Spengler and Duncan, states that population policy consists of 'a specific set of governmental objectives relative to the population magnitude and/or composition together with the instruments by which it may be possible to achieve them' (United Nations 1973, p. 632). Implicit in this definition is the expectation that government will establish an agency or organization to advance these goals, and will also provide resources to enable that agency to succeed. While the Spengler–Duncan definition provides a rigor that is missing in most definitions of population policy, it may at the same time not allow sufficient latitude for the give and take, negotiation, and compromise that characterize the policy process.

2. Structure Of Population Debate

The structure of the population debate was formed when Karl Marx bitterly attacked and condemned Malthus and his ideas more than a half a century after Malthus published his famous *Essay on Population* in 1798. Although Marx never formulated a 'theory of population,' he adamantly maintained that 'over population' was a consequence of the economic system, and would be remedied once the goal of socialism had been achieved. Marx recognized that if Malthus were correct to say that human poverty is anchored in human nature, then the intellectual edifice of Marxism was threatened. While the Marx–Malthus dispute provides a backdrop to the entire controversy over population policy, neither Marx nor Malthus had a profound affect on policies governments adopted until well into the twentieth century, when governments began to decide that their birth rates were either too high or too low to achieve national goals.

3. Population Policies in the Premodern Era

From earliest times, rulers and governors have attempted to increase the size and growth of their populations, which were periodically ravaged by famine, epidemic disease, and warfare. A large and growing population was universally regarded as a sign of a healthy and vital society while small or decreasing numbers signified weakness and decay. In classical Greece, the most martial city-state, Sparta, enacted laws that made marriage compulsory and imposed special taxes, civil indignities, and political disadvantage on celibates. Procreation was similarly encouraged in Rome when that city was still a city-state intent on conquest. Later, the famous population laws of Augustus, elaborated between 18BC and AD9, tried to address the low fertility of the senatorial class, on which the empire relied for the administration of its colonies (Hutchinson 1967). During this period pro-natalist policy had its origins in nationalism or imperialism, sentiments that would motivate pro-natalism through the mid-twentieth century. Before the modern era, however, the government lacked the means to reach into the hinterland, and it did not command the technologies required to enforce the law. Thus, it is doubtful that these policies were effective.

A second rationale for pro-natalist policy emerged in the Mercantilist states of Europe during the sixteenth to eighteenth centuries. To the Mercantilist rulers, a large and growing population was equated with a rich and prospering nation; therefore, the more people the better. While Mercantilist theorists disagreed over many points, most regarded labor as the most important factor of production. They also observed that high production density fostered industry and commerce and provided the basis for the establishment of colonies that would further enrich the nation. Mercantilist policies thus encouraged marriage and procreation, severely limited emigration, and encouraged immigration, especially of skilled tradesmen. Once again, effective techniques for implementation were lacking and it is unlikely that the laws had much effect (Spengler 1960).

4. Population Policies in Developed Countries

In the belief that population size and national power were closely related, political leaders in the late 1920s and 1930s saw a low birth rate and population decline as nothing less than a threat to national survival (Tietlebaum and Winter 1985). Concerned by the phenomenon of below replacement fertility, several governments adopted policies to raise their birthrates in an effort to alleviate severe economic depression and assure that the state would not decline in its relative power status in the future.

Mainly in the 1930s, several European governments formulated policy responses to declining fertility, utilizing both repressive and social welfare measures (Glass 1940). The first moves were to reactivate abortion legislation already inscribed in the criminal codes, which imposed Draconian penalties for convicted abortionists and forbade advice on birth control and the advertising and sale of contraceptives—except for condoms which were regarded as prophylactics. On the welfare side, child and family allowances, marriage and housing loans to subsidize maternity leave, benefits to reduce infant mortality, the establishment of nurseries and day care centers, and campaigns of pro-natalist propaganda were among the most significant measures introduced.

While France, Sweden, and other European nations relied mainly on welfare measures to encourage higher fertility, Nazi Germany initiated a vigorous pro-natalist policy for those thought to be of Aryan stock while simultaneously pursuing a policy of sterilizing the 'hereditarily unfit' and 'racially impure.' At the same time that Germany was urging its women to bear more children for the Fatherland, it invaded neighboring countries on the ground that it needed additional space, *Lebensraum*, for its people. Germany also utilized the power of the state and its extensive control apparatus to enforce its prohibition against abortion (Harmsen 1950). However, other than in Nazi Germany the restrictive legislation on abortion in other European nations had only a small effect, as judges and juries were loathe to convict because of the severity of the penalty and their own lack of commitment to the abortion law (Glass 1940).

Despite the anxiety over 'depopulation,' the liberal democracies of Europe could not achieve national consensus over the adoption of coherent pro-natalist policies until the late 1930s. By this time, however, birth rates were already starting to rise as preparations for war contributed to the end of the Depression and

the return of full employment. Pro-natalist measures were introduced in the Soviet Union immediately prior to World War II and spread to Eastern Europe after the war. Moreover, pro-natalism in the Soviet Union was buttressed by Marxist ideology that held that 'overpopulation' was only a problem under capitalism.

Fertility started to fall in Communist bloc states in the mid-1950s, soon after abortion was liberalized in these countries. Eastern Europe responded quickly by placing restrictions on abortion, with the most severe in Romania. Though legal, contraceptives were to all intents and purposes difficult to obtain or unavailable. Child allowances were increased and long-term maternity leave of one to three years was introduced (McIntyre 1975). A benefit was provided that varied between the minimum and average wage and re-employment was assured in the same or a similar job. These policies largely succeeded in maintaining fertility above the replacement level.

In Western Europe, by contrast, fertility fell below replacement level, and governments once again adopted pro-natalist policies. Recognizing that in the contemporary world national power no longer depended on numbers alone, policy makers gave increased emphasis to enhancing the quality of the population, that is, the health, education, and other factors that contribute to what has been called 'the effective population' (Organiski et al. 1972). Equally important as a gauge of national power was economic and technological superiority, as well as a nation's structure of alliances. For those nations intent upon encouraging couples to have at least an additional child, the most effective approach seemed to be a policy that lessened the difficulties of balancing paid work and motherhood (McIntosh 1983). Pro-natalism proved to be less than a successful undertaking for Western democracies as they were constrained by their own liberal democratic politics from implementing the multifaceted measures required for an effective policy aimed at increasing the birth rate.

5. *Population Policies in Developing Countries*

Although population had been a subject of philosophic speculation and debate since ancient times, it was not until the middle of the twentieth century that it became a cause inciting intense political controversy. There were several developments that moved population from speculation and theory to the policy arena. Of singular importance were demographic trends in less developed and colonial countries that challenged leaders of these nations, especially after they had achieved independence. Earlier, both fertility and mortality rates were high. After World War II the introduction of improved public health measures, especially clean water and improved sanitation, as well as the new availability of antibiotics, led to a decline in infant and child mortality and an unprecedented increase in the rate of population growth.

Beginning in the 1950s, leaders in developing nations came to regard rapid population growth as an obstacle to their country's economic development, a view that was reinforced by census results which showed that population size and growth exceeded earlier expectations. Four years after gaining its independence from the UK, India adopted a population policy as part of its initial five-year plan in 1951. In roughly the same time period, Pakistan and Sri Lanka instituted population policies. There were also leaders and political elites in many developing countries who, while they recognized the burden of rapid population growth, were nevertheless reluctant to adopt population control policies, partly because of the 'distrust the leaders had of the former metropolitan powers' projection of population as the greatest threat to national development.' (Made with reference to Sub-Saharan Africa in UNECA Population Policy Overview Report 2000).

Meanwhile China, the worlds most populous nation, captured the imagination of many Third World countries at the Bucharest Conference in 1974, by proudly proclaiming that socialist China could accommodate increasing numbers of people. Significantly, within five years China implemented the most vigorous population control program in modern history. The sharp changes in China's population policy were not merely a response to demographic conditions but were closely linked to the revolutionary fervor of China's leadership and their economic goals.

The One Child Per Family policy, adopted by the People's Republic of China in 1979, imposed penalties—at times severe—on those who abrogated the law. The policy succeeded and China achieved the sharpest fertility decline that any country had ever experienced. In achieving this goal, many women either sought or were compelled to undergo abortions in order to avoid a second child and avoid the possible legal consequences. A better understanding of demographic processes by Chinese officials combined with resistance to the one child policy within China and criticism from abroad, contributed to a *de facto* modification of the policy (Banister and Harbaugh 1994). Chinese officials realized that their population goals could be achieved, albeit more slowly, and with less public resentment, by implementing the one child policy less strictly.

Population policy came to be recognized internationally as emblematic of a nation's commitment to modernization. This trend was abetted and encouraged by the rich industrial nations, especially the USA, as well as by the United Nations, most particularly the United Nations Population Fund (UNFPA) and the World Bank. Industrial nations of the West felt that high fertility in poor countries would keep them poor forever. Consistent with this attitude, the USA, Canada, European nations, Japan, and

indeed all donor nations began to support efforts to limit population growth.

In 1967, partly as a result of a financial contribution from the USA, the UN began to address the issue of population growth. The Secretary General created a trust fund which eventually became the UN Fund for Population Activities (UNFPA), whose purpose was to encourage countries to identify their population problems and to attempt, with international assistance, to cope with them. Although many branches and organs of the UN would eventually be involved in curtailing rapid population growth, the creation of UNFPA was critical to the entire multilateral effort.

A point that is often neglected is that 'overpopulation' was attractive to both industrial societies as well as to developing countries as an explanation for the poverty of much of the Third World. It was much easier to attribute 'underdevelopment' to excessive population than to confront social and economic inequality, culture, religion, female subordination, or other plausible contributing factors. In the same spirit that western industrial nations found population control an appealing remedy for Third World poverty—for population control did not demand radical changes in the social and economic structure—scholars, intellectuals, and political leaders in developing countries also embraced population control as it did not require them to question fundamental attitudes and beliefs of their society. Thus, poor countries felt they could address their problems of development without first coming to grips with the social costs imposed by established structures of their societies. The discourse on the relationship between development and population has flowed from country to country and in the process has become universal rather than contained by national boundaries. The centrality of population to society assures that population issues will not be confined to any subsector of society but will become an issue that demands the attention of the polity. This has been the case in the past, and there is no reason to question its continuation.

6. *The US Influence in Population Policies*

The USA played a particularly significant role in shaping world population policy starting in the 1960s and in framing the modern debate over the politics of population. Until the 1960s, the USA was firmly opposed to playing any role, except a negative one, in the incipient population control movement. In words that he later came to publicly regret, President Dwight D. Eisenhower said, 'I cannot imagine anything more emphatically a subject that is not a proper political or governmental activity or function or responsibility' (Critchlow 1999, p. 44). Under Eisenhower's successors the American government became steadily more supportive of family planning at home and abroad.

Once the USA decided to assist poor nations to curtail their population growth, it became an American 'cause.' The USA soon became the world's largest donor to population assistance and used its political influence and financial resources to motivate the United Nations system to support family planning efforts in developing countries. Although the UK, Scandinavia, and other northern European states were important in the spread of the population movement, it was once again the political and economic influence of the USA that played a dominant role in persuading Third World countries that family planning programs were essential to their social and economic development. Equally important, the USA developed a strong professional cadre of population specialists covering every aspect of population and family planning from medicine to demography to management and communications, to assist family planning programs in almost every country in the developing world.

Perhaps of even greater importance in shaping population policy as well as in creating an international community of population experts was the investment made by the USA in training large numbers of men and women from the Third World at American universities. Upon returning home, graduates of these educational programs often filled critical positions in development programs and family planning. The USA also invested heavily in applied research designed to enhance the effectiveness of family planning programs as well as to provide an empirical base for health policy. One early research effort, the surveys of family planning and contraceptive knowledge, attitudes, and practice (KAP), demonstrated in country after country that women were having more children than they desired. The scientific contribution of KAP surveys was limited; however, they were believed to be politically valuable in persuading leaders to adopt national population policies.

As the USA exported its expertise and beliefs in voluntary family planning, so too did it export its profound political problems centering on the population issue. In the USA, the population dispute was entangled with such issues as foreign aid, state intervention in fertility decisions, sex education, contraception as a source of immorality, and, most serious of all, the abortion controversy, the most divisive issue in American society since the Civil War.

Family planners hold that abortion is a necessary backup procedure in the event of contraceptive failure. Other supporters of abortion argued that it is an essential component of any successful effort to limit population growth. A third group, led by but not limited to feminists, held that it is simply a woman's right to choose whether or not to have an abortion. Thus, while family planning and abortion are analytically distinct, there is a general tendency for family planning proponents to be pro-choice. Conversely, anti-abortion or pro-life advocates were prone to broaden their target to include opposition to family

planning. Both pro-life and pro-choice advocacy groups created international linkages whereby they have attempted to replicate the Western abortion controversy in developing countries as well as in the international arena.

7. *Population Policies and International Conferences*

United Nations conferences reflect the highly political and ideological quality of population policy at the international as well as the national level. Since the Stockholm Conference on Environment in June 1972—a conference generally regarded as highly successful—the United Nations has increasingly relied on the mechanism of international conferences to shape global policy on a variety of social and economic issues ranging from women, children, and human rights to population and development. In contrast to regular UN proceedings, these conferences have been more open to the inclusion and participation of nongovernmental organizations representing a broad spectrum of interest groups. To the extent that the policy process becomes more open and diverse groups acquire access to power, it is almost inevitable that political considerations are likely to play an important, if not a dominant, role in formulating policy at UN population conferences.

United Nations sponsorship of international population conferences actually began with gatherings in Rome in 1954 and Belgrade in 1965. The United Nations jointly sponsored these conferences and the International Union for the Scientific Study of Population (IUSSP), a highly respected international nongovernmental organization (NGO) with a comparatively small but influential worldwide membership of population professionals. The UN and IUSSP invited population experts to participate in the Rome and Belgrade conferences. These meetings, while of some intellectual significance, had only a modest impact on national population policies. Conference participants were invited as 'experts' rather than as representatives of government. Concerned about the lack of policy impact, the United Nations, as well as some of its more powerful members, changed the structure and composition of these early conferences. Their intent was to move the population debate out of the hands of scholars and experts and into the domain of government where population policy is made.

Under the leadership of the United Nations Population Fund (UNFPA), and with the strong support of the USA and other Western donor nations, the UN convened its first of a series of decennial population conferences in which governments rather than individual experts were represented. In contrast to the earlier World Population Conferences in Rome and Belgrade where participants were invited in their individual capacities as experts, delegates to the UN population conferences in Bucharest in 1974, Mexico City in 1984 and Cairo in 1994, were representatives of their government. Members of national delegations were not exclusively government officials and political leaders; governments also chose to select physicians, family planning activists, social workers, as well as feminists to represent their countries.

Contrary to expectations, these changes did not result in a stronger commitment by governments to population policy. Instead, they opened the door to a variety of interest groups to advance their own agendas and redefine population policy in terms that were more congenial to their interests. Governments also found these intergovernmental conferences a convenient host to their efforts to promote their own ideological and political agendas in collaboration with other friendly governments. Population conferences became far more controversial, interesting, and perhaps even more significant. For example, at Bucharest the Population Plan of Action, the consensus statement of the conference, urged a new and more equitable international economic order. In language that resembles a Marxist critique of development theory, the Background section of the Plan held that population problems are really symptoms of imbalances in the development process and a faulty international economic system. Twenty years later at Cairo, the new definition of population was one that gave primacy to women's reproductive health and the status of women. Nongovernmental organizations were far more prominent at Cairo than at either Bucharest or Mexico City, a move intended to gain the support of the growing universe of nongovernmental organizations. This too resulted in further politicization of the conference proceedings as NGOs at Cairo were articulate, well organized, and embraced an agenda that paid little attention to demographics.

8. *The Future of Population Policies*

Notwithstanding the near universality of fertility limitation programs in developing nations, world population reached 6.1 billion in mid-2000 and, according to UN estimates, was growing at an annual rate of 1.2 percent or 77 million people per year (United Nations Population Division 2001). According to the United Nations 'medium' projections, by 2050 world population will reach 9.3 billion. This growth is not evenly distributed throughout the world: almost all of it is taking place in developing countries. Clearly, on a global basis, the total fertility rate is above replacement level. Population growth will continue, if for no other reason than the momentum inherent in the age structure, that is, the large numbers of young people who have yet to complete their reproductive years (Bongaarts 1994). Thus, if world population is moving toward stabilization, it is moving at a much slower pace than the population community

predicted or desired. To the extent that societies and political leaders are dissatisfied with demographic conditions in their own countries, there remains little question that they will introduce population policies in an attempt to ameliorate these conditions. Whether they will set clear goals, create an organization to implement these goals, and provide sufficient resources to successfully carry out their policy remains, as in the past, a matter of political commitment.

See also: Environmental Policy; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Fertility Control: Eastern Europe; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Political and Political–Economic Perspectives; Malthus, Thomas Robert (1766–1834); Population Policy: International; Reproductive Rights in Developing Nations; United Nations: Political Aspects.

Bibliography

Banister J, Wu Harbaugh C 1994 *China's Family Planning Program: Inputs and Outcomes*. Center for International Research, Bureau of the Census, Washington, DC

Basu A M 1997 The 'politicization' of fertility to achieve non-demographic objectives. *Population Studies* **51**(1): 5–18

Bongaarts J 1994 Population policy options in the developing world. *Science* **263**: 771–6

Chamie J 1981 *Religion and Fertility: Arab Christian–Muslim Differentials*. Cambridge University Press, Cambridge, UK

Critchlow D T 1999 *Intended Consequences. Birth Control, Abortion, and the Federal Government in Modern America*. Oxford University Press, New York

Coale A J 1986 Population trends and economic development. In: Menken J (ed.) *World Population and US Policy: The Choice Ahead*. W. W. Norton, New York, pp. 96–104

Davis K 1971 The nature and purpose of population policy. In: Kingsley D, Styles F (eds.) *California's Twenty Million: Research Contributions to Population Policy*. Institute of International Studies, University of California, Berkeley, CA

Demeny P 1975 Population policy: The role of national governments. *Population and Development Review* **1**(1): 147–61

Dixon-Mueller R 1993 *Population Policy and Women's Rights: Transforming Reproductive Choice*. Praeger, Westport, CT

Eldridge H 1954 *Population Policies: A Survey of Recent Developments*. International Union for the Scientific Study of Population (IUSSP), Washington, DC

Finkle J L, McIntosh C A 1994 (eds.) The new politics of population: Conflict and consensus in family planning. *Population and Development Review* **20** (Suppl.)

Finkle J L, Crane B B 1975 The politics of Bucharest: Population, development, and the new international economic order. *Population and Development Review* **1**(1): 87–114

Glass D V 1940 *Population Policies and Movements in Europe*. Clarendon, Oxford, UK

Harmsen H 1950 Notes on abortion and birth control in Germany. *Population Studies* **3**(4): 402–5

Heidenheimer A, Heclo H, Adams C 1990 *Comparative Public Policy*, 3rd edn. St. Martin's Press, New York

Hutchinson E P 1967 *The Population Debate*. Houghton Miflin, Boston

McIntosh C A 1983 *Population Policy in Western Europe: Responses to Low Fertility in France, Sweden, and West Germany*. M E Sharpe, New York

McIntyre R J 1975 Pronatalist programs in Eastern Europe. *Soviet Studies* **27**(3): 366–80

McNeill W H 1990 *Population and Politics Since 1750*. University Press of Virginia, Charlottesville, VA

McNicoll G 1997 Governance of fertility transition: Reflections on the Asian experience. In: Jones G et al. (eds.) *The Continuing Demographic Transition*. Oxford University Press, Oxford, UK

Miro C A, Potter J E 1980 Population policy: Research priorities in the developing world. *Report of the International Review Group of Social Science Research on Population and Development*. St. Martin's Press, New York

Organiski A F K, Bueno des Mesquita B, Lamborn A 1972 The effective population in international politics. United States Commission on Population Growth and the American Future. Research Reports, Vol. 4, *Governance and Population*. United States Government Printing Office, Washington, DC, pp. 235–50

Spengler J J 1960 Mercantilist and physiocratic growth theory. In: Hozelitz B F (ed.) *Theories of Economic Growth*. The Free Press, New York, pp. 3–64, 299–344

Spengler J J, Duncan O D 1956 *Population Theory and Policy*. The Free Press, Glencoe, NJ, pp. 441, 456

Teitelbaum M S, Winter J M 1985 *The Fear of Population Decline*. Academic Press, London

United Nations, Department of Economic and Social Affairs 1973 The determinants and consequences of population trends: New summary of findings on interaction of demographic, economic, and social factors. *Population Studies* **1**: 50

United Nations, Department of Economic and Social Affairs 1999 Below replacement fertility. *Population Bulletin of the United Nations*. Special Issue Nos. **40/41**

United Nations 2000 United Nations economics commission for Africa 2000. *Population Policy Overview Report*. http://www.uneca.org

United Nations Population Division. February 28, 2001. *World Population Prospects: The* 2000 *Revision* (*Highlights*)

J. L. Finkle

Population Policy: International

There are two distinct but highly interrelated elements within international population policy. The policy set of goals and activities by states or international entities intended to influence and shape population outcomes in countries beyond their own national boundaries finds expression in both *bilateral* relationships and in *multilateral* activity. Bilateral activities seek to change population outcomes in specific countries. The collectivity of policies of the international community, expressed in multiparty international agreements, charters, covenants, and—most important for the population field—in international conferences and conference resolutions are *multilateral*. This activity is designed to standardize views and practices, and to

form a world-view of desirable goals and outcomes. The two are linked. Countries with active national or bilateral programs also seek to influence the direction of multilateral, international population policy.

1. The Unique Context of the Twentieth Century

Humans have probably always tried to manage fertility at the level of the individual and family. Rites and rituals to encourage fertility are time honored. There are also specific hieroglyphs in Egyptian antiquities citing herbal methods of contraception or abortion. More than a century before Christ, Polybius deplored both the tendency of comfortable urban Greeks to delay marriage and sharply reduce family size, and the resultant decline in population.

National concern and international preoccupation with population issues is very much a phenomenon of the last half of the twentieth century. While anxiety for the health impact on women of large families and unwanted pregnancy has been of long standing concern to health care providers, the major upsurge in interest in population issues originated in concern for the impact of rapidly accelerating global population numbers. Some appreciation of the remarkable quality of twentieth century demographic developments is therefore essential to understanding this policy evolution.

Never before the twentieth century did a human being live through a doubling of population; indeed the global population quadrupled between 1900 and 2000. The population increase in the 1990s alone was the equal of all of those who lived in the seventeenth century. In the preceding millennia, a good century for the human race was one in which the population grew at all. Plagues and normal high death rates in most societies ensured that this was not always the case.

Although everyone dies eventually, human death or mortality at an early age is what keeps population levels low. Absent high levels of mortality, migration, or decreasing fertility, population levels will increase. With the relative decline of famines and epidemics and increases in transport, communication, public health, and control of infectious diseases, nineteenth century Europe grew rapidly. About one-third of this increase was exported to colonies and overseas, a series of deliberate but minimally enunciated international population policies. The same spurt happened in post-W.W. II in the nonindustrialized countries emerging from colonialism, or 'developing world.' The increases in numbers were bigger, and the pressure valve of migration was no longer available. The world was regarded as settled by the major players who controlled immigration flows.

When the global population numbers really began to grow, when the ability to count them improved markedly, and when attitudes regarding sexuality and family issues began to change in the industrialized world, population issues arrived on the global scene and international population policy began.

Twentieth century population growth changed the shape and weight of international relationships. Over 95 percent of the population increase of the last decades took place in the developing world. At the beginning of the century, almost one person in five was European; at the end, less than one-tenth of the world lived in Europe. The six billionth child was born in the last year of the century. That child will very likely live long enough to see the peak of human population and the stabilization of population numbers. The trends that will cause the decline of high fertility are well begun. Although another huge population increase will take place in the first two decades of the twenty-first century, the end of the high fertility period is in most places now in sight. International population policy in the twenty-first century will likely focus on different issues, already emerging.

2. The Content of Population Policy: Sensitive, Fragmented, Confrontational

There has never been policy consensus on the nature, importance, or remedies in this field. The first stage of policy making revolved around the question of whether rapid population growth and high levels of fertility were in fact a threat, and what to do about them if they were. Many continued to insist that high levels of population growth were a transitory factor that would decrease with development and progress. The answer was seen to be investment in accelerated general development, not specifically targeted programs, i.e., 'development is the best contraceptive.'

Those on the other side of the issue used phrases such as 'population bomb' to describe what Malthus in the late eighteenth century foresaw as predictable crisis deriving from the geometric nature of human growth expansion, not sustainable by the foreseen linear growth of agricultural production. Concerns about the real and also postulated but never provable impacts of population growth expanded beyond food production.

Early advocates for international fertility reduction efforts voiced fears that such growth would lead to mass starvation, perpetual economic underdevelopment, internal conflict, and external wars to expand boundaries. In late 1960s there were confident assertions from certain US thinkers that the battle to feed humanity was over and that widespread famine would shortly begin. In later years, the list of concerns expanded to include the impact on the environment, climate change, and the battle to preserve biodiversity.

On the other side of the ledger, concern for the human rights implications of family planning programs and the emergence of feminist concerns in their

turn significantly reshaped the content of international debate and many population programs. Family planning and contraception services began in the industrialized world with Margaret Sanger and Marie Stopes in the USA and UK, respectively. These programs were delivered through the medical model of a doctor–patient relationship. What developed in the 1960s and 1970s was the novel idea that the state, or a public health service provider, should provide these services on a mass basis in order to bring down the number of births. Thus was born the idea of officially endorsed (or at least tolerated) family planning programs. Begun in Southeast Asia in the 1960s, family planning programs spread throughout the developing world in the subsequent 20–30 years. Internationalization occurred as foreign states and international actors intervened in such domestic programs, sometimes providing the impetus for their creation. The full impact on policy of the multilateralization efforts came somewhat later, with the creation of an international network, influenced by and influencing the existing programs.

The tools of the family planning part of population policy were largely unavailable until the 1960s. Such contraceptives as existed in the first half of the twentieth century were unwieldy, unmentionable, and often illegal. The advent of the contraceptive pill in 1960 provoked a revolution in behavior in the industrialized world. With it grew the conviction that new contraceptives, especially those which could be injected, implanted, inserted, or swallowed as opposed to being applied at the time of sexual relations, had the potential to stop or slow the rising levels of population in the developing world.

Yet another part of debate centered on the degree of importance that should be accorded to these tools and to the provision of contraception services vis-à-vis other fertility-influencing interventions such as girls education. This issue was characterized as the 'beyond family planning' debate. Later, with feminism came a strong concern that women's health issues must override concern for demographic increase in the population agenda. Included were preoccupations related to both the mechanism of program delivery, particularly targets, and several of the recently developed contraceptives. This discernibly altered and shaped population policy programs from the 1990s onward. The emergence of AIDS similarly pulled new elements related to the reproductive health of individuals into the population policy ambit.

These debates and uncertainties and the transitory coalitions that have endorsed particular actions at specific times have given changing goals and forms to international population policy.

While demographic trends can be discussed in statistical and abstract terms, and contraceptive techniques discussed in terms of their mechanisms of action, the reality of population change is grounded in the most personal human relationships, in human sexuality, and in highly sensitive social issues. These include the nature and power relationships that establish marriage, the role and status of women, the behavior of adolescents, definitions of morality and immorality, the ethical judgments related to sexual activity, and to the very value attached to life at various stages.

For this reason, the myriad of social institutions that define different societies, including organized religions and political forces, have strong interests in most population issues. The major religious have been supportive with the notable exclusion of official Roman Catholic policy and the orthodox elements of many established churches. The moral and ethical issues are complex and the introduction of state activity in these areas therefore often highly contentious. That foreign states might have an interest in the decision of a man and woman in another country to have or not have more children could not but raise very complicated issues indeed.

3. *Population Issues Enter Foreign Policy*

For most of human history, countries did not have population policies, domestic or international. There was a long tradition of hope for population increase as a sign that prosperity was edging out pestilence and disease. The 1901 Census reports for India expressed the cautious hope that population would increase in the decades ahead. Nongovernmental activity preceded both governmental and intergovernmental activities in the national and international spheres. A major step forward was the formation of the International Planned Parenthood Federation (IPPF) in 1952, which grew by the end of the twentieth century to be a federation of advocacy and service providing organizations in over 150 countries, supported by another 20. The Indian Family planning program began the same year, as did several US nongovernmental organizations (NGOs).

There was a major upsurge in population activities in the late 1960s. Governments became less timorous and more active. Several dozen heads of government agreed to come out in favor of moderating global population growth, and agreed to be featured with photos and quotations in a widely used poster campaign. The World Health Organization (WHO) began population activities. The UN Secretary General established in 1967 a Trust Fund which later became the UN Population Fund or UNFPA. The World Bank adopted a policy to lend for population activities after President Robert McNamara declared that population growth constituted a threat second only to the prospect of nuclear war. (Continuing ambivalence within the Bank about the actual impact of population growth on development has, however, meant that while loans are available in aid of Health Ministry programs, Finance and Planning ministries were never

pressed by bank officials to support or give any special priority to population programs. Many other aspects of national management have been subject to such suasion over the years.)

3.1 The International Community Response: A Wide Spectrum of Policy Perspectives, and Some Domestic Determinants

As foreign policy is the pursuit abroad of domestic policy goals, it could be postulated that even though the desired impact of bilateral population programs was to change outcomes abroad, donor countries would normally not have international population policies until they began to have domestic population policies. This was by and large the case. Most states approached the issue of their domestic population issues with a great deal of diffidence. President Eisenhower is widely quoted as saying that he could imagine 'no subject less fit for government activity' than family planning. The United States became the unquestioned global leader in this field (see Sect. 3.2), while a variety of stances characterized the roles and programs of other countries.

Sweden and the other Nordics countries were early financial supporters of family planning in developing countries, reflecting their relative lack of embarrassment in dealing with these issues domestically. In the later decades, as the influence of women's health advocates grew at home and abroad, Sweden in particular disclaimed linkage between providing contraceptives to women to support them in their own family size decisions and any demographically related outcome.

Successive French governments pursued strong pronatalist policies at home in France and perceived that strong support for family planning programs abroad could well be seen as somewhat racist. There has never been major French support. The UK bilateral programs in many areas included a vigorous population component. Italian governments were widely believed to have exchanged peace with the Vatican on the domestic contraception/abortion front in Italy, in exchange for Italian inactivity abroad. Canada was at one time the third largest supporter of population programs. Interest and support declined rapidly in the 1990s, probably reflecting aid officials' desires to distance themselves from US programs, and some political level wish to avoid exacerbating deemed Quebec sensitivities. Both are longstanding elements of Canadian domestic policy.

The Netherlands from the mid-1990s dedicated a full 4 percent of their aid programs to population and reproductive health. They often played the role of coordinator and host to conferences and meetings on a variety of population-related issues. Their domestic belief in frank discussion and the provision of open explicit information starting at an early age has translated into an active international program.

Japan, under heavy US pressure, dedicated fairly large amounts of funding to population programs, very broadly defined. Once again, the impact of domestic constraints for a long time meant that the nonavailability of hormonal contraceptives to Japanese women precluded their inclusion in the aid package.

The developing world presented a mixed picture. Thinkers in Mexico, Egypt, and India before independence worried about population gains destroying the positive impacts of their own development efforts. India's 1951 five-year plan had population program expenditures. The Southeast Asian countries were early and enthusiastic advocates with national programs established in Thailand, Taiwan, Singapore, and Korea. Over a two-decade period, most southern countries adopted some form of population policy, after a period of initial resistance on the part of many. Virtually all countries eventually formulated population policies, influenced by their own cultural and religious backgrounds, and development aspirations. Until the advent of the international conferences described below, only industrialized countries had international dimensions to their population policies. The advent of the conferences gave developing countries the opportunity to express views on population issues on the world stage.

Bilaterally, the Vatican has acted not as a state but as religious advisor to governments. It has been especially powerful in keeping official support away from these services in Latin America, and to some extent in Africa. This reluctance to endorse any form of contraceptive device continued to include proscriptions against condom use, even as the AIDS epidemic affected millions.

The Marxist countries during the Cold War period had an avowedly antifamily planning stance. The states of the former Soviet Union accorded little or no importance to the provision of contraception with subsequent high reliance on abortion. Population levels dropped sharply in many socialist countries and desperate, coercive measures were adopted by several. The pronatalist policies of Ceaucescu in Romania provided the sharpest example. The Marxian view was that difficulties caused by population growth really reflected maladjustment in wealth distribution. Although they did not proselytize to any major extent, this view was very strongly echoed by many developing countries in the early years. Particularly since it fit well with other themes of the New International Economic Order, the 'development is the best contraceptive' theme became the most commonly used rhetorical position in opposition to the new enthusiasm.

3.2 Global Leadership—the United States Role

Unquestionably, the global leader in this field has been the United States. American activity began in the 1960s and the US quickly became the most enthusiastic

proponent of population programs. A bipartisan, centrist coalition in Congress maintained support for international family planning programs, even in the Reagan years. Domestic NGOs appeared in the US for the express purpose of lobbying for greater global attention to this issue. The early name of one such group—ZPG or Zero Population Growth—indicates the intensity of public policy concern that underlay US policy aimed stabilization of global population levels.

Later with the resurgence of the religious right and the 1990s shift to political conservatism, the abortion issue re-emerged to cause violent schisms in US society. It defined presidential campaigns, filled the courts, and provoked violence against domestic abortion providers. The bipartisan consensus that had supported the family planning programs shattered. Republicans' public policy stance became the equation of family planning with abortion, and therefore opposition to all forms of public involvement in it. Democrats remained the champions of such programs.

The international policy reflection of this schism caused sharp reductions and severe restrictions on US support. Migration for economic improvement and to escape unfavorable political and social situations around the world will continue to be a growing population issue.

Specific measures of censure for China's one-child policy were included in funding conditions. Multilateral organizations were cut off from funding. A so-called gag rule was enacted. It prohibited the release of US family planning funds to international or foreign NGOs which use their own funds to provide abortion services or participate in policy debates on abortion outside the United States. This measure, which would be unconstitutional in the USA itself, was put into effect once by administrative decree in the Reagan years, and again by law in the final Clinton years.

The US flag has flown higher over population assistance in most countries than anyone else's. One of the advantages was the creation of unparalleled expertise in American government, nongovernmental, academic, and program delivery circles. The visibility was also a liability, especially in the early years when sensitivities were high about western enthusiasm for reducing southern births. Because of the intensity of US internal debate, and the rapidity of policy shifts, the US has often been at odds with the rest of the world. It was far out in front on global targets and Malthusian rhetoric in 1974, then pulled back a decade later when just about everyone else was finally coming to some policy consensus in favor of attenuating population growth.

3.3 The Tools of Population Policy: Aid Programs

The tools of bilateral population programming have included policy statements, speeches, persuasion, advertising, and funding. Assistance has included policy-influencing computer simulations, assistance to demographic and census departments, soap opera presentations, and research on a wide variety of reproductive health elements. Much program activity focused very directly on the provision of family planning services. Funds have been provided for contraceptive development, dissemination, and for the creation of national family programs.

Family planning programs had no real counterpart in developed countries, and were often 'vertical' or stand-alone programs, like some immunization campaigns. They occasionally became the focus of religious, traditional, and xenophobic criticism. These clinics often provided the only services available or available to poor women. In terms of impact on fertility, family planning programs were deemed to have succeeded; they probably accounted for about 40 percent of the decline in otherwise anticipated births that characterized the world as the demographic transition progressed in the countries of Asia and Latin America.

Because of the sensitivity of population and reproductive health issues, many donor countries directed a substantial percentage of their assistance through multilateral and international delivery mechanisms. Almost US$500 million flowed through multilateral agencies two years after the 1994 Cairo Conference. The programs of the primary multilateral agency, the UN Population Fund, reached $320 million in its high water years, usually providing a channel for about 25 percent of available donor funding. World Bank lending at one point reached $500 but declined at the end of the century. The regional banks have not been major players, with some exception for the Asian Development Bank. An increasingly important element of support was the assistance provided by (largely US) foundations to the population field, reaching as high as $150 million just after the 1994 Cairo Conference.

Aid or official development assistance to population/family planning was never large in relation to overall expenditures in the field or to overall levels of overseas development assistance. By the end of the twentieth century, developing countries were paying three-quarters of the costs of their own reproductive health and population programs.

Nor have population programs dominated the overall aid programs. Even within the US program, population assistance only ever represented about 7 percent of all US aid. Australia, Denmark, Finland, The Netherlands, Norway, and UK the population comprised about 3 percent and in France and Italy only 1 percent of their respective ODA programs. In dollar terms, when the totality of all countries' official development assistance was running around US $60 billion per year, population assistance (mostly family planning) never got above $2.0 billion in total. Eight countries almost always supplied 90 percent of all

population assistance. The US gave the biggest amount, usually about half of all bilateral aid to population. Denmark, Norway, Sweden, and The Netherlands gave higher percentages, relative to their own economic weight.

If not significant as a percentage of overall aid, or as a proportion of each countries' aid program, foreign aid to population has often been *very* significant in relation to the total health budget of many developing countries. It has had a catalytic impact in determining the scope, content, impact, and in some places existence of programs across the world.

4. *Multilateral Policy Making*

The second zone of international population policy activity is multilateral or many sided in nature. It is important for two reasons. As just noted, multilateral organizations have been important sources of technical and capital assistance, and vital to program delivery. The second key multilateral activity relates to international population policy making within the international community. Here, all countries have roles; the multilateral organizations provide the venue and organization for this policy-making activity.

4.1 The Population Conferences—Policy Tracers Par Excellence

It is through the phenomenon of the international population conferences that the changes and shifts of international policy in this area can be traced. Margaret Sanger organized an international population conference in Geneva in 1920. It resulted in the formation of the first international nongovernmental population organization, the International Union for the Scientific Study of Population (IUSSP), mostly but not exclusively demographic in purview.

Modern population conferences are not like those of the World International Trade Organization where treaties are negotiated and monitored. Unlike conferences such as the 1948 Universal Declaration of Human Rights, they do not require countries to take legal steps to bring domestic activity into line with new policy norms. They are rather international conclaves of policy expression, where undertakings of good faith are made by all, and statements are made and resolutions drafted which together give a comprehensive picture of how the world's thinking has evolved on an issue. They also give the world the opportunity to comment on individual programs: the one-child policy of China and the specific programs of Indonesia and India have from time to time drawn attention.

The preparatory work preceding such conferences is key, often including the collation of detailed country statements which can include reviews and revisions of laws, budgets, policies, and program descriptions. Most important, the preconference activity occasions consultation and negotiation on the domestic front in most countries. As the issue has internationalized, comparable activity takes place on a regional basis, and globally among NGOs. As NGOs and women's groups have become more important players, both the openness of the preparatory process and the difficulty in reaching consensus have grown. At the conferences themselves, NGO representatives have come to take part in the delegations and to play a role in formulating the consensus documents. Generally, countries reflect at these conferences the positions they have adopted domestically. The Vatican chooses on this issue to behave as a State in terms of attendance at international conferences. It works through lobbying, diplomatic demarches, and public diplomacy.

Population conferences as international policy making occasions therefore provide highly visible venues for intense debate about sensitive reproductive health issues of safe abortion, genital cutting, violence against women, and services for adolescents. These continue to provoke confrontation and contention.

4.1.1 Bucharest—1974. By the end of the 1960s, there was considerable enthusiasm—in the industrialized world—for serious attention to be given to population levels in the developing world. Serious attention in this era really meant a focus on family planning. In Bucharest, the developing countries both contested the view that population growth was deleterious, and asked to be paid more for going along with the industrialized country view. The socialist states and developing countries successfully insisted on bringing development, broadly and loosely defined, into the equation. The industrialized world did not disagree—John D. Rockefeller stated that '… the only viable course is to place population policy solidly within the context of general economic and social development in such a manner that it will be accepted at the highest levels of government and adequately supported.'

Post-Bucharest, program content did not change appreciably, concentrating primarily on family planning with some maternal and child health elements.

4.1.2 Mexico—1984. Many countries in the developing world, notwithstanding enunciated public positions, implemented population-limiting policies and programs in the largest growth period in family planning program activity, from 1974 to 1994. The last region to adopt population policies was Africa, following the landmark African regional conference in 1984, part of the lead-up to the Mexico City Conference. The 1984 world, except for the Vatican and allies, was reasonably united behind these programs, and was demanding more support for them. The US administration was by that time in full retreat from

former enthusiasm. 'Population growth is neither positive nor negative; it is a neutral phenomenon' became the leitmotif of the American administration position. Despite this policy level coolness, US budgetary and program support for continued growth in these programs were maintained.

4.1.3 Cairo—1994: a near universal consensus on a 20-year program of action. The preparatory sessions for Cairo were extremely contentious with severe criticisms of the existing approach to population programs coming from both conservative forces and from the new strong feminist voices. The pre-conference work was arduous, and the result was a complex, far reaching program of action that scarcely touched demographic forecasts and forces, but acknowledged at great length the complex context within which decisions about childbearing are made. The conference and action plan focused on health services, education, economic development, and livelihoods for women.

There was general acceptance that family planning must be voluntary and available, something never before achieved in conference documentation. The programs were to be centered on the needs of the individual and provided in a context which would take the client's entire reproductive health into account, particularly exposure to sexually transmitted diseases and AIDS. Targeted development interventions such as female education, credit for women, gender equity in inheritance and land ownership, infant and maternal mortality reduction, became explicit components of population policy.

Ninety percent of the 180 governments representing about 95 percent of the world's population approved the action plan in totality; a few conservative countries and the Vatican approved it in part. A price tag of $17 billion was established for achieving the program.

There is fairly general agreement that the Cairo Conference made an appreciable difference to the field. Five years after Cairo, progress under the Action Plan was reviewed. Although the same issues provoked the same dissent from many of the same sources, there was consensus that considerable progress has been made in policy, program redesign, increased partnership, and collaboration directed toward implementation of the action plan. International aid fell away from rather than toward the $7.5 billion target collectively undertaken in Cairo, but many developing countries were found to have made policy legislative and/or institutional changes in the area of population and development and reproductive health and rights. Improved transparency in governance, expanded activity of voluntary associations, improvements in communications, and legal and policy modifications were found to have advanced the prospects for the participatory approach seen to be at the center of successful implementation of reproductive health and rights programs. Much of course remained to be done: enormous problems of maternal mortality, violence against women, and unmet need for contraceptive services all persisted, and at the end of century, two-thirds of the world's illiterates were girls and women.

5. *The Future of International Population Policy*

Is international attention to population growth finished? One of the results of the Cairo Conference was a sharp refocusing of debate away from demographic concerns toward reproductive health issues. Analysis has continued but with a sharp diminution on discussion of several former key elements. In terms of program content, many now realized to be most important in addressing high levels of fertility, especially girls education, are being addressed for their own importance. As the world continues to urbanize, and as the situation of women improves in many areas, fertility levels will continue to decline, and with them, demographic growth. It is indeed possible that countries and groups will begin to discuss policy measures that have succeeded in *raising* family fertility intentions.

Judging from declining flows of development assistance, and of international attention, there will probably be relatively less attention paid to assisting with service provision. Some continuing attention will be given to the high levels of maternal mortality. Violence ainst women will continue to be addressed both nationally and internationally. The same levels of international assistance and interest do not, however, seem to be available in the service of reproductive rights, as could be mobilized to slow population growth, before it became politically incorrect to speak of doing so.

This is the case despite very high levels of measurable unmet need for contraceptive services and even though an additional two billion will be added to the global population before stability is reached, with most of that two billion added in areas where food and water are critical now. These pressing global issues will be addressed primarily by applying resources to freshwater availability, food production and availability, etc., and rather less to addressing the numbers of people in need of them.

There are three main issues that will likely form the core of international population policy in the future: reproductive health issues touching infectious disease, and the linked issues of migration and aging.

Reproductive health issues will occupy an expanding place in the global agenda, particularly those related to the spread of AIDS and other sexually transmitted diseases. The realization of the security threat posed by these diseases began to accelerate with the turn of the century.

Migration for economic improvement and to escape unfavorable political and social situations around the world will continue to be a growing population issue.

A different set of international institutions and government departments from those concerned with fertility are the central actors on this issue.

While the dramatic nature of fertility increase in the twentieth century is widely known, there is considerably less appreciation of the fact that the equally dramatic decline in fertility has radically changed the number of people within age groups in many countries. Aging may well be the dominant policy issue of the twenty-first century. The spread of aging populations is uneven. While there are more Europeans moving into the over-65 year old category than the under 16 year old group, at the turn of the century there was still a 26:1 ratio in the other direction in the totality of the developing world. In developing countries that initiated rapid declines of fertility in the 1960s, such as Taiwan, Korea, China, Thailand, Mexico, and Brazil, issues connected with aging are now rising on the agenda. This issue will internationalize primarily because of the impact on migration created by declining labor force size, and the worries that this creates vis-à-vis social security pressures and health system financing.

See also: Corporate Governance; Fertility Change: Quantum and Tempo; Fertility Control: Overview; Food Security; Greening of Technology and Eco-technology; Population Ecology; Population Pressure, Resources, and the Environment: Industrialized World

Bibliography

Bongaarts J 1994 Population policy options in the developing world. *Science* **263**: 771–76

Cohen J E 1995 *How Many People can the Earth Support?* W W Norton, New York

Brown L R 1995 *Who Will Feed China: Wake-up Call for a Small Planet*. W W Norton, New York

Erlich P R 1968 *The Population Bomb*. Ballantine, New York

Gulhati K, Bates L M 1994 Developing countries and the international population debate: Politics and pragmatism. In: Cassens (ed.) *Population and Development: Old Debates, New Conclusions*. Overseas Development Council, Washington, DC, and Transaction Publishers, New Brunswick, NJ

Jain A (ed.) 1998 *Do Population Policies Matter? (Fertility and Politics in Egypt, India, Kenya and Mexico)*. Population Council, New York

JeJeebhoy S J 1995 *Women's Education, Autonomy, and Reproductive Behaviour*. Clarendon Press, Oxford, UK

Malthus T R 1798 *An Essay on the Principle of Population*. J. Johnson, London

Myers N 1995 *Environmental Exodus: An Emergent Crisis in the Global Arena*. The Climate Institute, Washington, DC

Sen G, Germain A, Chen L C (eds.) 1994 *Population Policies Reconsidered: Health, Empowerment, and Rights*. Harvard University Press, Boston

Simon J L 1977 *The Economics of Population Growth*. Princeton University Press, Princeton, NJ

Sinding S W, Ross J A, Rosenfield A 1994 Seeking common ground: Unmet need and demographic goals. *International Family Planning Perspectives* **20**: 23–7, 32. Reprinted in Mazur LA (ed.) *1994 Beyond the Numbers: A Reader on Population, Consumption and the Environment. Island Press, Washington, DC*

United Nations Flows of Financial Resources in International Assistance for Population

M. Catley-Carson

Population Pressure, Resources, and the Environment: Industrialized World

1. Introduction

Few issues have attracted as much ink, and as much intellectual passion, as have debates about population, resources, and the environment. It is important to understand that many discussions of 'environment' have foundered due to at least four different uses of the term:

(a) As in 'the global environment' or the 'global ecosystem.' Explanations at this high level of abstraction often are so general as to defy contradiction or verification.

(b) As a shorthand reference for all natural resources, especially all sources of raw materials, energy, and food. Less general than the first use, but still combines very different settings and industries under one rubric.

(c) To connote the purity of air and water, especially as affected by urbanization and industrialization and their impacts on rural or 'natural' landscapes.

(d) As an antonym of 'heredity.' This is a quite different and often confusing meaning in which 'environmental' effects are those determined by social rather than biological processes, along with interaction between the two.

The disputes, which go back at least two centuries, touch upon theories and evidence central to much of modern intellectual history. Moreover, the participants in the debates have included many of the giants of Western scholarly and political thought, from Smith, Ricardo, and Malthus, to Marx and Engels, to Keynes and the Myrdals.

2. Two Fundamental Phases of Western Political Thought, 1750–1940

We can identify two fundamental phases of the developing debate on population, resources, and environment in the West. The first extends from 1750–

1880, the central period of industrialization in Western Europe and North America. The second covers the period 1880–1940, a period when industrial growth rates in Western Europe first peaked and then plateaued, Western imperial power peaked and then declined, and Western political instability and collective violence reached unprecedented levels. These two phases of ascending European political and economic power, followed by its slow decline, are reflected in the attendant debates about the balance among demographic growth, resources, and the environment.

During the expansionary phase, optimists from Godwin to Condorcet to Marx foresaw essentially unlimited expansion of material resources to support a growing population. What had to be changed was not the resource base, but rather the power of landowner and/or capitalist classes to place their demands above the needs of the majority. While Marx and his followers demanded revolutionary transformation of urban, industrial society, utopians sought to escape from its corruptions and create a new kind of society remote both physically and morally from the old. With the 20/20 vision of hindsight, we now can say that these utopian and Marxian views have not proven to be sustainable in the real world.

Meanwhile, pessimists of the period such as Malthus and his followers saw utopians as innocents unaware of the impossibility of human perfectibility. Malthus held that the human sexual instinct was a constant that would produce population growth ultimately leading to disaster. Idealistic efforts to reduce poverty (via 'transfer payments' in modern terminology) would prove self-defeating, he argued, because higher incomes would in turn lead to higher fertility and lower infant and child mortality. Hence, policies intended to relieve poverty would worsen the environment by putting additional pressure on the food supply through encouraging demographic increase. For Malthus, the only morally acceptable escape was by individuals exercising 'prudential restraint' of their fertility, by which he meant avoidance of marriage unless and until the couple had sufficient resources to support a family, accompanied by abstinence outside of marriage. Failing this, balance between population and resources would be imposed via the dreadful 'positive checks' of famine, plague, and war.

In similar hindsight we can say that Malthus's forecasts were mistaken too. He underestimated how fast food production and economic activity would increase, and dismissed as infeasible the conscious control of fertility within marriage that later came to be the norm in urban/industrial societies. His provocative arguments did, however, stimulate a passionate debate on population and the environment. In particular, his views were assaulted vigorously by Marx, thereby establishing a destructive 'right–left' dimension to more than a century of population/resource/environment debates that ensued.

Other conservative writers of the nineteenth century focused less on population issues and more on the 'destruction' of the 'natural environment' produced by capitalist industrial development and urbanization. In response to such claims, Marx and his followers acknowledged the validity of their criticisms, but argued that these developments were a function of capitalism, and hence transitional.

The second phase of the population/environment debate began toward the end of the nineteenth century, when the use of 'environmental' as an antonym to 'genetic' emerged in discussions of human decadence and perfectibility. Such arguments derived from enthusiasm about Darwinian arguments, though many of their 'eugenicist' proponents were far from well informed of the scientific evidence involved.

In general, they argued that the human gene pool was deteriorating, with powerful detrimental impacts on society and its betterment. The urban environment was itself genetically damaging, producing a cycle of residents who bore large numbers of stunted offspring who in turn raised large families of equally unhealthy children. Meanwhile, social elites were controlling their own family size, resulting in a collective 'enfeeblement.' For these proponents, this posed profound dangers to state and empire by reducing military and international power, while the increasing fertility differentials between social classes foretold the prosperous being swamped by a flood of urban disability. What was essential, they argued, was intervention to reverse the rot.

While many eugenicists were politically conservative, others were liberals or socialists. Though relatively few were fascists, the eugenics movement ultimately was compromised fatally when its notions of biological decadence were co-opted by the Nazis.

3. *Complex Phenomena, Simpler Analyses*

Any intersection among three forces as multifaceted as population, resources, and the environment—each of which is itself a subject of convoluted debate—is bound to be one of challenging complexity, required balanced and nuanced analyses. Yet, unhappily, there is an enduring tendency for arguments that anoint one of the three forces as dominant. Proponents on all sides tend toward single-cause explanations of complex social processes. Some emphasize the biological, others the economic, still others the ecological. Each group embraces one unifying theme to the exclusion of insights from other perspectives, and produce pronounced polemics of denunciation and retort.

Malthus's celebrated essay on the principle of population was itself such a response to earlier arguments from utopians (including his father). It in turn provoked passionate denunciations by utopian socialists and Marxists, some of whom argued that there were no natural limits on population growth.

Similarly, eugenicists of the early twentieth century exaggerated the importance of biological and genetic elements for social issues such as fertility, poverty, and illiteracy, and such arguments in turn led social scientists to exclude biological elements entirely from their discussion of the 'demographic transition.'

In short, there has been a mentality of the adversary underlying much of the discussion about population, resources, and the environment, redolent of the style of political, religious, and legal advocates who ignore the similarities of their approaches to those they condemn.

Another source of dissonance during the nineteenth and twentieth centuries was the tendency to search for or assert 'scientific laws' of population, resources, and/or environment. Marx and Engels claimed the mantle of science for their analyses of capitalism while attacking utopians seeing the world not as it was but as they wished it to be. Malthus too criticized utopians such as Godwin and Condorcet in the same way. The fact that Marx's arguments did not explain social conditions within his own lifetime, and that Malthus's distinction between geometric and arithmetic progressions of population and food was contradicted by trends underway as he wrote, failed to dissuade either from his claims of scientific validity.

Similarly, the near-appropriation of Darwinian theory by political proponents of social Darwinism, and later of then-modern genetics by those promoting eugenics as social policies, illustrates the deep mystique of science in political arguments of this period. Invoking 'scientific' evidence of biological limits to human progress, eugenicists derided moderate and revolutionary reformers' assurances that environmental or political change could ameliorate the lot of the poor. Such excessive arguments from eugenics then led social scientists from the 1930s onward to formulate a theory of the demographic transition that excluded biology.

Meanwhile, in the Soviet Union Lysenko and his associates developed a Soviet 'scientific' response to eugenics that asserted that environmental factors trumped genetic forces, and more specifically that hereditary characteristics could be acquired rapidly from the environment and subsequently transmitted reliably to later generations. This 'scientific' perspective offered Soviet officials the opportunity to use 'environmental' (i.e., social, economic, political) policies to manipulate the gene pool—an opportunity that produced devastating famines as a consequence.

4. *Economists and Demographers*

Economists and demographers of the twentieth century also tended to unduly limit the horizons of their analyses and data. Neo-classical economics flowered at a time when economic growth was so rapid that it was easy to ignore resource and environmental issues. Natural resources must have appeared unlimited to those at the center of expanding global empires, while the bulk transport allowed by steam power, railroads, and refrigeration rendered moot early concerns about limited agricultural land.

The emerging field of demography had multidisciplinary roots in economics, biology, mathematics, statistics, and epidemiology. Yet leaders of the nascent field were offended by the excesses of some biologists and others who sought to explain Western fertility declines in exclusively biological terms, and by the conservative political views that increasingly were associated with such explanations. The emergence of fascism exacerbated their concerns, leading them to relocate their 1936 scientific meeting from Rome to London and to boycott another meeting in Berlin.

At this moment in the intellectual history of demography a social and economic explanation of Western fertility coalesced into the 'theory of the demographic transition.' Though not really a theory in the sense of offering testable predictive hypotheses, the demographic transition provided a grand historical generalization that fertility declines could be wholly attributed to economic and social forces of the industrial revolution. In so doing, it ignored biological explanations then in vogue of the same phenomenon, as well as elements of culture, language, and religion unless these were driven by socioeconomic change (e.g., the effect of urban and industrial development in weakening the hold of the extended family on fertility behavior.)

The thinking of both economists and demographers was affected deeply by the Great Depression of the 1930s. To some degree they interpreted it as a consequence of 'underutilization,' both of capital and human resources in the economy and of the ability to bear children. Gunnar and Alva Myrdal in Sweden and Enid Charles in England produced gloomy, sometimes apocalyptic writings anticipating the terrible consequences of continuing low fertility. The Myrdals' prescription for dealing with 'the population crisis' was social legislation that evolved into the Swedish welfare state.

World War II focused concerns in different directions, but the war was followed by a long period of buoyant economic growth accompanied by much higher fertility levels. Expectations of unlimited economic growth replaced the gloom of the 1930s. Fertility trends ranged from powerful and sustained baby booms in the United States, Canada, Australia, and New Zealand to smaller and shorter-term 'boomlets' in most of Western Europe.

By the early 1960s, concerns about the negative environmental consequences of economic growth began, accompanied by worries about accelerating demographic increase both due to the Western baby booms and to rapid mortality declines with sustained high fertility rates in formerly colonial developing countries. Both concerns contributed to growing

awareness and activism about environmental problems, enhanced further during the 1970s by the oil crisis, the accident at the Three Mile Island nuclear plant, and the appearance of economic 'stagflation' that had no explanation in the canon of Keynesian economics. These events suggested there might, after all, be limits to natural resources, technological advance, and economic prosperity. In the United States and Western Europe, a new kind of political environmentalism emerged, including a direct electoral form in the German 'Green' party.

During the 1980s, and partially in reaction to environmental politics, there emerged in the United States an odd amalgam of earlier utopian, Marxian, and *laissez-faire* thought, deploying apparently scientific data and economic analysis, in what came to be called libertarianism or 'cornucopianism.' In its most enthusiastic variants, cornucopians argue that there are no natural limits to economic and population growth, and that the availability of natural resources is rising rather than declining. Some asserted that the 'ultimate resource' is human intellectual power to innovate technologically and create new resources, and hence the more humans on earth, the more resources there would be, *ad infinitum*.

Many of those who embraced such arguments did not seem to appreciate their intellectual origins, especially their Marxian pedigree. They were promoted strongly by opinion organs of American new right groups (the Heritage Foundation, Cato Institute, and the editorial board of the *Wall Street Journal*).

5. *Environmental and Ecological Approaches*

While many biologists, economists, demographers, and politicians have tended toward simple arguments about population, resource, and environmental issues, environmental scientists have emphasized the place of the human species in a complex ecological web of biological systems in constant interaction with the physical environment. In so doing, they have pointed appropriately to its vulnerability to the external insults of human technologies. However, there has also been a tendency to minimize or ignore the potentially positive environmental impacts that can also be produced by human technologies.

Consider for example the 'epidemiological transition' (see *Mortality, Epidemiological, and Health Transitions*). The triumph of human knowledge over food shortages and many endemic infectious diseases has produced dramatic improvement in human mortality since the eighteenth century. This was, of course, the primary reason for the acceleration of rapid human demographic growth since 1750, but it may also be seen as a fundamental technological transformation of the ecological systems involving humans, their food supply, and their 'predators.' Though environmental writers generally do not describe these particular technological achievements in negative terms, some prominent writers have expressed skeptical views of their long-term implications for the global ecosystem.

In general, environmental and ecological writers frequently have emphasized the negative effects of other elements of science and technology, such as those surrounding energy consumption and production and modern agriculture, and paid rather less attention to the potential of new technologies for expanding natural resources or ameliorating environmental damage from other human activities. Of course sophisticated ecologists take such potentials fully into account, but there remains among some elements of the environmental movement a tendency to de-emphasize the potential environmental positives of technological innovation—in some sense a polar extreme from the technological enthusiasms of some economists. To many ecologists, the technological optimism of much modern economics, and of both Marxists and the new right, ignores the laws and complexities of nature.

To some degree we have come full circle. When environmentalists denounce the rape of the rain forest or the despoliation of freshwater streams and lakes, they echo older and eloquent laments such as those of D. H. Lawrence at the ruination of the English countryside or Dickens's portrayal of the industrial wasteland in *Hard Times*. Concerns about the environment, like those about population and resources, have deep intellectual roots in Western intellectual traditions.

6. *Conclusion*

Given this rich but often contentious intellectual history, what might be expected of future discussions of population, resources, and environment? During the nineteenth century, most of the major political thinkers were forced by the powerful changes all around them to analyze population change and its interaction with resources and the environment. Typically they sought single (or at least simple) explanations of the complex transformations they could see, leading inevitably to an intensely ideological debate. Instead of building upon the work of earlier writers, they sought to refute their adversaries' errors. First the utopians, Malthus, Marx, and their allies clashed over the potential for demographic and economic growth with prosperity. Later biologists, eugenicists, and social scientists contested about the biological vs. environmental causes of declining fertility rates and increasingly 'enfeebled' urban populations.

These classical debates about population/resources/environment are a polemicists' dream: sweeping generalizations, exaggerated rhetoric, and emotional accusations. Those in the Marxian, socialist, and liberal-Christian traditions were certain

that when the majority took power from the reactionary elite, they would build a more humane environment unconstrained by demographic issues. Those following Malthus, and later the eugenicists, argued to the contrary that human progress is constrained by the laws of biology and demography.

The political and economic crises following World War I tended to mute these earlier debates about population growth, resource scarcity, and environmental degradation. After World War II, it was relief at the end of global war and the reappearance of prosperity that further eclipsed environmental and demographic questions. Only since the 1960s have these issues re-emerged, after decades of inattention.

There is good reason to fear that future debates may echo those of the past: ideological blinders, harsh rhetoric, narrow analytic perspectives. But there are also signs of hope. Resource economists have shown that it is possible to combine the power of economic theory and analysis with those of ecology and environmental perspectives. Demographers too have broadened their dependency upon purely socioeconomic analyses to embrace insights from biology, ecology, political science, and other relevant disciplines. The hope is that these more ecumenical trends will persist and flower, so that our collective understanding of population, resources, and the environment may someday be as subtle and complex as the phenomena themselves.

See also: Development: Sustainable Agriculture; Development Theory in Geography; Ecological Economics; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Economics; Environmental Vulnerability; Environmentalism: Preservation and Conservation; Population Policies, Politics of; Resource Geography; Resource Institutions; Sustainable Development

Bibliography

Daly H E 1996 *Beyond Growth: The Economics of Sustainable Development*. Beacon Press, Boston

Homer-Dixon T F 1999 *Environment, Scarcity, and Violence*. Princeton University Press, Princeton, NJ

National Research Council, Board on Sustainable Development 2000 *Our Common Journey: A Transition Toward Sustainability*. National Academy Press, Washington, DC

Ness G D, Drake W D, Brechin S R (eds.) 1993 *Population–Environment Dynamics: Ideas and Observations*. University of Michigan Press, Ann Arbor, MI

Polunin N (ed.) 1998 *Population and Global Security*. Cambridge University Press, Cambridge, UK

Ramphal S S, Sinding S W (eds.) 1996 *Population Growth and Environmental Issues*. Praeger, Westport, CN

M. S. Teitelbaum

Population Size Estimation and Capture–Recapture Methods

How many people were missed by the US census? How many injecting drug users are there in Glasgow? Homeless in Scotland? Industrial accidents in Maryland? Bugs in a computer program? How many words did Shakespeare know?

How do we estimate the sizes of such populations? If the population includes a subgroup of identifiable individuals and if their number is known, a simple approach is available. If a random sample can be taken from the whole population and the proportion of the subgroup that occur in the sample is calculated, that proportion can be used to scale up the sample size to provide an estimate of the unknown size of the population. Three important 'ifs' occur in the last two sentences that cause problems for this dual-system estimation (DSE) procedure, also known as capture- (or mark-) recapture (CR) from its widespread use in the study of wildlife populations by means of animals caught, given marks, released and subsequently recaptured.

Both in this simplest form, and in the many extensions proposed to overcome the problems posed by these 'ifs' in the real world, this CR method has many relationships with other statistical methods. The problem is one of sampling from a finite population. If an indicator variable y is defined to be 1 for an individual in that population, 0 otherwise, the unknown size of that unique population is the population total Y of y. The CR estimate is the classical ratio estimate based on the auxiliary indicator variable x which is 1 for an individual in the subgroup, 0 otherwise. Laplace's (1786) estimation of the population of France, with subgroup those registered as born in the previous three years and a cluster sample of over two million, can be regarded as the first use of CR as well as the first use of a superpopulation model (Pearson 1928, Cochran 1978). Two books discuss CR as sampling theory—by Thompson (1992) and Skalski and Robson (1992).

1. Statistical Theory

Early discussion, summarized in e.g. Seber (1982) or Cormack (1968), focused on distributions for the observations—hypergeometric, multinomial etc.—under different rules for defining the subgroup and stopping the sampling. Practical differences are minor. The distribution of the estimate, especially for small subgroups or samples, is more important. If the population size is N, the size of the known subgroup X, and the sample size n contains x from the subgroup, then the intuitive, also the maximum likelihood,

estimate of N is $[nX/x]$, known as the Petersen estimate after the Danish fisheries biologist who first proposed its use in biology. This is the observed number of different individuals $[X+n-x]$ plus the estimate $[(X-x)(n-x)/x]$ of the number U of unobserved individuals. As a ratio estimate this is biased: the recommended correction (Chapman 1951) adds 1 to the denominator of the estimate of U. The distribution of the estimate of U is skew, and if U is small, or based on small observed numbers, then a confidence interval based on assumed normality can cause embarrassment by including negative values for U. Likelihood can be used directly (Cormack 1992) to avoid this. Other approaches have been proposed, with different loss functions and/or Bayesian methods. The latter are discussed below.

The known subgroup often comprises individuals identifiable because they occur in another list or sample. In wildlife applications, in which the next developments took place, the two samples were often obtained by the same protocol at different occasions, with a long enough period between for thorough mixing of the mobile animals to be (questionably) assumed. Generalizations of this two-sample case were made, to more samples (Darroch 1958) and, as time periods extended, to allow for the population to change through birth, death, or migration between samples. In open populations, survival estimation from recapture data on marked animals is more reliable than estimation of population size. Open population models were applied to historical lists by James and Price (1976): the willingness of residents of a medieval town to contribute to the town's defense, and their ability to avoid paying a tax to the king, are clearly displayed, together with evidence of unsuspected epidemics.

The natural statistical way of representing counts in a number of categories, formed by the presence in or absence from a number s of lists, is as the cells in a 2^s contingency table. Models and analytical techniques are summarized in books such as Bishop et al. (1975). In CR one cell is not observed, a structural zero, and the aim is, by formulating a series of models for the observed categories in the table, to estimate the missing number by extension of these models. When additional information on individuals is available in the form of quantitative covariates, such as age or severity of disease, the use of logistic models at the individual level was suggested independently by Alho (1990) and by Huggins (1989). The Horvitz–Thompson estimator is used, with inclusion probabilities, of being in each list, estimated from the model, a procedure developed differently by Pollock and Otto (1983). With no covariate information, or with categorical information only—e.g., gender, ethnicity, geographical region—log linear models form the current standard approach. Maximum likelihood estimates (MLE), perhaps bias corrected, are then derived. In this framework it is natural to examine the data for interactions between categories, evidence that the samples were not random. Birth and death, unequal availability of different individuals, and behavioral change consequent on physical capture of animals, have all been expressed as interactions within the log linear framework.

Even thus far, numerous quirks arise whereby the corresponding statistical theory does not quite apply to the CR problem. Estimation from a log linear model does not maximize the complete likelihood: nuisance parameters of sampling probabilities are estimated from the observed cells, and N estimated conditionally on them. Normally such multinomial models for complete tables are fitted assuming a Poisson model, giving the estimates various desirable properties of the exponential family. With N unknown, the distributions of the estimate of N are different under multinomial and Poisson models though the point estimate is the same and estimates of nuisance parameters have the same properties. Regularity conditions for desirable properties of MLE do not hold in the multinomial model, but do in the Poisson. The Horvitz–Thompson correspondence shows that the estimate of N is calculated via estimates of $1/$(inclusion probabilities). Why then seek unbiased estimates of these probabilities? It is their inverses that are of interest.

2. *Health and Social Studies*

One widespread application in the social sciences is to ascertain how many individuals have some health or social problem, numbers needed for policy and budgetary decisions. Data are pre-existing lists from different agencies, each attempting to cover the whole population. Such lists have no or little random element in their formation. They will not, in general, be statistically independent. Individuals are referred from one agency to another. Conversely, two agencies may overlap, or be perceived to overlap, in their function so that an individual attending one may choose not to attend the other. Agencies primarily serving different sections of the population, geographically, by age, severity of problem, or whatever, will also be negatively dependent. Such correlation bias results in the standard estimate being an underestimate if lists are positively dependent, an overestimate if negatively dependent.

With two lists dependence is not identifiable: independence must be taken on trust. With more than two, log linear models can be used to elucidate the pattern of dependence. With three lists, eight models with different combinations of pairwise dependences are available. Extrapolation to the missing cell to estimate U and thus N requires the three-factor interaction to be zero or else expressible, from a model at the individual level, as a function of lower-order interactions. Most studies use some form of model selection to determine which dependencies must be

allowed for, usually based on log likelihood ratio statistics from the analysis of deviance—the residual deviance and the difference between deviances from nested models. The complexity of a model is balanced against its badness of fit. Many authors have advocated parsimony: others argue that a parsimonious model with independence between lists is difficult to believe *a priori*. With the latter attitude, it might seem best to always fit the saturated model, including all identifiable interactions. There are two objections to this. The first is that if all lower-order interactions are nonzero, considerable doubt is cast on the necessary act of faith that the highest-order interaction is exactly zero. The other is that the estimate is very sensitive to any small cell count, and gives either a zero or infinite estimate for the missing cell if any cell count is zero, an event which becomes more likely the more lists are included in the study.

Certain patterns of cells with small numbers cause problems for all analyses, in that they reveal that the data contain very little information about some dependencies, either having few overlaps between lists or, perhaps surprisingly, when one list has a very high coverage. Asymptotic distributions of criteria used for model selection no longer hold, and should be replaced by bootstrap resampling of data predicted by different models.

More than three lists can be analyzed similarly except that, with more than five or six, the small number problem usually dominates and routine application of a program to analyze log linear models becomes impossible. One approach to model selection is then to start with the model including all pairwise dependencies, perhaps use backward elimination to simplify, and add higher-order terms by forward selection. However, making inferences conditional on the truth of a model selected from the data has been recognized as the Achilles heel of statistical inference. The alternative of averaging over all models has been applied to CR by several authors (see Sect. 5).

In the study design a list should be sought which is likely to be independent of the others. In studies on drug abuse, lists from police arrests have sometimes appeared to fulfil this condition relative to lists from social and health agencies. Questions of commonality of definition do arise. Moreover, patterns of interactions have to allow for individual heterogeneity as well as direct dependence between lists. The resulting confounding is discussed below.

3. *Heterogeneity*

Animals, particularly humans, are not the exchangeable colored balls required by ideal mathematical theory. There are reasons why individual i is not observed in list j. To estimate the numbers missing from all lists requires exploration of these reasons. Exploration and thought are crucial to give even a qualitative impression of the effect of such individual heterogeneity on the estimate. To quantify it needs modeling of p^{ij}, the probability that individual i is observed in list j. An early attempt by Cormack (1966) considered $p^{ij} = \theta_i * \beta_j$, but this is superseded by the logit model: $\log \{p^{ij}/(1-p^{ij})\} = \theta_i + \beta_j$ introduced by Rasch (1960) for educational test scores. In the context of repeated samples with the same protocol and same effort, so that β_j may be assumed constant, Burnham and Overton (1978) introduced the idea of the individual effects θ_i being a random sample from some distribution. Some workers have developed the consequences of assuming a beta distribution, but, if no specific form is assumed, the set of capture frequencies—numbers of individuals seen k times—are jointly sufficient statistics for N after integration over the distribution.

Without θ_i these models represent independent lists: differences between individuals necessarily cause dependence between lists. The Rasch model for individual behavior corresponds to a log linear model for cell counts with quasisymmetry, all interaction terms of the same order having the same value (Fienberg et al. 1999). Assuming a random model for an individual contribution's to his probability of being in a list, and integrating over that distribution, creates a superpopulation model for a sample survey. With nonrandom lists, any inference must be model-based, and estimates from superpopulation models are standard in survey sampling. However, while it is easy to accept that an estimate of, say, mean income in a superpopulation gives a reasonable inference for the one finite population of interest, doubts surface when the parameter of interest is the size of that one finite population. Unfortunately, allowing direct dependence between lists in a Rasch model for an individual is not the same as a log linear model with only that one extra dependence term. CR has also been expressed as a latent class problem by Agresti (1994).

Heterogeneity can be reduced by analyzing subpopulations as separate strata. Differences in models selected and in estimated inclusion probabilities can be studied. Samples from some strata may however be too small to permit effective model selection.

4. *Coverage*

Closely related to CR is the classical species problem. How many are there in the world? As a specific example, how many words did Shakespeare know? Concordances list all occurrences of any word in each play. Some words are used in only one play. There must be many he knew, but never used, just as there are many undiscovered species of beetle. Some words occur much more frequently than others, stark evidence of heterogeneity between individuals (words).

Apart from elegant models for specific applications, debate about sampling distributions and behavior of estimates has paralleled, and often predated, that about CR. A good review has been provided by Bunge and Fitzpatrick (1993). A key concept is that of coverage, the proportion of individuals of species represented in existing lists, perhaps easier to estimate than N, the unknown number of species. Chao and colleagues (e.g., Chao and Lee 1992) have developed various nonparametric estimates of N via coverage. These can be linked to martingale approaches to CR by Yip (1991).

5. *Bayesian Methods*

Bayesian methods have a long history in mark-recapture, even if Laplace's inverse probability argument is not considered pure Bayes. Early formulations make the point that independence of population size N and capture probabilities may not be appropriate *a priori*. Not only prior beliefs about N but also model uncertainty can be built smoothly into the analysis. Madigan and York (1997) develop models, with some analytical tractability, for decomposable graphical models with hyper-Dirichlet priors for cell probabilities, respecting the model's pattern of conditional independence, and allowing informative priors for N. Bayesian analysis of hierarchical log linear models has also been developed, an approach linked to the Rasch model by Fienberg et al. (1999).

6. *Census Undercount*

Recent national censuses show nontrivial undercounts. The possibility of adjustment has been widely explored and debated, in the US courts as well as the scientific and social literature. Because of the scale and expense, there are at least four major differences from applications discussed above. One is that matching and other nonsampling errors, present in other applications, gain prominence. Another is that the second list, from follow-up field survey, is necessarily performed only on a sample of the population. A third is that only two lists are the norm. The fourth is that estimates are required not only for the population as a whole, but for particular minority groups in relatively small geographical areas. Statistical summaries of the US debate have appeared in special sections or issues of the *Journal of the American Statistical Association* (1993) and of *Statistical Science* (1994), and in Anderson and Fienberg (1999).

Matching errors tend to identify two records of the same individual as different rather than vice versa. With two lists this leads to overestimation of N. Attempts to correct this, by algorithms which declare individuals identical, even when differing in some identifiers, can give false positive matches, and hence underestimation. Probabilistic record linkage is discussed by Jaro (1995) and models for its effect on CR estimates by Ding and Fienberg (1996).

In recent usage, the second list comes from a postenumeration survey (PES), an intensive field survey of small geographical areas, defined slightly differently in different countries, in a complex randomized sample plan, typically with strata and clusters at two or more levels. The relationship between CR and ratio estimators allows CR models to be combined formally with the sample design, though sampling theory suggests other ways of using auxiliary variables.

With two lists, the required independence must be built into the design. Sampling units must be defined, small enough, to minimize heterogeneity. Small numbers cause instability, and attempts are made to correct this by spatial smoothing over neighboring areas. The multiplicity of ways in which sample-based inference can be integrated with model-based inference from multilevel data with CR models is a topic of much current debate.

7. *Summary*

The first difficulty with the use of CR is the definition of the population. If in a study of diabetes all observed individuals can be stratified by the medical care prescribed, then the missing numbers are not of undiagnosed diabetics, but of diabetics who have been diagnosed, but whose records have been lost. If different agencies have different definitions of homelessness, or cover different geographical areas, then which population is it hoped to estimate? No statistical analysis of the observed data resolves these problems. Second, there is the problem of matching individuals, a problem which grows with the size of the study, and the separation of the analyst from the data collection. The more sensitive the subject of the study, the more prevalent will be deliberate misinformation from an individual. Matching problems will vary between countries that have differing levels of confidentiality and anonymity. Third, there are the analytic problems caused by the unreality, in many cases, of the assumption of random sampling. Lists are not independent, individuals are not behavioral clones, and the two aspects of dependence and heterogeneity may be at least partially confounded unless care is taken to obtain data which allows disentanglement. Finally, there is the problem of the variety of analytical approaches—which to believe? With several lists, some models can be rejected because of lack of fit to the observed counts. However, failure to reject a model is not positive proof of its truth. Estimation of population size still requires extrapolation to the unobserved, the validity of which is untestable without randomness in the sampling.

Although absolute estimates of population size are often unjustifiable, it may be possible, by repeating the same protocol for list-formation at different times or places, to obtain reliable assessments of population change or geographical variation in incidence of a disease or social phenomenon.

See also: Censuses: History and Methods

Bibliography

Agresti A 1994 Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* **50**: 494–500

Alho J M 1990 Logistic regression in capture-recapture models. *Biometrics* **46**: 623–5

Anderson M J, Fienberg S E 1999 *Who Counts? The Politics of Census-taking in Contemporary America*. Russell Sage Foundation, New York

Bishop Y M M, Fienberg S E, Holland P W 1975 *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA

Bunge J, Fitzpatrick M 1993 Estimating the number of species: a review. *Journal of the American Statistical Association* **88**: 364–73

Burnham K P, Overton W S 1978 Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**: 625–33

Chao A, Lee S-M 1992 Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* **87**: 210–17

Chapman D G 1951 Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publications in Statistics* **1**: 131–60

Cochran W G 1978 Laplace's ratio estimator. In: David H A (ed.) *Contributions to Survey Sampling and Applied Statistics*. Academic Press, New York, pp. 3–10

Cormack R M 1966 A test of equal catchability. *Biometrics* **22**: 330–42

Cormack R M 1968 The statistics of capture-recapture methods. *Annual Review of Oceanography and Marine Biology* **6**: 455–506

Cormack R M 1992 Interval estimation for mark-recapture studies of closed populations. *Biometrics* **48**: 567–76

Darroch J N 1958 The multiple-recapture census I: estimation of a closed population. *Biometrika* **45**: 343–59

Ding Y, Fienberg S E 1996 Multiple sample estimation of population and census undercount in the presence of matching errors. *Survey Methodology* **22**: 55–64

Fienberg S E, Johnson M S, Junker B W 1999 Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society* **A162**: 383–405

Huggins R M 1989 On the statistical analysis of capture experiments. *Biometrika* **76**: 133–40

James T B, Price N A 1976 Measurement of the change in populations through time: capture-recapture analysis of population for St. Lawrence parish, Southampton, 1454 to 1610. *Journal of European Economic History* **5**: 719–36

Jaro M A 1995 Probabilistic linkage of large public health data files. *Statistics in Medicine* **14**: 491–8

Laplace P-S 1786. Sur les naissances, les mariages et les morts a Paris depuis 1771 jusqu'en 1784; et dans toute l'etendue de la France, pendant les annees 1781 et 1782. *Memoires de l'Academie des Sciences (1783)* 693–702

Madigan D, York J C 1997 Bayesian methods for estimation of the size of a closed population. *Biometrika* **84**: 19–31

Pearson K 1928 On a method of ascertaining limits to the actual number of marked members in a population of given size from a sample. *Biometrika* **20A**: 149–74

Pollock K H, Otto M C 1983 Robust estimation of population size in a closed animal population from capture-recapture experiments. *Biometrics* **39**: 1035–49

Rasch G 1960 *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press, Chicago

Seber G A F 1982 *The Estimation of Animal Abundance and Related Parameters*. Griffin, London

Skalski J R, Robson D S 1992 *Techniques for Wildlife Investigations: Design and Analysis of Capture Data*. Academic Press, San Diego, CA

Thompson S K 1992 *Sampling*. Wiley-Interscience, New York

Yip P 1991 A martingale estimating equation for a capture recapture experiment in discrete time. *Biometrics* **47**: 1081–8

R. Cormack

Populism

Populism is a form of mass politics. The core idea, as the etymology would suggest, rests with the claim to represent or act in the name of the people, understood as ordinary or common people, the majority, or the masses, as opposed to elites, privileged or special-interest groups, the establishment, or the power bloc. It is generally distinguished from two other forms of mass politics: class politics (based on class organizations of unions and social democratic parties) and pluralism (based on multiple, shifting, overlapping interest groups and a rejection of the populist dualism that stresses the representation of the general will of the people vs. some minority). 'Populism' has been used as a term of approbation by movements themselves and as a term of opprobrium by critics. Analysts have also used the term, applying it to diverse phenomena.

1. Empirical and Conceptual Variation

With origins in ancient Rome, where the Populares opposed the aristocratic Optimates, the first modern usage is probably the Russian Narodnik movement of the 1860s–70s, which never achieved mass adherents but represented a radical articulation of the idea that wisdom resides with the common people and that socialism could be based on the traditional peasant commune. American populism, beginning in the 1880s

and continuing particularly in the South through the twentieth century, has received much attention, as has Latin American populism of the mid-twentieth century. Currently there is a resurgence throughout the world of movements and leaders that have been dubbed populist: in Latin America (Collor, Fujimori, Chavez, Menem), in Western Europe (Haider, Le Pen), in the United States (Buchanan), and in post-Communist states (Meciar, Milosevic, Lukashenko). Many other movements have also been considered populist by at least some analysts, including those led by Nasser, Nyerere, Mao, Hitler, Father Coughlin, Gandhi, Chartists, and even Jimmy Carter.

The populist label has been attached to such a wide variety of political movements (right and left, from above and from below) that it is hard to stabilize any core meaning that can work rigorously as an analytical concept. Some analysts propose conceptualizations that are compilations of traits of specific movements, providing no basis for abstracting out a more parsimonious conceptualization that serves for broader, comparative analysis (Pipes 1964, Goodwyn 1976, Hicks 1955). Others, engaged in a more comparative quest, have highlighted a particular trait or subset of traits (Germani 1962, Di Tella 1965, Ionescu and Gellner 1969, Laclau 1977, Conniff 1999, Taggart 1995).

These approaches typically emphasize some components of the following.

(a) *Mobilization* or collective action from below, with deep roots in local communities.

(b) A *reform* or anti-status quo movement.

(c) A *support base* constructed by a discourse that attempts to promote an identity as the 'people' rather than a class, and sometimes aimed at neutralizing class identities. The political group constructed in this way is embedded in a particular adversarial juxtaposition to a vague 'other': elite (economic and/or political), alien, special interest, or minority group, often corrupt or conspiratorial. At the same time that a mass discourse seeks to construct an inclusionary base, populist movements in fact may be quite specifically supported by a particular class or classes.

(d) A multiclass or nonclass-based *coalition*.

(e) A *heteronomous* or politically subordinated, masssupport base, lacking autonomous power and subordinated to the leader and/or the interests of another class in the coalition.

(f) A particular *leadership style*, specifically strong, personalized, sometimes charismatic leadership.

(g) A leader–mass *linkage* that is direct, rather than mediated through organizations.

(h) Use of *rhetoric* and *oratory* that is anti-theoretical and anti-intellectual rather than coolly analytic, expository, explanatory, or abstractly elaborated in theory or ideology.

(i) An '*irrational*' form of politics in the sense that it attracts support on an emotive or moralistic rather than an interest or programmatic basis: sometimes identified as a politics of status-anxiety rather than economic demands.

(j) A set of *economic policies* or an economic model that emphasizes growth, distribution, and demand-side stimulation, and deemphasizes fiscal and market constraints. In this way it is seen as an economic policy that is undisciplined, inefficient, and distorted or contaminated by the goal of winning political support.

Confronted with the great number and diversity of both movements and traits, Tindall (1972) proclaimed a 'semantic identity crisis.' Some analysts have responded with an attempt to construct subtypes. Canovan adopted a 'phenomenological' and 'naturalist' approach—an admittedly atheoretical exercise that yields seven types. These accommodate diversity but 'do not really look like seven varieties of the same thing; on the contrary, some of them seem quite unconnected with others' (Canovan 1982, p. 551). Roberts (1995) attempted to retain a conceptual core or commonality while accommodating diversity by proposing a 'radial' conceptualization. Here the aim is to delineate a set of traits that defines a 'pure' populism, and to recognize that subtypes (or actually existing populisms) may not in fact have all the traits in the definition. The diversity of populist movements is thus accommodated by constructing a typology of 'diminished' subtypes—types differentiated in terms of traits in the core definition which they lack.

2. *Family Resemblance and Analytical Dimensions*

The important point about 'populism' as it has been used is not just that types of populism lack the full set of traits. Taken as a whole, conceptualizations of populism share a family resemblance. However, different combinations of traits describe a diverse set of cases (or types) that may share few, if any, traits. Thus, 'populism' can encompass quite incongruous, even 'opposite' or inverse movements.

In the above list of traits, each seems to follow the other unproblematically. However, as one moves from the beginning toward the end of the list, one travels from authentic reform movements from below toward demagoguery—from programmatic movements based in grass-roots collective action and/or organization to instrumental top-down mobilizations of support by political leaders to facilitate their own political ambitions, or, in a Marxist framework, to promote the interests of a superordinate class. In between, populism has been associated with quite standard electoral politics and the pragmatic strategies of political leaders, particularly in catch-all parties, in order to achieve electoral victory, strategies which can be presented positively, as rational politicians instrumentally adjusting positions to conform to the median voter, and negatively, as pandering or the marketing of

candidates. (Canovan (1982) includes politicians in catch-all parties as populists.)

More important than trying to define some common core of populism is delineating some dimensions for comparing very different forms of mass politics. The different traits printed in italics in the above list point to distinct 'domains' which can be treated as dimensions on which movements analyzed as populist vary. The range of variation is evident from a description of three of the most successful sets of populist mass movements: the historical, 'classical' cases of populism in the United States and Latin America, which have received the most attention in the literature, and the recent cases emerging in diverse settings throughout the world. These cases differ dramatically with respect to, for example, leader–follower linkage, composition and heteronomy of support base, class/mass discourse, programmatic/interest vs. 'irrational' orientation, and economic policy.

3. *Concrete Examples*

American populism of the 1880s and 1890s arose out of the economic distress of farmers and was based in a web of thousands of independent organizations formed to address grievances at a local level. In 1892 the two regional alliances which these organizations had formed founded the Populist Party to contest national elections. The Populist Party thus emerged from the autonomous mobilization and organization of farmers. Its platform is interesting for its programmatic nature and its class discourse. Though free silver was one of the most important and well-known demands, the platform included many other economic and political planks: government distribution of credit; nationalization of railroads; graduated income tax; direct election of president, vice-president, and senators; initiatives and referenda; and a number of demands of interest to urban labor and unions. The Populists saw themselves as embodying the will of all the 'toiler classes' against the 'criminal classes.'

After a poor showing, the party retreated and in 1896 backed the candidate of the Democratic Party, William Jennings Bryan, who adopted many Populist planks and whose campaign reached oratorical heights. Those (Hicks 1955, Goodwyn 1976) who see populism as a grass-roots agrarian movement identify 1892 as its peak, after which it was coopted by the Democrats under Bryan. To the extent they see an ongoing populist thread, the analysts who identify populism with agrarian radicalism trace it forward to other progressive movements.

The conceptual confusion about populism is reflected in the fact that those who identify populism with an unmediated, oratorical leadership style see Bryan not as the coopter but as the prototype (Hofstadter 1955). Emphasizing majoritarian and crusading traits that imply intolerance and express the irrationality of status anxiety, they trace American populism to the mobilization of an unorganized support base by paternalistic, demagogic Southern governors (e.g., Huey Long and George Wallace) and to McCarthyism, in which economic interests were even more distant, with the primary targeting of a foreign enemy and conspiracy against the nation. This alternative conception obviously involves very different sets of attributes.

'Classical' Latin American populism (of Cardenas in Mexico, Peron in Argentina, Betancourt in Venezuela, Haya de la Torre in Peru) emerged in the 1930s–40s as a strategy to win political support among the unionized proletariat in the context of the struggle by rising urban elites to attain or consolidate power. These movements consisted of urban, multiclass coalitions. It was a coalition between unequal partners, with the autonomy of workers limited to varying degrees, but generally workers had some bargaining power and political weight. The discourse was notably classist, and support mobilization occurred on the basis of real material and organizational concessions, not simply rhetorical appeal or charismatic leadership. Far from involving direct, unmediated leader–follower linkages, these were organizing periods when unions were strengthened, political parties were founded, and unions became affiliated to these parties. Union and party organizations played a central role in mobilizing support (Collier and Collier 1991).

If the first phase of early US populism is the autonomous expression and demands of farmer organizations and classical Latin American populism is the strategy of political leaders to construct an alliance with the unionized proletariat, contemporary populism is quite different. Like some US strands, contemporary populism ('neopopulism' in Latin America) generally consists of electoral support mobilization by leaders who make strong direct, rhetorical appeals to the 'people' unmediated by organizations. Yet unlike the 'classical' cases, contemporary populism occurs in the context not of a state interventionist economic model consistent with demand-side (re)distributive policies, but of the global move to marketizing policies, which are often unpopular. These policies have led to social–institutional change involving the growth of the informal sector and contingent workforce and the defensive position of organized labor. Politicians find fertile ground in the actual or threatened grievances and social dislocation resulting from new economic models, but once in office often adopt marketizing policies. Nonprogrammatic, nonclassist, noninterest-based appeals for support are thus a rational approach for sustaining governments pursuing marketizing policies.

The contrast of populisms is particularly evident in Latin America. There, classical populism represented a certain kind of inclusion and controlled mobilization of the new proletariat; neopopulism represents exclusion and demobilization. The one organizes classes;

the other disorganizes them or disarticulates their organizations. One is an alliance and bargains with the organized working class with broad payoffs (e.g., wage policy, food subsidies); the other is a rhetorical appeal to the unorganized masses (and particularly targets the growing informal sector) with selective, paternalistic distributions (e.g., local infrastructure). Postcommunist populism is a parallel form of 'post-organized' mass appeal when class solidarities and organizations have weakened or disappeared. In advanced countries, populist appeals have similarly targeted those threatened by the prospect or reality of the global relocation of production and labor, economic insecurity, and the pressure to flexibilize or liberalize labor regimes.

These examples suggest the diversity of 'populist' movements. The minimal commonality they share—that they seek to 'represent' or appeal to ordinary people in one way or another—is not very interesting; their differences are more important for elucidating the who, what, why, or how questions one might ask about them, and fitting them into larger theoretical and explanatory frameworks. If one is to use the term as more than political discourse or as a broad label, if the term is to do analytical work, it will be most helpful to delineate a common set of dimensions for scoring cases. Analysis of the resulting subtypes, with their distinct logics and dynamics, may be more useful than an ultimately elusive effort to achieve a common overall conceptualization of populism.

See also: Mass Media, Political Economy of; Political Culture; Political Discourse; Political History (History of Politics); Political Sociology; Reform: Political; Social Movements, Sociology of; Voting: Class

Bibliography

Canovan M 1982 Two strategies for the study of populism. *Political Studies* **30**: 544–652

Collier R B, Collier D 1991 *Shaping the Political Arena*. Princeton University Press, Princeton, NJ

Conniff M (ed.) 1999 *Populism in Latin America*. University of New Mexico Press, Tuscaloosa, NM

Di Tella T 1965 Populism and reform in Latin America. In: Veliz C (ed.) *Obstacles to Change in Latin America*. Oxford University Press, Oxford

Germani G 1962 *Politíca y sociedad en una época de transción*. Paidós, Buenos Aires

Goodwyn L 1976 *Democratic Promise: The Populist Movement in America*. Oxford University Press, New York

Hicks J 1955 *The Populist Revolt. A History of the Farmer's Alliance and the People's Party*. University of Nebraska, Lincoln, NE

Hofstadter R 1955 *The Age of Reform. From Bryan to FDR*. Knopf, New York

Ionescu G, Gellner E (eds.) 1969 *Populism. Its Meaning and National Characteristics*. Macmillan, London

Laclau E 1977 *Politics and Ideology in Marxist Theory. Capitalism, Fascism and Populism*. NLB, London

Ostiguy P 1997 Peronism y antiperonismo: Bases socioculturales de la identidad política en la Argentina. *Revista de Ciencias Sociales* **6**: 133–213

Pipes R 1964 Narodnichestvo: A semantic inquiry. *Slavic Review* **24**: 441–58

Roberts K H 1995 Neoliberalism and the transformation of populism in Latin America. The Peruvian case. *World Politics* **48**: 82–116

Taggart P 1995 New populist parties in Western Europe. *West European Politics* **18**: 34–51

Tindall G 1972 Populism: A semantic identity crisis. *Virginia Quarterly Review* **48**: 501–18

Weyland K 1996 Neopopulism and neoliberalism in Latin America: Unexpected affinities. *Studies in Comparative International Development* **31**: 3–31

R. B. Collier

Pornography

The topic of pornography has been discussed from many vantage points and often hotly debated by scholars and researchers. Some have focused on the functions and effects of pornographic mass media as they pertain to understanding the impact of pornography on individual consumers. For example, some theorists have argued that the particular type of pornography an individual is attracted to may serve as a 'window to the mind,' revealing core developmental conflicts of the psyche (Stoller 1976). Others have considered the beneficial and/or negative effects of pornography on individuals in the realms of fantasy, sex education, esthetic/artistic expression, sexual addiction, etc. (Gagnon 1977, Gordon 1980). Yet other scholars have approached the topic at a broader level, exploring such topics as the relevance of pornography to group relations, social institutions, and the role of government regulation (e.g., Strossen 1995, Weinstein 1999). Assertions have been made that pornography use contributes to unequal power relations between dominant and subordinate classes, particularly male subjugation of women (Itzin 1992). Others have claimed that pornography is a degrading threat to social institutions and values, such as the traditional family structure (Wojtyla 1981); and that it has functioned as an agent of social change (Sorokin 1956), even as a contributing factor to major social revolutions (Hunt 1993).

It is beyond the scope of this article to consider all of these diverse approaches to the topic of pornography. Instead, the focus here may be described as a 'social science-centered' approach revolving around two central questions: First, what are the findings of social scientific studies on the effects of pornography on men (who are much more likely than women to be the consumers of such media) in the areas of aggressive attitudes and behavior? Second, how can we under-

stand this major asymmetry in male vs. female consumption of pornography?

An extremely large media industry has existed for many years that sells sexually explicit material, including movies, magazines, books, videos, and, increasingly, Internet-based media. This 'pornography industry' was recently described by Forbes magazine as a $56 billion global industry that has become much more mainstream in recent years. Some companies marketing very explicit portrayals on the Internet are now listed on the Nasdaq stock exchange (Morais 1999).

1. Defining Pornography

Pornography refers to sexually explicit media that are primarily intended to sexually arouse the audience. Typical content consists of nudity and/or various sexual acts (e.g., intercourse, fellatio, etc.). Hunt (1993) notes that the production of media for this purpose is a relatively modern phenomenon, beginning toward the latter part of the nineteenth century. O'Toole (1999, p. 1) similarly argues that earlier in European history the use of sexually explicit depictions was primarily for political purposes, '… to satirise, criticise, to tilt at the Church, the state, the monarchy. … Porn was controlled during this period not because it was obscene but because it was seditious, blasphemous, or defamatory.'

Pornography does not include sexual content that is 'embedded' or interwoven with much nonsexual content (e.g., a movie that primarily includes nonsexual content but also has one or a few sexual scenes). The research literature in that area has been recently summarized by Malamuth and Impett (2000). Sometimes the distinction between 'embedded' and explicit sexual content is not so clear. For example, *Playboy* magazine regularly includes considerable nonsexual content (e.g., interviews with politicians and jazz artists) as well as nude portrayals; the Starr Report, an official state document focusing on President Bill Clinton, included much sexually explicit content. Nevertheless, the focus here is on content that is primarily designed to sexually arouse the consumer.

In terms of legal definitions, the terms 'pornography' and 'obscenity' need to be distinguished. Obscenity refers to pornographic content that has been judged by the legal system to be illegal. In the USA, most states use the 1973 Supreme Court decision in Miller vs. California as the legal definition of 'obscene.' To be judged obscene under this case the material must satisfy three conditions: (a) The average person, applying contemporary community standards, would find that the work, taken as a whole, appeals to the prurient interest; (b) the work depicts sexual conduct in a patently offensive way; and (c) the work, taken as a whole, lacks serious literary, artistic, political, or scientific value.

With increasing use of new types of media such as the Internet, there may be changes in the content of pornography and the feasibility that governments can restrict access to content considered by some as offensive or harmful. For example, Barron and Kimmel (2000) recently measured the levels of sexually violent content in representative samples of sexual magazine, video, and Internet sex stories sites. They found that the Internet content was significantly higher in violence than the videos or magazines. This difference was particularly high when considering 'extreme' forms of violence (e.g., use of weapon on a victim, torture, mutilation) where 17 percent of the Internet sites contained such material as contrasted with only 2 percent in the other media. Furthermore, the Internet content depicted men in dominant positions, as victimizer and not victim, in far greater proportion than magazines and videos. The authors note that the Internet is more 'democratic' than the other media, which are constrained by dependence on commercial advertisers. The researchers suggest the Internet content is '… as close as one can get to men's direct expressions of their own fantasies, unconstrained by the demands of the marketplace or the high costs of producing and distributing those fantasies to others' (p. 166). It should be noted, however, that the men who currently use such Internet sex sites may not be representative of the larger population and may have particular attraction to more aggressive content.

2. Research on the Effects of Sexually Explicit Media

2.1 Differences Among Studies

Varied methodologies and dependent measures have been used to study the effects of pornography. Much research has centered on the potential link between pornography exposure and aggression, particularly sexual aggression. Most of the studies conducted within North America, may be described along two orthogonal dimensions based on whether they used experimental (random assignment to conditions) or correlational methodology, and whether the dependent variable they assessed was a response presumably affecting sexual aggression (e.g., attitudes supporting such aggression) or some measure of actual aggression. Such studies may therefore be organized into four categories: (a) experiments manipulating exposure to pornography and examining the effect upon attitudes as the dependent measure; (b) experiments manipulating exposure to pornography and assessing the impact on aggressive behavior as the dependent measure (in laboratory settings); (c) studies of the correlations between individual differences in pornography exposure in naturalistic settings and men's attitudes supporting aggression; and (d) two

types of studies examining the correlations between pornography exposure and aggressive behavior in naturalistic settings, those comparing criminals vs. noncriminals on pornography use, and within samples of noncriminals, those comparing different levels of reported past sexual aggression and individual variability in use of pornography.

Several meta-analyses of pornography systematically synthesize the research literature in three of the four categories and partly in the fourth category of study. Research correlating pornography use and sexual aggression within the noncriminal population has not yet been synthesized by a metanalytic review. Here, fortunately, a recent publication using a national representative sample of all men in some form of posthigh school education is available to help fill this void

Below is a summary of the findings of the various meta-analyses and the findings of the national study.

2.1.1 Experiments on the effects of pornography on attitudes. Allen et al. (1995a, 1995b) conducted a meta-analysis of experiments, both laboratory and field studies, on the effects of pornography exposure on attitudes. The meta-analysis revealed that across the various studies conducted, there was a significant though modest effect, with violent pornography resulting in significantly greater increase in attitudes supporting aggression than exposure to nonviolent pornography (although there was also some effect for both types of pornography exposure). Therefore, the overall data showed that exposure to pornography, under controlled conditions, does cause an increase in attitudes supporting sexual aggression.

2.1.2 Experiments on the effects of pornography on laboratory aggression. The meta-analysis of Allen et al. (1995a) concluded that men exposed to nonviolent or violent pornography portraying sexual acts were more aggressive than those exposed to neutral content, in the laboratory research reviewed. In contrast, exposure to portrayals of nudity alone (without any sexual acts) resulted in reduced aggression in comparison to exposure to neutral content.

2.1.3 Pornography and attitudes favoring sexual aggression in naturalistic settings. Although correlational studies do not enable cause and effect conclusions, they often have the advantage of assessing responses occurring in naturalistic settings. Allen et al.'s (1995b) meta-analysis examined these studies. They found that there was a weak positive correlation that was judged not to be statistically significant between amount of exposure to pornography and attitudes favoring sexual aggression. Therefore, the correlational data for attitudes in naturalistic settings did not yield supportive data for the conclusions emerging from the experimental studies conducted in controlled settings.

2.1.4 Pornography and aggression in naturalistic settings. Allen et al. (2000) conducted a meta-analysis focusing on the pornography use of convicted sex offenders as compared to those of men from the noncriminal general population. They examined several types of dependent measures: (a) frequency of pornography use; (b) age of first exposure; (c) the degree to which pornography was a direct prelude to some sexual act (masturbation, consensual sex, or forced sex); (d) and degree of sexual arousal, measured by direct genital measures of such arousal. The findings across all of the studies and measures combined did not show that criminality was associated with frequency of pornography exposure nor with age of first exposure. After viewing pornography, criminals were more likely than noncriminals to engage in some sexual acts such as masturbation, consensual, or criminal sex.

Finally, these investigators conducted a related meta-analysis examining the degree to which different types of stimuli affect people in laboratory settings. They focused on studies of physiological sexual arousal (assessed by direct genital measures) to various types of content. The researchers concluded that sexual criminals were generally more sexually aroused than noncriminals in laboratory studies examining physiological arousal to sexual stimuli. However, studies which separated portrayals of consenting and nonconsenting sex found that in comparison to noncriminals, sex criminals were more aroused by violent sex. By contrast, the difference was in the opposite direction with consenting sexual portrayals, with nonrapists showing relatively more sexual arousal. It should be noted that although criminals may be less aroused sexually by consenting depictions than noncriminals, the data on 'sexual acting out' (see above) suggest that they may still be more likely than noncriminals to engage in some sexual activity following exposure to either types of pornography exposure.

Although a meta-analysis has not been conducted on the few available studies examining among noncriminals the potential relationship in naturalistic settings between use of pornography and individual differences in self-reported sexual aggression, these few studies do point to a significant association (for a review see Malamuth et al. 2000). However, these few studies have some methodological limitations, which were rectified in a national probability study recently reported by Malamuth et al. (2000). This study found that for the majority of men, high

pornography use is not indicative of high risk for sexual aggression: Among those classified as being at relatively low risk for committing sexually aggressive acts (based on various measures previously found to predict such risk), there was only a small difference in sexually aggressive behavior between those who reported different levels of pornography use.

In contrast, pornography use was indeed a very good 'marker' of higher sexual aggression levels when these researchers considered a particular group of men: those who, based on the other measures of risk for sexual aggression, were previously determined to be at high risk. It was found that among these high-risk individuals, those who additionally were very frequent users of pornography were much more likely to have engaged in sexual aggression than their counterparts who consumed pornography less frequently. The researchers cautioned, however, that such data alone cannot be used to infer cause and effect conclusions, even if they are useful for risk assessment. They did speculate that:

> associations between pornography consumption and aggressiveness toward women could be explained by a circular relationship between high coercive tendencies and interest in certain content in pornography, whereby aggressive men are drawn to the images in pornography that reinforce and thereby increase the likelihood of their controlling, impersonal, and hostile orientation to sexuality. The way relatively aggressive men interpret and react to the same pornography may differ from that of nonaggressive men. (Malamuth et al. 2000, p. 85)

2.2 Cross-cultural Comparisons

One of the most glaring apparent contradictions in the literature on pornography results from research conducted in different cultures (as well as using different methodologies). In research conducted primarily in Denmark (Kutchinsky 1970, 1991) and in Japan (Diamond and Uchiyama 1999) there has not been evidence of increased criminal sexual acts as a function of the wider availability of pornography. In commenting on such findings, Malamuth and Donnerstein (1984, p. 141) suggested some time ago that '… there may be considerable variations among individuals within a culture in susceptibility to media influences. Similarly, cultural factors may create major individual differences in the role and impact of media stimuli on members of differing societies.'

A similar point was also aptly made by Giglio (1985, pp. 289–90):

> The cultural environment in each country is a factor to consider in understanding the prevailing public attitudes towards pornography. The Danes appear to enjoy a more natural approach to sex in general. Public nudity, for example, is more acceptable in Denmark than in the United States … In a society where it is possible on a warm summer Sunday to visit the Rosenborg Palace in the heart of Copenhagen and see dozens of partially nude women in the surrounding gardens, usually accompanied by family or friends, without the slightest public disturbance, lends creditability to Kutchinsky's theory that the Danes lack the sustaining desire for pornography to make it a profitable domestic business.

In keeping with the position that aggressive-sexual predispositions may moderate the impact of exposure to certain types of pornography, it would be expected that within countries such as Denmark, fewer men than in the USA are at high risk for sexual aggression. If this is correct, then the lack of an association between pornography use and sexual aggression would map nicely on to the findings we have obtained. Although we do not have direct evidence bearing on this issue, there are some potentially relevant data. Zak and Knack (in press) compared the levels of trust that people had for each other in various countries throughout the world. They found strong differences, with countries such as Denmark being at the very highest levels of trust between people, and these levels were considerably higher than in the USA. To the extent that such general levels of trust may also be associated with trusting between men and women (an important component of the measures used to assess risk for sexual aggression), men may be at relatively low levels of risk for sexual aggression in a society such as Denmark. Therefore, when considering the possible impact of pornography on individuals within a cultural context, it may be essential to take into account the larger context within which such exposure is embedded.

3. Gender Differences in Pornography Consumption

What is clear is that across all of the cultures that have been studied, substantial differences exist in the type of sexually explicit media men and women are more likely to be attracted to (for a review, see Malamuth 1996, Oliver 2000). Many studies provide evidence that large gender differences exist in the consumption of and gratification derived from sexually explicit media. For example, in comparison to women, men consume more regularly, are more sexually aroused by, have more favorable attitudes to, and react with less negative affect to, portrayals featuring nudity of the opposite sex and/or sexual acts devoid of relationship context (e.g., Bryant and Brown 1989, Laumann et al. 1994, Mosher and MacIan 1994, Stauffer and Frost 1976). In contrast, men are less likely than women to consume sexually explicit media that emphasize 'sexual communion' (e.g., sex accompanied by emotional closeness) and romance (e.g., Faust 1980, Krentz 1992, Lawrence and Herold 1988, Perse 1994). Various other related gender differences found in pornography research are summarized by Malamuth (1996).

3.1 Adaptive Problems and Sexual Media

These differences can be discussed using the theoretical framework of evolutionary psychology, which predicts when gender differences are or are not expected, their direction, and why they occur (Buss 1999). Generally, males and females are expected to have the same psychological mechanisms in domains where natural selection has favored a universal human adaptive solution. However, different mechanisms are expected to exist in domains in which males and females have recurrently encountered different evolutionary problems and/or where the 'optimal solutions' differed for the genders. Evolutionary theory predicts that sexuality is one of these domains. Somewhat different mating strategies probably evolved for males and females due to differences in maximum reproductive capability and parental investment. The consumption of sexually explicit media might be the result of differing evolved psychological mechanisms interacting with environmental factors.

As Malamuth (1996) discusses in some detail, there appears to be a high correspondence between the adaptive problems that evolutionary psychologists have argued underlies a 'short-term' mating strategy and the recurrent content of sexually explicit media primarily consumed by men. According to evolutionary psychologists, males who engaged in ancestral environment in a short-term sexual strategy faced problems about how to get access to as many fertile females as possible while minimizing the commitment and investment to any one woman. Pornography marketed towards males primarily portrays casual sex with numerous, accessible women who display fertility cues through age and body shape. There is similarly a high level of correspondence between female-oriented sexual media (e.g., romance novels, portrayals of sex within a relationship context) and the adaptive problems that led females to evolve a long-term sexual strategy. These problems included identifying and securing commitment from a man who had the ability (e.g., high status) and willingness to invest heavily in her and her offspring. Additionally, due to the increased vulnerability associated with pregnancy and childbirth, females preferred to mate with men who could offer physical protection. Finally, women attempted to identify men who were kind and sensitive, responses that may be strong cues of parenting skills. Romance novels, one of the most popular forms of sexually explicit media for women, frequently portray a woman securing a relationship with a high status, physically protective male who also possesses parenting skills and the willingness to invest time and resources in a single female and her offspring.

In contrast, some researchers have argued that the gender differences described above can be explained exclusively as a product of the differing socialization of males' and females' cultural roles. While this process of socialization is probably relevant to these and many other gender differences, such a model needs to consider why such norms and rules of socialization emerged in the first place as well as to account for a large number of other related gender differences in the area of sexuality (Buss and Schmitt 1993).

4. Concluding Comments

Pornography research has often been influenced by ideological/political perspectives with vested interests in particular conclusions. This may have led to framing of research questions and design of some studies in ways that encourages simple conclusions, while not readily accommodating more nuanced conclusions (Linz and Malamuth 1993). Nevertheless, research in this area has progressed considerably. In future work, it is essential not to use an 'either–or' lens in which research is cast simply into questions such as whether pornography exposure is generally harmful or not. As suggested in the research summarized here, depending on such factors as the cultural milieu, the individual's background, the particular content of the stimuli, the type of responses focused on, the content of exposure, the consumer's environmental circumstances, and the way 'harm' is defined, differing conclusions may result. Some may wish for a 'one handed' set of conclusions, but the research more accurately justifies a 'multihanded' perspective that reveals the richness and complexity of the issues related to the study of pornography.

See also: Aggression in Adulthood, Psychology of; Freedom of the Press; Internet: Psychological Perspectives; Media Effects on Children; Media Ethics; Rape and Sexual Coercion; Regulation: Sexual Behavior; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Violence and Media

Bibliography

Allen M, D'Alessio D, Brezgel K 1995a A meta-analysis summarizing the effects of pornography II: Aggression after exposure. *Human Communication Research* **22**: 258–83

Allen M, D'Alessio D, Emmers-Sommer T M 2000 Reactions of criminal sexual offenders to pornography: A meta-analytic summary. In: Roloff M (ed.) Communication yearbook 22. Sage, Thousand Oaks, CA, pp. 139–69

Allen M, Emmers-Sommer T M, Gebhardt L, Giery M 1995b Pornography and rape myth acceptance. *Journal of Communication* **45**: 5–26

Barron M, Kimmel M 2000 Sexual violence in three pornographic media: Toward a sociological explanation. *The Journal of Sex Research* **37**: 161–8

Bryant J, Brown D 1989 Uses of pornography. In: Zillmann D, Bryant J (eds.) *Pornography: Research Advances and Policy Considerations*. Erlbaum, Hillsdale, NJ, pp. 25–55

Buss D M 1999 *Evolutionary Psychology: The New Science of the Mind*. Allyn & Bacon, Boston

Buss D M, Schmitt D P 1993 Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review* **100**: 204–32
Diamond M, Uchiyama A 1999 Pornography, rape, and sex crimes in Japan. *International Journal of Law & Psychiatry* **22**: 1–22
Faust B 1980 *Women, Sex, and Pornography: A Controversial and Unique Study*. Macmillan, New York
Gagnon J H 1977 *Human Sexualities*. Scott, Foresman, Glenview, IL
Giglio E D 1985 Pornography in Denmark: A public policy model for the United States? *Comparative Social Research* **8**: 281–300
Gordon G 1980 *Erotic Communications: Studies in Sex, Sin and Censorship*. Hastings, New York
Hunt L 1993 Introduction: Obscenity and the origins of modernity, 1500–1800. In: Hunt L (ed.) *The Invention of Pornography: Obscenity and the Origins of Modernity, 1500–1800*. Zone, New York, pp. 9–45
Itzin C (ed.) 1992 *Pornography: Women, Violence and Civil Liberties*. Oxford University Press, Oxford, UK
Krentz J A 1992 (ed.) *Dangerous Men and Adventurous Women*. University of Pennsylvania Press, Philadelphia, PA
Kutchinsky B 1970 Pornography in Denmark: Pieces of a jigsaw puzzle collected around New Year 1970. In: *Technical Reports of the Commission on Obscenity and Pornography*, US Government Printing Office, Washington, DC, Vol. 4
Kutchinsky B 1991 Pornography and rape: Theory and practice? Evidence from crime data in four countries where pornography is easily available. *International Journal of Law and Psychiatry* **14**: 47–64
Laumann E O, Gagnon J H, Michael R T, Michaels S 1994 *The Social Organization of Sexuality: Sexual Practices in the United States*. University of Chicago Press, Chicago
Lawrence K, Herold E S 1988 Women's attitudes toward and experience with sexually-explicit materials. *Journal of Sex Research* **24**: 161–9
Linz D, Malamuth N M 1993 *Pornography*. Sage, Newbury Park, CA
Malamuth N M 1996 Sexually explicit media, gender differences and evolutionary theory. *Journal of Communication* **46**: 8–31
Malamuth N M, Addison T, Koss M 2000 Pornography and sexual aggression: Are there reliable effects and how can we understand them? *Annual Review of Sex Research* **11**: 26–91
Malamuth N M, Donnerstein E 1984 Correlational and cross-cultural studies on pornography and aggression. In: Malamuth N M, Donnerstein E (eds.) *Pornography and Sexual Aggression*. Academic Press, New York, pp. 139–41
Malamuth N, Impett E A 2000 Research on sex in the media: What do we know about effects on children and adolescents. In: Singer D, Singer J (eds.) *Handbook of Children and the Media*. Sage, Newbury Park, CA, pp. 269–87
Morais R C 1999 Porn goes public. *Forbes Magazine*, 14 June, p. 214
Mosher D L, MacIan P 1994 College men and women respond to X-rated videos intended for male or female audiences: Gender and sexual scripts. *Journal of Sex Research* **31**: 99–113
Oliver M B 2000 The respondent gender gap. In: Zillmann D, Vorderer P (eds.) *Media Entertainment: The Psychology of its Appeal*. Erlbaum, Mahwah, NJ, pp. 215–34
O'Toole L 1999 *Pornocopia: Porn, Sex, Technology and Desire*. Serpent's Tail, London
Perse E M 1994 Uses of erotica and acceptance of rape myths. *Communication Research* **21**: 488–515
Sorokin P 1956 *The American Sex Revolution*. Sargent, Boston
Stauffer J, Frost R 1976 Male and female interest in sexually-oriented magazines. *Journal of Communication* **26**: 25–30
Stoller R 1976 Sexual excitement. *Archives of General Psychiatry* **33**: 899–909
Strossen N 1995 *Defending Pornography: Free Speech, Sex, and the Fight for Women's Rights*. Scribner, New York
Weinstein J 1999 *Hate Speech, Pornography, and the Radical Attack on Free Speech Doctrine*. Greenwood Press, Westport, CT
Wojtyla K (Pope John Paul II) 1981 *Love and Responsibility* [Willetts H trans.]. Farrar, Straus & Giroux, New York
Zak P J, Knack S in press Trust and growth. *Economic Journal*

N. M. Malamuth

Positivism, History of

Positivism is the name of a social and intellectual movement that tried to learn from the mistakes of the Enlightenment project that eventuated, first, in the Reign of Terror following the French Revolution of 1789, and second, in the irrationalism of the Weimar Republic following Germany's defeat in World War I. While it has been customary to distinguish between the quasipolitical movement called 'positivism' originated by Auguste Comte in the 1830s and the more strictly philosophical movement called 'logical positivism' associated with the Vienna Circle of the 1930s, both shared a common sensibility, namely, that the unchecked exercise of reason can have disastrous practical consequences. Thus, both held that reason needs 'foundations' to structure its subsequent development so as not to fall prey to a self-destructive scepticism. In this respect, positivism incorporates a heretofore absent empiricist dimension to the risk-averse orientation to the world historically associated with Platonism.

1. *Positivism's Platonic Political Roots*

In terms of Western intellectual history, positivism inflects Plato's original philosophical motivation through a secularized version of the Christian salvation story, in which Newton functions as the Christ figure. This captures both the spirit of Auguste Comte's original project and its residual effects in twentieth century logical positivism, which dropped the overt historicism of Comte's project, while retaining the fixation on Newton as the model for what it means to express oneself scientifically and a vague belief that greater scientific knowledge will deliver salvation. Indeed, positivism's core conceptual problem has been to define a scientific vanguard capable of both offering guidance to the unenlightened and itself

changing in light of further evidence and reflection. Its Platonic roots are worth recalling.

Plato was affected deeply by the volatile political environment created by the public dialectical encounters of Socrates and the other Sophists who mentored the leaders of his Athenian youth. Thus, after the fall of Athens, Plato was keen to restrict the dialectic to a cloistered setting, only to be unleashed to the larger society once initiates were sufficiently secure in their own knowledge and judgement that they would not be unduly swayed by popular demands. Where Plato had hoped to produce implacable philosopher-kings who would rule as absolute monarchs, positivists typically have envisaged a more differentiated but no less authoritative (authoritarian?) rule by experts, each an oligarch over their domain of knowledge. In this respect, positivism is bureaucracy's philosophical fellow traveler.

Not surprisingly, positivism's relationship to democracy has been chequered. Like Plato, positivists have feared protracted public disagreement most of all and hence have tended to demonize it as 'irrational' and 'noncognitive.' Their image of 'plural' authority presumes nonoverlapping competences, such that legislative questions are reduced to judicial ones concerning the expertise to which one should defer. This is in sharp contrast to most republican conceptions of democracy, which base representation on interest groups, each invested with a competence appropriate to the promotion of its interests but typically about matters in which other interest groups might also have competence. Since matters of public policy typically affect several such groups at once, collective decisions do not turn on identifying the right group whose judgement should hold sway. Rather, it is expected that public debate will issue in a solution that transcends all the groups' starting points but nevertheless manages to serve their respective interests.

Thus, there is a fundamental ambiguity in positivism's appeal to organized reason, or 'science,' in the public sphere. Sometimes this ambiguity is finessed by saying that positivists regard science as the main source of political unity. At the very least, this implies that it is in the interest of all members of society to pursue their ends by scientific means, as that may enable them to economize on effort and hence allow more time for the fruits of their labour to be enjoyed. Ernst Mach (discussed in Sect. 3) comes closest to defending this position in its pure form. It fits comfortably with the libertarian idea that democratic regimes should enable maximum self-empowerment.

However, many positivists have drawn a further conclusion that can thwart this libertarian impulse. From Comte onward, it has been common to argue that science can unify the polity by resolving, containing, or circumventing social conflict. Here a well-established procedure or a decisive set of facts is supposed to replace more 'primitive' and volatile forms of conflict resolution such as warfare and sometimes even open debate—all of which supposedly compromise the integrity of opposing viewpoints in the spirit of expedience. Accordingly, a scientific politics should not merely satisfy the parties concerned: it should arrive at the 'correct' solution.

To be sure, even this mentality admits of a democratic interpretation, as positivist social researchers have been in the forefront of presenting 'data' from parties whose voices are unlikely to be heard in an open assembly. Typically, this has occurred in surveys designed to represent the full empirical range of a target population. Nevertheless, the question remains of exactly who reaps the political benefits of these newly articulated voices: the people under investigation, the investigators themselves, or the investigators' clients? Moreover, once a target population has been registered empirically, do its members remain 'objects of inquiry' or are they promoted to full-fledged inquirers capable of challenging the original investigators' findings and methods? Probably the most sophisticated treatment of these questions in the context of positivistically inspired US social policy research is to be found in Campbell (1988).

These delicate questions arise because ultimately positivism turns Plato on his side by converting a static hierarchy into a temporal order. Where Plato imagined that authority flowed downward from the philosopher-king in a caste-based social structure, positivists have envisaged that all of humanity may pass (at a variable rate) through a sequence of stages that retrace the socioepistemic journey from captivity to autonomy. In the positivist utopia, it is possible for everyone to be an expert over their own domain. Moreover, there is a recipe for the conversion of Platonism to positivism. It proceeds by isolating a domain of inquiry from the contingencies surrounding its manifestations so that its essential nature may be fathomed. Whereas Plato reserved such inquiry to philosopher-kings, positivists have more often turned to state-licensed professional bodies. And instead of Plato's intellectual intuition (*nous*), positivists attempt to gain epistemic access by comparative historical and experimental methods.

This recipe can be illustrated in the work of Otto Neurath, an organizer of the Vienna Circle. He wanted to isolate the essence of the 'war economy' so that its efficient central planning mechanism can be transferred to environments where it would have more socially salutary consequences. Here Neurath anticipated what Alvin Gouldner would call the Janus-faced character of the 'welfare-warfare state,' whereby the same organizational structure (in this case, a concentration of resources in the nation–state) can have radically different consequences, depending on the supporting political environment. Nevertheless, as Neurath's many critics pointed out, positivism seems to have inherited Platonism's political naivete, which confuses the fact that, say, the 'war economy' can be identified analytically as a feature of many societies

and the analyst's ability to transfer it to new social environments—bar ring the imposition of sufficient force to hold all other environmental factors constant. If anything deserves the name of the 'positivist fallacy' it is this too easy assimilation of the forum to the laboratory. The history of positivism can be neatly captured as a Hegelian dialectic, the three moments of which are epitomized by the work of Auguste Comte (thesis), Ernst Mach (antithesis), and the Vienna Circle (synthesis). However, these moments have historically overlapped, occasionally coming together in figures such as Neurath. The career trajectories of positivism's standard bearers help explain the direction taken by their thought.

Comte was an early graduate of the *Ecole Polytechnique* who believed that its Napoleonic mission of rendering research a vehicle for societal transformation had been betrayed, once he himself failed to achieve a permanent academic post. Mach was a politically active physicist on the losing side of so many of the leading scientific debates of his day that his famous chair in Vienna, from which the logical positivists sprang, was awarded on the strength of his critical-historical studies, not his experimental work. Finally, the intellectual leader of the Vienna Circle, Rudolf Carnap, had to abandon physics for philosophy because his doctoral dissertation topic was seen as too 'metatheoretical' for a properly empirical discipline. For Carnap and others who came of age in World War I, physics had devolved into another specialized field of study, rather than—as it had still been for Einstein—natural philosophy pursued by more exact means.

2. *Thesis: Comtean Positivism*

Auguste Comte coined the word 'positivism' to refer to the positive side of the Enlightenment's essentially destructive approach to traditional social and intellectual orders. Comte agreed with the Enlightenment wits that scientific rationality is the model of social rationality, the exemplar of which is Newtonian mechanics. However, according to Comte, the French Revolution had refuted the wits' belief that people's 'natural reason' would come to light once they were freed from their captivity to religion and other traditional forms of authority. Comte thus designed positivism to replace both the church and the sovereign with scientific experts who together would function as a secular priesthood for modern society. 'Sociology,' another of Comte's coinages, referred at once to this political vision and the scientific project that would render it a legitimate form of authority.

Positivism in this basic sense has remained the philosophy underwriting science as a specialized profession of 'scientists' with an identifiable order of merit and career structure. Ironically, none of the movement's first phase protagonists—including Comte himself, John Stuart Mill, and Herbert Spencer—ever called a university home. Nevertheless, they all earned their living from producing knowledge (some would call it 'ideology'), unlike those for whom the pursuit of knowledge was a refined way of consuming wealth. In this important respect, positivism was aligned with the ascendant bourgeoisie and hostile to such aristocratic forms of learning as the gentleman-scientist and the academic humanist. Comte arrived at this position from his mentorship under Count Henri de Saint-Simon, a spiritual godfather of today's management gurus, who at the dawn of the industrial age anticipated that organized knowledge would be the principal factor of economic production and political control in what Daniel Bell would call, 150 years later, 'post-industrial society.'

Given its explicitly anticlerical posture, positivism was more often found in medical schools and polytechnic institutes than traditional universities. (French universities did not become fully secular institutions until the Third Republic in 1870.) The medical interest is noteworthy because, in the nineteenth century, positivism was used to legitimize everything from phrenological and psychosomatic research to the experimental approach to medicine championed by Claude Bernard. What they all shared was the recognition that physical constitution plays a significant role in defining psychic, or spiritual, life. This basic assumption was sufficient to challenge religion's ministerial monopoly.

More generally, positivism's anticlericalism has been among its most powerful export values, especially in Latin America (where Comte's motto 'Order and Progress' remains emblazoned on the Brazilian flag). Without a thoroughgoing Protestant Reformation, Latin Americans found Comte's secularized version of Catholicism much less utopian than did their European counterparts. To be sure, background cultural allegiances divided positivist loyalties, with Francophile Latin Americans (e.g., Brazil) preferring Comte and Anglophile ones (e.g., Argentina) Spencer. In either case, though, positivism was equally likely to justify authoritarian and democratic regimes. Perhaps the most lasting intellectual residue of positivism's Comtean origins has been the default association that social scientists make between reliable knowledge and statistically significant empirical correlations. Hempel (1942), the canonical positivist account of historical explanation, placed the Vienna Circle's distinctive stamp on this point. Nowadays, its import is easily lost in methodological discussions that trivialize correlations as deficient accounts of causation, proof that positivism owes too much to Francis Bacon. However, following Karl Pearson, perhaps the first self-declared positivist to hold a professorship in Britain, the appeal to correlations served at least four distinct functions.

First, correlations enabled an economy of thought by summarizing observations already made. This would decrease the likelihood of 're-inventing the

wheel,' thereby expediting overall human progress. Second, correlations forced investigators to look beyond the singular striking success to more generalizable, albeit mundane, tendencies. Here the positivism helped moderate the hype surrounding new inventions and discoveries. Third, the focus on correlations implied that scientific work could be mechanized by treating individual experiences as instances of more general types. This helped dispel the idea that scientists must have a special 'genius' for finding hidden causes beneath the welter of data. Fourth and probably most significant in Pearson's own turn-of-the-century context, the appeal to correlations echoed the 'uniformitarian' approach to natural history championed by Darwin, whose theory of evolution presupposed that all biological change has occurred by principles that emain in effect today. Thus, there was no need for either a divine creator behind the scenes whose actions transcend observed correlations or a mystification of statistically extraordinary events as 'miracles.'

3. *Antithesis: Machian Positivism*

Ernst Mach symbolized the second moment of positivism's dialectic: he gave the movement a dose of its own medicine, namely, a scientific sense of science's own limitations. Mach's counsel emerged in the midst of the ironic, perhaps perversely self-defeating, character of *fin de siecle* Viennese culture, whereby Karl Kraus's quip that psychoanalysis is the disease for which it is cure was eventually incorporated into positivistic lore as Ludwig Wittgenstein's dark saying that philosophy is the ladder that must be climbed in order to be discarded (Janik and Toulmin 1973).

In the long term, this sense of scientific self-restraint has made its strongest impression in the dual doctrines of 'value neutrality' and 'academic freedom,' which are most closely associated in the social sciences with Max Weber. Although these doctrines normally are treated as separate, it is difficult to motivate either without examining their historical interconnection. The missing link is the idea that, absent further empirical demonstration, the value of theoretical discourse—the most obvious epistemic marker distinguishing experts from the lay public—is purely heuristic. Theory enables those who already know something to know much more; however, it obscures the vision of those who have yet to know anything.

Thus, the various Mach-inspired positivist projects of 'reducing' theoretical discourse to its empirical bases have been motivated by a keen sense of the contingency of when and where new scientific insights arise. The distinction that Hans Reichenbach and Karl Popper drew between the contexts of discovery and justification canonized this point for contemporary philosophy of science. The unique means by which a discovery is made is neither necessary nor sufficient for demonstrating its validity. Elementary logic textbooks demonize the failure to abide by this point as the 'genetic fallacy.' For Mach, the contingent nature of discovery places a special burden on scientists to render their insights as widely as possible, without denying themselves the right to follow it up in their own way. The Vienna Circle positivists may be seen as having compromised this legacy by limiting the context of justification to the received canons of inductive and deductive reasoning. But perhaps they had good reason.

Mach may have been too keen to let the lay public draw whatever conclusions they wished from their studies of science. The second moment of positivism enabled the conversion of scientific knowledge into a pure instrument detached from any prior theoretical or normative commitment. In practice, this 'neutral' stance made science available to the highest bidder, who in turn could hold a monopoly over its use (or nonuse). Consider the fate of the law of diminishing marginal utility (LDMU), which states that if there is enough of a good to satisfy a basic level of need, then each additional increment of the good satisfies the need less than the previous increment. LDMU first appeared in John Stuart Mill's normative political economy, the source of modern welfare economics. The German translation of Mill's *System of Logic* in 1849 had introduced the first phase of positivism to Mach and such contemporaries as Wilhelm Wundt and Wilhelm Dilthey, initiating the long-standing debates over the reducibility of the methods of the *Geisteswissenschaften* (Dilthey's rendering of Mill's 'moral sciences') to those of the *Naturwissenschaften*. While Mill's later work, *On Liberty*, would inspire Mach's own views linking academic freedom to the right of dissent, Mach refused to follow Mill in drawing normative conclusions from putatively scientific facts, even though Mill's aim of removing the accumulated advantage of the rich was close to Mach's own heart.

Mill treated LDMU as a scientific basis for redistributionist social policies. He inferred that each additional increment of income earned by the rich would improve their lives less than the same increment transferred to the poor. Mill presumed a common standard of interpersonal utility and the idea that different interpersonal endowments could be explained as reversible historical accidents. Moreover, he modeled LDMU on a certain interpretation of Newton's laws, which holds them to be true in an ideal physical medium but not in actual physical reality. Thus, Mill made the ideality of the scientific laboratory stand for the normative basis of social reform. What is now called the 'marginalist revolution' in economics reinterpreted the Newtonian precedent for LDMU in the value-neutral terms that captured the social situation of Mach and other second generation positivists who called a university home.

In the 1870s, Mill's academic nemesis, William Stanley Jevons, had begun interpreting LDMU as the

emergent product of interactions between agents whose respective utilities are in principle incalculable, that is, an empirically stable resolution of unknowable constitutive forces. Today we would say that LDMU is a macrolevel market effect, in which the parameters of the relevant market are left unspecified. Thus, cases that fail to conform to LDMU do not force a difficult normative judgement about whether the law or the case is at fault; rather, one casts around for an appropriately defined market to which LDMU might apply, since one may not have considered all the effects over a large enough scale, over a long enough period.

Second generation positivists came to regard LDMU as holding no clear implications for social policy. This opened the door to the curious twentieth century 'noninterventionist' policies that Milton Friedman and others have used to oppose scientific justifications of the welfare state. Here positivism's antitheoretical and antinormative stance is explicitly identified with the inductively self-organizing order of the so-called invisible hand. The shift from Jevons to Friedman is indicative of what logicians call the 'modal fallacy': The denial that X implies Y is interpreted to mean that X implies not Y, where X = LDMU; Y = State-managed redistribution of income. Thus, an economic principle that was seen originally as not necessarily licensing a particular policy intervention came to be seen as an outcome that obviates the need for any policy intervention whatsoever.

Nevertheless, the transition between the first two moments of positivism was neither smooth nor clearly separated from the third. Free market liberals continued to invoke Comte's original sense of positivism to stigmatize large-scale, usually Marx-inspired, social planning of the economy (e.g., Hayek 1952). A target of these studies was Vienna Circle organizer, Otto Neurath, who supposed that a specialist understanding of the persistent underlying structures of political economy provides a better handle on social policy than the spontaneously aggregated experiences of individual producers and consumers. While Neurath's Marxist orientation to social planning was close to Comte's privileging of expert over lay judgement, he also accepted the Machian view that useful knowledge should be spread as widely as possible. In that sense, Neurath's interest in designing an ideographic language ('Isotype') to enable the populace to understand capitalism's inequities was born of the same impulse as his debates with Rudolf Carnap over the construction of the 'protocol statements' that constitute the fundamental language of science. Here the second moment of positivism naturally shades into the third.

4. *Synthesis: The Vienna Circle*

The Vienna Circle was started by Moritz Schlick, another physics-trained philosopher who succeeded Mach to the chair of the history of inductive sciences at the University of Vienna. Much has been made of the Vienna Circle's preoccupations with the conceptual and empirical foundations of science, the demarcation of science from nonscience, and the unification of the special sciences under a single metaphysical and/or methodological rubric. Less has been said to explain the Circle's partial return to positivism's communitarian and authoritarian origins, especially after Mach had made positivism virtually synonymous with the liberalization of epistemic authority. Nevertheless, there were few true liberals in the Vienna Circle: Schlick and Wittgenstein were conservatives, Carnap and Neurath socialists. Popper, the Circle associate closest to Mach's own politics, was not easily accepted into the fold.

This antiliberal turn can be explained by changes in science's social relations in the German-speaking world over the first quarter of the twentieth century. Seen against this background alone, the Vienna Circle appears utopian and even prone to nostalgia; but with Nazism's ascendancy in the 1930s, the 'logical positivist' movement quickly acquired a defiant and progressive cast. At first, the Circle recoiled from Machian positivism because of the ease with which Mach's neutral instrumentalism allowed scientists, especially chemists, to harness their knowledge for aggressive military–industrial ends. Consequently, Germany's humiliating defeat in World War I paved the way for the antiscientific backlash of Weimar culture that formed the backdrop to the Circle's meetings. In recognizably Platonic fashion, the Circle forged a philosophy that protected science from this backlash, while divesting it of any potentially corrupting links to technology. The key was the idea that a genuine science must satisfy certain logical and empirical criteria relating to the public testability of knowledge claims. The model for this requirement was ongoing debates over the foundations of mathematics, which Carnap and Reichenbach had already transferred to debates over the presuppositions of relativity and quantum theory.

However, one decidedly anti-Platonic insight that the Vienna Circle retained from Mach, which increased in importance over the years, was that esoteric forms of knowledge must be 'reducible' to their core 'cognitive' components so as to facilitate both their evaluation and utilization. Specifically, the Vienna Circle aimed to demystify 'metaphysical' jargon that encoded prejudice without extending insight. (Carnap famously cited Heidegger in this respect.) At the peak of its popularity in the 1940s, logical positivism was linked with such American best sellers of the 'general semantics movement' as Hayakawa (1978). Here the foe had shifted from Heidegger to Hitler, with the positivists arguing against the irrationalism of mass propaganda. Perhaps the most famous single work of the positivist movement, Ayer (1936), written by the British observer at the Vienna Circle, can be seen as a transitional work between these two phases.

Given this history, logical positivism appears in a rather ironic light in the highly publicized *Positivismusstreit* that transpired in 1961–8 between Karl Popper and Frankfurt School doyen, Theodor Adorno. It turns out that both Popper and Adorno demonized 'positivism' in social science for the exact same ills as the logical positivists had demonized Heidegger and Hitler a quarter-century earlier, namely, a 'jargon of authenticity' and its contribution to an 'authoritarian personality' (two relevant Adornian phrases). Moreover, the two antagonists presupposed the logical positivists' ideal of unified inquiry and the inadequacy of defining society as simply the totality of 'social facts.' Where Popper and Adorno clearly differed from the positivists was over the latter's promotion of a sharp fact-value distinction, whereby the presumed unity of factual knowledge offsets the irreducible plurality of value orientations. Here positivism courted a 'relativism' that both antagonists strongly opposed. Nevertheless, the ascendancy of relativism, in the guise of 'postmodernism,' would render Popper and Adorno's wranglings over 'critical reason' beside the point by the 1980s.

In detecting positivism's latent relativism, Popper and Adorno not only anticipated the future; they also recalled the 'value neutrality' that had originally turned Machian positivism into a vehicle of power politics. In the social sciences, Max Weber's distinction between charisma and bureaucracy as modes of legitimation enabled both politicians and scientists to turn a blind eye to each other's activities, as the two groups were supposedly driven by radically different normative orientations: the politician acts, while the scientist provides reasons. Similar criticisms have been made of other movements in the social sciences affiliated with positivism, most notably Hans Kelsen's legal positivism and B. F. Skinner's radical behaviorism. Yet, even Neurath was happy to call his project 'neutral Marxism.'

These cases are united by a refusal to blur the boundary between science and politics, which in practice has meant a rejection of social democratic, welfare-state policies in favor of more extreme political stances: either complete withdrawal or outright authoritarianism. Once again, positivism would seem to have fallen victim to the modal fallacy: because politics and science can contaminate each other, therefore they inevitably will, and hence each must be left to its own devices.

5. *Final Assessment: Does Positivism Have a Future?*

After the leading members of the Vienna Circle migrated to the US in the 1930s, logical positivism seeded that country's analytic philosophy establishment for the second half of the twentieth century. However, this is the only context in which positivism possibly dominated an established discipline. For the most part, positivism has been embraced by disciplines that have yet to achieve academic respectability, even in the natural sciences, where Mach found his strongest support among chemists, biomedical scientists, and psychologists—not physicists. (It is often overlooked that positivism's reliance on Newtonian mechanics as the model for all science generally was not appreciated by a physics community jealous to guard its guild privileges.) Unsurprisingly, positivism's most ardent supporters have been social scientists, not for the Comtean reason that sociology is the pinnacle of all science but for the more mundane reason that positivism seemed to offer a strategy for rendering one's activities 'scientific.'

These matters came to a head with the publication of Kuhn (1970), which Carnap enthusiastically endorsed as the final instalment of the logical positivists' *International Encyclopedia of Unified Science*. It provided an account of scientific change based largely on the history of physics that was quickly embraced by social scientists, as Kuhn stressed science's self-organization over its larger societal impact. However, unlike previous positivist accounts, Kuhn's explicitly fixated on particular scientific disciplines (or 'paradigms') without presuming that science as a whole is heading toward a unified understanding of reality. By implication, then, all scientific standards are discipline-relative. Kuhn's approach suited what is now called the 'postmodern condition,' which the *Positivismusstreit* had desperately tried to stave off.

In retrospect, the popularity of Kuhn (1970) is better understood as signifying positivism's decadent phase than, as it was originally seen, a fundamental challenge to positivism. The decadent state of contemporary positivism can be gleaned from the ongoing debates over the scientific status of 'sociobiology.' Each side of the debate captures part of the positivist legacy: proponents like E. O. Wilson hark back to positivism's Comtean vision of unified knowledge in aid of societal transformation, while opponents like Richard Lewontin draw on positivism's more recent restriction of the bounds of science to discipline-specific criteria.

If positivism has a future, it lies in rekindling a sense of 'Science' that transcends the boundaries of particular scientific disciplines. This is how Comte originally thought about the discipline he called 'sociology.' He claimed that it was the last to develop, not simply because of the complexity of its human subject matter, but more importantly, sociology had to reconstitute the (natural) sciences that historically preceded it. Too often the history of positivism's quest for unified science has been interpreted as exclusively a matter of applying the methods of physics to less developed sciences. The reciprocal movement is really more important, namely, the application of sociological findings to the future direction of science as a whole. Such interdisciplinary projects as 'social epis-

temology' and the 'science of science' have tried to fulfill this side of the Comtean promise, which may be redeemed once the present intellectual climate changes.

See also: Behaviorism; Comte, Auguste (1798–1857); Critical Theory: Frankfurt School; Durkheim, Emile (1858–1917); Empiricism, History of; Enlightenment; Kuhn, Thomas S (1922–96); Logical Positivism and Logical Empiricism; Mill, John Stuart (1806–73); Popper, Karl Raimund (1902–94); Positivism: Sociological; Wittgenstein, Ludwig (1889–1951)

Bibliography

Adorno T (ed.) 1976 *The Positivist Dispute in German Sociology* (trans. Adey G, Frisby D). Heinemann, London

Ayer A J 1936 *Language, Truth and Logic*. Gollancz, London

Campbell D T 1988 *Methodology and Epistemology for Social Science: Selected Papers*. University of Chicago Press, Chicago

Cartwright N, Cat J, Fleck L, Uebel T 1996 *Otto Neurath: Philosophy between Science and Politics*. Cambridge University Press, Cambridge, UK

Fuller S 1993 *Philosophy of Science and Its Discontents*, 2nd edn. (orig. 1989). Guilford Press, New York

Fuller S 2000 *Thomas Kuhn: A Philosophical History for our Times*. University of Chicago Press, Chicago

Hayakawa S I 1978 *Language in Thought and Action*, 4th edn. (orig. 1939). Harcourt Brace and Jovanovich, New York

Hayek F 1952 *The Counter-Revolution of Science: Studies on the Abuse of Reason*. University of Chicago Press, Chicago

Hempel C G 1942 The function of general laws in history. *Journal of Philosophy* **39**: 35–48

Janik A, Toulmin S 1973 *Wittgenstein's Vienna*. Simon and Schuster, New York

Kolakowski L 1972 *Positivist Philosophy From Hume to the Vienna Circle* (trans Guterman N). Penguin, Harmondsworth, UK

Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. (orig. 1962). University of Chicago Press, Chicago

Proctor R 1991 *Value-Free Science? Purity and Power in Modern Knowledge*. Harvard University Press, Cambridge, MA

Turner S P 1986 *The Search for a Methodology of Social Science*. Kluwer, Dordrecht, Netherlands

Zea L 1963 *The Latin American Mind* (trans. Abbott J A, Dunham L). University of Oklahoma Press, Norman, OK

S. Fuller

Positivism: Sociological

Positivism arose in the social sciences in an effort to legitimate sociology as a scientific discipline. In essence, positivism asserts that the social sciences, and sociology in particular, are natural sciences in which abstract laws of human organization can be formulated to explain the operative dynamics of the social universe. The plausibility of these laws are then to be assessed against systematically collected empirical data.

1. The Rise of Positivism

Advocacy for a positivistic social science began with Auguste Comte's (1830–42) *Course of Positive Philosophy*. Comte was, no doubt, influenced by Montesquieu's ([1748] 1900) *The Spirit of Laws*, where the notion of formulating laws about the operation of human society was first suggested, and by both Turgot and Condorcet, from whom Comte received, via his mentor, Saint-Simon, the notion of the three stages in the evolution of idea systems. For Comte, the various sciences developed at different rates, with sociology (and by implication the other social sciences as well) emerging last. As each science develops it goes through three stages—theological, metaphysical, and eventually, positivistic. By the early 1800s, the idea that sociology and other social sciences could take their place at the table of science was widespread, but it was Comte who forged the various lines of argument into a forceful advocacy for a science of society. Comte's preferred name for sociology was *social physics*, and this name demonstrates his desire to develop a sociology organized around basic laws of the social universe.

In *Course of Positive Philosophy*, Comte posits a view of sociology as a theory-driven enterprise. As a descendant of the French Enlightenment, and as one who admired greatly what Newton had done for astrophysics, Comte argued for a sociology in which abstract laws of human organization would explain the dynamics of the social universe. His exemplar was Newton's law of gravity, and yet he never developed any such law—although he saw the 'law of the three stages' as something like Newton's law of gravity. Still, from the opening pages of *Positive Philosophy*, a clear message is sent: 'The first characteristic of Positive Philosophy is that it regards all phenomena as subject to invariable natural *Laws*' (1830, pp. 5–6, emphasis in original). In this same passage, Comte argues against a search for '*Causes*, whether first or final'; rather the goal of positivism is the 'discovery of laws, with a view to reducing them to the smallest possible number.' In later portions of *Positive Philosophy*, Comte (1830–42) buttressed this emphasis on abstract laws by criticizing raw empiricism, for 'the next great hindrance to the use of observation is the empiricism which is introduced by those who, in the name of impartiality, would interdict the use of any theory whatsoever.' Comte concludes this passage with an even stronger statement: 'No real observation of any kind of phenomena is possible, except in as far

as it is first directed, and finally, interpreted, by some theory.'

Thus, in Comte's view, all observation is mediated through a theory, and the goal of positivism is to make the theories explicit and formal, while employing systematic methods to gather data on the plausibility of a theory. He advocated four basic methods: observation, experimentation, comparison, and historical analysis in order to collect the broadest array of data possible to assess the validity of laws that are presumed to explain the operation of the social universe in all times and places.

By 1830, then, Comte had laid out in bold and incisive fashion the basic tenets of positivism. As Comte's mental instability caused his star to fade, positivism was carried forward by Herbert Spencer in England and Emile Durkheim in France. The titles of almost all of Spencer's important works indicate the positivistic slant of his work in both the natural and social sciences: *First Principles*, and *The Principles of* (sequentially) *Psychology*, *Biology*, and *Sociology*. But nowhere was Spencer's advocacy more powerful than in his first published sociological work, *The Study of Sociology* (1873), where he not only reaffirms his commitment to seeing the universe as guided by forces whose dynamics can be captured by abstract principles, but also articulates in a systematic manner the sources of bias that can distort the formulation of these principles and their systematic assessment by data. Twenty-two years later, Emile Durkheim ([1895] 1938) made a similar plea in his *The Rules of Sociological Method* for a sociology based upon laws about 'social facts.' Both Spencer and Durkhein personified how positivistic sociology could be conducted, mixing sophisticated functional theory with presentations of data. Indeed, from Comte's early analogies to organisms, positivism was associated initially with functional sociology, a theoretical point of view that was to flower in the middle of the twentieth century, and then to die on the vine in the latter part of the century. And as functionalism faded, so did positivism as a coherent point of advocacy. Yet, much of the fate of positivism in the twentieth century is associated with intellectual ferment in Europe around what was to become known as the Vienna Circle.

2. *The Vienna Circle and the Transformation of Positivism*

During the first half of the twentieth century, a group of intellectuals located mostly in and around Vienna debated issues of thought and reasoning as well as connections between the empirical world and modes of representing this world (Johnson 1983). Two outside and somewhat marginal figures of the Vienna Circle—Ernst Mach (1893) and Karl Popper (1959, 1969)—are critical to the debate that raged within the Circle for the first half of the twentieth century. Mach had proposed a radical empiricism in which all conceptions of the universe as guided by natural laws were rejected in favor of mathematical descriptions of relations among empirically observable variables. This extreme position called for a radical empiricism, emphasizing that formal thought must always be tied to, and inducted from, observable empirical facts. At the other pole were those concerned with the structure and logic of formal thinking to either the exclusion or subordination of empirical observations to rules of logic. Karl Popper's famous solution to this impasse was to emphasize that theories can never be proven by empirical facts, only disproven. The goal of science is, therefore, to assess the empirical plausibility of hypotheses deduced formally from abstract principles by repeated assaults from empirical observations. For many decades, this argument became the new standard way to assert the viability of positivism; and it led to the notion of 'logical positivism' in which emphasis is on the logic of deduction from abstract principles to testable hypotheses. Comte had never conceptualized how one makes abstract laws available for empirical assessment, and so this point of emphasis was perhaps required in a more complete formulation of positivistic science.

3. *Vienna Circle Debates and US Sociology*

The Vienna Circle had much more influence on US social science than on disciplines in Europe where scholars were, in general, more philosophically sophisticated than their US counterparts. But much more was involved; US social science adopted Popper's argument because these Vienna-inspired arguments were compatible with the new social science emerging in the USA.

If we look at US sociology of the 1920s and 1930s, there was a vacuum of theory. Spencer's star had faded, and other figures in sociology's European pantheon—Weber, Durkheim, Simmel, and Pareto—remained largely untranslated and, hence, unknown to many US sociologists. And yet, US sociology had been very much committed to Comte's view of science, and many early texts refer to 'principles' of various subjects in a manner reminiscent of Spencer. At the same time, quantitative statistical methods were being adopted by US sociologists under the influence of Franklin Giddings (1901), and later his students and followers. Sociologists were thus very much interested in finding a legitimating mantra for statistical analysis of data that could make claims to being scientific (Turner and Turner 1990). One mantra was quantification, *per se*, but many early sociologists had a more Comtean–Spencerian view of science as a search for laws. Popper's solution provided the reconciliation of the debate within not

only the Vienna Circle but also within conflicting camps of US sociology as well: theories were to be tested against quantitatively arrayed data using new statistical techniques borrowed from the European agronomist, Karl Pearson. Through much of the mid-century, the theory part of this reconciliation took a back seat to quantification. And, not surprisingly, it is from this period that the misperception of positivism as 'raw empiricism' becomes prominent, despite the fact that this label is wholly inappropriate to positivist epistemology.

The radical empiricism of the Vienna Circle found a comfortable home in American quantitative research in the last quarter of the century, and for a brief time, the logical part of positivism was trumpeted through the effort to 'build' and 'construct' theories. A host of books on 'theory building' and 'theory construction' emerged in the late 1960s, beginning with Hans Zetterberg's (1965) *On Theory and Verification in Sociology*, and continuing through the 1960s and 1970s (e.g., Blalock 1969, Dubin 1969, Gibbs 1972, Hage 1972, Reynolds 1971). These books attracted great interest in an effort to overcome the radical empiricism of quantitative research, since they tried to reconnect methods and theory along the lines suggested by Popper. The problem was that these books had a cookbook and mechanical character, in essence trying to do for theorizing what standardized quantitative protocols had done for quantitative methods. As a result, they soon fell into disuse—at about the same time that functionalism, which had been allied with positivism since its origins, was also going out of style. The result was that as functionalism faded and as the rather mechanical and restrictive character of theory construction books was exposed, positivism as Comte had conceived of it also declined. Indeed, the very term 'positivism' is now seen as anachronistic or as an epithet for raw empiricism.

4. *The Critics of Positivism*

Beyond inviting the label of 'raw empiricist,' proclaiming oneself a positivist in the contemporary era is to invite an additional series of attacks on scientific sociology in general and, more specifically, on the prospects for formulating universal laws on the dynamics of the social universe. Some of the most common lines of criticism are enumerated below (Turner 1992).

4.1 *The Phenomenological Critique*

The world 'out there' can only be known through human consciousness, and so the nature of reality is ultimately filtered through the human mind. Hence, laws of human organization cannot be formulated without a prior understanding of the dynamics of the human mind.

4.2 *Hermeneutics*

The fundamental nature of the social world can be changed by human agency. Hence, since the nature of social organization is subject to change in its most fundamental character, universal laws about generic processes are, at best, time bound and, at worst, the illusions of social theorists. Moreover, even if a law of social organization is articulated, it soon becomes a tool for changing the fundamental processes described by the law, thereby obviating the validity of the law.

4.3 *Critical Discourse*

Any set of laws describing the social universe operates as a legitimating umbrella for current social arrangements, many of which are oppressive. Thus, science legitimates the status quo and closes off discourse on less oppressive alternatives.

4.4 *Historical Particularism*

Social reality is historically specific, reflecting particular confluences of events at a particular moment in time. Hence, any effort to generalize beyond a given historical epoch is doomed to failure, since history constantly remakes the very nature of social reality. The most recent version of this kind of argument can be found in postmodernism, where it is postulated that a fundamentally new reality has emerged, obviating the laws of modernist social organization.

4.5 *Scientific Politics*

Since science is an organized activity, the knowledge produced by scientists reflects organizational, professional, and political processes as much as reality 'out there.' Thomas Kuhn's (1970) famous work reinvigorated what had previously been known as 'the sociology of knowledge'—an approach revolving around analyses of the social context of knowledge production. Scientific knowledge is, therefore, distorted by the very social processes that organize scientists.

5. *The Defense of Positivism*

The five lines of criticism listed above make interesting points, and then take them to the extreme. But they represent an *ad hoc* coalition of positions that oppose positivism, seeing positivism as a 'failed epistemology' in the social sciences. Others (e.g., Turner 1985) have sought to defend positivism, addressing the specifics of these points of critique, while trying to re-establish the search of universal and abstract laws about generic social process.

The phenomenological and scientific politics critiques are undone by the fact that the same problems exist for the physical and biological sciences; and yet scientists in these areas still manage to generate verifiable knowledge. The hermeneutical critique poses a unique challenge to scientific positivism in the social sciences, but it is merely assumed by those making this critique that the universe is constantly changing its fundamental character. To make their point stick, critics have to demonstrate that the fundamental forces governing the social world have changed; to assert this to be the case without demonstration with data makes the critique less convincing. The historical particularism critique encounters similar problems, for while it is true that specific events and processes are the product of a particular configuration of events, the underlying dynamics of specific historical conditions may still be generic processes of the social universe. Every specific event is the product of a causally unique historical chain of causes—such is the case in both social and physical universes—but this fact does not mean that the forces that move this chain of causes along are unique historically. Indeed, positivists would argue that these forces are universal, operating in all times and places.

The critical discourse critique is perhaps the most important because, to the extent that generalizations from current empirical reality are viewed as 'laws' of how things are, these statements summarizing current empirical reality will assert implicitly that the status quo is the way things *must* be. But the critical theorists run into the opposite problem of postulating ideal states that must be pursued, even if they violate basic laws of human organization and, as a consequence, positively harm people. If ideology were informed by scientific laws, then it is the theory more than the ideology that is useful in reconstructing the social universe in humane ways.

6. *The Prospects of Positivism*

Positivism will perhaps never be accepted uncritically in sociology, and perhaps not in anthropology and even some branches of psychology and political science. In other social sciences, especially economics, the basic arguments of positivists are not even subject to debate. Positivism was a unique advocacy created by Comte to legitimate the scientific study of human organization by a discipline reluctantly labeled 'sociology,' but his advocacy was never accepted by all in the nineteenth or twentieth centuries, and it is unlikely to be accepted in the twenty-first century. Indeed, the term positivism might be best phased out, since it appears to raise such a bright red flag to an *ad hoc* coalition of positions critical of the scientific pretensions of some social sciences. In a sense, theoretical and empirical work dedicated to discovering and assessing general laws has been relegated to the status of yet another specialty in some social sciences, although one still can find positivistic elements within virtually all specialties of social science. Thus, there is some hope that science will persist in those disciplines, particularly sociology and anthropology, where doubts about science are greatest, but positivism in any guise will never be the only epistemology within sociology and anthropology.

See also: Comte, Auguste (1798–1857); Durkheim, Emile (1858–1917); Empiricism, History of; Hempel, Carl Gustav (1905–97); Hermeneutics, History of; Logical Positivism and Logical Empiricism; Popper, Karl Raimund (1902–94); Positivism, History of; Sociology, Epistemology of; Sociology: Overview; Spencer, Herbert (1820–1903); Theory: Sociological; Wittgenstein, Ludwig (1889–1951)

Bibliography

Blalock H M Jr. 1969 *Theory Construction: From Verbal to Mathematical Formulations*. Prentice Hall, Englewood Cliffs, NJ

Comte A 1830–42 *Cours de philosophie positive: Les préliminaires géneraux et la philosophie mathématique*. Bachelier, Paris

Dubin R 1969 *Theory Building*. Free Press, New York

Durkheim E [1895] 1938 *The Rules of Sociological Method*. Free Press, New York

Gibbs J 1972 *Sociological Theory Construction*. Dryden Press, Hinsdale, IL

Giddings F 1901 *Inductive Sociology*. Macmillan, New York

Hage J 1972 *Techniques and Problems of Theory Construction in Sociology*. Wiley, New York

Johnson W M 1983 *The Austrian Mind: An Intellectual and Social History, 1848–1938*. University of California Press, Berkeley, CA

Kuhn T 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

Mach E 1893 *The Science of Mechanics* [trans. McCormack T J]. Open Court, La Salle, IL

Montesquieu C [1748] 1900 *The Spirit of Laws*. Colonial, London, 2 Vols.

Popper K 1959 *The Logic of Scientific Discovery*. Hutchinson, London

Popper K 1969 *Conjectures and Refutations*. Routledge & Kegan Paul, London

Reynolds P D 1971 *A Primer in Theory Construction*. Bobbs-Merrill, Indianapolis, IN

Spencer H 1873 *The Study of Sociology*. Routledge & Kegan Paul, Boston

Turner J H 1985 In defense of positivism. *Sociological Theory* **3**: 24–30

Turner J H 1992 The promise of positivism. In: Seidman S, Wagner D G (eds.) *Postmodernism and Social Theory*. Blackwell, Oxford, UK, pp. 75–94

Turner S P, Turner J H 1990 *The Impossible Science: An Institutional Analysis of American Sociology*. Sage, Newbury Park, CA

Zetterberg H L 1965 *On Theory and Verification in Sociology*, 3rd edn. Bedminster Press, New York

J. H. Turner

Possession (Linguistic)

In linguistics, possession is conceived as a relationship of appurtenance between a possessor (P'or) and a possessed (P'ed, possessum, possessee) for which there are specific paradigms of expression in languages of the world. On the conceptual side, this relationship of appurtenance ranges from material possession via kinship and body-part relationships to the domain of abstract relations (someone's ideas, someone's activities, etc.). On the formal side, one distinguishes mainly between attributive possession (e.g., 'my house,' 'the man's house') and predicative possession ('the house is mine,' 'the house belongs to me,' 'I have a house'). However, it is not always easy to distinguish the conceptual domain of possession from other cognitive domains, and also on the formal side there is a continuum of expression rather than some strictly defined set of constructions. For instance, in many languages (e.g., in English or Icelandic) constructions like 'I have ploughed an acre' (expressing an event in time) historically go back to possessive expressions whose literal meaning was 'I have an acre as something ploughed,' and there are also numerous ways of crossing the boundary between attributive and predicative sentential constructions. For instance, many languages many alternate between attributive constructions like German *der Sohn des Mannes* (ART: NOM son: NOM ART: GEN man: GEN, 'the man's son'; where ART = article, GEN = genitive, NOM = nominative) and *dem Mann sein Sohn* (ART: DAT man: DAT his: NOM son: NOM, lit. '(to) the main his son'), where the dative (DAT) was originally some indirect object in predications like *sie entführten dem Mann seinen Sohn* (they: NOM kidnapped ART: DAT man: DAT his: ACC son: ACC, lit. 'they kidnapped (to) the man his son'; where ACC = accusative). These latter constructions (which are typically associated with some psychological involvement on the part of the P'or) are known as *dativus sympaticus*-constructions, or are referred to in the context of possessor raising or external possession. Since psychological involvement also correlates with the relative 'closeness' and inalienability' of the possessive relation, it is customary to distinguish between inalienable and alienable possession. Alienable possession normally refers to material belongings as opposed to kinship and body-part relationships, and there are many languages which have standardized ways of distinguishing these domains. In addition to expressing a concept which is not just relevant for linguistics, possessive categories also have specific functions in the discourse, which suggests that to a certain extent linguistic possession must be differentiated from non-linguistic concepts of possession.

1. Approaches to Possession

1.1 The Literature on Possession

The linguistic literature on possession concentrates on different aspects. Seiler's monograph (Seiler 1983) is still the most comprehensive attempt at dealing with the cross-linguistic evidence for the continuum of constructions in the domain of possession (see Sect. 1.2 below). Heine (1997) focuses mainly on predicative possession and on typical ways in which such constructions develop in the process of grammaticalization. Attributive possession is usually extensively described in traditional grammars under the headings of possessive pronouns and genitival constructions. McGregor and Chappell (1996) is a standard reference to the typology of inalienable possession, and Payne and Barshi (1999) focus on the typology of external possession. Other publications include discussions on the grammatical status of specific items from the domain of possession. For instance, it is sometimes unclear whether possessive pronouns behave like genitival arguments or like adjectives. The discussions in this area are rather technical and must be seen in the context of DP-analysis (for reference see Abney 1987, Löbel 1995). In what follows, this article will focus on some important issues from the typology of possessive constructions.

1.2 Details on Typological Research on Possession

The most difficult problem for the typological description of possession is the great variety of expression and the fact that there is no simple correspondence between a conceptual relationship of possession and traditional linguistic categories like possessive pronouns or possessive verbs (in fact, possessive verbs are cross-linguistically rather rare). For instance, a Latin construction like *mihi est liber* (me: DAT is book: NOM, lit. 'to me is (a) book') contains no 'classical' possessive category, but conceptually it figures as a possessive construction. Conversely, a construction like 'my train' in 'I've got to catch my train' does not really refer to the train which I own, though it does involve a 'classical' possessive pronoun. Finally, there are instances of so-called 'possessor deletion' as in the German *er hob die Hand* (lit. 'he raised the hand') vs. the English 'he raised his hand,' which unlike the German example contains an explicit possessive pronoun.

As a solution to these problems, Seiler (1983) proposes that the expression of possession constitutes an operational program ranging from minimal

explicitness of the relationship (indicativity) to maximal explicitness (predicativity). Indicativity literally means to 'point' at what one may take for granted, while predicativity establishes a relationship by overt means. The 'indicative' and less explicit techniques are mainly characteristic of the attributive pole (with constructions like 'my train'), while the most predicative type of construction is found in the area of possessive verbs (with constructions like 'I own a house'). For 'my train' the relationship is still very vague, while 'I own a house' is very specific about the relationship in question.

In this context, it is no coincidence that inalienable, inherent possessive relationships usually need less 'effort' to be encoded than alienable relationships. For example, in the Amerindian language Cahuilla the expression 'his eye' would just involve the direct juxtaposition of a personal pronoun and the lexical stem for 'eye,' while the expression 'his dog' needs the addition of a 'possessive classifier' meaning 'pet': compare *he-puc* (3.SG-eye, 'his eye'; where SG = singular) vs. *he-?as ?awal* (3.SG-pet dog, lit. 'his pet dog'). In other words, the (inalienably conceived) body-part term 'eye' is inherently relational, such that the possessive relation between 'he' and 'eye' is more or less 'self-understood'; for 'dog' the situation is different, and the relationship between 'he' and 'dog' must be established by some relational item like 'pet.' By way of contrast, one would need far more effort in Cahuilla to use a word like 'eye' in a 'derelationalized' sense. Thus, while 'the dog' is simply *?awal*, 'the eye' (in some abstract sense) would be *puc-il* (eye-DER; where DER = derationalizer). Seiler lists a number of techniques commonly employed for the expression of possessive relationships, which can all be placed on a continuum between maximal implicitness and maximal explicitness of the relation. These techniques include plain juxtaposition of P'or and P'ed ('N-N-constructions,' see above), connectives (CONN), possessive classifiers (possessive) cases, existential copula constructions, and possessive verbs. The Persian *iza:fa*-construction would be an example for connectives (cf. *ha:ne-je pedar* (house-CONN father, 'the father's house')), and for possessive classifiers see Cahuilla. Latin *domus patris* (house: NOM father: GEN) is an instance of a case construction. A case construction is a more 'explicit' technique than the connective technique in the sense that so-called 'cases' may allow far more oppositions than all-purpose connectives, etc.; thus, apart from using a possessive genitive, many languages also employ locatives in attributive possession (cf. Modern Icelandic *höfudid á mér* (lit. 'the head on me,' 'my head') which is fairly explicit way of specifying how the body-part relates to the P'or). Case constructions also include expressions like the German *der Mann mit dem Hut* ('the man with the hat'). Here the relationship of possession is expressed from a different perspective than *der Hut von dem Mann* (lit. 'the hat from the man,' 'the man's hat'); in the former construction, the nucleus is the possessor, while in the latter the nucleus is the possessed. (For the change of perspective also see Sect. 1.3 below.)

With the help of existential copulas, the case constructions can often be turned into literally predicative expressions of possession (cf. the German *der Hut is von dem Mann* (lit. 'the hat is from the man,' 'the hat belongs to the man') and Icelandic *ég er med bíl* (lit. 'I am with (a) car,' 'I have a car with me'). Finally, languages may also employ actual verbs of possession as in 'I have or own a book' and 'the book belongs to me,' etc. This technique allows the greatest number of oppositions (the number of lexical verbs is virtually limitless), and since this technique is the most predicative one in the sense of Seiler, it is hardly 'natural' in the context of inalienable and 'self-understood' possession. For example, clauses like 'I have a father' or 'I have a head' are considered deviant, unless modifications are added (cf. 'I have a nice father,' 'I have a swollen head'). Nevertheless, though in principle inalienable possession tends towards the 'indicative' pole, and alienable possession tends towards the 'predicative' pole, there are always possibilities of making the implicit explicit and the explicit rather implicit. For instance, Indo-European possessive pronouns may go back to locative constructions of much the same type as the ones which are now commonly used for attributive body-part possession in Icelandic (see above), and the Icelandic attributive constructions in question like *höfudid á mér* (lit. 'the head on me') originate in turn from predicative *dativus sympaticus*-constructions like 'he hit (on) me on the head.' Therefore, there appears to be some shift from the more explicit and predicative to the less explicit in the process of grammaticalization. At the same time, as a reaction to a decrease of explicitness, new explicit constructions may be coined. In this context, Seiler claims that indicativity and predicativity represent complementary principles which are in competition with each other in the operation of giving expression to conceptual possession.

Heine (1997), too, is interested in grammaticalization channels, but he concentrates on the lexical sources of (mostly non-attributive) possession and on the ultimate destiny of possessive items in other domains of grammar. As for non-attributive possession ('predicative possession' in a more restricted sense than in Seiler's framework), Heine (1997, p. 47) presents the following typology based on various conceptual 'schemata': 'X takes Y' ('action schema'), 'Y is located at X' ('location'), 'X is with Y' ('companion'), 'X's Y exists' ('genitive'), 'Y exists for/to X' ('goal'), 'Y exists from X' ('source'), 'as for X, Y exists' ('topic'), 'Y is X's (property)' ('equation'). As an example for the action schema Heine mentions Portuguese, while Turkish has a location schema. Somali has a companion schema, while the genitive schema is found again in Turkish. The goal schema is

manifested in Tamil and Hungarian, the source schema manifests itself in Ik (Nilo-Saharan), Cahuilla has a topic schema, and Russian has an equation schema. In many of these languages different schemata coexist. Again, there is no clear separation between possession and very many other kinds of conceptual domains.

For the different schematic types mentioned it is possible to identify very common paths of development and transitions to other categories. Especially postural verbs (e.g., 'stand,' etc.) often develop into locative copulas, and the resulting location schema ('X is located at Y') may develop into a companion schema ('X is with Y'); both the location schema as well as the companion schema then tend to be re-employed as expressions of possession (for companion cases ('be with') cf. Icelandic *ég er med bíl* (see above), and for location cf. Finnish *Liisalla on mies* (lit. 'at Liisa is (a) husband,' 'Liisa has a husband')). The same construction types also often grammaticalize to expressions of aspect (Swahili has a 'be with'-progressive, and Icelandic has a 'be at'-progressive). Aspectual categories also commonly develop from possessively interpreted action schemata like 'to take X' (> 'to have X' > 'to have gone'), and aspectual categories may ultimately be interpreted as tense categories ('past tense').

The shortcomings of traditional descriptions of possessive constructions are most noticeable in the area of inalienable possession. The *dativus sympaticus*-constructions are extremely common, and the cases used in possessive constructions for the P'or or noun phrase include not only genitives, datives, locative constructions, and the plain juxtaposition mentioned above, but occasionally even accusatives (e.g., Faroese kinship relations). As a consequence, it is necessary to give special attention to alternatives to traditional grammatical encodings of possession. The book edited by McGregor and Chappell (1996) presents a rich cross-linguistic survey of inalienable possession; common strategies associated with inalienables apart from 'he hit me on the head'-constructions are 'incorporating' strategies like 'he is bigheaded' (vs. ?'he is bigcarred'). Quite typically, the P'or may not be mentioned at all on the possessed noun phrase (cf. the phenomenon of 'possessor deletion' (see above)). But conversely, inherent relationality may sometimes have to be obligatorily marked (lit. 'his hands-his'). As a variant of possessor raising giving prominence to psychologically involved possessors, also different honorifics may be used in the context of body-part relations (cf. Japanese, see Tsunoda 1996).

Payne and Barshi's definition (1999, p. 3) of 'external possession' covers more than the constructions used in possessor raising: here external possession denotes any construction in which the P'or figures as a core argument of the verb and in a constituent separate from that which contains the P'ed. This definition includes applicatives (e.g., Mixean, see Zavala 1999), 'double subject constructions' (Japanese, see Shibatani 1994) and expressions where the P'or is the subject of a verb with an incorporated P'ed (Mohawk, see Baker 1999). What is common to all these constructions is some difference of 'empathy,' 'discourse-prominence,' and 'fore-grounding' in the relationship between P'or and P'ed as compared to the 'standard' possessive constructions. In fact, all kinds of possessive constructions can be ordered in terms of the comparative discourse prominence for P'or and P'ed. With this in mind, it is necessary to be aware of some specifically linguistic functions of possessive categories.

1.3 The Linguistic Side of Possession

Different degrees of discourse prominence can be tested, among other things, with criteria like accessibility for relativization, etc. Languages differ considerably in the extent they 'normally' give prominence to possessors and possessed items, but there is usually variation between various options in the same language, too. Compare English 'her dog *who (= she)' vs. 'this dog of hers, who (= she),' where the ordinary possessive pronoun cannot be taken up in a relative clause, but the phrase 'of hers' can. Normally, there is some continuum of possessor prominence from possessive pronouns via personal pronouns in the genitive to actual arguments of the verb (in the dative or accusative, etc., as in 'possessor raising' constructions). And while the P'ed is totally inaccessible for relativization in 'the blue-eyed man *which (= the eyes) …,' the P'ed can be relativized in 'the man with the blue eyes which (= the eyes) …,' etc.

In this context, possessive constructions do not simply express some extra-linguistic concept, i.e., the constructions in question must also be judged with respect to their function in the discourse. This is particularly conspicuous in connection with genitival constructions. It is true that 'normally' the genitive phrase 'expresses' the possessor (cf. 'the man's car'), but sometimes the situation is reversed. Compare German *das Auto des Mannes ist alt* (lit. 'the car (P'ed: NOM) of the man (P'or: GEN) is old') vs. *der Besitzer des Autos ist alt* (lit. 'the owner (P'or: NOM) of the car (P'ed: GEN) is old'). The answer why it is the P'or who is in the genitive in the first example, and why is it the P'ed which is in the genitive in the second ultimately has to do with discourse pragmatics. The noun phrase in the genitive typically encodes a known referent which can be used for the identification of something which is less inherently identifiable. 'Normally,' it will be human P'ors who are 'better known' than inanimate referents or dependent referents like body-parts, etc., but in examples like German *der Besitzer des Autos mit dem Kennzeichen XY wird gebeten, sich mit der Polizei in Verbindung zu setzen* ('the owner of the car with the registration number XY is requested to contact the police') it is the inanimate

referent which happens to be better identifiable. In this case, the P'or will have to be expressed by a special relational noun (*Besitzer* 'owner'), but the genitival phrase is still basically used for identification in the discourse, even though it now encodes the P'ed. These considerations may also explain why many languages have paradigms like 'the man's car,' 'John's car,' 'his car,' 'my car,' while there is often not a corresponding construction like '*I with the car,' '*John with the car' (which in principle just looks at the relationship from another perspective just like 'the man with the car'): 'Normally,' referents like 'I' and 'John' are identifiable without further specification, while dependent items are not default referents in the discourse. Therefore it is quite 'natural' for possessed items to carry some additional marking for making them identifiable in the discourse (e.g., with reference to an animate, 'ego-like' P'or).

Also the distinction between so-called *eius*- and *suus*-possession ('of somebody else' vs. 'of his own') reflects differences of discourse prominence. In Modern Icelandic, the *suus*-possessive *sinn* only refers to prominent participants (usually a subject) within the clause. Conversely, *hans* (*eius*) is used mainly when the pronoun refers to a possessor mentioned outside of the clause. Interestingly, *hans* distinguishes masculine, feminine, and neuter with reference to the P'or, while *sinn* distinguishes masculine, feminine, and neuter only with reference to the P'ed. Hence, the 'distant' P'or will be far better retrievable by means of *hans* than by means of *sinn*, which prefers 'near' and very salient noun phrase referents. Since discourse-relevance is a direct correlate of high animacy and conceptual independence, it is clear why the linguistic treatment of possessive relations will tend to be associated with distinctions like [+/−animate] and [+/−alienable]. As animacy also correlates with control just like any relationship of possession, it is not surprising that there are conceptual shifts from possession to distinctions of agency (cf. 'I have beaten' (agentive) vs. 'I am beaten' (non-agentive)) and vice versa. Agent (AG) and patient (PAT) are also differentiated in nominalized possessor-of-the-act constructions (cf. 'Peter's (AG) beating (action) of the dog (PAT),' or Tongan *ko 'eku* (alienable) *taa'i 'o* (inalienable) *e kulií* (lit. 'it is my (active) beating of the (passive) dog,' I am currently beating the dog')). Apart from that, agent and patient relations have also frequently been observed to coincide with differences of imperfective vs. perfective aspect. From this point of view, the typical correspondences between possession, voice, and aspect are not surprising. Like the voice distinction, possession also implies some shift of perspective from an 'ordinary' discourse referent (who is animate and independent) to some referent which is usually inanimate and dependent. It is these shifts of perspective and the many alternative forms of encoding which make possession a highly interesting domain of study for linguistics and cognitive science.

Bibliography

Abney S 1987 *The English Noun Phrase in its Sentential Aspect*. MIT Press, Cambridge, MA

Baker M 1999 External possession in Mohawk: Body parts, incorporation, and argument structure. In: Payne D, Barshi I (eds.) *External Possession*. Benjamins, Amsterdam, pp. 293–324

Heine B 1997 *Possession. Cognitive Sources, Forces, and Grammaticalization*. Cambridge University Press, Cambridge, UK

Löbel E 1995 Zur kategorialen Bestimmung der Possessiva in der NP/DP: Possessiva als Adjektive. In: Tappe T, Löbel E (eds.) *Struktur der Nominalphrase*. Bergische Universität-Gesamthochschule, Wuppertal, Germany, pp. 54–86

McGregor W, Chappell H (eds.) 1996 *The Grammar of Inalienability. A Typological Perspective on Body Part Terms and the Part-Whole Relation*. Mouton de Gruyter, Berlin

Payne D, Barshi I (eds.) 1999 *External Possession*. Benjamins, Amsterdam

Seiler H 1983 *Possession as an Operational Dimension of Language*. Narr, Tübingen, Germany

Tsunoda T 1996 The possession cline in Japanese and other languages. In: McGregor W, Chappell H (eds.) *The Grammar of Inalienability. A Typological Perspective on Body Part Terms and the Part-Whole Relation*. Mouton de Gruyter, Berlin, pp. 565–630

Zavala R 1999 External possessor in Oluta Popoluca (Mixean): Applicatives and incorporation of relational terms. In: Payne D, Barshi I (eds.) *External Possession*. Benjamins, Amsterdam, pp. 339–72

J. Broschart

Postal Systems

Postal systems are administratively coordinated communication networks that originated in antiquity, yet remain important today. During their 4,000-year history, they have changed significantly, making it hard to generalize about their scale, scope, and accessibility. Until fairly recently, for example, postal systems often conveyed people in addition to information and goods.

1. Overview

Most surveys of the world's postal systems are institutional genealogies. Postal systems beget one another in a grand procession of names, places, and dates. Such an approach has the merit of bringing together a wealth of anecdotal material not easily found elsewhere (Scheele 1970). Yet it exaggerates continuity, discourages systematic comparison, and obscures the role of the postal system as an agent of change.

This article takes a comparative institutional approach. The world's postal systems are divided into three groups: imperial, corporate, and national. The section on national postal systems is subdivided into three parts: origins, postal reform, and twentieth- and twenty-first-century developments. The article concludes with a brief discussion of postal systems and social science, with suggestions for future research.

2. Imperial Postal Systems

Imperial postal systems enabled political leaders to conduct diplomacy, plan military campaigns, and oversee routine administration. The most highly developed were those of the Persians and the Romans. The Persian postal system under Emperor Darius (521–486 BC) employed teams of mounted couriers to transmit clay tablets at 14-mile intervals along a 1,600-mile route that stretched from the Tigris River to the Aegean Sea. The Greek historian Herodotus praised its speed and regularity: 'Neither snow, nor rain, nor heat, nor gloom of night stays these couriers from the swift completion of their appointed rounds.'

The Roman postal system, the *cursus publicus*, was established by Emperor Augustus (29 BC–AD 14) and lasted 500 years. Relays of couriers with horse-drawn carts transmitted official dispatches along with government officials and their baggage. Routes extended throughout the empire and, during military campaigns, to generals in the field. Messages were written on papyrus, parchment, or wax-coated tablets, paper remaining unknown in Europe until after AD 1000. The relays explain the etymology of 'postal'; it was derived from *positus*, the past participle of *ponere*, 'to place.' Couriers traveled 50 miles per day—a speed rarely exceeded until the nineteenth century.

Access to the *cursus publicus* was restricted largely to government officials. For this reason, some scholars term it a government communication network, and reserve the term postal system for networks open to the general population. Merchants relied on private messengers, seafarers, or friends. Maritime links were particularly important for Christian apostles such as Paul, who used them to send epistles throughout the Mediterranean world.

3. Corporate Postal Systems

Long after the fall of Rome, imperial postal systems continued to flourish in Asia, South America, and the Middle East. In Europe, however, no comparable institution existed for centuries. When postal systems revived in the 1200s, their organizers were no longer imperial regimes, but corporate bodies, including the University of Paris, the monks of Cluny, and the German butchers' guild.

Corporate postal systems rarely maintained relays of mounted couriers and often lacked the trappings of power. Yet in several respects they marked a distinct advance. Many crossed political boundaries, establishing links *between* states as well as *within* them. Few were restricted, like the Roman *cursus publicus*, to government officials. Open access was facilitated by the recent introduction of paper, a light-weight writing medium that was easy to transport. The policy also helped defray costs, since travel was expensive and many corporate postal systems were self-supporting. Henceforth, the primary constraint for potential postal patrons was financial: postal rates were typically fairly high, limiting access to the well-connected and the well-to-do.

Among the most important corporate postal systems were those established by the Italian city-states. By the end of the thirteenth century, couriers linked Venice with Constantinople, and Florence, Genoa, and Siena with northern France. The most enduring was the *Corrieri Bergamaschie* that Amadeo Tasso established in Bergamo, near Milan, in 1290. This institution much impressed the Hapsburgs, who, in the 1400s, enlisted the Tasso and Delle Torre families to extend it throughout the Holy Roman Empire.

To oblige the Hapsburgs, the families Germanicized their names to Thurn and Taxis. Shortly thereafter, in 1489, Johann von Taxis became the Hapsburgs' first postmaster. Within the next few years, he and his brother Franz von Taxis established postal links between the Hapsburgs' far-flung domains and Rome, France, and Spain, and opened the service to the general population (Dallmeier 1977). Couriers carried a distinctive post horn, an instrument that remains to this day a symbol of the German Deutsche Post. Although the rise of the nation-state greatly restricted its scope, the Thurn and Taxis postal system survived until 1867, when it was bought out by the Prussian government (Behringer 1990).

4. National Postal Systems

4.1 Origins

Corporate postal systems stymied the rulers of the emergent nation-states in several ways. Most obviously, they created an unregulated communication channel that threatened political stability. In addition, they deprived rulers of a useful tool with which to monitor the flow of information and ideas. In response, nation-states throughout Europe, led by France and England, established postal systems under tight government control (Coase 1955).

The French and English postal systems originated, like the imperial postal systems of antiquity, as government communication networks. In France, the founder was Louis XI (1461–83), and in England,

Henry VIII (1509–47). In both countries, the early 1600s marked a decisive turning point. Within a few decades, both governments granted access to the general population, instituted a monopoly, established a rate structure and delivery schedule, and organized a surveillance mechanism to monitor their subjects' correspondence: the *cabinet noir* in France; the secrets office in England (Vaille 1946–55, Robinson 1948). Shortly thereafter, the English government mandated a major rate increase, transforming its postal system into a revenue source, which it would remain until the twentieth century (Daunton 1985).

Beginning in the 1600s, national postal systems did much to hasten the emergence of the newspaper—at the time, a new form of political communication. This relationship was particularly close in England, British North America, The Netherlands, and Prussia. By permitting the regular transmission of information on current events, postal systems became, in effect, an early broadcast media. Having gained control of the means of communication, governments used it to foster a distinctive world view. In so doing, they helped forge the national identities that are such a fundamental feature of the modern world (Mueller 1986).

Nowhere was the relationship of the postal system and the newspaper closer than in the USA. In the 1790s, following the adoption of the federal Constitution, the American government invested its postal system with an educational rationale far more expansive than the fiscal rationale that had previously prevailed. The new mandate, codified in the Post Office Act of 1792, hastened the circulation of newspapers by conveying them at extremely low rates while enormously expanding the number of localities to which they could be sent (John 1995). A landmark in postal history, this policy may well have marked the first time a government deliberately subsidized the transmission of a particular kind of information.

By 1828, the American postal system, with almost 8,000 post offices, was larger and more geographically extended than any postal system in Europe. At this time, the USA boasted 74 post offices per 100,000 inhabitants, Britain 17, and France four (John 1995). Prior to the coming of the railroad and telegraph, the American government had established the preconditions for the disembodied public sphere that French traveler Alexis de Tocqueville (1835, 1840) would explore so perceptively in *Democracy in America*.

4.2 Postal Reform

Victorian commentators often hailed the postal system of their day as a fundamentally *new* institution that was little older than the railroad and telegraph. For the majority of the population, this was often true. Prior to the 1840s, postal rates remained so high that letter writing was, in practice, confined to the few. Even in the USA, if a farmer or artisan received anything in the mail, it was more likely to be a newspaper than a letter.

In both Britain and the USA, the high cost of letter postage hastened the establishment of clandestine letter-delivery services in direct violation of the postal monopoly. Following the advent of the railroad in the 1830s, these operations expanded greatly, hastening the major rate reduction known as postal reform.

Postal reform began in Britain, where it was championed by educator Rowland Hill. Hill recognized that high letter rates encouraged illicit mail carriage, and posited that a major reduction would hasten a massive expansion in letter writing. To buttress his argument, he demonstrated in a famous pamphlet, *Post Office Reform: Its Importance and Practicality* (1837), that rates bore no relation to expenses. All but a tiny fraction of the cost of delivering a letter, Hill calculated, derived from its *routing* through the system rather than from its *conveyance* from sender to recipient.

Hill prevailed and, in 1840, the British government simplified the rate schedule and reduced the basic letter rate to one penny. Hill styled his reform uniform postage; it was commonly called penny postage. Hill's reform greatly increased mail volume, just as he had predicted. In addition, it hastened the advent of several familiar postal practices, including mandatory prepayment, mail boxes, and the postage stamp (Daunton 1985).

Many governments followed Hill's lead. The USA lowered its letter rates in 1847 and 1851; France in 1848. Japan, eager to emulate the best practices in the West, established in 1870 a postal system modeled closely on Britain's (Westney 1987). Increased mail volume, in conjunction with the advent of the ocean-going steamship, hastened the proliferation of transnational postal treaties. These diplomatic initiatives were regularized in 1878 with the establishment of the Universal Postal Union (UPU), today a branch of the United Nations. Closely related to postal reform was the substitution of the railroad for the stagecoach as the preferred mode of conveyance. By the 1830s, stagecoach service had attained a high level of development, especially in Britain and France. Interestingly, Rowland Hill based his famous cost calculations on the stagecoach, rather than the railroad, which had yet to be widely adopted. Stagecoach speeds often averaged 10 miles per hour, a standard rarely matched since the demise of the *cursus publicus* of imperial Rome. The use of the railroad to convey mail posed various scheduling dilemmas and could increase postal costs; in the USA, for example, it quickly ran the postal system into debt. Yet it proved indispensable, greatly increasing carrying capacity and revolutionizing postal routing. In Britain, the USA, and most industrialized nations, it was long customary to sort the mail on-board moving trains, a practice known as railway mail.

Postal reform ushered in a golden age of letter writing. In many countries, the exchange of holiday greeting cards became a popular custom. Soldiers' letters emerged, especially in wartime, as a notable mode of literary expression. Although letter writing would, in the twentieth century, face competition from telephony, the habit persists, and, since the popularization of E-mail, may well be enjoying a revival.

In the decades preceding World War I, each of the world's leading postal systems—Britain, France, Germany, and the USA—undertook major new responsibilities. Each established a parcel post and instituted a rudimentary banking service known as postal savings. And each, with the exception of the USA, took over the telegraph and telephone, establishing a ministry of Posts, Telegraphs, and Telephones, known in France and elsewhere as the PTT (Bertho 1981, Perry 1992). This expansive era in postal history reached its apogee around 1900, when it was common for journalists to hail the postal system as the greatest business in the world (Fuller 1972).

4.3 Twentieth and Twenty-first Century Developments

In both the developed and the developing world, postal systems remain indispensable elements of the information infrastructure (Sinclair 1984). In the latter part of the twentieth century, however, much changed. Beginning around 1970, rapid changes in communications technology led many governments to re-evaluate the union of the postal system, the telegraph, the telephone. Deregulation followed. In Britain, the break-up came in 1981, and in France, in 1991.

The Internet poses a further challenge, yet, in a sense, it is nothing new. Since the mid-nineteenth century, postal systems have co-existed with earlier modes of electric communications. The USA is the originator not only of the Internet, but also of 40 percent of all the world's mail. In the early twentieth century, pundits routinely predicted that telegraphy would soon render mail delivery obsolete, yet telegraphy has been everywhere supplanted, while postal systems thrive.

Additional challenges stem from the emergence of new competitors, e.g., DHL in Europe, Federal Express and UPS in the USA. Express firms rely on improvements in air travel to compete head-on with the national postal systems in key high-value markets. In response, many governments have re-examined—and, in several cases, narrowed—the scope of the postal monopoly (Campbell 1994).

Postal systems have met these challenges in various ways. Many have reinvented themselves as public corporations, a transformation sometimes called privatization. If this trend continues, postal systems in the twenty-first century may share less with the national postal systems out of which they evolved than the corporate postal systems of medieval Europe. The French postal system, known today as La Poste, has been a publicly traded corporation since the 1970s. The British postal system, the Royal Mail, has aggressively sought new markets. So, too, has the Deutsche Post, which issued its first securities in 2000, and the US Postal Service, which has not been a government department since 1970 and no longer receives a government subsidy.

5. Postal Systems and Social Science

Scholars agree that the printing press was an agent of change. Only recently, however, have postal systems been studied in a similar way. The last edition of the *International Encyclopedia of the Social Sciences* (1968) contained no entry on the postal system, and no references to it in the index.

The rise of the new institutionalism in history, sociology, and political science bodes well for postal studies. So too does the vogue for French literary theory, with its focus on the complexity of the communication process (Siegert 1993). Developments in information technology promise further insights, offering new perspectives on modes of communication that are seemingly being displaced. For postal scholars, Hegel's dictum holds true: the owl of Minerva flies at dusk.

Research opportunities abound. Comparative data remain scarce, especially for the period prior to 1840. Without it, it is hard to frame credible generalizations about what postal systems did and how they worked. Topics that invite exploration include the development and significance of scheduling; censorship and surveillance; postal privacy; communication networks and the rise of nationalism; challenges to the postal monopoly; changing rationales for postal systems; and the shifting meaning of the equal access ideal.

See also: Communication and Democracy; Communication and Transportation, History of; Communication: Electronic Networks and Publications; Communication: Geographic Aspects; Internet: Psychological Perspectives; Mass Communication: Empirical Research; Mass Communication: Normative Frameworks; Mass Communication: Technology; Telegraph; Telephone

Bibliography

Behringer W 1990 *Thurn und Taxis: Die Geschichte ihrer Post und ihrer Unternehmen*. Piper, Munich

Bertho C 1981 *Télégraphes et Téléphones de Valmy au Microprocesseur*. Librarie Générale Française, Paris

Campbell R M 1994 *The Politics of the Post: Canada's Postal System from Public Service to Privatization*. Broadview Press, Peterborough, Canada

Coase R H 1955 The postal monopoly in Great Britain: An historical survey. In: Eastham J K (ed.) *Economic Essays in*

Commemoration of the Dundee School of Economics, 1931–55. William Culross, Coupar Angus, UK, pp. 25–37
Dallmeier M 1977 *Quellen zur Geschichte des europäischen Postwesens, 1501–1806*. 3 parts. Lassleben, Kallmünz, Germany
Daunton M J 1985 *Royal Mail: The Post Office since 1840*. Athlone, London
Fuller W E 1972 *The American Mail: Enlarger of the Common Life*. University of Chicago Press, Chicago
John R R 1995 *Spreading the News: The American Postal System from Franklin to Morse*. Harvard University Press, Cambridge, MA
Mueller M 1986 The currency of the word: War, revolution, and the temporal coordination of literate media in England, 1608–1655. M.A. thesis, University of Pennsylvania, Philadelphia, PA
Perry C R 1992 *The Victorian Post Office: The Growth of a Bureaucracy*. Royal Historical Society, London
Robinson H 1948 *The British Post Office: A History*. Princeton University Press, Princeton, NJ
Scheele C H 1970 *A Short History of the Mail Service*. Smithsonian Institution Press, Washington, DC
Siegert B 1993 *Relais: Geschicke der Literatur als Epoche der Post, 1751–1913*. Verlag Brinkmann & Bose, Berlin; 1999 *Literature as an Epoch of the Postal System*. Stanford University Press, Stanford, CA
Sinclair J 1984 *Uniting a Nation: The Postal and Telecommunication Services of Papua New Guinea*. Oxford University Press, Melbourne, Australia
Vaille E 1946–55 (1947) *Histoire Générale des Postes Françaises*. 7 Vols. Presses Universitaires de France, Paris
Westney D E 1987 *Imitation and Innovation: The Transfer of Western Organizational Patterns to Meiji Japan*. Harvard University Press, Cambridge, MA

R. R. John

Postcolonial Geography

Historically speaking, a special relationship existed between geography and the making of empires. Pragmatic geographies of orienteering, surveying, mapping and description were the frontier technologies of the operations that brought distant territories under the control of metropolitan centers. As historian Edward Said famously noted in his book *Culture and Imperialism* (1993, p. 271), imperialism is 'an act of geographical violence through which virtually every space in the world is explored, charted, and finally brought under control.' At no time was the relationship between geography and colonialism more intimate than during the span of modern European expansion. Such worldly colonizations might be understood simply as metropolitan appropriations of the resources of other peoples and lands. Yet this material self-interest was motorized by a complex set of values that made these often audacious and violent appropriations make perfect moral sense to those whose interests were being served. Not least, deeply embedded assumptions about racial superiority and apportions of civility, gave legitimacy to these territorializations. The end of the nineteenth century marked the close of an era in which the discipline of geography was intoxicatingly enlivened.

The modern discipline of geography lives with this past. For a dwindling few it remains a source of nostalgic pride. For many it is source of considerable ambivalence. It is, in part, out of this ambivalence that some of the work that might claim the title 'postcolonial geography' emerges. The global authority assumed by imperial geographical knowledges was laid to rest finally by the self-reflexive shock wave that passed through the discipline in the mid-1980s. Geographers who had traditionally marauded across class, gender, race, and other divides began to interrogate the politics of representation associated with such reportage (see special issue of *Environment and Planning D: Society and Space* 1992). This was a symptom of a discipline that was hurriedly retreating from its own past.

To understand what postcolonial geographies are it is essential to understand what is meant by the term postcolonial. As the prefix suggests, postcolonialism refers to the state that comes after colonialism. Most simply, this is the political condition that arrives when a colony is granted independence from its colonial rulers. Many have questioned the 'post' of such postcolonialism. The movement from colony to independent nation–state is not a sharp break with the past. Postcolonial states remain complexly entangled in their colonial pasts, just as they are constantly negotiating various kinds of neocolonialism, be it in the form of geopolitical influence, cultural effect, or economic leverage. The persistence of various kinds of colonialism has led various scholars to note that the term postcolonialism is prematurely celebratory. Others are similarly skeptical, suggesting that the term does nothing more than describe a condition of a newly diasporic intelligentsia.

Those geographies to claim the term 'postcolonial' reflect a use common to the social sciences and humanities more generally. Here the term refers to an ongoing set of relationships and processes inaugurated by the experience of colonization. Postcolonialism is not then a state that arrives once colonialism ends, but a struggle that comes into being because of, and alongside, the fact of colonization. In this sense, colonial and postcolonial formations are cohabitors. For example, colonial authorities were always dealing with postcolonial effects produced by this cohabitation. This might have included certain subtle, but nonetheless unsettling hybridizations. More starkly, it might have included counter-colonial formations through which the colonized sought to reinstate their authority.

Not surprisingly, many different kinds of formations, periods, and circumstances have come to be described as postcolonial. Such elasticity in the term

postcolonial compromises both its descriptive and analytical worth. Yet postcolonialism is still a useful term to apply to a particular intellectual (and political) stance. Indeed, postcolonial studies are defined as much by their theoretical field of influence as they are their empirical concerns. Nowhere is this more evident than in the discipline of geography. Postcolonial geographies are varied in their subject matter, but may be drawn together as a loosely unified set of geographies intent on bringing into view a set of phenomena that work against colonial authority. This might include evident ruptures in the apparent coherence of colonial occupation, it might include the hybrid outcomes of cohabitation, the so-called 'minor' knowledges and alternative terrains of the colonized, and postcolonial formations that aim to wrest back control of territories.

Much of the current work within postcolonial geography, as with postcolonial studies more generally, is based on revisiting the site of colonialism in order that the silences that exist in the colonial record may be uncovered. This recovery project, it is assumed, will lead to a radical rethinking of the forms of knowledge and social identities authorized under colonialism. Historical geography has been much enlivened by this critical return to the imaginaries, logics, and practices of colonialism. In 1993 Livingstone (1993) published his elegant account of the geographical tradition, within which he paid particular attention to the relationship between European geography, race, and empire. In the same year Driver (1992) published an important article that 'unmasked' modern geography's relationship to European colonization. By demonstrating the way geographical knowledges framed other worlds, his paper drew the analytic terms set by Edward Said's landmark contribution, *Orientalism* (1978), into the heart of a rapidly postcolonizing discipline.

Critical appraisals of geographers and the knowledges they produced have continued as an important theme within the discipline. Harley's (1988) work on the relationship between cartographic knowledges and power, is exemplary in this regard. His re-appraisal of cartographic practices and colonialism contributed to an on-going scholarship (both within the discipline and outside) on cartographies and colonial power. The spatial emphasis in geography's critical engagement with the colonial archive is also evident in accounts of other, more populist, visual representational fields. Photography, for example, played an important role in bringing home the experiences, peoples and lands of those faraway new worlds (Ryan 1997).

Geographical practitioners, institutions, knowledge systems, and pedagogies of various kinds have succumbed to the scrutiny of this autocritique. Particular attention has been given to the systems of knowledge used to understand, order and benefit from the often exotic nature found in the new worlds. These knowledge systems made meaningful peoples and lands that, from a eurocentric worldview, were so radically different they could at first only be understood as 'savages' or 'wastelands.' Two showcase collections of studies detailing diverse examples of geography's relationship to various imperialisms (old and new, European, and other) appeared in the mid-1990s (Godlewska and Smith 1994, Bell et al. 1995).

Delivering a gender sensitive perspective to the geographical critique of imperialism has shown that this is seemingly masculinist project, is one in which women (colonial and colonized) are complexly embedded, both as fellow adventurers and as objects of desire. The letters and journals of women who traveled to colonized lands have been an important source of information for rich scholarship on the women in the colonial project.

The global scale expansions of imperialism ground themselves in micro-geographies of occupation and domestication. As Jacobs (1996, p. 19) notes, the pure idea of imperialism 'tumbled into the fractured and erratic everyday practices of the personalities … who were forced together in the making of colonies.' One of the key collective contributions of the now vast body of critical colonial geographies has been to point to the contingencies of colonialism as a located, place-specific practice. For some time a key concern of this work has been the way relations of differentiation and domination have been enacted in and through space, and how colonized subjects subvert or negotiate those spatial controls.

A significant subset of critical colonial and postcolonial geographies concerns the city. Although not a geographer, the work of Anthony King has been highly influential in shaping the direction of this inquiry within the discipline. His theory of colonial urban development outlined the economic, political, and cultural processes that gave rise to and defined the form of the cities of colonial territories. Colonial cities were important sites in the transfer of modern capitalist values to new worlds and functioned as important centers of power for administering indigenous people and resources. In these cities, town-planning ordinances about cleanliness and sanitation became the mechanism through which colonial assumptions about racial difference were translated into on-the-ground technologies of segregation and exclusion. Jacobs (1996) extends this urban postcolonial scholarship by documenting the ways in which imperial pasts and postcolonial struggles continue to shape contemporary cities. More recent scholarship has turned its attention to the way imperial geographies impacted upon the townscapes and spatial arrangements of the metropolitan cities that stood at the heart of empires. Driver and Gilbert (1999) offer an exemplary collection on the history of cross-mappings between empire and modern European cities.

The geographies accounted for thus far largely confine their attention to the archive or to historic

landscapes and territorializations. The intellectual project of making a critical return to the scene of colonialism is undertaken in the belief that it contributes to a decolonization of the discipline. Such a return can and does have its postcolonizing value, but there remains a certain ambiguity around just how postcolonial these modern adventures in critical colonial geography might be. Not least there is a tendency in these geographies for the colonized subject (the 'other') to be brought into view simply by way of the colonial discourses that framed them. Can attending to the colonial scene with a postcolonial lens dislodge the effects of that history? Or does it simply re-inscribe the colonial moment as the master referent of all that follows? Colonial constructs, practices, and landscape transformations have significant after effects, they flow into present times both sustaining vectors of domination and exploitation and providing the site against which alternate identities, places, and ways of being in the world are asserted. It is to the geographies that attend to the contemporary scene of postcolonialism that we now turn.

Racism and racialized difference rightly dominate the critical analysis of imperialism and the making of colonies, for it was on the basis of this register of difference that many of the most audacious imperial appropriations have been made. Despite this, there is an intellectual distance between the radical geographical tradition of studies of race and racism and geographies that describe themselves as 'postcolonial.' Some recent work has begun to converge around contemporary notions of whiteness, the postcolonial politics of antiracism and the emergence of new forms of postcolonial racism.

With geographical knowledges and technologies playing such a central role in the making of colonial territories, they have come to play an important part in the unmaking of these territories. There is, for example, a growing literature within geography on the use of cartographic technologies, place renaming, and alternative geographical knowledges, in various forms of postcolonial nationalism and indigenous rights movements, as well as newly commodified forms of conquest. Similarly, geographers have begun to show how modern environmentalisms are complexly entangled with colonialist assumptions about nature and race: bringing forth new kinds of (romantic) racisms and exclusions alongside of new possibilities for political alliance.

An exemplary body of self-consciously decolonizing geographical writing is the scholarship that has emerged around the racist spatial logic of settlement in Southern Africa. That lingering colonial construct of apartheid gave rise to stark geographies of differentiation and disadvantage based on racially defined 'group areas.' These racist spatialities called into being many passionately political geographies of reform. Similarly postcolonial geographies are emerging in the local scholarship other former colonial territories and settler states. Some of these geographies continue to remind us of the tenacity of colonial and neocolonial constructs and the way in which they are embedded in the physical and ideological landscapes of places and nations. This might include the ways in which the landscapes and identities of those once colonized are now being absorbed into new touristic circuits or commodity markets. Other postcolonial geographies genuinely reorient the concerns and subject matter of the discipline. They may install radically alternative evaluations of land and resources, that are directed away from dominant models of use and in so doing give legitimacy to previously marginalized ways of being in the world. They may chart hopeful examples of resistance to domination, be it in the form of minor subversions or stark oppositions. They may even be so optimistic as to speculate about new models of coexistence. What all of these emerging geographies of postcolonialism reveal is how specific historical and geographical contexts have delivered a peculiar political and moral imperative to the project of decolonizing the disciplinary gaze.

See also: Colonialism, Anthropology of; Colonialism: Political Aspects; Geography; Globalization and World Culture; Imperialism, History of; Imperialism: Political Aspects; Racism, History of; Subaltern History

Bibliography

Bell M, Butlin R, Heffernan M (eds.) 1995 *Geography and Imperialism, 1820–1940*. Manchester University Press, Manchester, UK and New York

Blunt A, Rose G (eds.) 1994 *Writing Women and Space: Colonial and Postcolonial Geographies*. Guilford Press, London and New York

Bonnett A 2000 *White Identities: Historical and International Perspectives*. Prentice Hall, Harlow, UK

Crush J 1994 Post-colonialism, de-colonization, and geography. In: Godlewska A, Smith N (eds.) *Geography and Empire*. Blackwell, Oxford, UK, pp. 333–50

Driver F (ed.) 1992 Geography's empire: Histories of Geographical knowledge. *Environment and Planning D: Society and Space* **10**: 23–40

Driver F, Gilbert D 1999 *Imperial Cities: Landscape, Display and Identity*. Manchester University Press, Manchester, UK and New York

Godlewska A, Smith N (eds.) 1994 *Geography and Empire*. Blackwell, Oxford, UK

Harley J B 1988 Maps, knowledge and power. In: Cosgrove D, Daniels S (eds.) *The Iconography of Landscape*. Cambridge University Press, Cambridge, UK, pp. 277–312

Jacobs J M 1996 *Edge of Empire: Postcolonialism and the City*. Routledge, London and New York

Livingstone D N 1993 *The Geographical Tradition*. Blackwell, Oxford, UK

Marcus G E (ed.) 1992 *Environment and Planning D; Society and Space* **10**(5) (whole issue)

Phillips R 1997 *Mapping Men and Empire: A Geography of Adventure*. Routledge, London and New York

Robinson J 1995 *The Power of Apartheid: State, Power and Space in South African Cities*. Butterworth-Heinemann, Oxford, UK and Boston
Ryan J R 1997 *Picturing Empire: Photography and the Visualization of the British Empire*. The University of Chicago Press, Chicago
Smith D M (ed.) 1992 *The Apartheid City and Beyond*. Routledge, London and New York
Sparke M 1994 White mythologies and anemic geographies—a review. *Environment and Planning D: Society and Space* **12**: 105–23
Yeoh B S A 1997 *Contesting Space: Power relations and the Urban Built Environment in Colonial Singapore*. Oxford University Press, Oxford, UK

J. M. Jacobs

Postcolonial Law

Postcolonial theory is largely associated with South Asian scholarship, subaltern and literary studies, analyses of resistance, and a general rethinking of a dominant European historiography that places the 'West' at the center of the world. Postcolonial theory plays a particularly important function in its acknowledging and recovering the ongoing significance of colonized peoples in shaping the epistemologies, philosophies, practices, and shifting identities of what is taken for granted as dominant Western subjects and subjectivities. Most importantly in the context of law, postcolonial theory, and the various colonial sites through which it is inscribed and substantiated, punctuates the degree of violence and the technologies of power involved in understanding concepts such as modernity and capitalism as well as contemporary state and trans-state nationalisms.

That being said, postcolonial theory rarely considers the importance of law and imperial governance (however, see Baxi 2000, Spivak 1999). Likewise, legal scholarship has largely ignored the significance of a postcolonial perspective (see Darian-Smith and Fitzpatrick 1999). Certainly sociolegal scholarship has gone some way to bring to the fore some of the implications of gender, ethnicity, race, class, age, religion, and so on, in the formulating of the rule of law and the maintenance of its legitimizing practices. However, the particularly uncomfortable position offered by a postcolonial perspective, which places side by side the oppressed and the oppressor, and their enduring psychological and material interdependence, is generally treated as a radically different line of inquiry from that which seeks to analyse economic and political inequalities promulgated through a particular legal system. What, after all, do cultural and literary studies that focus on the colonial situation have to do with an empirical study of law in places such as the United States? In an attempt to dismantle this prevailing attitude of mutual exclusivity between postcolonial and legal studies, discussed below are some of the insights postcolonial theory brings to bear on an understanding of contemporary legal processes in order to identify, analyse, and assess what is meant by the term 'postcolonial law.'

1. *Defining a 'Postcolonial' Perspective*

Throughout the late 1980s and 1990s, a considerable body of scholarship devoted to issues of postcoloniality emerged. And with this proliferation of research in Anglo/European universities, there was, and is still, much debate over the meaning and scope of postcolonial terminology (see Schwarz and Ray 2000, Loomba 1998, Williams and Chrisman 1994). Despite these ongoing deliberations, it is helpful to turn to the most significant implications of a postcolonial perspective which mark it as distinctly different from, yet complementary to, other contemporary theoretical investigations. This process will help to flesh out the contours of postcolonial studies and, at the same time, underscore its intellectual lineage.

Postcolonial studies emerged initially among Indian and British scholars in the wake of World War II and in an era of general demise of colonial regimes. Spurred on by radical Western thinkers such as Marx, Nietzche, Althusser, Fanon, Jameson, Foucault, and Derrida, Edward Said's important publication *Orientalism* (1978), these scholars of historiography critiqued Eurocentric and Indocentric interpretations alike for their assumptions of superiority in European countries vis-à-vis the past and future of Indian people and a modernist concept of the Indian nation (Prakash 1990, 1992, O'Hanlon and Washbrook 1992, Chakrabarty 1992).

Building on Marxist, poststructuralist, and feminist theories, postcolonial scholars have played a major role in bringing to the fore the chronological and psychological relations between the former colonized and colonizers, whom, they argue, cannot be understood except in conjunction with each other (see Gandhi 1998). Postcolonial theorists do not claim that colonialism was experienced in the same way under different regimes, just as they recognize that today colonialism operates in very different ways from its earlier configurations. Nonetheless, while paying attention to the details of specific contexts, postcolonial scholars agree that, in order to understand all contemporary histories of peoples and places, irrespective of whether there historically existed in any given site an explicit colonial regime, it is crucially important to remain aware of the enduring presence of former colonial discourses that posit 'civilized,' 'progressive,' and 'just' Europeans against 'barbaric,' 'static,' and 'unethical' indigenous populations.

According to these scholars, colonial assumptions of Western superiority endure across time and across attempts to blur racial and ethical boundaries. While racism is not equivalent to colonial

regimes, categories of racial difference were used by colonial governments to varying degrees as strategies of power and control over indigenous peoples (Stoler 1989, Dirks 1986). Today, despite claims of increasing acceptance of multiculturalism and cultural diversity, racial categories and differences continue to exist, though often in less overt manifestations. And, according to many postcolonial theorists, these boundaries of difference are insurmountable given the psychological intersubjectivity between former colonizers and colonized which constantly necessitates a desire for re-representations of difference (see Fitzpatrick 1999). Drawing on a variety of theoretical perspectives—including an Hegelian master/slave dialectic, phenomenological essentialism, and psychoanalytical insights gleaned from Fanon and Lacan—postcolonial theorists such as Homi Bhabha, Gayatri Spivak and Benita Parry bring to the fore complex understandings of how oppressed peoples resist and seek empowerment (see Gates 1991).

In recognizing processes of mutual desire and negation between the 'master' and 'slave,' postcolonial scholars are forced to recognize the profound irony in attempting to present alternative historical narratives and identities other than those conventionally supplied by the West (Trivedi and Mukherjee 1996). The very moment of self-determination by peoples formerly colonized requires the adoption of European knowledge such as notions of progress and development, concepts of self and personhood, as well as forms of government, state-building, and, perhaps most importantly of all, law. Thus, according to a postcolonial perspective, all assertions of freedom and self-awareness require elements of mimicry and voyeurism. As Douglas Robinson has noted:

Postcolonial or subaltern scholars claim it is at once essential and impossible to forge a 'new' postcolonial identity: *essential*, because those colonial constructs were at once alien and negative, because they came from the outside and destroyed much of value in the indigenous cultures, and because an effective postcolonial politics requires the development of more positive indigenous visions: but also *impossible*, because colonial discourse continues to inform even these postcolonial attempts to break free of it, and tends to condition even the imagination of a 'new' (postcolonial) identity along 'old' (colonial) lines. (Robinson 1997, pp. 19–20)

A prominent postcolonial theorist, Dipesh Chakrabarty, has responded to the ironies of alterity by calling for the 'provincializing,' or decentering, of Europe and European epistemological knowledge. Chakrabarty (and others) argue that non-Western knowledge has been historically ignored and precluded from historiographical accounts of humanist understanding and intellectual endeavor (Chakrabarty 1992, 2000a). In an effort to critique 'the "Europe" that modern imperialism and (third world) nationalism have, by their collaborative venture and violence, made universal,' Chakrabarty urges scholars 'to write into the history of modernity the ambivalences, contradictions, the use of force, and the tragedies and the ironies that attend it' (Chakrabarty 1992, pp. 20–21).

2. *Postcolonial Law*

Nowhere in the histories of modernity has the 'use of force, and the tragedies and the ironies that attend it' been so obvious as they are in the context of law. Law, in a variety of ways, was the formal mechanism and institutional frame through which colonial governments oppressed and controlled indigenous peoples. Of course, there are many varieties of imposition and reception of law, a few even categorized as voluntary. In some cases, such as in Africa (Mamdani 1996), this involved coopting native chiefs and traditional procedures of arbitration and dispute resolution. In other cases, such as Australia, the British declared the continent *terra nullius* or vacant, and on the basis that native populations were thought to be less than human they were deemed to have no law at all. As a result, traditional methods of peacekeeping were overlooked or deliberately obliterated. Law, emblematic of European rationalism, universal objectivity, individual property rights, and sovereign state authority, provided the justification and the discourse of capitalism, and its necessary corollary of social inequality, be it based on the grounds of gender, class, or ethnicity.

It is impossible to separate colonial from postcolonial law since each developed in collaboration and in conjunction against and through each other. The metropole has always been connected with its colonial peripheries, be these geographically distanced outposts beyond state boundaries, or socially and politically isolated enclaves (such as native reservations) within it. The process of decolonialization and the emergence of new states throughout the twentieth century clearly demonstrates the extent to which independent nations have had to adopt modern forms of state building, and the institutions, bureaucracies, constitutions, and legal ideologies of their former oppressors in order to participate in national and international legal communities. The rule of law, the handmaiden of Western capitalist enterprise, exemplifies the postcolonial dilemma that requires new states and non-Western states to mimic and adopt, at least to a certain extent, the legal structure of so-called democratic society.

3. *Postcolonial Law and Globalization*

Prevailing discussions about law and globalization epitomize the ironies presented by postcolonial law. For instance, The World Bank, the United Nations, the European Union, GATT, NAFTA, and other international legal arrangements all require that, in

order for a country to participate in a global political economy, it must demonstrate commitment to, and adherence with, the foundational values of modern Western law. This does not mean that Western law, or more specifically Anglo/American law, remains impervious to the challenges and collaborations of interacting with different legal values and processes. In a postcolonial world, just as in the colonial situation, there are always ongoing modifications and appropriations between legal systems and the various cultural values that legal processes represent. One only has to think of human rights to appreciate the extent to which its apparently universal actual application has to be constantly modified according to prevailing social, political, and economic values existing within particular peoples and places. That being said, however, the dominance of Anglo/American law is overpowering and not easily destabilized.

An assumption of Western superiority is evidenced by a majority of United States and European socio-legal scholars who continue to treat the rule of law as a discrete entity or thing, much like the nation-state (which law has historically served), and religion (which law has in effect replaced), and not as a dynamic product of historically contested cultural (and colonial) interactions. Thus up to the end of the twentieth century, in arguments both for and against the enduring significance of the nation-state in the increasing context of globalization, the meaning of law is taken as a given, and it is only its spheres of centralized legal activity and influence that are being brought into question. As a result, legal analysts tend to look only at the privileged domains of legal interaction amongst lawyers, judges, business people, and entrepreneurs, and not from the perspective of ordinary people (e.g., Dezalay and Garth 1996). An important point is that what happens in the offices of international and state law firms, as well as bureaucratic and regulatory institutions, may have very little to do with the operation of law on the ground in the barrio, reservation, schoolyard, parking lot, hospital, jail, home-cleaning service, bank, and employment office. These realms of legal interaction, while not so overtly important in macroeconomic and political terms, are, nonetheless, where the majority of ordinary people interface with the law and understand its meanings. And it is in these realms of everyday legal interaction where challenges to static legal ideologies and definitions coagulate, solidify, and, in some cases (such as the riots against the World Trade Organization in Seattle in 2000), become articulated.

4. *Challenges and Limits of Postcolonial Law*

To really accept the challenges and resistances presented by postcolonial law, be it from within or from without the boundaries of the nation-state, requires that legal scholars accept that there is no universal legal code and legal objectivity. This in turn would mean that the relative dominance of one set of legal values over another would be a constant topic of review and analysis. The naturalized centrality and superiority of an American/European legal perspective would be dislodged and, borrowing Chakrabarty's terminology discussed above, necessarily 'provincialized.' Given the relations between law, capitalism, and a global political economy, it is no surprise that Western legal scholarship has largely ignored, some would argue deliberately, the long-term implications of the presence of postcolonial law.

That being said, the ironies and limits of a postcolonial perspective applies to the context of law as it does elsewhere. Desires to elevate the singularity, independence, and significance of peoples formerly colonized, and their particular legal ideologies and practices, must also recognize the extent to which postcoloniality is a product—culturally, politically and economically—of the colonial enterprise. There is no obvious solution to how to move beyond the inherent limitations whereby postcolonial knowledge is posited as part of the enduring problem of Western oppression and domination of others. Still one thing seems sure: in order to reveal the intersections between legal processes and new forms of colonialism and imperialism that are emerging within the technological and economic innovations of the twenty-first century (such as the Internet), what is first required is a basic rethinking of the currently prevailing Eurocentric, modernist, state-bound understanding of what constitutes law.

See also: Colonialism: Political Aspects; Law and Development; Law: Imposition, Reception, and Colonial; Modernity; Modernity: History of the Concept; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: South Asia; Nationalism, Historical Aspects of: The West; Postcolonial Geography; Postcoloniality

Bibliography

Baxi U 2000 Postcolonial legality. In: Schwarz H, Ray S (eds.) *A Companion to Postcolonial Studies*. Blackwell, Oxford, pp. 540–55

Chakrabarty D 1992 Postcoloniality and the artiface of history: who speaks for 'Indian' pasts? *Representations* **37** (Winter): 1–26

Chakrabarthy D 2000a *Provincializing Europe: Postcolonial Thought and Historical Difference*. Princeton University Press, Princeton, NJ

Chakrabarty D 2000b A small history of subaltern studies. In: Schwarz H, Ray S (eds.) *A Companion to Postcolonial Studies*. Blackwell, Oxford, pp. 467–85

Darian-Smith E 1996 Postcolonialism: a brief introduction. *Social and Legal Studies* **5**(3): 291–99

Darian-Smith E, Fitzpatrick P (eds.) 1999 *Laws of the Postcolonial*. University of Michigan Press, Ann Arbor, MI

Dezalay Y, Garth B 1996 *Dealing in Virtue: International*

Commercial Arbitration and the Construction of a Transnational Legal Order. University of Chicago Press, Chicago and London
Dirks N B 1986 From little king to landlord: property law, and the gift under the Madras Permanent Settlement. *Contemporary Studies in Society and History* **28**: 307–33
Fitzpatrick P 1999 Passions out of place: law, incommensurability, and resistance. In: Darian-Smith E, Fitzpatrick P (eds.) *Laws of the Postcolonial*. University of Michigan Press, Ann Arbor, MI, pp. 39–60
Gandhi L 1998 *Postcolonial Theory: A Critical Introduction*. Colombia University Press, New York
Gates H Jr 1991 Critical Fanonism. *Critical Inquiry* **17** (Spring): 457–70
Loomba A 1998 *Colonialism–Postcolonialism*. Routledge, London and New York
Mamdani M 1996 *Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism*. Princeton University Press, Princeton, NJ
Ngugi wa T 1986 *Decolonising the Mind: The Politics of Language in African Literature*. James Currey, London
O'Hanlon R, Washbrook D 1992 After orientalism: Culture, criticism, and politics in the Third World. *Comparative Studies in Society and History* **34** (1): January, pp: 141–67
Praksh G 1990 Writing post-orientalist histories of the Third World: Perspectives from Indian historiography. *Comparative Studies in Society and History* **32** (2): April, pp. 383–408
Prakash G 1992 Can the 'Subaltern' Ride? A reply to O'Hanlon and Washbrook. *Comparative Studies in Society and History* **34** (1): January, pp. 168–84
Robinson D 1997 *Translation and Empire: Postcolonial Theories Explained*. St. Jerome Publishing, Manchester
Said E 1978 *Orientalism*. Vintage, New York
San Juan E 1998 *Beyond Postcolonial Theory*, 1st edn. St. Martin's Press, New York
Schwarz H, Ray S (eds.) 2000 *A Companion to Postcolonial Studies*. Blackwell, Oxford, UK
Spivak G C 1999 *A Critique of Postcolonial Reason*. Harvard University Press, Cambridge, MA
Stoler A 1989 Rethinking colonial categories: European communities and the boundaries of rule. *Contemporary Studies in Society and History* **31**(1): 134–61
Trivedi H, Mukherjee M (eds.) 1996 *Interrogating Post-Colonialism: Theory, Text and Context*. Indian Institute of Advanced Study, Rashrapati Nivas, Shimla, India
Williams P, Chrisman L (eds.) 1994 *Colonial Discourse and Post-Colonial Theory: A Reader*. Colombia University Press, New York

E. Darian-Smith

Postcoloniality

Postcoloniality is the unprecedented, complex predicament faced by individuals, societies, and their governments after the end of European colonial rule. An increasing number of social scientists, especially in the 1980s and 1990s, have found conceptions of postcoloniality useful for characterizing many contemporary societies. Theories of postcoloniality have emerged especially as more established alternatives—theories of modernization, of underdevelopment, of the Third World, and of revolution—have lost their power to explain the present and predict or guide the future. The premise of most postcoloniality theories is that 'The history of the West and the history of the non-West are by now irrevocably different and irrevocably shared' (Sangari 1995, p. 147).

Writings on postcoloniality range widely in discipline and scholarly methods, and might be typifiable first of all by an attitude, even a mood: disillusionment coupled with commitment, disillusionment with the conspicuous failure of decolonization to produce a world of peace, prosperity, and real liberation, coupled with commitment to diagnose that failure and seek improvements or alternatives. In a British colonial prison during World War II, Jawaharlal Nehru wrote that 'Out of this crucible of war, wherein so much is melting, we hope that something finer will emerge for the West as well as the East, something that will retain all the great achievements of humanity and add to them what they lacked' (Nehru 1982 [1946], p. 81). Thirty-seven years later, as he helped launch the 'Subaltern Studies' school of history, Ranajit Guha declared the 'historic failure of the nation to come to its own' in postcolonial India, its failure as either a bourgeois democracy on the European model or a socialist 'new democracy.' He declared the study of this failure 'the central problematic' for the historiography of India (Guha 1982, p. 7). Similarly, Frantz Fanon wrote in the late 1950s, in the midst of the Algerian anticolonial war, that 'Decolonization is the veritable creation of new men,' that 'the European game has finally ended; we must find something different,' and that 'It is a question of the Third World starting a new history of Man' (Fanon 1968, pp. 36, 312, 315). Thirty-five years later, in the Afterword added to his classic philological study *Orientalism*, Edward Said specifically abjured the idea of 'fixed identities battling across a permanent divide' of Europeans and others (Said 1994, p. 335), in favor of more nuanced multicultural critique of power. While some theorists voice a continuing optimism that, especially in Southern Africa, Eastern Europe, and Southeast Asia, the end of the twentieth century has inaugurated a new age of revolutions (e.g., Comaroff 1996), the dominant tendency in most studies of postcoloniality is critical dissection of utopian anticipations.

To date, discussion of postcoloniality has been more extensive in the Humanities than in the Social Sciences. Critical, often reflexive, theorizing about postcolonial identity and experience has been more extensive than social scientific research into political, economic, and cultural dynamics of postcoloniality. However, controversies in postcolonial theory have provoked fruitful research in anthropology, history, geography, sociology, political science, and other social sciences, characteristically entailing profound

political as well as analytic questions. Good available readers on postcolonial studies include Williams and Chrisman (1994) and Ashcroft et al. (1995). This article centers on efforts to address three powerful antinomies in postcolonial theory, concerning (a) the agency of dominated groups, whether they determine their own fate, (b) the power of the West, whether Western paradigms are inevitable, and (c) the goal of liberation, and when and how postcoloniality might ever end. The article concludes with the tensions between postcoloniality theory and postmodernism, and with efforts to relate postcoloniality and globalization, arguing against portrayals of globalization as the end of postcoloniality, in favor of depiction of the two as intrinsically related.

1. *Can the Subaltern Speak?*

In 1988, literary critic Gayatri Spivak posed the above question as a challenge to the 'Subaltern Studies' school of historians (Spivak 1988). The Subalternists, following Gramsci's theory of class politics in culture, have sought to uncover hidden realities of dominated people's experience and agency in South Asia. The 'Subaltern Studies' project, and Spivak's question, have led to debates and research globally on class, status, culture, and power especially in colonial and postcolonial societies.

The premises of the Subalternists have been, first, that prior histories of colonialism, decolonization, and the postcolonial nation-state have focused on elites, and second, that especially under colonial conditions, elite culture has not been hegemonic, i.e., has not established rituals and routines for social order unchallenged by 'subalterns' or lower class and status, dominated people. In order to explain why a unified postcolonial nation has failed to take shape, Subalternists seek to understand the social experience, collectivities and initiatives of 'subaltern' groups. The original South Asian 'Subaltern Studies' of peasant insurgencies and labor struggles have joined a much broader stream of scholarship on colonial and postcolonial politics, broader not only geographically but also topically, including especially in the 1990s many studies of sex and gender relations, institutions, and imaginaries. However, the issue of the failure of decolonization and the postcolonial nation-states remains central. Efforts have begun to identify positive, existing alternatives to 'the nation' as a form of social organization, not only in South Asia but also in East Asia (notably Duara 1995), the Americas and elsewhere, and also to understand the social legacies of specifically colonial and postcolonial state structures (e.g., in Africa (Mamdani 1996), in Indonesia (Pemberton 1994), and in India (Pandey 1990)).

In the context of research seeking to isolate colonial pathologies and celebrate successful resistances to them, Spivak's question attracted enormous attention because it challenged a second tier of scholarly premises. Phrased with a literary critic's focus on voice, 'Can the Subaltern speak?' quickly metamorphosed into a question about coherent, autonomous agency, and initiative. Is it possible for 'subaltern' people to shape their own society and culture? Precisely as a radical among philologists, Spivak had insisted that representation in a literary or semiotic sense must be reconsidered in connection with representation in politics, representation in the sense of any capacity for a person to be the agent of, to stand for, the will of other people. Do subalterns have any real power to represent themselves, or is that precisely the lack that constitutes them as subaltern?

Social scientists have come to recognize empirical as well as theoretical issues here. In research on postcolonial societies, the question of subaltern political agency has intersected with studies of class and ethnic structure of anticolonial movements and postcolonial governments, and with studies of democracy in postcolonial conditions. While one tendency in contemporary scholarship is to discover unnoticed ways and means by which subaltern groups preserve themselves, maintain autonomous self-control, and even augment their life chances (e.g., J. C. Scott 1985), another trend is to seek aspects of political structure, global and local, that might account for the reproduction of extremities of domination and violence in postcolonial societies (e.g., Tambiah 1996), and a third trend is to seek to reintegrate the questions about colonial and postcolonial subalterns with studies of domination in metropolitan societies (e.g., Stoler and Cooper 1997). Are postcolonial governments especially likely to be bound to the interests of urban, elite, and ethnically marked subpopulations? Are 'subaltern groups' in such societies less likely to exercise routine rights as citizens, and more likely to sustain whole institutional apparatuses of avoidance and resistance, 'informal' economies, even labor migrations kept deliberately invisible to the postcolonial state? No consensus has emerged. However, it does grow increasingly clear at least that, systematically, the nation-state does not 'come into its own' in postcolonial societies, and 'subaltern' classes in them are neither powerless nor as powerful as citizens in the West.

2. *Provincializing Europe*

Is the postcolonial world destined (or doomed) to follow the West's lead? To fit Western molds? Is the history of the West the central axis of world history? If it isn't, then how can sense be made of colonial history? Research assessing the successes and failures of postcolonial societies that sought alternatively to emulate and to refuse Western models (China, Singapore, Japan, Malaysia, Indonesia, India, and many sites in Africa), especially concerning successes of non-Western capitalisms, and the failures of self-proclaimed socialist revolutions, suggest that efforts to understand the West either as 'exceptional,' successful

for particular cultural or material reasons, or as 'universal,' a distillation of necessities, both fail to grasp the more complex dynamics of the actual history of capitalism, colonialism, and the rise of the nation-state. Anticipations kindred to those of Fanon and Nehru, hoping more or less radically for a new order to supplant the civilization of Europe, were common among anticolonial intellectuals of the mid-twentieth century. These anticipations frequently entwined in a variety of ways with socialist and communist projects for social reconstruction and/or revolution, an end to a capitalist era. For this and other reasons, decolonization became deeply enmeshed in the narrow frame of the Cold War, and was dramatically reconceived within it. The question of the future relation of excolonies and excolonizers became, temporarily, reduced to matters of allegiance in a grand global dualism. Social scientists operated under the premise that all 'Third World' new nations were destined as they 'modernized' to join either the 'free' or 'Communist' worlds (see Pletsch 1981). However, by the end of the century much that was taken for granted in this Cold War common sense was in doubt, was under scrutiny, or was repudiated outright, not only concerning the viability of a revolution against capitalism but also concerning modernization and the nation-state.

In the late-twentieth century, great popularity accrued among scholars for Benedict Anderson's vaguely Marxist theory that the nation as an 'imagined community' was central to the century's global culture (Anderson 1983), but at the same time, disquiet about actual postcolonial nations and states led to new critical questions. Partha Chatterjee mounted the most influential challenge to the premise of the nation-state for excolonies. Chatterjee amplified and reconsidered the implications of parts of Anderson's argument, that 'the nation' was a form for politics linked intrinsically to Enlightenment social theories, a political format pioneered in the Americas and Europe, and then 'pirated' in already 'modular' format as colonies were refashioned and new nations built. Insofar as this is what happened, then, decolonization was not actually a break from Western hegemony. 'Europe and the Americas, the only true subjects of history, have thought out on our behalf not only the script of colonial enlightenment and exploitation, but also that of our anticolonial resistance and postcolonial misery' (Chatterjee 1993, p. 5). Chatterjee doubts that 'our imaginations must remain forever colonized' (Chatterjee 1993, p. 5). However, the question of how to decolonize political imagination, and whether or where it has already been decolonized, has been a provocative one for postcolonial studies. While another influential critic has argued that postcolonial critique of Western universalisms has opened a 'third space' outside both local, cultural particularisms, and modernist universalism (Bhabha 1994, pp. 36–9), others have defined postcolonialism instead as the dream that this could be done, 'the need, in nations or groups which have been victims of imperialism, to achieve an identity uncontaminated by universalist or Eurocentric concepts and images' (During 1987, p. 3), or the quest for a new relationship with 'the intimate enemy' (Nandy 1983). In this sense, as Lenin once hoped that a United Front politics in colonies could use bourgeois nationalism as a stepping stone to a proletarian revolution, a postcolonial perspective is proffered by some as the key to a new political future. With or without such a dream, the matter remains of how to understand the impact of 'the West' and its (possibly parochial) universals on one-time colonies in both the past and the present. Touchstones for controversy have included 'world system' models of a globally coercive Western-centered economy, and models of globally hegemonic Western culture. World system economic models have largely been encompassed by approaches emphasizing the irreducibility of both global and local economic forces, and the necessity to understand both. While support increases for models of economic 'globalization' that see global economic forces and relations increasing in their influence over time, the causes of the rise of the West as the pre-eminent capitalist center, and the place of colonialism in that history, are still much debated. Similar controversies have followed models of globally hegemonic Western culture, notably Edward Said's landmark study of the kind of discourse he named 'Orientalism,' discourse that constitutes the East and West comparatively, the East as the inferior antithetical other. Said's study has been criticized, ironically, for itself overriding the possibility of non-Western political discourses, for itself undermining the possibility of a world history centered anywhere but the West. Studies have also begun of processes of cultural globalization and of the crucial process of the localizing intermixture of global and local cultural elements that clearly accompanies the development of global cultural connections, the process Hannerz has named 'creolization' (Hannerz 1992).

At their extremes, globalization models can be as simple in their symmetries as conceptions of postcoloniality can be stark in their asymmetries. Recharacterizations of global culture in terms of flows, links, and hybridities, and inquiry into 'modernity at large' (Appadurai 1996), have added an important complexity to studies of postcoloniality, undermining the starkness and simplicity of the colonizer–colonized dualism, and challenging the premise of continuing cultural superiority of the West. However, the situation of postcoloniality requires attention to asymmetries. Reconsidering the past, present, and future place of the West, postcolonial theorists have grown as skeptical of triumphalist visions of reversal, of a non-Western overcoming of Western culture, as they are critical of intrinsically colonial visions of Western superiority over the rest. In *Provincializing Europe*, Dipesh Chakrabarty suggests a way out of the anti-

nomies generated by efforts to reckon with Western domination (Chakrabarty 2000). The paradox in all postcolonialisms that seek a triumphant postcolonial succession to Europe at history's center, Chakrabarty shows, is that even in such a success, the historical structure would still be 'First in Europe, and then elsewhere.' Europe indeed pioneered the form of historicist explanation that gives time a single processural center, giving capital and its culture of Enlightenment a unified place, time, and pattern of development in Europe, while occluding all inassimilable complexities of its earlier and later history elsewhere. However, Chakrabarty argues for abandoning such historicism, in favor of a method which never confines universal history, and provincializes Europe not from a new center, but instead by understanding Western cultural history, also, as one of creolizations (to use Hannerz's term here). Following Chakrabarty (or the similar, 'anti-Hegelian,' argument of Duara (1995)), one can provincialize Europe, confront and unmake colonial asymmetries, without a triumphalism that will inevitably, fatally, unmask as mimicry of European colonialism's own naive historicist egocentrism. Reflection on how, in recent but past centuries, Europe encountered and came in particular, powerful, but not all-powerful, ways to embody and develop universals of capital, democracy, science, etc., might point toward alternative developments in the past and the present, and possibilities not yet recognized for the future, rather than renewing an imperative to overcome empires already dead. In any case, postcolonial theorists have grown weary of battle with colonial ghosts.

3. *When, and How, Can Postcoloniality End?*

It will now be useful, and easy, to clarify the distinction between postcolonialism and postcoloniality. If postcolonialism is the aspiration for a profound political perspective to emerge from the experience of decolonization, postcoloniality is the unprecedented political predicament that decolonization has delivered many societies. The antinomies that fuel the struggle to constitute a postcolonialist political vision are, in less ideological form, equally inescapable for those without postcolonialist aspirations. Postcolonialist quests for political mission via subaltern standpoints struggle over what to expect from and for subaltern political agency. However, vast, dominated, and politically active populations no doubt exist in subaltern positions in postcolonial societies. Any coherent postcolonialism would seem to require a history both centered on and decentering of Europe, but, no doubt, Europe did once dominate, in various and varying ways, the societies it colonized. The question of how, and whether, that domination can be (or has been) ended is crucial to the third antinomy for postcolonialism, the third aspect of predicament in postcoloniality.

Anticolonial theorist and activist Franz Fanon insisted on the goal of liberation, not mere decolonization, for European colonies. Decolonization was something done for, and to, one-time colonies, where the goal should be actual liberation. In fact, as Fanon perceived, in the aftermath of World War II, the European empires disintegrated with extraordinary rapidity, and (except for the French and Portuguese) with little violence: Fanon perceived why. In the atmosphere of the Cold War, Western powers and especially the newly pre-eminent United States greatly feared radical new states emerging from wars of liberation, and preferred to shape successor regimes and constitutions through a pre-emptive orderly process. As Fanon put it, the message flashing was 'for God's sake let's decolonize quick,' as the pieces of European empires were reconfigured into the nation-state form of 'self-determination' required by the new, US-fashioned United Nations (Fanon 1968, pp. 70, 79).

By many diagnoses, decolonization and not genuine liberation was what the new 'Third World' received. But the problem of what would constitute genuine liberation, and which groups merit it, has become increasingly complex. As Firth (2000) has demonstrated, rhetorics of decolonization and liberation have expanded and proliferated after the breakup of most of the European empires has already been accomplished, expanding to articulate aspirations both of groups within nation–states (Biafrans in Nigeria, Kurds in Iraq) and new kinds of groups (women, indigenous peoples, even environments). If the premise of the United Nations has been that the nation-state as a political form constitutes the ultimate form of self-determination for all peoples, that simple, symmetric premise has been called into question. While most reckonings of postcoloniality focus upon the need to overcome asymmetries past, another approach entirely is to reckon it the onset, rather than the aftermath, of something, the predicament constituted for capital-poor excolonies by their re-establishment within the formal symmetries, and limited liabilities, of the United Nations model (Kelly and Kaplan 2001). From this perspective, the goal of liberation itself has become a very specific trap, as formally independent and autonomous political standing comes systematically to entail perpetual clientage, expanding debts and ongoing substantive subservience to outside agencies and agendas. Like the 'free laborer' perceptively critiqued by Marx as free of all access to means of production, these liberated states are, by their very freedom and formal equality, radically limiting of the life chances of their citizens, especially their subalterns. In an interesting recent controversy, Tambiah (1996) rejects the arguments of Pandey (1990) explaining postcolonial communalist violence in South Asia as a legacy of colonial policies that emphasized and institutionalized communal political differences. Tambiah argues that endemic

political violence is a phenomenon of democracy especially in postcolonial conditions, conditions that engender a spiral of destructive, effervescent leveling crowds. The crowds seek not liberation but leveling, destruction of privileges perceived as unmerited, especially across ethnic lines.

Boundary issues have arisen concerning which societies are postcolonial, especially concerning East Asian societies never formally colonized, and the Americas where the legacy of earlier, Iberian domination and earlier independence movements must be calibrated with the history of European colonies in Africa, Asia, and elsewhere. Sociologist Haejoang Cho has argued forcefully that Korea and other East Asian societies were colonized by the West, with becoming colonized defined as coming to require foreign means, methods, and tools simply to live and to think (Cho 1992–94). For the Americas, Klor de Alva (1992) has argued that few or none of the societies outside the Caribbean are postcolonial in fact, that regardless of mestizaje ideologies, Latin American societies are descended not from colonized peoples but largely from the colonizers. As with most of Latin America, whether the United States is postcolonial depends upon the definition adopted. The question is most interesting when connected to different boundary issues concerning postcoloniality: its relations with postmodernism, and with globalization.

Is the post- in postcolonial the post- in postmodern? (So asked another literary critic, Appiah (1992).) Can either end? Dramatic, relativistic postmodernist proclamations that the Enlightenment universals are dead have been countered, Appiah argued, by the fact of misery in postcolonial societies, leading postcolonial political movements to ground their calls for new reforms in ethical universals. While postmodernism offered a first alternative vision to the stalled modernization logic of the Cold War, at present conceptions of both postcoloniality and globalization offer more dynamic alternative formulations. How, then, are postcoloniality and globalization connected? Some scholars envision globalization as the successor, the late capitalist liberator from postcoloniality, as the globe moves from Western modernity to modernity at large or to alternative modernities, negating all vestiges of asymmetric colonial relations. The counter argument (Kelly 1998) is that postcoloniality and globalization are two sides of the same coin, the coin of American power, phenomena not necessarily of a 'late' capitalism but certainly of the new form capitalism has taken under US leadership in the twentieth century. If this is the case, if decolonization and globalization are the fruit of US-led quests for political self-determination and economic open doors, then Europe has already been provincialized by an explicitly anti-imperial agenda, and the predicaments of postcoloniality are the consequences, precisely, of a postcolonialism already instituted. David Scott has perceptively observed, in the first book with 'After Postcoloniality' in its title, that in the late twentieth century, in the wake of the global resurgence of liberal democracy and the failures of socialist states both Soviet and non-aligned, 'it is no longer so clear what "overcoming" Western power actually means' (D. Scott 1999, p. 14). Perhaps, then, the time has come for dividing 'the West' into its colonial European and twentieth-century American versions. In that case, postcoloniality could end only when American power is as thoroughly confronted as European power has been, and the limitations intrinsic to the formal symmetries of the political present are as fully overcome as have been the formal asymmetries of the colonial past.

See also: Colonialism, Anthropology of; Colonialism: Political Aspects; Colonization and Colonialism, History of; Fourth World; Globalization and World Culture; Hegemony: Anthropological Aspects; Imperialism, History of; Imperialism: Political Aspects; Nationalism: Contemporary Issues; Postcolonial Law; Subaltern History; Subaltern Studies: Cultural Concerns; Third World

Bibliography

Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN

Appiah K A 1992 The postcolonial and the postmodern. In: Appiah K A (ed.) *In My Father's House: Africa in the Philosophy of Culture*. Oxford University Press, New York, chap. 7, pp. 137–57

Asad T 1993 *Genealogies of Religion: Discipline and Reasons of Power in Christianity and Islam*. The Johns Hopkins University Press, Baltimore, MD

Ashcroft W, Griffiths G, Tiffin H (eds.) 1995 *The Post-colonial Studies Reader*. Routledge, London

Bhabha H 1994 *The Location of Culture*. Routledge, New York

Chakrabarty D 2000 *Provincializing Europe: Postcolonial Thought and Historical Difference*. Princeton University Press, Princeton, NJ

Chatterjee P 1993 *The Nation and its Fragments: Colonial and Postcolonial Histories*. Princeton University Press, Princeton, NJ

Cho H 1992–94 *T'alsikminji Chisikin ui Kullikki wa Samilkki* [*Reading Texts, Reading Life: The Point of View of a Post-Colonial Intellectual*], 3 vols. Ttohaniuimunhwa Press, Seoul

Comaroff J 1996 Ethnicity, nationalism, and the politics of difference in an age of revolution. In: Wilmsen E N, McAllister P (eds.) *The Politics of Difference: Ethnic Premises in a World of Power*. University of Chicago Press, Chicago, pp. 162–83

Duara P 1995 *Rescuing History from the Nation: Questioning Narratives of Modern China*. University of Chicago Press, Chicago

During S 1987 Postmodernism or post-colonialism today. *Textual Practice* **1**: 32–47

Fanon F 1968 *The Wretched of the Earth*. Grove Press, New York

Firth S 2000 Decolonization. In: Borofsky R (ed.) *Remembrance of Pacific Pasts: An Invitation to Remake History*. University of Hawaii Press, Honolulu, HI
Guha R 1982 On some aspects of the historiography of colonial India. In: Guha R (ed.) *Subaltern Studies I*. Oxford University Press, Delhi, pp. 1–8
Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York
Kelly J D 1998 Time and the global: against the homogeneous, empty communities in contemporary social theory. *Development and Change* **29**: 839–71
Kelly J D, Kaplan M 2000 *Nation and Decolonization*. University of Chicago Press, Chicago, IL
Klor de Alva J 1992 Colonialism and postcoloniality as (Latin) American mirages. *Colonial Latin American Review* **1**: 3–23
Mamdani M 1996 *Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism*. Princeton University Press, Princeton, NJ
Nandy A 1983 *The Intimate Enemy: Loss and Recovery of Self Under Colonialism*. Oxford University Press, New York
Nehru J 1982[1946] *The Discovery of India*. Oxford University Press, New Delhi
Pandey G 1990 *The Construction of Communalism in Colonial North India*. Oxford University Press, New Delhi
Pemberton J 1994 *On the Subject of 'Java.'* Cornell University Press, Ithaca, NY
Pletsch K E 1981 The three worlds, or the division of social scientific labor, circa 1950–1975. *Comparative Studies in Society and History* **23**: 565–90
Said E 1994 *Orientalism*. Vintage Books, New York
Sangari K 1995 The politics of the possible cultural critique 7. In: Ashcroft W, Griffiths G, Tiffin H (eds.) *The Post-colonial Studies Reader*. Routledge, London, pp. 143–7
Scott D 1999 *Refashioning Futures: Criticism After Postcoloniality*. Princeton University Press, Princeton, NJ
Scott J C 1985 *Weapons of the Weak: Everyday Forms of Peasant Resistance*. Yale University Press, New Haven, CT
Spivak G C 1988 Can the subaltern speak? In: Nelson C, Grossberg L (eds.) *Marxism and the Interpretation of Culture*. Macmillan Education, Basingstoke, UK, 271–313
Stoler A L, Cooper F 1997 Between metropole and colony: Rethinking a research agenda. In: Cooper F, Stoler A L (eds.) *Tensions of Empire: Colonial Cultures in a Bourgeois World*. University of California Press, Berkeley, CA, pp. 1–56
Tambiah S J 1996 *Leveling Crowds: Ethnonationalist Conflicts and Collective Violence in South Asia*. University of California Press, Berkeley, CA
Williams P, Chrisman L (eds.) 1994 *Colonial Discourse and Postcolonial Theory: A Reader*. Columbia University Press, New York

J. D. Kelly

Post-Keynesian Thought

1. Introduction

For Joan Robinson (1978; *CEP*, in reprinted Vol. 5, 1979, p. 210), 'the expression *post-Keynesian*… [applied] to an economic theory or method of analysis which takes account of the difference between the future and the past': a helpful point from which to start but it must be remembered that the approaches to political economy reflecting Post-Keynesian thought are there for historical reasons and because of logical associations. Post-Keynesianism is a broad church. Overlaps at each end of a long spectrum of views are marginal, reflecting little more than hostility towards mainstream neoclassical economics and methodology, and conventional forms of Keynesian economics as opposed to the economics of Keynes. Some Post-Keynesians attempt to synthesize the principal strands. Others regard this as misguided. Post-Keynesianism should use a situation—and issue-specific method, a 'horses for courses' approach, itself an all-embracing structure at the methodological level.

In this essay historical origins are discussed first; it then deals with approaches to fundamentals issues political economists have tackled since the Physiocrats and Adam Smith theories of value, distribution, price-setting, accumulation, growth and the cycle, and the role of money and financial intermediaries, especially whether money is endogenous or exogenous. Post-Keynesian thought belongs to the tradition that the *raison d'être* of economics is its bearing on policy. Stress is put on the role of institutions in determining economic and political processes, the political philosophies of Post-Keynesians and the policies following from their theories.

All these issues, the methodologies implied in them, the particular 'vision' of the economy associated with each strand, are intertwined. This article concentrates on origins, pioneers, and issues. For a full cast list see, for example, Eichner and Kregel (1975), Hamouda and Harcourt 1988; reprinted in Sardoni 1992, Arestis 1992, Lavoie 1992.

2. Classical Roots

Keynes's new theory reached back over neoclassical economics to join up with the pre-occupations and approaches of our classical forebears. Much that was integral to Keynes's system, especially its Marshallian offshoots, was jettisoned, at least by Post-Keynesians sympathetic to the viewpoints of the classical economists and Marx. This rules out two prominent American Post-Keynesians, Sidney Weintraub and Paul Davidson, whose structures are amalgams of Marshall and Keynes and who are hostile to Marx, Michal Kalecki, and Piero Sraffa.

An important aspect of classical economics for Post-Keynesian thought is the organizing concept of the surplus—its creation, extraction, distribution, and use. It is the core of their system and its modern counterpart, dovetailing logically with a view of economic processes as dynamic, evolving, progressing organic systems, not the more static allocative process characterizing neoclassical economics until neoclassical growth theory emerged in the 1950s. Linked to the

concept is the notion of a class society, sometimes people, sometimes income, sometimes overlapping. Classes perform different functions. Each is treated as homogenous, with no need to start from the isolated individual economic agent of modern analysis.

Associated with the surplus concept is that of the long-period position, containing the centers of gravitation of the system (Smith's natural prices, Marx's prices of production, Marshall's long-period normal equilibrium prices (and quantities)). With them comes the view that *general* theory is analysis of interrelationships of dominant, persistent forces in society. Modern counterparts are Sraffa's work on value and distribution and Pierangelo Garegnani's (and others') long-period interpretation of Keynes.

3. *The Marshallian-Keynesian Roots*

Turning to the Marshall-Keynes connection, Marshall's system, as Keynes understood it and within which he perceived himself as operating up to *A Treatise on Money* (1930; *C.W.*, Vols. V, VI, 1971) is noted. Basic ingredients were emphasis on the long period, the real and money dichotomy, the logically prior need for a Say's Law long-period equilibrium position for the Quantity Theory of Money to explain the general price level. Keynes liberated himself from this to write the theory of a monetary production economy—*The General Theory* (1936; *C.W.*, Vol VII, 1973).

Effective demand occurs at the intersection of the aggregate demand and supply functions, usually implying involuntary unemployment. The money rate of interest (which clears the money, not the saving and investment market) rules the roost. Uncertainty is a permanent environment for important economic decisions, especially those concerning investment and holding financial assets for speculative and other purposes. The consumption function is stable, the investment function, volatile; they are connected by the Kahn-Meade multiplier. Finally (an aberration for most Post-Keynesians), the general price level is explained by short-period Marshallian analysis at the level of economy as a whole. These are the jumping-off points for American Post-Keynesians especially.

4. *Keynes's Theory of Accumulation*

Accumulation theories under our rubric start from, or incorporate, the ingredients of Keynes's theory. The details are usually stringently criticized. The recipes emerging mix the same ingredients in more acceptable ways. Within its own framework Keynes's theory was unacceptable. Keynes confused accumulation's stock and flow aspects by naming his principal concept the marginal efficiency of capital (*MEC*). Capital is a stock; he was mainly concerned with investment, a flow, and so the marginal efficiency of investment (*MEI*).

Keynes's arguments why higher rates of accumulation are associated with lower *MEI* values, with short-period flow equilibrium achieved when $MEI = i$, (i = the rate of interest), were far-fetched or non-clinching. Keynes assumed rising marginal costs and (usually) marginal-cost pricing. At higher values of accumulation in a given situation, there are higher supply prices and lower *MEI* values. Individual free competitors have, in effect, rational expectations; it is as if they use the eventual equilibrium, not the market price of capital goods in their calculations. Otherwise, their combined actions could not bring about that overall rate of production of capital goods which establishes the appropriate supply price and *MEI* value equal to i.

Second, Keynes argues that because we do not know the future, we project the present into the future unless there are good reasons for not doing so. Higher investment levels are associated with higher sales, prices, profits, output, and employment. Keynes's investment analysis assumes that expected longer-period demand curves for the products concerned remain stable while future short-period supply curves associated with possible levels of investment move to the right, implying lower expected prices, quasi-rents and *MEI* values for investment projects now. If Keynes were true to himself, expected demand curves would move too. Lower values of prices, etc. would not necessarily follow. Keynes used a static analysis to analyze a dynamic process, a criticism first made by Kalecki reviewing in Polish *The General Theory* (1936, *C.W.*, Vol. 1, 1990, pp. 223–32), subsequently by Joan Robinson 1951–1980 and Tom Asimakopulos 1971.

They reacted by setting out the two-sided relationship between accumulation and profitability—actual investment is a major determinant of current profits; current profits, a major determinant of expected profits, which determine planned accumulation in situations of given financial conditions and long-term expectations. This is the basis of Kalecki's theory of distribution, accumulation, growth and the cycle, and Joan Robinson's famous banana diagram, Robinson (1962, p. 48).

Pasinetti (1997) remains true to Keynes, pointing out that the *MEI* soon became identified with the marginal product of capital, that investment was treated as a 'deepening' process responding to different i values. Pasinetti argues Keynes made no such claim—he argued that in a given situation, lower i values make more of the existing investment projects seem profitable *regardless of capital-intensity orderings*. Davidson (1972), too, accepted Keynes's analysis, using Keynes's alternative presentation to describe investment projects by their *NPVs* at ruling i (s). He used Keynes's spot, forward, and futures analysis to go from investment possibilities—prices exceed supply prices—to determine equilibrium flows per period.

5. *Post-Keynesian Theories of Value, Distributions, and Price Setting*

Post-Keynesian theories of value, distribution, and price-setting by, usually, an oligopolistic price-leader firm are now examined. Both sets of theories arise from dissatisfaction with supply and demand theories. The theories criticized and proposed as alternatives are macroeconomic. They arose from the empirical observation *cum* theoretical deduction of classical political economy that, with free competition, there is a tendency to a uniform rate of profit in all activities. Two questions arose: what is the origin?, and what determines the size of the system-wide rate of profits (Sraffa's usage) (Sraffa 1960) to which individual rates tend to be equal? The most explicit, refined answers were Marx's. He emphasized that the ultimate source of value is the difficulty of reproduction in a system of production of commodities by means of commodities (as opposed to utility and scarcity in the alternative tradition). Reconciliation is needed between the embodied labor values of commodities in the sphere of production and prices of production in the sphere of distribution and exchange. The latter contain the overall rate of profits and are the centers of gravitation of the system.

The other major alternatives to the neoclassical marginal productivity explanation of returns to capital and labor and shares of wages and profits are those Kaldor (1955–56) called Keynesian. He was not the first to set forth such a theory. Kalecki (1933, reprinted in 1971, pp. 78–9) had long before: 'capitalists may decide to … invest more in a given period than in the preceding one, but they cannot decide to earn more … therefore, their investment … decisions … determine profits, … not vice versa.' Kaldor located their origins in the widow's cruse story of *A Treatise on Money* (1930; *C.W.*, Vol. V, 1971, p. 125) Joan Robinson (1977; *C.E.P.*, Vol. V, 1979, pp. 184–96) provided a lucid account of Kalecki's theory. Consider the case of no consumption by rentiers, no saving by wage-earners and no government or overseas sectors. Profits (II) exactly equal investment (I), causation runs $I \rightarrow II$. Kaldor's version is the best known. Common features are different saving behavior at the margin as between wages and profits (and/or wage-earners and profit-receivers) and the Keynesian/Kaleckian proposition that investment leads and saving responds. Kaldor's theory is, paradoxically, long-period and relates to a fully employed economy. Kaldor argued, never coherently, that a growing economy must be fully employed in the long-period sense. Planned saving and investment (and their share of full employment income) are equalized when the income distribution is such that the matching amount of saving is forthcoming.

The process bringing equality depends upon Kaldor's (claimed) empirical observations that while short-term money prices and money-wages are sticky, so that changes in income equalize saving with investment, longer-term prices are more flexible than wages. These differential changes continue until equality is secured, implying that modern wage-earners have real wage levels so far above the classical-Marxian subsistence levels that they passively accept what remains after profit-receivers have first bite of the cherry: the opposite of the older scenarios in which wages had to be met and the surplus available for profits (and thus accumulation) is what is left.

The other theories under this heading are short-period and not constrained to be at full employment. Saving is equalized with investment by changes in activity and income distribution as profit margins change relative to wage costs, a rest state usually exhibiting involuntary unemployment. (In the most developed models there are roles for the government and overseas sectors.)

The most rigorous presentation of modern classical theory is *Sraffa's Production of Commodities by Means of Commodities* (1960). Sraffa's model is formally similar to Leontief's input-output model; the *conceptual* interpretation is different. With production with a surplus and explicit wage rate (w) and rate of profits (r), Sraffa shows we must know from outside the value of one of the distributive variables. (The classical economists and Marx chose w, Sraffa chose r.) The prices of production and the value of the other distributive variable are simultaneously determined. The prices of production are centers of gravitation in the particular circumstances of the time, around which market prices fluctuate—or perhaps on which they converge. (Modern work suggests they do not necessarily do *either*.)

Sraffa starts with a circulating, single-use commodity model, then takes us through non-renewable and/or addable to resources (land and rent) and joint production (to analyze durable capital goods, not wool and mutton). He ends with the choice of techniques in the investment decision, showing the possibility of reswitching, destroying the conceptual foundation of a well-behaved demand curve for 'capital' in the alternative tradition.

6. *The Size of the Mark-up*

Alfred Eichner's version (1976) is the most widely known variant of the theory of mark-up size (apart from Kalecki's 'degree of monopoly' theory). Eichner related its size to the investment plans and financial requirements of oligopolistic firms. As his investment function is flawed, we present here as representative Adrian Wood's (1975) long-period steady-state model of an oligopolistic firm, steady-state because expectations of the values of relevant variables are realized. It is a partial analysis—with a given level of aggregate demand, the firm strives for its optimum share of

demand for the industry's products at the expense of other firms in the industry.

The firm wants its sales revenue to grow as fast as possible, subject to two constraints formalized as an opportunity and a finance frontier. The former reflects that there is eventually a trade-off between profit margin size on turnover and sales revenue growth. Increasing growth requires the margin to be squeezed by a price cut, a rise in sales costs, or both. Assuming one best-practice technique, and given external financial conditions, each growth rate of sales requires a particular margin to finance the capacity to make it possible. The two frontiers' intersection defines the margin associated with highest growth rate. Introducing choice of techniques defines a family of interacting frontiers. Because opportunity frontiers move out at a decreasing rate—greater investment per unit of sales lowers costs at a decreasing rate—while finance frontiers fan out proportionately—the margins, growth rates locus forms a concave curve containing a maximum margin, growth rate combination.

Wood's analysis is in Golden Age logical time—expectations realised, no incompatibilities present. Harcourt and Kenyon (1976, reprinted in Sardoni 1992, pp. 48–66) set the problem in historical time—disappointed expectations determine what happens next—using Salter's vintage analysis. The logical, historical time distinction is due to Joan Robinson (1962, pp. 23–9). Setting the price which finances the accumulation program simultaneously determines the accumulation needed to satisfy expected future sales and still profitable existing capacity (in the sense of covering variable costs) able to contribute to future sales. Frederic Lee (1998) tells the Post-Keynesian price theory story from the earliest times to the present.

7. *Post-Keynesian Accumulation: Moses and the Profits*

Post-Keynesian theories of accumulation were stimulated by Harrod's immediate pre- and post-second world war contributions towards a dynamic economics 1939; 1948. Harrod in effect rediscovered the systemic instabilities of capitalism Marx brought out in his reproduction schemas analysis. Marx asked: what conditions must be satisfied, period by period, in order that aggregate demand and supply, and their compositions, match? This threw up the impossibility of these being achieved by capitalists doing their own things with regard to production, employment, and accumulation.

Harrod obtained the same result by bringing out the long-period destabilizig effects of a momentary gap between planned saving and investment when short period stabilizing effects are abstracted from by assuming the immediate achievement of investment planned for each short period. The growth rate of income allowing planned saving to equal investment might be such that if business people had known it, they would not have planned those investment levels. The saving and investment functions are so related that excess demand (supply) situations are to the right (left) of the equilibrium position; the system gives the wrong signals as to what to do next. Unless the system is growing at the warranted rate (g_w) (the growth rate allowing expectations and plans to be realized), it moves away from it—all regardless of whether g_w coincided with the natural rate of growth (g_n), the growth rate implied by the economy's supply-side characteristics.

Solow and Swan interpreted Harrod as assuming a fixed capital-output ratio, and wondered whether this was responsible for his disturbing results. As pioneering Keynesians, they assumed an all-wise government removed effective demand puzzles. They asked whether in a long-period sense, Marshall's 'dynamical principle of substitution seen ever at work' allowed a competitive price mechanism to give appropriate price signals, causing the production methods chosen by accumulators to lead g_w towards g_n. The answer is 'yes' in the simplest one, all-purpose commodity models. These results attracted the Cambridge, UK growth theorists' attention. Already suspicious of smooth continuous substitution possibilities, they wished to bring in the 'book of blue prints' approach (long 'at home' at MIT and elsewhere within a different context). The Cambridge UK theorists were loathe to leave out aggregate and effective demand. Their writings mostly remained within the confines of steady-state growth models, learning to walk before running, except for Kaldor who used his distribution mechanism to equalize g_w with g_n.

Kalecki, Josef Steindl, and Richard Goodwin followed a different path, one eventually congenial to Kaldor and Joan Robinson. Kalecki initially developed trendless cycles; in his last paper (1968, reprinted in Kalecki 1971, pp. 165–83) on these themes, trend and cycle were indissolubly mixed: 'the long-run trend [is] but a slowly changing component of a chain of short-period situations…[not an] independent entity' [Kalecki 1968, reprinted in 1971, p. 165].

This view was developed independently by Goodwin, reaching maturity in 'A Growth Cycle,' Goodwin (1967). He used the Volterra prey-predator model; the analogy of 'the symbiosis of two populations—partly complementary, partly hostile—is helpful in…understanding…the dynamical contradictions of capitalism, especially when stated in a…Marxian form' (Goodwin 1967, reprinted in 1982, p. 167). He analyzed the wages, profits fight, and feedbacks, spawning an expanding literature.

Goodwin integrated effective demand and production interdependence in Goodwin and Punzo (1987). This impressive eclectic book reflects influences from Marx, Schumpeter, Keynes, von Neumann, Joan Robinson, Sraffa, and Kalecki, the developments of catastrophe theory, the concept of 'bifurcation,' and the

Volterra prey-predator biological analogy. Goodwin concentrated on the nature of evolutionary structures experiencing from time to time large jumps and breaks, his key to the cyclical development of economies characterized by production interdependencies.

Joan Robinson (1962) synthesized these ideas in her banana diagram. From Kalecki she derived a relationship between actual accumulation and achieved profitability; from Keynes she derived the 'animal spirits' function relating desired accumulation to expected profitability, itself a function of achieved profitability, in a situation of given financial conditions and long-term expectations. With plausible conjectures about the shapes and positions of the two curves, iteration takes the economy to a stable intersection point where the expected and desired are achieved, her version of Harrod's g_w. It is not a solution of Harrod's (or Domar's) problem; there is no reason why g_w coincides with g_n. Nor are there any mechanisms to take it to g_n. Even if the economy attains g_w, nothing ensures that it will not be driven away in the future—neither relationship can be expected to remain stable.

Discussing *The General Theory* twenty five years on, Joan Robinson signaled the change of method lying behind her analysis in 1956 and 1962.

'The short period is here and now, with concrete stocks of the means of production in existence. Incompatibilities in the situation … determine what happens next. Long-period equilibrium is not at some date in the future: it is an imaginary state of affairs in which there are no incompatibilities in the existing situation.' Joan Robinson (1962, reprinted in *C.E.P.*, Vol. 3, 1965, p. 101).

These developments constitute an attack on the procedure of setting up theoretical issues as searches for existence and uniqueness of equilibria, themselves determined by factors independent of those responsible for stability, local, and global. Similarly, the applied procedure of separating trend from cycle with independent factors responsible for each is also discarded.

8. *Post-Keynesian Method*

There are two views on method within historically defined Post-Keynesian thought. Increasingly, commentators put neo-Ricardians in a class of their own; The author does not agree. Sraffa's critique of neoclassical economics and positive contributions are integral parts of the historical and logical developments, especially when his underlying Marxism is recognized. The neo-Ricardian stance (on method) is identified as common to all strands of political economy until the 1920s when neoclassical economists, including Hayek, sensed the incoherence of the concept of the marginal product of capital and its relationship to the long-period method. They reacted by changing the questions and method to the temporary equilibrium analysis of, for example, J.R. Hicks's *Value and Capital*. Neo-Ricardians argue that *general* theory can only be written about the characteristics of long-period positions reflecting the ultimate outcomes of relationships associated with persistent forces. The economist's role is to identify these, making rigorously explicit their interrelationships. This, they argue, was common to Smith, Marx, and also Marshall who fitted the 'new' supply and demand theories into the old method, making long-period equilibrium prices and quantities the heart of the *Principles*. Therefore, Keynes's theory must be long-period to be a true revolution Garegnani, Eatwell, Milgate, Rogers, and Krishna Bharadwaj refer to passages in *The General Theory* supporting this. Short-period theory has no role in its own right, as opposed to situation-specific analysis at different levels of abstraction.

For Kalecki, Kaldor, and Joan Robinson and their followers, and Paul Davidson and his (the theories differ considerably), the long-period method is a non-starter for descriptive analysis. (For doctrinal debates even Joan Robinson granted it a role. If conjectures or concepts are shown to be incoherent under ideal conditions, this is a legitimate critique of approaches, conceptions, intuitions, and theories.) For descriptive analysis, we must start from a given short period with inherited historical circumstances and accompanying expectations. Long-term expectations and factors are relevant for some current decisions, e.g. investment, pricing; short-term expectations and factors are relevant for others, e.g. production, employment. The story unfolds from one short period to another. (We have simplified: some current activity results from past decisions; some current decisions are relevant for future activity. This led Keynes to abandon period analysis, despairing of finding a determinate time unit into which all relevant interrelationships could be fitted.) There is an element of convergence in the discipline recently because this philosophy underlies work on path-dependence even when the *theory* is neoclassical.

These new developments link back to old approaches to method. Keynes, following Marshall, argued that which variables were regarded as endogenous (determined) and which were exogenous (determining) were arbitrary on any absolute criteria. The crucial decider was the issue in hand.

'The object of our analysis is, not to provide a machine, or method of blind manipulation, [furnishing] an infallible answer, but to provide … an organized and orderly method of thinking out particular problems; and, after we have reached a provisional conclusion by isolating the complicating factors one by one, we … go back on ourselves and allow … for the probable interactions of the factors amongst

themselves.... the nature of economic thinking.' Keynes (1936; as in *C.W.*, Vol VI, 1973, p. 297).

Marshall hoped we would get definite, if partial, results and, if we went right round the economy, would be able to bring all results together in a full, overall picture. He forgot that the procedure was inconsistent with his deeper vision that economic processes were akin to systemic interrelated biological processes. (Jan Kregel argues that the *problem* is intractable.)

Sheila Dow (1991) stresses the unifying methodological aspect of Post-Keynesian thought, justifying diverse methods. She recalls to mind Keynes on method and theorizing: there is a continuum of languages which runs from poetry and intuition through lawyer-akin arguments to mathematics and formal logic. Each has a role depending upon the issues, or aspects of issues, discussed. Keynes's emphasis was championed by George Shackle, his favourite example was Keynes's use of 'sentiment.' It could not be defined mathematically, but it gave insights, conjured up atmosphere, invoked the environment of situations analyzed. Kregel (1976) identified in *The General Theory* three models of economic reality of differing degrees of complexity, the most complex being that of shifting equilibrium. The methods were common to Keynes and Post-Keynesians when modeling decision-making in uncertain environments.

Post-Keynesians following Goodwin, Kalecki, and Joan Robinson eschew the Marshallian (or even Keynesian) method; others combine a similar 'vision' of society with explicit use of the Marshallian/Keynesian method or, even earlier, the classical economists' implicit method. Outstanding examples include Pasinetti whose many years work on a multi-sectoral growth model came to fruition in 1981 (with a simple exposition in 1993). Pasinetti integrates Keynes's macroeconomic aggregates with the production interdependence of the classicals, Marx and Sraffa. He sets out the consistency conditions required to maintain full employment growth in a multi-sector model in which the demands for commodities over time follow their respective Engels curves, technical progress occurs and, in later stages of the analysis, the economy is open, trading with and borrowing from or lending to, the rest of the world. The best-known exposition of his insights is Pasinetti (1962); he obtained the 'remarkable result,' $r = g/s_c$. It has proved extremely robust as complications have been added—international trade, government, financial intermediaries, endogenously determined saving ratios.

Another major contribution is Marglin's contribution to the conflict inflation literature (1984). (Rowthorn (1977) derived the same results.) Marglin has a Keynesian-Marxian 'vision' but is Marshallian in method. His aim is to establish long-period 'equilibrium' relationships—short-period stations approach long-period crosses which stay 'put' until and after they are reached. His principal result is that sustained rates of inflation allow disappointed aspirations of wage-earners and capitalists to be shared in uneasy but sustainable truces.

Post-Keynesian thought starts from real world observations, for example, Kaldor's famous 'stylized facts': broad empirical generalizations holding, in a rough and ready way, for long runs of historical time, requiring situation-specific theories to explain them. Kaldor pioneered the theory of cumulative causation, a deep idea learnt from Allyn Young (and Gunnar Myrdal) Young alerted Kaldor to the pervasive existence of dynamic increasing returns in economic life. Kaldor exploited these illuminations, and inter-regional relationships, to analyze specific periods of economic history.

A wolf pack analogy illustrates the contrast between mainstream axiomatic, equilibrium theory and Post-Keynesian cumulative causation theory. The former thinks of a wolf pack running along. If some wolves stray, powerful forces return them to the pack. The latter identifies powerful forces making those leaving the pack get further ahead or fall further behind, at least for long periods of time—hence the rejection of the equilibrium concept in descriptive analysis.

9. *Endogenous Money*

Keynes, before and after *The General Theory*, was an endogenous money person, arguing that the demand for money, especially for credit, helped determine the interest rate and the supply of money. Was his stance in *The General Theory*, where he is interpreted as having an exogenous money supply with the interest rate, the price allowing it to be voluntarily held, an aberration? The answer (given most recently by Sheila Dow (1997) is that the money supply is *given*, not *exogenous*. By 1937 he reverted to type when discussing the finance motive for holding money (read, demanding credit)—he had banks creating credit in response to demands for finance for investment projects, a base from which start Post-Keynesian debates on money and credit demand and supply.

Within these discussions, there are claims that the money supply is entirely demand (not supply) determined, summarized in the phrase. Horizontalists versus Verticalists, see Moore (1988). Moore's position is extreme. Most reject the verticalist interpretation, a stance attributed to much mainstream theory, but allow important roles for credit rationing and rising interest rates in determining outcomes in the financial assets market. It is almost universally argued that meeting demand for funds for investment and now, with credit for all, consumption too, creates deposits. This contrasts with the mainstream view that deposits of present and past saving initiate potential lending by financial intermediaries. The former view was fervently expressed by Keynes: 'The investment

market can become congested through shortage of cash, never…through shortage of saving.' (*C.W.*, Vol. XIV, 1973, p. 222). Kalecki independently arrived at the same view. Post-Keynesians start from here, responding to new institutional developments and refining, extending the analysis. They reflect the major change James Meade attributed to Keynes: Keynes changed us from looking at the world as a saving dog wagging an investment tail to one in which an investment dog wagged a saving tail.

Within this set of discussions, some stress the stock aspects of money and financial assets, others, the flow aspect of credit as finance for important expenditures. Associated with this distinction are Post-Keynesians concentrating on one period only in their analysis of money—accomodationists. They contrast with structuralists, stressing period by period analysis in the manner of the mature Kalecki and John Hicks. The upshot is to preserve Keynes's liquidity preference theory when explaining demand for money and financial assets, levels and patterns of interest rates and banks extending credit guided by their liquidity preference, so that not all demand is necessarily accomodated, see Cottrell (1994).

10. Policy

The connection between Post-Keynesian anlaysis and the real world includes a healthy preoccupation with policy. The mean Post-Keynesian's policy proposals match the mean Bastard Keynesian's—a slightly left of center package deal of monetary, fiscal and incomes policies, combined with a longer-term commitment to freer trade while preserving protective positions until 'the time is ripe' and preference for fixed exchange rates (with institutional means allowing orderly realignments now and then). Revamping Bretton Woods's structures, to remove built-in contractionary biases, guides suggestions on the international stage. Deregulation of financial markets and 'freeing up' of exchange rates unleashed an era of instability following the tranquil experience of 'the Golden Age of Capitalism.' This has led to attempts, e.g. the Tobin tax on foreign exchange transactions, to curb the excesses of speculative movements in the foreign exchange markets, stock exchanges, and property markets.

Where Post-Keynesians feel they have the upper hand is in their argument that cumulative causation processes characterize the workings of markets and economic systems. If such processes operate it is harder to argue for beneficial systemic effects from speculation—that it reduces fluctuations and gets the market (system) to underlying long-period equilibrium positions more quickly than otherwise would be the case. If they are not there, speculation makes matters worse. We are led to middle way programs between ruthless *laissez faire* (but not that competitive) capitalism and authoritarian, inefficient, centrally planned economies. Heinrich Bortis (1997) profoundly analyzed the philosophical, political and economic aspects of these developments. He synthesized three major stands of Post-Keynesianism: the neo-Ricardian approach to capture the longer term aspects of economies' developments, the Kaleckian-Robinsonian, their cyclical aspects, and the American Post-Keynesians, short-period behavior in product and money markets under uncertainty. To this he allied a democratic socialist platform, with philosophical foundations and a comprehensive outline of policy, nationally and internationally. Another stalwart is John Cornwall. Over the years he developed his original theoretical critique of Harrod for supposing that g_w and g_n were independent of each other and supplemented his analysis of the implications for growth of systemic supply and demand interdependence, with a steady stream of down-to-earth, humane recommendations for institutional reforms, see Harcourt and Monadjemi (1999).

Bortis and Cornwall were influenced by Kaldor. Kaldor's last book (1996), contains his mature views on how the world works: policy proposals are in the last chapter. He has long emphasized the importance of industrial increasing returns and different pricing behavior as between the (export) products of primary producing countries and countries producing industrial goods. He combined policies for relative stability in exchange rates worldwide with international buffer stock schemes for primary commodities. The latter serve the dual purpose of reducing fluctuations in prices and incomes which damage the economies producing these commodities (and those buying them) while simultaneously providing worldwide liquidity for trade and capital movements, a provision free of disadvantages arising when one major country's currency performs this role, as happens now.

See also: Keynes, John Maynard (1883–1946); Keynesianism and State Interventionism; Marxian Economic Thought; Political Economy, History of; Robinson, Joan (1903–83)

Bibliography

Arestis P 1992 *The Post-Keynesian Approach to Economics*. E. Elgar, Aldershot, UK

Asimakopulos A 1971 The determination of investment in Keynes' model. *Canadian Journal of Economics* **4**: 382–8

Bortis H 1997 *Institutions, Behavior and Economic Theory. A Contribution to Classical-Keynesian Political Economy*. Cambridge University Press, Cambridge, UK

Cottrell A 1994 Post-Keynesian monetary economics. *Cambridge Journal of Economics* **18**: 587–605

Davidson P 1972 *Money and the Real World*. Macmillan, London

Dow S 1991 Keynes's epistemology and economic methodology. In: O'Donnell R M, Keynes J M (eds.) Philosophy, Economics and Politics. St Martin's Press, New York, pp. 144–167
Dow S 1997 Endogenous money. In: Harcourt C C, Riach P A (eds.) *A 'Second Edition' of The General Theory*. Routledge, London, Vol. 2, pp. 61–78
Eichner A S 1976 *The Megacorp and Oligopoly: Micro Foundations of Macro Dynamics*. Cambridge University Press, Cambridge, UK
Eichner A S, Kregel J A 1975 An essay on Post-Keynesian theory: a new paradigm in economics. *Journal of Economic Literature* **XIII**: 293–314
Goodwin R M 1982 *Essays in Dynamic Economics*. Macmillan, London
Goodwin R M, Punzo L F 1987 *The Dynamics of a Capitalist Economy: A Multi-sectoral approach*. Basil Blackwell, Polity Press, Oxford
Harcourt G C, Monadjemi M S 1999 The vital contributions of John Cornwall to economic theory and policy. In: Setterfield M (ed.) 1999. *Growth, Employment and Inflation. Essays in Honour of John Cornwall*. St Martin's Press, New York
Harcourt G C, Riach P A (eds.) 1997 *A 'Second Edition' of The General Theory*, 2 Vols. Routledge, London
Harrod R F 1939 An essay in dynamic theory. *Economic Journal* **49**: 14–33
Harrod R F 1948 *Towards a Dynamic Economics. Some Recent Developments of Economic Theory and their Application to Policy*. Macmillan, London
Kaldor N 1955–56 Alternative theories of distribution. *Review of Economic Studies*. **23**: 83–100
Kaldor N 1996 *Causes of Growth and Stagnation in the World Economy*. Cambridge University Press, Cambridge, UK
Kalecki M 1936 Some remarks on Keynes' theory. In: Osiatynski J (ed.) 1990 *Collected Works of Michal Kalecki*, Vol I. Oxford University Press, Oxford, UK, pp. 223–32
Kalecki M 1971 *Selected Essays on the Dynamics of Capitalist Economies 1933–71*. Cambridge University Press, Cambridge, UK
Keynes J M 1971–89 *The Collected Writings of John Maynard Keynes,* 30 Vols. for the Royal Economic Society. Macmillan, London
Kregel J A 1976 Economic methodology in the face of uncertainty: the modeling methods of Keynes and the Post Keynesians. *Economic Journal* **86**: 209–25
Lavoie M 1992 *Foundations of Post-Keynesian Economic Analysis*. E. Elgar, Aldershot, UK
Lee F S 1998 *Post Keynesian Price Theory*. Cambridge University Press, Cambridge, UK
Marglin S A 1984 *Growth, Distribution and Prices*. Harvard University Press, Cambridge, MA
Moore B J 1988 *Horizontalists and Verticalists: The Macroeconomics of Credit Money*. Cambridge University Press, Cambridge, UK
O'Donnell R M (ed.) 1991 *Keynes as Philosopher-Economist*. St Martin's Press, New York
Pasinetti L L 1962 'Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* **29**: 267–79
Pasinetti L L 1981 *Structural Change and Economic Growth: A Theoretical Essay on the Dynamics of the Wealth of Nations*. Cambridge University Press, Cambridge, UK
Pasinetti L L 1993 *Structural Economic Dynamics. A Theorey of the Economic Consequences of Human Learning*. Cambridge University Press, Cambridge, UK
Pasinetti L L 1997 The Marginal Efficiency of Investment. In: Harcourt G C, Riach P A (eds.) *Capital Controversy: Post-Keynesian Economics and the History of Economic Thought*. Routledge, London, **1**: pp. 198–218
Robinson J 1956 *The Accumulation of Capital*. Macmillan, London
Robinson J 1962 *Essays in the Theory of Economic Growth*. St Martin's Press, London
Robinson J 1978 Keynes and Ricardo. *Journal of Post Keynesian Economics* **1**: 12–18
Robinson J 1951–1980 *Collected Economic Papers,* 6 Vols. Blackwell, Oxford, UK
Rowthorn R E 1977 Conflict, inflation and money. *Cambridge Journal of Economics* **1**: 215–39
Sardoni C (ed.) 1992 *On Political Economists and Modern Political Economy. Selected Essays of G C Harcourt*. Routledge, London
Sraffa P 1960 *Production of Commodities by Means of Commodities. Prelude to a Critique of Economic Theory*. Cambridge University Press, Cambridge, UK
Wood A 1975 *A Theory of Profits*. Cambridge University Press, Cambridge, UK

G. C. Harcourt

Postmodern Urbanism

1. *Postmodernism in Urban Studies*

In the field of urban studies, three principal references are discernable in postmodern thought:

(a) a series of distinctive cultural and stylistic practices that are in and of themselves intrinsically interesting;

(b) the totality of such practices, viewed as a cultural ensemble characteristic of the contemporary epoch of capitalist society; and

(c) a philosophical and methodological discourse antagonistic to the precepts of Enlightenment thought, most particularly the hegemony of any single intellectual persuasion.

Implicit in all these approaches is the vision of a 'radical break,' i.e., a discontinuity between past and future trends (whether cultural, political-economic, or philosophical).

The pivotal point of departure in a postmodern urbanism is that just as the tenets of modernist thought have been undermined and a multiplicity of new ways of knowing put in their place, so, analogously, in postmodern cities the logics of previous urbanisms have evaporated; and in the absence of a single new imperative, multiple forms of (ir)rationality clamor to fill the vacuum. The localization, sometimes literally the concretization, of these multiple logics is creating the emerging time–space fabric of what may be called a postmodern society. The contradictions between a modernist and a postmodern urbanism are encoun-

tered most directly in the differences between the 'Chicago School' and the 'Los Angeles School' of urban structure.

1.1 The Chicago School

The most enduring of the Chicago School models was the zonal or concentric ring theory (Burgess 1925). Based on assumptions that included a uniform land surface, universal access to a single-centered city, free competition for space, and the notion that development would take place outward from a central core, Burgess concluded that the city would tend to form a series of concentric zones. At the core of Burgess's schema was the Central Business District (CBD), which was surrounded by a transitional zone, where older private houses were being converted to offices and light industry, or subdivided to form smaller dwelling units. The transitional zone was succeeded by a zone of workingmen's homes, which included some of the oldest residential buildings in the city, and stable social groups. Beyond this, newer and larger dwellings were to be found, occupied by the middle classes. Finally, the commuters' zone extended beyond the continuous built-up area of the city (where a considerable portion of the zone's population was employed).

Other urbanists noted the tendency for cities to grow in star-shaped rather than concentric form, along highways that radiate from a center with contrasting land uses in the interstices. This gave rise to a sector theory of urban structure, advanced in the late 1930s by Homer Hoyt (1939), who observed that once variations arose in land uses near the city center, they tended to persist as the city grew. Distinctive sectors thus expanded out from the CBD, often organized along major highways. The complexities of real-world urbanism were further taken up in the multiple nuclei theory of C. D. Harris and E. L. Ullman (1945). They proposed that cities have a cellular structure in which land uses develop around multiple growth-nuclei within the metropolis—a consequence of accessibility-induced variations in the land-rent surface and agglomeration (dis)economies.

Much of the urban research agenda of the twentieth century was predicated on the percepts of the concentric zone, sector, and multiple nuclei theories of urban structure. However, the specific and persistent popularity of the Chicago concentric ring model is probably related to its beguiling simplicity and the enormous volume of publications produced by adherents of the Chicago School.

1.2 The Los Angeles School

During the 1980s, a group of loosely associated scholars began to examine the notion that what was happening in the Los Angeles (LA) region was somehow symptomatic of a broader socio-geographic transformation taking place within the US. Their project was based on certain shared theoretical assumptions, as well as the view that LA was emblematic of some more general urban dynamic. One concern common to all adherents of the LA School was a focus on restructuring, which included deindustrialization, the birth of the information economy, the decline of nation-states, and the rise of the Pacific Rim. Such proliferating logics often involved multiple theoretical frameworks that overlap and coexist in their explanations of the burgeoning global/local order—a heterodoxy consistent with the project of postmodernism.

1.3 From Chicago to LA

Based on the differences between Chicago and LA, the distinction between a modernist and postmodern urbanism may be summarized as follows: whereas traditional concepts of urban form imagine a city based on a single central core that organized its hinterland (Chicago-style), in postmodern urbanism it is the hinterlands that organize what is left of the center (as in Los Angeles).

2. Postmodern Urbanism: Urban Form and Urban Process

Contemporary urban development in southern California provides a wealth of evidence that is symptomatic of a burgeoning postmodern urbanism.

2.1 Edge Cities

Joel Garreau (1991, p. 3) noted the central significance of Los Angeles in understanding contemporary metropolitan growth in the US. He asserts that: 'Every single American city that is growing, is growing in the fashion of Los Angeles,' and claims there are 26 edge cities within a five-county area in southern California. The classic location for contemporary edge cities is at the intersection of a suburban beltway and a hub-and-spoke lateral road. The central conditions that have propelled such development are the dominance of the automobile and the associated need for parking, the communications revolution, and the entry of women in large numbers into the labor market. One essential characteristic of the edge city is that politics is not yet established there. Into the political vacuum moves a 'shadow government'—a privatized proto-government that is essentially a plutocratic alternative to normal politics. Shadow governments can tax, legislate for, and police their communities; but they are rarely accountable, are responsive primarily to wealth (as opposed to numbers of voters), and are subject to few constitutional constraints (Garreau 1991, p. 187).

2.2 Privatopia

'Privatopia,' perhaps the quintessential edge-city residential form, is a private housing development based in common-interest developments (CIDs) and administered by homeowners' associations (McKenzie 1994, p. 11). Sustained by an expanding catalogue of covenants, conditions, and restrictions (or CC&Rs), privatopia has been fueled by ideology of 'hostile privatism' that has provoked a culture of non-participation. McKenzie notes how this 'secession of the successful' has altered concepts of citizenship, in which 'one's duties consist of satisfying one's obligations to private property' (1994, p. 196).

2.3 Cultures of Heteropolis

Provoked to comprehend the causes and implications of the 1992 civil disturbances in Los Angeles, Charles Jencks (1993, p. 32) zeroes in on the city's diversity as the key to LA's emergent urbanism. The consequent built environment is characterized by transience, energy, and unplanned vulgarity. Jencks views this improvisational quality as a hopeful sign, since it accepts the different voices that create a city, suppresses none of them, and makes from their interaction some kind of greater dialogue (1993, p. 75). This is especially important in a city where 'minoritization' is 'the typical postmodern phenomenon where most of the population forms the "other"' (Jencks 1993, p. 84).

2.4 City as Theme Park

California in general and Los Angeles in particular, have often been promoted as places where the American (suburban) dream is most easily realized. Many writers have used the 'theme park' metaphor to describe the emergence of its associated cityscapes. For instance, Michael Sorkin (1992) describes theme parks as places of simulation without end, characterized by a-spatiality, plus technological and physical surveillance and control. What is missing in this new cybernetic suburbia is not a particular building or place, but the spaces between; that is, the connections that make sense of forms (p. xii). What is missing, Sorkin maintains, is connectivity and community.

2.5 Fortified City

The downside of the southern Californian dream has been the subject of countless dystopian visions. In one powerful account, Mike Davis noted how southern Californians' obsession with security has transformed the region into a fortress. This shift is manifested in the physical form of the city, which is divided into fortified cells of affluence and places of terror where police battle the criminalized poor. These urban phenomena, according to Davis, have placed Los Angeles 'on the hard edge of postmodernity' (Davis 1992a, p. 155).

2.6 Interdictory Spaces

Elaborating upon the notion of fortress urbanism, Steven Flusty observed how various types of fortification have extended a canopy of suppression and surveillance across the entire city. His taxonomy of interdictory spaces (1994, p. 16–17) stresses how spaces are designed to exclude by a combination of their function and cognitive sensibilities. Some spaces are passively aggressive. Other configurations are more assertively confrontational. Flusty notes how combinations of interdictory spaces are being introduced 'into every facet of the urban environment, generating distinctly unfriendly mutant typologies' (1994, pp. 21–33). Some are indicative of the pervasive infiltration of fear into the home; others betray a fear of the public realm.

2.7 Historical Geographies of Restructuring

In his history of Los Angeles between 1965 and 1992, Ed Soja (1996) attempts to link the emergent patterns of urban form with underlying social processes. He identified six kinds of restructuring, which together define the region's contemporary urban process. In addition to 'Exopolis,' the city without (a massive simulation of what a city should be), Soja lists: 'Flexcities,' associated with the transition to post-Fordism, especially deindustrialization and the rise of the information economy; and 'Cosmopolis,' referring to the globalization of Los Angeles, both in terms of its emergent world-city status and its internal multicultural diversification. Three specific geographies are consequent upon these dynamics:

(a) 'Splintered Labyrinth,' which describes the extreme forms of social, economic, and political polarization characteristic of the postmodern city;

(b) 'Carceral' city, referring to the 'incendiary urban geography' brought about by the amalgam of violence and police surveillance; and

(c) 'Simcities,' the term Soja uses to describe the new ways of seeing the city—an epistemological restructuring that foregrounds the postmodern perspective.

2.8 Fordist and Post-Fordist Regimes of Accumulation and Regulation

Many observers agree that one of the most important underlying processes in the contemporary political

economy is represented by the shift from a Fordist to a post-Fordist industrial organization. Allen Scott's basic argument is that there have been two major phases of urbanization in the US. The first related to an era of Fordist mass production, during which the paradigmatic cities of industrial capitalism (Detroit, Chicago, Pittsburgh, etc.) coalesced around industries that were themselves based upon ideas of mass production. The second phase is associated with the decline of the Fordist era and the rise of a post-Fordist 'flexible production.' This is a form of industrial activity based on small-size, small-batch units of (typically subcontracted) production that are nevertheless integrated into clusters of economic activity (Scott 1993). Post-Fordist regimes of accumulation are associated with analogous regimes of regulation, or social control. Perhaps the most prominent manifestation of changes in the regime of regulation has been the retreat from the welfare state, which has led many people to the brink of poverty just at the time when the social welfare 'safety net' is being withdrawn.

2.9 *Globalization*

The global/local dialectic is another important (if somewhat imprecise) *leitmotif* of contemporary urban process. In his reference to the global context of LA's localisms, Mike Davis (1992b) claims that if LA is in any sense paradigmatic, it is because the city condenses the intended and unintended spatial consequences of post-Fordism. He insists that there is no simple master logic of restructuring, focusing instead on two key localized macro-processes: the over-accumulation in southern California of bank and real estate capital, principally from the East Asian trade surplus; and the reflux of low-wage manufacturing and labor-intensive service industries, following upon immigration from Mexico and Central America. Through such connections, what happens today in Asia and Central America will tomorrow have an effect in Los Angeles.

2.10 *Politics of Nature*

The natural environment of southern California has been under constant assault since the first colonial settlements. Human habitation on a metropolitan scale has only been possible through a widespread manipulation of nature, especially the control of water resources. On one hand, southern Californians tend to hold a grudging respect for nature, living as they do adjacent to one of the earth's major geological hazards, and in a desert environment that is prone to flood, landslide, and fire. On the other hand, its inhabitants have been energetically, ceaselessly, and sometimes carelessly unrolling the carpet of urbanization over the natural landscape for more than a century. This uninhibited occupation has engendered its own range of environmental problems, most notoriously air pollution; but it also brings forth habitat loss and dangerous encounters between humans and other animals. In this sense, a postmodern urbanism is axiomatically an environmental issue (Wolch and Emel 1998).

In sum, these diverse dynamics suggest an urban process:

(a) that is driven by a global restructuring and permeated and balkanized by a series of interdictory networks;

(b) whose populations are socially and culturally heterogeneous, but politically and economically polarized;

(c) whose residents are educated and persuaded to the consumption of dreamscapes even as the poorest are consigned to 'carceral' cities;

(d) whose built environment, reflective of these processes, consists of edge cities, 'privatopias,' and the like; and

(e) whose natural environment is being erased to the point of unlivability.

3. *The Postmodern Urban Condition*

In their model of postmodern urbanism, Dear and Flusty (1998) synthesize the diverse spatial manifestations of the postmodern urban process into a vision of 'keno capitalism.' According to their characterization, urbanization is occurring on a quasi-random field of opportunities. Capital touches down as if by chance on a parcel of land, ignoring the opportunities on intervening lots, thus sparking the development process. The relationship between development of one parcel and nondevelopment of another is a disjointed, seemingly unrelated affair. While not truly a random process, it is evident that the traditional, center-driven agglomeration economies that guided urban development in the past no longer apply. Conventional city form, Chicago-style, is sacrificed in favor of a noncontiguous collage of parceled, consumption-oriented landscapes devoid of conventional centers, yet wired into electronic propinquity and nominally unified by the mythologies of the (dis)information superhighway. The consequent urban aggregate is characterized by acute fragmentation and specialization—a partitioned gaming board subject to perverse laws and peculiarly discrete, disjointed urban outcomes.

The validity and full potential of the concept of postmodern urbanism can be judged only by a thorough comparative urban analysis that goes beyond Los Angeles (Dear 2000). Four contemporary emphases seem especially promising:

(a) world cities, or the emergence of a very few 'megacities' involved in key command and comfort functions in the globalizing political economy;

(b) dual cities, or the burgeoning polarizations between rich and poor, powerful and powerless, computer literate and illiterate, etc.;

(c) hybrid cities, or the fragmentation of existing social alliances and the rise of unprecedented hybrid categories and places; and

(d) 'cybercities,' or the radical undermining of conventional geographies by the telecommunications revolution.

See also: Critical Realism in Geography; Humanistic Geography; Postmodernism in Geography; Postmodernism in Sociology; Urban Anthropology; Urban Studies: Overview

Bibliography

Burgess E W 1925 The growth of the city. In: Park R E, Burgess E W, McKenzie R D (eds.) *The City: Suggestions of Investigation of Human Behavior in the Urban Environment*. University of Chicago Press, Chicago, IL, pp. 47–62

Davis M L 1992a Fortress Los Angeles: The militarization of urban space. In: Sorkin M (ed.) *Variations on a Theme Park*. Noonday Press, New York, pp. 154–80

Davis M L 1992b Chinatown revisited? The internationalization of downtown Los Angeles. In: Reid D (ed.) *Sex, Death and God in L.A*. Pantheon Books, New York, pp. 54–71

Dear M 2000 *The Postmodern Urban Condition*. Blackwell, Oxford, UK

Dear M, Flusty S 1998. Postmodern urbanism. *Annals, Association of American Geographers* **88**: 50–72

Flusty S 1994 *Building Paranoia: The Proliferation of Interdictory Space and the Erosion of Spatial Justice*. Los Angeles Forum for Architecture and Urban Design, West Hollywood, CA

Garreau J 1991 *Edge City: Life on the New Frontier*. Doubleday, New York

Harris C D, Ullman E L 1945 The nature of cities. *Annals of the American Academy of Political and Social Science* **242**: 7–17

Hoyt H 1939 *The Structure and Growth of Residential Neighborhoods in American Cities*. US Federal Housing Administration, Washington, DC

Jencks C 1993 *Heteropolis: Los Angeles, the Riots and the Strange Beauty of Hetero-architecture*. St. Martin's Press, New York

McKenzie E 1994 *Privatopia: Homeowner Associations and the Rise of Residential Private Government*. Yale University Press, New Haven, CT

Scott A J 1993 *Technopolis: High-technology Industry and Regional Development in Southern California*. University of California Press, Berkeley, CA

Soja E 1996 Los Angeles 1965–1992: The six geographies of urban restructuring. In: Scott A J, Soja E (eds.) *The City: Los Angeles & Urban Theory at the End of the Twentieth Century*. University of California Press, Berkeley, CA, pp. 426–62

Sorkin M (ed.) 1992 *Variations on a Theme Park: The New American City and the End of Public Space*. Hill and Wang, New York

Wolch J, Emel J (eds.) 1998 *Animal Geographies: Place, Politics and Identity in the Nature–Culture Borderlands*. Verso, London

M. J. Dear

Postmodernism in Geography

The impact of postmodernism on the field of geography has been particularly intense and far-reaching, perhaps more so than in any other social science. Although still controversial, with few geographers specifically identifying themselves as postmodernists, postmodern and related *poststructuralist* critiques of traditional modes of interpreting and explaining geographical phenomena have influenced scholarly research and teaching in nearly all branches of the discipline. Especially influential has been postmodernism's role in shaping the development of what has come to be called *Critical Human Geography* and in the rethinking of such core concepts as space, place, region, and environment (see also *Postmodern Urbanism*). At the same time, this postmodernization of geography has contributed significantly to expanding geography's impact on other disciplines, both in the social sciences and the humanities. Geographical writings on postmodernism and postmodernity have been an integral part of the recent resurgence of interest in concepts of space and place across a wider spectrum of academic disciplines than ever before in the twentieth century.

1. The Postmodern Turn in Geography

Postmodernism entered geography as an offshoot and extension of critical responses to the positivist 'spatial science' paradigm that began to dominate the discipline in the 1960s. These critical responses took two major forms. A radical or *Marxist Geography* consolidated in the 1970s as a critique of prevailing modes of explanation in spatial science and as a means of making geography more socially and political relevant. Rather than focusing primarily on statistically measurable surface appearances and searching for empirical regularities based on the frictions of distance and spatial covariation, how one geographical pattern or distribution correlates with another, Marxist geographers analyzed and explained what they called *specific geographies* as socially constructed outcomes of underlying processes such as class formation and capitalist accumulation. This carried with it a strong emphasis on structural explanation of geographical phenomena (see *Structuralism*) and gave rise to a specifically geographical (urban, regional, international, ecological) political economy perspective that would influence the discipline significantly in subsequent years.

The second stream of critical response, developing initially as *Humanistic Geography*, also concentrated on alternative modes of explanation, but emphasized the role of human agency and subjectivity much more than the effects of powerful structuring forces emanating either from society or the environment (Ley and Samuels 1978). Drawing on phenomenology, exis-

tentialism, and hermeneutics as well as traditional forms of cultural geography and landscape analysis, these geographical humanists were less concerned with space and political economy than with *place*, *nature–society relations*, and the interpretation of *cultural landscapes*. Continuing long-established links with anthropology, this reinvigorated form of *Cultural Geography* would also influence the development of postpositivist critical human geography over the next several decades as an important countercurrent to the Marxist critique.

By the early 1980s, the spatial science paradigm had passed its peak of influence and many new approaches were entering geography, each competing to define the leading edge of postpositivist research. At the same time, significant critiques were developing both within and between Marxist and humanistic geography (and within spatial science as well), adding further to the proliferation of new perspectives. Perhaps the most challenging critiques came from feminist geographers, who saw an exploitative and demeaning *masculinism* inherent in every branch of what would be labeled modern geography. This patriarchal order was not only expressed in the workplace (e.g., in university privileges, hiring, salaries) but also in the very practice of geography, whether theoretical or empirical, positivist, Marxist, or Humanist. The feminist critique echoed on a broader platform many other concerns over what was perceived as an excessive narrowing of perspectives affecting and constraining the development of (modern) critical human geography. It was in this atmosphere of increasing epistemological fragmentation and escalating internal critique of modern geographical practices that postmodernism first entered the disciplinary discourse.

The deepest divisions within postpositivist geography revolved around traditional philosophical dichotomies or binary oppositions: Subject–Object, Idealism–Materialism, Agency–Structure. Their particular expression, however, took the form of an opposition between approaches that emphasized spatial political economy vs. those that stressed place-based culture. In the late 1970s, there had been some efforts to bring these opposing positions together, most notably in the writings of Derek Gregory. Influenced by the social theorist Anthony Giddens, also then at Cambridge University, Gregory (1978) argued for a critical synthesis of 'structural' and 'reflexive' approaches to geographical explanation that would go beyond their traditional antagonism to create a new, socially and politically 'committed' critical human geography. Gregory was one of the first geographers to recognize the potential relevance of postmodern and closely associated poststructuralist debates to achieving this goal, and for the next two decades would be a leading voice in adding a postmodern, poststructuralist, and, later, postcolonial perspective to geographical research, theory, and teaching (Gregory and Walford 1989, Gregory 1994).

In addition to the influence of Gregory's work and the growing impact on geography of Giddens's structuration theory, there were other developments in the early 1980s that attempted to overcome the divisions and fragmentation of postpositivist geography by at least in part introducing some aspects of the postmodern and poststructuralist critiques (Olsson 1980). In retrospect, however, the key turning point in the postmodernization of geography came from outside the discipline, with the publication (in 1984) of Fredric Jameson's *Postmodernism, or the Cultural Logic of Late Capitalism*. Jameson, a widely recognized Marxist literary critic and cultural studies scholar, had recently become interested in continental European theories of space, place, and architecture, and especially in the work of the French philosopher and urbanist Henri Lefebvre. Then teaching in the History of Consciousness program at the University of California-Santa Cruz, Jameson had organized a lecture tour by Lefebvre in the previous year and hosted his stay as visiting professor at Santa Cruz in 1984.

While he saw architecture as the 'privileged' language of postmodernism, Jameson was aware of the new developments in critical human geography and gathered a group of geographers, urbanists, and architects, mainly from California, at a conference to celebrate Lefebvre's creative spatial theories. When the article appeared, presenting a postmodern and explicitly spatial critique that effectively bridged Marxist political economy and critical cultural studies, innovatively interpreting emerging forms of contemporary urban geography, and calling for a new 'political esthetic' based on what Jameson termed a 'cognitive mapping' of 'postmodern hyperspace,' its effect on geography was particularly powerful.

The impact was especially intense among geographers working in California and, in particular, in Los Angeles, from whose geography Jameson drew some of his most imaginative examples and where there already existed one of the largest clusters of critical human geographers in the US. In the ensuing years, the geography and the geographers of Los Angeles would play a leading role in defining postmodernism and extending its impact on the discipline.

2. *Postmodernism Seen from Los Angeles*

By the mid-1980s, after decades of relative neglect by urban studies scholars, Los Angeles had become the research focus for a group of geographers and urban planners interested in making theoretical and practical sense of the new urbanization processes that were transforming the contemporary city, not just locally but elsewhere as well. The earliest work on this *urban restructuring* emphasized the radical political economy perspective that was so central in Marxist geography, but there was an openness to alternative perspectives

and practices that arose in large part from the close connections that had developed between geography and urban planning at local universities such as UCLA and USC. These connections linked theorizing the political economy of urban restructuring to the practical demands of public policy, urban politics, and local community activism.

Developing from these early studies was a conceptual framework that saw in the restructuring of Los Angeles the emergence of a new era of urban and regional development affecting cities all over the world in the aftermath of the tumultuous urban crises of the 1960s. This new kind of urbanism and urban geography was interpreted as the product of deep changes in the local, regional, and national economy associated with the rise of *post-Fordism*, increasingly *flexible* methods of industrial production, and the powerful local effects of *globalization* and new information technologies (Scott 1988; see also *Economic Geography*; *Industrial Geography*; *Technology Districts*; *Economic Restructuring: Geographic Aspects*; *Information Society, Geography of*). Built in to this primarily economic conceptualization of urban restructuring was also a critical postmodern interpretation that saw in the dramatic changes taking place in the century's last three decades the rise of a distinctively postmodern as well as post-Fordist urbanism.

In 1986, a special issue devoted to recent research on Los Angeles appeared in the journal *Environment and Planning D: Society and Space*, with an editorial introduction by Scott and Soja describing Los Angeles as the most representative, if not paradigmatic, city of the late twentieth century. The journal, with its founding editor Michael Dear and close editorial ties to British geographers such as Derek Gregory and Nigel Thrift, had quickly become the major journal for current debates relating geography to social theory and cultural studies since its origins four years earlier. In this special issue, two articles appeared, one by Dear and the other by Soja, that were the first explicit statements arguing for a postmodern perspective in geography and planning. Drawing on the changing geography of Los Angeles and the innovative urban scholarship of Jameson and Lefebvre, the authors described postmodernism in three related ways: as a *method* of textual and visual representation, as a contemporary *condition* of everyday life, and as a comprehensive epistemological *critique*. In each of these approaches to postmodernism in geography, emphasis was placed on the distinctiveness of the present era, on what is new and different in the contemporary world, an emphasis that in the broadest sense defines a postmodern perspective.

Defined as a method of textual and visual representation, postmodernism aimed at the intentional *deconstruction* of what were seen as outmoded modernist practices of writing geography, with the intent of opening up creative new alternatives. Rather than accepting the impress of powerful continuities with the past, for example, there would be playful and disruptively sequenced historical references, accentuating the synchronic and the simultaneous rather than a sequential diachronic narrative. The orderliness of visual form and textual narrative was similarly deconstructed through collage, pastiche, interjected verse, and other unruly modes of literary and visual representation. Explanation and interpretation thus became increasingly involved in *discourse analysis*, with geographies seen as complexly constructed 'texts' to be discursively 'read' as simultaneously real and imagined, a spatial version of the language games and deconstructive practices of such leading postmodernist philosophers as Lyotard, Derrida, and Foucault.

Postmodernism defined as an empirical condition of contemporary life was much easier to accept in geography than what appeared to many to be merely a fashionable and unduly confusing style of writing. That at least part of the world had entered what could be described as a postmodern era and that there were *postmodern geographies* and a *postmodern urbanism* worthy of being studied as such posed relatively little challenge to most geographers. The growing role of economic and industrial geographers in making practical and theoretical sense of the restructuring processes affecting the contemporary world reinforced this widening impact, and helped to create a close association between the geographical condition of postmodernity and the impact of post-Fordist economies of flexible production, globalization processes, and the information society. Just how extensive this empirical postmodernization was, whether it was positive or negative or both, and, especially, how it should affect the practice of geography beyond the description of places like Los Angeles remained much more controversial.

The greatest resistance arose, not unexpectedly, over postmodernism as a deep epistemological critique, especially after subsequent writings by Dear (1988) and Soja (1989) pushed this critique into the very core of geography as a discipline. They argued that all forms of explanation in (modern) geography were fundamentally flawed, for, among other reasons, they rested on epistemological assumptions that excluded too many important issues and voices, especially related to gender, race, class, and what postmodern critics call the marginalized 'Other'; were based on all-encompassing or totalizing metanarratives that too confidently proclaimed the possibility of obtaining complete and accurate knowledge about human geographies; and were excessively bound up in oppositional binaries that left little room for alternative forms of knowledge production. Without abandoning modernism and modern geography entirely, Dear and Soja contended that the restructuring of empirical geographies that had been taking place in the last third of the twentieth century had so changed the contemporary world that established methods of analysis

and explanation in geography, postpositivist or otherwise, could no longer be so confidently relied upon. New and different approaches, an assertively postmodern*ist* geography rather than simply a descriptive geography of the postmodern condition, needed to be opened up and explored.

3. *Postmodernism in Geography: 1989–Present*

Two books published in 1989 were particularly influential in shaping the directions that would be taken by postmodernism in geography in the 1990s: Soja's *Postmodern Geographies: The Reassertion of Space in Critical Social Theory* and Harvey's *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Both authors were rooted in Marxist geography and political economy, and shared a deep commitment to the creation of a historical materialism that was simultaneously profoundly geographical and spatial. Where they differed most was in the degree to which this construction of a historical and geographical materialism required a critical deconstruction and reconstitution of the ontology and epistemology of Marxist analysis itself.

Harvey adaptively recognized the challenge to radical geographical analysis posed by the postmodern condition but remained confident in the persistent power of historical (and geographical) materialism as a critical methodology for understanding and changing the contemporary world. There were many aspects of the condition of postmodernity that were new and different from the past, but they could best be understood through a more geographical Marxism that viewed postmodernity as just the most recent phase of capitalist development rather than as a disruptive or epochal break with the past.

Soja was more avowedly a postmodernist, insisting that a much deeper restructuring of Marxism was necessary to engage effectively with contemporary geographical realities. He called for a radical rethinking of two fundamental or ontological premises embedded in Marxism, one relating the social and the spatial dimensions of capitalist societies and the other linking time and space, history and geography. Following Lefebvre, Soja argued that there was a more balanced and mutually causal relation between society and space, a sociospatial dialectic that had been submerged in Marxism's overemphasis on social relations of production and aspatial class analysis. Similarly, he saw a persistent privileging of history over geography, the temporal over the spatial, in a particular form of *historicism* that severely constrained the development of a balanced and mutually interactive historical and geographical materialism. He called instead for a more balanced and three-sided ontology and epistemology that dynamically related the spatial, social, and historical dimensions of human existence (spatiality, sociality, and historicality), with no one of the three inherently privileged over the others (Soja 1996).

The contrasting positions taken by Harvey and Soja on the implications of postmodernism were virtually ignored in the immediate reaction in geography to the two books. Instead, their publication by two prominent male geographers authoritatively writing about postmodernity provoked a concerted response from feminists to what was perceived as another round of masculinist appropriation of the discipline's leading edge of research and theory. Both were accused of 'flexible sexism' (Massey 1991), of living in 'Boy's Town' hopelessly 'lost in space,' of imposing their authoritative power too forcefully, no matter how different their positions on postmodernism. Many of the harshest feminist critics shared an interest in Marxist geography and in the postmodern critiques of modernism, but were suspicious of the particular directions suggested by Harvey and Soja for the future of critical human geography.

The end result of this confrontation and critique was both a deflection and a *refraction* of the impact of postmodernism. A new generation of geographers, especially in Great Britain, set the confrontation aside to work in many different ways at applying postmodernist and poststructuralist perspectives to geographical analysis and interpretation, avoiding both the narrowed historical materialism of Harvey's approach to the postmodern condition and the overly enthusiastic celebrations of postmodernism associated with the work of Soja and Dear.

To an unusual degree, some version of the postmodern epistemological critique (often subsumed under the safer sounding label of poststructuralism) was absorbed into an increasingly eclectic critical human geography in the 1990s, but more as a justification for theoretical and methodological pluralism than as an encompassing or integrative philosophy. Continuing debates about explanation in geography were significantly expanded in scope and increasingly focused on such issues as the construction of *difference* and *identity*; the relations between *space, knowledge, and power*; the pervasive problems of *representation*; and the increasing confusion of the real and the imagined in *cyberspace* and *hyper-reality*. Growing attention was given to different scales of geographical analysis, ranging from studies of the human body, the most intimate geography, to the new landscapes being produced by the globalization of capital, labor, culture, and information flows. Nearly every subfield became opened up to a multiplicity of interpretive perspectives, including cartography (see Harley in Barnes and Gregory 1997) and geographical information systems (Pickles 1994), with little concern for establishing an overarching theorization or disciplinary paradigm.

Feminist geographers such as Doreen Massey (1994), Gillian Rose (1993), and Linda McDowell

(1993) were particularly influential in shaping the continuing and open-ended postmodernization of geography, cautiously promoting a gender-based postmodern critical epistemology as part of the widening pluralism of interpretive methodologies. Similarly, Kathy Gibson and Julie Graham (writing as Gibson-Graham 1996) called for new directions in Marxist geography based on their reinterpretations of contemporary feminist, postmodernist, and postructuralist theory. The new cultural geography, building on the early work of Cosgrove and Daniels (1988), Duncan (1990), and Barnes and Duncan (1992), moved increasingly closer to critical cultural studies and deconstructive discourse analysis, with perceptive readings of cultural landscapes as texts and 'signifying practices.' Cultural and political economy perspectives were partially reconnected through a growing interest in the *postcolonial critiques* of scholars such as Edward Said, Gayatri Spivak, Arjun Appadurai, and Homi Bhabha (Gregory 1994, Soja 1996), while geographical studies of the body, sexuality, the unconscious, and the formation of human subjectivity, drawing on the psychoanalytical theories of Lacan as well as the ideas of Foucault and Lefebvre, created other kinds of bridges between subjective place-based culture and spatial political economy (Pile and Thrift 1995, Pile 1996).

There has also developed a vigorous antipostmodernism in the 1990s, even among those whose work most geographers would associate with a postmodern perspective. Many Marxist geographers, for example, continuing to follow David Harvey's lead (see Harvey 1996), have narrowed postmodernism almost entirely to being a politically diverting and divisive aspect of the contemporary *condition* of global capitalism. Efforts to continue to press a postmodern epistemological critique are therefore seen as politically retrogressive and antagonistic to the radical project of historical and geographical materialism. At the other end of the political spectrum, more conservative geographers frequently reduce postmodernism to superficial and often foolish playfulness, blunting any challenge to their established convictions. But very few geographers today are without an opinion on postmodernism or unaware of its impact on the discipline.

As the new century begins, it can be argued that postmodernism and the interpretive approaches associated with it (poststructuralism, postcolonial critiques, post-Marxism, postmodern feminism, methods of deconstruction and discourse analysis) have been absorbed into geography as an integral part of the contemporary discipline (see such recent overviews of modern and postmodern geography as Barnes and Gregory 1997, Benko and Strohmayer 1997, Peet 1998, Massey et al. 1999; see also Jones et al. 1994 and Watson and Gibson 1995). Postmodernism has been blunted and deflected in its boldest claims, continues to be actively resisted by many, but remains part of the consciousness and imagination of most human geographers.

The achieved and continuing impact of postmodernism can be summarized in several ways. First, it has helped to bring geography into closer contact with social theory and philosophy, and nearly every branch of the social sciences and humanities, from film studies and literary criticism to anthropology and economics. This has contributed in turn to what might be called a 'de-disciplining' of geography: an opening up of its traditional borders with other fields; a wider diffusion of its core concepts and ways of studying space, place, region, and environment; and the emergence of an extraordinary pluralism within geography with respect to theory, epistemology, and methods of empirical analysis. Intense frictions and disagreements remain, to be sure, but there may be no other time in its existence as a distinct discipline when geography has been so paradigm-free and yet so critically engaged with the major issues and events of our times. While these developments are not entirely due to the impact of postmodernism, it has certainly played a stimulating role.

Bibliography

Barnes T J, Duncan J (eds.) 1992 *Writing Worlds: Discourse, Text, and Metaphor in the Representation of Landscape*. Routledge, New York

Barnes T J, Gregory D (eds.) 1997 *Reading Human Geography: The Poetics and Politics of Inquiry*. Arnold, London

Benko G, Strohmayer U (eds.) 1997 *Space and Social Theory: Interpreting Modernity and Postmodernity*. Blackwell, Oxford, UK

Cosgrove D, Daniels S 1988 *The Iconography of Landscape*. Cambridge University Press, Cambridge, UK

Dear M 1988 The postmodern challenge: Reconstructing human geography. *Transactions I B. Geography* **13**: 262–74

Dear M 1986 Postmodernism and planning. *Environment and Planning D: Society and Space* **4**: 367–84

Duncan J S 1990 *The City as Text: The Politics of Landscape Interpretation in the Kandyan Kingdom*. Cambridge University Press, Cambridge, UK

Gibson-Graham J K 1996 *The End of Capitalism (As We Knew It): A Feminist Critique of Political Economy*. Blackwell, Oxford, UK

Gregory D 1978 *Ideology, Science, and Human Geography*. Hutchinson, London

Gregory D 1994 *Geographical Imaginations*. Blackwell, Oxford, UK

Gregory D, Walford R (eds.) 1989 *Horizons in Human Geography*. Macmillan, London

Harvey D 1989 *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Blackwell, Oxford, UK

Harvey D 1996 *Justice, Nature and the Geography of Difference*. Blackwell, Oxford, UK

Ley D, Samuels M (eds.) 1978 *Humanistic Geography: Prospects and Problems*. Maaroufa Press, Chicago

Jameson F 1984 Postsmodernism, or the cultural logic of late capitalism. *New Left Review* **146**: 53–92

Jones J P, Natter W, Schatzki T R (eds.) 1994 *Postmodern Contentions: Epoch, Politics, Space*. Guilford Press, New York
Massey D 1991 Flexible sexism. *Environment and Planning D: Society and Space* **9**: 31–57
Massey D 1994 *Space, Place and Gender*. University of Minnesota Press, Minneapolis, MN
Massey D, Allen J, Sarre P (eds.) 1999 *Human Geography Today*. Polity Press, Cambridge, UK
McDowell L 1993 Space, place and gender relations: Part II: Identity, difference, feminist geometries and geographies. *Progress in Human Geography* **17**: 305–18
Olsson G 1980 *Birds in Egg*, 2nd edn. Pion, London
Peet R 1998 *Modern Geographical Thought*. Blackwell, Oxford, UK
Pickles J (ed.) 1994 *Representations in an Electronic Age: Geography, G.I.S., and Democracy*. Guilford Press, New York
Pile S 1996 *The Body and the City: Psychoanalysis, Space and Subjectivity*. Routledge, London
Pile S, Thrift N 1995 *Mapping the Subject: Geographies of Cultural Transformation*. Routledge, London
Rose G 1993 *Feminism in Geography: The Limits of Geographical Knowledge*. University of Minnesota Press, Minneapolis, MN
Scott A J 1988 *Metropolis: From Division of Labor to Urban Form*. University of California Press, Berkeley, CA
Soja E W 1986 Taking Los Angeles apart: Some fragments of a critical human geography. *Environment and Planning D: Society and Space* **4**: 255–72
Soja E W 1989 *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. Verso, London
Soja E W 1996 *Thirdspace: Journeys to Los Angeles and other Real-and-imagined Places*. Blackwell, Oxford, UK
Watson S, Gibson K (eds.) 1995 *Postmodern Cities and Spaces*. Blackwell, Oxford, UK

E. W. Soja

Postmodernism in Sociology

Prominent fields of postmodernist research have been work and organizations, political action, science and technology, commodification, consumption, gender, media, and popular culture. Classical sociological topics like symbolic consumption have been further explored and enriched. Sociological postmodernism has also pioneered domains like computers and the Internet. It is to be expected that in the future postmodernist research will evolve in the direction of concrete field studies, with an emphasis on the cultural forms induced by computers and the Internet.

In sociology, postmodernism designates (a) a cluster of theoretical and meta-theoretical approaches; (b) an analysis of postmodernity, understood as encompassing the social and cultural features of late capitalism; (c) an extension of sociological inquiry to new domains; and (d) new forms of sociological expression. Sociological postmodernism is thus a form of sociological analysis, a kind of sociological sensibility, and a sociologists' social and intellectual condition at the same time. Its varieties share a series of theoretical and methodological premises, but differ in their conclusions and research programs. Core common elements are: (a) avoiding recourse to a set of universally valid assumptions as theoretical and methodological foundations, together with (b) the key role ascribed to notions like subject, identity, text, and symbol in the analysis of society.

1. *Origins of Postmodernism*

The term 'postmodernism' has been used in Latin-American literary criticism since the 1930s, and in Anglo-American debates since the 1940s, in order to designate new forms of expression in their relationship with the aesthetic of modernism. In history, the term was used by the British historian Arnold Toynbee in *A Study of History* in 1947 and designated the latest phase of Western civilization. In sociology, it was introduced by Amitai Etzioni's book *The Active Society* in 1968. Starting with the early 1970s, the term appeared more and more frequently in sociological texts in France, North America, and the UK. At that time, it was already common stock in literary theory and criticism. An essay written by the French philosopher Jean-François Lyotard, *The Postmodern Condition: A Report on Knowledge* (1979) played a seminal role in making this term widely known and used in the social sciences. Therefore, postmodernism was partially a conceptual import; its diffusion in the 1970s was part of a wider innovative movement in the discipline, marked, among others, by the rise of social constructivism and feminism. At the same time, it continued and reformulated a series of topics already present in sociology.

The intellectual roots of sociological postmodernism can be identified in the works of some key nineteenth and early twentieth century philosophers, sociologists, and linguists. Karl Marx's critique of capitalism, Georg Simmel's analysis of modernization processes, Friedrich Nietzsche's critique of epistemology, and Ferdinand de Saussure's theory of language play here a prominent role. The French structuralist movement (Claude Lévi-Strauss's anthropology, Roland Barthes's semiotic theory, Jacques Lacan's psychoanalysis), and the poststructuralist one (Michel Foucault, Jacques Derrida), Ludwig Wittgenstein's later works, and Clifford Geertz's anthropology have also played a considerable role. Among other influences there are the Frankfurt School (Walter Benjamin, Theodor Adorno, Max Horkheimer, and Herbert Marcuse), symbolic interactionism, ethnomethodology, and phenomenological sociology.

2. *The Tenets of Postmodernism*

There are two main tenets, mirroring the distinction between postmodernism and post-modernity: while the latter designates the social and cultural features of

late capitalism, distinguishing them from earlier phases (modernity), the former designates a theoretical and methodological program, as distinct from earlier (modernist) ones.

2.1 The Analysis of Postmodernity

The analysis of postmodernity developed at the beginning of the 1970s, continuing a preoccupation already found in the works of Georg Simmel, Thorstein Veblen, and the Frankfurt School. Postmodernity is a late phase in the evolution of the capitalist system, seen either as a continuation and intensification of modernity, or as a substantially new stage. It is characterized by the transition from manufacture to a decentralized services economy; science and complex technological systems play a key role here. Decentralization, together with the pervasiveness of expert and technological systems in all domains of everyday life lead to a fragmentation and diversification of social identities. The electronic media (television, computers, video) are prominent in the constitution and presentation of social reality; visual and linguistic symbols become pervasive, while popular cultural forms diversify and expand. Society is therefore fragmented and heterogeneous; social order is always local and contextual, and works as an assemblage of symbolic codes. Production and consumption, social stratification, as well as personal and collective identities become dependent on such codes. They are understood as a support, a medium for, and an outcome of social action. This notion becomes a general explanatory paradigm: every domain of societal organization can be seen as a symbolic system and analyzed accordingly.

Consequently, sociology (an outcome of modernity itself) should take fragmentation, decentralization, relativization, and the mediated character of social reality into account; it should adapt its own tenets, claims, as well as its own modes of analysis and representation to these changes. Starting from these premises, Zygmunt Bauman (1992) has argued that notions like sociality, habitat, self-constitution, and body are better suited for the analysis of contemporary social processes than the traditional concepts of society and order.

2.2 The Critique of Modernist Sociology

During the 1980s, a second theme made its way from philosophy and literary theory into sociological debates. It was a critique of the status of sociological theory as a modernist product, and it drafted a new methodological program. This debate raised a series of questions about the validity claims of scientific theories, the nature of scientific truth, and the status of the knowing subject, arguing that: (a) all knowledge is contextual and local; (b) the validity claims of any scientific theory are not to be found in some abstract, universal criteria, but are rather the results of either negotiated consensus or power struggles; and (c) as a consequence, the knowing subject does not dispose of universal criteria to ascertain the validity and truth of his or her knowledge. The subject should not be seen as autonomous with respect to an objective, given world, but rather as the product of social circumstances. Since knowledge has a contextual and local character, it follows that scientific theories are reinterpreted according to local conditions: they are open-ended and cannot be controlled by their authors. Scientific theories are not determined by external validity and truth criteria; they work as systems of symbols enabling certain courses of social action and promoting a certain worldview. On these grounds, there should be a structural similarity between texts and social order: both are locally determined, network-like, open-ended systems of symbols requiring certain kinds of action from the actors embedded in them.

Notions like text and performance should occupy a prominent place in the analysis of social phenomena. Sociology should focus on investigating and describing how symbol systems work in particular social contexts, instead of producing all-encompassing theories. This claim, together with the one that the knowing subject is the product of social circumstances, have sometimes been described in a more radical form as 'the end of sociological theory' and the 'death of the author,' respectively. A corollary of these arguments is that sociological theories should acknowledge the fragmentation and diversity inherent in the local character of knowledge; they should explore their consequences for the world we live in, and should experiment with new modes of representation—like poetry, or dialogical forms. To some extent, these epistemological claims overlap with those of the first strand of postmodernist sociology. However, we encounter an important difference: while the first strand derives the fragmentation, relativization, and diversification of knowledge from social changes, the second one takes them as stemming out of a number of epistemological, universally valid first principles.

At the end of the 1980s and in the early 1990s, such meta-theoretical topics were intensely debated in sociology and, for a short time, even eclipsed the first strand of postmodernism. However, the acceleration of globalization processes brought social analysis to the forefront again.

3. The Methodology of Postmodernist Sociology

One important methodological aspect is the anti-positivist and antiverificationist stance: there is no totalizing explanatory model and no universally valid verification procedure. Postmodernist sociology stresses the role of qualitative methods in the analysis of contemporary social phenomena, as a consequence

of the prominence of visual and linguistic symbols. While some authors (Richardson 1997) emphasize interpretation and subjectivity, others (Lemert 1997) argue that methodology must preserve objectivity and logical consistency. There has been an increased sociological interest for discourse analysis, understood as a cluster of techniques for grasping the emergence, functioning, and context-boundedness of symbolic codes. Ethnographic methods, earlier reserved for the study of 'exotic' societies or of marginal social phenomena, also play an important role. This has increased the interest for texts and writing processes as sources of conceptual enrichment in sociology. Poetry, drama, or dialog are promoted as legitimate complements to the established, academic expository modes. Hybrid forms of sociological expression have appeared too—for example, in works combining academically written texts, poetry, and artwork (Haraway 1997).

4. *Fields of Postmodernist Research*

Postmodernist analyses have been prominent in domains like work and organizations, political action, science and technology, commodification and consumption, gender, media, and popular culture. George Ritzer (1993), for example, has examined the new forms of work which emerge in the knowledge-based, global economy: they are characterized by fragmentation, local specialization, flexibility, and mobility, on the one hand, and by global dispersion and typification, on the other. Under these circumstances, the temporal and spatial conditions of work change substantially: while the lifespan of products shortens, they need to be replaced faster, and worktime expands. Tensions arise between requirements for flexibility and mobility, on the one hand, and typification, on the other. New directions in commodification and consumption complete this picture. Commodification is understood as the unrelentless expansion of market exchange, which reshapes social relationships according to its own logic; the pendant of this process is the increased social significance of the act of consumption. Jean Baudrillard (1972), for example, has argued that symbolic consumption becomes one of the main social mechanisms through which individual and collective identities, as well as social relationships are defined and reproduced.

In political sociology, the analysis has focused on political discourses and symbolisms, on the spectacular form taken by political events, and on the role of television as a medium for political action. Another direction has been to examine the strains put on democracy by the media-induced forms of political action, and by the globalized economy. Authors like Chantal Mouffe (1992) have asked whether the new economic and political conditions do not require devising new forms of democratic participation and of political action.

With knowledge lying at the core of the postmodern society, science has been a domain of sustained investigation. The focus has been on the changes brought about by the new information and biomedical technologies in social life, as well as on the role and effects of complex technological systems. Bruno Latour (1991) and Donna Haraway (1997), among others, have argued that such systems are so pervasive and deep-reaching, and that social life is dependent on them to such an extent that new forms of social symbiosis have emerged; these are described with the help of notions like the cyborg society, hybrids, or networks of human and nonhuman actors. These authors have called for a revision of the conceptual apparatus of sociology on the grounds of the relationships between human actors and technical systems.

A further domain of investigation, situated at the crossroads between feminist and postmodernist sociology, has been that of gender and sexuality. Postmodern diversification brings forth a fragmentation and multiplicity of gender and sexual identities. The distinctions between genders become blurred; some forms of sexual identity formerly considered marginal gain a new significance, and new sexual identities emerge (e.g., transsexuality). Owing to the local and contextual character of knowledge, the gender specificity of knowledge and experience has to be taken into account. Several authors (e.g., Smith 1990) have explored the ways and the implications of gender-specific knowledge in the workplace, the field of politics, popular culture, and the public sphere.

One of the key domains in which the impact of postmodernist sociology has been felt is that of the electronic media and communication. Since the social changes brought about by computers and television have been considerable, questions arise about new, media-induced forms of social life and organization. According to Jean Baudrillard, among others, large portions of the social world are now media generated, and whole series of political and social events have a media-supported existence: they exist as simulacra, as symbolic codes which can be reproduced endlessly, without making recourse to an original. Other domains of investigation have been computers in the workplace, or the media-induced forms of social life, like those arising in and around the Internet (Turkle 1995). Finally, postmodernist sociology has also considerably influenced the subfields of cultural and urban studies, reinvigorating a research tradition which goes back at least to the Chicago School of sociology.

5. *The State of Postmodernist Research*

Postmodernist sociology has been around for approximately three decades; while controversial and contested for some of its more radical theoretical stances, many postmodernist ideas have made inroads into mainstream sociology. At the same time, it has

continued and enriched a series of themes already present in classical sociology—like the sociology of consumption, or the study of popular cultures. It has also pioneered some new domains of research—the study of Internet cultures, of the media, or (together with ethnomethodology) that of the impact of computers. Together with feminist sociology, it has had a considerable influence on the sociology of gender and sexuality. Its empirical, domain-oriented impact has been more significant than that of its pure theoretical tenets.

6. New Research Directions

The cultural forms induced by the electronic media (e.g., the culture of software programmers, of Internet users) have not been studied in depth until now. The changes in consumption patterns and market relationships brought about by the Internet, the rapid rise of popular electronic exchanges, the Internet's role in the globalization processes are topics which await deeper exploration too. It is to be expected that postmodernist sociology will evolve more and more in the direction of empirical, concrete field studies, focusing on these phenomena.

See also: Culture; Consumption, Sociology of; Cultural Relativism, Anthropology of; Feminist Epistemology; Feminist Theory: Postmodern; Globalization, Anthropology of; Identity: Social; Modernity; Political Discourse; Postmodern Urbanism; Postmodernism in Geography; Postmodernism: Methodology; Postmodernism: Philosophical Aspects; Relativism: Philosophical Aspects; Sociology, Epistemology of; Structuralism, Theories of

Bibliography

Baudrillard J 1972 *Pour Une Critique de l'Économie Politique du Signe*. [*For a Critique of the Political Economy of the Sign*]. Gallimard, Paris

Bauman Z 1992 *Intimations of Postmodernity*. Routledge, London

Etzioni A 1968 *The Active Society: A Theory of Societal and Political Processes*. Collier-Macmillan, London

Haraway D 1997 *Modest Witness @ Second Millenium: FemaleMan® Meets OncoMouse™; Feminism and Technoscience*. Routledge, New York

Lash S (ed.) 1991 *Poststructuralist and Postmodernist Sociology*. Elgar, Aldershot, UK

Latour B 1991 *Nous n'Avons Jamais Été Modernes*. [*We Have Never Been Modern*]. La Découverte, Paris

Lemert C 1997 *Postmodernism Is Not What You Think*. Blackwell, Cambridge, MA

Lyotard J F 1979 *La Condition Postmoderne: Rapport sur le Savoir*. [*The Postmodern Condition: A Report on Knowledge*]. Editions de Minuit, Paris

Mouffe C 1992 *The Return of the Political*. Verso, London

Richardson L 1997 *Fields of Play*. Rutgers University Press, New Brunswick, NJ

Ritzer G 1993 *The McDonaldization of Society: An Investigation into the Changing Character of Contemporary Social Life*. Pine Forge Press, Newbury Park, CA

Smith D E 1990 *Texts, Facts, and Feminity. Exploring the Relations of Ruling*. Routledge, London

Toynbee A 1947 *A Study of History*. Oxford University Press, London

Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon and Schuster, New York

A. Preda

Postmodernism: Methodology

If methodology is defined as referring to the foundations of knowledge, as how we perceive and understand reality, as well as how we study it, then there is not only a single implicit postmodern methodology. There are several. But all challenge the methodological assumptions associated with rigorous, modern social science inquiry, be it qualitative or quantitative (Dickens and Fontana 1994). Indeed, there are serious philosophical differences between postmodernism and the enlightenment norms of traditional science. Some postmodernists are uninterested in generalization, definitive explanation, replication, validity, reliability, etc. Neither prediction nor theory building are a major concern. They look to deconstruction as method. Deconstruction tears a text apart, revealing its contradictions and assumptions. Other postmodernists criticize modern methodology and seek to revise it, to build upon it. They look for alternatives to what they call the old, past, closed conventional social science methods inherited from the natural sciences. But they do so without entirely rejecting reason, consistency and coherence of argument. Although they abandon foundations, they retain what they call preferences.

Many postmodernists see the need for new standards for evaluating knowledge. Their criteria are likely to be subjective in nature, including, for example, flexibility, sensitivity, and interactivity. They look for beauty, strength, or force in a text. Some value elegance of expression and style, and seductiveness of content while others say that such criteria are community specific and so cannot be enumerated. Still other postmodernists willingly employ more direct valuations, such as negation of oppression, exploitation, and domination, or the affirmation of liberation, freedom, insubordination, and resistance as key standards (Rosenau 1992).

1. A Postmodern Methodological Focus

Postmodern methodology focuses on the margins and postmodern social scientists highlight the unique and the unusual. They concentrate on the enigmatic

and appreciate the unrepeatable. They seek to multiply paradox, to invent ever more elaborate repertoires of questions as much as to determine answers. They look to 'differ*a*nce' (Derrida 1972) rather than sameness, regularity, and synthesis. 'Differ*a*nce' is a structuring principle that tells us that definition rests not on some inherent characteristics of a thing but rather in terms of its negative and positive references to other texts. It also suggests that in the end, definition is never final, always postponed, deferred forever (Derrida 1981).

Postmodern methodology seems driven by substantive concerns quite different from those modern social scientists focus upon. Postmodernists direct their attention to everyday life and they gravitate toward mininarratives or stories. Postmodern social science is a 'systematic commentary on the knowledge of daily life' (Bauman 1990). This means examining what is assumed to be quite ordinary and attributing a new importance to it. Postmodern social science makes room for the unexpected, the mystical, and the metaphysical. Concepts such as fate and spirituality are, once again, legitimate (Griffin 1988).

1.1 Linking Substance and Method

If there is a close association of substance and method, then conventional social science methods are not helpful given the postmodern focus on the margins, the unexpected, the paradoxical, and the unrepeatable. Empirical methods require repeatability so that careful measurement can be carried out and statistical norms met. But the subject matter of interest to postmodernists does not recur necessarily and its very interest may be in the fact that it is transient, constantly in flux, and difficult to capture. Because of their substantive foci, these postmodernists require methods that are open and flexible, rather than highly structured, codified, and technically detailed. They must be adaptable to a great variety of situations because of the range of potential application. There is no easy methodological recipe for doing this type of research.

1.2 Going Beyond Qualitative Methods

Postmodernists would reject any argument that their methodology is simply qualitative social science in a different guise. The gamut of modern qualitative approaches developed over the years (from observation to participant observation) retains a preoccupation with generalization and replication as much as an appreciation of complexity. Those employing conventional qualitative methods in social science have invested enormous energy in separating out what is procedurally acceptable from the unqualified. All this is abandoned in the postmodern methodology that embodies an attitude of humility and respect. It does not try to convince others that its own methodological perspective is right, superior, truthful, objective, the best assessment, or the source of the only correct answer. It does not settle on solutions or give privileged voice to any particular, singular point of view. It does not attribute authority to authorship. It queries and interrogates constantly. It follows its own advice: no new metanarrative. A metanarrative is a global worldview, a 'mastercode' that assumes the validity of its own truth claims.

2. An Upbeat Postmodern Methodology

What methodological alternatives does postmodernism have to offer vis-à-vis modern social science methodology? It emphasizes personal wisdom, lived experience, and knowledge as rich sources of material that inform writing and research. We are counseled to look to music, art, poetry, and literature instead of census data and public opinion polls. A postmodern social scientist would have more confidence in feelings, faith, empathy, emotion, insight, sensitivity, intuition, imagination, subjective judgment, and introspection as method. Postmodern social scientists are obviously and openly more comfortable with these than with an exclusive confidence in reason and logic. But by praxis, if not discourse, some postmodernists communicate that their alternative methodologies are of most value if disciplined by reason and logic. This amounts to coming to terms with the best of the Enlightenment heritage but without being intimidated by it. It suggests a refusal to accept the authority of modern science but also a desire to salvage from it whatever is of interest, such as the possibility of an interactive and reciprocal relationship between, say, reliable knowledge and meaningful participation in the public sphere, the creation of social life and political action, the enhancement of personal development, self-expression, self-determination and collective praxis (Jennings and Graham 1996). Doing social science might also have to be fun, enjoyable, entertaining rather than somber, listless, and mundane (Bakhtin 1973, Todorov 1984). It would not seek to demonstrate the obvious so much as to highlight the overlooked and the unrecognized.

2.1 Methodology as Art and Literature

Postmodern social science is often viewed as vision and art more than trained observation. It disperses any clear distinction between self and other; fact and values; author, text, and reader. It is inspired by the cultural and contextual preoccupations of anthropology where the distinction between literature and

science was never clearly established. Anthropology had to be content with differing 'evocative devices of metaphor and imagery' because it was so difficult to arbitrate between different narratives from the past (Richardson 1988). Postmodern social science methodology centers on 'listening to,' and 'talking with' the other (Tyler 1986). The reader 'writes' the postmodern text and in so doing postmodern social science gives voice to those previously silent. It empowers through the methods of interpretation and criticism.

Postmodern methodology employs a literary capacity and it re-invigorates the mininarrative as a story that is worthy of notice even if it is not necessarily true in the modern sense of science. Storytelling is both a method and a means to emancipatory power (Newman and Holzman 1999). The need to reconcile various stories, narratives or research findings may or may not be urgent. The postmodernism tolerance for diversity and its indulgence of discrepancy reflect the fact that there is no rush to create, to construct, to accumulate, and to impose new knowledge in the modern sense.

2.2 Methodology as Interpretation

Postmodern methodology is interpretive but postmodern interpretation bears but slight resemblance to the more conventional social science forms of interpretation. Modern social science understands interpretation to be the careful consideration of data with the goal of locating patterns. Meaning is not arbitrary and some interpretations are assumed to be better than others. But neither does postmodernism resemble hermeneutic interpretation that is more subjective and less ambitious than modern social science's interpretation. Hermeneutic interpretation, unlike postmodern interpretation, seeks to probe the 'silences,' to uncover a deeper meaning, masked and hidden perhaps, but waiting to be discovered (Dreyfus and Rabinow 1983, Madison 1988). There is a current within postmodernism that says there simply is no 'meaning' to uncover. In many ways hermeneutic interpretation accepts truth, universality, and a shared discourse (a community of understanding).

Postmodern interpretation is *introspective*, *intersubjective*, and *anti-objectivist*, a form of individualized understanding. Postmodern theories of interpretation no longer 'ground,' they rather 'mediate (Kugler 1988). They range from 'interpretative reason' (Bauman 1990) to prayer and chanting as 'interpretative communication.'

2.3 Methodology for a Complex Intertextual World

Postmodernists see the world as so complex that any direct causal statements relating to reality would necessarily be simplistic. They argue that it is impossible to established temporal priority in a simple fashion. They similarly are suspicious of modern versions of time that clearly delineate the past from the present, and include a strict version of linear time or geography. Their preference is for a present-oriented social science that erases linear time and ignores strictly bounded forms of geography. Their analysis emphasizes postmodern intertextuality, including, for example, chaos theory with its view of nonlinear social dynamics, and a definitely more complex version of reality (Young 1992).

If the social universe for these postmodernists is intertextual, this has an impact on their methodology. Relationships, interconnections, and interactions are areas of substantial interest. Conventional research tools are hardly appropriate. Quantitative methods isolate and simplify what is being studied, holding as many extraneous variables as possible constant so as to tease out causal, predictive associations. These postmodernists seek methodologies that integrate what is being studied with all that is related to it even if this complicates the research task, because in the end explanation in any absolute sense may be impossible.

2.4 Relativism and Questioning Postmodernists

Postmodernism is plagued, because of its European philosophical precursors, with the charge of relativism (Gellner 1992) but many postmodernists defy this heritage and reject relativism. This is especially true for the postmodernists who come from a philosophically committed heritage: Marxism, feminism, environmentalism, peace movements, ecology groups, and religious studies (Griffin 1988, 1990, Kumar 1995, Leonard 1997). This leads them to argue that while deconstruction is an interesting first step to a postmodern methodology, it is limited because its goal is to undo all constructions. Deconstruction's intent is not to improve, revise, or offer a better version of the text, so they reject it. Postmodernism, in this view, must ultimately move to reconstruction and by extension, methodology remains essential.

3. Methodology Defined Narrowly

If methodology is defined narrowly as the rules and regulations of inquiry, then postmodernists must question the whole undertaking, arguing that there is no such thing as a post-modern methodology. 'Rules,' as such, are modern, not postmodern in character because they imply hierarchy, a series of good–bad evaluative criteria, an assessment on the basis of a set of agreed upon standards. Postmodernists reject rule-bound approaches as 'metanarratives.' Critics of

postmodernism, however, respond that the absence of rules is itself a rule, the rule that 'anything goes' (Feyerabend 1975, Heller 1987) and that there is, thus, a postmodern methodology.

4. Rejecting Methodology Altogether

Postmodernists, inspired by Heidegger and Neitzsche, are antiscience and anti-Enlightenment, and this leads to an absolute relativism and to methodological nihilism. This postmodern orientation emphasizes the hopelessness and meaninglessness of life; theirs is the pessimistic, dark side of postmodernism (Rosenau 1992).

Some such postmodernists practice epistemological impossibilism. They do not intend to provide 'clear reasoned defense.' They tell us their texts offer 'no consistent argument' and that 'not all points made are "necessary."' They boast that by design 'some essential matters are forgotten' (Corlett 1989). To insist that these postmodernists develop a self-conscious methodology is to 'do violence' to their views, to impose one's own perspective, and thus diminish them by forcing them to acknowledge your views (Ashley and Walker 1990a, 1990b, Lyotard 1988). In the end, from this postmodernist methodological perspective, we can never really know anything, not even our own feelings and emotions. These postmodernists speak of the death of the author (as agent) and of the absence of the subject (as a real person) with a specific identify. This means that they escape agency, accountability, and responsibility. They deny causality, question efforts to represent reality (through social science theory), and they contend that all interpretations are equally interesting (Miller 1981).

All textual meaning, all interpretation, is undecidable within this postmodern perspective, because there is no final sense of a text and everything is a text in this postmodern world. No interpretation can be regarded as superior to any other (Latour 1988). There is a multiplicity of readings and these postmodernists refuse to privilege one statement over another. 'Every understanding is a misunderstanding' (Culler 1982, Norris 1988). Within this intellectual framework there is no need to bother about methodology at all.

5. Conclusion

Postmodern methodologies will be judged in the long term by their product. In the arts and literature, postmodern methods can be fascinating, challenging, and interesting. They stimulate creativity and innovation as exemplified by the writings of Umberto Eco, Italo Calvino, Milorad Pavic, and Salman Rushdie. They dissolve the boundary between fiction and social critique as the works of Jean Baudrillard and Michel Foucault so well illustrate. The fruitfulness of post-modern methodologies in the applied social sciences is yet to be demonstrated. Many feel them to be seriously flawed and epistemologically dubious. Post-modern methods contributed little toward the solution of the greatest social problems of the twentieth century. But then, that is not what they set out to do; it was never the intent.

See also: Deconstruction: Cultural Concerns; Feminist Theory: Postmodern; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Postmodernism in Geography; Postmodernism in Sociology; Postmodernism: Philosophical Aspects; Symbolic Interaction: Methodology

Bibliography

Ashley R K, Walker R B J 1990a Reading dissidence/writing the discipline: Crisis and the question of sovereignty in international studies. *International Studies Quarterly* **34**: 367–416

Ashley R, Walker R B J 1990b Speaking the language of exile: dissident thought in international studies. *International Studies Quarterly* **34**: 259–68

Bakhtin M 1973 *Problems of Dostoevsky's Poetics*. Ardis, Ann Arbor, MI

Bauman Z 1990 Philosophical affinities of postmodern sociology. *The Sociological Review* **38**: 411–44

Corlett W 1989 *Community without Unity: A Politics of Derridian Extravagance*. Duke University Press, Durham, NC

Culler J 1982 *On Deconstruction: Theory and Criticism after Structuralism*. Cornell University Press, Ithaca, NY

Derrida J 1972 *Marges de la philosophie*. Editions de Minut, Paris

Derrida J 1981 *Positions*. University of Chicago Press, Chicago

Dickens D R, Fontana A (eds.) 1994 *Postmodernism and Social Inquiry*. Guilford Press, New York

Dreyfus H L, Rabinow P 1983 *Michel Foucault, Beyond Structuralism and Hermeneutics*, 2nd edn. University of Chicago Press, Chicago

Feyerabend P 1975 *Against Method: Outline of an Anarchistic Theory of Knowledge*. New Left Books, London

Gellner E 1992 *Postmodernism, Reason, and Religion*. Routledge, London

Griffin D R (ed.) 1988 *The Reenchantment of Science: Postmodern Proposals*. State University of New York Press, Albany, NY

Griffin D R (ed.) 1990 *Sacred Interconnections: Postmodern Spirituality, Political Economy, and Art*. State University of New York Press, Albany, NY

Heller A 1987 Mouvements culturels et changements des modeles de vie quotidienne depuis le deuxieme guerre. In: Corten A, Tahon M B (eds.) *La Radicalite du quotidien*. VLB editeur, Montreal, PQ

Jennings L E, Graham A P 1996 Postmodern perspectives and action research: reflecting on the possibilities. *Educational Action Research* **2**

Kugler P 1988 From modernism to post-modernism: Some implications for a depth psychology of dreams. *Psychiatric Journal of the University of Ottawa* **13**: 60–5

Kumar K 1995 *From Post-Industrial to Post-Modern Society: New Theories of the Contemporary World*. Blackwell, Oxford, UK
Latour B 1988 *The Pasteurization of France*. Harvard University Press, Cambridge, MA
Leonard P 1997 *Postmodern Welfare*. Sage Publications, Thousand Oaks, CA
Lyotard J F 1988 *L'inhumaine Causerie sur le Temps*. Galilee, Paris
Madison G B 1988 *The Hermeneutics of Postmodernity: Figures and Themes*. Indiana University Press, Bloomington, IN
Miller J H 1981 The disarticulation of the self in Nietzsche. *The Monist* **64**: 247–61
Newman F, Holzman L 1999 Beyond narrative to performed conversation. *Journal of Constructivist Psychology* **12**: 23–40
Norris C 1988 *Deconstruction and the Interests of Theory*. Pinter, London
Richardson L 1988 The collective story: Postmodernism and the writing of sociology. *Sociological Focus* **21**: 199–208
Rosenau P M 1992 *Post-Modernism and the Social Sciences*. Princeton University Press, Princeton, NJ
Todorov T 1984 *Mikhail Bakhtin: The Dialogical Principle*. University of Minneapolis Press, Minneapolis, PA
Tyler S 1986 Post-modern ethnography: from document of the occult to occult document. In: Clifford J, Marcus G E (eds.) *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA
Young T R 1992 Reinventing sociology. In: *Mission and Methods in Postmodern Phenomenology*. The Red Feather Institute, Weidman, Michigan

P. V. Rosenau

Postmodernism: Philosophical Aspects

'Postmodernism' is a term that defies simple definition. Although earlier uses are recorded, the term has been widely used since the 1960s to refer to a range of antimodernist attitudes and strategies in art criticism, architecture, and literary criticism. References to postmodernism in philosophy and social theory did not become common before they appeared in the work of Lyotard and Rorty during the early 1980s. Philosophical postmodernism quickly became associated with the idea of a 'crisis' in representation, abandonment of the belief in a universal subject of reason and value, and the rejection of unitary schemas of progress inherited from the Enlightenment. In place of these foundational motifs of modern thought, postmodernism emphasized difference, diversity and irreducible conflict in human affairs rather than unity and the possibility of harmony.

There are several reasons why postmodernism should be 'a particularly unstable concept' (Bertens 1995, p. 12). First, the concept of postmodernism stands in a variety of complex relations to other concepts such as modernism, poststructuralism, deconstructionism, modernity and postmodernity. Since the meaning of these concepts is equally uncertain, so our understanding of 'postmodernism' will be affected by the particular meaning attached to these related concepts. Second, to the extent that postmodernism is considered a cultural movement alongside others such as romanticism, positivism, or structuralism, then it too will encompass a variety of distinct aesthetic and intellectual strategies. Bertens' *The Idea of Postmodernism: a History* begins with illustrations of the sometimes contradictory ways in which the term has been employed. For example, by 1970 modern painting had long abandoned the idea that representation or narrative were essential to its art. In explicit reaction against modernism understood as the self-reflexive exploration of the formal possibilities of painting, postmodern painting involved a reprise of representation and narrative. In a similar fashion, postmodern architecture rejected the formalism of the postwar International style and re-embraced historical and vernacular styles. By contrast, at least in some of its manifestations, literary postmodernism involved a move in the opposite direction away from narrative and representation towards a form of radical self-reflexivity. Bertens concludes that, depending on the artistic discipline, 'postmodernism is either a radicalization of the self-reflexive moment within modernism, a turning away from narrative and representation, or an explicit return to narrative and representation. And sometimes it is both' (Bertens, 1995, p. 5). In view of the many different phenomena to which it has been applied, it is common for discussions of postmodernism, to begin, as this one has, with the claim that the term has no clear meaning (Eco 1985, p. 65, Rorty 1999, p. 262).

1. *Modernity/Modernism*

Just as modernism is often regarded as a response to the process of modernization which transformed the material conditions of European society during the nineteenth century, so postmodernism is often presented as a consequence of the restructuring of capitalism at the end of the twentieth century. The historical period of modernity which extends from the middle of the eighteenth through to the middle of the twentieth century saw the development of industrial capitalist infrastructure and means of production such as railways, automobiles, air transport, electricity, and telegraphic and telephonic communication. It also saw the growth and modernization of cities, the spread of public schools, hospitals, and prisons, and the development of bureaucratic and rational procedures for the government of social life. 'Modernism' is commonly used to refer to an artistic sensibility which embraced innovation and change. In its early stages, modernism involved a positive sense of living at the dawn of a new era. To be modern implied an heroic commitment to action and to revolutionizing oneself

as well as the means of artistic production, hence the dynamic of successive avant-gardes in the particular arts. During the latter part of the nineteenth and the early years of the twentieth century, modernism meant experimentation with the style and form of the various arts.

Theorists of postmodernity assert a parallel connection between the changes associated with the restructuring and globalization of late capitalism and the emergence of a postmodern culture. Two of the leading theorists, Fredric Jameson and David Harvey, argue that changes in technology, urban space, and forms of consumption have created the conditions of a new mode of experience of self and others (Jameson 1991, Harvey 1989). In particular, they argue that the 'compression' of social space and time which results from new electronic and communications technologies is reflected in a new cultural sensibility. Moreover, just as the modernist commitment to the destruction of traditional ways of life and thought was often linked to a tragic sense of loss, so some theorists of postmodernity argue that a similar kind of homelessness is generated by the late twentieth century mutations in our experience of space and time. Jameson is among those who take a negative view of the cultural experience of postmodernity on the grounds that it does not sustain a sufficiently robust sense of self and history. He argues that the rapid evolution of postmodern urban space has overtaken 'the capacities of the individual human body to locate itself, to organize its immediate surroundings perceptually, and cognitively to map its position in a mappable external world' (Jameson 1991, p. 83). On his account, postmodernity is experienced as incapacity in the face of complex and discontinuous social spaces and as disorientation in the absence of a rich sense of the past or future. The result is a widely documented fragmentary and schizophrenic experience of subjectivity, along with a tendency towards affectless repetition or 'blank parody' of past forms of artistic and cultural production.

For others, modernist and postmodernist alike, the appropriate affective response to change is much less one of nostalgia for older ways of being at home and more a question of orientation towards the future. Nietzsche, who is in many respects the source of much postmodern philosophy, took the condition of 'homelessness' as a metaphor for the condition of those, like himself, who looked forward to 'the strengthening and enhancement of the human type': 'We children of the future, how *could* we be at home in this today?' (Nietzsche 1974, p. 338). In their efforts to make positive sense of the experience of being estranged from the social world in which they must live, postmodernists often have recourse to the concept of irony. Richard Rorty calls ironists those who are conscious of the contingency and mutability of the 'final vocabularies' employed to make sense of their lives, yet nonetheless committed to the values which those vocabularies embody (Rorty 1989, p. 73). Umberto Eco argues that the postmodern attitude is encapsulated in the irony which acts and speaks in full awareness of the degree to which all our ways of acting and speaking involve repetition of the past: 'the postmodern reply to the modern consists of recognizing that the past, since it cannot really be destroyed, because its destruction leads to silence, must be revisited, but with irony, not innocently' (Eco 1985, p. 67).

While postmodernism is supposed to involve rejection of the presuppositions of modern art, philosophy, and social theory, commentators and users of the term hesitate over whether postmodernism is primarily a temporal or a typological concept. The prefix and the manner in which the term has been employed in the arts suggest a temporal reference, even though estimates of the beginning of the postmodern era range from the 1860s to the 1980s. But some theorists of postmodernism in the arts and culture treat it as simply the radicalization of tendencies already present in the modern. In philosophy, Rorty argues that, to the extent that postmodernism stands for anything distinctive, it refers to a pluralism shared by a range of modern thinkers such as Nietzsche and William James (Rorty 1999). In similar fashion, Jean-Francois Lyotard argues that postmodernism in painting stands for an aesthetic of the sublime which embraces the ever-renewed attempt to 'present the unpresentable' (Lyotard 1983). For these reasons, Eco suggests that postmodernism should not be chronologically defined but rather treated as 'an ideal category—or better still, a *Kuntswollen*, a way of operating. We could say that every period has its own postmodernism, just as every period would have its own mannerism' (Eco 1985, p. 66).

2. *Unity and Plurality*

A feature of many cultural responses to the transient, fleeting and ephemeral character of modern life is their attachment to ideas of unity, whether in the object to be represented or the artefact to be produced. Thus, for example, modernists believed that a machine or a city could be a unified, smoothly functional object, and a novel or work of art could be a complete, self-contained meaningful entity. Where there is experimentation with diverse modes of representation, this is undertaken in the attempt to convey a single complex and multifaceted reality. And even where there is a dizzying succession of conceptions of the nature of a particular art such as painting, there is nevertheless a conviction at each stage that at last the truth has been found.

Even though 'modernism' does not have the same currency as a classificatory term in philosophy and social theory as it does in the arts, there are epistemological parallels to this modernist attitude towards

unity. Modernist thought would be that which tolerated perspectivism only on the assumption that there is a single, intelligible reality to be presented. Even when there is an acceptance of surface incoherence, as in Marx's account of the contradictions of capitalism or Freud's account of conscious behavior, this is balanced by the discovery of an underlying structure which produces such contradictory effects. In this sense, Harvey argues that modernism 'took on multiple perspectivism and relativism as its epistemology for revealing what it still took to be the true nature of a unified, though complex, underlying reality' (Harvey 1989, p. 30). By contrast, postmodernism is associated with the acceptance of plurality in both ways of being and ways of knowing. In opposition to the ideal of unified objects of knowledge, whether texts or artefacts, postmodernism refers instead to networks, open systems, or the dissemination of meaning. In opposition to the modernist ideal of unified science, postmodernism upholds the idea of irreducible difference and even incommensurability among the varieties of knowledge.

3. *Poststructuralism and Postmodernism*

Postmodernism coincides with the work of a number of French thinkers who came into prominence at the end of the 1960s, notably Baudrillard, Deleuze, Derrida, Foucault, and Lyotard. By different intellectual paths, these thinkers rejected the idea that theories and concepts simply represent a pre-existing reality. For some, such as Derrida or Baudrillard, this conclusion was reached via a radicalization of Saussurian theses about the arbitrariness of signs and the manner in which their sense is derived from differential relations to other signs. For others, such as Foucault, similar conclusions were reached via Kantian views of the role played by language and thought in the construction of its own objects. A consequence of Foucault's historicization of the Kantian idea that objects of knowledge are determined in part by the underlying rules governing discourse is that 'truth' acquires a history and can no longer be measured solely against an independent reality.

While it is true that the work of Foucault and others does imply a form of historical relativism with regard to truth, it is overstating the case to suggest that the resultant 'crisis in representation' amounts to 'a deeply felt loss of faith in our ability to represent the real, in the widest sense' (Bertens 1995, p. 11). It is not our ability to represent reality but rather what Deleuze calls the representational 'image of thought' that is impugned by poststructuralist epistemology (Deleuze 1994, pp. 129–67). Foremost among the influential philosophical assumptions which make up the representational image of thought is the idea that thought naturally seeks truth. For Rorty, too, what lies at the heart of postmodernism in philosophy is the abandonment of the correspondence theory of truth and the 'theologicometaphysical belief that Reality and Truth are One—that there is One True Account of How Things Really Are' (Rorty 1999, p. 262). A less negative characterization might say that postmodernism does not so much reject the idea that thought represents reality as embrace more complex views of what representation entails. In this manner, Foucault's genealogies of punishment and sexuality drew attention to the ways in which what passes for truth in the social sciences is bound up with the exercise of power over individuals and populations. More generally, postmodernism has drawn attention to new aspects of social reality, such as the systems of thought and language which condition public discourse. It has also given rise to new kinds of analysis of cultural and social phenomena: genealogical, deconstructive, narratological, and so on. In these ways, postmodernism has given rise to new research methods and new objects of analysis in anthropology, sociology, and other social sciences (Clifford and Marcus 1986, Dickens and Fontana 1994). In addition, postmodernism has been widely taken up in jurisprudence and the philosophy of law, where it has been described as the most influential trend at the end of the twentieth century (Litowitz 1997, p. 1).

4. *Universalism and Freedom*

One of the most influential formulations of postmodernism in philosophy was that provided in Lyotard's *The Postmodern Condition* (Lyotard 1984), first published in 1979. Here, Lyotard uses 'modern' to refer to any form of knowledge which seeks legitimation by appeal to some 'grand narrative, such as the dialectics of the spirit, the hermeneutics of meaning, the emancipation of the rational or working subject, or the creation of wealth.' He then defines the postmodern attitude as one of 'incredulity towards metanarratives' (Lyotard 1984, pp. 23–4). Lyotard is not concerned in this essay with postmodernity understood as an historical epoch, since his inquiry is limited to 'the condition of knowledge in the most highly developed societies' (Lyotard 1984, pp. 13). Rather, his aim is to show that the supposed transcendent status of scientific knowledge is an illusion sustained by a certain narrative language game. He argues that the privileged status of scientific knowledge is destined to disappear once it is accepted that this status depends upon a narrative which privileges science over other kinds of language game. This narrative of legitimation has either followed the French Enlightenment link between the pursuit of knowledge and human emancipation, or the German Idealist construal of history as the process of the realization of reason through the sciences. Lyotard asserts not only that these 'metanarratives' have lost their appeal but that the very idea of transcendent legitimation has given way to forms of legitimation which are local and immanent in particular social

practices and forms of life. The postmodern condition therefore is the state in which we must accept this plurality of micronarratives and address the conflicts that arise other than through the imposition of a single master narrative.

In *Modernity—An Incomplete Project*, Habermas (1983) provided an influential response to Lyotard and other poststructuralist thinkers. He characterized the philosophical project of modernity as the desire for unification of the three spheres of cultural value distinguished by Kant—the theoretical, the practical, and the aesthetic—and for the application of the results of these distinct kinds of judgment to the improvement of everyday life. According to this view, the project of modernity consisted in the effort of Enlightenment thinkers 'to develop objective science, universal morality and law, and autonomous art, according to their inner logic. At the same time, this project intended to release the cognitive potentials of each of these domains to set them free from their esoteric forms. The Enlightenment philosophers wanted to utilize this accumulation of specialized culture for the enrichment of everyday life—that is to say for the rational organization of everyday social life' (Habermas 1983, p. 9). The major obstacle preventing the realization of this project, according to Habermas, lay in the manner in which these spheres have become separated from one another, both institutionally and in their internal logics and forms of argumentation, and in the manner in which they have become increasingly separated from a lifeworld dominated by the instrumental rationality of the theoretical sphere.

Habermas's solution to this problem appealed to the procedural rationality implicit in language use, or what he calls 'communicative reason,' as a means to restoring the possibility of intersubjective consensus. Both his diagnosis and his solution to the problem of modernity are founded upon the Enlightenment belief that universal standards of judgment are necessary for the development of both reason and moral progress. By contrast, postmodern thinkers reject the idea of progress in human affairs in favor of the idea that there are many ways in which the present differs from the past. In common with Lyotard, they reject the idea of universal standards of judgment, value or 'legitimation.' Although Habermas characterizes French thinkers such as Foucault, Derrida, and Lyotard as 'young conservatives' (Habermas 1983, p. 14), most of them do not wish to abandon the ideals of freedom and emancipation but only the approach which links these to universality, unity, and consensus and does so in the name of the Enlightenment.

In response to Habermas's criticisms, Foucault presented his practice of genealogy as serving the cause of freedom, while suggesting that it represented a critical ethos or attitude towards the present which had always belonged to the Enlightenment. Genealogical criticism does not seek to identify the universal structures of human knowledge and moral action, but rather takes the form of an historical investigation into the means by which we have been constructed as certain kinds of subject. Against the presumption that human reason is an unchanging and undivided faculty, Foucault views reason as divided among the distinct rationalities embodied in particular systems of thought and practice such as 'medicine,' 'discipline,' or 'sexuality.' The aim of this type of historical investigation, Foucault suggests, is to determine 'in what is given to us as universal, necessary, obligatory, what place is occupied by whatever is singular, contingent, and the product of arbitrary constraints' (Foucault 1984, p. 45). Thus, in contrast to the idea that progress in human freedom depends upon the extension of the sphere in which universal rationality governs social life, Foucault presents progress in human freedom in negative terms as the escape from particular forms of domination. For many poststructuralist theorists, freedom is not linked to the achievement of universal standards of value or to the hope that dissensus will eventually give way to consensus. Rather, it is dissensus and the experience of irreducible difference which are the conditions of freedom and creativity or change in human affairs.

5. *Difference and Multiplicity*

In his later work, Lyotard introduced the concept of the differend in order to characterize the postmodern condition with respect to political and moral judgment as the situation in which conflict between two parties cannot be resolved for lack of a rule of judgment common to both (Lyotard 1988, p. 9). Foucault also described his critical attitude towards the present as one which endorsed differentiation from the past rather than movement towards some future identity (Foucault 1984). However, the recent origins of this concern with difference must be sought in the philosophical work of Deleuze and Derrida published during the 1960s. Both sought to elaborate a non-dialectical concept of difference, which would see it neither as the simple contrary of identity, nor as 'dialectically' identical with identity (Descombes 1980, p. 136). Hegel was a focus of their criticism because he represented the culmination of a metaphysical tradition which treated identity as primary and difference as the derivative or secondary term. In *Difference and Repetition*, Deleuze (1994) explicitly rejects the link which Hegel forged between difference and contradiction, arguing that contradiction is not the condition or ground of difference but on the contrary: 'It is not difference which presupposes opposition but opposition which presupposes difference, and far from resolving difference by tracing it back to a foundation, opposition betrays and distorts it' (Deleuze 1994, p. 51). Limitation or opposition is a distortion of difference, according to Deleuze, because difference in

itself implies 'a swarm of differences, a pluralism of free, wild or untamed differences' (Deleuze 1994, p. 50). Similarly, in *Of Grammatology*, Derrida (1976) drew upon Heidegger's concept of the ontological difference between being and beings and Saussure's concept of the differentiation which gives rise to units of signification in a sign system to draw attention to the movement of deferral and differentiation (*différance*) which underlies all production of meaning (Derrida 1976, pp. 56–65).

Both Deleuze and Derrida were concerned with the problem of how to conceive of the identity of objects such as meanings, texts or signs which are subject to continuous variation, or how to envisage a form of unity which is not identical with itself. To this end, Derrida relied upon a series of concepts such as *différance*, the trace and iterability, while Deleuze relied upon a concept of open or variable multiplicity derived from Bergson. The concept of multiplicity, he and Guattari suggest, 'was created precisely in order to escape the abstract opposition between the multiple and the one, to escape dialectics, to succeed in conceiving the multiple in the pure state, to cease treating it as a numerical fragment of a lost Unity or Totality or as the organic element of a Unity or Totality yet to come, and instead distinguish between different types of multiplicity' (Deleuze and Guattari 1987, p. 32). In his collaborative work with Guattari (Deleuze and Guattari 1987), the concept of abstract multiplicities provided the formal structure for the elaboration of a series of philosophical concepts of such things as desire, language use, and forms of social organization. Deleuze and Guattari defend their idiosyncratic practice of philosophy as the invention of new concepts in order to bring about new ways of speaking and acting.

A distinction is often drawn between 'postmodern' theorists, such as Deleuze–Guattari, Derrida, and Foucault, and theorists of postmodernity such as Jameson and Harvey. The theorists of postmodernity tend to adopt a broadly Marxist approach to society and culture, and to the cultural phenomena they see as representative of the social totality, in order to pass judgment. Their 'modernist' social theory presupposes that a total picture is possible and that cultural phenomena can be understood as effects of a certain economic 'logic.' They proceed on the assumption that critical distance is possible and that an appropriate interpretative framework will provide such distance. By contrast, the 'postmodern' theorists tend to abandon economic or other forms of historical determinism, along with the idea that a totalizing theory of society and history is possible or desirable, in favor of the idea of multiple, overlapping, and divergent histories written from the distinctive perspective of particular social groups, peoples, or interpretative frameworks (power, desire, the body, etc.). This distinction is not absolute, since there are those such as Baudrillard, Foucault, and Lyotard who might be included in both camps although at different stages of their careers. However, it is characteristic of postmodern theorists that their work involves a self-conscious attempt to theorize in a different manner, while also searching for critical strategies which do not rely upon a normative standpoint outside or apart from the object of criticism.

6. *Simulacra and Simulation*

Philosophies which seek to make difference an object of affirmation must do more than reverse the traditional hierarchy between identity and difference. The mere inversion of hierarchy does not change the fundamental relation between the elements involved, nor does it change the nature of those elements. For this reason, Derrida's practice of deconstructive philosophy always envisaged a further stage after the initial hierarchy has been overturned which would involve 'the irruptive emergence of a new "concept," a concept that can no longer be, and never could be, included in the previous regime' (Derrida 1981, p. 42). In the course of his essays he produced or drew attention to a series of such paradoxical concepts, including the trace, writing-in-general, metaphoricity, and Plato's *pharmakon*, all of which served to disrupt conceptual oppositions and hierarchies in philosophy. Similarly in *Plato and The Simulacrum* (Deleuze 1990, pp. 253–66), Deleuze argued that the concept of simulacra served to undermine or 'overturn' the primacy of identity as this was embodied in Plato's conception of a model or form of which reality is only an imperfect copy.

However, it was the sociologist Jean Baudrillard who took up the concept of simulacra and applied it to aspects of everyday life. In his earlier work, Baudrillard had argued that industrial society at the end of twentieth century had entered a new phase in which consumption rather than production served to structure social relations. He argued that the Marxian view of commodities as useful items produced in order to realize exchange value must be replaced by a semiotically informed view of commodities as signs, produced in order to realize their 'sign value' as differential elements of a social code of signification. In a series of essays published during the early 1980s, Baudrillard built upon this analysis and other accounts of the power of mass media to argue that society had entered a new era in which simulation had replaced representation as the dominant principle of signification. The essays translated in *Simulations* (Baudrillard 1983) combined a Nietzschean inspired critique of metaphysics with a descriptive account of the present stage of social life as a regime of simulacra in which the distinction between reality and the forms of its representation has broken down.

According to Baudrillard, there are many areas of social life, such as the media, where signs or other

representations do not faithfully represent a prior reality but function in its stead. For practical purposes we live in a hyper-real social world of representations without referents, images without originals. The proliferation of media transmitted information and the speed with which it circulates creates a perpetual cultural present in which contemporary society has begun to lose its capacity to retain its own past.

7. *Criticism Without Foundations*

While critics object to the nihilistic, impressionistic and highly selective nature of Baudrillard's descriptions, his analyses of technological, media, and political events as simulations served to draw attention to this pathological feature of the experience of social reality at the end of the twentieth century. His essays have been criticized for their failure to provide any stable foundation for criticism. However, Baudrillard openly acknowledges their ironic and hyperbolic style, describing them as 'theory–fictions' intended to provoke and to challenge contemporary common sense rather than convince by evidence or argumentation. Such a provocative function is often invoked in defense of postmodern theory. At best, it is argued, postmodern approaches serve to problematize established concepts and to draw attention to ways in which traditional universalist and humanist assumptions function to exclude or marginalize differences.

Postmodern theory is also criticized for its limited and negative approach to normative issues. Critics assert that the anti-universalism of Foucault, Lyotard, and others is incapable of providing grounds for moral or political judgment much less comprehensive programs for social reform. Defenders argue that their aim is both more limited and more modest, namely to work on the limits of what it is possible to think or to say and thereby to contribute to the emergence of new norms which may also become bases for judgment. To those who argue that the task of problematizing existing ways of thinking or challenging existing norms is ultimately of limited importance, postmodernists reply that not to do so is to limit the possibilities for thinking and acting differently. For this reason, they argue, the ethos or *Kunstwollen* of postmodernism should be seen as a necessary element of every genuinely critical approach.

See also: Foucault, Michel (1926–84); Phenomenology: Philosophical Aspects

Bibliography

Baudrillard J 1983 *Simulations* [trans. Foss P, Patton P, Beitchmann P]. Semiotext, New York

Bertens H 1995 *The Idea of Postmodernism: a History*. Routledge, London

Best S, Kellner D 1991 *Postmodern Theory: Critical Interrogations*. Macmillan, Basingstoke, UK

Clifford J, Marcus G E (eds.) 1986 *Writing Culture: the Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA

Deleuze G 1990 *The Logic of Sense* [trans. Lester M, Stivale C, Boundas C, ed.]. Columbia University Press, New York and Athlone Press, London

Deleuze G 1994 *Difference and Repetition* [trans. Patton P]. Athlone Press, London and Columbia University Press, New York

Deleuze G, Guattari F 1987 *A Thousand Plateaus: Capitalism and Schizophrenia*. [trans. Massumi B]. University of Minnesota Press, Minneapolis, MN and Athlone Press, London

Derrida J 1976 *Of Grammatology* [trans. Spivak G C]. Johns Hopkins University Press, Baltimore, MD

Derrida J 1981 *Positions* [trans. Bass A]. University of Chicago Press, Chicago

Derrida J 1982 *Margins of Philosophy* [trans. Bass A]. University of Chicago Press, Chicago

Descombes V 1980 *Modern French Philosophy* [trans. Scott-Fox L, Harding J M]. Cambridge University Press, Cambridge, UK

Dickens D R, Fontana A 1994 *Postmodernism and Social Inquiry*. Guildford Press, New York

Eco U 1985 *Reflections on the Name of the Rose*. Secker and Warburg, London

Foucault M 1984 What is enlightenment? In: Rabinow P (ed.) *The Foucault Reader*, 1st edn. Pantheon, New York, pp. 32–50

Habermas J 1983 Modernity—an incomplete project? In: Foster H (ed.) *The Anti-aesthetic: Essays in Postmodern Culture*. Bay Press, Port Townsend, WA, pp. 3–15 [Reprinted 1981 version of Modernity versus postmodernity. *New German Critique* **22:** 3–14]

Habermas J 1987 *The Philosophical Discourse of Modernity* [trans. Lawrence F G]. MIT Press, Cambridge, MA

Harvey D 1989 *The Condition of Postmodernity: an Enquiry into the Origins of Cultural Change*. Blackwell, Oxford, UK

Jameson F 1991 *Postmodernism, or, the Cultural Logic of Late Capitalism*. Duke University Press, Durham, NC

Litowitz D E 1997 *Postmodern Philosophy and Law*. University Press of Kansas, Lawrence, KS

Lyotard J-F 1983 Answering the question: what is postmodernism? In: Hassan I, Hassan S (eds.) *Innovation/Renovation: New Perspectives on the Humanities*. University of Wisconsin Press, Madison, WI, pp. 329–41

Lyotard J-F 1984 *The Postmodern Condition: a Report on Knowledge*. University of Minnesota Press, Minneapolis, MN

Lyotard J-F 1988 *The Differend: Phrases in Dispute*. University of Minnesota Press, Minneapolis, MN

Nietzsche F 1974 *The Gay Science* [trans. Kaufmann W]. Vintage Books, New York

Rorty R 1989 *Contingency, Irony, and Solidarity*. Cambridge University Press, Cambridge, UK

Rorty R 1999 Afterword: pragmatism, pluralism and postmodernism. In: *Philosophy and Social Hope*. Penguin, London, pp. 262–77

P. Patton

Postsocialist Societies

'Postsocialist societies' is the general term for the former socialist countries which are in the course of full or partial restructuring of social life. The problem

of their definition (in many cases as 'postcommunist societies'), which ignores substantial differences in their past, is examined, along with certain common and significant problems that have acquired a specific form in the 'process of transition.'

1. *Definition*

The term 'postsocialist societies' itself raises a serious problem: societies which were considered (call themselves or identified as) socialist, and which have renounced the 'socialist course of development' since the 1989/90 changes in government, have practically no content of their own. These societies are classified in one and the same group, defined in one and the same way from the perspective of their past—their existence as 'socialist societies'—which they have renounced. On the other hand, this definition from the perspective of their past suggests that they are presumed to be societies in the course of self-identification—seen as transition from totalitarianism to democracy, from centrally planned to market economy—which have embarked on the arduous course towards building civil society.

This raises a serious question of the possible criteria by which a concrete national society would shed its 'postsocialist' label, that is, would have ultimately defined itself and be defined from the perspective of the present rather than of the past. At what stage does the present existence of the respective society provide sufficient grounds for regarding it as a society with sustained development based on values asserted in it (and shared by its citizens)?

Changes in the form of government are a necessary but not a sufficient prerequisite for radical restructuring of social life, for significant and sustainable changes in the concepts of the world, of human relations, and individual actions. That is precisely why the term is 'postsocialist societies' rather than 'postsocialist states'—the form of state government may or may not be socialist, whereas the declared renunciation of the 'socialist course of development' in itself does not free a society of the values, ideas, and expectations accumulated over the years, nor of the established relations and groups, and the very attitudes to and mechanisms of their establishment. At the same time, the course of self-identification of the 'postsocialist societies' is indefinite—their point of departure is known, but one cannot know for sure what they will ultimately become.

Of course, this is a generalization only, since the majority of 'postsocialist societies,' those in Europe, are expected to embrace precisely democratic values, a market economy, and civil society as the pillars of self-identification— most of them have not merely declared this but are making real progress in this respect. They are apparently aware that they will no longer be labeled as 'postsocialist societies' once they complete the process (already under way with the majority) of accession to the European Union (EU). In other words, for those national societies the criteria of recognition of a completed process of self-identification are the same as the criteria of fully-fledged EU membership, which in itself means that they have already chosen a course of development, that their development is predetermined—all they have to do is attain the ultimate objective. This does not apply to the non-European 'postsocialist societies' such as China, or to many of the ex-Soviet republics (some of which are European in purely geographical terms); nor does it apply to the special case of 'rump' Yugoslavia.

2. *Differences from the Past*

The term 'postsocialist' creates a false impression of uniformity of the societies in question (even though this definition is indefinite itself) and, respectively, of uniformity in the process of departure from their previous development as socialist. This idea is largely based on the previous term 'socialist societies' which, along with everything else, has the ideological function of hiding or belittling the internal distinctions of the respective 'socialist societies,' as well as of rendering as entirely insignificant, from the perspective of their existence as 'socialist,' the differences in their previous development and the concrete process of their 'transformation' to socialism. They are 'socialist societies,' and all details of the history of their transformation to socialism and of the national specificity of their existence are insignificant. Yet it is precisely those formerly 'insignificant' differences (declared and considered insignificant, but to a large extent also virtually insignificant from the perspective of 'the general line' of development, of the rallying of the respective societies into the 'socialist camp'), which also largely determine the form of renunciation of the 'socialist course of development,' that have become significant accelerators or impediments—generally speaking, significant determiners of the national specificity of the existence of the concrete societies as 'postsocialist,' as well as of the pace and consistency of the transition period, of the ways of reaching national consensus on the parameters of the change, and of the actual concept and definition of 'national consensus.'

Forgetting the history of the formation of a particular nation-state and the way in which it 'embarked on the road to socialist development,' and/or failing to take into consideration the influence of national specificity on the existence of the respective society as socialist, usually results in overemphasizing the impact of international institutions (economic and financial, as well as political) whose actions are guided by economic and geopolitical interests, as well as of the

role of the personality of the respective political leaders (in the concrete 'postsocialist society'). Thus the changes and their pace in 'postsocialist societies' (especially in Europe) prove directly and crucially dependent on 'foreign' interests or the individual abilities of politicians. Consequently, individual/group responsible action proves impossible or, rather, pointless. Even if it were possible, its ineffectiveness is readily justified by counteracting 'external forces' or inactive 'internal factors.'

Social scientists, among others, are accountable for the formation of explanatory models centered on the two above-mentioned variables. In an effort to explain the transition period and its problems, they deviate from their scientific position under pressure from their partisan allegiance; concentrate on the present, 'forgetting' history; resort to biased selection of historical facts to legitimate their interpretation of events; build a macroexplanatory model of transition in general, ignoring national differences; or cite separate facts or data from opinion polls without the historical-theoretical framework requisite for their interpretation.

The production and public presence of such 'explanatory models' of the problems of the transition period, actually generate further tensions in the postsocialist society. First, they are explanatory and deliberately created for the purpose of introducing ultimate clarity among ordinary citizens about current developments; second, they are even more convincing on the strength of their authors' scientific prestige; third, they are qualitatively heterogeneous—rather than just different—interpretations, and at the same time each one of them claims to be the only truthful interpretation. This lopsided way of 'understanding' developments, history, the origin of problems in social life, of events and political actions, proves to be yet another factor (along with political allegiance and social position) that antagonizes individuals and groups.

3. *Present Similarities or Key Problems of Transition*

Two ranges of problems may be identified: problems associated with identity and problems associated with social interaction. Although they are not confined to postsocialist societies only, these problems have acquired special significance and a specific form. They may also be generalized as 'political,' 'economic,' or 'social' from a macrosociological perspective on social reality, on the processes occurring within the latter, and their specific intertwining. The microsociological perspective proposed here is based on the understanding that the restructuring of a society which has renounced the socialist principles and norms (or, generally speaking, any society in the course of self-identification as something that is qualitatively different) proceeds from the presumption of the major significance of the changes in the ways of thinking and action and is seriously impeded by the continuing effect of previous models of thinking of the world and the others, patterns of behavior, and mutual relations of individuals and groups.

3.1 *Problems of Identity*

These problems concern apparently quite heterogeneous phenomena which, however, are experienced by the individual or group in a similar way as problems of self-identification: from 'nationalism,' which has become indicative of postsocialist societies, to 'unemployment' and 'poverty,' which are rarely treated as problems of identity (defined foremost as social and economic problems).

The preoccupation with national/ethnic identity, which has apparently divided (albeit in quite different ways) the postsocialist societies, stems from two specific features of the ideology upheld in socialist society: national/ethnic specificities within the boundaries of a separate state and between the states from the 'socialist camp' are insignificant—the significant self-identification is 'socialist citizens' (which does not mean that ethnocultural differences have been forgotten within the separate states, or that they never were an undisclosed reason for actual social inequality); the significant difference is directly associated with significant self-identification—the localization of everything 'strange,' 'there,' beyond the 'iron curtain.'

The disintegration of 'the camp' and the fall of 'the wall' dividing the two worlds, inevitably rejecting the validity of significant self-identification and difference, generates the need of 'national' self-identification. Whereas the question 'Who are we?' focuses the attention inwards, on the 'national community' and back in history, on the national community's existence and formation before its 'depersonalization,' the attempt at defining 'us' (who are united within a state) has exposed differences and has led to intolerance and discrimination against different ethnic and/or religious minorities. This process, common to all postsocialist societies, has acquired qualitative specificity (largely determined according to the constitution of the respective 'socialist societies'): from violent disintegration of multinational states (the Soviet Union and Yugoslavia), to the peaceful division of Czechoslovakia, to the more or less conflict-ridden confrontations between 'national majority' and 'ethnic/national minorities' in Bulgaria, Romania, Hungary, and elsewhere. Yet the common problem of emphasizing internal differences (by origin, language, religion) has remained beyond the qualitative specificity: internal differences which are thought to be significantly distinctive of people to the point of their

existence within more or less closed groups, and perception of respective individuals as significantly predetermined by their 'belonging' to those groups. Regardless of the self-perception of the respective individuals, 'the others' ascribe identity to them and treat them (engage in or avoid relations with them) from the perspective of this identity.

The problematic process of identification with European (Western) values unfolds against the background of these 'internal' problems of self-identification and ascription of identity. This process is problematic mainly because of the complex and contradictory combinations of three commonplace ideas: 'We are European,' 'We have had a tragic destiny (not without the involvement of the West), now the West should help us become like them,' 'Once we are admitted to the European Union, we shall become Europeans.' Although the images and expectations underlying the apparently unambiguous images of 'Europe,' 'the West,' 'we are European,' 'we shall become Europeans' are quite different, their focus is arguably on 'become like them,' whereas the principal dimension of the process of 'becoming' is the expected well-being.

This is not accidental. And it is directly associated with another aspect of the identity crisis in postsocialist society—unemployment and poverty. These two related but different problems are doubtless indicative of the scope of the economic crisis, but their personal aspect concerns the deepest layers of individual identity, largely because the status of being unemployed is unfamiliar, virtually absent in the experience of the 'socialist individual,' whereas poverty is a process which individuals who have found themselves in a postsocialist society do not associate with democracy and a market economy.

3.2 Problems of Social Interaction

The change (or, rather, the orientation towards change) in the principles of organization of social life triggers tensions at all levels of social interaction. At the level of the individual, these tensions take the form of a large-scale crisis in human relations. At the level of institutionalized interactions, they are manifested in 'corruption' of significant proportions.

The 'crisis in human relations' could be reworded too—as a moral and a political crisis, as well as a problem of interethnic relations. For the everyday person, however (insofar as we can talk about an everyday person in a world where 'what is taken for granted'—if anything at all—is always questioned), this problem appears as tensions in relations, as a failure to understand the 'other' (or vice versa), avoidance of certain others, isolation of and from strangers, and so forth. That problem covers the whole range from the closest social environment—family, relatives, friends—to neighbors and colleagues, to functional relationships with anonymous others.

The problem generally defined as a 'crisis in human relations' actually results from the intertwining and superimposing of diverse manifestations and echoes of political, economic and cultural values and other changes on the macrosocial level. The realized opportunity of being different in our political preferences and allegiance, in our economic status, of choosing between different cultural models, of adhering to different values, has inevitably destroyed established models of family, friendly, neighborly, business, and interethnic relations. Yet 'new' models may appear only as a result of a long and complex process of construction, mutual adjustment, and assertion—they are not something that can be borrowed (imported, grafted), or something that becomes valid simply because it is (or appears to be) 'better.'

Similar to nationalism, corruption has become emblematic of postsocialist societies. And just like nationalism, corruption is not an unfamiliar phenomenon in Western societies. Moreover, corruption is an affliction of democratic society insofar as the latter has sufficiently large opportunities for responsible decision making, instead of the strictly prescribed rules for any action which are typical of any totalitarian society and relieve its members of responsibility. Yet the opportunity for responsible decision making is also an opportunity for irresponsible action. In the postsocialist society, the complex and contradictory intertwining of still effective old schemes of control and decision-making, and the deficiency of new norms (both on the institutional level of laws, regulations, ordinances, etc., as well as on the level of individual and group consciousness—role models, values, notions, etc.) turn corruption into an epidemic. The affliction acquires huge proportions due to, *inter alia*, the internally contradictory perception of corruption: on the one hand, it is seen as an inevitable relation in a situation of transition, of reformulating and reregulating of institutions, social positions, rules, and so forth, which are invisibly present in the social sphere and are manifested in status, security, possessions, and money of those who have effected the replacement; on the other, corruption is embodied in the bribe-taker, insofar as bribe-takers are those who have rare resources, by rule of power, which many other persons would like to use. However, the bribe-giver is simply one of many: bribe-givers do not control the 'rules' of replacement, the way of negotiations, the size of the bribe to the bribe-taker, they are simply 'this person' who has happened to be in this position and, in this sense, anyone may end up in this position. Thus the bribe-taker is regarded as an irresponsible power-holder and the bribe-giver as 'any one of us' forced by the circumstances, and therefore not responsible for their involvement in relations of corruption.

See also: Democratic Transitions; Eastern European Studies: Culture; Eastern European Studies: Economics; Eastern European Studies: Politics; Ethnic Conflicts; Identity: Social; Minorities; Nationalism, Sociology of; Poverty, Sociology of; Prejudice in Society; Social Change: Types; Socialist Societies: Anthropological Aspects; Soviet Studies: Culture; Soviet Studies: Economics; Soviet Studies: Politics

Bibliography

Andrusz G, Harloe M, Szelenyi I (eds.) 1996 *Cities After Socialism: Urban and Regional Change and Conflict in Post-socialist Societies*. Blackwell, Cambridge, UK

Bukowski C G, Racz B (eds.) 1999 *The Return of the Left in Post-communist States: Current Trents and Future Prospects*. Edward Elgar, Cheltenham, UK

Elster J, Offe C, Preuss, Boenker F, Goetting U, Ruel F W 1998 *Institutional Design in Post-communist Societies: Rebuilding the Ship at Sea*. Cambridge University Press, Cambridge, UK

Fowkes B 1999 *The Post-communist Era: Change and Continuity in Eastern Europe*. St Martin's Press, New York

Jawlowska A, Kempny M (eds.) 1994 *Cultural Dilemmas of Post-communist Societies*. IFiS, Warsaw, Poland

Kornai J 1995 *Highway and Byways: Studies on Reform and Post-communist Transition*. MIT Press, Cambridge, MA

Kornai J 1997 *Struggle and Hope: Essays on Stabilization and Reform in a Post-socialist Economy*. Edward Elgar, Cheltenham, UK

Michta A A, Prizel I (eds.) 1992 *Post-communist Eastern Europe: Crisis and Reform*. St Martin's Press, New York

Nagle J D, Mahr A 1999 *Democracy and Democratization: Post-communist Europe in Comparative Perspective*. Sage, London

Opalski M (ed.) 1998 *Managing Diversity in Plural Societies: Minorities, Migration and Nation-building in Post-communist Europe*. Nepean Forum Eastern Europe, Nepean, ON, Canada

Rose R, Mishler W, Haerpfer C 1998 *Democracy and its Alternatives: Understanding Post-communist Societies*. Johns Hopkins University Press, Baltimore, MD

Rupesinghe K, King P, Vorkunova O (eds.) 1992 *Ethnicity and Conflict in a Post-communist World: The Soviet Union, Eastern Europe, and China*. St Martin's Press, New York

Tismaneanu V 1998 *Fantasies of Salvation: Democracy, Nationalism, and the Myth in Post-communist Europe*. Princeton University Press, Princeton, NJ

M. Grekova

Post-traumatic Stress Disorder

1. Introduction

Psychological responses to the experience of traumatic events have long been recognized, although the meaning attached to and the terms used to describe such reactions have varied over the years, and continue to do so, influenced by the prevailing beliefs and interests of the time and depending on the culture studied. 'Post-traumatic stress disorder' (PTSD) is the term now given, at least in Western psychiatry, to a constellation of symptoms which may follow exposure to a traumatic event, and which cause clinically significant distress or impairment in social, occupational, or other important areas of functioning. The word 'trauma' is derived from ancient Greek surgical terminology, meaning an injury stemming from penetration of the body defense (intactness) such as skin. Exposure to major trauma does not cause psychiatric disorder in all those who experience it, possibly due, at least in part, to differences in vulnerability and resilience. The concept of PTSD has been extended from adults to include children, although it is still uncertain to what extent the developmental status of the child influences the symptoms (Scheeringa et al. 1995). The child's perception of its family's reaction to a disaster may also be relevant.

Unfortunately, the recent burgeoning of interest in PTSD has perhaps given some the idea that PTSD is the only response after trauma. This is not the case (see Udwin 1993 for a discussion regarding children). It should be emphasized at the outset that PTSD is only one of a number of possible psychological sequelae of traumatic experiences, and the concept is not without its critics.

2. History

Early descriptions of the reactions of individuals and societies during the pre-Christian period, including the Sumerian, Hebrew, and Ancient Greek eras, are described by Boehnlein and Kinzie (1992). More recent examples which have been described include the response of Samuel Pepys, as described in his diary after the Great Fire of London in 1666. Pepys wrote of his 'dreams of the fire and falling down of houses,' and even six months later described how he was unable to sleep 'without great terrors of fire' (Daly 1983). Charles Dickens, the English novelist, was involved in a railway accident in June 1865, and also helped in rescuing others. He subsequently became anxious about travelling, and wrote, 'I am not quite right within, but believe it to be an effect of the railway shaking' (Trimble 1984). Two years after the accident, he was still experiencing 'sudden vague rushes of terror' (Ackroyd 1999).

The development of the railways stimulated interest in the psychological responses to traumatic events. Doctors at that time noted that there were many victims of railway accidents who appeared to suffer no physical injury and yet who complained of a variety of symptoms. Indeed, the term 'compensation neurosis' was introduced in 1879 following the increase in

invalidism reported after railway accidents. Some argued that post-traumatic symptoms were the result of physical damage to the nervous system following relatively mild trauma. John Eric Erichsen, a professor of surgery at University College Hospital, London, used the term 'railway spine' to describe such injuries. A contrary view was taken by others. Herbert Page, a London surgeon, proposed in 1885 that accident victims were suffering from 'nervous shock,' attributable to a functional (psychological) disturbance rather than physical injury. Jean-Martin Charcot, a Parisian neurologist, supported the psychological model, arguing that the syndrome of 'railway spine' was a form of hysteria. He believed that, in the immediate aftermath of an accident, the victims were in a suggestible state and that their own idea of physical injuries influenced the development of symptoms. Charcot, in turn, influenced Sigmund Freud, who later suggested that the physical symptoms of hysteria were due to the repressed memories of traumatic events. The development (in many countries) of workmen's compensation acts, which provided for financial compensation to those injured during the course of work, also led to increased interest in the psychological effects of traumatic experiences (Trimble 1984).

A major influence on the development of the concept of PTSD has been war. The belief that the psychological responses to traumatic experiences were the result of physical injury or dysfunction of the central nervous system was strongly held and represented in the concept of 'shell-shock' which prevailed during the First World War in Europe. During the First World War, the British Army reported that seven to ten percent of the officers and three to four percent of the other ranks suffered 'mental breakdowns' (Gersons and Carlier 1992). Some soldiers were treated as cowards and deserters, and punished. The number of death sentences passed was 3,500, of which approximately 282 were executed prior to 1917. However, many more were considered ill and in need of treatment (Deahl 1997). Some, such as the Scottish doctor, William Rivers, saw the problem as a psychological one, and tried psychological techniques, such as abreaction and catharsis, as well as simply letting the person talk about their traumas. Rivers held that repression explained how an individual suppressed painful and terrifying experiences, and that these later became converted into physical and mental symptoms. World War II also produced large numbers of psychiatric casualties, despite more stringent recruit selection. Mental disorder accounted for 31 percent of medical discharges from the British Army.

After wars, increased rates of mental illness have been reported in service veterans who have seen combat. For example, between 1964 and 1975, 3.14 million US servicemen and women served in the war in Vietnam. Approximately 58,000 died and 300,000 were wounded. Of combatants, 20 percent witnessed atrocities or abusive violence. It has since been estimated that, of all US Vietnam veterans, approximately 15 percent (450,000) currently suffer from PTSD, whilst an additional 11.1 percent of male and 7.8 percent of female veterans suffer at least some symptoms of traumatic stress related to the experience in Vietnam (see Deahl 1997 for a review).

Psychological sequelae in both adults and children, including PTSD, have now been reported after a wide variety of traumatic events, such as a fire in a nightclub in Boston (Lindemann 1944), and, more recently, after a variety of both human-made and natural disasters, kidnapping, torture, and road traffic accidents, as well as in rape victims, after abuse, in refugees, after domestic violence, and even after childbirth. Epidemiological studies suggest that PTSD or subthreshold PTSD may be present in a significant proportion of a population, although it may not be recognized as such. Comorbidity may be present.

3. *Classification*

Attempts have been made to categorize responses to traumatic experiences. Publication of the first edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-I*) (American Psychiatric Association 1952) coincided with the Korean war, and included the categories of 'gross stress reactions' and 'anxiety neurosis/transient situational disturbance' with reference to 'response to overwhelming environmental stress.' The second edition (*DSM-II*) (American Psychiatric Association 1968), which developed during a period of relative global peace, eliminated the diagnostic category of gross stress reaction.

In 1980, diagnostic criteria for PTSD were described in the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III*) (American Psychiatric Association 1980). These criteria were based on information obtained from Vietnam veterans, and probably also reflected the experience of Holocaust concentration-camp survivors from World War II. The diagnostic criteria were modified in the revised edition (*DSM-III-R*) published in 1987. There was increasing emphasis on avoidance phenomena. It was acknowledged that children may be affected, that 'survivor guilt' was not a primary feature, and that symbolic reminders of the trauma may be important.

Meanwhile, in Europe, the ninth revision of the *International Classification of Diseases* (*ICD-9*) (World Health Organization [WHO] 1978) incorporated the subcategories of 'acute reaction to stress' and 'adjustment reaction' following exposure to acute stress. Acute reaction to stress described transient responses to exceptional physical or mental stress, lasting only a few hours to a few days. Adjustment reactions were longer lasting, but usually lasted only a few months.

The diagnostic criteria are presently defined in the *Diagnostic and Statistical Manual of Mental Disorders*

(*DSM-IV*), published by the American Psychiatric Association in 1994, and *The ICD-10 Classification of Mental and Behavioural Disorders*, published by the World Health Organization in 1992. There is also the publication *The ICD-10 Diagnostic Criteria for Research* (World Health Organization 1993). According to *ICD-10*, PTSD is a delayed and/or protracted response to a stressful event or situation (either short- or long-lasting) of an exceptionally threatening or catastrophic nature (including serious accident). Typical symptoms include episodes of repeated reliving of the trauma in intrusive memories or dreams, a sense of numbness and emotional blunting, with feelings of detachment from other people, inability to enjoy activities, and avoidance of activities and situations reminiscent of the trauma. Other symptoms, such as hyperarousal of the autonomic nervous system, hypervigilance, an enhanced startle reaction, difficulty sleeping, anxiety, and depression may also be present. The *ICD-10 Diagnostic Criteria for Research* are similar, but specify that there must be persistent remembering or 'reliving' of the stressor in intrusive 'flashbacks,' vivid memories, or recurring dreams, or in experiencing distress when exposed to circumstances resembling or associated with the stressor. In addition, for the diagnosis of PTSD to be made using these criteria, there should be an actual or preferred avoidance of circumstances which resemble or are associated with the stressor, which was not present before exposure to the stressor, and either an inability to recall, either partially or completely, some important aspects of the period of exposure to the stressor, or persistent symptoms of increased psychological sensitivity and arousal (not present before exposure to the stressor), as shown by any two of the following—difficulty sleeping, difficulty concentrating, irritability or outbursts of anger, hypervigilance, or an exaggerated startle reaction. These symptoms should have started within six months of the stressor.

According to *DSM-IV*, PTSD occurs after a traumatic event in which the person experienced, witnessed, or was confronted with an event or events that involved actual or threatened death or serious injury, or a threat to the physical integrity of self or others. The person's response should have involved intense fear, helplessness, or horror. For the diagnosis of PTSD, the individual should be persistently re-experiencing the traumatic event, such as experiencing recurrent and intrusive distressing recollections of the event (which, in children, may be manifest in repetitive play in which themes or aspects of the trauma are expressed), having recurrent distressing dreams of the event (in children, the content of the dreams may be less recognizable), acting or feeling as if the traumatic event were recurring, experiencing intense psychological distress at exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event, or experiencing physiological reactivity on exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event. There should be evidence of persistent avoidance of stimuli associated with the trauma and numbing of general responsiveness (not present before the trauma), as indicated by three (or more) of the following: efforts to avoid thoughts, feelings, or conversations associated with the trauma, efforts to avoid activities, places, or people that arouse recollections of the trauma, an inability to recall an important aspect of the trauma, markedly diminished interest or participation in significant activities, a feeling of detachment or estrangement from others, a restricted range of emotions, a sense of a foreshortened future. There should also be at least two persistent symptoms of increased arousal (not present before the trauma), such as difficulty falling or staying asleep, irritability or outbursts of anger, difficulty concentrating, hypervigilance, an exaggerated startle response. These symptoms should have lasted more than one month and be causing clinically significant distress or impairment in social, occupational, or other important areas of functioning. 'Delayed-onset' PTSD may be seen where a person initially appears to have dealt with a traumatic experience but, perhaps years later, symptoms of PTSD develop, possibly after a reminder of the original trauma or after a relatively trivial incident.

While diagnostic classification systems are useful, they are not a checklist, and clinical judgment is considered essential in their use. Criteria are not 'written in stone' but evolve in the light of clinical experience and information from research. For example, there is ongoing discussion about the validity of the 'stressor criterion,' since PTSD may occur after seemingly minor accidents. It seems likely that, in the future, there will be attempts to eliminate as many differences as possible between the ICD and DSM systems of classification (see *Nosology in Psychiatry*).

4. *Theoretical Models*

The concept of PTSD has been accompanied by a variety of theoretical models (neuropsychiatric, psychoanalytical, cognitive-behavioral) to explain its origins and maintenance in the traumatized individual. For example, a number of psychobiological findings in PTSD, such as altered autonomic function and alterations in the hypothalamic–pituitary–adrenal axis, and studies of endogenous opioids (such as encephalins) may be relevant (see *Endocrinology and Psychiatry*). Psychoanalytic models include those which regard the re-experiencing phenomena as primary, leading to defensive functioning, and those which regard the reaction to the stressor as a failure of adaptation. Common to all dynamic models is a concern with the meaning a person gives to a traumatic event and/or its aftermath. Behavioral models of PTSD have used the principles of conditioning theory to explain the development and maintenance of PTSD,

while cognitive models of PTSD have included the concepts of learned helplessness, the attributional style of individuals, cognitive appraisal, and, more recently cognitive-processing theory. However, no single model offers a comprehensive view of PTSD. It has also been suggested that an understanding of memory may assist in the understanding of the phenomenology of responses seen after traumatic events. It is to be hoped that a clearer understanding of the origins of PTSD will help in the development of effective therapeutic techniques.

5. *Assessment*

There are a number of questionnaires or scales which may be used to help assess the psychological responses of traumatic events. Davidson et al. (1997) describe a 17-item self-rating scale for PTSD, which measures each *DSM-IV* symptom on five-point frequency and severity scales. Costello et al. (1998) have described a new measure of life events and PTSD for children and adolescents aged nine to seventeen years, for use in both epidemiological and clinical studies. One of the most widely used questionnaires in either research into, or clinical work on, the psychological sequelae of traumatic events, is the Horowitz Impact of Event Scale (Horowitz et al. 1979). This is a 15-item self-rating scale with two subscales: intrusion and avoidance. It is normally used to assess the response after a single traumatic event, rather than after exposure to a chronic stressful situation. It gives a subjective estimate of the frequency of intrusive recall of the traumatic event and of attempts to avoid such recall, and thus a measure of psychological disturbance when compared to normative data. It does not include questions about physiological reactivity and cannot, therefore, be used to make any diagnosis. It was not originally designed for use with children but has been shown to be useful in those aged eight years and over. A score of 20 or above is often considered to be a cause for concern for civilians in peacetime, although others have used a score of 30 or above to indicate significant post-traumatic stress.

6. *Treatment*

A variety of approaches are presently used, both in adults and in children. In adults, behavioral and cognitive approaches have been increasingly used over the past 15 years or so, usually involving exposure (either real-life or imagined), cognitive therapy, or anxiety management training. Clinically, exposure and cognitive restructuring are often included in an anxiety management program. Psychotherapy may be individual, group, or with a family. A further technique, 'eye movement desensitization and reprocessing' (EMDR), may be a useful therapeutic technique, but its value and the mechanism of any effects that it has require further elucidation. There is little evidence that drug treatments have a central role in the treatment of PTSD.

The emphasis when treating children has been on cognitive behavioral approaches that aim to help survivors of a traumatic event make sense of what has happened and to master their feelings of anxiety and helplessness. Sometimes support may be appropriately offered in groups. Since children live with adults, whose reactions to events are part of the context of that event for the child, most treatment plans for children should involve both families and schools. This may be especially relevant when the adults have themselves experienced the traumatic event. Distress in the adults may mean that they are less able to recognize distress in the children for whom they have responsibility, or that they are relatively unavailable to respond to the children's needs (see *Child and Adolescent Psychiatry, Principles of*).

Further research is needed to evaluate the effectiveness of the differing therapies and also to determine the effectiveness of screening populations at high risk of PTSD (for example, after a major disaster) to find out whether any interventions may prevent the development of PTSD or other psychiatric morbidity in those identified as being especially at risk. However, there is no clear evidence that offering counseling routinely to all survivors of traumatic events is helpful—indeed, there is some suggestion that it may be harmful (Wessely et al. 2000).

7. *Areas of Controversy*

PTSD is an evolving concept. There has been continuing debate on whether individual, as opposed to stressor characteristics, are decisive in the development of PTSD. It may be inappropriate to use a single medical concept to describe the responses seen after different traumatic experiences. Distress after the experience of traumatic events has been variously interpreted in the past. The current emphasis on the 'medicalization of distress' in Western cultures may, perhaps, reflect the prevailing expectations of that society and culture. It is still unclear whether it is appropriate to use a concept which has developed in Western psychiatry to other cultures, and how appropriate it is to extend the concept to describe the responses seen after a wide range of traumatic events. The present classifications assume a dichotomous rule—disorder is either present or absent, with no opportunity to consider severity.

An important consequence of the diagnostic category of PTSD is that it has facilitated, and perhaps encouraged, claims for compensation after traumatic events. Clinicians undertaking forensic assessments are advised that diagnostic criteria are clinical guidelines, and that there may be an imperfect fit between questions of ultimate concern to the law and information contained in a clinical diagnosis.

8. Conclusions

The concept of PTSD is an evolving one, of uncertain value, and it seems likely that there will be further developments and refinements in our understanding of individual responses to traumatic events, and how best to describe and classify them. The importance of developmental influences, the significance of symptoms in differing cultures, and the risk and protective mechanisms involved require further attention.

See also: Mental Illness, Etiology of

Bibliography

Ackroyd P 1999 *Dickens*. Vintage, London

American Psychiatric Association 1952 *Diagnostic and Statistical Manual of Mental Disorders*, 1st edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1968 *Diagnostic and Statistical Manual of Mental Disorders*, 2nd edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1987 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd rev. edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association, Washington, DC

Boehnlein J K, Kinzie J D 1992 Commentary. *DSM* diagnosis of post-traumatic disorder and cultural sensitivity: A response. *Journal of Nervous and Mental Disease* **180**: 597–9

Costello J, Angold A, March J, Fairbank J 1998 Life events and post-traumatic stress: The development of a new measure for children and adolescents. *Psychological Medicine* **28**: 1275–88

Daly R J 1983 Samuel Pepys and post-traumatic stress disorder. *British Journal of Psychiatry* **143**: 64–8

Davidson J R T, Book S W, Colket J T, Tupler L A, Roth S, David D, Hertzberg M, Mellman T, Beckham J C, Smith R D, Davison R M, Katz R, Feldman M E 1997 Assessment of a new self-rating scale for post-traumatic stress disorder. *Psychological Medicine* **27**: 153–60

Deahl M 1997 Effect of conflict on combatants. In: Black D, Newman M, Harris-Hendriks J, Mezey G (eds.) *Psychological Trauma: A Developmental Approach*. Gaskell, London, pp. 134–47

Gersons B P R, Carlier I E 1992 Post-traumatic stress disorder: The history of a recent concept. *British Journal of Psychiatry* **161**: 742–8

Horowitz M, Wilner N, Alvarez W 1979 Impact of Event Scale: A measure of subjective stress. *Psychosomatic Medicine* **41**: 209–18

Lindemann E 1944 Symptomatology and management of acute grief. *American Journal of Psychiatry* **101**: 141–8

Scheeringa M S, Zeanah C H, Drell M J, Larrieu J A 1995 Two approaches to the diagnosis of post-traumatic stress disorder in infancy and early childhood. *Journal of the American Academy of Child and Adolescent Psychiatry* **34**: 191–200

Trimble M R 1984 Post-traumatic stress disorder: The history of a concept. *Psychiatry in Practice* **May**: 5–7

Udwin O 1993 Annotation—Children's reactions to traumatic events. *Journal of Child Psychology and Psychiatry* **34**: 115–27

Wessely S, Rose S, Bisson J 2000 Brief psychological interventions ('debriefing') for treating immediate trauma-related symptoms and preventing post-traumatic stress disorder. *Cochrane Database of Systematic Reviews* **2**: CD000560 [Medline]

World Health Organization 1978 *Mental Disorders: Glossary and Guide to the Classification in Accordance with the Ninth Revision of the International Classification of Diseases*. World Health Organization, Geneva, Switzerland

World Health Organization 1992 *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. World Health Organization, Geneva, Switzerland

World Health Organization 1993 *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization, Geneva, Switzerland

M. Newman

Potlatch in Anthropology

'Potlatch' refers to a ceremonial exchange complex of aboriginal peoples of the Northwest Coast of North America. Originating in precontact practices, the potlatch evolved in the nineteenth century into a quasicommercial system of acquiring prestige through large gifts—and occasionally destruction—of native and trade goods. Among the Kwakiutl of Vancouver Island and neighboring groups, the potlatch came to include a supplemental economy of loans and interest, allowing for prestations so large that they beggared the host. This extreme form, appearing between roughly 1870 and 1930, earned it a central place in the ethnographic imagination.

1. History and Origins

1.1 Precontact Trade and Ceremonial Exchange

'Potlatch' derives from the Chinook jargon word *patshatl*, to give. Chinook jargon was a trade language centered in the Columbia River basin, but extending north to the central Northwest Coast. It predates European contact, and served as a *lingua franca* among groups engaged in aboriginal trade. Archeological evidence suggests the existence of intergroup ceremonial exchange alongside precontact trading relations. However, it is clear that such exchange was both quantitatively and qualitatively different from the historic potlatch. It involved modest amounts of animal pelts and a few types of prestige goods, and was known under different labels in different areas. It reaffirmed but likely did not alter rank relations within and between groups. Thus, it is fair to say that the potlatch, as a translocal institution, was a product of the fur-trade era.

1.2 The Fur Trade: Influx of Wealth

The early fur trade of the 1770s and 1780s involved Russians, Spanish, British, and Americans. Trade focused on the sea otter and was conducted primarily from ships. This had a limited effect on aboriginal groups. However, by the late 1780s the sheer number of trading vessels, and the shift to land-based trading, had begun to bring changes. The transfer of the Spanish colony on Juan de Fuca Strait to the British in 1793 marked the beginning of a process that would see Hudson's Bay Company forts established all along the coast, and eventually a substantial colony at Victoria. To the north, the Russians established a fort at Sitka, in Tlingit territory, and to the south, Americans established posts on the Columbia in the 1810s. The shift to land-based trade brought with it an influx of wealth and trade goods, as well as epidemic disease, and a broad range of new cultural influences.

Despite the negative effects of contact, the early nineteenth century was characterized by an efflorescence of art and expressive culture on the Northwest Coast. Nowhere is this seen more clearly than in the plastic arts. The introduction of metal woodworking tools and, somewhat later, commercial paint, allowed for an explosion of art in the form of masks, architectural elements, storage vessels, and free-standing monumental sculpture. Such objects were commissioned with new commercial wealth; these objects, in turn, were vehicles by which this wealth was displayed and validated. Wealth had always been closely associated with rank. The problem arose in that rank was ideologically a closed system: a limited number of ranked positions, associated with titles, privileges, and symbolic property ('crests') were believed to exist, and these positions were filled in accordance with regular principles of descent and succession. However, during the fur trade period *nouveaux riches* rose to prominence. New crests were created for them by artists who were experimenting with both themes and techniques. Simultaneously, such persons also staked claim to unused traditional crests or those in dispute. The potlatch was the forum in which such claims were made public, and were either validated or rejected by the larger community.

1.3 The Dominion Period: Response to Traumatic Change

After the 1860s, which saw both Canadian Dominion and the US purchase of Alaska, aboriginal groups were increasingly encapsulated within Euroamerican political, social, and economic institutions. Simultaneously, the full effects of disease mortality were being felt, producing upheavals in aboriginal society. The potlatch, an effective means of dealing with demographic and social change, was mobilized to address issues such as the amalgamation of groups, the lack of personnel to fill important rank positions, changing relations with Euroamericans, and changes in the relative status of local groups. In the process, potlatches became larger and more expansive. A wide range of aboriginal practices, such as shamanic winter dances, became incorporated within the potlatch. Simultaneously, the potlatch replaced intergroup violence, which was suppressed, as a means for gaining prestige at neighbors' expense.

During this period, the potlatch was suppressed by Canadian authorities, who passed a 'Potlatch Law' in 1885. Although unenforceable for most of its existence, it did result in famous prosecutions and convictions of Kwakiutl potlatchers in 1923. In addition to criminal punishment, the state seized 'potlatch goods,' including ritual objects. These provided the basis of ethnological collections at museums in Ottawa and Toronto. In the 1970s and 1980s, these items were returned to Kwakiutl communities at Cape Mudge and Alert Bay, which have built museums to house them.

2. Geographical Range and Variation

2.1 The Northern Potlatch

The northern region, consisting of matrilineal societies, used the potlatch primarily in the context of funerals and succession. For the Tlingit and Haida, potlatches were between moieties, and stressed cooperation and solidarity between hosts and guests. A minor theme of competition was introduced, with two separate guest groups attempting to outdo each other in the degree of 'comfort' provided to the host. Tsimshian potlatches were similar, but, lacking moieties, were somewhat more agonistic. They were associated with marriages, as well as other life-crisis rituals; marriage was seen in oppositional terms. In addition, potlatches were used to make territorial claims.

2.2 The Southern Potlatch

South of the Tsimshian, social organization was cognatic, with a matrilineal orientation in the north becoming more patrilineal in the south. Among the Nootka, potlatches were essentially private entertainments sponsored by a chief for his people, and thus stressed social harmony. Elsewhere, however, potlatches were primarily agonistic. Marriage potlatches, including those held to repay bridewealth, were couched in military terms. As group membership and rank were less fixed in cognatic societies, they were more the object of political manipulation, and thus of representation in the potlatch.

2.3 The Kwakiutl Potlatch

The extreme example of agonistic prestation was provided by the Kwakiutl, who live in northeastern

Vancouver Island and the adjacent islands and mainland. During the 60 years after Dominion, the potlatch intensified to the point that Benedict (1934), in her flawed analysis of Boas' material (Boas 1966), described the Kwakiutl as 'megalomaniacal' and 'paranoid.' Fierce competition among brothers, cousins, and affines for succession to titles, the existence of numerous latent titles, the rise of a *nouveaux riche* class called 'Eagles,' and a system of intergroup ranking, all contributed to the intensity of the potlatch, which participants viewed as warfare by other means. Potlatches opened with guests arriving in war regalia, and a mock battle on the beach, reminiscent of the murderous raids that were endemic to this area.

Kwakiutl potlatches were predicated on rivalries between chiefs and, by extension, their groups. In addition to the usual means of amassing wealth, would-be hosts engaged in a system of loans with 100 percent interest, which allowed lenders to build fortunes quickly, and borrowers to obtain last-minute infusions. The currency was Hudson's Bay Company blankets, supplemented and eventually replaced by cash. Copper shields were the ultimate valuable; bid up in previous transactions, 'coppers' reached cash values of several thousand dollars. In addition, trade goods played an increasing role. By the start of the twentieth century, manufactured items such as shotguns, washing machines, and pool tables were given away, along with quantities of flour, rice, cloth, and utensils. In some of the most extreme potlatches, goods including coppers were destroyed.

3. *Functional Accounts*

3.1 *The Potlatch as Total Social Phenomenon*

The French sociologist Marcel Mauss (1990) noted that the potlatch is an integral element of the rank system of Northwest Coast societies. It is a mechanism by which claims to rank status are validated or challenged, and changes in social order are communicated to the larger community. The potlatch rests on opposition between host and guest. The host—a chief or high-ranking person and, by extension, his social group—was concerned with conducting certain types of business: naming heirs, contracting marriages, conducting funerals, proclaiming new titles and privileges. This business was witnessed, in a juridical sense, by the guests, who were paid for their service. Potlatch gifts went beyond mere payment for services, however, since they were also taken as *prima facie* evidence for the validity of the claim.

Wealth was a necessary concomitant of high rank. High-rank positions implied a certain degree of extraordinary power, similar to the *mana* of Polynesian chiefs, which necessarily resulted in wealth. Inability to provide an abundant potlatch would seriously undermine the legitimacy of the claimant's assertions. By the same token, an extravagant potlatch could 'fasten on' a title, merging the holder's social identity with the permanent qualities of the title. Such acts were seen as a sort of reincarnation, by which the living title holder became incorporated in the title, which encompassed the holder's predecessors and mythical founder. For chiefly positions, the ascendance of a new holder was often marked by events symbolic of the (re)creation of the social world, including human sacrifice, the razing of old chiefly houses and poles, and their replacement by new ones.

In chiefly succession potlatches, a broad range of political, legal, and economic issues were implicated. In all groups, chiefly titles involved territorial claims of some sort. These titles were based on mythical narratives, set locally, and taking place at the beginning of human history, thus staking aboriginal claim to land. In some groups, such as the Nootka, such territorial claims were very specific and resembled Western notions of ownership. In all Northwest Coast cultures, chiefs were minimally seen as stewards of the land and its inhabitants, and were responsible for regulating relations between human and nonhuman populations (which were believed to include more than the flora and fauna recognized by biological science). This was accomplished by ritual means, including prayers, attraction magic, and first fruit rituals. The latter were used to regulate the seasons on fish and game. Although the informing ideology was moral, not ecological, such regulation had the effect of conserving resources.

The potlatch thus involved relations between the social group and the nonhuman outside world, the prehuman myth world, and the group's ancestors. In this way, the group was situated within a larger universe, and was thus defined and delimited. Potlatches likewise served to delineate the internal organization of society, and the functional relations obtaining between its components. In a general sense, one potlatches to one's affines, who constitute the friendly other. In northern matrilineal groups, this opposition was absolutely clear. Especially in the moiety system of the Tlingit and Haida, one's affines were a separate corporate group; the cooperation of the two sides presented an image of social totality cosmological in scope.

3.2 *The Potlatch and Marriage Exchange*

Potlatches were associated with marriage in most areas; even where they were not, the potlatch was transacted between affines. Politically, potlatches were vital in renewing alliances among groups. Among the northern groups, whose aristocrats practiced preferential cross-cousin marriage, long-term stable exchange relations involving two or more local groups were established. Such groups provided political, military, and economic support, as well as external validation of chiefly claims. Among cognatic groups, no such long-term structural relations existed, but

were rather produced and reproduced in a series of exchanges tied to specific marriages (payment and repayment of bridewealth, naming of children, funeral of a spouse). The connection between marriage and potlatching was so strong that in groups such as the Kwakiutl, Bella Coola, and Heiltsuk, aristocrats would engage in serial monogamy and even fictive marriage in order to obtain prestige from potlatching. Potlatching varied along several dimensions (size, number, occasion, agonism, and others) depending on the precise configuration of kinship, descent, and marriage rules.

3.3 The Potlatch and Ecological Variation

Since the 1960s, materialist scholars have viewed the potlatch as a virtual homeostatic mechanism, distributing personnel and food efficiently through space and time. The earlier conception of the Northwest Coast as free from food pressures is justly rejected. The potlatch is seen as having four distinct, but interrelated functions: (a) spurring overproduction, (b) encouraging storage of food, (c) allowing for the exchange of food for prestige goods, and (d) regulating group membership and territorial boundaries, matching resources with personnel. This approach assumes local variation in resources, a condition that does not exist everywhere, but is more pronounced in the north. Item (d) was effected much more often by warfare than by potlatching. However, these remain at least potential functions of the potlatch.

3.4 Rank, Class, and Political Power

Northwest Coast society was stratified by both rank and class. Potlatching was the domain of chiefs and nobles; however, both commoners and slaves played important roles in amassing resources. By definition, only persons with titles were in a position to potlatch. Potlatching thus reinforced vertical, as well as horizontal, social divisions. Simultaneously, potlatching was the mechanism by which social mobility was integrated into the social system.

The potlatch both manifested and represented political power. By drawing together a large number of guests, especially if they were from far away or their chiefs were especially high ranking, a host was demonstrating his prestige and political network. The material goods and food distributed were an index of his land holdings and their productivity, which ultimately depended upon the chief's legitimacy. Potlatches were fora for authoritative political discourse, defined technically as monologic, multichannel, redundant communication. A significant aspect of this communication was an illocutionary speech act. This was supplemented with nonverbal media, including plastic arts and dance. The political success of a potlatch was usually measured in terms of the degree to which these utterances remained unchallenged. They may be challenged either directly, in rivalry gestures, or more subtly through various sorts of corrosive discourse, including humor. However, in some groups, such as the Kwakiutl, such elements were institutionalized within the structure of the potlatch. This had the effect of heightening tensions within the potlatch, and giving it a more game-like structure, in which a wide range of outcomes was possible.

4. Problems of Reference

4.1 Reification

Given the degree of variation over space and time, it is unclear that the potlatch, *as such*, exists. Rather, our understanding of it is conditioned by reifications imposed by government agents and missionaries, anthropologists, and aboriginal people themselves. The Potlatch Law of 1885 first tried to specify 'the Indian festival known as the "Potlatch".' Problems of reference significantly delayed successful prosecution of the law. During this intervening time, the symbolic labor of many people served to bring descriptions and practice closer together.

4.2 Invention of Tradition

Aboriginal people themselves reified the potlatch by defining their ceremonial praxis as potlatch, a foreign word and concept mediated through Euroamerican society. By the time Franz Boas recorded ethnographic data among the Kwakiutl in the 1890s (Boas 1966), the concept of potlatch was present. This consciousness of 'the potlatch,' on the part of both Boas and the Kwakiutl, was shaped largely by the existence of the Potlatch Law. Clearly, aboriginal practice and self-awareness of practice were conditioned by exogenous expectations and categories. It is not surprising that during a period of rapid change such an invention of tradition would occur. It allowed aboriginals to retain symbolic control of their destiny, by redefining novel structures and practices as traditional. Because of this useful quality, potlatching remains a common practice today.

4.3 The Potlatch as Anthropological Totem

The potlatch remains one of the most durable anthropological concepts, despite the reanalyses and deconstructions to which it has been subjected. Beginning with Boas, anthropologists defined potlatch as both self and other, transparent and exotic. This contradictory and mediational quality has made potlatch an especially useful metaphorical vehicle. Potlatch has been viewed as, *inter alia*, metaphorical banking, warfare, and mental illness. At the same time, various aspects of Western societies, including

politics and consumer capitalism, have been compared to potlatch. Its hybrid nature, and its origins in the desire on the part of Euroamericans to describe aboriginals, and aboriginals to represent change and continuity, make it an attractive concept, but one that resists critical usage.

See also: Boas, Franz (1858–1942); Colonialism, Anthropology of; Elites, Anthropology of; Exchange in Anthropology; Exchange: Social; Mauss, Marcel (1872–1950); Political Anthropology; Totemism

Bibliography

Adams J W 1973 *Gitksan Potlatch: Population Flux, Resource Ownership, and Reciprocity*. Holt, Rhinehart, and Winston, Toronto, Canada

Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston

Boas F 1966 *Kwakiutl Ethnography*. University of Chicago Press, Chicago

Bracken C 1997 *The Potlatch Papers: a Colonial Case History*. University of Chicago Press, Chicago

Codere H 1950 *Fighting with Property: a Study of Kwakiutl Potlatching and Warfare, 1792–1930*. Augustin, New York

Cole D, Chaikin I 1990 *An Iron Hand Upon the People: the Law Against the Potlatch on the Northwest Coast*. University of Washington Press, Seattle, WA

Drucker P, Heizer R F 1967 *To Make My Name Good: a Reexamination of the Southern Kwakiutl Potlatch*. University of California Press, Berkeley, CA

Harkin M E 1997 *The Heiltsuks: Dialogues of Culture and History on the Northwest Coast*. University of Nebraska Press, Lincoln, NE

Kan S 1989 *Symbolic Immortality: the Tlingit Potlatch of the Nineteenth Century*. Smithsonian Institution, Washington, DC

Lévi-Strauss C 1982 *The Way of the Masks*. University of Washington Press, Seattle, WA

Mauss M 1990 *The Gift: the Form and Reason for Exchange in Primitive Societies*. Hall W D (trans.) Routledge, London

Rosman A, Rubel P G 1971 *Feasting with Mine Enemy: Rank and Exchange Among Northwest Coast Societies*. Columbia University Press, New York

Schulte-Tenckhoff I 1986 Potlatch: *Conquête et Invention: Réflexion sur un Concept Anthropologique*. Lausanne: Éditions d'en Bas

Seguin M 1985 *Interpretive Contexts for Traditional and Current Coast Tsimshian Feasts*. National Museums of Canada, Ottawa, Canada

Walens S 1981 *Feasting with Cannibals: Metaphor and Morality in Nineteenth Century Kwakiutl Culture*. Princeton University Press, Princeton, NJ

M. E. Harkin

Poverty and Child Development

Income poverty is when a family's income fails to meet a federally established threshold; this threshold differs across countries. In general, living at or below a poverty threshold signifies that a family lacks basic financial resources, and thus children may lack access to necessary resources including food, shelter, and health care. Historically, the measurement of poverty as well as research linking poverty to child development has been an Anglo-Saxon endeavor (Smeeding et al. 2000). However, international organizations, such as the United Nations Children's Fund (UNICEF), have fostered global concern. A growing body of research suggests that poverty has consequences for children's well-being. Internationally, the prevalence of poverty among families with children has profound implications for researchers, policy makers, and practitioners concerned with children's development.

This article, divided into three sections, reviews the literature on the role of income poverty in the lives of children and families. The first section considers different measures of poverty as well as international trends in child poverty. Second, a review of recent studies examining the association between poverty and children's physical, cognitive, and emotional development is provided. The last section reviews the potential pathways through which poverty may influence children's health and well-being. The conclusion briefly highlights broad policy implications.

1. Measuring Poverty and International Trends

1.1 Income and Poverty

In most countries, income poverty typically is defined by official poverty thresholds. Those families whose income falls above the poverty line are considered 'not poor,' whereas, families whose incomes are at or below the threshold are considered 'poor.' Two general types of thresholds exist, absolute and relative. *Absolute poverty thresholds* are based on an absolute level of resources; as such, a family's poverty status is not influenced by other families' incomes. For instance, the USA has an absolute poverty threshold based on anticipated food expenditures (thrifty food basket), taking into account household size, and multiplied by three; the thresholds are adjusted annually for changes in cost of living. In contrast, Canada and most Western European nations employ relative poverty thresholds. A *relative poverty threshold* is calculated by comparing a family's resources to that of the average family. These thresholds are, generally, defined by half of the median or mean income of all households in the respective country. Thus, other families' incomes can change a family's poverty status, even if their income is relatively stable over time.

1.2 International Trends in Child Poverty

Using the poverty measures described, international comparisons of child poverty rates are considered. Comparisons of the proportion of children living in

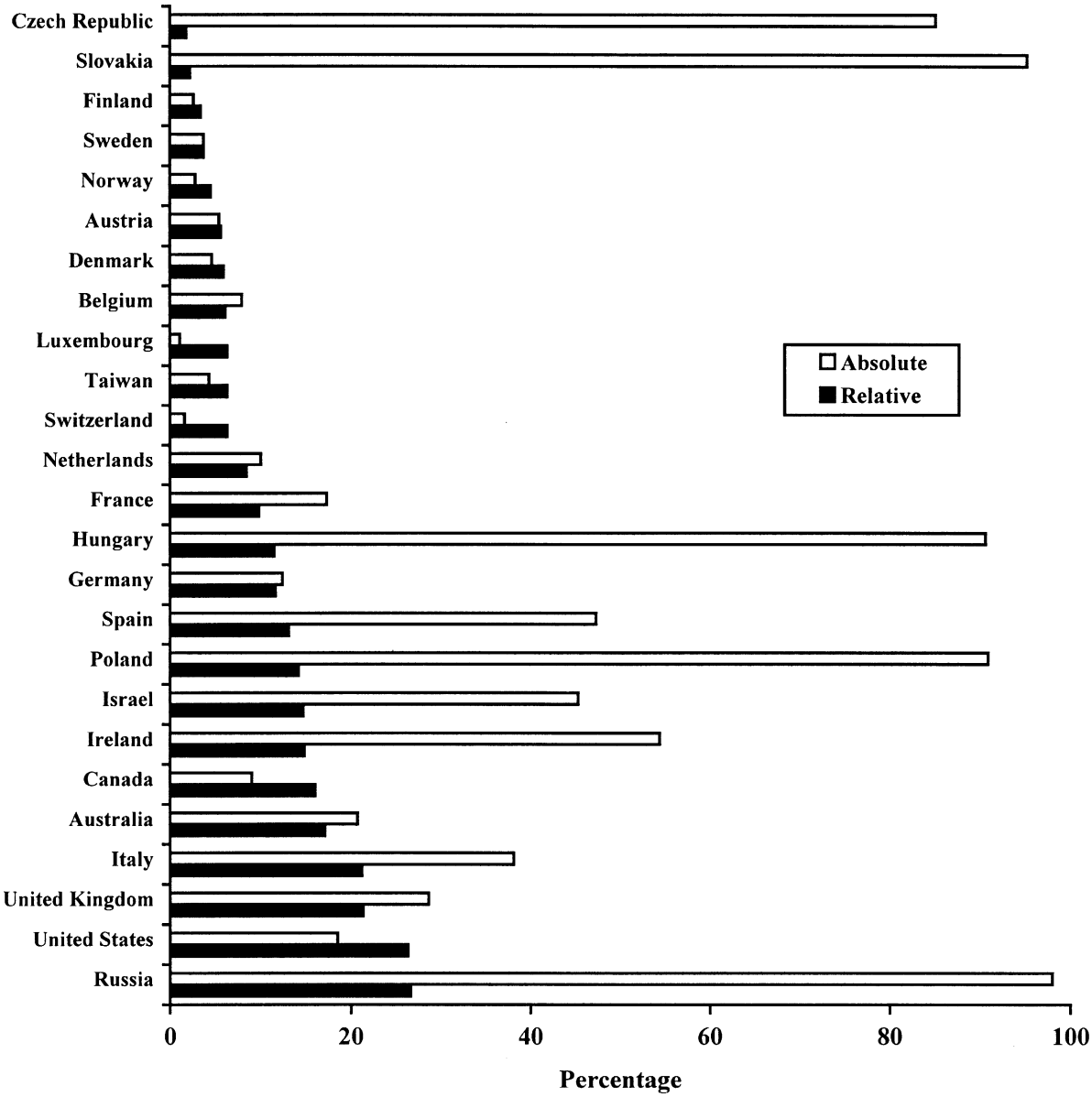

Figure 1[i]
Percentage of children living in poverty in industrialized nations using relative and absolute measures[ii]
[i]Adapted from: Bradbury and Jantti 1999
[ii]Relative poverty rate is 50 percent of overall national median income for survey year. Absolute rate is $15,299, based on 1995 US official poverty line for a two-parent family with two children

poverty in industrialized countries vary depending upon the type of poverty measure utilized (see Fig. 1, Bradbury and Jantti 1999). Using a relative measure, where families living at or below 50 percent of the national median income are considered to be poor, the USA ranks 24th out of 25 in terms of low child poverty rates, despite general economic prosperity. With the exception of the USA and UK, countries with higher levels of national income tend to have lower relative child poverty rates than countries in economic transition (Bradbury and Jantti 1999, Smeeding 1997). However, if an absolute measure of poverty is applied

using the 1995 US criteria, this pattern shifts. Although the child poverty rate in the USA declines only slightly, the country's ranking jumps to 12th place, while Russia and Hungry, countries in economic transition, have child poverty rates of 90 percent or higher. In contrast, child poverty rates remain low in the Nordic countries and Germany and are not appreciably altered by changing poverty measures (Bradbury and Jantti 1999).

The overall level of economic development of a country is not the only factor influencing child poverty trends, as over time, countries with similar levels of development fared differently. Diverse demographic differences between countries, such as the prevalence of single-parent versus two-parent families, account for some of the discrepancies, as well as different roles assumed by governments for fighting child poverty. International poverty rates also are affected by the inclusion or exclusion of social transfers including unemployment benefits, illness or disability assistance, income assistance, and housing benefits (Eurostat 1999). Within the European Union (EU) countries, the USA, and Canada, poverty rates drop from approximately 20 percent to 25 percent when transfers are included as part of a family's annual income (Eurostat 1999, Smeeding 1997). Countries generally considered 'welfare leaders,' namely the Nordic countries, have, on average, lower poverty rates than other countries with fewer social programs.

2. *Poverty and Child Health and Well-being*

This section provides an overview of the effects of poverty on child health and well-being. The findings are discussed for five developmental age groups: prenatal to two years, early childhood (3 to 6 years), late childhood (7 to 10 years), early adolescence (11 to 15 years), and late adolescence (16 to 19 years). Each age group incorporates at least one significant transition in a child's life, such as school entrance or exit or biological maturation. The developmental challenges confronted during each period are relatively universal and may entail new modes of adaptation to biological, psychological, or social changes. The domains of well-being addressed include: physical health; school readiness and achievement; and behavior, social, and emotional problems. Each age period is associated with somewhat different indicators.

The review presented draws heavily from several recent endeavors to investigate the influence of poverty on child development. The first initiative is a cross-national comparison of key child indicators reported by UNICEF (2000), which provides a macrolevel view of how countries' overall levels of economic development affect child well-being. The second effort is a comprehensive evaluation of research on poverty and children's physical, emotional, and cognitive development. This analysis by Brooks-Gunn and Duncan (1997) provides a 'risk assessment' of the effects of poverty on child health and well-being at a single point in time by examining outcomes for poor and nonpoor children drawn from large, national cross-sectional studies, primarily in the USA. Additional figures are reported from Canadian national data (Ross et al. 1989, Ross and Roberts 1999). In contrast, the final undertaking, an interdisciplinary initiative to conduct parallel evaluations across multiple data sets with income data measured over the life course, centers on understanding the role of the dynamics of poverty (depth, timing, and persistence) on a range of outcomes, after accounting for child and family background characteristics. These results on the long-term effects of poverty on children and youth were described in a volume, *Consequences of Growing Up Poor* (Duncan and Brooks-Gunn 1997); statistical criteria were used to determine the magnitude of income effects reported.

2.1 *Physical Health*

A country's level of economic development is associated with key indicators of child health; in developing countries (primarily those in sub-Saharan Africa and South Asia), over half of all deaths among children five years of age or younger are associated with malnutrition (UNICEF 2000). In many of these countries, AIDS/HIV is also contributing to rising infant mortality rates.

Research in the USA and Canada indicates that before children are born, poverty takes a toll on their health; children from poor families are approximately twice as likely to be low birth weight (less than 2,500 grams) than children who escape poverty (Brooks-Gunn and Duncan 1997, Ross and Roberts 1999, Ross et al. 1989). Other threats to poor children's health include lead poisoning and experiences with hunger as well as increased risk of mortality, hospitalization, and stunting (small for height). Poor children display more health problems and disabilities than do children from affluent families (Newacheck et al. 1993). When the longer term effects of poverty are considered, living in poverty during the childhood years has been found to be associated with stunting later in life (Duncan and Brooks-Gunn 1997).

2.2 *School Readiness and Achievement*

Access to basic education is associated with a country's level of economic development, such that in many developing countries, at least one-third of children are not enrolled in primary school, resulting in lower literacy rates compared with industrialized countries (UNICEF 2000). Among children of all ages, living in

poverty is also associated with deficits in cognition and achievement (Brooks-Gunn and Duncan 1997, Ross and Roberts 1999, Ross et al. 1989). Findings from US studies indicate that poor children 3 to 17 years of age are about one and a half times more likely than nonpoor children to display learning disabilities. Among school-age children, poverty increases the risk that children repeat a grade, perform poorly in school, are expelled or suspended from school, and obtain low scores on tests of school readiness and achievement. By late adolescence, poor youth are more likely than their nonpoor peers to drop out of school and to be unemployed. Taken together, these studies point to the disadvantage poverty bestows upon children's academic development; however, they do not shed light on the long-term effects of poverty on achievement.

A series of studies by Brooks-Gunn, Duncan, and colleagues have sought to understand the longer term effects of poverty on young children's cognitive and school outcomes. Findings have been highly consistent; after accounting for family and child background characteristics, income has large effects on young children's cognition, verbal ability, and achievement with a large effect defined as about five to six points on standardized tests. The depth of poverty is also important; children who were very poor (deep poverty defined as incomes 50 percent or less than the poverty threshold) scored particularly low on measures of cognitive and school achievement. The number of years a child lives in poverty matters as well for children's school readiness, particularly living persistently (all years) in poverty.

Longitudinal studies of young adolescents in the USA and Canada, generally, report small to modest income effects on indicators of school performance including grade placement and grade point average (Duncan and Brooks-Gunn 1997). Among studies of late adolescence, income was also found to have small to moderate associations with measures of completed schooling (Haveman and Wolfe 1994). The timing of poverty is important for late adolescent achievement; experiences with poverty during early childhood, as opposed to other developmental periods, are most salient for educational attainment (Duncan and Brooks-Gunn 1997).

2.3 *Behavior, Social, and Emotional Problems*

Similar to health and achievement, the threats to poor children's social and emotional health are great. A country's level of economic development places children at risk for exposure to hazardous and exploitative labor as well as conflict, war, and violence (UNICEF 2000). Even in the USA, poor children are twice as likely to be the victim of a violent crime and live in unsafe neighborhoods than nonpoor children (Brooks-Gunn and Duncan 1997). Among older children in Canada, poor children exhibit higher levels of emotional problems than nonpoor children, and are approximately twice as likely to have psychiatric and conduct disorders than their nonpoor peers (Ross et al. 1989). Finally, among adolescents in the USA, the rate of teenage unmarried parenthood is three times higher among poor female adolescents than nonpoor youth. These risk figures suggest that poor children and youth are faring at less than an optimal level.

In longitudinal studies, family income has small to moderate associations with the prevalence of young and school-age children's behavior problems as reported by parents and teachers (Brooks-Gunn and Duncan 1997). The depth, timing, and persistence of experiences with poverty during early childhood also are linked to children's socioemotional well-being. Finally, small to moderate income effects on adolescent anxiety and hyperactivity have been reported (Duncan and Brooks-Gunn 1997).

3. *The Pathways between Poverty and Child Development*

In the previous section, the direct effects of poverty on children's and youth's health and well-being were reviewed. Together, the research presents a consistent picture of the harmful effects of poverty, particularly during early childhood. To increase our knowledge about the influence of poverty on children's development, it is necessary to identify the pathways between poverty and child development. This exercise is particularly constructive for drawing policy implications. Several possible processes have been hypothesized to transmit the effects of income to child development; each is reviewed in turn.

3.1 *Home Environment*

The quality of the home environment is a salient mechanism through which income may alter child development. The home environment entails: emotional warmth displayed by parents while interacting with their children; provision of stimulating and learning experiences in the home; and physical surroundings, such as safety of play areas and cleanliness. Poor children are likely to be raised in less emotionally, cognitively, and physically enriching environments than children from more affluent families. Several studies of young children have found that quality of the home learning environment accounts for a substantial share of the effect of poverty on children's achievement and behavioral outcomes.

3.2 *Child Care and Schools*

Child care and schools are important contexts for children that may transmit the effects of low income to developmental outcomes. During young childhood,

high quality child care and early intervention programs have demonstrated short-term and long-term benefits for poor children's academic, social, and behavioral outcomes (Yoshikawa 1994). Moreover, these programs have the potential to enhance parenting outcomes as well, such as mental health and parenting skills. Among older children and youth, income effects on achievement outcomes may operate in large part through school characteristics, such as quality, climate, and demographics (Jencks and Mayer 1990).

3.3 Parenting Behavior

Economic hardship resulting from income loss, unemployment, job instability, and economic insecurity rather than low income *per se* is another potential mediator of income effects on child well-being. According to this 'family stress model,' economic hardship is associated with parental stress and depression which leads to more inconsistent, unsupportive, harsh, and punitive parenting practices. These reduced quality parent–child interactions, in turn, are associated with emotional and school problems in children (Conger et al. 1994, McLoyd et al. 1994).

3.4 Parental Mental Health

Maternal emotional well-being may be a path through which poverty affects child development; it may operate directly on child well-being or through parenting, as discussed under the family stress models. Substantial evidence exists that poverty has negative consequences on adult mental health (Kessler 1997). Additional studies reveal links between maternal emotional well-being and children's behavior (Velez et al. 1989). One study of young children found that income might have an indirect effect on children's behavior and cognition by means of maternal depressive symptoms and coping skills (Duncan et al. 1994).

3.5 Neighborhoods

An additional mechanism through which family income may impart its effects to child outcomes is neighborhood of residence. Inadequate income constrains the choice of where a family chooses to live, and as such, low income families may reside in very poor neighborhoods characterized by crime, unemployment, lack of resources, few role models, and absence of adult supervision of children and youth. A growing body of research indicates that neighborhood conditions affect children's development above and beyond family income, rather than indirectly (Leventhal and Brooks-Gunn 2000). Affluent (compared with middle income) or high SES neighbors appear to have beneficial effects for children's school readiness and achievement outcomes, while low income (compared with middle income) or low SES neighborhoods have negative consequences for children's and youths' emotional and social well-being. Neighborhood residence, notably income/SES, is also associated with quality of home environment, parenting practices, quality of child care and schools, and peer group attributes.

3.6 Health Care

Accessibility to health care is another pathway through which poverty may operate on child health and well-being. Although children from economically disadvantaged families are in poorer health, they are less likely to receive health care services or receive services of lower quality than children from more affluent families (Newacheck et al. 1993). Several studies suggest that access to quality care may mitigate health-related problems.

3.7 Exposure to Violence

The final process by which poverty may exert an influence on child well-being is exposure to violence inside as well as outside of the home. Children's exposure to violence as either a witness or a victim may account in part for the link between income and children's health. Both family and neighborhood income are associated with exposure to violence (Martinez and Richters 1993, Sampson et al. 1997). Exposure to violence, in turn, is associated with psychiatric symptoms.

4. Conclusion and Future Directions

In conclusion, large international discrepancies are seen in child poverty rates. Poverty rates, however, depend upon the type of measure used, demographic differences across countries, and government policies. The research indicates that poverty has negative short-term and long-term influences on a host of child outcomes. In addition, the timing, depth, and duration of experiences with poverty are also important; deep and persistent poverty during early childhood are especially detrimental to child well-being. Increasingly, researchers are beginning to explore the potential pathways through which poverty transmits it deleterious effects to children's development. There is growing evidence that the home environment and parental behavior are primary mechanisms of influence. Such knowledge is crucial for policies seeking to counter the negative consequences of poverty for poor children and their families. These policies can target

family resources directly or the mechanisms through which poverty operates on child development. In addition, policies can impact children directly through, for example, assistance with child care, or indirectly through a policy's impact on the family unit and neighborhood, for instance labor policies. Policies that coordinate the needs of both children and their parents are likely to be more efficacious than those that focus solely on one or the other. An important task for researchers will be to evaluate the effects of poverty-related policies on children's development.

See also: Child Care and Child Development; Early Childhood: Socioemotional Risks; Infancy and Childhood: Emotional Development; Infant and Child Development, Theories of; Parenting in Ethnic Minority Families: United States; Population, Economic Development, and Poverty; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Poverty, Culture of; Poverty in History; Poverty: Measurement and Analysis; Poverty Policy; Poverty, Sociology of; Urban Poverty in Neighborhoods

Bibliography

Bradbury B, Jantti M 1999 *Child Poverty Across Industrialized Nations*. Innocenti Occasional Papers, Economic and Social Policy Series No. 71. UNICEF, Sydney, Australia

Brooks-Gunn J, Duncan G J 1997 The effects of poverty on children. *The Future of Children* **7**(2): 55–71

Conger R D, Ge X, Elder G, Lorenz F, Simons R 1994 Economic stress, coercive family process and developmental problems of adolescents. *Child Development* **65**: 541–61

Duncan G J, Brooks-Gunn J (eds.) 1997 *Consequences of Growing up Poor*. Russell Sage Foundation Press, New York

Duncan G J, Brooks-Gunn J, Klebanov P K 1994 Economic deprivation and early-childhood development. *Child Development* **65**: 296–318

Eurostat 1999 Social benefits and their redistributive effect on the EU. *Statistics in Focus*, Population and Social Conditions, no. 13/99. Eurostat, Luxembourg

Haveman R, Wolfe B 1994 *Succeeding Generations: On the Effects of Investments in Children*. Russell Sage Foundation, New York

Jencks C, Mayer S E 1990 The social consequences of growing up in a poor neighborhood. In: Geary L L (ed.) *Inner-city Poverty in the United States*. National Academy Press, Washington, DC, pp. 111–86

Kessler R C 1997 The effects of stressful life events on depression. *Annual Review of Psychology* **48**: 191–214

Leventhal T, Brooks-Gunn J 2000 The neighborhoods they live in: The effects of neighborhood residence upon child and adolescent outcomes. *Psychological Bulletin* **126**: 309–37

Martinez J E, Richters P 1993 The NIMH Community Violence Project: II. Children's distress symptoms associated with violence exposure. *Psychiatry* **56**: 22–35

McLoyd V C, Jayaratne-Epstein T, Ceballo R, Borquez J 1994 Unemployment and work interruption among African American single mothers: Effects on parenting and adolescent socioemotional functioning. *Child Development* **65**: 562–89

Newacheck P W, Stoddard J J, McManus M 1993 Ethnocultural variations in the prevalence and impact of childhood chronic conditions. *Pediatrics* **91**: 1031–8

Ross D P, Roberts P 1999 *Income and Child Well-being: A New Perspective on the Poverty Debate*. Canadian Council on Social Development, Ottowa, ON

Ross D P, Shillington R, Lochhead C 1989 *The Canadian Fact Book on Poverty*. Canadian Council on Social Development, Ottowa, ON

Sampson R J, Raudenbush S W, Earls F 1997 Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* **277**: 918–24

Smeeding T M 1997 *Financial Poverty in Developed Countries: The Evidence from LIS*. Working Paper No. 155. Syracuse University, New York

Smeeding T M, Rainwater L, Burtless G 2000 *United States Poverty in a Cross-national Context*. Syracuse University, New York.

United Nations Children's Fund (UNICEF) 2000 *The State of the World's Children 2000*. UNICEF, New York

Velez C N, Johnson J, Cohen P 1989 A longitudinal analysis of selected risk factors for childhood psychopathology. *Journal of the American Academy of Child and Adolescent Psychiatry* **28**: 861–4

Yoshikawa H 1994 Prevention as cumulative protection: Effects of early family support and education on chronic delinquency and its risks. *Psychological Bulletin* **115**: 28–54

T. Leventhal and J. Brooks-Gunn

Poverty and Gender in Affluent Nations

Women are more likely to be poor than men in nearly all Western nations, with the greatest disparity found in the USA. The underlying causes for women's poverty vary across nations but generally fall into one of three main categories—demographic composition, economic conditions, and government policy. In this article, poverty is defined, and the advantages and disadvantages of different poverty measures are noted. Then, the recent scholarship on gender and poverty is briefly reviewed, and women's poverty rates in various Western, industrialized countries are compared. Finally, the three primary causes of women's poverty are discussed.

1. Definitions of Poverty

The concept of 'poverty' is relatively new, having only entered the public lexicon in the latter half of the twentieth century. Poverty can be defined in either absolute or relative terms. It is typically measured with respect to families—not individuals—and is adjusted for the number of persons in a family (because larger families are assumed to need more income but also to benefit from economies of scale). *Absolute* measures of poverty (typically used in Anglo countries such as the

USA and the UK) define poverty in relation to the amount of money necessary to meet basic needs such as food, clothing, and shelter. *Relative* poverty measures (typically used in European countries and by academic researchers) define poverty in relation to the economic status of other members of the society. One half of the median family income is often chosen as the poverty threshold. Absolute and relative measures of poverty reflect divergent ideas about the meaning of poverty—whether it represents the inability to afford basic necessities, or comparative economic disadvantage—and both have advantages and disadvantages. The advantage of an absolute poverty measure is that it clearly indicates whether or not a family's income is sufficient to sustain a minimal level of subsistence.

However, if an absolute threshold is adjusted only for inflation, during periods of economic growth that same threshold will increasingly reflect a greater level of relative destitution. Relative measures avoid this problem because they are linked to the average standard of living: when median income goes up, the poverty threshold goes up. Yet, because of this practice, 'poverty' in one country may imply a vastly different economic status than 'poverty' in another country. Relative measures are particularly affected by the income distribution: in countries with a more unequal distribution, poverty rates will be high; in countries with a more equal distribution, they will be low—regardless of the actual standard of living that 'poverty' represents.

2. *Women's Poverty in Comparative Perspective*

The salience of gender for understanding poverty was first noted by Diana Pearce (1978) who coined the term 'the feminization of poverty.' Pearce observed that women and children living in female-headed families were over-represented among the poor, and that the proportion of adults in poverty who were women had been increasing. Since that time, various scholars have used data to examine trends in men's and women's poverty rates—and the ratio between them—in order to explore how economic status may be affected by gender (see, for example, McLanahan et al. 1989, Casper et al. 1994). Feminist scholars have raised important issues about the gendered nature of welfare state policies (Folbre 1984, Gordon 1990, 1994, Orloff 1993, O'Connor et al. 1999) and the inequitable distribution of the costs of children (Folbre 1994a, 1994b).

Because poverty is measured at the family level, for married couples, poverty is by definition the same for women as for men: if a married-couple family is classified as poor, the husband, wife, and any children are all considered to be poor. Only for single individuals (defined by most researchers as neither married nor cohabiting) can the rates of poverty differ by gender. Thus, differences in the proportion of the population that is single, and differences in the poverty rates of single men and single women, underlie the overall gender differences in poverty both within and across nations.

Table 1 presents poverty rates in eight Western, industrialized nations. A relative measure of poverty is used, and an individual is counted as poor if he or she lives in a household with disposable money income that is less than half the median for households in the nation, adjusted for size. The table presents poverty rates for all women and men, for single women and men, and for two groups of single women that are particularly disadvantaged—single mothers and elderly women living alone.

Overall, the table highlights three fundamental observations about gender and poverty in highly industrialized nations. First, within each of the demo-

Table 1
Poverty rates for select demographic groups in eight industrialized nations[a]

Country (year)[b]	All women	All men	Single women	Single men	Single mothers[c]	Elderly women living alone
Australia (1994)	10.7	8.2	24.7	14.8	38.6	62.1
Canada (1994)	13.1	11.6	29.1	17.4	40.5	16.2
France (1989)	10.7	9.6	16.4	13.2	24.7	NA
Germany (1994)	9.8	8.3	22.6	11.8	39.1	12.7
Netherlands (1991)	5.6	4.9	13.2	8.1	25.7	3.4
Sweden (1992)	2.9	4.0	4.8	9.3	3.4	14.7
United Kingdom (1995)	10.9	9.1	21.4	13.6	31.9	50.1
United States (1994)	15.1	11.0	32.0	16.6	47.1	43.1

Source: Figures for the first five columns are from Christopher et al. (2001); data for the last column are from Smeeding (1998).
NA: Not available.
a Poverty is defined as less than 50 percent of a country's median income. All columns except the last are for nonelderly adults ages 25 to 54; the final column refers to elderly women ages 65 and older. 'Single' indicates that an individual is neither married nor part of a cohabiting male/female couple.
b Years of data for the first five columns are shown in parentheses. For the figures on elderly single women, years of data are as follows: Australia 1989, Canada 1991, Germany 1989, Netherlands 1991, Sweden 1992, U.K. 1991 and US 1994.
c Single mothers are women living with a child under the age of 18 and without an adult male.

graphic groups shown, poverty levels vary across the nations observed. For example, Sweden has very low poverty rates for most categories, while the USA has notably higher poverty rates. Second, in all but one country shown, women are more likely to be poor than men; the exception is Sweden, where men are slightly more likely to be poor than women. In the USA, the gap between women's and men's poverty rates is the highest—about four percentage points. As would be expected—since the single population drives the gender differences in poverty—the gender gap in poverty rates is larger for single individuals than for the adult population overall. Again, Sweden runs counter to the trend, with a lower poverty rate for single women than for single men. For all other countries, the difference in poverty rates between single men and women ranges from about three percentage points in France to more than 15 percentage points in the USA.

The third observation is that among single women, single mothers and elderly women living alone are two groups that face particular economic disadvantage. For single mothers, poverty rates are 25 percent or higher for all countries shown except Sweden; in Australia, Canada, and Germany, about 40 percent of single mothers are poor; and in the USA, nearly half of all single mothers live in poverty. For elderly women living alone, poverty rates range from 3 percent in The Netherlands to 62 percent in Australia, with rates of more than 40 percent observed in both the USA and the UK.

3. *Why are Women Poor?*

There is no single explanation for why women are poor—and why they have higher poverty rates than men—particularly as one looks across countries that have vastly different histories, cultures, and social and economic contexts. Three primary factors have emerged, and each will be discussed in turn.

3.1 *Demographic Composition*

Since differences in poverty rates by gender are fundamentally determined by the size and characteristics of the population that is single (neither married nor cohabiting), the demographic distribution in various countries has an important effect on women's poverty rates. Several major demographic trends have led to an increasing proportion of adult women (many of whom have children) living separately from men. First, differential mortality rates by gender imply that women live longer than men, and women are, thus, more likely to live alone in their later years. Since single elderly women have higher poverty rates than either men living alone or women living with a spouse (Smeeding 1997), in countries with a greater proportion of elderly women living alone, poverty rates for women will be higher. Yet, mortality rates among Western, industrialized nations are not highly variant; therefore, poverty among elderly women living alone has only a modest impact on overall gender differences in poverty compared to the second major demographic factor, the rise in single-parent families.

Table 2
Single mothers in eight industrialized nations

Country (year)	Single mothers as a percent of all women ages 25 to 54
Australia (1989)	7.8
Canada (1994)	8.5
France (1989)	5.8
Germany (1994)	7.0
Netherlands (1991)	6.0
Sweden (1992)	9.9
United Kingdom (1995)	9.8
United States (1994)	12.7

Source: Data are from Smeeding et al. (1999) using the Luxembourg Income Study (LIS).

Increases in divorce and nonmarital childbearing have led to notable growth in the number of single-mother families. Although the extent of single parenthood varies across countries, the trend has been remarkably similar—single parenthood has increased in nearly all Western, industrialized nations since the 1960s (McLanahan and Casper 1995). Table 2 shows single mothers as a percentage of all women ages 25 to 54 for eight industrialized nations. Of the countries shown, the proportion of single mothers ranges from 5.8 percent in France to 12.7 percent in the USA (Smeeding et al. 1999).

If the economic costs of rearing children were evenly distributed between mothers and fathers, then even if the parents were not married, the sex–poverty ratio would not be affected. However, in reality, single parenthood has a much greater affect on women's poverty rates than men's (McLanahan and Kelly 1999). Women are more likely to have custody of a child following divorce or nonmarital childbirth, and many fathers who live away from their children do not provide adequate financial support for them. Having co-resident children decreases a household's economic status by definition, because children require economic resources but do not contribute any additional income. Unless the government provides generous financial support when the noncustodial father fails to pay (as is the case in Sweden), in countries with a high proportion of single mothers, poverty rates for women will be high. For this reason, the prevalence of single motherhood represents the major demographic factor underlying the gender disparity in poverty rates.

Table 3
Economic characteristics of women and single mothers ages 25 to 54 in eight industrialized nations

	Labor force participation rates		Mean earnings[a]	
Country (year)	All women	Single mothers	All women	Single mothers
Australia (94)	63.0	49.3	$15,522	$11,904
Canada (94)	73.1	63.2	$17,391	$16,922
France (89)	57.2	70.3	$13,671	$14,083
Germany (94)	68.0	67.4	$15,033	$14,745
Netherlands (91)	50.0	38.1	$13,650	$10,946
Sweden (92)	88.0	83.8	$13,075	$12,362
United Kingdom (95)	60.5	37.7	$15,196	$13,597
United States (94)	74.2	71.8	$21,186	$18,547

Sources: Data are from special tabulations of the Luxembourg Income Study (LIS) data by Karen Christopher at the University of Arizona.
a For those who have non-zero earnings. Figures are in 1994 US dollars, adjusted for purchasing power parity of each country.

3.2 Economic Conditions

The rise in female labor force participation, especially among married women with children, represents one of the most dramatic socioeconomic changes in the West in the latter half of the twentieth century. Table 3 shows that labor force participation rates for all women (in eight countries) range from 50 percent in The Netherlands to 88 percent in Sweden; employment rates for single mothers vary between 38 percent in The Netherlands to 84 percent in Sweden. While women are increasingly entering the labor force, their employment rates remain lower than those of men. Further, women have average earnings that are significantly lower than those of men. As shown in Table 3, women's average earnings (among those who have any earnings) range from approximately $13,000 in Sweden to about $21,000 in the USA; earnings for single mothers vary between about $11,000 in The Netherlands to $18,500 in the USA.

Because, relative to men, women have less income from earnings to support themselves and their children, women are more likely to be poor. (This, however, is not a factor affecting elderly women's poverty as the elderly are no longer of working age.) Yet, earnings *per se* are not the only reason why women have higher poverty rates. Rather, the fact that women disproportionately bear the responsibility of rearing children has important indirect effects on their economic status. In particular, caring for children limits the amount of time that women are able to work (as well as the energy they may have available to invest in their work), and it also increases the costs of working because women must pay for out-of-home child care (not counted in the poverty measure). In addition, women who reduce their work hours or leave the labor force in order to care for children lose vital work experience which affects both their earnings and their seniority in the workplace.

3.3 Government Policy

All individuals and families receive income from two primary sources—the market and the state—in addition to any private transfers they may receive. Government policy has a significant effect on women's economic well-being because it can either obviate or intensify the inequalities that result from the operation of the market economy. Among single individuals living without children, government policy has little impact on women's poverty because men and women are largely treated the same. Although there may be some differences between elderly men and women in how pensions are allocated in various countries, differences in government policies primarily affect the economic status of single mothers and their children.

Welfare states vary dramatically in the extent to which they 'socialize' the costs of children. Feminist scholars have noted that many members of society benefit from children being brought up well (Folbre 1994a, 1994b), and yet market mechanisms fail to equitably apportion costs for this 'public good' (Christopher et al. 2001). In countries where more of the costs of childrearing are borne by governments instead of parents, women's poverty rates tend to be lower. In general, single mothers tend to benefit if a greater proportion of welfare assistance is universal rather than means-tested, because they are not faced with high marginal tax rates on their earnings that result from decreased welfare benefits. The overall level of income assistance provided to single mothers varies notably across industrialized nations; mean transfers for single mothers as a percentage of median equivalent income range from 22 percent in the USA to 71 percent in The Netherlands (unpublished tabulations of Luxembourg Income Study data by Timothy Smeeding and Lee Rainwater).

Numerous scholars have developed typologies that classify welfare states along particular dimensions.

Table 4
Poverty rates for children in single-mother families[a] before and after government programs[b]

Country (year)	Pretransfer poverty rate	Post-transfer, post-tax poverty rate	Percent decline
Australia (1989)	73.2	56.2	23.2
Belgium (1992)	50.7	10.0	80.3
Canada (1991)	68.2	50.2	26.4
Denmark (1992)	45.0	7.3	83.8
Finland (1991)	36.3	7.5	79.3
France (1984)	56.4	22.6	59.9
Germany (1989)	43.9	4.2	90.4
Ireland (1987)	72.6	40.5	44.2
Italy (1991)	31.7	13.9	56.2
Luxembourg (1985)	55.7	10.0	82.0
Netherlands (1991)	79.7	39.5	50.4
Norway (1991)	57.4	18.4	67.9
Sweden (1992)	54.9	5.2	90.5
Switzerland (1982)	33.7	25.6	24.0
United Kingdom (1986)	76.2	18.7	75.5
United States (1991)	69.9	59.5	14.9

Source: Rainwater and Smeeding 1995.
a Single-mother families are families where one female adult resides in the household; cohabiting-parent families are not included here. Poverty is defined as 50 percent of a country's median adjusted income.
b Government programs include income and payroll taxes and all types of government cash and near-cash transfers. Child care and other non-cash benefits are not included; this is particularly important for countries such as France which provide generous child care subsidies that are not reflected in these calculations.

Probably the best known is by Esping-Anderson (1990) which classifies welfare states into three categories according to decommodification (the ability of individuals to subsist apart from labor and capital markets) and stratification (differences between classes).

(a) *Social democratic states*, typified by the Scandinavian countries, promote gender equality and provide the most generous support to single-parent families. These nations emphasize full employment, and work and welfare are intricately intertwined, with some part-time public jobs provided for mothers. Many benefits and allowances are universal, supplemented by specific programs targeted to needy families. Women's employment is encouraged by the provision of high-quality childcare services and extensive parental leave. Also, the government guarantees that support is provided to children who do not live with both parents. Because of these policies, mothers are able to combine work and parental responsibilities and to 'package' income from both market and government sources.

(b) *Conservative-corporatist nations* also have generous transfer systems, but they are more traditional with respect to family values and expectations (largely due to the significant influence of the Church). Benefits are generally targeted toward families (rather than individuals), and inequalities across families and households may be sustained. An important component of welfare policy is social insurance, which is linked to labor force participation and may differ by occupation. Austria, France, Germany, and Italy are representative conservative-corporatist states.

(c) In the *liberal welfare states* (notably the Anglo countries), assistance is primarily means-tested with modest universal transfers, primarily paid to the elderly or disabled. Emphasis is on the market and individuals' labor force activity as the primary means for resource allocation. Benefit levels are typically meager when compared to average wage levels, and recipients are often stigmatized by nonrecipients. Supports for working mothers, including public childcare, are limited, with little or no paid parental leave. Parents who live away from their children are expected to pay child support, although the government does not provide any assistance for children of noncustodial parents who fail to pay. Because welfare benefits are income-tested, women have a difficult time combining work and welfare in order to escape the 'poverty trap.'

These variations in welfare state policy lead to dramatic differences in the economic status of women and children. Table 4 presents poverty rates for children living in single-mother families both before and after government taxes and transfers are included in income (while childcare and other non-cash benefits are not included). The table shows that, with several exceptions, the percentage decline in the poverty rates roughly clusters according to the three categories of nations in Esping-Anderson's typology. Government tax and transfer policies reduce the poverty rates in each of the four social democratic nations shown in the table (Denmark, Finland, Norway, and Sweden)

by 68 percent or more. In the corporatist nations, poverty reductions fall in the middle of the spectrum with a 60 percent decline in France and a 56 percent decline in Italy. Germany is an important exception, with its government policies reducing poverty for children in single-mother families by fully 90 percent. The lowest reductions in poverty rates are noted in the liberal states, such as 26 percent for Canada, 23 percent for Australia, and 15 percent for the USA. The UK represents another exception because its policies reduce poverty by 76 percent—a much greater decline than in the other liberal states.

Some feminist scholars have criticized Esping-Anderson's framework because it emphasizes workers' dependence on employers while ignoring women's dependence on men. Orloff (1993) argues, for example, that welfare regimes should be judged by the extent to which they allow women to establish independent households. In contrast, other feminist scholars have raised questions about women's dependence on the welfare state (Gordon 1990, 1994) and gender biases in the welfare state (Nelson 1990).

4. Conclusion

Ultimately, reducing women's poverty in industrialized countries requires the convergence of all three factors—demography, the market economy, and government policy—to better support women and their children. This is because women's economic status depends on whatever resources they can conjointly obtain from the family, the market, and the state (Christopher et al. 2001). First, women's poverty is closely tied to whether the responsibility for children is shared equally between mothers and fathers. Private transfers, regulated by the government, may be just as important as public transfers in compensating for the demographic changes that have resulted in women bearing a disproportionate share of the cost of children. Second, women benefit from market economies that pay men and women equally. Despite the recent gains in wages relative to men, women (especially mothers) earn less than men in all Western countries. Countries that promote gender equity in wages and that maintain high minimum wages reduce the poverty rate of all women, especially single mothers. Third, government policy can be assessed both in terms of the level of support provided to single mothers and the extent to which public support encourages or discourages labor force participation. In particular, childcare subsidies and parental leave policies enable women to balance their responsibilities as parents and workers, which ultimately reduces poverty and increases economic independence.

See also: Divorce and Gender; Domestic Violence: Sociological Perspectives; Economic Development and Women; Economic Globalization and Gender; Family and Gender; Family Bargaining; Gender, Economics of; Inequality; Labor Movements and Gender; Lone Mothers in Affluent Nations; Motherhood: Economic Aspects; Poverty and Gender in Developing Nations; Poverty Policy; Poverty, Sociology of; Power in Society; Race and Gender Intersections; Sex Differences in Pay; Sex Segregation at Work; Social Welfare Policies and Gender; Underclass; Urban Poverty in Neighborhoods; Welfare Programs, Economics of; Welfare State

Bibliography

Casper L M, McLanahan S S, Garfinkel I 1994 The gender-poverty gap: what we can learn from other countries. *American Sociological Review* **59**: 594–605

Christopher K, England P, McLanahan S, Ross K, Smeeding T M 2001 Gender inequality in poverty in affluent nations: the role of single motherhood and the state. In: Vleminckx K, Smeeding T M (eds.) *Child Well-Being, Child Poverty and Child Policy in Modern Nations*. The Policy Press, Bristol, UK, pp. 199–219

England P 1997 Gender and access to money: what do trends in earnings and household poverty tell us? *8th International Meeting. Comparative Project on Class Structure and Class Consciousness*. Canberra, Australia

Esping-Anderson G 1990 *The Three Worlds of Welfare Capitalism*. Polity Press, Cambridge, UK

Folbre N 1984 The pauperization of motherhood: patriarchy and public policy in the United States. *Review of Radical Political Economics* **16**(4): 72–88

Folbre N 1994a Children as public goods. *American Economic Review* **84**: 86–90

Folbre N 1994b *Who Pays for the Kids?: Gender and the Structures of Constraint*. Routledge, New York

Gordon L (ed.) 1990 *Women, the State, and Welfare 1890–1935*. University of Wisconsin Press, Madison, WI

Gordon L 1994 *Pitied But Not Entitled: Single Mothers and the History of Welfare*. Harvard University Press, Cambridge, MA

McLanahan S, Casper L 1995 Growing diversity and inequality in the American family. In: Farley R (ed.) *State of the Union: America in the 1990s, Vol. 2 Social Trends*. Russell Sage Foundation, New York, pp. 1–45

McLanahan S S, Kelly E L 1999 The feminization of poverty: past and future. In: Chafetz J (ed.) *Handbook of the Sociology of Gender*. Kluwer, Dordrecht, The Netherlands, pp. 127–45

McLanahan S S, Sorenson A, Watson D 1989 Sex differences in poverty, 1950–1980. *Signs:Journal of Women in Culture and Society* **15**: 102–22

Nelson B J 1990 The origins of the two-channel welfare state: workmen's compensation and mother's aid. In: Gordon L (ed.) *Women, the State, and Welfare*. University of Wisconsin Press, Madison, WI

O'Connor J S, Orloff A S, Shaver S 1999 *States, Markets, Families: Gender, Liberalism, and Social Policy in Australia, Canada, Great Britain, and the United States*. Cambridge University Press, New York

Orloff A S 1993 Gender and the social rights of citizenship. *American Sociological Review* **58**: 303–28

Pearce D M 1978 The feminization of poverty: women, work and welfare. *Urban and Social Change Review* **11**: 28–36
Rainwater L, Smeeding T M 1995 Doing poorly: the real income of American children in a comparative perspective. In: Skolnick J H, Currie E (eds.) *Crisis in American Institutions*. Allyn and Bacon, Boston, pp. 118–25
Smeeding T M 1998 Reshuffling responsibilities in old age: the United States in a comparative perspective. In: Gonyca J (ed.) *Re-Securing Social Security and Medicare: Understanding Privatization and Risk*. Gerontological Society of America, Washington, DC, pp. 24–36
Smeeding T M, Ross K, England P, Christopher K, McLanahan S S 1999 Poverty and parenthood across modern nations: findings from the Luxembourg Income Study. Luxembourg Income Study, Working Paper No. 194

S. S. McLanahan and M. J. Carlson

Poverty and Gender in Developing Nations

Poverty as used in contemporary literature is an economic measure of income or consumption. A person or household is considered below the 'poverty line' if such income or consumption falls below a minimum standard needed to provide basic human needs. Economic development in developing countries is considered successful when the gross national product (GNP) increases and raises the per capita income. Rapid population increase dilutes the impact of GNP and led to international family planning programs as one approach to poverty reduction. This direct connection between economic development and population decrease is often questioned. The insensitive implementation of family planning programs has been widely criticized by women who initiated the emphasis on women's education and health to reduce both poverty and family size at the 1994 United Nations International Conference on Population and Development in Cairo (see *Reproductive Rights in Developing Nations*).

The poverty line is linked to a minimum standard within each country which varies by country, region, or culture. Attempts to compare income poverty across societies has led the World Bank to utilize the Purchasing Power Parity (PPP) that accounts for currency fluctuation and local consumption habits and is commonly referred to as the market-basket approach.

Although economic measures dominate the literature on poverty and on development, critics have objected to such sterile methods of accounting. Additional indicators of wellbeing have been combined into more textured indices of progress or poverty. The social aspects of poverty do not lend themselves to such aggregation. To understand the multidimensional nature of poverty, scholars are today investigating a broad range of social and cultural practices as well as governmental policies. Only with this comprehension of the causes of poverty can more effective programs for poverty alleviation be designed. This effort has taken on increased urgency in light of current economic trends which have exacerbated income disparity within countries and pushed more people into poverty (Lundberg and Milanovic 2000).

Differential patterns of poverty are experienced by women and men due largely to their social roles and responsibilities. The increase in the number of women-headed households throughout the world, caused by rapid economic transition, migration, and urbanization, has led to the dissolution of many traditional support systems. As a result, women must both nurture and provide a livelihood for their children in an economy that privileges full-time work away from home. This double burden limits both a woman's choice of jobs and her mobility, thus reducing her income causing the feminization of poverty.

Even within needy families, women may be poorer than their husbands because they have little control over how income or food are allocated. Food aid programs now deliver food directly to poor women who show more equity in feeding their children Pinstrup-Andersen et al. 1999. Similarly, development agencies have focused more of their income projects on women (see *Economic Development and Women*). While some critics argue that loading additional work responsibilities on women has created a crisis of time, others demonstrate that income increases women's household bargaining power (see *Family Bargaining*). Conservatives object to programs that provide women with jobs because they believe that upsetting the gender division of labor encourages further disintegration of families. In contrast, feminists argue that the focus on welfare programs to alleviate the worst aspects of poverty will do nothing to reduce the subordination of women. Recent approaches to alleviating poverty emphasize asset creation and housing rights to empower women.

The first section explores the broadening of definitions, measures, and social indicators of poverty. The second section presents arguments for a gendered approach to poverty alleviation and discusses studies that elaborate on, or question, earlier assumptions. The last section describes significant programs designed to assist women out of poverty and reviews salient criticisms of them.

1. *Definitions and Causes of Poverty*

Defining poverty in economic terms alone raises issues of the accuracy of statistics in developing countries and encounters the range of moral objections to valuing life only in material terms. Development

agencies are aware of the unreliability of data from developing countries yet cognizant of the demands by policy makers for easily replicated measures. To compare poverty globally, the World Bank uses estimates of local income expressed as U$1 and US$2 per day to identify levels of poverty, estimating that in 1998, 1.2 billion people worldwide had consumption levels below $1 a day and 24 percent of the population of the developing world and 2.8 billion lived on less than $2 a day (World Bank 2000).

1.1 Indicators Beyond Income

As the GNP became the common method of measuring progress, critiques appeared. Alternative measures, such as the Physical Quality of Life Index (PQLI) promoted by the Overseas Development Council in Washington DC in the 1970s, used indicators for life expectancy, infant mortality, and literacy for an index that often reversed ideas of improved life through increased income. For example, Saudi Arabia rated high for income and low for PQLI while Sri Lanka had a low GNP and a high PQLI. These indicators actually measure women's status, though that was not the original intent, since poor health for women results in high maternal mortality that often also means high infant mortality while women's literacy reflects educational opportunities and cultural openness.

Following this trend, the Human Development Index (HDI) refines these indicators into three indices based on life expectancy, education, and gross domestic product (GDP). Utilizing these and other indices, the Human Development Report makes explicit the need for gender analysis in the Gender Development Index (GDI) and the Gender Empowerment Measure (GEM). Saudi Arabia is placed in the medium human development list but ranks low in GDI and has no entry for GEM. Three developing countries, Sri Lanka, Thailand, and Jamaica, have a GDI rank that is more than 10 points higher than the HDI rank (UNDP 1998).

In 1997 the Human Development Report introduced the Human Poverty Index (HPI) to emphasize that poverty is much more than a lack of material wellbeing, but also a denial of opportunities to live a healthy and creative life and enjoy both self-esteem and the respect of others. While the HDI measures national progress, the HPI measure the deprivations of those who are left out of progress (UNDP 1998 p. 25). Significant is the finding that the US has highest human poverty among the 17 developed countries listed although it has the highest per capita income measured in PPP.

1.2 Social Aspects of Poverty

As world poverty increases, development agencies are searching for more accurate social indicators for education, health, access to services and infrastructure. Other nonincome dimensions of poverty are being studied to tease out and evaluate illusive patterns of vulnerability, social exclusion, and gender inequities. Analyses of poverty in terms of access to social capital, entitlements to food, income, land, and houses are being conducted. These indicators do not lend themselves to aggregation into a single index. Rather this broadening of research reflects the complexities of addressing social aspects of persistent poverty (Chambers 1995). Reports on urban poverty in cities around the world and programs that improve the lives of the poor have been published in three issues of *Environment & Development*. In preparation for the World Development Report 2000/1 on Poverty and Development, the World Bank commissioned a wide range of working papers (World Bank 2000). Gendered poverty is the concern of studies produced by the United Nations Research Institute for Social Development (UNRISD 1999). The Inter-American Development Bank on causes of intergenerational transmission of poverty has resulted in an initiate on early childhood programs.

1.3 Structural Influences on Poverty

Economic and political forces external to culture and community also contribute to the current trend of increased poverty and income disparity around the world. The rapidity of the present economic transitions toward global trade and largely unfettered markets is exacerbating income disparity and contributing to the increase in poverty. Economic prescriptions to improve creditworthiness of countries, such as the Structural Adjustment Policies (SAP) of the International Monetary Fund and the World Bank, are widely blamed for the reduction of health and education budgets in those countries. Subsequently, these financial institutions have instituted changes in conditionality to require governments to balance cuts in the social safety net with other measures to reduce expenditures by trimming bloated bureaucracies or eliminating inefficient industrial projects.

Governments privilege different groups through macroeconomic choices regulating trade, investments, and food policy, among others (UNRISD 1999). Urban poverty is much influenced by the actions of city and municipal governments. Inefficient and corrupt bureaucracies contribute to income inequality and limit the effectiveness of poverty alleviation programs. In 1999, research teams in 23 countries held *Consultations with the Poor* (Narayan et al. 1999). A large majority of the poor said they were worse off, had fewer opportunities, and were more insecure than in the past. Most had little good to say about government personnel or services and turned to their own networks and institutions for support. The presence of nongovernmental organizations (NGOs) was uneven; their programs received mixed reviews, often seen as

irrelevant because, despite the goals of participation espoused by NGOs, the poor do not feel that their voices are heard.

2. Poverty as Gendered

The gender division of labor traditionally allocated domestic production to women, activities that included providing subsistence as well as care to their families. The shift from labor exchange to a monetized economy tended to leave women with tasks and responsibilities that are not paid and therefore are excluded from national accounts. Opportunities for men to earn money are more plentiful but may require migration for short or long periods of time, weakening family cohesion and kin networks while changes in traditional status hierarchies as well as income paths increases insecurity. Once men earn money they often feel entitled to spend it on drinking, gambling, or women rather than on their families. These events, drinking, insecurity, and lack of kin supervision, have led to a worldwide epidemic of domestic violence and AIDS (see *Domestic Violence: Sociological Perspectives*).

2.1 Women-headed Households

The same transitions encourage the increase of households headed by women (see *Economic Development and Women*; *Economic Globalization and Gender*). Definitions of and data on women-headed households are contested, as are most statistics in developing countries. Estimates of 30 percent have been found valid since 1975, but studies have refined categories to note the numbers of widows, or of women with children returning to parental homes and becoming invisible. Women-headed households are perceived as the poorest, though regional and class variations are noted. The International Food Policy Research Institute (IFPRI) reported that women-headed households in more than half of the countries examined in Asia and sub-Saharan Africa are poorer than households containing adult men. Recent household studies link urban poverty not only to gender but to life cycle, age, and housing options. Food security is of prime concern in poor households; women are more likely than men to direct more resources to food and children (United Nations 1995, p. 129). These findings have important implications for poverty alleviation programs.

2.2 Time Pressures of Women's Double Day

In order to purchase needed commodities or pay school fees, women are pushed to earn income as well as carry out domestic production. The dual responsibilities of women tend to limit their income opportunities and have generally resulted in women earning less than men in most of the world. Statistics on women's work, along with a discussion of problems of the definition of work and the imprecision of much data, are frequently published by the UN's Statistical Division (United Nations 1995). Time-use studies underscore the fact that women work longer hours than men in a large number of countries (see *Time-use and Gender*). This has led feminist scholars to promote appropriate technology to reduce women's work burdens (Tinker 1987). Others deplore the labor-intensive poverty programs since heavy physical work, called for in such poverty alleviation projects as road building, reduce the already limited nutritional reserves of poor women (UNRISD 1999). Government cost reduction programs that cut social spending generally add to women's time pressures as they increase hours of care-giving to children and relatives and often sacrifice income activities in the process (Moser 1996).

3. Poverty Alleviation Programming

Poverty reduction has been a major goal of economic development since the thrust toward basic human needs and redistribution with growth identified in the Second Development Decade, 1970–80 documents. The focus on basic needs encouraged development programs to address the concerns of people not just countries and promoted the women in development movement. Pressures from the growing global women's movement forced the debates over structural adjustment and free trade to consider their impact on women (see *Economic Development and Women*). At the end of the twentieth century, a New Policy Agenda emerged which considered market-led growth as the primary engine for reducing poverty. In developing countries, agriculture and labor-intensive activities were expected to provide employment. The role of the state was limited to service delivery, assisted by NGOs and other civil society organizations. The fact that the unfettered global market economy has led to more poverty has forced rethinking of poverty and its alleviation as noted above.

3.1 Programs to Alleviate Poverty

In the 1970s poverty alleviation programming was focused on rural areas. Appropriate technology was introduced to reduce women's energy and time required for subsistence activities such as collecting water and fuel, and processing food. Improved cookstoves and water-seal toilets also reduced the needs for water and fuel. Time saved became available for

literacy programs or income activities which at first were ill-adapted to poor women's skills (see *Economic Development and Women*).

3.2 Microcredit and Microfinance

Most current poverty alleviation programs focus on savings, credit for microenterprise, and the organizing home-based workers through NGOs. In combination these programs assist the poor women and men to accumulate assets and expand ties within the community which in turn encourages investments in human capital and creates social capital. The impact is more dramatic for women than men in those traditional societies where women and men are segregated socially, where women's mobility is severely limited, and where women's work is largely unremunerated. Not surprisingly, two of the most successful microcredit programs are in Bangladesh, the Grameen Bank, and India, SEWA (Self-Employed Women's Association). Savings are a prerequisite for credit which is utilized to start or improve tiny entrepreneurial activity from raising a goat to selling street foods to marketing fish. All programs require frequent attendance at group meetings which function both to break down caste, ethnic, and kin barriers and provide mechanisms for wide-scale organizing. SEWA has spearheaded the rapid expansion of Homenet, a worldwide organization for home-based workers which advocates for the extension of pensions, health benefits, and childcare to these workers. Meetings of these organizations promote the sharing of information on health, contraception, nutrition, and dealing with domestic violence as well as on business techniques. Women learn the benefits of investing in the health and education of their children, especially daughters. The Grameen Bank is proud of the number of children of members now attending secondary schools and colleges.

Profits from microenterprise and the savings programs have allowed women to purchase more durable assets such as a house or agricultural land. Many urban NGOs support programs by which the poor obtain housing through sweat-equity and credit. Women as the primary family caretakers need the security of shelter. Most laws or customary practices in developing countries still privilege male ownership of house and land (see *Land Rights and Gender*). Registration for these assets remains an obstacle though laws and traditions are changing, often under pressure from NGOs (Tinker and Summerfield 1999). The surplus of widows from war and AIDS has recently convinced both Uganda and Rwanda to allow women to inherit land. Globally, the rise of women-headed households and absentee partners, now exacerbated in postconflict regions, and their critical need for shelter has prompted the Gender Unit of the United Nations Centre for Human Settlements to mount a campaign for security of house tenure for women (UNCHS 1999).

Programs to provide income and housing to women empower them and enhance their household bargaining position (see *Family Bargaining*). The social capital gained through membership in organizations enhances women's community presence and provides them with a voice in elections. Such power addresses gender inequalities and provides ways for women to rise out of poverty.

See also: Boserup, Ester (1910–99); Development, Economics of; East Asian Studies: Gender; Economic Development and Women; Economic Globalization and Gender; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Southeast Asia; Gender and Technology; Latin American Studies: Gender; Near Middle East/North African Studies: Gender; Poverty and Gender in Affluent Nations; Poverty Policy; Poverty, Sociology of; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender; Southeast Asian Studies: Gender

Bibliography

Chambers R 1995 Poverty and livelihoods: whose reality counts? *Environment and Urbanization* **7**(1): 173–204

Lundberg M, Milanovic B 2000 Globalization and inequality: are they linked and how? Working paper. World Bank, Washington, DC

Moser C O N 1996 *Confronting Crisis: A Comparative Study of Household Responses to Poverty and Vulnerability in Four Poor Urban Communities*. Environmentally Sustainable Development Studies and Monographs Series No. 8. World Bank, Washington, DC

Narayan D, Chambers R, Shah M, Petesch P 1999 Global Synthesis: Consultations with the Poor. Discussion Draft, Poverty Group, World Bank, Washington, DC

Pinstrap-Andersen P, Pandya-Lorch R, Rosegrant M W 1999 *World Food Prospects: Critical Issues for the Early Twenty-First Century*. International Food Policy Research Institute, Washington, DC

Renner M, Sheehan M O 2000 *Vital Signs 2000: The Environmental Trends That Are Shaping Our Future*. Worldwatch Institute, Washington, DC

Tinker I 1987 The real rural energy crisis: Women's time. *The Energy Journal* **8**(87): 125–46

Tinker I, Summerfield G (eds.) 1999 *Women's Rights to House and Land: China, Laos, Vietnam*. Lynne Rienner, Boulder, CO

United Nations 1995 *The World's Women 1995: Trends and Statistics*. Prepared by the Statistical Division, ECOSOC, United Nations, New York

United Nations Centre for Human Settlements 1999 *Women's Rights to Land, Housing and Property in Postconflict Situations and During Reconstruction: A Global Overview*. Land Management Series #9. UNCHS, Nairobi, Kenya
United Nations Development Programme 1998 *Human Development Report 1998*. Published annually since 1990 with different emphases. Oxford University Press, New York
United Nations Research Institution for Social Development (UNRISD), United Nations Development Programme (UNDP), Centre for Development Studies, Kerala 1999 *Gender, Poverty and Well-Being: Indicators and Strategies*. Report of the UNRISD, UNDP and CDS International Workshop, November 1997. UNRISD, Geneva
World Bank 2000 *World Development Report 2000/1: Attacking Poverty*. Oxford University Press, New York

I. Tinker

Poverty, Culture of

The culture of poverty concept was developed in the USA during the 1960s primarily through the best-selling ethnographic realist publications of the cultural anthropologist Oscar Lewis, who tape-recorded eloquent life histories of the urban poor. He reprinted numerous versions of his definition of the term 'culture of poverty' in short journal articles and also in the introductions to his books on family life among Mexicans, Puerto Ricans, and Cubans living in shanty towns and ghettos (Lewis 1961, 1966a, 1966b, 1967). Lewis's culture of poverty struck an academic identity politics nerve, and at the turn of the millennium the concept remained enmired in a bitter polemic over how to analyze and engage politically the persistence of poverty in the midst of postindustrial plenty.

1. *Ideological Backdrop to the Culture of Poverty*

In the USA, irrespective of the theoretical orientation of researchers, most discussions on poverty polarize around value judgments concerning individual self-worth or around racial/ethnic stereotypes. US attitudes towards poverty are rooted in the country's colonial Calvinist/Puritanical heritage and are exacerbated by the historical importance of racialized hierarchies that have legitimized genocide, slavery, colonization, and immigration control. This helps explain why the culture of poverty concept continues to generate so much emotional heat while shedding so little conceptual light. The uses and misuses of the concept offer a fascinating case study in the sociology of knowledge illustrating the political interfaces between theory, empiricism, art, and ethnocentric moralizing in the social sciences.

Poverty research throughout history has been more successful at reflecting the biases of an investigator's society than at analyzing the experience of poverty. The state of poverty research in any given country emerges almost as a litmus for gauging contemporary social attitudes toward inequality and marginalization. For example, while Lewis's books are read by a US public as an individualistic interpretation of the persistence of poverty that blames victims, in France his work is interpreted as a critique of society's failure to remedy the injuries of class-based inequality under free market capitalism.

2. *Defining the Culture of Poverty*

The socialist sociologist Michael Harrington was the first prominent academic to use the phrase 'culture of poverty' in a major publication. His book, *The Other America*, documented rural poverty in Appalachia and represented a moral call to action that anticipated the War on Poverty initiated by President Johnson in 1964 (Harrington 1962). As a first-generation son of impoverished Jewish immigrants who was influenced by Marxism in his youth, Lewis shared Harrington's social democratic commitment to combating poverty (Rigdon 1988, Harvey and Reed 1996). Ironically, however, Lewis's popularization of the culture of poverty concept is said to have tolled an intellectual death knell to the optimistic idealism of the mid-1960s that advocated eradicating poverty through direct state intervention (Katz 1989). This is because Lewis's definition of the culture of poverty stressed that a significant minority of the poor (approximately 20 percent) were trapped in self-perpetuating cycles of dysfunctional behaviors and attitudes: 'By the time slum children are age six or seven they have usually absorbed the basic values and attitudes of their subculture and are not psychologically geared to take full advantage of changing conditions or increased opportunities' (Lewis 1965a, p. xlv).

This kind of psychological reductionist and individualistic interpretation of the persistence of poverty resonated with US popular blame-the-victim discourse. Ironically, in the same articles or book introductions in which he defined the culture of poverty, Lewis also included radical political statements contradictory to the implication that poverty is caused by self-perpetuating deficient value systems. For example, in his *Scientific American* version of 'The Culture of Poverty' he quotes Frantz Fanon, praises Castro's Cuba, and criticizes 'free-enterprise, pre-welfare-state stage capitalism' for spawning the culture of poverty (Lewis 1966b). At the same time he states that 'it is much more difficult to undo the culture of poverty than to cure poverty itself,' and advocates 'psychiatric treatment' for poverty in the USA (Lewis 1966b, p. 25).

In other words, the culture of poverty concept was confused theoretically at its inception. Unfortunately, Lewis never managed to clarify what he intended to mean. His published correspondence reveals that he was profoundly disturbed by the blame-the-victim interpretation of the causes of poverty that he triggered in the USA: 'There is nothing in the concept that puts the onus of poverty on the character of the poor' (Lewis 1967, p. 499).

The notion of a culture of poverty, consequently, should not be treated as a full-blown theory. As presented by Lewis, it was merely a bundle of some 70 traits which he did not link to a particular processual or dynamic logic. In fact, he never even listed all 70 of the traits that he claimed existed. The theoretical sloppiness of the culture of poverty concept may well be a product of the McCarthyist anticommunism in US academia that impinged on Lewis during his formative years in the 1950s. Four decades after their inception, his lists of culture of poverty character traits appear embarrassingly arbitrary, ethnocentric, and psychologically reductionist (Lewis 1966a, p. xlviii):

> … a high incidence of maternal deprivation, of orality, of weak ego structure, confusion of sexual identification, a lack of impulse control, a strong present-time orientation with relatively little ability to defer gratification and to plan for the future, a sense of resignation and fatalism … male superiority, and a high tolerance for psychological pathology of all sorts.

3. *The Policy Implications of the Culture of Poverty*

In the late 1960s and 1970s Lewis's culture of poverty concept produced an outpouring of political and academic reactions, primarily in the USA but also in Mexico and Puerto Rico (Leacock 1971, *Alteridades* 1994). Despite conceiving of his work as a call for the expansion of public-sector intervention on behalf of the poor, his concept took the popular intellectual spotlight off the need for structural economic reform, and glossed over the social power dynamics revolving around class, ethnic, gender, and colonial inequalities. Policy makers, if they paid any attention to the culture of poverty concept, interpreted it as advocating the need to rehabilitate the deficient cultural value systems of poor children through the agency of psychiatric social workers. For example, in the applied policy realm Lewis consulted in the development of the Head Start Program in the USA (Rigdon 1988), which has been criticized retrospectively as an attempt to 'take inner-city preschoolers who live in lead-painted, rat-infested tenements without steady heat or hot water, and metamorphose them into bright-eyed, upper-middle-class overachievers' (Bourgois 1995, p. 325). Significantly, in the year 2001, Head Start was still identified by both liberals and conservatives as one of the only successful antipoverty programs of the 1960s.

4. *The Theoretical Implications of the Culture of Poverty*

Unfortunately, most of the hostile academic responses to Lewis's culture of poverty concept have limited themselves to contradicting Lewis's empirical assertions, rather than to critiquing theoretically his psychological reductionism, his sloppy use of the culture concept, and his failure to link in a dynamic manner macrostructural political and economic forces—including gender power relations—to ideology, culture, and individual values (Valentine 1968, Stack 1974; for a political economy exception, see Katz 1989; for a feminist literary criticism exception, see Franco 1989). The bulk of the negative reaction hinges on a political concern for replacing the negative imagery of Lewis's painful but expressive ethnographic portraits of the everyday suffering of urbanized families, with positive images of the worthy poor, struggling for upward mobility against all odds. A late 1990s rehabilitating of the culture of poverty concept from a Marxist perspective dismissed the virulence of the US progressive reaction against the culture of poverty concept as a sectarian 'ultra Bolshevism' that swept the New Left when the general public was drifting ideologically to the Right following the War on Poverty. This precipitated a 'fruitless game of radical one-upmanship' among frustrated intellectuals, who were completely marginal to public political discourse, and who chose instead to devote their energies to proving their dedication to protecting the image of the poor (Harvey and Reed 1996). More importantly, the urgent righteousness of the anti-culture of poverty social science literature is comparable to the polemics against Moynihan's 1967 patriarchal attribution of the 'tangle of pathology' in the black family as being the central cause for the persistence of poverty among urban African-Americans (Rainwater and Yancey 1967).

The angry denial by academics of the existence of the types of violence and self-destructive behaviors described ethnographically by Lewis among the vulnerable families that he tape-recorded and described reveals how far removed intellectuals can be from the inner-city street. Although Lewis's writing deserves criticism for presenting his subjects in a decontextualized pornography of violence, sexuality, and emotional brutality, none of the behaviors or personalities described by Lewis should shock anyone who is familiar with everyday life in the US inner city or Latin-American shanty towns. On the contrary, Lewis's ethnographic realist descriptions, unfortunately, still ring true four decades after they were written. His disturbing material, however, demands

theoretical explanation and political contextualization, and that is where both Lewis and his critics and admirers have largely failed. By confining the debate to a worthy vs. unworthy poor dichotomy, the internecine squabbles between leftist, liberal, and conservative academics mimetically reproduce the right-wing hegemony in popular US culture that equates poverty with sinfulness.

Arguably, the polemics of righteousness that the culture of poverty prompted scared a generation of social scientists away from ethnographic analyses of inner-city poverty in the USA and, to a lesser extent, around the world (Wilson 1987, pp. 13–16). Indeed, accusations of supporting a 'culture of poverty interpretation' are still frequently invoked in polemical identity politics attacks between academics over any representation of poverty that is not flattering to poverty's victims (cf. Lassalle and O'Dougherty 1997). Hence the virulence of the 'underclass debate' in sociology spawned by William Julius Wilson's book *The Truly Disadvantaged* during the late 1980s through the 1990s (Katz 1989, Wacquant 1997).

From a theoretical perspective, the legacy of the culture of poverty debate has impoverished research in the social sciences on the phenomenon of social suffering, everyday violence, and the intimate experience of structural oppression in industrialized nations. Most importantly, by remaining mired in debates driven by identity politics, researchers have minimized the painful experience of day-to-day survival among the persistently poor. Epidemiological data on the associations between social class interpersonal violence, domestic violence, health outcomes, education outcomes, substance abuse, etc. are simply ignored by most poverty researchers in the USA.

The vacuum of critical intellectual engagement with the phenomenological experience of poverty has allowed right-wing academics subscribing to facile neoliberal blame-the-victim interpretations to capture popular imagination and policy debates—especially in the USA. Social or cultural reproduction theory, which emerged out of studies of poor youth at the intersection of the disciplines of education, sociology, and anthropology during the 1980s and early 1990s, offered a critical theoretical alternative. By focusing on the power dynamics of the interface between culture and social inequality, social/cultural reproduction theorists address the empirical reality of the existence of patterns of interpersonal self-destruction without obscuring structural political forces. Although vulnerable to critique for being overly functionalist, these theories allow for the reinscription of agency among the poor, as well as an autonomous role for culture in political economy (Bourdieu and Passeron 1977, Willis 1981).

Few serious social science researchers, if any, at the dawn of the twenty-first century would dare utilize the term 'culture of poverty' in their analysis, despite the fact that their empirical and theoretical work addresses cultural expressions of social suffering due to the entrenchment of urban poverty in industrialized and postindustrialized societies. The problematic analytical and political utility of the culture of poverty concept demonstrates how dangerously essentializing the phrase 'culture of …' can become with respect to any concept. Indeed, anthropologists cannot agree upon a useful definition for culture; nor do they understand how it operates without turning it into a black box of totalizing essences (Gupta and Ferguson 1997).

The culture of poverty furore reminds us that academics fight so hard over so little especially when marginalized political perspectives are at stake. At the turn of the millennium, much of the world's population survives precariously in shanty towns, housing projects, tenements, and homeless encampments where mind-numbing, bone-crushing experiences of poverty engulf the socially vulnerable. Meanwhile, concerned academics continue to fiddle in their ivory towers, arguing over how to talk correctly about the structural violence of poverty.

See also: Class: Social; Conflict Sociology; Income Distribution; Inequality; Poverty: Measurement and Analysis; Poverty Policy; Poverty, Sociology of; Racism, History of; Racism, Sociology of; Underclass; Urban Poverty in Neighborhoods; Wealth Distribution

Bibliography

Alteridades 1994 Oscar Lewis Revisitado. *Alteridades* **4** Special Issue

Bourdieu P, Passeron J-C 1977 *Reproduction in Education, Society, and Culture*. Sage Publications, London

Bourgois P I 1995 *In Search of Respect: Selling Crack in El Barrio*. Cambridge University Press, New York

Franco J 1989 *Plotting Women: Gender and Representation in Mexico*. Columbia University Press, New York

Gupta A, Ferguson J (eds.) 1997 *Culture, Power, Place: Explorations in Critical Anthropology*. Duke University Press, Durham, NC

Harrington M 1962 *The Other Poverty: Poverty in the United States*. Macmillan, New York

Harvey D L, Reed M H 1996 The culture of poverty: an ideological analysis. *Sociological Perspectives* **39**: 465–95

Katz M B 1989 *The Undeserving Poor: From the War on Poverty to the War on Welfare*. Pantheon Books, New York

Lassalle Y, O'Dougherty M 1997 Death without weeping: The violence of everyday life in Brazil. *Radical History Review* **69**: 243–60

Leacock E 1971 *The Culture of Poverty: A Critique*. Simon and Schuster, New York

Lewis O 1961 *The Children of Sanchez: Autobiography of a Mexican Family*. Random House, New York

Lewis O 1966a *La Vida: A Puerto Rican Family in the Culture of Poverty—San Juan and New York*. Random House, New York

Lewis O 1966b The culture of poverty. *Scientific American* **215**: 19–25

Lewis O 1967 The children of Sanchez—Autobiography of a Mexican family. *Current Anthropology* **8**: 480–500
Rainwater L, Yancey W L (eds.) 1967 *The Moynihan Report and the Politics of Controversy*. MIT Press, Cambridge, MA
Rigdon S M 1988 *The Culture Façade: Art, Science, and Politics in the Work of Oscar Lewis*. University of Illinois Press, Urbana, IL
Stack C B 1974 *All Our Kin: Strategies for Survival in a Black Community*. Harper and Row, New York
Valentine C A 1968 *Culture and Poverty: Critique and Counterproposals*. University of Chicago Press, Chicago
Wacquant J D 1997 Three pernicious premises in the study of the American ghetto. *International Journal of Urban and Regional Research* **21**(2): 341–53
Willis P 1981 *Learning to Labor*. Columbia University Press, New York
Wilson W J 1987 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago

P. Bourgois

Poverty in History

1. *The Concept of Poverty*

There is no universal concept of poverty which can be applied to all cultures, societies, or times. Nowadays, poverty is usually defined in relation to the medium-income level of a society, and the poverty line is drawn where a family has less than 50 percent of this medium income available. Such seemingly 'exact' definitions are not possible before general income statistics were generated by statistical offices in the twentieth century (with forerunners in some of the most developed countries in the nineteenth century, and even earlier in some cities).

Some researchers have sought to define an absolute level of poverty by drawing the borderline at the minimum input of calories which a person needs to survive. But even this definition cannot be used as a universal measure since this minimum differs with age, sex, climate, the individual body, health, and the kind of work a person does. Moreover, poverty certainly begins before a person dies of hunger. Therefore, 'poverty can be defined objectively and applied consistently only in terms of the concept of relative deprivation' (Townsend 1979, p. 31). But what is relative deprivation?

2. *The Size of Poverty*

If we were to take present standards of living in advanced societies as a yardstick, most people, probably more than 90 percent, in all cultures and ages, would have to be defined as paupers. Only rulers, their immediate entourage, and some large landholders and merchants were free of poverty. Most people working in handicrafts or on the land whom we would term today as the middle classes lived at least at the borderline of poverty. When the harvest failed, epidemics arrived, or war devastated a region—and this occurred frequently even in Europe up to the seventeenth and eighteenth centuries—they could be driven into misery. In other regions of the world, where periods of drought or torrential rains recur regularly, it is even today difficult for the majority of people to work themselves out of poverty.

For Western Europe we have sporadic estimates of the size of poverty for single cities and smaller regions since the later middle ages, and by Gregory King for a whole country, England, in 1688. They vary considerably, partly because the criteria used often remained vague and depended on the prejudice and the purpose of the author or institution taking the numbers. In some relatively well-to-do cities like Toledo, Lyons, Verona, Augsburg, Amsterdam, and Norwich in the sixteenth and seventeenth centuries, the poor numbered between 5 and 10 percent of the population in 'normal' times, rising to between 20 and 30 percent in times of wars, epidemics, or harvest failures. Similar figures have been found for some rural hinterlands of such cities, e.g., around Nuremberg or Zurich. (Jütte 1994, p. 53).

According to Gregory King, however, paupers constituted the majority of the population of England at the end of the seventeenth century. While half a million families (comprising about 2.7 million persons) increased the wealth of the country, about 850,000 families (comprising more than 2.8 million persons) decreased it by depending on others for their livelihood (Mathias 1983, p. 24). Other contemporaries give, however, different figures.

In eighteenth century France estimates of the size of poverty vary according to the definition between one-third and half of the population. When, during the French Revolution, the definition of poverty was sharpened by excluding beggars, vagabonds, and others, 'unwilling to work' the Comité de Mendicité stated that 39 percent of the working men were not able to support a family of five persons (Hufton 1974, p. 22).

One can distinguish at least three different measures used by officials in Western Europe in early modern times to count their paupers. The first was to count the recipients of welfare in one way or the other. This depended, of course, on the welfare policy applied at a particular location. Known figures vary from 2 percent for the city of Berlin in 1665 (but 7.2 percent in 1799) to 14.8 percent in Trier in 1623. But it is also known that in nearly all cities, the majority of people who applied for welfare were refuted. (In Amsterdam in 1799 37,500 received help, while 81,080 were refuted.) The second measure were censuses of the poor. These were rare, and the few available figures vary between

9.5 percent in Valladolid, Spain, in 1561 and 20–25 percent in Huddersfield, England, in 1622. The third method is to count the 'fiscal poor': the people (mostly heads of households) exempted from taxes because of poverty. Here we find a much higher level. The lowest available figure is 4.7 percent for Cologne in 1582 while the highest was 77 percent for Verona in 1635. Most other cities record shares of the tax-exempt households between 35 and 50 percent (Jütte 1994, pp. 46–56).

Taking into account that most of these figures stem from cities in the wealthier parts of Europe, it is easy to imagine the degree of poverty in other regions, particular the notoriously poorer mountainous regions, or the countryside in Eastern Europe, not to speak of the less developed continents.

3. *Who Are the Paupers?*

Certain groups of paupers can be found in all cultures and centuries, for instance, the bodily or mentally handicapped who are not able to work, and father- or motherless children (except those who live in extended families). Very often mothers with fatherless children are being reported amongst the poor. Finally, elderly people, particularly widows, frequently fell into poverty, at least before old age pensions were introduced in the late nineteenth and twentieth centuries. Even then they rarely received pensions high enough to support them. These people can be called 'structural poor.' They probably formed the 5–10 percent of paupers registered even in well-to-do cities and regions. In addition are those who have become temporarily unable to work because of sickness or injury, particularly during wars or natural disasters. One could term them 'cyclical poor.' Their numbers could swell easily to more than half of the population if events like epidemics, crop failure, or war devastated, a region. At such times usually all those who lived on the margin of subsistence—until very recently the majority of the population—fell into deprivation.

All these groups have been seen since the Middle Ages by both worldly and religious authorities as 'deserving poor,' to help whom was a task for church, local or central governments, and private charity. They are to be distinguished from the 'undeserving poor': able-bodied idlers, beggars, vagrants, vagabonds, and prostitutes, who had to be educated or penalized, being expelled or—since the sixteenth century in Europe—put into workhouses or prisons (often one and the same).

For modern observers the biggest group are, however, the 'working poor': small peasants or handicrafts-men day laborers in cities as well as in the countryside, and the unemployed. Nor should slaves be forgotten. Though not all slaves in ancient Europe or seventeenth to nineteenth century America were poor—some had learned crafts or worked as privileged servants, even as artists or managers—the great majority certainly were, and slavery existed also in other cultures and in some European countries as late as the seventeenth and eighteenth centuries.

4. *The Causes of Poverty*

For analytical purposes one can distinguish between different causes of poverty but one has to be aware that in reality they often were intertwined. Jütte (1994) distinguishes between accidental, cyclical, and structural causes. Accidental causes are, for instance, sickness, which counted according to the English Poor Law Commission in 1909 for 'at least one-half of the total cost of pauperism' (p. 21). In early modern Europe plague, influenza, typhus, typhoid fever, or smallpox did not only kill large segments of the population but also impoverished many of those who had survived. Warfare is another 'accidental' cause. Bad harvests, as well as periods of accelerated population growth and structural changes in the economy, Jütte counts as 'cyclical,' while he sees 'structural' reasons mainly in demographic patterns: persistent large families with many children and a non-secured old age. One would have to add seasonal causes to this typology. In northern countries each winter produced unemployment and poverty.

Another analytic scheme could be to distinguish between natural, demographic, sociopolitical, and economic causes of poverty. In all pre-industrial societies man was subject to natural conditions and disasters: drought or flooding, fire, unbearable heat in summer, or ice-cold storms in the winter time could stop his activities, and the natural fertility of his land limited his productivity. Richer or poorer soil, the availability of transport routes on rivers, lakes, and oceans, the availability of raw materials, like wood, copper, or iron ores determined the results of his labor. Demographic factors played an important role—as long as productivity could not be raised markedly, population growth could lead to increased poverty. Sociopolitical factors influenced the degree of and the distribution of wealth and poverty in all cultures, societies, and ages. Centralized or decentralized power, the strictness or looseness of social and political hierarchies, the degree and distribution of freedom, serfdom, or slavery amongst the population influenced also the living standards. Where geographic mobility was restricted, as in the case of slaves and serfs, people could rarely evade poverty by moving to better places. Where upward social mobility was restricted even hard work could help little to improve the lot of the lower strata of society. Here India with her caste system is probably the worst case, but Chinese and Eastern European peasants also fared worse than Western European ones, or Europeans in their new

settlements overseas. Economic conditions are closely connected to the political and social order. Where market conditions could develop a growing minority at least had the chance to escape poverty. But economic development did not occur harmoniously and evenly. The rise of one sector or region could cause impoverishment in another. When Dutch and English woolen products flourished in northwestern Europe in the sixteenth and seventeenth centuries, Italian cities suffered from the new competition. When Indian cotton products appeared on the European markets, the European woolen and linen producers had a difficult time, and when the British and Western European textile industries mechanized their production in the late eighteenth and nineteenth centuries Indian calico producers lost out.

Since the Industrial Revolution numerous such displacements took place and led to regional and branch- or industry-wide pauperization. Most, but not all, were temporary and could be stopped by adjustments to new technologies or other activities. Social and geographical mobility helped to absorb such shocks. There was, however, no overall impoverishment connected with industrialization. On the contrary, rather, the growth of productivity connected with mechanization and rationalization improved even the lot of the poorer segments of the population in the long run. The social problems at the beginning were caused rather by the population growth, which accelerated noticeably from the later eighteenth century. Without industrialization and the growth in productivity in agriculture too, Malthus' law that population grows faster than productivity would not have become obsolete.

There has been a long discussion amongst economic historians of Britain about whether the standard of living of the working classes during the Industrial Revolution improved or deteriorated. There is agreement that it improved from the 1830s when productivity grew recognizably and regularly and the growth of industry created more better-paid jobs. The pessimistic school maintained, however, that between about 1780 and 1830 the living standard of the workers deteriorated. This discussion suffers from a lack of adequate statistics, particularly of consumer prices and regional differences. As always when structural changes occur, there were winners and losers. The winners were skilled laborers and those who could move to better-paid jobs in growing regions; the losers were the immobile, unskilled people in areas not enjoying growth in industry, trade, and services (Fischer and Bajor 1967, and an ongoing discussion in the *Economic History Review*).

The discussion was first aired amongst British social reformers and Members of Parliament in the 1820s and 1830s, culminating in F. Engels' famous book on the situation of the working classes in England (1845) in which he painted an unrealistic picture of the idyllic life of the workers in the putting-out industry and contrasted it with that of the industrial workforce. It was also influenced by the last grave subsistence crisis which occurred in parts of Western Europe during the second half of the 1840s. Ireland in particular, but also regions on the continent like Silesia, was hit by bad harvests including a nearly total loss of potatoes, the poor man's main food. Millions died or emigrated. The not-yet-mechanized linen industry could not compete with mechanized cotton industry. Pauperism seemed on the rise again, but this was a cyclical rather than a structural phenomenon. Since the middle of the nineteenth century living standards of workers improved in most countries of Western Europe and North America without, however, erasing poverty. In many countries serious empirical work on the size and causes of poverty began only now, culminating (amongst others) in the seventeen volumes Charles Booth published between 1889 and 1897 about *Life and Labour of the People of London*.

5. *Welfare Policies*

There is agreement amongst students of poverty that (modern) welfare policy began only in the nineteenth century, with the onset of social security schemes. But to care for the poor was seen as a moral obligation in all major religions and it was institutionalized at very early stages of history. Compassion was, however, only one reason for this: the other was fear, for paupers were regarded as dangerous. This twin motivation runs right through the centuries. The second motive, fear, seems to have dominated public policy in the Roman Empire, but early Christians insisted that *caritas* was essential. The Early Fathers of the Church, like St. Augustine, preached that poverty was a virtue and some founders of Christian orders, as in the case of St. Francis of Assisi, obliged their brethren to adhere to it. The Church and its monasteries were thus the main bodies responding to the fate of the poor during most of the Middle Ages, though not the only ones. Self-help played a role not only amongst families and neighborhoods but also in the craftsmen's guilds, which cared for their sick people and sometimes for the widows and elderly masters. Brotherhoods existed not only for journeymen but sometimes also for unskilled laborers, and in the later Middle Ages in some cities such as Frankfort, Strasbourg, or Zurich even for 'the blind and the lame' (Fischer 1982, p. 29). In other countries, such as Spain, confraternities which enrolled men and, in some cases, women from different occupations, offered mutual aid alongside spiritual welfare to their members. Christian belief convinced wealthier citizens to secure their salvation by giving alms, establishing foundations for the poor or leaving parts of their fortunes to institutions such as hospitals.

Hospital pauperum already constituted small centers of welfare in the early Middle Ages. Often they cared for twelve persons, symbolizing the twelve apostles. In

the twelfth and thirteenth centuries many such hospitals were founded, not only by bishops and monasteries but also by kings, princes, merchants, guilds, and cities. About 600 of such hospitals existed in England in the middle of the fourteenth century. Some of them specialized, for example, in the care of children or of old people, or of those with certain maladies like leprosy. Often they supplied only minimal help, not even food was necessarily provided, and their inhabitants received permission to beg. Medical care was an exception since very few physicians were available. Most of the hospitals in the smaller towns remained all-round institutions for helpless people, but a few housed several hundred people.

In the later Middle Ages cities took over some such hospitals or founded new ones. They installed poor relief administrators, who began investigations of individual cases of poverty and made surveys for their towns or districts. Poor relief was to be made more efficient and rational; pure almsgiving was no longer enough. The 'deserving poor' were to be separated from the 'undeserving' ones. Orderly procedures were being introduced in administering poor relief. This was particularly marked in northern and central Europe while in France, Spain, and Italy worldly authorities remained more reluctant to take poor relief out of the hands of the Church.

This rationalization and bureaucratization began well before the Reformation, and afterwards Catholic and Protestant regimes did not necessarily follow different policies—both, for instance, tried to educate the poor, particularly the children, though with meager results. But there were differences as to the centralization or decentralization of such policies. Mainly Protestant citics centralized their institutions into a 'common chest' to exercise better control. In the German Empire the Imperial Diet of 1497 declared that each community should take measures to care for its own poor, and this principle was enacted in the Imperial Police Ordinance of 1530 which remained in force until the end of the Empire in 1806. Switzerland, The Netherlands, Britain, and the Scandinavian countries followed similar paths, while in the Mediterranean countries the many decentralized institutions remained more or less undisturbed.

While Britain is seen as the pioneer of a general poor tax in 1572, the municipality of Paris had introduced such a tax already in 1551 after the French King had given it the authority 'to supervise and conduct all things required for the treatment and the relief of the poor' (Jütte 1994, p. 119). But Britain pioneered early modern poor regulations in several respects. Already in the middle of the fourteenth century Parliament ordered that the working poor should be distinguished from beggars, and in 1388 tried to restrict the mobility of beggars. From the 1530s to around 1600 numerous statutes were enacted, of which those of 1572 and 1596 became known as the 'Old Poor Laws,' which remained in force until 1834. The Judges of the Peace had to select four men of substance to act as overseers of the poor. These men had to raise the poor tax, distribute help, and take care of the education of orphans. They were supposed to punish beggars and vagabonds, and could establish houses for correction. As in other countries, the result of their endeavors was very uneven. The poor taxes always fell short of the needs. It has been estimated that only 7 percent of the expenditure for the poor originated from taxes—by far the greater part came from private donations and charitable trusts which were established by wealthy citizens. One rich Londoner alone set up institutions for the poor in 216 communities (Fischer 1982, p. 43f.). Such trusts also ran hospitals and schools.

In the minds of the early modern poor administrators and the wider public, education to work became more important than poor relief. Therefore, workhouses were established in the major cities and some smaller towns. The hope was that they would both accustom the poor to work and produce some profit. Usually they failed in both respects. Particularly in The Netherlands, the Scandinavian countries, and in Germany they became rather a combination of workhouse and house of correction, the famous example being the Amsterdam *tuchthuis* founded in 1595. In smaller towns such houses often tried to perform four functions: educating orphans, getting adults to work, penalizing beggars and criminals, and taking care of handicapped or sick people.

The English Poor Law of 1834 was supposed to remedy the ages-old deficiencies in welfare policies. After a Parliamentary Commission had found that the 'Old Poor Law,' by prescribing assistance to every needy person, had actually provided a disincentive to work, the new law stipulated that only such paupers who were willing to live (and work) in poorhouses should get relief. It turned out that there were not enough such institutions to house everybody who did not find work on his own or did not earn enough to survive, particularly with a larger family. The new law did not remove the ancient causes of poverty, neither could it take care of the temporary unemployment of industrial workers, but it revived the discussion about social problems in the country. A whole set of measures had to be taken to improve the lot of the poorer classes: while more employment was to be created by industry, transport business, and services. Parliament regulated child and female labor, and began to enforce safety measures at the workplace, while local communities improved the housing and sanitary conditions. In 1876 about 19 percent of the population were insured by voluntary Friendly Societies, relatively more than in Germany's obligatory health insurance in 1885, but the most needy could not afford the contributions. Private charity remained, therefore, important in Britain. In 1897 Britain introduced obligatory accident insurance, in 1908 old age pensions, and in 1911 a combined health- and unemployment insurance.

Other European countries were faster in creating obligatory accident-, health-, and old age insurance. But unemployment, one of the most important reasons for poverty, was being regarded as non-insurable even at the end of the nineteenth century. In Germany, unemployment insurance came into force only in 1927. Most of these insurances were quite inadequate at the beginning. They covered industrial workers only in some branches, not day-, farm-, and home-workers, thus excluding the poorest part of the labor force. Bismarck's Old Age Insurance provided means only at the age of seventy when most people had died already, and widows received only 20 percent of the earning of their husbands in accident insurance. But the principles of a modern welfare state were established, and the basic institutions for it were created before or shortly after World War I.

See also: Income Distribution; Inequality; Inequality: Comparative Aspects; Poverty, Culture of; Poverty: Measurement and Analysis; Poverty Policy; Poverty, Sociology of; Social Inequality in History (Stratification and Classes); Social Security; Social Stratification; Social Welfare Policy: Comparisons; Wealth Distribution; Welfare; Welfare Programs, Economics of; Welfare State; Welfare State, History of

Bibliography

Abel W 1974 *Massenarmut und Hungerkrisen im vorindustriellen Europa. Versuch einer Synopsis* [*Mass Poverty and Subsistence Crisis in Pre-industrial Europe*]. Paul Parcy, Hamburg and Berlin

Aldcroft D H, Catterall R (eds.) 1996 *Rich Nations—Poor Nations. The Long Run Perspective*. Edward Elgar, Cheltenham, UK

Fischer W 1982 *Armut in der Geschichte. Erscheinungsformen und Lösungsversuche der 'Sozialen Frage' in Europa seit dem Mittelalter* [*Poverty in History*]. Vandenhoeck and Ruprecht, Göttingen, Germany

Fischer W (ed.) 1995 *Lebensstandard und Wirtschaftssysteme* [*Living Standards and Economic Systems*]. Fritz Knapp, Frankfurt, Germany

Fischer W, Bajor G (eds.) 1967 *Die Soziale Frage. Neuere Studien zur Lage der Fabrikarbeiter in den Frühphasen der Industrialisierung* [*The Social Question, Recent Studies about the Situation of Industrial Workers in the Early Phases of Industrialization*]. Koehler, Stuttgart, Germany

Fischer W, van Houtte J A, Kellenbenz H, Mieck I, Vittinghoff F (eds.) 1980–1993 *Handbuch der Europäischen Wirtschafts- und Sozialgeschichte* [*Handbook of European Economic and Social History*]. Klett-Cotta, Stuttgart, Germany, 6 Vols

Flynn M 1989 *Sacred Charity: Confraternities and Social Welfare in Spain, 1400–1700* , London.

Geremek B 1988 *Geschichte der Armut* [*History of Poverty*]. Artemis, Zürich, Switzerland

Gutton J P 1974 *La Société et les pauvres en Europe, 16e–18e siècle* [*Society and the Poor in Europe, 16th–18th Centuries*]. Presses universitaires, Paris

Himmelfarb G 1984 *The Idea of Poverty: England in the Industrial Age ….* London

Hippel W von 1995 *Armut, Unterschichten, Randgruppen in der frühen Neuzeit* [*Poverty, Lower Strata, Marginal Groups in Early Modern Times*]. Oldenbourg, Munich, Germany

Hufton O W 1974 *The Poor of Eighteenth-Century France 1750–1789*. Clarendon, Oxford, UK

Illife J 1987 *The African Poor. A History*. Cambridge University Press, Cambridge, UK

Jütte R 1994 *Poverty and Deviance in Early Modern Europe*. Cambridge University Press, UK

Landes D S 1998 *The Wealth and Poverty of Nations. Why Some Are So Rich and Some So Poor*. Norton, New York

Lindenmeyr A 1996 *Poverty is Not a Vice: Charity, Society and the State in Imperial Russia*. Princeton University Press, Princeton, NJ

Lis C, Soly H 1979 *Poverty and Capitalism in Pre-industrial Europe*. Harvester, Hassocks, UK

Mathias P 1983 *The First Industrial Nation. An Economic History of Britain 1700–1914*, 2nd edn. Methuen, London

Mollat M (ed.) 1974 *Etudes sur l'histoire de la pauvreté (Moyen Age–XVI siècle)* [*Studies in The History of Poverty: Middle Ages to 16th Century*]. Sorbonne, Paris, 2 Vols

Pullan B 1994 *Poverty and Charity, 1400–1700. Europe, Italy and Venice*. Variorum, Aldershot, UK

Riis T (ed.) 1981–1990 *Aspects of Poverty in Early Modern Europe*. Klett-Cotta, Stuttgart, Germany, 3 Vols

Rimlinger G 1971 *Welfare Policy and Industrialisation in Europe, America and Russia*. Wiley, New York

Salama P 1994 *Pauvretés et inegalités dans le tiers monde* [*Poverty and Inequality in the Third World*]. Paris

Thorp R 1998 *Progress, Poverty and Exclusion: An Economic History of Latin America in the 20th Century … …* Washington, DC

Tomaskovic-Devey D 1988 *Poverty and Social Welfare in the United States*. University of Colorado Press, Boulder, CO

Townsend P 1979 *Poverty in the United Kingdom. A Survey of Household Resources and Standards of Living*. Penguin, Harmondsworth, UK

Williamson J G 1991 *Inequality, Poverty and History*. Blackwell, Cambridge, UK

Woolf S 1986 *The Poor in Western Europe in the Eighteenth and Nineteenth Centuries*. Methuen, London

W. Fischer

Poverty Law

'Poverty law' refers to two distinct but related topics. One has to do with the welfare state, or more particularly, that part of the welfare state that seeks to address the needs of the poor. In the United States, as is true of all advanced industrial economies, many programs fit this definition. In this article, the definition will be restricted to income support for single mothers and their children—the program formerly called 'AFDC' (Aid to Families with Dependent Children) now replaced by 'TANF' (Temporary As-

sistance to Needy Families). Although not the largest program, by far, it is the most controversial. This is the program that is referred to as 'welfare,' which has become the code word for the stereotype inner-city, African-American woman, often a substance-abuser, having children to stay on welfare, and breeding a criminal class. 'Welfare' is the word used to describe the 'underclass' in America.

The second and related meaning of 'poverty law' refers to the legal rights revolution that blossomed in the 1960s that focused on welfare and other poverty issues. Although Americans have always resorted to the judicial system to try to further political goals, the modern era can probably be dated from the court victories of the National Association for the Advancement of Colored People (NAACP) and its offshoot, the Legal Defense Fund (LDF, also known as the 'Inc. Fund'). The LDF victory in *Brown vs. Board of Education* (1954), where the US Supreme Court declared school desegregation unconstitutional, ushered in an incredibly active period of law reform litigation brought by a wide variety of legal service, public interest, and private lawyers. In addition to civil rights (for women, other racial and ethnic minorities), consumers, and environmental protection, rights were won for the mentally ill, the disabled, children with learning disabilities, health care patients, criminal defendants, prisoners, gays and lesbians, and so forth. The particular concern of this article is with lawyers who worked on behalf of the poor.

The two parts of 'poverty law'—'welfare' and poverty lawyers—will be discussed together, because their developments were intertwined.

1. *Aid to Single Mothers Prior to the 1950s and 1960s*

Programs to provide income support for poor single mothers were enacted by the states between 1910–20. They were part of the Progressive Era Child Saving Movement. Social reformers, alarmed at the number of immigrant children growing up in the urban slums, expanded the definition of child neglect and removed children from what they considered to be baleful living conditions and sent them to either farms in the mid-West or reformatories. Juvenile courts were established at the turn of the century with jurisdiction over 'delinquent, dependent, and neglected' children. The latter two were considered 'pre-delinquent.' Then, the Child Savers decided that if the home was just poor, but otherwise 'fit and proper,' perhaps the mother could be supported with a grant as an alternative to removing the children. The programs were called 'Aid to Dependent Children,' (ADC) although popularly known as 'Mothers' Pensions.' In most jurisdictions, they were administered by the juvenile or county courts, assisted by local charities, to make sure that the homes and the mothers were 'fit and proper.'

This early history set the pattern for the next 40 years, The programs were locally administered, small, and basically restricted to white widows. Other single mothers—divorced, separated, never married, and of color—were not considered 'fit and proper.' The consequences of exclusion were severe. Those who were excluded were lumped together with the rest of the 'undeserving poor'—childless adults, the long-term unemployed, others who were excluded from other assistance programs. The 'undeserving poor,' including most single mothers, were thus required to support themselves as best they could in the paid labor market. The establishment of the American welfare state, in 1935, changed the pattern for single mothers only slightly. ADC became a grant-in-aid. The Federal Government assumed about half of the costs and set a few eligibility and procedural conditions, but substantive and administrative control remained with the states and the counties.

AFDC (Aid to Families with Dependent Children, 1962) changed radically starting the late 1950s and early 1960s. The previous system of exclusion collapsed and the rolls and costs exploded. In 1950, two million people were on the rolls; in 1960, three million; then 13 million by the end of the 1980s. Costs rose from about $500 million in 1950 to about $23 billion. Most dramatically, however, was the change in the demographic characteristics of the rolls. In streamed the previously excluded—divorced, single, never married, with increasing proportions of nonwhites. There were various reasons for the changes—the massive migration of African Americans out of the rural south to the northern cities; the entry of women into the paid labor force; the severe unemployment in the inner cities, with the accompanying increase in poverty; rising social unrest, disorder, crime, and riots (during the early 1960s, there were riots in 24 cities); the pressure exerted on local governments by the Federal Government's War on Poverty; the welfare rights movement, and the legal rights movement; the courting of innercity voters by the Democratic Party; and the rise of single mother families. This was the era of the Great Society. A variety of federally sponsored local initiatives and activist groups attacked local institutions and practices, including welfare. The federal courts, War on Poverty lawyers, and welfare rights organizations, such as the National Welfare Rights Organization (NWRO), forced open the state AFDC gates. Welfare became a 'right'; social movement groups actively recruited potential eligibles, reaching more than 90 percent in the early 1970s. Welfare was now in 'crisis.'

2. *The War on Poverty; Poverty Lawyers*

The first legal aid society was established in 1876 in New York City to ease the transition of German immigrants. Thereafter, with the prodding of the

American Bar Association, organizations spread serving the various needs of the poor. By the early 1960s, there were 249 offices. However, the individual offices were small and underfunded. The caseload was mostly landlord–tenant, consumer, and family matters, but few divorces, adoption, bankruptcy, and juvenile proceedings. Legal aid offices did not want to get involved in morally controversial issues. The work was individual service cases. There was no attempt at law reform, test case litigation. The first public defender system was established in 1910. The public defender system grew much more slowly than legal aid. In contrast to legal aid which was based on private charity, the public defender depended on public funding.

The big change in legal services for poor came with Kennedy Administration (1961–1963). Legal rights for the unrepresented was an integral part of the New Frontier and then President Johnson's War on Poverty. The overarching strategy was 'community action,' whereby the poor would help themselves (with lots of government support). If the poor were to be empowered, they would need legal help. There was a struggle, however, over the model and control of legal services. Local bar associations pressed for an expansion of their established individual service offices. The alternative model came from the more aggressive, test-case, law-reform activities, primarily, but not exclusively the NAACP and particularly, the LDF. It would be hard to overestimate the influence of the LDF class action, law-reform strategy during this period. Aided by an activist federal judiciary, at the stroke of the judicial pen, so it seemed, legal rights and legitimacy were given to disadvantaged groups. This prospect of significant social change via the legal system had an enormous appeal to activist, young, elite lawyers. The US Supreme Court, led by Chief Justice Earl Warren, and the NAACP litigation seemed the perfect example of what law and lawyers could do. This was a period of enormous social activism, particularly with civil rights. Although there were the usual compromises with the bar associations as to the organization and funding of legal services, the aggressive, law-reform approach became the most important influence in the development of the War on Poverty's Office of Economic Opportunity (OEO) Legal Services program (1965), as well as in foundation support for consumer, environmental, and public interest law. As we shall see, it also became the focus of the conservative attack on Legal Services.

The specific model for Legal Services was the Ford Foundation's Grey Areas program, where community-based legal services was visualized as an integral component of the tools necessary for the militant self-assertion of the poor and the disenfranchised. Legal services were offered for a range of activities—education, employment, welfare, health and mental health, individual service, and community development. The dominant approach was using test case litigation to establish new rights and the rule of law in dealing with government administrative agencies. The most important legal needs of the poor were thought to be those that concerned their relations with public service programs such as welfare and housing.

While many ideas and organizations contributed to the development of Legal Services, three stand out. One was the influential article by Charles Reich (1964), then a professor at Yale Law School, who argued that relationships with government should receive the same status and protection as traditional property rights. He coined the term, the 'new property.' The other was the articles published by Jean and Edgar Cahn (1964, 1970) which presented the case for neighborhood, activist legal services offices. The third was the support for Legal Services by the American Bar Association which became crucial not only in the contests with local bar associations, but also against the rising conservatism specifically directed at Legal Services.

By 1966, OEO was giving US $20 million to 130 local legal services organizations throughout America. The local offices, in turn, were supported by a half dozen backup centers which provided research support for the litigation activities. From the mid-1960s to mid-1970s, Legal Service lawyers and their allies won a significant number of cases in a variety of areas. The US Supreme Court held that welfare could not be denied simply because the mother cohabited with an unrelated male (the 'man-in-the-house' rule); that states could not withhold welfare or provide lower benefits to the poor who migrated from other states. The most significant welfare case was *Goldberg vs. Kelly* (1970) which held that welfare was a statutory entitlement and that before welfare could be terminated, the client had a due process right to a trial-type hearing at the administrative level with the right of judicial review. During this period store-front offices would advise welfare clients of their rights, both statutory and administrative. Aided by lawyers, clients were emboldened to confront the line staff, and, in certain parts of the country, organize mass demonstrations and sit-ins at local offices. Significant victories were won in both federal and state courts—for example, persons were entitled to a due process hearing to contest the repossession of property or eviction from public housing; the mentally disabled had the right to be represented by an attorney in commitment hearings and if confined to an institution, a 'right to treatment.' In addition, significant legislation was also passed—e.g. the Voting Rights Act, the Americans with Disabilities Act, the Education for All Handicapped Children Act, which, among other things, sought to guarantee parent participation in school special education decisions.

Goldberg vs. Kelly was the high-water mark. Legal Services lawyers tried to establish that the right to a basic income was a fundamental right that deserved the same protection as other fundamental rights, such

as speech, religion, and the right to be free from racial discrimination. But the Supreme Court almost immediately took back what it gave in *Goldberg*. The Court held that welfare regulation was similar to other economic and social matters and could be regulated according to a less exacting standard of reasonableness. Subsequent cases applied the 'reasonable regulation' standard in upholding maximum grants regardless of family size, percentage reductions of AFDC grants as compared to grants for the aged, blind, and disabled, lower amounts of school funding in poorer districts, and unannounced caseworker visits as not violating the Constitutional prohibition of unreasonable searches and seizure.

By this time, the reaction to the activism of the 1960s began to take hold. Not surprisingly, Legal Services' successes against government agencies aroused the hostility of state and local political leaders. Legal Services earned the undying enmity of Ronald Reagan, as governor of California (1967–69) when one of the offices successfully blocked his attempts to cut back on welfare and health care for the poor. From time to time, governors attempted to prevent Legal Services offices from bringing suits against government agencies or on behalf of particular clients (e.g., migrant workers).

3. *Change in Welfare*

Reflecting the liberal moment, the first response of the Kennedy Administration to the 'welfare crisis' was social services to strengthen families and help them towards self-support. The 'social service amendments,' however, never provided much money or services. Rolls continued to grow, and the social service approach was abandoned in 1967. Thereafter, the conservative approach took over. The perception was that AFDC, especially at the local level, was out of control. The program was overly complex, administratively impossible, and hence, subject to rampant 'waste, fraud, and abuse.' The first reaction of the conservatives was a major overhaul of administration. Many of the complex rules were abolished (most notably, those providing for special grants to meet special needs).

Caseworkers were downgraded to eligibility technicians (often with little more than a high school education) whose main task was to monitor rules and reduce errors. The Federal Government imposed strict error-rate standards on the states, and the states, in turn, imposed strict standards on local county administration. The various state and local programs became increasingly rationalized and computerized, all designed to check eligibility and financial payments. Grants to families were either reduced or, more commonly, not allowed to keep pace with inflation. Average grants declined by 45 percent between 1970 and 1996.

The second response was to return to first principles. Those previously excluded from AFDC were required to work in the paid labor market; they were part of the 'undeserving' poor. The conservatives argued that they did not lose their negative moral ascription merely because misguided liberals allowed them into the AFDC program. Beginning in 1967, the first federally-imposed work test for able-bodied AFDC recipients was enacted (WIN, the Work-Incentive Program). Liberals vigorously fought WIN, arguing that if nonwelfare mothers were not required to work, then it was unfair and punitive to require welfare mothers to work. The conservatives said that these historically undeserving women were morally different from mothers who were either man-dependent or self-sufficient. Over the next 20 years, there was a standoff. The work programs, both state and federal, never succeeded in the sense of either reducing rolls and costs through employment, or setting the poor to work, or imposing sanctions for failure to comply. It was administratively easier and less costly to simply defer the vast majority of work-eligible welfare recipients.

Both sides remained unhappy. The conservatives continued to attack the 'entitlement state' on the grounds that there are responsibilities as well as rights in the social contract. Then, in the late 1980s, the liberals changed. They now agreed that AFDC mothers should be expected to work. They gave two reasons. Since the majority of nonwelfare mothers were in the paid labor force, it is only reasonable to expect welfare mothers to work. Second, families are better off, both materially and socially, when the adults are gainfully employed. The Family Support Act of 1988 attempted to strengthen the work requirements. In the meantime, states began to experiment with welfare-to-work programs under waiver from the Federal Government.

The welfare 'crisis' continued to boil. President (then candidate) Clinton promised to 'end welfare as we know it.' The 1994 mid-term elections produced a strongly conservative Republican Congress, which enacted the Personal Responsibility and Work Opportunity Act of 1996. AFDC was abolished, replaced with Temporary Assistance to Needy Families (TANF). State authority over welfare, already considerable, was increased by converting the federal entitlement into block grants. The work requirements were strengthened significantly, now enforced by two sets of time limits: recipients can no longer receive aid for more than two consecutive years (states have the option of shorter limits); and there is a cumulative 5-year lifetime limit. States are required to move an increasing percentage of recipients into work or work-related activities; states are required to reduce grants for recipients who refuse to engage in work or work activities. In addition, there are many 'family values' conditions—for example, requiring mothers under 18 to live with their parents or in adult-supervised

settings; requiring cooperation in the establishment of paternity; states can refuse to pay for children conceived while the mother was on welfare; or deny cash assistance for life to persons convicted of a felony (which, in most states, could be possession of a small amount of marijuana). The Act also modifies several other programs which can impact on the wellbeing of the poor—health care, food stamps, disability, and so forth.

The consensus behind the work provisions (as well as the family values) is that welfare encourages dependency with all its attendant ills for both the adults and the children. Tough work requirements, enforced by time limits, will not only reduce welfare costs, but, more importantly, replace the entitlement status and permissiveness of the current system with values of responsibility and self-sufficiency. TANF explicitly rejects education and training in favor of the 'work first' strategy. The argument for this strategy is: (a) there are plenty of jobs for those who want to work; (b) by taking any job, even an entry-level job, and sticking with that job, a person will move up the employment ladder; (c) the problem with welfare recipients is that they do not have the motivation or the incentives to leave welfare and enter the paid labor market; and (d) state programs have shown that recipients can be moved from welfare to work.

Many states began implementing their programs between 1993 and 1995. In the meantime, starting in 1994, AFDC, nationwide, began a sharp decline—from 5 million families to 2.8 million, in 1998. State and Federal politicians, of course, are claiming 'success' for the welfare reforms. Aside from the fact that the declines in the rolls started well before the initiation of the welfare-to-work programs, the evidence from the state programs is that comparatively few recipients entered the labor market as a result of the programs and they earned very little more than the controls. For example, in the most 'successful' program—the standard-bearer for the 'work first' strategy—Riverside, California—at the end of the demonstration, two-thirds of the experimentals were not working and over half never worked during the entire 3-year period.

The reasons for the lack of success of the state programs are that the assumptions behind the 'work first' strategy are totally misconceived as to who welfare recipients are, why they are on welfare, and the characteristics of the low-wage labor market. Despite the impressive job growth, there has been a stagnation, if not decline, in the wages and benefits for the less skilled, less educated worker. Jobs are increasingly contingent or short-term. Most Americans may be better off, but the bottom fifth of the population is worse off; the average annual income of this group is US $8800 (1999), down from US $10,000 in 1997. Despite the stereotype, most welfare recipients are adults with small families (1.9 children on average), and are on welfare for relatively short periods—between 2 and 4 years. They have extensive labor market connections and many combine welfare with work. But work is uncertain. Jobs disappear, child care breaks down, there are family and transportation problems. They return to welfare because they are usually ineligible for unemployment benefits (leaving a job for family reasons is a 'voluntary quit').

The welfare reforms do little, if anything, to meet these problems. Despite the job growth, there is a shortage of jobs for the low-skilled in many inner cities. Welfare recipients are in tough competition. Most lack a high school degree; they have young children; and they often have transportation problems. They have to rely on informal child care, which is often uncertain and still costs between a fifth and a third of monthly income. They do not have health insurance. And many have additional barriers—lack of skills, and networks, poor health, depression, abuse and violence. Although ex-recipients much prefer work to welfare, they are rarely better off financially and still remain in poverty. As the rolls decline, more of the hard-to-employ are left. In the meantime, sanctions are increasing and the two-year time limits on welfare are due. While there is very little data on those who have been turned away from welfare but are not in regular employment, in general, poverty and hardship is increasing for families, with potentially very serious impacts on the children. In 2000, nationwide more than a third of the homeless are families, an increase of more than ten percent since 1985. In the past, AFDC provided at least some financial stability.

4. *A Crippled Legal Services*

When President Reagan took office in 1981, he proposed no funding for Legal Services for seven successive budgets. The program survived, but funds were sharply cut, forcing significant reductions in staff Consequently, groups of cases (e.g., bankruptcies, disability-benefits) had to be dropped by many offices. The Nixon Administration (1969–74) had also tried to abolish the program and distribute the funds to the states. When that plan failed, the compromise was to create an independent corporation, the Legal Services Corporation (LSC) (1974) and, for the first time, restrictions were imposed on the kinds of cases that Legal Services could handle—school desegregation, the draft (the Vietnam War), abortion, post-conviction civil challenges. The back-up centers were eliminated. On the other hand, the Corporation's budget was increased dramatically, from US $71 million in 1975 to US $321.3 million in 1981. In that year, the LSC made annual grants to 323 programs which employed 6218 lawyers in 1425 offices providing legal assistance in over 1.2 million matters. That was the last increase. The 1994 Republican Congress made deep funding cuts (from US $400 million to US $278) as part of a three-year plan to abolish Legal Services altogether,

and added the most severe restrictions to date. Congress prohibited all class actions; eliminated the collection of attorneys' fees; prohibited participation in administrative rulemaking, lobbying, litigation on behalf of prisoners, representing drug-related public housing evictions, representing certain categories of immigrants, and challenging state, welfare-reform laws under any circumstances—even if such laws violate the US Constitution or federal laws.

The restrictions went into effect in August 1996, and Legal Services had to withdraw from more than 600 class-action lawsuits across the country. In 1997, Legal Service programs employed about 3500 attorneys. They were assisted by about 59,000 private attorneys, mostly on a pro bono basis. The program closed 1.2 million cases and provided services for 1.9 million clients. In 1998, the budget for Legal Services was actually US $21 million less than in 1980. The American Bar Association has estimated that only about 20 percent of the legal needs of the poor are being met. Thus, at a time when there is growing inequality and poverty, when more than 40 million Americans lack health care, and when the poor are facing one of the harshest welfare reforms in recent history, they are also being denied effective legal assistance.

In 2001 the US Supreme Court (*Legal Services Corporation vs. Velazquez* (2001)), held that the restriction on legal services lawyers from challenging welfare reform statues or regulations in the course of representing individual clients seeking welfare benefits was unconstituional; perhaps a new chapter on representing the poor will open.

5. *The Welfare State in Western Europe*

In the US, the current welfare reform can be looked at as a continuation of American welfare policy. From the earliest days of the colonies, the central mission was to separate out the 'deserving' poor from the 'undeserving.' The latter were defined primarily in terms of the labor market—are they able-bodied, could they work? If so, then no welfare. The mission of welfare policy was to exclude, rather than include. Excluded were the deviants—the moral deviants—and this was carried out at the local level. In the USA there is no concept of 'social citizenship.'

Europe could not have been more different. In the decades following World War II, until the oil shocks of the 1970s, despite the differences between countries, the basic approach was universalism, inclusion. There were significant protections and supports for workers, and generous benefits for those who were not in the labor market. The basis for the European welfare state was full-employment prosperity, that is, the system was successful as long as most people were not drawing significant benefits, but rather, were working and contributing to the welfare state. It was this prosperity that allowed the nationalist, peak, corporatism bargaining—concertation—between labor and government. Labor would agree to wage restraints in return for jobs and benefits.

By 2000, much had changed. The demographics of the welfare state have changed to an aging population, greatly increasing pension and health care costs. In addition, for most of the Continent, the economies have been sluggish with persistent long-term employment. Increasing proportions of the long-term unemployed are the vulnerable—youth, immigrants, lone mothers. The current dilemma is that with the high rates of persistent, long-term employment, demands are being made on the welfare state. But under the European Union monetary restrictions, neither tax increases nor budget deficits can be used to finance the increased demands.

Thus far, efforts to develop social protections at the European level have not been successful. There is now pretty much of a consensus that negative integration—the removal of trade barriers, the free flow of labor—is far different than positive integration which requires fundamental changes in government structures, decision making processes, claims to legitimacy—and, most important, solidarities across borders. In fact, the present differences among European welfare states are so significant, that most commentators believe that national institutions are too embedded to converge on a common approach.

When the conservatives were in office, the argument was made that the reason for the sluggish economies is that the generous welfare state has made labor too costly, thus preventing job growth. In the global economy, the foundations of the welfare state—a society of steady, well-paying jobs—are being replaced by major changes in both economic structures (low paying, part time work) and demographics (declining proportion of workers as compared to retirees), the disabled, an increasing proportion of women in the paid labor force, the rise in lone parent families, and immigrants. The foundations of the present welfare state—the intact family, male breadwinner—are no longer valid. The 'passive' solutions of unemployment—early retirement and long-term unemployment and disability—are no longer feasible. Pointing to the US and UK experience, conservative politicians and EU officials argue that labor has to become more 'flexible' and labor market protections have to become more 'active'—that is, encourage people to re-enter the labor market by shortening the period of unemployment benefits, requiring training, education, and employment, raising the retirement age, and tightening disability.

The voters rejected these arguments and the Social Democrats have replaced most conservative governments. While promising to protect the welfare state, they are acutely aware of the constraints on budget deficits or raising taxes. Some countries—Sweden, Denmark, and the Netherlands—have been able to

emerge from the recession in the early 1990s, maintain or re-establish concertation between major employers and labor, and protect their systems of social protection. In other countries, the Social Democrats are struggling. There have been some changes—the proportion of unemployed who qualify for benefits has been declining, there is more use of means-tested benefits, the reduction of benefits in some jobs, the abandonment of centralized wage bargaining in some countries. Most countries are trying to replace 'passive' benefits with an 'active' labor market policy. But so far, these changes have been incremental, and the basic welfare states are still intact. It has proven easier to change markets than social welfare states.

While there is still strong social support for basic welfare state programs, the effective, political resistance, it is argued, basically applies to the existing workforce, not the excluded workers and the vulnerable groups. In the meantime, even though transfers may not have been significantly reduced, we see the quiet undermining of the welfare state—the break-up of national collective bargaining, the growing disparity in wages, the spread of low-wage work, and hostility towards immigrants. The deterioration of the labor market will undermine the social welfare provisions. As pointed out, it was the combination of earnings and benefits that accounted for the comparative equality in Europe. In the final analysis, the foundation of families rests on the economic viability of the adults. Two things are happening in the labor market. As noted, large numbers of the vulnerable population are increasingly facing more-or-less permanent exclusion from the labor market. Others will be increasingly forced back into the labor market, but good, permanent jobs will not be there. An effective active labor policy requires both a large public investment and an integrated infrastructure of labor market institutions, a commitment to some basic income equality, together with a strong safety net. As a growing number of people become dependent, the political pressure increases to restrict income support for the able-bodied. Although there is an increase in need-based and conditional benefits, and there seems to be still strong support for retirees and for children, the crucial question is how solid is the support for the able-bodied? At the present time, welfare policy in Western Europe is at an impasse.

See also: Justice, Access to: Legal Representation of the Poor; Lone Mothers in Affluent Nations; Lone Mothers in Nations of the South; Poverty and Child Development; Poverty in History; Poverty: Measurement and Analysis; Poverty Policy; Poverty, Sociology of; Social Welfare Policies and Gender; Social Welfare Policy: Comparisons; Urban Poverty in Neighborhoods; Welfare; Welfare Programs, Economics of; Welfare State; Welfare State, History of

Bibliography

Abramovitz M 1988 *Regulating the Lives of Women: Social Welfare Policy from Colonial Times to the Present*. South End, Boston

Bell W 1965 *Aid to Dependent Children*. Columbia University Press, New York

Davis M 1993 *Brutal Need: Lawyers and the Welfare Rights Movement 1960–1973*. Yale University Press, New Haven, CT

Cahn E, Cahn J 1964 The war on poverty: A civilian perspective. *Yale Law Journal* **73**: 1317–52

Cahn E, Cahn J 1970 Power to the people or the profession? The public interest in public law. *Yale Law Journal* **79**: 1005–48

Ellwood D 1988 *Poor Support: Poverty in the American Family*. Basic Books, New York

Esping-Andersen G 1999 *Social Foundations of Postindustrial Economics*

Garfinkel I, McLanahan S 1986 *Single Mothers and their Children: A New American Dilemma*. Urban Institute Press, Washington, DC

Gordon L 1994 *Pitied but not Entitled: Single Mothers and the History of Welfare, 1890–1935*. Free Press, New York

Handler J, Hasenfeld Y 1997 *We the Poor People: Work, Poverty, and Welfare*. Yale University Press, New Haven, CT

Handler J, Hollingsworth E, Erlanger H 1978 *Lawyers and the Pursuit of Legal Rights*. Academic Press, New York

Katz M 1986 *In the Shadow of the Poorhouse: A Social History of Welfare in America*. Basic Books, New York

Mink G (ed.) 1999 *Whose Welfare?* Cornell University Press, Ithaca, NY

Patterson J 1981 *America's Struggle Against Poverty. 1900–1980*. Harvard University Press, Cambridge, MA

Piven F, Cloward R 1977 *Poor People's Movements: Why they Succeed, How they Fail*. Pantheon, New York

Reich C 1964 The New Property. *Yale Law Journal* **73**: 733–87

Skocpol T 1992 *Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States*. Harvard University Press, Cambridge, MA

Trattner W 1989 *From Poor Law to Welfare State: A History of Social Welfare in America*, 6th edn. Free Press, New York

J. F. Handler

Poverty: Measurement and Analysis

The issue of low income and poverty has concerned social scientists for centuries. In 1776 Adam Smith linked economic poverty directly with the lack of 'not only the commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without' (Smith 1776, Book V, Chap. II, Pt. II, Article 4th). Interest in this problem was a natural accompaniment of efforts to understand why nations experience economic growth. If the economy of a nation grows, it is only natural to wonder how the fruits of this growth will be distributed among people and why some people are excluded from

the benefits of economic prosperity. A primary issue is to define poverty precisely, and then to document its extent consistent with this definition, that is, to count the poor. Having identified the poor population, questions regarding the nature, causes, and cures of poverty follow naturally. The process of measuring poverty and analyzing its causes and consequences has advanced social science research methods and techniques.

1. *The Measurement of Poverty*

The most prominent early effort to measure poverty was undertaken by B. S. Rowntree in 1902, in his attempt to assess the extent of poverty in York, England (Rowntree 1902). Since that time, there have been numerous attempts to measure poverty in a variety of countries or more narrowly defined areas.

Although reducing poverty is a nearly universal goal among both nations and scholars, there is no commonly accepted way of identifying who is poor. Some argue for a multidimensional poverty concept, such as being 'housing poor' or 'health poor' or lacking in social contacts, thought of as social deprivation. Economists tend to prefer a concept of hardship that reflects economic position or economic well-being, somehow measured; that concept has come to dominate the measurement of poverty.

The measurement of economic poverty seeks to identify those families whose economic position, defined in terms of *command over resources*, falls below some minimally acceptable level. In addition to requiring a definition of economic well-being, the measure must specify a *minimum level of well-being* (or 'needs') in terms that are commensurate with 'resources.' Sen (1983) considered the needs standard (or poverty line) to have 'some absolute justification of its own,' a level below which 'one cannot participate adequately in communal activities, or be free of public shame from failure to satisfy conventions' (p. 167). An ideal poverty measure would allow for differentiation according to household size and composition and would enable people to compare the level of access with resources over time.

Having determined the measure of household resources, the needs standard—or poverty line—appropriate for various households allows one to compare the level of household resources with its needs, and thereby to indicate if the family is poor or not poor. When this process is followed for all households in a nation (or some other jurisdiction), one obtains the *poverty rate*, or the percentage of the people (or households) in the jurisdiction who are living in poverty.

Within this economic perspective, a number of economic well-being indicators have some claim to identifying the economic position of households. For example, the official US poverty measure relies on the *annual cash income* of a family, and compares this with some minimum income standard or 'poverty line.' Alternative—and equally legitimate—measures suggest that the level of *annual consumption* better reflects a family's level of living, or that some measure of a family's *capability to be self-reliant* or a family's *own assessment of economic well-being* are superior concepts for measuring the level and composition of poverty. No matter which perspective is adopted, a large number of judgments and choices are necessary for developing a concrete poverty measure. For example, if income is taken to be the best indicator of economic status, is annual, multiyear, or lifetime income the appropriate measure? Should we examine pretax, pretransfer income or income after accounting for taxes and/or transfers? Should in-kind income be counted or not? Each decision implies both a different poverty rate and poverty population and a different set of antipoverty policies. Finally, irrespective of the decisions made, the resulting poverty measure can be either absolute or relative. The indicator is absolute if the definition of needs is fixed, so that the poverty line does not vary with the standard of living. A relative poverty measure reflects changes in the overall standard of living of the society.

1.1 *Absolute Income Poverty Measures*

The most prominent absolute income poverty measure is that of the USA, which is the only industrialized country to have an official poverty measure. See Ruggles (1990) and Citro and Michael (1995) for further discussion of the US definition and other definitions. Since the late 1960s, the nation has reported on the level of national poverty and the composition of the poor population. This indicator seeks to identify those families that lack sufficient annual cash income to meet what is judged to be their annual needs. The needs indicator is a fixed income threshold, constructed originally as a multiple of the food needs of households of different sizes and compositions. This threshold changes only as the price level changes, but not as average incomes in society change.

This and other income poverty indicators have a particular philosophical basis. The implied social objective is that, together, the community's efforts and those of the individual should ensure that some minimal level of living is attained. Although people may experience hardship in many dimensions—education, housing, food, social contacts, security, environmental amenities—only a sufficiently low level of money income matters in determining who is poor. Such a measure also presumes that money income is a good proxy for welfare (or utility), and that a particular year's income is an acceptable indicator of longer-run income.

Aside from taking exception to the social objective that underlies the official measure, most other criticisms of it focus on the adequacy of the annual cash income measure of 'command over resources.' While the current cash income numerator of the poverty ratio may reflect the extent to which the family has cash income available to meet its immediate needs, it indicates little about the level of consumption spending potentially available to the family. For many families, annual income fluctuates substantially over time owing to unemployment, decisions to change jobs, or changes in health. Even as an indicator of a family's ability to meet its immediate needs, the current income measure is flawed—it reflects neither the recipient value of in-kind transfers (e.g., food or medical care subsidies) which support the economic well-being of low-income families, nor the taxes for which the family is liable. Similarly, while current cash income—and hence the official poverty measure—reflects financial flows in the form of interest and dividends from the assets held by individuals, the assets themselves are not counted, nor is the value of leisure (or voluntary nonwork) time reflected in the measure. The arbitrary nature of the denominator of the poverty ratio—the minimum-income-needs indicator—has also been criticized. Given the weak conceptual basis and the crude empirical evidence on which the dollar cut-offs rest, including the equivalence scales designed to account for the income needs of different family sizes and structures, the official US poverty lines are viewed generally as arbitrary.

These weaknesses have stimulated efforts in the US to improve the official poverty measure. In 1995, a panel appointed by the National Research Council issued its report suggesting revisions in the official measure. The poverty thresholds were modified so as to reflect a broader commodity bundle of needs (relative to the food need basis of the current thresholds) and a new set of equivalence scales. Moreover, the measure of resources was revised to include the value of in-kind transfers received by a household and to exclude estimates of the labor and income taxes for which the household is liable. (see Citro and Michael 1995). While alternative versions of this new measure are being published, the new measure has not yet been adopted formally.

1.2 Relative Income Poverty Measures

Use of a relative income poverty measure takes into consideration changes in the overall economy (in terms of wages and prices); it stresses the relationship of the resources of a living unit to those available to other living units in the society.

One prominent relative poverty definition, used extensively by the European Community, considers those with incomes less than one-half of the median income to be in poverty, hence reflecting the view that poverty is only meaningful when compared with overall income or spending levels. Often, a one-half of median income standard is used in the analysis of poverty in the European Community, but more recently, a 60 percent of the median standard has been adopted (see Bradshaw 2001).

In addition to the arbitrariness of the 50—or 60—percent threshold, this measure has other weaknesses. Critics point out the weakness of a relative measure in assessing the efficacy of antipoverty efforts. Moreover, the relative income standard ensures that the poverty threshold will rise most rapidly in periods of economic growth, during which time those at the bottom of the distribution also experience real growth in both earnings and consumption. Hence, even though poor families may perceive themselves as better off during a prosperous period, the poverty rate may not fall, thus overstating the poverty problem. As Ruggles (1990) has stated: 'poverty cannot decline under a relative poverty measure without some change in the shape of the income distribution as a whole.'

1.3 Consumption-based Poverty Measures

A primary criticism of income poverty measures is that the annual income concept on which they rest is a poor indicator of the permanent income (or lifetime resources) of the family unit. Using such a measure, a wealthy family with a well-educated head and substantial assets, but a year of low income, would be classified as 'poor.'

One proposal designed to avoid this problem involves use of measured family consumption to determine poverty status, based on the argument that family consumption is a superior proxy for the family's permanent income, or family command over resources. Slesnick (1993) argues that income measures of poverty are 'severely biased' (p. 2) in that *temporarily* low-income families are regarded as *permanently* needy. The consumption-based poverty indicator proposed by Slesnick uses household real consumption expenditure per equivalent adult (taken to be the quotient of real household consumption and a household-specific cost-of-living index and an equivalence scale) as the indicator of resources.

Slesnick's consumption poverty measure has been criticized on several grounds. One particularly salient criticism concerns the arbitrary and potentially misleading equivalence scales developed by Jorgenson and Slesnick (1987) that are employed. (see Triest 1998). While other consumption-based indicators (see Cutler and Katz 1991) have shown a growing poverty rate over the last two decades, the Slesnick measure suggests that poverty in the USA has been decreasing secularly over that period.

In general, the largest impediment to utilizing a consumption-based index is the difficulty of obtaining complete and accurate family expenditure data, at

least relative to measuring a family's income. Furthermore, consumption may not fully reflect a family's true well-being; it is possible that simple frugality may be mistaken for poverty.

1.4 Subjective Measures of Poverty

An alternative view of poverty is the 'Leyden School' approach, which relies on the subjective responses of individuals to questions that seek their impressions of their own economic position or well-being, relative to some norm. (see Hagenaars 1986, van Praag et al. 1982). As with income and consumption poverty measures, subjective poverty indicators are based on an 'access to resources' concept, but because subjective thresholds change over time, these measures are relative poverty indicators.

Subjective poverty measures are based on survey responses to questions regarding the minimum level of income or consumption that respondents consider to be 'just sufficient' to allow them to live a minimally adequate lifestyle. If families respond that their own level of living exceeds that minimum, it is possible by observing their actual income to infer both a poverty line and a poverty rate. Because subjective measures are based on individual opinions of what constitutes 'minimally adequate' or 'enough to get by,' they require that individual perceptions of these notions reflect the same level of real welfare for all respondents.

A number of subjective measures have been developed and tested, mainly in Europe. Despite rather minor differences in terminology and question phrasing among these measures, they have yielded highly diverse results. The effectiveness of subjective measures is also limited by the nature of the data collection method. Most estimates are based on small sample sizes, yielding large standard errors.

1.5 Inability to be Self-reliant as a Poverty Measure

One additional poverty indicator should be mentioned—one that views those who are without the capability to be self-reliant as the ones with the lowest economic position. This view rests on two considerations, one conceptual and the other practical. The *conceptual reason* is the more basic. This position has its foundations in the writings of Sen (1992, 1997), who argued 'that the basic failure that poverty implies is one of having minimally adequate capabilities,' (1992, p. 111) and hence that 'poverty is better seen in terms of capability failure than in terms of the failure to meet the "basic needs" of specified commodities' (1992, p. 109).

There is also a *policy-related reason* for developing a capability measure of poverty. In recent years there has been renewed civic debate regarding appropriate norms and standards for individual responsibility and behavior, and hence the appropriate role of the state. A prominent viewpoint in this debate has emphasized the merits of individual independence (relative to reliance on government programs), the negative effects of government programs on individual behavior, and the desirability of a smaller economic and social policy role for government. If society is to base policy on the central social goal of 'economic independence,' then it would seem important to identify the size, composition, and growth of the population of citizens who do not have the capability to be independent in a market economy.

An example of such a capability-based poverty measure is based on the concept of a family's net earnings capacity (the amount that the adults in a family could earn if they worked full time, full year, adjusted for the constraints on working at capacity due to health problems or disability and the expenses required if all of a family's working-age adults did work at capacity). If earnings capacity is less than a poverty threshold, the family is defined as 'unable to be self-reliant,' and in poverty (see Haveman and Bershadker 2001). A related measure of family capability is Becker's (1965) concept of 'full income,' which includes both income realized through market work and the value of leisure time.

Measuring 'self-reliant poverty' requires several implicit conventions, norms, and assumptions. For example, such a measure is reliable only if the 'norm' of full-time, full-year work is an accepted, socially determined norm representing the full use of human capital. Of necessity, attribution of poverty status to any particular family requires prediction from statistical estimates rather than values measured in survey data (such as income), and hence may be viewed as less objective than, say, income-based measures.

2. The Analysis of Poverty

Since the mid-1960s and the US declaration of a War on Poverty, social science researchers have addressed a number of issues concerning the nature, causes, and cures of poverty. In the process, substantive additions to economic knowledge and to research methods have been developed.

2.1 Understanding the Determinants of Poverty: Demographics, Discrimination, Human Capital, Immigration, and Macroeconomic Performance

A major objective among social science researchers has been to understand the processes which combine to result in some people failing to attain minimal standards of economic well-being. Although the number of such processes and institutions is very large, and the interactions among them complex, research has been concentrated in several areas.

The underlying *demographic structure* of the population—age, gender, family arrangements, fertility choices—plays a major role in determining whether families or individuals are poor. For example, the share of the population that is elderly, and hence unable to be self-reliant through work and earnings, influences the prevalence of poverty in a society. Research has tracked the poverty status of the older population and evaluated the ability of older people to escape poverty given their own resources and social support. Similarly, changes in the prevalence of single-parent families caused by rising divorce and non-marital childbearing rates are also fundamental determinants of poverty. Families with children but only a single adult tend to be restricted in their ability to be self-reliant from own earnings. Social scientists have studied the trends and determinants of these choices in an effort to better understand the causes of poverty.

If the attitudes of the majority of the population toward minority groups are such as to lead to lower payments for equivalent work or to inferior schooling, housing, or mobility opportunities, poverty rates among minority groups would tend to be high. Given the prevalence of racial minorities among the poor population, the possible causal role of *labor market, education, and housing discrimination* has been an important topic of study, especially by economists. While the existence and extent of racial income gaps had been studied prior to the 1960s, later research delved deeper than previous work in measuring and explaining the sources of these income gaps. See Cain (1986) for a review of this line of analysis. Given that competition would be likely to erode the effects of discriminatory behavior by employers (Becker 1957), research turned to explanations involving 'statistical discrimination' and 'market signaling' as the basis for continued discrimination as a cause of poverty.

The role of *education and training*—human capital—in enabling people to attain self-sufficiency also became a focus of poverty research and analysis. The question of the effect of increases in schooling on individual earnings occupied the efforts of numerous researchers, and education came to be analyzed as a process that created human capital. Estimates of the rates of return on schooling seemed to justify diverting resources from private sector investments in physical capital to public investments in education. Rosen (1987) surveys the literature in this area. While economists focused on the earnings returns from schooling, other social scientists sought to document the effect of schooling on student attainment (such as test scores), social mobility, and economic status. This research documented the complexity of the processes by which some children succeed while others fail, and led to a search for improved data and statistical methods capable of disentangling the effects of education on attainments. This 'human capital' approach to schooling carried over to numerous studies of the effectiveness of governmental programs designed to provide skills and training to adult workers. See Heckman and Lochner (2000) for a review and interpretation of the evidence on the economic effects of education and training.

Concern with the problem of poverty in the USA coincided with a rapid increase in *immigration*, often of individuals with low levels of schooling or language competence. The potential impact of immigration in eroding wage rates for unskilled workers in urban labor markets containing large numbers of immigrants became an important policy and research issue.

The effect of the overall performance of the economy on the poverty rate was also the subject of substantial research and analysis, beginning in the 1960s. During the 1970s, economic growth and low unemployment seemed to be powerful instruments in reducing unemployment. However, during the 1980s and early 1990s, economic growth seemed to have lost its antipoverty effectiveness. Evidence suggests that during the 1990s, the level of poverty was again responding to macroeconomic performance. See Haveman and Schwabish (2000) for a review of this research and for recent evidence.

2.2 Understanding Social Mobility, Status Attainment, and Income Dynamics

Research on the causes of poverty have tended to focus on static issues—factors that appear to be causally related to low economic position. Simultaneously, a complementary line of research addressed questions of the dynamic processes by which people either succeed or fail to attain economic position—processes that indicate the openness or fairness of society. While research on these social mobility processes had been undertaken prior to the 1960s, those studies were limited by the absence of both a conceptual framework and data on individuals over time—longitudinal data—necessary to trace these dynamic patterns. The early 1960s witnessed a number of important research contributions designed to increase understanding of the process of social mobility. One of the most central of these presented a causal model of the socioeconomic life cycle, distinguishing the interactions among three primary life phases: family, schooling, and work (see Duncan and Hodge 1963). The socioeconomic status of the job was taken to be the ultimate indicator of success, with family back-ground and schooling (as an intervening variable) viewed as the determinants of that status. This framework stimulated several empirical studies of these relationships, yielding insights into changing intergenerational correlations involving occupations and income, changes in the degree of social mobility, changes in the role of education in fostering mobility, and changes in mobility patterns between whites and nonwhites. A variant of this approach sought to

identify the factors that contribute to social or economic status, distinguishing those that relate to exogenous family and background factors from those amenable to policy influence. Two prominent examples of this approach are Jencks et al. (1972, 1979). Family characteristics and choices and intellectual inheritance were seen as playing important roles, but school quality and the racial mix of neighbors and peers were found to be less important. Haveman (1987) reviews and critiques this research.

While these analyses were revealing, they demonstrated the difficulty of making progress in this area without detailed data on individuals and families that tracked their attributes and performance over time. The Panel Study of Income Dynamics, the National Longitudinal Surveys, and the Survey of Income and Program Participation are among the most prominent and high quality of these data collection efforts. Extensive use of them has enabled major advances in understanding the nature and causes of poverty, including the lengths (or persistence) of poverty and welfare stays among the poor, trends in income trajectories among families with various characteristics, the identification of the long-term or permanently poor population, and estimates of the relative roles of family background, schooling choices, neighborhood characteristics, and labor market discrimination in determining the probability of being poor. A complementary field of study has used these data to explore the determinants of these family characteristics (e.g., poverty-level incomes, single-parent status, low parental education, welfare recipiency), schooling quality, neighborhood characteristics, and own schooling and fertility choices on a variety of children's attainments (see Haveman and Wolfe 1995, Duncan and Brooks-Gunn 1997).

2.3 *Antipoverty Effects of Policy Interventions*

Poverty researchers have also sought to describe the various policy measures in place to assist poor families and to identify changes in or alternatives to these policies that might be effective in reducing poverty. Probably because the main antipoverty efforts were in the area of income support measures, analysts concentrated their assessments on such policies. A voluminous literature describing the nature of income support policies and critiquing their incentive (efficiency) and equity characteristics has influenced policy changes over the period since the 1960s. These analyses revealed low 'take-up' rates among people eligible for support, horizontally uneven support patterns among equally poor people, the uncertain rationale for and implications of providing support via in-kind benefits vs. tax subsidies vs. cash transfers, the relative effectiveness of alternative programs in reducing poverty rates or the intensity of poverty, and the presence of incentives for undesirable behaviors (e.g., reductions in work, increases in family breakups, and nonmarital childbearing) implicit in the structure or administration of the programs.

These descriptive and policy analysis studies played an important role in guiding numerous antipoverty policy reform efforts that have occurred over the past several decades, including proposals for 'negative income tax' plans, work-related subsidies, the substitution of in-kind for cash benefits, and 'welfare-to-work' initiatives. They also provided a catalog of important questions for which answers would be relevant in designing effective policy: how large are work effort reductions in response to income transfers? Has public income support reduced family stability, increased unemployment, reduced savings, or increased nonmarital childbearing? (see Haveman 1987).

2.4 *Labor Supply, Family, and Fertility Effects of Income Support Policy*

Concerns with the work incentive effects of income support policy have dominated debates over this antipoverty strategy. Numerous analysts responded to this concern with empirical studies documenting the extent of these behavioral responses to public income support programs. These studies, employing a variety of survey data and empirical methods, found that public benefits tend to reduce employment probabilities, lower work hours per year if working, and possibly reduce wage rates among both recipients and potential recipients of benefits.

Most of the early studies employed cross-sectional survey data in order to measure the responses of individuals to the increased income from transfer benefits and the reduced effective wage rate associated with working. Later studies relied on the results from carefully designed 'social experimental' studies of these potential responses (see below). The essential characteristic of these studies is the random assignment of families to alternative forms of transfer program designs (in terms of the generosity of income guarantees and the extent of 'benefit reduction' rates), and the observation of the behavior and performance of these families over subsequent years. Both forms of studies confirmed that, as expected, work effort falls as nonlabor income rises and as the effective wage rate falls. While the estimates varied widely, nearly all of them raised serious questions about the efficiency and effectiveness of this income support approach to helping poor people.

While these studies focused on the work effort responses of people to the availability of income support, other analyses (including randomized approaches) tested the family structure, childbearing, health care utilization, housing quality, and welfare utilization choices of low-income families in response to both welfare-type policies and reform efforts in the educational, housing, health insurance, and public employment areas. Danziger et al. (1981) and Moffitt (1992) have surveyed this research.

3. Poverty Research and Social Science Methods

In addition to adding to the body of knowledge regarding the nature, causes, and cures of poverty, the growth of poverty-related research since the 1960s has also contributed to the development of new research methods of general applicability in the social sciences (Haveman 1987). Four contributions deserve mention.

The first contribution consists of a new field of evaluation analysis and research. As analysts, in response to the demands of policy makers, focused on the poverty problem, an emerging analytical area, evaluation research, was also being spawned. Interest in this applied research approach extended far beyond the poverty issue, but research on social and antipoverty policy issues provided this then-nascent subdiscipline with some of its basic methods and case studies. Without the emphasis of policy makers on the need to evaluate the effectiveness of the numerous new antipoverty initiatives, it is unlikely that this field of analysis, with its journals and professional organizations, would have developed as rapidly as it has, and in the same directions.

Second, as noted above, randomized social experimental research was first applied to proposed policy interventions in the poverty area—the negative income tax proposals, school and housing vouchers, employment and training measures, and so on. Without the emphasis on the effect of antipoverty policy interventions on family economic status during the 1960s and 1970s, it is unlikely that this important social science research method would have attained the status it now holds among social science researchers.

One of the problems that analysts face in assessing the effects of policy and other changes on the behavior of people is the need to rely on samples of observations that may be composed of individuals who are selectively included in the sample, and hence unrepresentative of the larger population for which inferences are desired. This concern dominated the study of labor markets well before demands to assess the impacts of antipoverty measures, especially in the employment and training area, where factors that determined the presence of a person in a research sample were also determinants of the effect that was being studied (see Ashenfelter 1978). Several approaches to this problem were developed by econometricians (see especially Heckman 1979), and these statistical tools became a standard part of applied statistical analyses.

Finally, in part because of the development of large representative samples of national populations, analysts began developing computer models designed to simulate the effects on the income, well-being, and behavior of each observation in a data set of actual or proposed policy measures. The application of such microdata simulation models found its most welcomed audience among analysts concerned with the effects of changes in taxation and income transfer policy on low-income families. Without the impetus provided by the need to assess the impacts on poor people of proposed policies, this advance to predictive model building would not have occurred. This research methodology is assessed in Citro and Hanushek (1991).

See also: Education: Skill Training; Human Capital: Educational Aspects; Income Distribution: Demographic Aspects; Male Dominance; Mobility: Social; Policy History: State and Economy; Poverty and Gender in Affluent Nations; Poverty, Culture of; Poverty in History; Poverty Policy; Poverty, Sociology of; Social Inequality in History (Stratification and Classes); Social Mobility, History of; Social Stratification; Wealth Distribution

Bibliography

Ashenfelter O 1978 Estimating the effect of training programs on earnings. *Review of Economics and Statistics* **60**: 47–57

Becker G S 1957 *The Economics of Discrimination*. University of Chicago Press, Chicago

Becker G S 1965 A theory of the allocation of time. *Economic Journal* **75**: 493–517

Bradshaw J 2001 The measurement of absolute poverty. In: Schokkaert E (ed.) *Ethics, Poverty and Inequality and Reform in Social Security*. Ashgate, Aldershot, UK

Cain G C 1986 The economic analysis of labor market discrimination: A survey. In: Ashenfelter O, Layard R (eds.) *Handbook of Labor Economics*. North Holland, Amsterdam

Citro C, Hanushek E 1991 *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*. National Academy Press, Washington, DC

Citro C, Michael R 1995 *Measuring Poverty: A New Approach*. National Academy Press, Washington, DC

Cutler D M, Katz L F 1991 Macroeconomic performance and the disadvantaged. *Brookings Papers on Economic Activity* **2**: 1–61

Danziger S, Haveman R, Plotnick R 1981 How income transfer programs affect work, savings, and the income distribution: A critical review. *Journal of Economic Literature* **19**: 975–1028

Duncan G J, Brooks-Gunn J (eds.) 1997 *Consequences of Growing up Poor*. Russell Sage Foundation, New York

Duncan O D, Hodge R W 1963 Education and occupational mobility. *American Journal of Sociology* **68**: 629–44

Hagenaars A J M 1986 *The Perception of Poverty*. North Holland, Amsterdam

Haveman R 1987 *Poverty Policy and Poverty Research*. University of Wisconsin Press, Madison, WI

Haveman R, Bershadker A 2001 The 'inability to be self-reliant' as an indicator of poverty: Trends for the US 1975–1997. *Review of Income and Wealth* **47(3)**

Haveman R, Schwabish J 2000 Has macroeconomic performance regained its antipoverty bite? *Contemporary Economic Policy* **18**(4): 415–27

Haveman R, Wolfe B 1995 The determinants of children's attainments: A review of methods and findings. *Journal of Economic Literature* **33**(4): 1829–78

Heckman J 1979 Sample selection bias as a specification error. *Econometrica* **147**: 153–61

Heckman J, Lochner L 2000 Rethinking education and training policy: Understanding the sources of skill formation in a

modern economy. In: Danziger S, Waldfogel J (eds.) *Securing the Future: Investing in Children from Birth to College*. Russell Sage Foundation, New York
Jencks C, Smith M, Acland H, Bane M J, Cohen D, Gintis H, Heyns B, Michelson S 1972 *Inequality: A Reassessment of the Effect of Family and Schooling in America*. Basic Books, New York
Jencks C, Bartlett S, Corcorau M, Crouse J, Eaglesfield D, Jackson G, McClelland K, Mueser P, Olneck M, Schwartz J, Ward S, Williams J 1979 *Who Gets Ahead? The Determinants of Economic Success in America*. Basic Books, New York
Jorgenson D W, Slesnick D T 1987 Aggregate consumer behavior and household equivalence scales. *Journal of Business and Economic Statistics* **5**: 219–32
Moffitt R 1992 Incentive effects of the US welfare system: A review. *Journal of Economic Literature* **30**(1): 1–61
Rosen S 1987 Human capital. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics*. Palgrave Publishers, Hampshire, UK, Vol. 2, pp. 681–90
Rowntree B S 1902 *Poverty, A Study of Town Life*. Macmillan, London
Ruggles P 1990 *Drawing the Line: Alternative Poverty Measures and Their Implications for Public Policy*. Urban Institute Press, Washington, DC
Sen A 1983 Poor, relatively speaking. *Oxford Economic Papers* **35**: 153–69
Sen A 1992 *Inequality Reexamined*. Russell Sage Foundation and Harvard University Press, Cambridge, MA
Sen A 1997 *On Economic Inequality*. Clarendon Press, Oxford, UK
Slesnick, D T 1993 Gaining ground: Poverty in the postwar United States. *Journal of Political Economy* **101**(1): 1–38
Smith A 1776 *Wealth of Nations*. Modern Library Edition, 1993. Random House, New York
Triest R K 1998 Has poverty gotten worse? *Journal of Economic Perspectives* **12**(1): 97–114
van Praag B M S, Hagenaars A J M, van Weeren H 1982 Poverty in Europe. *Review of Income and Wealth* **28**(3): 345–59

R. H. Haveman

Poverty Policy

Poverty policy or the politics of poverty is a constituent element of social policy. It covers public action that aims to deal with poverty when and if poverty is seen as a public issue. While various forms of poverty or pauperism have always invited some action of close communities, the term poverty policy is reserved for interventions in which the modern state has played a role.

1. *The Political and Intellectual Context*

Poverty policy is informed by public policy in general or social welfare policy in particular. These policies are by their very nature ideologically loaded. The political labels to be applied may be debated. For the present purpose the conventional political categories of conservatism, liberalism, and social democracy appear useful. The ideology based on the supremacy of the market will be termed neoliberalism. Conservatism and neoliberalism may be situated more to the right, liberalism and social democracy more to the left of the political spectrum. By means of these categories an involved pattern will be presented in an oversimplified way to offer an intellectual context for poverty policy.

The political right perceives poverty as an individual rather than a social problem, attributing its causes to the variations in individuals' merits and achievements. One conservative position adds as a main cause of individual failure genetic inheritance that explains for instance the poverty of the intellectually 'inferior' races in America (Herrnstein and Murray 1994). The variation between individuals being natural, poverty itself is seen as a natural and inevitable phenomenon. Its alleviation for altruistic reasons may be approved of, but attempts at its eradication are rejected as futile. Conservatives may prefer voluntary charity as a means of poverty alleviation. Neoliberals may accept state intervention in the form of a well-targeted safety net for the truly needy. The political right usually regards poverty policy as the only acceptable type of social welfare policy.

In a left perspective, poverty is considered as the outcome of social inequalities generated by the economic, political, and social power structure, by the unequal distribution of all the forms of capital. Poverty policy is seen as a must to alleviate immediately the lot of the poor by means of public redistribution. The prevention of the continuous reproduction of poverty is also a major concern, though. This objective requires broader public policies. Public intervention should promote a fairer access to all socially relevant material and symbolic resources, which in turn requires curbing the domination of the market over society. The success of such policies depends on, among other things, the empowerment of people and strong economic and social rights. In this broader policy framework, poverty policy becomes residual.

2. *History*

Poverty policy followed closely the economic and social changes connected with unfolding capitalism in Europe from about the fifteenth century. Four periods may be distinguished, although their contours are inevitably vague.

2.1 *The Beginnings of a Public Policy Concerning the Poor*

Central intervention focusing on the poor was prompted by several factors. The closely knit primary communities—the family, the neighborhood, the parish, the feudal manor, later the guild—endorsed for a long time the care of the weakest members unable to survive without this protection, who would have

otherwise endangered the cohesion or the stable reproduction of the group. The ability of self-help of the small communities was slowly eroded by wars, epidemics, the laicization of society, the loosening of feudal bonds, new forms of mobility, and an increasing number of marginalized vagrants. Most of these processes were connected to urbanization and industrialization. Following these developments, the number of the unsupported poor multiplied while the traditional forms of spontaneous or organized charity were waning. The direct encounter with the misery and the physical suffering of the poor affected uncomfortably their betters, whose sensibility had increased with the Enlightenment (Mollat 1978). More importantly, the poor have become dangerous. They may have caused direct personal danger through unrest, banditism, or subsistence crime. The more indirect danger was that they represented or exuded physical and moral 'evils' (Swaan 1988). The two aspects of increasing poverty—physical suffering and threats to public safety—invited Janus-faced interventions. The politics of poverty combined from the start the functions of policing (control and punishment) and of helping (poverty alleviation) (Geremek 1987). Legal regulations at the local or central level combining these two functions appeared with increasing frequency from the fifteenth and sixteenth centuries onwards (Castel 1995) in most Western countries and even in Eastern Europe. Until the early seventeenth century, the policing element was probably stronger (under the reign of Henry VIII between 1509 and 1547, 72,000 vagrants were executed in England). The supportive function became more explicit in the first major Poor Law of Europe adopted in England in 1601. This was the Elizabeth 43 Act for the Relief of the Poor. In its Section 2 it declared: 'It is agreed and ordered by the present Assembly that each town shall provide carefully for the relief of the poor, to maintain the impotent, and to employ the able…' The French ordinance of Moulin of the same year applies almost identical wording. At this time even if regulation was central, its implementation was the duty of the local authorities in collaboration with the churches and voluntary organizations. Despite the more explicitly prescribed obligation to give assistance, the policing function never disappeared.

Poverty policy was beleaguered from the earliest times by several dilemmas. One of them was the separation of the genuinely helpless from the able-bodied poor. Those belonging to the first group—the sick, the invalid, the frail elderly, sometimes orphans or widows—were usually considered as deserving help. The members of the second category were mostly seen as undeserving of support. It was assumed that they could maintain themselves were they not work-shy or lazy. It was acknowledged only very late—towards the end of the nineteenth century or even later—that the majority of the able-bodied poor would have been willing to work had they had access to a job. Another lasting problem was the separation of the local poor from strangers. The locality was willing to endorse the responsibility for its own poor but was keen to exclude strangers; hence the various acts concerning the duties of provision and the rights to expulsion of the settlements, and also the practice of visible stigmatization of those getting assistance or of those allowed to beg in the locality. These regulations were ubiquitous from Pennsylvania in North America to the town of Györ in Hungary. A third problem centered on the level of assistance. This level was haphazard at least until the wide acceptance of the 'less eligibility' principle defined by the (English) New Poor Law of 1834. The principle implied that 'the pauper's condition should be rendered less eligible, or more unpleasant than that of the poorest paid independent laborer.' The workhouse was then conceived of so as to assure maximum feasible less-eligibility. The principle has also been applied to define the level and the modalities of access to other types of assistance. A further dilemma that lost its edge by now was the choice between so-called indoor and outdoor relief. Indoor relief meant the poorhouse or the workhouse not only in England but all over Europe and America (termed, for instance, Hôtels de Dieux and hôpitaux pour renfermement des mendiants in France) from the sixteenth century onwards. The regulation of poor relief in Elberfeld, Germany, became the model for organized outdoor relief only in the nineteenth century.

2.2 *The Handling of Industrial Poverty*

With the unfolding of capitalism, the character of poverty changed. The number of the urban poor multiplied, and urban destitution acquired tragic dimensions. Early nineteenth century legislation dealing with urban hygiene, or the Factory Acts protecting young children and mothers, were concerned less with utter poverty than with the survival of the whole community. Public assistance continued to play its role often with a strong punitive edge, but it also acquired new dimensions. The workers were affected by new insecurities as compared with the 'pre-industrial' poor. The risks of grave accidents, of unprovided old age, or of job losses all increased, and were slowly recognized as new causes of poverty. Moreover, the workers started to organize and mobilize themselves. The expansion of political rights forced the states to be more responsive to the pressures of vocal political thinkers and of popular movements. From the last third of the nineteenth century progressive taxation, various branches of social insurance, and in a few countries fully or nearly universal flat-rate benefits were added to the instruments of public redistribution. The monopolistic role of poverty policy within social welfare policy was thus ended. From the early twentieth century some countries—Sweden in the first

place—formulated more ambitious projects about the improvement of the working and living conditions of the whole community, and about the abolition of poverty as well as of a separate and autonomous poverty policy.

2.3 The Welfare State and Poverty Policy

During and after World War II, the extension of public responsibility for all-encompassing social security became an accepted, albeit never uncontested, principle in most developed capitalist states. The seminal document founding the idea of the Welfare State was the Beveridge Report, a proposal prepared for the British government published in 1942 under the unassuming title 'Social Insurance and Allied Services.' It proposed to fight by means of universal benefits and services the five 'giant evils,' Want, Disease, Ignorance, Squalor, and Idleness—that is, to end poverty and its root causes. The experiences of World War II, a widespread appeal of left ideals, and the enforced competition with the 'communist' world in Central and Eastern Europe triggered all over Western Europe the development of social protection, increasingly underpinned by labor law and social law.

The welfare state project was never and nowhere fully successful. Yet in its three ascending decades it considerably reduced social inequalities, deep and even shallow poverty, and improved the working, living, and housing conditions and the existential security of the majority. Poverty policy did not disappear, but in most countries its role—and the role of social assistance—have become marginal (Hatzfeld 1971, Esping-Andersen 1990). The situation was different in the USA where universal schemes were never accepted and their function was fulfilled by the targeted schemes devised within the framework of poverty policy. However, from the 1960s on, the War on Poverty and the project of a Great Society improved the welfare programs. For about 15 years they became more generous and better adjusted to needs. The results are contested by many on the right. Yet statistical evidence shows that the poverty rate fell by about half, some inequalities were reduced, and housing, health, and nutrition indicators rose also among those on the assistance rolls.

In Central-Eastern Europe, pre-war poverty policy had many elements in common with the most conservative, more punitive than generous systems of Western Europe. After World War II, the modern systemic differentiations were supplanted by a monolithic system in which a single party's politics dictated the rules of all subsystems such as the economy, law, or social policy. Welfare policy was absorbed by the economy, which was in its turn dominated by politics. Poverty policy did not exist for two more reasons. On the one hand, price policy, full employment policy, broad and almost all-encompassing health, pension, etc., schemes were constructed so as to produce a not too unequal income distribution, and to prevent relatively successfully the emergence of deep poverty. On the other hand, the dominant ideology denied the existence of the poverty that still existed, and therefore did not make any provision to handle it. It had no need and no space for a specific poverty policy, leaving many social ills unattended (Ferge 1986).

2.4 The Withdrawal of the Welfare State and the Revival of Poverty Policy

Following the oil shock and the advent to power of the Reagan and Thatcher administrations, the tide turned. A neoliberal agenda has unfolded requesting less state intervention in the operation of the free market and the retrenchment of the social state. The collapse of totalitarian state socialism bolstered this process.

The new trends affect the traditional institutions of social policy. Universal solutions with access tied to citizenship rights are losing ground. In the case of social insurance, there are efforts to privatize it, to weaken its solidary (nonactuarial) components, and to lower its standards. The efforts to deregulate the labor market have weakened labor rights and increased the number of working poor. The consequences are increasing poverty and unmet needs. The countries in 'transition' are inordinately hit by these changes. The solution offered everywhere is the increased role of poverty policy, particularly of means-tested benefits. The political support of programs targeted only to the poor may be so weak, however, as to threaten subsistence rights seriously.

The new ideology has affected social policy everywhere, albeit not to the same extent. The European welfare states having integrative programs with strong historical roots and wide public support were less affected than other parts of the world. In the USA even the welfare programs seem to have suffered severe cuts (Piven and Cloward 1982, Katz 1993). In many Latin American and 'postsocialist' countries, proposals figuring on the neoliberal agenda are replacing the former, more comprehensive solutions. In other parts of the world where welfare policies hardly existed before, they seem to appear mainly as variants of poverty policy. In East Asia this is explained by the Confucian ideology of self-reliance and familism (Goodman et al. 1998). In desperately poor third-world countries poverty policy seems to be the only alternative to improve the situation, albeit the mobilization and empowerment of the poor seem to acquire greater weight than in earlier poverty programs (World Bank 2000, United Nations 1998).

3. Current Theory and Research

Theory and research on poverty and on poverty policy are strongly interwoven. An attempt is made here to focus on the policy aspects. One school of thought

deals with the theoretical and political possibilities of poverty reduction on a world scale. It explores the possibilities and instruments of global redistribution. Supranational agencies, particularly the United Nations, the International Labor Office, and more recently the World Bank as well as some civil organizations, are stimulating this research. Many elements of these endeavors are summarized in the Report of the UN World Summit on Social Development held in 1995. A second trend deals with the role of poverty policy within public policy. It explores the respective roles of the state, of civil society, and of the market (economic growth) in decreasing poverty and social exclusion. The main question is whether the 'trickling down' of economic growth is a sufficient condition to lift the poor or whether additional public and civil efforts are needed. If the latter is the case (as research seems to show), the question is whether focusing on the poor is sufficient, or a broader approach is warranted. A third large field explores the instruments of poverty policy and of their evaluation. These are partly methodological issues mentioned in Sect. 4. Some of them have acquired momentous theoretical importance. The age-old problem of poor relief, the kernel of the 'social question,' was always how to separate those worthy of help and compassion from the undeserving poor. The growing number of able-bodied poor because of lasting unemployment and the increase in single parenthood and low-paid jobs put the issue again forcefully on the agenda. One direction of research—more to the left—looks for 'third way' solutions including investment in human capital and the creation of acceptable jobs. The conservative research agenda is involved in proving the harms of welfare programs that stifle incentives, destroy stable families and lead to a dependency culture.

4. *Distributional Issues*

If help is to be given, it has to be known at least who to give it to, what to give and how, and how to assess the results of the policy.

4.1 *Who Are to Be Helped?*

The poor and the needy may be defined in different ways in poverty research and in poverty politics. In research the definition may be based on objective or subjective, or on absolute or relative poverty criteria; it will depend on whether the research relates to one country or is comparative, whether it describes one point in time or covers processes—let alone the idiosyncratic outlook of the researcher. Poverty policy usually applies definitions which relate to one country and one time-point, albeit they may change with time. The main instrument of targeting help is the minimum income or assistance level, which becomes *de facto* a poverty line. Those under the line may get help. The poverty line (unlike in research) has to be unambiguous. It is usually defined by central or local governments on the basis of pragmatic considerations. It may be lower than a scientifically based adequate minimum income standard (Eardley et al. 1996).

The poverty line may relate to the individual, the family, or the household. In the last two cases different weights may be assigned to family or household members, that is, different equivalence scales may be applied (OECD 1997). In addition, poverty policy may target help also to those who are above the poverty line but whose resources are insufficient to cover some special needs. Examples are assistance offered to cover unusually high housing costs or health costs, or in the case of a crisis situation.

4.2 *Who to Give to, in What Form, and How Much?*

Help may not be offered to all those under the poverty line. The selection of the beneficiaries is in part a political issue discussed above. Methodologically it has to be decided whether to apply only an income test, or a means (resource) test comprising an income test and an asset test, or whether to add also various behavior tests (a work test, a cohabitation test, and such like). The form of the provision also matters. There are arguments for and against assistance in cash to be used freely, or tied in some way, given in kind or in intermediary forms such as vouchers or food stamps. When a poverty line is established, the policy may be to offer enough assistance to cover the whole poverty gap (the difference between the set line and the actual income) or to alleviate only the situation of the poor without eliminating the poverty gap. The adequacy of the provisions is a further major issue. The European Union has strongly and repeatedly recommended a 'right to sufficient resources' or a minimum income for all as a measure to reduce poverty and to promote social integration (Schulte 1997). In other parts of the world governments may accept poverty lines well below the adequacy level either because of insufficient resources or because of ideological reasons (Veit-Wilson 1998). The provisions may then still assure subsistence, or only mere survival, or not even that (Gans 1995). The 'less eligibility' principle appears still to survive.

4.3 *The Assessment of the Policies*

The methods of evaluation of poverty policy were developed partly in connection with the war on poverty in the USA, and partly by independent researchers in Britain and elsewhere in Europe. Evaluation deals with such issues as the efficiency and effectivity of the various programs (to what extent did they reach the target groups and did they solve the problems); the measurement of the impact of policies on horizontal and vertical equity; the costs of relief; the public acceptance of different programs; the reasons of varying take-up rates; the merits and

demerits of social assistance based on right or on discretion; the dangers of emergence of poverty traps, unemployment traps, or savings traps—all hampering efficient help. The assessment of the negative incentives mentioned above as one of the most important theoretical issues requires involved research methods, too (Haveman 1987).

5. *The Future*

The future of poverty and of the orientation of poverty policy depends to a large extent on political forces. Supranational agencies, the United Nations in the first place, are committed to poverty reduction on a global scale, with emphasis on the poorest countries. The attainment of this objective would require a stronger international state with economic mandates and a stronger international civil society to curb the strengthening market—conditions different to fulfil.

The future of individual countries will depend on the global trends, but also on their own policies. One may expect the continuation of the trends described in Sect. 2.4 with some variations and innovations (McFate et al. 1995). In the European Union and in some left-liberal circles there is a search for new instruments to combat poverty and exclusion such as the citizen's basic income (Parijs 1992), or a new integration of economic and social policy (Atkinson 1995). This orientation implies a reduction of the role of means testing in favor of more encompassing public policies. These proposals usually meet with the resistance of economically and politically strong groups.

All in all, at the start of the new millenium there are genuine threats of widening within- and between-country inequalities, of growing poverty and of the increasingly harsh treatment of the poor whose right to subsistence is questioned. The problem is not so much the scarcity of resources as their distribution. Vigorous public and poverty policies would be needed to achieve more equity in order to prevent major tensions.

See also: Feminist Economics; Industrialization; Inequality; Inequality: Comparative Aspects; Poverty, Culture of; Poverty: Measurement and Analysis; Poverty, Sociology of; Social Security; Underclass; Welfare; Welfare Programs, Economics of; Welfare State, History of

Bibliography

Atkinson A B 1995 *Incomes and the Welfare State*. Cambridge University Press, Cambridge, UK

Castel R 1995 *Les Métamorphoses de la Question Sociale: Une Chronique du Salariat* [*The Transformations of the 'Social Problem': A Chronicle of Wage-earners*]. Fayard, Paris

Eardley T, Bradshaw J, Ditch J, Gough I, Whiteford P 1996 *Social Assistance in OECD Countries*. HMSO, London

Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Polity Press, Cambridge, UK

Ferge Z 1986 *Fejezetek a Magyar Szegénypolitika Történetéből* [*The History of the Politics of Poverty in Hungary*]. Magvetö, Budapest, Hungary

Gans H J 1995 *The War Against the Poor. The Underclass and Antipoverty Policy*. Basic Books, New York

Geremek B 1987 *La Potence et la Pitié. L'Europe et les Pauvres du Moyen Age à Nos Jours* [*Gallows and Compassion. Europe and the Poor from the Middle Ages to the Present*]. Gallimard, Paris

Goodman R, White G, Kwon H J (eds.) 1998 *The East Asian Welfare Model: Welfare Orientalism and the State*. Routledge, London

Hatzfeld H 1971 *Du Paupérisme à la Sécurité Sociale* [*From Pauperism to Social Security*]. Librairie Armand Colin, Paris

Haveman Robert H 1987 *Poverty Policy and Poverty Research: The Great Society and the Social Sciences*. University of Wisconsin Press, Madison, WI

Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York

Katz M B (eds.) 1993 *The 'Underclass' Debate. Views from History*. Princeton University Press, Princeton, NJ

McFate K, Lawson R, Wilson W J (eds.) 1995 *Poverty, Inequality, and the Future of Social Policy: Western States in the New World Order*. Russel Sage Foundation, New York

Mollat M 1978 *Les Pauvres au Moyen Age*. Hachette, Paris

OECD 1997 *Family, Market and Community. Equity and Efficiency in Social Policy*. OECD, Paris

Parijs van P. 1992 *Arguing for Basic Income: Ethical Foundations for a Radical Reform*. Verso, London

Piven F F, Cloward R A 1982 *The New Class War: Reagan's Attack on the Welfare State and its Consequences*, 1st edn. Pantheon Books, New York

Schulte B 1997 Juridical instruments of the European Union and the European Communities. In: Beck W, van der Maesen L, Walker A (eds.) *The Social Quality of Europe*. Kluwer Law International, The Hague, pp. 51–76

Swaan de A 1988 *In Care of the State: Health Care, Education and Welfare in Europe and the USA in the Modern Era*. Oxford University Press, Oxford, UK and Polity Press, New York

United Nations 1998 *Poverty Reduction Strategies. A Review*. United Nations, New York

Veit-Wilson J 1998 *Setting Adequacy Standards. How Governments Define Minimum Incomes*. The Policy Press, Bristol, UK

World Bank 2000 *World Development Report 2000/2001: Attacking Poverty*. Oxford University Press, New York

Z. Ferge

Poverty, Sociology of

Sociologists have long written about the topic of poverty, as have economists and other social scientists. Indeed, this article is complemented by several others with similar themes. Readers may therefore want to also consult other entries as well (e.g., *Poverty:*

Measurement and Analysis; *Poverty and Gender in Affluent Nations*). Here we begin with a basic description of the nature and causes of poverty as seen by sociologists, and second with the techniques used to measure poverty by sociologists and other social scientists in both national and cross-national contexts. Our study also focuses on cross-national comparisons with rich countries.

1. Sociology and Poverty

Because the fundamental concept of poverty concerns itself with having too little resources or capabilities to participate fully in a society, the study of poverty has found a wide following in sociology. Sociologists have been more interested in the explanation of poverty than in its measurement, though almost all sociologists believe that poverty statistics are meaningful social indicators of basic needs (Townsend 1993, Piachaud 1987, Ringen 1985). Sociological interests in poverty center around the ideas of the 'culture' of poverty and the effects of 'place' on poverty. The urban—central city—ghetto aspects of this issue are well treated in the work of Wilson (1987), Massey and Denton (1993), and Massey (1996); Harrington's classic work, *The Other America*, covers rural deprivation as well (1962). Roles of culture, power, social structure, and other factors largely out of control of the individual are the main forces that sociologists use to explain poverty. The basic working hypothesis is that individuals are strongly influenced by the physical and cultural context in which they live. 'Neighborhood' exerts a strong influence on behavior and concentrated poverty in central city ghettos therefore has a strong effect on future life chances and long-run deprivation. Indeed, in the United States, this has spawned interest in the so-called underclass, which goes beyond poverty alone to include all persons with dysfunctional behaviors living in 'bad' neighborhoods (Mincy et al. 1990). The European term for a similarly disaffected population is 'social exclusion' which is also beyond poverty measurement *per se* (Room 1999, Hills 1999, Glennerster et al. 1999).

Gender and household structure also frequently play a role in sociological theories of poverty (e.g., Miller 1996, Ringen 1985). As women's roles in society often leave them with care-giving responsibilities, with low-paying careers and jobs, and therefore in poor economic circumstances, poverty is often high among single female parents and elderly widows. Thus, we find a preponderance of women in poverty in most rich nations (Casper et al. 1994, McLanahan et al. 1989, Christopher et al. 2001).

In general, sociologists are critical of the economist's perspective of free choice models whereby individuals control their own destiny and are therefore the cause of their own poverty (Piachand 1987). Some sociologists even feel that poverty has a functional role to play in capitalist society (e.g., Gans 1972). Ultimately, however, sociologists, like all other social scientists, need to first establish the breadth and depth of this social phenomenon called 'poverty' before they can meaningfully analyze it and explore its ultimate causes and remedies. Thus, we turn to measures and comparisons of poverty employed by sociologists and other social scientists within and across nations.

2. Poverty in Rich Nations: Concepts and Measures

We begin by noting that poverty measurement began as an Anglo-American social indicator. In fact, official measures of poverty (or measures of low income) exist in very few nations. Only the United States (US Bureau of the Census 1999) and the United Kingdom (Department of Social Security 1993) have official poverty series. Statistics Canada publishes the number of households with incomes below a low income cutoff on an irregular basis, as does the Australian government. In Northern Europe and Scandinavia, the debate centers instead on the level of income at which minimum benefits for social programs should be set. In other words, their concept of insufficient 'low income' is directly fed into programmatic responses to social needs (Björklund and Freeman 1997).

While poverty measurement is an exercise that is particularly popular in the English-speaking countries, most rich nations share the Anglo-Saxon concern over distributional outcomes and the well-being of the low-income population. There is no international consensus on guidelines for measuring poverty, but international bodies such as the European Statistical Office (Eurostat 1998), the United Nations Development Programme (UNDP 1999), the United Nations Children's Fund (UNICEF 2000), and the Organization for Economic Cooperation and Development (OECD 2000), have published several cross-national studies of the incidence of poverty in rich countries in recent years. The large majority of these studies are based on the Luxembourg Income Study (LIS) database which can be accessed remotely via Internet at www.lis.ceps.lu. Some examples of these studies include Förster (1993), Smeeding et al. (1990), Kenworthy (1998), Jäntti and Danziger (2000), and Smeeding et al. (2000), many of which are directly available online from the LIS web page.

The measure of poverty in rich nations involves the comparison of some index of household well-being with household needs. When command over economic resources falls shorts of needs, a household (or person or family) is classified as poor. Well-being refers to the material resources available to a household. The concern with these resources among most social scientists, including sociologists, is not with material

consumption, *per se*, but rather with the capabilities such resources give to household members so that they can participate fully in society (Sen 1983, 1992). These capabilities are inputs to social activities, and participation in social activities gives rise to a particular level of well-being (Coleman and Rainwater 1978, Rainwater 1990). Methods for measuring a person's or household's capabilities differ according to the context in which one assesses them, either over time or across nations or among subpopulations within a nation.

Most sociologists recognize that all advanced societies are highly stratified (e.g., Townsend 1979). Hence, some individuals have more resources than others. The opportunities for social participation are affected by the resources that a household disposes, particularly in nations like the United States where there is heavy reliance on the market to provide such essential services as health care, post-secondary education, and child care. Money income is therefore a crucial resource. Of course, there are other important kinds of resources, such as social capital, noncash benefits, primary education, and access to basic health care, all of which add to human capabilities (Coleman 1988). These resources may be available more or less equally to all people in some societies, regardless of their money incomes.

Sociologists believe that there are many forces in rich societies that reduce well-being by limiting capabilities for full participation in society, including inadequacies in neighborhoods where people live. Sociologists are therefore interested in racial discrimination, neighborhood violence, low quality public schools, lack of good jobs, and job instability, all of which increase economic insecurity, reduce human capabilities, and increase poverty (e.g., Massey 1996).

2.1 Absolute vs. Relative Poverty

An absolute poverty standard is defined in terms of a level of purchasing power that is sufficient to buy a fixed bundle of basic necessities. A relative standard, on the other hand, is defined relative to the typical income or consumption level in the wider society. The purchasing power of a relative poverty standard will change over time as society-wide income or consumption levels change; an absolute poverty standard will change only with the prices of commodities. Most sociologists inherently favor the relative definition of poverty.

And, in fact, all measures of poverty or economic need are relative, because context is important to the definition of needs. The World Bank uses poverty measures of \$1 to \$2 per person per day—or \$1,095 to \$2,190 per year for a family of three—for the developing nations of Africa or Latin America (Ravallion 1996). In contrast, the 1998 United States, 'absolute' poverty threshold was \$13,003 for a family of three—about 6 to 12 times the World Bank's poverty line—while one-half of median income, the preferred relative poverty standard, was another 25 percent above this poverty line or 8–15 times the poverty standard in poor countries. Moreover, as economic inequality has increased in most rich societies over the past 20 years, the study of relative deprivation and poverty has taken on a new life (Gottschalk and Smeeding 2000). The link between marginalization, relative poverty, and the globalization of trade is just beginning to emerge as a topic of prime interest among sociologists (e.g., Alber and Standing 2000).

In fact, cross-national comparisons of poverty in rich countries rely almost exclusively on a relative concept of poverty. Cross-national studies typically compare the percentage of persons living with income below some fraction, normally half, of the national median income. This kind of comparison is consistent with a well-established theoretical perspective on poverty (Townsend 1979, Sen 1983, 1992). However, some feel that such a standard is too low. For instance, the European Statistical Office Working Group on Poverty Measurement recently recommended using 60 percent of the national median income as a common poverty threshold for European Community poverty studies in the new millennium (Eurostat 2000).

2.2 Alternative Measures of Resource Poverty

Because there is no single commonly accepted way to measure poverty among sociologists, there is a desire to go beyond the popularly used income poverty definition. As a result, there exists a wide variety of additional poverty measures that substitute for or complement the preponderance of income-based measures used by quantitative sociologists and economists (e.g., Ruggles 1990, Haveman and Mullikin 1999, Boltvinik 2000). In principle, poverty is a multidimensional concept and should reflect several aspects of personal well-being. Forms of deprivation other than economic hardship can certainly be relevant to poverty measurement and to anti-poverty policy-making.

Even restricting consideration to economic poverty, measures other than annual income might be used to estimate household resources. For households with fluctuating incomes, annual income can provide a poor approximation for permanent or long-run average income. Some social scientists suggest that annual consumption represents a better proxy for permanent income and hence should be preferred over annual income for measuring household resources (Haveman and Mullikin 1999). However, broad-based consumption data are more difficult to come by than are income data. A number of authors have suggested that separate measures of needs ought to be developed for different goods and services (Aaron 1986). Housing and health care are often mentioned in this context, while the former is of much greater interest in the United Kingdom (DSS 1993).

2.3 Social Indicators

Social indicators present an entirely different method of poverty measurement. Researchers have a long tradition of using social indicators to measure both economic hardship and affluence. Analysts often define direct measures of deprivation, such as lack of food, heat, or access to health care (Bauman 1998, Mayer and Jencks 1993). Because each household's need is imperfectly measured by an income or consumption threshold, many households with incomes above the thresholds may still experience hardships, such as hunger, or face high rent payments or heavy medical expenses that prevent its members from buying enough food or fuel. Further, individual members of a household may suffer hardship because of unequal sharing of incomes within the household. In these cases, direct measures of material hardship might offer a useful indication of poverty, especially in a cross-national context, and many sociologists are therefore attracted to such measures (Mayer and Jencks 1993).

2.4 Time Poor

Finally, resources other than access to material goods might be important determinants of poverty. The literature in the 'time poor' has grown in recent years after a nascent period. The argument is that time is often the scarcest commodity in market work-oriented societies. Further, time spent on home production is often undervalued or not valued at all. The gender bias in time use is therefore at the heart of these types of analyses (Vickery 1997, Douthitt 1994). In particular, recent research has focused on the time crunch on single parents who must both earn a living in the labor market and care for their children (Folbre 1994, England and Folbre 1999).

3. Summary

Sociologists and other social scientists have used a variety of methods to define and measure poverty. Each approach has its strengths and weaknesses. Over the past few decades, income-based poverty measures using annual assessments of household income have provided the main statistics for tracking poverty over time, across nations, and across sub-populations within nations for sociologists and for other social scientists. They have also provided important tools for testing causal hypotheses concerning the explanations, causes, and cures for poverty, including place, space, race, gender, and other popular theories of poverty in sociology. They are complemented by social indicator-based measures of inadequacy, and by the adequacy of time for market work, home work, and the nurturing and care of family members. Sociological studies of poverty have both a rich heritage and have become an important area of research in increasingly unequal rich societies.

See also: Equality and Inequality: Legal Aspects; Income Distribution; Inequality; Inequality: Comparative Aspects; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Poverty, Culture of; Poverty in History; Poverty Law; Poverty Policy; Social Stratification; Wealth Distribution

Bibliography

Aaron H 1986 Comments on methods of measuring noncash benefits. Reported. In: *Noncash Benefits: An Evaluation of the Census Bureau's Measurement Conference*. US General Accounting Office, Washington, DC

Alber J, Standing G 2000 Social dumping, catch-up or convergence? Europe in a comparative global context. *Journal of European Social Policy* **10**: 99–119

Bauman K J 1998 *Direct Measures of Poverty as Iudicators of Economic Need: Evidence from the Survey Glucane and Program Participation*. Population Division Technical Working Paper No. 30. US Bureau of the Census, Washington, DC

Björklund A, Freeman R 1997 Generating equality and eliminating poverty—the Swedish way. In: Freeman R B, Topel R, Swedenborg B (eds.) *The Welfare State in Transition: Reforming the Swedish Model*. University of Chicago Press, Chicago

Boltvinik J 2000 *Poverty Measurement Methods: An Overview*. United Nations Development Programme, SEPED Series on Poverty Reduction, United Nations, New York

Casper L M, McLanahan S S, Garfinkel I 1994 The gender–poverty gap: What we can learn from other countries. *American Sociological Review* **59**: 594–605

Christopher K, England P, McLanahan S, Ross K, Smeeding T 2001 Gender inequality in poverty in affluent nations: The role of single motherhood and the state. In: Vleminckx K, Smeeding T (eds.) *Child Well-being, Child Poverty and Child Policy in Modern Nations: What Do We Know?* Policy Press, Bristol, UK, pp. 199–220

Coleman J S 1988 Social capital in the creation of human capital. *American Journal of Sociology* **94** (Suppl. 5): S95–S120

Coleman J, Rainwater L 1978 *Social Standing in America: New Dimensions of Class*. Basic Books, New York

Department of Social Security 1993 *Households Below Average Income, A Statistical Analysis, 1979–1993/94*. Government Statistical Service, London, UK

Douthitt R A 1994 *Time to do Chores? Factoring Home-production Needs into Measures of Poverty*. Discussion Paper No. 1030–94, Institute for Research on Poverty, University of Wisconsin, Madison, WI

England P, Folbre N 1999 Who should pay for the kids? *Annals of the American Academy of Political and Social Science* **563**: 194–207

Eurostat 1998 *Recommendations of the Task Force on Statistics on Social Exclusion and Poverty*. European Statistical Office, Luxembourg

Eurostat 2000 *Report of the Working Group: Statistics on Income, Social Exclusion and Poverty*. European Statistical Office, Luxembourg

Folbre N 1994 *Who Pays for the Kids? Gender and the Structures of Constraint*. Routledge, New York
Förster M 1993 Comparing poverty in 13 OECD countries: Traditional and synthetic approaches. *Studies in Social Policy #10*. OECD, Paris
Gans H 1972 The positive functions of poverty. *American Journal of Sociology* **78**(2): 275–89
Glennerster H, Lupton R, Noden P, Power A 1999 *Poverty, Social Exclusion and Neighborhood: Studying The Area Bases of Social Exclusion*. CASE paper 22, Neighbourhood Research Centre for the Analysis of Social Exclusion, London School of Economics, London UK
Gottschalk P, Smeeding T M 2000 Empirical evidence on income inequality in industrialized countries. In: Atkinson A B, Bourgignon F (eds.) *Handbook of Income Distribution*. North Holland, New York, pp. 261–308
Harrington M 1962 *The Other America: Poverty in the United States*. Macmillan, New York
Haveman R, Mullikin M 1999 *Alternatives to the Official Poverty Measure*. University of Wisconsin, Madison, WI. Presented at a Conference on 'Poverty: Improving the definition after 30 years'
Hills J 1999 *Social Exclusion, Income Dynamics and Public Policy*. Annual Sir Charles Carter Lecture, Northern Ireland Economic Council, Report No. 129, Belfast, UK
Jäntti M, Danziger S 2000 Income poverty in advanced countries. In: Atkinson A B, Bourguignon F (eds.) *Handbook of Income Distribution*. North Holland, Amsterdam, pp. 309–378
Kenworthy L 1998 *Do Social-welfare Policies Reduce Poverty? A Cross-national Assessment*. Luxembourg Income Study Working Paper #188, Center for Policy Research, Syracuse University, Syracuse, NY
Massey D 1996 The age of extremes: Concentrated poverty and affluence in the 21st century. *Demography* **33**: 395–412
Massey D S, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA
Mayer S E, Jenks C 1993 Recent trends in economic inequality in the United States: Income versus expenditures versus material well-being. In: Papadimitriou D B, Wolff E N (eds.) *Poverty and Prosperity in the USA in the Late Twentieth Century*. St. Martins Press, New York
McLanahan S, Sorensen A, Watson D 1989 Sex differences in poverty 1950–1980. *Signs: Journal of Women in Culture and Society* **15**: 102–22
Miller S M 1996 The great chain of poverty explanations. In: Øyen E, Miller S M, Samrod S A (eds.) *Poverty: A Global View*. Scandinavian University Press, Oslo, Norway, pp. 569–586
Mincy R, Sawhill I, Wolf D 1990 The underclass: Definition and measurement. *Science* **248**: 450–3
Organization for Economic Cooperation and Development 2000 *Trends and Driving Factors in Income Distribution and Poverty in the OECD Area*. Social Policies Studies Division Occasional Paper #42, OECD, Paris
Piachaud D 1987 The distribution of income and work. *Oxford Review of Economic Policy* **3**: 41–61
Rainwater L 1990 *Poverty and Equivalence as Social Constructions*. Luxembourg Income Study Working Paper #55, Center for Policy Research, Syracuse University, Syracuse, NY
Ravallion M 1996 *Economic Journal* **106**: 1328–44
Ringen S 1985 Toward a third stage in the measurement of poverty. *Acta Sociologica* **28**: 99–113
Room G J 1999 Social exclusion, solidarity and the challenge of globalization. *International Journal of Social Welfare* **8**: 166–74
Ruggles P 1990 *Drawing the Line: Alternative Poverty Measures and their Implications for Public Policy*. Urban Institute Press, Washington, DC
Sen A 1983 Poor, relatively speaking. *Oxford Economic Papers* **25**: 153–169
Sen A 1992 *Inequality Reexamined*. Harvard University Press, Cambridge, MA
Smeeding T M, O'Higgins M, Rainwater L 1990 *Poverty, Inequality and the Distribution of Income in a Comparative Context: The Luxembourg Income Study* (*LIS*). Wheatsheaf Books, London, UK and Urban Institute Press, Washington, DC
Smeeding T, Rainwater L, Burtless G 2000 *United States Poverty in a Cross-national Context*. Center for Policy Research, Luxembourg Income Study Working Paper #244, Syracuse University, Syracuse, NY
Townsend P 1979 *Poverty in the United Kingdom: A Survey of Household Resources and Standards of Living*. University of California Press, Berkeley, CA
Townsend P 1993 *The International Analysis of Poverty*. Harvester Wheatsheaf, London, UK
UNICEF Innocenti Research Centre 2000 *A League Table of Child Poverty in Rich Nations*. Innocenti Report Card No. 1, UNICEF, Florence, Italy
United Nations Development Programme 1999 *Human Development Report* 1999: *Globalization With a Human Face*. United Nations, New York
US Bureau of the Census 1999 Poverty in the United States: 1998. *Current Population Reports* P60–207, US Government Printing Office, Washington, DC
Vickery C 1997 The time-poor: A new look at poverty. *Journal of Human Resources* **12**: 27–48
Wilson J 1987 *The Truly Disadvantaged: The Inner City, the Underclass and Public Policy*. University of Chicago Press, Chicago

T. M. Smeeding

Power in Society

Power has been seen as an 'essentially contested' concept (Lukes 1974). Some of the major recent contours of this contestation will be addressed in this entry.

1. *Weber: Ideal Types of Rule and Power*

At the core of modernism, according to Weber, lies the rationalization of various institutional arenas such as the market, technology, law, state, the growing pervasiveness of rational calculation in all spheres of life, and the disenchantment of more spiritual or superstitious forms of calculation. Disenchantment means exposure to a world in which meaningfulness is never

Table 1
Power/authority dynamics

	Versus tradition	Versus charisma	Versus rational-legality
Tradition	Rival dynasties	Dynasty confronting a revolutionary hero	Privilege ranged against rights according to law
Charisma	Princess Diana versus the British Royal Family	Hitler against Stalin during World War II	Nelson Mandela against De Klerk's Apartheid Government of South Africa
Rational-legality	Oliver Cromwell leading the Parliament against the monarch, Charles I	The Marcos Government in the Phillipines against Cory Aquino	Two governments disputing each others claim to the same territory, such as India and Pakistan over Kashmir

'given.' Instead, meanings clash, struggle against the resistance of others and are never wholly secure. Rational action consists precisely in the capacity to respond to the uncertainties of a world without meaning. Analytically, he conceived of three main types of rule around which rational action might revolve: principles of charisma, tradition, or rational legality. Each of these three types is a source of a particular type of *Herrschaft* (authority, or in better translation, domination).

In Weber's view, rationality had only a limited number of sources. It might derive from those few individuals who choose to impose value and meaning upon the world by confronting the world as something other than resolutely unyieldingly, absolutely and unbreakably cast in a mold. There were precious few such people in Weber's vision, other than charismatic leaders. Charisma is often identified with historical figures regarded as ordained with exceptional grace and a sense of calling that they were able to transmit and project to others, such as Jesus. Charisma is not a purely historical category, however, but it is only the rare charismatic leader, whose personal grace, sense of calling, duty, and devotion to some ultimate ideal is unshakeable, who would be able to impose such meaning on life. Hence, charisma was one of the key bases of authority that Weber identified. Charisma is typically something above the grubby theater of mundane political wheeling and dealing. In terms of modern life we might think of celebrities such as Princess Diana who were able to parlay their celebrity into a passionate social advocacy on behalf of the victims of landmines and in so doing, effect substantial changes in government policies. Here was a charismatic celebrity for the modern age—especially in her death. Clearly, in modern times, celebrity and charisma are closely related but by no means the same. Often they become confused: the cult of the personality that modern leaders like Mao, Stalin, and Hitler cultivated were deliberate attempts to use tight control of the state to fashion and stage-manage a celebrity that transcended the political sphere and which sought to invest them with charisma.

Tradition functions as another ideal type, best seen in claims to authority based on lineage or dynasty, in principles of 'blood and breeding' or exclusion from opportunities to claim authority because of the lack of such blood and breeding. The succession principles of monarchy, or inheritance rights attached to property, are prime examples of rule by tradition. Most pervasive, however, as a form of authority, was what Weber took as the quintessential modality of modern rule: rational-legal bureaucracy, where rule is conducted according to abstract principles. The dominance of what Weber referred to as the rational-legal does not overwhelm all other bases for rule and authority. However, its pervasiveness is often summed up today in a popular fear of 'economic rationalism,' referring to those situations where only a legally mandated economic calculus is applied to policy formulation (such as competition policy), without regard for other ways of accounting for the impact of such policies on the human spirit, communities, and the fabric of society.

Each principle of rule posits a particular type of authority and it is a simple matter to construct a matrix of these relations in order to generate some idea of what power might be. Weber defined power in terms of social action that is goal-oriented, even against the resistance of others. An evident example of power thus arises when there are cases of conflict between principles of rule (see Table 1). Here, power is seen in terms of the opposed principles or ultimate values.

2. *Power as a Generic Concept*

While Weber's definition of power is quite generic, the concept need not be premised on the conflict of rival, ultimate values, as per the examples of Table 1. Steven Lukes (1974), for instance, in one of the most influential books on the subject, regarded Weber's definition as essentially similar to the generic definition of power contained in the work of Robert Dahl (1957). Here power is defined as the situation that prevails when A gets a B to do something that they would not

otherwise do. This sounds a methodologically sparse and modern behavioral definition but it is one that reaches back into the earliest development of the modern era—to Thomas Hobbes's conception of power in a causal, mechanical world.

The Community Power Debate that formed around Dahl's (1961) work concerned the distribution of power in postwar USA. Dahl (1961) insisted that such questions could only be answered empirically, rather than through a broad ranging essaying of concepts that were then current, such as their being a 'military-industrial complex.' Additionally, he thought them better answered at the local, community level rather than at the national level. To determine the distribution of power required an operational definition for empirical investigation. For Dahl, power was indissoluble from its exercise: Someone had to be seen doing something to another for us to say that power had been exerted. Unlike Weber, he was not concerned with ultimate values but he did think it important that the exercise of power should be determined by competition over issues that people thought important. Basically, he sought to determine what the key issues were in a community and then establish the outcome of the politics surrounding these decisions, to see who gained or yielded power.

3. *Power, Real Interests and Hegemony*

By the time that Lukes (1974) published his book on *Power: A Radical View*, Dahl's approach had already been criticized by Bachrach and Baratz (1970) for leaving out of consideration 'what does not happen.' Nonissues, nondecision making, the construction of agendas, and the mobilization of bias through everyday routines, had all been introduced to show how power is accomplished through covert means, through inaction, rather than through overt exercise of causal power.

Lukes took the critique one step further. He argued that the critics had only sought to qualify Dahl's behaviorism rather than transcend it through radical critique. What was required, he argued, was a conception of how power worked not only against people's expressed and implicit preferences but also their 'real interests.'

The notion of real interests came from Marxist discourse—originally designed to address situations of 'false consciousness'—where a class of people were unable to recognize that the system within which they were embedded was the source of their entrapment. Lukes followed the Italian Marxist theorist, *Antonio Gramsci*, in introducing a 'cultural spin' with the notion of hegemony. Where people were deluded about what they wanted and what they might authentically be, because of their inability to think in concepts other than those of domination and the dominated, then they suffered from hegemony. The basic idea was that they were subject to a 'ruling culture' disseminated through the media, the state, education, and other institutional arenas.

Theories of hegemony became variations upon the theme of working-class acquiescence secured through normative incorporation by dominant ideology (Abercrombie et al. 1980). The narrative was one of concealment: Marxist theory, in theory, should have provided revelation—but it did not. Only the theorist could see the truth of the mechanisms that, despite and even through resistance, secured and reproduced consent. The subjects of analysis had the nature of their real interests concealed from them. Western Marxism, the source of these insights, was impotent at best and at worst irrelevant, in lifting the veils of occlusion (Anderson 1976). In the West socialism became increasingly utopian (Bauman 1977); in the East increasingly dystopian (Fehér et al. 1983), as the collapse of 'dictatorship over needs' produced its implosion.

4. *The Discursive Turn*

The ideological arguments, with their fundamental denial of consumer sovereignty, receded in importance in the 1980s. One response to this situation within the Marxist camp came from Laclau and Mouffe 1985, who instead of thinking of hegemony as a state of mind regarded it as a set of practices. These practices were primarily discursive. Hegemony thus came to be seen not in terms of its substantive content but its form. Hegemony existed when the indefinite possibilities of signifying elements and their relations were strategically foreclosed (Laclau and Mouffe 1985). Again, the emphasis is on things that do not happen. Any stable set of articulations is thus hegemonic: Only permanent Cultural Revolution would not be—although the implications are not spelled out.

Real interests are what radical politics and theories claim to uncover and reveal. One way of doing this, suggested Haugaard (1998), might be to explain to actors the way in which their structured practices feed into the reinforcement of a regime of domination contrary to interests already known to them. What may be at issue are structural and institutional relations of domination rather than real but unknown interests. Once actors reflect upon their practical consciousness then they enjoy relative autonomy. Power thus consists not so much in the suppression of real interests but the denial of the relative autonomy required for reflection (Haugaard 1998, p. 141).

Huagaard (1998) notes that conceptions of interests form discursively, even to the extent that one can persuade some people that they have real but unrealized interests blocked by their present conditions of existence. Such discursive realities help shape them to

resist the sources of domination identified. Politics mobilize power and resistance through discursive expression of existing, past, or future conditions of existence. Thus, resistance is not endemic but depends on the practical consciousness of people and the categorical devices through which they lead their everyday lives in relation with the discursive consciousness proffered by various media of representation—parties, press, radio, television, the Internet, public intellectuals. Only in rare moments is such diversity represented or resisted in the name of some great unifying narrative or theme. Most probably these will be those relatively rare moments where a privileged identity is posited as a historical subject: the subject people, the subject party, the subject race, the subject nation, the subject church, the subject gender, the subject sexuality, the subject ecology, and so on.

A plurality of life-worlds, with interpretative resources, and translation devices for switching between them, minimizes the chances of hegemony. People from sophisticated cultures routinely have a multiplicity of interpretative horizons available to them, as a result of exposure to different life-worlds through various media, educational experiences, travel or migration, or interpersonal relations. Of course, to the extent that these experiences are compartmentalized, one way or another, then the ability to achieve critical distance from any posited identity is diminished.

The theorization of autonomy in terms of moves between interpretative horizons means that the ability of actors to transcend their social environments and structured context does not presuppose an ability to 'escape' the processes of socialization by adopting a neutral, culturally unbiased, position. Autonomy does not involve a flight into any transcendental realism of absolute Truth, nor does it involve the transcending of false consciousness. Critical faculty does not imply the objectivity of truth, nor does it entail an undersocialized concept of agency whereby autonomy is gained by transcending the basis of social culture. It requires only relative reflexivity on the part of subjects. Such a conception of the relation of power and knowledge, with its emphasis on conditions of relative autonomy, turns out to be a way of re-approaching the classical questions about power and community that were at the center of Dahl's work. The difference is that, this time, the approach has been through Foucault rather than Dahl.

5. *Foucault and Power*

Foucault provides a detailed history of some power practices and techniques that have characterized modernity in a way that is nonreductive and complements Weber (Dandekker 1991), although he understates the importance of Max Weber's (1978) argument concerning discipline. Nonetheless, there are common themes (see Szakolczai 1998). Different emphases on power occur through the course of Foucault's work. First, there is his work on genealogy as the natural history of specific practices. Second, the work on the 'care of the self,' where histories are used to critique the present, by uncovering the constituted nature of what have been taken as the here-and-now; third, the discussion of government or 'governmentality.'

In Foucault's (1979) most influential genealogical model of 'surveillance and discipline,' power is constituted in an intimate relationship with liberty. Because the exercise of power presumes a free subject, in a well-known formula, there will always be resistance. Where liberty is extremely limited, because power is stable and hierarchical, then Foucault terms it domination. (The parallelism of this account with that of Weber's 1978 discussion of *Herrschaft*, is evident.) Two models of power are compared and contrasted: the repressive and the positive. (Foucault's 1979 concern with the facilitative and productive aspects of power relates to the kind of arguments made most clearly by Parsons 1963.) In general Foucault does not offer a critique of power: It would be impossible to do so because, to coin a phrase, power has no other: It is inescapable. Liberation from power is only a dream (or a nightmare). However, specific forms of power may and do change. The 'punitive' power of corporal excess may be contrasted with the 'positive' power of bureaucracy. However, there is no space outside of power, no knowledge that is pure truth. (In this way are utopia and dystopia avoided.) In reading Foucault it is necessary to uncouple a sociological from a philosophical question concerning truth. The former deals with what passes for truth rather than asking what is truth—with how truth-effects are produced, in the Foucauldian vernacular. It is the latter that concerns Foucault, rather than the former.

In Foucault's model of power relationships through government—the structuring of action through conduct—which he defines as the capacity of people to regulate their own behavior through various rationalities, his treatment of liberalism suggests that it is conditions of relative autonomy that define the nature of political community.

6. *Community Politics after Foucault*

Foucault suggests that democracy is neither an ideal with which to measure actual practice nor a method for producing utopias. It is the messy and politically expeditious ways of making sense of politics its participants routinely use. It is this fundamental approach that Flyvberg (1998) takes from Foucault. He provides a deeply documented, rich narrative of one complex set of issues and processes associated with the many ways in which political communities were constructed and deconstructed in the laborious

pursuit of local interests and issues surrounding a proposed traffic-plan for the city of Aalborg in Denmark.

Aalborg was not unusual in respect of its power relations: They were sufficiently asymmetric as to verge on domination. Here, the authority of technical rationality was a weapon of the politically weak—those without other resources at their disposal. Hence, not surprisingly, technically rational plans were constantly undercut and attempts at democracy secured on that technically rational basis proved weak. Neither a normative emphasis on democracy nor rationality proved sufficient to avoid domination by superior power. Such power defined reality and, not surprisingly, the rationality it accepted depended on power rather than legality. One of the main strategies of the dominant group was to use techniques of rationalization as power: Indeed, the greater the power the lesser the rationality. Thus, we might conclude that power is more ordinarily stable than conflictual and that its relations are constantly being produced and reproduced in process. Where power is stable then power relations are more characteristically rational.

7. Conclusion

In conclusion, we may note that in the genealogy of power that has been constructed here, power is not defined as only one thing, or in one set of terms. The terms shift with debate but there is a family resemblance among them. The emphases have changed with the values underlying scholarship. Lukes' importation of the Gramscian notion of hegemony seemed to take us a long way from the political communities that Dahl studied with such methodological scruples. Laclau and Mouffe's introduction of Foucault to the debate, somewhat surprisingly, brings debate back full circle to issues of political community because of the centrality, their conception of power placed on discourse and identity. The theme of the relative autonomy of political processes dominated the late work of Foucault and was taken up empirically by Flyvberg. Thus, nearly 50 years after the Community Power Debate first kicked off contemporary sociological interest in the conceptual relation of power and society, the concepts have come back to their empirical point of departure—but much richer and more sophisticated than when they first embarked on that journey.

See also: Authority, Social Theories of; Charisma: Social Aspects of; Democracy; Foucault, Michel (1926–84); Political Sociology; Power: Political; Weber, Max (1864–1920)

Bibliography

Abercrombie N, Hill S, Turner B S 1980 *The Dominant Ideology Thesis*. Allen and Unwin, London

Anderson 1976 *Considerations on Western Marxism*. Verso, London

Bachrach P, Baratz M S 1970 *Power and Poverty: Theory and Practice*. Oxford University Press, Oxford, UK

Bauman Z 1977 *Socialism: The Active Utopia*. Allen & Unwin, London

Clegg S R 1990 *Modern Organizations: Organization Studies in the Postmodern World*. Sage, London

Dahl R A 1957 The concept of power. *Behavioural Science* **2**: 201–5

Dahl R A 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT

Dandekker C 1991 *Surveillance, Power and Modernity*. Polity Press, Cambridge, UK

Fehér F, Heller A, Márkus G 1983 *Dictatorship Over Needs*. Blackwell, Oxford, UK

Flyvberg B 1998 *Rationality & Power: Democracy in Practice*. Chicago University Press, Chicago

Foucault M 1979 *Discipline and Punish*. Penguin, Harmondsworth, UK

Haugaard M 1998 *The Constitution of Power*. Manchester University Press, Manchester, UK

Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy*. Verso, London

Lukes S 1974 *Power: A Radical View*. Macmillan, London

Parsons T 1963 On the concept of political power. *Proceedings of the American Philosophical Society* **107**: 232–62

Szakolczai A 1998 *Max Weber and Michel Foucault: Parallel Life–Works*. Routledge, London

Weber M 1978 *Economy and Society*. University of California Press, Berkeley, CA

S. R. Clegg

Power: Political

Leading political theorists, such as Machiavelli and Hobbes, have devoted much attention to 'political power.' As Robert A. Dahl noted (1957), 'power' has a certain universality, whether it is called 'power, influence, control, *pouvoir*, *puissance*, *Macht*, *Herrschaft*, *Gewalt*, *imperium*, *potestas*, *auctoritas*, *potentia*,' etc. This article focuses on the concept of 'power' defined as a type of social causation.

1. Power as Causation

From 1953 to 1975, several leading political scientists developed the view of power as social causation, converging with an earlier viewpoint of Max Weber. After rejecting alternative definitions of 'power,' Herbert Simon (1953) cited a definition by Lasswell and Kaplan (1950): 'The exercise of influence (influence process) consists in affecting policies of others than the self.'

In the formulation best known to political scientists, Dahl stated (1957): 'My intuitive idea of power, then, is something like this: A has power over B to the extent

that he can get B to do something that B would not otherwise do.' James G. March (1955) stated: '… the set of all influence relations is here defined to be that subset of all causal relations such that the behavior of an individual appears as the terminal point in the causal linkage. Alternatively, we can say that two individuals are in an influence relation if their behaviors are linked causally.'

This leads to distinctions among different types of causal relations as they might relate to power. First, 'political power' refers to causal relations among humans, not to inanimate forces changing human behavior, as a hurricane might, for instance. Second, a distinction causing much confusion, one needs to distinguish situations in which the changed behavior reflects the intentions of the powerful actor. One might change the behavior of another in a way not intended, as when an incompetent lobbyist changes a legislator's vote to one different from that favored by the lobbyist. Some political scientists, such as March, are more comfortable with 'power' as causing changes in the direction intended by the powerful actor, and referring to the more general category of causing any changes as 'influence.' However, the seminal articles by Simon and Dahl do not restrict 'power' to intended changes in others' behavior.

Other examples of such causal relations would include power-as-suggestion: when A proposes a behavior change, B adopts the change without resi stance. Another category of causal relations occurs when B resists, and A exercises rewards and sanctions, perhaps even force, to get B to change his or her behavior. This type of power may be called coercion (McFarland 1969).

In regard to power as coercive causation, John Harsanyi (1962) emphasized the costs of exercising power and also the costs of compliance with the exercise of power. For instance, a lobbyist might change a legislator's vote with little cost if it just means arranging a meeting with one of the legislator's staff. However, the cost is high if the lobbyist must mobilize hundreds of persons to petition the recalcitrant legislator. From the standpoint of the legislator, a vote may be 'cheap,' offending almost no one, or it may endanger victory in the next election. Still another causal situation occurs when A's preferences cause changes in B's preferences, causing B to act differently. This is widely known as 'anticipated reactions,' in which B, perhaps to please A and to avoid A's exercise of sanctions, changes behavior in line with an anticipation of what A wants him or her to do, even though A did not explicitly state such preferences (Nagel 1975).

An important, implicit aspect of defining power as causation is its link to empiricism. A statement of power then implies observations of instances of A changing B's behavior. It rules out defining 'power' in terms of the resources (wealth or force, etc.) used to get B to comply. One first has to observe the causal relationship; then one analyzes how power was wielded.

Such resources were called by Dahl (1957) a 'power base,' and both he and Simon (1953) stated that the power base should not be confused with 'power' itself. In addition, Dahl stated that power-as-causation, to be an empirically useful concept, had to be used with four concomitant dimensions of power: base, amount, scope, and domain. *Base*, then, refers to the resources or means that A uses to cause changes in others' behavior. *Amount* refers to the idea that some instances of power refer to greater changes in behavior than others. A lobbyist exercises more power when he or she changes the legislator's vote from 'no' to 'yes,' for instance, than if the change is only from 'abstain' to 'yes.' The *domain* of an actor's power consists of those persons subject to the actor's power. The *scope* of power consists of the matters subject to the actor's power.

Dahl's major contribution to the study of political power is his insistence that the domain and scope always be clearly stated in discussions of power. Or to put it another way, an actor's power in one situation should not be immediately generalized to postulating equivalent power over other persons in other situations. The pharmaceutical lobbyist may have power over just one or over a score of legislators regarding the regulation of generic drugs. The same lobbyist might not have equivalent power over votes about Medicare payments for prescriptions; this lobbyist may have no power at all over any legislators in regard to civil rights or to gun control.

Scope and domain of power were key issues in the still widely remembered community power debate of the 1960s. Following the example set by Dahl in *Who Governs*? (1961), political scientists studied political power in localities by conducting case studies of public policy-making and charting which persons changed the course of events. Such studies normally found that the domain and scope of power was fragmented; different persons had power in different areas of policy. On the one hand, other social scientists—tending to be sociologists—followed a different conception of power, conceiving power to be possession of wealth or high social status. Such observers usually described inequalities of wealth and status and stated this as a finding of 'power' in a community. Political scientists, on the other hand, for the most part did not consider wealth and status to be 'power' itself, but instead to be resources that could be used to win political power, if such wealthy or high-status persons intended to do so. But such intentions are an empirical question, in this view, and must be observed in case studies.

This famous (or notorious) controversy extended for at least a decade and was expressed in scores of books and hundreds of journal articles before the frustrated debate participants, unable to persuade one another, for the most part went on to study questions other than community power (Polsby 1980).

Proponents of power-as-resource generally overlook a strong criticism of the power-as-causation model. Social causation is ordinarily very complex; case studies are not easy to do; power as causation thus is ordinarily complex, requiring a great deal of effort to study.

Such writers as Simon (1957), March (1955), Harsanyi (1962), and Riker (1964) left us with elaborate statistical models for studying power, models which require costly and difficult data collection. Jack Nagel (1975) wrote a concluding work in this line of analysis, in which power-as-causation was stated as path analysis, a form of statistics.

2. *Power and Policy-making*

March (1966) and Riker (1964) were both noted for asking the question in respect to political analysis: 'what is the power of power?' Political scientists answered this question, not by flocking to do more studies of the holders of wealth and status, but by stating models of power structure applicable to political systems, both national and local. Dahl (1961) had pressed forward the definition of power as causation in the course of his empirical refutation of the power elite model (Mills 1956), as applied by Floyd Hunter (1953) to the city of Atlanta. Such writers were among those who described power by describing characteristics and activities of an elite group, whose actual exercise of power usually went undemonstrated. Scholars such as Dahl became known as stating a *pluralist* model of power, finding power to be fragmented among various scopes and domains, as opposed to the concentration of power within a single elite.

Scholars such as Edelman (1964), McConnell (1966), and Lowi (1979) agreed with Dahl that political power in America is fragmented, but argued that stable, unrepresentative elites often controlled policy within specific scopes of power—such as setting grazing fees on federal lands, developing nuclear power plants, setting airline and interstate trucking schedules and fees, and so forth. Such analyses, which flourished briefly, were also based on case studies of policy-making, and constitute a school which might be termed *multiple-elitist*, finding not a single power elite but hundreds of elites, each specific to a particular scope of action and not subject to control by elected officials. But multiple-elitism in turn faced a revisionist interpretation, *neopluralism*, as found in studies by James Q. Wilson (1980) and Jack Walker (1991), for instance, who found that while power within specific scopes of policy was normally unequal, it was nevertheless shared by so many organized interest groups, politicians, and governmental agencies that even within narrowly defined policy scopes, stable and controlling elites are seldom found.

The power-as-causation model was developed from 1953 to 1975, and the examples cited in this literature were almost always of power among individual persons. But given the great difficulties in gathering such data, political scientists seldom conducted such studies, and instead the analysis of political power flourished with the development of models of public policy-making just cited. Such scholars saw no need to worry about possible criticisms of ignoring the power bases of wealth and status, and continued to conduct case studies of policy-making, following Dahl's example, even as many argued with Dahl's pluralist model. Such case studies were not conducted according to the canons of path analysis or other forms of statistics; instead researchers used historical materials, interviews, and current documents to ascertain which groups, leaders, and government agencies caused changes in the course of public policy. Relatively simple empirical methods, assuming power-as-causation, have been enough to generate interesting models of power structure, but perhaps more complex methods will be needed to resolve differences among various schools of political power analysis.

The study of political power has developed satisfactorily since Simon's seminal article in 1953, but a great theoretical difficulty remains—the problem of *hegemony*. Hegemonic writers stress the great importance of the power to determine which issues are on the political agenda of a political system (Lukes 1974). Community power theorists Bachrach and Baratz (1962) argued that there are 'two faces of power,' one of which is power over the formation of political issues, and the other is the overt activity studied by Dahl. In arguing for the understanding of hegemony, the Marxist Antonio Gramsci argued (1971) that capitalism maintained its political power through socialization of persons into supportive values and norms, as much as through the machinations of 'bourgeois government.' Social values shape the preferences of the actors in their search for political power, but how can the scholar make empirical statements about one set of values displacing perhaps hundreds of conceivable alternative value systems? This seems to presuppose a comprehensive theory of social relations. Hegemony can be distinguished from the power of *anticipated reactions*, power exercised when B anticipates A's intentions, and acts according to them, perhaps fearing the costs of A's overt coercion. Republican members of Congress may have considered blocking appropriations for a government department, but anticipating President Clinton's effective reactions, they think better of the idea and do little but discuss the matter. But since the matter received some discussion, the theorist of political power may make an empirical observation. Similarly, in international relations, we speak of the deterrent power of weapons, even if these are never used. From the behavior of those deterred, scholars can find evidence that some probability of warlike behavior by one

nation was deterred (a change of behavior) through the possession of weapons by the other nation.

See also: Agendas: Political; Authority, Social Theories of; Causation (Theories and Models): Conceptions in the Social Sciences; Class: Social; Community Power Structure; Migration: Sociological Aspects; Power in Society

Bibliography

Bachrach P, Baratz M 1962 Two faces of power. *American Political Science Review* **56**: 947–52

Dahl R A 1957 The concept of power. *Behavioral Science* **2**: 201–15

Dahl R A 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT

Edelman M 1964 *The Symbolic Uses of Politics*. University of Illinois Press, Urbana, IL

Gramsci A 1971 Prison notebooks. In: Hoare Q, Nowell-Smith G (eds.) Selections from the Prison Notebooks of Antonio Gramsci. International Publishers, New York

Harsanyi J C 1962 Measurement of social power, opportunity costs, and the theory of two-person bargaining games. *Behavioral Science* **7**: 67–80

Hunter F 1953 *Community Power Structure*. University of North Carolina Press, Chapel Hill, NC

Lasswell H D, Kaplan A 1950 *Power and Society*. Yale University Press, New Haven, CT

Lowi T 1979 *The End of Liberalism*, 2nd edn. Norton, New York

Lukes S 1974 *Power: A Radical View*. Macmillan, London

March J G 1955 An introduction to the theory and measurement of influence. *American Political Science Review* **49**: 431–51

March J G 1966 The power of power. In: Easton D (ed.) *Varieties of Political Theory*. Prentice-Hall, Englewood Cliffs, NJ

McConnell G 1966 *Private Power and American Democracy*, 1st edn. Knopf, New York

McFarland A S 1969 *Power and Leadership in Pluralist Systems*. Stanford University Press, Stanford, CA

Mills C W 1956 *The Power Elite*. Oxford University Press, New York

Nagel J H 1975 *The Descriptive Analysis of Power*. Yale University Press, New Haven, CT

Polsby N W 1980 *Community Power and Political Theory*, 2nd edn. Yale University Press, New Haven, CT

Riker W H 1964 Some ambiguities in the notion of power. *American Political Science Review* **58**: 341–9

Simon H A 1953 Notes on the observation and measurement of power. *Journal of Politics* **15**: 500–16

Simon H A 1957 *Models of Man*. Wiley, New York

Walker Jr J L 1991 *Mobilizing Interest Groups in America*. University of Michigan Press, Ann Arbor, MI

Wilson J Q (ed.) 1980 *The Politics of Regulation*. Basic Books, New York

A. S. McFarland

Practical Reasoning: Philosophical Aspects

Practical reasoning is reasoning that aims to resolve the question of what one is to do. Such reasoning is practical in at least two senses. First, it is practical in its subject matter, insofar as it is concerned with human action. But it is also practical in its issue, insofar as reflection about action can itself directly move people to act. Reasoning of this kind raises two sets of philosophical problems. One concerns the question of how reflection can succeed in being practical in its issue. What assumptions must be made both about agents and about the processes of reasoning they engage in to make sense of the fact that such reflection can give rise to action? Can justice be done to this fact while preserving the idea that practical reasoning is genuinely a form of reasoning? A second set of issues concerns the content of the standards that are brought to bear in practical reasoning. Which norms for the assessment of action are binding on us as agents? Do these norms provide resources for critical reflection about our ends, or are they exclusively instrumental? Under what conditions do moral norms yield valid standards for reasoning about action? The first set of issues is addressed in Sects. 1–3, while Sects. 4 and 5 cover the second set of issues.

1. Practical and Theoretical Reasoning

Practical reasoning takes place from a distinctive standpoint. When agents deliberate about action, they think about themselves and their situation in characteristic ways. What are some of the salient features of the practical point of view?

One way to interpret this point of view is to contrast it with the standpoint of theoretical reflection. Theoretical reflection is sometimes understood as reasoning about questions of explanation and prediction. Looking backward to events that have already taken place, it asks why they have occurred; looking forward, it attempts to determine what is going to happen in the future. In these ways, theoretical reflection is concerned with matters of fact and their explanation. Furthermore it treats these issues in impersonal terms that are accessible (in principle) to anyone. The natural and social sciences can be viewed as giving paradigmatic expression to theoretical reasoning, so construed.

Practical reasoning, by contrast, takes a distinctively normative question as its starting point. It typically asks, of a set of alternatives for action, none of which has yet been performed, what one ought to do, or what it would be best to do. It is thus concerned not with matters of fact and their explanation, but with matters of value, of what is to be done. In practical reasoning

agents attempt to assess and weigh their reasons for action, the considerations that speak for and against alternative courses of action that are open to them. Moreover, they do this from a distinctively first-personal point of view, one that is defined in terms of a practical predicament in which they find ourselves (individually or collectively—people sometimes reason jointly about what they should do together; see *Joint Action*).

There is, however, a different way of understanding theoretical reasoning, stressing the parallels rather than the differences between the two forms of reflection. According to this interpretation, theoretical reflection is concerned with a normative rather than a factual question, namely with the question of what one ought to believe. It attempts to answer this normative question by assessing and weighing reasons for belief, the considerations that speak for and against the particular conclusions one might draw about the way the world is. Furthermore, it does this from a standpoint of first-personal reflection: the stance of theoretical reasoning in this sense is the committed stance of the believer, not the stance of detached contemplation of one's beliefs themselves (Moran 1997). Seen in this way, the contrast between practical and theoretical reflection is essentially a contrast between two different systems of norms: those for the regulation of action on the one hand, and those for the regulation of belief on the other.

Theoretical reflection, interpreted along these lines, addresses the considerations that recommend accepting particular claims as to what is or is not the case. That is, it is reasoning with an eye to the truth of propositions, and the reasons for belief in which it deals are considerations that speak in favor of such propositions' being true, or worthy of acceptance. Practical reasoning, by contrast, is concerned not with the truth of propositions but with the value of actions. The reasons in which it deals are considerations that speak in favor of particular actions being good, or worthy of performance in some way. This difference in subject matter corresponds to a further difference between the two kinds of reasoning, in respect of their consequences. Theoretical reflection about what one ought to believe produces changes in one's overall set of beliefs, whereas practical reason gives rise to action; it is, as noted above, practical not only in its subject matter, but also in its issue.

Two observations are in order about this side of practical reasoning. First, the contrast just drawn might suggest that there is a categorial difference in the effects of theoretical and practical reasoning, insofar as the former produces changes in our mental states, whereas the latter gives rise to bodily movements. But this would be misleading. Practical reasoning gives rise not to bodily movements *per se*, but to intentional actions, which are intelligible as such only to the extent they reflect our mental states. It would thus be more accurate to characterize the consequences of both theoretical and practical reasoning as *attitudes*, noting that the former leads to modifications of our beliefs, whereas the latter leads to modifications of our intentions (Harman 1986, Bratman 1987). Second, in neither case do the modifications in question occur infallibly. There is room for irrationality both in the theoretical and the practical sphere, which in its strongest form involves a failure to acquire the attitudes that one acknowledges to be called for by the reasons one has reflected on. Thus a person might end up watching another hour of television, while judging that it would be better on the whole to go back to work. Practical irrationality of this latter kind is known as *akrasia*, incontinence, or weakness of will, and its possibility makes clear that practical reasoning is not automatically practical in its issue. A better way to represent the consequences of practical reasoning would be to say that it generates appropriate intentions to the extent an agent is rational (Korsgaard 1996; see *Irrationality: Philosophical Aspects*).

2. *Naturalism and Normativity*

The connection of practical reasoning with intentional action raises large questions about its credentials as a genuine form of reasoning. As noted above, intentional action is not mere bodily movement, but reflects a mental state of the agent's, viz. intention. To be in this kind of mental state is to have a goal or aim of some kind, which one seeks to realize through one's action. Intention seems in this respect to be strikingly unlike belief. The propositions that give the content of beliefs have a representative function; they aim to fit the way the world is, so that if one discovers that the world is not how one previously took it to be, one's belief will automatically be modified in the process. With intentions, however, matters are otherwise (Smith 1987). The intention to go shopping on Tuesday, for instance, is not a state that would or should be abandoned upon discovering that one has not (yet) gone shopping on Tuesday; rather the person with such an intention will ordinarily try to bring the world into alignment with the intention, by going shopping when Tuesday comes around. Intentions are in this way more like blueprints than like sketches of an already completed structure.

Reflection on this contrast between belief and intention has led some philosophers to ask whether practical reasoning might not be something of a misnomer. The concern, in a nutshell, is how processes of thought that count as genuine reasoning could by themselves generate states with the peculiar, nonrepresentative function of intentions. Reasoning seems to consist of cognitive operations, whereas intentions are distinctively noncognitive states insofar as they do not aim to reflect the way things happen to be in the world.

Expressivist accounts give voice to this skeptical attitude about practical reasoning. These accounts offer interpretations of the normative language that distinctively figures in practical reflection about action. As was seen in Sect. 1, such reflection is an attempt to assess an agent's reasons for acting in one way or another; conclusions about such reasons are characteristically expressed in claims about what it would be *good* to do, or about the actions that one *ought* to perform. According to the expressivist, however, such claims do not represent genuine cognitive achievements, judgments that are literally capable of being true or false. Rather they give expression to desires, sentiments, and pro-attitudes, the kinds of noncognitive state that motivate people to action. Only if normative assertions are understood in this way—so the expressivist contends—can sense be made of the capacity of practical reflection to generate states with the peculiar structure and function of intentions.

Expressivism in this form can be seen as a naturalistic interpretation of practical reasoning, one that is appropriate to the enlightened commitments of the modern scientific world view. It is naturalistic metaphysically, insofar as no assumptions need be made about the objective existence in the world of such questionable entities as values, norms, or reasons for action. If normative claims do not represent genuine cognitive achievements, then their legitimacy does not depend on our postulating a realm of normative facts to which those claims might correspond. It is also naturalistic psychologically, insofar as it promises explanations of intentional human behavior that are basically continuous with explanations of the behavior of nonhuman animals. In both the human and the nonhuman case, behavior is seen as the causal product of noncognitive attitudes, operating in conjunction with the creature's factual representation of how things are in its environment. The difference is just that humans have much more sophisticated linguistic methods for giving voice to their motivating noncognitive attitudes. Indeed, many contemporary expressivists would contend that these expressive resources are sufficiently powerful to explain the features of practical reflection that initially give it the appearance of a genuine form of reasoning (Blackburn 1998, Gibbard 1990).

Other philosophers remain unimpressed with these naturalistic explanations. One ground for dissatisfaction with them is the following. The expressivist strategy relies on an initial contrast between practical reflection and the genuine forms of cognitive activity characteristic of theoretical reasoning. There has to be some basic sense in which such discourse does not come up to the standards of authentic cognitive discourse in the literal sense, otherwise the contention that normative discourse is expressive rather than cognitive will lack any significant content. But the contrast between theoretical and practical reflection required for this purpose seems elusive. As was seen in Sect. 1, theoretical reasoning is no less a normative enterprise than practical reasoning. It concerns itself with reasons for belief: the evidence and other considerations that speak for and against particular conclusions about the way things are in the world. In this respect, theoretical and practical reasoning would seem equally problematic from the naturalistic perspective—assuming, that is, that it leaves no place for such normative considerations as reasons. But if naturalism calls in question the credentials of theoretical reason, it thereby undermines the contrast between genuine reasoning and noncognitive normative discourse on which expressivists themselves rely (see *Emotion and Expression*; *Intentionality and Rationality: An Analytic Perspective*).

3. *Reasons and Motivation*

The capacity of practical reasoning to generate intentional action divides even those philosophers who agree in rejecting the expressivist strategy discussed above. Such philosophers are prepared to grant that there are reasons for action, and to accept the cognitive credentials of normative claims about such reasons. But they differ in their accounts of the truth conditions of these normative claims. Two approaches may be distinguished. The first of these, frequently referred to as internalism, holds that reasons for action must be grounded in an agent's prior motivations (Williams 1981). According to this position, a given agent s can have reason to do x only if x-ing would speak to or further some element in s's 'subjective motivational set.' There must be some rational route that connects s's x-ing with the subjective motivations to which s is actually already subject; otherwise the claim that s has reason to x must be rejected as false or incoherent. Behind this internalist position lies the idea that practical reasoning is practical in its issue. Internalists contend that sense can be made of the capacity of such reasoning to generate new intentions only if it is conditioned by motivational resources that are already to hand. Reasoning about action is a matter of working out the implications of the commitments contained in one's subjective motivational set; in this sense, motivation is prior to practical reason, and constrains it. Externalists reject this assumption, contending that one can have reasons for action that are independent of one's prior motivations. Most of them agree that practical reasoning must be capable of generating new intentional actions, and hence of giving rise to motivation. They agree, in other words, that if agent s has reason to do x, it must be possible for s to acquire the motivation to x through reflection on the relevant reasons. But they deny that such reasoning must in any way be constrained by s's subjective motivations prior to the episode of reasoning. Prac-

tical reasoning is not exclusively a process of working out the implications of one's existing commitments; it is also an attempt to get clear about what it would objectively be good to do. Normative considerations of this kind are independent of one's prior motivations, and can open up new motivational possibilities (Parfit 1997).

Underlying this dispute are diverging approaches to the explanation of intentional action. Internalists are impressed by the differences between intentions and the cognitive states that figure in paradigmatic examples of theoretical reasoning. Pointing to these differences, they ask how practical reasoning can succeed in producing new intentions if it is not grounded in something of the same basic psychological type: a motivation or desire that is already part of the agent's subjective motivational equipment. Many externalists find this contrast between intentions and cognitive states overdrawn. They observe that we need to postulate basic dispositions of responsiveness to reasons to account for the capacity of theoretical reflection to generate new beliefs, and question why these same dispositions cannot explain the fact that practical reasoning is practical in its issue. Cognitive or not, intentions belong to the broad class of attitudes that are sensitive to judgments, and this may account for the capacity of practical reflection to give rise to new intentions (Scanlon 1998). A third possibility is that intentions are products of the will, a capacity for self-determination distinct from the dispositions that render theoretical rationality possible. Depending on how it is developed, the theory of the will may offer a different way of accounting for the practical consequences of practical reflection, without assuming that reasons for action are grounded in an agent's subjective motivations (Wallace 1999).

4. *Instrumental and Maximizing Rationality*

Among the substantive norms of practical reasoning, those of instrumental rationality appear least controversial. Instrumental rationality, in its most basic form, instructs agents to take those means that are necessary to achieve their ends. In the modern era, this form of rationality has widely been viewed as the single unproblematic norm of practical reasoning. The instrumental principle makes no assumptions about the prospects for rationally scrutinizing peoples' ends. Rational criticism of this kind would seem to presuppose that there are objective values, standards for assessment of ends that are independent from psychological facts about what people happen to be motivated to pursue. In keeping with the naturalistic attitude sketched in Sect. 2, however, it may be doubted whether such standards can be reconciled with the metaphysical commitments of contemporary scientific practice (see *Value Pluralism*). A world that is shorn of objective values or norms leaves no room for rational criticism of people's ends, but only for Weberian *Zweckrationalität*: the rational determination of means to the realization of ends that are taken as given, as a matter of psychological fact.

It is important to bear in mind, however, that instrumental rationality is itself the expression of an objective normative commitment. The instrumental principle tells us that we have *reason* to take the means that are necessary to achieve our ends; if the principle is valid, then people are open to rational criticism to the extent they fail to exhibit this kind of instrumental consistency, regardless of whether they want to comply with the principle or not. If naturalism really leaves no place for objective norms or values, it may be wondered how an exception can possibly be made for the instrumental norm. The consistent naturalist should perhaps reject even *Zweckrationalität* in favor of a skeptical attitude towards practical reasoning in all its forms (Hampton 1998). Further questions can be raised about the plausibility of the instrumental norm as a self-standing principle of practical reason. The norm says that one should take the means that are necessary to achieve one's ends. But how can the fact that a given means is necessary, relative to some end, give a person reason to choose the means, if the end is not itself something it would be good to achieve in some way? The instrumental principle seems to function as a binding standard for practical reasoning only if it is taken for granted that there are additional, independent standards for the assessment of our ends (Korsgaard 1997).

One way to make room for the rational criticism of individual ends, without departing from the spirit of *Zweckrationalität*, is to expand one's view to encompass the totality of an agent's ends. Thus, even if there are no values that are ultimately independent of an agent's given ends, it would appear possible to criticize particular intrinsic desires by reference to others in the agent's subjective motivational set. An agent's desire for leisure, for instance, might be subordinated insofar as its satisfaction would frustrate the realization of other goals that are subjectively more important to the agent. Practical reasoning, it might be suggested, is a holistic enterprise, concerned not merely with the specification of means to the realization of individual ends, but with coordinated achievement of the totality of an agent's ends.

Many philosophers view this holistic perspective as the most fruitful way of conceptualizing the tasks of practical reasoning. It defines an important and difficult problem for practical reasoning to solve, without departing from the metaphysically modest assumption that there is no court of appeal for the rational criticism of an agent's ends that is independent of those ends themselves. The most sophisticated and influential expression of this holistic perspective is the maximizing conception of rationality. According to the maximizing conception, the fundamental aim of practical reasoning is to determine which course of

action would optimally advance the agent's complete set of ends. Thus it has been suggested that the rational action is the one whose subjective expected utility—reflecting both the utility of possible outcomes, from the agent's point of view, and the agent's beliefs about the probability of those outcomes—is the highest.

The maximizing conception of rationality finds its most complete expression in decision theory and in the theory of rational choice (as developed, for instance, in economics). These disciplines articulate with mathematical precision the assumption that one is acting rationally to the extent one does what is likely to bring about the best state of affairs, given both one's preferences and one's beliefs about the probable consequences of the alternatives that are open to one. Proponents of these theories often claim for them the additional advantage of empirical adequacy, contending that they are flexible enough to accommodate the full range of behaviors that human agents engage in, both within the marketplace and outside of it. Especially if one operates with the notion of 'revealed preferences'—preferences, that is, that are inferred solely on the basis of actual behavior—then virtually anything an agent might choose do to could be interpreted as an attempt to maximize expected utility. Decision theory, on this account of it, becomes an all-encompassing framework for the interpretation of human behavior, according to which agents are always striving to produce outcomes that would be optimal, relative to their preferences and beliefs (see *Rational Choice Explanation: Philosophical Aspects*).

If decision theory is interpreted in this way, however, then its relevance to the understanding of practical reasoning begins to seem correspondingly tenuous. Maximizing rationality initially looks to be a norm of practical reflection. In this guise, its appeal lies in the idea that there can be rational requirements on action, stemming from the totality of an agent's preferences and beliefs, even if it is not assumed that there are independent, substantive standards for the critical assessment of individual ends. But this normative understanding of maximizing rationality is tenable only if it is at least conceivable that individual agents might occasionally fail to comply with its requirements—an 'ought' that it is not so much as possible to flout is not really an 'ought' at all. Proponents of the maximizing approach need to be attentive to the latent tension between the empirical and the normative aspirations of their theory, for maximizing rationality cannot claim both perfect empirical adequacy and normative authority for practical deliberation.

Assuming this tension can be resolved, further questions arise about the plausibility of the normative requirement to maximize expected utility. Doubts have been raised, for instance, with regard to the assumption that it is necessarily irrational to fail to select that action that would be optimal, relative to one's preferences and beliefs. Perfectly rational agents often seem to be content with states of affairs that are 'good enough,' from the perspective of their aims and desires, even when they are aware of alternatives that promise a higher return; they 'satisfice' rather than seeking to maximize the value of the outcomes that they bring about (Slote 1989). They also treat their past intentions and plans as fixed constraints on their decision-making, rather than attempting to maximize subjective utility anew in every situation they confront (Bratman 1987). Some contend that maximizing rationality is flexible enough that it can accommodate alleged counterexamples of these kinds (Pettit 1984). If not, however, there may be grounds for doubting that it represents a fundamental and uncontroversial norm for reasoning about action.

5. *Specification, Expressive Rationality, and Moral Reasons*

If maximizing rationality is not the unproblematic requirement of practical reasoning that it initially seemed to be, what are the alternatives to it? Consider the assumption that critical assessment of an agent's individual ends is off-limits. This apparently truistic assumption has been questioned by some philosophers, who point out that a great many of one's aims in life are rather inchoate; people want, for instance, to be successful in their careers, and loyal to their friends, without being entirely clear about what these ends require of them. To the extent one's ends are indefinite in this way, they will not provide satisfactory starting points for instrumental, maximizing, or even satisficing reflection. One will need to specify such ends more precisely before one can begin to think about which means they require one to pursue, or to generate from them a rank-ordering of possible outcomes. Here is a possible task for practical reasoning that does not fit neatly into the categories of *Zweckrationalität*, however broadly construed (Wiggins 1987, Richardson 1994).

Practical reflection about ends is not an easy or well-defined activity. There are no straightforward criteria for success in this kind of reflection, and it is not even clear when it has been brought to a satisfactory conclusion. These considerations are no doubt partially responsible for the assumption—especially widespread in the social sciences—that there is no reasoning about final ends. On the other hand, how is one supposed to clarify one's largest and most important ends, if not by reasoning about them in some way? Rather than exclude such reflection because it does not conform to a prior and possibly inappropriate paradigm of proper reasoning, perhaps the conception of practical reasoning should be expanded to make room for clarificatory reflection about the ends of action. To do so would be to acknowledge that practical reasoning can be both objective and subjective at the same

time; by working out the meaning and implications of such commitments as loyalty or success, for instance, agents also help to get clear the values that define who they are (Taylor 1985; see *Hermeneutics, Including Critical Theory*).

A different point at which to question received models of practical reasoning is the theory of value to which they are committed. Maximizing and satisficing models alike seem committed to a consequentialist account of the relation of action to value: the thesis that value inheres ultimately in states of affairs, and that actions are to be judged rational to the extent they produce valuable states of affairs. This thesis seems at odds, however, with intuitive ways of thinking about value. It appears a distorted understanding of friendship, for instance, to say that what friends value fundamentally are states of affairs (involving, say, joint activities with the friend); what people value as friends are rather the concrete persons with whom they are befriended. This suggests a different way of conceptualizing the rationality of action. What makes actions rational, it has been claimed, is not the production of valuable states of affairs, but rather their success in expressing the attitudes that it is rational to adopt toward the true bearers of intrinsic value: people, animals, and things (Anderson 1993). A supposed advantage of this approach is its ability to explain the rationality of behaviors that seem intuitively sensible, but that are hard to fit into the consequentialist scheme (such as commitments deriving from one's past involvement in an activity or project, which can look like an irrational weighing of 'sunk costs' to the consequentialist; compare Nozick 1993).

Morality provides an especially fertile source of examples and problems for the theory of practical reasoning. One of the defining questions of moral philosophy is the question of the rational authority of moral norms: to what extent, and under what conditions, are people rationally required to comply with the demands of conventional morality? Reflection on this question has produced some of the most far-reaching and illuminating philosophical work on the nature of practical reasoning. Two divergent tend encies within this body of work should be singled out. Some approaches to moral reasoning proceed by relating it to patterns of reflection familiar from other, nonmoral domains, particularly the maximizing patterns canvassed in the preceding section. Thus it has been argued that, though morality imposes constraints on the direct pursuit of individual utility, these constraints can be justified in terms of the values expressed by ordinary economic rationality; a strategy of constrained maximization is recommended on grounds of enlightened self-interest, and this in turn accounts for the authority of moral norms to govern the practical reflection of individuals (Gauthier 1986). Other philosophers have sought to make sense of morality as a source of rational norms by assimilating it more directly to the maximizing conception. For instance, proponents of consequentialism as an ethical theory—the theory that agents are morally required to maximize the good, impartially conceived—consider it an advantage of this account that it represents moral requirements in terms of the maximizing conception of rationality familiar from other contexts (Harsanyi 1982; see *Consequentialism Including Utilitarianism*; *Economics and Ethics*).

A different approach stresses the discontinuities between moral and nonmoral patterns of reasoning. According to this approach, morality is a source of constraints (such as prohibitions on murder and deception) that cannot be represented accurately within the framework of maximizing rationality (for example, Scanlon 1998). If this is correct, then those who wish to make sense of moral requirements, as norms that appropriately govern the reflection of individual agents, will need to expand their conception of the forms and possibilities of practical reasoning (see *Ethics and Values*).

See also: Ethics and Values; Free Will and Action; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective.

Bibliography

Anderson E 1993 *Value in Ethics and Economics*. Harvard University Press, Cambridge, MA

Blackburn S 1998 *Ruling Passions*. Clarendon Press, Oxford, UK

Bratman M E 1987 *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA

Gauthier DP 1986 *Morals by Agreement*. Clarendon Press, Oxford, UK

Gibbard A 1990 *Wise Choices, Apt Feelings*. Harvard University Press, Cambridge, MA

Hampton J E 1998 *The Authority of Reason*. Cambridge University Press, Cambridge, UK

Harman G 1986 *Change in View*. MIT Press, Cambridge, MA

Harsanyi J C 1982 Morality and the theory of rational behavior. In: Sen A, Williams B (eds.) *Utilitarianism and Beyond*. Cambridge University Press, Cambridge, UK

Korsgaard C M 1996 Skepticism about practical reason. In: Korsgaard C M (ed.) *Creating the Kingdom of Ends*. Cambridge University Press, Cambridge, UK

Korsgaard C M 1997 The normativity of instrumental reason. In: Cullity G, Gaut B (eds.) *Ethics and Practical Reason*. Clarendon Press, Oxford, UK

Moran R 1997 Self-knowledge: discovery, resolution, and undoing. *European Journal of Philosophy* **5**: 141–61

Nozick R 1993 *The Nature of Rationality*. Princeton University Press, Princeton, NJ

Parfit D 1997 Reasons and motivation. *Proceedings of the Aristotelian Society* **77**: 99–130

Pettit P 1984 Satisficing consequentialism. *Proceedings of the Aristotelian Society* **58**: 165–76

Richardson H S 1994 *Practical Reasoning about Final Ends*. Cambridge University Press, Cambridge, UK

Scanlon T M 1998 *What We Owe to Each Other*. Belknap Press of Harvard University Press, Cambridge, MA
Slote M 1989 *Beyond Optimizing*. Harvard University Press, Cambridge, MA
Smith M 1987 The human theory of motivation. *Mind* **96**: 36–61
Taylor C 1985 What is human agency? In: Taylor C (ed.) *Human Agency and Language* Cambridge University Press, Cambridge, UK
Wallace R J 1999 Three conceptions of rational agency. *Ethical Theory and Moral Practice* **2**
Wiggins D 1987 Deliberation and practical reason. In: Wiggins D (ed.) *Needs, Values, Truth*. Blackwells, Oxford, UK
Williams B 1981 Internal and External Reasons. In: Williams B (ed.) *Moral Luck*. Cambridge University Press, Cambridge, UK

R. J. Wallace

Practice: Anthropological Aspects

The word practice carries a clear meaning in everyday English. A practice is a convention or institution, and in anthropological discourse the word often appears in its ordinary sense. Tibetans follow the practice of sky burial or South Indians practice cross-cousin marriage. But the word also enters anthropological discourse in more theoretized ways.

1. Sources

The idea that human action takes place against a background—constituted or tacit knowledge, *intersubjective meanings*, a form of life, or practice—has sources in Husserl, Wittgenstein, and Polanyi. But the notion has mostly influenced students of human behavior, from Merleau-Ponty to Michel Foucault and Pierre Bourdieu, by way of Heidegger's *Being and Time*. For Heidegger, the epistemological question—how do we know our understanding of the world matches the world itself—cannot be asked as Descartes did, by making the question of truth a matter of beginning with what a conscious subject can know of the world with absolute certainty and working from there. The Cartesian subject on Heidegger's account can know that he or she exists only because of the presence of a set of background practices which make knowing (or acting, judging, laughing, and so on) possible. The intelligibility of the world in other words depends on the actor's prior 'being-in-the world.'

Heidegger argues that understanding human action requires recognizing that it is practical twice over—it depends on this background of practices and it is purposive. In Heidegger's language *intentionality* is always practical rather than cognitive, but practices are purposive in an extended sense of the word. The actor may not have any purpose in mind because being-in-the world is a skilled activity that goes on whether the actor is aware or not (1962, p. 405). Intrinsic to human activity is the actor's comportment with things in the world, and everyday being resides in his or her doing, intending, and coping. Human beings are socialized into these practices relative to their historical moment and culture, yet the practices themselves cannot be made explicit. If such is the case, the idea of practice challenges hopes for a scientific account of behavior.

Mainstream political science, for instance, assumes that the *actor's point of view* is a thing in itself. It consists of beliefs, attitudes, and evaluations. These states of mind are an objective social reality and capable of furnishing the brute data on which the political scientist constructs a science of politics. The problem is that the actor's point of view cannot be reduced to that which can be attributed to him as a belief, attitude, or evaluation if the content of consciousness depends on intersubjective belief or practices that stand behind these mental states and make them possible (Taylor 1971). The intersubjective beliefs Japanese villagers hold about reaching public consensus, Taylor says, must be understood as residing in the relationship between actors and the society they inhabit. Critics of *positivism* argue that no amount of questioning could elicit these intersubjective beliefs from the Japanese themselves because they are ineffable and depend on still other practices. Yet without accounting for the bottomless matrix of practices in which those beliefs, attitudes, and evaluations are embedded, no comparative science of politics is possible.

2. Habitus, Practice, and Time

Another example of Heidegger's influence is Bourdieu's use of the word practice in the sense of *praxis*, *action*, or *agency*. While Taylor reacts to the assumption that social action can be explained by focusing on an actor's consciousness, Bourdieu denies that it can be understood as consciousness reduced to rules and norms. In his view, action is relational, recursive, and reflexive. Behavior has a cultural referent, but is considerably more complicated than the execution of a blueprint. Bourdieu's attempt to accommodate strategy and innovation as part of an account of social behavior appeared as nearly as his *Algeria 1960* which he organized around the distinction between dispositions and economic structures (English trans. 1979). He theorizes that relationship in a full-blown way in his *Outline of a Theory of Practice* (1977). Dispositions, Bourdieu writes, have two parts—the *habitus*—the ensemble of dispositions towards action and perception that operates from within social agents—and the practice which the habitus generates. The habitus is 'the durably installed generative principle of regulated improvisations' (1977, p. 78). Those structured, objectively unified practices

are subjective and changing, but they follow from the habitus in the same way a jazz band improvizes on a familiar melody.

Where a traditional anthropological account of kinship focused on genealogical rules and the constraints those rules exert on some activity such as marriage choice, Bourdieu saw genealogical forces in relation to other variables (1977). To regard Kabyle marriage as the realization of a parallel cousin rule for arranging a match is to remove one source of determination from the relationship in which it operates. Any calculation that an anthropologist might make of how often Algerians marry their parallel cousins depends on the extent of the social unit in relation to which the calculation is made. The size of that social unit is itself a strategic variable (1990, p. 15). Other variables pertain to the couple (age and age difference, previous matrimonial history, sibling order, relationship to the family authority-holder), but a great many more are family matters (the economic and social status of the families, their social affliations, ability to defend themselves, as well as sheer numbers). For Kabyles the habitus includes parallel cousin marriage as one disposition among many, and the strategies that emerge from all of those dispositions—weighing each in relation to others—constitute practice.

Families strategize by considering all of these considerations, and as they learn more about the other side, their strategy evolves. Because human behavior is recursive as well as relational, its analysis must take *time* into consideration. Bourdieu insists that detemporalizing science fails to understand what constitutes social behavior. Time gives behavior 'its form, as the order of succession, and therefore its direction and meaning' (1990, p. 98). Where Lévi-Strauss thought the gift exchange was organized according to deep principles unknown to the actor, Bourdieu argued that it was driven by 100 small calculations, idiomatic to actors who re-evaluate and change their strategies as the situation unfolds in time. Those calculations may be intuitive, but theory misconstrues its object when it takes the actor's agency as simply applicative.

There is a third part of practice where Bourdieu turns Heidegger in a sociological direction. Practices are *reflexive*. They carry with them their own interpretation, which means that they organize the relationship between individuals and society in a way that is circular. Heidegger puts the matter in distinctive language—existence is an understanding of being shared by individuals and society. Because practices are shared, they are the basis of mutual intelligibility, but this same condition means that there is never anything like a subjective inner sphere (1977, pp. 80–1). Individuals, whatever their self-conception, are social creatures. Individuality cannot be absolute, but a variant on the local class habitus and other dispositions.

In this sense Bourdieu blurs the distinction between the individual and society. 'The habitus—embodied *history*, internalized as a second nature and so forgotten as history—is the active presence of the whole past of which it is a product. As such, it is what gives practices their relative autonomy with respect to external determinations of the immediate presence. This autonomy is that of the past, enacted and acting, which …, makes the individual agent a world within a world' (1990, p. 56). *Symbolic interactionists* and *ethnomethodologists* assume that social action derives from individuals responding to immediate circumstances, and Bourdieu too wants to account for circumstances. But he does so by insisting that the actor's course is guided by a dialectic. He is always a social being subject to the orientations and limits of the habitus' 'operations of invention' as they develop amidst those objective structures (1977, p. 95).

That practices are reflective of larger social understandings (of what it is to be a person, a teacher, a woman) has implications for the way Bourdieu understands agency: 'it is because subjects do not, strictly speaking, know what they are doing that what they do has more meaning than they know' (1977, p. 77). It also has implications for the way Bourdieu understands agency as organized in time. He says that gift exchange is one of the social games that cannot be played unless the players ignore 'the objective truth of the game … and unless they are predisposed to contribute, with their efforts, their marks of care and attention, and their time, to the production of collective misrecognition. Everything takes place as if the agents' strategies and especially those that play on the tempo of action … were organized with a view to disguising to themselves and from others the truth of their practice, which the anthropologist brutally reveals simply by substituting the interchangeable moments of a reversible sequence for practices performed in time and in their own time' (1990, pp. 105–6).

3. *Practice and Domination*

The notion that the actor's point of view entails his or her misrecognizing the social game derives from a tradition of concern for consciousness that has affinities to both Marx and Heidegger. What Bourdieu takes from Marx (and what separates him from Heidegger) is the idea of structure, and specifically the economic structures in which actors act. The world is prior to the actor's agency in the world, but it bears on individuals relative to their position in the social organization of production. It is the objective character of social conditions that produces a class habitus, the harmony of dispositions, ethos, taste, and what he calls body *hexis*—while individuals are thoroughly social, they are related to society in ways appropriate to their peculiar identities.

Bourdieu has studied how relations of *domination* shake an actor's subjective aspirations in an extraordinary variety of contexts—from the Kabyles of

Algeria, peasants in the region of Bearn in southwestern France, the academic establishment, systems of classification in educational settings, museums, class relations, sports, to high fashion and the production of other cultural goods in modern industrialized societies. In its most cunning form domination operates in systems of classification. Whether he turns his attention to anthropological concerns with kinship, ritual, calendars, and domestic space or more sociological matters such as fashion, taste, and educational systems of classification, Bourdieu dismisses Kantian and Lévi-Straussian assumptions. All such classifications are not folk epistemologies but ideological mystifications of political realities.

Because Bourdieu assumes the nondistinction of the material and the symbolic, he rejects 'the spurious alternatives of social physics and social phenomenology' (1990, p. 140). In the same way that charisma is not a particular form of power, contra Weber, but the condition of all power, meaning grounds all human action. In studying *taste* as a discriminating practice in modern France, Bourdieu draws the study of culture in the normative sense of refinement or cultivation into the realm of culture understood anthropologically (1984). The endless conversion of economic capital into *symbolic capital* uses taste to enunciate class differences but requires the complicity of the entire society. That conversion is the condition which makes possible the continuation of domination, yet it 'cannot succeed without the complicity of the whole group … As Mauss puts it, the whole society pays itself in the coin of its dream' (1977, p. 195).

4. Practice and its Applications

Seen against the background of anthropological theory, practice theory was a response to the profession's recent discontent with the notion of culture, especially its neglect of power and strategy. Seen against a Heideggerian background, it was another critique of the error with which Western philosophy began, objectivism (1962, pp. 41–9). Anthropologists found much to admire in Bourdieu's work—his even-handed concern for both the actor's point of view and structure, efforts to transcend the conventional antimonies, and recognition of temporality as figuring in social life. Exploiting practice theory as a way to tie together consciousness and structure, action and its temporal horizon, as well as Weber and Marx, theorists have applied Bourdieu's thinking to case studies that range from gift-giving, saving, and consuming to gender, sexual difference, class relations, and colonial encounters.

If one counted anthropologists invoking the importance of practice, temporality, or simply Bourdieu's name, his influence would be striking. Counting only scholars who fully pursue his research program, the number would shrink to people who publish in the journal he founded, *Actes de la recherche en sciences sociales*, or work with him at the Centre de Sociologie Européenne in Paris. The most vibrant tradition occupying the space between these extremes are anthropologists who have seen practice theory as a vehicle for coping with history and temporality. Marshall Sahlins' work on Hawaii and Fiji carries on Bourdieu's interest in social reproduction, here fixed on situations where European actors intruded on South Pacific systems of domination. In accounting for the circumstances that surrounded Captain Cook's arrival in Hawaii, Sahlins characterizes events in a way that melds agency and structure, showing how mutual misrecognition created changes that look in retrospect entirely predictable (1985). In dealing with Hawaiians's initally worshipful attitude to Cook and his eventual murder, Sahlins pushes Bourdieu's commitment to agency and structure in a historical direction, arguing that those incidents evolved in a way better understood not as misrecognition and disaster, but the reproduction of culture—Hawaiians thought Cook was their god Lono—and its transformation. Sherry Ortner melds practice theory, structure, and history in characterizing the relationship of 'big' people and 'small' people and monks and nuns in Sherpa society and relations among that society, the Nepal state, and the British Raj (1989). Her confidence in the Bourdieuan project also organized her important article evaluating the state of anthropological theory (1984).

Arjun Appadurai's attempt to theorize globalization has even more inspirations, but Bourdieu's work is critical. Approaching consumer capitalism as a system as much symbolic as material, Appadurai extends the analysis in several directions. He understands consumption as a transnational phenomenon, which finds its locus at the level of body hexis, the individual, class, and nation-state. Stressing the way the media have unleashed the imagination, Appadurai argues that in the transnational world the habitus has been replaced by self-conscious choice, justification, and representation, often intended for far-flung diasporic communities and capitalist frontiers (1996, p. 44). But the role agency, strategy, and innovation play in Appadurai's account of local societies negotiating modernity would strike Bourdieu as the operations of the habitus under new circumstances.

See also: Interactionism: Symbolic; Phenomenology in Sociology; Phenomenology: Philosophical Aspects

Bibliography

Appadurai A 1996 *Modernity at Large*. University of Minnesota Press, Minneapolis, MN

Bourdieu P 1977 *Outline of a Theory of Practice* [trans. Nice R]. Cambridge University Press, Cambridge, UK

Bourdieu P 1979 [1963] *Algeria 1960* [trans. Nice R]. Cambridge University Press, Cambridge, UK

Bourdieu P 1984 *Distinction* [trans. Nice R]. Harvard University Press, Cambridge, UK
Bourdieu P 1990 *The Logic of Practice* [trans. Nice R]. Stanford University Press, Stanford, CA
Heidegger M 1962 *Being and Time*. Harper and Row, New York
Heidegger M 1981 *Basic Problems of Phenomenology* [trans. Hofstadter A]. Indiana University Press, Bloomington, IN
Ortner S B 1984 Theory in anthropology since the sixties. *Comparative Studies in Society and History* **26**(1): 126–66
Ortner S B 1989 *High Religion: A Cultural and Political History of Sherpa Buddhism*. Princeton University Press, Princeton, NJ
Sahlins M 1976 *Culture and Practical Reason*. University of Chicago Press, Chicago
Sahlins M 1985 *Islands of History*. University of Chicago Press, Chicago
Taylor C 1971 Interpretation and the sciences of man. *Review of Metaphysics* **25**(1): 3–51

S. Kemper

Pragmatics: Linguistic

Pragmatics is the study of the relationship between the structure of a semiotic system (notably language) and its usage in context, and, along with semantics (see *Semantics*), forms part of the general theory of meaning. Within the theory meaning, pragmatics is concerned especially with implicit meaning, with inference and the unsaid, and with the way in which language structure trades on this background of the presumed and the inferred. Pragmatics has come to play an important part in general linguistic theory, partly because it has substantial intrinsic subject matter, partly because it promises explanations for other linguistic phenomena, and partly as a response to over-idealization in contemporary grammatical theory (see *Generative Grammar*). It has also become an area of interdisciplinary concern, with fundamental contributions from philosophy of language, linguistics, psychology, and the sociology of language.

1. History

The use of the term 'pragmatics' in modern semiotics derives from the philosophical work of C. S. Peirce and R. Carnap, reflected in Morris's (1938) three divisions of semiotics, the study of sign systems: *syntax*, which investigates the relation of signs to signs; *semantics*, which investigates the relation of signs to the things referred to; and *pragmatics*, which studies the relation of signs to their users. Since then, the usage of the term has bifurcated into a broad use, which subsumes sociolinguistics (see *Sociolinguistics*) and discourse analysis (see *Conversation Analysis: Sociological*, and *Discourse, Anthropology of*), and a narrower use (associated especially with philosophy of language and approaches to the study of meaning which derive from it) in which pragmatics deals with those aspects of meaning that are systematically context-dependent.

2. Scope

Contemporary linguistic pragmatics is focused on a number of special relations between linguistic meaning and context. On the narrower scope of pragmatics, concerned with context-dependent meaning, the following topics have come to be central: deixis, presupposition, speech acts, implicature, and conversational inference, which are discussed briefly in turn.

2.1 Deixis

Deixis, from the Greek word for 'pointing or demonstrating' (the equivalent philosophical term 'indexicality' is from the corresponding Latin) is the phenomenon whereby some linguistic expressions are systematically dependent on the context for their interpretation. Take, for example, the utterance 'Put this book over there'—which book is being referred to, and in which place it is to be put, are determined by features of the context outside the utterance itself, here typically gestures: *this* and *there* act as place holders for contextually-specified parameters. A good test for whether an expression is deictic or not is whether an utterance in which it plays a part would be equally true (or false) regardless of speaker, addressee, time, or place of utterance: thus 'The square root of nine is three' has no obvious deictics, whereas 'We are coming soon' has only deictic expressions.

At first sight, deixis seems to be a simple phenomenon, a leftover from the direct 'here and now' relevance of animal communication systems. But the intersection of this context-dependence with the property of abstract symbolic representation in language (Hockett's displacement feature of language design) leads to deep complexities, and the phenomena turn out to be very puzzling, both philosophically and psychologically, for deixis introduces context dependency into almost every utterance (in English, for example, nearly all sentences are tensed, and tense is deictic, as in 'It is Tuesday' where *is* locates the reference time as today). One of the central philosophical puzzles is that deixis makes possible self-reference in utterances, thus introducing, for instance, many paradoxes of the 'Cretan liar' type: 'This sentence is false' is true if and only if it is false. One of the central psychological puzzles is that deixis makes clear that there is a mismatch between what a sentence means and the thought corresponding to its utterance. For example, what exactly is the thought corresponding to the utterance 'Today is Tuesday'? It is obviously not 'Tuesday is Tuesday,' nor (if the day of speaking is

31st December 2035) '31st December 2035 is a Tuesday' since the speaker may have no idea of the date. Closer might be 'The diurnal span in which I am now speaking falls on the calendrical second day of the week'—but now we have paraphrased one deictic term (*today*) in terms of others (*I*, *now*). And the puzzle goes on.

Languages incorporate deictic context-dependency in many different places in their grammars and lexica, but it is conventional to consider four main notional parameters.

First, there is *person deixis*, reflected in the traditional grammatical categories of first, second, and third person, where first and second person refer to members of the speech event (current speaker and current addressee, respectively) in contrast to the third person. These categories are reflected in pronoun systems, verbal agreement and elsewhere. Many languages make unfamiliar demands here: for example, in Southeast Asian languages such as Korean or Japanese have different word-forms for referring to entities such as kinsfolk, houses, or food, according to the relative status of speaker and addressee. Or in some Australian languages (such as Dalabon) pronouns make reference to whether the speaker and the referent, or the referents in the case of plural third person pronouns, are one vs. two generations apart genealogically. This encoding of the social relation between speaker, addressee, and third-party referents is often recognized as a separate deictic dimension, called *social deixis*.

A second important parameter is time. *Temporal deixis* is manifested directly in English in words such as *now*, and, as mentioned above, in tense. True tenses locate time relative to the moment of speaking, so that a past tense is used for events (or situations) preceding it, present for events including it, and future for events succeeding it. Not all languages have tenses (e.g., Malay or Chinese), and not all the grammatical categories that grammarians call tenses are truly deictic. Some languages (e.g., Bantu or Papuan languages) have up to six deictic tenses, however (specifying, for instance, that something happened the day before yesterday). There are many other manifestations of temporal deixis, from calendrical specifications like *yesterday* to more covert items like *ago* (as in *nine years ago*) or *Good night* which presumes that it is night and that the speaker is leaving.

A third parameter is *spatial deixis*, reflected in demonstratives such as *this* and *that*, or adverbs such as *here* and *there*, where spatial locations are indicated by reference to the place of speaking. Demonstrative systems vary greatly across languages, with some making distance from speaker primary, others making a contrast between 'this near speaker' vs. 'that near addressee,' and others invoking attentional issues, so that one has a contrast between, for instance, 'this we are talking about' vs. 'that to which I want to draw your attention.' Huge arrays of demonstratives are found in some languages (e.g., Inuit) where these sorts of parameters intersect with, for example, the shape of things referred to. Languages often have pairs of verbs, such as *come* vs. *go*, *bring* vs. *take*, which are often interpreted according to whether the motion is toward or away from the place of speaking. Many other expressions in language are covertly place-deictic. For example, 'the cat is behind the tree' implies that the tree is between the speaker and the cat, and 'Let's go to the local pub' implies that the pub is near to the place of speaking.

Lastly, *discourse deixis* involves the possibility of referring from one utterance backward or forward to others, as in 'That was a good speech' vs. 'It made a sound like this: goooo.' Discourse deixis grades into anaphora, the use of expressions such as pronouns to refer to entities introduced earlier, usually by fuller expressions. It is also closely related to *definiteness*, as in *the man* vs. *a man*, since all these phenomena rely on mental models of entities introduced into the discourse.

Considerable effort has gone into understanding how deixis works, and how it can be brought within a formal theory of meaning. But for the most part theoretical models underestimate the complexity and pervasiveness of deictic phenomena, and the richness of the contextual systems that support them. For example, good studies of actual deictic usage are rare, and the study of gesture is in its infancy (McNeill 2000).

2.2 Presupposition

Presupposition is a second major topic in pragmatics, and concerns the way in which propositions already presumed in a discourse context are usually not stated or questioned, but encoded in a more 'background' way. For example, 'Has he stopped bothering you?' presupposes the proposition that you and I know that he has been bothering you, and asks whether this has stopped. The classical test for presupposition is survival under negation: 'He hasn't stopped bothering me' and 'He has stopped bothering me' both presuppose that he was bothering me. Languages have complex systems for foregrounding and backgrounding information in this way. Thus mental attitude verbs, such as *know* or *regret*, or change of state verbs such as *start* and *stop*, presuppose their complements, definite descriptions (such as 'the King of Buganda') presuppose the existence of the entities referred to, iteratives like *again* (as in 'John did it again') presuppose earlier occurrences, and so on. The phenomena and corresponding explanations are complex, and are dealt with in a separate article (see *Linguistic Presupposition*), but the relevance to pragmatics is that presupposition implies clearly that natural languages are built to trade on, and signal, the dependency of utterances on propositions already taken for granted.

The pragmatic aspects of the phenomena are often underplayed in semantic accounts of the phenomena. For example, consider the sentence 'Sue cried before she finished her thesis'. This would normally presume that she finished her thesis, this being a presupposition from the *before*-clause. But the minimally different sentence 'Sue died before she finished her thesis' seems to make no such presumption, because of course we happen to know that the dead do not complete theses. This *defeasibility*, or cancelation of an inference in the context of contrary assumptions, is a hallmark of pragmatic inference.

2.3 Speech Acts

A third major topic in pragmatics concerns the general force or point of utterances, and is usually discussed under the rubric of speech acts. We can take a simple underlying proposition such as 'The wine is (put) on the table,' and we can embody it in utterances with different *illocutionary force* (in the terminology of the philosopher J. L. Austin). We can assert or vouch for the truth of the proposition, as in the assertion 'The wine is on the table.' We can ask whether it is the case, as in 'Is the wine on the table?' We can demand that the proposition be made true, as in 'Put the wine on the table,' and we can forbid anyone making it true, as in 'Do not put the wine on the table.' Languages typically indicate a handful of such illocutionary forces in the grammar of sentence types, as here in the English unmarked statement order, the marked inversion of yes–no questions, or the covert subjects of imperatives. Beyond statements, questions, and imperatives, most languages have in addition minor sentence types, e.g., exclamatives, as in 'What a wonderful dress,' or hortatives like 'Let's start,' or imprecatives like 'Damn you!' and other more exotic types (e.g., special forms for blessings, curses, warnings). Thus the grammars of languages seem to presume a set of fundamental uses for language, and there has been some philosophical and theoretical interest in these categories. However, actual usage is much more complicated than this discussion would suggest. For example, one can use the statements 'I suppose the wine is on the table' to question whether it is, or 'I would like the wine on the table' to request that it be put there, or 'The wine should not be placed on the table' to forbid it. Some of these uses can be explained as secondary or derivative uses of one sentence type to perform a speech act prototypically associated with another. For example, *I would like* … could be said to indirectly hint, rather than directly request, and can thus be described as an indirect speech act. What this discussion makes clear is that it is important to keep apart the grammar of sentence types from the illocutionary forces or speech acts they can be used to perform.

There has been much work on grammatical sentence types (and what the grammarians call *mood*), speech acts, and indirect speech acts, and their relationship to one another. This is the subject of a separate article, but here one major point is that these phenomena indicate directly that natural languages are designed not just as abstract systems, but as tools for human communication, a point of view underlined in functional approaches to language structure (see *Grammar: Functional Approaches*). Speech acts can also be seen to be fundamentally context dependent. First, speech acts are dependent on, and contribute to, the context in which speech is taking place. One way of seeing this is to construe speech acts as operations on the context (conceived of as sets of propositions taken for granted): a statement adds a proposition to the context, a question requests whether such a proposition is to be added, a denial removes one, and so on. Note too how a permission presumes that a prohibition would otherwise be in force. Second, the interpretation of a sentence as performing a specific speech act is obviously context dependent. Consider the sentence 'What are you doing tonight?' in, say, the following discourse context where it has a straightforward informational query status:

A: 'I can't wait for this evening.'
B: 'What are you doing tonight?'

Now compare its role in a different context:

A: 'What are you doing tonight?'
B: 'Oh I don't know, not much.'
A: 'Would you like to come to the movies?'

Here the utterance checks out a condition on an invitation, and in so doing can be seen as preluding such an invitation (a 'pre-invitation'), which is then (given an appropriate response) expected to follow. Sequences of utterances thus set up specific expectancies, crucial in interpretation—a matter dealt with more fully below.

2.4 Conversational Implicature

So far we have considered ways in which the structures of languages build in presumptions of use, as in pronouns, demonstratives, verbs of presupposition, or sentence types as indicators of speech acts. In these cases, the semantics of the expressions trigger the pragmatic inferences, which are themselves subject to pragmatic resolution. But another very important aspect of pragmatics is concerned with inferences that are invited—or *implicated*—rather than required or triggered by the semantics: often, synonymous expressions will have different implicatures. The philosopher H. P. Grice noted that there are a number of different ways in which such invited meanings can be invoked, all of which point to a set of background assumptions about how language should be used, which he formulated as a set of maxims of conversation. He sketched four such maxims, with attendant submaxims: the maxim of Quality ('Say what you believe to be true'); the maxim of Relevance ('Make what you

say relevant and timely'); the maxim of Quantity ('Don't say more or less than is required'); and the maxim of Manner ('Be brief and clear'). Then he showed that if a speaker violates them overtly, the speaker is generally credited with some implicit, indirect meaning—for example, if it is raining, and the speaker says 'Thank God I didn't bother to bring my umbrella,' thereby clearly violating the maxim of Quality, an ironic interpretation is generated. (Any such interpretations which are suggested by reference to the maxims, Grice called *conversational implicatures* to distinguish them from logical implications). If on the other hand the speaker just seems to be following the maxims, then additional inferences are generated. For example, if you say 'There are five gallons in the tank,' I'll assume that there are no more than five gallons in the tank, because otherwise by the maxim of Quantity ('say as much as is required') you should have said so (note that if there are more than five gallons, then of course what you have said is certainly true, but is also misleading). Grice went on to note a number of interesting applications of his theory of implicature. For example philosophers had long disagreed about just how close the ordinary language words '*if*,' '*or*,' and '*and*' are to their logical counterparts. Grice argued that the divergence between ordinary language and logic was due largely to the implicatures of language usage. For example, if I say 'Glenn was an astronaut or he was a senator,' I suggest that he was not both, but this does not mean that natural languages *or* is not equivalent to logical (inclusive) disjunction. That is because if I knew he was both, I should (following the maxim of Quantity) have said so; since I didn't, I imply that I believe he was one or the other, but not both.

Although modern versions of implicature theory use somewhat different 'maxims,' they follow Grice's idea that the presumption of background principles of language use can generate many detailed inferences that are suggested but not entailed by what has been said. There are two main branches of contemporary theory. One takes the line (following the trend in the cognitive sciences) that these background principles are innate cognitive mechanisms of information processing. The other branch follows Grice more closely, and argues that these principles follow from rational design characteristics of communication. The former line is more generally concerned with the nature of inference in communication, while the second 'neo-Gricean' line of argumentation has been developed more narrowly to give explanations of linguistic facts, and is thus of more direct relevance to linguistic theory, and so is the focus of the following remarks.

One central attraction of the Gricean approach is that it promises to simplify the kinds of meanings we attribute to expressions. For example, as mentioned above, English *or* seems normally to have an exclusive use (as in 'The book is either in the bedroom or the sitting room'), but obviously this is not always so (consider 'I lent you my pen or my pencil, or both'). Grice's tactic is to avoid positing an ambiguity (between inclusive and exclusive *or*), by assuming that *or* has a wide, general meaning (e.g., the inclusive meaning), which is then specialized in context where appropriate by an extra pragmatic inference—namely the Quantity implicature 'not both.' By generalizing Grice's observations we can note, first, that any pair of expressions (such as *and*, *or*), where the use of one expression would entail the other (as in *p and q* implying *p or q*) but not vice versa, form a scale of informativeness—the use of *and* is more informative than *or*, because *p or q* allows not only for the circumstance that *p*, or that *q*, but also for the circumstance that *p and q*. In short, *p and q* rules out more alternative states of affairs. Then we can also note that, for any such pair of expressions, asserting the weaker or less informative member will implicate the inapplicability of the stronger expression (for if you had meant that, you should have said it). Hence asserting *p or q* implicates that one is not in a position to assert *p and q*. Now consider many other such ordered pairs, where each is of the form ⟨strong, weak⟩:

⟨*and*, *or*⟩, ⟨*all*, *some*⟩, ⟨*certain*, *possible*⟩,
⟨*four*, *three*⟩, ⟨*must*, *may*⟩, ⟨*hot*, *warm*⟩

For each of these, asserting the weaker, less informative expression will implicate that the stronger expression does not apply, as in 'Some of the students are punctual,' which, other things being equal, will implicate that not all of the students are punctual. Modern treatments of these so-called *scalar implicatures* are quite developed (Atlas 2000, Gazdar 1979, Horn 1989, Levinson 2000).

Many further systematic observations can be made. Consider the contrast between 'John could solve the problem' vs. 'John had the ability to solve the problem'—the first suggests that he did solve it, while the second suggests he didn't. The opposition is one between a direct lexical coding with the verb *can* and a more periphrastic expression *have the ability to*. More generally, the opposition is between any simple, direct encoding, which invites a rich interpretation to the stereotypical scenario, and a less direct, unusual, more prolix or *marked* expression, which suggests a complementary interpretation. Consider, for example, 'He went downtown and bought a coat,' which suggests a single complex action (a coat-buying expedition), vs. 'He went downtown and in addition he bought a coat,' which suggests two independent actions. These opposing tendencies of interpretation can be related to *iconicity* in language (see *Linguistics: Iconicity*), but what gives them reliability is mutual expectations between speaker and addressee—that is, that they follow from general principles or maxims.

Contemporary neoGricean theory recognizes two or three maxims: for example, a maxim of Quantity (or

Q-principle, giving us the scalar implicatures mentioned above); a maxim that maximizes information from direct, unmarked expressions (an Informativeness or I-principle); and a maxim of Manner or markedness (or M-principle) that curtails those informative interpretations (Levinson 2000). Three such principles suffice to capture many detailed inferences which interact closely with linguistic structure, and they can be thought of as stable heuristics which serve to amplify message content with default inferences.

In this way, Grice's ideas can be given both precision and generality. The description of many crucial aspects of meaning (for example, those associated with the natural language connectives, quantifiers, and modals) can be simplified and improved by taking into account these inferences which are closely associated with expressions by general principle, but not actually encoded in them. Such inferences have a default or presumptive character, and may be confused with the semantics and syntax of expressions.

These observations have considerable importance for linguistic theory. If they are pushed home, they imply a rather different view about the relation of language competence to nonlinguistic abilities than is usually assumed. For example, it turns out that pragmatics cannot interface with semantics in the way originally imagined: as a module or system of principles that, given the proposition coded in an utterance, would calculate additional noncoded inferences. The problem is that pragmatic inferences clearly play a central role in deriving the proposition in the first place. For example, in 'It is better to drive home and drink than to drink and drive home' the proposition expressed depends on the antecedent construal of *and* as temporarily asymmetrical (i.e., as 'and then'), which is a matter of implicature. Thus what is coded semantically and what is inferred pragmatically are often necessarily integrated in the extraction of propositional content. The implication is that there is a single level of representation to which quite different kinds of principles contribute—semantics and logical inference on the one hand, and pragmatics and presumptive reasoning on the other.

Another possible radical implication is that many correlations between form and meaning that have been attributed to syntax are more properly ascribed to pragmatics. For example, it is normally supposed that in a sentence like 'The chairman voted for him,' the chairman and the man voted for must be distinct by a rule of grammar. But in 'No-one else voted for the incumbent chairman—only the chairman voted for him' it is possible to understand the chairman as voting for himself. Arguably, this is because normally one would express such coreference by the reflexive *himself*, and the use of the nonreflexive pronoun thereby implicates a non-coreferential reading, unless there are other reasons to use it. Here there are such other reasons—'Only the chairman voted for himself' would imply that everybody else voted for people other than themselves, and that may not be what is intended. Many such reductions of alleged syntactic patterns to general patterns in pragmatics may be possible (see Huang 1994, Levinson 2000).

2.5 Conversational Structure

Conversational inferences seem to be of two kinds: first the implicature type, which relies on general background presumptions or heuristics; and second, inferences that are tied to more specific details of conversational interchange. Conversation—or the normal, informal exchange of speech, where there is rapid alternation of speakers—can be contrasted with more formal, specialized types of speech exchange (as in chaired meetings, interviews, or religious services). Needless to say, conversation is the primary form of human verbal interaction, the context in which all primary language-learning is accomplished and many details of linguistic structure are tied to it intimately. For example, a primary property of conversation is rapid turn-taking, and with this turn-taking the deictic centering also alternates: what was 'I' is now referred to as 'you,' what was 'here' becomes 'there,' and so forth, an alternation that children initially find difficult.

The characteristics of turn-taking (and indeed conversation in general) have been studied primarily by sociologists (under the rubric of conversation analysis). In the model advanced by Sacks and Schegloff, conversational turn-taking is organized as a sharing system whereby the current speaker has rights only to the end of a sentential construction or turn-taking unit, at which point transition to another speaker is governed by an ordered set of rules—if the speaker has selected a next speaker, that party should speak next, if not, other speakers may start speaking, failing which the original speaker may continue. These rules explain many details of conversational patterning, such as just where overlaps occur (e.g., as competing starts). They also explain the differential significance of silence—if no-one decides to speak after a unit which selects no following speaker, silence is likely to be interpreted simply as a lapse in conversation, but after a next speaker is selected it is heard as a significant withholding of speech (Levinson 1983, p. 300, after Atkinson and Drew):

A: *Is there something bothering you or not?*
B: (1.0 second silence)
A: *Yes or no?*
B: (1.5 seconds silence)
A: *Eh?*
B: *No.*

Differential meanings attributed to silence eloquently make the point that the semiotic significance of utterances is very much conditioned by the locus in which they occur. Consider, for example, that after a

request, an acceptance or refusal is due, but these options are not equal: an acceptance is usually delivered immediately in simple form, whereas a refusal is typically delayed, indirect, and accompanied by excuses. Recollect too, that as noted above, many speech acts are preceded by an utterance that checks out whether the conditions for that action are met—thus one can hint at an upcoming request for an appointment by asking whether the addressee would be available at a certain time. Then we can readily understand the following, where a telephone caller C already interprets a two-second pause as a negative answer (from Levinson 1983, p. 320):

C: *So I was wondering would you be in your office on Monday*

(micro-second pause)

by any chance?

(2.0 second silence)

C: *Probably not*

Here the interpretation relies (a) on the asymmetry between what the conversation analysts call *preferred* and *dispreferred* responses to specific speech acts, and (b) on the understanding that the question is not idle, but functions as a prelude to a request.

The study of conversation thus shows that the attribution of meaning to utterances depends crucially on sequential location. Arguably, this is exactly what Grice had in mind by his maxim of Relevance, where the implicatures of an utterance would also derive from the expected goals of the speaker at a particular point in discourse. Goal-driven accounts of conversational understandings have been explored most thoroughly in artificial intelligence circles, where the construction of a machine that could enter into functional dialogues has been a long-term objective.

3. Conclusions

In addition to the topics reviewed here, there are pragmatic perspectives on many other phenomena, for example register, style, jargon, and other specializations of language, the contrastive use of languages by multilingual speakers (see *Code Switching: Linguistic*), and social constraints on the use of language (see, for example, *Politeness and Language*). From research on all these topics, an overall picture is emerging of the systematic relations between semiotic systems and their contexts of use. First, any sign system or language makes systematic provision for contextual features, as in deixis or presupposition. Second, what is actually coded in signs or linguistic expressions is only a small part of the meaning attributed in context, the remainder being generated by presumptions of use. Third, the coded meaning or semantics is usually specified much less fully than conventional wisdom implies. Consider a roadsign depicting falling rocks—the depicted content is specific yet hardly conveys the full message: we must add the illocutionary force ('Beware!'), and the deictic inference ('in this vicinity' or 'the next kilometer'). Such a sign contrasts with other warning signs—for instance, a simple exclamation mark—which indicates general danger but implies that the danger is other than rocks (since otherwise the rock sign would have been used). All such signs contrast with no sign, which implies no special foreseeable danger. In similar but much more complicated ways, pragmatic principles remain the essential basis for linguistic communication. Because lexical limitations and the bounds of learning and memory constrain natural language semantics deeply. In addition, the relatively slow articulation of speech compared with understanding processes puts a clear premium on pragmatic inference. A pragmatic perspective offers fundamental insights into both the structure and use of language.

See also: Communication and Social Psychology; Deixis; Lexical Semantics; Semantic Similarity, Cogni-tive Psychology of; Semiotics; Verbal Communication, Psychology of

Bibliography

Atlas J in press *Logic, Meaning and Conversation*. Oxford University Press, New York

Brown G 1995 *Speakers, Listeners and Communication*. Cambridge University Press, Cambridge, UK

Clark H H 1996 *Using Language*. Cambridge University Press, Cambridge, UK

Cole P (ed.) 1981 *Radical Pragmatics*. Academic Press, New York

Davis S (ed.) 1991 *Pragmatics: A Reader*. Oxford University Press, Oxford, UK

Fillmore C J 1997 *Lectures on Deixis*. CSLI Publications, Stanford, CA

Frawley W 1992 *Linguistic Semantics*. Lawrence Erlbaum Associates, Hillsdale, NJ

Gazdar G 1979 *Pragmatics: Implicature, Presupposition and Logical Form*. Academic Press, New York

Green G M 1996 *Pragmatics and Natural Language Understanding*, 2nd edn. Lawrence Erlbaum Associates, Mahwah, NJ

Grice H P 1989 *Studies in the Way of Words*. Harvard University Press, Cambridge, MA

Horn L R 1989 *A Natural History of Negation*. University of Chicago Press, Chicago

Huang Y 1994 *The Syntax and Pragmatics of Anaphora: A Study with Special Reference to Chinese*. Cambridge University Press, Cambridge, UK

Hutchby I, Wooffitt R 1998 *Conversation Analysis: Principles, Practices and Applications*. Polity Press, Cambrige, UK

Jarvella R J, Klein W (eds.) 1982 *Speech, Place and Action: Studies in Dexis and Related Topics*. Wiley, New York

Kasher A (ed.) 1997 *Pragmatics: Critical Concepts*. Routledge, London

Levinson S C 1983 *Pragmatics*. Cambridge University Press, Cambridge, UK

Levinson S C 2000 *Presumptive Meanings*. MIT Press, Cambridge, MA

McNeill D (ed.) 2000 *Language and Gesture: Window into Thought and Action*. Cambridge University Press, Cambridge, UK

Mey J L (ed.) 1998 *Concise Encyclopedia of Pragmatics*. Elsevier, Oxford, UK
Mey J L 1993 *Pragmatics: An Introduction*. Blackwell, Oxford, UK
Morris C W 1938 Foundations of the theory of signs. In: Neurath O, Carnap R, Morris C (eds.) *International Encyclopedia of Unified Science*. University of Chicago Press, Chicago (reprinted in Morris C W 1971 *Writings on the Theory of Signs*. Mouton, The Hague)
Nerlich B, Clarke D D 1996 *Language Action and Context: The Early History of Pragmatics in Europe and America, 1780–1930*. Benjamins, Amsterdam
Turner K (ed.) 1999 *The Semantics/Pragmatics Interface from Different Points of View*. Elsevier, Oxford, UK
Verschueren J 1999 *Understanding Pragmatics*. Arnold, London
Yule G 1996 *Pragmatics*. Oxford University Press, Oxford, UK

S. C. Levinson

Pragmatism: Philosophical Aspects

Pragmatism is a US philosophical movement which began in the late nineteenth century and exerted significant influence until around the time of John Dewey's death in 1952. Pragmatism then fell from favor but began to enjoy a revival at the end of the twentieth century. The 'classical' pragmatists are C. S. Peirce, William James, and John Dewey. Their philosophies differed a good deal but all defended broadly empiricist views of knowledge which reject many traditional empiricist psychological views, and emphasize the role of thought in guiding action. Richard Rorty is the most influential contemporary figure who aligns himself with pragmatism.

1. *'Classical' Pragmatism*

The most important pragmatist philosophers are C. S. Peirce, William James, and John Dewey. These three are often referred to as the 'classical' pragmatists. All three were born and worked in the USA. All were primarily philosophers but had strong interests in other disciplines. The views of the 'classical pragmatists' do have important common themes, but are also dissimilar in many ways.

Pragmatism is an empiricist view of thought and knowledge which rejects much of the psychological picture usually associated with empiricism. Pragmatists think traditional empiricism has had too passive a view of the mind, and pragmatists place great emphasis on the role played by thought in guiding action and solving practical problems. Pragmatism also rejects attempts to understand human knowledge by tracing its connections to some special set of 'foundational' beliefs. For pragmatism, both specific beliefs and methods of inquiry in general should be judged primarily by their consequences, by their usefulness in achieving human goals.

The nature of truth has been a central topic for pragmatists, and a source of much trouble for them. Pragmatists in general reject 'correspondence' theories of truth. Correspondence theories claim that a true belief or statement is one which represents the world as it really is. After rejecting correspondence, pragmatists have had a difficult time devising an alternative view of truth.

Pragmatism has an often complicated relationship with science, but it is fair to say that both Peirce and Dewey regarded science as a distinctive and superior way of acquiring knowledge. For both, the scientist is in some respects the model or ideal inquirer. Although pragmatists stress the role of knowledge in solving practical problems, it is definitely an error to see pragmatism as claiming that the direction of research in science, or choices between rival theories, should be guided solely by practical or commercial demands.

Pragmatists tend to argue that there is an important place in the universe for human choice and initiative; they oppose philosophical systems which regard the world as in some sense 'finished,' 'complete,' or impervious to the effects of human choice. This is a rather vague thing to oppose, but it is common for pragmatists to be constantly on the lookout for any hint of a denial of the significance of human choice. This concern pervades much pragmatist writing. Pragmatists also tend to hold generally 'humanistic' perspectives on matters involving morals and values, opposing both nihilistic views that reject all moral thinking as illusory, and views that locate moral and other evaluative facts outside the everyday world of human well-being.

2. *Charles Sanders Peirce (1839–1914)*

Peirce's career began well but ended very badly, without his ever having held a long-term academic job. Peirce never set out his ideas in an organized and systematic way. It is difficult and probably fruitless to try to form these ideas into a single coherent system. Peirce's most famous works are about the relations between doubt, inquiry, belief, and action. He also wrote a good deal about logic, the theory of signs (which he called 'semiology'), some speculative cosmological ideas, probability, measurement, and other topics in the philosophy of science.

Peirce argued that inquiry always begins with 'real and living doubt' which he distinguished from the feigned doubt of Descartes. Doubt prompts inquiry, which aims at belief. Peirce claimed that the essence of belief is the establishment of a habit of action. Alexander Bain had made this suggestion in the mid-

nineteenth century, within the framework of a more orthodox associationist psychology. Peirce and later pragmatists made this link between belief and action central to their theories of mind and knowledge.

Peirce believed that science is the most effective method for relieving doubt and acquiring useful habits of action. His defense of science stressed the social properties of scientific inquiry, especially the tendency scientific testing has to produce a 'convergence' of belief among people who start from different positions.

In his most influential paper, 'How to make our ideas clear,' Peirce argued for a principle which later became famous as 'the pragmatic maxim.' This principle is often seen as the single idea most central to pragmatism. Peirce said that to work out what the meaning of a 'conception' is, we should work out 'what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object.' This famous principle is quite obscure, and was later paraphrased by James, Peirce, and others in a number of conflicting ways. One reasonable view is to see the famous 'maxim' as expressing a fairly familiar empiricist view, that our concept of something is just our concept of its possible effects on our experience. Then the novelty in 'How to make our ideas clear' lies in Peirce's attempt to derive this empiricist principle from an unusual starting point—a theory of the role of belief in action.

In 'How to make our ideas clear' Peirce also gave a famous definition of truth. Peirce said that what we mean by the truth is the opinion which is 'ultimately fated to be agreed to by all who investigate,' or in a more modern formulation, the opinion that will be held at the limit of scientific inquiry. This is usually taken to be a novel *definition* of truth, in conflict with more familiar views such as the correspondence theory. In fact it is difficult to work out how Peirce intended his theory of truth to be taken. In some places it looks more like an expression of faith in the power of science to *find* the truth than an unorthodox philosophical theory of what it is for an idea to be true.

3. *William James (1842–1910)*

James made pragmatism famous, especially in his 1907 book *Pragmatism* (repr. 1979), which is the most widely read of all pragmatist writings.

James had a brilliant and influential career based at Harvard, in both philosophy and psychology. He wrote many important works before he began calling himself a 'pragmatist' and aligning his views with Peirce, in 1898. In these earlier works James argued that we have the right actively to choose sides on momentous, unresolved philosophical issues that will affect how we live our lives. Choices made on these issues will inevitably be strongly affected by our individual temperaments; we should not expect all individuals to find the same philosophies appealing. Problems James applied this doctrine to include the existence of God, the freedom of the will, and the reality of moral facts. In this earlier period James also published a landmark work in psychology, his two-volume *Principles of Psychology* (1890, repr. 1950).

After 1898 James began to embed his characteristic themes within an overtly 'pragmatist' framework. Unlike Peirce, James took pragmatism to be continuous with the mainstream empiricist philosophies of English thinkers such as David Hume and J. S. Mill. James publicized pragmatism with great energy, and transformed it in ways that induced some distress in Peirce.

James presented pragmatism as a way to avoid the errors of the two key rival philosophies of his day—overly scientistic materialism and empiricism on the one hand, and sentimental, over-optimistic religious idealism on the other. He saw pragmatism as a way of distilling the genuine human significance from obscure philosophical theories and debates. James also argued that there is an inevitable and appropriate role for values in choices about theoretical issues, defended the role of human choice as a factor in determining the ultimate nature and significance of the universe, and defended the legitimacy of individual religious experience.

James made pragmatism more 'practical' than Peirce did, linking it to problems of everyday life. James was a less scientistic thinker than Peirce, and also more individualistic than Peirce or Dewey. James talked of assessing ideas in terms of their 'cash value,' an unfortunate phrase which led many to interpret pragmatism as vulgar and anti-intellectual. James also invited a stream of criticism through his tendency to discuss the nature of truth in a vague and simplistic way. He said at various times that true ideas are just those that can be verified, or those that are useful. Some of the criticisms of James rested on misinterpretations, but there is no question that he wrote in an imprecise way and was not the most consistent of philosophers. Despite this, James's writings have endured. His discussions of the role of individual temperament in philosophical choices never seem dated, for example. Another reason for James's continuing appeal is the way his attractive personality, imagination, and boundless energy seem to radiate from the pages of his work.

4. *John Dewey (1859–1952)*

Dewey's work is the high point of the pragmatist tradition to date. In the course of a remarkably long and productive career, Dewey published in virtually all areas of philosophy and in psychology, education, and politics. He wrote both popular, accessible works and dense, uncompromising ones. Dewey spent his most important years at the University of Chicago, where he worked in philosophy, education, and psychology, and after that at Columbia, where he

focused more on philosophy and was also active in politics.

Dewey's work went through several phases. Early in his career he accepted an idealist philosophy of the type influenced by Hegel, and also had strong Christian interests. Around the turn of the century he turned toward a more scientific approach to philosophy, influenced by Darwin and by James's *Principles of Psychology*. Around this time Dewey also dropped his commitment to Christianity, although his attitude to religious ideas and to the psychological and social role of religion remained complex.

The scientifically oriented philosophy Dewey developed was initially reminiscent of James's pragmatism. Later Dewey began to recast some pragmatist themes within a more 'naturalistic' framework—a framework which sets out from a biological description of living organisms and their relations to their environments. Dewey argued that intelligence is a means for humans to transform their environments, in order to deal with the problems posed by uncertainty and change in natural events.

In *Experience and Nature*, perhaps his greatest work, Dewey both defended this naturalistic view of mind and knowledge, and criticized the philosophical tradition for its postulation of false divides or 'gulfs' between mind and matter, thought and object, and so on. Dewey thought that the philosophical tradition is plagued by 'dualisms' which lead to pseudoproblems. The source of these dualisms is a 'split in being' established by the ancient Greeks, a split between the 'perfect, permanent, self-possessed' and the 'defective, changing, relational.' Dewey sought to replace these dualisms with a view based on various kinds of 'continuity,' between mind and nature, between organism and environment, and between cognition and simpler biological capacities. For Dewey, these natural continuities also provide the material needed to resolve other traditional oppositions such as that between the theoretical and practical, and between fact and value. Much of Dewey's later work is an unusual mixture of scrupulously careful system building based on assertions of 'continuity,' and sweeping historical surveys, tracing the histories of key philosophical errors back and forward through thousands of years.

Dewey also defended a social theory of mind, claiming that because thought is symbolic and symbolism is social, thought only exists within a language-using community. He argued that the primary role of science is to help human societies deal with problems and control their environments. But, like Peirce, Dewey thought that it was unscientific for scientists to direct their work according to *specific* practical problems. Rather, science is the study of a special class of properties of natural processes—relations and connections between things, which Dewey called 'instrumental' properties. Science is most successful in expanding our capacities for problem solving when it is directed on the study of the instrumental properties of things in an open-ended way unconstrained by immediate practical applications.

Dewey thought that the proper role of knowledge is to enable humans to transform their situations in beneficial ways. This led Dewey to reject what he called the 'spectator theory of knowledge,' the view that the ideal knower is someone who registers what is going on but does not intervene. This also led Dewey into some of the same arguments James had with defenders of correspondence theories of truth. Dewey thought that theories of truth as correspondence were aimed at inventing magical relations between thought and the world in order to overcome what he saw as a nonexistent problem—the problem of how the mind and the external world could have any contact with each other. He also thought correspondence theories belong with a spectator theory of the role of knowledge.

Dewey wrote often of the role played by knowledge in 'transforming' or 'reconstructing' the world. This 'transformation' should not be understood as implying an idealist metaphysical view. Dewey thought that knowledge is a factor in changing things in the world because knowledge has a role in guiding *action*, which transforms things by means of ordinary physical relationships.

Dewey was a significant figure in US political and social thought. Here he defended a version of liberalism, and he wrote extensively about the proper structure of a democratic society. At various times he was attacked by both the right and the left. He hoped for a more 'democratic' economic order, but always kept Marxism and communism at arm's length. Dewey's best known political activity was his chairing of an international inquiry into Stalin's trial of Trotsky. The inquiry entitled its report 'Not guilty.'

At Chicago Dewey worked closely with George Herbert Mead, who became one of the founders of social psychology in the USA. Mead is sometimes grouped as a major pragmatist thinker (see, for example, Scheffler 1986, Joas 1993). Mead argued that human individuality is a product of social embedding; that cognition in general and specific mental developments such as a sense of self emerge as products of various kinds of symbolic behavior within a community. Dewey also wrote about art and took a lifelong interest in education. When he was at Chicago he established an experimental school. Dewey favored problem solving as an approach to all types of learning. His ideas remained influential in education for many years while his name had almost dropped out of mainstream English-speaking philosophy.

5. *After Dewey*

Dewey's ideas, and pragmatism in general, subsided from philosophical discussion for several decades after Dewey's death in 1952. A primary reason for this was

the growing focus on formal logic and the philosophy of language in US philosophy after World War II, prompted especially by the immigration of European philosophers and logicians who defended a rigorous 'analytic' style of philosophy. After Peirce, the pragmatists had not taken an interest in formal logic. Dewey also had idiosyncratic and rather opaque ideas on the philosophy of language, which rapidly came to seem dated.

More recently, interest in pragmatism has revived. In part, this has been due to the influence of Richard Rorty, who has both revived some of Dewey's ideas and become a significant pragmatist thinker in his own right. But aside from Rorty's explicit advocacy, quite a number of the key figures in mid- to late twentieth-century English-speaking philosophy have some connection with pragmatism. W. V. O. Quine is sometimes seen as a pragmatist, because of his opposition to foundationalist projects in epistemology, and his view of rational thought as a matter of making adjustments to a 'web of belief.' For Quine, these adjustments to the web are made in whatever ways seem to work best in dealing with experience; no beliefs are in principle immune to revision. In the closing sentences of his immensely influential paper 'Two dogmas of empiricism' (1953), Quine refers to these adjustments made to beliefs in the face of evidence as 'pragmatic.' But Quine's theory focuses on the role of beliefs in prediction, and does not discuss the link between thought and action in the detail associated with classical pragmatism.

Hilary Putnam is another influential defender of some broadly pragmatist ideas. Putnam's early work defended a strongly 'realist' set of views about the relation between thought and the external world. But Putnam changed his mind, and defends a position he calls 'internal realism,' which rejects the correspondence theory of truth and which Putnam sees as continuous with the pragmatist tradition. Putnam has defended views about truth which resemble (one reading of) Peirce's view, for example. Donald Davidson, another giant figure in late twentieth-century philosophy, is occasionally classified as a pragmatist, largely because of his opposition to some standard doctrines about the representational properties of language and thought, and his opposition to some versions of the correspondence theory of truth. But Davidson's connection to classical pragmatism is tenuous.

Rorty's work has carried the influence of pragmatism outside of philosophy into neighboring humanistic disciplines. Rorty views himself as carrying on the tradition of Dewey, and to a lesser extent James. Rorty is an 'antiessentialist' about traditional philosophical concepts such as truth, knowledge and justice. Antiessentialism in this sense holds that the only theory there can be of truth (etc.) is a theory of how the term 'true' functions in ordinary discourse; it is an error to give a philosophical analysis of truth that goes beyond this. The degree of continuity between Rorty's ideas and Dewey's is a matter of controversy. As outlined above, Dewey's mature work includes a mixture of sweeping historical stories aimed at dissolving philosophical problems, and careful system building of his own. Rorty endorses the historical dissolving but not the system building. However, it can be argued that Dewey's critical points and his attempts to 'get over' standard debates tend to depend on his positive philosophical theories. And James used pragmatism to breathe new life into metaphysical problems, not to deflate them. As Rorty's views on traditional philosophical questions are so focused on dissolving and deflating, it is in some ways more appropriate to associate Rorty with the later work of Ludwig Wittgenstein rather than with pragmatism.

In recent discussion, the term 'pragmatist' has come to be used in a very loose way, to signify endorsement of any one of a wide range of ideas that are taken to have a connection with Peirce, James, or Dewey. These ideas include holism about justification, opposition to correspondence theories of truth, opposition to representationalist theories of thought and language, opposition to traditional views about logic and rationality, and almost any kind of emphasis on the practical, action-guiding side of reasoning and belief. Contemporary writers who identify themselves with 'pragmatism' often have very little in common. These loose uses of the term have lead to confusion. Perhaps any contemporary 'pragmatism' should hold, at a minimum, that the role of thought in guiding action is of central importance to the philosophy of mind and epistemology, and as a consequence of this there is a need to rework standard views about knowledge and inquiry. But any summary one might give runs into the problem that even 'classical' pragmatism was a diverse and shifting collection of ideas.

See also: Dewey, John (1859–1952); Empiricism, History of; Irrationality: Philosophical Aspects; James, William (1842–1910); Mead, George Herbert (1863–1931); Social Science, the Idea of; Wittgenstein, Ludwig (1889–1951)

Bibliography

Buchler J (ed.) 1955 *Philosophical Writings of*. Peirce. Dover, New York

Dewey J 1969–72 *The Early Works* [ed. Boyston J A]. Southern Illinois University Press, Carbondale, IL

Dewey J 1976–83 *The Middle Works* [ed. Boyston J A]. Southern Illinois University Press, Carbondale, IL, especially Vols. 13 and 14

Dewey J 1981–90 *The Later Works* [ed. Boyston J A]. Southern Illinois University Press, Carbondale, IL, especially Vols. 1, 4, and 10

James W 1950 *Principles of Psychology*. Dover, New York

James W 1956 *The Will to Believe and Other Essays in Popular Philosophy*. Dover, New York

James W 1979 *Pragmatism and the Meaning of Truth*. Harvard University Press, Cambridge, MA
James W 1983 *The Varieties of Religious Experience*. Penguin, Harmondsworth, UK
Joas H 1993 *Pragmatism and Social Theory*. University of Chicago Press, Chicago
Mead G H 1934 *Mind, Self and Society* [ed. Morris C W]. University of Chicago Press, Chicago
Peirce C S 1931–58 *Collected Papers* [Vols. 1–4 ed. Hartshorne C, Weiss P, Vols. 7–8 ed. Burks A]. Harvard University Press, Cambridge, MA, especially Vol. 5. A new edition of Peirce's works is in preparation
Putnam H 1990 *Realism with a Human Face*. Harvard University Press, Cambridge, MA
Quine W V O 1953 Two dogmas of empiricism. In: (ed.) *From a Logical Point of View*. Harvard University Press, Cambridge, MA, pp. 20–46
Rorty R 1982 *Consequences of Pragmatism*. University of Minnesota Press, Minneapolis, MN
Rorty R 1991 *Objectivity, Relativism and Truth: Philosophical Papers Volume I*. Cambridge University Press, Cambridge, UK
Scheffler I 1986 *Four Pragmatists*. Routledge, New York
Schillp P A, Hahn L E (eds.) 1989 *John Dewey*, 3rd edn. Open Court, La Salle, IL
Thayer H S 1981 *Meaning and Action: A Critical History of Pragmatism*. 2nd edn. Hackett, Indianapolis, IN
Westbrook J 1991 *John Dewey and American Democracy*. Cornell University Press, Ithaca, NY

P. Godfrey-Smith

Pragmatist Social Thought, History of

Pragmatism is the name usually given to the classical philosophical movement in the USA. It emerged in the last decades of the nineteenth century and achieved a certain hegemony in the country's intellectual life during, and immediately after, the so-called Progressive Era (1896–1914). From the 1930s, and even more after 1945, it was largely displaced by other currents of thought in philosophy, the social sciences, and public political discourse. American pragmatism also attracted considerable attention in Europe, especially around 1910, though exposition and critical discussion of pragmatism displayed widespread misunderstandings and the tendency to depreciate it by reducing it to an expression of alleged American national characteristics.

1. *Pragmatism Defined*

The everyday sense of the term 'pragmatic' certainly contributed to these misunderstandings, connoting as it does a kind of 'muddling through' which is oriented to immediate requirements, disregards theoretical or moral principles, and treats the given features of the situation simply as part of a calculation. The term pragmatism has the same Greek roots as praxis, practical, etc. The founder of pragmatism, Charles Sanders Peirce (1839–1914), coined the term as a result of his reflections on Kant's use of the terms 'pragmatic' and 'practical.' Peirce's lectures and writings around 1878 are now regarded as the original documents of pragmatism. These first became known only to narrow circles of intellectuals in Cambridge, Massachusetts, and did not achieve wider notice until some 20 years later, when William James (1842–1910) referred to them in some lectures, but above all through his lectures on 'Pragmatism' published in 1907. Peirce came to distance himself from James's pragmatism and to call his own philosophy 'pragmaticism.'

Along with Peirce and James, the inner core of pragmatism is usually taken to include John Dewey (1859–1952) and George Herbert Mead (1863–1931). The differences between all these thinkers were so evident, not least to the 'pragmatists' themselves, that their subsumption into a single school or movement has constantly been questioned. The most widely accepted view, however, has been that despite the differences there is enough of a core of shared views to justify speaking of 'pragmatism' as a distinct philosophical orientation.

What are the basic themes of pragmatism? Peirce's approach starts from his critique of Descartes' methodological principle of radical doubt and the resultant program of making the thinking ego's certainty of itself into the firm foundation of a new philosophy. The pragmatist questions the meaningfulness of Cartesian doubt, not in order to defend unquestionable authorities against the emancipatory claim of the thinking ego, but arguing for a more substantial doubt, that is, the grounding of cognition in real problem situations. In pragmatism, the guiding idea of the doubting ego is replaced by the idea of a cooperative search for truth in order to overcome real problems of action. Real doubt occurs in action, conceived as a cyclical succession of phases. Thus every perception of the world and every action in it are anchored in an unreflective belief in self-evident conditions and successful habits. But these habitual ways of acting constantly come up against the resistance of the world, which is seen to be the source of the destruction of unreflective expectations.

The resultant phase of real doubt leads to a reconstruction of the interrupted context. Perception must grasp new or different aspects of reality; action must attach itself to other elements of the world or reorganize its own structure. This reconstruction is a creative accomplishment by the actor. If he or she succeeds, via a change in perception, in acting differently and thus getting going again, then something new has come into the world: a new way of acting which can be institutionalized and itself become an unreflectively followed routine. Thus pragmatists see all human action in the opposition between un-

reflective habits of action and creative accomplishments. This also means that creativity is seen here as characterizing achievements in situations which demand a solution, rather than the unconstrained creation of something new without a constitutive background in unreflective habits.

From this basic model of pragmatism, in which action and cognition are combined in a particular way, we can derive the other central claims of pragmatism. In the metaphysics of pragmatism, reality is not deterministic: rather, it enables and demands creative action. In the epistemology of pragmatism, knowledge is not the reproduction of reality but an instrument for dealing with it successfully. The semantics of pragmatism locates the meaning of concepts in the practical consequences for action resulting from their use or their difference from other concepts. Thus in pragmatism's theory of truth, the truth of sentences can only be determined by way of a process of agreement on the success of action based on them, and not on, say, their correspondence with an uninterpreted reality. The misunderstanding of pragmatism as principally a movement aiming at the destruction of the ideal of true knowledge resulted primarily from the isolation of individual sentences, such as some of William James's statements about truth, from the whole complex of pragmatist thought.

2. *Main Variants of Historical Pragmatism*

The leading representatives of pragmatism contributed to different fields of inquiry. Peirce was primarily interested in developing a general theory of scientific knowledge and a broadly conceived theory of signs of semiotics. His multifaceted work, which does not deal with questions of social theory, includes important thoughts on the intersubjective use of signs and on the creative production of hypotheses ('abduction'). This theory of signs, particularly in its emphasis on the 'discourse' of scientists in an experimental community, has become very influential.

William James at first worked mainly in psychology, in which he saw the prospect of a way out of the dilemma between a religiously based belief in the free will of the moral agent and the scientific image of the world as a universe governed by causal processes. The solution, for James, lay in the functionality for the survival of the human organism in its environment which he saw in the human capacity to pay deliberate attention to perceptual impressions and to choose between alternative courses of action. A 'functionalist' psychology could proceed by understanding all mental achievements in terms of their function for the organism's active mastery of its environment.

James elaborated his antideterminist stance in an influential article on *The Will to Believe*, in which he defended the legitimacy of religious experience against scientistic skepticism. This essay also became the point of departure for his wide-ranging study of the varieties of religious experience, in which he founded the empirical study of religion not on religious institutions or theological doctrines, but on the phenomenology of religious experience. In his last years, he began to use the phenomenology of experience as a guideline for the elaboration of a pluralist, nondeterminist metaphysics.

John Dewey, whose thought at first developed rather independently of the first pragmatists, abandoned his earlier neo-Hegelian approach during the 1890s and linked up with the logical and psychological aspects of pragmatism as developed by Peirce and James. He constructed a philosophy which extended the core ideas of pragmatism into all the traditional domains of philosophy (metaphysics, logic, ethics, aesthetics), but also and in particular the fields of educational theory, social and political philosophy. He gave more explicit emphasis than the others to a radical version of democracy as the normative core of pragmatism. Taking part in a great number of political controversies, he became one of the most influential American intellectuals ever and *the* representative thinker of the United States.

George Herbert Mead, a friend of Dewey, did most to develop James's strategy of the translation of pragmatist themes into the program of a biologically based empirical social science. His decisive contribution to social theory consists of his theory of the specific features of human communication and hence the attempt to thematize the constitution of personality structures in the dynamics of interpersonal relationships. Mead attacked the assumption of a presocial substantial self and replaced it by a theory of the genesis of the self in which even the interaction of a person with himself or herself is conceptualized as the result of social structures. Mead also pursued this line of thought in the direction of problems of cognitive development, such as the constitution of permanent objects in experience and the constitution of structures of time. His ethical theory and antideterminist philosophy of history remain rather neglected.

Thus it can be said that pragmatism has never been a clearly delimited school of thought with a fixed research program; it shows, however, the characteristics of an intellectual movement or, at least, constellation with the common elements of antifoundationalism, the fallible nature of truth, the social nature of the self, the importance of inquiry in all fields of human life, and (metaphysical) pluralism (see Bernstein 1992). During the first decades, pragmatism exerted a considerable influence on a variety of disciplines in the United States (functionalist psychology; institutionalist economics; educational reform; law; theology; political theory; the Chicago school of sociology). After the 1930s this influence clearly waned, reaching a nadir in the 1950s and 1960s. (An indicator of this development is probably the fact that the 1968 edition of this encyclopedia did not contain an entry on pragmatism.) One of the few currents

which kept the heritage of pragmatist social thought alive was the school of symbolic interactionism in sociology, which produced a rich amount of empirical work mostly on interpersonal relationships, processes of emergence and definition of deviant and subcultural behavior, social movements, and the division of labor in professional organizations. There are other main figures of American sociology at that time in whose work pragmatist orientations can easily be detected (C. Wright Mills, Philip Selznick).

3. Contemporary Pragmatist Thinking

Since the late 1970s a renaissance of pragmatist philosophy has been underway. Apart from thinkers like Richard Bernstein and John Smith, who had continuously pursued a pragmatist agenda, two thinkers have probably contributed most to this renaissance. One is Richard Rorty who had for a long time tried to articulate pragmatist motifs within the tradition of analytical philosophy before breaking (more or less) with it and proposing a variant of postmodern thinking he calls pragmatism. Though Rorty is a perceptive social critic, he has no social theory and actively repudiates the need for it. His philosophy has sparked a vast critical literature and reappropriation of the historical pragmatists; his main rival in these debates is Hilary Putnam in whose work one can find a thorough and significant restatement of pragmatist philosophy.

The other main originator of this renaissance is Jürgen Habermas. Together with Karl-Otto Apel, these two German thinkers gave a new twist to the winded history of the European reception of pragmatism. From the beginnings of the twentieth century on, very diverse schools of European thinking had used elements from American pragmatism for their own purposes (e.g., French Catholic Modernism; Italian proto-Fascism; German Philosophical Anthropology; and others). Other European thinkers came close to pragmatism in some aspects of their work (e.g., Max Scheler, Arnold Gehlen, Martin Heidegger, Ludwig Wittgenstein) without spelling out their relationship with it. Peirce's ideas about the 'discourse' of scientists has been of decisive importance for the 'discourse ethics' developed by Apel and Habermas. Mead's understanding of symbolically mediated interaction has been crucial for the 'Theory of Communicative Action' by Habermas; he declares Mead the main inspiration for 'the paradigm shift from purposive to communicative action' which he proposes in sociological theory.

Habermas' faithfulness to the pragmatist heritage has been disputed almost as much as Rorty's. There are obviously other possibilities for pragmatist social thought than his combination of a theory of communication with systems theory or than his restriction of the tasks of moral theory to the program of discourse ethics. One can claim (Joas 1996, 2000) that the most fundamental contribution of pragmatist social thought lies not in its emphasis on communication, but in its specific understanding of the creativity of human action. This understanding distinguishes pragmatist social thought from approaches focusing on either rational action or normatively oriented action, or from a systems theory or structuralism which disconnect the dynamics of human action from the development of social wholes. On the basis of pragmatism understood as a theory of the 'situated creativity' of human action, it becomes possible to reconceptualize rationality and normativity and to develop a theory of macrosocial processes and social change prepared by the political theory and philosophy of history in the works of the classical pragmatists.

See also: Aristotelian Social Thought; Communication: Philosophical Aspects; Critical Theory: Frankfurt School; Dewey, John (1859–1952); James, William (1842–1910); Marxist Social Thought, History of; Mead, George Herbert (1863–1931); Postmodernism: Philosophical Aspects; Sociology, History of; Utilitarian Social Thought, History of

Bibliography

Bernstein R J 1971 *Praxis and Action*. Duckworth, Philadelphia, PA

Bernstein R J 1992 The resurgence of pragmatism. *Social Research* **59**: 813–40

Blumer H 1969 *Symbolic Interactionism: Perspective and Method*. Prentice-Hall, Englewood Cliffs, NJ

Dewey J 1967–87 *Early/Middle/Late Works*. Southern Illinois University Press, Carbondale, IL

Dickstein M (ed.) 1998 *The Revival of Pragmatism. New Essays on Social Thought, Law, and Culture*. Duke University Press, Durham, NC

Habermas J 1984/1987 *The Theory of Communicative Action*. Beacon Press, Boston, MA, 2 Vols. [English translation by McCarthy T]

Habermas J 1992 *Postmetaphysical Thinking*. MIT Press, Cambridge, MA

James W 1987/1992 *Writings 1878–1899/1902–1910*. Library of America, New York

Joas H 1993 *Pragmatism and Social Theory*. University of Chicago Press, Chicago, IL

Joas H 1996 *The Creativity of Action*. University of Chicago Press, Chicago, IL [English translation by Gaines J, Kecrt P]

Joas H 2000 *The Genesis of Values*. Polity Press, Cambridge, UK [English translation by Moore G]

Mead G H 1934 *Mind, Self, and Society from the Standpoint of a Social Behaviorist*. University of Chicago Press, Chicago, IL

Mead G H 1964 *Selected Writings*. Bobbs-Merrill, Indianapolis

Peirce C S 1931–60 *Collected Papers*. Harvard University Press, Cambridge, MA

Rorty R 1979 *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ

Selznick Ph 1992 *The Moral Commonwealth: Social Theory and the Promise of Community*. University of California Press, Berkeley, CA

Thayer H S 1981 *Meaning and Action: A Critical History of Pragmatism*. Hackett, Indianapolis

H. Joas

Precautionary Principle

The precautionary principle is an intuitively appealing concept that has risen to prominence over the past decade. The most well-known definition of the principle is contained in the Rio Declaration of the 1992 United Nations Conference on Environment and Development (UNCED). Principle 15 of the Declaration states that 'Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation' (Cameron 1994, p. 269). The principle has been invoked so often in international environmental resolutions that it is considered by many to be a fundamental normative principle (Cameron 1994, Costanza and Cornwall 1992). Unfortunately, there is no consensual definition of the principle and few practical guidelines for its implementation in national and international environmental policy.

The precautionary principle's rapid rise to prominence can be attributed to several factors. First, the notion of precaution is intuitively appealing, perhaps because we have been told from an early age that 'an ounce of prevention is worth a pound of cure.' Second, with the growing realization that human activities are creating global problems, such as global warming, the depletion of the ozone layer and the reduction in biological diversity, such admonishments have become all the more urgent. As the scales of the problems we face have grown, so too have the stakes. The higher the stakes, the less we can afford to take chances with the future. Third, as the size and complexity of the problems have grown, so too have the uncertainties. Thus, traditional science is unable to provide unequivocal answers. Given the stakes involved and absent clear solutions, it seems only prudent to err on the side of safety and take action now rather than later. If we wait for clear scientific proof of harm, it may be too late to do anything.

The term precautionary principle (*Vorsorgeprinzip*) was first used explicitly with regard to environmental policy in the former Federal German Republic in the 1970s (O'Riordan 1994, p. 306, O'Riordan and Jordan 1995, p. 193, Boehmer-Christiansen 1994). Subsequently, the precautionary principle was incorporated into many national and international environmental policies and protocols. The earliest use of the precautionary principle in an international agreement was in the Ministerial Declaration of the Second International Conference on the Protection of the North Sea in London in 1987. This declaration stated that 'in order to protect the North Sea from possibly damaging effects of the most dangerous substances, a precautionary approach is necessary which may require action to control inputs of some substances before a causal link has been established by absolutely clear scientific evidence' (Cameron 1994, pp. 267–8). In 1989, the Nordic Council's Conference on Pollution of the Seas called for a precautionary approach 'eliminating and preventing pollution emission when there is reason to believe that damage or harmful effects are likely to be caused, even where there is inadequate or inconclusive scientific evidence to prove a link between emissions and effects' (Cameron 1994, p. 268). In 1992, the precautionary principle was boosted to prominence by its incorporation as Principle 15 in the United Nations Conference on Environment and Development (UNCED). Similar language was used in the 1992 United Nations Framework Convention on Climate Change, and the 1992 UNCED Biodiversity Convention (Cameron 1994, p. 269). Many other similar definitions have been proposed. For example, a group of academics and environmental activists called the Wingspread Group have concluded that there is compelling evidence of harm from toxic substances, global warming, ozone depletion and nuclear materials, and that existing environmental regulations are inadequate. The group proposed that 'When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically' (Risk Policy Report 1998).

The precautionary principle comprises three core ideas. First, it is necessary to take protective actions in advance of clear scientific proof, because scientific proof is unlikely to be forthcoming in the near future, and delay in action may ultimately prove more harmful and more costly to correct after the fact. Advocates of the precautionary principle argue that the uncertainties inherent in problems such as global warming and ozone depletion necessitate taking action prior to clear scientific proof. Because the uncertainties are inherent, traditional scientific methods can never eliminate them, so waiting for clear scientific evidence is foolhardy. Wynne (1992) identifies four kinds of uncertainty—risk, uncertainty, ignorance and indeterminancy. He proposes that a precautionary approach that recognizes these uncertainties and the limitations of traditional scientific methods is necessary. Funtowicz and Ravetz (1992) argue that the nature of environmental problems changes as the decision stakes and levels of uncertainty increase. Applied scientists and professional consultants (such as engineers) are able to provide adequate solutions for environmental problems with moderate to low decision stakes and levels of uncertainty. As we move into the realm of problems such as global warming, the stakes and the uncertainties are so high that a new kind of science is necessary. Post-normal science relies on the use of 'extended facts' and an extended peer community for

review. Thus, advocates of the precautionary principle call for greater use of civic science based on negotiation, quasi-formal predictions and evaluations, and more attention to the process rather than the content of the analysis (O'Riordan and Jordan 1995, Hunt 1994).

Second, traditional methods of economic analysis, such as the use of discount rates and cost benefit analysis, are inadequate because of the nature and likelihood of the harm, the extent of uncertainties, and the delay before the harm becomes evident (Perrings 1995). Traditionally, economists assume that the growth in capital resulting from current investments will more than compensate future generations for possible harm. Unfortunately, this means that little value is placed on impacts 25 years or more in the future, and it remains unclear what the impacts will be on natural capital. Advocates of the precautionary principle argue that the likelihood, nature and magnitude of the future costs of environmental problems, such as global warming or biodiversity loss, are unknown and cannot be calculated using traditional economic methods (O'Riordan and Jordan 1995).

Third, there is a need to shift the burden of proof to the 'developer' to show that the technology or activity is 'safe.' Under many current legal and policy frameworks, the presumption is that a technology or activity is 'innocent until proven guilty.' The precautionary principle would change this to presume that a technology or activity is 'guilty until proven innocent' (O'Riordan and Jordan 1995, O'Riordan 1994).

O'Riordan and Jordan (1995) identify several other elements that characterize precautionary approaches. These include the need to: (a) maintain proportionality between the likely harm and the cost of preventive measures; (b) take preventive measures to safeguard 'ecological space' to prevent pushing natural systems beyond thresholds that result in irreversible change; (c) recognize that natural systems have intrinsic value and there is a moral imperative to protect them; (d) shift emphasis to meso-term planning (i.e., 25–100 years); and (e) pay for the ecological damage caused by those who were less than cautious in the past.

The precautionary principle has been criticized on several fronts. O'Riordan and Jordan (1995, p. 192) fear that the principle runs the risk of following the 'dangerously successful pathway pioneered by sustainability.' The principle has broad intuitive appeal but lacks rigorous consensual definitions, and while it has been adopted widely as a laudable goal for environmental policy there are no clear guidelines for its implementation. Like sustainability, the precautionary principle is neither a well-defined principle nor a stable concept. It has become the repository for a jumble of adventurous beliefs that challenge the status quo of political power, ideology and civil rights. Neither concept has much coherence other than it is captured by the spirit that is challenging the authority of science, the hegemony of cost–benefit analysis, the powerlessness of victims of environmental abuse, and the unimplemented ethics of intrinsic natural rights and intergenerational equity (O'Riordan and Jordan 1995, p. 191).

Bodansky (1994) argues that the precautionary principle underlies many of the environmental laws enacted in the 1970s in the US, but that its impact has often been contrary to the intent of the framers of these laws. He argues that the precautionary approach of the original Clean Air and Clean Water Acts led to delays in the promulgation of regulations. He also suggests that shifting the burden of proof to the proponent of a technology or activity may do little to enhance safety, since harmful impacts are often discovered only belatedly, as in the case of DDT and CFCs. Furthermore, shifting the burden of proof may actually exacerbate harm by discouraging the introduction of beneficial technologies. Bodansky questions whether it is feasible to establish guidelines that determine what levels of harm and scientific certainty are necessary to trigger the precautionary principle, and whether we should always err on the side of safety even if the social costs are very high. Bodansky (1994, p. 225) concludes 'Simple rules promise simple answers. But, in the end, there may be no practical alternative to the imperfect, difficult, often dissatisfying task of muddling along.'

Perrings (1995) and Costanza and Cornwall (1992) are more optimistic about the practical application of the principle. Perrings (1995, p. 159) believes that in the absence of better (i.e., more certain) information the precautionary principle 'leads to a sequential approach in which the decision maker reserves judgment about the uncertain outcomes of each activity, assuming the worst-case outcome until evidence is provided to the contrary'. Thus, the decision-maker assigns worst-case values to the uncertain outcomes of current activities and attempts to minimize the maximum environmental costs. The pragmatic problem becomes choosing a plausible worst-case for decision-makers.

Costanza and Cornwall (1992) propose combining the precautionary principle with the polluter pays principle to develop an environmental deposit-refund or assurance bonding system. Under this system, an 'assurance bond,' equal to the best estimate of the largest potential future damages, would be levied on the proponent of the technology or activity. Portions of the bond would be returned as and when the proponent could show that the suspected damages had not occurred or they would be used to repair or compensate for any damages resulting. Costanza and Cornwall believe this system would shift the burden of proof and the cost of uncertainty from the public to the proponent or developer. The bond would also serve as an incentive for businesses to conduct thorough, publicly accessible research on the environmental impacts of their activities and to look for ways to minimize pollution. Unfortunately, many questions

about the practical value of such a system remain unanswered: who determines the worst-case scenario and the size of the bond, especially since there is so much uncertainty about the future environmental impacts of so many activities? Who decides if and when the bond should be returned, and how? Is it practicable to think of dealing with myriad environmental problems in this fashion? How would this be implemented at an international level?

In conclusion, the precautionary principle has enormous intuitive appeal and has become an integral part of many international environmental agreements. The future success of the principle, however, rests on the development of clear practical guidelines for its implementation at both the national and international levels.

See also: Environment and Development; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Economics; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; International and Transboundary Accords, Environmental; Planning Issues and Sustainable Development; Sustainable Development

Bibliography

Bodansky D 1994 The precautionary principle in US environmental law. In: O'Riordan T, Cameron J (eds.) *Interpreting the Precautionary Principle*. Earthscan, London, pp. 203–28

Boehmer-Christiansen S 1994 The precautionary principle in Germany—enabling government. In: O'Riordan T, Cameron J (eds.) *Interpreting the Precautionary Principle*. Earthscan, London, pp. 31–60

Cameron J 1994 The status of the precautionary principle in international law. In: O'Riordan T, Cameron J (eds.) *Interpreting the Precautionary Principle*. Earthscan, London, pp. 262–89

Costanza R, Cornwall L 1992 The 4P approach to dealing with scientific uncertainty. *Environment* **34**: 12–42

Funtowicz S, Ravetz J 1992 Three types of risk assessment and the emergence of post-normal science. In: Sheldon K, Golding D (eds.) *Social Theories of Risk*. Praeger, Westport, CT, pp. 251–74

Hunt J 1994 Social construction of precaution. In: O'Riordan T, Cameron J (eds.) *Interpreting the Precautionary Principle*. Earthscan, London, pp. 117–25

O'Riordan T 1994 The precautionary principle in environmental management. In: Ayres R U, Simonis U E (eds.) *Industrial Metabolism: Restructuring for Sustainable Development*. United Nations University Press, Tokyo, pp. 299–318

O'Riordan T, Jordan A 1995 The precautionary principle in contemporary environmental politics. *Environmental Values* **14**: 191–212

Perrings C 1995 Reserved rationality and the precautionary principle: technological change, time, and uncertainty in environmental decision making. In: Krishnan R, Harris J M, Goodwin N R (eds.) *A Survey of Ecological Economics*. Island Press, Washington, DC, pp. 158–62

Risk Policy Report 1998 Wingspread group offers new precautionary principle statement. *Risk Policy Report* **5**: 22–3

Wynne B 1992 Uncertainty and environmental learning: reconceiving science and policy in the prevention paradigm. *Global Environmental Change* **2**: 111–27

D. Golding

Prefectorial Government

Governments must govern their territory and ensure that decisions reached in the national capital are implemented throughout the country. Different political traditions emphasize policy coherence and uniformity across their territory to differing degrees, and therefore invest varying levels of political and administrative resources into controlling subnational governments. For countries more concerned with uniformity (such as France and countries that have adopted the French administrative tradition) prefects are crucial instruments for central governments. France and several other counties utilize these administrative officials as one component of their attempts to control policy implementation through subnational units of government, and also exercise control over the field staffs of other ministries in central government.

Prefectoral government refers to the utilization of central government officials—the prefect or *prefet*—to monitor and control policy within the area of the country for which each is responsible. For example, each department in France has a prefect, along with several *sous prefets* and other administrators responsible for controlling policy within the department. These officials are employees of the Ministry of the Interior and must review actions taken by both field services of most other ministries, as well as many decisions of department and commune officials. The relationships between the prefect and field staffs may be rather complex. For example, the prefect has little control over the pedagogical activities of the Ministry of Education, but does control construction of new schools.

The office of prefect has a long history, in France being derived from royal intendants in the *ancien* regime, and having persisted in one form or another since the Napoleonic era. As the office was developed during the Napoleonic era, it was seen as the means of creating a common nationhood in the face of substantial internal diversity within the country. The office briefly was called Commissioner of the Republic during the early stages of decentralizing reforms during the 1980s but the traditional title soon was restored to reflect the power of the State. Also, the creation of the regional planning system and regional governments with limited powers during the 1960s produced another level of government at which there is a need for prefectoral control in France, with some analogous development in other systems. Also, the

continued decentralization and deconcentration of policymaking in France (and other countries) means there is an additional necessity of integrating and coordinating policies at the implementation stage (Dupuy and Thoenig 1985, Thorval 2000). The system has been copied, with predictable variations, at various times in numerous other countries such as Italy, Greece, Spain, and Japan, as well as continuing to be utilized in several former French colonies. Few of the systems copied from the original have provided these officials with the potent regulatory functions characteristic of French administration.

Formally prefects exercise substantial powers within their department. The can reverse decisions made by other officials within their department and impose their own decisions in place of those decisions they consider inappropriate, or at least refer the issues involved back to Paris for final determination. Given the strong tutelary and control functions exercised by prefects, they rarely if ever are natives of the territories within which they work. In addition, the prefect has tended to have a virtual monopoly over communication between the department and the central government in Paris, so that the field staffs of most ministries have been channeled through the prefecturate. This monopoly of communication constitutes a major source of power for prefects, given that the field staffs responsible to other ministries might have few independent avenues for complaining about the decisions of the prefect. In addition to the powers over national ministries, prefects frequently are directly involved with local government and have been in essence the chief executive for local governments. With reforms in the administrative system these local governments have been given increased powers but the prefect still has numerous tutelary powers.

The French prefect, much as the President of the United States, can be conceptualized as fulfilling a number of functions (Bernard 1992):

(a) The guarantor of the heritage of the nation and of the continuation society;

(b) The guardian of public order, with ultimate responsibility to preserve security and safety;

(c) The representative of the State within the geographical area, and the guarantor of equality of citizens;

(d) The promoter of regional economic development and (increasingly) the link of the region with the European Union; and

(e) The promoter of social consensus and cooperation within the region.

The political climate favoring decentralization and enhanced local autonomy has minimized the central position of prefects in most political systems, but the office remains an active component of administration in political systems that remain concerned about creating uniformity, including France. Indeed, the role of the prefect as a coordinator of local services may be even more crucial with the substantial deconcentration of policymaking. The range of powers available to prefects is becoming more constrained, and local communities are more likely to assert their own policy and political concerns. Even with the increase in local democracy, prefects remain powerful actors in local government, preserving their political power in part through alliances with local political leaders. With the creation of regional governments concerned primarily with economic development there is now a new level of prefectoral government in France, although the degree of control available to these officials is substantially more limited. Further, the role that the prefect plays in the distribution of funds coming from the European Union has also enhanced the central role of the prefect in subnational policy-making.

All administrative systems that utilize this variety of administrative mechanism for control over local areas are not called prefectoral. For example, the governors appointed by central government in Scandinavian countries have lost much of their authority over local authorities, although there are some vestigial powers of that sort. Even local mayors in some administrative systems (Finland and the Netherlands) may be in effect central government officials responsible for ensuring policy delivery in the face of potentially recalcitrant local political forces.

Authoritarian governments also have employed a variety of administrative mechanisms to control local governments and policy implementation, often using party officials rather than administrators to exercise the same control functions exercised by prefects.

Likewise, all administrative systems that are called prefectoral do not function in the same ways. Fried (1963), for example, distinguished between integrated and unintegrated prefectoral systems, with the former being exemplified by the French model and the latter by the Italian. The Italian prefectoral system is substantially less restrictive than the French version of exercizing administrative control through prefects. For example, Italian field staffs from line ministries have had their own lines of contact with the government in Rome without being dependent upon the prefect. Also, Italian prefects have not exerted direct control over local governments in as overt a manner as they might in France, nor have they supplied accounting and other managerial services to the local governments. The prefect in Italian (and most other) administrative systems is more of a monitor than an effective regulator of subnational governments.

There are in many ways two prefects existing simultaneously in the office, and there is a great deal of ambiguity in the role as it has evolved (de Montricher 2000). On the one hand, much of what has already been described reflects the legal version of the prefectoral role. This official is the formally powerful representative of the authority of the State, resplendent in their gold-braided uniform, playing a central part in local ceremonies to demonstrate the involvement of central government in those activities. On the

other hand, the prefect is a bureaucratic politician who must build coalitions with local officials in order to be successful in governing. Although prefectoral systems originally were designed to exercise central control over subnational policies and administration, in practice local governments also have a great deal of political influence over the prefect, and may convince that official to promote their own local interests, rather than simply being forced to accept what they are being instructed from the center.

The prefect generally does not have to be persuaded very much to help the local community in pressing its demands on central government. The prefect is a career administrative official, generally coming from the very summit of the career structure. These officials naturally want to build a successful career, and prefects are also subject to removal from office more readily than are other civil servants. One measure of success for the prefect would be the ability to administer central policies without complaints from the localities, complaints that might be expressed through local political notables. Further, if prefects can construct a workable coalition with political elites in the localities over which they have responsibility they may be able to build an even more effective administration and thereby improve their career chances even more. Various examinations of the politics of these administrative relationships undertaken in France demonstrate that influence and control circulate in both directions and that most final policy decisions are the products of negotiation and bargaining rather than imposition of authority through the prefect (Crozier and Thoenig 1976, Worms 1966). Further, prefects engage in a number of consultation and consensus-building operations within their departments to both enhance the quality of the decisions made and improve the involvement of local citizens in their own government. In France the power of local politics is enhanced by the *cumul des mandats* through which the same politician will hold office at both the central and local levels.

Prefectoral control may appear almost quaint in the contemporary era that stresses local democracy and accepts greater diversity among policy outcomes. Further, prefectoral administration may appear anti-democratic, given that this instrument of governing substitutes administrative controls for the judgment of locally elected public officials (Behar and Estebe 1996). The overwhelming concern with uniformity and control also may limit the capacity of local and regional governments to innovate, while there is growing evidence of the efficacy of policy innovation at lower levels of government. In addition, uniformity in policy and administration may not conform well to managerialist ideas devolving authority to lower echelons of organizations.

Prefectoral policymaking systems are concerned strongly with centralized control and can be contrasted with federal systems that tend to involve states or provinces for implementation without the direct monitoring and control, and that are willing to accept wider disparities in policy. The prefectoral systems may also be contrasted with other types of unitary government that rely on the functional field staffs of central ministries to provide sufficient uniformity throughout their territory. Those functional administrative systems do not consider it necessary to impose an extra level of oversight, at least not to impose *ex ante* controls. Thus, the choice of this form of governing reflects some fundamental choices about the way in which laws are to be implemented and the extent to which uniformity is a requirement for justice.

See also: Administration in Organizations; Financial Accounting and Auditing; Governments; Public Administration, Politics of; Public Bureaucracies

Bibliography

Behar D, Estebe P 1996 *Villes, sciences sociales, professions*. L'Harmattan, Paris

Bernard P 1992 *Le prefet de la Republique: Le chene et l'olivier*. Economica, Paris

Crozier M, Thoenig J-C 1976 La regulation des systems organises complexes: Le case du systeme de decision politico-administratif en France. *Revue française de sociologie* **16**

Dupuy F, Thoenig J-C 1985 *L'administration en miettes*. Fayard, Paris

Fried R 1963 *The Italian Prefects*. Yale University Press, New Haven, CT

de Montricher N 2000 The prefect and the reform of the state. *Public Administration*

Thorval J 2000 L'avenir de la representation de l'Etat. *Pouvoirs Locaux* **44**(1): 84–9

Worms J-P 1966 Le prefet et ses notables. *Sociologie du travail* **3**

B. G. Peters

Preference Models with Latent Variables

1. Introduction

Much of human behavior may be viewed as successive choices among more or less well-defined options or alternatives. In some situations, choices are made over a continuum of possibilities; for example how much of a particular resource to allocate for a task or how much to work. In other situations, choices are made among a finite (discrete) number of options. Familiar instances are a citizen deciding whether to vote and if so for whom, a shopper contemplating various brands of a product category, or a couple weighing marriage. Frequently, these decisions are made in a sequential manner. Examples include choice of employment, domicile, transportation mode and investment. With an emphasis on discrete types of choice problems,

research on preference and choice has revolved about the development of mathematical models that predict how both observed and unobserved attributes of decision makers and of choice options determine decisions. By treating the unobserved attributes as latent (random) variables, these preference models focus mainly on choice outcomes and to a lesser extent on the underlying decision processes. As a result, the main purpose of preference models with latent variables is to summarize the data at hand and to facilitate the forecasting of choices made by decision makers facing possibly new or different variants of the choice options.

The prevalence of behavioral models with their objective of choice prediction originated with the work of Thurstone (1927). By invoking the psychophysical concept of a sensory continuum, Thurstone explained the observation that choices by the same person may not be deterministic under seemingly identical conditions. He argued that options should be represented along this sensory continuum by random variables that describe the options' effects on a person's sensory apparatus. The choice process is then reduced to selecting the option with the greatest realization. One may question the use of randomness as a device to represent factors that determine the formation of preferences but are unknown to the observer. However, because, in general, it is not feasible to identify or measure all relevant choice determinants (such as all attributes of the choice options of the person choosing, or of environmental factors) the question of whether the choice process is inherently random or is determined by a multitude of different factors may be of little relevance in at least one aspect: either position arrives at the same conclusion that choices are best described in terms of their probabilities of occurrence.

Thurstone's early conceptualizations gave rise to an extensive body of methodological and empirical research, mainly in economics and psychology. Surveys about developments in choice theories (e.g. see *Decision and Choice: Random Utility Models of Choice and Response Time*; *Luce's Choice Axiom*; *Ranking Models, Mathematics of*, Marley 1992, Suppes et al. 1992) and choice modeling (Amemiya 1985, Fligner and Verducci 1993, Manski and McFadden 1981, McFadden 1984) indicate that the merging of the literatures in these two disciplines produced a flexible and powerful framework for the structural investigation of choice data. In particular, this framework has proved to be useful for analyses of individual preference differences at a particular point in time and over time, multidimensional representations of choice alternatives, and the incorporation of variables that describe choice options and decision makers. Each of these areas of research is outlined in the following sections. Because emphasis is on conceptual issues, more technical topics such as sampling, estimation and inference will not be discussed (Albert and Chib 1993).

2. Theoretical Framework

We define a population $\mathscr{P}$ of decision makers ($n \in \mathscr{P}$) and a finite set $\mathscr{C}$ of choice alternatives ($j \in \mathscr{C}$). For each decision maker n and choice alternative j, a vector $\boldsymbol{x}_{nj}$ of observed variables is available that partially describe the pair (n, j). The underlying momentary utility of option j to person n is denoted by u_{nj} and a function of the observed and unobserved attribute values. Distributional assumptions are made to capture the variation of the unobserved attribute values and taste differences among the decision makers. In combination with a decision rule for selecting an option, the utilities determine the options' choice probabilities. As a result, the choice probabilities can be represented by an a priori specified member of a family of functions which is partially determined by the distributional assumptions about the unobserved attribute values and utility variations and indexed by a parameter vector $\boldsymbol{\theta}$.

Choice data may be collected in a variety of formats. For example, the *paired comparison* technique is the method of choice in laboratory settings. In large scale surveys or panel interviewer studies, incomplete and/or partial *rankings* or *preference ratings* are frequently collected because of their time-efficiency. In conjoint measurement studies (e.g., see *Measurement Theory: Conjoint*), the constant-sum technique is often used. In recurrent choice studies, we observe how often each choice option is selected within a certain time period. Each data type requires a different treatment to both investigate how decision makers differ in their preferences for choice options and to accommodate the distinct psychological processes that are specific to each choice task.

Consider, for example, a ranking experiment with the choice set $\mathscr{C} = \{i, j, k\}$ but without any attribute information. The first three columns of Table 1 represent the six possible ranking outcomes. Each number corresponds to one of the ranking positions of an option. For instance, the ranking outcome {2, 1, 3} indicates that options i, j and k are ranked second, first and third, respectively. Here the rankings are obtained based on the ordering (from largest to smallest) of the underlying options' utilities. The remaining columns of Table 1 depict these orderings of the utility values

Table 1
Outcomes of a ranking experiment

Ranks			
i	j	k	Ordering of utilities
1	2	3	$u_i \geqslant u_j \geqslant u_k$
1	3	2	$u_i \geqslant u_k \geqslant u_j$
2	1	3	$u_j \geqslant u_i \geqslant u_k$
3	1	2	$u_j \geqslant u_k \geqslant u_i$
2	3	1	$u_k \geqslant u_i \geqslant u_j$
3	2	1	$u_k \geqslant u_j \geqslant u_i$

(without the subscript n) for each of the six ranking outcomes. For instance, $u_j \geqslant u_i \geqslant u_k$ indicates that the utility of j is larger than the utility of i which, in turn, is larger than the utility of k. The probability of observing this ranking outcome can be written as $\Pr(j \succcurlyeq i \succcurlyeq k|\boldsymbol{\theta}) = \Pr[(u_j - u_i) \geqslant 0 \bigcap (u_i - u_k) \geqslant 0]$. If the individual utilities follow a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_i, \mu_j, \mu_k)'$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_i^2 & \sigma_{ij} & \sigma_{ik} \\ \sigma_{ij} & \sigma_j^2 & \sigma_{jk} \\ \sigma_{ik} & \sigma_{jk} & \sigma_k^2 \end{bmatrix}, \tag{1}$$

then probability of observing this ranking can be computed by evaluating the cumulative distribution function of the bivariate normal density with upper integration limits $(\mu_j - \mu_i)/(\sigma_j^2 - 2\sigma_{ij} + \sigma_i^2)$ and $(\mu_i - \mu_k)/(\sigma_i^2 - 2\sigma_{ik} + \sigma_k^2)$ and correlation coefficient $(\sigma_{ij} - \sigma_{jk} - \sigma_i^2 + \sigma_{ik})[(\sigma_j^2 - 2\sigma_{ij} + \sigma_i^2)^{\frac{1}{2}}(\sigma_i^2 - 2\sigma_{ik} + \sigma_k^2)^{\frac{1}{2}}]^{-1}$. It is important to note that because the ranking probabilities are based on the differences between the options' utilities, it is not possible to uniquely identify the underlying mean and covariance structure.

When relating attribute information to individual utilities, it is useful to distinguish between systematic and random sources of influence. Random sources of influence include (a) attributes of the choice options that are unobserved but influence preferences, (b) measurement errors, and (c) unobserved variables that account for taste variations among the decision makers (Manski & McFadden 1981). Systematic components of influence include the effects of the measured person and option characteristics $\boldsymbol{x}_{nj}$. The implication of this view is that u_{nj} can be written as $u_{nj} = \mu_{nj} + \varepsilon_{nj}$, where μ_{nj} and ε_{nj} represent the effects of observed and unobserved determinants, respectively, of an option's utility. Without prior information, we can only determine the distribution of the unobserved values across the population of decision makers.

2.1 *Modeling* μ

The specific functional relationship between μ_{nj} and known option and person characteristics may depend on the particular application. For example, in situations where the measured attributes of the choice options are of a 'more-is-better' type, a linear function may prove sufficient to describe the attributes' effects. This is the case for the *vector model* (Carroll and DeSoete 1991) which assumes that a decision maker combines the attribute values in a linear fashion. In contrast, when the systematic utility component is a *single-peaked* function of the measured variables, an *ideal-point* representation is more appropriate (Coombs 1964).

The attribute effects may vary from person to person in a systematic fashion, for example, by being related to such person-specific variables as age, gender, and educational achievement. However, in many cases reference to demographic variables is not sufficient to explain heterogeneity in preferences among individuals. Because psychological or economic theories provide little guidance about the role person-specific variables play in the formation of preferences, a random regression approach is required. Thus, it may be postulated that the attribute effects follow some a priori specified distribution. In particular, latent structure models are useful in capturing individual differences in the attribute effects by introducing the fairly unrestrictive assumption that individual variations can be described by a small number of 'prototypical' effect vectors (Böckenholt and Böckenholt 1991). Individual differences are accounted for by assigning each person to a preference class such that members of a preference class exhibit the same or similar attribute effects. For example, when preference data are collected by the method of first choices, a latent-class choice model is given by the following equation

$$\Pr_n(i|\boldsymbol{\theta}) = \sum_{s=1}^{S} \pi_n(s) \Pr_n(i|\boldsymbol{\theta}_s) \tag{2}$$

where

- $\Pr_n(i|\boldsymbol{\theta})$ = probability of person n choosing alternative i, and $\boldsymbol{\theta}$ is a parameter vector;
- $\pi_n(s)$ = probability of person n belonging to class s;
- $\Pr_n(i|\boldsymbol{\theta}_s)$ = probability of person n choosing alternative i given n belonging to class s.

The latent classes may represent (a) different *decision rules* adopted by the individuals; (b) variable choice sets considered by the individuals; or (c) subgroups of the population with varying tastes or, sensitivity to the options and their attributes.

2.2 *Modeling* ε

Attributes that influence the utility of an alternative but are not measured are represented by ε_j. The distribution of ε_j is frequently specified to be multivariate normal with arbitrary covariance matrix. This specification is useful in analyzing *similarity* relationships between choice options that are introduced by the unmeasured attributes. Because the estimation of the elements of the covariance matrix is computationally complex, more restrictive representations are also of interest. For example, Stern (1990) proposed a family of choice models based on independent gamma random variables which includes Luce's (1959) choice model as a special case.

Careful checks are necessary to determine whether the independence assumption is valid or whether unobserved attributes may contain information useful in identifying systematic sources of individual differences and option characteristics. If respondents per-

ceive a common set of latent attributes but differ in the weights they assign to the attribute values, the covariances of ε_j can be decomposed according to a linear *factor-analytic model* (Takane 1987). Other possibilities include the application of ideal-point models which are based on the assumption that individuals choose the option that is closest to their 'ideal option' where closeness is a function of the distance between the choice option and the person-specific ideal option (Coombs 1964). Distance may be defined in various ways, for example, by an Euclidean distance in a continuous latent attribute space or by a latent tree structure in which both choice and 'ideal' options are represented by nodes. However, in all of these cases, the interpretation of taste differences is much simplified provided the assumption is correct that decision makers use the same set of attributes in assessing the choice options.

2.3 Modeling Preferences Over Time

When collecting time-dependent choice data, it is important to take into account that individuals not only differ in their preferences at a particular point in time but also in the way they change their preferences over time. For instance, current preferences may depend on such diverse factors as previous choices, *interchoice times*, or the length of time a particular preference state has been occupied. Decision makers may also differ in their familiarity with features of the choice options with the result that some decision makers are stable in their preferences over time while others may change their preferences according to some learning process. A complicating factor in this context is that choice environments outside a laboratory setting are rarely stable. Typically, attributes of choice options (e.g., price) and decision environments (e.g., changing tax code) vary over time and influence choice behavior.

The development of time-dependent choice models requires significant modifications of the statistical framework for cross-sectional choice data. However, because choice processes are inherently dynamic and are studied best by taking their temporal nature into account, approaches that focus on these aspects are more likely to lead to important insights about decision making. Conceptually, it is useful to distinguish between *heterogeneity* and *state dependence* in the analysis of recurrent choice data. According to the 'heterogeneity approach,' individual choices are assumed to be conditionally independent given the person-specific parameters. For example, *latent structure choice models* represent time-dependent individual differences by switches among latent preference states. If preferences are stable, membership in a latent preference state is time-homogeneous. In contrast, systematic choice variability over time may indicate changes in the perception of (some of) the choice options. Stability and shifts in preference states (for, say, three equidistant time points) can be investigated by extending Eq. (2) to

$$\Pr_n(i_{t-1}, j_t, k_{t+1}|\boldsymbol{\theta}) = \sum_{a=1}^{S}\sum_{b=1}^{S}\sum_{c=1}^{S}\pi_n(abc)\,\Pr_n(i_{t-1}|\boldsymbol{\theta}_a)\,\Pr_n(j_t|\boldsymbol{\theta}_b)\,\Pr_n(k_{t+1}|\boldsymbol{\theta}_c), \qquad (3)$$

where

- $\Pr_n(i_{t-1}, j_t, k_{t+1}|\boldsymbol{\theta})$ = probability of person n choosing alternative i at time point $(t-1)$, j at time point t, and k at time point $(t+1)$; and
- $\pi_n(abc)$ = probability of person n belonging to classes a, b, and c at time points $(t-1)$, t, and $t+1$, respectively.

Special cases of (3) are obtained by imposing constraints on $\pi_n(abc)$ to describe the switching process among the latent preference states (Böckenholt and Langeheine 1996).

Although time-varying taste heterogeneity is perhaps the single most important source of dependency in recurrent choice data, the assumption of conditionally independent decisions may require the estimation of a large number of preference states in order to be satisfied. A more parsimonious approach for modeling state dependency is obtained by considering the outcomes of previous choices as possible explanatory variables for current choices. For example, if only the choice outcome at $(t-1)$ influences the current choice, the state-specific choice probabilities are specified as $\Pr_n(j_t|i_{t-1}, \boldsymbol{\theta}_s)$. This approach has been shown to be fruitful in capturing such psychological phenomena as habit persistence, learning, or variety seeking (Manski and McFadden 1981). Further generalizations are also possible by allowing for several sets of preference states, where each set follows a different switching process over time.

3. Conclusion

The formulation of choice models in terms of latent utilities has a long tradition in biometric, econometric, psychometric and sociometric disciplines. The numerous applications and further developments in each of these areas have led to many improvements in the statistical apparatus necessary for estimating and validating choice models. However, the unifying principle remains valid that observed individual choice behavior is an imperfect reflection of the underlying preferences of a person. This viewpoint leads to a probabilistic representation of the relation between the manifest and latent levels of choice analysis taking into account both the process by which an individual arrives at a choice and the factors that influence this process. Some of these factors are unobservable (such as the processed information about the choice alternatives), while other factors are known and measurable. By separating these two sources of influence due to measured and unobserved attributes at the individual level, a powerful modeling approach is obtained that

facilitates the identification and isolation of the net effects of choice option characteristics, individual characteristics, and interactions of option and individual characteristics as separate choice determinants.

Modifications of this choice modeling approach are necessary when the choice set is large or when a simultaneous presentation of the choice options is not feasible. In these cases, one may argue that decision makers tend to rely on *satisfying* rules and the idea of an *aspiration level* may be more appropriate than the principle of utility maximization (Simon 1957). In contrast to the large number of possible satisfying rules, however, only a few mathematical representations have been proposed. Perhaps the most popular model is due to Coombs' (1964) analysis of the 'pick any/m' task and Thurstone's (1959) work on the method of successive categories. In the 'pick any/m' task it is assumed that choice options are selected when their utilities exceed some specified minimum value. Thus, the underlying choice process is scalar valued and based on a binary comparison between the option's utility and a threshold value. More specifically, according to Coombs' (1964) unidimensional unfolding approach, the utility of item j for person n is given by the distance between two parameters representing j and n on a unidimensional continuum. This idea has been incorporated in several psychometric models by making the simplifying assumption that the perception of the item-specific parameters does not vary from person to person (Andrich 1997). In general, results of the *maximization* and 'aspiration level' hypotheses are not directly comparable because in the latter case we model the probability of belonging to some category (i.e., 'above the threshold') and in the former case we model order relations among the choice options.

In recent years, stochastic choice models have come under criticism because, frequently, they cannot capture context-sensitive behavior (Tversky and Kahneman 1991). For example, the models' underlying assumption that the assessment of an option's utility does not depend on comparisons drawn between it and other available alternatives has received little support in laboratory studies. Moreover, the deliberation process from the onset of the choice task to the final selection has been shown to play an important role that should be taken into account by the next generation of choice models (Busemeyer and Townsend 1993). Clearly, the latent utility approach has proved to be most useful in providing parsimonious descriptions of how judges differ in their preferences at a particular point in time or over time. However, by emphasizing interindividual differences, this approach can render only limited insights about cognitive processes influencing a choice.

See also: Bounded and Costly Rationality; Decision Research: Behavioral; Factor Analysis and Latent Structure: Overview; Unfolding and Vector Models

Bibliography

Albert J, Chib S 1993 Bayesian analysis of binary and polychotomous data. *Journal of the American Statistical Association* **88**: 669–79

Amemiya T 1985 *Advanced Econometrics*. Harvard University Press, Cambridge, MA

Andrich D 1997 A hyperbolic cosine IRT model for unfolding direct responses of persons to items. In: van der Linden W J, Hambleton R K (eds.) *Handbook of Modern Item Response Theory*. Springer, New York, pp. 319–414

Böckenholt U, Böckenholt I 1991 Constrained latent class analysis: Simultaneous classification and scaling of discrete choice data. *Psychometrika* **56**: 699–716

Böckenholt U, Langeheine R 1996 Latent change in recurrent choice data. *Psychometrika* **61**: 285–302

Busemeyer J R, Townsend J T 1992 Decision field theory: A dynamic cognition approach to decision making. *Psychological Review* **100**: 432–59

Carroll J D, DeSoete G 1991 Toward a new paradigm for the study of multiattribute choice behavior. *American Psychologist* **46**: 342–52

Coombs C 1964 *A Theory of Data*. Wiley, New York

Fligner M A, Verducci J S (eds.) 1993 *Probability Models and Statistical Analyses for Ranking Data*. Springer, New York

Luce R D 1959 *Individual Choice Behavior*. Wiley, New York

Manski C F, McFadden D (eds.) 1981 *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press, Cambridge, MA

Marley A A J 1992 A selective review of recent characterizations of stochastic choice models using distribution and functional equation techniques. *Mathematical Social Sciences* **23**: 5–30

McFadden D 1984 Qualitative choice models. In: Griliches Z, Intriligator M D (eds.) *Handbook of Econometrics*. MIT Press: Cambridge, Vol. II, pp. 1395–457

Simon H A 1957 *Models of Man: Social and Rational, Mathematical essays on rational human behavior in a social setting*. Wiley, New York

Stern H 1990 A continuum of paired comparison models. *Biometrika* **77**: 265–73

Suppes P, Krantz D, Luce D, Tversky A 1992 *Foundation of Measurement. Geometric, Threshold, and Probabilistic Representations*. Academic Press, San Diego, Vol. II

Takane Y 1987 Analysis of covariance structures and probabilistic binary choice data. *Communication and Cognition* **20**: 45–62

Thurstone L L 1927 A law of comparative judgment. *Psychological Review* **34**: 273–86

Thurstone L L 1959 *The Measurement of Values*. University of Chicago Press, Chicago

Tversky A, Kahneman D 1991 Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics* **107**: 1039–61

U. Böckenholt

Prefrontal Cortex

The prefrontal cortex is the cortex of association of the frontal lobes. It plays a major role in the temporal organization of behavior, reasoning and language.

The underlying mechanisms are not fully understood, but are known to support at least two cognitive functions of temporal integration that are essential to that role: active short-term memory and preparatory set.

1. Anatomy

In evolution and in individual growth, the prefrontal cortex develops more slowly, yet further, than most other brain structures, in both size and weight (Fig. 1). Phylogenetically and ontogenetically, it is one of the last regions of the cortex (neocortex) to develop (Brodmann 1909). The human prefrontal cortex does not reach full structural maturation of cells and connections until young adulthood. In the adult human, the prefrontal cortex constitutes almost one-third of the totality of the neocortex. For reasons that are not well understood, this region of the cortex, which is one of the latest to develop, is also one of the most vulnerable to involution and pathological aging (Raz 2000).

The primate prefrontal cortex (Petrides and Pandya 1994) is composed of three major parts or regions (Fig. 2): the dorsolateral prefrontal cortex, that is, the cortex of the dorsal and lateral convexity of the anterior portion of the frontal lobe (Brodmann cyto-architectonic area 46, and lateral parts of areas 8, 9,

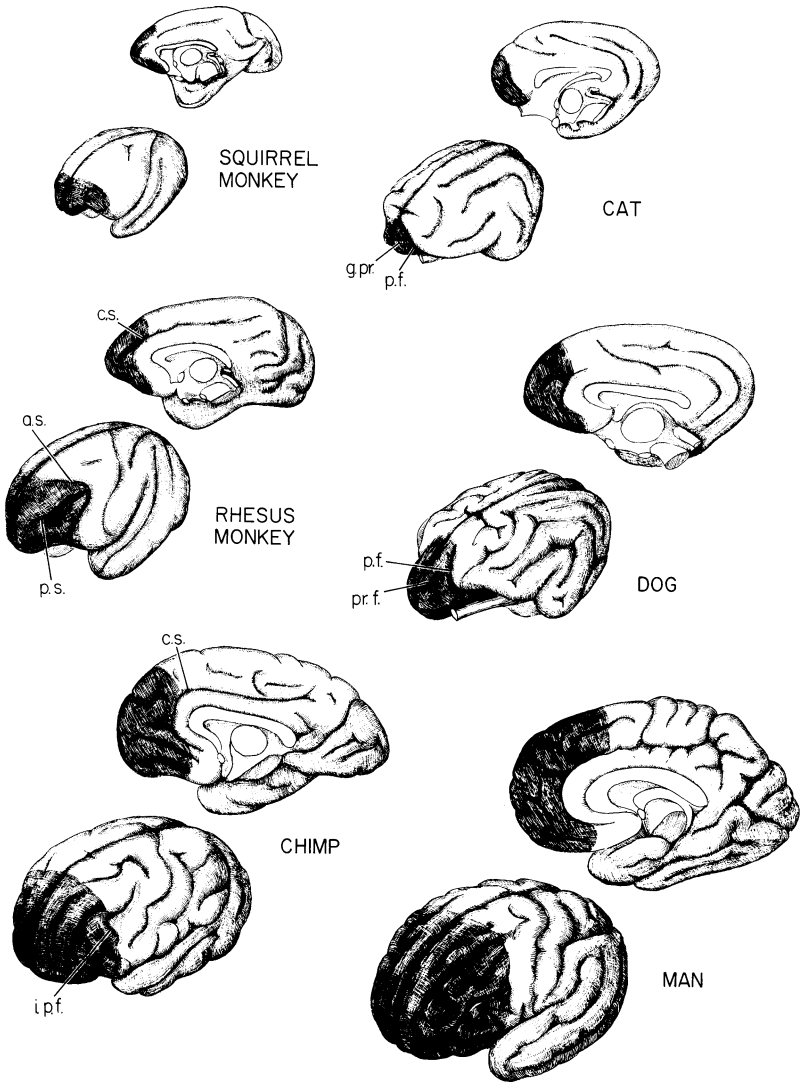

Figure 1
The prefrontal cortex (shaded) in six animal species

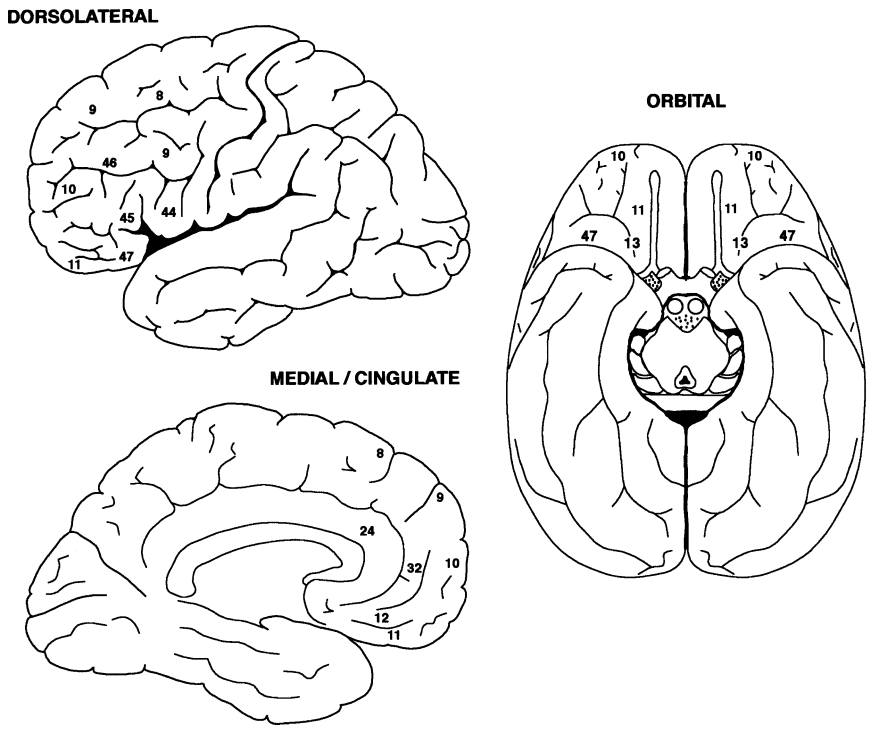

Figure 2
The three major regions of the prefrontal cortex in the human, with cytoarchitectonic areas numerated according to Brodmann's map

10, and 11); the medial and cingulate prefrontal cortex, which is nearly flat and faces the medial surface of the contralateral frontal pole (areas 12, 24 and 32, and medial parts of 8, 9, 10, and 11); and the inferior orbital prefrontal cortex, slightly concave and located directly above the orbit (areas 13, 47, and inferior parts of 10, 11, and 13).

Of the three regions, the dorsolateral develops last and most. This region is bordered posteriorly by the premotor cortex (area 6); the medial and orbital prefrontal cortices (often considered together and designated orbitomedial prefrontal cortex) are anterior and adjacent to the corpus callosum and to limbic structures (piriform cortex and amygdala, cingulate cortex, septum, and hypothalamus). In terms of cellular architecture, the posterior orbitomedial cortex is transitional, with features of both the limbic cortex and neocortex.

2. *Functional Connectivity*

The functions of the prefrontal cortex can be best understood as components of a general role of frontal cortex in the representation and execution of the actions of the organism. In its totality, the frontal cortex can be considered motor cortex in the broadest sense. It represents and coordinates actions in all domains of adaptation of the organism to its internal and external environment. In order to coordinate such a wide gamut of activity, the prefrontal cortex receives inputs from, and sends outputs to, many other parts of the brain.

Each of the three principal prefrontal regions has its own set of fiber connections mostly reciprocal and topologically organized—not only with the other two prefrontal regions, but with other cerebral structures, cortical and subcortical. Especially noteworthy for their number and functional implications are the connections of the three prefrontal regions with the thalamus, limbic structures, the basal ganglia, and other parts of the cortex. The orbital and medial prefrontal regions are well connected with the medial and anterior nuclei of the thalamus, the piriform cortex, the hippocampus, the amygdala, and the hypothalamus.

This connectivity with the thalamus and limbic structures falls into three major categories: (a) gustatory and olfactory inputs for the processing of information on taste and olfaction; (b) inputs from the internal milieu of the organism, notably from the hypothalamus, related to motivation, visceral functions, and affective states; and (c) efferent outputs to thalamic and limbic structures for the control of

instinctual, emotional, and social behavior, as well as certain visceral functions.

The dorsolateral prefrontal cortex, on the other hand, is extensively interconnected with other areas of the frontal lobe, homolaterally and contralaterally (through corpus callosum), with the hippocampus, and with several areas of the neocortex of the temporal and parietal lobes, in the posterior aspects of the cerebral hemisphere (Pandya and Yeterian 1985). All this cortical connectivity is essential for cognition. It serves the integrative and associative functions that support the cardinal role of the prefrontal cortex in executive memory and in the temporal organization of behavior. In addition, the dorsolateral prefrontal cortex sends important efferent fibers to the basal ganglia, including caudate nucleus, and the cerebellum. This connectivity serves the motor output functions of the prefrontal cortex through structures that are involved in skeletal and ocular motilities.

3. *Physiology*

Drive, emotional behavior and visceral actions, as well as their affective connotations, are the physiological purview of medial and orbital prefrontal areas. The medial prefrontal cortex plays an important role in basic motivation, though this role is still poorly understood.

In the human, large lesions of this medial cortex lead to apathy, inattention, and lack of spontaneity in speech, behavior, and ideation (Cummings 1985). The orbital prefrontal cortex, on the other hand, seems essential for the inhibitory control of internal drives. Animals or humans with lesions of this cortex often manifest disinhibition of eating, sex, and aggression, in addition to related abnormalities of social and emotional behavior. The human subject with orbitofrontal damage commonly shows uninhibited instinctual behavior, impulsivity, reckless risk taking, coarse and tasteless humor, and poor moral judgement (Stuss and Benson 1986, Damasio et al. 1994). Oftentimes he or she runs into conflict with the law. Further, because internal impulse control is necessary for steady attention and normal goal directed behavior, the orbitomedial prefrontal cortex serves the general role of the prefrontal cortex in the organization of behavior and cognition.

The dorsolateral prefrontal cortex is involved in both the representation and the enactment of temporally organized behavior and of cognitive operations, which are required in language and logical reasoning (Luria 1966, Fuster 1997). These representations fall under the general category of motor or executive memory, which includes the representations of complex schemes of temporally organized action, such as plans, programs, or scripts of behavior and cognition. The qualifications of novelty and complexity should be emphasized because, after its acquisition and consolidation, routine behavior is no longer represented or enacted by the prefrontal cortex, but by neural structures at lower stages of the motor hierarchy (e.g., motor and premotor cortices, basal ganglia).

A subject of much debate is the functional specialization of areas within the dorsolateral prefrontal cortex. Given that this cortex can be anatomically subdivided into a number of cytoarchitectonically disctinct areas (Petrides and Pandya 1994), each with a separate set of connections (Pandya and Yeterian 1985), the functional specialization of its areas can be, in principle, reasonably assumed. Accordingly, some single-cell studies with monkeys support a degree of seemingly modular commitment of various dorsolateral areas to the representation of different categories of sensory information, such as spatial location or visual configuration (Goldman-Rakic 1995). Other single-unit studies, however, point to the essentially associative character of all dorsolateral prefrontal areas.

In the pursuit of behavioral goals, dorsolateral prefrontal cells associate sensory information across modalities (e.g., sight and sound) and across time (Fuster et al. 2000). Furthermore, some cells associate sensory stimuli with subsequent motor actions (Rainer et al. 1998). These associative properties of dorsolateral prefrontal cells point to their essential temporal integrative functions. Thus, the apparent sensory or motor specificity of certain dorsolateral areas may be related to the specificity of their predominant inputs from sensory systems or outputs to motor systems. In any case, that specificity of areas—thus far fragmentary, and far from established—seems secondary to their basic integrative functions at the cognitive level. Those functions (see later) are crucial for the temporal organization of behavior.

It is likely, though not proven, that the connectivity of the prefrontal cortex with the hippocampus is essential for the formation of executive memory, including the schemas of behavioral and cognitive action. Once they have been formed, the executive memory networks of the cortex of the frontal lobe appear hierarchically organized. Represented in primary motor cortex, the lowest stage of that organization, are the simplest motor memories, elements of motor action defined by specific movement and muscle group (i.e., innate or phyletic motor memory). Immediately above, in the premotor cortex, are representations defined by trajectory and goal of movement. At the summit, in the dorsolateral prefrontal cortex, are representations of complex schemas and sequences of goal-directed action. In a manner still unknown, this cortical region represents the broad schemas of action and, in addition, is critically involved in the enactment of those schemas. A consistent component of the dorsolateral frontal-lobe syndrome in the human is the inability to formulate and to enact plans of action.

4. Temporal Structuring of Action

The main function of the dorsolateral prefrontal cortex is the temporal organization of behavior, language and reasoning. This cortical region appears to secure the orderly and timely execution of all new goal-directed sequences in those three domains of activity. As a precondition for its enactment, however, a new or recently learned sequence of acts is presumably represented in prefrontal networks in the form of a schema or plan of action (executive memory). The execution of that sequence is a highly complex process of temporal integration that probably engages many cortical and subcortical brain regions. The dorsolateral prefrontal cortex seems to orchestrate the functions of many other parts of the brain. It does so with the purpose and end-result of structuring behavior and cognition in time, of forming temporal gestalts (or melodies of action) toward the attainment of biological and cognitive goals.

In as much as intelligence is manifested and tested by performance of logical goal-directed directed tasks and sequences of behavior and language, it is reasonable to assume that the dorsolateral prefrontal cortex is an important component of the neural substrate of intelligence. Indeed, it has been argued on the basis of empirical data from the human, that this cortex is the foundation of general intelligence, as measured by Spearman (Duncan et al. 1996). This psychological metric reflects the proficiency of a subject in performance of a large variety of tests, all of which require the temporal integration of behavioral, verbal or cognitive contents.

Towards the end of the twentieth century, evidence has been accumulating for two temporally integrative functions of dorsolateral prefrontal cortex providing the neural and cognitive support to the process of temporal organization. Those two functions are temporally symmetrical and mutually complementary. The first is a temporally retrospective function of short-term memory; the second, a temporally prospective function of preparatory set. Both together help the organism to bridge time between sensory and motor events, and thus to mediate cross-temporal contingencies of behavior. Both seem to share the same or overlapping networks of dorsolateral prefrontal cortex. Both depend on the coordination of neural activity between this cortex and the association cortices of the posterior regions of the cerebral hemispheres.

4.1 Short-term Memory

Short-term memory is commonly conceptualized as the temporary retention of information before it becomes permanent or long-term memory. Recently, however, cognitive psychology has developed a different concept of short-term memory: the temporary retention of information, old or new, for the execution of an action in the short term, as is required for the performance of complex behavior or the solution of problems. This kind of short-term memory has been called working memory (Baddeley 1986). There is considerable evidence that both types of short-term memory share the same cortical neurons and networks, and that working memory consists in the sustained activation of a wide cortical network representing information to be used prospectively in the near term. That information may be newly acquired or may be an item of long-term memory reactivated for the short term as the occasion demands. As that memory contributes to form a temporal structure of behavioral or cognitive action, the neurons of dorsolateral prefrontal cortex are part of the activated memory network. There is conclusive evidence that the dorsolateral prefrontal region is critically involved in all forms of active short-term memory toward a goal, or working memory (Fuster 1997).

The evidence comes chiefly from two methodologies, selective cortical lesion (or inactivation) and single-cell recording, in the nonhuman primate performing short-term memory or delay tasks (e.g., delayed response, delayed matching). In these tasks, the animal is presented on successive trials with an item of sensory information—which varies from trial to trial—and is required to retain it in memory for electing and executing the appropriate motor response a few seconds or minutes later. Each trial amounts to a specific selection from a set repertoire of actions, a selection that depends on the sensory cue that has preceded it by seconds or minutes. The relative novelty of the relevant sensory–motor association for that particular trial, and the need to retain a sensory item temporarily in memory, place both the trial and the memory under prefrontal control. Monkeys with lesions of dorsolateral prefrontal cortex perform poorly in delay tasks. They seem unable to keep the working memory of the stimulus for each trial. In a normal animal performing a delay task, neurons of that cortex fire persistently—and often in stimulus-specific manner—during the delay (memory period) of every trial.

These observations denote the importance of cellular dynamics in dorsolateral prefrontal cortex for working memory and the temporal organization of behavior. Recent neuroimaging methods have provided further evidence, at a more global level, of the involvement of the dorsolateral prefrontal cortex in working memory (Smith and Jonides 1999). Unlike microelectrode studies, however, and because of its constraints of temporal and spatial resolution, neuroimaging does not yet reveal the dynamics of neuronal populations of this cortex in working memory. What defines prefrontal memory in the active state, or working memory, is the teleological quality of that memory, which is mobilized for the construction of future action. As noted in Sect. 3, one source of

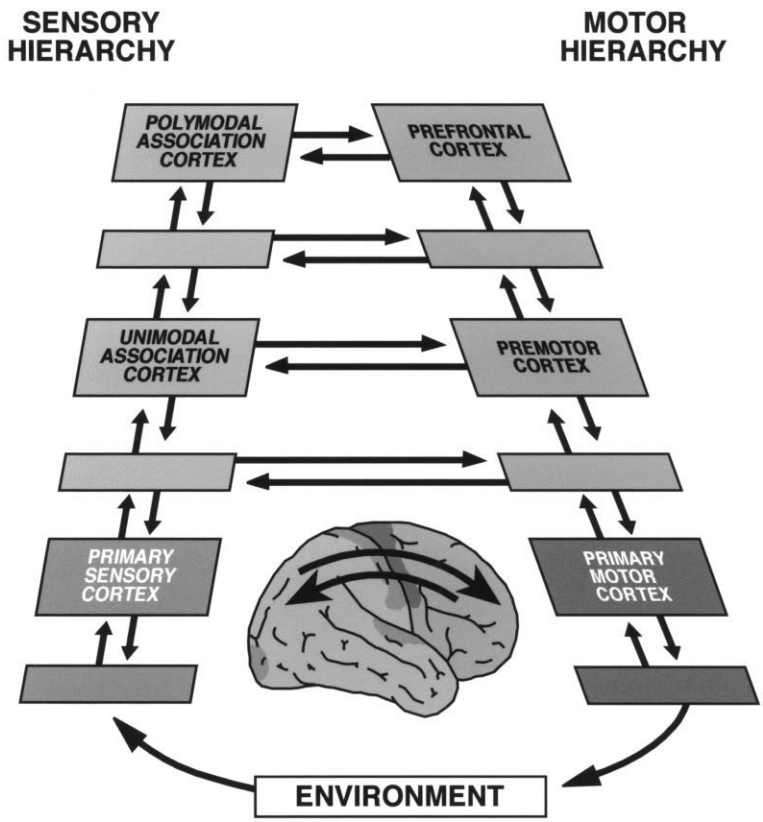

Figure 3
Cortical anatomy of the perception–action cycle depicted around a lateral view of the human brain. Unlabelled regions represent sub-areas of labelled regions or intermediate areas between them. All the arrows represent connections that have been demonstrated in the monkey

apparent areal specialization in dorsolateral prefrontal cortex is the nature of sensory inputs leading to associated action (visual, spatial, etc.). The other is the nature of the associated actions themselves (skeletal movements, ocular movements, etc.). Sensory- or motor-specific cell groupings and areas within that cortex are probably components of teleological associations, and thus the paths to prospective action. In any case, the integrative functions of dorsolateral cortex, namely memory and set, are supramodal and transcend the sensory or motor properties of any given prefrontal area.

4.2 Set

Also from the teleological character of the temporal structuring of action derives the prefrontal function of preparatory set. Setting receptors and effectors for prospective actions is the temporal mirror image of active short-term memory. This second integrative function, preparatory set, is the other side of the coin of the superordinate prefrontal function of temporal organization. It prepares the organism for anticipated percepts and actions (memory of the future) that will complete the bridging of a cross-temporal contingency. Although the mechanisms of set are not yet known, they probably include the priming of sensory and motor structures via efferent prefrontal connections. In the monkey and the human, the activation of dorsolateral prefrontal cortex in preparatory set is manifested by a progressive increase of cell discharge and slow surface-negative potentials before a stimulus-contingent act. The cellular evidence indicates that memory and set occur largely at the same time. Both seem to be processed simultaneously in the period between two mutually contingent events.

While some prefrontal cells engage in working memory of the cue, others in the same region do it in the preparation of the upcoming behavioral response. The activity of those cells shows that, while the representation of sensory information tends to wane over a delay, the representation of the prepared response tends to increase (Quintana and Fuster 1999).

The choice of one motor act among alternatives is, on the motor side, the equivalent of focusing sensory attention. Thus, preparatory set, the second cognitive function of the dorsolateral prefrontal cortex, can be rightly considered attention, the focusing of attention on a motor act to ensure its prompt and efficient execution.

In summary, the symmetry of the two temporal integrative functions of prefrontal cortex, memory and set, is paralleled by a symmetry of attentive processes, one sensory and the other motor. Active short-term memory is attention focused on the representation of a recent sensory event, whereas preparatory set is attention focused on a consequent and subsequent act. Although attention—especially perceptual attention in the human—is usually associated with consciousness, neither of the two temporal integrative functions of the prefrontal cortex—active memory or set—need be conscious.

5. *Perception–Action Cycle*

The organization of new and complex behavior commonly requires a continuous succession of temporal integrations of percepts and actions. In this succession, each integration depends to some degree on previous ones. Each act produces changes in the environment that will determine and modify subsequent perceptions, and these will determine and modify subsequent acts. The continuous operation of this cybernetic cycle of interactions of the organism with its environment is a basic principle of biology, the so-called perception–action cycle.

As complex new behavior evolves toward a goal, percepts and acts that are mutually contingent are often separated by time and by other intervening acts and percepts. Further, the pursuit of that goal requires the intervening attainment of lesser or subordinate goals. Thus, the temporal organization of the behavior requires a continuous mediation of contingencies across time. To mediate those cross-temporal contingencies is a function of the biological substrate of the perception–action cycle. The dorsolateral prefrontal cortex is at the apex of that substrate, which basically consists of two parallel hierarchies of neural structures, one sensory and the other motor, extending from the spinal cord to the cerebral cortex. Corresponding levels of each hierarchy are reciprocally interconnected, while each structure in either hierarchy is reciprocally interconnected with the one above and the one below. Figure 3 shows the cortical stages of those hierarchies. The contingencies of automatic and routine behavior (e.g., walking, reflex acts) are mediated at lower, subcortical, stages of the neural organization of the perception action cycle.

Complex new behavior, however, engages the cortical stages of the cycle neural hierarchies. Because this kind of behavior usually contains multiple and complex contingencies across time, the highest neocortical stages are called into action for its implementation, in particular the dorsolateral prefrontal cortex. This cortex, in functional coordination with the sensory association cortices of the posterior hemisphere, mediates cross-temporal contingencies at the top of the cycle. The prefrontal cortex supports that mediation across time through its two temporal integrative functions of active memory and preparatory set. Both require the functional interaction of prefrontal with posterior cortices, as suggested by combined use of behavioral, cortical-cooling and cell-recording methods (Fuster et al. 1985).

The understanding of the mechanisms of that interaction will lead to more precise knowledge than we now have of the functions of the prefrontal cortex in the temporal organization of behavior. For a number of empirical and theoretical reasons, the role of this cortex in the organization of language and reasoning is probably based on similar functions and mechanisms.

See also: Cognitive Control (Executive Functions): Role of Prefrontal Cortex; Prefrontal Cortex Development and Development of Cognitive Function; Short-term Memory, Cognitive Psychology of; Short-term Memory: Psychological and Neural Aspects; Working Memory, Neural Basis of; Working Memory, Psychology of

Bibliography

Baddeley A 1986 *Working Memory*. Clarendon Press, Oxford, UK

Brodmann K 1909 *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth, Leipzig, Germany

Cummings J L 1985 *Clinical Neuropsychiatry*. Grune and Stratton, Orlando, FL

Damasio H, Grabowski T, Frank R, Galaburda A M, Damasio A R 1994 The return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science* **264**: 1102–5

Duncan J, Emslie H, Williams P 1996 Intelligence and the frontal lobe: the organization of goal-directed behavior. *Cognitive Psychology* **30**: 257–303

Fuster J M 1997 *The Prefrontal Cortex: Anatomy Physiology, and Neuropsychology of the Frontal Lobe*, 3rd edn. Lippincott-Raven, Philadelphia, PA

Fuster J M, Bauer R H, Jervey J P 1985 Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Research* **330**: 299–307

Fuster J M, Bodner M, Kroger J 2000 Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature* **405**: 347–51

Goldman-Rakic P S 1995 Architecture of the prefrontal cortex and the central executive. *Proceedings of the National Academy of Science* **769**: 71–83

Luria A R 1966 *Higher Cortical Functions in Man*. Basic Books, New York

Pandya D N, Yeterian E H 1985 Architecture and connections of cortical association areas. In: Peters A, Jones E G (eds.) *Cerebral Cortex*. Plenum Press, New York, Vol. 4, pp. 3–61

Petrides M, Pandya D N 1994 Comparative architectonic

analysis of the human and the macaque frontal cortex. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*. Elsevier, Amsterdam, pp. 17–58
Quintana J, Fuster J M 1999 From perception to action: Temporal integrative functions of prefrontal and parietal neurons. *Cerebral Cortex* **9**: 213–21
Rainer G, Assad W F, Miller E 1998 Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* **393**: 577–9
Raz N 2000 Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In: Craik F I M, Salthouse T A (eds.) *Handbook of Aging and Cognition*. Erlbaum, Mahwah, NJ, pp. 1–90
Smith E E, Jonides J 1999 Storage and executive processes in the frontal lobes. *Science* **283**: 1657–61
Stuss D T, Benson D F 1986 *The Frontal Lobes*. Raven Press, New York

J. M. Fuster

Prefrontal Cortex Development and Development of Cognitive Function

Prefrontal cortex (PFC) is by far the largest cortical area in the human brain (see *Prefrontal Cortex*). It appears to be required when a task is novel or complicated and thus requires concentration. An example would be when you need to guide your actions by information you are holding in mind, and must pay attention so that you act according to that information and not your natural inclination. Its period of maturation is particularly protracted. PFC shows significant developmental changes as early as the first year of life, but does not reach full maturity in humans until young adulthood. In this article, developmental periods are broken up into 0–1, 1–3, 3–7, and 7–22 years. For each period, changes in cognitive abilities thought to depend on PFC are briefly described, as is evidence of maturational changes in PFC.

1. Behavioral Evidence of Improvements in Cognitive Functions that Depend on PFC During the First Year of Life

Piaget's 'A-not-B' task has been widely used to study infant cognitive development (Wellman et al. 1987). Under the name 'delayed response,' the almost identical task has been widely used to study the functions of a subregion of PFC in rhesus monkeys (Goldman-Rakic 1987). The subregion is dorsolateral PFC (DL-PFC). A participant watches as a reward is hidden to the left or right in one of two identical hiding places. A few seconds later the participant is encouraged to find the hidden treat. The participant must hold in mind over those few seconds where the treat was hidden and, over trials, must update his or her mental record to reflect where the treat was hidden last. The participant is rewarded for reaching correctly by being allowed to retrieve the treat. In this way, the behavior of reaching to that location is reinforced and hence the tendency to make that response is strengthened. When the reward is then hidden at the other location, the participant must inhibit the natural tendency to repeat the rewarded response and instead respond according to the representation held in mind of where the reward was hidden last. Hence, this task requires an aspect of working memory (holding information in mind) plus inhibition of a prepotent action tendency (the tendency to repeat a positively reinforced response).

By roughly $7\frac{1}{2}$ to 8 months of age, infants reach correctly to the first hiding location with delays as long as two or three seconds (Diamond 1985). When the reward is then hidden at the other hiding place, however, infants err by going back to the first location (called the 'A-not-B error'). Infants show marked improvements in their performance of the A-not-B/delayed response task at $7\frac{1}{2}$ to 12 months of age. For example, each month they can withstand delays approximately two seconds longer, so that by 12 months of age delays of roughly 10 seconds or longer are needed to see the A-not-B error (Diamond 1985).

In a transparent barrier detour task called 'object retrieval' (Diamond 1991) a toy is placed within easy reach in a small, clear box, open on one side. There is a very strong pull to reach straight for the toy through the side one is looking, which must be inhibited when an infant is looking through a closed side of the box. At 6 to 8 months of age, infants reach only at the side through which they are looking. They must look through the opening, and continue to do so, to reach in and retrieve the toy. As they get older, the memory of having looked through the opening is enough; infants can look through the opening, sit up, and reach in while looking through a closed side. By 11 or 12 months of age, infants do not need to look along the line of reach at all. Infants progress through a well-demarcated series of five stages in performance of this task from 6 to 12 months of age (Diamond 1991).

Although the A-not-B/delayed response and object retrieval tasks share few surface similarities, human infants improve on these tasks during the same age period (6–12 months) and so do infant rhesus monkeys ($1\frac{1}{2}$–4 months; Diamond 1991). Despite wide individual differences in the rate at which infants improve on these tasks, the age at which a given infant reaches 'Phase 1B' on the object retrieval task is remarkably close to the age at which that same infant can first uncover a hidden object in the A-not-B/delayed response paradigm (Diamond 1991).

There is no behavioral task more firmly linked to DL-PFC than the A-not-B/delayed response task (Goldman-Rakic 1987, Fuster 1989). This is one of the strongest brain–behavior relations in all of cognitive

neuroscience. DL-PFC lesions in the monkey also disrupt performance on the object retrieval task (Diamond 1991). MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) injections, which reduce the level of dopamine in PFC, also produce deficits on the task (Taylor et al. 1990, Schneider and Roeltgen 1993). (MPTP also affects the level of dopamine in the striatum, but lesions of the striatum do not impair performance on the object retrieval task (Crofts et al. 1999).)

Human infants of 7 to 9 months, infant monkeys of $1\frac{1}{2}$ to $2\frac{1}{2}$ months, adult monkeys in whom DL-PFC has been removed, infant monkeys of 5 months in whom DL-PFC was removed at 4 months, and adult monkeys who have received MPTP injections to disrupt the prefrontal dopamine system all fail the A-not-B/delayed response and object retrieval tasks under the same conditions and in the same ways (Diamond 1991). In human infants, changes in the pattern of electrical activity detected by electroencephalogram (EEG) over frontal cortex and in the coherence of electrical activity detected by EEG over frontal cortex and parietal cortex (A-not-B: Bell and Fox 1992, object retrieval: N A Fox, personal communication 1992) correlate closely with developmental improvements on the A-not-B/delayed response and object retrieval tasks. This does not prove that developmental improvements on these tasks rely in part on maturational changes in DL-PFC, but it is consistent with that hypothesis.

2. *Anatomical and Biochemical Evidence of Maturational Changes in PFC During the First Year of Life*

Nerve cells consist of axons, dendrites, and a cell body. Dendrites, which branch off from the cell body, represent the largest part of the receptive surface of a neuron. In humans, the period of marked growth in the length and extent of the dendritic branches of pyramidal neurons in layer III of DL-PFC is $7\frac{1}{2}$ to 12 months (Koenderink et al. 1994), exactly coinciding with the period when human infants are improving on the A-not-B/delayed response and object retrieval tasks. Pyramidal neurons of this PFC region have a relatively short dendritic extent in $7\frac{1}{2}$-month-old infants. By 12 months of age, they have reached their full mature extension. The surface of the cell bodies of these neurons also increases between $7\frac{1}{2}$ and 12 months (Koenderink et al. 1994). The level of glucose metabolism in DL-PFC increases during this period as well, and approximates adult levels by 1 year of age (Chugani et al. 1987).

Dopamine is an important neurotransmitter in PFC. During the period that infant rhesus monkeys are improving on the A-not-B/delayed response and object retrieval tasks (1.5–4 months), the level of dopamine is increasing in their brain (Brown and Goldman 1977), the density of dopamine receptors in their PFC is increasing (Lidow and Rakic 1992), and the distribution within their DL-PFC of axons containing the rate-limiting enzyme (tyrosine hydroxylase) for the production of dopamine is changing markedly (Rosenberg and Lewis 1994).

Acetylcholinesterase (AChE) is an enzyme essential for metabolizing another neurotransmitter, acetylcholine. In humans, the pattern of AChE staining in DL-PFC changes dramatically during the first year of life (Kostovic et al. 1988).

3. *Behavioral Evidence of Improvements in Cognitive Functions that Depend on PFC From Age 1–3 Years*

This is the period for which we know the least about changes in PFC anatomy or in cognitive functions dependent on PFC. Kochanska et al. (2000) report that the ability to inhibit a prepotent response in order to perform a modulated or different response improves markedly from 22 to 33 months of age and that the consistency across their various measures of inhibition also increases between 22 and 33 months.

Using a lazy susan to bring a toy within reach requires relating the lazy susan and its movement to the toy and its movement. It also requires inhibition of trying to reach on a direct line of sight (as the younger children try to do) and inhibition of the tendency to push the lazy susan in the direction you want the toy to go (you must push left to get the toy to approach on the right). Children improve in their ability to do this between 12 and 24 months (Koslowski and Bruner 1972). Case (1992) similarly reports marked improvements on a simple balance beam problem between $1\frac{1}{2}$ and $2\frac{1}{2}$ years.

4. *Evidence of Maturational Changes in PFC From Age 1–3 Years*

Almost nothing is known about changes in prefrontal cortex during this period. AChE reactivity of layer III pyramidal neurons begins to develop at this time (Kostovic et al. 1988), but that is surely not the only change in prefrontal cortex between 1 and 3 years of age.

5. *Behavioral Evidence of Improvements in Cognitive Functions that Depend on PFC From Age 3–7 Years*

Clear improvements in tasks that require holding information in mind plus inhibition are seen between 3 and 7 years of age. Three-year-olds make an error reminiscent of infants' A-not-B error, but with a more difficult task. On this task, 3-year-olds sort cards

correctly by the first criterion (whether it is color or shape: Zelazo et al. 1995), just as infants and prefrontally lesioned monkeys are correct at the first hiding place, and just as adults with PFC damage sort cards correctly according to the first criterion (Wisconsin Card Sort test: Milner 1964, Stuss et al. 2000). Three-year-olds err when they must switch to a new sorting criterion, e.g., when cards previously sorted by color must now be sorted by shape, just as infants of $7\frac{1}{2}$ to 9 months and prefrontally lesioned monkeys err when required to switch and search for the reward at the other location, and just as adults with PFC damage err when required to switch to a new sorting criterion. Although 3-year-old children fail to sort by the new sorting criterion, they can correctly state the new criterion (Zelazo et al. 1996), as is sometimes seen with adult patients who have sustained damage to PFC (Luria and Homskaya 1964, Milner 1964). Infants, too, sometimes indicate that they know the correct answer on the A-not-B task, by looking at the correct well, although they reach back incorrectly to the well that was previously correct (Hofstadter and Reznick 1996).

Another example of apparently knowing the correct answer, but not being able to act in accord with it is provided by work with 'go/no-go' tasks (Tikhomirov 1978, Livesey and Morgan 1991). Here, the participant is to respond to one stimulus but do nothing when shown another. Children between 3 and 4 years old repeat the instructions correctly, but they cannot get themselves to act accordingly; they respond even to the 'no-go' stimulus. By 5 years, they perform well on the card sort and go/no-go tasks.

The problem on the card sort task appears to be in (a) integrating two dimensions of a single stimulus (e.g., if children are used to focusing on whether stimuli are red or blue, they have trouble refocusing and concentrating instead on whether the same stimuli are trucks or stars) and (b) inhibiting the tendency to repeat their previously correct response. Similarly, children of 3 years have difficulty with 'appearance–reality' tasks (e.g., Flavell 1986) where, for example, they are presented with a sponge that looks like a rock. Three-year-olds typically report that it looks like a rock and really is a rock, whereas children of 4 or 5 years correctly answer that it looks like a rock but really is a sponge. The problem for the younger children is in relating two conflicting identities of the same object (e.g., Rice et al. 1997) and in inhibiting the response that matches their perception. Manipulations that reduce perceptual salience, by removing the object during questioning, enable children of 3 or 4 years to perform much better (e.g., Heberle et al. 1999).

'Theory of mind' and 'false belief' tasks are other tasks that require holding two things in mind about the same situation (the true state of affairs and the false belief of another person) and inhibiting the impulse to give the veridical answer. Here, as well, manipulations that reduce the perceptual salience of the true state of affairs aid children of 3 to 4 years (e.g., Zaitchik 1991). Carlson et al. (1998) reasoned that pointing veridically is likely to be a well-practiced and reinforced response in young children, and that children of 3 to 4 years have trouble inhibiting that tendency when they should point to the false location on false belief tasks. Carlson et al. (1998) found that when they gave children a novel response by which to indicate the false location, children of 3 to 4 years performed much better.

Many of the advances of Piaget's 'preoperational' child of 5 to 7 years over a child of 3 to 4 years, who is in the stage of 'concrete operations,' similarly reflect the development of the ability to hold more than one thing in mind and to inhibit the strongest response tendency of the moment. For example, children of 3 or 4 years fail tests of liquid conservation (they do not attend to both height and width, attending only to the most perceptually salient of the two dimensions) and they fail tests of perspective-taking where they must mentally manipulate a scene to indicate what it would look like from another perspective and must inhibit the tendency to give the most salient response (i.e., their current perspective). By 5 or 6 years, they can do these things. Since part of the difficulty posed by Piaget's liquid conservation task is the salience of the visual perception that the tall, thin container contains more liquid, placing an opaque screen between the child and the containers before the child answers enables younger children to perform better (Bruner 1964).

In the 'delay of gratification' paradigm, when faced with the choice of a smaller, immediate reward or a later, larger reward, children of 3 to 4 years are unable to inhibit going for the immediate reward although they would prefer the larger one. By 5 or 6 years of age, children are better able to wait for the preferred reward (Mischel and Mischel 1983).

Between 4 and 5 years of age children show improvement on several tests in the CANTAB battery that are designed to assess frontal lobe function (Luciana and Nelson 1998). For example, their improved ability to retain temporal order information in memory is demonstrated by better performance at 5 years on (a) a test of spatial memory span that is similar to the children's game, Simon, or the Corsi block task (the child must touch boxes displayed on a computer screen in the order in which he or she just saw them change color); and on (b) a spatial self-ordered pointing task where the child must keep track of which squares he or she has already touched.

The 'day–night' task (Gerstadt et al. 1994) requires that children hold two rules in mind and inhibit the tendency to say what the stimuli really represent; instead they must say the opposite ('Say "night" when shown a white card with a picture of the sun, and say "day" when shown a black card with a picture of the moon and stars'). Children of $3\frac{1}{2}$ to $4\frac{1}{2}$ years find the task very difficult; by 6 or 7 years it is trivially easy. If

abstract designs are used as the stimuli, even children of $3\frac{1}{2}$ have no difficulty correctly saying 'day' to one and 'night' to the other (Gerstadt et al. 1994). Hence, the need to learn and remember two rules is not in itself sufficient to account for the poor performance of young children on the task.

Luria's 'tapping' test (Luria 1966) also requires (a) remembering two rules and (b) inhibiting a prepotent response, making the opposite response instead. Here, one needs to remember the rules, 'Tap once when the experimenter taps twice, and tap twice when the experimenter taps once,' and one needs to inhibit the tendency to mimic what the experimenter does. Adults with large frontal lobe lesions fail this task (Luria 1966). Children improve on the task over the same age range at which they improve on the day–night task (Diamond and Taylor 1996). Moreover, performance on the two tasks is correlated so that children whose performance on the day–night task is delayed or accelerated show a corresponding delay or acceleration in their performance on the tapping task (Diamond et al. 1997, Diamond 2001).

The 'counting span' and 'spatial span' tasks of Case (1972) require finding target stimuli, counting them or remembering their locations, holding that information in mind while finding target stimuli in new arrays, and keeping track of the answers computed on each trial so that they can be repeated back in order at the end. Thus, these tasks require temporal order memory and resisting interference from prior and interpolated activity. A meta-analysis of 12 cross-sectional studies revealed strong linear improvements on both tasks between $4\frac{1}{2}$ and $6\frac{1}{2}$ years (Case 1992). Performance on these tasks is highly correlated with performance on Piaget's 'balance beam' task, 'Raven's matrices,' and 'concept acquisition.' On the balance beam and Raven's matrices tests, 4-year-olds tend to answer accordingly to what is perceptually salient. By 7 or 8 years of age, they are able to take two dimensions, or two perspectives, into account and relate one to the other. They can mentally manipulate, combine, recombine, order, and reorder information.

6. *Anatomical Evidence of Maturational Changes in PFC From Age 3–7 Years*

The density of neurons in human DL-PFC is highest at birth and declines thereafter. At 2 years of age it is 55 percent above the adult mean, but by 7 it is only 10 percent above adult levels (Huttenlocher 1990). Thus there is a dramatic change in neuronal density in DL-PFC between the ages of 2 and 7. The synaptic density of layer III pyramidal cells in DL-PFC increases after birth and reaches its maximum at about 1 year of age; by 7 years of age the decrease in synaptic density is significant, though not yet down to adult level (Huttenlocher 1979). Another change during this period is a marked expansion in the dendritic trees of layer III pyramidal cells in human DL-PFC between 2 and 5 years of age (Mrzlijak et al. 1990). In addition, the density of neuropeptide Y-immunoreactive neurons in human DL-PFC increases between the ages of 2 and 4 years and 6 to 7 years (DeLalle et al. 1997).

7. *Behavioral Evidence of Improvements in Cognitive Functions that Depend on PFC From 7 Years to Adulthood*

Some abilities dependent on DL-PFC appear to reach their adult levels by about 11 years of age. The Wisconsin Card Sort test is one of the classic tests for studying PFC function in adults (Milner 1964, Stuss et al. 2000). Here, the participant must deduce the rule for sorting cards, which can be sorted by color, shape, or number. Children do not begin to reach adult levels of performance on this task until they are about 11 years old (Chelune and Baer 1986). Continuous, marked improvement is seen on the counting and spatial span tasks through age 8, and then improvement is less steep until performance asymptotes around 10 or 11 years of age (Case 1972, 1992). Pascual-Leone's tasks (the compound stimulus visual information and the digit placement tests) show similar developmental progressions (Pascual-Leone 1970, Case 1972). Hale et al. (1997) report that resistance to interference improves from 8 to 10 years, when it approximates adult levels.

Other abilities dependent on PFC continue to show improvement into adulthood. Zald and Iacono (1998) report that spatial working memory improves from 14 to 20 years of age. They found that 20-year-olds were significantly more accurate than 14-year-olds at indicating the location of an object in space using memory even after brief delays of only half a second. (There was no age difference in the ability to accurately indicate an object's location using visual feedback.) The improvement with age was in the ability to accurately get the information into working memory, not in the ability to hold it in mind for a longer period. Indeed, they found little difference over age in the rate of degradation of the memory.

Another task that requires precise spatial memory is the pattern span task (Wilson et al. 1987). On any given trial some of the squares in a grid pattern are filled and some unfilled. After a quick look, the participant is presented with the grid with one change; the participant is to find that change. Performance on this task improves dramatically between 5 and 11 years, when it reaches adult levels (Wilson et al. 1987).

Several bodies of work indicate that the ability to inhibit a prepotent response tendency continues to improve until early adulthood. In the 'directed forgetting' paradigm, participants are directed to forget some of the words they are shown and to remember others. Even children as old as roughly 11 years show more intrusions of the to-be-forgotten words than do

adults (e.g., Harnishfeger and Pope 1996, Lehman et al.). The 'anti-saccade' task requires the participant to suppress the tendency to reflexively look to a visual stimulus in the periphery, and instead look away in the opposite direction. It depends especially on the frontal eye fields, and the supplementary eye fields and DL-PFC (O'Driscoll et al. 1995, Luna et al.). Performance on the task improves continuously from 8 until about 20 to 25 years of age (Munoz et al. 1998, Luna et al. (in press)). Luna et al. (in press) report that while activation in the frontal eye fields, supplementary eye fields, and DL-PFC increased during anti-saccade performance in participants of all ages, increased activation of the thalamus, striatum, and cerebellum was seen only in adults, suggesting that the circuit connecting PFC with subcortical regions might mature late.

8. *Anatomical Evidence that PFC is not Fully Mature Until Adolescence or Early Adulthood*

One of the functions of glial cells is to provide an insulating sheath, called myelin, around the axons of neurons, which increases the speed of transmission of communication between neurons and which acts as an electrical insulator, decreasing lateral transmission and interference, thus improving the signal:noise ratio. Myelination of PFC is protracted and does not reach adult levels until adolescence (Huttenlocher 1970, Giedd et al. 1999). For example, using MRI and following the same children longitudinally, Giedd et al. (1999) were able to show that the amount of white matter (i.e., myelinated axons) increased linearly in frontal cortex from 4 to 13 years of age.

Portions of the neuron that are unmyelinated, such as the cell body, have a gray appearance. In their longitudinal study, Giedd et al. (1999) found that gray matter in frontal cortex increased until adolescence, reaching its maximum extent at 12 years of age for males and 11 years for females. However, in cross-sectional volumetric studies, Jernigan et al. (1991) and Sowell et al. (1999a) report *reductions* in gray matter volume between childhood and adolescence, with the most dramatic changes occurring in dorsal frontal and parietal cortex. Sowell et al. (2001) related these gray matter changes to cognitive performance and found that, between 7 and 16 years of age, gray matter in frontal cortex (which included in their analyses not only PFC, but motor cortex, and all cortex in between) decreased in size, and the ability to accurately remember which words had and had not been presented earlier (i.e., the ability to remember which words had been seen in the present context and to discriminate them from other familiar words) improved. Impressively, gray matter thinning in frontal cortex was significantly correlated with this source memory, independent of age. Indeed, while the relation between frontal cortex gray matter thinning and this ability remained significant controlling for age, the relation between age and source memory was no longer significant controlling for frontal gray matter changes.

Synaptogenesis occurs concurrently with myelination. Huttenlocher (1979) reports that the synaptic density of layer III pyramidal cells in DL-PFC increases until about the age of 1 year, and then decreases, finally reaching adult levels at about 16 years of age. Huttenlocher and Dabholkar (1997) report that the formation of synaptic contacts in DL-PFC reaches its maximum after 15 months of age, and synapse elimination occurs late in childhood, extending to mid-adolescence for DL-PFC.

Developmental changes in PFC continue into adulthood. Sowell et al. (1999b) report a reduction in the density of gray matter in frontal cortex between adolescence (12–16 years) and adulthood (23–30 years). Kostovic et al. (1988) report that acetylcholinesterase (AChE) reactivity of layer III pyramidal cells in DL-PFC, which begins to develop after the first postnatal year, reaches its peak intensity in young adults.

See also: Brain Development, Ontogenetic Neurobiology of; Cognitive Development in Childhood and Adolescence; Cognitive Development in Infancy: Neural Mechanisms; Piaget's Theory of Child Development; Prefrontal Cortex

Bibliography

Bell M A, Fox N A 1992 The relations between frontal brain electrical activity and cognitive development during infancy. *Child Development* **63**: 1142–63

Brown R M, Goldman P S 1977 Catecholamines in neocortex of rhesus monkeys: Regional distribution and ontogenetic development. *Brain Research* **127**: 576–80

Bruner J S 1964 The course of cognitive growth. *American Psychologist* **19**: 1–15

Carlson S M, Moses L J, Hix H R 1998 The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development* **69**: 672–91

Case R 1972 Validation of a neo-Piagetian capacity construct. *Journal of Experimental Child Psychology* **14**: 287–302

Case R 1992 *The Mind's Staircase: Exploring the Conceptual Underpinnings of Children's Thought and Knowledge*. L Erlbaum Associates, Hillsdale, NJ

Chelune G J, Baer R A 1986 Developmental norms for the Wisconsin Card Sorting Test. *Journal of Clinical and Experimental Neuropsychology* **8**: 219–28

Chugani H T, Phelps M E, Mazziotta J C 1987 Positron emission tomography study of human brain functional development. *Annals of Neurology* **22**: 487–97

Crofts H S, Herrero M T, Del Vecchio A, Wallis J D, Collins P, Everitt B J, Robbins T W, Roberts A C 1999 Excitotoxic lesions of the caudate nucleus in the marmoset: Comparison with prefrontal lesions on discrimination learning, object retrieval and spatial delayed response. *Society for Neuroscience Abstracts* **25**: 891

Delalle I, Evers P, Kostovic I, Uylings H B M 1997 Laminar distribution of neuropeptide Y-immunoreactive neurons in

human prefrontal cortex during development. *The Journal of Comparative Neurology* **379**: 515–22

Diamond A 1985 Development of the ability to use recall to guide action, as indicated by infants' performance on ABBAR. *Child Development* **56**: 868–83

Diamond A 1991 Neuropsychological insights into the meaning of object concept development. In: Carey S, Gelman R (eds.) *The Epigenesis of Mind: Essays on Biology and Cognition*. L Erlbaum Associates, Hillsdale, NJ, pp. 67–110

Diamond A 2001 A model system for studying the role of dopamine in prefrontal cortex during early development in humans. In: Nelson C A, Luciana M (eds.) *Handbook of Developmental Cognitive Neuroscience*. MIT Press, Cambridge, MA, pp. 433–72

Diamond A, Taylor C 1996 Development of an aspect of executive control: Development of the abilities to remember what I said and to 'Do as I say, not as I do'. *Developmental Psychobiology* **29**: 315–34

Diamond A, Prevor M B, Callender G, Druin D P 1997 Prefrontal cortex cognitive deficits in children treated early and continuously for PKU. *Monographs of the Society for Research in Child Development* **62**(4): Monograph 252

Flavell J H 1986 The development of children's knowledge about the appearance–reality distinction. *American Psychologist* **41**: 418–25

Fuster J M 1989 *The Prefrontal Cortex*, 2nd edn. Raven Press, New York

Gathercole S E 1998 The development of memory. *Journal of Child Psychology and Psychiatry* **39**: 3–27

Gerstadt C L, Hong Y J, Diamond A 1994 The relationship between cognition and action: Performance of 3.5–7 year old children on a Stroop-like day-night test. *Cognition* **53**: 129–53

Giedd J N, Blumenthal J, Jeffries N O, Castellanos F X, Liu H, Zijdenbos A, Paus T, Evans A C, Rapoport J L 1999 Brain-development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience* **2**: 861–3

Goldman-Rakic P S 1987 Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In: Plum F (ed.) *Handbook of Physiology*. American Physiological Society, Bethesda, MD, Vol. 5, pp. 373–417

Guitton D, Buchtel H A, Douglas R M 1985 Frontal lobe lesions in man cause difficulties in supressing reflexive glances and in generating goal-directed saccades. *Experimental Brain Research* **58**: 455–72

Hale S, Bronik M D, Fry A F 1997 Verbal and Spatial Working Memory in School-Age Children: Developmental Differences in Susceptibility to Interference. *Developmental Psychology* **33**: 364–71

Harnishfeger K K, Pope R S 1996 Intending to forget: The development of cognitive inhibition in directed forgetting. *Journal of Experimental Child Psychology* **62**: 292–315

Herberle J F, Clune M, Kelly K 1999 *Development of young children's understanding of the appearance–reality distinction*. Paper presented at the Biennial Meeting of the Society for Research in Child Development, Albuquerque, NM

Hofstadter M, Reznick J S 1996 Response modality affects human infant delayed-response performance. *Child Development* **67**: 646–58

Huttenlocher P R 1970 Myelination and the development of function in immature pyramidal tract. *Neurology* **29**: 405–15

Huttenlocher P R 1979 Synaptic density in human frontal cortex—developmental changes and effects of aging. *Brain Research* **163**: 195–205

Huttenlocher P R 1990 Morphometric study of human cerebral cortex development. *Neuropsychologia* **28**: 517–27

Huttenlocher P R, Dabholkar A S 1997 Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology* **387**: 167–78

Jernigan T L, Trauner D A, Hesselink J R, Tallal P A 1991 Maturation of human cerebrum observed in vivo during adolescence. *Brain* **114**: 2037–49

Kochanska G, Murray K T, Harlan E T 2000 Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology* **36**: 220–32

Koenderink M J Th, Ulyings H B M, Mrzljiak L 1994 Postnatal maturation of the layer III pyramidal neurons in the human prefrontal cortex: A quantitative Golgi analysis. *Brain Research* **653**: 173–82

Koslowski B, Bruner J S 1972 Learning to use a lever. *Child Development* **43**: 790–99

Kostovic I, Skavic J, Strinovic D 1988 Acetylcholinesterase in the human frontal associative cortex during the period of cognitive development: Early laminar shifts and late innervation of pyramidal neurons. *Neuroscience Letters* **90**: 107–12

Lehman E, Srokowski S A, Hall L C, Renkey M E (submitted) Directed forgetting of related words: Evidence for the inefficient inhibition hypothesis

Lidow M S, Rakic P 1992 Scheduling of monoaminergic neurotransmitter receptor expression in the primate neocortex during postnatal development. *Cerebral Cortex* **2**: 401–16

Livesey D J, Morgan G A 1991 The development of response inhibition in 4- and 5-year-old children. *Australian Journal of Psychology* **43**: 133–7

Luciana M, Nelson C A 1998 The functional emergence of prefrontally-guided working memory systems in four- to eight-year-old children. *Neuropsychologia* **36**(3): 273–93

Luna B, Merriam E P, Minshew N J, Keshavan M S, Genovese C R, Thulborn K R, Sweeney J A (in press) Development of brain function for voluntary response suppression from childhood through adolescence. *Neuroimage*

Luria A R 1966 *The Higher Cortical Functions in Man*. Basic Books, New York

Luria A R, Homskaya E D 1964 Disturbance in the regulative role of speech with frontal lobe lesions. In: Warren J M, Akert K (eds.) *The Frontal Granular Cortex and Behavior*. McGraw Hill, New York, pp. 353–71

Milner B 1964 Some effects of frontal lobectomy in man. In: Warren J M, Akert K (eds.) *The Frontal Granular Cortex and Behavior*. McGraw Hill, New York, pp. 313–34

Mischel H N, Mischel W 1983 The development of children's knowledge of self-control strategies. *Child Development* **54**: 603–19

Mrzlijak L, Uylings H B M, van Eden C G, Judas M 1990 Neuronal development in human prefrontal cortex in prenatal and postnatal states. In: Uylings H B M, van Eden C G, de Bruin J P C, Corner M A, Feenstra M G P (eds.) *The Prefrontal Cortex: Its Structure, Function, and Pathology*. *Progress in Brain Research*. Elsevier, Amsterdam, Vol. 85, pp. 185–222

Munoz D P, Broughton J R, Goldring J E, Armstrong I 1998 Age-related performance of human subjects on saccadic eye movement tasks. *Experimental Brain Research* **121**: 391–400

O'Driscoll G A, Alpert N M, Matthysse S W, Levy D L, Rauch S L, Holzman P S 1995 Functional neuroanatomy of antisaccade eye movements investigated with positron emission tomography. *Proceedings of the National Academy of Sciences (USA)* **92**: 925–9

Pascual-Leone J 1970 A mathematical model for transition—in Piaget's developmental stages. *Acta Psychologia* **32**: 301–45

Rice C, Koinis D, Sullivan K, Tager-Flusberg H, Winner E 1997 When 3-year-olds pass the appearance-reality test. *Developmental Psychology* **33**: 54–61
Rosenberg D R, Lewis D A 1994 Changes in the dopaminergic innervation of monkey prefrontal cortex during late postnatal development: A tyrosine hydroxylase immunohistochemical study. *Biological Psychiatry* **15**: 272–7
Schneider J S, Roeltgen D P 1993 Delayed matching-to-sample, object retrieval, and discrimination reversal deficits in chronic low dose MPTP-treated monkeys. *Brain Research* **615**: 351–4
Sowell E R, Delis D, Stiles J, Jernigan T L 2001 Improved memory functioning and frontal lobe maturation between childhood and adolescence: a structural MRI study. *Journal of the International Neuropsychological Society* **7**: 312–22
Sowell E R, Thompson P M, Holmes C J, Batth R, Jernigan T L, Toga A W 1999a Localizing age-related changes in brain structure between childhood and adolescence using statistical parametric mapping. *Neuroimage* **9**: 587–97
Sowell E R, Thompson P M, Holmes C J, Jernigan T L, Toga A W 1999b In vivo evidence for post-adolescent brain maturation in frontal and striatal regions. *Nature Neuroscience* **2**: 859–61
Stuss D T, Levine B, Alexander M P, Hong J, Palumbo C, Hamer L, Murphy K J, Izukawa D 2000 Wisconsin Card Sorting Test performance in patients with focal frontal and posterior brain damage: Effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia* **38**: 388–402
Taylor J R, Elsworth J D, Roth R H, Sladek J R Jr., Redmond D E Jr 1990 Cognitive and motor deficits in the acquisition of an object retrieval detour task in MPTP-treated monkeys. *Brain* **113**: 617–37
Tikhomirov O K 1978 The formation of voluntary movements in children of preschool age. In: Cole M (ed.) *The Selected Writings of A R Luria*. M E Sharpe, White Plains, NY, pp. 229–69
Wellman H M, Cross D, Bartsch K 1987 Infant search and object permanence: A meta-analysis of the A-not-B error. *Monographs of the Society for Research in Child Development* **51**: 1–67
Wilson J T L, Scott J H, Power K G 1987 Developmental differences in the span of visual memory for pattern. *British Journal of Developmental Psychology* **5**: 249–55
Zaitchik D 1991 Is only seeing really believing?: Sources of the true belief in the false belief task. *Cognitive Development* **6**: 91–103
Zald D H, Iacono W G 1998 The development of spatial working memory abilities. *Developmental Neuropsychology* **14**: 563–78
Zelazo P D, Reznick J S, Piñon D E 1995 Response control and the execution of verbal rules. *Developmental Psychology* **31**: 508–517
Zelazo P D, Frye D, Rapus T 1996 Age-related dissociation between knowing rules and using them. *Cognitive Development* **11**: 37–63

A. Diamond

Prehistoric Art

As far as we know, people have been making objects, forms and images that we consider to be 'art' for at least 35,000 years and perhaps even longer. Some claim that evidence of red ochre, as a kind of coloring material, indicates human engagement with artistic activities (such as body painting) for more than 200,000 years. Others have pointed to individual objects, dated to even earlier than that, which they are convinced were modified by ancient humans with an intent to represent something, suggesting that 'art' has long been a part of the cultural repertoire of humans, even of different biological species than we anatomically modern humans. And when most lay people learn that they are talking to an archaeologist, they inevitably ask, 'What is the most fantastic thing you have ever found?,' hoping we will be able to tell them about a gold decorated necklace, a carved figurine, or magnificent painted jar: they want to know about 'prehistoric art.'

1. *Definitions*

The very term, 'prehistoric art,' is difficult to define. Technically speaking, it refers to the 'art' made by prehistoric peoples; that is, by people in those societies without, and usually before the development of, written texts. There have in recent times been human societies without writing systems (nonliterate) and with what might be called 'the arts,' but these are more likely to be what we call 'ethnographic arts' because they are still 'living' arts and can be described ethnographically. There are some problems with the way in which the term 'prehistory' is used, including the implication that, without writing systems, prehistoric groups have no history, and experienced little change. But we know this is not the case at all, and that there are myriad ways other than in writing by which 'history' is reckoned, transmitted, and experienced. Also, in relation to the idea of 'prehistoric art,' there are some archaeologically known societies, such as the Maya or Egyptians, who had writing systems, but whose 'arts' are sometimes referred to as prehistoric or, more often, as ancient art. In fact, it is perhaps better to think of the arts that we are interested in as the archaeological arts, rather than prehistoric ones. With such a term, we realize we are dealing with the art(s) from past human societies that we study by using the wide range of archaeological methods, theories, and techniques.

Then there is the term 'art,' which is itself notoriously difficult to define. Furthermore, it is a term that has had differing meanings even within the last two centuries of Western scholarship and culture. Since the nineteenth century, 'we' in the Western tradition have associated the term, 'art,' with aesthetics and aesthetically-pleasing attributes; as something that is usually produced by specialists who are often considered to be gifted individuals with admirable degrees of creativity, and who are often separated from real-life, everyday activities. Our present-day notions of art include the idea that art is to be held in awe while, at the same time, the production

and products of art are embedded in the capitalist system, with economic as well as cultural value; art objects function in their own domain, that of high culture and aesthetics.

However, all this is unlikely to be applicable to the art-making and art images and forms that we are concerned with in prehistoric and archaeological contexts. It is a well-known truism in anthropology that in most cultures which have been studied ethnographically, there is rarely a word, or even a concept, equivalent to our word and concept, 'art.' It is not really possible to try to separate out 'art' from other aspects of cultural life, such as art as something distinct from religion, social relations, and/or politics. And yet there are images, forms, and cultural practices in other cultural contexts, both ethnographic and archaeological, that we acknowledge unproblematically as 'art,' mostly because, if they were produced by our own culture, we would classify them readily as art: cave and rock paintings, sculpted figurines or statues, decorated bronze goblets, and the like.

Is there a better term: visual culture? Material culture? Images? Representations? Material symbols? The term or label we use, such as 'art,' is not itself the real problem; what is more of an issue is that if we use such a term in our own sense, without recognizing its historical specificity, then what we mean by it is often left unexamined, and the analysis or interpretation of the phenomenon is led in certain directions because of the label. If what we have is 'art,' then there must be artists and aesthetic meanings. However, as Morphy (1999) has so eloquently noted: we need to develop a concept of art that 'neither distances the objects of other societies from Western cultures nor appropriates the object to European categories.'

Books that treat the topic of 'prehistoric art' as a single entity are rare; most book-length treatments are concerned with the art and imagery of specific regions (e.g., Keyser 1992; see also *Sub-Saharan Africa, Archaeology of*), time periods (e.g., Leroi-Gourhan 1967) or in specific media (Dowson 1992). In 1966, Powell authored an overview of prehistoric art, which was divided into four chapters: the art of the hunters, that of the cultivators, that of the metalworkers, and what he called the 'art of a barbarian nation,' referring to the Celts of Europe (Powell 1966). His coverage was limited to Europe, and he treated each 'kind' of art by apparent categories of 'lifestyle,' but it is just as much a book about the 'progress' of art from the Paleolithic period (the hunters), through the Neolithic (the cultivators), the Bronze and Iron Ages (the metalworkers), to the 'archaic state' (Celts). As such, he presents a now-problematic evolutionist view, a view of 'progress' through different chronocultural periods. Indeed, for much of the nineteenth and twentieth century, the ethnographic (and archaeological) arts have been embedded in a progressivist notion (Errington 1998); we used to talk unquestioningly of 'primitive art' (Price 1989).

More recently, Bahn (1998) has authored an illustrated history of prehistoric art, but it is almost exclusively about 'rock art,' with a short chapter on 'body art' and one on 'objets d'art,' or portable art (items that could circulate). He returns again and again to a baseline of the Ice Age imagery of Europe, although this book is an important source for appreciating that rock art—paintings, engravings, and petroglyphs on rock surfaces, rock shelters and in caves—is truly global in its distribution. Bahn's map of sites is a map of the world, while Powell's is one of Europe. This is indicative of a real shift of scope in our understanding and treatments of 'early art' and/or rock art from all times. As we will see below, the search for the earliest art has been one of the major preoccupations of many who study 'prehistoric art.'

2. *How to Study Prehistoric Art*

First, it is always important to recognize that the corpus of materials that we consider as 'prehistoric art' is certainly only a sample of what was made in the human past. There is always the problem of preservation, and that what we have in the present is only a sample of past artistic production. It is no coincidence that rock art, ceramic sculpture and decorated pots, metal objects, and images made out of ivory, bone, antler, shell, and beads of various sorts are some of the primary forms of prehistoric art. The wood, hide, basketry, cloth, netting and twining, feathers, and body art are rarely, if ever, preserved. If the arts of ethnographic groups are any indication, we have an impoverished archaeological sample of prehistoric visual and material culture.

But given the preservation of at least some portion of prehistoric artistic production, how do we then go about studying it? First, we need some theoretical or conceptual frameworks within which even to think about what the images and forms might be about, especially if we cannot directly apply our own notions of 'art,' artists, and aesthetics. Much of this has to come from related fields, such as the anthropology of art, the ethnographic arts, and art history. We need to draw on understandings of what visual and material representation is all about (e.g., Gombrich 1961, Bryson 1983), of symbolism (e.g., Geertz 1983), and of humans as materialists and symbolists (Conkey 1993); of how technology and materials are integral to the artforms (e.g., Lechtman 1984), of style (Conkey and Hastorf 1990); and of the social contexts and production of art (e.g., Wolff 1981), and of the 'social life of things' (e.g., Appadurai 1986). Studies of 'art' in ethnographic contexts (e.g., Morphy 1991), general overviews of the anthropology of art (e.g., Layton 1981), and ethnoarchaeological studies (see *Ethnoarchaeology*) provide important resources.

Second, we need to recognize the importance of the archaeological context of the so-called art forms, and

we need to take context into consideration when we try to interpret or understand the meanings, uses, or even the making of the art. Do the preconceptions that female figurines—from about 7,000 years ago from sites in such places as southeast Europe—are objects of worship and a female-centered world change when we recognize that some, if not many, of the figurines have been recovered from trash deposits, often broken? Are the Ice Age paintings of southwestern Europe found deep in caves, away from daylight, comparable to those found on daylit shelters? Is there a difference in the uses and meanings between objects found in everyday places and living sites, and those found in architectural structures that are so different in size, form, and location that they must be other-than-'houses'? The individual images and objects may well take their specific meanings from the locations where they were made and/or found, or from the other artifacts or structures with which they are associated.

We can use the most up-to-date techniques and methods, especially those of materials science, to make inferences about the images and objects: we can identify the sources of the raw materials, such as clay, stones, or pigments, which in turn can inform on trade, exchange, circulation of goods, and the social relations that effect such practices. We can identify different ways in which pigments were processed by Ice Age cave 'artists' in Europe (e.g., Clottes 1993), and therefore suggest the social relationships that must have obtained in order to have images in different sites made by the same pigment 'recipes.' We can show how certain fibers were used to apply pigment in Australian rock paintings (e.g., Cole and Watchman 1992). We can infer the stages of manufacture of metal earspools (e.g., Lechtman 1984) and how these stages resonate with other cultural principles and practices. We can use infrared photography to 'see' color and images that have long since faded away. We can use high powered microscopic analysis to detect the precise strokes of a stone tool on the surface of a bone object; from such analyses have come the suggestion that lunar calendars may have been made at least 15,000 years ago (Marshack 1991). These are just a few examples of how our increasingly sophisticated methods of materials science that are common in the analysis of other archaeological materials can be applied to the study of prehistoric art images and objects.

Both of these approaches—knowing the archaeological contexts of the art, and the application of scientific techniques—have featured prominently in the increasing ability to date many objects and images. It has long been the case that objects that are found in stratigraphic context—that is, in the geological/archaeological layers of an archaeological site—can, if excavated and recorded carefully, be attributed to at least a relative date (see *Chronology, Stratigraphy, and Dating Methods in Archaeology*). They can also be given more secure dates if other artifacts or materials in the same stratigraphic layer are suitable for some of the many techniques of archaeological or chronometric dating (see *Archaeometry* and *Chronology, Stratigraphy, and Dating Methods in Archaeology*). However, many art forms, such as rock art or objects that are unfortunately out-of-context, cannot be dated in this associative manner. And many of these images themselves cannot be dated at all, although recent advances have changed this somewhat. Earlier generations of scholars, for example, dated all the cave art of Europe or the varying rock art styles of Australia by means of stylistic criteria, often based on assumptions as to what 'ought' to be earlier image-making and what 'ought' to be later. At the same time, many researchers were reluctant even to consider the art as a subject worthy of archaeological study, primarily because it could not be dated (see Whitley and Loendorf 1994).

That much art cannot be dated has, in combination with the fact that the interpretation of art is never straightforward (if only because it is the product of symbolic cultural practices that are not easily comprehended through the usual archaeological methods), led many archaeologists to dismiss art and images as viable archaeological data. This has been the case especially when it comes to objects and images from many thousands of years ago, where there are no real historic links between the archaeological 'cultures' and those known ethnohistorically or ethnographically, where there is no direct historical method that can be applied (see *Theory in Archaeology*). Some have suggested that the recent explosion of interest in the global study of rock art is due, in part, to new dating methods that have allowed researchers to begin to date imagery with more security. For this genre of prehistoric art—rock art—the most important new technique has been the development of accelerator mass spectrometry (AMS) carbon-14 dating, which can be done using a very minute sample of material, thus not destroying a painting or image, although it can only be used if there are organic materials in the pigments, such as charcoal (see *Archaeometry*).

3. *Some Major Questions*

At least two major questions come to mind when we sit down to address the subject of prehistoric art: first there is the question of the origins or the 'birth' of art; second, there is the challenging issue of how we can interpret the art and approach what it might mean. Until recently, it was thought that the cave paintings and other arts of the Upper Paleolithic period (c. 30,000–10,000 years ago) in Europe comprised the 'first' art, even though images from ancient cultures had been found all over the globe (see Bahn 1998 on the 'discovery' of prehistoric art). This was a compelling idea because there were many different caves

with images and many other objects and art forms; because it seemed to be associated with the earliest evidence for anatomically modern humans—our own species, *Homo sapiens*; and because it was in Europe, which allowed for the idea of a continuous tradition of art—a crucial essence of what it means to be human—from deep prehistory up into the twenty-first century.

However, work during the last few decades of the twentieth century has undermined seriously the idea that there is any single 'birthplace' of art: both southern Africa and Australia have yielded rock art images that can be attributed to the same general time period as the European Paleolithic imagery. And there are those who point to more isolated finds (see Bahn 1998, pp. 86–7), although many are debated hotly as being 'art' or intentional representations at even earlier dates. And we now recognize that anatomically modern humans have a deeper history, perhaps as much as 200,000 years longer, than previously thought (see *Sub-Saharan Africa, Archaeology of*). However, the exact nature of the relationship between being anatomically modern and the appearance of representational and/or symbolic imagery is hardly resolved, but the fact that this kind of imagery appears in multiple locations, in varying forms and amounts, calls for a more contextual approach: what is it about life and social relations at these times that set a context within which 'art'—at least in more durable forms that have been preserved—became a part of cultural practices? While the search for 'the origins' of many phenomena in the human past is always a seductive question, as far as prehistoric art is concerned, it is now clear that there was no single 'center' for art-making, but that the preserved evidence around the globe attests the engagement of ancient humans with materials, images, and technologies (of working stone, of coloring materials, etc.) for cultural purposes that still elude us.

And thus, the final, but most difficult question is the interpretive one: how can we infer some (if any) of the meanings of prehistoric art? This question is both the most difficult and most complex question, and what everyone wants to know. Since all of archaeology is interpretive, while drawing on as much empirical evidence as possible, and since human behavior is so multivocal, with so many potential meanings and ambiguities, it is unlikely to expect a single answer, a single interpretation. Even in known or ethnographic contexts, the so-called meaning of art-making or of the art itself is ever-shifting, contextual, and multiple, dependent upon viewers, users, makers, and situations. Do we assume that prehistoric peoples were somehow 'simpler' and therefore a (simple) single account can suffice?

Despite the complexity of the challenge, there are some things that we can say about interpreting prehistoric art. In some instances, we can infer information clearly about clothing from human depictions, about activities as depicted in 'scenes,' preferred symbols or forms, and what kinds of object were traded or used in certain contexts. In other instances, we can advance general theories, such as the role of shamans in making images that derive from their visions (e.g., Lewis-Williams and Dowson 1989, Turpin 1994), such as the ways in which ritual landscapes were created with artforms and megaliths, such as the idea that some arts appear to be 'about' interpersonal social relations, whereas other arts appear to be 'about' maintaining the corporate entity (e.g., Layton 1985). While we are unlikely ever to fix, with certainty, what a particular picture or object 'meant,' recent past decades have witnessed enormous gains, thanks to the application of scientific techniques, the innovative and expanded theories for interpretation, and a more widespread recognition that the archaeological arts are just as crucial to our understanding of past societies as are other kinds of archaeological data.

Bibliography

Appaduri A 1986 *The Social Life of Things*. Cambridge University Press, Cambridge, UK

Bahn P 1998 *The Cambridge Illustrated History of Prehistoric Art*. Cambridge University Press, Cambridge, UK

Bryson N 1983 *Vision and Painting: The Logic of the Gaze*. Yale University Press, New Haven, CT

Clottes J 1993 Paint analyses from several Magdalenian caves in the Ariège region of France. *Journal of Archaeological Science* **20**: 223–35

Cole N, Watchman A 1992 Painting with plants: Investigating fibers in Aboriginal rock paintings at Laura, North Queensland. *Rock Art Research* **9**

Conkey M 1993 Humans as materialists and symbolists. In: Rasmussen D T (ed.) *The Origins and Evolution of Humans and Humanness*. Jones and Bartlett, Boston, MA, pp. 95–118

Conkey M, C Hastorf (eds.) 1990 *The Uses of Style in Archaeology*. Cambridge University Press, Cambridge, UK

Dowson T 1992 *Rock Engravings of Southern Africa*. Witwatersrand University Press, Johannesburg, South Africa

Errington S 1998 *The Death of Authentic Primitive Art and Other Tales of Progress*. University of California Press, Berkeley, CA

Geertz C 1983 Art as a cultural system. In: Geertz C (ed.) *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York

Gombrich E 1961 *Art and Illusion: A Study in the Psychology of Pictorial Representation*, 2nd edn. Pantheon, New York

Keyser J D 1992 *Indian Rock Art of the Columbia Plateau*. University of Washington Press, Seattle, WA

Layton R 1981 *The Anthropology of Art*. Columbia University Press, New York

Layton R 1985 The cultural context of hunter-gatherer rock art. *Man* (*N.S.*) **20**: 434–53

Lechtman H 1984 Andean value systems and the development of prehistoric metallurgy. *Technology and Culture* **25**(1): 1–36

Leroi-Gourhan A 1967 *Treasures of Prehistoric Art*. Harry Abrams, New York

Lewis-Williams J D, Dowson T 1989 *Images of Power: Understanding Bushman Rock Art*. Southern Book Publishers, Johannesburg, South Africa

Marshack A 1991 *The Roots of Civilization*. McGraw-Hill, New York
Morphy H 1991 *Ancestral Connections: Art and an Aboriginal System of Knowledge*. University of Chicago Press, Chicago
Morphy H 1999 Traditional and modern visual art of hunting and gathering peoples. In: Lee R B, Daly R (eds.) *The Cambridge Encyclopedia of Hunters and Gatherers*. Cambridge University Press, Cambridge, UK, pp. 441–8
Powell T G E 1966 *Prehistoric Art*. Praeger, New York
Price S 1989 *Primitive Art in Civilized Places*. University of Chicago Press, Chicago
Turpin S (ed.) 1994 *Shamanism and Rock Art in North America*. Rock Art Foundation, San Antonio, TX
Whitley D, Loendorf L 1994 *New Light on Old Art: Recent Advances in Hunter-gatherer Rock Art Research*. Monograph No. 36, UCLA Institute of Archaeology, Los Angeles, CA
Wolff J 1981 *The Social Production of Art*. St. Martin's Press, New York

M. W. Conkey

Prejudice in Society

Prejudice is defined as cognitive beliefs, affects, and discriminatory behaviors towards members of a group on account of their membership to this group. When race, other groups, sex, or sexual preferences, for instance, are the targets of prejudice, it will be labeled racism, ethnocentrism, sexism, or homophobia. Many authors add that prejudice relies on undue generalizations and is irrational. Such addition implies that one knows what the objective truth is. It also overlooks the sophisticated ways that prejudice can take.

1. *Classical Explanations*

Personality was one of the first explanations to be proposed. After World War II, researchers explained fascism, ethnocentrism, and antisemitism by socialization (Adorno et al. 1950). Because children were obliged to obey parental authority blindly and to repress their hostile ideas, they later had to find scapegoats as a source of catharsis and represented the world into simplified categories: the good and the bad ones. This 'right-winged' authoritarian personality later was expanded to include all dogmatisms (Rokeach 1960). Whether authoritarian or dogmatic, personality cannot explain rapid changes in prejudice. Moreover, it is unable to account for regional differences in prejudice and it is mute about why certain groups, but not others, are the targets of prejudice.

Other psychologists influenced by behaviorism and psychoanalysis proposed a general theory of aggression that focused on situations (Dollard et al. 1939). According to them, aggression was a necessary consequence of frustration, and frustration was the necessary and sufficient antecedent of aggression. Applied to prejudice, this theory predicted that socio-economic factors (e.g., the price of cotton) would be determining in eliciting aggressive discriminatory behaviors (e.g., lynching of Blacks in the US). Besides being unfalsifiable, this theory is not better than personality ones in order to predict which group will be selected as a scapegoat.

Lack of contact was a third explanation (Allport 1954). It played a crucial role when the US Supreme Court ruled segregation in schools illegal. Socio-psychological experts testified that prejudice was due to lack of contacts between the groups. These psychologists were not naive to the point of thinking that contact *per se* was sufficient to eliminate prejudice. They listed a series of conditions under which contact would be effective. Unfortunately, in everyday life, these conditions are sometimes difficult to meet, such as contact restricted to atypical members of the stigmatized group and collaborative institutional environment. That will be discussed under different headings later on, but lack of contact is still a predominant explanation of prejudice. Sherif (1966) was one of the opponents to the contact hypothesis. For him, life is a competition for scarce resources and mere contact will render competition even harder. In order to be successful, Sherif argued, reduction of tension depends on supra-ordinate collaboration, that is, a collaboration that needs the efforts of the two otherwise competing groups.

In reaction to Sherif's conceptualization, proponents of Social Identity Theory stressed the fact that resources need not be real to induce conflict (Tajfel 1981). Symbolic resources may suffice. For these theoreticians, mere categorization into groups will lead to ethnocentrism, also called in-group favoritism bias, if the latter brings out a positive social identity to the discriminators.

2. *Modern Forms of Prejudice*

Prejudice or at least its expression has decreased over the years. When the measure of stereotypes, that is, measure of general beliefs about groups' behaviors and personality, was discovered in the early 1930s, it was a good thing for white Americans to express negative stereotypes about African-Americans. Whites gave more negative stereotypes about African-Americans when responding publicly than when answering anonymously. Now, the targets of open prejudice or stereotypes become rarer although they still exist (e.g., integrists). Such a situation does not mean that prejudice has disappeared. According to the authors, it is either the form or the conditions of expression that have changed. Here, we review briefly the different conceptualizations (see Dovidio and Gaertner 1986).

The most pessimistic view is probably the one of symbolic racism. For its proponents, prejudice has

remained the same but its expression has become subtler because of the pressure from the society. The most optimistic view is represented by regressive racism, which suggests that prejudice has decreased but that it comes again to the surface in stressful situations. Between these two extremes, aversive racism defends the idea that people persuade themselves that they are not prejudiced and try to rationalize what could be considered prejudiced in their behavior. Under the label of ambivalent amplification racism, other theoreticians hypothesize that people have ambivalent feelings about others. According to the circumstances, these ambivalent feelings are amplified either towards the absence of prejudice or towards its expression.

In fact, prejudice itself may take ambivalent forms. To the same extent that sociability and competence are the two dimensions organizing person perception, they may also guide the perception of groups. Groups may be perceived as likeable but incompetent (e.g., housewives) or they may be considered competent but unlikable (e.g., feminists). Competence would be triggered by the status of the group in the society (e.g., feminists are perceived as intellectuals). Likability, on the other hand, would result from the outcome dependence between the dominant and the dominated groups in society (e.g., the males 'need' housewives). Because of these differences in the perception of groups, people will try to explain them. The most radical explanation is to attribute them a different (genetic, cultural, linguistic, religious) essence. Because group members are all and foremost concerned with their in-group, they will attribute to it the human essence. It follows that they will attribute an infrahuman essence to 'others.' This infrahuman essence is characterized by a lack of typically human attributes. According to the circumstances, these deficient attributes can be the intelligence, the language, some typically human emotions such as love or sorrow, etc. (Leyens et al. 2000).

Finally, to the same extent that lack of contact was thought to reduce prejudice, many authors have argued that lack of cognitive resources may facilitate the expression of prejudice. All people know the negative stereotypes held by different groups and many do not share them. However, when they are overloaded by a demanding environment and in need of cognitive capacities, they may let prejudice leak. Such perspective makes a clear distinction between beliefs or stereotypes, on the one hand, and prejudice on the other (Devine 1989). It is obvious that, in normal circumstances, the attribution of an infrahuman essence is only possible under lack of cognitive control. In war times, on the contrary, this attribution often constitutes a conscious strategy.

Besides their differences, these different conceptualizations share one common point. Increasingly, prejudice is considered less a reaction against some other groups of individuals but more a measure of protection of the in-group. In general, people are less concerned by the other groups than by the well being of their own group.

3. *Implicit and Explicit Measures of Prejudice*

Prejudice typically was measured by questionnaires. To give but one example, the 'social distance scale' proposed different items of the form: 'I would accept a member of X (the group concerned) in the country, in the street, at work, as a friend, as a member of my family, etc.' A response at one level (e.g., at work) implied acceptance of the lower levels (e.g., country and street). As the expression of prejudice changed, its measures had to be adapted. From being blatant, questionnaire items have become more subtle but it is still easy to guess their purpose and, therefore, to control one's answers. These questionnaire measures are called explicit.

Some behaviors are obviously a more valuable source of information than questionnaires. Prosocial behaviors in the street or hiring rates in high-status positions on the job market may be taken as valid indicators of nonprejudice.

Laboratory implicit measures have also proliferated. Some of these measures have the advantage that the participants to the studies are not at all aware of what is measured. For instance, participants have to decide whether the string of letters they see on the screen of a computer is a word or a nonword. Among the words there are some positive or negative words. Before the presentation of some of the words, a social category is primed subliminally. The typical finding of such procedure is that the reaction times will be shorter for positive words when the social category is liked and for negative words when the social category is disliked.

The problem still to be solved concerns the correlations between these different measures. The questionnaire measures usually correlate highly between them. They rarely correlate, however, with behaviors and implicit measures. Also, different implicit measures do not necessarily correlate between themselves, which means that they probably measure different aspects of prejudice.

4. *Automatic and Controlled Prejudice*

The pervasive expression of prejudice and the development of implicit measures have launched an important and provocative debate about automatic vs. controlled behavior. One is reminded here of the unconscious in psychoanalysis.

Not long ago, most social psychologists believed that people made errors in their social judgments and, therefore, showed prejudice because they were unmotivated or incompetent to attend correct infor-

mation. Intelligence (or information) and effort were thought to be able to counterbalance these 'errors.' At present, some researchers do not hesitate to claim that people's first reactions are automatic and that neither abilities nor motivation can help. Automaticity does not imply inevitability of prejudice. It only means that people with prejudice will show it in their first reactions no matter how hard they try not to.

Automatic behavior will also depend on what is activated. A Black person is not only Black; she may also be tall, attractive, a physician, a woman, etc. The activation of the concept 'woman' may induce seduction. The activation of 'physician' may lead to help-seeking behavior. Among many persons, however, there is a limit to this flexibility. A lot of individuals will see a physician when this Black physician brings good news, and they will see a Black person when the physician is the bearer of bad news.

This link between a category and its attributes has implications. Active suppression of a category (e.g., skinhead) may, paradoxically, lead to a rebound effect. Stated otherwise, people who are instructed not to think of someone as a skinhead may, later on, perceive him as more prototypical of a skinhead. Indeed, by an ironic process, the monitoring of 'not-thinking-skinhead' activates attributes pertaining to the category. When the suppression is abandoned, all the activated attributes pop up; it is the rebound effect.

Up to now, we have reasoned as if activation of a category and of its attributes always led to an assimilation process (i.e., prejudice). It is true that people who are primed with the concept of professors do better at a Trivial Pursuit game than those who have been primed by the concept 'hooligan.' Priming specific exemplars of a category, like Einstein, may lead to a contrast effect. Indeed, everyone realizes that they are not Einstein and far from resembling him. This example leads us naturally to the attempts at reducing prejudice.

5. *Reduction of Prejudice*

One can distinguish between two approaches in the studies that looked at the reduction of prejudice. One approach focuses on information about individuals and the other concentrates on categorical representation of the world.

5.1 *The Focus on Individuals*

This approach relies mostly on the intellectual (and somewhat emotional) abilities to take into account evidence contradictory with one's prejudice. Three models have been proposed as possible ways to reduce prejudice:

5.1.1 The conversion model. According to this model, which could also be called the Sidney Poitier or Mandela model, the encounter with a very atypical member of a disliked category could provoke a conversion that would generalize to the whole category. Although anecdotal evidence is not lacking in favor of this model, no experimental demonstration has yet succeeded.

5.1.2 The book-keeping model. For the proponents of this model, people would integrate the different pieces of information that they receive from various members of the stigmatized group and that contradict their beliefs. Reduction of prejudice would be proportionate to the contradicting information. Empirical evidence in support of this model exists but it is not overwhelming.

5.1.3 The subtyping model. When confronted with contradicting information, people would try to subcategorize the exceptions in order to maintain their representation of the whole category intact. 'These women are unemotional; but they are not real women, they are businesswomen.' Needless to say that this strategy is often used.

5.2 *The Focus on Categories*

It is not surprising that this approach has been more elaborated than the previous one by psychologists because it goes back to the definition of prejudice, that is, reaction against individuals because of the category they belong to.

5.2.1 Decategorization. The early proponents of the contact hypothesis were members of the human relations movement and believed that solutions to conflicts would come from considering anybody as an individual rather than part of a given category. In spite of the disaster brought up by the compulsory desegregation of schools in the US, decategorization can be effective in eliminating the in-group favoritism bias.

5.2.2 Recategorization. It is often unrealistic to transform group members into mere individuals. Supporters of recategorization inspired themselves of Sherif's (1966) superordinate goal. This goal could only be achieved when all parties put their forces together and create some kind of new entity. Re-

categorization also eliminates the in-group favoritism bias, but it does it differently and more advantageously than decategorization. Whereas decategorization brings the attraction to the in-group down to the level of the out-group, recategorization carries the attraction of the out-group up to the level of the in-group. Recategorization transforms the 'we' and 'they' into 'us.'

5.2.3 Cross-categorization. It is said that Belgium never had bloody conflicts despite its history of linguistic problems because language was not the only divider in the country; religion and political orientations also played a role. Language cross-cuts religion and ideology. The problem with cross-categorization is that some people will find similarities with the other side, but that others will be completely different.

5.2.4 Harmonious differentiation. European researchers especially are aware that it may be neither feasible nor desirable to get away with categories. Some of them defend a blend of decategorization while at the same time proposing means to respect existing categories. This stream of research is still in need of convincing results.

6. *The Targets' Reactions to Prejudice*

Members of stigmatized group may suffer from prejudice and react to it in different ways (Crocker et al. 1998). When treated badly, for example, some may prefer to attribute the situation to discrimination rather than to their misconduct. Such a solution preserves their self-image. Others prefer to condemn their own behavior because this behavior can change whereas they will remain forever in their stigmatized group. The conditions that stir up the first or the second reaction are not yet well known.

Another solution is to find innovative ways to affirm a valuable identity. A typical example is the case of the African-Americans who invested in sport when the slogan 'Black is beautiful' lost its momentum. Simultaneously to this investment in a new domain, members of stigmatized groups may disengage themselves from goals that are denied to them. School or work may become unimportant for some of them. Indeed, very often, immigrants' children do poorly in school and their peer group is likely to disavow those who do well.

Besides the lack of social support, the atypical members of stigmatized groups suffer an additional burden in competitive situations like exams or job interviews. Not only do they feel the anxiety linked to the situation, but they have the additional anxiety to demonstrate that they are atypical of their group. It is, for instance, the case of women and mathematics, or of children from poor families and school achievement in general. In a way, it is a fulfilling prophecy.

See also: Adorno, Theodor W (1903–69); Allport, Gordon W (1897–1967); Anti-Semitism; Attributional Processes: Psychological; Authoritarian Personality: History of the Concept; Collective Beliefs: Sociological Explanation; Discrimination; Discrimination: Racial; Ethnic Identity, Psychology of; Ethnicity, Sociology of; Gender, Class, Race, and Ethnicity, Social Construction of; Identity: Social; National Socialism and Fascism; Racism, History of; Racism, Sociology of; Stereotypes, Social Psychology of

Bibliography

Adorno T W, Frenkel-Brunswick E, Levinson D J, Sanford R N 1950 *The Authoritarian Personality*. Harper and Row, New York

Allport G W 1954 *The Nature of Prejudice*. Doubleday Anchor, New York

Crocker J, Major B, Steele C 1998 Social stigma. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, Vol. 2, pp. 504–53

Devine P G 1989 Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* **56**: 5–18

Dollard J, Doob L, Miller N E, Mowrer O, Sears R 1939 *Frustration and Aggression*. Yale University Press, New Haven, CT

Dovidio J F, Gaertner S L 1986 *Prejudice, Discrimination, and Racism*. Academic Press, New York

Eberhardt J L, Fiske S T 1998 *Confronting Racism. The Problem and the Response*. Sage, Thousand Oaks, CA

Leyens J-P, Paladino M P, Rodriguez R T, Vaes J, Demoulin S, Rodriguez A P, Gaunt R 2000 The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and Social Psychology Review* **4**: 186–97

Leyens J-P, Yzerbyt V, Schadron G 1994 *Stereotypes and Social Cognition*. Sage, London

Rokeach M 1960 *The Open and the Closed Mind*. Basic Books, New York

Sherif M 1966 *In Common Predicament*. Houghton Mifflin, Boston

Tajfel H 1981 *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge University Press, Cambridge, UK

J.-P. Leyens

Pre-motor Cortex

The frontal lobe of primates is formed by two large regions, a rostral region involved in cognitive functions (prefrontal lobe) and a caudal one devoted to motor functions. Histologically, the caudal region is

characterized by the absence of granular cells. It is referred to also as the agranular frontal cortex (AFC). The AFC consists of two main sectors: area 4 (the primary motor cortex) and area 6. Recent data showed that area 6 is neither anatomically nor functionally unitary, but is formed by a mosaic of independent 'premotor' areas (see He et al. 1993, Rizzolatti et al. 1998). The aim of the present article is to review the general organization of the premotor areas in monkeys and to discuss in detail the operations of two of them. A comparison between the monkey and human premotor cortices will be made in the last section.

1. *The General Organization of the Premotor Area: Classes and Connectivity*

The existence of a cortex with motor functions lying rostral to the primary motor cortex (the 'premotor cortex') has been postulated since the beginning of the twentieth century. However, following the stimulation studies of Woolsey and Penfield, the prevalent view was that AFC functionally consists of two large motor areas: the 'primary motor cortex' located on the lateral brain convexity (MI) and the 'supplementary motor area' (SMA) located on its mesial surface (see Porter and Lemon 1993). Anatomical and functional evidence collected from 1980–2000 showed that this view is incorrect. The AFC is formed by seven independent areas (F1–F7). F1 corresponds to the Brodmann area 4 (the 'primary motor cortex'); the other areas are located within the Brodmann area 6 (the 'premotor areas'). This new subdivision of the AFC is shown in Fig. 1.

The premotor areas can be subdivided into two main classes: areas that receive their predominant cortical input from the parietal lobe (parieto-dependent premotor areas) and areas that receive their predominant cortical input from the prefrontal and the cingulate cortex (fronto-dependent premotor areas). Areas F2, F3, F4, and F5 belong to the first class, areas F6 and F7 to the second (Rizzolatti et al. 1998).

The organization of corticospinal projections is in accord with this subdivision (He et al. 1993). The parieto-dependent areas have direct access to the spinal cord, whereas the fronto-dependent areas send their output to the brain stem. Furthermore, the parieto-dependent areas have direct connections with F1, whereas the fronto-dependent areas lack this link.

Within the parieto-dependent group of premotor areas, each area is the target of a different set of parietal areas. Typically, each premotor area receives strong afferents from one parietal area. Conversely, each parietal area projects predominantly (but not exclusively) to a single premotor area. Thus, the parietal and premotor areas form a series of anatomical circuits largely independent one from another. The functional correlate of this anatomical organization is that each parieto-premotor circuit is dedicated to a specific sensorimotor transformation (see Rizzolatti et al. 1998). The result of this operation is the generation of a 'potential motor action,' that is of a representation of an action that may or may not be executed.

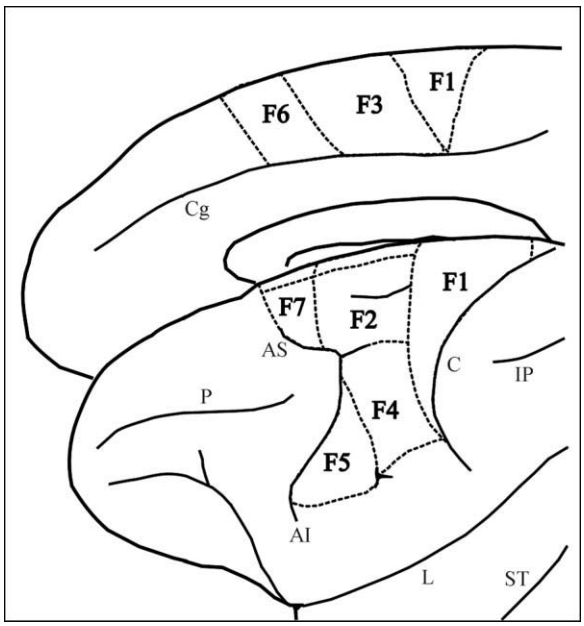

Figure 1
Frontal lobe of the monkey brain. The mesial and the lateral surfaces are shown in the upper and lower part of the figure, respectively. Dotted lines indicate the borders between the various areas forming the agranular frontal cortex (F1–F7). Area F1 roughly corresponds to area 4 of Brodmann Some authors refer to areas F2 and F7 as caudal and rostral dorsal premotor cortex, respectively. Areas F4 and F5 are also collectively referred to as ventral premotor cortex. Abbreviations: AI = inferior arcuate sulcus; AS = superior arcuate sulcus; C = central sulcus; Cg = cingulate sulcus; IP = intraparietal sulcus; L = lateral sulcus; P = principal sulcus; ST = superior temporal sulcus

The fronto-dependent premotor areas (F6 and F7) have only weak connections with the parietal lobe. F6 is mostly connected with the prefrontal area 46 and the rostral cingulate cortex (area 24c). F7 is richly connected with the dorsal sector of the prefrontal cortex. The functions of prefrontal and cingulate areas are much less known than those of the parietal cortex. There is consensus, however, that these areas are not involved in the analysis of sensory stimuli, but rather

play a role in cognitive functions such as working memory, temporal planning of actions, and motivation. It is clear, therefore, that the functions of the fronto-dependent premotor areas ought to be different from those of the parieto-dependent ones. The hypothesis presented here is that the fronto-dependent premotor areas relay information on the individual's motivations, long-term plans, and memory of past action from the prefronto/cingulate cortex to the parieto-dependent premotor areas. On the basis of this information, the potential motor actions, formed in the posterior premotor areas, are either executed or not.

2. *The Parieto-dependent Premotor Areas (Fig. 1: F2, F3, F4, and F5)*

As a model of information processing of the parieto-dependent areas, the functional properties of one of them (area F5) will be reviewed in some detail (see Rizzolatti et al. 1999). The properties of the other parieto-dependent areas will only be dealt with in short.

2.1 *Area F5*

Area F5 occupies the rostroventral part of the AFC (Fig.1). Its electrical stimulation evokes hand and mouth movements. F5 neurons typically discharge in association with goal-directed motor actions, such as object grasping, holding, tearing, and manipulating. The neuron firing correlates with specific actions (e.g., grasping) or action segments (e.g., hand opening, hand closure during grasping) rather than with the individual movements forming the action. Most of the F5 neurons code specific types of hand prehensions such as precision grip or whole-hand prehension.

A considerable part of F5 neurons respond to visual stimuli. Visually responsive F5 neurons fall into two classes: 'canonical neurons' and 'mirror neurons.' Canonical neurons discharge to the presentation of 3-D objects. Typically, their response is also present when there is no subsequent action on the presented object. Most canonical neurons are selective for one or a small set of visual objects similar in size and shape. Typically, the visual selectivity is congruent with the motor selectivity. A neuron that is selective for precision grip responds also to the presentation of small, but not of large objects. Mirror neurons do not discharge in response to object presentation but to action upon objects (see Sect. 2.2).

The parietal area that predominantly sends its input to area F5 is the anterior intraparietal area (AIP). Neurons of this area have been subdivided into three main classes: 'motor dominant,' 'visual dominant,' and 'visual and motor' neurons. 'Motor dominant' neurons are functionally similar to F5 motor neurons. Like F5 neurons, their discharge is associated with actions rather than individual movements. Often they are selective for object size, shape, and/or orientation. 'Visual dominant' neurons are active only when the task is performed in light, but not in dark. They respond also to object presentation in the absence of movements. Many of them are selective for size, shape, and/or orientation of 3-D objects. 'Visual and motor' neurons discharge in both light and dark, but less in the last condition. Like F5 canonical neurons, 'visual and motor' neurons show congruence between their visual and motor properties (Sakata et al. 1995).

From this summary of the AIP neuron characteristics, it is clear that the AIP and F5 have many functional properties in common. There are, however, also some important differences between the two areas. First, in F5 there are no 'visual dominant' neurons. Second, purely motor neurons are much more frequent in F5 than in the AIP. Third, most AIP neurons discharge during the whole action leading to the grasping of objects, being active often also during object holding. By contrast, most F5 neurons are active only during a specific phase of the grasping/holding action, not during the whole action.

On the basis of these findings and inactivation data, it has been proposed that areas AIP and F5 form the key elements of a circuit that transforms visual information on intrinsic (physical) properties of objects into grasping movements (Jeannerod et al. 1995). The following model has been proposed to explain how this transformation occurs.

AIP neurons code visual information on the object's intrinsic features. This information is then sent to specific sets of F5 neurons that transform the received information into potential hand movements appropriate to the size, shape, and orientation of the presented object. In turn, F5 neurons transmit this information to F1 (and the spinal cord) for movement execution. The movement occurs, however, only if there is a concomitant 'go' signal coming from the fronto-dependent premotor areas.

Objects can be grasped in different ways according to the use one wants to make of them. Because different AIP neurons are selective for different object features, F5 receives a multiple descriptions of each seen object from the AIP. The decision on how to grasp that object is determined by the motivations of the acting agent. This decision comes from the fronto/cingulate cortex and reaches the AIP via F5. Thus, a corollary discharge from F5 activates the AIP. In this view, the AIP 'visual and motor' neurons are memory neurons that maintain the representation of the selected object feature active during the entire execution of movement. Finally, the 'motor dominant' neurons would represent an intermediate stage in the transmission of F5 corollary discharge to the AIP 'visual-and-motor' neurons.

2.2 Mirror Neurons

F5 mirror neurons are neurons that discharge when the monkey makes a specific hand action and when it observes another individual making a similar action. The simple presentation of 3-D objects does not activate them. Similarly, the observation of actions that are not directed toward an object, or the observation of actions made using tools, fail to activate them (for a review, see Rizzolatti et al. 1999).

The visual and motor properties of mirror neurons are usually congruent. A neuron that discharges when the monkey grasps an object discharges also when the monkey observes another individual making the same action. In some neurons the way in which the object is precisely grasped does not matter (broad congruence). Other neurons fire only if the observed action is identical to that coded by that neuron (e.g., executed precision grip and observed precision grip; strict congruence).

There is evidence that mirror neurons are mostly located in the caudal part of F5. This sector is anatomically connected with parietal area PF, where, unlike in area AIP, there are neurons with mirror properties. It appears, therefore, that mirror neurons are part of a parieto-premotor circuit separate from that of canonical F5 neurons.

The discharge evoked by the observation of actions is most likely, as in the case of canonical neurons, not a pictorial description of the stimuli, but the motor representation of the action coded by the neuron. However, while canonical neurons use this representation to guide successive motor actions, this is not true for mirror neurons. This, obviously, raises the question of what this motor representation is for. Monkeys, unlike humans, are unable to imitate hand actions. Thus, the imitation hypothesis can be ruled out (at least for hand F5 neurons). The most accepted interpretation of mirror neuron activity is that it is used for action understanding. The assumptions underlying this hypothesis are: (a) individuals understand actions made by other individuals because they are able to react to them internally; and (b) the individuals know the outcome of their actions.

In conclusion, mirror neurons data indicate that the parieto-premotor circuits intervene in higher order cognitive functions and, very interestingly, that the same mechanisms that are at the basis of sensorimotor transformation underlie cognitive functions.

2.3 Areas F2, F3, and F4

Area F2 forms the caudal part of superior area 6 (Fig. 1). F2 receives a rich somatosensory input from area 5. Its ventro-rostral sector also receives visual information from the parietal areas MIP (medial intraparietal area) and V6A (see Rizzolatti et al. 1998). The dorsal part of F2 is related to hind limb movements, whereas the ventral part is related to forelimb movements. Single neuron studies of the arm field showed that some F2 neurons discharge during the preparation of reaching movements, others during movement execution. F2 neurons code the direction of reaching movement rather than a specific position in space (Weinrich et al. 1984). It is likely that F2 uses information on limb position in order to organize (and/or control) appropriate limb trajectories.

Area F3 (or SMA proper) is located on the mesial cortical surface. Its main input originates from somatosensory parietal areas (mostly mesial area 5). F3 contains a complete representation of body movements, with a prevalence of axial and proximal, arm and leg, movements. Electrical stimulation of F3 frequently evokes complex movements, involving both proximal and distal joints. Single joint movements are rarely evoked (Luppino et al. 1991). It has been proposed that F3 uses the knowledge of actual limb position to prepare the posture necessary for the incoming movements, thus playing a fundamental role in predictive postural adjustments. SMA also appears to contribute to the initial stages of the learning of motor sequences improving their performance (Hikosaka et al. 1999) and, according to others, in the execution of certain motor sequences (Tanji 1994).

Area F4 forms the caudal part of inferior area 6. F4 receives its main input from the ventral intraparietal area (VIP), a parietal area that contains visual and bimodal, visual and tactile, neurons. Electrical stimulation of F4 evokes mostly head and proximal arm movements. One of the most interesting characteristics of F4 is the presence of neurons with tactile and visual receptive fields. The tactile receptive fields are typically large and located on the face, arms, and chest. The visual receptive fields are limited in depth and are located in the space immediately around the animal. They are registered with the tactile fields. The visual receptive fields remain anchored to the tactile fields regardless of the eye and the body part position. These properties show that the F4 visual receptive fields are coded in egocentric coordinates. It has been suggested that F4 plays a crucial role in reaching movements, transforming spatial locations coded in egocentric coordinates into movement directed toward those locations (Graziano et al. 1994, Fogassi et al. 1996).

3. The Fronto-dependent Premotor Areas (F6, F7)

3.1 Area F7

Very little is known on the functional properties of area F7, except for its dorsal sector where eye movements are represented (supplementary eye field, SEF). The SEF is connected with the frontal eye field and other centers controlling eye movements. In contrast to the SEF, the F7 ventral part is linked to the premotor areas controlling body movements. In both

the SEF and ventral F7 many neurons respond to visual stimuli. The precise role of F7 in motor control is unknown.

Interestingly, joint ablation of F7 and (rostral) F2 determines a profound deficit in learning arbitrary stimulus-movements associations. This finding stresses once more that the premotor areas are involved in cognitive, as well as motor functions (Passingham 1993).

3.2 Area F6

Area F6 (or pre-SMA) corresponds to the rostral part of the SMA of Woolsey and Penfield. Motor responses can be evoked from it only with rather high current intensities. Typically, they consist of slow, complex arm movements (Luppino et al. 1991). Visual responses are common in F6, while somatosensory responses are rare. F6 neurons, recorded during active arm movements, show frequently a long leading activity preceding movements (Tanji 1994).

Studies carried out in a naturalistic condition where monkeys interacted with objects presented close to and far from them, showed that F6 neurons exert a global control on arm actions, rather than firing in association to specific, distal (like F5), or proximal (like F4) actions. Some discharge at visual presentation of graspable objects independent of their location, but increase their firing as the stimulus moves toward the monkey. Others, on the contrary, are inhibited at the presentation of the graspable objects, but start to discharge as soon as an object is brought close to the monkey. Summing up: there are several facilitation/inhibition combinations that characterize different F6 neurons. What appears to be constant is the modulation of the neuron discharge according to whether an object can or cannot be grasped (see Rizzolatti et al. 1998).

Recent data showed that F6 plays an important role in the early stages of motor learning, precisely when the cognitive aspect of the procedure to be learned is acquired. This conclusion is based on the behavior of single neurons and on inactivation experiments showing that the number of errors for learning new sequences increases after muscimol injections of this area. In contrast, the performance of learned sequences is not affected by pre-SMA inactivation (Hikosaka et al. 1999).

The data reviewed above suggest that F6 plays a double role in motor control. A first role consists in controlling the posterior premotor areas. Normally, even when the posterior areas are activated, movement does not start. Only when external contingencies and motivational factors allow it, F6 renders the movement onset possible. A second possible role of F6 is more cognitive. It consists in the control of visuomotor associations underlying the declarative phase of motor learning.

4. The Premotor Cortex in Humans

Data from brain imaging studies indicate that the organization pattern of the AFC in humans is similar to that of non-human primates. As in monkeys, there is a multiplicity of premotor areas in humans. Some of them become active during the execution of simple movements (Fink et al. 1997); others require more complex tasks.

The distinction between premotor areas that elaborate sensory information for action (parieto-dependent areas) and areas that control action execution (fronto-dependent areas) appears to hold true also for humans. For example, recent data (Binkofski et al. 1999) showed that in humans there is a specific circuit for hand-object interactions. As in the monkeys, this circuit includes a parietal area—the probable homologue of monkey area AIP—and a ventral premotor area, area 44, the probable homologue of monkey F5. Conversely, data on the pre-SMA (F6) indicate that this premotor area is involved in humans, as it is in the monkey, in complex actions, rather than in the execution of simple movements (Picard and Strick 1996, Hikosaka et al. 1999).

See also: Motor Control; Motor Cortex; Neural Representations of Intended Movement in Motor Cortex

Bibliography

Binkofski F, Buccino G, Posse S, Seitz R J, Rizzolatti G, Freund H J 1999 A fronto-parietal circuit for object manipulation in man: Evidence from an fMRI-study. *European Journal of Neuroscience* **11**: 3276–86

Fink G R, Frackowiak R S J, Pietrzyk U, Passingham R E 1997 Multiple nonprimary motor areas in the human cortex. *Journal of Neurophysiology* **77**: 2164–74

Fogassi L, Gallese V, Fadiga L, Luppino G, Matelli M, Rizzolatti G 1996 Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology* **76**: 141–57

Graziano M S A, Yap G S, Gross C G 1994 Coding of visual space by premotor neurons. *Science* **266**: 1054–7

He S Q, Dum R P, Strick P L 1993 Topographic organization of corticospinal projections from the frontal lobe: Motor areas on the lateral surface of the hemisphere. *Journal of Neuroscience* **13**: 952–80

Hikosaka O, Sakai K, Nakahara H, Lu X, Miyachi S, Nakamura K, Rand M K 1999 Neural mechanisms for learning of sequential procedures. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 553–72

Jeannerod M, Arbib M A, Rizzolatti G, Sakata H 1995 Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences* **18**: 314–20

Luppino G, Matelli M, Camarda R, Gallese V, Rizzolatti G 1991 Multiple representations of body movements in mesial area 6 and the adjacent cingulate cortex: an intracortical microstimulation study. *Journal of Comparative Neurology* **311**: 463–82

Passingham R E 1993 *The Frontal Lobe and Voluntary Action*. Oxford University Press, Oxford, UK
Picard N, Strick P L 1996 Motor areas of the medial wall: A review of their location and functional activation. *Cerebral Cortex* **6**: 342–53
Porter R, Lemon R 1993 *Corticospinal Function and Voluntary Movement*. Clarendon Press, Oxford, UK
Rizzolatti G, Fogassi L, Gallese V 1999 Cortical mechanisms subserving object grasping and action recognition: A new view on the cortical motor functions. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 539–52
Rizzolatti G, Luppino G, Matelli M 1998 The organization of the cortical motor system: New concepts. *Electroencephalography-and-clinical-neurophysiology* **106**: 283–96
Sakata H, Taira M, Murata A, Mine S 1995 Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. *Cerebral Cortex* **5**: 429–38
Tanji J 1994 The supplementary motor area in the cerebral cortex. *Neuroscience Research* **19**: 251–68
Weinrich M, Wise S P, Mauritz K M 1984 A neurophysiological study of the premotor cortex in the rhesus monkey. *Brain* **107**: 385–414

G. Rizzolatti and G. Luppino

Prenatal and Infant Development: Overview

Infancy extends from birth through 2 years of age. It has long been regarded as a special period of life in which numerous important transformations of structure take place, and the essential processes of information exchange with the world are laid down. Over this period, the infant is transformed from a condition of utter dependence on the caregiver for survival to a condition of relative independence or life-preserving agency—a transformation accompanied by marked physical, behavioral, and cognitive changes. While no characterization of this transition is entirely satisfactory, several elements set the infant apart from children of other ages. Despite disclaimers that what happens very early in life can be reversed or improved upon, most developmental psychologists celebrate the very considerable and enduring influence of early experience. This article paints a broad picture of infant development, its prenatal antecedents, and its relation to the behavior of older children and adults.

1. Prenatal Development

1.1 Physical and Cognitive Development

The 'prenatal period' extends from conception to birth. During the 'germinal stage' (fertilization–2 weeks) and 'embryonic stage' (2–8 weeks), the fertilized egg rapidly divides and attaches to the uterine wall, the major organs and systems of the body differentiate, and interconnections form between the developing components of the central and peripheral nervous systems. During the fetal stage (8 weeks–birth), the organism grows rapidly and the physical features that identify it as human become increasingly apparent.

During the prenatal period, the mechanisms that will facilitate the organism's adaptation postnatally become functionally mature. Infants born prematurely, for example, who readily learn responses that increase the amount of quiet sleep, reveal this anticipatory maturation. The characteristics of the mother's voice are also learned during the prenatal period. After birth, hungry infants orient and root in the direction of their mother's voice, thereby increasing their efficiency of breast-feeding irrespective of whether the mother is holding them on the left or right. Newborns will also learn to activate a tape-recording of their mother's voice—but not of another woman's voice—by sucking non-nutritively on a modified nipple. Moreover, 'prenatal learning' is highly specific: newborns will suck harder to listen to their mother reading a Dr. Seuss passage that she had read aloud daily during the last 6 weeks of gestation than to hear her reading a novel Dr. Seuss passage (DeCasper and Spence 1986). This specificity continues postnatally throughout the first half year. Similar learning prenatally familiarizes organisms with the linguistic environment into which they eventually will be born.

1.2 The Jacksonian Principle

The development and dissolution of prenatal motor function is described by the 'Jacksonian principle,' which states that the last reflexes to develop are the first to disappear when the organism undergoes demise, whereas the first to develop are the last to disappear (Jackson 1884, cited in Taylor 1958). The contralateral (avoiding) reflex, which moves a stimulated site away from a tactile stimulus, is supplanted weeks later by an ipsilateral reflex (Hooker 1952). This sequence is reversed by fetal anoxia. The reversion occurs because reflexes that develop first are well practiced and require less oxygen to be initiated than reflexes that are more recent. Similarly, the order in which sensory functions develop (tactile, vestibular, thermal, chemical, auditory, and visual) is subsequently reversed during injury, disease, or aging.

2. Mileposts of Infant Development

2.1 Zero–6 Months

Birth marks the organism's transition from a parasitic, waterborne existence in a thermally constant environment with constant access to nutrition to an in-

dependent, land-dwelling existence in a desiccating and cold environment with discontinuous access to food. The occasion of birth is unremarkable, important primarily for triggering the onset of the 'regulatory mechanisms' (respiration, blood flow, digestion) that are critical for this transition. These regulatory arrangements and their interactions are the major constraints on the newborn's behavior and the environmental conditions it can tolerate (Adolph 1968). The evolutionary requirement of all newborn mammals is to grow, which necessitates maximizing energy intake and minimizing energy expenditure. Not surprisingly, therefore, newborns primarily eat and sleep. Early postnatal learning also reflects this necessity: Newborns readily learn where and when food is accessible. Also, because full-term newborns do not physiologically thermoregulate before 4–9 weeks of age, they will not learn energetically costly responses that compete with behavioral thermoregulation.

Over the first 3 months of life, the birthweight of the infant doubles. During this period, brain development is very rapid and influenced appreciably by the kinds of experience to which the infant has been subjected. While most information about dendrite proliferation, the directionality of neuronal connections, cell death, and the effects of deprivation on neural and behavioral development comes from mammals other than humans, the universality of the effects described by neuroanatomists and neuropsychologists is impressive.

In the period following birth, much of the infant's behavior is reflexive, that is, specific stimuli call forth specific reactions (e.g., touches to the mouth elicit rooting and mouthing, abrasions of the skin elicit retraction of the injured body part). This early behavior is principally under the control of the subcortical portions of the brain, and the elicited behaviors are typically swift and obligatory.

Between 2 and 4 months of age, cortical maturation—especially dendritic proliferation—is increasingly evident, and neonatal reflexes disappear or weaken. Also during this period, activity undergoes a major transition from elicited to voluntary. The developmental sequence in which reflexes emerged during the prenatal period is recapitulated during this transitional period, and the sequence in which the postnatal reflexes appear is similarly reversed when the organism is challenged (Humphrey 1969). The postnatal repetition of fetal activity sequences is again due to the lower oxygen requirement of more primitive, hence more practiced, reflexes. Practicing neonatal reflexes reduces their oxygen requirement and speeds the transition.

The transitional process has been implicated in the sudden infant death syndrome (SIDS). McGraw (1943) characterized the period when reflexes are neither reliably elicited nor under voluntary control as a period of 'disorganized behavior.' The transitional period of the respiratory occlusion reflex corresponds to the time when SIDS is most likely to occur (95 percent of all crib deaths occur between 2–5 months). By 6 months, the occlusion reflex has come under cortically mediated voluntary control, and the infant is once again able to defend successfully against respiratory occlusion (Lipsitt 1979).

Also, over the first 3 months of age the behavioral repertoire of the infant sharply expands. Infants can acquire both classically conditioned and operantly conditioned responses at birth and can remember what they have learned for a surprisingly long time (Lipsitt 1969). An eyeblink reflex that is classically conditioned at 10 days of age, for example, is remembered for at least 10 days. Newborns can habituate to a repeatedly presented stimulus and discriminate it from others that they have not previously encountered. Imitation, a particularly efficient form of learning, can be demonstrated within hours of birth. Not only can newborns imitate an adult's facial gestures immediately after they were modeled, but by 6 weeks of age, infants can imitate facial gestures after a 24-hour delay. Piaget (1962) described early facial imitation as stimulus-bound and reflexive and hypothesized that it disappeared along with many other neonatal reflexes. However, it was later shown not to disappear (e.g., Meltzoff and Moore 1994). Infants' early imitation of facial expressions foreshadows their later dyadic interactions with adults and siblings.

By 3 months of age, infants can acquire arbitrary categories, learn the serial order of an arbitrary list, and pick up incidental information about events in their visual surround—including information about correlated features ('what goes with what'). Although some aspects of visual processing (e.g., accommodation and convergence, and feature search) are adult-like by 3–4 months, others (e.g., visual acuity and conjunction search) do not achieve adult levels for 1 year. Infants can detect depth by 2–3 months but do not fear depth until they begin to crawl. They can also recognize faces of live adults after 24 hours by 6 weeks of age and can recognize them from photos by 5 months.

Observers often remark that infants blossom into 'social creatures' between 2 and 4 months of age, cooing and smiling widely at individuals who interact with them, reaching to other persons, and intensely watching the behavior of others. By their fourth month, infants follow conversations between adults with their eyes, differentiate social from nonsocial objects (which they previously did not), and imitate vocalizations. Not surprisingly, social stimuli are highly effective in establishing and maintaining behavior at this time. The features that define a stimulus as social, however, change with age. Although newborns differentiate between facial expressions and prefer to look at faces than at other visual stimuli, this preference is not based on a social dimension but on contrast and movement—critical features of the human face which influence infants' scanning patterns over the

first 8 weeks. By 4 weeks of age, however, infants display greater pupillary dilation to social than non-social stimuli; by 6 weeks, they display differential cardiac responses to mothers and strangers; and by 12 weeks, they display differential pupillary responses to them and smile more frequently at the mother than at a stranger.

2.2 Six to 12 Months

During the second half-year, a normal infant becomes a walker, a speaker, and a social being who responds emotionally to parents and differentially to familiar persons and strangers. The simultaneous appearance of strong attachment to caregivers and wariness of unfamiliar individuals at 9–10 months of age is one of the most consistent benchmarks of this period (Ainsworth 1979). At the same time, infants begin to look at the target of a pointing finger rather than at the finger *per se* and begin to exhibit social referencing—a strategy by which infants assess the emotional valence of a new situation or person by checking how the caregiver has responded and responding in kind.

For a half-century, another developmental benchmark was attainment at 10 months of object permanence—the concept that an unseen object or person continues to exist. In the classic object search task, infants younger than 10 months consistently failed to retrieve an object they had just watched an adult hide. New paradigms using looking instead of reaching measures, however, have now revealed that even 3-month-olds can locate where an object was hidden.

The types of learning of which infants are capable and their understanding of the physical world rapidly expands over the last half-year. Infants can individuate objects, solve problems requiring causal reasoning, form learning sets ('learning how to learn'), comprehend some spoken speech, repeat a couple of syllables and perhaps produce a word, plan and execute a detour, and track objects through occlusion—a finding that has fueled heated debate over whether infants can count. Before infants begin to crawl (7–9 months), they have already learned what happens in what places. Once infants can locomote, they begin to form a cognitive map or representation of the spatial relationships between these places, which facilitates navigation from one place to another.

2.3 Twelve to 24 Months

Over this period, infants' linguistic, motoric, and cognitive skills continue to improve and broaden. Between 12 and 18 months of age, for example, infants can use landmarks to locate objects instead of using their own body as a reference point, and between 18 and 24 months, they can use geometric cues in the environment to reorient. Although 18 months was set as the minimum age for deferred imitation of actions on an object (Piaget 1962), the use of tasks particularly suited to younger infants has lowered this age to 6 months. Still, infants' ability to imitate televised demonstrations lags well behind their ability to imitate live demonstrations. The same live demonstration that 6-month-olds can imitate after 24 hours cannot be imitated from television before 18 months, although 15-month-olds can imitate some televised actions after 24 hours if the viewing conditions are more naturalistic. Generally speaking, the pervasiveness of reports that infants' cognitive and motoric performance is superior in natural settings calls into question most if not all minimum age limits that have been imposed by laboratory tests.

Within the last five years, research on the ontogeny of memory has revealed that retention increases linearly over the infancy period, and reminders protract retention even longer. At all ages, a single reactivation or reinstatement reminder, for example, doubles the life of a memory—a remarkable result given that memories of older infants are initially so long. The finding that a few reinstatements can maintain a memory acquired at 6 months through 2 years of age reveals that early experiences can be enduring and raises new questions about infantile amnesia (adults' inability to recollect experiences from early childhood). The two memory systems (declarative, non-declarative) found in older children and normal adults are also functional by 3 months (Rovee-Collier et al. 2001).

3. Risk Factors and Individual Differences

In recent years, advances in the successful delivery and survival of very small and premature babies, often born in jeopardy, have fueled new debate over the importance of early experience. Because the lower threshold for survival is very new, it is inevitable that time must pass before the numerous longitudinal studies, now under way, will yield critical information about the relation between specific prenatal deficits and neonatal hazards and particular developmental outcomes. Such investigations are unlikely to rule out the possibility that some affected infants will be indistinguishable from their age-matched peers in developmental competencies. It is a question of proportions and resiliency.

Although infants born 38 weeks after conception are, on average, less mature than those born at 40 weeks are, the level of physiological maturity can vary among infants born at 40 weeks by as much as 3–4 weeks. Sex differences are also present at birth, with girls being 1–2 weeks ahead of boys in neurological maturation. In normal infants, some of this variability disappears by the end of the first year, but many

individual differences persist, the full range and significance of which are not yet fully appreciated.

Early attempts to predict later intellectual performance from infant behavior focused on sensory and motor abilities and was unsuccessful. Modest-to-good correlations have now been found between measures of information processing beginning at 5–6 months (e.g., habituation rate, novelty preference, 1-week retention ratios) and scores on various intelligence tests at 3–5 years. The Fagan Test of Infant Intelligence (Fagan and Detterman 1992), which correlates later intelligence with infant novelty preference scores, is sensitive to early exposure to environmental toxins, substances of abuse, and nutritional deficits and supplements, as well as to maternal HIV infection, intraventricular hemorrhage, bronchopulmonary dysplasia, genetic anomalies, and various neurological abnormalities.

4. *Past Trends and Future Trajectory*

The origins of research on infant development can be traced to baby diaries published by Darwin and Preyer in the late nineteenth century. Over the first half of the twentieth century, detailed diaries of their own children's behavior stimulated Morgan, Stern, Guillaume, Valentine, and Piaget to propose theoretical accounts of early cognitive development (Wallace et al. 1994). At the same time, Pavlov's seminal studies spawned longitudinal studies of classical conditioning with infants in Russia and other Eastern-block countries. These early studies provided the first systematic data on infant learning and documented the sequential onset of sensory function from the age at which stimuli from different modalities became effective conditioned stimuli. During this period, Western researchers produced only scattered reports of infant conditioning and retention, emotional development, activity, and color vision.

In 1958, federal funding of a long-term prospective study of infant development ushered in the era of modern infancy research. Using sophisticated apparatus and innovative methods, researchers at 12 sites across the US began to systematically document the infant's basic sensory and learning abilities. These initial inquiries sparked subsequent generations of research on all aspects of perceptual, cognitive, and behavioral development across the entire span of infancy. By the end of the twentieth century, a large amount of developmental data was amassed on infant attention, perception, cognition, motor skills, and the sequellae of many early risk conditions.

The twenty-first century will witness intensive research that elaborates this knowledge and challenges other long-held beliefs about infant development. Also in this century, advances in cognitive and behavioral neuroscience should facilitate attempts to relate developmental changes in infant cognition to specific changes in the developing brain. This effort will require the development of new, noninvasive techniques and the adaptation of current ones (e.g., functional magnetic resonance imaging, positron emission topography) for use with very young infants. Finally, methodological advances should allow new insights into prenatal development and its contribution to postnatal development.

See also: Cognitive Development in Infancy: Neural Mechanisms; Prefrontal Cortex; Prefrontal Cortex Development and Development of Cognitive Function

Bibliography

Adolph E F 1968 *Origins of Physiological Regulations*. Academic Press, New York

Ainsworth M D S 1979 Attachment as related to mother-infant interaction. In: Rosenblatt J S, Hinde R A, Beer C, Busnel M (eds.) *Advances in the Study of Behavior*. Academic Press, New York, Vol. 9

DeCasper A J, Spence M J 1986 Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development* **9**: 133–50

Fagan J F III, Detterman D K 1992 The Fagan Test of Infant Intelligence: A technical summary. *Journal of Applied Developmental Psychology* **13**: 173–93

Hooker D 1952 *The Prenatal Origin of Behavior*. University of Kansas Press, Lawrence, KS

Humphrey T 1969 Postnatal repetition of human prenatal activity sequences with some suggestions of their neuroanatomical basis. In: Robinson R J (ed.) *Brain and Early Behaviour: development in the fetus and infant*. Academic Press, London

Lipsitt L P 1969 Learning capacities of the human infant. In: Robinson R J (ed.) *Brain and Early Behaviour: development in the fetus and infant*. Academic Press, London

Lipsitt L P 1979 Critical conditions in infancy: A psychological perspective. *American Psychologist* **34**: 973–80

McGraw M B 1943 *The Neuromuscular Maturation of the Human Infant*. Columbia University Press, New York

Meltzoff A N, Moore M K 1994 Imitation, memory, and the representation of persons. *Infant Behavior and Development* **17**: 83–99

Piaget J 1962 *Play, Dreams, and Imitation in Children*. Routledge and Kegan Paul, London

Rovee-Collier C, Barr R (in press) Infant cognition. In: Wixted J (ed.) *Stevens' Handbook of Experimental Psychology*. Wiley New York, Vol. 4

Rovee-Collier C K, Hayne H, Colombo M 2001 *The Development of Implicit and Explicit Memory*. Benjamins Publishing, Amsterdam

Smotherman W P, Robinson S R 1988 *Behavior of the Fetus*. Telford Press, Caldwell, NJ

Taylor J E (ed.) 1958 *Selected Writings of John Hughlings Jackson*. Hodder and Stoughton, London, Vol. 2

Wallace D B, Franklin M B, Keegan R T 1994 The observing eye: A century of baby diaries. *Human Development* **37**: 1–29

L. P. Lipsitt and C. Rovee-Collier

Copyright © 2001 Elsevier Science Ltd. 11997
All rights reserved.

 ISBN: 0-08-043076-7

Presidency: United States

1. Meaning

The term 'presidency' is applied to several distinct perspectives on the office of the President of the United States. Most commonly, and in its narrowest sense, it is employed as a label to refer to the incumbency of one particular President, as in the presidency of Abraham Lincoln or the presidency of Franklin Roosevelt. More broadly, the term applies to the historical development and legacy of a succession of presidencies since the office was established in 1789. In a more precise constitutional sense, the presidency is that branch of government established by Article II of the Constitution of the United States encompassing a set of constitutional and statutory powers and duties of the President together with additional 'implied' or 'inherent' powers and responsibilities not specifically enumerated in the Constitution or statute law. Finally, the term 'presidency' is used to denote the institutional structure of the office of President embracing the Executive Office of the President—a collection of staff units, originating from proposals contained in the Brownlow Report of 1937, that provide the President with the capacity to manage the work of the executive branch of government and to liaise with the principal constituents of the President, namely the Congress, the media, and interest groups.

These different usages of the term present no significant analytical difficulties. In the literature on the US presidency, the context is usually sufficient to clarify the sense in which the term is being employed and in none of its various forms is it conceptually or theoretically problematic. This is partly because the academic study of the presidency has not been encumbered by rigid theoretical or conceptual frameworks—a matter of concern to some presidency scholars who wish it were so, and a relief to others who see value in the broad, disparate, and eclectic approach that has characterized the study of the presidency to date.

2. Intellectual Context

The academic study of the US presidency has essentially been the preserve of historians and political scientists. Historians have generally confined their attention to individual presidents and presidencies whereas political scientists have traditionally concentrated on the constitutional and political power of the presidency and its institutional structure and development over time. Few historians, with some notable exceptions (see, e.g., Schlesinger 1973, McDonald 1994) have focused on the broader institutional development of the American presidency and only a handful of political scientists (see, e.g., Milkis 1993, Skowronek 1993) have borrowed from the discipline of history to further what political science has had to say about the presidency. Given the vast amount of scholarship on US presidents and the presidency produced by historians and political scientists, there has been little significant cross-fertilization between the two disciplines and it would not be unfair to say that the study of the US presidency, as distinguished from the study of US presidents, has been dominated by political scientists.

Until the publication in 1960 of Richard Neustadt's seminal work, *Presidential Power*, scholarship on the US presidency had primarily focused on the constitutional and statutory powers of the President, the various roles and functions of the President, and its relationship to the other branches of government as exemplified in the classic texts of Edward S. Corwin (1957), Clinton Rossiter (1956), and Wilfred Binkley (1962). There has also been a strong normative tradition in the literature the presidency, but this has produced various and divergent diagnoses of the weaknesses of the presidency and no uniform prescription for change, although the British system of cabinet government or variants of it has tended to be among the more popular reform proposals.

Neustadt's *Presidential Power* brought the study of the presidency into contact with the behavioral revolution in postwar political science. Neustadt defined power in terms of the president's persuasive or bargaining ability derived from his 'professional reputation' and 'public prestige.' Asserting that 'powers are no guarantee of power,' he proceeded to show that the leadership ability of presidents was not inherent in the office but was highly personal and dependent on the President's own capabilities as a seeker and wielder of effective influence upon others involved in governing the country. Neustadt's immediate impact was to shift the concerns of presidency scholars away from the formal, legal, and institutional boundaries of presidential power toward a more behavioral and theoretical effort to locate the reality of power and leadership in the White House.

While Neustadt's work was a major break with the tradition of presidency scholarship it did not represent a departure from the basic values of the leading presidency scholars who were firm in their belief that the presidency was and ought to be the center of action in American politics and who advocated a strong and dominant presidency as the only institution capable of responding to the social and political problems facing the country. They represented a generation reared during the presidency of Franklin D. Roosevelt and who made FDR, or what was often referred to as 'the New Deal presidency,' the normative model for their scholarship.

3. The Changing Focus of Presidency Scholarship

Notwithstanding Neustadt's impact, the political science of the American presidency lagged behind the behavioral revolution for much of the 1960s and 1970s

and not until the 1980s did the community of presidency scholars begin to exercise the kind of introspection and self-criticism on which behavioralism in America had thrived. That introspection produced a clear but belated appeal for the study of the presidency to connect with the basic tenets of the behavioral revolution—a more rigorous scientific methodology, the need for quantitative measures, and a greater concern for the conceptual and theoretical underpinnings of presidency scholarship (see, e.g., Edwards and Wayne 1983). That appeal did result in a redirection of the focus of presidency scholars and, since the mid-1980s, the political science of the US presidency has taken a quantum leap theoretically, methodologically, and conceptually.

The fundamental values of presidency scholars also changed around the same time. The combined effect of the Vietnam War and Watergate led to a significant reassessment of the New Deal model of presidential power and even those who continued to advocate strong presidential leadership became more sensitive to notions of accountability and restraint. By the end of the century, presidency scholarship had launched an assault on what Mark Peterson called the 'presidency-centered perspective' on American government (see Peterson 1990, also Jones 1994), a timely shift in a new era of divided government.

4. *Current Theory and Research*

The epistemological transformation of presidency scholarship from the mid-1980s could not properly be called a revolution because it did not succeed in displacing older, established approaches entirely. The 'public law' tradition, e.g., has flourished during the last two decades partly because of the practical relevance of issues such as the war power, the legislative veto, impoundments, presidential nominations, and impeachment. Scholars following in the footsteps of Corwin (see, e.g., Fisher 1995) have made significant contributions to the study of the presidency that have not been dependent on theoretical sophistication or quantitative methodology. Similarly, the 'institutional' school has also been reinvigorated by political developments in recent times. Watergate and the Iran-contra affair, for example, acted as a catalyst for substantial research on the role of the presidential staff and the structure and organization of the Executive Office of the President (see Burke 1992, Hart 1995, Walcott and Hult 1995) and, although much of this work is theoretically and conceptually grounded, it also builds on the descriptive and empirical approach that underlies the 'institutional' tradition in presidency scholarship.

The breadth of current research on the US presi dency would be difficult to categorize. In addition to those areas mentioned above, it encompasses the study of presidential psychology and character, leadership style, decision-making processes, popularity, professional reputation, bargaining and persuasion, media management, rhetoric, the impact of public opinion on presidents, the impact of presidents on public policy, and much more.

Current research also embraces a variety of approaches and techniques which, when combined with the range of subject-matter, makes presidency scholarship a broad, diverse, and pluralistic area of study open to the criticism that it lacks coherence in focus and methodology. But it would also be true to say that much current research on the presidency builds upon and extends the seminal research of the past and, while there may not be a single, coherent approach to the political science of the American presidency, there is continuity of research concerns co-existing with a willingness to utilize new methodological and theoretical developments.

Rational choice theorists, however, have staked a claim for the comprehensiveness of their approach and suggested that it may offer presidency scholarship the analytical and theoretical coherence it presently lacks. Applications of rational choice theory to the study of the presidency are still few in number but recent work by rational choice theorists, particularly in the field of presidential–congressional relations, has made a significant impact (see especially Krehbiel 1998, Cameron 2000) and laid the foundations for an extension of that approach to other hitherto less theoretically receptive aspects of the presidency

5. *Methodological and Theoretical Problems*

Although the cutting–edge of presidency research is in the hands of the rational choice theorists and their close companions, the new institutionalists, their approach has been contested by other presidency scholars who are less convinced of its general utility across the spectrum of presidency scholarship and who question the desirability of imposing a conceptual and methodological uniformity on future research. The debate is exacerbated by the tendency of the new institutionalists to distinguish between the 'institutional' and the 'personal' presidency, rejecting the latter in the interests of parsimony and theoretical clarity. While it is true that behavioral studies of the 'personal' aspects of the presidency, as exemplified by James David Barber's pioneering work on presidential personality (Barber 1972), revealed how difficult it is to theorize on the basis of a large number of complex variables, that has yet to be accepted as a sufficient argument for ignoring the complexity of the US presidency.

A more longstanding methodological problem for presidency scholars is the lack of access to their subject of study. Presidents of the USA generally do not

give interviews to political scientists, nor do they respond to questionnaires, nor, for the most part, can they be observed at work, neither do their memoirs or papers usually provide the kind of data that presidency scholars require. This is a principal reason why the study of the presidency has lagged behind other branches of the US government in terms of scientific rigor and why the gaps in the state of knowledge about the presidency are still quite substantial.

6. *Future Directions*

During the last 20 years, research on the presidency has been driven in large measure by the behavior of presidents themselves and the imperatives of the political science profession. In all probability, these two forces are likely to determine the direction of presidency research in the foreseeable future.

Presidency scholars have often been caught unprepared by the behavior of recent presidents and the events surrounding their presidencies. The Vietnam War was a catalyst for substantial research on the war power. The growth and expansion of the presidential staff system was barely studied before Watergate. The weakness of Jimmy Carter as President provided the impetus for Polsby's (1983) landmark study of the reformed presidential nomination process and its impact on the nature of the presidency. Ronald Reagan's public popularity prompted numerous studies of 'going public' strategies of presidential leadership (see especially Kernell 1986). As long as occupants of the White House continue to surprise scholars of the presidency they are likely to influence the range and nature of future research.

Similarly, the professional emphasis on theoretical sophistication and scientific rigor is likely to determine the priorities of future research on the presidency as it has done in the past. The danger is that such imperatives may overwhelm and marginalize those aspects of presidency scholarship where theory and quantitative methods are less applicable and less suitable to the object of study.

See also: Behavioralism: Political; Executive Branch, Government; Leadership: Political; Leadership, Psychology of; Presidential Government; Presidential Personality; Presidential Popularity

Bibliography

Barber J D 1972 *The Presidential Character: Predicting Performance in the White House*. Prentice-Hall, Englewood Cliffs, NJ

Binkley W E 1962 *President and Congress*. Vintage Books, New York

Burke J P 1992 *The Institutional Presidency*. Johns Hopkins University Press, Baltimore, MD

Cameron C M 2000 *Veto Bargaining: Presidents and the Politics of Negative Power*. Cambridge University Press, New York

Corwin E S 1957 *The President: Office and Powers* 1787–1957, 4th rev. edn. New York University Press, New York

Edwards G C, Wayne S J (eds.) 1983 *Studying the Presidency*. University of Tennesse Press, Knoxville, TN

Fisher L 1995 *Presidential War Power*. University Press of Kansas, Lawrence, KS

Hart J 1995 *The Presidential Branch: From Washington to Clinton*. Chatham House, Chatham, NJ

Jones C O 1994 *The Presidency in a Separated System*. Brookings Institution, Washington, DC

Kernell S 1986 *Going Public: New Strategies of Presidential Leadership*. CQ Press, Washington, DC

Krehbiel K 1998 *Pivotal Politics: A Theory of US Lawmaking*. Chicago University Press, Chicago

McDonald F 1994 *The American Presidency: An Intellectual History*. University Press of Kansas, Lawrence, KS

Milkis S M 1993 *The President and the Parties: The Transformation of the American Party System Since the New Deal*. Oxford University Press, New York

Neustadt R E 1960 *Presidential Power: The Politics of Leadership*. John Wiley, New York

Peterson M A 1990 *Legislating Together: The White House and Capitol Hill from Eisenhower to Reagan*. Harvard University Press, Cambridge, MA

Polsby N W 1983 *Consequences of Party Reform*. Oxford University Press, New York

Rossiter C L 1960 *The American Presidency*, 2nd edn. Harcourt, Brace & World, New York

Schlesinger A M 1973 *The Imperial Presidency*. Houghton Mifflin, Boston

Skowronek S 1993 *The Politics Presidents Make: Leadership from John Adams to George Bush*. Belknap Press, Cambridge, MA

Walcott C E, Hult K M 1995 *Governing the White House: From Hoover Through LBJ*. University Press of Kansas, Lawrence, KS

J. Hart

Presidential Government

Democratic governance takes two basic forms: parliamentarism and presidentialism; some people would add semipresidentialism, leaving aside the very unique Swiss institutional structure. Each of them presents a variety of forms that lead to question the dichotomy. There are a few defining characteristics from which many consequences can be derived. Each is identified with a paradigmatic case: the British Westminster parliamentarism, US presidentialism and the semipresidentialism of France. However, as Sartori (1994) notes, parliamentarism takes a variety of forms like the German Federal Republic—the Kanzler democracy—and the assembly parliamentarism of the Third and Fourth French Republics, and contemporary Italy.

Presidential systems are based on an executive with considerable powers in the constitution, and generally with full control of the composition of the cabinet and the administration, elected by the people (directly or by an electoral college elected for that purpose) for a fixed period of time, and are not dependent on a formal vote of confidence by the democratically elected representatives in a parliament. The president is not only the holder of executive power, but is also the symbolic head of state who cannot be dismissed between elections except in rare cases of impeachment.

Two features stand out in presidential systems:

(a) Both the president, who controls the executive and is elected by the people (or an electoral college elected by the people for that sole purpose), and an elected legislature (unicameral or bicameral) enjoy democratic legitimacy. It is a system of 'dual democratic legitimacy.'

(b) Both the president and congress are elected for a fixed term, the president's tenure in office is independent of the legislature, and the survival of the legislature is independent of the president. This leads to what we characterize as the 'rigidity' of the presidential system.

Most of the characteristics and problems of presidential systems flow from these two essential features. Some other nondefining features of presidentialism are often associated with it, such as term limits or no re-election, automatic succession by a vice-president, more or less extensive rule making, and emergency powers. One characteristic, so normal that it can be included in the definition, is that the presidency is a unipersonal office. There have only been two cases of directly elected pluripersonal 'presidencies': the two-person Cypriot administration (1960–1963) and the Uruguayan Colegiado (which governed twice —1918–1933 and 1952–1967).

The basic characteristic of the president is the full claim to democratic legitimacy. The claim very often has strong plebiscitary components, although sometimes it is based on fewer popular votes than are received by many prime ministers in parliamentary systems heading minority Cabinets. The presidency gives the incumbent, who combines the qualities of head of state representing the nation and the powers of the executive, a very different aura and self-image, and creates very different popular expectations than those redounding to a prime minister with whatever popularity he might enjoy after his party receiving the same number of votes.

The most striking fact is that in a presidential system, the legislators, particularly when they represent well-organized, disciplined parties that constitute real, ideological, and political choices for the voters, also enjoy a democratic legitimacy, and it is possible that the majority of a legislature might represent a different political choice from that of the voters supporting a president. Who, on the basis of democratic principles, is better legitimated to speak the name of the people: the president, or the congressional majority that opposes his policies? Since both derive their powers from the vote of the people, a conflict is always latent and sometimes erupts dramatically; there is no democratic principle to resolve it, and the mechanisms that might exist in the constitution are generally complex, highly technical, legalistic, and therefore of doubtful democratic legitimacy for the electorate. It is no accident that in some of those situations the military intervenes as 'poder moderador.'

It could be argued that such conflicts are normal in the USA and have not led to serious crisis. It would exceed the limits of this essay to explain the uniqueness of American political institutions and practices that have limited the impact of such conflicts, including the unique characteristics of the American political parties that lead many American political scientists to ask for a more responsible, disciplined ideological party system (Riggs 1998).

In a unipersonal presidency there is only one winner who gains full control of the executive. It has been argued that in a Westminster parliamentary system a party winning an absolute majority is also (or even more) a 'winner take all' while in a presidential system it is less likely that a president will have an absolute majority in congress for the entire period of his mandate. However, it is a president, one person that wins control of the executive rather than a party, and that person gains that power irrespective of the size of his plurality or in a runoff in a narrowed competition in which necessarily one candidate will have a majority. The loser in a presidential election ends with no public office in contrast to the leader of a defeated party in a parliamentary election likely to hold a seat in Parliament. Part of the unipersonal election is, in many presidential systems, the succession by a vice-president chosen by the nominee using criteria that would not have been those of the electorate or a party leadership.

The plurality rule remains the most common method for electing presidents. However, the fear of a president being elected by a small plurality has led to introduce a majority runoff election among the two leading candidates (when no candidate has received a majority). The plurality rule (Shugart and Carey 1992) favors a two-party competition or at least the formation of broad pre-election coalitions and discourages candidates from minor parties. The runoff election, on the other hand, encourages a larger number of candidates in the first run, either in the hope to make it to the runoff (even with a modest number of votes) or place themselves in the condition to bargain for their support by one or the other of the two front runners.

Each procedure can have negative consequences. The plurality rule in a fragmented party system can lead to the election with the support of a small minority. The runoff in addition to encouraging party

fragmentation in the first run, can create false coalitions (which will not hold after the election), or mortgages the candidate with allies which may exact a price or distort his image. It also generates in the winner the feeling that he represents the majority of the people, though in the first run he had weak support and his opponent might have been ahead.

In a multiethnic society the fear that a candidate from a populous state could be elected against those from other ethnic-religious communities, has led to a complex system of concurrent pluralities, to assure the election of a president with appeal across ethnic groups (as in Nigeria).

There are a few instances in which the popular vote, in the absence of a decisive majority, does not determine the winner. In Chile, before 1973, the Congress chose the president in the event that no candidate received a majority of popular votes among the top two contenders. Traditionally the one with the largest plurality was elected. In the USA, when in the nondeliberative electoral college no candidate obtains the vote of the majority of electors at the first count, the decision goes to the House of Representatives.

The fixed term of office of both presidents and congress introduces a particular rigidity. A failed president cannot be (normally) forced to relinquish power, but is obliged to continue governing, and he cannot (constitutionally) dismiss a congress committed to oppose his policies. A failed parliamentary prime minister can lose a vote of confidence, be replaced by another leader of his party or coalition, and if he feels that the opposition is wrong, call for an election hoping to discipline his party or coalition and/or revalidate his authority. The rigidity and the length of some presidential mandates leads to the frequent rule of no-reelection, or of no continuous reelections.

The dual legitimacy resulting from separate election of president and the legislature and their mutual independence can lead to 'divided government.' The congress and the presidency may be controlled by different parties (or factions of the same party) creating an impasse. How to govern?

Ackerman (2000) has developed three scenarios. The first emphasizes accommodation, based on bargaining, they may be called the Madisonian hope. Fiorina (1996) argues that American voters desire a divided government because they are uncertain about endorsing an ideological direction offered by either party. Mayhew (1991) argues that divided government has not impeded major legislation. The second involves constitutional breakdown, in which one or another power assaults the system and installs itself as a single lawmaker. A situation of this type is in Latin America, where traditionally the military as 'poder moderador' intervened temporarily, or more recently assuming complete power. Referring to my analysis (Linz 1994), Ackerman calls this breakdown the 'Linzian nightmare.' A third less dramatic alternative is backbiting, mutual recrimination, and partisan deadlock, which he calls 'crisis of governability.' To the use of constitutional or extraconstitutional decree powers, emergency legislation, patterns described in detail by Carey and Shugart (1998) and their collaborators in a chapter significantly entitled 'Calling Out the Tanks or Filling Out the Forms,' I would add a fourth alternative that explains the absence of recorded conflicts: the application of the 'rule of anticipated reactions.' The avoidance of conflict by not initiating policies is likely to lead to conflict even when desirable, necessary, and proposed by the president at the time of his election.

The disjunction between the image of the voters and the president himself as an initiator of major policies, reforms, and change, and the gridlock, mutual blaming, and/or of a frustrated president or congress confronted with *decretazos*, an avalanche of decrees as *faits accomplis*, reducing the legislature to impotence, does not contribute to political efficacy and the legitimacy of either or both institutions. The breakdown of democracy and the turn to unconstitutional rule or authoritarianism is only one possible outcome. 'Formal' democratic continuity, particularly in recent crises in Latin America (or the ex-USSR countries) is no proof that the system works.

The situation that Ackerman calls of full authority, when the presidency, the House and Senate are controlled by the same party and there is a favorable majority in the Supreme Court is rare and may last a short time. The result is a burst of legislation, lawmaking initiatives that will be strong in symbolic statement and difficult to change afterwards, with less concern about implementation. In a parliamentary system, where laws can more easily be changed with electoral defeat, governments are more interested in statutes framed in mid-sized concepts keyed to operational realities, to make them work.

Although the separation of executive and the legislative power, more accurately separate institutions sharing powers, implies that the bureaucracy (and the military) are the domain of the presidency, the fact is that they have two masters. Congress is perhaps increasingly adept to intervene in details of public administration and defense thanks to the prerogative of 'oversight,' of financing departments and programs, the long-standing freedom to investigate. A result is the dependence on both the presidency and the congress and indirectly the separation from both, and subordination to neither. A response to this situation has been 'politicized professionalism': the growing number of political appointees, the large number of in-and-outers, and the short tenure of those appointed who prefer to move to the private sector (and consequently the lack of accumulated experience and continuity in policies). The president turns to 'loyalists' rather than experienced civil servants who would be concerned about their relations with powerful congressmen or women.

In the ongoing debate over the success or failure of presidentialism it is important to distinguish:

(a) The basic theoretical issues derived from the defining characteristics of presidentialism (Linz 1994, Shugart and Carey 1992).

(b) The problems for presidentialism not derived from the basic characteristics; like no reelection, tensions with vice-presidents, mid-term congressional elections, etc.

(c) The problems with presidentialism and the actual breakdowns of presidential compared to parliamentary systems. A comparison plagued by a number of methodological problems; time period covered, regional comparisons, level of economic development (OECD vs. developing countries), and increasingly formal stability with 'defective' democracies (some actually authoritarian regimes) (Stepan and Skach 1994, Shugart and Carey 1992, Thibaut 1996).

(d) The absence of a possibility of a transition to parliamentarism and therefore the limited relevance of a theoretical comparison and the need to consider 'improvements,' modifications of presidentialism that correct some features contributing to some of its failures (Nohlen and Fernandez 1991).

(e) The need to keep the problems of semipresidentialism, offered as a more viable alternative to presidentialism, separate. In this case, the lack of consensus on which countries are semipresidential and the small number of cases (at least before the new postcommunist regimes), makes an analysis like (c) difficult.

The basic argument for the separation of powers is that it provides mutual checks, enhancing freedom, preventing the concentration of power enjoyed by a British prime minister. Mutual checks, the greater difficulty in enacting policies, the opportunity for all interests to be heard and taken into account, the slowing down of the process, and the opportunity for vetoes, generate weak executives and block many of their initiatives, and this is presumably what the voters want when they split their vote for congress and the president. Perhaps this is an advantage in a society that works well independently of public policies, in a wealthy market economy with no major need for redistributive policies, or with independent nonpolitical agencies, but is unlikely to fit societies facing serious problems of underdevelopment and social injustice, requiring more effective government action.

Among the positive aspects of presidentialism its advocates list:

The separation of powers of congress and the executive allows the 'representatives' to act as representatives of interests, particularly of those with a geographical basis, since they are not bound to support the policies of the executive and no party discipline is needed to support a government. In the absence of proportional representation, they enjoy considerable independence from the party leadership and are more likely to represent interests rather than broad policy or ideological alternatives. All politics become local. As Sartori (1994, p. 84) writes: three factors 'ideological unprincipledness, weak and undisciplined parties and locality centered politics, contribute to the working of US presidentialism.'

The separate election, particularly mid-term elections, and the limited influence of the president in many cases in the nomination or the making of lists of candidates (especially in federal systems) and in their electoral success, makes the representatives independent. Even their own partisans will owe them almost nothing and often doubt their usefulness in the elections. The selection by primaries has probably reinforced that tendency. This makes members of congress, compared to parliamentarians, better representatives of interests of their constitutents but less reliable in supporting a party (and presidential) policy program. In fact, that 'representative' function is seen (Shugart and Carey 1992) as one of the positive aspects of presidentialism.

One presumed advantage is the identifiability of presidents, compared to the incertitude of who will head the government in some parliamentary systems based on coalition-making among multiple parties after the election. However, in many parliamentary systems the voter knows well who will head the government if the party he votes for wins. Although presidential candidates identify themselves with certain policy goals, they are less bound to a program than a party and the prime minister it supports, or minor parties that make their coalition support dependent on certain demands (for example, national minority parties).

Shugart and Carey (1992, p. 44–45) consider the personal identifiability the basis for greater accountability. This might be true, except when reelection is barred and for the blaming game between the president and congress (particularly in the case of divided government).

Presidential elections allow the citizen to choose one individual to put his trust in—on account of real or perceived qualities and commitments without the mediation of complex institutions: a party, a legislature, or a plurality of actors like parties forming a coalition. That personalization simplifies the choice. It also permits outsiders, not identified with a party, who may not have held any office before, to appeal personally and directly to the voter: Collor, Fujimori, Tyminski, Perrot. That appeal requires access to the media, which in turn may involve access to funding—in some cases personal wealth—and considerable misunderstanding of what is necessary to govern. It may also appeal to 'being above parties,' perceived as divisive, somewhat like a monarch. The election of military leaders: Mannerheim, Hindenburg, De Gaulle, Eanes, even Eisenhower, reflects this desire. TV is particularly congruent with these patterns.

Latin American presidents greatly vary in terms of their constitutional powers, both reactive (veto power,

for example) and proactive (decree legislation) and in terms of their partisan support in congress. The possible combinations (Mainwaring and Shugart 1997, pp. 40–54) are too complex to summarize here.

The difficulties with presidentialism in Latin America have generated two opposite responses: attempts to curtail the powers of the president out of fear of the abuse of power and authoritarian predispositions and efforts to strengthen the presidency as an answer to the weakness of the presidency, particularly in relation to congress.

Many presidential systems have a vice-president, with limited functions except when he has to step in to succeed the president for reasons of death, resignation, impeachment, or incapacity. He generally is not elected independently but on a ticket with the presidential candidate, chosen by him, often for electoral or personal reasons without concern about his or her qualifications. The frequent destructive conflicts between presidents and vice-presidents have led to recent constitutions to not include that office.

1. *Semipresidentialism*

Semipresidentialism as a distinctive form of organizing democratic political life, gained importance as a constitutional model and political reality with the consolidation of the French Fifth Republic defined by Sartori (1994, p. 132):

(a) The head of state (President) is elected by popular vote—either directly or indirectly—for a fixed term of office.

(b) The head of state shares the executive power with a Prime Minister, thus entering a dual authority structure whose three defining criteria are (c) to (e).

(c) The President is independent from parliament, but is not entitled to govern alone or directly and therefore his will must be conveyed and processed via his government.

(d) Conversely, the Prime Minister and his Cabinet are president-independent in that they are parliament-dependent: they are subject to either parliamentary confidence or no-confidence (or both), and in either case need the support of a parliamentary majority.

(e) The dual authority structure of semipresidentialism allows for different balances and also for shifting prevalences of power within the executive, under the strict condition that the 'autonomy potential' of each component unit of the executive does subsist.

Shugart and Carey (1992) prefer to call this system 'premier presidentialism' and distinguish it from 'presidential parliamentary' (which others include under semipresidentialism) on the basis of the constitutional position of the president.

Since its inception in the constitution of the Weimar Republic (1918) and the Finnish Constitution, it has been the object of considerable dispute. Different terms were used to characterize it: bipolar executive, parliamentary presidential, and Duverger established the term semipresidentialism. The ambiguities surrounding it lead some authors to treat it as a variant of presidentialism, while others see it as a form of parliamentarism. This author, as Sartori, considers it a form *sui generis* distinct from both presidentialism and parliamentarism.

Bartolini (1984) has shown that semipresidentialism was introduced in countries that achieved independence and sought a symbol for the new nation (Ireland, Iceland, and earlier in Finland). In Portugal, it was an attempt to overcome the uncertainties of a transition to democracy.

The French semipresidentialism of the Fifth Republic after the introduction in 1962 of direct presidential election is a two-headed system. It makes possible the alternation—according to some authors—between presidential and parliamentary phases depending on the party or parties of the president having a majority in the chamber. Sartori, rightly, prefers to speak of 'oscillation,' since, in the case of split majorities—described by the French as 'cohabitation'—the president retains some specific powers, particularly in foreign affairs and is not reduced to the status of a president head of state in a parliamentary republic. Nor can a president cum government, that has to govern with and through another body, be assimilated to a 'pure' president that governs alone, that is the government. The oscillation assures considerable flexibility and until now 'cohabitation' has worked relatively well.

Elgin (1999, p. 14) lists countries that he considers semipresidential: 6 in Western Europe: France, Finland, Portugal, Austria, Iceland, and Ireland; 5 in Central and Eastern Europe: Poland, Bulgaria, Romania, Croatia, Macedonia, and Slovenia; 11 in the former USSR: Russia, Ukraine, Lithuania, Moldavia, and Belarus (of them in the Asian belt: Armenia, Azerbaijan, Kazakstan, Kyrgyzstan, and Uzbekistan; 5 in Asia and the Middle East: Lebanon, Maldives, Mongolia, South Korea, and Sri Lanka; 3 in America: Dominican Republic, Guyana, and Haiti; and 11 in Africa. A total of 42 countries, among which only 18 in 1999 were rated from 1 to 3 on the Freedom House index of political rights; 6 in Western Europe, 4 in Central and Eastern Europe, 1 (Lithuania) in the ex-USSR; 2 in Asia; 1 in the Americas; and 4 in Africa.

The conclusion about regimes that from a formal constitutional perspective and definition can be classified as semipresidential is that the material constitution (the actual working over time) varies enormously. In some, the president, although popularly elected and with considerable legal powers, acts more like a president in parliamentary regimes. One difficulty is that a large number of countries that have adopted semipresidential institutions are new democracies, and are politically unsettled (and doubtful democracies), and it is premature to characterize those

polities. Moreover, much depends on the personality and political goals of the founding presidents and their successors.

Though all conceptual classification of forms of government cover a wide range of variations, this is even truer about semipresidential given the ambiguities in the constitutions about the powers of presidents and prime ministers, the role of leadership and personality, and the fact that the extended list of such regimes includes many countries dubiously democratic.

On the basis of the long-time actual working of formally semipresidential regimes, Linz (1994) has excluded Austria, Ireland, and Iceland from the list and he and Sartori (1997) would also exclude Portugal after 1982.

The dual legitimacy resulting from the separate direct election of president and parliament persists in semipresidential regimes, but the principle that a prime minister requires the confidence of a majority in the parliament reduces—perhaps even eliminates—the conflicts derived from 'divided government.' The French constitutional practice in the situations of 'cohabitation' supports that interpretation, though one might still ask to what extent it reflects a long tradition of parliamentarism (in addition to other favorable conditions).

We might add to the complexity (and perhaps confusion) a distinction between formally semipresidential regimes in which the president exercises significant powers (and is expected to do so) and those where he has not done so (and is not expected to do so). We obviate the argument that the formal institutional structure is not irrelevant, and that in fact it might at some point become politically relevant. The formally but not materially semipresidential regimes on that account are not parliamentary though they may work as parliamentary. More complicated is the issue—not considered in the earlier studies of this regime type—of the president enjoying such formal and actual powers that the prime minister and his cabinet are fully dependent on the president and the parliament, though formally empowered to question that dominance, is unable to do so. Should such regimes be considered presidential rather than semipresidential? The question can be raised about Russia and Sri Lanka. Again, the complex constitutional rules cannot be ignored but so can the actual working of the system under Yeltsin not be fitted into a 'semipresidential mode.' With such a narrowing down we still would have a number of working stable pure semipresidential regimes.

Presidents generally are identified with a political party, a leader of a party, or after winning the nomination of a party become its leader (yet when defeated lose that position). However, in Russia, Yeltsin, though elected with the support of the parties, opted for not identifying with a party and the constitution formally forbids the president from being a party member. The same is true for other post-communist presidential or semipresidential constitutions.

The idea of president above party, not identified with a party and not campaigning for a party, is likely to make it more difficult for a president to have party support, and does not contribute to the building of strong parties.

Semipresidentialism can work well when the presidential majority and the majority in parliament are of the same party or coalition and when, without being the same, there is a majority in the chamber, supporting a prime minister of a different party (the French 'cohabitation'). The system becomes more problematic as Kalteefleiter (1970) has analyzed in detail when the parliament is unwilling or unable to provide support to the choice of prime minister or of the president but is also unable to support someone of its own choice. This can be the case of a highly fractioned parliament or of a polarized pluralism in which the extremes make it impossible for the center to govern. Parties might be ready to tolerate weak presidential cabinets (abdicating their responsibility) and the president governing by decree. The president, on the other hand, having the power of dissolution, may be tempted to call for elections to solve the impasse and construct a majority. Repeated elections in a climate of crisis might deepen the crisis and contribute to strengthen the antisystem disloyal or semiloyal oppositions. These patterns developed in Germany under the presidency of Hindenburg and his presidential cabinets, contributed to the breakdown of the Weimar Republic.

See also: Electoral Systems; Executive Branch, Government; Leadership: Political; Legitimacy: Political; Parliamentary Government; Parliaments, History of; Party Systems; Presidency: United States; Presidential Personality; Presidential Popularity

Bibliography

Ackerman B 2000 The new separation of powers. *Harvard Law Review* **113**(3): 634–729

Carey J M, Shugart M S (eds.) 1998 *Executive Decree Authority*. Cambridge University press, Cambridge, UK

Duverger M 1978 *Echec on Rev*. Albin Michel, Paris

Duverger M 1980 A new political system model: Semi-presidential government. *European Journal of Political Research* **8**(2): 165–87

Elgie R (ed.) 1999 *Semi-presidentialism in Europe*. Oxford University Press, Oxford, UK

Fiorina M 1996 *Divided Government*. Allyn and Bacon, Boston, MA

Kalteefleiter W 1970 *Die Funktiones des Staatsoberhauptes in der parlamentcnschen Demokratie*. Westdeutscher Verlag, Köln

Linz J, Stepan A (eds.) 1978 *The Breakdown of Democratic Regimes*. Johns Hopkins University Press, Baltimore

Mainwaring S 1990 Presidentialism in Latin America. *Latin American Research Review* **25**: 157–79

Mainwaring S 1993 Presidentialism, multipartism, and democracy: The difficult combination. *Comparative Political Studies* **26**(2): 198–228
Mainwaring S, Shugart M S (eds.) 1997 *Presidentialism and Democracy in Latin America*. Cambridge University Press, Cambridge, UK
Mayhew D R 1991 *Divided We Govern: Party Control, Lawmaking, and Investigations, 1946–1990*. Yale University Press, New Haven, CT
Sartori G 1994 *Comparative Constitutional Engineering: An Inquiry into Structures, Incentives, and Outcomes*. New York University Press, New York
Serraforo M D 1999 El poder y su sombra. los vice presidentes. Editorial de Belgrono, Buenos Aires
Shugart M S, Carey J M 1992 *Presidents and Assemblies: Constitutional Design and Electoral Dynamics*. Cambridge University Press, New York
Stepan A C, Skach C 1994 Presidentialism and parliamentarism in comparative perspective. In: Linz, Valenzuela, p. 36
Thibaut B 1996 *Prësidentialismus und Demokratie in Lateinamerika*. Leske and Budrich, Opladen, Germany

J. J. Linz

Presidential Personality

The structure of the American constitutional system and the realities of American politics combine to make it a virtual certainty that the personal qualities of the chief executive will have significant consequences for the nation and the world. The Constitution accords the president coequal status with the legislative and judicial branches of government. Barring impeachment by the House of Representatives and conviction by the Senate, he (and at some possible point she) is entitled to remain in office until the end of his term. Because the occupant of the Oval Office has power to veto legislation and is his nation's chief diplomatic officer, the commander and chief of its armed forces, and the administrator of its laws, he is a force to contend with even when his standing with the public and Congress is at low ebb.

One could imagine a nation in which the highest political leader had great formal powers but was of little personal consequence, because that individual was merely the agent of a powerful political party or some other extra-governmental force. That is emphatically not the case in the USA, where the political parties are highly decentralized and there are many centers of power. American presidents have placed their stamp on public policy since the founding of the Republic, but for much of American history the president tended to defer to Congress for the initiation of policy, the federal government played only a modest part in the lives of citizens, and the nation was inactive in global affairs. Then came the Great Depression, World War II, and the incisive response to these challenges of Franklin D. Roosevelt. During the three terms and three months of the Roosevelt presidency, the USA became a world power and a nascent welfare state, and the presidency itself underwent a metamorphosis. The president replaced Congress as the principal source of legislative proposals; presidents began to make an increasing amount of public policy by executive action; and the Executive Office of the President was created, providing the chief executive with the organizational support needed to carry out his obligations.

The qualities that distinguish one chief executive from another have their greatest potential impact in the realm of national security, where the president is custodian of a vast nuclear arsenal. No president other than Harry S. Truman has employed nuclear weapons, but there have been many presidentially-initiated conventional military commitments, the largest of which have been in Korea in the 1950s, Vietnam in the 1960s and 1970s, and the Persian Gulf in 1991. The president also is a major domestic force, not only because of his formal powers, but also because his great political visibility places him in a unique position to seek public support for his policies.

The two most influential efforts to identify the personal qualities that shape presidential performance are Neustadt's *Presidential Power* (1960), and Barber's *The Presidential Character* (1972). Neustadt's book, which has been described as the American equivalent of Machiavelli's *The Prince*, is an exercise in political psychology with the emphasis on the political. Taking it as his premise that the American political order places severe constraints on the chief executive, Neustadt argues that a president can best overcome the constraints on him by using his formal powers as levers for bargaining and persuasion and making it evident to the other members of the policy-making community that he has the support of the public and is a skilled, determined political operator. From this it follows that an effective president should be well grounded in the ways of American politics, sensitive to power relations, and self-assured about the merit of his policies and his ability to direct his nation.

Barber's book is a study of political psychology with the stress on the psychological. Barber identifies five broad determinants of a president's White House performance. Two relate to the political context—the contemporary *power situation* (e.g., the partisan balance in Congress) and *climate of expectations* (e.g., the state of public opinion). The other three bear on the president's inner characteristics—his *political style* (habitual way of carrying out his responsibilities), *world view* (political beliefs and assumptions), and *character* (the deeper layers of his psyche). In principle, this is an excellent framework within which to analyze presidential personality, in that it reminds the analyst that a president's actions will be a joint function of his personal predispositions and his political context

(Greenstein 1969). In practice, however, Barber's focus is virtually exclusively on character, which he classifies in terms of whether a president is active or passive in the conduct of his responsibilities and positive or negative in the feelings he invests in politics.

In the resulting four-part classification, the two passive types parallel David Riesman's distinction between other- and inner-directed personalities. (Reisman 1950). These are the *passive–positive* presidents, who take their clues from others, and the *passive–negative* presidents, whose motivations for public service are grounded dictates of their consciences. Neither type, Barber argues, is well suited for the activist demands of the modern presidency. Barber's two active types echo another important distinction. The *active–positive* presidents are marked by what Lasswell (1951) refers to as the democratic character structure, which is to say that they are free of emotional perturbations that impede their leadership. The *active–negative* presidents are of a piece with the class of politicians Lasswell identifies as emotionally troubled power seekers—those for whom politics is a means of venting inner insecurities. As Barber puts it, the active–negative presidents channel great energy into their political actions, 'but it is an energy distorted from within' (Barber 1972, p. 9). Examples of such distortions include Woodrow Wilson's stubborn refusal to make the modest compromises that would have permitted the ratification of the Versailles Treaty and Lyndon Johnson's persistent escalation of the war in Vietnam, despite mounting protest at home and lack of success in the field.

Barber performs a valuable service by directing attention to the deeper strata of presidential personality, but his assumptions about human motivation are overly simple. Indeed, the very dichotomies that determine Barber's typology fail to account for important complexities. Rather than being either positive or negative in their feelings about politics, presidents may be subject to mood swings, as was the case of Theodore Roosevelt and Lyndon Johnson. Rather than being either active or passive, they may be active in some spheres and passive in others. Thus, Ronald Reagan was a tireless public communicator but had a hands-off approach to the internal workings of his presidency.

Barber's taxonomy is an example of what A. L. George and J. L. George (1998) characterize as a 'file drawer' approach to personality analysis. As an alternative, they propose elucidating presidents and other political figures in all of their complexity, drawing on existing clinical typologies as sources of hypotheses about a subject's inner dynamics. Theirs is the strategy taken in the most persuasive presidential psychobiographies, including their own sharply etched portrait of Woodrow Wilson (A. L. George and J. L. George 1956), Glad's (1980) subtle account of Jimmy Carter, and Renshon's richly psychoanalytic interpretation of Clinton (Renshon 1996).

A related track can be found in Greenstein's *The Presidential Difference* (2000), in which each of the eleven presidents from FDR to Clinton is characterized on his own terms but also compared with the others on the basis of six criteria: *public communication ability*, *organizational capacity*, *political skill*, *policy vision*, *cognitive style*, and *emotional intelligence*. A number of instructive patterns emerge. Only three of those presidents prove to have been consistently outstanding public communicators, Roosevelt, John F. Kennedy, and Reagan. In each case the president's effectiveness was much enhanced by his rhetorical prowess. Organizational capacity proves to have been even less common than eloquence, but it is at least as important. The most dramatic example of what can result in its absence is the chaotically organized Kennedy administration's failed effort to land anti-Castro insurgents at Cuba's Bay of Pigs.

The presidency of Jimmy Carter illustrates the critical importance of political skill for presidential effectiveness. Carter displayed impressive political ability in his rise to the White House, but upon becoming president he resisted the standard reciprocities of Washington politics, with negative results for the success of his program. The presidency of Lyndon Baines Johnson offers a complementary illustration. Between November 1963 and the end of 1965, Johnson displayed prodigious political skills, presiding over a tidal wave of legislative enactments. But in 1965 Johnson also embarked on an open-ended military intervention in Vietnam, using his skill to minimize public debate and controversy. What Johnson lacked was a realistic policy vision, at least in international affairs. He failed to establish the precise goals of the intervention, its probable duration, and the likely troop requirements. By 1968 there were a half million American troops in Vietnam at which point Johnson halted further troop increases, entered into negotiations with the communists, and withdrew himself from contention for a second elected term.

The cognitive and emotional qualities of presidents also are central of their leadership. One way presidents differ in their cognitive styles is in their capacity for abstract thinking. Jimmy Carter, for example, had an impressive ability to master specifics, but was unable to embed them in a larger rationale. Richard Nixon, by way of contrast, had an outstanding capacity to envisage broad patterns and possibilities. Before taking office, Nixon established three major international policy goals: extricating the nation from the fighting in Vietnam, arriving at an accommodation with the Soviet Union, and establishing relations with the People's Republic of China. By the end of his first term he had accomplished them all. But during the same period Nixon initiated the program of domestic political espionage and sabotage which led to the destruction of his presidency.

Nixon was one of several modern presidents who had outstanding cognitive strengths, but lacked what

has come to be called emotional intelligence (Goleman 1995), a term that captures the ability to manage one's emotions constructively, rather than being under their domination. Others include Johnson, Carter, and Clinton. Nixon and Johnson presided over Vietnam and Watergate—two of the unhappy episodes in American history. The presidencies of Carter and Clinton were not marked by comparably dire events, but each of them suffered from the character flaws of the chief executive.

The American political system is said to be one of institutions rather than individuals. But it was not the institution of the presidency that authorized the use of the atom bomb against Japan or led the USA into Vietnam. It was Harry S. Truman and Lyndon B. Johnson. American history is replete with episodes in which presidents took actions that might not have been taken by another White House incumbent, some of them of momentous significance. It follows that the qualities that distinguish one president from another demand close and continuing attention.

See also: Clinical versus Actuarial Prediction; Executive Branch, Government; Leadership: Political; Leadership, Psychology of; Presidency: United States; Presidential Government; Presidential Popularity

Bibliography

Barber J D 1972 *The Presidential Character: Predicting Performance in the White House*. Prentice-Hall, Englewood Cliffs, NJ

George A L, George J L 1956 *Woodrow Wilson and Colonel House: A Personality Study*. John Day Co., New York

George A L, George J L 1998 *Presidential Personality and Performance*. Westview, Boulder, CO

Glad B 1980 *Jimmy Carter, In search of the Great White House*. W. W. Norton, New York

Goleman D 1995 *Emotional Intelligence*. Bantam Books, New York

Greenstein F I 1969 *Personality and Politics; Problems of Evidence, Inference, and Conceptualization*. Markham Pub. Co., Chicago

Greenstein F I 2000 *The Presidential Difference: Leadership Style from FDR to Clinton*. Martin Kessler Books, New York

Lasswell H D 1930 *Psychopathology and Politics*. The University of Chicago Press, Chicago

Lasswell H D 1951 *Democratic Character*, in *The Political Writings of Harold D. Lasswell*. Greenwood Press, Westport, CT

Neustadt R E 1960 *Presidential Power, The Politics of Leadership*. Wiley, New York

Riesman D 1950 *The Lonely Crowd; A Study of the Changing American Character*. Yale University Press, New Haven, CO

Renshon S 1996 *High Hoper: The Clinton Presidency and the Politics of Ambition*. New York University Press, New York

F. I. Greenstein

Presidential Popularity

A president's 'popularity' consists of the public's affection for him and endorsement of his policies. Popularity becomes significant to the extent that it can be used by a president to achieve his policy goals. For much of the twentieth century, commentators discussed the 'mandate' that a president has received in an electoral victory for him and his party in Congress. A more recent development has been the attempt to analyze and measure a president's popularity, defined as a resource enabling him to pursue goals that were not a product of that mandate. It is this latter point on which most recent scholarship has been focused following the publication of Richard Neustadt's *Presidential Power*, arguably the most influential book on studies of the presidency. Neustadt stated that a president's standing with the public, along with his professional reputation, was the base on which his influence and success would lie.

The need that White House officials have shown to organize and mobilize a president's popular standing stemmed from significant changes in the political environment during the second half of the twentieth century. Previously, the most important talent a president needed to gain support in the political system was to be able to bargain and negotiate with the independent political powers of Congress, the political parties, and the states. Without this collective mobilization of resources he did not have the ability to frame government policy. Important exceptions to this constitutional weakness of the office occurred during declared international wars such as World War I, when President Woodrow Wilson was able to direct policy without the normal deliberations and specific legislative powers granted by Congress, and during the extreme internal threat of the Civil War when President Lincoln asserted similar plenary authority.

The changes that led observers of the presidency to recognize the importance of his popularity as a major political resource occurred because of the erosion of hierarchy and authority among those in the political system whose support he needed to achieve his goals. After 1960, the fragmentation of institutional power led presidents and their advisers to try to gain and maintain the allegiance of a majority among the politically involved portions of the public. The diminished ability of a president to make arrangements within the small sphere of those who continuously engaged as primary actors in the political system required that he seek public support for almost all-important domestic policies before they could be instituted.

Beginning with the Administration of Richard Nixon, the White House established and continued to develop institutions at the service of the president so that he could maintain continuous support for his conduct of the office, could gather public support for

his policy direction, and combat his adversaries in the political system and the media whose activities weakened the president's status with the public. As Nuestadt elucidated, presidential power was now dependent not only on his professional prestige among national political influentials but on his popularity and prestige among the public at large.

Changes in technology have also played a central role in the importance of presidential popularity as a major foundation on which a president depends. Until the third decade of the twentieth century, however, a president had no way of reaching a national audience whose attention he could command directly and simultaneously. This changed dramatically with the arrival of radio, film with sound in the ubiquitous newsreels in the third and fourth decades of the twentieth century, and the establishment of television as the primary universal channel of communication and news in the fifth and sixth decades, when a president could sell his policies directly to an audience made up of large segments of the public, an audience that could increase his influence with formal apparatus of power in the system.

An early prediction of the importance of presidential popularity may be found in the pre-presidential work of Woodrow Wilson. In series of lectures delivered in 1908 while still in the academic world, Wilson stated that the president 'is the one person about whom a definite national opinion is formed, and, therefore, the one person who can form opinion by his own direct influence and act upon the whole country at once' (Wilson 1961, p. 127.) Other major commentators on the presidency through the 1950s put little or no emphasis on the issue of popularity. Among the best-known authors of texts on the presidency during this era, Clinton Rossiter provided only passing mention of the consequences for a president of his standing with the public (Rossiter 1954). Rossiter did suggest that communication with the public is an important tool of presidential leadership. Edward Corwin, author of the most widely used text on the presidency, made no mention at all of popularity (Corwin 1989).

By the 1960s and 1970s, political scientists, journalists, and other observers recognized that there was an important relationship between what a president had been able to accomplish and his personal support among the public. Especially after the end of the Cold War in the late 1980s, a president could no longer use his prestige as defender against a recognized adversary to gain support for his efforts.

Studying these phenomena, however, proved both difficult and elusive. What exactly was a president's popularity? Did it stem from his personal skills in presenting himself, or from the benefits the public believed could be attributed to him? Did events drive his popularity up or down, or did his popularity enable him to undertake difficult and risky objectives, which, if successful, would make him loved and respected by the public? Furthermore, although you could show statistical correlations between the president's standing with the public as measured by polls, surveys, and particular events of his administration, causal links would be much harder to establish. Attempts to analyze the hows and whys of the level of popular support a president received during his term in office did not necessarily establish the consequences of this support.

The issue of presidential popularity now consists of a body of work divided here into two categories:

(a) The first encompasses research into how to define and measure presidential popularity, and how to explain changes in that popularity over time. Some variables that are commonly offered as potential explanations for changes in presidential standing with the public are presidential actions, the president's abilities as a communicator, news events, and the presentation of the news by the media as well as the general level of stability and prosperity in the nation.

(b) The second category includes work that addresses the question of how presidential popularity serves as a resource or constraint on a president in his dealings with other political actors. In particular, researchers in this area focus on a president's institutional and personal efforts to use his popularity to garner support from other institutional and political supporters in the system.

1. *Measuring Presidential Popularity*

Efforts by scholars to measure and obtain data on the public's perception and approval of a president have relied on polls, particular those published by the Gallup organization, which has provided results since the 1930s. From 1975 through 2000, Richard Brody provided the most prolific and important analysis summarizing public attitudes toward the president (Brody 1991; this book provides the best bibliographic source on works by Brody and others through 1991). In particular, Brody looked for correlations in the popular standing of the president, and public support for the president's ability to handle economic, social, and international situations. Following the course of a president's popularity during several administrations, he found that 'news events are dynamic ingredients in the processes of opinion change.' Further, 'evidence of policy success and failure contained in daily news reports in the mass media' are the basis on which the American people form and revise their collective evaluation of the president. (Brody 1991, p. 9)

Students of presidential popularity have paid a great deal of attention to the so-called 'rally phenomenon,' first put forward by John Mueller in 1970 (Mueller 1985). He suggested that poll data indicated that presidents received a significant boost in their popular rating from a public affected by patriotism when news events suggested that an international

crisis could involve the safety of the nation or its military forces. The widespread dissemination of this theory during the Cold War led to the political charge that a president might undertake military action deliberately to distract the public from other failings. Subsequent research, however, has indicated that a president receives such a boost in support only if there is no counterelite opposing his action. Since the end of the Cold War, many international events have been greeted with debate rather than support by the opposition and the rally phenomenon seems much more ephemeral.

A continuing problem of this research has been the difficulty of determining the proximate causes of fluctuations in a president's popularity. For example, although one can point out that a particular level of popularity was measured at the same time as certain events occurred, it is not clear whether there was support for a president's policy because he was popular or because the public endorsed the policy. Further, as suggested by Paul Gromke and Brian Newman, poll standing may reflect an 'aggregation of personal preferences' rather than the 'preferences of aggregates' (see Gronke and Newman 2000; this paper contains an excellent bibliography of recent works in the field.)

2. The Fabric of Presidential Popularity

The second line of study involves descriptive research emphasizing the action and/or statements of a president and his entourage as they attempt to bolster his standing with the public. Using interviews, observations, and reviewing documents, these researchers present a portrait of the importance of popularity as a factor determining a president's status in the political system. In particular, they have focused on the methods the White House employs to enhance a president's success in gaining approval for his administration. They also examine the activities of those in politics or working for news organizations who provide an alternative and frequently critical version of what is going on.

In *Going Public*, Samuel Kernell describes the process by which presidents have needed to gain public support in a political universe where persuading Congress or the bureaucracy involves approaching them with a powerful backing (Kernell 1997). In a study of the relationship between news organizations and the White House, Michael Grossman and Martha Kumar describe the institutionalization of White House operations designed to maximize the president's standing through the news media and the efforts by news organizations to present an independent and alternative viewpoint (Grossman and Kumar 1981). In *Congress and the Presidency*, Nelson Polsby (1976) describes the difficult problem a president has in converting his public support to a body that has its own connection with the public: 'It is hard for a President to change all this while explaining the intricacies of a particular issue over television. He can, however, communicate his concern about the issue … and, to a certain extent, set the terms in which it will be debated' (p. 187).

Researchers of this type also have a problem determining whether a president's standing is due to their efforts or other factors in the political environment. In spite of this lack of generally applicable tools of measurement, however, all researchers agree that what they have found has provided important information about how a president's standing affects his success.

See also: Campaigning: Political; Executive Branch, Government; Media, Uses of; Party Identification; Political Parties; Presidency: United States; Presidential Government; Presidential Personality; Public Opinion: Political Aspects

Bibliography

Brody R A 1991 *Assessing the President: The Media, Elíte Opinion and Public Support*. Stanford University Press, Stanford, CA

Corwin E S 1989 *The President: Office and Powers 1787–1984*. New York University Press, New York

Edwards G C 1989 *At the Margins: Presidential Leadership of Congress*. Yale University Press, New Haven, CT

Gronke P, Newman B 2000 FDR to Clinton. A 'state of the discipline' review of presidential approval. Paper presented at the American Political Science Association Annual Meeting, September

Grossman M B, Kumar M J 1981 *Portraying the President: The White House and the Media*. Johns Hopkins University Press, Baltimore, MD

Kernell S 1997 *Going Public: New Strategies of Presidential Leadership*, 3rd edn. Congressional Quarterly Press, Washington, DC

Mueller J 1985 *Wars, Presidents, and Public Opinion*. Lanham, MD

Polsby N 1976 *Congress and the Presidency*. Englewood Cliffs, NJ

Rossiter C 1954 *The American Presidency*. Harcourt Brace, New York

Wilson W 1961 *Constitutional Government in the United States*. Columbia University Press, New York

M. B. Grossman

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Pri

Primary Elections

One of the major functions that political parties perform in democratic political systems is the recruiting and selection of candidates for public office. In most democracies, this task is performed by a small number of official party members, meeting together as a caucus, committee, or convention. In the USA, by contrast, candidate selection is done almost entirely by party primaries: preliminary elections, held under governmental auspices, in which virtually every interested voter is allowed to participate, regardless of their past or current support for the party and its purposes.

1. History and Development

From the 1830s through the early 1900s, American political parties nominated their candidates for public office through a multitiered system not unlike those still used in most other democracies. The system began with gatherings of the party membership in small local units such as towns, wards, or precincts, which would select delegates to county or district conventions. These county and district conventions would nominate some candidates themselves and might also select delegates to participate in a state convention.

In the first decade and a half of the twentieth century, however, this system was almost entirely abandoned, as most states enacted legislation requiring the major political parties to nominate their candidates through primary elections. The rapid and widespread adoption of primaries was the confluence of two distinct trends within the American political system. The first was a movement to subject the parties to a greater measure of legal regulation. Though parties clearly played a major role in electoral politics, they were, in the beginning, regarded as entirely private entities, such that fraud, intimidation, and corruption in party conventions was seen as a purely internal matter. As Key (1958) put it, 'it was no more illegal to commit fraud in the party caucus … than it would be to do so in the election of officers of a drinking club.' Shortly after the Civil War, however, a number of states began passing laws aimed at outlawing bribery and fraud and preserving some measure of order and procedural integrity in party affairs. By the end of the nineteenth century, almost every state had enacted some legislation of this kind (for a detailed discussion of this development, see Merriam 1908).

Even as this trend was sweeping the field, however, a second demand was being made: that the parties needed to be 'democratized.' To quote Key again, 'through the history of American nomination practices runs a persistent attempt to permit popular participation in nominations and thereby to limit or to destroy the power of party oligarchies.' In the early 1900s, this viewpoint had a particular resonance with the Progressive movement, which saw popular government as being threatened by an alliance of party machines and powerful economic interests. The push for party 'democracy' was given further impetus by the increasing sectionalism of the American party system. By the late 1890s, many parts of the country had become so completely dominated by one party or the other that receiving the major party's nomination was tantamount to being elected.

Though a number of southern Democratic parties had already used primary elections on their own initiative, and other states had made them available on a voluntary basis, the first comprehensive, mandatory, statewide primary law was adopted by Wisconsin in 1903. Wisconsin's example was quickly emulated in a number of other western and midwestern states, and then spread to the rest of the country. By 1917, 'the direct primary had been adopted for most nominations in all save a handful of states' (Key 1958). A small number of states resisted the trend, but the tides of historical circumstance were clearly against them. The last holdout was Connecticut, which finally enacted a primary law in 1955. As of the early twenty-first century, primary elections are clearly the dominant way of making nominations throughout the USA.

It is perhaps worth adding that most state election laws require nomination by primary election only for so-called major parties (usually defined as parties that received some minimum percentage of the vote in the last election). Minor and third parties, such as the Libertarians or the Socialists, still generally use a caucus-convention process to nominate their candidates for public office.

Students of comparative politics have generally regarded primary elections as a uniquely American institution, strikingly different from the nomination systems employed in every other democracy (see, for example, Ranney 1981). While this still holds true as a generalization, there are some signs that other countries are beginning to experiment with primaries. In 1999, for example, the PRI in Mexico nominated its presidential candidate through an open, national primary. In these other countries, however, the de-

cision to use a primary is generally made by the party itself; only in the USA is the primary imposed upon the major parties by law.

2. Presidential Primaries

The trend toward primaries also partially transformed the presidential nomination process. Since the 1830s, both major parties in the USA had nominated their presidential candidates at national conventions, to which each state sent its own delegation. These state delegations, in turn, were selected at state party conventions or by state party committees. Beginning in 1912, however, a sizable number of states started using primary elections to choose their national convention delegates. Some Progressives (including President Woodrow Wilson) wanted to go further, scrap the national conventions entirely (at least as a device for making nominations), and choose the parties' presidential candidates through a national primary.

Yet, for a variety of reasons, primaries never became as thoroughly institutionalized in the presidential nomination process as they did in making state and local nominations. Proposals for a national primary never came close to being enacted; and while 25 states had passed presidential primary legislation by 1916, over the next 15 years eight of them repealed these statutes. The result, when all the dust had settled, was a distinctive system for conferring presidential nominations that modern scholars generally call the 'mixed system.' Under the mixed system, about one-third of the states selected their national convention delegations via presidential primary, while the remaining two-thirds used state conventions or state party committees. National conventions thus had a partly popular, partly organizational basis, and remained significant decision-making bodies, though the presidential nomination itself was increasingly decided in a single ballot (the best account of presidential primaries during this period is Davis 1967).

The mixed system came to an end in the early 1970s. In 1968, at a remarkably divisive and chaotic national convention, the Democrats voted to create a commission to re-examine their presidential selection process. In just four years, this Commission on Party Structure and Delegate Selection (often called the McGovern–Fraser Commission, after the two men who served as its chairmen) completely rewrote the party's rules governing delegate selection—so much so that most scholars see the years since 1968 as constituting a new era in presidential nominations, often referred to as the plebiscitary system. While the plebiscitary system differs from the old, mixed system in a variety of ways, one of the most important differences has been a huge increase in the number of presidential primaries. Where the 1968 election featured 17 Democratic presidential primaries and 16 Republican primaries, by 2000 there were 39 Democratic primaries and 43 Republican primaries. Under the mixed system, it was quite possible for someone to win a presidential nomination without contesting a single primary; under the new system, the nomination essentially goes to whichever candidate wins the most primaries (on these points, see Hagen and Mayer 2000). Though both parties still hold national conventions, they are increasingly empty affairs, put on for the benefit of television, in which all the important questions have already been decided.

3. Who Can Participate?

Since primaries are a product of state rather than national legislation, primaries vary across states in a host of major and minor ways. One of the most important differences, for both its practical and theoretical consequences, concerns who is allowed to vote in the primaries. At one extreme are the so-called closed primaries, in which every registered voter is asked if they are supporters of or affiliated with any particular party, and then only those who have expressed an affiliation with a party in advance are allowed to vote in that party's primary. (Even in this 'extreme' case, it should be noted, the level of commitment required from the voter is exceedingly light. Voters are only asked to affirm a kind of general, psychological tie to the party; they are not required, in any state, to show that they have performed any previous service for the party or to make a binding promise of future support.)

A somewhat more permissive regime, that might be called semiclosed primaries, prohibits registered Democrats from voting in Republican primaries and vice versa, but allows unaffiliated or independent voters to vote in either party's primary. A related provision, found in many state statutes, allows voters to change their party affiliation when they show up at the polling place on election day. More permissive still are what might be called declarative systems, where there is no party registration and any voter may vote in either party's primary simply by asking for that party's ballot.

Declarative systems at least ask the voters to state publicly that they wish to participate in the Democratic or Republican party primary. Open primaries, by contrast, relieve the voters of even this encumbrance. In open primaries, each voter is, in effect, given a ballot for both parties and then allowed to make his or her decision about which primary to participate in in the privacy of the voting booth. Since open primaries at least require a person to vote in only one party's primary on any given election day, a number of western states have recently sought to remove even this last imposition on the long-suffering voter. So-

called blanket primaries allow voters to participate in both party's primaries on the same day, casting a vote in, say, the Democratic primary for the US Senate, the Republican primary for the US House, the Democratic primary for state representative, and so on. (In June 2000, the US Supreme Court declared the California blanket primary law unconstitutional.)

States frequently change their primary legislation in large and small ways, and many of the systems that emerge are not easily placed in one of the categories listed above. But according to a recent analysis by Jewell and Morehouse (2001), 11 states had closed primary laws, 15 used semiclosed systems, 10 had declarative systems, 10 used open primaries, and 3 (including California) had opted for the blanket primary. As has already been suggested, none of these systems demand very much from the voter in the way of past or future support for the party, and none except the closed primaries even do much to prevent members of one party from helping select the other party's candidates.

Does it matter what type of participation standard a state chooses to adopt? Much of the concern about having open or declarative primaries originally focused on the possibility of party 'raiding': cases where large numbers of persons affiliated with one party would vote in the other party's primary, with the deliberate intention of nominating the weakest candidate. But there is little evidence that this sort of thing occurs: it is difficult enough to get most American voters to the polls at all, much less to convince them to carry out a complicated strategic maneuver. A more likely consequence of open primaries is that they make it more difficult for the parties to present a cohesive ideology and set of candidates. Independents and members of the opposition party tend to be less committed to a party's core symbols and principles, and thus more willing to support a candidate who is out of step with large segments of his or her own party.

4. *Who Does Participate?*

In numerical terms, voter turnout in primary elections is substantially greater than the turnout at local caucuses (where these are still held), but smaller than the turnout at general elections. The best venue for comparing turnout rates is presidential elections, for besides having reliable data on the number of votes cast in presidential primaries and in the November general election, there are, as noted earlier, still about 10–15 states that continue to select their national convention delegates through a caucus-convention procedure. In general, contested presidential primaries attract about 20–35 percent of a state's voting age Democrats or Republicans, compared to a 50–55 percent turnout in the general election and a participation rate of just 1–3 percent in the presidential caucuses. In addition, there is far greater variation in primary election turnout rates, depending on how many seriously contested races are on the ballot.

One major exception to these generalizations occurred in the South when that region was still a one-party enclave. Since winning the Democratic nomination virtually guaranteed a candidate's triumph in the general election, southern Democratic primaries regularly drew a far larger turnout than the general elections that followed.

It was once widely believed that the people who voted in a party primary were not very representative of the party's typical voter or identifier. More recent research, however, suggests that this fear has been exaggerated. While there are some differences between primary voters and general election voters (the former are, for example, somewhat older and better educated), the differences are relatively small and do not appear to have much effect on either the type of candidate who gets nominated or the ideological views that get expressed.

5. *Consequences*

The widespread use of primary elections as a nominating device has had several major consequences for American politics. First, it greatly weakened party organizations by taking away from them the single most important function they once performed. Prior to the Progressive Era, it was very difficult for a candidate to get nominated by claiming to be an 'outsider' or insurgent, running in avowed opposition to the party establishment. Since the adoption of the primary, however, such candidacies have become quite common. This is not to say, of course, that candidates endorsed or supported by the regular party organization invariably lose primaries; but primaries have greatly increased the odds that a non- or anti-organization candidate will be successful.

Primaries also greatly lengthen the public phase of American election campaigns, since they generally guarantee that, in one party or the other, there will be two distinct campaigns: one for the nomination and one for the general election. Primary campaigns are also more expensive than those required in a caucus-convention system, since they must be targeted at a much larger and less attentive audience.

Though the point is disputed, there is also a long-standing concern that primaries make it more difficult for the parties to present a united front during the general election. According to this argument, often referred to as the 'divisive primary' hypothesis, primaries require the parties to air all their grievances and divisions in public, rather than settling them behind closed doors. In some cases, indeed, candidates may go out their way to create or emphasize divisions, as a way of distinguishing themselves from the other contenders in a primary. Primaries, it is further argued, may be good at assessing the first choices of party

voters, but are not very good at discovering compromise candidates that are acceptable to a variety of factions, or at promoting 'balanced tickets' that offer some measure of reward or recognition to a diverse set of interests. Other scholars, however, claim that primaries do not so much create divisions as reveal the ones that already exist (for a review of these issues, see Mayer 1996). At the very least, one can say that sharp intraparty conflict was certainly not absent from the days when American parties selected their candidates through caucuses and conventions, and that the primary has not prevented most Republican and Democratic candidates from receiving a substantial majority of the votes cast by their fellow partisans.

See also: Electoral Systems; Key, Valdimer Orlando (1908–63); Participation: Political; Party Systems; Political Money and Party Finance; Political Parties; Political Parties, History of; Voting: Turnout

Bibliography

Davis J W 1967 *Presidential Primaries: Road to the White House*. Thomas Y. Crowell, New York

Hagen M G, Mayer W G 2000 The modern politics of presidential selection. In: Mayer W G (ed.) *In Pursuit of the White House 2000: How We Choose Our Presidential Nominees*. Chatham House, New York

Jewell M E, Morehouse S M 2001 Political parties and elections in American States, 4th edn. *Congressional Quarterly*, Washington, DC

Key V O 1958 Politics, *Parties, and Pressure Groups,* 4th edn. Thomas Y. Crowell, New York

Mayer W G 1996 *The Divided Democrats*. Westview Press, Boulder, CO

Merriam C E 1908 *Primary Elections*. University of Chicago Press, Chicago

Ranney A 1981 Candidate selection. In: Butler D, Penniman H R, Ranney A (eds.) *Democracy at the Polls*. American Enterprise Institute, Washington, DC

W. G. Mayer

Primary School Education

1. Introduction

The primary grades are the first few grades of elementary school, in which students aged approximately 5–8 are socialized into the student role and taught basic tool skills. Evolving views on what constitutes appropriate curriculum and instruction during these grades have been influenced by developments in theory and research on the needs, interests, capabilities, and other characteristics of children in this age range. Emphasis has shifted from maturation of generic stages of cognitive development to learning of domain-specific knowledge networks, so that readiness to learn is now seen as dependent more on relevant prior knowledge and experience than on age or developmental level.

2. From Generic Developmental Stages to Domain-specific Learning

The term 'primary school education' refers typically to the first two or three grades of elementary school, in which curriculum and instruction emphasize basic literacy and numeracy skills. These grades also emphasize socialization into the student role: teaching children to pay attention to lessons and to work persistently on assignments; to follow prescribed conventions for participating in whole-class activities (e.g., raise hand and wait to be called on); to collaborate productively with peers in pairs and small groups; and to take good care of personal possessions and classroom equipment. In many countries, formal schooling in the primary grades is preceded by a transitional year (called kindergarten in the USA). It is a year in which students are oriented to school routines, engaged in activities that feature artwork and other multisensory representations that call for listening, speaking, counting, building vocabulary, making visual discriminations, exercising fine motor skills, or in other ways developing readiness for the more formal academic learning emphasized in subsequent grades.

At all levels of schooling, curriculum development involves negotiating a reasonable balance among four general approaches: (a) equipping students with knowledge that is lasting, important, and fundamental to the human experience (emphasizing but not limited to the academic disciplines), (b) following the natural course of child development by matching content to students' interests and learning needs at each grade level, (c) meeting society's needs by preparing students to assume adult work roles and citizenship responsibilities, and (d) combating social injustice and promoting social change by addressing social policy issues (Kliebard 1994). The principle of meeting children's developmental needs and interests has special relevance for the primary grades because substantial cognitive development occurs during these years. Children aged five–eight (especially five–six) are often considered to be not merely less knowledgeable, but qualitatively different in several respects from older children and adults. Philosophers traditionally have maintained that children reach the *age of reason* at about age seven, when developments in their understanding of cause-and-effect relationships, reciprocity and the Golden Rule, and other key concepts and principles enable them to distinguish right from wrong and engage in principled moral reasoning. Researchers studying children's cognitive development speak of a

developmental shift that occurs between ages five and seven as children become less like animals and infants and more like older children and adults in their performance on a broad range of perceptual and cognitive tasks (Sameroff and Haith 1996).

Jean Piaget (1983) described children aged two–seven as egocentric in purview and unstable in their cognitive structures and thought patterns. Within his comprehensive stage theory of cognitive development, such children were labeled as preoperational because they had not yet acquired reversibility of thinking, conservation, classification, seriation, and other *concrete operations* that stabilize older children's cognitive structures, and enable them to engage in systematic reasoning (see *Cognitive Development: Child Education*). Piaget's work implied that progression from the preoperational to the concrete operational stage was determined primarily by maturation and thus not open to significant acceleration through educational interventions. Consequently, primary schooling models influenced by his ideas tended to emphasize play, exploration, discovery learning, and sensorimotor activities over more formally academic and verbally mediated instruction. Ideas drawn from Arnold Gesell, Maria Montressori, and others who emphasized developmental stages or the presumed special needs and characteristics of young children similarly led to early childhood and primary-grade education models that emphasize short and varied activities, hands-on learning, concrete and integrated topical units rather than systematic treatment of disciplinary content, and attempts to be responsive to children's learning initiatives rather than to move them through preplanned curricula in lock-step fashion. Contemporary notions of developmentally appropriate practice continue this tradition (NAEYC 1991), as do the British Infant School model and the Italian Reggio Emilia model.

In recent years, the predominant theoretical emphasis in developmental psychology has shifted from the stage theory of Piaget to the sociocultural learning theory of Lev Vygotsky (1978), and research on children's cognition has shifted emphasis from general developmental stages to domain- and situation-specific learning. Findings indicate that many aspects of cultural learning, including much of what is taught in elementary school, can be acquired through instruction, in which teachers provide modeling and explanations for students and coach and 'scaffold' their learning efforts. Instead of emphasizing limitations in cognitive capacity, pending the unfolding of genetically programmed maturation of internal structures, contemporary thinking emphasizes connecting with and building on students' prior knowledge in the domain. Instruction focuses on the *zone of proximal development*: the range of knowledge and skills that students are not yet ready to learn on their own but can learn with help from teachers (Tharp and Gallimore 1988).

A related notion is that instead of being organized into hierarchical levels, knowledge is organized within networks structured around key ideas. This implies that, instead of having to begin with the lowest cognitive level (knowledge) and proceed systematically through successively higher levels (comprehension, application, analysis, synthesis, and evaluation), instruction can begin anywhere in the network and develop understandings from there. Initial possibilities will be limited by the extent of learners' prior knowledge in the domain, but in principle, learners can accomplish any domain-related learning goals if they experience whatever instruction and learning opportunities may be required (without having to wait for supposed maturation of cognitive structures to occur).

Studies of children's domain-specific knowledge have verified that children's knowledge and learning capacities in many areas are much more advanced than previously suspected, although children find certain concepts difficult to grasp and certain misconceptions are commonly developed (e.g., that the sun rotates around the earth instead of vice-versa). It appears that Piaget's notions about children's learning capacities were too pessimistic and that his stage concepts are more applicable to developments in children's mathematical and scientific thinking than to developments in their thinking about issues addressed in the humanities and social sciences. Children are able to understand historical and literary content that is encoded within narrative structures built around the goals and motives of central characters, as well as cultural comparisons or other social science content that focuses on food, clothing, shelter, transportation, or other cultural universals with which they have had personal experience (Downey and Levstik 1991).

Research supports the following generalizations about difficulty level of curriculum content: (a) difficulty increases with the number of items that must be addressed simultaneously (it is harder to compare or interpret three things than two things); (b) abstract content is more difficult than more concrete content; (c) fine distinctions are more difficult than gross ones; (d) learning from print alone is more difficult than learning from multiple media; (e) it is easier to develop skills in thinking about matters that students see as closely related to their own lives than about other matters; and (f) providing structure, cues, and props makes thinking easier (Fair 1977)

In summary, much of what Piaget claimed about stage-related limitations in children's cognitive capacities is incorrect or limited to certain knowledge domains. However, Piaget was correct in suggesting that young children are qualitatively different from older children and adults in certain respects, so that it is not possible (or at least, not cost effective) to teach them some of the more abstract and complex aspects of each school subject. Primary-grade children appear to be ready for certain forms of systematic instruction that go well beyond what usually are described as

'developmentally appropriate practices,' but not yet ready for discipline-based courses that emphasize abstract content organized conceptually.

3. Curriculum and Instruction in the Primary Grades

Children show the capacity for operational thinking at five or six, but usually do not become functionally operational (use operational thinking most of the time) until a year or two later. Thus, in effect, the preoperational period extends until second or third grade for most students. Preoperational students are ready to learn about familiar experiences and observable phenomena, especially through exploration and manipulation of concrete objects. They also learn basic literacy and mathematics skills and those aspects of content area knowledge that can be linked to their experience base. Instruction in basic skills should connect with knowledge about language, print, counting, and numbers that children develop through their preschool experiences. These tool skills should be taught in an integrated way that stresses explanations for meaningful understanding and opportunities for authentic applications (Adams 1990, Hiebert and Carpenter 1992, Hooper and DeFrain 1980). Execution of basic skills should gradually become automatized, but these skills should be learned as strategies, so that procedural knowledge (about what to do and how to do it) is developed in connection with related propositional knowledge (about the nature and purposes of the skill) and conditional knowledge (about when and why to use it).

Primary-grade students require demonstration and explanation of skills, guided practice with feedback, and frequent opportunities to develop mastery through application. Good demonstrations and visual aids provide concrete models for children to watch or imitate, so that they do not have to struggle to follow purely verbal instruction. Complex tasks are broken into subtasks that children can master in sequences that gradually connect until the ultimate task is learned. When they work on their own, they may need cues, reminders, or other learning *scaffolds* that will help them overcome their limited attention spans and difficulty in following spoken or written directions. Models of completed tasks can be placed in learning centers and worksheets can include features such as spacing and lines to keep things separate; boxes or arrows to indicate where responses should be placed or where to go next; and division of assignments into modular units that can be presented, reviewed, and discussed separately.

Teaching in subject areas other than basic skills should either emphasize familiar content or make the strange familiar by using concrete props or other media. Actual objects are best, but substitutes such as videos or photos are also useful. Because young children have limited experience and major gaps in their vocabularies, they may need such props to enable them to understand what the teacher is talking about. They are most likely to have difficulty when content is organized around the conceptual structures used in the academic disciplines rather than around their own emerging knowledge structures. Children tend to learn specifics first and build up abstract generalizations only gradually as they accumulate experience; they have trouble following the logical patterns (flowing from general rules to specific examples) that adults find helpful. They are predisposed to learn in various domains, and although they are inexperienced, they reason facilely with the knowledge they have (Bransford et al. 1999). However, it is important to focus instruction on familiar and observable events, or at least events that are easily assimilated into the students' existing schemes. For example, children have an intuitive sense of the past and an interest in people that can provide a basis for instruction in history and social studies. However, such instruction has to reach the operative level rooted in students' concrete experiences, e.g., by having them investigate their own family histories, explore the geography of the neighborhood, or examine photos and artifacts from another culture. If the instruction stays at a purely figurative level (associating tipis with Native Americans, sombreros with Mexicans, etc.), it will have little meaning or potential for inclusion in networks of knowledge structured around powerful ideas (and will engender stereotypes, rather than layered understandings).

There has been noteworthy cross-national variation in the age at which children begin school and in the scope of the primary-grade curriculum. As more has been learned about children's knowledge and learning capacities, the trend has been towards earlier school initiation and more demanding curricula. Kindergarten classes in the USA, for example, now feature less sensorimotor play and more verbally mediated instruction than in the past. Even so, primary-grade teachers tend to be nurturant individuals who are oriented more toward teaching children than teaching subject matter, and who are concerned about supporting their students' personal development and adjustment to school, along with meeting their academic learning needs. Subject matter specialization usually does not begin until at least the fourth grade, so primary-grade teachers tend to be generalists who stay with their home-room class all day and teach all subjects.

Some schools even use looping arrangements that keep intact classes together for two consecutive years with the same teacher, as part of a larger effort to provide emotionally safe and supportive classroom environments for young learners. The children often work in learning centers or sit together on a rug instead of always sitting at desks or in chairs. Teacher aides, adult volunteers, or older students often assist the teacher by working with individuals or small

groups. Assessment is mostly informal, without emphasis on tests or grades. Curricula often present an idealistic and optimistic view of the world, sanitized to eliminate disturbing or depressing content, in an effort to 'preserve childhood innocence' by delaying confrontation with threatening realities. In general, most primary education approaches focus on orienting young children to school and equipping them with basic knowledge and skills, but within a relatively informal and highly supportive learning community. This ethos has persisted through developments from penpals to KIDSNET; from mathematics flash cards to number lines and manipulables; from penmanship drills to authentic writing and publishing; from fixed desks to flexibly arrangeable children's chairs and tables, etc., apparently because it is well-suited to the educational needs and capacities of young students.

See also: Cognitive Development: Child Education; Cognitive Development in Childhood and Adolescence; Education (Primary and Secondary Schools) and Gender; Piaget's Theory of Child Development; Schooling: Impact on Cognitive and Motivational Development; Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Adams M 1990 *Beginning to Read: Thinking and Learning about Print*. MIT Press, Cambridge, MA

Bransford J, Brown A, Cocking R (eds.) 1999 *How People Learn: Brain, Mind, Experience, and School*. National Academy Press, Washington, DC

Downey M, Levstik L 1991 Teaching and learning history. In: Shaver J (ed.) *Handbook of Research on Social Studies Teaching and Learning*. Macmillan, New York, pp. 400–10

Fair J 1977 Skills in thinking. In: Kurfman D (ed.) *Forty-seventh Yearbook: Developing Decision-making Skills*. National Council for the Social Studies, Arlington, VA, pp. 29–68

Hiebert J, Carpenter T 1992 Learning and teaching with understanding. In: Grouws D A (ed.) *Handbook of Research on Mathematics Teaching and Learning*. Macmillan, New York, pp. 65–97

Hooper F, DeFrain J 1980 On delineating distinctly Piagetian contributions to education. *Genetic Psychology Monographs* **101**: 151–81

Kliebard H 1994 *The Struggle for the American Curriculum 1893–1958*, 2nd edn. Routledge, New York

National Association for the Education of Young Children (NAEYC) 1991 Guidelines for appropriate curriculum content and assessment in programs serving children ages three through eight. *Young Children* **46**(3): 21–38

Piaget J 1983 Piaget's theory. In: Mussen P (ed.) *Handbook of Child Psychology*, 4th edn. Wiley, New York, Vol. 1, pp. 703–32

Sameroff A, Haith M 1996 *The Five to Seven Year Shift: The Age of Reason and Responsibility*. University of Chicago Press, Chicago

Tharp R, Gallimore R 1988 *Rousing Minds to Life: Teaching, Learning, and Schooling in Social Context*. Cambridge University Press, Cambridge, UK

Vygotsky L 1978 In: Cole M, John-Steiner V, Scribner S, Souberman E (eds.) *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge

J. Brophy

Primate Behavior: Significance for Understanding Humans

Interest in the behavioral similarities between human and nonhuman primates can be traced to Huxley's (1863) assertion of a phylogenetic relationship between humans and African apes. From the outset, primate studies have been incorporated into diverse fields of inquiry, including biology, ecology, psychology, and anthropology. Physical anthropologists were intensely interested in the degree to which primate behavior could be viewed through an interpretive lens for insights into the behavior and mental capacities of ancestral humans. Contemporary primate studies continue to inform a broad array of fields in both the human and nonhuman behavioral sciences; this integrative approach contributes to our understanding not only of human evolution, but of the behavior of extant humans as well. In this article we briefly outline the history of primate behavioral studies within the field of physical anthropology, discussing past uses of primatology for building hypotheses of human evolution. We then discuss changes in the theoretical approach to the study of primate behavior and their effect on the understanding of social systems. We conclude by presenting current issues in primatology as they relate to the study of human social evolution and behavior.

1. Early Primate Models of Ancestral Hominin Behavior

Early primate models for the behavior of ancestral humans were based on either: (a) assumptions of similarity in ecological adaptation between extant savannah-living primates and ancestral humans, or (b) shared behavioral traits within the clade of African hominoids (including humans). Raymond Dart, the discoverer of the *Australopithecus* specimen at Taung, popularized the hypothesis of a saltatory shift to human bipedalism, tool use, and hunting as a result of moving from forests to savannahs (Dart 1957). Dart drew explicit parallels to baboons as contemporary savannah-dwellers that might share behavioral similarities to early humans. Brief field studies of wild baboons revealed distinct differences from modern humans. Nonetheless, early descriptions of baboon sociality, which emphasized male dominance and

aggression as important influences on social structure, were used to typify the human condition (Kinzey 1987). The historical prejudices and methodological weaknesses that gave rise to these models have received considerable critical scrutiny (Cartmill 1993). Despite their methodological weaknesses, these early studies planted the seeds for future primate field studies by asking the following questions: (a) In what way do ecological parameters shape social systems? (b) What were the relevant ecological parameters that shaped early human evolution?, and (c) What can extant primates tell us about human behavior and social evolution?

While ecological affinities led to the baboon model, phylogenetic affinities between human and nonhuman apes prompted field studies of chimpanzees, gorillas, orangutans, and bonobos (Hamburg and McCown 1979). Taking a comparative approach, Louis Leakey sponsored well-publicized field studies of three great ape species. Anthropologists focused most attention on the common chimpanzee, the closest living relative to humans (Sibley and Alquist 1987). Behaviors observed in both humans and chimps were assumed to be phylogenetically conservative and therefore likely to have been expressed in our shared common ancestor. Tool use, hunting/meat eating, and the savannah/woodland habitat of chimps at Gombe reinforced the confidence of some anthropologists in the importance of this suite of characters for the evolution of humans (Kinzey 1987).

Contrary to expectation, these studies revealed that, neither baboon nor chimpanzee models have strong explanatory power in reconstructing the behavior of ancestral humans. First, extant species evolved under unique selective pressures and are not satisfactory typological specimens of ancestral species. Second, there is significant intraspecific variation in the behavior of chimpanzees and baboons, with no adequate criteria for selecting which behaviors are applicable to models of ancestral humans. Third, if behaviors common to chimpanzees and humans are assumed to be phylogenetically conservative, there are significant problems with interpreting morphological evidence from the fossil record that is not shared by either modern chimpanzees or humans (see Potts, in Kinzey 1987).

These criticisms apply to the assumptions and methods employed by these modeling approaches and not to the usefulness of primate behavioral studies *per se* in elucidating past and current human behavior. Past mistakes have served to refine the approaches by which primate studies are conducted and applied. At the beginning of the twenty-first century, few primate studies make explicit use of typological models. Contemporary studies of primate behavior emphasize a comparative approach in which the functions of behaviors are investigated under different ecological and social conditions. The next section presents important conceptual changes in the study of primate behavior and their influence on the application of primate models to the understanding of human behavior.

2. *Conceptual Shifts in Primate Behavior Studies*

During the same time that primate behavior was being used to construct models of ancestral hominines, it was also an active area of interest in the zoological sciences, particularly in the subfield of behavioral ecology. Behavioral ecology investigates behavioral adaptations to ecological and social conditions. Although behavioral ecologists directly observe phenotypic traits, they use these observations to test falsifiable predictions generated within an evolutionary framework which states that individuals act in ways that increase their inclusive fitness. For example, studies testing optimal foraging theory use economic cost-benefit models to generate predictions of behavioral responses to ecological conditions that affect an individual's energy budget. Other studies, particularly those that focus on reproductive behavior such as mating and offspring care, examine an individual's decisions when situated in a particular social milieu (Krebs and Davies 1991). This approach to the study of primate behavior is distinctly different from that which was used previously in two important ways: (a) its goal is to test and further develop general evolutionary principles for behavior rather than create models of specific extinct species, and (b) comparative outgroups can be taken from all animals rather than those restricted to the primate order.

Within the field of anthropology, human behavioral ecologists applied evolutionary concepts to test economic models of the ecology of contemporary human hunter-gatherers. Human behavioral ecologists focused on contingencies that had been found to be important in non-human primates: predation, reproductive opportunities, and the distribution and predictability of resources. However, the application of primate studies to the interpretation of human behavior encountered significant difficulties following the publication of E. O. Wilson's *Sociobiology* (1975). Criticism of early sociobiological theory is well documented and expressed two main concerns. First, the effusion of speculations on the application of sociobiological principles to human behavior was untestable and therefore unfalsifiable (Caro and Borgerhoff Mulder 1987). The second body of criticism, which objected to sociobiological depictions of humans as inflexible executants of the interests of selfish genes, stemmed from a poorly explained differentiation between genotypic determinism and the plasticity of behavioral phenotypic expression. Suites of behaviors resulting from complex socioenvironmental conditions (e.g., 'rape') were incorrectly treated as simple units of evolutionary analysis (Sunday and Tobach 1985). Disagreement over the application of

sociobiological theory to the study of human behavior resulted in a division between cultural anthropologists, who emphasized human uniqueness, and those biological anthropologists who studied primate behavior with an eye toward understanding that of humans. While it is a truism that genetic make-up influences behavior, and genetic relatedness influences to whom particular behaviors may be expressed, genetic determinism has no place in behavioral ecology.

The contemporary use of primate behavior within anthropology has shifted from the use of primates as typological models of human origins to field studies and experiments that test predictions developed within a behavioral ecology framework. Significant methodological changes have accompanied this theoretical shift. First, standard methods for the collection of quantifiable data facilitate interpretable statistical tests of predictions and comparison across primate studies (Altmann 1974). Second, more primate species are studied, thereby providing a larger and more diverse sample for comparison. Third, long-term field studies provide information on seasonal and interpopulational variability in behavior, the ontogeny of and change in relationships among individuals, and the causes and consequences of demographic processes in long-lived species. Fourth, a more diverse representation of scientists expands the field of inquiry to include aspects of social life previously assumed ancillary to social organization, particularly female reproductive strategies and the energetics of offspring care (Altmann 1980, Hrdy 1999). Analyses of the resulting data sets demonstrate rich intra- and interspecific variability in behavior and relationships that parallel the intraspecific variability observed in humans.

3. *Current Primate Studies in Anthropology*

3.1 *Social Behavior*

Two general principles are commonly invoked to explain group size and composition in birds and mammals. First, female reproductive success is limited primarily by the availability of ecological resources, particularly food. With whom females form groups, and the competitive regime between females within those groups, is a function of the distribution of ecological resources in space and time (Wrangham 1980, van Schaik 1989). Male distribution, on the other hand, is patterned upon that of females (Andelmann 1986). The diversity of grouping patterns observed among primates across a range of environments supports this general framework.

Whether of the scramble or contest type, competition for resources underlies this ecological model (Krebs and Davies 1991). Competition, however, can be expressed in multiple ways. Aggression between group members certainly occurs, and individuals obtain inclusive fitness benefits by aiding kin in contests, but not all competition results in direct conflict. Unrelated individuals also engage in cooperative and affiliative behaviors, and reconciliation often occurs after conflicts (de Waal 1983). Several benefits accrue from affiliative interactions with nonkin. First, individuals who engage in post-conflict reconciliation experience more rapid reductions in stress response than those that do not (Aureli 1992). Second, initiating affiliative behaviors garners favor with powerful group members and develops relationships with potential future allies (de Waal 1983). Third, males who invest in nonsexual affiliative behavior toward females obtain future matings with their 'friend' (Smuts 1985). 'Peaceful' interactions are essential to the social cohesion of groups and provide an alternative means of exerting influence and eliciting support. Long life spans and the opportunity for repeated encounters between known individuals provide the conditions under which individuals can predicate present behavior on anticipated future social interactions (Byrne and Whiten 1988).

The study of behavioral ecology in humans is likewise concerned with the tactics employed by modern humans to survive and reproduce within a particular socioecological context. Using principles developed in the field of animal behavior, human behavioral ecologists collect data from modern hunter-gatherers to test predictions derived from ecological optimization models (Winterhalder and Smith 2000). Human behavioral ecologists are able to measure directly the economic results of behavioral traits thought to be important to human evolution such as the intersexual division of labor, long juvenile dependence, and females' lengthy post-reproductive life span. The results of such studies have led to important challenges to the primacy of hunting in human history, suggesting instead that the economics of female provisioning and care of offspring provide a more parsimonious explanation for the evolution of human social systems (Hawkes et al. 1997, O'Connell et al. 1999).

Because of their general applicability, the use of behavioral ecology models raises the question of whether primate studies in particular are any more informative than studies of other taxa. Several lines of evidence combine to produce a positive answer. While studies of all taxa, particularly social mammals, serve to test theories that may be relevant to humans regarding the causes and consequences of social systems, the complex nature of social relationships observed in nonhuman primates make them particularly useful as comparative species for understanding the function of human social relationships. In addition, the use of nonhuman primates as comparison species controls for traits more phylogenetically conservative than behavior that may influence behavioral options (e.g., life history parameters, functional morphology, and physiology). Confining the study of

behavioral ecology to humans within the field of anthropology limits the perspective from which to understand human behavior. Cross-cultural studies can demonstrate intraspecific variation, but they do nothing to reveal which traits or behaviors are uniquely human. Interspecific comparative studies among primates reveal the extent and degree to which certain behavioral patterns are confined to humans. Primate studies therefore both focus and expand the field of vision in the study of human behavior.

3.2 Functional Morphology

Anatomists have a rich tradition of inferring the behavior of extinct species based on fossil skeletons, and tooth morphology figures prominently in the reconstruction of hominin diets (Jolly 1970). Interpretation of fossils is only possible, however, with studies of analogous structures and their function in extant species. Recently, anatomists have teamed up with primate behavioral ecologists to combine direct observations of primate feeding behavior with dental morphology. This partnership helps refine the techniques used to determine ancestral diet and feeding patterns. For example, field studies of the feeding ecology of four sympatric Sumatran primates allowed Ungar (1996) to test whether incisor size corresponds with diet. His finding that incisor size and shape best reflect the mode of ingestion rather than diet *per se* resolves some conflicts in interpretation of the fossil record. Dental microwear analyses can reveal more detailed information about diet than can analysis of gross dental morphology. Recent field studies examine primate feeding behavior, measure the physical properties of foods, and cast the teeth of wild primates in order to better understand the relationship between food type, mode of ingestion, and microwear (Teaford 1994). Similar studies are being conducted on humans (Ungar and Spencer 1999). With further refinement of this new technique, analysis of dental microwear can play an important role in the reconstruction not only of the diet of ancestral hominines, but the use of teeth for non-feeding functions as well (Teaford 1994).

Studies of positional behavior investigate the relationship between postcranial musculoskeletal anatomy and locomotion. The means by which wild great apes travel in their natural environment reveal the function of bone and muscle in the limbs (Fleagle 1999), and comparative anatomy between great apes and fossil hominines provides a basis for inferring locomotor behavior in hominines. This approach has shown that bipedal hominines descended from an arboreal primate that moved primarily by climbing rather than brachiating (Fleagle 1999). Likewise, *Australopithecus afarensis*, the hominin most likely responsible for the famous Laetoli footprints (Leakey 1987), probably spent a good proportion of the day moving arboreally as well as bipedally (Stern and Susman 1983). Field observations also provide clues to possible sex differences in behavior. In apes, sexual dimorphism corresponds with differential use of arboreal and terrestrial substrates for moving, with males spending more time on the ground than females (Doran 1996). If the sexually dimorphic *A. afarensis* moved on the ground and in the trees, sex differences in substrate use may have likewise occurred.

3.3 Intelligence/Cognition

The comparative study of intelligence and cognition among nonhuman primates in the laboratory spans over half a century (Russon et al. 1996). The discovery of tool manufacture and use in wild chimpanzees (Goodall 1968) put to rest concerns that the development of the cognitive capacities necessary to use tools by laboratory primates was an artifact of intensive interaction with humans. Wild orangutans also manufacture and use tools with considerable complexity (Fox et al. 1999). In contrast, the proficiency with which capuchin monkeys use tools in both field and laboratory settings is significantly lower than that of great apes (Visalberghi and Trinca 1989). Great apes have crossed the line into a domain of intelligence once postulated to be uniquely human; furthermore, the intelligence required for ape-level tool use and manufacture evolved before the orangutan–African ape split (van Schaik et al. 1996).

Tool use behavior, however, is only one demonstration of intelligence. The decision-making processes required for large-bodied apes to travel arboreally is another (Povinelli and Cant 1995). Complexity in social interactions such as deception evinces cognition as well (Byrne and Whiten 1988). Primate studies provide empirical proof that many purported aspects of human uniqueness, such as intelligence and culture, are unique to humans only insofar as they are defined as such (Cartmill 1990). The corpus of work in this field demonstrates that, broadly defined, intelligence has evolved by degrees rather than as a result of a single selective pressure on early hominids. Diverse levels and types of cognitive capacities in an array of primates elicit two questions about humans: (a) is the higher intelligence demonstrated in humans an added *degree* of intelligence or is it a difference of *kind*, or both? and (b) what selected for human-level intelligence? Language is an obvious difference between humans and nonhuman primates, and it will undoubtedly figure prominently in answering these questions.

4. Future Directions in Research

Although most primates live in groups, a group is comprised of individuals with interests that often differ from those of their group-mates. In addition, those interests change as individuals grow to maturity, reproduce, and then, in some species, shift to a post-

reproductive phase of life. Building upon earlier research, future studies are needed to investigate the complexity of factors that influence individual decisions within the context of group life, in particular: (a) the strategies that individuals employ to satisfy their reproductive interests, even when those interests conflict with other group members, and (b) reproductive decisions that incorporate not just matings or fertilizations, but offspring survival as well.

Male aggression or threat of aggression toward females is observed among most species of primates, including humans (Smuts 1992, Smuts and Smuts 1993). Some of this conflict is with current or potential mating partners. However, significant variation in rates of sexual aggression occurs between and within species. For example, woolly spider monkey females copulate without aggression from their mates or other males, while male sexual aggression in rhesus macaques occurs frequently. In baboons, particular males develop 'friendships' with females while others do not (Smuts and Smuts 1993). There is even variation in individual males: some human males with pair-bonded mates force copulations with other females, and some pair-bonded males alternately exhibit affiliative and aggressive behavior toward their mate (Ellis 1989). Despite the seemingly widespread occurrence among primate species, however, extreme forms of direct sexual aggression in the form of forced copulations are uncommon in primates other than humans and orangutans. Some hypotheses to explain this striking contrast have been proposed and deserve further exploration.

What is the function of male sexual aggression? Under what conditions does it occur? Feminist scholars and biological and social scientists have long recognized that conflict between female and male reproductive interests influence individual behavior and shape the social systems in which individuals live (Hrdy 1977, MacKinnon 1989). Theoretical treatments of intersexual mating conflict in human and nonhuman primates provide a basis for predicting the conditions under which male sexual aggression, and conversely female countertactics to thwart males, occur (Smuts 1992, Smuts and Smuts 1993). Recent studies provide some preliminary basis for comparison (Smuts and Smuts 1993), but to strengthen the comparative method more studies conducted across a greater spectrum of primates are necessary. Of particular importance is quantifying the costs of sexual coercion, such as reduced foraging efficiency, in addition to infant or female mortality. New noninvasive hormonal assay techniques for field-collected samples are starting to be used (Whitten et al. 1998) and will allow researchers to measure endocrine response by females who are subject to male sexual aggression. Using these techniques, research can reveal the degree to which male aggression produces physiological effects that may negatively impact female fecundity.

Little is understood about why females resist mating by some males but not others. Equally little is known about female tactics to circumvent male attempts to restrict female reproductive choice. More studies are needed to determine the degree to which females select mates based on variation in quality among males (direct mate choice) or whether other factors, including male aggression, sufficiently constrain females to result in facultative, or indirect, female mate choice (Gowaty 1992, Wiley and Poston 1996). Whether for genetic quality, the quality of services, or constraints placed by males, with whom a female mates has an important reproductive outcome. Females are therefore predicted to exercise tactics, including sperm selection, that optimize their ability to retain control over reproductive choice and the timing of reproduction. Techniques for genetic analysis using noninvasive sampling methods have recently been developed (Borries et al. 1999) and will in the future facilitate evaluation of the success of both male and female reproductive strategies.

Variability among male reproductive decisions is likewise important. In contrast to the previously pervasive view within evolutionary studies and behavioral ecology, a recent shift recognizes that the acquisition of multiple mates is not the sole factor motivating male reproductive decisions. Whether they are excluded by other males, fail to attract mates, or are constrained by the need to allocate time to nonmating activities, male primates are limited in the number of females with whom they can actually mate and produce successful young. Males therefore optimize reproductive effort within operating constraints. While limitations on male mating opportunities have traditionally been viewed in terms of male–male competition and male mating aggression toward females, this is not always the case. Male choosiness for mates is likely to be one important tactic in species that have long periods of infant and juvenile dependency (Altmann 1997). Differences in female fertility, for example, may guide males away from young, females as mates. Furthermore, in primates conception and birth are merely the first steps to reproductive success. Offspring must survive into adulthood in order to reproduce, since reproductive success is most accurately measured in the second generation (Clutton-Brock 1988). Offspring survival is critically dependent on maternal competence, including mothers' ability to obtain access to resources necessary to rear young. Males are therefore predicted to prefer as mates those females who are higher ranking, older, and experienced (Altmann 1997).

Females with insufficient resources to rear young may abandon current offspring in order to invest in future reproductive opportunities. Alternatively, in order to meet the demands of rearing offspring, mothers in some species share the task of offspring care. Fathers sometimes significantly assist females, and the need for paternal care constrains male prom-

iscuity in some primates as it does in many species of birds (Clutton-Brock 1991). Likewise, related or unrelated individuals sometimes assist mothers. Related individuals gain inclusive fitness benefits resulting from the survival of grandchildren, sisters, and cousins, while both related and unrelated individuals may gain needed experience necessary for the survival of their future offspring (Hrdy 1999). Long juvenile dependency and the need for assistance in rearing offspring play an important role in the lives of extant hunter-gatherers, and it figures prominently in models of human social evolution (O'Connell et al. 1999). The function of extended juvenile dependency, how mothers meet the demands of extensive infant care, and the consequences for reproductive success are among the most fascinating questions of human and nonhuman primate life-history. Yet because primate life spans are so long, these questions are also among the hardest to study. As data from uninterrupted long-term field studies become available, questions concerning the fitness advantages to variation in parental and alloparental care can be addressed.

See also: Comparative Neuroscience; Comparative Psychology; Primate Families, Structure of; Primate Socioecology; Primates, Evolution of; Primates, Social Behavior of; Recognition Memory (in Primates), Neural Basis of

Bibliography

Altmann J 1974 Observational study of behavior: Sampling methods. *Behaviour* **49**: 227–67

Altmann J 1980 *Baboons Mothers and Infants*. Harvard University Press, Cambridge, MA

Altmann J 1997 Mate choice and intrasexual reproductive competition: Contributions to reproduction that go beyond acquiring more mates. In: Gowaty P A (ed.) *Feminism and Evolutionary Biology*. Chapman and Hall, New York

Andelmann S J 1986 Ecological social determinants of cercopithecine mating patterns. In: Rubenstein D I, Wrangham R W (eds.) *Ecological Aspects of Social Evolution: Birds and Mammals*. Princeton University Press, Princeton, NJ

Aureli F 1992 Post-conflict behaviour among wild long-tailed macaques (*Macaca fascicularis*). *Behavioral Ecology and Sociobiology* **31**: 329–37

Borries C, Laundhardt K, Epplen C, Epplen J T, Winkler P 1999 DNA analyses support the hypothesis that infanticide is adaptive in langur monkeys. *Proceedings of the Royal Society London* **266**: 901–4

Byrne R W, Whiten A (eds.) 1988 *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Clarendon Press, Oxford, UK

Caro T M, Borgerhoff Mulder M 1987 The problem of adaptation in the study of human behavior. *Ethology and Sociobiology* **8**: 61–72

Cartmill M 1990 Human uniqueness and theoretical context in paleoanthropology. *International Journal of Primatology* **11**: 173–92

Cartmill M 1993 *A View to a Death in the Morning: Hunting and Nature Through History*. Harvard University Press, Cambridge, MA

Clutton-Brock T H (ed.) 1988 *Reproductive Success: Studies of Individual Variation in Contrasting Breeding Systems*. University of Chicago Press, Chicago, IL

Clutton-Brock T H 1991 *The Evolution of Parental Care*. Princeton University Press, Princeton, NJ

Dart R A 1957 The osteodontokeratic culture of *Australopithecus prometheus*. *Transvaal Museum (Pretoria) Memoir* **10**: 1–105

DeVore I, Washburn S L 1963 Baboon ecology and human evolution. In: Howell F C, Bourlière F (eds.) *African Ecology and Human Evolution*. Aldine, Chicago, IL

Doran D M 1996 Comparative positional behavior of the African apes. In: McGrew W C, Marchant L F, Nishida T (eds.) *Great Ape Societies*. Cambridge University Press, Cambridge, UK

Ellis L 1989 *Theories of Rape: Inquiries into the Causes of Sexual Aggression*. Hemisphere, New York, NY

Fleagle J G 1999 *Primate Adaptation and Evolution*, 2nd edn. Academic Press, San Diego, CA

Fox E A, Sitompul A F, van Schaik C P 1999 Intelligent tool use in wild Sumatran orangutans. In: Parker S T, Miles H L, Mitchell R W (eds.) *The Mentalities of Gorillas and Orangutans*. Cambridge University Press, Cambridge, UK

Goodall J v L 1968 The behavior of free-living chimpanzees in the Gombe Stream Reserve. *Animal Behaviour Monographs* **1**: 161–311

Gowaty P A 1992 Evolutionary biology and feminism. *Human Nature* **3**: 217–49

Hamburg D A, McCown E R (eds.) 1979 *The Great Apes*. Benjamin/Cummings, Menlo Park, CA

Hawkes K, O'Connell J F, Blurton Jones N G 1997 Hadza women's time allocation, offspring provisioning, and the evolution of long post-menopausal life spans. *Current Anthropology* **38**: 551–77

Hrdy S B 1977 *The Langurs of Abu: Female and Male Strategies of Reproduction*. Harvard University Press, Cambridge, MA

Hrdy S B 1999 *Mother Nature*. Pantheon, New York

Huxley T H 1863 *Evidence as to Man's Place in Nature*. Williams and Norgate, London

Jolly C J 1970 The seed eaters: A new model of hominid differentiation based on a baboon analogy. *Man* **5**: 5–26

Kinzey W (ed.) 1987 *The Evolution of Human Behavior: Primate Models*. State University of New York Press, Albany, NY

Krebs, J R, Davies N B 1991 *Behavioural Ecology*. Blackwell Scientific, Oxford, UK

Leakey M D 1987 Introduction (to the hominid footprints). In: Leakey M D, Harris J M (eds.) *Laetoli: A Pliocene Site in Northern Tanzania*. Clarendon, Oxford, UK

MacKinnon C A 1989 *Toward a Feminist Theory of the State*. Harvard University Press, Cambridge, MA

O'Connell J F, Hawkes K, Blurton Jones N G 1999 Grandmothering and the evolution of *Homo erectus*. *Journal of Human Evolution* **36**: 461–85

Povinelli D J, Cant J G H 1995 Arboreal clambering and the evolution of self-conception. *Quarterly Review of Biology* **70**: 393–421

Russon A E, Bard K A, Parker S T (eds.) 1996 *Reaching into Thought: The Minds of Great Apes*. Cambridge University Press, Cambridge, UK

Schaik C P van 1989 The ecology of social relationships amongst female primates. In: Standen V, Foley R A (eds.) *Comparative*

Socioecology: The Behavioural Ecology of Humans and other Mammals. Blackwell Scientific, Oxford, UK
Schaik C P van, Fox E A, Sitompul A F 1996 Manufacture and use of tools in wild Sumatran orangutans. *Naturwissenschaften* **83**: 186–8
Sibley C G, Alquist J E 1987 DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. *Journal of Human Evolution* **26**: 99–121
Smuts B B 1992 Male aggression against women: an evolutionary perspective. *Human Nature* **3**: 1–44
Smuts B B, Smuts R W 1993 Male aggression and sexual coercion of females in nonhuman primates and other mammals: evidence and theoretical implications. *Advances in the Study of Behavior* **22**: 1–63
Stern J T, Susman R L 1983 The locomotor anatomy of *Australopithecus afarensis*. *American Journal of Physical Anthropology* **53**: 69–80
Sunday S R, Tobach E (eds.) 1985 *Violence Against Women: A Critique of the Sociobiology of Rape*. Gordian Press, New York
Teaford M F 1994 Dental microwear and dental function. *Evolutionary Anthropology* **3**: 17–30
Ungar P S 1996 Relationship of incisor size to diet and anterior tooth use in sympatric Sumatran anthropoids. *American Journal of Primatology* **38**: 145–56
Ungar P S, Spencer M A 1999 Incisor microwear, diet, and tooth use in three Amerindian populations. *American Journal of Physical Anthropology* **109**: 387–96
Visalberghi E, Trinca L 1989 Tool use in capuchin monkeys: distinguishing between performing and understanding. *Primates* **30**: 511–21
Waal F de 1983 *Chimpanzee Politics: Power and Sex Among Apes*. Harper and Row, New York
Whitten P L, Brockman D K, Stavisky R C 1998 Recent advances in noninvasive techniques to monitor hormone-behavior interactions. *Yearbook of Physical Anthropology* **41**: 1–24
Wiley R H, Poston J 1996 Indirect mate choice, competition for mates, and coevolution of the sexes. *Evolution* **50**: 1371–81
Wilson E O 1975 *Sociobiology*. Harvard University Press, Cambridge, MA
Winterhalder B, Smith E A 2000 Analyzing adaptive strategies: Human behavioral ecology at twenty-five. *Evolutionary Anthropology* **9**: 51–72
Wrangham R W 1980 An ecological model of female-bonded primate groups. *Behaviour* **15**: 262–300

E. A. Fox and J. Altmann

Primate Families, Structure of

We often talk about animal 'families' when referring to the breeding units of animals. The word 'family' however, is derived from the Latin word for 'household' in which a man controls the resources of his wife and offspring. Such entities seem absent in non-human primates. In contrast, most female primates and their dependent offspring forage largely independently of the males. Furthermore, mating and breeding occur in a large variety of ways. Biologists use the term 'mating system' to describe the patterns of behavior related to mate choice and reproduction. Such mating systems are intimately linked to the social system in which the animals live, but these systems are not identical. Mating partners include, but need not be restricted to, social partners. A key characteristic of female primates is their monthly ovulation. Breeding seasons do exist in many primates, but overall periods of female fertility are more frequent than in other mammals. Accordingly, sexual behavior is frequent and has acquired additional functions in many primates.

1. *Diversity of Primates and Their Mating Systems*

1.1 *Types of Mating Systems*

Primates include two hundred species ranging in size from mouse-sized prosimians and pigmy marmosets weighing less than 100 g to orangutans and gorillas with males weighing over 200 kg. Most primates are tree-dwelling but some, such as gorillas and baboons, are terrestrial. While some of the prosimians (a group of relatively 'primitive' primates such as galagos, pottos, lorises, and lemurs) are nocturnal and mostly solitary, the vast majority of primates are diurnal and extremely social. In fact sociality and sexuality are key primate adaptations. Primate mating systems represent a wide spectrum ranging from polyandry, where a single female mates with several males (South American tamarins), to monogamous pairs (owl monkeys, gibbons), to multi-male, multi-female groups (chimpanzees, spider monkeys, red colobus monkeys, and baboons) to harem type groups where a single male has exclusive reproductive access to multiple females (gorillas and langurs). Single-male groups can experience temporal visits by groups of bachelor males (Guenons) and solitary males and females can temporarily gather (orangutans). Size differences between males and females (called sexual dimorphism) seem to correlate with the degree to which males of a given species are polygamous. The males and females of the (serially) monogamous gibbons are the same size but males of the polygamous gorillas and mandrills can weigh more than twice as much as females. Primate mating systems are far more flexible and dynamic than previously thought. The same species may show different mating systems in different locations or at different times in the same location (Dunbar 1988). Changes in resource distribution or social changes can cause animals to alter the ways in which they find and choose mating partners.

1.2 Reproductive Costs for Male and Female Primates

Sexual reproduction refers to the process by which male and female germ cells fuse and produce individuals with unique combinations of genes. The germ cells carry half of the reshuffled genes that the parents inherited from the grandparents. Female germ cells (ova) are large and costly to produce whereas male germ cells (sperm) are tiny and much cheaper to produce. In animals with internal fertilization and gestation, this discrepancy in reproductive effort is exaggerated. Female mammals bear an even more disproportionate cost of reproduction because of lactation (the physiological cost of which is higher than gestation itself). Male investment often ends with successful insemination, after which males can look for further mates. The average number of offspring born to male and female species is roughly the same for both sexes, but the variance is much higher in males, where some highly successful males sire large numbers of offspring and many others none at all.

Biologists have repeatedly pointed at the imbalance between male and female reproductive costs to explain why males have a natural tendency to be promiscuous while females tend to be coy and choosier (Hrdy & Williams 1983). This caricature of male emphasis on quantity and female emphasis on quality is in part responsible for the fact that students of animal behavior have long overlooked the importance of female choice and multiple paternity (Gowaty 1992). Females, much rather than just being passive resources, for which males compete, are pursuing their own reproductive interests very actively. Their strategies are often less conspicuous and may involve long-term patterns, which explains why they have so often been overlooked.

Male primates can undoubtedly increase their lifetime reproductive success by maximizing the number of matings with fertile females. In primates, where mating often has more functions than simple fertilization, there may be strong advantages to multiple matings for females as well. Primates have long periods of development and infants are very dependent on parental assistance. Parental care is crucial to primate reproduction and is mostly the responsibility of mothers, though, in many cases males are also involved. The tamarins of the Amazon forest even have cooperative paternal care for twins born to the female. The young of many primates benefit from the protection and care from males with whom their mothers associate. Infant baboons and gibbons are often carried by males and chimpanzee males are known to adopt orphaned male youngsters. Wherever male help with the rearing of offspring is needed and there are several males present in the group, incertitude of paternity can cause males to assist offspring that they did not sire, resulting in a strong motivation for females to mate with many different males. Too much incertitude about paternity is potentially dangerous, as males can be a threat to offspring that they perceive as not their own.

1.3 Sex as a Social Tool

In bonobos, apes known for their frequent sexual interactions, sex is directly used to reduce tension in the group and is often exchanged for food. Chimpanzee females often exchange sex for meat that tends to be controlled by males, which do most of the hunting and use prey as a social currency. A female chimpanzee will copulate more than six thousand times with a very large number of males (10 to 30) while she can only produce a maximum of six surviving offspring. Similarly, the high frequency of same-sex sexual contact in bonobos plays important, entirely social roles. In fact, bonobos appear to be the one species of primates, where frequent female–female sex facilitates the existence of strong female alliances which are strong enough to keep male aggression in check.

1.4 Sexual Advertising in Multi-male Groups

Females in many species with multi-male groups exhibit conspicuous pink swellings of their genital area during several days around the time of their ovulation. These swellings are known as estrous swellings and they represent a formidable 'tool' for female primates. A female with a swelling becomes the focus of attention of all males in the group. Her social status increases and males will compete for access to her. Several females can have estrous swellings at the same time, thus making it virtually impossible for any given male to monopolize access to all of them. Estrous swellings appear to play an important role for species where females disperse and join new social groups after they become sexually mature. Swellings are a kind of 'social passport' which greatly facilitates a female's acceptance by unfamiliar males. There is evidence for certain amounts of disconnect between the advertising of fertility and actual time of ovulation. Such a disconnect could work to the advantage of the female because she could escape the controlling efforts of certain males and exert a freer choice of paternity for her offspring. The topic of 'concealed ovulation' takes on considerable importance as many scientists have suggested that humans have attained a stage of perfectly concealed ovulation (concealed even from a large proportion of women themselves). To what degree the inability of human males to detect the moment of peak fertility in women has played a role in leading to the high frequency of (serial) monogamy in our species is a much-debated issue.

Some primates use vocalization as a way to attract mates. Notable cases are the orangutan, gibbons, and

siamangs. In the orangutan, an interesting case of multiple male mating strategy includes sexually mature but small males, that do not sing to attract females but rather rely on forced matings between several males and a very reluctant female. Female criteria for choosing their mates remain poorly understood. Females can certainly learn about future social partners through their yearlong interactions with males, but for detecting 'good genes' they must rely on indirect signs such as body symmetry and relative absence of disease and parasites.

2. Determinants of Sociality in Primates

2.1 The Need for Social Grouping

The distribution of resources and the presence of predators play key roles for the distribution of individuals across space. When resources are clumped, groups of animals can live around a particular cluster of resources and defend a certain territory. Such groups can form cohesive social units. When resources are few and very dispersed, no single cluster of resources can be defended even by large groups of animals. Such environments tend to harbor animals living in much smaller groups or as dispersed individuals. Even highly social populations of chimpanzees are known to scatter into much smaller sub-groups during periods of low food abundance. The solitary orangutans on the other hand can gather in large numbers during periods of plenty (mast fruiting).

There are many good reasons for living socially. More eyes can see better and more ears can hear better. Living socially facilitates the search for food and the detection of predators. Another advantage is the phenomenon of the 'selfish herd' where the presence of other individuals spreads the risk of predation, as any of the others is as likely to be preyed upon as oneself. Also, cooperation in anti-predator defense is only possible in groups. Obviously, social life facilitates finding mates, even if it will increase the competition for them. Animals in groups can cooperate in defending resources against competitors. Such cooperative resource defense, when done by males, may have the added benefit of attracting females to the group, which may represent unwelcome competition to the females already resident in the group. Finally, the presence of individuals possessing different capabilities (such as hunting, raiding beehives, remembering the time and place certain foods can be harvested) in a group will benefit all group members. Unique individual experience and history shape large-brained, long-lived organisms such as primates. The mating strategies within a social group can vary dramatically between individuals. Some males, for example, may not show any ambition with respect to the dominance hierarchy. They spend their whole life as subordinate males. Their affiliative behavior and friendship with certain females in the group may in the end translate into numerous offspring. Certain females may be much more successful at forming social bonds with dominant males. Sheer force can allow a male to ascend in the dominance hierarchy, but the majority of animals in his group may not follow him. The cognitive capacities of primates appear to have coevolved with their tendency for sociality. Long-lasting relationships with many group members require good memory and social intelligence.

2.2 Costs of Living Socially

There are multiple costs to sociality. Competition between males can create social instability and conflict. Such conflict is often violent and can lead to harm to bystanders or even infanticide by males hoping to remove the offspring of competitors and gain rapid access to fertile females (the loss of an infant causes lactating females to resume ovulatory cycles). Competition occurs not only between males for access to females, but also between females for access to males, access that may lead to support (shared resources) and male protection during social conflicts. Competition between female may be the factor that precludes the formation of strong female coalitions against males in many species. Infanticide perpetuated by females has been observed in some wild primate groups and may represent the darkest side of female–female competition for resources. The study of female dominance hierarchies in natural primate populations is still in its infancy. Such hierarchies tend to be more subtle, nevertheless, female social rank can play an important role for reproductive success and change the way females invest in sons or daughters. Finally, higher risk of infection by pathogens and parasites is an important cost to sociality.

2.3 Most Primate Groups are Female Bound

In most primates, females and their female relatives remain in their native groups while most males disperse to other groups when they reach sexual maturity. Some notable exceptions to female residence are found in colobus and spider monkeys and in our closest living relatives, chimpanzees and bonobos, where females disperse upon reaching sexual maturity and males remain in their natal communities. The driving force behind dispersal is thought to be inbreeding avoidance and/or competition with animals of the same sex (Moore & Rauf 1984). Sexual mimicry in white-cheeked gibbons, where adolescent females look like males, and in red colobus, where young males have pseudo-swellings, indicates that same-sex competition is important in those species. The mechanisms

underlying inbreeding avoidance are poorly understood. Whether primates are able to recognize their paternal kin by features inherited from the father or whether they simply avoid mating with individuals from the same age cohort is still unclear (Alberts 1999). In humans, close proximity at an early age seems to be the crucial mechanism that triggers avoidance as sexual partners.

2.4 Do Males Follow Female Spacing?

The way females deploy themselves across time and space tends to determine how males can space themselves around females. Only if the ecosystem is capable of feeding a large number of females with enough tolerance for each other can males attempt to control such groups, either as the solitary reproductive male in a harem or as a group of males cooperating in female and territory defense and sharing the reproductive access to those females. All multi-female groups create the potential for female–female competition over access to resources and access to males and the support they may provide. Competition between potential mates may create the opportunity to enhance the visibility of positive traits or weaknesses. Because competition involves stress and stress is known to impact an animal's immune system, competition can exaggerate the visible differences in male quality. If female choice incites increased male competition, then female choice may indirectly lead to an increased size difference between males and females. Because strong and assertive males can be more successful at guarding resources that they share with females, females may favor the evolution of such males. Females need male protection from unwanted attention of other males, including, in the extreme case, forced copulation, as well as against the threat of male infanticide; both are part of male mating strategy in many primate species (Smuts 1993). Strong and protective males by the same token will be more successful at restricting the independence of females.

3. The Importance of Genes

3.1 The Benefit of Multiple Paternity to Females

Mating with multiple males can be advantageous for a female in various ways. As mentioned above, confusing paternity of her offspring can effectively protect them from male infanticide. She may gain support from her mating partners, both in terms of social support during conflicts with other group members and in terms of material support in the form of shared resources. More recently, the possibility of a genetic benefit has been proposed and some evidence is starting to accumulate (Zeh & Zeh 1996, 1997). Because the primate fetus develops inside the mother, the female and the fetus need to achieve a type of 'immunological détente.' That is, the fetus must not be rejected by the mother's powerful immune system but, at the same time, the demands of the fetus must be kept in check, not allowing it to exact too high a toll on maternal resources. Multiple paternity will produce more diverse offspring. Such diversity may have much to do with defense against disease. The vertebrate immune system is an immensely sophisticated and very costly system, in order to allow long-lived organisms such as primates to survive in a world in which very short-lived, rapidly evolving parasites and pathogens (ranging from worms and protozoa to bacteria and viruses) are a constant threat.

The immune system relies on genetic variation present in all populations, as well as generated anew in each individual in order to produce a flexible repertoire of molecular probes (antibodies and T-cell receptors) which detect specific intruding pathogens. It seems critical that each fetus inherits genes coding for two rather different immunological 'tool kits' in order for the offspring to be well equipped for its countless future encounters with pathogens. Conceiving offspring with different fathers would allow females to produce sets of offspring with more variation in their immune repertoires, strongly reducing the risk of losing all or most offspring to the same disease. The same obviously applies to males as well. Furthermore, paternal and maternal genes have to work in harmony in order to produce healthy offspring. Some combinations may simply fail to work and yet others may work but put the offspring at a disadvantage from the beginning. The detection of embryos with strongly limited immune repertoires (i.e., two almost identical sets of multi-histocompatibility complex or 'MHC' alleles) inherited from both parents, can cause spontaneous abortions in many mammal species including our own. Because of the importance of compatibility between genomes, the most promising genes for each female are found in different males. Finally, multiple paternity may allow females to reduce inbreeding where females cannot recognize close kin.

3.2 Molecules and Behavior

Ova and sperm engage in intense molecular dialogue, a process which is important for creating reproductive barriers between species but which is also suspected to involve an additional level of (post-mating) female choice. Female primates are known to produce a large quantity of antibodies against sperm in their genital tracts, a process which strongly favors diversity in paternity as the antibodies would be formed against the sperm of the 'usual' mating partner leaving sperm from unknown males at an advantage. The role of chemoreception (pheromones) in regulating some primate behavior remains very poorly understood.

The fact that primates may be able to smell important immune molecules (MHC) points towards potentially important roles of such molecules and their detection in mate choice and sexual behavior (Wedekind & Füri 1997).

3.3 Post-mating Female Choice and Sperm Competition

Certain sperm may be more likely be taken up by the egg as they approach it. Such cryptic female choice would represent a choice of individual sperm, possibly based on the genes carried by that sperm. Certain genes of crucial importance in early embryonic development are known to be expressed on the surface of sperm. Males can compete for fertilization of females via a process called 'sperm competition.' When females mate with many males within a short period of time, the sperm of different males will be available for fertilizing her egg. This exerts pressure on males to produce larger quantities of sperm in order to succeed in preference to other males. In chimpanzees, where multiple matings with many males give rise to strong sperm competition, males have conspicuously large testicles and produce large amounts of sperm. They even produce 'copulatory plugs,' a gelatinous mass blocking the entry to the uterus. In gorillas, on the other hand, where silver back males face very little competition in their harems, males have very small testicles. Female–female competition may take place via the depletion of sperm supply for other females.

3.4 Genomic Conflicts

Because different males can sire subsequent offspring, a conflict of interest can arise between the paternal and the maternal genes. A female will spread her total reproductive efforts across all her subsequent offspring. Males are likely to have sired offspring with other females and thus are spreading their reproductive effort across several females. Paternal genes in the embryo are likely to cause each embryo to be overdemanding from the point of view of the mother. Such paternal genes will increase the male's reproductive success at a detriment to each female, who will have invested too much in that particular offspring. It is this conflict which seems to be at the root of a phenomenon called 'imprinting.' Imprinting occurs when, among the two variants of each genes present in the embryo (the paternal and maternal alleles), only one of the two is expressed. Often, only the allele inherited from the mother is expressed, while the paternal allele appears to be silenced. Imprinting may have several functions but one of them may be female control of overdemanding genes from the father (Mochizuki et al. 1996, Spencer et al. 1998).

Primate mating systems represent the outcome of a delicate balance between various constantly changing forces. These include:

(a) ecological factors such as the distribution of resources and pressure by predators;

(b) female foraging and mating strategies, including the need to find multiple fathers of high genetic quality and good genetic compatibility and one or more supportive social partners;

(c) male mating strategies that may entail the access to the largest number of fertile females often in severe competition with other males; and,

(d) the need for parental care of infants and juveniles.

Compromises between male and female strategies seem to occur at every level, from gene expression to behavior. The existence of life-long friendships between adult primates of both sexes introduces an additional dimension, that of emotional attachment and trust between individuals.

4. Future Directions

Long-term field studies of primates in their natural habitat are being complemented by powerful molecular methods of analysis. Genetic analyses are providing detailed information on pedigrees and kinship, while hormone studies can provide knowledge of female reproductive state. Studies of hormones, such as oxytocin and vasopressin and their receptors, which seem to have dramatic effects on male social attachment behavior, are likely to produce new insights (Young et al. 1999). The study of genomic conflicts and imprinting is in its infancy. Similarly, further insight into the determinants of genetic compatibility and the maximizing of immune repertoires will produce much badly needed knowledge. The large number of longitudinal studies of natural primate communities under way is bound to reveal additional ways in which males and female primates strike compromises and successfully reproduce in ever-changing environments. Understanding mating systems is not only essential if one wants to understand evolution, because many primates are facing extinction, such an understanding is also crucial for conservation efforts in the wild and in captivity.

See also: Group Processes, Social Psychology of; Primate Behavior: Significance for Understanding Humans; Primates, Social Behavior of

Bibliography

Alberts S C 1999 Paternal kin discrimination in wild baboons. *Proceedings of the Royal Society of London, Biology* **266**: 1501–6

Dunbar R I M 1988 *Primate Social Systems*. Cornell University Press, Ithaca, New York

Gowaty P A 1992 Evolutionary biology and feminism. *Human Nature* **3**: 217–49
Hrdy S B, Williams G C 1983 Behavioral biology and the double standard. In: Wasser S K (ed.) *Social Behavior of Female Vertebrates*. Academic Press, New York, pp. 3–17
Mochizuki A, Takeda Y, Iwasa Y 1996 The evolution of genomic imprinting. *Genetics* **144**: 1283–95
Moore J J, Rauf A, Ali R 1984 Are dispersal and inbreeding avoidance related? *Animal Behaviour* **32**: 94–112
Smuts B B 1993 Male aggression and sexual coercion of females in non-human primates and other mammals: Evidence and theoretical implications. In: Slater P J B, Rosenblatt J S, Milinski M, Snowdon C T (eds.) *Advances in the Study of Behavior*. Academic Press, New York, pp. 1–63
Spencer H G, Feldman M W, Clark A G 1998 Genetic conflicts, multiple paternity and the evolution of genomic imprinting. *Genetics* **148**: 893–904
Wedekind C, Füri S 1997 Body odour preferences in men and women: Do they aim for specific MHC combinations or simply heterozygosity? *Proceedings of the Royal Society of London, Biology* **264**: 1471–9
Young L J, Nilsen R, Waymire K G, MacGregor G R, Insel T R 1999 Increased affiliative response to vasopressin in mice expressing the V1a receptor from a monogamous vole. *Nature* **400**: 766–8
Zeh J A, Zeh D W 1996 The evolution of polyandry I: Intragenomic conflict and genetic incompatibilty. *Proceedings of the Royal Society of London, Biology* **263**: 1711–7
Zeh J A, Zeh D W 1997 The evolution of polyandry II: Post-copulatory defences against genetic incompatibilty. *Proceedings of the Royal Society of London, Biology* **264**: 69–75

P. Gagneux

Primate Socioecology

Sociality is the primary behavioral adaptation of the anthropoid primates, the monkeys and apes. All of the anthropoid primates, except orangutans, spend most of their lives in stable groups of familiar individuals. However, the social organization among the anthropoid primates encompasses great diversity, ranging from the pair-bonded groups of gibbons and owl monkeys, to the polyandrous groups of tamarins; one-male groups of howlers and blue monkeys; multimale, multifemale groups of cebus monkeys and baboons; and fission-fusion communities of chimpanzees and spider monkeys. This great diversity in social organization shapes the behavioral options available to males and females and profoundly influences the evolution of social behavior and reproductive strategies in both sexes.

1. *The Evolution of Social Organization*

The great diversity in primate social organization has prompted researchers to ask a number of related questions: Why do primates live in groups? How should groups be structured? How big should groups be? These questions were first raised in the 1960s and 1970s by primatologists who conducted comparative analyses searching for broad ecological correlates of social organization (reviewed by Janson 2000). These comparative efforts yielded some consistent associations between diet, home range size, group size, body weight, and group composition. But the analyses were unsatisfying because they could not account for much of the observed variation in social organization among primates. Moreover, while these studies revealed a number of robust associations among various ecological variables, they did not produce much insight about the causal processes underlying these correlations. Finally, these comparative studies did not consider the possibility that some of the similarities between species were the result of common phylogenetic history, not independent adaptations to ecological conditions, a problem that has become a central methodological concern in comparative studies in behavioral ecology.

In the 1980s researchers returned to these questions, equipped with the theoretical and methodological tools of contemporary behavioral ecology (Krebs and Davies 1993). Behavioral ecologists try to explain how evolutionary processes shape behavior in particular ecological contexts. Sociality is presumed to reflect a dynamic balance between the advantages and disadvantages of living in close proximity to conspecifics. Sociality is beneficial because animals who live in social groups are better able to acquire and control resources or less vulnerable to predators. On the other hand, sociality is disadvantageous because group-living animals may face more competition over access to food, become more vulnerable to disease, and become more susceptible to cuckoldry, infanticide, and cannibalism. The size and demographic composition of the groups that we see in nature reflects a compromise between the costs and benefits of sociality for individuals who occupy particular ecological niches.

1.1 *Why do Primates Live in Groups?*

For primates, the primary advantages of sociality are thought to be enhanced access to resources, reduced vulnerability to predation, or some combination of both these factors. Many primates feed on foods, such as fruit, that occur in patches large enough to feed several individuals at the same time. Taking note of this fact, Wrangham (1980) suggested that primates aggregate because groups are more successful in defending access to resources than lone individuals, an argument that has also been applied in other taxa. Thus, the potential for between-group competition over resources favors sociality.

van Schaik (1983, 1989) questioned Wrangham's explanation of the evolution of sociality, arguing that the benefits accrued in between-group competition would be offset by the costs incurred from increased within-group competition. Instead, van Schaik argued, grouping evolved as a defense against predators. Primates are preyed upon by a variety of predators, including leopards, raptors, snakes, and other primates. Grouping provides safety from predators for three reasons. First, levels of vigilance are higher because there are more individuals scanning for predators. Second, groups are more effective in actively mobbing or deterring predators. Third, any single individual is less likely to be caught in a group if predators strike at random. Thus, primates aggregate to obtain protection from predators, and within-group competition for food or safe positions is an unavoidable side effect of sociality.

Initial efforts to evaluate the relative importance of between-group competition and predation among primates provided some support for both models (Janson 2000). Thus, the between-group competition model was supported by evidence that hostile intergroup interactions are common among primates, and larger groups are typically able to supplant smaller groups in intergroup encounters. In some populations, large groups are better able to defend higher quality home ranges than smaller groups. However, within-group competition seems to reduce foraging efficiency consistently in gregarious species (Janson 1992).

The predation model was supported by evidence that group sizes are reduced when predators are absent and that members of large groups spend less time scanning for predators and detect predators sooner than members of small groups do (Janson 1992). Nonetheless, vigilance levels do not decline monotonically as group size increases (Janson 2000), suggesting that groups are larger than needed for predator defense.

1.2 The Nature of Competition and Female Social Relationships

As researchers continued to debate the role of resource competition and predation in the evolution of sociality, they also began to think more carefully about the effects of competition on primate females. The distribution of food resources determines the distribution of females, and this in turn determines the distribution of males. Females come first in this scenario because female fitness depends mainly on their nutritional status (Wrangham 1980), while male fitness depends primarily on their ability to obtain access to fertile females.

Wrangham (1980) hypothesized that variation in the extent of intergroup competition would produce two distinct kinds of social arrangements. When resources were clumped and abundant enough to feed more than one individual, females would benefit from banding together to collectively defend access to them. Coalitionary alliances among kin are expected to be more beneficial than alliances among nonkin because kin share a common genetic interest (Hamilton 1964); thus females were expected to form nepotistic alliances. Female philopatry preserves matrilineal networks, and strong affiliate bonds strengthen alliances. Thus, female-bonded kin groups are expected to be associated with between group competition over resources. On the other hand, if resources are not clumped or occur in patches that are too small to feed several individuals, then the collective defense of resources will not be profitable. Nepotistic alliances will not enhance females' access to resources. In this situation, selective pressures favoring female philopatry and strong female bonds will be relaxed. When there is little between-group competition, non-female-bonded groups are expected to occur. Wrangham linked dietary specializations to social organization, arguing that fruit occurs in discrete patches that are often large enough to make collective defense profitable. He predicted that female-bonded groups would be mainly frugivorous.

Although it was rather easy to classify primate species as female-bonded or non-female-bonded, a number of empirical and theoretical objections to Wrangham's model were raised. However, all subsequent efforts to link ecology and social behavior have relied on Wrangham's insight that resource competition among females drives the evolution of social organization among primates.

van Schaik (1989) considered how the distribution of resources influences the nature of competition among females. Scramble competition occurs when individuals cannot effectively monopolize access to resources, and consequently do not compete over them directly. Contest competition occurs when resources are limited and can be monopolized profitably, generating direct confrontations over access to them. van Schaik suggested that within group contest competition (WGC), within group scramble competition (WGS), and between group contest competition (BGC) would have different effects on social organization (Table 1). (Between group scramble competition is assumed to be an automatic consequence of population density, but is not expected to influence social relationships directly).

When females compete over access to resources with group members, access to resources will be a function of individual dominance rank Van Schaik predicted that WGC would be associated with strict linear dominance hierarchies. Since dominance rank influences female access to resources and ultimately affects their fitness, females will provide coalitionary support to their kin in order to enhance their dominance rank. This will lead to matrilineal dominance hierarchies and create strong selective pressures favoring female philopatry. Nepotistic support will also enhance the

Table 1
Effects of competitive regime on social organization and behavior

	Competitive regime			
	WGC only	WGC + BGC	WGS + BGC	WGS only
Characteristics of dominance hierarchy	Strictly linear Nepotistic Very stable	Moderately linear Nepotistic Fairly stable	Egalitarian Individualistic Unstable	Egalitarian Individualistic Unstable
Coalitionary support	Nepotistic	Nepotistic in WGC, non-nepotistic in BGC	Non-nepotistic	Absent
Nature of female bonds	Highly differentiated	Highly differentiated	Poorly diffentiated	None
Dispersal from natal group	Males only	Males only	Males only	Males/Females
Social organization category[a]	Resident-nepotistic	Resident-nepotistic-tolerant	Resident-egalitarian	Dispersal-egalitarian

a Following Sterck et al. 1997.

stability of dominance hierarchies by dampening the effects of temporary variations in individual fighting ability. Affiliative interactions with coalition partners will reinforce alliances, so females are expected to form close and well-differentiated social relationships.

When WGS is the primary form of resources competition, female reproductive success will depend on the size of the female's social group, not her own rank. This means that there will be little incentive to aid relatives in disputes with other group members, and dominance hierarchies will not be nepotistic or strictly linear. If WGS is combined with strong BGC, females will rely on coalitionary support in intergroup encounters from all group members. Because females benefit more from supporting kin than nonkin, female philopatry will be favored. Since females' fitness depends on collective group action, females are expected to cultivate social ties with all other group members. Thus, females are expected to form relatively poorly differentiated social relationships.

If WGS is associated with weak BGC, then selective pressures favoring female philopatry will be relaxed. Female residence in social groups should reflect the balance between the benefits of living in groups of particular sizes and the costs of emigration from groups of suboptimal size.

1.3 Testing Socioecological Models in the Field

In the 1990s, many researchers set to work testing the predictions derived from socioecological models. This body of work confirmed many of the predictions derived from socioecological models. For example, Sterck et al. (1997) showed that coalitionary support, linear dominance hierarchies, and female philopatry are linked in a sample of 27 primate taxa. Others have compared closely related species, that differ in key features of their ecology. For example:

(a) Mitchell et al. (1991) compared two closely related species of squirrel monkeys in Costa Rica and Peru that had similar diets, formed groups of similar sizes, and experienced similarly high levels of predation. The two species differed mainly in the extent of within-group competition over access to food resources. As predicted by socioecological models, the species that experienced high within-group competition also formed stable, linear dominance hierarchies, kin-based coalitions, and exhibited strict female philopatry. In the species that experienced low levels of within-group competition, a dominance hierarchy was not detected among females, female coalitions were not observed, and females sometimes left their natal groups.

(b) Barton et al. (1996) compared the behavior and ecology of baboons at two sites. In Laikipia, Kenya, where within-group competition for food and predation pressures were high, groups tended to be large, females formed strong bonds with one another, and females rarely transferred from their natal groups. In the Drakensburg Mountains of South Africa, food was more evenly distributed and predation pressure was relaxed. There, groups were typically smaller, bonds among females were weaker, female dispersal was more common, and bonds between females and males were strengthened.

(c) Isbell and her colleagues have examined the relative importance of phylogeny and ecology in shaping the behavior of female monkeys in two closely related species, patas (*Erythrocebus patas*) and vervets

(*Cercopithecus aethiops*). If phylogeny is the primary factor influencing social organization, then patas and vervets should be very similar. However, if ecological factors are crucial, then patas and vervets should differ because their diets differ. Patas rely more heavily than vervets on arthropods and other food items that require little handling time and are consumed very quickly. On the other hand, vervets rely more heavily than patas on foods, such as gums, seeds, and fruits, that have longer processing times and take longer to eat. Patas have less linear dominance hierarchies than vervets do, and dominance rank has a less pronounced effect on nonagonistic behavior tightly linked to ecology (such as ranging, diet, and activity budgets) among patas than among vervets (Isbell et al. 1999).

(d) Koenig et al. (1998) compared two populations of Hanuman langurs (*Presbytis entellus*) living at different sites in India. The langurs of Ramnagar relied heavily on just three food plants. These plants were low in abundance and clumped in their distribution, conditions that are expected to lead to contest competition. The langurs of Kanha relied on a larger array of plant foods, and one of their most highly preferred foods is superabundant, conditions that are expected to lead to scramble competition. The langurs of Ramnagar behave much as expected when WGC is strong: they have linear dominance hierarchies and females are philopatric. The langurs of Kanha have poorly developed dominance hierarchies and females frequently emigrate from their natal groups.

1.4 Distribution of Males

In socioecological models, males generally play a secondary role. Ecological factors shape the distribution of females, and males go where females are (Kappeler 2000). Social factors are generally thought to be more important than ecological factors in determining the distribution of males because instrasexual competition over access to females contributes more to variation in male reproductive success than intrasexual competition over access to food. When females live in groups, they become a defensible resource for males. The more females there are, the more difficult it is for a single male to monopolize them. This difficulty is compounded if females are widely dispersed or if females become receptive simultaneously. Comparative analyses indicate that the number of males in primate groups is generally correlated with the number of females present and with the extent of synchrony in female receptivity (Mitani et al. 1996, Nunn 1999).

The availability of females explains some, but not all of the variation in the number of males in social groups. This raises the possibility that the distribution of males is also influenced by other factors, including ecological pressures. Comparative analyses suggest that the distribution of males in social groups is directly linked to predation pressure. Hill and Lee (1998) conducted a systematic survey of the Old World monkeys in an effort to assess the effects of predation risk on the distribution of males. They found that predation risk consistently influenced group size, as groups that faced the highest risks of predator attacks lived in the largest groups. In single-male groups, high predation risk was associated with an increase in the number of females, but in multimale groups predation risk influenced the number of males in social groups, but not the number of females. Thus, the distribution of males may be a joint product of the distribution of females and the intensity of predation pressures.

2. Beyond Socioecology

Socioecological models of primate social organization gained considerable currency in the 1990s. However, in the last few years, socioecological models of social organization have been subjected to a new set of empirical and theoretical challenges (Janson 2000). Not all the observed variation in social organization and behavior fit predictions derived from socioecological models. For example the comparison of squirrel monkeys by Mitchell et al. was complicated by subsequent observations of squirrel monkeys in Suriname (Boinski 1999). There, monkeys face intense WGC, but females have weak bonds with one another and do not form coalitions with other females. Comparative evidence indicates that males generally go where females are, but it has been surprisingly difficult to explain why individual males move from one group to another. Male baboons do not always enhance their access to females when they emigrate (Altmann 2000).

The role of environment in shaping social organization was challenged by new comparative studies which took phylogeny explicitly into account (di Fiore and Rendall 1994). Selected pairwise comparisons of closely related primate species provide support for socioecological models, but these comparisons may not be representative. Most of the characteristics associated with WGC, including linear dominance hierarchies, nepotistic alliances, and female philopatry are found in almost all of the extant Cerocipthecine species, even though they now occupy an extremely diverse range of habitats. Di Fiore and Rendall's analysis indicates that there is considerable inertia in social evolution. The social systems categorized by socioecologists may represent different peaks in the adaptive landscape (Pope 2000). If so, this seriously constrains the possible paths for change in social evolution.

Socioecological models assume that ecological factors play a primary role in shaping social organization, but a strong case can be made that demographic and social factors may also shape social systems. A

growing body of evidence suggests that the risk of infanticide may affect primate social systems (van Schaik 2000). There is compelling evidence that infanticide is an evolved reproductive strategy in many animal taxa. In many primate species, males selectively kill infants whose deaths enhance their own reproductive prospects (van Schaik and Janson 2000). The presence of multiple males in social groups throughout the year may be linked to infanticidal risks faced by females (van Shaik 2000).

In summary, present evidence suggests that a broad range of ecological, historical, demographic, and social factors influence the evolution of primate groups. It was probably naïve to expect simple models to explain the full range of variation in primate social organization and behavior. It is, however, essential to begin with simple models before constructing more elaborate ones. The task now is to elaborate and extend socioecological models to accommodate new data and ideas about why primates live in groups.

See also: Communication and Social Psychology; Evolutionary Social Psychology; Group Processes, Social Psychology of; Primate Behavior: Significance for Understanding Humans; Primate Families, Structure of; Primates, Evolution of; Social Psychology; Social Psychology, Theories of

Bibliography

Altmann J 2000 Models of outcome and process: Predicting the number of males in primate groups. In: Kappeler P M (ed.) *Primate Males*. Cambridge University Press, Cambridge, UK

Barton R A, Byrne R W, Whiten A 1996 Ecology, feeding competition, and social structure in baboons. *Behavioral Ecology and Sociobiology* **38**: 321–9

Boinski S 1999 The social organizations of squirrel monkeys: Implications for ecological models of social evolution. *Evolutionary Anthropology: Issues, News, and Reviews* **9**: 101–12

di Fiore A, Rendall D 1994 Evolution of social organization: A reappraisal for primates by using phylogenetic methods. *Proceedings of the National Academy of Science USA* **91**: 9941–95

Hamilton W D 1964 The genetical evolution of social behaviour I, II. *Journal of Theoretical Biology* **7**: 1–52

Hill R A, Lee P C 1998 Predation risk as an influence on group size in cercopithecoid primates: implications for social structure. *J. Zool. Lond.* **245** (part 4): 447–56

Isbell L A, Pruetz J D, Lewis M, Young T P 1999 Rank differences in ecological behavior: A comparative study of patas monkeys (*Erythrocebus patas*) and vervets (*Cercopithecus aethiops*). *International Journal of Primatology* **20**: 257–72

Janson C H 1992 Evolutionary ecology of primate social structure. In: Smith E A, Winterhalder B (eds.) *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, New York, NY

Janson C H 2000 Primate socioecology: The end of a golden era. *Evolutionary Anthropology: Issues, News, and Reviews* **9**: 73–86

Kappeler P M 2000 Primate males: history and theory. In: Kappeler P M (ed.) *Primate Males*. Cambridge University Press, Cambridge, UK

Koenig A, Beise J, Chalise M K, Ganzhorn J U 1998 When females should contest for food—testing hypotheses about resource density, distribution, size, and quality with Hanuman langurs (*Presbytis entellus*). *Behavioral Ecology and Sociobiology* **42**: 225–237

Krebs J R, Davies N B 1993 *An Introduction to Behavioural Ecology*, 3rd edn. Blackwell, Oxford, UK

Mitchell C L, Boinski S, van Schaik C P 1991 Competitive regimes and female bonding in two species of squirrel monkeys (*Saimiri oerstedi* and *S. sciureus*). *Behavioral Ecology and Sociobiology* **28**: 55–60

Mitani J C, Gros-Louis J, Manson J H 1996 Number of males in primate groups: Comparative tests of competing hypotheses. *American Journal of Primatology* **38**: 315–32

Nunn C L 1999 The number of males in primate social groups: A comparative test of the socioecological model. *Behavioral Ecology and Sociobiology* **46**: 1–13

Pope T R 2000 The evolution of male philopatry in neotropical monkeys. In: Kappeler P M (ed.) *Primate Males*. Cambridge University Press, Cambridge, UK

Sterck E H M, Watts D P, van Schaik C P 1997 The evolution of female social relationships in nonhuman primates. *Behavioral Ecology and Sociobiology* **41**: 291–309

van Schaik C P 1983 Why are diurnal primates living in groups? *Behaviour* **87**: 120–44

van Schaik C P 1989 The ecology of social relationships amongst female primates. In: Standon V, Foley R A (eds.) *Comparative Socioecology: the Behavioural Ecology of Humans and Other Mammals*. Blackwell, Oxford, UK

van Schaik C P 2000 Social counterstrategies against infanticide by males in primates and other mammals. In: Kappeler P M (ed.) *Primate Males*. Cambridge University Press, Cambridge, UK

van Schaik C P, Janson C H (eds.) 2000 *Infanticide by Males and its Implications*. Cambridge University Press, Cambridge, UK

Wrangham R W 1980 An ecological model of female-bonded primate groups. *Behaviour* **75**: 262–300

J. B. Silk

Primates, Evolution of

The Primates, one of 20 extant orders of placental mammals, are of special interest because they include the human species (*Homo sapiens*). It should be noted at once that tree shrews, included in the Primates for most of the last hundred years, are now generally thought to have only a distant connection (if at all) and are hence allocated to their own order Scandentia. A recent review (Rowe 1996) portrayed 238 living primate species (excluding tree shrews) and the number is now some 250 and rising, partly reflecting the discovery of entirely new species but also following subdivision of existing species because of new information, notably molecular evidence. This is a dramatic increase over the 150 species listed in a leading

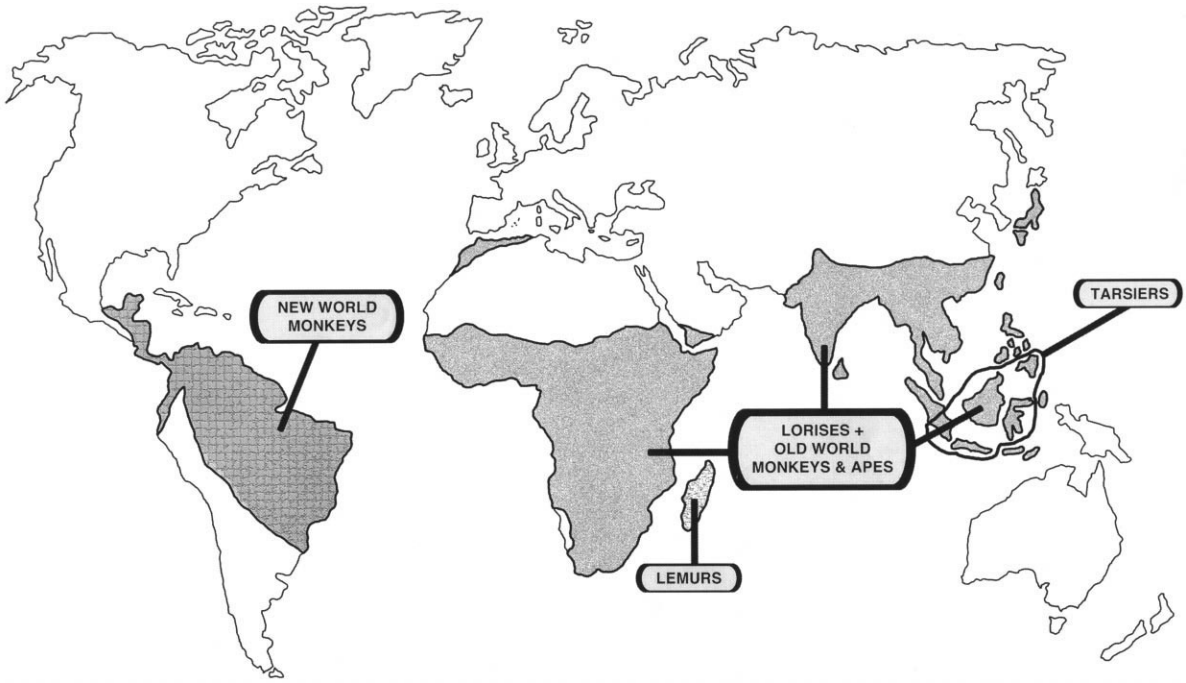

Figure 1
Geographical distribution of living nonhuman primates (excluding tree shrews). Shaded areas indicate the approximate ranges of the five natural groups, with the exception of the loris group (which occupies most but not all of the range occupied by Old World monkeys and apes)

handbook in the 1960s (Napier and Napier 1967). The order Primates also includes almost 450 fossil species (also progressively increasing), documenting the evolution of this group.

The evolutionary history of primates is a striking example of adaptive radiation, in which descendants from an original common ancestor with a particular set of characteristics diverged and adapted to occupy different geographical areas and ecological niches. Following Darwin's *Origin of Species*, this evolutionary radiation was reconstructed initially almost entirely from morphological (anatomical) features, but the comparative framework has been expanded gradually to include physiological, behavioral, ecological, chromosomal, and molecular characters.

1. *Living Primates*

The logical starting point for reconstruction of primate evolution is the array of living species, partly because recognition of the group arose from the study of living representatives, and partly because many features (notably chromosomal and molecular characters) are not fossilized. Known fossil relatives of living primates can be included subsequently in the overall framework, providing the benefit of a geological timescale. Later inclusion of fossil species in a preliminary phylogenetic tree based on living forms also permits use of palaeontological evidence to test its validity.

Living nonhuman primates can be divided into five 'natural groups', each with a restricted geographical range (Fig. 1). As geographical isolation plays a fundamental part in evolutionary divergence, these groups are inherently likely to reflect basic subdivisions within the primate tree. Indeed, they are recognizable immediately in all widely accepted phylogenetic reconstructions and classifications. The first such group contains the Madagascar lemurs (infraorder Lemuriformes). These are the only primates present on that island, and document a substantial local radiation (Tattersall 1982). There are at least 35 living species and the modern array also includes at least 17 subfossil species that died out over the last few millennia, following the first human inhabitation of Madagascar. The second group contains the lorises and bushbabies (infraorder Lorisiformes). They are distributed widely in Africa, South Asia and Southeast Asia, but the total number of species (about 20) is smaller than for lemurs, representing only a limited radiation. This applies even more drastically to the third group, containing the five closely related tarsier species of the South east Asian archipelago (infraorder Tarsiiformes). The fourth group contains the

monkeys of South and Central America (infraorder Platyrrhini), which are the only New World primates. This is a relatively large group of some 80 species, divided into claw-bearing marmosets and tamarins (family Callitrichidae), and true monkeys (family Cebidae), again documenting a major local radiation (Kinzey 1997). The fifth group, containing the Old World monkeys and apes of Africa, South Asia and South east Asia (infraorder Catarrhini) is the largest, with approximately 105 species. In fact, humans also belong to this group, but have expanded far beyond this original geological range. Within this fifth group, which is most directly relevant to human evolution, there is a clear division between Old World monkeys (superfamily Cercopithecoidea) and apes and humans (superfamily Hominoidea), despite extensive overlap in geographical distribution.

The first three natural groups of primates (lemurs, lorises and bushbabies, tarsiers) have remained relatively primitive and are collectively labeled prosimians (literally 'before the monkeys'). They are comparatively small (modal body weight approximately 500 g) and most species are nocturnal. Diurnal species (certain relatively large-bodied members of the families Lemuridae and Indridae) occur only among the lemurs. In fact, some lemurids (species of *Eulemur* and *Hapalemur*) show a highly unusual cathemeral pattern (Tattersall 1988), involving a mixture of nocturnal and diurnal activity. This is just one example of the great diversity of special adaptations found among the lemurs of Madagascar.

By contrast, members of the fourth and fifth natural groups of primates (monkeys, apes, and humans) share a distinctive set of advanced features in their dentitions, jaws, brain, and reproductive system, and can be labeled collectively simians. They are comparatively large (modal body weight approximately 5 kg), and virtually all species are diurnal. The only nocturnal simians are the New World owl monkeys (*Aotus*). The tenfold difference in modal body size between simians and prosimians, undoubtedly connected with the general distinction between nocturnal and diurnal habits, is just one indication of the pervasive importance of body size in primate biology. Availability of a compilation of body weights for almost all primate species, very largely derived from wild animals, therefore deserves special mention (Smith and Jungers 1997). Overall, living primates cover a 300-fold range of body sizes, extending from 40 g for the smallest of the mouse lemur species (*Microcebus*) to 120 kg for an adult male gorilla.

2. *Definition of Living Primates*

It is often stated that living primates represent a graded series, ranging from relatively primitive to very advanced, and that it is therefore difficult to demarcate them from other orders of mammals (notably Insectivora). However, this view was influenced heavily by the inclusion of tree shrews in the order Primates. Reexamination excluding tree shrews revealed that there is in fact a substantial set of characteristics shared by all living primates, which separates them fairly clearly from other placental mammals, leading to the following definition (Martin 1986):

> Primates are typically arboreal inhabitants of tropical and subtropical forest ecosystems. Their extremities are essentially adapted for prehension, rather than grappling of arboreal supports. A widely divergent hallux (big toe) provides the basis for a powerful grasping action of the foot in all genera except *Homo*, while the hand usually exhibits at least some prehensile capacity. The digits typically bear flat nails rather than bilaterally compressed claws; the hallux always bears a nail. The ventral surfaces of the extremities bear tactile pads with cutaneous ridges (dermatoglyphs) that reduce slippage on arboreal supports and provide for enhanced tactile sensitivity in association with dermal Meissner's corpuscles. Locomotion is hindlimb-dominated, with the center of gravity of the body located closer to the hindlimbs, such that the typical walking gait follows a diagonal sequence (forefoot precedes hindfoot on each side). The foot is typically adapted for tarsi-fulcrimation, with at least some degree of relative elongation of the distal segment of the calcaneus, commonly resulting in reverse alternation of the tarsus (calcaneo-navicular articulation).
>
> The visual sense is greatly emphasized. The eyes are relatively large and the orbits possess (at least) a postorbital bar. Forward rotation of the eyes ensures a large degree of binocular overlap. Ipsilateral and contralateral retinofugal fibres are approximately balanced in numbers on each side of the brain and organised in such a way that the contralateral half of the visual field is represented. Enlargement and medial approximation of the orbits is typically associated with ethmoid exposure in the medial orbital wall (though there are several exceptions among the lemurs). The ventral floor of the well-developed auditory bulla is formed predominantly by the petrosal. The olfactory system is unspecialized in most nocturnal forms and reduced in diurnal forms. Partly because of the increased emphasis on vision, the brain is typically moderately enlarged, relative to body size, in comparison to other living mammals. The brain of living primates always possesses a true Sylvian sulcus (confluent with the rhinal sulcus) and a triradiate calcarine sulcus. Primates are unique among living mammals in that the brain constitutes a significantly larger proportion of fetal body weight at all stages of gestation.
>
> Male primates are characterised by permanent precocial descent of the testes into a postpenial scrotum; female primates are characterised by the absence of a urogenital sinus (i.e. urinary and reproductive tracts entirely separate). In all primates, involvement of the yolk-sac in placentation is suppressed, at least during the latter half of gestation. Primates have long gestation periods, relative to maternal body size, and produce small litters of precocial neonates. Fetal growth and postnatal growth are characteristically slow in relation to maternal size. Sexual maturity is attained late and life-spans are correspondingly long relative to body size. Primates are, in short, adapted for slow reproductive turnover.
>
> The dental formula exhibits a maximum of 2.1.3.3/2.1.3.3. The size of the premaxilla is very limited, in association with the reduced number of incisors, which are arranged more

transversely than longitudinally. The cheek teeth are typically relatively unspecialized, though cusps are generally low and rounded and the lower molars possess raised, enlarged talonids.

In fact, an additional distinctive feature should be added to this definition (Sussman 1999): all living primate species live in fairly elaborate social networks. Although nocturnal primate species have commonly been labeled 'solitary,' numerous field studies have shown that they have well established patterns of social interactions based on shared ranges, communal nesting, and regular encounters during the night. There are no immediately recognizable social groups of the kind found in all diurnal primates except the orang-utan, but there are nonetheless long-lasting associations between individuals. The pronounced social tendency now recognized for all primates is presumably linked to the high investment in individual offspring (including intensive parental care), and relatively long lifespans that characterize primates.

The distinctive set of characteristics shared by living primates reflects a suite of adaptations developed in their common ancestor. Taken in combination with various primitive features inferred for the ancestral condition, such as relatively small body size and nocturnal habits (Martin 1990), they yield a plausible reconstruction of the earliest primates. These were probably small-bodied, nocturnally active arboreal creatures that were particularly well adapted for exploitation of the fine branch niche, where they fed on a mixed diet of insects and fruit (the latter accounting for certain modifications of their molar teeth). Leaping and running in the fine terminal branches of trees and among lianas would account for their typical nail-bearing, grasping extremities, notably the prehensile foot, and for emphasis of the visual sense, with large, forward facing eyes and specialization of the brain for highly effective visual orientation. Development of prehensile extremities also involved a shift toward use of the hands rather than the snout in grasping food items, thus explaining shortening of the snout and the associated reduction in number of incisors. Adaptations for low reproductive turnover and high investment in individual offspring can be interpreted as responses to life under the relatively stable and predictable conditions of tropical forests combined with comparatively low predation levels among fine branches.

3. *Phylogenetic Relationships among Living Primates*

In their broad outlines, reconstructions of the phylogenetic tree of primates based on morphological evidence have been confirmed by analyses based on chromosomal or molecular data, such that a general consensus is now emerging (Martin 1990, Fleagle 1999). Although certain relationships in the tree remain controversial, notably in cases where very early divergences are involved, such confirmation derived from largely independent analyses of new data provides valuable support for earlier phylogenetic reconstructions based exclusively on morphological evidence.

The substantial body of evidence now available indicates that the five natural groups of living primates (lemurs; lorises and bushbabies; tarsiers; New World monkeys; Old World monkeys, apes, and humans) are, in fact, phylogenetically distinct subunits within the tree (Fig. 2). In other words, each of the five natural groups is monophyletic (derived from a specific common ancestor within the tree). Although certain cranial features led some authors to suggest that the mouse and dwarf lemurs (family Cheirogaleidae) are more closely related to lorises and bushbabies than to other lemurs, chromosomal and molecular evidence indicates overwhelmingly that the lemurs are monophyletic (Yoder 1997). Furthermore, it has consistently emerged that the Old World monkeys, and the apes and humans, respectively, belong to two separate monophyletic groups at a lower level. The overall pattern of relationships between the major groups of primates has also become relatively clear. On the one hand, it has emerged consistently that the Lemuriformes and the Lorisiformes are sister groups (that is, their individual common ancestors were derived from an earlier common ancestor). Together, the lemurs, lorises, and bushbabies constitute a larger-scale monophyletic group known as strepsirrhines, because of their common retention of a rhinarium (a naked, moist area of skin surrounding the nostrils in the primitive mammalian condition). On the other hand, it has emerged equally consistently that the simian primates (monkeys, apes, and humans) together constitute a second larger-scale monophyletic group. There is also a considerable body of evidence indicating that tarsiers and simians are sister groups, together constituting a high-level monophyletic group known as haplorrhines, sharing complete suppression of the rhinarium (unique among mammals). Although some investigators have questioned an ancestral link between tarsiers and the simians, an overall review of evidence (Martin 1990) indicated that haplorrhines do, indeed, constitute a monophyletic group (Fig. 2).

4. *The Primate Fossil Record*

The known record of reliably identified fossil primates dates back to the beginning of the Eocene epoch, about 55 mya. An assemblage of earlier fossil forms mainly from the Palaeocene epoch (65–55 mya), collectively labeled 'archaic primates' (Plesiadapiformes), has traditionally been allocated to the order Primates.

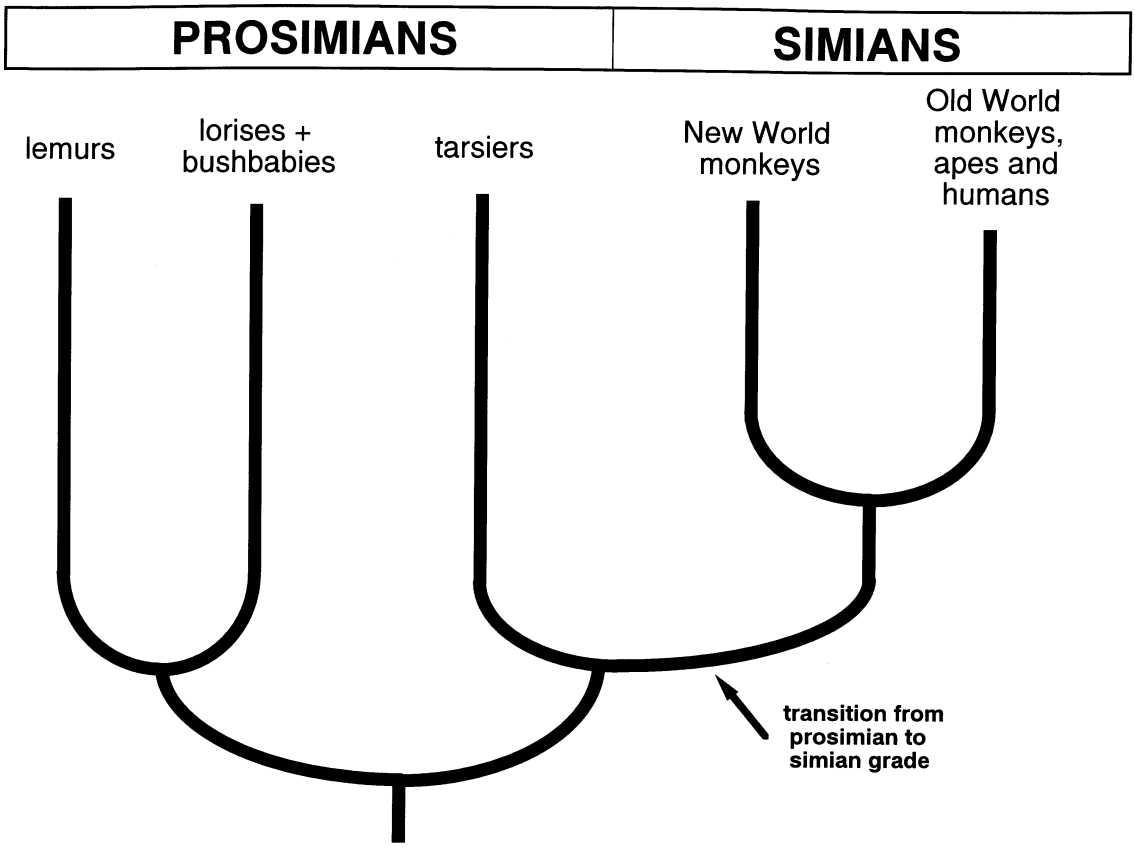

Figure 2
Outline phylogenetic tree of living nonhuman primates (excluding tree shrews) showing the likely pattern of relationships among the five monophyletic natural groups. Traditional classifications distinguish between the generally more primitive prosimians and the simians, which are characterized by a set of more advanced features

However, these early fossil forms (some of which survived into the Eocene) show few, if any, of the features in the definition given above and recent evidence indicates that their connection with primates is questionable. Even if Plesiadapiformes did have some early connection with primates, they surely diverged at such an early stage that they are of little relevance to the later adaptive radiation leading to modern primates and their direct fossil relatives ('primates of modern aspect').

The currently documented primate fossil record, excluding Plesiadapiformes, is divided fairly sharply into two segments separated by a gap of at least 6 million years in the middle of the Oligocene epoch (35–25 mya). Approximately equal numbers of fossil primate species (about 225) are known from each half of the record. Initially, primates of the Eocene epoch (55–35 mya) were documented only by early prosimians from Europe and North America. On both continents, they can be divided fairly clearly into two families, generally larger lemur-like Adapidae (e.g., *Adapis* in Europe, *Notharctus* in America), and generally smaller tarsier-like Omomyidae (e.g., *Necrolemur* in Europe, *Shoshonius* in America), and many authors have linked them to modern strepsirrhines and haplorrhines, respectively. Almost the entire skeleton is known for *Notharctus*; it is very lemur-like and certainly demonstrates that many defining features of modern primates were present at that stage.

This initial picture from America and Europe has been modified radically by subsequent fossil finds from North Africa, Asia and South east Asia. In North Africa, fossils from the latest Eocene and early Oligocene, notably from the Egyptian Fayum site, have documented the presence of a variety of early simians (e.g., *Aegyptopithecus* and *Apidium*) along with some potential relatives of modern strepsirrhines and tarsiers. More recently, Chinese middle Eocene fossil deposits have yielded evidence of early simians (*Eosimias*) and a putative direct ancestor of modern tarsiers (*Tarsius eoceanus*), along with adapids and omomyids. Furthermore, late Eocene sites in Myan-

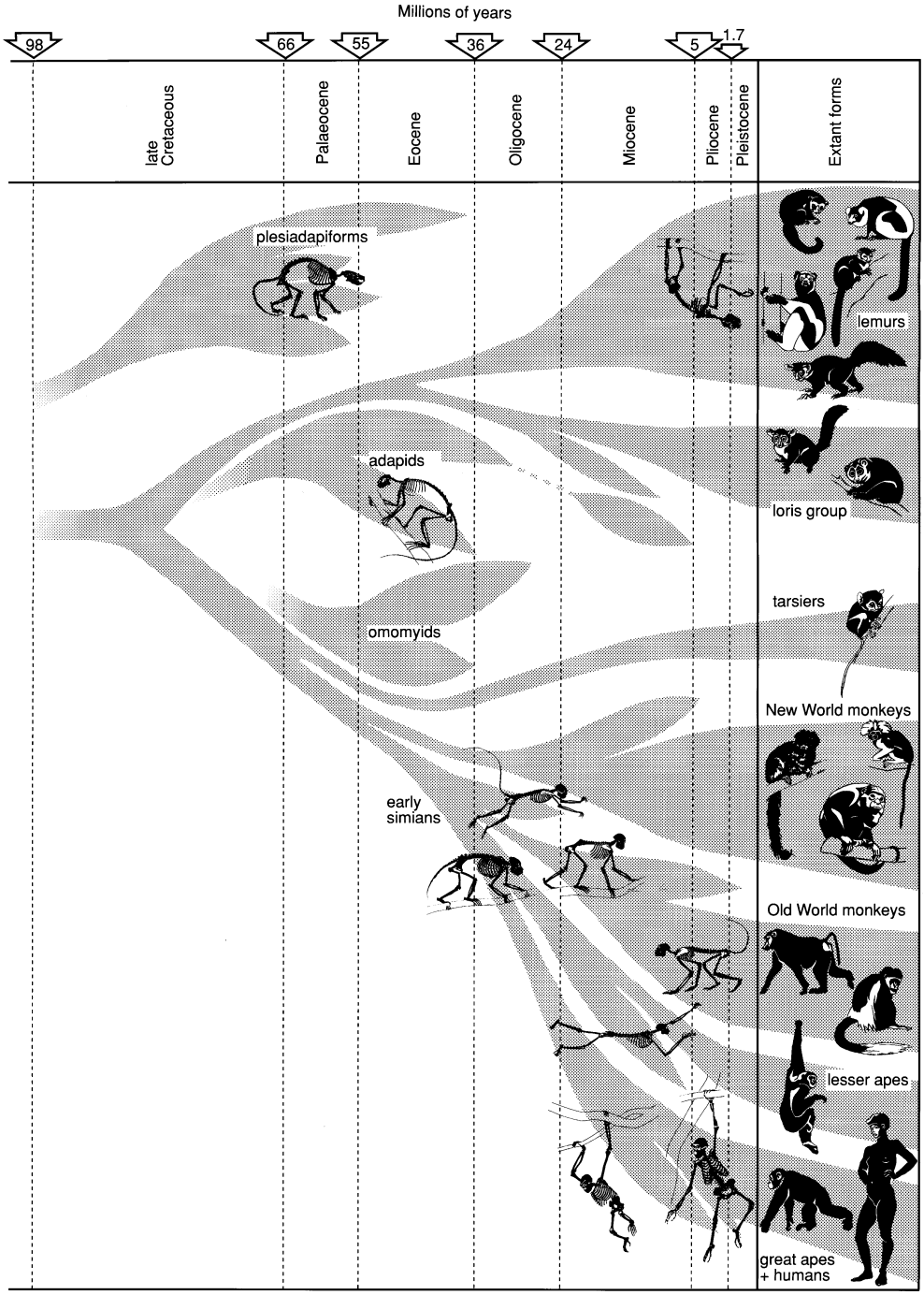

Figure 3
Overall phylogenetic tree of primates (excluding tree shrews) showing the likely pattern of relationships among the five monophyletic natural groups of living primates. This tree also includes major groups of fossil representatives (excluding plesiadapiforms). Note that the initial date of divergence for 'primates of modern aspect' is set at about 80 mya. Original illustration by Lukrezia Bieler-Beerli. (Reproduced by permission of *Nature* (www.nature.com) (Martin R D **363**: 223–34) © (1993) Macmillan Magazines Ltd.)

mar and Thailand have yielded several fossil simians (e.g., *Amphipithecus* and *Siamopithecus*). Thus, both prosimians and simians were present in the Eocene.

The situation is entirely different in the second half of the primate fossil record, extending from the late Oligocene to the present day. Direct fossil relatives of New World monkeys are known from South America, beginning in the latest Oligocene (*Branisella*) and extending into the Miocene (e.g., *Homunculus*). In the Old World (notably Africa), early Miocene deposits have yielded fossil relatives of bushbabies and lorises (e.g., *Progalago*), of tarsiers (albeit just one molar tooth from Thailand), and of monkeys and apes (e.g., *Victoriapithecus* and *Proconsul*). Thus, four of the five modern natural groups of primates can be traced back to about 25 mya, and one (tarsiers) can be tentatively traced back to 40 mya. Nevertheless, there are also notable gaps in the known fossil record. The most striking gap is in Madagascar, where no single fossil from the major radiation of lemurs is known prior to the recently extinct subfossils. Similarly, in the New World, no single fossil form can be linked reliably to the marmosets and tamarins, a substantial proportion of the modern species.

Gaps in the fossil record influence directly the inferred date for the origin of primates. The standard practice has been to assume that primates diverged during the Palaeocene epoch, perhaps 60–65 mya, just a few million years prior to the first known fossil primates. However, direct inference from the known fossil record will lead to serious underestimation if there are substantial gaps in the fossil record, as is true of primates (Martin 1993). Several recent studies that calibrated molecular phylogenies for mammals using various dates for fossils of well documented groups outside the primates have all concluded that primates diverged from other mammals about 90 mya. Given some allowance for the elapse of time between this initial divergence and initiation of the adaptive radiation leading to modern primates, their common ancestor can be dated at about 80 mya (Fig. 3).

See also: Human Cognition, Evolution of; Primate Socioecology; Primates, Social Behavior of

Bibliography

Fleagle J G 1999 *Primate Adaptation and Evolution*, 2nd edn. Academic Press, New York

Kinzey W G (ed.) 1997 *New World Primates: Ecology, Evolution, and Behavior*. Aldine de Gruyter, New York

Martin R D 1986 Primates: A definition. In: Wood B A, Martin L B, Andrews P (eds.) *Major Topics in Primate and Human Evolution*. Cambridge University Press, Cambridge, UK, pp. 1–31

Martin R D 1990 *Primate Origins and Evolution: A Phylogenetic Reconstruction*. Princeton University Press, Princeton, NJ

Martin R D 1993 Primate origins: Plugging the gaps. *Nature, London* **363**: 223–34

Napier J R, Napier P H 1967 *A Handbook of Living Primates*. Academic Press, London

Rowe N 1996 *The Pictorial Guide to the Living Primates*. Pogonias Press, East Hampton, New York

Smith R J, Jungers W L 1997 Body mass in comparative primatology. *Journal of Human Evolution* **32**: 523–59

Sussman R W 1999 *Primate Ecology and Social Structure. Vol. 1: Lorises, Lemurs and Tarsiers*. Pearson Custom Publishing, Needham Heights, MA

Szalay F S, Delson E 1979 *Evolutionary History of the Primates*. Academic Press, New York

Tattersall I 1982 *The Primates of Madagascar*. Columbia University Press, New York

Tattersall I 1988 Cathemeral activity in primates: A definition. *Folia Primatologica* **49**: 200–2

Yoder A D 1997 Back to the future: A synthesis of strepsirrhine systematics. *Evolutionary Anthropology* **6**: 11–22

R. D. Martin

Primates, Social Behavior of

The distribution of individuals within a species is related to such factors as the distribution and abundance of resources, predation pressure, and genetically determined behavioral propensities. In some animal species, individuals spend most of their time alone, whereas in others most or all of their time is spent in groups. With one exception, all diurnal primates live in social groups. In many animal taxa temporary aggregations form under certain conditions (e.g., schools of fish, flocks of birds, herds of ungulates) but in most of these cases, aggregations are made up of different individuals each time the animals congregate. One of the most common features of primate groups is the relative constancy of group membership. There are many theories attempting to explain why diurnal primates live in groups and the reasons for different types of groups (see Janson 2000).

1. Definition of Terms

There is a great deal of debate concerning what factors may have led to group living among diurnal primates and the relationships that may exist between primate social structure and ecology. However, before discussing some of these issues, it is necessary to define some terms.

1.1 The Term 'Social Group'

If one searches the primate literature for a definition of the term 'social group,' it is quite difficult to find. It seems that primatologists know a social group when

they see one but are reluctant to define the term generally. In the literature, there are few attempts to formally define the primate 'social group' (e.g., Altmann 1962, Kummer 1967, Struhsaker 1969, Sussman 1999) and few critical evaluations of these definitions. Altmann (1962) defines a group as consisting of conspecific, intercommunicating individuals bounded by frontiers of far less frequent communication. Kummer (1967) considers that a group as a 'spatial aggregation of primates that travel and rest together and at the same time avoid the proximity of other such aggregations is, at closer inspection, also a functional and reproductive unit' (p. 359). These are both broad definitions that emphasize social proximity, social interaction, and synchrony of activity. However, all of these criteria are not equally useful in defining the social group.

Reproduction was once thought to be a major *raison d'etre* of primate groups but a number of factors indicate that it is not a major common denominator of the primate group. For example, in many species breeding is seasonal but the group is stable throughout the year. In fact, in some species, behavior during the breeding season actually disrupts the social group. In the Madagascar sifaka, *Propithecus*, groups break up during the short mating season and reformulate after breeding takes place. Most species of primates have mechanisms that ensure continuous gene flow between groups. Groups exist within a breeding population (or deme) and genetic exchange is common between groups, at higher or lower frequencies, depending on the species.

Struhsaker (1969) offered a more detailed definition of the social group. He included in his definition the following factors: (a) membership in the same social network; (b) the majority (at least 80 percent) of nonaggressive social interactions take place within this social network; (c) membership is temporally stable; (d) group members occupy the same home range; (e) group members have distinct social roles; (f) progressions between members are synchronized; (g) group members exhibit different behavior toward nongroup conspecifics; and (h) social groups are relatively closed to new members.

Using some but not all of the above criteria, the following definition of a 'social group' in primates is proposed: a number of individuals who interact socially more frequently among themselves than with other individuals; group members exhibit different behavior towards nongroup members and occupy the same home range. Group membership is relatively stable over time and group members have distinct social roles but, although these two factors are characteristics of the group, the particulars of group stability and of social roles are extremely variable from species to species and are not necessary components of this definition.

It is also true that most social groups are relatively closed to new membership but this is also highly variable and would be difficult to include in a general definition. In fact, the only social groups that were thought to be closed to new membership are monogamous pairs, and this does not appear to be as true as was previously believed (Fuentes 1998, Strier 2000); individual exchange between groups is a common occurrence in most diurnal primate species. Synchronized progressions, mentioned in the definitions of both Kummer and Strushaker, are not common to all primate groups. A number of groups divide into subgroups for various periods of time and these subgroups do not necessarily synchronize movements or activity while separated. Because of the dynamics of individual social relationships, Dunbar (1988, p. 12) preferred to view primate groups or social systems in terms of their constituent relationships. 'Where the sets of relationships of several individuals coincide and are reciprocated, a formal group emerges as a stable cohesive unit. … These clusterings are sufficiently distinctive in their cohesion in time and space that we can define them as levels of groupings.'

There are many different types of social group found among the diurnal primates. At a very general level these include: monogamous pairs (i.e., one adult male, one adult female, and their immature offspring); multimale groups (groups containing more than one adult male); one-male groups (groups containing only one adult male); and fission–fusion groups (groups that form dispersing and coalescing subgroups). Of course, each of these superficial categories does not include groups of identical social structure or social organization. For example, multimale groups of lemurs are quite different from those of baboons in structure, function, evolutionary history, and ecological advantage. This is even more obvious when we compare species with 'fission–fusion' group structures.

1.2 The Terms 'Social Structure' and 'Social Organization'

The terms social structure and social organization have now been introduced. These terms are often used interchangeably; however, it is useful to distinguish between them. Following definitions similar to those proposed by Rowell (1972), the social structure of a species can best be divided into two components: group structure and population structure. Group structure is the size and composition (as to age and sex) of a primate group. Population structure is the demographic character of the population, population density, and the density and dispersal of groups within the breeding population or deme. One measures social structure using censusing techniques. Social organization is the pattern of social interactions that occurs between individuals within and between groups: it is a description of social behavior.

Thus, social structure refers to the components of the social system, whereas social organization refers to

the patterns of interaction between the components. Both patas monkeys and Hamadryas baboons live in one-male groups that can have essentially similar age and sexual composition. However, in Hamadryas one-male groups, the male is the center of attention of the group and directs group travel. In the patas monkey, the male takes a peripheral position in the group and the females initiate and direct group travel. Thus the patas monkey and the Hamadryas baboon have groups with similar structure but the social organization of the two species is very different. Furthermore, differences in degrees of tolerance between adult males result in very different distribution patterns of one-male groups. Hamadryas one-male groups often sleep and move together, whereas patas males are intolerant of each other and one-male groups are usually widely dispersed. Although the group structure is similar, the structure of the populations of these two species is radically different.

1.3 Spacing Between and Within Groups

When dealing with animals living in groups, two levels of social spacing must be considered: between groups of the same species and between individuals living within the same group. Each species has developed complex communication patterns for both intergroup and intragroup spacing.

These patterns are part of the social organization of a species and they help to maintain the species-specific social structure. In group-living animals, such as diurnal primates, the terms home range, core area, area of exclusive use, and territory relate to intergroup spacing, whereas dominance hierarchies, roles, kinship lines, and peer groups are usually pertinent to spacing patterns within a group.

In the above definition of a 'social group,' group members were said to occupy the same home range. Home range can be defined as the area over which a group normally travels in pursuit of its routine activities. Other terms such as day range and monthly range are used to denote the area over which the group travels in a given time period. If a group uses a particular part of its home range more frequently, or with greater regularity, than other parts, this zone is referred to as a core area. The core area often contains many of the major resources of the group. In some species of primates, the boundaries of different groups overlap only slightly, whereas in others home ranges overlap a great deal and each group may have a very small area that is not shared by other conspecific groups (referred to as an area of exclusive use). In a number of primate species, home ranges of groups may overlap completely.

Mechanisms used for intergroup spacing also vary between species. In some species, spacing between groups involves the maintenance of exclusive use of a particular area, but in others it seems to involve the maintenance of intragroup integrity and a patterned use, by two or more groups, of the same space. However, a unitary definition of territoriality and a simple dichotomy between territorial and nonterritorial species might do more to confuse the issue than to aid in an understanding of ecological parameters of spatial relationships between individuals and groups of conspecifics (see Sussman 1999). For example, active, ritualized, agonistic defense behaviors are used by many species of primates to maintain the integrity of home range boundaries or areas of exclusive use. However, in other species, spatial integrity is maintained between groups by exchange of vocalizations or with no apparent outward behavioral spacing mechanisms. In fact, very few species of primates could be considered territorial in the strict sense of the term.

A large number of species that do not maintain exclusive areas employ ritualized postures and gestures, and/or vocalizations, to maintain group integrity and to regulate spatial relationships within a shared home range. Some of these species were thought to be highly territorial until recent studies have shown this not to be the case. Thus, there appears to be two separate phenomena: (a) the behavioral mechanisms that control spacing and (b) the degree of isolation or overlap between groups. The use of the term territoriality often confounds these two factors and a more clear definition might be the active defense of individual or group home range boundaries by actual or ritualized agonistic encounters, thereby maintaining essentially exclusive use of the home range. This, then, describes a very specific set of behaviors that control spacing and a particular pattern of spacing. However, even using this strict definition may conceal important variation in conditions of spacing between two species labeled as territorial.

2. Natural Selection and Primate Behavior

An important question relating to group-living among diurnal primates is why, in fact, do they live in groups? What advantage does this habit confer on individuals? The theory of natural selection predicts that evolution will act to increase the frequency of traits that improve an individual's ability to survive and to reproduce more successfully than other members of its species. Traits that confer survival and reproductive advantages upon an individual are referred to as adaptations. There were, however, a number of characteristics that Darwin found difficult to explain when he developed the theory of natural selection. These included morphological and behavioral differences between the sexes that occur in many species, such as larger canines or body size, or greater aggressiveness in males than in females. In many cases it was not obvious how these secondary sexual characteristics gave any direct advantage to individuals in their attempt to obtain resources.

To explain these traits, Darwin formulated the theory of sexual selection. He proposed that many morphological and behavioral differences between the sexes evolved because they contributed directly to an individual's reproductive success, either by improving his or her competitive ability against members of the same sex, or by increasing attractiveness to the opposite sex. Some of the traits Darwin believed to be sexually selected were subsequently shown to proffer direct adaptive advantage to individuals. However, sexual selection is not an important mechanism in evolution.

A basic assumption of sexual selection theory is that there is a direct relationship between particular traits and genetics. If dominance in males is directly related to increased fitness, then males must be able to pass on this trait to their offspring. However, dominance is complex and involves a number of morphological and behavioral characteristics as well as a social and ecological context. This is true of many other traits proposed to be sexually selected. It is often extremely difficult, if not impossible, to explain the genetics of these complex traits.

Another phenomenon scientists found difficult to explain was cooperative behavior. Why would animals cooperate if this did not increase their evolutionary fitness? If an animal avoids being eaten by a predator by joining forces with others, cooperation easily fits within the rubric of individual selection. However, there are many cooperative behaviors, especially among group-living animals, that appear to be of no benefit or are even dangerous to the individual. To explain these behaviors, the concepts of kin selection and reciprocal altruism were formulated.

Darwin was unaware of Mendelian genetics and thus genetic inheritance did not figure in his explanation of natural selection. Hamilton (1964) examined the potential genetic consequences of cooperative behavior and developed the theory of kin selection to explain altruistic behavior between related individuals. An individual's traits, and the genes responsible for them, are more likely to be shared by close relatives than by unrelated individuals. Thus, by sacrificing oneself for a number of one's kin, an individual may be ensuring the survival of genetically inherited traits. Since more closely related individuals are more likely to share copies of the same genes, altruism is expected to be selectively directed toward kin, and more close kin will accrue more costly altruism (Silk 1987).

However, many altruistic acts are performed by unrelated individuals. Trivers (1971) formulated the theory of reciprocal altruism to explain cooperation and assistance in these cases. He believed that altruistic acts would only occur between unrelated individuals if they were likely to reciprocate in the future and had the opportunity to do so, if the benefits to the recipient were greater than the costs to the altruist, and if there was little chance of cheating. Thus, both parties improve their individual fitness in the long term. In practice both types of altruistic interactions are often difficult to identify, and predictions derived from these theories are difficult to evaluate empirically (see Silk 1987 for an excellent review).

These concepts formed the basis of the subdiscipline of Sociobiology (Wilson 1975), which was developed to explain the influences of biology on social behavior. Sociobiology was very controversial when first formulated because of the questionable genetic nature of many of the behaviors it attempted to explain. Were behaviors such as aggressiveness or dominance inherited or the result of a complex interaction between genetics and environment? However, many of the theories currently in vogue in socioecology are also based on sexual selection, kin selection, and reciprocal altruism. These include theories attempting to interpret variation in social structure, social organization, and social behavior in group-living primates, and the relationships between ecology and social systems. For example, one of the major theories attempting to relate individual spacing to that of resource distribution is based on sexual selection. In most vertebrates, females must invest a great deal of time and energy in successfully reproducing offspring, whereas males invest a small amount. Because of this, the distribution of females in the environment is presumed to be more closely related to resources than is that of males. The distribution of males is related to the dynamics of male competition for females (Wrangham 1987). Thus, females compete for energy resources and males compete for females, and the social system of a primate species can be understood by studying these relationships. In fact, this theory was proposed when it was believed that, in most primates, females remained in the same home range throughout their lives and males dispersed. Now researchers have shown that, in the majority of primate taxa, either females or both sexes disperse (Strier 2000).

To date it still is not possible to relate patterns of resource distribution to patterns of social spacing among primates. Very little is known about the factors that underly the enormous amount of variability in primate social systems. These theories so far have not led to a better understanding of the relationships that might exist between ecology and social structure.

3. *Ecology and Social Behavior Among Primates*

Many of the theories currently attempting to explain the relationship between ecology and social structure among primates are one-dimensional (e.g., distribution of resources, predation pressure, infanticide avoidance). However, there are probably a multitude of factors interacting to cause the great variety of social groups among these animals.

We would expect that there will be some relationship between the spacing of resources in the environment and the spacing of those individuals using the re-

sources. However, to explain the adaptive significance of specific differences between the types of social structure and organization found among primates, diachronic as well as sychronic factors must be considered.

A complete understanding of the relationship between ecology and social behavior among primates will most likely have to take into account at least the following factors: (a) the resources and periodicity of these resources within the total geographical range of the species, as well as those in any particular area; (b) the selectivity of the species—those resources chosen by the species vs. those available; (c) the evolutionary history of the species leading to its current phylogenetic (genetic) propensities (these should be related to the present and past geographic distribution of the species and the environments and resources which exist(ed) within these ranges); (d) the history of the particular population and group, since the conditions in a given area may select for population-specific or group-specific adaptations. This latter factor should take into account stochastic and demographic phenomena which can have a major influence on a population at any particular time (Altmann and Altmann 1979, Chan 1992, Rowell 1993).

See also: Communication and Social Psychology; Evolutionary Social Psychology; Group Processes, Social Psychology of; Primate Behavior: Significance for Understanding Humans; Primate Families, Structure of; Primate Socioecology; Primates, Evolution of; Social Psychology; Social Psychology, Theories of

Bibliography

Altmann S A 1962 A field study of the sociobiology of Rhesus monkeys, *Macaca mulatta*. *Annals of the New York Academy of Sciences* **102**: 338–435

Altmann S A, Altmann J 1979 Demographic constraints on behavior and social organization. In: Bernstein I S, Smith E O (eds.) *Primate Ecology and Human Origins*. Garland, New York

Chan L K 1992 Problems with socioecological explanations of primate social diversity. In: Burton F D (ed.) *Social Processes and Mental Abilities in Non-human Primates*. Edwin Mellen Press, Lewiston

Dunbar R I M 1988 *Primate Social Systems*. Comstock, Ithaca, NY

Fuentes A 1998 Re-evaluating primate monogamy. *American Anthropology* **100**: 890–907

Hamilton W D 1964 The genetical evolution of social behavior. *American Nature* **97**: 354–6

Janson C H 2000 Primate socio-ecology: The end of a golden age. *Evolutionary Anthropology* **9**: 73–86

Kummer H 1967 Dimensions of a comparative biology of primate groups. *America Journal of Physical Anthropology* **27**: 357–66

Rowell T E 1972 *Social Behaviour of Monkeys*. Penguin Books, Harmondsworth, UK

Rowell T E 1993 Reification of primate social systems. *Evolutionary Anthropology* **2**: 135–7

Silk J B 1987 Social behavior in evolutionary perspective. In: Smuts B B, Cheney D L, Seyfarth R M, Wrangham R W, Strushaker T T (eds.) *Primate Societies*. University of Chicago Press, Chicago

Strier K B 2000 *Primate Behavioral Ecology*. Allyn and Bacon, Boston

Strushaker T T 1969 Correlates of ecology and social organization among African cercopithecines. *Folia Primatologica* **11**: 80–118

Sussman R W 1999 *Primate Ecology and Social Structure, Volume 1: Lorises, Lemurs and Tarsiers*. Pearson Custom Publishing, Needham Heights, MA

Trivers R L 1971 The evolution of reciprocal altruism. *Quarterly Review of Biology*. **46**: 35–57

Wilson E O 1975 *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA

Wrangham R W 1987 Evolution of social structure. In: Smuts B B, Cheney D L, Seyfarth R M, Wrangham R W, Strushaker T T (eds.) *Primate Societies*. University of Chicago Press, Chicago

R. W. Sussman

Priming, Cognitive Psychology of

Priming refers to the phenomenon whereby the identification of a word or an object can be enhanced either by the presentation of a related stimulus or by a prior processing episode involving that word or object. In the case of priming with a related stimulus, enhancement is assessed relative to identification performance following an unrelated stimulus or some neutral stimulus. Priming effects are used to support inferences about the time course of component processes that contribute to stimulus identification and about the structure of knowledge representation. Priming induced by an earlier processing episode involving the target is referred to as *repetition priming*. Enhancement due to repetition priming is assessed relative to performance on targets that were not included in any earlier processing episode. Repetition priming effects provide evidence regarding the influence of memory on skilled task performance.

1. Priming with Related Stimuli

The widespread use of related primes to examine word and object identification has its origins both in the classic era of verbal learning and in the more recent study of visual attention.

Experiments on verbal learning tasks such as memorizing word lists showed that prior learning of a target word's associate (e.g., *joy* is an associate of *happy*) improved subsequent learning of that target

(Cofer 1967). These early studies foreshadowed some of the well-established aspects of priming obtained in more recent research (e.g., priming dissipates rapidly over time).

In the visual attention domain, it has been proposed that presentation of a prime stimulus that is of the same type as the target could benefit the processing of the target through two mechanisms (Posner and Snyder 1975). First, the prime automatically can activate the pathway used in processing the target. Second, the observer can be induced consciously to attend to activation produced by the prime, creating an expectancy and amplifying the benefit of the prime but also creating a cost if the prime is unrelated to the target and the expectancy is not met.

1.1 Semantic Priming

Building on the seminal work of Meyer and Schvaneveldt (1971), who provided one of the first demonstrations that a pair of semantically related words (e.g., *nurse–doctor*) was identified faster than a pair of unrelated words, early investigations of semantic priming examined automatic and conscious contributions to priming. These studies were motivated in part by the idea that semantically related concepts could automatically activate one another (Collins and Loftus 1975). This automatic spreading activation was the hypothesized basis for the facilitative effect of presenting a prime prior to identifying a semantically related target word. Experiments demonstrating this effect typically employed a word identification task such as lexical decision, in which word and nonword targets are presented and the observer is required to classify each target as a word or nonword. Accuracy in this task is high, so the aspect of performance that is of greatest interest is the time taken to make a response. The benefit of a related prime is revealed by a shorter response latency relative to when a neutral prime such as *XXX* is used.

By varying the time interval between the onset of the prime stimulus and the subsequent onset of the target (stimulus onset asynchrony, or SOA) it was shown that automatic activation of a target word by a semantically related prime can occur within 250 ms of the onset of the prime (Neely 1977). At an SOA of 400 ms or more, however, not only was there a facilitative priming effect, there also was an interference effect, such that relative to the neutral prime condition, a longer response latency was found when an unrelated prime was used. This pattern of benefit and cost suggested that in the earliest stages of processing a prime word, semantic information related to that prime was automatically activated. Shortly thereafter, attention was directed to the activated information allowing the benefit of activation to be maintained but also leading to a cost if the target was unrelated to the expected domain of semantic knowledge.

1.2 Models of Semantic Priming

The idea that semantic priming may be amplified by conscious attention, perhaps owing to expectations about what related target word might follow a particular prime, is generally accepted. In contrast, however, the proposal that activation spreads automatically from the prime concept to semantically related concepts in semantic memory has been challenged. One type of alternative theory, compound-cue theory, proposes that rather than activation spreading from the prime, the prime stimulus combines with the target to form a compound cue that is matched with the contents of long-term memory (Ratcliff and McKoon 1988). Related prime–target pairs are likely to be represented in memory owing to a prior history of co-occurrence, so this compound cue has a high degree of familiarity and permits a relatively rapid lexical decision. Unrelated prime–target pairs have lower familiarity and more time is needed to make a lexical decision. This theory correctly predicts that presenting an unrelated stimulus between the prime and target eliminates the priming effect.

Whittlesea and Jacoby (1990) extended the compound-cue idea by proposing that processing a target can induce automatic retrieval of a related prime event, producing an integrated or compound prime–target cue. They demonstrated that forming a prime–target compound is situation dependent. When a target is degraded (e.g., by using alternating case: dOcToR), the prime is more likely to form a compound with the target because independent identification of the target is more difficult. The result is a larger priming effect with degraded targets.

A second alternative account of semantic priming derives from neural network models of word identification. In many of these models, knowledge about words is represented by a pattern of activation across a collection of processing units. It is the instantiation of this pattern in the network that supports a response to a word target on tasks such as lexical decision. Related words have similar patterns of activation. Presentation of a related prime moves the network into a pattern similar to that associated with the target, giving the network a head start at instantiating the target's pattern of activation, thereby creating a priming effect. As in the compound-cue theory, presenting an unrelated word between the prime and target is expected to reduce or eliminate the priming effect, although neural network theory attributes this reduction to disruption of the pattern of activation created by the prime (Masson 1995).

Although semantic priming can be disrupted by an intervening unrelated word, prime events appear to modify long-term memory in a manner that supports semantic priming even when a period of minutes intervenes between prime and target. The survival of semantic priming across such time intervals requires that the task performed on the target demand sub-

stantial semantic processing, such as deciding whether the target represents a living or a nonliving object (Becker et al. 1997). A neural network model developed by Becker et al. accounts for this effect by assuming that the prime event modifies connection weights in the network to favor related words, which is equivalent to modifying long-term memory.

Recent work on semantic priming, then, is largely consistent with the idea that short and long timescale priming effects arise from different mechanisms: transient patterns of activation or compound cueing of memory in the former case and incremental learning in the latter.

1.3 Phonological and Other Forms of Priming

The role of *phonological recoding* in the reading of words has been studied by presenting phonologically related primes with target words. Perfetti and Bell (1991) used nonword primes that were phonologically related to target words (*baik–bake*) and control primes that were graphemically but not phonologically related to the targets (*bawk–bake*). In this paradigm, the target word was presented briefly (about 15–30 ms) and was *followed* by the prime, which served as a backward mask and was in view for a similarly brief time. Finally, a pattern mask consisting of a row of *X*s was presented. Subjects were more accurate in their report of the target word when phonologically related primes were used, even though the primes were usually not consciously perceived. Phonological information derived from the prime can thus contribute to the process of identifying a target. Results such as this one, which have been replicated using Hebrew and Chinese materials, suggest that phonological recoding of words is the default processing mode in reading.

Other types of primes bearing various relations to target words, such as orthographic (*cord–card*) or morphological similarity (*swim–swam*), have also been used to determine whether these knowledge sources play a role in reading. Related primes of these types usually improve performance on target identification tasks relative to primes that bear no relation to the target, implicating these sources of knowledge in reading. In addition, priming paradigms have been used to study clinical populations in an effort to understand the knowledge retrieval processes associated with particular disorders. For example, semantic priming of alcohol-related words is enhanced among subjects who suffer from alcohol dependence. Medications have been shown to reduce this priming effect, suggesting that such interventions are capable of reducing the retrieval of alcohol-related memories.

2. Repetition Priming

Stimulus identification can be enhanced not only by priming with a related item but also by an earlier presentation of the target itself. Repetition priming of words or objects has been demonstrated in experiments in which a preliminary presentation of an item is followed minutes, hours, or even days later by a repetition of that same item. Improved processing typically is evidenced either by a reduction in the time taken to identify the item or increased accuracy in its identification. In a particularly striking demonstration of this phenomenon, Kolers (1976) showed that even when over 1 year intervened between the first and second reading of text passages presented in inverted typography, subjects were able to read repeated passages faster than ones they had not read earlier. An important aspect of repetition priming that separates it from many other studies of memory is the fact that subjects are often unaware that their performance is being influenced by prior experience with the target stimulus (see *Implicit Memory, Cognitive Psychology of*).

2.1 Long-term Priming

The fact that repetition priming can persist over relatively long periods of time provides an important theoretical constraint. The proposal that repetition priming results from sustained activation of the mental representation of a target item has been criticized on the grounds that it would be implausible for activation to be sustained over long periods of time. Moreover, repetition priming effects involving novel stimuli such as nonwords and novel shapes indicate that activation of pre-existing mental representations is not a necessary condition for priming.

A more plausible account proposes that a representation of the first encounter with an item is encoded into long-term memory and is recruited, usually without conscious intention, when that item is repeated on a later occasion. To the extent that the processing operations performed on the item are similar on the two occasions, recruitment of memory for the original processing episode will benefit current task performance—a principle known as *transfer-appropriate processing*.

Because many of the tasks used to study repetition priming involve the identification of a stimulus, perceptual operations that support identification have been of particular interest in this work. Consistent with the transfer-appropriate processing principle, it has been shown that altering the perceptual characteristics of an item between its first and second presentations (e.g., spoken to printed presentation, or male to female voice) reduces or eliminates repetition priming.

The perceptual specificity of repetition priming has also been taken as support for the proposal that this form of priming is mediated by a neurologically distinct memory system that represents perceptual forms. This memory system is assumed to be in-

dependent of other memory systems, such as the one responsible for supporting conscious recollection of prior episodes. The independent system assumption receives support from the finding that amnesic subjects who are severely impaired on tasks requiring conscious recollection nevertheless show normal repetition priming on various tasks.

Although the processing and memory systems accounts of repetition priming constitute potentially opposing views of memory organization and function, these two approaches are seen by many researchers as complementary (Roediger 1990). In particular, consideration of the memory systems account has encouraged the integration of informative findings from cognitive neuroscience studies, particularly those involving cases of amnesia, into behavioral theories of priming.

2.2 Masked Priming

A second form of repetition priming appears to be short-lived. In *masked repetition priming*, a prime is presented briefly (e.g., 60 ms) and is both preceded and followed by a mask. The second mask is in fact the target item that the subject is required to identify. On repetition priming trials, the target is identical with the prime, except for a change in case (*house–HOUSE*), whereas on unrelated trials the prime is unrelated to the target. Presentation of a repetition prime enables subjects to identify the target item more quickly or accurately. This kind of repetition priming has been dissociated from long-term repetition priming in two ways. First, it appears to dissipate within less than 1 s if a delay is introduced between the prime and target. Second, masked repetition priming is equally strong for common, high-frequency words and for uncommon, low-frequency words, whereas in long-term priming low-frequency words show a larger priming effect than high-frequency words.

Dissociations between masked and long-term repetition priming have supported the view that whereas the latter type of priming may be due to a form of episodic memory, the former is due to temporary activation of lexical knowledge (Forster and Davis 1984). The apparent lack of masked priming for nonwords in the lexical decision task also is consistent with this proposal. More recent evidence, however, has challenged two of these dissociations. First, there are cases in which masked repetition priming is stronger for low-frequency words, just as in long-term priming. Second, masked repetition priming of nonwords can occur when the experimental design allows subjects to overcome the tendency to classify fluently processed repetition-primed nonwords as words.

Similarities between masked and long-term repetition priming encourage the view that both phenomena may reflect the operation of a form of episodic memory. Presentation of a prime is assumed to create an episodic representation that can later be recruited as a resource to assist in the identification of that item on a subsequent presentation. This interpretation of repetition priming fits with the retrieval interpretation of semantic priming described above and promises to provide a fruitful framework for interpreting priming effects in general.

2.3 Negative Priming

There is an interesting circumstance under which repetition of an item can, paradoxically, interfere with task performance. *Negative priming* arises when a pair of items is presented on a *prime* trial, and another pair is presented on the subsequent, *probe* trial. On each trial, the subject is cued to respond to one member of the pair and to ignore the other (e.g., a drawing of a dog in red and a drawing of a violin in green; name the green item). If the ignored item on the prime trial then appears as the cued item on the probe trial, a longer response time is observed than when unrelated items appear on both trials (Tipper 1985). One interpretation of this result is that the act of ignoring an item on the prime trial inhibits the mental representation of that item, rendering it more difficult to identify it on the subsequent probe trial.

The inhibition interpretation of negative priming has prompted some researchers to apply the technique as a measure of inhibitory processing in studies of special populations such as older adults and clinical groups (e.g., individuals with obsessive-compulsive disorder). These applications appear to be premature, however, given recent evidence for an interpretation of negative priming that does not involve inhibition.

On this alternative view, negative priming does not result from selecting against an item on the prime trial. Rather, the appearance of the target on the probe trial automatically retrieves memory for its appearance on the prime trial as in semantic and repetition priming. Because the target must be selected over a distracter item, however, the subject must discriminate between the current perceptual representation of the target and the retrieved memory representation of the target's appearance on the prime trial, among other events. It is this discrimination process that slows responding and produces negative priming (Milliken et al. 1998). Two important findings support this proposal. First, negative priming is found even when a single stimulus that requires no response is presented on the prime trial, eliminating the need for any inhibition of or selection against that stimulus. Second, when no selection is required on the probe trial, negative priming is not obtained. Thus, in keeping with the memory retrieval account, it is selection on the probe trial, not the prime trial, that induces negative priming.

3. Conclusion

Although the initial theoretical frameworks for understanding priming with related stimuli and repetition priming advocated the idea that priming reflects the relationship between stable mental structures such as a mental lexicon or semantic memory and the activation of those representations, an alternative class of theories is emerging in both domains. These alternative theories emphasize the role of specific episodes in modifying memory representations and in constructing resources that are recruited when a target event is to be identified. The emergence of this new class of theories holds significant promise for the development of a unified account of how memory and perception interact to enable skilled identification of words and objects.

See also: Attention, Neural Basis of; Cognitive Psychology: History; Cognitive Psychology: Overview; Implicit Learning and Memory: Psychological and Neural Aspects; Implicit Memory, Cognitive Psychology of; Semantic Knowledge: Neural Basis of

Bibliography

Becker S, Moscovitch M, Behrmann M, Joordens S 1997 Long-term semantic priming: a computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **23**: 1059–82

Cofer C N 1967 Conditions for the use of verbal associations. *Psychological Bulletin* **68**: 1–12

Collins A M, Loftus E F 1975 A spreading-activation theory of semantic processing. *Psychological Review* **82**: 407–28

Forster K I, Davis C 1984 Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **10**: 680–98

Kolers P A 1976 Reading a year later. *Journal of Experimental Psychology: Human Learning and Memory* **2**: 554–65

Masson M E J 1995 A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 3–23

Meyer D E, Schvaneveldt R W 1971 Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* **90**: 227–34

Milliken B, Joordens S, Merikle P M, Seiffert A E 1998 Selective attention: A re-evaluation of the implications of negative priming. *Psychological Review* **105**: 203–29

Neely J H 1977 Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General* **106**: 226–54

Perfetti C A, Bell L 1991 Phonemic activation during the first 40 ms of word identification: Evidence from backward masking and priming. *Journal of Memory and Language* **30**: 473–85

Posner M I, Snyder C R R 1975 Attention and cognitive control. In: Solso R L (ed.) *Information Processing and Cognition: The Loyola Symposium*. Erlbaum, Hillsdale, NJ

Ratcliff R, McKoon G 1988 A retrieval theory of priming in memory. *Psychological Review* **95**: 385–408

Roediger H L 1990 Implicit memory: Retention without remembering. *American Psychologist* **45**: 1043–56

Tipper S P 1985 The negative priming effect: Inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology* **37**: 571–90

Whittlesea B W A, Jacoby L L 1990 Interaction of prime repetition with visual degradation: Is priming a retrieval phenomenon? *Journal of Memory and Language* **29**: 546–65

M. E. J. Masson

Primitive Society

In popular usage, 'primitive society' distinguishes ourselves from other people who have made little progress towards what we understand as civilization. There are very few such people left today: they live in scattered communities in deserts or rainforests, and they interest us mainly because we think of them as the living relics of our own distant ancestors. The comparison is usually unflattering. Until the beginning of the nineteeth century 'primitive' simply meant 'first' or 'earliest,' but as the word was applied to the original inhabitants of territories colonized by European and American states, it acquired the connotations of inferior, backward, rude.

In the evolutionary social science which developed towards the end of the nineteenth century, 'primitive society' came to be regarded as the defining subject matter of the emerging discipline of anthropology. The new natural science of biology provided the basic inspiration, and physical anthropology remains one of the four subfields of the discipline. However, although physical (or 'biological') anthropologists are much concerned with comparisons among primates they no longer distinguish human bodies as more or less primitive. We vary from one population to another in average body size, skin color, hair type, etc., and people within each community (neighbors or family members, young and old) look unalike, but unless we wish to be regarded as 'racist' we studiously avoid implying that any of these differences connotes inferiority or even some sort of adaptive advantage. Physically we are neither less 'primitive' nor more 'modern' than our ancestors 50–100,000 years ago.

It is, therefore, somewhat ironic that Darwinian hypotheses of the 'ascent of man' by natural selection have provided such a tenacious metaphor for distinguishing among human societies and cultures. For most social scientists for most of the twentieth century, being 'primitive' was a relative judgment about how people think and behave, not about their genes, brains, or bodies. Nineteenth-century explanations of social evolution construed Western civilization not just as the most successful human adaptation, but also as an inevitable and morally necessary path of progress.

While metabolizing or copulating might have changed little over the millennia, there were certainly more and less civilized ways of dining or marrying. And it was significant that only civilized modern man had the capacity to understand and explain the difference.

The first generation of modern anthropologists sought a general, empirically valid definition of the 'primitive,' and used the word without inhibition. Tylor's article defining anthropology in the famous 11th edition of *Encyclopedia Britannica* (1911) may represent the apogee of the technical use of the term. Thereafter, 'primitive' continued to appear in the title of ethnographic monographs, more as a generic label for subject matter than as an analytical concept. By the middle of the twentieth century, even the label had gone out of favor in the social sciences. The dissolution of the European empires and the emergence of new, transnational political-economic interests brought a shift in intellectual attitudes. Ideas of 'underdevelopment' as a natural, primitive condition which would be relieved by technically-enhanced progress towards modernity, soon yielded to ideas of 'underdevelopment' as a political-economic predicament inflicted on citizens of the 'Third World' by the expansion of industrial capitalism from its metropolitan centers. Peoples in the new states who were once classified as 'primitive' were now identified as the least privileged members of a widely distributed 'Fourth World,' which included the impoverished citizens of 'developed' countries.

The only entry under 'primitive' in the 1968 counterpart of this *Encyclopedia* referred to naive styles in art. 'Primitive society' has resurfaced in the present work as a historical curiosity rather than as current terminology. The phrase says more about changing interests, attitudes, and theories in anthropology than about the peoples to whom it once referred. In trying to come to terms with their designated subjects during a century of radical social change, anthropologists have buried 'primitive' society in layers of definition. These include specification of a distinctive sort of social structure, mode of livelihood, spatial distribution, and mentality, all defined (tacitly or explicitly) by a contrast with, and presumed progress towards, civilization as we understand it. Very occasionally an anthropologist will resort to 'primitive' as a portmanteau term: Hallpike (1976) uses it 'simply as a convenient abbreviation for ''non-literate, small-scale, face-to-face-societies with subsistence economies.''' Although the pejorative implications are resisted sternly, there is still an evident need to refer contrastively to people who have not yet been absorbed within the ambit of 'literate, large-scale, impersonal societies with industrial economies.' Rejecting one word around which a discipline was constructed is not enough to dismiss the assumptions about evolutionary progress which linger on in our understanding of the differences between 'Us' and the sorts of people who are 'Our' anthropological subjects.

1. *Livelihood*

The simplest and least obviously offensive definition of the primitive has been the material circumstances of their lives. Anthropology as a discipline is still organized around a sequence of social types based on technical developments in the means of production. The prototypical human population of 'hunter-gatherers' lived in nature, taking what they needed day-by-day, whereas subsistence farmers were obliged to work on nature with tools and skills, taking possession of land and other resources, and thereby setting in motion the larger political and economic organization of the state. Technical complexity increased the division of labor, which produced societies in which larger numbers of people depended on more specialized functional relationships with each other. This was especially true of the second great technical revolution which began with the industrialization of cloth production in north-west Europe towards the end of the eighteenth century. With capitalism as its social vector, industrialism swept out to transform the world, creating among other things a new and palpable contrast between the modern and the primitive. It also produced the first generation of social scientists whose role in the ever-expanding division of labor was to explain what was going on.

2. *Social Structure*

In the pattern of evolutionary contrasts which emerged in the writing of Morgan, Tylor, Rivers, and the other early anthropologists, the social world of the primitive was the small, technically, and socially 'simple' community. In this evolutionary formulation, 'primitive society' was almost a contradiction in terms, ''society'' being seen as the multiplication of roles and relationships in expanding populations, with heterogeneous individuals bound by new forms of rational contract into complex entities like the modern nation state. The primitive, by contrast, was a 'society of clones' (Kuper 1988), held together 'mechanically' (Durkheim's term) by the relations of reproduction. Social anthropologists pieced together the logic of systems of descent and marriage which apparently gave these 'homogeneous' communities their cohesion. For example, the simple rule 'reckon that you are the child of your father rather than your mother' was enough to tease a whole population out into mutually-exclusive (patri-) lineages, conferring distinct political identities on people who might never actually meet each other. Debates raged about whether paternity or maternity was the more primitive mechanism of social solidarity. The 'elementary structures of kinship' (clans, demes, moieties, etc.) conferred statuses and roles by birth, but these 'ascribed' relationships were less useful as people were drawn into much larger fields of interaction in technically more advanced societies where

survival and progress depended on geographical movement and the exercise of choice and imagination.

Governed by the biologically-derived moral mechanisms of kinship, the primitive tribe could manage without officials and rulers. From an imperial perspective, these emerging definitions of the 'tribe,' boosted by romantic writings of the period, were both intellectually attractive and administratively useful. The tribe suggested a prototypical democracy, a predominantly male political structure, and a paradigm of nationalism—the great seductive and destructive passion of the twentieth century. After two world wars which revealed the barbarity of civilization and cast doubts on who was more primitive than whom, the natural logic of the 'tribe' (as expounded, e.g., by Evans-Pritchard for *The Nuer* (1940) of the southern Sudan) was appealing. In the 1968 *Encyclopedia*, 'Primitive society' was broken down into seven cross-references: Anthropology, Culture, Evolution, Hunting and Gathering, North American Indians, and—the principal entry—Tribal Society. However, the 'Tribal' label could not stick. The year 1968 also brought the revival of the intellectual left, and the association of 'tribalism' with the ideologies and practices of imperialism quickly made it taboo. Anthropologists were now being accused of complicity in the task not just of identifying but inventing the tribes which had been the units of European and North American colonial government.

At this time, the other substitution for 'primitive' was traditional, the preferred counterpart for modern in the evolutionary scheme. But very soon this too was found to be pejorative, mainly because the innumerable books juxtaposing the two words in their titles seemed to presume an ethnocentric convergence on a Euro–American paradigm of modernity. 'Traditional societies' were scattered and diverse, stuck in the inertia of custom until history drew them inexorably into the global monotype of 'modern society.' In the 1980s such teleologies of modernization were rejected by scholars who defined themselves philosophically as postmodern. They wrote critically of our historic and ideological self-centeredness and in praise of a diverse category of 'The Other', which stood mostly against the hegemony of (Western) scholarship itself. An emblematic study was Clastres's (1977) account of the Tupi–Guarani indians of the Amazon, whose profoundly primitive social structure supposedly set up antibodies for the development of inequality, leadership, capitalism, the state, and the other vices of modern civilization. Such writing paved the way for a revival of sympathetic interest in the 'primitive.' Postmodern characterizations of 'Otherness' invoked a much older assumption about primitive society: its spatial isolation. Earlier anthropologists had pursued the prototypical primitive to the ends of the earth, especially Australia and Oceania, and ethnographic accounts of these tiny encapsulated communities in turn fleshed out the definition of 'primitive.' Postmodern anthropologists pointed out that the (relatively) concerted forces of worldwide capitalism had tended to marginalize people physically remote from the metropolitan centers, making 'peripheralness' a common denominator of what had hitherto been called traditional, tribal, or primitive. The desire to see 'Others' entirely in their own terms as just 'different' not only meant abandoning pernicious comparisons with 'Us' but also the assumption that the 'Others' had anything specific in common among themselves—some generic sort of 'traditional' culture or 'primitive' social structure.

Most problematically, it was not even safe to assume that everyone had essentially the same sort of body, because biology or physiology themselves were only historically specific sets of ideas for coming to terms with the human condition. This deliberate affront to science has exacerbated a split in anthropology, set in motion by the invention of 'sociobiology' in the 1970s. With 'evolutionary psychology' as the link, a new genre of anthropology is pursuing genetically endowed physical mechanisms which shape culture, for example, through the processes of language acquisition or mate selection. From this point of view, 'primitive society' as represented by today's vanishing hunter populations are living testimony to the 'Environment of Evolutionary Adaptedness' in which modern man was physically constructed and mentally endowed. Virtual banking or an appreciation of Beethoven is fundamentally as 'primitive' as gnawing bones. These exciting speculations have captivated public attention, putting the denser, relativistic theories of idealist anthropologists at a disadvantage.

However, it is ironic that the radically Cartesian postmodern approaches to society have taken the psychic unity of human beings entirely for granted: it is not our minds which differ, but the way they are culturally made up. This returns us to some very old and persistent mental assumptions underlying the meaning of 'primitive' in its various transmutations.

3. *Mentality*

In drawing analytical metaphors from biology or botany the social sciences and humanities have not always kept pace with theoretical advances in the natural sciences themselves. The nineteenth century evolutionist idea of recapitulation was absorbed by the emerging social sciences, and lingers as a latent assumption about human behavior in forms which biological scientists would now barely recognize. The original formula, as specified by Ernst von Haeckel, was that the life history of the individual organism repeats, from the embryo onwards, the stages of its species evolution ('ontogeny recapitulates phylogeny'). Transposed to the social domain, this fostered

the notion that children are primitive—and, by a facile converse, that primitives are children. Thus, Lubbock (1875) was fascinated by 'The similarity of savages and children,' and struggled to prove that the childish trait of phonetic reduplication (words like *bonbon*) was an infantile characteristic of primitive languages, absent from civilized English, French, German, and Greek. For social philosophers like Henry Sumner Maine, primitive society was 'the infancy of man,' and as late as 1953 Robert Redfield commended the human race for its tendency 'to grow up.'

The metaphoric transfer of this early biological concept to society and culture continues to create much confusion. Although it is implausible that a modern child could be reliving the traumatic experiences of ancestral adults, the idea that children relive primitive urges remains at the heart of Freudian psychology. Indeed, in his study of *Totem and Taboo* Freud (1938 [1919]) pursued the argument to 'resemblances between the psychic lives of savages and neurotics.' The logic of recapitulation leads inexorably to the conclusion that the metamentality of 'culture' is, for primitive people, childish and crazy, and that their world is a kindergarten which can only benefit from some civilized parenting—a good excuse for colonization and imperialism. This elaborate equation of the mad, the childish, and the primitive is extraordinarily tenacious, despite many attempts to sort out the confusion. Margaret Mead, for example, sought to demonstrate that primitive Melanesian children were actually 'more rational' than their 'animistic' parents. This subverts Enlightenment assumptions that rationality increases as a function of both of individual and species development. But for such a devout idealist and anti-evolutionist as Mead, it also raises disturbing question about how children everywhere come ready-equipped with this sort of scientific rationality, and think straight until 'enculturation' fills their primitive minds with superstitious rubbish.

Anthropologists have always traded on public fascination with the magic and superstition we associate with primitive (and childish) thought. Given its persistence in modern society, religion itself has not been an acceptable marker of the primitive. The discriminating variable here is, once again, scale: the universalistic 'World Religions' (Christianity, Judaism, Islam) deal with the mass of individual souls, while the narrow-minded particularistic religions of primitive society connect people spiritually to community and locality ('animism', 'fetishism') often through the biosocial idiom of descent ('ancestor worship,' 'totemism'). These peculiarities of primitive thought have interested French anthropologists particularly, immersed as they are in the idealist tradition. Inspired by Franz Boas, Kroeber, Mead, Benedict and other American anthropologists tried very explicitly to reject the evolutionary premises, and pursued a relativistic interpretation of culture (rather than 'society') which sought to avoid the prejudices of a superior western civilization. However, the central conundrum of relativism seems inescapable: how can we distinguish one culture from another if we have no agreed definition (some sort of 'metaculture') of what culture is in the first place? Simply to insist that all human beings have the mental capacity to 'do' culture evades the issue of how that capacity is physically embodied, how it constrains behavior, and (the evolutionary question) how and why we need it.

The most fundamental belief, which very few modern anthropologists would dare challenge, is of psychic unity, a conception of the human mind as inhabiting a standard brain which may determine how but not what we think. In the idealist tradition, what we think is determined outside of ourselves in an integral, self-reproducing system of ideas—culture, with a 'mind' of its own. The trouble is how to reconcile the enormous differences in culture with the primitive sameness of our minds. A symptom of lingering evolutionism has been the association of this primitive mentality with a closeness to nature.

This is captured in the old term savage, a synonym for primitive which appeared in anthropological titles well into the twentieth century, though increasingly in ironic quotation marks (Lévi-Strauss (1966) *The Savage Mind*; Goody (1971) *The Domestication of the Savage Mind*; LaFontaine (1981) *The Domestication of the Savage Male*, etc.). Although in French 'sauvage' still has a more benign, Rousseauian ring, the English connotations of fierceness and intemperateness latterly made it at least as stigmatic as 'primitive.'

Seeking a general and objective axis of intellectual difference, many anthropologists have settled on literacy as a key technical distinction between primitive and modern mentality. Thus, in the early 1930s, Margaret Mead uses the word 'primitive' to refer to those peoples who are completely without the benefits of written tradition, and in 1953 Robert Redfield declared 'I shall use "primitive" and "preliterate" interchangeably' (he also used a range of cognate terms including 'folk' and 'precivilized'). By the 1970s pre- or nonliterate had become the favored term for the societies traditionally discussed by anthropologists, and it remains a key universal classification for those working in the idealist tradition.

4. *Nostalgia*

A final reason for the persistence of notions of the primitive within and beyond anthropology is plain old-fashioned nostalgia—the old image of the noble savage as a counterpoint to the vices of civilization. Anthropologists are notorious for their populist commitment to the people among whom they have their main (usually first) ethnographic experience. Indeed, it has often been said that anthropologists are people

seeking a special sort of redemption in the notion that the moral virtue lacking in so-called civilized society may be rediscovered among simple people who are more in tune with nature. Defense of primitive peoples, usually in conjunction with the protection of their forest or desert environments, has become a respectable political commitment for many anthropologists. Despite Marx's rage against populist utopianism, his own material laws of history imagined a return to the perfect altruism of primitive society. This image of our natural selves before we were corrupted by the rise of capitalism was inspired by the ethnographic vision of 'Ancient Society' in North America constructed in the mid-nineteenth century by one of the founding fathers of anthropology, Lewis Henry Morgan. Marxists have been rebuked for urging the violent recovery of this old society without offering any plausible account of its structure. That it should somehow combine the benefits of the scale, technical expertise, occupational specialization and intellectual achievements of modern industrial society directly contradicts most of what has been learned about primitive society.

Today, the idea of the primitive lives on in a schizophrenia of admiration and disgust. On the one hand, popular writers like Robert Kaplan have promised the return of the modern world to barbarism, while anthropologists like Paul Richards explain that what appears as primitive lapses in Sierra Leone or Cambodia is a wholly modern circumstance—the global failures of industrial capitalism. In whatever guise, the concept of the primitive seems irrepressible in our imagination of ourselves, lurking in the shadows as we struggle to detach our vile, ancient bodies from the dream of a transcendent modern mind.

See also: Evolution, History of; Evolutionary Approaches in Archaeology; Evolutionism, Including Social Darwinism; Hunter–Gatherer Societies, Archaeology of; Hunting and Gathering Societies in Anthropology; Tradition, Anthropology of; Tribe

Bibliography

Clastres P 1977 *Society Against the State*. Blackwell, Oxford, UK

Durkheim E 1933 *The Division of Labor in Society*. Free Press, Glencoe, IL

Evans-Pritchard E E 1940 *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People*. Clarendon Press, Oxford, UK

Freud S 1938 [1919] *Totem and Taboo: Resemblances Between the Psychic Lives of Savages and Neurotics*. Penguin, Harmondsworth, UK

Goody J 1990 *The Oriental, the Ancient and the Primitive: Systems of Marriage and the Family in the Pre-industrial Societies of Eurasia*. Cambridge University Press, Cambridge, UK

Hallpike C R 1976 Is there a primitive mentality? *Man* **11**: 253–70

Kaplan R D 1994 The coming anarchy: How scarcity, crime, overpopulation, and disease are rapidly destroying the social fabric of our planet. *Atlantic Monthly* **February**: 44–76

Kuper A 1988 *The Invention of Primitive Society: Transformations of an Illusion*. Routledge, London

Lewis I M 1968 *Tribal Society*. *International Encyclopedia of the Social Sciences*. Free Press, Glencoe, IL, pp. 146–51

Lowie R 1920 *Primitive Society*. Boni and Liveright, New York

Lubbock J 1875 *The Origin of Civilisation and the Primitive Condition of Man: Mental and Social Conditions of Savages*, 3rd edn. Longmans, Green, London

Maine H J S 1893 [1874] *Lectures on the Early History of Institutions*, 6th edn. Murray, London

Mead M 1932 An investigation of the thought of primitive children, with special reference to animism. *Journal of the Royal Anthropological Institute* **62**: 173–90

Redfield R 1953 *The Primitive World and its Transformations*. Cornell University Press, Ithaca, NY

Stocking G W Jr 1982 *Race, Culture, and Evolution: Essays in the History of Anthropology*, Phoenix edn. University of Chicago Press, Chicago

Terray E 1972 *Marxism and 'Primitive' Societies*. Monthly Review Press, New York

Tylor E B 1913 [1871] *Primitive Culture*, 5th edn. Murray, London

A. F. Robertson

Printing as a Medium

Printing is a technique for producing all kinds of paper-based records (books, periodicals, musical scores, ephemera, etc.) by means of ink, a press, and reusable characters. It permits the mechanical production of many identical copies, thereby eliminating the corruption inevitable in hand-copying. Printing has had profound consequences in all spheres of social life. In particular, it has often been ascribed a major role in the transition to modernity that took place in the Western world after the Renaissance.

1. The History and Impact of Printing

1.1 The Invention of Printing

Techniques for the mechanical reproduction of texts are not a uniquely Western invention. Such techniques existed in Korea and China hundreds of years before they did in the West. Yet it is in the West that the press had its greatest impact. It is customary to divide the history of printing there into two periods: the hand-press era, lasting from the mid-fifteenth century to the early nineteenth; and the industrial era, beginning shortly after 1800 with the invention of the steam press and lasting into the present. Social scientists' arguments about the effects of printing as a medium tend to

concentrate on the former, in an effort to help explain modernization in terms of a putative 'printing revolution.' But, as will be suggested below, they would often be more aptly applied to the latter.

There is no plausible route by which Asian technologies may have influenced Johann Gutenberg (c. 1398–1468), a goldsmith from the Rhineland town of Mainz who built the first European press, so he may legitimately be called its inventor. At first, his intent was anything but revolutionary. When he constructed his first press, Gutenberg did so as a calculated intervention in the world of commercial manuscript production that already existed. His earliest printed books were designed to resemble manuscripts, and he intended to sell them as such. He also swore his workmen to confidentiality about his new technique, apparently, in the manner of apprentices in any late-medieval craft. But the strategy did not last. In 1462 Mainz was besieged and sacked, and Gutenberg's printing house broke up. His workers went to other towns and set up their own operations. Before long, printers had appeared in a score of German, Dutch, Italian, and French cities, often fostering legends that printing had been invented locally. Venice, Avignon, Strasbourg, and Haarlem created the most plausible of such histories, as a result of which determining the true origins of the press remained a notoriously intractable problem well into the nineteenth century. William Caxton set up his printing house in London at the end of this first wave of expansion. By that time, in 1480 or so, a new European craft was firmly established.

1.2 The Practice of Early Printing

That craft centered on the *press*—a simple, largely wooden machine, capable of being made by semi-skilled workmen and, if small enough, amenable to concealment and transport from place to place. This press would normally be operated in a place called a printing house. Here, a compositor would set type into a frame called a *form*, which would then be placed in the press itself, which two pressmen would operate to make up to 1,000 impressions per day. The resulting sheets would then be dried, and collated to make books. It sounds, and in essence it was, a simple process. But in practice it was hard to sustain. Most printing houses were severely under-capitalized, and depended for their day-to-day survival not on the printing of books, but on that of ephemera and pamphlets—anything, in fact, that furnished regular and frequent influxes of cash. Almost all of these products were destroyed soon after manufacture. The population of books that has lasted to the present day to fill our great research libraries therefore represents an extreme distortion of the reality of early modern print culture.

This being so, printing raised an immediate question: whether culture could be allowed to be defined by commercial necessity. Early modern Europe attempted to answer that question by constraining printing on three social levels: the household, the craft, and the polity (where the polity might be a religious one, such as the Catholic church).

In the first place, a printing house, like most places of work in early modern Europe, was normally supposed to be a residence as well. The master printer generally lived in the same building—as, quite frequently, did his journeymen and apprentices. There was a reason why this was held to matter. It was conventional to believe that the moral structure of the patriarchal household largely guaranteed the legitimacy of practices pursued within that household, and printing was to be subjected to the same moral oversight.

In the second place, printing was also organized into guilds, one of which existed in every major city. To be a master printer was to be a respected member of such a group. Each maintained its own conventions, expected to be honored by all members of the book trade. Violations were identified and remedied by book-traders themselves, generally behind closed doors, so that for outsiders the image of an intrinsically harmonious order of print was sustained. Again, there was nothing strange about this; it was how all early modern crafts operated.

And in the third place, such guilds operated in fragile harmony with regulations imposed by church and state. Chief among these were 'privileges' (patents) and licenses. Privileges were grants of literary property, to use an anachronistic phrase, given out by the monarch or equivalent authority. They conflicted in principle with titles allocated by the guilds internally, which was one reason why they were eventually replaced by copyright conventions. A license, on the other hand, was a permission to print (in Latin, *imprimatur*, 'it may be printed'), granted by an authorized individual to the printer or bookseller. In principle, all works had to be vetted by such an official before publication, but in practice this was rather less severe than the censorship sometimes portrayed by modern historians. Most books were never licensed, and only a tiny minority were pursued for lack of an *imprimatur*. The Counter-Reformation Catholic Church extended this principle of licensing into a systematic *Index of Prohibited Books*, which listed titles the ecclesiastical licensers had decreed too unsound for general readers. Again, while this undoubtedly affected the availability of certain books and authors (Machiavelli and Copernicus, for example) in Catholic lands, protestant publishers quickly realized that appearing on the Index virtually guaranteed best-seller status in their own countries, and they rushed to print large impressions.

In these ways, then, the social order of early modern Europe sought to accommodate the potentially revolutionary new craft. It was never an entirely stable arrangement, and by the late seventeenth century it

was breaking down. In England, first, licensing lapsed in 1695. With it went much of the authority of the local guild, called the Stationers' Company. Nobody wanted outright deregulation, however, and the major result was a hard-won new compromise between state, authorial, and trade interests. It was articulated in the so-called 'Copyright Act' (1710—the term copyright never actually appears in it). This Act became the basis for the entire subsequent history of copyright and intellectual property law. By the late eighteenth century, agitation for reform of printing and literary property on similar lines had occurred across Europe, and in the years around 1800 broadly similar copyright provisions were established in all major Western states; they extended beyond Europe and the colonies in the nineteenth century. The modern conventions of printed authorship thus owe their origin to the existence and instability of early modern attempts to define and regulate printing: to licensing, patents, and the craft practices of printers themselves (Chartier 1992).

1.3 Industrial Printing

The technology of printing remained virtually unchanged from soon after Gutenberg's invention until around 1800. At that time, however, change came rapidly. Not only did the legal and conventional practices governing the industry alter once and for all; so did the technology of printing itself. First came Stanhope's metal press. This was soon followed by the steam press, which at a stroke increased production rates by an order of magnitude. Stereotyping finally made the dream of truly identical copies a reality. And at the same time, the technology of papermaking changed dramatically with the success of Henry and Sealy Fourdrinier's machines in Hertfordshire. Although it never had the iconic status of the steam press, the papermaking machine had as great a role in the changes of this period, and Brunel himself hailed it as 'one of the most splendid inventions of our age.' With the advent of all these new devices occurring within a generation, between 1800 and 1840 printing underwent its own industrial revolution.

The first commercial steam printing was done at the office of *The Times* in London. It vastly increased the print run of the newspaper, quite literally overnight (the management had the machines installed in secrecy to prevent unrest among the soon-to-be-redundant workers, and unveiled them only when they produced their first day's issue). Nor did the improvements cease there: by 1827 *The Times* had new presses that were producing 5,000 copies an hour. Steam printing was understandably slower to take hold in smaller publishing markets, since it required a large capital outlay—Cambridge University Press, for example, was unusually avant-garde among specialist houses in adopting the technology, yet still did not do so for more than a generation. And in general the book publishing business was less immediately affected by industrialization than the newspaper and periodical press. But the advent of industrial technology nonetheless changed printing forever, and vastly increased its social impact.

In fact, many of the attributes that historians and social scientists attribute to the hand press may more properly be assigned to the steam press, in conjunction with mechanical papermaking and stereotyping. Typographical fixity in particular was never as clear a component of hand printing as we tend, anachronistically, to assume—in an edition of perhaps 300–1,000 copies, any number of variations would occur across different copies, as correction would be done on the fly. This kind of variability now ceased. Print runs 10 or 100 times as great as had previously been viable also represented a quantitative leap as great as that associated with Gutenberg. They made possible, indeed in economic terms virtually necessitated, the first national daily press. The development of mass literacy, too, was really much more a phenomenon of the nineteenth and especially twentieth centuries than of the sixteenth and seventeenth; mass education of the kind parodied by Dickens in *Hard Times* relied on steam printing. At the same time, since industrial printing required a great deal more concentration of capital than had hand printing, it encouraged agglomeration and uniformity within the publishing industry itself. The press barons of the twentieth century, with their incalculable influence on social and political life, owed their baronies to the printing techniques introduced in the nineteenth century. In effect, printing had become a branch of factory culture.

2. A Printing Revolution?

Gutenberg's invention had profound social and cultural consequences, and modern historians and social scientists have tried hard to discover what they were. First and foremost, the press produced texts in unprecedented quantities. Lucien Febvre and Henri-Jean Martin (1958), in the first modern work to focus on the history and impact of printing, calculated that the number of printed books soon exceeded the number ever produced in manuscript up to that time. Sheer quantity of this order had its own consequences. Books were suddenly available in unprecedented numbers, in more places, and at lower costs. As they spread through social ranks, people—especially town-dwellers—made these new objects their own. Literacy increased markedly, and that alone constitutes one of the major turning-points in the development of modern social order. But mere numbers cannot explain the more specific cultural consequences of the press. Major *qualitative* changes were also brought about by print. These have been described most exhaustively by Elizabeth Eisenstein (1979). They were of two major

kinds: transformations in the making and appearance of the page, and innovations in the uses of books by writers and readers.

After Gutenberg's brief experiment with imitation, the printed page soon began to look very different from its manuscript predecessor. Typefaces and layouts became standardized. Cross-referencing, indexes, errata, and notes came into use, according to conventions that themselves grew more uniform. Woodcuts and engravings allowed the use of repeatable images, permitting diagrams, charts, and scientific illustrations to be communicated with a security never before possible (Ivins 1953, McKenzie 1986). Hitherto, it had been very difficult to express sophisticated claims in pictorial form, so rapidly did images degrade in the hands of copyists. The simple ability to juxtapose and compare reliable representations of competing claims revolutionized the possibilities for learned work.

This textual 'fixity,' as Eisenstein calls it, affected a number of cultural realms, among which religion and science may be singled out. It is not a coincidence that Luther's reformation lasted, while previous attempts at reform (such as those of Hus, or the Lollards) had withered or been wiped out. The permanence and cornucopic plenty of print made Luther's program impossible to suppress. Indeed, it is in the furore over Lutheranism that some of the most characteristic forms of mass communication were invented, such as the pamphlet and the polemical cartoon (Scribner 1981). And, most fundamentally, only such a machine could truly hope to permit a 'priesthood of all believers' by making the Bible available as an object to innumerable readers.

In the sciences too the advent of print permitted great upheavals. The contemplative natural philosophy of the schoolmen gave way before a newly active interrogation of natural processes, modeled, allegedly, on the practices of craftsmen. Print not only circulated this new natural knowledge but also helped to create it, since printers' homes became 'polyglot households' where scientists of all nationalities could meet and correspond. Print thus helped forge the norm of open communication that is so central to science. And men like Henry Oldenburg secured the status of the new science in novel forms of printed communication. Oldenburg's *Philosophical Transactions*, launched in 1665, survived to rank today as the oldest of all scientific journals.

But printed texts alone did not change the world; printed texts plus new ways of using them did. Not least, the sheer mass of printed matter required new classification techniques—in effect, a science of bibliography—to master a worldwide 'library without walls.' But more profound was the change in reading practices that print facilitated. From its early days, the *Philosophical Transactions* was the central plank of a new, transnational readership of naturalists, gentlemen, mathematicians, and scholars—the ancestor of today's scientific community. The common readership of such periodicals was rapidly recognized as uniting a novel kind of collective, called the *republic of letters* or, rather later, the *public sphere*. This 'public' was defined by its engagement with print. And with Baconian ambition, it took all knowledge to be its province (Habermas 1962). It marked a decisive break from traditional and absolutist societies. In Habermas' very influential account, it therefore laid the foundations for true socio-political transformation, most notably in the form of the French revolution. In that sense, the 'printing revolution' ushered in by Gutenberg warranted its name.

3. *Theories about Print*

It was in the context of the Jacobins that printing was first recognized as an explicitly revolutionary force. True, Francis Bacon's salute to the press (along with gunpowder and the mariner's compass) was much cited; yet it was little more than an aphorism. The first fully articulated and contextual interpretation of printing's impact was that of the Marquis de Condorcet (1795), who was seeking to explain the overthrow of the French monarchy. Condorcet sketched out a history of 'the human spirit' in a series of stages, with the invention of the press constituting a major turning-point and effectively making the deposition of the Bourbons into 'the revolution that the discovery of printing must bring about.' In composing this sequence, he became the first writer to chart a trajectory of print-based modernization that saw its effects manifested first in the sciences, and only later in general social life. That trajectory has retained its persuasive power in modern times, but its meaning has been construed in very different ways.

Heirs to Condorcet, modern historians, and social scientists have continued to concentrate their interpretative efforts on the typographical *ancien régime* that came to an end with the dual inventions of copyright and the steam press. Compared with this, the period of industrial printing has received comparatively slight attention. There have been many empirical histories of printing, publishing, and reading in the nineteenth and twentieth centuries, but nothing as influential in social-scientific terms as Habermas or Eisenstein. Indicative is the reception of Habermas's work on the public sphere: most Anglophone readers neglect its second half, which tells a depressing story of the replacement of Enlightenment public culture with industrial mass culture. His story is far more elegiac than celebratory, but one would not realize this to read most of the Anglo-American commentators on *The Structural Transformation of the Public Sphere* (1962).

Where they do not ignore the later history of printing, most scholars simply conflate it with the earlier history, and speak of 'print' as though it were a

single, uniform entity. Probably the most famous and notorious of construals along these lines has been that of Marshall McLuhan (1911–1980), the Canadian literary critic and scholar. McLuhan's *The Gutenberg Galaxy* (1962) and *Understanding Media* (1964) set the pattern for a series of works published in the 1960s that made extravagant claims for the cultural and even psychological impact of print. His works used what he called 'probes'—unsystematic, radical aphorisms delivered scattergun fashion to jolt readers out of the formulaic thinking that, he believed, typography inevitably instilled. The probes worked, in at least one sense. McLuhan himself became a cult figure: Tom Wolfe remarked that he sounded like 'the most important thinker since Newton, Darwin, Freud, Einstein, and Pavlov.'

What McLuhan achieved was to make 'media' the subject of analysis in their own right. He argued that these media, of which print was the archetype, brought the world together, dissolving boundaries. The most important of these boundaries were socio-psychic in character. Indeed, McLuhan's very definition of a medium was that it acted as an 'extension of man,' transcending the material constitution of the human frame. In effect, he believed that the state of media determined not only how humans lived, but what humans were. This he presented as a matter of evolution. In particular, McLuhan portrayed what he called 'typographic man,' namely the kind of being who lived according to the cultural logic of print. Typographic man thought in terms of linear logic and objectivity, because his reference points were fixed texts, in what later became the Eisensteinian sense of fixity. Typographic man was both individualist (he could be sure who he was) and nationalist (he could see the bounds of his community, and tell the difference consistently between it and others). There had been no such being before about 1450. At that point, typographic man had begun to displace 'tribal man,' who had been restricted to local contacts and communication by hand. And in his turn typographic man was now becoming extinct, to be replaced by something else that McLuhan himself left nameless, but who could easily be called 'electronic man.'

McLuhan remarked that it was in the nature of electronic media to integrate with the nervous system itself, establishing humans as nodes on a worldwide network. In effect, the skin itself would dissolve as a social constraint, as neurology became inseparable from sociology. Written in the early 1960s, before the first computer network, it is easy to see why McLuhan has recently been rediscovered by the digerati and labeled a 'prophet of the Web.'

Yet McLuhan's arguments were ultimately too faddish in their tone, and too determinist in their content, to remain convincing beyond their bare outlines. Criticisms soon started to mount up (an influential example is Jonathan Miller's short 1971 introduction, which leaves little doubt where its author stands). There are few social scientists or historians today who would openly acknowledge McLuhan as a leading influence on their representations of print and its consequences. But that influence is nonetheless real. For the most part, it is refracted through the historical analysis of Eisenstein. Eisenstein's careful arguments about the impact of the hand press gave empirical weight to McLuhan's aphorisms. Through her, his proposals have gained academic respectability. An example is Benedict Anderson's widely-praised account of nationalism (1983), which rests on an explicitly Eisensteinian argument linking print to time reckoning, and thence to consciousness of national identity. Similarly once-removed arguments of this kind can be found in many of the analyses of 'print culture' to have appeared in the 1980s and 1990s.

Yet there is one important respect in which current trends in sociological and historical research are starting to depart consciously from this approach. This is the move towards an empirical history of the practice of reading. The immediate origins of this trend lie in France, where the postwar enterprise of the history of the book was born. The spur came from the responses of both history and sociology to the dominant *Annales* school. In both fields, a realization took hold in the later 1970s that quantitative social-scientific accounting failed to capture something fundamental about the cultural impact of print. At the same time, Pierre Bourdieu (1979) was emphasizing the importance of active cultural appropriation by readers of the press rather than passive cultural reception by those readers (Hoggart (1957) had earlier made similar arguments for an English readership). By the early 1980s it looked as though attention to the varied ways of using books, building on Bourdieu's approach, might stand a chance of seeing what the quantitative *Annalistes* were clearly missing. Since then the history of reading has become a burgeoning field in its own right.

The leading proponents of the history of reading have been cultural historians like Roger Chartier in France and Robert Darnton in the USA, who have pioneered the reinterpretation of the French Revolution in these terms. Chartier in particular objects to the determinism about print that is implicit in the work of Habermas, and would direct attention more to changes in reading practice in the eighteenth century (Chartier 1990). In particular, he argues that it is hard to understand how the scatological and pornographic literature that admittedly flooded eighteenth-century France was intrinsically more devastating to the priesthood, say, than the rich seam of libels that appeared in Luther's day. Instead, Chartier would point to a newly skeptical and extensive practice of reading that was devoted to these materials, making them far more damaging in use. This practice contributed substantially to the desacralization of kingship. It hence made regicide possible, and with it the origins of modernity. In this light, the printing revolution

needs to be redefined as part of a broader reading revolution (Cavallo and Chartier 1999).

As a result, research into printing as a medium is today leaving behind quasi-determinist accounts of print exerting some kind of cultural 'logic' on societies, and towards empirical and historical research into the different ways in which societies have made use of the technology and its products. Important examples include the national histories of the book now being published in Britain, the USA, and other countries (Chartier and Martin 1982–1986; Amory and Hall 1999; Hellinga and Trapp 1999). D F McKenzie's 1984 study of a decisive cultural encounter mediated by print in his native New Zealand shows how this modern approach can influence anthropological understandings as well as historical. In McKenzie's terms, excessively schematic talk of 'print logic' is being replaced by a properly contextual 'sociology of texts' (McKenzie 1986). It is interesting to speculate whether, with print logic thus banished, the use of 'print culture' as a catch-all explanatory device may eventually fall into abeyance too. And that may happen, not coincidentally, just as the advent of electronic media signals the end of half a millennium in which print was the predominant communicative medium.

See also: Communication: Electronic Networks and Publications; Literacy and Illiteracy, History of; Mass Media, Political Economy of; Media and History: Cultural Concerns; Media Effects; Media, Uses of; News: General; Radio as Medium; Television: General

Bibliography

Amory H, Hall D D (eds.) 1999 *A History of the Book in America: Vol. I, The Colonial Book in the Atlantic World*. Cambridge University Press, Cambridge, UK

Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Bourdieu P 1979 *La Distinction: Critique Sociale du Jugement*. Éditions de Minuit, Paris [Also published as 1986 *Distinction: A Social Critique of the Judgement of Taste*. Routledge, London]

Cavallo G, Chartier R (eds.) 1999 *A History of Reading in the West*. Polity Press, Cambridge, UK

Chartier R 1990 *Les Origines Culturelles de la Révolution Française*. Seuil, Paris [also published as 1991 *The Cultural Origins of the French Revolution*. Duke University Press, Durham, NC]

Chartier R 1992 *L'Ordre des Livres: Lecteurs, Auteurs, Bibliothèques en Europe entre XIVe et XVIIIe Siècle*. Alinea, Paris [also published as 1994 *The Order of Books*. Polity, Cambridge, UK]

Chartier R, Martin H-J (eds.) 1982–1986 *Histoire de l'Edition Française*. 4 vols. Promodis, Paris

Condorcet M J A N, de Caritat, Marquis de 1795 *Esquisse d'un Tableau Historique des Progrès de l'Esprit Humain*. Agasse, Paris [also published as 1955 *Sketch for a Historical Picture of the Progress of the Human Mind*. Noonday Press, New York]

Darnton R 1982 *The Literary Underground of the Old Regime*. Harvard University Press, Cambridge, MA

Eisenstein E L 1979 *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-modern Europe*. Cambridge University Press, Cambridge, UK

Febvre L P V, Martin H-J 1958 *L'Apparition du Livre*. Michel, Paris [also published as 1984 *The coming of the Book: the Impact of Printing 1450–1800*. Verso, London]

Habermas J 1962 *Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. Luchterhand, Berlin [also published as 1989 *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. MIT Press, Cambridge, MA]

Hellinga L, Trapp J B (eds.) 1999 *The Cambridge History of the Book in Britain: 1400–1557*. Cambridge University Press, Cambridge, UK; New York

Hoggart R 1957 *The Uses of Literacy: Aspects of Working Class Life with Special Reference to Publications and Entertainments*. Chatto and Windus, London

Ivins W M 1953 *Prints and Visual Communication*. Harvard University Press, Cambridge, MA

McKenzie D F 1984 The sociology of a text: Orality, literacy and print in early New Zealand. *The Library* **6**(6): 333–65

McKenzie D F 1986 *Bibliography and the Sociology of Texts*. British Library, London

McLuhan M 1962 *The Gutenberg Galaxy: The Making of Typographic Man*. Routledge and Kegan Paul, London

McLuhan M 1964 *Understanding Media: the Extensions of Man*. Routledge and Kegan Paul, London

Miller J 1971 *Marshall McLuhan*. Viking, New York

Scribner R W 1981 *For the Sake of Simple Folk: Popular Propaganda for the German Reformation*. Cambridge University Press, Cambridge, UK

A. Johns

Prison Life, Sociology of

The sociology of prison life concentrates on thick descriptions of the actions and accounting practices that make up the distinctive social organization of a prison. Chief amongst its preoccupations have been how administrators and staff strive to maintain an authority which can never be wholly legitimate in the eyes of their captives; how, in turn, those captives contend individually and collectively with conditions of deprivation and adversity, including the adversities inflicted by their fellow inmates (Morris and Morris 1966, pp. 168–9); how a negotiated social order can emerge out of the exchanges between the two; and how, as a result, prisons manage to persist at all.

The proper study of prison had to await the emergence of universities and other institutions that could sustain social research, deploy some form of fieldwork as part of its methodology, and appreciate

that the prison is a strange institution which 'challenges the anthropological imagination' (Sparks et al. 1996, p. 33). It was not until the 1940s and beyond that those three strands converged in what are commonly held to be the first significant pieces of writing about the sociology of prison life, Clemmer's (1940) *The Prison Community* and, perhaps more important, Sykes's (1958) *The Society of Captives* and Goffman's (1961) *Asylums*. Because they set a pattern and foil for much of what was to come, it is useful to recite their principal arguments.

The Prison Community reported research conducted in the 1930s in a penitentiary for adult felons by a sociologist attached to the mental health staff. Clemmer used the methods of anthropology to study what he called the 'facts of social life,' the groupings, patterns of leadership, controls, leisure pursuits, sex, and work of prison, and to illustrate the complex relations, and the informal relations in particular, of a predatory, suspicious and atomized world of inmates banded publicly against authority but privately riddled by informers and mistrust.

The Society of Captives depicted inmates suffering 'the pains of imprisonment' and striving for the goals of better food, more rewarding work, fuller information, access to sex and pornography, security, status, and autonomy, formally denied them by an administration that purported to treat all prisoners more or less alike and certainly completely distinct from staff. It was in this sense that prison was held to engender anomic pressures that led, in their turn, to the production of illegitimate stratagems, informal social structures and a code of conduct that could satisfy a number of the inmates' aspirations. Inmates were enjoined not to interfere in one another's business, not to 'lose their head,' not to exploit one another, not to weaken, and not to trust the guards. Like many another code, it was largely exhortatory, a form of 'figurative action' rather than a reliable predictor of behavior and its accompanying social relations generated neither 'a perfect solidarity … [nor] a warring aggregate' (Sykes 1958, p. 83). In practice, some prisoners were forced or encouraged to service the sexual demands of others; some asserted their social standing by violence, defiance or conspicuous example; some supported illegal markets in the provision of goods; and overlapping hierarchies emerged.

The ensuing social order was ostensibly clandestine and based on a rejection of the rejecters, and the staff were supposed to suppress it, but the staff confronted their own difficulties. They themselves wielded great power in principle but little in effect, they were constrained by rules and the practical exigencies and limits of control, they were outnumbered, troubled by the precariousness of their authority, and discouraged from drawing their weapons lest they be disarmed by inmates. Their reply was to replace a formal equality of treatment by a *de facto* recognition and rewarding of the stronger inmates and inmate groupings, and they did so in exchange for a devolved social order in which inmates effectively regulated one another. The consequence was what Sykes called 'functional corruption,' but it could as readily be seen in functionalist language as a drift towards system integration, equilibrium, or homeostasis, an exemplification of the doctrine that the latent functions of action can work in seemingly paradoxical fashion to promote stability in social systems.

Goffman (1961) worked as a participant–observer in a mental hospital in Washington, DC, to construct a model of asylums, prisons, and like organizations as 'people-processing' 'total institutions' that were substantially cut off from the larger world by symbolic and physical boundaries; a continuous round of activity lived publicly within the walls; a relative lack of internal differentiation of space, role, and temporal schedule; an unbridgeable divide between staff and inmates; and, through what he called a 'mortification of the self,' the discarding of old identities and the assumption of new, degraded, and formally homogeneous identities peculiar only to the institution.

Five main processes have shaped the character and development of much of the work that followed. First has been the adoption of a somewhat critical stance. It was unavoidable that criminologists, penologists, and sociologists came motivated to the study of prison life, sometimes wishing to reform prisons, sometimes to abolish them, and sometimes—more rarely—to analyze them with the dispassion that they might extend to any other social institution. But it has been a strong distaste for prisons that has colored much of what they have written, and a number have been avowedly partisan and, indeed, polemical in their work. That stance has, in its turn, led to methodological difficulties in winning access to, and trust from, a number of key groups in prison, and prison officers especially.

Work has, second, selectively incorporated interpretations of transformations occurring in the political, administrative, and demographic conditions of penal institutions. In the United States, for instance, the sociology of prison life was born in descriptions of autocratic, *laissez-faire* regimes and the resistance that was waged by conservative inmates who feared they might lose power in the face of administrative reform. It moved on through a discussion of American penology's flirtation with rehabilitation in the early 1960s and the dilemmas that arose when the particularistic ethos of treatment confronted the universalism of disciplinary control. It looked at the radical politicization of prisoners, and particularly of black prisoners, in the late 1960s and early 1970s. It analyzed the efforts at legalistic reform, the organizational instabilities, fragmentation of inmate communities, and new waves of riots of the 1970s and early 1980s. And it has culminated most recently in an analysis of the new, racialized politics of mass incarceration and incapacitation that, in Simon's phrase,

is the consequence of 'governing through crime' (Simon 1993).

Internally, American prisons were to become 'Balkanized' into antagonistic racial camps based in some measure on a prior membership of street gangs. Externally, Simon has claimed, the logic of the penal system was to be extended out from the prisons through parole, policing, and probation to subject large blocks of the population, and particularly the young, black male population of the cities of the United States, to a tight control based on actuarial and risk-driven calculations of propensities to offend. Prison life has, in effect, spread beyond the walls of the American penitentiary to the ghetto.

Third, like the sociology of almost any other substantive area, the study of prison life has tended to edit and import some of the more general and evolving theoretical preoccupations of sociology at large as they have appeared to illuminate, and be illuminated by, those structural changes. It has already been observed, for instance, that work was launched inside a context set by the structural-functionalism and anomie theory of the 1940s and 1950s, a context suited to the study of willful, autocratic repression. Current work focuses on issues of resistance and authority, gender and power, 'governmentality' and 'structuration,' and other matters now engaging a number of the wings of sociology.

Fourth, like any other method, and aided by those transformations, interpretations and theoretical preoccupations, its focus has advanced restlessly from one phenomenon to another, continually introducing new 'facts' with which to re-interpret prison life. Above all, it has produced what is in effect a running series of footnotes or qualifications to the foundational models of the 1950s and 1960s.

It was easy enough at the beginning to overlook some of the finer and more nuanced details of structure. The social anthropologist in the American penologist would have been fascinated by the inmate community's vividly phrased argot and roles. What was less often reported—because it was less obtrusive and less interesting at first sight—was the presence of those who were relatively disengaged and played no part in the clandestine life of the prison, the 'retreatists' and others: who tried to serve their time uneventfully and discreetly and who eventually became interesting precisely because of their earlier neglect. What was also partially obscured at first was the lives of those apart from the inmate. Sociologists moved away from the inmate and towards chaplains, medical officers, and others: for example, the male prison officer beset by problems of violence, and the female prison officer and her gendering of control.

Studies report not only the features that were neglected in the earliest work, but also new features that have emerged since that work was done, but much remains unexplored, for instance, the structural repercussions of the social class of prisoners, aging and the age composition of prison populations, the world of prison administration, and such groupings as gay prisoners and the prison untouchables (the 'nonces' or nonsenses of English prison argot) imprisoned for sexual and violent offences against children or the elderly.

Most important, the political control of discreditable information and the absence of a sense of anthropological curiosity and of a robust ethnographic tradition in much of the world have ensured that there have been almost no studies of prison life outside Britain, North America and a few small pockets of northern Europe. Sociologists began to emphasize how the early models had tended to extrapolate from one form of time-bound regime alone, the American, authoritarian, maximum security prison for men. It is a model that works well enough where power is arbitrary and formally unchecked, controllers are outnumbered and resistance is fragmented, and it applies broadly to German concentration camps and Soviet labor camps as well as to some of the prisons of the United States. But doubts have been raised about how far or strictly it should otherwise be generalized. It is clearly a limited model. Take Ketchum (1965), who described how in Ruhleben, an internment prison camp established in Berlin for the citizens of countries hostile to Germany in the First World War, there could be seen almost none of the informal responses to incarceration listed by Clemmer, Goffman or Sykes. To the contrary: the inmate population behaved cooperatively, peacefully, and without mutual exploitation, and its conduct invites questions about the impact of class, shame, and pre-existing patterns of organization on the social construction of prison life.

Again, take research on a number of therapeutic or rehabilitative regimes outside the United States. Sparks (1994) studied Barlinnie, a small unit for long-term prisoners in a Scottish prison, and Genders and Player (1995) examined Grendon Underwood, an English therapeutic prison to which serving prisoners in other establishments could elect to go. The three authors declared that neither Barlinnie nor Grendon was critically troubled by the legitimation crises or conflicts of other prisons, but both institutions did benefit from being regimes within regimes—inmates could always be transferred back to the larger and more coercive system from which they had come—and it would be wrong to assume that therapy alone is a proof against disorder or confrontation. So it was that Mathieson's (1965) *Defences of the Weak* described a medium-secure Norwegian prison, organized on therapeutic lines, whose disunited and cynical inmates challenged those who guarded them in a language of 'censoriousness' that turned therapeutic ideology into a medium of criticism, defense, and attack. Theirs was a response very dissimilar to that of other models.

A second, allied form of qualification to the foundational model attacked the proposition that prisons are

total institutions which sever biographies and create new men and women within an indigenous culture (Sparks et al. 1996, p. 44). Instead of the fractured identities described by Goffman, study after study has revealed that inmates actually bring with them beliefs, associations, and styles of behavior from the outer world.

The importation into prison of social forms from the street gangs and antagonistic racial politics of America has already been noted, and there are other borrowings that have presumably been colored by local and national traits in the world outside the walls. Thus, Irwin and Cressey (1962) talked in America of *thieves* who carried a wider criminal culture into prisons, treating imprisonment as a career contingency; *convicts* whose behavior was a localized adaptation to the special travails of prison life; and s*quare Johns* who had an allegiance to neither convict nor thief but to a show of obedience to authority. Also, Cohen and Taylor (1972) talked about the affinities between the manner in which long-term inmates in Durham Prison in England managed problems of serving time and the offenses for which they had been convicted, armed robbers worrying about physical deterioration and concentrating on body-building exercises, and fraudsters engaging in copious attempts at litigation.

The discovery of the gendered character of imprisonment is as revealing. The very first studies dwelt on institutions for men. Whilst it was recognized from the outset that male inmates confronted largely male problems stemming from a loss of liberty, status, and heterosexual contact, gender itself was not treated as especially problematic or interesting. It was only when ideas about sexual identity were applied in explanations of the behavior of women inmates in the 1960s, women who were said to form themselves into pseudo-families in response to a loss of emotional sustenance, that the influence of gender itself became notable. Other sociologists have subsequently ex-extended and modified that early work, but the thesis that women combine together differently remains. Later, ideas about masculinity were themselves turned back on the social organization of men's prisons to translate them into places ordered by a 'hegemonic masculinity.' Male prisons, it was argued, must be understood as institutions that are dominated by a very masculine exercise of violence, reliance on hierarchy, ritualized combativeness, and the suppression of displays of emotion and vulnerability.

Last, it must be stressed how closely, although never absolutely, the study of prison life has been bound up with the use of ethnographic methods (Terence and Pauline Morris said of their base in Pentonville Prison, London, that it was 'like the anthropologist's hut on the village street' (Morris and Morris 1966, p. 8)). Its history has, by extension, roughly paralleled the rise and fall of sociological ethnography itself. It was probably at its most vigorous between the 1950s and early 1970s, when, according to Janowitz, it 'contain[ed] some of the most outstanding research monographs in sociology' (Foreword to Jacobs 1977, p. ix). Its decline may be linked in part to the seemingly inevitable half-life enjoyed by any body of theory and research, and to problems of access and publication that began to surface in the 1970s when prisons were turbulent and their managers became politically defensive and defeatist. But it may also be attributed to the ascent to influence of Michel Foucault (1979), whose political economy of discipline prompted scholars to turn *en bloc* to reflections about what they called 'penality' in a markedly more abstract and empirically disengaged manner.

Foucault took Bentham's plans for a concentric, panopticon prison to be an especially telling representation of the new technologies of control that were emerging in society at large. Prisoners in the brightly lit rim could never know whether they were being observed by guards in the shadowy inspection tower at the center, and the result was supposed to be that they would discipline themselves, colluding in their own subordination in a decentralized, self-sustaining, economical and dispersed system of power. The panopticon was never finally constructed in England, and Foucault's history was flawed, but so seductive and wide-reaching was his metaphor that it became an archetype for a large new family of ideas, and sociologists came to adopt Foucault rather than Clemmer, Sykes or Goffman as their mentor. Jonathan Simon (1999, p. 4) observed that, 'the great tradition of studies of prison social organization that begins with Clemmer and Sykes … has largely ceased in the last twenty years … Much of the research now is conducted from afar [and] … the new distance introduces important interpretive problems … .'

See also: Bentham, Jeremy (1748–1832); Control: Social; Courts and Adjudication; Crime and Class; Crime and Ethnicity (Including Race); Crime, Sociology of; Criminal Justice, Sociology of; Delinquency, Sociology of; Foucault, Michel (1926–84); Goffman, Erving (1921–82); Imprisonment: Sociological Aspects; Law: Economics of its Public Enforcement; Police, Sociology of; Prisons and Imprisonment; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Rehabilitation of the Offender

Bibliography

Clemmer D 1940 *The Prison Community*. Holt, Rinehart and Winston, New York

Cohen S, Taylor L 1972 *Psychological Survival*. Penguin, Harmondsworth, UK

Foucault M 1979 *Discipline and Punish*. Penguin, London

Genders E, Player E 1995 *Grendon: A Study of a Therapeutic Prison*. Oxford University Press, Oxford, UK

Goffman E 1961 *Asylums*. Doubleday Anchor, Garden City, NY

Irwin J, Cressey D 1962 Thieves, convicts and the inmate culture. *Social Problems* **10**: 145–55
Jacobs J 1977 *Stateville: The Penitentiary in Mass Society*. University of Chicago Press, Chicago
Ketchum J 1965 *Ruhleben: A Prison Camp Society*. Oxford University Press, Oxford, UK
Mathieson T 1965 *Defences of the Weak*. Tavistock, London
Morris T, Morris P 1966 *Pentonville: A Sociological Study of an English Prison*. Routledge, London
Simon J 1993 *Poor Discipline: Parole and the Social Control of the Underclass*. University of Chicago Press, Chicago
Simon J 1999 The 'society of captives' in the era of hyper-incarceration. Unpublished paper
Sparks R 1994 The Barlinnie Special Unit as prison and escape. *Prison Service Journal* **95**: 2–6
Sparks R, Bottoms A, Hay W 1996 *Prisons and the Problem of Order*. Clarendon Press, Oxford
Sykes G 1958 *The Society of Captives*. Princeton University Press, Princeton, NJ

P. Rock

Prisoners' Dilemma, One-shot and Iterated

The Prisoners' Dilemma is the formal structure of a common sort of collective action problems. It is also a vivid scenario embodying that structure. It offers a useful framework for the analysis of interaction where there are certain conflicts of interests.

1. Two Agents, One Meeting

Adam and Eve have been arrested. The District Attorney (DA) now tells them this. They can either talk or keep silent. If they both talk, they will each get a 10-year sentence. If one of them talks (confessing for both) and the other does not, the one who talked will go free; the other will get a 20-year sentence. If they both keep silent, each will get a one-year sentence on some trumped-up charge. The DA lets them discuss it. Then he puts them in separate cells and goes to take their pleas.

Say that neither Adam nor Eve cares about how the other makes out or thinks he will meet the other again. Each considers only the length of his or her own sentence. For each, that person's talking (D) dominates that person's being silent (C); that is, whatever Eve will do, Adam prefers the outcome of his talking to that of his *not* talking, and correspondingly for Eve. So if both Adam and Eve are rational, both of them will talk. But since each prefers the outcome of C,C (both of them silent) to that of D,D, both will be sorry for it.

This is the familiar prisoners' scenario, first presented by Albert Tucker in 1950. The story speaks of two people as agents, but the structure of the interaction allows the agents to be teams or nations—some of the earliest applications were to the Cold War race (on these matters, see Poundstone 1992). The arms story is meant to demonstrate that, in certain conflicts of interests where there is some agreement too, rational agents won't cooperate though it would be better for each if they both did than if they did not.

2. Two Agents, Many Meetings

Where the agents are rational and they think they won't meet again, there is no way out for them. Neither party will cooperate (C). Both will defect (D), and both will be sorry. But say that they think they *will* meet again, that their present interaction is only the first of many just like it. (In place of the possible jail-sentences, let the payoffs be sums of money reflecting the same ranking of the outcomes.) Here it may seem that the prospect each faces of having to live with the other's resentment ought to deter defection. Still, it often is argued that, where the number of rounds is finite and known to both of the agents, if they both are rational (and think each other to be rational …), both will defect from the start to the finish.

The argument first appeared in Luce and Raiffa (1957). Suppose that the number of rounds is known by both Adam and Eve to be n, and that they both are rational. At the nth round, each will know that there will be no further meetings, no need to guard against penalties, so they both will defect. Let Adam think that Eve is rational—let him think this throughout. He will then think in the $(n-1)$st round that Eve will defect in the nth round whatever *he* does in the $(n-1)$st, that his cooperating in the $(n-1)$st, wouldn't be rewarded by Eve in the nth. He will therefore defect in the $(n-1)$st round, and Eve, thinking likewise about Adam, will too. The same in round $n-2$: each expecting the other to defect in the round that follows, each will here defect, though we must now also assume that Adam thinks that Eve thinks *him* rational—let him think this too throughout—and that Eve thinks he thinks *her* rational. So we move stepwise back to round 1 (both agents defecting all the way), though with a heavier load of assumptions at each preceding stage.

This backward induction rests on assumptions about these people's beliefs about each other. It calls for supposing that, from the start to the finish, Adam believes Eve to be rational, believes she believes *him* to be rational, believes she believes he believes *her* to be rational, etc., and so too with the names transposed—call this Iterated Mutual Belief, or IMB. Some writers have noted that IMB isn't credible (Basu 1977, Bicchieri 1989, Pettit and Sugden 1989). Where in fact it fails, rational people might cooperate in at least some of the rounds; they need *not* defect throughout. Both parties may thus want it to fail, and one or the other may act in some way he thinks would get it to fail.

What sort of action would do that? If either cooperated in any round, the other would have to make sense of that, and this alone would suffice. Let Adam cooperate at $n-1$. Then Eve will either think (a) that Adam isn't rational (that he thinks this best for him however Eve responds to it) or (b) that he thinks *she* isn't rational (that he expects her to respond by cooperating at n). If he cooperates at $n-2$, she will think (a) or (b) or (c), that he thinks *she* thinks *he* isn't rational (that he expects her to respond by cooperating at $n-1$ in the hope of getting *him* to cooperate at n), etc. Whether Eve thinks (a), (b), or (c) (or the next-level d at $n-3$), IMB will have failed.

But would it ever be rational for Adam or Eve to cooperate? More generally, how should two rational people act in a sequence of Prisoners' Dilemmas (the number of rounds being known)? The answer depends on what each thinks the other party would do. Let B be the strategy of starting with C and then tit-for-tatting, doing in every subsequent round what the other did just before. Also let B_x be starting with C, then tit-for-tatting in all but the last x rounds, and then defecting in those. It would be rational for Adam to adopt B_x if he thought Eve will defect in round 1 but would then adopt B_{x-1}, and *that* would be rational for Eve if she took Adam's cooperating in round 1 to mean that he was pursuing B_{x-2}. There may be many such pairs of courses. In each, both Adam and Eve cooperate in some (or many or almost all) of the rounds, and what they are doing is rational. For each person's course is the best response to what that person thinks is the other's. But they still will defect in some (perhaps in many) rounds.

3. *Two Agents, Equilibria*

Facing a sequence of interactions, people seldom know how many there will be. So let us turn to cases in which their number is unknown to the agents. Here the backward-induction argument cannot get a foothold. We cannot begin with the round both parties know will be the last—none is here known to be last. But neither can we argue for the possibility of cooperation as we just did, for that too assumed that the agents reason backwards from some known final round. In game-theoretic terms, a Prisoners' Dilemma is a *game*. Where two people consider what strategy of action they should follow in some sequence of such games, they are facing what is called a Prisoners' Dilemma *supergame*. We are here speaking of *open-ended* supergames of this sort.

The fullest treatment is Taylor's (1987). Let each agent have a number of open-ended strategy options. If he is rational, he will choose among them in light of how he ranks the outcomes of their different pairings with the other agent's options, every such outcome being the sequence of the results of each agent's acting in each round as his strategy requires. Write each outcome of, say, C,C as $[C,C]$. The outcome of Adam's taking B where Eve does too (of both starting with C and then tit-for-tatting) is thus $[C,C]$, $[C,C]$, $[C,C]$ … . This is also the outcome of their both cooperating every time—of their both taking C^∞. Let B' be starting with D and then tit-for-tatting; where Adam takes B and Eve takes B', the outcome is $[C,D]$, $[D,C]$, $[C,D]$ … . Where Adam takes D^∞ (defecting each time) and Eve takes B, the outcome is $[D,C]$, $[D,D]$, $[D,D]$ … .

How Adam ranks these outcomes depends on how he values future benefits and costs. The more the future now matters to him—the smaller his *discount rate*—the more will it figure in his rankings. (If $1-d$ is Adam's discount rate, $0 \leqslant d \geqslant 1$, he values some benefit r rounds in the future at d^r times its worth to him now.) Where the future matters enough, he will prefer $[C,C]$, $[C,C]$, $[C,C]$… to $[D,C]$, $[D,D]$, $[D,D]$…, the prospect of the long run of $[D,D]$'s offsetting the appeal of the initial $[D,C]$. Thus if the future matters enough, he will prefer his taking B where Eve does too to his taking D^∞ where Eve is taking B.

If he is rational, what strategy will he follow? A supergame is a game, and the central concept of game theory is that of (Nash) equilibria. An equilibrium in a two-person game is a pair of options, each a best response to the other. The idea is that, if Adam knows the situation (including the operative discount rates), he will choose a strategy in an equilibrium pair. What if there are several equilibria? Call an equilibrium *admissible* if the outcome of no other equilibrium is either preferred to it by both agents or preferred by one of them while the other is indifferent. Taylor suggests that a rational person will choose an option in an admissible equilibrium pair.

He notes that several equilibria are possible in a supergame of the Prisoners' Dilemma: D^∞,D^∞ is always an equilibrium, and some other pairings of strategies are equilibria where the agents' discount rates meet certain conditions. His basic conclusion is that where the discount rates of both agents are sufficiently small, B,B and D^∞,D^∞ are equilibria and all other strategy pairs he thinks would be considered are not. Since only B,B is admissible (both agents prefer its outcome to that of D^∞,D^∞), each of the agents will choose B. It follows that, where their discount rates are sufficiently small, both agents will cooperate in every round.

This last may overstate the case. Whether it is rational to cooperate in every round depends not only on the discount rates but also on the strategies considered. Where only B,B',C^∞,D^∞, and certain others are in view, Taylor's conclusion holds. But still others might come up. Let the agents consider E^m (being an *exploiter*): taking B only after defecting m times while the other cooperated. Let them also consider P^m (being *patient*): cooperating m times and then taking B. The E^m,P^m pair is an admissible equilibrium (where the future matters enough to the

patient)—if Adam followed E^m and Eve followed P^m, they would both be rational. But Adam would be exploiting Eve in each of the first m rounds.

4. *Many-person Interactions*

Say that there are many people each of whom must either cooperate (pay their dues, pull their weight, etc.) or defect (*not* pay, *not* pull). Each would be better off defecting, whatever the others did. But if all defected, or more than a certain number, all would be worse off than they would be if they all had cooperated. Here we have a *many-person* Prisoners' Dilemma.

A frequently cited scenario is Hardin's (1968) commons story. Each farmer in the village sends his cows to graze on the just barely adequate commons. Each knows that his sending out a new cow would degrade that commons. Since each knows too that his share of the cost would be less than the benefit to him of having an extra cow, each will send out another cow and the commons will collapse. This is the logic of the underprotection (and underprovision) of *public goods*, of the *free rider* problem, the commons (the oceans, the forests, clean air) being the public good, the overgrazing (overfishing, polluting) being free riding.

Refinements are possible here. In a many-person Prisoners' Dilemma, each agent would be better off defecting than cooperating, whatever the others did. Each would also be better off if all cooperated than if all defected. But say that there is some lone defector who exploits all the others. Where these others remain better off than they would be if all defected, the defector is a mere nuisance—he is a free rider proper. Where he leaves some others worse off than they would be if all defected, the defector is injurious—he is a *foul dealer*. (For this distinction, see Pettit 1986.)

Let all the parties expect to be facing the same sort of cases many times over and let each consider what to do in the whole sequence of them. They are then facing a many-person Prisoners' Dilemma supergame. Much has been written about this (see Taylor 1987 for references). It appears that here, as above, cooperating in each round may be rational. This too is argued by Taylor, who shows that everyone's cooperating in round 1 and then also in all the rest if each of the others cooperated just before is an admissible equilibrium if each agent's discount rate is sufficiently small. But so is a lone defector's taking E^m (moving to B only after defecting m times while all the others cooperated) and the others taking P^m—the conditional cooperation after round m being dependent on everyone's cooperating in the round just before. So here too rationality allows for extensive exploitation.

The question of the rationality of cooperation in such many-person contexts has some resonance in political theory. The claim that it *isn't* rational (and so can't be expected) provides the principal backing for the idea that people need a strong governing state. Hobbes and Hume based their case for authority, for a *Sovereign* on this, and Hardin argues for strict enforcement of access rights to commons (for limiting procreation and entry into wealthy countries). Taylor uses his proof to the contrary to argue for the viability of voluntarism and for communal rule. So some large issues are involved.

5. *Evolution and Stability*

The above has been about questions of the rationality of various strategies. There is also a second sort of question. How well in the long run would various strategies fare in different contexts of others? Which (if any) would gain more adherents—which would *evolve*? And which (if any), if pursued by all, would resist invasion by others—which would then be *stable*? Much of the recent discussion derives from Axelrod (1984), a study relating Prisoners' Dilemma to current work in evolutionary biology.

Axelrod ran computer simulations of two round-robin tournaments of two-agent supergames, each supergame pairing a strategy with either itself or some other in the pool. Strategy B won both tournaments, and Axelrod notes that it would have won too in many other contexts (other strategy pools). Suppose that a tournament were repeated many times, the agents in each a new *generation*. Letting the change in the numbers of those pursuing the different strategies in each generation be in proportion to the relative fitness (the yield divided by the average yield) of those who pursued them in the one just before, Axelrod also ran a simulation of a many-fold repetition of tournaments. He found that B was the strategy pursued most often in the end. Conditional cooperation can thus *evolve*.

Would any strategy, if pursued by all, resist an invasion by others? Would any strategy be *stable*? Axelrod held that B would be stable (if the future mattered enough). This has been challenged by Boyd and Lorberbaum (1987), who hold that no strategy is ever stable. Which of these claims is correct?

Resistance to invasion might be *weak*, the invaders being kept from multiplying. Or it might be *strong*, their being pushed to extinction. Biologists speak only of *monotone* dynamics, in which greater fitness implies faster increase in numbers. Under such a dynamic, no strategy is *strongly* stable: none can strongly resist invasion. Say that the natives now pursue B. Let $B2$ be tit-for-*two*-tats, that is, starting with C and then cooperating unless the other defected in the preceding *two* rounds. $B2$ is a *neutral mutant* of B: it yields exactly what B yields in all interactions with either B or $B2$. Since B can't gain on $B2$, can't *strongly* resist an invasion by it, it isn't strongly stable. And neither is any other strategy, for every strategy has neutral mutants.

What about weak stability? Let us allow for *heterogeneous* invasions, not all the invaders pursuing the same strategy, and let us allow for (monotone) dynamics other than that of proportional fitness. Say

that *B* is the native strategy, that some of the invaders follow *B2* and the rest *B′*, and that the dynamic is that of imitating the preceding tournament's winner. *B* and *B2* do well against themselves and against each other, *B2* doing slightly worse against *B′*, against which *B* does badly. *B′* does badly against *B*, *very* badly against itself, and slightly better than *B* against *B2*. *B2* will win the first tournament, and having won, will be imitated and become the only strategy followed. (It in turn will be vulnerable to the alternation of cooperation and defection!) So *B* isn't weakly stable under the imitation dynamic, and neither is any other strategy, for there always is some other, inept (like *B′*) against itself, against which certain neutral-mutant invaders do better than the natives.

No strategy is strongly stable under any monotone dynamic, and (allowing for heterogeneous invasions) none is weakly stable under *every* such dynamic. Thus Boyd and Lorberbaum are partly right. Still, it can be shown that *B* is weakly stable under the proportional-fitness dynamic: it is (weakly) stable *in that special context*. So Axelrod is partly right too. (These results are in Bendor and Swistak 1998.)

The basic thesis here is that greater fitness implies faster increase in numbers. This is true of the genes that render their carriers fitter to reproduce. But are we here speaking of genes? Are there genes (or alleles) for *B*'ing and for *B2*'ing and the like? If not, the thesis must refer to the numbers of people *X*'ing or to the numbers of replicas of the *X*'ing-*meme* (for memes, see Dawkins 1976 and Dennett 1995). Increases in these are independent of reproductive fitness. They depend on imitation (by progeny or by others), and people's reproductive fitness needn't increase the number of their imitators.

What sort of fitness promotes imitation? There may be no single answer, for sets of memes co-adapt to their situations. 'Natural selection among genes chooses those that prosper in the presence of certain other genes … [and] gene pools come to consist of genes that do well in each others' company … .' (Dawkins 1980, p. 354). The same holds for selection among memes. So the prospects of *X*'ing may hinge on the support of co-adapting memes, among them the agents' ideologies, their concepts of morality. In the context of some common ideologies of behavior toward others, *B*'ing may be imitated. In the context of more *patient* ideologies, *B2*'ing or *B3*'ing or P^m'ing may be imitated. Where it goes along with an *exploiter* ideology, even E^m'ing may be imitated. What fits a strategy for imitation depends (at least partly) on the co-evolving ideology. Whether or not this is good news depends on the ideology.

See also: Economics and Ethics; Evolution, Natural and Social: Philosophical Aspects; Game Theory: Noncooperative Games; Rational Choice Explanation: Philosophical Aspects; Surprise Examination

Bibliography

Axelrod R M 1984 *The Evolution of Cooperation*. Basic Books, New York

Basu K 1977 Information and strategy in iterated Prisoner's Dilemma. *Theory Decision* **8**: 293–8

Bendor J, Swistak P 1998 Evolutionary equilibria: Characterization theorems and their implications. *Theory and Decision* **45**: 99–159

Bicchieri C 1989 Self-refuting strategies of strategic interaction: A paradox of common knowledge. *Erkenntnis* **30**: 69–85

Boyd R, Lorberbaum J P 1987 No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. *Nature* **327**: 58–9

Dawkins R 1976 *The Selfish Gene*. Oxford University Press, Oxford, UK

Dawkins R 1980 Good strategy or evolutionarily stable strategy? In: Barlow G W, Silverberg J (eds.) *Sociobiology: Beyond Nature/Nurture*. Boulder, Westview, CO

Dennett D C 1995 *Darwin's Dangerous Idea*. Simon & Schuster, New York

Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8

Luce R D, Raiffa H 1957 *Games and Decisions*. Wiley, New York

Pettit P 1986 Free riding and foul dealing. *Journal of Philosophy* **83**: 361–89

Pettit P, Sugden R 1989 The backward induction paradox. *Journal of Philosophy* **86**: 169–82

Poundstone W 1992 *Prisoner's Dilemma*, 1st edn. Doubleday, New York

Taylor M 1987 *The Possibility of Cooperation*. Cambridge University Press, Cambridge, UK

F. Schick

Prisons and Imprisonment

Imprisonment is the principal form of punishment in most countries for people convicted of serious crimes. People charged with crimes are often held in pretrial confinement until the charges are resolved. The institutions in which convicted offenders and pretrial detainees are held are usually called prisons. In the United States, however, the word 'prisons' refers only to state and federal institutions for offenders sentenced to terms of a year or more. Pretrial detainees and offenders sentenced to shorter terms are held in 'jails' operated by local governments. This article examines the history of prisons, historical and comparative patterns of imprisonment, and the collateral and direct effects of imprisonment on prisoners, their families, and their communities, and on crime rates and patterns.

1. Origins

The modern prison dates from the period 1780–1840, though recognizable precursors existed. Imprisonment as punishment for convicted criminals dates from classical times, but such incarceration occurred only

intermittently and on a small scale. By the sixteenth and seventeenth centuries, many Western countries had created debtors' prisons and houses of correction for paupers, and gaols for holding people pending trial.

Development of the Penitentiary is usually attributed to American Quakers. Penitentiaries were to be places of solitude and introspection where, guided by religious instruction, convicted offenders would repent, realize their errors, and prepare for law-abiding futures. The earliest such regime was established in 1790 in the Walnut Street Jail in Philadelphia. The first purpose-built American penitentiaries were in New York, and in Pennsylvania (see *Punishment: Social and Legal Aspects*).

In its extreme form, the Quaker regime, known as the 'silent system' and associated with Pennsylvania prisons, required prisoners to live in solitary confinement with only limited contacts with staff and clergy. Later, work was permitted in the one-person cells. Because the silent system led to the mental breakdown of many prisoners, it was soon abandoned for the 'congregate system' associated with New York prisons, in which inmates lived alone but worked together in enforced silence.

Europeans including Charles Dickens and Alexis de Tocqueville visited America to examine the penitentiaries and to consider their feasibility. Millbank, the first national penitentiary in England, was built in 1816, although the first on the American model, Pentonville, was not built until 1842. The first French penitentiary, La Petite Rocquette, opened in 1836. Penitentiaries soon existed in every European country.

A second wave of innovation occurred in the late nineteenth century, exemplified by the 1877 opening of New York's Elmira Reformatory. This prison, for male offenders up to the age of 30, was premised on rehabilitative ideas. It strove to reform offenders by providing a wide range of educational and vocational programs together with discipline and religious instruction. These rehabilitative ideas led to the first parole release systems, since there was no point in holding prisoners after they had been reformed. They also precipitated creation of indeterminate sentencing systems in which judges, prison managers, and parole boards exercised broad, standardless discretion over the lives of offenders.

Historians point out that the bursts of enthusiasm were short, that even in their own eras the innovative institutions were atypical, and that most prisons at most times have been brutal, inhumane places. Marxist accounts tied the development of the prison to the emergence of the industrial economy and portrayed the prison as a mechanism for labor discipline (see *Marxism and Law*). Michel Foucault in *Discipline and Punish* (1978) argued that the prison, like public education and the army, was a reaction to the breakdown of the *ancien régime* and served as a socializing institution, producing disciplined individuals suited to bureaucratic and conformist roles provided by the modern state. David Rothman, in *The Discovery of the Asylum* (1971), described prison developments in Jacksonian America as an attempt by philanthropic reformers to respond to social dislocation and rapid change by trying to reform the by-products of a disordered society: the deviant, disorderly, and deranged. In *Conscience and Convenience* (1980), Rothman considered why new Progressive Era institutions were created and why they endured. 'Conscience,' an idealistic wish to improve the lives of troubled people, explains their creation, and 'convenience,' the autonomy and flexibility given managers by indeterminate sentencing, explains their durability. Michael Ignatieff, in *A Just Measure of Pain* (1978), emphasized the humanitarian intentions of prison reformers but, like Foucault, saw the penitentiary as a product of a series of nineteenth-century social crises and the emergence of the modern state.

During most of the twentieth century, researchers, the media, and the public knew little of life inside prisons. The laws, traditions, and rehabilitative rationale of indeterminate sentencing gave prison and parole officials broad authority over the day-to-day lives of prisoners. In the United States, a 'hands-off doctrine' led courts to refuse even to consider prisoners' complaints, reasoning that prisoners were 'slaves of the state' and that courts should defer to the professional expertise of prison administrators.

Developments in the 1960s and 1970s undermined indeterminate sentencing and transformed American prisons. The prison's rehabilitative rationale lost credibility. Reviews of rehabilitative programs concluded that there was little credible evidence that correctional treatment programs reduced recidivism. The civil rights movement and social ferment of the 1960s led prisoners, especially minority prisoners, to become more assertive. This resulted in high-visibility prison riots, most famously in New York's Attica Prison in 1971, and led to abandonment of the hands-off doctrine. Conditions in many American prisons were then racist, squalid, and brutal. By the late 1970s, federal courts declared conditions in many jails and prisons to be unconstitutional. Prisons in more than 40 states came under federal court control in the 1980s.

Beginning in the 1960s, crime and punishment became central political issues in the United States. Many politicians ran for office on 'law and order' platforms, and supported harsher policies. Fifteen states abolished parole release. Many states enacted laws to make conditions in prisons less comfortable, including reducing or eliminating access to televisions and recreational facilities; others eliminated recreational and treatment programs. In the 1990s, nearly 30 states enacted 'three-strikes-and-you're-out' laws requiring mandatory life sentences for people convicted of a third serious crime. Most prosecutors requested

harsher penalties, most judges imposed them, and most parole boards released fewer prisoners.

Other changes affected prisons. By the late 1990s, private corporations operated hundreds of institutions and provided comprehensive services, such as medical care, to others. Efforts to build new facilities before 1980 were often stymied by not-in-my-backyard movements; communities thereafter competed for new prisons as local economic development initiatives. In the 1960s, there were no labor unions for prison guards; in California in the 1990s, the guards' union was a major campaign contributor and an active lobbying group.

Most developments affecting American prisons also occurred elsewhere. Compared to the USA, prison-guard unions developed earlier and were stronger in Australia, England, and France. Prison privatization occurred extensively in Australia, New Zealand, and England, though nowhere else in Europe. Riots precipitated by prisoners' complaints occurred in many countries. In England and in some Australian states, conservative legislators in the 1990s enacted harsher laws and mandated more Spartan prison conditions, but nowhere as harshly as in the United States. Prisoners' rights movements arose in many countries but, because US constitutional traditions give more legal protections to citizens than are recognized in most countries, no other courts intervened extensively in prisons. The exception is in Europe, where prisoner grievances can be redressed under the European Convention on Human Rights, and where the European Convention Against Torture and Inhuman Treatment influences improvements in prison conditions (see *Torture: Legal Perspectives*).

In 2000, prisons outside the Western countries were typically inhumane, unhygienic, overcrowded, and understaffed, with notable exceptions in a few Asian countries, especially Japan. The United Nations Declaration on Human Rights pertains to prisons, but no international body has authority to enforce such provisions. As a result, reform pressures must come from local groups, with occasional support from human rights groups such as Amnesty International and Human Rights Watch.

2. *Imprisonment Patterns*

Countries vary widely in how they use imprisonment. In the United States, from 1925 to 1972 the incarceration rate fluctuated between 90 and 130 per 100,000 residents (state and federal prisons only), leading some scholars to argue that self-balancing processes maintained relative stability over time in the incarceration rate. However, since 1972, American incarceration rates have increased every year, quintupling by the century's end. Trends in other countries vary widely, though generally the 1990s were a period of rising incarceration rates. In Holland, the incarceration rate increased continually after 1976, although the 1998 rate was still only 80 per 100,000. In Finland, however, the incarceration rate dropped continually after 1975 and in 1998 was around 60 per 100,000, only 40 percent of the rate in the mid-1960s. In Japan, the incarceration rate fell rapidly in the 1950s and then fluctuated around 35 per 100,000. In Southern European countries, including France, Italy, and Spain, incarceration rates varied erratically with sharp rises and declines.

National incarceration rates vary greatly. The United States' rate in 1998, 680 per 100,000 (prison and jail inmates combined), was among the highest in the world. Only Russia had a comparable rate (685 per 100,000 in 1998). As a region, Eastern Europe had the world's highest rates in the late 1990s, typically varying between 200 and 500 prisoners per 100,000. England, Ireland, Australia, Canada, and New Zealand had rates of 100–130 per 100,000. Most Western and Southern European countries had rates between 50 and 100 per 100,000, and most Scandinavian countries incarcerated approximately 60 people per 100,000. With the notable exception of Singapore (465 per 100,000 in 1997), most Asian countries had incarceration rates of 20–50 per 100,000.

Very different penal strategies can lie behind incarceration rates. Some Scandinavian countries, notably Sweden and Finland, have low absolute rates (55–60 per 100,000 in the 1990s), but large percentages of convicted offenders receive short-term prison sentences. In Germany, a much smaller percentage of offenders are sentenced to prison, and sentences shorter than six months are disapproved. In most European and English-speaking countries, only 1 to 3 percent of prison sentences are for 2 years or longer. In the United States in 1991, 56 percent of state prisoners were serving sentences of 10 years or longer. In part, these differences result from different penal ideologies. Many American policy makers believe crimes are primarily the product of wrongful moral choices of criminals, that punishment policies can affect crime rates, and that prison sentences should be harsh to provide deterrent disincentives, or lengthy to incapacitate incorrigible offenders. Many European policy makers believe that decisions to commit crimes are powerfully influenced by social conditions and inadequate socialization, and that punishment policies are unlikely to influence crime rates.

Imprisonment is used in most countries primarily as a punishment for violent crimes, serious property crimes, and drug trafficking, and seldom for white collar, regulatory, or environmental crimes. In many countries, prison sentences for violent and drug offenders are getting longer, and populations of drug offenders are increasing. Most prisoners in every country are young, male, unmarried, and economically and socially disadvantaged, and have extensive criminal records. Large percentages are drug dependent and committed their crimes under the influence of

alcohol or other drugs. Women constitute only 3 to 8 percent of prisoners in most countries but their numbers and percentages are increasing faster than men's (see *Crime and Gender*). Disadvantaged minority groups are disproportionately present in most prison systems, though the identity of the minority groups varies widely. Relative to population, blacks in the USA and Afro-Caribbeans in England are eight times more likely than whites to be in prison. In Australia and Canada, the aboriginal/nonaboriginal differences are as high as 15 and 20 to 1. Similar disproportions affect North Africans in Belgium, Holland, France, and Italy, Turks and Eastern Europeans in Germany and Scandinavia, and Koreans in Japan (see *Race and the Law*).

3. *Collateral Effects of Imprisonment*

At least five kinds of collateral effects can be identified. First, what are the effects on prisoners' later lives? Economic and ethnographic literature in the United States and Europe show that imprisonment reduces ex-offenders' subsequent incomes and employment. In the United States, state and federal laws deny ex-offenders the right to vote or to hold office, the opportunity to engage in certain occupations, and the right to receive various public benefits and services.

Second, what are the effects on offenders' spouses or partners, and their children? The relevant literature is thin and fragmented, and the most pressing task is to pull together existing knowledge in order to formulate plausible, testable hypotheses. Hypotheses would address the effects of imprisonment on the financial and social stability of prisoners' families while they are in prison and afterwards, on the maintenance of prisoners' relationships with families, and on the short- and long-term well-being and social functioning of prisoners' children. Sometimes families and partners benefit from the removal of abusive, disordered, or dysfunctional parents and spouses. However, given the strong negative effects on children's well-being of being raised in disadvantaged, single-parent households, the major effects of imprisonment on spouses and children are often likely to be negative.

Third, what are the effects on prisoners' later physical and mental health? Controlling for age, sex, and social variables, suicide rates are higher in prison than elsewhere. Psychological literature on coping and adaptation concludes that even long-term imprisonment has few lasting mental health effects, although some researchers doubt this. It would be surprising if established adverse effects on income, employment, and family functioning were unrelated to former prisoners' mental and physical health. However, a long-term, multiple-measure, longitudinal study is needed to answer such questions, and none has ever been done.

Fourth, what are the effects on prisoners' later crime involvement? The negative effects on ex-prisoners' incomes, employment prospects, and family involvement are predictive of increased offending probabilities. At least since the time of John Howard, the eighteenth-century prison reformer, the proposition has been put forward that prisons are 'schools for crime,' that inexperienced prisoners are socialized into antisocial and oppositional attitudes and leave prison more likely to commit crimes than when they entered. Sociologists refer to this as 'prisonization.' No informed person doubts that this process sometimes happens, and at least partly offsets any crime-reductive effects of imprisonment achieved through deterrence, incapacitation, and rehabilitation. Many European scholars accept as proven that prison is criminogenic. They contend that for crime-prevention reasons, prison terms should be avoided whenever possible. Such arguments explain why German and Austrian policies impose prison terms only for serious crimes and discourage imposition of short sentences (under six months).

Fifth, what are the effects on the larger community? Plausible hypotheses can be offered about effects of American prison expansion on the diversion of public resources (from higher education particularly, and from other public services, to prisons), on economic development (from inner cities where most offenders and victims live, to rural communities where prisons are built), and on community cohesion (in disadvantaged minority communities, large fractions of the young men are or have been in prison and are thereby disabled from working, parenting, marrying, or otherwise contributing to the community). In some disadvantaged communities, imprisonment is so common that it may no longer carry meaningful stigma, and it may have lost whatever deterrent effects it once had.

4. *Crime-control Effects of Imprisonment*

The literature on crime control effects is inconclusive, small, and for the most part old. First, has increased imprisonment reduced crime rates? Most people probably assume that a quarter-century's quintupling of the American prison and jail population has reduced crime rates. There has, however, been relatively little research in recent years on deterrence and incapacitation effects (see *Deterrence: Legal Perspectives*). Most authoritative reviews of the subject, including the famous 1978 report of the National Academy of Sciences Panel on Research on Deterrent and Incapacitative Effects, conclude that such effects exist, but are modest (Blumstein et al. 1978). Similar conclusions were reached after an exhaustive survey of deterrence research carried out in 1998 by Andrew von Hirsch et al. for the English Home Office.

Second, can prison treatment programs reduce recidivism rates? The pessimism associated with the 'nothing works' findings of the 1970s and 1980s has passed, and there are grounds for cautious optimism about the positive effects, under some conditions, of some cognitive-skills, drug-treatment, vocational training, educational, and other programs in adult prisons. The conditions, however, are stringent: program eligibility must be carefully matched to prisoners' needs and risks, programs must be well implemented and adequately funded, and compatible aftercare programs in the community must sustain treatment.

Third, without distinguishing among deterrent, incapacitative, and rehabilitative effects, does increased imprisonment reduce crime rates in general, and have the increases of recent years done so? The prevalent answer appears to be yes, but the literature, much of it by economists, is technical and inaccessible to nonspecialists. Moreover, research mostly dates from the 1970s and suggests that effects are more modest than is widely understood.

Fourth, in cost-benefit terms is increased imprisonment worth it? This is the smallest and most primitive of the prison effect literatures, and provides few policy-relevant findings. The work to date is useful mostly for methodological reasons, by showing how not to pursue such inquiries. For example, the most widely publicized recent work greatly exaggerates the costs of crime by means of inflated imputed costs of victim 'pain and suffering,' and takes no account of suffering by offenders or pains of imprisonment borne by their partners, children, or communities. It may be, as Franklin Zimring and Gordon Hawkins (1991) suggested, that cost-benefit assessments of the effects of imprisonment require weighing of inherently incommensurable values and have reached a dead end.

5. *Prisoners and Prison Staff*

In 1970, a sociological literature on prison and prisoner subcultures was the best known body of prison scholarship. Since then, prisons have changed a great deal, as have inmate and staff subcultures and interactions between them, and the learning of earlier times may no longer be valid. American prisons are much larger, inmate populations and staffs are more disproportionately black and Hispanic, more line and management staff are women, gangs are larger and their influence more pervasive, more prison staff are unionized, a large and growing fraction of prisons is under private management, and judicial oversight and intrusion are greater. Because larger fractions of prisoners are serving very long sentences, we have witnessed increased need for medical care and geriatric services.

Behavioral, cultural, and social changes in the larger society impinge on life inside prisons. AIDS and HIV, for example, are more prevalent inside prisons than outside. The proportions of mentally ill and mentally defective prisoners increased as a result of the 1960s and 1970s movement to deinstitutionalize the mentally ill. The civil rights and women's movement have affected prisons, as has the political conservatism of recent years. Programs premised on restorative justice ideas are beginning to appear inside prison walls.

Studies of women prisoners and life inside women's prisons are conspicuously underdeveloped. There is a small quality literature, but a paucity of research on life inside prisons since 1975 has had an especially impoverishing effect on traditionally under-studied subjects.

See also: Deterrence; Deterrence: Legal Perspectives; Deterrence Theory: Crime; Imprisonment: Sociological Aspects; Legal Issues: Public Opinion; Prison Life, Sociology of; Punishment, Comparative Politics of; Rehabilitation of the Offender

Bibliography

Adams K 1992 Adjusting to prison life. In: Tonry M (ed.) *Crime and Justice: A Review of Research*. University of Chicago Press, Chicago, Vol. 16, pp. 275–359

Blumstein A, Wallman J (eds.) 2000 *The Crime Drop in America*. Cambridge University Press, Cambridge, UK

Blumstein A, Cohen J, Nagin D (eds.) 1978 *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. National Academy Press, Washington, DC

Foucault M 1978 *Discipline and Punish*. Pantheon, New York

Ignatieff M 1978 *A Just Measure of Pain: The Penitentiary in the Industrial Revolution, 1750–1850*. Pantheon, New York

King R, McDermott K 1995 *The State of Our Prisons*. Clarendon Press, Oxford, UK

Morris N, Rothman D (eds.) 1995 *The Oxford History of the Prison: The Practice of Punishment in Western Society*. Oxford University Press, Oxford, UK

Rothman D 1971 *The Discovery of the Asylum*. Little, Brown, Boston

Rothman D 1980 *Conscience and Convenience: The Asylum and its Alternatives in Progressive America*. Little, Brown, Boston

Stern V 1998 *A Sin Against the Future: Imprisonment in the World*. Northeastern University Press, Boston

Tonry M (ed.) 1998 *The Handbook of Crime and Punishment*. Oxford University Press, New York

Tonry M, Petersilia J (eds.) 1999 *Prisons*. University of Chicago Press, Chicago

von Hirsch A, Bottoms A, Burney E, Wikström P-O 1998 *Criminal Deterrence and Sentence Severity: An Analysis of Recent Research*. Hart Press, Oxford, UK

Walmsley R 1996 *Prison Systems in Central and Eastern Europe*. European Institute for Crime Prevention and Control, Helsinki, Finland

Zimring F, Hawkins G 1991 *The Scale of Imprisonment*. University of Chicago, Chicago

M. Tonry

Privacy: Legal Aspects

Serious social and legal concern with privacy is a relatively modern matter, though themes of privacy and its importance form part of all known human civilizations.

1. *The Meaning of Privacy*

The definition of privacy is a matter of some controversy. Its characterizations move from the all-encompassing 'right to be left alone,' covering a general interest in not being interfered with in any way that violates human dignity, advocated by Edward Bloustein (1964), to the more limited interest in control over personal information and accessibility, argued for by Westin Allen (1967) and Parent (1983). Somewhere in the middle we find characterizations of privacy as only those aspects of liberty that relate to intimate and highly personal aspects of one's life (e.g., Rubenfield 1989), such as those involved in sexual preferences and decisions about whether and when to procreate. The uses of 'privacy' in everyday language and in various legal systems are broad enough to cover all of these approaches, so the analysis of ordinary language cannot, by itself, decide the issue. The controversy is found in philosophical analyses of privacy as well as in the contexts of its legal protection. It has led some scholars to argue that the term be abolished in order to promote clarity of thought and analysis (Prosser 1960, Thomson 1975).

Conflating privacy with all claims for noninterference, or with the value of an 'inviolate personality,' is indeed much too broad, and threatens the distinctiveness and the practical utility of privacy. All concede that there are ways to violate dignity (e.g., subjecting one to torture) and to interfere with one's important liberties (e.g., freedom of speech or freedom of religion) without raising concerns of privacy. It is also possible to violate privacy (by watching or by eavesdropping) without limiting the person's liberty. The need to stress this distinction between privacy and liberty is one of the main reasons for defining privacy in terms of control over information, or, in the somewhat broader analysis suggested by Gavison (1980), of accessibility. According to this approach, the hard core of concerns and interests covered by the label 'privacy' is related to the interest of individuals in limiting access to themselves. Privacy can be lost through the invasion of a person's physical space by either presence, wiretapping, or clandestine supervision; the acquisition and dissemination of information about a person; or the 'lifting' of a person's anonymity, which in turn may encourage loss of privacy in either of these other ways. Privacy in this sense is thus distinct from liberty, but it is connected to liberty and to dignity in a variety of ways. These connections to liberty and to 'inviolate personality,' to be discussed below, are in part responsible for the importance of privacy as a central value in modern life.

The need to distinguish privacy from a general interest in liberty is challenged by those who identify issues such as abortion, contraception, and choice of sexual expression as matters of privacy. While scholars like Rubenfield (1989) and DeCew (1997) concede that these claims concern liberty, they argue that the special strength of these claims derives from the highly personal and intimate nature of these decisions, and that these features of the decisions are what makes them matters of privacy. This argument, which is supported by our linguistic and moral intuitions, is based on an important ambiguity in the term 'privacy' and the related distinction between 'private' and 'public.' There is a sense in which the 'private' is that which is intimate and personal. But 'private' is also that which is held, or used, or owned, or financed by individuals (as compared to that which is held, used, or financed by the public or at least by some group). In addition, 'private' also means that which one is entitled to keep secret or inaccessible, but also that which the individual is entitled to claim as belonging to her own realm of autonomous decision-making. Abortions and sexual practices are both matters of great intimacy and personal import and matters that most people desire to keep unknown and unpublicized. In addition, the principle of offense suggests that we may prohibit the public performance of activities that are fully legitimate when conducted 'in private' (e.g., sexual intercourse).

Against this background, any attempt to provide one definition of privacy and to reject others must be arbitrary to some extent. Those who use the term in descriptive or normative statements should be aware of these ambiguities and clarify the sense(s) in which they use the terms. In the remainder of this essay, only concerns of privacy in terms of accessibility and information will be discussed. Issues of personal autonomy, important as they are, are better discussed under different headings. The importance of stressing the distinctive definition of privacy has grown as modern technology has created a variety of new challenges to privacy. Modes of surveillance, eavesdropping, photographing, and processing of such information, as well as the development of computerized data banks with huge abilities of storage, retrieval, and processing and linking of information, and the growing possibilities of quick dissemination of such information, both verbal and pictorial, have made the threats of loss of privacy more serious and

acute. It is thus important to connect these urgent practical concerns with the reasons for privacy's importance.

2. The Importance of Privacy

While most individuals seek privacy, at least some of the time and for some purposes, it is not clear whether this wish for privacy should always be protected or granted. Some scholars point out that privacy, at least in the sense of keeping personal information unknown, is simply the scoundrel's best friend (notably, Posner 1978). While these scholars may concede the coherence of the concept of privacy, they argue that privacy is in fact undesirable, encouraging tendencies to individualism and downplaying the importance of community and accountability. Some feminist writers argue that treating privacy as an ideal is just a way to perpetuate the exclusion of women from public life, and to facilitate their continued exploitation and vulnerability within the realm of the private (notably MacKinnon 1989 but see also Gavison 1993).

Proponents of privacy (see, e.g., Bloustein 1964, Westin 1967, Gavison 1980, Fried 1968, Benn 1971) do not argue that all wishes to limit access to information about one or of one's realm should be respected. They may well concede that total privacy is both impossible and highly undesired and undesirable. They do insist, however, that interests in privacy are important, and that privacy is related, and contributes, to other important values such as freedom, autonomy, and intimacy. These values are an integral part of our ideals of personhood and society, and our normative landscape will be improverished if we do not recognize both the uniqueness and the importance of privacy claims.

It is easier to see the many ways privacy functions in our lives if we try to imagine a world in which we have lost all the privacy we enjoy at present. Imagine the loss in terms of intimacy, friendship, trust, and love that we would have suffered if we could not choose in whom to confide, and with whom to share personal information about ourselves. Or if we had known that whatever we said to any individual would immediately be broadcast on the Internet. Imagine the loss in learning and acquisition of skills we would have suffered if all our drafts and early attempts at mastering a skill had been made public, open to the judgment and ridicule of all. Imagine the loss in terms of self-respect and autonomy that would ensue from a life in which we would not have the benefit of experimenting and trying things alone, or in a protected and supportive environment, which could teach us and give us the courage to face audiences less supportive. Imagine democracy without a secret ballot and the freedom to develop ideas and associations without constant police surveillance, or rehabilitation without the ability to start afresh. It is no accident that people have sought privacy from the inception of human society, and that extensive loss of privacy (of the kind that is suffered in total institutions like prisons or mental hospitals) is regarded by many as degrading and humiliating.

Even if this argument is conceded, and people accept that some claims to privacy should be morally protected, this does not say that claims to privacy or limitation of accessibility are absolute. Like all other claims, they must be balanced against other important interests and rights. The importance of privacy means only that we should be aware that the costs we pay in terms of privacy are important, and that we should try, as far as possible, to decrease them. While total loss of privacy is dangerous and bad, some claims of privacy are merely self-serving and instrumental. Naturally, a candidate for a job does not want her prospective employer to know her weaknesses. A con-man does not want his victims to know of his past record. In all these cases we may well prefer the interest in knowledge and thus refuse to defend the claim of secrecy or privacy. As a matter of fact, in many situations of conflict, privacy gives way to other claims. Some scholars, have concluded from this fact that it is therefore misleading to talk about a 'right to privacy' in morality or in law (Prosser 1960, Thomson 1975, Davis 1959) They argue that privacy claims can never trump other claims on their own. The cases in which we feel that claims of privacy have the upper hand are ones in which there is another interest involved, such as the interest in reputation, or in physical seclusion, or in property. It is the presence of these other interests, they argue, not the claim of privacy, that explains (and justifies) the fact that we find the privacy argument persuasive.

A closer analysis of cases of conflict, in both mortality and law, does not support this generalization. Norms defending the rights of prominent gay people not to be 'outed,' and laws prohibiting the publication of rape victims' names, are two examples in which claims of privacy are seen as morally justifying, on their own, the limitation of the freedom of others to disseminate information. Moreover, even if it is the case that in many situations privacy claims win only when they are reinforced by other interests, this does not mean that the costs in terms of privacy are not important. It may mean only that claims of privacy, on their own, cannot defeat some conflicting claims. It is misleading to conclude that the costs in terms of privacy are nonexistent.

3. Legal Protection of Privacy and its Limits

Part of the debate around the concept of privacy and the justification for granting a right to privacy was triggered by claims for legal protection against various sorts of losses of privacy. One of the consequences of the conceptual ambiguity mentioned above is that the

extension of the term 'privacy' may be different in different legal systems. Thus, the initial decisions concerning the use of contraceptives and abortion in the US were made in terms of privacy (Grisworld 1965 and Roe 1973). While the US Supreme Court has moved to discuss abortion as an issue of liberty rather than one of privacy (Planned Parenthood 1992), there are still scholars seeking to justify dealing with issues, such as the liberty to behave in certain intimate matters, in terms of privacy (Rubenfeld 1989, DeCew 1997, Allen 1988) . Other scholars see these decisions as based on conceptual and normative confusion between issues of liberty and issues of privacy (Gross 1971, Henkin 1974, Gavison 1980).

Legal systems vary a great deal in their positive protection of privacy. Most Western countries now try to regulate computerized data banks in some ways. The legal frameworks seek to make sure that data is accurate, that individuals know what information about them is held in such data banks, and that acquisition and dissemination of information are controlled. The first two issues, while of extreme practical importance, are not—strictly speaking—privacy concerns. My privacy is not enhanced by the fact that I know what information about me is available to others. In fact, such knowledge may threaten my sense of privacy, because it may reveal to me that I am in fact more exposed than I had thought. Most legal systems include in their protection of property rights the liberty and the power to exclude others from entrance. Most legal systems now contain regulation of practices such as wiretapping and electronic eavesdropping. While the details of legal regulation differ, all systems seek to reduce abuses resulting from the existence of extensive data banks. The law has been slow to respond to the implication of technological innovations in the realm of dissemination of information. Cable TV and satellite technologies, as well as the Internet, create fascinating and difficult problems for imposing legal liability for the dissemination of information, including privacy-violating publications.

Legal systems differ also in the extent to which they are willing to grant legal protection to individuals' concerns that true information about them not be published against their will. In the European continent, legal systems protect this interest through a 'right to personality.' England still has no recognition of a general right to privacy, and no legal liability for such publication, while in the US a right to privacy is recognized, but it usually gives way to the paramount interest in free speech. Finally, the question of the extent to which public figures are entitled to the protection of their privacy remains controversial, with different answers given by different legal systems and cultures. The US seems to take the lead in denying its public figures their right to privacy, and in a willingness not to stop short of explicit and detailed public discussion of their most intimate affairs. However, this area of privacy claims illustrates that, very frequently, social realities are not dictated only by the law. The actual privacy enjoyed by public figures in Western countries depends on cultural and journalistic norms much more than on the legal arrangements existing in them.

Violations of privacy, especially in the field of dissemination and publication of personal information, also illustrate the limits of law in effective protection of values. Once the information is published, or the embarrassing picture aired, there is nothing that can delete it from memory or consciousness. Money awards cannot undo the harm, and the publicity of the trial just adds to the magnitude of the original injury. We encounter here a paradox of law in some areas: Laws may, to some extent, serve as the basis of injunctions, or they may deter people from violations. If they fail in doing that, they are unlikely to be invoked to impose subsequent liability.

See also: Confidentiality and Statistical Disclosure Limitation; Information Society; Personal Privacy: Cultural Concerns; Privacy of Individuals in Social Research: Confidentiality; Public Sphere and the Media; Telecommunications and Information Policy

Bibliography

Allen A 1988 *Uneasy Access: Privacy for Women in a Free Society*, Totowa, NJ
Benn S 1971 Privacy, freedom and respect for persons. Nomos **13**: 1
Benn S Gaus G (eds.) 1983 *The Private and Public in Social Policy*, NY
Bloustein E 1964 Privacy as an aspect of human dignity. **39** *New York University Law Review* 962
Davis F 1959 What do we mean by the right to privacy? *South Dakota Law Review*. **4**
DeCew J 1997 *In Pursuit of Privacy: Law, Ethics and the Rise of Technology*. Cornell
Ferdinand S (ed.) 1984 *Philosophical Dimensions of Privacy*, Cambridge University Press, Cambridge, UK
Fried C 1968 Privacy, *Yale Law Journal* **77**: 475
Gavison R 1980 Privacy and the limits of law. *Yale Law Journal* **89**: 421
Gavison R 1993 Feminism and the private–public distinction. *Stanford Law Review* **45**: 1
Griswold vs. Connecticut 1965 381 US 479
Gross H 1971 Privacy and autonomy. *Nomos* **13**: 169
Henkin L 1974 Privacy and autonomy. *Columbia Law Review* **74**: 1410
MacKinnon C 1989 *Towards a Feminist Theory of the State*. Cambridge, MA
Parent W 1983 Privacy, morality and the law. *Law and Philosophy* **2**: 269–88
Pennock R, Chapman J (eds.) 1971 Privacy, *Nomos* **13**
Planned Parenthood vs. Casey 1992 505 US 833
Posner R 1978 The right to privacy. *Georgia Law Review*. **12**: 393
Prosser W 1960 'Privacy', *California Law Review* **48**: 383
Roe vs. Wade 1973 410 US 113

Rubenfield J 1989 The right of privacy. *Harvard Law Review* **102**: 73
Thomson, J J 1975 The right of privacy. *Philosophy and Public Affairs* **4**: 295–314
Westin Allan 1967 *Privacy and Freedom*. NY

R. Gavison

Privacy of Individuals in Social Research: Confidentiality

1. *Definitions*

The word privacy here refers to the state of the individual, including individuals' right to control information about them at times. Confidentiality refers to the state of information, that is, information in an individual's record may not be disclosed without the individual's consent except under certain conditions. Security here refers to the state of physical or electronic records, especially protection against unauthorized access to the record.

2. *An Ethical and Legal Context*

People may provide information in social surveys, randomized field trials, and in other studies. Their doing so may be voluntary. Or it may be mandatory, that is, required by law.

Generally, the researcher is under an explicit ethical obligation to assure the privacy of the individual, the confidentiality of information generated on the individual, and the security of the record. These obligations may be reflected in professional codes of ethics. The obligations may be verified through government regulations and review processes that foster assurance of privacy, such as Institutional Review Boards in the United States. Or, the obligation may be verified in law. Statutes that govern the activity of census bureau interviewers and statisticians in countries, such as Sweden, the United States, Canada, and elsewhere, are plain on this account and include penalties for failure to meet the obligation.

3. *Assessing Privacy in Surveys and Randomized Field Trials*

The social scientist's statutes must be met in a variety of ways at each stage in the process of a study. At the study's design stage, for instance, one minimizes the information that is asked of a respondent and minimizes the intrusiveness of questions, that is, one restricts questions so as to meet the study's object. One also minimizes the number of people who are engaged in a study. This is determined partly by technical concerns such as the level of statistical precision or statistical power that is needed.

Under many codes of ethics and certain broad legal regulations, individuals to whom questions are put must be told whether their participation in the research is voluntary. People usually must be told the purpose of the research, and the disposition and functions of the data that they provide. Usually, they are assured that the information they provide will remain confidential, that is, the assurance is that information will be used only for research purposes and that the information they provide will not be disclosed with their identity attached to anyone else.

Under certain circumstances, the information provided by individuals can be assured of confidentiality even in the face of government or private threat. In the United States, for example, researchers who are involved in studies on alcohol and substance abuse, and criminological research can receive a certificate of confidentiality. The rules are such that researchers are prohibited under penalty of law from disclosing identifiable microrecords, and are legally immune from any obligation to provide microrecords to law enforcement agencies.

More generally, statues that govern statistical agencies in the United States and other countries require that the agencies refrain from disclosing records on individuals to other people, that is, the statutes say that records on identifiable individuals cannot be appropriated legally by anyone else, including the courts, legislatures, and law enforcement agencies.

The researcher is under an ordinary obligation to cooperate with courts and other government agencies in providing information. This obligation becomes problematic if the researcher made promises of confidentiality and then, when confronted with a government demand, discovers that the promise has no legal standing. Keeping track of such episodes and their frequency is important. Shield laws that assure that the social scientist cannot be compelled to disclose information about identifiable individuals who have agreed to participate in the research are then important.

4. *Privacy and Researchers' Use of Statistical Data Bases*

Producing high quality data in social research can be costly irrespective of the benefits of the information. To the extent that the same data set can be put to work frequently to benefit the public, the cost/effectiveness ratio decreases.

For instance, good scholars undertook multiple reanalysis of data from the statewide class size randomized trial in Tennessee. These analyses informed

policy debates about reducing the size of classrooms in 20 states of the United States.

Making local, regional, or national data available to statisticians, however, may put one ethical standard against another: public interest in the results of statistical analysis versus the possibility that records will be used for nonresearch purposes.

Learning how to make individual records available for statistical analysis in ways that do not compromise the privacy of the individuals who are the source of the record information is not easy. Since the 1970s, serious research on this topic has been done by federal statistical agencies in many different countries. Virtually all such agencies concern themselves with surveys rather than active randomized trials on the effects of social programs. Nonetheless, the lessons from the survey research community are applicable to people who attempt to mount evaluations of program effects, such as randomized field trials.

The approaches that have been developed to assure researchers' access to existing data sets without compromising privacy concerns can be classified into two categories. The first involves simple procedures that focus on assuring that records cannot be used for learning about an individual. The second involves legal arrangements, including statutes, regulations, and contractual agreements.

Consider, first, the approaches to making data available and protecting individual privacy. Typically, the data custodian who releases a file for the public's use deletes the direct and clear identifiers from data files, that is, the addresses and other unique information about individuals are deleted. Further, because individual records are often detailed and external collateral information is available, sophisticated statistical approaches may be used to eliminate the possibility of deductive disclosure. Such approaches include deletion of certain variables from the file, eliminating certain microrecords that allow disclosure on small samples of particular individuals, reducing the precision of entry by making the reporting categories crude, introducing controlled random perturbation, controlled random swapping, and other techniques. These technical approaches fall under the rubric of statistical disclosure limitation technology. Advances in this technology by statisticians in the Netherlands, the United Kingdom, the United States, Israel, and Luxemburg are given in Fienberg and Willenborg (1998).

Data enclaves are organizations created to (a) permit intensive sophisticated analyses by people other than the original data custodian, and (b) protect the confidentiality of microrecords with the same rules that govern the original custodian. At the University of California at Los Angeles and at Berkeley, for instance, enclaves have been established that permit university-based researchers to analyze data that would be accessible ordinarily only to federal civil servants. The rules that govern the data's use, including security procedures and the threat of criminal and civil sanctions, apply to the enclave's researcher as it does to civil servants.

A variant on the enclave idea involves licensing agreements. A specific example of this arrangement, developed by the National Center for Education Statistics, is a license granted to particular individuals to analyze specific data sets that would not otherwise be accessible on privacy grounds. The justifications for such privilege must be made. The licensing agreement includes requirements for security of the microrecords, sanctions and rules or disclosure, and so on.

5. *Privacy and the Use of State Data Bases*

In the United States, each state of the 50 states has laws governing third party access to certain administrative record systems and administrative data bases. The state laws differ appreciably in their provisions for permitting a researcher (and certain others) to access individually identifiable microrecords. The administrative records that are laws and regulations also differ across systems within schools, hospitals, child welfare and adult welfare agencies, crime and justice agencies, runaway shelters, and others.

There exists no national compendium for identifying which administrative record systems are legally open to legitimate researchers and which systems are closed to them. A compendium for each country and for each political jurisdiction and agency within each country is important as a way of understanding how privacy of individuals and confidentiality of microrecords can be assured in a research context.

Despite the absence of a national, regional, and local compendium, there exist useful compendiums that have been constructed for special purposes. They identify the State and its laws on particular record systems and summarize access rules for certain individuals including perhaps researchers (see, for instance, the Office of Juvenile Justice and Delinquency Research 1995 on missing children and the relevant databases and the National Criminal Justice Information and Statistics Service (1978, 1988, 1992) on criminal history information. Among federal agencies in the United States, the Bureau of Justice Statistics (BJS) has been consistent in preparing and updating related information. It has produced eight reports since 1972; more recent reports supplant earlier ones. They all focus on state laws regarding criminal history record information.

In the BJS (1992) compendium, each state's laws are cited, the full text of laws being given in a national repository that is itself accessible to researchers. The compendium outlines information about each state law on 28 dimensions. One such dimension is whether the particular law contains 'provision for and regulation of' access to criminal justice information by outside researchers (BJS 1992, p. 12).

The compendium identifies 29 states that have laws or regulations that permit research access and the relevant statute or regulation. No analysis is provided. But researchers can obtain full text documents for the statutes and regulations from the criminal justice clearinghouse.

6. *Privacy and Record Linkage*

Linking records from different databases in the interest of research presents privacy problems. The rules governing one data custodian, such as a census bureau, may differ from the rules of another statistical agency such as an education statistics or a labor statistics bureau. Such rules differ across agencies for a less centralized government such as the United States. They may not differ across agencies for a more centralized government such as Canada, Sweden, or Japan. The matter is very complex and generalizations are hard to make.

Certainly, the case for linkage has to be made. Identifying the discernable benefits to society of a linkage and the absence of discernable costs to individual privacy is important. This can be done through example and systematic study. An early example of an effort to discover the benefits and costs of linkage for administrative purposes, not research purposes, is given in USGAO (1986a, 1986b). The bottom line was that certain administrative linkages are not worth the effort. Recent examples, more to the point for social scientists, are given in the work of the National Research Council (1999).

See also: Confidentiality and Statistical Disclosure Limitation; Ethical Dilemmas: Research and Treatment Priorities; Research Conduct: Ethical Codes; Research Ethics, Cross-cultural Dimensions of; Research Subjects, Informed and Implied Consent of

Bibliography

Boruch R F 199 *Randomized Experiments for Planning and Evaluation*. Sage, Thousand Oaks, CA

Boruch R F, Cecil J S 1979 *Assuring the Confidentiality of Social Research Data*. University of Pennsylvania Press, Philadelphia, PA

Bureau of Justice Statistics 1992 *Criminal History Record Information: Compendium of State Privacy and Security Legislation, 1992, Current Status by Category and State*. US Department of Justice, Washington, DC

Cecil J S, Wetherington G T (eds.) 1996 Court ordered disclosure of academic research: A clash of values of science and law. *Law and Contemporary Problems* **59**(3): 1–7

Donaldson M S, Lohr K (eds.) 1994 *Health Data in an Information Age: Use, Disclosure, and Privacy*. National Academy Press, Washington, DC

Duncan G T, Jabine T, deWolf V A (eds.) 1993 *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academy Press, Washington, DC

Fienberg T, Willenborg L C R J (eds.) 1999 Disclosure limitation methods for protecting the confidentiality of statistical data. Journal of Official Statistics **14**(4)

Fox J A, Tracey P E 1986 *Randomized Response*. Sage, Beverly Hills, CA

National Research Council 1999 Committee on National Statistics Workshop on Confidentiality of and Access to Research Files, October 14–15, 1999 Work Shop Papers. National Research Council, Washington, DC

Sigma Xi, The Scientific Research Society. *The Responsible Research: Paths and Pitfalls*. Sigma Xi, Research Triangle Park, NC

Soecken K L 1987 Randomized response methodology in health research. *Evaluation Health Professions* **10**(1): 58–66

Takas M, Bass D (eds.) 1995 *Addressing Confidentiality of Records in Searches for Missing Children: Final Report*. Office of Juvenile Justice and Delinquency Prevention, Washington, DC

US General Accounting Office 1986a *Computer Matching: Factors Influencing the Agency Decision-making Process* (PEMD – 87 –3 BR). USGAO, Washington, DC

US General Accounting Office 1986b *Computer Matching: Assessing Its Costs and Benefits* (PEMD –87 –2). USGAO, Washington, DC

R. F. Boruch

Privatization of the Economy

State ownership has characterized most market and, *a fortiori*, planning economies in the twentieth century, the United States being the most notable, albeit partial, exception. To an extent that has somewhat varied across countries and decades, financial, strategic, political, and technological reasons have been advanced to justify either the nationalization of existing private-sector firms or the *ex novo* creation of state-owned enterprises (SOEs). State ownership of utilities has been justified on the ground of technological conditions leading to natural monopolies, external economies, and diverging social and private discount rates. Moreover, many governments argued that only the state could overcome critical 'bottlenecks' in terms of physical, financial, and human capital investments required by industrialization. A further motive has been the desire to keep domestic control on monopoly rents produced exploiting nonrenewable natural resources. On weaker economic grounds, SOEs have been established to build a coalition in support of new regimes, to assure national independence in strategic industries, or to avoid that unpopular minorities could eventually control the economy. SOEs were used for different distributional

aims—to make available essential goods and services, to create new jobs, to reduce geographical concentration of economic power—but they were also crucial elements for playing patronage politics through jobs and the servicing of constituencies. A further explanation for the diffusion of state ownership in semiperipheral countries, especially in Latin America, was advanced by the dependency literature. While accepting the primacy of private ownership, 'bureaucratic-authoritarian regimes' considered that the role of the state had to go beyond that of a subsidiary character to that of supplying the complementary inputs to the process of private capital accumulation (Evans 1978).

Since the rise of oil prices in the 1970s, the whole postwar model of economic development has been heavily criticized, and SOEs have not been exempted from this swing in the ideological pendulum from the structuralist to the neo-liberal end. The appropriateness of public management of productive units was put in doubt, and SOEs were found to have negative income distribution effects, to worsen budget and trade balances, and to divert resources from much more needed social goals, such as education and health expenditures. The efficiency of SOEs was found to be lower *ceteris paribus* than in the private sector. Inefficient SOEs do not face the risk of bankruptcy, and the market for corporate control, because of the state's tight grip through majority stakes, cannot act as an adequate device for disciplining managers. In theory, this concentration of ownership in the hands of a single majority investor could have served to circumvent the collective action problem that impedes small, dispersed shareholders in listed companies to wield efficient monitoring on corporate managers. This relative advantage of public ownership, however, has often been offset by agency problems stemming from the multiplicity of ties linking the government, ministries with specific competencies, Parliament, political parties, and the management of SOEs, all with different goals, possibly at cross-purpose. Moreover, SOEs' managers are often appointed for political reasons, rather than for their management skills. As regards the utilities more specifically, fast technological advances have substantially weakened the argument that public ownership of integrated monopolies is necessary to assure investment growth, service quality, and price declines in real terms.

Over the last two decades, the initially timid progress recorded in some pioneering countries (United Kingdom, Chile, New Zealand) were matched and sometimes even surpassed by new countries joining the privatization bandwagon, first in the developing world (Argentina, Mexico, Malaysia), then in Europe (France, Italy), and finally in transition economies. (Possibly the first large-scale sale of a public enterprise was the Volkswagen issue in the early 1960s. Nonetheless, the dramatic fall of the share price after markets were hit by the Cuban missile crisis had a lasting negative impact on equity ownership culture in Germany.) Yet, while a central issue in the debate for almost two decades, privatization has been a rather restricted phenomenon from a quantitative point of view (Shirley 1998). Selling SOEs is part of the political game: it requires the creation of supportive coalitions and it impacts on the distribution of power resources. Strong vested interests, such as organized labor, the military, parts of the urban bourgeoisie, and most economic groups, mount in defense of nationalization. Moreover, privatization is a typical example of a collective action situation: actors more likely to gain from it are widely dispersed, and it is thus simpler for beneficiaries of the status quo to organize an opposition than it is for future beneficiaries to create a support group. Finally, in developing countries where capital markets are thin and scarcely organized, and where it is thus difficult to realize popular capitalism through public placing of SOEs, the potentialities of privatization in terms of coalition building are rather remote.

In practice, the intrinsic allure of free-market economics has proven less important a factor for the success of privatizations than more conjunctural conditions (Armijo 1998). The nature and the seriousness of crisis deeply influence both the perceived need of radical changes and the central themes of the political debate. The tighter the fiscal constraint, and the greater the contribution of SOEs to the deficit, the quicker the recourse to divestiture to solve problems. An important issue is therefore the choice of firms to be sold and the method of sale. Understanding the economic goals to be met through divestiture, in particular the choice between an immediate reduction of macroeconomic imbalances and the medium-term achievement of efficiency improvements by liberalization and more effective regulation, is crucial. The least productive enterprises are harder to dispose of, which might make it necessary to concede easy terms well beyond a low price, when not liquidating them altogether. At the same time, countries in dire economic straits need quick 'leading cases' to build their reputation *vis-à-vis* foreign investors. The choice has thus often been to start from the best managed and more profitable SOEs, such as telecommunications, at the risk of making learning mistakes in those very cases where state assets should be more coveted by international investors. Moreover, such industries present technological and tariff complexities, requiring sophisticated tools for asset valuation and regulation that developing countries seldom possess.

Timing and organizational structure also matter. New political incumbents have higher chances of implementing reforms, thanks not only to the 'post-election honeymoon,' but also because, according to the electoral cycle theory, there are less incentives to put off adopting risky and painful actions. Likewise, although the relationship between regimes and econ-

omic policies is indeterminate, being able to use emergency powers opens smoother roads on the reforms' path. 'Insulated change teams' or 'technopols' are important, especially where governments came to power with the backing of the lower class and need to prove themselves credible in the face of international investors' fears. However, not all technopols are alike, and the priority they assign to privatization (and the resources they demand politicians to earmark for this as opposed to other structural measures) also depends on their ideological values and on their appraisal of their country's riddle.

Third, given the short-term adjustment costs of economic reforms, it is necessary to introduce compensatory mechanisms. While scholarly attention has been focused on specific, targeted measures to face the reduction of fiscal subsidies to the poor, large business groups that dominate the industrial sector in most developing countries are hit by falling trade barriers, positive real interest rates, and more careful awarding of public credit. In order to win their support, some countervailing measures may be needed, such as cheap (relative to book value) sales of SOEs (Schamis 1998).

Studies of privatization in industrial countries have reported very positive results in firms' performance (e.g., Meggison and Netter 1999). As for developing economies, there are significant increases in profitability, operating efficiency, capital spending, output, and employment, which are usually greater in countries with higher per capita income (Boubakri and Cosset 1998). (In the case of transition economies, empirical tests of the relationship between enterprise performance and ownership generally refute the hypothesis that privatization *per se* is associated with improved performance (e.g., Estrin and Rosevear 1999). The belief that mass privatization, by providing powerful incentives for efficient restructuring, would release entrepreneurial endeavors has also proved naif (e.g., Spicer et al. 1999).) The substitution of private to public ownership, however, has been accompanied by the emergence of new regulatory challenges for policy-makers. First, in the hitherto natural monopolies that constituted the core of public enterprises (telecoms, energy, water, transport), how to prevent the new private owners from simply pocketing monopoly rents. Especially in the electricity industry, weaknesses in the regulatory framework have sometimes reduced the benefits of privatization and deregulation (Nicoletti 2000). Public authorities must devise sectoral policies that introduce and maintain competition; establish and maintain a sound regulatory framework for the remaining monopolies, public and private; keep transparency in transactions and convince investors that their investments are secure; negotiate, monitor, and enforce contracts with private suppliers of management and financing; ensure that resources from privatization sales are put to productive uses; and manage the inevitable political and social tensions that arise as enterprise reforms are implemented, especially the critical issues of foreign ownership and labor layoffs. There is, however, much less agreement on how to approach the next set of challenges (second-generation issues) for countries facing the consolidation of initial reforms. In general, these issues are related to postprivatization disputes and renegotiations between governments and the private sector and to the mechanisms necessary to promote competition in the reformed industries.

Second, in view of the goal of creating a people's capitalism, how to provide appropriate mechanisms of corporate governance that protect small shareholders, while also allowing management the required flexibility to pursue long-term corporate goals. Advances in the theory of corporate finance and industrial organization, as well as the ongoing debate over the 'best' capitalist model, have made the analysis of corporate governance—the mechanisms whereby economic systems (in their broadest possible sense) cope with the information and incentive problems inherent in financing investments and facilitate the intertemporal transfer of income claims—a burgeoning theme in comparative institutional economics. Key elements of this literature are the sources of financial resources for corporations, the concentration of ownership, the relevance of listing, the role and composition of boards of directors, the rules governing the market for corporate control and the obstacles that they may pose to corporate control activity, and the relative importance of the voice and exit mechanisms (boardroom pressures and takeovers, or internal and external control, respectively) in disciplining managers.

Links between privatization and corporate governance are of two main kinds: selling SOEs exposes them to the takeover and bankruptcy threats, thereby easing the corporate governance problems proper of public ownership, and it provides an opportunity to modify the distribution of ownership rights among different classes of investors by extending public listing among large firms, increasing the number of small shareholders, and reducing ownership concentration. But while creating 'people's capitalism' has always ranked high among governments' goals, tackling the issues that remain after ownership is transferred from public to private hands and when no (absolute) majority shareowner emerges has seldom been an overriding concern. Potential improvements in technical efficiency following control transfer may be jeopardized if corporate control is not contestable: in this case, and especially if conduct regulation proves insufficient to open up protected markets, managers can exploit rents accruing from market position without having to worry about the threat of takeovers. Privatization policies should take such elements into account, making it imperative to introduce reforms and redress perceived inefficiencies. Governments have great discretion in pricing the SOEs they sell, especially those being sold via public share offering, and they use

this discretion to pursue political and economic ends. Most experiences, however, suggest the limited power of privatization in changing the modes of governance which are prevalent in each country's large private companies. Further, those countries which have chosen the mass (voucher) privatization route have done so largely out of necessity and face ongoing efficiency problems as a result. In the UK, a country whose privatization policies are often referred to as a benchmark, 'control [of privatized companies] is not exerted in the forms of threats of take-over or bankruptcy; nor has it for the most part come from direct investor intervention' (Bishop et al. 1994, p. 11). After the steep rise experienced in the immediate aftermath of privatizations, the slow but constant decline in the number of small shareholders highlights the difficulties in sustaining people's capitalism in the longer run. In Italy, for example, privatization was accompanied by a legislative effort aimed at providing noncontrolling shareholders (i.e. both individual and collective investors) with more adequate safeguards and at introducing the necessary conditions to allow them to monitor managers. But successive governments were unsuccessful in broadening the number of large private business groups, whereas enhancing the mobility of control to investors outside the traditional core of Italy's capitalism was explicitly included among the authorities' strategic goals. On the other hand, experiences such as those of the UK and Chile underscore that mass sell-offs require the development of new institutional investors such as pension funds that may later play an active role in corporate governance. (In Chile, for example, the takeover of the country's dominant electricity utility, Enersis, one of the largest in emerging markets, was stalled for some months in 1998 as pension funds' questioned lucrative additional terms that the management had negotiated for themselves based on important agreements concerning the future strategic direction of Enersis that they had never told other shareholders about.)

To conclude, although privatization's promise has been frequently oversold, not least by international organizations, its ills have also been greatly exaggerated. When ownership transfer has been accompanied by market liberalization and proper implementation, in OECD and non-OECD countries alike, consumers and end-users have benefited in terms of choice, quality, and prices. If public opinion and policy-makers wish so, much remains to be sold and the future challenges for the regulatory state remain substantial to ensure maximization of welfare benefits.

See also: Business Law; Capitalism: Global; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; Law and Economics; Law and Economics: Empirical Dimensions; Liberalism; Markets and the Law; Monetary Policy; Multinational Corporations; Regulation, Economic Theory of; Regulation: Working Conditions; Smith, Adam (1723–90); Socialism; Venture Capital

Bibliography

Armijo L 1998 Balance sheet or ballot box? Incentives to privatize in emerging democracies. In: Oxhorn P, Starr P (eds.) *The Problematic Relationship between Economic and Political Liberalization*. Lynne Rienner, Boulder, CO

Bishop M, Kay J, Mayer C 1994 Introduction: privatization in performance. In: Bishop M, Kay J, Mayer C (eds.) *Privatization and Economic Performance*. Oxford University Press, Oxford, UK

Boubakri N, Cosset J-C 1998 The financial and operating performance of newly privatized firms: evidence from developing countries. *Journal of Finance* **53**: 1081–110

Estrin S, Rosevear A 1999 Enterprise performance and ownership: The case of Ukraine. *European Economic Review* **43**: 1125–36

Evans P 1978 *Dependent Development*. Princeton University Press, Princeton, NJ

Meggison W, Netter J 1999 From state to market: a survey of empirical studies on privatization. *Fondazione Eni Enrico Mattei Working Paper*, No. 01-99

Nicoletti G 2000 The implementation and the effects of regulatory reform: past experience and current issues. *Mimeo*. OECD Economics Department, Paris

Schamis H M 1998 Distributional coalitions and the politics of economic reform in Latin America. *World Politics* **51**: 236–68

Shirley M M 1998 Bureaucrats in business: the roles of privatization versus corporatization in state-owned enterprise reform. *World Development* **27**: 115–36

Spicer A McDermott G Kogut B 1999 Entrepreneurship and privatization in Central Europe: the tenuous balance between destruction and creation. The Wharton School. *Reginald H. Jones Center Working Paper*, No. 99-04

A. E. Goldstein

Probabilistic Grammars and their Applications

1. Introduction

Natural language processing is the use of computers for processing natural language text or speech. Machine translation (the automatic translation of text or speech from one language to another) began with the very earliest computers (Kay et al. 1994). Natural language interfaces permit computers to interact with humans using natural language, for example, to query databases. Coupled with speech recognition and speech synthesis, these capabilities will become more important with the growing popularity of portable computers that lack keyboards and large display

screens. Other applications include spell and grammar checking and document summarization. Applications outside of natural language include compilers, which translate source code into lower-level machine code, and computer vision (Fu 1974, 1982).

Most natural language processing systems are based on formal grammars. The development and study of formal grammars is known as computational linguistics. A *grammar* is a description of a language; usually it identifies the sentences of the language and provides descriptions of them, for example, by defining the phrases of a sentence, their interrelationships, and perhaps also aspects of their meanings. *Parsing* is the process of recovering a sentence's description from its words, while *generation* is the process of translating a meaning or some other part of a sentence's description into a grammatical or well-formed sentence. Parsing and generation are major research topics in their own right. Evidently, human use of language involves some kind of parsing and generation process, as do many natural language processing applications. For example, a machine translation program may parse an input language sentence into a (partial) representation of its meaning, and then generate an output language sentence from that representation.

Modern computational linguistics began with Chomsky (1957), and was initially dominated by the study of his 'transformational' grammars. These grammars involved two levels of analyses, a 'deep structure' meant to capture more-or-less simply the meaning of a sentence, and a 'surface structure' which reflects the actual way in which the sentence was constructed. The deep structure might be a clause in the active voice, 'Sandy saw Sam,' whereas the surface structure might involve the more complex passive voice, 'Sam was seen by Sandy.'

Transformational grammars are computationally complex, and in the 1980s several linguists came to the conclusion that much simpler kinds of grammars could describe most syntactic phenomena, developing Generalized Phrase-Structure Grammars (Gazdar et al. 1985) and Unification-based Grammars (Kaplan and Bresnan 1982, Pollard and Sag 1987, Shieber 1986). These grammars generate surface structures directly; there is no separate deep structure and therefore no transformations. These kinds of grammars can provide very detailed syntactic and semantic analyses of sentences, but even today there are no comprehensive grammars of this kind that fully accommodate English or any other natural language.

Natural language processing using handcrafted grammars suffers from two major drawbacks. First, the syntactic coverage offered by any available grammar is incomplete, reflecting both our lack of understanding of even relatively frequently occuring syntactic constructions and the organizational difficulty of manually constructing any artifact as complex as a grammar of a natural language. Second, such grammars almost always permit a large number of *spurious ambiguities*, that is, parses which are permitted by the rules of syntax but have unusual or unlikely semantic interpretations. For example, in the sentence 'I saw the boat with the telescope,' the prepositional phrase 'with the telescope' is most easily interpreted as the instrument used in seeing, while in 'I saw the policeman with the rifle,' the prepositional phrase usually receives a different interpretation in which the policeman has the rifle. Note that the corresponding alternative interpretation is marginally accessible for each of these sentences: in the first sentence one can imagine that the telescope is on the boat, and in the second, that the rifle (say, with a viewing scope) was used to view the policeman.

In effect, there is a dilemma of coverage. A grammar rich enough to accommodate natural language, including rare and sometimes even 'ungrammatical' constructions, fails to distinguish natural from unnatural interpretations. But a grammar sufficiently restricted so as to exclude what is unnatural fails to accommodate the scope of real language. These observations led, in the 1980s, to a growing interest in stochastic approaches to natural language, particularly to speech. Stochastic grammars became the basis of speech recognition systems by outperforming the best of the systems based on deterministic handcrafted grammars. Largely inspired by these successes, computational linguists began applying stochastic approaches to other natural language processing applications. Usually, the architecture of such a stochastic model is specified manually, while the model's parameters are estimated from a *training corpus*, that is, a large representative sample of sentences.

As explained in the body of this article, stochastic approaches replace the binary distinctions (grammatical vs. ungrammatical) of nonstochastic approaches with probability distributions. This provides a way of dealing with the two drawbacks of nonstochastic approaches. Ill-formed alternatives can be characterized as extremely low probability rather than ruled out as impossible, so even ungrammatical strings can be provided with an interpretation. Similarly, a stochastic model of possible interpretations of a sentence provides a method for distinguishing more plausible interpretations from less plausible ones.

The introduction sets out formally various classes of grammars and languages. Probabilistic grammars are introduced in Sect. 2, along with the basic issues of parametric representation, inference, and computation.

2. Grammars and Languages

The formal framework, whether used in a transformational grammar, a generalized phrase-structure grammar, or a more traditionally styled context-free grammar, is due to Chomsky (1957) and his co-

workers. This section presents a brief introduction to this framework. But for a thorough (and very readable) presentation we highly recommend the book by Hopcroft and Ullman (1979).

The concept and notation of a formal grammar is perhaps best introduced by an example:

Example 1: Define a *grammar*, G_1, by $G_1 = (T_1, N_1, S, R_1)$, where $T_1 = \{\text{grows, rice, wheat}\}$ is a set of words (a 'lexicon'), $N_1 = \{\text{S, NP, VP}\}$ is a set of symbols representing grammatically meaningful strings of words, such as clauses or parts of speech (e.g., S for 'Sentence,' NP for 'Noun Phrase,' VP for 'Verb Phrase'), and $R_1 = \{\text{S} \to \text{NP VP}, \text{NP} \to \text{rice}, \text{NP} \to \text{wheat}, \text{VP} \to \text{grows}\}$ is a collection of rules for rewriting, or instantiating, the symbols in N_1. Informally, the nonterminal S rewrites to sentences or clauses, NP rewrites to noun phrases and VP rewrites to verb phrases. The *language*, L_{G_1}, generated by G_1 is the set of strings of words that are reachable from S through the rewrite rules in R_1. In this example, $L_{G_1} = \{\text{rice grows, wheat grows}\}$, derived by $\text{S} \Rightarrow \text{NP VP} \Rightarrow \text{rice VP} \Rightarrow \text{rice grows}$, and $\text{S} \Rightarrow \text{NP VP} \Rightarrow \text{wheat VP} \Rightarrow \text{wheat grows}$.

More generally, if T is a finite set of symbols, let T^* be the set of all strings (i.e., finite sequences) of symbols of T, including the empty string, and let T^+ be the set of all nonempty strings of symbols of T. A language is a subset of T^*. A *rewrite grammar* G is a quadruple $G = (T, N, S, R)$, where T and N are disjoint finite sets of symbols (called the *terminal* and *nonterminal* symbols respectively), $S \in N$ is a distinguished nonterminal called the *start* or *sentence symbol*, and R is a finite set of productions. A *production* is a pair (α, β) where $\alpha \in N^+$ and $\beta \in (N \cup T)^*$; productions are usually written $\alpha \to \beta$. Productions of the form $\alpha \to \varepsilon$, where ε is the empty string, are called *epsilon productions*. This entry restricts attention to grammars without epsilon productions, that is, $\beta \in (N \cup T)^+$, as this simplifies the mathematics considerably.

A rewrite grammar G defines a *rewriting relation* $\Rightarrow_G \subseteq (N \cup T)^* \times (N \cup T)^*$ over pairs of strings consisting of terminals and nonterminals as follows: $\beta A \gamma \Rightarrow \beta\alpha\gamma$ iff $A \to \alpha \in R$ and $\beta, \gamma \in (N \cup T)^*$ (the subscript G is dropped when clear from the context). The reflexive, transitive closure of $\Rightarrow$ is denoted $\Rightarrow^*$. Thus $\Rightarrow^*$ is the rewriting relation using arbitrary finite sequences of productions. (It is called 'reflexive' because the identity rewrite, $\alpha \Rightarrow \alpha$, is included.) The language generated by G, denoted L_G, is the set of all strings $w \in T^+$ such that $S \Rightarrow^* w$.

Rewrite grammars are traditionally classified by the shapes of their productions. $G = (T, N, S, R)$ is a *context-sensitive* grammar if for all productions $\alpha \to \beta \in R$, $|\alpha| \leqslant |\beta|$, that is, the right-hand side of each production is not shorter than its left-hand side. G is a *context-free* grammar if $|\alpha| = 1$, that is, the left-hand side of each production consists of a single non terminal. G is a *left-linear* grammar if G is context-free and β (the right-hand side of the production) is either of the form Aw or of the form w where $A \in N$ and $w \in T^*$; in a *right-linear* grammar β always is of the form wA or w. A right or left-linear grammar is called a *regular* grammar. G_1, in Example 1, is context sensitive and context free.

It is straightforward to show that the classes of languages generated by these classes of grammars stand in strict equality or subset relationships. Specifically, the class of languages generated by right-linear grammars is the same as the class generated by left-linear grammars; this class is called the *regular* languages, and is a strict subset of the class of languages generated by context-free grammars, which is a strict subset of the class of languages generated by context-sensitive grammars, which in turn is a strict subset of the class of languages generated by rewrite grammars.

It turns out that context-sensitive grammars (where a production rewrites more than one nonterminal) have not had many applications in natural language processing, so from here on we will concentrate on context-free grammars, where all productions take the form $A \to \beta$, where $A \in N$ and $\beta \in (N \cup T)^+$.

An appealing property of grammars with productions in this form is that they induce tree structures on the strings that they generate. And, as Sect. 3 shows, this is the basis for bringing in probability distributions and the theory of inference. We say that the context-free grammar $G = (T, N, S, R)$ generates the labeled, ordered tree ψ, iff the root node of ψ is labeled S, and for each node n in ψ, either n has no children and its label is a member of T (i.e., it is labelled with a terminal) or else there is a production $A \to \beta \in R$ where the label of n is A and the left-to-right sequence of labels of n's immediate children is β. It is straightforward to show that w is in L_G if G generates a tree ψ whose *yield* (i.e., the left-to-right sequence of terminal symbols labeling ψ's leaf nodes) is w; ψ is called a *parse tree* of w (with respect to G). In what follows, we define Ψ_G to be the set of parse trees generated by G, and $\mathscr{Y}(\cdot)$ to be the function that maps trees to their yields.

Example 1 (continued): The grammar G_1 defined above generates the following two trees, ψ_1 and ψ_2.

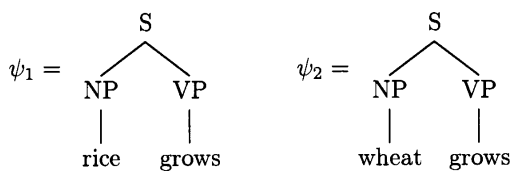

In this example, $\mathscr{Y}(\psi_1) =$ 'rice grows' and $\mathscr{Y}(\psi_2) =$ 'wheat grows.'

A string of terminals w is called *ambiguous* if w has two or more parse trees. Linguistically, each parse tree of an ambiguous string usually corresponds to a distinct interpretation.

Example 2: Consider $G_2 = (T_2, N_2, S, R_2)$, where $T_2 = \{$I, saw, the, man, with, telescope$\}$, $N_2 = \{$S, NP, N, Det, VP, V, PP, P$\}$ and $R_2 = \{$S $\rightarrow$ NP, VP, NP $\rightarrow$ I, NP $\rightarrow$ Det N, Det $\rightarrow$ the, N $\rightarrow$ NP, PP, N $\rightarrow$ man, N $\rightarrow$ telescope, VP $\rightarrow$ V, NP, VP $\rightarrow$ VP, PP, PP $\rightarrow$ P, NP, V $\rightarrow$ saw, P $\rightarrow$ with$\}$. Informally, N rewrites to nouns, Det to determiners, V to verbs, P to prepositions and PP to prepositional phrases. It is easy to check that the two trees ψ_3 and ψ_4 with the yields $\mathcal{Y}(\psi_3) = \mathcal{Y}(\psi_4) =$ 'I saw the man with the telescope' are both generated by G_2. Linguistically, these two parse trees represent two different syntactic analyses of the sentence. The first analysis corresponds to the interpretation where the seeing is by means of a telescope, while the second corresponds to the interpretation where the man has a telescope.

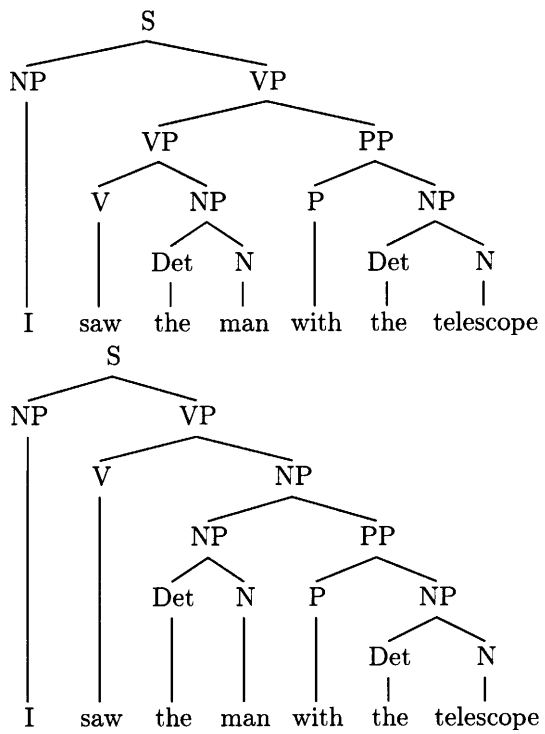

3. Probability and Statistics

Obviously broad coverage is desirable—natural language is rich and diverse, and not easily held to a small set of rules. But it is hard to achieve broad coverage without massive ambiguity (a sentence may have tens of thousands of parses), and this of course complicates applications like language interpretation, language translation, and speech recognition. This is the dilemma of coverage that we referred to earlier, and it sets up a compelling role for probabilistic and statistical methods.

We briefly review the main probabilistic grammars and their associated theories of inference. We begin in Sect. 3.1 with probabilistic regular grammars, also known as hidden Markov models (HMM), which are the foundation of modern speech recognition systems—see Jelinek (1997) for a survey.

In Sect. 3.2 we discuss probabilistic context-free grammars, which turn out to be essentially the same thing as branching processes. Finally, in Sect. 3.3, we take a more general approach to placing probabilities on grammars, which leads to Gibbs distributions, a role for Besag's pseudolikelihood method (Besag 1974, 1975), various computational issues, and, all in all, an active area of research in computational linguistics.

3.1 Regular Grammars

We will focus on right-linear grammars, but the treatment of left-linear grammars is more or less identical. It is convenient to work with a *normal form*: all rules are either of the form $A \rightarrow bB$ or $A \rightarrow b$, where $A, B \in N$ and $b \in T$. It is easy to show that every right-linear grammar has an equivalent normal form in the sense that the two grammars produce the same language. Essentially nothing is lost.

3.1.1 Probabilities. The grammar G can be made into a probabilistic grammar by assigning to each non terminal $A \in N$ a probability distribution p over productions of the form $A \rightarrow \alpha \in R$: for every $A \in N$

$$\sum_{\substack{\alpha \in (N \cup T)+ \\ \text{s.t. } (A \rightarrow \alpha) \in R}} p(A \rightarrow \alpha) = 1 \tag{1}$$

Recall that Ψ_G is the set of parse trees generated by G (see Sect. 2). If G is linear, then $\psi \in \Psi_G$ is characterized by a sequence of productions, starting from S. It is, then, straightforward to use p to define a probability P on Ψ_G: just take $P(\psi)$ (for $\psi \in \Psi_G$) to be the product of the associated production probabilities.

Example 3: Consider the right-linear grammar $G_3 = (T_3, N_3, S, R_3)$, with $T_3 = \{$a, b$\}$, $N_3 = \{$S, A$\}$ and the productions (R_3) and production probabilities (p):

S$\rightarrow$aS	$p = 0.80$
S$\rightarrow$bS	$p = 0.01$
S$\rightarrow$bA	$p = 0.19$
A$\rightarrow$bA	$p = 0.90$
A$\rightarrow$b	$p = 0.10$

The language is the set of strings ending with a sequence of at least two *b*s. The grammar is ambiguous: in general, a sequence of terminal states does not uniquely identify a sequence of productions. The sentence *aabbbb* has three parses (determined by the placement of the production S $\rightarrow$ bA), but the most

likely parse, by far, is S → aS, S → aS, S → bA, A → bA, A → bA, A → b ($P = 0.8 \cdot 0.8 \cdot 0.19 \cdot 0.9 \cdot 0.1$), which has a *posterior* probability of nearly 0.99. The corresponding parse tree is shown below.

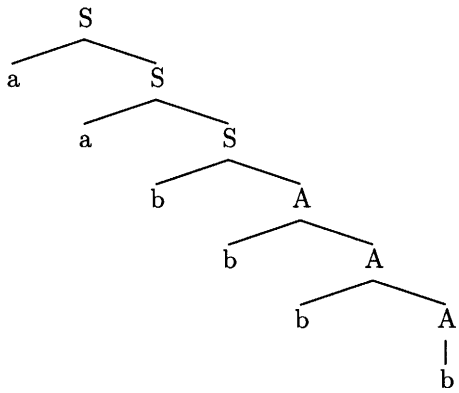

3.1.2 Inference. The problem is to estimate (see *Estimation: Point and Interval*) the transition probabilities, $p(\cdot)$, either from parsed data (examples from Ψ_G) or just from sentences (examples from L_G). Consider first the case of parsed data ('supervised learning'), and let $\psi_1, \psi_2, \ldots, \psi_n \in \Psi$ be a sequence taken iid according to P. If $f(A \to \alpha; \psi)$ is the counting function, counting the number of times transition $A \to \alpha \in R$ occurs in ψ, then the likelihood function (see *Likelihood in Statistics*) is

$$L = L(p; \psi_1, \ldots, \psi_n) = \prod_{i=1}^{n} \prod_{A \to \alpha \in R} p(A \to \alpha)^{f(A \to \alpha; \psi_i)} \tag{2}$$

The maximum likelihood estimate is, sensibly, the relative frequency estimator:

$$\hat{p}(A \to \alpha) = \frac{\sum_{i=1}^{n} f(A \to \alpha; \psi_i)}{\sum_{i=1}^{n} \sum_{A \to \beta \in R, \beta s.t} f(A \to \beta; \psi_i)} \tag{3}$$

The problem of estimating p from sentences ('unsupervised learning') is more interesting, and more important for applications. Recall that $\mathcal{Y}(\psi)$ is the yield of ψ, that is. the sequence of terminals in ψ. Given a sentence $w \in T^+$, let Ψ_w be the set of parses which yield w: $\Psi_w = \{\psi \in \Psi: \mathcal{Y}(\psi) = w\}$. Imagine a sequence $\psi_1, \ldots, \psi_n$, iid according to P, for which only the corresponding yields, $w_i = \mathcal{Y}(\psi_i)$ $1 \leqslant i \leqslant n$, are observed. The likelihood function is

$$L = L(p; w_1, \ldots, w_n) = \prod_{i=1}^{n} \sum_{\psi \in \Psi_{w_i}} \prod_{A \to a \in R} p(A \to \alpha)^{f(A \to \alpha; \psi_i)} \tag{4}$$

As is usual with hidden data, there is an EM-type iteration for climbing the likelihood surface—see Baum (1972) and Dempster et al. (1977):

$$t+1(A \to \alpha) = \frac{\sum_{i=1}^{n} E_{\hat{p}_t}[f(A \to \alpha; \psi) | \psi \in \Psi_{w_i}]}{\sum_{A \to \beta \in R} \sum_{i=1}^{n} E_{\hat{p}_t}[f(A \to \beta; \psi) | \psi \in \Psi_{w_i}]} \tag{5}$$

Needless to say, nothing can be done with this unless we can actually evaluate, in a computationally feasible way, expressions like $E_{\hat{p}}[f(A \to \alpha; \psi) | \psi \in \Psi_w]$. This is one of several closely related computational problems that are part of the mechanics of working with regular grammars.

3.1.3 Computation. A sentence $w \in T^+$ is *parsed* by finding a sequence of productions $A \to bB \in R$ which yield w. Depending on the grammar, this corresponds more or less to an interpretation of w. Often, there are many parses and we say that w is ambiguous. In such cases, if there is a probability p on R then there is a probability P on Ψ, and a reasonably compelling choice of parse is the most likely parse:

$$\arg\max_{\psi \in \Psi_w} P(\psi) \tag{6}$$

This is the *maximum a posteriori* (MAP) estimate of ψ—obviously it minimizes the probability of error under the distribution P. (Of course, in those cases in which Eqn. (6) is small, P does little to make w unambiguous.)

What is the probability of w? How are its parses computed? How is the most likely parse computed? These computational issues turn out to be more-or-less the same as the issue of computing $E_{\hat{p}}[f(A \to \alpha; \psi) | \psi \in \Psi_w]$ that came up in our discussion of inference. The basic structure and cost of the computational algorithm is the same for each of the four problems—compute the probability of w, compute the set of parses, compute the best parse, compute $E_{\hat{p}}$. In particular, there is a simple dynamic programming solution to each of these problems, and in each case the complexity is of the order $n \cdot |R|$, where n is the length of w, and $|R|$ is the number of productions in G—see Jelinek (1997), Geman and Johnson (2000). The existence of a dynamic programming principle for regular grammars is a primary reason for their central role in state-of-the-art speech recognition systems.

3.2 Context-free Grammars

Despite the successes of regular grammars in speech recognition, the problems of language understanding and translation are generally better addressed with the more structured and more powerful context-free grammars. Following our development of probab-

ilistic regular grammars in the previous section, we will address here the interrelated issues of fitting context-free grammars with probability distributions, estimating the parameters of these distributions, and computing various functionals of these distributions.

The context-free grammars $G = (T, N, S, R)$ have rules of the form $A \to \alpha$, $\alpha \in (N \cup T)^+$, as discussed previously in Sect. 2. There is again a normal form, known as the Chomsky normal form, which is particularly convenient when developing probabilistic versions. Specifically, one can always find a context-free grammar G', with all productions of the form $A \to BC$ or $A \to a$, A , B, $C \in N$, $\alpha \in T$, which produces the same language as G: $L_{G'} = L_G$. Henceforth, we will assume that context-free grammars are in the Chomsky normal form.

3.2.1 Probabilities. The goal is to put a probability distribution on the set of parse trees generated by a context-free grammar in Chomsky normal form. Ideally, the distribution will have a convenient parametric form, that allows for efficient inference and computation.

Recall from Sect. 2 that context-free grammars generate labeled, ordered trees. Given sets of nonterminals N and terminals T, let Ψ be the set of finite trees with:

(a) root node labeled S;
(b) leaf nodes labeled with elements of T;
(c) interior nodes labeled with elements of N;
(d) every nonterminal (interior) node having either two children labeled with nonterminals or one child labeled with a terminal.

Every $\psi \in \Psi$ defines a sentence $w \in T^+$: read the labels off of the terminal nodes of ψ from left to right. Consistent with the notation of Sect. 3.1 we write $\mathscr{Y}(\psi) = w$. Conversely, every sentence $w \in T^+$ defines a subset of Ψ, which we denote by Ψ_w, consisting of all ψ with yield w ($\mathscr{Y}(\psi) = w$). A context-free grammar G defines a subset of Ψ, Ψ_G, whose collection of yields is the language, L_G, of G. We seek a probability distribution P on Ψ which concentrates on Ψ_G.

The time-honored approach to probabilistic context-free grammars is through the production probabilities $p : R \to [0, 1]$, with

$$\sum_{\substack{\alpha \in N^2 \cup T \\ \text{s.t.}(A \to \alpha) \in R}} p(A \to \alpha) = 1 \tag{7}$$

Following the development in Sect. 3.1, we introduce a counting function $f(A \to \alpha;\ \psi)$, which counts the number of instances of the rule $A \to a$ in the tree ψ, i.e., the number of nonterminal nodes A whose daughter nodes define, left-to-right, the string α. Through f, p induces a probability P on Ψ:

$$P(\psi) = \prod_{(A \to \alpha) \in R} p(A \to \alpha)^{f(A \to \alpha; \psi)} \tag{8}$$

It is clear enough that P concentrates on Ψ_G, and we shall see shortly that this parameterization, in terms of products of probabilities p, is particularly workable and convenient. The pair, G and P, is known as a probabilistic context-free grammar, or PCFG for short. (Notice the connection to branching processes—Harris (1963): Starting at S, use R, and the associated probabilities $p(\cdot)$, to expand nodes into daughter nodes until all leaf nodes are labeled with terminals—i.e., with elements of T.)

3.2.2 Inference. As with probabilistic regular grammars, the production probabilities of a context-free grammar, which amount to a parameterization of the distribution P on Ψ_G, can be estimated from examples. In one scenario, we have access to a sequence $\psi_1, \ldots, \psi_n$ from Ψ_G under P. This is 'supervised learning,' in the sense that sentences come equipped with parses. More practical is the problem of 'unsupervised learning,' wherein we observe only the yields, $\mathscr{Y}(\psi_1), \ldots, \mathscr{Y}(\psi_n)$.

In either case, the treatment of maximum likelihood estimation is essentially identical to the treatment for regular grammars. In particular, the likelihood for fully observed data is again Eqn. (2), and the maximum likelihood estimator is again the relative frequency estimator Eqn. (3). And, in the unsupervised case, the likelihood is again Eqn. (4) and this leads to the same EM-type iteration given in Eqn. (5).

3.2.3 Computation. There are again four basic computations: find the probability of a sentence $w \in T^+$, find a $\psi \in \Psi$ (or find *all* $\psi \in \Psi$) satisfying $Y(\psi) = w$ ('parsing'); find

$$\arg\max_{\substack{\psi \in \Psi \text{ s.t.} \\ \mathscr{Y}(\psi) = w}} P(\psi)$$

('maximum *a posteriori*' or 'optimal' parsing); compute expectations of the form $E_{\hat{p}_t}[f(A \to \alpha;\ \psi) | \psi \in \Psi_w]$ that arise in iterative estimation schemes like (5). The four computations turn out to be more-or-less the same, as was the case for regular grammars (Sect. 3.1.3) and there is again a common dynamic-programming-like solution—see Lari and Young (1990, 1991), Geman and Johnson (2000).

3.3 Gibbs Distributions

There are many ways to generalize. The coverage of a context-free grammar may be inadequate, and we may hope, therefore, to find a workable scheme for placing probabilities on context-sensitive grammars, or perhaps even more general grammars. Or, it may be preferable to maintain the structure of a context-free grammar, especially because of its dynamic programming principle, and instead generalize the class of probability distributions away from those induced (parameterized) by production probabilities. But

nothing comes for free. Most efforts to generalize run into nearly intractable computational problems when it comes time to parse or to estimate parameters.

Many computational linguists have experimented with using Gibbs distributions, popular in statistical physics, to go beyond production-based probabilities, while nevertheless preserving the basic context-free structure. Next we take a brief look at this particular formulation, in order to illustrate the various challenges that accompany efforts to generalize the more standard probabilistic grammars.

3.3.1 Probabilities. The sample space is Ψ_G, the set of trees generated by a context-free grammar G. Gibbs measures are built from sums of more-or-less simple functions, known as 'potentials' in statistical physics, defined on the sample space. In linguistics, it is more natural to call these features rather than potentials. Suppose, then, that we have identified M linguistically salient features $f_1, \ldots, f_M$, where f_k: $\Psi_G \to R$, through which we will characterize the fitness or appropriateness of a structure $\psi \in \Psi_G$. More specifically, we will construct a class of probabilities on Ψ_G which depend on $\psi \in \Psi_G$ only through $f_1(\psi), \ldots, f_M(\psi)$. Examples of features are the number of times a particular production occurs, the number of words in the yield, various measures of subject-verb agreement, and the number of embedded or independent clauses.

Gibbs distributions have the form

$$P_\theta(\psi) = \frac{1}{Z} \exp \left\{ \sum_{i=1}^{M} \theta_i f_i(\psi) \right\} \tag{9}$$

where $\theta_1, \ldots, \theta_M$ are parameters, to be adjusted 'by hand' or inferred from data, $\theta = (\theta_1, \ldots, \theta_M)$, and where $Z = Z(\theta)$ (known as the 'partition function') normalizes so that $P_\theta(\Psi) = 1$. Evidently, we need to assume or ensure that $\sum_{\psi \in \Psi_G} \exp \{\sum_1^M \theta_i f_i(\psi)\} < \infty$. For instance, we had better require that $\theta_1 < 0$ if $M = 1$ and $f_1(\psi) = |\mathcal{Y}(\psi)|$ (the number of words in a sentence), unless of course $|\Psi_G| < \infty$.

3.3.2 Relation to probabilistic context-free grammars. Gibbs distributions are much more general than probabilistic context-free grammars. In order to recover PCFG's, consider the special feature set $\{f(A \to \alpha;\ \psi)\}_{A \to \alpha \in R}$: The Gibbs distribution Eqn. (9) takes on the form

$$P_\theta(\psi) = \frac{1}{Z} \exp \left\{ \sum_{A \to \alpha \in R} \theta_{A \to \alpha} f(A \to \alpha;\ \psi) \right\} \tag{10}$$

Evidently, then, we get the probabilistic context-free grammars by taking $\theta_{A \to \alpha} = \log_e p(A \to \alpha)$, where p is a system of production probabilities consistent with Eqn. (7), in which case $Z = 1$. But is Eqn. (10) more general? Are there probabilities on Ψ_G of this form that are not PCFGs? The answer turns out to be no, as was shown by Chi (1999) and Abney et al. (1999): given a probability distribution P on Ψ_G of the form of Eqn. (10), there always exists a system of production probabilities p under which P is a PCFG.

3.3.3 Inference. The feature set $\{f_i\}_{i=1,\ldots,M}$ can be accommodate arbitrary linguistic attributes and constraints, and the Gibbs model Eqn. (9), therefore, has great promise as an accurate measure of linguistic fitness. But the model depends critically on the parameters $\{\theta_i\}_{i=1,\ldots,M}$, and the associated estimation problem is, unfortunately, very hard. Indeed, the problem of unsupervised learning appears to be all but intractable.

Let us suppose, then, that we observe $\psi_1, \psi_2, \ldots, \psi_n \in \Psi_G$ ('supervised learning'), iid according to P_θ. In general, the likelihood function, $\prod_{i=1}^n P_\theta(\psi_i)$, is more or less impossible to maximize. But if the primary goal is to select good parses, then perhaps the likelihood function asks for too much, or even the wrong thing. It might be more relevant to maximize the likelihood of the observed parses, *given the yields* $\mathcal{Y}(\psi_1), \ldots, \mathcal{Y}(\psi_n)$:

$$\prod_{i=1}^{n} P_\theta(\psi_i | \mathcal{Y}(\psi_i)) \tag{11}$$

Maximization of Eqn. (11) is an instance of Besag's remarkably effective pseudolikelihood method (Besag 1974, 1975), which is commonly used for estimating parameters of Gibbs's distributions. The computations involved are generally much easier than what is involved in maximizing the ordinary likelihood function. Take a look at the gradient of the logarithm of Eqn. (11): the θ_j component is proportional to

$$\frac{1}{n} \sum_{i=1}^{n} f_j(\psi_i) - \frac{1}{n} \sum_{i=1}^{n} E_0 \left[f_j(\psi) | \mathcal{Y}(\psi) = \mathcal{Y}(\psi_i) \right] \tag{12}$$

and $E_\theta[f_j(\psi)|\mathcal{Y}(\psi)]$ can be computed directly from the set of parses of the sentence $\mathcal{Y}(\psi)$. (In practice there is often massive ambiguity, and the number of parses may be too large to feasibly consider. Such cases require some form of pruning or approximation.)

Thus gradient ascent of the pseudolikelihood function is (at least approximately) computationally feasible. This is particularly useful since the Hessian of the logarithm of the pseudolikelihood function is nonpositive, and therefore there are no local maxima. What is more, under mild conditions pseudolikelihood estimators (i.e., maximizers of Eqn. (11)) are consistent; (Chi 1998).

Bibliography

Abney S, McAllester D, Pereira F 1999 Relating probabilistic grammars and automata. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* Morgan Kaufmann, San Francisco, pp. 542–609

Baum L E 1972 An inequality and associated maximization techniques in statistical estimation of probabilistic functions of Markov processes. *Inequalities* **3**: 1–8
Besag J 1974 Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series D* **36**: 192–236
Besag J 1975 Statistical analysis of non-lattice data. *The Statistician* **24**: 179–95
Chi Z 1998 *Probability models for complex systems*. PhD thesis, Brown University
Chi Z 1999 Statistical properties of probabilistic context-free grammars. *Computational Linguistics* **25**(1): 131–60
Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague
Dempster A, Laird N, Rubin D 1977 Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**: 1–38
Fu K S 1974 *Syntactic Methods in Pattern Recognition*. Academic Press, New York
Fu K S 1982 *Syntatic Pattern Recognition and Applications*. Prentice-Hall, Englewood Cliffs, NJ
Gazdar C, Klein E, Pullum G, Sag I 1985 *Generalized Phrase Structure Grammar*. Blackwell, Oxford, UK
Geman S, Johnson M 2000 Probability and statistics in computational linguistics, a brief review. Internal publication, Division of Applied Mathematics, Brown University
Harris T E 1963 *The Theory of Branching Processes*. Springer, Berlin
Hopcroft J E, Ullman J D 1979 *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Reading, MA
Jelinek F 1997 *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA
Kaplan R M, Bresnan J 1982 Lexical functional grammar: A formal system for grammatical representation. In: Bresnan J (ed.) *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA, Chap. 4, pp. 173–281
Kay M, Gawron J M, Norvig P 1994 *Verbmobil: A Translation System for Face-to-face Dialog*. CSLI Press, Stanford, CA
Lari K, Young S J 1990 The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language* **4**: 35–56
Lari K, Young S J 1991 Applications of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language* **5**: 237–57
Pollard C, Sag A 1987 *Information-based Syntax and Semantics*. CSLI Lecture Notes Series No.13. University of California Press, Stanford, CA
Shieber S M 1986 *An Introduction to Unification-based Approaches to Grammar*. CSLI Lecture Notes Series. University of California Press, Stanford, CA

S. Geman and M. Johnson

Probabilistic Thinking

1. Introduction

Some people have enjoyed playing games that involved at least a component of chance ever since civilizations—as we know them—emerged in human history. Many of these people apparently had a fairly good intuitive idea of what we might now term the 'probabilities' or 'odds' of various outcomes associated with these games. Alternatively, at least some people (whom in hindsight we might now label 'superstitious') believed that the outcomes of these games could indicate something about the favorability of the world in general toward the gambler (a belief that may exist even today—at least on an implicit basis—for chronic gamblers), specifically the favorability or unfavorability of a particular god that was involved with a particular game.

In the last 600 years or so people have proposed systematic ways of evaluating probabilities—e.g., in terms of equally likely outcomes in games of chance or in terms of subjective belief. In the last 150 years or so scientific endeavor has started to make specific use of probabilistic thinking, and in the last 100 years or so probabilistic ideas have been applied to problems of everyday life. (see Gigerenzer et al. 1989). Currently, for example, jurors are often asked to determine manufacturer or company liability on the basis of differential rates of accidents or cancers; in such judgments, jurors cannot rely on deterministic reasoning that would lead to a specification of exactly which negative outcome is due to what; nor can they rely on 'experience' in these contexts in which they have had none. Thus, it is extraordinarily important to understand—if indeed generalizations about such thinking are possible—how it is that people think about and evaluate probabilities, and in particular the problems that they (we) have. The major thesis presented here is that both the thinking and the problems can be understood in terms of *systematic deviations* from the normative principles of probabilistic reasoning that have been devised over the past five or so centuries. These deviations are not inevitable; it is certainly possible for people to think about probabilities coherently (at least on occasion); nevertheless, *when* deviations occur, they tend to be of the systematic nature to be described here; they occur most often when people are engaged in what is termed 'automatic' or 'associational' thinking, as opposed to 'reflective' and 'comparative' thinking. (These terms will be made more precise later in this article.)

2. A Brief History (of Thought)

Beginning in early infancy, we develop intuitions about space, time, and number (at least to the extent of 'one, two, three, many'). These intuitions are essentially correct. They remain throughout our lives basically unaltered except for small variations, and they form the foundation of far more sophisticated mathematical systems that have been developed over the past thousands of years. Moreover, their importance in our lives has been clear throughout those millennia.

In contrast, probabilistic thinking is often unintuitive, and it generally can be taught only after we are familiar with fractions, so that analogs to relative frequency are clear. Moreover, our intuitions about probabilities are often just plain wrong. For example, belief in the 'gambler's fallacy' (negative recency) can even be explicitly rejected, but nevertheless implicitly influence decision making—such as when people demand a greater amount of money in preference to gambling on an outcome that is 'due' (e.g., a tail after five straight heads) than in preference to gambling on one that isn't (e.g., an additional head); see Gold (1997).

The fairly recent, and somewhat 'iffy,' use of probabilistic judgment does not imply, however, that people did not gamble throughout the centuries. As mentioned previously, the outcomes of games of chance could provide not just monetary winners and losers but 'signals' from 'the gods' about favorability toward the gambler; this belief is well captured by the narrative of the novelist Robert Graves (1943) about Caligula's gambling shortly before his assassination. This tyrant is both watching gladiator games, often putting to death people he has bribed to lose, and rolling a form of dice called 'astragali'—which are four-sided dice made from pig or lamb bones.

Caligula is playing with his Uncle Claudius, who in order to avoid winning (a dangerous outcome for Claudius) hands the emperor beautiful new dice loaded to come up with a 'Venus roll,' which consists of four separate sides (generally with the numbers 1, 3, 4, and 6). Caligula's subsequent success convinces him that Venus is favorable toward him that day, and he does not take the usual precautions upon leaving the games. Throughout his life, Claudius feels guilty that the loaded dice sent the wrong message to Caligula about how Venus felt about him that day—an inappropriate intervention between a god and a man. (According to Hacking 1975, the interpretation of such events as 'signals' from 'on high' is common up to the Renaissance, even after people in Europe had given up polytheism for monotheism; the reader might also note how often in the *Old Testament* people decide to 'cast lots' to decide what to do—apparently in the belief that God would determine the outcome.)

Of course, not everyone believed in such intervention. For example, Cicero is quoted as claming that 'Nothing is so unpredictable as a throw of the dice [modern translation],' and that 'both now and then indeed he will make it [the Venus roll] twice and even thrice in succession. Are we going to be so feeble-minded then as to aver that such a thing happened by the personal intervention of Venus rather than by pure luck?' (see David 1962, p. 25).

Among those not so superstitious (such as Cicero, who was himself assassinated), the intuitive understanding gained through experience of which bets are good ones and which bets are bad ones led to fairly good implicit estimates of probability. Even now, however, Wagenaar (1988) finds chronic gamblers to be more reliant on intuition and experience than on mathematical analysis.

In contrast, explicit probabilistic thinking goes beyond experience. We can, for example, indicate what are good or bad bets in gambles we have just now constructed, with which we have had no previous experience. Probabilistic thinking can also be extended to 'personal belief' about uncertain events, it can be used to evaluate risks, and it can serve as a guide for social policy.

Systematic, probabilistic thinking appears to have started with questions about gambling (in particular about gambles that have not yet been played multiple times). Despite the fact that people gambled a lot, however, the formal theory did not begin until the Italian Renaissance, specifically in correspondence and argumentation between Geronimo Cardano (1501–76) and his associates. The approach that has been ascribed to Cardano involved *comparative* rather than associational thinking. He proposed that the likelihood of success of a gamble could be *computed*, or at least approximated, by comparing the number of outcomes of the device favorable to the gambler divided by the total number of outcomes. In situations where the gambling device is set up to make each possible outcome equally likely, this ratio assessment of probability is used today. The ratio assessment leads to certain general principles, often termed 'axioms,' which in turn provide a more abstract notion of probability, and allow different interpretations of probability to be developed—just provided they lead to measures consistent with the axioms. For example, Ramsey (1931) demonstrated that when explicit probabilities are used to determine fair gambling odds, then the gambler will follow the basic principles of probability theory if and only if it is not possible to make a 'Dutch Book' (see de Finetti 1974–5, p. 87) such that the gambler always loses; that is, the gambler will be consistent if and only if it is impossible for a highly intelligent opponent to make a series of bets such that the opponent always wins.

After Cardano, the idea of probabilistic inference developed quite rapidly. Many French and Dutch intellectuals corresponded with each other about what constituted better or worse bets—not just in games that had been played chronically, but in games of their own invention. Interestingly, when they discussed probability, they often did so in terms of the expected value of the game, and questions were often posed as to which of the two games had a higher expected value. Then there was rapid development of probability theory—for example, Bernoulli's' Law of Large Numbers,' (see *Statistical Methods, History of: Pre-1900* and *Statistical Methods, History of: Post-1900*) which specified the relationship between the probability of a single outcome and the probability distribution of multiple outcomes of the same type (for example, the relationship between the probability a fair coin comes

up heads and the probability that it comes up heads at least three times in 10 trials); the 'law' itself derived the limit of the proportion of heads as the number of trials becomes indefinitely large. To jump ahead to recent times, von Neumann and Morgenstern published their classic work *Theory of Games and Economic Behavior* (1944/1947), in which again the question was one of choosing between different gambles, but in terms of expected utility rather than expected value. This work was quite consistent with the earliest work, in which the question of probability was secondary to the question of choice between alternative courses of action.

3. Probability Assessments

All probabilistic analysis is based on the idea that (suitably trained and intelligent) people can at least *recognize* good probabilistic arguments presented by someone else, or discovered or thought of by themselves, but not necessarily generate good assessments. The very fact that there was correspondence about the gambles—and occasionally some disputes about them—indicated that people do not automatically assess probabilities in the same way, or accurately (e.g., corresponding to relative frequencies, or making good gambling choices).

The von Neumann and Morgenstern work, however, does involve some *psychological* assumptions that people can engage in 'good' probabilistic thinking. First, they must do so implicitly, so that the choice follows certain 'choice axioms' that allow the construction of an expected utility model—i.e., a model that *represents* choice as a maximization of implicit expected utility; that in turn requires that probabilities at the very least follow the standard axioms of probability theory. It also implies considering conditional probabilities in a rational manner, which is done only when implicit or explicit conditional probabilities are consistent with Bayes' theorem. Thus, the von Neumann and Morgenstern work required that people be 'Bayesian' in a consistency sense, although that term is sometimes used to imply that probabilities should at base be interpreted as degrees of belief. Another way in which probability assessment must be 'good' is that there should be some at least reasonable approximation between probabilities and long-term relative frequencies; in fact, under particular circumstances (of interchangeability and indefinitely repeated observations), the probabilities of someone whose belief is constrained by Bayes' theorem must approximate relative frequencies (Dawid 1982).

Are people always coherent? Consider an analogy with swimming. People do swim well, but occasionally we drown. What happens is that there is particular systematic bias in attempting to swim that makes it difficult. We want to hold our heads above water. When, however, we raise our heads to do so, we tend to assume a vertical position in the water, which is one of the few ways of drowning (aside from freezing or exhaustion or being swept away in rough waters).

Just as people drown occasionally by trying to hold their heads above water, people systematically deviate from the rules of probabilistic thinking. Again, however, the emphasis is on 'systematic.' For example, there is now evidence that people's probabilistic judgments are 'subadditive'—in that when a general class is broken into components, the judgmentally estimated probabilities assigned to disjoint components that comprise the class sum to a larger number than the probability assigned to the class. That is particularly true in memory, where, for example, people may recall the frequency with which they were angry at a close friend or relative in the last month and the frequency with which they were angry at a total stranger, and the sum of the estimates is greater than an independent estimate of being angry period (even though it is possible to be angry at someone who is neither a close friend or relative nor a total stranger; see Mulford and Dawes 1999). The clever opponent will then bet against the occurrence of each component but on the occurrence of the basic event, thereby creating a Dutch Book.

4. The Heuristics and Biases Approach to Problems of Probabilistic Thinking

The basic conception is that we all use a series of 'heuristics' in our everyday estimation (explicit or implicit) of probabilities, and these heuristics lead to biases. These heuristics provide a type of 'bounded rationality' (Simon 1955, 1956, 1979)—in that they substitute for a very detailed or logical analysis, an analysis that may be beyond our 'computational capacity.' Just as associational thinking serves us well in many contexts, however, so do the heuristics. Occasionally, however, they lead to grossly erroneous—systematically erroneous—probabilistic thinking. The conception that heuristics have some type of 'generalized validity,' but that they often lead us astray, was well argued in a classic *Science* article by Tversky and Kahneman (1974); unfortunately, the principle has been lost on some critics, who can point out many situations in which these heuristics 'work.' Naturally, they do. If they didn't have some sort of generalized utility, they would not have survived over generations of people trying to make probabilistic judgments.

One of the most important of these heuristics is termed the 'representativeness' heuristic, which was first proposed by Tversky and Kahneman in an article concerning *Belief in the Law of Small Numbers* (1971). The idea supported in that article is that we expect very small samples to 'represent' the populations from

which they were drawn, where the 'we' includes even people with a fair amount of formal training in statistics. That expectation leads to judgment of probability on the basis of the degree to which an instance or a sample *matches* a population or general idea, rather than on a judgment of how it can be generated. For example, when tossing a coin, a result consisting of six heads appears to be much less likely than one of a head followed by two tails, followed by two heads and finally a tail. But of course each of the sequences is equally likely; each occurs with probability 1/64. In fact, the sequence of a head followed by two tails followed by two heads and a tail is seen as more likely than a sequence of strict alternation of heads and tails—even though in general people believe that there is more than a chance amount of alternation when they judge what is truly 'random' (and in fact are most apt to pick as random the sequence that has a transition probability of 2/3 rather than 1/2). This incorrect expectation of alternation leads to such erroneous beliefs as that in the 'hot hand' in basketball when observing clusters of similar outcomes (e.g., successful shots). See, for example, Gilovich et al. (1985).

Another heuristic is termed an availability one. Briefly, we estimate frequencies on the ease with which instances 'come to mind,' or on our judgment of the ease with which we can recall or generate instances. It is, in fact, generally true that when there are more instances as opposed to fewer in the world, we find it easier to recall or generate more as opposed to fewer. The problem is, however, that there are determinants of ease of recall or generation that are independent of—or sometimes even in conflict with—actual relative frequency in our environment. For example, vivid instances are more readily recalled than pallid ones. (See again the result on 'subadditivity' in memory.) Moreover, instances that are easily imaginable (even if we haven't actually experienced them) are more readily 'brought to mind' than are those difficult to imagine.

The availability heuristic creates a bias when we are unaware of the factors other than frequency that can result in ease of recall or generation, but use it as a way of estimating actual frequency. Sometimes, the bias is so strong that we even ignore instances that we *cannot* sample—a 'structural availability' problem. For example, clinical psychologists often claim that no child sexual abusers stop on their own without therapy. When asked to substantiate this claim, they will cite many of the child abusers with whom they have had contact, often mandated by a judge to enter therapy with them. The problem is, however, that they would not have had contact with any who had stopped on their own, and by definition they were seeing people only in a therapy setting, hence could not observe any of them stopping 'without therapy.' Nevertheless, the experience with such people is so salient and compelling that even while acknowledging the logical problem with the conclusion, many of these clinical psychologists simply revert to the belief a day or a week after having publicly recognized the availability bias (see Dawes 1988, p. 102).

Following the ideas of Fischhoff and Beyth-Marom (1983) as elaborated by Dawes (1998), the standard cognitive biases and heuristics can be understood by alluding to simple forms of Bayes' theorem. Consider, for example, the relationship between a symptom S and a disease D, and suppose a diagnostician observes this symptom S. If the probability of the disease is assessed on the basis of a pure matching or association between the symptom and the disease, independent of considerations of conditional probability, there is no normative structure to which the judgment corresponds. More often, however, the judgment will be made on the basis of the conditional probabilities—a normatively correct judgment if it is based on the conditional probability of the disease given the symptom, P(D/S), but a representative judgment if it is the probability of the symptom given the disease, P(S/D). Unfortunately, there is a lot of evidence that the judgment is often made on the basis of the latter relationship—when conditional probabilities are considered at all.

The relationship between these two probabilities is given by rewriting Bayes' theorem as:

$$P(D/S) = \frac{P(S/D)P(D)}{P(S)} \tag{1}$$

It can also be written in terms of the 'ratio rule':

$$\frac{P(D/S)}{P(S/D)} = \frac{P(D)}{P(S)} \tag{2}$$

Because the ratio rule follows from the very definition of conditional probability, it is simply incoherent to equate the probability of the disease given the symptom with the probability of the symptom given the disease in the absence of a simultaneous belief that the probability of the disease and the probability of the symptom are identical. (Again, if someone were to make a series of bets based on 'fair betting odds' in such a belief, an opponent could make a Dutch Book against that person.) Classic representative thinking—e.g., 'she gave a typical schizophrenic response, therefore she must be schizophrenic'—embraces such an identity. What is critical, of course, are the *base rates* of D and of S, yet extensive research has shown that people underutilize such base rates in their intuitive judgments—or at least fail to incorporate them to a sufficient extent into their prior beliefs, given that they are often seen as 'not relevant.' (Why, people often ask, should the *general* probabilities of these symptoms and diseases be relevant to a particular judgment about a particular person with a particular symptom? Such an argument is sometimes even followed by

bland assertions to the effect that 'statistics do not apply to the individual'—a wholly erroneous conclusion, as can be attested to by the heavy cigarette smoker who develops lung cancer.)

People often do not ignore these base rates *entirely* in the judgments, particularly if the probabilistic nature of the judgment is made clear (see Gigerenzer et al. 1989). Nevertheless, the 'underutilization' of such base rates is ubiquitous. In the psychological literature it was first decried by Meehl and Rosen (1955), who noted that in case conferences they attended, clinical psychology and psychiatry judges tended to ignore base rates completely; later, underutilization has been established in a variety of experimental situations using a wide variety of subjects, especially by Kahneman and Tversky (1972, 1973). For example, when asked to judge whether a description is of a particular engineer or a particular lawyer, subjects make the same judgment whether the experimenter states that the description was drawn from a pool of 70 percent engineers and 30 percent lawyers, or vice versa. As Bar-Hillel (1990) points out, the 70/30 split does not appear to be 'relevant,' but it's relevance would most likely be clear if the split were 99/1, and certainly if it were a 100/0. (Again, we can hypothesize that people can recognize probabilistic arguments—although it's possible to be fooled on occasion—but often not make them spontaneously.) Thus, for example, someone who is said to be interested in sailing and carpentry and sharing model train sets with a child is judged to be much more likely to be an engineer than a lawyer, whether this person was drawn from a pool of 70 percent engineers or 70 percent lawyers. Making the sampling procedure absolutely 'transparent' does, however, reduce that tendency.

The use of Bayes' theorem to understand the relationship between symptoms and disease leads naturally to its use to consider the general relationship between any hypotheses and any bit of evidence. Let e refer to some bit of evidence about whether a hypothesis h is or is not true; for example, a symptom may be regarded as a bit of evidence, and having a particular disease a hypothesis; or the hypothesis may concern who is going to win the election in the United States in the year 2020 and a particular bit of evidence may be the margin of victory or defeat of the potential candidate in a more local election. Bayes' theorem expressed in terms of evidence and hypotheses is presented as

$$P(h/e) = \frac{P(e/h)P(h)}{P(e)} \tag{3}$$

This form of Bayes' theorem, while true, often leads to complications when trying to evaluate the probability of the evidence. There is, however, a way to avoid having to estimate that probability. What we can do is to consider *odds* that the hypothesis is true. These odds are the probability the hypothesis is true given the evidence divided by the probability that this hypothesis is false given the evidence—that is, by the ratio $P(h/e)/P(-h/e)$. Once we know the odds, it is trivial to compute the probability. The advantage of considering odds is that the denominator in Bayes' theorem cancels out when we compute these odds. That is

$$\frac{P(h/e)}{P(-h/e)} = \frac{P(e/h)P(h)}{P(e/-h)P(-h)} \tag{4}$$

Now we are in a position to categorize the most common forms of representative thinking. Pseudodiagnosticity refers to making an inference about the validity of hypothesis h on the base of evidence e without considering hypothesis $-h$. Another way of stating pseudodiagnosticity is that it involves considering only the numerator of the first term on the right-hand side in Eqn. (4). People do that when they state that a bit of evidence is 'consistent with' or 'typical of' some hypothesis without concerning themselves about alternative hypotheses, or in particular the negation of the hypothesis. For example, in attempting to diagnose whether a child has been sexually abused, alleged experts in court often refer to symptoms 'typical' of such abuse—without concerning themselves with how typical these symptoms are of children who have *not* been abused, or the frequency with which children have been sexually abused.

Another type of representative thinking involves considering the probability of the evidence given the hypothesis or hypotheses without looking at the second terms in Eqn. (4), that is, the prior odds. The reason that these odds are important is that they yield the *extent* of the evidence and hypotheses considered. For example, being a poor speller may be evidence of dyslexia in the sense that it has a very high probability given dyslexia. However, in order to diagnose dyslexia on the basis of poor spelling, we have to know something about the base rates of poor spelling and the base rates of dyslexia in the population from which the person whom we wish to diagnose was chosen.

Dilution effects occur when evidence that does not distinguish between hypotheses in fact influences people to be less convinced of a most likely hypothesis; the problem is similar to that found in pseudodiagnosticity in that people do not realize that evidence that is not very likely given the hypothesis may be equally unlikely given the negation of that hypothesis, but may in fact believe a hypothesis less as a result of collecting evidence that is unlikely if it is true. Dilution is simply the converse of pseudodiagnosticity.

Finally, it is possible to describe *availability* biases as well by Eqn. (4). People believe that they are sampling evidence given the hypothesis, when in fact they are sampling this evidence given the hypothesis *combined with* the manner in which they are sampling. For example, when private practice clinical psycho-

logists claim that they are sampling characteristics of people who fall into a certain diagnostic category on the basis of their experience, what they are really sampling is people who fall in that category *and who come to them*. Similarly, when we sample our beliefs about 'what drug addicts are like,' most of us are sampling on the basis of how the media presents drug addicts—both in 'news' programs and in dramatizations (Dawes 1994b). Doctors often sample on the basis of their contact with addicts when these addicts are ill, perhaps gravely so, while police are often sampling on the basis of their experience with these same addicts during arrests and other types of confrontation. It is not surprising, then, that doctors are in favor of a 'medical' approach to drug addiction including such policies as sterile needle exchanges, while police are in favor of much more punitive policies. Both are sampling evidence *combined with* their exposure to it.

All these anomalies have been demonstrated experimentally. It is important to point out, however, that the investigation of these anomalies did not spring simply from understanding Bayes' theorem and then from creating very clever experimental situations in which subjects' inferences systematically violate it. The motive to study these anomalies of judgment arises from having observed them on a more informal basis outside the experimental setting—and trying to construct an experimental setting, which yields a greater measure of control, that will allow them to be investigated in a systematic manner. The reader is referred, for example, to the essay of Meehl (1977) or the book of Dawes (1994a) for descriptions of these biases in the clinical judgment of professional psychologists and psychiatrists who rely on their own 'experience' rather than (dry, impersonal) 'scientific' principles for diagnosis and treatment.

5. *Analyzing an Important Example*

This section ends with a discussion of pseudodiagnosticity, a bias that is prevalent in 'expert' proclamations encouraging 'child sex abuse hysteria' (Dawes 2001). The main problem here is that hypotheses are not compared: instead, single hypotheses are evaluated in terms of the degree to which evidence is 'consistent with' them; in addition, evidence is often *sought* in terms of its consistency with or inconsistency with 'favorite hypotheses'—rather than in terms of its ability to distinguish *between* hypotheses. This type of pseudodiagnosticity has made its way into the legal system, where experts are allowed to testify that a child's recanting of a report of sexual abuse is 'consistent with' the 'child sexual abuse accommodation syndrome' (Summit 1983). The finding is that many children who are *known* to have been sexually abused deny the abuse later (recant), therefore, the

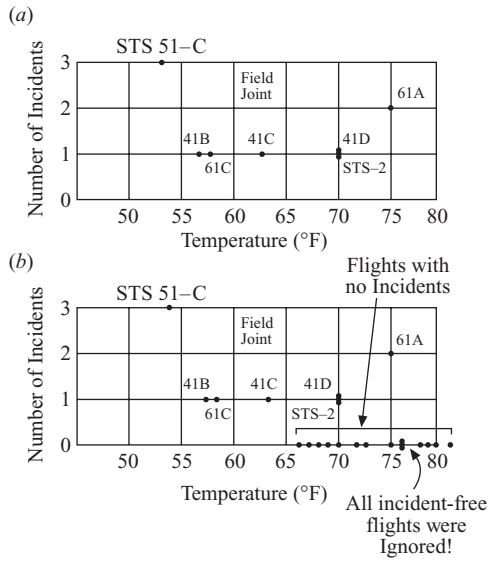

Figure 1
Partial and total analyses of the relationship between (a) data reviewed prior to launch and (b) the full story from past flights (source: Russo and Schoemaker 1989, pp. 197–8)

probability of the evidence (the child recants) given the hypothesis (actual abuse) is not as low as might be naively assumed; therefore, recanting is 'consistent with' having been abused (i.e., can be considered part of a 'syndrome' of accommodation to abuse).

The problem with this pseudodiagnostic reasoning is that it compares the probability of the evidence given the hypothesis to the probability of the *negation of the evidence* given the hypothesis, whereas a rational comparison is of the probability of the evidence given the hypothesis to the probability of the evidence given the *negation of the hypothesis*. When considering this latter comparison, we understand immediately that the recanting would be diagnostic of actual abuse only if the probability of recanting given actual abuse were *higher* than the probability of recanting given such abuse had not occurred—a highly implausible (not to mention paradoxical) conclusion.

Such pseudodiagnostic reasoning is involved in other disasters as well, for example, the explosion of the US Space Shuttle Challenger on January 22, 1986. When considering the possibility that cold temperatures might cause the O-rings to malfunction, 'the worried engineers were asked to graph the temperatures at launch time for the flights in which problems have occurred' (Russo and Schoemaker 1989, p. 197). Thus, these engineers attempted to evaluate the hypothesis that there would be O-ring problems by asking what the temperature was given the existence of these problems. As indicated in Fig. 1(a) (Fig. 6 in

Russo and Schoemaker 1989, p. 198), there appeared to be no relationship. When, however, the temperature of launches in which there were *no* problems were considered, a very clear relationship between problems and temperature is evident (see Fig. 1(b), from Russo and Schoemaker 1989, Fig. 7, p. 198.)

Moreover, people actively seek evidence compatible or incompatible with a hypothesis rather than evidence that distinguishes between hypotheses, as has been extensively studied by Doherty and his colleagues (see for example Doherty et al. 1979). Consider, for example, subjects who are asked to play the role of medical students (and actual medical students prior to specific training, see Wolf et al. 1985) who must determine whether patients have one of two conditions. The conditions are explained, and subjects are told that a patient has two symptoms. The judges are then presented with the conditional probability of one of these symptoms given one of the diseases (e.g., that it is very likely that the patient has a high fever given the patient has meningitis). They can then choose to find out the probability of the other symptom given the same disease, the probability of the first symptom given the second disease, or the probability of the second symptom given the second disease. While virtually no subjects choose a different symptom and a different disease, the majority choose to find the probability of the *second* symptom given the first disease (the 'one that is focal' as a result of being told the first symptom). But for all these subjects know, of course, the first symptom may be more or less typical of the *alternative disease*, as may the second symptom. Finding out that the probability of these two symptoms given only one disease in no way helps to distinguish between that disease and some other. *Of course, in real medical settings doctors may have prior knowledge of these other disease/symptom relationships*, but because the diseases are not even identified in the current setting, such knowledge would be of no help.

Occasionally, pseudodiagnosticity has been confused with what has been termed a 'confirmation bias'—that is, a desire to search for evidence that is very probable given a focal hypothesis. As Fischhoff and Beyth-Marom (1983), Trope and Bassok (1982), and Skov and Sherman (1986) point out, however, such a search does not necessarily imply pseudodiagnosticity (or any other particular irrationality) because it may be possible that the evidence is *not* found—that is, that instead of finding the anticipated evidence, the diagnostician finds its negation. In fact, Skov and Sherman (1986) demonstrate that in searching for evidence—at least in their experimental context—subjects often became quite rational *when they are informed of the conditional probabilities*. The irrationality occurs in the search for which probabilities to seek, and in 'jumping' to a conclusion on the basis of only one probability (i.e., on the basis of the first term in the numerator in the odds form of Bayes' theorem).

The cognitive explanation of pseudodiagnosticity and the other heuristic-based biases covered in the section is simple. People often think by association; in fact, whole computer programs have been built to simulate human judgment on the basis of 'matching' characteristics to schemas. The point of pseudodiagnosticity is that only one 'schema' (i.e., hypothesis) is considered, whereas *coherent diagnostic judgment is always comparative*. In fact, all the biases covered here are based on associational thinking, whereas clear normative probabilistic thinking always involves comparisons.

6. *Summary and Implications*

When we deal with probabilistic situations in everyday life, we can often 'muddle through,' but occasionally not appreciating the comparative nature of valid probabilistic thinking can lead to judgmental disasters. Similarly, experience with games of chance can lead to some pretty good intuitive estimations of the likelihood of various payoffs, but once again there can be systematic biases, such as belief in the gambler's fallacy. Explicit probabilistic thinking has developed in our cultures only in the last 500 years or so, and in the last 150 years or so it has become increasingly important in both scientific reasoning and in our day-to-day living—and even in legal contexts (where as jurors we may be asked to judge the liability of a particular company by comparing the probability of developing cancer among those exposed to it or its product as opposed to the 'base rate' probability, or may be asked to evaluate what an expert witness states about the relationship between a symptom and a diagnosis). Here, if we rely on our intuitive grasp of probabilities, we often find that we have systematic biases, much as novice swimmers may attempt to keep their heads above water and thereby drown. These biases—which have been briefly described and catalogued in this article—involve explicitly defined deviations from the conclusions of valid probabilistic reasoning, which follow from the simplest forms of Bayes' theorem. They are not random deviations, which could be modeled as analogous to an engineering context involving a 'signal' plus 'noise.'

Of course, we are also capable of precise probabilistic thinking—just as swimmers are capable of assuming a horizontal position in the water and consequently not drowning. The point is, however, that when we deviate from the rules ('axioms') of probability theory or from coherent judgments, these deviations follow the patterns described here—as has been documented in everyday observation and studied experimentally by those advocating the 'heuristics and biases' approach to such judgment. For a greater elaboration of this approach than is found here, see the chapters in Kahneman et al.'s book *Judgment Under Uncertainty: Heuristics and Biases* (1982). These

systematic departures themselves in turn may be connected to thinking in terms of associations, whereas good probabilistic judgment always requires comparative thinking.

See also: Information Theory; Information, Economics of; Probability and Chance: Philosophical Aspects; Probability: Interpretations; Risk, Sociological Study of

Bibliography

Bar-Hillel M 1990 Back to base-rates. In: Hogarth R M (ed.) *Insights in Decision Making: A Tribute to Hillel J. Einhorn*. University of Chicago Press, Chicago

Dawes R M 1988 *Rational Choice in an Uncertain World*. Harcourt Brace Jovanovich, San Diego, CA

Dawes R M 1994a *House of Cards: Psychology and Psychotherapy Built on Myth*. Free Press, New York

Dawes R M 1994b AIDS, sterile needles, and ethnocentrism. In: Heath L, Tinsdale R S, Edwards J, Posavac E, Bryant F B, Henderson-King E, Suarez-Balcazar Y, Myers J (eds.) *Social Psychological Applications to Social Issues. III: Applications of Heuristics and Biases to Social Issues*. Plenum, New York, pp. 31–44

Dawes R M 1998 Behavioral decision making and judgment. In: Gilbert D, Fiske S, Lindzey G (eds.) *The Handbook of Social Psychology, 11*. McGraw-Hill, Boston, MA, pp. 497–548

Dawes R M 2001 *Everyday Irrationality: How Pseudoscientists, Lunatics and the Rest of Us Systematically Fail to Think Rationally*. Westview Press, Boulder, CO

David F N 1962 *Games, Gods, and Gambling*. Hafner Pub. Co., New York, Chap. 4

Dawid A P 1982 The well-calibrated Bayesian. *Journal of the American Statistical Association* **77**(379): 605–10

de Finetti B 1974–5 *Theory of Probability: A Critical Introductory Treatment*. John Wiley and Sons, London, Vol. 1

Doherty M E, Mynatt C R, Tweney R D, Schiavo M D 1979 Pseudodiagnosticity. *Acta Psychologica* **43**: 111–21

Fischhoff B, Beyth-Marom R 1983 Hypothesis evaluation from a Bayesian perspective. *Psychological Review* **90**: 239–60

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L 1989 *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, Cambridge, UK

Gilovich T, Vallone R, Tversky A 1985 The hot hand in basketball: On the misperception of random sequences. *Journal of Personality and Social Psychology* **17**: 295–314

Gold E 1997 The Gambler's Fallacy, Doctoral dissertation submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy, Carnegie Mellon University

Hacking I 1975 *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge University Press, New York

Kahneman D, Tversky A 1972 Subjective probability: A judgment of representativeness. *Cognitive Psychology* **3**: 430–54

Kahneman D, Tversky A 1973 On the psychology of prediction. *Psychological Review* **80**: 237–51

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK

Meehl P E 1977 Why I do not attend case conferences. In: Meehl P E (ed.) *Psychodiagnosis: Selected Papers*. W. W. Norton, New York, pp. 225–302

Meehl P E, Rosen A 1955 Antecedent probability in the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin* **52**: 194–216

Mulford M, Dawes R M 1999 Subadditivity in memory for personal events. *Psychological Science* **10**(1): 47–51

Ramsey F 1931 Truth and probability. In: Ramsey F P (ed.) *The Foundations of Mathematics and other Logical Essays*. Harcourt, Brace and Co., New York, pp. 156–98

Russo J E, Schoemaker P J H 1989 *Decision Traps: Ten Barriers to Brilliant Decision Making and how to Overcome Them*. Simon and Schuster, New York

Simon H 1955 A behavior model of rational choice. *Quarterly Journal of Economics* **69**: 99–118

Simon H 1956 Rational choice and the structure of the environment. *Psychological Review* **63**: 129–38

Simon H 1979 *Models of Thought*. Yale University Press, New Haven, CT, Vol. 1

Skov R B, Sherman S J 1986 Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology* **22**: 93–121

Summit R C 1983 The child sexual abuse accommodation syndrome. *Child Abuse and Neglect* **7**: 177–93

Trope Y, Bassok M 1982 Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology* **43**(1): 22–34

Tversky A, Kahneman D 1971 Belief in the law of small numbers. *Psychological Bulletin* **76**: 105–110

Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* **185**: 1124–31

von Neumann J, Morgenstern O 1944/1947 *The Theory of Games and Economic Behavior*, 2nd edn. Princeton University Press, Princeton, NJ

Wagenaar W 1988 *Paradoxes of Gambling Behavior*. Lawrence Erlbaum Associates, Hillsdale, NJ

Wolf F M, Gruppen L D, Billi J E 1985 Differential diagnosis and the competing hypothesis heuristic—a practical approach to judgment under uncertainty and Bayesian probability. *Journal of the American Medical Association* **235**: 2858–62

R. M. Dawes

Probability and Chance: Philosophical Aspects

The central philosophical issue involving probability concerns the interpretation of ordinary probability claims, such as 'there is a 30% probability that it will rain today.' *Epistemic* interpretations of probability understand this to be a claim about belief under uncertainty: it describes the extent to which a certain type of rational agent would (or ought to) believe that

it will rain, given the available evidence. *Objective* interpretations of probability read the claim as describing some quantifiable feature of the objective world, sometimes called 'chance.' The concept of chance is thought by some to be intimately connected with *indeterminism*. A variety of epistemic and objective interpretations of probability will be discussed below.

1. The Problem of the Interpretation of Probability

The problem of the interpretation of probability can be posed using a familiar example: what does a meteorologist *mean* when he or she says that there is a 30 percent probability of rain on a given day? To give an account of what a proposition means is to say what must be the case in order for the proposition to be true. If the meteorologist says that it *will* rain, this assertion is true if and only if it does rain. Whether it rains or not, however, neither outcome by itself renders the *probabilistic* claim true. So what must be the case in order for this claim to be true?

An adequate solution to this problem must provide some way of making sense of the numerical values that appear in probability claims (and must do so in such a way that those values are in accord with the basic principles of probability—see *Probability: Formal*). The claim that 30 percent of the balls in a given urn are blue has a well-understood meaning: the number of blue balls in the urn, divided by the total number of balls in the urn, is equal to 0.3. In the case of the weather report, however, no comparable formula is immediately available.

Proposed interpretations of probability tend to fall within one of two broad camps, *objective* (or *empirical*) and *epistemic* (or *subjective*) interpretations. Objective interpretations interpret probability claims as describing some feature of the world. Epistemic interpretations of probability, by contrast, interpret probability claims as claims about the extent to which the claim is or ought to be believed. Note that this standard terminology is somewhat misleading: some epistemic interpretations (such as the logical interpretation) take probabilities to be objective in the sense that they do not vary from person to person. The various interpretations of probability canvassed below are often presented as *rivals*, but it may be more appropriate to think of them as complementary. Most researchers who have grappled with interpretive issues have been of the opinion that the concept of probability has both an objective and an epistemic dimension. Accordingly, philosophers often talk of 'objective probabilities' and 'epistemic probabilities.' Objective probabilities are sometimes called 'chances.'

Philosophers (as well as many philosophically minded statisticians) have become interested in the problem of interpretation of probability in part because of the ubiquity of probability claims in science as well as ordinary life, and in part because of the utility of probability theory in analyzing a variety of concepts of interest to philosophers, such as causation (see *Causation: Physical, Mental, and Social*), confirmation of hypotheses by evidence (see *Atomism and Holism: Philosophical Aspects*), and rationality (see *Rational Choice Explanation: Philosophical Aspects*).

This article will focus on the philosophical, rather than the mathematical, dimensions of the problem of interpreting probability. Readers interested in this topic would also do well to read the article on *Probability: Interpretations*. Other surveys of this topic include Salmon (1967), Sklar (1993), and Howson (1995).

2. Determinism and Chance

In a famous thought experiment, the French physicist Pierre Simon de Laplace (1819) imagined a demon who could observe the exact position and momentum of every particle in the universe at a given time. Such a demon could plug this information into Newton's equations of motion and predict the state of the universe at all future times. In this thought experiment, Laplace was articulating the concept of *determinism*. A system is *deterministic* if its state at a given time, together with the laws of evolution for that system (see *Causes and Laws: Philosophical Aspects*), uniquely determine the state of the system at later times. For Laplace, determinism meant that all *objective* probabilities would have to have values of zero or one. Laplace thus thought that probabilistic claims, such as the meteorologist's prognostication described above, reflect human ignorance about which outcome is determined to occur.

At present, there is good reason to believe that certain physical systems described by *quantum mechanics* behave indeterministically. For example, if a vertically polarized photon strikes a polarizer that is oriented with an angle of θ, quantum theory entails that the photon has a probability of $\cos^2\theta$ of passing through. According to the orthodox interpretation of quantum theory (which is not uncontroversial), this probability is *irreducible*: there is no further fact about the photon that determines what it will do on this occasion. If this is correct, then the probability value $\cos^2\theta$ does not merely reflect human ignorance about what the photon will do: it is an objective feature of the world. (For a thorough discussion of philosophical issues involving determinism, see Earman (1986).) One important philosophical issue is whether only such irreducible probabilities are genuine chances. If so, then it would be safe to infer that *none* of the probability claims made within the social sciences describe genuinely objective probabilities.

The word 'chance' is also sometimes used as a singular noun to refer to a chance that is not equal to zero or one. For example, the thesis of determinism is sometimes expressed by saying that there is no genuine chance in the world; 'the USA soccer team has a chance of winning the next world cup' says that the chance of their winning is greater than zero and less than one; to 'take a chance' is to embark on a venture that has a chance of success that is greater than zero but less than one; and so on.

3. *Epistemic Probability*

Laplace claimed that probability claims do not describe objective features of the world, but the state of human knowledge (and ignorance). Epistemic interpretations of probability all share this general feature.

3.1 *The Classical Interpretation*

If asked the probability of rolling an odd number on a six-sided die, most people would unhesitatingly respond with the answer one-half: there are six possible outcomes of the die roll, three of which are odd. This illustrates the *classical interpretation* of probability, often associated with Laplace (1819). According to this interpretation, the possible outcomes of a particular trial are divided into 'equally possible' cases, and the probability of outcome F on a particular trial is equal to the number of such cases in which F is true, divided by the total number of possible cases. Which cases are equally possible is determined by the *Principle of Indifference*, which states that two cases are equally possible in the absence of any reason to prefer one outcome over the other.

There are a number of difficulties with the principle of indifference. When the weather reporter claims that the probability of rain is 0.3, there seem to be at least two different types of outcome (rain and no rain) but it is hard to see what the 10 equally possible outcomes (three of which involve rain) could be. Even in much simpler cases, the principle of indifference is not precise enough to specify the equally possible cases. An urn contains both red balls and blue balls; some of the red balls are light red, and some are dark red. What is the probability that the next ball drawn is blue: are there two equally possible cases (red and blue) or three (light red, dark red, and blue)? This problem is further exacerbated when there are infinitely many possible outcomes.

3.2 *Subjective Interpretations*

According to the subjective interpretation of probability (whose proponents are variously called subjectivists' personalists, or Bayesians), probability measures the degrees of belief or *credences* of a certain type of idealized rational agent. Subjectivists regard probability theory as analogous to *formal logic*. Logic alone never tells an agent whether to believe a proposition, but it does require that an agent's beliefs be consistent with one another. For example, the rules of logic forbid an agent to believe 'it will rain today,' 'if it rains today, it will not snow today,' and 'it will snow today' at the same time. According to subjectivists, probability theory is an extension of logic that allows an agent to have degrees of confidence in a proposition that range in value between zero and one. Probability theory alone will not tell the agent how much confidence to place in the proposition 'it will rain today,' but rather imposes a consistency-like constraint (usually called *coherence*) on the totality of the agent's degrees of belief.

The numerical value of an agent's degree of belief in some proposition is connected with the agent's evaluation of bets on the truth of that proposition. For example, if an agent has a degree of belief of 0.3 for the proposition 'it will rain today,' that means that he or she would find a bet on rain at 3:7 odds—a bet that costs the agent three units, and buys a chance to win 10 units (for a profit of seven units) if it rains—to be fair. This connection between degrees of belief, utility, and evaluation forms the foundation of a theory of rational choice wherein an agent always acts so as to maximize expected utility (see *Rational Choice Explanation: Philosophical Aspects*; *Elicitation of Probabilities and Probability Distributions*; *Game Theory and its Relation to Bayesian Theory*).

This connection between degrees of belief and evaluation is often deployed in arguments that rational degrees of belief conform to the formal rules of probability (see *Probability: Formal*). When an agent's degrees of belief conform to the probability calculus, then his or her overall system of evaluations will have a number of desirable features. For example, Ramsey (1926) showed that having one's degrees of belief conform to the rules of probability is a necessary condition for having one's overall system of *preferences* satisfy a set of plausible constraints. One such constraint is transitivity: if the agent prefers A to B and B to C, then the agent prefers A to C. A second class of results comprises the so-called *Dutch Book* theorems. (See, for example, de Finetti (1964) and Kemeny (1955).) These results show that if and only if an agent's degrees of belief violate the rules of probability, the agent will be vulnerable to a 'Dutch Book.' A Dutch Book is a system of bets such that the agent finds each individual bet fair, but where the agent is guaranteed of suffering a net loss no matter what the outcome. For example, a bet on heads at 2:1 odds together with a bet of equal stakes on tails at 2:1 odds would constitute a Dutch Book, for the agent purchasing the two bets would be guaranteed of losing 2 units on one bet, and winning 1 unit on the other. (For more details, see *Probability: Interpretations*.)

The requirement that rational degrees of belief satisfy the rules of probability theory is a *synchronic* constraint on one's degrees of belief: it concerns an agent's degrees of belief at one particular time. Some subjectivists impose a further *diachronic* constraint on allowable methods for changing one's degrees of belief over time. Suppose that a rational agent's degrees of belief at some particular time are represented by the probability function P. If the agent then acquires some new piece of information I, then the agent's new degree of belief in any proposition A is equal to the *conditional probability* $P(A/I)$ (see *Probability: Formal* for the definition of conditional probability). This rule of updating degrees of belief by *conditionalizing* can also be motivated by a form of Dutch Book argument (proved by David Lewis and reported in Teller (1973)).

This rule for changing degrees of belief has been championed by so-called *Bayesians* as a theory of confirmation. According to this theory, a hypothesis H is confirmed or supported by a piece of evidence E just in case conditionalizing on E increases the probability of H. This probabilistic approach to confirmation theory has proven to be very fruitful. (See *Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects*; see also Howson and Urbach (1993).)

The various theorems invoked by subjectivists are *formal* results, and there is considerable debate over their interpretation. In particular, there is disagreement about how, or even whether, the Dutch Book theorems really show that coherence or updating by conditionalization are canons of rationality. Critics of subjectivism point out that even an incoherent agent would be capable of following the proofs of these theorems, and hence of foreseeing the dangers of Dutch Books; thus coherence and updating by conditionalization are not really necessary for avoiding financial loss (see Howson and Urbach (1993) for further discussion).

A second focus of criticism is the *subjective* nature of degrees of belief. Rational agents may differ in the degree to which they believe a proposition such as 'it will rain today'; indeed their degrees of belief may range all the way from zero to one. Moreover, the constraints of synchronic coherence plus updating by conditionalization do not rule out the possibility that one's degrees of belief are *dogmatic* in a certain respect. It is consistent with these constraints that an agent have degree of belief of 0.99 that each toss of a given coin will land heads, and after observing a thousand tails in a row, *still* believe that the *next* toss will almost certainly be heads. Subjectivists have been able to prove many interesting results about how an agent's degrees of belief *would* evolve in the face of new evidence *if* the agent's original degrees of belief (or *prior probabilities*) were nondogmatic in various ways (de Finetti's (1964) representation theorem is one important example), but there is nothing in subjectivism that would *require* the agent to have such nondogmatic prior probabilities. For many, this latitude makes subjective degrees of belief ill-suited as a foundation for theories of rational choice or rational inference.

3.3 Logical Interpretations

Logical interpretations aim to avoid the problems of subjectivist interpretations by imposing further constraints on an agent's prior probabilities. Prior probabilities are determined by the logical structure of the space of outcomes, much as they are in the classical interpretation. Unlike the classical interpretation, however, probability assignments can reflect the bearing of evidence that discriminates among the possible outcomes. Thomas Bayes (1763) and Laplace (1819), who are commonly associated with the subjectivist and classical approaches (respectively), both endorsed a version of the logical approach to probability. Twentieth century proponents include Carnap (1950), Jeffreys (1961), and Jaynes (1973).

Logical interpretations of probability inherit many of the difficulties of the classical interpretation discussed above: logical considerations alone do not determine a unique prior probability distribution. Moreover, many of the most plausible candidates for logical probabilities—such as one in which a sequence of coin tosses is *independent and identically distributed* (see *Probability: Formal* or *Distributions, Statistical: Special and Discrete*)—turn out to be dogmatic in just the manner described in Sect. 3.2.

3.4 The Need for Objective Interpretations

Although few would doubt that our ordinary concept of probability has an epistemic dimension, it seems implausible that *all* probability claims can be interpreted epistemically. Consider the claim that a photon has a certain probability of passing through a polarizer; it would be very odd indeed if the truth of this claim (and hence of quantum mechanics) depended on the extent to which some agent(s) may believe that the photon will pass through or on the extent to which available evidence confirms this hypothesis. At best, epistemic interpretations of probability will have to be supplemented by an objective interpretation of at least some probability statements.

4. Objective Interpretations of Probability

In practice, probabilities are often inferred from frequency data. To estimate the probability that a given coin will land heads when tossed, for example, one could toss it many times and observe the relative frequency with which the coin lands heads. Such

relative frequencies are *objective* features of the world. Hence relative frequencies provide a natural resource for objective interpretations of probability claims.

4.1 Finite Relative Frequency Interpretation

Perhaps the simplest interpretation of a probability claim would be to understand it as a claim about an actual finite relative frequency. For example, if a coin were tossed exactly 10 times and then destroyed, resulting in 'heads' on six of those 10 tosses, then the probability that the coin lands heads when tossed would be 0.6. The central difficulty with this definition of probability is that it rules out unrepresentative samples by definition. For example, if a coin is fair, i.e., has a probability of 0.5 of landing heads when tossed, that should not be taken to rule out the possibility of it landing hands on six of the 10 occasions on which it is tossed. According to the finite frequency definition of probability, however, if the probability of heads is 0.5, then *by definition* the coin will land heads on exactly half of its tosses. Thus, while finite relative frequencies provide indispensable evidence for probabilistic claims, probabilities cannot be identified with finite relative frequencies.

4.2 Limiting Relative Frequency Interpretation

In order to avoid these problems with finite frequencies, some philosophers have suggested that probabilities be identified with frequencies in an *infinite* sample. If the first n tosses of a coin yields $f(n)$ heads, then the relative frequency of heads in the first n tosses is $f(n)/n$. The *limiting relative frequency* of heads in the infinite sequence of tosses is the limit as n goes to infinity of $f(n)/n$. This approach to interpreting probability was pioneered by Venn (1866); leading twentieth century proponents of this interpretation (hereafter called *frequentists*) include Richard von Mises (1957) and Hans Reichenbach (1949).

One important problem with limiting relative frequencies is that they seldom, if ever, exist: no coin is actually tossed an infinite number of times. Thus the limiting relative frequencies that determine the probability of an outcome are really *hypothetical limiting relative frequencies*: to say that the probability of heads on a toss of a certain coin is 0.5 is to say that *if* the coin were to be tossed an infinite number of times, the limiting relative frequency of heads *would be* 0.5. This is a hypothetical or *counterfactual* claim; by itself, it says nothing about what features of the *actual world* make the probabilistic claim true. The problem of interpretation of counterfactuals is itself a thorny philosophical problem (see *Counterfactual Reasoning, Quantitative: Philosophical Aspects*; *Counterfactual Reasoning, Qualitative: Philosophical Aspects*).

One important issue that divides frequentists is whether it is possible to ascribe a probability value to the outcome of a single trial. One tradition, following von Mises (1957), views probability as a property of a sequence of trials, or a 'collective.' Any single coin toss belongs to infinitely many different sequences: the sequence of all coin tosses, the sequence of all tosses of that particular coin, the sequence of all tosses of that coin by imparting a particular force from a particular height above the floor, and so on. The limiting relative frequency of 'heads' may be different in each of these sequences. Thus, there is no unique probability of 'heads' on a given toss, but only a probability for 'heads' in a collective of tosses. While frequentists in this tradition deny that there is one privileged sequence of coin tosses, not just any sequence of tosses is fit to bear probability values. For example, if an infinite sequence of coin tosses were to yield an alternating sequence of 'heads' and 'tails,' the relative frequency of 'heads' would be 0.5. The regularity of the pattern of outcomes would nonetheless render this sequence inappropriate for characterizing a probability value. Rather, probabilities are limiting relative frequencies in a *random* sequence of trials. There is no universally accepted mathematical characterization of randomness, although this is currently an area of active research (see Earman 1986, Chap. VIII). For a more detailed exposition of this version of the limiting relative frequency interpretation, see *Probability: Interpretations*.

One feature of this approach is that it allows chances to take on values other than zero or one even if determinism is true. It may be that on any particular toss of a coin, the linear and angular momentum imparted to the coin together with its height above the floor and so on determine the outcome of the coin toss. Nonetheless, there may be a random sequence of coin tosses where the limiting relative frequency is not equal to zero or one, for the tosses comprising the collective need not be identical with respect to such determining factors.

The relegation of probabilities to properties of collectives would seem to be a drawback of this approach to probability. When deciding whether to bet on a coin toss, or to postpone a picnic, one wants to know the probability of heads *on that particular toss*, or the probability of rain *on this particular day*. Some frequentists, following Reichenbach, believe that it is possible to assign probability values to the outcomes of individual trials. This requires that a choice be made from among all of the sequences to which the trial in question belongs. For example, a particular day belongs in an infinite number of sequences of days, and the limiting relative frequency of rain need not be the same in all of them, so which sequence determines the probability of rain on the day in question? Presumably not just the sequence of all days, for then the probability of rain would not vary from day to day. A natural thought is that what is

wanted is a sequence of days *just like this one*. No two days will be exactly alike in *all* respects, however; the best that can be hoped for is a sequence of days that are alike in all *relevant* respects. The key problem, then, is to say what makes some respects of similarity between days relevant and others irrelevant. One detailed proposal is developed in Salmon (1984, Chap. 3).

On this approach, determinism is compatible only with chances of zero or one. Suppose, for example, that on a given day, climactic conditions determine that it will rain. Then the appropriate sequence will presumably consist of days in which the same climactic conditions prevail, and hence the limiting relative frequency of rain within that sequence will be one.

4.3 Propensity Interpretation

The propensity interpretation of probability, first introduced by Karl Popper (1959), is motivated by some of the difficulties with the limiting relative frequency interpretation. According to this interpretation, a coin, in conjunction with the conditions in which it is tossed, has a certain *disposition* or *propensity* to land heads. The probability that the coin will land heads is a measure of the strength of this propensity. This propensity is a property of the particular experimental set-up, hence, it is appropriate to assign a probability to an individual trial. Moreover, this propensity provides grounds for the truth of claims about what the coin *would do* if tossed infinitely many times in identical (or relevantly similar) conditions.

While this is certainly a useful way of thinking about single case probabilities within a frequentist framework, critics wonder whether anything has really been added. A coin is about to be tossed: there is some property of this experimental set-up in virtue of which the coin *would* land heads with a certain limiting relative frequency if this experiment were repeated infinitely often; calling this property a 'propensity' of a certain strength does little to indicate just what this property is. Moreover, if propensities are Pnot simply identified with hypothetical limiting relative frequencies, it is hard to see what is supposed to ground the claim that propensities have the mathematical structure of probabilities.

4.4 Best System Interpretation

The philosopher David Lewis (1986) has proposed an interpretation according to which an objective probability claim is true when it is entailed by a probabilistic *law* (see *Causes and Laws: Philosophical Aspects*). On this view, objective probability is intimately connected with indeterminism. A probabilistic claim is a law if it belongs to the set of propositions that provides the best overall systematization of the world. In order to be the best systematization, a set of propositions must obtain the optimal balance between informativeness, simplicity, and fit with data. For example, consider the claim that a photon will pass through a polarizer with a probability of $\cos^2\theta$. Although the frequency with which photons do pass through when the experiment is performed may diverge from this number, the probability value is firmly grounded in the principles of quantum mechanics, which overall provides an excellent fit with experimental data, and is a powerful and elegant theory.

Unlike the frequency interpretations, however, the truth conditions for probabilistic claims provided by the best system interpretation are not very precise. It is not clear what makes one set of propositions simpler than another, nor how simplicity is to be balanced against informativeness and goodness of fit.

4.5 Theoretical Role Interpretations

One further line of interpretation is influenced by the philosophical school of pragmatism (see *Pragmatism: Philosophical Aspects*). In this approach, objective probability claims are to be understood in terms of their role in our overall conceptual system. In particular, to understand a probability claim is to know what kinds of evidence would support or undermine that claim, and how to act accordingly if one accepts that claim as true. This approach has been developed by Mellor (1971), Skyrms (1980), and Lewis (1986).

Consider an agent whose degrees of belief conform to the rules of probability (see Sect. 3.2), and let *Ch* be a function that assigns numerical values to propositions. The agent takes *Ch* to be objective chance precisely if his or her degrees of belief are guided by beliefs about the values of *Ch*. For example, suppose that the agent believes *Ch* ('it will rain today') to be 0.3. If the agent understands *Ch* to be objective chance, then his or her degree of belief in the proposition 'it will rain today' will also be 0.3. This degree of belief, together with the agent's utility function, can then be used to decide whether or not to postpone a picnic. In this way, then, the agent's beliefs about objective chances influence his or her decisions. Moreover, since objective probability claims are now propositions about which the agent has degrees of belief, the apparatus of Bayesian confirmation theory shows how those degrees of belief will change in the face of new evidence.

This approach has a number of virtues, not least of which is that it explains why objective chances have the formal properties of probability theory: in order for beliefs about chances to guide degrees of belief in the appropriate way, objective chances must have the same formal structure as rational degrees of belief. There is, however, an important drawback to this approach. In analyzing the meaning of objective probability claims in terms of their theoretical role instead of their truth conditions, this account does not

answer what many have taken to be the central question of objective probability: what must be the case in the world in order for the (objective) probability of rain to be 0.3?

5. Conclusion

Interest in the philosophical aspects of probability dates back to the origins of the formal theory of probability in the seventeenth century. Epistemic and frequency interpretations of probability have been around for well over a century. In the second half of the twentieth century, interest in objective interpretations of single-case probabilities has dramatically increased the range of interpretive options available. Given the current lack of consensus, and the increasingly widespread use of probabilistic concepts in both philosophy and everyday life, it is safe to predict that interest in the problem of the interpretation of probability will continue into the twenty-first century.

See also: Bayesian Statistics; Bayesian Theory: History of Applications; Decision Theory: Bayesian; Probabilistic Thinking; Probability: Formal; Probability: Interpretations; Problem Solving and Reasoning, Psychology of; Subjective Probability Judgments; Utility and Subjective Probability: Contemporary Theories

Bibliography

Bayes T 1763 An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* **53**: 370–418

Carnap R 1950 *Logical Foundations of Probability*. University of Chicago Press, Chicago

de Finetti B 1964 Foresight: its logical laws, its subjective sources. In: Kyburg H E, Smokler H E (eds.) *Studies in Subjective Probability*. Wiley, New York

Earman J 1986 *A Primer on Determinism*. Reidel, Dordrecht, Chaps. II, VIII

Howson C 1995 Theories of probability. *British Journal for the Philosophy of Science* **46**: 1–32

Howson C, Urbach P 1993 *Scientific Reasoning: The Bayesian Approach*, 2nd edn. Open Court, Chicago

Jaynes E T 1973 The well-posed problem. *Foundations of Physics* **3**: 477–93

Jeffreys H 1961 *Theory of Probability*, 3rd edn. Clarendon Press, Oxford, UK

Kemeny J G 1955 Fair bets and inductive probabilities. *Journal of Symbolic Logic* **20**: 263–73

Laplace P S de 1819 *Essai Philosophique sur les Probabilités*. Mme. Ve Courcier, Paris

Lewis D 1986 *Philosophical papers*. Oxford University Press, Oxford, UK, vol. II, see especially the introduction and Chap. 19

Mellor D H 1971 *The Matter of Chance*. Cambridge University Press, Cambridge, UK

Popper K 1959 The Propensity interpretation of probability. *British Journal for the Philosophy of Science* **10**: 25–42

Ramsey F 1926 Truth and probability. In: Mellor D H (ed.) *Philosophical Papers*. Cambridge University Press, Cambridge, UK

Reichenbach H 1949 *The Theory of Probability*. University of California Press, Berkeley, CA

Salmon W C 1967 *The Foundations of Scientific Inference*. University of Pittsburgh Press, Pittsburgh, PA, see especially Chap. V

Salmon W C 1984 *Scientific Explanation and the Causal structure of the world*. Princeton University Press, Princeton, NJ, see especially Chap. 3

Sklar L 1993 *Physics and Chance*. Cambridge University Press, Cambridge, UK, see especially Chap. 3

Skyrms B 1980 *Causal Necessity*. Yale University Press, New Haven, CT

Teller P 1973 Conditionalization and observation. *Synthese* **26**: 218–58

Venn J 1866 *The Logic of Chance*. Macmillan, London

von Mises R 1957 *Probability, Statistics, and Truth*, 2nd rev. English edn. Allen and Unwin, London

C. Hitchcock

Probability Density Estimation

Associated with each random variable, X, measured during an experiment is the cumulative distribution function (cdf), $F(x)$. The value of the cdf at $X = x$ is the probability that $X \leqslant x$. Features in a set of random data may be computed analytically if the cumulative distribution function is known. The probability that a random sample falls in an interval (a, b) may be computed as $F(b) - F(a)$. This formula assumes the values of X are continuous. Discrete random variables taking on only a few values are easy to handle; thus, this article focuses on continuous random variables with at least an interval scale.

The same information may be obtained from the probability density function (pdf), $f(x)$, which equals $\Pr(X = x)$ if X takes on discrete values, or the derivative of $F(x)$ if X is continuous. If either $F(x)$ or $f(x)$ is known, then any quantity of interest such as the variance may be computed analytically or numerically. Otherwise, such quantities must be estimated using a random sample $x_1, x_2, \ldots, x_n$. The relative merits of estimating $F(x)$ vs. $F(x)$ are an important part of this article.

Density estimation takes two distinct forms—parametric and nonparametric—depending on prior knowledge of the parametric form of the density. If the parametric form $f(x|\vec{\theta})$ is known up to the p parameters $\vec{\theta} = (\theta_1, \ldots, \theta_p)$, then the parameters may be estimated efficiently by maximum likelihood or Bayesian algorithms. When little is known of the parametric form,

nonparametric techniques are appropriate. The most commonly used estimators are the histogram and kernel methods, as discussed below.

1. Histograms and CDFs

Grouping and counting the relative frequency of data to form a histogram may be taught at the beginning of every statistics course, but the concept seems to have emerged only in 1662 when John Graunt studied the Bills of Mortality and constructed a crude life table of the age of death during those times of plague (Westergaard 1968). Today the histogram and its close relative, the stem-and-leaf plot, provide the most important tools for exploring and representing structure in a set of data. Studying the shape of a well-constructed histogram can verify an assumption of normality, or reveal departures from normality such as skewness, kurtosis, or multiple modes. The recent study of the theoretical properties of histograms has led to concrete methods for computing well-constructed figures as discussed below.

With only a random sample of data, the cdf may be estimated by the empirical cdf (ecdf), denoted by $F_n(x)$. The value of the ecdf at x is the fraction of the sample with values less than or equal to x, that is, $\{\# x_i \leqslant x\}$. Now the probability any one of the x_i is less than x is given by the true cdf at x. Hence, on average, the value of $F_n(x)$ is exactly $F(x)$, and $F_n(x)$ is an unbiased estimator of $F(x)$. The inequality $\{X_i \leqslant x\}$ is a binomial random variable, so that the variance of

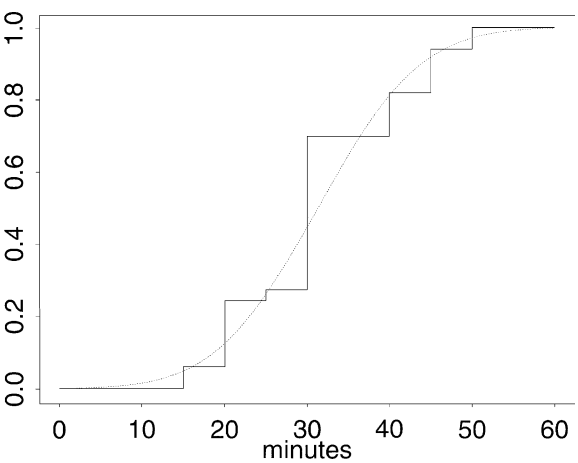

Figure 1
Empirical cumulative distribution function of physicians' estimated treatment times for a patient suffering migraine headaches. The dotted line is the cdf of the normal fit

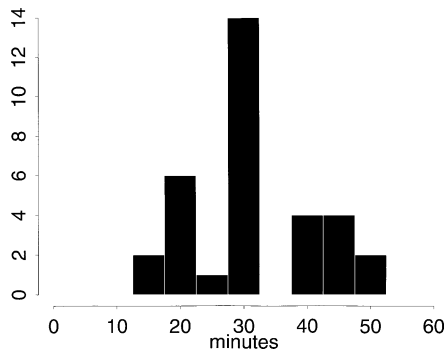

Figure 2
Frequency histogram of treatment times

$F_n(x)$ may be computed as $F(x)(1-F(x))/n$. In fact, there is no other unbiased estimator with smaller variance.

The graph of the empirical cdf looks like a staircase function. Figure 1 displays the ecdf of the treatment times estimated by 33 physicians when presented with a hypothetical patient suffering from extreme migraine headaches (Hebl and Xu 2000). The dotted line is the cdf of the fitted normal density $N(31.3, 9.9^2)$, which appears to fit reasonably well. Figure 2 displays a histogram of the same data, which does not appear to be normal after all. Social scientists often are taught to favor the ecdf over the histogram, but in fact aside from computing probabilities, the ecdf is not well-suited for the easy understanding of data.

The probability density function, $f(x)$, carries the same information as the cdf. Since the derivative of the empirical cdf $F_n(x)$ is not well-defined, estimators that rely on grouping have been devised. A convenient interval, (a, b), containing the data is chosen, and subdivided into m 'bins' of equal width $h = (b-a)/m$. The value h is called the bin width (and is often called the smoothing parameter). A 'well-constructed' histogram must specify appropriate values for a, b, h, and m. Given such parameters, the number of samples falling into each of the m bins is tallied. Label the m bin counts $v_1, v_2, \ldots, v_m$ from left to right. If the interval (a, b) contains all the samples, then $\Sigma_{k=1}^{m} v_k = n$. A *frequency histogram* is simply a plot of the bin counts, as in Fig. 2. Although the values are all multiples of 5 min, the data may be treated as continuous. However, the bin width should only be a multiple of 5 and care should be exercised to choose a and h so that no data points fall where two bins adjoin. Here, $a = 2.5$, $b = 57.5$, and $h = 5$. The histogram suggests a departure from normality in the form of three modes. Re-examine the ecdf in Fig. 1 and try to identify the same structure.

A true probability density is a non-negative function that integrates (sums) to 1. Hence, the *density his-*

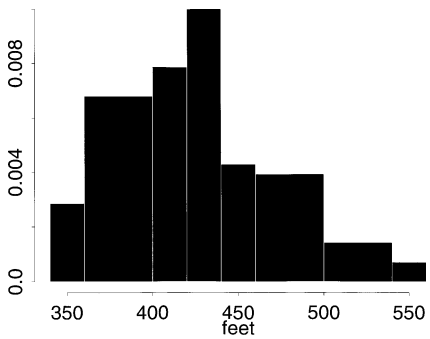

Figure 3
Density histogram of the distances of Mark McGwire's 70 home runs in the 1998 season

togram is a simple normalization of the frequency histogram in the kth bin,

$$\hat{f}(x) = \frac{v_k}{nh} \tag{1}$$

The normalization is important when the data are heavily skewed and unequal bin widths are constructed. In such cases, the interval (a, b) is partitioned into m bins of width $h_1, h_2, \ldots, h_m$. The bin counts $v_1, v_2, \ldots, v_m$ should *not* be plotted as they are not relatively proportional to the true density histogram, which is now given by $\hat{f}(x) = v_k/(nh_k)$ in the kth bin. Figure 3 displays a density histogram of the distances of the 70 home runs hit by Mark McGwire during his record-setting 1998 American baseball season. A bin width of 20 ft was chosen, and then several adjacent bins were collapsed to make the histogram look unimodal. The long right tail reflects the booming home runs for which McGwire is known.

2. *Bin Width Construction*

Rules for determining the number of bins were first considered by Sturges in 1926. Such rules are often designed for normal data, which occur often in practice. Sturges observed that the mth row of Pascal's triangle, which contains the combinatorial coefficients $\binom{m-1}{0}, \binom{m-1}{1}, \ldots, \binom{m-1}{m-1}$, when plotted, appears as an ideal frequency histogram of normal data. (This is a consequence of the central limit theorem for large m.) The sum of these 'bin counts' is 2^{m-1} by the binomial expansion of $(1+1)^{m-1}$. Solving the equation $n = 2^{m-1}$ for the number of bins, we obtain Sturges' rule

$$m = 1 + \log_2 n \tag{2}$$

This rule is widely applied in computer software.

The modern study of histograms began in Scott (1979) and Freedman and Diaconis (1981). The optimization problem formulated was to find the bin width h to make the average value of the 'distance' $\int[\hat{f}(x) - f(x)]^2 \, \mathrm{d}x$ as small as possible. The solution for the best bin width is

$$h^* = \left[\frac{6}{\int f'(x)^2 \mathrm{d}x} \right]^{1/3} n^{-1/3} \tag{3}$$

While the integral of the derivative of the unknown density is required to use this formula, two useful applications are available. First, the *normal reference rule* gives

$$h^* = 3.5\sigma n^{-1/3} \tag{4}$$

since the value of the integral for a normal density is $(4\sqrt{\pi}\sigma^3)^{-1}$. The second application was given by Terrell and Scott (1985) who searched for the 'easiest' smooth density (which excludes the uniform density) and found that the number of bins in a density should always exceed $\sqrt[3]{2n}$ or that the bin width should always be less than

$$h^* \leqslant 3.73\sigma n^{-1/3} \tag{5}$$

These choices are called *oversmoothed rules*, as the optimal choices will never be more extreme for other less smooth densities. Notice that the normal reference rule is very close to the upper bound given by the oversmoothed rule. For the McGwire data, the oversmoothed rule gives $h^* \leqslant 42'$.

More sophisticated algorithms for estimating a good bin width for a particular data set are described in Scott (1992) and Wand (1997). For example, a cross-validation algorithm (Rudemo 1982) that attempts to minimize the error distance directly leads to the criterion

$$\mathrm{CV}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)n^2 h} \sum_{k=1}^{m} v_k^2 \tag{6}$$

The bin counts are computed for a selection of bin widths (and bin origins), the criterion computed, and the minimizer chosen, subject to the over-smoothed bound. The criterion tends to be quite noisy and Wand (1997) describes alternative plug-in formulae that can be more stable.

All of these rules give bin widths which shrink at the rate $n^{-1/3}$ and give many more bins than Sturges' rule. In particular, most computer programs give histograms which are *oversmoothed*. The default values

should be overridden, especially for large datasets. Similarly, the error of the best histogram decreases to zero at the rate $n^{-2/3}$, which can be improved to $n^{-4/5}$ as described below. However, the histogram remains a powerful and intuitive choice for density estimation and presentation.

3. Advanced Algorithms

Density estimators with smaller errors are continuous functions themselves. The easiest such estimator is the frequency polygon, which connects the midpoints of a histogram. A separate error theory exists. For example, the oversmoothed bin width for a histogram from which the frequency polygon is drawn is given by

$$h^* \leqslant 2.33\sigma n^{-1/5}$$

Note that this bin width is wider than the rules for the histogram alone.

Many other estimators exist, based on splines, wavelets, Fourier series, or local likelihood, but all are closely related to the kernel estimator of Rosenblatt (1956) and Parzen (1962), which is described here. The kernel can be any probability density function, but the normal density is a common and recommended choice. Here the smoothing parameter h is the standard deviation of the normal kernel. If the normal density $N(\mu,\sigma^2)$ at a point x is denoted by $\phi(x|\mu, \sigma)$, then the kernel estimator is

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n}\phi(x|x_i, h)$$

This form shows that the kernel estimator is a mixture of normal densities which are located at the data points. The oversmoothed smoothing parameter value is

$$h^* \leqslant 1.14\sigma n^{-1/5}$$

An example is given below.

4. Visualization

The primary uses of density estimation are to display the structure in a set of data and to compare two sets of data. The histogram is adequate for the first task, but overlaying two histograms for comparison is difficult to interpret. The continuity and improved error properties of the kernel estimators excel for this purpose.

In the great home run race of 1998, Sammy Sosa finished second with 66 home runs. Figure 4 displays the empirical cdfs of the distances for the two home run kings (Keating and Scott 1999). This graph clearly shows that McGwire hits longer home runs than Sosa

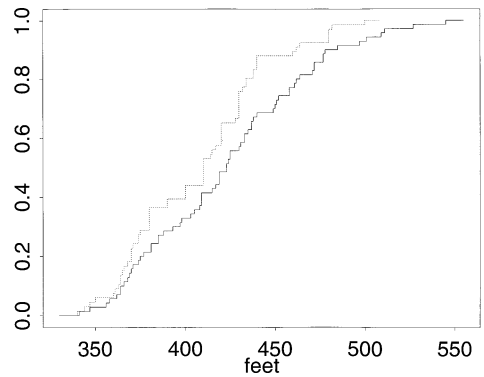

Figure 4
Empirical cumulative distribution functions of the distances of Mark McGwire's 70 (solid line) and Sammy Sosa's 66 (dotted line) home runs in 1998

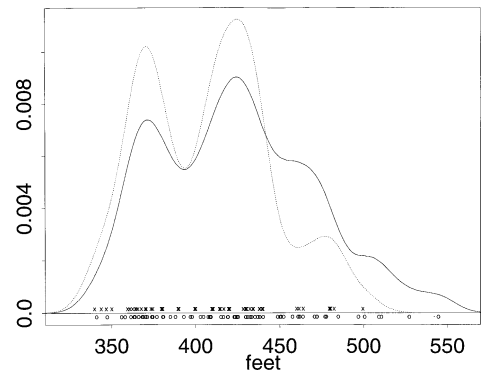

Figure 5
Kernel density estimates of the distances of Mark McGwire's 70 (solid line) and Sammy Sosa's 66 (dotted line) home runs in the 1998 season. McGwire's data are plotted below the horizontal line and Sosa's above

at every distance. In Fig. 5, kernel estimates of the McGwire and Sosa data with $h = 10$ are overlaid. The possible structure in these data is reinforced by the similarity of the two densities. Each has two primary modes at 370 ft and 425 ft, with smaller feature around 475 ft. In addition, McGwire's density clearly shows the right tail of his longest homers. In contrast, the modes in the original histogram of Fig. 3 did not appear convincing, and were eliminated by locally adjusting the bin widths. The continuity of the kernel and other estimators makes the assessment of modes much easier.

5. Multivariate Densities

Histograms can be constructed in the bivariate case and displayed in perspective, but cannot be used to draw contour plots. Kernel methods are well-suited

for this purpose. Given a set of data (x_1, y_1), …, (x_n, y_n), the bivariate normal kernel estimator is given by

$$\hat{f}_K(x, y) = \frac{1}{n}\sum_{i=1}^{n}\phi(x|x_i, h)\phi(y|y_i, h)$$

A different smoothing parameter can be used in each variable.

For three and four-dimensional data, the kernel estimator can be used to prepare advanced visualization renderings of the data. Scott (1992) provides an extension set of examples.

6. Other Materials

A number of other monographs exist that delve into various aspects of nonparametric approaches. A thorough survey of all topics by top researchers is contained in Schimek (2000). Early works popularizing these ideas include Tapia and Thompson (1978) and Silverman (1986). Wand and Jones (1994) and Simonoff (1996) focus on kernel-methods. Fan and Gijbels (1996) and Loader (1999) focus on local polynomial methods. General smoothing ideas are discussed by Härdle (1991) and Bowman and Azzalini (1997). Many other excellent references may be followed in these books.

See also: Probability: Formal

Bibliography

Bowman A W, Azzalini A 1997 *Applied Smoothing Techniques for Data Analysis*: The Kernel Approach with S-Plus. Clarendon Press, Oxford

Fan J, Gijbels I 1996 *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London

Freedman D, Diaconis P 1981 On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57**: 453–76

Härdle W 1991 *Smoothing Techniques: With Implementation in S*. Springer, New York

Hebl M, Xu J 2000 In: Lane D (ed.) Rice Virtual Lab in Statistics. http://www.ruf.rice.edu/~lane/rvls.html

Keating J P, Scott D W 1999 A primer on density estimation for the great home run race of '98 *Stats* **25**: 16–22

Loader C 1999 *Local Regression and Likelihood*. Springer, New York

Parzen E 1962 On estimation of probability density function and mode. *Annals of Mathematical Statistics* **33**: 1065–76

Rosenblatt M 1956 Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**: 832–837

Rudemo M 1982 Empirical choice of histogram and kernel density estimators. *Scandinavian Journal of Statistics* **9**: 65–78

Schimek M G (ed.) 2000 *Smoothing and Regression: Approaches, Computation, and Application*. John Wiley & Sons, New York

Scott D W 1979 On optimal and data-based histograms. *Biometrika* **66**: 605–10

Scott D W 1992 *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York

Silverman B W 1986 *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London

Simonoff J S 1996 *Smoothing Methods in Statistics*. Springer, New York

Tapia R A, Thompson J R 1978 *Nonparametric Probability Density Estiamtion*. Johns Hopkins University Press, Baltimore, MD

Terrell G R, Scott D W 1985 Oversmoothed non-parametric density estimates. *Journal of the American Statistical Association* **80**: 209–14

Wand M P 1997 Dated-based choice of histogram bin width histogram. *The American Statistician* **51**: 59–64

Wand M P, Jones M C 1994 *Kernel Smoothing*. Chapman and Hall, New York

Westergaard H 1968 *Contributions to the History of Statistics*. Agathon, New York

D. W. Scott

Probability: Formal

Probability theory is a branch of mathematics that has evolved from the investigation of social, behavioral, and physical phenomena that are influenced by randomness and uncertainty. For much of its early life, probability theory dealt almost exclusively with gambling games, and, even though there were important contributions made by such distinguished mathematicians as Pierre de Fermat, Blaise Pascal, and Pierre-Simon de Laplace, the field lacked respectability and failed to attract sustained attention.

One cause of the slow development of probability theory in its early days was the lack of a widely accepted foundation. Unlike the geometry of Euclid, or even the analytical investigations of Newton and Leibnitz, the theory of probability seemed to be eternally tied to the modeling process. In this respect, probability theory had greater kinship with the theories of heat or elasticity than with the pristine worlds of geometry or algebra.

Over time, foundations for probability were proposed by a number of deep thinking individuals including von Mieses, de Finetti, and Keynes, but the approach that has come to be the most widely accepted is the one that was advanced by Andrey Nikolayevich Kolmogorov (1933).

Kolmogorov's approach to the foundations of probability theory developed naturally from the theory of integration that was introduced by Henri Lebesgue and others during the first two of decades of the twentieth century. By leaning on the newly developed theory of integration, Kolmogorov demonstrated that probability theory could be simply viewed as another branch of mathematics. After Kolmogorov's work, probability theory had the same relationship to its applications that one finds for the

theory of differential equations. As a consequence, the stage was set for a long and productive mathematical development.

To be sure, there are some philosophical and behavioral issues that are not well addressed by Kolmogorov's bare-bones foundations, but over the years, Kolmogorov's approach has been found to be adequate for most purposes. The Kolmogorov axioms are remarkably succinct, yet they have all the power that is needed to capture the physical, social, and behavioral intuition that a practical theory of probability must address.

1. Kolmogorov's Axiomatic Foundations

Central to Kolmogorov's foundation for probability theory was his introduction of a triple $(\Omega, \mathscr{F}, P)$ that is now called a probability space. The triple's first element, the 'sample space' Ω, is only required to be a set, and on the intuitive level, one can think of Ω as the set of all possible outcomes of an experiment. For example, in an experiment where one rolls a traditional six-faced die, then one can take $\Omega = \{1, 2, 3, 4, 5, 6\}$.

The second element of Kolmogorov's probability space is a collection $\mathscr{F}$ of subsets of Ω that satisfy three basic consistency properties that will be described shortly. On the intuitive level, one can think of the elements of $\mathscr{F}$ as 'events' that may occur as a consequence of the experiment described by $(\Omega, \mathscr{F}, P)$. Thus, to continue with the example of rolling a die, the set $A = \{1, 3, 5\} \subset \Omega$ would correspond to the event that one rolls an odd number.

Two of the three consistency properties that Kolmogorov imposes on the $\mathscr{F}$ are quite trivial. First, $\mathscr{F}$ is required to contain Ω. Second, $\mathscr{F}$ must be closed under complementation; so, for example, if $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$ where $A^c = \{\omega : \omega \in \Omega \text{ and } \omega \notin A\}$. This third condition is only a bit more elaborate; the collection $\mathscr{F}$ must be closed under countable unions. Thus, if $A_k \in \mathscr{F}$ for $k = 1, 2, \ldots$, then one requires that the union $A_1 \cup A_2 \cup \ldots$ of all the events in the countable set $\{A_k : 1 \leqslant k < \infty\}$ must again be an element of $\mathscr{F}$.

The most interesting element of Kolmogorov's triple $(\Omega, \mathscr{F}, P)$ is the 'probability measure' P. Formally, P is just a function that assigns a real number to each of the elements of $\mathscr{F}$, and, naturally enough, one thinks of $P(A)$ as the probability of the event A. Thus, one can specify a probability model for the outcome of rolling a fair die, by 'defining' P for $A \subset \Omega$ by $P(A) = \frac{1}{6}|A|$, where $|A|$ denotes the number of elements of the set A.

For sample spaces Ω that are not finite, more care is required in the specification of the probability measures P. To deal with general Ω, Kolmogorov restricted his attention to those P that satisfy three basic axioms:

Axiom 1. For all $A \in \mathscr{F}$, one has $P(A) \geqslant 0$.

Axiom 2. $P(\Omega) = 1$.

Axiom 3. For any countable collection $\{A_k \in \mathscr{F} : 1 \leqslant k < \infty\}$ with the property that $A_j \cap A_k = \Phi$ for all $j \neq k$, one has

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Axioms 1 and 2 just reflect the intuitive view that $P(A)$ measures the frequency with which A occurs in an imagined sequence of repetitions of the experiment $(\Omega, \mathscr{F}, P)$. For most mathematicians, axiom 3 also just reflects the most basic intuition about the way probabilities should behave. Nevertheless, axiom 3 is more subtle because it deals with an infinite collection of sets, and in some ways, such collections are outside of our direct experience. This has lead some philosophers to examine the possibility of avoiding Kolmogorov's third axiom, and over the years various attempts have been made to replace Kolmogorov's third axiom with the simpler assumption of 'finite additivity.'

2. Random Variables

In most applications of probability theory, the triple $(\Omega, \mathscr{F}, P)$ that gives life to probability as a rigorous mathematical subject is almost invisible. In practice, builders of probability models take advantage of various shortcuts that have the effect of keeping the probability triple at a respectful distance. The most important of these shortcuts is the notion of a 'random variable.'

On the intuitive level, a random variable is just a number that depends on a chance-driven experiment. More formally, a random variable X is a function from Ω to the real numbers with the property that $\{\omega : X(\omega) \leqslant t\} \in \mathscr{F}$ for all t. What drives this definition is that one inevitably wants to talk about the probability that X is less than t, and for such talk to make sense under Kolmogorov's framework, the set $\{\omega : X(\omega) \leqslant t\}$ must be an element of $\mathscr{F}$. Random variables make the modeler's job easier by providing a language that relates directly to the entities that are of interest in a probability model.

3. The Distribution Function and Related Quantities

There are several ways to specify the basic probabilistic properties of a random variable, but the most fundamental description is given by the 'distribution function' of X, which is defined by $F(t) = P(X \leqslant t)$. Knowledge of the distribution function tells one everything that there is to know about the probability theory of the random variable. Sometimes it even tells too much.

The knowledge one has of a random variable is often limited, and in such cases it may be useful to

focus on just part of the information that would be provided by the distribution function. One common way to specify such partial information is to use the 'median' or the 'quantiles' of the random variable X. The median is defined to be a number m for which one has $P(X < m) \leqslant 1/2 \leqslant P(X \leqslant m)$. Roughly speaking, the median m splits the set of outcomes of X into two halves so that the top half and the bottom half each have probability that is close to one-half.

The quantile x_p does a similar job. It splits the possible values of X into disjoint sets, so that one has $P(X < x_p) \leqslant p \leqslant P(X \leqslant x_p)$. When $p = 1/2$, the quantile x_p reduces to the median, and, when $p = 1/4$ or $p = 3/4$, then x_p is called the 'lower quartile,' or the 'upper quartile,' respectively.

4. Mass Functions and Densities

If a random variable X only takes values from a finite or countable set S, then X is called a 'discrete' random variable. In such cases, one can also give a complete description of the probabilistic properties of X by specification of the 'probability mass function,' which is defined by setting $f(x) = P(\{x\})$ for all $x \in S$. Here, one should note that the probability mass function f, permits one to recapture the distribution function by the relation

$$F(t) = \sum_{x \leqslant t, x \in s} f(x) \quad \text{for all } t$$

There are many important random variables whose values are not confined to any countable set, but for some of these random variables, there is still a description of their distribution functions that is roughly analogous that provided for discrete random variables. A random variable X is said to be 'absolutely continuous' provided that there exist a function f such that the distribution function of X has a representation of the form

$$F(t) = \int_0^t f(x)\,\mathrm{d}x \quad \text{for all } t$$

The function f in this representation is called the 'density' of X, and many random variables of practical importance have densities. The most famous of these are the 'standard normal' (or, 'standard Gaussian') random variables that have the density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for all } -\infty < x < \infty$$

which has a key role in the Central Limit Theorem that will be discussed shortly.

Much of the theory of probability can be developed with just the notion of discrete or absolutely continuous random variables, but there are natural random variables that do not fit into either class. For this reason, the distribution function remains the fundamental tool for describing the probabilities associated with a random variable; it alone is universal.

5. Mathematical Expectation: A Fundamental Concept

If X is a discrete random variable, then its 'mathematical expectation' (or just expectation, for short) is defined by

$$E(X) = \sum_{x \in S} x f(x)$$

where f is the probability mass function of X. To illustrate this definition, one can take X to be the outcome of rolling a fair die, so that $f(x) = 1/6$ for all $x \in S = \{1, 2, \ldots, 6\}$ and

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

In a way that is parallel—yet not perfectly so—the expectation of an absolutely continuous random variable X is defined by the integral

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

Here one needs to work a bit harder to illustrate the definition. Consider, for example, an experiment where one chooses a number at random out of the unit interval [0, 1]. From the intuitive meaning of the experiment, one has $P(A) = a$ for $A = [0, a]$, and, from this relationship, one can calculate that the density function for X is given by $f(x) = 1$ for $x \in [0, 1]$ and by $f(x) = 0$ otherwise. For the density f, one therefore finds that

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \int_0^1 x \cdot 1\,\mathrm{d}x = \frac{1}{2}$$

From this formula, one sees that the expected value of a number chosen at random from the unit interval is equal to one-half, and for many people this assertion is perfectly reasonable. Still, one should note that the probability that X equals one-half is in fact equal to zero, so the probabilistic use of the word 'expected value' differs modestly from the day-to-day usage.

The probability distribution function and the expectation operation provide almost all of the language that is needed to describe the probability theory of an individual random variables. To be sure, there are

several further notions of importance, but these may all be expressed in terms of the distribution or the expectation. For example, the most basic measure of dispersion for the random variable X is its 'variance,' which is defined in terms of the expectation by the formula

$$\mathrm{Var}(X) = E(X-\mu)^2 \quad \text{where } \mu = E(X)$$

Finally, the 'standard deviation' of X, which is defined to be the square root of Var(X), provides a useful measure of scale for problems involving X.

6. Introducing Independence

The world of probability theory does not stop with the description of a single random variable. In fact, it becomes rich and useful only when one considers collections of random variables—especially collections of 'independent' random variables.

Two events A and B in $\mathscr{F}$ are said to be independent provided that they satisfy the identity,

$$P(A \cap B) = P(A)\,P(B)$$

What one hopes to capture with this definition is the notion that the occurrence of A has no influence on the occurrence of B—and vice versa. Nevertheless, the quantities that appear in the defining formula of independence are purely mathematical constructs, and any proof of independence ultimately boils down to the proof of the defining identity in a concrete model $(\Omega, \mathscr{F}, P)$.

Consider, for example, the experiment of rolling two fair dice, one red and one blue. The sample space Ω for this experiment may be taken to be the 36 pairs (j, k) with $1 \leqslant j \leqslant 6$ and $1 \leqslant k \leqslant 6$, where one thinks of j and k as the number rolled on the red and blue die, respectively. Here, for any $A \subset \Omega$ one can set $P(A) = |A|/|\Omega|$ in order to obtain a model for a pair of fair dice. Under this probability model, there are many pairs of events that one can prove to be independent. In particular, one can prove that the event of rolling an even number on the blue die is independent of rolling an odd number on the red die. More instructively, one can also prove that the event of rolling an even number on the blue die is independent of the parity of the sum of the two dice.

7. Extending Independence

The concept of independence can be extended to random variables by defining X and Y to be independent provided that the events $\{X \leqslant s\}$ and $\{Y \leqslant t\}$ are independent for all real s and t. One easy consequence of this definition is that for any pair of monotone functions ϕ and ψ the random variables $\phi(X)$ and $\psi(Y)$ are independent whenever X and Y are independent. This mapping property reconciles nicely with the intuitive assertion: if X and Y have no influence on each other, then neither should $\phi(X)$ and $\psi(Y)$ have any influence on each other.

The importance of independence for probability theory will be underscored by the theorems of Sect. 8, but first the notion of independence must be extended to cover more than just pairs of random variables. For a finite collection of n random variables $X_1, X_2, \ldots, X_n$, the condition for 'independence' is simply that

$$\begin{aligned} &P(X_1 \leqslant t_1, X_2 \leqslant t_2, \ldots, X_n \leqslant t_n) \\ &= P(X_1 \leqslant t_1)\,P(X_2 \leqslant t_2) \cdots P(X_n \leqslant t_n) \end{aligned}$$

for all real values $t_1, t_2, \ldots, t_n$. Finally, an infinite collection of random variables $\{X_s : s \in S\}$ is said to be independent provided that every finite subset of the collection is independent.

8. The Law of Large Numbers

Any mathematical theory that hopes to reflect real-world random phenomena must provide some rigorous interpretation of the intuitive 'law of averages.' Kolmogorov's theory provides many such results, the most important of which is given in the following theorem.

Theorem 1 (Law of Large Numbers). Suppose that $\{X_i : 1 \leqslant i < \infty\}$ is a sequence of independent random variables with a common distribution function $\mathscr{F}(\cdot)$, so $P(X_i \leqslant t) = \mathscr{F}(t)$ for all $1 \leqslant i < \infty$ and all real t. If the expectation $\mu = E(X_1)$ is well-defined and finite, then the random variables

$$\frac{1}{n}\{X_1 + X_2 + \cdots + X_n\}$$

converge to their common mean μ with probability one. More precisely, if one lets

$$A = \left\{ \omega : \lim_{n\to\infty} \frac{1}{n}\{X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)\} = \mu \right\}$$

then the event A satisfies $P(A) = 1$.

9. The Central Limit Theorem

The second great theorem of probability theory is the famous Central Limit Theorem. Although it is not tied as tightly to the meaning of probability as the Law of

Large Numbers, the Central Limit Theorem is key to many of the practical applications of probability theory. In particular, it often provides a theoretical interpretation for the 'bell curve' that emerges in countless empirical investigations.

Theorem 2 *(Central Limit Theorem)*. Suppose that $\{X_i: 1 \leqslant i < \infty\}$ is a sequence of independent random variables with a common distribution function $\mathscr{F}$. If these random variables have a finite variance $\mathrm{Var}(X_i) = \sigma^2$, then

$$\lim_{n\to\infty} P\left(\frac{1}{\sigma\sqrt{n}}\{X_1+X_2+\cdots+X_n-n\mu\} \leqslant x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}\mathrm{d}u$$

10. Stochastic Processes

The most fundamental results of probability theory address the behavior of sums of independent random variables, but many applications of probability theory lead to sequences of random variables $\{X_n: 1 \leqslant n < \infty\}$ that may not be independent. Such sequences are called 'stochastic processes,' and they are of many different types.

The simplest nontrivial stochastic process is the 'Markov chain,' which is used to model random phenomena where X_{n+1} depends on X_n, but, given X_n, the value of X_{n+1} does not depend on the rest of the past $X_{n-1}, X_{n-2}, \ldots, X_1$. To construct such a process, one most often begins with an $n \times n$ matrix $T = \{p_{ij}\}$ with entries that satisfy $0 \leqslant p_{ij} \leqslant 1$ and with row sums that satisfy $p_{i1}+p_{i2}+\cdots+p_{in} = 1$. One then generates the Markov chain by making sequential selections from the set $\mathscr{S} = \{1, 2, \ldots, n\}$ in accordance with the rows of the 'transition matrix' T. Specifically, if $X_n = i$, then X_{n+1} is obtained by choosing an element of $\mathscr{S}$ in accordance with the probabilities (p_{ij}) given by the ith row of T.

A second group of stochastic processes that is of considerable practical importance is the set of 'martingales.' Roughly speaking, these are stochastic process that have the property that the expectation of X_{n+1} given the values of $X_n, X_{n-1}, \ldots, X_1$ is just equal to X_n. One reason that martingales are important is that they provide a model for the fortune of an individual who gambles on a fair game. Such gambling games are relatively unimportant in themselves, but many economic and financial questions can be reframed as such games. As a consequence, the theory of martingales has become an essential tool in the pricing of stock options and other derivative securities.

The theory of stochastic processes is not confined to just those sequences $\{X_n: 1 \leqslant n < \infty\}$ with the discrete index set $\{1 \leqslant n < \infty\}$, and, in fact, almost any set $\mathscr{S}$ can serve as the index set. When one takes $\mathscr{S}$ to be the set of non-negative real numbers, the index is often interpreted as time, and in this case one speaks of continuous time stochastic processes. The most important of these are the 'Poisson process' and 'Brownian motion.' Brownian motion is arguably the most important stochastic process.

11. Directions for Further Reading

For a generation, Feller (1968) has served as an inspiring introduction to probability theory. The text only assumes a modest background in calculus and linear algebra, yet it goes quite far. The text of Dudley (1989) is addressed to more mathematically sophisticated readers, but it contains much material that is accessible to readers at all levels. In particular, Dudley (1989) contains many informative historical notes with careful references to the original sources.

For an introduction to the theory of stochastic processes, the text by Çinlar (1975) is recommended, and for an easy introduction to Brownian motion, martingales, and their applications in finance, one can consult Steele (2000).

For background on the early development of probability theory, the books of David (1962) and Stigler (1986) are useful, and the article by Doob (1994) helps make the link to the twenty-first century. Finally, for direct contact with the master, anyone does well to read Kolmogorov (1933).

See also: Probability and Chance: Philosophical Aspects; Probability Density Estimation; Probability: Interpretations

Bibliography

Çinlar E 1975 *Introduction to Stochastic Processes*. Prentice-Hall, Englewood Cliffs, NJ

David F N 1962 *Games, Gods, and Gambling: The Origins and History of Probability from the Earliest Times to the Newtonian Era*. Griffin, London

Doob J L 1994 The development of rigor in mathematical probability (1900–1950). In: Pier J-P (ed.) *Development of Mathematics 1900–1950*. Birkhauser-Verlag, Basel, Switzerland

Dudley R M 1989 *Real Analysis and Probability*. Wadsworth-Brooks/Cole, Pacific Grove, CA

Feller W 1968 *An Introduction to Probability and Its Applications*, 3rd edn. Wiley, New York, Vol. 1

Kolmogorov A N 1933 *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer-Verlag, Berlin (trans. Morrison N 1956 *Foundations of the Theory of Probability*. Chelsea, New York)

Steele J M 2000 *Stochastic Calculus and Financial Applications*. Springer, New York
Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA

J. M. Steele

Probability: Interpretations

1. Qualitative and Quantitative Probability

1.1 Introduction

It is entirely unremarkable that the commonplace word 'probability' and its cognates display a wide variety of distinct meanings in ordinary language. It is used as a classificatory concept, as in 'It is *probable that* it will snow this winter' or 'It is *improbable that* a fair coin will land heads 10 times in succession.' It is used as a (qualitative) comparative relation, as in 'It is *more probable that* someone will be killed by lightning *than* by a falling meteorite next year.' And, of course, it is used as a quantitative concept throughout science, e.g., Mendel's law, 'The probability is 0.75 that the offspring of a cross between simple hybrids will show the dominant trait.' What is worth noting, however, is that despite longstanding controversies over the legitimate interpretation of 'probability,' these diverse usages can nonetheless be linked easily to a few formal theories of mathematical probability.

The received formal theory of quantitative probability, due to Kolmogorov, is elegant. Mathematical probability is a (real-valued) function, $\boldsymbol{P}$, with values between 0 and 1 (inclusive), defined over a field of events closed under the Boolean operations of union, intersection, and complementation. The sure event Ω is certain, $\boldsymbol{P}(\Omega) = 1$, and $\boldsymbol{P}$ is additive over unions of disjoint events. The qualitative formal theory of probability uses a binary relation $\langle$ also defined on an algebra of events: $E \langle F$ means that event F is more probable than event E. The qualitative formal theory is presented through axioms that mimic the quantitative one.

The outline for this article is as follows. Sections 1.2 and 1.3 review relevant details of the received, formal (mathematical) theories of quantitative and qualitative probability, and some basic connections between them. (This section includes some technical results that the reader might bypass initially, to return to on a second reading.) Section 2 surveys several interpretations of the quantitative theory, including subjective probability (2.1), logical probability (2.2), objective probability (2.3), and an approach (2.4) here labeled the Fisher–Kyburg theory, relating to Fisher's enigmatic fiducial probability. Additional readings are suggested in Sect. 3.

1.2 The Formal Theory of Quantitative Probability

Begin with Kolmogorov's (1956) mathematical theory for quantitative probability, dating from the 1930s. A measure space $\langle \Omega, \mathscr{F}, \boldsymbol{P} \rangle$ is triple comprised by:

(1) a set of possibilities, Ω, whose elements, ω, are points of the space. A field $\mathscr{F} = \{E \text{ of subsets of } \Omega\}$. Elements of $\mathscr{F}$ are the events of the space. It is assumed that $\Omega \in \mathscr{F}$. Being a field, $\mathscr{F}$ is closed under finite unions, intersections, and complementation with respect to Ω. When $\mathscr{F}$ is infinite, usually it is required

(2) that $\mathscr{F}$ is a σ-field, closed under countable unions and intersections.

(3) A probability function $\boldsymbol{P}$ over $\mathscr{F}$ satisfying these axioms:

(3.1) $0 \leqslant \boldsymbol{P}(E) \leqslant 1$

(3.2) $\boldsymbol{P}(\Omega) = 1$

(3.3) $\boldsymbol{P}(E \cup F) = \boldsymbol{P}(\mathrm{E}) + \boldsymbol{P}(\mathrm{F})$, whenever $\mathrm{E} \cap \mathrm{F} = \varnothing$.
(finite additivity)

(3.4) $\boldsymbol{P}(\cup E_i) = \Sigma_i \boldsymbol{P}(E_i)$ whenever $E_i \cap E_j = \mathrm{F}\ \varnothing$ for $i \neq j$. (σ-additivity)

Moreover, when $\boldsymbol{P}(F) > 0$, a conditional probability given F, $\boldsymbol{P}(\cdot \mid F)$, is well defined over $\mathscr{F}$ by the equation

(4) $\boldsymbol{P}(\cdot \mid F)\boldsymbol{P}(F) = \boldsymbol{P}(\cdot \cap F)$.

Kolmogorov's mathematical theory provides a basis for each of the two nonquantitative uses of probability, illustrated above.

The classificatory concept of probability may be explicated as:
'E is *probable*' provided that $\boldsymbol{P}(E) > c > 0.5$, for some value c, determined by the context of the assertion. Then, 'E is *improbable*' means that the complementary event E^c is probable.
The (qualitative) comparative concept of probability may be explicated as:

'E is *more probable than* F' provided that
$$\boldsymbol{P}(E) > \boldsymbol{P}(F)$$

1.3 The Formal Theory of Qualitative Probability

A mathematical theory of qualitative probability is readily obtained from the quantitative one. The binary relation $\langle$ over events is said to be a (qualitative) comparative probability relation if it satisfies the following three axioms that parallel the first three axioms of Kolmogorov's theory.

(5.1) $\langle$ is an acyclic, transitive relation over $\mathscr{F}$. That is, if $E \langle F$ and $F \langle G$ then $E \langle G$, and it is never the case

that $E \langle E$. The qualitative relation that E and F are equiprobable events, denoted $E \approx F$, is defined by the condition that neither $E \langle F$ nor or $F \langle E$. Then $\approx$ is transitive and reflexive, i.e., $E \approx E$ always obtains.

(5.2) $\varnothing \langle \Omega$ and for no event E, $E \langle \varnothing$.

(5.3) Provided that $E \cap (F \cup G) = \varnothing$, $F \langle G$ if and only if $(E \cup F) \langle (E \cup G)$.

Say that a quantitative probability $\boldsymbol{P}$ agrees with the comparative probability relation $\langle$ on the condition that, $\boldsymbol{P}(E) < \boldsymbol{P}(F)$ if and only if $E \langle F$. DeFinetti (1931) asked whether every comparative probability relation $\langle$ has an agreeing quantitative probability. Kraft et al. (1959) established that the answer is no, even when $\mathscr{F}$ is a finite space. However, they also showed that for finite spaces, the following (stronger) additivity condition is both necessary and sufficient for an agreeing quantitative probability to exist.

(5.4) Let $(E_1, \ldots, E_m)$ and $(F_1, \ldots, F_m)$ be two sequences of events from $\mathscr{F}$ such that, for each point ω in Ω, the same number of events from each sequence contain ω. If it is not the case that $F_i \langle E_i$ $(i = 1, \ldots, m-1)$, then it is not the case that $E_m \langle F_m$.

Aside: It is interesting to note that the very same condition, expressed for a quantitative probability, serves as a necessary and sufficient condition for extending the domain of a quantitative probability, Horn and Tarski (1948).

A different relationship between quantitative and qualitative probabilities is also worth consideration. Say that a quantitative probability $\boldsymbol{P}$ almost agrees with the comparative probability relation $\langle$ on the condition that $\boldsymbol{P}(E) < \boldsymbol{P}(F)$ only if $E \langle F$. That is, quantitative strict inequalities entail the corresponding qualitative (strict) inequality. However, when $\boldsymbol{P}$ almost agrees with $\langle$, it may be that $\boldsymbol{P}(E) = \boldsymbol{P}(F)$ and $E \langle F$. Kraft et al. (1959) provide an example of a finite space equipped with a qualitative probability relation that admits a continuum of almost agreeing quantitative probabilities but no agreeing quantitative probability. For finite spaces, the analog to (5.4) is the following necessary and sufficient condition for a qualitative probability to admit an almost agreeing quantitative probability.

(5.5) Let $(E_1, \ldots, E_m)$ and $(F_1, \ldots, F_m)$ be two sequences of events from $\mathscr{F}$ such that, for each point ω in Ω, fewer events from the first sequence contain ω than do events from the second sequence. If it is not the case that $F_i \langle E_i$ $(i = 1, \ldots, m-1)$, then $E_m \langle F_m$.

On infinite spaces, the situation is more complex and more interesting. The following illustrates how a qualitative probability may make finer discriminations than any quantitative probability, i.e., the qualitative relation $\langle$ is non-Archimedean.

Example: Let $\mathscr{F}$ be the set of all subsets of the positive integers $\{1, 2, \ldots\} = \Omega$. Let $\langle$ be a qualitative probability subject to the following two constraints:

(a) Each integer, i.e., each point is as probable as any other. That is, for each pair of integers i and j, $i \approx j$.

(b) The qualitative probability relation respects proper subsets. That is, whenever $E \subset F$ (but $E \neq F$) then $E \langle F$.

(*Aside*: To establish the existence of a qualitative probability satisfying these constraints requires an appeal to the Axiom of Choice.) Condition (a) ensures that no countably additive probability agrees with $\langle$, since any agreeing probability must assign each point of Ω probability 0. Condition (b) insures that no finitely additive probability agrees with $\langle$, because, by the fact just noted, $\boldsymbol{P}\{i, j\} = \boldsymbol{P}\{i\} + \boldsymbol{P}\{j\} = 0$. But by (b) $\{i\} \langle \{i, j\}$. There exists an almost agreeing (uniform) finitely additive probability for $\langle$, where $\boldsymbol{P}\{i\} = 0$ for each integer i. Hence, this qualitative probability ordering is finer than any quantitative probability can be.

2. *Interpretations of Quantitative Probability*

2.1 *Subjective Probability*

The contemporary theory of subjective (or personal) probability has its roots in the writings by Ramsey (1926) and deFinetti (1937), from the 1920s and 1930s. Savage (1954) gives what to many is a definitive account of this line of reasoning. The subjective interpretation of the mathematical theory treats probability as a (rational) agent's assessment of uncertainty about a truth-value bearing proposition, E. (Here, equate the proposition asserting E and the event E of the mathematical theory.) The agent's uncertainty about E is expressed as her/his degree of belief in E, and under appropriate assumptions may be measured by the agent's disposition to wager on E (see *Bayesian Statistics*).

The so-called 'Dutch Book' argument offers a prudential account of why the rational agent who is willing to engage in betting (under conditions made explicit, below) will have degrees of belief that conform to the formal probability calculus. Suppose that betting is in monetary units.

Definitions:

A bet on E, at odds of $r:(1-r)$, with total stake $S > 0$ pays off to the agent with a gain of $(1-r)S$ in case E obtains, and with a loss of $-r(S)$ in case E fails to obtain.

A bet against E, at odds of $r:(1-r)$, with total stake $S > 0$ pays off to the agent with a loss of $-(1-r)S$ in case E obtains, and with a gain of $r(S)$ in case E fails to obtain.

The agent's betting odds of $r:(1-r)$ for E are fair just in case the agent is indifferent between betting on and betting against E at those odds, at least for moderate size stakes.

A (finite) set of bets constitutes a Dutch Book against the agent if the agent suffers a net loss from these bets regardless which state of affairs obtains.

The agent's fair betting odds are coherent if no Dutch Book is possible against them.

An elementary version of the 'Book' argument is captured in the following result:

The Dutch Book Theorem (deFinetti and Ramsey): Suppose that the rational agent offers fair odds for each event E in the space $\mathscr{F}$ and for moderate size stakes is willing to accept finite combinations of bets made at her/his fair odds. Then the agent is coherent, that is, immune to having a Dutch Book made if and only if her/his fair odds satisfy the axioms of mathematical (finitely additive) probability, (3.1)–(3.3).

Conditional probability, too, can be made the subject of the Dutch Book criterion by introducing called-off bets, e.g., a bet on E given F, which is annulled if F fails to occur.

Definition: A called-off bet on E given F, at odds of $r:(1-r)$, with total stake $S > 0$ pays off to the agent with a gain of $(1-r)S$ in case $E \cap F$ obtains, pays off with a loss of $-r(S)$ in case F obtains but E fails to obtain, and results in no change (0 payoff) in case F fails to obtain.

The concept of a called-off bet against E given F is likewise defined in parallel with the respective concepts for betting against an event, as is the condition of fair called-off odds. When called-off bets are included among those to be wagered on, the following improved result ensues.

Dutch Book Theorem for conditional probability: Assume that the agent offers fair odds and fair called-off odds, and for moderate size stakes is willing to accept finite combinations of wagers at these fair odds. Then the agent is coherent if and only if the fair odds satisfy the axioms of mathematical (finitely additive) probability (3.1)–(3.3), and the definition of conditional probability, (4).

● *Remark*: The reader is alerted to the sad fact that when $\boldsymbol{P}(F) = 0$, condition (4) does not establish that $\boldsymbol{P}(\cdot \mid F)$ is a probability over $\mathscr{F}$, i.e., it may fail axioms (3.1)–(3.3). Hence, the Dutch Book argument for conditional probability does *not* establish that, given an event F of probability 0, the coherent agent's called-off fair odds given F constitute a probability.

● *Remark*: Some writers use a dynamic interpretation of the conditional probability, so that $\boldsymbol{P}(\cdot|F)$ designates the agents subsequent degrees of belief upon learning that event F obtains. The Dutch Book argument for conditional probability does not underwrite this dynamic interpretation. It is part of the static constraints on a rational agent's degrees of belief at a fixed time. The called-off bets may, however, serve as a model of the agent's hypothetical reasoning at the fixed time.

Shimony (1955) introduces a more restrictive standard of coherence, which he calls strict coherence. An agent's fair odds are strictly coherent provided that there is no possibility of making a 'book' against her/him where the agent stands no chance of winning on net, but risks losing for at least one point of Ω. Shimony establishes that an agent's fair odds are strictly coherent if and only if they are coherent and only the impossible event $\varnothing$ receives probability 0. Despite the desirability of having one's odds be strictly coherent, the resulting standard precludes continuous probability distributions, unfortunately.

Savage (1954) gives an elegant version of personal probability in which the agent's probability is reduced to her/his rational preference over acts. In the briefest outline, Savage's axioms include the assumption that the agent's preferences reveal a qualitative probability, as follows. Let b and w be prizes and assume that the agent prefers receiving the better prize b to the worse one w. For an event E, let the *wager-on-E* be the act that rewards the agent with prize b if E occurs and with w if E fails. Thus, these wagers are generalization of the bets used in deFinetti's Dutch Book argument. Savage postulates that the agent strictly prefers the wager-on-F to the wager-on-E if and only if the qualitative probability relation $E \langle F$ obtains. Savage introduces additional postulates to insure that exactly one (finitely additive) quantitative probability agrees with this qualitative probability. That quantitative probability is offered as the agent's personal probability over events (see *Decision Theory: Bayesian*).

2.2 Logical (or Classical) Probability

Dating to Laplace's (1951) eighteenth-century work, at least, is the view that quantitative probability is determined once the space $\langle \Omega, \mathscr{F} \rangle$ is fixed. In its simplest version, LaPlace's *Principle of Insufficient Reason* states that the points of Ω determine equiprobable states. However, this principle leads quickly to a variety of inconsistencies, which are worth noting.

For one, LaPlace's principle makes the probability of an event depend upon which partition of states is used to describe it. A die's landing with the 1-spot up is made equiprobable with its not landing that way when Ω is just the binary partition; in which case, by LaPlace's principle, the probability is $\frac{1}{2}$ for each outcome. However, if the state space is partitioned into six points depending upon how many spots land facing upwards, LaPlace's principle yields a probability of 1/6 for the outcome with 1-spot showing uppermost. We may continue to refine the state space, to include the information that the sum of the visible side faces on the die (as seen from a given perspective) is either less than, or is equal to, or is greater than the face showing uppermost. Then there are 14 distinct possibilities, but only one of these has a 1-spot uppermost. Then, by LaPlace's principle, the probability of a 1-spot uppermost is 1/14.

When the state space is continuous, the problem of applying LaPlace's principle grows more serious as a transformation to an equivalent random variable may lead to a different probability. For example, if Ω is the set of real numbers in the unit interval [0,1] and $\mathscr{F}$ is

the space of Borel measurable subsets of the identity variable $X(\omega) = \omega$, then LaPlace's principle yields a uniform distribution over X. However, if we transform the space so that $\Omega' = [0, 1]$ with $\omega' = \omega^3$ and define the equivalent random variable $Y(\omega') = \omega'$, then Laplace's principle yields a uniform distribution when applied to the equivalent random variable Y. But a uniform probability distribution over X is evidently inconsistent with a uniform probability distribution over Y.

In an early-twentieth-century variant of LaPlace's idea, Keynes (1921) strove to revitalize the principle by applying it to a formalized language in the fashion of Russell and Whitehead's *Principia Mathematica*. Moreover, Keynes's theory has the added subtlety that takes qualitative probability as primitive, and that relation is only partially defined: Not all propositions can be compared even by a qualitative probability relation in Keynes's system. Following Keynes, Carnap (1950) developed quantitative probability for a logical calculus. At first, Carnap sought to determine a single, well-defined quantitative probability for the monadic predicate calculus. But soon enough his approach, Carnap and Jeffrey (1971), became one best described as subjectivist, favoring symmetry conditions for atomic propositions similar to deFinetti's concept of exchangeability. Papers by Scott and Krauss (1966) and Gaifman and Snir (1982) continue this tradition.

A more productive alternative than appeal to a formalized language to revive LaPlace's principle is the use of mathematical symmetries as a basis for fixing the necessary equiprobable states. Most notable is Jeffreys's (1961) theory of *Invariance*, dating from the 1930s. Jeffreys's sophisticated Bayesian treatment appeals to transformational invariants of parametric statistical models of the observed data to determine a 'prior' probability for the parameter. For example, a uniform 'prior' probability is the invariant distribution for a location parameter, e.g., for the mean μ of data from a normal $\mathbf{N}[\mu, 1]$ distribution. Fraser (1966) uses mathematical groups to determine the necessary invariances for fixing a probability, e.g., Haar measure. Each of these theories provides a *local*, not a *global*, interpretation of probability. That is, these theories provide necessary conditions for probability defined with respect to a specific inference problem. When multiple inference problems are considered at once, e.g., when the same theoretical quantity is a location parameter for one data set and a scale parameter for a second data set, difficulties ensue. These theories then generate what must be seen as either inconsistent or non-unique solutions. Furthermore, the invariant probabilities are not always countably additive. So called 'improper' priors, e.g., a uniform distribution over a real-valued parameter, correspond to a purely finitely additive probability by assigning equal probability to each unit interval for the parameter, which intervals constitute the elements of a countable partition. Dawid et al. (1973) provide an important litany of 'paradoxes' for these theories, involving both transformations of parameters and multiple, simultaneous inferences.

Yet another approach for reviving LaPlace's principle is found in the original work of E. T. Jaynes (1983), who uses entropy as a guide to uncertainty. His principle of *Maximum Entropy* dictates that, subject to a set of constraints (i.e., expectations for bounded random variables), a rational agent should fix probability over the space by choosing that probability that simultaneously satisfies the constraints while maximizing entropy. (For discrete spaces this requires choosing the probability $\boldsymbol{P} = \{p_i : i = 1, \ldots\}$ to maximize $-\sum_i \log(p_i)\, p_i$ subject to the constraints.) For example, when Ω is finite and the constraint set is vacuous, the uniform distribution over Ω maximizes entropy. Unfortunately, the Maximum Entropy principle inherits the problems of LaPlace's principle, as it is sensitive both to repartitioning the space Ω and, in the continuous case, equivalent transformations.

2.3 Objective Probability

2.3.1 Limiting and finite frequency interpretations. Since Venn's (1876) writings in the nineteenth century, at least, it has been popular to attempt to link probability with limiting frequency in infinite sequences of repeated trials, thereby underwriting an objective interpretation. Von Mises' (1931, 1939) work in the early twentieth century offered a mathematical program for this approach. For von Mises, probability applies to a collective, which is a suitably random sequence of repeated trials. For example, a (hypothetically infinite) sequence of repeated flips with a particular coin that can land either only heads-up or tails-up forms a collective and supports the fact that the objective probability of heads up is $\frac{1}{2}$ provided that:

(a) the limiting frequency in the sequence of the outcome heads up is $\frac{1}{2}$.

(b) the sequence of outcomes is random, i.e., this limiting frequency for heads up is invariant over selections of infinite subsequences.

Several important mathematical problems attend this approach. As is well known, the set of events that have a limiting frequency in a given sequence may fail to form a field (let alone a σ-field). Moreover, when Ω is infinite, even when limiting frequencies do exist they may fail to be σ-additive. However, the formal aspect of von Mises' program that has been given most scrutiny is the theory of randomness.

Evidently, for a sequence to pass muster as random in von Mises's sense, not all infinite subsequences are relevant. That is, assuming the limiting frequencies of the binary events {heads up, tails up} exist and are not either 0 or 1, trivially, there will be, e.g., the infinite subsequence consisting of outcomes with only heads

up, whose limiting frequency for that event is 1, not $\frac{1}{2}$. To sidestep this reduction of the randomness concept, von Mises proposed that only certain families of subsequences are relevant to objective randomness. These subsequences are determined by combinations of place selection rules, e.g., choose every k^{th} trial, and tests for after-effects, e.g., consider the subsequence of outcomes preceded by two-consecutive 'heads up' outcomes. One intuition von Mises offered for these restrictions is that the random sequence should be shown to be immune to sure profits from any feasible gambling system that might be used against it. Wald (1937) showed that a formally consistent theory was possible if the tests for randomness were limited to a countable infinity of rules. Church (1940) suggested using the denumerable class of recursive functions, in an attempt to make the concept of a random sequence language invariant. However, Ville (1939) showed that collectives based on denumerably many selection rules fail to validate some of the basic 'strong law' theorems of Kolmogorov's theory, e.g., the law of the iterated logarithm fails. Also, by appeal to martingale theory, Ville showed that such collectives fail von Mises's intuitive criterion for randomness: they do not preclude successful gambling systems.

Reichenbach (1938, 1949) proposed a rival frequency theory that avoided von Mises's randomness conditions, even permitting its application to finite sequences. In doing so, however, the resulting theory depends ever more heavily on epistemic considerations for its warranted application. For example, a (sequence) that alternates successively between heads up on odd numbered flips and tails up on even numbered flips has a limiting frequency of $\frac{1}{2}$ for each outcome. But, given that fact about the order of outcomes, we would be hard pressed to find a good use for this 'objective' probability when making inferences about such a process.

A more interesting finite-frequency interpretation is based on Kolmogorov's (1963, 1968, 1984) idea for interpreting randomness as informational complexity in finite sequences. A proper account of this approach is beyond the scope of this entry. For first-rate discussions see Fine (1973) and van Lambalgen (1987).

2.3.2 Chance and propensity interpretation. The philosophers Braithwaite (1953), Hacking (1965) and Levi (1980) propose and defend an objective interpretation of probability based on the idea that an objective chance is a theoretical, quantitative disposition for a process or kind of trial to respond when activated. That is, consider a process or kind of trial described as flipping this coin by hand. The chance of its landing heads up on such a process is taken to be an unobserved, quantitative disposition for the coin to respond when flipped. The propensity interpretation is thin, on purpose, in that chances satisfy the mathematical theory of probability by stipulation!

What is particularly novel in these approaches about making chance a theoretical concept is that probability is *not* reduced to observed frequencies, contrary to the frequency interpretation. Rather, the meaning of the theoretical term *chance* is given by rules for testing chance claims (Braithwaite) or by rules linking chance with subjective probability (Hacking and Levi). In the latter case, central to the meaning of chance is a so-called *Direct Inference* principle. Put simply, that principle asserts: the personal probability $\boldsymbol{P}$ for a specific outcome O of the kind of the trail occurring is p, given the information that the chance is p for outcome O on that kind of trial. To wit: $\boldsymbol{P}$ (coin lands heads up on a toss | chance is p that the coin lands heads up on a toss) $= p$.

This treatment of chance as a theoretical term, with it meaning supplied by rules for application of the concept, accords with the Pragmatic turn in philosophy as proposed by Peirce (1932) in the nineteenth century, though Peirce himself favored a frequency interpretation.

2.4 The Fisher–Kyburg Interpretation

Fisher (1956) and Kyburg (1961) introduce an original semantics for probability, focused on the single-case assertion, e.g. (*) The probability is $\frac{1}{2}$ that the next flip of this coin lands heads up.

For this theory, probability is fundamentally epistemological, grounded on both knowledge and ignorance of class-frequencies. The statistician who asserts (*) is required to have:

(a) *knowledge* of frequencies of outcomes of heads in the population of flips with this coin, a population that includes the next flip.

(b) *ignorance* of the statistics of outcomes in a smaller statistical population of coin flips that includes the next flip.

In this sense, their theory takes instances of direct inference (based on frequency information about populations) as the definition of probability.

For both Fisher and Kyburg, this interpretation of probability underwrites an important statistical analysis, e.g., Fisher's fiducial probability. For example, let the random variable X take values from a population with known frequencies approximated by a normal distribution, $\mathbf{N}(\mu, 1)$, but with parameter μ unknown. Then the pivotal quantity $V = (\mu - X)$ has a known distribution in this population. Its values are distributed according to a normal $\mathbf{N}(0, 1)$ distribution. Suppose $X = 7$ is observed on a particular trial. The Fisher–Kyburg theory then asserts that on this same trial, given $X = 7$, the statistician's probability distribution for V is $\mathbf{N}(0, 1)$. This claim rests on the assertion of ignorance on the part of the statistician: the statistician knows of no smaller population that includes this trial with statistics for V different from those of the $\mathbf{N}(0, 1)$ population. The upshot is a

fiducial **N**(7, 1) probability distribution for μ, given $X = 7$. When a random variable X has a continuous, parameteric distribution $F(X, \theta)$, then F itself may be treated as the pivotal quantity, having a known uniform **U**[0, 1] distribution. Fiducial inference about **F**, given the observation that $X = x$, yields probability for the unknown parameter satisfying Fisher's enigmatic formula: $\boldsymbol{P}(\theta|X) = -\partial F/\partial\theta|_{X=x}$. Kyburg's theory has the added generality over Fisher's in that he uses interval-valued frequency information, leading to a theory with lower and upper probability for single-case assertions. Despite the considerable novelty of this approach, it is challenged severely by nuisance parameters. That is, this interpretation of probability provides only a limited account of marginal probability, as its theory of conditional probability is not well defined, at least not as required by Bayes's theorem.

3. Additional Readings

Nagel's (1936, 1939) essays on quantitative probability are excellent reviews, still, of philosophical issues relating to rival interpretations. In particular, one important issue not addressed here is whether objective probability requires a commitment to indeterminism in science. Nagel makes clear why there is no conflict between objective probability and deterministic laws in science. A more elementary presentation of rival interpretations of quantitative probability is found in Kyburg's (1970) textbook.

Fine's (1973) book gives a good overview of complexity theories and randomness, relating in particular to Kolmogorov's theory from the 1960s.

Fishburn's 1968) article offers a wealth of insights on both philosophical and formal questions relating qualitative and quantitative concepts of probability. The bibliography there is particularly noteworthy. Also valuable is the discussion of comparative probability given in Walley and Fine (1979).

See also: Probabilistic Thinking; Probability and Chance: Philosophical Aspects; Probability Density Estimation; Probability: Formal

Bibliography

Braithwaite R B 1953 *Scientific Explanation*. Cambridge University Press, Cambridge, UK

Carnap R 1950 *Logical Foundations of Probability*. University of Chicago, Chicago

Carnap R, Jeffrey R (eds.) 1971 *Studies in Inductive Logic and Probability*. University of California Press, Berkeley, CA, Vol. 1

Church A 1940 On the concept of a random sequence. *Bulletin of the American Mathematical Society* **46**: 130–5

Dawid P, Stone M, Zidek J V 1973 Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society* **35**: 189–233 (with discussion)

deFinetti B 1931 Sul significato suggettivo della probabilita. *Fund Math* **17**: 298–329

deFinetti B 1937 *Foresight: Its Logical Laws, Its Subjective Sources*. Trans. in Kyburg and Smokler (eds.) 1964

Fine T 1973 *Theories of Probability*. Academic Press, New York

Fishburn P C 1986 The axioms of subjective probability. *Statistical Science* **1**: 335–58 (with discussion)

Fisher R 1956 *Statistical Methods and Scientific Inference*. Hafner Press, New York

Fraser D A S 1968 *The Structure of Inference*. John Wiley, New York

Gaifman H, Snir M 1982 Probabilities over rich languages, testing and randomness. *Journal of Symbolic Logic* **47**: 495–548

Hacking I 1965 *Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK

Horn A, Tarski A 1948 Measures in Boolean algebras. *Transactions of the American Mathematical Society* **64**: 467–97

Jaynes E T 1983 In: Rosenkrantz R (ed.) *Papers on Probability, Statistics, and Statistical Physics*. D. Reidel Publishing, Dordrecht, The Netherlands

Jeffreys H 1961 *Theory of Probability*, 3rd edn. Oxford University Press, London

Keynes J M 1921 *Treatise on Probability*. Macmillan, London

Kolmogorov A N 1956 *Foundations of Probability*. Chelsea Publishing, New York

Kolmogorov A N 1963 On tables of random numbers. *Sankhya* **25**: 369–76

Kolmogorov A N 1968 Three approaches to the definition of the concept of 'amount of information'. *Sel. Transl. Math. Stat. and Prob* **7**

Kolmogorov A N 1984 The logical basis for information theory and probability theory. *IEEE Trans. Inf. Theory*, **IT 14**: 662–4

Kraft C, Pratt J, Seidenberg A 1959 Intuitive probability on finite sets. *Ann. Math. Stat.* **30**: 408–19

Kyburg H E 1961 *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middleton, CT

Kyburg H E 1970 *Probability and Inductive Logic*. Collier–Macmillan, London

Kyburg H E, Smokler H E (eds.) 1964 *Studies in Subjective Probability*. John Wiley, New York

LaPlace P S 1951 *A Philosophical Essay on Probabilities*. Dover, New York

Levi I 1980 *The Enterprise of Knowledge*. MIT Press, Cambridge, MA

van Lambalgen M 1987 *Random Sequences*. University of Amsterdam, Amsterdam

Mises R von 1931 *Wahrscheinlichkeitsrechnung und ihre Anwendungen in der Statistik und theoretischer Physik*. Deuticke. Leipzig

Mises R von 1939 *Probability, Statistics, and Truth*. Macmillan, New York (Translation of [1928] *Wahrscheinlichkeit, Statistik, und Wahrheit*. Wien.)

Nagel E 1939 *Principles of the Theory of Probability*. University of Chicago Press, Chicago

Nagel E, Margenauf L, Ducasse C J, Wilks S S 1936 The meaning of probability. *Journal of the American Statistical Association* **31**(193): 10–30

Peirce C S 1932, 1936 *Collected Papers*. Harvard University Press, Cambridge, MA, Vols, 2 and 6

Ramsey F P 1926 Truth and probability. In: Braithwaite R B (ed.) *The Foundations of Mathematics 1950*. Humanities Press, New York

Reichenbach H 1938 *Experience and Prediction*. University of Chicago Press, Chicago
Reichenbach H 1949 *Theory of Probability*. University of California Press, Berkeley, CA
Savage L J 1954 *The Foundations of Statistics*. John Wiley, New York
Scott D, Krauss P 1966 Assigning probabilities to logical formulas. In: Hintikka J, Suppes P (eds.) *Aspects of Inductive Logic*. North-Holland, Amsterdam
Shimony A 1955 Coherence and the axioms of confirmation. *Journal of Symbolic Logic* **20**: 644–60
Venn J 1876 *The Logic of Chance*, 2nd edn. Macmillan, London
Ville J 1939 *Etude critique de la notion de collectif*. Gauthiers-Villars, Paris
Wald A 1937 Die Widerspruchsfreiheit des Kollektivsbegriffs der Wahrscheinlichkeitsrechnung. *Ergebnisse eins mathematisches Kolloquiums* **8**: 38–72
Walley P, Fine T 1979 Varieties of modal (classificatory) and comparative probability. *Synthese* **41**: 321–74

T. Seidenfeld

Problem Selection in the Social Sciences: Methodology

Social methodology textbooks state techniques of proper empirical analysis. For instance, they tell how partial coefficients of association are to be computed, adding that these coefficients provide a more severe test of multifactor explanations than bivariate associations. Books on theory building in sociology set standards for determining whether certain hypotheses are better than others. Logically unfalsifiable ones are out, future propositions should provide more information than current ones. Any science starts with problems. Are there precepts for selecting problems too? Yes, and this entry draws upon principles intimated by sociologists (Lenski 1966, p. 20, Blau and Duncan 1967, pp. 9–10, 194–9, 402), proposed by philosophers of science (Bunge 1967, pp. 165–221), and applied by sociologists to sociology (Merton 1959, Ultee 1980). These standards are elucidated by cases from sociology's past and present.

1. *Overarching Problems and Subproblems*

Durkheim 'was not interested in suicide at all,' so says Collins' 1994 textbook on sociological traditions. But then, why did Durkheim dub his 1897 study *Le suicide*? Weber in 1910 grumbled that his critics understood him to explain the rise of capitalism fully from the ethos of protestantism, and retorted that he only dealt with the question of whether this ethos fitted the spirit of entrepreneurial capitalism. Yet in the 1920 preface to his collected papers on world religions, Weber's interest was rationalization processes, including the rise of capitalism. These changes had gone furthest in the West, and Weber wondered why.

If these two stories about sociology's questions carry a moral, the lesson is not that persons nowadays regarded as sociology's founders did not know what they were up to. The message amounts to a methodological command on meritorious problems. Durkheim grasped that seemingly small questions become pertinent when taken as subproblems of larger ones, and Weber understood that it is wise to divide mammoth problems into more modest ones, and those in their turn into even more tractable questions. Durkheim and Weber entered sociology's hall of fame because they pointed towards a bunch of smaller questions falling under one and the same heading. The subtitle of *Le suicide* runs *Étude de sociologie*, lifting a phenomenon from psychology into that science which studies society—and the extreme case of leaving it.

2. *Problem Structures and Their Dynamics*

For Durkheim, a society's suicide rate tells about its cohesion: the higher the rate, the less cohesive a society. Weber observed that a society's inhabitants do not always use their economic freedoms, for instance those of enterprise and labor, and supplemented the problem of how these rights became included in a society's code of law, with the problem of the strength of a person's motivation to use these liberties. Both questions were part of the question of why entrepreneurial capitalism arose in the West. In its turn, the question of the rise of capitalism, including bankers' as well as entrepreneurial capitalism, was a subquestion of the even larger question of the speed of rationalization processes in the West during the last couple of centuries.

In Durkheim's *oeuvre*, the question of a society's degree of cohesion also comprises long-term trends in the prevalence of economic exchanges when comparing ancient Israel, the Roman Empire, and medieval Burgundy, and the degree to which assemblies of hunters and gatherers are religious in nature. Weber's question of rationalization led to less grand ones on the rise of science and the bureaucratization of states. Any methodology of problem selection should contain the rule that questions should not be judged on their own, but according to their place in a structure of overarching problems and subproblems. However small a problem seems, it has more appeal if it is part of an overarching problem. And however towering a question appears, it grows less scary if it is broken down into subquestions.

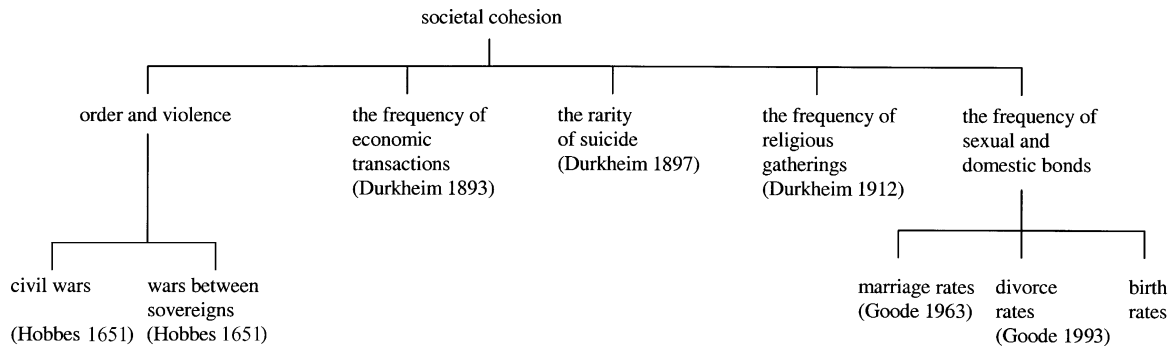

Figure 1
The Durkheimian problem structure and its dynamics

2.1 The Durkheimian Problem Structure

Figure 1 depicts the Durkheimian problem structure. The novelty of *Le suicide* was not simply that Durkheim turned the question of suicide into a subquestion of the larger one of societal cohesion. Hobbes (1651) took order and its opposite violence as the master problem of those studying the body politic. For him, civil wars and wars between sovereigns were part of the problem of order and violence. Next to it, Durkheim placed the problem of the presence and absence of ties between a society's members. The counterpart of the question 'why do people not always live in peace?' is not only 'why do they sometimes use violence?' It also is 'why do they sometimes live relatively isolated from each other?' For Durkheim, suicide was the supreme example of severing the ties between persons and the society in which they live.

Apart from *Le suicide*'s question, Fig. 1 encompasses the questions guiding Durkheim's two other main studies. *De la division du travail social* from 1893 focuses on the frequency of economic transactions between a society's members at various points in history. It narrows this question down to that of the degree to which cooperative law (civil law regulating economic contracts by restitution) is more widespread than criminal law (penal law involving repressive sanctions). *Les formes élémentaires de la vie religieuse* from 1912 notes that Australian aboriginals disperse when hunting and gathering their food, and asks to what extent their leisurely meetings celebrate a totem and turn them into members of one society. Durkheim's questions involve a shift from questions on economic ties, by way of questions on co-living, to questions on religious ties, yet they always pertain to societal cohesion.

Given the Durkheimian problem structure, later scholars can raise questions on types of attachment to which Durkheim paid limited or no attention. Examples are questions on changes in the rates of marriage, birth and divorce in industrial societies. A wedding creates a sexual and domestic bond between two members of the opposite sex, divorce dissolves it, and a higher birth rate amounts to more parent-child relations in a society. Goode (1963, 1993) focuses on the making and undoing of sexual and domestic ties. These monographs amount to a new branch of the Durkheimian problem structure.

2.2 The Weberian Problem Structure

The Weberian problem structure and its dynamics are given in Fig. 2. The overarching problem is that of rationalization, and Merton (1938) pursued the subproblem of the rise of science in the West, and Blau (1974) that of bureaucratization. The art historian Gombrich (1960) studied questions on the skills of painters to furnish a faithful representation.

The relation between problems as depicted in Figs. 1 and 2, are logical in nature. An overarching problem, usually by way of uncontroversial auxiliary assumptions, implies subproblems. It is another, and less important, matter whether a person raising a subproblem of the overarching problem of an earlier scholar, studied under that master or read a book by him or her. Merton and Blau took the lead from Weber, Gombrich was not influenced by him. Weber invokes Heinrich Wölfflin once when discussing the question of representational skills, Gombrich quotes Wölfflin frequently.

As stated, the step from an overarching question to a subquestion involves largely undisputed assumptions. Few would argue that capitalism is less efficient than slavery. Yet this proposition was belied by Fogel and Engerman's (1974) test for the American South shortly before slavery was abolished. A scholar working on a question can make progress by replacing these auxiliary assumptions by better approximations. Weber did not do so, but Durkheim did. In countries prescribing suicide under certain circumstances, suicide is not an instance of low cohesion, but of

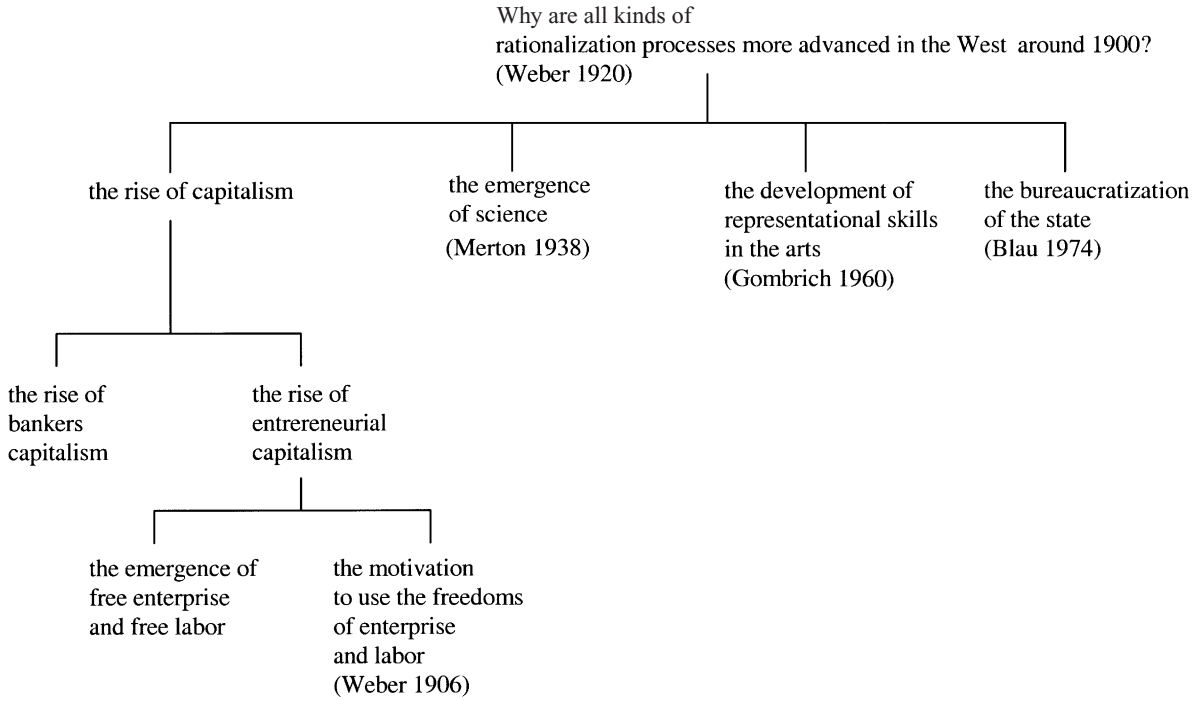

Figure 2
The Weberian problem structure and its dynamics

strong solidarity. In that case, Durkheim spoke of altruistic suicide.

Behind Weber's distinction between bankers' and entrepreneurial capitalism lurks evidence speaking against the hypothesis that the rise of capitalism is always accompanied by the growth of banking. Protestant businessmen did not take out loans. Their religion put a premium on frugality, and contributed to reinvestment of profits in their enterprise. The background of Weber's distinction between laws granting people economic freedoms, and the use of these freedoms by a society's members, forms the observation that people sometimes do not use their rights, as well as the hypothesis that they exercise these rights only if they are motivated to do so. Classical economics assumed that every person had this drive, Weber held that protestantism strengthened it. A wise rule is to state the background of a question as fully as seems necessary for bringing out a problem structure.

2.3 The Weberian Problem Structure Radically Revised

The division of an overarching problem into subproblem may drastically alter. Studies by Smith (1776), Elias (1939), Samuelson (1954), Van den Doel (1979), and De Swaan (1988) imply the revision of the Weberian problem in Fig. 3. Weber never made clear what state bureaucracies were doing. States possess the legitimate monopoly on the means of power, Elias added that states wield the legitimate monopoly on taxation. Were all taxes spent on armies, police and prisons? This obviously is not the case. Here the hypothesis is useful that markets make for the optimal production of individual goods (Smith), but not for the optimal production of collective goods (Samuelson). The production of collective goods approaches optimality in a state with universal suffrage and a bureaucracy which executes its laws (van den Doel).

De Swaan listed collective bads and laws contributing to their elimination and to the production of collective goods. Firstly, poverty is a collective bad. Poor people are more likely to trespass the law, and crimes form a general threat to life and property. Following De Swaan, this danger is alleviated by state insurance against disability and unemployment. Contagious diseases amount to a second collective bad. If local authorities provide each household with piped water and an outlet on the sewers, public health improves. Thirdly, compulsory schooling lowers the percentage of children who cannot read and write, and makes for a more optimal functioning of markets and bureaucracies. Weber's conception of the state is saved by the assumption that armies, police and prisons

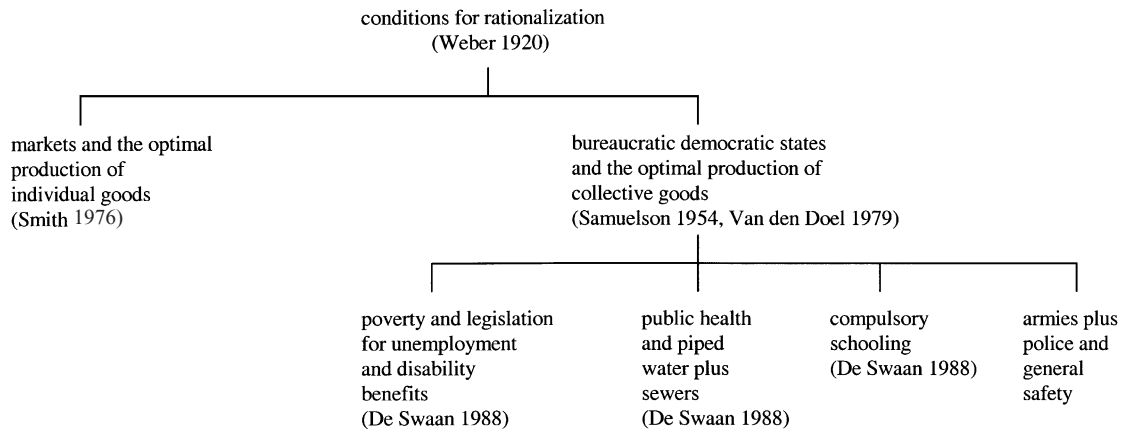

Figure 3
The Weberian problem structure radically revised

avert collective bads. Armies lessen a threat to a state's independence, police patrols contribute to safety on the streets, prisons deter.

2.4 Problem Structures and Theories

As the saying goes, a question well put is a question half answered. If it is clear that some problem is part of a larger problem, and if a solution for the overarching problem is at hand, a possible solution to the smaller problem is available too. But then, if an answer to a particular smaller question of a larger one turns out to be false, the opportunities to disestablish a solution to another subproblem of the larger one increase. Additionally, the notion that problem structures exist and have a dynamic, makes the project of sociology have more continuity than suggested by Boudon (1979) and Gove (1995). They accepted that sociology's questions to outsiders show a bewildering variety, and bypassed the possibility that one and the same larger question may harbor a host of seemingly disparate questions.

3. Contradictions and Issues

One way of formulating a problem in a more useful way is to ascertain its place in a problem structure. Are there much injunctions for problem choice? Sociologists debating sociology have argued about the primacy of theory vs. the supremacy of research. In contrast, Popper (1962) asserted that every science, including sociology, most fruitfully starts with contradictions between general hypotheses and specific research findings.

Contradictions—or, to use Kuhn's 1962 terms, puzzles and anomalies—contribute to sociology's growth. For Merton's (1973) sociology of science, the prevalence of priority disputes in a field was surprising in several respects. Its explanation was not the egotism of the scientists having made discoveries, since their part in the disputes was small. Nor was it to be accounted for by the increasing competition between scientists, since the frequency and intensity of priority disputes in the West decreased over time. A rule of method runs: single out contradictions.

Incompatibilities between hypotheses also invite new research and theorizing. They sometimes are called issues, but that term also refers to false alternatives. An example is the question of whether intelligence is determined by social or by genetic factors. Intermediate positions are possible, the list of social factors may be incomplete, and the location of intelligence on genes may be unknown. Popular science states issues in a metaphorical way: are we born with a brain hardwired with rules for language, or is the brain a big sponge, capable of creating rules for language by mopping up what we hear?

The phrasing of a contradiction problem sometimes contains seeds for its solution. For instance, Huizinga (1941) asked how its was possible that painting and other forms of culture flourished in the Netherlands during the seventeenth century, so shortly after this society became independent from Spain. His answer was that Dutch culture blossomed because the Netherlands was a new state. In states which have recently become independent, the upper social strata have risen from the lower and the more open a country's elite, the more its culture will flower. Trevor-Roper (1976) wondered how it was possible that an emperor as preoccupied with expanding his territory as Charles V (1500–58), furthered the art of painting so much. In the end Trevor-Roper surmised that Charles V commissioned paintings to the extent that, and precisely because, they legitimated his political power.

Apart from the rule for theory building to mobilize the information in a contradiction problem, there is the precept to turn the problem around. Rather than asking why the number of unruly passengers on airlines is increasing, it might be worthwhile to ponder why this number did not increase much more strongly until now. A more academic example is as follows. Marxism noted that a lot of workers refrain from joining unions and strikes, and raised the question of how workers come to have a false consciousness. Olson (1965) held that unions and strikes produce collective rather than private goods, applied the *prisoners' dilemma* from game theory in mathematics predicted that rational workers would not strike, and raised the question of why people participate in social movements. However, it should be clear that any theory with some substance to it should explain that some people join labor unions and certain others do not.

4. *Strings of Questions*

The terms problem and question sometimes are used interchangeably, referring to queries of any specificity. However, everyday usage of expressions like object of investigation, task, question and problem, suggest that some inquiries are more focused than others. What does the articulation of problems involve?

4.1 *Problem Articulation*

If it is maintained that sociology is about societies (and not about individuals), it sometimes is said that the object of research has been stipulated. Yet no single feature of this unit has been singled out for explanation. If this property is taken to be, say, the 'functioning' of a society, the object of inquiry turns into what has been called a task. Yet the circumscription of this aspect remains limited, since the term does not focus any particular way of functioning. If a distinction is made between optimal and suboptimal functioning, the study involves a question, with the term question being taken in a narrow sense. Of course, other types of functioning may be distinguished too. Questions in their turn do not allude to possible causes. If research on some sample of societies determines the extent to which less stability in a society's environment goes together with a less optimal functioning of this society, the matter at hand is articulated so fully that the problem now is whether some logically falsifiable hypothesis is true. Putting a question mark behind a falsifiable proposition amounts to posing a problem in some narrow sense of that word.

The example of a string of increasingly articulated queries just presented neglects any difficulty in obtaining measurements for the degree of stability in a society's environment, and the extent to which the functioning of a society approaches optimality. It has been held that questions should refer to directly observable phenomena only, and therefore in the human sciences to outward behavior and not to mental states. This rule is a leftover from a wrongheaded empiricist philosophy of science. All terms to some extent are theoretical, and any proposition whatsoever, whether about acts or thoughts of persons, or processes and situations within of societies, never can be definitely proved.

4.2 *From Descriptive to Explanatory Questions*

There are strings of questions of another kind. A researcher's early questions often are descriptive: how many of the people living in 1972 Britain and born to working-class parents had moved up to the higher classes? A follow-up question might be whether this probability to climb equals the chance to descend from the upper to the lower classes. A yet later question presupposes the answers to both earlier questions, and seeks to ascertain whether upward and downward mobility were as prevalent in 1949 as in 1972. By raising additional descriptive questions, narratives become richer.

At some point researchers leave descriptive questions behind them, to raise explanatory questions. The explanatory question following up on the previous string of descriptive questions ran: why during the 1970s was upward mobility less widespread in Britain than in Sweden, and more widespread than in France? The obvious answer to that question was that left-wing governments make for more upward mobility (Heath 1981, Goldthorpe 1987).

4.3 *From Superficial to Deeper Explanatory Questions*

Some explanations 'scratch the surface only,' others delve deeper. Why would particular ideologies make for more upward mobility? Just as individuals, governments cannot always attain their goals. By which means then do social democrats attain the goal of more upward mobility? Does Sweden have more upward mobility than Britain, because social democrats make children branch out at a later age to schools which demand more or less of pupils? Or is this the case because social democrats pursue a policy of full employment? And then, if the second question fares better in empirical research, why would there be more upwardly mobile persons in a society without unemployed persons?

There is no obvious point at which the hunt for deeper explanations should stop. Some theoreticians

in effect defend that 'the buck stops here.' No one nowadays seriously defends propositions about societies being logically independent from propositions about individuals. However, whether rock-bottom explanations of the features of societies should always refer to the motives and aversions of a society's members, as well as the opportunities they have and the barriers they face, is controversial. In the discussion on 'methodological individualism' (O'Neill 1973) the borderline between methodological rules and substantive requirements was violated. Perhaps it is wise to state this debate no longer in terms of theories explaining societal phenomena, but to settle it by recognizing that shorthand statements of sociology's questions refer to societal features only, and that these questions, if they are stated fully, always refer to the persons making up a society's population, that is, to *individuals* and *societies* (Ultee 1997).

4.4 From Meagre to Thick Descriptive Questions

Some descriptive questions—and their ensuing explanatory questions can be meagre, whereas others aim for lavishness. Thus, it often has been ascertained that no society ever has been or ever will be completely egalitarian. As Lenski (1966, p. 20) maintained, in the debate between the conservative and the radical tradition in sociology, it is possible to make headway by transforming problems stated in a categorical way into problems phrased in variable terms. It may indeed be the case that full equality never has been observed in any society, but is the lowest degree of inequality observed until now the lowest that is possible? And do long-standing differences between highly industrialized countries in their ruling ideology explain differences in their degree of income inequality? Questions about equality that demand 'yes' or 'no', are less precise than questions about degree of inequality.

Following Lenski, progress also can be made by breaking down the compound concepts contained in problems into their constituent elements. Thus the question of whether power—in the singular—determines privilege—in the singular—will not do. Rather, the question is to what extent differences in the consumption of a certain good in a particular society depend on the control of various resources. When answering this question, Lenski distinguished the skills of hunters and gatherers, the size of gardens in horticultural societies and of fields in agrarian societies, the possession of metal weapons, the private ownership of machines, the technical skills of the people designing machines, the range of legitimate command accruing to a political position, and the right to vote in elections. Lenski also suggested that the change from agriculture to industry was accompanied by the shift, from the right to vote in infrequent political elections for only the propertied classes, to universal suffrage in periodic elections. By breaking down compound problems into their constituents, thick descriptions and detailed explanations are obtained.

4.5 Erroneous and Well-stated Questions

If explanatory questions are logically prior to descriptive questions, explanatory questions may be wrong. Answers to questions can turn out false, questions themselves may be mistaken. The question of why in the 1970s social mobility was more widespread in France than in Sweden is erroneous, since in that decade mobility occurred more often in Sweden. The Olson question, of exactly which selective incentive makes persons protest against the hunting of seals, assumes that some selective incentive is always at work, which may not be the case. This example suggests that if a rock-bottom explanation slips into a question, that problem turns into a leading question.

Methodological rules which stipulate conditions banishing erroneous questions, and guaranteeing well-stated ones, seem doomed. It is wise to review the presupposed description, but checking has to stop somewhere. In this respect it is interesting that the answer to an explanatory question always implies that the original question may not have been fully right. When Durkheim asked why protestantism goes together with more suicide than catholicism, he devised the hypothesis that more tightly knit religions show a lower suicide rate. To derive what he wished to explain, Durkheim assumed that catholics are more integrated than protestants. He knew that this assumption was off the mark for England, since the Anglican Church was a state church. Yet in this way he could account for England as the deviant case to the regularity he wished to explain. Malewski (1964) showed with examples from social psychology that every explanation of a regularity predicts conditions under which exceptions occur to the regularity being explained.

Duncan held that the question of factors influencing the chances that the members of a society move up or down poses the issue poorly. The prime factor is the father's occupation. However, the proposition that the lower the occupational prestige of a person's father, the higher this person's chances of upward mobility, forms a truism. After all, the hypothesis thrives on bottom- and ceiling-effects. To test more interesting hypotheses, according to Duncan, questions on change in a person's life should not be stated in terms of some difference score, but as questions on the strength of the relation between a person's earlier and later score, and on the factors influencing that relation.

4.6 Undertheorized Questions

Schütz (1933) raised the question of why individuals understand one another, and did so with a certain

urgency by phrasing his question as 'how it is possible that people understand each other?' This query sounds like a contradiction, but although casual observation suggests that some people in some places and at some times indeed do understand one another, no particular hypothesis has been derived from a specific theory that holds that it is impossible for human beings to understand each other. Thus the theory rendering the question compelling is lacking. The Parsonian problem of how order is possible (Parsons 1937) is undertheorized in a similar way. Any derivation of the hypothesis that society is violent, from the principle that persons follow their interests, involves additional assumptions. These can be wrong, which makes the question of how order is possible less pressing than it sounds. Indeed, a full statement of the question 'how is order possible?' implies the conditions under which order occurs, and those under which it does not.

Recent research addresses the question of whether the association observed in a cross-sectional sample of a society's inhabitants between their year of birth and a particular characteristic, is to be accounted for as an age, cohort, or period effect. As methodologists point out, this question is undertheorized too. The available hypotheses do not predict some age effect, some cohort effect and some period effect. Those hypotheses are so vague that their conjunction is almost, if not fully, logically unfalsifiable. The hypotheses at hand have more substance: if people marry, grow up in lean times, and live through a war, then their values will be less post-materialist (Inglehart 1977). Those variables can be independently ascertained, making statistical models identifiable.

5. *Conclusion*

Ziman (1987) argued that the question of how scientists choose their problems is erroneous, since most researchers do not have a choice. Their problems are pregiven by the organization which employs them. This observation does not do away with the question of problem choice, but makes the question of the formal rules for problem choice by the higher authorities of these organizations more pressing. The methods outlined above should not only be taken as principles for researchers who have a choice, but also as precepts for institutes run on a budget, and for foundations which award subsidies.

In everyday life researchers consider some problems difficult, others significant, and yet others exciting. These notions are difficult to put into formal rules. However, a more difficult problem might be taken as an explanatory question to which several solutions have been proposed, none of which worked or led to closer approximations. Thus, difficult problems are specific kinds of Popperian contradictions. Also, a structure consisting of overarching problems and subquestions, goes some way towards formalizing the intuition that a problem is significant. Perhaps a problem area is called exciting, if, after many a failed attempt, the solution of an overarching problem seems near, because progress is being made on subproblems.

See also: Generalization: Conceptions in the Social Sciences; Normative Aspects of Social and Behavioral Science; Prisoner's Dilemma, One-shot and Iterated; Quantification in the History of the Social Sciences; Theory: Conceptions in the Social Sciences; Traditions in Sociology

Bibliography

Agassi J 1975 *Science in Flux*. D Reidel, Dordrecht

Blau P M 1974 *On the Nature of Organizations*. Wiley, New York

Blau P M, Duncan O D 1967 *The American Occupational Structure*. Wiley, New York

Boudon R 1979 *La logique du social*. Hachette, Paris [1981 *The Logic of Social Action*. Routledge and Kegan Paul, London]

Bunge M 1967 *Scientific Research I, The Search for System*. Springer, Berlin

Collins R 1994 *Four Sociological Traditions*. Oxford University Press, New York

De Swaan A 1988 *In Care of the State*. Polity Press, Cambridge, UK

Durkheim É 1893 *De la division du travail social*. F Alcan, Paris [1933 *The Division of Labour in Society*. Macmillan, New York]

Durkheim É 1897 *Le suicide, étude de sociologie*. Alcan, Paris [1951 *Suicide. A Study in Sociology*. Free Press, Glencoe, IL]

Durkheim É 1912 *Les formes élémentaires de la vie religieuse*. Alcan, Paris [1995 *The Elementary Forms of Religious Life*. Free Press, New York]

Elias N 1939 *Über den Prozess der Zivilisation*. Hans zum Falken, Basel [1994 *The Civilizing Process*. Basil Blackwell, Oxford, UK]

Fogel R W, Engerman S L 1974 *Time on the Cross. The Economics of American Negro Slavery*, 1st edn. Little, Brown, Boston

Goldthorpe J H 1987 *Social Mobility and Class Structure in Modern Britain*, 2nd edn. Clarendon Press, Oxford, UK

Gombrich E H 1960 *Art and Illusion, A Study in the Psychology of Pictorial Representation*. Phaidon, New York

Goode W J 1963 *World Revolution and Family Patterns*. Free Press, Glencoe, New York

Goode W J 1993 *World Changes in Divorce Patterns*. Yale University Press, New Haven, CT

Gove W R 1995 Is sociology the integrative discipline in the study of human behavior? *Social Forces* **72**: 1197–206

Heath A 1981 *Social Mobility*. Crooke, London

Hobbes T 1651/1968 *Leviathan*. [Penguin, Harmondsworth, London]

Huizinga J 1941 *Nederland's beschaving in de 17de eeuw*. Tjeenk Willink, Groningen [1969 *Dutch Civilization in the 17th Century*. Harper and Row, New York]

Inglehart R 1977 *The Silent Revolution*. Princeton University Press, Princeton, NJ

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Lenski G E 1966 *Power and Privilege*. McGraw-Hill, New York

Malewski A 1964 Generality levels of theories concerning human behavior. In: Zetterberg H L, Lorenz G A (eds.) *Symposium on Theoretical Sociology*. Bedminster, Totowa
Merton R K 1938 *Science, Technology and Society in Seventeenth-Century England*. Saint Catherine Press, Bruges
Merton R K 1959 Notes on problem finding in sociology. In: Merton R K, Bloom L, Cottrell L S Jr (eds.) *Sociology Today: Problems and Prospects*. Basic New York
Merton R K 1973 *The Sociology of Science*. University of Chicago Press, Chicago
Olson M Jr 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
O'Neill J (ed.) 1973 *Modes of Individualism and Collectivism*. Heinemann Educational, London
Parsons T 1937 *The Structure of Social Action*. McGraw-Hill, New York
Popper K R 1962 Die Logik der Sozialwissenschaften. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* **14**: 233–48. [1976 The logic of the social sciences. In: Adorno T W, Albert H, Dahrendorf R, Habermas J, Pilot H, Popper K R. (eds.) *The Positivist Dispute in German Sociology*. Harper and Row, New York]
Schütz A 1933 Der sinnhafte aufbin *der sozialen Welt*. Springer, Wien [1972 *The Phenomenology of the Social World*. Heinemann, London]
Samuelson P 1954 The pure theory of public expenditure. *Review of Economics and Statistics* **36**: 387–9
Smith A 1776/1970 *The Wealth of Nations*. Penguin, Harmondsworth; Whitestone, Dublin
Trevor-Roper H 1976 *Princes and Artists*. Thames and Hudson, London
Ultee W C 1980 *Fortschritt und Stagnation in der Soziologie*. Luchterhand, Darmstadt, Germany
Ultee W 1997 Do rational choice approaches have problems? *European Sociological Review* **12**: 167–79
Ultee W 1997 Bringing individuals back into sociology. In: Blossfeld H P, Prein G (eds.) *Rational Choice Theory and Large-scale Data Analysis*. Westview Press, Boulder, CO
Van den Doel H 1979 *Democracy and Welfare Economics*. Cambridge University Press, Cambridge
Weber M 1922–3 *Gesammelte Aufsätze zur Religionssoziologie*. Mohr, Tübingen [1930 *The Protestant Ethic and the Spirit of Capitalism*. Allen and Unwin, London]
Winckelmann J (ed.) 1978 *Max Weber. Die protestantische Ethik II. Kritiken und Antikritiken*. Gütersloh, Mohn, Germany
Ziman J M 1987 The problem of 'Problem Choice'. *Minerva* **25**: 92–106
Zuckerman H 1978 Theory choice and problem choice in science. In: Gaston J (ed.) *Sociology of Science*, 1st edn. Jossey-Bass, San Francisco, CA

W. C. Ultee

Problem Solving and Reasoning: Case-based

Case-based reasoning (CBR) is a paradigm of *artificial intelligence* and *cognitive science* that models the reasoning process as primarily memory based. Case-based reasoners solve new problems by retrieving stored 'cases' describing similar prior *problem-solving* episodes and adapting their solutions to fit new needs. As case-based reasoners process each problem, they learn by storing records of their problem solving as new cases to guide future reasoning. CBR research studies the CBR process both as a model of human cognition and as an approach to building intelligent systems. Principles from CBR research serve as a foundation for applied computer systems for tasks such as supporting human decision making, aiding human learning, and facilitating access to electronic information repositories.

1. Case-based Reasoning as Cognitive Model

Initial research on CBR was motivated by observations of how humans remember, reason, and learn from experiences. People who encounter surprising events are often reminded of similar past events and use these remindings to help them understand and respond to the new situations (Schank 1982). For example, a doctor whose patient has an unusual reaction to a medicine might be reminded of a previous patient who had similar reaction. Remembering the treatment used for the prior patient suggests a candidate treatment for the new patient; remembering the outcome of the previous treatment helps to predict the new outcome and warns of potential pitfalls to avoid.

Psychological studies provide evidence supporting human use of CBR for a wide range of problem-solving tasks such as learning programming, performing mathematical problem solving, explaining anomalous events, and decision-making (see Leake 1998) for examples of this work). Another use of CBR is for classification or *interpretation*. For example, in *legal reasoning* in the American legal system, prior cases are the basis for legal judgments: lawyers build arguments about how to interpret rules of law by comparing and contrasting precedent cases to the current situation (see Ashley and Rissland (1988) and *Legal Reasoning Models*).

Research on CBR as a cognitive model studies the knowledge and processes that underlie human CBR, addressing topics such as *memory organization*, *similarity assessment*, and *analogical mapping* (Kolodner 1993, Leake 1998). Computational models developed in this work illuminate functional constraints on reasoning processes and provide testable hypotheses about human reasoning for psychological investigation.

2. Steps in Case-based Problem-Solving

Computer CBR systems have been developed for numerous tasks including planning (Hammond 1989, Veloso 1994), design (Pu and Maher 1997), and explanation (Schank et al. 1994). Regardless of the specific reasoning task, any CBR system retrieves relevant prior cases, adapts their lessons to fit the new situation, evaluates and applies the result, and retains

the results of processing as new cases to guide future reasoning. Thus, computer models of CBR must account for four main processing steps: *retrieval*, *adaptation*, *evaluation/application*, and *storage*.

2.1 Retrieval

In order to facilitate retrieval of relevant cases, stored cases are organized by *indices* characterizing circumstances in which the cases are likely to be relevant. For example, in a medical CBR system to suggest treatments for illnesses, treatment cases could be indexed by the diseases they treat and the characteristics of patients (e.g., age) that tend to determine whether a given treatment is appropriate. Given a new problem to solve, a case-based reasoner first performs *situation assessment* to analyze the input problem and describe it in terms of the indices used to organize cases. Retrieval algorithms use the description to retrieve the cases in memory expected to be most relevant. For example, nearest-neighbor retrieval algorithms select the most similar prior cases according to a metric that weights features by their relative importances. Alternatively, case libraries may be organized into discrimination networks to enable retrievals based on only a small number of feature comparisons. After candidate cases have been retrieved, a finer-grained similarity assessment process is often used to identify the best candidates and to determine differences requiring adaptation

2.2 Adaptation

The case adaptation step revises the lessons of the retrieved case (or revises and combines the lessons of multiple retrieved cases) to generate an appropriate new solution. For example, a doctor might have to adopt a prior treatment by reducing the dosage if the prior patient was an adult and the current patient is a child, or might combine elements of prior treatments to address multiple complaints. Computer CBR systems often use rule-based reasoning to determine how to transform prior solutions. In interpretive CBR, the case-based reasoner does not modify the retrieved case, but instead develops justifications for why the prior interpretation applies (or does not apply) to the new situation.

2.3 Evaluation/Application

Tentative results of case adaptation are evaluated to assess the appropriateness of the new solution and to determine whether additional changes are needed, possibly resulting in a cycle of repeated adaptation and evaluation. Once a solution is judged acceptable, it is applied to the real problem. If the solution fails, it may be repaired, reevaluated, and reapplied.

2.4 Storage

Each processing episode of a CBR system, whether resulting in success or failure, is stored to guide future reasoning. Cases containing successful solutions suggest solutions for similar future problems, while cases storing failures provide warnings of problems to avoid and information on how to avoid them. The storage step packages information about the problem, solution and its results in a new case. It also assigns indices for the case and places it in the case library (also called the 'case base') for future use.

3. Research Issues

3.1 Retrieval and Similarity Assessment

Effective case retrieval depends on appropriate retrieval algorithms, well-organized case bases, and indices that are useful for the current task. In case-based problem solving, cases are indexed by information about the problems they solve. For example, a case-based planning system may index plan cases by the goals they achieve and constraints they satisfy. Indices should be predictive of the relevance of cases and should involve features of problems that are likely to be known to the reasoner; both abstract and concrete features may be useful. A crucial research issue in CBR is developing appropriate *indexing vocabularies* for particular reasoning tasks. In order to retrieve useful cases for novel situations, methods are also needed for reformulating or transforming indices to widen search when no directly relevant cases are available. There is also considerable interest in replacing traditional measures of similarity with similarity criteria that directly reflect pragmatic needs by predicting which prior cases will be easiest to adapt to fit the new situation (e.g., Smyth and Keane 1998). Research on retrieval algorithms investigates methods for achieving rapid retrievals from large case bases (e.g., by exploiting parallel algorithms).

3.2 Case Adaptation

Developing case adaptation rules is often difficult because of knowledge acquisition problems: the needed knowledge may not be available or may be difficult to codify. Consequently, improving case adaptation methods is a fundamental challenge for CBR. Research areas include applying machine-learning techniques to automatically generate case adaptation rules; applying CBR to a CBR system's internal processing, to adapt cases based on prior adaptation experiences rather than rules; using a *derivational analogy* process to account for new circumstances by replaying a prior solution-generation *process*, rather than revising a prior solution (Veloso 1994); and applying artificial intelligence methods

such as constraint-based reasoning to guide the adaptation process. It is important to note that sufficiently powerful case adaptation can enable cases to be used in very novel ways; research on flexible retrieval and adaptation processes has given rise to CBR models of creative problem solving.

3.3 Case-base Maintenance

As CBR systems receive long-term use, keeping the case base accurate becomes an important problem. Likewise, as cases are accumulated, it may be necessary to reorganize or update the case base in order to improve its coverage of important problems and to avoid slow retrievals. This has led to considerable interest in methods for improving case quality and retrieval efficiency by strategic deletion, revision, and addition of cases to a case base. Because cases are increasingly generated by multiple distributed sources (e.g., different offices of an organization), additional issues concern how to update and access distributed case information.

3.4 Integration with Other Methods

CBR systems may be integrated with other systems and approaches—both artificial intelligence systems and other systems—to leverage off the strengths of each. For example, case-based and rule-based reasoning may be integrated with each other, with cases used to guide interpretation of rules or to suggest targets for rule-based reasoning. CBR may also be used for speed-up learning, with initial rule-based problem solving building up a case library that can then be used instead of rules, bypassing costly reasoning from scratch, or to avoid errors by storing exceptions to rules. Conversely, rule-based reasoning may be useful to augment the case library and extend the range of problems that a CBR system can solve.

4. CBR Technology

Applications of CBR include both autonomous and interactive systems. As observed by Kolodner (1993), human users are often adept at evaluating the relevance of prior cases and applying their lessons, but may lack the experience needed to acquire a sufficient set of cases. On the other hand, case adaptation and evaluation are difficult for CBR systems, but practical methods exist for accessing large case libraries. These complementary strengths have led to increasing focus on case-based 'retrieve and propose' systems that support human reasoners by providing them with relevant prior cases to aid their decision making (cf. *Decision Support Systems*). For example, case-based 'help desk' systems are used widely to aid telephone support personnel who diagnose customer problems, and commercial CBR 'shells', are available to facilitate the development of such applications (Watson 1997). As novel problems and solutions are added to the case base during use, the system builds a sharable *corporate memory*, of cases, aiding corporate *knowledge management*.

CBR is also increasingly applied to *electronic commerce*. For example, case-based support systems are used during on-line purchases of computers, to help buyers configure their systems by presenting similar past configurations as a starting point for them to modify. This helps users find a good match to their needs, even if they do not know all the important features of the product they are ordering.

In addition, the CBR cognitive model and CBR principles are guiding the design of educational systems by, for example, suggesting when useful learning is most likely to occur and how to retrieve examples to support that learning. Likewise, CBR is being applied to accessing hypermedia and on-line repositories of information—providing indexing vocabularies and example-based query interfaces—and is being used to focus information retrieval, in order to extract task-relevant information from large-scale information sources. Another active research area, textual CBR, develops methods for facilitating knowledge acquisition by automatically capturing and indexing cases in textual form. Leake (1996) and Lenz et al. (1998) present projects illustrating these applications areas.

5. Relationship to Other Artificial Intelligence Methods

Much artificial intelligence research focuses on methods for generating new solutions from scratch by chaining together general rules (for example, see *Expert Systems in Medicine*). CBR is an interesting alternative for a number of reasons, including: (a) simplifying knowledge acquisition (experts often find it difficult to provide rules for their actions but can easily provide examples, and CBR systems can be fielded with small case libraries and learn additional cases during use); (b) providing natural explanations of decisions (by accounting for decisions in terms of actual previous examples, rather than in terms of rules that users may not understand or trust); (c) increasing problem-solving efficiency by reusing the results of prior reasoning; and (d) facilitating system development in domains that are hard to codify in terms of general rules (e.g., because the rules would have too many exceptions). In CBR, traditional rule acquisition issues are replaced by issues of capturing and representing reasoning episodes, selecting similarity and indexing criteria, and capturing case adaptation knowledge.

The learning done by CBR systems strikes a middle ground between knowledge-poor and knowledge-rich

learning approaches. Unlike knowledge-poor inductive learning methods, CBR exploits domain knowledge (e.g., many CBR systems use explanations to guide indexing and adaptation). Unlike knowledge-rich methods, such as explanation-based generalization, CBR can still be in the absence of a rule-based domain theory.

In its emphasis on building up knowledge by learning, CBR research is in the same spirit as research on neural networks (see *Connectionist Approaches*). However, CBR contrasts in storing and reasoning from examples rather than storing only learned generalizations. Also, CBR performs 'lazy learning.' Rather than immediately generalizing training instances and discarding the examples, as in, for example, explanation-based and neural network learning, CBR stores cases with minimal processing and processes them further only if and when adaptation is needed to solve new problems. In addition, unlike neural networks, case-based reasoning systems can change behavior immediately in response to new examples: a single correct example of dealing with a situation, and no further training, is all that is needed for the system to process that situation correctly in the future.

6. Conclusion

A case-based problem solver solves new problems by retrieving stored problem cases and adapting their lessons to fit new circumstances; it learns by storing new episodes for future use. This process can provide robust reasoning and learning even in domains that are hard to codify in rule-based form. CBR research has provided computational cognitive models of human information processing, artificial intelligence techniques, and information technology for a wide range of task domains.

The CBR cognitive model is discussed in Kolodner (1993) and Leake (1998). A volume edited by Leake (1996) surveys the field with chapters on issues and applications in tasks such as decision aiding, knowledge navigation, design, diagnosis, and education; Kolodner (1993) provides an in-depth examination of key issues; and a volume edited by Riesbeck and Schank (1989) includes chapters on important early CBR research projects and 'micro-versions', of the computer programs those projects developed. The process of developing CBR applications is described by Watson (1997). A volume edited by Lenz et al. (1998) presents chapters on CBR technology for tasks including supporting electronic commerce and textual CBR.

See also: Case Study: Logic; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning

Bibliography

Ashley K D, Rissland E 1988 A case-based approach to modelling legal expertise. *IEEE Expert* **3**(3): 70–7

Hammond, K J 1989 *Case-based Planning: Viewing Planning as a Memory Task*. Academic Press, Boston, MA

Kolodner, J 1993 *Case-based Reasoning*. Kaufmann, San Mateo, CA

Leake D B (ed.) 1996 *Case-based Reasoning: Experiences, Lessons, and Future Directions*. AAAI Press, Menlo Park, CA

Leake D 1998 *Cognition as Case-based Reasoning*. In: Bechtel W, Graham G (eds.) *A Companion to Cognitive Science*. Blackwell, Oxford, UK, pp. 465–76

Lenz M, Bartsch-Spörl B, Burkhard H D, Wess S (eds.) 1998 *Case-based Reasoning Technology: From Foundations to Applications*. Springer, Berlin

Pu P, Maher M L (eds.) 1997 *Issues and Applications of Case-based Reasoning in Design*. Erlbaum, Mahwah, NJ

Riesbeck C K, Schank R C 1989 *Inside Case-based Reasoning*. Erlbaum, Hillsade, NJ

Schank, R C 1982 *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, Cambridge, UK

Schank R C, Riesbeck C K, Kass A (eds.) 1994 *Inside Case-based Explanation*. Erlbaum, Hillsdale, NJ

Smyth, B, Keane, M T 1998 Adaptation-guided retrieval: Questioning the similarity assumption in reasoning. *Artificial Intelligence* **102**(2): 249–93

Veloso M M 1994 *Planning and Learning by Analogical Reasoning*. Springer, Berlin

Watson, I 1997 *Applying Case-based Reasoning: Techniques for Enterprise Systems*. Kaufmann, San Francisco, CA

D. B. Leake

Problem Solving and Reasoning, Psychology of

This article discusses two important aspects of human thinking: how people solve problems and how they use systematic thought to reach conclusions from premises. The two processes are intermingled closely in most everyday human thinking.

Woodworth (1938) wrote:

> Two chapters (of 30!) will not be too many for the large topic of thinking.

Woodworth here understates somewhat the importance he assigned to 'higher mental processes,' for another chapter examines reading, and nine others, memory, and learning. In the second half of the twentieth century, the small role of high-level cognition in experimental psychology has been enlarged greatly by treating thinking as symbolic information processing, formulating theories of thinking as non-numerical computer programs, and introducing new observational methods and instruments to capture

increased temporal detail in thought processes (Newell and Simon 1972).

Woodworth (1938) divided his final two chapters 'according to the historical sources of two streams of experimentation … One stream arose in the study of animal behavior and went on to human problem solving; the other started with human thinking of the more verbal sort.' Two branches of the first stream investigated (a) search as the basic problem solving process (trial-and-error, then selective, search; e.g., Morgan 1894, Thorndike 1898), and (b) insight (e.g., Köhler 1917). Little attention was given to expert problem solving.

The second stream of research, on verbal thinking, included concept attainment (e.g., Hull 1920), logical deduction ('reasoning'), creativity (mostly anecdotal reports on inventors, scientists, and artists), and verbal problem solving, the latter initiated by Selz (1913, 1922) and further developed by Duncker (1935) and De Groot (1946). All of these lines of work were important harbingers of the new information processing psychology, but the connections among them only became apparent later. This article describes today's knowledge of thought processes within the information processing framework down to, but not including, the neural implementation.

1. *Problem Solving*

Modern theories of problem solving have gone through three stages. Most early research viewed problem solving as a search process. In the 1960s, attention began to turn to expert problem solving and its dependence on recognition that evoked knowledge from memory. A decade later, it had become evident that solving complex problems usually also required interacting search in several distinct problem spaces.

1.1 *Problem Solving as Search*

Problem solving is often described as search: moving in a problem space from an initial situation toward a goal situation that satisfies all the conditions for a solution. Problem difficulty was initially thought to vary with the size of the problem space. A space with b choices at each point in the search, and paths s steps in length, would contain b^s different paths.

Choosing a chess move was very difficult, for from a typical current position 10^{20} or more branching paths might lead to stronger or weaker positions, and ultimately to won, lost, or drawn games. As a human being could search only an insignificant fraction of these paths, skill depended on discovering just those that might lead to favorable outcomes. DeGroot (1946), showed that stronger players did not search much more extensively than weaker players, and that even grandmasters rarely looked at more than 100 positions before making a move.

These findings led to a theory of problem solving postulating: (a) an 'understanding' process that generates a problem space from the problem description, (b) a selective (heuristic) search process that uses rules of thumb to follow the most promising paths, and (c) a position evaluation function to assess the relative merits of the positions reached. The process can be repeated, re-evaluating the situation after each step. Several chess programs that deliberately limit themselves to searches of the magnitudes that humans can make demonstrate that chess of credible quality can be played in this way, confirming and extending experimental findings about human players (Simon and Schaeffer 1992).

The same scheme applies to problems generally, not just to chess. Methods for generating paths include: generate-and-test (i.e., trial-and-error), hill-climbing (moving toward higher values of an evaluation function), and means-ends analysis. At each step in means-ends analysis: (a) one or more differences are discovered between the current situation and the goal conditions; (b) using past knowledge and experience, an operator is applied that usually eliminates or reduces differences of the kind discovered; (c) the process is repeated until all differences have been removed and the goal attained.

Suppose the goal is to solve an equation for x: $7x+10 = 4x+22$. The operators add or subtract the same expression from both sides of the equation or multiply or divide both by the same expression. The differences between the given expression and the solution ($x = N$) are the presence of a stand-alone number on the left side of the equation, an expression in x on the right, or a coefficient other than 1 for x on the left. In the equation above, means-ends analysis would discover an unwanted number, 10, on the left, then an expression in x, $4x$, on the right, and would subtract both from the equation, leaving $3x = 12$. Now it would notice x's coefficient, 3, and would divide, obtaining $x = 4$.

1.2 *Knowledge for Solving Problems*

Some problems are hard even though their problem spaces are small. For example, the Chinese ring puzzle, which requires freeing a metal bar from a series of rings, has an unbranching problem space, so that the solution can be reached simply by never reversing direction. Nevertheless, this problem, with seven rings, is seldom solved in less than half an hour. It has been found that people apply 'reasonable' heuristics, e.g., never reconnect a ring that has once been removed. But adherence to this plausible means-ends principle guarantees that the problem will not be solved. Gradually, recognition that rings sometimes have to be reconnected uncovers the solution strategy.

Comparing the problem-solving by experts and novices soon demonstrated that domain-specific knowledge (of which this is a simple example) lay at

the basis of expert skills, connecting problem solving with memory. A large discrimination net in the expert's memory permits sensory recognition of common stimulus patterns in the domain. Thus, a chess grandmaster can recognize perhaps 300,000 configurations of pieces that occur repeatedly in games (e.g., 'fork,' 'open file,' 'weak bishop,' etc.). Associated in memory with each such pattern is knowledge of actions that should be explored whenever the pattern is present. Most 'leaps' in thought called intuitive, insightful, or creative are produced by this knowledge-based pattern recognition (Richman et al. 1996).

The expert, then, recognizes familiar patterns, evoking knowledge that guides search to strong moves in a limited time, searching only a little more than the novice, but along productive instead of unproductive paths.

1.3 Multiple Problem Spaces

As empirical studies advanced, it became apparent that problem solving search employs many different spaces, including a space of potential problem spaces. Searching for solutions may also be preceded by searching for effective generators of actions. Problems of scientific discovery require search in spaces of potentially useful instruments, experimental designs, and hypotheses. The Soar program developed by Newell, with Laird and Rosenbloom (Newell 1990), searches in multiple spaces of these kinds.

2. Reasoning

Reasoning encompasses thinking that is orderly or 'logical': for example, the processes of formal logic and of concept attainment (see *Problem Solving: Deduction, Induction, and Analogical Reasoning*; *Problem Solving and Reasoning: Case-based*). Hence, most problem solving processes can also be called 'reasoning.' In logic, reasoning begins with premises (a situation), and uses rules of inference (actions) to derive conclusions (attain a goal). It may be inductive or deductive.

2.1 Forms of Reasoning

In inductive logic, the premises usually are empirical assertions and the rules of inference produce generalizations of these using terms like 'all,' 'most,' 'probable.' For example, from observation of many white swans, and none of other colors, one may reason inductively to: 'All swans are white,' 'Probably all swans are white,' 'The probability that any swan is white is 99 in 100,' or 'Most swans are white.' At one time or another, all of these sorts of induction have been proposed.

In deductive logic, the initial premises are taken as axioms and the rules of inference are tautologies (e.g., If all A are B and c is an A, then c is a B), so that, in any possible world in which the premises hold, the conclusions must hold.

Reasoning may be used to discover either a concept, an inductive conclusion, or a deductive conclusion; it may also be used to verify such a concept or conclusion, once reached. Discovery is a wholly different process from verification.

Verifying a proof is a deductive process: simply examining each step to assure that it was obtained from previous steps by an application of the rules of inference. It requires no real search. However, discovering a proof is an inductive problem solving process, which involves searching for a path from the premises to the conclusion, using the rules of inference. (The task of the first computer that simulated problem solving, the *Logic Theorist*, was to find, inductively, proofs for theorems in symbolic logic (Newell and Simon 1956)).

In concept attainment, the task is to find, by examining instances, a rule that distinguishes instances from noninstances of the concept: swans from nonswans. This is an inductive task, where the search proceeds in two spaces: a space of possible instances and noninstances and a space of rules (e.g., feature descriptions that can distinguish instances from noninstances). Search in the space of rules generates possible concepts; search in the space of instances generates objects for testing the rules.

2.2 Everyday Reasoning

Much remains to be learned about how human beings reason in their everyday affairs. One attractive hypothesis, only partially substantiated by the evidence, is that they use the same processes that they use for structured problems. Each situation a person encounters evokes from memory a constantly changing array of knowledge, beliefs, and goals, with varying degrees of mutual consistency and inconsistency. Current goals and the problem spaces they evoke trigger search that leads to actions and thence again to changes in the situations and the materials evoked from memory.

The difficulty of understanding such a system largely derives from the size and complexity of memory, with its idiosyncrasies of what it does or doesn't recognize at any moment and does or doesn't know, believe, like, or dislike about the things recognized. The closest approaches to theories of such systems have been simulations (unified theories) of sizable subsets of the cognitive processes (see Richman et al. 1996).

2.3 Reasoning Mechanisms

Sometimes it has been proposed that specific kinds of reasoning employ specialized machinery in the brain,

just as sensory and motor processes have their specialized organs. Chomsky (1976), for example, has claimed that there is a special language facility (not merely language 'regions,' but biological structures specialized to processing language). Rips (1994) has made claims for an inherited machinery for logical reasoning. Comparable claims have been made for 'pictorial' reasoning, using mental imagery.

The alternative hypothesis does not deny the evidences for brain localization and specialized functions but claims that all these faculties rest upon the same basic associative neural processes and organizations.

The latter alternative has been explored by constructing simulations that are capable of performing these functions without specialized mechanisms. Thus, without any specialization for logic, syllogistic reasoning will be carried out by all systems that employ if-then rules (production systems), because whenever all the conditions ('ifs') of a rule are satisfied in memory, the actions ('then') will be executed—a generalization of the classical stimulus-response mechanism. Production systems, conjoined with associative memories, provide a completely general architecture for symbolic processing, 'logical' or not (Eisenstadt and Simon 1997). They can reason about mental diagrams and can combine verbal with visual reasoning.

Concept attainment, as well as various kinds of verbal learning and expert memory performance, can be achieved by a discrimination net which grows, by exposure to stimuli and feedback of results, a network of tests of stimulus features that classify the stimuli. Linked to the discrimination net, and indexed by it, a large associative (semantic) memory holds information about each of its concepts, permitting the kind of problem solving by recognition that was discussed in Sect. 1.2 (Richman et al. 1996).

The evidence available at the turn of the twenty-first century allows no clear choice between specialized mechanisms for special domains of reasoning vs. regional localizations of functions that all employ production systems capable of reasoning and learning (i.e., acquiring new productions) and discrimination nets for recognition, and share an associative semantic memory of knowledge and beliefs, indexed by the net. Whichever turns out to be the actual architecture, there will remain the important task of finding the neural mechanisms that implement it.

See also: Cognitive Theory: ACT; Cognitive Theory: SOAR; Deductive Reasoning Systems; Problem Solving and Reasoning: Case-based; Problem Solving: Deduction, Induction, and Analogical Reasoning; Scientific Discovery, Computational Models of

Bibliography

Chomsky N 1976 *Reflections on Language*. Temple Smith, London

De Groot A D 1946 *Het Denken van den Shaker*. North Holland, Amsterdam [trans. 1965 as *Thought and Choice in Chess*. Mouton, The Hague]

Duncker K 1935 *Zur Psychologie des productiven Denkens*. Springer, Berlin [trans. 1945 as *On Problem-solving*. Psychological Monographs, Vol. 38, no. 5. American Psychological Association, Washington, DC]

Eisenstadt S A, Simon H A 1997 Logic and thought. *Minds and Machines* **7**: 365–85

Ericsson K A, Simon H A 1993 *Protocol Analysis: Verbal Reports as Data*, rev. edn., MIT Press, Cambridge, MA

Hull C L 1920 Quantitative aspects of the evolution of concepts: an experimental study. *Psychological Monographs* **28** (whole No. 123)

Köhler W 1917 *Intelligenzprüfungen an Menschenaffen*. Springer, Berlin

Morgan C L 1894 *An Introduction to Comparative Psychology*, 1st edn. Scribners, New York

Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

Newell A, Simon H A 1956 The logic theory machine. *Transactions on Information Theory*, IT-2(3), 61–79

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Richman H B, Gobet F, Staszewski J J, Simon H A 1996 Perceptual and memory processes in the acquisition of expert performance: the EPAM model. In: Ericsson K A (ed.) *The Road to Excellence*. Lawrence Erlbaum Associates, Mahwah, NJ

Rips L J 1994 *The Psychology of Proof*. MIT Press, Cambridge, MA

Selz O 1913 *Über die Gestze des geordneten Denkverlaufs*. Spemann, Stuttgart

Selz O 1922 *Zur Psychologie des producktiven Denkens und des Irrtums*. Friedrich Cohen, Bonn

Simon H A, Schaeffer J 1992 The game of chess. In: Aumann R J, Hart S (eds.) *Handbook of Game Theory*. Elsevier, The Netherlands, Vol. 1, pp. 2–17

Thorndike E L 1898 Animal intelligence: an experimental study of the associative processes in animals. *Psychological Review*, Monograph Supplements, 2 (serial No. 8)

Woodworth R S 1938 *Experimental Psychology*. Holt, New York

H. A. Simon

Problem Solving: Deduction, Induction, and Analogical Reasoning

1. The Closing of Gaps in Human Problem Solving

Deduction, induction, and analogical reasoning, the three basic forms of logical thought, were derived from natural or everyday situations, where they were originally embedded in commonplace forms of interaction and communication. The differences are explained in the respective paragraphs.

Depending on the context, the German word '*schließen*' ('conclusion') may mean to 'close,' 'conclude,' 'reason,' or 'infer.' Generally speaking, the term implies the bridging or closing of a gap, be this a *literal* gap between two entities or, in the metaphorical sense, a *logical* gap between a problem and its solution. The closing of such metaphorical gaps by means of reasoning or inference usually entails a series of cognitive 'steps.'

Problems can be defined in terms of three characteristic stages: (a) an initial state (e.g., the arrangement of the figures on a chess board, an algebra or geometry exercise, or a balance of trade), (b) an end state (e.g., putting the opponent's king in checkmate, solving the maths exercise, or finding ways of improving the balance of trade), and (c) a well-defined series of cognitive 'steps' by which the initial state is transformed to the end state: the solution to the problem (or one of a number of solutions).

They are still to be found in their original, undifferentiated form as IF-THEN relations: in barter, trade, the manufacture of implements, and the planning of collective action, for example. Once raised to the metalevel of pure thought, these forms of reasoning became a subject of philosophical inquiry in the fields of mathematical logic, linguistics, and psychology. Over the course of history they even became independent areas of scientific systematics. It was only at a very late stage, in the time of the great Greek philosophers, that the three types were differentiated, defined, and each traced back to its logical structure. This represented the first step on the path to European or Western thought and the scientific advances that were entailed. And the first to tread this path were the great trio of Ancient Greek thought—Socrates, Plato, and Aristotle.

2. *Three Great Greek Philosophers and the Foundations of Logic*

In the early Socratic *Dialogues*—preserved in Plato's writings—the processes of reasoning are dominated by IF-THEN relations. These and other dialogues demonstrate the inevitability of human thought. When we are told that all men are mortal and that Socrates is a man, our cognitive processes preclude us from harboring any doubts that Socrates is mortal. We will return to this aspect. First, it should be noted that this kind of logic is not rooted in words. It was one of Plato's fundamental theories that the objects perceived by the senses are but imperfect copies of Ideas, the mental representations now known as concepts. Socrates had already used inductive arguments to formulate definitions of concepts by gradually identifying their characteristic properties or features. Regardless of the countless possible variations in the shape perceived as a triangle, for example, the sum of the angles must always be equal to two right angles, and half the product of baseline and height must always be equal to the area. This is Plato's concept—or Idea—of triangularity.

By applying the principle of dieresis—the breaking down of a concept into ever more narrowly defined subconcepts—Plato developed a general taxonomy of concepts and a system of concept-subconcept relations. This laid the foundation for the Aristotelian system of classification: first the broad general category (*genus proximum*), then the more specific (*differentia specifica*)—for example, a tree is a plant with branches and leaves or needles.

In his treatise *On Interpretation*, Aristotle stripped away the semantic content of propositions, revealing the comparable structures behind the words. 'First we must define the terms "noun" and "verb," then the terms "denial" and "affirmation," then "proposition" and "sentence." Spoken words are the symbols of mental experience and written words are the symbols of spoken words.' It was another 2000 years before Wittgenstein again addressed this issue in such clear terms. Aristotle's approach enabled him to formulate the figures of classical logic, the syllogisms. His theory of syllogistic reasoning drew on Plato's hierarchical ordering of concepts, and a syllogism was defined as 'discourse in which, certain things being stated, something other than what is stated follows of necessity from their being so.' A well-known example is the syllogism: A is true of all Bs. C is an A. Therefore, B is true of C. 'All humans are mortal. Socrates is a human. Therefore, Socrates is mortal.' Negation operates on similar lines: 'No self-luminous bodies are planets. All fixed stars are self-luminous. Therefore, no fixed stars are planets.' These syllogisms were central to philosophical and logical inquiry throughout the Middle Ages and on into modern times.

Although examples with natural terms were mainly used to analyse and dispute the universality of the logical structures, it was not until 1679–90 that a new level of knowledge was attained in the field of logic. Scholars began to ask which inferences can be made when the premises of an argument are derived from prior experience, and do not allow universally valid conclusions to be drawn. This question was first formulated in clear terms by Gottfried Wilhelm Leibniz. Leibniz envisaged a universal formalized language, a '*Mathesis universalis*,' for the general analysis and description of human knowledge, a kind of algebraic representation of logical reasoning in human thought. The system was to encompass even those cases in which the given circumstances and available data did not allow definitive conclusions to be drawn. Leibniz suggested calculating the degree of probability that a logical conclusion would be reliable. Bolzano (1970) followed this line of thought, postulating that the prediction of a conclusion is governed by the conditions of occurrence in a sequence of events. This approaches the twentieth century definition of probability as the likelihood of an event. These are

examples from an era in which the ground was prepared for a new method for acquiring the precise definition of inductive reasoning.

3. *Drawing Conclusions from Insufficient Information: Induction*

Inductive reasoning occurs when a conclusion does not follow necessarily from the available information. As such, the truth of the conclusion cannot be guaranteed. Rather, a particular outcome is inferred from data about an observed sample B′, $\{B'\} \subset \{B\}$, where {B} means the entire population. That is, on the basis of observations about the sample B′, a prediction is made about the entire population B. For example: Smoking increases the risk of cancer. Mr X smokes. How probable is it that Mr X will develop cancer? As this example illustrates, inferential statistics (t-tests, analyses of variance, and other derived forms) are based on this kind of inductive reasoning.

The problems associated with the use of induction in scientific reasoning have been addressed from both philosophical and the mathematical perspective. From the philosophical point of view, the question arises as to what this process—the drawing of universal conclusions about a given phenomenon on the basis of the probability in a simple—actually entails. Logicians, as we now discuss, have responded in a variety of ways (v. Mises, Reichenbach, Keynes, Jeffrey, and Carnap).

Carnap's influential approach to this issue represents a form of compromise. Carnap calculates probability on the basis of accumulated experiential data, where confirmations of one hypothesis as opposed to another are entered into the equation as weights, basically resulting in a kind of weighted mean.

4. *Psychological Assumptions from Bayes' Theorem*

In psychological research (artificial or experimental) many attempts have been made to analyse beliefs about probability on the basis of examples. For example, marbles are randomly selected from a jar containing an unknown number of red and black marbles, and study participants asked to predict the next color to be drawn. Participants' predictions are sometimes based on *a priori* probabilities, i.e., unconditional, fixed relative frequencies. In central Europe, for example, the probability that it will rain in July is very high; the probability that it will snow is very low. These predictions would have to be modified for the month of April, and may even have to be reversed for January. Everyday human decision-making is rarely determined by such absolute probabilities however. This can be illustrated by the example of a horse-racing enthusiast's gambling behavior. Let us assume that the gambler is familiar with the horses taking part in the race and decides to bet that the favorite, horse E, will win. He goes to the tote and sees that horse E is injured and will not be taking part in the race. Our gambler immediately modifies his predictions, selecting the second favorite from the list of starters and placing a bet on this horse instead. This is not necessarily the horse he had previously predicted would come second, however. He might even decide not to bet on second place. Either way, his decisions are still based on estimates of probability, but as a result of the change in circumstances there are now fewer possible choices available. This points to the conditioned probabilities. The *hypothesis set H* for the decision H (horse E) has been reduced by the conditional probability $p(E|H)$. This can be seen in relation to the probability of the alternative hypothesis.

In calculating the so-called a posteriori probability, a comparison is drawn between these two quantities. $p(E|H)$ is weighted with the *a priori* probability:

$$p(E|H) \times p(H)$$

This is then placed in relation to the set of all possible positive or negative outcomes according to probability, resulting in the theorem named after Thomas Bayes:

$$p(H|E) = p(E|H)P(H)/P(E)$$

Many experiments in the field of psychology (see Gluck and Bower 1988) have been conducted to test whether the *a posteriori* probability is indeed as effective in human decision-making behavior as this calculation suggests. While some findings confirm the effectiveness of these components, some significant deviations from the theorem have also been registered. These deviations have provided important psychological insights, however. Edwards and Schmidt (1966), for example, conducted in-depth investigations into the ways people adapt their estimates and decision-making behaviour to the changed circumstances when the given information is modified—H-D Schmidt on the basis of high-risk hits of balls, Edwards using gaming chips and estimates of chance. Both found high levels of interindividual variation in the adaptability of study participants. While some people are flexible with respect to their predictions, many others—the so-called 'conservative estimators'—try to retain their initial predictions for as long as possible.

The findings reported by Kahnemann and Tversky (1973) are also very interesting in this context. They found that study participants largely ignore *a priori* probabilities in their predictions of specified variables, preferring to orientate themselves to conditional probabilities. This result is not counterintuitive, as the

conditional probabilities are the most salient. The so-called Monte Carlo effect is another example of the insensitivity to *a priori* probabilities. Suppose that marbles are drawn with replacement from an urn containing an equal number of red and black marbles. The probability of randomly drawing a marble of either colour is 50 percent. If a red marble has been drawn on a number of consecutive occasions, however, study participants regularly predict that it is more likely that a black marble will be selected next. They overlook the fact that each draw is an independent event, and that the *a priori* probability is independent of the previous draws, i.e., remains constant at 50 percent.

In a fictitious diagnosis experiment, Gluck and Bower (1988) found a high level of correspondence between the estimates made by study participants and the odds calculated using Bayes' formula. On the basis of a medical chart, study participants were asked to judge which of two diseases their (hypothetical) patients were suffering from. The *a priori* probability of one of the diseases was three times higher than that of the other. The conditional probabilities were then varied by the listing of symptoms relevant to the diagnosis. Gluck and Bower found a high degree of correspondence with the Bayes' calculations. When asked to estimate the frequency with which a symptom is associated with one of the two diseases (the rarer one), however, the study participants were unable to make reliable predictions. Hence, the 'diagnosticians' came to a Bayesian conclusion without using Bayesian methods. Gluck and Bower refer to this as implicit knowledge.

5. *Implicit Knowledge in Reasoning*

The question remains of how such a data compilation functions on the cognitive level. It can be assumed that such processes involve the compression of accessible information and that they are of a non-algebraic character. They constitute abstractions about data sets, leading to a kind of blurred conceptual knowledge. Examination of the effects of quantifiers in logical arguments has provided insights into these processes. The *modus ponens* is quite categorical: If P, then A holds. In other words, P implies ($\rightarrow$) A. Given P then A, hence $P \rightarrow A$. If it rains (P), then the lawn will be wet (A). It has rained, therefore the lawn is wet; $P \rightarrow A$. Is the reverse (not P) also true? That is: $\neg P \rightarrow \neg A$? Not necessarily; the lawn may well have been watered with a sprinkler or a hosepipe. It has been shown that, in more abstract contexts, even adults have difficulties in understanding this so-called *modus tollens* (Wason and Johnson-Laird 1972). But does the *modus ponens* also hold for conclusions beginning with the quantifier 'some'? For instance does it follow from 'some As are Bs, and some Bs are Cs,' that, 'some As are Cs'? Not necessarily. For instance, substituting 'women' for As, 'doctors' for Bs and 'men' for Cs, we would arrive at the conclusion that some women are men. Alternatively, assuming that some Bs are As and some Cs are Bs, does it follow that some As are Cs? Again, this is not entirely true, as is immediately apparent when terms from natural language are entered into the equation (cf. Anderson). Wason and Johnson-Laird (1972) have investigated the acceptance of such logical arguments in detail. Cosmides and Tooby (cf. Horgan 1995) have attempted to provide an explanation, postulating that such halo effects on techniques acquired over the course of evolution help to reap the maximum benefits. For example, they suggest that the need to detect deception on the part of one's interactants is an essential incentive for the development of human intelligence. The well-known prisoner's dilemma is another such phenomenon. At issue here is whether players trust in the other player having a positive attitude to sharing, or whether they attempt to appropriate the entire winnings at the risk of losing everything. There is no doubt that the maximum-benefit motive plays a role in numerous decision-making processes. In general, however, the evolutionary influences on logically motivated decision-making behavior are located at a deeper level—in strategies which remain stable over the course of evolution and account for both individual and group interests. And not only in the sense of reaping the maximum social benefits. The so-called Monte Carlo effect (see earlier discussion) is but one example. Since time immemorial, humans have experienced that even prolonged natural processes such as weathering, drought, rain, floods, epidemics, disease, etc. eventually come to an end. And it generally holds that the longer these processes have been operating, the more imminent (i.e., the more probable) their conclusion. Such an approach appears to be the heritage of experiences gathered over the thousands of years of hominid history. The fact that humans prefer to extrapolate to the future—as implied by the *modus ponens* with its deterministic IF-THEN relations—is a favorable predisposition, increasing the chances of survival. It is very simple: It is more important to react to potential future events than to past occurrences. As there is only limited storage space in a modular nervous system, some bodies of knowledge become overgeneralized, perhaps as a result of being classified in line with the *modus ponens*. A similar pattern emerges for the weaker inductions of the *modus tollens*: In order to behave appropriately in one's environment, it is less important to know what can be inferred about the antecedent conditions from the absence of an event than vice versa.

6. *Reasoning and Analogies*

While deductive reasoning, in logic, refers to the *necessary* outcomes of a set of conditions, inductive reasoning is concerned with determining the likelihood of an outcome. Only *possible* conclusions can be

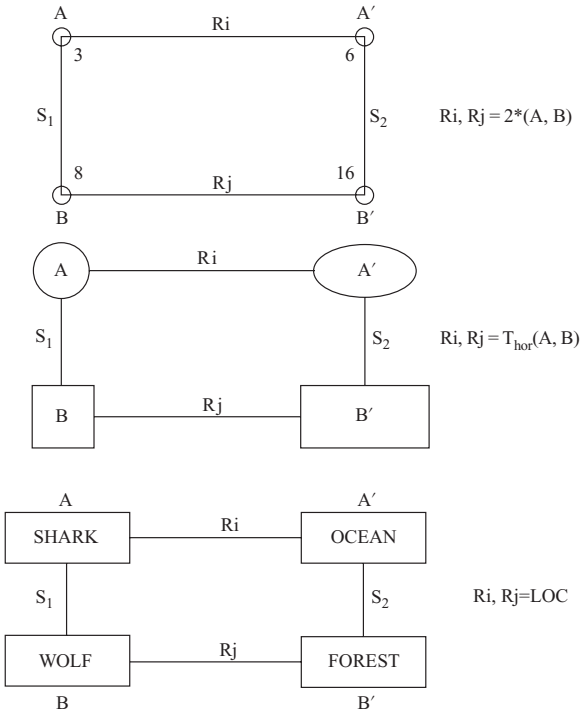

Figure 1
The basic structure of analogical reasoning. A–A′: the source area with relation (s) R, between them. B–B′: the target area with similar relations R_1. S_1 and S_2: mappings between source and target area

drawn, since not all of the conditions influencing or determining the outcome of a given situation are known. Analogical reasoning has a different character; it is basically a more precise form of conjecture. It is based on the mapping of structures or relations valid in one domain of knowledge onto another—less familiar or unknown—domain. The structures to be mapped are selected on the grounds of the similarity between the novel problem and the familiar one. The novel area is defined as the target domain of analogy construction, the familiar area as the source domain. With respect to the truth values of the conclusion, analogical reasoning spans all variants—from the absolute certainty of the *modus ponens* to *a posteriori* probabilities according to Bayes' Theorem, some of which are very weak. Suppose, for example, a second pair of terms is sought for a pair of numbers such as 3:6. Pairs such as 8:16 and 64:128 among many other combinations meet the requirements of the analogy.

As shown in Fig. 1, analogy construction can also be applied to geometric figures. It may be, but does not have to be the same factor of horizontal stretch which is mapped to produce a rectangle from a square or an oval from a circle. The boundaries within which geometric analogies are accepted are fuzzy, but when these boundaries are overstepped, the analogies are rejected. The same holds for meaningful words in natural languages. Let us imagine that analogous pairs with the same relational correspondence are to be generated for the following pairs of words: (a) collie: dog, (b) walk: run, (c) oak: birch. A good number of satisfactory analogies can be found, for example: (a) rose: flower, (b) speak: shout, (c) carp: trout. In each pair, the first term has all the relevant properties of the second (animal, voice, plant), as well as some of its own. The first pair is characterized by a sub-super-concept relation, the second by the intensity of the characteristic feature (the semantic relation of the comparative). The third pair shares a super-concept with common properties, but each term has specific properties of its own (the semantic relation of coordination). Pairs of terms such as dress and bullet or flea and fog do not allow analogies to be drawn on the basis of common properties, as these are non-existent. Clearly, similarity between the terms is an integral element in the construction of analogies. This similarity is not necessarily rooted in common properties, however. For us pairs of terms such as (a) ship and ocean, (b) hunter and deer, and (c) surgeon and scalpel, for example, do not share common properties. Nevertheless, the relational connections between the concepts are comparable with those found in other pairs of terms. In example (a) the shared relational connection is the location (as in flower and garden), in (b) it is the object of the action (as in butcher and cow), and in (c) the instrument (as in farmer and scythe). More complex relations can also be involved, as in 'learn: know' paired with 'fight: win.' Here, the common semantic relation is the objective or purpose. Such relational connections are the basis for analogy construction: 'The shark is the wolf of the ocean,' for example, or 'May is the Mozart of the calendar' (Erich Kästner). Many such examples are to be found in the Bible, particularly in the parables of the New Testament.

7. *The Origins and Original Function of Analogies*

Analogies are evidently not isolated examples of mental games, but a constant cognitive law. But what is the point of these transfers between bodies of knowledge? Is it just a case of juggling with combinations of words that match in some way? Or do analogies reflect a deeper (perhaps even social) need of *Homo sapiens*? Viewed from this evolutionary perspective, the question arises of where, exactly, the origins of analogical thought are to be found.

Analogical thought is primarily to be found in language use. Indeed, the key word 'language' puts us on the right track, for the evolutionary origins of

analogy are rooted in the need to communicate, and similes and metaphors are its direct evolutionary source. The cognitive function of analogy is to facilitate discourse; when an interiocutor does not understand a particular term or statement, the speaker attempts to clarify its meaning. This he or she does by referring to common knowledge that is as similar as possible to the term or statement in question. Similes and metaphors are essential ingredients of this procedure. The comparative phrase 'is like' is often added as an explicit reference to the existing similarity (Rumelhart 1979). For example, it is quite possible to imagine the following explanation being given in the course of prehistoric dialog:

Question: 'What's a quagmire?'

Answer: 'It's like a trap in the mud.'

The second term functions as a source of new knowledge. Indeed, the very purpose of this dialog is to impart (or acquire) new knowledge. This is also the case when adults explain things to children: 'The heart is like a motor driving the body,' 'pathologists are like butchers,' or 'rules are like corsets,' for example. Complete analogies can be viewed as extensions of such metaphors—for example, 'Rock in the opera—that's like heavy metal in cathedral,' an executioner in the Salvation Army is like a monk in the artillery,' or 'a June meadow in the mountains is like a bride in her wedding dress.' The list is never-ending. Such examples are no longer explanations or supplements to existing knowledge, but this is the source from which they originated. And, as illustrated by Eichendorff reference, they display originality and creativity of thought: 'It was as if the Sky had silently kissed the Earth, so that she, in the glimmer of blossoms, now had to dream of him.' A somewhat more mundane example can be drawn from one of General Moltke's cadet examinations. When Moltke asked a cadet, 'How wide is the Thames in Paris?,' the boy replied, 'Just as wide as the Seine in London, your Excellency.' By inverting the terms, the cadet demonstrated that he was well aware of the existing relations. Analogies can thus also be a source of smiles and approval.

8. *Cognitive Procedures in Analogy Construction*

Without a doubt, the recognition of structural similarities between the source and target domains plays a pivotal role in the acceptance of proposed analogies, as well as in their completion (e.g., supplying a fourth term when three are given), and creation (suggesting a new and possibly unique second pair of terms). Similarity is a necessary precondition in all three cases.

Different approaches to the definition of conceptual similarity have emerged in the literature definition in terms of (a) shared conceptual properties, as proposed primarily by Tversky Lit. and (b) structural similarity, as suggested by Gentner and Stevens (1983). A compromise between these two approaches has been formulated by Bachmann (1989) after Jaccard and Tanimoto.

Tversky defines the similarity between two objects in terms of the relationship between the features common to both structures and the features distinct to each, be they perceptual structures or sets of conceptual properties. As in Tversky's experiment, both components were then aggregated and compared to the estimated values. This resulted in a significant improvement on the correspondence between feature overlap and estimated similarity. The degree of improvement varied from concept, but correspondence calculated on the basis of common features alone was significantly lower.

The weighting on each of the semantic relations in similarity judgments is another matter of interest, and can be inferred from the sequence in which the individual relations are selected and entered into the analogous structure. Participants usually chose the actor and recipient first (e.g., hunt: hunter $\rightarrow$ game), followed by the instrument (e.g., rifle), the location (e.g., forest) and, often at a comparatively late stage, the objective (e.g., to bag a deer). This sequence of concept selection points to the existence of saliency weightings in the components of the individual concepts. Moreover, these results confirm that a moderate degree of similarity is essential not only for felicitous metaphors, but for creative analogies. Our experiments have shown that particularly characteristic features or relations can have preferential status in both the recognition and construction of analogies (Klix and Bachmann 1992). The originality of an analogy is dependent on factors such as the age and intelligence of its constructor, and analogy completion tasks are often employed in intelligence tests such as the Raven 'progressive matrices,' Sternberg (1977) has linked the factor structure of human intelligence to the solution of analogy tests. Moreover, Gentner and Sternberg have constructed algorithms which realize analogy detection. All of these are based on the comparison of features or semantic relations (Greeno and Riley 1984).

9. *Models of Analogy Detection*

Following the initial, very simple approaches to the processes of analogy detection, the dominant paradigm is now based on a model devised by Rips et al. (1973) which has been modified and extended by Bachmann (1998). According to this model, the following three aspects must be checked before an analogy can be accepted:

(a) Do the terms to be compared share a common feature which is much more salient than others? In train: railway and bus: truck; for example, the means of transport is so prominent that the analogy is immediately accepted.

(b) If this is not the case, the global impression is assessed. This entails a kind of compression of the available information, as discussed above in connection with Gluck and Bower's experiments.

(c) If this global comparison does not result in a decision, the concepts are closely scanned for common features and relations (deliberate checking). If no acceptable similarities are found, the analogy is rejected.

Unfortunately, the question of how much processing time is required for each of these components is as yet unanswered.

10. The Role of Analogy Construction in Scientific Innovation

The relationship between analogical reasoning and intelligence has been investigated in countless experiments and tests, with children and young adults of various ages, using plant, animal, object, and abstract analogies. However, there has been far less investigation of the ways in which analogy construction has impacted on the history of science, and the momentous mental achievements and creative insights which have resulted from this (see *Scientific Reasoning and Discovery, Cognitive Psychology of*).

Many significant examples can be found to illustrate how analogical reasoning has helped to break new ground in the history of science. The origins of analogy-based scientific innovation can be traced back to the early stages of Greek thought, with Archimedes one of its most eminent representatives. The Egyptian shadoof—a long-armed water-raising device used to draw water from wells—was invented long before Archimedes' time, and it was known that levers could be used to amplify physical force. It was left to Archimedes, however, to discover the relationship between the two arms, and thus to establish the principle of the lever. Hence mechanics—an essential precursor of physics—was founded on the basis of a number-pair analogy between the respective distances from the fulcrum and force relations according to the principle of levers. This development was founded on a paradigm of analogical thought, i.e., load arm: effort arm = length A: length B. Using the same method, i.e., in analogy to the principle of the lever, Archimedes was also able to calculate the surface area of any segment of a parabola (Serres 1994).

Eratosthenes is thought to be the first person to have calculated the earth's circumference. People had known how to determine the circumference of a circle using 360° since Babylonian times (influenced by the Sumerian sexagesimal system). Eratosthenes applied this knowledge in the following analogy: The distance between Syene (now Aswan) and Alexandria was approximately 800 km (in today's system of measurement). Once a year, the sun was at its zenith directly over a well in Syene and hence there are no shadows at that location at that time. At the same time in Alexandria, it cast a shadow from which it could be calculated that the sun was 7° from the zenith. This represents a difference of about 1/50 of a complete circle, hence the distance between Alexandria and Syene is about 1/50 of the earth's circumference. By multiplying the distance (800 km) by 50, Eratosthenes calculated the earth's circumference to be approximately 40,000 km.

In his famed *Dialogue on the Two Chief World Systems* (the Ptolemaic system and the Copernican system), Galileo contradicted the Aristotelians of the time. The Aristotelians advocated the Ptolemaic hypothesis, and presented the following argument against the Copernican model: Assuming that Copernicus is right in saying that the earth rotates around its own axis, the circumference of the earth is such that this must happen at high speed—40,000 km in 24 hours (or 920 m in some 2 seconds' free fall). As a test of the Copernican system, one can climb a high tower, drop a stone from the top, and note the position at which it hits the ground. Since the earth is turning while the stone is falling, the stone should be noticeably removed from the foot of the tower when it makes contact with the earth. This is not the case, however. The stone falls perpendicular to the tower and hits the ground at its foot. According to the Aristotelians, this disproved the Copernican system. Using the character of Salviati in the *Dialogue* as his mouthpiece, Galileo refuted this argument with the following analogy: Imagine a sailing boat traveling at high speed. If a stone is dropped from the top of the mast, it will hit the deck at the foot of the mast, since it retains the speed of the boat until it hits the ground. Galileo thus invalidated the anti-Copernican argument by means of an analogy.

The most impressive known example of analogical reasoning in scientific discovery is provided by the equations describing electromagnetism derived by James Clerk Maxwell in 1864. Maxwell's contemporary, Michael Faraday, had demonstrated the existence of electric and magnetic fields, showing that iron filings sprinkled on a thin piece of paper held above a magnet are attracted to its 'lines or force' or its poles. Maxwell investigated the properties of the lines of force described by Faraday and discovered similarities between the behavior of these lines of force and the flow of a liquid in a pipe. The equations for the latter were already known. Maxwell transferred this knowledge to magnetic fields and electric currents, and was thus able to define the behavior of magnetic lines of force at and between poles. Writing on this means of discovery, Maxwell postulated that in order to form an idea of physical laws for which no theory is available, it is necessary to familiarize oneself with the existence of physical analogies. The discovery of the interaction between electric currents and magnetic fields laid the foundations of the theory and practice of electromagnetism, and hence of the dynamo tech-

nology at the close of the nineteenth century. Moreover, it was the precursor of radio and television wave transmission, since one of Maxwell's analogous equations had suggested that radio waves must also exist. Twenty years after Maxwell presented his basic equations, Heinrich Hertz indeed succeeded in producing radio waves under laboratory conditions—another demonstration of how creative thought in the form of analogical reasoning can impact dramatically on the history of science and the fate of humankind.

See also: Analytic Induction; Informal Reasoning, Psychology of; Practical Reasoning: Philosophical Aspects; Problem Solving and Reasoning, Psychology of; Problem Solving (Everyday), Psychology of; Problem Solving (Joint), Psychology of; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Bachmann T 1998 Die Ähnlichkeit zwischen analogen Wissensdomänen [The similarity between analogous domains of knowledge]. Doctoral dissertation, Humboldt University, Berlin

Becker O 1990 *Grundlagen der Mathematik in geschichtlicher Entwicklung* [Foundation of mathematics in historical development]. STW, Freiburg, Germany

Beyer R 1991 Untersuchungen zum Verstehen und zum Gestalten von Texten [Investigations of the comprehension and organisation of texts]. In: Klix F, Roth E, van der Meer E (eds.) *Kognitive Prozesse und geistige Leistung*. Deutscher Verlag der Wissenschaften, Berlin

Bolzano B 1970 *Wissenschaftslehre* [Scientific model]. Akademie Verlag, Berlin

Gentner D, Stevens A L 1983 *Mental Models*. Lawrence Erlbaum, Hillsdale, NJ

Gluck M A, Bower G H 1988 From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General* **117**: 227–47

Greeno J G, Riley M S 1984 Prozesse des Verstehens und ihre Entwicklung [Processes of comprehension and their development]. In: Weinter F E, Kluwe R (eds.) *Metakognition, Motivation und Lernen*. Kohlhammer, Stuttgart, Germany, pp. 252–74

Hoffmann J 1993 *Vorhersage und Erkenntnis* [*Prediction and Knowledge*]. Hogrefe, Göttingen, Germany

Horgan G 1995 Die neuen Sozialdarwinisten [The new Social Darwinists]. *Spektrum der Wissenschaft* **12**: 80–90

Kahnemann D, Tversky J A 1973 On the psychology of prediction. *Psychological Review* **80**: 207–51

Klix F, Bachmann T 1992 Analogy detection—analogy construction: An approach to similarity in higher order reasoning. *Zeitschrift für Psychologie* **206**: 125–43

Leibniz G W 1960 *Fragmente der Logik* [*Fragments of Logic*]. Akademie Verlag, Berlin

Rips L J, Shoben E J, Smith E E 1973 Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior* **12**: 1–20

Rumelhart D E 1979 Some problems with the notion of literal meanings. In: Ortony A (ed.) *Metaphor and Thought*. Cambridge University Press, Cambridge, UK

Schmidt H-D 1966 *Leistungschance, Erfolgserwartung und Entscheidung* [*Chance, Expectation and Decision*]. Deutscher Verlag der Wissenschaften, Berlin

Serres M 1994 *Elemente einer Geschichte der Wissenschaften* [*Elements of a History of Science*]. Suhrkamp, Frankfurt, Germany

van der Meer E 1991 Die dynamische Struktur von Ereigniswissen [The dynamic structure of event-related knowledge]. In: Klix F, Roth E, van der Meer E (eds.) *Kognitive Prozesse und geistige Leistung*. Deutscher Verlag der Wissenschaften, Berlin

Wason P C, Johnson-Laird P N 1972 *Psychology and Reasoning*. Harvard University Press, Cambridge, MA

F. Klix

Problem Solving (Everyday), Psychology of

This article reviews developments in the psychology of everyday problem solving. The adjective 'everyday' is interpreted by contrast with the puzzles and academic tasks that have been the main focus of psychological problem solving research. This leads to an emphasis on the relation between planning and action, and between the mental representation of the problem and the physical situation in which problem solving takes place. The role of artifacts in support of everyday problem solving is discussed, and the idea that problem solving can be understood as an adaptation to environmental constraints is reviewed.

1. Background

The psychology of human problem solving was transformed by the work of Newell and Simon (1972) and all subsequent theorizing has used their work as a platform. Newell and Simon (1972) developed a very general abstract conception of problem solving as search through a problem space. A problem is defined by a goal state, a start state, and a set of operators, and the problem solver must achieve the goal state by applying a sequence of operators so as to transform the start state into the goal state. The set of states that are generatively defined by the start state and the operators is called the 'problem space,' and problem solving can be viewed as a search of this problem space for the goal state.

Newell and Simon's early empirical work within this framework concentrated on discovering the heuristics that people use to direct search through the problem space. This led them and their followers initially to

concentrate on well-defined puzzles, with a predetermined set of operators and a completely specified problem space. This program of research was very effective, leading at least one prominent reviewer (VanLehn 1989) to claim that the psychology of human problem solving had been 'solved' (leaving as a residual mystery the question of what is learned from problem-solving experience).

But however valuable puzzles like the Tower of Hanoi or letter-arithmetic might be for the study of human problem solving, they are clearly not 'everyday' problems. There is no doubt that the problem space theory and search heuristics such as means–ends analysis and hill-climbing are vital to a full understanding of everyday human problem solving. But equally, the salient and manifold differences between many everyday problems and well-defined, self-contained puzzles raise important issues for the problem solver and for the psychologist.

Consider the following examples of everyday problems (from Anderson 1990, p. 193, given by his students as answers to the question 'name a problem that you tried to solve in the last 24 hours').

Finding a route from a hotel to a restaurant
Debugging someone else's code
Getting subjects to sign up for my experiment
Getting two Diet Cokes from a Coke machine
Getting my car registered

Perhaps the most obvious special property of puzzles like Tower of Hanoi compared to these problems is that all the knowledge needed for a solution is provided in the problem statement. (This is not quite true of letter-arithmetic, of course, but even then the additional knowledge required, that is, arithmetic, is well-defined and strictly demarcated.) In particular the set of applicable operators is constrained and predetermined.

An extensive literature has extended the study of human problem solving to 'knowledge-intensive' domains, in which the application of the solver's experience (in particular by choosing operators from long-term memory) is the critical issue. The most extensively studied domains have been academic domains like mathematics, physics, and computer programming (as well as chess, which is self-contained in principle, but reliant on long-term memory in practice), and the most extensively used methodology has been a comparison of expert and novice problem solving. The classic established finding of this research is that experts are better problem solvers than novices not because they are better able to use search heuristics, but rather because their knowledge of the domain allows them to identify and classify problems in terms of the methods that will be useful to solve them, thereby obviating the need for extensive search of the problem space. Experts tend to work forwards from problem statements to solutions, whereas novices tend to work backwards from the goal, searching memory for applicable knowledge (Larkin et al. 1980).

Most everyday problems are even less well-defined than academic problems in terms of the applicable operators. If you want to register your car and you don't know how, the problem is discovering the applicable operators, not merely ordering them. Furthermore, in contrast to puzzles and to academic problems, most everyday problem solving involves interacting with (so as to change) the everyday physical and social environment.

2. *The Problem of Everyday Problems*

Perhaps the single most important issue that the study of everyday problem solving prioritizes is the nature of the interplay between mental activity and physical activity, or between the mental problem space and the physical setting in which problem solving takes place. Within the traditional domains of problem solving research noted above, this has seldom been treated as a serious issue (although it should be noted that Newell and Simon recognized its importance, and have contributed some important analyses of the relation between external and internal representations, e.g., Simon 1975). Search of a problem space is primarily a mental activity: the current state is mentally represented, operators are considered, and imagined next-states are evaluated. Perhaps hierarchies and sequences of operators are mentally considered ('planning' or 'look-ahead'). Eventually an operator or sequence of operators is chosen, and then enacted. Many computational models of human problem solving gloss any distinction between mental and physical application of an operator. And early AI planning systems, building on Newell and Simon's analyses of search, computed entire sequences of operators from start state to goal state before acting on the world. People are seldom so patient. Instead they iteratively interleave planning and action. Since the early 1980s a body of literature has shown that in many everyday problem-solving situations the specifics of the physical situation and the problem solver's management of the relation between planning, perception, and action are crucial determinants of problem-solving success.

Consider one observation from some very influential empirical studies by the cognitive anthropologist Jean Lave and her colleagues (1984). A supermarket shopper picked up one slab of cheese from a shelf of similar cheeses to be surprised by its price. She guessed that it had been wrongly labeled but wanted to check. She might have used mental division to compute its price per unit weight and then performed the same computation on one of the other slabs (as usual the cheeses varied somewhat in weight and therefore price). Instead she searched the shelf for a cheese of very similar weight and directly compared the prices of the two cheeses.

A similar observation of in-situ problem solving was made by Scribner (1984). She observed dairy workers completing orders for partially full crates of

milk bottles. For example, a packer might receive an order of 1-6, meaning one crate-six bottles. Instead of completing this order by taking a full crate (16 bottles) and removing six bottles, the packer might find a crate has eight bottles in it and add two. This 'nonliteral' strategy involves mental effort, using arithmetic to recode the original order, but saves physical effort. (Scribner found that nonliteral strategies were common among skilled packers, and could almost always be accounted for by the saving of physical effort.)

Both the packer and the shopper are evidently managing a tradeoff between physical and perceptual activities (such as searching a display for relevant items) and mental activities (arithmetic operations). To fully understand their problem-solving behavior it is necessary to understand the opportunities for this tradeoff that are provided by the situation, and the cognitive mechanisms by which the tradeoff is managed.

3. *Cognitive Artifacts*

The first issue becomes particularly salient when coupled with the observation that much everyday problem solving is supported by 'cognitive artifacts' or tools of thought, ranging from pencil-and-paper through conventional representations like diagrams and graphs to computational systems. How can we understand the role of such artifacts in problem solving?

One theoretical approach is to simply redraw the boundaries of the cognitive agent away from the human problem solver to include the artifacts being used (Hutchins 1995). This approach has led to some insightful descriptions of how various human–machine systems perform their specific tasks (e.g., how a cockpit remembers its speeds). In the work of Zhang and Norman (1994) it has been developed into a general framework for the understanding of distributed cognitive tasks, in which the relation between external and internal representations is the primary focus. Zhang and Norman (1994) argue that when aspects of the representation of a problem can be offloaded into the physical environment, problem solving becomes easier.

A second approach focuses on how the artifact changes the cognitive task of its user. For example, Larkin and Simon (1987) have analyzed how diagrams facilitate certain problem-solving strategies by allowing simple perceptual judgments (such as whether two lines intersect) to replace more demanding mental computations.

Diagrams might be referred to as 'read only' cognitive artifacts. Artifacts that are 'read-write' can have more complex mediating effects on cognition. Consider the use of a simple appointment diary (Payne 1993). Recording an appointment in a diary changes the task of remembering that appointment in several ways. First, the fact that the appointment has been named and written will affect the internal memory trace for the appointment. And of course, the diary can act as an external memory, allowing its user to retrieve the appointment by reading. Furthermore, in conventional paper diaries the act of writing a new appointment leads to the inevitable rehearsal of already noted appointments.

A third approach to understanding cognitive artifacts is to focus on the properties of the artifact itself as an information system, and to ask how readily it allows certain information-processing operations. This approach has been pioneered in the field of human–computer interaction by Green (1989) under the banner 'cognitive dimensions of notations.' Green's enterprise is to define a limited number of widely applicable dimensions on which cognitive artifacts can be compared in terms of the extent to which they support flexible patterns of use. (Green has been particularly concerned with artifacts that are used in design tasks, such as spreadsheets, programming languages, document processors, and musical notations, and with supporting opportunistic design, in which initial drafts are iteratively revised.) One cognitive dimension is 'role expressiveness,' the extent to which a notation exposes meaningful structure. As a simple example, the use of meaningful variable names in a program will tend to increase role expressiveness. Another dimension is 'viscosity,' the extent to which a notation is resistant to local change. Consider a text with section numbers. Insert a section and all subsequent numbers must be changed.

All these approaches are set to continue in various guises. Understanding the cognitive effects of artifacts is one of the most important open problems for the theory of human problem solving.

4. *Rational Approaches to Problem Solving*

Let us now turn to the second crucial issue raised above, that is, how the tradeoff between mental and physical activity is managed. Scribner (1984) in her analysis of milk crate packing, suggested that packers might be explicitly trading mental and physical costs and benefits. More recently this basic idea has been developed, particularly in the work of John Anderson (1990) on what he has called the rational analysis of cognition. The general idea is to analyze cognition as an optimal adaptation to the probabilistic structure of the environment. Such an analysis asks *why* cognitive behavior is the way it is, and may help structure the subsequent investigation of *how* cognition approximates the optimal solution.

This approach has already thrown some light on the critical characteristics of everyday problem solving.

First, consider the interplay between planning and acting. From a rational perspective, a problem solver should plan before acting to the extent that the estimated benefits of planning outweigh the costs. The benefits of planning will relate to a reduction in the number of actions required and the individual costs of those actions. We might suppose that Scribner's dairy workers would have been less inclined to transform their orders arithmetically if moving bottles around had been somehow much easier. This idea has been tested in a series of experiments by O'Hara and Payne (1998) who showed that clumsy user interfaces to puzzles (i.e., interfaces that made each move more costly in terms of keystrokes or time) produced fewer-move solutions to puzzles than did more efficient interfaces.

The use of cognitive artifacts has also been illuminated by an adaptive analysis. A simple case is the electronic calculator. There are many problems that can be computed either mentally or by calculator. An experiment by Siegler and Lemaire (1997) showed that people are able to make this choice adaptively according to an accurate model of their own capabilities: that is, they reliably choose to do a task mentally when that is indeed the quickest way to do it, and to use the calculator instead when that has real benefits.

Finally, Pirolli and Card (1999) have shown that aspects of people's use of complex computer-based information retrieval systems can be well modeled by equations based on optimal foraging theory. In other words, people behave as if they could compute and use accurate estimates of their current and previous rates of information gain (cf. an animal's intake of energy from a food supply) and of the costs of searching different information resources (cf. an animal's journey to a new food supply).

5. *Conclusions*

Some theorists have argued that the situational contingencies of everyday problem solving render the abstract cognitive theory of Newell and Simon useless. However, an alternative view is that instead these complexities simply emphasize certain theoretical questions over others, and the required reworking is best considered as an elaboration of aspects of Newell and Simon's framework. This second approach has the advantage of allowing incremental development in our understanding of the psychology of human problem solving.

See also: External Memory, Psychology of; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning

Bibliography

Anderson J R 1990 *The Adaptive Character of Thought*. Erlbaum, Hillsdale, NJ

Green T R G 1989 Cognitive dimensions of notations. In: Sutcliffe A, Macauley L (eds.) *People and Computers V*. Cambridge University Press, Cambridge, UK, pp. 443–60

Hutchins E 1995 *Cognition in the Wild*. MIT Press, Cambridge, MA

Larkin J, McDermott J, Simon D P, Simon H A 1980 Expert and novice performance in solving physics problems. *Science* **208**: 1335–42

Larkin J H, Simon H A 1987 Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science* **11**: 65–100

Lave J, Murtaugh M, de la Rocha O 1984 The dialectic of arithmetic in grocery shopping. In: Rogoff B, Lave J (eds.) *Everyday Cognition: Its Development in Social Context*. Harvard University Press, Cambridge, MA, pp. 67–94

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Payne S J 1993 Understanding calendar use. *Human–Computer Interaction* **8**: 83–100

O'Hara K P, Payne S J 1998 The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology* **35**: 34–70

Pirolli P L, Card S K 1999 Information foraging. *Psychological Review* **106**: 643–75

Scribner S 1984 Studying working intelligence. In: Rogoff B, Lave J (eds.) *Everyday Cognition: Its Development in Social Context*. Harvard University Press, Cambridge, MA, pp. 9–40

Siegler R S, Lemaire P 1997 Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General* **126**: 71–92

Simon H A 1975 The functional equivalence of problem solving skills. *Cognitive Psychology* **7**: 268–88

VanLehn K 1989 Problem solving and cognitive skill acquisition. In: Posner M I (ed.) *Foundations of Cognitive Science*. MIT Press, Cambridge, MA

Zhang J J, Norman D A 1994 Representations in distributed cognitive tasks. *Cognitive Science* **18**: 87–122

S. J. Payne

Problem Solving (Joint), Psychology of

'Joint problem solving' refers to the situation where two or more people work together to solve a shared problem, where agreement on a single solution may or may not be required. The proverb says 'two heads are better than one,' but why so, or when so, is not explicitly understood yet. Despite this folk knowledge, studies have shown that joint solving does not always result in better performance than that of the most capable solos. Rather, careful studies of the processes have revealed that there are clear cognitive process gains in the joint endeavor. In joint situations, the

participants tended to deepen their conceptual understandings of the topic problem (Miyake 1986), or the subjects generated a greater number of experimental hypotheses and had better chances to duplicate a scientific discovery (Okada and Simon 1997). The reasons for such gains are still under investigation, but research findings suggest that the joint situation allows the participants to naturally externalize their solving processes, and this externalization provides chances for flexible revisions of their trials, leading them to more conceptual understandings and newer ideas.

1. *Gains of Joint Processes*

Joint problem solving has been extensively studied in social psychology, mainly with an interest in whether people in groups, thus in 'social' situations, would perform better than solos. Hastie (1986) gives an extensive review, where joint cases are compared against solos on tasks like number estimation, vocabulary tests, and arithmetic brainteasers. Hastie summarizes the field and concludes, 'these studies tended to find that group performance exceeded the performance of the average individual, [but] for most of these problems groups did not match the performance of their best members' (p. 150). He also suggests that 'the critical characteristic of performance on the tasks is solution demonstrability, with higher levels of group performance associated with tasks where correct solutions are easily demonstrated and communicated to group members, and relatively low levels of performance associated with difficult-to-demonstrate solutions' (p. 150). His interest was mostly in categorizing group performances on tasks and group membership, not on cognitive aspects of the solving processes.

Miyake (1986) tried to clarify the cognitive gains associated with paired performance. She asked three pairs of subjects to understand how a sewing machine made its stitches and analyzed their verbal protocols. The sewing machine problem has interesting characteristics of having different levels of answers. One answer to the question would just mention that the machine uses two intertwined pieces of thread to fasten the cloth. On hearing this, one can further ask about the exact physical configuration of the seam, which turns out to be a loop-in-loop. Discovering this could open up a new puzzle because it is physically impossible to form a loop-in-loop with two threads without free ends. (At a middle point in sewing, for both upper and lower threads, one end of them is in the spindle or in the bobbin, while the other end is tightly sewn to the cloth.) In fact, while the upper thread comes down into the machine, its loop goes entirely around the bobbin. This is a perfect answer until one starts wondering how, then, the bobbin itself is attached to the machine.

This hierarchical nature of the possible answers to the sewing machine problem made it possible to analyze the protocols in terms of when they reached the intermediate understandings and how and when they noticed the insufficiency of their solutions and started to ask the 'one level deeper' questions. Miyake found that each of her subjects descended this hierarchy according to his or her individualistic formulation of the question. This often resulted in situations where members of the pair worked at different levels at the same time: while one person may worry about the physical configuration of the seam, the other could try to find how the loop-in-loop configuration is possible with two threads without free ends. Because of this disagreement, the subjects were almost always in conflict, which contributed to push each individual to refine his or her own models so that 'the other member would understand my points.' In summary, Miyake argues that the process gain of joint problem solving does not necessarily lie in the performance, but rather in raising the necessity and motivation for deepening and refining each individual's understandings.

In the 1990s, the amount of research on joint problem solving increased, showing more positive results for both pairs and groups. Conditions for successful collaborations were identified where earlier studies focused on degrees of mutual participation, while the more recent studies identified explicit types of talk that contributed to the success (e.g., Teasley's 'transactive' talks, Teasley 1997). A similar style of research has been extended to cover topics of learning and development (e.g., Rogoff 1990). Hatano and Inagaki (1991), for example, analyzed a science classroom where the children selected among the alternative hypotheses provided by the teacher and debated among them. By analyzing the discussion they identified what they named 'the partisan motivation,' shared by the naturally grouped members with the same alternative, that acted as an enhancer for the self-motivated articulation of their explanations.

2. *Joint Problem Solving in the Real World*

Okada and Simon (1997) took a slightly different course of identifying causes for pairs' success in an experimental scientific 'discovery.' Using a computer-simulated task of making a discovery in molecular genetics, paired subjects were found to be superior to single subjects. This success was associated with higher levels of explanatory activities of the pairs, and further, such explanatory activities were facilitated when paired subjects made explicit requests of each other for explanations and testable hypotheses.

This change occurred with the search into more ecologically valid research in the domain of problem solving. Daily problems often get solved in groups,

and there are pragmatic needs and interests in finding out how groups should be organized to be productive. This led cognitive problem solving research to take an 'in vivo' form (Dunber 1995), where workers and scientists in realistic situations were carefully studied. Dunber observed four laboratories carrying out immunization research and followed their daily scientific activities for a year, with various methods from attending the research meetings and taking notes, reading their lab diaries, to extensive interviews. His findings are diverse, and factors he found positive include frequent exchange of ideas, nonauthoritative membership, positive attitudes toward even negative intermediate results, use of local analogies, etc. In an interview-based study of a real R & D team in a detergent company, Ueda and Niwa (1996) found that having an outsider's view at the right time was crucial for the team to change the course, which led them to produce a new successful product. All such reported factors tend to confirm the folk expectation of what should contribute to high group performance. Yet how we can coordinate these factors to make them all work simply remains to be discovered. Furthermore, the overall conclusion we could draw from these sets of research is that the quality of group performances depends heavily on the organization and activity control of the groups.

3. *Joint Problem Solving from a Cognitive Anthropology Perpective*

In contrast, cognitive anthropologists have maintained more positive perspectives. Hutchins (1990) observed a team navigating a large ship and found that there was overlapping knowledge shared among the team members that strongly coordinated the teamwork. In the reported case of a 'fix,' or a joint activity of finding the exact location of a ship, at least three roles are involved. The bearing-taker measures the angles between the ship's direction and selected landmarks using a gyroscope, and calls out the number through a tube to the charting house. Then this number is repeated by the bearing timer-recorder who writes down the number into the log book. The number then gets converted to the angle on the device that allows the plotter to make lines on the chart so that they can pinpoint the location of the ship onto the map. Staff start working in this team as the bearing-taker, who after some experience is promoted to the timer-recorder job, who would later move into the plotter position. Thus the plotter knows everything the other two know, and the timer-recorder knows all of what the bearing-taker knows. It is also possible for the lower-rank staffs to closely observe what the staff one rank higher is engaged in. While a cross-sectional, one-shot observation might make this teamwork appear to be highly coordinated across divisions of work, in reality what coordinates this team is the shared knowledge of a longitudinal, historical nature. Laboratory-oriented, puzzle-based experimental studies of joint problem solving have been quite weak on this aspect.

Hutchins' report maintains one more important observation on how the teamwork is divided and carried out. In the above example of taking a fix, there are three subtasks, each of them producing some intermediate results. Careful inspection reveals that each of them takes quite a different, and redundant form of representation of the intermediate results of a single process. The first outcome of the bearer is by voice, which is then written down in number symbols, which is then again converted into lines on the chart. These different forms, at different times of representation, provide different and redundant forms for people involved to check the process from different perspectives. In this sense this is a real-world documentation of what Hastie called 'demonstrability,' which, in realistic situations, does not just provides the information concerning the correctness of the solution but opens up the possibility for other interpretations and possibly flexible redesigning of the process itself.

4. *Toward Further Research into the Mechanisms of Joint Gain*

Paired subjects could, in fact, take advantage of externally shared objects in the course of simpler experimental tasks. Shirouzu et al. (1999) asked subjects to indicate two-thirds of three-quarters of the area of a square, folding paper measuring 15 cm × 15 cm. They report that in various conditions people strongly opted (over 90 percent on average) to use external resources to do this arithmetic, either by folding the paper or by marking the edges with their fingers and partitioning, etc. Because this reliance on external resources is quite strong, most of the single subjects did not change the solution method (from folding to calculating, for example) even when they were consecutively asked the second question, i.e., to indicate the area of three-quarters of two-thirds of the same sized paper. The subjects did not change it even after seeing the 'externalized' result to be exactly one-half of the paper. They compared this to pair performances and reported that the pairs are a lot more flexible in choosing different solutions to the second question (70 percent of the pairs changed the solution methods while less than 20 percent of the single subjects did so). By analyzing the solving processes and the protocols of such subjects they found that the flexible pairs were in fact gradually responding to the constantly changing externalized forms of the paper while they worked together, rather than each randomly contributing newer ideas at their turns.

As a future course of joint problem-solving research, new learning studies focus on collaborations. As indicated above, joint situations tend to provide the

participants with verbal expressions and tangible externalized objects that each member can manipulate to concretely suggest newer, or at least different, ideas. In rapidly changing societies, the goal of learning is not only digesting what is taught. Learners are required to acquire skills to learn, to become self-regulated learners. In order to fulfill this goal, it is necessary to give learners more chances to be reflective than currently enforced, to scrutinize upon and revise what they themselves do and how others do similar tasks. Studies on joint problem-solving situations are expected to provide rich resources to help develop effective learning situations (e.g., Bransford et al. 2000).

See also: Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Processes in Organizations; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Joint Action; Problem Solving and Reasoning, Psychology of

Bibliography

Bransford J D, Brown A L, Cocking R R (eds.) 2000 *How People Learn*, expanded version. National Academy Press, Washington, DC

Dunber K 1995 How scientists really reason: Scientific reasoning in real world laboratories. In: Sternberg R J, Davidson J E (eds.) *The Nature of Insight*. Cambridge University Press, New York

Hastie R 1986 Review essay: Experimental evidence on group accuracy. In: Grofman B, Owen G (eds.) *Decision Research, Vol. 1, Information Pooling and Group Decision Making: Proceedings of the Second Conference on Political Economy*. University of California, Irvine, CA, pp. 129–57

Hatano G, Inagaki K 1991 Sharing cognition through collective comprehension activity. In: Resnick L B, Levine J M, Teasley S D (eds.) *Perspectives on Socially Shared Cognition*. APA, Washington, DC

Hutchins E 1990 The technology of team navigation. In: Galegher J, Kraut R E, Egido C (eds.) *Intellectual Teamwork: Social and Technological Foundations of Cooperative Work*. Erlbaum, Hillsdale, NJ

Miyake N 1986 Constructive interaction and the iterative process of understanding. *Cognitive Science* **10**: 151–77

Okada T, Simon H A 1997 Collaborative discovery in a scientific domain. *Cognitive Science* **21**(2): 109–46

Rogoff B 1990 *Apprenticeship in Thinking: Cognitive Development in Social Context*. Oxford University Press, Oxford, UK

Shirouzu H, Miyake N Masukawa H 1999 Roles of cognitive externalization for joint problem solving. In: *Proceedings of the 2nd International Conference on Cognitive Science and the 16th Annual Meeting of the Japanese Cognitive Science Society Joint Conference*, pp. 337–42

Teasley S D 1997 Talking about reasoning: How important is the peer in peer collaboration. In: Resnick L B, Saljo R, Pontecorvo C, Burge B (eds.) *Discourse, Tools, and Reasoning: Essays on Situated Cognition*, NATO ASI Series: Series F: Computer and systems sciences. Springer, Vol. 160, pp. 361–84

N. Miyake

Procedure: Legal Aspects

1. The Subject Matter of Procedure

Procedure addresses the work of adjudicatory processes that are based in courts. While the subject may appear dry (if not boring), the rules of any procedural system are interesting because they must respond to difficult political and social questions.

Traditionally the world of procedure has been divided into three categories—civil, criminal, and administrative. For example, civil cases are those in which no criminal penalties (often but not exclusively incarceration) are sought. Administrative law proceedings are those in which (typically) the government is a party and the case is processed in administrative agencies devoted to a particular subject matter rather than in generalist courts. Courts in turn are defined as institutions that handle civil and criminal cases. However, these lines are blurring. For example, while civil disputes often involve private citizens, in a substantial proportion of civil cases the government is a party, as either plaintiff or defendant. Concerns about overcrowded court dockets have prompted some commentators to recommend that mandates to agencies be enlarged—turning more agencies in essence into specialized courts.

What is shared among civil, criminal, and administrative proceedings is that someone (an individual, a group, or a representative of the government) claims that another has violated a legal norm. In all three categories a third party, empowered by the state, is called upon to determine the validity of the claim—whether a legal right exists ('liability') and what form of response ('remedy') is appropriate. In all three categories, the parties have some degree of autonomy to fashion the contours of the dispute and yet are also limited by conventions and rules. In all three categories, questions exist about the power of decision-makers, the constraints imposed upon them, the kinds of sanctions or remedies that might be imposed, and the ability to seek reconsideration of the decisions rendered. And, in all three categories, the procedures are assessed as to whether they are fair and efficient.

What makes for fair and impartial judgments? Should everyone have a 'right' to an authoritative adjudication, or may courts insist that litigants settle or resolve cases by other means? If trials are permitted, should they be decided by judges (professional adjudicators) or by lay juries? If judges are central, how is their impartiality and independence assured? Does the gender, race, or ethnicity of a decision-maker matter to the perceived or actual fairness of the outcome? Should state subsidies, including free lawyers, be provided for those too poor to litigate to facilitate access to justice?

Of course, some of the questions for proceduralists are more technical, such as: What injuries are grounds for a lawsuit? How does one actually start a case? In

which court? Whom should one sue? Yet responses to even these questions depend on judgments about the role of government, the degree to which individuals should be permitted autonomy as rights seekers, and the harm that litigation itself can inflict on unwilling participants. Therefore, rules of procedure embody, express, and help create values and attitudes about what constitutes fair decision making by, or through, the powers of the state. Procedure is laced with questions of legal policy and political philosophy.

In some countries, those questions are also framed in constitutional terms. For example, the US Constitution deploys the phrase 'due process of law' in its Fifth and Fourteenth Amendments, guaranteeing protection against deprivations of life, liberty, or property. Also detailed are rights of jury trials for criminal defendants and for civil damage litigation, as well as rights of counsel for criminal defendants. Other countries expressly address these and yet other rights, such as providing criminal defendants with a right of translation into their own language.

The set of problems facing proceduralists is rich, not only because rules of procedure inevitably implicate basic social and political values. Procedure is also compelling because those rules intersect with all the major problems of contemporary life. Under the umbrella of 'cases' can come the full range of social problems—disputes over personal assets, and familial relationships; claims of physical injuries from products like asbestos or tobacco; disagreements about commercial contracts made or breached; regulatory arguments about possible hazards to the environment or the illegal pricing of products; assertions of violations of state or federal statutes providing for civil rights, consumer information, or personal security. In short, Procedure must address the theories, aspirations, and values that animate procedural systems; the kinds of disputes that may prompt litigation; the means by which lawsuits are started and stopped, as well as what happens in between and who can participate; consideration of the identity and characteristics of the participants (parties, lawyers, witnesses, judges and juries); the powers of decision makers; what the structure of courts is, has been, and how it might be changed; and what role courts in turn play in a particular governmental structure or as trans-national institutions.

2. *Lawsuits: From Injuries to Litigation*

When thinking about Procedure, it is important to remember that not all moments of discord result in the filing of lawsuits, let alone in the kinds of proceedings described in newspapers or the trials that are televised. In fact, individuals and groups disagree a lot without 'naming' the problem a conflict. In contrast to popular assertions, for example, that the US is awash with litigation, scholars of dispute resolution instruct otherwise. When injured, individuals have a variety of responses (from 'lumping it' and never seeking redress to 'claiming' injury and then 'blaming' others); a numerically small percentage of such occurrences results in the actual filing of civil, criminal, or administrative proceedings. How often people file suit, in turn, is related to what forms of redress regulatory and social welfare regimes provide, and to conceptions of what role courts should take in a given polity. The role litigation plays in countries therefore differs.

Moreover, litigation should not be equated with adjudication through trial. In some countries, of cases that are filed, very few conclude by trials. Take the federal courts of the US as an example. Of 100 civil cases filed in the year 2000, about a third will be resolved by judicial rulings in advance of trial; most of the remaining two-thirds will be concluded through settlement. Trials will begin in only three or four cases.

The rate of trial in the US declined during the last decades of the twentieth century, as courts built into their processes a 'pretrial phase' that becomes the focus of lawyering and that offers a means of dispute resolution without trial. Many countries have given litigants rights of access to information in advance of trial; 'discovery' and 'disclosure' are terms of art in the US to refer to such pretrial exchanges. Further, negotiation, mediation, and arbitration (sometimes called 'alternative dispute resolution' or 'ADR') are concurrent options—or requirements—throughout the course of disputes.

Some procedural systems are styled 'inquisitorial,' by which is meant that judges determine a good deal of the shape of the inquiry. Others, described as more 'adversarial' and associated with common law countries, authorize parties (or more accurately their lawyers) to determine the scope of a controversy, to develop the information about the events at issue, and to frame legal theories. During the twentieth century, many common law countries revamped their procedural systems—as did civil law regimes—such that the two forms of adjudication are moving closer to each other.

In the US, federal rule makers undertook major reorganization of the procedural system in the 1930s. They spoke about aspirations for a simple, accessible justice system and they crafted a 'trans-substantive' set of procedures, by which it is meant that the same rules were to be used, regardless of the kind of case. By the 1970s, some litigants, judges, and lawyers became concerned that the system created decades earlier had generated its own problems. They complained about lawyers exploiting pretrial opportunities for personal gain and to inflict misery on opponents. Critics charged abusive discovery, incivility, and adversarial excess. Some reformers urged lessening reliance on lawyers by reeducating clients to look for resolutions

of disputes without, or with less lawyering. Others proposed rules aimed at constraining lawyers.

Many common law countries turned to judges to take more responsibility for the shape of lawsuits. Judges and other court personnel began to take an active role in scheduling and monitoring the pretrial phases of cases—to move cases along, to urge settlement, to attempt to supervise attorney behavior. Judges came to act as 'managers,' working with lawyers to set time frames in which parties may be added or dropped from cases, to resolve disagreements about the relevance of information sought from opponents, and to urge that disputes be ended by mandatory settlement conferences, court-annexed arbitration, and mediation. Procedures once termed 'extrajudicial' have been built into rules, and are now regular features of civil and criminal processes.

The term of art—in the US, England, Wales, Canada, and Australia—for such work is 'case management' but the content of such work varies by country. In the US, a good deal of the focus is on managing lawyers. In this respect, rules of procedure are connected to rules of professional conduct. Rules of lawyering ethics and rules of procedural systems increasingly overlap as both are aimed at structuring the behavior of attorneys.

Late twentieth-century reforms seek to lower the costs of process and to shorten the time to decision. Some reformers believe that these changes will help lower-income individuals gain access to courts. Critics view the rules as less friendly to small claimants because of limitations on discretion and flexibility. These debates point to a more general feature of procedural rule making. Procedural reforms of any era need to be understood as not 'only' about procedure; they are related to the creation of professionalized judiciaries, to country-specific social movements, and to institutionalization of courts as corporate actors within governments. Moreover, rule making is often driven by 'repeat player' litigants who have the economic wherewithal and the political interest and insight to seek to shape procedural regimes.

3. *The Many Configurations of Lawsuits*

Given the diversity of problems, not surprisingly the configurations of lawsuits vary. Some cases involve two disputants seeking attention from a single judge, readily envisioned as a pyramidal triangle, with the judge at the top and the two disputants on either side. Other litigations are multi-party conflicts, for which a Calder mobile offers a better visual image; the many parties realign depending on what issue is to be determined in what forum.

Many questions have emerged, including: Should one-on-one cases be paradigmatic of adjudicatory processes? How should aggregate litigations affect one's understanding of the purposes and forms of adjudication? The costs of lawyers to individuals make pursuit of remedies expensive; evolving norms of judicial process discourage labor-intensive attention for 'small' claims. Further, the paradigmatic car crash, sometimes assumed to involve individual disputants, is often a dispute controlled by two insurance companies rather than by individual litigants. Many commentators urge that such conflicts be turned over to 'no-fault' administrative systems that would provide compensation according to schedules and that other 'small cases' be resolved through lower cost and lower visibility processes located in administrative agencies.

At the opposite end of the spectrum are large-scale cases, some involving tort law arising from accidents (like airplane crashes) and others related to allegedly defective or injurious products (such as diet drugs, nicotine, and asbestos). Some large-scale cases involve contracts and consumer claims, from shareholders of companies to purchasers of goods; and yet others relate to group-based rights claiming (about issues involving environmental harms to air, food, or water or the fair treatment of women). Such litigation also raises questions of whether administrative models are more appropriate modes of disposition than adjudicatory ones.

Given widespread allegations of injury and hundreds or thousands of lawsuits involving similar or the same parties, courts have devised means, both formal and informal, to process cases *en masse*. Procedural systems respond through mechanisms such as 'consolidation,' 'class actions,' 'interpleader,' 'bankruptcy,' 'multi-district litigation'—all procedures designed to assist in reaching resolutions across categories of claimants.

4. *A Variety of Venues*

Not only do lawsuits vary in size and scale, they also go to an array of places for resolution, both public and private. Once a decision is made to file a lawsuit, a choice often exists about where to file a lawsuit. Some courts have 'general jurisdiction'—they can handle all kinds of cases. Other courts have 'limited jurisdiction,' tied to a particular subject matter or amount in dispute.

In some instances, the jurisdictional grants overlap, and courts may have 'concurrent jurisdiction.' When more than one venue is available, common law systems give the power to the parties to decide the court in which to proceed. Upon occasion, defendants argue that plaintiffs have made the wrong selection—either because a particular tribunal lacks authority over the subject matter of the litigation ('subject matter jurisdiction') or because the tribunal has no power to compel a particular defendant to appear and defend or

is an inappropriate place to rule on the merits ('personal jurisdiction' or '*forum non conveniens*').

More institutions than courts are available for dispute resolution. Many government agencies are also quasi-courts. In the US for example, the Social Security Administration, the Equal Employment Opportunities Commission, and the Immigration and Naturalization Service process thousands of disputes annually, far in excess of the number decided by the federal court system. Further, court systems in courthouses or in administrative agencies often do not exhaust the set of options. Again using the US as an example, Indian tribes have court systems as well. Given the overlap among courts and the demographic mobility of the citizenry, inter-jurisdictional cooperation among state, federal, and tribal systems has begun to develop.

An increasingly global economy makes international decision-making tribunals ever more important to litigants within the US. One project, based at the American Law Institute, is to develop transnational rules of procedure to which countries or litigants might subscribe voluntarily. Another approach is to create universal jurisdiction for certain kinds of offenses; the International Criminal Court, a product of the Treaty of Rome, exemplifies that effort.

Governments are not the only source of courts. Associations may decide to establish their own 'private' courts and may compel those who contract with them or those employed by them to rely on such tribunals in lieu of proceeding in state-based courts. Further, governmental agencies and courts may also urge litigants to devise means of resolution of disputes without adjudication and may contract for services with 'private' institutions. During the second half of the twentieth century, a number of private providers of dispute resolution have come to the fore. As the market of private providers of dispute resolution services has expanded, some worry that publicly based and financed systems may lose support as those with the resources go elsewhere.

5. *Tiers of Decision Making*

If cases are decided within formal court or administrative systems (by judges, juries, or hearing officers), losing parties may have a 'right' to appeal to a higher court, which has (often limited) authority to review decisions made below. In many jurisdictions, there are three or four tiers of decision-makers, some working at the trial level, which has 'original' jurisdiction, and others in an appellate system. Some appellate systems are obliged to review lower court rulings when presented with an appeal, whereas others enjoy discretion to decide whether to review cases and if so, which ones. At each stage, parties typically have the option to settle or withdraw.

The concept of 'preclusion' bears on these tiers of decision making. The idea is that at some point, procedural systems prohibit further consideration of a problem—because participants have (knowingly or inadvertently) 'waived' their rights by failing to raise certain claims in a specified time or because participants have made and lost claims along the way. '*Res judicata*,' '*estoppel*,' 'double jeopardy,' and 'statutes of limitations' are some of the legal terms that capture these principles of closure and finality. Yet most procedural systems also have safety hatches—means by which to reopen and reconsider issues because of new information, changing circumstances, or the passage of time. Child custody decisions are typically subject to reopening, and in criminal cases, the availability of '*habeas corpus*' marks a narrow window of reconsideration. Herein lies an exemplary tension of procedural rules, attempting to capture both the impulse toward closure and the recognition of limited knowledge and the imperfections of judgment.

6. *The Demography of the Court House*

During the twentieth century, the demographics of the participants of the legal system changed. Women gained juridical rights to bring lawsuits, as did peoples of many colors. Further peoples of all colors and both genders gained rights of access to the professions of lawyering and judging. These individuals prompted professional interest in the demographics of the court house to understand more about the racial, ethnic, and gender composition of judges, jurors, lawyers, court personnel, witnesses, and litigants, and to learn whether the processes of decision making are affected by the demographics.

Judiciaries and/or bar associations in countries around the world convened special task forces on the effects of gender, race, and ethnicity in the courts and in the legal profession. These joint bench and bar efforts fall within a long tradition of concern about the administration of justice; judges and lawyers have engaged in studies ranging from the problems of prison litigation and the death penalty to how to handle large numbers of asbestos filings. Proceduralists thus have new knowledge and perspectives from which to consider the processes of justice.

The work on race, ethnicity, and gender in the courts has generated data about the composition of the judiciaries and the bar, as well as about the experiences of male and female lawyers of all colors, and more limited information on litigants. Most of the reports conclude that court systems—like the culture in which they exist—reflect stereotypic assumptions about race, ethnicity, and gender. Women judges and women lawyers are often in dramatic minorities as contrasted with the number of women litigants and, in systems with juries, of women jurors. Women and men

who are lawyers and judges, and peoples of differing colors report different perceptions of the relevance of gender, race, and ethnicity to processes of adjudication. Studies in the US have found that women attorneys believe that gender affects both process and outcome, and describe its effects on their professional lives and the legal system. In contrast, male judges and lawyers are less likely to report gender as relevant to the administration of justice or the legal profession. Studies of minority judges and lawyers have parallel data: people of color report the relevance of color while non-minorities describe its unimportance to the justice system. As a consequence of these projects, topics once considered gender- and race-neutral—from the specifics of jury instructions and courtroom conduct to larger questions about the role of the judge, the nature of lawyering, and the allocation of work among the courts—are being revisited to consider how to make good on the promises of fairness.

7. *The Role of Resources*

Because providing legal process costs money, takes time, and requires energy, the role of resources is critical to the study of procedure. Rules that treat all individuals as similarly situated belie the reality that some of us are richer than others, some of us have more access to help than do others, and some of us are more likely to be harmed than are others. Procedure must therefore consider how rules affect the costs of lawyering and the impact of rules that shift costs from one side to the other or that provide for government subsidies for lawyers, investigation, expert witnesses, filing fees and other costs associated with litigation. Economic analysis of many areas of law has become commonplace. In this arena, such inquiries consider the incentive structures created by procedural rules, the means by which adversaries can inflict costs on opponents, the mechanisms by which bargaining can occur, how rules about sanctions for failure to settle affect differently those who are 'risk prone' and those who are 'risk averse,' and the impact of attorney fee-shifting rules. Constitutional and political theorists, analyzing the meaning of the 'equal protection' and 'due process' guarantees, consider whether the government has an obligation to make courts readily accessible; if so, for what categories of claims, and how, given economic disparity, to craft rules to implement such principles.

8. *The Constant Questions*

Despite all the variation in the resources of litigants, the kinds of cases, the people and problems involved, the courts systems, and the procedures themselves, the underlying problems that procedural systems have to solve are the same: Who can seek redress? For what kinds of injuries? Requesting what remedies? Based on what kind of information? Presented to what decision-maker? With any review or possibility of reconsideration? Obviously these questions do not admit to easy, unvarying answers; ideas about procedure are deeply rooted in the social–political fabric, and the issues central to procedure are value-laden. The questions to be addressed include: Why have process at all? Why care about how a decision is made? If process is required, how much should be provided? Are opportunities to be heard adequate? How formal should a procedure be? How expensive should it be? Who should pay? How free should the parties be to initiate lawsuits? Who gets to litigate and who is foreclosed? And what about the decision-makers—what kinds of information should be required prior to decision? What remedies should courts be able to order? How much power should judges have, and when should that power be constrained? When may decisions be reconsidered, and when should they be considered 'final'? Over and over again, procedure returns to these basic, and central, issues.

See also: Administrative Law; Civil Law; Conflict Sociology; Courts and Adjudication; Criminal Law and Crime Policy; Law: Overview; Laws, Conflict of; Legal Process and Social Science: United States; Parties: Litigants and Claimants; Rules in the Legal Process

Bibliography

Cover R M 1975 *Justice Accused*. Yale University Press, CT

Cover R M 1986 Violence and the word. *Yale Law Journal* **95**: 1601

Damaska M 1997 *Evidence Law, Adrift*. Yale Press, CT

Eisenberg T, Yeazell S C 1980 The ordinary and the extraordinary in institutional litigation. *Harvard Law Review* **93**: 463

Fiss O 1978 The forms of justice. *Harvard Law Review* **93**: 393

Fiss O 1984 Against settlement. *Yale Law Journal* **93**: 1073

Galanter M 1972 Why the 'haves' come out ahead: speculations on the limits of legal change. *Law & Society Review* **9**: 95

Galanter M, Cahill M 1994 'Most cases settle': Judicial promotion and regulation of settlement. *Stanford Law Review* **46**: 1339

Heinz J P, Laumann E O 1994 *Chicago Lawyers: The Social Structure of the Bar*, rev. edn. Northwestern University Press, Chicago

Hensler D R 1989 Resolving mass toxic torts: myths and realities. *University Illinois Law Review* **89**

Hensler D R, Dombrey-Moore B, Giddens B, Gross J, Moller E, Pace N M 2000 *Class Actions Dilemmas: Pursuing Public Goals for Private Gain*. RAND, Santa Monica, CA

Langbein J 1985 The German advantage in civil procedure. *University of Chicago Law Review* **52**: 823

Lind E A, Tyler T R 1988 *The Social Psychology of Procedural Justice*. Plenum Press, New York

Menkel-Meadow C 1985 Portia in a different voice: speculations on a women's lawyering process. *Berkeley Women's Law Review* **1**: 39

Ninth Circuit Gender Bias Task Force 1994 The effects of gender: the final report of the ninth Circuit Gender Bias Task Force, republished. *Southern California Law Review* **67**: 727
Purcell E A Jr 2000 *Brandeis and the Progressive Constitution: Erie, the Judicial Power, and the Politics of the Federal Courts in the Twentieth-Century America*. Yale Press, CT
Resnik J 1982 Managerial judges. *Harvard Law Review* **96**: 372
Resnik J 2000a Trial as error, jurisdiction as injury. *Harvard Law Review* **113**: 924
Resnik J 2000b Money matters: judicial market interventions creating subsidies and awarding fees and costs in individual and aggregate litigation. *University of Pennsylvania Law Review* **148**: 2119
Schuck P 1987 *Agent Orange on Trial*. Harvard Press, NY
Shavell S 1982 Suit, settlement, and trial: a theoretical analysis under alternative methods for the allocation of legal costs. *Journal Legal Studies* **11**: 55
Subrin S H 1987 How equity conquered common law: the federal rules of civil procedure in historical perspective. *University of Pennsylvania Law Review* **135**: 909
Trubek D M, Sarat A, Felstiner W F, Kritzer H M, Grossman J B 1983 the costs of ordinary litigation. *UCLA Law Review* **31**: 72
Yeazell S 1994 The misunderstood consequences of modern civil process. *Wisconsin Law Review* 631
Woolf H 1996 *Access to Justice: Final Report to the Lord Chancellor on the Civil Justice Systems in England and Wales*. The Stationery Office, London

J. Resnik

Procrustes Analysis

Procrustes analysis is concerned with matching two, or more, configurations of points. Imagine two maps of a region, each giving the positions of the same towns. The two maps are oriented differently, are of different sizes, differ in their accuracies, and use different map projections. At a first glance it may be difficult to see much resemblance between the two maps. Procrustes analysis addresses such problems, usually taking one of the configurations as a target and fitting the other to it, optimally, by using appropriate transformations. The 'maps' may be derived in many ways and may be in several dimensions but, for the concept of matching to be meaningful, it is important that the configurations refer to two representations of the same things: the same towns, the same cases (in multidimensional scaling), the same variables (in factor analysis), the same products (in marketing), the same landmarks (describing biological organisms).

The name comes from Greek mythology where Procrustes was an innkeeper who had a bed (sometimes two). Travelers were stretched or limbs chopped to make them fit the bed; with two, adjustment was ensured. The pejorative implications are no worse in Procrustes analysis than in other forms of data-analysis where, misguidedly, data may be fitted into the straightjacket of an unrealistic model.

1. Factor Analysis

The origins of Procrustes analysis (see Hurley and Catell 1962) are in factor analysis, where loadings $\mathbf{X}_1$ of k factors $\mathbf{f}$ occur in the form $\mathbf{X}_1\mathbf{f}$ and so may be replaced by $(\mathbf{X}_1\mathbf{T})(T^{-1}\mathbf{f})$ showing that they are determined only up to the arbitrary $k \times k$ matrix $\mathbf{T}$. By choosing a suitable $\mathbf{T}$, we may match $\mathbf{X}_1$ 'the traveler' to a previously determined, or hypothetical set of, loadings given in a matrix $\mathbf{X}_2$ 'the bed'; $\mathbf{T}$ represents the 'chopping/stretching transformation. Usually the degree of match is judged by least squares, minimizing $\|\mathbf{X}_1\mathbf{T} - \mathbf{X}_2\|$ the sum of the squared distances between like points of the target and transformed configurations. Without further constraint, this is the multivariate multiplenegression problem with well-known solution

$$\mathbf{T} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2.$$

Two of the most important constraints on $\mathbf{T}$ are:

(1) $\mathbf{T}$ must be orthogonal (denoted here by $\mathbf{T} = \mathbf{Q}$); and

(2) $\mathbf{T}$ must be a direction cosine matrix (denoted here by $\mathbf{T} = \mathbf{C}$)

Both choices relate to the interpretation that $\mathbf{f}$ is expressed relative to an orthogonal system of coordinate axes and the columns of $\mathbf{T}$ give direction cosines of new axes relative to the original orthogonal axes. When $\mathbf{T} = \mathbf{Q}$, a new set of orthogonal axes is being sought that give a better match to $\mathbf{X}_2$; when $\mathbf{T} = \mathbf{C}$ we imagine $\mathbf{X}_2$ as being referred to oblique axes. The orthogonal Procrustes least-squares solution is $\mathbf{Q} = \mathbf{V}\mathbf{U}'$ where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ is the singular value decomposition of $\mathbf{X}_2'\mathbf{X}_1$. With $\mathbf{C}$, minimization is intrinsically more difficult and is less well defined than in the orthogonal case. This is because there are at least two ways of expressing the position of a point relative to oblique axes and also because there are good arguments for replacing ordinary leastsquares by a weighted form which measures distance in the metric appropriate to the chosen system of oblique axes, giving four variants in which $\mathbf{T}$ is either $\mathbf{C}$, $\mathbf{C}'$, $\mathbf{C}^{-1}$ or $(\mathbf{C}')^{-1}$. The simplest case, $\mathbf{T} = \mathbf{C}$, was first discussed by Mosier (1939) but not finally solved until Browne (1967) gave a practical numerical method for solving the normal equations. Special algorithms have been proposed by Browne and Kristof (1969) and Gruvaeus (1970) for the case $\mathbf{T} = \mathbf{C}^{-1}$. However, all problems may be tackled uniformly by proceeding as if the directions of all but one of the axes are known and then estimating the optimal direction of the remaining axis; cycling through all the

axes iteratively until a stable solution is found for all the directions.

Rather than handling the indeterminacy of factor loadings as above, the modern tendency is to use an internal criterion of simple structure such as Varimax (Kaiser, 1958) that has no reference to an external target $\mathbf{X}_2$. Alternatively, $\mathbf{X}_2$ may be coded in binary form where $x_{ij} = 1$ indicates a desired high loading of the ith variable on the jth factor and $x_{ij} = 0$ indicates a low loading. This latter form remains a type of Procrustes problem but least squares is no longer appropriate and is replaced by choosing $\mathbf{T}$ to maximize the ratio of the sum of squares of the squared high loadings to the sum of squares of the squared low loadings (see Lawley and Maxwell 1971).

2. Variants and Later Developments

The problems outlined above are referred to as two-sets Procrustes problems because two configuration matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are involved. There are many variants:

(1) alternative constraints on $\mathbf{T}$ such as $\mathbf{T}$ is a rectangular column, orthonormal, or is a permutation matrix;

(2) two transformations: minimize $\|\mathbf{X}_1\mathbf{T}_1 - \mathbf{X}_2\mathbf{T}_2\|$;

(3) isotropic scaling: minimize $\|\rho\mathbf{X}_1\mathbf{T} - \mathbf{X}_2\|$;

(4) anisotropic scaling: minimize $\|\mathbf{X}_1\mathbf{S}_R\mathbf{T} - \mathbf{X}_2\|$, or $\|\mathbf{X}_1\mathbf{T}\mathbf{S}_R - \mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{T} - \mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{S}_R\mathbf{T} - \mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{T}\mathbf{S}_R - \mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{S}_{R1}\mathbf{T}\mathbf{S}_{R2} - \mathbf{X}_2\|$ for unknown diagonal $\mathbf{S}_R$ and $\mathbf{S}_L$;

(5) replace least squares by e.g., L1-norm, inner-product trace $(\mathbf{X}_2'\mathbf{X}_1\mathbf{T})$, robust estimation method;

(6) two-sided Procrustes problems $\|\mathbf{R}\mathbf{X}_1\mathbf{T} - \mathbf{X}_2\|$ where $\mathbf{R}$ and $\mathbf{T}$ are to be estimated, usually under specified constraints;

(7) weighted forms—minimize trace $\mathbf{N}(\mathbf{X}_1\mathbf{T}_1 - \mathbf{X}_2)'\mathbf{W}(\mathbf{X}_1\mathbf{T}_1 - \mathbf{X}_2\mathbf{T}_2)$ where $\mathbf{N}$ and $\mathbf{W}$ are given matrices, one of which may be ignored;

(8) handling missing values.

Brief comments on the above list must suffice. It should be clear that $\mathbf{X}_1$ and $\mathbf{X}_2$ no longer represent factor loadings. Isotropic scaling allows for differences in size; the more esoteric versions of anisotropic scaling verge on the more pejorative interpretation of the Procrustes legend: real violence being done to the data. The product of two orthogonal transformation matrices is equivalent to one and the inner product and least-squares criteria then give the same result. Weights $\mathbf{N}$ are useful to cover replication or missing rows of $\mathbf{X}_1$ and need not be diagonal, though they usually are. Note that weights are given matrices whereas anisotropic scaling is expressed in terms of unknown diagonal matrices requiring estimation. However, sometimes, $\mathbf{W}$ may be a function of the unknown $\mathbf{T}$, as with oblique Procrustes analysis in a metric where $\mathbf{W}$ is $\mathbf{C}'\mathbf{C}$ or its inverse.

Changes in orientation are given by an orthogonal matrix $\mathbf{Q}$, which includes rotations and reflections, leaving intact the essential distance and angular properties of the configuration. A $p \times q$ orthonormal matrix $\mathbf{T} = \mathbf{P}$ with q columns represents a projection from p to q dimensions. Projections may be interesting:

(1) because $\mathbf{X}_2$ has fewer dimensions than $\mathbf{X}_1$;

(2) a match may be required in some specified number of dimensions, say, $q = 2$.

When $q < p$ we may still match in p dimensions by rotating in the higher dimensional space; e.g., we may rotate three points lying on a line to fit three points lying in a plane. All we have to do is to augment $\mathbf{X}_2$ by $p - q$ zero columns. That $\|\mathbf{X}_1\mathbf{P} - \mathbf{X}_2\|$ can be less than $\|\mathbf{X}_1\mathbf{Q} - \mathbf{X}_2\|$ (where $\mathbf{X}_2$ is augmented by zero columns) has been seen as a justification for preferring projections to rotations but this is not comparing like with like. I believe that projection Procrustes analysis is rarely justified.

Nowadays, applications of Procrustes analysis are more often than not more to do with matching configurations other than factor loadings (Commandeur 1990). For example, $\mathbf{X}_1$ and $\mathbf{X}_2$ may be the two sets of points given by different multidimensional scalings (MDS) of the same data, or they may refer to MDS solutions based on different proximity/distance matrices measured between the same p objects, perhaps based on different variables. Fruitful applications have been found in the food sciences, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are derived from scores given by two judges on traits of several products (e.g., Dijksterhuis and Gower 1991/2). Applications in the biological sciences include the analysis of shape where the configurations refer to the directly measured coordinates of landmark features on organisms (see e.g., Bookstein 1991, Dryden and Mardia 1999). In shape analysis the matching of one configuration to another is often termed registration.

3. Generalizations

Given k configurations $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_k$ we may match them in pairs by doing the corresponding two-sets Procrustes analyses, minimizing $s_{ij} = \|\mathbf{X}_1\mathbf{T}_1 - \mathbf{X}_2\mathbf{T}_2\|$ for all pairs (i, j). Each s_{ij} may then be assigned to a matrix $\mathbf{S}$ and analyzed by some form of MDS (see Gower 1971 for an example from physical anthropology).

Alternatively, we may minimize $\Sigma_{i<j}^{k}\|\mathbf{X}_i\mathbf{T}_i\mathbf{X}_j\mathbf{T}_j\|$, which for the least-squares criterion, rejecting the solution $\mathbf{T}_i = 0$, is equivalent to minimizing $\Sigma_{i=1}^{k}\|\mathbf{X}_i\mathbf{T}_i\mathbf{G}\|$ where $\mathbf{G} = \frac{1}{k}\Sigma_{i=1}^{k}\mathbf{X}_i\mathbf{T}_i$, often termed the group-average. Thus, generalized Procrustes analysis is a three-mode method of analysis. For some information on algorithms, see ten Berge (1977) To help interpretation, each dimension of the analysis may be summarized in an analysis of variance, partitioning

the total into terms for the group average and for departures from the group average. Some dimensions may be identified with noise and some with signal, further subdivided into displayed and undisplayed parts (Dijksterhuis and Gower 1991/2).

The concept of a group average plays an important part in several psychometric methods (see e.g., Individual Differences Scaling) as a useful summary of k data sets. Perhaps, for this reason, it has been suggested that maximizing the group average of Procrustes analysis should be a main objective of Procrustes analysis. For orthogonal $\mathbf{T}$, both approaches are equivalent, but for projections, maximizing the group average is not the same as minimizing the criteria given above. The situation gets very convoluted, especially when the variants listed in Sect. 2 are admitted. This is because when focusing on the group average, the transformations, isotropic, and anisotropic scaling may be regarded as being attached to $\mathbf{G}$ rather than to the $\mathbf{X}_i$, giving two versions of every variant.

The leading eigenvector of the matrix with (i, j)th element trace $(\mathbf{T}_i'\mathbf{X}_i'\mathbf{X}_j\mathbf{T}_j)$ estimate isotropic scaling factors and also when $\mathbf{T}_i = \mathbf{I}$, this matrix is the basis of the STATIS method. This and similar links between Procrustes analysis and other methods that involve a group average, are likely to be an area for research in the immediate future. Research relating to biological landmark data and shape analysis is currently active and will continue (see Kendall et al. 1999.). Other research will focus on algorithms, the effects of estimation criteria and the difficult area of inference.

See also: Configurational Analysis; Factor Analysis and Latent Structure, Confirmatory; Factor Analysis and Latent Structure: IRT and Rasch Models; Factor Analysis and Latent Structure: Overview; Latent Structure and Causal Variables; Multidimensional Scaling in Psychology; Multivariate Analysis: Discrete Variables (Correspondence Models); Multivariate Analysis: Discrete Variables (Overview); Scaling: Correspondence Analysis; Scaling: Multidimensional

Bibliography

Bookstein F L 1991 *Morphometric Tools for Landmark Data*. Cambridge University Press, Cambridge, UK

Browne M W 1967 On oblique Procrustes rotation. *Psychometrika* **32**: 125–32

Browne M W, Kristof W 1969 On oblique Procrustes rotation of a factor matrix to a specified pattern. *Psychometrika* **34**: 237–48

Commandeur J J F 1991 *Matching Configurations*. DSWO Press, Leiden, The Netherlands

Dijksterhuis G B, Gower J C 1991/2 The interpretation of generalised Procrustes analysis and allied methods. *Food Quality and Preference* **3**: 67–87

Dryden I L, Mardia K V 1998 *Statistical Shape Analysis*. Wiley, Chichester, UK

Gower J C 1971 Statistical methods for comparing different multivariate analyses of the same data. In: Hodson J R, Kendall D G, Tautu P (eds.) *Mathematics in the Archaeological and Historical Sciences*. Edinburgh University Press, Edinburgh, Scotland

Gruvaeus G T 1970 A general approach to Procrustes pattern rotation. *Psychometrika* **35**: 493–505

Hurley J R, Catell B B 1962 The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioural Science* **7**: 258–62

Kaiser H F 1958 The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**: 187–200

Kendall D G, Barden D, Carne T K, Le H 1999 *Shape and Shape Theory*. Wiley, Chichester, UK

Lawley D N, Maxwell A E 1971 *Factor Analysis as a Statistical Method*. 2nd end. Butterworths, London

Mosier C I 1939 Determining a simple structure when loadings for certain tests are known. *Psychometrika* **4**: 149–62

Berge J M F 1977 Orthogonal Procrustes rotation for two or more matrices. *Psychometrika* **42**: 267–76

J. G. Gower

Production Systems in Cognitive Psychology

A *production system* consists of a collection of if–then rules that together form an information-processing model of some cognitive task, or range of tasks. Production systems have some special properties that make them highly suited to modeling cognition. From their origins as models of problem solving, production systems have grown to become a major formalism for modeling cognitive skill and aspects of learning. Since the 1990s, they have become increasingly identified with the topic of integrated cognitive architectures.

1. What is a Production System?

A production system is a model of cognitive processing, consisting of a collection of rules (called *production rules*, or just *productions*). Each rule has two parts: a *condition* part and an *action* part. The meaning of the rule is that when the condition holds true, then the action is taken. The basic idea is best illustrated with a nonpsychological example. Consider the following simple production system, with just two rules, describing the behavior of a thermostat in controlling a heating system:

Rule 1: if temperature $< 20\,°\mathrm{C} \Rightarrow$ turn on heating.

Rule 2: if temperature $> 20\,°\mathrm{C} \Rightarrow$ turn off heating.

When the room temperature is below 20 °C, the condition part of Rule 1 is true, so the thermostat takes the action specified by the rule and turns on the heating. (The rule is said to *fire*.) When the temperature is above 20 °C, Rule 2 similarly fires and turns

off the heating. Taken together, the two rules specify a process which describes the behavior of the thermostat.

A production system serving as a cognitive model usually has many more than two rules, perhaps as many as thousands. The production system operates in a cyclic fashion. First, a rule whose conditions are satisfied is identified. Then, that rule is fired, i.e., its action or actions are carried out. Usually, the actions will change the current situation in some way, so that now a different rule has its conditions satisfied, and the cycle repeats. This process is referred to as the *recognize–act cycle*, because the condition parts of the rules can be regarded as forming an active network which monitors the current situation and recognizes when each individual rule is applicable, triggering the act of firing the rule.

1.1 Production System Architecture

A particular production system operates within a definite structure of memories and processes known as its *architecture*. (For clarification of the term 'architecture' used in that sense). The architecture of a production system includes at least two memories: the *production memory*, and the *data memory*.

The production memory holds the rules. It is taken to be a long-term memory, in that its contents persist over time. Indeed, usually it is assumed that a production rule, once acquired, is never lost from memory.

The data memory holds the dynamic information about the task being worked on. (In the research literature, this memory is often described as the 'working memory,' but that terminology is avoided here because of the possible confusion with the psychological notion of working memory, with which it has a close but not straightforward relationship, see *Working Memory, Psychology of*.) The data memory is taken to be a short-term memory whose contents disappear with time, indeed possibly during the processing of a task. It holds the information which is tested by the condition part of production rules. In other words, the conditions of production rules consist of tests or patterns which are matched against the contents of the data memory.

1.2 Conflict Resolution

In a production system with more than a small number of rules, it is almost inevitable that sometimes more than one production rule has its conditions satisfied. Such an occurrence is known as *conflict*. Some production system architectures simply fire, in parallel, all the rules whose conditions are satisfied. Most architectures, however, insist that a single rule be chosen to fire.

The procedure for selecting a single rule to fire from those that have their conditions satisfied, is known as *conflict resolution*. Different production system architectures use different principles for conflict resolution, which take account of factors such as the recency of the data being matched against, the activation levels of the data, the complexity of the rule conditions, the order in which rules were acquired, and so on.

2. Origins in Problem Solving

Production systems were introduced into psychology by Newell and Simon (1972) in their study of human problem solving. There is a natural connection between production systems and the way that Newell and Simon analyse their data.

Newell and Simon (1972) recorded participants thinking aloud as they solved puzzles in logic, chess, or cryptarithmetic (where digits have to be substituted systematically for letters to make a correct arithmetic problem). Such recordings are known as *verbal protocols* (see *Protocol Analysis in Psychology*). Newell and Simon use the protocols to reconstruct the problem-solving steps taken by the participant. They then look for regularities in the participant's behavior. If, for example, whenever the participant assigns a digit to a letter, he or she then scans for another place where the letter is used, it is natural to express that regularity in the form of a production-like rule: 'if digit newly assigned to letter, then scan for occurrence of letter.' With some further attention to detail, a collection of such rules can be treated as a production system, and it then becomes an empirical question as to how well the production system models the participant's problem solving.

3. Properties of Production Systems

A production system specifies a computational process, and is therefore a kind of programming language. However, production systems differ from ordinary computer programming languages in a number of ways.

(a) *Parallel and serial processing*. Production systems exhibit aspects of both parallel and serial processing, corresponding to the two phases of the recognize–act cycle. The recognize phase, i.e., the testing of conditions, occurs in parallel, across possibly hundreds of thousands of rules. In the act phase, a single rule is chosen for firing, so the actions are carried out one at a time.

(b) *Independence of the rules*. Each rule in a production system is potentially independent of all the other rules. Each rule can be written to be self-

contained, expressing an item of knowledge which the problem solver is postulated to know.

(c) *Interruptible and flexible control.* A production system is not forced to follow a fixed control structure, such as a rigid hierarchy of routines and subroutines. Because the processing at each step is determined by what the rule conditions 'recognize' in the data memory, the control can be very flexible. For the same reason, it is interruptible. If relevant information appears in the data memory, the production system can abandon its current line of processing and follow a new one.

These and other properties make production systems well suited to the modeling of cognitive performance and cognitive skill.

4. Models of Cognitive Skill

Production systems have been written to model many different cognitive skills. Although only a few researchers have actually written production system models, their influence, as with other forms of computer modeling, has been far beyond their number. Production systems have now become a medium of choice for theorizing about cognitive skill, especially for skills requiring a large body of knowledge. This section summarizes just a small sample of the published work.

4.1 Written Subtraction

A model of errors in children's written, multicolumn subtraction illustrates how some of the properties of production systems support their application as cognitive models. Young and O'Shea (1981) present a simple production system, of around a dozen rules, for performing subtraction correctly. They proceed to show how small perturbations of the model reproduce many of the errors commonly found in 10- to 12-year-old children's performance.

Several of the most common errors arise by simply omitting one or more rules from the production system. For example, if the rule is omitted that initiates borrowing when the lower digit is larger than the upper, then the model never borrows but instead just subtracts the smaller from the larger digit in each column—a mistake common among children. The implication is that some children have failed to acquire the appropriate rule. Other errors arise by adding erroneous rules, or erroneous versions of rules, to the production system. For example, children who usually subtract correctly but simply copy the lower digit when subtracting from zero can be modeled by adding the corresponding zero-pattern rule to the production system. Although inappropriate for subtraction, that rule is appropriate for addition, suggesting that some children have not acquired sufficiently distinctive conditions for the rule.

The ability to add and remove rules in this fashion and still leave a model which runs and faithfully reflects children's behavior, depends critically upon several of the properties of production systems mentioned above. Young and O'Shea (1981) discuss the crucial role of the independence of the rules, of the flexible control structure, and of the particular conflict-resolution principles used in the model.

4.2 Reading

Reading is another practically important cognitive skill which has been modeled with production systems. Just and Carpenter (1987) present an ambitious model of the reading and comprehension of single sentences and connected text. Their production system provides an integrated model of the cognitive processes involved in reading, from word encoding and lexical access, through syntax and semantics, to textual understanding. Many aspects of the model are based upon data from participants' eye movements (see *Eye Movements in Reading*).

4.3 Knowledge-intensive Cognitive Skill

Production systems are well suited to building models that encode and deploy large amounts of knowledge. For this reason, they are frequently used in the construction of *expert systems*, a branch of applied artificial intelligence that developed in the 1970s at around the same time as production systems were making their appearance in psychology (see *Expert Systems in Cognitive Science*). On the same grounds, production systems are much used in psychological models of expert skill, such as medical diagnosis, chess playing (see *Chess Expertise, Cognitive Psychology of*), and many other domains of knowledge.

5. Learning and Development

The ability to add rules to a production system, mentioned in Sect. 4.1 in the context of subtraction, raises the possibility of modeling certain aspects of learning and cognitive development by 'growing' a production system, one rule at a time, to correspond to the gradual acquisition of a cognitive skill. Some of the earliest applications of production systems were in the field of learning and development (see Klahr et al. 1987). It continues to be an active area of research. Simon and Halford (1995), for example, report on several computational models of the mechanisms of developmental transition, among which production systems figure prominently.

6. The Emergence of Cognitive Architectures

Since the 1990s, production systems have become increasingly identified with the newly emerging topic of integrated cognitive architectures, of which the

main exemplars are SOAR (Newell 1990) and ACT (Anderson and Lebiere 1998)*Cognitive Theory: SOAR*; *Cognitive Theory*: *ACT*).

The primary advantage of locating a production system within a broader, more encompassing theory is that the architectural details of the production system are determined in a principled way, and fixed. In consequence, the architecture of the production system becomes theory-laden, no longer serving just as a convenient and expressive notation but also carrying part of the burden of explanation. This removes extraneous degrees of freedom from the modeler, thereby leading to stronger explanations and theories.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence in Cognitive Science; Cognitive Theory: ACT; Cognitive Theory: SOAR; Connectionist Approaches; Connectionist Models of Concept Learning; Connectionist Models of Development; Connectionist Models of Language Processing; Discrete State Models of Information Processing; Information Processing Architectures: Fundamental Issues

Bibliography

Anderson J R, Lebiere C 1998 *The Atomic Components of Thought*. Erlbaum, Mahwah, NJ

Just M A, Carpenter P A 1987 *The Psychology of Reading and Language Comprehension*. Allyn and Bacon, Boston

Klahr D, Langley P, Neches R (eds.) 1987 *Production System Models of Learning and Development*. MIT Press, Cambridge, MA

Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Simon T J, Halford G S (eds.) 1995 *Developing Cognitive Competence: New Approaches to Process Modeling*. Erlbaum, Hillsdale NJ

Young R M, O'Shea T 1981 Errors in children's subtraction. *Cognitive Science* **5**: 153–177

R. M. Young

Professionalization of Social and Behavioral Scientists: United States

This article focuses on social and behavioral scientists as professionals in the United States. It addresses the nature of these professions, the process of professionalization, the role of credentials, patterns in degree production, and the contours of the labor force in social and behavioral science (hereinafter called social/related science) fields. In this context, professionalization refers to the processes through which individuals become socialized as professionals. Certification refers to the authorization or authentication of credentials. And, labor force refers to that portion of the population (in this case of these professions) that is economically active; that is, either employed or, if unemployed, available to work.

1. The Social and Behavioral Sciences as Professions

The social and behavioral sciences constitute more than a single profession. Disciplines like economics, political science, psychology, and sociology have their own specialized training, expertise, reward structures, associations, networks, and norms. As scientific fields dedicated to the study of human social phenomena, they share features in common. Yet, because of their arenas of knowledge, occupational cultures, collegial interactions, labor markets, and how they are perceived by others, they comprise different professional groups. Moreover, these professions are not the same as the academic profession—although many persons with advanced degrees in the social/related sciences (especially with doctoral degrees) are employed in colleges and universities and, as such, are also academic professionals.

2. The Process of Professionalization

Professionalization involves the development of skills, identities, norms, and values associated with becoming part of a professional group. Through this process, recruits to the social and behavioral sciences acquire both substantive and methodological knowledge and develop understandings of their roles that permit them to function as professionals in these fields. Also, by training newcomers, these professions seek to ensure that the work of their sciences will continue congruent with certain principles and practices.

Research on professionalization dates back to the 1950s. Early research focused on the medical (e.g., Becker et al. 1961, Merton et al. 1957) and legal (e.g., Lortie 1959, Warkov and Zelan 1965) professions. Scholars also began examining graduate education in relation to academic or scientific careers (e.g., Pease 1967, Wright 1964). As these and recent studies (see, e.g., Keith and Moore 1995) indicate, although professionalization continues during the early years of employment, graduate departments are the primary socialization sites. In departments, students learn what is expected and rewarded and, conversely, what constitutes unacceptable practices (e.g., Anderson et al. 1994).

Across scientific disciplines, graduate education aims to develop in students a level of scholarly

proficiency through formal (e.g., courses, qualifying exams, dissertations) and informal (e.g., attendance at scholarly meetings) training. A curriculum that initiates students progressively into more independent work and autonomous thinking helps develop professionals who can both master knowledge and produce original work (Sullivan 1991). Also important is a curriculum that examines the range of roles integral to being a professional scientist (e.g., teaching, engaging in peer review of others' work). Similarly, a curriculum needs to address what it means to be a professional in terms of oral and written communication, ethical conduct, and other career skills. Finally, quality mentoring is key (e.g., Keith and Moore 1995, Anderson et al. 1994). It makes a difference for students' productivity, belief in the norms of appropriate conduct, professional activity, self-confidence, and satisfaction with graduate school.

Through the professionalization process, graduate students also develop shared norms of science that transcend any one discipline. Merton (1973) saw these to be 'universalism'—research should be judged based on scientific merit; 'communality'—findings should be made public; 'disinterestedness'—research should be done to advance knowledge; and 'organized skepticism'—the scientific community should suspend judgment until the evidence is in. While scholars debate the influence of these norms and the presence of countervailing norms (Zuckerman 1988, Ritzer and Walczak 1986, pp. 225–35), few would dispute that there is a culture of science and that socialization in graduate school aims to engender support for and adherence to its norms.

Professionalization is important as well to the development of expertise, patterns of activity, and values that are discipline-specific. Scientific disciplines differ as to their intellectual tasks, their methods of inquiry, and how their work is done and rewarded (e.g., Biglan 1973, Lodahl and Gordon 1972). For example, chemistry in comparison to sociology has a longstanding tradition of intensive collaboration on more tightly defined problems. Thus, chemistry students are more closely supervised, work in research teams, and are involved in joint publication; sociology students have more intermittent contact with advisors, work more independently, and may or may not engage in collaboration (see Anderson et al. 1994, Babchuk et al. 1999). Sociology students, however, work and publish more collaboratively than do anthropology students, who are in a field more influenced by single-investigator, ethnographic research (Babchuk et al. 1999).

Not unexpectedly, departments vary within and across fields of science in how they structure training programs and the quality of what they do. While successful recruits exit with the same formal credential (a doctoral degree), their actual experiences differentially prepare them to work productively and navigate successful careers. Above and beyond any effects of department prestige on graduates' careers, the nature of graduate training has far reaching implications for professional development.

3. *Credentials and Certification in the Social/Related Sciences*

Credentialing generally refers to a system or process that establishes the status, legitimacy, or expertise of persons. With respect to individuals, credentials may be issued in the form of: (a) degrees, diplomas, or other certificates by educational organizations; (b) licenses regulated by states; or (c) certificates by professional associations. Institutions too are credentialed by governmental agencies (e.g., incorporation charters) or private accreditation processes (e.g., regional accreditation associations), which create standards for operation and attest to the quality of what is being offered (Freidson 1986, Chap. 4).

For academic and scientific professions, university-level training is key. In these professions, completion of a program of training culminating in the award of a degree is the most basic credential that attests to expertise. Credentials such as degrees formally bestow on individuals the right to practice a profession or use a professional title (Freidson 1986). The legitimacy of the degree, as Freidson pointed out, relies on the fact that educational institutions themselves are credentialed (i.e., accredited) by the state or, formally or informally, by professional associations.

In the social and behavioral sciences, the advanced degree awarded to an individual constitutes the basic professional credential. The masters degree provides a measure of stature for an individual as a professional in these sciences. While there is a growing market of interest for MA-level social/related scientists (especially outside of the academic workplace), the doctoral degree is still considered to be the full credential.

Psychology is the only social/related science discipline with a well-established process for certifying individuals beyond the advanced degree or for accrediting programs. To engage in clinical practice, psychologists need to be licensed by the states in which they intend to practice. Licensure in psychology requires demonstrating proficiency in designated areas and passing a standard examination created by the Association of State and Provincial Psychology Boards, an alliance of agencies across North America responsible for licensure and certification of psychologists. In terms of accreditation, the American Psychological Association (APA) has a widely used, voluntary process for accrediting doctoral programs in the specialty areas of clinical, counseling, or school psychology programs (or combinations thereof). Some applied or practice associations in other social/related science disciplines are initiating such efforts (e.g., in 1995, the Society for Applied Sociology joined with

the Sociological Practice Association), but none has substantially taken hold.

Despite intermittent calls within some social/related sciences, there has been no strong impulse to further attest to individuals' expertise through additional credentials or certification programs. The American Sociological Association (ASA), for example, established but quickly abandoned a certification program for social psychologists in the early 1960s due to concerns about their certification by APA. In 1986 and 1990, respectively, Ph.D. and MA certification programs were launched by the ASA but generated little interest. In 1992, the ASA's MA Certification Program Committee expressed doubts about continuing the Program, noting that certification of sociologists is handled through receipt of an MA or Ph.D. In 1998, the ASA Council terminated the Association's program.

In the end, the conferral of advanced degrees through the university seems to provide a sufficient credential across fields of science. For each social/related science discipline, only a small proportion of academic departments has the authority to prepare the next generations for careers in their profession. In 1997, approximately 240 of the 4,009 institutions of higher education awarded doctorates. Except for psychology, there are only 110–50 Ph.D.-conferring programs and 90–170 MA-only programs in the social sciences (psychology, with its clinical specialty, has approximately 435 doctoral- and 260 MA-only programs). Thus, while scientific professions generally do not control the market through elaborate mechanisms of certification and accreditation, the size and structure of their educational systems affect the supply of trained professionals.

4. Degree Production

4.1 BA Degrees

Undergraduate majors in social/related science fields constitute the potential pipeline of those who might be attracted to and eligible for becoming social and behavioral scientists. They are also an important 'client' group that creates demand for trained social/related scientists as faculty members in the academic workplace. And, finally, even with only a bachelor's degree, they are a sector of the future workforce that may ultimately be employed in jobs classified as social/related science occupations (ostensibly among those that do not require more advanced expertise or skills).

Compared to mathematical, physical, and biological sciences, undergraduates declare majors in the social/related sciences later in their academic careers. This is primarily attributable to the scarce attention given to the social/related sciences as fields of science in the elementary and secondary school curricula. Therefore, many undergraduates have their first exposure to these fields as part of fulfilling their general education requirements.

Despite later exposure to these fields, the social and behavioral sciences confer a substantial number of BA degrees. In 1967, even excluding psychology, there were twice as many BA degrees in the social sciences (72,398) as in the next highest field of science and engineering (36,197 engineering BAs). By 1997, the biological/related sciences had more than doubled, but there continued to be more BA degrees conferred in the social sciences than in other fields (see S&ED 2000, Table 5).

The number of BA degrees also varies within the social/related sciences. A comparison of four disciplines—economics, political science, psychology, and sociology—makes the point. As shown in Table 1, in 1967 these fields attracted similar numbers of majors, but by 1997 there was much more dispersion. Most obvious is the substantial increase, almost four-fold, in the number of BA degrees in psychology between 1967 and 1997. Also, there are fluctuations in levels of interest in different disciplines at different points in time. Over the current period—1990 to 1997 (not shown in table), sociology and psychology had the largest proportional increase in BAs (55 and 38 percent, respectively) while political science and economics conferred a declining number of degrees (about −10 and −30 percent, respectively). These data are based on S&ED 2000, Tables 50, 52–4.

4.2 MA and Ph.D. Degrees

As with BA production, both MA and Ph.D. production has generally increased across fields of science since the late 1960s (see S&ED 2000, Tables 12, 19). For most social and behavioral sciences, substantial growth was evident between 1967 and 1972, followed by a dip in the 1980s, and continued growth in the 1990s. At the MA level, the social/related sciences continued to put more MA-trained persons into the labor force than did other science and engineering fields (although there are also a substantial number of MA engineers). At the Ph.D. level, despite growth, there were also differences over time among fields. In the most recent 5-year period, doctoral degree production in the social sciences, biological sciences, and engineering continued to increase, but physical sciences and mathematical/computer sciences declined or remained level.

Within the social/related sciences, there were also variations in Ph.D. production by discipline (see Table 1). Between 1967 and 1997, the number of doctoral degrees awarded in psychology was substantially higher than in other disciplines in absolute numbers and in growth. In 1997, psychology conferred about 3,600 doctoral degrees, about six times more than

Table 1
Degrees awarded in Social/Related Sciences, select years 1967–1997

	Year						
	1967	*1972*	*1977*	*1982*	*1987*	*1992*	*1997*
Bachelor's Degrees							
Economics	13,829	16,408	16,674	21,880	24,402	24,976	17,701
Political Science	19,332	30,677	30,972	31,630	33,876	47,214	37,807
Psychology	19,435	43,421	47,794	41,539	43,195	64,033	74,734
Sociology	17,751	35,630	25,008	16,333	12,359	19,644	24,750
Master's Degrees							
Economics	2,147	2,636	2,662	2,506	2,321	2,564	2,804
Political Science	3,326	5,573	10,162	10,210	9,933	10,767	13,368
Psychology	2,898	5,293	8,320	7,849	8,165	9,852	13,633
Sociology	1,193	1,987	1,851	1,183	970	1,379	1,752
Doctoral Degrees							
Economics	691	1,056	983	940	960	1,051	1,161
Political Science	501	911	879	709	647	804	970
Psychology	1,295	2,279	2,990	3,159	3,173	3,263	3,566
Sociology	331	639	725	568	449	512	598

Source: Data are from the Department of Education/National Center for Education Statistics, Integrated Postsecondary Education Data System Completions Survey (IPEDS), and National Science Foundation/Division of Science Resources Studies (NSF/SRS) Survey of Earned Doctorates (SED). Data are tabulated by NSF/SRS, and published in *Science and Engineering Degrees: 1966–97*. (NSF00–310, NSF/SRS), Tables 50, 52–4 (S&ED 2000)

sociology and three times more than economics (although about 35 percent of these were clinical doctorates). Because of a decline in clinical doctorates in recent years, psychology has grown only modestly over the 1990s (about 10 percent) in comparison to political science and sociology that have both increased their Ph.D. degree production by about 30 percent (based on S&ED 2000, Tables 50, 52–4).

4.3 Women and Minorities and the Ph.D. Degree

4.3.1 Gender. Since the 1960s, there has been a dramatic increase in women receiving Ph.D. degrees across all fields of science and engineering (see S&ED 2000, Tables 19, 23, 26, 35, 49–51). In 1997, 8,921 women received their Ph.D. degrees in contrast to 1,096 women in 1967. These numbers represent a substantial difference in the shares of Ph.Ds. awarded to women (32.8 percent in 1997 and 8.4 percent in 1967). Fields, however, vary considerably in the proportion of Ph.Ds. awarded to women, with the social/related sciences and the biological sciences conferring a much higher proportion than the physical sciences and engineering.

There are notable differences within the social and behavioral sciences as well (see S&ED 2000, Tables 50, 52–4). For example, in sociology and psychology, women received approximately 20 percent of the Ph.Ds. in 1967, and 55 and 66 percent, respectively, in 1997. In comparison, in economics and political science, even with increasing proportions of women awarded Ph.Ds., in 1997 women still received no more than about 30 percent of these degrees.

4.3.2 Race/Ethinicity. The data on earned doctorates for racial and ethnic minorities were first reported in the 1970s. By 1977, across all fields of science and engineering, the number of persons of color receiving Ph.Ds. (about 750) was quite small (see SED 1997, Table 8). In 1977, proportions of minorities receiving Ph.Ds. ranged from 5 to 7 percent in the life sciences, physical sciences, social/related sciences, and engineering. While the proportions of persons of color receiving Ph.Ds. increased over time, by 1997, they were still quite low—ranging from 12 to 18 percent across fields of science.

Important differences are evident among fields of science in the percentage of doctorates awarded to racial/ethnic minority groups (see SED 1997, Table 8). In 1997, in the physical sciences, Asian–Americans received the highest proportion of Ph.Ds. at over 7 percent and blacks the lowest proportion at under 2 percent. The social sciences included much higher proportions of black and Hispanic Ph.Ds. Overall, across the social/related sciences, somewhat more than 5 percent of the Ph.Ds. were awarded to blacks, slightly less than 5 percent to Hispanics, and slightly less than 4 percent to Asian–Americans.

There are also differences within the social/related sciences (see S&ED 1991, Table 3, SED 1997, Table A-

Table 2
Percentage of scientists with degree training in field of occupation by employment in field of occupation, 1997

	Computer/Math Scientists		Life/Related Scientists		Physical/Related Scientists		Social/Related Scientists		Engineering	
Degree	%	N	%	N	%	N	%	N	%	N
BA	41.5	(722)	10.4	(885)	23.3	(382)	2.6	(2,316)	53.9	(1,380)
MA	51.8	(245)	29.0	(157)	40.3	(115)	21.6	(485)	58.8	(431)
Ph.D.	77.5	(37)	59.9	(163)	60.5	(122)	67.6	(161)	66.6	(98)

Source: National Science Foundation, Scientists and Engineering Statistical Data System (SESTAT 1997): Table A-3
*Number is rounded in thousands of scientists employed at highest degree attained (e.g., 2,316 is 2,315,500 social/related scientists employed with highest degree attained at BA level)

2). For example, in 1997, economics and sociology awarded 28 and 25 percent, respectively, of their Ph.Ds. to racial and ethnic minorities, but varied as to which minorities received Ph.Ds. In economics, 16 percent of the Ph.Ds were awarded to Asian–Americans and 6 percent to African–Americans; in sociology, almost 10 percent were awarded to African–Americans and 6 percent to Asian–Americans. The overall small number of Ph.Ds. awarded to racial/ethnic minorities signals the need for more intentional outreach to *all* persons of color at earlier stages in education. Despite increases in awarding Ph.Ds., the proportion conferred to racial/ethnic minorities is still below their proportion in the US population.

5. *The Social and Behavioral Science Labor Force*

As noted in the introduction, the labor force includes those persons who are economically active—either employed, or, if unemployed, available for work. In 1998, according to the Current Population Survey (CPS) conducted by the US Census Bureau, approximately 137 million persons were active in the civilian workforce of the US, including 6 million unemployed persons. Social and behavioral science professionals constitute only a miniscule segment of this overall labor force (approximately one-half of 1 percent) if such professionals are defined as persons having an advanced degrees (an MA or Ph.D.) in a social/related science.

5.1 *Employment Status of Social and Behavioral Scientists*

The overall employment status of those with advanced degree training in the social and behavioral sciences is quite favorable. According to the National Science Foundation's Scientists and Engineering Statistical Data System (SESTAT), there are some 656,100 social/related science professionals in the labor force (494,100 and 162,000 at the MA and Ph.D. levels, respectively). Employment rates for these individuals are strong. In 1997, only 0.9 percent of all Ph.Ds. and 1.8 percent of all MAs in the social/related sciences were unemployed (see SESTAT 1997, Table B-2).

Of the 160,500 employed Ph.Ds. with social and behavioral science training, the vast majority (87 percent) was employed full-time (see SESTAT 1997, Tables B-2 and B-5). Also, approximately 117,000 of these Ph.Ds. were employed in science and engineering occupations. This proportion (73 percent) is consistent with other fields of science (see SESTAT 1997, Table A-3). Also, as might be expected, most of these Ph.Ds. (108,000) were working directly in social/related science occupations.

5.2 *Social and Behavioral Science Occupations and Professional Training*

In terms of standard approaches, the labor force in the social/related sciences can be conceptualized as: (a) those persons *trained* in the social and behavioral sciences, or (b) those persons *working* in social and behavioral science *occupations*. In contrast to a focus on degree training *per se*, classification into occupational categories includes persons trained in other fields who are working in these areas but excludes persons with relevant training who are working in other occupational areas. For example, social/related science occupations would include persons with Ph.Ds. in business who are teaching sociology in a university but would exclude persons with Ph.Ds. in sociology managing a business (see National Science Foundation 1999).

SESTAT classifies persons into occupations in terms of both work and education. Based on the SESTAT surveys, in 1997, approximately 350,000 persons with a BA degree or higher were classified as being in a social/related science occupation. This occupational category includes jobs that vary widely as to the nature of the work and required expertise. As might be expected, most persons classified as working in this occupation received their highest degree in a social/related science (approximately 80 percent across all degree levels). Furthermore, having an advanced degree beyond the BA also matters: the vast majority of persons working in a job classified as a social/

related science occupation have an advanced degree (see SESTAT 1997, Table A-3).

The value of an advanced degree can further be seen by looking at the relationship between BA, MA, or Ph.D. training in a field and working in an occupation classified as being in that same field (i.e., 'in-field' occupation). As shown in Table 2, for all fields of science, persons at successively higher degree levels are more likely to be working in occupations classified as 'in-field.' In the social and behavioral sciences, 2.6 percent of those persons with BAs was working in a social/related science occupation compared to 21.6 percent and 67.6 percent, respectively, of those with MA and Ph.D. degrees. By the Ph.D. level, there is also far less variability across fields of science in the proportion of persons working in 'in-field' occupations.

5.3 *Sectors of Employment*

Examination of Ph.D.-level professionals in the social and behavioral sciences shows the centrality of the academic workplace in relation to other work sectors. Across scientific fields, more Ph.Ds. were employed by 4-year colleges and universities than by any other employment sector (in 1997, 43.4 percent for all degree fields and 47.4 percent for the social/related sciences). The next highest work sector for all Ph.D. fields was business at 30.5 percent overall, although not as high a proportion of social science/related scientists was employed by the private sector as in other fields (see SESTAT 1997, Table C-2).

While specific disciplines within the social and behavioral sciences followed this overall pattern, there are important differences. As shown in Table 3, in 1997, a substantially greater proportion of sociologists (72 percent) than psychologists (34 percent) were employed by 4-year colleges and universities. In contrast, greater proportions of psychologists and economists were employed in the business sector. Also, for psychology, a larger proportion of Ph.Ds. is self-employed (17 percent), ostensibly due to the sizable number of psychologists who are in clinical practice.

5.4 *Women and Minorities in the Social/Related Science Labor Force*

5.4.1 Gender. Based on the 1997 SESTAT surveys, of the 160,500 employed Ph.Ds. with social/related science training, 62 percent were male and 38 percent were female (see SESTAT 1997, Table C-2). As could be anticipated from changes in the gender composition of Ph.D. degrees conferred over the years, substantially more employed female Ph.Ds. received their degrees in recent years. Of those who received their degrees 20–4 years ago, approximately 70 percent were male. In contrast, of those who received Ph.Ds. within the last 5 years, approximately 45 percent were male (see SESTAT 1997, Table C-10). This pattern holds across social/related science disciplines—from psychology that substantially increased in the production of female Ph.Ds. to economics that continued to produce more male Ph.Ds. (see SESTAT 1997, Table C-9).

Males and females with Ph.Ds. in the social/related sciences vary only modestly in terms of sectors of employment. Gender differences within disciplines are more striking than any overall trends. For example, a higher proportion of male than female economists was employed by 4-year colleges and universities (56 versus 49 percent) while a higher proportion of female than male economists was in the government sector (20 versus 12 percent) (see SESTAT 1997, Table C-2).

5.4.2 Race/Ethinicity. Of the 160,500 Ph.Ds. in the social and behavioral sciences employed in the workforce in 1997, the vast majority across social/related science disciplines was white (88 percent). The data on Ph.Ds. of color in the workforce are consistent with what is known about the production of minority Ph.Ds. since the mid-1970s (see SESTAT 1997, Table C–9). Of those who received their Ph.D. degrees 20–4 years ago, 92 percent were white and 2.9, 2.5, and 1.7 percent were black, Asian-American, and Hispanic, respectively. In comparison, of those receiving degrees within the past 5 years, 82 percent were white and 4.4, 9.5, and 3.7 percent were black, Asian-American, and Hispanic, respectively (see SESTAT 1997, Table C-11).

The 1997 data on employment sector reveal an important difference between minority and white Ph.Ds. in the social/related sciences (see SESTAT 1997, Table C-4). Higher proportions of Ph.Ds. of color were employed in 4-year colleges and universities than were their white counterparts (57, 56, and 54 percent, respectively, for blacks, Asian–American, and Hispanic Ph.Ds. compared to 46 percent for white Ph.Ds.). That higher proportions of minority Ph.Ds. are employed in higher education compared to white Ph.Ds. likely reflects strong efforts by the academic sector to diversify its workforce in the context of a very low number of Ph.Ds. of color being trained.

5.5 *Current Patterns for New Ph.Ds.*

Analysts examining the current employment picture for social/related scientists are variously optimistic, with some being quite favorable (Spalter-Roth et al. 2000) and others being more cautious (Siegfried and Stock 1999). A recent survey of graduates awarded their Ph.D. degrees between July 1996 and August 1997 provides important data. The survey focused on 14 scientific fields, including economics, political science, psychology, and sociology.

Table 3
Employed Social/Related Scientists in selected fields where highest degree attained is Ph.D. by employment sector, 1997

	Social/Related Scientists			
Employment Sector	Economists Percent	Political Scientists Percent	Psychologists Percent	Sociologists Percent
Universities and 4-year colleges	57.1	67.4	34.3	71.7
Other educational	0.6	3.1	5.6	3.9
Private not-for-profit	5.0	5.0	10.1	7.6
Private for-profit	16.7	9.5	21.9	6.2
Self-employed	2.2	3.6	16.9	3.4
Government	13.9	11.3	11.2	7.1
Total	**95.5	100.0	100.0	100.0
Total number*	20,080	15,820	79,320	13,230

Source: National Science Foundation, Survey of Doctoral Recipients (SDR 1997). Data are published in *Characteristics of Doctoral Scientists and Engineers in the United States 1997*, NSF00-308, NSF/SRS, Table 17
*Number is estimated number of persons employed at highest degree attained. Numbers may not add up to 100 percent due to rounding
** Total is not 100 percent because 4.5 percent of economists were classified in 'Other' employment sector (not shown on this table). Also, 0.1 percent of psychologists were so classified

As shown in Table 4, unemployment is generally quite low for newly exiting graduates. With the exception of chemistry and political science at 4.6 and 7 percent, respectively, other disciplines were all below 4 percent. Furthermore, Ph.Ds. in the social/related sciences fared very well. While compared to other scientists, a somewhat higher proportion of social and behavioral scientists were employed part-time (in particular political science and psychology), a lower proportion were in temporary positions. Most striking is the relatively high proportion of social/related science Ph.Ds. employed in the academy, with sociology being the highest at 83 percent. Also, these data suggest a favorable academic market for social/related scientists. Newly minted Ph.Ds. in economics, political science, psychology, and sociology were finding employment in the academy in proportions comparable to, if not higher than, their overall presence in this sector (compare Table 3).

6. *Continuing Issues for the Social and Behavioral Science Professions*

For at least the next 10 years, the academic market appears to be positive. Despite the complexities of making predictions (see, e.g., Breneman 1988, Smelser and Content 1980, Schuster 1992), Bowen and Sosa (1989) anticipated these trends. In particular, the picture for social/related scientists is consistent with their projections of an undersupply of Ph.Ds. due to such factors as high undergraduate enrollments, stable to diminished graduate enrollments, anticipated faculty retirements, and increases in nonacademic employment. Bown and Sosa projected more demand than supply through 2007 and believed this pattern would continue, though at a slower rate.

It is important to keep in mind that the academic marketplace is broad and undergoing a restructuring. In reality, it consists of multiple and overlapping markets (e.g., research universities, liberal arts colleges) that have different 'buyers' and 'sellers' (see Smelser and Content 1980, Schuster 1992) and that vary in prestige and mobility. Furthermore, these markets themselves are not fixed. The community college, for example, is currently the fastest growing sector of academic employment. Finally, potential changes in the structure of these markets (e.g., increases in student-faculty ratios, more use of part-time or contractual teaching appointments, increased use of distance learning) could affect Ph.D. shortages.

The market for social/related scientists outside of the academy also merits more attention. For certain disciplines, like economics and psychology, the proportions working in nonacademic settings are sizable. Furthermore, the data indicate that social/related scientists are increasingly located in these settings. A National Academy of Sciences report (1969) observed a slow but evident trend to greater nonacademic employment. Between 1968 and 1997, there was a 12 to 20 percent decrease in the proportion of sociologists, political scientists, economists, and psychologists employed in educational institutions. Consistent with commentators (e.g., Rosenfeld and Jones 1988, Schuster 1992) noting a 'pull' not a 'push' to these work sectors, the steady increases in nonacademic employment seem to occur irrespective of the tightness of academic markets.

Table 4
Employment status of recent Ph.D. graduates

Field of Study	Percent			
	Unemployed	Employed Part-time	Employed in temporary positions	Employed in academe
Biochemistry and Molecular Biology	4.0	2	79	73
Chemistry	4.6	2	58	57
Chemical Engineering	3.2	0.5	26	25
Computer Science	2.4	1	17	43
Earth and Space Sciences	3.9	8	53	53
Economics	2.3	5	19	59
Engineering	2.7	2	24	32
Mathematics	2.4	5	46	70
Microbiology	2.2	1	81	70
Physics	1.8	3	57	45
Physiology	2.7	2	75	82
Political Science	7.0	14	41	78
Psychology	0.6	24	25	37
Sociology	3.4	8	35	83

Source: Prepared by Commission on Professionals in Science and Technology (1998); Employment of recent doctoral graduates in science and engineering: Results of professional society surveys. www.cpst.org/web/site/pages/pubs/pdf_pubs/98data-1.pdf. Data on sociology were updated by the American Sociological Association based on a more complete response

As the social and behavioral sciences move into their second century as professions, the overall environment for these fields is currently strong. The challenges before them, however, are also many. They are at a juncture where their ambitions could enlarge and these professions could grow to address some of the most challenging issues of human behavior and society. How they organize themselves and their work and how they prepare their next generations for contributing knowledge and expertise will, more than any other factors, shape what these professions are and become (see *Infrastructure: Social/Behavioral Research (United States)*).

See also: Education and Employment; Education (Higher) and Gender; Education: Skill Training; Educational Institutions and Society; Educational Institutions, History of; Educational Productivity; History, Professionalization of; Knowledge, Education, and Skills, Geography of; Professionalization/Professions in History; Professions in Organizations; Professions, Sociology of

Bibliography

Anderson M S, Seashore L K, Jason E 1994 Disciplinary and departmental effects on observations of faculty and graduate school misconduct. *Journal of Higher Education* **65**: 330–50

Babchuk N, Bruce K, Peters G 1999 Collaboration in sociology and other scientific disciplines. *The American Sociologist* **30**: 5–21

Becker H S, Geer B, Hughes E C, Strauss A L 1961 *Boys in White: Student Culture in Medical School*. University of Chicago Press, Chicago

Biglan A 1973 Relationships between subject matter characteristics and the structure and output of university departments. *Journal of Applied Psychology* **57**: 204–13

Bowen W G, Sosa J A 1989 *Prospects for Faculty in the Arts and Sciences*. Princeton University Press, Princeton, NJ

Breneman D W 1988 Research on academic labor markets: Past and future. In: Breneman D, Youn T I K (eds.) *Academic Labor Markets and Careers*. Falmer Press, New York

Commission on Professionals in Science and Technology 1998 Employment of recent doctoral graduates in science and engineering: Results of professional society surveys. [www.cpst.org/web/site/pages/pubs/pdf_pubs/98datas-1.pdf]

Freidson E 1986 *Professional Powers: A Study of the Institutionalization of Formal Knowledge*. University of Chicago Press, Chicago

Keith B, Moore H A 1995 Training sociologists: An assessment of professional socialization and the emergence of career aspirations. *Teaching Sociology* **23**: 199–214

Lodahl J B, Gordon G 1972 Structure of scientific fields and the functioning of university graduate departments. *American Sociological Review* **37**: 57–72

Lortie D C 1959 Laymen to lawmen: Law school, careers, and professional socialization. *Harvard Educational Review* **29**: 352–69

Merton R K 1973 The normative structure of science. In: Merton R K (ed.) *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press, Chicago (First published Merton R K 1942 Science and technology in a democratic order. *Journal of Legal and Political Sociology* **1**: 115–26)

Merton R K, Reader G G, Kendall P L 1957 *The Student-Physician*. Harvard University Press, Cambridge, MA

National Academy of Sciences, Social Science Research Council 1969 *The Behavioral and Social Sciences: Outlook and Needs*. National Academy of Sciences Press, Washington, DC

National Science Foundation, Division of Science Resources Studies 1999 Counting the S&E workforce—It's not that easy.

Issue Brief-NSF-99-344, May 3,1999. National Science Foundation, Washington, DC
Pease J 1967 Faculty influence and professional participation of doctoral students. *Sociological Inquiry* **37**: 63–70
Ritzer G, Walczak D 1986 *Working*: *Conflict and Change*, 3rd edn. Prentice Hall, Englewood Cliffs, NJ
Rosenfeld R A, Jones J A 1988 Exit and re-entry in higher education. In: Breneman D, Youn T I K (eds.) *Academic Labor Markets and Careers*. Falmer Press, New York
Schuster J 1992 Academic labor markets. In: Clark B R, Neave G R (eds.) *The Encyclopedia of Higher Education, Vol. 3, Analytical Perspectives*. Pergamon Press, New York
S&ED 1991 *Science and engineering doctorates: 1960–90. Detailed statistical tables, NSF-91-310 Final*. National Science Foundation, Washington, DC
S&ED 2000 *Science and engineering degrees: 1966–97. Detailed statistical tables*. National Science Foundation, Washington, DC
SESTAT 1997 *Characteristics of scientists and engineers in the United States:1997 Detailed statistical tables*. National Science Foundation, Washington, DC
Siegfried J J, Stock W A 1999 The labor market for new Ph.D. economists. *Journal of Economic Perspective* **13**: 115–34
Smelser N J, Content R 1980 *The Changing Academic Market: General Trends and a Berkeley Case Study*. University of California Press, Berkeley, CA
Spalter-Roth R, Thomas J, Levine F J 2000 New doctorates in sociology: Professions inside and outside the academy. *Research Brief*, Vol. 1, no. 1. American Sociological Association, Washington, DC
Sullivan T A 1991 Making the graduate curriculum explicit. *Teaching Sociology* **19**: 408–13
Survey of Doctoral Recipients 1997 *Characteristics of Doctoral Scientists and Engineers in the United States, 1997 Detailed statistical tables*. National Science Foundation, Washington, DC
Survey of Earned Doctorates 1997 *Doctorate Recipients from United States Universities: Summary Report 1997*. National Science Foundation, Washington, DC
Warkov S, Zelan J 1965 *Lawyers in the Making*. Aldine, Chicago
Wright C R 1964 Success or failure in earning graduate degrees. *Sociology of Education* **38**: 73–97
Zuckerman H 1988 The sociology of science. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA

F. J. Levine

Professionalization/Professions in History

In modern historical scholarship, the professions include the occupations of doctors, lawyers, professors, and engineers, among others. Professions are characterized by special functions and structures, motives, representations, forms of knowledge, and sociocultural styles. Professionalization refers to processes affecting the social and symbolic construction of occupation and status. The history of the professions investigates the ways in which functional groups and status groups are regulated and institutionalized. It shows how meanings, functions, and structures develop in social and temporal context and how professions, in the restricted sense of the term, differ from the status groups and occupations of entrepreneurs, merchants and clerks, skilled manual workers (crafts, *métiers*) and the so-called semi-profession. Professionalization involves the formation of an occupation, on one hand, and interrelated developments regarding the social division of labor, structures of authority, and sociocultural inequality, on the other. Historical research on professionalization concentrates on such issues as the distribution of scarce resources at a particular point in time, the social definition of behavioral prerogatives, and the regulation of central values and functions. It is concerned with the motives, interests, and strategies of actors who either promote or hinder processes of professionalization.

Recent historical research on professionalization has tended to oscillate between individualizing historical perspectives and systematic social science perspectives. The former emphasize the peculiarities of particular professions or the special features of the professions in a particular historical context. The latter concentrate on general features and tendencies. Historical and comparative research shows that, while the professions are often concerned with the same or similar functions, problems, and tasks, these may be interpreted in different ways and institutionalized in different contexts. Similarly, depending on the characteristics of a particular profession and on the particular historical circumstances, the paths, cycles, and types of professionalization and deprofessionalization may vary. By taking these points into consideration, scholars have relativized both teleological understandings of professionalization and static conceptions of the professions.

1. Definitions of Profession and Professionalization

Abstract and temporally undifferentiated concepts of the professions have only a heuristic value in historical research. Therefore, one usually begins with a historical explication of the concept and then asks how professionalization is articulated in a specific context with regard to the relevant dimensions in particular domains. The individual dimensions and processes are more or less closely linked in concrete historical cases. The result is that the professions and professionalism find expression in a graded series of variations. The professions are characterized ideally by the following series of interrelated dimensions. Capabilities and skills are justified scientifically or systematically. Professional knowledge is described in such terms as 'exclusive,' 'more profound,' 'inaccessible to lay persons,' and 'not easily understandable.' It is acquired in special institutions such as universities, professional schools, and internships with professionals. It forms the basis of the superiority of the professional over lay

persons and clients. Also included in this body of knowledge are rules and attitudes which govern its application in a way which promotes trust between the professions and their social environment. These include formal procedures, ritual, titles, a professional habitus of studiousness, collegiality, and altruism, and a general orientation toward the common good. The professions claim for themselves a specific style of earning a living that is characterized by the principle of remuneration 'befitting ones rank' and through specific forms of contract and recompense. The professions demand for themselves higher prestige. They develop specific organizations for collective integration, self-administration, self-control, and for the representation of their interests, such as the 'chartered professions' in England, the disciplinary councils and chambers on the European continent, and, more generally, professional associations of various kinds. The professions strive for 'autonomy.' Viewed historically, this means, first, exiting from relations of patronage with traditional secular and clerical elites and, second, the reduction of external control and dependence on the state, on the rich and powerful, on clients or beneficiaries, and on employers.

Professionalization is a social project, which articulates itself in discourses on science, performance, honor, public welfare, and society and, in the ideology of professionalism. It is intended to secure for professionals a high or mid-range status in the system of social and cultural inequality. The historical literature distinguishes between professionalization that is externally controlled and professionalization that is controlled by the occupational groups themselves. Similarly, it distinguishes between heteronomous and autonomous professions. In the classical research on the 'free professions,' especially in the Anglo-Saxon literature, professionalization is understood largely from the perspective of members of the occupational groups themselves. In recent research, however, the analytical concept of professionalization is used increasingly for professionals who are occupied in state and private administrations and enterprises as well (Burrage and Torstendahl 1990, Perkin 1989, Siegrist 1988). Previously, and especially on the European continent, scholarly interest in professions such as these was often guided by concepts such as 'higher education,' 'meritocracy,' 'bureaucracy,' and 'civil servants.'

The historical connection between occupational role and social status is articulated concretely in concepts such as the 'gentlemanly professions,' the 'learned professions,' or, more simply, the 'professions' in England and the United States; '*professions libérales*,' '*bourgeoisie à talents*,' '*capacités*' in France; '*gelehrte Stände*,' '*höhere Berufe*,' '*bürgerliche Berufe*,' '*mittelständische Berufe*,' '*Akademiker*,' and '*Bildungsbürgertum*' in the German-speaking lands of Europe; '*professioni maggiori*,' '*professioni nobili*,' *and* '*borghesia unmanistica*' in Italy; 'intellectuals' in Western Europe and '*intelligentsia*' in Eastern Europe (Siegrist 1988, 1996; Freidson 1986; Conze and Kocka 1985; Malatesta 1995; Charle 1996; McClelland et al. 1995). In the twentieth century, professionals express claims to membership in the middle classes through discourses on the 'loss of middle-classness' and the 'threat of proletarianization.' Under National Socialism and Fascism, the meaning of the 'free professions' was hollowed out, as professionals were redefined as militant 'thinking workers,' who took part in the 'struggle of the Volk and the master race.' In the communist world, the concept of the socialist intelligentsia implied the transcendence of class relations, the end of occupational privileges, and the subordination of professionals to the state and the party. Modern western concepts such as 'technocrat' and 'expert' define the role and position of professionals independently of social or economic class.

2. *Conceptions and History of Professionalization and Professions*

The history of the professions begins in the Middle Ages and in the early modern era with the autonomous corporate colleges of the higher professions (lawyers, physicians, university professors) and the lower professions such as surgeons and notaries (Prest 1986, Betri and Pastore 1997). The various professions have specific orders of knowledge. They enjoy privileges with regard to education, the control of access to the professions, the practice of the professions, and ethical monitoring. The specific way of earning a livelihood distinguishes the higher professions from commercial occupations and handicrafts. The institutional pattern and the mentalities of the old European corporate occupational culture varied according to the degree of the abstraction of knowledge and the social proximity to the nobility, the patriciate, and the urban bourgeoisie. As a rule, the right to practice a profession was limited to a specific territory.

The traditional culture of the professions came under pressure and began to dissolve in the eighteenth and early nineteenth centuries. In the authoritarian and bureaucratic societies of Europe, the reformed professions were ruled under the control of the state. In radical-liberal, democratic, and egalitarian societies such as France in the revolutionary era, the United States, and Switzerland in the mid-nineteenth century, the professions were largely deregulated or were established as free trades (Ramsey 1984; Siegrist 1988, 1996). The general orientation of the different professions varied between (a) loyalty to and dependence on the state, (b) a progressive-liberal worldview, and (c) the desire to defend or revitalize traditional professional cultures and corporate mentalities. Different professions displayed many similarities with regard to the level of education, the orientation toward science and scholarship, and the social position of their

members. Nevertheless the various occupational groups often competed with one another for cultural hegemony, social prestige, and political power.

2.1 Institutionalization of Professional Training

The specific way in which education and training are institutionalized plays a central role in processes of professionalization. The states of the European continent led the way in the creation of theoretically and scientifically based training programs and in the transfer of such programs to (reformed) universities and elite institutions of higher learning such as the French Grandes Écoles. On the other hand, it is also true that (complementary) training through practice survived and even retained a central importance (trainees in state agencies, mandatory internships with practitioners, or, for example, when the first professional years serve, in effect, as a sort of apprenticeship). In Anglo-Saxon countries, the transition from 'training on the job' or 'shop culture' to 'school culture' has intensified since the late nineteenth century (Lundgreen 1990). A comparison of the length of time required for professional training in different countries reveals significant differences, even for the same occupation. Some of these differences may be attributed to variations in the contents of training programs, while others depend on factors such as tradition, restrictive state policies affecting the labor market, or the exclusive strategies of the professions themselves.

2.2 The Institutionalization of Professions

Beginning in the eighteenth century, it is important to distinguish between the legal institutionalization of the professions by the state, on one hand, and the formalization and regulation that is intended and implemented by the occupational groups themselves, on the other. With regard to the latter, one must also distinguish among (a) the continual development of a premodern profession, which, in principle, was never dissolved (such as the English 'bar'); (b) the restructuring of professions, which had been temporarily deinstitutionalized (as was the case with physicians and lawyers in the United States); and (c) the formation of new professions by members of functional and occupational groups that first arose in the nineteenth century (especially engineers but also a number of other 'new' professions). All of these involve long processes, which typically culminate in accreditation through the state or the legislation.

In much of Europe, the state or the government or legislation took the initiative to regulate and institutionalize the professions (Siegrist 1996; Ramsey 1984; McClelland et al. 1995). The state, as the guarantor of the legal order, the promoter of culture, and as an intervening force in society, needed professional services and provided the professions with knowledge, rights, and duties. In this way, the state distinguished the professions symbolically from the occupations and functional roles of business, commerce, and agriculture, which had been deregulated through the abolition of guilds and the liberalization of the trades. Henceforth, the professions included only those vocations that were occupied with central goods and values or with securing property, honor, health, public order, morality, and science. In this new configuration of the professions, the old European conceptions of profession, guild and vocation were combined with the ideas of progress, order, and freedom.

In contrast to liberal and democratic societies, monarchical, bureaucratic, and statist societies tend to define professionals as civil servants and to establish a specific hierarchy of the professions: at the pinnacle of this hierarchy are professionals in the civil service, who are followed by state-appointed professionals, then by highly qualified members of the 'free professions' and finally by those professionals whose training and practice are not prescribed exactly. In liberal societies, the hierarchy is less pronounced, and the free professions stand at the pinnacle (Siegrist 1996). Since the late nineteenth century, professionals have been increasingly employed in private enterprises such as industrial firms, banks, insurance companies, and research laboratories. The occupational and status groups of professionals who are employees develop their self-understanding with reference to the social types of the civil servant, the free professional, and the entrepreneur.

The professional and social types of the nineteenth century may be characterized as follows:

(a) Professionals who are 'civil servants' are to a great extent bound in the service of the king or the state. If necessary, the individual lawyer, physician, engineer, or professor must disregard professional criteria and follow the directives of his superior. Due, however, to the increasing recognition of professional competence and of the indefinite tenure of higher civil servants, specific ideas and forms of professional autonomy also come into play in the judiciary, in the administration, and in higher education. Professions such as the engineers who were educated in the elite technical colleges in France and then entered into the civil service developed a special esprit de corps (Schweitzer 1999);

(b) Members of 'state-appointed professions' are holders of public office. Professionals such as these enjoy a kind of protected market, but they are also monitored closely. Political appointment strengthens the authority of the expert in society but limits his autonomy, impedes the formation of a self-defined collective consciousness of the profession, and makes the organization of professional associations more difficult. Examples include the Prussian, Bavarian, and Austrian lawyers well into the late nineteenth

century, the French and Italian court lawyers (*avoués*, *procuratori*), the Prussian county physicians, and the community physicians of Italy. In the late nineteenth century, some of these state-appointed professions were able to emancipate themselves from state protection, which had come to be viewed as patronizing. Linking discourses on political freedom and the free market with discourses on professional autonomy, they expressed the desire to become 'truly free professions.' Since clients increasingly made similar demands, many states changed the corresponding laws and regulations in the last third of the nineteenth century;

(c) The 'free professions' are characterized by more elaborate forms of training and examination and distinguished titles. Credentials are derived mostly from the state, but the free professional must be able to adapt to one or more segments of the market for professional services. Minimal standards for both professional and ethical behavior are often regulated by law, at least in principle. In practice, such regulations are enforced either by professional courts of honor (disciplinary bodies with the authority of public law or executive boards of associations with informal authority) or by ordinary courts and clients;

(d) In those areas where professional functions are exercised in the form of free trade, the constraints of the market and the increasing authority of science and scholarship encourage informal processes of professionalization. Professionals are themselves motivated to improve their own qualifications. An elite defines professional qualifications, founds an association, and attempts to marginalize or discredit so-called semi-professional practitioners.

2.3 Cognitive and Emotional Orientations and Strategies

In the eighteenth and nineteenth centuries, the contents, forms, and meanings of professional knowledge changed. As was the case with other societal elites, professionals formulated new scientific discourses and new models for professional conduct. The physician achieved greater recognition as a scientifically and practically trained healer and health expert (Huerkamp 1985; Léonard 1981), while the lawyer came to be viewed as a legal expert and 'priest of justice,' who was responsible for the realization of the rule of law and for securing the rights, the property, and the honor of the citizens (Siegrist 1996). The engineer was the chief proponent of scientific and technical progress, and was long considered to be an expert for administration, supervision, and organization (Schweitzer 1999). The university professor and the teacher in the gymnasium or *lycée* were the experts for education, aesthetics, and morality. Finally, the Protestant minister was the educated theologian, whose legitimacy was no longer derived from faith and piety alone.

The degree of the systematization of knowledge increased over the course of the nineteenth century, but the superiority of the professions and the efficiency of professional conduct remained controversial for an extended period of time. This is evident in the debates over the competence of university-trained 'physicians,' practical 'surgeons,' 'herbalists,' 'quacks,' nuns, midwives, and orderlies, and, more generally, over 'professional knowledge' and the knowledge of 'lay' persons. The professions had enduring difficulties in their efforts to prove their greater efficiency and to legitimize their autonomy with reference to science. Therefore, they attempted to support their claims to professional autonomy with supplementary appeals to values and attitudes that they shared with most of their bourgeois clients and employers. The leading concepts in these efforts were independence, responsibility, impartiality, altruism, public service, acting 'in good faith,' and 'delicacy' in the sense of tact and discretion. These values and attitudes were often codified in laws, professional regulations, codes of ethics, and professional manuals. In many cases, such moral standards applied to the whole person of the professional, i.e., not only to his occupational activities but also to his behavior as a private person and citizen as well.

Compulsion and persuasion served to bridge the gap between professional culture and lay culture. Studies that conceive of professionalization as an element in the historical development of disciplinary processes and the growth of class domination emphasize the compulsive character of the social action of professionals. With the spread of medical, legal, technical, moralizing, and educational practices, professional culture as a combination of knowledge and power penetrated all aspects of modern life and thought (Goldstein 1984, 1987). The professions are said to have used their knowledge in order to secure for themselves social and economic advantages. On the other hand, empirical social and cultural historical studies, in which concrete interactions between experts and lay persons are investigated, often conclude that experts were hardly able to realize their cultural projects in a pure form. For quite some time, lay persons avoided contact with professionals by seeking advice with regard to health from pastors or priests and by seeking legal council from union secretaries.

The historical record shows that the professions that were close to the state or involved with legally regulated cases and procedures were, as a rule, more successful in compelling recognition of their competence. In contrast, clients who made a short visit to the doctor's office, to a lawyer, or to a priest were not exposed to experts to the same degree. On the whole, the free professions had to rely more heavily on persuasion. This explains why symbolic performances, clothing, decoration of work spaces, rituals, and rhetoric are such an essential part of professional culture.

2.4 Professions, Markets, and Economic Situation

The rise of the professions in the nineteenth and twentieth centuries was due largely to the increased and increasingly differentiated demand of private and public customers for goods and services. To a degree, professionals were able to stimulate this demand as well. They fulfilled a continually growing need for the explanation of the world, the formulation and attribution of meanings, for edification, and for entertainment. Through their influence, questions of law and property and of health and death were viewed less fatalistically. Many new opportunities arose for them in the construction trade, commerce, industry, banking and insurance, research, higher education, and also in agriculture. The professions profited from a general increase in prosperity and in the cultural needs of the middle classes. In the twentieth century, there was a rising demand for medical, social, and cultural services on the part of lower social strata, which had previously been neglected. Due to developments in the fields of education and insurance in the welfare state, and because of a general rise in income levels, these strata developed into a new mass market. The changing profile of the clients led to a revision of the traditional attitudes and practices of the professions.

The economic situation of the professions has at no time and in no place been homogeneous. The spectrum ranges from the modest conditions of the lower middle class (*petite bourgeoisie*) to those of the upper middle class (*haute bourgeoisie*). Income and wealth vary according to stages in career development. They depend upon individual abilities and performance, on regulations governing fees and honoraria, on conditions in the various markets, and on general economic cycles. In the nineteenth and early twentieth centuries, it was not uncommon for income to be augmented by supplementary, non-professional activities, by doweries and inheritance, by earnings from real estate, or by stocks and bonds. This was typical of the bourgeois and petit bourgeois milieus to which the professionals belonged through birth, marriage, or social relations. In the twentieth century, the income of professionals was still significantly higher than that of workers or than the national average. Due to increasing competition, to increasing recruitment from families of humbler background, and to economic and political crises, professionals have often become convinced that their heyday is behind them and that they are threatened with decline (Jarausch 1990). Under these circumstances, there is often an intensification of conflict within the professions between protectionists and those favoring a more aggressive market orientation.

Beginning in the late nineteenth century, as the model of the free professions continued to spread, all of the professions became increasingly dependent upon the market. In reaction to this, there were many debates among professionals over the 'overcrowding' of the market and the closure of the various professions. There was a general trend toward lengthening training programs and making them more difficult, in order to stem the influx of young aspirants. The pattern of discrimination against female candidates was widespread. In principle, women had been admitted to higher education since c. 1900, but they were effectively hindered in practicing the classical professions for quite some time (Jacobi 1994). In Central and Eastern Europe, members of religious and ethnic 'minorities,' who had been emancipated in the nineteenth century, still experienced discrimination when they attempted to gain admission to the professions. Due to their supposed 'overrepresentation' in universities and academic occupations, Jews were often made responsible for the worsening of the working conditions and the standard of living among professionals (Jarausch 1990; McClelland et al. 1995). It was not until the second half of the twentieth century that discrimination on the basis of religion, ethnicity, and gender gradually came to be forbidden.

3. Particularism Versus Universalism

Depending on varying perspectives among both contemporaries and historians, the modern history of the professions may be understood in terms of the nationalization of culture and society or in terms of the universalization of knowledge and of the culture of professionalism. In the process of nation-building, members of the professions became elites, which supported the development of national society, politics, science, and culture. The national system provided them with opportunities and assigned rights and duties to them. As functional and cultural elites of the nation, they had a common interest in maintaining the interconnection of power, status, knowledge, and earnings. In both the public arena and in more exclusive spheres, they represented themselves through professional and bourgeois discourses and symbols; these included historical accounts, myths, and *lieux de mémoire* as well as cultural artifacts such as busts, statues, and portraits of famous professionals in public buildings. Professionals have tended to view their own structures, practices, and status as being nationally specific, even in those cases, where, in objective terms, there is a dominant international pattern or style, which is subject only to slight variations and nuances.

Comparative historical research and the history of institutional and cultural transfer has relativized national interpretations and led to significant revisions of national views of history. Such approaches have shown how states and national professions imitate and vary foreign symbol systems and institutions, when they are convinced that this will lead to an improvement in their performance and in their status in international competition (Ramsey 1984; Lundgreen

1990; Siegrist 1996; McClelland et al. 1995). In some cases, occupational patterns that were originally imposed by European empires were maintained after the end of imperial domination. French models for advocates and for engineers were imposed on other lands during the Napoleonic era; but subsequently they were adopted voluntarily, because they were found to be compelling. The Habsburg monarchy influenced the structure of the professions in Southeastern Europe. Prussian prototypes exerted a strong influence in the development of civil service and the professions in northern Central Europe. During the colonial period, the model of the English and French professions diffused worldwide, and they have retained their attractiveness in the postcolonial era as well. The Soviet model of the socialist intelligentsia spread after 1945 to the lands of the Eastern Bloc. In Europe, American patterns have competed with national ones at least since the 1980s.

Since the late nineteenth century, the European or universal character of knowledge and professional culture has often been celebrated at international congresses, but national patterns survived into the twentieth century. In Europe, these have only begun to weaken in recent decades, in the wake of liberalization, the integration of the European market, and globalization. This is a clear indication of the fundamental difference between the professions and other occupations, which were liberalized at an earlier date and more thoroughly.

4. *Historiography and Emphases in Current Research*

Traditional historiography tended to be focused on a single profession. It served primarily to affirm a tradition and to establish an identity, to further the understanding of ones own profession and even to celebrate it, to delimit the profession vis-à-vis other occupations and semi-professions, and to justify the status quo and lend support to demands made in the public arena (Burrage 1990; Siegrist 1996). The oldest genres in the historiography of the professions include works which defend traditions, privileges, and customs in the face of threatening political, social, and legal developments or, alternatively, polemical works by outsiders and civil servants who want to reform the professions. Since the nineteenth century, there has been a proliferation of biographies and memoirs of exemplary professionals, of *festschriften* on the occasion of anniversaries, celebrations, and congresses, of memoranda and surveys with historical aspects, and of learned historical treatises in encyclopedias on general developments and various special aspects of the professions. More ambitious historical monographs on particular professions first appeared in the late nineteenth century. The authors were typically either academic or leading members of the profession in question with an interest in history. Until recently, academic historiography has only showed a moderate interest in the history of the professions. Prior to the 1960s, both amateur and academic historians were more interested in individual professions than in typical aspects of the various professions and rarely took the approaches and theses of the sociology of the professions into account.

Precursors of the systematic historical comparison (and sociology) of the professions include works which, beginning around 1800, compared local and regional occupational ordinances and occupational conditions with foreign equivalents. This sort of comparison and historical reflection was conducted in connection with the preparation of national laws pertaining to professions. Around 1900, representatives of the new discipline of sociology began to take abstract and systematic approaches to individual professions, to groups of professions, and to specific aspects of processes of professionalization. In the efforts of sociologists to discover the 'essence' of the professions or to develop a 'general theory' of the professions, historical aspects and cultural peculiarities were neglected. In the 1960s, research based on the sociology of class and social conflict and pioneer contributions to international comparison reintroduced a historical perspective to the study of the professions (see the historical overviews on the sociology of professions by Freidson 1986 and Burrage 1990). These works inspired the new historical research on the professions, in which the development of the professions is historicized and recontextualized, without, however, sacrificing the goals of systematizing, conceiving of historical types, and contributing to the formulation of a general theory. Today, the history of the professions is an established and dynamic field of social history, cultural history, the history of the sciences, and the history of education. Historical approaches to gender have shown why professionalization has, until very recently, remained a male phenomenon and how professionalization and gender have entered into specific kinds of symbiotic relationships.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Civil Service; Ethical Codes, Professional: Engineering Codes; Expertise, Acquisition of; Legal Professionalism; Medical Profession, The; Professions in Organizations; Professions, Sociology of

Bibliography

Betri M L, Pastore A (eds.) 1997 *Avvocati, Medici, Ingegneri. Alle Origini Delle Professioni Moderne* (Secoli XVI–XIX). Clueb, Bologna, Italy

Burrage M, Torstendahl R (eds.) 1990 *Professions in Theory and History*. Sage, London

Burrage M 1990 The professions in sociology and history. In:

Burrage M, Torstendahl R (eds.) *Professions in Theory and History*. Sage, London
Charle C 1996 *Les Intellectuels en Europe au 19e Siècle*. Éditions du Seuil, Paris
Conze W, Kocka J (eds.) 1985 *Bildungsbürgertum im 19. Jahrhundert* pt. I: Bildungssystem und Professionalisierung in internationalen Vergleichen. Klett-Cotta, Stuttgart, Germany
Freidson E 1986 *Professional Powers. A Study of the Institutionalization of Formal Knowledge*. University of Chicago Press, Chicago
Goldstein J 1984 Foucault among the sociologists. The 'disciplines' and the history of the professions. *History and Theory* **23**: 170–92
Goldstein J 1987 *Console and Classify. The French Psychiatric Profession in the Nineteenth Century*. Cambridge University Press, Cambridge, UK
Huerkamp C 1985 *Der Aufstieg der Ärzte im 19. Jahrhundert. Vom gelehrten Stand zum professionellen Experten*. Vandenhoeck & Ruprecht, Göttingen, Germany
Jacobi J (ed.) 1994 *Frauen zwischen Familie und Schule. Professionalisierungsstrategien bürgerlicher Frauen im internationalen Vergleich*. Böhlau, Köln, Germany
Jarausch K H 1990 *The Unfree Professions. German Lawyers, Teachers and Engineers, 1990–1950*. Oxford University Press, New York
Léonard J 1981 *La Médecine entre les Savoirs et les Pouvoirs. Histoire Intellectuelle et Politique de la Médecine Française au 19e Siècle*. Aubier, Paris
Lundgreen P 1990 Engineering education in Europa and the USA 1750–1930. The rise to dominance of school culture in the engineering profession. *Annals of Science* **47**: 33–75
Malatesta M 1995 *Society and the Professions in Italy 1860–1914*. Cambridge University Press, Cambridge, UK
McClelland C, Merl S, Siegrist H (eds.) 1995 *Professionen im modernen Osteuropa/Professions in Modern Eastern Europe*. Duncker & Humblot, Berlin, Germany
Perkin H 1989 *The Rise of Professional Society in England Since 1880*. Routledge, London
Prest W R 1986 *The Rise of the Barristers. A Social History of the English Bar 1590–1640*. Clarendon Press, Oxford, UK
Ramsey M 1984 The politics of professional monopoly in nineteenth-century medicine. The French case and its rivals. In: Geison G L (ed.) *Professions and the French State 1700–1900*. Pennsylvania University Press, Philadelphia
Schweitzer S 1999 Der Ingenieur. In: Frevert U, Haupt H-G (eds.) *Der Mensch des 19. Jahrhunderts*. Campus, Frankfurt am Main, Germany
Siegrist H (ed.) 1988 *Bürgerliche Berufe. Zur Sozialgeschichte der freien und akademischen Berufe im internationalen Vergleich*. Vandenhoeck and Ruprecht, Göttingen, Germany
Siegrist H 1996 *Advokat, Bürger und Staat. Sozialgeschichte der Rechtsanwälte in Deutschland, Italien und der Schweiz (18.–20. Jh.)*. Klostermann, Frankfurt am Main, Germany

H. Siegrist

Professions in Organizations

The study of professions in organizations is about the intersection of two areas of enquiry that have quite distinct histories, namely, the study of professions and the study of organizations, but that converge over the course of those histories. There have been times when they have been related strongly and other times when they have gone their distinct ways. In fact, the study of professions in organizations has gone through three, overlapping, phases since the mid-twentieth century. The first phase, started in the 1950s and particularly strong in the 1960s, and still apparent today, was that of a critical examination of the role of professionals in bureaucratic organizations. The second phase, begun in the 1970s and again, continuing until today, critically examined the orthodox view of professionals and gave a more power-based, labor market analysis of their position and how it affected organizational labor markets. The third phase, which has its origins in the 1960s and 1970s, and blossomed in the 1980s and 1990s, emphasized the dynamic nature of professionals in organizations and acknowledged the variety of organizational locations. This article outlines these phases, dealing with the central concepts and findings, examining the questions that each one raises, and pointing the way to the next phase of research in this rich field.

1. Professionals and Bureaucracy—1950s Onwards

The origins of the initial discussion of the relationship between professionals and organizations lies in the sociology of the professions and the discussions of defining characteristics of a profession. There are a number of contributions, but one that is often seen as central is that of Greenwood (1957) who suggested that professions have: a basis in systematic theory; prolonged training and certification; recognized authority; community sanction and legitimation; professional culture and codes of ethics. In essence, he came up with a set of traits or components that, when present in their entirety, defined the complete profession. This trait approach was echoed in a variety of ways by Wilensky (1964) and Hickson and Thomas (1969). There were two thrusts to these discussions. The first was to establish the characteristics that make up a profession. The second was to determine empirically which occupations were 'truly' professional and which were 'partly' professional and which 'aspired' to the status of a profession.

Out of these discussions came a relatively agreed upon set of characteristics of a profession that had structural and value aspects. The structural aspects were such things as a full-time occupation, a training and certification system, a professional association, and an ethical code. But, more importantly for the study of professionals and organizations, was the extension of Greenwood's notion of a professional culture into a set of values which encompassed peer reference, vocation, public service, self-regulation, and autonomy. When this work in the sociology of the professions connected with the sociology of organi-

zations it was in terms of the relationship between the values held by professionals and those promulgated by the modern, bureaucratic organization. This relationship was seen primarily as antagonistic (cf. Hall 1968).

Through the 1950s and 1960s the great influence on the sociological study of organizations was Weber (1949). Weber's principal contribution to this field was his characterization of organizations in terms of the authority relations within them (systems of imperative coordination) and the ways in which such systems had evolved historically. Each ideal type of authority, charismatic, traditional, and rational–legal, had its associated organizational form. Weber saw the rational–legal authority system with its organizational form of bureaucracy as the dominant institution of modern society. The authority system is rational because means are designed expressly to achieve certain goals and it is legal because authority is exercised through an office with its associated rules and procedures. For Weber the bureaucratic organization was technically the most efficient form of organization possible.

In terms of Weber's evolutionary history of social systems, bureaucracies represent the final stage in depersonalization. Offices (jobs) are held by experts who are arranged in a hierarchy. Rules and procedures provide predictability and consistency. Information is recorded and stored. Personal and business affairs are separated. There is 'the methodical attainment of a definitely given and practical end by means of an increasingly precise calculation of means.' Starting in the economic sphere, bureaucracy is such a potent method of organizing that it becomes characteristic of all areas of society such as education, government, politics, religion, and so on. Much of organizational analysis of the 1950s, 1960s and 1970s was essentially concerned with establishing, extending, and critiquing Weber's notion of bureaucracy (e.g., Pugh et al. 1969)

As Abbott (1988) in his influential work has pointed out, this 1950s and 1960s view of professions had an underlying assumption of professionalization as a natural process; that is, a regular sequence by which an occupation moved to 'full' professionalism (Wilensky 1964). Essentially, this view, together with 'the search for bureaucracy' produced a unitary view of both professional occupations and organizations. Professionals, it was suggested, were socialized into occupations where the key values were autonomy, peer control, and vocation. For example, Hall (1968) operationalized professional values as the professional organization as a reference, a belief in service to the public, a belief in self-regulation, a sense of calling to the field, and feelings of autonomy. Bureaucracy was operationalized as hierarchy of authority, division of labor, rules, procedures, impersonality, and technical competence. However, his conclusion was 'an assumption of inherent conflict between the professional or the professional group and the employing organization appears to be unwarranted.'

Of course, both sides of this coin, the professional and the organization, could be, and were opened up to scrutiny. On the organizational side of the coin, Scott (1965) had put forward the notion of Autonomous and Heteronomous Professional Organizations. Pugh et al. (1969) distinguished between Full Bureaucracies, Workflow Bureaucracies, and Personnel Bureaucracies as well as Implicitly Structured organizations. The point of these taxonomies and typologies was to open up the possibility of there being organizational forms in which the values espoused by individual professionals would be embraced and protected, thus potentially transforming the nature of the debate beyond an inherent conflict.

On the other side of the coin, the changing nature of professions was being stressed together with their organizational locations. Two issues are important. First, there was a trend of increasing professionalism within bureaucratic organizations. Occupations were arising that claimed professional status, such as social workers, nurses, and managers. These were bureaucratically produced and located occupations. Second, Hastings and Hinings (1970) suggested, from their work on chartered accountants, that there were established professions that carried out tasks that were centrally located in bureaucratic organizations and that, as a result, there was little or no conflict between professional and bureaucratic values in these situations.

This 'opening up' of both organization and profession led to new developments in understanding professionals in organizations. However, this theme of potential conflict between certain forms of organization and professionalism continues today. For example Raelin (1991) has examined the management of professionals in organizational contexts, writing about 'the clash of cultures' when generic managers are responsible for professionals. In the UK, and Europe more generally, there has been a concern with 'The New Public Management' and its organizational impact on professionals employed in healthcare, government at all levels, and education. The concern or thrust is still that there are incompatibilities between professional ways of working and values and certain kinds of organizational principles and practices. It continues to be a fruitful area of study.

2. *Professionals in the Labor Market—1970s Onward*

The second phase, begun in the 1970s and again, continuing until today, critically examined the orthodox view of professionals and gave a more power-based, labor market analysis of their position and how it affected organizational labor markets. The basis for this was laid down in critiques of the 'professionals in organizations' approach in the previous decades. The previous research questions had been anchored firmly

in issues about the structure and values of professionalism with notions of an underlying process of professionalization arising from the nature of work in modern economies.

The problems that were raised continued the opening up of professional occupations away from a homogeneous view of them. The uneven nature of professionals was stressed; that is, they could not and should not be considered as a unitary set of occupations. Similarly, the rise of the bureaucratic, organizationally based professions both in the context of the state and the private sector was recognized more fully. Abbott (1988) calls this the 'monopoly school' which attributes developments in professionalization 'to a desire for dominance or authority' (p. 15). The occupational/professional project is to control the conditions of work (the labor process and the labor market) and to increase and maintain occupational status and power (Johnson 1972, Larson 1977, Friedson 1986).

The issues raised by this approach 'dislocate' the relationship between professionals and organizations both because it makes our understanding of professional occupations more complex, and also because it shifts the locus of debate away from values and structure to power and control. Indeed, the connection of this critique of the functions of professions is not as strongly related to the theme of professionals in organizations as the theories being critiqued. The emphasis becomes recentered on the nature of a profession and the professional project. In particular, the claim to professional status by an occupational group, for example, social workers, is seen as a way to control entry to the occupation and the demand for membership, and to establish an authoritative position within the labor market.

In a sense, the connection to organizations is primarily through an analysis of the rise of new professions and a continuation of the theme of 'the professionalization of everyone' (Wilensky 1964). Because a key question to be answered is why new occupations make the claim to be professional, that is, the professional project, the examination is, in part, one of organizationally based occupations. Thus, the issue of professionals in organizations is not one of 'traditional,' autonomous professionals coming to terms with the rise of the bureaucratic organization, but the establishment of functionally and bureaucratically derived new occupations as professional. The issue for the new professional in this organizational setting is to acquire authority and control the conditions of their occupation's existence.

Thus, the professional/organizational relationship is not one of adaptation between two different sets of structure and values, but of an occupational group that is functionally defined and organizationally embedded acquiring the legitimacy and dominance that comes from professional status. The professional/organizational relationship is one of power plays, control of the labor process and fights over legitimacy that are played out both in the organizational setting and the institutional/regulatory setting within a particular society. This approach continues to be part of the discussion of professions in organizations and is to be particularly found in discussions of the healthcare sector (Friedson 1986, Larson 1977) and of the new public management. One of the issues that is raised by this stream of theorizing is that of the 'semiprofessions.' This was an attempt to distinguish between fully professionalized occupations and other occupations that aspired to professional status. These latter occupations tend to be primarily those that are organizationally produced and located and where the individual occupation members do not work as independent practitioners. Johnson (1972) sees this distinction (and the use of other descriptors such as 'quasi,' 'marginal,' or 'limited' professions) as an illustration of power relations between occupational groups. For him, professions retain positions of authority vis-à-vis the semiprofessions, whose claim to professional status is contested. As he puts it, 'the pervasiveness of the ideology (of professionalism) is not indicative of the extent of professionalism (p. 59).

The theoretical and research issue of competition between groups for professional status continues to be an important topic. Several writers have observed the phenomenon of stratification within the professions. Abbott (1988) observes that interprofessional competition is a fundamental fact of professional life with jurisdictional disputes between occupational groups over claims to particular expertise and knowledge bases. And with more and more occupational groups that lay claim to such expertise and knowledge being strongly based in organizations, then the dynamic between professionals and organizations changes, on both sides.

3. *Professionals in Professional Organizations—the 1980s Onwards*

The third phase, which has its origins in the 1960s and 1970s, and blossomed in the 1980s and 1990s, emphasized the dynamic nature of professionals in organizations and acknowledged the variety of both professional and organizational locations. The issues of the variety in professional settings had been increasingly raised and, at the level of the professional system, Abbott (1988) has pointed out that there is continuous competition and jurisdictional disputes between professional groups as they dispute claims to expertise. Similarly, on the organizational side, the typologies and taxonomies of organizations both differentiated between types of bureaucracy and introduced new organizational forms.

This discussion has emerged in two different forms. The first is setting the discussion within the changes occurring in national and international economies. As

thinking about 'the new economy' and 'globalization' has developed (Aharoni and Nachum 2000), so two issues have been emphasized which bring together much of the previous theorizing about professions. On the one hand there is the rise of new occupations which claim professional status. For example, in the software development industry there are computer engineers, software engineers, and systems analysts who are highly educated but operate in paraprofessional environments across boundaries of time and space. Financial services have grown rapidly, as has the whole sector of business services, and along with this have come investment analysts, financial planners, and mergers and acquisitions specialists whose professional status is unclear although one sees the same processes as were written about in the 1960s and 1970s, namely those of educational qualifications, professional associations, examinations, and designations and the ideological claim to professional status.

On the other hand there is the rise of the 'knowledge-based organization' (Alvesson 1995) together with the development of concepts such as the flexible organization and the learning organization. The suggestion is that organizations of the new economy are increasingly different from the various kinds of bureaucracy of the old economy. One of the reasons for this is that they are more service-based and that those services require highly educated employees who, in value terms, are much more like the 'old' professions, being committed to autonomy, vocation, and collegiality. These new organizational forms are seen as much more friendly and hospitable to professionals. Also, it is suggested that these forms are spreading into old economy organizations as they respond to the rigors of global competition.

The second form of the discussion is a renewed interest in the professional service firm, for two different reasons, each connected to new organizations and new occupations. Peters (1992) and Alvesson (1995) have suggested that the professional service firm is an example of the flexible, knowledge-based organization; that there is a wealth of experience in these organizations from which the newly developing knowledge-based firms can learn. But, at the same time, with the explosion of knowledge-based services, and, in particular, the tremendous expansion of business services in OECD economies, there is much more competition over professional jurisdictions. If one takes the example of the 'Big 5' accounting firms (formerly the Big 8 and the Big 10), there has been a continuous expansion of services within these firms and more and more integration across services and across nations (Greenwood et al. 1998).

The same processes are at work within other professional jurisdictions and professional service firms. In law, for example, there are strong internationalization pressures. Also, there is much discussion over the multidisciplinary practice, allowing non-lawyers to have a controlling interest in law firms. In 2000, the Law Society of New South Wales (calling itself 'the most progressive legal jurisdiction in the world') voted to allow multidisciplinary practices in law firms and also to allow them to incorporate.

Part of the process of defining these issues and also a result of them being articulated has been a rise in interest in the professional service firm (PSF). There were classic studies by Smigel (1964) and Montagna (1968) that examined the ways in which professionals operated in the professional service firm. But because of the concerns of theorists with issues of the rise of the bureaucratic organizational form and the response of professionals to that, and the concern with the monopoly critique, there was little or no interest in studying the professional service firm *per se*, even though it was the implicit, professionally hospitable setting contrasted with the bureaucratic form. It is this which has become the focus of much interest and research in the past two decades and in its own way it reflects issues of professionalization and bureaucratization, although the latter has been recast as managerialism and located in the context of knowledge-based organizations and globalization.

The theoretical and empirical location of studies of professional organization has been, primarily, in either the professional association and system or nonprofessional organizations. Very little study had been done of professionals in professional organizations, organizations that can be described as PSF. One important dimension in understanding professions in organizations is the extent to which they are primarily located in professionally dominated organizations or in other kinds of organizations, the so-called bureaucratic professions. For example, practicing lawyers are employed primarily in law firms that are organized as professional partnerships; only a small proportion are employed outside legal practices; for example, as corporate counsel or as lawyers within the public defender justice system. Accountants tend to be split about 50/50 between private practice and employment in private and public sector organizations. However, almost all accountants are trained in private practice. Engineers are employed overwhelmingly in private and public sector organizations with only a small proportion of the profession being employed in engineering consulting firms. And engineers receive their postuniversity professional certification training in a wide variety of settings.

There are three points to be taken from this. First, professionals work in a variety of organizational settings, something that has been ignored in much previous work. Second, there are some professionals who primarily work in bureaucratic settings and experience little or no conflict. Hastings and Hinings (1970) showed that chartered accountants working in industrial settings experienced little role conflict and they argued that part of the activity of accountants is to bureaucratize others. Similarly, the debate over the semiprofessionals such as social workers, teachers,

and nurses centered around the fact that they were located entirely in bureaucratic organizations and did not have independent, autonomous practices. But, third, the arena of professional work that had been ignored in particular was that of the professional service firm in spite of the fact that implicit in many analyses of professionals, their location in independent professional organizational settings was considered to be very important as the archetypal setting for an autonomous professional practitioner.

Much of the work of the 1980s and 1990s has been concerned with understanding the nature of the professionally based, private practice organization (Nelson 1988, Brock et al. 1999). As this work developed it also embraced the theme of bureaucratization although in a different guise, as managerialism.

This initial work was concerned with elaborating the nature of the traditional professional service firm, which Greenwood et al. (1990) typified as the P^2 form. The P^2 stands for the two key elements of professionalism and partnership. The PSF is an organization whose core workers are professionals. Traditionally they are from one dominant profession such as accounting or law. And the governance system of the organization is a partnership where the senior members of the professional group jointly and severally take responsibility for the direction of the firm.

There are a number of organizational consequences to these twin defining features. In human resource terms, one is the 'up or out' system or tournament. Young professionals either make it to partner or leave the firm. Organizationally, they are relatively loose structures, representative and collective decision-making with management being something that is done after dealing with clients. Issues of management and methods of organizing are not issues that are seen as important to discuss in the P^2 firm. The emphasis is on the individual professional and their ability to do good professional work. There are clearly strong echoes in this of the kinds of discussions that have arisen in the 1980s and 1990s about the knowledge-based firm with its employment of highly qualified individuals who are applying new, and sometimes esoteric knowledge, and, it is suggested, require freedom and flexibility to achieve results. In the P^2 organization, interesting work and the freedom and flexibility to do it are prized.

However, the nature of the professional firm has been changing due to changes in the institutional and market contexts of professional work. Because of the development of the new economy and the rise of services generally and business services in particular, there have been increasing pressures on the established professions such as law, accounting, and medicine. For more than a decade now, Western economies have been in a period of deregulation with the ideology of the market being dominant. Thus, jurisdictional boundaries have blurred and the level of *inter*-professional competition has increased. The rapid rise in the numbers of management consultants, financial services advisors, software development engineers, information technology consultants, media and marketing occupations, etc. has changed the dynamics of professional organizations. Similarly, there have been substantial changes in markets. On the one hand the demand for these professional business services has increased markedly, making it one of the most rapidly growing sectors in many economies. But the demands for these services are from increasingly sophisticated clients, operating on a global scale.

It has been suggested that a major consequence of these pressures and changes is that the professional, P^2 organizational form is in decline (Brock et al. 1999), and that it is being replaced by forms that place a greater emphasis on managerialism (Cooper et al. 1996). Cooper et al. (1996) named this new organizational form in professional firms the Managed Professional Business (MPB). In such an organization professionalism becomes centered on developing lines of business, hierarchies develop, strategic planning is introduced, targets are more binding, and decision-making rests in the hands of management teams rather than the partnership at large. So, the argument goes, the terms and conditions of work in the MPB move towards those of the regular corporation. However, this can only go so far because of the knowledge-based work of professionals (Rose and Hinings 1999).

What this suggests is convergence for professionals in organizations, but of a new kind. It is not convergence around either professionalism or managerialism (bureaucratization), *per se*, but around a combination of these which is organized around the increasing role of knowledge in 'postindustrial,' service-based organizations. As yet, we have not sufficiently tracked these kinds of developments to write authoritatively about them.

4. *Conclusion*

So what can we conclude about research in this rich field and the status of our current understanding? There are a number of things. A very important one is the continuing theme of the potential conflict between professionalism and bureaucratization and/or managerialism. That is, the notion of a professional encompasses values and cultural statements that are inimical to the forms and cultures of hierarchically organized, profit-driven corporations. However, this continuing theme has been modified in two important ways, both of which require much more study.

The first is the opening up of the idea of the professional. There is still a tendency to work with ideas of 'true,' 'full,' or 'traditional' professions that are seen as archetypal or at least worked with as ideal types. But, much of the original work on the professional–bureacratic relationship, and, in particu-

lar, the ideas of the monopoly school, suggested that there are a range of professions exhibiting varying degrees of compatibility with bureaucratic working. More research is needed which starts from the *diversity* between professionals, rather than the assumption of homogeneity. There are still incompatibilities between some professional ways of working and their associated values and some organizational principles and practices. But whether those professions that have evolved from bureaucratic organizations have more compatibility with those forms is still very much an open question, and it needs to be connected to issues of claims to professional status and competition between occupational groups for professional status.

The issue that Abbott (1988) sees as central, namely, interprofessional competition and jurisdictional disputes between occupational groups, should be related to the types of organizations within and between which such competition and disputes take place. The healthcare setting is a particularly good one for examining these kinds of issues because of the numbers of professional groups involved.

A second modification to the professional–bureaucratic opposition thesis lies in the variety of organizational settings and their changing nature. While the death of bureaucracy is much over-stated, there is no doubt that new organizational forms are emerging that emphasize knowledge acquisition, use, and dissemination, together with flexibility and autonomy in organizational working. We might argue that the organic, flexible design is the 'new' organizational form. But there is a range of established and emerging forms within which there are differing responses to expertise and professionalism.

A further development of these themes comes from our increased understanding of the role and nature of the professional service firm. Within the 1990s more and more interest and research has been centered on, *inter alia*, accounting firms, law firms, management consulting firms, and engineering consulting firms. These are, of course, the archetypal sites for the study of professionals within settings which should be the most accommodating because they are organized around professional work and managed by the professionals themselves. But they are also operating in changing institutional and market environments and face intra- and interprofessional competition. More research needs to be done on the way in which these professional fields are being restructured, new occupational grouping emerging, new organizational forms developing, and the organizational–professional relationship being changed. Part of this is also that there are important differences *within* as well as *between* professionals in terms of organizational setting.

Finally, an important and intriguing topic to be examined is the possible convergence of professionals in organizations and organizations of professionals. That is, on the one hand, many 'nonprofessional' organizations are becoming more knowledge-based and are experimenting with more flexible and autonomy-based organizational forms. On the other hand, professional organizations are moving into new markets, facing more competition, changing their knowledge base, and growing in size. In doing this they are becoming more managed, more strategic, and possibly decreasing some aspects of autonomy and flexibility. What they can offer is challenging work with major clients. This suggests a new kind of convergence, combining professionalism and managerialism (bureaucratization), due to the increasing role of knowledge in 'postindustrial,' service-based new economies. These changes present a real challenge for research.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Organization: Overview; Organizational Culture; Organizational Culture, Anthropology of; Organizations, Sociology of; Professionalization/Professions in History; Professions, Sociology of; Weber, Max (1864–1920)

Bibliography

Abbott A 1988 *The System of Professions*. University of Chicago Press, Chicago

Aharoni Y, Nachum L 2000 *Globalization of Services: Some Implications for Theory and Practice*. Routledge, London

Alvesson M 1995 *Management of Knowledge-intensive Companies*. De Gruyter, New York

Brock D, Powell M, Hinings C R (eds.) 1999 *Restructuring the Professional Organization: Accounting, Healthcare and Law*. Routledge, London

Cooper D J, Hinings C R, Greenwood R, Brown J L 1996 Sedimentation and transformation in organizational change: The case of Canadian law firms. *Organization Studies* **17**: 623–47

Friedson E 1986 *Professional Powers*. University of Chicago Press, Chicago

Greenwood E 1957 Attributes of a profession. *Social Work* **2**: 44–55

Greenwood R, Hinings C R, Brown J L 1990 The P2-Form of strategic management: Corporate practices in the professional partnership. *Academy of Management Journal* **33**: 725–55

Greenwood R, Rose T, Cooper D J, Hinings C R, Brown J L 1998 The global management of professional services: The example of accounting. In: Clegg S, Ibarra E, Bueno L (eds.) *Theories of Management Process: Making Sense Through Difference*. Sage, Thousand Oaks, CA

Hall R H 1968 Professionalization and bureaucratization. *American Sociological Review* **33**: 92–104

Hastings A, Hinings C R 1970 Role relations and value adaptation: A study of the professional accountant in industry. *Sociology* **4**: 353–66

Hickson D J, Thomas M 1969 Professionalization in Britain. *Sociology* **3**: 37–54

Johnson T 1967 *Professions and Power*. Macmillan, London

Larson M S 1977 *The Rise of Professionalism*. University of California Press, Berkeley, CA

Montagna P 1968 Professionalization and bureaucratization in large professional organizations. *American Journal of Sociology* **74**: 138–45
Nelson R 1988 *Partners with Power: The Social Transformation of the Large Law Firm*. University of California Press, Berkeley, CA
Peters T 1992 *Liberation Management*. Macmillan, London
Pugh D S, Hickson D J, Hinings C R 1969 An empirical taxonomy of structures of work organizations. *Administrative Science Quarterly* **14**: 115–26
Raelin J 1991 *The Clash of Cultures: Managers Managing Professionals*. Harvard Business School Press, Boston
Rose T, Hinings C R 1999 Global clients' demands driving change in global business advisory firms. In: Brock D, Powell M, Hinings C R (eds.) *Restructuring the Professional Organization: Accounting, Healthcare and Law*. Routledge, London
Scott W R 1965 Reactions to supervision in a heteronomous professional organization. *Administrative Science Quarterly* **10**: 65–81
Smigel E O 1964 *The Wall Street Lawyer*. Free Press, New York
Weber M 1949 *Theory of Economic and Social Organization* (trans. Henderson A). Free Press, Glencoe, IL
Wilensky H 1964 The professionalization of everyone? *American Journal of Sociology* **70**: 137–58

C. R. Hinings

Professions, Sociology of

The sociology of professions is a branch of the sociology of work concerned with the analysis of expert occupations. It analyzes their patterns of organization, types of work, and social status. Because the sociology of professions first arose in the UK and the USA, it took its name and much of its early conceptual apparatus from the pattern of expert work characteristic of those countries. In recent years, however, the field has become more international, a change that is transforming both its empirical base and its theoretical superstructure.

1. *Expert Occupations and the Sociology of Professions*

The sociology of professions first arose to study the particular social form taken on by expert occupations in England and the USA in the course of the nineteenth century. This form was characterized by a combination of independent individual practice with some type of collective association. It usually involved specialized education, examination or licensure, service fees, and some form of autonomous discipline. By the early twentieth century, such a form was characteristic of medicine, law, accounting, architecture, and various other fields in the Anglophone world.

The first systematic analysis of these 'professions' was by Carr-Saunders and Wilson in 1933, who set forth a typological analysis of occupations resting on various 'traits' that were taken to characterize the professions. The Carr-Saunders and Wilson analysis had a hard time with the military and the clergy, which were certainly 'professions' in the English sense of the term. Moreover, it made no effective contact with the quite different organization of expert work on the continent.

The French, for example, have no word with the denotation of the Anglo-American 'profession.' The French do distinguish *professions libérales*, but these include only experts in free individual practice, not all experts of a certain type. *Profession* itself has much broader purview than the English 'profession,' as does the parallel *métier*. Moreover, the organization of French experts has historically been far more state-dependent than has that of Anglo-American ones. Nor was France the only problem. The word and concept of profession were no closer to the German culture of expertism than to the French. Indeed, the characteristic pattern for organizing expert workers on the continent has more often been through quasi-civil service arrangements than through autonomous 'professional' ones. The sociology of professions thus began as a study of a particular kind of expert occupation. Only gradually have historical and comparative study made the limiting character of that approach clear.

2. *Historical Development of the Field*

From the Carr-Saunders book of the 1930s up into the 1970s, typology remained the dominant strategy of the sociology of professions. Millerson (1964) presented a comprehensive summary of all the various qualities that had been used to define 'professionhood.' The typological approach eventually strangled inquiry, as investigators worried whether or not various groups really were professions. Indeed, a substantial literature grew up in this period on the 'semiprofessions,' usually felt to include teaching, librarianship, nursing, and social work. These fit the typological model in some ways (advanced education, licensing) but missed it in others (independent practice). The so-called semiprofessions were all largely female occupations, a fact that would later become more emphasized, but engineering, although largely male, shared many of the same characteristics. Within typological thinking, salaried employment of professionals presented a problem. A small literature in the 1950s and 1960s considered this problem under the heading of 'professionals in bureaucracies,' viewing such professionals as being under strong role strain because of the conflicting demands of profession and organization.

The typological model enabled the sociology of professions to come to a rapprochement with structural functionalism. In a celebrated essay, Parsons (1954) argued that professions were rationalist (as

opposed to traditional), that they exercised functionally specific authority (as opposed to undifferentiated authority), and that they were universalist (as opposed to particularist). The later Parsonian tradition saw the professions as squarely anchoring the pattern maintenance sectors of crucial social subsystems. The typologists' focus on the organization rather than the content of professional work accorded better with the early Parsons, however, which kept them apart from the later structural functional interest in the professions.

Throughout the early postwar period, students of Everett Hughes and others in the Chicago tradition continued to amass case studies of professions and semiprofessions. Many of these fit the typological model badly. Their dependent nature often mocked the high role structural functionalists assigned to professions. One way around this problem was to argue that professionalization was a characteristic process that many occupations had not yet fully traversed. In the 1960s and 1970s, sociologists of professions argued for a characteristic professionalizing process, a sequence of stages through which all expert occupations developed. (Wilensky in 1964 provided the most famous analysis of such an argument.)

In the later 1960s, the sociology of professions, like many other subfields, turned towards skeptical interpretations. Major reinterpretations came from Johnson (1967), Freidson (1970) and Larson (1977). This new 'professional power' interpretation saw the professions in a different light. Freidson brought a generation of detailed studies of medical workplaces and medical knowledge together with a theory of professional organization to argue that the structures of professional organization were hardly the direct outcomes of rational and functionally specific universalism. The power of doctors over nurses suddenly seemed less a matter of advanced training than an enforced local division of labor. At a much more general level, Johnson recast professionalism as one among a number of modes of occupational control. This argument was carried further by Larson's reinterpretation of professions as collective mobility projects in which expert groups sought rewards through control of certain markets for services.

The rise of the power school coincided with a wide new interest in professions among historians, particularly in the USA. Scattered works had previously considered some major professions—English law and US medicine, for example. But the 1970s and early 1980s brought a flood of works covering various professions in the USA, England, and to a lesser extent France and Germany. These various works covered periods from the early twentieth century as far back as the seventeenth and eighteenth centuries. Although there were (and remain) important lacunae, theorists of professions began to work with a new empirical base. Perhaps more important, the historians turned to the sociological literature for conceptual guidance and in many cases found little of use to them.

A number of other trends helped recast study of professions in the 1970s and 1980s. Most important was the steady drift of professionals away from independent practice and into corporate settings. On the one hand, this was simply a continuation of the trend towards larger and larger professional firms. In law, accounting, and architecture, sizable firms dated from the first half of the twentieth century, at least in the USA. Legal restrictions generally prevented this development both in the UK and on the continent. On the other hand was the development of multiprofessional firms, a pattern that had begun in architecture and now spread into areas like accounting, finance, and consulting. These largely professional firms were complemented by a much larger body of salaried professionals scattered throughout commercial and nonprofit organizations in the USA. By 1980, few US professions indeed were more than 50 percent made up of independent practitioners.

The rise of salaried professional work in the USA accomplished three things. First, it brought the US professional world much closer to the European pattern. The pattern of large-scale salaried employment by professionals had fascinated the French particularly, who had developed, in the concept of *cadres*, the notion of a technocratic intelligentsia as a new mode of occupational structure (see Boltanski 1982). Second, the new salaried professional work reawakened the interest in the 'professionals in bureaucracies' problem of role conflict. Third, the new pattern led to new concepts of 'deprofessionalization' and 'proletarianization of professionals.' Although the latter term obviously derived from the attempt to develop a Marxist analysis of professions, both concepts attempted to come to terms with the notion—an absurdity under the 'professionalization' concept—that occupations could lose the perquisites and possibly even the form of professions once these had been gained. A wide variety of studies appeared in the 1970s and 1980s examining the degree to which professions and professionals were losing control of work and cultural authority as well as enduring more mundane changes in salaries and tenure.

By the mid-1980s, then, the sociology of professions looked quite different. It had lost nearly all touch with structural functionalism. There was confusion over which professions would develop in which directions at which times. A large new body of research opened to theorizing not only a wide variety of quasi-professional occupations in the USA and England, but also the expert occupations of the continent. In many cases, this new data had extraordinary historical depth (see, e.g., the comparative work in Cocks and Jarausch 1990). In 1988, Abbott attempted to synthesize the entire area by refocusing the sociology of professions on actual professional work (both cultural and practical), turning away from the strong emphasis

on the social structures of professionalism that had become the center of the literature. He envisioned a world of interprofessional competition over jurisdiction of work. Changes flowed into this world both from within professions and from beyond them, recasting some professions, privileging others, destroying yet others. Professions attempted to claim control of work in the workplace, before the public, and within the state, and the course of professions' histories was determined by the ensemble of this melee of interaction. Abbott explicitly attacked the still-continuing typological tradition for eliding large areas of work. For him a profession was anything that competed like one.

Although Larson and others had attempted some international comparisons, Abbott's heavy reliance on European cases, particularly French ones, presaged what became the major development in studies of professions in the 1990s. This was the trend toward examining the relation of professions and the state. Much of the most interesting work came from Europe, for example, Jarausch's (1990) comparative analysis of lawyers, teachers, and engineers in Germany in the first half of the twentieth century. Krause (1994) has made the most general attempt at analyzing relations between the state and professions (and also capital), studying the USA, the UK, France, Germany, and Italy by means of case studies of law, medicine, engineering, and academics in each country. Krause's analysis focuses on the relative decline of professional autonomy and power in all these societies in the years after 1970. Both state and capital have increased their authority over professions.

The relation of state and professions has been underscored as an empirical question by the vast complications of creating standardized professional credentials in the European Union. The future of the sociology of professions will be greatly shaped by the theoretical developments necessary to analyse this process of internationalization of professional life. There is also a 'relations with capital' side to internationalization, encompassed in the largely unstudied complexities of the transnational corporations' use of professional services. There has been some study of 'offshoring' of professional services (e.g., of engineering services in telecommunications), but this particular capitalist attack on professional organization has not been studied in detail. In general, there is a startlingly small amount of work on expert occupations in the third world, with the possible exceptions of lawyers and, of course, the military.

3. *Current Problems*

The central present issue for the sociology of professions, then, is the question of how the professions will be reshaped by the economic reorganization that arose out of the crisis of the 1970s. Many of the recent transformations have long antecedents, particularly the drift to salaried employment with its questioning of professions' cultural authority. Other trends, however, are new or newly important because of the economic conjuncture. Internationalization has been crucial. By the late 1980s the restrictions on professional firm size began to fall in Europe, largely under the onslaught of US firms, particularly law firms and multiprofessional consulting firms, which were the only firms able to manage some of the huge legal and financial projects characteristic of the new international economy. The new gigantism has transformed relations both within the professions, and, more important, between the professions and the state. It is in the international context, too, that the question of the profession–state relation is most clearly posed.

But there have been important internal effects to rethink as well. The decline of medical power from the pinnacle portrayed by Freidson has given new meaning to the phrase 'proletarianization of professionals.' This decline, driven by reorganizations of insurers and regulators, requires revising general theory about how expertise is organized in society. In that sense, the sociology of professions is again confronted with a functionalist question: how is expertise to be produced, controlled and reproduced in modern societies?

The issue of state and capitalist control is only one side of this question. On the other is the issue of commodification of professional knowledge. It is at present much too early to know the effects of computer-aided design systems, medical diagnosis systems, and other expert systems on the future of professions. But it is clear that new theory is required. It is also clear that commodification will be used to further reduce the status of professionals (see Abbott 1991).

Indeed, the field faces again the boundary question that has so long obsessed it. Earlier in this century, the boundary issue was 'what is a profession?' Given the new structure of employment relations, even in such highly professionalized surroundings as technological research firms and law firms, it is not clear what kinds of workers ought to be the subject of the sociology of professions. If the pattern of associational professionalism that gave rise to the field itself is either dead or transformed, then the field must retheorize its objects of scrutiny. Abbott's test—a profession is an occupation that competes by retheorizing others' work—presupposes fixed and organized occupations of a kind that simply may not exist under modern conditions of employment.

While the main focus of the professions literature has been on organizational and structural levels, some have insisted on a focus at the individual level, on the professionals themselves. The most general theme in this area has been a recurrent attempt to resolve the problem professionals (or expert workers) represent to Marxist theories of work. Originating in Djilas's book, *The New Class* (1957), the theme was continued in

Bell's *Coming of Post-Industrial Society* (1973), with its predictions of a knowledge world of the future. Attempting to synthesize this literature Brint (1994) recognizes the intense pressures brought on older structures of professionalism by modern economies. He sees a decline of 'social trustee professionalism' as professionals become mere private citizens with particular interests. Perkin's (1989) much broader analysis argues that the idea of profession became a model for modern societies generally in the middle fifty years of this century. The ideal of an individual with personal 'knowledge capital,' combined with the notion of just reward, was in Perkin's view fundamental to the conception of the modern welfare state. In that state's fall, argued Perkins, fell the professional ideal as well.

Perkins's argument provides a useful conclusion. In his article on 'professions' in the prior edition of this encyclopedia, Talcott Parsons (1968) closed with the confident assertion that 'the professional complex has already not only come into prominence but has even begun to dominate the contemporary scene in such a way as to render obsolescent the primacy of old issues of political authoritarianism and capitalistic exploitation.' Now that history has shown him wrong, the sociology of professions must develop a theory that can stand the test of major historical change.

See also: Bureaucracy and Bureaucratization; Doctor–Patient Interaction: Psychosocial Aspects; Law, Sociology of; Lawyers; Legal Professionalism; Medical Expertise, Cognitive Psychology of; Medical Profession, The; Medical Sociology; Organizations, Sociology of; Parsons, Talcott (1902–79); Professionalization/Professions in History

Bibliography

Abbott A 1988 *The System of Professions*. University of Chicago Press, Chicago

Albott A 1991 The future of professions. *Research in the Sociology of Organizations* **8**: 17–42

Bell D 1973 *The Coming of Post-Industrial Society*. Basic, New York

Boltanski L 1982 *Les cadres*. Editions de Minuit, Paris

Brint S 1994 *In an Age of Experts*. Princeton University Press, Princeton, NJ

Carr-Saunders A P, Wilson P A 1933 *The Professions*. Oxford University Press, Oxford, UK

Cocks G, Jarausch K H 1990 *German Professions*. Oxford University Press, New York

Djilas M 1957 *The New Class*. Praeger, New York

Freidson E 1970 *Profession of Medicine*. Dodd Mead, New York

Jarausch K H 1990 *The Unfree Professions*. Oxford University Press, New York

Johnson T J 1967 *Professions and Power*. Macmillan, London

Krause E A 1994 *Death of the Guilds*. Yale, New Haven, CT

Larson M S 1977 *The Rise of Professionalism*. University of California Press, Berkeley, CA

Millerson G 1964 *The Qualifying Associations*. Routledge, London

Parsons T 1954 The professions and social structure. *Essays in Sociological Theory*. Free Press, New York, pp. 34–49

Parsons T 1968 Professions. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 12, pp. 536–47

Perkin H 1989 *The Rise of Professional Society*. Routledge, London

Wilensky H 1964 The professionalization of everyone? *American Journal of Sociology* **70**: 137–58

A. Abbott

Program Evaluation

Evaluations are, in a broad sense, concerned with the effectiveness of programs. While commonsense evaluation has a very long history, evaluation research which relies on scientific methods is a young discipline but has grown massively in recent years. Program evaluation has made important contributions to various social domains; for example, to education, health, economics, and the environment.

This article gives an introduction to program evaluation by dealing with the following three questions: What is program evaluation? What standards can be set to enhance quality and fairness of evaluation? How to conduct a program evaluation?

1. What is Program Evaluation?

1.1 Definition, Distinction from other Activities, Users, and Evaluators

Evaluation, broadly construed, has a very long history. For example, in ancient Rome, tax policies were altered in response to observed fluctuations in revenues, and many thousands of years ago, hunters developed complicated procedures for fishing and trapping game. In recent years, commensense program evaluation has evolved into evaluation research which relies on scientific research methods. The article focuses on the latter type of evaluation (for introductory texts on evaluation research see, e.g., Berk and Rossi 1998, Fink 1995, Pawson and Tilley 1997, Rossi and Freeman 1993). In the article, the terms evaluation, program evaluation, and evaluation research are used synonymously. Concerns for the first large-scale national evaluations of social programs in the USA had its origins in the War of Poverty, and the way of monitoring programs is still evolving rapidly. Evaluation has proven its usefulness to the extent that

most federal and state agencies require evaluations of new programs.

Programs are the objects of evaluation, sometimes also called projects or interventions. Program evaluation derives from the idea that social programs should have demonstrable merits and explicit aims by which success or failure may be empirically judged; for example, literacy programs for high-school students should lead to measurable improvements in their reading skills. Introductory texts on evaluation describe programs as systematic activities that are provided on a continuing basis to achieve preplanned purposes; for example, a national program to train teachers to work in rural and underserved areas, a campaign to vaccinate all children in a school district (e.g., Fink 1995, JCSEE 1994).

Evaluation systematically investigates characteristics and merits of programs. It is the purpose of evaluation to provide information on the effectiveness of programs so as to optimize outcomes, quality, and efficiency. Program evaluations are conducted by applying many of the same methods that social researchers rely on to gather reliable and valid evidence. These include formulating the study questions, designing the evaluation, setting standards to assess the program impact, ongoing monitoring of the program functioning, collecting and analyzing data, assessing program effectiveness relative to the costs, and reporting results. For examples of evaluation in various fields of clinical and applied psychology see, for example, Bickman et al. (1995), Gueron and Pauly (1991), Hope and Foster (1992), and Spiel (2001).

While few authors (e.g., Scriven 1991) take the view that theories are a luxury for the evaluator, since they are not even essential for explanations, the majority of researchers called for injections of theory into every aspect of evaluation design and analysis (e.g., Pawson and Tilley 1997, Berk and Rossi 1998). This implies that program evaluation capitalizes on existing theory and empirical generalizations from the social sciences, and is, therefore, after all, applied research.

By relying on the scientific method, program evaluation differs from e.g., newspaper reporting and social commentary. Program evaluation differs from the social sciences in its goals and audience.

Evaluation goals are defined by the client; that is the individual, group, or organization that commissions the evaluator(s), for example, policymakers. However, often programs and clients do not have clear and consistent goals that can be evaluated for effectiveness. Therefore, to meet evaluation standards, program evaluation should be conducted by professional evaluators who are experts in social sciences, using a broad repertoire of concepts and methods, but who have also expertise in project and staff management, and have at least basic knowledge of the evaluated domain.

While impetus for evaluation rests in policy, information from evaluations is used by the *stakeholders*, all individuals or groups that may be involved in or affected by a program evaluation. There are at least five groups (Fink 1995): (a) federal, state, and local government, for example, a city, the National Institute of Education; (b) program providers, for example, the healthcare services, the curriculum department in a school; (c) policymakers, for example, a subcommittee on human services of the state; (d) program funders, for example, the Office of Education; (e) social researchers, for example, in government, public agencies, and universities. Therefore, program evaluation is also an interdisciplinary task where evaluators and stakeholders work together (see Cousins and Earl 1992).

1.2 Key Concepts and Assignment Criteria

Evaluation can and should be done in all phases of a program. *Baseline data* collected before the start of the program are used to describe the current situation; for example, the need for an intervention. In addition, baseline data are used to monitor and explain changes. For example, if the medical emergency services attempt to reorganize their service to better care for children, they may collect interim data to answer questions such as: In what cases did the emergency medical services fail to meet the needs of critically ill children? What are the differences to the needs of adults? What is the current performance of medical staff?

Before the start of the program a *prospective evaluation* can be employed to determine the program's potential of realisation, its effectiveness, and impact; that is, the scope of its effects.

In *formative evaluation*, interim data are collected after the start of a program but before its conclusion. It is the purpose of formative evaluation to describe the progress of the program and, if necessary, to modify and optimize the program design.

A *process evaluation* is concerned with the extent to which planned activities are executed and, therefore, is nearly always useful. For example, process evaluation can accompany implementation of a community-oriented curriculum in medical schools.

Outcome evaluation deals with the question of whether programs achieve their goals. However, as mentioned before, goals are often stated vaguely and broadly and, therefore, *program effectiveness* is difficult to evaluate. In practice, the concept of effectiveness must always address the issue 'compared to what?' Often it is difficult to distinguish program effects from chance variation and from other forces affecting the outcome. Berk and Rossi (1998) distinguish between *marginal effectiveness*, *relative effectiveness*, and *cost effectiveness*. Marginal effectiveness concerns the dosage of an intervention measure. For example, to prove the assumption that a better student–teacher ratio would improve student performance, the class size has to be reduced. In this case,

two or more different student–teacher ratios are the doses of intervention and their outcomes are compared.

Relative effectiveness is evaluated by contrasting two and more programs, or a program and the absence of the program. For example, lecture-based learning can be contrasted with problem-based learning in their effects on students' practical performance. Cost-effectiveness considers effectiveness in monetary units. Evaluation usually is concerned with the costs of programs and their relationship to effects and benefits. For example, in the evaluation of a physical education program, one can ask how much gain in monetary units can be expected from the reduced use of health services.

However, the entire impact of a program, for example, the extent of its influence in other settings, is very difficult, maybe impossible to assess. Therefore, *summative evaluations* are applied, that is, historical reviews of programs that are performed after the programs have been in operation for some period of time.

2. *What Standards Can be Set to Enhance Quality and Fairness of Evaluation?*

Because of the increasing number of evaluations and the high impact that decisions based on results of evaluations may have, there has been since the early 1970s an ongoing discussion about guidelines for effective evaluations. In the absence of any clear definition of what constitutes a reasonable program evaluation, the Joint Committee on Standards for Educational Evaluation (JCSEE 1994), in cooperation with evaluation and research specialists, compiled knowledge about program evaluation and established a systematic progress for developing, testing, and publishing evaluation standards. The Joint Committee began its work in 1975. In 1989 the progress for developing standards was adopted by the American National Standards Institute (ANSI), and became available worldwide.

Originally, the standards were developed to guide the design, implementation, and assessment of evaluations of educational programs, projects, and materials. However, in the new *Program Evaluation Standards*, published in 1994, new illustrations were included that feature applications of the standards in such other settings as medicine, nursing, the military, business, law, government, and social service agencies (JCSEE 1994). The standards are published for people who commission and conduct evaluations as well as for people who use the results of evaluations.

The Joint Committee defined an *evaluation standard* as a principle mutually agreed on by people engaged in the professional practice of evaluation, that, if met, will enhance the quality and fairness of an evaluation (JCSEE 1994).The standards provide advice on how to judge the adequacy of evaluation activities, but they do not present specific criteria for such judgments. The reason is that there is no such thing as a routine evaluation and the standards encourage the use of a variety of evaluation methods.

The 30 standards are organized around the four important attributes of sound and fair program evaluation: *utility*, *feasibility*, *propriety*, and *accuracy* (JCSEE 1994).

Utility Standards guide evaluations so that they will be informative, timely, and influential. They require evaluators to keep in any phase of the evaluation the stakeholders' needs in mind. Seven standards are included in this category; for example, Stakeholder Identification, Evaluator Credibility, Report Clarity, and Evaluation Impact.

Feasibility Standards call for evaluations to be realistic, prudent, diplomatic, and economical. They recognize that evaluations usually are conducted in natural settings. The three standards included in this category are Practical Procedures, Political Viability, and Cost Effectiveness.

Propriety Standards are intended to facilitate protection of the rights of individuals affected by an evaluation. They urge evaluators to act lawfully, scrupulously, and ethically. In this category eight standards are included; for example, Service Orientation, Formal Agreements, Disclosure of Findings, Fiscal Responsibility.

Accuracy Standards are intended to ensure that an evaluation will reveal and transmit accurate information about the program's merits. They determine whether an evaluation has produced sound information. The 12 accuracy standards include Program Documentation, Valid Information, Reliable Information, Justified Conclusions, and Metaevaluation.

Validity refers to the degree to which an information or measure assesses what it claims to assess, while *reliability* refers to the degree to which an information is free from 'measurement error'; that is, the exactness of the information. *Meta-evaluation* is evaluation of an evaluation.

For each standard the Joint Committee summarized guidelines for application and common errors. In addition, one or more illustrations of the standard's application are given (JCSEE 1994).

However, the Joint Committee acknowledges that standards are not all equally applicable in every evaluation. Professional evaluators must select to identify those that are most applicable in a particular context.

3. *How to Conduct a Program Evaluation?*

Program evaluation typically consists of a chronological sequence of activities or tasks. The main tasks or activies are (see, e.g., Fink 1995, JCSEE 1994): Posing questions about the program and defining evaluation

goals, setting standards of effectiveness, designing the evaluation, collecting and analyzing information, reporting the results. In addition, an evaluation plan must be worked out that specifies each of the tasks to be accomplished and the personnel, time, and resources needed.

3.1 Posing Evaluation Questions and Setting Standards of Effectiveness

Based on the program's goals and objectives, evaluation questions are formulated. While goals are often relatively general, e.g., to improve the organization of an institution, objectives refer to the specific purposes of a program; for example, to produce a guidebook about how to review the literature. Evaluation questions may also concern the magnitude, duration, and distribution of effects.

Because program evaluations should provide convincing evidence concerning the effectiveness of a program, specific effectiveness criteria are set. Because these *standards of effectiveness* determine the evaluation design, they should be set prior to any evaluation activity. Information about standards came from the literature, from experts, normative data, and statistical analyses (see, e.g., Donabedian 1982).

Consider, for example, the program goal to teach students how to review the literature (see Fink 1995). Then the evaluation question is, whether students learned to do this. The standard of effectiveness can set, for example, (a) that 90 percent of students learn to review the literature or, (b) that a statistically significant difference in learning is observed between students in Schools A and B, with students in School A participating in a new program and students in School B serving as controls. The decision for either standard (a) or (b) can have consequences for the evaluation design, the selection of participants, and the statistical analyses of the data.

3.2 Designing the Evaluation

An *evaluation design* is a structure created to produce an unbiased appraisal of a program's benefits. The decision for an evaluation design depends on the evaluation questions and the standards of effectiveness, but also on the resources available and on the degree of precision needed. Given the variety of research designs there is no single 'best way' to proceed (for an overview see, e.g., Cook and Campbell 1979, von Eye and Spiel 1996).

Essential criteria for evaluation designs are internal and external validity. *Internal validity* is attained when the evaluator can decide whether a finding is due to the program and cannot be caused by some other factors or biases. *Randomized experiments* are the method of choice to reach this criterion; for example, when participants are by chance assigned either to program A, to program B, or to no program. *External validity* concerns the generalizability of findings; for example, to participants of the program in other places and at other times.

3.3 Collecting Information

For collecting information an arsenal of methods is available; for example, questionnaires, performance tests, various kinds of interviews, record reviews, observations, etc. (see, e.g., Kosecoff and Fink 1982, Rosenbaum 1995). When selecting the methods of data collection, the evaluator must consider the evaluation questions, the available technical and financial resources, and the time needed to develop a new instrument. A 'good' measure is, in common terms, one that is likely to measure accurately what it is supposed to measure (Berk and Rossi 1998). In methodological terms, a good measure is reliable and valid.

Mostly, more than one measure is used to obtain the relevant information. For example, when health-related behavior of 14-year-old students is the variable of interest, possible methods of data collection include surveys of students, surveys of parents and teachers, observations of students, and reviews of health and medical records (Fink 1995).

3.4 Analyzing and Reporting Data

Program evaluators and social scientists employ statistical methods to analyze and summarize data. However, evaluators have to keep in mind that statistical significance does not always imply practical significance or educational meaningfulness. Standards of effectiveness that have been set in advance of the evaluation (see above) should guarantee that analyses of the data provide also evidence about practical effectiveness of the program (see, e.g., Fink 1995).

To combine results of evaluations that address the same questions or goals, meta-analyses can be conducted. A *meta-analysis* is an analysis of evaluation data from several studies. It is a statistical procedure which attempts to provide estimates of program impact by combining the results from a number of distinct evaluations to increase statistical power and the generalizability of results (see, e.g., Rosenthal 1991).

Reporting results is the last task for the evaluator. The evaluation report answers the evaluation questions, describes how the answers were obtained, and translates the findings into conclusions and recommendations about the program.

4. Concluding Remarks

The demands for sound program evaluation continue to grow in a way that some authors (Pawson and Tilley 1997) fear may become mandatory that everything

needs evaluating. Therefore, it is essential that quality control procedures and standards be introduced for all facets and tasks of evaluation and that the evaluation itself is evaluated against pertinent standards. However, there is also the risk of oversimplifying and only using a checklist of standards. There is no fixed recipe of how to run a successful evaluation.

See also: Curriculum Evaluation; Educational Assessment: Major Developments; Educational Evaluation: Overview; Industrial–Organizational Psychology: Science and Practice; Job Design and Evaluation: Organizational Aspects; Performance Evaluation in Work Settings; Psychological Treatment, Effectiveness of

Bibliography

Berk R A, Rossi P H 1998 *Thinking about Program Evaluation*. Sage, Thousand Oaks, CA

Bickman L, Guthrie A R, Foster M, Lambert E W, Summerfelt W T, Breda C, Heflinger C 1995 *Managed Care in Mental Health: The Fort Bragg Experiment*. Plenum, New York

Cook T, Campbell D 1979 *Quasi-experimentation*. McGraw-Hill, New York

Cousins J B, Earl L M 1992 The case for participatory evaluation. *Educational Evaluation and Policy Analysis* **14**: 397–418

Donabedian A 1982 *The Definition of Quality and Approaches to its Assessment*. Health Administration Press, Ann Arbor, MI

Fink A 1995 *Evaluation for Education & Psychology*. Sage, Thousand Oaks, CA

Gueron J, Pauly E 1991 *From Welfare to Work*. Sage, New York

Hope T, Foster J 1992 Conflicting forces: Changing the dynamics of crime and community on 'problem' estates. *British Journal of Criminology* **32**: 488–504

Joint Committee on Standards for Educational Evaluation (JCSEE) 1994 *The Program Evaluation Standards*, 2nd edn. Sage, Thousand Oaks, CA

Kosecoff J, Fink A 1982 *Evaluation Basics*. Sage, Newbury Park, CA

Pawson R, Tilley N 1997 *Realistic Evaluation*. Sage, London

Rosenbaum P R 1995 *Observational Studies*. Springer-Verlag, New York

Rosenthal R 1991 Meta-analysis: A review. *Psychosomatic Medicine* **53**: 247–71

Rossi P H, Freeman H E 1993 *Evaluation. A systematic approach*, 5th edn. Sage, Newbury Park, CA

Scriven M 1991 *Evaluation Thesaurus*, 4th edn. Sage, Newbury Park, CA

Spiel C (ed.) 2001 *Evaluation universitärer Lehre—zwischen Qualitätsmanagement und Selbstzweck* [Evaluating University Teaching—Between Quality Management and End in Itself]. Waxmann, Münster, Germany

von Eye A, Spiel C 1996 Research methodology: Human development. In: Tuijman A C (ed.) *International Encyclopedia of Adult Education and Training. Concepts, Theories, and Methods*, 2nd edn. Sect. 1. Pergamon Press, New York

C. Spiel

Progress: History of the Concept

The idea of progress was integral to the project of creating social sciences in eighteenth-century Europe, and the relationship has continued since. During the nineteenth century, the idea that 'civilization has moved, is moving, and will move in a desirable direction' (Bury 1932) was the historical framework and central problematic of social science theory and practice. The idea of progress as inevitable collapsed in the vicissitudes of the twentieth century, although in mitigated forms, it continued, waxing and waning with historical expectations. One component, the idea of the progress of the social sciences, figured prominently as the ideology of the social science disciplines.

1. Eighteenth-century Origins

The idea of progress and the idea of social science were centrally involved in each other's origin. Both emerged as part of the historicist understanding of history as a realm of human construction, propelled ever forward in time and taking new qualitative forms. Modern society was thus decisively different from its feudal and ancient forerunners, engaged upon a novel historical course. Replacing earlier cyclical and supernatural views of historical time, progress instantiated in human history the teleological movement that governed the Christian drama of redemption. Progress was a universal journey, and the array of human differences sharpened for European thinkers the progressive character of their own society and the stages by which they had reached modernity (Meek 1976). The emergence of the social sciences was itself a lynchpin of progress. Science was understood as the most advanced form of reason and the extension of the progress of science into the realm of government and morals—barely begun—was central to the expectation of civilizational advance. By the same token, the task of the social sciences was to understand the forces that were propelling the modern European world and to guide it into the future.

The theories of progress propounded in the eighteenth century drew upon, and were often coincident with, the early formulations of social science. Associationist psychology explained the ability of human mental capacities to improve over time, while the studies of historians and economists and the observations of travelers set the trajectory of progress toward greater complexity and refinement: with increasing social differentiation and intercourse, came improving manners, morals, knowledge, and rationality. The guiding role of providence in this process was still visible, but it was understood to work through nature. Particularly in the UK and France, natural principles assured the existence and continuation of progress.

Several major theories of progress emerged in this context. Scottish 'philosophical historians' like Adam Smith and Adam Ferguson developed four- or five-stage theories of history in which the economic mode of production at each stage shaped political organization, customs, and moral temper. The expansion of population, the division of labor, and technological advance nonetheless emerged within a complex social process, as history was often the result of unintended consequences. In contrast, the *philosophes* of the French Enlightenment like Condorcet believed reason to be the key factor in historical progress. The rational acts of enlightened individuals, and particularly of utilitarian science, moved history forward, while ignorance and war held it back. In Germany, Kant and Herder made historical progress the work of philosophical and moral reason. History was the progressive fulfillment of humanity's moral and spiritual destiny, realized through the unique development of each historical people.

Emerging in contested political arenas, the early theories of progress staked out ideological positions of longstanding influence in politics and the social sciences. Designedly, Smith's view lent itself to a liberal presumption against state action; Condorcet's view, to a justification of the authority of social scientists in the guidance of society; and Kant's view, to a liberalism subordinated to idealist ends. If the reliance on progressive reason assigned all evil to its reactionary enemies, the Scottish view recognized that progress involved secondary losses. Smith's critique of the diminished life of industrial labor began a longstanding discourse that made progress itself responsible for the problems of modern society.

2. *The Nineteenth-century Triumph of Progress*

The French Revolution raised the stakes for both progress and social science as the millennial hopes it aroused were projected into the secular future. Progress became the dominant view of history among the educated classes of Europe and North America, while contesting versions of progress proliferated. Responding to rapid capitalist development and the continuing possibility of political and social revolution, theorists of political economy, sociology, and anthropology recast the eighteenth-century theories of progress and sought to specify the factors that controlled the course of history.

The classical political economy that developed from Smith focused on the natural laws of economic growth and, of equal importance, on how the expanded product would be distributed. The question of whether the average workman would be better off, despite growing inequality, became urgent as egalitarian, utopian versions of progress appeared. The Reverend Thomas Malthus argued that food supply would keep wages at subsistence level. His political economy, informed by an Anglican reading of history, supported a cautious view of progress, incorporating cyclical movement and the necessity of suffering. The theories of David Ricardo were often pressed into this mold, though most practicing economists, including Ricardo, maintained a more optimistic scenario of growth based on capital accumulation. By mid-century, recognition of the role of technology in increasing productivity shifted the weight of economic opinion toward more optimistic predictions of growth (Berg 1990).

American exceptionalists and socialist radicals transformed this liberal political economy in opposite directions. Henry C. Carey argued that the uncultivated American continent would abrogate altogether the action of Malthusian law, vindicating America's utopian historical course (Ross 1991). Socialists, charging the capitalist economy with the degradation of workmen and the destruction of community, made capitalism an intermediate stage, to be superceded by socialism. In Karl Marx's formulation, history progressed by a process of historical contradiction between the forces and social relations of production: capitalism produced within itself the revolutionary proletariat that would destroy it and the technological basis of abundance that would support socialism (Cohen 1978).

August Comte, inheriting Condorcet's identification of historical progress with progress of the human mind, made social science its central agent. Civilization moved by necessary law from the theological to the metaphysical to the ultimate stage, the scientific, still struggling to be born. Although this trajectory was inevitable, action in accord with historical law could ease the transition. A new science of society, sociology, would objectively examine the world as it is and, based thereon, set out a plan for the rational ordering of society, opinion, and feeling (Lenzer 1975). Herbert Spencer's version of sociology, influenced by liberal political economy and Scottish philosophical history, removed the central planning role from sociology, but gave it the same power to discern the laws of historical progress to which successful social action must conform. More complacent about his own era, Spencer believed that the industrial age would encourage moral, intellectual, and social improvement. Spencer turned historical progress into a process of social evolution, drawing on biological notions of adaptation to the environment, the inheritance of acquired characteristics, and an embryological model of differentiation to form a cosmic law of evolution (Peel 1971).

Charles Darwin's *The Origin of Species* (1859) reinforced both the idea of progress and the use of biological analogy. Despite the radically nonprogressivist implications of natural selection, Darwin's theory, with its suggestion of the continuity of animal and human life, was often read in this era of faith in

progress—even by Darwin himself—as a proof of the progressive character of both evolution and history. The contemporaneous recognition of the long age of the earth created a space for 'prehistory,' solidifying the uniquely advanced position of European civilization (Segal 2000). Social scientists developed a 'comparative method' that sorted peoples and races on an evolutionary grid, assigning them to stages of a single evolutionary process by Euro-centered hierarchical standards, although that gradient was always subject to Romantic subversion (Bowler 1989).

3. *The Problem of Progress, 1890–1914*

During the late nineteenth and early twentieth century, the social sciences formed specialized, university-based disciplines, in part to grapple with the accelerating problems of modernity. As industrialization, democracy, and imperialism transformed society, Enlightenment hopes for modernity took chastened forms. At the same time, more stringent conceptions of science outlawed deterministic laws of history like those of Comte and Spencer as metaphysical constructions and called attention to the moral valuation inherent in the concept. By the early twentieth century, the rediscovery of Mendelian genetics attenuated the link between evolutionary biology and historical change, and critics, notably anthropologist Franz Boas, attacked the comparative method and its assumption of a single line of evolutionary advance. Progress no longer appeared to leading social scientists as an inevitable, universal process, guaranteed by natural or historical law. Yet the question of what form, and how beneficial a form, modern society would take remained central as did many nineteenth-century terms of discussion.

In economics, marginalism shifted the focus of attention from economic growth to allocation. Those who joined the mixed neoclassical discourse that resulted grounded capitalism in laws analogous to the laws of physical nature and retained optimistic assumptions of economic growth and its central role in social advance, but those assumptions no longer framed their specialized and dehistoricized discourse. Other economists who sought a qualitative change in capitalism worked within a self-consciously historical or institutional framework, like the socialist Thorstein Veblen, who asked in 1899 'Why Is Economics Not an Evolutionary Science?' (Ross 1991).

Sociologists narrowed their focus to the tendencies of modern society. In Germany, where antimodernist attitudes flourished, sociologists were most pessimistic. For Ferdinand Tönnies and Max Weber, traditional community (*Gemeinschaft*) and modern capitalist society (*Gesellschaft*) marked polar modes of social organization. The impersonality and loss of spirituality of modern society were inevitable consequences of its organizing principles. Weber, more than Tönnies, saw the positive advantages of instrumental rationality, yet termed the framework it built for modern society an 'iron cage' (Liebersohn 1988).

Emile Durkheim, like American sociologists, retained the more hopeful nineteenth-century formulation that placed the present troubled society in transition to a more harmonious modernity. Durkheim (1933, [1893]) argued that a new kind of functional and normative social integration was forming; the anomie of individuals thrown outside the regulative norms of society by the upheavals of social transformation would subside as modern society reached equilibrium. Edward A. Ross posited a transition from the natural social control of face-to-face community to the more fragile, 'artificial' controls of modern society, similar to the German model of *Gemeinschaft/Gesellschaft*, but he too expected that a harmonious adjustment would be reached (Ross 1991).

It was primarily in thinking about their own role in modern society that social scientists returned to, and reaffirmed, the central role of science in progress. In his presidential address to the American Psychological Association in 1899, John Dewey expressed both the anxieties of modernity and the solution favored by his colleagues. 'With tremendous increase in control of nature,' he noted, 'we find the actual realization of ends, the enjoyment of values growing unassured and precarious.' But, he added 'the entire problem is one of the development of science and of its application to life.' When the social sciences are further developed, 'we can anticipate no other outcome than increasing control in the ethical sphere' comparable to 'the revolution that has taken place in the control of physical nature' (Dewey 1976, **1**: 149–50). That science itself progressed, that it produced increasing control over nature and social life, and that such control worked to the benefit of humanity remained unquestioned. This faith served as the ideological basis for the social science disciplines.

4. *Progress Dethroned, 1914–1945*

The most severe challenges to the idea of progress occurred in the decades spanned by the two world wars: economic depression, totalitarian politics, and wars that unleashed mass destruction made possible by modern science and technology. At the same time, the intellectual and cultural changes that had undermined the idea of progress in the late nineteenth century gathered force after World War I. A sharp sense of historical discontinuity, breaking off the new world of mass production, mass consumption, and mass politics from the past, also broke the sense of cumulative gain in civilization on which progress depended. Historians mark World War I as the end of progress both as an unquestioned public faith and as

an idea that social scientists could presuppose or hope to demonstrate.

What remained open was the possibility of limited lines of progress, as human constructions, and the hope, still carried by the Marxist and liberal traditions and in the USA by an exceptionalist historical faith, that limited the gains worked toward the larger progress of history. Thus the American sociologist William F. Ogburn explained that social scientists had dropped the term progress and were instead studying '*social change*,' as a 'term free from dogmatic or moral implications.' Social change is caused by inventions, Ogburn argued, but inventions occur unevenly, creating 'lags' until society 'adjusts.' In his analysis, however, adjustments can always be made and society was always better off when they were. Progress was not so much abandoned as forced underground (Ogburn 1930).

The point at which it remained still vibrantly alive, especially in the USA, was the faith in science. The new understanding of science itself suggested that if science did not mirror reality, it could more readily reconstruct reality to human purposes (Porter 1994). While Ogburn was moderate in his hopes for prediction and control, other American social scientists of the 1920s, like the behaviorist psychologist John B. Watson and the political scientist Charles Merriam, expressed utopian faith in science and technology (Ross 1991). A more sober tone emerged throughout the West in the 1930s, but the possibility of historical progress through science remained a central feature of American social science theory and ideology.

5. *The Return of Progress, 1945–1970*

In the decades after World War II, the vision of a progressive modern society guided by science gained energy and urgency from the defeat of fascism, the disintegration of colonial empires, and the threat of communism. As the strongest power to emerge from the war, the USA embraced that vision with renewed self-confidence. In what was experienced as the 'American Moment' of an 'American Century,' the USA already appeared to stand at the summit of world history ready to embody the values its exceptionalist history promised.

This static notion of historical time was mirrored in social science theories of functionalist, cybernetic, and equilibrium systems that nonetheless incorporated the idea of progress. The sense of progress achieved was embodied most visibly in modernization theory, which measured the distance the underdeveloped world still needed to traverse to achieve the American or Western norm (Latham 1999).

Modernization theory also marked one of the many points in these decades at which the social scientists' ideology of progress through social science entered into the substance of social science theory. Thus modernization was defined as a 'process of social change that seeks to govern itself by rational policy planning' for which social scientists were 'indispensable' (Lerner 1968). Talcott Parsons suggested that academic intellectuals and the practical professionals they trained had become 'the most important single component in the structure of modern societies,' displacing from leadership the older political and newer capitalist class (Parsons 1968).

The possibility of progress, along with the liberal and Marxist traditions in which it was imbedded, revived in Western Europe with postwar reconstruction, although with far greater circumspection. In Germany, Jürgen Habermas blended native Marxist traditions with American theories of pragmatism and democracy to project not a utopian future, but an ideal against which modern society might measure its defects (A. Giddens on Habermas, in Skinner 1985). The dehumanizing character of modern society remained a powerful, and sometimes dominant, theme in German sociology.

The idea of progress and its drawbacks also played a conspicuous role in the development of the social sciences in the non-Western world. As an arm of secular, Western-style development, the social sciences had been since the nineteenth century a major site of debate over whether Western or native paths to modernity should be chosen. After World War II, the developing world also served as a major source of critical revision of Western conceptions of modernization, recasting modernization as economic dependency, or redefining tradition and modernity as mutually transformative rather than diametrically opposed processes.

6. *Postmodernism and Progress*

Political conflicts in Europe and the USA in the 1960s shattered the American Moment, opening the way for longstanding discontents with modernity and reconsideration of the idea of progress (Salomon 1955). The most radical challenge—the postmodern critique of modernity—emerged in the home of the rationalist Enlightenment. Francois Lyotard attacked the idea of progress, with its universalist and rationalist premises, as the basis of the most oppressive, totalitarian impulses of modern Western society. Michel Foucault pointed specifically to the professions and social sciences as the source of the knowledge/power that disciplined the unruly masses of modern society in the name of enlightened rationality. The pluralistic and libertarian values that animated this critique gained power from the increasing reach of Western capitalism into every area of the world and the proliferation of multi-ethnic societies. Globalization in the postmodern reading was producing not a single narrative of universal progress, but a pluralistic, hybrid world of many historical narratives. However, as the term postmodern suggests this vision can be read as a liberationist narrative of progress, drawn from Romantic,

rather than Enlightenment, strains of modernity (McGowan 1991).

This pluralistic vision had some influence on the social sciences as a support of feminism and multiculturalism within the USA and as a counter to Americanization elsewhere. The liberal, Western-centered idea of progress revived, however, on the strength of the free-market neoclassical paradigm in economics, with its presumption of capitalist-led progress, and a revival of the narrative of American exceptionalism after the fall of communism. Globalization in that context defined progress as the Americanization of the world.

The most fundamental aspect of the postmodern critique of progress, however, was its critique of scientific reason. Here, postmodern critics joined broader currents of philosophical criticism of positivism and analysis of the linguistic and social construction of knowledge. Although Thomas Kuhn argued that *The Structure of Scientific Revolutions* (1962) did not destroy the idea of scientific progress, it did radically remove the sciences from a universalistic narrative of the progressive discovery of truth and re-situate them in a social-historical world of shifting purposes and uncertain directions (B. Barnes on Kuhn, in Skinner 1985). Most social scientists in the USA have not paid attention to the philosophical debates around this work. Indeed, in those segments of social science identified most closely with mathematics and natural science, like economics, older positivist understandings of science are especially powerful (Backhouse 1997). However, the context of critique has weakened the self-confidence, and with it, the ideological construction of the progress of social science, among other segments of the social science disciplines. While the eighteenth-century formulation of historicism as the story of progress gave the social sciences their charter, the critical historicism of the late twentieth century called for a reformulation of both the historical assumptions and the self-understanding of the social science disciplines.

See also: Comte, Auguste (1798–1857); Differentiation: Social; Enlightenment; Evolution, History of; Evolutionism, Including Social Darwinism; Functionalism, History of; Kuhn, Thomas S (1922–96); Modernity; Modernity: History of the Concept; Modernization and Modernity in History; Modernization, Sociological Theories of; Postmodernism: Philosophical Aspects; Rationalism; Spencer, Herbert (1820–1903)

Bibliography

Backhouse R E 1997 *Truth and Progress in Economic Knowledge*. Edward Elgar, Cheltenham, UK

Berg M 1990 Progress and providence in early nineteenth-century political economy. *Social History* **15**: 365–75

Bowler P J 1989 *The Invention of Progress: The Victorians and the Past*. Blackwell, Oxford, UK

Bury J B 1932 *The Idea of Progress*. Macmillan, New York

Cohen G A 1978 *Karl Marx's Theory of History*. Princeton University Press, Princeton NJ

Dewey J 1976 Psychology and social practice. In: Boydston J (ed.) *John Dewey, The Middle Works*. Southern Illinois University Press, Carbondale, IL, Vol. 1

Durkheim E 1933 [1893] *The Division of Labor in Society*. Macmillan, New York

International Encyclopedia of the Social Sciences 1968 Macmillan and The Free Press, New York

Latham M E 1999 *Modernization as Ideology: American Social Science and 'Nation-building' in the Kennedy Era*. University of North Carolina Press, Chapel Hill, NC

Lenzer G 1975 *Auguste Comte and Positivism*. Harper & Row, New York

Lerner D 1968 Modernization: Social aspects. In: *International Encyclopedia of the Social Sciences*. Macmillan and the Free Press, New York

Liebersohn H 1988 *Fate and Utopia in German Sociology, 1870–1923*. MIT Press, Cambridge, MA

McGowan J 1991 *Postmodernism and its Critics*. Cornell University Press, Ithaca, NY

Meek R L 1976 *Social Science and the Ignoble Savage*. Cambridge University Press, Cambridge, UK

Ogburn W F 1930 Change, Social. In: *Encyclopaedia of the Social Sciences*, **3**: 330–4. Macmillan, New York

Parsons T 1968 Professions. In: *International Encyclopedia of the Social Sciences*. Macmillan and the Free Press, New York

Peel J D Y 1971 *Herbert Spencer: The Evolution of a Sociologist*. Basic Books, New York

Porter T M 1994 The death of the object: *Fin de siecle* philosophy of physics. In: Ross D (ed.) *Modernist Impulses in the Human Sciences 1870–1930*. Johns Hopkins University Press, Baltimore, MD

Ross D 1991 *The Origins of American Social Science*. Cambridge University Press, Cambridge, UK

Salomon A 1955 *The Tyranny of Progress*. Noonday Press, New York

Segal D A 2000 'Western civ' and the staging of history in American higher education. *American Historical Review* **105**: 770–805

Skinner Q 1985 *The Return of Grand Theory in the Human Sciences*. Cambridge University Press, Cambridge, UK

D. Ross

Progressive Education Internationally

Progressive education (*éducation nouvelle*, *Reformpaedagogik*) is an international movement which originated in Western Europe and North America by 1900 in different national contexts. It can be seen as a reaction to the decisive educational fact of the nineteenth century that, with the introduction of compulsory schooling, education had been achieved for everyone in the younger generation, even the inhabitants of the smallest rural villages, even the inhabitants of the working-class areas of growing big cities.

Education thus institutionalized as a system for the masses showed its deficiencies. Progressive education tried to overcome these defects while maintaining education for all.

1. Origins and International Spread

Progressive education started in the UK with the 'new schools' such as Abbotsholme, founded by Cecil Reddie in 1889; Bedales, founded by John H. Badley in 1893; and King Alfred's School, set up by a group of parents in northwest London in 1898. A generation later, Alexander S. Neill, a former teacher at King Alfred's, who had been in contact with child-oriented education in Homer Lane's self-governed 'Little Commonwealth,' and who had grounded his experiences in psychoanalytical theory, founded Summerhill in 1921 as an experiment of antiauthoritarian education.

Progressive education in the USA is linked with the practical and theoretical activities of educators, such as the Columbia University teacher trainer William H. Kilpatrick with his famous project method; the school superintendent of Winnetka, Illinois, Carleton Washburne, who successfully modeled the schools of his district on progressive principles; the Harvard psychologist William James, who was influential on progressive pre- and in-service teacher training; Helen Parkhurst, who in her Dalton Plan developed an institutional model of individualized learning; Jane Addams of Hull House, Chicago, who introduced progressivism into social work; and above all John Dewey (see *Dewey, John (1859–1952)*), who gave progressive education a sound foundation in pragmatist philosophy.

In the French-speaking countries of Europe the French primary schoolteacher Célestin Freinet, the inventor of the methodology of the *école moderne*, a prominent feature of which is 'the printing press in school,' the Swiss cofounder of the Geneva International Bureau of Education, Adolphe Ferrière, with his idea of the *école active*, where the children were to make their own textbooks as a result of their learning processes, and the Belgian medical doctor Ovid Decroly, the founder of the reform school L'Eremitage in Brussels, are key figures of the *éducation nouvelle*. In Italy, the most prominent representative of progressive education is the preschool educator Maria Montessori, who through her specifically designed didactic materials introduced the idea of self-directed learning into early childhood education. In Germany, the basic patterns of the English new schools were institutionalized by Hermann Lietz and Paul Geheeb in their 'rural education centers'; the Munich school superintendent Georg Kerschensteiner brought learning and practical work more closely together in his 'working school' concept; and the director of the Hamburg art gallery, Alfred Lichtwark, inspired the art education movement. In the Scandinavian countries, the Swedish teacher Ellen Key published the 1900 programmatic pamphlet of progressive education 'The century of the child.'

With the school Lev Tolstoy had erected at his estate in Yasnaya Polyana, progressive education was prepared in the USSR. Stanislav T. Šackih, who had transferred the Anglo-American idea of the settlements into summer camps for children even before the Soviet revolution, built the bridge between progressive and Marxist ideas in the education of the early USSR. The main propagator of melting progressivism into a Marxist–Leninist perspective of education was Pavel P. Blonsky, with his program of the production school. For realizing these efforts in educational policies, even more influential was Lenin's wife, Nadeszhda Krupskaya. The Polish physician and orphanage educator, Janusz Korczak, sealed his view of the dignity of the child by accompanying the Jewish children entrusted to him into the Holocaust.

By the end of World War I, progressive education had grown from local and individual initiatives into national movements and even, despite the nationalist cleavages the war had evoked, into an international network. The founding of the American Progressive Education Association in 1919 and of the international New Education Fellowship in 1921 mark the beginning of that new phase of the movement.

Progressive education now spread from its countries of origin into other parts of the world. Theoretical forerunners, like Tomeri Tanimoto and Kanjiro Higuchi, transferred progressive ideas to Japan which were later on institutionalized in experimental schools. Small group education, self-directed learning, and the working school were tried out by educators like Heiji Oikawa, Kishie Tezuka, and Taneichi Kitazama. Progressive education with its more individualist though socially responsible approach became an oppositional force against the nationalist imperialism that led Japan into World War II. Brazil is a good example of the adaptation of new education to the Latin American context. It was especially the 'three cardinals'—Lourenço Filho, who made efforts for a psychologically grounded theory of instruction, Fernando de Azevedo, who was envisaging the structural reform of the Brazilian education system, and Anísio Teixeira, who was striving for a democratic educational policy—that were defending educational progressivism against the conservative Catholic educational theory of the country. Later on, these progressive impulses merged with elements of the 'educaçao popular.'

On the African continent a mass education system in the 1920s was not yet developed. Consequently there was only a scarce influence of progressive education on the educational activities of the African countries. The school farms and school workshops of different missionary societies linked primary schooling for students and adult literacy programs to manual and agricultural activities. The Phelps–Stokes Com-

mission, which in 1919 evaluated the education situation in the British territories of Africa recommended a practical orientation toward basic education. The skills taught at school should be directly useful in family and community life. In Kenya in 1925 there was an attempt to train teachers as community development workers.

On the Australian continent progressive education was 'very much a series of side eddies, and not the main stream of education' (Connell 1995, p. 351), but teacher educators such as G. S. Browne of the University of Melbourne had acquainted their students with the Dalton Plan or the project method, or the educational philosophy of Dewey.

A quantitative content analysis of educational journals proved that by 1930 progressive education concepts had become the leading educational semantics worldwide; progressive education had indeed developed into an international movement (Schriewer et al. 1998).

2. *Core Elements and Comprehensive Principles*

The intertwining educational philosophies, political semantics, pedagogical programs, theories, projects, foundations of institutions, and teaching methodologies which progressive education consists of are not easily summarized. In spite of Cremin's verdict concerning a capsule definition of progressive education—'None exists, and none ever will' (Cremin 1961, p. x)—the common structural origin and the mutual international criticism of progressive educators, which distinguished between that which was pluralistically proven and that which was blinded by ideology, allow for pointing out some core elements or comprehensive principles and practical approaches of progressive education. These can be reconstructed by looking at the classic historical accounts of Cremin (1961), Boyd and Rawson (1965), Mialaret et al. (1976), or Roehrs (1998b), which together with Roehrs and Lenhart (1995) at the same time contain biographical and work-related references to the aforementioned progressive educators:

(a) Progressive education stood for the democratization of the industrial society by opening up chances of participation on an educational basis. This referred to the induction into technical-industrial processes of production and their social structuring as well as to the authorization to the exercise of political rights. Progressive education brought (sub)cultural group identities into focus. The concept of enabling the individual to cooperate included social feedback on the individual possibilities of action that were opened up by education. The interrelations of society as a whole were thus changed via the spheres of proximity of everyday life. An expression of this is the community-oriented approach of progressive educational endeavors. Education is conceived as an aid for everyone to cope with life.

(b) Progressive education is distinguished by its orientation toward the child's individuality. This means a child-centeredness that provides the educands with enough time to find and do things themselves while playing and learning. In life philosophy's metaphor of the 'opening of the eye,' the pedagogical effort is to let the children themselves experience and recognize 'things.'

Progressive education's intellectuality includes a new relationship to the body. The model is a simple, healthy life-style: there is no cult of athletics, but no neglect of the body either. The concept of human nature is rather pragmatic or positivist—Mialaret et al. (1976, p. 12), following Freinet, 'mechanize the functions (of the body) with the aim to free the vigor of life for the use of the higher activities of the mind.' It does not rule out an ultimate religious bond, but does not put it to the fore either.

(c) As a classical-modern reform movement, progressive education is (social and behavioral) science-oriented. It holds that a scientific foundation for and advice on practical education are both possible and necessary. The concept of science here is pluralistic, including different paradigmatic orientations. Progressive education above all refers to psychology, especially developmental psychology as it was constituted in empirical research and theoretical construction, for example in the works of Piaget (e.g., 1963). Next it refers to a sociology as it was put in concrete forms by J. Dewey (e.g., 1966) or G. H. Mead (e.g., 1967; see *Mead, George Herbert (1863–1931)*), for instance, with a pragmatic intention. Finally, it refers to education as an academic discipline as well, as this has, in its idealist-humanistic variant, become almost an accompanying theory of the German pedagogical reform projects.

(d) Progressive education at its heart is a school reform movement. The evolution of school as an educational institution is not completely to be reversed; still, teaching and learning at school is to be brought closer to the learning processes taking place in the family, workshop, workplace, and community. At the same time, the roles of those who encounter each other at school are changing. There is a friendly relationship between teachers and pupils, based on partnership and companionship, and the opportunities of involvement stretch as far as to the rights of participation ('government') in school communities. Progressive education disseminated into further educational fields apart from school: parallel to school and partly supported by the reform movement, kindergarten education, adult education, and social work were irreversibly institutionalized as fields of pedagogical action in their own right.

(e) As heritage of the philanthropist education of the Enlightenment, progressive education is concerned with a free relationship within educational interaction. This is especially manifest in teaching. The reformers invent, or at least accentuate newly, 'the teaching

dialogue, small group teaching, the various sorts of projects, (the) use of teaching aids, (they emphasize) self-governed activity…, (they aim at) a plurality of teaching methods, that (cause) as an "education by peers in the classroom" …a change in the direction of the activity' (Roehrs 1998, p. 319). The change of direction finds expression especially in those didactic models that transfer the planning of work and learning at least partly to the pupils. While acquiring and applying a basic knowledge on methodology the children prove their self-governing abilities.

3. *Progressive Education in Practice—Two Examples*

What internal reform of the school on progressive lines means in detail can be illustrated by the famous project method. This approach was theoretically explained by Kilpatrick (1918) and in the following decades practically spelled out worldwide in many variations (cf. Kliebisch and Sommer 1997). For Kilpatrick, a project was a 'wholehearted purposeful activity proceeding in a social environment' (Kilpatrick 1918, p. 330). This formal definition covers different project types which the educationist structures according to learning objectives and content matter, but he also gives hints to classify them into the categories of individual and group projects. A project has four phases: 'purposing, planning, executing, and judging.' The learners take all steps autonomously; the teacher only interferes if they are asked for advice or if the learning process is threatened by total failure. Later on, Kilpatrick embedded the project method into his philosophy of the school and of teaching procedures:

> (1) Count that the school is properly a place of living. Work for the finest quality of living that teacher and pupils can together devise, reaching out as much as feasible into the community… (2) Let the principles of interest and purposeful activity set the predominant pattern… (3) Run everything on a basis of democracy and respect for personality… (4) Work in season and out, depending on pupil maturity and development, for the finest and most inclusive aims of life, these to be increasingly exemplified in everything done in school and out.' (Kilpatrick 1959, p. 429)

Progressive education in its linkage to social work can be illustrated by the example of Jane Addams. Although she wrote several books and contributed articles to academic journals (e.g., the *American Journal of Sociology*), the focus of her activities was not the production of academic theory but practical political, educational, and social work efforts. These activities had their institutional basis in Hull House, a community center which Addams had set up in a Chicago slum area in 1889 after the model of the English 'settlements.' For 46 years the house under her leadership has been a center that has intervened in local educational policies to improve the educational situation for immigrant children, organized its own educational activities—such as a Shakespeare Club or a Working People's Science Club—provided meals for the poor, and improved hygiene standards by establishing public baths. Production and consumption cooperatives were operating; for some 'residents' common households were organized. Hull House built the first public playground in Chicago, and coinitiated the Juvenile Protection Association, which, through social activities and political pressure (e.g., on shop owners who sold alcoholic beverages to minors), tried to prevent youth delinquency. Addams and her friends succeeded in having a youth court set up that was no ordinary criminal court but acted *loco parentis*. The house was active in reducing child labor, and it fought against the 'white slave trade,' the 'import' of young women for prostitution. Hull House supported the trade unions, backing up the laborers in conflict situations, such as the big Pullman strike of 1894. The activities of Hull House finally contributed to the differentiation of sociology and social work theory and—against Addams's intentions—to transform social work into a profession (Eberhart 1995).

4. *History or Ongoing Movement*

Historians of education dispute the question whether progressive education was a historical movement in its time or whether it is a continuous ongoing stream of theories and activities. US theorists such as Sol Cohen (1995) tend to see it as history (and the US school reform efforts of the 1990s even as a countermovement to restore the moral-intellectual discourse of education), while the German contribution of Oelkers (1996) doubts whether there was an epoch of progressive education at all, and that of Roehrs (1998b) conceives of it as a continuing story. The overt history of progressive education in the USA ended in 1955 when the Progressive Education Association was dissolved. In fascist Italy and Germany progressive education with its liberty-oriented and democratic tendencies was politically banned, and some of its German proponents, such as Kurt Hahn, the director of the famous Salem School, were driven out of their home country. A few others, however, such as Andreesen, Lietz's successor in the leadership of the rural education centers, tried to establish links to Nazi ideology, thus enabling several measures of progressive education to be abused by totalitarian educational practice. In the USSR progressive education was suppressed by Stalin's dictatorship from the early 1930s. He even threatened Krupskaya with using another woman to 'replace' her as Lenin's widow, unless she stopped publishing about polytechnic education.

Conversely, the New Education Fellowship, as the organization of the progressivists, made important contributions to the establishment of UNESCO (United Nations Educational, Scientific, and Cultural

Organization) within the United Nations (UN) system after World War II. Many progressive education institutions have continued their activities into the present. More important, perhaps, is the fact that since the 1950s progressive ideas and practices have trickled down into the day-to-day activities of education worldwide. In this reduced form the movement is still a reservoir of innovation, and this is reflected in the alternative schools that have been established from the 1970s onward. The progressive movement has even made contributions toward meeting the most recent educational challenges, such as peace education, multicultural education, and human rights education. These must now be transformed creatively to build the 'four pillars of education' enumerated by the International Commission on Education for the Twenty-first Century: 'learning to know,' 'learning to do,' 'learning to live together, to live with others,' 'learning to be' (UNESCO 1996, pp. 86–96).

Bibliography

Boyd W, Rawson W 1965 *The Story of the New Education*. Heinemann, London

Cohen S 1995 The influence of progressive education on school reform in the USA. In: Roehrs H, Lenhart V (eds.) *Progressive Education Across the Continents*. Lang, Frankfurt, Germany, pp. 331–2

Connell W 1995 The influence of progressive education on Australian schools. In: Roehrs H, Lenhart V (eds.) *Progressive Education Across the Continents*. Lang, Frankfurt, Germany, pp. 351–6

Cremin L A 1961 *The Transformation of the School. Progressivism in American Education 1876–1957*. Knopf, New York

Dewey J 1966 *Democracy and Education. An Introduction to the Philosophy of Education*. Free Press, New York

Eberhart C 1995 *Jane Addams (1860–1935)*. Schaeuble, Rheinfelden, Germany

Kilpatrick W H 1918 The project method. *Teachers College Record* **19**: pp. 319–35

Kilpatrick W H 1959 *Philosophy of Education*. Macmillan, New York

Kliebisch U, Sommer P 1997 *Projektarbeit. Konzeption und Beispiele*. Schneider, Baltmannsweiler, Germany

Mead G H 1967 *Mind, Self, and Society*. University of Chicago Press, Chicago

Mialaret G, Chenon-Thivet M, Fabre A, François-Unger C, Gal R, Gloton R, Monod G, Seclet-Riou F 1976 *Education nouvelle et monde moderne*. Presses Universitaires de France, Vendôme, France

Oelkers J 1996 *Reformpädagogik. Eine kritische Dogmengeschichte*. Juventa, Weinheim, Germany

Piaget J 1963 *La naissance de l'intelligence chez l'enfant*. Delachaux Niestle, Neuchâtel, Switzerland

Roehrs H 1998a *Reformpädagogik und innere Bildungsreform. Gesammelte Schriften 12*. Deutscher Studienverlag, Weinheim, Germany

Roehrs H 1998b *Die Reformpädagogik. Ursprung und Verlauf unter internationalem Aspekt*. Deutscher Studienverlag, Weinheim, Germany

Roehrs H, Lenhart V (eds.) 1995 *Progressive Education Across the Continents*. Lang, Frankfurt, Germany

Schriewer J, Henze J, Wichmann J, Knost P, Taubert J, Barucha S 1998 Konstruktion von Internationalität: Referenzhorizonte pädagogischen Wissens im Wandel gesellschaftlicher Systeme (Spanien, Sowjetunion/Russland, China). In: Kaelble H, Schriewer J (eds.) *Gesellschaften im Vergleich. Forschungen aus Sozial- und Geisteswissenschaften*. Lang, Frankfurt, Germany, pp. 151–258

UNESCO 1996 *Learning: The Treasure Within 1996. Report to UNESCO of the International Commission on Education for the Twenty-first Century*. UNESCO, Paris

V. Lenhart

Projective Methods in Psychology

Projective techniques is a term that encompasses any test or procedure designed to increase insight into individuals by allowing them to respond freely to ambiguous stimuli. The underlying assumption is that, when faced with unstructured or ambiguous stimuli, people will reveal aspects of their personality in their attempt to structure the material. In other words, the vague quality of the stimulus evokes free fantasy play and triggers projection on the part of the subject while simultaneously revealing aspects of psychological functioning in the nature of their responses. Many experts believe that everything said or written in response to a stimulus has dynamic meaning and that people reveal conflicts, anxieties, attitudes, beliefs, and needs in their attempt to structure or complete an incomplete stimulus. Projective techniques tend to have brief, general instructions that, when coupled with ambiguous stimuli, tend to elicit less defensive responses than more structured testing. It should be noted that a concern about conflicts, anxieties, attitudes, beliefs, and needs, and the usual attempts to blend these into a picture of a person, is most closely associated with one approach to personality, the psychoanalytic approach. A preference for psychodynamic theory is not required in order to use projective techniques, but it is frequently found. Perhaps this is because one of the hallmarks of the psychoanalytic approach is the prominent role assigned to unconscious processes, and projective techniques are designed with the intention, satisfied or not, of eliciting unconscious material. (see *Psychoanalysis in Clinical Psychology*).

1. Historical Overview

Historically, the turn of the twentieth century was a time when psychological testing was applied predominantly for the study of intelligence. The trend was

toward finding data or norms descriptive of groups of people (nomothetic data). At this time, personality measures tended to examine only a single trait at a time (e.g., introversion/extroversion, flexibility). The scope began to broaden about two decades later when interest surfaced in tests that yielded specific data about an individual (idiographic data). In the early 1920s, projective testing emerged when Rorschach ([1921], 1942) developed his now famous inkblots in an attempt to aid in the diagnosis of schizophrenia.

The notion of projection with these tests gave birth to the progression of projective techniques throughout the rest of the twentieth century and the belief that ambiguous stimuli can be used to gain insight into the dynamics of the respondent. Frank (1939) coined the term 'projective methods' to describe a variety of techniques used by clinicians with ambiguous stimuli that pull for projection. He postulated a projective hypothesis, namely that the world is structured according to internal needs and external demands, so that ambiguous external stimuli will elicit structuring based on internal factors. The 'Projective Movement' grew tremendously during the 1940s and 1950s, as clinicians moved away from the nomothetic approach to studying their patients toward the more idiographic approach. More recently, because of psychometric considerations, and possibly because of a reduction in the number of clinicians trained in psychodynamic approaches, the use of projective techniques has waned, leaving an open question and considerable controversy over the value of the projective approach.

Perhaps the most commonly used projective techniques are the Rorschach, the Thematic Apperception Test (TAT), figure drawings, and sentence completion tests.

1.1 Rorschach

The use of inkblots was not a new concept in the 1920s. In fact, a popular game among children at the time was Blotto, in which children used bought or handmade blots of their own. Rorschach's clinical interest in the use of blots was sparked in 1911, while working with Eugen Bleuler, his professor and supervisor, who coined the term schizophrenia and instigated the need to differentiate schizophrenia from other forms of psychosis. Rorschach believed that the process of perception, as a respondent struggles to make sense of an ambiguous stimulus, reveals the threshold of each respondent to assimilate or integrate the new stimulus with existing mental images. Rorschach happened to observe in his work that schizophrenic patients tended to respond differently in Blotto than patients with other forms of psychosis or dementia, and decided to study this phenomenon more systematically. By 1921, he had administered the test often enough to believe in its usefulness as a diagnostic aid in identifying schizophrenia, but noted unexpectedly that increased frequencies of certain kinds of responses were consistent with certain personality traits or styles. Rorschach died in1922, but his successors saw the potential of his tool as an idiographic, projective measure and worked over the next two decades to develop a standardized scoring system for responses to the inkblots.

Samuel Beck was a psychology graduate student who attempted, shortly after Rorschach's death, to do a standardization study as a dissertation topic. He began his work at the Institute for Child Guidance in New York and later took positions at both Boston Psychopathic Hospital and Harvard Medical School. Over the years he gained extensive experience with the use of the inkblots and became acutely aware of the need for more intensive work toward standardization.

Around the same time, Bruno Klopfer came to Zurich to work as a technician at the Zurich Psychotechnic Institute where Rorschach inkblots were routinely used in the screening of candidates for employment positions. After working there for several years, he emigrated to the USA in 1934 and began working at Columbia University. The psychology graduate students at Columbia, who were frustrated by the lack of formal training available in the use of inkblots, were eager to have him share his experience with them. Klopfer began giving informal seminars to students and, in his work with these students, began to develop new scoring codes that seemed necessary, especially those for unusual locations or shading responses. Thus, by 1936, Klopfer was working hard to expand the existing coding system.

At the same time, Beck was busy writing papers that criticized the existing scoring systems for their lack of empirical soundness. He approached Klopfer's attempts to expand the system with criticism. Over the next 20 years, the arguments continued and different scoring systems were attempted. Despite this controversy over scoring, the method of using the Rorschach continued to burgeon. By 1957, five vastly different scoring systems had emerged.

John Exner began to study the difficulty with using inkblots in 1968. Between 1968 and 1972, he found that most clinicians were scoring the tests in a haphazard manner. Many did not score it at all, and those who did tended to use features of different scoring systems. Only a small minority used one system to score an entire test. Exner set out to develop a system that integrated features of all the available scoring systems that had empirically defensible data, and a system that was not wedded to any single personality theory. Thus, he developed the comprehensive system that is still used by many clinicians and that includes the successful work and research that accumulated from the 1920s onward. The system developed by Exner (1993) is probably the most widely used at the beginning of the twenty-first century, but many clinicians continue to treat the Rorschach purely as a projective device, rather than as a psychometric

instrument, and simply interpret the content of the responses according to their understanding of personality. It should be noted that there are several sets of ambiguous stimuli, usually inkblots, that are in use, but the Rorschach is by far the most popular.

1.2 Thematic Apperception Test

In 1935, Morgan and Murray introduced the Thematic Apperception Test (TAT; Murray 1943), a series of provocative pictures used to explore the respondent's dynamics. The stimuli are more highly structured than inkblots and respondents are required to give more meaningfully organized verbal responses. Respondents are asked to describe the scene in the picture, the apparent emotional content and thoughts of the people involved, what led up to the scene, and what the outcome will be. This test was developed on the premise shared by other projective techniques that people reveal something of their own personality (i.e., their motivational and emotional condition), as well as their cognitive structure, when ascribing thoughts and feelings to ambiguously drawn characters. Just as with the Rorschach, many sets of pictures have been developed for use as an apperceptive approach, but the TAT remains the most widely used.

Perhaps the most popular scoring method for this test is that devised by David McClelland (McClelland et al. 1953). McClelland's work was devoted to measuring achievement motivation and he scored the TAT responses across seven dimensions: (a) stated need for achievement (someone in the story expresses a desire to achieve a goal); (b) instrumental activity (whether the outcome of the instrumental activity is successful, doubtful, or unsuccessful); (c) anticipatory goal status (whether the person in the story anticipates achieving the goal or being frustrated by it); (d) obstacles or blocks (whether the goal in the story is blocked by an obstacle); (e) nurturant press (whether there are nurturing forces aiding the characters in goal achievement); (f) affective states (emotions present associated with achieving or being frustrated by the goal); and (g) achievement themes (when achievement imagery is the focus of the story).

TAT responses have been studied and scored in a variety of ways, with McClelland's approach being only one example. By far the easiest and most commonly used method among clinicians is simply to read the responses, look for common themes, treat the stories as psychologically meaningful communications, and consider them as projective tools for understanding the patient's underlying conflicts and needs.

1.3 Figure Drawings

Figure drawings are a series of drawings by the respondent that the clinician uses to understand nonverbal expressions of the self. Respondents are asked to draw a full person (gender not specified ahead of time), and then to draw a person of the opposite sex of the first drawing. The absolute and relative size of the persons are examined as well as details such as the position on the page, the quality of the lines, the sequence of parts drawn, the stance, position of the arms, and grounding effects. Clinicians have often found these drawings helpful in understanding the self-perception, motor expression, anxieties, and needs of the respondents. Variations of this task ask the respondent to produce other drawings, including drawings of objects such as a house or a tree. In general, regardless of what is drawn, the research with these approaches has not been very encouraging (Swenson 1968).

1.4 Sentence Completion

Sentence completion is another method for gathering information about personality dynamics. Respondents may be asked, orally or in writing, to complete sentences that are vague and ambiguous (e.g., What worries me … ; Women … ; My mother …). In addition, there have been standardized written sentence completion tests developed by authors such as Rotter (Rotter and Rafferty 1950). As with the TAT and other apperceptive measures, the value of the task often is dependent on the particular set of items that is employed, and the more closely the items approximate the target being studied, the more valuable the responses can be shown to be.

2. Psychometric Considerations

A psychological test or assessment device is subject to evaluation on standard dimensions and, regardless of how compelling a rationale may be, must demonstrate that it meets the usual psychometric criteria before it can be considered to be useful. Foremost among these criteria are reliability and validity (for a more detailed explication of these critical concepts, see Wiggins 1973).

2.1 Reliability

Reliability refers to the extent to which a test provides a consistent measure of what it is measuring. If a test lacks reliability it is akin to using a rubber ruler, providing different measurements each time it is applied, and there is little credence that can be given to any of those measurements. One approach to the measurement of reliability is through retest, with the expectation that a test given on two occasions will produce the same response on both testings. A second approach is internal consistency, which assesses the extent to which every item in a test is measuring the same characteristic.

Although most projective techniques employ many items, the items often are designed to be responsive to different dimensions, and so a projective test may often be seen as a collection of disparate single items rather than a collection of redundant items. Unfortunately, in measurement, redundancy is often desirable, and it is well demonstrated that single items are subject to far more unreliability than collections of items, in which errors may cancel each other out. If there are indeed collections of single items, it can be argued that the test really is a collection of tests, each of which measures a single dimension (and therefore requires a separate assessment of reliability and validity), and does so in a way that reduces the reliability of each. A proponent of these tests would counter that the object being measured is personality, which contains several dimensions, and that the items should converge to produce a clear picture of the multidimensionality of the respondent. Whether or not this is so is a matter of validity.

2.2 Validity

Reliability is necessary but not sufficient for a test. An unreliable test cannot be valid, but a reliable test may not necessarily be valid. Validity is an assessment of the extent to which a test measures what it is intended to measure. This leads to a very interesting problem in assessing the validity of projective techniques. Most projective techniques purport to measure many different aspects of the person, and it is necessary to indicate the validity of every inference that a test draws. Thus, rather than asking, for example, whether the Rorschach is valid, it is necessary to ask, for each inference drawn from the Rorschach, whether or not that inference is valid. One statistical estimate cannot represent the validity of the test, and the general question cannot be answered (Stricker and Gold 1999).

It is not unusual for a research project to compare the validity of projective techniques with more objective, self-report, measures of personality (e.g., McClelland et al. 1989). The usual finding of such studies is that the two approaches do not correlate with each other. Some people have interpreted this as indicating a failure in the validity of the projective approach, a conclusion that assumes the validity of objective tests. Others have determined that this lack of correspondence indicates that each approach is valid, but that they are measuring different aspects of the person. This view suggests that objective tests are superior at measuring objective criteria, such as behavior, whereas projective approaches are superior at measuring more covert, unconscious aspects of the person (Hiller et al. 1999). This interpretation is consistent with the theoretical position that usually underlies each of the approaches to measurement.

It also is reasonable to ask whether the incorporation of a projective approach to a test battery adds anything to the process of assessment (incremental validity). If the conclusion about measuring different aspects of the person is accepted, then projective approaches do contribute incremental validity insofar as the assessor is interested in less overt aspects of the respondent. Given the length of time necessary for the administration and interpretation of projective techniques, the question of cost effectiveness also can be raised. The answer to this question probably is dependent on the individual circumstance. It should be clear, however, that validity, incremental validity, and cost effectiveness are separate questions, and the answer to one does not determine the answer to the others.

Finally, it should be noted that the subjectivity of the projective approach places great stress on the skill of the examiner, and any question of validity is equally a function of the ability of the approach to elicit material and the ability of the examiner to interpret it.

3. Concluding Comments

It can be concluded, with some certainty, that the original promise of the projective hypothesis (Frank 1939) has not been sustained, and that there are many limitations on the conclusions that can be drawn from ambiguous stimuli. However, there are instances in which such stimuli can be used to produce valid inferences, and inferences that cannot be drawn from other approaches to measurement. Whether these inferences are useful depends on the theoretical approach and goals of the examiner, whose skillfulness also must be considered in reaching a conclusion about the utility of projective methods. Because projective techniques do show the promise of providing valid information that is different than that produced by objective measures, it is a path worth pursuing. However, any pursuit will have to be marked by attention to psychometric considerations, a neglect of which has led to the excessive promise of past attempts.

See also: Bleuler, Eugen (1857–1939); Content Validity; Measurement Theory: History and Philosophy; Mental and Behavioral Disorders, Diagnosis and Classification of; Personality Assessment; Personality Theories; Personality Theory and Psychopathology; Psychoanalysis in Clinical Psychology; Psychology of the Real Self: Psychoanalytic Perspectives; Psychometrics; Reliability: Measurement; Unconscious: History of the Concept

Bibliography

Beck S J 1950 *Rorschach's Test. I. Basic Processes*, 2nd edn. Grune and Stratton, New York

Exner J 1993 *The Rorschach: A Comprehensive System*, 3rd edn. Wiley, New York, Vol. 1

Frank L K 1939 Projective methods for the study of personality. *Journal of Psychology* **8**: 389–413
Hiller J B, Rosenthal R, Bornstein R F, Berry D T R, Brunell-Neuleib S 1999 A comparative meta-analysis of Rorschach and MMPI validity. *Psychological Assessment* **11**: 278–96
Klopfer B, Ainsworth M D, Klopfer W G, Holt R R 1954 *Developments in the Rorschach Technique. Vol. 1, Technique and Theory*. World, New York
McClelland D C, Atkinson J, Clark R, Lowell E 1953 *The Achievement Motive*. Appleton-Century-Crofts, New York
McClelland D C, Koestner R, Weinberger J 1989 How do self-attributed and implicit motives differ? *Psychological Review* **96**: 690–702
Morgan C D, Murray H A 1935 A method for investigating fantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry* **34**: 289–306
Murray H A 1943 *Thematic Apperception Test Manual*. Harvard University Press, Cambridge, MA
Rorschach H [1921]1942 *Psychodiagnostics: A Diagnostic Test Based on Perception*, 4th edn. Grune and Stratton, New York
Rotter J B, Rafferty J E 1950 *The Rotter Incomplete Sentences Test*. Psychological Corporation, New York
Stricker G, Gold J R 1999 The Rorschach: Toward a nomothetically based, idiographically applicable configurational model. *Psychological Assessment* **11**: 240–50
Swenson C H 1968 Empirical evaluations of human figure drawings. *Psychological Bulletin* **70**: 20–44
Wiggins J S 1973 *Personality and Prediction Principles of Personality Assessment*. Addison-Wesley, Reading, MA

G. Stricker and K. Somary

Pronouns

Noun referents can be substituted by personal pronouns. In most contexts the pronoun is coreferential with and interchangeable with the noun to which it refers. Other pronouns determine nouns, some have the capability of doing both. In English, the personal pronouns substitute, the articles determine, and the demonstratives do both. These, together with indefinites, negatives, locatives, interrogatives, and relatives are grouped together as non-personal pronouns.

The word *tree* refers to an object in the real world (exophoric use), but *it* can have meaning only in referring to some noun such as *tree* by being coreferential with it (endophoric use). Within the linguistic context *it* can replace *tree* or any other noun of the same class—but not, for example, *man* or *trees*, which require special pronominal forms in English.

Pronoun sets can also be distinguished on the basis of case and emphatic differences. Not all pronouns are in fact interchangeable with nouns, such as the reflexives, the logophorics, and the sequentials. Pronouns may be morphologically free or bound, that is, they may be separate words or affixes.

Instead of treating pronouns as part of the lexicon, Everett (1996) argues for considering them to be composed of phi-features, such as case, gender, number, tense, and person. These phi-features are stored in the lexicon, their phonological shapes are given by postlexical or precompiled spell-out rules. (see *Clitics, Linguistics of*; *Morphology in Linguistics*; *Word, Linguistics of*.)

1. Personal Pronouns

The personal pronouns can be studied under the parameters of person, number, and gender distinctions, which vary considerably from one language to another.

1.1 Person

In the person parameter, the third persons (*he/she/it/they*) are usually used endophorically, whereas the first and second persons (*I*, *we*, *you*) are frequently used exophorically.

The only language with person but no number distinctions reported to date is Mura Pirahã, an unclassified Amerindian language (Brazil). First, second, and third constitute the only formal pronoun distinction this language has. One language has been found which appears to distinguish in the pronominal system two persons only: first and nonfirst. In the *Balante* variety spoken in Guinea-Bissau, *N*-(homorganic nasal) is used for 1sg, *ha*- for 2sg or 3sg, and *be*-for plural in any of the three persons.

Many languages have pronouns that combine 1 + 2 persons; that is, speaker and hearer, into an *inclusive* first person. This is in contrast to the 1 + 1 or first person *exclusive* category, in which the hearer is excluded explicitly. Some languages combine other categories as well, such as 1 + 3—that is, speaker plus one person not present; 2 + 3—grouping together one hearer and one person not present; and 3 + 3—grouping together two persons not present. In addition to binary combinations, the tertiary combination of 1 + 2 + 3 is sometimes used, where everybody is included explicitly (particularly the ancestors).

1.2 Number

The person distinctions are usually combined with number distinctions, singular and plural being the most common. When person categories are combined, making the smallest possible number two, or even three, it is more sensible to talk about a minimum-plural distinction. The most complex person-number combinations so far reported are found in Ghomala'

Table 1
Ghomala' subject pronouns

Person		Minimum	1pl	2pl	3pl
	1	gə́	pyə́		
Mono	2	o		po	
	3	e			wap
	1+2	pu		pə	
Double	1+3	pyə́	pyayú		pyəapu
	2+3	poé		poayú	poapu
	3+3	pué			wap
Triple	1+2+3	pəayú	pəayú	pəayú	wəkpampu

(Bameleke, Cameroon). The plural categories have hierarchical order: first person comes before second, which comes before third.

Sursurunga (New Guinea) has what is probably the most complex number system in operation with mono, dual, triple, and quadruple, as well as four-person distinctions (Hutchisson 1986).

1.3 Gender

Gender distinctions are basically of two kinds: natural gender, such as masculine, feminine, and neuter (or animate/inanimate, human/non-human, for example), and grammatical gender having to do with noun classes. The latter is common in African and some Asian languages. Normally, gender distinctions are restricted to the third person, but may be found in other persons as well, particularly in possessive function.

1.4 Honorific Use

Many languages have special uses of pronominal forms for showing respect or social distance. Instead of the second singular, the plural may be used. Completely different pronouns can also be found. For Bangkok Thai, 17 different forms for first, 19 for second, and 10 for third person are in use, distinguished by such categories as power and status, kinship, friendship, occupation, ethnic–religious groups, age, sex, genealogical distance, solidarity, formality, presence of certain persons, emotional manifestations, and combinations of all these (Palakornkul 1975).

1.5 Special Meanings

There may also be unusual uses of person distinctions. Exclusiveness may be an added dimension to any of the personal pronouns, though there are no special forms to mark the distinction. In several Amerindian languages the 1sg has the exclusive meaning of 'only I (speaker), but not anybody else,' whereas the second means 'only you (hearer) but not I' (speaker). The only pronoun which can be used inclusively is the 1pl form.

Other pronouns can have meanings which are not apparent from examining the system. In Kaingang (Je, Brazil) whenever a family unit is involved, the 3pl fem. form is used after the name of the man; the 3pl masc. is reserved for a group of men only (see *Classifiers, Linguistics of*; *Grammatical Number*; *Language and Gender*; *Language and Social Class*; *Numeral Systems*; *Politeness and Language*).

2. Nonpersonal

Nonpersonal pronouns are used in the third person only, but may vary in number and gender.

2.1 Definite

Demonstratives and articles are included in the series of definites. A distinction is often found according to where the determined noun is placed in relation to the speaker and/or the hearer, which gives a two- or three-way distinction, as when the locative adverbs are used pronominally.

In English:
(a) *the* man is anaphoric: the already mentioned;
(b) *this* man is close to speaker (often anaphoric);
(c) *that* man is far from speaker (often anaphoric);
(d) *this* makes me wonder if … .
This and *that* may be used as substitutes (exophorically) if the object is in sight (and in that case not necessarily anaphorically) or to replace arguments and other parts of discourse. In German the articles may substitute for the noun.

In Didinga (Nilo-Saharan, Sudan) separate forms exist for information that is generally known as against

exclusive knowledge of the speaker; or whether the pronoun refers to a happening long ago or recently, or whether the person was referred to recently or much earlier.

2.2 *Indefinite*

A number of languages also have indefinite pronouns, referring to no one in particular, but rather to everyone or certain unspecified individuals, such as *they* in English or *on* in French. The second person, singular or plural, might be used in this way also: *you don't beat children!*

Among the indefinite forms can be classed such words as *a*, *certain*, *other*, *such*, *few*, *some*, *many*, *any*, and others. Some languages class the distributives among the indefinite, such as *all*, *every*, *each*, *both*, *half*, *part of*, *some*, *another*, *other*, *either*, and others.

2.3 *Negative*

Certain forms are entirely negative. They are often subsumed under indefinite. In English they include *none*, *nobody*, *nothing*, *no*, *neither*. In many languages the negation is not expressed in a pronoun (or noun) but rather in the verbal expression.

2.4 *Interrogative*

In all languages, people ask questions with the aid of interrogative substitutes or determiners. Some of these are simple, others are constructed from simple forms. In English, the simple forms are *how? what? when? where? which? who? whom? whose? why?*, constructed forms *how many? whatever?*, and others.

Some languages form interrogatives from demonstratives, such as:

Fulfulde (Cameroon)

(a) dewtere nde'e 'this book' (demonstrative)
(b) dewtere ndeye 'which book? (interrogative)
(c) gorko o'o 'this man' (demonstrative)
(d) gorko oyo 'which man?' (interrogative)

2.5 *Relative*

Relative constructions usually contain a relative pronoun, which may be a demonstrative or an interrogative, or an entirely different form, restricted in use to relative constructions. English: 'Here is the one *who* could help.' In Konzimé (Cameroon), the pronoun comes at the end of a relative construction and can be:

- either a special relative pronoun which varies according to the noun classes; or
- the demonstrative *nɣ* which is invariable.

In some languages the relative pronouns also indicate differences in tense and form part of the verbal paradigm:

Denya (Ekoid, Cameroon)

(a) à-wá 'he killed' – á-wà-né 'he who killed'
(b) à-wà-nè 'he kills' – á-wá-nè 'he who kills'

2.6 *Pronominally Used*

A number of adverbs have pronominal use, especially the indicators of space and time. The locative adverbs indicate at least three different localities: close to speaker, close to hearer (but far from speaker), and far from both, as in English *here—there—yonder*. There may also be other types of information, such as *inside*, *outside*, *underneath*, *above*, *before*, *after* that could be counted among them.

Nouns can also be used pronominally, especially (but not only) as honorifics. In Brazilian Portuguese, *a gente* 'the people' is commonly used instead of *nós* 'we'. (see *Adverbial Clauses*; *Anaphora*; *Grammaticalization*; *Negation, Linguistics of*; *Quantifiers, in Linguistics*; *Relative Clauses*).

3. *Set Distinctions*

In noun phrases and clauses, pronouns may be marked for any case distinction found in the language. Case indicators including circumstantial information may be fused to the pronouns causing different sets. English distinguishes *I–me*, *he–him*, *she–her*, *we–us*, *they–them*.

Yaouré (Ivory Coast) shows pronoun fusion with postpositions (Hopkins 1986):

í + *lɛ̄* → *jíɛ́* '2sg + of → of you sg'
ká + *lɛ̄* → *kɛ́* '2pl + of → of you pl'

3.1 *Genitive*

In genitive constructions the possessor can be replaced by a pronoun. The possessed item often takes the mark of the clause level function, whereas the mark on the possessor identifies it as such by the genitive case. The agreement rules of the associative noun phrase can be quite complex and varied.

English:

(a) John's book – *his* book (agrees with John)
(b) Mary's book – *her* book (agrees with Mary)

French:

(a) le livre de 'the book of
Jean – *son* livre John – his (masc) book'
(agrees with book)

Table 2
Kabiyè *subject + mode*

	Singular	Plural	Singular	Plural
1	mɔ̀n	dÍ	máá	dÍÍ
2	ǹ	*ɛ́*	ńń	*ɛ́ɛ́*
3	*ɛ̀*	pà	*ɛ́ɛ́*	páá

Table 3
Rendille (Kenya) *subject + aspect*
(verb root: táx-, *-e* marks perfective, *-a* imperfective)

	Perfective	Imperfective	Person markers
1sg	táxe	táxa	
1pl	táxne	táxna	-n
2sg	táxte	táxta	-t
2pl	táxten	táxtan	-t..-n

(b) la case de Jean – *sa* case (agrees with hut) 'the hut of John – his (fem) hut'
(c) les livres de Jean – *ses* livres (agrees with books) 'the books of John – his (pl) books'

African class languages show a wealth of pronominal forms in all the cases. In the genitive, agreement holds in first and second person as well. The agreement may be with the possessor (as in English) or with the possessed item (as in French), or with both.

There may be different kinds of genitives, such as alienable and inalienable possession. The differences may be marked in the pronoun sets:
Lendu (Zaire)
(a) *mà* *ndzá* *búkù* 'my book' alienable/possessive
my of-cl. book
(b) *nì* *dhí* *búkù* 'your book'
your of-cl. book
(c) *mā* djō 'my brother' inalienable/qualificative
my brother
(d) *n'ī* djō 'your brother'
your brother

In the alienable construction, the tone of the possessor pronoun is different from the inalienable construction, and the alienable particle 'of' (which changes in form according to noun class) is absent in the inalienable construction.

3.2 *Noun Phrase Substitution*

The whole possessive noun phrase can be substituted by a pronoun or a nominalized or adjectivized pronoun, which functions on clause level and may carry case markers, as in German:
(a) Mutters Garten – *ihr* Garten – *ihrer* – der *ihrige*
(b) *mein* Garten – *meiner* – der *meinige*

Table 4
Attie (Ivory Coast) *subject + mode + aspect*

	Perfective	Imperfective	Negative Perfective	Subjunctive
1sg	mɛ̃̀	mɛ̃	mɛ̃	mɛ̃
2sg	bù	bɔ̀	bù	bú
3sg	ò	wɔ̀	ò	ó

Table 5
Mundani (Cameroon) emphatic pronouns (simplified)

Person	1. Degree	2. Degree	Authoritative
1sg	m̀mɤ̀	m̀mɤ̀ m̀bɔ̀ŋ	m̀mɤ̀ m̀bɔ̀ŋ
1pl	bàá	bàá m̀bɔ̀ŋ	bàá tà m̀bɔ̀ŋ

3.3 *Subject + Distinctions*

Subject pronouns may contain, in addition to person, number, and gender information, a case marker, combined with verbal information such as mode—as illustrated from Kabiyè (Togo), or aspect as illustrated from Rendille (Kenya), or a combination of these, as illustrated from Attié (Ivory Coast).

The Kabiyè irrealis pronouns seem to be a fusion of the first series plus a suffix *-ka*, which is seen in its full form when following nouns.

The Attie imperfective can be reconstructed as a fusion of the perfective plus a suffix *-a*. In the negative perfective series, only two forms are different from the perfective: the 1sg and 2pl forms have high instead of low tone. Nevertheless, the whole set constitutes a new series. The subjunctive forms all have high tone.

3.4 *Emphatic*

Emphatic pronouns can be found in all languages, especially in subject function. These are usually focus forms, which can also be differentiated by degree. All of these differences can result in different pronoun series.

In Bafut (Grassfields, Cameroon), in addition to the location in insistence distinction adds members to the demonstrative set. These distinctions go from 0, the 'normal' form, to 3, where the people are actually shouting at each other. Since Bafut is a class language, the whole set includes corresponding forms for each class; class 5 forms are presented here. The location distinctions are: *nû* 'close to speaker'—*nî* close to hearer'—*nyáa* 'far from both,' and all have corresponding forms:
(a) nìbá nû 'this (here) wing' (0)
(b) nìbá núà '(no,) *this* (here) wing!' (1)
(c) nìbá núlà '(no, no,) *this (here)* wing!' (2)
(d) nìbá núànuù '(of course not!) *this (here) wing*!' (3)

Table 6
Kobon SS/DS pronouns

	SS			DS		
Person	Singular	Dual	Plural	Singular	Dual	Plural
1	-em	-ul	-un	-nö	-lo	-no
2	-mön	-mil	-mim	-ö	-lö	-be ∼ -pe
3	-öm	-mil	-öm	-ö	-lö	-ö

(see *Grammatical Agreement*; *Possession (Linguistic)*; *Serial Verb Constructions*; *Sign Language*; *Tense, Aspect, and Mood, Linguistics of*).

4. *Noninterchangeable*

Some pronouns are not interchangeable with the nouns with which they are coreferent. *John hurt himself* cannot be replaced by **John hurt John* without changing the meaning.

4.1 *Reflexive and Reciprocal*

Reflexive pronouns are used to show that the agent and the patient of an event are coreferent:
(a) John$_i$ hurt himself$_i$
(b) John$_i$ and Mary$_k$ hurt themselves$_{i+k}$
(c) John$_i$ and Mary$_k$ hurt each other
/one another$_{i<->k}$
If more than one participant is involved, and in the course of the action they exchange roles, the construction is reciprocal.

The relationship can hold between the subject (or agent) and the possessor:
(d) John$_i$ watched his (own)$_i$ son reflexive
(e) John$_i$ watched his$_k$ son not coreferent

4.2 *Logophoric*

Logophoric indicators, usually pronouns, are found in many African languages. Their primary use is in indirect reported speech. They express coreference between the quoted speaker and any participant in the quote content, but primarily the subject or agent.
Tupuri (Cameroon)
(a) Pɔl$_i$ riŋ Jaŋ$_k$ gá *sɛ́*$_i$ lɛ'ɛ
Paul$_i$ said to-John$_k$ that LOG$_i$ fell
'Paul said to John that he (Paul) fell'
(b) Pɔl$_i$ riŋ Jaŋ$_k$ gá *a*$_{k,l}$ lɛ'ɛ
Paul$_i$ said to John$_k$ that he$_{k,l}$ fell
'Paul said to John that he (John or someone else) fell'

In all the languages that have such a logophoric indicator, the relationship holds between the quoted speaker and the subject of the quote content. Some languages use the same particle to show coreference between the speaker and other functions in the quote content, such as direct object, indirect object, or possessor.

The logophoric indicator is normally used in the third person, always singular, but sometimes plural as well. Some languages also use it in the second person, and in Lélé (Chad) it is used for all three persons, singular and plural. In languages with nominal classes, the marker may be distinguished according to the class of the noun to which it refers (Wiesemann 1986).

Logophorics are used with certain verbs related to speaking, especially the equivalent of 'say'. Verbs of thinking and feeling may also permit their use. In some languages, the logophoric pronouns can also be used with any verb, in which case the clause is one of intention or cause, which can be interpreted as truncated indirect speech.
Babungo (Grassfields, Cameroon)
ŋwɔ́ nyíŋ lāā Kɩ́ vǎŋ sáN *yí* mē
3s$_i$ flee that NEG 3p$_k$ beat LOG$_i$ not
'He fled so they would not beat him'

4.3 *Sequential*

In a sequence of clauses one may be the principal and the others subordinate to it. This relationship can be made explicit by the form of the subject pronoun. Any dependent clause in which the subject is coreferential with that of the principal one, the same subject (SS) is marked by a special sequential form of the pronoun. Or the converse may be true: any subject that is not coreferent with that of the main clause, the different subject (DS) is marked in some way. This relationship is also called switch-reference (Munro 1979).
Bafut (Grassfields, Cameroon)
(a) *á*-ghɛ́ɛ̀ ɩ̄ndá ŋ̲́-kw ɛ́rɔ́ fōrə
3s$_i$-go house SS$_i$-take rat
'… then he went home and caught a rat.'
(b) *á*-ghɛ́ɛ̀ ɩ̄ndá ŋ̲́-kw ɛ́rɔ́ fōrə
3s$_i$-go house gwa 3s$_k$-take rat
'… then he went home and gwa caught a rat'
In Bafut, the SS–DS difference is restricted to the third person subject. In Kobon (Papua New Guinea) all the persons have a different form for SS and for DS, as well as a 'normal' pronoun form. The actual use of such forms can be illustrated by:

kobon (Papua New Guinea)
(a) kabö mɨjanu lau-*ö* in-a
stone that heat-DS3sg hot-3sg
'he heated the stones and they were hot'
(b) yad kaj pak-*em* ram ud ar-nab-in
1sg pig strike-SS2sg house take go-FUT-1sg
'I will kill a pig and take it home.'

Reflexives, Logophorics, and Sequentials can all be marked in the same language, either by the same markers or by different ones. In Mundani (Grassfields, Momo, Cameroon) the sequential pronoun is a homorganic nasal *N* in imperative and realis mode, and *e* in irrealis mode, the logophoric pronoun is *yè*, and the relfexive *vi*. They work together to keep the participants in a text straight (Parker 1986).

See also: Generative Syntax; Language and Gender; Linguistics: Incorporation; Serial Verb Constructions; Word Classes and Parts of Speech; Word, Linguistics of

Bibliography

Delisle G 1972 Three universal personal pronoun features. *Stanford University Working Papers in Language Universals* **9**: 181–3

Everett D L 1996 *Why There Are No Clitics. An Alternative Perspective on Pronominal Allomorphy*. The summer Institute of Linguistics and The University of Texas, Arlington, TX

Hopkins E B 1986 Pronouns and pronoun fusion in Yaouré. In: Wiesemann U (ed.) *Pronominal Systems*. Gunter Narr Verlag, Tübingen, Germany

Hutchisson D 1986 Sursurunga pronouns and the special uses of quadral number. In: Wiesemann U (ed.) *Pronominal Systems*. Gunter Narr Verlag, Tübingen, Germany

Ingram D 1978 Typology and Universals of Personal Pronouns. In: Greenberg J H (ed.) *Universals of Human Language*. Stanford University Press, Stanford, CA

Munro P (ed.) 1979 *Studies in Switch Reference*. University of California, Los Angeles, CA

Palakornkul A 1975 A sociolinguistic study of pronominal usage in spoken Bangkok Thai. *Linguistics* **165**: 11–41

Parker E 1986 Mundani pronouns. In: Wiesemann U (ed.) *Pronominal Systems*. Gunter Narr Verlag, Tübingen, Germany

Wiesemann U 1986 Grammaticalized coreference. In: Wiesemann U (ed.) *Pronominal Systems*. Gunter Narr Verlag, Tübingen, Germany

U. Wiesemann

Property: Legal Aspects

Property encompasses the governance of resources—real property (land), natural resources, and personal property (chattels and intangibles, including copyrights, patents, stocks, claims of right). Natural law theory, beginning with Aristotle and perhaps best articulated in the writings of John Locke (1698), posits that 'life, liberty, and property' exist naturally as a matter of fundamental principles of justice, irrespective of the state. Echoing Locke's formulation, the United States Constitution provides that a person may not 'be deprived of life, liberty, or property, without due process of law; nor shall private property be taken for public use, without just compensation.' By contrast, legal positivists, such as Jeremy Bentham (1789), assert that property rights arise only to the extent that they are formally recognized by the state: 'Property and law are born together, and die together. Before laws were made there was no property; take away laws, and property ceases.' Both views give content to what we understand as property. Legal positivism captures the practical reality that legislatures and courts define property rights in most societies, and that property rights have practical meaning only to the extent that they are enforceable. Nonetheless, natural law (and other philosophical frameworks) inform the content of property rules by providing the justification for state-created rights.

1. Philosophical Justifications for and Critiques of Private Property Rights

Waldron (1988) suggests that property is best viewed as a 'system of rules governing access to and control of material resources.' These rules are often characterized as a bundle of rights—in the case of land, the rights to possess (or occupy), to exclude, to use, to transfer, to enjoy fruits, and to alter or destroy—which can be allocated in different ways. Particular bundles represent different conceptions of property, which derive from one or more philosophical justification.

1.1 Labor Theory

John Locke (1698) offered a strong natural rights justification for private property, which remains a central pillar of property theory today. Beginning with the proposition that all humans possess property in their own 'person,' Locke argued that

> [t]he 'labour' of his body and the 'work' of his hands, we may say, are properly his. Whatsoever, then, he removes out of the stat[e] that Nature hath provided and left it in, he hath mixed his labor with it and joined to it something th[a]t is hi[s] own and thereby makes it his property. It being by him removed from the common state Nature placed it in, it hath by this labour something annexed to it that excludes the common right of other men. For this 'labour' being the unquestionable property of the labourer, no man but he can have a right to what that is once joined to, at least where there is enough and as good left in common for others.

Locke's theory is widely viewed as justifying first possession as a basis for acquiring property rights (Rose 1985, Epstein 1979). While useful as a basis for establishing rights in land and raw materials in a world of abundant unclaimed resources, Locke's theory significantly qualifies the laborer's claim in a world of scarcity—requiring 'enough and as good left in com-

mon for others.' Waldron (1988), Rose (1987), Ryan (1984), Becket (1977), and Schlatter (1951) elaborate and critique Locke's justification for private property rights.

1.2 Libertarianism

Drawing upon Lockean principles, libertarians emphasize individual freedom to control and alienate their private property. Nozick (1974) asserts that the state should play no role in distributing or redistributing property acquired properly, apart from respecting the voluntary transfers of property owners. Epstein (1985) argues for a broad interpretation of the government's obligation to compensate owners of private property for regulation and other governmental decisions interfering with private property rights.

1.3 Utilitarianism

First articulated by Jeremy Bentham (1789), utilitarianism judges the allocation of resources and the design of governance institutions on the basis of the utility or individual satisfaction that they produce. In applying this theory to property rights, utilitarian scholars have focused principally upon the inefficiency that can result from incomplete or shared resource ownership. In a now classic work, Hardin (1968) describes the 'tragedy of the commons,' whereby regulated access to a common pasture (or any other common resource such as air, water, or roads) results in overuse. Hardin suggested that private property or government regulation is needed to govern the commons. Economists generally prefer private property institutions (Posner 1998), although the efficacy of private property depends critically upon the costs of creating, monitoring, and enforcing such rights (Lueck 1995, Anderson and Hill 1975, Demsetz 1967). Government regulation often entails direct costs as well as inefficiencies attributable to rent-seeking by those interests affected by regulation. Although private property regimes often ameliorate the tragedy of the commons, shared governance institutions can maximize utility in a range of circumstances depending upon the cohesion and history of the community affected and the nature of the resources in question (Libecap 1994, Ellickson 1991, Ostrom 1990, Rose 1986).

1.4 Personhood

The personhood justification for private property, derived from Kant's Philosophy of Law and Hegel's Philosophy of Right, has been elaborated in modern legal discourse in the work of Radin (1982, 1993). 'The premise underlying the personhood perspective is that to achieve proper development—to be a person—an individual needs some control over resources in the external environment. The necessary assurances of control take the form of property rights' (Radin 1982). The personhood justification for private property emphasizes the extent to which property is personal as opposed to fungible: the justification is strongest where an object or idea is closely intertwined with an individual's personal identity and weakest where the 'thing' is valued by the individual at its market worth. The personhood justification for private property is limited and highly subjective (Schnably 1993).

1.5 Distributive Justice

Theories of distributive justice seek to allocate society's resources on the basis of just principles. The process of determining such principles is the focus of considerable debate. Many theories incorporate utilitarian and Lockean principles. In perhaps the most influential work on distributive justice of the past century, Rawls (1971) offers an 'ideal contractarian' theory of distributive shares in which a just allocation of benefits and burdens of social life is determined by what rational persons would choose from behind a 'veil of ignorance,' which prevents them from knowing what abilities, desires, parentage, or social stratum they would occupy. Rawls concludes that people behind such a veil would adopt what he calls the 'difference principle': 'primary goods'—not only wealth, income, and opportunity, but also bases of self-respect—would be distributed to the maximal advantage of a representative member of the least advantaged social class. Rawls (1993) emphasizes that the difference principle does not call for simple egalitarianism but rather measures to assure that 'the basic needs of all citizens can be met so that they can take part in political and social life.'

1.6 Civic Republicanism

Republicanism rejects the view that individuals are wholly autonomous and that the public interest is just the aggregation of individual preferences. Republicanism stresses civic virtue and the role of participation in public discourse as essential to the realization of the common good. Civic republicanism relies upon the propertied independence of its citizens. Such ownership of capital affords citizens the competence to engage in meaningful discourse and protects against the subversion of democracy by a wealthy oligarchy or an aggressive state (Simon 1991).

1.7 Marxist Theory

Karl Marx viewed human history as an evolutionary process of class struggle leading, ultimately, to a classless society (communism) in which the state will wither away. In the capitalist phase of this evolution, private property provides the basis for the most important social relation: ownership of the means of

production. According to Marx, the class owning the means of production (bourgeoisie) acquires the property (capacity to work) of the labor class (proletariat) through the employment relation, thereby obtaining exclusive control of and hence power over the worker class. Such subordination, exploitation, and alienation sow the seeds of revolution, leading to the establishment of a proletarian dictatorship, socialism (productive resources owned by workers through participatory institutions), and eventually communism, a classless society without exploitation or inequalities in which property in land and rights of inheritance are abolished (Marx and Engels 1848).

1.8 Ecological Theories

As understanding of the connections between technology, land development, and the environment has deepened, many environmental scholars have increasingly viewed private property rights more skeptically (Sax 1983). Traditional reformist theories—which are anthropocentric in nature and draw heavily upon utilitarian and other traditional philosophical frameworks—favor internalization of the adverse impacts of development (Menell and Stewart 1994), Ellickson (1977, 1973). Building upon Leopold's (1949) call for a land ethic which seeks to 'preserve the integrity, stability, and beauty to the biotic community,' Freyfogle (1996), Sprankling (1996), and Sax (1993) propose a shift away from private property toward greater regulation and coordination of land and natural resource development. At the other extreme, Anderson and Leal (1991), drawing upon property rights and public choice literatures, advocate complete reliance on private property and free market principles as the best institutional choice to address substantially all environmental problems. Menell (1992) highlights the institutional limitations of both extremes and urges careful attention to the strengths and limitations of the broad spectrum of institutional alternatives in governing resources.

1.9 Pluralist and Hybrid Theories of Property Rights

The insights of the many philosophical perspectives have led some scholars to synthesize multiple strands into new and comprehensive theories for justifying and evaluating property. Munzer (1990), for example, constructs a theory of property built upon principles of utility and efficiency, justice and equality, and rewards based on labor. Simon (1991) synthesizes civic republicanism and socialism to advocate an intermediate form of property emphasizing enterprise and community-based democracy (cf. Hansmann 1996), offering a utilitarian explanation for worker owned and managed firms and organization).

2. The Emergence of Private Property Rights

Although no single philosophical theory provides a complete explanation for private property rights, utilitarianism has emerged as the dominant framework for understanding property rights systems. In a seminal article, Demsetz (1967) posited that private property regimes develop to internalize externalities. As the benefits of internalizing externalities rise (e.g., where a common resource becomes more valuable) or the cost of internalizing externalities fall (e.g., as technology reduces the costs of fencing or monitoring property rights), private property systems emerge to promote efficient management of the resource. As support for this thesis, Demsetz cites anthropological evidence suggesting that native Americans established private hunting territories following the arrival of Europeans and the development of the fur trade. By spurring the hunting of beavers, the fur trade threatened the productivity of this renewable resource, thereby creating social pressure to institute property rules that internalize the effects of resource exploitation.

Demsetz's theory explains the emergence of private property institutions across a broad range of contexts, including air (sulfur oxide emissions in the United States), water allocation, fishing, telecommunications spectrum, wetlands, roads (tolls), trademarks (causes of action for diluting famous marks), and publicity (commercial exploitation of a person's image, likeness, or voice). Nonetheless, by focusing exclusively upon private property institutions, Demsetz's theory overlooks much of the richness of property institutions and the complex social processes that affect their evolution. Libecap (1994), Ostrom (1990), Ellickson (1991), Rose (1986), and others contribute to the explication of a broad spectrum of property institutions lying between the poles of pure private property and the commons where informal and shared governance emerge and succeed in allocating resources.

The emergence and nature of property rights also reflect the political economy driving the legislative process and government agencies that administer property rights (Buchanan and Tullock 1962). Interest groups tend to form in those areas where government policies create concentrated benefits or impose concentrated costs (Olson 1965). Laws and agencies creating or affecting property rights (e.g., through regulation) attract substantial rent-seeking activity. Thus, the design of such property regimes often reflect political, as well as economic, considerations.

3. Evolution of Jurisprudential and Analytical Conceptions of Property

Although property rights are as old as human civilization, the modern jurisprudential conception of property traces its roots to the agrarian and early mercantilist economies that developed during the

decline of feudalism. Sir William Blackstone (1765), author of the first widely accepted legal treatise on English law, regarded property as an absolute right

> which consists in the free use, enjoyment, and disposal of all acquisitions, without any control or diminution, save only by the laws of the land …
>
> So great moreover is the regard of the law for private property, that it will not authorize the least violation of it; no, not even for the general good of the whole community … .

Although some have questioned Blackstone's rigor and cogency (Burns 1985) few doubt his influence upon the jurisprudence of property in England and America. Maine (1861) reinforced the superiority of private property institutions through his historical account of the progress of civilization from primitive forms of shared ownership among kinship groups (status) to individual ownership and alienability of property interests (contract).

The industrial revolution and the shift from an agrarian economy to urban and other community structures changed the ways in which land and other resources were owned and exploited. The desire to restrict land through contract pushed courts to liberalize constraints on non-possessory interests, presaging the disintegration of the unitary bundle of sticks. Property came to be seen increasingly as a malleable concept that could be tailored to serve utilitarian purposes.

In the early twentieth century, Hohfeld (1913) developed a detailed descriptive framework for capturing the richness of legal rights reconceptualizing property as *relationships* among people relative to things. Recognizing four primary legal entitlements—rights, privileges, powers, and immunities—Hohfeld postulated four basic legal relations ('correlatives'): 'right–duty,' 'privilege–absence of right,' 'power–liability,' and 'immunity–disability.' Hohfeld provided the vocabulary and structure for representing the complexity of property relations—bundles of rights (and relations)—as well as the conceptual menu from which legislatures and courts could craft property rights.

Blackstone's conception of property as an inviolate, unified bundle of rights came under increasing assault with the proliferation of zoning and other public land use controls upon private property rights and the expansion of the role of government at all levels following the New Deal in the USA. Property rights, like much of the law governing the legal relations, came increasingly within the purview of political institutions, and accordingly gave way to the instrumentalist (utilitarian) nature of politics.

In a seminal article, Coase (1960) developed a new, instrumentalist framework for analyzing property rights. In much the way that a frictionless surface enables physicists to isolate and study the laws of physics, Coase hypothesized a world of zero transaction costs (costless bargaining and enforcement of agreements), which provided valuable insights into the interplay of legal and market institutions. Coase demonstrated that in a world of costless transactions, an efficient allocation of resources will result regardless of how property rights are initially allocated. The desire of conflicting property owners to maximize their joint profit will lead them to bargain to an efficient allocation of resources. The initial allocation of entitlements will affect the distribution of wealth, but not the total level of wealth.

As with the frictionless surface, Coase recognized that costless transactions do not exist in the real world. His article fueled a deluge of scholarship exploring the implications of costly transactions for the allocation and design of property rights. Building upon Coase's construct, Calabresi and Melamed (1972) developed a framework that divides the analysis of property rights into two inquiries: (1) how should property rights be allocated; and (2) how should such entitlements be enforced—by a property rule (enabling property owners to enjoin interference with their entitlement and requiring consent for violation or transfer of the entitlement), a liability rule (allowing property owners court-determined damages for invasion of their property right, as in the case of eminent domain), or by a rule of inalienability (preventing transfer of a property right, as with prohibitions on the sale of body parts).

Calabresi and Melamed propose that entitlements and enforcement rules be based on economic efficiency, distributional goals, and other considerations of justice. Where transactions are costly, entitlements should be allocated so that the party who is best able to minimize harm (the least cost avoider) bears responsibility for the harm in order to promote economic efficiency. With regard to the enforcement of entitlements, Calabresi and Melamed demonstrate that the use of liability rules can minimize the costs of making an erroneous decision about the allocation of entitlements where transactions are costly. Subsequent scholars have refined and extended the framework and applied it in a range of contexts (Symposium 1997, Kaplow and Shavell 1996, Merges 1996, Radin 1996, Krier and Schwab 1995, Rose-Ackerman 1985, Polinsky 1980, Ellickson 1977, 1973).

In a provocative essay published in the early 1980s, Grey (1980) argued that the field of property 'ceases to be an important category in legal and political theory' as a result of the 'disintegration' of the conception of property. In his view, the dividing of property ownership into a 'bundle of rights' and the development of differing conceptions of property across disciplines has 'fragmented' property theory 'into a set of discontinuous usages.' Numerous scholars have questioned Grey's claims on a range of bases (Munzer 1990), yet the observation that provided the point of departure for Grey's essay, the 'disintegration' of property, resonates today. Property has lost the simplicity of Blackstone's conception—unitary, ex-

clusive, and omnipotent control—but a new conception of property integrating the many institutional structures governing resources, the cultural contingency of governance regimes, and the dynamism of governance institutions has emerged.

4. Property as a Triadic Relation

Menell (2001) and Dwyer and Menell (1998) present property as a dynamic system comprising three components: resources, institutions, and culture. Resources may be governed by background legal institutions (such as common law rules, or in civil law regimes, codes), social norms, markets (contractual agreements supported by the background legal institutions), political institutions (e.g., zoning), or some combination thereof. At any point in time, the mix of governance rules and institutions reflects the nature of the resources governed and the society or culture (knowledge, values, cohesion, technology) in which the resources are situated. Thus, in smaller, more cohesive social settings, we might observe more commonly particular resources are governed through social norms or other common property institutions whereas the same resources might be governed through formal private property institutions in a larger, more heterogeneous community. In this way, the triadic framework captures the economic and cultural dependency of resource governance as well as the richness of property regimes.

The triadic conception of property also incorporates the dynamic nature of resource governance. As reflected in Fig. 1, the framework incorporates causal connections and feedback mechanisms. Changes in the society (such as the value attributed to the protection of endangered species or technological advances reducing the costs of monitoring private property rights) generate changes in the design and mix of governance institutions. Such changes, by affecting the resource base, may generate further reactions within the overall system. The system may also be affected by external shocks, such as charges in the larger political atmosphere. Demsetz's (1967) positive account represents a special case of this dynamic process, but the triadic relation generalizes the dynamism of institutional evolution to account for a broader mix of governance institutions.

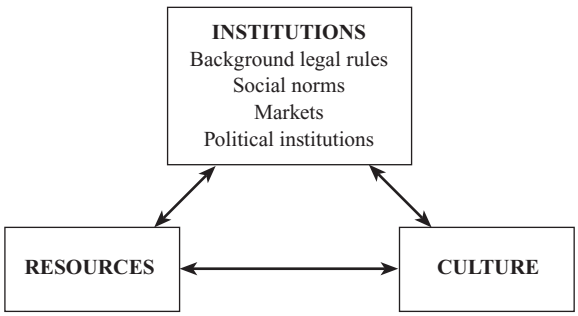

Figure 1
Property

5. New Frontiers

Notwithstanding the settlement of substantially all significant land masses on the planet, the frontiers of property continue to expand and evolve. Among the most important resources in today's economy are intangibles—ideas, genetic code, cyberspace, telecommunication spectrum (Merges et al. 2000). The internet has significantly expanded opportunities to distribute expressive works, but has also threatened the ability of copyright owners to control their works. This new medium has reshaped perceptions about rights to reproduce and distribute expressive works. Advances in bioscience have spawned disputes over rights and control of human genetic code, cell lines, and indigenous plants. And even in the more conventional realm of land management, increasing pressures on natural resources, greater knowledge of ecosystem effects, and changes in attitudes toward the environment continue to alter the balance among private property rights, public ownership, and regulation (Dwyer and Menell 1998, Menell and Stewart 1994). Public rights in artistic treasures, historical documents, architectural works, archeological discoveries, and scholarly research present a further set of challenges for a system of governance built significantly upon a foundation of private property rights (Sax 1999). Just as the advent of rights in government benefits expanded our conception of property four decades ago (Reich 1964), advances in technology and changing social values continue to expand and alter our conception of property.

See also: Intellectual Property: Legal Aspects; Legal Positivism; Natural Law

Bibliography

American Law Institute 1936 *Restatement of the Law of Property*. American Law Institute Publishers, St Paul, Minnesota

Anderson T L, Hill P J 1975 The evolution of property rights: A study of the American West. *Journal of Law and Economics* **12**: 163–79

Anderson T L, Leal D R 1991 *Free Market Environmentalism*. Pacific Research Institute for Public Policy, San Francisco

Bentham J 1789 *An Introduction to the Principles of Morals and Legislation*.

Becker L 1977 *Property Rights: Philosophic Foundations*. Routledge & Kegan Paul, London
Blackstone W 1765 *Commentaries*.
Buchanan J, Tullock G 1962 *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. University of Michigan Press, Ann Arbor
Burns R P 1985 Blackstone's theory of the 'absolute' right of property. *University of Cincinnati Law Review* **54**: 67–86
Calabresi G, Melamed A D 1972 Property rules, liability rules, and inalienability: One view of the Cathedral. *Harvard Law Review* **85**: 1089–128
Coase R 1960 The problem of social cost. *Journal of Law and Economics* **3**: 1–28
Demsetz H 1967 Toward a theory of property rights. *American Economic Review Papers and Proceedings* **57**: 347–58
Dwyer J P, Menell P S 1998 *Property Law and Policy: A Comparative Institutional Perspective*. Foundation Press, Westbury, NY
Ellickson R C 1973 Alternatives to zonoing: Covenants, nuisance rules, and fines in land use controls. *University of Chicago Law Review* **40**: 681–781
Ellickson R C 1977 Suburban growth controls: An economic and legal analysis. *Yale Law Journal*. **86**: 385–511
Ellickson R C 1991 *Order Without Law: How Neighbors Settle Disputes*. Harvard University Press, Cambridge, MA
Ellickson R C 1993 Property in land. *Yale Law Journal* **102**: 1315–1400
Epstein R A 1979 Possession as the root of title. *Georgia Law Review* **12**: 1221–43
Epstein R A 1985 *Taking: Private Property and the Power of Eminent Domain*. Harvard University Press, Cambridge, MA
Freyfogle E T 1996 Ethics, community, and private land. *Ecology Law Quarterly* **23**: 631–61
Grey T 1980 The disintegration of property. *Nomos* **22**: 69–85
Hansmann H 1996 *The Ownership of Enterprise*. Harvard University Press, Cambridge, MA
Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8
Hohfeld W N 1913 Fundamental legal conceptions as applied in judicial reasoning. *Yale Law Journal* **23**: 16–59
Kaplow L, Shavell S 1996 Property rules versus liability rules: An economic analysis. *Harvard Law Review* **109**: 713–90
Krier J, Schwab 1995 Property rules and liability rules: The cathedral in another light. *New York University Law Review* **70**: 440–83
Leopold A 1949 *A Sand County Almanac*. Oxford University Press, New York
Libecap G 1994 The conditions for successful collective action. *Journal of Theoretical Politics* **6**: 569–98
Locke J 1698 *Second Treatise of Government*.
Lueck D 1995 The rule of first possession and the design of the law. *Journal of Law and Economics* **38**: 393–436
Maine H S 1861 *Ancient Law*.
Marx K, Engels F 1848 *The Communist Manifesto*.
Menell P S 1992 Institutional fantasylands: From scientific management to free market environmentalism. *Harvard Journal of Law and Public Policy* **15**: 489–510
Menell P S 2001 Property as a triadic relation. (manuscript in progress)
Menell P S, Stewart R B 1994 *Environmental Law and Policy*. Little, Brown, Boston, MA
Merges R P 1996 Contracting in liability rules: Intellectual property and collective rights organizations. *California Law Review* **84**: 1293–393
Merges R P, Menell P S, Lemley M A 2000 *Intellectual Property in the New Technological Age*, 2nd edn. Aspen Law & Business, New York
Munzer S R 1990 *A Theory of Property*. Cambridge University Press, Cambridge, UK
Nozick R 1974 *Anarchy, State, and Utopia*. Blackwell, Oxford, UK
Olson M Jr 1965 *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge, MA
Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK
Polinsky A M 1980 Resolving nuisance disputes: The simple economics of injunctive and damage remedies. *Stanford Law Review* **32**: 1075–112
Posner R 1998 *Economic Analysis of Law*, 5th edn. Aspen Law & Business
Radin M J 1982 Property and Personhood. *Stanford Law Review* **34**: 957–1015
Radin M J 1993 *Reinterpreting Property*. University of Chicago Press, Chicago
Radin M J 1996 *Contested Commodities*. University of Chicago Press, Chicago
Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA
Rawls J 1993 *Political Liberalism*. Harvard University Press, Cambridge, MA
Reich R 1964 The new property. *Yale Law Journal* **73**: 733–87
Rieser A 1997 Property rights and ecosystem management in US fisheries: Contracting for the commons? *Ecology Law Quarterly* **24**: 813–32
Rose C 1985 Possession as the origin of property. *University of Chicago Law Review* **52**: 73–88
Rose C 1986 The comedy of the commons: Custom, commerce, and inherently public property. *University of Chicago Law Review* **53**: 711–81
Rose C 1987 'Enough, and as good' of what? *Northwestern University Law Review* **81**: 417–42
Rose C 1998 The several futures of property: Of cyberspace and folk tales, emission trades and ecosystems. *Minnesota Law Review* **83**: 129–82
Rose-Ackerman S 1985 Inalienability and the theory of property rights. *Columbia Law Review* **85**: 931–69
Ryan A 1984 *Property and Political Theory*. Blackwell, Oxford, UK
Sax J L 1983 Some thoughts on the decline of private property. *Washington Law Review* **58**: 481–96
Sax J L 1993 Property rights and the economy of nature: Understanding *Lucas v. South Carolina Coastal Council*, *Stanford Law Review* **45**: 1433–55
Sax J L 1999 *Playing Darts with a Rembrandt: Public and Private Rights in Cultural Treasures*. University of Michigan Press, Ann Arbor
Simon W H 1991 Social republican property. *UCLA Law Review* **38**: 1335–413
Schlatter R 1951 *Private Property: The History of an Idea*. George Allen & Unwin
Schnably S J 1993 Property and pragmatism: A critique of Radin's theory of property and personhood. *Stanford Law Review* **45**: 347–407
Sprankling J P 1996 The antiwilderness bias in American property law. *University of Chicago Law Review* **63**: 519–89

Symposium 1997 Property rules, liability rules, and inalienability: A twenty-five year retrospective. *Yale Law Journal* **106**: 2083–213
Waldron J 1988 *The Right to Property*. Clarendon Press, Oxford, UK

P. S. Menell

Property: Legal Aspects of Intergenerational Transmission

1. *Wills and Succession Throughout History*

Property exists essentially in two forms: movable (including intangible) and immovable (land). The transmission of that property at death generally occurs in one of three ways: by will, by customary or legal rules of succession, or by a combination of the two. An heir's right of succession by law or custom is an ancient and powerful right rooted in kinship ties. It was, for example, an attribute of early European feudal life, allowing family members to share that portion of the land vacated by a kinsman's death (see Chester 1982). By the twelfth century in England and many parts of Western Europe, lands were not partitioned, but descended to the eldest son of the landowner under the concept of primogeniture.

The right of a beneficiary to take property under a will is dependent on the legal power or right of the decedent to transmit this property at death to that beneficiary. This right was recognized relatively recently in Western thought and has often been regarded, at least by the positivist tradition, as having been created for society's convenience. Natural rights advocates by contrast might consider the right to bequeath an aboriginal attribute of property itself (see Chester 1998b).

In Western society, the legitimization of the right of bequest symbolized the shift from feudal to individualistic conceptions of property. Unknown to the early law of Greece, India, Egypt, Babylon, and Israel, the will was by no means a universal institution. The testamentary conception built upon by Europeans had its roots in Roman law and was resurrected after feudalism by ecclesiastical courts applying Romano-canonical law (see McMurray 1919).

In England the customary right of testamentary disposition was recognized in very early times, but land—the primary source of wealth—could not be willed until after the Norman invasion and the establishment of feudal tenures. Finally, in 1540, the Statute of Wills declared the right of an individual to leave real property to whomever the individual pleased by means of a duly executed will. This statute confirmed a customary right long recognized for personal property (Chester 1982).

Once established, the right of an individual to leave property outside the family had to be reined in by providing protection from disinheritance for certain family members. For example, in England, primogeniture protected the eldest sons of landed decedents until well into the nineteenth century. Common law rights of dower and curtesy for surviving wives and husbands provided them with interests in the land of the decedent (see Chester 1982). Even though English women of the sixteenth, seventeenth and eighteenth centuries could not will property or own it in their own name, it could descend through them. However, married women's assets were also protected to some extent by courts of equity by means of trusts and other devices, especially toward the end of that period.

Children who were omitted from wills were not protected directly, except those benefiting under primogeniture. However, the wills of their parents were strictly scrutinized to ensure that they had not been left out unintentionally. One English rule derived from Roman law even stated that to cut off an heir 'without a shilling' raised a presumption that the testator was legally incompetent to make a will. Thus even in England, the family's right to inherited property was protected not only when the decedent died without a will (intestate) but to some extent when the decedent did leave such a will (see Chester 1982).

The entrenchment of the right of testation in England was associated with the movement from feudalism, with its emphasis on family and kinship, to free-market capitalism, with its emphasis on the individual. By the mid-nineteenth century, the adult male, Anglo-American owner of both realty and moveable property had nearly unlimited power of testamentary disposition. In contrast, countries such as France, Spain, and Germany on the European continent made the move to capitalism with a right of testation that, even when recognized, was much more sharply curtailed than that in Anglo-America by the requirement of fixed shares for surviving spouse and children.

In the eighteenth and nineteenth centuries, the principal difference between the Continental and Anglo-American systems of succession was the different attitude each displayed toward the right of family members to succeed to the decedent's property. Although the Spanish principle of mandatory heirship appeared in theory to be a significant curb on the Anglo-American concept of testamentary freedom, in practice the differences were not so marked. This reality stems from two separate phenomena: (a) in the majority of instances free testamentary power among the Anglo-Americans was exercised in favor of those who would be forced heirs under a mandatory system such as that in Spain; and (b) even where testamentary power appears absolutely undiluted, far more people

die without wills than with them, resulting in the passage of their property to those who, under Spanish law, for example, would be designated forced heirs (see McKnight 1996).

Spanish lawyers of the eighteenth and nineteenth centuries were trained in the Germanic tradition, which accepted only an extremely narrow testamentary power. Thus, while Anglo-American lawyers thought of both testate and intestate succession, the basic distinction for the Spanish was between those who were forced heirs and those who were not.

What is more, married women in the Spanish system not only had a much more extensive right to property than did those in Anglo-America, but they also had will-making power, which they did not have in Anglo-America. American women, for example, did not gain the power of will-making until the passage of the Married Women's Property Acts of the late nineteenth century.

In Spain the surviving spouse was also entitled to half of the gains of the marriage under what is today termed the community property system. Most of the property that was the decedent's alone passed to descendant forced heirs, despite any contrary testamentary provisions, and if there were no descendants, ascendants took as forced heirs, although in smaller shares than descendants. If the decedent died without a will, the estate passed to the forced heirs, or, if none, to other heirs as dictated by law (see McKnight 1996).

French law also maintained forced heirship under the concept of legitim. During the French Revolution, this concept was viewed as a means of eliminating the feudal principle of primogeniture in order to break up feudal estates. There was some relaxation in the strict adherence to forced heirship by the time of the Napoleonic Code, which announced the general French view of succession that is held even today. Under the Code, half of the testator's estate (after marital property has been distributed to the surviving spouse) was freely disposable if there was one child; one-third if there were two children, and one-fourth if there were three or more children (see McKnight 1996).

In general, since the early fifteenth century, the difference between Continental and Anglo-American views of succession, seen most strikingly in the Hispanic tradition, is that the Spanish adhered to the traditional Germanic tribal concept of familial right to property, whereas the Anglo-Americans took an individualistic view of who should succeed to the decedent's property (see McKnight 1996). By the mid-nineteenth century, this English individualistic notion had triumphed throughout the USA, with the exception of Louisiana which followed the French law of legitim. The Anglo-American system was, however, tempered by the concept of community property in states with a Spanish tradition and by dower and curtesy (mentioned above) in the others. The former gave each spouse an immediate half interest in any property acquired during the marriage, while dower provided a surviving wife with a life interest in one-third of her dead husband's real property, and curtesy gave a surviving husband a life interest in all of the wife's real property.

2. *Trusts*

A particularly ingenious device for the intergenerational transfer of wealth was developed in Anglo-American courts of equity. Called the trust, it has sometimes been copied in other legal systems, for example that of modern Japan (see *International Law Digests* 1999). This form of property holding enabled the decedent to pass legal title to a third party who held the property for the benefit of its intended beneficiary. This third party trustee was subject to the equitable interest of the beneficiary, who generally received the income from the property for a period of time up to life. At the end of the designated period, the equitable title of that beneficiary would cease, transferring absolute ownership to that beneficiary or another chosen by the decedent.

The temporary division of legal and equitable title to property in trust allows, among other things, professional management of trust property assets. It also allows decedents to delay the actual vesting of full title to the property for a period of time up to and including a generation or more, during which period a beneficiary may receive a managed stream of income. Trusts in the USA that pay income to a decedent's children during their lives, and upon their deaths pay the principal to the grandchildren are called generation-skipping trusts. Estate tax advantages of these trusts (achieved because the children never received full title) were cut back by the Tax Reform Act of 1986 and subsequent amendments (which imposed a generation skipping transfer tax (see Johnson and Eller 1998, Dukeminier and Johanson, 1995).

3. *Wills and Succession Among Nations in the Twentieth Century*

The USA took freedom of testation one step further than did England and the countries of the British Commonwealth. While in varying degrees these nations developed flexible systems of family maintenance to protect surviving spouses and children (at least adult, competent children) from disinheritance, the USA maintained the right of testator to disinherit his or her children completely, even those children who were minors and/or had need. Thus, while surviving spouses in the USA are protected either by the elective share (in 'common law' states following the English tradition) or community property laws (in

those states following the Spanish tradition), children are theoretically left to the whims of unrestrained individualism.

Although this disparity has been attacked, it has also been noted that flexible concepts such as lack of testator competence, undue influence, and 'unintentionally' omitted children indirectly protect surviving children to some extent in most 'common law' states (see Chester 1997, 1998a).

That the USA's focus on the individual rights of the decedent, even to the exclusion of children and other descendants, is unusual in the world can also be seen by examining the systems of the Soviet Union (before its dissolution) and China. Traditionally, the Soviet law of inheritance did not differ substantially from that of the civil law countries of Western Europe (see Dukeminier and Johanson 1995). Although there was an attempt by the Bolsheviks to change this in their takeover of power in 1918, it failed. Nearly unique in the modern history of humankind, a 1918 Soviet Union law read: 'inheritance, testate and intestate, is abolished. Upon the death of the owner his property (moveable and immovable) becomes the property of [the state]' (1 *Sob. Uzak.*, RSFSR, No. 34, item 456, Apr. 26, 1918). This attempt to abolish inheritance proved so unpopular with the people that inheritance was reestablished within four years (see Foster-Simons 1985).

At present, inheritance law in Russia and the republics of the former Soviet Union generally follows the civil law views of Western Europe. In this way, it is similar to the law in modern China (see Dukeminier and Johanson 1995). In that country, if first order heirs (spouse, children and parents) inherit, then no second-order heirs (brothers, sisters, and grandparents) will inherit. Following a kind of community property system, all property in China acquired during marriage is joint property, unless the marital partners have executed a contract to the contrary (see *International Law Digests* 1999). Since much of China is still rural and nonindustrial, custom plays an important part in the actual distribution of estates (see Foster-Simons 1985), just as it does for example, in tribal Africa. This distribution depends partly on the law and the terms of any will and partly on the social and economic facts relevant to the intestate, or the testator and his heirs (see *International Law Digests* 1999).

4. *Free Testation and Capitalism*

It is tempting to view the relatively free testation of the USA as coincident with that country's status as the exemplar of capitalism. Except for the elective share (typically one-third) provided for surviving spouses, testation in America is predicated on protecting the free will of the testator. While England and the Commonwealth countries are certainly capitalistic, their laws show, on the whole, more of a socialist's regard for protection of society (and family) than do those of the USA. While family protection is achieved in these countries by reposing a great deal of discretion in probate judges to ensure family maintenance (as distinguished from the use of fixed shares in the Continental system), this still provides a strong check on the whims of the testator. The governmental intrusion into testation is, of course, even pronounced on the continent where fixed shares for the family are the norm.

In its 1987 decision in *Hodel vs. Irving* (481 US 1987), the USA Supreme Court even seemed ready to protect *constitutionally* the right of an individual to transmit property at his death in any manner that he or she chooses. While that decision has been criticized and not extended beyond its facts (which involved government confiscation of noneconomic land parcels descending at the death of American Indians), it symbolizes America's fierce protection of individualism, even after the individual's death (Chester 1995, 1998b). Interestingly, the occasion for that decision was the government's attempt to save Indians from the disastrous economic effects of the Anglo-American individual property system which had been forced on the various tribes. Without imposition of the notion of individually held property in the first place, functioning Indian tribes, much like the tribes in sub-Saharan Africa, might have continued to treat property, which is chiefly in the form of land, as a family, and by extension community asset, which was never the individual's to transmit at death in the first place.

5. *Implications of Inheritance and Taxation Policy for Equal Opportunity and the Distribution of Wealth*

Allowing large amounts of capital to pass between generations without significant taxation is problematic for those countries such as the USA which profess a belief in equal opportunity. Taxation of estates and gifts is widely seen as an intrusion on the freedom to transmit property at death—a freedom that Americans particularly seem to cherish. While the American estate and gift tax has high nominal rates, its high exemption levels and numerous loopholes, including a deduction of up to 100 percent for all property left to a surviving spouse, may make it in effect a 'voluntary tax' (see Cooper 1979).

Countries such as France and Spain have both an inheritance tax and an annual tax on accumulated wealth. Perhaps not coincidentally, these countries do not have so strong a tradition as does the USA of allowing testators to do almost anything they want to do with their money at death.

Commentators (e.g., see Ascher 1990, Chester 1976, 1982, 1998a, 1998b) have noted that the policies of free wealth transmission and equality of opportunity

coexist very awkwardly in the USA. Although legal systems cannot control differences in natural ability or the luck of being born into a family that values, e.g., culture and education, they *can* curb that form of luck represented by inheritance. If equality of opportunity is truly a key goal, then children fortunate enough to have been reared and educated by wealthy parents need not be allowed the additional luck of inheriting their parents' property. Strong and effective taxation of inheritances would contribute significantly to this goal of equal opportunity.

Whatever the logic of this point of view, in countries like the USA there appears to be little public support for more effective taxation on postmortem (or even lifetime) transmission of wealth to heirs. In fact, at the beginning of this twenty-first century bills in Congress relating to the estate tax generally call for its softening or its abolition. Those supporting such bills argue that low or nonexistent taxation of inherited wealth is good for the economy because it preserves large blocks of investment capital in private rather than government hands. However, it might also be said that untrammeled inheritance saps the initiative both of those who receive it and of those who cannot possibly compete on equal terms with those who do (see Chester 1976, 1982).

6. *The Future*

In looking toward the future of intergenerational wealth transmission in the twenty-first century, one might begin by considering the changing nature of the family itself. Particularly in advanced capitalist nations such as the USA, the family is showing signs of disintegration. Traditional heterosexual marriages often end in divorce and several forms of non-traditional relationships are being legitimized. Since advanced capitalist nations offer many forms of wealth that are obtainable by individuals, there is no longer a critical need for children to succeed to a parent's land. Thus, the strict family protection practiced by the inheritance systems of many capatalist nations may seem outdated to some.

US law, unlike that of many nations, does not directly protect children from disinheritance by their parents. However, other nations may not desire to aggravate by such a legal stance the atomization of family and society toward which many of them are already heading. Their common judgment may be that, even in the twenty-first century, some homage to the family, traditional or otherwise, still needs to be paid in the field of intergenerational wealth transfer. Whether the restraints on individualism that such a judgment entails will become the norm as this century progresses remains an open question.

See also: Children and the Law; Family Law; Gender and the Law; Intergenerational Justice; Property: Legal Aspects

Bibliography

Ascher M L 1990 Curtailing inherited wealth. *Michigan Law Review* **89**: 69–151

Chester R 1976 Inheritance and wealth in a just society. *Rutgers Law Review* **30**: 62–101

Chester R 1982 *Inheritance, Wealth, and Society*. Indiana University Press, Bloomington, IN

Chester R 1995 Is the right to devise property constitutionally protected?—The strange case of *Hodel v. Irving*. *Southwestern University Law Review* **24**: 1195–213

Chester R 1997 Should American children be protected against disinheritance? *Real Property, Probate, and Trust Journal* **32**: 405–53

Chester R 1998a Disinheritance and the American child. *Utah Law Review* 1998 : 1–35

Chester R 1998b Inheritance in American legal thought. In: Miller Jr. R K, McNamee S J (eds.) *Inheritance and Wealth in America*. Plenum, New York, pp. 23–43

Cooper G 1979 *A Voluntary Tax?* Brookings Institution, Washington, DC

Dukeminier J, Johanson S M 1995 *Wills, Trusts, and Estates*, 5th edn. Little, Brown, Boston

Foster-Simons F 1985 The development of inheritance law in the Soviet Union and the People's Republic of China. *American Journal of Comparative Law* **32**: 33–62

International Law Digests 1999 Martindale-Hubbell, Summit, NJ

Johnson B W, Eller M B 1998 Federal taxation of inheritance and wealth transfers. In: Miller R K, McNamee, S J (eds.) *Inheritance and Wealth in America*. Plenum, New York, pp. 61–90

McKnight J W 1996 Spanish legitim in the USA—its survival and decline. *American Journal of Comparative Law* **44**: 75–107

McMurray O K 1919 Liberty of testation and some modern limitations thereon. *Illinois Law Review* **14**: 96–123

Rheinstein M, Glendon M A 1995 Inheritance and succession. In: *The New Encyclopedia Britannica (Macropaedia)*. Encyclopedia Britannica, Inc., Chicago, IL, Vol. 21, pp. 638–47

R. Chester

Property, Philosophy of

The philosophy of property concerns itself with the distribution of benefits, entitlements, and responsibilities associated with goods in a society. Most major Western political philosophers have constructed theories of property rights and property ownership including the limits and justification of property (e.g., Plato, Aristotle, Hume, Locke, Mill, Marx). Philosophical inquiry into property rights is closely associated with distributive justice, the role of the state, and the fundamental tensions between contemporary socialist and liberal political theory. Philosophical interest in property and property rights involves conceptual analysis of what property, own-

ership, and property rights are, as well as critical evaluation of the justifications given for property systems. In this entry the meaning of property, property rights, and ownership are addressed, before providing an examination of the contexts (historical and intellectual) within which property theories arise, an overview of competing approaches to justifying property and, finally, a discussion of current (and potential) debates concerning property in political philosophy and philosophy of law.

1. The Term 'Property'

Property is probably best understood relationally. Property rights involve relations among people concerning benefits and responsibilities which often are (but need not be) associated with particular objects in the world. In ordinary speech, however, the term 'property' is often used to describe the objects over which property relations are held themselves. Although it is common to speak of property as concerning objects in the world, philosophical discussions of property concern themselves with property understood as social, legal, and moral relations with regard to objects, benefits, responsibilities, and their control, within given legal and social contexts. The law of property concerns the legally recognized distribution of rights, responsibilities, and benefits associated with objects of property (both physical and immaterial). Social and moral relations of property, however, are not bound by legally recognized understandings of the limits and extent of property rights.

The range of legally recognized property rights and responsibilities vary from jurisdiction to jurisdiction. Similarly, social and moral property rights vary between cultures and across history. It can be difficult to identify whether a particular claim to a benefit or attribution of responsibility is best characterized as a property right or some other form of legally or socially recognized benefit or responsibility. In jurisprudence, clear distinctions between property, contract, and tort bases for legal entitlements or legally protected interests may be impossible to make (see *Property: Legal Aspects*). In this way, philosophical discussions of property are bound inevitably with broader discussions of distributive justice, social justice and with normative discussions of the role of the state in protecting or promoting certain kinds of interests.

The extent and limits on property rights vary along three dimensions: the range of objects over which property rights can be held (for example, whether water, air, or titles can be held as property); the range of specific property rights over an object that are recognized (for example, whether property can be bequeathed on death or whether property rights can be leased to others); and the entities that can bear specific property rights (for example, individuals and corporations, communities, the state). In capitalist states, there is a presumption that anything could be conceived of as property and that anything which is property could be alienated, although there may be good moral grounds for the state to limit property rights in the public interest (for example, prohibitions against the sale of human beings, or prohibitions against the purchase of insurance against some legal responsibilities like legal culpability for criminal conduct). In pre-capitalist and socialist states there may be limits to the kinds of property which can be privately owned (for example there may be state or collective ownership of land, factories, or other means of production), with some presumption of a state interest in limiting private property rights for the purpose of redistribution of benefits and burdens through, for example, taxation, control over prices, limits on rights of sale, and control over wages.

1.1 Property as a 'Bundle of Rights'

In the conceptual analysis of property, it is generally understood as a term which refers to a bundle of rights, or rights and responsibilities. In liberal democratic theory and jurisprudence, specification of the bundle starts with the full array of rights and responsibilities enjoyed by a private owner of a thing. Minimally these are taken to include Hohfeldian rights of exclusion, rights of use, the power to alienate property and immunity from expropriation (Hohfeld 1919). Theorists have provided a number of ways of enumerating the different aspects (rights and responsibilities) of ownership, most are now heavily influenced by Honoré's (1961) list of the 11 'incidents' which together constitute 'full, liberal ownership.' These incidents are the rights of use, possession, management, income, capital, security, absence of term, residuarity, and the responsibilities of prohibition of harmful use and liability to execution. Honoré's analysis applies most readily to property in material objects, but is intended also to apply to immaterial property such as future production, control over labor and copyright. The terms 'thing' and 'object over which property rights are held,' then, are not meant to exclude these forms of property.

Honoré's analysis allows that the constitutive rights which together make up ownership could be identified differently (that finer distinctions could be made, for example between the right to capital through sale and through consumption of property). He also recognizes that for any given object over which property rights may be held (for example, a car) the range of property rights which a person must hold to be deemed the owner of that object may differ from what is required to own another kind of object (for example, a pharmaceutical company). Other theorists provide distinct analyses of property which are less extensive, but draw on Honoré's list: for example, Becker (1977)

and Snare (1972) (see *Property Rights*; *Property: Legal Aspects*).

1.2 'Ownership'

Within liberal capitalism, philosophical discussion of property is often understood as a discussion of 'ownership'. Ownership can be seen as a kind of legal trump with regard to other property rights, but not necessarily with regard to other kinds of moral or social claim.

Many writers seeking to analyze the concepts of property and ownership as social and legal institutions have focused on the rights an owner has regarding a thing, with which nonowners may not interfere. 'Ownership' is the relation one is in (characteristically in a private property system), when one has 'the greatest possible interest' in a thing (Honoré 1961). That is it is the title given to the person who has the widest range of property rights over the thing. It may be more correct to say that the owner is the person who holds the most socially or legally significant property rights over a particular kind of thing. The specific property rights which are seen as most significant will vary with the kind of thing owned and the possible uses of those rights: rights of control over production and rights to income from profits will be relevant in identifying the owner of an industrial business, while rights over control of access and use may be more significant in identifying the owner of a retail space; rights of possession and use may be more important in identifying the owner of personal property (such as a car). In legal contexts, the person who is deemed to be an 'owner' has a stronger legal claim than any other holders of property rights over the thing. Nonetheless, there are many cases where it is unhelpful to try to determine who has rights of ownership, as property rights may be separated and limited in a huge number of ways. Further, the terms 'owner' and 'ownership' most naturally arise in discussion of private property. In the justification of property theories where relative merits of private, communal, state or mixed systems of property are to be assessed, it is frequently more helpful to refer to the range of property relations recognized within the theory, rather than attempting to specify relations of ownership.

1.3 Property Systems

Property systems, or the legal and social institutions concerning property a society are frequently divided into three kinds: private, collective, and common property systems. Waldron (1988) describes property systems as systems of 'rules governing access to and control of material resources' and distinguishes between private, collective, and common property systems in terms of the different 'organizing idea' around which each kind of system is framed. 'In a system of private property, the rules governing access to and control of resources are organized around the idea that resources are on the whole separate objects each assigned and therefore belonging to some particular individual.' Collective property systems are characterized by the organizing idea or a 'social rule that the use of material resources in particular cases is to be determined by reference to the collective interests of society as a whole.' In common property systems, the organizing idea of the rules of allocation of rights to resources is the idea that 'each resource is in principle available for the use of every member alike' (Waldron 1988). For Waldron, kinds of property systems are distinguished according to their organizing ideas, yet Waldron's concept of an organizing idea for property systems is not meant to be reductive. Most property systems in actual societies contain elements of each system: some resources, e.g., sunlight, are understood as common property; some, e.g., state-run utilities, are held as collective property and some resources, e.g., clothing, are held privately.

2. The Context of Debates Surrounding Property

Within political philosophy debate about property rights and the justification of property are closely connected with debate about the legitimacy of state authority, national sovereignty, individual rights against the state and justice. Historically the shift from feudalism to representative democracies in Europe coincided with debate about the source and nature of the justification of property rights. Broadly speaking, feudal modes of production were agricultural. In feudal regimes all property in land was held by the sovereign and rights to use of that land were allocated in accordance with the sovereign will. The sovereign then levied taxes or required some form of payment for the use of the land and had ultimate authority to determine the use of the land and the income generated from the working of that land. Land obtained through conquest or treaty similarly came under the sovereign power. With the rise of representative democracies, the scope for exercising sovereign rights over land and determining taxes and other levies on the use of the land became constrained by the popular will and claims of the citizenry. What remained, however, was the idea that ultimate authority over grants of land, the use of land, and taxation rested with the sovereign, whether that be an individual monarch or the sovereign authority of a parliamentary government.

2.1 Property, Patriarchy, and God

An early debate about the source of legitimate property rights and its relation to political authority was

between Filmer (1680) and Locke (1690). Filmer had argued that the king's sovereignty could be traced directly back to God's gift of the earth and its bounty to Adam. The patriarchal lineage of the king gave him and no other person ownership of the world, and authority over all those living therein. Locke's refutation of Filmer rested on three propositions: the fundamental moral equality of all men; the gift by God of the earth and its fruits to all men in common (and not exclusively to Adam and his descendants); and that individual appropriation of property from the common could be legitimate, in the absence of the universal consent of all the other common owners. Filmer's argument tied the ownership of land and the political authority of the king together through Adam's sovereignty. Locke's argument separated the two, as he argued that legitimate property rights could be held in the state of nature (outside of political societies). However, he argues that political societies would not be entered into with the consent of the citizens unless the sovereign authority protected the pre-political property rights. Thus, the legitimacy of the state depends, in part, on its protection of property rights.

2.2 *Property and Liberalism*

Within liberal political philosophy more generally, the recognition, enforcement, and protection of private property rights is viewed as a function of the state (Rawls 1971), and as a source of its legitimacy. Private property is understood by many liberals as providing a realm within which individuals can exercise personal liberty, free from interference by others. If something is owned by a person, others must not interfere with the harmless exercise of the property rights held over the thing. Some liberal theorists, particularly those who emphasize the significance of 'negative freedom' argue that expansion of personal liberty depends on private possessions through which liberty can be exercised. Other liberals, however, recognize that the unequal distribution of private property rights limits the equal enjoyment of liberty and therefore argue that property rights must be limited to protect against gross inequalities in freedom (see *Freedom/Liberty: Impact on the Social Sciences*).

2.3 *Equality and Justice*

Because systems of property distribute goods among individuals within a state, discussion of property is closely linked with concerns about distributive justice and equality. Significant inequalities between individuals or groups which result from the unfettered exercise of property rights can be addressed within the justificatory structure of a property system (placing limits on what can be owned or how property rights can be exercised) or through separate appeals to equality or justice within a broader political theory. Rawls (1971) provides the 'difference principle' as part of his two principles of justice to protect against significant inequalities which might emerge through the various political institutions of his just state. The difference principle can be applied to any form of inequality within the state; applied to property it holds that differences in property rights are justifiable only if they are of the greatest benefit to those who are least advantaged. This would allow some inequalities within a property system, such as those which are required to provide sufficient incentive for some people to engage in risky, dangerous, or socially stigmatized activities where those activities benefit the least advantaged. It would also require that there be some decent minimum level of welfare for all members of the society and would limit the extent of difference between the most and least advantaged (see *Justice: Philosophical Aspects*; *Equality: Philosophical Aspects*).

3. *Property and Justification*

Although conceptual analysis of property, property rights, and ownership have concerned philosophers and students of jurisprudence, by far the greater philosophical interest in property concerns its justification. The array of property rights identified by Honoré as jointly constitutive of full ownership could be justified as a whole or separately. Some justifications of property are for mixed systems; so some specific property rights are justified by appeal to one kind of grounding while others may be justified by appeal to another grounding.

Proudhon (1890) famously argued that no private property could be justified. In writing '*La propriété c'est le vol*' (property is theft), he argued that any plausible justification of private property would both demand equality and would, thereby, prove self-defeating. On Proudhon's view, whatever justified one person's exclusive right to a thing, equally justified a similar right for all others. Given the scarcity or at least physical limits on potential objects of property, exclusive rights to property entailed unequal rights, hence one person's exercise of private property rights constituted theft of another's (potential or actual) property.

Despite Proudohn's skepticism that any adequate justification of property could be developed, many political theorists have attempted to provide normative justifications for property. The justification of property rights has taken a number of (sometimes overlapping) forms, these include property rights justified by natural law; property rights justified by acquisition; property rights grounded in the desert of the laborer; property rights justified by appeal to

utility; and property rights as necessary for the development of personality.

3.1 Natural Law

Locke (1690) offers both a natural law theory of property and a labor theory of property in the Second Treatise. First he argued that the earth and its bounty were given to mankind in common by God, such that each individual had the right to take some of the common bounty of nature to provide for life, preservation and comfort. The individual comes to own what is taken from the common, through the exercise of his natural liberty, the labor of his body. In the state of nature, property rights are limited: fruits taken from the common are for the individual's use and must not be allowed to waste in a person's possession. Further, the appropriator must leave 'enough and as good' bounty in the common for others. In the transformation from the state of nature to the civil state, property rules are formalized and the state takes on the responsibility of enforcing the property rights of citizens. Macpherson (1962) and Tully (1980) offer two interpretations of Locke's argument: Macpherson arguing that Locke's argument is essentially possessive individualist, so that the natural law grounding of property is de-emphasized and the protection of exclusive private property is emphasized. Tully, in contrast, argues that natural law limits to property extend into the political state, thus providing significant limitations to individual acquisition.

3.2 Acquisition

Nozick (1974) offers a contemporary libertarian justification for property drawing on Locke's argument, with three significant differences from Locke's argument. For Nozick, the earth and its products are initially unowned. Each individual comes to have legitimate private property rights in those things which he or she has legitimately acquired. Without the premises of common ownership or the aim of legitimate acquisition being the life, preservation and comfort of the acquirer, Nozick's argument justifies largely unrestricted property rights and unlimited property acquisition. Nozick's sole limitation of property rights is what he calls his 'lockean proviso' property rights can be limited if 'the position of others no longer at liberty to use the thing is thereby worsened' (Nozick 1974, p. 178). The situation of others is not worsened, however, by the restriction of liberty which they experience as a result of the thing becoming privately owned. Moreover, 'worsening' the situation of others through acquisition requires compensation, it does not curtail acquisition of the thing in question. Nozick is notoriously vague about which acts constitute acts of legitimate initial acquisition of unowned objects. Nozick's view is further criticized for failing to provide an argument for the unilateral imposition of duties on all other individuals which follows from an act of acquisition (see *Libertarianism*).

3.3 Labor and Desert

Contemporary arguments which justify limited property rights grounded in labor and desert can be found in Becker (1977) and more recently Munzer (1990). On both accounts, a person's (morally legitimate) labor to create, make available, or to improve something which was unowned grounds a claim to property. The person deserves some recompense for the labor expended, and this will often justify property rights in the improved thing or in the surplus value so generated. For Munzer, the laborers' use of their bodies in the production of something or in the provision of a service gives them the basis for a property claim on the product or on wages. These claims are *prima facie* claims, they are qualified by the needs or rights of others, and by scarcity. Munzer recognizes that the work involved in production is a social activity, and thus argues that payment of wages should be commensurate with desert. It is unclear, however, how desert is to be measured against labor, given the great differences between individuals in capacity, talent, and opportunity.

3.4 Utility and Efficiency

Hume (1740) and Mill (1891) offer utilitarian justifications for property. A system of property rights which distinguishes what is one person's from another's is required for the efficient use of the earth's resources. What the extent, limit, and scope of property rights should be, on a utilitarian account, is a matter of expediency and frequently involves balancing stability of expectations against efficient production and distribution. Hume favored custom and a fixed prescription of property rights, even if that approach allowed for considerable inefficiency. For Hume, inheritance and rights to alienate (sell, bequeath, or give) property acted as incentives for individuals to develop useful habits and accomplishments. Mill, by contrast, argued that the law of property ought to be used by the state to promote greater efficiency and a greater level of communal production of the goods needed within a society. Mill favored private property, but with limited rights of exclusive access and enjoyment of land so as to promote efficient exploitation. Mill recognized that property rights formed a bundle; each property right associated with ownership of a particular kind of property needed to be justified independently on the grounds of utility.

More contemporary utilitarian justifications of property concern themselves primarily with questions of efficiency in allocation of property rights. Economists in particular favor property systems grounded in efficiency, and argue for (somewhat moderated) capitalist relations of production and exchange as the most efficient way to respond to individual wants and needs. Philosophers emphasizing utility, rather than efficiency, are likely to justify mixed property systems (some public as well as some private property) which are roughly egalitarian, as the principle of utility can be directed to make interpersonal comparisons of individual preference satisfaction (see *Economics and Ethics*).

3.5 *Property and Personality*

Hegel (1821) argued that property is justified as an expression of human freedom. It is important to note, however, that Hegel does not ground property in natural freedom as natural law theorists, like Locke, sought to do. Property relations involve both a relationship between material objects and persons and relations among persons. In taking up an unowned object as property, the acquirer expresses his or her personality upon the object and gives the object value. In the absence of human actions of acquisition and use, material objects have no value and are useless. A human being can only develop—as a free moral agent—in the context of a community of other persons. To be a person is necessarily communal, it depends on recognition by others (Ryan 1984, p. 121). Each individual's will can only be recognized by others through the individual's impression of it on objects of property. Private property, then, is essential for personality. Ownership involves the impression of one's personality on what is possessed, to the exclusion of others. One's will or personality is embodied in the object. Waldron argues that Hegel's account is a general-right based justification for private property: that private property is something which it is important for every person to have, so that the poverty and propertylessness of some individuals is something a property system ought to protect against (Waldron 1988). Property rights justified on Hegel's account included possession, use, and alienation.

Radin (1993) offers a contemporary personality justification of property rights. Radin does not provide a fully developed argument for private property, but rather uses her argument to describe and evaluate existing property relations. Radin argues that at least some private property is required for the development of personhood, and that some things are so closely tied to personhood that they ought not to be commodified as property at all. Radin emphasizes a distinction between property relations based on personal attachment, property constitutive of personality as opposed to the more readily commodified, 'fungible' relations of property. Radin's view is not entirely Hegelian, as it is strongly influenced by more traditional liberal conceptions of freedom and individualism, yet it is clearly distinguishable from utilitarian emphasis on economic efficiency and Lockean labor justifications of property.

3.6 *Marxist Critiques*

Most contemporary theorists writing on property, write against the backdrop of the Marxist critique of private property and its centrality to capitalist modes of production. Marx (1971) argued that private property systems presuppose and reproduce inequalities in power between those who have amassed sufficient wealth to become owners of capital (the means of production of goods) and those who are forced to sell their labor power in order to purchase those goods needed for life. Given the equal needs of all humans for the material necessities of life, workers who sell their labor power are open to exploitation by those who own capital. The capitalist pays the worker at the rate required for the subsistence and reproduction of labor, but sells the product produced through the worker's labor at a price which exceeds the cost of the labor and the cost of the resources required to produce the product. The value added by the laborer's efforts (surplus value) is taken by the owner of capital as profit. The degree to which workers are exploited depends on how much surplus value is appropriated by the capitalist. Further, capitalist modes of production involve various forms of alienation including the alienation of people from nature, of workers from the products of their collective labor, and of workers from their own labor power. This alienation, for Marx, is a distortion of human relations. The supposed expansion of freedom offered by private property which some liberal and libertarian theorists argue to defend private ownership, must be reconciled with the alienation and loss of freedom suffered by the propertyless under capitalism (Cohen 1981) (see *Analytical Marxism*).

4. *Ongoing Debates About Property*

Current debates about property rights raise both fundamental questions about the relationship between property rights and state sovereignty and contextual questions about the appropriate limitation or expansion of property rights. One source of this discussion is response to the demand for recognition of indigenous people's prior ownership of colonized lands in the New World and Australasia. Tully (1995), for example, argues that proper recognition of indigenous peoples, their forms of government, and their legal systems, including their property systems is required for the current legitimacy of states such as

Canada, Australia, and the USA. This area of debate ties the justification of property systems to state legitimacy as well as demanding an account of international relations between states and justice in recognition of other states' property systems.

Another current area of debate concerns conflicting interests in protection of property rights and protection of environmental values. Environmental ethicists have argued for greater limitation of property rights where those rights risk serious damage to the environment and have also argued for increasing the responsibilities of owners for long term damage to environmental values (Dodds 1994). Debate about whether 'pollution taxes' on factories that produce effluent unfairly constrain property rights or can act as an effective deterrent against pollution are among the issues concerning property and the environment.

A third area of considerable current debate about property involves the range of entities which are properly understood as property and as commodities, including ideas, human body parts, reproductive capacities, organisms, and electronic commodities. Radin (1996) challenges the ready extension of property rights allowing for sale and purchase of human tissue, embryos, sexual services, and other forms of property which she views as incompatible or threatening to personality. Other theorists argue that whether property rights are formally recognized over these things or not is irrelevant as there exists a market for these 'goods' and that will determine what is property, regardless of legal or moral attempts at constraining property (e.g., see *Market and Nonmarket Allocation*). Finally, with the growth of electronic commerce, a new field of property debate is also opening up as there will inevitably follow a need for legal determinations about property relations involving purely electronic commodities.

See also: Contractarianism; Economics and Ethics; Equality: Philosophical Aspects; Freedom/Liberty: Impact on the Social Sciences; Intergenerational Justice; Justice: Philosophical Aspects; Law and Morality: A Continental-European Perspective; Law and Morality: An Analytic Perspective; Libertarianism; Market and Nonmarket Allocation; Property: Legal Aspects; Property Rights; Rights; Welfare: Philosophical Aspects

Bibliography

Becker L C 1977 *Property Rights: Philosophic Foundations*. Routledge & Kegan Paul, Henley, UK and New York

Cohen G 1981 Illusions about private property and proletarian unfreedom. In: Mepham J, Hillel Ruben D (eds.) *Issues in Marxist Philosophy*. Harvester Press, Brighton, UK, Vol. 4, pp. 223–42

Dickenson D 1997 *Property, Women and Politics*. Polity Press, Cambridge, UK

Dodds S 1994 Property Rights and the Environment. In: Cosgrove L, Evans D G, Yenken D (eds.) *Environmental Values, Knowledge and Action*. Melbourne University Press, Melbourne, Australia, pp. 47–58

Filmer R 1680 Patriarcha or the natural power of kings. In: Laslett P (ed.) *Patriarcha and Other Political Works of Sir Robert Filmer*. Blackwell, Oxford, UK

Harris J W 1996 *Property and Justice*. Clarendon Press, Oxford, UK

Hegel G W F 1821 *Grundlinien der Philosophie des Rechts*. Berlin [1942 Knox T M (trans. with notes) *Hegel's Philosophy of Right*. Oxford University Press, Oxford, UK]

Hohfeld W N 1919 *Fundamental Legal Conceptions as Applied in Judicial Reasoning*. [reprinted 1978. Greenwood Press, West Port, CT]

Honoré A M 1961 Ownership. In: Guest A G (ed.) *Oxford Essays in Jurisprudence* (*First Series*). Oxford University Press, London

Hume D 1740 *Treatise Concerning Human Nature* [1888 Selby-Bigge L A (ed.) Clarendon Press, Oxford, UK]

Locke J 1690 *Two Treatises of Government* [1960 Laslett P (ed.) Cambridge University Press, Cambridge, UK]

Macpherson C 1962 *The Political Theory of Possessive Individualism: Hobbes to Locke*. Clarendon Press, Oxford, UK

Marx K 1971 *Critique of the Gotha Programme*. Progress Publishers, Moscow

Mill J S 1891 *Principles of Political Economy*. Routledge, London

Munzer S R 1990 *A Theory of Property*. Cambridge University Press, Cambridge, UK

Nozick R 1974 *Anarchy, State, and Utopia*. Basic Books, New York

Proudhon P-J 1890 Qu'est ce que la propriété? [1970 *What is Property?* Tucker B R (trans.) Dover Publications, New York]

Radin M J 1993 *Reinterpreting Property*. University of Chicago Press, Chicago and London

Radin M J 1996 *Contested Commodities: The Trouble with Trade in Sex, Children, Body Parts, and Other Things*. Harvard University Press, Cambridge, MA and London

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Ryan A 1984 *Property and Political Theory*. Blackwell, Oxford, UK

Snare F 1972 The concept of property. *American Philosophical Quarterly* **9**: 200–6

Tully J 1980 *A Discourse on Property: John Locke and his Adversaries*. Cambridge University Press, Cambridge, UK

Tully J 1995 *Strange Multiplicity: Constitutionalism in an Age of Diversity*. Cambridge University Press, Cambridge, UK

Waldron J 1988 *The Right to Private Property*. Clarendon Press, Oxford, UK

S. Dodds

Property Rights

1. *The Nature of Property Rights*

A property right is the right to use a particular item of property in a particular way. An item of property can be used in more than one way, as long as those

different uses do not exclude each other. Therefore, there can be more than one property right attached to one item of property.

There are various types of property over which property rights can be exercised: water (e.g., extracting water from a river), intellectual property (e.g., authorship of a book), moveable property ('chattels'), and immovable property (land and buildings). For this article (property rights in connection with urban studies and planning), we restrict ourselves to immovable property (alternatively called landed property, real estate, or real property).

It is customary to classify the various types of rights in immovable property which can be attached to one item:

(a) according to the type of use (e.g., easements);

(b) according to the division of the space (i.e., special purpose boundaries within the general purpose boundaries of the item of property, such as an office in an office block);

(c) according to the time (temporary, such as usufructs or leases, or in perpetuity).

There are some other classifications of property rights which cut across those given above:

(a) superior/inferior (the owner of the superior right is entitled to acquire the inferior rights under certain circumstances);

(b) alienable/inalienable (the owner of an alienable right is entitled to transfer it to someone else of their own choosing);

(c) material/contractual (this distinction is derived from Roman law. Material rights, or 'rights in rem,' are attached to the property irrespective of the (legal) person exercising the right and can therefore be freely traded. Contractual rights, or 'rights in persona,' are attached to the property but may be exercised only by the (legal) person specifically named in connection with that right: they cannot, therefore, be traded).

2. *Separate Property Rights*

It follows from this that it is not an item of property which is owned or exchanged but rights in property and, because there can be many different rights attached to one item, it is possible that one right is owned separately from other rights in the same item (e.g., owning a lease to use a building 'owned' by another).

Where the property right may be traded (and this does not apply to rights which are inalienable and 'rights in persona'), this will have a market value equal to the 'present worth of rights to future benefits arising from ownership' (Jaffe and Sirmans 1989). The owner of more than one right over an item of property might have the choice between selling those rights to one person (keeping the bundle of rights intact) or to several persons (splitting or decomposing the rights). If the total value is greater undivided than divided, the seller will probably decide not to split the rights. Conversely, if the rights over one item are divided and if the total value of all the rights would be greater if they were amalgamated (there is a 'marriage value'), then one person will try to gain possession of as many rights as possible.

3. *The Right(s) of Property Ownership*

Because it is a right in property rather than property itself which is owned, it is possible to conceive of a situation in which rights over one item of property were so dispersed that no one person could direct or manage the way in which the item was used. This is not found often: there are inferior and superior rights and there are temporary rights which revert to the owner of perpetual rights. Clearly, it is found to be socially (and economically) desirable that one (legal) person should be able to act as owner of the property (and for this purpose it does not matter whether the legal person is a private or a public entity).

In countries where the law on property is still strongly influenced by the feudal past, property ownership is regarded as being ownership of a full 'bundle of property rights' (fee simple, in Anglo-Saxon legal terms): the starting point is many rights in one item of property, which have been bundled and are in possession of one person. In countries where the Napoleonic legal code shapes the present system, the starting point is that property ownership is complete and that partial rights can then be split off from it. For example, in Dutch law, which is based on that code, ownership is defined as 'the most comprehensive right that a person can have over an object' (Civil Code, article 5.1).

The particular set of rights which constitute 'property ownership' varies from society to society. Nevertheless, it has been suggested that ownership of a property item always includes:

- the right to exclude others from that property;
- the right to dispose of (to alienate) that property;
- the right to derive income from that property.

4. *Restrictions on the Exercise of Property Rights*

There are many restrictions on the exercise of property rights, most of them accepted voluntarily by the owner when entering into a contract (e.g., someone acquires a building lease from the owner of the fee simple). In addition, there are restrictions imposed by the public administration. These latter can be politically very contentious, when the right(s) of property ownership are constrained.

The owner of an item of property might wish that, 'It is my land and I can do what I like with it' but in practice the legislature always restricts the right of

ownership. Again we take the example of Dutch law (Civil Code, article 5.1) where the definition of ownership quoted above is followed by:

– 'the owner is entitled freely to use the object and to exclude others from that object, as long as this use does not infringe the rights of others, nor the constraints imposed by written and unwritten law';

– 'the owner of the object is owner also of the benefits arising from that object, apart from any rights that others may have.'

A restriction on property rights commonly found is expropriation (the state may acquire the ownership against the wishes of the owner). Another restriction is land-use planning by a public authority (see Sect. 7). It is interesting to note that although, under the US Constitution (5th amendment), 'nor shall private property be taken for public use, without just compensation,' the courts have ruled that public actions which restrict the rights of property owners, as long as this does not take 100 percent of the economic value of the property, falls under 'police powers' which allow regulation in the interests of 'public health, safety and morals.' Such 'taking' does not need to be compensated.

5. *Property Rights as a Social Creation*

In order that a property right can be exercised, it is necessary to delineate the right (define the limits of the property item and the limits of the particular property right), to register who is entitled to dispose of the right, and to determine how the right can be enforced or defended. Establishing those regulations is a matter for the legislature, enforcing them a matter for the courts. It follows from this that property rights are a 'social creation': there can be no property rights without agreement on what they entail and without the possibility of enforcing them legally. It is not that something is protected because it is a property right, but that something becomes a property right by being protected (Bromley 1998, p. 25).

From this it follows that property rights can be changed by act of the legislature. This can take the form of restricting an existing right. Introducing a law on noise zones provides an example: henceforth, the owner of land may build upon it only if noise levels on that land are lower than a certain level. Property rights can also be changed by identifying a use which was previously included in a wider right, and making of it a separate right. A well known example is the nationalization of development rights in Britain. The right to develop property was part of the right to own property: in 1947 the British parliament separated the right to development from the right to ownership, nationalized the right to development (which became ownership of the state), and compensated owners of the ownership right for the loss of the right to development. Yet another way of changing property rights is to establish private rights over something which previously could not be owned. It has been suggested, for example (and see Sect. 7), that the owner of a right to use land in a particular location should have also the right to clean air at that location, whereby that is a right which could be traded.

6. *Property Rights Regimes*

The statement that property rights are a social creation does not mean that they can be created only by the formal legislature (the law-making institutions). Many societies had property rights before they had such a legislature. The Dutch, for example, set up water boards with legislative powers over rights to use land and water long before they had a general purpose government at any level.

Moreover, a society has often developed its formal legal systems in close connection with its system of property rights: the wish to define and enforce property rights has influenced the legal system. Alternatively, a society may reform its legal system deliberately in order to change the existing system of property rights (as happened after the French Revolution and in the United States after the declaration of independence).

It is for such reasons that it is common to talk about a society having a 'property rights regime' (or a land regime), an amalgam of law and custom as distinct from the formal law on property. A regime can be characterised by the way in which it regulates three aspects of property rights:

– who may exercise the right (and may exclude others from exercising it);

– how the right may be used (including restrictions on that use); and,

– how the right may be transferred.

(It will be seen that these are the same three aspects used to define the minimum set of rights that constitute property ownership—see Sect. 1.) Two sorts of regime have already been mentioned: the Anglo-Saxon and the Napoleonic. Others have distinguished in addition ('legal families' in Europe, see Newman and Thornley 1996): the Germanic, the Scandinavian, and the East European regimes.

It is sometimes claimed that there are social laws which explain the creation of property rights regimes and changes in them (that is, the regimes do not arise accidentally). These laws have the form of a 'positive' theory which can be tested empirically. In an influential article, Ellickson (1992/3) put forward an 'efficiency thesis' that 'land rules within a close-knit group evolve as to minimise its members' costs.' The costs he identifies as 'deadweight' (net external costs of individual actions) and transaction costs. A change in land rules is efficient when it reduces the sum of these two sets of costs.

7. The Politics of Property Rights

Because a property rights regime is a social creation, it can then be evaluated socially, with a view to changing it in some desired way. That is not new: property rights have always attracted intense political attention.

One strand in the political argument may be called the dyadic view of property rights: private versus public. At one level, the issue is: who should own property rights (in particular, who should own land), the individual (or the family unit) or some collectivity such as the state? Some regard private ownership of property as a bulwark against state despotism, at the other extreme are those who claim (with Proudhon, 1809–1865) that 'property is theft' and that it should be owned collectively.

At a more refined level, the issue in most present-day Western societies is: to what extent and in what ways should the state intervene in the exercise of private property rights? Most Western societies agree that the private exercise of property rights cannot be unconstrained (see Sect. 4) and even the most fervent defenders of private property rights usually want the state (or some collectivity) to own land and build roads upon it if this gives access to their land.

Another strand in the political argument about property rights regimes can be called triadic. Then there are not two parties—a private person versus the state—but three—one person who, in the exercise of their property rights can affect another person, and the state as an arbiter in that private/private relationship. Most property law and custom is, in fact, civil law: a collectivity regulates the relationships between (private) legal persons in the use of their property.

Much of the current debate about property rights and land-use planning can be interpreted in terms of dyadic/triadic. To the extent that land-use planning by a public body is justified in terms of welfare economics, it is usually the arguments of Pigou (1932) which are invoked. The market in which private persons exchange property rights has certain characteristics which mean that a welfare optimum (in a Paretian sense) will not necessarily be reached. There are 'market failures.' This is seen as a problem, which is resolved when a public body intervenes in the market for property rights. Examples include: restricting the right to use a plot of land in the way that the land owner wants, expropriating the ownership right, public ownership to provide 'public goods.'

In 1960, Coase argued that market failures arise mainly because of transaction costs and because property rights are not clearly delineated. Examples of this latter are the right to clean air (see Sect. 5) and the right to graze cattle on common land. Incomplete separation of rights creates common property, i.e., in the public domain (Barzel 1898) and, if these rights are valued, resources are spent on capturing them. This can lead to rent-seeking and 'common-property dissipation' (of which the best known description is Hardin's (1968) 'tragedy of the commons'). According to this argument, the problem of market failures can be tackled not only by government intervention to restrict property rights (Pigovian measures: moreover, these can make matters worse by creating 'public failures,' see Wolf 1979) but alternatively by public measures to enable markets in property rights to function well, to reduce transaction costs and to delineate exclusive property rights.

The economic argument is that helping the market in private property rights to work better will reduce risks, which will stimulate investment, which will achieve a better allocation of resources than through public land-use planning. The political argument is that a property rights regime should be created in which—in dyadic terms—most property rights are held privately and in which—in triadic terms—the role of the public administration is to regulate private transactions in property rights in ways described above. If this is not possible (because, for example transaction costs are high and cannot be reduced, as is the case with roads in urban areas) then 'Pigovian' measures can be justified. This approach to land-use planning is called the 'property rights model' (see e.g., Lai 1997).

See also: Land Rights and Gender; Land Tenure; Property: Legal Aspects; Property, Philosophy of

Bibliography

Barzel Y 1898 *Economic Analysis of Property Rights*. Cambridge University Press, Cambridge, UK

Bromley D W 1998 Rousseau's revenge: The demise of the freehold estate. In: Jacobs H M (ed.) *Who Owns America?* The University of Wisconsin Press, Madison, WI, pp. 19–28

Coase R H 1960 The problem of social cost. *Journal of Law and Economics* **3**: 1, 1–44

Ellickson R C 1992/3 Property in land. *Yale Law Journal* **102**: 1315–400

Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8

Jaffe A J, Sirmans C F 1989 *Fundamentals of Real Estate Investment*. Prentice Hall, Englewood Cliffs, NJ

Lai L W C 1997 Property rights justification for planning and a theory of zoning. In: Diamond D, Massam B H (eds.) *Progress in Planning* **48**: 161–246

Newman P, Thornley A 1996 *Urban Planning in Europe*. Routledge, London

Pigou A C 1932 *The Economics of Welfare*. 4th edn. Macmillan, London

Wolf C J 1979 A theory of non-market failure, framework for implementation analysis. *Journal of Law and Economics* **22**: 107–39

D. B. Needham

Prophetism

Prophetism may be defined as a religious current or movement promoted by the activities of a prophet. A prophet in turn may be defined as a leader of or key participant in a movement or group who proclaims a religious mission on the basis of personal revelation. While many prophets have been founders of religious movements they are not necessarily so but may be significant figures within a revitalization of a religious group or simply play a central role in the religious life of a group or community by pronouncing upon religious matters through revelatory insights or experiences. Neither is the foretelling of future events a necessary component of prophecy but is not uncommonly associated with it. A prophet may or may not convey a message from the gods or be the mouthpiece of divinity. They may express the word of God or the gods or may simply convey religious truths to their followers, which are not claimed to stem from such supernatural sources. Buss (1980) puts the point usefully in the following terms: 'The role of a prophet or seer can be characterised as the receiving and transmitting of communications not available to ordinary conscious sensitivity which are held to come from a source (the divine) or through a form of perception transcending normal spatio-temporal limitations.' Prophets may or may not claim spiritual or supernatural powers themselves and may or may not receive revelation through the experience of states of trance or possession.

What is essential to the notion of prophecy, then, is the relatively emotional proclamation of a religious message that has been revealed personally in a non-routine and nonmundane manner. Prophets are called to their mission and compelled to follow it whether voluntarily or, as is not uncommon, very reluctantly.

The category of prophet as a religious specialist or type of religious figure is one which owes much to Weber who gives it a central role in his sociology of religion (see *Religion, Sociology of*). For Weber a prophet is 'a purely individual bearer of charisma, who by virtue of his mission proclaims a religious doctrine or divine commandment' (1965). Weber distinguishes the prophet from the category of priest. The priest commands religious authority on the basis of a position within a sacred tradition or religious organization and dispenses salvation by virtue of holding an office while the prophet commands religious authority on the basis of personal charisma and recognition of a personal calling to a mission through revelation. The term 'charisma' is used by Weber to refer generally to extraordinary powers essentially of a magical or non-mundane nature (see *Charisma and Charismatic*). The priest for Weber is thus the upholder and guardian of tradition while the prophet is a religious innovator or at least a radical revitalising force (see *Religious Specialists*).

Weber is careful to distinguish the prophet from a number of other types of religious specialist. While the authority and following of prophets may be based upon magical charisma they are not primarily magicians or workers of miracles. The prophet is concerned primarily with doctrine and commandment. It is this that distinguishes the prophet from the mystagogue whose following is based purely upon attribution of magical powers. Yet it is rare that prophets are not expected to give proof of their particular gifts and such proof has more often than not taken the form of possession of magical abilities. Reflecting the fact that charisma may be either inherent or acquired through undergoing extraordinary experiences or practices, these capacities may be inherent in the person or may be acquired through, e.g., ascetic regimes or cultivation of ecstatic states.

Again, while prophets engage, in many instances, in divinatory activities, as was the case, e.g., with the Biblical *nebiim* (singular *nabi*), they may be distinguished from diviners on the basis that it is normal for the latter to be economically recompensed for their services. Prophets provide their spiritual revelations gratuitously and, indeed, often unbidden. To do otherwise would be to undermine their distinctive claim to religious and spiritual influence.

Likewise, Weber distinguishes the prophet from related or similar religious specialists such as legislators, arbitrators, social reformers, orators and teachers on the grounds that lacking in the case of all of them is the proclamation of a religious truth on the basis of personal revelation. Later usages of the term 'prophet' are often less precise and looser than Weber's but in general at the root of them is this specific emphasis upon a purely personal revelatory impulse and calling.

Weber distinguished between two fundamentally different types of prophecy. In the 'ethical' or 'emissary' type, exemplified most clearly by figures such as Zoroaster or Muhammad, the prophet is an instrument for the proclamation of the message or commandments of a god. The doctrines proclaimed by such a prophet take the form of ethical duties. In the 'exemplary' type, most clearly exemplified by the Buddha, the prophet proclaims a soteriological message by personal example and demonstration of the means or path to salvation which others are free to follow to their personal benefit if they wish or ignore to their personal detriment.

In both cases the prophetic revelation acts to promote a meaningful understanding of life in relation to the whole of reality to which conduct must be brought into alignment if salvation is to be attained. It involves the construction of a cosmos as a meaningfully ordered totality. As Parsons (1965) says 'the prophet is above all the agent of the process of breakthrough to a higher, in the sense of more rationalised and systematised, cultural order, an order at the level of religious ethics, which in turn has

implications for the nature of the society in which it becomes institutionalised.' While there is such a stress in Weber's characterization of the prophet upon the construction of a more rationalized and systematized meaningful order, prophecy, Parsons is careful to point out, is not fundamentally an intellectual or theological enterprise but frequently takes emotional and sometimes even violent forms.

Weber's most extensive analysis of specific prophetic activity and prophetic movements occurs in his study of the development of Judaism, in his *Ancient Judaism* (1952), a development in which the Biblical prophets are given a crucial role (see *Judaism*). Weber traces the emergence of a new type of prophet towards the end of the period of the confederacy that he distinguishes from the earlier ecstatics or *nebiim* and the *roeh* or oracle givers. These new prophets saw themselves as called upon to transmit the will of Yahweh to the people and to the rulers. Early prophets of this type did not pronounce unless solicited but progressively they spoke independently and increasingly critically of the political authorities. From the time of Elijah onwards prophets spoke as much to the masses as to the rulers culminating in a situation in which the prophets became the intellectual leaders of the opposition against all that monarchy was perceived to stand for—centralization of power and government, heavy taxation, and social injustice.

Not that Weber considered social justice to be their primary concern. It was the destiny of the state and of the people that motivated their essentially religious yet inextricably political message. It was their purely religious concerns with fulfilling the commandments of Yahweh that brought upon them the opposition of the authorities. For Weber the Biblical prophets represented the values and ethos of the presettlement confederacy.

Classical Biblical prophecy in Weber's analysis is characterized by independence, spontaneity, and disregard for popularity among any sector of Israelite society. It was indifferent to economic, material, or other reward in contrast to other types dependent for their livelihood upon prophecy and prediction. They were at best tolerated and not infrequently persecuted.

While not primarily diviners and seers the classical Biblical prophets nevertheless stressed the dire fate that Israel would suffer if it did not keep the terms of its covenant with God. Such was the precarious situation of the Israelite states in Palestine that gloomy predictions were not uncommonly fulfilled. The religious conduct of the Israelite people was linked intimately by the activities of the prophets to their material and political fate and to that of the state.

The conditions which promoted the development and impact of prophetic currents of this type were those of military and political insecurity and unpredictability. Weber considers that new religious insights rarely occur in the main centers of civilization but rather in areas on their periphery which are subject to their influence and pressure. Those who live in such regions repeatedly are given reason to question the meaning of events that affect them. Unlike the inhabitants of long stable, politically secure, and settled regions who tend to take the world for granted, those outside these areas constantly are puzzled by the turn of events; 'the possibility of questioning the meaning of the world presupposes the capacity to be astonished about the course of events' (Weber 1952).

Subsequent research on Biblical prophecy has often been critical of aspects of Weber's analysis and by implication of his analysis of prophecy and prophetism in general. Berger (1963) and Turner (1974) have questioned whether the Biblical prophets were the independent force based upon pure charismatic authority and motivated entirely by religious concerns that Weber portrays them as being. In fact, the classical Biblical prophets are more likely, according to some recent research, to have occupied official positions in the cult of Yahweh. The conclusion that writers such as Turner draw from this, that it requires us to reconsider Weber's understanding of prophetic charisma as an autonomous and personal force and to recognize that it arises from the incumbency of positions in an established social institution is, perhaps, rather extreme. It is likely that there were several different types of prophet in ancient Israelite society. Emmet (1956), who considers that Weber equates prophecy with charismatic authority which she finds too restrictive, usefully differentiates a number of types of prophet. First, there is the institutional prophet who fulfils a traditional and established role. Second, corresponding to Weber's charismatic prophecy, there is the leader of a movement which seeks radical change. Finally, there are prophets who do not attract a personal following but who act as independent critics, advocates of reform or who provide moral leadership. While the prophets associated with the Temple in Jerusalem, whose function was an institutionalized one within the Temple cult, were perhaps more numerous, the classical Biblical prophets with which Weber was concerned may be considered to be of a different type. Whether they enjoyed greater autonomy and independence stemming from personal charisma is still, however, open to question.

Weber's stress upon the charismatic authority of the classical Biblical prophets has been challenged by Petersen (1979) on the grounds that there is no evidence that they were leaders with a group of followers, which he claims is implied by Weber's notion of charismatic authority. Neither is it useful, in his view, to characterize Biblical prophets as office holders. The dichotomy between charismatic leader and office holder is, in any case, he argues, far too strictly drawn. We are led, therefore, to question whether the Israelite prophets could have provided the breakthrough that Weber attributes to them. In a later work Petersen (1981) argues that we may get a better

understanding of the Biblical prophets if we see them as having a specific role. He applies role theory in an analysis of their position in Israelite society based upon a close examination of the use of a variety of terms in Hebrew used to refer to such figures. The results show that there was more than one type of prophetic role in this society. Some held established positions in the official cult and operated also by offering a personal service for a fee. Others, termed central morality prophets, pronounced as individuals at times of crisis but in order to uphold established values and institutions. Still others, peripheral prophets, operated outside institutional contexts and established values.

Also doubtful of the simple dichotomy of functionary vs. autonomous charismatic is Lang (1983). The prophetic tradition, he argues, was largely an institution of the land-owning nobility that enabled them to express publicly their political views. He attempts to show that the prophets were largely from a land-owning, noble background. While they were not chosen by formal procedures but simply emerged, they nevertheless, according to Lang, had to learn how to be an acceptable prophet and to observe the rules governing appropriate behavior for prophets. They were closer, however, to Emmet's third type, in Lang's view, and not wholly established functionaries within the cult. The contrast between charisma and tradition in discussions of prophetism is indeed often overstated. Lang points out that they are often bound up with one another in complex ways such that prophetic charisma can be seen as a force for renewal which emerges within traditional structures at certain times. It is a force that changes and reforms without destroying.

While the Weberian view of Biblical prophecy has, in recent decades, been questioned by some, it has for others stood the test of time and deeper scrutiny rather more robustly. Williams' analysis of sources (1969), e.g., supports the view that many of the classical prophets were independent of the official cult. A detailed study by Wilson (1980) is alert to the complexities and differences between periods and regions. He contrasts the premonarchical Ephraite tradition of the North with the monarchical period in both the Northern and Southern kingdoms and also different periods and circumstances in the South. In the premonarchical North, prophetic behavior was expected to conform to familiar patterns regarding the stereotypical linguistic style in which pronouncements were made and to customary behavioral patterns. This clearly suggests that these prophets occupied a defined position and role within the central religious institutions of the society. In the monarchical period prophets seem to have become figures of less central significance but rather more peripheral to the established religious institutions and rather more concerned with producing change. In Judah, however, political circumstances and alliances differentially affected the type of role and significance that prophets had. Some, standing outside the established cult, were more critical of the status quo while others exercised a more supportive function. This reminds us that prophetism, at least of this type and in these circumstances, may not simply refer to a movement or development led by a prophetic figure but to a dynamic situation of struggle to define or lay claim to the accepted interpretation of a religious tradition or interpretation. Lang (1983) points out that, like many prophets in various traditions and cultures, Jeremiah engaged in conflict with rival prophets opposed to his policy of accommodation with the Babylonians and who advocated total resistance.

Such debates highlight the extent to which Biblical prophecy, and it is an observation applicable to all prophecy, is associated with crises which, while differing extensively in kind, generally call forth a questioning of the status quo and of established institutions and patterns of power and authority. This point has been made by Carroll (1979) in the case of ancient Palestine and Weber's analysis. The religious-political factions that tend to emerge in such circumstances may understandably become aligned with pro-regime and anti-regime types of prophecy.

Not that this necessarily means that prophetism is really a form of politics under the guise of religion. As Albrektson (1972) notes, Weber may well be correct in attributing to the Biblical prophets motives which were primarily religious. Circumstances often made their pronouncements political in tone but at root they were essentially utopian and idealistic rather than pragmatic or designed to stir up political opposition for its own sake.

The political element in prophetism is a factor that comes to the fore most notably in studies of those forms of prophetism that can be described as millennial. It is in the context of the quite extensive studies of such movements that prophetism has been most systematically examined (see *Millennialism*). Not all prophetic movements are millennial and not all millennial movements are prophetic, though most are. It is prophetic millennial movements, however, that have received most attention (see, e.g., Wilson 1975). The role of the prophet in millennial movements varies. Such movements may be initiated by the visions, revelations, and activities of a prophet. Once underway they may call forth a prophet who focuses relatively disordered and uncoordinated activities of participants and who gives coherent expression to previously vague, inchoate, and diverse ideas. The appearance of the prophet transforms millennial ferment into a millennial movement.

The role of the prophet in the millennial movement generally has been found to reflect Weber's emphasis upon prophetism being associated with social and political crisis or radical social change. On the other hand, it is rarely a complete break with the past and more often a process of bridging past and present, the

traditional and the new. The prophetic message in such movements tends to reinterpret old understandings in ways appropriate for an altered situation while at the same time adding new elements and opening up the imagination to new possibilities. The prophet gives concrete expression to what are often inarticulate feelings and ideas and channels them into a program of action promising redemption from current ills. The prophet helps those who are confused by circumstances of change, uncertainty, and instability to comprehend and grasp the meaning of their situation by offering insight derived from special revelation which through understanding points the way forward for remedying what is unsatisfactory. Prophetic movements frequently are found in circumstances of disaffection in which a population is subject to foreign or colonial rule or have little comprehension of the mechanisms and structures of power or how to operate within them in order to ameliorate their situation. The prophet provides the missing knowledge and means of finding solutions. It is significant that many prophetic leaders of such movements are more knowledgeable than their followers as a result of wider experience through travel, limited education, or access to information. The prophet, by providing a focus, is often able to bring together otherwise fragmented communities and diverse groups previously unable to operate in concert in the pursuit of common aims. What is promised, however, is not simply material but also usually, and more significantly, moral and emotional redemption through the recovery of lost personal and cultural dignity and sense of self-worth.

See also: Charisma and Charismatic; Charisma: Social Aspects of; Millennialism; Monasticism: West and East; Religion: Definition and Explanation; Religion, History of; Religious Organizations; Religious Specialists; Weber, Max (1864–1920); Weberian Social Thought, History Of

Bibliography

Albrektson B 1972 Prophecy and politics in the Old Testament. In: Biezais H (ed.) *The Myth of the State*. Almqvist and Wiksell, Stockholm, Sweden

Berger P 1963 Charisma and religious innovation: the social location of Israelite prophecy. *American Sociological Review* **28**: 940–50

Buss M J 1980 The social psychology of prophecy. In: Emerton J A (ed.) *Prophecy*. W. de Gruyter, Berlin

Carroll R P 1979 *When Prophecy Failed: Cognitive Dissonance in the Prophetic Traditions of the Old Testament*. Seabury, New York

Emmet D 1956 Prophets and their societies. *Journal of the Royal Anthropological Institute* **86**: 13–23

Lang B 1983 *Monotheism and the Prophetic Minority: An Essay in Biblical History and Sociology*. Almond Press, Sheffield, UK

Parsons T 1965 Introduction. In: Weber M (ed.) *The Sociology of Religion*. Methuen, London

Petersen D L 1979 Max Weber and the sociological study of ancient Israel. In: Johnson H M (ed.) *Religious Change and Continuity*. Jossey Bass, San Francisco

Petersen D L 1981 The Role of Israel's Prophets. *Journal for the Study of the Old Testament*, Supplement Series, **17**

Turner B S 1974 *Weber and Islam*. Routledge, London

Weber M 1952 *Ancient Judaism*. Free Press, Glencoe, IL

Weber M 1965 *The Sociology of Religion*. Methuen, London

Williams J G 1969 The social location of Israelite prophecy. *Journal of the American Academy of Religion* **37**: 153–65

Wilson B R 1975 *Magic and the Millennium*. Paladin, St Albans, UK

Wilson R 1980 *Prophecy and Society in Ancient Israel*. Fortress Press, Philadelphia, PA

M. Hamilton

Prophylaxis in Psychiatry

In psychiatry, prophylactic drug therapy includes the use of medication for both prevention and maintenance. For the sake of clarity prevention is measures used in order to prevent the occurrence of a first illness-episode in a group of high-risk individuals (primary prevention), or the occurrence of subsequent illness episodes following the occurrence of a first one (secondary prevention). While secondary prevention is used in a variety of psychiatric disorders, primary prevention, especially for the psychotic disorders, is still under investigation. In order to familiarize the reader with both the concept of primary and secondary prevention, the following paragraphs will discuss a variety of strategies suggested for implementation of primary and secondary prevention focusing on their use in schizophrenic illness.

In the absence of current curative treatments in psychiatry, the use of maintenance therapy—which is associated with long-term control of symptoms and illness-episodes—is especially important. Some of the current drug-prophylactic-maintenance strategies will be described that are used to control abnormal behavior and symptoms in three major psychotic disorders: schizophrenia, major depression, and bipolar illness. Since a knowledgeable choice of a proper maintenance drug treatment necessitates a set of predictors for the success or failure of a given drug, this issue will be exemplified below while discussing maintenance treatment for bipolar illness.

Drug therapy, whether preventive or maintenance, depends on patient compliance. This in turn depends on several variables and certain measures can be taken to maximize compliance. This issue will be discussed as well.

Finally, future directions in prophylactic medication treatment for psychiatric disorders will be discussed focusing on strategies for primary prevention.

1. Primary Prevention in Schizophrenia

There is as yet no consensus as to the nature of a subgroup of individuals carrying a high risk for the development of schizophrenia. Research designs used toward this aim involve both retrospective and prospective methods.

Prospective research projects mainly apply long-term follow-up of selected high-risk groups of subjects. Such prospective strategies assume that biological and psychological data obtained from subjects who will later develop the disease will help to identify markers associated with this disease. This, it is hoped, will enable a knowledgeable prediction as to who will develop a schizophrenic illness. One example of such a prospective approach is the follow-up of offspring of schizophrenic patients or their family members (Mednick et al. 1987, Weintraub et al. 1987, Asarnow et al. 1988, Erlenmeyer-Kimling et al. 1993). Another is the examination of psychomotor and neuropsychological variables in large cohorts of individuals while identifying variables associated with later development of psychotic illness and schizophrenia. Yet, another prospective strategy is the study of a variety of factors and early symptoms associated with a transition phase—that is, the phase leading from 'normal' into a psychotic state—in order to identify relevant risk factors. All the above strategies require long-term follow-up. Interestingly, there is also a suggestion for a short-term prospective strategy employed by Yung et al. (1998) aiming at identifying a very high risk group of subjects for the development of psychosis. This strategy utilizes a 6–12 months follow-up of selected subgroups of individuals demonstrating high risk for a manifestation of a schizophrenic episode.

In general, retrospective and prospective studies have identified a host of markers increasing the risk for future psychosis including a family history of schizophrenia, distressed family functioning, early separation and institutional rearing, neurobehavioral deficits, delivery complications, maternal exposure to influenza during the second trimester of pregnancy, along with certain behavioral and personality patterns. The latter are mainly reported by teachers and caregivers. Also, a variety of biological findings were identified, all of which still need to be verified in further research. An example of a biological finding is that of small amygdala-hippocampal complex or thalami in people at high risk of developing schizophrenia (Lawrie et al. 1999). Another finding is that of increased plasma homovanillic acid that is negatively correlated with performance in arithmetic tasks in a group of subjects demonstrating prodromal phase of schizophrenia (Shimushi et al. 2000). All the markers discussed above should be taken at the moment with caution until further research is done. Also, it is our opinion that the use of certain combinations of the above markers for the initiation of antipsychotic treatment in high-risk individuals—although suggested by some (Phillips et al. 1999)—is not recommended at this time as routine treatment.

2. Maintenance Treatment in Schizophrenia

Maintenance therapy in schizophrenia is provided by long-term treatment. Based on accumulated experience, it was suggested that schizophrenic patients could be looked at as two groups. One in which maintenance therapy seems to play a 'suppressive role,' and its discontinuation leads to a rapid reappearance of psychotic symptoms, and another group in which the treatment has more of a 'prophylactic' role. In this last group discontinuation of treatment leads to a more gradual and remote in time relapse of symptoms (see Gelenberg et al. 1991).

It is estimated that about 10 percent of remitted medication free schizophrenic patients will relapse in any given month. That means that after 6 months roughly about half of these patients will relapse whereas after 12 months about 70 percent will do so and only a small portion of these patients will not relapse at the end of 2 years.

Several strategies of maintenance treatment are suggested for these patients. The first consists of a continuous maintenance treatment. This strategy applies relatively lower doses of antipsychotic treatment as compared with that used for acute exacerbation.

The second maintenance strategy includes 'drug holidays,' that is, treatment periods interrupted by drug free periods. This strategy was suggested in order to lower the incidence of side effects while retaining drug efficiency. However, such a strategy is not recommended nowadays since research studies could not support its effectiveness. A third strategy suggests 'intermittent' treatment or 'targeted dose' for chronic schizophrenic patients. In this case, patients are treated during the acute episode only, while monitored for the appearance of prodromal symptoms (first symptoms anticipating the appearance of an exacerbation). Once prodromal symptoms appear, pharmacological treatment is initiated.

In general pharmacological treatment in schizophrenia includes the use of a group of drugs collectively called 'antipsychotic drugs.' These drugs have the ability to block neurons using the neurotransmitter dopamine to transmit messages to other neurons.

In the 1980s, a second generation of drugs came into use. The prototype was clozapine, a drug that besides its ability to block some types of dopaminergic

neurons can also exert a blocking effect on certain serotonergic neurons—(labeled as 5HT-2 neurons) in the central nervous system. This drug was shown to have a lower side-effect profile compared to the more traditional drugs, especially it had minimal extrapyramidal signs (a set of abnormal motoric signs that interfere with daily activities) and it did not elevate plasma levels of prolactin (the lactating hormone). However, clozapine also has several distressing side effects including the lowering of white blood cell count that may expose these subjects to severe infections. Clozapine seems to affect not only the positive illness symptoms but also the negative symptoms of schizophrenia (a set of symptoms characterized by impaired volition and motivation, apathy and paucity of speech) and it may also improve some neurocognitive impairments demonstrated in this disease (see Kane et al. 1988). Beside clozapine, other second-generation antipsychotic drugs were also developed of which mainly olanzepine and risperidone are in use. However, there is still a question whether all these second generation antipsychotic drugs or only some of them can mimic clozapine's effect and if so to what extent.

3. *Maintenance in Bipolar Illness*

This illness course is characterized by the appearance of manic episodes with or without the occurrence of depressive episodes. It is estimated that the majority of patients who had a manic episode will have one or more recurrences. Since both manic and depressive episodes may be associated with progressive deterioration of interpersonal and occupational functioning and may also carry the risk of harming oneself or others, an adequate maintenance drug treatment for bipolar illness is of utmost importance. Several mood stabilizers—that is, drugs designed for maintenance of the euthymic (normal) state while preventing the appearance of both manic and depressed episodes—are now available. The first mood stabilizer to be used was lithium, a drug still considered by many to be the gold standard treatment for bipolar illness. Other commonly used mood stabilizers are valproic acid and its derivatives, and carbamazepine whereas newer drugs such as the anticonvulsants lamotragine, gabapentin, and topiramate are still under study. This article will mainly address data published for lithium.

Three open trials suggested that lithium may be effective in the prophylaxis of bipolar disorders (Baastrup 1964, Hartigan 1963, Baastrup and Schou 1967). During the 1970s, a series of double blind studies were conducted to examine the possible efficacy of lithium in the prophylaxis of bipolar disorder. Goodwin (1994) summarized these studies calculating the recurrence rates over a one year period for placebo (a nonactive treatment) to be 81 percent, but only 34 percent for lithium. These reports included two types of studies. Prospective studies treating patients with either lithium or placebo and discontinuation studies where one group of patients continues with lithium treatment for a given period of time, whereas the other discontinues treatment for a similar time period.

From the 1980s and on more prospective and discontinuation studies were conducted. These studies basically confirmed lithium's efficacy as maintenance treatment for bipolar disorder. They also suggested that there may be an advantage in keeping maintenance lithium blood levels at about 0.6 meq/L as it may lead to less side effects and increase compliance (Coppen et al. 1983) and that lithium plasma levels of 0.8–1.0 meq/L seem to be associated with less relapse (Gelenberg et al. 1989). In general, although many authors still suggest that lithium is an efficient drug for bipolar illness, especially in the classic bipolar patient (Grof 1999), some authors now point to the fact that there is a group of patients who are less responsive to this drug or that demonstrate a decreased response throughout the years.

One important issue associated with long-term maintenance therapy is the finding of predictors for a beneficial, or lack of, response for a given drug. Several predictors of lithium response were offered. Grof et al. (1983a, 1983b, 1999) suggested a better response to lithium treatment in a group of patients demonstrating a course of illness with periods of euthymia (normal range of mood), followed by mania, depression, and then euthymia. These authors also suggested that the absence of rapid cycling (four and more illness episodes per year), the absence of personality disorder, and the diagnosis of primary bipolar disorder (i.e., not secondary to street drugs or organic cause) predict a good prophylactic response to lithium. Faedda et al. (1991) in agreement with Grof suggested that mania-depression-euthymia course is associated with a good response to lithium. Others suggested that good response to lithium in relatives and a family history of mood disorders carry a good response. Aagaard and Vestegaard (1990) suggested that bipolar patients tend to demonstrate decreased response to lithium treatment when they have a lengthy episode, they are young or of female gender. Abou-Soleh and Coppen (1990) suggested that positive prophylactic response to lithium over 6–12 months predicts further good response in the future. Patients with dysphoric and mixed mania (manic episode associated with some depressive symptomatology) were suggested to respond less to lithium treatment and some studies suggest this is also true for rapid cycling bipolar disorder. Hopefully, a knowledgeable use of predictors may help the physician to choose selected subgroups of patients, which will benefit from lithium maintenance treatment.

4. *Maintenance in Major Depression*

Depression is a serious illness. It carries high suicidal risk. It is estimated that the recurrence of major depression is between 70–85 percent. Two types of

studies report the efficacy of antidepressants to prevent relapse in the first year after remission from a depressive episode. The first is that of a cross-over design, whereas the second is that of a continuation design. In the first design patients are first brought into remission with a given medication and are then maintained on this medication for a period of 2–3 months. Patients who respond are then entered into a double blind design for up to another year, in which they receive randomly either the same antidepressant or placebo. In the continuation design, acutely depressed patients are assigned randomly to either active drug or placebo. Patients who respond continue with the same treatment. Eighteen studies compared a group of antidepressants collectively called tertiary amine tricyclics (this name describes the common molecular structure characterizing this group of drugs), finding that while 50 percent relapsed on placebo only 23 percent relapsed on the active drugs. Three studies compared another group of newer antidepressants collectively called selective serotonin reuptake inhibitors (SSRIs) (this name describes the hypothesized mechanism of action of this group of drugs) finding that the relapse rate for the active drug was between 13–26 percent whereas for placebo it was 46–57 percent (see Janicak et al. 1997).

In general, it is agreed that prevention of relapse in major depressive illness is of major importance and every effort must be taken to address this. Contrary to previous reports that supported the use of low doses in the prophylactic treatment of major depression, several authors now recommend that the full dose be used, that is, a similar dose to that used for the acute treatment of a major depressive episode.

In general, it is recommended to continue maintenance treatment after a first episode of major depression for about 6 months and then to taper the drug gradually. If another episode occurs the recommendation is to continue treatment for several years and after a third episode a lifetime maintenance treatment is recommended.

5. *The Issue of Compliance and its Importance in Prophylactic Drug Treatment*

Compliance to drug treatment is defined as adherence to an agreed treatment schedule between a physician and the patient. It is estimated that no more than 50 percent of patients fully comply with their treatment and we suspect that the magnitude of non-compliance may be even greater in the case of long-term maintenance treatment. Various factors seem to influence compliance including the nature of the patient-physician relationship, the patient's set of beliefs and expectations as well as the occurrence of side effects. Issues relevant to compliance in psychiatric disorders are: (a) the occurrence of delusions and severe cognitive distortions about the index drug, its effect and side effects, (b) the patient's cognitive functioning and motivation, which may be impaired in certain psychiatric disorders, and (c) the patient's general attitude toward mental illness. Especially important in this context is the level of stigma felt by the patient; the greater the stigma, the less compliant the patient will be.

Several remedies can be offered to enhance compliance and they should be tailored according to the patient's condition and level of understanding. First, an empathic approach, one that builds basic trust with the patient and their family is of utmost importance. Second, whenever possible, the patient with the guidance of their physician should be an active participant in decisions made for the treatment plan. This does not mean that the patient decides or guides such decision making, but that they are presented with several treatment alternatives and an educated decision is agreed upon with the guidance of the physician. Third, the patient should be educated as to side effects and ways to cope with them once they appear (i.e., lower the drug by a certain degree and contact your physician). Treatment diaries and involvement of family members whenever necessary can also be of help. Since maintenance treatment is a long-term treatment, it is our opinion that a responsible physician should renew such a drug treatment 'agreement' at least every several months, discussing it with the patient and perhaps with relevant family and community members if needed. The physician should also actively ask in each meeting with the patient about side effects as well as examine whether there is any change in the patient's set of beliefs about the drug and its action that may need renewed education.

6. *Future Directions for Primary Prevention of Psychiatric Disorders*

Genetic epidemiological studies have shown that genes play a major role in several psychiatric disorders including schizophrenia, bipolar illness, and major depression. It is now believed that these disorders are not related to a single gene, but rather are determined polygenetically. In general, it seems that gene-environment interactions are involved in the causation of many psychiatric disorders. Interestingly, the vulnerability to environmental influences may also be genetically determined (see Owen et al. 2000 for review).

The human genome project, now completed, is expected to contribute to the finding of genes responsible for the susceptibility, occurrence, and maintenance of these disorders. The sorting out of genetic variations in the population, which are related to these psychiatric disorders, may open new avenues to genetic medical interventions prior to the manifestation of these diseases. Such interventions may alter the course or even prevent these diseases from

occurring. Although this project is still far off, one cannot ignore it at the dawn of this new era, which started with the recent completion of the decoding of the human genome.

See also: Antidepressant Drugs; Bipolar Disorder (Including Hypomania and Mania); Depression; Drugs and Behavior, Psychiatry of; Patient Adherence to Health Care Regimens; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Public Psychiatry and Public Mental Health Systems; Schizophrenia; Schizophrenia and Bipolar Disorder: Genetic Aspects; Schizophrenia, Treatment of

Bibliography

Aagaard J, Vestegaard P 1990 Predictors of outcome in prophylactic lithium treatment: a two-year prospective study. *Journal of Affective Disorders* **18**: 259–66

Abou-Saleh M T, Coppen A J 1990 Predictors of long term outcome of mood disorder on prophylactic lithium. *Lithium* **1**: 27–35

Asarnow J R 1988 Children at risk for schizophrenia: Converging lines of evidence. *Schizophrenia Bulletin* **14**: 613–31

Baastrup P C 1964 The use of lithium in manic depressive psychosis. *Comprehensive Psychiatry* **5**: 396–408

Baastrup P C, Schou M 1967 Lithium as a prophylactic agent. *Archives of General Psychiatry* **16**: 162–72

Coppen A, Abou-Saleh M, Milln P, Bailey J, Wood K 1983 Decreasing lithium dosage reduces morbidity and side effects during prophylaxis. *Journal of Affective Disorders* **5**: 353–62

Erlenmeyer-Kimling L, Cornblatt B A, Rock D, Roberts S, Bell M, West A 1993 The New York High Risk Project: Anhedonia, attentional deviance and psychopathology. *Schizophrenia Bulletin* **19**: 141–53

Faedda G L, Baldessarini R J, Tohen M, Strakowski S M, Waternaux C 1991 Episode sequence in bipolar disorder and response to lithium treatment. *American Journal of Psychiatry* **148**: 1237–9

Gelenberg A J, Bassuk E L, Schoonover S C (eds.) 1991 *The Practitioner's Guide to Psychoactive Drugs*, 3rd edn. Plenum, New York, pp. 170–6

Gelenberg A L, Kane J M, Keller M B, Lavori P, Rosenbaum J F, Cole K, Lavelle J 1989 Comparison of standard and low serum levels of lithium for maintenance treatment of bipolar disorder. *New English Journal of Medicine* **321**: 1489–93

Goodwin G M 1994 Recurrence of mania after lithium withdrawal. *British Journal of Psychiatry* **164**: 149–52

Grof P 1999 Excellent lithium responders people whose lives have been changed by lithium use. In: Birch W J, Gallictio V S, Becker K W (eds.) *Lithium: 50 Years of Psychopharmacology*. Weidner Publishing Group, Cheshire, CT

Grof P, Alda M, Grof E, Fox D, Cameron P 1999 The challenge of predicting response to stabilizing lithium treatment. *British Journal of Psychiatry* **163**(21): 16–19

Grof P, Hux M, Grof E, Arato M 1983a Prediction of response to stabilizing lithium treatment. *Pharmacopsychiatry* **16**: 195–200

Grof P, Lane J, Hux M 1983b Predicting the response of long term lithium treatment: Interim results of a prospective study, in the prediction of lithium response. In: Dufour H, Pringuey D, Milech T (eds.) *Proceedings from Symposium in Marseille, France; June 2–3, 1980*. France University, Marseille, pp. 43–53

Hartigan G P 1963 The use of lithium salts in affective disorders. *British Journal of Psychiatry* **109**: 810

Janicak P G, Davis J M, Preskorn S H, Ayd F J 1997 *Principles and Practice of Psychopharmacotherapy*, 2nd edn. Williams & Wilkins, Baltimore, MD, pp. 295–301

Kane J, Honigfeld G, Singer J, Meltzer H 1988 Clozaril collaborative study group: clozaril for the treatment-resistant schizophrenic: a double-blind comparison with chlorpromazine. *Archives of General Psychiatry* **45**: 789–96

Lawrie S M, Whalley H, Kestelman J N, Abukmeil S S, Byrne M, Hodges A, Rimmington J E, Best J J, Owens D G, Johnstone E C 1999 Magnetic resonance imaging of brain in people at high risk of developing schizophrenia. *Lancet* **353**(9146): 30–3

Mednick S A, Parnas J, Schulsinger F 1987 The Copenhagen High-Risk Project, 1962–1986. *Schizophrenia Bulletin* **13**(3): 485–95

Owen M J, Cardo A G, O'Donovan M C 2000 Psychiatric genetics: Back to the future. *Molecular Psychiatry* **5**: 22–31

Phillips L, Yung A, Hearn N, McFarlane C, Hallgren M, McGorry P 1999 Preventive mental health care: accessing the target population. *Australia and New Zealand Journal of Psychiatry* **33**: 912–17

Shimushi T, Kurachi M, Kurokawa K, Yotsctsugi T, Urehara T, Itch H, Saitoh O 2000 Plasma homovanillic acid in the prodromal phase of schizophrenia. *Biological Psychiatry* **47**: 428–33

Weintraub S 1987 Risk factors in schizophrenia: The Stony Brook High-Risk Project. *Schizophrenia Bulletin* **13**(3): 439–50

Yung A R, Phillips L J, McGorry P D, McFarlane C A, Francey S, Harrigan S, Patton G C, Jackson H J 1998 Prediction of psychosis. A step towards indicated prevention of schizophrenia. *British Journal of Psychiatry* **172**(Suppl.): 14–20

J. Levine and S. Gershon

Proportional Representation

As a standard for making political evaluations, 'proportional representation' implies that members of a group (R_g) participate in a governmental entity in numbers that bear the same ratio to the total membership of that body (R) as persons identifying with the group (P_g) have to the total population governed (P)—thus, $R_g/R = P_g/P$. The group in question may be supporters of a political party, adherents of an ideological or policy tendency, or persons sharing an identity based on race, ethnicity, religion, language, or gender. 'Proportional representation' also, and more commonly, refers to methods of electing representatives designed to achieve, or at least approximate, the ideal of proportionality in converting citizens' votes into legislative seats. In this article 'proportional representation' (PR) will indicate electoral systems, while the standard at which they aim will be referred to as 'proportionality.' (For related articles, see *Electoral Systems*; *Parliamentary Government*; *Political Representation*; *Voting: Tactical*; *Voting: Turnout*.)

1. PR Electoral Systems

There are two basic types of PR electoral systems—list PR and the single-transferable vote (STV). All PR systems require multimember districts—constituencies from which voters elect more than one representative. Thus, except in the trivial sense that the winner should have more votes than any loser, proportional representation does not apply to elections that choose a single winner, such as for executive posts or legislatures based on single-member districts (SMDs). The degree to which a PR electoral scheme can approximate perfect proportionality depends on the number of representatives elected per constituency, or district magnitude (M). In the limiting case, the entire nation constitutes a single constituency, as in Israel and The Netherlands, where $M = 120$ and 150, respectively.

1.1 List PR

In the pure form of list PR, the closed list, political parties designate before the election a rank-ordered list of M (or fewer) candidates. Electors vote for a party, not for individual candidates. Parties winning seats select representatives from their lists according to the predetermined ranking.

The number of seats each party wins is determined by a formula that employs one of two basic mechanisms—quotas or divisors. In quota (or largest remainder) formulas, parties receive an initial allocation of seats equal to the number of quotas achieved in the vote for their list. The quota is determined by dividing total votes by M (Hare or natural quota), $M+1$ (Droop quota), or $M+2$ (Imperiale quota). Any remaining seats are assigned according to largest remainders—i.e., by the size of parties' leftover votes after subtracting the quotas they have used. In divisor (or highest average) methods, seats are assigned sequentially. Each seat goes to the party which, at that point in the allocation, has the largest quotient after its votes are divided by a number that varies with the seats the party has already received (S). Well-known divisors include $S+1$ (1, 2, 3, ...), the *d'Hondt* formula, and $2S+1$ (1, 3, 5, ...), the *Sainte-Laguë* formula.

Formulas inspire political controversy, because different quotas and divisors favor large, medium, or small parties. Nevertheless, district magnitude has a more powerful effect on parties' fortunes; small parties generally do better as M rises. Some list PR systems incorporate another element to exclude small parties—a threshold of votes (e.g., five percent) that a party must achieve to win seats.

List PR systems are employed for national parliamentary elections in most countries of Latin America and continental Europe.

1.2 The Single-transferable Vote

STV, also known as the quota-preferential system, depends on a preferential ballot, in which voters rank individual candidates from most to least preferred. A quota of votes, Q, is established, usually by dividing the total number of voters by $M+1$. All candidates receiving Q or more first preferences are elected. To determine the remaining winners, two types of transfers are carried out. In downward transfers, surplus votes (those above Q) received by winning candidates are reallocated to voters' second (or, if necessary, subsequent) choices. In upward transfers, the candidate ranking last in votes is eliminated, and ballots giving first preferences to him or her are reassigned to the voters' second (or subsequent) choices. Transfers are iterated until M candidates have achieved the quota.

Perhaps because of its complexity, both in balloting and selecting winners, STV is employed in national elections by just a few countries—Ireland, Malta, and (for the Senate only) Australia.

1.3 Parties and Candidates in PR Voting Systems

In its pure form, list PR assigns control over the fate of individual candidates to parties' internal decision processes, whereas STV reserves that power for voters. In addition, list PR forces voters to choose just one party (or sometimes an electoral alliance of parties), whereas STV allows voters to move back and forth across parties while ranking candidates. Thus, list PR strengthens party organization and promotes party discipline in the legislature, whereas STV is conducive to intraparty electoral competition and individualistic behavior among legislators.

In practice, however, these differences may be blurred. Many list PR countries use open lists, in which electors are permitted (or, as in Finland, required) to indicate preferences for candidates as well as a party. In the *panachage* system of Switzerland and Luxembourg, electors may even distribute votes for candidates across different party lists. Conversely, in STV systems, unless district magnitude is kept low (on the order of three to five representatives per constituency), the task of ranking numerous candidates becomes so difficult that most voters will depend on instructions from parties. Thus, if it imposes excessive cognitive demands, a system intended to maximize the autonomous power of individual voters may end up functioning rather like list PR.

2. Mixed-member Proportional

In an effort to combine the virtues of proportionality and local representation, Germany after World War II pioneered the mixed-member proportional (MMP)

electoral system. In MMP, a share of representatives (half in Germany) are elected from single-member districts, and the remainder are elected from party lists. Electors cast two votes—one for an individual constituency representative and one for a party. Seats that parties win from constituencies are subtracted from their list allocations to make overall delegations proportional to party votes. This crucial compensatory feature makes MMP a truly proportional system.

3. *Semi-proportional Systems*

Another mixed type, easily confused with mixed-member proportional, is the parallel system. Like MMP, parallel systems elect some legislators from single-member constituencies and some from PR lists. Parallel systems diverge from MMP by not compensating for constituency victories in allocating seats to party lists. In effect, the legislature is chosen by an amalgamation of two separate electoral systems, only one of which aims at proportionality. If, as is usually the case, seats chosen by PR comprise no more than half the assembly, then the parallel system is appropriately classified as semiproportional. Japan and Russia adopted parallel systems in the 1990s.

The term semiproportional is also applied to electoral systems that are intended to achieve a degree of proportionality, but do not employ PR formulas to do so. These systems use multimember districts (of low magnitude), voting for individual candidates, and the plurality formula. The opportunity for proportionality arises from their balloting rules. The limited vote requires electors to cast fewer than M votes. In the extreme version case, the single-nontransferable vote (SNTV), electors may vote for only one candidate, even though M will be elected. Japan employed SNTV until 1993. Cumulative voting systems give electors M votes, but allow them to cast multiple votes for a favored candidate or candidates. Limited voting and cumulative voting, which have similar underlying strategic properties, enable well-coordinated minorities to win seats. However, because parties must avoid dispersing votes among too many candidates, these systems tend to reduce competition, and such competition as does occur tends to be within parties or voting blocs, causing internal conflict and factionalism.

4. *Effects and Evaluations*

The obvious intended benefit of PR systems is to establish what advocates consider an essential condition of electoral fairness—a close correspondence between parties' votes and seats. The use of a PR formula, however, does not preclude significant deviations from proportionality when thresholds exist or district magnitudes are low, especially if parties have not yet adapted strategically to the parameters of the electoral system.

Although the immediate effect of list PR is to promote proportionality for political parties rather than underlying population groups, most parties will compete for votes of minority groups by placing members in electable list positions. If they do not, groups large enough to surpass the threshold can win representation by organizing their own parties. Consequently, list PR systems usually offer improved representation to minorities. The same incentives apply to electing women. Where they have become politically active and ambitious, women generally win more legislative seats under list PR than under single-member district (SMD) systems.

Because PR is less likely than SMD to waste votes cast for smaller parties, it usually results in multiparty parliaments (see *Party Systems*). The association between PR and multipartyism must be regarded as permissive rather than automatic. Within the constraints imposed by the electoral framework (district magnitude, threshold, and formula), the number of parties depends on the number of societal cleavages and on the willingness and ability of politicians to exploit them by organizing parties (see *Cleavages: Political*). Proportionality and multipartyism together make one-party majorities rare, so parliaments elected by PR typically form coalition governments or minority governments that must get support from other parties to pass legislation.

Critics of PR usually concede its representational advantages, but they contend that it produces a fragmented party system, unstable coalition governments, and excessive influence by extremist parties. Their favorite negative examples include Italy (before 1993 reforms), Israel, and Weimar Germany. In response, defenders of PR recommend 'moderate PR' with significant thresholds or districts of intermediate magnitude, rather than the pure forms those three cases exemplify. They defend the need for coalitions under PR as fostering consensus politics and stable policies, if not always stable cabinets.

5. *Historical Development*

Because PR methods are non-obvious and complex, their development stands as one of the key intellectual achievements in the history of democracy. The Marquis de Mirabeau enunciated the principle of proportionality in 1789, but methods designed to achieve it were not refined until well into the nineteenth century. In 1834 Victor Considérant devised a list PR system that Antoine Morin of Switzerland developed further in 1862. In 1885 an international conference on proportional representation in Antwerp recommended the list PR method proposed in 1882 by the Belgian Victor D'Hondt. In 1899, Belgium adopted D'Hondt's system and thus became the first country to use list

PR. By 1920 most countries of western Europe had adopted versions of PR (Carstairs 1980).

Beside the intellectual and moral appeal of proportionality, two practical motives spurred the turn to PR in Europe. In societies divided along linguistic or religious lines, such as Switzerland, Belgium, and The Netherlands, PR offered all groups a stake in democracy (see *Consociationalism*). For countries about to expand the franchise, PR assured ruling conservative parties that they would not be shut out of parliament, as might have occurred under existing electoral systems once previously disenfranchised citizens became a majority of voters. These motives also influenced later decisions to adopt PR, including the notable case of South Africa in 1994.

Perhaps because of their continuous history of elections predating the development of PR, the UK and the USA (as well as some countries they influenced) have clung to the older, simpler method of single-member districts and plurality rule (see *First Past the Post, in Politics*). Proponents of PR in English-speaking countries have usually favored STV, reflecting its invention by the Englishman Thomas Hare in 1857 and John Stuart Mill's enthusiastic advocacy. That tradition was broken in 1993 when New Zealand adopted MMP.

Agreement on a list PR system was often a key ingredient in democratizations during the late twentieth century, as in Portugal, Spain, South Africa, and Indonesia. However, in the 1990s the global trend shifted to mixed electoral systems, both the semi-proportional parallel version and fully proportional MMP. By 1997, of the world's 211 independent states or semiautonomous territories, 75 (35 percent) used PR systems (list, MMP, or STV) to elect the principal chamber of the national legislature. Countries using PR comprised 46 percent of states respecting political rights and civil liberties, and 59 percent of well-established democracies (Reynolds and Reilly 1997). In many other nations, PR is used for elections to upper chambers, regional and local councils, or supranational authorities, such as the European Parliament.

See also: Electoral Systems; Party Systems; Political Representation; Voting: Tactical

Bibliography

Carstairs A M 1980 *A Short History of Electoral Systems in Western Europe*. Allen & Unwin, London

Cox G W 1997 *Making Votes Count*. Cambridge University Press, Cambridge, UK

Lijphart A 1994 *Electoral Systems and Party Systems*. Oxford University Press, Oxford, UK

Reynolds A, Reilly B 1997 *The International IDEA Handbook of Electoral System Design*. International Institute for Democracy and Electoral Assistance, Stockholm

Taagepera R, Shugart M S 1989 *Seats and Votes*. Yale University Press, New Haven, CT

J. H. Nagel

Propositional Representations in Psychology

In cognitive psychology the term 'proposition' refers to a truth-valued unit of meaning that is represented symbolically in the human cognitive system. Propositions have two important components. First, they consist of semantic information structures that represent aspects of experience (e.g., events, states), real and hypothetical situations in the world, and abstract conceptual knowledge. Second, they include truth value and modality information that represents an attitude or belief on the part of a speaker, writer, or any holder of the proposition, with respect to the truth or falsity of the semantic relationships that are represented by the proposition. Propositions function as a semantic base for natural language where they represent types of semantic information that are distinguished in natural languages; and they function as a base for reasoning, serving as truth-valued and quantified predicates to which logical reasoning operations may be applied. This article examines propositional representations from a psychological point of view, the functions of propositions in the human cognitive system, models of propositional representations, and a closely related discourse analysis and model-building technique in cognitive science known as 'propositional analysis.'

1. Propositional Representation in Cognitive Systems

1.1 Propositions as Cognitive Representations

Like all cognitive representations, propositions are a form of semantic representation. As such, they may be used to represent concrete situations of cognition and action in the world, or they may represent more general or abstract conceptual knowledge that is a part of one's long-term store of knowledge about the world and how to act in it. However, in contrast to propositions, systems of cognitive representation are usually thought of as large networks of concepts and relations that reflect a person's comprehension and knowledge. Such networks are often referred to as 'mental models,' 'situation models,' 'conceptual or semantic networks,' 'schemata,' 'knowledge frames,' or in terms of other similar constructs (see Rumelhart and Norman 1988 for a review).

For example, when a visual scene is interpreted, it is represented semantically in the cognitive system as a

network in which objects are identified and represented (as nodes), their properties are represented by links to attributes, and their relationships to each other are represented by particular types of semantic links. To talk about a scene requires the construction of propositions. These consist of truth-valued assertions about information from this network. The validity of these assertions can be evaluated and they can consequently serve as a basis for discussion, inferences, and reasoning. A major task for a theory of cognitive representation is to identify the types of concepts and relations that are used by humans to interpret and represent their physical world as it is interpreted and described through natural language.

An essential property of systems of cognitive representation is their relationship to the particular situation or situations in the world that they represent. Thus, systems of cognitive representation must support evaluations of their validity as representations of (real or hypothetical) situations in the world, or in relation to bodies of shared abstract or concrete knowledge (e.g., in a scientific field). This semiotic property of cognitive representations underlies the truth value and modality information present in propositions, since it is information about asserted truth conditions that permits discussion and reasoning about perceived situations in the world. Consequently, the study of propositional representations in natural language processing, and of reasoning and inference through natural language, has provided psychologists with models of semantic representation that have had a large impact on theories of cognitive representation in general (see Frederiksen 1975, 1986, Kintsch 1974, Rumelhart and Norman 1975, van Dijk and Kintsch 1983).

1.2 The Influence of Other Disciplines on Cognitive Theories of Propositional Representation

Propositions are closely associated with the production and comprehension of natural language, where they represent semantic information that is asserted by a speaker or writer and explicitly encoded in language structures and discourse. Researchers studying language comprehension have shown that in addition to interpreting the propositions encoded in lexical and syntactic structures, understanding discourse also involves construction of inferred propositions. Inferences are used to complete, elaborate, and connect local propositional information into a coherent text representation, and relate a text's meaning to prior knowledge and to its situational context (Denhière and Rossi 1991, Graesser and Bower 1990, Kintsch 1998). The construction of propositional representations, whether through interpretation of natural language or through reasoning about situations in the world, is a fundamental aspect of the cognitive processes that enable us to understand and interpret the world. Thus, propositions are a fundamental aspect of the human cognitive system where they function in a variety of ways by providing the basic representational units of meaning in cognition and language.

Cognitive theories of propositions as units of semantic representation have been influenced by work from many fields including: linguistics (case grammar, linguistic semantics, text linguistics, pragmatics, systemic linguistics); logic; model theory; situation semantics; philosophy of language (speech act theory, semiotics); and computer science (artificial intelligence and language understanding, conceptual graph and semantic network representations, frame theory, scene analysis, and computational linguistics). Other relevant disciplines include ethnography of communication (conversational inference); cognitive sociology (conversational analysis, ethnomethodology); and cognitive linguistics. For further reading on work in computer science, see Brachman and Levesque 1985; for a multidisciplinary overview, see van Dijk 1997. Fauconnier (1997) discusses recent work in cognitive linguistics and related fields.

1.3 The Role of Propositions in the Architecture of the Cognitive System

The architecture of the cognitive system consists of its general properties as an information-processing system (e.g., Simon 1996 and other papers in the same volume). These include systems for: processing input information; constructing and storing internal symbolic codes; interpreting input information; accessing, using, and modifying stored prior knowledge; operating on internal representations (e.g., to understand, reason, and solve problems); and producing appropriate actions and results (including language) as output. The cognitive architecture that supports natural language processing capabilities consists of a stratified semantic processing system in which propositions are intermediate semantic units that are involved in producing and comprehending natural language in relation to its situational context and to the participants' long-term store of conceptual and relational knowledge. Figure 1 identifies the major elements of this stratified processing system and the central role that propositions play in comprehending and producing natural language discourse. The major levels of representation and processing are:

Natural language discourse: phonetic or lexical patterns; word/morpheme sequences; syntactic structures; linguistic forms signaling discourse cohesion; linguistic and extralinguistic markers (e.g., gestures, stress, intonation) that relate discourse constituents to events, states, and representations in the spatio-temporal environment and signal the situational cohesion of the discourse; and (if the discourse is interactive) speech acts and conversational structures.

Propositional representations: representations of concepts and relational semantic information explicitly encoded in natural language expressions (the text

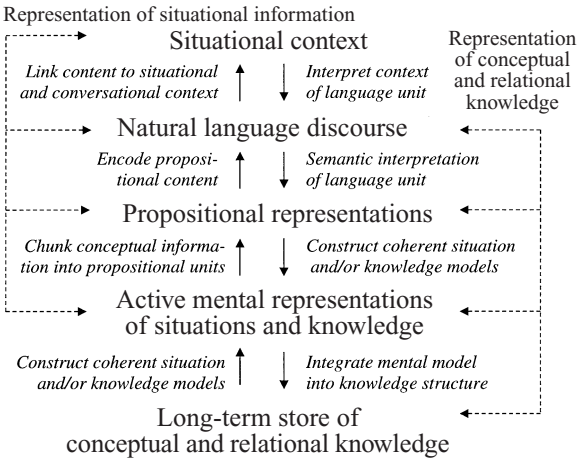

Figure 1
Levels of processing in the production and comprehension of situated discourse

base); inferred propositions that elaborate and connect local propositions in the text and resolve anaphoric references; coherence inferences; macropropositions that summarize content; propositions generated by logical rules to represent the logical structure of arguments; and propositions that resolve extralinguistic references to the situational context.

Active mental representations of situations and knowledge: conceptual networks constructed to link propositional information into a contextually appropriate connected and coherent cognitive representation—a currently active mental model, situation model, or knowledge structure that is constructed based on the text propositions, situational information, and prior knowledge.

In a stratified semantic processing system, these multiple representations are constructed in a situational context that includes a physical environment (of things, states, actions, events, processes, sequences, etc.) and a representational environment (of symbols, language, icons, images, texts, etc.). They also reflect an individual's use of prior conceptual and relational knowledge that is potentially relevant to the situation and content of the discourse.

2. *Cognitive and Communicative Functions of Propositions*

2.1 *The Semantic Content of Linguistic Expressions*

Propositions represent the semantic content of linguistic expressions. For example, in 'The airplane was rising swiftly,' the proposition would represent the structure of an event involving an action (rising), an object undergoing change as a result of the action (the airplane), and an attribute of the action (swiftly). It would also specify how the airplane is determined (definite) and quantified (singular), and the truth assertion about the proposition (positive truth value). To see how this can become more complex, in 'Laura felt the airplane rising swiftly,' the previous proposition would have to be embedded so as to reflect its semantic role as the thematic content of a process (feel) of a patient or experiencer (Laura), and this proposition would have its truth value specified and objects (Laura) quantified. As can be seen in Fig. 1, producing utterances in a language involves the specification of propositions that 'chunk' conceptual and semantic information taken from situation or knowledge models into propositional units so that the information can be encoded morphologically and syntactically in a particular language.

2.2 *Units of Meaning for Comprehension, Inference, Reasoning, Thinking, and Learning Processes*

In comprehension, the end product is a mental representation of a situation and knowledge. The direction of processing can be data-driven (when the flow of processing is from semantic interpretation of language units, to propositional inferences, to the construction of a situation model), or it can be knowledge-driven (when prior conceptual knowledge or knowledge frames are applied to guide and constrain the construction of a situation model that incorporates propositional information from a text base). Processes of inference and logical reasoning can operate on propositional predicates, or they may be based on mental models that are constructed to facilitate a more inductive style of reasoning based on concrete examples (Johnson-Laird 1996). Thinking processes involve the construction, activation, and modification of situation models and propositions, and learning involves the incorporation of propositions and information from active situation and knowledge models into a more permanent knowledge structure.

2.3 *Building Blocks for Systems of Cognitive Representation*

Propositions provide the building blocks for systems of cognitive representation. If knowledge is to be communicated through language or treated as an object of logical reasoning, then it must at some point be propositionalized. Consequently, there should be a close relationship between propositions and systems of cognitive representation. Thus, cognitive psychologists have adopted the strategy of building models of knowledge structure to incorporate semantic relationships that are present in the propositional semantic structures that are expressed in natural language (Frederiksen and Breuleux 1990, Rumelhart and Norman 1988).

2.4 The Semantic Representation of Situations and Knowledge

The propositional information that we construct to describe situations through language is similar to the semantic structures we use to interpret and construct cognitive representations of situations and knowledge. Thus, a primary function of propositions and situation models is to represent and continually update our representations of knowledge and of the currently experienced situational contexts in which we are participating.

2.5 Communication, Conversational Inference, and Situated Cognition

In social environments of collaborative action, reasoning, and problem solving, the main channel of communication is conversational dialog. Participation in such task-oriented dialogs demands conversational inferences by the participants, and these are based on comprehension of the propositional content of the speech as it occurs within the situation and with reference to relevant domain knowledge. Communication and collaborative action in such environments depends on comprehension of the propositions asserted by the participants in the conversation. Thus, propositional representations play a central role in such situations of conversational interaction and socially situated cognition (see Koschmann 1999 for an example).

3. Models of Propositional Representations

3.1 Types of Propositions

Propositions represent a variety of types of semantic information that may be expressed in language. Types of propositions have been defined in terms of the particular semantic relations and concepts that are involved in a particular type of situation that is being represented. A listing of some of the types of propositions that have been identified can help convey the richness of propositional representations.

Events: Events are structures involving an action that produces a change, its causes, results, goals, instruments, and other information specifying the participants and objects effected, the internal structure of an event, and properties of the action, e.g., 'The man walked from the top of the hill to the cemetery gate.'

Processes: Processes identify a process and the objects or persons who experience or are involved in the process; they include cognitive processes having thematic content (e.g., know, believe) and represent the structure of a complex process and its properties, e.g., 'The war raged.' 'John knows Sally.' 'Laura felt the airplane rising gradually.'

States: States identify objects and specify properties of the objects (e.g., attributes, parts, set relations) as well as their determination and quantification, e.g., 'The painting is beautiful.' 'All of the visitors were tourists.' 'The painting is on the wall.'

Abstract relations: Abstract relations identify abstract concepts (corresponding to propositions) and their properties, e.g., 'This theory is complex.' 'The diagnosis was not obvious.'

Identities: Identity relations link objects or propositions into identity sets, e.g., 'The American President is George W. Bush.'

Algebraic relations: Order and equivalence relations are involved in comparative statements and apply to variables that represent values identified in other propositions, e.g., 'John is taller than Mary.' 'His temperature is normal.'

Functions: Functions represent operations defined on variables that return values, e.g., 'Ottawa is close to Montreal.' 'Montreal is closer to Toronto than to New York.' 'The mean height of the team members is six feet.'

Dependency relations: Dependency relations make one proposition depend on another, such as by causal, conditional, and logical implication relations, e.g., 'He won because he had trained intensively.' 'If you finish your supper, then you can have dessert.'

Conjoint relations: Conjoint relations, such as 'and,' alternating 'or' and exclusive 'or,' specify intersecting, alternating, or mutually exclusive sets of objects or propositions.

Tense and aspect components: These specify temporal properties of propositions, e.g., 'He was attending the opera.'

Modality and truth value components: These specify the asserted truth conditions of propositions, e.g., 'He may not have been at the Opera.'

Determination and quantification components: These specify the determination and quantification of concepts, e.g., objects, and of abstract concepts, e.g., 'Six of the men volunteered,' 'All people are mortal.'

3.2 Predicate-argument vs. Conceptual Graph Structures

Two sorts of formalisms have been used to represent propositions: predicate-argument notation, and conceptual graphs (Sowa 1984). Predicate-argument formalisms use a list notation in which a proposition is represented as a predicate (e.g., event) followed by a list of its arguments (the components of the event: the action, the object of the action, the result, etc.). Information about the determination and quantification of objects may be embedded in its argument slot. For example, consider the following sentence taken

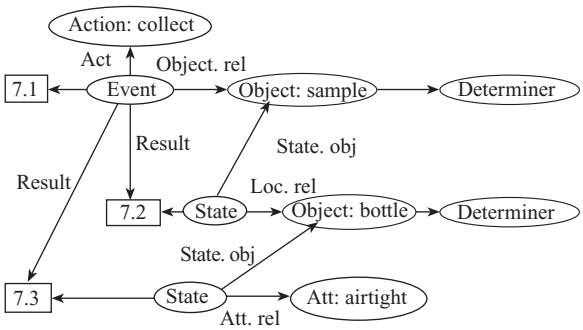

Figure 2
Conceptual graph of 'Collect the samples in an airtight bottle' (adapted from Frederiksen and Breuleux 1990)

from a procedural text in chemistry: 'Collect the samples in an airtight bottle.' An event proposition for this sentence might be as follows (in predicate notation):

7.1 Event: Act (*collect*), Object (*samples*: Def, Plural), Result (7.2, 7.3), Truth.Value (Positive);

The result consists of pointers to two embedded state propositions representing the location of the samples after the action has taken place and a property of the bottle:

7.2 State: Object (*samples*: Def, Plural), Locative. Relation (Object (*bottle*: Token, Singular)), Truth. Value (Positive);

7.3 State: Object (*bottle*: Token, Singular), Attribute.Relation (*airtight*), Truth.Value (Positive)

The alternative conceptual graph notation for this example is given in Fig. 2. Conceptual graphs are equivalent but have the advantage of helping to visualize the semantic relationships being represented by the propositions. This notation is particularly useful in developing models of knowledge structures in particular domains being studied.

3.3 Propositions as Canonical Frames or Semantic Grammars

If propositions are constructed in interpreting and producing language, and in making inferences, how can we explain how they are generated? Two types of explanation have been considered: the canonical frame explanation, and the semantic grammar explanation. It is possible to think of propositions as a vast collection of frame structures which contain 'slots' (or variables) that can be 'bound to' particular conceptual information.

If we replace the arguments in the event proposition by variables, we have an event frame for the action collect which can be applied to many different situations by 'binding' the variables to particular 'values': 'Collect the butterflies in the jar' (Y = butterflies, Z = a state representing the butterflies' location in the jar).

Event: Act (*collect*), Object (Y: Determiner, Quantifier), Result (Z), Truth.Value (Positive);

A conceptual frame theory requires a very large store of such frames in memory and processes to apply them to new linguistic or situational data.

Others have argued that representing propositions as a set of frames is insufficiently powerful to represent the full range of propositional representations that can be constructed. It is argued that like natural language structures, propositional structures reflect rules of formation, and propositional representations need to be defined in terms of knowledge representation languages or semantic grammars. Semantic grammars provide a way of precisely defining a language of propositional representation, and provide rules to guide the generation of propositions and semantic parsing of linguistic expressions. Both canonical frames and semantic grammars appear to be plausible as cognitive mechanisms for the construction of propositional representations (see Brachman and Levesque 1985, Frederiksen 1986, Rumelhart and Norman 1988 for discussions of both perspectives).

4. Propositional Analysis

Theories and models of propositional representation have provided a principled basis for the development of precise methods for analyzing the semantic content of texts as representations of knowledge, developing models of expert domain knowledge, and analyzing the cognitive representations and processes underlying cognitive processes in discourse comprehension and production, inference, reasoning, and other cognitive processes that are reflected in verbal protocols. Propositional analysis is essentially a method for analyzing the propositional content of texts, verbal protocols such as text recalls, responses to questions, reasoning, or talk-aloud protocols, and even task-oriented dialog in collaborative situations of learning or performance (see Frederiksen 1986 and Kintsch 1998).

Propositional analysis normally begins with transcription and segmentation of textual material. Linguistic analysis of the text segments may be undertaken in terms of the linguistic structures previously cited. Semantic parsing rules or propositional frames are applied to specify the propositions explicitly encoded in the sentences. The generation of propositions can be guided by computer tools that enable the application of semantic parsing rules or apply frames to analyze the propositional content of text segments. Inferred propositions may be added to the model to fill recognized gaps in the text information or argument structure, relate the text to aspects of the context or task, summarize the macrostructure of the text, provide responses to questions, etc. Inference analysis can

be based on data from subjects who verbally recall, discuss, or analyze a text. Models of situation and knowledge models may be constructed applying conceptual network formalisms to the propositions. Finally, texts can be analyzed in terms of high-level conceptual frame structures such as underlie procedural texts, problem-solving or reasoning protocols, argument structures, narratives, and story plot structures.

See also: Cognitive Psychology: History; Cognitive Psychology: Overview; Critical Thinking, Cognitive Psychology of; Imagery versus Propositional Reasoning; Informal Reasoning, Psychology of; Logic and Cognition, Psychology of; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Brachman R J, Levesque H J 1985 *Readings in Knowledge Representation*. Morgan Kaufmann, Los Altos, CA

Denhière G, Rossi J-P 1991 *Text and Text Processing*. North-Holland, Amsterdam

Fauconnier G 1997 *Mappings in Thought and Language*. Cambridge University Press, Cambridge, UK

Frederiksen C H 1975 Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology* **7**: 371–458

Frederiksen C H 1986 Cognitive models and discourse analysis. In: Cooper C R, Greenbaum S G (eds.) *Studying Writing: Linguistic Approaches*. Sage, Beverly Hills, CA, pp. 227–67

Frederiksen C H, Breuleux A 1990 Monitoring cognitive processing in semantically complex domains. In: Frederiksen N, Glaser R, Lesgold A, Shafto M G (eds.) *Diagnostic Monitoring of Skill and Knowledge Acquisition*. Erlbaum, Hillsdale, NJ, pp. 351–91

Graesser A C, Bower G H (eds.) 1990 *The Psychology of Learning and Motivation: Inferences and Text Comprehension*. Academic Press, New York, Vol. 25

Johnson-Laird P N 1996 The process of deduction. In: Steier D S, Mitchell T M (eds.) *Mind Matters: A Tribute to Allen Newell*. Erlbaum, Mahwah, NJ, pp. 363–99

Kintsch W 1974 *The Representation of Meaning in Memory*. Erlbaum, Mahwah, NJ

Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK

Koschmann T (ed.) 1999 Edge of many circles: making meaning of. *Discourse Processes* **27**(2): 103–17

Rumelhart D E, Norman D A 1975 *Explorations in Cognition*. W H Freeman, San Francisco

Rumelhart D E, Norman D A 1988 Representation in memory. In: Atkinson R C, Herrnstein R J, Lindzey G, Luce R D (eds.) *Stevens' Handbook of Experimental Psychology*, 2nd edn. Wiley, New York, pp. 511–87

Simon H A 1996 The patterned matter that is mind. In: Steier D S, Mitchell T M (eds.) *Mind Matters: A Tribute to Allen Newell*. Erlbaum, Mahwah, NJ, pp. 407–31

Sowa J F 1984 *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, MA

Van Dijk T A (ed.) 1997 *Discourse as Structure and Process*. Sage, London

Van Dijk T A, Kintsch W 1983 *Strategies of Discourse Comprehension*. Academic Press, New York

C. H. Frederiksen

Prosecution

In its broadest sense, 'prosecution' signifies an aggrieved party's initiation and pursuit of a legal case through the court system. It begins with the filing of a complaint document and continues through trial, awarding of remedies, and appeal. While this term technically applies to both civil and criminal matters, it is most commonly used in the criminal law context. In a criminal case, the term 'prosecution' may refer to the process of handling a dispute in criminal court or to the aggrieved party who files the complaint in criminal court.

This article addresses both definitions of criminal prosecution using comparative data to illuminate differences and similarities among prosecution systems throughout the world. Prosecutors have been called the 'gate-keepers of criminal justice' because without their actions, judicial intervention and sanctions could not occur (di Federico 1998). However, a study of prosecution should not be limited to assessments of the prosecutor's role in individual cases. One must instead consider the ways in which countries deal with competing needs and limited resources when they establish and operate their prosecution systems. Because prosecution contributes substantially to the definition and implementation of criminal policy, any comprehensive examination of prosecutors' functions must include an analysis of their relationship to the state.

This article considers six of the most salient features of prosecution systems: (a) the degree to which prosecution is handled by public officials or private parties; (b) the degree to which the prosecutor's office is centralized or decentralized; (c) the political accountability of the prosecutor; (d) the prosecutor's relationship to and with the police; (e) the prosecutor's filing standards; and (f) the prosecutor's role in court. The content of this essay will be limited to the role of prosecution in democratic countries; for a discussion of criminal courts in nondemocratic nations, see *Trials: Cultural*; *Nazi Law*; *Socialist Law*; *Political Trials*; *War Crimes Tribunals*.

1. Public or Private Prosecution

In many legal systems, criminal law was once a branch of what scholars now call 'torts,' (see *Torts in Law*) and prosecution was initiated and pursued by the injured parties. With the rise of the modern state, crime became a distinct offense against the state itself, and responsibility for prosecution was assumed by the

state (Langbein 1974). This trend emerged earlier and more sharply in Europe than in England, where crimes were defined as offenses against the state, the system of privately initiated and financed prosecution remained in place well into the nineteenth century.

1.1 Prosecutors as Public Officials

On the European continent, public prosecutors play a much larger role in the criminal process than they do in the United States and England. Continental prosecutors control the investigation of crime, often having direct supervisory powers over the police, and typically oversee the execution and disposition of almost all criminal arrests. This tradition derives largely from the institution of the *ministère publique* from the Napoleonic Code. Furthermore, some civil law countries combine the role of prosecutor and judge, allowing officials to move back and forth between the two positions at different points in their careers and sometimes within the same case. A number of non-western countries that have based their criminal justice systems on German or French models (such as Japan) also permit role flexibility to varying degrees (see also *Trials: Cultural*; *Judges* for the relationship between prosecutors and judges). Recent developments in France and Italy, however, suggest the use of more clearly defined boundaries between criminal justice system actors.

Common law countries reject this expansive conception of the prosecutor's office and formally prohibit any degree of fluidity in roles, regarding such interchangeability as in conflict with due process and separation of powers principles (see also *Due Process in Law*; *Civil Rights*). In common law countries such as the United States, the public prosecutor unilaterally (with very few exceptions) makes filing and case-processing decisions. The investigative work is typically left to the police and judging is the domain of elected or appointed members of the judiciary. (Of course, in actual practice, the discretion with which each of these actors is vested allows some flexibility in task performance; see also *Plea Bargaining* and *Policing* for examples.) In England and Wales, the newly developed Crown Prosecution Service, comprised exclusively of professional prosecutors under the leadership of the Director of Public Prosecutions and the Attorney General, handles the prosecution of most crimes. Like their American counterparts, British judges and police officers function independently of prosecutors, although they do coordinate their activities.

1.2 Elements of Private Prosecution

Since the mid-nineteenth century, public prosecutors everywhere have become more powerful and independent. However, since the 1970s, with the growth of the victims' rights movement in many countries, legislatures have begun to provide victims with a greater voice in criminal proceedings. The degree of victim empowerment varies tremendously among countries. For example, Belgian law recognizes that certain crimes (i.e., defamation) require a formal complaint by the victim in order for the public prosecutor to initiate a criminal case. Even in that event, however, the public prosecutor is not obligated to file charges. In Germany, a victim may act as a private prosecutor if the public prosecutor does not file charges, or he may join the public prosecution as a joint plaintiff and support the prosecutor's case. A French crime victim who feels ignored by the authorities may file a civil case based on the criminal behavior. This action automatically initiates a criminal case, thereby forcing an otherwise reluctant public prosecutor to become involved.

American crime victims generally have no formal right to participate directly in criminal cases; their role is limited to testifying as witnesses and to receiving restitution as part of the final case disposition. Nonetheless, victim enthusiasm or willingness to participate (or lack thereof) may greatly influence the prosecutor's decision to proceed with a case; cooperative victims are not a prerequisite to a successful prosecution (given the expansive laws of evidence), but they often bolster an otherwise shaky case. Victims of crime in the United States can also file a civil lawsuit against the defendant based on the same conduct at issue in the criminal case, although the remedies available remain limited to civil damages. Interestingly, the role of private interests in prosecution has made a comeback in the United States, as scholars have recently debated the legitimacy of allowing private victims to finance investigations or prosecutions which are conducted by public officials (Kennedy 1997).

2. Centralized or Decentralized Office

Centralization or decentralization of oversight affects all aspects of the prosecutorial function, from the hiring and training of staff, to the implementation of policies for filing and processing cases, to levels of responsiveness to public opinion. Scholars can distinguish between those highly centralized systems that place prosecutors in an national civil service hierarchy led by the Ministry of Justice, and those decentralized systems that consider prosecutors to be locally selected (and often elected) political officials. Overall these structural and institutional features are likely to be closely related to the larger structure of government and to make a difference (Damaska 1981). Needless to say, variations exist between these extremes, and even within each category, according to hiring practices or filing control.

At one extreme is the American experience. The vast majority of criminal cases—both serious and minor—in the United States are filed and processed at the state level. However, 'State Control' is something of a misnomer, because most American state prosecutors are county officials who enforce state law. The chief prosecutor of each office (usually called the district attorney) is elected by the people, while the deputies or assistants may be either career civil servants or employees of a specific administration. State attorneys general have little control over their behavior or filing practices, handling mostly training and appeals (Flemming et al. 1992). Each county operates its own hiring system and typically recruits graduates from local law schools. There is little advance training for the job; on-the-job training occurs locally, within the office or in short courses run by the state. As a result, each office sets its priorities according to the demands of its electorate and according to the interests of its leaders; this can produce great diversity in policies among prosecutors' offices, even within the same state (Utz 1979, Feeley 1979).

At the other end of the spectrum lie countries like France and Japan, where the procuracy is tightly controlled and pyramidal in shape. For example, at the apex of the Japanese system sits the Supreme Public Prosecutors Office, followed by eight High Public Prosecutors Offices and then fifty District Public Prosecutors Offices. At the base are 452 Local Public Prosecutors Offices, located in major cities throughout Japan. The Local Offices prosecute only relatively minor cases, while the District Offices handle all serious crimes. Moreover, Japanese prosecutors are an elite subset of the population; they have graduated from top law schools, passed an extremely rigorous bar exam, and endured two years of training at the Supreme Court's Legal Training Institute. Because they are overseen by the Supreme Court, Japanese prosecutors at all levels have close association with judges; indeed, prosecutors can transfer to judicial work if they like and vice versa. As a result of their training and the hierarchy within which they work, and perhaps their close association with the judiciary, Japanese prosecutors value consistency in case-handling to a degree which other nations' prosecutors find almost inconceivable (Johnson 1996, 1998).

Centralization also affects the extent to which prosecutors can be influenced by public opinion. Officials in decentralized offices are often vulnerable to intense local political pressures and media criticism. In contrast, where public prosecutors' offices form part of a civil service hierarchy headed by the Ministry of Justice, prosecutors may consider themselves more immune from public opinion. They contend that this immunity is crucial in order for the prosecutor to remain objective and impartial; it allows him to more easily dismiss cases for insufficient evidence, or to prosecute popular or famous citizens, without fear of retaliation on election day. These claims to immunity should not be taken at face value, however; some modern prosecutors in Italy, Spain, France, and Portugal will actively court the media to generate public interest and support for their cases. Prosecutors in civil service countries are also much less likely than American prosecutors to leave government work for private practice; hence, they are less concerned about keeping good relations with the local defense bar. However, the existence of tight governmental controls provides a different sort of pressure: at least one scholar has noted that Japanese prosecutors have been severely criticized for failing to prosecute corrupt politicians (Johnson 1996).

3. *Accountability to the Public*

As the previous comments regarding public opinion and corrupt politicians reveal, prosecutors in most countries must answer to someone for their unpopular decisions. As discussed above, some countries have responded to this issue by providing victims with the right to initiate private prosecutions after the public prosecutor declines to file charges. While there is not much data on the degree to which these provisions are used by private citizens, the existence of these policies serves (at least in theory) as a check on the prosecutor's monopoly over charging decisions.

Where the prosecutor is an elected official, as in the United States, accountability is likely to come from another source: re-election pressures. A prosecutor who wishes to be re-elected must present herself as an ethical, fair, and supremely competent attorney and administrator who is responsive to the demands of the local community (Flemming et al. 1992). She is thus not likely to file charges for technical violations about which the local community cares little (such as prostitution in some places), and she may press for rigorous prosecution and sentencing for those crimes which inspire publicity, political lobbying, or protest (such as drunk driving or domestic violence in the latter part of the twentieth century).

Where the prosecutor is a career civil servant, accountability is usually to his superiors or to the Ministry of Justice. Decisions not to file charges will be reviewed by prosecutorial superiors, who tend to be conservative in matters of discretion not to prosecute (Damaska 1981). Japanese prosecutors, for example, must seek the approval of two or three superiors before making any charge decision (Johnson 1996, 1998). Yet some civil servant prosecutors have found another source of leverage; as noted above, prosecutors in Latin European countries often attempt to use media coverage to pressure office superiors into filing or pursuing controversial criminal cases. Hierarchical control is not unproblematic, though, particularly when it affects a prosecutor's career advance-

ment. France, for example, enacted reforms in 1993 and 1994 to limit the discretionary powers of the Minister of Justice, with regard to both his hierarchical supervision of prosecutorial activities and his decisions concerning the status, promotion, and discipline of public prosecutors (di Federico 1998).

Assessments of accountability must acknowledge context: prosecutors are, *inter alia*, employees of large organizations. As such, they must advance the interests of the organization in order to advance their own careers. Hence, prosecution policy and behavior are likely to follow bureaucratic dictates or public demands, despite (or within) the constraints provided by formal law.

4. *Relationship to and with Police*

Is a prosecutor a justice official or a member of law enforcement? While Langbein (1974) reports that English justices of the peace served as both investigators and prosecutors in the fourteenth and fifteenth centuries, most countries today have formally separated the roles of police and prosecutor. Nonetheless, prosecutors vary in the degree to which they control or participate in the investigative phase of a case. A prosecutor may bring a high level of control to police activities by formally supervising a police investigation (before or after bringing charges), by interviewing witnesses, or by attending the execution of search warrants. In contrast, a prosecutor may manifest low levels of control by simply filing charges in response to documents and evidence collected and submitted by investigating police officers.

In code-based countries with strongly centralized prosecution agencies, high control is the norm. Police may be statutorily required to notify prosecutors of an ongoing investigation or the detention of suspects, and the prosecutor can directly interrogate suspects or collect evidence. While prosecutors often delegate these decisions to police officers, the existence of prosecutorial investigative authority may serve as a useful check on police power (Johnson 1996, Frase 1990, 1999).

Low levels of control characterize common law countries like the United States; police typically handle the investigation and arrest phases of a criminal matter and notify the prosecutor only after arrest. While federal prosecutors often become far more involved in the investigatory phases, coordinating wire-taps, search warrants, and witness interviews, most state prosecutors do not interrogate suspects or assist in the collection of evidence (Leo 1996). Hence, police often control the flow of information on which prosecutors must rely in order to decide whether and what crimes to charge. This has led some commentators to conclude that prosecutors do little more than ratify police decisions (Feeley and Lazerson 1983), rendering the arrest itself far more significant than the prosecutor's filing decision. The same tendency has been observed in Canada (Grosman 1969).

The division of labor between police and prosecutors is not always clear-cut or unproblematic, as police behavior can influence both the onset and eventual disposition of a criminal case (Skolnick 1975). These tensions are manifest in the British experience. The police were Britain's public prosecutors from 1829 to 1986, due to the absence of a suitable alternative; in this time period, they developed a vested interest in controlling all stages of criminal cases—from investigation to arrest to prosecution (Fionda 1995). The Crown Prosecution Service was created to separate police and prosecutorial functions in order to remedy a perceived lack of police objectivity in prosecutorial duties, such as the filing and disposing of cases. However, in the fifteen years since the creation of the CPS, police officers have continued to fight for control over filing decisions and have been accused of manipulating files and witnesses to influence prosecutors' options (Nduka-Eze 1995). (See *Policing* for more about police discretion.)

5. *Filing Standards and Guidelines*

With respect to filing policies, scholars are principally concerned with whether a nation imposes a program of mandatory or compulsory prosecution on the one hand, or officially allows its prosecutors discretion in filing charges on the other. Nonetheless, whether a country actually requires its prosecutors to file charges is often hard to ascertain; the official codes may dictate one practice while informal policies allow others. For example, the American criminal prosecution system gives vast discretion to prosecutors; many continental systems, in contrast, maintain a formal policy of 'full prosecution' and explicitly make no provision for individual discretion. However, most American systems have placed some restrictions on their prosecutors' discretionary powers, while most continental systems permit discretion in hidden (or not-so-hidden) forms. This paradox has led observers to conclude that 'the rule of compulsory criminal prosecution … may foster a wider use of discretion without responsibility if no control is at work' (Guarneri and Pederzolo 2001).

Mandatory prosecution, which requires a prosecutor to bring a charge whenever there are sufficient grounds to suspect a person of having committed an offense, is a more common feature of civil law countries, whose provisions and requirements are strictly controlled by the criminal codes. Civil law countries typically choose to address undesired consequences by amending their codes, rather than openly permitting the prosecution to use discretion in enforcement. For example, Germany, Sweden, and other countries have handled their increased crime problem by legislatively downgrading or decriminalizing certain offenses, and by making

others subject to administrative (rather than criminal) sanctions (Sessar 1979, Feeley 1979). However, Germany also recognizes the 'principle of opportunity' for petty offenses, which gives the prosecutor discretion to drop the charge, to issue a penal order (which allows a person to pay a fine in a minor case), or to bring the case for summary trial in order to expedite proceedings (Herrmann 1992, Felstiner 1979).

In France, the opportunity principle works somewhat differently. While the public prosecutor is not obligated to bring a prosecution, once she initiates a prosecution she no longer has a choice regarding whether to pursue the case (Frase 1999). In Italy, compulsory prosecution exists without the concurrent opportunity principle, but discretion reigns nonetheless (Gaurneri 1996, Corso 1993). When a prosecutor does not wish to file a case, she may neglect to investigate and ask the judge to declare the evidence insufficient, or she may allow the statute of limitation to lapse.

The Italian prosecutor's manipulation of formal criteria to justify a non-filing decision raises the issue of discretion: in countries which officially allow some discretion, are prosecutors subject to any formal rules regarding its exercise? For example, Israeli and British prosecutors must prosecute when the evidence is sufficient, unless they find that prosecution would not be in the public interest (Harnon and Stein 1999, Fionda 1995). Notably, British prosecutors may be using this public interest criterion as an informal presumption against prosecution in relatively minor cases, especially where the defendant has no previous criminal record (Fionda 1995).

In the United States, where there is no official 'public interest' component to the filing decision, prosecutors have an enormous amount of discretion; they need to justify filing decisions only to their superiors, as victims have no legal rights to intervene. As a result, prosecutors allow informal social criteria, including victim characteristics, to influence their charging decisions. Cases involving victims who are related to defendants, or who are otherwise 'undesirable' people (such as prostitutes or gang members) trigger larger societal issues, and thus appear less distinctively 'criminal' to American juries. Because these cases pose conviction problems, prosecutors concerned about their personal 'win rates' are reluctant to file them (Frohmann 1997, Vera Institute of Justice 1981). American prosecutors have also been accused of overcharging cases, filing multiple charges in the complaint with an eye towards plea bargaining later. Overcharging gives the prosecutor leverage in the plea bargaining process, but it has been called a coercive, corrupting, and deceptive technique to get the defendant to agree to virtually the same sentence he would receive after trial (Alschuler 1968) (see *Plea Bargaining*).

Discussion of filing policy necessarily leads to consideration of alternatives to prosecution, because many cases which are not technically prosecuted are also not fully dismissed. Rather, they are redirected into other channels, alternatively called diversion, penal orders, or summary procedures. These programs arose in response to demands to keep minor offenders and unimportant cases out of the already overworked court systems. However, they have all been accused of 'widening the net' of criminal social control, as prosecutors are more inclined to divert a minor offender into this type of alternative program than to drop his case altogether.

6. *Roles in Court*

The final point of comparison concerns the prosecutor's role during the trial of a criminal case. Her role will largely be determined by the type of trial system in which she works (adversarial, inquisitorial, or some mixture thereof; see *Trials: Cultural*). She may be an active adversarial participant striving to win a conviction, or a more passive, impartial actor charged with ensuring that justice results.

Where the trial depends largely on the contents of a written dossier (in countries such as Japan, Argentina, France, and Germany), the prosecutor is generally more passive; she submits the dossier containing the investigative reports to the judge (or to the panel of judges) before trial, but she generally relinquishes control over questioning of witnesses. In many of these systems, the prosecutor is obligated to ask the judge for an acquittal following trial if the evidence merits this result, and the prosecutor typically has the right to appeal both the outcome of the case and the sentence.

In systems that require oral testimony (including Italy, Great Britain, Israel, and the United States), the prosecutor is the principal player in the courtroom. He may sit near the factfinders or judges to establish rapport; he controls the majority of the witnesses and the presentation of evidence; he argues his case both first and last, providing both the initial opening statement and the final closing argument. For this reason, prosecutors in oral-trial systems feel personally invested in the outcome of a case and often place tremendous pressure on themselves to win convictions (Heilbroner 1990, Baker 1999). They generally have no right to appeal acquittals.

A final consideration is the prosecutors' input into the sentencing decision. Prosecutors are often said to have tremendous influence over sentencing based on their charging and plea bargaining policies. Because each crime is associated with a specific minimum and maximum sentence, the prosecutor can control the type of sentence a defendant will receive by filing or making him plead guilty to certain charges. This power is exacerbated by mandatory sentencing guidelines; by eradicating the judge's discretion to sentence appropriately after conviction, these guidelines give

the prosecutor the ability to control the sentence through his charge-bargaining practices. Furthermore, some nations, such as Great Britain and Germany, limit the ability of certain courts to impose certain punishments; hence, the mode of trial predetermines the maximum sentencing powers of the court in a given case. The prosecutor may thus indirectly control sentencing by filing the case in a particular type of court.

Prosecutors might also influence sentencing directly, although nations differ concerning the degree to which they allow prosecutors to make sentencing recommendations after trial. In many European countries, prosecutors can assist judges in navigating complex sentencing formulas, while in the United States the sentencing hearing is regarded as another phase of the adversarial process in which the prosecutor may present both witnesses and arguments. In Great Britain, the prosecutor has no legal authority to recommend an appropriate sentence; indeed, some commentators believe such a power would be an unconstitutional infringement on judges' power (Fionda 1995). However, British prosecutors have a duty to correct a judge who is in danger of imposing an illegal sentence, and the British Attorney General, like prosecutors in South Africa and the United States, may appeal any sentence imposed by the trial court on the grounds of undue leniency or statutory error.

7. *Conclusion*

Given the vast array of criminal prosecution systems existing at the close of the twentieth century, it no longer seems useful to rely on the broad categories of adversarial and inquisitorial systems to illustrate similarities and differences. A number of axes exist that better describe the elements of various prosecution systems; to ignore these dimensions in favor of simple dichotomies provides us with too superficial a view of the prosecutor's role in court, in public life, and in political affairs. Moreover, the prosecutor is not an isolated actor: her responsibilities, relationships, and roles depend on social, legal, and political contexts. In order to truly understand the role of prosecution from a social scientific perspective, we must commit ourselves to understanding these environments and to accepting the cultural differences which exist among the world's nations.

See also: Common Law; Courts and Adjudication; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Death Penalty; Integration: Social; Law: Economics of its Public Enforcement; Plea Bargaining; Police, Sociology of; Policing; Punishment: Social and Legal Aspects; Trials: Cultural

Bibliography

Alschuler A W 1968 The prosecutor's role in plea bargaining. *University of Chicago Law Review* **36**: 50–112

Baker M 1999 *D.A.: Prosecutors in Their Own Words*. Simon and Schuster, New York

Corso P 1993 Italy. In: Van Den Wyngaert C (ed.) *Criminal Procedure Systems in the European Community*. Butterworths, London, pp. 223–60

Damaska M 1981 The reality of prosecutorial discretion: Comments on a German monograph. *American Journal of Comparative Law* **29**: 119

di Federico G 1998 Prosecutorial independence and the democratic requirement of accountability in Italy. *British Journal of Criminology* **38**: 371–87

Feeley M M 1979 *The Process is the Punishment*. Russell Sage, New York

Feeley M M, Lazerson M H 1983 Police–prosecutor relationships: An interorganizational perspective. In: Boyum K, Mather L (eds.) *Empirical Theories about Courts*. Longman Press, New York, pp. 216–43

Felstiner W L F 1979 Plea contracts in West Germany. *Law and Society Review* **13**: 309–26

Fionda J 1995 *Public Prosecutors and Discretion: A Comparative Study*. Clarendon Press, Oxford

Flemming R B, Nardulli P F, Eisenstein J 1992 *The Craft of Justice: Politics and Work in Court Communities*. University of Pennsylvania Press, Philadelphia, PA

Frase R S 1990 Comparative criminal justice as a guide to American law reform: How do the French do it, how can we find out, and why should we care? *California Law Review* **8**: 539–686

Frase R S 1999 France. In: Bradley C M (ed.) *Criminal Procedure: A Worldwide Study*. Carolina Academic, Durham, NC, pp. 143–84

Frohmann L 1997 Convictability and discordant locales: Reproducing race, class, and gender ideologies in prosecutorial decisionmaking. *Law and Society Review* **31**: 531

Goldstein A, Marcus M 1977 The myth of judicial supervision in three 'inquisitorial' systems: France, Italy, and Germany. *Yale Law Journal* **87**: 240–83

Goldstein A, Marcus M 1977 Comment on continental criminal procedure. *Yale Law Journal* **87**: 1570–7

Grosman B 1969 *The Prosecutor: An Inquiry into the Exercise of Discretion*. University of Toronto Press, Toronto, Canada

Guarneri C 1996 Prosecution in two civil law countries: France and Italy. In: Nelken D (ed.) *Comparing Legal Cultures*. Aldershot, Brookfield, VT

Guarneri C, Pederzolo P 2001 *The Power of the Judge*. Oxford University Press, Oxford, UK

Harnon E, Stein A 1999 Israel. In: Bradley C M (ed.) *Criminal Procedure: A Worldwide Study*. Carolina Academic, Durham, NC, pp. 217–44

Heilbroner D 1990 *Rough Justice: Days and Nights of a Young D.A.* Pantheon Books, New York

Herrmann J 1992 Bargaining Justice—A bargain for German criminal justice. *University of Pittsburgh Law Review* **53**: 755–76

Johnson D 1996 The Japanese way of justice. Doctoral Dissertation. University of California at Berkeley, Berkeley, CA

Johnson D 1998 The organization of prosecution and the possibility of order. *Law and Society Review* **32**: 247–308

Kennedy J 1997 Private financing of criminal prosecutions and

the differing protections of liberty and equality in the criminal justice system. *Hastings Constitutional Law Quarterly* **24**: 665
Langbein J H 1974 *Prosecuting Crime in the Renaissance: England, Germany, France*. Harvard University Press, Cambridge, MA
Langbein J H, Weinreb L L 1977 Continental criminal procedure: 'myth' and reality. *Yale Law Journal* **87**: 1549–69
Leo R 1996 Inside the interrogation room. *Journal of Criminal Law and Criminology* **86**(2): 266–303
Nduka-Eze C 1995 The CPS and independence from the police. *New Law Journal* 1843–4
Sessar K 1979 Prosecutorial discretion in Germany. In: Macdonald W F (ed.) *The Prosecutor*. Sage, Beverly Hills, CA, pp. 255–76
Sigler J 1979 The prosecutor: a comparative functional analysis. In: Macdonald W F (ed.) *The Prosecutor*. Sage, Beverly Hills, CA, pp. 53–74
Skolnick J 1975 *Justice Without Trial: Law Enforcement in Democratic Society* (2nd edn.), Wiley, New York
Utz P 1979 Two models of prosecutorial professionalism. In: Macdonald W F (ed.) *The Prosecutor*. Sage, Beverly Hills, CA, pp. 99–124
Vera Institute of Justice 1981 *Felony Arrests: Their Prosecution and Disposition and New York City's Courts*. Longman, New York
Weigand T 1999 Germany. In: Bradley C M (ed.) *Criminal Procedure: A Worldwide Study*. Carolina Academic Press, Durham, NC, pp. 187–211

K. Levine and M. Feeley

Prosocial Behavior and Empathy: Developmental Processes

Research on helping others in distress took off with the brutal murder of a young woman in Queens, New York in the mid-1960s, which was witnessed by 30 people from their apartment windows. The initial research on why witnesses do not help (Latane and Darley 1970) focused on the context and found that the presence of other bystanders can interfere with a person helping by activating certain assumptions—'diffusion of responsibility' (I'm sure someone has called the police) or 'pluralistic ignorance' (no one is doing anything, so it can't be an emergency). There was also research on when people do help, which showed among other things that the tendency of bystanders to help increases with age.

Recent research has focused more on prosocial motivation and emotion. It highlights empathic emotion, sympathy, and guilt. This article is concerned with developmental processes in empathic emotions and their contribution to helping. Also covered is the role of cognition in arousing, developing, and shaping empathic emotion and enlarging the individual's perspective on prosocial behavior. Finally, the article deals with the contribution of socialization especially by parents and peers, and gender differences.

1. *Definition of Empathy*

Empathy has been defined in two ways: (a) the cognitive awareness of another person's internal states (thoughts, feelings, perceptions, intentions); (b) the vicarious affective response to another person. This chapter deals with affective empathy, because of its motive properties (see Ickes 1997 for cognitive empathy).

Affective empathy was initially defined as a match between the observer's and the model's feeling. It is now defined as it is here, more in terms of the processes that make an observer's feelings congruent with the model's situation not his own (Hoffman 1978, 2000). There may be a match, but not necessarily as when one feels empathic anger on seeing someone attacked even when the victim feels sad or disappointed rather than angry. Most of the research on affective empathy as a prosocial motive focuses on empathic distress because prosocial moral action usually involves helping someone in discomfort, pain, danger, poverty, and the like.

2. *Modes of Empathic Arousal*

Five empathy-arousal modes have been identified (Hoffman 1978). Three are primitive, automatic, and preverbal. (a) Mimicry, which has two steps: one spontaneously imitates another's facial, vocal, or postural expressions of feeling, producing changes in one's own facial or postural musculature; these changes trigger afferent feedback to the brain, which results in feelings that resemble the other's feelings. (b) Classical conditioning, in which empathic distress becomes a conditioned response to distress in others through observing others in distress at the same time that one is in actual distress. (c) Direct association of cues from others or their situation with one's own painful past experience. The empathy aroused by these three modes is passive, involuntary, based on the pull of surface cues, and requires little cognitive processing. Nevertheless, these modes are important because (a) they show that a person's distress can indeed result from someone else's painful experience; (b) they make empathy arousal possible in the early preverbal years of childhood; and (c) they give empathy an involuntary dimension throughout life.

Two empathy-arousing modes are higher order cognitive. (a) Verbally mediated association: language communicates the other's distress and connects the other's situation to one's own painful past experience. (b) Perspective-taking: one feels something of the victim's distress by imagining how one would feel in the victim's place and/or trying to imagine directly how the victim feels by using information one has about him or her. These higher order cognitive modes may be drawn out over time and subject to voluntary control, but they may also be triggered immediately on witnessing another's distress. What they contribute to

empathy is scope; they also allow one to empathize with others who are not present.

Multiple modes are important because they enable observers to respond empathically to whatever distress cues are available (Hoffman 2000). Cues from the victim's face, voice, or posture can be picked up through mimicry, situational cues, through conditioning or association. If a victim expresses distress verbally or in writing, or someone else describes the victim's situation, one can empathize through verbally mediated association or perspective-taking. Empathic distress is thus a multidetermined hence reliable human response, which fits the argument that it survived natural selection and has a hereditary component (Hoffman 1981, Zahn-Waxler et al. 1992).

3. *Developmental Stages*

Mature empathy has a metacognitive dimension: one knows one's feeling of distress results from another's plight and how the other presumably feels. One thus has a sense of oneself and others as separate beings with independent inner states (that are only partly reflected in outward behavior), separate identities, and separate life conditions. Before 6 or 7 years children can empathize but without this metacognitive dimension. This suggests that empathic affect develops along with cognitive self/other development. Five stages of empathy development have been hypothesized (Hoffman 1978, 2000).

(a) Global empathic distress: newborns show a vigorous, agitated distress response to another infant's cry. (b) Egocentric empathic distress: one responds to another's distress as though oneself were in distress. This occurs when children can be empathically aroused by the three preverbal modes but still lack a clear self-/other distinction. (c) Quasiegocentric empathic distress: one knows the other is in actual distress but confuses the other's inner states with one's own and tries to help by doing for the other what would comfort oneself. (d) Veridical empathic distress: one's feeling is closer to the other's because one realizes the other has inner states independent of one's own. (e) Empathy for another's experience beyond the immediate situation (e.g., chronic illness, economic hardship, deprivation): this high empathy level is possible when children realize that others have an identity and a generally sad or happy life condition; and empathy with an entire group (homeless; Oklahoma City bombing victims).

4. *Causal Attribution's Shaping of Empathic Affect*

Humans spontaneously attribute causality to events (Weiner 1985) and they presumably do this when witnessing someone in distress. If they blame the victim, empathic distress is of course reduced. Otherwise, depending on the attribution, empathic affect may be shaped into four empathy-based moral affects (Hoffman 2000). (a) Sympathetic distress, when the cause is beyond the victim's control (illness, accident, loss). (b) Empathic anger, when someone else is the cause. (c) Empathic feeling of injustice, when a discrepancy exists between the victim's character and the victim's fate (a good person fares badly). (d) Empathy-based guilt over inaction, when one does not help or one's efforts to help fail and the victim's distress continues; here one's empathic distress combines with blaming oneself not for causing the victim's distress but allowing it to continue.

5. *Empathy as a Prosocial Motive*

Empathy correlates positively with helping. There is also experimental evidence that arousal of empathic distress leads to helping; and that observers help more quickly the greater the victims' pain and the higher the observers' empathic distress. Furthermore, observers' empathic distress becomes less intense and they feel better if they help, but not if they don't help or if despite their best efforts and through no fault of their own, the victim's distress is not alleviated, which implies that the main objective of empathy-based helping is to alleviate the victim's distress. There is also evidence that empathy reduces both aggressive and manipulative (Machiavellian) behavior. It thus seems clear that empathic distress is a prosocial motive. See reviews by Hoffman (1978, 2000) and Eisenberg and Miller (1987).

Empathy has limitations, however (Hoffman 1984, 2000). (a) Egoistic motives may intervene and one may not help, even at the price of feeling guilty afterwards. (b) Empathic distress increases with the level of victims' distress, but empathic distress can become aversive enough to shift one's attention from the victim's to one's own personal distress (empathic overarousal). (c) People empathize with almost anyone in distress (the victims in research are usually strangers), but they empathize more with kin, in-group members, friends, and people who share their personal concerns (familiarity bias)—which may not be a problem except when intergroup rivalry contributes to intense empathic anger (hatred, ethnic cleansing) toward outgroup members. (d) People empathize more with victims who are physically present (here-and-now bias), probably because most empathy-arousing processes require immediate situational and personal cues.

6. *Socialization by Parents and Peers*

The role of parent as socializer has three facets: model, disciplinarian, nurturer. Each contributes to children's empathy and prosocial behavior, although discipline has been the most studied (Hoffman 2000). The

discipline method most likely to contribute to empathy, helping, and guilt over harming others is a type of reasoning called induction, often used by educated middle-class American parents when children harm or are about to harm someone. Inductions direct the child's attention to the victim's distress and may thus engage the empathy-arousing modes noted above. They also highlight the child's role in causing the victim's distress, which may lead to empathy-based transgression guilt. Power-assertive disciplines (physical force, threats, commands) are associated with low empathy, low helping, and low guilt, although some power assertion may be needed at times to get children to attend and process an induction's message.

Parents' behavior outside discipline encounters may provide prosocial models that reinforce children's empathic proclivity and legitimize helping (Eisenberg and Fabes 1998, Hoffman 2000). Examples are expressing compassion toward people in difficult straights such as the homeless; linking a television protagonist's feelings or situations with the child's own experience; pointing out similarities among all humans such as sadness due to separation and loss. Parent models may thus reinforce children's empathic dispositions especially when they are bystanders and someone needs help. Parent models may also make children more receptive to inductions, and, along with inductions, make children more receptive to claims of their peers (Hoffman 2000).

Piaget suggested in the 1930s that interaction with peers (not parents, who are too powerful) fosters prosocial moral development in children (Piaget 1932). Cognitive-developmental psychologists have recently elaborated this view (Turiel 1998), briefly as follows: children 'co-construct' and discover moral norms in the give-and-take of peer interaction (disputes, arguments, negotiations). These peer conflicts put pressure on children to take different points of view and to consider and coordinate the needs and rights of self and others (for example, regarding claims to ownership and possession of objects). This is a compelling hypothesis that merits research. Hoffman (2000) suggests that the peer interaction processes involved can serve to add the finishing touches to the prosocial outcomes of having inductive, nurturant, prosocial models as parents.

7. *The Role of Cognition*

Though we deal with empathic affect, cognition's importance is evident in two empathy arousal processes (verbally mediated association and perspective-taking): in the key role of self/other differentiation in empathy development; in the production of empathic anger and empathy-based guilt through causal attribution; and in children processing the information in inductions. Beyond that, cognitive development enables humans to form images, represent people and events, and imagine themselves in another's place; and because represented people and events can evoke affect, victims need not be present for empathy to be aroused in observers. Empathy can be aroused when observers imagine or read about victims, discuss or argue about economic or political issues that involve victims, or make moral judgments about hypothetical moral dilemmas (see *Moral Reasoning in Psychology*) in which there are victims. See Hoffman (1987, 2000) for discussion of the importance of cognition, including the cognitive dimension of moral principles, in empathy's development and prosocial functioning.

8. *Gender Differences*

Girls are socialized more than boys to be kind and feel responsible for others' well-being. Girls are thus expected, increasingly with age, to be more empathic and prosocial and less likely to harm others. The research supports these expectations but in varying degrees (Eisenberg and Fabes 1998). Girls show more kindness and consideration for others and, to a lesser extent, are more likely to help and share. They are less likely to harm others and more apt to feel guilty when they do. They are also more empathic than boys, but this is clear only with self-report, especially questionnaire measures, in which it is obvious what is being assessed and responses are under the subject's control. It is not clear with unobtrusive naturalistic observations or physiological indices of empathy. The empathy findings may thus reflect subjects' conceptions of what boys and girls are supposed to be like. Whether they reflect gender differences in actual behavior is a matter for future research to decide.

See also: Adulthood: Prosocial Behavior and Empathy; Altruism and Prosocial Behavior, Sociology of; Antisocial Behavior in Childhood and Adolescence; Moral Reasoning in Psychology; Social Cognition and Affect, Psychology of; Social Support, Psychology of

Bibliography

Eisenberg N, Fabes R A 1998 Prosocial development. In: Damon W (ed.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, vol. 3

Eisenberg N, Miller P A 1987 The relation of empathy to prosocial and related behaviors. *Psychological Bulletin* **101**: 91–119

Hoffman M L 1978 Empathy, its development and prosocial implications. In: Keasey C B (ed.) *Nebraska Symposium on Motivation*, University of Nebraska Press, Lincoln, NV. Vol. 25, pp. 169–218

Hoffman M L 1981 Is altruism part of human nature? *Journal of Personality and Social Psychology* **40**: 121–37

Hoffman M L 1984 Empathy, its limitations, and its role in a comprehensive moral theory. In: Kurtines W, Gewirtz J (eds.)

Morality, Moral Behavior, and Moral Development. John Wiley, New York, pp. 283–302
Hoffman M L 1987 The contribution of empathy to justice and moral judgment. In: Eisenberg N, Strayer J (eds.) *Empathy and its Development*. Cambridge University Press, New York, pp. 47–80
Hoffman M L 2000 *Empathy and its Development: Implications for Caring and Justice*. Cambridge University Press, New York
Ickes W 1997 *Empathic Accuracy*. Guilford Press, New York
Latane B, Darley J M 1970 *The Unresponsive Bystander: Why Doesn't He Help?* Appleton-Century Crofts, New York
Lennon R, Eisenberg N 1987 Gender and age differences in empathy and sympathy. In: Eisenberg N, Strayer J (eds.) *Empathy and its Development*. Cambridge University Press, New York, pp. 195–217
Piaget J 1932 *The Moral Judgment of the Child*. Harcourt, New York
Turiel E 1998 The development of morality. In: Damon W (ed.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*. Wiley, New York, Vol. 3
Weiner B 1985 'Spontaneous' causal thinking. *Psychological Bulletin* **97**: 74–84
Zahn-Waxler C, Emde R N, Robinson J L 1992 The development of empathy in twins. *Developmental Psychology* **28**: 1038–47

M. L. Hoffman

Prosopagnosia

Prosopagnosia is a neuropsychological condition involving loss of ability to recognize familiar faces. Prosopagnosic patients instead rely on voice, context, name, or sometimes clothing or gait to achieve recognition of people they know. The problem in face recognition is not due to blindness (the person with prosopagnosia can still see) or general intellectual impairment (people are still recognized from nonfacial cues). Yet even the most familiar faces may go unrecognized: famous people, friends, family, and the person's own face when seen in a mirror.

Face recognition impairments are socially disabling, and most patients develop strategies for tackling the problem. They may note carefully what clothes someone is wearing, ask relatives always to wear a particular distinctive item, or become adept at initiating or maintaining a conversation whilst they work out who they are talking to. These strategies are not completely effective. One prosopagnosic patient lost a legal action when he discussed his case with his opponent's lawyer—he thought it was his own advocate because he was wearing similar court clothes.

Prosopagnosia should not be confused with other types of impairment that can compromise face recognition (Young 1992). In prosopagnosia, recognition from other cues than the face is usually possible. In contrast, cases of inability to recognize people from face, voice, and name have been described—these seem to reflect loss of semantic information about the identities of individuals. There are also reports of problems of name retrieval (anomia), in which familiar people are successfully recognized and appropriate semantic information is accessed, but their names cannot be recalled. Face recognition impairments therefore reflect damage to an underlying system which can break down in different ways.

Although reports of problems affecting face recognition after brain injury can be traced back to the nineteenth century, prosopagnosia was identified as a distinct neuropsychological problem by Bodamer in 1947 (Ellis and Florence 1990). Whilst it is generally considered a rare deficit, there are now several hundred case descriptions in the literature. These include cases where brain injury early or late in life has led to inability to recognize previously familiar faces, and developmental cases where the disorder is present from birth. People with prosopagnosia know when they are looking at a face, and can describe its features, but the loss of any sense of overt recognition is often complete, with no feeling of familiarity. In contrast, other aspects of face processing, such as the ability to interpret facial expressions, or to match views of unfamiliar faces, can remain remarkably intact in some (though not all) cases. The ability to recognize everyday objects other than faces may also remain good, and many prosopagnosic patients are able to read without difficulty. In one of the cases of early acquired deficit, a child who had remained unable to recognize any faces since suffering meningitis and subsequent complications in infancy was nonetheless able to learn to read (Young and Ellis 1989); this underlines the implication that reading and face recognition depend on different types of visual analysis.

Deficits commonly associated with prosopagnosia include a visual field defect in the left upper quadrant, achromatopsia (loss of color vision), and topographical disorders (inability to find one's way about). These are useful clinical pointers, but they are thought to be due to anatomical proximity of otherwise unrelated processes rather than having any direct functional significance; cases of prosopagnosia without each of these associated deficits have been described, and they have also each been reported in the absence of prosopagnosia.

Studies of prosopagnosia initially concentrated on the location of the lesions responsible, and to what extent the problem is specific to face recognition. More recent work has asked whether there exist different underlying forms of deficit, and has explored the implications of findings of covert recognition.

1. *Lesions Causing Prosopagnosia*

The underlying pathology involves lesions affecting ventromedial regions of occipitotemporal cortex; these include the lingual, fusiform, and parahippocampal

gyri, and more anterior parts of the temporal lobes. Functional imaging studies of normal observers have confirmed the importance of these regions to face perception (Kanwisher et al. 1997). Bilateral lesions are usually present in the relatively small number of cases that have come to autopsy (Damasio et al. 1982), but cases involving unilateral lesions of the right cerebral hemisphere have been reported in several computerized tomography (CT) and magnetic resonance imaging (MRI) studies, and the functional imaging findings with normal people also emphasize the importance of the right hemisphere.

2. *A Face-specific Deficit?*

Prosopagnosia involves a severe impairment of face recognition, yet a number of prosopagnosic patients can read and can recognize many everyday objects, so the deficit does not compromise all aspects of visual recognition. But is it only face recognition that is impaired?

In most cases, the answer is clearly 'no.' So, for example, recognition of familiar buildings, flowers, types of car, or animal species may all pose difficulties. Like faces, these are all visual categories with a high degree of interitem similarity.

This has led to the hypothesis that what is lost in prosopagnosia is the ability to recognize the individual members of categories that contain several items of similar appearance. Ability to identify the general category to which the items belong (face, building, flower, etc.) is preserved, but within-category recognition is defective. Although attractive, this idea has been undermined by (rare) reports of highly circumscribed impairments. One of De Renzi's (1986) cases could find his own belongings when they were mixed in with similar objects, identify his own handwriting, pick out a Siamese cat among photographs of other cats, recognize his own car without reading the number plate, and sort domestic coins from foreign coins. In all of these tasks he was therefore able to identify individual members of visual categories with high interitem similarity, yet he could not recognize the faces of relatives and close friends. Just as strikingly, the deficit can affect mainly *human* faces: when McNeil and Warrington's (1993) prosopagnosic patient took up farming, he was able to learn to recognize his sheep, and correctly identified several of them from photographs of their faces!

Although they are exceptionally rare, such cases imply that the possibility of face-specific deficits must be taken seriously. However, it may be important to reconsider the underlying issue. The presumption behind the interest in face-specific deficits has been that these might indicate whether the brain has an evolved neural substrate for the important social task of face recognition. But this rationale would only be valid if the organization of the evolved substrate was so rigid that it could not become involved in other important visual discriminations. Studies of the properties of visual cells in the primate brain (Logothetis and Pauls 1995) and acquired expertise in humans (Diamond and Carey 1986) suggest, in contrast, a more plastic organization. It is possible that much of our visual expertise involves extending the sphere of involvement of cells that would otherwise become dedicated to face recognition.

3. *Different Forms of Prosopagnosia*

There has been much discussion as to whether prosopagnosia can have more than one different underlying cause. Often, a distinction is made between cases that seem to reflect defective face perception and those that are more like a loss of face memories. When face perception is defective, patients may be poor at telling male from female faces, or matching photographs of unfamiliar faces, whereas patients with a face memory problem may remain able to perform such tasks fairly normally.

For many years, this debate proved inconclusive because it was conceivable that the so-called perceptual cases simply reflected a more severe form of the same impairment. However, studies of ability to form mental images of premorbidly familiar faces have shown that someone suffering from the perceptual form of prosopagnosia can perform better than someone with a loss of face memories (Young et al. 1994). A plausible interpretation is that this happens because we need the same representation to conjure up a mental image of someone's face that we need to recognize that face. In perceptually based cases of prosopagnosia, these stored representations of familiar faces could remain intact, but a seen face could not access them because it is not perceived sufficiently accurately; imaging faces which cannot be recognized would then remain possible. In memory-based cases, stored memories cannot be used for imagery or for recognition—either because they are themselves damaged or because access to the store is defective.

4. *Covert Recognition*

The interpretation of some cases of prosopagnosia as a domain-specific impairment of face memory is strengthened by findings of covert recognition, which show clear parallels to priming effects found more generally in amnesia. Prosopagnosic patients usually fail all tests of overt recognition of familiar faces. They cannot name the face, give the person's occupation or other biographical details, or even state whether a face is that of a familiar person (all faces seem unfamiliar).

Surprisingly, however, there is substantial evidence of covert recognition from physiological and behavioral measures (Bruyer 1991).

Bauer (1984) measured skin conductance whilst a prosopagnosic patient, LF, viewed a familiar face and listened to a list of five names. When the name belonged to the face LF was looking at, there was a greater skin conductance change than when someone else's name was read out. Yet if LF was asked to choose which name in the list was correct for the face, his performance was at chance level. Hence, there was a marked discrepancy between LF's inability to identify the face overtly and the relatively good recognition shown using the indirect skin conductance measure.

A number of other indices of nonconscious, covert recognition in prosopagnosia have also been developed. Matching of familiar faces is better than for unfamiliar faces, priming has been found from familiar faces onto the recognition of name targets, and learning of correct face-name pairings is better than learning of incorrect pairings (Bruyer 1991). All of these effects can be demonstrated when overt recognition of the faces is at chance level.

Not all prosopagnosic patients show covert recognition, however. As might be expected, when there is evidence of substantial impairment of face perception, covert as well as overt performance can fall to chance level.

There is a parallel between preserved covert recognition abilities and those aspects of recognition that operate automatically for normal people. We cannot look at a familiar face and decide not to recognize it—the mechanisms responsible for visual recognition are not open to conscious control in this way. It seems that some of these automatic aspects of recognition continue to function in some cases of prosopagnosia. A simulation of this pattern can be made by halving some of the connection strengths in a computer model of the architecture of the recognition system (Young and Burton 1999). The network is then no longer able to classify face inputs as familiar, yet it continues to display priming effects. This helps us to understand how covert responses can be preserved when there is no overt discrimination, but it does not solve the more difficult issue of what is involved in being aware of recognizing a face.

Despite the extensive range of covert effects demonstrated in the laboratory, prosopagnosic patients do not act as if they recognize faces in everyday life. Instead, most are acutely conscious of their problems in face recognition. However, some patients lack insight into their disability and show unawareness of impairment (anosognosia). Case SP was very poor at recognizing familiar faces, yet she maintained that she recognized faces 'as well as before,' even when directly confronted with her failure to recognize photographs of familiar faces (Young et al. 1990). In contrast, SP showed adequate insight into other physical and cognitive problems produced by her illness—her lack of insight into her face recognition impairment involved a deficit-specific anosognosia. Such deficit-specific anosognosias may reflect impairment to mechanisms used for monitoring performance in everyday life; we can all make occasional errors of recognition, so we constantly monitor our own performance to correct these errors quickly and avoid embarrassment whenever possible.

Remarkably, the deficits of overt recognition noted in prosopagnosia need not be absolute. Sergent and Poncet (1990) found that their patient, PV, could achieve overt recognition of some faces if several members of the same semantic category were presented together. This only happened when PV could determine the category herself; otherwise she continued to fail to recognize the faces overtly even when the occupational category was pointed out to her. Sergent and Poncet (1990) suggested that the simultaneous presentation of several members of the same category may have temporarily raised their activation above the threshold at which recognition could be achieved.

Such findings in prosopagnosia show that the boundary between awareness and lack of awareness is not as completely impassable as it seems to the patients' everyday experience. However, the circumstances under which this has been found to happen are at present very limited, and it has not yet proved possible to find a way to provide significant remedial assistance in real life.

See also: Face Recognition Models; Face Recognition: Psychological and Neural Aspects

Bibliography

Bauer R M 1984 Autonomic recognition of names and faces in prosopagnosia: A neuropsychological application of the guilty knowledge test. *Neuropsychologia* **22**: 457–69

Bruyer R 1991 Covert face recognition in prosopagnosia: A review. *Brain and Cognition* **15**: 223–35

Damasio A R, Damasio H, Van Hoesen G W 1982 Prosopagnosia: Anatomic basis and behavioral mechanisms. *Neurology* **32**: 331–41

De Renzi E 1986 Current issues in prosopagnosia. In: Ellis H D, Jeeves M A, Newcombe F, Young A (eds.) *Aspects of Face Processing*. Martinus Nijhoff, Dordrecht, The Netherlands, pp. 243–52

Diamond R, Carey S 1986 Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General* **115**: 107–17

Ellis H D, Florence M 1990 Bodamer's (1947) paper on prosopagnosia. *Cognitive Neuropsychology* **7**: 81–105

Kanwisher N, McDermott J, Chun M M 1997 The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* **17**: 4302–11

Logothetis N K, Pauls J 1995 Psychophysical and physiological evidence for viewer-centred object representations in the primate. *Cerebral Cortex* **5**: 270–88

McNeil J E, Warrington E K 1993 Prosopagnosia: A face specific disorder. *Quarterly Journal of Experimental Psychology, Series A* **46**: 1–10
Sergent J, Poncet M 1990 From covert to overt recognition of faces in a prosopagnosic patient. *Brain* **113**: 989–1004
Young A W 1992 Face recognition impairments. *Philosophical Transactions of the Royal Society, London, Series B* **335**: 47–54
Young A W, Burton A M 1999 Simulating face recognition: Implications for modelling cognition. *Cognitive Neuropsychology* **16**: 1–48
Young A W, De Haan E H F, Newcombe F 1990 Unawareness of impaired face recognition. *Brain and Cognition* **14**: 1–18
Young A W, Ellis H D 1989 Childhood prosopagnosia. *Brain and Cognition* **9**: 16–47
Young A W, Humphreys G W, Riddoch M J, Hellawell D J, De Haan E H F 1994 Recognition impairments and face imagery. *Neuropsychologia* **32**: 693–702

A. W. Young

Prosopography (Collective Biography)

Prosopography, or collective biography, has developed considerably since the 1960s, expanding in the fields of modern and contemporary history, after the tool was invented and applied primarily in ancient and medieval history. The method's principle is simple: defining a population on the basis of one or several criteria and designing a relevant biographical questionnaire containing a range of variables or criteria which serve to describe it in terms of social, private, public and/or cultural, ideological or political dynamics, depending on the population under scrutiny and the questionnaire that is being used. In the words of Lawrence Stone: 'Prosopography is the investigation of the common background characteristics of a group of actors in history by means of a collective study of their lives (Stone 1971, p. 46) Once the documentation has been compiled, which is the most time-consuming aspect of the work, the data can be evaluated with the help of a number of techniques, both quantitative and qualitative, including manual counting or computerized processing, statistical tables or factor analyses, depending on the scope and sophistication of the questionnaire and the sample.

To appreciate the success of this method and demonstrate its extension to different branches of history, several historiographic references are cited in Sect. 1. To gain insight into its merits and limitations, it is necessary to analyze a few examples of its application, taken primarily from recent periods see Sect. 2 where other social history methods can be used, permitting us to weigh the advantages and disadvantages of both alternatives. An attempt is made to define the heretofore untapped potential of the method in the area of comparative studies in Sect. 3. This possibility has been the least explored to date.

1. Historiographic References

Ancient history scholars claim to be the fathers of prosopography (Chastagnol 1970, Nicolet 1970). The term was conspicuously used in the modern sense by T. Mommsen in the announcement of the *Prosopographia Imperii Romani*, published in 1897 (Groag and Stein 1897, Lalouette 1999). This serial was imitated for other periods of ancient history throughout the twentieth century, both in collective works and in individual monographs. Prime examples are *The Prosopography of the Later Roman Empire* by A. H. M. Jones and J. R. Martindale (1971–92), and *Prosopographie chrétienne du Bas-Empire*, of which two volumes have appeared, edited by André Mandouze (1982) and Charles Pietri and Luce Pietri (1982). Countless monographs contain prosopographies of elites of antiquity. One of the earliest is Jean Kirchner's *Prosopographia attica* (1901), followed by Paul Poralla's 1913 work on the Lacedaemonians and the classics by Friedrich Münzer (1920), Ronald Syme (1939) and Claude Nicolet (1966), etc. The application of prosopography to medieval times started to develop between the world wars. Studies focus on administrative personnel and the entourage of the sovereigns of the principal Western European states. Since the 1960s, the body of literature has became quite copious, as various surveys (Bulst 1986, Griffiths 1986) reveal. The same goes for modern history, where the initiator was Sir Lewis Namier (1929, 1930), who studied English political elites. His works inaugurated an ongoing series of studies on the House of Commons from the Middle Ages to 1832 (Cruickshanks 1986). French and Italian investigation has concentrated on diverse selections of officials taken from public administration and the judiciary as well as on ecclesiastical, intellectual, financial and commercial elites (see Charle et al. 1980, Nagle 1986).

In the area of contemporary history, work began during the second half of the 1960s. Most of it was inspired by the perspective on elites introduced by sociologists, particularly American, influenced by the theses of Pareto and Mosca beginning in the 1950s. In France, one of the issues that stimulated research of this kind was the debate on the French Revolution opened by the theses of F. Furet and D. Richet (Richet 1969, discussed by Vovelle 1974), proposing an alternative to the Marxist interpretation then dominant in France, founded on prosopographic study of elites before and after the upheaval. This is the background of the investigation by L. Bergeron and G. Chaussinand-Nogaret on notables of the Napoleonic Era (1979). Similar works on prominent figures of the nineteenth century and the bourgeoisie since the 1960s have relied on the prosopographic method. This approach was inspired by Ernest Labrousse (Tudesq 1964) (for an overall survey, see Charle 1992). French sociology of education, influenced by the works of Pierre Bourdieu (Bourdieu and Passeron 1970,

Bourdieu 1989), has also been the source of numerous intellectual and academic prosopographies (Karady 1972, 1973). In Britain, Germany, and the USA and more recently in Italy, Spain, and Switzerland, collective biographies of political, economic, religious, intellectual, artistic, and militant elites have proliferated. They are an outgrowth of great national historiographic debates on subjects such as the foundation of political rifts in the British parliament, the evolution and function of English nobility, the distinguishing characteristics of German, Swiss, and Italian bourgeoisies (Kocka and Frevert 1988, Tanner 1990, Augustine 1994, Malatesta 1995), the formation and reproduction of American elites (Baltzell 1958, Jeher 1973), and so on. In contemporary history, the method has also been applied to larger groups than elites, which are the sole focus of prosopography in the spheres of ancient, medieval and modern history because records on other segments of the population are not available. The history of women, the middle classes, lower classes and even marginals have also availed themselves of collective biography in recent decades thanks to abundant printed sources, oral history and new assessments of large biographic files which already existed (Dreyfus et al. 1996, Omnès 1997).

2. *Merits and Limitations*

The proliferation of studies in contemporary history and historical sociology using the prosopographic method is due to intellectual trends and techniques developed since the 1970s. The concomitant decline of global statistical methods and of the intellectual impact of Marxism, which featured social classes with huge populations, and the desire, typical of societies focused more and more on the individual, to come to grips with individual experience and the diversity of social patterns have inclined historians of recent times to emphasize social microhistory, which gives priority to prosopography or collective biography. Finally, the widespread advance of computerization has facilitated complex processing of bigger and bigger databases (Stone 1971, Millet 1985). Whereas in former times the lack of mass documentation required patient gathering of data on the individual level before any collective generalizations could be made, in recent periods resorting to prosopography has been construed as a change of perspective from traditional social history based on serial documentation.

Thanks to this technique, since the 1970s historians have increasingly set their sights on gaining intimate knowledge of small problem-related constituencies. Thus, S. W. Serman's study of officers of the Second Republic and the Second Empire traced the long-lasting confrontation in nineteenth century France between the aristocratic ideal and the rising democratic tide (Serman 1978). Investigations of elites (high-ranking personalities of the July Monarchy (Tudesq 1964, Girard et al. 1976), the business leaders of the Second Empire (Plessis 1985, Barjot et al. 1991–1999), ruling circles of the Third Republic (Estèbe 1982, Charle 1987), and deputies to the national assemblies in Paris and Frankfurt in 1848–49 (Best 1990)) were attempts to understand the evolution of models of reproduction dominant in one regime toward those prevailing in other systems, the disparity between myths justifying the social order and sociological mechanisms, the possible social impact of political revolutions, reasons for the unique French road to industrialization, the peculiar German path to an incomplete parliamentary system, and the relations between English entrepreneurs and previously dominant elites (Berghoff 1991). In short, these collective biographies allow us to revamp our answers to some major questions as well as to conduct an in-depth reassessment of cherished tenets of social history without confining ourselves to the preconceived structure of serial and quantifiable sources. In these new approaches, groups define themselves with reference to relational characteristics or their reciprocal images or even their capacity to impose an image of themselves on others and on the majority of their members. The notion of construction of the object thus becomes decisive in this microhistory based on collective biographies.

This is precisely the point which since 1971 has triggered premature criticism of prosopography, summed up by Lawrence Stone with reference to modern history studies inspired by Sir Lewis Namier: the biases of official sources on which the biographies rely lead to a partial view of reality, the delimitation of populations is largely arbitrary, and criteria applied for the biographies are often restrictive. All this contributes to an elitist, cynical and conformist perspective on the spheres of leadership and their relations with society at large (Stone 1971). These criticisms are not valid unless the historians confuse the method with its objective and forget that they can never consider more than a fraction of reality as determined by the sources which they select and the limitations of their biographical questionnaires. They are less likely to fall into these traps for recent periods, where the findings of collective biographies can be confronted with other sources, where partial prosopographies can be compared and merged, and where the diversity of groups readily available for investigation goes well beyond the scope of the elites to which earlier studies were confined.

The prosopographical historian navigates permanently between two reefs: the undefined biography of individuals (with the risk of losing sight of the collective dimension) and the opposite: the expansion of prolific samples to the scale of the entire society (with the risk of reducing the questionnaire to its lowest common denominator). To escape from this dilemma, collective investigation by means of large

research teams, adopting common principles, has been applied with varying success in different countries. Examples include the ongoing study of the British House of Commons (Namier and Brooke 1964), the study of mayors in France (Agulhon et al. 1986), the investigation of members of parliament of the French Third Republic (Corbin and Mayeur 1995), the collective biographies of legislators by the Zentrum für Historische Sozialforschung in Cologne (Schröder et al. 2000, Best and Cotta 2000). These surveys require a long-term institutional financial commitment, which has become more and more difficult to find since profitability criteria were introduced in academic research. They also assume long-term involvement of loyal research personnel, which is less and less compatible with normal levels of staff turnover. This is why the principle of coordinated monographs, being more flexible and decentralized, appears to be more realistic in the contemporary university environment. It also overcomes the need to sacrifice the originality of each individual's work in the anonymity of a hierarchic endeavor. However, to avoid the risk of dissipation or non-comparability of partial results, it posits a minimum of harmonization and entails meetings to fine-tune the common questionnaires, coding systems and evaluations. In exploiting collective biography, one is not only forced to gauge the conditions of sample validity. One must also recognize the limits of the categories applied *a priori* to the data. Since the responses to the questions which are included involve social stakes, the process of interpretation of prosopographic data is characterized, more so than in other methods, by historians who consciously or unconsciously put their mark on the results at all stages of the study: sampling, data gathering, coding, and processing.

3. *Toward Comparative Prosopography*

In France and increasingly in most other developed countries, nearly all groups that lend themselves to collective biography have been adopted by a biographer: aristocracies, dignitaries, urban bourgeoisies, administrative elites, financial, commercial and entrepreneurial elites, intellectual and university elites, artists, middle-class professionals (physicians, jurists, journalists, professors), students, officeholders at all levels, labor leaders and militant feminists, even marginals, etc. Is a decline of the method of collective biography imminent? Will it be stopped in the name of the unyielding law of diminishing returns and as a result of the resurgence of historiographic topics which are less amenable to this approach (cultural history, history of memory, history of collective sensibilities)? Three considerations allow us to reject this premature diagnosis of a decline in the advantages of a method which has been so fruitful in the area of recent social history.

First, apart from elites, other social groups which are beginning to revise their social history thanks to insiders' collective biographies (middle classes and, increasingly, lower classes, not only through the eyes of their militant elites, but also in the form of life stories of privileged witnesses) (Ozouf et al. 1992, Pudal 1991, Gribaudi 1987) are far from being thoroughly investigated. Application of the method to these new fields requires new reflection on the composition of samples (which can no longer be exhaustive) and on the relations between individual paths and morphology: the internal diversity of the configurations is directly proportional to the size of the target population. This again raises the issue of sample representativeness, which had disappeared when investigators employing prosopography of elites resorted to exhaustive polls. All these groups are transitory, while elites, by definition at the top of the heap, are target groups; hence characteristics of the individuals of which such groups are composed cannot be viewed in isolation, but must be understood as vectors of multiple strategies.

Second, European historiographers are acquainted with the adjustments required to enable them to merge complementary studies to arrive at a homogeneous scrutiny of comparable groups from different societies. Thus, the comparative research coordinated by Jürgen Kocka on European bourgeoisies has demonstrated that not all of the difficulties in analyzing this group at an international level are due to the unequal advancement of studies; above all they arise from the collateral persistence of divergent investigations—some inspired by classical methods of social history, particularly in Germanic and Anglo-Saxon countries, others, particularly in France, primarily involving research based on collective biographies, which result in different approaches to different problems, and at times to inappropriate comparisons (Kocka and Frevert 1988, Charle 1990).

Collective biography, therefore, still has a nearly unexploited field at its disposal: that of comparative prosopography. The principal, eternal objection to comparative historical study is the necessity, given the abundance of material, to conduct secondhand research, using monographs prepared by other scholars, with the compounded risks of hasty generalization of conclusions taken out of their unique context and of the permanent lack of correspondence in the compared data, particularly in social history, on account of the recurring coding problem. Despite these obstacles, German researchers have already forged ahead, and certain French historians and specialists in England have followed suit since the 1990s (Best 1990, Berghoff and Möller 1994, Charle 1994, Siegrist 1995, Cassis 1997, Ruggiu 1997). Collective biography is an extraordinarily demanding methodology in terms of research time and rigorous organization of data. Comparative prosopography multiplies the time it takes to assemble the data by at least two—for two

countries—or more, if one wants to compare a greater number of cases. The problem is compounded by the extra time needed to design the research and evaluate the data. In order to test the initial hypothesis of feasibility and fruitfulness of a comparative project, it is convenient to start in a domain where one of the two countries involved already has a critical mass of usable data allowing researchers to define an analogous scope of investigation in the other selected country, which must possess a comparable, evident interest in the chosen elite. This explains the fact that comparative prosopographies deal primarily with the best-documented and most frequently studied elites (in big business, politics, higher education and high society), as in the cases mentioned previously. Collective biographies at a national level have permitted us to better understand the internal rifts in the various groups which have been studied as well as their social and generational dynamics. In certain favorable cases, it has been possible to relate these rifts to political, ideological or religious stances or apply them to better understand certain failures or successes. Comparative prosopography should allow us to go even further in this interpretive analysis by relativizing correlations which were taken for granted in a given national or social setting. They should account for common trends which transcend borders and the peculiarities of one moment in time, of one milieu or one nation. For instance, the generally accepted notion that the expansion of higher education translates into an opening of university teaching positions to a wider segment of society is disproved if one compares the collective biographies of the professors of the University of Berlin (Arts and Sciences) with their counterparts at the Sorbonne at the turn of the nineteenth and twentieth centuries. The major divergence accounts for the misunderstandings between the two academic communities even when they attempt to cooperate, and goes a long way toward explaining their widely varying relations with other elites during political crises (Charle 1994).

The goal of comparative prosopography, as with any kind of prosopography, is to employ collective biographies to portray the real social behavior of the institutions or the milieus in which the individual subjects of the study interact (Genet and Lottes 1996). The main danger in applying comparative prosopography is therefore, as with many erudite ventures, that one will discover nothing but well-established or generally acknowledged truths in each country concerned. One usually arrives at this impasse or this disappointing result when one deals with samples that are too vast to lend themselves to controlled comparisons which would allow the study to qualify any dominant feature or put it into perspective. In light of this danger it is essential to make a particularly judicious choice of variables to be compared. We face a series of common methodological issues which transcend historical periods. Their resolution would allow us to launch a wider debate on the possible foundations of a comparative sociocultural history of European society.

See also: Antiquity, History of; Biographical Methodology: Psychological Perspectives; Biography and Society; Biography: Historical; Comparative History; Ethnoarchaeology; Historiography and Historical Thought: Current Trends; History: Forms of Presentation, Discourses, and Functions; History: Theories and Methods; Memory: Collaborative; Middle Ages, The; Quantification in History; Social History; Social Inequality in History (Stratification and Classes)

Bibliography

Agulhon M et al. 1986 *Les Maires en France du Consulat à nos jours*. Publications de la Sorbonne, Paris

Augustine D L 1994 *Patricians and Parvenus. Wealth and High Society in Wilhelmine Germany*. Berg, Oxford

Autrand F (ed.) 1986 *Prosopographie et genèse de l'État moderne*. L'École normale supérieure des jeunes filles, Paris

Baltzell E D 1958 *Philadelphia Gentlemen. The Making of a National Upper Class*. Free Press, Glencoe, IL

Barjot D (ed. in chief) 1991–1999 *Les patrons du Second Empire*. Cenomane, Picard, Le Mans, Paris

Bergeron L, Chaussinand-Nogaret G 1979 *Les Masses de Granit*. EHESS, Paris

Berghoff H 1991 *Englische Unternehmer 1870–1914. Eine Kollektivbiographie führender Wirtschaftsbürger in Birmingham, Bristol und Manchester*. Vandenhoeck & Ruprecht, Gottingen, Germany

Berghoff H, Möller R 1994 Tired pioneers and dynamic newcomers? A comparative essay on English and German entrepreneurial history, 1870–1914. *Economic History Review* XLVII **2**: 262–87

Best H 1990 *Die Männer von Bildung und Besitz. Struktur und Handeln parlamentarischer Führungsgruppen in Deutschland und Frankreich, 1848/49*. Droste Verlag, Düsseldorf, Germany

Best H, Cotta M (eds.) 2000 *Parliamentary representatives in Europe, 1848–2000. Legislative Recruitment and Careers in Eleven European Countries*. Oxford University Press, Oxford, New York

Bourdieu P, Passeron J-C 1970 *La reproduction*. Minuit, Paris

Bourdieu P 1989 *La Noblesse d'État, grandes écoles et esprit de corps*, Paris, Minuit Paris

Bulst N, Genet J-Ph (eds.) 1986 *Medieval Lives and the Historian*. Studies in Medieval Prosopography, Kalamazoo, MI

Bulst N 1986 La recherche prosopographique récente en Allemagne (1250–1650). Essai d'un bilan. In: Autrand F (ed.). Prosopographie et genèse de l'État moderne. L'École normale Supérieure des jeunes filles. Paris, pp. 35–52

Cassis Y 1997 *Big Business: the European Experience in the Twentieth Century*. Oxford University Press, Oxford, New York

Charle C et al. 1980 *Prosopographie des élites françaises (XVIè-XXè siècles). Guide de recherche*. CNRS, Institut d'histoire moderne et contemporaine, Paris

Charle C 1987 *Les Élites de la République (1880–1900)*. Fayard, Paris

Charle C 1990 A la recherche des bourgeoisies européennes. *Le Mouvement Social* **153**: 91–7

Charle C 1992 The present state of research on the social history of elites and the bourgeoisie: a critical assessment on middle class and elites *Contemporary European History* **1**(1): 99–112

Charle C 1994 *La République des universitaires (1870–1940)*. Le Seuil, Paris

Charle C 1996 Du bon usage de la biographie sociale comparée. In: Dreyfus M, Pennetier C, Viet-Depaule N (eds.) *La part des militants, Biographie et mouvement ouvrier: Autour du Maîtron, Dictionnaire biographique du mouvement ouvrier français*. Editions de l'Atelier, Paris, pp. 51–72

Chastagnol A 1970 La prosopographie méthode de recherche sur l'histoire du Bas Empire. *Annales ESC* **25**(5): 1229–35

Corbin A, Mayeur J-M (eds.) 1995 *Les Immortels du Sénat*. Publications de la Sorbonne, Paris

Cruickshanks E 1986 Multi-biographical analysis as an approach to parliamentary history. In: Autrand F (ed.) *Prosopographie et genèse de l'État moderne*. L'École normale supérieure de jeunes filles, Paris, pp. 335–44

Dreyfus M, Pennetier C, Viet-Depaule N (eds.) 1996 *La part des militants, Biographie et mouvement ouvrier: Autour du Maîtron, Dictionnaire biographique du mouvement ouvrier français*. Editions de l'Atelier, Paris

Estèbe J 1982 *Les Ministres de la République (1871–1914)*. Presses de la Foundation Nationale des Sciences Politiques, Paris

Genet J-Ph, Lottes G (eds.) 1996 *L'État moderne et les élites XIIIè–XVIIIè siècles. Apports et limites de la méthode prosopographique*. Publications de la Sorbonne, Paris

Girard L, Serman W, Gossez R 1976 *La Chambre des députés en 1837–39*. Publications de la Sorbonne, Paris

Gribaudi M 1987 *Itinéraires ouvriers. Espaces et groupes sociaux à Turin au début du Xxème siècle*. Editions de l'EHESS, Paris

Griffiths R 1986 Bureaucracy and the English State in the Later Middle Ages. In: Autrand F (ed.) *Prosopographie et genèse de l'État moderne*. L'École normale supérieure des jeunes filles, Paris, pp. 52–65

Groag E, Stein A (eds.) 1897 *Prosopographia Imperii Romani Saec I. II III Pars I*. Walter de Gruyter, Berlin and Leipzig, Germany

Jaher F C (ed.) 1973 *The Rich, the Well born, and the Powerful. Elites and Upper Classes in History*. University of Illinois Press, Urbana, IL

Jones A H M, Martindale J R (eds.) 1971, 1980, 1992 *The Prosopography of the Later Roman Empire*. Cambridge University Press, Cambridge

Karady V 1972 Normaliens et autres enseignants à la Belle Époque. Note sur l'origine sociale et la réussite dans une profession intellectuelle. *Revue francaise de sociologie* **13**(1): 35–58

Karady V 1973 L'expansion universitaire et l'évolution des inégalités devant la carrière d'enseignant au début de la IIIème République. *Revue française de sociologie* **14**(4): 443–70

Kocka J, Frevert U (eds.) 1988 *Bürgertum im 19. Jahrhundert*. Deutsche Taschenbuch Verlag, Munich, Germany

Lalouette J 1999 De l'exemple à la série: histoire de la prosopographie. *Siècles, cahiers d'histoire des entreprises et des communautés* **10**: 3–20

Malatesta M (ed.) 1995 *Society and the Professions in Italy, 1860–1914*. Cambridge University Press, Cambridge

Mandouze A (ed.) 1982 *Prosopographie chrétienne du Bas-Empire 1 Prosopographie de l'Afrique chrétienne (303–533)*. Editions du CNRS, Paris

Millet H (ed.) 1985 *Informatique et prosopographie*. Editions du CNRS, Paris

Münzer F 1920 *Römische Adelsparteien und Adelsfamilien*. J. B. Metzlersche Verlagsbuchhandlung, Stuttgart, Germany

Nagle J 1986 Prosopographie et histoire de l'État: la France moderne XVIè-XVIIIè siècles. In: Autrand F (ed.) *Prosopographie et genèse de l'État moderne*. L'École normale supérieure des jeunes filles, Paris

Namier L 1929 *The Structure of Politics at the Accession of George III*. Macmillan, London

Namier L 1930 *England in the Age of the American Revolution*. Macmillan, London

Namier L, Brooke J 1964 *The House of Commons, 1754–1790*. Published for the History of Parliament Trust by her Majesty's Stationery Office, London

Nicolet C 1966 *L'Ordre équestre à l'époque républicaine (312–43 avant J.-C)*. De Boccard, Paris

Nicolet C 1970 Prosopographie et histoire sociale: Rome et l'Italie à l'époque républicaine. *Annales ESC* **25**(5): 1209–28

Omnès C 1997 *Ouvrières parisiennes. Marchés du travail et trajectoires professionnelles au 20è siècle*. Éditions de l'EHESS, Paris

Ozouf J et al. 1992 *La République des instituteurs*. Gallimard, Le Seuil, Paris

Pietri C, Pietri L (eds.) 1982 *Prosopographie chrétienne du Bas-Empire. 2 Prosopographie de l'Italie chrétienne (303–533)*. Ecole française de Rome, Rome

Plessis A 1985 *Régents et gouverneurs de la Banque de France* (1852–1870). Droz, Geneva, Switzerland

Poralla P 1913 *Prosopographie der Lakedaimonier bis auf die Zeit Alexanders des Grossen*. Kommissions-Verlag von Max, Breslau, Germany

La prosopographie, problèmes et méthodes 1988. In: *Mélanges de l'École française de Rome, Moyen Age, Temps modernes* **100** (1): 5–284.

Pudal B 1991 *Prendre parti, pour une sociologie historique du PCF*. Presses de la Foundation Nationale des Sciences Politiques, Paris

Richet D 1969 Autour des origines idéologiques lointaines de la Révolution française. *Annales ESC* **24**(1): 1–23

Ruggiu F-J 1997 *Les élites des villes moyennes en France et en Angleterre XVIIè-XVIIIè siècles*. L'Harmattan, Paris

Schröder W-H, Weege W, Zech M 2000 Historische Parlamentarismus-, Eliten- und Biographieforschung, Forschung und Service am Zentrum für Historische Sozialforschung, Köln, Zentrum für Historische Sozialforschung. *Historische Sozialforschung Supplemente/Beiheft* **11**

Serman S-W 1978 *Le corps des officiers sous la Deuxième République et le Second Empire*. Atelier de reproduction des thèses de Lille III, Lille, France

Siegrist H 1995 *Advokat, Staat und Bürger*. V. Klostermann, Frankfurt, Germany

Stone L 1971 Prosopography. *Daedalus* **100**(1): 46–79

Syme R 1939 *The Roman Revolution*. Clarendon Press, Oxford

Tanner A 1990 Aristokratie und Bürgertum in der Schweiz im 19. Jahrhundert: Verbürgerlichung der 'Herren' und aristokratische Tendenzen im Bürgertum. In: Brändli S et al. (ed.) *1990 Schweiz im Wandel. Studien zur neueren Gesellschaftsgeschichte*. Helbing & Lichtenhahn, Basel & Frankfurt, pp. 209–28

Tudesq A-J 1964 *Les grands notables en France (1840–49). Étude historique d'une psychologie sociale*. Presses Universitaires de France, Paris

Vovelle M 1974 L'élite ou le mensonge des mots. *Annales ESC* **29**(1): 49–72

Kirchner J 1901 *Prosopographia attica*. Reimer, Berlin, Germany

C. Charle

Prospective Memory, Psychology of

Prospective memory refers to remembering to perform an action in the future such as remembering to pick up a loaf of bread on the way home, to give a friend a message, to send a birthday card to your aunt, to take your medication, and to pick someone up at airport at 4 p.m. Ellis (1996, p. 1) defines it somewhat more precisely as the 'realization of delayed intentions and associated actions.' Prospective memory can be contrasted with retrospective memory, which is memory for past events.

Examples of retrospective memory include remembering the name of your eighth grade teacher, the causes of the World War I, the formula for the area of a circle, what you had for lunch yesterday, and a list of words learned in an experiment 15 minutes earlier. In thinking about the distinction between prospective and retrospective memory, it is worthwhile considering that many of the same processes may be involved in retrospective and prospective memory tasks. It is interesting to note that all prospective memories are to some extent retrospective in the sense that you have to have retrospective memory for the action that is to be performed and when it is to be performed in order to remember to perform it on time. For example, in order to remember to give a friend a message, you have to remember the message and to whom it is to be given. In addition, however, successful prospective memory requires that you actually remember to do this when you see your friend, and this is the prospective component of the memory. One fairly common error that is made in this situation is that upon seeing your friend, you remember that you have a message to give but forget the message (a retrospective memory problem). Another is that you forget to give the message even though you have good memory for it (a prospective memory failure).

1. Contrasts with Retrospective Memory

When studying retrospective memory (at least explicit forms of retrospective memory) in the laboratory, the experimenter presents some materials for learning. After a delay of some duration, the experimenter puts participants in a retrieval mode and asks them to recollect intentionally as they perform a designated criterial task such as free or cued recall. By contrast, when studying prospective memory, the experimenter asks participants to perform an action or several actions at a later point in time (either at a particular time or when a certain event occurs; see Sect. 2.2 for examples). The interest is in whether participants remember to perform the task at the appropriate time/event and without an external agent (e.g., the experimenter) prompting participants to remember at the critical period. A key difference then between retrospective and prospective memory is that while the process of remembering is initiated by the experimenter in retrospective memory, it is initiated by the participant in prospective memory.

This self-cueing feature of prospective memory, along with the fact that retrieval needs to occur sometime in the future, places a different and perhaps more complex combination of demands (than typical retrospective memory tasks) on our cognitive systems. Indeed, many researchers have argued that prospective memory involves much more than memory processes (Dobbs and Reeves 1996, Ellis 1996, Burgess and Shallice 1997, Marsh et al. 1998b). They make the case that plan creation, plan modification, monitoring or rehearsal of the conditions for performing the action, and output monitoring, as well as individual factors such as working memory resources and personality variables, are all essential to a complete understanding of prospective memory.

Despite the differences just mentioned, it is likely that prospective and retrospective memory share similar memory processes and, accordingly, our vast understanding in the retrospective memory arena will be useful for understanding prospective memory. For example, given that many prospective memory tasks involve realizing a connection between an event (seeing a friend) and an intended action (giving a message), our basic understanding of associative memory processes should prove useful in understanding prospective memory. Also, as in retrospective memory, it seems that the level of processing of the target event is critical for good prospective memory as is the overlap between processing of the target event at encoding and retrieval (McDaniel et al. 1998).

2. Research in Prospective Memory

2.1 Historical Overview

Nearly all memory research conducted thus far has focused on retrospective memory, and this seems inconsistent with the nature of our everyday memory demands. Our lives are replete with prospective memory demands such as fulfilling an appointment, giving a message, remembering to flip the pancakes at the appropriate time, and returning a library book. To get some sense of the prevalence of prospective memory in everyday life, for several years G. O. Einstein asked students in his classes to write down the last thing they remembered forgetting. Interestingly, 62 percent of college-age students and 54 percent of older students (60 years and older) listed responses that were prospective memory failures. Given its prevalence in

everyday living, there is a real gap in our scientific understanding of prospective memory. It is difficult to overemphasize this gap as Kvavilashvili and Ellis (1996) in a recent review found only 45 prospective memory papers published over the prior 20 years, and only a small percentage of these papers consisted of rigorous experimental work in major journals. However, as evidenced by the inclusion of prospective memory in this volume, there now appears to be strong interest in the area.

2.2 Experimental Paradigms

Initial research in prospective memory required prospective memory responses to be initiated outside of the laboratory. For instance, subjects were asked to return postcards (e.g., Meacham and Singer 1977) or make telephone calls at specified times (Moscovitch 1982). Although these paradigms were valuable in producing the first investigations of factors thought to influence prospective remembering (like a person's age), they did not allow control of extraneous factors that affect prospective remembering. For instance, in these experiments there may have been variation in the degree to which subjects used external aids such as calendars to help complete the prospective memory task.

Empirical and theoretical work in prospective memory has increased significantly with the introduction of laboratory paradigms that bring experimental control and precision to the study of prospective memory. One influential paradigm, developed by Einstein and McDaniel (1990), requires that the subject be busily engaged in a primary task that serves as a cover task for the prospective memory component. During the instructions for the primary task, subjects are also instructed that they are to remember on their own to perform another action (e.g., press a special function key on the keyboard) at some designated point after they begin the primary task.

Sometimes the designation is an event embedded in the primary task (e.g., a particular word, a particular type of question) and other times the designation is a period of time that has elapsed (e.g., after 5 minutes). In this paradigm, the retrospective memory demands are designed to be simple and well remembered so that the major focus is whether participants remember to perform the intended action at the appropriate point. Other paradigms that focus on planning involve having participants remember to perform multiple sets of activities, each with their own constraints (e.g., Kliegel et al. 2000).

2.3 Principal Findings and Theoretical Perspectives

The general theoretical orientation guiding prospective memory research has turned on potential differences between prospective memory processes and retrospective memory processes. Unique features of prospective memory seem possible at encoding, storage, and retrieval stages of memory.

Encoding refers to rehearsal and elaboration of information so as to enter or commit that information to memory. People may use different encoding strategies for prospective memory tasks than they use for retrospective memory tasks. Consistent with this possibility, Koriat et al. (1990) found that actions encoded to be executed later are better recalled than actions that are encoded for a later recall test. It is likely that additional encoding processes found in prospective memory tasks are related to the requirement to remember to perform the intended action. Most prominently, during encoding of a prospective memory intention people often construct plans for how and when the intended activity will be accomplished (Ellis 1996, Marsh et al. 1998b). These plans may also anticipate the conditions or arrange the conditions under which the intended action must be retrieved, information that may not typically be considered when engaged in encoding for a retrospective memory task.

Only a few studies have examined the planning processes in prospective memory, but it is already clear that individuals vary in their prospective memory planning in several ways. People vary in the extent to which they use external aids like notes or lists to keep track of their plans (Marsh et al. 1998b). The complexity of the plans also varies and for older adults especially, more complex plans seem to be related to the effectiveness with which the intention is eventually executed (Kliegel et al. 2000).

Storage refers to maintaining information in memory during the retention interval. It would be adaptive if the storage of prospective memory were somewhat special so that the intended action would be remembered at the time it needed to be performed. Evidence to date suggests there may indeed be several ways in which prospective memory storage is different from retrospective memory storage. First, it may be that prospective memories (intended actions) enjoy greater and more persistent activation in memory (Goschke and Kuhl 1993, Marsh et al. 1998a).

Theoretically, memories in a more activated state more easily come to mind when one comes across cues in the environment that are related to those memories. It would also be adaptive if these memories for intentions would be deactivated after the activity was performed, so that they would not come to mind later. Preliminary findings suggest that such deactivation may occur (Marsh et al. 1998a), although for older adults under demanding conditions it appears that the memory for the intention may persist so that they inappropriately repeat their performance of the intended action (Einstein et al. 1998).

Another special storage feature of prospective memory suggested in one study is that prospective mem-

ories may actually become stronger as the retention (storage) time lengthens (Hicks et al. 2000). This is in dramatic contrast to the well-known finding that retrospective memory often declines with increased retention intervals. Additional research is needed to explain this provocative finding that prospective memories become better as the retention time increases. One clear possibility is that people remind themselves about the intended activity during the retention interval, especially as the duration increases between the formation of the initial intention and the point at which the intention can be performed. Remindings would presumably refresh the memory of the intended activity, thereby increasing the memorability of the intended action and its likelihood of being successfully performed at the appropriate time (Guynn et al. 1998).

Retrieval refers to accessing information in long-term memory so that it can be brought into awareness and used in a conscious fashion. As noted earlier in this article, a critical feature of prospective remembering is that retrieval must be initiated by the individual performing the prospective memory task. By contrast, retrieval in a retrospective memory task is initiated by somebody else (e.g., an experimenter) who prompts the individual to retrieve the target information. Accordingly, retrieval processes in prospective memory are thought to be somewhat different from those that are characteristic of retrospective memory retrieval.

Cognitive psychologists are not in agreement about how prospective memory retrieval operates. Consistent with the above analysis, one view proposes that prospective memory tasks are very high in self-initiated retrieval (Craik 1986) and therefore require a high level of cognitive resources to support retrieval. Another view proposes that initiating an intended action is not so much a memory process as it is an attentional process in which the person monitors the environment for signals that indicate that it is appropriate to perform the action (e.g., Burgess and Shallice 1997). By these views, prospective memory tasks should be especially difficult for older adults because of declines in cognitive or attentional resources that accompany aging. In support of this expectation, age-related decrements have been reported in time-based prospective memory tasks (Einstein et al. 1995, Park et al. 1997), tasks for which the intended action must be performed after a certain period of time has elapsed.

However, in some studies, older and younger subjects have surprisingly demonstrated equivalent prospective memory performance in event-based prospective memory tasks (Einstein and McDaniel 1990, Einstein et al. 1995). In an event-based task an external event signals that it is appropriate to perform the intended action. This event (or cue) might be the appearance of a friend to whom you have to give a message or a particular word in an experiment to which you have to give a special response. The absence of age differences in event-based prospective memory suggests that prospective memory retrieval can be a relatively automatic process. One idea is that the environmental cue may automatically stimulate the retrieval from memory of the intended action, after which the action can be performed (McDaniel et al. 1998).

More research is needed to understand completely prospective memory retrieval. Based on the literature it seems likely that prospective memory tasks may vary such that those with very few environmental cues (such as time-based tasks) may rely heavily on self-initiated retrieval and attentional processes, whereas prospective memory tasks with salient environmental cues may be supported by more automatic retrieval mechanisms. In addition, future work in prospective memory will no doubt increasingly focus on the planning processes that are involved in prospective memory (e.g., Dobbs and Reeves 1996, Burgess and Shallice 1997, Kliegel et al. 2000) and the potential special storage dynamics of prospective memories.

See also: Citizenship: Political; Elaboration in Memory; Incidental versus Intentional Memory; Memory Retrieval; Recognition Memory, Psychology of

Bibliography

Burgess P W, Shallice T 1997 The relationship between prospective and retrospective memory: neuropsychological evidence. In: Conway M A (ed.) *Cognitive Models of Memory*. 1st MIT Press edn., MIT Press, Cambridge, MA, pp. 247–72

Craik F I M 1986 A functional account of age differences in memory. In: Klix F, Hagendorf H (eds.) *Human Memory and Cognitive Capabilities: Mechanisms and Performances*. Elsevier, Amsterdam, pp. 909–22

Dobbs A R, Reeves M B 1996 Prospective memory: more than memory. In: Brandimonte M, Einstein G O, McDaniel M A (eds.) *Prospective Memory: Theory and Applications*. L. Erlbaum, Mahwah, NJ, pp. 199–225

Einstein G O, McDaniel M A 1990 Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**: 717–26

Einstein G O, McDaniel M A, Smith R E, Shaw P 1998 Habitual prospective memory and aging: remembering intentions and forgetting actions. *Psychological Science* **9**: 284–8

Einstein G O, Richardson S L, Guynn M J, Cunfer A R 1995 Aging and prospective memory: examining the influences of self-initiated retrieval processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 996–1007

Ellis J A 1996 Prospective memory or the realization of delayed intentions: a conceptual framework for research. In: Brandimonte M, Einstein G O, McDaniel M A (eds.) *Prospective Memory: Theory and Applications*. L. Erlbaum, Mahwah, NJ, pp. 1–22

Goschke T, Kuhl J 1993 Representation of intentions: persisting activation in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**: 1211–26

Hicks J L, Marsh R L, Russell E J 2000 The properties of

retention intervals and their affect on retaining prospective memories. *Journal of Experimental Psychology*: *Learning*, *Memory and Cognition* **26**: 1160–9

Kliegel M, McDaniel M A, Einstein G O 2000 Plan formation, retention, and execution in prospective memory: a new approach and age-related effects. *Memory and Cognition* **28**: 1041–49

Koriat A, Ben-Zur H, Nussbaum A 1990 Encoding information for future actions: memory for to-be-performed tasks versus memory to to-be-recalled tasks. *Memory and Cognition* **18**: 568–78

Kvavilashvili L, Ellis J 1996 Varieties of intention: some distinctions and classifications. In: Brandimonte M, Einstein G O, McDaniel M A (eds.) *Prospective Memory: Theory and Applications*. Erlbaum, Mahwah, NJ, pp. 23–51

Marsh R L, Hicks J L, Bink M L 1998a Activation of completed, uncompleted, and partially completed intentions. *Journal of Experimental Psychology*: *Learning*, *Memory*, *and Cognition* **24**: 350–61

Marsh R, Hicks J L, Landau J D 1998b An investigation of everyday prospective memory. *Memory and Cognition* **26**: 633–43

McDaniel M A, Robinson-Riegler B, Einstein G O 1998 Prospective remembering: perceptually driven or conceptually driven processes? *Memory and Cognition* **26**: 121–34

Meacham J, Singer J 1977 Incentive effects in prospective remembering. *Journal of Psychology* **97**: 191–7

Moscovitch M 1982 A neuropsychological approach to memory and perception in normal and pathological aging. In: Craik F I M, Trehub S (eds.) *Aging and Cognitive Processes*. Plenum, New York, pp. 55–78

Park D C, Hertzog C, Kidder D P, Morell R W, Mayhorn C B 1997 Effect of age on event-based and time-based prospective memory. *Psychology and Aging* **12**: 314–27

M. A. McDaniel and G. O. Einstein

Prostitution

Prostitution is simultaneously a sexual and an economic institution, and it is also highly gendered. The majority of prostitutes are female, and an even larger majority of customers are male. While men's prostitute use is widely tolerated, female prostitution is popularly viewed as a form of social and sexual deviance, and mainstream social scientists have traditionally reproduced such attitudes in their research on the topic. Feminist theorists, by contrast, have long been concerned to explore parallels between marriage, prostitution, slavery, and wage labor, as well as the sexual, political, and economic relations that underpin these institutions. They do not speak as one on the subject, however, and the division between feminists who are concerned with prostitution as a sexual institution and feminists who approach it first and foremost as a form of economic activity is particularly sharp. This article provides a brief overview of prostitution in the contemporary world and highlights the theoretical and actual problems it poses for gender scholarship.

1. *Prostitution: A Stigmatized and Criminalized Activity*

A profoundly negative stigma is almost universally attached to prostitute women. Religious thinking on men's prostitute use varies, but there is no major world religion that actively sanctions female prostitution and, in secular societies, 'scientific' thinking has done little to displace traditional attitudes. Mainstream medical, psychological, psychoanalytic, and sociological research on the topic has generally assumed that while men's prostitute use is based on natural, biologically determined sexual drives, women who prostitute are somehow abnormal, unnatural, a threat to public health and order. Prostitution law varies from country to country and even within individual nation states, but typically enshrines this kind of stigma by treating female prostitutes as a distinct class of persons, separate from other workers and/or women in terms of their rights to protection, privacy, and/or self-determination (Walkowitz 1980, Bindman 1997).

Law enforcement almost invariably focuses upon female prostitution, ignoring men's prostitute use (often also male prostitution), and is generally informed by one of two basic models: prohibition/abolition, or regulation/registration. The prohibitionist model has historically led to an emphasis on controlling public manifestations of prostitution through the criminalization of (usually street working) prostitute women. The regulation/registration model, which legalizes prostitution providing it takes place in licensed brothels or designated geographical zones, has been associated with a variety of civil rights violations, such as requirements for prostitute women to register with the police and/or other authorities, have compulsory health checks, and so on. In some countries, there have been moves to deregulate prostitution either explicitly or by default, a shift which is taking place largely for reasons of 'financial exigency, prosecutor indifference, court deadlocks, and resistance to the overreach of the criminal law among some social sectors' (Davis 1993, p. 8).

2. *The Economic Significance of Prostitution*

Because female prostitution is ideologically and legally construed as a form of deviance, states do not officially recognize or regulate prostitution as an economic sector like any other. It is therefore extremely difficult to obtain accurate data on the size and earnings of the sex trade. However, research indicates that sex commerce is a significant feature of economic life in many nations, regardless of their overall level of economic development. Studies suggest that, for example, in London, prostitution generates around £200 million annually (Matthews 1997); that in New Zealand, almost one in every 150 women aged between 18 and 40 is employed in some form of sex work (Lichtenstein 1999); and that in Thailand, the Philippines, Malaysia,

and Indonesia, the sex sector accounts for between 2 percent and 14 percent of gross domestic product (Lim 1998). Research further suggests that a fairly consistent set of factors underpin prostitution labor markets. Gender discriminatory social practices are structural 'push' factors in economically developed and underdeveloped nations alike (Davis 1993).

As part of an unofficial and often criminalized economic sector, prostitution is frequently connected with corruption, organized crime, and drug abuse, but it is often simultaneously integrated into mainstream economic structures, such as the entertainment and the tourism industries (Lim 1998). This draws attention to the increasingly international nature of the sex industry.

3. *Global Politics and Markets and Prostitution*

Prostitution is affected by global politics and market trends in three main ways. First, there are countries in which the development of the sex sector has been directly shaped by the military and economic interests of external powers. Thailand, for example, was used for 'Rest and Recreation' by the US military in the 1950s, 1960s, and 1970s, and a highly sexualized entertainment sector developed to meet demand from legions of young men. Western financial institutions and businesses soon recognized that such an entertainment industry was also suited to serve one particular segment of the long haul tourist market, and economic initiatives based on a 1975 World Bank report 'led to what is routinely described today as a $4-billion-a-year business involving fraternal relationships among airlines, tour operators, and the masters of the sex industry' (Bishop and Robinson 1998, p. 9). Similar developments have occurred elsewhere in South East Asia, as a result of either US or Japanese military or economic interventions in the region.

Sex commerce in poorer countries has also been affected by a set of linkages between international debt, price fluctuations in global commodity markets, economic development policy, and prostitution. Since the 1970s, world financial institutions have encouraged indebted nations to respond to economic crisis by developing tourism and/or 'nontraditional' export industries such as gold, diamonds, and timber. One side effect of such development policies is the creation of highly concentrated, effective demand for prostitution: affluent tourists seeking 'entertainment' and predominantly male, migrant workers in isolated mining and logging regions with cash to spend on 'recreation.' Meanwhile, structural adjustment measures have expanded the prostitution labor market. It is women and children who have been most adversely affected by the sudden introduction of cash economies into what were formerly subsistence economies, currency devaluations and concomitant falls in the price of labor, and the redirection of subsidies away from social spending and basic commodities towards debt servicing. In such climates, prostitution has often become the best or only economic alternative for large numbers of women and teenagers (Chant and McIlwaine 1995, Clift and Carter 1999, Kempadoo 1999). Third, the sex trade is international in the sense that many of those working in prostitution have either voluntarily migrated in order to prostitute or have been trafficked by third parties for that purpose. Migrant prostitutes are especially vulnerable to abuse and exploitation by third parties and, as both illegal migrants and prostitute women, particularly likely to be denied legal protection and civil rights (Bindman 1997).

Ideological, as well as economic and political factors underpin the globalization of sex commerce. Racist discourses, which typically construct the 'Other' as sexually debased and/or exotic, have long fueled desire for 'racially' Other prostitutes and continue to shape demand today (Shrage 1994).

4. *The Diversity of Prostitution*

Prostitution varies enormously in terms of its social organization and the power relations it involves. In virtually every country of the world, hierarchies exist within both female and male prostitution (West 1992, Aggleton 1999). At their apex are independent, self-employed adult prostitutes who exercise a relatively high level of control in transactions with clients and whose earnings are relatively high. At their base are individuals who gain little or no financial reward from their prostitution and exercise little, if any, control over it (including control over condom use). This may be because they are physically forced into a given work regime by a third party pimp or brothel-keeper, but it may also be because extreme economic and social marginalization, and/or youth and/or drug addiction renders them horribly powerless within transactions.

Between the two extremes are prostitutes who either work independently or enter into some form of employment relation with a third party. The degree of control they exercise over whether, when, how often, and on what terms they prostitute varies according to a range of factors, including their level of economic desperation; the contractual form of the prostitute–client exchanges they enter into; and the specific legal, institutional, social, political, and ideological context in which they prostitute. There are, for example, settings in which prostitution law so heavily penalizes independent prostitution that it effectively operates as a pressure on prostitutes to enter and remain in third party controlled prostitution no matter how exploitative the third party may be. Equally, in contexts where certain social groups (for instance, women, children, homosexuals, migrants, particular 'castes' or racialized minorities) are generally devalued, and/or denied full juridical subjecthood, and/or denied independent access to welfare support, their

vulnerability to prostitution and to third party exploitation and abuse within it is increased. It should also be noted that people come to prostitution as individuals of a certain age and with particular personal histories and experience, and this leaves some more open to exploitation and at higher risk of violence and sexually transmitted disease than others (O'Connell Davidson 1998).

5. *Conceptualizing Prostitution*

One approach to the task of conceptualizing prostitution is to examine how it simultaneously resembles and differs from other social institutions. In prostitution, as in the institution of marriage, the sexual and the economic are conjoined. Unlike marriage, however, prostitution does not construct long term, mutual, or diffuse relations of dependency between individuals and their kin. Instead, sexual interaction between prostitute and client takes place on the understanding that some immediate economic benefit will accrue to either the prostitute or a third party as a result, and the client's obligations are thus discharged by payment in cash or kind. To this extent, relations within prostitution resemble market relations.

However, prostitution does not involve any ordinary commodity exchange. The client does not wish to purchase the disembodied means to an orgasm, a 'thing' that can be detached from the prostitute's person. Rather clients part with money and/or other material benefits in order to secure certain powers of sexual command over the person of the prostitute. Thus, although prostitution is often thought of as the exchange of *sex* or *sexual services* for money and/or other benefits, it is better conceptualized as an institution which allows certain powers over one person's body to be exercised by another. In this sense, prostitution has something in common with slavery, wage labor, and relations between buyers and self-employed sellers of services. All are institutions through which certain powers of command over the person are transferred from one individual to another, conferring a right upon the client/slave holder/employer/buyer to require that the prostitute/slave/worker/seller acts in ways that she/he would not otherwise choose to act.

Yet the relationship between prostitute and client differs from that between slave and slave holder or worker and employer in crucial respects. Not only is it generally brief and transitory, but also, clients enter into it as consumers, not producers. Clients do not depend upon prostitutes for their economic well being in the way that slave-holders and employers depend upon slaves and workers for their material existence. And even where the prostitute is self-employed, the prostitute–client exchange is made unlike other exchanges between independent sellers and purchasers of services by the social meanings that are popularly attached to sexuality.

In monetary economies, sexual and economic relations are typically imagined, explained, and justified in very different ways. The products of human labor, human labor power, sometimes even human beings themselves, have been and are imagined as commodities to be bought and sold on the basis of rational economic calculation and/or in pursuit of status and social prestige. Human sexual interaction, by contrast, is generally regulated and given meaning through reference to premarket or nonmarket ideas, such as those pertaining to honor, shame and duty, and/or romantic love, and/or recreation, pleasure, and desire. Prostitution thus occupies a troubled and troubling space between two quite different symbolic domains. It does not readily fit into popularly understood categories of 'sex' or 'work,' and this tension has been central to much feminist debate on the subject.

6. *Feminist Debates on Prostitution*

Though sharing a common concern with the welfare of prostitute women, feminists are often deeply divided on the question of whether to view prostitution as a sexual or an economic institution. For 'radical' feminists, who foreground the sexual domination of women by men in their analyses of women's political subordination, prostitution is the unambiguous embodiment of male oppression. In this view, as a sexual institution, prostitution reduces women to bought objects, allows men temporary, but direct, control over the prostitute, and increases their existing social control over all women by affirming their masculinity and patriarchal rights of access to women's bodies. For such commentators, prostitution is always and necessarily damaging to women, there can be no distinction between 'forced' and 'free choice' prostitution, and in tolerating, regulating, or legalizing prostitution, states permit the repeated violation of human rights to dignity and sexual autonomy (Barry 1995, Jeffreys 1997).

Feminists who adopt a sex workers' rights perspective generally reject the idea that prostitution is intrinsically or essentially degrading, and view links between prostitution and patriarchal domination as contingent, rather than necessary. Treating prostitution as an economic institution, they make a strong distinction between 'free choice' prostitution by adults and all forms of forced and child prostitution. While they believe the latter should be outlawed, they hold the former to be a type of paid work, a job like any other. Since sex workers' rights feminists view free choice prostitution as a mutual voluntary exchange, they see state actions which criminalize or otherwise penalize those adults who make an individual choice to enter prostitution as a denial of human rights to self-determination (see Nagel 1997).

Because hierarchies exist within prostitution in terms of earnings, working conditions, vulnerability to violence and exploitation, and autonomy as regards specifying, limiting, and retracting from contracts with clients, these two very different interpretations of prostitution can each be partially supported empirically. Radical feminists point to research which indicates that around the world, the average age of entry into prostitution is very low (often under 18 years); that it is often precipitated by the experience of rape, sexual assault, and/or incest; and that the experience of prostitution can be extremely damaging psychologically, as well as associated with drug and solvent abuse, various forms of sexual and physical violence, and suicide (Hoigard and Finstad 1992, Kelly et al. 1995). Radical feminism's emphasis on a link between prostitution and systems of patriarchal domination is supported by the strong historical and contemporary association between the military and organized prostitution, while the persistence of highly abusive and slavery-like practices within prostitution and the significant presence of children aged between 12 and 18 in mainstream prostitution, even where prostitution is legally regulated rather than proscribed, lends further credence to a view of prostitution as a hugely problematic institution (Ennew 1986, Enloe 1989)

Sex workers' rights feminists point out that slavery-like practices and child labor persist in other industries, and that historically, such abuses have been most successfully addressed through collective struggle for workers' political representation, international conventions guaranteeing minimum rights to various groups of workers/persons, national employment laws outlawing abusive practices, and so on. They cite research which suggests that prohibitionist legislation increases, rather than reduces, prostitutes' vulnerability to violence, third party coercion, and extortion, and emphasize empirical studies in which adult women and men in prostitution describe themselves as perfectly content with their chosen occupation (Bindman 1997, Kempadoo and Doezema 1998).

Though certain feminist commentators remain rigidly wedded to a view of prostitution as either work or sexual violence, many theorists and activists (including some feminist abolitionists and sex workers' rights feminists) place a growing emphasis on the need to develop analyses of prostitution which can embrace its diversity and its particularity as both a sexual and an economic institution (see, for example, Shrage 1994, Chapkis 1997, O'Connell Davidson 1998).

See also: Feminist Theory: Radical Lesbian; Heterosexism and Homophobia; Male Dominance; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Rape and Sexual Coercion; Sexual Harassment: Social and Psychological Issues; Sexuality and Gender; Transsexuality, Transvestism, and Transgender

Bibliography

Aggleton P (ed.) 1999 *Men Who Sell Sex*. UCL, London

Barry K 1995 *The Prostitution of Sexuality*. New York University Press, New York

Bindman J 1997 *Redefining Prostitution as Sex Work on the International Agenda*. Anti-slavery International, London

Bishop R, Robinson L 1998 *Night Market: Sexual Cultures and the Thai Economic Miracle*. Routledge, London

Clift S, Carter S 1999 *Tourism and Sex: Culture, Commerce and Coercion*. Pinter, London

Chant S, McIlwaine C 1995 *Women of a Lesser Cost: Female Labour, Foreign Exchange & Philippine Development*. Pluto, London

Chapkis W 1997 *Live Sex Acts*. Cassell, London

Davis N (ed.) 1993 *Prostitution: An International Handbook on Trends, Problems and Policies*. Greenwood, Westport, CT

Enloe C 1989 *Bananas, Beaches, Bases*. Pandora, London

Ennew J 1986 *The Sexual Exploitation of Children*. Polity Press, Cambridge, UK

Hoigard C, Finstad L 1992 *Backstreets: Prostitution, Money and Love*. Polity, Cambridge, UK

Jeffreys S 1997 *The Idea of Prostitution*. Spinifex, Melbourne, Australia

Kelly L, Wingfield R, Burton S, Regan L 1995 *Splintered Lives: Sexual Exploitation of Children in the Context of Children's Rights and Child Protection*. Barnardo's, Ilford, UK

Kempadoo K (ed.) 1999 *Sun, Sex and Gold: Tourism and Sex Work in the Caribbean*. Rowman and Littlefield, Boulder, CO

Kempadoo K, Doezema J (eds.) 1998 *Global Sex Workers: Rights, Resistance and Redefinition*. Routledge, New York

Lichtenstein B 1999 Reframing 'Eve' in the AIDS era: The pursuit of legitimacy by New Zealand sex workers. *Sexuality and Culture*. **2**: 37–60

Lim L 1998 *The Sex Sector: The Economic and Social Bases of Prostitution in Southeast Asia*. International Labor Organization, Geneva, Switzerland

Matthews R 1997 *Prostitution in London: An Audit*. Middlesex University, London

Nagel J (ed.) 1997 *Whores and Other Feminists*. Routledge, London

O'Connell Davidson J 1998 *Prostitution, Power and Freedom*. Polity Press, Cambridge, UK

Shrage L 1994 *Moral Dilemmas of Feminism*. Routledge, London

Walkowitz J 1980 *Prostitution and Victorian Society: Women, Class and the State*. Cambridge University Press, Cambridge, UK

West D 1992 *Male Prostitution*. Duckworth, London

J. O'Connell Davidson

Protein Synthesis and Memory

A role for proteins in memory formation was invoked by early neuroscientists in their search for the 'engram', an envisioned physical manifestation of stored memory. It was recognized that measured electrical activities were too transient to endure for a

lifetime, and chemically stable macromolecules such as proteins seemed more likely candidates, perhaps as metabolically stable bricks that comprised the structures of altered synapses (Gerard 1953). It eventually became evident that proteins are constantly being produced and broken down, necessitating more complex models of the molecular basis for memory storage than a simple accretion of structural molecules.

The nature of protein synthesis in the brain became of particular interest for behavioral scientists as a result of demonstrations in a number of vertebrate and invertebrate species that chemical agents that inhibit protein synthesis in the brain block the formation of long-term memory (LTM) of a variety of training tasks, yet do not block acquisition of the task (Agranoff et al. 1999). More recently, analogous effects of protein synthesis inhibition have been reported in a reductionist model of learning and memory, termed long-term potentiation (LTP) (see *Long-term Potentiation (Hippocampus)*).

For the purposes of this article, learning and short-term memory (STM) formation are defined by experimental measures of improved performance of an adaptive response observed during or shortly after a training session (acquisition), while LTM formation is evidenced by retention of the learned response in a subsequent training session, typically administered some days or weeks later. An experimental advantage of using an interventive agent such as a protein synthesis blocker in studying learning and memory is at the same time its great disadvantage: while it can establish the timing of an inferred molecular mechanism taking place in an otherwise intact animal, it does not specify cells or even brain regions critical for LTM formation, nor does it identify specific proteins on which memory formation is presumed to be dependent. Such proteins could hypothetically be identified by complementary correlative studies, for example, by examining the nature of newly synthesized proteins on the basis of their incorporation of radiolabeled precursor amino acids during the period of LTM formation. Together, the interventive and correlative approaches have proven useful in establishing the molecular bases of other complex cellular processes.

1. Proteins Defined

Proteins are biological macromolecules consisting of a linear array of 20 component L-amino acids. The amino acids are joined in chemically stable peptide linkages that are not easily disrupted by thermal, chemical or mechanical stresses encountered by living cells. A protein typically consists of several hundred amino acids in a precise sequence (termed its primary structure), much like a very long word written in a 20-letter alphabet. The secondary and tertiary structures of a protein refer to how the polypeptide chain is folded into helices, ribbons and sheets. This overall conformation of proteins is largely dictated by the chemical properties of the individual amino acids in its primary structure.

Proteins play both structural and functional roles in cells. Membranes, consisting of lipid bilayers in which proteins are embedded, define a neuron and its intracellular organelles, such as the nucleus, mitochondria, synaptic vesicles, and endoplasmic reticulum. Proteins may be primarily structural, adding integrity to the membrane, or serving as a bridge to other membranes or to one of a group of fibrillar proteins including tubulin, actin and neurofilament proteins, known collectively as the cytoskeleton. Functional roles of proteins are exemplified by the enzymes, which are generally soluble proteins, present in a globular conformation. Hundreds of enzymes have been characterized and shown each to be a chemical catalyst that accelerates the rate of a specific biochemical reaction, without itself being consumed. Structural and functional proteins together catalyze and otherwise mediate the business of the cell (Voet and Voet 1995), expressing its genetic phenotype, including its appearance, its metabolism and its repertoire of responses. Knowledge of the primary structure of a protein is useful in determining its relation to other known proteins, in some instances yielding information about its evolutionary lineage as well as providing insight into its function. Short amino acid sequences within proteins (domains) identify probable catalytic active sites, the number of times a protein chain spans a membrane and the probable site of a protein's binding to other known proteins. There are also short amino acid sequences that identify the nature and likely sites of 'decoration' of proteins with small chemical groups following their synthesis. In addition to acting as catalysts, proteins can mediate recognition of a variety of small molecules that are intracellular chemical messengers. Membrane proteins on the cell surface may also act as receptors that recognize a variety of small molecules, including extracellular neuromodulators, hormones, neurotransmitters and environmental chemical signals (odorants) (see *Neurotransmitters*; Siegel et al. 1999). Proteins can recognize other proteins, even xenobiotic (novel) ones, as in the immune response. While proteins are clearly the arbiters of all cellular activities, not all cellular actions depend on ongoing protein synthesis, as will be seen.

1.1 Proteins are Synthesized on Ribosomes from Messenger RNA

The primary structure of a protein ultimately is determined genetically by linearly coded triplet deoxynucleotides in DNA in the cell nucleus, which are

transcribed to a complementary sequence of nucleotide triplets in the messenger RNA (mRNA) for the protein. The newly synthesized mRNA is *translated* ('decoded') into the new protein on polyribosomes in a cellular region near the nucleus, the RNA triplets establishing the primary structure of the new protein. The entire process, beginning with the availability of DNA template for transcription and ending with protein synthesis, generally takes minutes to hours (Voet and Voet 1995). Antibiotic inhibitors of protein synthesis such as puromycin and acetoxycycloheximide are structurally unrelated, and block translation by different molecular mechanisms, yet each has been shown to block LTM formation, confirming that their behavioral effects are indeed the result of their inhibition of protein synthesis rather than of some unknown side effects of the agents. Inhibition of RNA synthesis has also been found to block LTM formation.

Upon release from the polysomes, newly synthesized proteins may be subjected to *post-translational modification*, typically the addition of stable, small chemical substitutents, such as fatty acid, isoprenoid, carbohydrate, and phosphate or methyl groups. These modifications are catalyzed by specific enzymes, and may also require cofactors such as calcium ions or cosubstrates such as ATP. Phosphorylation of a protein by ATP requires a protein kinase, although some proteins can autophosphorylate themselves. As a consequence of such modifications, a protein may be tagged with a chemical 'zip code' that designates its intracellular destination. For example, a lipid substituent will target a protein to a membrane site, while phosphorylation could direct a protein to a nuclear site. The overall conformation of proteins can be affected by these post-translational changes. A newly formed protein may be inactive until it is cleaved by a specific protein-splitting enzyme that removes a built-in inhibitory sequence.

Many proteins are part of intracellular messenger systems that act sequentially to produce multiplicative cascades, resulting ultimately in a cellular response, such as hyperpolarization, depolarization, contraction, or secretion. Unlike the ribosomal synthesis of proteins, in which the translation of proteins is measured in tens of minutes, post-translational modification can occur rapidly, even in milliseconds, and unlike typical protein synthesis, is not restricted to the perinuclear region of the cell. Post-translational modification of proteins can occur at any time in the life of a protein, and can even target proteins for degradation.

1.2 Brain Proteins are in a Dynamic State of Formation and Degradation

The thousands of protein species present in a brain cell are in a metabolically dynamic state, i.e., they are continuously being synthesized by ribosomes and are then broken down into their component amino acids by the action of enzymes called proteases. The released amino acids may be metabolized or can be reincorporated into new proteins. Turnover studies employing radiolabeled amino acid precursors show that a given cellular protein in a specific cell type is turning over with a characteristic half-life of minutes to weeks. Each protein is present at a characteristic steady state level, i.e., the rate of synthesis of the protein is equal to its rate of degradation. Even though most neurons are formed early in life and remain in a postmitotic state through adulthood, their component structures are nevertheless in a dynamic metabolic equilibrium at a molecular level. A neuron may thus maintain an unchanging microscopic appearance throughout life, yet, with the exception of DNA, every component molecule, including all of its neuronal proteins, is constantly being replaced. There is considerable evidence that the cell's morphology and other cellular modifications of long duration are maintained by continual feedback to nuclear DNA, signaling initiation of transcription of mRNAs for constitutive proteins as they are depleted, or the induction of novel proteins, as occurs in cellular differentiation during development and in nerve regeneration. Thus, changes in synaptic connections might be similarly maintained by such transcriptional feedback loops in LTM storage. Translational and post-translational feedback loops that mediate memory formation have been proposed (Lisman and Fallon 1999).

1.3 Temporal Considerations in Establishing the Role of Protein Synthesis in Memory Formation

Since acquisition of a new behavior can occur within seconds, it is immediately apparent that it cannot be either directly or indirectly mediated by the relatively slow process of *de novo* protein synthesis. Acquisition could, however, involve a sustained post-translational modification at the synapse, which can occur within milliseconds and at any cellular locus. As noted, the macromolecular process of protein synthesis requires initiation of transcription in the nucleus followed by ribosomal translation, and takes minutes to hours, a time frame compatible with LTM, but not with STM formation. Additional time may be required for a putative chemical message from the presumed site of STM formation in the synaptic region (see *Memory: Synaptic Mechanisms*) to be transported to the nucleus in the neuronal cell body and to initiate transcription and translation. Still more time may be required for newly synthesized proteins to be transported from the cell body to a synaptic destination, where they could mediate long-term neuroplastic changes.

Rates of protein flow away from the cell body (anterograde axonal transport) and from the dendrites and axons toward the cell body (retrograde axonal

transport) were established in neurons with long processes. The amount of time required for trafficking of a chemical signal from a synapse to the nucleus and of a newly synthesized protein back to the synapse hypothetically could be extrapolated from known rates of axonal transport, together with knowledge of the morphology of the neuroplastic neurons hypothesized to be mediating the process. Ribosomal protein synthesis at a distance from the nucleus (Steward 1997) could shorten the calculated transport times. As will be addressed in the following section, the advent of antibiotic agents that selectively block protein synthesis offered the opportunity to study these complex issues experimentally.

2. *Agents that Inhibit Protein Synthesis Block Formation of Long-term Memory*

In the early 1960's, the excitement over the identification of the genetic code spurred renewed speculation and experimentation directed at the possible molecular nature of the engram. In 1962 Flexner and co-workers reported that intracerebral injection of the antibiotic puromycin, an aminonucleoside that prematurely cleaves the growing peptide chain from ribosomes, blocked brain protein synthesis in the mouse brain and could also block formation of memory of a Y-maze shock avoidance task (Flexner et al. 1962). This was followed by studies in the goldfish in 1964 (reviewed in Agranoff et al. 1999), and subsequently in a variety of other vertebrates and invertebrates.

2.1 *Effects of Protein Synthesis Blockers on LTM Formation in Experimental Animals*

Injection of puromycin into the cranial fluid (intracranially; IC) surrounding the goldfish brain produced a profound block in brain protein synthesis for several hours. Studies on acquisition (STM) and on LTM formation of a light-coupled-to-shock shuttlebox task revealed that a block in brain protein synthesis had no effect on acquisition of avoidance behavior during an initial multitrial training session, but eliminated retention of the improved performance that would otherwise have been seen in a second session administered a week later. The depth and duration of the inhibition of brain protein synthesis and the effectiveness of the block of LTM formation correlated well. Injection of the blocker into the fish body cavity rather than IC was without effect on LTM formation. IC injection of acetoxycycloheximide, another inhibitor of protein synthesis, was equally effective in blocking LTM formation.

The behavioral effects of inhibition of protein synthesis differ from that of a number of neurodisruptive agents and treatments, such as electroconvulsive shock, in that animals exhibit normal behavior during the action of the interventive agent. This permitted an exploration of STM decay. It was known that fish injected IC with puromycin before the first training session would show increasing avoidance scores, comparable to those seen in uninjected control fish; yet in a second training session a week later, such fish performed at a naïve level, i.e., had formed no LTM of the first session. Groups of fish trained under conditions of blocked protein synthesis (and that nevertheless demonstrated normal acquisition, as noted above) were then retrained at various times following the initial session. It was found that performance gradually dropped to the level of naïve fish over a 2–3 day period, providing evidence for the decay of STM for this task. Block of protein synthesis immediately after training also blocked LTM formation, whereas if animals were returned to their home tanks and injected with puromycin a few hours later, there was no detectable effect on LTM. This period of LTM susceptibility was comparable in its duration to that seen with electroconvulsive shock and other disruptive agents administered after training. When fish were detained in the training apparatus following the first training session, the consolidation time was prolonged, suggesting that the onset of LTM formation required an 'environmental trigger.' This phenomenon, in which continued exposure of an experimental subject to its training environment delays the onset of LTM formation, has also been termed 'detention.'

Behavioral studies with numerous other species, including young chicks and rats (for review, see Davis and Squire 1984), support the general hypothesis that LTM formation requires ongoing protein synthesis, while STM formation does not. Among vertebrates, studies in Hermessenda and in Aplysia (Bailey et al. 1992) support this hypothesis (see Agranoff et al. 1999). Drosophila have proven useful in behavioral studies, with genetic mutants in which selected proteins are mutated, deleted or overexpressed, implicating various neurotransmitter-related and second messenger system proteins in memory formation (see Agranoff et al. 1999). Effects of protein blockers on human memory have not been reported, but the possibility did not escape the imagination of a novelist (MacDonald 1968).

2.2 *Protein Synthesis Inhibition in Long-term Potentiation, a Reductionist Model of Learning and Memory*

Attempts to identify specific proteins, the synthesis of which are crucial to LTM formation in the brain, have been beset with many experimental problems, not the least of which is knowledge of the precise anatomical sites of the presumed altered molecular events underlying an observed altered behavioral performance

interpreted as LTM formation. LTP (see *Long-term Potentiation (Hippocampus)*) can be induced and studied *ex vivo* in localized tissue preparations such as brain hippocampal slices, and has been proposed as a cellular model of LTM formation. Blockers of protein synthesis have been shown to prevent LTP formation, a result taken by some as its validation as a model of LTM formation. Numerous pharmacological, biochemical, molecular, and genetic studies have been performed to implicate one or another structural protein, enzyme, receptor, or intracellular messenger pathway in the development of LTP, as a result of which over 100 known molecules have been implicated in LTP formation, including intracellular messengers, vesicular trafficking proteins, proteases, kinases, etc. (Sanes and Lichtman 1999). Eventually it may prove to be the case that the ongoing synthesis of one or only a few of these proteins is indeed necessary and sufficient for LTP formation, and can account for the observed block of LTM formation by protein synthesis inhibitors. It would seem more likely that the synthesis of many such proteins is necessary for the complex process of LTM formation to proceed. This embarrassment of riches in candidate compounds continues to grow as additional previously undiscovered cellular proteins become implicated in neuroplasticity, in that their selective deletion or inactivation prevents LTP or LTM formation. Further information regarding the temporal sequence of the observed molecular events accompanying LTP may prove useful in establishing whether they are indeed causal or epiphenomenal in its formation.

2.3 *Protein Nutrition and Mental Function*

Half of the 20 amino acids of which proteins are composed can be synthesized in the human body, while the remaining 10, termed 'essential amino acids' (EAAs), must come from the diet, principally in the form of ingested proteins that are degraded to their component amino acids in the gut and absorbed into the bloodstream. Eventually some will cross the blood/brain barrier, which has efficient EAA uptake systems. Also, as the brain's own proteins are broken down, the released amino acids are conserved and become a rich source of EAAs for new brain protein. Thus, under conditions of dietary EAA deficiency in adults, as occurs in starvation, brain protein metabolism is relatively protected. This is not true for the developing brain in infancy and early childhood, periods during which the brain must grow. For this, the brain relies on an exogenous source of EAAs, without which permanent cognitive deficits can ensue, as in kwashiorkor, a protein deficiency disease of children encountered in impoverished tropical countries. A blood amino acid imbalance is seen in the genetic defect phenylketonuria in which there is a high blood level of the amino acid phenylalanine, while blood levels of the other EAAs are normal. This imbalance overloads and blocks the amino acid transport system at the blood/brain barrier and effectively prevents entry of the other EAAs, eventually resulting in mental retardation. If infants with this genetic defect are raised on a diet deficient in phenylalanine, the brain develops normally, and as they mature, the toxic effect of dietary phenylalanine is much reduced.

3. *Summary*

It is apparent from the foregoing that in a number of animal species, disruption of protein synthesis in the brain for a period of hours has no effect on acquisition of a behavioral task, but blocks the formation of LTM. Since all brain proteins turn over metabolically, it is presumed that positive feedback loops are induced during LTM consolidation that involve transcription and/or translation of new protein, but how and where these loops function and are distributed in the complex structure of the brain is not yet known. Selective inhibition or elimination of various proteins by genetic means has been shown to block memory formation, but the salience of each new finding is dampened by the increasing number of identified genetic defects or deletions that can be shown to lead to a defect in memory and learning. The synthesis of many proteins each may prove necessary (but not sufficient) for LTM formation.

STM formation, as defined in human and primate learning and memory studies, may be measured in fractions of seconds, with lifetimes that are compatible with the durations of electrophysiological measurements of neuronal activity. Other forms of STM formation, primarily in lower vertebrates and invertebrates, may be long-lived (seconds to hours or even days) as indicated in Sect. 2.1. Intermediate stages in STM formation have been proposed in studies on taste aversion in the chick (Gibbs and Ng 1979). It is possible that the longer lasting forms of STM reside in synaptic regions and are mediated by post-translational feedback loops between proteins, and thus would not be affected by the protein synthesis blockers, as has been observed. It must be remembered that the various molecular, pharmacological, behavioral, and electrophysiological interpretations of memory formation are not mutually exclusive and in many instances may reflect observation of the same neurobiological phenomenon from dissimilar vantage points.

See also: Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Long-term Potentiation and Depression (Cortex); Long-term Potentiation (Hippocampus); Memory,

Consolidation of; Memory: Synaptic Mechanisms; Short-term Memory: Psychological and Neural Aspects; Working Memory, Neural Basis of

Bibliography

Agranoff B W, Cotman C W, Uhler M D 1999 Learning and memory. In: Siegel G J, Agranoff B W, Albers R W, Fisher S K, Uhler M D (eds.) *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, 6th edn. Lippincott-Williams & Wilkins, Philadelphia/New York, Chap. 50

Bailey C H, Montarolo P, Chen M, Kandel E R, Schacher S 1992 Inhibitors of protein and RNA synthesis block structural changes that accompany long-term heterosynaptic plasticity in Aplysia. *Neuron* **9**: 749–58

Davis H P, Squire L R 1984 Protein synthesis and memory: A review. *Psychological Bulletin* **96**: 518–59

Flexner J B, Flexner L B, Stellar E, de la Haba G, Roberts R B 1962 Inhibition of protein synthesis in brain and learning and memory following puromycin. *Journal of Neurochemistry* **9**: 595–605

Gerard R W 1953 What is memory? *Scientific American* **189**: 118–25

Gibbs M E, Ng K T 1979 Neuronal depolarization and the inhibition of short-term memory formation. *Physiological Behavior* **23**: 369–75

Lisman J E, Fallon J R 1999 What maintains memories? *Science* **283**: 339–40

MacDonald J D 1968 *The Girl in the Plain Brown Wrapper*. Fawcett, Greenwich, CT

Sanes J R, Lichtman J W 1999 Can molecules explain long-term potentiation? *Nature Neuroscience* **2**: 597–604

Siegel G J, Agranoff B W, Albers R W, Fisher S K, Uhler M D (eds.) 1999 *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, 6th edn. Lippincott-Williams & Wilkins, Philadelphia/New York

Steward O 1997 mRNA localization in neurons: A multipurpose mechanism? *Neuron* **18**: 9–12

Voet D, Voet J G 1995 *Biochemistry*, 2nd edn. Wiley, New York

B. W. Agranoff

Protestantism and Gender

At the beginning of the twenty-first century, the term 'Protestant' is used to designate any group of Christians that does not acknowledge direct loyalty and connection either to the Roman Catholic hierarchy, or to the many catholic churches rooted in Euro-Asian cultures known collectively as Eastern Orthodox Christianity. In contemporary discourse the term 'Protestant' has come to apply to a broad range of spiritual communities. While this discussion of Protestantism and gender focuses on older Protestant 'denominations,' it should be remembered that by the late nineteenth century, Protestantism had expanded and changed greatly. By then, the majority of Protestants were chiefly in movements whose historical development occurred outside of Europe, particularly in the Northern Hemisphere of the so-called 'New World.' Beginning in the nineteenth century, new forms spontaneously generated indigenous Christianities emerged in the Western Hemisphere, and by the early twentieth century, on the continents of Latin America, Africa, and Asia. The newer indigenous movements, identifiable as Christian by their emphasis on personal conversion, by the centrality of Jesus to their doctrine and practice, and by their fidelity to those texts broadly recognized as Christian scriptures (often called 'The Bible'), nevertheless bear little resemblance to the movements and cultures of Protestantism that had their roots in European history. Such Protestants now constitute the fastest growing edge of Christianity. It is too early to describe how such faith-communities will affect the future of gender relations, or reshape the historical directions of the original Protestantisms described here. And as will become clear, all Protestantisms, including these newer forms, are racked with conflict over gender issues, including the roles of women in churches and the nature of Christian attitudes toward sexuality. Today, strong forces within Christianity seek to reverse the clear directions toward gender equality and greater parity between men and women that have slowly shaped most Protestant belief and practice.

The origins of Protestantism are much debated. Deep resistance to the existing western Catholic sacramental system, loosely organized through the leadership of the bishops of Rome, appeared in late medieval Christianity. Much of this resistance presaged issues raised in the Protestant Reformation. However, Protestant beginnings are usually dated from the work of the German Reformer, Martin Luther, early in the sixteenth century. The four main trajectories that shaped the social formations of early Protestantism became differentiated and broadly identifiable 'traditions' of Protestant Christian belief and practice. The earliest, emerging under Luther's leadership, prospered chiefly in regions of Central Europe and quickly moved northward into Scandinavia. Almost simultaneously, a second major tradition emerged from Swiss Reform movements led by Zwingli, and was quickly joined by French-speaking groups under the leadership of John Calvin. These churches came to be known as 'Reformed.'

A third Continental stream of Protestantism emerged from the even more ardent and diverse dissenters within agrarian and peasant communities. These reforming groups were known in the first eras of Protestantism as Anabaptists. They were antihierarchical and egalitarian, and would eventually profoundly shape Protestant culture. However, initially the Anabaptist movements and communities generated hostility both from Lutheran and Reformed groups and from the Roman Catholic Church. The tensions generated by religious conflict on the Con-

tinent also reached the UK and also sparked a prolonged English Reformation struggle. The fourth Protestant tradition emerged from the settlement that created the Church of England and Anglicanism, which claimed to be a 'bridge' between Catholic and Protestant spirituality. It needs to be remembered that for three centuries Europe experienced a prolonged period of political, economic, and cultural turmoil as Christianity changed dramatically. This period is sometimes referred to as 'the Wars of Religion.'

By the late eighteenth century these realignments and newly-reordered ecclesial communities stabilized into clear cut 'denominations' with differences in doctrine and ritual practice that became matters of ongoing public dispute in cultures where Christianity predominated. Importantly, doctrinal differences were paralleled to differing normative views of how reforming Christian ecclesiastical culture and institutions should be related to political order and state power. Such differences are important to the topic of gender because Protestantist groups that were opposed to religious establishment and church support of state power would also become more egalitarian about issues of gender. But where the new Protestantisms became aligned with state power, they joined in suppressing dissent from their doctrine and practice as Roman Catholics frequently did during the Counter Reformation. As a result Anabaptist and Reformed Protestants migrated to seek refuge in any area of Europe or the New World of the Americas where they could find toleration or freedom from persecution.

Through these historical dynamics, Protestant reform seeded and reseeded itself, both in Europe and by the eighteenth century in North America. Migrations of many language groups also created hybrid Protestant churches that were Lutheran and Reformed. Lutheranism came to speak all the northern European languages, and 'Reformed Churches' from Switzerland, France, and the Netherlands dramatically flourished in the UK where the Church of England, or Anglicanism, had many rivals. Early European settlements in the Western Hemisphere often also brought multiple 'denominations' of the same Reformation traditions to North America. Dutch Reformed groups and Scottish and British Reformed communities created several separate denominations. Later European colonialism carried the four European-generated forms of Protestantism to areas of Africa, Asia and to parts of the sections of the Western Hemisphere previously colonized by Roman Catholic nations.

Reformed and 'Anabaptists,' sometimes called left-wing Protestants, in particular migrated in search of religious freedom and to escape state religious coercion. These traditions especially prospered in North America. Their emphasis on 'believers' communal equality' in the task of overseeing church life, in doctrinal discernment, and in scriptural interpretation, fit well with the democratizing tendencies of immigrant life. As a result, North American Protestantism took on a dynamic life of its own that quickly produced new 'schisms of reform' that rebounded back to Europe. For example, Episcopalians, Anglicans in the USA, quickly generated a major division led by John Wesley that became known as 'Methodism.' Anabaptist or 'free church' movements that especially emphasized the priesthood of all believers and local autonomy of believing congregations came forth in great profusion, often spontaneously. Protestant denominationalism came to have indescribable diversity. Many such groups came to be known as Baptists. Baptists included forms of 'slave Christianity' that enabled African-Americans in slavery to generate a religious culture independent of the slave owners' 'white Christianity' imposed on them and designed to keep them in their place. Methodism and white Baptist practice spread in rural and frontier communities throughout the northern hemisphere. Even in the twenty-first century, the two largest Protestant churches in North America are Black and white Baptist denominations. Cultural historians now recognize that gender was never a theorized social reality until the modern period of western cultural history. This means that an accurate assessment of Protestantism's impact on gender cannot be uncovered through exercises in intellectual history alone. More recent methods of historical-material historiography enable us to see that the major impacts on gender relations within Protestantism were actually the result of changes in the religious practice initiated by the new Protestant churches. Slowly these changes produced concrete redirections in gender relations and, more slowly, reshaped how Protestants theorized gender.

Not surprisingly, early Protestantism continued what may be called 'traditionalist' Christian notions of gender. After the high Middle Ages, Catholic (now Roman Catholic) Christianity had made the moral theology of Thomas Aquinas its formal moral-theological standard. Thomistic thinking formalized Catholic moral teaching so that it came to presume two normative sources of moral knowledge: revelation and reason. Though what was meant by either revelation or reason differed in specific theological accounts, western Christians came to assume that both of these sources contributed to a proper knowledge of gender. What the resulting 'traditionalist Christian' view of gender presumed, roughly following Aristotelian biology, was that males and females differed by nature (science or reason) and divine will (revelation). Such thinking supported notions widely held in popular culture that made deep challenges to traditional gender relations difficult. Throughout Christian history women's roles varied and were frequently changed by resistance to existing traditionisms. Well before the modern advent of women's organized efforts at political change, forms of what today are called Christian feminism existed and should be read as expressions by women of refusal of domination. Contemporary critical historiography confirms these continuous chal-

lenges but given the strength of patriarchal control, the point of visible institutional strain was never reached.

Even so, the full relevance of the Reformation to gender change is frequently missed because only formal teaching on gender is noted. The alternatives in communal form, ritual, and teaching in the Protestant traditions under discussion here were simultaneously shifts both in ecclesial/social forms and in Christian communal life. It was these new social formations and forms of ritualized communal practice that slowly changed gender relations in churches and in the broader culture.

The closing of the clergy and monastic-centered penitential systems, including the break with clerical celibacy brought the Christian parish, or in left-wing parlance, the Christian congregation, into the center of Protestant Christian faith practice. As a result, the primary units of Protestant spirituality soon became the local community (parish or congregation) and especially the family. In particular the patterns used to socialize persons as Protestant Christians dramatically altered gender relations. The Protestant family was theorized as an arena where both men and women, albeit initially in varied ways, were called to be agents of the formation of Christian communal life. Men were chiefly responsible for the public and the most authoritative ecclesial roles, but men and women alike were teachers of the faith, leaders of prayer, and witnesses to divine presence.

Protestantism particularly flourished in tandem with the invention of the printing press, which made Luther's, Calvin's and the Anabaptists' formal emphasis on scripture as the central theological source for Christians an actualizable reality. In all Protestant communities the locus of prayer, ritual practice, and theoretical (i.e. theological) formation shifted to the congregation and the family. None of the new Christian duties articulated by the Reformers were considered more important than that of transmitting the new 'knowledge of faith,' whether as scriptural literacy, or as newer catechism (doctrine). While none of this resulted in immediate transformations in female social roles because of the deeply gendered nature of the sex role systems, the new practices increased women's direct participation in matters of faith discernment. They also situated women within the central sacral systems of Protestant life. Prior to the Reformation, in spite of women's creativity and resistance, a lack of formal clerical power and women's remoteness from male-monopolized institutionalized hierarchies of ecclesial power prevailed. Pre-reformation Christian women certainly had created powerful gender alternatives, in particular monastic institutions for women. These enabled some women ('nuns') to live in an alternative social space where intellectual, artistic, and spiritual gifts could be cultivated. The impact of such efforts was severely marginalized structurally, however, and could be easily manipulated by male hierarchies. Historians have also recognized that some initial Reformational changes had negative impacts on women's roles and status. For example, Luther's closing of nunneries and his and Calvin's insistence on clerical marriage, especially when gender dislocations led to arranged marriages between priests and nuns, robbed women of crucial existing social space. However, the need for alternatives also led the Protestant movements to create new gendered sub-cultures for women, usually in the form of Deaconess Orders or catechetical offices that enabled women to work with women and children in initiating them into 'the fullness of faith.'

The trajectory of changes in gender practice and, with the rise of modernity, gender theory, within these Protestant movements cannot be traced in detail. However, by the late nineteenth and early twentieth centuries, when gender began to become a contested political issue and calls for gender justice emerged widely in public discourse, calls for justice for women were voiced in all but the most 'traditionalist' Protestant Christian communities. It is erroneous, as some 'secular' historians as well as conservative Christians have done, to imagine that demands for change in gender roles occurred primarily outside of, and apart from, the life and practice of Protestant communities. Christian feminisms fed public political feminism and Christian anti-feminist resistance testifies to the militant presence of feminist agitation within Protestantism throughout the modern period. Feminism is perceived as secular in origins only because contemporary historians tend to discount the role of religion in positive social change and because feminist discourse within Protestantism was initially less confrontative than later feminist political rhetoric. But in Protestant cultures by the late nineteenth century, women's energies were powerful within the churches and pressure for change escalated simultaneously with the rising public efforts of women for political change. It is not too much to say that by the end of the nineteenth century Protestant churches were thoroughly feminized and women had gained considerable financial and organizational power. Within Christianity, antifeminism should be read as a symptom of increased pressure for change from women. A review of historical gains for women make clear that change came fastest in nations where Protestantism was strong but where the churches did not aspire to state-enforced establishment.

Even as many male church leaders reasserted Christian gender traditionalism against the claims of Christian feminisms, gender parity slowly increased in Protestant communities, and on-going feminist contentions did not disappear. On the contrary, by the twentieth century, contentions over gender exploded into nearly all Christian communities. Because feminist pressures were more muted in Anglican and Lutheran traditions until well into the century, and because controversies over ordination when they did

erupt were especially strident there, public awareness of church gender debates became widespread only at that historical moment. However, nearly all nations where Protestantism has had a clear religio-cultural role were the first to adopt explicit legal and moral norms of equality between men and women both within churches and in the wider society. Furthermore, because the congregation and the family have remained the cultic center of Protestant life and practice, active patriarchical misogyny is rarely overtly expressed within it. Protestant pastoral practice is invariably more egalitarian than official theologies would suggest so women are more likely to be pedestilized than criticized in Protestant rhetoric.

The ways in which clergy-lay ritual practice and power were realigned and the patterns by which Protestants groups related to state political power are also crucial to the pace of gender change. Where Protestant denominations replicated Roman Catholic efforts at church-state establishment as in the Church of England, in sectors of Lutheranism and in some Reformed churches such as those in early New England, USA, or predemocratic South Africa, Protestant leaders sought to reassure Christian rulers that the churches would officially conform to 'traditional values' and to the stability of social order. Efforts in support of keeping the Christian family 'traditional' were frequent among male leaders, though much more common in established churches than elsewhere. Gender change came much more rapidly where Reformed churches broke with theories of the divine right of kings and endorsed democratic political theory and where the 'free churches,' actively refusing religious establishment, sought to actualize a theology of universal priesthood. In fact, it was these free-church Protestants who demanded, on theological grounds, that religious toleration become the public stance of the just state. Left-wing Protestants militantly insisted that church-state separation is a matter of constitutional or political 'right' and in the USA they were successful.

Today it is often these Protestant movements, the first to ordain women and the strongest critics of identification with state power, who have become the most conservative on gender. This is because in the late twentieth century the locus of Protestant traditionalism in relation to gender change shifted dramatically. Until recently, theological progressivism had slowly carried the day in the older Protestantisms under discussion here. Over time, newer forms of knowledge prevailed and the historical-critical or hermeneutic sciences prevailed. The main dynamic of Protestant theological development was towards recognition that the 'two' sources of religious and moral knowledge—revelation and reason—supported the central Protestant historical/theological emphasis on 'continuing reformation.' Such 'theological liberalism,' as conservatives called it, was comprehended by most Protestants as the heart of their faith-claims. Even when Protestant theological polemics included charges that gender justice was 'heresy' the trend was toward equality. In recent debates, however, the peculiar Protestant theological reaction of Biblicism has begun to alter older historical directions that embraced historical-critical readings of scripture.

In the late nineteenth century Protestant reaction to change began to take the form of what is called, somewhat misleadingly, 'biblical literalism.' So-called literalists do not seek to replicate ancient sex role patterns, but they project current traditionalist views to suggest that the Bible does not support gender justice. Such change, they contend, is impious because women's 'place' is forever set by what 'the Bible' teaches. This refusal of all truth except that found in 'the Bible' produced what in modern parlance is known as 'Fundamentalism.' Fundamentalism is a reactionary historical dynamic intrinsic to Protestantism as a text-based religion. Today many claim that other major religious movements are also becoming 'fundamentalist,' but such charges should be approached cautiously. Fundamentalism was generated by the particularities of Protestant inter-religious conflict. The refusal of gender justice within Protestant Christianity has parallels in other contemporary religions that accounts for the glib extension of the term. However, such usage is seriously misleading as to what is occurring within other religion. Even when reactionary Protestants and the Papacy attempt to enlist conservatives within Judaism, Islam, and even the great Asian religions against gender justice movements, Christian fundamentalist justifications for resisting such justice are not readily generalizable to other groups. Political and economic dynamics profoundly different than those occurring in older Protestant cultures are taking place elsewhere. For example, many former progressive left-wing Protestants in North America, are today for the first time, strongly aligned with global capitalist political power, whereas Islamic religious reaction is often aligned with anti-colonialism and resistance to capitalist 'modernizing' dynamics. Failure to notice such disparities may pit justice-seeking activists within the religions against each other.

In spite of the rising tide of fundamentalist anti-feminist attack within Protestant cultures and communities the struggles for justice continue in all of them. However, today even some moderate Protestants portray debates on gender justice as 'trivial.' Arguments about women's ordination, Christian sexual ethics, including Christian positions on gay, lesbian and transsexual lifestyles, and debates about reproductive choice and abortion are not understood, as earlier Protestant thought understood them, as aspects of the Protestant theological struggle towards genuine moral discernment and responsibility in a complex world. Most Protestants continue to espouse 'continuous and on-going reformation,' as the proper theological norm of Christian ecclesial and communal

life but the new fundamentalism has been chillingly successful in obscuring the connection between progressive Protestant theological principles and these particular issues. This is why in spite of the historical directions outlined here the question of whether continued movement toward equality in male and female relations among Protestants remain salient. Progressive moral change is always difficult to sustain both within religious communities and in the wider society. The fundamentalism now aligned with conservative political and economic power enjoys considerable success in part because of better political mobilization on the right. Even reputable media-spokespersons and scholars, who surely should know better, occasionally mislead in their interpretation of these contemporary controversies. For example, until the last decades of the twentieth century most Protestant churches from the traditions discussed here, emphatically embraced family planning and took no role in efforts to criminalize abortion. In ongoing debates legal abortion came to be embraced on various theological and moral grounds by major Protestant churches. However, the Christian Right has now succeeded in efforts to construct an 'unbroken' Christian anti-abortion tradition out of a pastiche of traditionalist Biblical pronatalism, Roman Catholic teaching, and post-sonogram technical imagery. To be sure, a few Protestant moral-theologians have been strongly anti-abortion on Biblicist grounds, but until recent efforts to 'reconstruct' Protestant history, it is not accurate to portray Protestants as anything other than committed to family planning and as moderately supportive of safe and legal abortion. Today, no reputable theological spokesperson would criticize family planning or embrace Roman papal doctrines that sex is chiefly for procreation, but fewer and fewer Protestant leaders recall the active debates that led to the embrace of reproductive choice for women in their communities. As a result the slow but dramatic march of Protestantism toward inclusion of women in 'the priesthood of all believers' may be reversed and the historical vocation of protestant theological reform rendered unrecognizable.

See also: Accidents, Normal; Buddhism and Gender; Catholicism and Gender; Christianity: Evangelical, Revivalist, and Pentecostal; Christianity: Liberal; Christianity Origins: Primitive and 'Western' History; European Religion; Feminist Theology; Fundamentalism (New Christian Right); Islam and Gender; Judaism and Gender; Reformation and Confessionalization; Religion and Gender; Religion, History of; Religiosity: Modern; Women's Religiosity

Bibliography

Connery J R 1977 *Abortion, the Development of the Roman Catholic Perspective*. Loyola University Press, Chicago

Coward H, Maguire D C (eds.) 2000 *Visions of a New Earth: Religious Perspectives on Population, Consumption, and Ecology*. State University of New York Press, Albany, NY

Dombrowski D A, Deltete R 2000 *A Brief, Liberal, Catholic Defense of Abortion*. University of Illinois Press, Urbana, IL

Feldman D M 1998 *Birth Control in Jewish Law: Marital Relations, Contraception, and Abortion as Set Forth in the Classic Texts of Jewish Law*. J. Aronson, Northvale, NJ

Harrison B W 1983 *Our Right to Choose: Toward a New Ethic of Abortion*. Beacon Press, Boston

Harrison B W 1985 In: Robb C S (ed.) *Making the Connections: Essays in Feminist Social Ethics*. Beacon Press, Boston

Harrison B W (ed.) 1995 Rhetorics, rituals, and conflicts over women's reproductive power. *Journal of Feminist Studies in Religion. Scholars Press* **11**(2) Special issue

Gilson A B 1999 *The Battle for America's Families: a Feminist Response to the Religious Right*. Pilgrim Press, Cleveland, OH

Gordon L 1990 *Woman's Body, Woman's Right: Birth Control in America*. Penguin Books, New York

Kertzer D I 1993 *Sacrificed for Honor: Italian Infant Abandonment and the Politics of Reproductive Control*. Beacon Press, Boston

Macdonald L O 2000 *A Unique and Glorious Mission: Women and Presbyterianism in Scotland 1830 to 1930*. J. Donald Publishers, Edinburgh, UK

Maguire D C (ed.) 2001 *Sacred Choices: The Right to Contraception and Abortion in Ten World Religions*. Fortess Press, Minneapolis, MN

Petchesky R P 1990 *Abortion and Woman's Choice: The State, Sexuality, and Reproductive Freedom*. Northeastern University Press, Boston

Steffen L H 1994 *Life/Choice: The Theory of Just Abortion*. Pilgrim Press, Cleveland, OH

Taub N 1994 *International Conference on Population and Development*. American Society of International Law, Washington, DC

Walker W, Norris R A, Lotz D W, Handy R T 1986 *A History of the Christian Church*, 4th edn. T. & T. Clark, Edinburgh, UK

B. W. Harrison

Protocol Analysis in Psychology

Protocol analysis is a rigorous methodology for eliciting verbal reports of thought sequences as a valid source of data on thinking. When psychology emerged as a scientific discipline at of the end of the nineteenth century, the majority of psychologists were interested in consciousness. They sought to examine the structure and elements of their thoughts and subjective experiences through introspective analysis. Within a few decades, the introspective method was discredited, and introspective descriptions were rejected as scientific evidence. Psychology turned away from studies of individuals' descriptions of their thoughts and first-hand experiences, and focused on observations of how other adults performed tasks in the laboratory. The rejection of introspection made the study of thinking come to a virtual halt until the 1950s, when the

technical innovations such as the computer led to the emergence of cognitive and information processing theories of psychological phenomena. In the new research approach to the study of thought processes, subjects were asked to 'think aloud,' leading to a new type verbal report of thinking that differed from the earlier introspective methods and became the core method of protocol analysis. The next section describes the fundamental problems of introspection and general solutions offered by protocol analysis.

1. *Problems with Introspective Descriptions of Thinking and Proposed General Solutions*

Adults describe their subjective experience, such as dreams, memories, and mental slips, on a regular basis, so why are observations of one's own experiences and thinking problematic as psychological data? The fundamental problem with descriptions of inner experience is that they are private and known only to the person who experienced them. There are no methods to verify the accuracy of these experiences independently. Consequently, protocol analysis is restricted to situations where the scientist has other types of information about the participant's thinking that can be used to validate the verbal reports.

The other major objection to introspection is that the mere act of observation of one's thoughts and experiences changes them. For example, when someone examines a single visual image appearing during thinking to search for specific associated details, this analysis tends to make the desired details emerge and thus changes the content and structure of the original image. This extended analysis also interrupts the normal flow of visual images, so it is impossible just to resume the thought process where it was interrupted. To minimize this type of interference in protocol analysis, research participants are instructed to think aloud while problem solving so they can concentrate on the experimental task, and let verbalization of thought be secondary.

2. *Historical Background*

Self-observation of thinking has a long history, which helps explain the controversies surrounding introspection, as well as the development of less controversial methods for eliciting verbalizations of thinking. The first systematic attempts to understand the structure of thinking by observing one's own spontaneous thoughts have been traced back at least as far as the Greeks. Based on observation of his thoughts, Aristotle argued that thinking corresponds to a sequence of thoughts (see Fig. 1), where the brief transition periods between consecutive thoughts (see the arrows in Fig. 1) do not contain any reportable information.

Aristotle's general account of thinking as a thought sequence has never been seriously challenged (Ericsson and Crutcher 1991). The controversies about introspection have all involved attempts to go beyond the immediately observable aspects. For example, Aristotle tried to infer the unobservable processes that control the thought sequence (see the arrows in Fig. 2). By examining his memory of which thoughts tended to follow each other, Aristotle concluded that previously experienced associations between thoughts were the primary determining factor. Philosophers in the seventeenth, eighteenth and nineteenth centuries (Ericsson and Crutcher 1991) asked questions that required more involved inferences and introspective analyses. For example, can every thought be described as an amalgam of memories? In order to evaluate these claims empirically, philosophers would typically relax and daydream, allowing their thoughts to wander. Once in a while they would try to freeze an image and inspect it carefully to assess its sensory modality and components. Based on such detailed observations, the famous British philosopher David Hume came to the surprising conclusion that thoughts were as detailed as original perceptions—we cannot have a visual image a printed page of text without imaging every letter on the page at the same time.

By the nineteenth century, philosophers generally rejected Hume's claim and agreed that the process of analyzing a visual image changes its original appearance. For example, when a person has an image of a rose and then decides to inspect its petals, the whole mental image changes. Consequently, the process of examining mental images is fundamentally different from observing solid objects in the real world. A real rose can be looked at repeatedly from different directions without being changed, even its petals can be inspected with a looking glass without any change. In contrast, extended introspective analysis of a mental image changes its content and structure and cannot therefore be used to study thoughts as they emerge during spontaneous thinking.

In the beginning of the twentieth century, psychologists at the University of Würzburg tried to minimize this form of reactivity by training introspective observers. These observers were given thought-provoking questions and asked to respond as fast as possible. Only after their answers had been given would the observers start to recall the thoughts that they had had while answering the questions. The observers were trying to recall the most basic elements of their thoughts and visual images. Most reported thoughts consisted of visual and auditory images, but some subjects claimed to have experienced thoughts without any corresponding imagery–imageless thoughts. The alleged existence of imageless thoughts had far-reaching theoretical implications and led to a heated exchange between the psychologists who claimed to have observed them and others who argued that these reports were artifacts of inappropriate

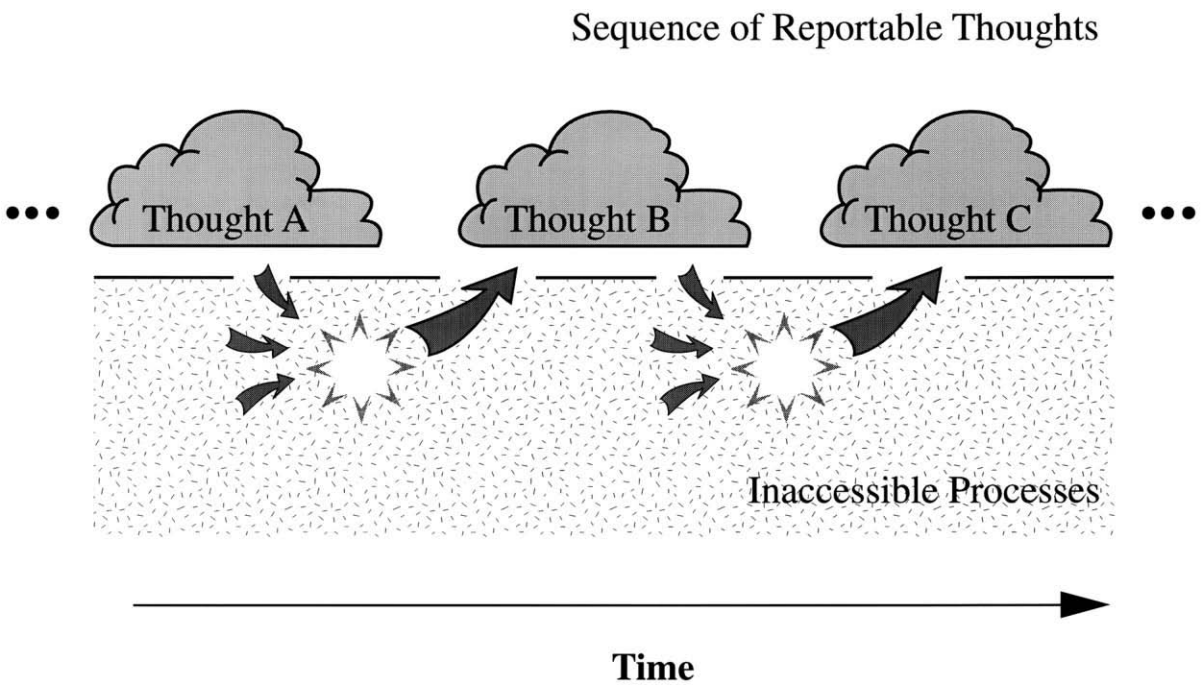

Figure 1
An illustration of Aristotle's view of thinking as a sequence of states with reportable thoughts, while the processing responsible for transitions between consecutive thoughts remains inaccessible

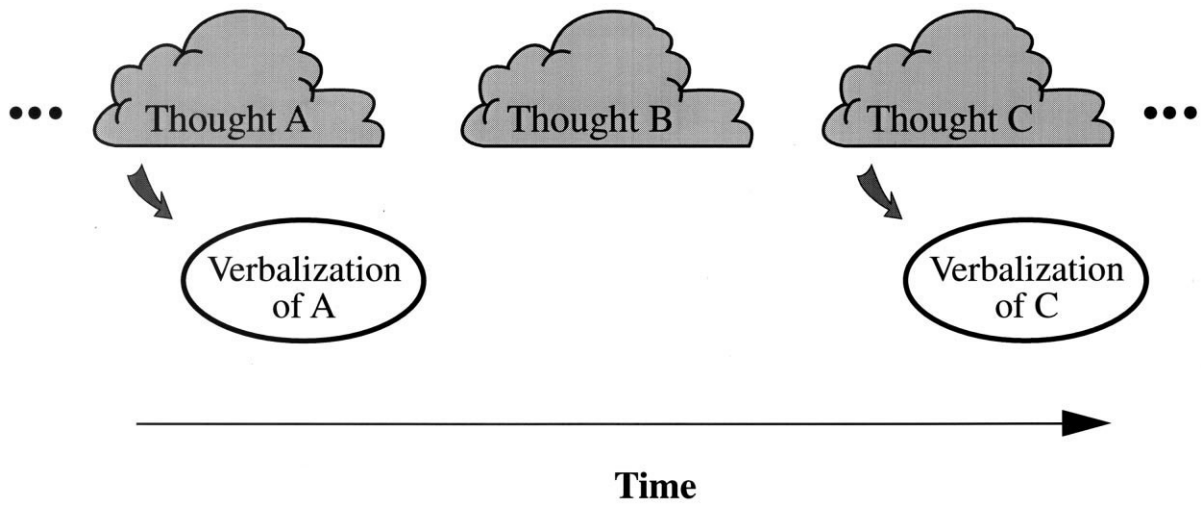

Figure 2
An illustration of a covert sequence of thoughts and the associated overt verbalizations of most thoughts passing through attention during the performance of a task

reporting methods and theoretical bias of the observers. A devastating conclusion arose from this controversy: the existence of imageless thoughts could not be resolved empirically. This finding raised fundamental doubts about analytic introspection as a scientific method.

Reacting to the methodological and theoretical problems of the introspective method, psychologists redirected their research from conscious experience and thinking and towards observable behavior, and focused on basic learning processes shared by adults, children, and animals. For example, some of the tasks involved learning the path from a starting location to a goal location in an arbitrary maze, where the benefits of prior experience and knowledge would be minimal. Similar learning tasks required participants to memorize 'meaningless' materials, such as nonsense syllables like XOK and ZUT, where the memory performance was measured objectively by scoring the sequence of recalled syllables for accuracy. The hypothesis that memory for nonsense syllables was based only on basic associations, and was later rejected in studies by cognitive psychologists showing that during memorization subjects reported thinking of words that were similar to the studied nonsense syllable, such as 'cozy' for KOZ and 'payment' for PYM (Prytulak 1971). When subjects were able to associate the nonsense syllable with a word, the speed and accuracy of memorization improved (Montague 1972), which provides validation for the reported thoughts.

The cognitive revolution in the 1960s renewed interest in higher-level cognitive processes and how thinking allows individuals to generate solutions to novel tasks. Cognitive theories describe how individuals are able to apply acquired knowledge and procedures to novel problems, such as mental multiplication of any combination of two two-digit numbers. Information processing theories (Newell and Simon 1972) proposed computational models that could reproduce the observable aspects of human performance on well-defined tasks through the application of explicit procedures.

One of the principle methods of the information processing approach is task analysis. Task analysis specifies the range of alternative procedures that people could use, in light of their prior knowledge of facts and procedures, to generate correct answers to a task. Let me illustrate how task analysis can be applied to mental multiplication. Most adults have only limited mathematical knowledge: they know their multiplication table and only the standard 'pencil and paper' procedure taught in school for solving multiplication problems. Accordingly, one can predict that they will solve a specific problem such as 36×24 by first calculating $4 \times 36 = 144$ then adding $20 \times 36 = 720$. More sophisticated subjects may recognize that 24×36 is equivalent to $(30+6) \times (30-6)$ and use the formula $(a+b) \times (a-b) = a^2 - b^2$, thus calculating 36×24 as $30^2 - 6^2 = 900 - 36 = 864$. The choice of alternative procedures participants use to generate the answer can be inferred by the time needed and verbal reports of their thoughts during problem solving.

In conclusion, the theoretical and methodological controversies about verbal reports have never cast doubt people's ability to recall part of their thought sequences. The controversies have centered around efforts to go beyond the sequence of thoughts (see Fig. 1), to analyze their detailed structure through introspection, and to infer the processes controlling the generation of new thoughts. In fact, all major theoretical frameworks concerned with thinking have advocated the use of verbally reported sequences of thoughts (Ericsson and Crutcher 1991). For example, the behaviorist John B. Watson pioneered the use of 'think aloud,' and the gestalt psychologist Karl Duncker established it as a major method.

3. *Protocol-analysis Methodology*

The central assumption of protocol analysis is that it is possible to instruct subjects to verbalize their thoughts in a manner that doesn't alter the sequence of thoughts mediating the completion of a task, and can therefore be accepted as valid data on thinking.

3.1 *Elicitation of Nonreactive Verbal Reports of Thinking*

Based on their theoretical analysis, Ericsson and Simon (1993) argued that the closest connection between thinking and verbal reports is found when subjects verbalize thoughts generated during task completion (see Fig. 2). When subjects are asked to think aloud, some of their verbalizations seem to correspond to merely vocalizing 'inner speech,' which would otherwise have remained inaudible. Nonverbal thoughts can also be often given verbal expression by brief labels and referents. For example, when one subject was asked to think aloud while mentally multiplying 36 by 24 on two test occasions one week apart, the following protocols were obtained:

(a) OK, 36 times 24, um, 4 times 6 is 24, 4, carry the 2, 4 times 3 is 12, 14, 144, 0, 2 times 6 is 12, 2, carry the 1, 2 times 3 is 6, 7, 720, 720, 144 plus 720, so it would be 4, 6, 864.

(b) 36 times 24, 4, carry the—no wait, 4, carry the 2, 14, 144, 0, 36 times 2 is, 12, 6, 72, 720 plus 144, 4, uh, uh, 6, 8, uh, 864.

In these two examples, the reported thoughts are not introspectively analyzed into their perceptual or imagery components, but merely verbally expressed and referenced, such as 'carry the 1,' '36,' and '144 plus 720.' Similarly, subjects are not asked to describe or explain how they solve these problems. Instead, they are asked to remain focused on solving the problem and merely to give verbal expression to those thoughts that emerge in attention while generating the solution under normal (silent) conditions.

If the act of verbalizing subjects' thought processes does not change the sequence of thoughts, then subjects' task performance should not change as a result of thinking aloud. In a comprehensive review of

dozens of studies, Ericsson and Simon (1993) found no evidence that the sequences of thoughts (accuracy of performance) were changed when subjects thought aloud as they completed the tasks, compared with subjects who completed the same tasks silently. However, some studies showed that think-aloud subjects would take somewhat longer to complete the tasks—presumably due to the additional time required to produce the overt verbalization of the thoughts.

The same theoretical framework can also explain why other types of verbal report procedures consistently change cognitive processes. For example, when subjects are instructed to explain or carefully describe their thoughts, they are not able merely to verbalize each thought as it emerges, they must engage in additional cognitive processes to generate the thoughts corresponding to the required explanations and descriptions. This additional cognitive activity changes the sequence of mediating thoughts. Instructions to explain and describe the content of thought are reliably associated with changes in ability to solve problems correctly (Ericsson and Simon 1993).

In sum, after brief training in giving verbal reports, subjects can think aloud without any systematic changes to their thought process (see Ericsson and Simon 1993 for detailed instructions and associated warm-up tasks recommended for laboratory research). This means that subjects must already possess the necessary skills for verbalization of thoughts.

3.2 Validity of Verbalized Information from Think Aloud

When adults are able to perform tasks while thinking aloud without sacrificing accuracy and speed, the verbalized information would almost have to reflect some aspect of the regular cognitive processes. By analyzing the information expressed as verbalized thoughts, it is possible to assess the validity of the verbalized information. In protocol analysis, the verbalized thoughts are compared with intermediate results generated by different strategies that are specified in a task analysis (Ericsson and Simon 1993). The sequence of thoughts verbalized while multiplying 24 by 36 mentally (reproduced above) agrees with the sequence of intermediate thoughts specified by one, and only one, of the possible strategies for calculating the answer.

Even when subjects think aloud with its close connection between thoughts and reports, correspondence between verbalized thoughts and intermediate products predicted from the task analysis is not perfect (see Fig. 2). The lack of one-to-one correspondence is due primarily to the fact that not all thoughts which pass through attention are verbalized and some processing steps (thoughts) may be short-circuited with acquired skill. However, there is persuasive evidence of validity for the thoughts that are verbalized (Ericsson and Simon 1993). Even if a highly skilled participant's think-aloud report in the multiplication task only consisted of '144' and '720,' the reported information would still be sufficient to reject many alternative strategies, because these strategies did not involve generating both of the reported intermediate products. The general finding that a task analysis can identify *a priori* the specific intermediate products that are later verbalized by subjects during their problem solutions, provides the strongest evidence that concurrent verbalization reflects the processes that mediate the actual generation of the correct answer.

More generally, verbal reports are only one indicator of the thought processes that occur during problem solving. Other indicators include reaction times (RTs), error rates, patterns of brain activation, and sequences of eye fixations. Given that each kind of empirical indicator can be separately recorded and analyzed, it is possible to compare the results of such independent data analyses. In their review, Ericsson and Simon (1993) found that longer RTs were associated with verbal reports of a larger number of intermediate thoughts than those corresponding to shorter RTs. Furthermore, there seemed to be close correspondence between subjects' thoughts and what information that they looked at—when subjects verbalized thoughts about objects in the environment they very frequently looked at them.

Finally, the validity of verbally reported thought sequences depends on the time interval between the occurrence of a thought and its verbal report, where the highest validity is observed for concurrent, think-aloud verbalizations. For tasks with relatively short response latencies (less than 5–10 s), subjects are able to recall their sequences of thoughts accurately immediately after the completion of the task, and the validity of this type of retrospective reports remains very high. However, for cognitive processes of longer duration, the problems of accurate recall of prior thoughts increases, with a corresponding decrease in validity of the verbal reports.

3.3 Explanations of Validity Problems with Other Types of Verbal Reports

Protocol analysis specifies the constrained conditions necessary for valid, nonreactive verbalizations of thinking. Many of the problems with verbally reported information obtained by other methods can be explained as violations of protocol analysis methodology. The first problem arises when the investigators ask subjects to give more information than is contained by the their recalled thought sequences. For example, some investigators ask subjects why they responded in a certain manner. Sometimes subjects will have deliberated and their recalled thoughts will

provide a sufficient answer, but typically the subjects need to go beyond any retrievable memory of their processes to give an answer. As subjects can only access the end products of their cognitive processes during perception and memory retrieval, they cannot report why only one of several logically possible thoughts entered their attention, so they must confabulate to answer such questions. In support of this argument, Nisbett and Wilson (1977) found that subjects' responses to 'why' questions were in many circumstances as inaccurate as those given by other subjects who merely observed and tried to explain another subject's behavior.

Second, scientists are interested primarily in the general strategies and methods subjects use to solve a broad class of problems in a domain, such as mathematics or text comprehension. They often ask subjects to describe their general methods after solving a long series of different tasks, which often leads to misleading summaries or after the fact reconstructions of what subjects think they must have done. In the rare cases when subjects have deliberately and consistently applied a single general strategy to solving the problems, they can answer such requests easily by recalling their thought sequence from any of the completed tasks. However, subjects typically employ multiple strategies, and their strategy choices may change during the course of an experimental session. Under such circumstances subjects would have great difficulty describing a single strategy that they used consistently throughout the experiment, thus their reports of such a strategy would be poorly related to their averaged performance. Hence, general strategy descriptions are usually inadequate, even when reports after each trial are valid descriptions of thought sequences.

Similar problems have been encountered in interviews of experts (Hoffman 1992). When experts are asked to describe their general methods in professional activities, they have difficulties, and there is often poor correspondence between the described methods and their observed behavior. However, when experts perform challenging representative problems from their domain of expertise while thinking aloud, their verbalizations consistently reveal a valid and informative account of their thought sequences during complex thinking and planning (Ericsson and Lehmann 1996). In sum, to obtain the most valid and complete trace of thought processes, scientists should strive to elicit those verbalizations that most directly reflect the subjects' spontaneous thoughts generated while they perform tasks that are representative of the studied phenomenon.

4. *Range of Application*

Protocol analysis has emerged as one of the principal methods for studying thinking in cognitive psychology (Crutcher 1994), cognitive science (Simon and Kaplan 1989), and behavior analysis (Austin and Delaney 1998). As further evidence of its validity, protocol analysis now plays a central role in applied settings, such as in the design of surveys and interviews (Sudman et al. 1996) and user testing of computer software (Henderson et al. 1995). Finally, several interesting adaptations of verbal-report methodology are emerging in the study of text comprehension (Pressley and Afflerbach 1995) and education (Renkl 1997).

See also: Behavior Analysis, Applied; Cognitive Science: Overview; Critical Thinking, Cognitive Psychology of; Education: Skill Training; Educational Assessment: Major Developments; Ergonomics, Cognitive Psychology of; Introspection: History of the Concept; Job Design and Evaluation: Organizational Aspects; Language and Thought: The Modern Whorfian Hypothesis; Organizational Behavior, Psychology of

Bibliography

Austin J, Delaney P F 1998 Protocol analysis as a tool for behavior analysis. *Analysis of Verbal Behavior* **15**: 41–56

Crutcher R J 1994 Telling what we know: The use of verbal report methodologies in psychological research. *Psychological Science* **5**: 241–44

Ericsson K A, Crutcher R J 1991 Introspection and verbal reports on cognitive processes-two approaches to the study of thinking: A response to Howe. *New Ideas in Psychology* **9**: 57–71

Ericsson K A, Lehmann A C 1996 Expert and exceptional performance: Evidence on maximal adaptations on task constraints. *Annual Review of Psychology* **47**: 273–305

Ericsson K A, Simon H A 1993 *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA

Henderson R D, Smith M C, Podd J, Varela-Alvarez H 1995 A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics* **38**: 2030–44

Hoffman R R (ed.) 1992 *The Psychology of Expertise: Cognitive Research and Empirical AI*. Springer-Verlag, New York

Montague W E 1972 Elaborative strategies in verbal learning and memory. In: Bower G (ed.) *The Psychology of Learning and Motivation*. Academic Press, New York, Vol. 6, pp. 225–302

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Nisbett R E, Wilson T D 1977 Telling more than we can know: Verbal reports on mental processes. *Psychological Review* **84**: 231–59

Pressley M, Afflerbach P 1995 *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading*. L Erlbaum Associates, Hillsdale, NJ

Prytulak L S 1971 Natural language mediation. *Cognitive Psychology* **2**: 1–56

Renkl A 1997 Learning from worked-out examples: A study on individual differences. *Cognitive Science* **21**: 1–29

Simon H A, Kaplan C A 1989 Foundations of cognitive science. In: Posner M I (ed.) *Foundations of Cognitive Science*. MIT Press, Cambridge, MA, pp. 1–47

Sudman S, Bradburn N M, Schwarz N (eds.) 1996 *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco, CA

K. A. Ericsson

Psychiatric Assessment: Negative Symptoms

Negative symptoms comprise, among others, flattening of affect, reduction of speech, and social withdrawal. They occur in different psychiatric diseases, but especially in schizophrenia, where they are traditionally seen as an important part of the symptomatology. In the last decade, they have evoked a special research interest due to the fact that novel/atypical antipsychotics have been developed with an improved therapeutic efficacy in negative symptoms of schizophrenia.

1. *Definition of Negative Symptoms*

Negative symptoms in schizophrenia comprise broadly of blunted affect, apathy, attentional impairment, poverty of thought or speech, incoherence, loosening of association, thought blocking, and social withdrawal. However, there is no unanimous agreement on the concept of negative symptoms in schizophrenia: some authors adhere to a broad interpretation (Meltzer and Zureick 1989, Rösler and Hengesch 1990), while others prefer a narrow approach (Crow 1985, Strauss et al. 1974, Wing 1989). The controversy centers on incoherence, loosening of association, and thought-blocking, since these are also attributes of positive symptoms. This disparity in the selection of items also manifests itself in the outcome of patient testing with different rating scales. Different authors use different concepts and definitions for the type and number of symptoms implicated (Rösler and Hengesch 1990).

These problems of definition can also be identified in the rating scales used for measuring negative symptoms.

2. *History of the Term 'Negative Symptoms'*

The origins of the positive/negative distinction are usually attributed to Jackson (1887), the British neurologist who drew the distinction in order to delineate primary from secondary neurological phenomena. Accordingly, positive symptoms are defined as the presence of a behavioral function that is otherwise not normally present in an individual. Negative symptoms are defined as the absence of a behavioral function that is usually present in the normal individual. Negative symptoms all appear to be characterized by the loss of an important cognitive or behavioral function, while positive symptoms represent an exaggeration or distortion of a normal function.

3. *Standardized Assessment and Dimensional Differentiation of Negative Symptoms*

To improve the reliability of diagnostic assessment of negative symptoms, several scales have been developed, mostly in the last two decades (SANS: 1981, PANSS: 1987, SAPS: 1991).

Standardization of the assessment of negative symptoms through rating scales had a positive effect on reliability. For example, the interclass correlations for the Scale for the Assessment of Negative Symptoms (SANS) and the Scale for the Assessment of Positive Symptoms (SAPS) complex global ratings range between 0.70 and 0.88. Principally, the negative symptoms can be rated with as much reliability as positive symptoms.

Rating scales applied in the past decade include, among others, the anergia score of the Brief Psychiatric Rating Scale (BPRS), the Manchester Scale and Crow's modification of this scale, the negative score of the Positive and Negative Scale (PANSS), the SANS, and Abrams and Taylor's Rating Scale for Emotional Blunting. Nowadays, the SANS and the PANSS are the most frequently used rating scales.

In nearly all studies in which a broad spectrum of negative symptoms was investigated by a factor analytical approach, two or three dimensions of negative symptoms were found. This emphasizes the inadequacy of a unidimensional concept of negative symptoms in the complex phenomenology of this domain (Meltzer and Zureick 1989).

4. *Symptomatic Overlap with Parkinsonism and Depression*

The differentiation between depression and negative symptoms is difficult on the basis of both clinical phenomenology and rating scales, and therapeutic effects on depressive symptoms might be misinterpreted as successful treatment of negative symptoms and vice versa. Gaebel (1989) reported a correlation of $r = 0.34$–0.57 at different measurement points for the BPRS scores depression and anergia. In a follow-up study on schizophrenic patients, the correlation between the scores for depression and retardation/apathy in the Inpatient Multidimensional Psychiatric Scale (IMPS) amounted to $r = 0.35$ (Möller and von Zerssen 1986).

Another problem in the interpretation of rating scale data from studies in this field is that the negative symptoms score correlates positively with the drug-

induced Parkinson syndrome. Furthermore, there is a correlation between negative symptoms and hyperkinetic neurological symptoms such as akathisia and dyskinesia. One could therefore conclude that drugs without extrapyramidal side effects in general may appear to be more advantageous in treating negative symptoms.

5. *Negative Symptoms in Schizophrenia*

Both Kraepelin and Bleuler described a core of schizophrenic symptoms corresponding to the modern concept of negative symptoms (Table 1). Later, this position was gradually displaced by an emphasis on productive features as being pathognomonic of schizophrenia, in fact, as being a prerequisite for its diagnosis. This is exemplified in the diagnostic criteria for schizophrenia put forward by Schneider 1959, Feighner et al. (1972), in the research diagnostic criteria (Spitzer et al. 1978), and in the *Diagnostic and Statistical Manual DSM-III*, *DSM-III-R*, and *DSM-IV* criteria.

The need to incorporate both negative and positive features into a model of schizophrenia was brought out by the factor analytic studies of the 1960s and in particular by the research of Strauss et al. (1974), who adapted Jackson's (1887) terminology for classifying different profiles in schizophrenia.

Whereas Crow's (1980b) model viewed the positive and negative symptoms as annulated facets of schizophrenia that may reflect separate pathological processes, Andreasen (1982) proposed instead a typological model. They described positive and negative symptoms as opposing features that characterize different subtypes.

Which symptoms of schizophrenia are considered positive or negative varies across different authors. Nearly all authors include flat affect and poverty of speech among the negative symptoms, and hallucinations and delusions among the positive symptoms. Most regard anhedonia, apathy, avolition, or abolia as negative symptoms. Thought disorder, bizarre behavior, and inappropriate affect, however, are variably classified as positive or negative.

6. *Negative Symptoms in other Psychiatric/Mental Disorders*

Anergia, anhedonia, poverty of speech, and affective flattening, which occur as part of the negative symptoms in schizophrenia, were also observed in patients severely ill with depressive disorder. Although negative symptoms of schizophrenia may be different from depressive symptoms in depression, they do overlap, thus making negative symptom complexes and psychomotoric retardation, lack of energy, anhedonia, and other symptoms of depression significantly associated. On the other hand, it was observed that a distinct hypokinetic syndrome involving anhedonia, depressed mood, and cognitive impairment existed not only in negative schizophrenia and retarded depression, but also in Parkinson's disease, suggesting that reduced dopamine turnover might play a role in the pathophysiology of each.

7. *Primary and Secondary Negative Symptoms*

The differentiation between primary and secondary negative symptoms, as proposed by Carpenter et al. (1985), seems of great theoretical and clinical relevance. While primary negative symptoms are the consequence of a morbogenic deficit state, secondary negative symptoms are the consequence of:

(a) positive symptoms (e.g., social withdrawal because of paranoid ideas);

(b) extrapyramidal effects (e.g., motor retardation as an indicator of neuroleptogenic akinesia);

(c) depressive symptoms (e.g., postpsychotic or pharmacogenic depression); or

(d) social understimulation (e.g., hospitalization).

Table 1
Relationship between the concept of basic symptoms and the concept of negative symptoms (Klosterkötter 1990)

Kraepelin (1904)	Bleuler (1911)	Wing and Brown (1970)	Strauss et al. (1974)
Cognitive decline ('Stultification,' 'Dementia')	Association disorder	Schizophrenic thought disorders Slowness of thought Poverty of speech	Certain kinds of formal thought disorders, such as blocking
Flatness of affect	Affective disorder Autism	Flatness of affect	Blunting of affect
Loss of intentional power	Ambivalence Abulia Lack of initiative	Social withdrawal Low motivation	
Loss of energy	Lack of interest	Lack of initiative Underactivity	Apathy

According to this distinction, the treatment of negative symptoms has to be differentiated concerning the different causes.

Angst et al. (1989) underlined the necessity of discriminating negative symptoms according to different stages in the course of schizophrenia, i.e., the prepsychotic period, the productive psychotic period (characterized by a mixture of positive and negative symptoms, whereby negative symptoms may be hidden at this stage), and the postpsychotic period (characterized by residuum or defect with or without positive symptoms, postpsychotic endogenous, or reactive depression, postremissive exhaustion).

Crow (1980a) postulated the stability of a specific subtype of schizophrenia with positive symptoms (type I) and a subtype with negative symptoms (type II), explaining this by different biological causes. Most other authors stressed the point that negative and positive symptoms can be combined cross-sectionally or longitudinally and cannot be taken as indicators for a specific biological or phenomenological subtype of schizophrenia.

8. *Relationship between Positive and Negative Symptoms in Schizophrenia*

The positive/negative distinction may represent different manifestations of the common underlying process. It may also suggest the existence of different pathophysiologies in schizophrenia. Hints about the inter-relationships of these processes may be gleaned from the intercorrelations among their final common expressions as symptomatic phenomenology. An initial effort by Andreasen (1982) showed that the positive and negative symptoms clusters were significantly correlated within clusters and inversely correlated between clusters, suggesting a bipolar construct with positive and negative symptoms at the opposite ends of a continuum. This was replicated by another group. However, later investigations showed little or no relationship between positive and negative symptoms, strongly suggesting that they are independent processes within schizophrenia.

Since both positive and negative symptoms are frequent in schizophrenia, they are likely not to be completely independent (Möller et al. 1995). However, some findings have been controversial. For example, Rosen et al. (1984) failed to find a positive/negative symptom correlation.

9. *Course of Negative Symptoms in Schizophrenia*

Findings on the course of positive and negative symptoms over time have aroused considerable controversy. In acute and/or first episode cases, negative symptoms are sparse and linked tightly to the presence of other manifestations of psychoses, especially to positive symptoms (Möller and von Zerssen 1986). In the North American Risperidone Study (Marder and Meibach 1994), in which risperidone was compared to haloperidol, both drugs improved positive and negative symptoms significantly. Similar results were obtained in clinical trials on other novel/atypical neuroleptics such as olanzapine, sertindole, amisulpride, etc. Clearer results may be obtained if more sophisticated designs are used for psychopharmacological trials on negative symptoms. A European collaboration group therefore made some suggestions (Möller et al. 1994), focusing on patients with predominantly negative symptoms (Table 2).

During the acute phase, negative symptoms can often be unstable. For example, McCreadie (1989) studied schizophrenic patients suffering a first episode at three points in time (admission, one- and two-year follow-up). The percentages of patients with positive and negative symptoms were 96 and 23 percent at admission, 22 and 40 percent at the one-year follow-up, and 21 and 25 percent at the two-year follow-up, respectively. Other studies with acute patients have shown that negative symptoms are more stable over time. Ragin et al. (1989) found an increase in the level of poverty of speech between index admission and seven months after discharge. Wagman et al. (1987) found that the anergia score of the BPRS was significantly correlated across four months of repeated measures in patients with deficit symptoms, but not in patients without deficit symptoms.

Overall, negative symptoms appear to be more stable in subacute/subchronic patient samples. Pogue-Geile and Harrow (1985) recorded high and stable levels of anhedonia one and three years after episodes of schizophrenia. Biehl et al. (1986) found that negative symptoms had significant retest stability between years two and five in a sample from the International Pilot Study of Schizophrenia.

In a follow-up study of schizophrenia carried out in Munich (Möller and von Zerssen 1986), patients were rated using the IMPS at admission, discharge, and at a five-year follow-up. A decrease was observed in the mean score of positive symptoms and also in syndrome retardation as an indicator of negative symptoms. The correlation for the retardation score between discharge and follow-up was 0.36. In a similar 15 years follow-up study on first hospitalized patients with schizophrenia, a decrease of negative symptoms was found between admission and discharge, and an increase between discharge and follow-up (Möller et al. in press).

Many authors came to the conclusion that, in contrast to the views of Crow (1980b), the negative subtype and the positive subtype of schizophrenia are not stable conditions during the short- or long-term course of the disease. Most authors stressed the point that negative and positive symptoms can be combined cross-sectionally or longitudinally, and cannot be taken as indicators of a specific biological or phenomenological subtype.

Table 2
Suggestions of the working group on negative symptoms in schizophrenia (Möller et al. 1994)

Aim: evaluating treatment effects on negative symptoms in schizophrenia
Patient selection.
Positive symptoms not dominating the actual clinical picture, e.g. PANSS negative type
Duration of negative symptoms > 6 months
Stable condition of the schizophrenic illness > 6 months
Flat affect and poverty of speech as core symptom of negative symptoms
No/low depression score
Design: double-blind comparison to placebo or active drug
Efficacy parameters: BPRS or SANS or PANSS
Other scales: depression scales, EPS scales
Statistical analyses
End-point comparison
Interaction with productive symptoms, depression, EPS

BPRS: Brief Psychiatric Rating Scale; EPS: extrapyramidal symptoms; PANSS: Positive and Negative Syndrome Scale; SANS: Scale for the Assessment of Negative Symptoms

10. Biological Correlates of Negative Symptoms in Schizophrenia

The most prominent biochemical finding in schizophrenic patients with negative symptoms appears to be the reduction in central dopaminergic, serotonergic, and noradrenergic activity, and, if at all, a general common pattern can be described. This decrease in amine activity tends to be associated with structural brain abnormalities, i.e., cortical atrophy or enlarged ventricles. There are indications that typical neuroleptics reduce at least those negative symptoms of schizophrenia that are secondary to positive symptoms when the positive symptoms are effectively treated. However, negative symptoms of schizophrenia that occur independently of positive symptoms may also be reduced, especially through treatment with atypical antipsychotics or with antidepressants. The efficacy of drug treatment of negative symptoms seems to be related to their pharmacological profile, i.e., their interference with dopaminergic, noradrenergic, and serotonergic receptor systems. As to atypical antipsychotics, among others the 5-HT_{2A} blockade was hypothesized as being a relevant mechanism for the efficacy of atypical antipsychotics in treating negative symptoms.

As mentioned above, there is a considerable overlap in the symptomatology of psychiatric disorders with respect to negative symptoms. Also, different types of negative symptoms in schizophrenia have to be considered. In addition, the lack of homogeneity in the definition and assessment of negative symptoms in schizophrenia can be a problem. All these difficulties have consequences for research and explain some of the inconsistencies in biochemical and pharmacologic schizophrenia research (Meltzer and Zureick 1989). The results of biochemical investigations as well as of positron emission tomography and magnetic resonance imaging studies on negative symptoms, warrant further research efforts in this field to find, among others, a rationale of neuroleptic drug treatment in negative symptoms of schizophrenia.

11. Treatment of Negative Symptoms in Schizophrenia

There has been a strong debate in the literature on whether neuroleptics can really reduce negative symptoms, and especially whether they can reduce primary negative symptoms and not just secondary negative symptoms. An earlier review found evidence that some neuroleptics, especially some atypical neuroleptics, can reduce negative symptoms. However, since most of the trials were carried out in patients with negative and positive symptoms during acute exacerbations of schizophrenic psychoses, the question arises whether only those negative symptoms were reduced that were secondary to positive symptoms. In addition, the overlap between negative symptoms of schizophrenia and extrapyramidal side effects and depression has to be taken into account. As an alternative hypothesis to the direct effect of a drug on negative symptoms, the possibility that a drug reduces depression, akinesia, sedation, etc., and in this way also reduces negative symptoms, must be considered seriously.

One possibility of differentiating between direct and indirect drug effects on negative symptoms is to use statistical procedures. Formerly, only two studies had used an elementary statistical approach to account for the possibility of indirect effects on negative symptoms. Meltzer (1985) used a covariance analysis and found that the neuroleptic effect on negative symptoms could not be explained via a neuroleptic effect on positive symptoms alone. Van Kammen et al. (1987) used a regression analysis which showed that

the neuroleptic effect on negative symptoms could be explained by the effect on positive symptoms.

More complex statistical approaches were not used until the mid-1990s. Möller et al. (1995) applied a complex statistical model, a path analysis, to the data of the North American Risperidone Study (Marder and Meibach 1994), including effects on positive, negative, and extrapyramidal symptoms to determine whether risperidone had a more potent direct effect on negative symptoms than haloperidol. The aim was to determine whether a (direct) effect on negative symptoms remains after the effects which are possibly caused through the reduction of positive and extrapyramidal symptoms have been excluded. The estimate from the path model showed that even after statistical control for indirect effects on secondary negative symptoms, risperidone improved negative symptoms to a greater extent than haloperidol.

See also: Depression; Differential Diagnosis in Psychiatry; Mental and Behavioral Disorders, Diagnosis and Classification of; Nosology in Psychiatry; Schizophrenia; Schizophrenia: Neuroscience Perspective; Symptom Awareness and Interpretation

Bibliography

Andreasen N C 1982 Negative vs. positive schizophrenia. Definition and validation. *Archives of General Psychiatry* **39**: 789–794

Angst J, Stassen H H, Woggon B 1989 Effect of neuroleptics on positive and negative symptoms and the deficit state. *Psychopharmacology* **99**: S41–S46

Bandelow B, Muller P, Gaebel W, Kopcke W, Linden M, Muller-Spahn F, Pietzcker A, Reischies F M, Tegeler J 1990 Depressive syndromes in schizophrenic patients after discharge from hospital. *European Archives of Psychiatry and Clinical Neuroscience* **240**: 113–120

Biehl H, Maurer K, Schubart C, Krumm B, Jung E 1986 Prediction of outcome and utilization of medical services in a prospective study of first onset schizophrenics. Results of a prospective 5-year follow-up study. *European Archives of Psychiatry and Clinical Neuroscience* **236**: 139–147

Carpenter W T, Heinrichs D W, Alphs L D 1985 Treatment of negative symptoms. *Schizophrenia Bulletin* **11**: 440–452

Crow T J 1980a Molecular pathology of schizophrenia: more than one disease process? *British Medical Journal* **280**: 66–8

Crow T J 1980b Positive and negative schizophrenic symptoms and the role of dopamine. *British Journal of Psychiatry* **137**: 383–6

Crow T J 1985 The two-syndrome concept: origins and current status. *Schizophrenia Bulletin* **11**: 471–86

Feighner J P, Woodruff-R A, Winokur G, Robins E, Guze S B, Munoz R 1972 Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry* **26**: 57–63

Gaebel W 1989 Indikatoren und Prädiktoren schizophrener Krankheitsstadien und Verlaufsausgänge [Indicators and predictors of stages in schizophrenic illness and course outcomes]. Thesis, Free University of Berlin

Gaebel W 1990 Erfassung und Differenzierung schizophrener Minussymptomatik mit objektiven verhaltensanalytischen Methoden [Recording and differentiation of schizophrenic negative symptoms with objective behavior analytical methods]. In: Möller H J, Pelzer E (eds.) *Neuere Ansätze zur Diagnostik und Therapie schizophrener Minussymptomatik [Newer approaches to the diagnosis and treatment of schizophrenic minus symptoms]*. Springer, Berlin, Heidelberg, New York, pp. 79–92

Jackson J H 1887 Remarks on evolution and dissolution of the nervous system. *Journal of Mental Science* **23**: 25–48

Klosterkötter J 1990 Minussymptomatik and kognitive Basissymptome [Minus symptoms and basic cognitive symptoms]. In: Möller H J, Pelzer E (eds.) *Neuere Ansätze zur Diagnostik und Therapie schizophrener Minussymptomatik Newer approaches to the diagnosis and treatment of schizophrenic minus symptoms*. Springer, Berlin, Heidelberg, New York, pp. 15–24

Marder S R, Meibach R C 1994 Risperidone in the treatment of schizophrenia. *American Journal of Psychiatry* **151**: 825–35

McCreadie R G 1989 The Scottish first episode schizophrenia study. VII. Two-year follow-up. *Acta Psychiatrica Scandinavica* **80**: 597–602

Meltzer H Y 1985 Dopamine and negative symptoms in schizophrenia: Critique of the type I-II hypothesis. In: Alpert M (ed.) *Controversies in schizophrenia*. Guilford, New York, pp. 110–36

Meltzer H Y, Zureick J 1989 Negative symptoms in schizophrenia: A target for new drug development. In: Dahl S G, Gram L F (eds.) *Clinical pharmacology in psychiatry*. Springer-Verlag, Berlin, pp. 68–77

Möller H J, von Zerssen D 1986 *Der Verlauf schizophrener Psychosen unter den gegenwärtigen Behandlungsbedingungen [The Course of Schizophrenic Psychoses under the Current Treatment Conditions]*. Springer-Verlag, Berlin Heidelberg New York

Möller H J, van Praag H M, Aufdembrinke B, Bailey P, Barnes T R, Beck J, Bentsen H, Eich F X, Farrow L, Fleischhacker W W, Gerlach J, Grafford K, Hentschel B, Hertkorn A, Heylen S, Lecrubier Y, Leonhard J P, McKenna P, Maier W, Pedersen V, Rappard A, Rein W, Ryan J, Sloth Nielsen M, Stieglitz R D, Wegener G, Wilson J 1994 Negative symptoms in schizophrenia: Considerations for clinical trials. *Psychopharmacology* **115**: 221–8

Möller H J, Müller H, Borison R L, Schooler N R, Chouinard G 1995 A path-analytical approach to differentiate between direct and indirect drug effects on negative symptoms in schizophrenic patients. A re-evaluation of the North American risperidone study. *European Archives of Psychiatry and Clinical Neuroscience* **245**: 45–9

Möller H J, Bottlender R, Groß A, Hoff P, Wittmann J, Wegner U, Strauß A in press The Kraepelinian dichotomy: Preliminary results of a 15-year follow-up study on functional psychoses: Focus on negative symptoms. *Schizophrenia Research*

Pogue-Geile M F, Harrow M 1985 Negative symptoms in schizophrenia: Their longitudinal course and prognostic importance. *Schizophrenia Bulletin* **11**: 427–39

Ragin A B, Pogue-Geile M, Oltmanns T F 1989 Poverty of speech in schizophrenia and depression during in-patient and post-hospital periods. *British Journal of Psychiatry* **154**: 52–7

Rosen W G, Mohs R C, Johns C A, Small N S, Kendler K S, Horvath T B, Davis K L 1984 Positive and negative symptoms in schizophrenia. *Psychiatry Research* **13**: 277–84

Rösler M, Hengesch G 1990 Negative Symptom in AMDP-System [Negative symptoms in the AMDP system]. In: Baum

U, Fähndrich E, Stieglitz R D, Woggon B (eds.) *Veränderungsmessung in Psychiatrie und klinischer Psychologie [Measurement of Changes in Psychiatry and Clinical Psychology]*. Profil, Munich, pp. 329–39
Schneider K 1959 *Clinical Psychopathology*. Gune & Stratton, New York
Spitzer R L, Endicott J E, Robins E 1978 *Research Diagnostic Criteria (RDC) for a Selected Group of Functional Disorders, 3rd edn*. New York State Psychiatric Institute, New York
Strauss J S, Carpenter Jr W T, Bartko J J 1974 The diagnosis and understanding of schizophrenia. Part III. Speculations on the processes that underlie schizophrenic symptoms and signs. *Schizophrenia Bulletin* **11**: 61–9
van Kammen D P, Hommer D W, Malas K L 1987 Effect of pimozide on positive and negative symptoms in schizophrenic patients: Are negative symptoms state dependent? *Neuropsychobiology* **18**: 113–17
Wagman A M I, Heinrichs D W, Carpenter Jr W T 1987 Deficit and nondeficit forms of schizophrenia: Neuropsychological evaluation. *Psychiatry Research* **22**: 319–30
Wing J K 1989 The concept of negative symptoms. *British Journal of Psychiatry* **155**: 10–14

H.-J. Möller

Psychiatric Care, Economics of

A discussion of the economics of mental health care cannot take place without considering the epidemiological characteristics of mental illness and its cost to society.

1. Epidemiology of Mental Illness

1.1 The Burden of Mental Illness

Mental disorders are among the most prevalent and disabling illnesses worldwide. Unipolar major depression is the leading cause of disability globally in 1990, both in high-income and in low- and middle-income countries. It has been estimated that depressive and anxiety disorders account for 20–30 percent of all primary care visits worldwide. Approximately 29.5 percent of the US population is estimated to experience a diagnosable mental or addictive disorder in a 12-month period, which is comparable to the 20 percent rate for cardiovascular disorders (Kessler et al. 1994). Prevalence estimates of schizophrenia from studies in low-income countries vary from 0.8 per thousand reported from rural China to 5.9 per thousand in Calcutta, India. There were about 23 million people with schizophrenia in the world in 1985, three quarters in the less developed countries. Severe and persistent mental illnesses (as defined by diagnosis, disability, and duration) affect about 4 percent of the population each year. These disorders are long lasting, produce significant level of disability, reduce an individual's ability to function in major social roles, and interfere with the creation and maintenance of social networks (Desjarlais et al. 1995).

These national and international statistics do not reflect the enormous toll of misery from mental disorders. Neuropsychiatric conditions have been ignored for a long time as they are absent from cause of death lists. However, disease priorities change dramatically as measurement of disease burden shifts from simple mortality indicators to indicators that incorporate disability. When disease-burden measurement includes time lived with disability, several of the neuropsychiatric disorders become a leading cause of disease burden worldwide. The World Bank estimates that neuropsychiatric diseases account for 6.8 percent of the Global Burden of Disease (GBD), measured in disability adjusted life years (DALYs), a toll greater than that exacted by tuberculosis, cancer, or heart disease. For adults aged 15–44 in developing economies, neuropsychiatric diseases account for 12 percent of the GBD. When intentional, self-inflicted injuries are added, the total constitutes 15.1 percent of the GBD for women and 16.1 percent for men. Because of the limited mortality consequences, this burden has been previously underestimated (World Bank 1993).

According to a 1999 estimate, neuropsychiatric conditions account for 10 percent of the burden of disease measured in DALYs in low and middle-income countries and 23 percent of DALYs in high-income countries. In comparison, cardiovascular diseases are responsible for 10 percent of DALYs in low and middle-income countries and 18 percent of DALYs in high-income countries. In high-income countries, one out of every four DALYs is lost to a neuropsychiatric condition. Of the ten leading causes of disease burden in young adults (in the 15–44 year age group) four are neuropsychiatric conditions. Alcohol dependence, bipolar disorder, and schizophrenia are among the leading causes of disease burden in this age group in 1998 (World Health Organization 1999).

1.2 The Social and Treatment Costs of Mental Illness

Society pays a high price for mental and addictive disorders—in both direct and indirect costs. Indirect costs (lost productivity, lost earnings due to illness or premature death, accidents, absenteeism) are twice the direct costs of care, adding to the actual economic impact of mental and addictive disorders. The total costs for the United States for 1990 for all mental disorders was $148 billion (Rice et al. 1990). The direct cost of care for all mental disorders was $67 billion, which represents 10 percent of the total $670 billion direct cost of all health care in the Unites States

in 1990 (Levitt et al. 1991). In comparison, the total cost for all cardiovascular system diseases in US for 1990 was $159 billion. More recent estimates show the full economic cost of alcohol and drug abuse in USA in 1992 to be $246 billion. The cost burden from depression is about the same as that from heart disease. In the experience of one Fortune 100 company, mental health was ranked number two in the company's 10 most costly diseases, topped only by ischemic heart disease in terms of direct cost. If the same 10 diseases were ranked according to the full burden of disease, indirect costs drove mental health to the number one position and ischemic heart disease to the number 10 position (NAPHS 1999).

Spending on treatment is concentrated on those people with the most disabling conditions. Although individuals with severe and persistent mental illness represent only 5–10 percent of all the individuals who have mental illness, they account for approximately 40 percent of the specialty care expenditures (Steinwachs et al. 1992). Another estimate shows that 30 percent of the spending on mental health and substance abuse care is accounted for by 5 percent of the users of care (Frank et al. 1994). The total cost of severe and persistent mental illness in 1990 was estimated to be $74 billion (accounted for by the dollar costs of treatment, lost productivity, shortened lives, and the costs incurred in the criminal justice and social services systems), with $44 billion of the total being indirect cost (Rice at al. 1990).

However large, the dollar cost of mental illnesses does not account for the enormous emotional cost and pain that these illnesses exert not only on the affected individual but also on their family and on society. Mental disorders are often chronic conditions that create substantial disability. It is estimated that approximately 20–25 percent of the single adult homeless population suffers from some form of severe and persistent mental illness (Jencks 1994). Mental illness is correlated with other costly social problems: unemployment, crime and violence (Steadman et al. 1998), child abuse and neglect (Kelleher et al. 1994), and auto accidents (Rice at al. 1990).

1.3 Utilization of Services

It is not clearly established what proportion of individuals with mental disorder need, or want, treatment. Unlike other physical illnesses, not everyone with diagnosable mental illness perceives a need for treatment, and not all who seek treatment have a diagnosable disorder. Thus the under treatment and the over treatment of mental health problems can coexist within the same system (Regier et al. 1998).

In the United States, of the 30 percent of the population that has a mental illness at some point during the year, only 17.3 percent get some treatment in the health care sector with an additional 7.4 percent getting their only treatment from a human service agency or a self-help group. Thus, only about 25 percent of those with a diagnosable condition get some form of treatment over a twelve-month period. Low rates of treatment characterize the most severely ill. Only 36 percent of the people diagnosed with bipolar disorder or major depression in a year are treated in any health sector, only 25 percent of those with substance abuse are treated, and 57 percent of the patients with schizophrenia are treated. At the same time 4.5 percent of individuals with no disorder, as assessed with a diagnostic instrument, receive mental health care (Kessler et al. 1997). According to Frank and McGuire (1999) nearly 38 percent of the users and 28 percent of all visits for mental health are not associated with a diagnosable disorder.

Two conclusions particularly relevant to the economic analysis of mental health care can be drawn from the above data on the epidemiology of mental illness and it's cost to the society. First, mental and addictive disorders are prevalent and disabling and associated with a variety of social costs. Second, the social costs of mental and addictive disorders are concentrated in the 4 percent of the population that experiences severe and persistent mental illness, making this subgroup of people with mental illness 'undesirable' to insurers, employers, and neighbors.

2. The Market for Mental Health Care Services

Economics is the science and art of the distribution and allocation of scarce resources. The allocation of resources and rationing of scarce goods and services are guided by the market economy prices. The appropriate price for a product is determined by the relationship between *demand* and *supply* for this product. In the context of the market economy, health care in general and mental health care in particular can be conceptualized as a product. The classic relationship between the demand and the supply can be applied to mental health care as well. However, mental health care does not conform to the traditional rules of a marketplace. Several factors account for this. The nature of mental illness, the use of treatment services, and the supply side of the market differ from the general health sector. Plus, there are a greater variety of organizations and professionals with different theoretical orientation that supply mental health services. These include the specialty mental health sector (psychiatrists, psychologists, psychiatric social workers, psychiatric nurses, and counselors), the general medical sector (primary care physicians), self-help groups, and the clergy. Despite efforts to quantify mental health services and make them more objective, there is greater clinical uncertainty compared to the rest of medical care. Many mental disorders affect a person's ability to make decisions in their own best interest. Markets for other goods normally consist of a straightforward transfer of goods between the provider and the consumer. However, in the health sector,

in addition to this arrangement, funds from patients often flow through an intermediate funding body (either the government or an insurance agency). Market economics in mental health care are strongly influenced by *externalities*, that is the value that society derives from the health of its individual members.

Health care markets suffer from some kind of *market failure* but the mental health care market is subject to more failure than most. The most common market failures affecting the health care sector are due to lack of incentive to supply 'public goods,' neglect of externalities, and absence of well informed consumers. Failures in the market can be explained also by *adverse selection* (persons more likely to use coverage choose better coverage) and *moral hazard* (those who are insured will tend to overuse insured services). Evidence suggests that the forces of adverse selection and moral hazard work are at work in mental health (Frank and McGuire 1999). The special effort that is needed to reduce moral hazard and adverse selection in mental health care markets is one of the reasons why mental health services are organized and paid for differently than other types of health care (for review see Sharfstein et al. 1992, 1995).

3. *Financing of Mental Health Services*

Health care in general and mental health care in particular can be financed by public or private sources, the funds can be managed by public or private entities. Funds can be raised through taxes, mandates, private health insurance, out-of-pocket payments, grant assistance, charitable contributions, and borrowing. Funds can be managed by government, for-profit or not-for-profit private organizations, or consumers. Funds are then used to purchase health services that can be publicly or privately provided. A number of issues related to both public and private sources of financing are predicated on government's goals (allocational, distributional, and economic) and the policies utilized to correct market failures and instabilities.

3.1 *Relationship Between Finance and Provision of Mental Health Care*

It is important to distinguish between financing and provision of services. There is no absolute connection between the way mental health care is paid for and the way it is delivered. Public finance does not imply public provision. It is possible for services to be publicly financed yet privately provided or vice versa. Three major forms of relationships exist between the providers of mental health care and those who finance it, reimbursement, contract, and integrated approaches. Under the reimbursement approach, providers receive retroactive payments for services that they have supplied. Reimbursement can be found in systems with multiple private and public insurers and multiple suppliers as in the United States. In low- and middle-income countries reimbursement is rarely combined with public finance with the exception of Chile. Contract involves an agreement between third-party payers (insurers) and health care providers. It is found in social insurance systems. Preferred provider organizations (PPOs) in the United States also use this form. In integrated health care systems the same agency controls the funding and the provision of health services. Medical providers are paid salaries and the budget is the main instrument for allocating resources. Integrated public systems are used in the Nordic countries and in developing countries as well. Health maintenance organizations (HMOs) in the United States are an example of an integrated private system. In countries with national health systems, the purchaser of care and the provider are commonly one and the same, namely, the government. In Europe there has been increasing interest in separating provider and financing functions in national health systems in order to promote competition.

3.2 *Inequities in the Financing of Mental Health Care Services*

A key characteristic of the mental health care service system is the inequity in financing mental health services compared to the coverage for other disorders. Resources available to mental health care have been eroding over the past decade (Buck and Umland 1997). Discriminatory limits on treatment of mental disorders characterize both the public and private sector. This discrimination against mental illness is based on arbitrary and often artificial distinctions between mental and physical illness.

3.3 *Insurance and Managed Care*

Health insurance is a system in which the prospective consumers of care make payment to a third party, which in the event of future illness will pay for some or all of the expenses incurred. Health insurance is a mixed source of finance, as it often draws contributions from both employers and employees, and sometimes from government. The existence of risk is the fundamental rationale for insurance. Insurance reduces risk by improving the predictability of adverse events through the pooling of a large number of similar risks. The few who become ill are covered for care by the many who remain well. Thus, risks that are unpredictable for the individual become predictable for the group and can be estimated.

Insurance markets are characterized by inherent instability. Adverse selection occurs in private insurance markets. Private insurance markets attempt to counter adverse selection by excluding the riskier customers and adjusting premiums which undermines the value of insurance altogether. This is one of the main arguments in favor of *public insurance* which can

be made universal and force everyone to share the risks. Cost sharing (deductibles, copayment), limits of benefits, total expenditures limits on policies, all are ways to deal with moral hazard. Despite the use of these mechanisms, the increased cost of health care and mental health care in particular has led to the introduction of other cost-containment mechanisms known as *managed care*. Managed care includes various organizational arrangements, such as HMOs and PPOs in the Unites States. Managed care organizations integrate the three functions of managing, financing and providing care. Managed care also includes financing mechanisms (capitation and risk-sharing contracts) that have strong incentives to reduce the amount of mental health care provided. The essence of managed care is its ability to ration care without imposing undue financial risk on consumers. Managed care principles can be utilized by public and private financing entities (see Mechanic 1997, 1998).

Managed care mechanisms are intended to affect rationing through establishment of a network of selected providers, contracts with providers that include incentives to limit care, introduction to the concept of medical necessity, and directing individuals to levels of care. Managed care plans also use various strategies to select good (profitable) risks from an insurance pool thus undermining the very existence of insurance and precluding effective risk pooling (Cutler and Zeckhauser 1998). They may prevent or discourage high-cost individuals from joining the plan or distort the services they provide in order to attract good and deter the bad risks. Supply-side management mechanisms to reduce costs include preauthorization requirements, utilization review, case-management for high utilizers, reducing access to specialty care providers, and substituting low cost providers for high cost providers.

Managed care has been shown to reduce overall mental health and substance abuse costs and service utilization. However, there has been concern about how the cost-containment strategies affect the access to care and the quality of care for patients. Cost-containment strategies can lead to excessive denial of care and under utilization of services, thus limiting access to care. Financial incentives to providers to reduce specialty referrals, hospital admissions, and length and amount of treatment can ultimately contribute to lowered quality of care.

Some demand-side cost-containment mechanisms, such as higher deductibles, copayments, visit- and day-limits, used by managed care can be reduced by introducing parity for mental health services with all other medical services. However, the introduction of parity generates incentives for insurance plans to replace the demand side cost-containment mechanisms with such mechanisms on the supply side. According to Frank and McGuire (1998) parity in insurance benefits for mental health care is not sufficient to guarantee equality in access to mental health services if managed care-rationing devices such as 'medical necessity' denials cannot be controlled. Nevertheless, introducing parity for mental health services is a step towards overcoming the discrimination of mentally ill, reducing the financial barriers to treatment of mental disorders and improved access to services. Parity in private insurance benefits may result in some public-to-private sector cost shifting.

3.4 Private Financing

Private financing can be direct (personal payments) or indirect (payments for health services by employers). Private insurance, out-of-pocket payments for direct purchase of services, medical savings accounts, and charitable contributions are all sources of private financing of health services. Private financing is an essential part of health care financing given government's limited ability to raise enough resources to finance health services. Reliance on private financing is a necessity in many low-income countries.

The combination of price competition and the ownership incentives that the private sector offers may contribute to improved efficiency in the sector as a whole. However, instabilities in private insurance markets necessitate public financing of services or, in some cases, government regulation of private insurance markets. Except for a few countries (Switzerland, the United States) private health insurance is a minor source of finance. In developing countries this is because most of the population cannot afford private insurance. In high-income countries it is because the state has assumed most of the insurance function (Schieber and Maeda 1997).

No industrial country relies exclusively on free markets to produce and allocate health care in general and mental health care in particular. Government intervention is justified, in order to improve the efficiency and equity with which the private market produces and allocates its services. Public regulation of private markets is also essential to ensure effective risk pooling, affordability, informed choice, and continuity of coverage.

3.5 Public Financing

The main argument for public financing of health services is the achievement of universal access. This argument is based on efficiency (arising from market failures of the private market) and equity considerations (financing should be according to the ability to pay, and the distribution according to need). Public financing of health services can be of two types, a national health-service approach and social insurance funds. In both models collective risk pooling is achieved through compulsory taxation. There are a few major differences between the two systems.

National health services are usually financed by general taxes and other public revenue sources and thus they are subject to annual budget processes. Social insurance funds are generally financed through payroll taxes and tend to be more independent of annual political machinations. Public finance accounts for 75 percent of health care financing in high-income countries with the exception of The United States where public finance covers about 50 percent of the overall health care financing. In developing countries public finance is less important and accounts for about 20 percent of the total health care financing with out-of-pocket payments accounting for most of it. The composition of public finance varies in high-income countries. France, Germany and The Netherlands rely on social insurance while Canada, the Nordic countries and the United Kingdom use general taxation. United States operates both kinds of publicly financed systems (Johnson and Musgrove 1997).

3.6 Financing of Services for Severe and Persistent Mental Illness

Direct public services for the most severely ill people is a common feature of health systems that may otherwise be organized to provide and pay for other health care with a range of approaches (Hollingsworth 1992). The public mental health system in the United States, Germany, Canada, UK, and France are quite similar despite vast differences in their approaches to health care financing. The responsibility for mental health care is assigned to sub-national government and tends to rely on local tax funding. The role of local government-provided mental health care tends to emphasize providing care to the poor and disabled. The commonality can be explained partially by the fact that provision of mental health care has a public safety component and, therefore, the locality has a greater interest in assuming more direct control over the delivery of certain forms of mental health care to fulfill its obligation of public protection. Most countries have laws that tend to favor society's interests (externalities) over individual liberties of the mentally ill. This can explain partly the greater role of the government in provision of mental health care, including public hospitals and clinics.

Treatment programs aimed at people with severe and persistent mental illness are concerned not only with the acute care of symptoms but address also housing needs, income support, rehabilitation, social contacts, and social control of the affected individuals. The nature of severe and persistent mental illness requires coordination of services from various providers. Many of the successful innovations in the treatment of this group of people with mental illness involve new ways of organizing and coordinating the various elements of care.

4. Setting Priorities and the Role of the Government

There is little consensus or empirical evidence as to the right level of financing for health care and for mental health care in particular. Policymakers face the perpetual challenge of raising sufficient revenue for the health sector in an equitable and efficient way. Decisions about health spending cannot be isolated from a country's political, social, and economic characteristics. People's values, the responsiveness of the political system to public preferences, and the resources available are all reflected in the pattern of financing of health care. In the context of scarcity, mental health services are seldom given high priority by national governments or by international aid programs. Despite changing public attitudes towards mental illness and trends towards integration of the mentally ill into the mainstream of the health care system, the fact remains that, in terms of public and private policy, mental illness and substance abuse are not treated the same as other illnesses. In most countries today priorities are set in ways that exclude large numbers of people from access to organized care. Ability to pay appears increasingly to be the mechanism rationing access to care. In middle and upper income countries, health-financing policy is driven frequently by the need to increase coordination, reduce fragmentation, and exert better control over total health care costs. Countries in this group are often worried that their level of health spending will threaten economic growth and competitiveness by making their labor force, and therefore goods and services, more expensive. If universal access to affordable and effective mental health care is to be achieved, priorities have to be agreed upon and implemented. Choosing public priorities is how economic reality becomes an integral part of mental health system development and reform.

See also: Health Economics; Health Insurance: Economic and Risk Aspects; Psychology and Economics; Psychopharmacotherapy, Costs of

Bibliography

Anonymous. Health Care Reform for Americans with Severe Mental Illness: Report of the National Advisory Mental Health Council 1993. *American Journal of Psychiatry* **150**: 1447–65

Buck J A, Umland B 1997 Covering mental health and substance abuse services. *Health Affairs* **16**: 120–26

Cutler D, Zeckhauser R 1998 The anatomy of health insurance. In: Newhouse J P, Culyer A Y (eds.) *Handbook of Health Economics*. North Holland Press, Amsterdam

Desjarlais R, Eisenberg L, Good B, Kleinman A 1995 *World Mental Health. Problems and Priorities in Low-Income Countries*. Oxford University Press, New York

Dorwart R A, Epstein S S 1993 *Privatization and Mental Health Care. A Fragile Balance*. Auburn House, Westport, CT
Frank R G, McGuire T G 1998 The economic functions of carve-outs in managed care. American Journal of Managed Care **4** (suppl.): SP31–SP39
Frank R G, McGuire T G 1999 Economics and mental health. In: Newhouse J P, Culyer A Y (eds.) *Handbook of Health Economics*. North Holland Press, Amsterdam, pp. 10
Frank R G, McGuire T G, Regier D A, Manderscheid R, Woodward A 1994 Paying for mental health and substance abuse care. *Health Affairs* **13**: 337–42
Hollingsworth E J 1992 Falling through the cracks: Care of the chronically mentally ill in the United States, Germany, and the United Kingdom. *Journal of Health Politics, Policy and Laws* **17**: 899–928
Jencks C 1994 *The Homeless*. Harvard University Press, Cambridge, MA
Johnson B, Musgrove P 1997 Government financing of health care. In: Schieber G J (eds.) *Innovations in Health Care Financing*. World Bank, Washington, DC
Kelleher K, Chaffin M, Hollenberg J, Fischer E 1994 Alcohol and drug disorders among physically abusive and neglectful parents in a community-based sample. *American Journal of Public Health* **84**: 1586–90
Kessler R C, Frank R G, Edlund M, Katz S J, Lin E, Leaf P 1997 Differences in the use of psychiatric outpatient services between the Unites States and Ontario. *New England Journal of Medicine* **336**: 551–7
Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshleman S, Wottehen H U, Kendler K S 1994 Lifetime and 12-month prevalence of DSM-III-R Psychiatric Disorders in the United States. *Archives of General Psychiatry* **51**: 8–19
Levitt K R, Lazenby H C, Cowan C A, Letsch S W 1991 National health care expenditures. *Health Care Financing Review* **13**: 29–54
Mechanic D 1997 Muddling through elegantly: finding the proper balance in rationing. *Health Affairs* **16**: 83–92
Mechanic D (ed.) 1998 *Managed Behavioral Health Care: Current Realities and Future Potential*. Jossey-Bass Publ., San Francisco
NAPHS 1999 Trends in behavioral healthcare systems: a benchmarking report. *1998 Annual Survey Report*. The National Association of Psychiatric Health Systems, Washington, DC
National Advisory Mental Health Council 1998 *Parity in financing mental health services, managed care effects on cost, access and quality 1998*.US Department of Health and Human Services, NIH Publication 98–4322
Regier D A, Kaelber C T, Rae D S, Farmer M E, Knauper B, Kessler R C, Norquist G S 1998 Limitations of diagnostic criteria and assessment instruments for mental disorders: Implications for policy. *Archives of General Psychiatry* **55**: 109–115
Rice D et al. 1990 *The Cost of Alcohol and Drug Abuse and Mental Illness*. USGPO, Washington, DC
Schieber G, Maeda A 1997 A curmudgeon's guide to financing health care in developing countries. In: Schieber G J (eds.) *Innovations in Health Care Financing*. World Bank, Washington, DC
Sharfstein S S 1991 Prospective cost allocations for the chronic schizophrenic patient. *Schizophrenia Bulletin* **17**: 395–400
Sharfstein S S, Goldman H H, Arana J 1992 The market for mental health care: New rules of reimbursement and new delivery systems. In: Talbott J A, Hales R E, Keill S L (eds.) *Textbook of Administrative Psychiatry*, 1st edn. American Psychiatric Press, Washington, DC pp. 91–115
Sharfstein S S, Webb W L, Stoline A M 1995 Economics of psychiatry. In: Kaplan H I, Sadock B J (eds.) *Comprehensive Textbook of Psychiatry/VI*, 6th edn. William & Wilkins, pp. 2678–89
Steadman H J, Mulvey E P, Monohan J, Robbins P C, Appelbaum P S, Grisso T, Roth L H, Silver E 1998 Violence by people discharged from acute psychiatric inpatient facilities and by others in the same neighborhood. *Archives of General Psychiatry* **55**: 393–401
Steinwachs D M, Kasper J D, Skinner E A 1992 Patterns of use and costs among severely mentally ill people. *Health Affairs* **11**: 178–85
World Health Organization 1993 *Evaluation of Recent Changes in the Financing of Health Services*. Report of a WHO Study Group 829. World Health Organization, Geneva, Switzerland
World Health Organization 1999 *The World Health Report 1999 – Making a Difference*. World Health Organization, Geneva, Switzerland
World Bank 1993 *World Development Report 1993: Investing in Health*. Oxford University Press, New York

S. S. Sharfstein and S. Tyutyulkova

Psychiatric Disorders: Culture-specific

1. *Introduction: Development of the Concept and Definition*

The concept of 'culture-bound syndromes' has been the focus of an ongoing debate in the field of transcultural or comparative cultural psychiatry between psychiatric universalists who interpret these conditions as cultural elaborations of universal neuropsychological or psychopathological phenomena, and cultural relativists who see them as generated and expressive of distinctive features of a particular culture. The 8th edition of Kraepelin's textbook of psychiatry in 1909 was the first handbook describing clinical pictures which were later labeled 'culture-bound syndromes,' namely *latah*, *koro*, and *amok*. In the past the term 'exotic' was used to emphasize the non-Western character of culture-related conditions. It was not anticipated then that a disorder such as *anorexia nervosa* in young women would become common due to sociocultural factors operant in modern Western society, so that it has now been recognized as a Western culture-reactive syndrome (DiNicola 1990). The concept of culture-specific psychiatric disorders was originally introduced into psychiatric literature in the 1950s and 60s by Yap, who also made the first attempt to order what he called 'culture-bound reactive syndromes,' known under a great variety of folk names, in a diagnostic classification schema (Yap 1967).

A list of 168 so-called culture-bound syndromes was compiled in a glossary by Hughes (1985), while the *Diagnostic and Statistical Manual* of the American Psychiatric Association, 4th edition, contains a glossary with 25 entries. These listings also include colloquial and local names for some well-known ubiquitous conditions. The Diagnostic Criteria for Research of the 10th Revision of the *WHO International Classification of Diseases* list 12 examples of 'culture-specific disorders' and suggest diagnostic codes while also stating that these conditions cannot be easily accomodated in established psychiatric classifications. Indeed, attempts at a nosological classification of culturally related syndromes have never been quite satisfactory due to (a) the overlap of diagnostically relevant symptoms, and (b) the fact that the various indigenous terms for similar syndromes are charged with different culture-specific meanings and traditional interpretations that codetermine the illness behavior of the patient and the reaction of the human environment.

On the basis of the data to be presented in this article, culture-specific psychiatric disorders are defined as clusters of symptoms and behaviors that can be plausibly related to specific cultural emphases, or to specific sociocultural stress situations, which are typical of a particular population.

2. *Local Folk Idioms of Distress*

Local folk idioms of distress for psychologically stressful situations that can occur in any culture are not signifiers of culturally specific disorders although the manifestations are influenced by the local culture. *Susto*, *espanto*, *miedo* are colloquially equivalent Spanish terms implying an experience of fright or scare, widely used among Hispanic populations of the Western hemisphere for a great variety of symptoms (Rubel et al. 1985). The conditions so labeled frequently resemble neurotic or somatoform syndromes but may also be caused by other kinds of disorders, including organic pathology. The Spanish word *nervios* is often used among Hispano-American and Caribbean populations, as is the English term *nerves* among rural Anglo-American groups, to denote chronic dysphoric mood states with various somatic manifestations especially in postmenopausal women under psychosocial stress, as is commonly encountered also in other cultures. The Spanish term *ataque de nervios* is, among Hispanic populations of Central and North America and the Caribbean, applied in reference to women showing a brief episode of behavioral discontrol as immediate reaction to any emotionally traumatizing experience (Oquendo 1994). Middle-aged Korean women, like those in other cultures, may react to stressful life situations and ongoing marital or family problems, with depressed mood, feelings of anger, and somatizing complaints, typically of a sensation of an epigastric mass. This chronic dysthymia is known among Koreans and Korean–Americans as *hwa-byung*, 'fire disease,' and traditionally treated by female shamans (Lin 1983).

3. *Culturally Stereotyped Reactions to Extreme Environmental Conditions*

Certain behavioral reactions to extreme environmental conditions are culturally stereotyped although they are not truly culture-specific. This is the case with conditions generally subsumed under the heading 'arctic hysteria,' a term used by outside observers for the often dramatic behavioral reactions shown by indigenous inhabitants of arctic and subarctic areas in stress situations in which the person affected experiences a temporary state of dissociated consciousness induced by acute anxiety (Foulks 1972). Physical deprivation, mineral and vitamin deficiencies, as well as the psychological stresses of surviving in an extreme climate, were adduced to explain these phenomena which have practically disappeared as the traditional Eskimo way of life has largely been abandoned.

Frequently cited in the ethnological literature are *pibloktoq* (Inuit, 'crazy') among Eskimo in Greenland, arctic Canada, Alaska, and Siberia, and *kayak-svimmel* (Danish, 'kayak dizziness') also called *kayak-angst*, among seal-hunting Greenlanders. *Pibloktoq*-type attacks usually affected younger women who after a prodromal phase of anxious and irritated behavior started imitating animal calls, then suddenly became agitated and, screaming loudly, rushed into the snow or even into icy water, albeit not too far off so that they were easily rescued. The motor excitation ended with collapse or sometimes with convulsive movements, followed by deep sleep. There was only a vague recollection of the event which, in traditional culture, was attributed to possession by evil spirits.

Attacks of *kayak-svimmel* occurred among seal hunters experiencing extreme isolation and sensory deprivation when sitting lonely and motionless in the kayak, surrounded by the monotonous whiteness of the arctic sea, waiting for a seal to appear. In this situation the hunters might suddenly fear the kayak was filling with icy water when they had a sensation of coldness in the legs. In a state of mounting anxiety they felt dizzy and unbalanced; were trembling and sweating, paralyzed with panic anxiety and subject to illusional perceptions. With help coming in time they could reach shore safely but more intense attacks followed on other trips so that they had to give up kayak hunting.

In the past, frequently referred to as a culture-specific condition of the Algonquin Indian tribes of northern Canada, was the '*windigo* (*witiko*) psychosis,' described as an obsessive murderous desire for human

flesh under the delusion of being possessed by, or turning into, the mythical cannibalistic *Windigo* monster with a heart of ice, said to appear during severe subarctic winters. The struggle for survival in the harsh northern climate may have led to isolated instances of famine cannibalism but a critical examination of the many reports of '*windigo* psychosis' could not confirm a single case of cannibalistic compulsion (Marano 1982). The term *windigo* is today used among the northern Ojibwa to denote a state of deep depression, without implication of cannibalistic impulses.

4. *Syndromes related to Specific Cultural Emphases or Specific Socio-cultural Stress Situations*

4.1 Syndromes Related to a Cultural Emphasis on Fertility and Procreation

4.1.1 'Genital shrinking' syndrome. The Malayo-Indonesian term *koro* and the Mandarin Chinese term *suo-yang* (Cantonese *suk-yang*) denote 'shrinking of the penis.' These terms refer to an acute state of panic anxiety prompted by the belief of the affected male person that his penis is shrinking and will disappear, and of the affected female person that her breasts and/or labia are shrinking. The 'genital shrinking' is believed not only to lead to impotence and/or sterility but to death once the organs are completely retracted. This syndrome came to the attention of European psychiatry in the late nineteenth century through reports by colonial physicians working in South East Asia and through the growing interest in traditional Chinese medicine. Chinese traditional medicine interprets *suo-yang* as yin–yang imbalance due to deficiency of the 'warm' male principle yang and excess of the 'cold' female principle yin, so that vitalizing yang remedies are indicated.

For the theorists of psychoanalysis, *koro* served as the concrete example of Oedipal castration anxiety, while some Chinese authors like Yap saw in *suo-yang* the paradigm of a true culture-bound syndrome, assuming that the disorder itself was generated by the suggestive effect of traditional Chinese concepts. However, it has since been ascertained that the typical 'genital shrinking' syndrome also occurs not infrequently among Asian and African populations without Chinese cultural influence. Complaints of genital shrinking are also occasionally verbalized by Western patients but these cases are different from the culturally related 'genital shrinking' syndrome as they are associated with other, usually chronic, neurotic, psychotic, or organic conditions; and the patient's ethnic–cultural group in Western society does not share the belief in the reality of genital disappearance and ensuing death. *Koro/suo-yang* is the only culturally related syndrome that has occured in major epidemic outbreaks. Epidemics of 'genital shrinking' have been repeatedly reported in recent decades (Jilek 1986). Outbreaks in Singapore, Thailand, North East India, and South China were precipitated by collective anxieties over perceived threats to the security or even survival of the afflicted population; in Nigeria by fears of magic robbing of the genitals.

4.1.2 'Semen loss' syndrome. Known in the traditional medicine of China as *shen-k'uei*, 'weakness of the kidneys,' in the Ayurvedic medicine of India as *dhat* (from Sanskrit *dhatu*, essential body substance) or *jiryan*, in Sri Lanka as *sukra prameha*, the complaint of 'semen loss' is not uncommon among young men of these countries in which a high premium is placed on reproductive capability and semen is considered a precious elixir of life. The presented symptoms have a hypochondriac quality; they are associated with an anxious and dysphoric mood state and typically attributed to the imagined loss of sperm in urine or to natural nocturnal emissions (Paris 1992). Sperm loss is assumed to cause the symptoms of general weakness, loss of energy, exhaustion, difficulty concentrating, sexual dysfunction, and rather vague somatic complaints. Many of these patients consult practitioners of traditional medicine as well as medical doctors; as immigrants they are also increasingly seen in European and North American clinics.

4.2 Syndromes Related to a Cultural Emphasis on Learnt Dissociation

4.2.1 Latah-*type reactions*. The peculiar *latah* reactions have been cited by cultural relativists as a paradigm of culturally generated behavior (Kenny 1978), and by biopsychological universalists as cultural elaborations of the neurophysiological startle reflex (Simons 1996). *Latah* reactions were interpreted by Gilles de la Tourette as a variant of his neurological syndrome, and by Kretschmer as a phylogenetic mechanism. The complete *latah* syndrome consists of an initial startle response provoked by a person of certain social status and triggered by a visual, acoustic, or tactile stimulus, the sensitivity to which is both individual and culture-specific. The startle response induces an altered state of consciousness, associated with repeated impulsive utterance of sounds or inarticulate words, often also swearing with sexual or obscene words ('coprolalia'). Characteristic are echopraxia, echomimia, and echolalia, namely the exact imitation of the provoking person's movements, facial and oral expression; even of words and sentences spoken in foreign languages. Automatic

obedience to commands may occur in hypnoid trance and serve the amusement of onlookers. The *latah* syndrome can sometimes culminate in a brief state of cataleptic rigidity. Because of the typical echo symptoms, *latah* reactions are known in French psychiatry as *névroses d'imitation*.

Automatic imitation behavior has also repeatedly been observed in normal indigenous people during first contact with Europeans, as already reported by Charles Darwin from Tierra del Fuego and more recently from remote areas of Melanesia (Gajdusek 1970) and East Africa (Jilek-Aall 1979). The typical *latah* syndrome occurs mostly in females over two geographic zones (Winzeler 1995): the South East Asian zone, encompassing Malaysia, Indonesia (*latah*, 'nervous, ticklish'), Thailand (*bah-tshi*, 'tickle-crazy'), Burma-Myanmar (*yaun*, 'ticklish'), the Philippines (*mali-mali*); the North Eurasian zone, encompassing the Ainu people of Hokkaido and Sakhalin (*imu*, 'startle'), aboriginal peoples in Mongolia and Siberia (Russian, *miryachit*), and the Sami of northernmost Europe ('Lapp panic'). The varied ethnic groups of these two *latah* zones have no culture traits in common other than (a) learnt dissociation which is practiced from early youth on in religious, social, and therapeutic ceremonies, trance rituals, and shamanic seances; (b) learning by imitation and coping by copying the behavior of persons who appear to be more powerful. Sporadic occurence of *latah* symptoms has been reported among some males in Arabic-Islamic populations of North Africa and Yemen, and among South African Bantu people. While European cases are rarely encountered elsewhere, the endemic and familiar incidence of so-called 'jumping,' a reaction similar to *latah*, among males of French Canadian background in certain rural areas of Maine, New Hampshire, and Quebec, has attracted scientific interest since the 1870s.

4.2.2 Amok-*type reactions*. The term *amok* is originally derived from the Portuguese–Indian *amuco*, referring to heroic warriors ready to die in battle who are immortalized in Malayan epics, similar to the *berserkr* in ancient Norse sagas. Today, 'running amok'—likewise 'going berserk'—has become a colloquial label for any episode of apparently unprovoked randomly aggressive and homicidal behavior, usually engaged in by attention-craving angry young men in modern Western society. Although such behavior has certain features in common with the *amok* reaction observed among Asian *amok*-runners, it does not represent the specific dissociative syndrome described under this term (Pfeiffer 1994). Typical *amok* reactions are preceded by a dysphoric-tense mood state in the context of perceived or imagined social adversity or slight, loss of prestige or status, interpersonal or socioeconomic problems. A sometimes trivial event then triggers the actual *amok* reaction. After a phase of meditative brooding an alteration of the state of consciousness ensues in which visual perception changes: the outside world turns into darkness or becomes red, threatening visions may appear, causing feelings of fear or rage. The affected person is suddenly seized by hypermotility and embarks on the *amok* run, a discharge of aggression leading to random destruction and killing, and eventually to suicide or suicide attempt. In case of survival, the attack is terminated by stuporous sleep; afterwards amnesia is claimed for actions committed.

Amok reactions have been presented as specific to the male ethos of Malayo-Indonesian culture. Such reactions also occur among related ethnic–cultural groups in the Philippines and in Laos. However, there are also typical *amok* cases in Papua New Guinea (Burton-Bradley 1975), besides the common non-violent pseudo-*amok*, the psychodramatic 'wild man behavior' of New Guinea highlanders. In historical analysis, *amok* in Southeast Asia appears to have changed from a glorified warrior ethos and means of social protest of the precolonial era, to a dissociative reaction, which in more recent times is mainly found in marginal figures or as an episode during the course of a chronic mental disorder (Murphy 1973).

4.3 Syndromes Related to a Cultural Emphasis on Presenting a Pleasing Physical Appearance: Taijin-Kyôfu *Reactions*

In contrast to the condition commonly diagnosed in Western psychiatry as social phobia, the Japanese form of anthropophobia, *taijin-kyôfu* ('fear in relation to others'), is not a phobic avoidance of social contacts in order to avoid unpleasant feelings for oneself; rather it is the fear that one's external appearance may be disturbing or offending to the other who as a rule is neither a stranger nor a close relation but a respected acquaintance. Symptoms of *taijin-kyôfu* were already mentioned in the eleventh century Genji saga by Lady Murasaki Shikibu, and detailed clinical descriptions suggest this condition to be typical of Japanese culture (Kimura 1995). However, this type of anthropophobia is also not uncommon in Korea (Lee et al. 1994). The clinical syndrome *taijin-kyôfu* was introduced into psychiatric literature in the 1920s by Shoma Morita as a subtype of the nervous temperament (*shinkeishitsu*) that he considered to be the main indication for his 'Morita Psychotherapy.' The syndrome is not uniform but occurs as one of several phobic anxiety manifestations: as erythrophobia (i.e., fear of blushing in front of others); as eye contact phobia (i.e., fear that others feel bothered when looked at); as body odour phobia (i.e., fear that one's own assumedly offensive body odour will bother others); as dysmorphophobia (i.e., fear of having an extremely unattractive, repulsive face or defective physique); or as fear that one

will grimace or otherwise behave embarrassingly. The sufferers tend to be convinced of their imagined defects to a degree that their phobias appear delusional, and the presence of other psychotic symptoms has to be ruled out in order to delimit *taijin-kyôfu* from a schizophrenic disorder. However, the anthropophobic patient's relation to the human environment is fundamentally different from that seen in chronic schizophrenic psychoses. *Taijin-kyôfu* patients are neither autistic nor paranoid; they suffer from a fear of presenting themselves.

4.4 Syndromes Related to Acculturative Stress

4.4.1 Brain fag. The term 'brain fag,' coined by Nigerian students, refers to a condition first reported by Prince (1960). The brain fag syndrome has since been recognized in a high percentage of Nigerian students at home and abroad, and also in other African countries (Liberia, Ivory Coast, Uganda, Tanzania, Malawi, Swaziland) in students exposed to the acculturative stress of a Western-type education system emphasizing theoretical book knowledge, quite different from the practical know-how and tribal lore acquired through oral traditions by older generations in Africa. It is noteworthy that subnormal intelligence, malnutrition, or physical disease do not account for the symptoms of brain fag which are also seen at educational institutions in Europe and North America among African students in good state of nutrition and general health. The brain fag syndrome is categorically different from the general study stress and examination crises experienced by students from Western or Asian cultures with centuries of literary tradition and education systems already long-based on reading and writing and on memorizing texts for formal examinations. The brain fag syndrome is also clinically different from the common complaints everywhere related to study stress: The mere effort of reading may cause blurred vision and flow of tears. There are also inability to concentrate on reading material and lectures, impairment of retention and memory, daytime fatigue and disturbed night sleep. Characteristic symptoms are the feeling of burning heat on head and body and the sensation of crawling as if caused by insects on, or by worms under, the skin.

4.4.2 Bouffée délirante-*type reactions*. Clinical findings as well as research data indicate that acute transient psychotic and psychosis-like reactions occuring in African and Afro-Caribbean populations are disorders directly related to culture and culture change, and are encountered with increasing frequency in Africa (Jilek and Jilek-Aall 1970). Such reactions are characterized by intense fear of magical persecution or retribution and are usually triggered by an event or experience evoking acute anxiety in the context of beliefs in sorcery or witchcraft that persist or even increase under rapid sociocultural changes. *Bouffée délirante*-type reactions are of sudden onset and brief duration, they are associated with paranoid delusions, oneroid (i.e., dream-like) confusion, frequently also visual and auditory hallucinations, and often lead to dramatic and sometimes dangerous acting-out behavior.

Anglophone psychiatrists have described transient psychotic and psychosis-like reactions in Africans under various terms while their francophone colleagues use the diagnostic label *bouffée délirante*, which was introduced into French psychiatry in the 1880s by Magnan and later redefined by French-speaking authors in Africa in accordance with their clinical experience (Collomb 1965). The well-known Nigerian psychiatrist Lambo introduced the English term 'frenzied anxiety' and equated it with *bouffée délirante* (Lambo 1960). As repeatedly stated by Lambo and by other researchers, the main underlying causes for these states are not toxic–organic but sociocultural factors. Among these factors, of paramount importance is the stress exerted by rapid culture change which is leading to the social marginalization of many Africans in the process of 'modernization,' as recently also confirmed in Swaziland (Guinness 1992). Under situations of urbanization and Westernizing acculturative pressures the traditional communalistic society is disintegrating and the supportive kin network is breaking down. The individual then experiences an increasing economic rivalry and social isolation which intensifies the old fears of witchcraft and sorcery, never obliterated by Christianity, while the traditional protective and remedial resources are no longer readily available.

5. Culture-specific Disorders vs. Traditional Ritual Behavior

Culture-specific psychiatric syndromes have to be differentiated from extracurricular but not pathological behavior in the traditional rituals of non-Western cultures. A Western-trained clinician who attaches pathology labels to behaviors and ideas that are considered normal in a non-Western culture, commits a 'eurocentric fallacy.' As a general statement it can be said that the culture-specific disorders are not part of ceremonials in which ritualized altered states of consciouness are induced and practiced for religious or therapeutic purposes. While such institutionalized trance and 'possession' states constitute culture-congenial behavior in certain ceremonials, the culture-specific syndromes described above are considered as abnormality or illness also by the indigenous experts of the culture and are explained in terms of a particular folk etiology which determines traditional therapeutic management.

See also: Body: Anthropological Aspects; Cross-cultural Psychology; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Obesity and Eating Disorders: Psychiatric

Bibliography

Burton-Bradley B G 1975 *Stone Age Crisis: A Psychiatric Appraisal*. Vanderbilt University Press, Nashville, TN

Collomb H 1965 Bouffées délirantes en psychiatrie africaine. *Psychopathologie Africaine* **1**: 167–239

DiNicola V F 1990 Anorexia multiforme: Self-starvation in historical and cultural context. *Transcultural Psychiatric Research Review* **27**: 165–96, 245–86

Foulks E F 1972 *The Arctic Hysterias of the North Alaskan Eskimo*. Anthropological Studies No. 10. American Anthropological Association, Washington, DC

Gajdusek D C 1970 Physiological and psychological characteristics of stone age man. *Engineering and Science* **33**: 26–33, 56–62

Guinness E A 1992 Patterns of mental illness in the early stages of urbanization. *British Journal of Psychiatry* **160**: Supplement 16: 12–72

Hughes C C 1985 Glossary of 'culture-bound' or folk psychiatric syndromes. In: Simons R C, Hughes C C (eds.) *The Culture-bound Syndromes: Folk Illnesses of Psychiatric and Anthropological Interest*. Reidel, Dordrecht, pp. 469–505

Jilek W G 1986 Epidemics of 'genital shrinking' (koro): Historical review and report of a recent outbreak in South China. *Curare* (Heidelberg) **9**: 269–82

Jilek W G, Jilek-Aall L 1970 Transient psychoses in Africans. *Psychiatria Clinica* **3**: 337–64

Jilek-Aall L 1979 *Call Mama Doctor: African Notes of a Young Woman Doctor*. Hancock House, Saanichton, BC

Kenny M G 1978 Latah: The symbolism of a putative mental disorder. *Culture, Medicine and Psychiatry* **2**: 209–31

Kimura B 1995 *Zwischen Mensch und Mensch: Strukturen japanischer Subjektivität*. Wissenschaftliche Buchgesellschaft, Darmstadt

Lambo T A 1960 Further neuropsychiatric observations in Nigeria. *British Medical Journal* **2**: 1696–1704

Lee S H, Shin Y C, Oh K S 1994 A clinical study of social phobia for 10 years. *Journal of the Korean Neuropsychiatric Association* **33**: 305–12

Lin K-M 1983 Hwa-byung: A Korean culture-bound syndrome? *American Journal of Psychiatry* **140**: 105–7

Marano L 1982 Windigo-psychosis: The anatomy of an emic-etic confusion. *Current Anthropology* **23**: 365–412

Murphy H B M 1973 History and the evolution of syndromes: The striking case of latah and amok. In: Hammer M, Salzinger K, Sutton S (eds.) *Psychopathology: Contributions from the Biological, Behavioral and Social Sciences*. John Wiley, New York, pp. 33–55

Oquendo M 1994 Differential diagnosis of ataque de nervios. *American Journal of Orthopsychiatry* **65**: 60–5

Paris J 1992 Dhat: The semen loss anxiety syndrome. *Transcultural Psychiatric Research Review* **29**: 109–18

Pfeiffer W M 1994 *Transkulturelle Psychiatrie*, 2nd edn. Thieme, Stuttgart

Prince R 1960 The 'brain fag' syndrome in Nigerian students. *Journal of Mental Science* **106**: 559–70

Rubel A J, O'Nell C W, Collado R 1985 The folk illness called susto. In: Simons R C, Hughes C C (eds.) *The Culture-bound Syndromes*. Reidel, Dordrecht, pp. 333–50

Simons R C 1996 *Boo! Culture, Experience and the Startle Reflex*. Oxford University Press, New York

Winzeler R L 1995 *Latah in Southeast Asia: The History and Ethnography of a Culture-bound Syndrome*. Cambridge University Press, Cambridge

Yap P M 1967 Classification of the culture-bound reactive syndromes. *Australian and New Zealand Journal of Psychiatry* **1**: 172–9

W. G. Jilek

Psychiatry and Sociology

Sociology and psychiatry have a longstanding relationship which has never been free of tensions and can at least be characterized as ambivalent. On the one hand, psychiatry as a subspecialty of medicine, which has to solve concrete problems in everyday life, has become a subject matter of sociology, which often criticizes psychiatry's concepts and practices. On the other hand, psychiatry has profited from cooperating with sociologists by making use of sociological research methods. The article sets out with describing the relationship between the two disciplines, and then proceeds to discussing the contributions made by sociology towards the understanding of how psychiatry functions ('sociology *of* psychiatry') before describing the uses of sociology within the field of psychiatry ('sociology *in* psychiatry').

1. Sociology and Psychiatry—An Uneasy Relationship

Sociology and psychiatry have a common topic: deviant behavior. Sociology, as a scientific discipline which examines the rules by which society functions, must have a genuine interest in phenomena which threaten these rules. Emile Durkheim (1895), one of the founding fathers of sociology, points out in his famous book, *The Rules of the Sociological Method*, that the standards and rules defining what is pathological help to reinforce the norms and values of society. For psychiatry, in turn, a specific form of deviant behavior, called mental illness, is its very '*raison d'être*.' In this respect, the relationship between sociology and psychiatry could be regarded as a relationship of equal partners joining their intellectual forces in analyzing a topic of common interest, though using different approaches. However, there is also a different type of relationship between these two disciplines, in so far as one of them, psychiatry as a subspecialty of medicine, has become the subject matter of the other, sociology.

When discussing the relationship between sociology and psychiatry, a profound difference between these two disciplines has to be kept in mind: Psychiatry is not only a scientific discipline—like sociology—but also a practice field, akin to medicine in general, or to education. Sociologists might aim at changing things, but sociology has never been a major player in practical problem-solving, in contrast to psychiatry, which from its very beginning 200 years ago was assigned the societal task of treating the mentally ill or managing (controlling, as sociologists would say) the problem of mental disorders. Van der Geest (1995), in comparing medicine and social sciences (exemplified by cultural anthropology), speaks of two 'ethnocentrisms,' i.e., two different cultures, which are dissimilar in three aspects. The culture of medicine is practical and problem-oriented (doctors are supposed to find concrete solutions to concrete problems), whereas the social sciences culture is theoretical; doctors have no time to lose (they have to act promptly before it is 'too late'), social scientists are not in a hurry to finish their work; finally, doctors measure their success in maintaining and restoring people's health (they are accountable to people); for social scientists the fulfilment of their task does not lie in an improvement of the lives of the people studied, but in the production of texts about them. The 'ethnocentrism' of these two disciplines makes communication difficult and prone to misunderstandings.

Psychiatry's relationship to sociology is ambivalent. On the one hand the application of sociological research methods to psychiatric problems have undoubtedly contributed to raise psychiatry's societal visibility and importance—by furthering epidemiological research, which has shown that between one quarter and up to nearly half of the US population suffer from at least one diagnosable mental disorder during a person's life time (e.g., Robins and Regier 1991, Kessler et al. 1994). On the other hand, sociologists have been, and still are, among the most radical critics of psychiatry, concerning both its concepts and its practice.

One further aspect renders the discussion of the relationship between sociology and psychiatry difficult: Both disciplines are extremely diverse within themselves. Sociology comprises a multitude of different philosophies (e.g., the 'social constructivist' and the 'social realist' perspectives) and research methods (e.g., the 'quantitative' and the 'qualitative' approaches). Psychiatry, both as a scientific discipline and a practice of identifying, explaining, and treating behavioral problems, which are deemed to be mental disorders, is even more heterogeneous, since psychological and sociological, as well as biological, approaches live under its roof. There is also a substantial overlap between each of the two disciplines and other fields. Sociology can be broadened to mean social sciences in general, and incorporate social anthropology, social psychology, and other disciplines. Furthermore, fields such as epidemiology, public health, social medicine, health economics, and others, if applied to psychiatric problems, would have to come in. Also, no attempt can be made here to discuss the boundaries of psychotherapy, clinical psychology, and social psychiatry in relation to clinical psychiatry.

For the purpose of this review, which addresses a broad audience, the author—who is a psychiatrist with strong interests in sociology—has found the dichotomy of the relationship between sociology and medicine, proposed by Robert Straus (1957), useful. Following Straus' distinction between 'sociology *of* medicine' and 'sociology *in* medicine,' the ensuing part of the text will be divided into the 'sociology *of* psychiatry' and the 'sociology *in* psychiatry.' The first approach discusses sociological contributions toward the understanding of psychiatry as a whole, i.e., of how psychiatry behaves in terms of developing and using concepts and applying them in practice. The second approach deals with the application of sociological methods within the field of psychiatry. While these two approaches are sometimes intertwined, their separation is nevertheless useful for the purpose of orientation. This distinction can, at least implicitly, also be found in two recent reviews of the sociology of mental health and illness, to which the reader of the present review is referred for more detailed discussions of the issues raised here (Pilgrim and Rogers 1999, Busfield 2000). It has to be noted, though, that 'sociology of mental health and mental illness' is a broader concept than the two sociologies (*of* and *in* psychiatry) discussed here. In the following two sections these two approaches are illustrated by examples.

2. *Sociology of Psychiatry*

The sociology of deviant behavior is central to sociology. Analyzing how boundaries are drawn between the normal and the abnormal, in other words, how the normal and the abnormal are classified, and analyzing how society reacts to the abnormal is one of the main interests of sociology. Since psychiatry has obtained an important definitional power over the specific type of deviant behavior called mental illness, and since it also has substantial influence on how such behavior is dealt with, it is obvious that psychiatry itself has become a subject matter of sociology. Sociologists have dealt both with psychiatric concepts and with the practice of psychiatry, with a special emphasis on the professional nature of psychiatry, on psychiatric services and the issue of legal vs. medical control of mental disorders.

2.1 *Psychiatric Concepts*

Sociologists have always expressed doubts about the coherence, credibility, and validity of psychiatric

concepts and knowledge. The most profound sociological criticism of psychiatry concerns the concept of mental illness and its classification into discrete diagnostic entities. While such criticism has come forward from many quarters, including the anti-psychiatric and consumer movements as well as scientific psychiatry itself (where some advocates favor dimensional rather than categorical concepts), the specific and original contribution from sociology was the labeling theory in the 1960s (Scheff 1999), a variant of what is called 'social constructivism,' i.e., a theory assuming that reality is a product of human definitions and not something waiting out there to be discovered (Brown 1995). More specifically, using 'symbolic interactionism' as their framework (a theory about how people map out social reality in everyday life), the advocates of 'labeling theory' claimed that the same deviant behavior may either go unnoticed or be labeled as mental illness, depending on the circumstances in which it occurs and is perceived by others. For instance, in a now famous study, Rosenhan (1973) observed that healthy test persons ('Pseudopatients') who presented themselves to psychiatric hospitals and told the doctor that they were hearing voices (saying 'empty,' 'hollow,' and 'thud') were admitted and received a psychiatric diagnosis. Once labeled, so the theory goes, the person's identity and social status is altered significantly. Part of the process is that the labeled person internalizes the new identity ascribed to them and becomes trapped in the new role of a mental patient.

A radical social constructivist perspective creates hostility among doctors, patients, and relatives, since they get the feeling that it denies the reality of suffering connected with mental disorders. While the idea that mental disorders, especially schizophrenia, are caused by labeling is more or less abandoned in recent thinking, labeling is certainly a powerful mechanism when it comes to discussing the role of stigmatization and discrimination (Link et al. 1987). A more moderate form of criticizing psychiatric diagnosis is the notion that through the diagnostic process illnesses are 'socially framed' (Rosenberg 1992). Mirowsky and Ross (1989), while acknowledging that psychological problems are real, criticize that psychiatric diagnosis treats attributes as entities, and that psychiatrists often speak of psychological problems 'as if they are discrete entities, entering the bodies and soul of helpless victims.' They target especially the operationalization of psychiatric diagnosis in modern diagnostic systems, such as the *Diagnostic and Statistical Manual III* (DSM-III) (and onwards) of the American Psychiatric Association (1994). It has to be acknowledged, though, that the construction of DSM-III was partly a reaction to the criticism that psychiatrists were not able to diagnose mental disorders reliably (Rosenhan 1973). However, while operationalization might have increased the reliability of psychiatric diagnosis, the question of validity is still debatable. Brown (1995) has suggested that instances of medicalization of normal life problems (such as 'late luteal phase dysphoric disorder—LLPDD'; now called 'pre-menstrual dysphoric disorder—PMDD') are mainly a feature of professional 'expansionism.'

2.2 *Psychiatry as a Profession*

Sociologists have not only dealt with psychiatric concepts but also with the process of professionalization of psychiatry, which is linked closely to both conceptual issues and to the practice of psychiatry. The rise of the psychiatric profession during the nineteenth century has been regarded alternatively as having been due to its functional value for economic order and efficiency under capitalism (Scull 1979), and to its role for maintaining the moral, rather than the economic order (Foucault 1965).

Psychiatrists certainly have difficulties in agreeing with these views, since they have always seen themselves as primarily a helping profession, whose practice is based on scientific advances. There is a different strain of sociological analysis of psychiatry as a profession, going back to the German sociologist Max Weber, who postulated that professionals want to advance their own social status, persuade clients and potential clients about the need for the service they offer, corner the market in that service, and exclude competitors (Freidson 1970, Pilgrim and Rogers 1999). Other sociological approaches to the issue of professionalization include the sociology of deviance (crises in personal lives are reframed as mental illness), symbolic interactionism (regarding the professional power in negotiating the meaning of mental illness), and feminist sociology (men occupy higher-status positions within the more powerful professions).

The concept of power is central to this analysis, and the specific knowledge base of a profession (which can only be obtained by a difficult process of socialization) gives the profession power over its clients, its recruits and related occupational groups. With the emergence of nonpsychiatric mental health professionals, professional power has become a complex issue, since clinical psychologists, community psychiatric nurses, psychotherapists, to name only a few, are also increasingly professionalized and state registered. Also, in the era of managed care, professional autonomy and decision-making power are threatened increasingly by third-party authorization mechanisms.

Finally, those receiving care from professionals start to threaten traditional professional identities. They are themselves becoming more powerful in forming self-help and advocacy groups, and tend to use terms such as 'clients,' 'consumers,' and even 'survivors of psychiatry,' instead of the term 'patients.' These terms have different implications of how the traditional

doctor–patient relationship is redefined. 'Client,' unlike 'patient,' implies partnership rather than submissivity and passivity. The term 'consumer' goes even further in claiming that patients have certain entitlements for service for which they can sue, and the concept of 'survivor of psychiatry' denies the right of psychiatry to intervene in a person's problems. Pilgrim and Rogers (1999) give a detailed account of sociological thinking about these issues, and how the situation has changed since Talcott Parsons (1951) formulated the paradigm of the 'sick role.' According to Parsons, the sick role allows people to be exempted temporarily from social responsibilities, but, in exchange, they must accept that they cannot get well on their own, and that they have to look for professional help. Foucault's (1977) concept of 'productive power' provides an interesting sociological framework for the evolving discourse between professionals and the (potential, actual, or ex-) users of their services.

2.3 Psychiatric Practice

The old mental hospital was the mainstay of psychiatry for over 150 years. At the end of this period, in the 1950s, the mental hospital (or the 'asylum' as it had been called) became a favorite topic of sociology. There were at least two reasons for this interest. Mental hospitals could be regarded as a 'small society,' an experimental situation, so to speak, in order to study sociological phenomena in general (Caudill 1958), and, probably more important, sociologists became critics of the impersonal way in which patients were treated. Goffman (1961), in his famous study based on participant observation, coined the concept of the 'tal institution' for the mental hospital and similar organizations which disregard the personal needs of their inmates. In a more rigid research design, Wing and Brown (1961, 1970) showed a strong association between the typical poverty of the social environment of the mental hospital and the 'clinical poverty' of psychopathological phenomena such as blunted affect and social withdrawal—a view which has been challenged later (Curson et al. 1992). In a different approach, social historians attributed the growth of the asylum during the nineteenth century to more general developments in society, such as the growth of humanitarianism (Jones 1960), of capitalism (Scull 1979), and of surveillance tendencies (Foucault 1977).

Beginning during the 1950s, in countries with large numbers of beds in psychiatric hospitals—such as the USA and the UK—the numbers of psychiatric beds, which had been steadily rising over the first half of the twentieth century, started to fall. While, for instance, in the USA, in the mid-1950s about 600,000 patients were in mental hospitals, this number had come down to below 100,000 some 50 years later. The debate about the reason for this fall is still not resolved. Psychiatrists attribute it mainly to the introduction of the neuroleptic drugs (which can control delusions and hallucinations). Sociologists have suggested different explanations: That, with the growth of the welfare state, controlling by segregation had become too costly, and that the public's attitude toward psychiatric patients had changed in the sense of becoming more favorable (for an overview, see Busfield 1986). Furthermore, differences in the extent and kind of implementation of community psychiatry in different countries have been related to different health and social welfare systems (Goodwin 1997).

Forster (2000) examined three sociological theories (political-economy: cost containment; professional dominance: psychiatric 'expansionism'; and post-structuralism: maintaining order and control) about the process of de-institutionalization, and concludes that none of them explains sufficiently the move away from hospital to community psychiatry. He proposes that the kind and degree of the medicalization of psychiatry has changed profoundly: While mainstream psychiatry is now more medicalized than before, the treatment of less serious mental health problems cannot be monopolized any more by medicine, and is becoming increasingly the practice field of other professions.

Compulsory admission and compulsory treatment are an intrinsic part of psychiatric services. Since the beginning of the twentieth century the legal professions have been involved increasingly in the process of admissions to mental institutions in order to counteract what has been regarded as mismanagement by the medical professions. This has prompted sociologists to analyze the social processes involved. There are basically two sociological positions: Those seeing the legal and the medical systems as two competing institutions (Mechanic 1978), and those who argue that, in practice, they join to become one actor, a 'treatment-control' system which 'takes care' of residual cases viewed as problematic for society, although the written law distinguishes between treatment and control (Dallaire et al. 2000). In the UK and other countries, a new dimension has been added to this issue with the concept of a 'community treatment order,' which would allow compulsory medication for patients living in the community and would let the pendulum swing back to medical control (Pilgrim and Rogers 1999).

Finally, sociologists have also analyzed and criticized specific therapeutic approaches. An example is the classical antipsychotic drugs and the barriers to the professional recognition of their side effects, especially of tardive dyskinesia, as an 'iatrogenic disease' (Brown and Funk 1986).

3. Sociology in Psychiatry

Sociologists working within the field of psychiatry are often criticized by their colleagues that they accept

psychiatric concepts naively. Robert Straus (1957) has noted that 'the sociologist *in* medicine risks a good relationship, if he tries to study his colleagues.' Work presented in this section is therefore clearly separated in nature from that discussed in the previous section. In principle, psychiatric diagnosis is taken for granted, or at least not regarded as problematical by these authors.

Typical variables investigated in psychiatric studies using sociological methods include gender, age, social class, and inequality, as well as ethnicity and race (Pilgrim and Rogers 1999). An overlap with the sociology *of* psychiatry can sometimes not be avoided, especially if questions of disadvantage, stigma, and mental health policy are the topic of a study. Gender is a typical example of such a topic belonging to both sociologies (Busfield 1996).

Sociologists *in* psychiatry have been, and are, active in a large number of fields. They study, or help to study, the distribution of mental disorders in society, and the social factors and processes which may be causal for mental illness or influence its course. They investigate which social factors prevent the early detection, and thus the early treatment, of mental disorders, and which factors influence 'illness behavior' or 'help-seeking behavior' (here conceptual issues such as the professional and lay definition of illness, i.e., aspects of the sociology of psychiatry, are involved), and how effective psychiatric treatments are under different institutional and social conditions. They have introduced new outcome measures for psychiatric interventions, which are less 'illness-centered' and more 'life-centered' (such as quality of life, social functioning, disability, social support, burden on the family) and have promoted the use of these measures to assess the 'reality of life' of persons with mental disorders, including the effects of stigma. Finally, in times of increasing restrictions of funds for health services, sociologists apply these concepts to health policy and health economics.

The epidemiology of mental disorders is discussed elsewhere in this volume (see *Mental Illness, Epidemiology of*). It must be noted though that the epidemiological survey methodology cannot only be used to study the influence of *social* factors, but also of *biological* and *genetic* factors. In fact, before the advent of the molecular biological approach, but also still at the time of writing, many conclusions drawn by psychiatric geneticists were based on epidemiological studies on the distribution of specific mental disorders among the relatives of mentally ill persons.

A few important historical examples of 'sociology in psychiatry' are worth mentioning. During the 1930s, the Chicago school of urban sociology studied the distribution of psychiatric disorders, such as schizophrenia, across the city alongside other forms of deviant behavior, and found higher rates in deprived areas (Faris and Dunham 1939). In the 1950s and 1960s, groups of sociologists and psychiatrists carried out a number of seminal studies: Cumming and Cumming (1957) investigated public attitudes toward mental illness and tested the efficacy of a large intervention program (which turned out not to be successful), Hollingshead and Redlich (1958), and Langner and Michael (1963) found an association between lower social class and mental illness. Srole et al. (1962) demonstrated the astonishingly high prevalence of mental disorders in metropolitan New York. Star (1955) found that public attitudes towards the mentally ill were stereotyped and quite unfavorable. Finally, Brenner (1973), in a time series analysis covering 117 years, demonstrated that in times of economic recession the yearly rates of admissions to mental hospitals in New York increased.

A few more recent studies and topics applying sociological research methods are described here by way of illustration. Brown and Harris (1978) conducted a detailed sociological analysis on the etiology of depression in women and found that, among women who experienced a major life event, only those who had no close confidant, or in whom other social vulnerability factors were present, became depressed. Concerning the presumed dangerousness of psychiatric patients, Steadman et al. (1998) reported that psychiatric patients who did not abuse drugs or alcohol had the same rates of violence as their non-patient neighbors, and that, if violence occurred, it was directed toward family and friends rather than strangers. A longstanding debate concerns the interpretation of the finding of a higher prevalence of mental disorders in the lower social classes and in deprived city areas, where the 'social causation' and the 'social selection/drift' hypotheses rival for explanation (Fox 1990). Recently Löffler and Häfner's (1999) study in Germany supported the social drift theory, while Fox (1990) had concluded earlier that there was very little, if any, empirical support for social selection/drift processes in serious mental illness. The epidemiological approach to suicide using published statistics has been a topic of much controversy since Durkheim's (1897) seminal study. Douglas (1967) has denied the usefulness of this approach, and suggested that the only scientific topic worthwhile studying is the social meaning of suicide, i.e., the motives of those who kill themselves. A recent review on the sociological literature on suicide is provided by Stack (2000).

Typical problems on the microsocial level which have been studied by sociologists (partly in conjunction with psychiatrists) are: Stigma, quality of life and satisfaction, illness behavior, life events, the emotional climate within the family ('expressed emotion') and its relation to relapse in schizophrenia, social support, coping, compliance and the process of caregiving for the chronic mentally ill, homelessness, and the work environment. Phelan et al. (2000) found that people with psychosis are more feared today than they were in the 1950s. Markowitz (1998) studied the effects of

stigma on the psychological well-being and life satisfaction of persons with mental illness, and Link et al. (1987) looked into the effects of stigma on recovery. Rosenfield (1992) found that the single most powerful influence on the improvement of subjective quality of life of the chronic mentally ill was the improvement of the individual's actual power in terms of economic status. Katschnig et al. (1997) have reviewed the concept of quality of life in relation to mental disorders, and Jackson et al. (2001) have summarized research on predictors of patient satisfaction. Mechanic (1995) relates these issues to the concept of illness behavior, arguing that the disability process is specifically relevant to promoting function and maintaining patients' quality of life, especially in primary care settings, where patient motivation should be increased, helpful attitudes encouraged, coping strategies taught, family members and employers educated, and support provided.

Finally, services research as well as studies which evaluate the functioning of psychiatric services in terms of quality control (Scheid and Greenley 1997) and mental health policy papers (Mechanic 1994) are further examples of sociological involvement in psychiatric issues.

4. Conclusion

As this review has shown, there are numerous points of contact between sociology and psychiatry, and each can learn much from the other. This is especially true at a time when the focus of psychiatry is moving further back into the community, and consumer participation and a 'civil dialogue' are becoming more important. Qualitative research methods, which seek to explain quantitative findings, generate hypotheses and ensure that questionnaires are relevant, are an appropriate tool for accompanying this process (Chard et al. 1997). Also, it might be useful to develop a sociology of the relationship between sociology and psychiatry. This review could serve as modest point of departure for such an undertaking.

See also: Culture as a Determinant of Mental Health; Epidemiology, Social; Health and Illness: Mental Representations in Different Cultures; Mental Health: Community Interventions; Mental Illness, Epidemiology of; Psychiatry, History of; Psychoanalysis: Overview; Public Psychiatry and Public Mental Health Systems; Social Intervention in Psychiatric Disorders

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition. Washington, DC

Brenner H 1973 *Mental Illness and the Economy*. Harvard University Press, Cambridge, MA

Brown G W, Harris T 1978 *The Social Origins of Depression: A Study of Psychiatric Disorder in Women*. The Free Press, New York

Brown P 1995 Naming and framing: The social construction of diagnosis and illness. *Journal of Health and Social Behavior* (extra issue): 34–52

Brown P, Funk S C 1986 Tardive dyskinesia: Barriers to the professional recognition of a iatrogenic disease. *Journal of Health and Social Behavior* **27**: 116–32

Busfield J 1986 *Managing Madness*. Hutchinson, London

Busfield J 1996 *Men, Women and Madness: Understanding Gender and Mental Disorder*. Macmillan, London

Busfield J 2000 Introduction: Rethinking the sociology of mental health. *Sociology of Health and Illness* **22**: 543–58

Caudill W 1958 *The Mental Hospital as a Small Society*. Harvard University Press, Cambridge, MA

Chard J A, Lilford R J, Court B V 1997 Qualitative medical sociology: What are its crowning achievements? *Journal of the Royal College of Medicine* **90**: 604–9

Cumming E, Cumming J 1957 *Closed Ranks: An Experiment in Mental Health Education*. Harvard University Press, Cambridge, MA

Curson D A, Pantelis Chr, Ward J, Barnes Th R E 1992 Institutionalism and schizophrenia 30 years on. *British Journal of Psychiatry* **160**: 230–41

Dallaire D, McCubbin M, Morin P, Coehen D 2000 Civil commitment due to mental illness and dangerousness: The union of law and psychiatry within a treatment-control system. *Sociology of Health and Illness* **22**: 679–99

Douglas J D 1967 *The Social Meanings of Suicide*. Princeton University Press, Princeton, NJ

Durkheim E 1950 *The Rules of the Sociological Method* [1895 French original]. Free Press of Glencoe, Glencoe, IL

Durkheim E 1951 *Suicide* [1897 French original]. Free Press of Glencoe, Glencoe, IL

Faris R E, Dunham H W 1939 *Mental Disorders in Urban Areas*. University of Chicago Press, Chicago, IL

Forster R 2000 The many faces of deinstitutionalisation—a discussion of sociological interpretations. *Psychiatric Praxis* **27** (Suppl. 2): 39–43

Foucault M 1965 *Madness and Civilization* [French original 1961]. Random House, New York

Foucault M 1977 *Discipline and Punish: The Birth of the Prison* [French original 1975]. Pantheon, New York

Fox J W 1990 Social class, mental illness, and social mobility: The selection-drift hypothesis for serious mental illness. *Journal of Health and Social Behavior* **31**: 344–53

Freidson E 1970 *Profession of Medicine*. Harper & Row, New York

Goffman E 1961 *Asylums*. Doubleday, New York

Goodwin S 1997 *Comparative Mental Health Policy: From Institutional to Community Care*. Sage, London

Hollingshead A B, Redlich F C 1958 *Social Class and Mental Illness*. Wiley, New York

Jackson J L, Chamberlin J, Kroenke K 2001 Predictors of patient satisfaction. *Social Science and Medicine* **52**: 609–20

Jones K 1960 *Mental Health and Social Policy 1845–1959*. Routledge & Kegan Paul, London

Katschnig H, Freeman H, Sartorius N 1997 *Quality of Life in Mental Disorders*. Wiley, Chichester, UK

Kessler R C, McGonagle K A, Zhao S, Nelson C B, Hughes M, Eshleman S, Wittchen H U, Kendler K S 1994 Lifetime and

12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**: 8–19
Langner T, Michael S T 1963 *Life Stress and Mental Health*. The Free Press, New York
Link B, Cullen F 1989 A modified labelling theory approach to mental disorders, an empirical assessment. *American Sociological Review* **54**: 400–23
Link B, Cullen F, Frank J, Wozniak J 1987 The social rejection of ex-mental patients: Understanding why labels matter. *American Journal of Sociology* **92**: 1461–500
Löffler W, Häfner H 1999 Ecological pattern of first admitted schizophrenics in two German cities over 25 years. *Social Science and Medicine* **49**: 93–108
Markowitz F 1998 The effects of stigma on the psychological well-being and life satisfaction of persons with mental illness. *Journal of Health and Social Behavior* **39**: 335–47
Mechanic D 1978 *Medical Sociology*, 2nd edn. The Free Press, New York
Mechanic D 1994 Establishing mental health priorities. *Milbank Quarterly* **72**: 501–14
Mechanic D 1995 Social dimensions of illness behavior. *Social Science and Medicine* **41**: 1207–16
Mirowsky J, Ross C E 1989 Psychiatric diagnosis as reified measurement. *Journal of Health and Social Behavior* **30**: 11–25
Parsons T 1951 *The Social System*. Free Press, Glencoe, IL
Pilgrim D, Rogers A 1999 *A Sociology of Mental Health and Illness*. Open University Press. Philadelphia, PA
Phelan J C, Link B G, Stueve A, Pescosolido B A 2000 Public conceptions of mental illness in 1950 and 1996: What is mental illness and is it to be feared? *Journal of Health and Social Behavior* **41**: 188–207
Robins L, Regier D A 1991 *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study*. Free Press, New York
Rosenberg C E 1992 Introduction. In: Rosenberg C E, Gooden J (eds.) *Framing Diseases. Studies on Cultural History*. Rutgers University Press, New Brunswick, NJ
Rosenfield S 1992 Factors contributing to the subjective quality of life of the chronic mentally ill. *Journal of Health and Social Behavior* **33**: 299–315
Rosenhan D 1973 On being sane in insane places. *Science* **178**: 250–58
Scheff T J 1999 *Being Mentally Ill: A Sociological Theory*, 3rd edn. [1st edn. 1967]. De Gruyter, New York
Scheid T L, Greenley J R 1997 Evaluations of organizational effectiveness in mental health programs. *Journal of Health and Social Behavior* **38**: 403–26
Scull A 1979 *Museums of Madness*. Penguin, Harmondsworth, UK
Srole L, Langner T S, Michael S T, Opler M K, Rennie T A C 1962 *Mental Health in the Metropolis: The Midtown Manhattan Study*. McGraw-Hill, New York
Stack S 2000 Suicide: A 15-year review of the sociological literature. Part I: Cultural and economic factors; Part II: Modernization and social integration perspectives. *Suicide and Life Threatening Behaviour* **30**: 145–62, 163–76
Star S A 1955 *The Public's Ideas About Mental Illness*. National Opinion Research Center, Chicago, IL
Steadman H J, Mulvey E P, Monahan J, Robbins P C, Appelbaum P S, Grisso T, Roth L H, Silver E 1998 Violence by people discharged from acute psychiatric inpatient facilities and by others in the same neighborhoods. *Archives of General Psychiatry* **55**(5): 393–401
Straus R 1957 The nature and status of medical sociology. *American Sociological Review* **22**: 200–4
Van der Geest S 1995 Overcoming ethnocentrism: How social science and medicine relate and should relate to one another. *Social Science and Medicine* **40**: 869–72
Wing J K, Brown G W 1961 Social treatment of chronic schizophrenia: A comparative survey of three mental hospitals. *British Journal of Psychiatry* **107**: 847–61
Wing J K, Brown G W 1970 Institutionalism and schizophrenia. A comparative study of three mental hospitals. Cambridge University Press, London

H. Katschnig

Psychiatry: Anthropological Aspects

Psychiatry—if we think for instance of psychoanalysis and of medical anthropology—has had an essential influence on philosophical anthropology, determining humankind in its world and its being-with-others from the perspective of manifold risks and disturbances. However, this article is concerned with exactly the opposite relationship, namely that of anthropology to psychiatry (see Blankenburg 1980a, Kraus 1999, Lang 1982).

What mental illness is in its various manifestations and which methods of investigation and therapies are appropriate for mentally ill patients from the aspect of their humanity, is also a question of anthropology as a comprehensive framework. A particular anthropology determines object theory to a high degree, that is, how mental illness can become objective, as well as the methodological accessibility of this object in psychiatry. Anthropological assumptions do not only determine norms usually applied in psychiatry to differentiate between mental illness and health but also determine ethics in psychiatry and address the relationship between body and soul and problems of free will and of self-determination. Because psychiatry as a science and therapeutic practice often latently implies an anthropology, by which psychiatry is deeply determined, psychiatry cannot avoid reflecting these premises. Above all, this is valid for the question of the adequacy of its methods for the respective object. A particular anthropology, however, does not only contribute to the development of what in psychiatry can become the object and its methodological accessibility, but it can also even prevent this development. Thus, certain results of research can be enabled, but can also be prevented.

As yet, it has not been sufficiently realized that psychiatry can be restricted in its scientific and therapeutic possibilities by its anthropology and can even be endangered by dogmatic and overgeneralized

basic assumptions. Problems of reductionism, however, of nothing-but theories, of the methodological preconditions of experience, are increasingly acknowledged. What Jaspers (1965) said about philosophy is also valid for anthropology. Stressing its significance for psychiatry, it does not only free the inner space for new possibilities of knowledge, but can also open horizons more adequate to a better understanding of the patient. A switching between anthropological and empirical research (Mundt and Spitzer 1999) has decisive significance for innovation and thus for the development of psychopathology as well as of psychiatry in its therapeutic and rehabilitative practice.

Ever since the separation of natural and cultural sciences we have differentiated between anthropology in the sense of natural science and philosophical and cultural anthropology. Whereas scientifical or biological anthropology, identical with physical anthropology or human biology, is concerned with the biological nature of humankind, philosophical anthropology makes statements about our personal or spiritual essence. Cultural anthropology, mostly interchangeable with ethnology, is concerned with the question of which information is given by cultural statements about what is variable and constant in human beings. To what extent is it possible to start with a previously given essence of human beings and in which way does the human being as the 'nondetermined animal' (Gehlen 1966) determine what he or she is. This question became one of the main topics of existential philosophy, where this human self-determination became the most important feature of a person's essence.

Immanuel Kant ([1798] 1978) already differentiated between a physical and a pragmatic anthropology. Whereas the first is concerned with what humans are by nature, the second, identical with philosophical anthropology, asks what they can make of themselves. Kant's philosophy became important for psychiatry not only because he developed interesting ideas about mental illness, seeing its general features in a loss of communal sense (*sensus communis*), that is, the loss of the ability of a person to be corrected by others (§ 53, p. 117), but especially because he has influenced psychiatry with respect to his methodology. Showing the dependency of experience on *a priori* ideas and thus overcoming the traditional splitting of subject and object, Kant's philosophy ([1781/87] 1965) stood for an opposition of idealism to realism which maintains the independence of the recognizing subject from the object. This opposition has had a constant effect not only on many philosophical schools of thought but also on psychiatry. With his concept of the transcendental subject he paved the way for Husserl's phenomenology and for existential approaches of philosophy. These new concepts and methods on their part made possible approaches like *Daseinsanalyse* (existential analysis), existential psychiatry, and cognitive approaches. Because above all in psychoses the comprehension of the world by the experiencing subject has changed, these psychiatric approaches claim a better understanding of the cognitive processes in these patients.

The following authors of a philosophical anthropology have become important for psychiatry: J. G. Herder (1744–1803), emphasizing the human as a being gifted with language; G. W. F. Hegel (1770–1831) with his historical philosophical critique of anthropology; M. Scheler (1874–1928), the founder of modern anthropology in his work *The Place of Man in the Universe*; H. Plessner (1892–1985), who speaks of the eccentric position of human beings, who have only to make out of themselves what they are; and A. Gehlen (1904–67) with his determination of humans as deficient beings because of lack of adjustment to their surroundings and reduction of instinct. At the same time an openness to the world corresponds to this lack. For a psychiatry which calls itself anthropological, the following philosophers were above all influential: S. Kierkegaard (1813–55) who initiated the development of existential philosophy; E. Husserl (1859–1938), the main founder of the phenomenological approach in philosophy; M. Heidegger (1889–1976) with his *Daseins* analytic; J. P. Sartre (1905–80) with existential psychoanalysis; and M. Merleau-Ponty (1908–61) with his phenomenology of the body. But on the other hand there has been an influence of medicine, especially of psychiatry, on philosophical anthropology: here we could mention Heinroth (1773–1843), V. von Weizsäcker (1886–1957), and M. Binswanger (1881–1966).

Multiple influences on psychiatry also arose from biological anthropology (including paleoanthropology, human genetics, research on human bodily constitution), from social anthropology and cultural anthropology, from cognitive and evolutional anthropology. Social anthropology and role theory (see Kraus 1980) had a strong impact on the development of social psychiatry and antipsychiatry followed by many psychopathological concepts and therapeutic measures.

Cultural anthropology examining the cultural variation of the human being as well as human constants independent of culture and epoch shows very clearly the double significance of the concept of anthropology already mentioned above. Showing the variability of the actual being and behavior of human beings, at the same time it marks the boundaries of anthropology as a universally valid theory of the essence of humankind. Cultural anthropology has cultivated a special approach of a transcultural or cultural psychiatry which investigates the variability and frequency of symptoms of disturbance entities, nowadays defined in glossaries, as well as of these entities themselves in different cultures. It has also led to the question of whether regarding the classical entities of disturbances like schizophrenia and manic-depressive illness we are actually dealing with entities of natural diseases or if

these are entities only coming from concepts of understanding. Transcultural or cross-cultural psychiatry is thus fundamentally involved in the contemporary crisis of nosology in psychiatry.

With the notion of a cognitive anthropology, Emrich (1999) sums up different approaches of a psychological, physiological (Singer 1994, Roth 1994), and philosophical constructivism. In cognitive anthropology the leading idea is the general quality of constructivity ascribed to systems generating consciousness (Roth 1994). The progress of modern evolutional biology in the future could enable a phylogenetic and ontological understanding of mental disturbances.

Heinroth can be looked at as the founder of an anthropological medicine in its true sense in the context of psychiatry, not to be confused with a medical anthropology. Many anthropological approaches in general medicine, especially those of Weizsäcker and Wyss, have also had effects on psychiatry.

Janzarik (1988) has presented an approach of his own to anthropological psychiatry. In his structural-dynamic concept he sees psychopathological phenomena in their neurobiological (dynamic) foundation as coherently connected with biographical structurally bound significances.

Under the heading of a phenomenological-anthropological psychiatry, different approaches can be summarized which have a phenomenological method as well as an anthropological orientation in common. These psychiatric approaches of an eidetic phenomenology of essence and a constitution-phenomenology as well as of *Dasein* and existence analysis relate to different authors of philosophical anthropology in its broader sense such as E. Husserl, M. Heidegger, J. P. Sartre, and M. Merleau-Ponty.

Even if the understanding of phenomenology and phenomenological in these different approaches is not always the same, the notion of phenomenon aims at that 'which-shows-itself-in-itself' and as such is different from the notion of the symptom which points to something which does not just show itself directly, namely to the illness underlying it. An approach strongly oriented to the phenomenon enables us to describe the original individual subjective experience as well as the respective lifeworld of the patient more adequately and differentiated than otherwise possible.

'Anthropological' signifies on one hand what is common to all people and on the other hand what in contrast to the animal is specifically human. In the context of anthropological psychiatry with this term the relationship of psychiatric disturbances to basic human structures is meant, as for instance consciousness and subjectivity, intersubjectivity and individuality, history and temporality, freedom, corporality and so on. These approaches not only take into consideration human essence because of ethical or humanistic reasons, but above all of methodological reasons. The aim is to understand psychiatric disturbances as regular variations of these basic structures, that means of human essence itself. This is done not so much to investigate the factual conditions for the development of psychopathological phenomena but to investigate the conditions originating from human essence which made them possible. For instance hallucination can be understood out of the normal act of perception as an uncoupling of spontaneous imaginative acts from a normal interaction with receptivity (Blankenburg 1980a, 1980b).

Binswanger's (1955) approach with 'anthropological proportion' has turned out to be especially fertile in psychiatry. The partly oppositional dimensions of height and breadth in human existence, of self-realization and world-realization, of individuation and relationship to community, of need for continuity and openness to innovation, of tolerance and intolerance of ambiguity, of overidentification and underidentification normally stand in a certain proportion to each other. These proportions are shifted in psychiatric disturbances. Relating to this alteration of the respective proportions some pictures of illness are oppositional. Thus an overidentification with their respective being in melancholics can be opposed to an underidentification in hysterics (Kraus 1996a). Overidentification with social role for example, with environment, etc., can understandably precipitate a melancholic phase in the case of their loss. One speaks of an empty-nest depression if a parent becomes depressed when children leave home. Other examples are a depression in the situation of retirement or of moving house. Hysterics, on the contrary, tend rather not to identify with their respective being or their actual feelings. They are striving for being more and something different than they actually are or to show feelings they have not. The abnormal as well as the normal can in this way be located in a polar field which is structured through ideal norms which oppose one another in a field in which 'deviations' from at least two sides are possible, as Blankenburg (1980a) put it. The pathological element can in this way be traced back to healthiness. On the other hand healthiness shows itself as something which always holds the possibility of becoming pathological by an alteration of proportions.

Such a concept of mental disturbances offers a special possibility of identification with the patient in the therapeutic process. Beyond a purely descriptive psychopathology in the sense of Jaspers (1965) it is thus possible to develop a dynamic phenomenology of psychiatric disturbances which has a dialectic character. Jaspers in his subjective psychopathology emphasized the special experience of the patient as opposed to a mere behavioral description in the sense of an objective psychopathology.

An eidetic essence-oriented phenomenology described above goes a step further in trying to grasp the essence of altered experiences and behaviors in a so-

called eidetic reduction in the sense of Husserl. In a process of free imaginative variation here one tries to get from a fact to the essence of a fact, from a sensual to a categorical presentation. Categorical presentation means for instance to recognize a particular species of tree in an actual tree, or in different kinds of behavior and experiences of hysterics the element of under-identification, of the pretence of being instead of actual being. Such an eidetic reduction can also be carried out in relation to Heidegger's notion of '*In-der-Welt-sein*' or Merleau-Ponty's '*Etre-au-monde*' showing the special relationship of patients to themselves and to the world which is expressed in their experiences and behaviors. A constitutive or transcendental phenomenology in the sense of Husserl is to be differentiated from an eidetic phenomenology. Instead of going from the fact to the essence of a fact as in eidetic phenomenology, constitutive phenomenology examines the constitutive achievements of consciousness, which let us perceive the concrete as it appears to us.

This approach of a constitutive phenomenology and the following philosophy of existence became important for psychiatry because in the perception of psychotic patients—but not only for them—the world and its significances constitute themselves in another way than in healthy people. In psychiatry this has resulted not only in making the experience of objects, for instance classical symptoms, into research topic, but also the preobjective experience, that is, the constitutive processes out of which self and world develop.

The preobjective experience is especially important in the intuitive diagnostics, for instance in the so-called precox feeling. Such preobjective experiences are for instance, the spatialization, temporalization, intersubjectivity, and corporality of a patient. Classical symptoms such as delusion, hallucination, phobic anxieties, and obsessive-compulsive symptoms assessed by a psychopathology oriented to the object, turned out to be constructed on such preobjective experiences.

Phenomenology of constitution, as Husserl put it, recently entered into a manifold relationship with cognitive theory and cognitive neuroscience (Wiggins and Spitzer 1997, Gallagher 1997). Varela (1996) even proposed a neurophenomenology. Several authors of the phenomenological-anthropological approach showed how in the contents of delusion and hallucination vs. the psychotic tendency of dissolution of their world, patients try again to give sense to their new world, even if a distorted one (Feldmann and Schmidt-Degenhardt 1997, Müller-Suur 1980, Kraus 1994a).

Husserl's concept of an intersubjectively constituted lifeworld became of increasing significance for psychiatry, above all for the understanding of schizophrenic modes of existence (Callieri and Castellani 1981, Blankenburg 1979). People with schizophrenia are characterized by a damaged rootedness in the lifeworld, connections of significances and relevancies becoming confused, which shows itself in extravagance and oddity but also in the delusions of these patients. Blankenburg (1971) speaks of a lack of natural self-evidence in schizophrenics. Whereas schizophrenics often lack an inner standard which would allow them a behavior appropriate to the social norms, melancholics being too adapted to these norms are too strongly fixed to the common sense and to their lifeworld (Blankenburg 1979, Stanghellini 1997, Kraus 1996b). This is already shown premorbidly by their hypernomic behavior which is characterized by a lack of distance to normative expectations (Kraus 1996b). This is also shown by the fact that they find their identity above all in their social roles. For this reason they are especially in danger of becoming depressed in the case of the loss of a social role (Kraus 1996b).

Binswanger has developed his existential analysis above all under the influence of Heideggers *Dasein* analytic in *Being and Time* (Heidegger 1927), which, against the ontologic intention of the author, Binswanger understood as an anthropology. With existential analysis we are concerned with a hermeneutic of the inner life history which is asked for its underlying individual concept of world which determines the concrete experience and behavior of a person.

Sartre's (1962) existential psychoanalysis has also contributed to the better understanding of certain pathological modes of existence, such as those of hysterics, melancholics, and patients with anxiety and obsessive-compulsive disorders. In this way we can establish certain existential types which can have an impact on the classification of these disorders (Kraus 1994b).

The phenomenological-anthropological orientation in psychiatry has promoted a series of therapeutic measures, for instance the anthropological-integrative psychotherapy of Wyss (1982) and existential analysis in the sense of Boss (1951) and Condrau (1963). But it has also influenced other therapies.

Summing up we can say that all anthropological approaches in psychiatry have as their aim to bring psychiatric disturbances in contact with the whole of human reality to achieve better understanding of those disturbances in a way that is appropriate to the essence of humankind. Considering anthropological prepositions in this way allows us to reflect the fundamentals of psychiatry in research as well as in therapy by which the phenomenological-anthropological approaches contribute to further development.

See also: Anthropology, History of; Cognitive Anthropology; Health and Illness: Mental Representations in Different Cultures; Mental Health and Normality; Mentally Ill: Public Attitudes; Mind–Body Dualism; Psychological Anthropology

Bibliography

Beck A 1970 The core problem in depression: The cognitive triad. In: Massermann J H (ed.) *Depression. Theories and Therapies*. Grune & Stratton, New York

Binswanger L 1955 Vom anthropologischen Sinn der Verstiegenheit. In: *Ausgewählte Vorträge und Aufsätze*. Franke, Bern, Germany, Vol. 2, pp. 235–42

Binswanger L 1992 Drei Formen missglückten Daseins. In: Herzog M (ed.) *Ludwig Binswangers ausgewählte Werke*. Asanger, Heidelberg, Germany, Vol. 1, pp. 1–443

Blankenburg W 1971 *Der Verlust der natürlichen Selbstverständlichkeit*. Enke, Stuttgart, Germany

Blankenburg W 1979 Phänomenologische Epoché und Psychopathologie. In: Sprondel W M, Grathoff R (eds.) *Alfred Schütz und die Idee des Alltags in den Sozialwissenschaften*. Enke, Stuttgart, Germany, pp. 125–39

Blankenburg W 1980a Anthropologisch orientierte Psychiatrie. In: Peters U W (ed.) *Ergebnisse für die Medizin, 2: Psychiatrie. Die Psychologie des 20. Jahrhunderts*. Kindler, Zürich,Vol. 10, pp. 182–97

Blankenburg W 1980b Phenomenology and psychopathology. *Journal of Phenomenological Psychology* **11**: 50–78

Boss M 1951 Beitrag zur daseinsanalytischen Fundierung des psychiatrischen Denkens. *Schweiz Archives of Neurology and Psychiatry* **67**(1): 15–19

Callieri B, Castellani A 1981 On the psychopathology of the lifeworld. In: Bello A A (ed.) *Analecta Husserliana*. Reidel, Dordrecht, Germany, pp. 172–202

Condrau G 1963 *Daseinsanalytische Psychotherapie*. Huter, Bern, Germany

Emrich H M, Schiefenhovel W 1999 Philosophische Anthropologie als Grundlagenwissenschaft der Psychiatrie. In: (ed.) *Psychiatrie der Gegenwart*, 4th edn. Springer, Heidelberg, Germany, Vol. 1

Feldmann H, Schmidt-Degenhard M 1997 Strukturale Affinität des Unverständlichen im schizophrenen Wahn. *Nervenarzt* **68**: 226–30

Gallagher S 1997 Mutual enlightment: Recent phenomenology in cognitive science. *Journal of Consciousness Studies* **4**(3): 195–214

Gehlen A 1966 *Der Mensch*, 8th edn. Athenäum, Frankfurt, Germany

Heidegger M 1963 *Sein und Zeit*. Niemeyer, Tübingen, Germany

Janzarik W 1988 *Strukturdynamische Grundlagen der Psychiatrie*. Enke, Stuttgart, Germany

Jaspers K 1965 *Allgemeine Psychopathologie*. Springer, Berlin

Kant J [1781/87] 1965 *Critique of Pure Reason* (trans. Kemp Smith N). St. Martin's Press, New York

Kant J [1798] 1978 *Anthropology* (trans. Dowdell V L, rev. and ed. Rudnick H H). Southern Illinois University Press, Carbondale, IL

Kraus A 1980 Bedeutung und Rezeption der Rollentheorie in der Psychiatrie. In: Peters U H (ed.) *Ergebnisse für die Medizin, 2: Psychiatrie. Die Psychologie des 20. Jahrhunderts*. Kindler, Zürich,Vol. 10

Kraus A 1994a Phenomenology of the technical delusion in schizophrenics. *Journal of Phenomenological Psychology* **25**(1): 51–69

Kraus A 1994b Phenomenological and criteriological diagnosis: Different or complementary. In: Sadler J, Wiggings O P, Schwartz M (eds.) *Philosophical Perspectives on Psychiatric Diagnosis and Classification*. Johns Hopkins University Press, Baltimore, MD, pp. 148–62

Kraus A 1996a Identitätsbildung Melancholischer und Hysterischer. In: Seidler G H (ed.) *Hysterie Heute*. Ferdinand Enke, Stuttgart, Germany, pp. 103–10

Kraus A 1996b Role performance, identity structure and psychosis in melancholic and manic-depressive patients. In: Mundt C, Goldstein M J, Hahlweg K, Fiedler P (eds.) *Interpersonal Factors in the Origin and Course of Affective Disorders*. Gaskell, London, pp. 31–47

Kraus A 1999 Phänomenologisch-anthropologische Psychiatrie. In: Helmchen H, Henn F, Lauter H, Sartorius N (eds.) *Grundlagen der Psychiatrie*. Springer, Berlin

Lang H 1982 Phenomenologisch-anthropologisch orientierte Psychiatrie und Psychologie im französischen Sprachraum. In: Eicke D (ed.) *Tiefenpsychologie, Vol. 4 Individualpsychologie und Analytische Psychologie*. Weinheim, Basel, Switzerland

Müller-Suur H 1980 *Das Sinn-Problem in der Psychose*. Hofgreve, Göttingen, Germany

Mundt C, Spitzer M 1999 Psychopathologie heute. In: (ed.) *Psychiatrie der Gegenwart*, 4th edn. Springer, Heidelberg, Germany, Vol. 1, pp. 4–44

Roth G 1994 Gehirn und Geist. In: Schiefenhövel W, Vogel C, Vollmer G, Opolka U (eds.) *Vom Affen zum Halbgott*. Thieme, Stuttgart, Germany

Sartre J P 1962 *Das Sein und das Nichts*. Rowohlt, Hamburg, Germany

Singer W 1994 *Gehirn und Bewusstsein*. Spektrum, Heidelberg, Germany

Stanghellini G 1997 For an anthropology of vulnerability. *Psychopathology* **30**: 1–11

Varela F 1996 Neurophenomenology. A methodological remedy for the hard problem. *Journal of Consciousness Studies* **4**: 330–49

Wiggins O P, Spitzer M 1997 Cognitive science. In: Embree L, Behnke E A, Carr D et al. (eds.) *Encyclopedia of Phenomenology*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 101–4

Wyss D 1982 *Der Kranke als Partner*. Vandenhoeck und Ruprecht, Göttingen, Germany, Vols. 1 and 2

A. Kraus

Psychiatry, Evidence-based

No country can afford health services to meet all the possible needs of the population, so it is advisable to establish criteria for which services to provide. Two basic criteria are the size of the burden caused by a particular disease, injury or risk factor, and the cost-effectiveness of interventions to deal with it (Bobadilla et al. 1994).

Health services are rationed. Access in most countries is controlled by money, waiting lists, and by the clinician's decision about what should be done. Health services are concerned to avoid the explicit rationing that would follow if they adopted the position outlined above and so they have become concerned with maximizing efficiency. They have done this by limiting

the input of funds, and by using scarcity to encourage improvements to the process of care. They keep asking 'can science inform practice?' an issue beset by many questions and too few answers. Yet it does seem relatively simple. Research is concerned with discovering the right thing to do; clinical audit with ensuring that the right things are done. The link that can close the gap is said to be evidence-based medicine: 'the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients' (Sackett et al. 1996).

The evidence-based medicine movement seems to have originated from Archibald Cochrane's set of essays 'Effectiveness and Efficiency: Random Reflections on Health Services,' (Cochrane 1998) first published in 1971 and reprinted by BMJ Books in 1998. He argued that no new treatments should be introduced unless they had been shown in randomized controlled trials to be more effective than existing treatments, or equally effective and cheaper; and that one should evaluate all existing therapies, slowly excluding those shown to be ineffective or too dangerous. Cochrane actually meant efficacy when he wrote effectiveness (he acknowledged hating the word efficacy) but it is now usual to distinguish the three terms: efficacy, does the treatment work; effectiveness, does it work in practice; and efficiency, is it a wise expenditure of funds? The popular press is preoccupied with efficacy (can the doctors cure disease X?), not with effectiveness (do they?) or with efficiency (can we afford the cost?). Evidence-based medicine is concerned with identifying the efficacy of treatments and encouraging their implementation. The goal is increased effectiveness and efficiency (Andrews 1999).

Efficacy: 'Does the treatment work?' means does it produce benefit reliably over and above that to be expected from just being in treatment? Randomized placebo-controlled trials are valuable because they control for the uplifting effect of being in treatment (the true placebo effect). They also control for changes in symptoms and disability attributable to spontaneous remission, including the remission that occurs because people with chronic and fluctuating disorders commonly delay consulting until their disorder is particularly severe and can be expected to improve whatever happens. In some disorders the extent of this overall 'placebo response' is very large. In depression, it is two to three times as large as the change related to the present specific therapies, be they drugs or psychotherapy (see the discussion in Enserink 1999, pp. 238–40). That is, people given a sugar tablet will improve, not because of the tablet, but because of these other processes. People taking the active antidepressant tablet will improve more, but only one third of the overall change will be due to the antidepressant. It is for this reason that changes in symptoms with a specific treatment are suspect, unless they are shown to exceed the changes expected from a nonspecific placebo treatment.

With a few significant exceptions, most randomized controlled trials in psychiatry, even multicentre trials, have relatively small numbers of subjects. It is therefore usual to defer judgment until a number of trials have been done in different centers. If they agree that a certain treatment is better than placebo, then it probably is, but as most use different measures of outcome it has been necessary to develop a method for estimating the average degree of improvement with this treatment. Meta-analysis employs standard data analysis techniques to compare the results of studies by converting, if necessary, the results onto a standard metric, called an effect size. Even so, the wide range of outcome measures in use makes comparison difficult, and it is not always adjusted by expressing change in effect size terms. For example, while trials in depression almost always use the Hamilton scale so that results can simply be averaged, a meta-analysis of panic disorder found over 300 different measures in 58 studies, which has to be evidence of social irresponsibility in science. While each investigator will need to use a disorder-specific measure of symptoms, there is some real advantage in all using a generic measure of disability such as the short forms from the RAND medical outcomes study.

Structured systematic reviews are now common. In memory of Cochrane, a center at Oxford has been established to encourage and disseminate systematic reviews of the medical literature. These reviews abide by strict methodological guidelines and identify the level of confidence that the reader can place on any finding. In mental health, the dominant US publication is a book by Nathan and Gorman (1998). Any clinician, asked by a client what other treatments there are, would be obliged ethically to mention the treatments listed in that book as supported by type 1 evidence before offering advice about the treatment to be preferred. The self-help version of that book should be in every mental health clinician's waiting room (Nathan et al. 1999).

Effectiveness: Does the treatment work in practice when applied by the average clinician to the average client? There is a paucity of such data. Such data as there are show that effectiveness is reduced as one moves from experimental practice to routine clinical practice. Too often treatment as usual seems almost ineffective. Thus the problem is not so much defining the treatments that work as overcoming the degradation of benefit that occurs in practice. There are four issues that undermine the potential effectiveness of a treatment: deficits in coverage, deficits in compliance, deficits in clinical competence, and treatment integrity. Epidemiological studies in the USA, Canada, the UK, and Australia show consistently that only a fifth to a half of people who meet criteria for a mental disorder get help for their disorder (Andrews and Henderson 2000). Coverage is therefore low, variably limited by access to services and by the expectation that there is little point in attending. Poor

coverage when access is limited is understandable, but why should there be poor coverage when a national health service provides access? The Australian national mental health survey asked why the burden of depression persisted when access to care was free. Half the people who identified depression as their main complaint (and were disabled as a result) did not consult because they 'had no need for treatment.' There are data on the mental health literacy of Australians that show that the majority of the population think that relaxation and tonics are good for depression, and that medication is not. It is clear why they do not consult their doctors.

Deficits in compliance and treatment integrity are difficult to estimate, simply because auditing of either will produce a temporary change in behavior. There is evidence that asking people if they have taken all their medication is a valid measure of compliance; at least those who say they haven't can be counseled. One advantage of providing patient treatment manuals for psychological treatments is that the manuals contain explicit sheets on which the homework can be recorded and reviewed. Like homework at elementary school, if the clinician doesn't ask, the client doesn't do. The patient treatment manual also serves to ensure treatment integrity. After all, the manual is the curriculum for the treatment program and clinicians have little opportunity but to follow the framework, personalizing it for each patient, and following the sequence in every case. All psychological therapies should be manualized, because psychotherapy is about learning, and most learning programs have text books.

Coverage and compliance concern the system and the consumer. Clinical competence and treatment integrity concern the physician. How do you make sure the physician knows what to do, and then does what should be done. The evidence-based medicine movement has clear ideas about how clinicians should practice. Ideally, clinicians should constantly be asking 'what is the evidence about treatment for this type of patient with this type of disorder?' Searching the Internet they should source systematic reviews, such as those from the Cochrane Collaboration, before deciding which treatment is appropriate. There are three problems. The first is time; busy practitioners, especially those in private practice, have no opportunity to consult sources in this way. The neediness of patients simply consumes the day. The second is that systematic reviews are about people who satisfied the inclusion and exclusion criteria for RCTs, and who are unlike the pregnant woman, for whom information is needed. The third issue is that the methodological formalism required for such reviews may remove the very data or comments that would be helpful. Systematic reviews are essential, but they may need to be supplemented with more clinically relevant material.

Review journals are flourishing. For example, the aim of the British Medical Association's journal *Evidence Based Mental Health* was to print one-page summaries of research that had met rigorous standards. Clinicians, with no time to do their own literature searches, could rely on the summaries to guide practice. To a large part this happens, but methodological formalism may be limiting the scope and obscuring meaning. In January 2000, the journal's website had a summary of an RCT in which three sessions of CBT and three sessions of psychodynamic therapy were equally effective in relieving subsyndromal depression. The take-home message for the busy practitioner was unclear. Psychotherapy that requires transference can be completed in three meetings? CBT that teaches how to change thoughts and behaviors can be achieved in three sessions? That it is important to treat mild symptoms that do not constitute a disorder? Which is strange in a country like the UK, in which 70 percent of people with major depression do not see a doctor and most who do are treated inadequately. The review journals are important, but the editors need to think about the burden of the disease and the implications for practice, as well as about the probity of research designs. Complex problems are unlikely to have simple one-page answers.

Clinical practice guidelines are the traditional avenue used to put research findings into practice. Properly written they can combine a continuum of information from good research to good clinical wisdom, differentiating between statements supported by research, and statements, no less valuable, supported by opinion. The steps in preparing clinical practice guidelines are illustrated in Fig. 1. In summary: when there is a clinical issue and there are no existing guidelines, a panel of all stakeholders should be convened to identify the desired health outcome. They should proceed by preparing a systematic review of efficacy of interventions in relation to this outcome, differentiating clearly between interventions supported by the differing levels of the weight of evidence (level 1a: meta-analysis of randomized controlled trials; level 1b: at least one RCT; level 2: at least one controlled or quasi-experimental study; level 3: non-experimental descriptive studies; level 4: expert committee reports or opinions). Each statement in the guideline is qualified traditionally in accord with the quality of the supporting research evidence. Increasingly, the principal recommendations of guidelines require support from level 1a evidence, lower levels of evidence being valuable to fine-tune that principal recommendation for special groups of patients, or atypical clinical situations.

Most clinicians have access to at least half a dozen such guidelines, and many doctors in primary care have mountains of guidelines, so many that their ability to use any is questionable. While most originators claim that regular revisions will be issued, this has not proven to be the case. The problem is that guidelines have largely been written by advocates keen to support a position and, understandably, they are

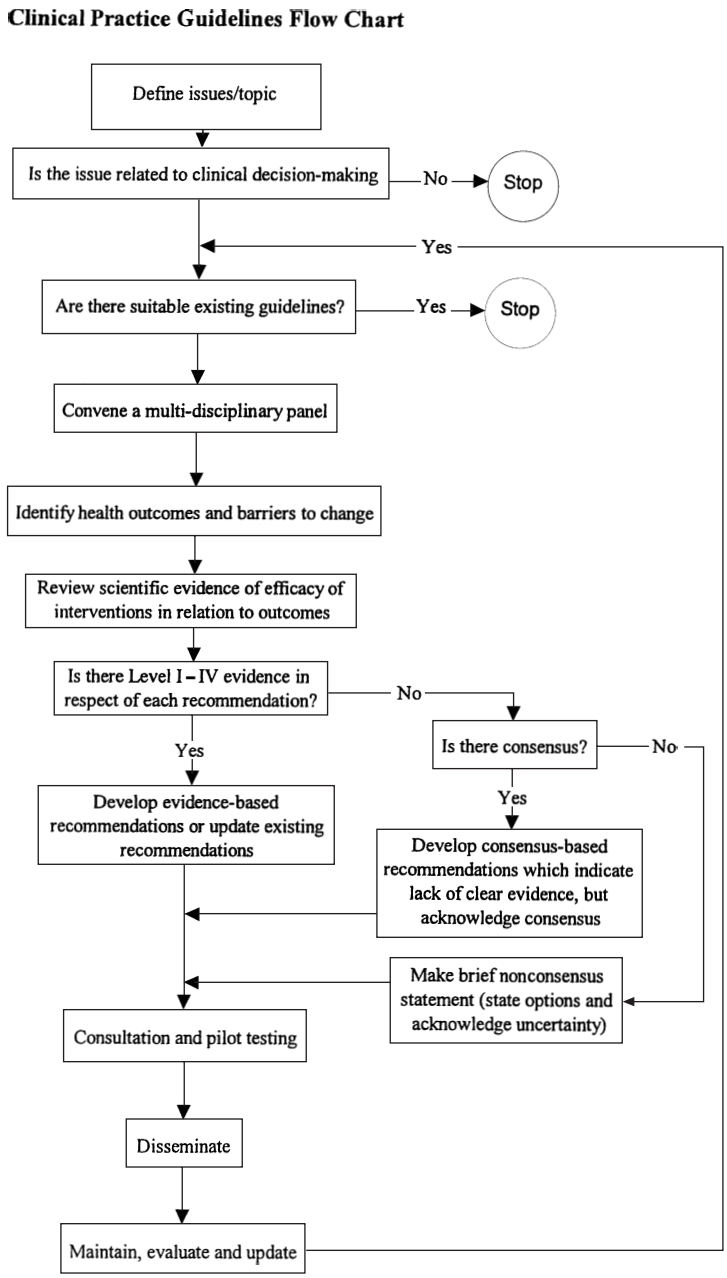

Figure 1
Clinical practice guidelines flow chart

not keen to update the data in terms of new developments. The Royal Australian and New Zealand College of Psychiatrists produced guidelines for 1981–91 (Quality Assurance Project 1983), and in 2001 they at last began working on a revision of some of these. The American Psychiatric Association and the UK Royal College of Psychiatrists have also produced guidelines for their members. An attempt to produce mental health consensus guidelines that incorporated the views of all stakeholders—the various providers, the consumers and carers, and the public and private funders—appears to have run out of support: who wants to be fair when there is so much to be gained by being partisan? Moreover, when guidelines are prod-

uced by independent bodies such as the Agency for Health Care Policy and Research they can run foul of vested interests so that the impact is nullified (cf. breast cancer screening, spinal fusion) (Deyo et al. 1997). Clinical practice guidelines should be a flowering of the scientific method—find out what works and implement it—but too often they are political acts that represent the interests of an advocacy group.

Leaving aside the problem of the accuracy of the information, how is it possible to introduce guidelines that busy practitioners will follow. The Australian National Health and Medical Research Council recently convened a meeting to learn how 14 studies in medicine, surgery, and psychiatry fared as they sought to implement their guidelines. None fared well. Simply publishing the guidelines had no effect at all. The key elements seemed to be to produce guidelines that were simple, short, accurate, and current. Short meant on two sides of a plastic card, even if backed by a source book that contained the science. The content should be sponsored by a respected professional organization and the guidelines should represent a collaboration between all the stakeholders likely to use them, or be affected by them. They gain credence by being published in a reputable journal.

The rules for successful implementation of clinical practice guidelines are not at all clear, but look like a combination of hard work and marketing, done at a local hospital or regional professional organization level. *Getting Research Findings into Practice* (Haines and Donald 1998) is an excellent source book. The process is as follows: involve the key stakeholders and the clinical and management opinion leaders, have an accurate idea of what one wants to change, and a clear idea of the changes to organizational procedures, medical equipment, and clinical skills that will be necessary if implementation is to occur. Provide academic detailing to the clinicians so that all will have a chance to discuss and question the guidelines before implementation starts. Involve the consumers in the preparation of the guidelines and then in their implementation. Put the guidelines on the hospital or group's homepage, and 10 percent of consumers will consult these before attending. Use clinical audit to show that change is occurring and to show improved clinical outcomes. If the change does not improve clinical outcomes, ask whether further implementation is wise. Sustain the whole process by showing an improved cost effectiveness profile and arrange that some of the savings benefit the clinicians who brought this about. Expect that the implementation procedure will cost a considerable amount in staff time; the time costs for staff to attend meetings to produce a simple change to the way acute asthma was managed in one hospital exceeded $50,000. The costs of implementation were perceived as inordinate even though that money was quickly recouped from the savings that accrued from the improved outcomes once the guidelines were implemented. If originators of the guidelines are not prepared to do all this work then they should simply enjoy the process of producing them, for the guidelines will not influence practice. This is not to deny that identifying empirically supported treatments carried out by a profession does not have important professional and political advantages for the profession. Funders, providers, and consumers all like to pretend that efficacy is the same as effectiveness, and lists of empirically supported treatments feed this delusion.

Efficiency: Efficiency or cost effectiveness is about the wise expenditure of scarce health resources. Costing studies are now common; the problem is that most are based on efficacy calculations in which coverage, compliance, and treatment integrity are all believed to be perfect. As effectiveness is much, much less than efficacy, the true costs per unit of benefit are much, much more than those identified in the costing studies. Australia was the first country to require costing data to be submitted when pharmaceutical companies sought to have a new drug placed on the list of subsidized medications. It has rightly been praised for this. Enquiries as to whether the data submitted are cost effective data or cost efficacy data are met with the assurance that the former is the case. As the data are 'commercial—in confidence' they are not amenable to 'freedom of information' enquiries. None are published and one might suspect that most are estimates of cost effectiveness under ideal conditions in which effectiveness is presumed to equal efficacy. As psychological therapies are not subsidized they are not subject to regulation, and randomized controlled trials about efficacy and cost are not required by any country. If we are to be anything more than a medicated society we do need such data.

League tables of the 'cost effectiveness' of medical interventions are beginning to appear. Few include mental health interventions. The median medical intervention cost in a study of 500 life-saving interventions was estimated as $19,000 per life year saved. In mental disorders, the principal burden comes from the years lived with a disability, not from the years of life lost. Nevertheless, shadow prices under $30,000 per disability-adjusted life year averted or quality-adjusted life year gained are said to be of interest to health planners. In specialized clinics, where costs and outcomes are known, it appears that drug and psychological therapies for the common anxiety and depressive disorders can be provided at little more than $5,000 per life year gained. These figures are competitive and it might prove more advantageous if accurately costed league tables of empirically supported therapies in mental disorders were to be prepared by an organization of high repute. I cannot think of a greater stimulus to the implementation of evidence-based psychiatry than some appreciation of the good bargains available. After all, treatment as usual that produces little benefit is very costly to both funder and consumer. The providers benefit in the

short term, making the majority of their income from people who do not get better, but in the long term they become disillusioned by their lack of effectiveness. Evidence-based psychiatry can close the gap between research and practice, but it is not a trivial task. It can improve effectiveness to resemble efficacy and so result in better outcomes for patients, and that is not a trivial achievement. Isn't that what mental health services are for?

See also: Behavior Therapy: Psychiatric Aspects; Clinical Treatment Outcome Research: Control and Comparison Groups; Cognitive and Interpersonal Therapy: Psychiatric Aspects; Health Outcomes, Assessment of; Psychiatric Care, Economics of; Psychological Treatment, Effectiveness of; Psychological Treatments: Randomized Controlled Clinical Trials; Truth and Credibility: Science and the Social Study of Science; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Andrews G 1999 Efficacy, effectiveness and efficiency in mental health service delivery. *Australian and New Zealand Journal of Psychiatry* **33**: 316–22

Andrews G, Henderson S (eds.) 2000 *Unmet Need in Psychiatry*. Cambridge University Press, Cambridge, UK

Bobadilla J L, Cowley P, Musgrove P, Saxenian H 1994 Design, content and financing of an essential package of health services. In: Murray C J L, Lopez A D (eds.) *Global Comparative Assessments in the Health Sector*. World Health Organization, Geneva, pp. 171–80

Cochrane A L 1989 *Effectiveness and Efficiency: Random Reflections on Health Services*. British Medical Journal, London

Deyo R A, Psaty B M, Simon G, Wagner E H, Omenn G S 1997 The messenger under attack: Intimidation of researchers by special-interest groups. *The New England Journal of Medicine*. **336**: 1176–80

Enserink M 1999 Can the placebo be the cure? *Science* **284**: 238–40

Haines A, Donald A 1998 *Getting Research Findings into Practice*. BMJ Books, London

Nathan P E, Gorman J M 1998 *A Guide to Treatments that Work*. Oxford University Press, New York

Nathan P E, Gorman J M, Salkind N J 1999 *Treating Mental Disorders: A Guide to What Works*. Oxford University Press, New York

Quality Assurance Project 1983 A treatment outline for depressive disorders. *Australian and New Zealand Journal of Psychiatry* **17**: 129–46

Sackett D L, Rosenberg W M C, Gray J A M, Haynes R B, Richardson W S 1996 Evidence based medicine: What it is and what it isn't. *British Medical Journal* **312**: 71–2

G. Andrews

Psychiatry, History of

Attempts to heal madness go back in Western medicine at least to Hippocrates. However psychiatry as a medical specialty—that is, a gainful form of medical practice that occupied its practitioners full-time and to the exclusion of other branches of healing—and, hence, as a self-conscious scientific project, was the invention of the late eighteenth century. It had, at the level of ideas, two major prerequisites: a new technique for treating madness, called the moral treatment, believed to effect cure more reliably than anything that had preceded it; and a new institution, the asylum, conceptualized both as the appropriate setting for the cure of madness and as itself an active instrument of cure. While the name eventually given to the specialty in most countries was of German coinage (*Psychiatrie*, first used in 1808 by Johann Christian Reil), the practical initiatives critical in establishing psychiatry came from England and France.

1. Moral Treatment

Psychiatry's founding therapeutic strategy was of combined Anglo-French provenance and as such became the subject of a muted, international priority dispute. Cursorily defined, the moral treatment for the cure of insanity meant the use of methods that engaged or operated on the intellect and emotions, as opposed to the traditional methods of bleedings, purges, and emetics applied directly to the lunatic's body. By all accounts, the British cleric and physician Francis Willis had employed moral methods when in 1788–9 he assumed the weighty responsibility of treating the madness of King George III. His contemporary, the French physician Philippe Pinel, duly credited Willis and others with having inspired his own efforts. But when codifying the moral treatment in his *Traité médico-philosophique sur l'aliénation mentale* (1801), Pinel also accused his British colleagues of willfully failing to spell out the method they so enthusiastically recommended. For their part, British reviewers of Pinel's *Traité* dismissed the Frenchman's supposed innovations as pale imitations of British practice.

Wherever priority lay, the moral treatment proved notoriously resistant to precise definition. Contrary to his avowed intention, Pinel never provided an explicit definition in his *Traité*. He instead presented a series of case histories of his own use of the treatment, leaving to his readers the task of inferring its central features. These included:

(a) A preference for 'gentleness' as a therapeutic technique. This choice was rooted in an optimistic Enlightenment belief that insanity never involved a complete annihilation of reason, reducing the mad person to animality, but that a glimmer of 'humanity'

always remained and could be appealed to and reinforced.

(b) The use of 'repression' only as a last resort and then without anger or cruelty. When acting in a repressive capacity, the healer of the insane was expected to exude authority through his physical presence—Willis, for example, was known for his ability to command the obedience of the mad by his piercing gaze.

(c) The grounding of the treatment in an etiology of insanity based on the sensationalist psychology of John Locke and the abbé de Condillac. Regarded as a set of erroneous ideas that had become implanted firmly in the mind, insanity could be cured by a jolting experience that dislodged those ideas and enabled their replacement by sound ones. Hence the theatrical aspects of Pinel's moral treatment: vivid scenes were staged deliberately to impress on mad persons the falsity of their pathological beliefs (Goldstein 1987, Porter 1987).

Clearly, such a therapeutic strategy stamped psychiatry as a humanitarian endeavor and a product of the Enlightenment faith in science. In France, these affiliations were strengthened further by the link between psychiatry and the French Revolution. The Revolutionary government put Pinel in control of the Paris institutions for the insane, and according to a myth that took shape over the course of the nineteenth century, and has only recently been debunked by historians, he inaugurated the moral treatment by literally unchaining the inmates of Bicêtre, an emancipatory gesture that cured them instantaneously. In the USA, the chief proponent of the moral treatment, Benjamin Rush, was so ardent a supporter of the American Revolution that he even figured among the signatories of the Declaration of Independence (Binger 1966). In Britain, ratification of the moral treatment took place in the context of parliamentary investigations into the civil liberties violations entailed in the maltreatment of incarcerated lunatics (Scull 1975).

But if the moral treatment had strong associations with Enlightenment and political liberty, it also lent psychiatry an unscientific taint that proved difficult to eradicate. For, as its proponents liked to point out, it had originated among healers without official educational credentials. Pinel was a bona fide physician but, in keeping with the hostility of the French Revolution toward all privileged corporations, he proudly announced that neither he nor any other member of a medical faculty had devised the moral treatment, that it was the work of a humble layman who served as hands-on caretaker of the insane at Bicêtre. His own contribution was, he said, to have rendered the caretaker's version of the treatment scientific, translating it into the language of sensationalist psychology and demonstrating its efficacy with the new mathematical tool of statistics. Reliance on the moral treatment left Pinel's successors vulnerable to competition from traditional Catholic healers of the insane, who argued persuasively that medical men had no special skills in administering a noncorporeal remedy. Indeed in France the dispute between alienists and clerics for control of the treatment of insane continued until the end of the nineteenth century, when the anticlerical policies of the Third Republic decided conclusively in favor of the former (Goldstein 1987).

In Britain, it was a layman, the tea merchant William Tuke, who instituted the moral treatment at the York Retreat, the influential Quaker facility for the insane. Tuke was highly vocal about his distrust of medical practitioners pretending to treat insanity with Physick. After a long period of uneasy debate between the models of medical and moral therapy, the official line of British psychiatry became in the 1840s that a judicious combination of both had greater efficacy than the sole use of either. In adopting this position, British psychiatrists shrewdly obviated the protest that clergymen were better equipped than physicians to handle insanity (Scull 1975).

2. *The Asylum*

The insane asylum is one of a group of institutions, which, in the 1960s, the American sociologist Erving Goffman (1961) labeled 'total' and, in the 1970s, the French philosopher Michel Foucault (1975) described as 'panoptic.' For Goffman, such institutions are characterized by locating the ordinarily compartmentalized activities of sleep, play, and work under a single roof and a single authority and by erecting barriers, both legal and architectural, to social intercourse with the outside world. For Foucault, they are characterized by the incessant visibility of the inmates to the authorities, a visibility redounding to the power of the latter. Both theorists stressed the tight schedule of activities enforced in such institutions. Both noted their proliferation in the West during the late eighteenth and early nineteenth centuries, the period that saw the birth of prison and poorhouse as well as insane asylum.

In itself, the incarceration of deviant populations was no novelty. At least since the 'Great Confinement' of the mid-seventeenth century, western European governments, especially the absolutist monarchies of the continent, had made a concerted effort to establish good order by removing troublemakers from social circulation. The novelty lay rather in an unprecedented confidence in the rehabilitative power of certain forms of incarceration and in a fresh taxonomic sensibility that led to the division and strict segregation of deviants according to type. By contrast, the French *hôpital général* of the age of Louis XIV and its counterparts, the British workhouse and the German *Zuchthaus*, had entertained no therapeutic ambitions and had lumped together madness, idleness, and

debauchery as fully comparable species of unreason (Foucault 1961).

The classical scientific rationale for the asylum was provided in 1819 by Pinel's student J. E. D. Esquirol, who also appropriated the old term 'asylum' (*asile*), meaning a place of refuge, to underscore the benevolent ethos of the new institution. The asylum could cure insanity, said Esquirol, because it temporarily removed mad persons from their habitual surroundings. By dint of this 'isolation,' as he called it, the institution operated as an instance of the moral treatment: the psychic jolt of sensory stimuli arising from a radically altered physical and interpersonal environment dislodged the patient's erroneous ideas, which could then be replaced by rational ones through the expert interventions of the psychiatric doctor. The latter's therapeutic authority was enhanced by the fact that, in the artificial, self-contained world of the asylum, he ruled absolutely (Goldstein 1987).

Like the embrace of the moral treatment, so, too, the wedding of psychiatry to the institutional form of the asylum imposed special professionalization problems on the emergent specialty. In order to survive collectively, psychiatrists needed the support of a governmental or private agency willing to make large capital investments in the construction of asylums and motivated to place the psychiatrist at the top of the asylum's hierarchy of personnel. These goals were first achieved in France, where Esquirol and his disciples joined with moderate liberal politicians to secure passage of the Law of 1838. The Law mandated the establishment of a nation-wide network of asylums headed by psychiatrists, who thereby became salaried functionaries of the French state. In the less centralized British state, psychiatric victory came more gradually. An Act of 1808 providing for the construction of asylums financed by the county rate made no provision for the inclusion of a medical man on the staff. Only with the passage of the Lunatics Act of 1845, which placed institutions for the insane under a permanent national inspectorate composed of lay, legal, and medical commissioners, did the British introduce even an implicit force for the medicalization of the asylum (Jones 1955, Scull 1975).

A combined impulse toward rational state-building and enlightened humanitarianism led the disunified German lands of the first half of the nineteenth century to a spate of new-style, public asylum foundings as well as the medical restructuring of old-style, multifunctional absolutist institutions of incarceration (Goldberg 1999). In the USA, the asylum-building vogue of the same period occurred in two phases: first, emulation of European models in new, privately funded eastern-seaboard institutions; then, goaded by the Jacksonian-era critique that the frenetic competition for wealth and power in democratic America actively bred insanity, the establishment of public asylums by most state legislatures. American asylums were run by medical superintendents—the early name for psychiatrists in the USA—and were so plentiful that an Association for Medical Superintendents was formed in 1844 (Rothman 1971).

3. *Late Nineteenth-century Reconfigurations*

The second half of the nineteenth century saw important changes in the institutional and intellectual models that had launched psychiatry. Most dramatically, the prevailing optimism about the therapeutic effect of the asylum gave way to deep pessimism as asylum populations swelled with chronic patients demonstrably impervious to moral treatment (Lantéri-Laura 1972, Scull 1979, Rothman 1971). At about the same time, psychiatrists abandoned their earlier conception of insanity as a curable psychological or psychosomatic disorder and theorized it as an irreversible brain condition and often as a product of 'degeneration.' This degeneration was defined as a pathological departure from the norm initially caused by a noxious environment, poor nutrition, or alcoholism, and subsequently transmitted in the Lamarckian manner through heredity, becoming more severe with each generation. Every European nation had its *fin-de-siècle* theorists of degeneration: Bénédict-Augustin Morel and Valentin Magnan in France, Cesare Lombroso in Italy, Henry Maudsley in Britain, Richard von Krafft-Ebing and Max Nordau in Austria (Pick 1989).

The shift to an organicist conception of insanity was coupled with a shift in psychiatry's international center of gravity from France to Germany. By the mid-nineteenth century, German physicians had thoroughly repudiated an earlier affinity for Romantic, holistic *Naturphilosophie*. Their new taste for positivism was epitomized in psychiatry by Wilhelm Griesinger's textbook, *Pathologie und Therapie der psychischen Krankheiten* (1845) that equated psychiatry with the study of cerebral lesions. Griesinger also epitomized the era's new psychiatric career pattern: he was not an asylum doctor but a university psychiatrist, having been appointed in 1865 to the recently created Chair of Psychiatry and Neurology at the University of Berlin (Decker 1977). Later adopted in other countries, the university psychiatry model increased the attractiveness of the psychiatric career, offering the practitioner a less reclusive lifestyle than was afforded by residence in an asylum. It also changed the intellectual content of the career: psychiatric knowledge was now assumed to be generated less from clinical contact with the insane than from neuroanatomical research in a university laboratory.

A second late nineteenth-century reconfiguration likewise encompassed both psychiatry's scientific theory and its career structure. Asylum psychiatry had worked with a simple binary opposition between sanity and insanity. After 1850, however, psychiatrists began to speak of a large intermediate zone occupied by the *demi-fous*, or half-mad. This conceptual move

eventually led to the articulation of two categories fundamental to the enterprise of twentieth-century psychiatry: the 'psychotic,' whose contact with reality is severely ruptured, and the 'neurotic,' whose minimal maladaptations do not preclude getting on in ordinary society. The former, requiring institutionalization, fit the older model of asylum psychiatry. The latter, who could safely remain in the outside world but would benefit from scheduled visits to the doctor's office, made possible the practice of outpatient psychiatry.

A rough international division of labor governed research on the neuroses and psychoses. The Parisian Jean-Martin Charcot and his many students dominated the study of the less severe forms of mental illness, especially hysteria. Munich-based Emil Kraepelin and his school continued German psychiatric preeminence through their domination of the study of the psychoses. Kraepelin made a psychiatric staple of the diagnostic category, dementia praecox, a state of emotional detachment and vacuousness that declared itself in adolescence and carried an invariably bleak prognosis. (The Swiss psychiatrist Eugen Bleuler would in 1911 rename the condition schizophrenia and insist that it might have a favorable outcome.) The Kraepelinian approach held sway in many countries, including the USA, through the first quarter of the twentieth century.

4. *Twentieth-century Trends*

In broad outline, three trends distinguished psychiatry in the twentieth century: the emergence of Freudian psychoanalysis; the discovery and use of psychotropic drugs; and the antipsychiatry movement.

Created by the Viennese neurologist Sigmund Freud as an outpatient treatment for neurotic disorders, psychoanalysis was initially distant from institutional psychiatry. Indeed, Freud's professional overtures to Bleuler and Carl Jung, Bleuler's assistant at the Burghölzi asylum in Zurich, can be read as efforts to obtain a psychiatric imprimatur for the new discipline of psychoanalysis (Freud 1914). Freud never fulfilled this ambition in Europe, but in the less tradition-bound scientific environment of the USA, many professors and medical superintendents welcomed his ideas. Eventually, American psychiatry fully incorporated psychoanalysis. In 1924, when William Alanson White, who had introduced psychoanalytic psychotherapy for schizophrenic patients at St. Elizabeth's Hospital, became president of the American Psychiatric Association, that organization and the American Psychoanalytic Association began holding their annual meetings together. This arrangement lasted until 1974, by which time the Freudian ardor of American psychiatrists had cooled (Barton 1987).

Waning interest in psychoanalysis on the part of American psychiatrists coincided with rising expectations for psychopharmacology on the part of the international psychiatric community. The first psychotropic drug, lithium, was used successfully to manage manic-depression in Australia in 1949. A variety of antipsychotic and antidepressant compounds, most notably the phenothiazines and imipramine, were developed in France and Switzerland in the early 1950s (Baldessarini 1977). The fluoxetine family of antidepressants, whose first member went under the brand name Prozac, was introduced in the USA in the late 1980s and became among the most widely prescribed drugs in the world. Psychopharmacology brought a new wave of self-confidence and therapeutic optimism to psychiatry, promising a relatively safe, speedy, and cost-effective method of alleviating mental illness. It also revived psychiatry's identity as a 'hard' and exact science, the identity cultivated in the second half of the nineteenth century.

Somewhat ironically, then, the reliance on psychotropic drugs was one of the causes of the antipsychiatry movement of the 1960s. Since the beginning of the specialty, psychiatry's incarcerative practices had given rise to sporadic criticism, and some psychiatric abuses, notably the 'euthanasia' program mounted by the Nazi regime during World War II, have only been fully uncovered after the fact (Burleigh 1994). But the antipsychiatry movement of the 1960s, centered on the London-based psychiatrist R. D. Laing, questioned the specialty even more radically and globally. It charged that madness was neither an objective biochemical nor behavioral condition, but a context-dependent label; that Western rationalism was predicated on a denial of its value; and that psychopharmacology robbed patients of the potentially enriching and self-healing experience of full-blown psychosis. Laing's views were supported by the psychiatrist Thomas Szasz in the USA and to some degree by the philosopher Michel Foucault in France. Antipsychiatry gave impetus to the deinstitutionalization of the insane, which occurred in the late 1970s and the 1980s. A less controversial outcome of the movement was the recognition of the patient's right to refuse antipsychotic drugs and electroshock therapy (Sedgwick 1982).

The antipsychiatry movement has now largely spent its force. But as the twenty-first century begins, psychiatry, while among the oldest of the medical specialties, lacks the stability that age might seem to confer. It remains hostage to the mind–body problem, buffeted between psychological and physical definitions of its object and its methods.

See also: Deinstitutionalization, Psychiatry of; Mental Health and Normality; Philosophy, Psychiatry, and Psychology; Psychiatric Care, Economics of; Psychiatry and Sociology; Psychiatry: Anthropological Aspects; Psychoanalysis, History of; Psychoanalysis in Clinical Psychology; Psychoanalysis in Sociology; Psychologism, History of; Psychology: Overview; Psychometrics; Psychophysics; Psychotherapy, History of: Psychiatric Aspects; Social Psychology

Bibliography

Baldessarini R J 1977 *Chemotherapy in Psychiatry*. Harvard University Press, Cambridge, MA

Barton W E 1987 *The History and Influence of the American Psychiatric Association*. American Psychiatric Press, Washington, DC

Binger C 1966 *Revolutionary Doctor: Benjamin Rush, 1746–1813*, 1st edn. Norton, New York

Burleigh M 1994 *Death and Deliverance: 'Euthanasia' in Germany 1900–1945*. Cambridge University Press, Cambridge, UK

Decker H S 1977 *Freud in Germany*. International Universities Press, New York

Foucault M 1961 *Histoire de la Folie*. Plon, Paris [English translation 1965 *Madness and Civilization: A History of Insanity in the Age of Reason*. Random House, New York]

Foucault M 1975 *Surveiller et Punir: Naissance de la Prison*. Gallimard, Paris [English translation 1977 *Discipline and Punish: The Birth of the Prison*. Pantheon, New York]

Freud S 1914 On the history of the psycho-analytic movement. *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London, vol. 14

Goffman E 1961 *Asylums: Essays on the Social Situation of Mental Patients and Other Inmates*, 1st edn. Anchor Books, Garden City, NY

Goldberg A 1999 *Sex, Religion and the Making of Modern Madness: The Eberbach Asylum and German Society, 1815–1849*. Oxford University Press, New York and Oxford, UK

Goldstein J 1987 *Console and Classify: The French Psychiatric Profession in the Nineteenth Century*. Cambridge University Press, New York and Cambridge, UK

Jones K 1955 *Lunacy, Law and Conscience, 1744–1845*. Routledge & Kegan Paul, London

Lantéri-Laura G 1972 La chronicité dans la psychiatrie française moderne. *Annales ESC* **27**: 548–68

Pick D 1989 *Faces of Degeneration: A European Disorder, c. 1848–1918*. Cambridge University Press, Cambridge, UK

Porter R 1987 *Mind Forg'd Manacles: A History of Madness in England from the Restoration to the Regency*. Harvard University Press, Cambridge, MA

Rothman D J 1971 *The Discovery of the Asylum: Social Order and Disorder in the New Republic*, 1st edn. Little, Brown & Co., Boston

Scull A T 1975 From madness to mental illness: medical men as moral entrepreneurs. *Archives Européennes de Sociologie* **16**: 218–51

Scull A T 1979 *Museums of Madness: The Social Organization of Insanity in Nineteenth-century England*. St. Martin's Press, New York

Sedgwick P 1982 *Psycho-Politics: Laing, Foucault, Goffman, Szasz and the Future of Mass Psychiatry*, 1st edn. Harper and Row, New York

J. Goldstein

Psychiatry: Informed Consent and Care

Informed consent is a philosophical and legal doctrine that has come to serve as the cornerstone of ethically sound clinical care in Western society. This doctrine requires that individuals truly understand and freely make decisions regarding their treatment. Informed consent is an ideal that emphasizes respect for the individual and places primacy on autonomous decision-making. Furthermore, it is inherently relational, that is, it occurs in the context of a professional relationship in which the fundamental values of the medical profession find expression. Fulfilling the ideal of informed consent in practice is an elusive aim, however. As suggested by intuition, clinical phenomenology, and empirical evidence, it is particularly challenging when consent is sought from individuals who suffer from mental illness. For this reason, it is important to consider the specific issues arising in informed consent in the context of mental health care. This chapter provides a clinically-oriented overview of informed consent in mental illness by: (a) characterizing the core conceptual elements of informed consent; (b) applying informed consent concepts in mental health clinical practice; and (c) giving a brief overview of key empirical data related to mental illness in clinical consent decisions.

1. Elements of Informed Consent

Informed consent does not occur in isolation or in a moment, nor is it reducible to a signature on a legal document. It is a shared, significant process between people—one of whom faces a series of evolving choices critical to his or her mental and physical health, the other of whom is a professional person entrusted with helping to define these choices astutely and present them truthfully, respectfully, and compassionately. In clinical care, promoting the health and alleviating the suffering of the individual patient will be the basis of any treatment considered in the consent process. Consent in this situation focuses on the patient's personal values, is based on the overall goals for care, and hinges on trust in the clinician–patient relationship.

1.1 Information

The sharing of appropriate information is the first of three core elements of informed consent. The patient should be provided with accurate and balanced information in understandable language. This information should include the anticipated risks and benefits of a given treatment, the procedures involved, and its alternatives and their expected outcomes. Alternatives may include implementation of another treatment or additional treatment, or it may involve no intervention upon the course of illness at all. While every minute detail cannot be reviewed in this discussion, rare but serious risks should be characterized in a manner that conveys their relevance and reality. Furthermore, the omission of information in the disclosure process should not be intended to deceive or mislead the person undertaking the decision.

Ideally information related to a matter requiring consent will be disclosed in a manner that fosters dialogue, allows for clarification of complex issues and the identification of misunderstandings, and keeps opportunities open for subsequent discussion as new issues emerge. In obtaining informed consent for treatment of mild depression, for instance, the clinician should discuss the variable course of depressive illnesses, including the potential for symptom resolution in the absence of treatment and the chance of recurrence without intervention. The clinician should review the rationale for a proposed antidepressant agent given the symptoms of the patient and the patient's attitudes and values. The discussion should cover the medication's known possible side effects, potential adverse effects, and their relative likelihood. The clinician should explicitly anticipate the choices that will be faced by the patient in the future in light of the therapeutic alternatives discussed and the nature of the illness. This discussion should be documented by the clinician as an important element of the care process.

Consent forms that accurately and completely describe a treatment and outline patient rights represent an important tool in the consent process. Consequently, consent forms should be developed carefully in terms of their understandability (e.g., simple but precise language), readability (e.g., size of font), content (e.g., purpose, risks and benefits, rights, responsibilities of the clinician/researcher), and related issues. Consent forms, furthermore, signify the importance of a new decision for the patient, and they give specific focus to the dialogue between the professional and the patient. Even general consent forms such as those for hospitalization or for initiating medication—which may or may not confer legal protection when a medico-legal conflict occurs—have this value. From a legal perspective, notes written by the clinician reflecting the disclosure dialogue with the patient may be far more salient and can carry greater weight.

1.2 Decisional Capacity

Decisional capacity, the second element of informed consent, represents a sophisticated clinical assessment of an individual's global and specific cognitive abilities within a particular context and in relation to a particular choice. It is distinct from 'competence' which is a legal determination, i.e., a formal assessment by a judge or other officer of the court regarding a person's ability to function within a general domain of one's life such as financial or medical decisions (Appelbaum and Gutheil 1991). The presence of a serious symptom (e.g., delusions, memory loss, suicidal ideation, diminished cognitive functioning) or of a serious diagnosis (e.g., schizophrenia, dementia, major depression, developmental disability) does not mean that an individual lacks decisional capacity. Similarly, being a young adolescent or a very elderly person does not, *a priori*, preclude one from being decisionally capable. In the context of mental illness, decisional capacity assessment may be subtle, complex clinical work that entails constructing a cohesive sense of the individual's strengths and deficits in light of a number of factors. The type of decision to be made, the nature, time course, and severity of the person's symptoms, the patient's situation, and past experiences, personal values, psychological defenses or coping 'style,' and other factors should all be considered.

Decisional capacity itself has four principal components, each of which should be viewed on a continuum and not as 'all or none' phenomena (Appelbaum and Grisso 1988). The first component is the individual's ability to express a preference (i.e., capacity for expression or communication). On the extreme end of the continuum will be severely ill psychiatric patients who are mute or catatonic, or who manifest complete language and thought disorganization, and therefore may be intermittently 'incapable' of communication. In the medical context, a patient who is comatose due to metabolic disturbances, aphasic secondary to a stroke, delirious, or unconscious after a head injury will also lack this most fundamental ability. There are times when this component of decisional capacity is very difficult to assess, such as in the spinal cord injured patient who communicates only through eye blinks, or the dysphoric person with Broca's aphasia who is frustrated by the process of generating words but has the capacity to understand and indicate choices. For these reasons, while the ability to communicate is often obvious, this assessment in many situations will require careful interactions with patients and astute observational and listening skills on the part of the clinician.

The second component of decisional capacity is the individual's ability to understand relevant information (i.e., capacity for comprehension). In talking with the clinician, the patient with this ability will be able to take in factual information regarding the nature of the illness, the need for intervention, the kinds of treatments under consideration, their relative risks and benefits, and the possible consequences of not intervening. Comprehension is influenced by intelligence, educational level, and prior experience. In assessing this dimension, it is critical that the patient state his or her understanding of the situation and the choices at hand in their own words. Otherwise it is difficult to know if the patient is using terms and phrases of the clinicians without truly knowing their meaning.

Third is the individual's ability to think through the decision in a way that is 'rational' and deliberate (i.e., capacity for reasoning or rational thinking). This component of decisional capacity relates to the person's analytical abilities given a specific set of facts or data and, like comprehension, is affected by

intelligence and educational level. For instance, the 'rational' person, all other factors being equal, will consistently choose a medication that has a 75 percent likelihood of treating the illness effectively over one that has a 50 percent chance. Similarly, the 'rational' person will consistently choose a medication that poses fewer risks and promises fewer adverse side effects than another medication. In essence, this dimension of decisional capacity has to do with the ability to use objective information.

Fourth, and often most important, is the person's ability to make meaning of the decision and its consequences within the context of his or her personal history, values, emotions, and life situation (i.e., capacity for appreciation). Deeply dependent upon psychological insight, this is the most sophisticated component of decisional capacity. It can be very difficult to assess clinically. In some individuals with early dementia, for instance, a person's social graces and cognitive skills may be only superficially affected and yet these individuals may be unable to think through the meaning of an important treatment decision and its repercussions. In some patients suffering from either anxiety or affective disorders, the negative cognitive distortions inherent to their illnesses may interfere with their abilities to interpret information within a personal context, although other components of decisional capacity may be intact. Importantly, in people with severe mental illnesses such as schizophrenia, the simple presence of psychotic symptoms such as hallucinations or delusions may *not* interfere with the individual's overall decisional capacity, but the cognitive and insight deficits associated with the illness may interfere with their capacity for appreciation of the situation at hand. Alternatively, even the most gravely ill individuals may have difficulty expressing themselves and an incomplete sense of the factual information but may truly appreciate the severity of the situation and possess a deep and consistent sense of who they would like to make decisions for them.

Decisional capacity is therefore a rich construction, not simply reducible to an individual choosing a predetermined 'right answer' when making a decision of importance. It is the person's ability to gather, assimilate, and use information, to truly see the reasons for and against a given option, and to develop a meaningful decision in the larger context of one's life. The seriousness and complexity of any specific decision will govern how rigorously these criteria are applied. In the context of serious mental illness, a person may be fully capable of deciding whether or not to begin a new psychotropic medication but, on a given day, may be incapable of deciding whether to enroll in a clinical protocol involving a medication-free interval and potential randomization to placebo. The standard for decisional capacity is also robust when a very ill patient refuses treatment that is highly recommended because of the benefit it is likely to bring to him (see discussion in Sect. 2). This is not to say that patients cannot refuse treatment; when the risks of refusal are great, however, there is a greater responsibility on the caregiver to assure that the patient is genuinely capable of this decision and to manage its consequences.

1.3 Voluntarism

Beyond information disclosure and decisional capacity, the final element of informed consent is voluntarism or autonomy. Autonomy represents the expression of the authentic preferences of the individual in the absence of coercive pressures. Voluntary preferences may be tempered by the realities of the situation such as the presence of a serious diagnosis and the lack of 'curative' treatments for many mental illnesses, but should reflect the person's sustained and genuine wishes as evidenced in their expressed values and prior life decisions. Autonomous wishes may be deeply influenced by social and cultural expectations and psychological history. For instance, a patient who has been in the military may be oriented to hierarchical power relationships and may view a clinician's recommendation as an 'order' to be obeyed. Similarly, a seriously mentally ill person who has been institutionalized involuntarily in the past may see themselves as no longer possessing the right to their own choices. Illness processes may also affect patient preferences as well as the ability to express autonomous wishes. A person who suffers deeply either psychologically or physically may feel desperate and desolate, phenomena that can generate inherent internal coercive pressures. A person who has been the victim of violence may see themselves as relatively powerless. Finally, clinical syndromes that involve apathy, alexithymia, amotivation, or cognitive slowing may render ill individuals less able to identify and express their true wishes.

2. Clinical Application of Informed Consent Concepts

Translation of informed consent concepts to clinical practice is affected by the clinical status and situation of the patient. In emergencies, for instance, there may be little time for information-sharing and mutual decision-making and the standard for consent is less rigorous than in other contexts. Acting beneficently, i.e., to save the life or preserve the health of the patient in an acute situation, is the clinician's duty in such circumstances. This is held to be critical ethically and justifiable legally; it is presumed that in such emergency situations the patient wishes to be helped and that the clinician has permission to act with the patient's well-being and best interests at heart. Often the patient, if alert and conscious, will explicitly assent to treatment in these circumstances, i.e., voluntarily accept the choice proffered by the clinician at the time. If the patient is unconscious or is unable to com-

municate at all, the clinician acts with the 'implied consent' of the patient.

When the patient refuses treatment in an emergent situation, however, a very rigorous standard for consent must be applied because of the likelihood of an adverse outcome and because there are many clinical factors that may be adversely affecting the judgment of the patient. Undiagnosed illness, delusional thinking, severe physical pain, emotional distress, and interpersonal pressure all may undermine the ability of the patient to make a fully informed, autonomous choice in an acute situation. Evaluation, hospitalization, and treatment, even against the apparent wishes of the patient, may be essential in such circumstances. Consider two illustrative patient cases:

Case Example 1. A man diagnosed with a serious mental illness characterized by psychotic and mood symptoms was brought to a psychiatric crisis unit after being found wandering on the street near his home. It was discovered that he had superglued his eyelids shut, and it was imperative that he receive medical attention immediately. He was outwardly calm, did not speak, answered questions through nodding, writing, and hand gestures, but actively refused to get into an ambulance to be taken to the medical emergency room. The clinician on duty that evening knew the patient well and was able to deduce that the patient's peculiar behavior was linked to a recurrent delusion he had when entering a manic period. In this case, the patient's ability to make a fully informed decision was impaired by his delusional belief that if he looked directly at people, they would be instantly destroyed by powerful lasers that had been implanted in his eyes. Once the clinician initiated a conversation about these beliefs with the patient, it was possible to clarify that the patient wanted help and that he was willing to accept treatment, but was terrified that he would hurt people or himself be killed. The patient thus had some decisional ability and could offer assent for emergency care. The clinical team opted to pursue a legal process that would allow involvement of a time-limited treatment guardian due to concerns over the patient's compromised insight and judgment and the severity of his delusions.

Case Example 2. An agitated man demanded discharge from a surgical unit two days after an emergency appendectomy. He had elevated heart rate, blood pressure, and temperature, had signs of abdominal infection, and yet was insistent that he leave the hospital. The medical team was conflicted about granting the patient an 'AMA discharge' ('Against Medical Advice')—the nurses and residents believed this to be the patient's 'right to refuse treatment' but the attending clinician did not agree. On initial psychiatric evaluation, the patient showed irritability and very subtle indications of diminished attention, was oriented except to day of the week, and had no other apparent abnormalities on an abbreviated mental status examination. He was noted to be tremulous. He was placed on involuntary 'hold' status. Within an hour, the patient was agitated, disoriented, and expressed disorganized thoughts. Additional history was then obtained that the patient drank one to two pints of vodka daily prior to admission. Post-surgical infection in combination with alcohol withdrawal had triggered a significant delirium, an early indication of which, ironically, was his treatment refusal and discharge demand. His most distressing symptoms and disorientation resolved rapidly with treatment, and he withdrew his request to leave the hospital.

When a patient refuses treatment there may be a significant clinical issue that requires close attention, and involuntary treatment may be imperative. Clarification of underlying clinical considerations in such circumstances is the caregiver's primary responsibility—a duty grounded in the bioethics principles of respect for persons, clinical standards of care, fidelity, and beneficence. Careful evaluation of the patient's true strengths and abilities is important so that he or she may exercise choice in every area possible. The intent is not to undermine the patient's autonomy. Rather, it is to empower the clinician only insofar as to help advance the health and well-being of the patient and to assure that the patient is optimally able to express his or her true wishes to accept or decline a given therapeutic approach. In sum, in situations where clinical decisions may have very serious or irreversible health consequences, including death, informed refusal of treatment must fulfill the most stringent standards.

The goal of informed, voluntary decision-making may be more clearly fulfilled in the context of care for chronic illness. Over the course of time, the clinician may come to know the patient—their symptoms, their values and concerns, their responsiveness to treatment interventions, and their personal supports. In turn, the patient may have a greater sense of trust, express themselves with greater openness, and gather more complete information about the illness and potential therapeutic interventions. The patient with a chronic illness also lives out the consequences of side effects, adverse reactions, and benefits of treatments in a manner different from the patient with a time-limited health problem. In this context, the patient may better appreciate how their illness and treatment affect other aspects of their life, and it is possible to come to know their preferences and concerns with greater certainty. Expressing autonomous wishes and deciding upon specific treatment approaches and the overall goals of therapy through mutual understanding and dialogue are more intrinsically possible in the context of chronic illness. The standard for consent in chronic illness, thus, is very rigorous in terms of the patient's level of information, opportunities to enhance decision-making, and creating a circumstance in which patients may indicate their wishes and hopes.

In clinical practice, psychiatric advance directives may be valuable in translating concepts of informed

consent to patient care decisions. Advance directives articulate patient preferences for future care. Advance directives are often formal written papers created and signed by the patient; they may also consist of a body of documentation developed by the patient, family members, and clinicians regarding his or her wishes and values concerning personal health care. Advance directives are helpful in that they serve as a reference for the clinician if the patient becomes decisionally incapable or compromised. For example, knowing that they had a pattern of developing compromised judgment during periods of mania, a patient with bipolar disorder completed an advance directive requesting that treatment be instituted despite their (anticipated) protests. Advance directives also provide evidence of the patient's past personal healthcare preferences that may result in controversial clinical decisions, e.g., when patients refuse treatment based on deeply and consistently held personal convictions such as the refusal of blood transfusions by a patient who is a Jehovah's Witness. Ideally, these issues will not be addressed for the first time in a moment of great crisis. They should be anticipated, discussed, and documented carefully beforehand. Advance directives do not, under any circumstances, replace the wishes expressed by a patient who is currently decisionally capable. Moreover, there should be opportunities to adapt, or even reverse, earlier decisions as one cannot always tell prospectively how one will respond in a crisis situation. For these reasons, advance directives may guide but can never supplant prudent, reflective clinical decision-making.

In surrogate decision-making, healthcare choices are made for the decisionally impaired patient by another person. This person may be a spouse, sibling, adult child, friend, or a court-appointed guardian. The surrogate decision-maker has the responsibility of serving the best interests and well-being of the patient while seeking to make decisions that are faithful to the personal values and prior wishes of the patient. This can be difficult. In some situations, the patient's wishes are not known. In other situations, the patient is too young or too cognitively compromised to have demonstrated their personal values. In yet other situations, these values and wishes are known and appear to conflict with accepted standards of clinical care. Consequently, in working with surrogate decision-makers, it is important for clinicians to clarify the goals and value system governing the choices made on behalf of the patient. It is also crucial that the patient's decisional capacity be assessed accurately and, in most circumstances, the patient's symptoms be treated aggressively so that they can contribute to the choices made to whatever extent possible.

3. *Empirical Studies on Consent in Clinical Care*

Nearly 400 studies performed since 1970 suggest that informed consent is an ideal that is often difficult to achieve in clinical and research practice involving people with physical and mental illness (Sugarman et al. 1999). These data indicate, however, that issues faced in clinical care for mental illness are not radically different in degree or kind from those encountered in the treatment of physical illness (Stanley et al. 1988). Moreover, an extensive body of evidence on consent consistently demonstrates the poor readability, understandability, and ineffectiveness of formal consent documents in clinical care (Grossman et al. 1994). An important recent advance in the informed consent field has been the development of the MacArthur Competence Assessment Tool for Treatment (MacCAT-T) (Grisso et al. 1997). This is a structured interview designed to assess patients' clinical care decision-making abilities based on empirical work conducted over ten years by Appelbaum andN Grisso. While this instrument is best utilized as a tool for approaching decisional capacity evaluation, not as a definitive 'test' of competence, it will introduce greater consistency of consent standards to both scientific and clinical arenas.

Evidence on consent in the care of mental illness reveals many key challenges for clinicians (Roth et al. 1982). Studies of psychotic, affective, and dementia disorders, for instance, have shown that symptoms of mental illness may adversely impact information-based or more strictly cognitive aspects of consent (Roberts 1998). In the 1995 MacArthur Treatment Competence Study (Appelbaum and Grisso 1995), the decisional abilities of patients with schizophrenia or major depression, patients with ischemic heart disease, and non-patient community volunteers (total $n = 498$) were assessed. When acutely ill, the psychiatric patients in this study had greater difficulties with decision-making on several formal cognitive measures than did the medically ill patients and community participants. The problems were more severe for individuals with psychotic symptoms than for depressed individuals. Clinical treatment led to significant improvement of these deficits, resulting in closer resemblance across the three groups.

Smaller studies echo these findings. An early study by Schachter et al. (1994) involving 38 schizophrenia patients revealed that the presence of conceptual disorganization, delusions, and hallucinations was associated with patients' poor performance on the information elements of consent. Appelbaum et al. (1981) assessed 50 people admitted to a mental health inpatient unit and found that most were unable to consent to hospitalization. Later this team evaluated 100 voluntarily hospitalized psychiatric patients, finding that a significant minority did not understand information presented to them on the purpose of hospitalization and discharge (Appelbaum et al. 1998). Beck (1988) evaluated the capacities of 56 psychiatric patients on antipsychotic medications, revealing that they had limited understanding and appreciation of their treatment. Stanley et al. (1988) compared the

reasoning ability of elderly individuals suffering from major depression ($n = 45$) and dementia ($n = 38$) with elderly control subjects ($n = 20$). The cognitively impaired dementia patients and, to a lesser extent, the depressed patients had significant deficits on a number of decision-making items. Similarly, Marson et al. (1995) used a clinical vignette-based instrument to measure the level of competence exhibited in medical treatment decisions by individuals with dementia and healthy comparison subjects. Not unexpectedly, this work revealed that the severity of cognitive impairment in dementia correlated with lack of ability to formulate logical reasons for treatment choice, i.e., addressing the 'rational reasons' dimension of decisional capacity. Interestingly, global measures of dementia severity such as the Mini-Mental Status Exam were not predictive of clinical decision-making capacity.

Studies of advance directives and surrogate decision-making also have important implications for mental health care. Several survey and interview studies have revealed that advance directives are underutilized (Bradley et al. 1998) and that patient preferences are rarely identified prospectively (Seckler et al. 1991). In terms of surrogate decision-making, several studies have shown that most elderly patients would prefer that a family member such as a spouse or adult child serve as an alternative decision-maker (High 1989). Studies of hypothetical decision-making by proxies suggest that they may not assess patients' wishes accurately, however Sachs et al. (1994) studied 42 decisionally capable dementia patients, 64 proxy decision-makers, and 60 well elderly individuals and found that the decision-makers' predictions of patient preferences were overall discordant with the patients' own stated wishes. Similarly, Warren et al. (1986) studied participation choices made by proxy decision-makers and 168 decisionally capable nursing home patients with respect to hypothetical, minimal risk research projects. The results demonstrated that the surrogate decision-makers used criteria other than the patient's predicted preferences in making these choices.

Clear evidence is also emerging that the barriers to consent may be amenable to intervention, offering significant hope (Roberts et al. 2000). Interventions in which information is presented verbally and in writing, using clear and understandable language, have been shown to improve patients' understanding of consent decisions (Carpenter et al. 2000). The timing of consent information and the provision of information in a relaxed manner or on repeated occasions is important to patients' abilities to integrate and apply this data (Roberts 1998). Treatment of symptoms, ranging from hallucinations to physical pain, has been observed to improve the quality of patients' decision-making. Finally, clinician behaviors that reflect how the consent process is valued, e.g., by giving time to the consent dialogue, have been found to be valued by patients and surrogate decision-makers (Roberts and Roberts 1999). The observation that educational efforts, symptom treatment, and attitudinal obstacles can be addressed suggests a number of constructive approaches to consent in the context of clinical care for the mentally ill.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Civil Rights; Ethical Dilemmas: Research and Treatment Priorities; Ethics and Psychiatry; Ethics and Values; Ethics for Biomedical Research Involving Humans: International Codes; Free Will and Action; Fundamental Rights and Constitutional Guarantees; Placebos, Use of: Ethical Aspects; Privacy of Individuals in Social Research: Confidentiality; Psychotherapy: Ethical Issues; Research Conduct: Ethical Codes; Research Subjects, Informed and Implied Consent of; Responsibility: Philosophical Aspects

Bibliography

Appelbaum B C, Appelbaum P S, Grisso T 1998 Competence to consent to voluntary psychiatric hospitalization: A test of a standard proposed by APA. American Psychiatric Association. *Psychiatric Services* **49**: 1193–6

Appelbaum P S, Grisso T 1988 Assessing patients' capacities to consent to treatment. *New England Journal of Medicine* **319**: 1635–8

Appelbaum P, Grisso T 1995 The MacArthur Competence Study I, II, III. *Law and Human Behavior* **19**: 105–74

Appelbaum P, Gutheil T 1991 *Clinical Handbook of Psychiatry and the Law*, 2nd edn. Williams & Wilkins, Baltimore, MD

Appelbaum P, Mirkin S, Bateman A 1981 Empirical assessment of competency to consent to psychiatric hospitalization. *American Journal of Psychiatry* **138**: 1170–76

Beck J 1988 Determining competency to consent to neuroleptic drug treatment. *Hospital Community Psychiatry* **39**: 1106–8

Bradley E H, Peiris V, Wetle T 1998 Discussions about end-of-life care in nursing homes. *Journal of the American Geriatric Society* **46**: 1235–41

Carpenter W T Jr, Gold J M, Lahti A C et al. 2000 Decisional capacity for informed consent in schizophrenia research [see comments]. *Archives of General Psychiatry* **57**: 533–8

Grisso T, Appelbaum P S, Hill-Fotouhi C 1997 The MacCAT-T: A clinical tool to assess patients' capacities to make treatment decisions. *Psychiatric Services* **48**: 1415–19

Grossman S A, Piantadosi S, Covahey C 1994 Are informed consent forms that describe clinical oncology research protocols readable by most patients and their families? *Journal of Clinical Oncology* **12**: 2211–15

High D M 1989 Standards for surrogate decision making: What the elderly want. *Journal of Long Term Care Administration* **17**: 8–13

Kaufmann C L, Roth L H, Lidz C W, Meisel A 1981 Informed consent and patient decisionmaking: The reasoning of law and psychiatry. *International Journal of Law and Psychiatry* **4**: 345–61

Lambert P, Gibson J M, Nathanson P 1990 The values history: An innovation in surrogate medical decision-making. *Law, Medicine, and Health Care*. **18**: 202–12

Lidz C 1984 *Informed Consent: A Study of Decisionmaking in Psychiatry*. Guilford Press, New York

Marson D C, Cody H A, Ingram K K, Harrell L E 1995 Neuropsychologic predictors of competency in Alzheimer's disease using a rational reasons legal standard. *Archives of Neurology* **52**: 955–9

Roberts L W 1998 The ethical basis of psychiatric research: Conceptual issues and empirical findings. *Comprehensive Psychiatry* **39**: 99–110

Roberts L, Roberts B 1999 Psychiatric research ethics: An overview of evolving guidelines and current ethical dilemmas in the study of mental illness. *Biological Psychiatry* **46**: 1025–38

Roberts L W, Fraser K, Quinn D, McCarty T, Severino S K 1999 Persons with mental illness. In: Bennahum D (ed.) *Managed Care: Financial, Legal, and Ethical Issues*. Pilgrim Press, Cleveland, OH

Roberts L W, Warner T D, Brody J L 2000 Perspectives of patients with schizophrenia and psychiatrists regarding ethically important aspects of research participation. *American Journal of Psychiatry* **157**: 67–74

Roth L H, Lidz C W, Meisel A, Soloff P H, Kaufman K, Spiker D G, Foster F G 1982 Competency to decide about treatment or research: An overview of some empirical data. *International Journal of Law and Psychiatry* **5**: 29–50

Sachs G A, Stocking C B, Stern R, Cox D M, Hougham G, Sachs R S 1994 Ethical aspects of dementia research: Informed consent and proxy consent. *Clinical Research* **42**: 403–12

Schachter D, Kleinman I, Prendergast P, Remington G, Schertzer S 1994 The effect of psychopathology on the ability of schizophrenic patients to give informed consent. *Journal of Nervous and Mental Disorders* **182**: 360–2

Seckler A B, Meier D E, Mulvihill M, Paris B E 1991 Substituted judgment: How accurate are proxy predictions? [see comments] *Annals of Internal Medicine* **115**: 92–8

Stanley B, Stanley M, Guido J, Garvin L 1988 The functional competency of elderly at risk. *Gerontologist* **28** (Suppl): 53–8

Sugarman J, McCrory D C, Powell D, Krasny A, Adams B, Ball E, Cassell C 1999 Empirical research on informed consent. An annotated bibliography. *Hastings Center Report* **29**: S1–42

Warren J W, Sobal J, Tenney J H, Hoopes J M, Damron D, Levenson S, Deforge B R, Muncie H L, Jr. 1986 Informed consent by proxy. An issue in research with elderly patients. *New England Journal of Medicine* **315**: 1124–8

L. W. Roberts

Psychoanalysis: Adolescence (Clinical–Developmental Approach)

Until relatively recently, the study of adolescence was synonymous with the theory and practice of psychoanalysis. The major contributors to adolescent development came from eminent clinicians and theorists, such as Anna Freud (1936), Peter Blos (1962), Erik Erikson (1950). The adolescent crisis, psychosexual maturation, identity formation, and the second separation–individuation phase are all psychoanalytic concepts that have remained alive ever since they were first introduced during the twentieth century. Since then, the fields of adolescent development and adolescent psychopathology have differentiated greatly, and a myriad of theories and models, many developed in antagonism to psychoanalysis, have evolved. Grand theories and clinical observations have far less currency at the start of the twenty-first century than do systematic crossectional and longitudinal studies.

The discovery of adolescence by psychoanalysts became the entry point of the study of development across the life cycle. If an essential new set of tasks develops in adolescence, such as the creation of identity, then there was reason also to explore similarly significant transformations in adulthood. The important initial point was to overcome the singular psychoanalytic focus on early childhood, especially the oedipal years. In this vein, Erikson and others discovered stages of adulthood that reach well into old age. Thus adolescent development has to be viewed in the context of a shifting conception of psychoanalytic developmental psychology which views early childhood as only one important period of development.

But there is certainly no consensus on the issue of development in psychoanalytic circles: Even now, most psychoanalysts believe that the formative periods in life are the early attachment years, and the oedipal period between the ages of 5 and 7. If there was a shift, it was less to explore the later stages of life and more to understand the earliest, pre-oedipal stage of development. It is this fundamental notion of early childhood that has denied the life cycle theorists, with all their creativity, a very significant role in psychoanalytic practice.

However, adolescence, as the intermediary space between childhood and adulthood, has become a significant aspect of most psychoanalysts' thinking about development and psychopathology. This phase of life represents a theoretical compromise, a recognition that new transformations occur beyond the oedipal years, without needing to give up the primacy of early childhood for personality development and symptom formation.

With the majority of innovations about adolescence and the life cycle coming from outside psychoanalysis, it would be easy to dismiss the analytic approach. But many modern adolescent clinicians and researchers recognize that psychoanalysis has changed greatly, and that, when applied carefully, it continues to be an important interpretive method in the social and behavioral sciences. But such modern versions of psychoanalytic theory and research must include the many innovations of adjacent fields (e.g., Smelser 1999), especially the cognitive sciences, psychiatric epidemiology, anthropology, and academic developmental psychology.

The approach taken in this article is to integrate psychoanalytic concepts in an integrative framework called clinical-developmental psychology. This approach represents a synthesis of psychoanalytic concepts, especially Eriksonian identity theory, object

relations psychoanalysis as well as cognitive-developmental theory. Thus, the integrated theoretical, clinical, and research perspective does not represent psychoanalysis proper, nor does it represent traditional cognitive psychology in the tradition of Jean Piaget. It is, instead, a working out of the contradictions of two significant paradigms of human development that have formed the building blocks of a new synthesis. This new approach forms a dynamic and transformational theory of adolescent development and the lifespan, within a research paradigm and a differentiated treatment approach.

In general, psychoanalysis has become far more open to these kinds of syntheses than ever before. Bowlby's contribution (1969), for example, represents an integration of psychoanalytic and ethological thinking. The result is neither psychoanalysis nor ethology, but a theory of attachment that is one of the most significant advances of psychoanalytic thought. Modern psychoanalysis has become far more open, because of the rapid developments in academic developmental psychology and other fields and an increasing openness to integrate these approaches.

1. *Psychoanalysis*

The popular notion is that Freud's theory was about unconscious motivation, sexuality, aggressive drives, and the three structures, id, ego, and superego. Freud is, furthermore, credited with having created a comprehensive theory of human emotions. These general conceptions about Freud's contributions are not incorrect, but they are incomplete. Freud was equally dedicated to cognition and logic. He was convinced that psychiatric symptoms and psychological conflicts could be deciphered following a logical, hermeneutic strategy of interpretation. He also described in the 'Introductory Lectures' (Freud 1933), as well as in 'Mourning and Melancholia,' a theory of human relationship, one that was to become the dominant modern psychoanalytic perspective called 'object relations theory.'

These two Freudian approaches, cognitive and relational, have expanded over the decades and were joined by a major third psychoanalytic perspective, the life cycle approach. Interestingly, the cognitive branch of psychoanalysis came to overshadowed by the far more elegant and researchable cognitive contributions of academic developmental psychology, especially the work of cognitive and social cognitive psychologists. The life cycle approach, building primarily on Erikson's work (e.g., Erikson 1950) was acknowledged in psychoanalytic circles but never really shaped central psychoanalytic theory and practice. However, as mentioned above, it did help to extend the traditional focus on the early oedipal and pre-oedipal years at least into adolescence, as a second period of separation.

The life cycle approach in psychoanalysis changed into a research paradigm primarily through the work of George Vaillant (1977). Vaillant introduced a hierarchy of defenses in the tradition of Anna Freud, extended their development beyond adolescence into adulthood, and defined 'families of defenses,' defined not only in negative terms but also as adaptive styles. Vaillant showed, for example, how projection, with its pathological preoccupation with others, is related to a higher level adaptive defense which he calls 'empathy.' Vaillant applied this hierarchy of defenses to a fascinating set of longitudinal studies, especially the so-called Grant Study of Harvard men selected for their health. These men have now reached their eighties, and the research on them has demonstrated that a set of psychoanalytic constructs that were widely viewed as unresearchable, such as defense mechanisms, can be studied rigorously, have a developmental core, and are one way to trace the evolution of development in adulthood.

The psychoanalytic approach to adolescent development and the lifespan requires us to address C. G. Jung as well. His insights have been impressive and continue to garner a larger readership interested in development, meaning, and spirituality. Jung was really the first psychoanalytic thinker to dedicate significant effort to understand the transformations a person goes through during life. He was also fascinated by the collective symbols and unconscious processes of development and transformation in human development. While Jung had some interest in adolescents, his central developmental contribution was about midlife, what he viewed as a typical shift from the external world to an inner and spiritual life. It is not difficult to create a bridge between these ideas to a theory of adolescent development. Many of Jung's ideas apply to adolescent themes of creating a sense of collective belonging to a generation, to a preoccupation with 'making it in the world,' and a preoccupation with being unique.

But it is the so called object relations strand of psychoanalysis of Fairbairn (1952) and Winnicott (1965) and others that has become the organizer of modern psychoanalysis, in the form of Kohut's self-psychology, as relational psychoanalysis, and as a research paradigm in the form of attachment theory. These relational approaches have reshaped the psychological and psychoanalytic understanding of adolescence. In the earlier version, the adolescent crisis was viewed as stemming from the need to separate, to overcome childhood identifications, and to create autonomy. For that reason, Blos called this phase the second separation–individuation phase (the first one being the one that Margaret Mahler described for the toddler). In the modern relational version of the adolescent, separation is far less important than are new relational capacities, and new forms of interconnections between the adolescent, the family, and society. The psychoanalytic adolescent of the twenty-

first century is far less of a rebel who pushes the parents away than a person yearning for support and closeness, and wishing to find a new balance between family relationships and new intimate relationships. It is impossible to disaggregate what represents a new reality for families, what is a new adolescent ethos, and what is a reframing of psychoanalytic theory. All forces are at play, and while there is no single psychoanalytic scholar who is developing a new theory of adolescence, many clinicians and researchers are contributing to this relational shift.

Erikson's contribution, while having had limited resonance in psychoanalysis, continues to exert considerable influence in developmental psychology, despite the fact that he did not use traditional empirical methods of sampling and proof. He showed that understanding adolescence means to go deeper, to understand the cognitive and emotional dimensions of meaning-making. At the time of writing, no one has yet been able to combine these kinds of keen observations and theoretical synthesis when studying adolescent life. Erikson's conceptualization of adolescent identity formation, and the processes it includes (crisis, diffusion, foreclosure, moratorium), remain the standard by which subsequent contributions are measured, even if those contributions might be far more precise and valid empirically.

Since Erikson introduced his work, much has changed in the world of adolescents. Given the social dimension of the construct, one would expect cultural and historic changes (what Erikson (1968) called 'the historic moment') to influence the very nature of the adolescent experience. But also, new theories and research evidence have reshaped our understanding of how meaning-making and identity formation develop in the second decade of life. It is these cultural changes of adolescence, and the new developments inside and outside psychoanalysis, that gave impetus to a clinical-developmental synthesis of adolescent theory, research, and practice (see *Developmental Sciences, History of*).

2. *Clinical-developmental Assumptions of Adolescence and the Life Cycle*

2.1 *Development of Representations of Self and Other*

Clinical-developmental psychology of adolescence has a foundation in a relational and interactionist perspective. Such a psychoanalytic and cognitive perspective views interpersonal relationships and their internalizations as essential dimension of human development, symptomatology, and more recently, psychological treatment. Some of the pioneers of cognitive-developmental psychology (e.g., Piaget 1926) took the interactionist idea and applied it to an abstract, epistemic, logical subject of self in relation to an equally cognitive and abstract general 'other.'

Evidence has accumulated since then, that significant relationships with less abstract 'others,' such as parents, teachers, siblings, and friends are constructed actively and interpreted by the self, and that one result is an internal set of self and object representations. The concept that relationships and experiences with important others become represented internally is common, despite many conceptual differences, to a number of theories of psychology and psychoanalysis. These theories include classical psychoanalysis and branches of the modern psychoanalytic schools, such as self-psychology, object relations theory, and attachment theory. They also include the developmental theories built on the ideas of Vygotsky, Piaget, Mead, and Baldwin.

Although psychoanalysts and attachment theorists all share the view that the weaving of various significant relationships creates the inner fabric of the self, the clinical-developmental psychologist examines the *typical transformations* of relationships over the course of the lifespan (Noam and Fischer 1996). There is now evidence from a number of cross-sectional and longitudinal studies for the clinical-developmental view that internal representations of the self and others change over time in childhood, adolescence and even adulthood (e.g., Broughton 1978, Noam et al. 1995).

This view owes much to Piaget and Kohlberg, whose early work on the moral judgment of the child (Kohlberg 1969, Piaget 1932) revealed that children's constructions of social rules and relationships change systematically as they grow. Piaget and Kohlberg's ideas have been refined; new cognitive-developmental perspectives on the child's relationship to the social world in general, and to important others, now exist throughout the developmental literature.

While clinical-developmental psychology is built on an understanding of normative cognitive development, significant theoretical and clinical extensions of the traditional cognitive models have occurred. The psychoanalytic frame plays a very important role in these theoretical extensions. For example, there is greater awareness that representations of significant relationships can be quite resistant to change, despite their potential for transformation. Even an adult tends to represent the parents as relatively powerful figures, at times continuing to view them from the perspective of a child. But although the inner representational life is conservative in nature, most people are capable of changing their perspectives on significant relationships, and are able to do what attachment theorist have called 'updating the working model.'

Clinical-developmental psychology recognizes the fluctuations between progress and regress in development, the resistance to changing one's early representations, and the evolution and progression that often take place simultaneously. These fluctuations are especially strong during the adolescent years, when

close relationships with the family have to be brought into a new balance with peer influences and identity development.

2.2 Normative Development

In order to understand the developmental dimensions inherent in adolescent psychopathology, it is essential to understand what occurs typically in 'normal development.' How else can we know which challenges arise naturally during this phase, and which represent the beginnings of serious deviation and maladaptation? For example, anxieties in early childhood are common (e.g., fears of separation, monsters, darkness, etc.) and are not necessarily a cause for clinical concern. But severe separation anxiety in middle childhood and adolescence is not typical, and such anxiety often requires clinical attention. Another example is presented by Offer and colleague who have argued against Erikson's notion of serious adolescent crisis as a part of normal development. They present evidence that 'normal' adolescence is far less conflictual and risky than thought previously.

While theory development and clinical observation, the traditional psychoanlytic tools, can help us to understand what is developmentally normal and what is not, it is only through empirical studies with community samples that these distinctions can properly be made. Such studies can also help us to determine what the prevalence of anxiety, aggression, depression, suicidality, etc., is in the child and adolescent population. How do they vary by socioeconomic class, geographical location, ethnic group, and gender? And how do these problems change over time? Answers to some of these questions have already come from large-scale longitudinal epidemiological studies (for a review, see Costello and Angold 1995).

These findings give us the tools to apply, in a meaningful way, the concepts of delay and aberrant pathways that have proved to be so useful to studies of mental retardation, delinquency, and psychosis. We can, for example, trace a delinquent's moral development, allowing us to see that it is delayed in comparison to that of his nondelinquent peers (e.g., Gibbs et al. 1984). Findings such as these give us a developmental focus; they allow us to see that it is not only behavior or brain functioning that distinguishes delinquents from nondelinquents, but also level of adolescent socio-moral maturity. More importantly, once we understand the moral delay, we can, among other interventions, begin to focus our interventions on supporting the process of maturation of the cognitive system that guides moral development.

Clinical-developmental psychology, however, is not just an application of the study of normal development to the realm of psychological dysfunction. It is also a reformulation of a number of traditionally held assumptions of normative theory. Key among these reformulations is the view that adolescent symptoms may be expressed differently depending on the individual's level of development (e.g., Noam 1988). Typically, in early phases of development, impulsive, action-oriented symptoms are more frequent, while later in development, internalizing symptoms such as depression arise with greater frequency. Therefore, the idea that higher stages of development are necessarily more adaptive, as is suggested by many developmental theories, requires serious reconsideration. The clinical-developmental psychologist does not view development as progressing simply from egocentric, oral, unsocialized, or impulsive stages to those defined by the ability to take another person's perspective and show tolerance, empathy, and 'phallic' strength. Developmental theories have assumed this view for too long.

But we now know that a similar developmental course that leads to increased self-knowledge and strengths in some people leads to alienation, self-deception, and self-destructive behavior in others (e.g., Noam et al. 1995). Reaching Piagetian formal operations, for example, can throw the adolescent into a dangerous identity crisis, providing the adolescent with so many potential 'selves' that no choice seems possible (e.g., Perry 1970). Chandler (1973) referred to this cognitive crisis of adolescence as the 'epistemological loneliness of formal operations' (p. 381).

2.3 Age Chronology

The clinical-developmental approach is far less focused on chronological age than are most research approaches to development and psychopathology. Unfortunately, it has become quite common to label a model or study as being an example of research in 'developmental psychopathology' if age is introduced as a significant variable. But chronological age is a very crude indicator of the underlying processes of biological maturation, cognitive capacity, peer and family relationships, and one's understanding of one's self. Age does, of course, play a role in development in most of these domains, but there is frequently a significant variation in maturity in each of these areas within a single age group.

There are, of course, also age trends in some forms of psychopathology. For example, suicide rates rise dramatically in adolescence, as do other disorders such as depression, and conduct problems. Schizophrenia is typically an illness that begins in late adolescence or young adulthood. Age, therefore, can serve as a simple way to organize an understanding of many underlying pathogenic processes. However, the simplicity of this approach is deceptive (see Rutter 1989).

In psychopathological development, as in normal development, we often find that chronological age is

by no means a guarantor that basic cognitive and social-cognitive processes have occurred. For example, despite textbook claims to the contrary, many adolescents and adults never achieve formal-operational thought (i.e., tables in a typical psychology textbook list adolescence next to formal operations, as if all adolescents function at that cognitive stage). While many adolescents have achieved a formal operational level of thinking, others continue to function at the level of concrete operations, and indeed may remain at this level for the rest of their lives.

Consequently, if we use chronological age as the principle developmental marker, then the important variations that occur in every group of normally-developing individuals will receive insufficient attention. For the clinical-developmental psychologist, this suggests that chronological age needs to be supplemented with a domain-specific understanding of a person's capacities in different areas of development.

But, although capacities are often measured in terms of chronological age, the more sophisticated measurements of development that we have as we enter the twenty-first century call on us to subsume the idea of chronological age into our understanding of progressive adaptation and skill acquisition. In so doing, our research finds significant links between development and psychopathology in adolescents, where chronological age plays only a minor role. Level of social cognitive functioning and the shape of the internal self and object representations, rather than age, are often associated with specific disorders.

2.4 *Lifelong Development and Recovery*

Significant developmental changes can, and regularly do, occur long after childhood and adolescence. The lifelong process of development has been described eloquently by Erikson (1950), Vaillant (1977), Baltes (1997), and others. But despite a great deal of progress in understanding the achievement of salient life tasks, meaning making, and cognition, little progress has been made in applying this understanding to psychological treatment and recovery. Indeed, little is known even now about typical problems and psychological dysfunctions that arise with development throughout life.

Critical to clinical-development is the understanding that at each point in life, fundamental shifts in cognitive, emotional, social, and interpersonal development are possible, and that these shifts follow a certain logic and organization. For example, adults who identify strongly with work, authority, and control, can become disappointed when others do not live up to their expectations. These inclinations to disappointment, perfectionism, and excessive control might be long-standing, stemming from compulsive character traits that were formed in childhood. But they might also stem from the way people, as adults, define their basic categories of self, their world, and their relationships. Perhaps the categories most central to them are responsibility, predictability, and accountability. These categories and their salience can change over time into a greater acceptance of human frailty and of the limitations of control and predictability. As complex ways of understanding and experiencing change, past rigidities are also revisited. Each new chapter in a person's life typically rewrites earlier chapters. Understanding the ways in which people can transform their representational world allows the developmental therapist and developmentally-oriented psychoanalyst to help the patient revisit old relationships, overcome old vulnerabilities, work through past traumas, and help to prevent new ones from occurring. But the clinical-developmental perspective puts a great deal of emphasis on the self-righting and healing potential that is triggered under the right therapeutic conditions.

3. *Extending the Psychoanalytic Paradigm: Differential Therapy with Adolescents*

The next brief section will deal with using clinical-developmental principles to differentiate adolescent functioning and briefly introduce differential treatment implications. For a more detailed discussion see (Noam 1988), where four different levels of adolescent functioning and intervention strategies are described. Because of space considerations only two adolescent positions will be differentiated here. This section will show that there is no single treatment approach for adolescents, even youth who share the exact same age, but that the approach will differ fundamentally, not only by diagnosis, but also by developmental organization of self and relationship representation.

3.1 *Reciprocal–Instrumental Self-complexity*

Establishing treatments for adolescents who function at the reciprocal–instrumental self-complexity level is difficult. First, these adolescents are viewed by society as delayed developers, and often have many problems adjusting to the demands placed on them. They tend to be oppositional, even delinquent, and have a concrete and instrumental approach to relationships. Adding to these problems is the fact that play therapy is not appropriate at this age, because it seems infantilizing to them. Many intervention specialists take account of the verbal capacities of these adolescents in devising a treatment strategy. Because these adolescents are often quite verbal, therapists expect the youngsters to talk about their problems. But talking therapies generally do not work, nor do treatments that require a great

deal of self-reflection and insight. The therapist who tries such an approach is usually greeted with contempt and hostility, and can expect communication to break down.

And yet, it is these adolescent girls and boys who are at risk of developing problems at school, dropping out, using drugs and alcohol, behaving aggressively, and running away from home. They need a great deal of support and mentoring, in the forms of individual therapy and group interventions, to learn how to negotiate life, and how to develop goals that do not 'get them into trouble.' If adults are conducting the intervention, it is important they orient the work toward joint goals, giving the adolescent a great deal of room to decide what should happen. It is very important that therapists do not get caught up with questions such as 'what kind of therapy am I pursuing,' but instead that they remain extremely flexible, leaving the office when it is appropriate, to play, eat, or walk. It is not by chance that coaches and child-care workers are often successful with these adolescents; they are often perceived as nonthreatening and available when the adolescent needs them.

It is also essential to create contexts of peer learning and therapy for these adolescents. These environments can help the adolescents to learn how to overcome their difficulties in empathizing with others. A combination of developing group intimacy and confronting problem behavior by peers is powerful in helping the adolescent overcome an over-reliance on self-protection and move toward a more mature developmental understanding of the self and the world.

3.2 *Mutual–Inclusive Self-complexity*

At this point, talking therapy has a real potential for making the adolescent feel understood and supported. Sharing one's inner life with adults and peers, and framing it in psychological terms, becomes a powerful motivation toward intimacy and recovery. Group conformity is at its height, and so is the constant fear of losing the 'relationship base,' of abandonment by parents, teachers, and friends. It is this fear of losing those who define the self, and of not fitting in, that is a core feature of this adolescent's adolescence.

The ability of these youngsters to frame experiences in psychological terms, to describe feeling states and interactions with important others, encourages most therapists to explore patterns and motivations, autonomous self-observation and critical judgment about self and relationships. These explorations usually lead to surprisingly superficial descriptions, which frustrate both therapist and client. Surprised by this 'lack of self,' the therapist may interpret this state as an early, and primitive, manifestation of a separation–individuation problem. The clients, consequently, feel incompetent, sensing that they are not living up to expectations. Feeling inadequate leads to feeling hopeless, and clients are thereby silenced. The silence, in turn, makes the therapist and client often feel quite uncomfortable, increasing the potential for low self-esteem and depression.

There is no therapeutic model that clearly fits this stage of adolescence, which is surprising in light of the fact that so many adolescents who enter therapy function at this level. One reason might be that we did not know about this developmental gestalt before the emergence of social cognition. Existing therapy methods that use supportive strategies run the risk of under-challenging the patient. The relationship can easily become stale as the adolescent waits for some challenge or some guidance in framing the problems in new ways.

Insight-oriented therapy, on the other hand, requires an observation of systematic patterns of the self which is not available to adolescents at this developmental stage. Following a set of goals that is experienced as self-chosen, questioning existing conventions, and creating new ones, requires a differentiation between the part of the self steeped in prescribed pathways and conventions, and the part that is reflecting, doubting, and questioning. Therapists can be sure that they have an ally in the patient, who can contain cognitively, at least under supportive conditions, both sides of the self. For that reason, more traditional psychodynamic therapy can be applied, which becomes even more viable as the person grows into full adulthood.

4. *Conclusion*

The psychoanalytic perspective continues to provide important impulses for the study and treatment of adolescents and their families. Erikson provided significant theoretical tools, and Bowlby introduced a research paradigm that has been amazingly productive. But further integrations of perspectives, especially pertaining to the cognitive dimensions of adolescent thought, have proven necessary. The clinical-developmental approach not only expands the theoretical tools for the study of adolescent development and adolescent risk: It also introduces a developmental typology of intervention and prevention strategies. Intervention advances with adolescents have been slow, despite the introduction of many strategies. It is possible that progress has been slow in part because the emphasis has been primarily on differentiating disorders, rather than on understanding these disorders in the context of adolescent development.

See also: Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Adolescent

Vulnerability and Psychological Interventions; Bowlby, John (1907–90); Erikson, Erik Homburger (1902–94); Freud, Sigmund (1856–1939); Psychoanalysis, History of; Psychoanalysis in Clinical Psychology

Bibliography

Baldwin J M 1902 *Social and Ethical Interpretations in Mental Development*. Macmillan, New York

Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist* **52**: 366–80

Blos P 1962 *On Adolescence: A Psychoanalytic Interpretation*. Free Press, New York

Bowlby J 1969 *Attachment and Loss, Vol. I: Attachment*. Basic Books, New York

Broughton J 1978 *The Development of Concepts of Self, Mind, Reality, and Knowledge. New Direction in Child Development: Social Cognition*. Jossey-Bass, San Francisco, CA

Chandler M J 1973 Egocentrism and anti-social behavior: The assessment and training of social perspective taking skills. *Developmental Psychology* **9**: 326–32

Costello E J, Angold A 1995 Developmental epidemiology. In: Cicchetti D, Cohen D J (eds.) *Developmental Psychopathology*. Wiley, New York, Vol. 1, pp. 23–56

Erikson E 1950 Growth and crisis of the healthy personality. *Psychological Issues* **1**: 50–100

Erikson E 1968 *Identity, Youth, and Crisis*. Norton, New York

Fairbairn R 1952 *Psychoanalytic Studies of the Personality*. Routledge & Kegan Paul, London

Freud A 1936 *The Ego and the Mechanisms of Defense*. International Universities Press, New York

Freud S 1933 The anatomy of the mental personality, Lecture XXXI. In: *New Introductory Lectures on Psycho-analysis*, trans. W J H Sprott. W W Norton, New York, Inc. 3

Gibbs J C, Arnold K D, Ahlborn H H, Chessman F L 1984 Facilitation of socio-moral reasoning in delinquents. *Journal of Consulting and Clinical Psychology* **52**: 37–45

Kohlberg L 1969 Stage and sequence: The cognitive-developmental approach to socialization. In: Gloslin D (ed.) *Handbook of Socialization, Theory and Research*. Rand McNally, New York, pp. 347–480

Mead G H 1934 *Mind, Self, and Society*. University of Chicago Press, Chicago

Noam G G 1988 A constructivist approach to developmental psychology. In: Nannis E, Cowan P (eds.) *Developmental Psychopathology and Its Treatment*. Jossey-Bass, San Francisco, CA, pp. 91–122

Noam G G, Chandler M, LaLonde C 1995 Clinical-developmental psychology: Constructivism and social cognition in the study of psychological dysfunctions. In: Cicchetti D, Cohen D J (eds.) *Developmental Psychopathology, Vol. 1: Theory and Methods*. Wiley, New York, pp. 424–64

Noam G G, Fischer K W 1996 (eds.) *Development and Vulnerabilities in Close Relationships*. Erlbaum, Hillsdale, NJ

Offer D, Schonert-Reichl K 1992 Debunking the myths of adolescence: Findings from recent reviews. *Journal of the American Academy of Child and Adolescent Psychiatry* **31**: 1003–14

Perry W S 1970 *Forms of Intellectual and Ethical Development in the College Years*. Holt Rinehart, New York

Piaget J 1932 *The Moral Judgement of the Child*. Harcourt, New York

Rutter M 1989 Age as an ambiguous variable in developmental research: Some epidemiological considerations from developmental psychopathology. *International Journal of Behavioral Development* **12**: 1–34

Smelser N 1999 *The Social Edges of Psychoanalysis*. University of California Press, Berkeley, CA

Vaillant G E 1977 *Adaptation to Life*. Little, Brown, Boston

Winnicott D W 1965 *The Maturational Processes and the Facilitating Environment*. International Universities Press, New York

G. G. Noam

Psychoanalysis: Current Status

1. *Freud's Theory of the Mental Apparatus*

Freud's (1953a) starting point was his study of hysterical patients and the discovery that, when he found a way to help these patients piece together a coherent account of the antecedents of their conversion symptoms, dissociative phenomena, and pathological affective dispositions, all these psychopathological phenomena could be traced to traumatic experiences in their past that had become unconscious. In the course of a few years Freud abandoned his early efforts to recover repressed material by means of hypnosis, and replaced hypnosis with the technique of 'free association,' an essential aspect of psychoanalytic technique until the present time.

Practicing this method led Freud to several lines of discovery: to begin, he conceptualized unconscious mechanisms of defense that opposed the recovery of memories by free association. He described these mechanisms, namely repression, negation, isolation, projection, introjection, transformation into the opposite, rationalization, intellectualization, and most important, reaction formation. The latter involves overt, chronic patterns of thought and behavior that serve to disguise and disavow opposite tendencies linked to unconscious traumatic events and the intrapsychic conflicts derived from them. The discovery of reaction formations led Freud to the psychoanalytic study of character pathology and normal character formation, and still constitutes an important aspect of the contemporary psychoanalytic understanding and treatment of personality disorders (for practical purposes, character pathology and personality disorders are synonymous concepts).

A related line of development in Freud's theories was the discovery of the differential characteristics of conscious and unconscious thinking. Freud differentiated conscious thinking, the 'secondary process,' invested by 'attention cathexis' and dominated by

sensory perception and ordinary logic in relating to the psycho-social environment, from the 'primary process' of the 'dynamic unconscious.' That part of the unconscious mind he referred to as 'dynamic' exerted constant pressure or influence on conscious processes, against the active barrier constituted by the various defensive operations, particularly repression. The dynamic unconscious, Freud proposed, presented a general mobility of affective investments, and was ruled by the 'pleasure principle' in contrast to the 'reality principle' of consciousness. The 'primary process' thinking of the dynamic unconscious was characterized by the absence of the principle of contradiction and of ordinary logical thinking, the absence of negation and of the ordinary sense of time and space, the treatment of a part as if it were equivalent to the whole, and a general tendency toward condensation of thoughts and the displacement of affective investments from one to another mental content.

Finally, Freud proposed a 'preconscious,' an intermediate zone between the dynamic unconscious and consciousness. It represented the storehouse for retrievable memories and knowledge and for affective investments in general, and it was the seat of daydreaming, in which the reality principle of consciousness was loosened, and derivatives of the dynamic unconscious might emerge. Free association, in fact, primarily tapped the preconscious as well as the layer of unconscious defensive operations opposing the emergence of material from the dynamic unconscious.

This model of the mind as a 'place' with unconscious, preconscious, and conscious 'regions' constituted Freud's 'topographic theory.' He eventually replaced it with the 'structural theory' namely, the concept of three interacting psychic structures, the ego, the superego and the id (Freud 1961).

1.1 The Id

The id is the mental structure that contains the mental representatives of the 'drives,' that is, the ultimate intrapsychic motivations that Freud (1955) described in his final, 'dual drive theory' of libido and aggression, or metaphorically, the sexual or life drive and the destruction or death drive. In exploring unconscious mental processes, what at first appeared to be specific traumatic life experiences turned out to reflect surprisingly consistent, repetitive, intrapsychic experiences of a sexual and aggressive nature.

Freud (1953a) was particularly impressed by the regularity with which his patients reported the emergence of childhood memories reflecting seductive and traumatic sexual experiences on the one hand, and intense sexual desires and related guilt feelings, on the other. He discovered a continuity between the earliest wishes for dependency and being taken care of (the psychology, as he saw it, of the baby at the mother's breast) during what he described as the 'oral phase' of development; the pleasure in exercising control and struggles around autonomy in the subsequent 'anal phase' of development, (the psychology of toilet training); and, particularly, the sexual desire toward the parent of the opposite gender and the ambivalent rivalry for that parent's exclusive love with the parent of the same gender. He described this latter state as characteristic of the 'infantile genital stage' (from the third or fourth to the sixth year of life) and called its characteristic constellation of wishes and conflicts the positive Oedipus complex. He differentiated it from the negative Oedipus complex, that is, the love for the parent of the same gender, and the corresponding ambivalent rivalry with the parent of the other gender. Freud proposed that oedipal wishes came to dominate the infantile hierarchy of oral and anal wishes, becoming the fundamental unconscious realm of desire.

Powerful fears motivated the repression of awareness of infantile desire: the fear of loss of the object, and later of the loss of the object's love was the basic fear of the oral phase, directed against libidinal wishes to possess the breast; the fear of destructive control and annihilation of the self or the object was the dominant fear of the anal phase directed against libidinal wishes of anal expulsion and retentiveness, and the fear of castration, 'castration anxiety,' the dominant fear of the oedipal phase of development, directed against libidinal desire of the oedipal object. Unconscious guilt was a dominant later fear, originating in the superego and generally directed against drive gratification (see Sect. 1.3). Unconscious guilt over sexual impulses unconsciously equated with oedipal desires constitutes a major source of many types of pathology, such as sexual inhibition and related character pathology.

Freud described the oral phase as essentially coinciding with the infantile stage of breast feeding, the anal phase as coinciding with struggles around sphincter control, and the oedipal stage as developing gradually during the second and through the fourth years, and culminating in the fourth and the fifth years of life. This latter phase would then be followed by more general repressive processes under the dominance of the installation of the superego, leading to a 'latency phase' roughly corresponding to the school years, and finally, to a transitory reactivation of all unconscious childhood conflicts under the dominance of oedipal issues during puberty and early adolescence.

The drives represent for human behavior what the instincts constitute for the animal kingdom, that is, the ultimate biological motivational system. The drives are constant, highly individualized, developmentally shaped motivational systems. Under the dominance of the drives and guided by the primary process, the id exerts an on-going pressure toward gratification, operating in accordance with the pleasure principle. Freud initially equated the drives with primitive

affects. After discarding various other models of unconscious motivation, he ended up with the dual drive theory of libido and aggression.

He described the libido or the sexual drive as having an 'origin' in the erotogenic nature of the leading oral, anal, and genital bodily zones; an 'impulse' expressing the quantitative intensity of the drive by the intensity of the corresponding affects; an 'aim' reflected in the particular act of concrete gratification of the drive; and an 'object' consisting of displacements from the dominant parental objects of desire.

The introduction of the idea of an aggressive or 'death' drive, arrived at later in Freud's writing, stemmed from his observations of the profound self destructive urges particularly manifest in the psychopathology of major depression and suicide, and of the 'repetition compulsion' of impulse driven behavior that frequently seemed to run counter to the pleasure principle that supposedly governed unconscious drives. He never spelled out the details of the aggressive drive as to its origins. Freud described drives as intermediate between the body and the mind; the only thing we knew about them, Freud (1957a) suggested, were 'representations and affects.'

1.2 The Ego

While the id is the seat of the unconscious drives, and functions according to the 'primary process' of the dynamic unconscious, the ego, Freud (1961) proposed, is the seat of consciousness as well as of unconscious defense mechanisms that, in the psychoanalytic treatment, appear as 'resistance' to free association. The ego functions according to the logical and reality based principles of 'secondary process,' negotiating the relations between internal and external reality. Guided by the reality principle, it exerts control over perception and motility; it draws on preconscious material, controls 'attention cathexes,' and permits motor delay as well as selection of imagery and perception. The ego is also the seat of basic affects, particularly anxiety, as an alarm signal against the danger of emergence of unconscious, repressed impulses. This alarm signal may turn into a disorganized state of panic when the ego is flooded with external perceptions that activate unconscious desire and conflicts, or with overwhelming, traumatic experiences in reality that resonate with such repressed unconscious conflicts, and overwhelm the particularly sensitized ego in the process.

Freud originally equated the 'I,' that is, the categorical self of the philosophers, with consciousness. Later, once he established the theory of the ego as an organization of both conscious and unconscious functions, he at times treated the ego as if it were the subjective self, and at other times, as an impersonal organization of functions. Out of this ambiguity evolved the contemporary concept of the self within modern ego psychology as well as in British and American object relations and cultural psychoanalytic contributions (Kernberg 1984). An alternative theory of the self was proposed by Kohut (1971), the originator of the self-psychology approach within contemporary psychoanalysis.

Nowadays, an integrated concept of the self as the seat of subjectivity is considered an essential structure of the ego, and the concept of 'ego identity' refers to the integration of the concept of the self. Because developmental processes in early infancy and childhood are now better understood, an integrated self concept usually goes hand in hand with the capacity for an integrated concept of significant others. An unconscious tendency toward primitive dissociation or 'splitting' of the self-concept and of the concepts of significant objects runs counter to such integration.

Character, from a psychoanalytic perspective, may be defined as constituting the behavioral aspects of ego identity (the self-concept) and the internal relations with significant others, (the internalized world of 'object relations'). The sense of personal identity and of an internal world of object relations in turn reflects the subjective side of character. It was particularly the ego psychological approach—one of the dominant contemporary psychoanalytic schools—that developed the analysis of defensive operations of the ego, and of pathological character formation as a stable defensive organization that needed to be explored and resolved in the psychoanalytic treatment. In the process, ego psychology importantly contributed to the psychoanalytic treatment of personality disorders.

Freud (1957c) conceptualized narcissism as the libidinal investment of the ego or self, in contrast to the libidinal investment of significant others ('objects'). In proposing the possibility of a withdrawal of libidinal investment from others with an excessive investment in the self as the basic feature of narcissistic pathology, he pointed to a broad spectrum of psychopathology, and thus first stimulated the contribution of Abraham (1919), and later those of Grunberger (1979), Jacobson (1964), Kernberg (1975), Klein (1957), Kohut (1971), and Rosenfeld (1964). Thus crystallized the description of the narcissistic personality as a disorder derived from a pathological integration of a grandiose self as a defense against unbearable aggressive conflicts, particularly around primitive envy.

1.3 The Superego

In his analysis of unconscious intrapsychic conflicts between drive and defense, Freud regularly encountered unconscious feelings of guilt in his patients, reflecting an extremely strict, unconscious infantile

morality which he called the superego. This unconscious morality could lead to severe self-blame and self-attacks, and particularly to abnormal depressive reactions, which he came to regard as expressing the superego's attacks on the ego. It was particularly in studying normal and pathological mourning, that Freud (1957c) arrived at the idea of excessive mourning and depression as reflecting the unconscious internalization of the representation of an ambivalently loved and hated lost object. In unconsciously identifying the self with that object introjected into the ego, the individual now attacked his own self in replacement of the previous unconscious hatred of the object; and the internalization of aspects of that object into the superego reinforced the strictness of the individual's preexisting unconscious infantile morality.

Freud traced the origins of the superego to the overcoming of the Oedipus complex via unconscious identification with the parent of the same gender: in internalizing the oedipal parent's prohibition against the rivalry with him or her and the unconscious death wishes regularly connected with such a rivalry, and against the incestuous desire for the parent of the other gender, this internalization crystallized an unconscious infantile morality. The superego, thus based upon prohibitions against incest and parricide, and a demand for submission to, and identification with the oedipal rival, became the guarantor of the capacity for identification with moral and ethical values in general.

Freud also described the internalization of the idealized representations of both parents into the superego in the form of the 'ego ideal.' He suggested that the earliest sources of self esteem, derived from mother's love, gradually fixated by the baby's and small child's internalizations of the representations of the loving mother into the ego ideal, led to the parental demands also becoming internalized. In other words, normally self-esteem is maintained both by living up to the expectations of the internalized idealized parental objects, and by submitting to their internalized prohibitions. This consideration of self-esteem regulation leads to the clinical concept of narcissism as normal or pathological self-esteem regulation, in contrast to the theoretical concept of narcissism as the libidinal investment of the self.

2. *Psychoanalytic Treatment*

2.1 The Psychoanalytic Theory of Psychopathology

This theory proposes that the clinical manifestations of the symptomatic neuroses, character pathology, perversions, sexual inhibitions, and selected types of psychosomatic and psychotic illness reflect unconscious intrapsychic conflicts between drive derivatives following the pleasure principle, defensive operations reflecting the reality principle, and the unconscious motivations of the superego. Unconscious conflicts between impulse and defense are expressed in the form of structured conflicts between the agencies of the tripartite structure. There are ego defenses against impulses of the id; the superego motivates inhibitions and restrictions in the ego; at times the repetitive, dissociated expression of id impulses ('repetition compulsion') constitutes an effective id defense against superego pressures. The resolution of unconscious conflicts implies the analysis of all these intersystemic conflicts.

All these conflicts are clinically expressed by three types of phenomena: (a) inhibitions of normal ego functions regarding sexuality, intimacy, social relations, and affect activation; (b) compromise formations between repressed impulses and the defenses directed against them; and (c) dissociative expression of impulse and defense. This latter category implies a dominance of the splitting mechanisms referred to before. These have acquired central importance in the understanding of severe character pathology as reflected in contemporary psychoanalytic thinking.

2.2 The Structural Formulation of the Psychoanalytic Method

Psychoanalytic treatment consists, in essence, in facilitating the reactivation of the pathogenic unconscious conflicts in the treatment situation by means of a systematic analysis of the defensive operations directed against them. This leads to the gradual emergence of repressed impulses, with the possibility of elaborating them in relation to the analyst, and their eventual adaptive integration into the adult ego. Freud (1953c) had described the concept of 'sublimation' as an adaptive transformation of unconscious drives: drive derivatives, converted into a consciously tolerable form, are permitted gratification in a symbolic way while their origin remains unconscious. The result of this process is an adaptive, non-defensive compromise formation between impulse and defense. In analysis, the gradual integration into the patient's conscious ego of unconscious wishes and desires from the past and the understanding of the phantasied threats and dangers connected with them facilitates their gradual elaboration and sublimatory expression both in the consulting room and in everyday life.

2.3 The Object Relations Theory Formulation of Psychoanalytic Treatment

In the light of contemporary object relations theory, the formulation based upon the structural theory (resolution of unconscious conflicts between impulse and defense) has changed, in the sense that all

unconscious conflicts are considered to be imbedded in unconscious internalized object relations. Such internalized object relations determine both the nature of the defensive operations and of the impulses against which they are directed. These internalized object relations constitute, at the same time, the 'building blocks' of the tripartite structure of id, ego, and superego. Object relations theory proposes that the gradual analysis of intersystemic conflicts between impulse and defense (structured into conflicts between ego, superego, and id, decomposes the tripartite structure into the constituent conflicting internalized object relations. These object relations are reactivated in the treatment situation in the form of an unconscious relation between self and significant others replicated in the relation between patient and analyst, that is, the 'transference.'

The transference is the unconscious repetition in the 'here and now' of unconscious, conflicting pathogenic relationships from the past. The transference reflects the reactivation of the past conflict not in the form of a memory, but in the form of a repetition. This repetition provides essential information about the past, but constitutes at the same time, a defense in the sense that the patient repeats instead of remembering. Therefore, transference has important informative features that need to be facilitated in their development, and defensive features that need to be therapeutically resolved once their nature has been clarified. Transference analysis is the fundamental ingredient of the psychoanalytic treatment.

2.4 *The Psychoanalytic Treatment Process*

Gill (1954) in a classical definition that is still relevant today, proposed the definition of psychoanalysis as a treatment that facilitates the development of a 'regressive transference neurosis' and its resolution by means of interpretation alone, carried out by the analyst from a position of technical neutrality.

The concept of a regressive transference neurosis refers to the gradual gathering into the relationship with the analyst of the patient's most important past pathogenic experiences and unconscious conflicts. The concept of a regressive transference neurosis has been largely abandoned in practice because, particularly in patients with severe character pathology, transference regression occurs so early and consistently that the gradual development of a regressive transference neurosis is no longer a useful concept.

Gill's proposal that the resolution of the transference be achieved 'by interpretation alone,' refers to 'interpretation' as a set of the psychoanalyst's interventions that starts with 'clarification' of the patient's subjective experiences communicated by means of free association, expands with the tactful 'confrontation' of aspects of the patient's patterns of behavior that are expressed in a dissociated or split off manner from his subjective awareness, and thus complements the total expression of his intrapsychic life in the treatment situation, and finally evolves into 'interpretation per se.' Interpretation per se implies the formulation of hypotheses regarding the unconscious meanings in the 'here and now' of the patient's material, and the relation of these unconscious meanings with the 'there and then' of the patient's unconscious, past pathogenic experiences.

The concept of 'technical neutrality' refers to the analyst's impartiality regarding both impulse and defense, with a concerned objectivity that provides a helpful collaboration with the patient's efforts to come to grips with his intrapsychic conflicts.

This definition of the nature of psychoanalytic treatment needs to be complemented with the contemporary concepts of 'transference,' 'countertransference,' 'acting out' and 'working through.'

2.5 *An Object Relations Theory Model of the Transference and Countertransference*

Modern object relations theory proposes that, in the case of any particular conflict around sexual or aggressive impulses, the conflict is imbedded in an internalized object relation, that is, in a repressed or dissociated representation of the self ('self representation') linked with a particular representation of another who is a significant object of desire or hatred ('object representation'). Such units of self-representation, object representation and the dominant sexual, dependent or aggressive affect linking them are the basic 'dyadic units,' whose consolidation will give rise to the tripartite structure. Internalized dyadic relations dominated by sexual and aggressive impulses will constitute the id; internalized dyadic relations of an idealized or prohibitive nature the superego, and those related to developing psychosocial functioning and the preconscious and conscious experience, together with their unconscious, defensive organization against unconscious impulses, the ego. These internalized object relations are activated in the transference with an alternating role distribution, that is, the patient enacts a self representation while projecting the corresponding object representation onto the analyst at times, while at other times projecting his self representation onto the analyst and identifying with the corresponding object representation. The impulse or drive derivative is reflected by a dominant, usually primitive affect disposition linking a particular dyadic object relation; the associated defensive operation is also represented unconsciously by a corresponding dyadic relation between a self representation and an object representation under the dominance of a certain affect state.

The concept of countertransference, originally coined by Freud as the unresolved, reactivated transference dispositions of the analyst is currently defined

as the total affective disposition of the analyst in response to the patient and his/her transference, shifting from moment to moment, and providing important data of information to the analyst. The countertransference, thus defined, may be partially derived from unresolved problems of the analyst, but stems as well from the impact of the dominant transference reactions of the patient, from reality aspects of the patient's life, and sometimes from aspects of the analyst's life situation, that are emotionally activated in the context of the transference developments. In general, the stronger the transference regression, the more the transference determines the countertransference; thus the countertransference becomes an important diagnostic tool. The countertransference includes both the analyst's empathic identification with a patient's central subjective experience ('concordant identification') and the analyst's identification with the reciprocal object or self representation ('complementary identification') unconsciously activated in the patient as part of a certain dyadic unit, and projected onto the analyst (Racker 1957). In other words, the analyst's countertransference implies identification with what the patient cannot tolerate in himself/herself, and must dissociate, project or repress.

At this point, it is important to refer to certain primitive defensive operations that were described by Klein (1952) and her school in the context of the analysis of severe character pathology. Primitive defensive operations are characteristic of patients with severe personality disorders, and emerge in other cases during periods of regression. They include splitting, projective identification, denial, omnipotence, omnipotent control, primitive idealization, and devaluation (contempt). All these primitive defenses center on splitting, i.e., an active dissociation of contradictory ego (or self) experiences as a defense against unconscious intrapsychic conflict. They represent regression to the phase of development (the first two to three years of life) before repression and its related mechanisms mentioned are established.

Primitive defensive operations present important behavioral components that tend to induce behaviors or emotional reactions in the analyst, which, if the analyst manages to 'contain' them, permit him to diagnose in himself projected aspects of the patient's experience. Particularly 'projective identification' is a process in which: (a) the patient unconsciously projects an intolerable aspect of self experience onto (or 'into') the analyst; (b) the analyst unconsciously enacts the corresponding experience ('complementary identification'); (c) the patient tries to control the analyst, who now is under the effect of this projected behavior; and (d) the patient meanwhile maintains empathy with what is projected. Such complementary identification in the countertransference permits the analyst to identify himself through his own experience with the aspects of the patient's experience communicated by means of projective identification. This information complements what the analyst has discovered about the patient by means of clarification and confrontation, and permits the analyst to integrate all this information in the form of a 'selected fact' that constitutes the object of interpretation. Interpretation is thus a complex technique that is very much concerned with the systematic analysis of both transference and countertransference.

2.6 Contemporary Trends of the Psychoanalytic Method

Contemporary psychoanalytic technique can be seen as having evolved from a 'one person psychology' into a 'two person psychology' and then into a 'three person psychology.' The concept of 'one person psychology' refers to Freud's original analysis of the patient's unconscious intrapsychic conflicts by analyzing the intrapsychic defensive operations that oppose free association. The 'two persons psychology' refers to the central focus on the analysis of transference and countertransference. In the views of the contemporary intersubjective, interpersonal and self psychology psychoanalytic schools, the relationship between transference and countertransference is mutual, in the sense that the transference is at least in part a reaction to reality aspects of the analyst, who therefore must be acutely mindful of his contribution to the activation of the transference. The so-called 'constructivist' position regarding transference analysis assumes that it is impossible for the analyst to achieve a totally objective position outside the transference/countertransference bind.

In contrast, the contemporary 'objectivist' position, represented by the 'three person psychology' approaches of the Kleinian school, the French psychoanalytic mainstream, and significant segments of contemporary ego psychology proposes that the analyst has to divide himself between one part influenced by transference and countertransference developments, and another part that, by means of self reflection, maintains himself/herself outside this process, as an 'excluded third party,' who, symbolically, provides an early triangulation to the dyadic regression that dominates transference developments. This triangulation in the treatment situation becomes particularly important in the treatment of severe personality disorders.

The 'repetition compulsion' as a resistance of the id is very likely a form of acting out as a defense against emotional containment of an extremely painful or traumatic set of experiences. 'Working through' refers to the repeated elaboration of an unconscious conflict in the psychoanalytic situation. It is a major task for the analyst, who has to be alert to the subtle variation in meanings and implications of what on the surface

may be appear to be an endless repetition of the same conflict in the transference. The patient elaboration of the conflict that presents itself with these repetitive characteristics also implies the function of 'holding' originally described by Winnicott (1965). It consists of the analyst's capacity to withstand the onslaught of primitive transferences without retaliation, abandonment of the patient, or a self devaluing giving up, and the maintenance of a working relationship (or 'therapeutic alliance') that addresses itself consistently to the healthy part of the patient, even when the latter is under the control of his most conflicting behaviors. Bion's (1967) concept of 'containing' is complementary to 'holding,' in the sense that holding deals mostly with the affective disposition of the analyst, and containing with his cognitive capacity to maintain a concerned objectivity and focus on the 'selected fact,' permitting the integration in the analyst's mind what the patient can only express in violently dispersed or split off behavior patterns.

'Dream analysis' developed in the context of the method of free association, and constituted, in Freud's (1953b) view, a 'royal road to the unconscious.' Freud's discovery of primary process thinking derived from his method of dream analysis. By now, psychoanalytic thinking has evolved into the view that there are many 'royal roads' to the unconscious. The analysis of character defenses, for example, or of particular transference complications, may be equally important avenues of entry into the patient's unconscious mind.

The technique of dream analysis consists, in essence, in asking the patient to free associate to elements of the 'manifest content' of the dream, in order to arrive at its 'latent' content, the unconscious wish defended against and distorted by the unconscious defensive mechanisms that constitute the 'dream work,' and have transformed the latent content into the manifest dream. The latent content is revealed with the help of the simultaneous analysis of the way in which the dream is being communicated to the analyst, the 'day residuals' that may have triggered the dream, the unconscious conflicts revealed in it, and the dominant transference dispositions in the context of which the dream evolved. Dreams also provide some residual, universal symbolic meanings that may facilitate the total understanding of the latent content.

'The analysis of character' may be the single most important element of the psychoanalytic method in bringing about fundamental characterological change. Character analysis is facilitated by the patient's use of reaction formations, that is, his defensively motivated character traits, as transference resistances. Thus, the activation of defensive behaviors in the transference, reflecting the patient's characterological patterns in all interpersonal interactions, facilitates both the analysis of the underlying unconscious conflicts, and in the process, the resolution of pathological character patterns. The result is an increase in the patient's autonomy, flexibility, and capacity for adaptation.

The overall objective of psychoanalytic treatment is not only the resolution of symptoms and pathological behavior patterns or characteristics, but fundamental, structural change, that is, the expansion and enrichment of ego functions as the consequence of resolution of unconscious conflict and the integration of previously repressed and dynamically active id and superego pressures into ego potentialities. Such change is reflected in the increasing capacity for both adaptation to, and autonomy from, psychosocial demands and expectations, and an increased capacity for gratifying and successful functioning in love and work.

3. *Treatment Results: Research on Outcome*

The Menninger Psychotherapy Research Project, a naturalistic study comparing psychoanalysis, psychoanalytic psychotherapy, and supportive psychotherapy, showed psychoanalysis to be the most effective of these approaches with patients presenting relatively good ego strength, while patients with severe ego weakness—what nowadays would be described as presenting severe personality disorders or borderline personality organization—improved most with psychoanalytic psychotherapy (Kernberg et al. 1972). That research also showed how important supportive elements were throughout all modalities of treatment (Wallerstein 1986). A comprehensive review of outcome studies on psychoanalytic psychotherapy and psychoanalysis by Bachrach and collaborators (Bachrach et al. 1985) concluded that the improvement rates are in the 60–90 percent range, but it also pointed to limitations and problems in the methodology utilized.

Recently, studies regarding the treatment process and outcome of psychoanalysis and psychoanalytic psychotherapy have become more precise in defining the specific treatment variables of psychotherapeutic and psychoanalytic treatments, and several systematic studies on psychoanalytic psychotherapies and psychoanalysis are in progress (Fonagy 1998). A recent study by the Stockholm Outcome of Psychoanalysis and Psychotherapy Project (STOPPP) has found, on the basis of a relatively large patient population, that psychoanalytic treatment, in comparison with psychoanalytic psychotherapy, obtained a significantly higher degree long range symptomatic improvement (Sandell et al. 1997).

See also: Conventions and Norms: Philosophical Aspects; Ethical Issues in the 'New' Genetics; Normative Aspects of Social and Behavioral Science; Norms; Research Conduct: Ethical Codes; Research Ethics: Research; Research Funding: Ethical Aspects; Value Pluralism

Bibliography

Abraham K 1919 A particular form of neurotic resistance against the psychoanalytic method. In: *Selected Papers on Psycho-Analysis*. Brunner/Mazel, New York, 1979, pp. 303–11

Bachrach H M, Weber J J, Murray S 1985 Factors associated with the outcome of psychoanalysis. Report of the Columbia Psychoanalytic Research Center (IV). *International Review of Psychoanalysis* **12**: 379–89

Bion W R 1967 *Second Thoughts. Selected Papers on Psychoanalysis*. Basic Books, New York

Fonagy P 1998 *An Open Door Review of Outcome Studies in Psychoanalysis*

Freud S 1953a *Three Essays on the Theory of Sexuality*, standard edn. Hogarth Press, London, **7**: 125–245

Freud S 1953b *Fragment of an Analysis of a Case of Hysteria*, standard edn. Hogarth Press, London, **7**: 3–122

Freud S 1953c *The Interpretation of Dreams*, standard edn. Hogarth Press, London, **4**: 1–338, **5**: 339–625

Freud S 1955 *Beyond the Pleasure Principles*, standard edn. Hogarth Press, London,

Freud S 1957a *Instincts and their Vicissitudes*, standard edn. Hogarth Press, London, **14**: 109–40

Freud S 1957b *Mourning and Melancholia*, standard edn. Hogarth Press, London, Vol. 14, pp. 237–58

Freud S 1957c *On Narcissism*, standard edn. Hogarth Press, London, Vol. 14, pp. 69–102

Freud S 1961 *The Ego and the Id*. Standard edn. Hogarth Press, London, Vol. 19, pp. 3–66

Gill M 1954 Psychoanalysis and exploratory psychotherapy. *Journal of the American Psychoanalytic Association* **2**: 771–97

Grunberger B 1979 *Narcissism: Psychoanalytic Essays*. International Universities Press, New York

Jacobson E 1964 *The Self and the Object World*. International Universities Press, New York

Kernberg O F 1975 *Borderline Conditions and Pathological Narcissism*. Jason Aronson, Inc., New York

Kernberg O F 1984 *Severe Personality Disorders. Psychotherapeutic Strategies*. Yale University Press, New Haven, CT

Kernberg O F, Burnstein E, Coyne L, Appelbaum A, Horwitz L, Voth H 1972 Psychotherapy and Psychoanalysis: Final Report of the Menninger Foundation's Psychotherapy Research Project. *Bulletin Menninger Clinic* **36**: 1–275

Klein M 1952 Notes on some schizoid mechanisms. In: Riviere J (ed.) *Developments in Psycho-Analysis*. Hogarth Press, London, pp. 292–320

Klein M 1957 *Envy and Gratitude*. Basic Books, New York

Kohut H 1971 *The Analysis of the Self*. International Universities Press, New York

Racker H 1957 The meaning and uses of countertransference. *Psychoanalytic Quarterly* **26**: 303–57

Rosenfeld H 1964 On the psychopathology of narcissism: A clinical approach. *International Journal of Psycho-Analysis* **45**: 332–7

Rosenfeld H 1987 *Impasse and Interpretation*. Tavistock, New York

Sandell R, Blomberg J, Lazar A 1997 When reality doesn't fit the blueprint: doing research on psychoanalysis and long-term psychotherapy in a public health service program. *Psychotherapy Research* **7**: 333–44

Sandell R, Schubert J, Blomberg J, Carlsson J, Lazar A, Bromberg J 1997 The influence of therapist factors on outcomes of psychotherapy and psychoanalysis in the Stockholm Outcome of Psychotherapy and Psychoanalysis Project (STOPP). *Annual Meeting of the Society for Psychotherapy Research*, Geilo, Norway

Wallerstein R 1986 *Forty-two Lives in Treatment: A Study of Psychoanalysis and Psychotherapy*. Guilford Press, New York

Winnicot D 1965 *The Maturational Processes and the Facilitating Environment*. International Universities Press, New York

O. F. Kernberg

Psychoanalysis, History of

With the centenary of the birth of psychoanalysis, historians have begun to perceive that the biography of Sigmund Freud (1856–1939) and the interpretation of psychoanalysis constitutes an important chapter in twentieth-century cultural and intellectual history both as a body of theory and as cultural practice. An ongoing attempt to determine the meaning of Freud and psychoanalysis has led to the publication of hundreds of volumes, and the historiography of psychoanalysis has found its own place within the history of science. Historical literature about psychoanalysis ranges from the self-representations and autobiographical works of Freud to the first official biography, from moderate revisionism and contextualized approaches to the harsh, and at times hostile, clinical, feminist, epistemological, hermeneutical, and psychobiographical critiques of today. Each culture has absorbed Freud and psychoanalysis differently; but, a general history of psychoanalysis as a single science can be catalogued in three ways: (a) as a history of the psychoanalytic theories from Freud to the present; (b) as a history of the psychoanalytic movement, its institutions and practices; (c) as a self-reflection on the historiography of psychoanalysis itself.

1. *A History of the Psychoanalytic Theories from Freud to the Present*

The development of Freud's theory of the unconscious and his psychotherapeutic method can be understood on the basis of his professional development in the context of a process of differentiation and specialization of the sciences, in particular medicine and philosophy, in the late nineteenth century. Historical works of the kind, for example studies of the scientific and social context and the institutional structure of the discipline, have created a methodologically sophisticated and broad view of the psychoanalytic discipline, including empirical research and epistemological interpretation.

Various types of hermeneutical studies of psychoanalysis itself, that is, visions of it as a system of

symbols and as a method for culture-decoding rather than as a type of natural science, have broadened the view of the history of psychoanalysis. Jean Laplanche and Jean-Baptiste Pontalis's *Vocabulaire de la Psychanalyse* (1967), the first conceptual dictionary of psychoanalysis, is based on the techniques of the history of philosophy and the French tradition of the epistemology of the sciences. As the basic concepts of the psychoanalytic theory have undergone changes that reflect the development of the sociopolitical climate, the authors have incorporated the later concepts of Freud and his followers as well as critics who created new schools and theories of psychoanalysis (e.g., Melanie Klein, Jacques Lacan). Their attempts have stimulated a whole universe of separate hermeneutical and epistemological studies.

Freud's psychoanalysis with its keywords *subjectivity* and the *scientific empirical method* has been interpreted as a characteristic product of Viennese modernity, and has been seen as a parallel to phenomena like Zionism and Austromarxism. Bourgeois society and European modernity experienced significant crises in Vienna of the *fin de siècle*, where psychoanalysis was founded in the last decades of the Austro-Hungarian Monarchy. On a sociopolitical level, the multinational state tended to search for a renewal that included different forms of living. During the period between 1890 and 1910, extraordinary achievements were made in the areas of music, literature, art, and science. These were fundamentally influential for new developments in observation and thought in the twentieth century, and the psychoanalytic theory enjoyed its first heyday in the years before World War I. The following factors were generative forces behind the *fin de siècle* creativity: the rise of new cultural founding figures of the liberal bourgeousie, the influence of the Jewish intelligentsia, and the upward mobility of immigrants from the Austro-Hungarian Empire; the resistance against the political dirigism, bureaucracy and the repression of the Church and censorship; the cultural networks established with European centers and circles abroad; the import of foreign ideas; and the intellectual exchange between the submetropoles of the countries of the monarchy (Nautz and Vahrenkamp 1993).

From 1859 until 1938, Sigmund Freud lived and worked in Vienna, and the significance of Vienna as the site of psychoanalysis' foundation and development is evident. Socio-politico-cultural studies have dealt with Freud's *fin de siècle* (see Schorske 1981, McGrath 1986), viewing psychoanalysis within Viennese culture, for example at the juncture of a specific time and place. Special interest has focused on Freud's Jewish identity and the relationship between Judaism and psychoanalytic theory, ranging from minimalistic to maximalistic explanations.

Psychoanalytic theory and Freud's biography were and in many ways still are intertwined in a unique way. Therefore, the historian is in the position of having to treat psychoanalytic concepts as both strictly personal documents and as data governed by the architecture of the theoretical system. In Freud's early texts, the evolution of psychoanalytic theory is presented linearly, as though it were fundamentally the same as any other scientific field, but there are serious historiographic difficulties entailed in approaching psychoanalysis in this fashion. Freud's self-analysis, his autobiographical writings, and self-created image of psychoanalysis as a science and various theoretical concepts of psychoanalysis (e.g., seduction theory, dream interpretation, Oedipal conflict, resistance, repression, transference, and countertransference) have become prominent in the field of psychoanalytic historiography, including Freud's own changes of theory throughout his life, as well as post-Freudian developments (e.g., drive-psychology, ego-psychology, the development from defense to adaptation, from instinctual to motivational diversity).

2. *A History of Freud's Biography, the Psychoanalytic Movement, and its Institutions and Practices*

In the historiography of psychoanalysis, the genre of Freud studies has been differentiated from the historical examination of the psychoanalytic movement in its local national centers and within specific cultures. Biographical studies dominate over sociohistorical works, and the focus has been on the struggle over legacy and legitimacy rather than on a social history of the profession.

2.1 *Freud Biographies*

Three groups of historical literature about Freud have been described: (a) publications of primary sources, for example Freud's correspondences; (b) publications of historical materials concerning case studies and theoretical works of Freud; and (c) critical-academic studies about psychoanalysis. The history of Freud biography and the psychoanalytic movement began with Freud's autobiographical works *On the History of the Psychoanalytic Movement* (1914) and *An Autobiographical Study* (1925), and in his own historical papers, Freud—in his one-sided but coherent story—suggests that the history of psychoanalysis is basically the history of refusals, resistances, and the necessary regroupings of the orthodoxies (ideas, institutions and people/practitioners themselves). His story of psychoanalysis appears to be synonymous with the history of its splits, schisms, and dissidents. But, just as the proliferation of schools and the consequent battles among and splittings within psychoanalytic institutes stimulated abundant polemics, these upheavals also stimulated reworkings of almost every important item in the history of Freud biography.

Dreams from *The Interpretation of Dreams*, patients from *The Studies on Hysteria* and later case studies have been analyzed and reanalyzed from new perspectives; all the schisms and quarrels in the history of psychoanalysis were rediagnosed by the latter-day schismatics and schism-haters. Based on newly accessed archival material and historical information, early speculations and widely differing interpretations have been reconsidered. Freud biographies have addressed almost every aspect of Freud's life (profession, social status, political views, culture, religion, affection-emotion), sometimes in a general fashion, sometimes in a very specific manner, and reflecting the time and the standpoint of their authors. To forestall future falsifications in popular biographies and early attempts at studying Freud's biography by his pupils and followers, Freud's daughter Anna Freud felt prompted to authorize Ernest Jones to write a first full-scale biography (Young-Bruehl 1998). Freud studies proper began with the pioneering research of Siegfried Bernfeld; his scholarship was invaluable to Ernest Jones when putting together his trilogy *The Life and Work of Sigmund Freud* (1953–57), yet, at least some of Bernfeld's findings were suppressed in Jones'. As a member of the first generation of loyal followers, Jones' monumental work became the precipitant for a third type of historical accounts: critiques of Jones's biography, or what might be called second-order biographies.

For a long time, Freud's prominence nurtured the illusion of an isolated, ahistoric development of psychoanalysis. A cultural-historical approach to Freud became more intense as the 1960s unfolded, as biography was in general too role-model-oriented. With Henri Ellenberger's *The Discovery of the Unconscious* (1970), a mode of Freud biography contextualizing Freud began to flourish, taking the precursors of Freud's ideas more into account than Jones did.

2.2 The Psychoanalytic Movement

Interest in supporters, students, and contributors of Freud has long been marginal or limited to only the most well-known representatives. With the publication of *The Minutes of the Vienna Psychoanalytic Society* (Nunberg and Federn 1962–75), a new focus was put on the institutional framework, the form in which these individuals cooperated scientifically, and how they organized their exchanges with other social structures. Works on Freud's friendships and professional relationships include the dynamical and sociological aspects of the formation of groups, the production of scientific knowledge, the structure of mastery and slavery (often basic to the formation of psychoanalytic institutions), as well as more contextually embedded accounts of specific psychoanalytic cultures. Further, *Freud and his Followers* (Roazen 1975) stimulated studies of the family networks and subnetworks; as a consequence, biographies of early pupils and later analysts have emerged. Their adherent or schismatic relations to Freud figured largely in these accounts. A special body of historical literature has focused on the role and influence of female patients and analysts. The founding of new psychoanalytic schools has created a new intellectual-historical approach, including the professionalization and institutionalization of the discipline in different cultures and at different times. From the early days of psychoanalysis, the internationalization of psychoanalysis had its own motor and effect on theory and practice (e.g., Kutter 1992, 1995). Taking geographical and temporal particularities into account, a Western (European and North American) perspective has been criticized.

Economic and political circumstances, as well as cultural and sociohistorical connections, have all influenced the theoretical, practical, and organizational developments of psychoanalysis as a discipline and are part of the production of psychoanalysis as a corpus of knowledge. Fenichel (1998) has observed the impacts of outside reality on the production of theory and practice during the years when Central European Jewish psychoanalysts, candidates in training and their families were forced to flee National Socialist Germany and Austria and the occupied countries in Europe. The production of scientific findings and the transformation and transfer of knowledge as social processes in which internal and external factors are in a constant exchange does not minimize the role of the founder of the science of the unconscious in its beginning, but it destroys the illusion of a single isolated discovery. The shifts in the history of psychoanalysis can be seen and interpreted as a complex, dialogical, multi-author enterprise and depend on the studies of the special scientific tradition and the study of psychoanalytic microcultures (see Hale 1995, Lockot 1985, Roudinesco 1982).

3. A Self-reflection on the Historiography of Psychoanalysis Itself, and the Problems of Methodology

The history of psychoanalysis over the past generation has been a rapidly growing, controversy-ridden, and attention-attracting area. Over the past 30 years, each of the historical visions has inspired and/or provoked a revision, and each revision a counter-revision. The history of psychoanalysis has to take into account the peculiar, complex, and powerful interplay among psychoanalysis, historiography, and social, cultural, and political ideology; because this has not been recognized, many widely disparate accounts of psychoanalytic history have been proferred. Neighboring fields, such as the history of medicine, philosophy, psychology and science, women's studies, the history of law, crime, and deviance, the history of

professions, the history of sexuality, the history of the body, and cultural studies, literature, queer/gay/gender studies, have had their impacts on the historiography of psychoanalysis as well.

Studies in the area of Freud biography and the pre- and early history of psychoanalysis have received more attention. Because connections with later periods seem less powerful and convincing, the fixation on Freud himself and, even more particularly, on the originary decade of Freud's discoveries have meant that later developments during and especially since Freud's lifetime are receding in prominence.

The various schools of psychoanalysis have fashioned for themselves idealized genealogies, lines of intellectual filiation between the great minds of the past and contemporary practitioners, tracing a suitable line of evolution, and have been selective of their usable past. They may also assist in sharpening contested interdisciplinary or intradisciplinary boundaries in which 'positive' episodes are memorialized, taken as characteristic of the essence, while 'negative' ones are omitted, serving to legitimate particular theories and methodologies and delegitimizing others. Historical accounts of psychoanalytic theory and practice have to include thoughts and speculations about the special acquisition and transmission of psychoanalytic knowledge.

A 'coherent' picture of Freud and the psychoanalytic movement has undergone a profound change. In the historiography of psychoanalysis, at least two different schools of thought have been described: the 'official' history of Freud, which has been largely though not exclusively produced and consumed by practicing psychoanalysts, and a 'revisionist' approach that has emerged in the last three decades and has been mainly produced by people working outside of institutional psychoanalysis (e.g. historians, psychologists, philosophers, sociologists, political scientists). For a long time, the two groups have debated separately—with a trend to specialization and fragmentation thereby repeating the fragmentation of the object they are analyzing—and have tried to avoid other findings. But lately, a dialogue between historians and clinicians—both interested in history—concerning the history of psychoanalysis has arisen, and there have been attempts to overcome prejudices. One outcome of this dialogue has been the formation of local and international organizations and discussion groups and the establishment of special journals and periodicals. Still, the issues of language, including the problems of translation and the questions concerning authorship and audience remain central, as in other scientific disciplines. Language questions have been correlated with speculative histories on Freud and the psychoanalytic movement.

After the historical optimism of the mid-century period, the height of the professional prestige of psychoanalysis, when there was remarkably little criticism from historians of psychiatry, a revisionist historiography of psychoanalysis was stimulated; traditional biography and the general longitudinal narrative were followed by the study of institutions introducing the ideal of objective, systematic, and quantitative analysis, having in mind methodologically sophisticated, empirically substantial, and sociologically-oriented scholarship. On the other hand, as scholars have argued, the demystifications of hagiographical idealizations have often only generated 'heroic' neo-marxist and Foucauldian remystifications. Revisionist and hagiographical literature have been diagnosed by contemporary scholars as both reductionistic and lacking self-reflection.

Psychoanalysis as a theory and a profession emphasizes the value and virtue of self-awareness and in its therapeutical practices often employs a developmental mode of inquiry; nevertheless, with a few exceptions, it has remained resolutely unintrospective about the methodological and ideological conditions under which it writes the narrative of its own past. That is all the more peculiar in light of the past decade's extensive and sophisticated discussion concerning different conceptual, methodological, and epistemological strategies for writing the history of science generally. Two approaches have been observed, the internalist approach, stressing cognitive development, and the externalist approach, focusing on the social, economic, political, and professional determinants of scientific theory and practice. All history writing has been diagnosed perspectivally as the expression of a particular historical environment, culture, background, and self-consciousness among cultural-critics. LaCapra (1988) emphasizes the Freudian point that any historian's relationship with the past, and through that with the present, is a transference relationship, and that, in this transference, repetition and denial come more easily than working through and mourning (Smith and Morris 1992).

Jones' idealization of Freud has been rejected, and, particularly in the medium of psychobiography, a much more contextualized, conflicted, and complicated founder of psychoanalysis has emerged in the historical literature. Jones did not interpret blanks in his documentation as secrets or resistances, but nowadays psychologically informed biographers consider the fact that many of the most important things in people's lives are never recorded.

Contrary to the sophisticated histories of Freud's unfolding discoveries, historians have lamented the difficulties of psychobiography, and some of the key methodological problems concerning childhood experiences, the workings of the mind, and the character of thought, have been widely recognized—although such recognition has not yet informed a major historiographical work on psychoanalytic biography and history writing.

Despite the progressive de-Freudianization of psychiatry and psychology in general, Freud's place as one of the supreme minds of the twentieth century

remains secure for most historians. Psychoanalysts have altered psychoanalytic theory over the course of time, their theoretical and methodological findings depending on their analysands, their social context, and the historico-political changes in general. Historiography has altered accordingly, considering and tolerating the complexity instead of simplistic and popular versions of psychoanalysis.

See also: Cognitive Psychology: History; Defense Mechanisms; Ego Psychology and Psychoanalysis; Feminist Theory: Psychoanalytic; Freud, Sigmund (1856–1939); Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Jung, Carl Gustav (1875–1961); Klein, Melanie (1882–1960); Personality Disorders; Personality Theories; Personality Theory and Psychopathology; Psychiatry, History of; Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis in Clinical Psychology; Psychoanalysis in Sociology; Psychoanalysis: Overview; Psychologism, History of; Psychology: Overview; Psychotherapy, History of: Psychiatric Aspects; Social Psychology; Transference in Psychoanalysis; Unconscious: History of the Concept

Bibliography

Ellenberger H F 1970 *The Discovery of the Unconscious. The History and Evolution of Dynamic Psychiatry*. Allen Lane, London

Fenichel O 1998 *119 Rundbriefe Europa (1934–1938), Amerika (1938–1945)* [Mühlleitner E, Reichmayr J (eds.)]. Stroemfeld Verlag, Basel, Frankfurt am Main

Forrester J 1994 A whole climate of opinion: rewriting the history of psychoanalysis. In: Micale M S, Porter R (eds.) *Discovering the History of Psychiatry*. Oxford University Press, New York, Oxford

Freud S 1914 Zur Geschichte der psychoanalytischen Bewegung. In: *Gesammelte Werke X*, S Fischer, Frankfurt am Main [1957 On the history of the psychoanalytic movement. In: *The Standard Edition of the Complete Psychological Works of Sigmund Freud XIV*. Hogarth Press, London]

Freud S 1925 Selbstdarstellung. In: *Gesammelte Werke XIV*. S Fischer, Frankfurt am Main [1959 An autobiographical study. In: *The Standard Edition of the Complete Psychological Works of Sigmund Freud XX*. Hogarth Press, London]

Grubrich-Simitis I 1993 Zurück zu Freuds Texten. Stumme Dokumente sprechen machen. S Fischer, Frankfurt am Main [In: Grubrich-Simitis I 1996 *Back To Freud's Texts: Making Silent Documents Speak*. Yale University Press, New Haven, CT]

Hale N G Jr 1995 *The Rise and Crisis of Psychoanalysis in the United States: Freud and the Americans, 1917–1985*. Oxford University Press, New York

Haynal A, Falzeder E (eds.) 1994 *100 Years of Psychoanalysis: Contributions to the History of Psychoanalysis*. H. Karnac, London

Jones E 1953–57 *The Life and Work of Sigmund Freud* 1st edn. Basic Books, New York, 3 Vols.

Kearns K 1997 *Psychoanalysis, Historiography, and Feminist Theory. The Search for a Critical Method*. Cambridge University Press, Cambridge, UK

Kurzweil E 1989 *Freudians: a Comparative Perspective*. Yale University Press, New Haven, CT

Kutter P (ed.) 1992, 1995 *Psychoanalysis International: a Guide to Psychoanalysis Throughout the World*. Frommann-Holzboog, Stuttgart, Bad Cannstatt

LaCapra D 1988 History and psychoanalysis. In: Meltzer F (ed.) *The Trail(s) of Psychoanalysis*. University of Chicago Press, Chicago

Laplanche J, Pontalis J-B 1967 *Vocabulaire de la Psychanalyse*. Presses Universitaires de France, Paris [1973 *The Language of Psychoanalysis*. Hogarth Press and the Institute of Psycho-Analysis, London]

Lockot R 1985 *Erinnern und Durcharbeiten: Zur Geschichte der Psychoanalyse und Psychotherapie im Nationalsozialismus*. Fischer Taschenbuch Verlag, Frankfurt am Main

Lorenzer A 1984 *Intimität und Soziales Leid*. Archäologie der Psychoanalyse. S Fischer, Frankfurt am Main

Loewenberg P 1995 *Fantasy and Reality in History*. Oxford University Press, New York

McGrath W J 1986 *Freud's Discovery of Psychoanalysis*. Cornell University Press, Ithaca, NY

Mühlleitner E 1992 *Biographisches Lexikon der Psychoanalyse. Die Mitglieder der Psychologischen Mittwoch-Gesellschaft und der Wiener Psychoanalytischen Vereinigung 1902–38*. Edition Diskord, Tübingen

Nautz J, Vahrenkamp R (eds.) 1993 *Die Wiener Jahrhundertwende. Einflüsse, Umwelt, Wirkungen*. Böhlau, Wien

Nunberg H, Federn E 1962–75 *The Minutes of the Vienna Psychoanalytic Society*. International Universities Press, New York, 4 Vols

Reichmayr J 1994 *Spurensuche in der Geschichte der Psychoanalyse*. Fischer Taschenbuch, Frankfurt am Main

Roazen P 1975 *Freud and His Followers*. Knopf, New York

Roudinesco E 1982–86 *La bataille de cent ans. Histoire de la psychanalyse en France*. Editions Ramsay, Paris, 2 Vols. [1990 *Jacques Lacan & Co: A History of Psychoanalysis in France 1925–1985*. University of Chicago Press, Chicago]

Schorske C E 1981 *Fin-de-Siècle Vienna*. Vintage Books, New York

Smith J, Morris H (eds.) 1992 *Telling Facts. History and Narration in Psychoanalysis*. Johns Hopkins University Press, Baltimore, MD

Young-Bruehl E 1998 *Subject to Biography: Psychoanalysis, Feminism, and Writing Women's Lives*. Harvard University Press, Cambridge, MA

E. Mühlleitner

Psychoanalysis in Clinical Psychology

1. Historical Background to Psychoanalysis

As is well known, Freud's discovery of the talking cure was really that of an intelligent patient (Anna O) and her physician (Breuer). The patient found out for herself that certain symptoms disappeared when she

succeeded in linking up fragments of what she said and did in an altered state of consciousness (which we would now call dissociative) with forgotten impressions from her waking life. Breuer's remarkable contribution was that he had faith in the reality of the memories which emerged and did not dismiss the patient's associations as products of a deranged mind. Breuer, of course, left the therapeutic battlefield as soon as his patient's speech and behaviors started to go beyond the boundaries marked out by turn-of-the-century Viennese society.

Psychotherapy had to wait for a more powerful intellect who would not take fright at the instinctual elements in the psychological world of so-called civilized man. At first Freud sought to unearth the traumas which he believed lay at the origins of the neuroses. Later, when confronted by evidently incorrect statements, he modified his theory assuming consistency between recollection and childhood psychic reality rather than physical reality. The issue of accuracy of memories of childhood sexual trauma remains controversial, although its relevance to psychoanalytic technique is at best tangential. Freud's technique, however, was dramatically modified by his discoveries. The intense emotional relationship between patient and physician, which had its roots in catharsis following hypnotic suggestion, had gradually subsided into what was principally an intellectual exercise to reconstruct the repressed causes of psychiatric disturbance from the fragments of material derived from the patient's associations. It was a highly mechanistic approach reminiscent of a complex crossword puzzle. In the light of therapeutic failures, however, Freud once more restored the emotional charge into the patient–physician relationship. However, in place of hypnosis and suggestion, he used the patient's emotional reactions to the therapist. He thought that these reactions were explicable as transferences of the patients' feelings about their past relationships on to the current analytic relationship. Instead of seeing the patient's intense emotional reaction to the therapist as an interference, Freud came to recognize the importance of transference as a representation of earlier relationship experiences which could make the reconstruction of those experiences in analysis highly meaningful to that individual (see *Therapist–Patient Relationship*. Freud's early clinical work lacked some of the rigor which came to characterize classical psychoanalysis. His occasional encouragement to his patients to join him on holiday might now be considered boundary violations. What is perhaps less well known is that Freud remained somewhat skeptical about the effectiveness of psychoanalysis as a method of treatment. Indeed, autobiographies of some of his patients testify to his great flexibility as a clinician and use of nonpsychoanalytic techniques, including behavioral methods. Nor was Freud the only clinician to use psychoanalytic ideas flexibly. The Hungarian analyst Sandor Ferenczi was one of the first to discover the treatment of phobic disorders by relaxation and exposure, although his discovery had little impact on subsequent clinical practice.

The technique of psychoanalysis after Freud's death came to be codified with those (such as Alexander and French, and Frieda Fromm-Reichmann) who attempted to revive or retain Freud's original clinical flexibility being subjected to powerful intellectual rebuttals. In reality, psychoanalysts probably continued to vary in the extent to which they observed the ideals of therapeutic neutrality, abstinence, and a primarily interpretive stance, but these deviations could no longer be exposed to public scrutiny for fear of colleagues' forceful condemnation. Personal accounts of analyses with leading figures yield fascinating insights into variations in technique, principally in terms of the extent to which the analyst made use of personal relationship. There has been an ongoing dialectic throughout the history of psychodynamic approaches between those who emphasize interpretation and insight and those who stress the unique emotional relationship between patient and therapist as the primary vehicle of change. The controversy dates back to disputes concerning the work of Ferenczi and Rank but re-emerged with the first papers of Balint and Winnicott in London opposing a Freudian and Kleinian tradition, and somewhat later in the United States with Kohut opposing classical ego psychology.

Since the early 1990s, the pluralistic approach of modern psychoanalysis has brought out into the open many important dimensions along which psychoanalysts' techniques may vary. In particular, the recent trend to consider analyst and patient as equal partners engaged in a mutual exploration of meaning has directly challenged many of the classical constructs. This controversy is too fresh to permit a conclusive assessment. It is evident, however, that fairly fundamental reconsideration of the aims and techniques of psychoanalytic therapy is currently underway.

2. *The Basic Assumptions of Psychoanalysis*

Notwithstanding the theoretical heterogeneity of psychoanalytically based therapeutic approaches, there is probably a core set of assumptions to which all psychodynamic therapists would, to a greater or lesser extent, subscribe. These can be basically summarized as follows: (a) intrapsychic conflict is an all-pervasive aspect of human experience; (b) the mind is organized to avoid displeasure arising out of conflict and maximize the subjective sense of safety; (c) defensive strategies are adopted to manipulate ideas and experience in order to minimize displeasure; (d) psychological disturbance arises developmentally, so an adaptation that was rational and reasonable at an earlier developmental phase may leave residues that

cause maladjustment in adulthood; (e) social experience is organized into relationship representations; self and other are depicted in specific interactions charged by powerful emotional content; (f) these relationship representations inevitably re-emerge in the course of psychoanalytic treatment.

2.1 Technique—Principal Features

2.1.1 Neutrality and abstinence. Based on the classical framework of libidinal theory, Freud explicitly enjoined against the analyst giving in to the temptation of gratifying the patient's sexual desire. Obviously, this is primarily an ethical issue. However, within the psychoanalytic context it also justifies the analyst's stance of resisting the patient's curiosity or using the therapeutic relationship in any way that consciously or unconsciously could be seen as motivated by the need to gratify their own hidden desires. Within this classical frame of reference, the patient must also agree to forgo significant life changes where these could be seen as relevant to current psychotherapeutic work. In practice, such abstinence on the part of the patient is rare. Yet long-term psychodynamic treatment may founder if the emotional experiences of the therapy are obscured by the upheavals of significant life events.

Abstinence primarily ensures the neutrality of the therapist. The therapist must resist the temptation to direct the patient's associations and must remain neutral irrespective of the subject matter of the patient's experiences or fantasies. It is perhaps this aspect of the psychoanalyst's therapeutic stance that makes psychoanalysts most vulnerable to ridicule, but it is probably genuinely critical for the therapist to retain emotional distance from the patient to a degree which enables the latter to bring fantasies and fears of which they feel uncertain. Nevertheless, neutrality at its worst denies the possibility of sensitivity; recent literature on the process and outcome of psychotherapy makes it clear that the therapist's genuine concern for the patient must be manifest if significant therapeutic change is to be achieved.

2.1.2 Mechanisms of defense. Classical psychoanalytical theory and its modern equivalent (ego psychology) see conflict at the core of mental functioning. They see defenses as adaptations to intrapsychic conflict. Within object relations theories, defenses are seen as helpful to the individual to maintain an authentic or 'true' self representation or a nuclear self (see *Defense Mechanisms*. Attachment theory understands defenses as assisting in the maintenance of desirable relationships. The Klein–Bion model makes limited use of the notion of defense mechanisms but uses the term in the context of more complex hypothetical structures called defensive organizations. The term underscores the relative inflexibility of some defensive structures which are thus best conceived of as personality types. For example, narcissistic personality disorder combines idealization and destructiveness; genuine love and truth are devalued. Such a personality type may have been protective to the individual at an earlier developmental stage, and has now acquired a stability or autonomy which must be rooted in the emotional gratification which such a self-limiting form of adaptation provides.

Irrespective of the theoretical frame of reference, from a therapeutic viewpoint clinicians tend to differentiate between so-called primitive and mature defenses based on the cognitive complexity entailed in their functioning. In clinical work, primitive defenses are often noted together in the same individual. For example, individuals loosely considered 'borderline' tend to idealize and then derogate the therapist. They maintain their self-esteem by using splitting (clear separation of good from bad self-perception) and then projection. Projective identification is an elaboration of the process of projection. An individual may ascribe an undesirable mental state to the other through projection but when the other can be unconsciously forced to accept the projection and experience its impact, the defense becomes far more powerful and stable. The analyst's experience of a fragment of the patient's self state has in recent years been considered an essential part of therapeutic understanding.

Through projective identification the patient can experience a primitive mode of control over the therapist, whether in fantasy or in actualized form. Bion argued that when the self is experienced as being within another person (the therapist), the patient frequently attempts to exert total control over the recipient of the projection as part of an attempt to control split-off aspects of the self. Bion also argued that not all such externalizations were of 'bad' parts of the self. Desirable aspects of the self may also be projected, and thus projective identification can be seen as a primitive mode of communication in infancy. There are other aspects of projective identification which we commonly encounter clinically. These include the acquisition of the object's attributes in fantasy, the protection of a valued aspect of the self from internal persecution through its evacuation into the object, and the avoidance or denial of separateness. It is thus a fundamental aspect of interpersonal relationships focused on unconscious fantasy and its appreciation is critical for the adequate practice of long-term psychotherapy.

Some mechanisms of defense are thought to be more characteristic of the less severe psychological disorders (e.g., depression, anxiety, obsessive–compulsive disorders, etc.). It is beyond the scope of this article to consider the various defense mechanisms in detail. Since Anna Freud's classical work, these defense mechanisms have become fairly generally

accepted, if only as terms of mild rebuke between mental health professionals. Thus we generally accept that motivated repression may be associated with momentary forgetting of conflictual contents, that denial may be invoked by individuals wishing to disavow the emotional significance of an experience, that reaction formation is helpful in turning terror into aggression, and that identification with the aggressor may be the only solution available to a maltreated child who becomes an abuser in adolescence.

The diagnostic significance of defense mechanisms is controversial. Some researchers have claimed that an individual's habitual mode of defense has predictive value beyond that of psychiatric diagnosis. Given the theoretical ambiguity that surrounds the concept, it is unlikely that its use as part of a diagnostic formulation is justified. It does, however, assist the psychoanalytic clinician in conceptualizing the patient's reactions. In particular, these defenses are likely to be encountered in the course of psychoanalytic treatment at moments of intense emotional resistance. For example, it is common to observe patients experiencing considerable difficulty in recalling the contents of treatment sessions, yet their memory for other less central aspects of their lives appears to be exceptional. Being alert to the presence of defenses is important, not because it provides an opportunity for confrontation with ideas which the patient seems reluctant to acknowledge, but rather because it alerts the analyst to the presence of underlying anxieties which need to be tackled if resistance to the therapy is to be overcome.

2.1.3 Modes of therapeutic action. The primary mode of the therapeutic action of psychoanalytic psychotherapy is considered to be *insight*. Insight may be defined as the conscious recognition of the role of unconscious factors in current experience and behavior. Unconscious factors encompass unconscious feelings, experiences, and fantasies. Insight is more than mere intellectual knowledge. Thomä and Kächele (1987) consider insight to be equidistant from emotional experience and intellect. Etchegoyen (1991) distinguished descriptive insights from demonstrated (ostensive) insights which represent a more direct form of knowing, implying emotional contact with an event one has experienced previously.

Although specific formulations of the effect of insight depend on the theoretical framework in which explanations are couched, there is general agreement that insight's therapeutic effect is to integrate mental structures in some way (Thomä and Kächele 1987). Kleinian theorists see the healing of defensively created splits in the patient's representation of self and others as crucial. One may be more specific by specifying split or part-objects as isolated representations of intentional beings whose motivation is insufficiently well understood for these to be seen as coherent beings. In this case insight could be seen as a development of the capacity to understand internal and external objects in mental state terms, thus lending them coherence and consistency. The same phenomenon may be described as an increasing willingness on the part of the patient to see the interpersonal world from a third-person perspective.

A simple demonstration to the patient of such an integrated picture of self or others is not thought to be sufficient. The patient needs to work through a newly arrived at integration. Working through is a process of both unlearning and learning: actively discarding prior misconceptions and assimilating learning to work with new constructions. The technique of working through is not well described in the literature, yet it represents the critical advantage of long-term over short-term therapy. Working through should be systematic and much of the advantage of long-term treatment may be lost if the therapist does not follow through insights in a relatively consistent and coherent manner.

In contrast to the emphasis on insight and working through are those clinicians who, as we have seen, emphasize the relationship aspect of psychoanalytic therapy (Balint, Winnicott, Loewald, Mitchell, and many others). This aspect of psychoanalytic therapy was perhaps most eloquently described by Loewald when he wrote about the process of change as: 'set in motion, not simply by the technical skill of the analyst but by the fact that the analyst makes himself available for the development of a new "object-relationship" between the patient and the analyst …' (Loewald 1960 pp. 224–5). Sandler and Dreher (1996) have recently observed 'while insight is aimed for, it is no longer regarded as an absolutely necessary requirement without which the analysis cannot proceed.' There is general agreement that the past polarization of interpretation and insight on the one hand and bringing about change by presenting the patient with a new relationship on the other was unhelpful. It seems that patients require both and both may be required for either to be effective.

It has been suggested that change in analysis will always be individualized according to the characteristics of the patient or the analyst. For example, Blatt (1992) suggested that patients who were 'introjective' (preoccupied with establishing and maintaining a viable self-concept rather than establishing intimacy) were more responsive to interpretation and insight. By contrast, anaclitic patients (more concerned with issues of relatedness than of self-development) were more likely to benefit from the quality of the therapeutic relationship than from interpretation.

2.2 Psychoanalysis in Clinical Psychology

While intensive long-term psychoanalysis is relatively rarely offered nowadays as a therapy except to those engaged in psychotherapeutic training, long-term psychoanalytic psychotherapy, using the same principles but with session frequency at once or twice a week is

perhaps the most common long-term treatment offered for personality and emotional problems by clinical psychologists. Clinical psychology is increasingly concerned with the evidence base of the treatments that psychological therapists engage in. What hope is there in this era of empirically validated treatments, which prizes brief structured interventions, for a therapeutic approach which defines itself by freedom from constraint and preconception, and counts treatment length not in terms of number of sessions but in terms of years? Can psychoanalysis ever demonstrate its effectiveness, let alone cost-effectiveness?

There can be no excuse for the thin evidence base of psychoanalytic treatment. In the same breath that psychoanalysts often claim to be at the intellectual origin of other talking cures (e.g., systemic therapy, cognitive behavior therapy), they also seek shelter behind the relative immaturity of the discipline to explain the absence of evidence for its efficacy. Yet the evidence base of these 'derivatives' of psychoanalytic therapy has been far more firmly established than evidence for psychoanalysis itself. Of course, there are reasons given for this—reasons such as the long-term nature of the therapy, the complexity of its procedures, the elusiveness of its self-declared outcome goals, and the incompatibility of direct observation with the need for absolute confidentiality. None of these reasons stands up to careful scrutiny, however.

What is surprising, given this unpropitious backdrop is that there is in fact some suggestive evidence for the effectiveness of psychoanalysis as a treatment for psychological disorder.

2.2.1 The evidence base of psychoanalytic treatment. Psychoanalysts have been encouraged by the body of research which supports brief dynamic psychotherapy (see *Psychological Treatment, Effectiveness of*). A meta-analysis of 26 such studies has yielded effect sizes comparable to other approaches (Anderson and Lambert 1995). It may even be slightly superior to some other therapies if long-term follow-up is included in the design. One of the best-designed Randomized Controlled Trials (RCTs), the Sheffield Psychotherapy Project (Shapiro et al. 1995), found evidence for the effectiveness of a 16-session psychodynamic treatment based on Hobson's (1985) model in the treatment of major depression. There is evidence for the effectiveness of psychodynamic therapy as an adjunct to drug dependence programs (Woody et al. 1995). There is ongoing work on a brief psychodynamic treatment for panic disorder (Milrod et al. 1997). There is evidence for the use of brief psychodynamic approaches in work with older people (Thompson et al. 1987). There are some excellent studies confirming the cost-effectiveness of psychoanalytic brief therapy in a naturalistic clinical setting.

There are psychotherapy process studies which offer qualified support for the psychoanalytic case. For example, psychoanalytic interpretations given to clients which are judged to be accurate are reportedly associated with relatively good outcome (Crits-Christoph et al. 1988). There is even tentative evidence from the reanalysis of therapy tapes from the NIMH (National Institutes of Mental Health) Treatment of Depression Collaborative Research Program that the more the process of a brief therapy (CBT, cognitive behavior therapy, IPT, Interpersonal therapy) resembles that of a psychodynamic approach, the more likely it is to be effective.

Evidence is available to support therapeutic interventions which are clear derivatives of psychoanalysis. However, most analysts would consider that the aims and methods of short-term, once-a-week psychotherapy are not comparable to 'full analysis.' What do we know about the value of intensive and long-term psychodynamic treatment? Here the evidence base becomes somewhat patchy.

The Boston Psychotherapy Study (Stanton et al. 1984) compared long-term psychoanalytic therapy (two or more times a week) with supportive therapy for clients with schizophrenia in a randomized controlled design. There were some treatment-specific outcomes, but on the whole clients who received psychoanalytic therapy fared no better than those who received supportive treatment. In a more recent randomized controlled study (Bateman and Fonagy 1999), individuals with a diagnosis of borderline personality disorder (BPD) were assigned to a psychoanalytically oriented day hospital treatment or treatment as usual. The psychoanalytic arm of the treatment included therapy groups three times a week as well as individual therapy once or twice a week over an 18-month period. There were considerable gains in this group relative to the controls and these differences were not only maintained in the 18 months following discharge but increased, even though the day hospital group received less treatment than the control group. A further controlled trial of intensive psychoanalytic treatment of children with chronically poorly controlled diabetes reported significant gains in diabetic control in the treated group which was maintained at one-year follow-up (Moran et al. 1991). Experimental single case studies carried out with the same population supported the causal relationship between interpretive work and improvement in diabetic control and physical growth (Fonagy and Moran 1991). The work of Chris Heinicke also suggests that four-or five-times-weekly sessions may generate more marked improvements in children with specific learning difficulties than a less intensive psychoanalytic intervention (Heinicke and Ramsey-Klee 1986).

One of the most interesting studies to emerge recently was the Stockholm Outcome of Psychotherapy and Psychoanalysis Project (Sandell 1999). The study followed 756 persons who received national

insurance funded treatment for up to three years in psychoanalysis or in psychoanalytic psychotherapy. The groups were matched on many clinical variables. Four- or five-times-weekly analysis had similar outcomes at termination when compared with one to two sessions per week psychotherapy. However, in measurements of symptomatic outcome using the SCL-90, improvement on three-year follow-up was substantially greater for individuals who received psychoanalysis than those in psychoanalytic psychotherapy. In fact, during the follow-up period, psychotherapy patients did not change but those who had had psychoanalysis continued to improve, almost to a point where their scores were indistinguishable from those obtained from a nonclinical Swedish sample.

Another large pre–post study of psychoanalytic treatments has examined the clinical records of 763 children who were evaluated and treated at the Anna Freud Centre, under the close supervision of Freud's daughter (Fonagy and Target 1996). Children with certain disorders (e.g., depression, autism, conduct disorder) appeared to benefit only marginally from psychoanalysis or psychoanalytic psychotherapy. Interestingly, children with severe emotional disorders (three or more axis I diagnoses) did surprisingly well in psychoanalysis, although they did poorly in once or twice a week psychoanalytic psychotherapy. Younger children derived greatest benefit from intensive treatment. Adolescents appeared not to benefit from the increased frequency of sessions. The importance of the study is perhaps less in demonstrating that psychoanalysis is effective, although some of the effects on very severely disturbed children were quite remarkable, but more in identifying groups for whom the additional effort involved in intensive treatment appeared not to be warranted.

The Research Committee of the International Psychoanalytic Association has recently prepared a comprehensive review of North American and European outcome studies of psychoanalytic treatment (Fonagy et al. 1999; also see http://www.ipa.org.uk). The committee concluded that existing studies failed to unequivocally demonstrate that psychoanalysis is efficacious relative to either an alternative treatment or an active placebo, and identified a range of methodological and design problems in the 50 or so studies described in the report. Nevertheless, the report is encouraging to psychoanalysts. A number of studies testing psychoanalysis with 'state of the art' methodology are ongoing and are likely to produce more compelling evidence over the next years. Despite the limitations of the completed studies, evidence across a significant number of pre–post investigations suggests that psychoanalysis appears to be consistently helpful to patients with milder (neurotic) disorders and somewhat less consistently so for other, more severe groups. Across a range of uncontrolled or poorly controlled cohort studies, mostly carried out in Europe, longer intensive treatments tended to have better outcomes than shorter, nonintensive treatments. The impact of psychoanalysis was apparent beyond symptomatology, in measures of work functioning and reductions in health care costs.

3. *Conclusions*

Psychoanalytic notions are no longer considered 'outlandish' from the point of view of empirical psychology. New discoveries from the experimental human sciences have yielded much data consistent with psychoanalytic ideas. None of these studies can 'prove' psychodynamic ideas as the two pertain to independent realms of human functioning: the laboratory and the consulting room. However, taken as a whole, they point to the existence of underlying psychological processes posited by psychodynamic theoreticians when offering their model of human development and the pathological outcomes that may arise from it. Psychodynamic theory is unlikely ever to be comprehensively validated in empirical studies. All that we can hope for is that the general model of a person, painstakingly assembled on the basis of neuropsychological, social, and cognitive–psychological investigations, will not be too dissimilar from the vision of human functioning which has emerged from the clinical endeavors of psychodynamically oriented practitioners.

See also: Defense Mechanisms; Ego Psychology and Psychoanalysis; Feminist Theory: Psychoanalytic; Freud, Sigmund (1856–1939); Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis: Current Status; Psychoanalysis, History of; Psychoanalysis in Sociology; Psychoanalysis: Overview; Psychological Treatment, Effectiveness of; Psychology of the Real Self: Psychoanalytic Perspectives; Therapist–Patient Relationship; Transference in Psychoanalysis

Bibliography

Anderson E M, Lambert M J 1995 Short-term dynamically oriented psychotherapy: A review and meta-analysis. *Clinical Psychology Review* **15**: 503–14

Bateman A, Fonagy P 1999 The effectiveness of partial hospitalization in the treatment of borderline personality disorder—a randomised controlled trial. *American Journal of Psychiatry* **156**: 1563–9

Blatt S J 1992 The differential affect of psychotherapy and psychoanalysis with anaclitic and introjective patients: The Menninger Psychotherapy Research Project revisited. *Journal of the American Psychoanalytic Association* **40**: 691–724

Crits-Christoph P, Cooper A, Luborsky L 1988 The accuracy of therapists' interpretations and the outcome of dynamic psychotherapy. *Journal of Consulting and Clinical Psychology* **56**: 490–5

Etchegoyen H 1991 *The Fundamentals of Psychoanalytic Technique*. Karnac, London
Fonagy P, Kachele H, Krause R, Jones E, Perron R, Lopez L 1999 *An Open Door Review of Outcome Studies in Psychoanalysis*. International Psychoanalytical Association, London
Fonagy P, Moran G S 1991 Studies of the efficacy of child psychoanalysis. *Journal of Consulting and Clinical Psychology* **58**: 684–95
Fonagy P, Target M 1996 Predictors of outcome in child psychoanalysis: A retrospective study of 763 cases at the Anna Freud Centre. *Journal of the American Psychoanalytic Association* **44**: 27–77
Heinicke C M, Ramsey-Klee D M 1986 Outcome of child psychotherapy as a function of frequency of sessions. *Journal of the American Academy of Child Psychiatry* **25**: 247–53
Hobson R F 1985 *Forms of Feeling: The Heart of Psychotherapy*. Basic Books, New York
Loewald H W 1960 On the therapeutic action of psycho-analysis. *International Journal of Psycho-analysis* **41**: 16–33
Milrod B, Busch F, Cooper A, Shapiro T 1997 *Manual for Panic-focused Psychodynamic Psychotherapy*. American Psychiatric Press, Washington, DC
Moran G, Fonagy P, Kurtz A, Bolton A, Brook C 1991 A controlled study of the psychoanalytic treatment of brittle diabetes. *Journal of the American Academy of Child and Adolescent Psychiatry* **30**: 926–35
Sandell R 1999 *Long-term Findings of the Stockholm Outcome of Psychotherapy and Psychoanalysis Project* (*STOPPP*). Paper presented at the Psychoanalytic Long-term Treatments: A Challenge for Clinical and Empirical Research in Psychoanalysis, Hamburg, Germany
Sandler J, Dreher A U 1996 *What do Psychoanalysts Want? The Problem of Aims in Psychoanalysis*. Routledge, London and New York, Vol. 24
Shapiro D A, Rees A, Barkham M, Hardy G, Reynolds S, Startup M 1995 Effects of treatment duration and severity of depression on the maintenance of gains after cognitive–behavioral and psychodynamic–interpersonal psychotherapy. *Journal of Consulting and Clinical Psychology* **63**: 378–87
Stanton A H, Gunderson J G, Knapp P H, Vancelli M L, Schnitzer R, Rosenthal R 1984 Effects of psychotherapy in schizophrenia: I. Design an implementation of a controlled study. *Schizophrenia Bulletin* **10**: 520–63
Thomä H, Kächele H 1987 *Psychoanalytic Practice. I: Principles*. Springer-Verlag, New York
Thompson L W, Gallagher D, Breckenridge J S 1987 Comparative effectiveness of psychotherapies for depressed elders. *Journal of Consulting and Clinical Psychology* **55**: 385–90
Woody G E, McLellan A T, Luborsky L, O'Brien C P 1995 Psychotherapy in community methadone programs: A validation study. *American Journal of Psychiatry*. **192**: 1302–8

P. Fonagy

Psychoanalysis in Sociology

Psychoanalysis, as developed by Sigmund Freud and his followers, has had a major impact upon sociological theory and modern sociology. Freud's central discoveries—the unconscious, sexual repression, the Oedipus complex, and the like—have been deployed by sociologists to interpret and discuss the self and human subjectivity, gender and sexuality, the family and socialization, language and ideology, as well as the formation of cultural identities and forms of political domination. Notwithstanding this impact, sociology as a discipline has long had a difficult, and indeed fraught, relationship with psychoanalysis. Sociologists have criticized psychoanalytical theory on the grounds of its methodology, epistemology, and ontology. Despite these criticisms, many contemporary sociologists remain engaged with, and some strongly committed to, the psychoanalytic tradition in order to conceptualize the relation between the individual and society, especially the complex, contradictory ways that human subjects acquire and reshape the ideas, values, symbols, beliefs, and emotional dispositions of the wider society. This has been particularly evident over recent decades, in which Freudian themes and psychoanalytic motifs have been used to analyze sexual politics, issues of identity and lifestyle, as well as the debate over modernity and postmodernism.

1. *Freud and the Sociological Imagination*

Freudian psychoanalysis shares with sociology a primary preoccupation with the fate of the individual or self in the context of social relationships and the wider cultural process. While Freud's own writings were primarily derived from his clinical work with patients, and to that degree are at variance with the core methodologies of mainstream social science, his characterization of human personality and self-identity has clear parallels with, say, Thomas Hobbes' theory of human nature or Karl Marx's account of the self-seeking individual in organized capitalist society. Yet whereas both Hobbes and Marx stress the impact of social forms in the constitution of the self, Freud's methodological starting point is the individual psyche, principally the instinctual impulses and libidinal longings that shape the human imagination. The self for Freud is radically fractured or divided, split between consciousness of identity and the repressed unconscious. The biographical trajectory of the self, according to Freud, is carried on against the backdrop of this radical otherness of the unconscious, a domain of the psyche that infuses three agencies of the psyche: id, ego, and superego. The id, lying at the root of unconscious desire, is that which cannot be symbolized yet constantly strives for expression in our daily lives, manifesting itself in dreams, day-dreams, slips of the tongue, and the like. It is a hidden area of the self which knows no reality, logic, or contradiction; the unconscious is at the root, says Freud, of how people for example can simultaneously love and hate their

parents. The superego, as internal anchoring of cultural prohibition, is founded in this id. Desire, according to Freud, infiltrates all human intentions, ideals, and imperatives. Similarly, Freud sees parts of the ego as interwoven with the force of the id, the self arising as a product of the unconscious.

While Freud's writings on human personality and the constitution of the self have proven of interest to sociologists, it is his account of the relations between self and society—primarily his late writings on civilization—that has most influenced the development of modern sociology. In his late writings, Freud comes to see human beings living under the destructive force of a terrifying death drive, as well as strict cultural prohibitions on sexual desire and enjoyment. These themes are set out in his magisterial books *Beyond the Pleasure Principle* (1920) and *Civilization and its Discontents* (1930). Civilization, writes Freud, is repressive. Society imposes severe psychic demands upon individuals to achieve cultural conformity, demands that produce intense personal misery and neurotic suffering. Some sociologists have interpreted this aspect of Freud's sociology as at one with the conservative social thought of Hobbes, likening it to the theorem that cultural reproduction demands the overcoming of chaotic passion. But it is clear that this involves a radical misunderstanding of Freud's position. The dualisms that Freud uses to analyze modern civilization (consciousness/unconscious, desire/repression, the pleasure principle/the reality principle, Eros/Thanatos) suggests a disjunction between self and world. According to Freud, ambivalence is at the core of the individual's relation to the self, to others, and to society.

Freud's writings on the fate of the self in contemporary culture have strongly influenced sociological debates—from Herbert Marcuse to Michel Foucault. Too much repression, Freud says, leads to intense unconscious anguish, hostility, and rage. At such a point, the intensification of unconscious desire can release the 'mental dams' of sexual repression in a far-reaching way. The issue of the subjective seeds of social and political transformation are thus at the heart of Freud's theoretical contribution to sociology and social theory.

2. *The Integration of Psychoanalysis and Sociology*

With regard to social analysis and sociological theory, Freudian ideas have loomed large in the sociological conceptualization of human subjectivity and interpersonal relationships, sexuality and modern culture, as well as the mix of reason and irrationality in politics and history. There have been three key sociological developments in this connection: first there has been several early and influential articulations of psychoanalysis and sociology developed in Europe (particularly as represented in the work of the Frankfurt school of sociology) and North America (especially in the writings of Talcott Parsons); second, structuralist and poststructuralist appropriations of psychoanalysis as refracted in sociological debate; and finally, feminist and postmodern social theory has been significantly shaped by psychoanalysis.

2.1 *Early Disciplinary Articulations*

In Europe, Herbert Marcuse and Theodor Adorno—leading members of the Frankfurt school of critical theory—turned to Freud in order to reconceptualize the relation between self and society. The political motivation prompting this turn to Freud had its roots in Marcuse's and Adorno's attempts to conceptualize the rise of fascism, Nazism, and also the spread of bureaucratic capitalism. From Freud's theory of the Oedipus complex, and particularly his theorem concerning the repression of infantile sexuality, Marcuse and Adorno developed the notion of 'the authoritarian personality.' Driven by a desire for conformity and clear rules, the authoritarian personality was viewed by Marcuse and Adorno as a character type strongly prevalent in the German middle classes, a character type who hungered for strong leadership, social order, and regulation. Not only in Nazi Germany, however, was this personality type to be found; in the advanced liberal societies of the West, tendencies towards authoritarianism and conformism are increasingly evident. Marcuse thus spoke of the emergence of 'one-dimensional man.'

Marcuse's radical Freudianism, in particular, won a wide audience in the 1960s—not only in social science circles, but among student activists and sexual liberationists. Arguing that the so-called 'sexual revolution of the 1960s' did not seriously threaten power structures of the established social order, Marcuse sought to show how demands for freedom were routinely rechanneled for commercial interests. The core of his analysis rested upon the distinction he drew between basic and surplus repression. Basic repression he defined as the minimum level of psychological renunciation demanded by the social structure and cultural order. Repression that is surplus, by contrast, refers to the intensification of self-restraint demanded by asymmetrical relations of power. Marcuse describes the 'monogamic-patriarchal' family, for example, as a site of surplus repression. Interestingly, while Marcuse saw signs of surplus repression increasingly everywhere in late capitalist society, he remained remarkably optimistic about the possibilities for social and cultural change.

On the other side of the Atlantic, a different approach to the integration of psychoanalysis and sociology was fashioned. The core thematic of this

approach concerned social order, socialization, and the reproduction of the social system. The grand theorist of American sociology, Talcott Parsons, employed Freudian ideas to understand how basic symbols and values are internalized by human subjects throughout the socialization process. According to Parsons's appropriation of Freud for social theory, the structure of human personality is an outcome of an internalization of desired objects, role relationships, and ethico-cultural values that make up the broader social network. In this approach, it is the linkage of personality structure, the social system, and the cultural system that is stressed. Unlike Marcuse's and Adorno's emphasis on the social manipulation of the unconscious, Parsons finds a kind of pre-established harmony between the individual and society.

2.2 Structuralist and Poststructuralist Psychoanalytic Sociology

For many years the integration of psychoanalysis and social analysis developed by the Frankfurt school was commonly regarded as the most sophisticated and important work in this subfield of modern sociology. From the late 1960s onwards, however, the impact of French theory, particularly structuralist and post-structuralist philosophy, became increasingly influential in terms of theorizing the social dimensions of psychoanalysis. The key figure in this connection was Freud's French interpreter, Jacques Lacan. Seeking to rework the core concepts of psychoanalysis in the light of modern linguistics, Lacan argued that the unconscious exemplifies key features of language; as Lacan famously argues, 'the unconscious is structured like a language.' The subject or 'I,' according to Lacan, is not self-transparent, but is rather located in a system of signification from which identity is fashioned. For Lacan, intersubjectivity is at the center of psychological functioning and its disturbances; distortions or pathologies at the level of the self are, says Lacan, located in 'the discourse of the other.'

It is perhaps Lacan's essay 'The mirror stage as formative of the function of the I' (1949) that has come to exert most influence on contemporary sociological theory. In the essay, Lacan conceptualizes the small infant's initial recognition of itself in a mirror or reflecting surface, and of how this generates a sense of identity. Through the mirror, says Lacan, the infant makes an imaginary identification with its reflected image, an identification which the infant reacts to with a sense of jubilation and exhilaration. But the mirror image of the self for Lacan is, in fact, a distortion; the mirror lies. The mirror stage is radically 'imaginary,' in Lacan's theorization, since the consolingly unified image of selfhood which it generates is diametrically opposed to the bodily fragmentation and lack of coordination of the child. These imaginary traps and distortions are a universal and timeless feature of self-organization, and Lacan sees such illusions as directly feeding into and shaping pathologies of the self in contemporary culture.

Lacan was not especially interested in the social applications of psychoanalysis; it was one of his followers, the French Marxist political philosopher Louis Althusser, who brought Lacanian theory into the center of key debates in sociology. In his important essay '*Ideology and Ideological State Apparatuses*' (1971), Althusser analyzed ideology in terms of the process by which individuals come to understand and relate to themselves in a manner which supports dominant class relations. According to Althusser, ideology provides an imaginary identity, an imagined map for locating oneself in the wider social network. Echoing Lacan, Althusser uses the notion of the mirror stage to deconstruct ideology. There is a duplicate mirror-structure at the heart of ideology says Althusser, a structure that grants to the self an ideological mirror in which it can recognize itself and other people. Althusser calls this process 'interpellation'—the capturing of the individual within the net of received social meanings.

The Lacanain/Althusserian account of the decentering of the subject has been highly influential in recent sociological theory, and has impacted upon debates concerning agency, structure, class, social fragmentation, and cultural order. The social-theoretical work of Paul Hirst, Barry Hindess, Stuart Hall, Etienne Bailbar, Pierre Machery, Fredric Jameson, and Slavoj Zizek is, in differing ways, indebted to the Lacanain/Althusserian theory of the subject and its ideological subjection.

2.3 Feminism, Postmodernism, and Psychoanalytic Sociology

It is not only in studying politics and social change that psychoanalysis has become an important theoretical tool for sociology, but also in debates concerning gender and sexual politics Freudian ideas have been incorporated into social theory. In the sociology of sexuality, the sociology of the family, and especially the social theory of gender transformation, Freudian psychoanalysis has played a vital role in expanding sociologists' understanding of the subjective and affective components of human social relationships.

The psychoanalytic perspective made a forceful entry into contemporary feminist sociological theory in Juliet Mitchell's pioneering book, *Psychoanalysis and Feminism* (1974). Mitchell deployed Freudian and Lacanian psychoanalytic ideas as a means of connecting a discussion of gender power with an Althusserian/Marxist theory of late capitalist society. Against this theoretical backdrop, she asserted that definitions of masculinity and femininity are framed through linguistic and historical structures—with man as a self-determining, autonomous agent, and woman as a lacking other. Such gender dualism, according to Mitchell, is highly conducive to capitalist social

regulation—the split between private and public, the pathologies of the familial life, and the like. Mitchell's ideas, while criticized in some feminist sociological circles, had a lasting impact on psychoanalytically-oriented feminist sociology. Especially in terms of Lacanain feminist approaches, the writings of authors such as Julia Kristeva, Luce Irigaray, Jacequline Rose, and Judith Butler have significantly influenced sociological debate in recent years.

In the United States, the feminist theories of Nancy Chodorow, Jessica Benjamin, and Jane Flax have been influential in contemporary sociology. These feminist authors draw from the psychoanalytic perspective, but rather than turn to Lacan and French psychoanalysis their work selectively incorporates the insights of Freudian and post-Freudian (especially object-relational) theory. In Chodorow's work, it is part of an attempt to understand the psychic components of female and male socialization, especially in terms of the unconscious forces that shape gender roles. In Benjamin's work, psychoanalysis is deployed to rethink the dynamics of domination and submission within the wider frame of gender, society, and history. In Flax's discussion of psychoanalysis, feminism, and postmodernism, it is primarily a set of philosophical observations about the development of gender relations and the sociology of sexuality and intimacy.

Related to the intertwining of psychoanalysis and feminism, postmodernist appropriations of psychoanalysis have also been influential in recent sociological interventions. This has been especially true for the contested modernity/postmodernity debate, in which sociologists as diverse as Anthony Giddens, Alaine Touraine, and Zygmunt Bauman have drawn from Freud to analyze anew the self and self-identity. In the writings of Michel Foucault, Jean-Francois Lyotard, and Jacques Derrida, which have also had a strong influence on sociological theory, there has been much debate about the fate of the individual or 'death of the subject' in postmodernist culture. In all of these discussions, psychoanalysis has provided sociology with conceptual tools for questioning and deconstructing the enlightenment and rationality of progress.

See also: Critical Theory: Contemporary; Critical Theory: Frankfurt School; Ego Psychology and Psychoanalysis; Feminist Theory; Feminist Theory: Psychoanalytic; Freud, Sigmund (1856–1939); Gellner, Ernest (1925–95); Parsons, Talcott (1902–79); Postmodernism in Sociology; Psychoanalysis, History of; Structuralism, History of; Structuralism, Theories of; Unconscious: History of the Concept

Bibliography

Althusser L 1984 Ideology and ideological state apparatuses. *Essays on Ideology*. Verso, London

Chodorow N 1978 *The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. University of California Press, Berkeley, CA

Elliott A 1999 *Social Theory and Psychoanalysis in Transition: Self and Society from Freud to Kristeva*. Free Association Books, London and New York

Freud S 1935–74 *The Standard Edition of the Complete Psychological Works of Sigmund Freud* [trans. Stachey J]. Hogarth Press, London

Lacan J 1977 *Ecrits: A Selection*. Tavistock Press, London

Marcuse H 1956 *Eros and Civilization*. Ark, London

Mitchell J 1974 *Psychoanalysis and Feminism*. Allen Lane, London

Smelser N J 1998 *The Social Edges of Psychoanalysis*. University of California Press, Berkeley, CA

A. Elliott

Psychoanalysis: Overview

By psychoanalysis is generally understood, in the first instance, the psychological theory and corresponding clinical practice developed by Sigmund Freud in the first four decades of the twentieth century, and in most usage the term refers also to at least some of the developments that proceed directly from Freud, such as Jungian, Kleinian, and Lacanian theory, though the proper boundaries of psychoanalysis are hard to mark and subject to dispute. This article will, after a brief sketch of Freud's development, survey the principal theoretical issues raised by psychoanalytic theory, the main developments in psychoanalysis after Freud, and the relation of psychoanalysis to other fields of human enquiry.

1. Freud's Life and Writings

Freud was born of a Jewish family in Freiberg (now in the Czech republic) in 1856. He qualified as a physician at the University of Vienna, becoming early involved in neurological research under the tutelage of the physiologist Ernst Brücke, a positivist and Darwinian. After three years of medical practice at the General Hospital in Vienna, Freud received an appointment as lecturer in neuropathology. A period spent at the Salpêtrière in Paris exposed Freud to the French neurologist Jean Charcot's use of hypnotherapy to treat mental disorder. In 1886 Freud set up in private practice in Vienna, applying Charcot's methods, and in 1893 he published with the Viennese physician Josef Breuer *Studien über Hysterie* [*Studies on Hysteria*], an account of the use of hypnosis to effect the cathartic recall of forgotten traumatic experience, underpinned by a schematic theory of mental dissociation. This led Freud to attempt to formulate a general theory of mental functioning in strict neurological terms, his

'Project for a scientific psychology' of 1895, which he abandoned without completing. Thereafter, Freud concentrated on the analysis of clinical material provided by his patients and undertook a self-analysis. The explorations of these years led to the publication in 1900 of *Die Traumdeutung* [*The Interpretation of Dreams*], Freud's first full statement of psychoanalytic ideas. Here Freud defended the substitution of free association for hypnosis as a clinical method. Over the following decade Freud published *Zur Psychopathologie des Alltagslebens* [*The Psychopathology of Everyday Life*] (1904) and *Drei Abhandlungen zur Sexualtheorie* [*Three Contributions to the Theory of Sexuality*] (1905), and began to acquire a following, leading to the founding in 1910 of the International Psychoanalytic Association. The spread of psychoanalysis in Europe and the United States was accompanied by internal disagreement and the formation of alternative schools, most notably those of Alfred Adler and Carl Jung. Freud's later writings focused on theoretical questions and the application of psychoanalysis to cultural issues. Forced by the Anschluß to flee Austria, Freud died in London in 1939. (Jones 1953–7 and Gay 1988 are classic intellectual biographies of Freud. Decker 1977 analyses Freud's intellectual historical context.)

2. *Fundamental Conceptual Issues in Psychoanalytic Theory*

Though subject to fierce criticism from many quarters, Freud's theories achieved over the course of the twentieth century a massive degree of influence in all of the humanities and human sciences, and contemporary publication suggests that their power to stimulate is by no means exhausted. The theoretical dissemination of psychoanalytic ideas far exceeds the prevalence of psychoanalysis as a clinical practice: though institutes of psychoanalysis are established and thrive in most Western countries, psychoanalytic therapy has been consistently repudiated by the medical establishment and accorded no recognized place in psychiatric treatment. The greater clinical influence of psychoanalysis has been achieved through dilution, its transformation into less demanding and theoretically committed forms of treatment: psychotherapy in its myriad contemporary forms, though it most often represents itself as wholly distinct from psychoanalysis, is barely conceivable without its historical relation to Freud.

Both the influence exerted by psychoanalysis and the chronic controversy surrounding it are closely connected to the highly ambiguous, indeterminate status of psychoanalytic theory, reflected in the host of epistemological, methodological, and conceptual problems which it sets. Freud presented his theories, which he revised constantly, in terms that display very clearly the importance for his development of diverse strands in nineteenth-century thought, including philosophy and a number of the natural sciences (Darwinian biology, sexology, evolutionary anthropology, brain anatomy, associationist psychology; see Sulloway 1979 and Kitcher 1992). Furthermore, the range of phenomena that Freud regarded psychoanalysis as subsuming grew continuously, from specific forms of individual pathology (hysteria, neurosis) and abnormal psychological explananda (dreams, parapraxes) to mental functioning in general, and thence to morality, religion, art, and the social order. These factors—the continual development and interdisciplinary scope of Freud's theories, and the heterogeneity of the theoretical sources on which Freud drew—help to account for the paradoxical appearance that psychoanalysis gives of being at once an attempt to extend naturalistic, mechanistic, and reductionist modes of explanation to the human sphere, and a hermeneutical enterprise fundamentally more akin to literary criticism and philosophical reflection than to biology.

The conceptual indeterminacy of psychoanalytic theory may be analyzed under three headings.

2.1 *The Structure of Psychoanalytic Theory*

Though Freud appealed increasingly to the application of psychoanalytic ideas to social and cultural phenomena as a source of corroboration, and some post-Freudian developments have sought to shift the emphasis in this direction, psychoanalytic theory is primarily a theory of individual mental functioning. There is, therefore, a basic distinction to be drawn between the trans-individual applications of psychoanalysis and the psychoanalytic theory of the individual mind, which has logical priority in all contexts of psychoanalytic explanation.

The psychoanalytic theory of the individual mind may be divided for analytical purposes into three levels: the clinical level at which interpretations of individuals are formulated and case histories composed; low-level hypotheses concerning particular kinds of psychological configuration, such as Freud's theories of psychosomatic hysterical disorders, obsessional neurosis, psychosis, mourning and melancholia, disturbances of memory, sexual orientation, formation of male and female identity, and character types; and the highest, unifying level of psychoanalytic theory, called by Freud 'metapsychology,' consisting of abstract, semiphilosophical assumptions about the nature and general form of the mind.

At the center of the metapsychology is Freud's concept of the dynamic unconscious. The notion of the unconscious had an important place in nineteenth-century thought (Ellenberger 1970, Whyte 1979), versions of the concept being discoverable in philos-

ophers such as Arthur Schopenhauer, Friedrich Nietzsche, and Eduard von Hartmann, and in the nascent empirical psychology of Friedrich Herbart, Gustav Fechner, Hermann von Helmholtz, and Wilhelm Wundt. Occurrences of the concept of the unconscious in nineteenth-century writings tended, however, to be either theoretically undeveloped (as in Schopenhauer), or to play a merely ancillary role in metaphysical speculation (as in Hartmann), or to be qualified by uncertainty regarding the reference and explanatory value of the concept (as in the empirical psychologists). Freud may be said to have given the concept a fully scientific character, insofar as he both brought the philosophical conception of the unconscious into close relation with problems of psychological explanation and gave the concept a theoretical and empirical determinacy which it lacked previously.

Freud conceived the unconscious as an instinctually grounded source of motivation which is directly responsible for the ubiquity of conflict in normal mental life and substantially independent from cognition in its mode of operation. The nature of unconscious mental activity is explicated through Freud's theories of wish-fulfillment, of primary (as opposed to secondary) process thinking, and of the pleasure (as opposed to the reality) principle, these theories being closely associated with a developmental story. Though Freud asserted the autonomy of psychological characterizations of the unconscious from neurological speculation in *The Interpretation of Dreams*, and always claimed that the unconscious is composed of 'ideas' rather than somatic states, i.e., that it has an intentional character, at all periods of his development Freud regarded 'economic' descriptions of mental life—descriptions in terms of quantities of neural energy—as also essential to the metapsychology. Psychoanalytic theory thus incorporates suggestions of both dualism and physicalist reductionism.

Within this framework, the metapsychology underwent continual modification, the most important development being Freud's move from an account that ties the existence of the dynamic unconscious directly to repression—the first topographical model of *Ucs*., *Pcs*., and *Cs*., articulated in Freud's metapsychological papers of 1915—to the structural model of ego, id, and superego presented in *Das Ich and das Es* [*The Ego and the Id*] (1923), according to which the unconscious is aboriginal, and distributed across the major portion of the ego as well as the id. Whereas the first model implies that unconscious mentality is created only through aberrations in conscious mental life, which remains the core of subjectivity, the second model implies that consciousness requires a special 'making conscious' of mental states and that human subjectivity by its nature extends beyond consciousness. The structural model may thus be argued to impact in a quasi-philosophical manner on the traditional and commonsensical conception of the self or person. (For scholarly expositions of Freud's metapsychological concepts, see Laplanche and Pontalis 1967.)

Though Freud's own writings stress the interdependence of the three levels, there is some vacillation in Freud between regarding the metapsychology as the crowning and defining achievement of psychoanalysis, and reducing it to a mere superstructure that might be radically overhauled and is replaceable, in principle if not in practice, by neurological science. Questions may thus be raised concerning the relations of the three levels, and in this context various possibilities present themselves. Psychoanalysis may be identified with either a theoretical edifice offering a general, philosophically significant model of the mind; or a cluster of lower-level theories aiming at the formulation of psychological laws in specific subdomains of psychology; or an accumulated body of case studies of individuals serving primarily to specify a therapeutic technique. These options—in numerous variations and combinations—have been defended and explored by theorists in close connection with different views of the epistemological foundations of psychoanalysis.

2.2 Epistemological and Methodological Issues

Because of the physicalist–reductionist strain in psychoanalysis, and the positivistic character of the conception of knowledge appealed to by Freud in his claims for the scientificity of psychoanalysis, there has been a natural (and, in the English-speaking world, dominant) tendency to consider the question of the empirical warrant for psychoanalytic theory in the context of scientific methodology. This mode of approach to psychoanalysis is associated with a negative estimate of the objectivity and cognitive value of psychoanalysis.

Two major writings in this sphere are Karl Popper (1963, Chap. 1) and Adolf Grünbaum (1984). Popper conducted an enormously influential attack on psychoanalysis in the course of defending falsifiability as the key to scientific method. According to Popper, psychoanalysis does not open itself to refutation, and so fails to satisfy the condition of falsifiability which, in his account, provides the necessary and sufficient condition for determining the scientific standing and candidacy for rational acceptability of a theory. In view of its immunity to counterevidence, Popper holds, psychoanalysis must be classified (alongside Marxism) as a 'pseudo-science.' Grünbaum, operating with a different conception of scientific method, maintains in opposition to Popper that psychoanalysis does meet the conditions for evaluation as a scientific theory, but proceeds to offer a detailed critique of what he takes to be Freud's inductive reasoning. Freudian theory reposes, Grünbaum argues, on the claims that only psychoanalysis can give correct

insight into the cause of neurosis, and that such insight is causally necessary for a durable cure. Grünbaum then emphasizes the empirical weakness of psychoanalysis' claim to causal efficacy, and presses the frequently voiced objection that the therapeutic effects of psychoanalysis may be due—for all that Freud is able to demonstrate to the contrary—to suggestion. Grünbaum argues further that, even if the clinical data could be taken at face value, the inferences that Freud draws are unwarranted.

Different conceptions of science carry different implications for the epistemological status of psychoanalysis, and a sufficiently anti-rationalistic conception of science, such as Paul Feyerabend's, may succeed in eliminating the epistemological discrepancy between psychoanalysis and the natural sciences; but this strategy carries obvious costs. It is generally agreed that, without repudiating altogether the concept of a distinctive scientific method in the manner of Feyerabend, the prospect of defending the epistemological credentials of psychoanalysis in the terms supplied by the philosophy of natural science is poor. Attempts to test psychoanalytic hypotheses experimentally in controlled, extraclinical contexts have been inconclusive (see Eysenck and Wilson 1973). The comparison with cognitive science, which has successfully claimed for itself empirical support, may appear to aggravate the epistemological situation of psychoanalysis. (For discussion of psychoanalysis' scientificity, see Hook 1964 and Edelson 1984.)

An alternative approach to the epistemological issue is to abandon the assumption that psychoanalysis must be shown to conform to the model of explanation of the natural sciences in order for it to be held to have cognitive value. This involves breaking with Freud's own understanding of psychoanalysis, but it may draw support from the fact that Freud also on numerous occasions emphasizes the alignment of psychoanalysis with, and its dependence on, commonsense principles of reasoning: the everyday assumptions about the mind that are embedded in commonsense ('folk') psychology.

A diversity of approaches falls under this heading, but they share agreement that the right way to understand psychoanalysis is to regard it as reworking the routes of psychological understanding laid down in our ordinary, everyday grasp of ourselves, and validated independently from scientific methodology.

Some philosophers, encouraged by Ludwig Wittgenstein's writings on the philosophy of psychology and remarks on Freud (Wittgenstein 1978), explored the notion that psychoanalytic explanation trades off the concept of a *reason* rather than that of a cause, and tried to suggest that psychoanalytic interpretation is directly continuous with the practice of explaining agents' actions by citing their reasons. Though psychoanalysis is thereby freed from the burden of having to appeal to strict causal laws, the epistemological grounds of which prove so problematic, an evident difficulty faces this approach, arising from the tension between the assumption of the agent's rationality which is presupposed by the very application of the concept of a reason, and the nonrational or irrational character of the mental connections postulated in psychoanalytic explanation. The resulting concept of 'neurotic rationality' employed by some philosophers reflects this discomfort.

However, more satisfactory solutions along these lines can be developed by making the relation of psychoanalysis to commonsense psychology less direct. One such approach regards psychoanalysis as a *theoretical extension* of commonsense psychology. It is argued that the familiar schema of practical reason explanation, formalized in the practical syllogism, is in psychoanalytic explanation fundamentally modified by the substitution of the concepts of wish and fantasy for those of belief and desire, and by the replacement of a direct, intrapsychic relation between desire and mental representation for the syllogistic relation between practical reasoning and action. The basis for assigning content and explanatory role to wishes and fantasies—thematic linkages, relations of meaning—remains, however, the same. Melanie Klein's development of Freud's theories, which attributes enormous importance to the role of fantasy in mental life, is standardly appealed to by proponents of this approach. In this view, the overarching epistemological ground for psychoanalysis lies in its capacity to offer a unified explanation for phenomena—dream, psychopathology, mental conflict, sexuality, and so on—that commonsense psychology is unable or poorly equipped to explain (see *Irrationality: Philosophical Aspects*), while exploiting the same interpretative methodology as commonsense psychology and complementing commonsense psychological explanations. (See Wollheim 1974, 1991, Wollheim and Hopkins 1982, Hopkins 1992, Neu 1991, Cavell 1993, Levine 1999.)

An alternative epistemological route from commonsense psychology to psychoanalysis is to proceed by way of *philosophical theory*: given an independently elaborated philosophical theory of psychoanalytically relevant aspects of human subjectivity, intersubjectivity, and representation, it becomes possible to interpret psychoanalysis in the terms provided by this theory and thereby extend to psychoanalysis whatever cognitive authority the theory possesses. Proponents of this approach typically draw on different philosophical traditions from the commonsense-extension theorists, whose philosophical allegiances characteristically reflect a mixture of empiricism and Wittgenstein. (The two approaches are, however, not necessarily incompatible.)

Jürgen Habermas and Jacques Lacan provide two clear examples of this mode of approach. Habermas' (1968, Chaps. 10–11) hermeneutic account asserts a complete separation of psychoanalysis from the natural sciences, this association being attributed to a

naturalistic and scientistic misconception of psychoanalysis on Freud's part, and seeks instead to integrate psychoanalysis with communication theory and to set it in the context of a nontraditional conception of rationality and cognitive interest. The unconscious and its symbolism are understood by Habermas in terms of specific forms of distortion of communication and interruptions to self-communication; psychoanalytic interpretation and treatment are understood correspondingly as directed to retrieving elements that have suffered exclusion from the field of public communication, and to engendering a form of self-reflection that will undo self-alienation and thereby emancipate. (Ricœur 1965 offers an alternative hermeneutic approach.)

Lacan's reading of Freud (see Evans 1996) evinces a similar methodological structure: a general theory of subjectivity and representation, in Lacan's case deriving from multiple sources including Ferdinand de Saussure and G. W. F. Hegel, is employed to elucidate psychoanalytic concepts. The concept of the unconscious is explicated by Lacan in terms of paradoxes arising necessarily from the attempt to represent the self and its objects, and constraints put on interpersonal desire by its symbolic mediation, while Freud's theories of the nature of unconscious mental processing are recast in terms borrowed from structural linguistics.

2.3 Further Theoretical Issues

An issue that has plagued psychoanalysis since its inception concerns the conceptual possibility of mental phenomena being unconscious. Nineteenth-century philosophy of psychology raises frequently the question whether the posit of unconscious mental phenomena is warranted or even intelligible, Franz Brentano for example devoting to it a chapter of his *Psychologie vom empirischen Standpunkte* [*Psychology from an Empirical Standpoint*] (1874). Freud's own response was to emphasize the dogmatism of those who reject the notion of unconscious mental phenomena in advance of considering what explanatory gains may be made by employing it, these being of course, in Freud's view, decisively favorable. It is fair to say that beyond this point little progress has been made with the problem, which to that extent remains without a satisfactory solution. This does not, however, constitute a serious embarrassment for psychoanalysis, since the identity of mental phenomena with consciousness maintained by opponents of psychoanalysis is as hard to supply with a philosophical proof as the concept of unconscious mental phenomena is hard to make intuitively acceptable. The issue has in any case been largely overtaken, or at any rate moved down the contemporary philosophical agenda, by the confident positing of unconscious mental processing witnessed by Noam Chomsky and cognitive psychology.

In its place a new question has emerged concerning what it may mean to speak of mental content which is prelinguistic, unconceptualized, or nonpropositional. Psychoanalytic theory, most markedly in its Kleinian development but also in Freud, appears to require the existence of such content, in view of what it claims about the mental life of infants and primary process thinking. The opposition here is between those who wish to take such talk at face value, as referring to mental states with no less reality than the beliefs and desires attributed in commonsense psychology, and those who regard it as violating *a priori* conditions on the possession of mental states, and are consequently obliged to either reject or offer nonrealistic readings of psychoanalytic accounts of the infant's mind and primary process (Cavell 1993). The prominence of this issue in contemporary discussion of psychoanalysis reflects the continuing shift in twentieth-century philosophy, also evidenced in Habermas's and Lacan's reconstructions of psychoanalysis, away from Car tesianism towards a consideration of mental states in terms of their conditions of attribution and thus in relation to language.

A further set of questions regarding the nature of psychoanalysis is raised by the recent ascent of cognitive science (see *Cognitive Science: Philosophical Aspects*). Freud's unpublished *Project* allows itself to be recast in contemporary, computational terms, and it has been argued that the same sort of reconstruction should be applied to Freud's properly psychoanalytic theories. This would open up the possibility of integrating psychoanalytic theory with the models and hypotheses of present-day cognitive psychology, and further theoretical integration with evolutionary psychology may also be envisaged.

This raises the conceptual question whether psychoanalysis is essentially a subpersonal psychological theory, or whether it is as much a theory of the fully personal level as commonsense psychology. The former is necessary if there is to be direct theoretical integration of psychoanalysis with cognitive psychology; if psychoanalysis is a personal-level psychology, then its relation to cognitive psychology will prove as problematic as that of commonsense psychology. Examination of Freud's writings shows that the question is hard to answer. On the one hand, a move in the direction of subpersonality would appear to be implied in the explanatory shift away from consciousness, and to be required if psychoanalysis is to succeed in offering explanations where commonsense psychology gives out. On the other hand, to regard unconscious motivation and content as on a par with involuntary bodily processes, and as no more intrinsically tied to the self or subjectivity than a neural event, would appear to reflect a deep misunderstanding of the nature and purpose of psychoanalytic explanation. Plausibly, psychoanalysis needs

to be read as containing both personal and subpersonal elements, but to say this is, of course, still to leave the matter unclear.

Questions of possible theoretical integration aside, cognitive psychology poses an epistemological challenge to psychoanalysis. If cognitive psychology should prove capable of explaining the same set of explananda as psychoanalysis, and if its explanations are not isomorphic with those of psychoanalysis, then psychoanalytic theory would lose its claim to explanatory value: Freudian theory would simply have been supplanted, and the commonsense-extension theorist's defense of psychoanalysis would collapse. In such circumstances the only prospect of legitimating psychoanalytic theory would lie in allying the theory with some anti-naturalistic philosophical outlook. Critics of psychoanalysis who share Popper's misgivings about psychoanalysis' scientificity, and those who believe that Freud remained too close to commonsense psychology and failed to carry through a sufficiently strict program of scientific naturalism in psychology, regard this *dénouement* as the most likely.

As this section has shown, discussion of psychoanalysis is inseparable from a multitude of broad philosophical issues: it cannot be maintained that psychoanalysis is unproblematically grounded on experience, nor that there is any one self-evidently correct reading of the conceptual structure of psychoanalytic theory. No other major psychological theory finds itself in the same position. Consequently, even if it is agreed that psychoanalysis should be removed from the context of the philosophy of science, no determinate account of the nature or epistemology of psychoanalytic theory can be arrived at without importing general philosophical commitments. The inevitable result is that the estimation of psychoanalytic theory is embroiled with a number of fundamental philosophical antinomies, above all the conflict between scientific naturalism and anti-naturalism in philosophy and the human sciences.

3. Post-Freudian Developments in Psychoanalytic Theory

The vast range of developments of Freud's ideas may be roughly distinguished into those that attempt an explicit synthesis of psychoanalysis with some other body of theory in the human sciences and those that seek to remain within the terms of Freud's specifically psychological program, aiming either to carry forward Freud's own theories, or to formulate rival, non-Freudian forms of psychoanalysis.

The former will be surveyed in the next section. Post-Freudian developments of the latter kind (see Eagle 1984, Frosh 1987) have in some cases been driven by purely theoretical considerations, reflecting particular views of the relation of the clinical to the theoretical components in psychoanalysis, of the relation of psychoanalytic theory to the natural sciences, and of the epistemology of psychoanalysis. Instances include the attempts of American theorists such as Merton Gill, Karl Pribram, Emmanuel Peterfreund, David Rapaport, and B. B. Rubinstein to reconstruct Freud's metapsychology in the terms of the natural sciences or to discard it altogether; and, at the opposite end of the spectrum, the endeavors of Jean-Paul Sartre and Ludwig Binswanger to re-express what they take to be Freud's central insights in a philosophical idiom wholly purified of naturalistic elements, 'existential psychoanalysis.'

The more fruitful and influential post-Freudian developments in psychoanalytic theory, however, have taken place in close association with clinical work, and thus engage with Freud's empirically significant, low-level hypotheses, as much as with his metapsychology. The earliest instances are provided by Freud's dissenting contemporaries (see Brown 1961). Jung's 'analytical psychology' disputes Freud's view of the importance of sexuality, and more fundamentally asserts the direct and determining participation in psychic life of unconscious mental contents with a wholly impersonal content and origin, psychological 'archetypes' composing a 'collective unconscious' (see Nagy 1991). Again, Adler's monistic interpretation of human behavior in terms of responses to feelings of inferiority, and Otto Rank's theory of the trauma of birth as the source of mental disorder compete directly with Freud's view of the determinants of mental life.

Only Jung succeeded in founding a school of psychoanalysis that is fully independent of Freud; the ideas of Adler, Rank, and other early figures in the psychoanalytic movement, such as Sandor Ferenczi, were instead either discarded or incorporated into later developments in Freudian theory. The most presently active traditions in psychoanalytic thought are those of Lacan, Klein, and the school of ego psychology which descends from Anna Freud. These schools are geographically anchored in, respectively, France, Britain, and the United States.

As indicated above, the tendency in Lacan's reading of Freud is to raise psychoanalysis to the level of a philosophical discourse, an approach which—encouraged also by the work of Jacques Derrida—has been endorsed by many French philosophers (see, e.g., Henry 1985) and literary theorists who have made use of Lacanian ideas. While some elements in Lacan, such as his claim that the unconscious is structured like a language, may be regarded as reconstructive, other parts of Lacanian theory, such as his account of the infant as acquiring its ego through imaginary identification with a mirror image, offer novel extensions to Freudian theory. Other theses of Lacan's, such as his conception of psychoanalytic treatment as directed towards the acceptance of a fluid self-identity rather than firmly integrated ego, and his view of human discontent as deriving from the metaphysics of

subjectivity rather than the hedonic costs of renouncing instinctual gratification, seem indisputably revisionary.

The Kleinian development of Freud, referred to also as the British school of object-relations, rests on an innovation in clinical practice. The psychoanalysis of children undertaken by Klein greatly expanded the body of data available to psychoanalytic theory. On the basis of the new evidence concerning the mental life of very young children afforded by her employment of the play technique, Klein hypothesized that relations to external objects are standardly—in childhood and throughout adult life—facilitated and mediated by representations of objects in fantasy. On that basis Klein developed a synchronic and diachronic model of the mind as incorporating a fantasized 'inner world' composed of 'internal figures' and exhibiting certain key structures (called by Klein the 'depressive' and 'paranoid–schizoid positions') and patterns of activity (splitting, projective identification) determining crucial aspects of the subject's relation to objects in reality (see Segal 1989, Hinshelwood 1990, and *Object Recognition: Theories*). Some aspects of Kleinian theory remain highly controversial—particularly Klein's affirmation of Freud's concept of the death instinct, and the richness of the mental life that her account attributes to the child in its first few months of life—but it has been vigorously extended in the post-Kleinian work of Wilfred Bion, Donald Winnicott, Hannah Segal, and others, especially with reference to psychosis (an area largely neglected by Freud) and symbolism (see Anderson 1992).

Ego psychology developed out of work by Anna Freud, whose *Das Ich und die Abwehrmechanismen* [*The Ego and the Mechanisms of Defence*] (1936) sought to shift the emphasis of psychoanalytic interpretation away from the contents of the id toward the modes of operation employed by the ego in its task of mediating the relations of drives to external reality on the grounds that only in this way can the specificity of an individual's psychology be grasped and addressed therapeutically. (Anna Freud and her followers clashed with Kleinian members in the course of schismatic 'Controversial Discussions' held at the London Institute for Psycho-Analysis in the 1940s.) This theoretical reorientation became more pronounced in the work of Heinz Hartmann in the United States, who produced a systematic account of the ego as comprising a sphere of autonomy and as serving the function of adaptation. Later developments stemming from Hartmann include Margaret Mahler's work on the infant–mother relation and Heinz Kohut's 'self-psychology.'

The diversity of post-Freudian schools is frequently cited as an argument against the cognitive value of psychoanalysis. Whether this implies a lack of convergence of the kind that may properly be taken to impugn the objectivity of a theory is, however, a moot point. Plausibly, the diversity of psychoanalytic theory derives in large part from their adjustment to different therapeutic techniques—entailing different explanatory foci—and at least some of the disagreement that appears to remain may be attributed to differences of formulation rather than of substance (implicit versions of Kleinian ideas, for example, can be detected readily in Lacan and ego psychology). The degree to which different psychoanalytic theories are intertranslatable, and to which consilience may be discovered among the clinical interpretations of differently schooled analysts, has, however, yet to be determined.

4. The Relation of Psychoanalytic Theory to other Human Sciences

Wherever the human sciences conceive their object in intentional or semantic terms— that is, in all but the most austerely reductionist contexts—the possibility of establishing a continuity with psychoanalytic theory exists and has been explored to some degree; in most of the social sciences, a variety of interfaces with psychoanalysis have been formulated. Freud's own writings on culture, religion, and social theory—*Totem und Tabu* [*Totem and Taboo*] (1913), *Massenpsychologie und Ich-Analyse* [*Group Psychology and the Analysis of the Ego*] (1921), *Die Zukunft einer Illusion* [*The Future of an Illusion*] (1927), *Das Unbehagen in der Kultur* [*Civilization and its Discontents*] (1930), *Der Mann Moses und die monostheistische Religion* [*Moses and Monotheism*] (1939)—provide models for the trans-individual application of psychoanalysis but have not been adopted in mainstream social and cultural thinking (similarly Jung, whose exploration of cultural anthropology was more intensive than Freud's); probably the element in Freud to have most influenced these spheres has been, closer to the psychology of the individual, his theory of moral development (Deigh 1996).

The most extensive applications and developments of psychoanalytic ideas beyond the psychology of the individual have taken place in contexts where psychoanalysis has promised to fill an explanatory lacuna in some other body of theory and vice versa; here the possibility of a genuine synthesis is created.

Marxism found itself in just such a situation, and elicited highly original developments in the critical theory of the Frankfurt school. A basic consonance of Marxism (regarding its theory of ideology, alienation, and the social superstructure) with psychoanalysis is established by their shared concern with the analysis of forces conditioning and limiting rationality and freedom. The theoretical need of Marxism for psychoanalysis derives from the absence from classical Marxist theory of any account of the subjective process whereby ideology is instantiated and rendered effective at the level of the individual. The corresponding need of psychoanalysis consists in its failure to take account of or make conceptual provision for the culturally

variable mediation of human desire and the permeation of sociohistorical conditions to unconscious mental functioning (Freud's insensitivity to the role of cultural variables was asserted early on by psychoanalytic theorists such as Karen Horney).

In this light, the social holism of Marx and the methodological individualism of Freud could be regarded as complementary rather than opposed, and capable of yielding in conjunction a form of social psychology which would unify the instinctual and cultural, and individual and social, dimensions of human existence. Theories that sought such a unification—specifically, a comprehensive theory of capitalism and of features of modernity such as mass culture and fascism—were advanced by Erich Fromm, Wilhelm Reich, Herbert Marcuse, Max Horkheimer, T. W. Adorno, and others (see Held 1980, Chap. 4). Habermas's reading of Freud also belongs to this tradition.

The project of critical theory has been partially recapitulated in the intense preoccupation of feminist theory with psychoanalysis over recent decades (Mitchell 1975 and Chodorow 1978 are early, seminal works). Here the theoretical alliance has been more ambivalent, however. On the one hand psychoanalysis has seemed to offer feminist theorists a unique and necessary theoretical route to a deep understanding of women's subjectivity; no other psychological theory appears to carry a similar promise of allowing a unified understanding of sexuality, gender, and power relations. On the other hand, psychoanalysis has appeared to be antithetical to the purposes of feminism, on account of its reflections of Freud's own patriarchiality and conservatism, and the putative biological determinism and essentialism of Freudian theory, whereby existing social relations and identities are universalized. Resolutions to this dilemma have typically taken the form of a partial historicization of Freudian theory (sometimes appealing to the work of Michel Foucault) and an exploration of post-Freudian (especially Lacanian) theory, aimed at constructing a form of psychoanalysis adequately oriented towards the cognitive–emancipatory interests of women.

See also: Ego Psychology and Psychoanalysis; Feminist Theory: Psychoanalytic; Freud, Sigmund (1856–1939); Klein, Melanie (1882–1960); Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis: Current Status; Psychoanalysis, History of; Psychoanalysis in Clinical Psychology; Psychoanalysis in Sociology; Psychology of the Real Self: Psychoanalytic Perspectives; Transference in Psychoanalysis

Bibliography

Anderson R (ed.) 1992 *Clinical Lectures on Klein and Bion*. Routledge, London

Assoun P-L 1976 *Freud: la philosophie et les philosophes*. Presses Universitaires de France, Paris

Brown J 1961 *Freud and the Post-Freudians*. Penguin, Harmondsworth, UK

Cavell M 1993 *The Psychoanalytic Mind: From Freud to Philosophy*. Harvard University Press, Cambridge, MA

Chodorow N 1978 *The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. University of California Press, Berkeley, CA

Clark P, Wright C (eds.) 1988 *Mind, Psychoanalysis and Science*. Blackwell, Oxford, UK

Decker H 1977 *Freud in Germany: Revolution and Reaction in Science, 1893–1907*. International Universities Press, New York

Deigh J 1996 *The Sources of Moral Agency: Essays in Moral Psychology and Freudian Theory*. Cambridge University Press, Cambridge, UK

Eagle M 1984 *Recent Developments in Psychoanalysis: A Critical Evaluation*. Harvard University Press, Cambridge, MA

Edelson M 1984 *Hypothesis and Evidence in Psychoanalysis*. University of Chicago Press, Chicago

Ellenberger H 1970 *The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry*. Basic Books, New York

Evans D 1996 *An Introductory Dictionary of Lacanian Psychoanalysis*. Routledge, London

Eysenck H, Wilson G (eds.) 1973 *The Experimental Study of Freudian Theories*. Methuen, London

Frosh S 1987 *The Politics of Psychoanalysis: An Introduction to Freudian and Post-Freudian Theory*. Macmillan, London

Gay P 1988 *Freud: A Life for Our Time*. Dent, London

Grünbaum A 1984 *The Foundations of Psychoanalysis*. University of California Press, Berkeley, CA

Habermas J 1968 *Erkenntnis und Interesse*. Suhrkamp Verlag, Frankfurt am Main [1971 *Knowledge and Human Interests*. Beacon, Boston]

Held D 1980 *Introduction to Critical Theory: Horkheimer to Habermas*. University of California Press, Berkeley, CA

Henry M 1985 *Généalogie de la psychanalyse: le commencement perdu*. Presses Universitaires de France, Paris [1993, *The Genealogy of Psychoanalysis*. Stanford University Press, Stanford, CA]

Hinshelwood R 1990 *A Dictionary of Kleinian Thought*. Free Association Press, London

Hook S (ed.) 1964 *Psychoanalysis, Scientific Method and Philosophy*. New York University Press, New York

Hopkins J 1992 Psychoanalysis, interpretation, and science. In: Hopkins J, Savile A (eds.) *Psychoanalysis, Mind and Art: Perspectives on Richard Wollheim*. Blackwell, Oxford, UK

Jones E 1953–57 *Sigmund Freud: Life and Work*. Hogarth Press, London, 3 Vols.

Kitcher P 1992 *Freud's Dream: A Complete Interdisciplinary Science of Mind*. MIT Press, Cambridge, MA

Laplanche J, Pontalis J B 1967 *Vocabulaire de la psychanalyse*. Presses Universitaires de France, Paris [1973 *The Language of Psychoanalysis*. Methuen, London]

Levine M (ed.) 1999 *Analytic Freud: Philosophy and Psychoanalysis*. Routledge, London

Mitchell J 1975 *Psychoanalysis and Feminism*. Penguin, Harmondsworth, UK

Nagy M 1991 *Philosophical Issues in the Psychology of C. G. Jung*. State University of New York, Albany, NY

Neu J (ed.) 1991 *The Cambridge Companion to Freud*. Cambridge University Press, Cambridge, UK

Popper K 1963 *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge & Kegan Paul, London
Ricœur P 1965 *De l'interprétation: essai sur Freud*. Éditions du Seuil, Paris [1970 *Freud and Philosophy: An Essay in Interpretation*. Yale University Press, New Haven, CT]
Segal H 1989 *Klein*. Fontana, London
Sulloway F 1979 *Freud: Biologist of the Mind*. Basic Books, New York
Whyte L 1979 *The Unconscious Before Freud*. Friedman, London
Wittgenstein L 1978 *Lectures and Conversations on Aesthetics, Psychology and Religious Belief*. Blackwell, Oxford, UK
Wollheim R (ed.) 1974 *Freud: A Collection of Critical Essays*. Anchor Doubleday, New York [Reprinted 1977 as *Philosophers on Freud: New Evaluations*. Aronson, New York]
Wollheim R 1991 *Freud*, 2nd edn. Fontana Collins, London
Wollheim R, Hopkins J (eds.) 1982 *Philosophical Essays on Freud*. Cambridge University Press, Cambridge, UK

S. Gardner

Psychobiology of Stress and Early Child Development

Stressors that threaten well-being activate physiological systems that, while fostering immediate survival, can interfere with repair of the body, defense against illness, and development. Two systems play major roles in stress: the sympathetic adrenomedullary (SAM) system that produces adrenaline and other catecholamines, and the limbic–hypothalamic–pituitary–adrenocortical (LHPA) system that produces the steroid hormone cortisol in humans. Of these two systems, the LHPA system exerts slower, more persistent actions, and is often the focus of research on stress physiology. A growing literature indicates that conditions associated with stress, maltreatment and neglect during fetal and early development alter regulation of the LHPA system in ways that could heighten vulnerability later to stress-related disorders. During early life, individual differences in temperament interact with qualities of the caregiving environment to predict cortisol responses to psychosocial stressors. Combined, these growing literatures on normal and atypical development promise to explicate mechanisms through which early experiences affect physical and psychological health.

1. Psychobiology of the LHPA System

Animal studies constitute the bulk of research on this system. Determining their implications for human development is a major challenge. Current perspectives drawn from the animal research will be presented in this section.

Cortisol is the end product of a cascade of events that begins with stimulation of cells in an area of the brain called the hypothalamus. These cells secrete peptide hormones (CRH and AVP) that, in turn, cause cells in the pituitary gland (once thought of as the master endocrine gland) to produce yet another biochemical, ACTH. ACTH is secreted into the bloodstream where it is picked up by receptors on the cortex of the adrenal glands, stimulating the adrenal to secrete cortisol (Johnson et al. 1992). Humans produce cortisol even when they are not stressed. Nonstressed or basal cortisol production follows a daily rhythm. Like a thermostat, basal levels are maintained through negative feedback loops; the primary feedback sites are located in two areas of the brain, called the hypothalamus and hippocampus. Through multiple pathways in the brain, internal and external threats to well-being cause increased activity in this LHPA system that raise cortisol levels over baseline (de Kloet 1991). Responses to psychological stressors operate through pathways involving an area of the brain called the amygdala that also plays important roles in organizing behavioral reactions to threat. CRH, the peptide that begins the cascade of events that stimulates cortisol production, is produced in many brain areas, including sites such as the amygdala, where it is believed to play a major role in regulating threat reactive behavior and in stimulating both the LHPA and the SAM systems to react to psychological stressors. One major theory relating stress to depression focuses on repeated stressors and the orchestrating role of extra-hypothalamic CRH (Nemeroff 1996). The cumulative impact of stressors on health has been described as allostatic load, or the health-impairing costs of survival-facilitating activity of the LHPA system (McEwen 1998).

Understanding how the LHPA system promotes both health and disease is a core problem in stress research. Both hypo- and hyper-responsivity of the LHPA system are associated with disorder (de Kloet et al. 1998). Disorders of hypo-responsivity include those related to failure to contain immune inflammatory responses and to maintain the responsivity of areas of the brain involved in emotions and information processing; hyper-responsivity disorders include those related to immune suppression, cardiovascular disease, and affective psychopathology. One line of reasoning traces these contrasting impacts through the two types of receptors for cortisol that mediate two modes of the hormone's effects (de Kloet et al. 1998). The first, a 'proactive' or 'permissive' mode, maintains basal activity of the system, regulates the threshold of the system's response to stressors, promotes coordination of the daily rhythm in the hormone, and helps maintain the viability of brain systems involved in selective attention, sensory integration, and response selection. Low levels of circulating cortisol affect this mode through stimulating MR or Type I receptors in the central nervous system. The second, a 'reactive' or 'suppressive' mode, operates through GR or Type II receptors that mediate highly catabolic events promoting increases in energy availability, modulation

of the immune system, and threat-related learning. GRs are occupied increasingly as cortisol rises above basal levels into stress ranges. Both proactive and reactive modes support health and survival. However, the reactive mode, if sustained, threatens health. In the hypothalamus and hippocampus, occupation of GR or Type II receptors also terminates the stress response, serving as a brake on the duration of the reactive/suppressive stress mode. In animals, early experiences appear to alter the balance of MRs/GRs, lowering the level of hormone at which the system shifts from proactive to reactive functioning and allowing the reactive, suppressive mode to continue for longer periods before cortisol returns to basal levels.

A recent line of developmentally relevant research focuses on mechanisms relating the LHPA system to systems involved in growth and development of the brain and body. Of particular note are studies examining the interaction of glucocorticoids, MR/GR balance and serotonin (5-HT) receptor activity in the hippocampus (López et al. 1998), and studies examining the interaction of glucocorticoids, growth hormone, and insulin growth factors (e.g., IGF-1) (Johnson et al. 1992). In animals, conditions that elevate glucocorticoids chronically also increase basal glucocorticoid levels, affect MR/GR ratios, and lower serotonin activity in the hippocampus. These conditions include stress during pre- and postnatal development. These changes are associated with heightened vulnerability to stressors later in life, as well as problems with behavior regulation, response inhibition, and sensitivity to self-administration of drugs of abuse (de Kloet et al. 1998). Critically, again in animals, conditions associated with responsive caregiving during postnatal development reduce the impact of prenatal stress and buffer the impact of postnatal stressors (Caldji et al. 1998, Levine 1994).

2. *Studies in Infants and Young Children*

2.1 *Measurement Issues*

Research in children has been stimulated by the development of cortisol assays that use samples of free-flowing saliva (Kirschbaum and Hellhammer 1994). Unfortunately, salivary sampling does not allow the assessment of activity higher in the system (i.e., at the pituitary and brain levels). Blood sampling is needed to assess ACTH concentrations, and a spinal tap is typically needed to measure CRH. Particular drugs are usually given to probe the physiology of the LHPA system in ways that bypass individual differences in how the person perceives and interprets events. Because of their invasiveness, none of these techniques is used typically with normally developing children.

In addition, the researcher who studies the LHPA system must attend to many confounding variables that constrain how studies can be conducted. Because of the daily rhythm in the hormones, all samples must be conducted at the same time of day. Careful attention needs to be given to the child's health, what he or she has eaten and drunk recently, and the familiarity of the context (Kirschbaum and Hellhammer 1994). These constraints have limited the study of this system in children. Despite this, a literature is developing which promises to inform our understanding of early risks for later physical and emotional pathology.

2.2 *Developmental Changes in Responsivity*

Prenatally, the LHPA system produces cortisol by the beginning of the second trimester. In animals, the mother's cortisol affects the fetus, resulting in the development of fewer MRs and larger LHPA responses to stressors (Barbazanges et al. 1996). In humans, fetal vulnerability (low birthweight, prematurity) increases with increased maternal psychosocial stress and LHPA activity (Sandman et al. 1997).

The full-term neonate exhibits robust, graded elevations in cortisol that sensitize to repeated pain-eliciting events (e.g., heel lances) and habituates to repeated handling stressors (e.g., physical examinations) (Gunnar 1992). Responsiveness to handling stressors appears to decrease at around 3 months of age, after which stressors such as physical examinations fail to provoke elevations in cortisol for most infants. Another marked decrease in cortisol responsivity is observed between 6 and 12 months of age, after which during infancy it becomes difficult to provoke elevations in cortisol, on average, to many mild stressors (Gunnar 2000).

2.3 *Relationships as Regulators of the HPA system*

By 6–12 months, social interactions with caregivers mediate cortisol responses. Thus, brief maternal separations fail to elevate cortisol if substitute caregivers are sensitive and responsive, while elevations are noted when substitute caregivers are not (Gunnar 2000). Similarly, the availability of the parent in secure attachment relationships buffers elevations in cortisol to events that elicit wariness and behavioral distress, while elevations are noted for toddlers exhibiting similar wariness/distress if the attachment relationship is insecure (Spangler and Schieche 1998). Elevations in cortisol have also been noted for toddlers with the type of disorganized/disoriented attachment more typically noted among infants of depressed and/or abusive parents. Higher home baseline levels of cortisol have also been reported among 3-year-olds whose

mothers were clinically depressed during the child's first year of life. These data are consistent with the literature in animals cited above relating maternal care to stress reactivity and regulation.

Social relations with peers are potent stimuli of the LHPA system in human adults and animals (Kirschbaum and Hellhammer 1994). For children, this appears true at least as early as the preschool years. For toddlers and preschoolers, childcare involving long hours of interaction with like-aged playmates produces elevations in cortisol for many children (Dettling et al. 1999). Even short periods of interaction in half-day preschool programs may elevate cortisol for some children. Children who show activation of the LHPA system in peer settings are those with problematic peer relations; peer rejection, low dominance status, and social withdrawal tend to characterize these children. Problematic styles of interacting (e.g., anger, aggression, and poor self-control) that foster peer rejection also correlate with higher and more responsive patterns of cortisol production in preschool peer groups (Gunnar et al. 1997).

3. *Temperament*

Both extreme shyness and problems in effortful control have been associated with LHPA system activity in young children. Extreme shyness has been hypothesized to reflect altered thresholds for activation of stress physiology involving central fear circuits in the brain (i.e., amygdala–bed nucleus pathways). Thus, increased responsiveness of the LHPA system has long been hypothesized for extremely shy, socially anxious children (Kagan et al. 1987). Strong evidence for this hypothesis is largely lacking when children are assessed under conditions of social challenge (Gunnar et al. 1997); however, shyness, whether measured by teacher report or direct observation, has been found to correlate positively with cortisol sampled at home (de Haan et al. 1998). It is not clear why cortisol correlates more readily with shyness when home hormone concentrations are examined. It may be that under conditions of challenge, moderating factors such as coping strategies and social support need to be taken into account, while home levels are more reflective of tonic influences.

Rothbart has identified effortful control as a higher-order dimension of temperament that emerges from factor analyses of temperament questionnaires during the preschool years. Attention focusing and the capacity to inhibit prepotent responses form the core of this dimension, which Posner and Rothbart (1994) argue reflects the development of an attentional network involving the anterior cingulate gyrus which in animals also plays a role in containing the cortisol stress response (Diorio et al. 1993). In children, effortful control modifies relations between negative emotionality and behavior (Eisenberg et al. 1994). In preschool-aged children, parent and teacher reports of effortful control tend to be negatively correlated with cortisol levels and measures of cortisol responsivity in classroom settings (Gunnar et al. 1997). Whether these correlations reflect mechanisms linking the LHPA system to the neurobiology of effortful control or the impact of effortful control on the behaviors that create stressful social encounters has not been determined.

4. *Neglect, Maltreatment, and Trauma*

Based on the animal research, early neglect and abuse should alter the activity of the developing LHPA system. So far, there is only a small literature on this in humans. Because of the multiple genetic and experiential risk factors typically noted in children with maltreating backgrounds, isolating the impact of adverse early experiences on the human LHPA system promises to be difficult. A full discussion of this emerging literature is beyond the scope of this article. Two findings related to physical and sexual abuse seem consistent enough to mention. First, during periods of active abuse young children do exhibit elevated cortisol levels and heightened responsivity to CRH challenge suggestive of a hyper-responsive LHPA system. Once the child's life circumstances are stabilized, however, for most children this hyper-responsiveness diminishes. The exceptions may be (a) depressed maltreated children who show a flattening of the daily rhythm in cortisol production (Hart et al. 1996), and (b) maltreated children with chronic post-traumatic stress disorder (PTSD) and who have higher baseline cortisol and catecholamine levels that are correlated with the duration of abuse (De Bellis et al. 1999).

In addition, the relations between growth failure, maltreatment, and LHPA system bear comment. Physical abuse and neglect sometimes produce significant growth delays due to reduced pituitary production of growth hormone (GH), cell insensitivity to GH, and blunted GH responses to provocative challenges, alterations that all revert to normal when the child is removed from the maltreating circumstances (Alanese et al. 1994). Theoretically, the LHPA system may be involved in the etiology of this syndrome, called variously psychosocial short stature or psychosocial dwarfism (Johnson et al. 1992). Children adopted from deprived, orphanage conditions exhibit growth failure, and six years after adoption have elevated baseline cortisol levels that were correlated with length of orphanage experience (Gunnar 2000).

5. *Future of Research*

Psychobiological studies of stress in young children have burgeoned since the development of techniques to measure cortisol noninvasively in saliva. Com-

munity studies of children negotiating the normative stressors of childhood are beginning to make connection with studies of children from highly stressful, adverse life circumstances and those with clinical disorders. As in studies of adults and animals, individual differences in response to stressors are common and related to the child's social circumstances and temperament. The challenge for the future is to determine the relation between LHPA function in early development and child outcomes and to explicate the relevant molecular processes.

See also: Attachment Theory: Psychological; Biopsychology and Health; Shyness and Behavioral Inhibition; Stress, Neural Basis of; Stress: Psychological Perspectives; Temperament and Human Development

Bibliography

Albanese A, Hamill G, Jones J, Skuse D, Matthews D R, Stanhope R 1994 Reversibility of physiological growth hormone secretion in children with psychosocial dwarfism. *Clinical Endocrinology* **40**: 687–92

Barbazanges A, Piazza P V, Le Moal M, Maccari S 1996 Maternal glucocorticoid secretion mediates long-term effects of prenatal stress. *Journal of Neuroscience* **16**: 3943–9

Caldji C, Tannenbaum B, Sharma S, Francis D, Plotsky P M, Meaney M J 1998 Maternal care during infancy regulates the development of neural systems mediating the expression of fearfulness in the rat. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 5335–40

De Bellis M D, Baum A S, Birmaher B, Keshavan M S, Eccard C H, Boring A M, Jenkins F J, Ryan N D 1999 A. E. Bennett Research Award. Developmental traumatology. Part 1. Biological stress systems. *Biological Psychiatry* **45**: 1259–70

de Haan M, Gunnar M R, Tout K, Hart J, Stansbury K 1998 Familiar and novel contexts yield different associations between cortisol and behavior among 2-year-olds. *Developmental Psychobiology* **33**: 93–101

de Kloet E R 1991 Brain corticosteroid receptor balance and homeostatic control. *Frontiers in Neuroendocrinology* **12**: 95–164

de Kloet R, Vreugdenhil E, Oitzl M S, Joels A 1998 Brain corticosteroid receptor balance in health and disease. *Endocrine Reviews* **19**: 269–301

Dettling A C, Gunnar M R, Donzella B 1999 Cortisol levels of young children in full day childcare centers: Relations with age and temperament. *Psychoneuroendocrinology* **24**: 519–36

Diorio D, Viau V, Meaney M J 1993 The role of the medial prefrontal cortex (cingulate gyrus) in the regulation of hypothalmic–pituitary–adrenal responses to stress. *Journal of Neuroscience* **13**: 3839–47

Eisenberg N, Fabes R A, Nyman M, Bernzweig J, Pinuelas A 1994 The relations of emotionality and regulation to children's anger-related reactions. *Child Development* **65**: 109–28

Gunnar M R 1992 Reactivity of the hypothalamic–pituitary–adrenocortical system to stressors in normal infants and children. *Pediatrics* (*suppl.*) **90**: 491–7

Gunnar M R 2000 Early adversity and the development of stress reactivity and regulation. In: Nelson C A (ed.) *The Effects of Adversity on Neurobehavioral Development. The Minnesota Symposia on Child Psychology*. Lawrence Erlbaum Associates, Mahwah, NJ, Vol. 31, pp. 163–200

Gunnar M R, Tout K, de Haan M, Pierce S, Stansbury K 1997 Temperament, social competence, and adrenocortical activity in preschoolers. *Developmental Psychobiology* **31**: 65–85

Hart J, Gunnar M R, Cicchetti D 1996 Altered neuroendocrine activity in maltreated children related to symptoms of depression. *Developmental Psychobiology* **8**: 201–14

Johnson E O, Kamilaris T C, Chrousos G P, Gold P W 1992 Mechanisms of stress: A dynamic overview of hormonal and behavioral homeostasis. *Neuroscience and Biobehavioral Reviews* **16**: 115–30

Kagan J, Reznick J S, Snidman N 1987 The physiology and psychology of behavioral inhibition in children. *Child Development* **58**: 1459–73

Kirschbaum C, Hellhammer D H 1994 Salivary cortisol in psychoneuroendocrine research: Recent developments and applications. *Psychoneuroendocrinology* **19**: 313–33

Levine S 1994 The ontogeny of the hypothalamic–pituitary–adrenal axis: The influence of maternal factors. *Annals of the New York Academy of Sciences* **746**: 275–88

López J F, Chalmers D T, Little K Y, Watson S J 1998 A. E. Bennett Research Award. Regulation of serotonin 1A, glucocorticoid, and mineralocorticoid receptor in rat and human hippocampus: Implications for the neurobiology of depression. *Biological Psychiatry* **43**: 547–73

McEwen B S 1998 Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York Academy of Sciences* **840**: 33–44

Nemeroff C B 1996 The corticotropin-releasing factor (CRF) hypothesis of depression: New findings and new directions. *Molecular Psychiatry* **1**: 336–42

Posner M I, Rothbart M K 1994 Attentional regulation: From mechanism to culture. In: Bertelson P, Elen P, d'Ydewalle G (eds.) *International Perspectives on Psychological Science*. Lawrence Erlbaum Associates, Mahwah, NJ, Vol. 1, pp. 41–54

Sandman C A, Wadwha P D, Chicz-DeMet A, Dunkel-Schetter C, Porto M 1997 Maternal stress, HPA activity, and fetal-infant-outcome. *Annals of the New York Academy of Sciences* **814**: 266–75

Spangler G, Schieche M 1998 Emotional and adrenocortical responses of infants to the strange situation: The differential function of emotion expression. *International Journal of Behavioral Development* **22**(4): 681–706

M. R. Gunnar

Psychohistory

1. Its Meaning

The term *psychohistory* (originally written *psychohistory*) emerged in America in the 1950s to designate historical writings utilizing psychology, for the most part depth psychology, as their primary frame of

reference or mode of explanation. Two of its tenets in particular distinguish psychohistory sharply from other systematic approaches to the human past. The first of the two, which is rarely spelled out even in methodological writings, is that the facts of human history consist entirely of what people did whether spontaneously or under pressure of circumstances. To the psychohistorian, an intellectual revolution thus signifies thinkers changing their minds; a decline in fertility translates as women bearing fewer children; urbanization conveys people living increasingly in cities. Other historians may agree that even such seemingly impersonal historic developments all come down to human doings, yet whenever other historians narrate or explain developments of the sort they tend in the first instance to see belief systems or birth rates or urban:rural ratios changing rather than people thinking or acting differently. The other highly distinctive tenet of psychohistory is, by contrast, fully and even emphatically explicit in the psychohistorical literature: that the determinants of human conduct, whether individual or collective, are largely unconscious. Historians of other schools will often freely acknowledge deeply subjective causation of historic human behavior only to shy away from it in their own research as too dauntingly difficult to reconstruct, let alone demonstrate. To the psychohistorian, however, whatever the difficulties involved, there is simply no other way for history to be understood (see *History: Theories and Methods*).

2. *Its Development*

Psychohistory originated as applied psychoanalysis, with Freud and his disciples drawing on their clinical findings now and again to illuminate historical facts (see *Psychoanalysis: Overview*). At first, those clinical findings bore heavily on the patterns and pitfalls of individual emotional development, with a strong emphasis on pregenital sexuality as it emerges in childhood bit by bit and then undergoes repression even while continuing to inform emotional life. Accordingly, the earliest applications of psychoanalysis to history—including those by Freud himself, who long set the standard for his followers—were biographic and focused on the infantile concerns latent in the words and deeds of eminent or paradigmatic historic figures (see *Biography: Historical*). A landmark work in this category was psychoanalyst Erik H. Erikson's *Young Man Luther* of 1958, which argued that Luther's bold quest for spiritual autonomy was that of his whole age writ large (see *Erikson, Erik Homburger (1902–94)*). Unlike biography in relation to historiography as a whole, psychobiography remains to this day the basic as well as the leading form of psychohistory. At the same time, it was as quick as psychoanalysis itself to move beyond its initial Freudian orthodoxy (see *Freud, Sigmund (1856–1939)*). Post-Freudian concepts utilized in psychobiography when this is practiced as applied psychoanalysis have notably included C. G. Jung's 'individuation,' Alfred Adler's 'inferiority complex,' Erik H. Erikson's 'identity crisis,' Heinz Kohut's 'narcissistic rage,' and Jacques Lacan's 'language of the id.'

Psychohistory long continued to show its derivation from psychoanalysis even where it broadened its scope beyond psychobiography. As late as the 1970s psychohistorians who explored social groups, whole cultures, or long-range collective developments, did so as a rule only in terms of the individual psychodynamics thought to have determined them, tracing their salient characteristics to formative experiences widely shared by the individuals involved, and most often to pathogenic childrearing practices on the order of abrupt weaning or intrusive parenting. Correspondingly, the preferred subjects under the head of group behavior were instances of social pathology such as witchcraft, or of political pathology such as Nazism. The most systematic psychohistorical approach to group behavior to emerge on these individualistic premises is prosopography duly slanted: the profiling of the likes of saints or witches or Nazis in terms of variables of psychoanalytical significance in their individual family histories—ages of parents and siblings, for instance, or again early parent or sibling losses (see *Prosopography (Collective Biography)*).

A prominent early alternative to the individual personality as the basic unit even of large-scale, broad-gaged psychohistorical accounts or analyses was the study of a mass leader whose following interacted with him like a single composite being. Freud himself pioneered this departure from the classic individualistic model that underlay his own science of psychoanalysis. In his *Group Psychology and the Analysis of the Ego* (*Massenpsychologie und Ichanalyse*) of 1922 he treated mass phenomena for the most part as simultaneous, parallel regression by individuals, but in a couple of highly suggestive passages he speculated on human consciousness as originally shared by a Darwinian primal horde from which a leader eventually emerged with a mind and will of his own; such shared group consciousness, he considered, might revive at critical historic junctures under the spell of a charismatic leader who threw back to that aboriginal tribal chief. Freud (1939) worked this speculative construction into a highly imaginative reading of the Moses story, with Moses reincarnating the tribal father supposedly slain and later exalted by his sons acting in concert. Given the very nature of this concept of group consciousness as a reversion to a tribal leader and following, subsequent psychohistorical treatments of leader and led have mostly dealt with political collectivities such as Lincoln's or Nixon's America, or again William II's and especially Hitler's Germany.

On the Freudian model of group process, then, group identity constitutes a regressive throwback to an original shared consciousness in the face of a strong

leader. But a rival conception of shared mental life, one that makes it out as permanent and leaderless, came into psychohistory independently. Its *locus classicus* lies outside of psychohistory: in the 'collective consciousness' (*conscience collective*) that Emile Durkheim in his *Suicide* of 1897 inferred from the patterned regularity of suicide rates over and beyond the particular subjective causation of each individual suicide. An early specimen of such a leaderless group model in psychohistory was Caroline E. Playne's *The Neuroses of the Nations* of 1925, which owed nothing to psychoanalysis and indeed wholly ignored it. When William L. Langer, in a resounding presidential address to the American Historical Association in 1957, called the mastery and use of depth psychology the historian's 'next assignment,' he expressly meant its application to leaderless group phenomena, which he exemplified by the epidemic of religious and cultural pathology that followed in the wake of the Black Death of 1348. (Oddly, Langer made no reference to a pioneering work in this mold just then published by his Harvard University colleague Walter Abell (1957) about how the waxing and waning insecurities of medieval European life were projected into Europe's public art of the time.) By the 1970s the dynamics of 'group process,' as it came to be called, topped the psychohistorical agenda, with collective fantasy as its privileged medium of analysis. An exemplary early product of this approach to group process was Paul Monaco's (1976) work, which teased the deep-lying national anxieties of France and Germany in the 1920s out of the recurrent motifs and techniques of each country's most popular movies of those years. Another choice area, still under exploited, for ferreting out the unconscious underpinnings of group activity is historical demography (see *Historical Demography*), especially the adjustments of aggregate birth rates to the felt needs of a population with regard to its overall numbers.

As the term 'group process' itself implies, explanation in psychohistory gradually has shifted from unconscious motives and meanings to unconscious mechanisms governing thought and conduct, while conscious purposes have been even more marginalized in the process. This shift reflects the continuing self-distancing of psychohistory from its parent discipline of psychoanalysis, focused as the latter is on latent impulses originating in childhood. Nowhere is this shift more evident than in the concept of traumatic reliving, which matured within psychohistory well before it was taken up in psychoanalysis. Psychoanalysis had first confronted adult psychic trauma during World War One in the form of battle shock for which, as for other emotional ailments, it sought a childhood referent that in this case was not forthcoming. Comparative psychobiography established two patterned, alternative reactions to emotional trauma: either a continual recall accompanied by the full, tormenting original affect, or a repression of the traumatic affect, and sometimes even of the traumatic memory itself, followed by a contrived, controlled, disguised reexperience or reliving of the trauma to the apparent end, never attained, of mastering the attendant affect that could not be mastered at the time of the experience itself. These same mechanisms have subsequently been shown to operate as well on the collective level in instances of group trauma, and to operate independently of the widely varying reactions of the individuals composing the traumatized group. Precursively, William L. Langer in handing out his 'next assignment' conceived of the after effects of the Black Death, notably its morbid reenactments such as the Dance of Death, as just such a case of group traumatic reliving.

3. *Its Problematic*

Psychohistorical research is uniquely demanding of its practitioners because of its concern with subjective phenomena that are rarely if ever directly recorded. To understand a piece of the human past psychohistorically, researchers must first master the relevant facts at least as thoroughly as the practitioners of any other mode of inquiry. But much more, they must become immersed in those facts, make them vicariously their own, and feel out their inner determinants by dint of this imaginative identification with their original experience. The method is strikingly like the one promoted by Henri Bergson under the name of 'intuition': to study a subject intellectually from the outside until one enters into it subjectively through a core insight around which everything known about it will then fall into place. Such intimate, personal self-projection into a historic subject is trebly problematic. For one thing, it presupposes a basic equivalence of mental and emotional constitution as between researchers and their human subjects, however remote those subjects may be from them historically, and even if they are groups rather than individuals like themselves. For another, researchers' identification with their subject must be thoroughgoing, yet unless they can also sustain some critical awareness outside of that identification the insights gained through it will be lost. And for a third, pieces of the researchers' own mental underworld may slip unnoticed into their reading of a subject. This danger, often called 'countertransference' by loose analogy with psychoanalysts projecting into their patients, is the risk most cautioned against by psychohistorians themselves. Actually, an affinity in researchers for their subject's inner experience can, on the positive side, sensitize or alert them to connections and meanings they might otherwise miss. Besides, a purely projective finding will not stand up in any case inasmuch as it will fail to fit all the evidence, let alone satisfy anyone apart from the researchers themselves.

Researchers will feel the rightness of their inner

grasp of a subject when it not only pulls together all the relevant facts at their disposal, but points the way to unsuspected additional corroborative evidence. Yet the discovery of such new evidence, however suggestive, is in itself no proof that their inner grasp of their subject is universally valid. Because of the very nature of psychohistorical explanation, to validate a psychohistorical thesis takes more than just a wealth of corroborative material, even if the thesis itself is what turned up key pieces of that material. To pass muster, a psychohistorical thesis must also carry subjective conviction, beyond the researchers themselves, with others willing and able to put themselves in the subject's place in their turn and see whether the proposed unconscious source of the thought or action at issue checks out subjectively with them too.

This second, subjective criterion of verification means in brief that a psychohistorical thesis must be emotionally as well as intellectually compelling. Consequently it can rely on no pregiven theory for support. Yet however direct its appeal, its need for subjective on top of objective verification demands too much effort, not to add goodwill, of critics accustomed instead to merely material and logical criteria of validation. As a substitute test of validity I therefore propose that, in psychohistory as in any science, a thesis is confirmed if, first of all, the known evidence all runs its way, if moreover it could potentially be refuted by new evidence, and if finally any piece of what it purports to explain cannot very well be explained otherwise. To take a simple example of this last stipulation from my own thesis that Hitler was traumatized by his mother's iodoform poisoning during her terminal cancer treatment: why else did he put it to his generals in late 1944 that his Ardennes offensive had caught the Allies off guard because they believed he was dead already 'or at all odds am down with cancer somewhere and can't live any more or drink any more' except that deadly iodoform poisoning induces a burning thirst together with an inability to drink? (Binion 1976, p. 143).

4. *Its Crisis Years*

As a historical school in its own right, psychohistory flourished above all in the US during the 1960s and 1970s. Even then, like all new departures, it was highly controversial, especially within the historical professions, because of its constitutive tendency to undercut all other modes of historical explanation. It has since suffered considerably from the decline of psychoanalysis, with which it is rightly associated in that the two still share at least the premise of unconscious motives and meanings. Experimental graduate programs in the field have all folded, while course offerings on college campuses have dwindled to the vanishing point. The ranks of researchers in the field are fast thinning even as the stringent standards of verification invoked for it by critics and adepts alike are discouraging prospective practitioners. Fewer psychohistorical books and articles are appearing year by year. Of the two major specialized journals, *The Psychohistory Review* ceased publication in 1998 while *The Journal of Psychohistory* has lost most of its initial circulation. Perhaps worst of all for the discipline, the consensus of outsiders is that its scholarly output is falling off qualitatively as well.

On the other hand, psychohistory has left its impress on most other kinds of historical inquiry, especially those that emerged or developed later than psychohistory itself. Biographies and even general histories now routinely incorporate psychohistorical materials and interpretations. The study of mentalities often verges on psychohistory. So does that of historic memory, especially in the concern with memory gaps or distortions that has dominated the field. Gender history and gay history, as well as the history of emotions, lean heavily on psychohistorical concepts and findings. Many a term launched in psychohistory, such as Erikson's 'identity crisis,' has become common coinage elsewhere, while the mechanism of traumatic reliving by individuals and groups alike has found its way into a broad variety of scholarly contexts, including even a quantitative history of testation in late medieval Tuscany (Cohn 1992) (see *Identity in Childhood and Adolescence*).

The fact that psychohistory has lost much of its vitality of the 1960s and 1970s, or again that it is being put to subordinate uses by historians of other orientations, does not affect its constitutive claim to provide a privileged access to causality in history. Given this inherently persuasive claim, neither the recent loss of momentum by psychohistory as a method in its own right nor its present auxiliary status in historical research can very well be regarded as permanent.

See also: Emotions, History of; Historiography and Historical Thought: Current Trends; Psychoanalysis, History of

Bibliography

Abell W 1957 *The Collective Dream in Art: A Psycho-historical Theory of Culture Based on Relations between the Arts, Psychology, and the Social Sciences*. Harvard University Press, Cambridge, MA

Aldrich R 1993 *The Seduction of the Mediterranean: Writing, Art, and Homosexual Fantasy*. Routledge, London

Binion R 1976 *Hitler among the Germans*. Elsevier, New York

Binion R 1994 Fiction as social fantasy: Europe's domestic crisis of 1879–1914. *Journal of Social History* **27**: 679–99

Brink A 1989 *Bertrand Russell, A Psychobiography of a Moralist*. Humanities Press International, Atlantic Highlands, NJ

Cohn S K Jr. 1992 *The Cult of Remembrance and the Black Death. Six Renaissance Cities in Central Italy*. Johns Hopkins University Press, Baltimore, MD, pp. 116–17, 288
Demos J P 1970 *A Little Commonwealth: Family Life in Plymouth Colony*. Oxford University Press, New York
Erikson E H 1959 *Identity and the Life Cycle*. International University Press, New York
Freud S 1939 *Der Mann Moses und die Monotheistische Religion*. Verlag Albert de Lange, Amsterdam
Graziano F 1992 *Divine Violence: Spectacle, Psychosexuality, & Radical Christianity in the Argentine 'Dirty War.'* Westview Press, Boulder, CO
Kohut T 1985 Mirror image of the nation: An investigation of Kaiser Wilhelm II's leadership of the Germans. In: Strozier C B, Ofer D (eds.) *The Leader: Psychohistorical Essays*. Plenum, New York
Langer W L 1958 The next assignment. *American Historical Review* **63**: 283–304
Mack J E 1977 *A Prince of our Disorder: The Life of T. E. Lawrence*. Weidenfeld and Nicolson, London
Markowitz J 1969 *The Psychodynamic Evolution of Groups*. 1st edn. Vantage Press, New York
Monaco P 1976 *Cinema and Society: France and Germany during the Twenties*. Elsevier, New York
Roper L M 1994 *Oedipus and the Devil: Witchcraft, Sexuality, and Religion in Early Modern Europe*. Routledge, New York
Saslow J M 1986 *Ganymede in the Renaissance: Homosexuality in Art and Society*. Yale University Press, New Haven, CT

R. Binion

Psycholinguistics: Overview

1. Introduction

Psycholinguistics is a field that combines methods and theories from psychology and linguistics to derive a fuller understanding of human language. From psychology, it inherits experimental methodology and a body of knowledge about processes in perception, memory, attention, learning, and problem-solving. From linguistics, it derives detailed descriptions of specific languages, rigorous accounts of the shape of grammar, and ideas about the nature of human language (Osgood and Sebeok 1967).

The basic issue that motivated the establishment of psycholinguistics as a separate field of study was the problem of the 'psychological reality' of linguistic concepts. For example, speakers of English can form the plurals of nouns by adding the suffix *-s*. This process allows us to form the plural *chandeliers* from *chandelier*. But do we actually use a productive rule to produce this plural or do we simply retrieve the plural chandeliers from our long-term memory as a unit? Psycholinguistic research shows that, in fact, both rote and rule are operative at various times in language production (Pinker 1999).

This issue of psychological reality applies to all levels of language structure and usage, including articulatory phonetics, auditory phonetics, phonology, morphology, lexicon, syntax, semantics, and pragmatics. It applies to all of the constructs of linguistics, including rules, grammars, paradigms, trees, segments, words, and morphemes. When we hear words, do we break them up into their component phonemes and distinctive features (see *Phonology*), or do we recognize them as wholes? When we listen to sentences, do we actually construct treelike representations of the types proposed by linguists or do we somehow extract meaning without building up formal structures? If we do use formal grammars to listen and speak, what is the exact shape of the grammars that we use?

Exploration of the psychological reality of linguistic structures immediately leads us to two related fields of study. The first if developmental psycholinguistics, or the study of child language acquisition (Fletcher and MacWhinney 1995). If we believe that adults form the plurals of nouns by adding *-s*, we need to consider how young children can learn to apply this rule to produce 'cats' and not 'foots' or 'tooths.' If we argue that they simply learn each form by rote, how can we account for the fact that they make errors like 'feets' and the fact that they can produce the plural for a new word like 'wug' even without having been given that form explicitly?

An even more difficult issue involves how rules of grammar might be processed in the brain. Addressing this question has led psycholinguists to explore issues in neurolinguistics (Stemmer and Whitaker 1998) and cognitive neuroscience (Gazzaniga 1997). When the field of psycholinguistics first developed in the 1950s, psychologists knew little about the detailed functioning of the human brain and were forced to treat it as a 'black box.' However, as our understanding of the functioning of the human brain grew during the 1980s and 1990s, it became clear that a precise understanding of the functioning of human language would have to make reference to neural mechanisms. The interactive approach to cognition (McClelland and Rumelhart 1986) used artificial neural networks to model the processing of human language. Directly opposed to connectionism was Fodor's (1983) modular approach to cognition that emphasized the independence of separate cognitive modules for each level of linguistic structure. Researchers have attempted to test the contrasting predictions of the interactive and modular approaches using standard experimental methodology. However, this work has indicated that neither of the strong positions can be maintained (Simpson 1994). To better understand the mechanisms involved, psycholinguists are now trying to link experimental methodology to methods for the imaging of the human brain during language processing. On this level, it appears that processing works in terms of interactive modules.

This article will examine research in six core areas of psycholinguistics: spoken word recognition, sentence comprehension, sentence production, message construction, memory limitations, and cross-linguistic comparisons. In addition to these core areas, psycholinguists are involved in the study of reading, conversational interaction, figurative language, text comprehension, aphasia, child language disorders, gesture, prosody, neurolinguistic imaging, animal communication, and language evolution. However, our analysis here will focus on these six core areas.

2. *Word Recognition*

When we listen to speech, we perceive words as following each other in clear temporal succession, like the beads on a string or the boxcars on a railway train. However, these perceptions underestimate the extent to which words are actually being blurred together by coarticulation and assimilation. Extreme examples of this blurring occur in forms such as '*supchu*' for 'What's up with you?' However, even nonphrasal sequences such as 'I owe you a yo-yo' show similar blurring. Even without blurring, the job of segmenting sentences into words would be a tough one. Consider a phrase such as 'my catalog value for Mark's piece.' Within this phrase, there are fragments that match a variety of other possible words, such as 'mike,' 'eye,' 'I,' 'cat,' 'at,' 'cattle,' 'log,' 'you,' 'Val,' 'ark,' 'arks,' 'are,' and 'form.' The reason why we tend not to hear these alternative forms is fairly simple. If we decided to commit ourselves to having heard the word 'mike,' we would end up with the nonsensical segmentation of 'mike at a log value for Mark's piece.' At each point during sentence perception, many of these alternative words are partially active in a short list of competitors. Competitors that recognize larger segments such as 'catalog' are preferred over those that recognize smaller strings or which break up larger strings into pieces, such as 'cat a log.' If a competitor leaves an unrecognized fragment, then its own recognition is weakened. These constraints work together to guarantee a maximally satisfactory segmentation. Although segmentation is primarily driven by the competition of words for matches, it is also facilitated by stress patterns, pauses, and other prosodic patterns (Norris 1994).

Segmentation relies on word recognition, but this process is also highly dynamic. Words do not have clear and invariant forms. Consonants leave no clear and invariant imprint on the auditory stream, since they are heavily blended into vowels. Because of differences in their vocal tracts, men and women produce vowels in radically different ways. Dialect differences and variations in speech level and further variability to word forms. Again, the solution to this problem involves lexical competition. For example, in the NAM model (Goldinger et al. 1989), the sound corresponding to a word like 'deep' will activate a neighborhood of similar words, such as 'deal,' 'dear,' 'peep,' or 'keep.' Within this domain of rhyming forms, competition is particularly keen between words that share the same initial segments, such as 'deep,' 'deal,' 'dear,' and 'deed.' The ability of a word to dominate in this competition is a function of the strength of its match to the features of the input.

These competitive models can be given a specific neuronal instantiation in terms of a neural network model based on self-organizing feature maps (Miikkulainen 1993). In these models, an unorganized, but interconnected, flat sheet of simulated neural tissue is trained to recognize a set of input words. As learning progresses, words that share features move to adjacent areas in the feature map. In our example, the words 'deep,' 'deal,' 'dear,' and 'deed' would occupy a small neighborhood on the larger map. The network operates in a winner-take-all fashion so that, when 'deep' starts to receive the highest level of activation, it will inhibit the activation of the competitors through a series of lateral connections. Maps of this type mimic actual cortical tissue in their use of position to represent featural structure and their reliance on lateral inhibition to sharpen the outcome of a competition.

The study of the precise mechanics of word recognition has relied on studies of work reading rather than listening. The advantage of studying visual word recognition is that experimenters can use the laboratory computer to control tightly the display of words on the computer screen. The tasks most commonly used in this area are naming and lexical decision. Both tasks are sensitive to frequency, neighborhood effects, grammatical relations, and priming effects. For example, the word 'doctor' will prime a subject's naming of the word 'nurse' and the word 'govern' will prime naming of the words 'government' and 'governing,' but not 'misgovern.' This means that the time to read the word 'nurse' out loud will be less when it follows 'doctor,' than when it follows a control word like 'house.' The details of these various priming effects provide psycholinguists with a powerful method for mapping out the shape of the mental lexicon or dictionary (Balota 1994).

3. *Sentence Comprehension*

At one time, psycholinguists thought that the process of word recognition preceded the process of sentence comprehension. The idea was that we must first identify all the words in a sentence before we can feed these words to a comprehension mechanism that decides what it all means. In the late 1970s, Marslen-Wilson (1975) and others showed that sentence processing is not blocked out in this way. Instead, processing keeps up with word recognition in a fully

incremental, or online fashion. Both word recognition and sentence comprehension attempt to keep up with speech exactly as it comes in, although sometimes there is a slight lag, particularly for sentence comprehension. You might notice this lag if you find that you are trying to read a newspaper or watch a television program while someone is talking to you. You hear their words, but you may not be processing exactly what they say. They may then stop and ask you whether you have been listening. Somewhat dishonestly, you reply that you have and to prove this you repeat the last seven or eight words they have said. In fact, you were not really understanding the message in any very deep way. Rather, you were squirreling away the words in a superficial form in your short-term memory, hoping not to have to actually do the work of fully comprehending the message.

3.1 *Incrementalism*

The previous example is interesting, because it illustrates the exception to the general rule of incrementalism. Generally speaking, we process incoming material incrementally, both lexically and conceptually. Take as an example the sentence, 'The boy chased the baboon into the bedroom.' As soon as we hear 'the' and 'boy,' we immediately begin to relate them to each other. We then relate this unit to the following verb 'chased.' Milliseconds after hearing 'chased,' we begin to construct an interpretation of the activity in which there is a boy doing some chasing. By the time we hear 'baboon,' we can begin to sketch out the figure that the boy is chasing. We do not need to wait until we have heard all the words in the sentence to begin to extract these meanings. In this sense, sentence processing is both interactive and incremental—we tend to make decisions about material as soon as it comes in, hoping that the decisions that we make will not be reversed. Although processing is incremental, it only builds as much structure as it needs in order to keep words related. The real job of comprehension is delayed until more of the message is heard.

3.2 *Garden-pathing*

There are times when the initial decisions that we have made take us down the garden path. A classic example of garden-path processing occurs with sentences such as 'The communist farmers hated died.' It often takes the listener awhile to realize that it was the 'communist' that died and that it was the 'farmers' who hated the 'communist.' Inclusion of a relativizer to produce the form, 'The communist that farmers hated died' might have helped the listener sort this out. A somewhat different example is the sentence, 'The horse raced past the barn fell.' Here, we need to understand 'raced past the barn' as a reduced relative clause with the meaning 'The horse who was raced past the barn.' If we do this, the appearance of the final verb 'fell' after 'barn' no longer comes as a surprise.

Garden paths arise when a word or suffix has two meanings, one of which is very common and one of which is comparatively rare (MacDonald et al. 1994). In a sentence like 'The horse raced past the barn fell' the use of the verb 'raced' as a standard transitive verb is much more common than its use as the past participle in a reduced passive. In such cases, the strong meaning quickly dominates over the weak meaning. By the time we realize our mistake, the weak meaning is fully suppressed by the strong meaning and we have to try to comprehend the sentence from scratch. A classic garden-path example from Lashley (1951) is the sentence 'Rapid righting with his uninjured left hand saved from destruction the contents of the capsized canoe.' When this sentence is read aloud, listeners find it extremely difficult to understand the second word as 'righting' rather than 'writing.'

3.3 *Lexical Effects*

Current models of sentence processing emphasize the extent to which lexical and syntactic ambiguities of individual words trigger competing syntactic structures (Trueswell and Tanenhaus 1994). For example, there are two readings of the sentence 'Flying planes can be dangerous.' The planes can be either dangerous to their pilots and passengers or dangerous to onlookers down on the tarmac. Both interpretations of the participle 'flying' are fairly strong. Because the two readings are of similar strength, they can compete with each other and no garden-pathing arises. Another example of this type is 'He bought her pancakes' in which 'her' can be either the possessor of the pancakes or the recipient of the pancakes. Both meanings are strong and can compete with each other during sentence processing, yielding a clear ambiguity.

Sometimes lexically based expectations can be fairly complex. Consider these sentences in which the verbs 'criticize' and 'apologize' set up contrasting expectations:

John criticized Mary, because she hadn't delivered the paper on time.

John apologized to Mary, because he hadn't delivered the paper on time.

John criticized Mary, because he hadn't delivered the paper on time.

John apologized to Mary, because she hadn't delivered the paper on time.

Processing of the first two sentences is quick and easy, because the gender of the pronoun matches that of the expected agent. However, the processing of the second pair is more problematic, because the gender of

the pronoun forces the selection of an unexpected cause or of the criticism or the apology (McDonald and MacWhinney 1995).

3.4 Modularity

Although processing on individual linguistic levels is highly incremental, the interaction between levels is not immediate. During the first 300 milliseconds after hearing a word, we attend primarily to its auditory shape, rather than to the degree to which it fits into some grammatical context. Take as an example the sentence 'The sailors took the port at night.' Here the word 'port' could refer to either the wine or the harbor. We can ask subjects to listen to sentences like this while watching a computer screen. Directly after subjects hear the word 'port' we can present one of these three words on the computer screen: 'wine,' 'harbor,' and some control word such as 'shirt.' If we do this, we will find that the recognition of both 'wine' and 'harbor' is facilitated in comparison to the control word 'shirt.' If we change the sentence to something like 'The sailors drank the port at night,' we might expect that the context would bias the subject to respond more quickly to 'wine' than to 'harbor,' because one is not likely to drink a harbor. However, there is evidence that both 'wine' and 'harbor' are facilitated in comparison to the control word 'shirt,' even when the context tends to bias the 'wine' reading of 'port.' This facilitation is fairly short-lived and the contextually appropriate reading soon becomes dominant.

This type of result indicates that, in the first fraction of a second after hearing a word, we rely most strongly on auditory cues to guide our processing. This is not to say that context is not present or not being used as a cue. However, during the first fraction of a second, we need to focus on the actual auditory form in order to avoid any 'hallucinatory' effects of paying too much attention to context too soon in processing.

4. Sentence Production

There are many similarities between sentence comprehension and sentence production. In both activities, we rely heavily on the words in our lexicon to control syntactic structures. Both activities make use of the same patterns for determining grammatical structures. The most important difference is that, during sentence production, we are in full control of the meanings we wish to express. In comprehension, on the other hand, we are not in control and have to follow the ideas of others.

The production of sentences involves at least four processes (Levelt 1989). The first process is message construction. This process takes our goals and intentions and builds up a thread of ideas to be articulated. The process of lexical access then converts these ideas in word form. The third process uses positional patterning to order words into phrases and clauses. The fourth process activates a series of verbal gestures through articulatory planning. As in the case of comprehension, these four stages are conducted not in serial order, but in parallel. Even before we have finished the complete construction of the message underlying a sentence, we begin the process of articulating the utterance. Sometimes we find out in the middle of an utterance that we have either forgotten what we want to say or do not know how to say it. It is this interleaved, incremental, online quality of speech production that gives rise to the various speech errors, pauses, and disfluencies that we often detect in our own speech and that of others.

Speech errors come in many different forms. Some involve simple slurring of a sound or retracing of a group of words. Others provide more dramatic evidence of the nature of the language planning process. Some of the most entertaining speech errors are spoonerisms, which owe their name to an English clergyman by the name of William Spooner. Instead of 'dear old queen,' Spooner produced 'queer old dean.' Instead of 'ye noble sons of toil,' he produced 'ye noble tons of soil.' Instead of 'I saw you light a fire in the back quad, in fact you wasted the whole term,' he said 'I saw you fight a liar in the back quad, in fact you tasted the whole worm.' These errors typically involve the transpositions of letters between words. Crucially, the resulting sound forms are themselves real words, such as 'liar,' 'queer,' and 'worm.' The tendency of these errors to produce real words is known as the 'lexical bias' in speech errors and indicates the extent to which the lexicon itself acts as a filter or checker on the articulatory process.

Another illuminating group of errors is named after a character named Mrs. Malaprop in the play *The Rivals* by Sheridan. Some examples of malapropisms include 'Judas Asparagus' for 'Judas Iscariot,' 'epitaphs' for 'epithets,' or 'croutons' for 'coupons.' Some malapropisms arise when uneducated speakers attempt to produce unfamiliar words, but many of these errors are true slips of the tongue. In a malapropism, the two words have a similar sound but a very different meaning. The fact that speakers end up producing words with quite the wrong meaning suggests that, at a certain point during speech planning, the output processor handles words more in terms of their phonological form than the meaning they express. It is at this point that malapropisms can occur.

When two words end up competing directly for a single slot in the output articulatory plan, the result is a lexical blend. For example, a speaker may want to describe both the flavor and the taste of some food and end up talking about its 'flaste.' Or we might want to talk about someone's 'spatial performance' and end up talking about their 'perfacial performance.' These

errors show the extent to which words are competing for placement into particular slots. When two words are targeting the same slot, one will usually win, but if there is no clear winner there can be a blend.

Another remarkable property of speech errors is the way in which grammatical markers seem to operate independently of the nouns and verbs to which they attach. Consider an exchange error such as 'the floods were roaded' for 'the roads were flooded.' In this error, the plural marker *-s* and the past tense marker *-ed* go to their correct positions in the output, but it is the noun and the verb that are transposed. This independent behavior of stems and their suffixes indicates that words are being composed from their grammatical pieces during sentence production and that grammatical markers contain clear specifications for their syntactic positions.

Some of the earliest thinking on the subject of speech errors was done by Sigmund Freud at the beginning of the century. Freud believed that slips of the tongue could provide evidence for underlying anxieties, hostilities, and worries. From this theory, arose the notion of a 'Freudian slip.' We now know that the majority of speech errors are not of this type, but it still appears that at least some could be viewed in this way. For example, if a speaker is discussing some activities surrounding a local barber, he might say 'He made hairlines' instead of 'He made headlines.' These errors indicate contextual influences from competing plans. However, even rare examples of this type of contextual influence seldom reveal underlying hostilities or neuroses.

5. *Message Construction and Perspective*

The basic goal of both sentence comprehension and sentence production is the linking of spoken forms to underlying conceptual meanings. In the 1970s, psycholinguists thought of these conceptual structures in terms of links between simple sentences or propositions. These propositions represented relations that had no clear relation to the actual physical perspective of the speaker. More recently, work in philosophy, robotics, psychology, and cognitive linguistics has recast this view of meaning structures in terms of dynamic activities that are linked to the embodied perspective of the speaker.

Consider what occurs when a speaker relates a fable, such as the story of 'the Tortoise and the Hare.' In this story, there are basically three perspectives: one for the tortoise, one for the hare, and one for the narrator. The narrator says 'the hare decided to rest.' At this point, the narrator is taking the perspective of the hare and describing the hare's actions from that perspective. Next, the narrator shifts to the perspective of the tortoise and describes how he passes the sleeping hare. These various shifts and blends of perspective are important aspects of all narrative structures. However, we can see their impact everywhere in language. For example, if we want to point out that there is a dent in the side of the refrigerator, we may use the impersonal construction, 'someone knocked a dent in the refrigerator' or we may use the pseudo-passive, 'the refrigerator got a dent knocked into it.' In the impersonal we are forced to point a finger at an unspecified perpetrator of the deed. In the pseudo-passive, we background or de-emphasize the doer of the activity, focusing instead on the result. These choices all reflect variations in perspective.

Another choice we make is between alternative perspectives on a single action. For example, we could describe a picture by saying, 'the girl gave a flower to the boy.' In this case, we are taking the viewpoint or perspective of the girl and describing the activity of giving from her perspective. Alternatively, we could say, 'the boy got a flower from the girl.' In this alternative form, we view the action from the perspective of the recipient. Depending on whether we choose to view the action from the viewpoint of the giver or the receiver, we will use either the verb 'give' or the verb 'get' Language is full of choices of this type, and sentence production can be viewed as involving a continual competition between choices for perspectives. The selection of pronouns is determined by the need to link up referents to earlier perspectives. Shifts from the present tense to the past also reflect movements in perspective to various points in time.

6. *Cognitive Resources*

Our ability to process language is shaped by our basic cognitive capacities. Even people with severe mental retardation acquire a basic control over spoken language. On the other hand, normal speakers vary in the degree to which they can remember long strings of words. The notion of a verbal working memory (Gathercole and Baddeley 1993) has figured heavily in attempts to derive the psychological reality of linguistic rules. For example, a sentence with multiple center embeddings, such as 'the man the dog the cat scratched bit smiled' are impossible for most people to understand. Psycholinguists believe that the problem here lies not in the violation of a linguistic rule, but in the fact that this structure places unrealistic demands on memory. Variations in memory span or capacity have also been related to differences in ability to learn words in a new language (Gupta and MacWhinney 1997), process complex sentences (Just and Carpenter 1992), and acquire reading skills in school.

Capacity limitations also shape the ways in which we form long-term memories from language. In general, the exact surface form of sentences often fades within seconds after we hear them (Potter and Lombardi 1990). This is particularly true of the

commonplace, predictable material that is used in textbooks or in experiments on sentence processing. However, if one examines memory for more lively, charged, interpersonal communications, a very different picture emerges. For this type of discourse, memory for the specific wording of jokes, compliments, insults, and threats can be extremely accurate, extending even over several days (Keenan et al. 1977). Often we are careful to note the exact phrasing of language spoken by our close associates, since we assume that this phrasing contains important clues regarding aspects of our interpersonal relations. For material that contains no such interpersonal content, we focus only on underlying meaning, quickly discarding surface phrasings.

7. *Cross-linguistic Comparisons*

All languages have sound segments, syllables, words, phrases, and sentences. In all languages, processing is online, incremental, and interactive. However, languages differ markedly in the particular grammatical contrasts they choose to mark and the devices that they use to mark them. In languages such as Warlpiri or Hungarian, word order is remarkably flexible. In languages such as Navajo or Eskimo, a single complex word may express what English does in a seven-word sentence.

Psycholinguistic comparisons of processing in different languages (MacWhinney and Bates 1989) use a simple common task to examine the effects of radically different structures. In one type of experiment, subjects listen to a sentence such as 'The eraser is pushed by the cat' and have to point to a picture that shows who is the actor. Consider a comparison of these sentences from English and Spanish:

The lion kisses the cow.

El león besa la vaca. (The lion kisses the cow).

A major difference between these two languages revolves around the use of variable word orders. In Spanish, it is possible to invert the word order and produce *la vaca besa el león*, while still meaning that the lion is kissing the cow. This inversion is even clearer if the particle *a* is added to mark the direct object, as in these variant orderings in which the two nouns are moved into different places around the verb *besa*.

El león besa a la vaca.

A la vaca besa el león.

Besa el león a la vaca.

Besa a la vaca el león.

These differences between English and Spanish can be traced to the relative cue validities of word order and object marking in the two languages. In English, it is virtually always the case that the noun that appears before the verb is the subject of the sentence. If the verb is an active verb, the preverbal noun is almost always the actor. This means that the cue of preverbal positioning is an extremely reliable guide in English to assignment of the actor role. In Spanish, there is no such simple rule. Instead, the best cue to assignment of the actor role is the presence of the object marker particle *a*. If one of the nouns in a two-noun sentence is marked by *a*, then we know that the other noun is the agent or subject.

Other languages have still other configurations of the basic cues to sentence interpretation. For example, Hungarian makes reliable use of a single case marking suffix on the direct object. German uses the definite article to mark case and number. The Australian language Warlpiri marks the subject with a variety of complex case markings. Navajo places nouns in front of the verb in terms of their relative animacy in the Navajo 'Great Chain of Being' and then uses verbal prefixes to pick out the subject. These languages and others also often rely on number agreement between the verb and the noun as a cue to the subject. English also requires subject-verb agreement, but this cue is often missing or neutralized. In languages such as Italian or Arabic, subject-verb agreement marking is extremely clear and reliable.

When a monolingual speaker comes to learning a second language, they need to fundamentally retune their sentence-processing mechanism. First, they need to acquire a new set of grammatical devices and markings. Second, they need to associate these new devices with the correct cue validities. Third, they need to reorganize their expectations for particular sequences of cues and forms. Initially, the learner simply transfers the cues, cue validities, and habits from the first language to processing of the second language. Over time, the cue validities change smoothly, until they eventually match that of the native monolingual (de Groot and Kroll 1997).

8. *Summary*

In this article we have examined six core issues being addressed by psycholinguists. These areas are central to the study of psycholinguistics, because they allow us to evaluate the psychological reality of linguistic formalisms. The interaction between psycholinguistics and linguistic theory has been intense, reciprocal, and dynamic. We have learned that language processing is interactive and incremental in regard to word recognition, sentence comprehension, sentence production, and message construction. We know that cross-linguistic differences in processing are closely related to differing levels of reliability for the cues involved and the ways in which structures impose demands on verbal memory capacity. Some versions of linguistic theory have moved to adapt to some of these findings from psycholinguistics (Pollard and Sag 1994). However, a satisfactory resolution of the core

issue of the psychological reality of linguistic structures has not yet been obtained.

See also: Computational Psycholinguistics; Connectionist Models of Language Processing; First Language Acquisition: Cross-linguistic; Language Acquisition; Language and Thought: The Modern Whorfian Hypothesis; Second Language Acquisition

Bibliography

Balota D A 1994 Visual word recognition: The journey from features to meaning. In: Gernsbacher M (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA

de Groot A M B, Kroll J F 1997 *Tutorials in Bilingualism*. Erlbaum, Mahwah, NJ

Fletcher P, MacWhinney B J (eds.) 1995 *The Handbook of Child Language*. Blackwell, Oxford, UK

Fodor J 1983 *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press, Cambridge, MA

Gathercole V, Baddeley A 1993 *Working Memory and Language*. Erlbaum, Hillsdale, NJ

Gazzaniga M (ed.) 1997 *The Cognitive Neurosciences*. MIT Press, Cambridge, MA

Goldinger S D, Luce P A, Pisoni D B 1989 Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language* **28**: 501–18

Gupta P, MacWhinney B J 1997 Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language* **59**: 267–333

Just M, Carpenter P 1992 A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* **99**: 122–49

Keenan J, MacWhinney B J, Mayhew D 1977 Pragmatics in memory: A study in natural conversation. *Journal of Verbal Learning and Verbal Behavior* **16**: 549–60

Lashley K 1951 The problem of serial order in behavior. In: Jeffress L A (ed.) *Cerebral Mechanisms in Behavior*. Wiley, New York

Levelt W J M 1989 *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA

MacDonald M C, Pearlmutter N J, Seidenberg M S 1994 Lexical nature of syntactic ambiguity resolution. *Psychological Review* **101(4)**: 676–703

MacWhinney B J, Bates E (eds.) 1989 *The Crosslinguistic Study of Sentence Processing*. Cambridge University Press, New York

Marslen-Wilson W 1975 Sentence perception as an interactive parallel process. *Science* **189**: 226–7

McClelland J L, Rumelhart D E 1986 *Parallel Distributed Processing*. MIT Press, Cambridge, MA

McDonald J L, MacWhinney B J 1995 The time course of anaphor resolution: Effects of implicit verb causality and gender. *Journal of Memory and Language* **34**: 543–66

Miikkulainen R 1993 *Subsymbolic Natural Language Processing*. MIT Press, Cambridge, MA

Norris D 1994 Shortlist: A connectionist model of continuous speech recognition. *Cognition* **52**: 189–234

Osgood C, Sebeok T A (eds.) 1967 *Psycholinguistics: A Survey of Theory and Research Problems With a Survey of Psycholinguistic Research, 1954–1964*. Indiana University Press, Bloomington, Indiana

Pinker S 1999 *Words and Rules: The Ingredients of Language*. Basic Books, New York

Pollard C, Sag I 1994 *Head-driven Phrase Structure Grammar*. Chicago University Press, Chicago

Potter M, Lombardi L 1990 Regeneration in the short-term recall of sentences. *Journal of Memory and Language* **29**: 633–54

Simpson G B 1994 Context and the processing of ambiguous words. In: Gernsbacher M (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA

Stemmer B, Whitaker H 1998 *Handbook of Neurolinguistics*. Academic Press, New York

Trueswell J C, Tanenhaus M K 1994 Toward a lexicalist framework for constraint-based syntactic-ambiguity resolution. In: Clifton C, Rayner K, Frazier L (eds.) *Perspectives on Sentence Processing*. Erlbaum, Hillsdale, NJ, pp. 155–79

B. J. MacWhinney

Psychological Anthropology

Psychological anthropology considers connections between the individual and sociocultural milieu, including cultural influences on personality and psychological foundations of society and culture. The field comprises studies of child rearing, self-representation, emotion, motives, moral reasoning, cognition, dreaming, mental disorder, gender relations, violence, racism, and cultural symbolism, among other things. Theoretical and methodological perspectives include cultural psychology, ethnopsychology, psychoanalytic anthropology, cross-cultural studies of child development, evolutionary psychological anthropology, and cognitive anthropology (see *Cognitive Anthropology*).

Although it shares concerns with several related fields (i.e., social psychology, cultural psychology (in psychology), cross-cultural psychology, and 'applied' psychoanalysis), psychological anthropology differs from these enterprises in being strongly committed to ethnographic fieldwork in diverse cultures. This engagement with culture (and cultures) leads psychological anthropologists to put two demanding questions to theories of mind and personality: they ask if such theories adequately consider cultural influences on the individual, and whether they can in any way illuminate cultures, particularly the symbolic content of expressive culture and the logic of local knowledge. Whereas the first question poses doubts about the universal validity of Western psychology, the second, pointing in the opposite direction, encourages interest in psychoanalysis and cognitive science, psychoanalysis because it addresses content and cognitive science because it addresses logic.

Some cultural determinists or culturalists define their work as 'cultural psychology.' In their view,

'psychological anthropology' implies universalism and reductionism and, correlatively, too little concern with local cultures. Concurring with this appraisal, still others narrow their concerns to 'ethnopsychology,' to local cultural ideas about the mind or person. Yet for many others 'psychological anthropology' remains a comprehensive enterprise that allows for both cultural influences on subjectivity and for psychological interpretations of society and culture. It may also include evolutionary perspectives on human nature and attention to precultural social environment in child development.

1. *Culture and Personality*

Franz Boas, the founder of modern American anthropology, was interested in general psychology or human nature as well as 'folk psychology,' the way in which the 'genius of peoples' is shaped by local history and culture (Stocking 1992). Many of his students, however, were more influenced by his interest in this second, cultural psychology than by his interest in general psychology. Margaret Mead argued for the plasticity of sexuality and gender relations, and she rebutted Piaget and others effectively who were asserting similarity between the thinking of 'primitives' and children. Ruth Benedict contended that cultures inculcate radically different worldviews and, therefore, that subjectivity varies substantially from one culture to another. In yet another version of this cultural determinism, the neo-Freudian psychoanalyst Abram Kardiner, working with several anthropologists, suggested that the 'primary institutions' of a culture (i.e., family organization, child-rearing practices) produce a distinctive 'basic personality structure.' This shared personality core, in turn, informs 'secondary institutions' (e.g., art, myth, ritual, politics) through projection. The model guided attempts to measure basic personality structure with projective personality tests and to document the effects of child rearing with cross-cultural studies, such as those carried out by John W. M. Whiting and his many students.

The principal tenets of this 'culture and personality' theorizing and research were congruity between personality and culture, and reproduction of shared personality and culture through child-rearing practices. By the late 1960s, these tenets were in doubt. Life histories and projective personality testing had demonstrated shared psychological traits in communities, but they had also revealed much individual variation. The assumption that child-rearing practices have robust consequences for personality and culture also seemed less certain. Edward Sapir, Boas's most brilliant student, expressed reservations about culture and personality well before these empirical results. In his view, personality and culture are interrelated, partly because they have a common nexus in symbolic communication and social interaction, but he also emphasized that individual personality is not simply a passive effect of social environment. Observing that cultural knowledge and practices are unevenly distributed in groups and that the same cultural practices can have diverse meanings and emotional entailments for individuals, Sapir inferred that the individual is the creative locus and foundation of culture. Although he had misgivings about it, Sapir concluded that psychoanalysis was the best available language for describing individual psychology.

More in keeping with Sapir's thinking, some psychoanalytically oriented anthropologists focused on expressions of human nature and individual personality in cultural beliefs and practices (e.g., Géza Róheim, George Devereux, Weston La Barre, and Melford E. Spiro). Róheim, the first of these psychoanalytic anthropologists, was especially critical of culture and personality for ignoring human nature and for underestimating cultural and psychological diversity within societies and communities.

2. *Social Anthropology and Psychoanalysis*

In the early twentieth century British anthropologists turned their attention from social evolution to actual social life, including questions about social actors. Accordingly, they were interested in the newly emerging fields of social psychology and psychoanalysis. Both C. G. Seligman and W. H. R. Rivers read Freud. Bronislaw Malinowski especially was intrigued by the conflict between sexual impulses and social norms in psychoanalytic theorizing. A pragmatic empiricist, Malinowski nonetheless had difficulty grasping hidden structures and processes, whether social or psychological. Assuming that intrapsychic complexes were evident in observable social relations, he argued that the Oedipus complex takes a quite different form in the matrilineal Trobriands. When the response among psychoanalysts to this intervention was less than positive, Malinowski's already ambivalent interest in psychoanalysis waned considerably (Stocking 1986).

Various expressions and degrees of ambivalence about psychoanalysis were in fact common in British social anthropology. While social anthropologists sought to understand culture in terms of social relations, it was apparent that the drama of social life involves motivated social actors with personal histories. They could not help but notice, moreover, that expressive culture often includes oedipal themes and bodily imagery (e.g., Meyer Fortes, Edmund Leach, and Victor Turner but especially John Layard, Kathleen Gough, T. O. Beidelman, and L. R. Hiatt).

3. *Psychological Anthropology*

As the basic tenets of culture and personality became more questionable, the field began in the 1960s to reinvent itself under the heading of 'psychological anthropology.' There was less interest in typifying

entire cultures and more concerted efforts to explore specific issues. Tensions between culturalist and psychoanalytic or otherwise psychological approaches have persisted, however. Recently, the theoretical scene has been further complicated by the emergence of evolutionary psychological anthropology, a perspective that competes with psychoanalytic anthropology in some respects and complements it in others.

3.1 Human Nature

Culturalists contend that human nature lacks content, except perhaps for a capacity for learning language and culture. Many psychoanalytically oriented anthropologists are convinced, however, that human beings are desiring, emotional, and moral creatures as well as creatures capable of learning language and culture (Ingham 1996). For their part, evolutionary psychological anthropologists argue that human beings come equipped with inclinations and capabilities oriented toward ensuring inclusive fitness. The evolutionary and psychoanalytic perspectives converge in some respects, particularly in their attention to deception and sexuality. Proposals by evolutionary psychological anthropologists and evolutionary psychologists that childhood experience may shape reproductive strategies also converge in interesting ways with thinking in psychoanalytic anthropology. The ethnographic evidence for these varying views of human nature is ambiguous. While cultural diversity argues against an essential human nature, recurrent themes in the ethnographic data appear to have just the opposite implication. In support of a common human nature, it is noticed, for example, that oedipal themes and culture heroes and mischievous tricksters are nearly everywhere present in myth and folklore (Johnson and Price-Williams 1996).

3.2 Incest Aversion

Although there are attempts to explain incest aversion in cultural terms, the most lively discussions involve psychoanalytic and evolutionary, neo-Westermarckian perspectives. Wolf's work in particular evokes debate, partly because it purports to have devastating consequences for psychoanalysis. Wolf adduced data on Chinese *simpua* marriage (a son marries an adopted daughter with whom he has a 'sisterlike' relationship) to argue that familiarity in childhood inhibits sexual desire (the *simpua* marriages result in higher rates of infertility and divorce). Spain (1987) responded that both Freud and Westermarck predict no incestuous behavior as a result of normal child development. Subsequently, Wolf (1995), again dismissing Freud, drew on attachment theory to argue that the familiarity effect begins as early as the second year of life, presumably before the oedipal period. Nonetheless, a further response to Wolf seems likely. Wolf did not investigate mother–child relationships or childhood sexuality, and he does not consider that attachment theory is a branch of psychoanalysis and that there is clinical literature that implicates early attachment in the development and resolution of childhood oedipality.

3.3 Embodiment and Desire

It has become increasingly common to assert that the body and desire are socially or culturally constructed. 'Embodiment' in this thinking refers not to the bodily grounding of the self or mind but to the way in which culture or discourse shapes ideas about the body and bodily experience. The experience of embodiment is then expressed in 'narratives' about the body.

A different approach to embodiment assumes a more active, motivated subject in which motivation and body image reflect unconscious object relations and internalized moral values as well as the more immediate effects of social discourse. In a study of a recovered anorexic woman, for example, Banks (1998) shows that the woman's changing personal narratives about eating had more to do with her emotional experiences, personal relationships, and religious convictions than dominant cultural discourses about beauty and thinness. A fantasy story about a cookie jar, written in a diary at 11 years of age, implies that the woman's anorexia during adolescence was a way of organizing memories of maternal neglect in early life and religiously inflected anxieties about the dangers of sexuality. The temptation of the cookies and the decision to forego them in the story anticipated the way in which anorexia would become a defense against sexuality during adolescence. Far from being a passive self, the woman was an active subject who responded to discourses about the body in terms of her personal experience and deeply internalized religious convictions.

Culturalists and discourse theorists say that sexual desire and romantic love are 'socially constructed' or 'invented' along with the body, and, to be sure, they provide ample ethnographic material in support of this contention. It is not clear, however, how discourse fashions motivation or affect in desire and love. Cognitive and culturalist anthropologists theorize that motives are represented in the mind as cultural schemas (D'Andrade and Strauss 1992). A case has been made for this approach but, again, it is uncertain how schemas 'stick' in the mind or how they acquire affective force. As Banks's study suggests, cultural schemas may take hold in intrapsychic object relations, that is, in unconscious representations of emotionally significant relationships, especially with parents. If this is so, it does not suffice to say that the

body or desire is socially constructed. Developmental studies of individuals challenge simplified versions of social constructionism but so, too, does the evidence for cultural universals. One of these challenges is the finding that more or less durable passionate attachment is virtually universal in human societies (Jankowiak 1995).

3.4 Attachment, Oedipus Complex, Socialization

Observations in various cultures suggest that warm, nurturing parenting and paternal presence in early life lay an emotional basis for sociability and positive self-regard. They also suggest that secure early attachment postpones and reduces sexual activity in adolescence and lays an emotional foundation for durable love relationships and warm, empathic parenting. An evolutionary argument relates these findings to reproductive strategy, that is, it is theorized that during the course of human evolution secure attachment in childhood promoted durable passionate attachment and high-quality parenting in adulthood, whereas insecure attachment promoted increased sexual activity and multiple relationships. The first pattern supposedly favored reproductive success where supportive environments resulted in lower death rates, whereas the second pattern was more likely to lead to reproductive success where adverse material conditions resulted in higher mortality (Chisholm 1992). Another formulation finds the Oedipus complex in these scenarios. In this argument, secure attachment results in early onset and resolution of oedipal attachments and, eventually, in an emotional capacity for enduring passionate attachment, whereas insecure attachment promotes more persistent and sexualized forms of oedipal attachment and, consequently, promiscuity and other forms of sexualizing and objectifying sexual partners. Accordingly, the secure form of the Oedipus complex supported the transition to pair-bonding and increased paternal investment in offspring during early human evolution (Ingham 1999).

Attachment is probably everywhere part of early child development, but this need not imply that secure attachment is everywhere produced in the same way. On the contrary, evidence suggests that diverse parenting styles and family arrangements can promote secure attachment. Whether it can be added that attachment and the Oedipus complex vary greatly from society to society continues to be a subject for debate, however. Some writers think they do (Spain 1992), while others find the evidence less than persuasive (Spiro 1992). Attachment and oedipality aside, there is no doubt that children are socialized into their cultures. Many studies show that children learn their cultures from an early age, and that this learning informs perception, emotional expression, and moral reasoning as well as social attitudes and worldview. Many such studies contest assumptions about what constitutes 'normal' development. Shweder (1991), for example, argues that theories of moral development may say more about Western concerns with individual rights than anything about the universal course of moral development.

3.5 Self and Emotion

Culturalists and postmodernists argue that the self and emotion, like embodiment and desire, are culturally determined or socially constructed, and, consequently, that they too vary greatly from one culture to another. One often-repeated idea is that the self is more individualistic or 'egocentric' in modern societies and more communal or 'sociocentric' in traditional societies. Implicit in these arguments is an assumption that what people think about themselves and how and what they feel are heavily constrained by words and idioms. It is argued, for example, that Russians are more expressive emotionally than Americans. In support of this contention, it is noted that Russians have a richer store of words and phrases for describing emotion and more linguistically marked emotional bodily expression than Americans (Wierzbicka 1998).

It can be objected, however, that linguistic constructionism risks typifying whole peoples and underestimating psychological complexities. Although self-descriptions may differ somewhat from culture to culture in how much they refer to social ties, studies in various cultures show that individuals are often better described as having both individualistic and communal inclinations, and, moreover, that this is so even in cultures that would seem to be highly communal (see, e.g., McHugh 1989). Research in the US suggests that the common assertion that Americans are unusually individualistic may be doubly simplistic, not only because Americans are communal as well as individualistic but also because their individualism assumes different forms depending on social class and neighborhood.

What is proving more interesting for many researchers than local lexicons is the way in which self-representation and emotional expression are organized through the stories people tell about themselves and to themselves. Research is being done, for instance, on how individuals use self-narratives to reconstruct and negotiate social identities in diasporic communities. Interest in self-narratives is part of a broader trend toward methodologically rigorous discourse-centered ethnography in psychological anthropology. Speech productions are tape-recorded, transcribed with technical notation, and examined for narrative style and speech style. They are then further examined for what they express about the self and culture. Since discourse is inherently dialogical—interlocutors are both speakers and listeners—attention to discourse translates the abstract relation

between self and society, psyche and culture, into a more concrete mutual relation between interlocutors. In studies of personal narratives about religious conversion, Stromberg (1991) shows how conversion makes culture personally meaningful by transforming referential symbols into constitutive symbols. In the conversion process, symbols that refer to the external world become part of self-identity. At the same time, they may become a personal language for managing intrapsychic conflict, say between desire and moral duty. In one of his studies a man feels powerless and insufficiently masculine. Speech style, body posturing, and cultural symbolism imply that this self-organization involves wishes for masculine power, memories of a brutal father, and feelings of having been overpowered by women, especially his mother and motherlike wife. His dreams represent a masculine power as phallic, flying horses, and the mother figure as a vaginalike tornado. Recommitment to Christian faith occurs after the man identifies in his dreams with a potent, athletic Christ (Stromberg 1991).

Another set of issues concerns how universal emotions or common emotional predicaments are organized in particular cultures. One example is grieving, arguably a universal experience. In some cultures mortuary rituals involve multiple burials and reburials or cremations and re-cremations in which the deceased is figuratively and literally deconstructed and reconstructed. Beliefs associated with such practices sometimes link the mother with the deceased, and the rituals themselves may involve eating the flesh of the deceased and giving the flesh of the deceased or its symbolic equivalent to relatives. In a comparative study of these practices, Stephen (1998) draws on Melanie Klein's object-relations psychoanalysis to theorize that many emotional losses resonate with separation from the mother at some level and may, therefore, activate unconscious traces of child rage and guilt and related needs to make reparation.

3.6 Unconscious Meaning and Motivation in Myth and Ritual

Because it seems to speak to symbolic content in this way, efforts to understand the unconscious meaning and motivation of expressive culture draw more on psychoanalysis than any other psychology. Psychoanalytically oriented anthropologists have often argued that relations with supernatural figures recreate or repair imaginary relations with parents (see Obeyesekere 1981). Various studies suggest that puberty rites and adolescent initiation rites may be oriented toward resolving oedipal ties to parents and orienting young people toward marriage. Ancient Greek myths can be read as implying that adolescent transition rites function in this way. Perhaps the exception that proves the rule, the story of Oedipus may be a deviant form of more common myths in which the hero marries and becomes king after passing through an initiatory trial with a monster (see Goux 1993). Other elements in Western culture may have similar socializing functions. In a rigorously close reading of the text and its midrashic commentaries, Paul (1996) shows that mythic patricide and incestuous desire are recurrent though disguised themes in the five books of Moses. The reading suggests the Torah activates unconscious oedipal fantasies and related guilt feelings but that it also represents both paternal prohibitions and a substitute for the maternal object. In effect, the Torah may serve the growing boy or girl in the Hebrew community as a model for normative displacement and transformation of childhood attachments.

Culturalists frequently complain that psychoanalytic interpretations of cultural symbolism are simplistically universalistic and reductionistic. This may be a fair criticism in some cases—there are examples of overly mechanical and insensitive uses of psychoanalytic interpretation in the anthropological literature—but it seems unfair in many others, including those cited here. Psychoanalytic anthropologists are usually careful to consider ethnographic details and conscious personal and cultural meanings. Indeed, they contend that the evidence for unconscious meanings and motives lies precisely in such ethnographic details and conscious meanings.

4. Basic Dilemma, Contemporary Perspectives

As in the past, psychological anthropologists are struggling with an empirical and theoretical bind involving cultural particularism and psychological universalism, the influence of the sociocultural milieu on the individual, on the one hand, and the motivated agency of psychological subjects, on the other. Advocates of cultural psychology and ethnopsychology try to resolve the dilemma on the side of culture. They sometimes say their viewpoint allows for an active self—this is the point of their replacing the language of cultural determinism with that of social and cultural constructionism—but it is often difficult to see how their self is more than a passive product of culture. While their culturalist approach opens productive questions about how individuals are shaped or constrained by their sociocultural milieus, usually it leaves motivation and the perplexing, enigmatic features of those milieus in question. Universalistic psychoanalytic and evolutionary approaches have a reverse set of strengths and weaknesses. They offer theories and hypotheses about motivation but they are less well equipped to deal with social context and cultural influence.

Cognizant of both horns of the dilemma, many psychological anthropologists are exploring ways of conceptualizing the interplay between self and other, personality and culture. On the culturalist/constructionist side of the discussion, the growing interest in discourse-centered ethnography is arguably the most

noteworthy development (Besnier 1994). While it may sidestep motive and volition, it encourages attention to symbolic expression; it takes the culturalist approach beyond reified notions like individualism and sociocentrism to more detailed understanding of the complexities of subjectivity and culture; and, finally, because speech is inherently dialogical, it conduces toward a view in which the subject is not only the locus of cultural influence but also an active, speaking agent in the production and reproduction of culture. On the psychoanalytic side, there is effort to renew the encounter between psychoanalysis and social anthropology (Augé 1999, Heald and Deluz 1994). In this approach, culture mediates social relations and is not reducible to individual psychology but may tap into unconscious fantasy and emotion. In a dialogical version of this approach, individuals and groups influence each other through pragmatic discourse. More particularly, their rhetorical practices exert social and political effects by suppressing, mobilizing, and redirecting common sense, moral reasoning, and unconscious motivation, especially around issues of attachment, sexuality, aggression, and love. Finally, attention is given to the ways in which individuals resist these tactics or accede to them according to their social positions and psychological dispositions (Ingham 1996). In an attempt to integrate discourse theory with culturalist, cognitive, and psychoanalytic approaches, Nuckolls (1996) argues that cultural knowledge everywhere emerges in conversation and that it is always a paradoxical, dialectical mixture of diverse desires, goals, and rationalizations. At the same time, the dissonance in cultural knowledge engenders desires for coherence and synthesis. Unity and agreement remain elusive, however, and when achieved soon give way to further dialogue and disagreement. What is true of cultural knowledge in these respects is evidently true of psychological anthropology, itself an ongoing conversation. However appealing synthesis may be, it seems that psychological anthropology is entrapped in paradox and dialectic no less than its subject matter.

See also: Benedict, Ruth (1887–1948); Boas, Franz (1858–1942); Cultural Psychology; Emotion: History of the Concept; Incest Prohibition, Origin and Evolution of; Malinowski, Bronislaw (1884–1942); Myth in Religion; Psychoanalysis: Overview; Ritual; Sapir, Edward (1884–1939)

Bibliography

Augé M 1999 *The War of Dreams: Exercises in Ethno-fiction* (trans. Heron L). Pluto Press, London

Banks C G 1998 Illness narratives and human development: Changing metaphors about the body in a case of anorexia nervosa. In: Denzin N K (ed.) *Cultural Studies: A Research Annual,* Vol. 3. JAI Press, Stanford, CT

Besnier N 1994 The evidence from discourse. In: Bock P K (ed.) *Handbook of Psychological Anthropology*. Greenwood Press, Westport, CT

Chisholm J S 1992 Putting people in biology: Toward a synthesis of biological and psychological anthropology. In: Schwartz T, White G M, Lutz C A (eds.) *New Directions in Psychological Anthropology*. Cambridge University Press, Cambridge and New York

D'Andrade R G, Strauss C A (eds.) 1992 *Human Motives and Cultural Models*. Cambridge University Press, Cambridge, UK

Goux J-J 1993 *Oedipus, Philosopher* (trans. Porter C). Stanford University Press, Stanford, CA

Heald S, Deluz A (eds.) 1994 *Anthropology and Psychoanalysis: An Encounter Through Culture*. Routledge, London

Ingham J M 1996 *Psychological Anthropology Reconsidered*. Cambridge University Press, Cambridge, UK

Ingham J M 1999 The Oedipus complex and the signifying chain: A view from anthropology. *Journal for the Psychoanalysis of Culture & Society* **4**: 50–62

Jankowiak W (ed.) 1995 *Romantic Passion: A Universal Experience?* Columbia University Press, New York

Johnson A, Price-Williams D 1996 *Oedipus Ubiquitous: The Family Complex in World Folk Literature*. Stanford University Press, Stanford, CA

Obeyesekere G 1981 *Medusa's Hair: An Essay on Personal Symbols and Religious Experience*. University of Chicago Press, Chicago

McHugh E L 1989 Concepts of the person among the Gurungs of Nepal. *American Ethnologist* **16**: 75–86

Nuckolls C W 1996 *The Cultural Dialectics of Knowledge and Desire*. University of Wisconsin Press, Madison, WI

Paul R A 1996 *Moses and Civilization: The Meaning Behind Freud's Myth*. Yale University Press, New Haven, CT

Shweder R A 1991 *Thinking Through Cultures: Expeditions in Cultural Psychology*. Harvard University Press, Cambridge, MA

Spain D H 1987 The Westermarck-Freud incest-theory debate: An evaluation and reformulation. *Current Anthropology* **28**: 623–45

Spain D H 1992 Oedipus Rex or edifice wrecked? Some comments on the universality of oedipality and on the cultural limitations of Freud's thought. In: Spain D H (ed.) *Psychoanalytic Anthropology After Freud: Essays Marking the Fiftieth Anniversary of Freud's Death*. Psyche Press, New York

Spiro M E 1992 Oedipus redux. *Ethos* **20**: 358–76

Stephen M 1998 Devouring the mother: A Kleinian perspective on necrophagia and corpse abuse in mortuary ritual. *Ethos* **26**: 387–409

Stocking Jr G W 1986 Anthropology and the science of the irrational: Malinowski's encounter with Freudian psychoanalysis. In: Stocking Jr G W (ed.) *Malinowski, Rivers, Benedict and Others: Essays on Culture and Personality*. University of Wisconsin, Madison, WI

Stocking Jr G W 1992 Polarity and plurality: Franz Boas as psychological anthropologist. In: Schwartz T, White G M, Lutz C A (eds.) *New Directions in Psychological Anthropology*. Cambridge University Press, Cambridge, UK

Stromberg P G 1991 Symbols into experience: A case study in the generation of commitment. *Ethos* **19**: 102–26

Wierzbicka A 1998 Russian emotional expression. *Ethos* **26**: 456–83

Wolf A P 1995 *Sexual Attraction and Childhood Association: A Chinese Brief for Edward Westermarck*. Stanford University Press, Stanford, CA

J. M. Ingham

Psychological Climate in the Work Setting

Psychological climate is defined as 'employees' perceptions of their work environment' and has been the focus of considerable research in organizational psychology. Measures of psychological climate are intended to assess work environments as they are perceived and interpreted by employees (James et al. 1990). Examples of psychological climate dimensions include, but are not limited to, role clarity, job importance, leader support, and workgroup cooperation.

Research has linked psychological climate variables to work-related outcomes such as job satisfaction, job involvement, and job performance. Furthermore, research has attempted to determine whether there is a hierarchical model to psychological climate. That is, do the individual climate variables all represent assessments of the degree to which organizational events can contribute to desired ends (as described by one's personal values) and reflect an underlying higher order construct?

The next sections will examine the development of psychological climate research from its conception to its current state of affairs and probable future research directions.

1. *The Origins of Psychological Climate Research*

Beginning with Lewin et al.'s (1939) discussion of 'social climates' in the workplace, assessing employee climate perceptions became increasingly interesting to organizational researchers. However, the idea of psychological climate, as it is studied now, was not first comprehensively discussed until the late 1960s. Litwin and Stringer (1968) presented a model of the determinants of motivated behavior in organizations that included the concept of organizational climate. It should be pointed out that in the early stages of psychological climate research it was referred to as organizational climate, though that term is now used to describe shared perceptions of organizational policies and practices, rather than individual employees' perceptions of characteristics of the workplace. According to Litwin and Stringer (1968), the organization is seen as generating an organizational climate, which in turn either positively or negatively affects particular motivational patterns of employees. They theorized that climate could provide managers with the link between an organization's procedures/practices and the concerns and needs of individual employees. This link would help managers determine what practices would stimulate employee motivation. Litwin and Stringer (1968) defined several important dimensions of organizational climate, including structure (perception of formality and constraint in the organization), challenge (perception of challenge, demand for work, and opportunity for sense of achievement), reward and support (emphasis on positive reinforcement rather than punishment), and social inclusion (sociability, belonging, and group membership). Finally, Litwin and Stringer (1968) tested their hypotheses in an experiment that simulated eight days of actual organizational life in three companies with very different organizational climates. The organizations differed in terms of the amount of structure (strict procedures vs. more flexible procedures), responsibility given to employees (e.g., low vs. high), reward and punishment (e.g., only punishment vs. a combination of both punishment and rewards), warmth and support (isolated employees vs. friendly workgroups), cooperation and conflict (e.g., cooperation is praised or not), and risk and involvement (e.g., organization stressing conservatism vs. stressing challenge). Litwin and Stringer's (1968) results indicated that type of organizational climate was related to employee satisfaction and employee performance. Overall, they found that what they called an achieving climate (e.g., high employee responsibility, cooperation among employees is stressed, tolerate conflict and support risk taking) led to the best outcomes for the employees (i.e., high satisfaction) and the organization (i.e., high performance).

Litwin and Stringer's (1968) work provided the impetus for increased research on psychological climate in work settings. This subsequent research identified several problems that confounded much of the original psychological climate research. The most important of these was a 'level of analysis' problem.

1.1 *Distinction between Psychological Climate and Organizational Climate*

James and Jones (1974) conducted a review of the climate literature and commented that, while organizational climate research had received a considerable amount of attention, the 'conceptual and operational definition, measurement techniques, and ensuing results' (James and Jones 1974 p. 1096) were highly diverse and contradictory. These problems were causing organizational climate to be viewed as a somewhat nebulous construct. James and Jones (1974) argued that previous climate research could be organized into three different, but not mutually exclusive, approaches to defining and measuring organizational climate. They designated these approaches as: (a) the multiple-measurement-organizational attribute approach,

which views climate as a set of organizational attributes measured by a variety of methods; (b) the perceptual measurement-organizational attribute approach, which uses only perceptual tools to measure climate but still sees climate as an organizational level variable; and (c) the perceptual measurement-individual attribute approach, which views climate as perceptual and an individual level phenomenon. James and Jones (1974) believed that researchers needed to differentiate between climate measured at the organizational level and at the individual level. They proposed that psychological climate was an individual attribute, whereas organizational climate was a situational attribute of the organization. Furthermore, by establishing these new definitions, researchers would be able to focus on either psychological climate or organizational climate which would, in turn, lead to research that would better be able to define both constructs and how they both related to other variables.

Later studies helped shape the boundaries between psychological climate and organizational climate (Schneider 1975). Psychological climate was defined as individual descriptions of organizational practices and procedures, which are useful in understanding the influence of the organizational environment on individual performance and satisfaction. Organizational climate, on the other hand, refers to a collective description of the organizational environment, and is thought to relate to organizational level outcomes. This distinction allowed psychological climate researchers to focus squarely on individual level issues, such as the relationship between psychological climate and various outcome variables (e.g., individual job performance). However, as the definition of psychological climate became somewhat solidified, researchers began to debate whether there was a distinction between psychological climate and job satisfaction. This debate was settled when researchers reached the consensus, through theoretical arguments (James and Jones 1974) as well as empirical research (Schneider and Snyder 1975), that psychological climate reflected descriptions of the work environment whereas job satisfaction reflected an affective evaluation of the work environment.

1.2 Hierarchical Model of Psychological Climate

As the definition of psychological climate became generally accepted, researchers began to expand the list of perceptual variables they used to measure individuals' descriptions of their work environment. However, research on the measurement properties of psychological climate also began to indicate that these psychological climate variables often clustered within four higher order factors: role, job, leadership, and workgroup (James and James 1992). While the exact number and specific names of psychological climate variables clustered under these higher order factors changes depending on the scale used, the following examples should provide a general illustration of these higher order factors. The 'role' higher order factor is often made up of role ambiguity, role conflict, and role overload. The 'job' higher order factor often comprises job challenge, job autonomy, and job importance. The 'leader' higher order factor is most often made up of leader trust and support, leader goal facilitation, and leader hierarchical influence. Finally, the workgroup higher order factor often comprises work group cooperation, work group friendliness, and workgroup pride. Furthermore, this early research viewed these four psychological climate factors as 'distinct cognitive organizing principles for perceptual variables' (James and James 1989 p. 93).

However, James and James (1989) argued that there is a further hierarchical relationship among psychological climate dimensions. Specifically, they suggested that the judgment process of appraising organizational events in relation to one's personal values is common to each climate dimension. These value-based appraisals represent assessments of the degree to which organizational events can contribute to desired personal ends. Moreover, these judgments were hypothesized to reflect an underlying general factor (PCg), which represents an overall evaluation of whether the work environment is personally beneficial or detrimental to one's well-being. Research by James and James (1989) as well as others (e.g., Brown and Leigh 1996) has shown support for the idea of PCg. The PCg model, because it presents climate as one overarching construct, subsequently made examining the relationship between psychological climate and other variables much simpler.

2. Current Research

Current research falls into two categories: research that focuses on the relationship between psychological climate variables and other constructs (e.g., job performance); and research that focuses on the hierarchical structure of psychological climate.

2.1 Psychological Climate and Work-related Outcome Variables

Research has demonstrated that psychological climate perceptions provide an important mediating link between organizational characteristics and consequential work outcomes such as employee attitudes, motivation, psychological well-being, and performance. A recent meta-analysis (Parker et al. 1998) examined the relationships between individual-level higher order climate perceptions (e.g., role, job,

workgroup, leader) and work outcomes such as employee work attitudes, psychological well-being, motivation, and performance. It was found that individuals' climate perceptions had significant relationships with their job satisfaction, work attitudes, psychological well-being, motivation, and performance. Furthermore, each of the climate categories exhibited a somewhat different pattern of relationships with the various outcomes. For example, employees' job (e.g., importance, autonomy) and role (e.g., overload, clarity) perceptions appeared to have the weakest relationships with all of the outcome variables. Interestingly, the leader (e.g., goal emphasis, support) and work group (e.g., cooperation, warmth) constructs were most predictive of employees' work attitudes. A similar pattern was evident for the relationships of the various psychological climate dimensions with employee motivation and performance. For psychological well-being, however, job and leader perceptions provided the strongest relationships. Overall, it appears that employees' climate perceptions have stronger relationships with their work attitudes than with their motivation and performance. This pattern of relationships indicates that the effects of climate on motivation and performance may be mediated by employees' work attitudes. This hypothesis was tested in the study and supported. Thus, the relationship between employees' motivational and behavioral reactions to perceptions of their work environment and their motivation and performance does seem to be mediated by their overall evaluations of these perceptions.

Most studies that have considered the relationship between psychological climate variables and work-related outcome variables have assumed a linear relationship. Current research is investigating whether this is the case.

2.2 Nonlinear Relationships Between Psychological Climate and Outcome Variables

Researchers have debated over whether the relationships between certain climate dimensions (e.g., job autonomy) and outcome variables (e.g., job satisfaction) are linear or nonlinear in nature. For example, job autonomy is assumed by some researchers to follow an inverted U shape. That is, while medium levels of job autonomy are thought to be more desirable than low levels, very high levels of job autonomy are potentially harmful since it implies difficulty in decision making and high responsibility on the job. Other researchers, however, have assumed a linear relationship between job autonomy and job satisfaction: the more autonomy workers experience, the more satisfied they are with the job.

Previous studies examining psychological climate variables have shown support for both the linear and nonlinear points of view. However, the nonlinear relationships found in prior research have been somewhat weak, often explaining only a small amount of additional variance (1–2 percent) over the linear model. In summary, the evidence is inconclusive as to which type of relationship exists.

However, research examining methodological problems associated with using a common response format may provide an explanation for these weak curvilinear findings. The use of a common response format has been shown to introduce shared item variance (i.e., common methods variance, or CMV) and elevate relationships among job perceptions (Glick et al. 1986). However, one can measure constructs using a multimethod approach. By measuring constructs with more than a single method one should be able to address the common response format problem mentioned above. A recent study (Baltes et al. 1999b) used both multitrait-multimethod (MTMM) data and monomethod data to test the relationships between climate dimensions that have been hypothesized to have nonlinear relationships with job satisfaction. Strikingly different results were found between the monomethod and MTMM data. Whereas the monomethod results mirrored earlier inconclusive findings, the MTMM data indicated that nonlinear equations explained significantly more of the relationship between all three climate dimensions and job satisfaction. These results suggest the following. First, the use of MTMM data allows one to test for nonlinear effects more effectively. Second, nonlinear quadratic effects better model the relationship between role clarity, role overload, job autonomy, and the outcome variable job satisfaction. These results, of course, have implications for both measurement and interpretation of psychological climate variables and work-related outcomes.

2.3 Hierarchical Structure of Psychological Climate

As mentioned earlier, recent climate research has used the James and James PCg model to justify collapsing multiple indicators of psychological climate into one overarching appraisal of the work environment (Brown and Leigh 1996, James and James 1992). However, it may be premature for psychological climate researchers to accept the construct validity of PCg because the James and James (1989) study relied on a single response format (i.e., traditional Likert scale). As mentioned earlier, the use of a common response format (i.e., monomethod data) has been shown to introduce shared item variance, which can elevate relationships among constructs. As a result, the intercorrelations among climate dimensions that generate support for PCg may be the result of nonrandom measurement errors caused by common methods variance. The hypothesis that support for

PCg is the result of common methods variance has received some support (Parker 1999). However, research efforts that use multiple measures of psychological climate are required to more appropriately test whether PCg is the result of common methods variance (James and James 1989). A study conducted by Baltes et al. (1999b) used multitrait–multimethod data collection to examine the hierarchical model of psychological climate (PCg). Two different self-report response formats were used to measure 12 psychological climate constructs. The PCg model, when tested with multitrait–multimethod data, did not fit the data. The lack of support for the PCg model appears to be caused by a strong common methods bias, which caused a 35 percent bias (i.e., inflation) in the observed correlations among the constructs. These results do not support the practice of collapsing psychological climate dimensions into an overarching construct. Furthermore, support was also not found for collapsing to the aforementioned four higher order constructs: role, job, leader, and workgroup. This result is perhaps the most troublesome given the extent of prior research that has considered the relationship between these higher order constructs and work-related outcome variables. These results are the product, of course, of one study and need to be replicated before previous research linking higher order psychological climate factors to outcome variables is considered obsolete.

3. *Future Research*

The results of current and past research will probably lead to three streams of future research. First, with respect to the relationship between psychological climate and various work outcomes, future research needs to test more complicated mediation and/or moderation models that include other work-related variables, such as employee' work attitudes and/or other individual difference variables. By testing more complicated models researchers will be better able to discern how psychological climate fits into the workplace environment.

Second, if the relationships between these climate variables and job satisfaction are nonlinear then perhaps the important substantive question to ask employees is not the traditional one of how much they have of, for example, role clarity, but how what they have differs from what they would like. That is, the discrepancy between what they would like and what they have becomes the important question to ask since this discrepancy would indicate where they fall along the nonlinear curve. Knowing where an employee falls on the regression curve would allow one to predict what effect more or less of a climate dimension would have. Just knowing how much an employee has of, for example, job autonomy, would not allow a researcher to effectively hypothesize what effect more or less job autonomy would have on the job satisfaction of that employee. Future research should not just ask the traditional question of how much an employee has of certain climate dimensions, but focus on the discrepancy between what they have and what they desire to better understand and predict outcome variables such as job satisfaction.

Finally, future research should assess psychological climate dimensions with other methods, such as essays or interviews, and test the PCg model. Whereas some current research, as mentioned earlier, has called into question the existence of a higher order psychological climate factor, more research with multiple measurement methods needs to be completed to address this issue more adequately.

See also: Organizational Climate

Bibliography

Baltes B B, Bauer C, Altmann R 1999a Psychological climate and job satisfaction: Multitrait-multimethod evidence for nonlinear relationships. Paper presented at the 107th annual meeting of the American Psychological Association, Boston, MA, August 20–7

Baltes B B, Parker C P, LaCost H, Altmann R, Huff J, Young S 1999b A Multitrait-Multimethod (MTMM) examination of hierarchical models of psychological climate. Poster presented at annual meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA, 14th Annual Meeting, April 24–May 2

Brown S P, Leigh T W 1996 A new look at psychological climate and its relationship to job involvement, effort, and performance. *Journal of Applied Psychology* **81**: 358–68

Glick W H, Jenkins G D, Gupta N 1986 Method versus substance: How strong are underlying relationships between job characteristics and attitudinal outcomes? *Academy of Management Journal* **29**: 441–64

James L R 1982 Aggregation bias estimates of perceptual agreement. *Journal of Applied Psychology* **67**: 219–29

James L R, James L A 1989 Integrating work environment perceptions: Explorations into the measurement of meaning. *Journal of Applied Psychology* **74**: 739–51

James L R, James L A 1992 Psychological climate and affect: Test of a hierarchical dynamic model. In: Cranny C J, Smith C, Stone E F (eds.) *Job Satisfaction: How People Feel About Their Jobs and How it Affects Their Performance*. Lexington Books, New York, pp. 89–117

James L R, James L A, Ashe D K 1990 The meaning of organizations: The role of cognition and values. In: Schneider B (ed.) *Organizational Climate and Culture*. Jossey-Bass, San Francisco, CA, pp. 40–89

James L R, Jones A P 1974 Organizational climate: A review of theory and research. *Psychological Bulletin* **81**: 1096–112

Lewin K, Lippitt R, White R K 1939 Patterns of aggressive behavior in experimentally created 'social climates.' *Journal of Social Psychology* **10**: 271–99

Litwin G H, Stringer R A 1968 *Motivation and Organizational Climate*. Harvard University Press, Cambridge, MA

Parker C F 1999 A test of alternative hierarchical models of

psychological climate: PCg, satisfaction, or common method variance? *Organizational Research Methods* **2**: 257–74
Parker C F, Young S, Baltes B B, Altmann R, Huff J, Lacost H 1998 A meta-analysis of relationships between climate perceptions and work outcomes. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX, 13th Annual Conference, April 23–26
Schneider B 1975 Organizational climates: An essay. *Personnel Psychology* **28**: 447–79
Schneider B, Snyder R A 1975 Some relationships between job satisfaction and organizational climate. *Journal of Applied Psychology* **60**: 318–28

B. B. Baltes

Psychological Development: Ethological and Evolutionary Approaches

1. *Introduction: The Science of Animal Motives*

Classical ethology, the 'biology of animal behavior' (Eibl-Eibesfeldt 1970, Hinde 1982), compares descriptions of how different species act in their natural environment, and in families and societies. Lorenz proposed that genetically determined motor programs or Innate Release Mechanisms (IRMs), triggered by environmental Sign Stimuli, confer 'dispositions to learn' that benefit the species. By comparing spontaneous behavior rituals and sequences of display or response between individuals, ethologists seek evidence of innate impulses in different species, and determine the genetic relationships and evolutionary lineage (see *Comparative Method, in Evolutionary Studies*). Ethology developed from zoology and field studies of wild animals. It accepts that animals have motives and awareness that can be understood by accurate description of their behaviors in nature, and how these develop. Darwin's *Expression of the Emotions in Man and Animals* (1872) inaugurated ethological science (see *Darwin, Charles Robert (1809–82)*). He related human feelings to those one can infer in animals from their expressions.

Tinbergen (1951) identified innate signaling behaviors of communication—in mating, nest building, care of young, territorial defence, hunting or escape from predators, sharing environmental resources, etc. (see *Evolutionary Social Psychology*). He defined four time-scales of behavioral determination: over evolutionary time; in development of the individual; in control of behavioral tendencies over hours, days, or months; and in adaptive functional response to immediate environmental events. He used field experiments to identify that stimuli or behavior signals are 'releasers' of innate activities.

Scientists research animal senses, often very different from ours, and the ways animals move to investigate their environment. Communicative signals are instructive, because they have evolved to be perceived. Rhythmic light flashes, visible movements of body parts, vocal or other sounds, or vibrations, convey effective messages, even in primitive species. How animals react to signal patterns gives further evidence about internal processes that guide attention, movements, and emotions. In humans, innate hearing and vision has been explored by studying the vocal sounds, face expressions, and body movements that infants prefer, and the emotions infants express (Fernald 1992, Papousek and Papousek 1987, Stern 2000, Trehub 1990, Trevarthen 1998, 1999).

As in Darwin's day, evidence for species-specific motives and feelings in humans comes from comparing spontaneous expressions of people in different cultures, and by observing infants before they have had the chance to learn (Eibl-Eibesfeldt 1989). Newborn infants communicate and imitate, and children show systematic changes in behaviors and interests (Stern 2000, Trevarthen 1998). As children's motives develop, they change the behavior of caregivers, stimulating 'intuitive parenting' and the sharing of meanings (Papousek and Papousek 1987).

2. *Learning, Cognition, and Communication*

Efforts have been made to relate ethological theory to concepts of experimental psychology and 'conditioning' in simplified laboratory situations (Hinde 1982). After decades of rejection by psychologists of 'mentalistic' interpretations, the idea that animals may share motives, feelings, and conscious experiences comparable to those of human beings has again become scientifically acceptable (see *Human Cognition, Evolution of*; *Evolution of Cognition: An Adaptationist Perspective*). Psychologists postulate internal 'representations' that process perceptual information and control actions of subjects in complex circumstances. More recently, attention has turned to the evolved functions of emotional regulatory systems in the brain, and their role in self-awareness and intelligence (Damasio 1994) (see *Emotions, Evolution of*).

Cognitive psychology acknowledges that human behavior is regulated by conscious states and intentions, and that learning integrates emotions and mental representations. Nevertheless, cognitive science remains individualistic and rational—focused on reality. Complex intelligence is analyzed in terms of perceptual categorizations and hierarchical representational systems. Different modes of intelligence are thought to be formulated anatomically in identifiable brain regions. In particular, social intelligence is considered to pass, in evolution and development, through grades of cognitive representation, and levels of awareness of cognition in other individuals—from immediate, 'thoughtless' sensorimotor reactions, to verbalizable, rationally formulated 'theories of mind'

(see *Evolution of Cognition: An Adaptationist Perspective*). Dynamic Systems Theory, opposed to linear causal explanation, attributes behavior patterns and mental representations to 'emergent properties' thrown up in complex systems comprising large numbers of elements—neural connections, perceptual feature detectors, sensorimotor loops, motoric forces, representations in memory, social transactions, etc. In neither of the dominant contemporary theories of complex mental life and behavioral coordination is evolved anticipatory motivation given a determinant role.

No adaptive behavior of an animal subject is without innate 'constraints' due to prefunctional configuration (morphogenesis) of sensory and motor structures, brain tissue, and nerve connections, which map both the body and its future engagement with the outside world in a coherent, subjective space and time. What is innate is an active neurobiological initiative that investigates the environment, applying specific anticipatory, probing processes of behavior. 'Critical periods,' or ages in which active investigations of experience change, correspond to brain growth periods that are generally more open to learning (see *Lifespan Development: Evolutionary Perspectives*).

3. *Behavior from Dyads to Large Social Groups*

Subjective motives communicated to other individuals become intersubjective (Bråten 1998). In the more intelligent, behaviorally complex species there is a rich capacity for display of mental impulses—of attention, intention, and emotion. Intersubjective relations assume central importance in reproduction, feeding, establishment of cooperative groups in families or much larger societies. Interspecific collaborations, including predator–prey interactions, control behavior in dyads and larger groups, regulating ecological relations between animals psychologically.

Sociobiology has formulated theories to explain the evolutionary advantages of various forms of conflict and cooperation within and between communities, and special complementary adaptations of individuals in the service of the group, or to propagate a genetic type. Reductive cost-benefit analyses, of energy balance sheets for motor and metabolic investment, or of transmission and survival of genes in populations (reproductive success), have prevailed, which can be explained by an unwillingness of quantitative biologists to attribute self-regulatory motives and emotions to animals as individuals, or members of families and cooperative groups (see *Evolution, Natural and Social: Philosophical Aspects*). They see behaving animals as driven by forces of competition for future success that are indifferent to any aspirations for, or pleasures in, life in the present. The genetic history of humans indicates that there has been rapid adaptation to an extremely large range of environments, and new adaptability by cooperative learning of behavioral traditions (Cavalli-Sforza and Cavalli-Sforza 1995) (see *Genes and Culture, Coevolution of*; *Cultural Evolution: Overview*).

The benefits of competition have been exaggerated. Research on the psychopathology of infanticide or mass killing by individuals indicates that direct competition by assault results from a loss of pleasure in the kinds of altruistic and affectionate motives of play that lead to creative invention of new and beneficial motives, ideas, or activities in a group (Bekoff and Byers 1998). Communication between members of a species and between differently adapted species without concern for gain leads to intricate dynamic ecological systems and mutual dependency. Some behaviors of predator and prey in natural ecological equilibrium are cooperative, or mutually supportive. Skilled collective understanding of the environment by intelligent social species depends on positive motives for sharing enjoyment of experience and skilled action, such as are abundantly shown by the young of many species when they play together.

4. *Attachment, Companionship, and Parental Support of Early Development*

Emotions linked to innate behaviors regulate parent–offspring attachments and caregiving (see *Attachment Theory: Psychological*). In many species, mutual adaptations of motives in parents and young facilitate cooperative encounters that nurture the young, and parenting strategies limit the offspring's dependency. The food, care, and support that nestlings or nurslings receive sustain the developing body and brain. Human infants show attachment to the physiological or bodily benefits from mothering (Schore 1994), the biological importance of which was first demonstrated experimentally in monkeys. Infants also respond to psychological expressions of pleasure and interest carried by the voice, facial expressions, and gestures of the parent, and these lead to 'protoconversations' and companionship (Trevarthen 1998, 2001b; Trevarthen and Aitken 2000). Infants and parents seek contact and affective communication, and deprivation of sympathetic parenting has detrimental effects on both emotional and cognitive development of a child.

Pleasure and pain may be the main emotions or neurochemical systems seeking parental nurturance in mammals, including humans (Panksepp 1998), but more complex emotions, including self–other conscious states and relational emotions such as pride and shame, regulate intersubjective motives and assist the learning of a wide range of socially and culturally important skills in infants (Stern 2000, Draghi-Lorenz

et al. 2001) (see *Self-conscious Emotions, Psychology of*). Infants seek emotional confirmation to develop knowledge and skills (Trevarthen and Aitken 2000).

5. *Motives and Benefits of Play*

Animal motives are revealed not only in behaviors that immediately benefit the individual, or the propagation and flourishing of the group or species. Their essential creative aspect is vivid in play (Bekoff and Byers 1998). Ethological and psychological theories of play appear confused by a desire to find a function outside motivation itself. However, by its nature, and by any definition, play behavior largely is determined by the anticipatory processes of behavioral control, and not any immediate adaptive consequence. Play behaviors are more apparent in the more social species that have elaborate intersubjective capacities or social awareness. By giving strong and continuous messages of exuberant motives, play invites sympathetic attention and, indeed, participation by others in changing outcomes and risks. Human children, like young of other species, can play alone, practising and shaping perceptuo-motor skills. But behaviors exhibited in self-satisfying play have a potential to be social, and they are responsive to provocative or 'teasing' interactions with other individuals. Play fighting between young in a cooperative family group is affiliative, consolidating relations and training awareness of mutual intentions. Many play behaviors anticipate functions of mature 'innate' behavior patterns. Thus, play expresses motives for behaviors that require exercise of motor and cognitive processes for their maturation. It would seem pointless to attempt a cognitive rationalization of play if play is the prior behavior, adapted to generate intelligence and regulate cognition.

Research with infants confirms that play develops relationships and skills, as well as creating rituals and 'formats' of cooperation. The imaginative or 'symbolic' play of young children foreshadows the creative efforts made by young adult athletes, and the striving of outstanding craftsmen, artists and scientists throughout their lives (Bekoff and Byers 1998). Art, appreciation of beauty in artefacts and ceremony in human communities may be related to universal rhythms and modes of intimate communication between infants and their partners (Dissanayake 2000).

6. *'Communicative Musicality' and the Pulse of Human Thought and Imagination*

Animal communication employs precisely timed rhythms, modulations of expressive activity, and accurate interaction of movements in communicating individuals (Tinbergen 1951). Infants' expressions, and their refined discrimination of temporal, prosodic, harmonic, and melodic features of vocalization or musical sounds, beginning before birth, indicate that human language communication draws on rich preverbal talents for expressive gesture and perception (Trehub 1990). Human sensitivity for the emotive rhythms, qualities, and narrative sequences of expressive song and dance, may be evolved with bipedal standing and walking, which free the erect body, and especially the hands and voice, for flexible polyrhythmic gestures that can richly express dynamic emotions and their blends and changes in interpersonal engagements (Trevarthen 1999). Emotions are conveyed in regulated 'contours' or 'melodies' of muscle activity. They are held together by a coherent time-generating process in the brain that creates a hierarchy of acts and succession of perceptual events in one 'mind time' (Pöppel 1994). Infants discriminate the pulse, expressive quality, narrative succession, and contingency of response in adult expressions.

7. *The Biology of Human Imitation, Cultural Learning, and Language*

Monkeys and apes learn in social play, and imitate, but parental support for learning is limited in subhumans, and the most elaborate monkey play is between peers. Apes play over a wider range of ages, and their play assists development of family and group bonds, and they imitate how adults forage and use transformed objects as 'tools' for getting food. No apes show the intricate intersubjective exchanges of human vocal and gestural play with infants soon after birth, or the constructive and narrative-making fantasy games of toddlers. Donald (1991) argues from comparative cognitive psychology and the fossil record that it is likely that the first bipedal hominids were communicating nonverbally, by gestural mimetic narratives, about nonpresent experiences (see *Cultural Evolution: Theory and Models*). He concludes that grammar of language evolved out of the rhythmic syntax of gestural mimesis (see *Evolution and Language: Overview*). Study of developments towards language of infants and toddlers supports this theory (Locke 1993).

Human beings have unique motives for cultural learning—the acquisition of artificial knowledge and skills. These govern progress from preverbal forms of communication and play (Trevarthen 1998). Human ethology is necessarily a science concerned with these motives. All cultural cooperation among humans, however artificial, depends on a uniquely productive capacity for 'mirroring' feelings, interest, and purposes of other persons, and for solving cognitive, practical, moral, and artistic problems collaboratively (Bruner 1996). Studies of imitation reveal that a neonate human has a remarkably complete image of move-

ments in the expressive parts of another person's body, an image that is both receptive through the constellation of all senses, and capable of forming a matching movement in communicative interactions (Bråten 1998, Nadel and Butterworth 1999).

Human language learning depends upon joint attention and sympathetic understanding of intentions (Tomasello and Farrar 1986). Toddlers acquire use of words by following convergent interest to grasp meanings in other persons' speech (Locke 1993). Long before verbal communication, an infant is proficient in 'protoconversational' games, and in participation in rhythmic rituals of action games and songs, which parents are highly motivated to share, intuitively matching and complementing the emotions expressed by the infant (Trevarthen 1998).

8. *The Evolutionary Neuroembryology of Motives and Emotions*

Discoveries in neurobiology enhance our knowledge of adaptive motive systems and emotions and their role in self awareness and communication (Schore 1994). Brain mechanisms of emotion in brainstem regions influence the higher functions of imagination, thought, and language (Panksepp 1998) and exercise an influence over the maturation of cortical cognitive systems in both prenatal and postnatal development of mammals, including human beings (Trevarthen 2001a). Cortical mechanisms for perceiving postural, facial, and vocal expressions, as well as manual gestures, for remembering experiences associated with different expressed emotions, and for generating such expressions as elaborate narratives of internal impulses and feelings, are enlarged greatly in humans as compared to homologous brain areas of other species (Damasio 1994, Trevarthen 2001a). Play is associated with manifestations of neurochemical well-being or 'happiness' (Panksepp 1998). Often young animals play vigorously with body movements at a particular age, and it has been noted that in this 'sensitive period' the cerebellum of the hind brain is undergoing accelerated growth and acquiring refinements of circuitry necessary for accurate and rapid motor control.

9. *Future Directions*

Only recently has over two centuries of research on the comparative anatomy of brains and their development been comprehensively integrated with psychological theories of cognition and learning. A focus on the forebrain cortex, and the effects of cortical lesions on human awareness, memory, and motor skills had left the core systems that incorporate motivational and emotional controls out of the picture. Recognition that acquired skills and knowledge are directed, focused, selected, and integrated by subcortical systems leads to questions about how adaptations of behavior and awareness that anticipate experience and shape learning are generated. Ethological theory directly supports new enterprises in human genetics, neuroethology, and developmental neuroscience, and applied fields, such as education and psychiatry. We anticipate a more balanced natural science of mental life integrating the processing of empirical experience and reason with inherited emotive impulses for both individual experience and cooperative social life. Important questions will arise concerning what determines human affairs and our relation to the natural world.

See also: Animal Cognition; Behavioral Genetics: Psychological Perspectives; Comparative Psychology; Primate Behavior: Significance for Understanding Humans; Psychological Development: Ethological and Evolutionary Approaches; Sensitive Periods in Development, Neural Basis of

Bibliography

Bekoff M, Byers J A 1998 *Animal Play: Evolutionary, Comparative and Ecological Approaches*. Cambridge University Press, New York

Bråten S (ed.) 1998 *Intersubjective Communication and Emotion in Early Ontogeny*. Cambridge University Press, Cambridge, UK

Bruner J S 1996 *The Culture of Education*. Harvard University Press, Cambridge, MA

Cavalli-Sforza L L, Cavalli-Sforza F 1995 *The Great Human Diasporas: The History of Diversity and Evolution*. Addison-Wesley, Reading, NY

Damasio A R 1994 *Descartes Error: Emotion, Reason and the Human Brain*. Putnam, New York

Dissanayake E 2000 *Art and Intimacy: How the Arts Began*. University of Washington Press, Seattle and London

Donald M 1991 *Origins of the Modern Mind*. Harvard University Press, Cambridge, MA

Draghi-Lorenz R, Reddy V, Costall A 2001 Rethinking the development of 'non-basic' emotions: A critical review of existing theories. *Developmental Review* (in press)

Eibl-Eibesfeldt I 1970 *Ethology: The Biology of Behavior*. Holt, Rinehart and Winston, New York

Eibl-Eibesfeldt I 1989 *Human Ethology*. Aldine de Gruyter, New York

Fernald A 1992 Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In: Barkow J H, Cosmides L, Tooby J (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, Oxford, UK

Hinde R A 1982 *Ethology: Its Nature and Relation with Other Sciences*. Fontana, London

Locke J L 1993 *The Child's Path to Spoken Language*. Harvard University Press, Cambridge, MA and London

Nadel J, Butterworth G (eds.) 1999 *Imitation in Infancy*. Cambridge University Press, Cambridge, UK

Panksepp J 1998 *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, New York
Papousek H, Papousek M 1987 Intuitive parenting: A dialectic counterpart to the infant's integrative competence. In: Osofsky J D (ed.) *Handbook of Infant Development*, 2nd edn. Wiley, New York
Pöppel E 1994 Temporal mechanisms in perception. *International Review of Neurobiology* **37**: 185–202
Schore A N 1994 *Affect Regulation and the Origin of the Self: The Neurobiology of Emotional Development*. Erlbaum, Hillsdale, NJ
Stern D N 2000 *The Interpersonal World of the Infant: A View from Psychoanalysis and Development Psychology*, 2nd edn. with new Intro. Basic Books, New York
Tinbergen N 1951 *The Study of Instinct*. Clarendon Press, Oxford, UK
Tomasello M, Farrar M J 1986 Joint attention and early language. *Child Development* **57**: 1454–63
Trehub S E 1990 The perception of musical patterns by human infants: The provision of similar patterns by their parents. In: Berkley M A, Stebbins W C (eds.) *Comparative Perception: Vol. 1, Mechanisms*. Wiley, New York
Trevarthen C 1998 The concept and foundations of infant intersubjectivity. In: Bråten S (ed.) *Intersubjective Communication and Emotion in Early Ontogeny*. Cambridge University Press, Cambridge, UK
Trevarthen C 1999 Musicality and the intrinsic motive pulse: Evidence from human psychobiology and infant communication. *Musicae Scientiae*, Special Issue (1999–2000): 157–213
Trevarthen C 2001a The neurobiology of early communication: Intersubjective regulations in human brain development. In: Kalverboer A F, Gramsbergen A (eds.) *Handbook on Brain and Behavior in Human Development*. Kluwer Academic Publishers, Dordrecht, The Netherlands
Trevarthen C 2001b Intrinsic motives for companionship in understanding: Their origin, development, and significance for infant mental health. *Infant Mental Health Journal* **22**(1–2): 95–131
Trevarthen C, Aitken K 2000 Intersubjective foundations in human psychological development. *Journal of Child Psychology and Psychiatry* **42**(1): 3–48

C. Trevarthen

Psychological Therapies: Emotional Processing

It is now well established that systematic exposure to phobic stimuli markedly reduces associated fear and several theories about the psychological mechanisms underlying the efficacy of exposure therapy have been advanced (Craske 1999). Emotional processing theory (e.g., Foa and Kozak 1986) initially was advanced as one explanation of the effectiveness of exposure therapy in the treatment of anxiety disorders. This article traces the origin of the concept of emotional processing, describes emotional processing theory and how it accounts for the effectiveness of psychological therapies in treating anxiety disorders, and discusses recent developments in the concept of emotional processing.

1. *History of the Emotional Processing Concept*

The concept of emotional processing has it origin in Lang's (1977) analyses of fear-relevant imagery in the context of behavior therapy and fear reduction. In studying the procedure of systematic desensitization, Lang et al. (1970) found three predictors of successful treatment: greater initial heart-rate reactivity during fear-relevant imagery, greater concordance between self-reported distress and heart rate elevation during fear-relevant imagery, and systematic decline in heart-rate reactivity with repetition of the imagery. On the basis of these findings, Lang (1977, p. 863) suggested that 'the psychophysiological structure of imagined scenes may be a key to the emotional processing which the therapy is designed to accomplish'. While Lang did not expand on the concept of emotional processing, in subsequent papers he further elaborated on the theory of emotional imagery in behavior therapy (e.g., Lang 1984).

Lang's (1977, 1984) theory holds that a fear image is a cognitive structure containing stimulus, response, and meaning information that serves as a program to avoid or escape from danger. For example (see Fig. 1), the fear structure of a dog phobic includes representations of dogs, representations of physiological and behavioral responses in the presence of dogs (e.g., rapid heart-rate, sweating, running away), and representations of meanings associated with dogs (e.g., 'Dogs are dangerous') and with responses to dogs (e.g., 'Rapid heart rate and sweating mean that I am afraid'). The fear structure is activated by environmental input that matches some of the information stored in the structure. According to Lang (1977, p. 867), 'the aim of therapy could be described as the reorganization of the image unit in a way that modifies the affective character of its response elements.' He further proposed that for image modification to occur, the fear structure must be at least partially activated. Accordingly, it is this aspect of the theory that accounts for the relationship between physiological measures (e.g., elevated heart-rate reactivity) and good outcome in exposure therapy for fear. However, Lang did not specify how activation of the fear structure brings about alterations in the structure.

The concept of emotional processing was subsequently taken up by Rachman (1980, p. 51) who proposed a working definition of emotional processing as 'a process whereby emotional disturbances are absorbed, and decline to the extent that other experiences and behaviors can proceed without disruption.' Accordingly, successful processing would be evidenced by a person's ability to confront a previously distressing event without evidencing signs of emotional distress. By contrast, several clinical phenomena

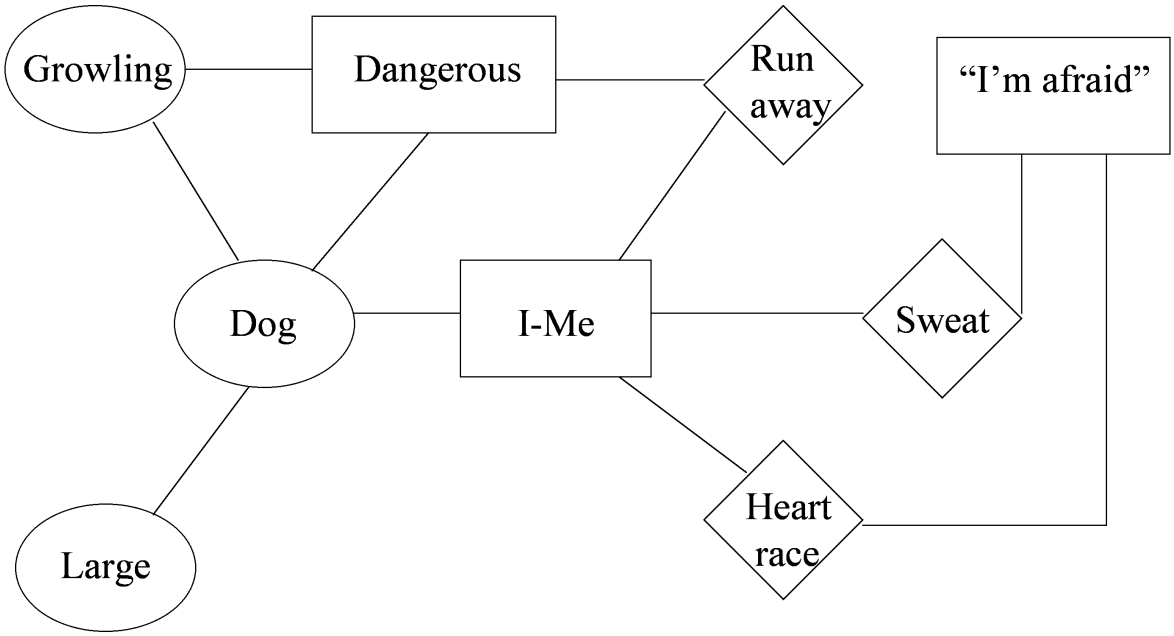

Figure 1
A schematic diagram of the fear structure of an individual with a dog phobia. The ovals indicate stimulus representations, diamonds indicate response representations, and rectangles indicate meaning representations. Lines indicate associations among representations

such as persistent fears and obsessions, failures to benefit from treatment designed to reduce fears, and the return of fear after apparently 'successful' treatment would be evidence of failed or incomplete emotional processing.

Rachman (1980) noted that the majority of people are able successfully to process the upsetting events that occur in their lives and proposed a number of situational, personality, and contextual variables that are likely to produce difficulties in processing an emotionally disturbing event. For example, difficulties in emotional processing are likely to occur in generally anxious and introverted individuals who confront intense, dangerous, unpredictable, and uncontrollable events. Rachman further discussed factors that can facilitate emotional processing, such as prolonged exposures to fear relevant stimuli or images without distraction. Rachman's analysis does not constitute a theory of emotional processing. Rather, it is an attempt to describe the sort of phenomena for which an adequate theory of emotional processing may provide an explanation. Indeed, as noted by Foa and Kozak (1986), Rachman's approach to defining and analyzing emotional processing suffers from circular reasoning: the fear reduction that is attributed to successful emotional processing is the same evidence used to infer that successful emotional processing has occurred.

2. *Emotional Processing Theory: Basic Concepts and Mechanisms Underlying Treatment*

2.1 *The Cognitive Structure of Fear*

Building on Lang's (1977, 1984) theory, Foa and Kozak (1986) conceptualized fear as a memory structure that serves as a blue print for escaping or avoiding danger that contains information about the feared stimuli, fear responses, and the meaning of these stimuli and responses. When a person is faced with a realistically threatening situation (e.g., a car accelerating at you) the fear structure supports adaptive behavior (e.g., swerving away). However, a fear structure becomes pathological when the associations among stimulus, response, and meaning representations do not accurately reflect reality in that harmless stimuli or responses are erroneously interpreted as being dangerous. Furthermore, characteristic differences between different anxiety disorders are thought to reflect the operation of specific pathological fear structures (cf. Foa and Jaycox 1999). For example, the fear structure of individuals with panic disorder with agoraphobia is characterized by fear of physical sensations associated with their panic symptoms (e.g., rapid breathing, heart palpitations) and interpretation of these sensations as meaning that they are having a heart attack or are going crazy. As a result of this

misinterpretation, panic disordered individuals avoid locations they anticipate will give rise to panic attacks or similar bodily sensations, such as physical exertion.

By contrast, the fear structure of individuals with posttraumatic stress disorder (PTSD) involves associations among trauma-related harmful stimuli and similar but non-harmful stimuli, erroneous interpretation of safe stimuli as dangerous (e.g., 'All men are rapists'), and erroneous interpretations of responses to mean self-incompetence (e.g., 'My symptoms mean I can't cope with this'). Accordingly, when encountering a trauma reminder, the PTSD sufferer experiences strong emotional reactions similar to those that occurred during the trauma. Support for these hypotheses comes from a study comparing traumatized individuals with and without PTSD to one another and to nontraumatized individuals using the Posttraumatic Cognitions Inventory (PTCI), a measure of negative cognitions about the self, negative cognitions about the world, and self blame. Individuals with PTSD showed higher scores on all three scales than did the other two groups who did not differ from one another (Foa et al. 1999).

2.2 Emotional Processing as Modification of the Fear Structure

If fear behavior reflects the activation of an underlying cognitive fear structure, then changes in the fear structure will result in corresponding changes in behavior. Foa and Kozak (1986) proposed that psychological interventions known to reduce fear, such as exposure therapy, achieve their effects through modifying the fear structure. They further suggested that, 'The hypothesized changes in the fear structure may be conceptualized as the mechanism for what Rachman (1980) has defined as emotional processing' (Foa and Kozak 1986, p. 22).

2.3 Necessary Conditions for Emotional Processing

According to emotional processing theory (Foa and Kozak 1986), two conditions are necessary for therapeutic fear-reduction to occur. Similar to Lang (1977, 1984), Foa and Kozak suggested the fear structure must be activated in order for it to be available for modification. Activation occurs when a person is exposed to information that matches some of the information stored in the fear structure. But activation by itself is not sufficient. New information that is *incompatible* with the pathological aspects of the fear structure must also be available and incorporated into the existing structure for emotional processing to occur. Indeed, exposure to information that is *consistent* with the existing fear structure would be expected to strengthen existing associations. Although experiences may have the effect of either strengthening or weakening existing fear structures, the term emotional processing is reserved for modifications that result in fear-reduction.

2.4 Indicators of Emotional Processing and Relation to Treatment Outcome

Fear and emotional processing are theoretical constructs that cannot be observed directly. Rather, they must be inferred through observable variables. Because emotional processing is being offered as an *explanation* of fear reduction, it is important that the indicators of emotional processing be different from the indicators of fear reduction. To avoid circularity, Foa and Kozak (1986) distinguished between measures of emotional processing during the course of exposure therapy and treatment outcome. Specifically, they proposed that emotional processing can be inferred from: (a) self-reports (e.g., 'I'm afraid') and physiological reactivity (e.g., elevated heart rate) that are indicative of fear activation; (b) gradual decreases of these measures during continued exposure to fear-relevant information; and (c) diminished initial fear reactions across successive exposures. In contrast to these indicators, Foa and Kozak viewed treatment outcome as a much broader notion that encompasses variables not specifically targeted in treatment, such as the quality of a person's job performance, personal relationships, and their degree of life satisfaction.

Several studies are consistent with the hypothesized relationship between indicators of emotional processing and treatment outcome. For example, individuals with PTSD experience distressing intrusive thoughts about their traumatic experience; carefully avoid people, places, and activities that remind them of the traumatic event; and experience disruptions in memory, concentration, and sleep. An effective treatment for PTSD is imaginal exposure to memories of the traumatic experience (Shalev et al. 1996). It has been found that individuals who exhibited fear activation during imaginal reliving, followed by decline across sessions achieved better outcome (e.g., fewer intrusive thoughts, less avoidance, and improvements in memory, concentration, and sleep) than individuals who did not exhibit fear activation or who exhibited fear activation with minimal habituation across sessions (cf. Foa 1997). The validity of these predictors of outcome has been demonstrated in other anxiety disorders as well, such as obsessive-compulsive disorder and specific phobias (Foa and Kozak 1986).

2.5 Sources of Corrective Information

It is easy to see how exposure therapy, in which fearful individuals are helped to confront feared images or situations, results in activation of a fear structure. But what kind of information does exposure therapy

provide to correct erroneous associations among the stimulus, response, and meaning representations in the fear structure?

One important source of corrective information is *habituation*, by which repeated exposure to a stimulus that reliably elicits a specific response results in a decline in that response. Foa and Kozak (1986) proposed that habituation of physiological responses within an exposure therapy session, one of the indicators of emotional processing, constitutes an important source of information for modifying aspects of the fear structure. For example, habituation generates new response information about the *absence* of physiological responding in the presence of the phobic stimulus that is incompatible with prior response information. Physiological habituation also counters erroneous meaning representations underlying the cognition that fear will persist until escape occurs. Repeated and prolonged exposure may also provide corrective information as to the realistic likelihood of feared consequences. For example, the fear structure of an individual with obsessive compulsive disorder may include meaning representations with unrealistically high estimations of the probability of contracting a severe illness from touching doorknobs. As a result, doorknobs are carefully avoided thereby preventing the person from learning that touching doorknobs does not pose a serious health risk. Treatment through exposure (e.g., touching dirty doorknobs) and response prevention (e.g., refraining from washing) provides the corrective information that the probability of becoming sick after touching a doorknob is actually extremely low.

3. *Recent Advances in Emotional Processing Theory*

3.1 *Mechanisms Underlying Natural Recovery from Trauma*

Studies investigating patterns of reactions to a trauma over time indicate that individuals differ in their ability to successfully process the traumatic event and thus recover naturally with the passage of time. While most trauma victims process the trauma successfully, a substantial minority fails to do so and consequently develops chronic PTSD (Foa 1997). Why do some traumatized individuals recover and others go on to develop chronic disturbances? Foa (1997) suggested that three common mechanisms underlie both natural recovery from trauma and reduction of PTSD severity via treatment. The first mechanism is *emotional engagement* with the trauma memory, which in the context of exposure therapy has been termed fear activation. In relation to natural recovery, emotional engagement refers to fear activation that occurs upon encountering a trauma reminder in the natural environment. Just as fear activation has been found positively associated with treatment outcome, the absence of emotional engagement has been found to be associated with poor natural recovery following a traumatic event. For example, individuals who *dissociate* during or shortly after experiencing a traumatic event experience greater PTSD symptom severity later on, compared to those who do not dissociate. The dissociation phenomenon connotes the absence of emotional engagement with the trauma memory and hence, according to emotional processing theory, will retard emotional processing. A delay in the peak reaction to a traumatic event may also be conceptualized as the absence of emotional engagement and thus is expected to hinder emotional processing and subsequent recovery. Consistent with this hypothesis, individuals whose peak symptom severity occurred within two weeks of the traumatic event were less symptomatic 14 weeks later than were individuals whose symptoms peaked between 2–6 weeks after the event. A second mechanism hypothesized to underlie both natural recovery and improvement with treatment is the *degree of organization or articulation* of the trauma narratives (Foa 1997). In support of this hypothesis, a higher degree of narrative articulation shortly after the trauma predicted greater recovery (lower PTSD symptom severity) three months later. In the same vein, exposure therapy for PTSD was associated with increased organization and decreased disorganization of the trauma narrative over the course of treatment, and those changes were correlated with symptom improvement.

The third hypothesized mechanism for recovery is a *change in core cognitions* about the world and the self. As discussed earlier, two basic dysfunctional cognitions are thought to underlie PTSD: the world is completely dangerous and the self is totally incompetent. It follows that recovery would be associated with less negative cognitions (Foa and Jaycox 1999). Indeed, in a longitudinal study of crime victims, individuals who retained their PTSD diagnosis three months after their assault exhibited more negative cognitions compared to those who no longer had PTSD. Similarly, cognitions about the world and self were less negative after treatment for PTSD than they had been before treatment (see Foa 1997, for discussion of the preceding two studies).

The ways in which exposure therapy is thought to promote fear activation and modification of erroneous cognitions have been explicated earlier. But how does emotional processing theory account for natural recovery? It is proposed that natural emotional processing and the consequent recovery occurs via repeated activation of the trauma memory in daily life through engagement with trauma-related thoughts and feelings, sharing them with others, and confronting trauma-reminder stimuli. In the absence of repeated traumatization or 'emotional breakdown,' these natural exposures contain information that disconfirms the

common post-trauma cognition that the world is a dangerous place. For example, a rape victim may repeatedly encounter men that are physically similar to the perpetrator, thus activating the trauma memory structure. As these encounters do not lead to additional assaults, the natural initial postrape inclination to view the world as entirely dangerous is not confirmed and thus gradually subsides. In addition, repeated thinking or talking about the traumatic event promotes habitation of negative emotions and the generation of an organized trauma narrative. By contrast, individuals who avoid engaging with the traumatic memories by either suppressing related thoughts and feelings or avoiding trauma-related situations, do not have the opportunity to form an organized narrative or disconfirm the initial post-trauma negative cognitions and thus develop chronic disturbances. Indeed, mothers who lost their babies to Sudden Infant Death Syndrome and were able to share their loss with a supportive social network resolved their grief more satisfactory 18 months after the trauma than did mothers who were not able to share their grief with supportive others (Lepore et al. 1996).

3.2 Emotional Numbing and Emotional Processing in PTSD

In an apparent paradox, individuals with PTSD frequently report experiencing intense negative emotions when reminded about the traumatic (re-experiencing) while also having difficulty experiencing and expressing emotions in situations that call for an emotional response (emotional numbing). To explain this paradox, Litz et al. (2000) hypothesized that the strong emotional reactions cued by trauma reminders among individuals with PTSD temporarily depletes their emotional resources, thereby reducing emotional responsiveness to subsequent events. Consistent with this hypothesis, veterans with PTSD who were first shown a video tape of combat images and then asked to rate a series slides depicting emotionally positive images (e.g., a baby) were emotionally more reactive to the combat video, but less reactive to the positive slides than veterans without PTSD. In contrast, no differences between groups emerged when the veterans viewed neutral images (e.g., furniture) before rating the positive slides. Thus, veterans with PTSD displayed emotional numbing for the positive slides, but only after first experiencing negative emotional arousal.

The research by Foa and co-workers (see Foa 1997) on the consequences of emotional nonengagement and by Litz et al. (2000) on the relationship between arousal and emotional numbing suggests two ways in which extreme emotional reactivity following a traumatic event may interfere with emotional processing and recovery. According to Foa acute emotional reactivity can lead to extensive avoidance, thereby preventing exposure to corrective information. Litz and co-workers proposed that chronic emotional reactivity can lead to emotional numbing, thereby reducing emotional engagement with the trauma memory.

3.3 The Effects of Corrective Experiences: Alteration of Existing Cognitive Structures or Development of New Cognitive Structures?

Foa and Kozak (1986) defined emotional processing as a change in the pathological elements of an existing fear structure following exposure to corrective information. A plausible alternative is that exposure to corrective information results in the formation of a *new* memory structure that coexists with the old fear memory rather than replacing it (Bouton 1988). Accordingly, a person's behavior in a given situation will be determined by which of the two structures is activated. From Foa and Kozak's perspective, the return of fear following successful treatment represents a further change in the fear structure. Bouton's theory would view relapse as a reduction in the accessibility of the new nonfearful memory. Support for Bouton's hypothesis comes from research conducted by Craske and co-workers (cf. Craske 1999), demonstrating that certain procedural variations in treatment by exposure therapy produced less between-session fear reduction, but prevented the long-term return of fear at follow-up.

3.4 Worry and Depressive Rumination

Emotional processing concepts have been utilized in the analysis of worry (Borkovec and Lyonfields 1993) and depressive rumination (Teasdale 1999). Although somewhat different analyses have been offered for these two phenomena, both Borkovec and Teasdale suggested that cognitive processing in one channel or at one level of meaning (linguistic/conceptual or propositional) may interfere with processing in another channel or level of meaning (imagery or implicational). Specifically, on the basis of studies showing that verbal articulation of fearful material results in less physiological arousal than imagining the same material (e.g., Vrana et al. 1986), Borkovec and co-workers have proposed that worry serves to inhibit emotional imagery. The short-term consequence of worry is to minimize emotional distress by limiting activation of the fear structure. However, the long-term consequence of avoiding full activation of the fear structure, as discussed above in the context of emotional numbing and emotional nonengagement, is to inhibit emotional processing and thereby maintain the worry.

Although the concept of emotional processing has generally been limited to the field of anxiety, Teasdale and co-workers (cf. Teasdale 1999) have proposed an

emotional processing model of depression. The model distinguishes between two different, but interacting levels of cognitive meaning, and proposes three different configurations of the interactions between these levels of meaning, also referred to as the 'three modes of mind.' One level of meaning is propositional, corresponding to the kind of information that can be conveyed in a single sentence. The other, implicational, represents higher order meaning derived from recurring patterns or themes. In the case of depression, the implicational meaning of experiences such as failing a class or the break-up of an important relationship may be a failure in all aspects of one's life. Importantly, the model posits that emotions are directly generated by the implicational level, while the propositional level of meaning only effects emotional states indirectly through the continual cycling of information between the two levels of meaning. Depressive rumination involves the cycling of depresogenic information between these two levels of meaning, thereby generating and maintaining dysphoria.

Teasdale (1999) further suggested that each meaning level can operate in one of two modes. In the *direct* mode, processing occurs in real time, whereas in the *buffered* mode information is allowed to accumulate before it is processed further. Although both levels of meaning can be operating in direct mode at the same time, only one level can be buffered at a given time. This yields three possible configurations, each one having different consequences for the person's emotional experience. When neither level of meaning is buffered, 'mindless emoting' occurs, in which people identify themselves with their emotional reactions and engage in little self-reflection. When the propositional level of meaning is buffered, individuals are said to be in the 'conceptualization/doing' configuration in which they think about themselves and their emotions as objects and develop goal-oriented strategies to cope with their problems and emotions. For example, depressive rumination frequently involves the person thinking about the nature, causes, and consequences of their depression. The third configuration, called 'mindful experiencing/being,' occurs when the implicational level is buffered and is characterized by 'integrated cognitive-affective inner exploration, use of present feelings and "felt senses" as a guide to problem solution and resolution, and a non-evaluative awareness of present subjective-self-experience' (p. S77).

Within this model of depression, Teasdale (1999) has defined effective emotional processing as 'processing that leads to modification of the affect-related schematic models [i.e. implicational level of meaning] that support processing configurations maintaining dysfunctional emotion' (p. S65). He further hypothesizes that emotional processing will occur only when the implicational system is buffered, thereby allowing the integration of new and old information. Accordingly, two general strategies can be used to promote changes in implicational meaning. One strategy is to help depressed individuals create and encode in memory new implicational meanings, and many current cognitive therapy techniques (e.g., cognitive restructuring) help in doing so, although such techniques may run the risk of maintaining the conceptualizing/doing configuration. A second strategy, unique to Teasdale's theorizing, is to help people learn to shift from either of the two modes of mind that do not promote emotional processing (mindless emoting and conceptulizing/doing) to the one that does (mindful experiencing/being). This involves exercises in mindfulness, similar to some forms of meditation, in which individuals learn to become aware of, and acknowledge depresogenic thoughts without exclusively focusing on them. Teasdale found that a treatment program combining traditional cognitive therapy interventions with mindfulness exercises was more effective in preventing relapses among individuals with recurrent major depression than was cognitive therapy alone.

4. *Concluding Comments*

Emotional processing theory was advanced by Foa and Kozak (1986) primarily to account for the effectiveness of exposure therapy in the reduction of anxiety. However, consideration of its historical antecedents (Lang 1977, Rachman 1980) indicates the concept of emotional processing originally was conceived more broadly. Rachman (1980) in particular proposed that the concept should encompass a range of phenomena that includes, but is not limited to, anxiety and the efficacy of exposure therapy. In keeping with this, recent advances in such topic areas as natural recovery following trauma (Foa 1997), PTSD symptom structure (Litz et al. 2000), worry (Borkovec 1993) and relapse following recovery from depression (Teasdale 1999) have extended the domain of emotional processing theory and demonstrate its continued heuristic value.

At the same time, questions remain for future research as to the nature and limits of emotional processing, two of which are briefly noted. First, does exposure to corrective information alter existing cognitive structures, as proposed by Foa and Kozak (1986), or does it result in the acquisition of a new memory (Bouton 1988, Craske 1999)? Resolution of this issue has implications for understanding the process of relapse and development of relapse-prevention strategies. Second, to what extent do procedures that promote emotional processing of fear and anxiety promote the emotional processing of other problematic emotional reactions, such as sadness/depression, anger, or guilt? In exposure therapy, an important source of corrective information is habituation. However, it may be that these other emotions do not habituate (Pitman et al. 1991) and that alternative procedures may be needed to promote their

emotional processing, such as traditional cognitive restructuring techniques and mindfulness strategies (Teasdale 1999). Thus, resolution of this issue has implications for the design of interventions for other problematic (nonfear) emotional reactions.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Behavior Psychotherapy: Rational and Emotive; Behavior Therapy: Psychiatric Aspects; Behavior Therapy: Psychological Perspectives; Emotion, Neural Basis of; Emotions and Health; Emotions, Psychological Structure of; Fear Conditioning; Habituation; Panic Disorder

Bibliography

Borkovec T D, Lyonfields J D 1993 Worry: Thought suppression of emotional processing. In: Krohne H W (ed.) *Attention and Avoidance*. Hogrefe & Huber, Seattle, WA, pp. 101–18

Bouton M E 1988 Context and ambiguity in extinction of emotional learning: Implications for exposure therapy. *Behaviour and Research Therapy* **26**: 13–149

Craske M G 1999 *Anxiety Disorders: Psychological Approaches to Theory and Treatment*. Westview Press, Boulder, CO

Foa E B 1997 Psychological processes related to recovery from a trauma and an effective treatment for PTSD. In: Yehuda R, McFarlane A (eds.) *Psychobiology of PTSD Annals of the New York Academy of Sciences*, New York, pp. 410–24

Foa E B, Ehlers A, Clark D M, Tolin D F, Orsillo S M 1999 The Posttraumatic Cognitions Inventory (PTCI): Development and validation. *Psychological Assessment* **11**: 303–14

Foa E B, Jaycox L H 1999 Cognitive-behavioral theory and treatment of posttraumatic stress disorder. In: Speigel D (ed.) *Efficacy and Cost-effectiveness of Psychotherapy*. American Psychiatric Association, Washington, DC, pp. 23–61

Foa E B, Kozak M J 1986 Emotional processing of fear: Exposure to corrective information. *Psychological Bulletin* **99**: 20–35

Lang P J 1977 Imagery in therapy: An information processing analysis of fear. *Behavior Therapy* **8**: 862–86

Lang P J 1984 Cognition in emotion: Concept and action. In: Izard C, Kagan J, Zajonc R (eds.) *Emotion, Cognition and Behavior*. Cambridge University Press, New York, pp. 193–206

Lang P J, Melamed B G, Hart J 1970 A psychophysiological analysis of fear modification using an automated desensitization procedure. *Journal of Abnormal Psychology* **76**: 220–34

Lepore S J, Silver R C, Wortman C B, Wayment H A 1996 Social constraints, intrusive thoughts, and depressive symptoms among bereaved mothers. *Journal of Personality and Social Psychology* **70**: 271–82

Litz B T, Orsillo S M, Kaloupek D, Weathers F 2000 Emotional processing in posttraumatic stress disorder. *Journal of Abnormal Psychology* **109**: 26–39

Pitman R K, Altman B, Greenwald E, Longpre R E, Macklin M L, Poire R E, Steketee G S 1991 Psychiatric complications during flooding therapy for posttraumatic stress disorder. *Journal of Clinical Psychiatry* **52**: 17–20

Rachman S 1980 Emotional processing. *Behaviour and Research Therapy* **18**: 51–60

Shalev A, Bonne O, Eth S 1996 Treatment of posttraumatic stress disorder: A review. *Psychosomatic Medicine* **58**: 165–82

Teasdale J D 1999 Emotional processing, three modes of mind and the prevention of relapse in depression. *Behaviour and Research Therapy* **37**: S53–S77

Vrana S R, Cuthbert B N, Lang P J 1986 Fear imagery and text processing. *Psychophysiology* **23**: 247–53

F. B. Foa and S. P. Cahill

Psychological Treatment Delivered Nontraditionally

The term 'psychotherapy' generally connotes an interactive, face-to-face process in which a trained mental health professional attempts to help a client or group of clients effect personal change. Authority and responsibility for facilitating such change reside with the professional, whose presence and expertise are deemed necessary to guide and control the meetings in accordance with sound clinical and ethical practices.

Since the 1970s an array of interventions for delivering psychological assistance have come into being that deviate significantly from the traditional psychotherapy format. In general these newer formats enjoy widespread popularity among the general public and their use is less directly controlled by a mental health professional. Such alternative psychological interventions include self-help and mutual support groups, radio and TV shows dispensing psychological advice, interactive computer software packages for addressing a wide variety of personal problems, self-help books, self-help audio and video tapes, and most recently a spate of cyberspace offerings. The latter category includes self-help guides, single session e-mail and bulletin board advice, ongoing e-mail counseling, real-time chat groups, and most recently, web telephony and videoconferencing.

A small body of research evaluating effectiveness focuses primarily on paraprofessionals and self-help books, and shows these alternatives to be sometimes as effective as more classic psychotherapy. However, much more research is needed. Added benefits of alternative formats are lower cost, convenience of access and use, and the availability of psychological aid for people who otherwise either would not, or could not, utilize traditional psychotherapy. Examples include the homebound, those in remote areas, people who consider psychotherapy stigmatizing, and individuals with a strong need for privacy and independence when dealing with their psychological problems.

This article first examines some of the social and cultural factors contributing to the development and use of these newer forms of psychological treatment. The treatments themselves are then reviewed along two dimensions. The first dimension is the source of the help: is it professional, paraprofessional, nonpro-

fessional, or some mix? The second dimension is the nature and degree of contact between help giver and help seeker. Finally, the status of research, along with issues about quality, safety, and protection of the public are briefly discussed.

1. Social Climate and the Rise of New Forms of Psychological Treatment

Viewed as a group, the newer models of psychological assistance, as compared with traditional forms of psychotherapy, tend to involve less direct professional contact and control, even when the treatment is designed or delivered by a mental health professional. This shift away from strictly professional delivery of mental health services reflects an intersection of several social developments. Observations about social trends are based on developments in the USA and may differ in applicability to other parts of the world.

At a broad societal level, considerable value has come to be placed on empowering the individual, as opposed to almost total reliance on institutions and professionals for information, support, and assistance. This stems in part from a post-Watergate, post-Vietnam weakening of US public confidence in its major institutions, including the corporate world, government, and the medical establishment. Various activist movements (e.g., those concerning civil rights, women, consumer rights, gays and lesbians), beginning in the 1960s and continuing to the present, also reinforce values about personal entitlement and empowerment. A result of this trend is that people now demand a greater measure of involvement and control over their physical and mental health care.

Technological advances in communications are a further factor in stimulating individual desire for greater control of over personal well-being. Before the 1980s many personal issues such as alcohol and drug abuse, child molestation, family violence, sexual orientation, bulimia and anorexia nervosa, cancer, sexually transmitted diseases, Alzheimer's disease, fetal alcohol syndrome, mental illness, to name but a few, were topics rarely spoken of in public. Early television talk shows played an important role in shining a spotlight on such previously closeted issues. As people became increasingly sophisticated about the relationship between stress and health, and the potential damage that various stressful life events can cause, the public also demonstrated its eagerness for help.

However, scientific surveys of the prevalence and incidence of mental health problems, at least insofar as the US public is concerned, make it clear that we do not now have, and probably never will have, a sufficient number of trained mental health professionals to meet the needs of all those reporting diagnosable psychological problems. Meanwhile, the need for psychological services continues to grow, fed in part by dramatic advances in medical science. More people with serious chronic medical conditions are surviving, many into old age, putting numerous stresses on both patients and their families.

At the same time, the growth of managed health care, with its focus on cost containment, has resulted in limiting access to care, especially for mental health concerns (see *Managed Care and Psychotherapy in the United States*). This is so despite a growing body of research from the field of psychoneuroimmunology (Hafen et al. 1996), demonstrating that psychological stressors have a negative impact on the immune system, general health, and overall functioning (see *Psychoneuroimmunology*). The same body of research also shows the social support and the opportunity to discuss problems constructively can positively affect health, enhance the quality of life, and even promote longevity. Western medicine, though desirous of promoting health and preventing the occurrence or the worsening of disease, fails fully to capitalize on what is known about the mind–body connection and its impact on health and well-being.

Further, under managed care, doctors and nurses, traditionally an important source of informal psychological assistance, now have less time to spend with their patients than in the past. The ironic result is a public increasingly aware of the connection between stress and physical and psychological illness, but finding affordable psychological services growing more scarce.

To summarize, four factors stand out as particularly influential in promoting nontraditional psychological treatments: (a) the growing public awareness and desire for help at the very time that for many the availability of affordable professional psychotherapy is decreasing; (b) the value placed on personal entitlement and empowerment in today's society; (c) an ongoing revolution in electronic forms of communication, including the very rapid growth and easy accessibility of information and services via the Internet; (d) the rapid dissemination of information by today's highly competitive mass media businesses. Given these forces, it is not surprising that enterprising people have, and no doubt will continue to find, alternative, accessible, less costly (although sometimes financially quite profitable) ways of delivering psychological treatments.

One such example of a break from traditional psychotherapy began in the late 1970s, as the number and type of self-help/mutual support groups multiplied rapidly (Goodman and Jacobs 1994). Such groups had existed for decades, the best known being Alcoholics Anonymous and Recovery, Inc. (see *Alcoholics Anonymous*). But suddenly groups sprang up for almost every conceivable personal problem from substance abuse, to parenting concerns, to being a victim of violence, to life transitions such as divorce, retirement, or bereavement, to lifestyle issues such as sexual orientation, transvestitism, open marriages, as well as groups for almost every disease listed by the

World Health Organization. Millions of people joined them and, although some groups utilize professional leaders or professionally developed ideas, many have little or no professional input. Sociologist Frank Reissman (Gartner and Reissman 1984) terms the heart of the self-help group movement the 'helper-therapy' principle—the idea that when people suffering from a common problem come together for mutual support, they not only receive help from others, particularly from those more experienced in coping with the problem, but also give help, and in so doing are themselves strengthened and empowered. In the 1980s many health professionals held negative views about self-help groups, asserting that lay people left on their own were likely to harm each other by dispensing bad advice. Less openly discussed were their concerns over loss of professional control and income. Over time professionals have become more supportive of self-help groups as both an adjunct to traditional treatment and as a useful, member-governed intervention in its own right. Today many professionals are directly involved with groups in a variety of roles such as consultants, researchers, leaders, co-leaders with group members, and even as participants who share the group's common concern.

More recent alternatives to traditional psychotherapy have been generated as a result of the ongoing technological revolution in electronic communication, which continues to offer new means of interacting to large numbers of people. On the Internet, some formats take advantage of real-time communication to deliver psychological aid, while others necessitate time delays between the posting of messages requesting psychological help and the responses to those messages. One recent review of psychological assistance on the Internet (Barak 1999), listed 10 types of interventions: (a) information resources on psychological concepts and issues; (b) self-help guides; (c) psychological testing and assessment; (d) help in deciding to undergo therapy; (e) information about specific psychological services; (f) single-session psychological advice through e-mail or e-bulletin boards; (g) ongoing personal counseling and therapy through e-mail; (h) real-time counseling through chat, web telephony, and videoconferencing; (i) real-time and delayed-time support groups, discussion groups, and group counseling; (j) psychological and social research. The next section offers a framework for organizing the many and varied nontraditional intervention formats offered to the public.

2. *A Framework For Viewing Nontraditional Forms Of Psychological Assistance*

Two basic elements in most traditional psychotherapy are (a) live, real-time, face-to-face contact between a help seeker and a help giver and (b) the help giver being a trained mental health professional. This section outlines alternative psychological treatments in terms of their departure from these two elements, that is, alternative formats for delivering psychological aid are shown according to two major categorizations.

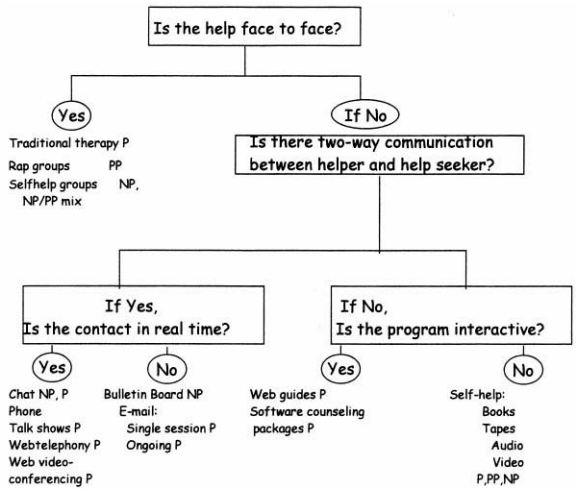

Figure 1
Nature and type of psychological treatments

The first categorization is the nature of the contact between help seeker and the source of the help: is it live and face-to-face? If not, is there some type of two-way communication between the helper and help seeker? If so, is it in real time or delayed time? (Note that in either case, the true identity of either party may not be known.) If not, does the help seeker interact and receive feedback from a packaged program such as those that can be bought and used on personal computers or accessed via the Internet? If not, what is the format for the delivery of the noninteractive aid?

The second categorization refers to the training of the person offering the mental health aid: is the person a trained mental health professional, a paraprofessional, or a nonprofessional? The intention is to provide a framework for categorizing any type of psychological treatment. For the examples listed in Fig. 1, P = professional, PP = paraprofessional, and NP = nonprofessional. The listing is not intended to be exhaustive.

The relatively unregulated nature of the various forms of nontraditional psychological treatment being offered to the lay public has raised concerns about their safety, quality, and effectiveness. Research into the use of paraprofessionals vs. professionals, self-help books vs. therapist-delivered services, and computer-assisted vs. therapist-delivered treatment generally shows these alternatives to be similar in effectiveness to professionally delivered psychotherapy (Christensen and Jacobson 1994). But not only is that body of scientific studies relatively small, some newer

computer-based formats have never been studied. Thus, many important questions remain unanswered. For example, are some nontraditional formats more appropriate for certain types of people, or certain types or levels of severity of problem?

Further, the relatively unregulated nature of the Internet leaves open the possibility of poor quality offerings, untested claims, and even outright charlatanism among those designing and offering psychological assistance. In the interest of assuring quality psychological services for the public, priority needs to be given to research, as well as education for both Internet users and service providers, about appropriate standards.

See also: Psychotherapy: Ethical Issues; Psychotherapy Integration; Psychotherapy Process Research; Social Support and Health; Social Support and Recovery from Disease and Medical Procedures; Social Support and Stress; Social Support, Psychology of; Support and Self-help Groups and Health

Bibliography

Barak A 1999 Psychological applications on the Internet: A discipline on the threshold of a new millennium. *Applied and Preventative Psychology* **8**: 231–45

Christensen A, Jacobson N S 1994 Who (or what) can do psychotherapy: The status and challenge of nonprofessional therapies. *Psychological Science* **5**: 8–14

Gartner A, Reissman F 1984 *Self-help Revolution*. Human Services Press, New York

Goodman G Jacobs M K 1994 The self-help, mutual support group. In: Fuhriman A, Burlingame G M (eds.) *Handbook of Group Psychotherapy*. Wiley, New York, pp. 489–526

Hafen B Q, Karren K J, Frandsen K J, Smith N L 1996 *Mind/Body Health The Effects of Attitudes, Emotions and Relationships*. Allyn and Bacon, Boston, MA

M. K. Jacobs

Psychological Treatment, Effectiveness of

1. Introduction

Despite its relative youth, the field of psychotherapy has faced many challenges, including one that disputed the very purpose it intended to accomplish. In a review of the effectiveness of psychotherapy, Eysenck (1952) concluded that 74 percent of patients, drawn from 24 studies, and diagnosed as neurotic, were found to make progress considered equivalent to a similarly diagnosed group of individuals who received no treatment. This finding stimulated years of controversy regarding the effectiveness of psychotherapy, as well as a determined and focused effort on the part of researchers to establish empirically the beneficial effects, if any, of psychotherapy. In addition to reporting the findings supporting the efficacy of psychotherapy treatment, the remainder of this discussion will present related findings from this effort: (a) the durable effects of psychological treatment, (b) the factors leading to patient improvement, and (c) the disturbing evidence that suggests that not all patients benefit from treatment. Of related importance to this discussion is the paradoxical outcome that often accompanies many scientific investigations; namely, that as each question is sufficiently answered, more questions take its place.

2. The Definition of Psychotherapy

In general, psychological treatment is a method designed to change a person's thoughts, feelings, or behaviors with the intention of facilitating increased functioning and life adjustment. Approaches to psychotherapy are numerous and diverse including such orientations as cognitive, behavioral, client centered, gestalt, and existential therapies. All of the various approaches involve a confiding relationship with a trained therapist focused on a patient in need of assistance.

3. The Effectiveness of Psychotherapy

As mentioned above, the conclusion made by Eysenck stimulated both the frequency and quality of studies designed to evaluate the outcome of therapy. Researchers pursuing evidence to refute Eysenck's claim have investigated a number of domains that are considered relevant to distinguishing the effects of psychotherapy. Assessments of the benefits of psychotherapy have focused on such domains as an increase in problem-solving skills, reduction in symptoms, improvements in interpersonal and social functioning, and health. The changes noted in patients treated with psychotherapy have been compared to groups of individuals who have not been exposed to therapy. Reviews of this literature have confirmed the beneficial effects of therapy, concluding that treated patients consistently fare better than untreated patients (Lambert and Bergin 1992, 1994).

An often used statistical technique, meta-analysis, has enabled researchers to summarize individual collections of empirical data, comparing the effect sizes of treatment, placebo, and control groups. For example, using meta-analysis, Smith et al. (1980) reported that the average person in treatment is better off at the end of treatment than 80 percent of the untreated

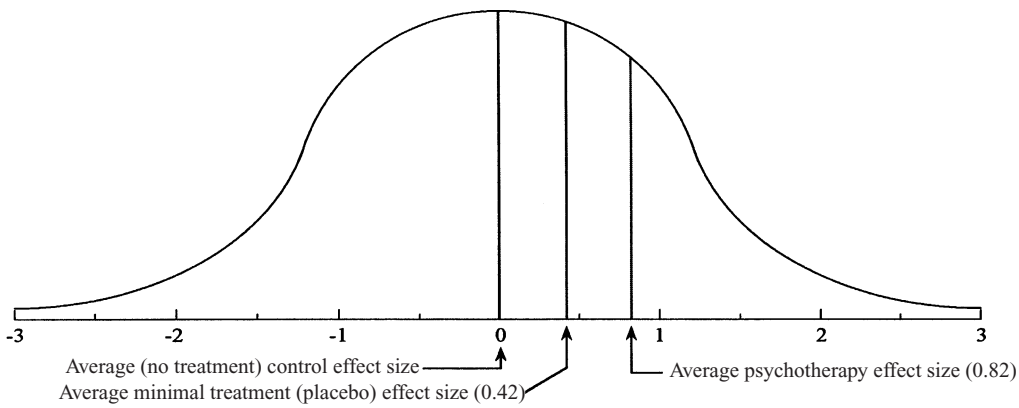

Figure 1
Comparison of placebo and psychotherapy effects in relation to no-treatment control

persons. A graphic representation of the effect sizes (treatment, placebo, and control) found in a number of meta-analytic reviews is presented in Fig. 1.

4. *Treatment Duration and Lasting Effects*

In contrast to the length of treatment associated with early theories of psychoanalytic psychotherapy, the extant literature suggests that substantial benefits from treatment occur in relatively short durations, and equally important, these gains are maintained. Dose-effect designs, derived from the dosage model introduced by Howard et al. (1986), have confirmed a positive rate of effect as the dose (i.e., number of sessions) increases, with diminishing returns at higher dosage levels. Studies from the dose-effect literature suggest that, for many clinical disorders, a range of 10 to 20 sessions is sufficient for 50 percent of patients to return to normal functioning. Moreover, promising results occur as the dosage level is increased to the extent that 75 percent of patients can be expected to recover substantially after 26 to 50 sessions (Lambert et al. 2001).

Although a majority of patients experience significant benefits from psychotherapy, there is also evidence that different clusters of psychological symptoms respond to treatment at distinct rates of improvement. It appears that symptoms of poor morale respond most quickly, followed by chronic symptoms of distress, with symptoms intrinsic to the character of the patient recovering the slowest. Observed trends in patient improvement have led some researchers to propose a phase model of recovery in which personal welfare, symptom reduction, and interpersonal and social functioning improve in sequence.

In addition to results supporting relatively short recovery periods for a majority of patients, there is evidence that beneficial gains from psychotherapy are enduring. Although adjustment from psychotherapy treatment does not prevent recurrences of psychological disturbance, it appears that many patients maintain healthy functioning for prolonged periods of time subsequent to treatment. However, exceptions to the notion of lasting treatment gains are noted in the conclusions of a review by Nicholson and Berman (1983), which included patients suffering from symptoms of substance abuse, eating, recurrent depression, and personality disorders.

5. *Factors Important to Psychotherapy Outcome*

In addition to the research focused on supporting the effectiveness of psychotherapy, considerable emphasis has been placed on the factors attributable to positive patient outcome. Findings from this research fall into three general categories: patient variables, common variables, and specific intervention variables.

5.1 *Patient Variables*

Although a number of studies have investigated the impact therapists have on psychotherapy outcome, there is surprisingly little evidence that suggests therapists or their techniques are central and essential factors affecting patient outcome. On the contrary, it is in fact the patient that is most likely to determine successful outcome. The literature focusing on the study of patient variables is too voluminous to repeat, and thus only the most important findings are reported here.

Each patient entering treatment brings a diverse array of factors when presenting to the clinician. These include a variety of psychological disorders, historical backgrounds, stresses, and the quality of their social support networks to name just a few. Patient variables found to have relationships with outcome are severity

of disturbance, motivation and expectancy, capacity to relate, strength of ego, psychological mindedness, and ability to recognize relevant problems (Garfield 1994). Therefore, patients suffering from challenging symptoms as well as relative deficits in numerous patient variable categories are likely to experience negligible improvement during treatment. For example, the borderline, alcoholic patient with suicidal tendencies who is forced into treatment, believing that most marital problems are a result of an insensitive spouse, is likely to benefit less from treatment than the depressed patient who voluntarily begins treatment and is determined to make personal changes that will lead to marital harmony.

Although patient variables have been conveniently categorized, the impact each has on the outcome of therapy is not equal. In general, the severity of disturbance is the patient variable that is most related to the outcome of the treatment. Patients with serious mental disorders, such as the schizophrenic, schizoaffective, and bipolar disorders are typically treated primarily with psychopharmacological management, suggesting the challenge these disturbances provide for psychotherapy alone. Personality disturbances also appear immutable to change in the short term use of psychotherapy.

5.2 Common Factors

Because outcome researchers have consistently failed to find sufficient evidence indicating superior effect sizes for particular theoretical modalities, some have concluded that psychotherapy equivalence is a result of factors common to all theories of psychotherapy. Whereas some view these beneficial ingredients as both necessary and sufficient for change, others perceive that these ingredients act in coalition with variables that are unique to particular interventions. It is important to note that regardless of the position held, factors common to all treatments are likely to account for the greatest amount of the improvement in patients undergoing psychotherapy.

In a review of the empirical research, Lambert and Bergin (1994) grouped many of the shared active ingredients of a variety of psychotherapies into three categories: support, learning, and action factors. The factors included in these categories seem to operate most actively during the in-session process of therapy, resulting in a collaborative working relationship that allows clients to confront and cope with reality more effectively. Factors that engender a collaborative relationship include a client's increased sense of trust, security, and safety, along with a reduction in tension, threat, and anxiety that in turn promotes a client's awareness of problems and ultimately new and more healthy perspectives on fears, appropriate risk taking, and motivations for resolving problems in interpersonal relationships.

Factors receiving the most attention in the literature are the common factors considered to be the core of therapy (empathy, warmth, and positive regard). A number of related terms such as acceptance, tolerance, therapeutic alliance, working alliance, and support are also used to describe these ingredients (see *Therapist–Patient Relationship*). In a study designed to explore which of several therapist variables distinguished most consistently between more effective and less effective therapist trainees, Lafferty et al. (1991) concluded that a client's perception of empathy is important to outcome, because less effective therapists, those with patients experiencing more distress as a result of treatment, were found to have lower levels of empathy.

Findings from studies investigating the therapeutic relationship have suggested that a positive working alliance between the patient and therapist is a necessary condition for patient improvement. When asked for factors that were most helpful in their treatment, patients tended to endorse the personal qualities of their therapists, using adjectives such as 'sensitive,' 'honest,' and 'gentle.' The results of similar studies confirm the import that patients place on the characteristics of therapists (Horvath and Greenberg 1994).

Additional evidence supporting the necessity for a strong working relationship is the positive association repeatedly found between scores on measures of early treatment alliance and psychotherapy outcome. In clinical trial studies comparing multiple treatments for particular psychological disorders, patients' positive assessment of the therapeutic alliance is one of the best predictors of psychotherapy outcome (e.g., Krupnick et al. 1996). Although a significant amount of evidence points to common factors as mediators of patient change, the notion of beneficial ingredients unique to specific therapies cannot be ruled out, as occasionally such interventions can be shown to contribute to successful outcomes in patients independent of common factors (see *Psychological Treatments: Randomized Controlled Clinical Trials*).

5.3 Specific Interventions

The drive to empirically validate the efficacy of specific interventions is due, in part, to the field of psychotherapy's commitment to theoretically based approaches, as well as the influence of political and economic forces. Collectively, these factors have led to the use of comparative outcome studies, which generally avoid the ethical and methodological problems associated with no-treatment and placebo controls, while enabling researchers to study the effectiveness of particular techniques or theories by comparison.

5.3.1 Anxiety disorders. Specific behavioral treatments have been found to be effective in treating

patients with anxiety disorders. To date, two treatments, systematic exposure and applied relaxation techniques, have been shown to be efficacious in treating a variety of anxiety disorders. Exposure is a broad term that encompasses both interoceptive and situational exposure interventions. Interoceptive procedures attempt to reliably produce panic-type sensations, with the intent of severing or weakening the patient's associations between certain bodily cues and their respective reactions to panic. The technique of situational exposure, also known as in vivo exposure, refers to the patient's repeated intentional encounters with a feared stimulus. Given the focus placed on the catalyst of fear, it is not surprising that clearly identifying the stimulus that induces the patient's reactions is crucial for the success of these procedures.

Evidence suggests that the most effective treatment strategy includes identification of feared stimuli, confrontative periods of exposure sufficient in length to reduce the patient's level of anxiety, and a focus on thoughts providing self-efficacy or performance accomplishment. A general strategy employing the use of exposure techniques appears to be the only efficacious intervention for agoraphobia (DeRubeis and Crits-Christoph 1998). Similar techniques have been shown to be successful for social phobic, panic, and post-traumatic stress disorder patients as well.

The results of a form of relaxation, referred to as applied relaxation, appear to show promise in treating specific anxiety disorders. Applied relaxation entails instructing patients in the skill of progressive muscle relaxation. Patients are generally taught how to progressively relax specific muscle groups. Whereas the evidence supporting the use of applied relaxation techniques alone is mixed in the panic disorder literature, it appears there is support for treating generalized anxiety disorder with relaxation training.

Additional evidence suggests that the treatment of specific disorders is most successful when particular interventions are combined with specific theory-based therapies. For instance, panic disorder is treated most effectively when a cognitive-behavioral theory of therapy is used in tandem with behavioral techniques. Barlow et al. (1989) compared relaxation training (RT), imaginal exposure plus cognitive restructuring (E+C), and a combination of these two interventions (RT+E+C). Outcome was determined by the percentage of patients experiencing panic attacks. Results indicated that 36 percent of the WL (wait-list control) patients, 60 percent of the RT patients, 85 percent of the E+C patients, and 87 percent of the RT+E+C patients were free of panic attacks at the end of treatment.

5.3.2 Depression. Like anxiety-based disorders, there is some evidence supporting the use of specific behavioral techniques in the treatment of depression; however, in the case of depression, these procedures are no more effective than cognitive, interpersonal, and psychodynamic-interpersonal therapies. For example, behavioral activation alone, which refers to a patient scheduling activities that promote mastery as well as pleasure experiences, has been found to be as effective as cognitive therapy. Whereas behavioral activation is a procedure that makes concerted efforts to involve patients in constructive and pleasurable activities, cognitive therapy emphasizes the modification of maladaptive thoughts, beliefs, attitudes, and behaviors of individuals. Although some researchers predicted the superiority of cognitive therapy's ability to successfully treat depressive patients (Hollon and Beck 1986), the results of large well-controlled studies suggest that interpersonal and psychodynamic-interpersonal therapies are efficacious as well. In contrast to the focus that cognitive therapy places on maladaptive thoughts, beliefs, and behaviors, the interpersonal theory perceives depression as the consequence of problems in the interpersonal history of the patient.

5.3.3 Eating disorders. As eating disorders are generally characterized by harmful cognitions surrounding weight and physical image, cognitive and behavioral interventions have been the most extensively researched treatments. In particular, cognitive behavior therapy has been the focus of much research in the treatment of bulimia nervosa, showing promising results. However, it appears that interpersonal therapy provides generally equivalent results when long term follow-up findings are considered (Fairburn et al. 1993).

5.3.4 Substance abuse disorders. Reviews of the treatment of substance abuse disorders suggest that various therapy modalities have been found effective in treating patients. Therapies that provide generally equivalent results are cognitive, cognitive behavior, and psychodynamic derived therapies. Whereas patients abusing alcohol have been treated, with similar success, using psychodynamic, cognitive behavior, and psychoeducational treatment modalities, opiate dependent patients have been treated with cognitive and supportive-expressive psychodynamic therapies with generally equivalent results.

Some caution should be applied when interpreting the above findings, however, as the results mentioned are typically based on randomized clinical trials. To date, there is little evidence suggesting that these findings generalize to the clinical setting. The question of whether one technique is comparatively unique and more effective than another is a consistent topic in the literature of psychotherapy. There is currently little

evidence supporting superiority of particular interventions, although future research may note unique contributions. For instance, there is a growing body of evidence suggesting that the use of therapy manuals allow for the detection of discernable differences in the behavior of therapists, while at the same time enhancing the effects of specific therapy procedures. Perhaps these contributions will allow for more accurate comparisons in outcome studies. The search for specific techniques and common factors that distinguish successful treatment has proved to be difficult because of the myriad factors that likely exist in the interactions between therapists and clients.

6. *The Negative Effects of Psychotherapy*

Findings from the study of the progress and outcome of patients undergoing psychotherapy have resulted in both expected and unexpected conclusions. Most patients receiving treatment improve, a minority of patients remain unchanged, and contrary to the intent of psychotherapy, some patients actually worsen as a result of treatment. Independent reviews of the literature suggest that there are patients who fail to achieve beneficial gains from treatment. Although accurate estimates of patient worsening are difficult to make because the results of such incidents are not often the focus of psychotherapy outcome studies, estimates of patient worsening vary between 0 and 15 percent (Lambert and Bergin 1992). Similar to patient improvement, it appears that the phenomenon of deterioration is equivalent across theoretical orientations. In addition, treatment failures are reported in all client populations, treatment interventions, and group and family therapies.

Despite what appears to be an equitable distribution of treatment failures, specific factors have been associated with deterioration. Consistent with the influence that patient variables often have in treatment success, patient characteristics appear closely linked to treatment failure. Severity of mental illness alone is predictive of unsuccessful outcome, and severe levels of mental illness in combination with specific interventions designed to breach, challenge, and undermine entrenched coping strategies and defense mechanisms are more likely to result in treatment failure. Consequently, psychotic, borderline, schizophrenic, and bipolar patients are those patients most frequently experiencing deterioration during the course of treatment. Similar findings have been reported in the psychoanalytic literature, as patients lacking quality interpersonal relationships, low anxiety tolerance at the inception of treatment, and low motivation were more likely to worsen when treated with psychoanalysis and purely supportive therapies, whereas patients treated with supportive-expressive psychotherapy enjoyed better outcomes. Deterioration is not unique to individual therapy, however. Reviews of group psychotherapy outcome have reported a positive correlation between treatment failure and patient variables such as low participation, poor self-esteem, poor self-concept, and more significant needs for fulfillment. Additional personal characteristics associated with patient deterioration include hostility, interpersonal dysfunction, and negative expectations of treatment.

The above findings suggest that it is important to consider the type of intervention most appropriate for a particular client; however, harm is not the fault of patient and intervention alone. The results of studies investigating therapists' contributions to the phenomenon of deterioration have suggested that specific therapists' behaviors can lead to treatment failure. For instance, the results of a study investigating the relationship between therapist's emotional adjustment, empathy, directiveness, support, and credibility, and patient outcome found that a therapist's empathic abilities were the most predictive of patient outcome (Lafferty et al. 1991). Additional findings of the study suggested that effective therapists were also likely to provide more direction and support than ineffective therapists. These researchers were also interested in the similarity in values held by the therapists and patients, and found that less-effective therapists emphasized values, such as the importance placed on a life of comfort and excitement, whereas effective therapists considered intellectual values, such as reflection, as more important.

Another study suggested that patient deterioration is related to therapist countertransference reflecting disappointment, hostility, or irritation (Mohr 1995). In addition, there was some indication that client deterioration is more prevalent when clinicians underestimate their client's illness or overestimate their client's progress in treatment. Research exploring the importance of leadership styles in group therapists suggests that confrontational and aggressive group therapists tended to have more treatment failures in their groups. These therapists were observed to be more likely to insist upon immediate client disclosure, emotional expression, and change in attitude.

7. *Conclusion*

Years of clinical research have produced compelling evidence for the value of psychotherapy. Contrary to Eysenck's controversial publication regarding the efficacy of psychotherapy, outcome studies have confirmed that patients benefit in a variety of domains from treatment. According to the dose-effect literature, benefits derived from therapy transpire in relatively brief periods of time. For many clinical disorders, generally one-half of all patients return to normal functioning in 10 to 20 sessions. An additional 25 percent of patients experience substantial improvement when dosage levels are increased to 30 to 50 sessions. The benefits from treatment appear to be

durable, as the results of follow-up studies suggest that the effects of treatment endure for at least one to two years subsequent to treatment.

Not all patients experience the beneficial effects of treatment, and this finding has inspired the investigation of factors facilitating both positive and negative change. The results of such studies have indicated that a patient's level of disturbance is most predictive of outcome; however, desire to change, interpersonal skills, ego strength, psychological mindedness, and awareness of relevant problems have all been found to be related to outcome. In addition to patient variables predictive of successful outcome, there are characteristics of a therapeutic relationship that facilitate a patient's sense of trust, safety, and security. According to patients' reports, therapist empathy and a strong therapeutic alliance are most predictive of treatment outcomes.

Although the use of specific interventions has showed promising results in the case of particular disorders, in general the search for superior treatment strategies has been disappointing, and according to some researchers unnecessary, given the complexity of potential interactions between therapists, clients, and modes of therapy. Regretfully, not all patients improve from treatment, indeed, some actually worsen. Knowledge of this phenomenon is somewhat limited; however, there is evidence that suggests that treatment failure is best predicted by severity of mental illness. Furthermore, factors inhibiting the development of a strong therapeutic alliance also predict negative outcome.

Although researchers have contributed empirical evidence supporting the effectiveness of psychotherapy, they are just learning to understand why patients tend to improve. In time, it is likely that focused and determined research will eventually lead to greater insights into this important question.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Program Evaluation; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Clinical Trials; Psychotherapy and Pharmacotherapy, Combined; Psychotherapy: Clinical Practice Guidelines; Psychotherapy Process Research; Therapist–Patient Relationship

Bibliography

Barlow D H, Raske M G, Cerny J A, Klosko J S 1989 Behavioral treatment of panic disorder. *Behavior Therapy* **23**: 507–28

DeRubeis R J, Crits-Christoph P 1998 Empirically supported individual and group psychological treatments for adult mental disorders. *Journal of Consulting and Clinical Psychology* **66**: 37–52

Eysenck H J 1952 The effects of psychotherapy. *Journal of Consulting and Clinical Psychology* **16**: 319–24

Fairburn C G, Jones R, Peveler R C, Hope R A, O'Conner M 1993 Psychotherapy and bulimia nervosa: Longer-term effects of interpersonal psychotherapy, behavior therapy, and cognitive behavior therapy. *Archives of General Psychology* **50**: 419–28

Garfield S L 1994 Research on client variables in psychotherapy. In: Bergin A E, Garfield S L (eds.) *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York, pp. 190–228

Hollon S D, Beck A T 1986 Cognitive and cognitive-behavioral therapies. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York, pp. 443–82

Horvath A O, Greenberg L S (eds.) 1994*The Working Alliance: Theory, Research, and Practice*. Wiley, New York

Howard K I, Kopta S M, Krause M S, Orlinsky D E 1986 The dose-effect relationship in psychotherapy. *American Psychologist* **41**: 159–64

Krupnick J L, Sotsky S M, Simmens S, Moyer J, Elkin I, Watkins J, Pilkonis P A 1996 The role of the therapeutic alliance in psychotherapy and pharmacotherapy outcome: findings in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology* **64**: 532–9

Lafferty P, Beutler L E, Crago M 1991 Therapeutic relationship and improvement as perceived by clients and therapists. *Journal of Consulting and Clinical Psychology* **57**: 76–80

Lambert M J, Bergin A E 1992 The achievements and limitations of psychotherapy research. In: Friedman D K (eds.) *The History of Psychotherapy: A Century of Change*. American Psychological Association, Washington, DC, pp. 360–90

Lambert M J, Bergin A E 1994 The effectiveness of psychotherapy. In: Bergin A E, Garfield S L (eds.) *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York, pp. 143–89

Lambert M J, Hansen N B, Finch A E 2001 Patient-focused research: using outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology* **69**

Mohr D C 1995 Negative outcome in psychotherapy: A critical review. *Clinical Psychology Science Practice* **2**: 1–27

Nicholson R A, Berman J S 1983 Is follow-up necessary in evaluating psychotherapy? *Psychological Bulletin* **93**: 261–78

Smith M L, Glass G V, Miller T I 1980 *The Benefits of Psychotherapy*. Johns Hopkins University Press, Baltimore

Wampold B E 1997 Methodological problems in identifying efficacious psychotherapies. *Psychotherapy Research* **7**: 21–43

M. J. Lambert and E. J. Hawkins

Psychological Treatments, Empirically Supported

Changes in the US healthcare system, set in full motion during the past decade and continuing to evolve, are likely to have a profound impact on the way psychotherapy is practiced. Two primary forces are driving the changes: (a) the increasing penetration of managed care and other types of cost-conscious medical plans, and (b) the development and proliferation of clinical practice guidelines (see *Psychotherapy:*

Clinical Practice Guidelines). The most notable impact of these trends has been the requirement for greater accountability on behalf of service providers. Essentially, managed care and practice guidelines ask the question: does a treatment work and, if so, is it cost effective? Although these forces are influencing every area of healthcare, the current paper will discuss their specific impact on the practice of psychotherapy, and explore why the identification, promotion, and dissemination of empirically supported treatments (ESTs; i.e., those with controlled research data supporting their efficacy, see below for the specific criteria used to make this determination) is essential to the survival of psychological interventions as a first-line treatment for emotional disorders.

1. *Managed Care*

More than any previous event, managed care organizations (or any other system monitoring the utilization and cost of service such as health maintenance organizations, capitated contracts with providers, etc.) are reshaping the practice of psychotherapy (see *Managed Care and Psychotherapy in the United States*). In the traditional fee-for-service model, decisions about the cost and length of treatment were primarily a function of choices made by the doctor and patient. The allocation of resources (i.e., cost of psychotherapy) was of little concern to the clinician. In fact, the fee-for-service model encourages the provision of service, as more service creates more income. However, in response to the increased costs of psychotherapy, and in particular to the perceived 'endless' nature of psychotherapy, managed care organizations are pressuring clinicians to allocate decreasing (less than usual) amounts of service.

While to date, the focus of managed care organizations cost cutting has been almost entirely on limiting the number of sessions a patient receives, ultimately, in order to compete, managed care organizations will have to focus on the quality and efficacy of the psychotherapeutic interventions as well. Managed care organizations are clearly motivated to cut costs. However, they must also satisfy the consumer (i.e., patient), and payor (e.g., employer providing health benefits) and therefore have to balance cost cutting with patient outcome. Simply reducing the length of care will not accomplish this goal, because many, and perhaps the majority, of psychotherapists are trained in 'long-term' therapeutic interventions which by nature require more sessions than managed care organizations are willing to allocate. Simply reducing the length of therapies devised to take place over a longer period of time is likely to result in ineffective outcomes, leading both to consumer dissatisfaction and increased costs down the road as the severity of the disorder may increase, becoming less responsive to treatment. Thus, concern for the effectiveness of an intervention will eventually temper the managed care organizations focus on economics. Ultimately, managed care organizations will be interested in clinicians providing the 'optimal intervention … the least extensive, intensive, intrusive and costly intervention capable of successfully addressing the presenting problem' (Bennett 1992, p. 203).

2. *Practice Guidelines*

A second force that will have a significant impact on the practice of psychotherapy is the establishment of clinical practice guidelines and treatment consensus statements. As noted by Smith and Hamilton (1994), 'Guidelines are now being developed because there is a perception that inappropriate medical care is sometimes provided and that such inappropriate care has both health and economic consequences' (p. 42). It is well known that in all medical fields there is significant variability in treatments delivered to patients with various illnesses. Patients are not necessarily receiving the most effective treatment. Consequently, in order to ensure that patients uniformly receive the optimal intervention—whether it be a type of medication, surgical procedure, or psychotherapeutic intervention—clinical practice guidelines are being developed. For the most part, the development of practice guidelines relies on systematically evaluating scientific and clinical input.

One organization developing practice guidelines is the federal Agency for Healthcare Policy and Research (AHCPR). In order to create impeccable guidelines, only treatments with documented efficacy from randomized controlled trials are considered. Therefore, for an intervention to be included as a first-line treatment, research evidence attesting to its efficacy must be available. Without this evidence, one cannot conclude if a treatment is effective or ineffective, or make judgments about its relative efficacy compared with other available treatments.

The recommendations of the clinical practice guidelines are quite clear. Consider the wording from the AHCPR guideline for Depression (Depression Guideline Panel 1993). When psychotherapy is going to be selected as the sole treatment, the guideline states 'Psychotherapy should generally be time-limited, focused on current problems, and aimed at symptom resolution rather than personality change as the initial target. Since it has not been established that all forms of psychotherapy are equally effective in major depressive disorder, if one is chosen as the sole treatment, it should have been studied in randomized controlled trials' (Depression Guideline Panel 1993, p. 84). In addition to endorsing specific treatments for depression that had sufficient empirical evidence (e.g., cognitive, behavioral, interpersonal), the report goes on to state 'Long-term therapies are not currently indicated as first-line acute phase treatments' (p. 84).

Along the same line as treatment guidelines, consensus statements are likely to carry a significant

amount of weight in determining what treatments have been shown to be effective as well. For example, in 1991 a Consensus Development Conference on the Treatment of Panic Disorder was held, sponsored by the National Institute of Mental Health. As was true in the process of developing the AHCPR treatment guidelines, the merit of various treatments was determined by the available scientific evidence. The format of the conference was to let 'each camp put its best data forward' and the panelists would evaluate and weigh the evidence to formulate a consensus statement (Wolfe and Maser 1994). As in the case of treatment guidelines, specific treatments were judged to be effective for panic disorder: Several pharmacological compounds and cognitive behavioral psychotherapy (cf. Panic Consensus Statement published in Wolfe and Maser 1994). Once again, as is true of treatment guidelines, the wording of the Panic Consensus Statement as to the use of treatments not supported by empirical evidence, whether they be psychological or pharmacological, is quite clear, 'One risk of maintaining individuals in nonvalidated treatments of panic disorder is that misplaced confidence in the therapy's potential effectiveness may preclude application of more effective treatment.' The statement spells out a specific concern about the use of psychotherapies without demonstrated effectiveness for panic disorder, 'The nature of the therapeutic relationship makes it difficult for the patient to seek additional or alternate treatment.'

As one can imagine, as clinical guidelines and treatment consensus statements continue to emerge for a wide array of emotional disorders, they will have a significant impact on the way clinicians practice psychotherapy. In effect, these documents set standards of care, which if ignored, leave the clinician both ethically and legally vulnerable. Health insurance companies and managed care plans are now providing their practitioners with copies of guidelines and asking that they be followed. The implications are clear: failure to provide an empirically supported treatment from these guidelines, when one exists, may constitute malpractice in the eyes of the payor. For example, a number of states have passed, or are in the process of passing, legislation that give guidelines the 'force of law' by protecting clinicians who follow the guidelines from malpractice litigation (Barlow and Barlow 1995).

3. *Identifying, Promoting, and Disseminating Empirically Supported Treatments*

Recognizing the lack of utilization of ESTs and considering the trends noted above, the Society of Clinical Psychology (Division 12 of the American Psychological Association) established a task force to identify, promote, and assist in the dissemination of ESTs. The original Task Force on Promotion and Dissemination of Psychological Procedures (now named the Committee on Science and Practice; CSP) was created by David H. Barlow, then president of the Society. The CSP is now a standing committee in the Society of Clinical Psychology, and the work will continue into the future (the most up-to-date reports and a list of committee members are available on the Internet at www.apa.org/divisions/div12).

As noted in a recent paper summarizing the accomplishments and future directions of the CSP (Weisz et al. 2000), there are two broad goals guiding the CSP: to influence both practice and science. Ideally, the influence on practice would be to improve mental health care for the public by encouraging and supporting empirically supported practice and by fostering clinical training that emphasizes reliance on ESTs. As for science, the CSP hopes to stimulate improvement of the quality and clinical utility of treatment research so that it is most relevant to the concerns of clinicians (see Weisz et al. 2000 for specific details).

3.1 *Identifying Empirically Supported Treatments*

The first mission of the original Task Force was to define criteria to determine whether treatments qualified as ESTs. The following criteria were established (Chambless et al. 1998). For treatments to be considered by the Task Force, research must have been conducted with treatment manuals. This requirement was set because treatment manuals make for better research design, and the use of a treatment manual allows readers to know what specific procedures have been supported. For example, although many interventions fall under the rubric of behavior therapy (e.g., exposure, relaxation training, cognitive restructuring), only exposure and response prevention has been demonstrated as efficacious for obsessive-compulsive disorder. Stating that behavior therapy is effective for obsessive-compulsive disorder does not allow the reader to know what specific strategies were tested and how they were applied. Use of a treatment manual solves this dilemma.

As for whether or not the treatment qualifies as an EST, the Task Force distinguishes between two levels of EST 'well-established' and 'probably-efficacious.' EST status is determined by the results from group design and single-case design studies. With regard to group design studies, to qualify as a well-established treatment, at least two studies conducted by different investigators must qualify as supporting studies. The following criteria must be satisfied in order to qualify as a supporting study: (a1) treatment is superior to pill placebo or psychological placebo, or (a2) treatment is equivalent to an already established treatment; (b) a treatment manual was used in the study; (c) characteristics of the client sample must be clearly specified. A treatment may not meet criteria for well-established

Table 1
Examples of empirically supported treatments

Well-established treatments
Anxiety and stress:
Cognitive behavior therapy for panic disorder with and without agoraphobia
Cognitive behavior therapy for generalized anxiety disorder
Exposure treatment for agoraphobia
Exposure/guided mastery for specific phobia
Exposure and response prevention for obsessive-compulsive disorder
Stress inoculation training for coping with stressors

Depression:
Behavior therapy for depression
Cognitive therapy for depression
Interpersonal therapy for depression

Health problems:
Behavior therapy for headache
Cognitive–behavior therapy for bulimia
Multicomponent cognitive-behavior therapy for pain associated with rheumatic disease
Multicomponent cognitive-behavior therapy with relapse prevention for smoking cessation

Problems of childhood:
Behavior modification for enuresis
Parent training programs for children with oppositional behavior

Marital discord:
Behavioral marital therapy

Probably efficacious treatments
Anxiety:
Applied relaxation for panic disorder
Applied relaxation for generalized anxiety disorder
Cognitive behavior therapy for social phobia
Cognitive therapy for obsessive-compulsive disorder
Couples communication training adjunctive to exposure for agoraphobia
EMDR for civilian post-traumatic stress disorder (PTSD) Rothbaum
Exposure treatment for PTSD
Exposure treatment for social phobia
Stress inoculation training for PTSD
Relapse prevention program for obsessive-compulsive disorder
Systematic desensitization for animal phobia
Systematic desensitization for public speaking anxiety
Systematic desensitization for social anxiety

Chemical abuse and dependence:
Behavior therapy for cocaine abuse
Brief dynamic therapy for opiate dependence
Cognitive–behavioral relapse prevention therapy for cocaine dependence
Cognitive therapy for opiate dependence
Cognitive–behavior therapy for benzodiazepine withdrawal in panic disorder patients
Community reinforcement approach for alcohol dependence
Cue exposure adjunctive to inpatient treatment for alcohol dependence
Project CALM for mixed alcohol abuse and dependence
Social skills training adjunctive to inpatient treatment for alcohol dependence

Depression:
Brief dynamic therapy
Cognitive therapy for geriatric patients
Reminiscence therapy for geriatric patients
Self-control therapy
Social problem-solving therapy

Table 1
Continued

Probably efficacious treatments (continued)
Health problems:
Behavior therapy for childhood obesity
Cognitive–behavior therapy for binge eating disorder
Cognitive–behavior therapy adjunctive to physical therapy for chronic pain
Cognitive–behavior therapy for chronic low back pain
EMG biofeedback for chronic pain
Hypnosis as an adjunct to cognitive–behavior therapy for obesity
Interpersonal therapy for binge-eating disorder
Interpersonal therapy for bulimia
Multicomponent cognitive therapy for irritable bowel syndrome
Multicomponent cognitive–behavior therapy for pain of sickle cell disease
Multicomponent operant–behavioral therapy for chronic pain
Scheduled, reduced smoking adjunctive to multicomponent behavior therapy for smoking cessation
Thermal biofeedback for Raynaud's syndrome
Thermal biofeedback plus autogenic relaxation training for migraine
Marital discord:
Emotionally focused couples therapy for moderately distressed couples
Insight-oriented marital therapy
Problems of childhood:
Behavior modification of encopresis
Cognitive-behavior therapy for anxious children
Exposure for simple phobia
Family anxiety management training for anxiety disorders
Sexual dysfunction:
Hurlbert's combined treatment approach for female hypoactive sexual desire
Masters and Johnson's sex therapy for female orgasmic dysfunction
Zimmer's combined sex and marital therapy for female hypoactive sexual desire
Other:
Behavior modification for sex offenders; Marshall et al. (1991)
Dialectical behavior therapy for borderline personality disorder
Family intervention for schizophrenia
Habit reversal and control techniques
Social skills training for improving social adjustment of schizophrenic patients
Supported employment for severely mentally ill clients

Source: Chambless et al. (1998)
Citations for evidence supporting each treatment can be found in Chambless et al. (1998). Updates of the EST list will be published on the Internet at www.apa.org/divisions/div12

but still satisfy the less conservative criteria for a probably efficacious treatment: two studies show that the treatment is more effective than a waiting list control group, or only one supporting study exists, or if more than one exists, all were conducted by the same investigator. (For details concerning the definition of EST, including use of single subject designs, and resolution of conflicting data, the reader is directed to the Task Force report itself.)

Based upon these criteria, a list of ESTs was published in 1995 (Task Force on Promotion and Dissemination of Psychological Procedures 1995) and two subsequent updates have appeared (Chambless et al. 1996, 1998). As can be seen in Table 1, the Task Force has identified more than 70 ESTs (Chambless et al. 1998). It is important to note that it is not assumed that treatments without empirical evidence supporting them are ineffective, but instead, in the absence of scientific data, no conclusion can be drawn about its efficacy or lack thereof.

4. *Promotion and Dissemination*

Deciding upon criteria and identifying ESTs is only the first step in the process towards increasing their

utilization so that the general public benefits. The second step is to call attention to ESTs so that practitioners, health insurers, and the general public is aware of their efficacy (promotion), and to be sure that practitioners are able to use these treatments (dissemination).

4.1 Promotion

As noted by Barlow (1994), despite the considerable data that exist demonstrating the efficacy of psychological interventions, they seem to be taking a backseat to pharmacological approaches in clinical practice guidelines. Why? Perhaps one reason is that, unlike pharmaceutical companies that spend a significant amount of money promoting the use of their treatments to consumers and providers, no such profit motivated organization exists for psychotherapeutic interventions. As a result, many patients are not aware that effective psychological treatments exist, and thus, do not seek them out. Clearly, more effort needs to be focused educating the public about ESTs. As a first step towards this goal, the CSP is about to launch a website listing a description of ESTs for a range of disorders so that 'consumers' of mental health services can find out what types of treatments have been shown to be effective for the problem/disorder for which they are seeking treatment.

4.2 Dissemination

Identifying and promoting ESTs will not solve all the problems noted above. The next step is to ensure that the treatments are available. Unfortunately, as suspected by the CSP, a recent survey of clinical psychology doctoral programs and internship programs revealed that there is insufficient training in ESTs (Crits-Christoph et al. 1995). For example, as noted above, the two of the first-line treatments for depression listed by the AHCPR Depression Treatment Guideline were cognitive therapy and interpersonal therapy. The survey by Crits-Christoph et al. (1995) revealed that among clinical psychology internship programs—the place where clinical psychologists receive the bulk of supervised clinical experience—only 59 percent of programs provided supervision in cognitive therapy for depression, and a mere 8 percent of programs provided supervision in interpersonal psychotherapy. These are two psychotherapeutic treatments listed as first-line interventions for depression in the AHCPR Depression Treatment Guideline, thus, the lack of training in these areas is of significant concern. University doctoral programs were slightly better; 80 percent providing supervision in cognitive therapy and 16 percent in interpersonal psychotherapy. It is important to note, however, that having supervision available does not require students to receive it. As a result, these numbers do not indicate that 80 percent of students are in fact receiving training in cognitive therapy, and thus, the percentages are an overestimation of the number of students receiving supervision in that approach. For example, only 14 percent of internship programs required training in cognitive therapy for depression and only 3 percent in interpersonal therapy. Unfortunately, training in ESTs for depression is the highest. As noted in the Crits-Christoph et al. (1995) survey, only a handful of programs require training in a range of ESTs. For example, the survey revealed that only 7 percent of internship sites provide supervision in interpersonal therapy (IPT) for bulimia, 5 percent in CBT for social phobia. The numbers for doctoral programs, though higher, are still not all that promising: 21 percent offer supervision in IPT for bulimia, 24 percent in CBT for social phobia. Each of these treatments represents a first-line intervention for the respective disorder (with some of the treatments being the only empirically supported psychotherapeutic intervention), so it is not as though training in other EST approaches is occurring. Unfortunately, training in other ESTs is similar to the low numbers for depression, bulimia, and social phobia.

If training at the graduate and internship level is lacking, one can imagine how difficult it will be to disseminate these treatments to practitioners already practicing in the field. To date, no formal process exists to disseminate and train ESTs to practitioners. A perfect example of the consequence of this is that despite the strong data supporting cognitive behavioral treatment for panic disorder, a recent survey found that the majority of patients with panic disorder receiving psychological treatment receive dynamic psychotherapy (which has not been empirically supported as a first-line treatment) (Goisman et al. 1999). Indeed, as noted by Calhoun et al. (1998) in a recent paper discussing training in ESTs, reaching professionals out in the field is the most problematic area. With a rapidly changing data base, clinicians trained 10 years ago are unlikely to be up-to-date with the newer, empirically supported psychotherapies. Indeed, the majority of references supporting ESTs have appeared in the past 10 years (Calhoun et al. 1998). How can clinicians become motivated to learn ESTs? Even if they are motivated, how can treatments be disseminated in a way to provide high integrity supervision to insure they are properly trained? The CSP plans to devote a considerable amount of effort to develop a model that would facilitate training professionals in ESTs, because as discussed above, this is an essential activity for the survival of psychotherapy as a viable first-line treatment for emotional disorders. For those interested, a list of manuals outlining ESTs is available (cf. Sanderson and Woody 1995; available at www.apa.org/divisions/div12).

5. Criticism of the EST Movement

The movement towards ESTs are not welcomed by everyone within our profession, and indeed, several authors have raised important concerns and criticisms (e.g., Garfield 1998, Kovacs 1995, Nathan 1998) regarding the identification of 'validated' therapies. Although space does not permit a discussion of the specific issues here, in fact, the CSP agrees that there are difficulties and risks in establishing so-called psychotherapy guidelines—or any treatment guidelines for that matter (cf. Pincus 1994 concerning the development of psychiatry guidelines)—yet the CSP and other professional organizations believe the benefits clearly outweigh the risks (Pincus 1994, Sanderson 1995), and thus, will continue the work outline above (cf. Weisz et al. 2000 for a discussion of the future directions of the CSP).

See also: Adolescence, Psychiatry of; Adolescent Development, Theories of; Adolescent Vulnerability and Psychological Interventions; Antisocial Behavior in Childhood and Adolescence; Clinical Treatment Outcome Research: Control and Comparison Groups; Conduct Disorder; Managed Care and Psychotherapy in the United States; Psychological Treatment, Effectiveness of; Psychological Treatments: Randomized Controlled Clinical Trials; Psychotherapy: Case Study; Psychotherapy: Clinical Practice Guidelines; Social Competence: Childhood and Adolescence

Bibliography

American Psychiatric Association 1993 Practice guidelines for major depressive disorder in adults. *American Journal of Psychiatry* **150**(4): 1–26 (Suppl. 5.)

Barlow D H 1994 Psychological interventions in the era of managed care. *Clinical Psychology: Science and Practice* **1**: 109–22

Barlow D H, Barlow D G 1995 Guidelines and empirically validated treatments: Ships passing in the night? *Behavioral Healthcare Tomorrow* **4**: 25–9

Bennett M J 1992 The managed care setting as a framework for clinical practice. In: Feldman J L, Fitzpatrick R J (eds.) *Managed Mental Healthcare: Administrative and Clinical Issues*. American Psychiatric Press, Washington, DC, pp. 203–18

Calhoun K S, Moras K, Pilkonis P A, Rehm L P 1998 Empirically supported treatments: Implications for training. *Journal of Consulting and Clinical Psychology* **66**: 151–62

Chambless D L, Sanderson W C, Shoham V, Johnson S B, Pope K, Crits-Christoph P, Baker M, Johnson B, Woody S R, Sue S, Beutler L, Williams D, McCurry S 1996 Empirically validated therapies: A project of the division of clinical psychology, American psychological association, task force on psychological interventions. *The Clinical Psychologist* **49**: 5–15

Chambless D L, Baker M J, Baucom D, Beutler L E, Calhoun K S, Crits-Christoph P, Daiuto A, DeRubeis R, Detweiler J, Haaga D A F, Johnson S B, McCurry S, Mueser K T, Pope K S, Sanderson W C, Shoham V, Stickle T, Williams D A, Woody S R 1998 Update on empirically validated therapies: II. *The Clinical Psychologist* **51**: 3–16

Crits-Christoph P, Frank E, Chambless D L, Brody C, Karp J F 1995 Training in empirically validated therapies: What are clinical psychology students learning? *Professional Psychology: Research and Practice* **26**: 514–22

Depression Guideline Panel 1993 *Depression in Primary Care: Volume 2. Treatment of Major Depression. Clinical Practice Guideline, Number 5*. Rockville, MD (Department of Health and Human Services, Public Health Service, Agency for Healthcare Policy and Research. AHCPR Publication No. 93-0551). US Government Printing Office, Washington, DC

Garfield S 1998 Some comments on empirically supported treatments. *Journal of Consulting and Clinical Psychology* **66**: 121–5

Goisman R, Warshaw M, Keller M B 1999 Psychosocial treatment prescriptions for generalized anxiety disorder, panic disorder, and social phobia. *American Journal of Psychiatry* **156**: 1819–21

Kovacs A 1995 We have met the enemy and he is us! *The Independent Practitioner* **15**(3): 135–7

Nathan P E 1998 Practice guidelines: not yet ideal. *American Psychologist* **53**: 290–9

Pincus H A 1994 Treatment guidelines: Risks are outweighed by the benefits. *Behavioral Healthcare Tomorrow* **4**: 40–45

Sanderson W C 1995 Can psychological interventions meet the new demands of health care? *American Journal of Managed Care* **1**: 93–8

Sanderson W C, Woody S 1995 Manuals for empirically validated treatments. A project of the division of clinical psychology, American psychological association, task force on psychological interventions. *The Clinical Psychologist* **48**: 7–11

Smith G R, Hamilton G E 1994 Treatment guidelines: Provider involvement is critical. *Behavioral Healthcare Tomorrow* **4**: 40–45

Task force on promotion and dissemination of psychological procedures 1995 Training in and dissemination of empirically-validated psychological treatments. *The Clinical Psychologist* **48**: 3–23

Weisz J R, Hawley K M, Pilkonis P A, Woody S R, Follette W C 2000 Stressing the other three Rs in search of empirically-supported treatments: Review procedures, research quality, relevance to practice and public interest. *Clinical Psychology: Science and Practice* **7**: 243–58

Wolfe B E, Maser J D (eds.) 1994 *Treatment of Panic Disorder: A Consensus Statement*. American Psychiatric Association Press, Washington, DC

W. C. Sanderson

Psychological Treatments: Randomized Controlled Clinical Trials

The randomized controlled trial is one method of ascertaining the effectiveness of a treatment. In its simplest form, a controlled trial consists of the recruitment of a group of participants with particular characteristics and their allocation by chance to one of two treatment conditions. These conditions should

differ only in the treatment procedure A whose effects are being studied. Thus one condition will include A, and the other condition will not. Randomization is used to ensure that events other than treatment that might affect outcome will be equally balanced across the two treatment conditions. If measures of key features of the disorder, for example, the principal symptoms, are taken before and after treatment, then it is possible to test whether the null hypothesis of no effect of treatment A is true or not. While more complex experimental designs are often used, and the choice of the control condition is often difficult, all controlled clinical trials adhere to this basic conceptual form.

1. *Development of Controlled Clinical Trials in Psychology*

Parallel developments in two clinical disciplines, medicine and psychology, beginning in the 1930s, edged both fields toward the use of controlled trials. In an early study spanning both medicine and psychology (Chappell and Stephenson 1936), the authors demonstrated that an educational group program resulted in greater improvement and less relapse in a group of peptic ulcer patients than the control group, both groups having received the same medical care. Another study, the Cambridge-Somerville Youth Project, begun in the late 1930s and continuing with a long-term follow-up into the 1950s, was one of the earliest trials of a psychological treatment (Teuber and Powers 1953). This large-scale, well-designed study of 650 boys thought to be at high risk of developing delinquency was aimed at evaluating a prevention program. The outcome, as these individuals were followed over time, demonstrated that the treatment group fared less well than the control group in terms of a range of delinquent behaviors. Such negative results, while disappointing, are important because they allow ineffective approaches to treatment to be abandoned.

These early controlled trials were the exception rather than the rule. Most studies of the effectiveness of psychological therapies consisted of treating a series of patients without a control group. Hence, it was impossible to be sure whether the results were due to spontaneous recovery or to the treatment, let alone specific aspects of the treatment rather than a nonspecific (placebo) effect. Two developments catalyzed the use of the controlled trial. Hans Eysenck, Professor of Psychology at London University, reviewed the psychotherapy treatment literature in 1952 and again in 1966 (see Eysenck 1952, 1966). On both occasions Eysenck concluded that there was little evidence that psychotherapy was effective, that is, that the results did not exceed the expected rate of spontaneous recovery. While this opinion resulted in a vigorous debate and defense of psychotherapy, including the addition of a few treatment trials that demonstrated some success for various forms of psychotherapy that Eysenck had not considered, Eysenck's conclusions were widely accepted, reaffirming the need for well-conducted controlled trials. Some, however, argued that the effects of psychological therapies, because they were aimed at 'inner' essentially unmeasurable changes, could never be adequately assessed in a controlled trial. This problem was largely overcome by the next development.

The second, and most important, catalyst for the use of controlled trials of psychological treatments, was the emergence of behavior therapy (see *Behavior Analysis, Applied*; *Behavior Therapy: Psychological Perspectives* for further details). Eysenck 1966, p. 39) had concluded that 'With the single exception of the psychotherapeutic methods based on learning theory, results of published research … suggest that the therapeutic effects of psychotherapy are small or nonexistent … .' Because behavior therapy was based on learning theory, the focus of treatment was directly on the problem behaviors comprising the clinical syndrome. This allowed for the development of objective measures of treatment outcome (see *Behavioral Assessment* for a fuller discussion of objective evaluation methodology). In addition, because treatment was scientifically based, its practitioners were more likely to use scientific methods, including the randomized controlled trial, to assess the effectiveness of these new treatments.

2. *The Flow of Research*

Although the controlled trial is the definitive method of ascertaining the effectiveness of a treatment, it is not the only method used in the research process aimed at the development of a new therapy. The first step in the development of a new therapy, the initial formulation of its conceptual basis and procedures, may stem from basic psychological research or theory, and/or from clinical observation. The first step in the development of a new therapy is usually an uncontrolled observational study, namely treatment of a series of patients. If a sufficient number of patients respond favorably to the new treatment, then the next step is to document the treatment procedures in a manual so that it can be replicated by others. The first controlled study may compare treatment to no treatment, a step that is relatively easy to carry out with a small number of patients, providing information as to whether the new treatment is better than doing nothing, and allowing the results to be roughly compared with those of standard therapy for the same condition, if such a therapy exists. One objective method of comparing the effects of different therapies is to calculate the effect size (the change from pretreatment to post-treatment in a group of patients divided by the standard

deviation of the change). An effect size of 0.3 would be regarded as small, one of 0.5 as moderate, and 1.0 as large.

The next step is to ascertain whether the promising new treatment has specific effects over and above what is usually termed the placebo or nonspecific effect (see Sect. 3). Hence, it is important to compare the new treatment to a second treatment containing nonspecific features common to all therapies, but not containing the specific procedures aimed at ameliorating the behaviors being studied. Single case controlled studies may also be useful at this stage of development to more precisely delineate the active components of treatment (see *Single-case Experimental Designs in Clinical Settings*).

Once the new treatment has been demonstrated to have specific effects on a particular disorder, the next step will probably consist of a comparison with the standard therapy for the condition. At this stage related studies are likely to be carried out at different centers. The results of such studies can be compared using a procedure known as meta-analysis in which effect sizes are compared across studies. It may be clear from the results of such an analysis that the new treatment is better or worse than standard treatment, or that it is useful for a particular subgroup of individuals, or that it has advantages such as being less costly but similarly effective. On the other hand the results of such comparisons may be unclear. Some studies may demonstrate positive effects while others may show no effect. Such conflicting results are often due to the initial studies having too few patients enrolled to determine accurately the effectiveness of the new treatment as compared with other treatments. In this case a study involving several clinical sites may be needed to recruit a sufficient number of patients in a short enough time, to obtain more definitive results.

2.1 Multisite Studies

Multisite studies are necessarily more complex to carry out than a study at a single site. They are also more expensive than single-site studies and for that reason should only be entered upon when the preliminary work described above has been completed, and the need for such a study has been precisely formulated. The investigators involved must agree to a common experimental protocol. The treatment(s) being studied must be well documented in a treatment manual and therapists at each of the sites trained in the delivery of the treatments. Moreover, the therapists should be carefully supervised throughout the study to ensure that they are delivering the treatment(s) in accord with the protocol. Tape recordings of treatment sessions may also be audited to ensure fidelity to the treatment. In addition, the assessments used must be carried out in a standard manner and done comparably between sites. This may mean that assessors will also be trained and supervised across sites in the particular assessment methods. One advantage of a multisite over a single-site study is that the effects of treatment are examined in different sites that may have unique characteristics in terms of patient population, therapist characteristics, and so on. Hence, generalizability of the results of treatment to the clinical world is greater than in the single-site study.

Another reason to conduct a multisite study is when treatment for a relatively rare disorder is being studied. If the study were to be done at only one site it might take many years before sufficient participants entered the study. A multisite study would provide more rapid results. Another reason for considering a multisite study is when the question of interest shifts to what treatment or treatments might benefit those who do not respond to an initial effective treatment. Because only treatment failures will be entered into such a study, a large number of individuals must be treated in order to generate a sufficiently large sample for the study.

2.2 Effectiveness Research

As noted above, a distinction has arisen between efficacy and effectiveness controlled trials. Trials to establish efficacy are considered to be carried out in refined settings with carefully chosen patients and expert therapists. Effectiveness studies on the other hand are considered to be carried out in clinical practice settings with 'real life' patients and possibly less expert therapists. Some would consider efficacy studies to be of limited value in assessing the outcome of psychological therapies in contrast to effectiveness studies. However, if the flow of research described above is taken into consideration it is evident that efficacy and effectiveness studies lie on a continuum rather than being different types of studies, and that research along the entire continuum is needed. Moreover, the scientific questions being studied differ along the continuum that constitutes the flow of research.

Once a psychological treatment has been established as useful for a particular disorder in a number of single-site studies (efficacy studies) and, if needed, a multisite study, research as to how to transfer the treatment to other settings may be needed (effectiveness studies). This step in the development of a treatment may involve controlled studies of modifications of the treatment. Such modifications may include simplification of the treatment for wider dissemination to less skilled therapists and investigation of the necessary clinic environment to support the therapy. For example, in most controlled trials therapists are carefully supervised, and the question arises whether such supervision is essential to the effectiveness of the therapy. Additionally, the type of

patient encountered in different clinical settings may differ, again requiring modification of the therapeutic procedures. Once these questions have been resolved, often using controlled trials, an effectiveness study can be designed in one or more clinical settings using the staff of such clinics as the therapists, and enrolling all the patients with the particular disorder into the study as they appear in the clinic. A control condition in such cases might be the usual care that participants would receive in such a clinic.

3. Essential Elements of a Controlled Trial

While the basic design of two groups differing only in A (the components of a new treatment) suffices for early stages of research, other questions will need to be addressed. Among the most problematic is how to control for nonspecific elements of psychological treatments. Treatments can be regarded as a package containing both specific procedures aimed at the particular problem, together with elements common to all therapies, such as therapist attention, and therapeutic instructions that include the rationale for therapy together with engendering an expectancy that treatment is likely to be successful.

Research has demonstrated that nonspecific elements lead to significant behavior change. Hence, it is important to know to what extent procedures specific to the therapy contribute to a successful outcome over and above nonspecific elements. This raises problems for psychological therapies because there is no placebo pill. This problem has been solved in several ways. The first solution is to use a nonspecific therapy, that is, one without the specific therapeutic procedures being examined but which also has face validity as a therapy with which to compare the active treatment. Some controlled trials may compare a psychotherapeutic procedure with medication and with a pill placebo as a control condition, thus solving the nonspecific effect question in a different manner. Finally, it is possible to compare a new therapy with a standard therapy for the condition which has already been demonstrated more successful than a nonspecific therapy.

3.1 The Use of Treatment Manuals

The reports of early studies of psychological therapies usually described the treatment in a paragraph or two. Such descriptions are obviously insufficient to allow other investigators or therapists to provide the same treatment to their patients. The treatment manual was developed to solve this problem. Such manuals, now regarded as an essential element of a controlled trial of psychological therapies, describe the treatment procedures in sufficient detail for others to replicate them. Therapists are carefully trained in the procedures, which the manual facilitates.

There is some controversy regarding the clinical use of treatment manuals, which have now become a common method of disseminating new therapies to clinicians. Some argue that manuals lock the therapist into a rigid method of providing treatment, thus eliminating the opportunity to individualize the treatment. Others suggest that manuals provide essential guidelines for therapists and that the experienced therapist will be able to individualize treatment to the extent necessary. There is relatively little experimental data on the effectiveness of the use of manuals in the clinic; however existing data suggests that manuals are likely to enhance therapy probably by ensuring that the therapist stays on track during the treatment process. See *Clinical Psychology: Manual-based Treatment* for a fuller discussion of this important issue.

3.2 Controlling the Fidelity of Treatment

A further essential element of a clinical trial is a method to ensure that the therapy provided is faithful to the treatment manual. This is usually ensured in two ways. First, therapists are trained in the use of the manual and practice the treatment on several patients before the controlled trial begins. During this period the therapists are supervised by an expert. Second, once the trial begins, therapy sessions are audiotaped, and a random sample of such tapes is audited. The auditor will use a checklist to ensure that all the elements of treatment are being provided, and that no foreign elements have crept in. Regular feedback is provided to the therapists, and the degree of therapist compliance with the treatment protocol should be reported in the paper describing the study.

4. Outcomes of Controlled Trials

There are several potential outcomes from a randomized clinical trial. First and foremost is whether the treatment is effective as compared to the control group; second, whether there are specific predictors as to which patients do best with the treatment under study; third, whether variables that mediate the outcome of treatment can be identified. Such mediators may throw light on how a given therapy works.

4.1 Assessment of the Primary Outcome

The primary outcome for a controlled trial is the main indicator or indicators that characterize the syndrome.

For example, for panic disorder, the number and severity of panic attacks would comprise the main outcome; for depression a measure of the severity of depression; for bulimia nervosa, a measure of binge eating and purging. Because these measures test the primary hypothesis addressed, that is, the null effect of the treatment in question as compared with a control group, the study should stand or fall depending on whether the primary outcome is positive or not. If the primary outcome is positive then secondary measures—usually associated psychopathology—are also examined for the effect of treatment. Because some symptoms such as mild anxiety or depression are normative in a community, the extent to which the treated population now approximates community norms may also be reported.

4.2 Examination of Predictors

It is usual in a controlled trial to assess a series of variables that are hypothesized to affect the outcome of treatment. Examples of potential predictors include: the duration and severity of the problem, the age and social class of the patient, and associated psychopathology. If predictors of sufficient strength are found they may be used in future trials to select patients most likely to succeed with the treatment, thus refining the indications for treatment. In general, the search for such predictors has not been very successful. Usually such predictors are sufficient for a general description of the patient most likely to improve, but not specific enough to identify such patients before treatment begins. An alternative method of prediction is to examine the progress of patients during treatment. In many psychological treatments patients obtain significant benefit early in therapy. Patients who do not demonstrate this early response tend to do poorly. Hence, it may be possible to identify those most likely to do badly with a particular psychotherapy early in treatment and to use a different treatment approach with such patients.

4.3 Identifying Treatment Mediators

A mediator is a variable that is responsible for an effective treatment outcome. Hence, it is a result of a therapeutic procedure that occurs prior to the benefits of treatment and that is significantly related to the outcome. Rather than measuring such variables before treatment begins, as is the case with predictors, potential mediators are measured early in the course of treatment. Potential mediators are usually selected based on precise hypotheses about how the treatment works. The identification of mediators can throw important light on the mechanisms underlying the effectiveness of a given psychological treatment and may provide insight as to how to improve the treatment. See also *Psychotherapy Process Research* in which interactions between therapist and patient, and their effect on outcome, form the question of interest.

5. Are Controlled Psychological Treatment Trials Useful?

While controversy over the use of controlled trials to evaluate psychological treatments continues, the argument has shifted as to whether such trials can reliably inform clinical work and social policy. Some argue that uncontrolled trials conducted in 'real' clinical settings with real patients would be more informative. This argument ignores the necessity for a sequenced flow of research in the development of a new psychological therapy. Moreover, without a control group, it is impossible to determine whether it is the treatment or some other variable that is responsible for the outcome. Despite this ongoing controversy, the controlled trial has become the accepted method of evaluating the effectiveness of psychotherapy. Most individuals would prefer that their medical treatment was informed by controlled therapeutic trials; similarly, consumers of psychotherapy should demand that these procedures are well tested.

6. Future Developments

It is likely that the use of randomized controlled trials to evaluate the effectiveness of psychological treatments will accelerate. At present a relatively large number of treatment procedures have been identified as effective for specific psychological problems (see *Psychological Treatment, Effectiveness of*; *Psychological Treatments, Empirically Supported* for a detailed review of such treatments). Many, but not all, such therapies fall under the rubric of behavior therapy. These research efforts have now generated a small, but growing number of multisite trials in which the treatment is put to a more definitive test, compared with another effective psychotherapy, or with medication either separately or in combination with psychotherapy. As is the case in medical research, it would be expected that this trend will continue, and that multisite trials will be increasingly used in psychotherapy research.

Because no one psychotherapy is effective with all those suffering from a particular psychological disorder, research is beginning to examine the efficacy of a second treatment approach for those who fail initial therapy. Moreover, there is much interest in the effectiveness of therapist-supported bibliotherapy.

Such developments may ultimately lead to algorithms guiding the manner in which psychological treatments should be sequenced for the treatment of a given condition. For example, treatment might begin with the least costly treatment followed by increasingly expensive (and probably more effective) treatments for those who fail one or more initial treatment.

In the somewhat further future the biological mechanisms through which psychotherapies work will become a focus of research, ultimately bringing together psychological and biological mechanisms underlying behavior change.

See also: Clinical Treatment Outcome Research: Control and Comparison Groups; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Single-case Experimental Designs in Clinical Settings

Bibliography

Beutler L E, Crago M 1991 *An International Review of Programmatic Studies*. American Psychological Association, Washington, DC

Campbell D T, Stanley J C 1963 *Experimental and Quasi-Experimental Designs for Research*. McNally, Chicago, IL

Chappell M N, Stephenson I T 1936 Group psychological training in some organic conditions. *Mental Hygiene*, **20**: 588–97

Eysenck H J 1952 The effects of psychotherapy: An evaluation. *Journal of Consulting and Clinical Psychology* **16**: 319–24

Eysenck H J 1966 *The Effects of Psychotherapy*. International Science Press, New York

Goldfried M R, Wolfe B E 1998 Toward a more clinically valid approach to therapy research. *Journal of Consulting and Clinical Psychology* **66**: 143–50

Kazdin A E 1998 *Research Design in Clinical Psychology*, 3rd edn. Allyn and Bacon, Boston, MA

Kraemer H C, Pruyn J P 1990 The evaluation of different approaches to randomized clinical trials. *Archives of General Psychiatry* **47**: 1163–9

Persons J B, Silberschatz G 1998 Are results of randomized clinical trials useful to psychotherapists? *Journal of Consulting and Clinical Psychology* **66**: 126–35

Russell R L, Orlinsky D E 1996 Psychotherapy research in historical perspective: Implications for mental health care policy. *Archives of General Psychiatry* **53**: 708–15

Teuber N L, Powers E 1953 Evaluating therapy in a delinquency prevention program. *Proceedings of the Association for Research in Nervous & Mental Disease* **31**: 138–47

Ullman L P, Krasner L 1965 *Case Studies in Behavior Modification*. Holt, Rinehart and Winston, New York

Wilson G T 1998 The clinical utility of randomized controlled trials. *International Journal of Eating Disorders* **24**: 13–29

W. S. Agras

Psychologism, History of

The history of psychologism is the history of an accusation and of a threat. To label an author an advocate of psychologism is to claim that he or she uses psychological concepts and ideas in the wrong place. The accusation first became popular in the German-speaking world with the rise of experimental psychology in the second half of the nineteenth century. The most famous episode in the history of the accusation is the German 'psychologism controversy' at the beginning of the twentieth century. It will be the central topic of this article. It is widely accepted that this controversy played an important role in shaping the character of Western philosophy in general and analytic philosophy in particular. It also influenced subsequent debates over 'sociologism.'

1. The Psychologism Controversy in Early Twentieth-century Germany

From the 1880s onwards, a number of German logicians and epistemologists started to attack some of their philosophical colleagues for presenting and endorsing the following kinds of 'psychologistic' arguments:

Logic studies specific 'laws of thought.' Psychology is the study of all kinds of laws of thought. *Therefore*, logic is a part of psychology.

Normative disciplines must be based upon descriptive-explanatory disciplines. Logic is a normative discipline with respect to human thinking. Psychology is *the* descriptive-explanatory discipline with respect to human thinking. *Therefore*, logic must be based upon psychology.

Logic is concerned with judgments, concepts, and inferences. Judgements, concepts, and inferences are human mental entities. *Therefore*, logic is a theory about human mental entities.

The touchstone of logical truth is self-evidence. Self-evidence is a human mental experience. *Therefore*, logic is about a human mental experience (and thus a part of psychology).

We cannot conceive of a radically different logic. The limits of conceivability are our mental limits. *Therefore*, logic is relative to the human species, and thus a subfield of human psychology.

In some cases there was considerable dispute concerning the question who actually held these views. The first argument can be found in Theodor Lipps and Gerardus Heymans, the second in Wilhelm Wundt, the third in Wilhelm Jerusalem and Christoph Sigwart, the fourth in Theodor Elsenhans, and the fifth in Benno Erdmann. All the accused had published widely on psychological issues, and some of them openly advocated treating psychology as the foundation of philosophy.

Arguments against these and other allegedly psychologistic authors were provided by many writers, amongst them the logician and mathematician Gottlob Frege, the phenomenologist Edmund Husserl, and the Neokantians Hermann Cohen, Paul Natorp, Heinrich Rickert, and Wilhelm Windelband. Of these and the many other critics, Husserl was undoubtedly the most successful. Husserl's arguments appeared in the *Prolegomena* to his *Logische Untersuchungen*. The following arguments figured centrally in the debate:

If logical rules were based upon psychological laws, then all logical rules would have to be as vague as the underlying psychological laws. *But*, not all logical rules are vague. *Therefore*, not all logical rules are based upon psychological laws.

If laws of logic were psychological laws, then they could not be known *a priori*. They would be more or less probable rather than certain, and justified only by reference to experience. *But*, laws of logic are known *a priori*; they are justified by apodictic self-evidence, and certain rather than probable. And *therefore*, laws of logic are not psychological.

If logical laws were psychological laws, they would refer to psychological entities. *But*, logical laws do not refer to psychological entities. And *therefore*, logical laws are not psychological laws.

All forms of psychologism imply relativism. Psychologistic thinkers take logic to be true only for the human species. But this contradicts the meaning of truth. Truths are unchanging, and independent of the human mind.

A good deal of the psychologism controversy concerned the validity and originality of Husserl's arguments. Several commentators maintained that Husserl's first three arguments all begged the question. Others challenged Husserl's claim that the laws of logic did not imply matters of fact. Husserl's critics also denied that truths were eternal or independent of human thought.

The debate over psychologism was vigorous. This can be seen, for instance, from the fact that more than 130 philosophers were accused of advocating psychologism and that there were as many symptoms and definitions of psychologism as there were philosophical schools (Kusch 1995). It was not merely the relationship between logic and psychology that was at issue, but the relationship between psychology and philosophy as a whole. This was a pressing issue not least because the practitioners of the then new (experimental) psychology worked in philosophy departments, and because between 1873 and 1913 the number of full professorships held by these 'psychologists' increased from one to ten. Of the 39 full professorships in philosophy in 1892, practitioners of experimental psychology held three; of the 42 full professorships in 1900 they occupied six; while of the 44 full professorships in 1913 they had already gained 10. Practitioners and advocates of the new psychology sometimes claimed that psychology provided the key to real progress in philosophy or that it was the heart of philosophy (Ash 1995, Kusch 1995, 1999).

This last-mentioned contention was opposed by a quickly growing group of 'pure' philosophers—led by the Neokantians, and by Eucken, Dilthey, and Husserl. They deemed it harmful to philosophy and its progress that experimental psychology should be regarded as part and parcel of philosophy. To put it another way, these philosophers believed that the role of the philosopher, as suggested by the academic success of experimental psychology, needed purification.

The prolonged debate between the advocates and opponents of psychology as a philosophical discipline was eventually brought to an end by World War I. The war brought about an atmosphere in which attacks on one's colleagues were regarded as utterly inappropriate. Moreover, it also led to a clear division of labor between pure philosophers and psychologists: while pure philosophy concentrated on the ideological task of celebrating the German 'genius of war,' experimental psychology focused on the training and testing of soldiers. After the war, both academic pure philosophy and experimental psychology had to cope with, and accommodate to, an intellectual environment that was hostile to science, rationality, and systematic knowledge. The project of a scientifically minded philosophy with experimental psychology as its central pillar quickly lost support. Advocates and practitioners of experimental psychology therefore had to find new ways of justifying their work. Many of them continued their wartime involvement with applied psychology (Kusch 1995).

2. *The Influence of the Psychologism Controversy*

2.1 *Psychologism and Sociologism*

The way the psychologism controversy was fought by the defenders of philosophical purity became an important model for subsequent disputes between advocates and opponents of 'naturalism' (the view that philosophical questions can be answered using the results of the natural and social sciences). Many later arguments against 'sociologism' or 'biologism' closely mirror Frege's and Husserl's ideas.

For instance, in the German-speaking world the psychologism controversy of the first two decades of the twentieth century was followed by the dispute over the nature and philosophical aspirations of the sociology of knowledge. Accusations of 'sociologism' surfaced quickly and 'sociologism' was seen by many as the 'successor' to psychologism. Again many philosophers felt that a special science—this time sociology—was trying to occupy philosophical terrain.

Critics of the sociology of knowledge accused it of both psychologism and sociologism, and in so doing, they drew on Husserl's work in particular. Defenders of the sociology of knowledge tried a similar feat in presenting their work as a bulwark against psychologism (Meja and Stehr 1990). This idea has become a general feature of the field to this day (Bloor 1983). The close link between the two 'isms' has been transmitted to subsequent Anglophone thinking mainly through the work of Sir Karl Popper.

2.2 Psychologism and Philosophy

German-speaking philosophy of the turn of the century exerted a powerful and lasting influence on philosophy in many other countries. Frege's and Husserl's (1970) antipsychologism thus became the dominant view in Western philosophy for much of the twentieth century. This is not to deny that other countries also had their own debates over the relation between psychology and philosophy. In the UK, for instance, George Frederick Stout and James Ward emphasized the philosophical significance of psychology; British Idealists like Francis Herbert Bradley and Brand Blanshard were sympathetic to psychology even though they denied its philosophical credentials; and realists like John Cook Wilson, Bertrand Russell, Herbert Joseph, and Herbert Prichard insisted on a separation. For Prichard, psychology was not even a 'proper science' (Passmore 1994).

The institutional separation of psychology from philosophy eventually removed one important source of conflict. Once institutional separation had been achieved, 'pure' philosophers and 'impure' philosophers (i.e., psychologists) were no longer in direct competition over academic chairs. On the plane of ideas, however, the psychological challenge has remained. Throughout the twentieth century, psychology exerted a powerful pull on philosophers who wished to put philosophy on the sure path of a natural science. Willard Van Orman Quine explicitly called for a return to psychologism in the 1960s. Such calls have kept the psychologism accusation alive. Amongst the accused one can find the names of Rudolf Carnap, Michael Dummett, Peter Geach, Nelson Goodman, Thomas Kuhn, John McDowell, Karl Popper, Wifrid Sellars, and Ludwig Wittgenstein.

It is now common amongst historians of philosophy and psychology to lament the outcome of the psychologism controversy. The separation of philosophy from psychology delayed the development of research programs like 'cognitive science' that combine psychological experimentation with philosophical aspiration.

See also: Determinism: Social and Economic; Experimentation in Psychology, History of; Husserl, Edmund (1859–1938); Knowledge, Sociology of; Logic and Cognition, Psychology of; Psychology: Historical and Cultural Perspectives; Psychology: Overview

Bibliography

Ash M G 1995 *Gestalt Psychology in German Culture 1890–1967: Holism and the Quest for Objectivity*. Cambridge University Press, Cambridge, UK

Bloor D 1983 *Wittgenstein: A Social Theory of Knowledge*. Columbia University Press, New York

Frisby D 1983 *The Alienated Mind: The Sociology of Knowledge in Germany 1918–33*. Heinemann, London

Haaparanta L (ed.) 1994 *Mind, Meaning and Mathematics: Essays on the Philosophical Views of Husserl and Frege*. Kluwer, Dordrecht, The Netherlands

Husserl E 1970 *Logical Investigations*, Vol. I, trans. Findlay J N. Routledge, London

Kusch M 1995 *Psychologism: A Case Study in the Sociology of Philosophical Knowledge*. Routledge, London

Kusch M 1999 *Psychological Knowledge: A Social History and Philosophy*. Routledge, London

Meja V, Stehr N 1990 *Knowledge and Politics: The Sociology of Knowledge Dispute*. Routledge, London

Notturno M A 1985 *Objectivity, Rationality and the Third Realm: Justification and the Grounds of Psychologism: A Study of Frege and Popper*. Nijhoff, Dordrecht, The Netherlands

Passmore J 1994 *A Hundred Years of Philosophy*, 2nd edn. Penguin, London

Popper K 1945 *The Open Society and Its Enemies, Vol. II, The High Tide of Prophecy: Hegel, Marx, and the Aftermath*. London: Routledge

M. Kusch

Psychology and Economics

> Economics without psychology has not succeeded in explaining important economic processes and 'psychology without economics' has no chance of explaining some of the most common aspects of human behavior (Katona 1951).

Psychology as the science of human behavior naturally also encompasses the study of economic behavior. From this perspective, there is a considerable overlap between the two disciplines of psychology and economics, as manifested in such subdisciplines as economic psychology, behavioral or psychological economics, economic cognition, and other hybrids. Despite this multiple overlap, both disciplines use rather different approaches, rest on entirely different

assumptions, and focus on different phenomena of interest. This article first portrays the points of contact and difference between psychology and economics. A generalized exchange assumption is then identified as a theoretical interface connecting both disciplines. The implications of this basic exchange assumption afford an organizing framework for a review of empirical research findings.

1. Comparison of the Two Disciplines

1.1 Points of Contact

As disciplines psychology and economics are multiply interrelated. Both serve each other as explanans and explanandum. First and foremost, psychological theories and knowledge have been amply applied to explain problems genuinely studied in economics. Vice versa, economic theory has also been applied to some fields located traditionally more within the realm of psychologists rather than economists. In most general terms, one might say that both disciplines meet in the area of decision making.

Psychological theories and knowledge help to explain how scarce resources are handled so as to satisfy needs. This approach focuses on psychological mechanisms and processes underlying consumption, investment, saving, and other economic behavior. Although early classics of economic theory explicitly or implicitly acknowledged psychological variables, there was no attempt to incorporate the psychological research systematically. Only Katona (1951) recognized that purely economic variables such as income insufficiently explain economic behavior such as consumption. He assumed that psychological variables such as affect, attitudes, expectations, and others intervene between economic stimuli and consumer reactions. Until today, the study of consumer behavior including marketing and advertising (see *Consumer Psychology*; *Psychology and Marketing*; *Advertising, Psychology of*) has remained by far the most successful and most pervasive application of psychology to economic behavior. This is particularly so in the area of consumer information processing (CIP), which exclusively relies on psychological theory. CIP deals with how consumers attend to, encode, store, and retrieve product relevant information and how they integrate this information to arrive at a judgment or choice. Examples of other interdisciplinary applications are fiscal psychology, that deals with attitudes towards taxes and the behavior of tax evasion, environmental behavior such as energy conservation or contingent valuation, pricing in business, household saving, and investing.

Whereas psychologists study stock market behavior, Gary Becker (1976) applied a purely economic approach to explain marriage, crime, and having children—all behaviors that have traditionally been of more interest to psychologists than to economists. By assuming a maximizing utility behavior, stable preferences, and a market equilibrium—the key assumptions in economics—a wide range of human behaviors can be integrated and explained in terms of costs and benefits (Levinger 1965).

Another point of contact concerns the impact of economic variables on states and behaviors that are significant for the individual's well-being such as poverty, debt, or unemployment as related to happiness, physical and psychological health, and suicide. Conversely, more psychological variables such as achievement motivation or religious conviction can have an impact on economic success.

1.2 Points of Difference

As described above, psychology and economics differ not so much in their content domains but in the research methods and theoretical premises and frameworks. First of all, whereas psychologists study *individual* behavior, the research object of economists are *markets*, not individuals or even households. This distinction is most important as other major disagreements and often seemingly irreconcilable differences stem from it. Economists presuppose that markets behave rationally, following rules that maximize utilities. Although individuals may not adhere to these rules—as psychology has shown over and over—the assumption is that irrationalities on the individual level will cancel each other on the market level, resulting in rational markets. This assumption is unrealistic from psychology's view. A whole research tradition has accumulated evidence for how individuals deviate from normative rules not at random but in a systematic fashion. Given such systematic biases in judgment and decision-making, cancellation at the aggregate level becomes an illusion.

Psychology is inductive whereas economics is deductive. Psychologists observe individual behavior in order to test hypotheses concerning the underlying mediating processes. Although these hypotheses are often derived from higher-order theories, these theories are typically domain-specific and controversial, rather than universal and commonly shared. Economists on the other hand are more interested in predictors rather than causes, are more interested in outcomes rather than processes, and share widely accepted key assumptions such as utility maximizing behavior. Based on these key assumptions, economists deduce formalized models to predict particular outcomes. Whether the predictors are causally relevant at the individual level is only of interest to the extent that such causal mediators result in accurate models.

Clearly, this juxtaposition describes the extremes. Economists and psychologists have developed some common ground, combined interests, and a fruitful

dialog, mainly incorporating psychology into economic decision making. Some of this work will be reviewed in the following. However, given that psychology is mainly concerned with individual behavior and economics with aggregate behavior, it is interesting that the mediating link—how individual behavior translates into aggregate behavior—has largely been neglected by both sides and their respective interdisciplinary approaches. Aggregate behavior, however, is clearly not simply aggregated individual behavior. For example, although consumer sentiment predicts consumer spending at the aggregate level—and is indeed a better predictor than income or other economic variables—it does not do so at the individual level.

Research methods constitute an additional source of divergence. Psychological research relies mainly on highly controlled experiments mostly conducted in the laboratory. The experimental situation is quite often rather different from the setting where the corresponding behavior naturally occurs. In particular, economists have questioned whether the use of deception in experiments leads to valid research inferences. Nevertheless, a considerable part of economic research has adopted the experimental method from psychology, with and without deception.

2. *Exchange Rules as an Interface for Research in Psychology and Economics*

The common denominator of most empirical research in the interface of psychology and economics lies in the assumption of a basic *exchange paradigm* that governs all kinds of research topics such as pricing, marketing, consumer choice, fairness of negotiation, coalition formation, investment, etc. The ultimate goal of economical exchange is to minimize the total input to actions and transactions and to maximize the total output combined over several dimensions or resources (money, effort, time, security, etc.). Within this paradigm, success means to obtain high output from low input, failure means that the output is low relative to the input. Fairness means that different parties' (actors,' companies') output–input ratio is balanced. Note that in order to calculate an output–input ratio, it is essential that resources of all kinds be translatable to a joint scale so that effort can be paid in money, security can be traded against reduced profit, information costs are compensated by higher quality, and innovative ideas increase product quality, market price, and organizational efficiency (i.e., reduce workload and coordination loss). Only when all these different kinds of resources can be exchanged and can compensate each other, different action alternatives or decision options can be represented on a monotonic preference scale. Monotonicity or equivalence is indeed a precondition for all rational behavior within this basic exchange paradigm.

The rationality principle that underlies models of economic behavior as well as the irrational tendencies emphasized in psychological approaches revolve around diverse aspects of monotonicity. Indeed, a review of research on psychology and economics can be organized in terms of rules and violations of monotonicity on the one hand and problematic aspects of the underlying normative monotonicity assumptions on the other.

3. *Rules and Violations of Monotonicity: Research Examples*

3.1 *Transitivity of Preference*

Any quantitative scaling of enonomic goods (e.g., products, shares, investment options, job applicants) a, b, c must fulfill the transitivity constraint: If $a < b$ and $b < c$, then $a < c$. For instance, if the price that can be obtained for object b is higher than the price for a, and the price for c is even higher than for b, then c must also obtain a higher price than a. Otherwise, rules of exchange would either break down completely, or the joint scaling of prices, preferences, and attitudes would be severely complicated. However, in reality, this seemingly obvious normative rule can be violated systematically. Psychological models predict intransitivities under specified conditions. In a multi-attribute utility framework, each object a, b, c is characterized by multiple aspects, describing advantages or disadvantages. When comparing a pair or subset of objects, some aspects are more salient and receive higher weights than other objects. Intransitivities can arise when the preference of b over a and of c over b is due to different aspects than the comparison of a and c. For instance, when choosing among three meals in a restaurant, somebody may prefer b over a because b seems more delicious, and c over b because c seems even more delicious. But when comparing c to a, the price difference may become more salient, and a may now be preferred to c because a is less expensive.

3.2 *Independence of Context*

The independence principle implies that preference orders be invariant over context conditions; if $a < b$ in context x, then $a < b$ should also hold in context y. Again, empirical evidence and multi-attributive theories are in conflict with this premise of consistent exchange. Let a and b represent two cars the value of which is mainly determined by two attributes, performance and economy. Let a be superior in performance and b superior in economy. The relative advantage of either car in one attribute exactly compensates for its disadvantage in the other attribute so that a and b are represented on the same iso-

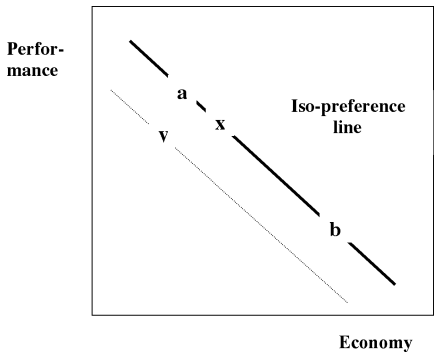

Figure 1
Graphical illustration of similarity and attraction effects

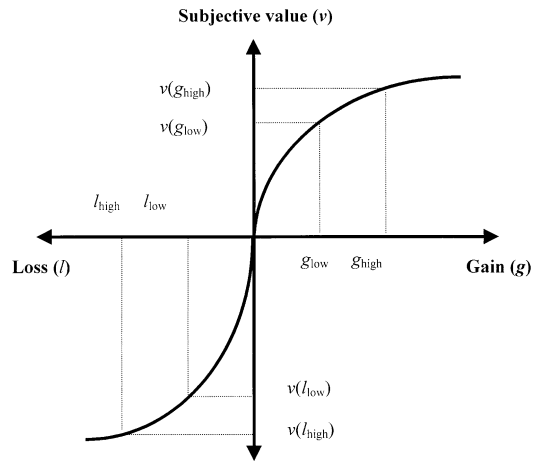

Figure 2
S-shaped subjective value function, according to prospect theory

preference line (Fig. 1). Given such a choice situation, the likelihood of consumers to choose *b* rather than *a* depends crucially on the presence of further alternatives. If the comparison is made in the context of a third car *x* that lies on the same iso-preference line but much closer to *a* than to *b*, then the context stimulus *x* becomes a rival for cars like *a*, inducing a tendency to prefer *b* over *a*. In contrast to this so-called 'similarity effect,' a context stimulus *y* that also lies close to *a*, but on a slightly inferior iso-preference line, functions like a decoy that serves to highlight the dominance of *a* over context stimuli and thereby increases the likelihood to choose *a* rather than *b*. This opposite influence of context has been termed 'attraction effect' (Huber and Puto 1983). Context effects like these have strong implications for marketing, consumer choice, and investment decisions involving similarly attractive multi-attribute options. For example, by including a new car in the range of an automobile manufacturer, the attraction of an already existing car may increase (attraction effect) or decrease (similarity effect).

3.3 Framing of Decision Options

Verbal labeling or circumscribing of economic targets must not affect their monotonic ordering for rational exchange to work properly. However, countless studies demonstrate preference reversals as a function of mere verbal framing. Decisions under uncertainty often involve the choice between two largely equivalent options, a P-option with a relatively high probability of success but a relatively low outcome, and an O-option with a higher outcome in case of success but a relatively low success probability. The tendency to choose the more risky O-option depends strongly on whether the verbal framing focuses on positive or negative outcomes. If the decision problem is framed positively in terms of possible gains, a majority of people would show risk aversion and choose the more secure P-option. If, however, the same decision problem is framed negatively in terms of the possible losses, preference for the O-option becomes more likely.

To account for such framing effects, psychologists and economists refer to prospect theory (Kahneman and Tversky 1984)—a very prominent model of economic choice. In particular, outcome framing effects are commonly explained by the assumption of an S-shaped subjective value function in which both the positive value of gains and the negative value of losses increase in a negatively accelerated fashion, as portrayed in Fig. 2. Twice as extreme an outcome is not worth twice as much, psychologically. The value of winning or loosing $200 is less than twice the value of winning or loosing $100, respectively. Such an S-shaped value function is consistent with the finding that positive outcome framing increases the tendency to choose the P-option (because the higher value of the O-option is worth less, due to the S-shaped function, than the higher probability of the P-option). Negative framing favors the choice of the O-option (because the higher negative outcome of the O-option is less extreme subjectively than the higher probability of the P-option).

3.4 Exchangeability and Compensability of Resources

An essential premise of exchange is that qualitatively different resources can be exchanged and aggregated in a compound value function. Thus, when calculating the lower limit of the price for a product, the costs for

development, production, advertising, and taxing have to be translated onto the same joint scale as the aesthetic appeal, the sonorous brand name, and the scarcity of the product. Monotonicity should not only hold within the singular components, but must also be conserved after exchanges involving different resources.

3.5 Pricing and Choice

One apparent implication of the exchangeability of resources is that differences in attractiveness should correspond to differences in the price that can be obtained for products or goods. A preferred or chosen object should justify a higher price than a less preferred or not-chosen object. However, many experiments show a preference reversal such that whereas the P-option is consistently preferred to an O-option (of the same expected value) on a *choice task*, the O-option is preferred over the P-option on a *pricing task* (Slovic 1995). Although most people prefer the secure P-option when making a choice, they are willing to pay a higher price for the O-option—an often-replicated finding that cannot be derived from prospect theory. One interpretation for the disparity of pricing and choice, though not the only one, is that the pricing task makes the outcome dimension (typically indicated in monetary terms) more salient, whereas the choice task makes the probability of winning more salient.

3.6 Effort and Accuracy

Another implication of considering qualitatively different resources as exchangeable is that amount of work and effort can be paid in money and, likewise, that work load is commonly considered a positive monotonic predictor of quality or returns. Accordingly, investing additional work should increase, and must not decrease, the quality of a product or the accuracy of an organizational decision. Otherwise, the economical value of labor and effort expenditure would be negative, creating a severe exchange paradox. Nevertheless, at least under specified conditions, working or thinking too much can impair rather than improve rational decisions (Wilson and Schooler 1991).

3.7 Information Costs and Expertise

Extending the same principle to information costs, a related implication is that the value of a product, or the future gains to be expected from rational decisions, should increase monotonically with the amount of information gathered, or with the expertise of consultants and decision makers. Information costs, much like production costs, must be subtractable from the gross income on the same monetary scale.

Of course, there is ample evidence to support this plausible assumption as, for instance, evidence showing that the reliability of decisions (e.g., choices among job applicants, prediction of growth, market analyses) increases with the number of diagnostic indicators. However, there is also a notable body of opposite evidence pointing out the less-is-more principle. As Gigerenzer and Goldstein (1996) have shown in pertinent computer simulations, increasing the number of predictors and increasing the complexity and sophistication of algorithms may actually reduce the validity of predictions and decisions. This seemingly paradoxical result can be expected in highly uncertain environments when indicators are error-prone and fallible so that reliance on an increasing number of indicators often means increasing reliance on error variance. To understand the less-is-more effect, the logical principle of rationality has to be replaced by the notion of ecological rationality. Moreover, bounded rationality highlights the need to pit accuracy against information costs and effort load and to be content with 'satisficing' rather than optimal actions and decisions.

One particularly intriguing variant of the less-is-more effect pertains to personalized knowledge or expertise. The additional knowledge acquired by experts as compared with lay-people, and the increased costs of expert consultation, may not be justified by a corresponding improvement of products or decisions. For example, at the stock market, experts who are highly informed about shares and market rules may be outperformed by novices simply because experts try to utilize too many fallible indicators (Borges et al. 1999). At the same time, the degree of overconfidence is likely to increase with increasing expertise. Thus, as experience and qualification increases, subjective confidence may increase at a faster rate than accuracy and actual achievement (Yates et al. 1998).

3.8 Monotonicity of Indices

Economical decisions must often rely on indices, such as the daily value of a share on the stock market, as an index of the corresponding company's (potential) performance. It is commonly assumed that indices correlate highly with the corresponding true properties on the same monotonic joint scale. For instance, it is likely that those companies whose share value index recently outperformed the market are expected to be more successful in the near future than those companies that lately failed to match the market. Indices are thus often regarded as monotonic functions of underlying true properties. However, this seemingly plausible assumption fails to take statistical regression into account, which is the source of many illusions and erroneous predictions. Regression is an inevitable result of error variance. To the extent that the reliability of an index is less than perfect, high index

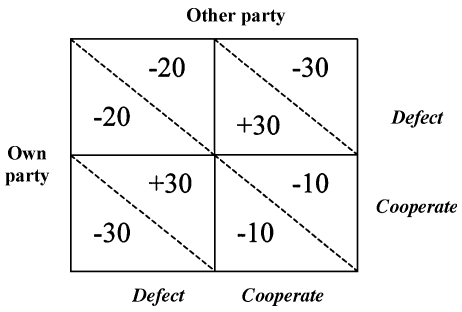

Figure 3
Pay-off matrix exemplifying a social dilemma situation

values are more likely to overestimate and low values are more likely to underestimate the true attribute to be assessed. Therefore, high values (of shares at the stock market) can be expected to be less successful after a while, whereas low values can be expected to be more successful. Ironically, then, investing in funds that recently outperformed the market may lead to less profit than investing in recently less successful funds (Hubert 1999). Thus, the assumption of a positive monotonic relation between an index and the true value of the underlying attribute is counteracted by an inverse relation between the error component of the index and the underlying attribute.

3.9 Absolute and Relative Gains and Losses

When the context, or state of the world, changes, monotonic orderings must be conserved, as already explained with reference to the independence principle. If decision option a yields a higher relative payoff than option b under condition x, and the relative payoff of a also dominates b under the complementary condition y, a is assumed to dominate b in general. Game theory is concerned with dramatic violations of this rule within social and economical dilemmas. The payoff structure of many negotiation and competition dilemmas is such that the payoff that one party gains from action a is higher than the payoff gained from action b, regardless of whether another party chooses strategy x or y. But although $a > b$ given either x or y, it is wiser and more profitable to choose action b rather than a. Such a payoff matrix is shown in Fig. 3, with reference to an embargo situation involving two parties each having two options: to cooperate (not delivering any goods) or to defect (violating the embargo to increase own profit). The situation can be exploited if one defects and the other party cooperates (yielding a high $+30$ payoff). However, if the other party also defects, violations of the embargo are sanctioned (-20). Nevertheless, defecting leads to higher payoff (i.e., lower embargo costs) than cooperating regardless of whether the other party defects (-20 vs. -30) or cooperates ($+30$ vs. -10). In spite of the *relative* dominance of defecting over cooperating, given both opponent strategies, cooperation can have a strong *absolute* advantage in the long run, just because defecting is likely to be reciprocated by the other party so that the absolute costs will be clearly higher (-20 or -30) than when the other party cooperates ($+30$ or -10). Therefore mutual fairness and solidarity is in the long run a more profitable strategy (moderate costs of -10 for both parties) than reciprocal defection (stable pattern of -20 for both parties).

4. Conclusion

The above examples represent only a small subset of possible monotonicity breaks. They illustrate that at the crossroads of psychology and economics interesting research problems emerge wherever the interchangeability principle is violated. This offers a rather fruitful perspective for the role of psychology in economic research. Economists often refer to psychology when the real data do not meet the theoretical predictions. The connotation is that psychology represents irrationality or some source of error which is responsible for the incongruency of economic models and real market behavior. However, the above examples demonstrate that violations of economic assumptions are not necessarily irrational. There may be good reasons for focusing on different features when comparing different products; the regressiveness of indices is an unavoidable problem; and the very fact that buying more information may sometimes lower the quality of decisions reflects the intrinsic uncertainty of the probabilistic world. Rather than attributing violations of rational norms to human deficits, these violations can be as well interpreted as reflecting unrealistic simplifications of the normative models used to investigate economic behavior.

See also: Entrepreneurship, Psychology of; Money and Finances, Psychology of

Bibliography

Becker G S 1976 *The Economic Approach to Human Behavior*. University of Chicago Press, Chicago

Borges B Goldstein D G Ortmann A LGigerenzer G 1999 Can ignorance beat the stockmarket? In: Gigerenzer G, ToddP M, ABC Research Group (eds.) *Simple Heuristics that make us Smart*. Oxford University Press, Oxford, UK, pp. 59–72

Gigerenzer G, Goldstein D G 1996 Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* **103**: 650–69

Huber J, Puto C 1983 Market boundaries and product choice: Illustrating attraction and substitution effects. *Journal of Consumer Research* **10**: 31–44

Hubert M 1999 The misuse of past-performance data. In: Lifson L E, Geist R A (eds.) *The Psychology of Investing*. Wiley, New York, pp. 147–57

Kahneman D, Tversky A 1984 Choices, values, and frames. *American Psychologist* **39**: 341–50
Katona G 1951 *Psychological Analysis of Economic Behavior*. McGraw-Hill, New York
Levinger G 1965 Marital cohesiveness and dissolution: An integrative review. *Journal of Marriage and the Family* **27**: 19–28
Slovic P 1995 The construction of preference. *American Psychologist* **50**: 364–71
Wilson T D, Schooler J W 1991 Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology* **60**: 181–92
Yates J F, Lee J-W, Shinotsuka H, Patalano A L, Sieck W R 1998 Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence?*Organizational Behavior and Human Decision Processes* **74**: 89–117

K. Fiedler and M. Wänke

Psychology and Marketing

Marketing is an applied field. It is synonymous in everyday parlance with the economic activity of selling, and especially with advertising. The field, however, is construed differently by most professional practitioners. The American Marketing Association defines marketing as, 'the process of planning and executing the conception, pricing, promotion, and distribution of ideas, goods, and services *to create exchanges that satisfy individual and organizational goals*' (emphasis added). While marketing draws upon a wide range of social sciences, including anthropology, economics, and sociology, the emphasis on satisfying goals implies that insight into the psychology of the customer is particularly important for the marketer. Thus a company might offer a new microwave oven that is the product of work by electrical and industrial engineers, home economists, financial analysts, etc. A marketer would bring to this effort a focus on how consumers will think about the product and how it will fit with their behavior.

Marketing is hence an applied field in that interventions are designed with regard to the psychological responses of consumers. Marketing research is commonly used to learn more about these responses. In practice this can be as simple, and as particularistic, as surveys asking about product preferences or usage and qualitative research such as focus groups eliciting more in-depth reactions from consumers. Academic researchers typically conduct more complex studies. An example is the work of Simonson (1990) on the variety of brands consumers choose for consumption over multiple occasions. He finds that the set of brands chosen reflects greater variety when the choices for multiple consumption occasions are made at a single point in time (i.e., simultaneously) vs. being made just prior to each consumption occasion (i.e., sequentially). This observation can be applied to marketing situations such as the design of retail stores. If a store, for instance, is intended as a convenience store to be used for items consumed immediately, it will not be necessary for it to carry as large a selection of products as with, say, a traditional grocery store that is used to build the household pantry inventory.

Psychology and marketing are thus intimately related. Marketing is based on an assessment of psychological response. The goal of marketing research is to enhance the applicability of these assessments to marketing situations of interest. There are two approaches to achieving this goal and each leads to a different view of exactly how psychology and marketing are connected.

1. *Effects Application*

One approach is that marketing research can focus on *effects application* (EA) (Calder et al. 1981, Calder and Tybout 1999). Research starts with psychological notions and proceeds by creating a research setting that allows the mapping of observed effects directly into situations of marketing interest. If the research setting is similar enough to the 'real' situation, the research is thought to generalize to these situations. From the perspective of this approach, it is important to conduct research with people who match the demographics and other characteristics of target consumers. Likewise the research setting, including independent variable treatments, should match the types of experiences people would have in the marketplace. Effects application studies are applied by extrapolating observed effects to a situation of interest.

The Simonson study described above can be viewed as an example of EA research. The research was no doubt motivated by psychological considerations, such as notions of consumers' desire for variety and their motivation to simplify complex tasks. The important thing is that although this research is motivated by psychological considerations it is the *effect*, greater variety with simultaneous versus sequential purchases, that is of primary interest. And the issue of application hinges on whether the effect can be directly generalized to marketing situations of interest. Because the original study was conducted with student subjects in a laboratory setting, the applicability to purchases in the marketplace might be viewed as suspect. For this reason, Simonson and Winer (1992) sought to enhance applicability by demonstrating the same effect with actual consumer purchase data from in-store scanner records. They looked at whether consumers bought a greater variety of yogurt flavors when they were buying larger quantities of yogurt. The argument from an EA perspective is that scanner data, in which actual

consumers spent their own money in real stores, makes the effect more directly applicable to marketing contexts than the laboratory study.

Other studies in the marketing literature use a related but somewhat different effects strategy. These studies seek to demonstrate the robustness of effects by demonstrating them across a variety of settings. Thus Strahilevitz and Myers (1998) draw upon notions of mental accounting to predict that the promise of a donation to charity will be a more effective incentive for consumer purchase when the product for which the incentive is offered is frivolous rather than practical. They find support for their prediction across three experiments that employ different stimuli. The robustness of the phenomenon, rather than the correspondence between the demonstrations and naturally occurring settings, provides the rationale for applying the effect to marketing contexts.

The EA approach views the relationship between psychology and marketing as follows: psychology provides important examples of psychological phenomena. It is the task of marketing research to investigate these effects in special research settings designed to be applicable to marketing situations of interest. Marketers thus draw from the psychological literature possible effects that must be observed in research settings that, by virtue of their similarity, are applicable to marketing situations. Or, alternatively, enough different settings are used to suggest robustness.

2. *Theory Application*

A second major approach leads to a very different view of the relationship between psychology and marketing, as well as what indeed it means for a discipline like marketing to be an applied field. This approach is called *theory application* (TA) (Calder et al. 1981, Calder and Tybout 1999). TA studies are designed to be applicable in a different way than are EA studies. The focus is on theory as an explanation of observed effects. It is not the effects that are expected to generalize; it is the explanation. Effects are used to test the theory. It is the theory that is expected to generalize to marketing situations.

This focus on theory testing leads in practice to marketing research studies that look quite different from EA studies. There is no requirement that research participants match consumers on demographic or other characteristics (unless this is relevant to the test of theory), nor is it important that things such as real money or natural marketing contexts be used in TA studies. Moreover, replicating the same effect over many situations is not regarded as particularly informative. This is because TA studies seek only to establish a strong explanation, not to generalize by virtue of similarity of the research setting and marketing situation or on the basis of robustness. Instead of focusing on whether the greater variety with simultaneous choices effect can be obtained in a setting that sufficiently resembles marketing situations, a TA approach would concentrate on developing the theoretical explanation.

It might, for example, be hypothesized in connection with the variety-seeking research discussed above that the simultaneous choice setting is more cognitively demanding than the sequential choice setting and, thus, leads consumers to employ cognitive strategies that simplify the task. Simplification might be achieved by expanding the set of selected items, thereby obviating the need to make finer discriminations among the items. Here we have an explanation of the variety-seeking effect as opposed to the effect itself. Evidence for this explanation, and against rival explanations (i.e., that simultaneous choice makes potential satiation more salient and thereby encourages variety seeking), is the goal of TA research. Application to marketing situations is based on the rigor with which the theory has been tested and how well the theory describes the marketing situation.

Research by Meyers-Levy and Peracchio (1995) provides a good example of TA research. These authors examined the processes that underlie consumers' responses to full color, black and white, and color-highlighted advertisement pictures. Petty and Cacioppo's (1983) elaboration likelihood model (ELM) served as a starting point for Meyers-Levy and Peracchio's theorizing. ELM describes two routes to persuasion: a central route where people are motivated by relevance to process detailed arguments and a peripheral route where people are not motivated to process detailed arguments and instead respond to peripheral cues. Meyers-Levy and Peracchio reasoned that the correspondence between the cognitive resources made available and those required to process the advertisement claim might moderate processing via the central route.

Specifically, when processing motivation is high and the advertisement claim requires few resources to substantiate it, advertisements that include color will be more persuasive than black and white ones. This is expected because individuals will have sufficient cognitive resources to process the color and the color will substantiate and reinforce the advertisement claim. However, when more resources are needed to assess the advertisement claim, either a black and white picture or one in which color is used to highlight only those aspects of the picture that are relevant to the claim, is expected to be more persuasive than a full color advertisement. This is because individuals will lack sufficient resources to substantiate the advertisement claim *and* process full color information.

Meyers-Levy and Peracchio's hypotheses received support in two laboratory experiments conducted with student subjects. As a result, their integration of ELM

with the notions of cognitive resource matching is preferred to an explanation based on ELM alone. Further, they demonstrate that cues are not *per se* either central or peripheral, but rather depend on the match between the resources required and available. It is Meyers-Levy and Peracchio's theoretical explanation of their effects, rather than the effects themselves, that is applied to marketing contexts beyond the particular one they examined.

As the above discussion makes clear, the research setting and subjects are not of primary interest in TA research, it is the explanation being tested. Meyers-Levy and Peracchio's studies would not generalize to marketing situations based on similarity between the research setting and subjects. However their theorizing might be applied to any number of marketing contexts. This application would use the theory to design an intervention for the situation of interest that according to the theory would work in a desired way. The explanation provided by the theory would predict an outcome but it would do so based on the design of the intervention suggested by the theory.

An example of this kind of theory application may be helpful. Suppose there is an interest in communicating with consumers who are in the market for a new car. Presumably such individuals are motivated to engage in detailed processing of message claims. Meyers-Levy and Peracchio's theorizing would suggest that full color advertisements be used to substantiate and reinforce claims related to styling (which may be apprehended with relatively few resources) but black and white advertisements be employed when presenting detailed information about the car's features and warranty. The design of the advertisements thus constitutes an application of the theory and leads to different expectations for the ads.

The important point about the above application is, again, that it does not depend on matching the settings of studies to the particular concern with advertising. It only matters that the theory has received support and that it describes the marketing situation of interest. Note also that there is always the possibility that there are other, unanticipated factors involved in the situation of interest that may not be taken into account by the theory. In our advertising example, for instance, the color resolution or the particular colors chosen might be unappealing, in which case the use of color might undermine rather than reinforce the style claim. Such possibilities have two implications. One is that any application via TA research has the potential to fail (just as with EA research where it is impossible to match on all differences between the research setting and situations of interest). The other is that it may be desirable to conduct *intervention testing* (Calder et al. 1981, Calder and Tybout 1999), in which more factors are allowed to vary freely than in the typical theory test-oriented TA study. This can increase confidence in the application, but because this will ordinarily be a weak test of the theory, it does not increase support of the theory *per se*, and the application still depends most critically on the prior TA research.

3. Conclusion

To return to the relationship between psychology and marketing, as we have seen, one account of this relationship is provided by the EA approach. Psychological phenomena motivate EA studies that are applied by virtue of the similarity of the research setting to marketing situations of interest. The alternative TA approach, however, provides a very different perspective. On this view the relationship of psychology and marketing is much more reciprocal and synergistic, as demonstrated in the discussion of Meyers-Levy and Peracchio's (1995) work.

With the TA approach marketing research is primarily engaged in testing psychological theories. This work is no different from the theory-testing work that goes on in psychology. It is only that marketing researchers have selected certain theories for further development and testing; these theories are ones that are considered especially promising for application. This applicability is based on the relevance of the theory to marketing situations. The success of the application depends most critically on the rigor of the TA studies supporting the theory, both within psychology and marketing. As experience with such applications increases, assisted in some cases by intervention testing, there will be an additional source of feedback about theories. It is the cooperative aspect of testing theories, through both TA research and the process of theory application, which constitutes the relationship between psychology and marketing. At the beginning of the twenty-first century, marketing continues to explore both paths to application.

See also: Advertising: General; Advertising, Psychology of; Decision and Choice: Economic Psychology; Marketing Strategies; Services Marketing

Bibliography

Calder B J, Phillips L W, Tybout A M 1981 Designing research for application. *Journal of Consumer Research* **8**: 197–207

Calder B J, Tybout A M 1999 A vision of theory, research, and the future of business schools. *Journal of the Academy of Marketing Science* **27**: 359–66

Meyers-Levy J, Peracchio L A 1995 Understanding the effects of color: How the correspondence between available and required resources affects attitudes. *Journal of Consumer Research* **22**: 121–38

Petty R E, Cacioppo J T 1983 *Attitude Change: Central and Peripheral Routes to Persuasion*. Springer/Verlag, New York

Simonson I 1990 The effect of purchase quantity and timing on variety seeking behavior. *Journal of Marketing Research* **27**: 150–62

Simonson I, Winer R S 1992 The influence of purchase quantity and display format on consumer preference for variety. *Journal of Consumer Research* **19**: 133–8
Strahilevitz M, Myers J G 1998 Donations to charity as purchase incentives: How well they work may depend on what you are trying to sell. *Journal of Consumer Research* **24**: 434–46

B. J. Calder and A. M. Tybout

Psychology: Historical and Cultural Perspectives

Psychology occupies a middle position among the sciences, between methodological orientations derived in part from the physical and biological sciences and a subject matter extending into the social and human sciences. The struggle to create a science of both subjectivity and behavior and the effort to develop professional practices utilizing that science's results provide interesting examples for the reach, and also the limits, of scientific ideals in modern life. This article presents a brief account of the social and cultural relations of psychological thought and practice from the eighteenth through the twentieth centuries.

1. Routes to Institutionalization, 1700–1914

From the earliest times philosophical arguments about the nature of humans and of human knowledge have included claims about mental states and psychological processes. Protestant scholastics first used the term 'psychology' in discussions of the relation of soul and body in the sixteenth century. Christian Wolff's systematic accounts of 'rational' and 'empirical' psychology published in 1732 and 1734, respectively, marked out a distinct subject of scholarship, with journals and practitioners, a subject matter and intended methods of study. But that subject generally remained a teaching and research field within philosophy and pedagogy. The introduction of so-called 'physiological' psychology in the middle third of the nineteenth century and the institutionalization of laboratory instruction on the model of the natural sciences soon thereafter had a major impact on the development of the discipline. However, experimental psychology never succeeded in dominating the field entirely. The institutionalization of psychology took different forms in different parts of Europe and the USA.

1.1 Britain and France

Britain was the home of the statistical research practices pioneered by Francis Galton. These practices targeted not psychological processes assumed to be essentially similar in all individuals, but rather distributions of performances *among* individuals. In *Essay on Human Faculty* (1883) and other works, Galton attempted to show that both physical and mental capabilities are quantitatively distributed in the same way, and that both are therefore inherited to the same—large—extent. In a 1904 paper, Charles Spearman took the approach a step further by distinguishing between 'general intelligence,' or 'g,' a factor underlying all performances in a test series and presumed to be hereditary, and so-called 's' factors accounting for differential performance on specific tests, presumed to be teachable. By the second third of the twentieth century, this emphasis on individual differences expressed in statistical terms had become a predominant research mode in both academic and applied psychology in the English-speaking world. Even so, as late as the 1920s there were only six university chairs for psychology in England. Psychological practitioners of various kinds far outnumbered academics in the membership of the British Psychological Association at the time of its founding in 1901.

In France, philosophers Hippolyte Taine and Théodule Ribot shared a coherent vision of the field as a synthesis of medical and philosophical approaches. However, Ribot's university course in psychology, first offered at the Sorbonne in 1885, was located in the Faculty of Letters rather than the Faculty of Sciences, while the laboratory demonstrations took place in the Faculty of Medicine. Despite this fragmented situation, he encouraged younger scholars, such as the physician Pierre Janet and the biologist Alfred Binet, to adopt a natural scientific approach.

When he became director of the first psychological laboratory in France, located in the Sorbonne's Faculty of Sciences, in 1890, Binet attempted to establish an explicitly biological science of higher mental processes, which he called 'individual psychology.' In 1894 he founded France's first scientific psychological journal, *L'Année Psychologique*. A commission from the Ministry of Education led to his publication, with Theodore Simon, of the first intelligence tests in 1905. The tests' purpose was not to measure intelligence directly—Binet doubted that this was possible—but rather to establish practical criteria for separating 'subnormal' from normal children, in order to provide the former with special education. But schoolteachers opposed the use of the tests in France, and here, as in England, extensive academic institutionalization of psychology did not follow.

1.2 Germany and the USA

Germany is generally regarded as the homeland of scientific psychology. In response to Immanuel Kant's claim that mental events, lacking the attribute of space, could not be measured, Johann Heinrich Herbart presented a program for the measurement of

sensations in 1816. In 1860, physicist and philosopher Gustav Theodor Fechner realized the program, with modifications, in his psychophysics—the measurement of relations between external stimuli and just-noticeable differences in sensation. Working in part with Fechner's techniques and in part with methods derived from laboratory physiology, Wilhelm Wundt established the first continuously operating laboratory for 'physiological psychology,' as he called it, in Leipzig in 1879. Other approaches widely received at the time included the 'medical psychology' of Rudolph Hermann Lotze, Franz Brentano's descriptive *Psychology from an Empirical Standpoint* (1874) and the *Völkerpsychologie* launched in 1860 by Moritz Lazarus and Haim Steinthal, which took an ethnological, linguistic, and historical approach. Other experimenters, such as Georg Elias Müller and Carl Stumpf, worked parallel to Wundt, but pursued quite different research programs.

In Wundt's and other German laboratories, in contrast to the situation in the UK or France, experimenter and subject were generally equal in status and often changed roles. German experimenters employed mechanical apparatus to control stimulus presentation, but they supplemented their data charts with extended records of their subjects' self-observations. In this way they engaged in an instrument-aided version of the self-discovery traditional to members of the German educated middle classes. Thanks to journals such as the *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, founded in 1890, as well as the Society for Experimental Psychology, founded in 1904 with Müller as first chairman, the 'new psychology' was on firm ground in Germany by 1905. But there was no agreement on the subject matter and methods of the discipline.

One reason for this was what Kurt Danziger (1990) has called the 'positivist repudiation' of Wundt by experimenters, including Müller, Hermann Ebbinghaus and Oswald Külpe. Though Wundt denied that experimental methods were sufficient to study the higher mental processes, these scientists extended apparatus-driven experimental techniques from sensation and perception to memory and thinking. Wundt opposed premature efforts to apply laboratory techniques in real-world situations, but William Stern, Karl Marbe, and others assessed the veracity of witnesses' testimony in court and tested the performances of schoolchildren at different times of day or the skills of industrial workers. In addition, an explicitly humanistic philosophical tradition persisted, with competing conceptions of the subject matter, method, and practical uses of psychology.

At the turn of the century, neo-Kantians and other philosophers attacked the intrusion of 'psychologism' into their field; in 1912, over 110 German teachers of philosophy signed a public statement opposing the appointment of psychologists to professorships in philosophy. But this protest failed, because state officials were not convinced that the discipline had any obvious link to professional or civil service training. Until the Nazi era, experimenting psychologists in Germany maintained their own laboratories, journals, and association, but generally continued to compete for chairs in philosophy.

Wundt's US students transferred the new experimental psychology from Germany to the USA in the 1880s and 1890s, but the positivistic concepts they employed to justify using such tools were quite different from Wundt's. The sheer size of the country as well as the decentralized structure of the emerging US university supported rapid institutionalization. By 1910 there were more psychological laboratories in the USA than universities in Germany. The founding of the American Psychological Association in 1892 predated that of corresponding European societies. Despite this rapid growth, the new discipline exhibited some continuity with the past. Instruction in psychology had long been part of the required philosophy courses taught by college presidents such as James McCosh at Princeton. These courses and their teachers encouraged an orientation toward moral issues and concentration on useful knowledge rather than the emphasis on empirical foundations for philosophy of mind prevalent in Germany.

Also formative of the discipline in the USA was the work of Darwin and Spencer. Evolutionary thinking reinforced the emphasis on biological functions vs. mental faculties, and also encouraged comparison of adult humans with children and animals. Such views supported evolutionary theories of cognition like those of James Mark Baldwin, and also granted psychologists so inclined the authority to present themselves as agents of human betterment. Education and child study thus came to be of central concern to US psychology; here John Dewey, G. Stanley Hall, and Edward Thorndike were the opinion leaders, though they advanced different research and reform programs.

William James combined science and reform in his own way. Himself an evolutionist in certain respects, he created, with Dewey and James Rowland Angell, a distinctly American functional psychology. Though committed to psychology as a natural science, he criticized the tendency to substitute psychologists' conceptions of reality for their subjects' reported experiences in his classic text, *The Principles of Psychology* (1890). He favored a more expansive conception of the subject matter of psychology, but his later efforts to study the experiences of psychics and mystics with the same objectivity as those of 'normal' adults was not widely accepted. Instead, an emphasis on social usefulness that presupposed an engineering model of science became central to the establishment of psychology in the USA.

Disagreements on the scope and methods of psychology in the USA paralleled those in Germany. A group of experimentalists led by Edward Bradford

Titchener, which began to meet separately from the APA in 1904, did not oppose applied work *per se*, but insisted on rigorous method both within and outside the lab. In contrast, activists like Hall, who pioneered the use of questionnaires in the USA, were less concerned with laboratory-style rigor than with translating moral issues into scientific ones.

In this atmosphere, intelligence testing became more popular in the USA than in France. Henry H. Goddard, director of a training school for so-called 'feeble-minded' children, propagated the tests as instruments of human betterment. Lewis M. Terman revised the Binet–Simon scale for use in US schools in 1915, and thus linked 'mental age' to school class years. This move proved well suited to US schools as sorters of a socially and ethnically diverse population.

1.3 Common Features of the 'New' Psychology

Despite these multiple routes to institutionalization and differences in approach, certain common features of the 'new' psychology can be identified. One of these common features was a reliance on measuring instruments to establish objectivity and acquire scientific standing. With their apparatus for the controlled presentation of stimuli and measuring reaction times, experimental psychologists reconstituted the object to which their efforts were addressed. What had been mental and moral capacities became psychical functions; and the sensing, perceiving, conscious mind became an instrument that functioned, or failed to function, in a measurably 'normal' way.

A second common feature of the 'new' psychology was the use of physiological analogies based in turn on mechanical physics and technology. The term 'inhibition,' for example, blended organic and machine metaphors and applied them to both human action and to society. Soon after scientists and engineers applied the idea of energy conservation to human labor in order to create a science of work intended to make the 'human motor' run more efficiently, Emil Kraepelin and others extended the effort to 'mental work'; Hugo Münsterberg gave the result the name 'psychotechnics.'

A third common feature of the 'new' psychology was a studied vagueness about the mind–body relationship. Many psychologists asserted some version of psychophysical parallelism or claimed a functional relationship of mind and brain, but few were very precise about the nature of that relationship. A fourth common feature was the contested use of the term 'experimental' itself. Until late in the century, the term *psychologie expérimentale* referred in both France and Germany to seances. The experimentalists actively opposed spiritualism and attempted to expose quack practitioners, but James, Pierre Janet, and others studied altered mental states in psychics and mystics.

This broader view was not widely accepted, due to a fifth common feature of the 'new' psychology—a tendency to restrict its subject matter to topics that could be addressed by the natural scientific methods and apparatus then available, such as psychophysics, sensory psychology, attention span, and retention. One result was an uneasy tension between efforts to preserve the notion of a volitional, active mind and the actual stuff of experimental research—measurable reactions to external stimuli. Another result was the exclusion of social or 'crowd' psychology; measuring instrument experimental methodology was plainly not applicable to group behavior.

A sixth common feature of the new psychological science was its gendered dimension. The head–heart dichotomy and the worship of the (female) 'beautiful soul' persisted through the nineteenth century; but its role in the 'new' psychology was ambivalent.The generalized, 'normal' adult mind that the experimentalists usually claimed to be their subject matter was at least implicitly the common property of both sexes, but the vocabulary and practices of objective science carried unmistakably masculine symbolism.

2. The Era of Competing Schools, 1910–45

The struggle for intellectual dominance in early twentieth-century psychology has been depicted since the 1930s as a battle of competing 'schools.' This view has its uses, but it conveys the false impression that all schools competed on an equal basis everywhere. Behaviorism captured both expert and popular attention in the USA in the 1920s, but the new approach was hardly taken seriously in other countries until after 1945. The 'reflexology' of Russian physiologists Ivan Pavlov and V. M. Bekhterev did not become a dominant approach in psychology even in the USSR until the 1940s. Gestalt psychology and other initiatives from Germany were received with interest but also with skepticism in other countries. Psychoanalysis had established itself as an international movement by the 1920s, but acquired few academic adherents at the time.

In German-speaking Europe, the 'crisis of psychology' announced by Vienna professor Karl Bühler in a 1927 book reflected continuing disagreement about subject matter and methods. The most widely received view internationally was that of Gestalt psychology, developed by Max Wertheimer, Wolfgang Köhler, and Kurt Koffka. They claimed, among other things, that immediately perceived structures (*Gestalten*) and relationships rather than isolated sensations are the primary constituents of consciousness, and employed holistic vocabulary to ground a rigorously natural-scientific world view. Felix Krueger, head of the Leipzig school of 'holistic psychology' (*Ganzheitspsychologie*), emphasized the role of feeling in perception and espoused Neo-Romantic cultural conservatism. William Stern's personalism focused on the individual as a 'psychophysical whole' in a way consistent with political liberalism.

The situation of psychology as a profession in Germany was equally contested. The challenge of philosopher Eduard Spranger's 'humanistic' psychology, as well as alternative professional practices such as handwriting analysis advocated by Ludwig Klages and typological personality diagnostics such as that of Ernst Kretschmer, raised the pressure to develop modern research instruments congruent with German cultural tradition. In Austria, the work of the Vienna Psychological Institute formed a bridge between old and new, theory and practice, Europe and America. Karl Bühler's department of general psychology conducted basic cognition research in the tradition of Brentano. At the same time, in rooms located at the city's adoption center, the institute's department of child and youth psychology, led by Charlotte Bühler, created performance measures for assessing the behavioral development of infants modeled on those of Arnold Gesell at Yale. In addition, Rockefeller Foundation funding supported the sociographic and survey research of the institute's Research Center for Economic Psychology under Paul Lazarsfeld in the late 1920s and early 1930s. All this put the Vienna institute, along with those in Jena and Hamburg, in the forefront of the transition to practice-oriented basic research in German-speaking psychology.

In the USA, multiple versions of behaviorism competed for attention in the 1920s. As proclaimed by John B. Watson in 1913, radical behaviorism excluded consciousness altogether from psychological science in favor of 'prediction and control' of behavior; in his later writings Watson advocated Pavlovian conditioning as a form of social engineering. Far more significant within the discipline at the time were the social science and child development programs generously funded by the Laura Spellman Rockefeller Foundation. The workers in these programs were not doctrinaire behaviorists, but they believed that measuring children's growth and IQ test scores over time would produce scientific norms of human development, and hoped to utilize this knowledge to rationalize society. Critics of behaviorism called on Gestalt psychology for support, while Harvard professor Gordon Allport and other prominent psychologists advocated a person-centered conception of psychology.

The 1930s were dominated by competing versions of neobehaviorism. Edward Tolman tried to integrate purposive motivation and cognitive processes into behavior theory, going so far as to claim that white rats framed 'hypotheses' as to which maze route would yield an expected food reward. Clark Hull developed an elaborate model for learning theory based on what he took to be Newton's physics, and tried to expand the model from the habit hierarchies of classical conditioning to personality theory. Finally, B. F. Skinner developed operant conditioning in the 1930s, producing careful measurements of the relative likelihood of simple behaviors such as rats or pigeons pressing a bar to obtain pellet, of food under rigorously controlled conditions and suspending all efforts to explain such behavior.

A prominent nonbehaviorist effort to bring systematic theorizing into psychology was that of Berlin émigré Kurt Lewin. Lewin advocated experimental study of ideal-typical behavioral situations, exemplified in his Iowa studies of 'democratic' and 'authoritarian' leadership in children's groups. Lewin and his US competitors shared an admiration for classical physics and a willingness to draw upon philosophy of science, especially operationism and logical positivism, to legitimate their positions. They differed in their basic conceptual foundations and also in the physics they chose to emulate. The competition was resolved only by the rapid fragmentation of the discipline in the 1950s.

In Britain and France psychology remained relatively weakly institutionalized in this period in comparison with Germany and the USA. Yet precisely this situation enabled a wide range of alternatives to US behaviorism to flourish. In the UK, Cyril Burt, originally a London school official and later professor at University College, London, adapted and expanded Charles Spearman's concept of general and specific intelligence in studies of educational performance, delinquency, and so-called 'backward' children. He then developed a mathematical basis for factorial approaches to intelligence and personality testing. Controversy over accusations that he manipulated or even invented some of his twin research and other data did not emerge until after his death. During the same period, Cambridge professor Frederick Bartlett published the pioneering study *Remembering* (1932), in which he established the role of learned schemata in retention and laid the foundations for considering memory as a process of active reconstruction rather than rote recall.

In France, psychology remained divided between medicine and philosophy; there was no separate degree until 1947. One result was that alongside the strictly experimental work of Henri Piéron, Binet's successor as director of the Psychological Laboratory at the Sorbonne, philosophers and sociologists considered psychological issues in broader and less positivistic ways. One example was the debate over Lucien Lévy-Bruhl's concept of 'primitive' mentality, which contributed to the emergence of the 'mentalities' concept of the *Annales* school in history. In French-speaking Switzerland, biologist and philosopher Jean Piaget, building in part upon Eduard Claparède's functional psychology, began his pioneering studies of cognitive development in children.

2.1 Dynamics of Professionalization to 1945

The turning point for the public visibility of professional psychology in the USA came with the mass

use of intelligence tests in the US Army during World War II. The route of application ran here not from the 'normal' to the 'pathological,' but rather from socially marginal populations—the so-called 'feeble-minded' and schoolchildren—to 'normal' adults. The interaction of applied psychology and the professional officer corps reshaped both the aims of intelligence testing, the test instrument itself, and ultimately conceptions of the objects being assessed. Intelligence became not intellectual or problem-solving capacity alone, but a sum of skills and (presumably hereditary) aptitudes for certain kinds of learning.

'Binet testing,' as it was then called, fueled the professionalization of psychology in both the USA and UK during the 1920s. Quantitative assessment and classification instruments spread rapidly in both basic research and professional practice, primarily because the products thus created supported the functions required by administrators, initially in schools and later also in industry and social service agencies. During this period the field became more open to women; but a gender hierarchy emerged, with industrial psychology remaining male-dominated, while female 'Binet testers' and social workers took on more people-oriented functions. During World War II, fields of application included the employment of social psychology in morale research and applied human relations, incorporation of psychophysics and experimental psychology into studies of human-machine interactions, and diagnostic testing in clinical psychology, as well as the use of intelligence and personality testing in personnel management. All this led in turn to significant basic research programs in the postwar period.

The professionalization of psychology in Germany took a rather different course. During World War I, efforts focused on the adaptation of techniques from psychophysics to develop sound-ranging devices, and to test the visual discrimination ability of drivers and pilots. More important after the Nazi takeover of power were the rapid growth of military psychology as a result of German rearmament, and the resulting shift from psychotechnical skills testing to 'intuitive' character diagnosis. The primary purpose here was elite officer selection, rather than sorting large numbers of average recruits. Paper-and-pencil and skills tests were secondary to the extended observation of officer candidates in simulated command situations. The personality characteristics sought had considerable affinities to the traditional virtues of the Prussian officer—the will to command and the ability to inspire troop loyalty. In contrast, diagnostic efforts based on Nazi 'race psychology' could not be translated into professional practice.

In the USA, too, personality diagnostics ultimately became a road to professionalization. However, in contrast to Germany, quantitative methods based on techniques of factor analysis developed by L. L. Thurstone and others predominated, despite competition from 'projective' tests such as the Rorschach in the 1930s and 1940s. The history acquired a gendered dimension in the construction of 'female' and 'male' traits in early personality research.

3. *The Postwar Era: 'Americanization' and the Alternatives*

In the USA, the postwar years saw explosive expansion and differentiation in both the scientific and professional realms. The establishment of a divisional structure within the APA in 1947—already negotiated during the war—reflected this process. Despite the optimism of the time, it proved difficult to subsume all aspects of psychology's protean identity within single university departments or graduate programs. Fragmentation was most obvious in the different research practices institutionalized in experimental, social, and personality psychology.

Cognition returned to the laboratory in this period, supported in part by reliance on methodological conventions patterned on those of the behaviorists, such as rigorous separation of independent and dependent variables, in part by computer-centered models of mental processes. In contrast to the statistical inference techniques and the corresponding computational models of mind that came to be preferred in cognition research, the preferred research tools in educational psychology were the correlational methods pioneered by Galton. A comparable methodological split distinguished neobehavioristic learning theory from experimental social psychology and personality theory. Nonetheless, experimental studies of social influence on perception by Solomon Asch and of prejudice by Gordon Allport, as well as T. W. Adorno and colleagues' *The Authoritarian Personality* study (1950) captured the imagination of many in the field. The popularity of such studies reflected a widespread tendency of the period to psychologize, and thus individualize, social problems. Meanwhile, developmental psychology took the work of Piaget as a touchstone for numerous studies closely related to the practical needs of schools for age-related developmental norms.

By the 1970s, the sheer numbers of psychologists (over 70,000, over 100,000 by the end of the century) had reached levels that could not have been imagined 50 years earlier. The growth was worldwide, but more than two-thirds of the total were Americans. The openness of both discipline and profession to women continues; today more than half the doctorates in the field go to women. Gender divisions also continue, with women being most numerous in developmental and educational psychology and men in experimental, industrial, and personnel psychology. Nonetheless, the broad institutional anchorage of psychology in the USA was more than sufficient to assure that the

research and professional practices instituted there would spread throughout the world, and that alternative viewpoints coming from Asia, Africa, or Latin America would generally be marginalized.

The most important exceptions to the overall trend were the impact of Piaget in developmental psychology, and the reception of work by British psychologists Hans Eysenck and Raymond Catell in personality testing and diagnostics. In cognition research, the work of Bartlett and the achievements of Soviet researchers such as Alexander Luria were mobilized to lend respectability and theoretical sophistication to the resurgent field in the USA. Nonetheless, the all-pervasive influence of computer metaphors and the associated information-processing models were plainly of Anglo-American origin.

The history of the psychological profession after 1945 continued to be affected by contingent local circumstances. The rise of clinical psychology in the USA was originally driven by the need to deal with large numbers of mentally ill veterans after World War II. The new field ultimately brought forth its own basic research in both clinical and academic settings, which led to the emergence of scientific communities based on methodological norms quite different from those of experimental or developmental psychologists. In addition, an eclectic, so-called 'humanistic' psychology movement arose in opposition to both behaviorism and psychoanalysis, and became widely popular in psychotherapy, social work, and the emerging field of counseling psychology.

In Europe, the rise of clinical psychology came approximately 10 years later than in the USA. There, in contrast to the USA, the supremacy of personality diagnostics and its quantitative tools had already been established in basic research before the professionalization of the clinical field. Another important difference was that clinical training in academic settings in Europe was based far more on cognitive and behavioral techniques than on psychoanalysis. Barriers to the academic institutionalization of psychoanalytic research and training were surmounted only in exceptional cases.

What had been a predominantly European field at the beginning of the twentieth century has become deeply dependent on US research styles and professional practices. US predominance has been contested by dissident local-language movements, most notably in France and Germany. Most significant, however, is the contrast between US predominance worldwide and the insecure standing of trained psychologists in the USA itself. Vagueness and confusion in the use of the term 'psychologist' in public discussion have been remarkably consistent over time; the term itself lacks legal protection in any case. The fact that the popularity of self-help books does not depend on whether their authors are psychologists or not indicates that even in the USA, where most of the world's psychologists live and work, trained academics and professionals can hardly claim hegemony over psychological discourse in the public sphere to the degree that physical scientists can in their fields.

4. *Conclusion*

Given this situation, it might well be asked why such a shakily legitimated field has acquired such an important role in the twentieth-century. Roger Smith (1998) suggests that the discipline drew its authority from and simultaneously gave voice to a culture and society in which everyone 'became her or his own psychologist' (p. 577). Nikolas Rose (1996) argues that psychological practices make possible particular kinds of social authority, assembled at first *ad hoc*, then grafted onto all activities aimed at simplifying the administration of modern life by producing calculable individuals and manageable social relations, from law and penal administration to education and parenting. No single science or profession has monopolized the codification and certification of these activities.

The predominance of behaviorism in the USA in the middle third of the twentieth century was an episode in a much larger story. However, it is a characteristic episode, for both the discourse of prediction and control and its associated practices have persisted, even as the so-called cognitive revolution reintroduced mentalistic vocabularies. One reason for this appears to be that not only the members of the discipline and profession called psychology, but the modern culture and society in which they function, require and may even desire both technocratic discourse and the instruments that embody and enact it.

See also: Allport, Gordon W (1897–1967); Authoritarian Personality: History of the Concept; Behaviorism; Behaviorism, History of; Binet, Alfred (1857–1911); Broadbent, Donald Eric (1926–93); Campbell, Donald Thomas (1916–96); Cattell, Raymond Bernard (1905–98); Cognitive Psychology: History; Cognitive Science: History; Disciplines, History of, in the Social Sciences; Erikson, Erik Homburger (1902–94); Freud, Sigmund (1856–1939); Galton, Sir Francis (1822–1911); Hall, Granville Stanley (1844–1924); Heider, Fritz (1896–1988); Helmholtz, Hermann Ludwig Ferdinand von (1821–94); James, William (1842–1910); Janet, Pierre (1859–1947); Jung, Carl Gustav (1875–1961); Kant, Immanuel (1724–1804); Kohlberg, Lawrence (1927–87); Khler, Wolfgang (1887–1967); Lewin, Kurt (1890–1947); Lorenz, Konrad (1903–89); Luria, Aleksander Romanovich (1902–77); Pavlov, Ivan Petrovich (1849–1936); Piaget, Jean (1896–1980); Psychology: Overview; Rogers, Carl Ransom (1902–87); Savage, Leonard J (1917–71); Skinner, Burrhus Frederick (1904–90); Stevens, Stanley Smith (1906–73); Thorndike, Edward Lee (1874–1949); Tversky, Amos

(1937–96); Vygotskij, Lev Semenovic (1896–1934); Watson, John Broadus (1878–1958); Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Ash M G 1995 *Gestalt Psychology in German Culture 1890–1967: Holism and the Quest for Objectivity*. Cambridge University Press, Cambridge, UK

Benetka G 1995 *Psychologie in Wien. Sozial- und Theoriegeschichte des Wiener Psychologischen Instituts 1922–1938*. Wiener Universitätsverlag, Vienna

Benjamin L (ed.) 1992 The history of American psychology. *American Psychologist* **74**: 109–335

Capshew J H 1999 *Psychologists on the March: Science, Practice and Professional Identity in America, 1929–1969*. Cambridge University Press, Cambridge, UK

Carson J 1999 *Talents, Intelligence and the Construction of Human Difference in France and America, 1750–1920*. Princeton University Press, Princeton, NJ

Danziger K 1990 *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge University Press, Cambridge, UK

Dehue T 1995 *Changing the Rules: Psychology in the Netherlands, 1900–1945*. Cambridge University Press, Cambridge, UK

Gardner H 1985 *The Mind's New Science: A History of the Cognitive Revolution*. Basic Books, New York

Geuter U 1992 *The Professionalization of Psychology in Nazi Germany* (trans. Holmes, R) Cambridge University Press, Cambridge, UK

Gould S J 1981 *The Mismeasure of Man*. W. W. Norton, New York

Herman E 1996 *The Romance of American Psychology*. University of California Press, Berkeley, CA

Joravsky D 1989 *Russian Psychology: A Critical History*. Blackwell, Oxford, UK

Kusch M 1995 *Psychologism*. Routledge, London

Lück H, Grünwald H, Geuter U, Miller R, Rechtien W 1987 *Sozialgeschichte der Psychologie*. Leske und Budrich, Opladen, Germany

Morawski J G (ed.) 1988 *The Rise of Experimentation in American Psychology*. Yale University Press, New Haven, CT

O'Donnell J M 1985 *The Origins of Behaviorism: American Psychology, 1870–1920*. New York University Press, New York

Oppenheim J 1985 *The Other World: Spiritualism and Psychical Research in England, 1850–1914*. Cambridge University Press, Cambridge, UK

Richards R J 1987 *Darwin and the Emergence of Evolutionary Concepts of Mind and Behavior*. University of Chicago Press, Chicago, IL

Rose N 1985 *The Psychological Complex: Psychology, Politics and Society in England, 1839–1939*. Routledge, London

Rose N 1996 *Inventing Our Selves: Psychology, Power and Personhood*. Cambridge University Press, Cambridge, UK

Ross D 1972 *G. Stanley Hall: The Psychologist as Prophet*. University of Chicago Press, Chicago, IL

Scarborough E, Furumoto L 1989 *Untold Lives: The First Generation of American Women Psychologists*. Columbia University Press, New York

Smith L 1986 *Behaviorism and Logical Positivism: A Reassessment of the Alliance*. Stanford University Press, Stanford, CA

Smith R 1992 *Inhibition: History and Meaning in the Sciences of Mind and Brain*. University of California Press, Berkeley, CA

Smith R 1998 *The Norton History of the Human Sciences*. W. W. Norton, New York

Sokal M M (ed.) 1987 *Psychological Testing and American Society*. Rutgers University Press, New Brunswick, NJ

Van Ginneken J 1992 *Crowds, Psychology and Politics 1871–1899*. Cambridge University Press, Cambridge, UK

Vidal F 1994 *Piaget Before Piaget*. Harvard University Press, Cambridge, MA

Zenderland L 1998 *Measuring Minds: Henry Herbert Goddard and the Origins of American Mental Testing*. Cambridge University Press, Cambridge, UK

M. G. Ash

Psychology of the Real Self: Psychoanalytic Perspectives

Theories of the self have led to more confusion than understanding in the development of psychoanalytic thinking. Freud's ambiguous use of the term *Ich*, referring to the self both as a whole person—subjective experiential—and to the ego (Strachey 1966) theoretical, more objective and mechanical—led to considerable confusion about the relationship between the self and the ego which persists to the beginning of the twenty-first century.

The matter was further confused (Bettelheim 1982) by the abstract mistranslations of Freud's use of the word soul. An additional factor contributing to the confusion was that Freud and the pioneer analysts mostly worked with neurosis and, therefore, with oedipal levels of conflict. This is a developmental level at which the self had already become autonomous and assumed its functions and capacities. As a result, it did not cause enough of the kind of clinical problems that might have drawn attention.

However, when analysts turned their attention to patients with psychoses and character disorders, they were confronted with disorders of the self. These patients, development was arrested not at the oedipal but at the preoedipal level and, therefore, their deep difficulties were predominantly with the functioning of the self. The lack of a fully developed concept of the self, which had not previously been an obstacle, now became one.

From this basic conflict there emerged two mainstreams of thought with regard to the concept of the self, each of which emphasized one of the two senses in which Freud used the term—either the whole person or as a special part of the mind with its own unique capacities and functions. Those who emphasized *Ich* as ego tended to be working mostly with neurotics and oedipal conflicts and placed heavy emphasis on the structural theory (ego, id, superego).

Those who emphasized the other meaning of the term *Ich*—the self as the whole person—tended to be working with psychoses and character disorders, and preoedipal levels of conflict and difficulties with the self.

This latter group, who emphasized the self as a whole person, further fractionated into a number of subgroups based on varying definitions of the self, its nature, and its functions as follows:

(a) Jung (1953) saw the self as a primordial image or archetype expressing the individual's need for unity, wholeness, and its highest aspirations.

(b) Rank (1929) focused on the psychological aspects of the emergence of the physical self at birth and emphasized separation anxiety as springing from birth trauma.

(c) Adler (1940) saw the life-style (repetitive patterns derived from early childhood) as a defensive overcompensation for inferiority feelings about the self, therefore blocking the emergence of the real, creative self.

(d) Horney (1950) postulated an idealized self which produced a pseudo unity or wholeness that blocked the emergence of what she called the real self.

(e) Sullivan (Mullahy 1970) felt that the essence of the self was determined by its functions: (i) the fulfillment of needs; (ii) the maintenance of security.

A reconciliation of both senses of the use of the word *Ich* or self—whole person and ego—has more recently been undertaken by theoreticians in ego psychology and object relations theory. Hartmann (1958) distinguished between the self as a whole person, the ego as a system (part of the tripartite structure of id, ego, superego), and the self representation as the intrapsychic representation of the self. He also defined narcissism as the libidinal investment of the self representation. Then Jacobson (1964), in order to work with psychotics, carried the concept further by defining the self representation as the unconscious, preconscious, and conscious intrapsychic representations of the bodily and mental self in the system ego. Kernberg (1982) used the term character to refer to the self as a whole person and preferred to use the term self to refer to the sum total of intrapsychic self representations in intimate contact with the sum total of object representations in the ego. In contrast to Hartmann, he felt that the self and object representations are both libidinally and aggressively invested.

Kohut (1971), from his work with narcissistic personality disorders, viewed the self as consisting of fused self objects which were invested with narcissistic libido that had its own line of development separate from the development of object representations and object libido. In other words, there was no separation of self and object, but rather maturation of self objects from primitive to mature forms.

This focus on the self as an intrapsychic entity, when kept in concert with the other perspectives of developmental object relations theory, can lead to a broader, more inclusive and comprehensive concept of the real self and its disorders. This view reconciles both definitions of *Ich*—the self and the ego—bringing together the personal, subjective experiential with the objective, mechanistic theoretical.

1. *The Real Self: Structure, Development, and Function*

1.1 *Clinical*

A few definitions will facilitate discussion of the structure of the real self.

1.1.1 Self-image. The image an individual has of himself at a particular time and in a specific situation. It consists of his body image and the mental representation of his state at the time. This image may be conscious or unconscious, realistic or distorted (Moore and Fine 1968).

1.1.2 Self-representation. A more enduring schema than self-image, constructed by the ego out of the multitude of realistic and distorted self images which an individual has had at different times. It represents the person as he consciously and unconsciously perceives himself and may be dormant or active (Moore and Fine 1968).

1.1.3 Supraordinate self-organizations. Because subjective experiences may be organized by multiple self representations, the 'I' of one state of mind may not necessarily be the same as the 'I' of another state of mind. This term is used for the organization and patterning of the various subordinate self-images and representations (Horowitz and Zilben 1983). It connects them and provides a continuity between them and a sense of unity and wholeness.

2. *The Real Self*

In keeping with ego psychology and object relations theory, the term 'real self' is used in the intrapsychic sense of the sum of self and object representations with their related affects.

The term 'real' is synonymous with healthy or normal. It is used to emphasize that the representations of the real self have an important conscious reality component even though unconscious and fantasy elements occur. It also indicates that the real self has an important reality function: to provide a

vehicle for self-activation and the maintenance of self-esteem through the mastery of reality.

The term real self also helps to differentiate it from the false selves of the borderline and narcissistic patients that are based primarily on defensive fantasy, not on reality, and are directed toward defense, not toward self-expression in reality. When using the term real self, both the collective subordinate self-representations and their supraordinate organization are referred to.

The real self exists as a parallel partner of the ego and has its own development, its own capacities, and its own psychopathology. The self and the ego develop and function together in tandem, like two horses in the same harness. If the ego is arrested in development, so is the self. One aspect of the self could be viewed as the representational arm of the ego, although it is obviously more than that. Similarly, one aspect of the ego, since it deals with volition and will and with the activation and gratification of individuative wishes, could be viewed as the executive arm of the self. However, it is also obviously more than that as its primary function is maintaining intrapsychic equilibrium.

Erickson (1968) referred to the dual and inseparable nature of the self-ego as follows: Identity formation can be said to have a self aspect and an ego aspect. What could be called the self identity emerges from experiences from which temporarily confused (subordinate) selves are successfully reintegrated in an ensemble of roles which also secure social recognition. One can speak of ego identity when one discusses the ego's synthesizing power in the light of its central psychosocial function and a self identity when the integration of the individual's self role images are under discussion.

To differentiate between personal identity [*or in my terms the real self*—JFM] and ego identity: personal entity is a conscious feeling based on two simultaneous observations: the perception of the self sameness and continuity of one's existence in time and space, and the perception of the fact that others recognize one's sameness and continuity. Ego identity, however, is the quality of that existence in its subjective aspect; it is the awareness of the fact there is a self sameness and continuity to the ego synthesizing methods.

The real self, then, provides an internal repertoire that, although finite and fixed, is varied and flexible enough to blend the need for real self-expression with the external roles required by adaptation.

3. *Development of the Real Self*

The ingredients or building blocks of the real self consist of the biologic and genetic endowment, the child's experience of proprioceptive and sensory sensations from his or her own body, as well as pleasure in increasing mastery in coping with the environment plus the interactions with a caring object as described by Mahler and McDevitt (1982).

I will assume the infant has two basic points of reference from which he builds up his self schema: one, his own inner feelings or states forming the primitive core of the self on the one hand and, two, his sense of the care given by the libidinal object on the other hand. Insofar as the infant's development of the sense of self takes place in the context of the dependency on the mother, the sense of self that results will bear the imprint of her caregiving.

Psychoanalytic child observation researchers have traced the emergence of the self in normal children in the first three years of life. The issues of this child observation research are quite complex and beyond the scope of this paper. The contributions of four researchers: Mahler (1975), Stern (1985), Bowlby (1969), and Schore (1994) are briefly summarized.

Mahler (1975) derived intrapsychic hypotheses from observations of the child's behavior. Her work with child psychotics led her to theorize that the child's self emerges from a symbiotic relationship with the mother. She's described this occurring in four stages: autistic, symbiotic, separation–individuation and on the way to object constancy. Each stage brought further separation of the self from the maternal object representation. She emphasized the vital importance of the mother's support for the child's emerging self.

Stern (1985) similarly based his theories of intrapsychic development from observations of the children's behavior. He found that there was no symbiotic stage that the child was prewired to see the mother perceptually as separate from birth and that the child was an active co-creator of its own development in relationship with the primary caretaker. He also described four stages: emerging self, core self, intersubjective self, and verbal self.

Bowlby (1969), unlike Mahler and Stern, made no intrapsychic hypotheses but limited his findings to what can be observed from the child's behavior, thus providing a firmer scientific base that could be replicated. He theorized that at birth the child had no attachment to the mother, and that the primary task was for the child to find a secure attachment. He divided that attachment into two types—the secure and the insecure. This insecure was then subdivided into avoidant, resistant, disorganized, and disoriented.

Schore (1994), based on neurobiological rather than psychoanalytic research, reported that the development of the brain is experience-dependent on the interaction with the primary caretaker or the mother. The genes are responsible for the physical neurons but it is the interaction with the mother during the crucial early stages of development between 10 to 12 months and 14 to 16 months that is responsible for the wiring of those neurons. It is the wiring that occurs during these crucial stages that helps the self to emerge and to take on its functions in the regulation of affect and socio-emotional relationships. Thus, the mother is

found to be as important neurologically as psychologically.

Although the self becomes autonomous and able to perform its function at three years, further increments are added during the phallic–oedipal phase, latency, and adolescence. As the child passes through each later psychosocial stage, the resolution of it provides certain capacities for the self such as industry, initiative, and generativity as described by Erickson (1968). The whole process of development of the self culminates in the normal identity crisis of adolescence, where the adolescent must test, select, and integrate the self images derived from the crises of childhood in the light of the climate of youth and sexual maturity.

3.1 Capacities of the Self

Simultaneously and in parallel with the maturation of ego functions, and with self and object representations becoming whole, the now whole, separate, real self becomes autonomous and takes on its capacities. These capacities were clinically identified mostly by their impairment and repair in borderline and narcissistic patients. This list is probably far from complete but forms an adequate clinical working scheme:

(a) Spontaneity and aliveness of affect. The capacity to experience affect deeply with liveliness, joy, vigor, excitement, and spontaneity.

(b) Self-entitlement. From early experiences of mastery coupled with parental acknowledgment and support of the emerging self, the sense builds up that the self is entitled to appropriate experiences of mastery and pleasure, as well as to the environmental input necessary to achieve these objectives. This sense, of course, is sorely deficient in the borderline disorder, and pathologically inflated in the narcissistic disorder.

(c) Self-activation, assertion and support. The capacity to identify one's unique individuative wishes and to use autonomous initiative and assertion to express them in reality and to support and defend them when under attack.

(d) Acknowledgment of self-activation and maintenance of self-esteem. To identify and acknowledge to oneself that one's self (in both senses of the term) has coped with an affective state and/or an environmental issue or interaction in a positive, adaptive manner. This acknowledgment is the vehicle for autonomously fueling adequate self-esteem.

(e) Soothing of painful affects. The capacity to autonomously devise means to limit, minimize, and sooth painful affects.

(f) Continuity of self. The recognition and acknowledgment through an effective supraordinate organization that the 'I' of one experience is continuous over time and related to the 'I' of another experience.

(g) Commitment. To commit the self to an objective or a relationship and to persevere, despite obstacles, to attain the goal.

(h) Creativity. To use the self to change old, familiar patterns into new, unique, and different patterns.

(i) Intimacy. The capacity to express the self fully in a close relationship with minimal anxiety about abandonment or engulfment.

(j) Autonomy. The capacity to autonomously regulate self-esteem and socio-emotional relationships.

4. The Healthy and Impaired Real Self

For a clinical example of the operation of the real self, consider a young psychologist who submits his first version of his PhD thesis to his advisory committee and is told that in some areas it lacks information and in others the writing is poor.

Since the cornerstone of his real self is his sense of self-entitlement, he feels entitled to mastery and whatever input that this requires, plus his capacities for self-activation, assertion, and commitment. He realizes at the outset that the purpose of submitting his thesis at that time was to learn its deficiencies so that he could correct them. He then applies his creativity, commitment, and self-assertion to find ways to learn the necessary knowledge and improve his writing. By mastering his reality problem, he has provided an autonomous feedback to support his self-esteem or self-image, a support which is also soundly based on reality.

To extend the metaphor of the psychology student to contrast the operation of the borderline, false defensive self with the real self: Because of his difficulty with self-activation, assertion and commitment, and creativity, the student might procrastinate and delay submitting his thesis way beyond the deadline. The critical comments would be felt as a reinforcement or proof of his inadequate self and would precipitate an abandonment depression to which he would respond by helpless clinging to a teacher or other authority, or by avoiding the project altogether. This, of course, would further reinforce his sense of an inadequate self.

See also: Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Ego Psychology and Psychoanalysis; Person and Self: Philosophical Aspects; Personality and Conceptions of the Self; Psychoanalysis: Overview; Self: History of the Concept; Self: Philosophical Aspects

Bibliography

Adler A 1940 *The Theory and Practice of Individual Psychology*. Harcourt Brace, New York

Bettelheim B 1982 *Freud and Man's Soul*. Alfred A. Knopf, New York
Bowlby J 1969 *Attachment and Loss: Attachment*. Basic Books, New York, Vol. 1
Erickson E 1968 *Identity, Youth, and Crisis*. W W Norton, New York, pp. 208–31
Freud A 1937 *The Ego and the Mechanisms of Defense*. Hogarth Press, London
Freud S 1964 New introductory lectures on psychoanalysis. Lecture XXXI: The dissection of the psychical personality. In: Stracheys (ed.) *Standard Edition*. Hogarth Press, London, Vol. 22, pp. 57–80
Hartmann H 1958 *Ego Psychology and the Problem of Adaptation*. International Universities Press, New York
Horney K 1950 *Neurosis and Human Growth*. W W Norton, New York
Horowitz M, Zilben N 1983 Regressive alterations in the self concept. *American Journal of Psychiatry* **140**(3), March
Jacobson E 1964 *The Self and the Object World*. International Universities Press, New York
Jung C G 1953 Two essays on analytic psychology. In: *Collected Works* **7**, Pantheon, New York
Kernberg O 1982 Self, ego, affects, drives. *Journal of the American Psychoanalytic Association* **30**: 893–917
Kohut H 1971 *The Analysis of the Self*. International Universities Press, New York
Mahler M S 1975 *The Psychological Birth of the Human Infant*. Hutchinson and Co., London
Mahler M, McDevitt J 1982 Thoughts on the emergence of the self with particular emphasis on the body self. *Journal of the American Psychoanalytic Association* **30**: 827–47
Masterson J F 1983 *The Real Self: A Developmental, Self, and Object Relations Approach*. Brunner/Mazel, Inc., New York
Moore B, Fine B (eds.) 1968 *Glossary of Psychoanalytic Terms and Concepts*. American Psychoanalytic Association, New York
Mullahy P 1970 *Psychoanalysis and Interpersonal Psychiatry: The Concepts of H. S. Sullivan*. Science House, New York
Rank O 1929 *The Trauma of Birth*. Harcourt Brace, New York
Schore A N 1994 *Affect Regulation and the Origin of the Self*. Lawrence Erlbaum Associates, NJ
Stern D 1985 *The Interpersonal World of the Infant*. Basic Books, New York
Strachey J (ed.) 1966 In collaboration with Freud A. *The Complete Psychological Works of Sigmund Freud*. Hogarth Press, London

J. F. Masterson

Psychology: Overview

Psychology is both an academic discipline and a profession. The concept of psyche originates from the Greek language and is frequently translated as 'soul.' The term has been used in religious and philosophical treatises since the antiquities, before an academic discipline explicitly designated as psychology was established in the sixteenth century. Until the nineteenth century, psychology was a special field within philosophy. During the first half of the twentieth century, it was institutionalized in universities as an independent science. In the second half of the century, an increasing number of positions for university graduates with a degree in psychology were provided in public service, business organizations, and private practice. Today, there are approximately 300,000 scientifically qualified psychologists worldwide, most of whom are employed outside academic institutions.

Although psychology often presents itself as a unitary field, it comprises a considerable diversity of domains, theories, and methods. Despite its independent status, psychology shares historical roots and current concerns with other disciplines and professions. Both the ramifications of psychology into diverse branches and its position among related disciplines and professions will be outlined in turn.

1. Conceptions and Fields ('Psychologies') within Psychology

The soul was conceived as the principle which distinguishes the living body from the dead and manifests itself in consciousness and movement, or mind and behavior. The soul and its mental and behavioral manifestations have been studied from various perspectives. The paradigmatic perspective emphasizes theoretical and methodological principles and affiliations to 'schools.' The universal perspective seeks for knowledge that is valid for the human species, or even all vertebrates endowed with intelligence and mobility. In contrast, the differential perspective explores the variations occurring within and between species. In the analytical perspective, researchers refine the existing knowledge by discovering ever more details, while adherents of the holistic perspective endeavor to comprehend higher-order structures. The theoretical perspective calls for the description, classification, and causal explanation of mental and behavioral phenomena. The pragmatic perspective views the soul as the source of happiness, sociability, and efficiency, and raises questions of utility and morale.

Specialized research programs and practical applications have emerged from these perspectives. They have been institutionalized as curricular units, classifications for books and journals, and as divisions within and between scientific and professional associations. Some programs have competed for dominance and presented themselves as complementary 'psychologies' within psychology.

1.1 Paradigmatic: The Three Major Schools

Scientific paradigms are defined by their explanatory principles, the concomitant concepts and terms, and their preferred methods and values. They are established and maintained by a community of scientists that endorse one and only one paradigm. Modern psychology has spawned three major theoretical paradigms: cognitivism, behaviorism, and depth

psychology, all of which continue longstanding philosophical traditions.

In the tradition of rational philosophy, cognitivism regards human beings as reasonable, conscious, reflexive, and self-organizing. In this view, each person is endowed with insight into natural, logical, and social laws, and may utilize their free will and foresight to plan and execute actions directed towards valued goals. Moreover, humans are seen as sociable and communicative. Language is a tool for representing imaginations and intentions both in a person's own mind and in his or her interaction with others. Authors with a cognitive orientation (e.g., Broadbent, Köhler, Piaget) rely on introspection and self-reports (e.g., think-aloud protocols in studies of problem solving), but also collect behavioral data (e.g., operations in the course of problem solving). Typically, cognitivistic studies employ methods from mathematics, physics (e.g., field theory), or engineering (e.g., information theory) to elucidate the intricacies of mind and behavior.

In the tradition of materialism and utilitarianism, behaviorism (in Russia: reflexology) focuses on overt behavior (e.g., locomotion, instrumental acts) elicited by external stimuli (e.g., color signals) and internal stimuli (e.g., thirst), or emitted to take advantage of opportunities in the natural and social environment (e.g., food, praise). According to this conception, a human being is an automaton capable of forming stimulus-response associations (habits) and of adapting to contingencies between acts and their rewards or punishments (reinforcements). Behaviorism assumes innate, species-specific habits, or instincts (e.g., mating). Authors such as Hull, Pavlov, and Skinner have studied the acquisition of new habits and the shaping of behavior by reinforcements. Orthodox behaviorism opposes introspection, and relies exclusively on objective behavioral data.

In the tradition of mysticism, depth psychology posits a pervading influence of unconscious ideas and affects on conscious thought, language, and action. Freud's psychoanalysis, the most popular paradigm within depth psychology, contends that the unconscious originates from personal conflicts in early childhood when anxiety-provoking tendencies (primarily hostility towards the parent of opposite sex) are repressed and rejected from the conscious mind. In his analytical psychology, Jung even postulates an innate, species-specific unconscious consisting of archetypes. These archetypes are explained as fundamental experiences (e.g., motherhood, fire) which have been acquired and transmitted over generations. Depth psychology believes in the power of the enlightened human mind to reveal unconscious contents. However, the revelation of unconscious contents requires skillful interpretations. Beyond straightforward observations, interpretations should offer explanations for the meaning of symbols from the unconscious (e.g., the moon as a symbol of femininity) which are encountered in dreams and other fantasies, as well as in achievements (e.g., in the arts) and failures (e.g., slips of the tongue).

The three major theoretical paradigms gained wide recognition during the first half of the twentieth century; they were also engaged in mutual antagonisms. However, no final verdict has been reached in the battles for dominance. There were some partially successful attempts to reconcile the conflicting paradigms. At present, conflicts between 'schools' have attenuated and orthodoxy has been superseded by eclecticism. For instance, numerous behaviorists now regard cognitions (e.g., attitudes and beliefs) as determinants of behavior, despite their subjective nature, while cognitive theorists accept the notion of unconscious ideas (e.g., subliminal perceptions).

1.2 Universal, Analytic: General Psychology, Biological Psychology

General psychology investigates universal and basic features of mind and behavior. Originally, it was launched in the spirit of philosophical idealism that acknowledged as scientific only universal and basic statements probing into the essence of matters. Ever since, general psychology has continued the quest for invariances and fundamental notions. It has predominantly pursued four objectives:

(a) Taxonomies. Cognitive functions have been categorized as sensation and perception, imagination and memory, concept formation, reasoning, problem solving, and language. Similarly, behavioral functions were classified as motives, drives, intentions, emotions, motor patterns, and actions.

(b) Structural analysis. Both cognitive and behavioral structures have been researched extensively. An example is semantic networks, that is, organizations of concepts representing domains of knowledge (e.g., automobiles), and action sequences (e.g., mountain climbing).

(c) Functional analysis; that is, the elaboration of the conditions which enable, prevent, and modify cognitive and behavioral processes and their outcomes (e.g., task complexity as a condition of learning).

(d) Formal (e.g., mathematical) models. General psychology designs models of internal mechanisms underlying psychic phenomena (e.g., the model of working memory). These models also serve to explain the effects of salient conditions.

In pursuit of these four objectives, general psychology is striving for progressive refinement. For structural and conditional analyses as well as for tests of formal models, experimental methods have become indispensable. Therefore, general psychology is sometimes equated with experimental psychology.

General psychology can be extended to encompass the exploration of the biological bases of cognitive and behavioral phenomena—the domain of biological

psychology. Using advanced measurement techniques, biological psychology locates and monitors mechanisms and processes of the brain, the peripheral nervous system, the endocrine glands, and other physiological systems which constitute the biological substrate of psychological functions such as language, memory, perception, and emotion, and which mediate cognitive and motivational states such as wakefulness and anxiousness.

1.3 Differential, Analytic: Animal, Developmental, Social Psychology, Individual Differences

Since the eighteenth century, the monopoly of general psychology has been challenged by romanticism, which emphasized the principle of organic growth in the natural world and in cultural settings, and by evolution theory (Darwinism). Both movements viewed variation as a fundamental natural principle. Romanticism and evolution theory provided the intellectual climate for the emergence of new psychological approaches that made variations their object of study:

(a) Animal psychology, also designated as comparative psychology, compared the behavior and the intelligence of species, including humankind. Much of this work was based on the assumption of a mental evolution (i.e., a gradual increase in intelligence from lower animals to humans), and a behavioral evolution (primary instincts like love or hate, transmitted from prehistoric man to modern-day humans, and secondary instincts acquired individually).

(b) Child psychology commenced with the study of mental and behavioral changes in humans during the first years of life, and was later extended to developmental psychology covering the whole life-span. In the beginning, child psychology frequently had an evolutionary background, since child development was interpreted as a replication of the evolution of the human species. Subsequently, evolution theory lost its major impact, and the study of human development continued with a generic interest in the changes that occur during people's lives.

(c) Social psychology documented and reflected the variations in cognitions, habits, and rules that are attributable to culture and social groups within cultures (e.g., the family). It drew cross-cultural comparisons of social structures, above all language, art, customs, power, education, kinship. Favorite topics were the evolution of social structures (social Darwinism) as well as the impact of external (e.g., geographical) factors on social life. Another influential branch was mass psychology that attributed to crowds the loss of rationality. Crowds were described as emotional, expressive, irresponsible, and addicted to charismatic leaders. Finally, social psychology restricted itself to the analysis of three basic issues: social relations (e.g., centralized vs. decentralized groups), social cognitions (e.g., attitudes, stereotypes), and social processes (e.g., communication, group problem solving).

(d) The psychology of individual differences emerged from theories of human faculties, both intellectual (e.g., reasoning) and moral (e.g., altruism). The study of individual differences benefited from sociometrics which assessed frequency distributions of crime, income, etc. in populations. Measures of intelligence, motivation, and performance were found to be distributed as the Gaussian bell-shaped curve with the mean representing the most frequent case. From the genetic point of view, the positions of parents and their offspring in distributions were compared. Obvious correlations between parents and their children could be attributed to hereditary factors. However, sociocultural factors (e.g., educational style) were also examined as causes for transmissions of mental and behavioral dispositions over generations.

The studies on variations have modified the taxonomies and models from general psychology to account for inter- and intra-individual differences. New models were constructed which consider the ontogenetic, phylogenetic, social, and individual factors governing mental and behavioral variations. Thus, developmental and animal psychology have designed stage models of ontogenetic and phylogenetic changes, and social psychology has developed models of group formation and group cohesion.

Methodologically, differential approaches have conducted experiments, but have also collected a large body of data by field observations and surveys (e.g., longitudinal studies in developmental psychology).

1.4 Holistic: Personality Psychology

While human life manifests itself in a diversity of phenomena that vary with social context and over the life-span, the experience of a unitary mind and of consistency in behavior prevails in human consciousness. Upon this experience the concepts of self and ego are based. These two concepts replace the concept of soul in modern psychology. Self and ego represent an internal system which integrates diverse and even conflicting functions (e.g., inhibition of recollections), developmental stages (e.g., transitional stages from being protected in childhood to assuming responsibility in adulthood), and social roles (e.g., private vs. professional) to establish a sense of personal identity in consciousness and activity.

The unitary structure of mind and the consistency of behavior are the central issues of personality psychology. The concept of personality originally designated the coherent and integrated organization of persons, reflecting their life-long identity and individuality. Thus far, personality psychology has followed the philosophy of holism, that truth and valuable action can only be attributed to complete,

integrated structures. Presently, postmodern approaches question the assumptions of continuous integrality and self-identity. They prefer the assumption of a patchwork personality consisting of discontinuous and even contradictory features (e.g., role conflicts, different attitudes for different settings).

The method of understanding has been advocated as the most useful approach to personality. Understanding has been conceived as a hermeneutic technique that enables direct communication between an observer and his or her object. Adherents of the hermeneutic position have contributed in-depth character studies and biographies. Influential and popular hermeneutic approaches to personality development were provided by depth psychology (see Sect. 1.1). Numerous authors have acknowledged the artistic but have criticized the scientific relevance of hermeneutic contributions. As alternatives, they have proposed more rational models for personality theory. Factor theories of personality and theories of self-regulation have met with great success. Factor theories are based on measurements of individual differences. Employing mathematical techniques, they reduce the variety of measures to a minimum of personality dimensions. Extraversion–introversion and neuroticism are among the most frequently identified dimensions. Theories of self-regulation attempt to conceptualize the processes by which individuals maintain their identity and consistency, as well as those mechanisms through which persons induce or cope with changes in their personality. Central to these theories are self-concept, self-awareness, and self-management.

Due to its diversity of methods and theoretical principles, personality psychology is the most heterogeneous among the fields treated in this article. Its holistic claims, however, have merited its frequent appraisal as the most substantial field in psychology.

1.5 Pragmatic: Applied Psychology

Psychology has served three practical purposes: assessment, intervention, and evaluation. A large number of testing methods are available to assess the intelligence and the personality, as well as the specific aptitudes and attitudes of individuals. Based on these assessments, persons can be selected (e.g., for vocational careers), and classified for specific treatments (e.g., therapy plans). Besides persons, objects (e.g., sales products) and situations (e.g., work places) are also subjected to psychological assessments. Interventions include training, therapy, rehabilitation, and work organization programs. Evaluation refers primarily to public programs (e.g., educational projects, health campaigns). Evaluation methods are designed to monitor outcomes and costs of these programs, and to offer advice on how to improve their efficacy.

Some have claimed that a generic applied psychology can be adapted to all domains of civilization. Indeed, psychological services have been offered for a plethora of specialized fields such as architectural psychology, sports psychology, and traffic psychology. Firmly established are the following:

(a) Clinical psychology for the diagnosis and therapy of mental and behavioral disorders in individuals, couples, and small private groups. Clinical psychologists typically treat neurotic and psychosomatic disorders (e.g., anxiety, compulsion, high blood pressure), as well as behavioral and social deviations (e.g., aggressiveness, substance abuse).

(b) Educational psychology is predominantly applied in educational institutions (preschool and school programs, special schools, vocational training, etc.), where psychologists advise educators and offer guidance to students. Educational psychologists also deal with educational processes resulting from peer interactions, media exposure, etc.

(c) Work and organizational psychology emphasizes personnel development, training, and work organization in various industries and on the administrative level.

(d) Law psychology is divided into several branches. Forensic psychologists assess witnesses and defendants in court cases; in criminal psychology, the objective is the prevention of delinquency and psychological support for inmates of penal institutions. Moreover, psychologists act as consultants to administrative courts (e.g., by rendering opinions in custody disputes).

(e) Economic psychology encompasses market research, consumer communication, and macroeconomics (i.e., the relationship of income, consumption, and savings).

The traditional view of applied psychology is based on the scientist–practitioner model which assumes that practical expertise is an extension of basic theoretical and methodological expertise. Alternatively, applied psychology may be conceptualized as representing an independent field of practical expertise that originates from the direct interaction with the problems encountered in the various branches of psychology. In this latter view, the scientific value of applied psychology rests on the documentation and analysis of practical problems, and on the validation of psychological procedures and theories. For instance, intelligence tests are not directly derived from basic theories of reasoning. They are carefully selected compilations of various tasks that have long been familiar in educational institutions and workplaces. Psychologists have standardized and validated these tasks to meet specific diagnostic requirements (e.g., school and job entry examinations).

1.6 Conclusions

The programs and paradigms of psychology form a fairly coherent and stable structure. Nevertheless, they exhibit inclinations towards isolating themselves. Typically, fields such as biological and social psy-

chology establish their own research programs, and each field holds its own conferences. Deficits in interdisciplinary exchange have also been claimed to exist between basic psychological research and applied psychology. On the other hand, there is a considerable overlap between various fields of specialization. For instance, research in social psychology may be duplicated in developmental psychology, as development occurs in social contexts and is interpreted to a large extent as a socialization process.

A pervasive feature of psychology is the fragmentation due to differences in perspective and historical background. Traditionally, boundaries between different 'psychologies' have been lamented and have given rise to efforts at unification and integration. Recently, however, disintegration and deregulation have also been interpreted as the results of diversity and subjectivism, and the fragmented state of psychology has been defended as a desirable adaptation to postmodern culture.

2. *The Position of Psychology among Academic Disciplines and Professional Fields*

Psychology has often been hailed as the heart of the life sciences. Psychological problems, methods, and theories relate to the humanities as well as to the social and the natural sciences. From the humanistic point of view, psychology is the basis of all scientific inquiry, since it studies the human mind as the source of knowledge. Furthermore, it provides the foundation for moral conduct as it elucidates the human condition.

Until the nineteenth century, scientific research of mind and behavior, and of individuals and cultures was conducted primarily under the label of philosophy. Moreover, psychological doctrines have been cultivated in law schools, medicine, and theology. In the late nineteenth century, psychology was institutionalized as a specialized academic discipline, along with other methodologically and theoretically advanced fields from the philosophical faculty such as linguistics, history, ethnology, sociology, and biology. A similar development took place in the practical domain. When psychology became a profession, some related branches of practice had already been firmly established (e.g., teachers in public schools). With an academic and professional system steadily growing and splitting into branches, psychology followed two courses. As an independent scientific discipline and profession, psychology develops specific theories, methods, and practical techniques to cultivate an exclusive expertise. In terms of its interdisciplinary position, various related disciplines and professions continue to contribute psychologically relevant findings and methods.

In the second part of this article, the position of psychology among academic disciplines and social service professions will be examined. Both disciplinary and interdisciplinary projects will be regarded within four contexts.

2.1 *The Metaphysical Context: Psychology, Theology, Occultism, Epistemology*

Soul is a central concept in psychology and in theology. Theology defines the soul as an immaterial substance inherent in individuals. As far as the theological conception of the soul is holistic, it corresponds to the notion of a unitary self in personality theory (see Sect. 1.4). Unlike psychology, however, theology assumes a metaphysical world of external wisdom and ultimate values as the origin and destination of the soul. Moreover, theology posits interactions between the soul and metaphysical agents (e.g., gods, angels). This belief extends to miracles beyond rational explanation (e.g., resurrection from the dead), and to particular vicissitudes of the soul (e.g., creation, condemnation, and redemption).

While psychology as a modern doctrine refrains from metaphysical speculations, it does consider religious experiences and practices from a theoretical and a practical point of view. With regard to the former, a field designated as psychology of religion analyzes religious experiences and practices as social cognitions and customs. Regarding the latter, ministers are trained in pastoral psychology that is aimed at improving their ability to cope with the demands of their profession. Psychological programs, especially psychotherapy, may include religious notions and exercises to serve clients who are believers. Most churches agree with psychology in respect to the desirability of mundane peace and harmony. Therefore, clerical counseling services include or cooperate with psychological agencies. During and after natural and technical disasters (e.g., earthquakes, airplane accidents), members of the clergy and psychologists trained in emergency psychology join to provide immediate help to victims and their families.

While religious beliefs were discounted as superstitions in the era of enlightenment, some authors continued to believe in the existence of immaterial beings (e.g., spirits, demons) and rationally inexplicable phenomena (e.g., accurate prophecies, communication with spirits, movement of physical objects by spiritual forces). The movement of occultism suggested that these phenomena may be based on natural conditions that have thus far eluded all scientific scrutiny (e.g., 'animal magnetism'). Parapsychology has resumed the investigations of occult phenomena. Numerous studies have revealed deception as the origin of allegedly supernatural events. Nevertheless, there have remained some unusual demonstrations and reports that could not be explained scientifically.

In accord with epistemology, the field of metaphysics has not been fully abandoned in modern

psychology. Beyond religious beliefs, metaphysics has been interpreted as cognitive contents that transcend personal experience. In this sense, the metaphysical approach is maintained in nativistic theories of psychology such as the assumption of archetypes in analytical psychology (see Sect. 1.1) that allegedly are transmitted to the unconscious from preceding generations. Moreover, fundamentals of perception and reasoning (e.g., the dimensions of space and time, the concept of causality) have been postulated which serve as prerequisites for the acquisition of new experiences, but are not the result of individual experience.

2.2 The Cultural Context: Psychology, the Humanities, and the Social Sciences

Psychology has shared topics with all the humanities and social sciences. Linguistics is an example. The relationship between language, mind, and society has long been deemed worthy of scrutiny. Consequently, the interest in language as the object of research is common to a variety of disciplines (e.g., sociology, history). While their priorities differ, their efforts converge in providing a comprehensive, interdisciplinary understanding of language. Psychological principles are posited by many authors as providing the basis for the structure of languages (e.g., memory capacity, interactive minds). In psychology itself, language is also an important topic. Cognitive, developmental, social, and other fields of psychology investigate verbal representations and verbal behavior (e.g., words and concepts, language development in children). Within some psychology departments, the study of speech comprehension and production (both oral and written) has been institutionalized as a special field called psycholinguistics. However, the intricacies of specific languages (e.g., Dutch, Chinese) usually remain unexplored in psychological research. Expanding expertise pertaining to the grammar, vocabulary, morphology, and literature of a plethora of modern-day, as well as ancient languages has spawned language-specific research and training. This has been institutionalized in language departments that conduct extensive studies on specific languages. Moreover, many language departments pursue a program of linguistics that elaborates general theories of language.

A similar example from the social sciences is ethnology. Culture-specific lifestyles and beliefs have been salient topics in the discussion of human nature and human interactions on all levels of complexity (e.g., parent–child relationships, government of states). Much emphasis has been placed on comparing 'primitive' with 'civilized' lifestyles, in the hope that such studies would reveal the evolution of humankind. Frequently, cultural studies have been assigned to psychology, and psychological principles were invoked to explain diverse cultural phenomena. Jung (see Sect. 1.1), for instance, has found evidence for archetypes in folk mythology. Widespread interest in foreign and primitive cultures has inspired research expeditions to distant countries. Such projects have provided invaluable information on human cultures (including their geography, language, technical skills, etc.). Moreover, they have led to the development of advanced research methods. A result of these efforts has been the establishment of fields such as ethnology and anthropology. Related disciplines, including psychology, have contributed to and benefited from ethnological research. Within psychology, there still exists the subdivision of cross-cultural psychology.

2.3 The Natural Context: Psychology, Biology, Physiology

Some psychological research has been initiated during the nineteenth century by biologists and physiologists in order to study brain functions which eluded the research methods of the time. Physiological psychology was promoted as a new scientific discipline that employed experiments to investigate higher mental functions (e.g., sensation, attention, learning). By the end of the twentieth century, brain research, neurophysiology, and biochemistry have come closer to the explanation of how cognition and behavior are represented in and controlled by the nervous system. Above all, brain research has determined the location of psychologically relevant neural functions (e.g., centers for speech production and comprehension), and has elucidated the organization of neural networks that interact during psychological processes (e.g., links between cortical correlates of perceptions and subcortical correlates of emotions in the appraisal process). Equally vital is the understanding of the hormonal control of arousal in the nervous system (e.g., due to adrenaline discharge), including the regulation of the sleep–wake cycle, alternations between activity and relaxation, and mood changes. In order to continue the research in this area, a new field called neuropsychology has been established.

The study of individual differences was instigated by efforts to ascertain the inheritance of character traits and skills (including deviant character and ingenuity). Studies on phenotypes have demonstrated the substantial influence of genetic factors on personality and behavior. Biological research programs are close to unraveling the mystery of the genetic code that is the substrate of inherited cognitive and behavioral dispositions.

Ethology deserves to be mentioned as a traditional interface between psychology and biology. It focuses on species-specific behavior with an emphasis on adaptation to the environment. Central issues are instincts as long-term, species-specific adaptations of drives and related behavior patterns (e.g., sex drive and mating behavior), and key stimuli eliciting instinctive behavior (e.g., the scheme of childlike characteristics which triggers affectionate behavior). Currently, ethological studies in biology outnumber

those in psychology, as the latter discipline has turned to questions of individual learning and socialization. However, psychology, in particular animal psychology and biological psychology (see Sect. 1.2) has continually incorporated ethological findings into its theories. The existence of neural and hormonal correlates of instinctive behavior has long been postulated, and neurophysiological research is in the process of providing the evidence for these assumptions.

2.4 The Context of Healthcare: Psychology and Medicine

Health is a physical and a mental state. Physical ailments have a psychological impact, and psychological disorders may cause physical diseases. The role of psychosocial factors has long been recognized in medicine, and special categories of neurotic and psychosomatic diseases have been defined. Neuroscience has demonstrated the effects of psychological stress on the immune system, and the new field psychoimmunology has been established.

The treatment of neurotic and psychosomatic disorders has become the domain of certified therapists trained in clinical psychology (see Sect. 1.5) rather than medicine. These therapists implement psychological diagnostic procedures and therapies such as cognitive and behavior therapy. A rapidly growing field in the health services is health psychology that designs, conducts, and validates programs of disease prevention (e.g., dental care, breaking nicotine addiction) and rehabilitation (e.g., after myocardial infarctions).

While all these psychological fields center on patients, another approach focuses on physicians: medical psychology. Obviously, diagnostic and therapeutic procedures may engender unconventional and emotional interactions between physicians and patients, and the outcome of treatments may depend on the social skills of physicians. Therefore, medical psychology studies and attempts to enhance doctor–patient relationships. For instance, doctors are advised on how to cope with stressful experiences (e.g., death of patients), and with problems of intimacy (e.g., during urological examinations).

2.5 Conclusions

There exists a tripartition of psychologically relevant research and practice: (a) generic involvement in topics which various fields of psychology (e.g., developmental, social psychology) share with other disciplines (e.g., sociology, pedagogy, biology), (b) specializations in object-centered disciplines (e.g., linguistics, ethnology, neurophysiology), and (c) specializations within psychology (e.g., psycholinguistics, cross-cultural psychology, neuropsychology). A complex network of academic disciplines and social service professions has flourished in recent years, and within this network psychology occupies a respectable position with links to a variety of related fields.

However, the existing interdisciplinary network has also been criticized on the grounds of insufficient cooperation and of causing paralyzing conflicts. In particular, overlapping research in psychology and other independent disciplines has been indicted as a waste of resources. Why must there be linguistics and psycholinguistics, ethnology and cross-cultural psychology, neurophysiology and neuropsychology, organizational psychology and sociology, pedagogy and educational psychology, etc.? There are voices that advocate a re-integration of the presently diverging branches of the humanities, the social and the natural sciences. Some view psychology as the designated leader in this movement. Presently, however, the vision of a department of psychology pervading the entire university is highly unrealistic.

In regard to interdisciplinary integration, the attitudes diverge also within psychology itself. Some psychologists pride themselves in the academic and professional identity which psychology has achieved by its own methodological and theoretical approaches, and therefore stand opposed to a comprehensive integration, while others emphasize the benefits to be gained by merging resources with related disciplines. Two new interdisciplinary fields are exemplary for the movement towards integration: cognitive science and neuroscience. The highly sophisticated methodology of these disciplines (e.g., computer technology, high-resolution measurements of brain processes) may provide answers to fundamental psychological problems and may facilitate the development of revolutionary practical applications (e.g., decoding the memory traces in the brain could advance more effective learning procedures). However, major objections against a full integration of psychology into the new fields of cognitive science and neuroscience remain: mind may be more than a complex web of biological and technical structures and processes.

See also: Behaviorism; Broadbent, Donald Eric (1926–93); Cognitive Psychology: Overview; Developmental Psychology; Freud, Sigmund (1856–1939); Jung, Carl Gustav (1875–1961); Köhler, Wolfgang (1887–1967); Motivation and Actions, Psychology of; Pavlov, Ivan Petrovich (1849–1936); Personality Psychology; Piaget, Jean (1896–1980); Skinner, Burrhus Frederick (1904–90); Social Psychology; Watson, John Broadus (1878–1958)

Bibliography

Hilgard E R 1987 *Psychology in America*. Harcourt Brace Jovanovich, New York

Koch S (ed.) 1959/1963 *Psychology: A Study of Science*. McGraw-Hill, New York, Vols. 1–6

Schönpflug W 2000 *Geschichte und Systematik der Psychologie* (*History and Structure of Psychology*). Beltz/Psychologie Verlags Union, Weinheim, Germany

W. Schönpflug

Psychometrics

Modern test theory makes free use of the idea of a latent variable to explain answer choices in each test item. A latent trait can be transformed by any smooth order-preserving function, and thus does not have the measurement structure that is often mistakenly attributed to it. The notion of a manifold is used to explain this one-dimensional structure. This manifold is not a latent variable, however, and arc length or distance along this manifold from a fixed point is a measure of ability or trait level that is manifest. This arc length measure is suggested as an alternative to latent trait values in order to avoid the confusion that the idea of a latent trait often causes.

1. Introduction

Terms like 'psychometrics' are constantly-moving targets for one seeking a definition, but some frequently encountered features of psychometric data are indicated in the next section. But many readers will think of test scores as prototypical data. What characteristics of the people taking tests do test scores and the data that are used to compute them reflect? Test scores are numbers, but not all the structure of numbers are meaningful representations of something about examinees. Section 3 considers how test scores define nearness or proximity of examinees, and the status of the algebraic operations of summing and differencing. Section 4 relates the psychological concept of a trait to the mathematical notion of a curve in high-dimensional space or a manifold, and considers arc length along this curve as a type of measurement.

2. Definitions and Common Themes

According to one popular dictionary, psychometrics means 'mental measurement.' But we need a better definition, since the status of both mind and measurement as concepts are no longer as unequivocal as they were are the beginning of the twentieth century. This article considers as psychometric most quantitative models for the behavior of humans and other animals that is controlled by the central nervous system.

Defined as broadly as this, it may seem surprising that there are any themes that seem to be common to psychometric research. Nonetheless, there are a few, and they may be motivated by the following questions:

What variables should be observed? In the natural sciences, where seven constants are said to define the universe, which variables to observe is seldom an issue. But in behavioral research, variables are made to measure by researchers, often in the form of tests or scales developed to reflect vaguely defined concepts such as 'depression' or 'ability.' For any popular concept, there are likely to be dozens and even hundreds of such measures available.

What does a variable value mean as a measure? That is, what information can be considered useful in a set of test scores, response times, or other observations, and what can be considered irrelevant? Stevens (1946) proposed a classification into nominal, ordinal, interval, and ratio scales according to the amount of algebraic structure usable in the data. Viewing test scores as nominal scales only treats numbers as labels for examinees. Ordinal scales represent their order structure or ranking. Interval scales also reflect a stable property of the examinees in the ratio of differences, and ratios of scores are meaningful for ratio scales. But so many writers have criticized this system as inadequate that this issue needs revisiting here, at least briefly.

How many variables should be observed? Judging by research reported in widely read journals, a great number. Because it is difficult to decide which measure of, say, intelligence to use, researchers tend to use several, hoping to cover the ground and perhaps to head off complaints from supporters of particular measures. Moreover, access to replicates, usually people, is often expensive and may involve ideological and political sensitivities. Consequently, psychometric data are often high dimensional, especially relative to sample sizes typically used. Researchers may hope that they can compensate for limited sample sizes by observing more variables; but this belief is both naive and dangerous.

Is there some simple low-dimensional structure in the high-dimensional data? Often a one-dimensional structure is assumed, as when choices on tens or even hundreds of test items are reduced to a single test score, but in other cases special psychometric technology is applied to data to reveal both the low dimensionality implicit in the data, and how the data are approximated within this low-dimensional structure. Principal components or factor analysis is, therefore, a standard tool in behavioral research.

3. What does a Psychometric Measurement Mean?

It is fair to say that most researchers don't concern themselves with this question, but it is a hard issue to avoid, since sooner or later numbers are interpreted, and usually by people whose limited exposure to these issues makes them vulnerable to misinterpretation of

the results. Is a change in score from 10 to 5 on a depression scale as important as one from 30 to 25? Probably not, but is there at least some useful information in score changes such as these?

A psychometric measurement system is a mapping from a set of people into some system of numbers. It has a purpose, namely to use some of the structural characteristics of numbers to reflect or represent analogous characteristics of people. For example, assigning a social security number just distinguishes one person from another, but assigning a test score tends to mean much more, and at least is intended to reflect an order relation presumed to hold among people. But numbers, and especially real numbers, are structured in many different ways, and the question is; which of these ways should be considered as reflective of characteristics of people, and which should be ignored?

Mathematics, at least as an applied science, is the study of structured sets. By sets, of course, we mean collections of people, drawn from some population X, including the set of nobody, which we indicate by $\varnothing$. Sets A and B can be combined in two ways: unions $A \cap B$, meaning the people who belong to either A, B, or both; and intersections $A \cup B$, meaning only the people who belong to both.

There are three main types of structure that preoccupy mathematicians:

(a) topological: structure that defines nearness or neighborhoods,

(b) measure or probability: structure that defines size, and

(c) algebraic: rules for connecting two or more elements, that is, identifying certain pairs, triples, etc.

These structures are increasing in level of organization, with topological being in a sense minimal, and algebraic being the most sophisticated.

We shall have a look at topological structure because it seems important in explaining why test scores are so widely accepted as psychometric variables, and is an aspect of measurement that Stevens and other writes have neglected. But measure and algebraic structure are left for others to others to deal with. (See *Memory Models: Quantitative*.) Test scores are by and large accepted by both ordinary citizens and researchers as useful summaries of human behavior. Why this is so, in spite of their murky status among philosophers, mathematicians, and statisticians, probably derives from their topological structure and its relation to analogous properties of behavior that seem to us reasonable.

Topology has a bad reputation for being abstruse, but it deals with a type of structure corresponding to intuitions rooted in experience. Topology is about nearness or neighborhoods defined within a population X of people. One may also think of the topology of X as its *texture*, and in these remarks we will consider whether test scores have gained their accepted place in most cultures because they reflect assumptions of a textural nature about how people vary.

A topology is a family T of subsets drawn from a set X such that:

(a) both nobody $\varnothing$ and everybody X belong to T;

(b) the union of no matter how many sets in T is also in T; and

(c) the intersection of a finite number of sets in T is also in T.

Sets in T are called *open*, and any open set containing a specified person is called a *neighborhood* for that person. For example, a person's neighborhood can be defined as those other persons sharing the same postal code, or having gone to the same university, or inhabiting the same dwelling, or of the same age, or any combination of these characteristics. Or, in this context, as those who have been assigned the same score on a test. The second part of the definition of a topology ensures that neighborhoods can be expanded as far as needed, even to include everybody, and the third says that neighborhoods can be shrunk, but only a finite number of times. What happens when we expand and shrink neighborhoods is the essential practical issue that the topological structure seeks to address.

Since the topology of test scores may reflect something of the topology of the people assigned the scores, let's switch to considering the topology of test scores, and ask whether this reasonably captures something about the population to which they are assigned. Because test scores are real numbers, an open set is defined as the set of test scores between two limits, and this topology is of a specialized variety called *metric*. Of course, test scores also define a topology of the examinees that have taken them, since we can define an open set of examinees as one containing all examinees whose test scores are within specified limits.

Figure 1 illustrates examinee neighborhoods for a score of 2 on a 50-item test, with the neighborhoods being defined by three different test-scoring strategies. The test is simulated, but this does not matter. Nor does it matter that we have chosen the relatively extreme score of 2; neighborhoods for more average scores will show the same features as shown in these three plots, but with less compelling graphics.

The points labeled E display what can be called an *even* topology on examinees. They indicate the number of examinees among 1000 whose scores differ from 2 by varying amounts. Each score separation value therefore defines a neighborhood. Everybody is, for example, no more than 48 points from 2, and the size of these neighborhoods shrinks smoothly as we reduce the score separation, until we see that only a few people are less than 5 points away from 2. We call this an *even* topology because neighborhoods shrink smoothly.

Suppose, now, we arbitrarily add 10 to all scores of 30 or above. This doesn't change the order of the

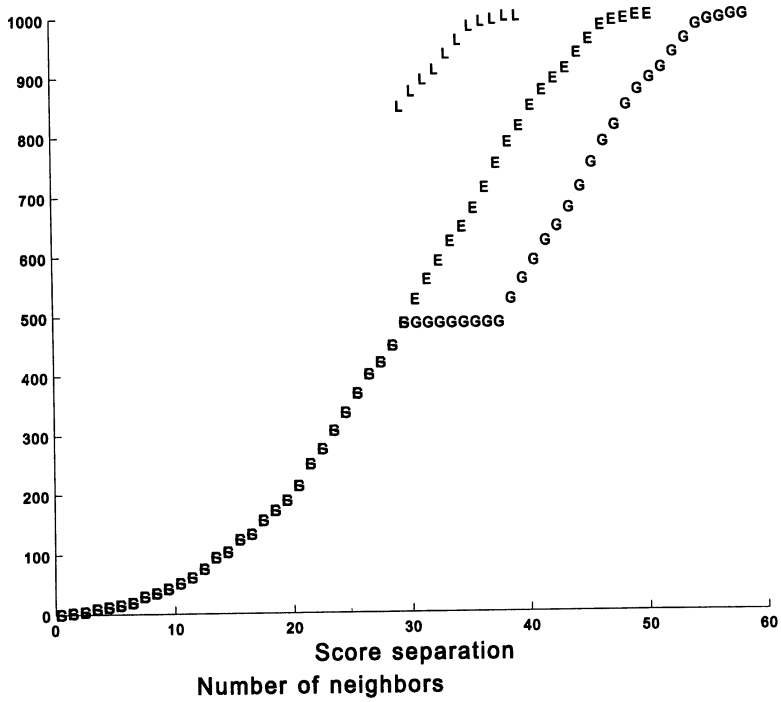

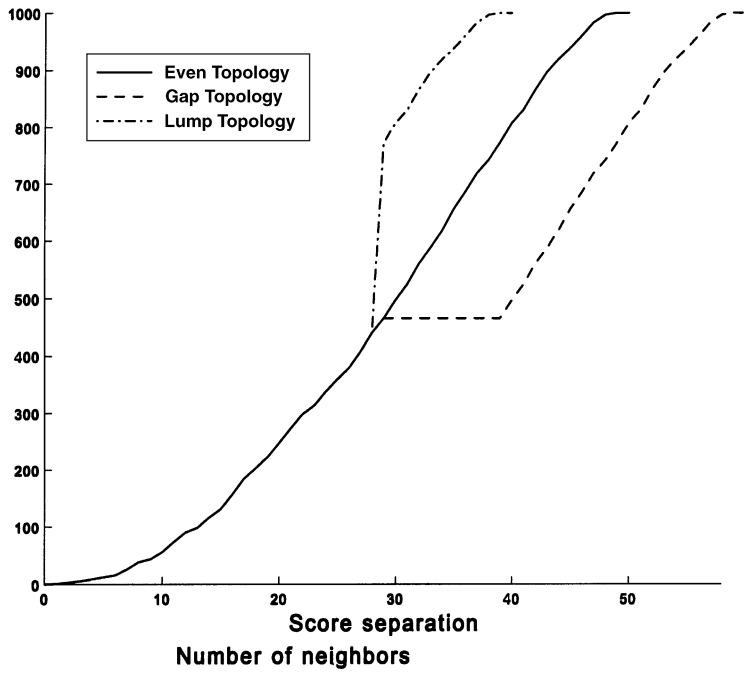

Figure 1
Number of examinees with scores separated by various amounts from a score of 2 according to three different topologies for a 50-item test given to 1000 examinees. Points labeled E have the even topology, corresponding to scores without gaps or lumps, those labeled G have the gap topology, and those labeled L have the lump topology

scores, but it certainly changes the neighborhoods. The points labeled G show that neighborhood size fails to shrink for score separations between 28 and 38 because of the gap in the scores. This is a *gap* topology. If the population being scored had such a gap, it would mean that there would be nobody capable of producing a set of choices on items corresponding to a test score between 30 and 40.

Or, alternatively, let us combine all scores between 30 and 40 into 30, and subtract 10 from all scores above 40. The neighborhood plot is shown as points labeled L, and we see that neighborhood size shrinks abruptly for score separations above 28 with respect to a score of 2. This is a *lump* topology because, put in people terms, there is a large group of examinees who will necessarily produce identical scores. Do people lump up in this way with respect to their behavior? Perhaps sometimes, but mostly not.

Both the gap and lump topologies that are shown in Fig. 1 are defined by test-scoring methods that preserve the order of test scores, but most of us would not be happy attributing either of these textures to subsets of people. We like to imagine that people are evenly distributed in the sense that (a) if we observe enough behavior, only a very small number of people will display identical behavior, and (b) nevertheless all behavior patterns are potentially exhibited by someone. This is the sense in which the even topology determined by most test scoring methods defines a texture for the population that has neither gaps nor lumps, and tends to correspond to intuitions that we have about how performance on a test is related to the ability that it is supposed to reflect.

A transformation that leaves topology unchanged is called *continuous*, a term that we may roughly translate into *smooth*. If there are no gaps or lumps, then, for example, substituting $y = \exp(x)$ or $z = \exp(x)/[1 + \exp(x)]$ for score x will not introduce lumps or gaps, whereas the two transformations indicated above will, and are clearly not smooth.

Stevens' ordinal scale is, then, too simple. It says too little about the texture of either the data or the population generating the data, and something like the concept of an even topology is needed to explain why those who work with test scores are clearly so comfortable with them as descriptions of the behavior of people.

4. *Latent Traits and Manifolds*

Few who have worked with colleagues in departments of psychology or faculties of education, or taught statistics to students in these places, will quarrel with the statement that finding low-dimensional structure in high-dimensional data is perhaps the central statistical challenge in these fields. Consequently, statistical procedures such as stepwise regression, principal components analysis, and structural equation modeling are essential in most contemporary research, and new methods for this problem appear regularly in our methodology journals.

The latent trait is one way to motivate the concept of a low dimensional structure. Test and scale designers often appeal to constructs such as 'proficiency,' 'intelligence,' and 'creativity' in describing what scores on their instruments mean, and freely appeal to a conceptual link between usually unidimensional test scores and these constructs. Of course, it is only test scores, and the choices that generate them, that we can actually observe, and so these constructs are said to be latent or unobservable variables whose structure is reflected in these test scores. Similarly, early factor analysts liked to refer to collections of such latent traits that generate the observed tendency for rather large number of test scores to exhibit low dimensional correlational structure. The five variables in the NEO Personality Inventory, introversion, agreeableness, conscientiousness, neuroticism, and openness, (Costa and McCrae 1985) used in personality research are one of the more successful outcomes of this type of thinking.

Figure 2 displays three *item characteristic curves* for an actual test, these being defined as the relation between the probability $P_i(\theta)$ of getting test item i correct to one's position along a one-dimensional continuum θ. Continuum θ is not observed, but rather inferred from the data. These three curves were estimated by a nonparametric technique described in Ramsay (1991). Probabilities of success on these three items are displayed as functions of four different types of continuum: the first extending over the whole real line; the second containing only nonnegative numbers; the third within the unit interval [0,1]; and the fourth to be explained below. Now, since we are not actually observing θ, there is no deep reason for preferring one of these continua over the others, so that the transformations taking us from one to any other are all legitimate modifications of this model. What are not arbitrary, however, are the probability values $P_i(\theta)$ themselves. These are anchored in the data, and must conform to actual proportions of success for groups of people who have neighboring values of θ or, since smooth transformations preserve topology, neighboring values of either a transformation of θ, or of test scores themselves.

Consequently, the upper panel of Fig. 3 shows what is invariant with respect to transformations of θ. By plotting probabilities against each other, we see that probabilities of success migrate smoothly within the unit cube. The sinuous curve that this figure displays also has the even topology in the sense that it has no gaps or lumps. In this plot, the role of θ or any of its transformations has disappeared, except of course, for its topology, since the curve has inherited the even topology of real numbers.

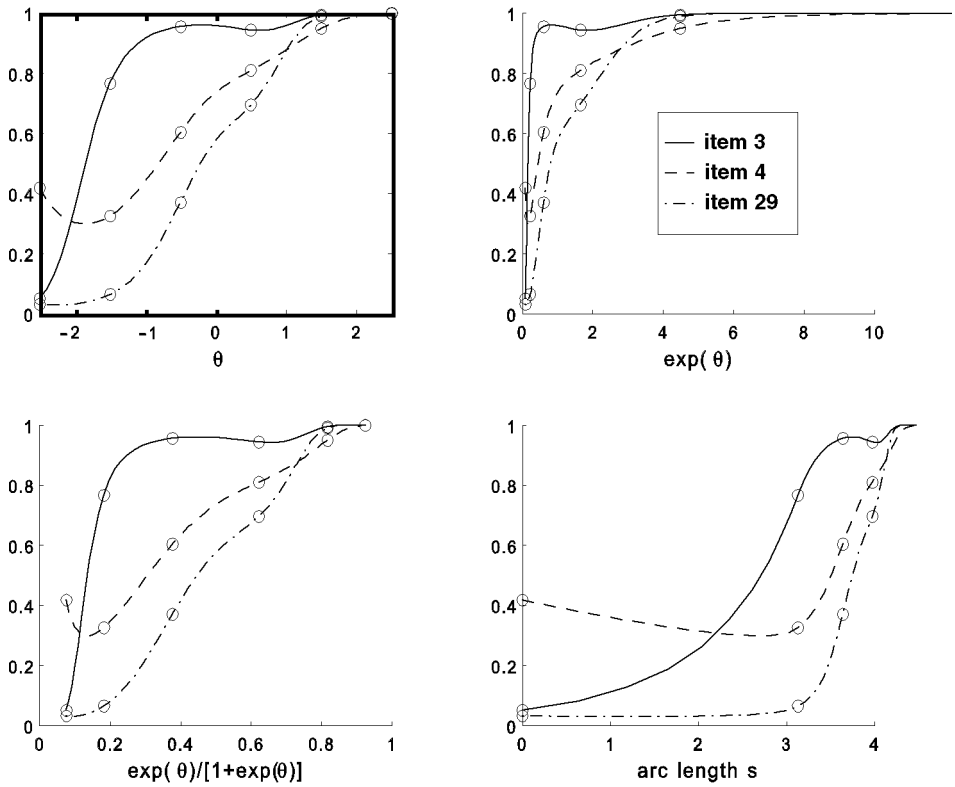

Figure 2
The probabilities of success on three test items as a function of positions on four types of traits. Traits in the upper right and lower left are transformations of that in the upper left, and the trait in the lower right, arc length, is observable rather than latent

The curve in Fig. 3 is a *manifold* in mathematical jargon. Indeed, a smooth manifold, and, if we can compute derivatives of item characteristic curves, a *differentiable manifold*. The subdiscipline dealing with such structures is *differential geometry*. The variable, or any smooth transformation of it, is called a *chart* for the manifold. And with good historical reason, too, since a familiar example of a two-dimensional manifold is the surface of the earth, and latitude and longitude are used to chart this structure, but this is only one example among a large number of alternative charting systems used by cartographers.

What is essential here is that the notion of a manifold captures both the even topological texture discussed in Section B, and something else that is much stronger. Any latent variable charting the manifold can be freely transformed by a smooth and order-preserving function. But there is one charting system that is intrinsic to the manifold, and therefore not at all latent. This is *arc length s*, the distance along the manifold from its beginning, or from any designated point within the manifold. It measures how far someone at the lowest level of performance has to go within probability space to get to a specified performance criterion. Sums and differences of arc lengths can be calculated, and this makes s a ratio scale variable in Stevens' sense. Indeed, it is more, since s cannot be multiplied by an arbitrary positive constant, and it therefore is an *absolute* scale. Each examinee has an s-value, and, since values of s also possess a measure structure, we can approximate the distribution of these values. The probability density function for s is shown at the bottom of Figure 3. It may be argued that arc length s is the *natural* or *intrinsic* measure of test performance. For further discussion of the manifold perspective on modern test theory, see Ramsay (1996). (See also *Fechnerian Psychophysics* for another application of this concept.)

Of course, this is a rather elementary, albeit enormously important, example of a manifold. What comes out of a principal components analysis of p variables is a flat k-dimensional manifold, although there are certainly researchers who have considered how to construct curved manifolds. Multidimensional

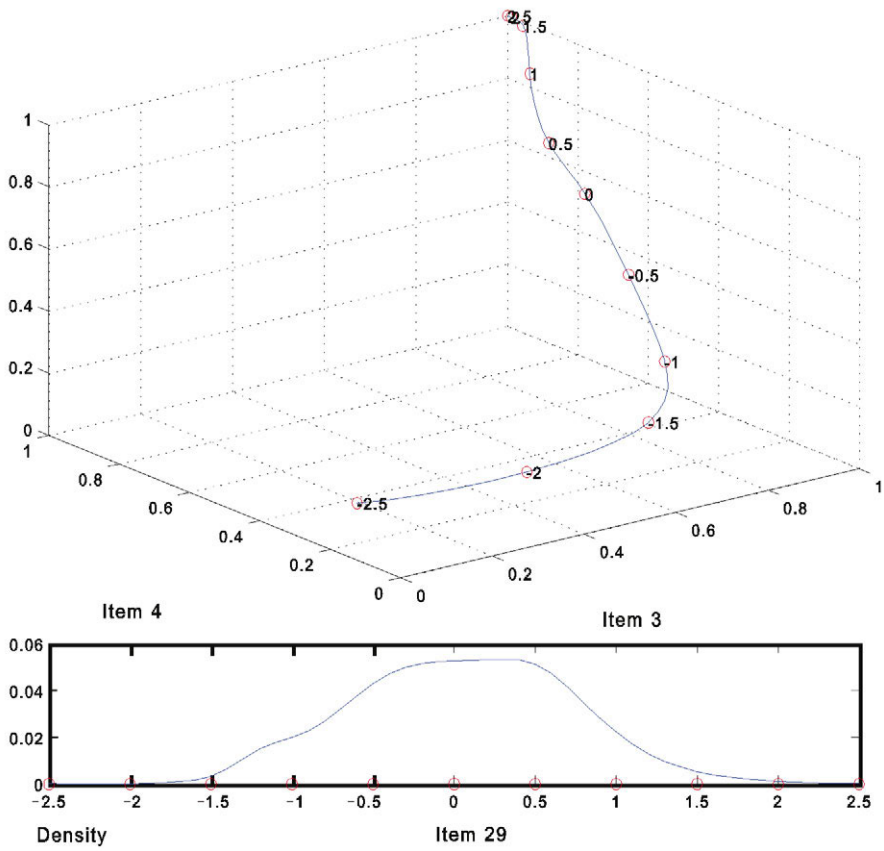

Figure 3
The curve in the upper plot results from plotting probabilities of success for the three items in Fig. 2 against one another. This curve is called a probability manifold. Numbers indicate various trait values on the scale of θ indicated in Fig. 2. The lower panel displays of the distribution of trait values

scaling uses distance information, which is essentially topological, to identify similar flat structures, but within which interesting curved substructures are often easily visible.

The advantages of manifolds, as opposed to latent variables, are:

(a) Manifolds are manifest. One can understand how they relate to the data.

(b) Manifolds have topological, measure-theoretic, metric, and algebraic structures that justify using operations such as computing derivatives, for example, to measure the amount of information conveyed by an item about an examinee.

(c) Manifolds do not confuse people. Nobody is misled in using sums, differences, and other operations within these manifolds, whereas the same operations applied to latent variables do not survive the legitimate (i.e., continuous, differentiable, etc.) transformations that may be applied to them.

(d) There is an established, and even rather old, body of mathematical theory, namely differential geometry, that can be used to develop new methods for data analysis.

It may be that we need to replace models expressed in terms of latent traits by explicit references to the manifolds that they implicitly define.

See also: Classical (Psychometric) Test Theory; Combinatorial Data Analysis; Conjoint Analysis Applications; Dimensionality of Tests: Methodology; Measurement Theory: History and Philosophy; Measurement Theory: Probabilistic; Memory Models: Quantitative; Multidimensional Scaling in Psychology; Scaling and Classification in Social Measurement; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Costa P T Jr., McCrae R R 1985 *The NEO Personality Inventory Manual*. Psychological Assessment Resources, Odessa, FL

Hambleton R K, Swaminathan H 1991 *Fundamentals of Item Response Theory*. Sage, Newbury Park CA
Lord F M 1980 *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, Hillsdale, NJ
Ramsay J O 1991 Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika* **56**: 611–30
Ramsay J O 1996 A geometrical approach to item response theory. *Behaviormetrika* **23**: 3–17
Stevens S S 1946 On the theory of scales of measurement. *Science* **103**: 677–80
van der Linden W J 1997 *Handbook of Modern Item Response Theory*. Springer, New York

J. O. Ramsay

Psychoneuroimmunology

In defending the organism against the invasion of foreign material, the immune system is critical for the maintenance of health. In times past, the immune system was considered a self-regulating, autonomous agency of defense. Research conducted since the mid-1970s, however, has revealed that immunoregulatory processes are influenced by the brain and that neural and endocrine functions and behavior are influenced by the immune system. Psychoneuroimmunology, then, is the study of the interactions among behavior, neural and endocrine function, and immune system processes (Ader 1981a). The basic premise underlying psychoneuroimmunology is that adaptive responses reflect the operation of a single, integrated network of defenses. While each component of this defensive network evolved to serve specialized functions at special times, each also monitors and responds to information derived from the others. Thus, it is not possible to achieve a complete understanding of immunoregulatory processes without considering the organism and the internal and external environment in which immune responses take place. The evidence for brain-immune system interactions, with particular but selected emphasis on the role of behavioral and psychosocial factors in modulating immune function, is the subject of this article.

First, however, a brief reminder that the immune system's defense of the organism against foreign, 'nonself' material (antigens) is executed by white blood cells, primarily T and B lymphocytes, capable of responding clonally to antigens and retaining the 'memory' of that encounter. Antibody-mediated immunity refers to the exposure of antigens to bone-marrow derived B cells which produce antibodies that protect the organism against extracellular infectious agents. Cell-mediated immunity is provided by thymus-derived T cells that defend against intracellular parasitic and viral infections. An integrated immune response actually involves interactions among specialized subpopulations of T cells (e.g., helper, cytotoxic), B cells, other white blood cells such as macrophages, and substances (cytokines) secreted by activated immune cells. Lymphocyte proliferation in response to *in vitro* stimulation by mitogens (plant lectins) that preferentially influence T and/or B cells is a measure of the capacity of these lymphocytes to divide in response to nonspecific foreign stimuli. Natural killer (NK) cells are large, granular lymphocytes that are implicated in tumor metastases and in the recognition and defense against viral agents without having had prior experience with that antigen.

1. Historical Perspective

The first sustained program of research on brain-immune system interactions was initiated by Russian investigators in the 1920s. Derived directly from a Pavlovian perspective, these investigators studied the conditioning of immune responses. There was reliable evidence for conditioned alterations in nonspecific immunologic reactions; the evidence for a conditioned change in antibody production was less convincing (Ader 1981b). Early studies on the effects of brain lesions on anaphylactic responses began in the 1950s and were pursued by investigators in the former Soviet Union and in the United States who studied the effects of hypothalamic lesions and electrical stimulation on several parameters of immunologic reactivity.

In the 1950s and 1960s, Fred Rasmussen, Jr., a virologist, and Norman Brill, a psychiatrist (the first such interdisciplinary collaboration), began a series of studies on the effects of stress (avoidance conditioning, physical restraint, electric shock, social 'crowding') on susceptibility to experimentally-induced infectious diseases (e.g., Rasmussen et al. 1957). Depending on the nature of the stressor, these stimuli increased susceptibility to some viral infections and decreased susceptibility to others. These experiments stimulated only a few other investigators to study the effects of early life experiences and stressful psychosocial circumstances on susceptibility to infectious and neoplastic diseases in animals.

In the 1970s, things began to change. Several dramatic observations led to sustained research on the relationship between the brain and the immune system. Hadden et al. (1970) documented the existence of receptors on lymphocytes, providing the first observations linking the immune system to the sympathetic nervous system; Ader and Cohen (1975) demonstrated behaviorally conditioned suppression of the immune system; Bartrop et al. (1977) described immunologic effects associated with the bereavement that followed the death of a spouse; Besedovsky and Sorkin (1977) began their studies of a neuroendocrine-immune system network and the effects of immune responses on neural and endocrine function; the sympathetic innervation of lymphoid tissues was documented

(Felten and Felten 1991); and Blalock and Smith (1980) showed that lymphocytes, themselves, were capable of producing neuropeptides. These seminal studies provided the impetus for new research on receptors for hormones and neurotransmitters on lymphocytes, the effects of sympathetic nervous system (SNS) activity on immunity, conditioned immune responses, depression and immunity, the effects of stressful life experiences on immune function and resistance to disease, psychosocial factors in the progression of AIDS, and immunologic effects on behavior.

2. *Brain–Immune System Communication*

Two pathways link the brain and the immune system: autonomic nervous system activity and neuroendocrine outflow via the pituitary. Primary (bone marrow, thymus) and secondary (e.g., spleen, lymph nodes) lymphoid organs are innervated with sympathetic nerve fibers that form close neuroeffector junctions with lymphocytes and macrophages and release a variety of neuropeptides. The stimulation or interruption of these connections can influence immune responses. Chemical sympathectomy, for example, suppresses antibody- and cell-mediated immunity; hypothalamic lesions generally inhibit and stimulation potentiates immunologic reactivity.

Conversely, activation of the immune system is accompanied by changes in hypothalamic, autonomic, and endocrine processes. There is, for example, an increase in the firing rate of neurons in the hypothalamus that corresponds to the time of peak antibody production; sympathetic activity (turnover of norepinephrine) is increased in the spleen and hypothalamus; and immune responses initiated by viral infections are associated with increases in circulating levels of ACTH and corticosterone.

Because lymphocytes carry receptors for a variety of hormones and neuropeptides, the cellular interactions that mediate humoral and cellular immunity can be modulated by the neuroendocrine environment in which these responses occur. Lymphocytes bear receptors for corticotropin-releasing factor (CRF), ACTH, and endogenous opioids. There are direct immunomodulatory effects of CRF and ACTH, but their major *in vivo* effects are exerted through interactions with other hormones and immune system products, the most conspicuous of which is the release of adrenocortical steroids. There is an extensive literature documenting the anti-inflammatory and immunosuppressive effects of corticosteroids—at least in pharmacologic doses. At physiologic doses, adrenocortical steroids are essential for normal immune function and, depending on steroid concentration, antigen concentration, and the temporal relationship between hormonal and antigenic stimulation, corticosteroids can be immunoenhancing.

Recent data have also shown that lymphocytes activated by antigenic stimuli are capable of producing neuropeptides and hormones. Thus, the brain is capable of detecting and responding to signals released by an activated immune system. Cytokines, nonantibody messenger molecules released by activated immune cells, besides regulating cellular interactions, are another means by which the immune system communicates with the CNS and, thereby, influences behavior. The precise site(s) at which cytokines act within the brain has not been elaborated. Nonetheless, it is known that cytokines are electrophysiologically, endocrinologically and behaviorally active. Symptoms of sickness, for example, decreased activity, loss of interest in the environment, reduced food intake, fatigue, and cognitive disorders and affective changes can be induced in healthy subjects by the central or peripheral injection of proinflammatory cytokines. Indeed, the recent therapeutic use of interferons in human disease has been associated with neurologic and psychiatric side effects (Meyers 1999).

At the neural and endocrine levels, then, there is abundant evidence of interactions between the brain and the immune system.

3. *Behavioral Influences on Immune Function*

3.1 *Effects of 'Stress' on Immune Function*

The link between behavior and immune function goes back to the earliest observations of a relationship between psychosocial factors and susceptibility to immunologically mediated disease processes. There is compelling evidence that psychosocial factors influence the onset or severity of rheumatoid arthritis and other autoimmune diseases, diseases effected by immune reactions generated against an individual's own tissues. Similarly, stressful life experiences are associated with increased susceptibility to infectious diseases and the reactivation of latent viral infections. The death of a spouse, for example, has been associated with depression and an increased morbidity and mortality in response to diseases presumed to involve immune defenses. Bereavement and/or depression is also accompanied by a depression of immune function (reduced lymphoproliferative responses, impaired NK cell activity). In none of these instances, however, has it been demonstrated that these health effects are the result of stressor-induced changes in immune function (see *Stress and Health Research*).

Other 'losses,' such as marital separation and divorce and other chronic stressors such as caregiving for disabled patients, have also been associated with decreases in lymphocyte function and NK cell activity and a depressed antiviral response to an influenza immunization. In individuals seropositive for Epstein-Barr virus (EBV), there is also an increase in anti-EBV

antibody titers reflecting compromised cell-mediated immune system control over the latent virus. Even more seemingly innocuous circumstances such as taking examinations results in a transient increase in levels of distress and a corresponding decrease in lymphoproliferative responsiveness, NK cell activity, percentage of helper T-lymphocytes, and interferon production by stimulated lymphocytes. Exam periods are also associated with elevated EBV titers and an increase in the incidence of self-reported symptoms of infectious illness.

Clinical studies have described an association between some immunologically mediated diseases and clinical depression. Multiple sclerosis (MS), a T cell mediated autoimmune disease, is associated with depressive states. Systemic lupus erythematosus (SLE) is a B cell mediated autoimmune disease in which depression is frequently reported. Depressive features have also been described in HIV-infected individuals. Thus, considerable attention has been given to the effects of clinical depression on immune function. Based on a meta-analysis of the literature (Herbert and Cohen 1993), depression is reliably associated with a decline in both enumerative and functional measures of immunity such as lymphoproliferative responses and NK cell activity. Depressed patients also show elevated antibody titers to herpes simplex virus (HSV) and cytomegalovirus and a decreased ability to mount a specific cell-mediated response to varicella zoster virus. The decline in cellular immunity is comparable to the decline observed in elderly patients who are more susceptible to herpes zoster infection than younger adults (Irwin et al. 1998). Other measures of immune function in depressed patients are characterized by a high variability within and between studies.

Typically, acute laboratory stressors such as unsolvable puzzles or public speaking have been reported to increase numbers of cytotoxic lymphocytes and the number and activity of NK cells along with a decrease in lymphoproliferative responses. These data, too, are characterized by high variability that is, at least in part, related to individual differences in autonomic nervous system reactivity. For example, subjects who are characterized as high SNS reactors in response to an acute laboratory stressor show a greater decrease in lymphoproliferative responses, a greater increase in the number of suppressor/cytotoxic T cells and, perhaps, a greater increase in NK cell activity than low reactors.

It is clear that, while stressful life experiences influence immune function in humans, there are large individual differences due, presumably, to personality and dispositional factors that influence the individual's perception and appraisal of the environmental circumstances, the individual and social resources available to cope effectively with the demands of the environment, and the biological substrate upon which the stressful stimuli are superimposed.

Most of the evidence for stress-induced alterations in immunity comes from basic research on animals. Stressors such as restraint, electric shock, avoidance conditioning, and noise can disrupt immune function and increase susceptibility and/or mortality to a variety of infectious agents (e.g., HSV, influenza, Coxsackie virus, Sindbis virus) or allow an otherwise inconsequential exposure to a pathogen to develop into clinical disease (Moynihan and Ader 1996, Sheridan et al. 1994).

In contrast, though, some of these same stressors *decreased* susceptibility to poliomyelitis virus infection, experimental allergic encephalomyocarditis and *Plasmodium berghei*, a rodent malaria. Stressful life experiences also exert a protective effect on the development of experimental allergic encephalomyelitis, a murine model of MS. In the case of experimentally induced arthritis, increased or decreased susceptibility depends on the experimental model of arthritis and the nature of the stressor.

Prenatal and early life experiences such as interruptions of mother-young interactions, the social environment in which animals live, exposure to predators or pheromones emitted by stressed conspecifics, and subjecting animals to environmental conditions over which they have no control (e.g., physical restraint, electric footshocks, forced exercise) are among the psychological and social manipulations that alter neuroendocrine states and modulate immune responses. Antibody production is generally suppressed by such stressful stimulation.

Although differential housing (frequently mislabeled as 'crowding' or 'isolation') may not qualify as a stressor, the manner in which animals are housed or a change in housing conditions is sufficient to influence the antibody response to different antigens. Primary and/or secondary antibody responses in rodents are suppressed by restraint or electric shock stimulation, depending on the nature of the antigen and the timing of the stressor in relation to the antigenic stimulation. Other stressors, however, are associated with enhanced antibody production. Similarly, cell mediated processes (e.g., graft rejection, delayed type hypersensitivity reactions) and nonimmunologically specific responses (e.g., lymphocyte proliferation in response to mitogens, NK cell activity, cytokine production) generally are suppressed in animals subjected to stressors but, again, the nature and timing of the stressor, the particular response being measured as well as the species, strain and sex of the animal will determine the direction and magnitude of the effects.

A few attempts have been made to discriminate between stressors over which the individual does or does not have control, for example, avoidable vs. unavoidable, escapable vs. inescapable, signaled vs. unsignaled electric shock in animals. Although differences have been observed in some studies, the results have been contradictory. In humans, too, studies involving control—or perceived control—over stress-

ful stimulation have yielded inconsistent results. Given the number of variables identified as having an effect on the assessment of immunologic reactivity, however, there are insufficient data to warrant any conclusions at this time.

While many hypotheses have been proffered, the neuroendocrine mechanisms presumed to mediate the effects of stressful life experiences on immune function remain elusive. For one thing, our understanding of neuroendocrine-immune system interactions under normal and stressful conditions is incomplete. Glucocorticoids, for example, are generally immunosuppressive, and it is generally assumed that corticosteroid elevations, the most common manifestation of the stress response are responsible for the immunosuppression associated with stressful life experiences. There are many examples of stress-induced, corticosteroid-mediated alterations of immune responses. But, there are very many other observations of stress-induced alterations of immunity that are independent of adrenocortical activity. Intermittent and continuous schedules of inescapable electric footshocks, both of which elevate corticosterone levels, effect an analgesic response to subsequent footshock. Only the intermittent shock, however, is an opioid-mediated analgesia and only the intermittent footshock suppresses NK cell activity, illustrating the immunomodulating potential of endogenous opioids.

Even without being able to detail the extensive literature, it is clear that the response to stressful life experiences involves complex neural, endocrine, and immune response interactions (Glaser and Kiecolt-Glaser 1994, Rabin 1999). Different stimuli commonly referred to as stressors—and commonly thought to elicit a representative 'stress' response—can have different effects on the same immune response. Conversely, the same stressor can have different effects on different immune responses. Also, the intensity and chronicity of the stressor are among the several parameters of stimulation which influence immune responses. In general, then, we have an incomplete picture of the direction, magnitude, and duration of the effects of stressful life experiences on immunity. We do, however, know something about the factors that contribute to the observed variability. They include:

(a) the quality and quantity of stressful stimulation;

(b) the capacity of the organism to cope effectively with stressful circumstances;

(c) the quality and quantity of immunogenic (or pathogenic) stimulation;

(d) the temporal relationship between stressful stimulation and immunogenic stimulation;

(e) the parameters of immune function and the sample times chosen for measurement;

(f) the experiential history of the organism and the prevailing social and environmental conditions upon which stressful stimulation and immunogenic stimulation are superimposed;

(g) a variety of host factors such as species, strain, age, sex, nutritional state and current disease; and

(h) interactions among these several variables.

3.2 *Conditioning Effects*

Behaviorally conditioned alterations of immune function provide one of the more dramatic lines of evidence linking the brain and the immune system. In a prototypical study (Ader and Cohen 1975), a novel, distinctively-flavored drinking solution (e.g., saccharin), the conditioned stimulus (CS) is paired with the injection of an immunomodulating drug, the unconditioned stimulus (UCS). When animals are subsequently injected with antigen, some of them are re-exposed to the CS, alone. When the immunosuppressive drug, cyclophosphamide (CY), was used as the UCS, conditioned animals re-exposed to the CS showed an aversion to the saccharin solution and an attenuated antibody response relative to conditioned animals that were not re-exposed to the CS or nonconditioned animals that were exposed to saccharin. Several independent investigators have verified and extended these findings.

The effects of conditioning have been consistent under a variety of experimental conditions (Ader and Cohen 2001). Studies have documented the acquisition and/or extinction of conditioned nonspecific host defense responses such as NK cell activity and different cell-mediated as well as antibody-mediated immune responses using different unconditioned and conditioned stimuli. Also, conditioning is not limited to changes associated with taste aversion learning; several studies have described the acquisition and extinction of conditioned 'stress' effects. Moreover, there is no consistent evidence of a relationship between conditioned behavioral and conditioned immune changes. Taste aversions can be expressed without concomitant changes in immune function, and conditioned changes in immunologic reactivity can be obtained without observable avoidance conditioning. Also, the available data indicate that conditioned immunosuppressive responses cannot be attributed to either stress-induced or conditioned elevations of adrenocortical steroids.

The acquisition and extinction of a conditioned *enhancement* of immunologic reactivity have been observed using antigens rather than pharmacologic agents as UCSs. In the first such study (Gorczynski et al. 1982), CBA mice were anesthetized, shaved, grafted with skin from C57BL/6J mice, and left bandaged for 9 days. In response to the grafting of allogeneic tissue, there was, as expected, an increase in the number of precursors of cytotoxic T lymphocytes (CTLp) that could react against the foreign (nonself) tissue. The conditioning procedures were repeated when CTLp numbers returned to baseline levels. On the fourth such trial, the procedures were repeated again, but the tissue graft was omitted. Approximately half the

animals in each of several experiments showed an increase in CTLp in response to the grafting procedures alone. Half of the 'responder' mice were given additional conditioning trials and half received extinction trials (sham grafting). When tested again, all the animals given additional conditioning trials showed the conditioned increase in CTLp whereas none of the 'responder' mice that experienced unreinforced trials showed a conditioned response.

In another study on conditioned immunoenhancement, mice were immunized repeatedly with low doses of a common laboratory antigen following exposure to a gustatory CS. Three weeks after the last trial, half the animals in each of the several groups were re-exposed to the CS, alone, or were re-exposed to the CS in the context of a low-dose booster injection of the same antigen. The latter groups were added to determine if conditioning could be more sensitively indexed by asking if a CS could potentiate the effects of a minimally effective dose of antigen rather than initiate the production of antibody by itself. In this experiment, neither the CS nor the booster dose of antigen, alone, were sufficient to elicit a robust antibody response. However, antibody production was enhanced when conditioned mice were re-exposed to the CS in the context of re-exposure to a minimally immunogenic dose of that same antigen. Recently, others have observed a conditioned enhancement of antibody production after a single conditioning trial and without the need for a booster injection of antigen.

Data on conditioning in humans is limited. The anticipatory (conditioned) nausea that frequently precedes immunosuppressive cancer chemotherapy was also associated with anticipatory suppression of proliferative responses to T cell mitogens. In healthy subjects, the pairing of a distinctive flavor with injections of adrenaline resulted in an enhancement of NK cell activity when subjects were subsequently re-exposed to the CS alone. Preliminary data on cell-mediated immunity derived from a study in which subjects were repeatedly injected with tuberculin obtained from a green vial and with saline drawn from a red vial. On the test trial, the contents of the vials were switched. Saline administered to the arm previously treated with tuberculin did not elicit a skin reaction, but there was an attenuated response in the arm previously injected with saline. These studies lend credence to earlier anecdotal reports of what would now appear to be examples of conditioned alterations of immunologic reactivity in human subjects or patients.

4. *Biologic Impact of Behaviorally-induced Alterations of Immune Function*

Many studies have related stressful life experiences to changes in immunity and/or changes in the onset or progression of disease (Biondi and Zinnino 1997), and these correlational studies are, indeed, provocative. The association between stressful life events and increased morbidity and the association between stressful life events and alterations in immune function do not, however, establish a causal chain linking stressful life events, immune function, and health or disease. There are, at present, few, if any, human studies in which altered resistance to disease has been demonstrated to be a direct result of biologically relevant changes in immune function induced by psychosocial or stressful life experiences. Such data are being collected from animal models of spontaneously occurring disease or models in which it is possible to precipitate adaptive responses to immunogenic stimuli experimentally.

In mice infected with influenza virus, for example, there is a delay in the production of virus-specific antibody in restrained relative to unrestrained animals. In mice inoculated with HSV, prolonged periods of restraint or repeated periods of electric shock suppressed NK cell activity and the development of HSV-specific cytotoxic T lymphocytes. Also, higher titers of infectious virus were recovered from the restrained mice. Other recent studies have shown that psychosocial factors can reactivate latent HSV. While restraint was ineffective, disruption of the social hierarchy within a colony of mice by repeatedly introducing an intruder into a group cage increased aggressive behavior, activated the hypothalamic–pituitary–adrenal axis (HPA), and resulted in reactivation of latent HSV in more than 40 percent of the latently infected animals.

In an NK-sensitive tumor model in which lung metastasis is related to NK cell function, it has been found that, depending upon when the stressor was imposed in relation to the effect of NK cells on the metastatic process, forced swimming, acute alcohol intoxication and social confrontations decreased NK cell activity and increased lung metastases. This finding is consistent with strain, age, and sex differences in NK cell activity and the development of lung metastases.

In defending the organism against infection, inflammatory processes are an essential component in the healing of wounds. Both the SNS and the HPA axis are capable of modulating inflammatory and immune processes. Thus, stressful life experiences would be expected to influence wound healing as, in fact, it does. Chronically stressed caretakers of Alzheimer's patients took 24 percent longer than control subjects to heal a standard, experimentally produced wound in the skin, and there was a 40 percent delay in wound healing in dentistry students tested just before an examination period (when perceived stress was high) compared to healing during a summer vacation (when perceived stress was low). Paralleling this human research, recent animal studies have found that mice restrained for 3 days before and 5 days after wounding showed a reduced inflammatory response, an elevated

corticosterone level and a dramatic delay in the rate at which the wounds healed.

These animal models of wound healing, NK-sensitive tumors and reactivation of latent virus infection will enable research designed to examine the mechanisms underlying the effects of stressful life experiences and determine whether they are, as suggested, due to behaviorally-induced alterations in immune function.

The spontaneous development of a systemic lupus erythematosus-like disease in New Zealand mice was used to assess the biologic impact of conditioned alterations in immune responses since a suppression of immunologic reactivity would be in the biological interests of these animals. Half the weekly treatments received by the lupus-prone mice consisted of saccharin plus an ip injection of saline (the CS) rather than the CS plus an injection of cyclophosphamide. Under this pharmacotherapeutic regimen, the onset of disease was delayed using a cumulative amount of immunosuppressive drug that was not, by itself, sufficient to alter progression of the autoimmune disorder.

Consensual validity for these observations was provided by a study in which re-exposure to a CS previously associated with CY accelerated mortality in conditioned mice in response to a transplanted plasmacytoma. It has also been shown that protection against adjuvant-induced arthritis could be achieved by exposing animals to a CS previously paired with an immunosuppressive drug or with a stressful (immunosuppressive) experience. Among mice previously conditioned by the pairing of a CS with an immunosuppressive drug, re-exposure to the CS following transplantation of allogeneic tissues prolonged survival of the allografts. Clearly, conditioned immunologic effects can have a profound biologic impact on the development of autoimmune disease, neoplastic processes, and graft survival. Thus far, there is only one clinical case study in which a conditioning protocol was successfully used to reduce the total amount of cytoxan used in the treatment of a child with SLE (Olness and Ader 1992).

5. *Conclusions*

'Psychoneuroimmunology' is an interdisciplinary field of study that has developed and now prospers by exploring and tilling fertile territories secreted by the arbitrary if not illusory boundaries of the biomedical sciences. Disciplinary boundaries and the bureaucracies they created are biological fictions that can restrict imaginative research and the transfer and application of technologies. They lend credence to Heisenberg's assertion (1958) that 'What we observe is not nature itself, but nature exposed to our method of questioning' (Ader 1995).

A paradigm shift is occurring, however. Based on research conducted since the mid-1970s, it has become evident that immunoregulatory processes can no longer be studied as the independent activity of an autonomous immune system. We now know of at least two pathways through which the brain communicates with the immune system: autonomic nervous system activity and neuroendocrine outflow via the pituitary. Both pathways generate chemical signals that can be perceived by the immune system via receptors on the surface of lymphocytes and other immune cells. Conversely, an activated immune system generates chemical signals that can be perceived by the nervous system. Thus, there is a bidirectional exchange of information between the brain and the immune system and all immunoregulatory processes take place within a neuroendocrine milieu that is responsive to the individual's perception of and adaptive responses to physical and psychosocial events occurring in the external world.

It should not be surprising, then, that stressful life experiences or Pavlovian conditioning could influence immune function. The conditioned suppression or enhancement of immune responses has been demonstrated in different species and under a variety of experimental conditions. Also, stressful life experiences influence immune function. Both conditioning and stressful stimulation exert biologically meaningful effects in that they appear to be implicated in altering the development and/or progression of immunologically mediated disease processes. While these are reproducible phenomenon, the extent to which one can generalize from one stressor or from one parameter of immune function to another is limited. Also, the direction and/or magnitude of the effects of behavioral factors in modulating immune responses depend on the nature of the behavioral circumstances; the nature of the antigenic stimulation and the temporal relationship between them; the immune response and time at which it is measured; a variety of host factors; and the interactions among these variables.

The neuroendocrine pathways underlying the effects of conditioning or stressful life experiences on immune responses are not yet known. It would appear, though, that the neural and endocrine changes associated with changes in behavioral states and the network of connections between the brain and the immune system provide multiple pathways through which behavioral processes could influence immune defenses. The existence of these bidirectional pathways reinforces the hypothesis that immune changes constitute an important mechanism through which psychosocial factors could influence health and disease.

See also: Endocrinology and Psychiatry; Stress and Cardiac Response; Stress, Neural Basis of; Stress: Psychological Perspectives

Bibliography

Ader R (ed.) 1981a *Psychoneuroimmunology*. Academic Press, New York

Ader R 1981b A historical account of conditioned immunobiologic responses. In: Ader R (ed.) *Psychoneuroimmunology*. Academic Press, New York

Ader R 1995 Historical perspectives on psychoneuroimmunology. In: Friedman H, Klein T W, Friedman A L (eds.) *Psychoneuroimmunology, Stress and Infection*. CRC Press, Boca Raton, FL

Ader R, Cohen N 1975 Behaviorally conditioned immunosuppression. *Psychosomatic Medicine* **37**: 333–40

Ader R, Cohen N 2001 Conditioning and immunity. In: Ader R, Felten D L, Cohen N (eds.) *Psychoneuroimmunology*, 3rd edn., Academic Press, New York

Bartrop R W, Lazarus L, Luckhurst E, Kiloh L G, Penny R 1977 Depressed lymphocyte function after bereavement. *Lancet 1* **i**: 834–37

Besedovsky H, Sorkin E 1977 Network of immune-neuroendocrine interactions. *Clinical and Experimental Immunology* **27**: 1–12

Biondi M, Zannino L-G 1997 Psychological stress, neuroimmunomodulation, and susceptibility to infectious diseases in animals and man: A review. *Psychotherapy and Psychosomatics* **66**: 3–26

Blalock J E, Smith E M 1980 Human leukocyte interferon: Structural and biological relatedness to adrenocorticotropic hormone and endorphins. *Proceedings of the National Academy of Science, USA* **77**: 5972–74

Felten S Y, Felten D L 1991 Innervation of lymphoid tissue. In: Ader R, Felten D L, Cohen N (eds.) *Psychoneuroimmunology*, 2nd edn. Academic Press, New York

Glaser R, Kiecolt-Glaser J K (eds.) 1994 *Handbook of Human Stress and Immunity*. Academic Press, San Diego, CA

Gorczynski R M, Macrae S, Kennedy M 1982 Conditioned immune response associated with allogeneic skin grafts in mice. *Journal of Immunology* **129**: 704–9

Hadden J W, Hadden E M, Middleton E 1970 Lymphocyte blast transformation. I. Demonstration of adrenergic receptors in human peripheral lymphocytes. *Journal of Cellular Immunology* **1**: 583–95

Heisenberg W 1958 *Physics and Philosophy: The Revolution in Modern Science*, 1st edn. Harper, New York

Herbert T B, Cohen S 1993 Depression and immunity: A meta-analytic review. *Psychology Bulletin* **113**: 472–86

Irwin M, Costlow C, Williams H, Artin K H, Levin M J, Hayward A R, Oxman M N 1998 Cellular immunity to varicella-zoster virus in patients with major depression. *Journal of Infectious Diseases* **178S**: 104–8

Meyers C A 1999 Mood and cognitive disorders in cancer patients receiving cytokine therapy. *Advances in Experimental Medicine and Biology* **461**: 75–81

Moynihan J A, Ader R 1996 Psychoneuroimmunology: Animal models of disease. *Psychosomatic Medicine* **58**: 546–58

Olness K, Ader R 1992 Conditioning as an adjunct in the pharmacotherapy of lupus erythematosus. *Journal of Developmental and Behavioral Pediatrics* **13**: 124–5

Rabin B S 1999 *Stress, Immune Function, and Health*. Wiley-Liss, New York

Rasmussen A F, Marsh J T, Brill N Q 1957 Increased susceptibility to herpes simplex in mice subjected to avoidance-learning stress or restraint. *Proceedings of the Society for Experimental Biology and Medicine* **96**: 183–9

Sheridan J F, Dobbs C, Brown D, Zwilling B 1994 Psychoneuroimmunology: Stress effects on pathogenesis and immunity during infection. *Clinical Microbiological Review* **7**: 200–12

R. Ader

Psychopharmacotherapy, Costs of

Psychopharmacotherapy is one of the mainstays of the treatment of psychiatric disorders besides social therapy and psychotherapy. Psychiatric disorders essentially comprise the affective disorders (depression, bipolar, i.e., manic-depressive disorder), the schizophrenias, the anxiety disorders (panic disorder, generalized anxiety, post-traumatic stress disorder, obsessive–compulsive disorder, social phobia), the eating disorders (bulimia and anorexia nervosa), the substance abuse and dependence disorders including alcoholism, the personality disorders, and the dementias (with Alzheimer-type dementia being the cause in about 70 percent of cases). They are highly prevalent (Table 1) and frequently take a chronic or chronic remittent course. Therefore, it is easily understandable that the total costs of psychiatric disorders compare very well with the costs of the leading somatic illnesses (Table 2). The prevalences of depression and of the eating disorders are increasing. The WHO burden of illness study (Murray and Lopez 1997) suggests that depression will be the leading cause of disability in the developed countries in 2020. Psychiatric disorders (except the dementias) manifest in early adulthood thus limiting the abilities of patients for social achievement.

Psychopharmacotherapy essentially comprises seven classes of drugs:

(a) neuroleptics for the treatment of psychotic states, especially the schizophrenias;

(b) antidepressants for the treatment of depression of whatever origin as well as for the long-term treatment of anxiety disorders;

(c) mood stabilizers (lithium, some anticonvulsants) for the short-term treatment of mania and for the long-term relapse and recurrence prevention of affective disorders;

(d) minor tranquilizers (essentially benzodiazepines) for the short-term treatment of anxiety and agitation of whatever origin and of alcohol withdrawal syndromes;

(e) hypnotics (essentially benzodiazepines and benzodiazepine-like drugs) for short-term treatment of insomnia;

(f) anticraving compounds (bupropion for smoking cessation, acamprosate for alcohol abstinence, opiate antagonists for alcoholism and opiate dependence) which, however, are not yet approved in all countries; and

Table 1
Annual rate of behavioral disorders in the USA (adapted from NIMH)

Disorder	Percent afflicted
Anxiety disorders (all)	13
Depressive disorders (all)	10
Alcohol abuse/dependence	7
Drug abuse/dependence	3
Severe cognitive impairment	3
Antisocial personality	2
Schizophrenia	1
Somatization	0.2

(g) cognition enhancers (nootropics such as piracetam, ginkgo biloba) for the nonspecific treatment of dementias and acetylcholinesterase inhibitors specifically for Alzheimer-type dementia.

Concerning the eating disorders, only one drug, the antidepressant fluoxetine, is approved for supportive treatment of bulimia.

Psychopharmacotherapy competes with psychotherapy in the treatment of mild to moderate de pression and the anxiety disorders. The relative merits of both alternatives are subject to debate which is fueled by a lack of scientific data, ideology, and economic interests of the health care providers involved. As a rule of thumb, pharmacotherapy and psychotherapy may be thought not to be alternatives but supplements whenever a psychiatric disorder is moderate to severe, although a synergistic action remains to be definitely proved.

Life expectancy increased dramatically in the twentieth century, and this was accompanied by an increase in the prevalence of age-associated diseases. In the field of psychiatry these are the dementias. Thus, although the number of healthy life years is generally increasing, the prevalence of chronic diseases is increasing in parallel. Consequently, the demand for health care services is increasing. Costs for drug therapy are increasing overproportionately because drug development has become increasingly complicated and costly. The mean investment necessary for developing a single new chemical entity to approval amounts to about US$500 million.

Worldwide, health care is one of the most rapidly growing markets with positive impacts on employment. Nevertheless, governments internationally tend to restrict the growth because at least part of it is publicly funded. The fundamental but unanswered question is what proportion of the gross domestic product society is willing to invest in health care. Obviously, however, most people value health and thus health care very highly. Thus, disorders, i.e., the patients suffering from them, compete for limited resources. Economic criteria have become a third dimension besides efficacy and effectiveness as well as tolerability and safety when selecting between treatment options. It is easily conceivable that psychiatric disorders, faced with persistent stigma, have a hard standing in this competition.

Pharmacoeconomics, a young scientific discipline, has developed a sound scientific basis for the evaluation of the economic dimension (Gold et al. 1996). This is indispensable in the interests of an ethically justifiable allocation of resources. The question is

Table 2
Estimated annual total costs of some psychiatric and somatic diseases (American Diabetes Association 1996, American Heart Association 1998, Davidson et al. 1999, Greenberg et al. 1993, 1999, Whetten-Goldstein et al. 1998)

Disease	Estimated total cost ($US billion)	Year considered
Schizophrenia	33	1991
Depression	43.7	1990
Bipolar (manic-depressive) disorder	25.6	1991
Alcoholism	98.6	1991
Anxiety disorders	42.3	1990
Alzheimer's disease	87.6	1991
Coronary heart disease	43	1990
Coronary heart disease	99.8	1999
Cancer	104	1991
Acquired immunodeficiency syndrome (AIDS)	66	1991
Multiple sclerosis	100.8	1994
Diabetes mellitus	91.8	1992
Hypertensive disease	33.3	1999
Congestive heart failure	21	1999
Stroke	45.3	1999

whether a new, more costly drug allows one to achieve better outcomes so that it is worth its higher costs. In principle, a new drug might even spare money in that the better outcome might help avoid costly, higher-level health care such as inpatient treatment and reduce indirect costs, e.g., by less absenteeism.

1. *Basic Terms of Pharmacoeconomics*

Pharmacoeconomics is the science of measuring *costs* and *outcomes* associated with the use of pharmaceuticals in health-care delivery. Its objective is to improve public health through rational decision making when selecting among alternative therapies, e.g., for formularies and their impact on costs and outcomes. Several methodological guidelines have been published (e.g., CCOHTA 1994).

Outcomes are the possible end result of a particular treatment. Outcomes depend on the *effectiveness* of the treatment under the influence of all the confounding variables in real life. Therefore, costs have to be measured under real-life conditions. In contrast, the experimental *efficacy* trials are designed to eliminate as completely as possible all confounding influences. The outcome parameter chosen should have high everyday relevance, e.g., the rate of full remissions of depression.

Costs are the resources consumed by the illness and by its treatment. *Direct costs* are those directly related to treatment such as prescribed medication, self-medication, consulting the physician, blood tests, transportation, psychotherapy, occupational therapy, informal care by relatives, etc., as well as costs for hospital and nursing-home treatment. *Direct costs* include those for diagnosis and treatment of side effects and complications of the initial therapy. There is methodological controversy as to whether the costs of social support (welfare, unemployment benefit, disability pension, housing benefit, etc.) should be included in the direct costs. A complete evaluation of direct costs should consider the opportunity costs, i.e., the lost potential of the resources invested to provide more efficient benefits in other fields.

Indirect or *future costs* are those related to lost productivity due to illness, i.e., due to absenteeism, premature pensioning because of disability, and premature death (e.g., by suicide) as well as those related to health care consumption during the years of life gained owing to the intervention. Concerning psychiatric disorders, costs of violence and crime also have to be considered. *Intangible costs* comprise those attributable to impairments of the quality of life, e.g., of the patients themselves and their relatives and acquaintances involved in the care.

The economic impact of treatment extends into the future. If costs and benefits accrue during different periods, costs and savings have to be discounted for the remaining life-span of the patient sample in question. Discount rates are supposed to reflect the rates of return on private-sector investments and are generally set at 3–5 percent.

If the costs are low and the outcomes excellent the decision in favor of the treatment is obvious as is its rejection if the costs for poor outcomes are high. Therefore, pharmacoeconomic analyses make sense only when both the costs and the quality of treatment are high or low. The economic *efficiency* then is the ratio of costs divided by the outcomes. When allocating resources the drug yielding the most favorable outcomes at the lowest cost will be favored.

2. *Types of Pharmacoeconomic Analysis*

The basic requirement of a pharmacoeconomic analysis is the definition of the perspective, e.g., that of the patient, the health care provider, the health insurance as the payer, or the societal perspective of government. The categories of costs to be considered depend on this perspective. For example, from a provider's perspective, the value of an antidepressant drug under capitation may be expressed in terms of cost per reduction in some depression scale score. From the patient's perspective, the value may be in cost per full treatment response with least side effects. From the societal perspective, the value may be in cost per patient returning to continuous employment at or above the level of previous education. Simply to compare the daily costs of medication does not fulfill the criteria of pharmacoeconomic analysis.

The principle types of analysis include cost-minimization analysis, cost-effectiveness analysis, cost-utility analysis and cost-benefit analysis. *Cost-minimization* analysis (CMA) assumes that the interventions in question yield identical outcomes; this is rarely valid. *Cost-effectiveness* analysis (CEA) compares the costs of alternative treatments in terms of a specific health outcome such as life-years gained, e.g., by reducing mortality from suicide, or responder rates. *Cost-utility* analysis (CUA) adjusts the life-years gained by some measure for quality of life. Thus, the results of CUA are typically expressed as costs per quality adjusted life years (QALYs). Quality of life is assessed either by some rating scale for health-related quality of life or, more typically, by determining the preferences of patients suffering from specific diseases for one or the other disease state as opposed to perfect health where perfect health is set as '1' and death as '0.' In *cost-benefit* analysis (CBA), the QALYs gained are transformed into monetary terms using the human capital approach or willingness-to-pay approach.

As there is no one-to-one relationship between therapeutic intervention and outcome and its financial sequelae, the impact of these uncertainties has to be estimated. This is done by *sensitivity analysis*, where key variables are systematically modified and the economic consequences calculated yielding best- versus worst-case scenarios.

3. *Pharmacoeconomic Study Designs*

There are two principle approaches to pharmacoeconomic analysis: the clinical trial approach (experimental or observational), and the modeling or systems approach. The former means gathering cost data (retrospectively or, preferably, prospectively) in clinical trials. The latter '*in vitro*' approach applies decision-tree analysis to data available from published literature and databases.

Current guidelines favor the clinical trial approach as the 'gold standard.' Modeling, however, can overcome limitations implicit in clinical trials. These include the selection bias concerning the patients recruited in randomized, strictly controlled, and therefore artificial efficacy trials, thus limiting generalizability. Furthermore, their observation period rarely exceeds a few months whereas economic consequences may extend far into the future. Efficacy trials necessarily are underpowered for economic analysis because sample size estimation is based on the critical difference and variability of the efficacy parameter. Obviously, the critical difference and variability of economic data will fundamentally differ. Clinical trials using the mirror image design, i.e., comparing costs and benefits in a defined period before and after administering a specific drug, are of limited value because at least in psychiatric disorders there is a spontaneous decrease of the costs per outcome during the course of illness irrespective of the mode of treatment.

The putatively best experimental design is a prospective, randomized, 'naturalistic,' i.e., non-blind, adequately powered clinical trial. Both principle approaches, the clinical trial approach and modeling, have their weaknesses and strengths so that they are not really exclusive alternatives but complementary.

Friedberg et al. (1999) found that studies sponsored by drug companies were less likely to report unfavorable results than studies funded by nonprofit organizations. This finding could be due to various sources of bias which, however, if present, still have to be identified.

The organization of health-care supply, the health-care utilization patterns, and the prices differ considerably between and even within cultures. Therefore, the results of pharmacoeconomic analyses cannot simply be transferred from one health care market to another.

4. *Pharmacoeconomic Findings with Neuroleptics*

Since the 1990s new neuroleptics have been introduced (amisulpride, clozapine, olanzapine, quetiapine, risperidone, sertindole, zotepine), although not yet approved in all countries. These compounds are called 'atypical' because they have a lower incidence of extrapyramidal side effects (e.g., Parkinsonism) than conventional neuroleptics (e.g., haloperidol, fluphenazine). Moreover, they possibly have superior efficacy concerning cognitive dysfunction and so-called negative symptoms as well as quality of life in schizophrenia. These advantages might improve compliance, which is most important in long-term relapse and recurrence prevention. The costs of daily treatment with atypical neuroleptics exceed those of conventional neuroleptics by a factor of 2–8 depending on the compound and the country considered. On the other hand, neuroleptic medication costs generally make only about 4–8 percent of direct costs.

The pharmacoeconomic analyses actually available for neuroleptics are limited to amisulpride, clozapine, olanzapine, risperidone, and sertindole in schizophrenia. The majority of studies address clozapine, olanzapine, and risperidone. The costs are mostly considered from the payer's perspective, i.e., addressing direct costs. Most are cost-effectiveness analyses, some cost-utility analyses. The majority follow the modeling approach. The economic analyses based on clinical efficacy trials have the limitations discussed above, especially those restricting the analysis to study completers. This is far from the naturalistic, real-world approach necessary. There is only one randomized, naturalistic, adequately powered trial comparing clozapine with conventional neuroleptics, where clozapine significantly reduced the frequency and duration of inpatient treatment; the cost data are to be reported. All in all, the current evidence suggests that clozapine is cost-effective for neuroleptic-refractory schizophrenia and that olanzapine and risperidone therapy may be cost-neutral despite higher medication costs (Revicki 1999). Nevertheless, the access of patients to the new therapeutic chances is restricted on economic grounds, although this varies internationally.

5. *Pharmacoeconomic Findings with Antidepressants*

Since the 1980s, a number of new antidepressants have been introduced (selective serotonin reuptake inhibitors (SSRI): citalopram, fluoxetine, fluvoxamine, paroxetine, sertraline; the selective norepinephrine reuptake inhibitor reboxetine; the serotonin–norepinephrine reuptake inhibitor venlafaxine; the norepinephrine auto- and heteroreceptor blocker mirtazapine; the reversible and selective inhibitor of monoamine oxidase moclobemide), although not yet approved in all countries. Their efficacy is comparable to that of traditional tricyclic antidepressants (TCAs such as amitriptyline, imipramine, desipramine, nortriptyline), but owing to selectivity their tolerability and safety are superior. Safety is of specific concern in view of the suicide risk, although antidepressants are involved only in the minority (< 10 percent) of suicides. These advantages help reduce drop-out rates although to a limited degree (about 25 percent versus

29 percent). Nevertheless, this might be particularly relevant in long-term recurrence prevention which is necessary in about 60 percent of patients in whom antidepressants are indicated. There is a dramatic underrecognition and undertreatment worldwide (Davidson et al. 1999, Tylee et al. 1999). Undertreatment causes a doubling of direct costs due to increased utilization of other medical services (Manning and Wells 1995) and increases indirect costs due to absenteeism. The advantages of innovative antidepressants might increase acceptability and thus utilization of antidepressants. The costs of daily treatment with selective antidepressants, however, exceed those of TCAs by a factor of 2–20 depending on the compound and the country considered. On the other hand, antidepressant medication costs generally constitute only about 3–6 percent of direct costs.

Pharmacoeconomic analyses published cover all innovative antidepressants except reboxetine. All but one (Gould et al. 1995) are restricted to depression. The costs are mostly considered from the payer's perspective. Most are cost-effectiveness analyses, only a few cost-utility and cost-benefit analyses. There are no randomized, open-label, naturalistic, adequately powered trials but some studies analyzed large, complete samples of HMO participants. By far the majority follow the modeling approach. There are some outstanding analyses in which classes of compounds are compared (e.g., CCOHTA 1997, Casciano et al. 2000). Overall, the current evidence suggests that extended-release venlafaxine might be more cost-effective than SSRIs and that the cost-effectiveness of SSRIs might equal or even exceed that of TCAs, despite higher medication costs (Conner et al. 1999, Casciano et al. 2000). However, more research is needed before this can be taken for granted. Accordingly, there is debate as to whether the access of patients to innovative antidepressants should be restricted or whether they should be recommended as first-choice drugs (CCOHTA 1997).

The limited evidence available suggests that psychotherapy is more likely to be cost-effective the more severely afflicted the patients are (Gabbard et al. 1997). The cost of psychotherapy is about 3–6 times higher than that of antidepressant drug therapy. One modeling study (Antonuccio et al. 1997) suggests, however, that cognitive behavioral therapy (CBT) might be more cost-efficient than fluoxetine treatment of depression owing to lower recurrence rates after psychotherapy. In a CUA based on a randomized controlled trial (Lave et al. 1998), nortriptyline was slightly superior to interpersonal psychotherapy (IPT). One meta-analysis (Gould et al. 1995) suggests that CBT is superior to imipramine and cost-effective in panic disorder. Economic analyses addressing the relative merits of pharmacotherapy and psychotherapy are completely lacking for prominent diseases such as obsessive–compulsive disorder, generalized anxiety, and social phobia.

6. *Pharmacoeconomic Findings with Mood Stabilizers*

Mood stabilizers by definition reduce the risk of relapses and recurrences and thus putatively hospitalization. Therefore, they may be expected to reduce direct costs (about 60 percent for hospitalization). Lithium is the only drug reducing mortality essentially by reducing the suicide rate (Ahrens 1997). Only a few pharmacoeconomic analyses have been published. Lithium was more cost-effective than carbamazepine from the payer's perspective in a modeling approach based on retrospectively gathered data from a randomized controlled clinical trial (Dardennes et al. 1999). Valproate was more cost-effective than lithium over 1 year from the payer's perspective in a modeling approach (Keck et al. 1996); the prophylactic efficacy of valproate, however, remains to be established.

7. *Pharmacoeconomic Findings with Tranquilizers and Hypnotics*

There are almost no economic analyses published on tranquilizers or hypnotics. Yuen et al. (1997) found that hospital stay was longer and costs were higher in elderly people if sedative–hypnotic medications were prescribed, even if prescription adhered to established guidelines. Research is needed especially to clarify whether the higher costs of the newer hypnotics (zaleplon, zolpidem, zopiclone) compared with benzodiazepines are economically justified.

8. *Pharmacoeconomic Findings with Anticraving Drugs*

Long-term acamprosate adjunctive to conventional therapy was found to be cost-effective in the treatment of alcoholism in the German setting in a single, retrospective modeling study (Schädlich and Brecht 1998). Opiate antagonists (Howard et al. 1996), opiate agonist (methadone) substitution, and bupropion have not been evaluated economically.

9. *Pharmacoeconomic Findings with Cognitive Enhancers*

There are no economic analyses published on nootropics. The acetylcholinesterase inhibitors donepezil (Foster and Plosker 1999) and rivastigmine (Fenn and Gray 1999) seem to be at least cost-neutral, especially by delaying the need for institutionalization. The current evidence is restricted, however, to a few observational studies and modeling analyses using the cost-effectiveness approach from the payer's perspec-

tive. Economic analyses on galanthamine, an acetylcholinesterase inhibitor and nicotinic agonist expected to be approved, remain to be published.

10. Conclusions and Some Precautions

The innovative neuroleptics and antidepressants not only have the potential for a balance of costs despite higher costs of medication, but might even save direct costs essentially by reducing the risk of institutionalization owing to lower drop-out rates in comparison with traditional compounds. This conclusion, however, is based essentially on decision analyses, i.e., modeling. More research, particularly large-scale, randomized trials under naturalistic conditions, is needed. From a systemic perspective, the potential cost savings can only be realized if the institutions concerned, e.g., hospitals, reduce their capacities accordingly. About 80 percent of the costs of hospitals are common costs, mostly wages and salaries. Whereas in a free health-care market the savings might be realized by market forces, in systems bound in economic welfare a political will is needed to enforce the inevitable dismissal of staff.

See also: Alzheimer's Disease: Antidementive Drugs; Antidepressant Drugs; Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Health Economics; Placebo Studies (Double-blind Studies); Prophylaxis in Psychiatry; Psychiatric Care, Economics of; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychopharmacotherapy: Side Effects; Psychotherapy and Pharmacotherapy, Combined

Bibliography

Ahrens B 1997 Mortality studies and the effectiveness of drugs in long-term treatment. *Pharmacopsychiatry* **30** (Suppl. 1): 57–61

Antonuccio D O, Thomas M, Danton W G 1997 A cost-effectiveness analysis of cognitive behavior therapy and fluoxetine (Prozac) in the treatment of depression. *Behavior Therapy* **28**: 187–210

Bacaltchuk J, Trefiglio R P, Oliveira I R, Hay P, Lima M S, Mari J J 2000 Combination of antidepressants and psychological treatments for bulimia nervosa: a systematic review. *Acta Psychiatrica Scandinavica* **101**: 256–64

CCOHTA 1994 *Canadian Coordinating Office for Health Technology Assessment 1994. Guidelines for the Economic Evaluation of Pharmaceuticals: Canada*. Canadian Coordinating Office for Health Technology Assessment (CCOHTA), Ottawa, ON

CCOHTA 1997. *Canadian Coordinating Office for Health Technology Assessment 1997. Selective Serotonin Reuptake Inhibitors (SSRIs) for Major Depression. Part II. The Cost-effectiveness of SSRIs in Treatment of Depression: Canada.* Canadian Coordinating Office for Health Technology Assessment (CCOHTA), Ottawa, ON

Casciano R, Arikian S R, Tarride J-E, Casciano J, Doyle J J 2000 Antidepressant selection for major depressive disorder: the budgetary impact on managed care. *Drug Benefit Trends* **12**: 6–16

Conner T M, Crismon M L, Still D J 1999 A critical review of selected pharmacoeconomic analyses of antidepressant therapy. *Annals of Pharmacotherapy* **33**: 364–72

Dardennes R, Lafuma A, Watkins S 1999 Prophylactic treatment of mood disorders: a cost-effectiveness analysis comparing lithium and carbamazepine. *Encephale* **25**: 391–400

Davidson J R T, Meltzer-Brody S E, Goldberg D, Wittchen H-U, Magruder K M, Lecrubier Y, Ballenger J C, Tylee A, Schulberg H C, Nutt D J 1999 The underrecognition and undertreatment of depression: what is the breadth and depth of the problem? Discussion. *Journal of Clinical Psychiatry* **60** (Suppl. 7): 10–11

Fenn P, Gray A 1999 Estimating long term cost savings from treatment of Alzheimer's disease: a modelling approach. *PharmacoEconomics* **16**: 165–74

Foster R H, Plosker G L 1999 Donepezil: pharmacoeconomic implications of therapy. *PharmacoEconomics* **16**: 99–114

Friedberg M, Saffran B, Stinson T J, Nelson W, Bennett C L 1999 Evaluation of conflict of interest in economic analyses of new drugs used in oncology. *Journal of the American Medical Association* **282**: 1453–7

Gabbard G O, Lazar S G, Hornberger J, Spiegel D 1997 The economic impact of psychotherapy: a review. *American Journal of Psychiatry* **154**: 147–55

Gold M R, Russell L B, Siegel J E, Weinstein M C (eds.) 1996 *Cost-effectiveness in Health and Medicine*. Oxford University Press, Oxford, UK

Gould R A, Otto M W, Pollack M H 1995 A meta-analysis of treatment outcome for panic disorder. *Clinical Psychological Review* **15**: 819–44

Greenberg P E, Sisitsky T, Kessler R C, Finkelstein S N, Berndt E R, Davidson J R T, Ballenger J C, Fyer A J 1999 The economic burden of anxiety disorders in the 1990s. *Journal of Clinical Psychiatry* **60**: 427–35

Greenberg P E, Stiglin L E, Finkelstein S N, Berndt E R 1993 The economic burden of depression in 1990. *Journal of Clinical Psychiatry* **54**: 405–18

Howard M O, McGuffin R W, Saxon A J, Sloan K L, Walker R D 1996 Clinical issues related to the costs of alcoholism. *PharmacoEconomics* **9**: 134–45

Keck P E Jr, Nabulsi A A, Taylor J L, Henke C J, Chmiel J J, Stanton S P, Bennett J A 1996 A pharmacoeconomic model of divalproex vs. lithium in the acute and prophylactic treatment of bipolar I disorder. *Journal of Clinical Psychiatry* **57**: 213–22

Lave J R, Frank R G, Schulberg H C, Kamlet M S 1998 Cost-effectiveness of treatments for major depression in primary care practice. *Archives of General Psychiatry* **55**: 645–51

Manning W, Wells K B 1995 How can care for depression become more cost-effective? *Journal of the American Medical Association* **274**: 1931–4

Murray C J L, Lopez A D 1997 Alternative projections of mortality and disability by cause 1990–2020: global burden of disease study. *Lancet* **349**: 1498–504

Revicki D A 1999 Pharmacoeconomic studies of atypical antipsychotic drugs for the treatment of schizophrenia. *Schizophrenia Research* **35** (Suppl.): S101–9

Schädlich P K, Brecht J G 1998 The cost effectiveness of acamprosate in the treatment of alcoholism in Germany:

economic evaluation of the prevention of relapse with acamprosate in the management of alcoholism (PRAMA) study. *PharmacoEconomics* **13**: 719–30

Tylee A, Gastpar M, Lepine J-P, Mendlewicz J 1999 DEPRES II (Depression Research in European Society II): a patient survey of the symptoms, disability and current management of depression in the community. *International Clinical Psychopharmacology* **14**: 139–51

Whetten-Goldstein K, Sloan F A, Goldstein L B, Kulas E D 1998 A comprehensive assessment of the cost of multiple sclerosis in the United States. *Multiple Sclerosis* **4**: 419–25

Yuen E J, Zisselman M H, Louis D Z, Rovner B W 1997 Sedative–hypnotic use by the elderly: effects on hospital length of stay and costs. *Journal of Mental Health Administration* **24**: 90–7

J. Fritze

Psychopharmacotherapy: Side Effects

Psychopharmacotherapy refers to the treatment implications of the science of psychopharmacology. Psychopharmacology is the study of drug-induced alterations in the cognitive function and behavior of humans that aims to develop treatment strategies for neurological and psychiatric disorders. Rapid technological progress has led to an abundance of scientific techniques in neurochemical and animal studies on the interaction between drug action and behavior (Feldman et al. 1997).

Most psychoactive drugs exert their effects first by interfering with the communication between nerve cells (neurons). At the molecular level this occurs by their interaction with receptors of neurotransmitters located on the surface or inside the neuron. This step initiates a cascade of events that eventually lead to alterations in the neurotransmitter and/or hormonal systems. Nonspecific drug action at any level in this process results in side effects.

The term side effect implies an undesirable (adverse) drug effect; a particular side effect of a drug for a given condition may be a desirable one for another condition. For instance, benzodiazepines may be prescribed for anxiety disorders, but as a side effect they cause sedation. On the other hand, they may also be prescribed for the treatment of insomnia as a primary indication. Indeed, the psychoactive potential of several drugs was discovered as a result of unexpected but beneficial side effects on psychiatric patients who were taking these drugs for other therapies (see below). The spectrum of drug effects originate largely from their effects on multiple neurotransmitter systems, and the interactions of these systems with each other and other sites of action in the central nervous system.

At times it is a challenge in the practice of psychiatry to be able to determine if a given behavior is a symptom of an illness or a side effect of drug treatment, as the latter may mimic the former. For instance, untreated patients with schizophrenia frequently display decreased spontaneity, lack of motivation, and social isolation. These symptoms resemble some of the extrapyramidal system side effects induced by conventional antipsychotics. A practical management of this potential difficulty involves careful assessments of the patient's functions e.g., sleep patterns, daily living routines, and sexual function prior to treatment.

1. *Historical Perspective*

Treatment with substances that alter mood, thinking, and cognition date back to prehistoric times (Spiegel 1996). Various psychoactive substances have been used in religious ceremonies, rituals, and for healing of pain.

Alcohol is the oldest drug used for its calming and sedative effects. The disinhibition caused by alcohol is described in the first book of the Old Testament. Apart from consumption in social settings, alcohol was introduced for medical use in the nineteenth century as a general anesthetic. The narrow therapeutic index (the ratio of toxic over desired dose) of alcohol limited its usefulness for this purpose.

Opium (*papaver somniferum*), *hellebore*, and *rauwolfia* roots are among the oldest compounds that were extracted from plants in prehistoric and ancient times. Morphine was isolated from opium in the beginning of the nineteenth century and has been available as an analgesic with considerable addiction potential.

In ancient times, different forms of hellebore root were used as emetics, laxatives, and for mental conditions such as mania and melancholy. Risk of seizures limited its use.

Rauwolfia serpentina roots were used by Indians to treat snake bites and also in some mental conditions. Reserpine, a rauvolfia alkaloid was noted to have a calming effect in patients with schizophrenia. Because reserpine induces severe depression as a side effect, its use as an antipsychotic has been abandoned.

Cocaine is an alkaloid derived from the shrub *Erythroxylon coca*; it originated in South America where it was used for its stimulating effects. It was isolated in the nineteenth century and is still used as a local anesthetic. Paranoid delusions, hallucinations, seizures, and dependency are some of the adverse effects of cocaine.

The discovery of psychoactive medications has resulted from the side effect profiles of drugs used in medicine. For instance, the antipsychotic chlorpromazine was initially used to help patients with presurgical preparation and as an adjuvant to anesthetics. Its sedative and calming effects easily found its way to treatment in psychiatric disorders as a primary

indication soon after its antipsychotic properties were discovered. Similarly, iproniazid (an antidepressant of the monoamine oxidase inhibitor (MAOI) type) was originally developed as a drug for the treatment of tuberculosis. In clinical drug trials, patients treated with this drug were observed to become stimulated and cheerful. This observation lead to the investigation of the antidepressant effect of iproniazid.

2. *Genetic Factors and Side Effect Vulnerability*

Side effects are not uniformly observed in everyone. While some patients never develop a certain side effect, others are affected from day one of treatment. The strongest determinant of this interindividual variation is believed to be in the genetic complexion. Variations in the sequence of the DNA called polymorphisms are believed to play a role in expression of a wide spectrum of reactions to drugs: from full response to non-response; minimal to severe side effects. The completion of the Human Genome Project will aid in the identification of patient subgroups according to their drug response patterns. For instance, certain variations in the gene that codes for the enzyme CYP2D6 lead to slow metabolism of some drugs such as tricyclic antidepressants, therefore these patients run higher levels of drugs and are more vulnerable to side effects (Schatzberg and Nemeroff 1998). The application of pharmacogenetics will be the identification of drugs that will be both efficacious and well tolerated for each individual according to his or her genetic makeup.

3. *Assessment and Management of Side Effects*

Compared to the rating scales used to assess the symptoms of diseases, there are relatively few rating scales for the assessment of side effects. The most comprehensive of these is the Systematic Assessment for Treatment Emergent Events which has been designed for use in clinical drug trials. Udvalg for Kliniske Undersogelser Side Effect Rating Scale is also comprehensive and was developed for use in clinical trials and in routine practice. The Rating Scale for Extrapyramidal Side Effects (Simpson–Angus Extrapyramidal System Scale), Barnes Akathisia Rating Scale, and Abnormal Involuntary Movement Scale were developed to measure the extrapyramidal system side effects, akathisia, and dyskinetic movements, respectively, that are associated with antipsychotic drug use (American Psychiatric Association 2000). Even though each side effect ultimately requires individual management, a comprehensive approach is vital for general management. Such an approach involves a baseline assessment, careful choice of medication, patient and family education, minimization of the number of drugs (to decrease potential side effects and to avoid drug interactions), single bedtime dosing, the availability of clinicians to discuss the side effects, adjustment in dosage, counteractive drug treatment, and continuous assessment.

4. *Noncompliance*

Among factors that play a role in treatment noncompliance, such as denial of the illness and lack of supervision, one of the strongest predictors is early onset and persistence of side effects. Noncompliance by itself is a predictor of relapse in patients with chronic mental disorders who are recommended to receive lifelong maintenance treatment. One of the studies in this area found that half of the schizophrenic patients took less medication than recommended, and the reluctance was associated with the side effects of antipsychotic drugs (Van Putten 1974). On the positive side, many studies indicate improvement in treatment compliance with patient education. Particular approaches in patient education have been proposed that maximize the information conveyed while avoiding suggestibility (Balon 1999).

5. *Classification of Side Effects*

One of the common characteristics of psychoactive drugs is their capability to cross the blood-brain barrier and reach the central nervous system. Consequently, side effects of psychoactive drugs may be generated both centrally and peripherally. Here, the distinction between drug action and drug effect is also relevant: drug action refers to alterations produced by the drug at the molecular level, whereas drug effect (as well as side effect) refers to the observed reactions of the organism. Below, we will examine the behavioral side effects of psychoactive medications; a review of physiological systems is beyond the scope of this chapter and can be found elsewhere (Schatzberg and Nemeroff 1998).

5.1 *Autonomic*

The autonomic nervous system controls homeostasis in the different functional systems in the body mostly through involuntarily mechanisms. Major neurotransmitter systems involve those of acetylcholine and norepinephrine (β or α receptors).

Autonomic side effects of psychoactive drugs are most commonly due to *anticholinergic effects*. Because acetylcholine is widely distributed in the central and peripheral nervous systems, adverse anticholinergic effects can be observed both behaviorally and systemically.

Most antiparkinsonian (benztropine, trihexiphenidyl) and the majority of the antidepressant and antipsychotic drugs produce anticholinergic effects.

Among the tricyclic antidepressants, tertiary amine tricyclics (amitriptyline, clomipramine, doxepin, imipramine, trimipramine) produce more anticholinergic side effects compared to secondary amine tricyclics (desipramine, nortriptyline, protriptyline) and tetracyclic antidepressants (amoxapine, maprotiline). Compared to tricyclic antidepressants, anticholinergic side effects with selective serotonin reuptake inhibitors (SSRIs) are minimal. Conventional low-potency antipsychotics (drugs in this category, such as chlorpromazine, thioridazine, and mezoridazine require higher doses to produce equal efficacy than the high-potency antipsychotics) have significant anticholinergic properties compared to high-potency antipsychotics (fluphenazine, haloperidol, perphenazine, pimozide, thiotixene, trifluoperazine). Among the new generation antipsychotics, clozapine has the most potent anticholinergic effects, as well as having properties of acetylcholine release. Olanzapine also possesses anticholinergic effects, but less than clozapine. Risperidone, queitapine, and ziprasidone have minimal anticholinergic activity. Peripheral anticholinergic effects include warm skin, flushing, dryness, dilated pupils (and blurred vision), increased intraocular pressure, dry mouth (if unrecognized this can cause dental problems, such as caries), tachycardia (increased heart rate), reduced acid secretion in the stom-ach, reduced pulmonary secretions, diminished bowel motility (constipation), urinary retention, and delayed or retrograde ejaculation.

Central anticholinergic effects of psychoactive drugs involve features such as memory impairment, confusion, disorientation, delirium, visual hallucinations, and worsening of preexisting psychotic symptoms. Combination of psychoactive drugs with anticholinergic effects, such as a tricyclic antidepressant, and a low-potency antipsychotic will highly increase the likelihood of an adverse reaction. Geriatric patients are particularly sensitive to anticholinergic side effects and deserve special consideration in the choice of psychoactive drugs and evaluation of a deteriorating course of their clinical situation (see *Antidepressant Drugs*).

5.2 Dependence

Dependence is an important component of the side effect spectrum of some of the psychoactive drugs (barbiturates, benzodiazepines). The *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* defines substance dependence as a maladaptive pattern of substance use, leading to clinically significant impairment or distress, as manifested by three or more of the following criteria: tolerance, withdrawal, consumption of large amounts or over a longer than intended time period, persistent desire to cut down, significant effort to obtain the substance, social, occupational sacrifice, and continued usage despite known hazards. It is of note that in this definition a biopsychosocial approach is undertaken. The treatment models for dependency are conducted in a multidisciplinary fashion that address physical, psychological, and environmental factors. Chronic drug use involves tolerance, physical dependence, and psychological dependence. Even though related, tolerance and physical dependence are different concepts:

Tolerance refers to the need of increased amount of a drug with repeated usage in order to achieve the same magnitude of effect. Interestingly, not all of the behavioral effects of drugs show similar tolerance patterns; in the case of benzodiazepines, for instance, tolerance develops for the sedative effects, but not for the antianxiety effects. Tolerance is a reversible condition; abstinence from the drug over a period of time will reduce the amount of drug needed to exert the particular effects. Cross-tolerance is a phenomenon where the tolerance for a particular drug effect operates for another drug effect as well. It is demonstrated by the diminished anticonvulsant effects of the barbiturates in people with chronic alcohol use. Both changes in cellular physiology and behavioral conditioning systems have been postulated as underlying mechanisms for the development of tolerance.

Physical dependence refers to the condition where withdrawal symptoms emerge upon discontinuation of the drug. Withdrawal symptoms are specific to different drugs and quite unpleasant, so that continued use is sought to avoid them. Most withdrawal symptoms are separate from the ones in the intoxication state and can be identified without difficulty.

5.3 Endocrinological and Metabolic

Most antipsychotics elevate serum prolactin secondary to the blockade of dopamine. Increased prolactin may cause menstrual irregularities in females, gynecomastia (abnormal enlargement of the breast) in males, and galactorrhea (spontaneous milk flow) in both females and males. Amenorrhea (complete cessation of menses) occurs frequently in women taking antipsychotics. These changes persist with treatment and resolve with discontinuation. Galactorrhea and amenorrhea may also be observed with tricyclic antidepressants.

Hypothyroidism (decreased function of the thyroid gland) may be associated with long-term lithium therapy. The elderly and females are at greater risk for lithium-induced hypothyroidism. Weight gain is a common side effect of antipsychotics, particularly of new generations (clozapine, olanzapine, risperidone, sertindole, quetiapine), cyclic antidepressants, MAOIs, and SSRIs. Impact of weight gain on general health in the long run remains to be determined. In contrast, stimulants (amphetamine, dextroamphet-

amine, methylpenidate), and certain SSRIs are associated with appetite suppression and weight reduction. Stimulants are also associated with adverse growth effects (height and weight suppression); however, these effects are compensated with drug holidays.

5.4 Insomnia

Insomnia is the most common transient side effect of stimulants and MAOIs.

5.5 Neurological/behavioral

5.5.1 Cognitive effects. Even though the motor and cognitive functions of the brain do not operate independently from each other, specific neuropsychological tests that capture primarily motor or cognitive functions have been applied to measure the effects of psychoactive drugs in these domains. While some studies of conventional antipsychotics showed worsening performance on psychomotor speed, attention, and working memory, recent publications do not support a significant worsening in working memory, but even indicate improvements in some attentional measures (Goldberg and Weinberger 1996). The only double-blind study comparing the new generation antipsychotic clozapine and the conventional antipsychotic haloperidol revealed a significant interaction on measures of executive/attention and visualspatial functions. Further analysis revealed that these effects resulted from worsening performance of the haloperidol treated patients. Patients treated with clozapine showed an insignificant improvement (Buchanan et al. 1994). Anticholinergic doses have been associated with poor verbal learning, verbal fluency, and motor speed. Lithium may also interfere with memory and motor speed functions. Benzodiazepines may cause slowed cognitive function, impaired memory, and behavioral disinhibition.

5.5.2 Motor effects. Conventional antipsychotic drugs (haloperidol, fluphenazine), and to a lesser extent, new generation antipsychotic drugs (risperidone, olanzapine), some of the selective serotonin reuptake inhibitors (sertraline, fluoxetine), and cyclic antidepressants (amoxapine) affect motor functions of the extrapyramidal system. Acute conventional antipsychotic treatment interferes with simple repetitive motor tasks such as finger tapping. These effects are related to the drugs' dopamine blocking properties in the basal ganglia. Lithium and valproic acid may cause hand tremors. Stimulant use may be associated with development or worsening of tics.

Acute motor effects of the conventional antipsychotics include parkinsonism, dystonic reactions, and akathisia. The major chronic effect of antipsychotic use is tardive dyskinesia. *Parkinsonism* is a word derived from Parkinson's Disease, which is characterized by tremor, rigidity, and bradykinesia (slowed movement). All three features are observed in susceptible patients in the initial phases of treatment. Tremor with a regular rate may be observed in the hands and feet, and is present at rest. Tremor around the mouth also occurs. Akinesia (absence of movement) of the facial muscles sometimes lead to lack of expression which is known as masked face. Patients walk slowly, sometimes with shuffling, short steps, and the normal arm swing is diminished. Rigidity of muscles can be felt on examination; cogwheeling (increased tone in all directions of movement and superimposed ratchet-like resistance) may be elicited. Handwriting is characteristically small and tremulous.

One of the most disturbing side effects of antipsychotic medications (mostly of the high-potency conventional ones and some of the new generation antipsychotics) is akathisia, a subjective experience of motor restlessness. Patients complain of not being able to sit still, they pace in hallways, constantly move their legs to help the discomfort. They describe it as '… I want to run a hundred miles …' Significant akathisia may lead to insomnia because of having trouble lying in bed. It may be accompanied by a feeling of anxiety and agitation. Restlessness and agitation are frequently observed as part of psychosis in the absence of any drug treatment. Drug-induced akathisia therefore is very important to diagnose differentially for appropriate treatment. If it is mistaken as psychosis-related agitation, increased antipsychotic drug dose will only make it worse. Akathisia has been identified as a predictor of antipsychotic noncompliance. Several treatment options are available.

Acute dystonic reaction is another side effect of the conventional antipsychotics that requires immediate treatment. Dystonic reactions consist of abruptly developing muscle contractions which frequently involve the muscles of the face, tongue (protrusions, twisting), neck (spasmodic torticollis), jaw (trismus), eyes (oculogyric crisis, blephoraspasm), and limbs. It is most common in young men in the initial phases of treatment. It is treatable with anticholinergic and antihistamine drugs. Prophylactic treatment with these drugs may increase the chances of compliance.

Tardive dyskinesia is a long-term side effect of conventional antipsychotics characterized by involuntary choreiform (random, brisk) or athetoid (slow, regular continuing twisting) movements of the head, face, tongue, limbs, and the trunk. These movements may range from mild and unrecognized to severe and incapacitating. Movements around the mouth are the most common. Dyskinesia may be exacerbated with stress and disappears with sleep. About 10–20 percent of patients treated with conventional antipsychotics

for more than a year develop tardive dyskinesia. Children, women, the elderly, patients with brain damage or mood disorders constitute high-risk groups. While tardive dyskinesia is associated with antipsychotic use, 1–5 percent of patients with schizophrenia who lived before the antipsychotic era have been reported to exhibit similar dyskinetic movements. Thus, tardive dyskinesia may be related to the pathophysiology of schizophrenia. The new generation antipsychotics (clozapine, olanzapine, risperidone, quetiapine) have significantly lower incidence of extrapyramidal system side effects and tardive dyskinesia compared to the conventional antipsychotics. This is a feature of the new generation antipsychotics that distinguish them from the conventional antipsychotics. Numerous animal studies have demonstrated that clozapine fails to induce extrapyramidal system related movements, and it is the only antipsychotic that might have a potential to prevent tardive dyskinesia from deteriorating (see *Schizophrenia, Treatment of*).

5.5.3 Sedation. Sedation is a common side effect with most antipsychotics, mood stabilizers, benzodiazepines, barbiturates, and antidepressants (tertiary amine cyclic antidepressants, MAOIs, trazodone, sometimes with other SSRIs, venlafaxine, and nefazodone). In particular clinical situations, such as severe agitation, sedation may be desirable, whereas in others such as a negative symptoms of schizophrenia (affective flattening, lack of spontaneous speech, avolition), or depression with hypersomnia, it would not be. The choice of drug therefore largely depends on the clinical necessity. In undesirable conditions (night time dosing, reduction in dose) switching to a less sedating medication is considered.

5.5.4 Seizure. Antipsychotic drugs, particularly the ones with low potency reduce the seizure threshold. In a patient with no history of seizure disorder, incidence is low and no prophylactic treatment is warranted. Patients with a history of seizure disorder deserve special consideration. Clozapine increases risk of seizures in a dose-related fashion. Anticonvulsants such as valproic acid can benefit the patient to continue treatment with clozapine. Among the antidepressants, clomipramine and amoxapine may lower the seizure threshold while incidence of seizures with SSRIs is low, but present. Tricyclic antidepressants, MAOIs, anticonvulsants, lithium, and antipsychotics may also induce myoclonic (irregular rapid muscle contraction) movements.

5.5.5 Hypomania/mania. Even though it remains controversial, one of the important side effects of antidepressant treatment is the induction of hypomania/mania. In a patient with no history of mania, undergoing antidepressant treatment, emergence of manic features merits consideration of the possibility of drug-induced mania. In patients with bipolar disorder, on the other hand, the risk of mania induction with sole antidepressant (in the absence of a mood stabilizer) treatment is established.

5.5.6 Delirium. Infrequently, worsening of psychosis, confusion, and agitation may occur with antipsychotic treatment. Most cases are, however, associated with concomitant anticholinergic use.

5.5.7 Neuroleptic malignant syndrome. This is a potentially life-threatening complication of the antipsychotic treatment characterized by muscle rigidity, hyperthermia, mental state changes (stupor, agitation, confusion), and autonomic dysfunction (blood pressure instability, increased pulse, respiration, sweating). The incidence is about 0.1 percent of patients exposed to conventional antipsychotics. High-potency antipsychotics may have a higher risk than low-potency antipsychotics, even though rarely it has also been reported with clozapine. Rhabdomyolysis (muscle necrosis) may occur that leads to myoglobinuria (passing a muscle-related protein—myoglobin—in urine) and subsequent renal failure. Mortality may approach 12 percent. Immediate intervention is necessary.

5.6 Placebo Effects

A placebo is a pharmacologically inactive agent administered in a group of patients in a double-blind clinical trial to improve the true estimate of the active agents efficacy and side effects. Placebo effects may be relevant to the concept of side effects from the perspective of experiences of drug effects in the absence of drug action. Substantial evidence suggests emergence of physiological changes associated with placebo administration. While placebo drug efficacy approaches significant percentage of patients in some clinical trials, incidences of side effects are also variably present (see *Placebo Studies (Double-blind Studies)*).

5.7 Sexual

In males impotence, decreased libido, changes in quality of orgasm, priapism (painful erection), and erectile and ejaculatory disturbances have been reported with antipsychotic use. In women orgasmic dysfunction and decreased libido are frequently observed with antipsychotic treatment. Sexual dysfunction associated with SSRI use impairs quality of life, and is a reason for noncompliance. Buproprion and nefazodone may have a more favorable profile com-

pared to TCAs, MAOIs, and venlafaxine. In contrast, some antidepressants are prescribed for the treatment of premature ejaculation.

6. Quality of Life and Side Effects

The World Health Organization has defined Quality of Life as:

> 'An individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept incorporating in a complex way the person's physical health, psychological state, level of independence, social relationships, personal beliefs and their relationship to salient features of the environment' (WHOQOL 1995).

This definition emphasizes a comprehensive approach in the assessment of quality of life. Disciplines of medicine, particularly cancer research have concentrated on global measures of quality of life, as symptom improvement by itself does not account for it. In psychiatry, the slow pace of research in this area may be due to difficulties of valid and reliable quality of life assessments from psychiatric patients. Pioneered by Lehman (Lehman et al. 1982), renewed interest has led to research in debilitating psychiatric disorders such as schizophrenia and demonstrated the significance of a multidisciplinary treatment model. While the strongest predictor of symptom improvement is identified as compliance with recommended drugs, its sole impact on quality of life in the long term is yet to be elucidated. The issue gets complicated due to the side effect profile of psychoactive drugs, which may mimic the symptoms of diseases, and also due to the direct impact of side effects on quality of life. For patients with schizophrenia, symptoms, their severity, and subjective distress caused by side effects accounted for half the variance in quality of life (Awad et al. 1997).

In one study where 53 schizophrenic outpatients treated with depot neuroleptics were studied, 70 percent complained about side effects when 94 percent were found to experience them. Among the patients who denied having side effects, 87 percent had some form of it (Larsen and Gerlach 1996).

The recognition of the negative impact of tardive dyskinesia on, for instance, patients' self esteem, incapacity, and perception by society has been essential towards targeting more favorable treatments. The introduction of new generation psychoactive drugs, particularly antidepressants and antipsychotics is therefore encouraging. Such drugs have not only led to superior efficacy, but also have more favorable side effect profiles. However, there remains significant room for further improvement in side effect profiles of psychoactive drugs.

See also: Alzheimer's Disease: Antidementive Drugs; Antidepressant Drugs; Drugs and Behavior, Psychiatry of; Medicalization: Cultural Concerns; Patient Adherence to Health Care Regimens; Prophylaxis in Psychiatry; Psychopharmacotherapy, Costs of; Psychotherapy and Pharmacotherapy, Combined

Bibliography

American Psychiatric Association 2000 *Handbook of Psychiatric Measures*. APA, Washington, DC

Awad A G, Voruganti L N, Heslegrave R J 1997 A conceptual model of quality of life in schizophrenia: Description and preliminary clinical validation. *Quality of Life Research* **6**(1): 21–6

Balon R 1999 *Practical Management of the Side Effects of Psychotropic Drugs*. Dekker, New York

Bernstein J G 1995 *Handbook of Drug Therapy in Psychiatry*. Mosby-Year Book, St Louis, MO

Bloom F E, Kupher D J (eds.) 1994 *Psychopharmacology the Fourth Generation of Progress*. Raven, New York

Buchanan R W, Holstein C, Breier A 1994 The comparative efficacy and long-term effect of clozapine treatment on neuropsychological test performance. *Biological Psychiatry* **36**(11): 717–25

Feldman R S, Meyer J S, Quenzer L F 1997 *Principles of Neuropsychopharmacology*. Sinauer Associates, Sunderland, MA

Goldberg T E, Weinberger D R 1996 Effects of neuroleptic medications on the cognition of patients with schizophrenia: A review of recent studies. *Journal of Clinical Psychiatry* **57**(Suppl. 9): 62–5

Kaplan H I, Sadock B J 1999 *Comprehensive Textbook of Psychiatry*. Williams and Wilkins, Baltimore, MD

Larsen E B, Gerlach J 1996 Subjective experience of treatment, side-effects, mental state and quality of life in chronic schizophrenic out-patients treated with depot neuroleptics. *Acta Psychiatrica Scandinavica* **93**: 381–8

Lehman A F, Ward N C, Linn L S 1982 Chronic mental patients: The quality of life issue. *American Journal of Psychiatry* **139**(10): 1271–6

Schatzberg A F, Nemeroff C B 1998 *The American Psychiatric Press Textbook of Psychopharmacology*. American Psychiatric Press, Washington, DC

Spiegel R 1996 *Psychopharmacology: An Introduction*, 3rd edn. Wiley, Chichester, UK

Van Putten T 1974 Why do schizophrenic patients refuse to take their drugs? *Archives of General Psychiatry* **31**(1): 67–72

World Health Organization 1999 *WHOQOL Annotated Bibliography*. WHO, Geneva, Switzerland

N. R. Schooler and H. Gunduz

Psychophysical Theory and Laws, History of

1. Introduction

Psychophysics is a scientific approach to the measurement of mental processes. The empirical laws of psychophysics are based on such measures as the magnitudes of stimuli just noticeably different from

each other, the proportion of errors made when comparing two stimuli, the time required to complete such judgments, and self-reports of the confidence in comparative judgments or the strength of a stimulus. Some laws define relations between physical events and their mental representations such as sensations and feelings. The theories of psychophysics provide explanations of these empirical relations and suggest neural mechanisms that give rise to experimental results.

2. *Nineteenth-century Psychophysics*

The first law defined the 'just noticeable difference' (jnd) in sensation. A jnd occurred, phenomenally, when an increment in the size of a physical stimulus evoked a sensation of a difference in magnitude. The German physiologist Ernst Heinrich Weber (1834, 1846/1978) discovered, for many sense modalities, the increment in a physical stimulus needed to evoke a jnd. Leipzig physicist Gustav Fechner (1860) quantified Weber's observations and created the First Law of Psychophysics, Weber's Law,

$$\Delta s/S = c \tag{1}$$

where S is the magnitude of the physical stimulus, Δs is the increment, and c is called Weber's Constant. For discriminating lifted weights $c = 0.043$. Thus, an initial weight of 1000 g will require an increment of 43 g (to $S+\Delta s = 1043$ g,) before the new weight will feel just noticeably different from 1000 grams.

The subjective nature of these self-reports led Fechner to wonder whether the person experiencing a just noticeable difference could identify which of the two stimuli, S or $S+\Delta s$, was larger. Fechner discovered that two stimuli may feel just noticeably different, but may be identified correctly in only about 75 percent of the comparisons made. Fechner described these experiments and officially established the field of psychophysics in the two-volume *Elemente der Psychophysik* (1860/1966).

In *Elemente* Vol. 1, Fechner created a theory of how humans comparing two sensory stimuli, such as a pair of lights for brightness, two weights for heaviness, or two lines for length, make errors in their judgments. The theory assumed that the sensory system provided the mind with measurements of the external world. However, the process by which the sensory/nervous system created an internal sensory representation of an external physical stimulus introduced error into the measurement. A fixed weight might at one time feel heavier, or lighter, depending upon the amount of error added to, or subtracted from, the true sensory value for the external stimulus. Extending the theory of error created by Gauss in 1809 for physical measurement devices, Fechner invented a theory of judgment about invisible mental events that allowed him to measure the size of the error in the nervous system.

Figure 1 illustrates the essential features of Fechner's theory as it applies to the comparison of two just noticeably different weights, W_a and W_b. Along the abscissa are true sensory values, μ_a and μ_b, for the two weights undergoing comparison by lifting each once. Two Gaussian (Normal) probability distributions describe the error of sensory measurement and surround the true sensory values of μ_a and μ_b. Fechner postulated that a (decision) threshold, T, for deciding which weight was the larger existed exactly midway between μ_a and μ_b, that is at

$$T = \mu_a + (\mu_b - \mu_a)/2 \tag{2}$$

A lifted weight sensed to be greater than this threshold was thought to be the heavier weight.

The shaded area of Fig. 1 to the right of T represents the probability of making an error of judgment in deciding that the smaller of two stimuli was in fact the larger. The symmetric nature of Fechner's theory predicts that the probability of judging the larger of two stimuli to be the smaller equals the same probability, and is shown by the shaded area to the left of the threshold, T. The two error probabilities increase as the variability in the sensory system increases. Such an increase in variability causes a greater spread in each Gaussian probability distribution.

Fechner's goal was to measure the invisible variability of the sensory system in units of the physical stimuli. As a measure of variability Fechner chose the value $\sqrt{2}\sigma$, Gauss's measure of the spread of the Gaussian distribution. The value of σ, a parameter of the Gaussian probability distribution, creates a scale of measurement units along the abscissa of Fig. 1. The unknown value for T is located an unknown number of σ units above the true sensory value μ_a. But, once the probability of an error of judgment is known, one can determine from tables (many created by Fechner) the corresponding distance in σ units of the (decision) threshold T above the true sensory value μ_a. For example, if the probability of an error in judging which weight is larger equals 0.25, then the (decision) threshold, T, is 0.674σ units greater than the true sensory value μ_a.

To determine the value of σ, Fechner replaced the unknown values of μ_a and μ_b by the physical values that gave rise to them, W_a and W_b. The distance from μ_a to T is defined as $t = 0.675\sigma$. That is,

$$t = 0.675\sigma = (W_b - W_a)/2, \quad \text{or}$$
$$\sigma = (W_b - W_a)/(2{*}0.675) = (W_b - W_a)/1.35 \tag{3}$$

Using the example above, the smaller of two just noticeably different weights equaled 1000 g and the larger 1043 g (and the probability of an error equaled 0.25). Thus the measure of variability is

$$\sigma = (1043\,\text{g} - 1000\,\text{g})/1.35 \approx 32\,\text{g} \tag{4}$$

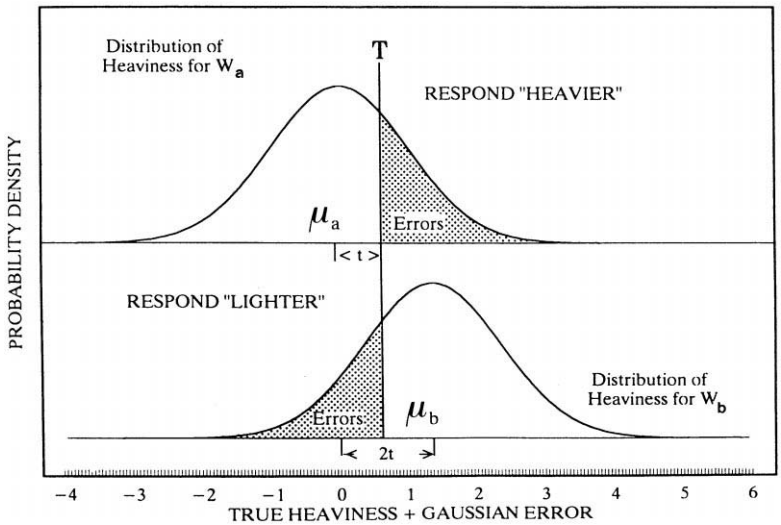

Figure 1
Fechner's theory of mental judgment proposed that Gaussian variability perturbed nervous system transmission of peripheral sensory stimuli. Weight W_a was incorrectly judged heavier than weight W_b whenever the sensory value for W_a exceeded the value T. Similarly, W_b was incorrectly judged lighter than W_a whenever the sensory value for W_a was less than T. The probabilities of an error are represented by the shaded areas to the right and left of T

Fechner summarized these ideas in a famous formula that is the Second Law of Psychophysics:

$$t = hD/2 \tag{5}$$

Here $h = 1/\sqrt{2}\sigma$ and D equals the difference between two physical stimuli, $W_b - W_a$

The term h measures the sensitivity of the system. When σ is large, indicating great sensory variability and a small value of h, many errors occur and the system is said to be insensitive. A small value for σ leads to few errors, a larger value for h, and the system is said to be more sensitive. This first method of measuring the invisible, internal, variability of the sensory system established psychophysics as a truly scientific discipline. Measuring the unknown variability of the sensory/nervous system in units of physical stimuli marks a revolution in the scientific investigation of the human mind. This powerful theory became the basis for the fields of experimental psychology and statistical hypothesis testing.

Fechner also theorized, in *Elemente* Vol. 2, about a measure of how much sensation a stimulus generated. Placing each jnd in a sequence is a way of adding together phenomenally equal units of sensation. Using Weber's Law as a base, Fechner argued that on a continuous scale of sensation Weber's Constant,

$$c = \Delta s/S \tag{6}$$

is a differential equation. Integrating the differential equation is equivalent to adding together the jnds of sensation increments to obtain a total amount of sensation. The result of the integration is called Fechner's Law of Sensation: the total amount of sensation generated by a stimulus is

$$\xi = k^* \log_e(S) \tag{7}$$

where S is the magnitude of the stimulus and ξ is the amount of sensation.

Fechner's Law is perhaps the most widely known of any psychophysical law. Yet criticisms of Fechner's idea never abated (Krueger 1989, Lockhead 1992, Murray 1993). The notion of adding together jnds to create a yardstick of sensory measurement is appealing, but the 'law' provides for no independent experimental means for testing its prediction. How is internal sensation to be measured? Or as Laming (1997) wonders, 'Can sensation be measured?'

The father of American Philosophical Pragmatism, C. S. Peirce, put Fechner's ideas to another test. Peirce argued that the sense of confidence in a judgment was a sensation directly related to the probability of making a correct judgment. Peirce's Law states that if C is the average confidence in a judgment and p is the probability of making a correct judgment, then

$$C = k^* \log_e[p/(1-p)] \tag{8}$$

Like Fechner's Law there is a logarithmic relation between two variables. But, in this case each variable can be measured through direct experimentation. Peirce and Jastrow (1884, see Link 1992, Vickers 1979), and later experiments, confirmed this prediction.

The profound impact of Fechner's Psychophysics led to Wilhelm Wundt's 1879 creation of the first psychological laboratory at Leipzig University. A new student, James McKeen Cattell, proposed a new measure of the amount of sensation generated by perfectly discriminable stimuli, as opposed to jnd stimuli. Reaction time, RT, the time taken to respond to the presence of a stimulus, could measure stimulus strength. Cattell's experiments showed remarkable decreases in response time as stimulus magnitude increased. Later Piéron (1914, see Welford 1980, Luce 1986) proposed the general principle: Piéron's Law:

$$RT = a + kS^{-\beta} \tag{9}$$

where a is a constant due to non-sensory effects, k and β are constants, and S is physical stimulus intensity.

3. *Twentieth-century Psychophysics*

Yet, measuring the magnitudes of perfectly discriminable stimuli remained elusive. Thurstone (1955, Bock and Jones 1968) proposed that every stimulus projects a value on an internal psychological continuum. For example, stimuli such as crimes may vary on a continuum of the seriousness of that crime. When asked to judge which of two crimes is more serious, a subject will compare the two crime's seriousness values and choose the crime with the greater seriousness value. This is the famous Law of Comparative Judgment.

Crimes vary in their degree of seriousness for each subject, but for a fixed subject each crime's seriousness is fixed. For a fixed crime, individual differences across subjects created variability in the seriousness of crime. Thurstone represented this variability by a Gaussian distribution of seriousness values. For a group of subjects the theory is formally identical to Fechner's, but the emphasis on measuring stimulus values, rather than variability, led to a powerful new direction in psychophysical research.

Thurstone proved that by comparing all stimuli against each other, each stimulus could be measured on the psychological continuum. Now, mental phenomena as elusive as attitudes could be measured. In an application to classical psychophysics, Guilford (1954) used the method to determine heaviness values for difficult-to-discriminate lifted weights.

Perplexing irregularities in the ordering of probabilities in the full paired comparison matrix were the basis for Clyde H. Coombs to propose an important extension of the Law of Comparative Judgment, the Theory of Ideal Comparisons. Rather than comparing two stimuli against each other, Coombs (1964) proposed that each stimulus was compared against an ideal defined on a psychological continuum. Choosing between two stimuli equally distant from the ideal led to indifference with respect to them and the peculiar response proportions in the paired comparison matrix. For example, a continuum of shades of gray may contain an ideal gray. But, two shades of gray equally distant on either sides of the ideal would be equally preferred. The subject faced with choosing either of these stimuli would be indifferent to them and chose either with probability 0.50.

In spite of its immense value the Law of Comparative Judgment said nothing about how response time or confidence should vary when stimuli are compared against each other. D. M. Johnson (1939, see Link 1992) combined choice, confidence, and response times (RT) in an experiment on judgments of differences between line lengths and discovered Johnson's Law:

$$RT = a(10)^{b(1-C)} \tag{10}$$

where $0 \leqslant C \leqslant 1$ is a measure of confidence and a and b are constants. Johnson's Law, together with Peirce's Law, suggests a relationship between confidence and response probability.

S. S. Stevens (1975) dismissed the theoretical reasoning leading to Fechner's Law and the Law of Comparative Judgment, and proposed an alternative formulation for determining the amount of sensation generated by a single stimulus. For Stevens 'equal stimulus ratios produce equal sensation ratios.' This idea is the basis for the famous Power Law of Sensation,

$$\xi = \lambda S^{\gamma} \tag{11}$$

where ξ is the strength of response to a stimulus of intensity S, λ is a non-sensory scaling constant and γ measures the effects of peripheral sensory transducers that are unique to a sensory modality. Stevens (1975) determined Power Law exponents for many sensory quantities and gave convincing evidence of the satisfactory representation of stimulus strength provided by the Power Law.

Although the Power Law contradicts Fechner's Law, the Power Law must be consistent with Weber's Law. Link (1992) derived from the Wave Theory of Difference and Similarity the prediction that Stevens' exponents and Weber's Constants are reciprocals of each other. Using values reported by Teghtsoonian (1971, see Link 1992) the prediction held for sensory quantities investigated by Stevens such as shock, heaviness, loudness, finger span, vibration, saturation, taste, and brightness. Ekman's Law provides yet another relation between jnds and the Power Law by asserting that increments in sensation on a psychological continuum are a linear function of sensation magnitude, as if Weber's Law applied to both the psychological continuum and the physical stimulus continuum. Gescheider (1997) shows how Ekman's Law leads to Stevens' Power Law.

Cognitive factors also affect the magnitude of response. Baird (1996) suggests that beyond the sensation effects caused by the peripheral nervous system are cognitive factors that demand a reinterpretation of the Power Law. Proposing a Comple-

mentarity Theory of Psychophysics, Baird accounts for a wide variety of experimental results unaccounted for by the Power Law. In a different approach, Norwich (1993) developed a general theory of the sensory language of the brain. Using the fundamental equation,

$$F = kH \tag{12}$$

where F is a perceptual variable, k is a positive constant and H is a measure of entropy, Norwich derives both the Power Law and Fechner's Law from the same basis.

Fechner's proposal that the (decision) threshold, T, was halfway between the means of the two sensory distributions shown in Fig. 1 proved too restrictive. A major advance in traditional psychophysical modeling occurred when Tanner, Swets, and Birdsall (1954, see Swets 1964) allowed the decision threshold to vary. Calling the decision threshold a response criterion, c, the authors ushered in a new era of psychophysical studies. Experiments showed how a nonsensory component of performance, response bias, affected the position of c. This Theory of Signal Detection and Recognition greatly expanded the applications of psychophysical methods.

For choices between many stimuli Luce (1959, see Falmagne 1985) created the Choice Axiom. Let $S_1, \ldots, S_n$ be a set of commensurate stimuli. Luce's Choice Axiom states that the probability of choosing stimulus S_i is

$$W(S_i)/(\Sigma_n W(S_n)) \tag{13}$$

where W_i is the psychophysical value of stimulus S_i. For example, given five breakfast cereals, $B_1, \ldots, B_5$, the probability of choosing cereal B_i is

$$p_i = W(B_i)/(\Sigma_h W(B_n)) \tag{14}$$

Luce's Choice Axiom receives support from numerous experiments in psychology and economics.

The majority of psychophysical theories and experiments focus on the ability to discriminate between stimuli rather than on the similarity between them. Shepard (1957, see Borg and Lingoes 1987) proposed that similarity between two stimuli, (S_i, S_k), is a continuously changing negative exponential function of the distance, d_{ik}, between them. The probability of correctly identifying a particular stimulus from a set of n stimuli is often assumed to be

$$p_{ik} = \exp(-d_{ik})/\Sigma_n \exp(-d_{in}) \tag{15}$$

a form similar to Luce's Choice Axiom.

Following Thurstone's suggestion that there are many underlying psychological continua, Anderson (1981) proposed an Information Integration Theory of how sensory values on different dimensions are combined to create perceptual experience. Anderson's cognitive algebra creates perceptual magnitudes from a linear model of sensory values defined on the dimensions on which a stimulus projects its values. Subjectively, a fine dinner is a sum of the psychological values of its many components.

4. *Conclusion*

During nearly 185 years, the field of psychophysics introduced new measurements of, and theories about, mental processes. Following Weber's discovery of the jnd, Fechner created Weber's Law, the field of psychophysics, and a theory of error in sensory comparisons. Thurstone introduced the idea of psychological continua on to which stimuli project psychological values. Comparisons between two psychological values led to the Law of Comparative Judgment and measurements of stimuli not defined on a physical scale. Anderson proposed a cognitive algebra for adding psychological values combined across several such psychological continua.

Fechner's Law, that the amount of sensation equaled a logarithmic transformation of the physical stimulus intensity, provoked continual controversy. Piéron showed that the time required to respond to the presence of a stimulus obeyed a power law of stimulus intensity. Stevens proposed that the amount of sensation obeyed a Power Law with an exponent unique to a sensory modality.

But cognitive factors were also shown to affect judgments of stimuli. Coombs proposed that psychological values were compared against ideals rather than against each other, and suggested a method of locating an ideal. Signal Detection Theory provided for a decision criterion that could vary in response to cognitive factors. Theoretical ideas suggested new experiments. New empirical discoveries challenged old theories. Now, at the beginning of the twenty-first century, psychophysics is poised for major theoretical advances toward more refined and more extensive measurements of mental processes.

See also: Cognitive Psychology: History; Cognitive Psychology: Overview; Fechnerian Psychophysics; Memory Psychophysics; Sensation and Perception: Direct Scaling; Signal Detection Theory, History of; Signal Detection Theory: Multidimensional

Bibliography

Anderson N H 1981 *Foundations of Information Integration Theory*. Academic Press, New York

Baird J C 1996 *Sensation And Judgment: Complementarity Theory of Psychophysics*. Mahwah, Hillsdale, NJ

Bock R D, Jones L V 1968 *The Measurement and Prediction of Judgement and Choice*. Holden-Day, San Francisco

Borg I, Lingoes J 1987 *Multidimensional Similarity Structure Analysis*. Springer-Verlag, New York

Coombs C H 1964 *A Theory of Data*. Wiley, New York

Falmagne J C 1985 *Elements of Psychophysical Theory*. Oxford University Press, New York

Fechner G T 1966 [1860] In: Boring E G, Howes D H (eds.) *Elements of Psychophysics*. Holt, Rinehart & Winston, New York
Gescheider G A 1997 *Psychophysics: The Fundamentals*. Erlbaum, Mahwah, NJ
Guilford J P 1954 *Psychometric Methods*, 2nd edn. McGraw-Hill, New York
Krueger L E 1989 Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences* **12**: 251–67
Laming D R J 1997 *The Measurement of Sensation*. Oxford University Press, New York
Link S W 1992 *The Wave Theory of Difference and Similarity*. Erlbaum, Hillsdale, NJ
Lockhead G R 1992 Psychophysical scaling: Judgments of attributes or objects? *Behavioral and Brain Sciences* **15**: 543–58
Luce R D 1986 *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press, New York
Murray D J 1993 A perspective for viewing the history of psychophysics. *Behavioral and Brain Sciences* **16**: 115–37
Norwich K H 1993 *Information, Sensation and Perception*. Academic Press, San Diego, CA
Stevens S S 1975 In: Stevens G (ed.) *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*. Wiley, New York
Swets J A (ed.) 1964 *Signal Detection and Recognition by Human Observers: Contemporary Readings*. Wiley, New York
Thurstone L L 1955 *The Measurement of Values*. University of Chicago Press, Chicago
Vickers D 1979 *Decision Processes in Visual Perception*. Academic Press, New York
Weber E H 1978 [1846] Der Tastsinn. In: Ross H E, Murray D J (eds.) *E. H. Weber: The Sense of Touch*. Academic Press, London
Welford A T (ed.) 1980 *Reaction Times*. Academic Press, New York

S. W. Link

Psychophysics

How might the human mind be studied, scientifically?

It is a commonplace notion that the brightness of a light, the loudness of a sound, or the heaviness of a weight is one thing, and the luminance, the sound pressure level, or the physical weight in gm quite another. If a way could be found to measure brightness, loudness, and heaviness—not the physical attributes of luminance, sound pressure level, and weight, but the subjective experiences they engender—then such measurements might provide the foundation for a scientific study of the mind, analogous to physics.

The name of this enterprise is *psychophysics*. It has led to two quite distinct traditions of experimental research that I call *sensory discrimination* and *sensory judgment*.

1. Sensory Discrimination

Fechner ([1860] 1966) conceived the idea of using the 'just noticeable difference' (JND) between two stimulus magnitudes as the unit of subjective measurement. Earlier, Weber ([1834, 1846] 1978) had enunciated the law that now bears his name. If X_1 and X_2 are two sensory magnitudes such that X_2 is just noticeably greater than X_1, then

$$(X_2 - X_1)/X_1 = \text{constant}. \quad (1)$$

Writing the constant, which is known as the Weber Fraction, as Θ,

$$X_2 = X_1(1+\Theta) \quad (2)$$

or

$$\ln X_2 - \ln X_1 = \ln(1+\Theta), \quad (3)$$

another constant, which Fechner identified as the subjective value of a JND. It follows that the subjective experience of stimulus magnitude increases as $\ln X$, a relation known as Fechner's Law.

1.1 Psychophysical Methods

The project of measuring subjective sensation by means of Eqn. (3) depends on accurate measurement of JNDs. Fechner directed systematic research towards the development of three methods.

1.1.1 The method of limits. Envisage that X_2, greater than X_1, is reduced progressively by small amounts until it is judged 'equal' and then by further small amounts until it is 'less than', or the process might be reversed with X_2 initially less than X_1. This procedure identifies 'thresholds' above and below X_1 at which the judgment changes.

1.1.2 The method of average error. Envisage that X_2 is adjusted until it is judged 'equal' to X_1, which is fixed. The probable error of adjustment can then be used as an alternative to the upper and lower thresholds identified by the method of limits.

1.1.3 The method of constant stimuli. Envisage that various X_2s are presented together with a fixed X_1 and the subject asked to say which is greater. It turns out that the probability of a correct judgment increases smoothly with the difference between X_1 and X_2. That relationship is known as the *psychometric function*.

1.2 Signal-detection Theory

The hundred years following Fechner's seminal work were spent arguing, first, how properly to analyze the judgments from these three psychophysical methods and, second, what the relationship between them was. That argument was resolved only with the advent of signal detection theory, which showed that these were the wrong questions to be asking. To put the matter succinctly, the presentation of a stimulus, be its magnitude X_1 or X_2, provides the subject with information whether the stimulus magnitude is, in fact, X_1 or X_2. This information is a random variable with a different distribution depending on which stimulus is presented. The optimum strategy for the subject is to report 'X_2' when the information in favour of X_2 exceeds a certain criterion and 'X_1' otherwise.

Signal-detection theory brought the idea of hypothesis testing by bisection of the likelihood-ratio (Neyman and Pearson 1933) into psychological theory. It identified two fundamental factors in sensory discrimination; there is (a) the information afforded to the subject (a property of the stimuli to be discriminated), and (b) the decision criterion (which is at the subject's disposal). These two factors are confounded in the traditional psychophysical methods.

It is common to model signal detection in terms of two normal distributions (unit variance, difference of mean d'), but it should be borne in mind that this is a *model* of the information and its function is simply to facilitate calculations. Nevertheless, in that spirit, let the mean of the distribution representing magnitude X be $\ln X$ and let the variance be $[\ln(1+\Theta)]^2$. If a JND be defined as that stimulus difference which gives $d' = 1$, Eqn. (3) is obtained.

We cannot say, however, that signal-detection theory validates Fechner's Law (Eqn. (3)) because the signal-detection model is no more than a model and any other model delivering the same numerical predictions, more or less, will do as well. Such another model is composed of χ^2 distributions multiplied by scaling factors proportional to the magnitudes of the stimuli. It arises in this manner.

1.3 Differential Coupling

Weber's Law (Eqn. (1)) holds rather accurately for discrimination between two separate stimuli (two flashes of light in darkness, two bursts of noise in silence), and for many stimulus attributes, down to, but not below, absolute threshold. The amplitude of a pure tone (but not Gaussian noise) and the contrast of a sinusoidal grating are known exceptions. But if the task is to detect a flash of light added to a pre-existing background, the relationship is more complicated. For the detection of a sinusoidal grating van Nes and Bouman (1967) demonstrated a square-root relationship up to a transition illuminance, above which detection conformed to Weber's Law, and the transition illuminance increased as the square of the wavenumber of the grating. In addition, the psychometric function for detection is steeper than that for discrimination between two separate stimuli; and the Weber fraction for discrimination increases when one of the stimulus magnitudes lies below the absolute threshold. In short, Weber's Law, on which Fechner's Law is based, is just one component of a large body of related phenomena, a body sufficiently large and complex as to admit substantially only one explanation.

(a) The sensory pathway is differentially coupled to the physical stimulus so that only changes in sensory input are available as a basis for discrimination. When the two stimuli to be distinguished are presented separately, that means intrinsic random variability only. The energy of that intrinsic variability (a sample of Gaussian noise) has a χ^2 distribution multiplied by the noise power density, which, in this case, is proportional to the stimulus magnitude X. The logarithm of a χ^2 variable is, to a very good approximation, normal with variance independent of X. Weber's Law falls directly out of differential coupling.

(b) The steeper gradient of the psychometric function for detection is equivalent to the interposition of a square law transform, which, in turn, is directly analogous to the phenomenon of low-signal suppression in radio reception (Davenport and Root 1958). The stimulus to be detected adds a small perturbation to the Gaussian noise giving a noncentral χ^2 distribution of energy. The resultant central/noncentral χ^2 model accommodates additional phenomena within the general theory and is to be preferred to the logarithmic model underlying Fechner's Law. The distributions in this model are everywhere scaled in proportion to physical stimulus magnitude, so that there is no empirical justification (arising from the study of sensory discrimination) for any measure of sensation distinct from the physical (Laming 1997, Chaps. 3, 4).

(c) The differential coupling is realized in the balanced excitatory/inhibitory receptive fields of sensory neurons. The low-signal suppression is a consequence of transmission by all-or-none action potentials of one polarity only. The theory of sensory discrimination describes the flow of information through the early stages of sensory pathways and annexes that preconscious component of perception to the physical domain. There is no possibility of basing a scientific study of the mind on resolving power.

2. Sensory Judgment

A very wide range of sound energy levels are encountered in nature and sound pressure level is measured on a logarithmic scale; 10 decibels equates to a 10-fold

increase in power. If Fechner's Law were valid, people should say that 100 dB (a subway train entering a station) should sound twice as loud as 50 dB (quiet conversation in a library). Most people would say that the subway train is much more than twice as loud. Around 1930 this posed a problem for acoustic engineers explaining sound levels to their clients and led to the development of the *sone* scale (Stevens 1936). Loudness in sones increases twofold for each 10 dB increase in sound power.

2.1 Stevens' Power Law

Stevens (1957) returned to the problem of subjective measurement in the 1950s with three new experimental methods.

(a) *Magnitude estimation*, in which subjects assigned numbers to stimulus magnitudes;

(b) *Magnitude production*, in which subjects adjusted a variable stimulus magnitude to match a given number; and

(c) *Cross-modality matching*, in which subjects adjusted a stimulus magnitude on one continuum to match a given magnitude on another.

Taking the numbers assigned to the stimuli to be direct measures of subjective sensation, Stevens found that

$$N = aX^{\beta}, \tag{4}$$

where N is the mean assignment, a a constant, and β an exponent typical of the stimulus attribute. Equation (4) is known as *Stevens' Power Law* and its validity has been demonstrated for more than 30 different attributes.

2.2 Explanation of the Power Law

It is tempting to suppose, as Stevens did, that the sheer volume of experimental demonstration cracks the problem of subjective measurement. But those experiments invariably employed a geometric ladder of stimulus values. If subjects could be induced to assign numbers in like manner (to 'judge ratios'), then a power law relationship is inevitable, simply by regression of log number onto log stimulus magnitude. There are several reasons for supposing that this is all Eqn. (4) tells us.

(a) The power law relation can be perturbed by spacing the stimuli other than geometrically.

(b) The subjects in Stevens' laboratory consistently used numbers extending over a range of about 1 to 30. The estimate of the exponent β can therefore be calculated from the log range of stimulus values employed.

(c) Instructions to subjects are invariably illustrated with numerical examples. A wider spread of numbers in those examples, but examples which conform to the same nominal scale, produces a wider spread of estimates in the experiment and therefore an increased value of β. Subjects accept guidance about how wide a range of numbers they should be using.

(d) The variance of the numerical assignments in experiments from Stevens's laboratory is about 100-fold greater than the equivalent variability implicit in sensory discrimination.

(e) But that variability is much reduced when two similar stimulus values (± 5 dB) are presented in succession. The correlation between successive log numerical assignments (then about $+0.8$) implies that each stimulus is used as a point of reference for the succeeding judgment.

So magnitude estimates are very imprecisely determined and are easily influenced by extraneous factors. Good power law data depends on balancing these extraneous influences by appropriate randomization of the schedule of stimulus presentations.

2.3 Relative Judgment

The correlation (e) above supports a theory of relative judgment in which each stimulus is compared to its predecessor. More to the point, that comparison is little better than ordinal; we can say that this sound is louder/softer/about the same as the previous one, but that is all. This idea (Laming 1984) accounts for several otherwise quite unrelated phenomena, including the quantitative variation of the autocorrelation (e) with stimulus differences. By focusing on the precision with which such judgments are made, this idea also relates magnitude estimation to category judgment and absolute identification—the stimuli and the subject are the same, only the response categories are different—and especially to the often repeated finding that the amount of information in a category judgment is limited to 2.3 bits, which is equivalent to the use of no more than five categories without error (Garner 1962, Chap. 3). But if human judgment is ultimately no better than ordinal, it affords no possibility of establishing a metric internal to the mind that is distinct from the physical.

3. How Might the Human Mind be Studied?

There are three ways in which these traditions of research might be continued.

3.1 Information Theory

(Note that I refer here to the mathematical theory of statistical information, not to the information theory (properly *communication theory*) of Shannon (1949), nor to the colloquial use of 'information' which has subsequently become fashionable in psychology.)

Discriminations between two separate stimuli conform accurately to the normal, equal variance, signal detection model (Tanner and Swets 1954). In that

model the parameter d'^2 measures the mean information afforded by a single presentation of either stimulus. A discrimination between two indistinguishable (identical) stimuli provides a natural zero; two looks at the same stimulus (cf. Swets 1964), or two simultaneous, correlated, signals, provide a natural concatenation of separate means; and a comparison between the informativeness of two different discriminations can easily be fashioned from statistical analysis. These are the criteria needed to establish ratio-scale measurement, and it is ratio-scale measurement of fundamental quantities such as mass, length, and time that has enabled the physical sciences to develop their present-day precision.

With that lesson in mind, psychologists have long striven to establish their own scales of measurement, but hitherto have always lacked empirical foundations sufficient to establish a ratio scale (Michell 1999). The parameter d'^2 bids fair to provide the first ratio-scale measure of specifically psychological origin; other ratio scales in psychology have been borrowed from extraneous disciplines. By this means experimental performance may be related to physical parameters of the stimuli (see Laming 1986, 1988).

3.2 The Precision of Sensory Judgment

Magnitude estimation and category judgment may be meaningfully analyzed using a multistimulus elaboration of the normal signal detection model, with a separate normal distribution for each stimulus value (the Law of Categorical Judgment, Case IC; Torgerson 1958). The inverse variance in units of log stimulus magnitude measures the precision of the judgments. That precision varies inversely with the square of the log stimulus range and also depends on the randomness of the stimulus presentation schedule. An 'analysis of variance' of sensory judgment provides a second vehicle of enquiry.

Inverse variance is Fisher's (1922) measure of information and is closely related to d'^2. So this second suggestion employs the same mathematical theory as the first, but now directed toward the relationships between different experimental paradigms, including the classical psychophysical methods listed above. The lesson to be absorbed is that a magnitude estimate provides some limited information about the subject's assessment of the stimulus, but is never to be taken at its face value.

3.3 Sequential Interactions

It has been known since 1920 that successive psychophysical judgments are statistically related, but there is, as yet, no theoretical understanding what those interactions signify. Aggregate performance in an experiment results from a pattern of sequential interactions driven by a (usually) random sequence of stimulus values. Trial-by-trial analysis provides a third way to study the mechanics of the human mind.

3.4 Internal Sensation

But none of these approaches measures subjective sensation. Internal experience is private to the subject; it cannot be shared with the experimenter except to the extent that it can be expressed in terms of some feature of the external world perceptible to both. Experimental paradigms are simply schemes for expressing that otherwise private experience. The strictly limited precision of any and every subject's judgments suggests that there is no more precise internal metric to be expressed.

See also: Fechnerian Psychophysics; Memory Psychophysics; Psychophysical Theory and Laws, History of; Signal Detection Theory; Signal Detection Theory, History of; Signal Detection Theory: Multidimensional

Bibliography

Davenport W B Jr., Root W L 1958 *An Introduction to the Theory of Random Signals and Noise*. McGraw-Hill, New York

Fechner G T [1860] 1966 *Elements of Psychophysics*. Holt, Rinehart and Winston, New York, Vol. 1

Fisher R A 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* **222A**: 309–68

Garner W R 1962 *Uncertainty and Structure as Psychological Concepts*. Wiley, New York

Laming D 1984 The relativity of 'absolute' judgements. *The British Journal of Mathematical and Statistical Psychology* **37**: 152–83

Laming D 1986 *Sensory Analysis*. Academic Press, London

Laming D 1988 Précis of *Sensory Analysis* and A re-examination of *Sensory Analysis*. *Behavioral and Brain Sciences* **11**: 275–96, 316–39

Laming D 1997 *The Measurement of Sensation*. Oxford University Press, Oxford, UK

Michell J 1999 *Measurement in Psychology: Critical History of a Methodological Concept*. Cambridge University Press, Cambridge, UK

Neyman J, Pearson E S 1933 On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London* **231A**: 289–337

Shannon C E 1949 The mathematical theory of communication. In: Shannon C E, Weaver W (eds.) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL, pp. 1–91

Stevens S S 1936 A scale for the measurement of a psychological magnitude: Loudness. *Psychological Review* **43**: 405–16

Stevens S S 1957 On the psychophysical law. *Psychological Review* **64**: 153–81

Stevens S S 1975 *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. Wiley, New York
Swets J A (ed.) 1964 *Signal Detection and Recognition by Human Observers: Contemporary Readings*. Wiley, New York
Tanner W P Jr., Swets J A 1954 A decision-making theory of visual detection. *Psychological Review* **61**: 401–9
Torgerson W S 1958 *Theory and Methods of Scaling*. Wiley, New York
van Nes F L, Bouman M A 1967 Spatial modulation transfer in the human eye. *Journal of the Optical Society of America* **57**: 401–6
Weber E H [1834, 1846] 1978 *The Sense of Touch*. Academic Press, London

D. Laming

Psychophysiology

1. Introduction

1.1 What is Psychophysiology?

Psychophysiology, as the term implies, is the study of the interrelationships between mind and body. Although clearly related to behavioral and cognitive neuroscience, psychophysiology differs from traditional neuroscience both methodologically and conceptually. Methodologically, psychophysiologists study primarily human subjects rather than lower animals, they usually measure physiological responses non-invasively, and the physiological responses are more molar as opposed to molecular. Conceptually, most psychophysiologists focus primarily on understanding psychological events and secondarily on understanding the underlying physiological events.

1.2 Goals and History of Psychophysiology

The notion that one can understand psychological states and processes by means of observing or measuring physiological changes dates from the origins of the human race (e.g., blushing indicates embarrassment, sweating is associated with fear, etc.). However, it was not until approximately 1960 that psychophysiology was defined as a specific scientific discipline. The Society for Psychophysiological Research was founded in 1960 and the journal *Psychophysiology* was first published in 1963 under the leadership and vision of Albert Ax. The stated purpose of the society was and continues to be 'to foster research on the interrelationships between the physiological and psychological aspects of behavior.' Because only some of the highlights of psychophysiology can be identified here, the reader is referred to an introductory text by Hugdahl (1995), a more advanced handbook edited by Cacioppo et al. (2000), and the journals *Psychophysiology* and *International Journal of Psychophysiology*.

1.3 Typical Psychophysiological Measures

A great many physiological responses are used as psychophysiological measures. These fall broadly into four categories. The first, and historically oldest, category includes responses of the autonomic (or involuntary) nervous system (ANS), that branch of the nervous system, which controls the actions of internal organs such as the heart, blood vessels, sweat glands, and digestive organs. It is divided into the sympathetic branch, which mobilizes body resources to prepare the organism to respond to the environment and deal with stressors and threat (e.g., by the increasing heart rate, respiration rate, and palmar sweating), and the parasympathetic branch, which acts to store and conserve energy by reducing heart rate and increasing digestive activity. As can be seen, the ANS controls many aspects of physiological function that serve as indicators of emotion or activation. Thus, a psychophysiologist might use heart rate or palmar sweating as a measure of how anxious a person felt in a particular social situation, such as having to give a speech to an audience.

The second category of psychophysiological measures reflects the activity of the skeletal nervous system, as indicated by motor activity. Of particular interest have been general measures of overall muscle tension, such as tension in the forearm muscles, which can reflect broad states of arousal and activation, and measures of the activity of facial muscle activity, which can indicate states of emotion.

The third category of psychophysiological measures consists of electrophysiological indices of central nervous system activity, primarily electroencephalogram (EEG) and evoked response potential (ERP) measures. For instance, one might use the amount of alpha (8–12 Hz) rhythm in the EEG as an index of relaxation, or the magnitude of a particular ERP component elicited by a stimulus as a measure of attention to that stimulus. The fourth category of measures is derived from brain scanning techniques, such as positron emission tomography (PET), magnetic resonance imaging, MRI, or functional MRI scans. For example, one could use a PET-derived measure of metabolic activity in the frontal lobes to examine differences in executive function between patients with schizophrenia and normal subjects.

The third and fourth categories of measures are the subjects of individual articles in this encyclopedia (see *Event-related/Evoked Potentials*; *Electroencephalography: Basic Principles and Applications*; *MRI (Magnetic Resonance Imaging) in Psychiatry*). There-

fore, the present discussion is confined to ANS activity and skeletal muscle activity. These measures have been used to investigate long lasting tonic states, transient phasic responses, and individual differences among people.

2. Long-lasting States

2.1 Alertness/Arousal/Activation

The terms alertness, arousal, and activation have a long and useful, albeit a bit checkered, history in psychophysiology. These terms generally denote a continuum of energy mobilization, from the low end of coma or deep sleep to the high end of excitement and extreme agitation. Primary psychophysiological indices of the state of arousal are levels of activity of the ANS, skeletal electromyographic (EMG) tension, and EEG. In keeping with the focus of this chapter, we will limit our discussion to the ANS and EMG measures.

Arousal has been manipulated in psychophysiological laboratories in a variety of ways, including having subjects perform a difficult arithmetic task; be threatened with impending electric shock; blow a balloon to bursting; and hold their hands in extremely cold water. In each of these tasks there is typically a significant increase in activity of the sympathetic division of the ANS (e.g., increased palmar skin conductance and heart rate) and EMG. In more 'real world' situations, arousal has been observed in parachutists just before jumping from an aircraft, and from students just before taking critical exams. Again, there is evidence of increased activity in the sympathetic division of the ANS and EMG.

Some problems with the simple global arousal concept were emphasized by Lacey (1967). Most important for the present discussion, he and his colleagues showed that the various psychophysiological indices of arousal (e.g., skin conductance and heart rate) were often not significantly correlated and, even more embarrassing for a simple arousal concept, these measures would go in opposite directions in many situations. For example, when subjects watched colored lights or listened to tones, skin conductance would increase but heart rate would decrease. Thus, different stimulus situations would reliably produce different patterns of physiological responding ('situational stereotypy' or 'stimulus specificity').

To this day arousal is a useful descriptive concept, but psychophysiologists remain careful not to overgeneralize from one response system to another, or from one stimulus situation to another. There is a far greater degree of specificity of response systems and situational reactions than was recognized by the early global arousal/activation concepts.

2.2 Emotion

The study of emotion has also had a long history in psychology and psychophysiology. Advances with both EMG and ANS measures since the 1960s have shed new light on our understanding of emotions.

Movement of facial muscles is a common way of expressing and communicating emotion and therefore facial EMG offers one measure of emotions. Even small covert changes in facial expression can be detected as small electrical changes with surface EMG recordings from various muscles of the face. Several researchers have found that the muscle that pulls up the corner of the lip as in a smile (*zygomaticus major*) is more active when subjects report positive emotion, whereas the muscle that pushes the eyebrows together as in a frown (*corrugator supercilii*) is more active when subjects report negative emotion. Whether facial EMG can distinguish various specific emotions (e.g., anger, fear, jealousy, disgust, etc.) is more controversial. Even more controversial is whether ANS responses can distinguish between different specific emotions. It is generally agreed that ANS activity occurs during emotional states, but the degree of specificity of that activity to different emotions remains an interesting and unsettled issue.

Another specific EMG measure that has proved useful in the study of emotions is the acoustic startle reflex. The acoustic startle response includes a series of involuntary muscle movements following a loud abrupt noise. One of the fastest and more reliable of these movements in humans is the eyeblink reflex. The eyeblink reflex can be measured easily by placing electrodes on the surface of the skin over the *orbicularis oculi* muscle just below the lower eyelid. For an overview of the methods and many implications associated with this measure, see Dawson et al. (1999). An important characteristic of the eyeblink startle reaction is that its magnitude is enhanced in the presence of unpleasant negative emotions and is diminished in the presence of pleasant positive emotions. This has been demonstrated many times by startling human subjects while they view unpleasant pictures (e.g., mutilated face) or pleasant pictures (attractive nude). It has been shown that changes in the acoustic startle eyeblink reflex can provide valid psychophysiological indices of the emotional state of the individual at the time that the reflex is elicited. A review of this literature and a proposed theoretical account can be found in Bradley et al. (1999).

3. Transient Stimulus-elicited Responses

3.1 Orienting, Defensive, and Startle Responses

Psychophysiologists who focus on ANS and skeletal motor activity distinguish among three broad types of

transient responses seen in reaction to a wide variety of stimuli: the orienting response, the defensive response, and the startle response (Graham 1992). The orienting response is the response to any novel or significant, but non-aversive, stimulus. It is thought to signal a shift of attention to that stimulus and to facilitate stimulus processing. Its components include a transient increase in palmar skin conductance, heart rate deceleration, constriction of the blood vessels in the skin (usually measured in the fingertips), and dilation of the blood vessels of the head. The defensive response is the response to aversive stimuli, including painful or threatening stimuli. Its components include an increase in palmar skin conductance, an increase in heart rate, and vasoconstriction in both the blood vessels of the skin and the head. The startle response is the reaction to unexpected stimuli that are intense and have a rapid onset. While there are ANS components of the startle response, the component that has been most investigated, because of its reliable occurrence and slow habituation, is the startle eyeblink, as discussed above.

3.2 Classical Conditioning

A paradigm in which stimulus-elicited responses are studied that has been of continuing interest to psychophysiologists is classical conditioning. In paradigms in which the conditioned stimulus (CS) is several seconds long, with the unconditioned stimulus (UCS) occurring at CS offset, the form of the conditioned response (CR) is generally observed to be that of a heightened orienting response to the CS (Hugdahl 1995). A general question frequently investigated has concerned the relationship between classical conditioning, a very simple form of learning, and the more complex forms of learning of which humans are capable, including verbally mediated processes. For instance, Dawson and his colleagues (Dawson and Schell 1985) studied the relationship between acquisition of the conditioned skin conductance response and the human subject's conscious awareness of the relationship between the CS and the UCS, asking whether a person who is unaware of the CS–UCS contingency will show conditioning. In a series of studies in which awareness of the CS–UCS relation was prevented or delayed by a distracting secondary task, they found that a CR was seen only in subjects who became aware of the CS–UCS relation. Moreover, CRs developed only at or after the point in time in a series of CS–UCS trials when subjects indicated awareness. Thus, these results indicate that conscious relational learning is necessary for human classical conditioning of ANS responses with neutral CSs. On the other hand, the relationship between awareness and conditioning may be different with certain paradigms used to establish the skeletal eyeblink conditioned response (Clark and Squire 1999).

In most psychophysiological studies of classical conditioning, the CSs have been fairly neutral stimuli such as simple tones or colored lights. In contrast, Öhman and his colleagues (see Hugdahl 1995) conducted a series of interesting studies of a different class of CSs, those variously characterized as potentially phobic, fear-relevant, or biologically prepared, such as pictures of spiders, snakes, or angry faces. Öhman and his colleagues have found that CRs (usually the conditioned skin conductance response) conditioned to potentially phobic stimuli (a picture of a snake) are harder to extinguish than are CRs conditioned to neutral stimuli (a picture of a flower). They require more non-reinforced presentations of the CS, and are resistant to instructional manipulation. That is, while responses conditioned to neutral stimuli can usually be abolished by instructing the subject that the UCS will no longer be delivered, so that the subject no longer has any cognitive expectancy that the UCS will follow the CS, responses conditioned to potentially phobic stimuli remain after such instructions, even in the absence of conscious expectancy of the UCS. Thus, psychophysiological techniques can be used in the laboratory to study the dynamics of 'irrational' conditioned responses, such as one may encounter in clinical phobias where the patient fears an object while realizing rationally that the object is harmless.

4. Psychopathology

Psychophysiology is particularly useful in the study of psychopathology because it offers nonverbal measures of psychological states and processes. It thus opens a window onto events that psychiatric patients are unlikely to be able to report verbally. Many psychophysiological measures have been studied with psychopathological patients but space limitation requires that we mention briefly only a few measures in schizophrenia and psychopathy. For a general review of a greater variety of measures and a greater variety of psychopathological conditions, see Keller et al. (2000).

4.1 Schizophrenia

Schizophrenia is a psychotic disorder characterized by hallucinations, delusions, and disorganized thinking and behavior. With ANS measures, especially skin conductance activity, there is usually tonic sympathetic hyper-arousal. On the other hand, with phasic orienting response ANS measures, there is usually hypo-reactivity, at least in sizable sub-groups of patients (for a review see Öhman 1981). Among the many unsettled issues is whether these ANS abnormalities are state-related and/or trait related. That is, if ANS abnormalities are present even before the schizophrenia patients show symptoms, and persist

after the symptoms disappear, they are considered trait measures. However, if the abnormalities are present only when the patients are exhibiting psychotic symptoms then they are considered state related. Although this issue is still unsettled, preliminary evidence suggests that tonic autonomic hyper-arousal is state related, whereas phasic autonomic hypo-reactivity may be trait related, especially when it is expressed relative to their general arousal state (Dawson et al. 1994). These findings suggest that hypo-reactivity may be related to a continuing vulnerability to schizophrenia, whereas the hyper-arousal may be a separate phase associated with the development of symptoms.

4.2 Psychopathy

Psychopathy is a serious personality disorder characterized by antisocial behavior, untruthfulness, irresponsibility, and lack of remorse or empathy. It has been hypothesized that the psychopath is primarily deficient in fear reactivity, and consistent with this hypothesis is the reliable finding that psychopaths show less fear conditioning as indicated by poor skin conductance classical conditioning (Lykken 1957).

Another psychophysiological measure of fear and negative affect mentioned earlier involves the acoustic startle eyeblink reflex. Recall that normal subjects give enhanced eyeblink startle reactions while observing unpleasant pictures, compared to startle while observing neutral or pleasant pictures. This is not true of psychopaths, or at least of significant subgroups of psychopaths (Patrick et al. 1993). This finding also is consistent with the hypothesized deficit in fear in some types of psychopathy.

5. Applied Psychophysiology

In addition to their use in investigating many questions of broad theoretical interest in behavioral and social sciences, psychophysiological techniques have also been used in a number of ways to deal with specific applied problems. The two most often encountered areas of applied psychophysiology are biofeedback and the detection of deception (lie detection).

5.1 Biofeedback

In the early development of learning theory, it was generally believed that while responses of the skeletal motor system could be instrumentally conditioned, responses of the ANS could only be classically conditioned. Because they were not under conscious, voluntary control, ANS responses would not respond to operant reinforcement contingencies. However, beginning in the early 1960s, investigators such as H. Kimmel, B. Engel, N. Miller, and D. Shapiro began to demonstrate that rewarding organisms, including humans, for specific ANS activity increased the amount of such autonomic activity, and that this could occur whether or not a human subject was aware of the reinforcement contingency. Thus, it was demonstrated that people could learn to increase the size and number of skin conductance responses, or to increase or decrease heart rate or blood pressure, if they were rewarded for doing so by the arrival of pleasant stimuli or the avoidance of unpleasant ones. It was soon discovered that people could learn these things if their only reward was being told that they were acquiring a desired response or performing a task set for them by, for instance, lowering the pitch of a tone that was controlled by their heart rate, or moving the needle on a dial that indicated blood pressure.

These findings were soon applied to treating the symptoms of many psychosomatic disorders. Such disorders typically involve ANS dysfunction (migraine headache, hypertension, Raynaud's syndrome) or loss of control of skeletal muscle tension (tension headache). In clinical biofeedback applications, the patient is given an external stimulus, such as a tone or meter reading, that indicates the level of muscle tension, cerebral vasoconstriction, blood pressure, etc., and the patient's task is to lower the pitch of the tone, move the meter indicator to the left, etc. This approach has been used with varying degrees of success as a therapeutic method in a wide variety of psychosomatic disorders (Carlson et al. 1994), and the training parameters that maximize its effectiveness have been widely studied.

5.2 Detection of Deception

Use of psychophysiological techniques in the detection of deception began in the early part of the twentieth century. There are principally two different techniques in use, each with its strong advocates and critics. Both rely upon the fact that stimuli or information that are of significance or importance to a person will elicit greater ANS activity than will less significant, important stimuli. The Guilty Knowledge Technique, developed by Lykken (see Lykken 1981), relies on there being information known to the guilty person in a crime and to the polygraph examiner, but not to the innocent person. Thus, if a victim were wearing a blue shirt, a guilty person will orient more to the word 'blue' than to other colors when asked the color of the shirt and presented with alternatives, whereas the innocent person will not show differential orienting. The Control Question Technique, first well described by Reid and Inbau (1977) and extensively investigated by Raskin and his colleagues, compares responses to 'critical' questions about a particular crime ('Did you steal that red Porsche last Friday'?) with 'control' questions that are presumed to elicit greater arousal among those innocent of that crime, and in fact that are presumed to elicit lying by the innocent, than do the critical questions ('Have you

ever stolen anything of value'?). Laboratory investigations with 'mock crimes' have indicated 90 percent or better accuracy rates with these techniques, with the most common error by far being the false guilty finding. However, a number of investigators have raised questions about the accuracy rate in real-world situations, particularly with respect to whether or not adequate control questions could be found when a person is accused of a serious crime.

6. Conclusion

Psychophysiology offers potentially powerful nonverbal, noninvasive, methodologies with which to study psychological processes in real time. Cognitive processes and emotional processes are studied in both normal and abnormal populations using psychophysiological measures and concepts. Thus psychophysiology lies at the interface of cognitive science, clinical science, and neuroscience and offers the capability to bridge these disciplines.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Autonomic Classical and Operant Conditioning; Classical Conditioning, Neural Basis of; Cognitive Neuropsychology, Methodology of; Electroencephalography: Basic Principles and Applications; Emotion, Neural Basis of; Event-related/Evoked Potentials; Fear: Potentiation of Startle; Magnetoencephalography

Bibliography

Bradley M M, Cuthbert B N, Lang P J 1999 Affect and the startle reflex. In: Dawson M E, Schell A M, Böhmelt A H (eds.) *Startle Modification: Implications for Neuroscience, Cognitive Science, and Clinical Science*. Cambridge University Press, Cambridge, UK

Cacioppo J T, Tasinary L G, Bernston G G 2000 *Handbook of Psychophysiology*. Cambridge University Press, Cambridge, UK

Carlson J C, Seifert A R, Birbaumer N 1994 *Clinical Applied Psychophysiology*. Plenum, New York

Clark R E, Squire L R 1999 Human eyeblink classical conditioning: Effects of manipulating awareness of the stimulus contingencies. *Psychological Science* **10**: 14–18

Dawson M E, Nuechterlein K H, Schell A M, Gitlin M, Ventura J 1994 Autonomic abnormalities in schizophrenia: State or trait indicators? *Archives of General Psychiatry* **51**: 813–24

Dawson M E, Schell A M 1985 Information processing and human autonomic classical conditioning. In: Ackles P K, Jennings J R, Coles M G H (eds.) *Advances in Psychophysiology*, JAI Press, Greenwich, CT

Dawson M E, Schell A M, Böhmelt A H 1999 *Startle Modification: Implications for Neuroscience, Cognitive Science, and Clinical Science*. Cambridge University Press, Cambridge, UK

Graham F K 1992 Attention: The heartbeat, the blink, and the brain. In: Campbell B A, Hayne H, Richardson R (eds.) *Attention and Information Processing in Infants and Adults: Perspectives from Human and Animal Research*, Lawrence Erlbaum, Hillsdale, NJ

Hugdahl K 1995 *Psychophysiology: The Mind-Body Perspective*. Harvard University Press, Cambridge, MA

Keller J, Hicks B D, Miller G A (2000) Psychophysiology in the study of psychopathology. In: Capioppo J T, Tassinary L G, Bernston G G (eds.) *Handbook of Psychophysiology*, Cambridge University Press, New York

Lacey J I 1967 Somatic response patterning and stress: Some revisions of activation theory. In: Appley M H, Trumbull R (eds.) *Psychological Stress: Issues in Research*, Appleton-Century-Crofts, New York

Lykken D T 1957 A study of anxiety in the sociopathic personality. *Journal of Abnormal and Social Psychology* **55**: 6–10

Lykken D T 1981 *A Tremor in the Blood*. McGraw-Hill, New York

Öhman A 1981 Electrodermal activity and vulnerability to schizophrenia: A review. *Biological Psychology* **12**: 87–145

Patrick C J, Bradley M M, Lang P J 1993 Emotion in the criminal psychopath: Startle reflex modulation. *Journal of Abnormal Psychology* **102**: 82–92

Reid J E, Inbau F E 1977 *Truth and Deception: The Polygraph 'Lie-detector' Technique*, 2nd edn. Williams and Wilkins, Baltimore, MD

M. E. Dawson and A. Schell

Psychosomatic Medicine

In its simplest sense, 'psychosomatic medicine' merely indicates that there is a (psyche) mind–body (soma) relationship relevant to human health. Drawing from this, many nonprofessionals believe that it is a medical specialty that concentrates on the psychological influences on disease, particularly certain 'psychosomatic diseases.' Actually, it is a health science that is concerned with the nature of the mind–body relationship, including its clinical applications.

1. Definition

More specifically, psychosomatic medicine is a field of health science based on the concept that the human organism is a *psychosomatic unity*, that biological and psychological processes are inextricably inter-related aspects of its function. This stands in contrast to the still predominant, traditional biomedical model (see Engel 1977) from the nineteenth century which holds that diseases represent cellular or organ pathology arising from intrinsic defects or external bio-physicochemical agents. Psychosomatic medicine conceptualizes diseases as occurring in persons, not their components, as a consequence of their transactions with an environment that is informational, social, and cultural, as well as bio-physicochemical. It notes that people are not simply biological organisms, but sentient ones with motives, thoughts, feelings, and relationships.

Mind–body unity should not be misunderstood to imply that mental and somatic phenomena cannot be

separately studied. Such reductionism is a legitimate tactic that permits the study of scientific questions limited to that system. Each organ or part function merits study by techniques appropriate to it. To understand *organismic* function, however, requires a perspective that includes data from all body systems, including the psychological. Psychosomatic medicine provides no theories about the nature of the underlying processes. What it provides is its holistic perspective for interpreting them. It draws on the data, theory, and techniques of all other relevant fields, integrating these in unique ways informed by its conceptual base. Thus, contrary to common lay belief, psychosomatic medicine is not a medical specialty. Nor is it a scientific discipline. Its practitioners come from a range of disciplines and clinical specialties.

Some have accused psychosomatic medicine of overweighting the importance of psychosocial factors in disease, trivializing biology. This had legitimacy during the 1940s and 1950s when a prominent segment of the field fell into this trap (see Sect 2.3). If this seems still true, that is because its main effort is directed to clarifying the effects of psychosocial factors on the body. But that emphasis is necessitated by the neglect of this work by the remainder of biomedicine. Psychosomatic medicine is concerned with *all* mind–body relationships, giving equal importance to biology.

Doubts have been raised about the aptness of the term. To some, the bipartite nature of 'psychosomatic' undermines the intended unity. But it is one, unhyphenated word. Others question the absence of reference to the social, noting that many relevant factors are such. Engel's (1977) alternative, 'biopsychosocial model' was meant to resolve that, as well as to circumvent the residual misinterpretation that biology is devalued. But if social is included, what about cultural? Indeed, the list could include economic and other factors. Moreover, these factors can impinge on the organism only via the psychological apparatus. 'Psychosomatic medicine' also has the virtue of historical continuity that includes a mass of research documentation.

2. History

2.1 Origins

From earliest times, people have been struck by instances in which physical symptoms followed noteworthy life events, and reports of these appear in ancient writings. With the beginnings of medical science, physicians added more such reports as well as describing illness outbreaks during wars and major upheavals. The idea that stressful life experiences caused illness gained wider currency. During the nineteenth century, philosophical conjectures about the holistic quality of human nature developed. This was the period when 'psychosomatic' came into use by physicians, although in a sense partly different from our own.

2.2 Early Twentieth Century

With the turn of the century, Cannon's (1915) studies on the physiology of the emotions added key data revealing how these physiological responses were *adaptive* to the external threats that had triggered the emotions. Building on Claude Bernard's earlier work, he indicated (1932) how his findings fit into the process by which organisms maintain their integrity by adapting to environmental change via internal responses to maintain internal equilibrium: 'homeostasis.'

Clinical and epidemiological studies of better quality increased. By 1935, (Helen) Flanders Dunbar was able to cite some 2,251 references in her groundbreaking work, *Emotions and Bodily Disease*, which ushered in the modern, scientific era of psychosomatic medicine.

Further stimulus came from World War II. With astonishing foresight, the (US) National Research Council, anticipating casualties from war stress, convened a series of conferences among psychosomatic experts during the mid-1930s. This kindled additional research which culminated in 1939 with publication of the journal *Psychosomatic Medicine*, still the preeminent periodical in the field. Three years later, the American Psychosomatic Society was founded to foster scientific interchange and education. US military physicians observed how war trauma led to physical as well as mental disorders, and many returned with an interest in psychosomatic medicine. Psychosomatic societies arose in other countries and, later, international meetings began.

Psychoanalysis provided further impetus. Freud and his followers were greatly interested in physical symptoms. Much of their theorizing involved a simplistic transfer of the psychoanalytic psychology of hysterical symptoms, without regard to physiology. Nevertheless, as psychoanalytic activity flourished, it stirred interest among intellectuals, adding to the popularity of psychosomatic medicine. With this came the naive belief that many medical disorders were primarily psychogenic.

2.3 Specificity Theories

Paralleling major biomedical advances that elucidated the etiology and pathogenesis of disease, psychosomatic medicine undertook a search for specific psychological causative factors. Dunbar was a pioneer here too, observing that patients with several medical disorders had distinctive personality traits which she concluded were etiological. Others developed similar formulations. The identified diseases were termed

'psychosomatic.' This effort peaked with the contributions of Franz Alexander (1950). As a psychoanalyst, he focused on unconscious processes, but his work was far more elaborate and sophisticated than others. Most importantly and enduringly, he clarified that no purely psychological mechanism could explain the physiological processes of internal organs. Instead, Alexander reported that 'specific dynamic constellations' of motivations, defenses, and emotions characterized patients with each of a series of 'psychosomatic' diseases, and that life experiences which matched the vulnerabilities inherent in the particular constellation precipitated that illness. He also linked each constellation to specific physiological processes mediated through autonomic neural pathways that plausibly explained the pathogenesis of that disease. The brilliance of this work captured the imagination of many, and it became a major force in psychosomatic medicine. Belief developed that psychotherapy might cure 'psychosomatic diseases.'

2.4 *The Last Third of the Twentieth Century*

More careful studies called into question much specificity research, which had been uncontrolled for observer bias, and utilized selected subjects. Skepticism grew and spread from specificity to involve all of psychosomatic medicine. Moreover, remarkable advances in biomedicine made other factors seem irrelevant: 'With penicillin, who needs psychology?' The influence of the field on the rest of medical science and practice withered.

Experiencing self-doubt and loss of external support, the field contracted and changed. Grand theories were eschewed. Research focused on better-defined questions and became methodologically tighter. Clinical studies became less common, especially those of therapy. Laboratory projects concentrated on elucidating basic psychophysiological relationships, often in normal subjects or even animals. If less exciting, this shift proved salutary, strengthening the scientific base and the quality of data.

Over time, the field recovered. Increasing studies of clinical problems embodied a new methodological rigor gained from models provided by the continuing basic research. A variety of subgroups within the psychosomatic society grew to a critical size and formed associations of their own, expanding work further. Acceptance by clinical biomedicine grew more slowly, however.

3. *Theoretical Models of Mind–Body Relationships*

The evidence of correlations between psychological and somatic processes is extensive. But how can this be understood? Five explanatory models can be adduced from scientific reports.

3.1 *Concomitant Effects*

This model rejects causality in the relationship. It postulates that both the psychological and physical findings are products of another, precedent factor. Commonly, this is said to be genetic. Another explanation draws on the fact that life events often have multiple stimulus properties. Thus, one stimulus is said to cause the psychological findings while another produces the physical one.

3.2 *Somato → Psychic Causation*

This model posits that the relationship follows entirely from the effects of somatic processes on the mind. This is the traditional biomedical view, which sees all disease as 'physical' in nature and origin. In a modification, a feedback loop is added going back from the mind to the body, acknowledging that psychological changes may have physical effects, but only secondarily. (If this is true, why cannot they do so primarily?)

3.3 *Psycho → Somatic Causation*

This is the counterpart of the preceding. It states that psychological responses to external events cause the somatic changes. Most commonly, stress or strong emotions are invoked as intervening mechanisms. Weiner (1977) noted that it may be the behavioral concomitants of the psychological processes which act on the body: 'lifestyle factors' such as smoking or substance abuse. Except for 'specificity' theories already described, the 'choice' of the resultant disease is usually left to other, biological factors. Again, secondary feedback is often added, bodily changes acting back on the mind.

3.4 *Bidirectional Psycho ⇔ Somatic Causality*

This is a combination of the last two models, allowing for causation in both directions and the feedback variations of each. Thus, longer causal chains are possible: … psycho … somato … psycho … somato … (etc.); but these remain linear.

3.5 *The Modern Psychosomatic Model*

This is a holistic model that includes all the preceding models and their variants, indicating that each explains some psychosomatic processes. Beyond this, it suggests that several are usually operating simultaneously, and to this adds that there are usually *webs of relationships* involving multiple stimulus–response linkages. To explain this, it invokes data about the operation of the mechanisms which integrate human function.

4. Integrative Mechanisms

From the biological side, *the brain* is the organ of the mind; mentation is subject to the same physical influences as all other organ functions. Beyond this, it is the body's chief regulatory and coordinating organ. Inputs to the brain come from both the body and environment. Similarly, its output alters bodily functions and, via behavior, the external world. These inputs from inside and out converge on many of the same brain centers, and the resultant messages alter both psychological and somatic functions.

From a psychological perspective, input consists not merely of sensations but of information organized into meanings. Meanings have individual as well as shared components. The same event or somatic message has different nuances, at least, for each person. The origins of these meanings, as well as those shared in the culture, derive from life experiences: they are learned. Some visceral responses to psychological stimuli also are 'learned,' but the learning that establishes these psychosomatic linkages is not the familiar, conscious process, but *conditioning*, of both Pavlovian and operant types. Most of this occurs in early life. Manipulations during early development can alter brain anatomy, neurotransmitters, and hormone levels, and modify subsequent stress responses.

When important meanings are attached to inputs, emotions are activated. These further affect body functions. The original response becomes a stimulus, too. In studying organismic function, what is defined as stimulus, and what as response, is arbitrary; it depends on what part of the process is being examined and at what point in time. Nor is this a simple chain of stimulus–response pairs. A network of actions follow, in which some responses affect one organ, some another, and some both, while some of these influences are stimulating and others inhibitory; and the succeeding responses become additional stimuli of the same variegated nature. Because many life events have more than one meaning, several such networks may be in process simultaneously. In most instances, then, a stimulus sets off *a transacting web of responses and new stimuli with both psychological and somatic components*.

Ordinarily, inhibitory processes finally dominate, and the person returns to a resting state. Sometimes, especially if the stimulus persists, inhibition may fail, so that the reaction continues or recurs. These extended responses are homeostatically maladaptive, and likely to cause organ damage and disease.

The *endocrine and immune* systems also serve as integrators of organismic function, directly and in their inter-relationships with each other and the brain. Pituitary secretions, which control many other endocrines, are themselves subject to control by hormones and neurotransmitters from the hypothalamus. Psychological stress is associated with stimulation of adrenocortical hormones via the pituitary, as well as neurally mediated adrenaline release from the adrenal medulla. The distinction between neurotransmitters and hormones is blurred: active neuropeptides and neuroamines are present in many peripheral organs. Hormones act back on the brain to inhibit their own overproduction, but also modify cognitive and emotional states.

Stress also affects both cellular and humoral immune functions, partly through the action of adrenal steroids; depression may have especially potent effects. Animals subjected to a variety of psychological stressors have increased morbidity and mortality to injected pathogens along with changes in immune functions. Hypnosis can cure warts and alter the hypersensitivity reaction to injected antigens. The immune system can 'learn' too: immune responses to previously inert substances can be conditioned.

5. Basic Psychosomatic Concepts

5.1 Health and Disease

Psychosomatic medicine rests on the fundamental physiological concept of homeostasis: the human organism exists in a state of constantly dynamic equilibrium subject to both endogenous shifts and perturbations induced by an ever-changing environment. These trigger homeostatic responses which keep the changes from exceeding viable limits, thereby maintaining organismic integrity and permitting continuing function in the environment. If the change is excessive or the compensatory adjustments become so, damage results. When damage significantly impairs adaptation, disease occurs: disease is failure of adaptation. At any given moment, every person is more or less healthy and more or less diseased.

Psychosomatic medicine enhances this concept to take account of our inherent psychological and physiological properties as well as the sociocultural and informational content of the physical world. Moreover, although 'disease' may seem to have a fixed definition, different times, places, and persons have provided differing criteria for deciding what disease exists and whether one is sick at all. (Epilepsy in ancient Egyptian royalty was a manifestation of divinity.)

5.2 The Nature of Disease; 'Psychosomatic Disease'

Many physicians and others dichotomize disease states as 'organic' (physical) *or* 'psychological' (sometimes extending these, respectively, as 'real' or 'imaginary'). Specificity theorists separated off 'psychosomatic' diseases in which psychogenesis played a uniquely central role. Psychosomatic medicine takes a wholly different view, conceptualizing *every disease in every*

person as physical AND psychological. There are no 'psychosomatic diseases' because *all diseases (and health) are psychosomatic*.

Even in the simple instance of accidental injury, it is the overall impact of the event on the person, not just the injured part, that defines the illness—including instances where behavioral factors played no causal role and no psychic 'trauma' results. The extent of pain, bleeding, swelling, etc., will be determined not only by the injury but by the victim's entire prior psycho-physiological state; and the victim's reactions to each of these effects, as well as his attitudes about the injured part and its functions, will alter that state. (For example, if anger arises, blood vessels may dilate, increasing bleeding.) The reactions of others, including health care providers, will also affect the psycho-physiological responses, as will the meanings of the treatment given. The somatic problems will be paramount medically, at least until late consequences require chronic care. Conceptually, however, even this extreme example of 'pure physical illness' can be fully understood only from a psychosomatic perspective. Optimal care will result only if health care providers are responsive to psychosocial needs throughout the course of treatment.

5.3 The Etiology of Disease

Similarly, the causes of all diseases are psychosomatic. Take malaria for example. The plasmodium, usually mosquito-borne, is prerequisite. But, what is to be said of the US soldier in the tropics who contracts malaria after failing to take prescribed prophylactic medication or to sleep beneath issued netting, or who fishes in a swamp designated 'off limits.' It is no less valid to categorize such behaviors as causative than the biological factors. Sociocultural factors, even political ones can enter in too. Examples include being a parentless child, or living in a culture ignorant of the mosquito's role or under a political system too poor or uncaring to provide public health programs. Any such factors or other psychological events can be involved in any given person. Often several are.

The choice of an infectious disease was deliberate. Traditional medicine views these as the prototype of external causation. However, as Dubos (1960) indicated, all people are incubators for various microorganisms growing in their throat, intestine, etc.; and in these circumstances, why are people not sick all the time? The real question is why some people become sick with what they do, when they do. External factors explain only part of this. Host resistance variables, including differences in immunity, are necessarily involved too; and psychosocial factors affect these.

5.4 Pathogenesis; the Onset of Disease

Among the psychosocial causes of illness, *stress* is the best documented. Selye (1950) demonstrated that the response to any noxious stimulus precipitated a nonspecific 'General Adaptation Syndrome' superimposed on the specific effects of the stimulus. He focused on the adrenocortical response, but it became clear that many other systems are involved in the syndrome, especially other hormones, the cardiovascular system, blood elements, including those active immunologically, and the psychological apparatus. Psychological stressors have proved as potent as physical ones. Indeed, some effects of physical stressors have proved to result from their psychological properties. Multiple clinical studies have documented that a variety of illnesses arise in the context of accumulated life stress.

Specific emotional states seem to operate in some circumstances. Anger and anxiety have been implicated, as well as depression, especially that associated with feelings of helplessness and hopelessness.

But disease is a process with a course that evolves variably, and often preceded by a period in which covert, even undetectable, lesions are developing; and a different superimposed process may be required for the disease to appear. Psychosocial factors have been implicated in each of these phases, but the specifics may differ from one to another, as Reiser (1975) clarified. For example, the gradual accretion of coronary atherosclerosis precedes thrombosis and infarction ('heart attack'). Anger and depression have been implicated in infarction, whereas nonspecific stress has been related to alterations in cholesterol metabolism which can lead to atherosclerosis. Of course, the events that provoke emotions differ among people, depending on their psychological makeup. Research provides the broad possibilities, but only individual study can delineate the pathogenic process in a given person.

The specific 'choice' of illness likely rests in genetic and other purely biological factors in many instances. Individually patterned psychosomatic links may be implicated in others. Either may be responsible for 'response specificity,' the tendency for individuals to have particular somatic responses to a variety of stressors. Moreover, discredited but not disproved psychological 'specificity' remains a possibility. The remarkable consistency among various specificity theorists regarding the psychological characteristics in people with certain disorders remains suggestive. Formidable methodological problems must be overcome to study this adequately, but such research is merited.

6. Clinical Applications

The application of psychosomatic medicine to patient care rests primarily on its basic concepts. As every patient is partly unique, only by identifying the issues specific to that individual can appropriate care be provided. The fundamental need both in diagnosis and treatment is establishing a trusting professional

relationship that allows patients to relax defenses and reveal personal information and feelings. Also required are interviewing skills to facilitate patients' communications of their life experiences, along with physical symptoms and the life context of the latter. This must be accompanied by sufficient knowledge of psychology and social science to identify the significance of these communications, and also of major psychosomatic research findings in order to identify correlations likely to be significant.

Treatment requires the ability to utilize this understanding of the patient in order to select the biological therapies that are most suitable for the individual, and to provide this in a manner that helps that patient use it best. On the psychological side, the need is to quantify and facilitate the extent of emotional release most helpful and tolerable to that patient, as well as the level of insight to pursue the relevant psychosomatic correlations.

All this seems complex, and it is. But the knowledge and skills can be acquired in the course of existing professional education, provided that psychosomatic medicine is integrated throughout. This is true for *all clinical specialties*. As noted initially, there is no specialty of psychosomatic medicine.

At the height of interest in psychosomatic medicine, a cadre of experts began to serve as educators of physicians, medical students, and others in academic health centers. Most were psychiatrists, although some internists and other specialists participated. This became a new subspecialty, named 'Consultation-Liaison Psychiatry,' as the service function became melded with education (see *Liaison and Consultation Psychiatry*). In the late twentieth-century, however, as psychiatry became more biological and positioned within the biomedical model, the psychosomatic element diminished. Given the widespread public dissatisfaction with the 'over(bio)scientific,' 'dehumanized' quality of medical care, this has been a significant loss.

But clinical psychosomatic medicine should not be confused with humanism or doing 'good'; it is applied *science*. Although its clinical practitioners act in a decent, humane fashion, it is not this but their skills and knowledge as applied scientists that make them effective.

7. Recent Developments

Substantial research has flourished under the rubrics of 'health psychology,' 'behavioral medicine,' and 'behavioral psychology' as well as psychosomatic medicine itself. Consequently, high-quality research has been advancing vigorously. Psychophysiological studies have begun to use highly sophisticated probe techniques to measure meaningful physiological parameters. A particularly new and useful approach involves the application of randomized clinical trials that elucidate both causal factors and the effectiveness of interventions in a number of diseases. New research designs have incorporated more real-life situations, particularly in research on stress. Information particularly relevant to actual clinical and life situations has emerged, giving meaningful promise for the future.

Bibliography

Alexander F 1950 *Psychosomatic Medicine*. Norton, New York

Anderson N B, Armstead C A 1995 Toward understanding the association between socioeconomic status and health: A new challenge for the biopsychosocial approach. *Psychosomatic Medicine* **57**: 213–25

Cannon W B 1915 *Bodily Changes in Pain, Hunger, Fear and Rage*, 2nd edn. Appleton, New York

Cannon W B 1932 *The Wisdom of the Body*. Norton, New York

Cobb S 1976 Social support as a moderator of life stress. *Psychosomatic Medicine* **38**: 300–14

Dubos R 1960 *The Mirage of Health*. Allen and Unwin, London

Dunbar H F 1935 *Emotions and Bodily Change*. Columbia University Press, New York (see also the 4th edn., 1954)

Engel G L 1962 *Psychological Development in Health and Disease*. Saunders, Philadelphia, PA

Engel G L 1977 The need for a new medical model: A challenge for biomedicine. *Science* **196**: 129–36

Grinker R R 1953 *Psychosomatic Research*. Norton, New York

Kiecolt-Glaser J K, Glaser R 1988 Methodological issues in behavioral immunology research with humans. *Brain Behavior and Immunology* **2**: 67–78

Kleinman A, Becker A E 1993 'Sociosomatics': The contributions of anthropology to psychosomatic medicine. *Psychosomatic Medicine* **60**: 389–93

Lazarus R S 1993 Coping theory and research: Past present and future. *Psychosomatic Medicine* **55**: 234–47

Mason J W 1968 Organization of psychoendocrine mechanisms. *Psychosomatic Medicine* **30**(Suppl.): 365–808

McEwen B S 1998 Protective and damaging effects of stress mediators. *New England Journal of Medicine* **338**: 171–9

Moynihan J A, Ader R 1996 Psychoneuroimmunology: Animal models of disease. *Psychosomatic Medicine* **58**: 546–58

Oken D 1989 Current theoretical concepts in psychosomatic medicine. In: Kaplan H I, Sadock B J (eds.) *Comprehensive Textbook of Psychiatry/V*, 5th edn. Williams and Wilkins, Baltimore, MD, Chap. 25.2

Reiser M F 1975 Changing theoretical concepts in psychosomatic medicine. In: Reiser M F (ed.) *American Handbook of Psychiatry, Volume 4*, 2nd edn. Basic Books, New York

Reiser M F 1984 *Mind, Brain, Body: Toward a Convergence of Psychoanalysis and Neurobiology*. Basic Books, New York

Selye H 1950 *Stress*. Acta, Montreal, Canada

Stainbrook E 1952 Psychosomatic medicine in the 19th century. *Psychosomatic Medicine* **14**: 211–27

Weiner H 1977 *Psychobiology and Human Disease*. Elsevier, New York

Weiner H, Hofer M A, Stunkard A J (eds.) 1981 *Brain, Behavior and Bodily Disease*. Raven Press, New York

D. Oken

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Psychotherapy and Hermeneutics

Well into the 1970s the atmosphere in the field of psychotherapy was characterized by a war of the schools. Psychoanalysis in its various forms, the humanistic movement, behaviorism, and cognitive therapy were the major players in a battle that included an ever growing number of contenders. The main assumption was that one theory was right and all the others were wrong; the question was which one would carry the day.

The statistical meta-analyses of psychotherapy outcome studies produced a result that called this basic assumption into question. It turned out that by and large all psychotherapeutic methods were helpful, and that the theory and technique applied was a very poor predictor of outcome. The question therefore became what metatheoretical framework could make sense of what seemed a conundrum: It simply didn't seem to matter how you worked. If anything it was important that client and therapist had a good rapport, and that the therapist had some convincing narrative framework that made sense of the client's suffering and problems (e.g., Orlinsky and Howard 1986).

This in turn squared well with positions taken by anthropologist Claude Lévi-Strauss ([1949] 1963) and psychiatrist Jerome Frank (1963). Both claimed that twentieth century Western psychotherapy was just a subcase of practices that were to be found in almost every existing culture. Lévi-Strauss's classic paper argued that the shamanic healer and the psychoanalyst did essentially the same thing: They integrated meaningless and therefore frightening phenomena into a symbolic system. Frank's model was that psychotherapy included three factors: a socially established status of healer, a myth or narrative that was applied to the patient's condition, and a ritual that structured the application of the myth to the patient's condition. Essentially the difference between Western and other forms of therapy only concerned the institutions that conferred status of healership, the myths that were applied, and the rituals that were enacted. These theoretical positions foreshadowed the empirical result of psychotherapy research. It turned out that a hermeneutic understanding of psychotherapy was one of the better candidates to present the profession with a new self-understanding.

1. *The Nature and History of Hermeneutics*

Hermeneutics originally denoted the discipline of interpreting texts. The place of origin of hermeneutics was the interpretation of the holy scriptures of the various religions, which entailed a variety of problems: how to derive laws from texts, and how to find out what the—presumably divine—intention behind the text was. There were several attempts to codify hermeneutic rules that were guided by specific religious, political, and social needs. The assumption of classical hermeneutics was that there was one right interpretation; the question was how to reach it.

It was in the nineteenth century, with the rise of philology and the *Geisteswissenschaften* as academic disciplines that methodological self-reflection of hermeneutics achieved a different turn. The sharpening of historical consciousness made it evident that the model of the one meaning behind the text (determined by the author's intention) was problematic. Furthermore it became evident that interpretation varied according to the historical position of the interpreter.

In the twentieth century the groundwork for philosophical hermeneutics was laid by Martin Heidegger's *Being and Time* ([1927] 1962). In it Heidegger claimed that the idea of the human subject as a mind contemplating the world was flawed and misleading. Human beings were primarily involved in what he called 'being-in-the-world.' This mode of being involved a pre-reflective interpretation of the world within which subjects projected themselves into a future on the basis of an understanding of their past.

Interpretation was therefore not an activity confined to the reading of texts, but was essential to the human condition as such. There could be no such thing as meeting an object (or another subject) outside an interpretive framework. This preunderstanding was not created by individual subjects through conscious intention. Instead it was the horizon within which all action and understanding took place.

Heidegger's student Hans-Georg Gadamer can be credited with having written the major text of philosophical hermeneutics, *Truth and Method* ([1960] 1975), taking Heidegger's ideas as its foundation. Gadamer claimed that human beings always lived within a horizon of intelligibility, a preunderstanding that guided their whole mode of being. The understanding of texts was just a special case of the activity of understanding essential to human life.

The question of interpretation did not arise if two subjects moved in the same horizon of intelligibility. Understanding each other was then a matter of course as in everyday interactions with people and texts (like newspapers or bank statements). Interpretation becomes problematic only when the speaker's (or author's) and the listener's (or reader's) horizon or intelligibility diverges.

If they did, the reader needed to integrate the text into his or her own horizon of intelligibility, as there can be no understanding outside this context. But in doing so, the reader expands this very horizon of intelligibility. If for example contemporary readers try to understand a Platonic dialogue, they may find out that their horizon of intelligibility simply does not fit some of the assumptions needed to make the Platonic text intelligible. They may therefore try to reconstruct the horizon of intelligibility implicit in the Platonic text.

By doing so, they may come to the point of understanding Plato better than Plato understood himself. Plato (like any one of us) could not be fully aware of the horizons of intelligibility, the pre-understanding of the world that made his world and his life meaningful. Human beings are never completely transparent to themselves, as we are embedded in historical contexts that are not fully understood.

Historical distance therefore creates both the necessity for interpretation and the perspective that allows the interpreter to understand aspects of meaning that were not accessible to the author. At the same time the activity of interpretation also changes the interpreter. The integration of historically (or culturally) distant horizons of intelligibility makes the interpreter more conscious of his or her own point of view, and thus changes and widens it.

Philosophical hermeneutics provided an interpretive frame that made it clear why there could never be a final understanding of any text. Every historical period, and every cultural position, needs to integrate texts and cultural practices that are historically or culturally distant into its own horizons. The interpretive act is therefore much closer to a dialog between interpreter and text (or cultural practice) than to the unearthing of a pre-existing meaning.

2. *Hermeneutics and Psychotherapy*

The hermeneutic approach to psychopathology and psychiatry was already introduced at the beginning of the twentieth century primarily through Karl Jaspers (1913) and the existential school of psychotherapy. The latter (Ellenberger and May 1958) was primarily based on the direct application of Heidegger's philosophy to the understanding of psychopathology. Psychotherapy was essentially viewed as the elucidation of the patient's modes of being-in-the-world. The point of psychotherapy was therefore to interpret and change the patients' interpretation of their *Lebenswelt*.

The European existentialist tradition in psychotherapy lost its cohesiveness as a movement, and philosophical hermeneutics began to be reapplied to psychotherapy within psychoanalysis first in Europe: Paul Ricoeur ([1965] 1970) in France, Jürgen Habermas (1971) in Germany, and then in the USA, Donald Spence (1981) and Roy Schafer (1983). Their common denominator was to argue that psychoanalysis had misunderstood its own nature by formulating itself in terms of a natural science model.

Psychoanalysis, they argued, was not a theory that discovered causal laws that connected between childhood events and psychopathological symptoms. It was the activity of helping patients to make themselves intelligible to themselves (Strenger 1991).

Schafer and Spence used literary theory in addition to hermeneutics in their attempt to clarify the nature of psychoanalysis (see *Psychoanalysis in Clinical Psychology*). In Schafer's model patients are seen as those whose narrative of themselves is overly narrow, simplified, and closed. This narrative (the term would correspond to Gadamer's 'preunderstanding') narrows down the patient's options of living, because life, human encounters, and actions are interpreted in overly rigid terms (e.g., 'every interaction is a power-struggle').

Psychoanalytic interpretation in this view is not the uncovering of a historical cause or pre-existing unconscious wish. It is the making explicit of the patients' interpretation of their world, their mind, and their interpersonal relationships. In other words: The object of psychoanalytic interpretation is the patient's self-interpretation. Its goal is to change this interpretation by making it conscious. The very act of making it conscious will lead to a change of the patient's self-interpretation.

In this manner Schafer achieved two objectives: first he freed psychoanalysis from the need of extraclinical validation of causal claims (e.g., unresolved Oedipal issues lead to psychoneurotic symptoms). Second, he made it possible to understand why analysts of different theoretical orientations would come to interpret the patient's mental life in different terms. In Gadamer's terms the analyst could only work from a particular horizon of intelligibility. Understanding the patient always meant the integration of the patient's self-understanding into a new horizon of intelligibility.

Quite quickly the application of the hermeneutic point of view began to spread to other psychotherapeutic approaches. Collections like *Hermeneutics and Psychological Theory* (Messer et al. 1988) showed that the hermeneutic point of view was, in fact, fruitfully applicable to almost every discipline in psychology. Anything from behavioral-cognitive through family therapy to social psychology could be reinterpreted in hermeneutic terms.

Since the 1990s the term 'hermeneutics' has not often been used explicitly in the psychotherapeutic literature. Nevertheless it could be said that its tenets have turned into something like a metatheoretical implicit consensus in the field. Most authors implicitly or explicitly assume that psychotherapy intervenes in the patient's meaning system. Furthermore it has become something like an implicit orthodoxy to accept that there are always many ways to construe the patient's experience.

The hermeneutic understanding of psychotherapy tends to go along with a constructivist understanding of psychotherapeutic activity: Psychotherapy does not uncover (or change) the single cause of symptoms, maladaptive character traits, feeling states, etc. It reinterprets the patient's actions, thoughts, and emotions in new terms. It is taken to be a matter of course that it is conceptually impossible to specify one correct such construal, since construals of the world are always perspectival (cf. Stolorow et al. 1987).

There are two major beneficial implications of the hermeneutic understanding of psychotherapy: first, it has led to a more ecumenical atmosphere in the field. The wars of the schools are largely over, and instead different approaches can enter fruitful dialogs. Second, it has led to greater sensitivity to the patient's right to a subjectivity of his or her own. If all psychotherapy is hermeneutic construction, it is not fruitful and legitimate to impose such construal on the patient. Current writers (e.g., Aron 1996, Strenger 1998) emphasize the necessity of turning the psychotherapeutic interaction into a more egalitarian and dialogical activity.

Meanwhile there have been extensive efforts to detect trends that favor some forms of therapy for particular disturbances (e.g., Roth and Fonagy 1996) that seem to reverse the original verdict that theory and technique do not make any difference whatsoever (e.g., obsessive-compulsive disorder favors a combination between cognitive-behavioral therapy and antidepressant medication; generalized, nonspecific character disorders favor interpersonal and psychodynamic approaches; see *Psychological Treatments, Empirically Supported*). Nevertheless these efforts have maintained the ecumenical trend in psychotherapy. The hermeneutic understanding of the nature of psychotherapy as an activity that generates meaning might well continue to serve as a framework that will enable the continuation of this discussion under the aegis of cooperation rather than competition between schools.

See also: Constructivist Psychotherapies; Hermeneutics, History of

Bibliography

Aron L 1996 *A Meeting of Minds. Mutuality in Psychoanalysis*. Analytic Press, Hillsdale, NJ

Ellenberger H, May R (eds.) 1958 *Existence*. Basic Books, New York

Frank J D 1963 *Persuasion and Healing*. Johns Hopkins University Press, Baltimore, MD

Gadamer H-G 1960 1975 *Truth and Method*. Seabury Press, New York

Habermas J 1971 *Knowledge and Human Interest*. Beacon Press, Boston

Heidegger M 1927 1962 *Being and Time*. Harper and Row, New York

Jaspers K 1913 *Allgemeine Psychopathologie*. J. Springer, Berlin

Lévi-Strauss C 1949 1963 The effectiveness of symbols. In: *Structural Anthropology*. Basic Books, New York, Vol. 1

Messer S B, Sass L A, Woolfolk R L (eds.) 1988 *Hermeneutics and Psychological Theory*. Rutgers University Press, New Brunswick, NJ

Orlinsky D E, Howard K I 1986 Process and outcome in psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York, pp. 311–84

Ricoeur P 1970 *Freud and Philosophy*. Yale University Press, New Haven, CT

Roth A, Fonagy P 1996 *What Works for Whom*. Guilford Press, New York

Schafer R 1983 *The Analytic Attitude*. Basic Books, New York

Spence D P 1981 *Narrative Truth and Historical Truth*, 1st edn. Norton, New York

Stolorow R D, Brandchaft B, Atwood G E 1987 *Psychoanalytic Treatment, An Intersubjective Approach*. Analytic Press, Hillsdale, NJ

Strenger C 1991 *Between Hermeneutics and Science. An Essay on the Epistemology of Psychoanalysis*. International Universities Press, Madison, WI

Strenger C 1998 *Individuality, the Impossible Project: Psychoanalysis and Self Creation*. International Universities Press, Madison, WI

C. Strenger

Psychotherapy and Pharmacotherapy, Combined

1. Background and Introduction

Combined psychopharmacology and psychotherapy is a treatment form quite accepted and ubiquitous in modern psychiatric practice (Beitman 1983). It may have its origins in the writings of Fromm-Reichman wherein the role and responsibilities of the analytic treating therapist were differentiated from the managerial therapist, a feature of long-term psychotherapy in the 1940s (Fromm-Reichman 1947). Further explicated by Firman, the rationale of this division of labor was based on analytic principles: the 'psychotherapist' should ideally not be in any managerial relationship with respect to the patient so that the patient's fantasies about the therapist could be viewed and analyzed as material emanating from the patient's own self (Firman 1982). According to Firman, the usefulness of this relationship is two fold: the 'outsider' (managerial) therapist can serve as a negative split object upon whom the patient can focus her own negative features. This serves to allow the tenuous whole-object relationship with the treatment therapist to proceed without jeopardy. Firman viewed the splitting that would occur in such collaborative relationships to be in a therapeutic setting and thus under better and safer control for the patient (Balon 1999, Firman 1982).

Since then, the techniques used in analytic practice and with patients in long-term psychotherapy have been transferred to almost all types of psychotherapy. Fueling this have been newer and simpler types of psychotropic medication which physicians, other than psychiatrists, feel comfortable prescribing (Horgan

1985, Regier et al. 1978, Schurman et al. 1985) and systems of care which have advocated for lower costs for patient care. Additionally, numerous research studies confirm that the combination of psychopharmacology and psychotherapy treatments are more effective than each alone (Goldsmith et al. 1999, Karasu 1992, Koenigsberg 1994, Manning et al. 1992) adding incentives for therapists to seek consultation for their patients from physicians regarding medication.

In modern psychiatry, collaborative treatment wherein a psychiatrist or other physician provides the medication and manages medical issues while a nonphysician therapist provides the psychotherapy is a common and growing treatment modality. While the field has not definitively defined this paradigm, the American Psychiatric Association published guidelines describing the responsibilities in a collaborative relationship, noting that the psychiatrist and therapist share responsibility for the care of the patient in accordance with the qualifications and limitations of each person's discipline and abilities (American Psychiatric Association 1980). Various terms have been used in the literature and in clinical care to describe this practice which include medication backup; split treatment; combined treatment; medical management; co-treatment; triangular treatment; collaborative treatment; medical backup, and parallel treatment. For the purposes of this section we will call this practice 'collaborative treatment.'

What has not been definitively researched, however, are the best methods for providing collaborative treatment. Recognizing the cognitive and emotional problems for all parties, Goldsmith and co-workers and others have sought to better explicate the potential difficulties and sources of conflict (Berg 1986, Bradley 1990, Chiles et al. 1984, Goldsmith et al. 1999, Kingsbury 1987). Others have described the positive aspects of this form of treatment (Balon 1999, Michels 1995).

The remainder of this article will focus on some of the highlights of collaborative treatment: the positive and negative aspects, legal and ethical issues, and suggestions for maximizing the collaborative process.

2. *Positive and Negative Aspects of Collaborative Treatment*

The combination of psychopharmacology and psychotherapy, or collaborative treatment, has some positive and negative aspects. As noted by Beitman (1983), collaborative treatment is based on a triangular relationship between psychiatrists, patients, and nonmedical therapists. In an ideal world, such a triangle should have just positive aspects, e.g., utilization of special talents. Triangular relationships, however, could also be a breeding ground for difficulties or negative aspects, such as splitting. The negative aspects of collaborative treatment do not arise only from patients' pathology. As Brill (1977) pointed out, 'More and more nonmedical mental health professionals began to resist and resent working under the direction of a psychiatrist. They believe that their own professions and therapeutic skills have advanced to the point that independence and self-determination were justified.' He emphasized the emergence of territorial conflicts arising in collaborative 'team' treatment. The territorial or turf conflicts and concerns are a bilateral issue—psychiatrists, as much as nonmedical therapists, frequently view the 'other side' as a competitor rather than a collaborator.

Collaborative treatment could benefit the patient only if it is truly collaborative, when positive aspects are maximized and negative aspects minimized. Awareness of the possible positive and negative aspects of collaborative treatment is the first step in maximizing the collaborative relationship (see below).

2.1 *Positive Aspects of Collaborative Treatment*

2.1.1 Utilization of special talents of both therapist and psychiatrists. As Balon (1999) pointed out, both professionals may possess expertise in special areas, e.g., mood disorders, eating disorders. Firman (1982) also pointed out that a psychiatrist who is a skillful psychotherapist could also serve as a second therapist, or a negative split object, on whom the patient focuses many of his/her own disowned negative features.

2.1.2 Cost-effective utilization of all resources. This is an issue especially important to patients. Due to various economic pressures, most patients cannot afford psychotherapy by psychiatrists unless they have good mental health care benefits. In many cases, insurance companies either do not cover psychotherapy by psychiatrists or pay an amount that is not acceptable by psychiatrists. Psychotherapy provided by nonmedical psychotherapists is considered cheaper (a contentious issue; see Goldman et al. 1998) and frequently is covered by insurance companies. Most patients thus can still benefit from the combination of pharmacotherapy and psychotherapy.

2.1.3 Better opportunity for patient to select therapist with similar ethnic and/or gender background. Some authors (e.g., Foulks and Pena 1995) have advocated that matching a patient with a therapist or

physician of a similar ethnic or cultural background may facilitate the process of psychotherapy or pharmacotherapy. Others (Yamamato et al. 1993) have warned that this matching could have its pitfalls and could be discriminatory at times. Nadelson and Zimmerman (1993) have recommended that a patient's request for a therapist or psychiatrist of a specific gender, ethnicity, or race or a patient's uneasiness with a particular therapist's or psychiatrist's gender, ethnicity, or race should be an issue explored and addressed in therapy.

2.1.4 Increased time and resources available for patients. The combination of a 20-minute medication review and 50-minute therapy session represent more resources than just one of these modalities and is an unquestionable benefit. The combination of resources is also advantageous during the temporary absence of a psychiatrist or therapist.

2.1.5 Increased amount of clinical information. There may be information that the patient feels more comfortable sharing with their therapist and vice versa. Thus, psychiatrist and therapist should set time aside to share information, impressions, and concerns and to support each other in their efforts to help patients (Pilette 1988).

2.1.6 Enhancement of compliance. As Paykel (1995) pointed out, psychotherapy may assist or possibly hinder, medication compliance. Similarly, collaborative treatment may also increase compliance with psychotherapy.

2.1.7 Professional and emotional support for both therapist and psychiatrist. As Pilette (1988) pointed out, such support is especially important when the patient is difficult to treat and also during times of crisis.

2.2 Negative Aspects of Collaborative Treatment

2.2.1 Inappropriate prescribing decisions without knowledge of the content of therapy. The prescribing psychiatrist may not always be aware of all the inner conflicts or external pressures felt by the patient. For instance, the psychiatrist may prescribe medication for anxiety that arose after an important intrapsychic conflict was addressed in therapy with the nonmedical therapist.

2.2.2 Potentially discrepant information given by the patient to each clinician. Frequently, the discrepant information provided by the patient to both treating professionals could lead to decisions with serious consequences. This issue is directly related to the next one—splitting.

2.2.3 Splitting of clinicians. This is probably the most frequently discussed negative aspect of collaborative treatment. Treatment provided by two persons provides a fertile ground for a patient to develop negative transference, to introduce problematic countertransference and to split the prescribing psychiatrist and therapist. Splitting may lead to noncompliance with any of the treatments.

2.2.4 Unclear confidentiality. As Pilette (1988) pointed out, is the collaborative treatment an occasion for free communication between a therapist and psychiatrist without the patient's consent? What if the patient asks that some important information not be related to the other treating professional? These issues are complicated and need to be discussed and resolved.

2.2.5 Clouded legal and clinical responsibility. It is not always clear who is legally responsible for various aspects of collaborative treatment. It is rather safe to assume, however, that the psychiatrist will most likely be found legally responsible for all aspects of clinical care including inappropriate or negligent supervision (Balon 1999).

2.2.6 Lack of reimbursement for collaboration. The fact that third party payors do not reimburse for collaboration somewhat hinders the collaborative treatment.

2.2.7 Various misconceptions of patient, therapist, and psychiatrist. These include issues such as therapists and psychiatrists ignoring the psychological meaning of medication, psychiatrists' belief that therapists should always agree with and support physicians' decisions, and patients' common percep-

tion that psychiatrists are only interested in prescribing medication (Goldsmith et al. 1999).

This is only a brief summary of positive and negative aspects of collaborative treatment. More thorough discussion of both positive and negative aspects is provided in the two introductory chapters of *Psychopharmacology and Psychotherapy. A Collaborative Approach* (Goldsmith et al. 1999, Riba and Balon 1999).

3. Legal and Ethical Issues

3.1 Legal

The major worry in collaborative treatment is that with two clinicians delivering the care, there will be a loss of information or understanding of what may be occurring for the patient. There are currently no standards in place for how frequently or for how long patients should be seen by physicians if collaborative treatment is provided. There is, therefore, a broad range of frequency of visits and communication patterns between the patient, therapist, and physician, which range from ideal to inadequate. Because medications are involved in collaborative treatment and historically have been the source of significant legal exposure, the issue of medical practice is an important topic for discussion. Negligence underlying malpractice liability is of two types: (a) carelessness in gathering necessary data about a patient and (b) negligence in assessing that information in determining appropriate treatment modalities, interventions, providing precautions, etc. (Macbeth 1999). Psychiatric medications, like all medications, are associated with side effects. Physicians who prescribe medications become liable if there is a deviation in the accepted uses of the medications and there results harm to the patient. It is critical for physicians to obtain informed consent from patients when prescribing medication and also to ascertain liability for patient suicide.

If patients are going to be seen infrequently by the prescribing physician, there must be a way for the physician to set up a communication pattern where side effects and suicidality are made known, to the extent possible, to the physician, on a regular basis. For example, if a patient were being seen by the physician every six months but the therapist is seeing the patient weekly, then the therapist and physician must discuss how issues and developments about the patient can be communicated.

In collaborative treatment, it must be determined how the treatment plan is shared between the clinicians; how emergencies including hospitalization will be managed; and who is directing the psychotherapy. The setting may play a large role in the extent to which the physician, therapist, and patient communicate. In many public settings, for example, the psychiatrist only sees the patient infrequently. The patient, however, has a case manager whose job it is to communicate to the various clinicians about the patient. In managed care settings, psychiatrists may find themselves in relationships with therapists whom they don't know and who are unaware of their skills, talents, experience, etc. There may be disincentives for the physician to see the patient often, even when more psychosocial problems accumulate for the patient. In collaborative treatment in certain settings, patients sometimes don't even know who the doctor is and often the physician has not had the chance or taken the opportunity to meet with the patient's family, since others (therapists, case managers, nurses, a previous physician) might have already done so.

There is, therefore, a range of potential legal problems for physicians who engage in collaborative treatment. Physicians are usually viewed as 'the deep pocket' when there is a lawsuit regarding a patient's treatment. The physician usually carries more insurance than the nonphysician therapist and is viewed by juries as 'being in charge' by virtue of education and experience. Even if the plaintiff's attorney does not believe that the psychiatrist was central to the matter, it is often likely that the psychiatrist will be part of the suit to extend the limits of insurance and to try to reach 'the deep pocket.' There is also the problem that occurs when there is more than one defendant. One defendant can try to make their case stronger by blaming the other defendant, thereby turning against one another. Another problem is that even if medications or the medical management provided by the physician is of minor importance to the reason for the suit, regardless of the realities, the physician is often viewed as being in charge while the therapist is viewed as subordinate to the physician.

Finally, physicians are increasingly signing contracts to provide services for patients in managed care. These contracts are usually written by the companies themselves and the fine print is not read carefully by the physician, causing them to be implicitly involved in relationships with therapists that are not collaborative in nature but more supervisory. This practice, therefore, often puts the physician at further risk.

There are some suggestions for clinicians to heed in order to decrease legal risk in collaborative treatment:

(a) Clinicians should understand the settings they are practicing in, including its operations and routines.

(b) Care should not be compromised because of the setting or because the patient is being treated in a collaborative model.

(c) Use good communication techniques with professionals with whom you are sharing responsibility for the patient. Document these discussions.

(d) If supervision is part of the collaborative relationship, then set regular supervision times and amend depending on the patient's condition and needs.

(e) Read the fine print in all managed care and other contracts that are signed.

For a more thorough discussion of these issues, please see the work of Macbeth (1999).

3.2 Ethical

When patients receive collaborative treatment because that is what is deemed best for their care, then there is not an ethical problem. An ethical and clinical dilemma however, may exist if such care is provided because of cost considerations (Lazarus 1999). While some patients clearly benefit from team approaches to care, others do not. Research is starting to help us discriminate between such patients and determine mental health outcomes in different practice and fee arrangements (Rogers et al.1993, Wells et al. 1992). Goldman et al. (1998), for example, has shown that for many patients, psychotherapy and psychopharmacology provided by a psychiatrist alone is as effective and less costly than collaborative treatment.

There are four bioethical principles that need to be considered when one is providing collaborative treatment (Beauchamp and Childress 1994):

(a) Beneficence: the covenant between professional and patient to do what is in the patient's best interest;

(b) Nonmaleficence: the principle of doing no harm;

(c) Autonomy: respect for the individual rights of patients, especially in the area of informed consent and treatment refusal; and

(d) Justice: the fair treatment of patients.

As noted by Lazarus, when determining ethical dilemmas related to collaborative treatment, any two or more of the above principles may be compromised or in conflict (Lazarus 1999). In this era of cost containment and lack of standards regarding collaborative treatment, all of these principles are being tested. The American Psychiatric Association (APA 1993) and the American Medical Association (AMA 1994) have addressed ethical principles and have recognized the importance of collaborative treatment. Section 5 of the AMA principles states: 'A physician shall continue to study, apply and advance scientific knowledge, make relevant information available to patients, colleagues, and the public, obtain consultation, and use the talents of other health professionals when indicated' (AMA 1994).

It is a complex area where psychiatrists and other mental health professionals engaged in collaborative treatment can find themselves in ethical dilemmas. Some of this has to do with the settings and the rules within those settings; physician–therapist relationships that should be modified because of the patient's condition but may not be; need for differential types of supervision and monitoring between therapist and physician but which are not taken into consideration; fee schedules that cause disincentives for providers to see patients more often, even when needed; lack of understanding by the patient of the nature of the collaborative relationship between the providers; etc.

In order for health professionals to provide psychotherapy and psychopharmacology in an ethical manner, Lazarus (1999) suggests:

(a) The physician and therapist keep themselves adequately and appropriately informed of the nature of the treatment and the progress of those patients cared for in this type of treatment.

(b) The physician and therapist assure themselves that the treatment is being carried out competently and adequately.

(c) The physician and therapist provide enough supervision, consultation, or collaboration to ensure that their ethical, medical, and legal responsibilities towards the patient are satisfied.

(d) The collaborative treatment is consistent with any local, state or national guidelines outlining psychiatrists' responsibilities to patients and in conformity with licensure requirements and the scope of practice of all the healthcare professionals involved.

4. Maximizing the Collaborative Relationship

Providing combined psychopharmacology and psychotherapy is a complex enterprise, much more difficult than a single therapist providing either the psychotherapy; psychotherapy and pharmacotherapy; or pharmacotherapy. There are, therefore, some general suggestions offered that might be useful.

It is incumbent upon the clinicians and patients involved in such treatment to be aware of the complicated nature of this form of care. While combining psychopharmacology and psychotherapy offers distinct and unique advantages, there are problems, as noted in the above section. Time and attention must, therefore, be given to the process of treatment as well as the content. Written contracts between clinicians have even been suggested as a way to delineate the specific duties and responsibilities of providers (Appelbaum 1991). Other authors have suggested that the patient be part of the three way contract and so be asked to sign such agreements for care (Chiles et al. 1984, Woodward et al. 1993). Specific guidelines for combined care based on the phases of treatment (initial, mid, termination) have been described by Rand (1999). Kahn (1990) has delineated the ideological differences that the psychotherapist and psychopharmacologist often have and the identified roles they must play in order to provide such treatment.

It is probably prudent for clinicians, both physicians and nonphysician therapists, to review their case mix from time to time. Goals would be to limit the numbers of such collaborative treatments based on the amount of time it takes to speak with other therapists; review and develop joint treatment plans; read other clinicians' notes; determine the background and treat-

ment orientation of the clinicians who are collaborating on patient care; work with patients' families and significant others; develop vacation and coverage schedules with collaborating clinicians; discuss such arrangements with patients; etc. Such thoughtful processes may lead to a decrease in malpractice exposure and ultimately more efficient and useful clinical care.

Attention must be paid carefully to the issue of confidentiality and communication between clinicians and with the patient. Consent to discuss treatment plans as well as content of sessions must be clarified. Since it is impossible to initially predict what will be important or relevant, the patient must be cognizant of these communication patterns and provide consent to clinicians at the beginning of treatment. Any barriers to such communication should be discussed with the patient as well as any relevant family members. It should be determined how the information will be transmitted, that is by telephone, electronic communication, mail, in person, etc. and how often. Communication must be regular, not just when there is a crisis (Vasile and Gutheil 1979). It is critical that the patient understands which type of problems should be discussed with which clinician. Further, patients should not be placed in the position of needing to relay information between clinician and the other.

Even with that said, it will sometimes be difficult to sort out. As an example, often what seems like uncomfortable medication side effects could be resistance to taking medications. What seems like a dynamic issue in therapy could in fact be a medical problem. Clinicians need to be supportive of the treatment plan; listen for resistances to either psychotherapy or pharmacotherapy; try to delineate the boundary of the treatment being provided by the individual clinician; and talk regularly with one another about the patient's progress. Because of financial remuneration patterns and modern contractual agreements between clinicians and payers, it is easier than ever for a clinician to find themselves responsible for large numbers of patients, many of whom are not being seen frequently or regularly. This can lead to problems with the inherent doctor–patient relationship that is the hallmark of providing care, whether it is pharmacotherapy, psychotherapy, or both. The clinician is well advised to review the ethical boundaries of the work situation and the nature of the tasks required.

Certain patient populations may be at higher risk in such combined therapies. For example, patients with personality disorders are notoriously difficult to treat given their problems with interpersonal relationships. As summarized by Silk (1999) and others, patients with Cluster B diagnoses are prone to splitting (Kernberg 1975); often do not tell the entire story equally to all clinicians (Main 1957); and tend to externalize problems (Silk et al. 1995). When there is concomitant substance abuse, emotional lability, and self-mutilating behaviors, combined treatment with more than one clinician can be enormously difficult and complex.

Finally, psychiatry residency training programs and graduate programs in social work and psychology must develop education and training in the specific area of combined treatment. Teaching psychopharmacology alone or psychotherapy without integrated pharmacotherapy is no longer sufficient. There needs to be curricula developed and teachers who are trained themselves in how to deliver such care with emphasis on learning to communicate clearly with other professionals; thinking clearly about diagnosis and treatment strategies; and developing leadership abilities (Spitz et al. 1999).

5. *Conclusion*

Providing combined pharmacotherapy and psychotherapy is a growing and important part of psychiatric care. It is a complex dynamic treatment, fraught with potential pitfalls but offering many advantages. Importantly, it offers a unique opportunity for a biopsychosocial and fiscal model of care and interesting areas for future research and clinical development.

Bibliography

American Medical Association Council on Ethical and Judicial Affairs 1994 *Code of Medical Ethics: Current Opinions with Annotations*. American Medical Association, Chicago, IL

American Psychiatric Association 1980 Guidelines for psychiatrists in consultative, supervisory, or collaborative relationships with nonmedical therapists. *American Journal of Psychiatry* **137**: 1489–91

American Psychiatric Association 1993 *The Principles of Medical Ethics with Annotations Especially Applicable to Psychiatry*. American Psychiatric Association, Washington, DC

Appelbaum P S 1991 General guidelines for psychiatrists who prescribe medication for patients treated by nonmedical psychotherapists. *Hospital and Community Psychiatry* **42**: 281–2

Balon R 1999 Positive aspects of collaborative treatment. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 1–31

Beauchamp T L, Childress J F 1994 *Principles of Biomedical Ethics*. Oxford University Press, New York

Beitman B D 1983 The demographics of American psychotherapists: A pilot study. *American Journal of Psychotherapy* **37**: 37–48

Berg M 1986 Toward a diagnostic alliance between psychiatrist and psychologist. *American Psychology* **41**: 52–9

Bradley S S 1990 Nonphysician psychotherapist-physician pharmacotherapist: A new model for concurrent treatment. *Psychiatric Clinics of North America* **13**: 307–22

Brill N Q 1977 Delineating the role of the psychiatrist on the psychiatric team. *Hospital and Community Psychiatry* **28**: 542–4

Chiles J A, Carlin A S, Beitman B D 1984 A physician, a nonmedical psychotherapist, and a patient: The pharmacotherapy-psychotherapy triangle. In: Beitman B D, Klerman G L (eds.) *Combining Psychotherapy and Drug Therapy in Clinical Practice*. SP Medical and Scientific Books, New York, pp. 89–101

Firman G J 1982 The psychiatrist-nonmedical psychotherapy team: Opportunities for a therapeutic synergy. *Journal of Operational Psychiatry* **13**: 32–6

Foulks E F, Pena J M 1995 Ethnicity and pharmacotherapy: a component in the treatment of cocaine addiction in African Americans. *Psychiatric Clinics of North America* **18**: 607–20

Fromm-Reichman F 1947 Problems of therapeutic management in a psychoanalytic hospital. *Psychoanalysis Quarterly* **16**: 325–56

Goldman W, McCulloch J, Cuffel B et al. 1998 Outpatient utilization patterns of integrated and split psychotherapy and pharmacotherapy for depression. *Psychiatric Services* **49**: 477–82

Goldsmith R J, Paris M, Riba M B 1999 Negative aspects of collaborative treatment. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 33–63

Horgan C M 1985 Specialty and general ambulatory mental health services: Comparison of utilization and expenditures. *Archives of General Psychiatry* **42**: 565–72

Kahn D A 1990 The dichotomy of drugs and psychotherapy. *Psychiatric Clinics of North America* **13**: 197–208

Karasu T B 1992 The worst of times, the best of times: Psychotherapy in the 1990s. *Journal of Psychotherapy Practical Research* **1**: 2–15

Kernberg O F 1975 *Borderline Conditions and Pathological Narcissism*. Jason Aronson, New York

Kingsbury S J 1987 Cognitive differences between clinical psychologists and psychiatrists. *American Psychology* **42**: 152–6

Koenigsberg H W 1994 The combination of psychotherapy and pharmacotherapy in the treatment of borderline patients. *Journal of Psychotherapy Practical Research* **3**: 93–107

Lazarus J 1999 Ethical issues in collaborative or divided treatment. Risk. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 159–77

MacBeth J E 1999 Divided treatment: Legal implications and risk. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 111–58

Main T F 1957 The ailment. *British Journal of Medical Psychology* **30**: 129–45

Manning D W, Markowitz J C, Frances A J 1992 A review of combined psychotherapy and pharmacotherapy in the treatment of depression. *Journal of Psychotherapy Practical Research* **1**: 103–16

Michels R 1995 Commentary: Medication use during psychoanalysis: A survey. *Journal of Clinical Psychiatry* **56**: 179

Nadelson C C, Zimmerman V 1993 Culture and psychiatric care of women. In: Gow A C (ed.) *Culture, Ethnicity and Mental Illness*. American Psychiatric Press, Washington, DC, pp. 501–15

Paykel E L 1995 Psychotherapy, medication combinations, and compliance. *Journal of Clinical Psychiatry* **56**(Suppl. 1): 29–30

Pilette W L 1988 The rise of trepany treatment relationships. *Psychotherapy* **25**: 420–3

Rand E H 1999 Guidelines to maximize the process of collaborative treatment. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 353–80

Regier D A, Goldberg I D, Taube C A 1978 The *de facto* US mental health services system: A public health perspective. *Archives of General Psychiatry* **35**: 685–93

Riba M B, Balon R (eds.) 1999 *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC

Rogers W H, Wells K B, Meredith L S et al. 1993 Outcomes for adult outpatients with depression under prepaid or fee-for-service financing. *Archives of General Psychiatry* **50**: 517–25

Schurman R A, Kramer P D, Mitchell J B 1985 The hidden mental health network: Treatment of mental illness by nonpsychiatrist physicians. *Archives of General Psychiatry* **42**: 89–94

Silk K R 1999 Collaborative treatment: Personality disorder patients. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 221–77

Silk K R, Lee S, Hill E M et al. 1995 Borderline personality-disorder symptoms and severity of sexual abuse. *American Journal of Psychiatry* **152**: 1059–64

Spitz D, Hansen-Grant S, Riba M B 1999 Residency training issues in collaborative treatment. In: Riba M B, Balon R (eds.) *Psychopharmacology and Psychotherapy: A Collaborative Approach*. American Psychiatric Press, Washington, DC, pp. 279–306

Vasile R G, Gutheil T G 1979 The psychiatrist as medical backup: Ambiguity in the delegation of clinical responsibility. *American Journal of Psychiatry* **136**: 1292–6

Wells K B, Burnam M A, Rogers W H et al. 1992 The course of depression in adult outpatients: Results from the medical outcomes study. *Archives of General Psychiatry* **49**: 788–94

Woodward B, Duckworth K S, Gutheil T G 1993 The pharmacotherapist-psychotherapist collaboration. In: Oldham J M, Riba M B, Tasman A (eds.) *American Psychiatric Press Review of Psychiatry*, Vol. 12. American Psychiatric Press, Washington, DC, pp. 631–39

Yamamoto J, Silva J A, Justice L R et al. 1993 Cross-cultural psychotherapy. In: Gow A C (ed.) *Culture, Ethnicity and Mental Illness*. American Psychiatric Press, Washington, DC, pp. 101–24

M. B. Riba and R. Balon

Psychotherapy and Religion

It has been suggested that the reason psychotherapy is consistently linked with medicine and a medical model is an artifact of Freud's profession as a neurologist. Had he come from another discipline, what is considered to be therapeutic might be altogether different. If, for example, he had been a member of the clergy then perhaps the relationship between psychotherapy and religion that has persisted in the twentieth century might be less adversarial. The politics, economics, and

social forces that have been behind the formation of the relationship between religion and psychotherapy are beyond the scope of this article. The intent here is to present the parallel paths of religion, particularly that with Judeo–Christian roots, and psychotherapy as it has developed in the industrialized West. We will consider at which points these paths have crossed, where they have diverged, and, ultimately, what effect this has had on the practice of psychotherapy as it is understood at the beginning of the twenty-first century.

1. Establishing Common Language

Prior to our exploration of the paths of religion and psychotherapy, a common language for health and pathology must be established in order to accurately read the road markers. Miller and Thoresen (Miller 1999) have proposed a construct of health that includes the domains of suffering, function, and coherence. As they point out, suffering is a state that can be understood as a 'form of unhealth and a common reason for seeking help, from either spiritual or psychological professionals.' Second, functional ability also is a term that is understood in either discipline. Third, the domain of coherence or inner peace is also defined in the language of both disciplines, although perhaps more subjective and less easily observed than the other two.

Due to a lack of specific focus on behavior, these overlapping domains often add more confusion than clarity, as in the case of certain extreme behaviors diagnosed as psychotic when considered from a therapeutic standpoint. Within this context, this cluster of behaviors suggests that an individual's normal affective and cognitive functions have gone awry. Common manifestations include audio and/or visual hallucinations, delusions that may appear grandiose or paranoid in nature, hypersensitivity, and an apparent disregard for those personal care habits consistently linked with health. However, these same behaviors in the context of religious tradition might be deemed a mystical experience or even the pinnacle of spiritual health. William James conceptualized the mystical state as merely another form of consciousness, albeit rare, that could be stimulated either by internal states or external substance. This mirrors Hunt's (1985) sentiment that mysticism is a direct manifestation of normal but usually inaccessible semantic operations. Therefore, it would appear then that specific target behaviors are unlikely to be the most fruitful ground for establishing common definitions for health and pathology. Perhaps a more useful approach would focus not on the specific behavior itself, but on the individual's experience of it, which is completely in line with the Miller and Thoresen (Miller 1999) domains.

James described four aspects of the truly mystical experience that appear to be rare in the experience of psychosis. The first is presentience. Mystical transcendence is more similar with an intuitive sense of knowing than it is with knowledge derived from either thought or sensation. Additionally, this transcendental experience affords the individual a new awareness of 'self' and 'other.' Armstrong (1989) adds that true mystics then demonstrate that this inner identity has come to fruition in outer world behavior. In other words, to be 'an integrated inner person is to be a productive outer person,' which is rarely the experience of the person in the midst of a psychotic episode. The second characteristic has been termed a noetic quality, or that of a being knowing about itself. This knowledge can be similar to a successful outcome in psychotherapy, but rarely of the psychotic experience. The third quality is that the mystical experience is usually transient, even in the most devout and saintly. It apparently is impossible to hold on to the presence of the Absolute for any length of time. The final characteristic of the mystical that was suggested by James was that of passivity. While an individual can dispose oneself to incorporate a deeper understanding and knowledge of the Divine, this wisdom cannot be called up like a specter. James noted that it was not uncommon for mystical union to be methodically cultivated as an element of religious life, similar to the actions of medieval mystics. But, he added, it is possible to stumble into the mystical, perhaps if one is inherently of the appropriate nature. What is also important in this regard, is that both Eastern and Western traditions have suggested that the mystical experience is not pure if the person is unprepared for or unwilling to lead a better life based on the knowledge imparted. This is a vital distinction given the common and profound sense of unpreparedness of persons experiencing severe psychopathology.

As has been previously suggested, when these four attributes of the mystical are compared to psychotic experiences, one may find some common observable traits. However, in studies comparing the experiences of people with schizophrenia and mystics, Buckley (1981) found that while both share a sense of noesis, heightened perceptual abilities, communication with the divine, and exultation, only in psychosis was there the experience of thought disruption, and therefore an inability to incorporate the knowledge in an effective way. It is also true that individuals with schizophrenia can go for extended periods of time without any relief from these experiences, where the mystic, as noted above, sustains this experience only briefly. And, in spiritual people, mystical experiences correlate significantly with psychological well-being, self-actualization, and optimism. Conversely, psychotic episodes tend to be marked with more negative emotional affect.

To summarize, what is aberrant and what is healthy, regardless of whether one is following a psychological

or religious model, is not so much the events themselves but the experience of the events. Although psychospiritual experiences can be present in many different contexts, it would seem that both psychotherapists and religious practitioners can agree on an optimally healthy state for any individual. While cultural variation will no doubt exist, this target state would allow the person to function, for the most part, within the affective and cognitive norm, supported by and supportive of a community that could foster these goals. This common definition of health will allow better exploration of the paths of religion and psychotherapeutic practice, regardless of psychological sophistication or religious beliefs.

2. *Historical Perspective*

In the late 1800s there was little divergence in the paths of religion and psychotherapy. The period was replete with schools of thought that outlined methods generally considered to be faith healing or mind cures. These approaches had strong support from some of the most influential writers of the time, such as Alcott, Emerson, and Thoreau. Overwhelmingly positive in their beliefs about an individual's potential for rekindling the divine spark believed by some to be inherent in all of us, one can well imagine their appeal when contrasted to the methods and practices common to the nascent psychiatric science of the time. These approaches held that illness, whether mental or physical, was the direct result of one being out of sync with some universal law. In contrast, the faith healing interventions, whether called down by external sources or fostered through self-help, merely required the realignment of one's beliefs and actions to be more attuned with the ultimate reality and truth. The writings of William James, the Mental Hygiene movement, and the resurgence of interest in the use of the group as an optimal setting for the healing of the individual all were well in line with religious beliefs of the day. However, this openly parallel course did not last.

By the early century, psychoanalytic and behavioral theories of the time seemed to suggest that humans were either unsuspecting victims of various drive states or unformed blocks to be molded by external cues or forces. Humans had lost their status as the pinnacle of creation. In retrospect, it is interesting to note that this portrait of human beings as rather hapless characters, lacking control over some of their most basic qualities, was not so different from some of Fundamentalism's tenets about the inherent evilness of humankind in the absence of God's grace.

Convergence began again toward the middle of the twentieth century. In response to the effects of the Second World War and the coming of the nuclear age, Western thinkers, both religious and psychotherapeutic, again began to re-evaluate the nature of humankind, its inherent qualities, and its potential for change. Important to both was a movement toward a more positive view of human nature with less of the pessimism that had marked some of psychotherapy's models in recent history. The existential shift that had affected so much of the Western world was incorporated into current thought. The nationalism that underlay much of the global conflict gave way to greater acceptance of 'that which is different' as being something other than 'that which is worse.' Prevalent again was the idea that humans were capable of virtuous acts, but this time the element of choice was worked strongly into the equation. And, while a number of religious leaders appeared to have understood that religion and psychotherapy were moving toward another confluence, many of today's psychotherapists question whether religion and psychotherapy can be reconciled. It is on this discrepancy that we now focus.

3. *Current Perspective*

The tradition of professional detachment regarding the spiritual beliefs of patients is reflected in current survey data indicating a profound disconnection between the religious beliefs and desires of the general population and mental health providers. For example, while 72 percent of the population view religion as an important influence on their lives, only 33 percent of psychologists respond similarly. In academic settings, psychologists are twice as likely as physicians to report that they have no religious preference (Shafranske 1996). Clinically, the issue of client religious beliefs typically is ignored or may be viewed as pathological. In contrast, survey data consistently reveal that religious beliefs are important to the general population, those presumably treated by mental health practitioners. Over 86 percent claim that they believe in God and 70 percent believe in a God who answers prayer (Barna 1992). Importantly, over half of responders state that they would prefer religious to secular mental healthcare regarding certain issues, especially marital and family problems (Privette et al. 1994).

There is reason to believe, however, that mainstream professional indifference to religious beliefs in psychotherapy has been overestimated. While mental health providers are less likely to report that religious beliefs are a major influence on their lives, 77 percent of them do claim to try to live their lives according to religious beliefs, which closely mirrors the general population (Bergin and Jensen 1990, p. 5). According to Miller and Thoreson, there is a growing recognition in the mental health field that religious beliefs in clients are correlated with longer life, and lower incidence of suicide, divorce, and alcohol abuse (Miller 1999). When the impact of religious beliefs is measured upon 'hard' variables, meaning those that are observable and easily

measured such as longevity, the correlation almost always indicates a protective effect of such beliefs. When the impact of religious variables is measured upon 'soft' variables, those that are not observable and easily measured such as rigidity, the results are mixed. Several researchers have noted that the anti-religious bias of mental health providers may lead to a lack of attention to the obvious benefits conferred by religious beliefs in important domains of human functioning, while overemphasizing the negative effects of religious beliefs in intrapsychic domains (Bergin et al. 1996).

Mental health providers clearly are moving toward a more egalitarian approach to the spiritual and religious beliefs of their clients, with momentum. For example, in 1992 the American Psychological Association included religious beliefs among the spectrum of diversity which is to be protected by psychologists, who are expressly forbidden to discriminate or tolerate discrimination on the basis of religious beliefs (Bergin et al. 1996). Additionally, a suggestion to add a new code to the *DSM* diagnostic formulation to indicate psychoreligious or psychospiritual problems (Lukoff et al. 1992) has been accepted. Perhaps more importantly, psychologists and other mental health providers are beginning to examine empirically the impact of religious beliefs upon the process of psychotherapy. Initial findings indicate that integrating religious beliefs into psychological treatments may offer superior results for anxiety (Propst et al. 1992) and depression (Razali et al. 1998) in controlled studies with random assignment, for clients with religious beliefs. The impact of incorporating a spiritual dimension to psychotherapy is perhaps nowhere more visible than in the treatment of addictions. The long tradition of incorporating spiritual renewal found in the 12-step approach has received empirical support in a growing number of correlational studies (Miller 1999) and rigorous research studies (Project MATCH 1997), regardless of client's pre-existing religious beliefs.

Another factor that may be influencing psychology's consideration of religious beliefs and practices is the large number of assessment instruments, with psychometric validation, currently available for the measurement of a wide variety of religious and spiritual constructs (see Hill et al. 1999). Many of these are relatively brief, but provide useful information on such topics as coherence, purpose in life, and spiritual well-being. The work of Pargament and Brant (1998) has addressed systematically an individual's coping strategies, their roots in either religious or spiritual beliefs, and where they are related to negative or positive health outcomes.

It appears, therefore, that the roads have once again begun to converge and that more practitioners of psychotherapy are recognizing the potential importance of assessing religious beliefs and incorporating them into practice. This is a marked change from the belief that client's spiritual beliefs are pathological, irrelevant, or damaging to the therapeutic process (Shafranske 1996).

4. *Future Directions*

Ultimately, however, if the relationship between religion and psychotherapy is to continue to grow, there are several areas that will continue to challenge us. First, it is no longer the case that the vast majority of persons in industrialized Western countries are descendent from Anglo–European lineage. By the middle of this new century it is estimated that Whites will be a racial minority in the United States. Subsequently, both psychology and religion will have to incorporate this divergence. This will mean that the concepts of health and spiritual well-being probably will come to include philosophies prominent in Asian and African cultures and religions. These cultures have often looked at the inherent value of a single human being, the role of suffering, and the connection with the environment in a manner not typical of current psychological and Judeo–Christian tenets. Second, the role of women has changed radically since the early days of both Judeo–Christian religions and psychotherapy. In this case, feminist voices from both disciplines are likely to continue to express conflict with the goals of therapeutic intervention and mechanisms of change. Last, as scientific understanding of the working of the human brain and nervous system continues to progress, it will be important for both psychotherapy and religion to remain flexible enough to incorporate this new knowledge.

See also: Clinical Psychology in Europe, History of; Counseling and Psychotherapy: Ethnic and Cultural Differences; Psychotherapy: Ethical Issues; Psychotherapy, History of: Psychiatric Aspects; Psychotherapy Integration; Religion and Health; Religion: Psychiatric Aspects; Religion, Psychology of; Religiosity: Modern; Religiosity, Sociology of

Bibliography

Armstrong R 1989 The psychology of inner and outer space. *Pastoral Psychology* **37**: 161–4

Barna G 1992 *What Americans Believe: An Annual Survey of Values and Religious Views in the United States*. Regal Books, Ventura, CA

Bergin A E, Jensen J P 1990 Religiosity of psychotherapists: A national survey. *Psychotherapy* **27**(1): 3–7

Bergin A E, Payne I R, Richards P S 1996 Values in psychotherapy. In: Shafranske E (ed.) *Religion and the Clinical Practice of Psychology*. American Psychological Association, Washington, DC

Buckley P 1981 Mystical experience and schizophrenia. *Schizophrenia Bulletin* **7**: 516–21

Hill P, Butman R, Hood R 1999 *Measures of Religiousness*. Religious Education Press, Birmingham, AL

Hunt H 1985 Relations between the phenomena of religious mysticism (altered states of consciousness) and the psychology of thought: A cognitive psychology of states of consciousness and the necessity of subjective states for cognitive theory. *Perceptual and Motor Skills* **61**: 911–61

Lukoff D, Lu F, Turner R 1992 Toward a more culturally sensitive *DSM-IV*. Psychoreligious and psychospiritual problems. *Journal of Nervous and Mental Disease* **180**(11): 673–82

Miller W R (ed.) 1999 *Integrating Spirituality into Treatment: Resources for Practitioners*. American Psychological Association, Washington, DC, pp. 3–18

Pargament K I, Brant C R 1998 Religion and coping. In: Koening H G (ed.) *Handbook of Religion and Mental Health*. Academic Press, San Diego, CA, pp. 112–28

Privette G, Quackenbos S, Bundrick C M 1994 Preferences for religious or non-religious counseling and psychotherapy. *Psychological Reports* **1**(2): 539–40

Project MATCH Research Group 1997 Project MATCH secondary *a priori* hypotheses. *Addiction* **92**(12): 1671–98

Propst R L, Ostrom R, Watkins P, Dean T, Mashburn D 1992 Comparative efficacy of religious and nonreligious cognitive–behavioral therapy for the treatment of clinical depression in religious individuals. *Journal of Clinical and Consulting Psychology* **60**: 94–103

Razali S M, Hasanah C I, Aminah K, Subramaniam M 1998 Religious–sociocultural psychotherapy in patients with anxiety and depression. *Australia and New Zealand Journal of Psychiatry* **32**(6): 867–72

Shafranske E P 1996 Religious beliefs, affiliations and practices of clinical psychologists. In: Shafranske E (ed.) *Religion and the Clinical Practice of Psychology*. American Psychological Association, Washington, DC, p. 52

V. A. Waldorf and T. B. Moyers

Psychotherapy, Brief Psychodynamic

Brief psychodynamic psychotherapy (BPP) describes a 'family' of psychotherapeutic interventions that share in common a short-term adaptation of psychoanalytically inspired psychotherapy. Other terms such as time-limited dynamic psychotherapy and short-term dynamic psychotherapy have also been used to describe this approach to treatment (Crits-Christoph and Barber 1991, Messer and Warren 1995).

The number of psychotherapies classified as brief psychodynamic has increased remarkably in recent years. As these numbers continue to grow, the clarity of features that constitute a psychotherapeutic system as brief dynamic becomes blurred. In an effort to provide a prototypical definition of BPP, we have gathered criteria from various authors, including those presented by Crits-Christoph et al. (1991), as follows.

(a) The theory of psychopathology explaining the cause and maintenance of a psychological disorder has a psychoanalytical basis (and thereby makes use of concepts related to such theories as Freudian, neo-Freudian, interpersonal, object relations, and self-psychology). Among the most important constructs one should include unconscious motivations, the formative role of early years on character formation, the tendency to repeat early behavior, the pervasiveness of transference, the importance of anxiety as a signal of underlying conflict, and the role of defenses.

(b) The main treatment techniques employed by the psychotherapist are those described in the psychoanalytic literature such as making clarifications, interpretations, and confrontations and noting transference and countertransference. Despite sharing these techniques, most authors do not consider BPP a brief or condensed form of analysis but rather a distinct treatment of its own. In general, the use of advice and homework, often practiced in behavioral and cognitive psychotherapies, is not recommended.

(c) The length of treatment is planned, often predetermined, and always time-limited. It usually lasts from 12 to 40 sessions depending on the patient's problems.

(d) Because of its brevity, patients are selected for BPP. Patients have to be screened to ensure that they meet at least some of the requirements for this form of treatment. In the literature, the following requirements have often been mentioned: psychological mindedness, recognition that present problems are in part of a psychological nature, willingness to change, existence in the present or past of at least one good quality interpersonal relationship, the ability to circumscribe present complaints to one symptom, the capacity to interact flexibly with the interviewer, and adequate ego strength (including the ability to tolerate frustration and ambivalence, adequate reality testing, and the use of neurotic but not psychotic defenses).

(e) In order to maintain brevity, a focus for treatment is implemented, and issues not related to the focal point or central conflict are often considered as having secondary priority. In general, the focus of treatment is developed collaboratively either before therapy starts or immediately following the initiation of therapy. The therapist actively works to maintain the treatment focus on the agreed upon core issue.

(f) BPP is intended to alleviate specific symptoms and provide limited personality or character change. This is in contrast to psychoanalysis and long-term dynamic therapy where the goal is typically pervasive character change.

Because most brief dynamic psychotherapies do not fulfill all of the aforementioned requirements, these criteria cannot be considered necessary. However, the criteria can and should be viewed as a prototypic definition whereby the more criteria a therapy meets, the more likely it will be designated as brief psychodynamic psychotherapy.

1. *Historical Background*

Modern psychotherapy can be traced to Sigmund Freud's work that began in the late nineteenth century.

During that time, Freud developed a way to investigate the mind that ultimately evolved into a psychological intervention known as psychoanalysis. In psychoanalysis, the analyst attempts to bring repressed unconscious conflicts to the patient's awareness. In other words, by providing patients an understanding ('insight') of the unconscious aspects of their problems, patients have the opportunity to work them through and subsequently master these difficulties. Although Freud originally worked with patients for a small number of sessions, psychoanalytic treatment became increasingly longer over the years. Gradually, psychodynamically oriented psychotherapy, also referred to as psychoanalytically inspired psychotherapy, emerged from the modified use of psychoanalytic principles and therapeutic techniques.

Similar to psychoanalysts, long-term psychodynamic psychotherapists tend to be quiet, waiting for patients to free associate and observing how the patient–therapist relationship (or transference) develops, rather than actively pursuing psychologically meaningful material and addressing transferential material early in the treatment. Despite this similarity, however, psychodynamically oriented psychotherapy differs from traditional psychoanalysis in significant ways. For example, in contrast to psychoanalysts who see patients on the couch four to five times a week, long-term dynamic psychotherapists generally see patients face to face once or twice a week. Nevertheless, consistent with psychoanalysis, psychoanalytically inspired psychotherapy has tended to be a long-term therapy.

In addition to the long-term approach to dynamic therapy, there has also been a small group of therapists who have attempted to keep psychoanalytically inspired psychotherapy brief and concise. These clinicians continued the work originated by Ferenczi (1950), Rank (1936), and Alexander and French (1946), who recommended the maintenance of an active stance in therapy to hasten the exploration of unconscious material. In spite of their efforts, however, most psychodynamic clinicians continued to view brief dynamic therapy as inferior to the lengthier psychoanalytic treatment. It was not until Malan (1976) Mann (1973), Sifneos (1979), and Davanloo (1980) that brief psychodynamic psychotherapy was deemed a valuable treatment option. Malan's emphasis on the importance of careful patient selection through the screening of inappropriate referrals and through trial interpretations attuned therapists to the fact that BPP can be effective for a subset of the patient population. Malan and Sifneos were also among the first to stress the significance of defining and maintaining a therapeutic focus and the relevance of such focus for being an appropriate candidate for BPP. Malan explicitly illustrated some of the cornerstones of psychodynamic treatment through the descriptions of two triangles. The first triangle is the 'triangle of conflict.' Its apexes correspond to the labels 'defense,' 'anxiety,' and 'underlying feeling or impulse.' The second triangle is the 'triangle of person' and has the labels of 'relationships with current figures,''relationship with the therapist (representing transference),' and 'relationships with important figures from the past' (e.g., parents) for its apexes. According to Malan, the therapist's task is to expose the underlying feelings and impulses that the patient has been protecting via defense mechanisms and to elucidate the role of the defenses in reducing the anxiety that the feelings create. He posited that the patient's hidden feelings were originally experienced in relation to the parental figures at some time in the past and since then have frequently recurred with other significant figures in the patient's life including the therapist. During therapy, the patient must understand the hidden impulses underlying each of the relationships described in the triangle of person. Typically, insight into a current relationship (either with a significant other or the therapist) is achieved first and is then related back to the parental figures. Today, the writings of Malan (1976), Mann (1973), Sifneos (1979), and Davanloo (1980) are considered the four 'traditional' approaches to brief dynamic therapy (Crits-Christoph et al. 1991). Their work was the initial driving force for the development of modern BPP and fueled the debate that brief therapy can be as successful as longer-term therapies.

In the last 20 years, a new generation of brief dynamic therapists appeared on the clinical scene. Among the most important are Luborsky (1984), Horowitz et al. (1984), Strupp and Binder (1984), and Weiss et al. (1986). This new group of brief dynamic therapists can be distinguished from previous generations by their interest in the empirical status of their treatment approaches. Perhaps resulting from their interest in research, this new generation has written intricately detailed descriptions of their clinical approaches that are very helpful for training and monitoring clinicians in the adequate use of their techniques. In fact, many of these books are considered treatment manuals and have been used in empirical research addressing the efficacy of these therapeutic methods.

2. *Efficacy of BPP*

During the 1980s, researchers began to explore the efficacy of BPP relative to that of other psychotherapies. Representing the trend toward methodological rigor that has prevailed throughout the 1980s and 1990s, manualized BPPs have slowly been incorporated into research studies to provide the standardization of treatment that is needed for the studies to be credible (see *Psychological Treatment, Effectiveness of*). Despite this methodological improvement, very few high quality studies have been conducted to examine the efficacy of BPP.

Anderson and Lambert (1995) published a comprehensive meta-analysis of 26 BPP efficacy studies that indicated that although BPP is more effective than minimal or no treatment, it is neither superior nor inferior to other forms of therapy. In spite of the fact that these data indicate that BPP is an effective treatment, it is not often recognized as such. One possible reason for the lack of recognition is the small number of studies that have examined the efficacy of BPP for specific disorders. For example, only one study has examined the efficacy of BPP for depression (Gallagher and Thompson 1982) and that study only included elderly depressed patients in its subject pool. Hence it is hardly surprising that many therapists do not find the data supporting the efficacy of BPP sufficiently convincing.

In addition to the interest in examining the efficacy of brief dynamic therapy, researchers and clinicians have consistently been curious about the processes of change underlying BPP. Strong findings in the literature on process research indicate that both nonspecific therapeutic factors, such as a positive therapeutic alliance, and specific therapeutic factors, such as accurate interpretations of the patient's core interpersonal conflicts, are associated with good therapeutic outcome.

3. *The Future of BPP*

During the last two decades, the popularity of BPP has grown. This shift can be attributed to the increased financial pressure from managed health care to shorten the average number of sessions in a treatment program, the patients' wish to receive help only for their specific complaints rather than for changing their personality, and the patients' wish to save money.

There is little reason to believe that the multiple causes influencing the trend in the late 1990s toward increased implementation of brief treatments including BPP will subside any time in the early part of the twenty-first century. Rather, it seems that efficiency and empirically demonstrated efficacy define future ideals. Consequently, it is likely that therapists will see an even larger increase in the use of brief therapy. In order to ensure that BPP will be one of the modalities recommended by the various governmental, insurance, and professional entities that are 'certifying' treatment modalities, there is an need for evidence of its efficacy for specific psychiatric disorders.

Without a doubt, ensuring that the quality of psychotherapy remains high is a major objective of all psychotherapists; however, other factors must also be considered with regard to the future status of brief dynamic psychotherapy. First, a large proportion of the training of psychiatrists and psychologists remains in the realm of long-term psychotherapy, although the trend toward training psychotherapists for short-term interventions might be occurring. Nevertheless, increased training of clinicians in the domain of brief therapy is required for patients to receive maximum benefit from this form of treatment. Also, psychotherapists should strive to implement the BPPs designed for specific diagnostic patient populations in their treatment of these patients. This task may prove to be especially difficult since the descriptions of symptoms and syndromes of particular disorders according to psychodynamic theory do not always correspond to those listed in the most recent version of the American Psychiatric Association's *Diagnostic and Statistical Manual* (*DSM*) *of Mental Disorders*.

Modern psychotherapy, including BPP, is becoming increasingly more integrative and eclectic in its treatment approach (see *Eclectic Psychotherapy: Psychological Perspectives*). As a result of this greater eclecticism, examining the efficacy of the actual treatments used in clinical settings and defining a given treatment delivered by a specific clinician as BPP are becoming perpetually more difficult tasks. Psychotherapy, including BPP, must constantly evolve to fit with the fast paced nature and current trends of the modern world. Thus, maybe future practice of BPP will include more components from other psychotherapies.

See also: Eclectic Psychotherapy: Psychological Perspectives; Psychoanalysis: Current Status; Psychoanalysis in Clinical Psychology; Psychoanalysis: Overview; Psychological Treatment, Effectiveness of; Transference in Psychoanalysis

Bibliography

Alexander F, French T M 1946 *Psychoanalytic Therapy*. Ronald Press, New York

Anderson E M, Lambert M J 1995 Short-term dynamically oriented psychotherapies: a review and meta-analysis. *Clinical Psychology Review* **15**: 503–14

Crits-Christoph P, Barber J P (eds.) 1991 *Handbook of Short-term Dynamic Psychotherapies*. Basic Books, New York

Crits-Christoph P, Barber J P, Kurcias J S 1991 Historical background of short-term dynamic therapy. In: Crits-Christoph P, Barber J P (eds.) *Handbook of Short-term Dynamic Psychotherapy*. Basic Books, New York pp. 1–16

Davanloo H (ed.) 1980 *Short-term Dynamic Psychotherapy*. Jason Aronson, New York

Ferenczi S 1950 The further development of an active therapy in psychoanalysis. In: Ferenczi S, Suttie J I (trans.) *Further Contributions to the Theory and Technique of Psychoanalysis* 2nd edn. Hogarth Press, London, pp. 198–217 (original work published 1926)

Gallagher D E, Thompson L W 1982 Treatment of major depressive disorder in older adult outpatients with brief psychotherapies. *Psychotherapy: Theory, Research and Practice*. **19**: 482–90

Horowitz M, Marmar C, Krupnick J, Kaltreider N, Wallerstein R, Wilner N 1984 *Personality Styles and Brief Psychotherapy*. Basic Books, New York

Luborsky L 1984 *Principles of Psychoanalytic Psychotherapy: a Manual for Supportive–Expressive (SE) Treatment*. Basic Books, New York
Malan D H 1976 *The Frontier of Brief Psychotherapy*. Plenum, New York
Mann J 1973 *Time-limited Psychotherapy*. Harvard University Press, Cambridge, MA
Messer S B, Warren C S 1995 *Models of Brief Psychodynamic Therapy: a Comparative Approach*. Guilford Press, New York
Rank O 1936 *Will Therapy*. Knopf, New York
Sifneos P 1979 *Short-term Dynamic Psychotherapy: Evaluation and Technique*. Plenum Medical Book Co., New York
Strupp H S, Binder J L 1984 *Psychotherapy in a New Key: a Guide to Time-limited Dynamic Psychotherapy*. Basic Books, New York
Weiss J, Sampson H, Mount Zion Psychotherapy Research Group 1986 *The Psychoanalytic Process: Theory, Clinical Observation and Empirical Research*. Guilford Press, New York

J. P. Barber and R. Stratt

Psychotherapy: Case Study

'Case study' typically refers to the investigation of the onset and course of a disorder on the basis of personal reports. Frequently used synonyms are case study methodology, case study research, case report, and case description. The acronym '$N = 1$' also pertains to case studies. Case study terminology is applied in an array of very different methodological approaches, ranging from nonstructured explorations to experimental testing of single-case hypotheses.

Description, documentation, and evaluation of a particular therapy are elements of 'controlled practice' (Petermann and Müller 2001), which in turn is based on patient-centered assessments made throughout the duration of therapy and on a well-organized documentation system. This approach allows for continuing control and optimization of the treatment process.

1. *Example of a Case Study Within 'Controlled Practice'*

At the outset of therapy the psychotherapist questions the patient in such areas as biography, current life situation, known problems, former experiences with psychotherapy, and the demand for therapy. The therapist documents answers by recording audiotapes and taking notes. A structured clinical interview for the purpose of classifying the disorder forms the foundation of the diagnosis. The client's current problems and therapeutic aims are organized in what is called a functional analysis. The therapist devises a written plan for the therapy detailing, for example, time frame and partial goals. The patient and therapist then sign a contract outlining rules and guidelines to remain in effect throughout the therapy. Prior to the actual onset of therapy, psychotherapist and patient negotiate the benchmarks and indicators for a successful outcome, for instance, how frequently undesirable behaviors may occur on a daily basis. Throughout the course of therapy the patient is asked regularly to fill in a brief questionnaire on major outcome variables. Direct observation by the therapist accompanies the written information. Finally, all collected documents, including the final interview, are summarized, discussed, and reviewed in cooperation with the patient. The psychotherapist's appraisal of the intervention and its outcome is based on comparisons of several case studies, thus allowing for continuous improvement in clinical decision making.

2. *Single-case: Different Methodological Approaches*

Yin (1994) distinguishes between descriptive, exploratory, and explanatory case studies. Typically, in case studies questions concerning the 'why' and 'how' dominate over questions that address the 'what,' 'who,' and 'where.'

2.1 *Descriptive Case Studies*

Descriptive case studies contain descriptions of individual cases but dispense with interpretation. Indications for psychotherapy, the course of therapy, and its outcomes are summarized in descriptive case studies in a verbal format—case reports—which have important documentation and communication functions for therapists. The lucidity of descriptive case studies, in contrast to abstract theory, makes them particularly useful in psychotherapy education. A variant of the descriptive case study yields a detailed description of the therapeutic process. Documentation of crucial outcome parameters is a necessary condition for process monitoring and has practical implications for optimization of therapy. If documentation is kept in the foreground of therapy reporting, treatment integrity can be better controlled. 'Treatment integrity' refers to the degree to which actual therapeutic work is concordant with the therapeutic method chosen (e.g., psychoanalysis, behavioral therapy). The more precisely the method of treatment can be defined, for example as a manual (see *Clinical Psychology: Manual-based Treatment*), the more exactly the degree of concordance can be assessed.

2.2 Explanatory Case Studies

Explanatory case studies attempt to explain observable therapy-induced changes in behavior. Critical life events in the patient's biography and experimentally varied factors may serve as explanations. In general, planned (prospective) case studies must be distinguished from unplanned (mostly retrospective) case studies. Planned case studies require a clear design, standardized and replicable data collection, and statistical methods of analysis. These studies are time-consuming and labor-intensive and consequently are not used very often. This type of process research addresses the decisive procedural mechanisms in effect during psychotherapy (see *Psychotherapy Process Research*). Experimental designs are strictly controlled types of prospective analysis allowing for comparisons between suggested causes in a systematic manner (see *Single-case Experimental Designs in Clinical Settings*).

2.3 Exploratory Case Studies

Exploratory case studies are used to receive one's first impression and to generate hypotheses. Data are collected in interviews or from biographical material. Hypotheses are derived from data of the individual by developing a model of the causal and sustaining factors pertinent to a particular individual's behavior ('func-tional behavior analysis,' see *Behavior Analysis, Applied*). Functional analyzes may be classified as exploratory or explanatory case studies. They contain descriptions of the patient's problem(s) and statements concerning factors that trigger and maintain problem behavior(s).

3. History of the Single-case Methodology

In the early days of modern psychology, when Edward Bradford Titchener, William James, and Wilhelm Wundt rose to prominence, single-case analysis was a well-respected research strategy. Not uncommonly the researchers chose themselves as the subjects for investigation. Introspection and single-case experiments largely dominated psychological research in those pioneering times. The German philosopher Karl Jaspers introduced a special form of case study, called *Biografik* (biographic), which focused primarily on the course of individuals' lives. With the introduction of what were considered more objective data sources and analytical methods, introspection and single-case experiments—and, as a consequence, the single-case approach—were sidelined and eventually replaced. The reasons were many (Danzinger 1990). Psychology as a new field of research had to gain ground and hold out against well-established traditional disciplines such as medicine, physics, and biology. The decision to adopt an experimental methodology was made in the hope of enhancing the reputation of psychological research. The questionnaire as an economic data collection tool was also making inroads into psychology. This methodology, adopted from the social sciences, allowed for the collection and analysis of large, standardized samples. Experiments could now be carried out involving substantial numbers of people.

The case study underwent a renaissance in the 1960s and 1970s. Awareness had arisen in psychotherapy research that certain questions could no longer be answered properly using group-based studies. Zubin (1950) had blamed the dire situation in clinical research on the lack of appropriate statistical tools for single cases, and called for the individual to be viewed as a distinctive universe. Kiesler (1966) complained about the absence of adequate research strategies in clinical psychology. He also challenged certain myths, including the 'patient uniformity assumption' (i.e., at the outset of therapy patients tend to be similar rather than different from one another) and the 'therapy uniformity assumption' (i.e., therapists treat individuals uniformly rather than variably).

New methods for the collection and analysis of data were developed in the 1960s. Goal-attainment scaling (Kiresuk and Sherman 1968), for example, opened up the possibility of collecting data that reflected individually defined outcomes. Innovations in statistics (e.g., Markoff chains and ARIMA models) now allowed for the analysis of data sampled from repeated measurements. Once again the case study was on its way to reclaim its reputation as a scientifically sound and respected methodology.

4. Single-case and Psychotherapy Research

Eysenck's (1952) provocative conclusion linking positive outcomes in the treatment of neurosis to spontaneous remissions sparked fervent clinical research seeking to prove the effectiveness of different treatments. The 'Dodo Verdict' (so called from the Dodo in *Alice in Wonderland*), according to which all therapies are equally successful, concluded the first phase of psychotherapy research. This view came under scrutiny during the second phase, when more differentiated analyzes were employed and differences in the indications for psychotherapy could be found for certain disorders. At the end of the second phase, knowledge of the actual mechanisms at work during different psychological therapies was still poor, a fact that eventually led to the third phase of clinical research, known as microanalysis. In microanalysis the interaction between therapist and patient became the primary focus. In an analogy to the use of a microscope, crucial moments during the therapy sessions were scanned with 'high methodological resolution' in order to uncover the crucial therapeutic mechanisms at work.

Gradually clinical practice research moved from a primary focus on outcomes to the model of process research. With each increase in the 'microscopic' resolution of details, single-case methodology gained in significance. Yet despite the fact that the theoretical rationale behind single-case study methodology was widely accepted in the 1980s, its impact on clinical practice remained limited. Not surprisingly, experts concluded: 'At present, clinical research has little or no influence on clinical practice' (Barlow 1981, p. 147).

5. *Criticism of Single-case Methodology*

Single-case methodology has always drawn criticism from different quarters. Some investigators felt that the presentation of qualitative case studies was too arbitrary; others have remained skeptical of the methodology in its entirety. Several problem areas with regard to case study research must be addressed:

(a) objectivity in the collection of data,

(b) degree of rigor in attempts to prove cause and effect, and

(c) interpretation and generalizability of outcomes.

Case studies in clinical research usually deal with personal experiences (e.g., quality of life, well-being), which can be evaluated solely by means of self-reports. However, assessment of self-reported experience is problematic since any subjective report automatically compromises an assessment's validity. On the other hand, standardized measures are of limited use when tapping into a patient's subjective appraisals. This problem of control in clinical practice remains unresolved.

The validity of causality statements depends primarily on the degree to which contextual factors in data collection remain constant. Causal relationships are determined by testing hypotheses while excluding factors that may serve as alternative explanations. It is a classical problem of empirical research, however, that alternative influences can never be ruled out entirely in data analysis. While controlling for several potentially confounding variables increases internal validity on the one hand, it limits the generalizability of findings on the other. This problem is reflected in research terminology: the term 'proof' is avoided, while 'evidence' or 'tendency' are preferred, thus allowing for case-specific statements about a treatment's effectiveness while avoiding statements of more general applicability.

Given these factors, the following prerequisites must be met in order to appraise the quality of a case study:

(a) Only validated or previously tested assessment tools should be employed,

(b) The decision to use prospective vs. retrospective data collection must be explained, and

(c) It must be decided to what extent contextual factors need to be controlled.

The accuracy of statements will increase with the precision and replicability of the data analysis as well as with the amount of detail provided on factors that may compromise findings.

6. *Integrating Case Study and Evidence-based Clinical Practice*

Legal and economic considerations increasingly require therapists to provide documentation of their interventions. Scientific evaluation is increasingly less limited to demonstrating the *efficacy* of certain therapeutic strategies and now must also show the *effectiveness* of the process of clinical practice as a whole.

The optimization of therapy serves both economic and humanitarian goals, as it is of utilitarian value for both patients and society in its entirety. In future the case study approach will form an integral part of what is known as 'evidence-based practice.' Particular difficulties include the problem of objectivity with regard to data collection and analysis, pressure on therapists to organize their interventions, and the integration of results with daily clinical practice. These demands on clinical practice control (see *Managed Care and Psychotherapy in the United States*) can only be met if data are analyzed in an exact and economical fashion. Recent software innovations may accelerate this trend.

See also: Single-case Experimental Designs in Clinical Settings

Bibliography

Barlow D H 1981 On the relation of clinical research to clinical practice. Current issues, new directions. *Journal of Consulting and Clinical Psychology* **43**: 147–55

Danzinger K 1990 *Constructing the Subject. Historical Origins of Psychological Research.* Cambridge University Press, New York

Eysenck H J 1952 The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology* **16**: 319–24

Kiesler D J 1966 Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin* **65**: 110–36

Kiresuk T J, Sherman R E 1968 Goal attainment scaling. A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal* **4**: 443–53

Petermann F, Müller J M 2001 *Clinical Psychology and Single Case Evidence. A Practical Approach to Treatment Planning and Evaluation.* Wiley, Chichester, UK

Yin R K 1994 *Case Study Research. Design and Methods.* Sage Publications, Thousand Oaks, California

Zubin J 1950 Symposium on statistics for the clinician. *Journal of Clinical Psychology* **6**: 1–6

F. Petermann and J. M. Müller

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Psychotherapy: Clinical Practice Guidelines

Since 1994, the process of delivering health and behavioral healthcare has been dramatically reshaped by the 'managed care revolution' (Strosahl 1994, 1995, see also *Managed Care and Psychotherapy in the United States*; *Managed Care*). Concerned by steadily escalating health care costs in the absence of convincing cost effectiveness data, both federal policy makers and the purchasers of health care have developed a variety of methods for containing health care costs. Most of these strategies (i.e., utilization review, precertification, case rates, capitation based financing) have focused on shifting the financial risk to vendors and providers, with the effect of restricting the services that are offered to consumers (Strosahl 1994). At the same time, health care reform has evolved from an overemphasis on supply side strategies (i.e., restricting access to services) to a cost and quality-oriented movement. This has stimulated a wide variety of system-level strategies for improving the quality of care while either reducing costs or holding them steady. Clinical practice guidelines evolved as one such strategy for quality improvement (Noonan et al. 1998). This article will examine the genesis of the practice guideline movement and different types of guidelines, discuss current parameters for evaluating the quality of guidelines, illustrate some of the major shortcomings of contemporary guidelines, and evaluate the empirical literature concerning the dissemination and impact of clinical guidelines.

1. *The History and Context of Practice Guidelines*

The most commonly accepted definition of a practice guideline is 'systematically developed statements to assist practitioners and patients make decisions about appropriate health care for specific clinical circumstances,' (Noonan et al. 1998). In this sense, practice guidelines have been developed and implemented informally at the agency and clinic level for years. Names such as treatment algorithms, clinical protocols, and practice standards have been used to describe such efforts. In contrast to contemporary practice guidelines, these early efforts were highly localized, often relied on clinical lore or the influence of a few key opinion leaders, did not systematically incorporate scientific evidence about the efficacy of clinical interventions, and addressed only the most controversial/problematic aspects of treating a patient with a particular condition. For example, many agencies and clinics have guidelines for assessing an acutely suicidal patient during the index crisis, primarily for the purpose of managing risk. At the same time, few guidelines have ever been developed for treating suicidal behavior when it is not of crisis proportions, even though the bulk of outpatient treatment occurs with less acutely suicidal patients. Due to their parochial nature, early practice guidelines—some developed only miles apart and for the same clinical condition—differed markedly in their emphasis and clinical recommendations.

The resurgence of interest in practice guideline stems from the growing influence of outcomes-based medicine and the associated philosophies of population and evidence-based care (Noonan et al. 1998, Strosahl 1994). Based on a public health model, population-based care is concerned with the health status of the entire population, not just those patients with current psychological disorders. The goal is to build consistent processes of care that reduce variation in clinical outcomes for the population served, thereby increasing the cost effectiveness of the overall system of care. A major area of concern with both medicine and behavioral health services is that systems of care vary wildly in the types of services available and the services rendered to treat a particular condition. The goal of population care is to narrow this variability across systems and providers, ensuring better access to, and quality of, care for consumers.

The related notion of evidence-based care is noteworthy because it suggests that the best outcomes across a population are achieved by the consistent use of interventions that have been shown to work in randomized clinical trials (see *Psychological Treatments: Randomized Controlled Clinical Trials*). Many clinical experts and the purchasers of health services argue that there is excessive variation among behavioral health providers in the types and amounts of treatment that are delivered for common psychological problems. In discussing the decision-making behavior of physicians, it has been suggested that only 25 percent of health services are objectively based upon the medical status of the patient. Primary nonmedical factors related to health services include availability of needed services, patient's beliefs and preferences and the physician–patient relationship. Most providers are ill-equipped to meet the diverse array of psychosocial and medical needs posed by their patients. This 'mismatch' between what the provider can offer and what the patient expects is a major source of cost overruns in both health and behavioral health.

2. *Types of Guidelines*

Broadly conceived, there are two types of practice guidelines. The system guideline is widely used (Noonan et al. 1998). It addresses the parameters of population-based care by specifying who is eligible to enter the system of care (i.e., medically necessary treatment), what setting of care is appropriate (i.e., in-patient, residential, partial hospital, out-patient) and

the program to which a patient with a particular set of symptoms and functioning should be assigned (i.e., vocational rehabilitation, drugs and alcohol, dual diagnosis). Note that a system guideline functions to define *who is eligible* to receive services and *what level of care* is appropriate to address the patient's condition. It does not attempt to guide the decision making or practice patterns of providers within that level of care. However, system guidelines can and should be evidence based, evolving from the scientific literature concerning the type and intensity of treatment mostly likely to work for a patient with a given set of symptoms and specified functional status. System guidelines are currently the focus of a major developmental initiative by the Substance Abuse and Mental Health Services Administration of the United States Government. System guidelines can promote the goals of population-based care by specifying and 'minimum services menu' that is needed to properly address the diverse mental health needs of a population. System guidelines help identify gaps in programs or facilities that can then be brought to the attention of policy makers, funding agencies, and foundations. One of the chief criticisms of clinical practice guidelines by providers is that they recommend settings of care or types of programs that are not accessible. While system guidelines often receive less discussion, they are a critical component of a comprehensive population care system.

Clinical practice guidelines are a second and more widely discussed entity. Ideally, they weave scientific evidence into clinical practice (Clinton et al. 1994). There has been an unprecedented proliferation of clinical practice guidelines in the beginning of the twenty-first century. Some are discipline specific such as the series of practice guidelines published by the American Psychiatric Association (Herz et al. 1997). Others, such as the Agency for Health Care Policy Research Guideline for Treating Depression in Primary Care, focus on treatment recommendations for a particular care setting, such as primary care (Depression Guideline Panel 1993). Some guidelines, such as those published by the Academy of Adolescent and Child Psychiatry are cross-disciplinary and designed to be applied across settings (Bernstein et al. 1997). Managed behavioral health care companies have also developed proprietary guidelines as a mechanism for supporting utilization reviewers. Finally, there are a growing number of 'off the shelf' guideline packages that have been developed privately and marketed to the provider community. The promulgation of clinical practice guidelines is promising in that it testifies to the growing conviction among policy makers, administrators, and clinicians that using scientific evidence to guide clinical practice is the only way to assure consistent quality. However, there is growing concern that many existing practice guidelines are biased toward truncating care, promoting guild interests, or simply lacking scientific validity (Nathan 1998).

3. *Functions of a Clinical Practice Guideline*

Properly constructed, a practice guideline should provide easily accessed, scientifically and clinically sound decision support that can aid the provider and the consumer at key decision points in the assessment and treatment process (Practice Guideline Coalition 1997), including:

(a) critical symptoms and/or functional attributes that make a patient eligible to be served under that guideline (i.e., diagnostic guidelines) and severity criteria that determine suitability for service setting (i.e., levels of care guidelines);

(b) conditions and/or syndromes that have similar symptom patterns and methods for distinguishing these conditions (differential diagnostic guidelines);

(c) relevant epidemiological and demographic information that helps providers identify different forms of clinical presentation related to geographic, racial, ethnic, or cultural factors;

(d) clinical procedures shown to be effective with the condition, rates of remission when such procedures are properly administered, comparative effectiveness of different treatments, treatment side effects, and other known risks as well as the expected time frame for clinical/functional response;

(e) commonly available and easily used self-report and/or interview methods for assessing clinical response, criteria for determining whether a procedure(s) is working and, if not, guidance in selecting an alternative procedure or treatment; and

(f) relapse rates and maximum relapse risk periods, effective maintenance, and relapse prevention strategies.

Note that the concept of assisting the provider and consumer in selecting an appropriate intervention for a specific condition suggests that guidelines need to exist in a format that can support use by both the provider and consumer. In addition to the core clinical function of improving the quality and appropriateness of care to the individual consumer, the Institute of Medicine (Field and Lohr 1990) has identified other guideline functions, including:

(a) decrease service costs produced by the provision of ineffective and/or unnecessary services,

(b) assist consumers and their families in advocacy efforts,

(c) assist payers in establishing rates and mechanisms for reimbursement,

(d) encourage the collection of standardized data for the purpose of 'validating' a guideline, and

(e) improve formal and post-graduate clinical training.

4. *Methods for Constructing Practice Guidelines*

As mentioned previously, early methods for constructing practice guidelines used either an informal or formal consensus method. Whereas informal con-

sensus might involve a group of providers agreeing to a care protocol at a lunchtime meeting, formal consensus methods require a structured development and ratification process (Woolf 1992). In the Delphi method, guideline development representatives determine rules about collecting and analyzing opinions from experts and distilling results into a set of recommendations. In iterative fashion, experts provide their opinions, then respond to the opinions of other experts until a convergence of opinion is obtained. In the Nominal Process method, a facilitator systematically elicits opinions from each group participant about the most important recommendations on each participant's list. This process continues until all recommendations from all lists have been cataloged, at which point the recommendations are distilled into a composite list. Each participant then rates this list in terms of priority, areas of maximum agreement are identified, and the resulting 'consensus' is put back to the group for ratification. The primary advantage of consensus-based methods is that guidelines can be rapidly generated, often ahead of the existing science. Innovations in clinical practice that have not been scientifically tested are more likely to be incorporated. The major, and often fatal, drawbacks are that the quality of consensus guidelines are entirely tied to the professional skills, qualifications, and biases of the group, tend to omit controversial topics where clinicians need the most guidance, produce bland recommendations that provide no advancement in clinical practice, or institutionalize locally favored practices that in fact are scientifically unsound. One of the best-known consensus-based guideline projects is the Expert Consensus Consortium, which has used the Delphi method to develop a wide variety of practice guidelines that are available nationally.

The limitations in consensus models for building guidelines have created a corresponding interest in using more objective, data-based methods. In contrast to consensus models, empirically-based methods stress that the quality of a guideline is directly related to its scientific underpinnings. For many years, the US Preventive Services Task Force has used the evidence-based guideline method. This approach involves conducting a comprehensive review of all scientific studies and reports pertaining to the condition under consideration, rating the amount and quality of the evidence, and then producing graded recommendations based upon the strength of the underlying scientific evidence. A 'Grade A' recommendation is based upon a number of well-conducted studies that meet specific and rigorous methodological criteria (i.e., sample size, randomization, well-specified, replicable interventions). In contrast, a 'Grade C' recommendation would have only limited evidence to support it, but still be seen as clinically important. Strict adherence to evidence-based methods tends to produce many fewer recommendations that are based on 'expert opinion.' However, providers and consumers can be left with no guidance in very core areas because the science is not available, and this may limit the clinical usefulness of the resulting product. Another frequent criticism of evidence-based methods is that cutting-edge innovations in clinical practice are left out of guidelines because scientific validation of effective procedures may lag years behind the introduction of new procedures in a field.

Eddy (1996) proposed an Explicit Guideline Development method that involves conducting an analysis of condition base rates within a system of care, along with costs and outcomes associated with existing care processes. Using cost modeling data derived from a review of scientific data regarding alternative procedures, a balance sheet is created which outlines the projected clinical benefits, costs, and potential risks associated with implementing a guideline. Generally, this approach is designed to identify conditions that occur with a high enough frequency and/or associated cost to justify changes in a whole system of care. In essence, this method helps one determine whether a particular condition is appropriate for a clinical practice guideline. Normally, systems and agencies that employ the initial decision-making method will employ an evidence-based method for constructing the actual guideline.

5. *Criticisms of Practice Guidelines*

Considering that guidelines vary greatly in both structure and content, it is interesting that a consistent set of criticisms have been forthcoming from key stakeholders (Noonan et al. 1998, Practice Guideline Coalition 1997).

Consumers have expressed concerns that guidelines are used to artificially limit care, are developed with little or no consumer involvement, ignore, and/or minimize consumer preferences and limit consumer access to information critical to the treatment decision-making process.

Clinicians complain that guidelines are rigidly applied to limit services and reimbursement, are too complicated to be useful in daily practice, recommend procedures or services that the provider is not trained to administer, neglect the important role of clinical judgment, function to suppress innovative clinical practices, and promote guild or discipline specific interests.

Scientists claim many guidelines incorporate unsubstantiated expert opinion, improperly characterize or ignore the scientific literature, recommend cost-oriented interventions that produce short-term benefits only, and are not kept current with recent advances in science.

Managed health care executives complain that guidelines cost too much and take too long to develop, exceed the skill and training level of most providers,

are too guild oriented, are not 'user friendly' in daily practice, and ignore the issue of cost effectiveness.

Concerns about the varying quality of practice guidelines have led to several efforts to identify criteria for judging the quality of a guideline. The American Psychological Association (1995) has developed standards for evaluating the validity of guidelines. Generally, these standards require that guidelines incorporate valid and reproducible scientific data, be written clearly, recommend procedures that have proven clinical effectiveness in the field, allow for flexible decision making, provide clear and precise recommendations, and be 'refreshed' regularly to keep up with changes in the scientific literature. In addition, the guideline development process should be multidisciplinary and well documented.

6. *Impact of Practice Guidelines*

At this point in time, there are few scientific studies concerning the impact of clinical practice guidelines for behavioral health conditions. Most reports are anecdotal and suggest that properly implemented, guidelines can alter the process of delivering care particularly if provider involvement is sought. Much less is known, even at the case report level, about the extent to which practice guidelines improve clinical outcomes.

Studies of practice guidelines for physicians have shown that dissemination of a guideline is not an end in itself. Training in use of the guideline is often required, as is the 'translation' of the guideline to fit local resource and practice realities. In addition, physicians may benefit from concurrent feedback and enforcement mechanisms. Even under the best of implementation circumstances, physicians may show enhanced awareness of the guideline, yet fail to modify their medical practices (Lomas et al. 1989). In essence, medical practice guidelines have not been consistently shown to alter the process of care, improve health outcomes, or improve consumer satisfaction (Grimshaw and Russell 1993). This may be due to variations in the quality of practice guidelines, differences in dissemination and training methods, or reflect the immediate financial incentives influencing physician behavior.

7. *Future Developments*

It is clear that the next generation of practice guidelines will need to be less cumbersome to use, less expensive and time consuming to develop, and reflect a consensus of all the key stakeholders as to what constitutes scientifically sound, feasible, cost-effective treatment. The Practice Guideline Coalition (PGC) represents one such effort to bring the key stakeholders together to form scientifically sound, user-friendly practice guidelines. The PGC is a coalition of approximately 80 scientific, clinical, managed care, and consumer groups, who have developed a common framework for the development, evaluation, and dissemination of practice guidelines. PGC guidelines have a very short development cycle, incorporate the views of scientists, clinicians, managed care representatives, and consumers, focus on core procedures that can be delivered by any qualified mental health provider and are designed for easy access in the context of daily practice (PGC 1997).

A second future theme will be the development of empirically validated methods for disseminating guidelines to the practitioner community. This very likely will involve the creation of training methods that are affordable and incorporate the results of early studies of guideline dissemination. The existing literature clearly suggests that it is one thing to put a guideline in front of a practitioner. It is quite another to have it impact the process or outcome of care. Dissemination and training represents perhaps the greatest challenge to the success of practice guidelines as a quality improvement tool.

Finally, there will be increasing pressure to scientifically validate practice guidelines in the field. The assumption underpinning interest in practice guidelines is that they will reduce unexplainable variations in the type of behavioral health care that is delivered, thus improving the overall quality. As suggested in the previous section, the evidence supporting this assumption is mixed, although the conceptual stance is intuitively appealing. Nevertheless, guidelines must be evaluated in field-based randomized studies that examine their impact on both processes and outcomes of behavioral health care. There is a need to demonstrate that guidelines produce their intended impact. Otherwise, the behavioral health community may be compelled to look for other methods for improving the quality of their care.

See also: Clinical Psychology: Manual-based Treatment; Managed Care; Managed Care and Psychotherapy in the United States; Medical Experiments: Ethical Aspects; Pecuniary Issues in Medical Services: Ethical Aspects; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Clinical Trials; Psychotherapy: Ethical Issues; Psychotherapy: Legal Issues (United States)

Bibliography

American Psychological Association Task Force 1995 *Template for Developing Practice Guidelines: Interventions for Mental Disorders and Psychological Aspects of Physical Disorders.* American Psychological Association, Washington, DC

Bernstein G A, Shaw K, Dunne J E 1997 Practice parameters for the assessment and treatment of children and adolescents with anxiety disorders. *Journal of The American Academy of Child Psychology* **36**: S49–S84

Clinton J J, McCormick K, Besteman J 1994 Enhancing clinical practice: The role of practice guidelines. *American Psychologist* **49**: 30–3

Depression Guideline Panel 1993 *Depression in Primary Care: Detection, diagnosis and treatment*. (Clinical practice guideline No. 5, AHCPR publication No. 93–0553.) US Department of Health and Human Services. US Public Health Service. Agency for Health Care Policy and Research, Rockville, MD

Field M, Lohr K (eds.) 1990 *Guidelines for Clinical Practice: From Development to Use*. Institute of Medicine Committee on Clinical Practice Guidelines. National Academy Press, Washington, DC

Eddy D 1996 *Clinical Decision Making: From Theory to Practice. A collection of Essays from SAMA*. Jones & Bartlett, Boston

Grimshaw J M, Russell I T 1993 Effect of clinical guidelines on medical practice: A systematic review of rigorous evaluations. *Lancet* **342**: 1317–22

Herz M I, Liberman R P, Liberman J A 1997 Practice guideline for the treatment of patients with schizophrenia. *American Journal of Psychiatry* **154**(4): 1–63

Lomas J, Anderson G M, Dominick-Pierre K, Vayda E, Enkin M, Hannah W 1989 Do practice guidelines guide practice? The effect of a consensus statement on the practice of physicians. *New England Journal of Medicine* **321**: 1306–11

Nathan P E 1998 Practice guidelines: Not yet ideal. *American Psychologist* **55**: 290–9

Noonan D, Coursey R, Edwards J, Frances A, Fritz T, Henderson M, Krauss A, Leibfried T, Manderscheid R, Minden S, Strosahl K 1998 Clinical practice guidelines. *Journal of the Washington Academy of Sciences* **85**: 114–12

Practice Guideline Coalition 1997 *Mission Statement and Framework for Developing Practice Guidelines*. Available from Practice Guideline Coalition, Department of Psychology/296, University of Nevada-Reno, Reno, NV. 89557

Strosahl K 1994 Entering the new frontier of managed health care: Gold mines and land mines. *Cognitive and Behavioral Practice* **1**: 5–23

Strosahl K 1995 Three 'gold-mine land-mine' themes in generation of managed care. *Behavior Therapist* **18**: 52–54

Woolf S H 1992 Practice guidelines: A new reality in medicine, Part II. Methods of developing guidelines. *Archives of Internal Medicine* **152**: 946–52

K. Strosahl and P. Robinson

Psychotherapy: Ethical Issues

When attempting to survey the landscape of ethical standards designed to deal with the practice of psychotherapy, one quickly encounters a plethora of formal codes fraught with, what at least initially appears to be, unrelated design, length, and content. Each of these codes reflects the official position taken by a professional body regarding ethical conduct on the part of the membership of their respective organizations (psychologists, psychiatrists, social workers, etc.). These differences, however, often create confusion among psychotherapists regarding just what is right or what is ethical. The problems created by these inconsistencies only become more serious when one is searching for ethical guidance on how to behave when dealing with ethical problems. Because of the inherent confusion that can frequently come from these codes, and the laws that come from them, perhaps the best way to address the topic of ethics in psychotherapy would be by turning away from the specific codes themselves to a study of the fundamental principles that form the foundation for ethical conduct in healthcare. Then, with an improved understanding of ethical theory one can move to a review of how these principles are applied to the practice of psychotherapy and to how they can give ethical guidance to the individual psychotherapist (Eberlein 1987).

A good source for information regarding ethical issues in healthcare can be found in the *Principles of Biomedical Ethics*, by Beauchamp and Childress (1994). These authors take the position that all ethical conduct in healthcare is based upon what they call a four-principled approach to biomedical ethics. They believe that the four primary moral principles that drive ethics in healthcare are: respect for autonomy, beneficence, non malfeasance, and justice. Simply put, healthcare professionals are ethically bound to try to help those they treat (beneficence) while avoiding making things worse (non malfeasance). In addition, healthcare practitioners should have a fundamental respect for the patient's right to choose (autonomy) and should practice their profession in a system that is equitable and fair (justice). Other authors have argued for the addition of a fifth principle, that of fidelity, to this list (Kitchener 1984; Koocher and Keith-Speigel 1999). As applied to healthcare, fidelity deals with the healthcare provider's obligation to act in good faith, be honest, keep promises and to fulfill obligations to those they treat. Since these five fundamental principles are generally thought to compose the foundation of ethics in healthcare, they warrant closer scrutiny.

1. *Respect for Autonomy*

Respect for autonomy is tied to the belief that people have a right to hold views and make choices that reflect their own beliefs and that these beliefs do not need to be shared with, or approved of by, others. Thus, autonomy is a fundamental component of the concept of freedom and, as applied to healthcare, this freedom allows the patient, and not the provider, to control treatment.

Perhaps the best place to find evidence of the principle of autonomy as applied to the practice of psychotherapy is with informed consent to treatment. In its basic form, informed consent requires that the provider tell the patient about the risks and benefits of

a specific treatment prior to commencing the treatment itself. In addition, the person consenting must have the capacity to consent and must do so freely (Stromberg et al. 1992). In application, appropriate informed consent serves to level the field in treatment, and allows the patient to become the decision-maker when dealing with his or her well-being. It also places the obligation to fulfill the spirit of autonomy in psychotherapy upon the therapist, who must make sure the decision-maker has enough information to make an informed choice that is free from coercion.

Many authors have argued that good informed consent is not just limited to an exchange that occurs at the outset of therapy but is actually a concept that is applied throughout the treatment process (Pope and Vasquez 1998, Younggren 2000). That is, individuals who find themselves in therapy have a right to understand their treatment, its direction and any changes that occur in its direction. So as the direction of therapy shifts and as the types of interventions used change, the therapist must make sure the patient understands and agrees to these changes. Once again, it is the patient who must be in control of the decision-making process.

Satisfying the requirements of informed consent does not end with fulfilling these obligations to the patient. A good informed consent additionally requires honesty on the part of the psychotherapist regarding comprehensive treatment issues and related matters. If the information shared with the patient is not truthful or complete, then the whole structure of informed consent is threatened since one could not consent to engage in treatment without an honest explanation of that treatment by the therapist.

The requirement of honesty in informed consent raises a number of interesting questions for providers who work in today's complex healthcare system. For example, a therapist may work with a managed care company whose contract, in application, is more limited than the patient believes. Would it be ethical for this therapist not to inform his or her patients of that reality? Are not a patient's rights compromised by a therapist's decision to not tell the patient of the limitations of that insurance when, in fact, the patient might truly not choose to proceed under such circumstances? Is this purposeful lack of clarity unethical? It is clear that the the principle of autonomy requires that the patient understand these types of realities prior to commencing treatment. However, fulfilling this requirement can create serious conflicts of interest for the psychotherapist because the detailed explanation of insurance limitations may be frowned upon by some managed care companies who see such conduct as noncooperative. It goes without saying the providers who are viewed as not being cooperative with the expectations of a managed care company run the risk of losing their provider contracts. All the same the ethical reality remains, that purposefully keeping realities about therapy away from a patient truly violates one of the fundamental tenets of informed consent and that this conduct is unethical. Thus, to conform with the standards of good ethical conduct, therapists must engage in a description of psychotherapy with their patients that is not only detailed, understandable, and noncoercive, but also open, honest and free from conflicts of interest.

Also inherent in the principle of autonomy is a respect for the rights of patients to control the exchange of private information about themselves and to understand the limitations of that control. Once again, patients must, as a part of informed consent, understand how the psychotherapist is going to deal with their personal and, frequently, very private information. This includes an explanation of what information will be shared with insurance companies, other professionals and other family members, to name a few. The spirit of this principle even extends to minors whose rights, in reality, are limited legally. Psychotherapists who work with children have an ethical obligation to make sure that those children understand what is going to be shared with others and what their legal rights are. While children may not be of the legal age to consent, the philosophy of informed consent clearly demands that minors are asked to give age-appropriate informed consent and the child therapist who fails to fulfill this obligation with his or her patients violates the minor's autonomy rights.

Interestingly, the implementation of the principle of autonomy in the practice of psychotherapy does not allow the therapist to condone destructive behaviour that may result from patients' expressing their autonomy. Rights to autonomy exist in balance with other moral principles. Thus, ethical mental health professionals understand that destructive choices that damage oneself, like suicide, or damage another, like homicide or child abuse, become expressed exceptions to autonomy rights. In addition, there are settings, like prisons, where individuals have by definition had their freedom of choice restricted and, consequently, these individuals may not be allowed the full benefit of the principles of autonomy. Regardless of these types of exceptions, the general exceptions to autonomy should be clearly understood by all patients prior to beginning psychotherapy.

2. *Beneficence*

Fundamental to all healthcare is a belief that the provider is attempting to improve the patient's condition through the delivery of a professional service. This belief is also key to the practice of psychotherapy. Upon first assessment, the value of this principle seems so obvious that some would say it should stand without explanation. However, when applied to the psychotherapy loosely and without being carefully defined, the implementation of beneficence as a principle can become confusing and complex and can raise

many questions. These questions include the following: Is it an expression of beneficence to help a patient gain insight into a very painful, historical reality consistent with an insight-oriented model of therapy, with a risk that such insight will result in psychological distress? Where is beneficence found in an extensive, analytically-oriented psychotherapy that can have no concrete or well-defined goal? Is beneficence evident when the therapist's intentions are good but when the model of intervention has no scientific merit? The answers to the question are elusive and frequently cause extensive, passionate debate among professionals. Psychotherapists should, at a minimum, have a theoretical rationale that answers these and related questions.

Theoretical debates aside, there are some aspects of beneficence that should universally be found in the ethical practice of psychotherapy. Ethically-based psychotherapists who practice with an eye toward beneficence, practice therapies that are, at a minimum, arguably good treatments for the problems their patients bring to them. This type of therapist does not do the same type of therapy with each person who comes through the door. Implied in this is the corollary that the ethical therapist does not see the same problem in every patient he or she treats. The ethical psychotherapist plans therapy with a respect for the empirical and theoretical support that exists, showing that a specific intervention is the right thing to do for the specific problem being addressed.

Respect for beneficence also requires that the therapist engage in activities that improve the quality of their treatments. That is, they engage in consultation, continuing education and literature reviewing in order to stay current with the changing status of their profession. The ethical therapist also constantly combines the law, current literature, and community standards in using a model of intervention that has the highest probability of success and the ethical therapist is secure in his or her ability to justify their professional conduct. This style of practice assures the ethical psychotherapist that the intervention in a specific case is both consistent with the standard of care for a psychotherapist and in the best interests of the patient.

It goes without saying that practicing an outmoded model of intervention, in the face of clear evidence that another treatment could be more effective, would likely be a violation of the principle of beneficence in psychotherapy. There is little justification for a psychotherapist to apply a model of intervention that is less efficient, more costly and less likely to be successful when compared with another treatment, without a clear understanding or rationale for doing so. Once again, the reality of managed care constraints lends itself well to a discussion of the application of the ethical principle of beneficence to psychotherapy. That is, ethical psychotherapists do not allow the constraints placed upon them by insurance benefits or managed care contracts to lead them away from what they believe would be the most effective intervention. The ethical therapist, at a minimum, must make sure that both the patient and the insurance company know that effectiveness of therapy is being administratively threatened by insurance constraints and all parties must know what the recommended treatment is. Most importantly, the ethical psychotherapist attempts to resolve any administrative insurance conflicts with a constant eye toward what is in the patient's best interest.

Finally, the ethical psychotherapist does not continue a treatment that is no longer benefiting the patient. When faced with an intervention that is leading nowhere, ethical constraints would demand a change in direction that would result in either a change in the type of therapy or, if absolutely necessary, the termination of treatment. To do otherwise is potentially exploitive of the patient, and denies the patient access to other, potentially more effective interventions: a clear violation of the principle of beneficence.

3. *Non Malfeasance*

It comes as no surprise that ethical psychotherapists are supposed to avoid doing bad or destructive things to and with their patients. In today's litigious society this has become a very powerful principle and is felt by some authors to take precedence over all other ethical principles (Marsh 1997, Harris 2000). It is obvious that therapy is not supposed to be a damaging process and it is the responsibility of the ethical therapist to minimize risk in order to avoid, at all costs, engaging in a process that is destructive. This prohibition against doing destructive things is very broad and not only applies specifically to the treatment aspects of psychotherapy but can also include administrative aspects of the treatment relationship.

Once again defining exactly what is destructive and damaging about psychotherapy can be quite subtle and elusive. As pointed out earlier, risk can frequently become a matter of theoretical debate among mental health professionals from differing theoretical orientations. Regardless of this lack of consistency and conceptual clarity, in general, therapists should constantly evaluate whether their interventions are potentially damaging to their patients. This is not only in the patient's best interests but also protects the therapist. However, this does not mean that ethical therapy must be void of risk, since there is probably some degree of risk in all treatments. What it does mean is that therapists do not engage in delivering treatments to patients that have risk without the patient understanding, as much as possible, the nature of those risks. If the term 'risk' comes from controversy in the field, and/or the lack of strong empirical support, it becomes all the more important that the patient receive a detailed explanation of these issues.

Non malfeasance principles would also prevent a therapist from engaging in extratherapeutic conduct with a patient that has no theoretical or empirical treatment value whatsoever, or that has even been shown to be potentially destructive. For example, a patient could never give informed consent to have sex with a therapist because research has already shown such conduct to be potentially very destructive to patients and to be of no treatment value. Even if the therapist informed the patient of these risks, agreement on the part of the patient would not reduce the serious nature of this type of ethical violation. A patient also could never give consent to allow the therapist to serve conflicting roles in a case regardless of the rationale for so doing. This prohibition against engaging in anything that is potentially destructive in psychotherapy includes patient exploitation, physical abuse, boundary violations, multiple relationships, and more. The seriousness of these types of violations is only made more obvious by the reality that this type of potentially dangerous conduct usually has nothing to do with psychotherapy and that it is frequently the result of designing the treatment relationship to meet the needs of the therapist, and not the patient.

4. *Justice*

The ethical principle of justice is the foundation for other very valuable healthcare concepts like fairness and equality. Justice should be evident whenever persons are due benefits from others because of their particular circumstance, as when they are ill or when they are struggling emotionally. The fulfillment of this principle is based upon the professional's respect for the fair and equal distribution of services regardless of societal differences like race, status, religious orientation, etc. (Kitchener 1984).

The principle of justice, as a concept applied to psychotherapy, takes on differing forms depending upon circumstance. Limited assets, limited time, limited staffing all are variables that have direct impact upon the implementation of justice in a treatment setting. Under some circumstances the lack of justice in the provision of professional services becomes clear; however, in others, it becomes clouded. For example, it would be obvious to most that a therapist who changes services based upon a patient's race would be unethical. In this circumstance, the lack of justice is easy to evaluate. However, what about when people cannot afford care because of poverty, societal unfairness or when benefits are limited or simply not available? How does the ethical psychotherapist pursue justice under these circumstances? Or, is the ethical therapist even responsible for addressing these types of issues when delivering care? Under these types of circumstances true justice becomes quite difficult to pursue.

In reality, things are not often just, and inequality is a part of the human circumstance. Suffice it to say that ethical psychotherapists are those who work toward justice in treatment but are realistic in so doing. That is, they aim to do the best they can in psychotherapy, to be fairest to the persons they are treating. In fact, this is the psychotherapist's fiduciary obligation in psychotherapy.

In application, the ethical psychotherapist might pursue justice in therapy by providing a few extra termination sessions to the patient who has lost insurance benefits. He or she could offer a few pro bono visits to the patient who does not have the financial means to continue treatment or the fee for therapy could be reduced so that the patient could afford it. Or he or she could also volunteer to give time to those who do not have the means to pay for the services they receive. Regardless of the actual case, the ethical psychotherapist pursues justice in treatment and makes this one of the key principles in ethical decision-making.

5. *Fidelity*

The final principle of ethical decision-making is that of fidelity. Fidelity addresses a person's responsibility to be loyal and truthful in their relationships with others. It also includes promise keeping, fulfilling commitments, and trustworthiness (Welfel and Kitchener 1992). In healthcare, however, the implementation of this principle extends beyond the regular responsibilities of business or contractual fulfillment to the creation of relationship based upon trust: the trust the patient has that the professional will always operate in the patient's best interests and will always fulfill their professional obligations to operate in his or her best interests. Some authors have taken the position that fidelity, along with non maleficence, are the most legally salient moral principles (Bersoff and Koeppl 1993).

As applied to psychotherapy, fidelity forces the professional to orient toward the patient's needs and not his or her own. Thus, the therapist who allows a sexual relationship to develop in treatment would be violating this principle. Clearly such a relationship is designed to meet the therapist's needs and becomes exploitive. This principle also applies to patient abandonment since such conduct would violate the trust the patient places in the psychotherapist, trust that the therapist will see the patient through the treatment process. Finally, fidelity also applies to circumstances where professionals allow administrative factors, like benefits limitations, fee capitation, and contractual rebates to affect their treatment decisions.

Fidelity, as applied to healthcare, is a dynamic process: and the way in which a therapist maintains fidelity in treatment changes as treatment changes. Fidelity also blends with the other principles of ethical decision-making to require that the therapist be open

and honest in the treatment relationship and to make sure that the patient's expectations are consistent with what the therapist can provide. Fidelity requires psychotherapists to make the best choices they can for those that they treat.

6. Summary

The ethical psychotherapist is one who understands the fundamentals of ethical philosophy and knows how to apply them to individual cases. If a psychotherapist is capable of viewing ethical dilemmas from the perspective of ethical philosophy, decisions regarding the correct choice, if you will, become much easier to find. This approach is much more effective than a rule-based system. Rule-based guidance will never give the psychotherapist ethical security because the profession is complex and is fraught with theoretical debate and inconsistency. The search for rules, versus an understanding of principles that generate those rules, will only create a complex, awkward, and laboriously large system that would truly obscure a professional's ability to evaluate problems and to then choose the best course of action. Therefore, the answers to ethical questions can be found in the thoughtful implementation of the five principles of autonomy, beneficence, non malfeasance, fidelity, and justice in psychotherapy.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Ethics and Psychiatry; Ethics in Organizations, Psychology of

Bibliography

American Psychological Association 1992 Ethical principles of psychologists and code of conduct. *American Psychologist* **47**: 1597–611

Beauchamp T L, Childress J F 1994 *Principles of Biomedical Ethics*, 4th edn. Oxford University Press, New York

Bersoff D N, Koeppl P M 1993 *Ethics and Behavior* **3**: 345–57

Eberlein L 1987 Introducing ethics to beginning psychologists: A problem solving approach. *Professional Psychology: Research and Practice* **18**: 353–59

Harris E 2000 *Ethical Decision Making and Risk Management in Clinical Practice*. A workshop presented by the American Psychological Association Insurance Trust at the meeting of the American Psychological Association, Washington, DC

Kitchener, K S 1984 Intuition, critical evaluation, and ethical principles: The foundations for ethical decisions in counseling psychology. *The Counseling Psychologist* **12**(3): 43–55

Koocher G P, Keith-Spiegel P 1998 *Ethics in Psychology*, 2nd edn. Random House, New York

Marsh, D T 1997 Ethical issues in professional practice with families. In: Marsh D T and Magee R D (eds.), *Ethical and Legal Issues in Professional Practice with families*, pp. 3–26 John Wiley, New York

Pope, K S, Vasquez M J T 1998 *Ethics in psychotherapy and Counseling*, 2nd edn. Jossey-Bass, San Francisco

Stromberg C D, Haggarty D J, Leibenluft R F, McMillan M H, Mishkin B, Rubin B L, Trilling H R 1988 *The Psychologist's Legal Handbook*. The Council for the National Register of Health Service Providers in Psychology, Washington, DC

Welfel E, Kitchener K 1992 Introduction to the special section: Ethics education—an agenda for the 90s. *Professional Psychology: Research and Practice* **23**: 179–81

Younggren J N 2000 Is managed care really just another unethical Model T? *The Counseling Psychologist* **28**(2): 253–62

J. Younggren

Psychotherapy, History of: Psychiatric Aspects

Psychotherapy—the systematic application of psychological principles to accomplish symptomatic or more substantial personality change—has its origins in two healing traditions—the magico-religious and the medico-scientific (Bromberg 1959). The idea that human experience is influenced by supernatural forces and that certain people have the power to intercede with these forces dates back to antiquity. Shamans and sorcerers have resorted to amulets, magical potions, and incantations for centuries. They have also conducted exorcism and induced altered states of consciousness.

With the Age of Reason, interest was aroused in the newly discovered phenomena of electricity and magnetism. Anton Mesmer (1734–1815), a brash physician, drew on the latter when propounding his theory of animal magnetism. Since the body contained 'magnetic fluid,' and this could become disturbed, by applying magnets to various parts a patient could be relieved of pain and other ailments. While Mesmer's work was ultimately denounced by a Royal Commission and he was forgotten, the observation that his methods appeared to help patients continued to attract attention. James Braid (1795–1860), a Manchester doctor, noted that he could induce a similar trance-like state by getting people to fix their gaze on a luminous object. Labeling the phenomenon hypnotism, he proceeded to demonstrate its effectiveness in treating a range of conditions.

The doyen of French neurology, Jean Martin Charcot (1835–93), considered hypnosis a neurophysiological process, and set about its study in states like somnambulism and hysteria. His pupil Pierre Janet (1859–1947) suggested that the latter was brought about by weakening of a higher brain function, resulting in a constriction of consciousness. In this 'dissociated' state thoughts could not be integrated, and symptoms were beyond the reach of consciousness.

In an alternative explanation, Ambroise Liebeault (1823–1904), and Hippolyte Bernheim (1840–1919) in France posited a narrowing of attention in hypnosis

which rendered the patient vulnerable to the therapist's influence by suggestion. Far from invalidating hypnosis, the power of suggestion could be investigated scientifically and accepted as a bona fide therapy.

Freud (1856–1939) had studied briefly with Charcot in 1885–6 and became enthusiastic about new possible treatments for hysteria. Freud's mentor, a respected Viennese physician, Josef Breuer (1842–1925), had described the treatment of a young woman who suffered from an array of marked hysterical symptoms. Instead of suggesting disappearance of the symptoms, Breuer had encouraged Anna O (the pseudonym given to her) to talk freely about her life under hypnosis. She did so and over time began to share memories about her father's illness. Breuer noted that after each session, his patient became less distressed and her symptoms improved. Anna herself referred to this approach as her 'talking cure' or 'chimney sweeping'; Breuer termed it 'catharsis.' Whereas Charcot had described physically traumatic events as possibly causal, in his hysterical patients, Breuer and Freud's interest focused on psychological trauma such as humiliation and loss. Freud also became less keen on applying hypnosis.

Their volume, *Studies in Hysteria*, which included the case of Anna O, and four others treated by Freud himself, also contained ideas on defense mechanisms and an account of various techniques including suggestion, hypnosis, and catharsis. Most importantly, however, was the new concept of 'free-association': the patient was instructed to disclose whatever came into their mind, without any censoring, while the analyst adopted an attitude of 'evenly-suspended attention' in listening to and interpreting the material.

On the basis of his subsequent self-analysis and further clinical experience, Freud elaborated his theory of infantile sexuality. Sexual fantasies in young children centered around the triangular relationships of love and rivalry with their parents (the Oedipal complex). Promoting free association of this theme, as well as of dreams, slips of the tongue, and other unconsciously-based thoughts, feelings, and fantasies, Freud emphasized the centrality of the complex in the neuroses. While catharsis and insight (making the unconscious conscious) seemed pivotal to therapy, Freud soon realized that other factors also operated, particularly the role of transference, i.e., how feelings, thoughts, and fantasies stemming from childhood experiences and revived in current relationships are 'transferred' onto the analyst.

Interpretations were directed at this threefold nature of the analytic experience. The therapist's response to transference (countertransference) led to analysts having their own personal therapy, the aim being to minimize it, and thus to be thoroughly objective in attending to the patient's disclosures (Ellenberger 1970).

In summary then, psychoanalytic psychotherapy elaborated the following processes: the patient's unbridled disclosure of whatever enters his or her mind—free association; the transference of infantile and childlike feelings and attitudes to the therapist which were previously directed to key figures in the patient's earlier life; interpretation of the transference as well as of defenses the patient applies to protect him/herself and the resistance he/she manifests to self-exploration; and finally, the repeated working through of the discoveries made in treatment. The ultimate aim is insight with translation into corresponding change in behavior and personality.

1. *Variations on a Theme*

Freud revised his theories at numerous points in his long professional life. In addition, many colleagues extended the boundaries of psychoanalytic thinking. But differences of opinion soon surfaced, some radical. The nature of psychoanalysis then, especially Freud's intolerance of dissent, led notable figures to leave the fold and to evolve their own models on which corresponding schools of therapy were established. Jung and Adler were two foremost European dissidents; Karen Horney, Erich Fromm, and H. S. Sullivan were pioneering neo-Freudians in the US; and W. D. Winnicott and Melanie Klein were prominent in Britain (Mitchell and Black 1995).

Carl Gustav Jung is perhaps the most critical dissenter since he was clearly being groomed to take over leadership of the psychoanalytic movement. But he was also celebrated because of his own contribution. Specifically with regard to psychotherapy, he advanced the notion of 'individuation'—the aim was to discover all parts of oneself and one's creative potential. Jung was less concerned than Freud with the biological roots of behavior, especially infantile sexual development, but rather emphasized social and cultural factors. The role of transference was replaced by a more adult type of collaboration.

Furthermore, the unconscious for Jung is not merely the repository of the individual's history but also a wider social history, a phenomenon Jung labeled the collective unconscious. He arrived at this notion through a study of myths, legends, and symbols in different cultures, in different epochs. That these are shared by a variety of cultures is not fortuitous but reflects cosmic mythical themes or archetypes, a salient feature of man's collective history.

Alfred Adler (1870–1937), a Viennese physician, was like Jung a prominent dissenter. Adler joined Freud's circle in 1902 but broke away to form his own school of 'individual psychology' after nine years. An important tenet concerns individual development. We begin life in a state of inferiority, weak, and defenseless, for which we compensate by striving for power and by evolving a lifestyle to make our lives pur-

poseful. The pattern that emerges varies and may include such goals as acquisition of money, procreation, high ambition, or creativity. The drive for power and choice of lifestyle may, however, go awry in which case a neurosis results. A path is followed which leads to ineffective efforts to cope with the feeling of inferiority and assumption of a facade or false self.

Adler regarded therapy as a re-educative process in which the therapist, who serves as model and source of encouragement, engages in a warm relationship with the patient and enables him or her to understand the lifestyle he/she has assumed. Unconscious determinants of behavior are less crucial than conscious ones and the term 'unconscious' is used only descriptively to refer to aspects of the person that are not understood by him.

Freud's insistence on innate drives and infantile sexuality not only led to the schisms with Jung and Adler but also spurred a new generation of analytically oriented analysts to concentrate on interpersonal aspects of psychological experience. In the US, Sullivan pioneered the school of 'interpersonal psychiatry,' in which the therapist adopted the role of participant observer in treatment, the transference providing one opportunity to explore communication and its breakdown in the patient's interpersonal world.

We can consider the chief features in the approach of the neo-Freudians by looking briefly at Karen Horney (1885–1952). She became disenchanted with Freud's rigid focus on instinctual biological factors, arguing that cultural factors were more salient, as reflected in the differences in psychological development and behavior among different sociocultural groups. Indeed, behavior regarded as normal in one culture could be viewed as neurotic in another.

In line with her emphasis on culture, Horney advanced the role of parental love in the life of a young child. Children typically suffer anxiety, a consequence of feeling helpless in a threatening world. The child reared in a loving atmosphere succeeds in overcoming basic anxiety. By contrast, the deprived child comes to view the world as cruel and hazardous. An inevitable result is low self-esteem.

The task of therapy is to examine the patient's defective patterns in relating to others. In part this is achieved through study of what transpires between patient and therapist. But there is no emphasis on transference as occurs in Freudian analysis. Therapy aims to enable the patient to move *with* others, by engaging in relationships which are reciprocal and mutual. Another goal is greater self-realization, with freedom from determined modes of thought and action.

While the neo-Freudians were establishing a new pathway for psychoanalysis, a group was equally innovative in the UK and France. In the UK, Melanie Klein (1882–1960), a physician hailing from Berlin, concluded from her study of children's play that the Oedipal complex, as described by Freud, was a late expression of primitive unconscious anxieties which children experienced in the first 18 months of life. Given that these states could be reactivated in adult life, Klein labeled them 'positions' rather than stages of psychological development. She also supported the move away from Freud's formulation of instincts and psychosexual energy toward one of object relations, namely the representations of the perception and experience of significant others. Similarly, transference was to be understood not only as a repetition of past or current relationships, but as evidence of relational patterns in the patient's current *internal* world.

In tandem with Klein, other analysts in Britain proposed ways in which significant early relationships, particularly between infant or young child and its mother, exerted a formative influence on the developing psyche. John Bowlby (1907–90), in particular, observed that children who had suffered a prolonged separation from their mother underwent a grief experience which predisposed them to a range of psychopathology in later life. Donald Winnicott (1896–1971), a pediatrician and analyst, proposed that a mother's attunement to her child, enabling him to feel psychologically 'held' and understood, was an essential ingredient for a stable sense of self. The American psychoanalyst Heinz Kohut advanced a similar concept, but highlighted the place of empathy whereby the mother functions as a 'self object' for the child.

Wilfred Bion, a student of Klein, suggested that the analyst functions as a 'container' into whom the patient projects their 'unprocessed' feelings and experiences; the analyst returns them later in a form accessible to him to work on. This experience is deemed to be crucial to the development of a person's ability to examine his or her own state of mind and that of others.

Michael Balint, an analyst at the Tavistock Clinic, worked along similar lines to Winnicott, focusing on the opportunity for the patient to 'regress' in the safety of the therapeutic relationship to a state of mind in which differentiating between patient and therapist may blur. The features of the encounter—its regularity, predictability, and the therapist's empathic attitude—creates circumstances for such regression which can then be used to examine and alter defense mechanisms.

These models have implications for practice in that the therapist expresses moment-to-moment empathic understanding of the patient's inner experiences and as this is repeated time and again the process promotes the evolution of a coherent identity.

Jacques Lacan (1901–81), the maverick French analyst, claimed a 'return to Freud,' his early work in particular, but through the prisms of semiotics and linguistics. These disciplines, he contended, do not see language as a value-free means to express ideas and convey meaning but as a system of signs that communicate symbolically tacit rules and power arrange-

ments. Language imposes on the child awareness of a separateness from its mother who represents a fantasized internal object (the imaginary mother), in whose eyes the child mirrors itself in an attempt to achieve a state of psychological unity. Lacan's controversial proposals regarding variable duration of treatment and trainees themselves determining their competence to qualify as psychoanalysts led to his expulsion from the International Psychoanalytic Association, albeit with a dedicated following.

Humanistic experientially oriented schools of therapy became popular in the 1960s and 1970s, mainly in the US. Their chief features were: challenging patients with the way they avoid emotionally significant matters in the 'here-and-now'; increasing awareness of nonverbal aspects of communication; facilitating emotional arousal; and providing a forum in which patients were encouraged to experiment with new behaviors.

Some of these schools have incorporated aspects of psychoanalytic theory. Transactional analysis, for instance, makes playful use of the concept of ego states, represented as parent, adult, and child; primal scream therapy is predicated on heightened arousal of the purported emotional trauma of birth; and psychodrama encourages people to enact their conflicts by personifying various roles in their psychological lives.

The most influential humanistic school has been the client-centered therapy of Carl Rogers, although it has been absorbed into psychotherapy generally and not always with appropriate acknowledgment. Its premise is that if the therapist creates an atmosphere that is nonjudgmental, empathic, and warm, people can realize their potential for self-expression and self-fulfillment.

Feminist-based psychotherapies, still evolving since the 1980s, reflect a range of opinion from those which condemn Freud as a misogynist who deliberately recanted his discovery of the actuality of childhood sexual abuse, to those who appreciate his linkage of life experience, innate drives, and corresponding defense mechanisms in influencing vulnerability to psychopathology (Elliot 1991, Appignanesi and Forrester 1992). The latter point of view, echoing a broader debate within psychoanalysis of the respective roles of internal and external reality, has led some feminist therapists to stress social reality, and others to stress physical reality.

2. *The Development of Other Modes of Psychotherapy*

Our focus, hitherto, has been on psychotherapy of the individual patient. The second half of the twentieth century saw the evolution of treatments conducted with more than one person: groups of strangers and families. In this section, we provide a brief account of the development of these two modes.

Although therapy of stranger groups was experimented on early in the century, it was not until World War II that major developments ensued. The exigencies of war were the spur in that the group format proved highly economical for dealing with huge numbers of soldier-patients. The Northfield Military Hospital in the UK was a center of remarkable innovation, led by psychoanalytically oriented therapists like Wilfred Bion and S. N. Foulkes. Bion's (1961) influence pervaded subsequent work at the Tavistock Clinic, while Foulkes (1965) was the founding father of the Institute of Group Analysis (both in London).

In the US, the move to the study of group process was spearheaded by social psychologists, particularly Kurt Lewin. A national center was established to train people in human relations and group dynamics. Participants from diverse backgrounds studied group functioning in order to act more effectively in their own work settings. Paralleling the human potential movement in the 1960s, the group movement transferred its focus from group to personal dynamics, and before long the encounter group movement had evolved with its thrust of promoting greater self-awareness and personal growth. Encounter groups became widespread during the 1960s and 1970s but after an initial fervor, declined both in terms of membership and appeal.

Formal group therapy under professional leadership then assumed a more prominent role. The most popular model was fathered by Irvin Yalom who drew upon an interpersonalist approach, one originally molded by Harry Stack Sullivan who proposed that personality is chiefly the product of a person's interaction with other significant people. Yalom pioneered the study of therapeutic factors specific to the group rather than following the pattern of transposing psychoanalytic or other theories from the individual setting. His *Theory and Practice of Group Psychotherapy* (Yalom 1995) became exceedingly influential, attested to by the appearance of four editions.

Constrained perhaps by Western medicine's focus on the individual patient, psychiatry was slow to develop an interest in the family (Gurman and Kniskern 1991). Scattered through Freud's writings are interesting comments about family relationships and their possible roles in both individual development and psychopathology. His description of processes like introjection, projection, and identification explained how individual experiences could be transmitted across the generations in a family. Influenced by the work in the UK of Anna Freud, Melanie Klein, and Donald Winnicott, the child guidance movement developed a model of one therapist working with a disturbed child and another with the parents, most often the mother on her own. The two clinicians

collaborated in order to recognize how the mother's anxieties distorted her perception and handling of her child, which were added to the child's own anxieties.

Things took a different turn in the US. There, Ackerman (1958) introduced the idea of working with the family of a disturbed child using psychoanalytic methods in the 1950s. An interest in working with the family, including two or more generations, arose concurrently. Thus, Murray Bowen (1971) found that the capacity of psychotic children to differentiate themselves emotionally from their families was impaired by the consequences of unresolved losses, trauma, and other upheavals in the lives of parental and grand-parental generations.

Boszormenyi-Nagy and Spark (1984) also addressed this transgenerational theme by describing how family relationships were organized around a ledger of entitlements and obligations which conferred upon each participant a sense of justice about their position. This, in turn, reflected the experience in childhood of neglect or sacrifices for which redress was sought in adult life.

Bowen also introduced the principles of Systems Theory into his work but it was Salvador Minuchin, working with delinquent youth in New York, who highlighted the relevance of systems thinking to their interventions. The youngsters often came from emotionally deprived families, headed by a demoralized single parent (most often the mother) who alternated between excessive discipline and helpless delegation of family responsibilities to a child. Minuchin's Structural Family Therapy deploys a series of action-oriented techniques and powerful verbal metaphors which enable the therapist to join the family, and to re-establish an appropriate hierarchy and generational boundaries between the family subsystems (marital, parent/child, siblings).

Another major development took place in Palo Alto, California, where a group of clinicians gathered around the anthropologist Gregory Bateson (1972) in the 1950s. They noted that implicit in communication were *meta-communications*, which defined the relationship between the participants. Any contradiction or incongruence carried great persuasive, moral, or coercive force and formed part of what they labeled a 'double-bind'; they proposed this as a basis for schizophrenic thinking.

All these system-oriented views assume that the family is a system observed by the therapist. However, therapists are not value-neutral. They may take an active role in orchestrating change in accordance with a model of family functioning. Yet these models ignore therapists' biases as well as the relevance of their relationships with families. This probably reflected the determination of some American family therapists to distance themselves from psychoanalytic theory, and also led them to neglect the family's history, how it changed throughout the lifecycle, and the significance of past traumatic and other notable life events.

In response to these criticisms there was a move away from the problem-focused approach which had characterized most behavioral and communication views of psychopathology. The Milan school (1980) whose founders were psychoanalysts developed a new method of interviewing families in conjunction with observers behind a one-way screen formulating and then presenting to the family and therapist hypotheses about 'their' system.

Family therapists also began to consider that families might be constrained from experimenting with new solutions because of the way they had interpreted their past experiences or internalized explanatory narratives of their family, the expert's, or society at large. This led to a shift from considering the family as a social to a linguistic system.

The narrative a family conveys about their lives is a construction which organizes past experience in particular ways; other narratives are excluded. When a family with an ill member talks to a professional, conversations are inevitably about problems (a problem-saturated description). The participants ignore times when problems were absent or minimal, or when they successfully confined problems to manageable proportions. A different story might be told if they were to examine the context that might have led, or could still lead, to better outcomes.

A number of narrative approaches have applied these concepts. Philosophically, they align themselves with postmodernism, which challenges the notion of a basic truth known only by an expert.

Many criticisms of systems approaches have been leveled including:

(a) Disregard of the subjective and intersubjective experiences of family members.

(b) Neglect of their history.

(c) Denial of unconscious motives which affect relationships.

(d) Although family members are reciprocally connected, the power they exert on one another is not equal; this is highlighted in the violence against women and in child abuse.

(e) Inequality and other forms of injustice based on societal attitudes toward differences in gender, ethnicity, class, and the like, are uncritically accepted as 'givens'.

(f) Minimizing the role of the therapeutic relationship, including attitudes members develop toward the therapist and her feelings toward them.

This critique has led to growing interest in integrating systems-oriented and psychoanalytic ideas, particularly those derived from object-relations theory. One variant is John Byng-Hall's (1995) synthesis of attachment theory, systems-thinking, and a narrative approach.

A further criticism of systems-oriented approaches is that they minimize the impact of external reality such as physical disability or biological factors, in the etiology of mental illness, and sociopolitical phenom-

ena like unemployment, racism, and poverty. One result is the 'psycho-educational' approach, which has evolved in the context of the burden schizophrenia places on the family and the potential for its members to influence its course. This has led to a series of interventions including educating the family about the nature, causes, course, and treatment of schizophrenia; providing them with opportunities to discuss their difficulties in caring for the patient, and to devise appropriate strategies; and helping them to resolve conflict related to the illness, which may be aggravated by the demands of caring for a chronically ill person.

3. *The Scientific Era*

Systematic research in the psychotherapies was a low priority for many decades; instead, practitioners' interests focused on theory and technique. Investigation of the subject only took off in earnest in the early 1950s. A notable impetus was the critique by H. J. Eysenck (1952) in which he argued that the treatment of neurotic patients was no more effective than no treatment at all. Two-thirds of both groups showed improvement with time. In later reviews, Eysenck was even more damning: the effects of psychoanalysis in particular were minuscule. The attack led to much rancor and an extended battle between the psychoanalytic and behavioral camps (Eysenck featured prominently in the latter). Fortunately, a positive repercussion was the sense of challenge experienced by the analytic group. They had been riding high for several years, particularly in the US, and barely questioned whether their concepts and practice required scientific appraisal. Thus, although Eysenck's interpretation of the research literature was flawed and biased, he had stirred a hornet's nest.

Since the 1950s research has burgeoned and yielded much knowledge about whether psychotherapy works (outcome research) and how it works (process research). These developments were not without incident. Many therapists challenged the relevance of conventional research methodology to psychotherapy. A key argument was that the encounter between therapist and patient is unique, involving two people in a complex collaborative venture, and cannot be subject to the same form of scrutiny as occurs in the natural sciences. Moreover, the latter approach is necessarily mechanistic and reductionistic.

There is merit to the argument but actual research practice reveals that methodology can accommodate, at least in part, respect for the highly personal nature of psychological treatment. Sophisticated statistical procedures can also contribute to maintaining an appreciation that many characteristics in both patient and therapist are relevant, and they cannot be viewed as homogeneous groups. An illustration is the multidimensional measurement of outcome. Instead of restricting this to one or two variables only, several can be examined concurrently which together encompass the patients' internal and interpersonal world, e.g., quality of life, self-awareness, authenticity, self-esteem, target problems, and social adjustment.

Another criticism of outcome research is examining effectiveness without adequate treatment being given. Duration of research-bound therapy in fact is often much briefer than in customary practice because of funding constraints. A similar objection relates to the follow-up period usually set. A robust test of effectiveness entails examining outcome one or more years following the end of therapy in order to judge clinical projects independent of the therapist.

4. *The Effectiveness of Psychotherapy*

Several reviews have been conducted since the 1970s. A sophisticated method—meta-analysis—which relies on a rigorous statistical approach, points to psychotherapy overall generating benefits for patients who have been suitably selected. Encouragingly, better designed studies show positive results more commonly than inadequate ones. In one trail-blazing meta-analysis, levels of effectiveness were computed from 375 studies in which some 25,000 experimental and 25,000 controlled patients were tested (Smith et al. 1980). The average patient receiving psychological treatment of one kind of another was better off than 75 percent of those who were untreated—a clear demonstration of the beneficial effects of treatment in general.

With tens of 'schools' available, many of them claiming a distinctive approach, the obvious question arises as to whether some are superior in producing benefit. This question of comparative effectiveness is complicated in that not only may 'schools' be compared but also specific procedures such as setting a time limit or not or conducting therapy individually or within a group. Notwithstanding these difficulties, considering research on comparative effectiveness is worthwhile. Several reviews show a consistent pattern—that 'everyone has won and must have prizes.' This was of course the judgment of a race handed down by the dodo bird in *Alice in Wonderland*. In the psychotherapy stakes, it appears that everyone has won too, a finding probably attributable to factors common to all therapeutic approaches. These factors, set out originally by Jerome Frank (1973), include a confiding relationship with a helping person, a rationale which provides an account of the patient's problems and of the methods for remedying them, instillation of hope for change, opportunity for 'success experiences' during treatment, facilitating emotion in the patient, and providing new knowledge, so

promoting self-awareness. These 'nonspecific' factors comprise a significant theme in psychotherapy; they probably serve as a platform for benefits from all treatments.

Any consideration of effectiveness brings up the issue of harm. In other words, treatment may be for better or for worse. After a long gestation, the concept of a 'negative effect' attracted widespread attention from the 1970s (Hadley and Strupp 1976). A growing sense of confidence perhaps permitted therapists to be more open to a potential harmful impact.

To distinguish between a patient becoming worse because of their intrinsic condition or following an adverse life event, a negative effect has been defined as deterioration directly attributable to treatment itself. Such a causal link is difficult to prove but genuine negative effects certainly do occur. The definitional difficulty, however, leads to estimates of prevalence ranging from rare to common. The type of therapy appears to influence the rate. Evidence points to worsening in about 5–10 percent of cases in psychotherapy generally. The reasons for deterioration are not well established although a common view is that patient factors probably contribute, particularly selecting people for treatments for which they are unsuited. In this situation, refining assessment, thus enhancing clinicians' ability to predict response to specific treatments, is deemed as helpful to reduce the casualty rate.

Another obvious facet is the therapist and/or technique. Here, inadequate training has emerged as salient, with poor judgment leading to inappropriate interventions. It has also been recognized that a proportion of therapists perform poorly because of their personality traits, no matter what their level of training or experience. In other words, the role of therapist does not suit all those who wish to practice. In a noteworthy study of different forms of encounter group, a substantial percentage of 'casualties' were produced by four leaders, who typically pummeled members to disclose personal information and express intense emotion.

During the half century of systematic research, attention has also been devoted to the processes that take place in the therapeutic encounter.

Process and outcome research are inevitably linked. If we can identify factors that promote or hinder effectiveness, we may be able to modify these and then note the result. The advent of recording techniques such as audio and video has facilitated studies of process and a wide range of observations of both verbal and non-verbal behavior, in both therapist and patient and their interaction, have been made. One illustration is the premise that a necessary condition for effective group therapy is group cohesiveness. It follows that anything enhancing cohesiveness could be advantageous. Compatibility between members has been proposed as pertinent and indeed been shown to relate to cohesiveness. Group therapy therefore could conceivably be more effective if, in selecting members, compatibility were taken into account. For example, a patient incompatible with all his peers would presumably not be placed in that group.

Research on process has been shown to be as salient as that on outcome since only with the study of what occurs in treatment can therapists appreciate its inherent nature and the factors for optimal improvement of participants. Diligent process research has provided solid foundations for establishing hypotheses about outcome.

We have dealt in general terms with the theme of psychotherapy research. The second half of the twentieth century has been a fertile period and seen much achieved; to do it justice is beyond our remit. The interested reader is recommended to consult the four editions of the *Handbook of Psychotherapy and Behavior Change* (Bergin and Garfield 1994) which has served the field in distilling work done and providing a critique of its quality.

See also: Bowlby, John (1907–90); Freud, Sigmund (1856–1939); Janet, Pierre (1859–1947); Jung, Carl Gustav (1875–1961); Klein, Melanie (1882–1960); Psychiatry, History of; Psychoanalysis, History of; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy and Religion

Bibliography

Ackerman N W 1958 *The Psychodynamics of Family Life*. Basic Books, New York

Appignanesi L, Forrester J 1992 *Freud's Women*. Basic Books, New York

Bateson G 1972 *Steps to An Ecology of Mind*. Chandler Publishing Co., San Francisco

Bergin A E, Garfield S L (eds.) 1994 *Handbook of Psychotherapy and Behaviour Change*, 4th edn. Wiley, New York

Bion W 1961 *Experiences in Groups*. Tavistock, London

Bromberg W 1959 *The Mind of Man: A History of Psychotherapy and Psychoanalysis*. Harper, New York

Boszormenyi-Nagy I, Spark G M 1984 *Invisible Loyalities: Reciprocity in Intergenerational Family Therapy*. Brunner-Mazel, New York

Bowen M 1978 *Family Therapy in Clinical Practice*. Jason Aronson, New York

Byng-Hall J 1995 *Rewriting Family Scripts. Improvisation and Systems Change*. Guilford, New York

Ellenberger H 1970 *The Discovery of the Unconscious*. Basic Books, New York

Elliot P 1991 *From Mastery to Analysis, Theories of Gender in Psychoanalytic Feminism*. Cornell University Press, Ithaca, NY

Eysenck H J 1952 The effects of psychotherapy. *Journal of Consulting Psychology* **16**: 319–24

Foulkes S H, Anthony E J 1965 *Group Psychotherapy: The Psychoanalytic Approach*. Penguin, Harmondsworth, UK

Frank J D 1973 *Persuasion and Healing: A Comparative Study of Psychotherapy*. Johns Hopkins University Press, Baltimore, MD

Gurman A S, Kniskern D P (eds.) 1991 *Handbook of Family Therapy, Vol. II*. Brunner/Mazel, New York
Hadley S W, Strupp H H 1976 Contemporary views of negative effects in psychotherapy. *Archives of General Psychiatry* **33**: 1291–302
Mitchell S A, Black M J 1995 *Freud and Beyond: A History of Modern Psychoanalytic Thought*. Basic Books, New York
Selvini-Palzzoli M, Boscolo L, Cecchin G, Prata G 1980 Hypothesizing-circularity-neutrality: Three guidelines for the conductor of the session. *Family Process* **19**: 3–12
Smith M L, Glass G V, Miller T I 1980 *The Benefits of Psychotherapy*. Johns Hopkins University Press, Baltimore, MD
Yalom I D 1995 *The Theory and Practice of Group Psychotherapy*, 4th edn. Basic Books, New York

S. Bloch and E. Harari

Psychotherapy Integration

As an international movement, psychotherapy integration represents a revolutionary departure from the discipline's historical organization around single and competing theories. Integration is characterized by a desire to look across and beyond rival theories to see what can be learned—and how patients can benefit—from multiple ways of practicing psychotherapy. Although the movement faces several obstacles and has not yet been convincingly supported by empirical research, some form of integration will probably constitute the future of the discipline.

1. *The Historical Context*

The practice and training of psychotherapy has been historically organized around distinct and competing theoretical orientations, such as psychoanalysis, humanistic persuasions, behavior therapies, and family systems models. In the infancy of the field, therapy systems, like battling siblings, competed for attention and affection in a 'dogma eat dogma' environment. Clinicians traditionally operated from within their own particular theoretical frameworks, often to the point of being blind to alternative conceptualizations and potentially superior interventions.

Although this 'ideological cold war' may have been a necessary developmental stage for the discipline, psychotherapy since the mid-1980s has witnessed a decline in ideological struggle and the stirrings of rapprochement. The theoretical substrate of each system is undergoing intensive reappraisal, as psychotherapists acknowledge the inadequacies of any one system and the potential value of others. Distinctive of the present era is the tolerance for and assimilation of formulations that were once viewed as deviant, even heretical.

2. *Integration Defined*

As the field of psychotherapy has matured, integration has emerged as a dominating force. The dramatic growth of interest in psychotherapy integration during the 1990s has crystallized into a formal movement, or more dramatically, a 'metamorphosis' in mental health (London 1988).

Psychotherapy integration is characterized by a dissatisfaction with single-school approaches and a concomitant desire to look across and beyond school boundaries to see what can be learned—and how patients can benefit—from other ways of thinking about psychotherapy and behavior change. The ultimate outcome of doing so, not yet fully realized, is to enhance the effectiveness, efficiency, and applicability of psychotherapy.

Any number of indicators attest to the unprecedented interest in psychotherapy integration. Eclecticism or the increasingly favored term 'integration' is the modal theoretical orientation of English-speaking psychotherapists and is gradually increasing in prevalence throughout Europe. Leading psychotherapy textbooks increasingly identify their theoretical persuasion as integrative, and an integrative chapter is routinely included in compendia of treatment approaches. Handbooks on psychotherapy integration have appeared throughout the world, including several in Europe and two in the USA (Norcross and Goldfried 1992, Stricker and Gold 1993).

As with many relatively new movements, integration has been plagued by confusing vocabulary and multiple goals. Various labels are applied to the movement—convergence, rapprochement, treatment selection, combined therapies, prescriptionism, differential therapeutics, and eclecticism, among them—but *psychotherapy integration* has emerged as the overarching consensual term. Similarly, a plethora of integrations have been advanced: combining different therapy formats/modalities (such as individual and family therapy); blending self-help and psychotherapy; integrating pharmacotherapy and psychotherapy; synthesizing research and practice; integrating Occidental and Oriental perspectives; and so on. This article confines itself to the traditional and primary meaning of integration as the blending of diverse theoretical orientations or schools of psychotherapy.

3. *Reasons for Recent Growth*

Integration as a point of view has probably existed as long as philosophy and psychotherapy. In philosophy, the third-century biographer, Diogenes Laertius, referred to an eclectic school that flourished in Alexandria in the second century AD. In psychotherapy, Freud consciously struggled with the selection and integration of diverse methods.

Formal ideas on synthesizing the psychotherapies appeared in the literature as early as the 1930s (see Goldfried and Newman 1992 for a history), but it is only within the latter decade of the twentieth century that integration has developed into a clearly delineated area of interest. A number of interacting, mutually reinforcing motives have fostered the recent growth of integration (Prochaska and Norcross 1999):

(a) *Proliferation of therapies*. The field of psychotherapy has been staggered by over-choice and fragmented by future shock. Which of 400 plus therapies on the market should be studied, taught, or bought? This might also be called the Exhaustion Theory of Integration: the cynical but accurate observation that peace among warring schools is the last resort.

(b) *Inadequacy of single theories for all patients and problems*. No single system has cornered the market on utility. Underlying the ecumenical spirit is the stark realization that narrow conceptual positions do not begin to explain the evidence in psychotherapy. Clinical realities have come to demand a more flexible, if not integrative, perspective.

(c) *Ascendancy of short-term, problem-focused treatments*. The brief, problem focus of psychotherapy, particularly in the US, has brought formerly different therapies closer together. Integration responds to the pragmatic time-limited injunction of 'which therapy works better—and quicker—for this patient with this problem.'

(d) *Opportunities to observe and experiment with various treatments*. The introduction of treatment manuals, the availability of psychotherapy videotapes, and the establishment of specialized clinics for the treatment of specific disorders have all afforded exposure to other theories and therapies, and stimulated some practitioners to consider other orientations more seriously.

(e) *Recognition that commonalties contribute heavily to outcome*. The equivalent outcomes found among the empirically evaluated psychotherapies in many studies has led to increased emphasis on factors common to successful cases. Therapeutic success can best be predicted by patient characteristics and the therapeutic relationship, both transtheoretical features of psychotherapy, while only 15 percent of outcome variance is accounted for by specific techniques (Lambert 1992).

(f) *Identification of specific treatments for particular disorders*. While commonalties contribute heavily to therapeutic success, controlled research has identified several treatments of choice for particular disorders; for example, cognitive therapy and interpersonal therapy for depression, behavior therapy for specific phobias and childhood aggression, and exposure for PTSD. This has led to selective prescription of different treatments, and combinations of treatments, to individual patient needs.

(g) *Development of a professional network for integration*. Reflecting and engendering the burgeoning field have been the establishment of interdisciplinary psychotherapy organizations devoted to integration, notably the Society for the Exploration of Psychotherapy Integration (SEPI).

4. *Pathways to Integration*

There are numerous pathways toward the integration of the psychotherapies. The four most popular routes are technical eclecticism, theoretical integration, change processes, and common factors. All four attempt to look beyond the confines of single theories, but they do so in rather different ways and at different levels.

Technical eclecticism is the least theoretical of the four, but should not be construed as either atheoretical or antitheoretical (Lazarus et al. 1992). Eclecticism seeks to improve our ability to select the best treatment based primarily on the evidence of what has worked for others in the past with similar problems and similar characteristics. Proponents of technical eclecticism use procedures drawn from different sources without necessarily subscribing to the theories that spawned them (see *Eclectic Psychotherapy: Psychological Perspectives*).

In theoretical integration, two or more therapies are synthesized in the hope that the result will be better than the constituent therapies alone. This entails a commitment, as the name implies, to a conceptual or theoretical creation beyond a technical blend of methods. The various proposals to integrate psychoanalytic and behavioral theories, such as the influential work of Wachtel (1977), illustrate this path, as do the grander schemes to meld the major systems of psychotherapy, such as the transtheoretical approach of Prochaska and DiClemente (1984).

A level of abstraction down from theory is that of the change process or change strategy. This intermediate level of abstraction between global theory and specific techniques is a heuristic that implicitly guides the efforts of experienced therapists and that is increasingly guiding the teaching of psychotherapy. Instead of smushing theories fraught with epistemological incompatibilities on the one hand, or melding techniques without an overarching theory on the other, proponents of change processes suggest this intermediate level as the ideal for conceptualizing effective behavior change (Goldfried 1980).

The fourth pathway to psychotherapy integration is the common factors approach, which seeks to determine the core ingredients that different therapies share in common, with the eventual goal of creating more parsimonious and efficacious treatments based on those commonalties. This search is predicated on the belief that commonalties are more important in accounting for therapy outcome than the unique factors that differentiate among them, as eloquently maintained by Frank (1973). Literature reviews have discovered that the most consensual commonalties

across theories are the development of a therapeutic alliance, opportunity for catharsis, acquisition, and practice of new behaviors, and clients' positive expectancies (Grencavage and Norcross 1990).

Integration, as is now clear, comes in many guises and manifestations. It is neither a monolithic entity nor a single operationalized system; to refer to the integrative approach is simply a fallacy. At the same time, these multiple integrative pathways are not mutually exclusive. No technical eclectic can totally disregard theory, and no theoretical integrationist can ignore technique. Without some commonalties among different schools of therapy, theoretical integration would be impossible. And even the most ardent proponent of common factors cannot practice 'non-specifically' or 'commonly'; specific methods must be applied.

5. *Future Directions*

Psychotherapy integration is a vibrant international movement that has begun to make encouraging contributions to the field. Transtheoretical dialogue and cross-fertilization fostered by the integrative spirit have produced new ways of thinking about psychotherapy and researching behavior change. This closing section outlines five probable directions of the movement.

First and foremost, psychotherapy integration, in some guise, will likely emerge as the psychotherapeutic Zeitgeist. A dispassionate panel of psychotherapy experts, for example, portend its escalating popularity. Indeed, integrative and eclectic therapies were two of the four orientations (along with cognitive therapy and family systems therapy) predicted to expand the most in the future (Norcross and Freedheim 1992).

Second, integrationists will increasingly attend to the potential obstacles and tradeoffs of the process of integration. These will include establishing ambitious training programs and processes that ensure student competence in multiple theories and interventions, an unprecedented task in the history of psychotherapy; acknowledging the underlying epistemological and ontological differences embodied by the different theories of psychotherapies, which precludes facile integration; and addressing the language problem that confounds understanding of disparate jargons and, in some cases, leads to active avoidance of theoretical constructs.

A third direction for the integration movement is to document through outcome research the effectiveness of various integrative treatments. The empirical evidence supporting integration remains largely anecdotal and preliminary (Lambert 1992, Kazdin 1996, Wolfe and Goldfried 1988).

Fourth, psychotherapy integration will evolve past its initial step of widening therapeutic horizons and embracing multiple techniques and theories. The next step is to specifically know when and where to use these multiple techniques and theories in order to enhance therapeutic efficacy, a process also known as differential therapeutics or prescriptive matching. Psychotherapies will thus be matched to multiple client variables, including but not limited to diagnosis or problem.

In this matching process lies the fifth direction for psychotherapy integration: customizing the therapeutic relationship to the individual patient. While the historical emphasis has been on matching techniques to specific disorders, the integration movement is more and more committed to the prescriptive use of the therapeutic relationship. One way to conceptualize the issue, paralleling the notion of 'treatment of choice' in terms of techniques, is how clinicians determine the 'relationship of choice' in terms of interpersonal stances for individual clients. Again, this customization will occur on the basis of both diagnostic and cross-diagnostic considerations, such as the patient's treatment goals, arousal level, coping style, personality traits, and stage of change (Beutler and Clarkin 1990).

In this manner, the psychotherapy integration movement promises to be instrumental in customizing research-informed treatment methods and therapeutic relationships to individual patients and, in so doing, enhance the effectiveness, applicability, and efficiency of psychotherapy.

See also: Behavior Psychotherapy: Rational and Emotive; Constructivist Psychotherapies; Counseling and Psychotherapy: Ethnic and Cultural Differences; Experiential Psychotherapy; Feminist Psychotherapy; Group Psychotherapy, Clinical Psychology of; Interpersonal Psychotherapy; Person-centered Psychotherapy; Psychological Treatment Delivered Nontraditionally; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy, Brief Psychodynamic; Psychotherapy Process Research

Bibliography

Beutler L E, Clarkin J F 1990 *Systemic Treatment Selection: Toward Targeted Therapeutic Interventions*. Brunner/Mazel, New York

Frank J D 1973 *Persuasion and Healing*, rev. edn. Johns Hopkins University Press, Baltimore, MD

Goldfried M R 1980 Toward the delineation of therapeutic change principles. *American Journal of Psychotherapy* **35**: 991–9

Goldfried M R, Newman C 1992 A brief history of psychotherapy integration. In: Norcross J C, Goldfried M R (eds.) *Handbook of Psychotherapy Integration*. Basic Books, New York

Grencavage L M, Norcross J C 1990 Where are the commonalities among the therapeutic common factors? *Professional Psychology-Research and Practice* **21**: 372–8

Kazdin A E 1996 Combined and multimodal treatments in child

and adolescent psychotherapy: Issues, challenges, and research directions. *Clinical Psychology: Science and Practice* **3**: 69–100
Lambert M J 1992 Psychotherapy outcome research: Implications for integrative and eclectic therapists. In: Norcross J C, Goldfried M R (eds.) *Handbook of Psychotherapy Integration*. Basic Books, New York
Lazarus A A, Beutler L E, Norcross J C 1992 The future of technical eclecticism. *Psychotherapy* **29**: 11–20
London P 1988 Metamorphosis in psychotherapy: Slouching toward integration. *Journal of Integrative and Eclectic Psychotherapy* **7**: 3–12
Norcross J C, Freedheim D K 1992 Into the future: Retrospect and prospect in psychotherapy. In: Freedheim D K (eds.) *History of Psychotherapy*. American Psychological Association, Washington, DC
Norcross J C, Goldfried M R (eds.) 1992 *Handbook of Psychotherapy Integration*. Basic Books, New York
Prochaska J O, DiClemente C C 1984 *The Transtheoretical Approach: Crossing the Traditional Boundaries of Therapy*. Dow Jones-Irvin, Homewood, IL
Prochaska J O, Norcross J C 1999 *Systems of Psychotherapy: A Transtheoretical Analysis*, 4th edn. Brooks/Cole, Pacific Grove, CA
Stricker G, Gold J R (eds.) 1993 *Comprehensive Handbook of Psychotherapy Integration*. Plenum, New York
Wachtel P L 1977 *Psychoanalysis and Behavior Therapy: Toward an Integration*. Basic Books, New York
Wolfe B E, Goldfried M R 1988 Research on psychotherapy integration: Recommendations and conclusions from an NIMH workshop. *Journal of Consulting and Clinical Psychology* **56**: 448–51

J. C. Norcross

Psychotherapy: Legal Issues (United States)

1. The Term 'Legal Issues in Psychotherapy'

Legal issues in psychotherapy can be understood as the ways in which the law affects providers of mental health services (hereinafter, mental health professionals or MHPs, in all stages of their careers and in all aspects of their practices. This article discusses the importance of MHPs knowing the law, the laws pertaining to MHP business matters and responsibilities, and psychotherapy services that are affected by the law. While comprehensive, this article provides only a general introduction to the law from the perspective of the United States's legal system (for detail on the issues discussed herein see, Sales and Miller 2001; Sales et al. in preparation). The US experience has been seen as predictive of future trends in other countries (e.g., United Kingdom, Bell-Boule 1999). It also does not discuss the provision of forensic services, covered elsewhere in this volume (see *Forensic Psychology*; *Expert Witness and the Legal System: Psychological Aspects*; *Courtroom Consultation: Psychology*; *Rehabilitation of the Offender*; *Treatment of the Repetitive Criminal Sex Offender: United States*; *Crime and Delinquency, Prevention of*).

2. Importance of MHPs Knowing the Law

Despite the law's pervasive influence, most MHPs do not know about or understand most of the laws affecting the operation of their practices, their services and their clients. This is due in part to the difficulty of understanding legal terminology, and in accessing the many different and applicable sources of law. This state of affairs is unfortunate because MHPs may not gain the benefits that the law may provide to them or their clients, and they may not be able to fully meet their clients' needs (e.g., Rothstein 1997). In turn, legal knowledge can increase MHPs' capability for handling their clients' distress (e.g., Shuman and Smith 2000). For example, MHPs who obtain legal information are better equipped to facilitate the appointment of a guardian for an abused child or a legally incompetent person (Tor and Sales 1996), and can help correct distorted beliefs about what may happen during their legal proceedings.

Not knowing the law is also potentially harmful in an era of increasing professionalization and litigiousness (e.g., Stromberg et al. 1988). MHPs run the risk of liability for professional practice that the law designates as inappropriate (e.g., divulgence of certain professional information when the law specifies it as confidential, or for not divulging certain professional information that the law requires to be reported).

3. The Law Affecting MHPs Business Matters and Responsibilities

Before engaging in the practice of psychotherapy, MHPs need to attend to their business matters and responsibilities, including: ensuring that they are properly licensed, deciding on the form or structure of their business, and implementing procedures for dealing with financial matters and confidential client information. Some of these business issues arise during the course of a MHP's practice (e.g., release of confidential information). The law typically affects all of these business issues.

3.1 Legal Credentialing and Licensure to Practice

In the United States, specific statutory and administrative laws govern each of the licensed mental health professions (e.g., Hogan 1979). The law of credential-

ing and licensure to practice is typically organized by discipline (e.g., psychiatrists, psychiatric nurses, psychologists, subdoctoral and unlicensed psychologists, school psychologists, social workers, professional counselors, chemical dependency/substance awareness counselors, school counselors, marriage and family therapists, hypnotists, polygraph examiners).

These laws have a number of different functions. First, they establish licensing boards for MHP professions. In addition, some States have created overarching boards that have oversight of MHPs from all disciplines, in order to standardize practices across disciplines. In general, overarching and discipline-specific boards operate similarly. Second, the laws specify licensure qualifications (i.e., education, experience, testing of competence, good character), and procedures and exceptions to licensure (e.g., for those unlicensed MHPs who are supervised by a licensed MHP).

Finally, the laws regulate MHPs' conduct, and prescribe sanctions for violations. The regulation of practice generally specifies what a MHP can and cannot do. As to what MHPs can do, the law will typically define the permissible scope of practice for licensed professionals (e.g., conduct psychotherapy, dispense medications). As to what MHPs cannot do, the licensure law will describe the powers of the agency or board to investigate and adjudicate violations of the law, specify reasons for sanctioning those found in violation of the law, and provide penalties for sanctioned professionals.

3.2 Legally-defined Business Arrangements

MHPs can organize their practices in a number of different forms, many of which are subject to legal regulation (see Sales and Miller 2001). Sole proprietorships involve a MHP being the only owner of his or her mental health practice, and usually do not require a legal application for a special business form. Certain MHPs may form professional corporations that may offer certain tax, insurance, and other practice benefits but require following legal procedures.

More types of business forms exist for those MHPs who choose to work with other MHPs. When two or more people enter into a 'for-profit' agreement as co-owners, a partnership is formed. These partners agree to combine their resources in order to undertake projects that would be financially difficult for one person, and to share responsibility for each other's acts. Recent developments in the law of some States allow for limited liability partnerships or corporations, in which there are a number of limited partners or shareholders whose financial commitment and liability is limited. Finally, MHPs may form a nonprofit or not-for-profit organizations to provide mental health services, including education, research, and/or inpatient or outpatient therapy, in a community health center, private hospital, or other setting.

The structure and organization of MHPs' practices are also affected by the health care entity with which MHPs and their clients are affiliated. The law may control certain health care provider arrangements in the US (e.g., MHPs who work for managed care companies), and government-run health services in other countries (e.g., Great Britain) (see *Adolescent Vulnerability and Psychological Interventions*; *Managed Care and Psychotherapy in the United States*). For example, in the US, Health Maintenance Organizations (HMOs) provide services to patient-subscribers (an individual or group) in exchange for a predetermined fixed premium. The law encourages HMOs to provide persons with competitively priced, alternative healthcare delivery and requires HMOs to offer basic, comprehensive healthcare services. The law varies by State as to whether HMOs are also required to cover mental health services, and whether only certain professionals are allowed to provide services to HMOs.

The liability of HMOs is an area of the law currently in development, a matter of much controversy. An HMO can be sued, and possibly may be held liable, for harm caused by the acts of participating providers in the course of furnishing such services. Another potential source of liability may arise from an HMO's refusing to allow for certain services in a particular case (e.g., utilization review that result in patients not being admitted or treated, or being discharged early). This could be a particular problem for MHPs if HMO precertification reviewers refuse to approve hospitalization, and a distraught patient then assaults another or commits suicide.

3.3 Financial and Contractual Matters

There are economic and contractual components to the relationships of MHPs with their clients as well as with the entities that may pay for or reimburse them for services. These issues are raised when MHPs are forming a relationship with a client and/or third-party payor; considering which services to provide; wanting financial payment for their services; and doing their taxes.

Since the MHP–client relationship is economic and contractual as well as therapeutic, MHPs should prepare a formal written agreement in order to protect the rights and interests of both parties at the outset of services. The client has the right to receive confidential, competent, and appropriate services that are performed according to professional standards. In return, the MHP should receive the client's cooperation with the service's plan (e.g., treatment) and payment for

services rendered. The written contract, therefore, should be drafted according to local laws and clearly establish (a) the fee structure; (b) policies regarding charges and payment (e.g., charges per session, charges for missed sessions); (c) the client's commitment to paying the MHP; (d) the nature and scope of services to be provided; and (e) the general rule of confidentiality and its exceptions (see Sect. 3.4).

Such contracts are made with the informed consent of the client. Informed consent should be obtained before administering mental health services, disclosing information concerning the client to a third party, or taking any other action that may impact the client. Informed consent basically involves the process of discussions between a MHP and a potential client that leads to an informed choice by the client. In general, informed consent has three major parts: (a) *competency* of the person to be able to consent to services; (b) disclosure by the MHP of relevant and material *information* concerning the nature and risks of the proposed professional services; and (c) consent that is *voluntarily* given to the MHP. Even if not legally defined, the components of informed consent are generally agreed upon (see, e.g., Kitamura et al. 1999 for a Japanese example). The failure to obtain informed consent can lead to legal liability for malpractice.

In general, when people voluntarily contract with a MHP to receive services, they have a right to receive the services that they pay for. Conversely, competent adults have the right to refuse treatment that they voluntarily sought. Legal issues become more complicated with regard to persons who are unable to consent to treatment. In some cases, a legally competent person can execute an advance directive that identifies his or her treatment wishes if she or he becomes incompetent, and specifies whom the healthcare proxy decision-maker should be during this period. If no advance directive exists, a court can appoint a guardian (substitute decision-maker) or guardian *ad litem* (guardian over a specific issue such as making treatment decisions) after the person has been determined to be legally incompetent to make treatment decisions.

Because it is increasingly common that clients do not pay MHPs directly for services, MHPs may be reimbursed by third party payors, such as insurance agencies or government agencies. As discussed above, third party payors can legally affect the nature and length of services that MHPs can provide. However, laws are changing to prohibit both reimbursement denial for mental health services in certain cases and disparity in coverage between mental and physical health services.

Finally, the law also affects how MHPs must calculate their income and deductions to pay their taxes. In the US, MHPs are allowed to deduct expenses only once they have begun employment as a professional. For example, they may deduct expenses from subsequent educational activities or courses that: (a) are actually related to their business; (b) help the MHPs maintain or meet minimum requirements of their jobs or performance; (c) involve improving or maintaining their individual skills; or (d) are required by law (e.g., continuing education requirements for licensure). The law is unsettled regarding deductions for personal psychotherapy and psychoanalysis. The general trend is to allow MHPs who receive mental health services themselves to deduct them as educational business expenses if the MHPs have begun their practice, and if it is a necessary part of their practice (e.g., because the experience improves their job skills).

3.4 Maintaining and Disclosing Information

The doctrine of confidentiality originated in professional ethics codes (see *Psychotherapy: Ethical Issues*). It was based on the belief that effective psychotherapy required a guarantee from the therapist that no information obtained in the course of patient evaluation or treatment would be given to others without the client's consent. Thus, any written or verbal information conveyed by the client to an MHP in the course of a professional relationship can be considered a confidential communication (DeKraai and Sales 1984).

Maintaining confidential client records is not only necessary for the conduct of a good and ethical psychotherapy practice, but licensure law generally requires it. The law may specify a minimum period of time for which records must be kept, and the necessary procedures to follow to maintain the confidentiality of client records. As a general rule, MHPs should take reasonable precautions (i.e., precautions that similar MHPs in your area would take) to protect patient records. These procedures typically ensure that the records are preserved in the case of MHP relocation, separation from a group practice, retirement, or death. MHPs, who generally own their records, are also legally required to make records available to their clients upon request. However, laws may vary as to whether the request needs to be in writing, if fees can be charged for the service (e.g., for time spent copying records), and if there are limits on what type of information and in what format the requested information can be provided.

However, MHPs' duty to maintain client confidentiality is not absolute. Thus, it is important that MHPs inform clients of the limits of confidentiality at the initiation of services. Several types of penalties may be imposed for violations of confidentiality. First, as noted above, licensure law provides that the Board may suspend or revoke the license of any person betraying professional confidences, or it may impose other penalties such as probation. The same offending

actions may also subject the MHP to criminal penalties or a civil lawsuit by the client.

There are a number of threats to the privacy of psychotherapy, including exceptions to confidentiality and competing duties. The law may require MHPs to divulge confidential client information to legal authorities: (a) under search and seizure law, (b) according to a subpoena, and (c) if the client sues the MHP. Generally, confidentiality protections do not apply in legal proceedings unless a privileged communications law protects the information. Confidential information may also be disclosed to third parties if the client gives his or her written consent to a release of information.

Reporting laws impose a competing, affirmative duty on MHPs to disclose specific information obtained in therapy. These laws are based on the premise that the duty to maintain private information revealed in therapy is outweighed by society's need to protect the safety of its members, including the client. For example, confidential information may be divulged if the client presents a clear and imminent danger to himself or herself, and refuses to accept further appropriate treatment. Other laws that require confidential disclosures include dangerous person reporting laws, and child abuse reporting laws. While commentary on such reporting laws has been critical (e.g., laws are based on faulty premise that MHPs can accurately predict dangerousness; see *Clinical Psychology: Validity of Judgment*), research has not shown that the laws are antitherapeutic, and courts continue to uphold the laws.

4. *MHPs Psychotherapy Services that Are Affected by the Law*

Because MHPs provide psychotherapy services to a variety of individuals and groups with different needs in different settings, this section discusses how the law directly and indirectly affects MHPs' psychotherapy practices with certain populations.

4.1 *Psychotherapy Services for Individuals with Special Needs*

Individuals with special needs include people with mental illness, who abuse alcohol and other substances, with developmental disabilities, who receive treatment and care because they have been victims of domestic violence or other crimes, and those individuals who are incarcerated in jail or prison. Knowing the laws allows MHPs to provide clients with the services to which they are legally entitled.

First, US State law provides for the care of people with mental illness in a number of ways (e.g., Brakel et al. 1985). Some people may be eligible to receive inpatient and outpatient mental health services through their heath care plans. The law also provides for voluntary civil admission and involuntary civil commitment of mentally ill persons to state operated facilities. Generally, voluntary commitment may be permitted for those people with a variety of mental, emotional, or personality disorders, while involuntary commitment is only authorized for those people who meet a set of more restrictive standards listed in the law (e.g., Perlin 1994). The law may explicitly favor voluntary treatment over involuntary commitment because involuntary commitment generates substantial costs, and involves a loss of liberty to the person being committed. In such cases, many legally-mandated procedures must be followed to ensure that the person's rights are protected. For example, regular or long-term commitment generally requires a petition for a court-ordered commitment, a court review of the petition, an order for commitment, and an appeals process. In addition, the committed patient has numerous rights during treatment, including the rights to be free from excessive or unnecessary medication, and to not be subjected to certain procedures (e.g., experimental research; shock treatment) without informed consent, and/or a court order. MHPs can evaluate the person for admission and provide therapeutic services within the facility.

The law also provides for the voluntary treatment and involuntary commitment of those persons who are seriously disabled by alcoholism, and/or those persons with major substance abuse problems (i.e., drug-dependent persons or addicts). The law in many areas has changed from a punitive model to a continuum of care model, involving treatment and accountability (e.g., drug treatment courts). MHPs may be part of a multidisciplinary evaluation and treatment team under these laws, and may provide services (see *Relapse Prevention Training*; *Alcoholics Anonymous*; *Psychotherapy Process Research*; *Substance Abuse in Adolescents, Prevention of*) in a variety of settings (inpatient, outpatient, residential).

Third, the law recognizes that people with developmental disabilities (DD) have the capacity to be productive citizens, many of whom can achieve at least some or complete independence (e.g., Stefan 2000). Thus, some laws specifically provide for and protect rights of people with DD, including the rights to be given a nutritionally-adequate diet, and regular and sufficient medical care, and to not be physically or chemically restrained or isolated, except under strict regulations in emergency situations. Also, the law provides various inpatient and outpatient services (including evaluation and treatment by MHPs) for such persons (see *Mental Retardation: Clinical Aspects*; *Autistic Disorder: Psychological*). However, because such services are often not required, the provision of services is heavily dependent on available

funding, and upon a government's commitment to provide a certain level of services.

Fourth, some laws provide services, including those of MHPs, for victims of crimes, including the crime of domestic violence. For example, Victim Assistance Boards and Victim–Witness programs help victims put their lives back in order after the crime, and help them to testify at trial. Beyond counseling abused partners, MHPs may also become involved in these cases by identifying an abusive or battering relationship. MHPs can encourage victims to obtain legal orders to stop the abuse, and help clients to develop safety plans for their and their children's safety (see *Childhood Sexual Abuse and Risk for Adult Psychopathology*).

Finally, some laws support and require mental health services in jails and prisons because they are recognized as a vital part of the overall healthcare program for incarcerated persons. While MHPs may provide services as employees or consultants to these institutions, very little law specifically governs their activities. In some States, the law provides specialized services for sex offenders as part of their sentence, or through a more controversial, and more regulated, civil commitment treatment program. Some States require sex offender evaluation and treatment providers to be certified and to be subject to standards of professional conduct unique to the certification (see *Treatment of the Repetitive Criminal Sex Offender: United States*; *Sex Offenders, Clinical Psychology of*).

4.2 Psychotherapy Services with Children and Families

MHPs who work with children and families often find themselves courted by the legal system to aid it in administration, as when they provide evaluations and psychotherapy to court-referred juveniles and adults. Juvenile and adult clients can be referred for treatment either before trial to evaluate and/or restore their ability to participate in the process (e.g., competency to stand trial), or after trial/adjudication as part of their disposition (e.g., juvenile determined to be delinquent/in need of supervision and services) (see *Transportation, Psychology of*; *Rehabilitation of the Offender*; *Psychotherapy: Clinical Practice Guidelines*; *Conduct Disorder*). Court cases can also clarify laws and regulations regarding the psychotherapeutic, educational, and other services which children (especially those who are disabled, gifted and talented, in need of guardianship/conservatorship, or in foster care or government custody) should be receiving.

Finally, MHPs find themselves indirectly affected by the law when their child and family clients become involved in legal entanglements that involve mental status issues (see generally, DeKraai et al., 1998), such as proceedings that concern the termination of marriage or parental rights. For example, legal child custody arrangements can affect not only the therapeutic needs of the child (see, *Psychological Therapies: Emotional Processing*; *Childhood Depression*; *Clinical Assessment: Interview Methods*) and adult (see *Anxiety Disorder in Children*) family members, but also the procedures that the psychotherapist must follow (e.g., the law may determine whether the MHP can evaluate and treat children at the request of a non-custodial parent).

5. Future Legal Developments

Despite the complexity of current legal regulations relating to psychotherapeutic practice, the law will continue to grow and change for a number of reasons. First, no longer do consumers unquestioningly accept MHP recommendations about treatment, or MHP assertions about the causes for unfavorable or unexpected treatment effects. Consumers are beginning to view psychotherapeutic services as similar to many other business arrangements. They want providers to be accountable for their work, and are no longer afraid to ask the courts and legislatures to impose new legal oversight and obligations to protect consumer interests. Second, increased legal regulation is likely to occur because many societies continue to fear, or at least stigmatize (e.g., Perlin 2000), the consumers of psychotherapeutic services. These perceptions will continue to fuel increased legal supervision of mental health services, which will likely be intrusive to the psychotherapeutic process.

Third, with MHP disciplines continuing to evolve, and as consumer concerns over cost-effectiveness in health care increase, new forms of practice are emerging both within and across disciplines. The inevitable result is the need for more laws to regulate these new forms of business practice and the unexpected consequences that sometimes ensue from them. Finally, because lawmakers are constantly looking for better ways to respond to societal needs (e.g., violent juvenile crime), we will continue to see new civil and criminal laws that partially rely on the services of MHPs. Each law brings new opportunities, requirements, and obligations for practitioners. What all this means for MHPs is that there will be increasing pressure on them to become conversant with the laws that affect both their practices and the consumers of their services

See also: Health Care: Legal Issues; Socioeconomic Status and Health Care

Bibliography

Bell-Boule A 1999 Psychotherapy and the law. *International Journal of Psychotherapy* **4**(2): 193–203

Brakel S, Parry J, Weiner B 1985 *The Mentally Disabled and the Law*. The American Bar Foundation, Chicago
DeKraai M, Sales B 1984 Confidential communications of psychotherapists. *Psychotherapy* **21**: 293–318
Dekraai M, Sales B, Hall S 1998 Informed consent, confidentiality, and duty to report laws in the conduct of child therapy. In: Kratochwill T, Morris R (eds.) *The Practice of Child Therapy*, 3rd edn. Allyn and Bacon, Boston, pp. 540–59
Hogan D 1979 *The Regulation of Psychotherapists*, Vols. 1–4). Ballinger Publishing, Cambridge, MA
Kitamura T, Kitamura F, Mitsuhashi T, Ito A, Okazaki Y, Okuda N, Katoh H 1999 Image of psychiatric patients' competency to consent to treatment in Japan: I. A factor analytic study. *International Journal of Law and Psychiatry* **22**(1): 45–54
Perlin M 1994 *Law and Mental Disability*. Michie Law Publishers, Charlottesville, VA
Perlin M 2000 *The Hidden Prejudice: Mental Disability on Trial*. American Psychological Association, Washington, DC
Rothstein L 1997 *Disabilities and the Law*, 2nd edn. West Group, St. Paul, MN
Sales B D, Miller M O (series eds.) 2001 *Law and Mental Health, Professionals Series*. American Psychological Association Books, Washington, DC [A series of volumes that comprehensively reviews and integrates the law that affects mental health professionals in the United States. 〈http://www.apa.org/books/lmhptest.html〉]
Sales B, Miller M, Hall S (in preparation) *Clinical Practice Law*. American Psychological Association, Washington, DC
Shuman D, Smith A 2000 *Justice and the Prosecution of Old Crimes: Balancing Legal, Psychological, and Moral Concerns* American Psychological Association, Washington, DC
Stefan S 2000 *Unequal Rights: Discrimination Against People With Mental Disabilities and the Americans with Disabilities Act*. American Psychological Association, Washington, DC
Stromberg, C, Haggarty D, Leibenluft R, McMillian M, Mishkin B, Rubin B, Trilling H 1988 *The Psychologist's Legal Handbook*. The Council for the National Register of Health Service Providers in Psychology, Washington, DC
Tor P, Sales B 1996 Guardianship for incapacitated persons. In: Sales B, Shuman D (eds.) *Law, Mental Health, and Mental Disorder*. Brooks/Cole Publishers, Pacific Grove, CA, pp. 202–18

S. R. Hall and B. D. Sales

Psychotherapy Process Research

Psychotherapy process research is the scientific study of what happens in psychotherapy. It involves the precise, systematic, and controlled observation and analysis of the events that constitute psychotherapy. The basic elements of psychotherapy are simple—a patient, the distress or illness experienced by the patient, a therapist, and the treatment that the therapist provides—but psychotherapy occurs in a wide variety of forms and settings, involves a wide diversity of practitioners, and is conducted according to widely divergent theories about the causes and cures of psychological illness and distress. Since psychotherapy process research takes all these variations as its subject, a brief overview of the field it studies may serve as an introduction.

1. *Psychotherapy*

The forms of psychotherapy most commonly include individual therapy (one therapist with one patient), couple therapy (one therapist with two married or emotionally committed patients), family therapy (one or more therapists with members of a multigenerational family), and group therapy (one or two therapists with several patients who were previously strangers).

The settings in which psychotherapy is practiced include community mental health centers, private professional offices, university or college counseling centers, inpatient or outpatient psychiatric facilities, social work agencies, specialized schools, correctional facilities, consulting firms, and church-affiliated counseling programs.

The practitioners of psychotherapy are not members of a single profession, but instead are individuals who have typically been trained in one of the following professions: psychiatry, clinical or counseling psychology, clinical social work, psychiatric nursing, marital counseling or other counseling specialties, or the religious ministry. The level of involvement of different professions in psychotherapeutic practice, and their relative standing, varies from country to country.

Psychotherapeutic practitioners also differ in terms of the theoretical models they use to guide their work with patients. Broadly, these include analytic and psychodynamic theories, behavioral theories, cognitive and cognitive-behavioral theories, experiential and humanistic, and systemic or family-systems theory. Traditionally, these theories have formed the bases for rival 'schools' of psychotherapy. More recently, many therapists have adopted an eclectic or integrative attitude, seeking to broaden their effectiveness by combining the strengths of several models.

According to the theory followed, the treatment setting and format, and the professional identity of the therapist, these varied mental health practices may be known as psychotherapy, therapeutic counseling, psychoanalysis, behavior modification, personal training, growth facilitation, or some other term. Their practitioners may be referred to as psychotherapists, counselors, psychoanalysts, or simply by the name of the profession in which they were trained (e.g., psychologist, psychiatrist, social worker). The persons who seek help from these practitioners may be called patients, clients, counselees, or analysands. For simplicity in discussion, I refer to all those treatments as psychotherapies, to the persons who practice them as psychotherapists, and to the persons who experience them as patients.

2. Psychotherapy Research

Scientific research on the psychotherapies emerged as a field of inquiry in the middle of the twentieth century, following World War II, through the work of teams led by investigators such as C. Rogers, J. Frank, D. Shakow, M. Lorr, and others (Freedheim 1992, Chaps. 9–12, Russell 1994, Chap. 7). In this field, a distinction soon grew up between studies that focused mainly on treatment effectiveness, or the impact of therapy on patients, which were called 'outcome' research, and studies concerned primarily with the interactions of patients and therapists during the course of therapy, which were called 'process' research. Even from the outset, however, some studies included measures both of process and outcome, and in the years that followed process–outcome research has continued to be a vital focus of inquiry (Orlinsky et al. 1994, Orlinsky and Howard 1986).

Although this simple distinction between process and outcome served adequately for three decades, an increasing volume and complexity of studies and findings led eventually to the proposal of more finely differentiated frameworks for research. Based on extensive reviews of process–outcome, one of these (Orlinsky and Howard 1986) offered an overview of the field based on two distinctions: first, between therapeutic process (the observable events of psychotherapy) and its functional contexts (the social, cultural, and psychological circumstances in which therapy takes place); second, between context states that exist before and after the process events being studied, which influence and are influenced by them.

The influences of prior contextual states on events in therapy constitute determinants of therapeutic process. These would include, most immediately, the personal problems and psychological characteristics of the patient, and the therapeutic model, professional skills, and personal characteristics of the therapist. However, they also (if indirectly) include the organizational characteristics of the treatment setting, the nature and state of the mental health service delivery system, current societal pressures and stressors, and prevailing cultural standards of individual effectiveness, maturity, and well-being.

The reciprocal impacts of therapeutic events on subsequent contextual conditions comprise the consequences of therapeutic process. Researchers have been most directly concerned with the impacts of therapeutic events on the patient's life and personality (i.e., clinical 'outcome'). However, this broader framework also directs attention to the impact of therapeutic events on the professional and personal life of the therapist, and to the aggregate influence of patients' and therapists' experiences on cultural values and social institutions.

Between the contextual determinants of psychotherapy and its contextual consequences lies the set of psychological interactions and transformations that comprise the psychotherapeutic process. Researchers who would observe and analyze therapeutic process must, like photographers, first make a number of decisions about framing and focusing their studies. Framing a psychotherapy process study entails decisions about the location of the phenomenon (where to look), the temporal unit of observation (how long to look), and the observational perspectives to be employed (from which angle to look). Focusing the study involves making decisions about the facets and phases of therapeutic process to assess (which 'variables' or changes to measure).

3. Process Research Frames

3.1 Location

While therapeutic process may be defined abstractly in terms of psychological interactions and transformations, process researchers have differed over where, concretely, they can be best observed. Many researchers view therapeutic process as evidently consisting of the interactions and communications that take place between patient and therapist during the regular meetings in therapy sessions. On this view, therapeutic process includes all of the events that can be observed and recorded during therapy sessions. However, other researchers define therapeutic process in terms of the changes which patients undergo as they improve. This view directs observations primarily to the patient's experience and activity in and outside of session, and regards much that occurs during therapy sessions as therapeutically inert or inessential. To distinguish them clearly, these contrasting but somewhat overlapping concepts will be referred to as therapy session processes and therapeutic change processes, respectively. Investigators who frame their studies in terms of session process generally take a descriptive, theoretically nonpartisan, approach to defining therapy. By comparison, investigators who frame their research in terms of change processes usually hold, and seek to test, some particular theoretical conception of therapy (e.g., Dahl et al. 1988, Rice and Greenberg 1984).

3.2 Temporal Units

The concept of process refer most generally to activities and events that proceed or evolve over time. However, some activities and events proceed or evolve so rapidly, or so slowly, that they normally cannot be noticed. Examples of the former are interactions that physicists observe among subatomic particles; or, more familiarly, the rapid projection of still photographs that we do not notice separately but experience sequentially as 'moving' pictures. Examples of

the latter include the gradual cooling of romantic affairs into stable companionate relationships through years of shared routine; or, more remotely, the imperceptible erosion of mountains by millennia of wind and rain. The point is commonplace but basic and relevant.

Observations of process must be framed within the temporal span of human consciousness, but there is still a range of noticeable rates at which events occur. In observing therapeutic process, the researcher may choose to examine very brief events, like changes in facial expression or vocal intonation; or somewhat longer events, like the statement by one person of a coherent line of thought involving several sentences, or dialog between patient and therapist on a particular topic that includes successive statements by each; still longer events, like the quality of whole therapy sessions; or even longer events, like the opening, middle, and closing phases of an extended course of treatment. The temporal units or frames of reference that researchers select for study inevitably determine the kind of therapeutic events they can observe, and the descriptions of therapeutic process they give (Orlinsky et al. 1994, pp. 275–6).

3.3 Observational Perspectives

Any complex event can be viewed from several angles, and each observational perspective has the potential to reveal a different aspect of the event. In social interactions, including the interactive events of psychotherapy, the observational perspectives include those of the participants (called 'participant observers') and of immediate or remote bystanders (nonparticipant observers). In psychotherapy, there are two distinct participant observational perspectives, that of patients and that of therapists, who experience the events of therapy directly but from different vantage points. (The simplest case is that of individual therapies, which constitute the most frequent form of treatment; the same principle becomes more complex in practice when the treatment format includes more than one patient or more than one therapist, as in group or family therapies.)

To the patient's and therapist's natural observational perspectives, psychotherapy research has added that of nonparticipant observers by using clinical raters who are not part of the events being observed, and whose view of therapeutic process is generally mediated by audio or video recordings of therapy sessions or transcriptions of the verbal content of sessions. (Nonparticipant observations are sometimes also made by mechanical means, e.g., by using clocks to time speech durations.) Participant observers have a 'first hand' and 'inside' but also fleeting view of the events they experience. They have direct awareness of their own thoughts, feelings, and intentions, and immediate impressions of their own and others' behaviors, but these are typically organized around their goals in action. By contrast, nonparticipant observers just gain a 'second hand' knowledge of therapeutic process, and can only infer the covert internal states of the participants, but they have the advantage over participant observers of being able to review the recorded events many times, without the situational pressures to respond that participants experience.

Researchers may, and have, used one, two, or all of these observational perspectives in their studies, depending on their interests and the questions they investigate. However, since patients, therapists, and outside observers have different perspectives on the events of therapy, they are bound to see somewhat different aspects of process. It should not seem strange to researchers, or an indication of methodological weakness, if observers who view therapeutic phenomena from different perspectives sometimes produce discrepant observations, since those discrepancies may validly reflect the inherent complexity of social phenomena.

4. Process Research Focus

When undertaking a study of therapeutic process, researchers understand that they are dealing with a set of events that is too complex to be observed in its entirety, and must simplify the focus of the study if it is to be successful. However a study is framed in terms of session process or change process, temporal units, and observational perspectives, the researcher must focus its observations on a few specific facets of therapeutic process: the 'variables' that are to be qualitatively assessed or quantitatively measured. For the simplification to be meaningful, these variables must be defined in terms of some conceptual model of therapeutic process, and to be widely accepted by other researchers that model must be generally accepted within the research community.

4.1 Research Variables

Traditionally, the variables examined in psychotherapy process studies have been differentially defined according to whether they focus on the instrumental-adaptive (goal-directed, or task) aspect of therapist and patient behavior, or the social-emotional (expressive-interpersonal, or relational) aspect of therapist and patient behavior. Therapist instrumental or task-oriented behaviors involve the technical interventions, methods, or procedures that various clinical theories indicate therapists should employ to help patients. Research variables include the nature, timing, skillfulness, and appropriateness with which therapeutic

techniques are used. Examples of specific technical interventions in psychoanalytic therapy are use of free associations to interpret dreams, interpretation of patient resistances, and transference interpretations. A typical therapist intervention in client-centered therapy would be empathic reflection of feeling; in behavioral therapy, relaxation training to assuage phobic anxiety; in gestalt therapy, the 'two-chair' technique to explore and resolve internal conflicts; and so on. Patient instrumental-adaptive or task-oriented variables generally focus on the patient's response to technical interventions, such as cooperativeness or resistance, as well as specific effects of interventions such as enhanced insight following an interpretation, deeper self-exploration and 'experiencing' in response to therapist empathy, or decreased anxiety when facing previously frightening stimuli as a result of relaxation training. Major research methods focused on instrument aspects of therapeutic process include those described in Greenberg and Pinsof (1986, Chaps. 2–8, 14).

A second traditional type of process variable that has been studied by psychotherapy researchers is defined in terms of the social-emotional or expressive-interpersonal aspect of therapist and patient behavior, and their respective contributions to the formation and development of the therapeutic relationship (see *Therapist–Patient Relationship*). Relationship variables generally focus on the ways that patient and therapist feel and behave toward one another: the manner or style of their relating, complementary to the content of their interactions. Again, specific qualities of relationships have been emphasized by different theoretical orientations. Researchers influenced by psychoanalytic thought have focused on two issues related to the clinical concept of transference: core-conflictual themes in relationships and therapeutic alliance. Others, influenced by client-centered theory, have focused on therapist-offered relationship conditions such as genuineness, empathy, and positive regard. Social psychological theories of interpersonal behavior have provided another conceptual approach to research on relationships, emphasizing the reciprocal influences between patient and therapist, focusing on dimensions of behavior such as acceptance vs. rejection and control vs. autonomy. Useful research methods for studying therapeutic relationships are described by their proponents in Greenberg and Pinsof (1986, Chaps. 9–12, 15).

A more detailed, systematic classification that has emerged from repeated reviews of the research literature relating process to outcome (Orlinsky et al. 1994) distinguishes the following categories of process variables: therapeutic contract, therapeutic operations, therapeutic bond, participant self-relatedness, in-session impacts, and sequential flow. These are aspects of process that can be observed simultaneously in any unit of process, viewed from any observational perspective.

The *therapeutic contract* includes variables indicating the goals, tasks, and norms of patient and therapist roles, as defined by the therapist's treatment model; the number and qualifications of persons occupying each role; and the terms, setting, and schedule of their meetings. *Therapeutic operations* refer to the specific instrumental actions that patients and therapists perform in implementing the therapeutic contract, including the patient's presentation of a problematic condition (e.g., symptoms), the therapist's expert understanding of that condition (diagnosis), the therapist's ameliorative interventions (techniques), and the patient's response to those interventions (e.g., cooperation). The *therapeutic bond* is the personal connection that develops between the participants in therapy as they perform their respective roles as patient or therapist more or less in accordance with their therapeutic contract: their 'personal compatibility' (mutual feelings, and expressive attunement or rapport), and their 'teamwork capacity' (personal commitment to their roles, and interactive coordination). *Participant self-relatedness* reflects the intrapersonal aspect of patients' and therapists' involvement in the therapeutic process, including variables of self-awareness (e.g., openness vs. defensiveness), self-efficacy (e.g., confidence vs. embarrassment), and self-control (e.g., expressiveness vs. constriction). The *in-session impact* of these interdependent and reciprocally influential aspects of therapeutic process consists of the patient's 'therapeutic realizations' (e.g., hope, relief, understanding, courage) as well as the therapist's reinforcements (satisfactions and frustrations). The *sequential flow* of these process aspects within the therapy hour is indicated by variables reflecting session development, and the flow over longer periods of time is indicated by variables reflecting the course of treatment.

4.2 Research Methods

The events of psychotherapy occur in a series of meetings between patient and therapist called therapy sessions. The most directly observable aspects of those events consist of conversation. In some types of therapy, conversation may be supplemented by inviting patients to engage in certain activities or exercises designed to provide significant learning or corrective emotional experiences for the patient. Given the prominence of verbal and nonverbal communications, researchers have relied heavily on audio and video tapes of therapy sessions to record the events of therapy. Recording and replay technology have provided investigators with voluminous records for detailed examination from an external or nonparticipant observational perspective, but have also constrained them to concentrate on temporal units that have generally been briefer than whole sessions, and sometimes no longer than one complete statement or

interchange between a patient or therapist. Examples of relevant methods can be found in Greenberg and Pinsof (1986, Chaps. 2–8).

But conversation is just the outward side of therapeutic events. The inner side of process is comprised of patients' and therapists' experiences of themselves and one another during their sessions, and during the imagined conversations or fantasy interactions that they may have in the times between sessions (e.g., Greenberg and Pinsof 1986, Chap. 13). This aspect of therapeutic process has been studied mainly through the use of postsession questionnaires. These retrospective ratings utilize the participant observational perspectives of patients and therapists to provide research data, and necessarily take the whole session as the temporal unit of observation.

Chapters providing further discussion of process research methods can be found in Bergin and Garfield (1994), Greenberg and Pinsof (1986), Horvath and Greenberg (1994), Kiesler (1973), and Russell (1994).

5. *Main Process–Outcome Findings*

A half century of research on psychotherapy has cumulatively demonstrated that qualities of the therapeutic bond or relationship, observed during the course of therapy from varied observational perspectives, robustly predict treatment outcomes as assessed by patients, therapists, clinical judges, and psychometric measures, although there is still some debate about whether these relationship qualities produce that outcome or merely reflect its on-going attainment. Global ratings of relationship quality are demonstrably linked to outcome from all observational perspectives. In terms of specific relationship qualities, therapists are better able to distinguish the patient's contribution to therapeutic process, whereas patient as well as therapist process contributions are associated with outcome from the distinct observational perspectives of patients, clinical judges, and psychometric measures. The patient's personal investment, collaborativeness, expressiveness, emotional responsiveness, and positive feelings towards the therapist are particularly important, as is the therapist's empathy and credibility. This has led to a widespread view of the therapeutic bond or alliance as the most salient aspect of therapeutic process, and the 'common factor' that enables different types of psychotherapy to be more or less equivalent in effectiveness.

With regard to therapeutic operations, some types of technical interventions, when skillfully focused on patients' life problems, have also been shown persuasively to be associated with positive outcomes (e.g., interpretation, experiential confrontation, and paradoxical intention), although this too is evident from the observational perspectives of patients and clinical judges rather than therapists. In terms of participant self-relatedness, assessments of patient openness (vs. defensiveness) have been very consistently associated with outcome, from all observational perspectives. Likewise, patients' positive in-session impacts ('therapeutic realizations') have consistently been found to predict positive treatment outcomes.

6. *Process Research and Practice*

Psychotherapy process research thus far has proven itself a scientifically successful field of inquiry. Initially, it had to overcome an exclusion rule, stated by Freud (1953, p.185), that 'the "analytic situation" allows the presence of no third party. … An unauthorized listener who hit upon a chance one of them would as a rule form no useful impression; he would be in danger of not understanding what was passing between the analyst and the patient.' Thus in the earliest days, investigators were mainly concerned to demonstrate the feasibility of systematic and methodologically objective research on therapeutic process. Once that had been established (e.g., Rogers and Dymond 1954), another 15 years or so were devoted primarily to devising reliable measures and rigorous methods of process research, during which time some therapists who had taken an interest in the field were disappointed by the methodological preoccupations of researchers and the paucity of clinically applicable results (e.g., Talley et al. 1994, Chaps. 2, 4, 5). However, the rate of research on process and process–outcome patterns increased steadily from about 1970 onwards. As the 1990s approached, a considerable body of well-replicated findings had been amassed, from which implications for therapeutic practice could begin to be drawn (e.g., Talley et al. 1994, Chap. 6). Current pressures for accountability and quality assurance from managed care have forced therapists of all schools to become interested in outcome research, in order to demonstrate the effectiveness of their treatments. Yet the knowledge that most psychotherapists have of process and process–outcome research findings lags behind the times. As progress in process research continues, a major challenge will be to increase clinical awareness of the relevance and potential usefulness of these research findings.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Behavior Psychotherapy: Rational and Emotive; Behavior Therapy: Psychological Perspectives; Cognitive and Interpersonal Therapy: Psychiatric Aspects; Cognitive Therapy; Experiential Psychotherapy; Person-centered Psychotherapy; Psychological Treatment, Effectiveness of; Psychotherapy, Brief Psychodynamic; Psychotherapy: Case Study

Bibliography

Bergin A E, Garfield S L (eds.) 1994 *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York
Dahl H, Kaechele H, Thomae H (eds.) 1988 *Psychoanalytic Process Research Strategies*. Springer-Verlag, Berlin
Freedheim D K (ed.) 1992 *History of Psychotherapy: A Century of Change*. American Psychological Association Press, Washington DC
Freud S 1953 The question of lay analysis. In: Strachey A (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press and the Institute of Psychoanalysis, London, Vol. 20
Greenberg L S, Pinsof W M (eds.) 1986 *The Psychotherapeutic Process: A Research Handbook*. Guilford, New York
Horvath A O, Greenberg L S (eds.) 1994 *The Working Alliance: Theory, Research, and Practice*. Wiley, New York
Kiesler D J (ed.) 1973 *The Process of Psychotherapy: Empirical Foundations and Systems of Analysis*. Aldine, Chicago
Orlinsky D E, Grawe K, Parks B K 1994 Process and outcome in psychotherapy—noch einmal. In: Bergin A E, Garfield S L (eds.) *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York
Orlinsky D E, Howard K I 1986 Process and outcome in psychotherapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York
Rice L N, Greenberg L S 1984 *Patterns of Change: Intensive Analysis of Psychotherapy Process*. Guilford, New York
Rogers C R, Dymond R F (eds.) 1954 *Psychotherapy and Personality Change*. University of Chicago Press, Chicago
Russell R L (ed.) 1994 *Reassessing Psychotherapy Research*. Guilford, New York
Talley P F, Strupp H H, Butler S F (eds.) 1994 *Psychotherapy Research and Practice: Bridging the Gap*. Basic Books, New York

D. E. Orlinsky

Puberty, Psychosocial Correlates of

1. Introduction

This article aims to review recent research on the ways in which the biological changes of puberty influence and are influenced by the social and psychological contexts in which young people live and develop. While the biology of puberty has presumably been relatively constant since the emergence of Homo sapiens, the context in which young people are developing has changed dramatically. The evidence for these effects and interactions will be briefly reviewed, with some attention to the processes thought to be involved. It is important to note that almost all of this research is descriptive and correlational rather than experimental, although more work is being done with animal models to clarify causal explanations.

The phrase 'pubertal youth' is likely to evoke images of a group of children at least some of whom are beginning to show manifestations of physical maturity in terms of adult size and shape. Similarly, 'pubertal behavior' would conjure images related to interest in the other sex. A photograph often used in textbooks on adolescence in which somewhat taller girls are paired as dancing partners with later-developing boys depicts both 'pubertal youth' and 'pubertal behavior.' That some youngsters depicted in the photograph do not yet show physical changes of puberty demonstrates that some will engage in 'pubertal behavior' for social reasons. Further, the probabilistic nature of any complex developmental process implies that what generally occurs in a cohort of youngsters will not apply to some individuals.

The psychosocial correlates of puberty, together with the biological process of puberty itself, are usually considered to be a single biopsychosocial process (that is, a process involving biological, psychological, and social aspects) in development. In this encyclopedia, 'puberty' is used to refer to the biological aspects of the process rather than the biopsychosocial process. In this article, psychosocial correlates will be considered in two ways: (a) those that appear to influence the timing or nature of the biological process and (b) those that appear to be results or effects of the process. The term 'behavioral' will be used for both kinds of psychosocial correlates. It is assumed that influences from the social context (including structural influences) have effects through behavior.

2. Behavioral Influences on Puberty

2.1 Influences on Pubertal Timing

There is evidence that several behavioral influences affect the timing of puberty, with some altering physical appearance. These include eating behavior, exercise, stress, and sexual behavior. Starvation and other extreme stress may prevent the pubertal process from ensuing, presumably to prevent reproductive capacity in an organism unable to sustain it. Limited nutritional intake along with high physical demands on the body result in a delay of pubertal onset. For example, puberty is delayed by 1–2 years in ballet dancers, figure skaters, and gymnasts who practice intensely and restrict eating. Because prepubertal growth of the long bones continues, these youngsters may have longer limbs than would otherwise be the case (Eveleth and Tanner 1990).

Recent hypotheses have linked familial stress to acceleration or delay of puberty. For example, evolutionary theory is used to posit that children of parents under high stress will engage in behavior that will speed their exit from the family (Belsky et al. 1991). Parents under high family stress will engage in insensitive, inconsistent, and affectively negative

parental care. Young children in this circumstance then develop insecure attachments to parents and an opportunistic style of interacting with others. These children are more likely to develop behavior problems such as aggression or social withdrawal in childhood and risky sex or substance use behaviors in adolescence. While the exact process linking family relationships to pubertal timing has not been identified, it is argued that earlier pubertal timing is adaptive in this circumstance.

Delay of puberty has also been proposed to result from stressful family contexts. In this hypothesis, environmental stress increases adrenal hormones (i.e., the hypothalamic–pituitary–adrenal axis) which inhibit sex steroids (i.e., the hypothalamic–pituitary–gonadal axis) (Malo and Tremblay 1997). Research supporting this hypothesis has found that families with paternal alcoholism or with negative emotions expressed by parents were more likely also to have sons with delayed puberty. The presumptive mechanism is a stress response in the sons that interferes with hormonal mechanism for initiating puberty resulting in delayed pubertal onset.

Another line of work links prepubertal sexual abuse to earlier puberty (Trickett and Putnam 1993). The underlying mechanism may be pheromones that trigger pubertal onset, or some other triggering mechanism related to the stimulating effects of sexual contact.

This work is all relatively new and more work is needed to refine understanding of precipitating factors and mechanisms. It is possible that behavioral influences operating directly on the gonadal system accelerate pubertal onset whereas those that operate primarily through the adrenal system delay puberty. However, at this point there is no definitive research to clarify why some stressors accelerate while others delay pubertal onset.

In any case, it is clear that even timing of pubertal onset, an aspect of puberty with strong evidence of genetic influence, can be affected by behavior within the social contexts in which young people live and develop.

2.2 *Influences on Other Aspects of Puberty*

The example of exercise above is one in which effects have been seen on the adult shape that results from pubertal growth, in addition to delaying pubertal onset. Young ballet dancers who train many hours a day develop longer limbs—both arms and legs—compared with sisters and mothers. A similar effect has been observed among swimmers. While there surely are selection effects confounding the results for these elite athletes, the comparison with other family members provides some degree of control.

Although there are likely to be other examples of effects of behavior on pubertal outcomes, effects on gross aspects of physical development are easiest to attribute. Even here, though, there is the continuing confound of experience during development that makes causal attribution to pubertal change more difficult to establish. For example, different shapes surely result from different eating patterns (e.g., overeating versus undereating), but in this discussion of puberty the focus is on those outcomes that are pubertal rather than those that would occur at any point in development. As complicated as these issues become with physical attributes, they are even more complicated with psychological or social outcomes.

One effect of context on puberty should be mentioned. The increased life span and generally better health and nutrition of many people today have led to an earlier age at puberty, clearly among more developed nations in the world and increasingly among developing nations (Garn 1992). This changes the context for puberty in that it occurs in otherwise less mature organisms. Some had thought that all aspects of development might proceed in concert, so that earlier puberty would be accompanied by earlier cognitive, psychological, and social development. This does not appear to be the case, probably because of the experience or training (sometimes called socialization) needed to become a mature human. Given the existence of psychologically or socially immature adults, we should not be surprised about physically mature but otherwise immature adolescents.

The fact that puberty occurs a few years earlier now than even in the mid-1900s, and is now closer to 10 years of age among girls than to 20 years as was more typical around 1900, has tremendous significance for our understanding of and also the experience of childhood. Biological maturity permitting procreation occurs at ages much earlier than those typically considered appropriate for taking on adult work and family roles. Most societies have not adjusted to this disjunction, either to protect children from becoming parents too early or to change socialization practices and social institutions to support earlier entry into adult roles. Instead, some societies have blamed the victim in this situation, suggesting that early childbearing represents a moral failure on the part of the child or their families.

3. *Pubertal Influences on Behavior*

As with the earlier discussion of effects on puberty, pubertal timing is much more important than are other aspects of pubertal change as a factor influencing behavior. Both kinds of effects will be summarized.

3.1 *Pubertal Timing Effects*

Three hypotheses or models have been proposed for pubertal timing effects on psychosocial adjustment,

the outcomes that have been the focus of most research: the stage termination hypothesis, the deviance hypothesis, and the goodness-of-fit hypothesis (Petersen and Taylor 1980, Alsacker 1995). The stage termination hypothesis places pubertal change within the context of other developmental stages, and suggests that early puberty could cause premature termination of prepubertal development. The deviance hypothesis suggests that negative effects could accrue from being deviant for one's age and gender, as with earlier-developing girls and later-developing boys. The goodness-of-fit hypothesis posits that negative effects will result when the social context differs from the individual's developmental status, as with an aspiring dancer who may be developing on time but producing a body shape different from the prepubertal one valued for dance. There is some support for each of these hypotheses, although developmental stage has receded in importance as a psychological construct.

3.1.1 Body image. Early-developing girls tend to have poorer body images during adolescence, with increasing effects by the end of the adolescent decade, compared with girls developing later or on time relative to their peers. In contrast, early-developing boys had the most positive body image in early adolescence, with effects diminishing over adolescence (Graber et al. 1996). Although body dissatisfaction is most strongly related to short stature among boys and heavier weight among girls, preference for body types has a strong cultural component, with variations seen cross culturally (Silbereisen et al. 1989).

3.1.2 Internalizing behaviors. Early-developing girls and, to some extent, late-developing boys appear to be more vulnerable to a range of mental health indicators, including internalizing problems (e.g., depression, phobic disorders) (Graber et al. 1997). Research has also shown that pubertal timing effects can be compounded by other life changes occurring at around the same time, such as change to a new school context (Petersen et al. 1991).

3.1.3 Externalizing behaviors. Earlier pubertal timing among both boys and girls is related to more problems with externalizing behaviors (e.g., aggression, delinquency). This is a strong effect from almost all the studies of problem behavior, including substance use and abuse, with effects found on lifetime rates of substance dependence and disruptive behavior disorders (Graber et al. 1997). There is some evidence that this pubertal timing effect is especially strong when early developers begin associating with older peers.

3.1.4 Sexual behaviors. Recent research studying hormones, pubertal timing, and sexual behavior demonstrates that for boys high levels of testosterone represent a marker for early puberty, and lead to earlier sexual behavior. For girls, high levels of testosterone also lead to earlier initiation of coitus but do not seem to be markers of early puberty. This research also demonstrates that social control (e.g., religious restriction of social behavior) can moderate biological effects (Halpern et al. 1997). Many other studies find that earlier developers are more likely to be sexually active, consistent with the pattern reported above for externalizing behaviors.

3.1.5 Summary of pubertal timing effects. For both body image and internalizing behaviors, the deviance hypothesis appears to predict the pattern of results; early-developing girls and, to some extent, late-developing boys are at greater risk of problems with these outcomes. For externalizing behaviors and sexual behavior, early development in both boys and girls leads to earlier and more extensive involvement in these behaviors. To the extent that these youngsters become involved with older peers, these results fit the goodness-of-fit hypothesis, but most studies find that early pubertal onset precedes peer choice. Hence, these results seem to suggest that a different model is operating, one more directly biological, although with social moderation of effects.

3.2 Pubertal Status Effects

As noted earlier, very few effects of pubertal status have been reported in the literature. The young person's stage of pubertal development seems to have little effect on behavior. The one area that is an exception is that of parent–child relationships. As the young person becomes pubertal, there is an increase in parent–child conflict and a decrease in warm interactions peaking midpuberty (or around the time of menarche for girls) and then declining (Paikoff and Brooks-Gunn 1991). The nature of conflict is mild bickering, frequently over chores or other expectations.

4. Conclusion

As this review demonstrates, becoming pubertal has modest significance for relationships with parents but otherwise has little consistent effect on behavior. Whether one becomes pubertal earlier or later than

one's peers is of much greater significance. Deviant development relative to the group of same-aged peers—that is, early development for girls and late development for boys—is related to internalizing behaviors, or the way one feels about oneself. Early development for both boys and girls is related to problems with externalizing behavior, or the way one feels about others. The latter finding suggests that in this time of earlier puberty for youngsters generally, those who are even earlier are especially at risk. Early-developing girls are at risk for both internalizing and externalizing problems while early-developing boys are primarily at risk for externalizing problems.

In general, puberty constitutes an important change in a young person's life. The effects of developing adult size, shape, and reproductive potential are significant. Even more important to the developing individual, however, is the meaning of these changes for the society in which the young person is developing. If these biological changes signal marginality in the society, the impact on the individual is likely to be negative. In contrast, if these changes are welcomed and even celebrated, as was the case in many traditional societies in which puberty was celebrated with rites of passage into adult society, the impact on the individual is likely to be positive. Therefore, societies can choose to have more positive, constructive behavior among their young people, or they can increase the likelihood of negative and life-damaging behaviors among young people.

There is much yet to learn about effects on and effects of puberty. Recent advances in genetics and neuroscience provide tremendous opportunities to link these advances to understanding of development, especially at puberty. Equally important is research illuminating the effects of social and cultural contexts on biological processes such as puberty and the ways in which the biological process of puberty interacts with individual psychology and social context. Especially needed are studies that establish the generality of current findings, or the limits of their generalizability, with broader populations.

Policy efforts are needed to apply what we already know about puberty. There is a need for greater understanding, and recognition in policy and practice, of the fact that puberty is occurring significantly earlier now for young people than it did in the current generation of parents and grandparents. Different effects for boys and girls also require some broader attention in practice if not policy.

See also: Adolescence, Psychiatry of; Adolescence, Sociology of; Antisocial Behavior in Childhood and Adolescence; Childhood and Adolescence: Developmental Assets; Cognitive Development in Childhood and Adolescence; Identity in Childhood and Adolescence; Social Competence: Childhood and Adolescence; Socialization in Adolescence

Bibliography

Alsacker F D 1995 Timing of puberty and reactions to pubertal changes. In: Rutter M (ed.) *Psychosocial Disturbances in Young People: Challenges for Prevention*. Cambridge University Press, Cambridge, UK, pp. 37–82

Belsky J, Steinberg L, Draper P 1991 Childhood experience, interpersonal development, and reproductive strategy: An evolutionary theory of socialization. *Child Development* **62**: 647–70

Eveleth P B, Tanner J M 1990 *Worldwide Variation in Human Growth*, 2nd edn. Cambridge University Press, Cambridge, UK

Garn S M 1992 Physical growth and development. In: Friedman S B, Fisher M, Schonberg S K (eds.) *Comprehensive Adolescent Health Care*. Quality Medical Publishing, St. Louis, MO, pp. 18–23

Graber J A, Lewinsohn P M, Seeley J R, Brooks-Gunn J 1997 Is psychopathology associated with the timing of pubertal development? *Journal of the American Academy of Child and Adolescent Psychiatry* **36**: 1768–76

Graber J A, Petersen A C, Brooks-Gunn J 1996 Pubertal processes: Methods, measures, and models. In: Graber J A, Brooks-Gunn J, Petersen A C (eds.) *Transitions Through Adolescence: Interpersonal Domains and Context*. Lawrence Erlbaum, Mahwah, NJ, pp. 23–53

Halpern C T, Udry J R, Suchindran C 1997 Testosterone predicts initiation of coitus in adolescent females. *Psychosomatic Medicine* **59**: 161–71

Malo J, Tremblay R E 1997 The impact of paternal alcoholism and maternal social position on boys' school adjustment, pubertal maturation and sexual behavior: A test of two competing hypotheses. *Journal of Child Psychology and Psychiatry* **38**: 187–97

Paikoff R L, Brooks-Gunn J 1991 Do parent–child relationships change during puberty? *Psychological Bulletin* **110**: 47–66

Petersen A C, Sarigiani P A, Kennedy R E 1991 Adolescent depression: Why more girls? *Journal of Youth and Adolescence* **20**: 247–71

Petersen A C, Taylor B 1980 The biological approach to adolescence: Biological change and psychological adaptation. In: Adelson J (ed.) *Handbook of Adolescent Psychology*. Wiley, New York, pp. 117–55

Silbereisen R K, Petersen A C, Albrecht H T, Kracke B 1989 Maturational timing and the development of problem behavior: Longitudinal studies in adolescence. *Journal of Early Adolescence* **9**: 247–68

Trickett P K, Putnam F W 1993 Impact of child sexual abuse on females: Toward a developmental, psychobiological integration. *Psychological Science* **4**: 81–7

A. C. Petersen

Public Administration: Organizational Aspects

Public administration is the term used to define the formal procedural and organizational arrangements under which public employees serve a government, by

implementing and advising on policy, and managing resources. Organizational aspects refer to both the overall structures and relationships within public administrations. This could include: The organizations that make up a civil service, sometimes referred to as the machinery of government; internal organizational arrangements; and/or organizational behavior. Thus, organizational aspects can be studied in a broad sense, or within several defined fields. This contribution covers organizational aspects, widely interpreted.

Beyond public administration as a discrete body of knowledge, organizational aspects can be examined further through other theories and practices relating to, for example, political science, public policy, sociology, economics, and management. In this sense, each area of study has its own theories and concepts. Furthermore, each state has its own history, organizational form, and approach, although there are many universal, common elements, which have developed through the international transfer of ideas. New governments, formal review processes, focused research, and events have often stimulated notable change. Therefore, the area of public administration is a difficult area to research, and over the years studies have been largely descriptive, rather than experimental or empirical.

For the purpose of this article, organizational aspects of public administration are examined as four phases of development and evolution, representing important paradigm shifts, or prospective shifts. These are the premodern, modern, postmodern, and virtual/supranational phases. Critical issues relating to the theories, models, trends, and practices of organizational aspects of public administration during these times are explored.

1. *Premodern Public Administration, Organizational Aspects* (*Pre-1850s*)

The organizational aspects of public administration have been apparent from ancient times, through many civilizations, such as Greek, Roman, and Chinese dynasties, and have often been the subject of criticism and analysis. Indeed, the ancient Greek satirist and playwright Aristophanes was probably one of the earliest in the literate world to highlight the lack of accountability and fairness in relation to public office. The works of ancient philosophers, such as Plato and Aristotle, address pertinent issues of interest in this regard.

Military histories over many years from the ancient Greeks and Romans to Machiavelli, to Frederick the Great, and to Napoleon provide further insights into the evolution of organizational aspects of public administration. For example, in the 1700s, in the context of militarism and empire building, the Prussian leader, Frederick the Great, and his successor, Frederic William III, structured an exemplary public administration. Furthermore, France's public administration, which had been evolving for centuries through constitutional reform following the French Revolution of 1789, was organized on the basis of similarly functioning districts, or *départements*. In the late 1700s to early 1800s, Napoleon oversaw the central appointment of officials charged with the administrative responsibility for the *départements*. He also developed bodies which undertook administrative review, and involved the separation of juridical and public administrative powers. The civil service was made up of a number of distinct corps supported by training schools.

In the context of royalty, an early patronage system of public administration developed in England, primarily, as a support function for the royal Anglo-Saxon courts. The later structuring of a distinct, centralized public administration was related to a concept of a parliamentary cabinet. Some senior government ministers had portfolio responsibilities and a small supporting staff. Separate departments eventually existed to provide administrative functions to the ministers in charge of the Treasury, the Exchequer, the Chancellery, and the State. Accountabilities of the ministers were divided, between the Parliament and the monarch. While ministers undertook many of the administrative duties because of a general lack of suitably qualified men [*sic*], a small, elite corps of public administrators was established with a resultant continuity of tenure from government to government. To a small extent, these public administrations served a legitimate purpose, but for the most part they existed to meet the needs and ends of the administrators themselves.

Some of the public administration ideas from Europe were also adopted in New World states. In the USA, in the context of a revolutionary federal constitution which came after the War of Independence, their early patronage system was based on political connections, rather than royalty or class. Not surprisingly, these arrangements meant that the administrative organizations of government became highly corrupted, as Woodrow Wilson observed (1887). The dissatisfaction with this type of system, and a promoted interest in reform processes, led to the eventual development of professional public administrations in many states.

1.1 *Modern Public Administration, Organizational Aspects* (*1850s–1970s*)

Organizational aspects of public administration continued to evolve because of constitutional and parliamentary reform processes and, in some cases, shifts from royal courts to representative governments. The industrial revolution, the rise of capitalism, and the

new bourgeoisie all contributed to pressures for professional civil services. Thus, in modern times, organizational aspects of public administration have been based largely on the need for professionalism and have often come about because of administrative review.

In Britain, for example, the Northcote–Trevelyan Report (1853) recommended major changes in civil service recruitment and other processes. Fundamental to the changes was the establishment of a central Board of Public Administration, which eventually became a Civil Service Commission in 1870. The Board was responsible for the competitive recruitment, based on merit, of professional civil servants, and their appointment to departments; it was largely independent of government. From an organizational aspect, the idea of a central and independent agency of government responsible for recruitment and other personnel matters was instituted from that time. The best educated men [*sic*] with knowledge of history, classics, language, and mathematics were appointed to the service. Three classes of civil servant also applied: the elite administrative classes; clerical support classes; and the technical or professional classes.

In the USA, European and British theories and practices continued to influence reform of public administration, especially through the Civil Service (Pendleton) Act of 1883. However, there was a determination, as Woodrow Wilson (1887) notes, to 'Americanize' the system. The state was regarded as an instrument of social cooperation, and a more bureaucratic public administration structure was needed to manage the state. However, there was always a fundamental rejection that the senior administration could perform in a non-partisan way. This meant that there were dual streams of career civil servants and political appointees at the high levels of the civil service. Therefore, it was recognized that there was a blurred delineation between the political and public administration environments. High-level entry requirements were embedded into the career civil service, and a political education was favored as the main qualification. Taylor's (1912) theory of scientific management and related developments in the field were promoted as appropriate models of organization and work distribution.

Another major contributor to the evolution of ideas about public administration and organizational aspects was the German economist and sociologist, Max Weber. Weberian principles of bureaucracy have had a universal impact on the organizational aspects of public administration. These principles included the matching of fixed job responsibilities to defined, permanent positions; a hierarchy of official authority; the requirement for appointed, suitably qualified, and anonymous staff with continuity of tenure until retirement; recognition that high level appointments bring social status; and entitlement for a salary and a retirement pension (Weber 1946).

As an example of Weber's force, his ideas were evident as far away as Japan. Public administration and organizational aspects were based largely on German bureaucracy and Weberian principles, with other local attributes. This included the preferential acceptance of law graduates into the administrative elite stream, which has meant that the critical decision processes have been rule and precedent bound. However, the developmental role of the state has led to the dominance of certain elite governmental organizations, such as the Ministry of International Trade and Industry.

Furthermore, major events and trends, such as world wars and the growth of global trade, have impacted upon organizational aspects of modern public administration. Consequently, public sectors have increased in certain periods to address major reconstruction initiatives. Cycles of nationalization of certain industries have also impacted upon organizational aspects. Political history is important, too, and organizational change at a micro level will usually occur when new governments are elected. Thus, there is a strong link between significant history and highly dynamic organizational aspects of public administration.

Thus, public administration has evolved as a distinct area of research and study in modern times. This is supported organizationally where, in concept, there is meant to be a clear separation of powers and responsibilities between political executives (governments); public administrations (civil or public services); and, the juridical arena (the law and courts). However, there are obvious accountabilities between governments and public organizations, as well, given that there is usually a correlation between the portfolio responsibilities of members of the government and the way public administrative arrangements are determined. For example, members of a government are allocated portfolio responsibilities pertinent to particular public policy fields, such as health, defense, or education. Typically, there is a defined public administration department dedicated to the implementation of government policy and/or service provision in the corresponding policy field.

Moreover, within public administrations, organization can take a number of forms. While categorization may vary from government to government, in general terms, five types of organization may be apparent. First, there are central agencies dedicated to coordinating and supporting government effort, responsible for areas such as financial and human resource management and management improvement. Second, there are departments charged with the responsibility for direct service provision to the public, across a broad range of policy fields. Third, there are government business enterprises (GBEs) which operate at a distance from government, in a more business-like way, usually providing some essential or commercial service. Fourth, there are the review

and regulatory agencies, such as auditors-general, ombudsmen [*sic*], and anti-corruption and whistle-blowing protection agencies. Fifth, there are the more peripheral boards and agencies, often semi-autonomous entities that conduct other, sometimes more obscure, aspects of government business.

In general, these organizational aspects still apply, but the public administration–bureaucratic paradigm which endured for over a hundred years has to a significant extent been rejected by governments in postmodern times. By the 1960s, especially in Anglo-American polities, it was apparent that the modern paradigm of public administration, as it had evolved, was no longer serving governments, as intended. Public administrations were seen by critics and reviewers to be too big, too inefficient, and too powerful in comparison to governments. These criticisms were also linked to the perceived failure of Keynesian economic principles, which had underpinned governments' strategic policy decisions since the 1930s.

1.2 Postmodern Public Administration, Organizational Aspects (1979–Present)

It took until the late 1970s and early 1980s for fundamental changes to be implemented in a major paradigm shift toward a market-based model of public administration. Conservative political leaders, Prime Minister Thatcher in Britain and President Reagan in the USA were the first politicians demonstrably to sustain such an approach, which led to adoption in other Anglo-Saxon polities. Similar earlier attempted changes had been evident in Latin America, where free-market concepts of the Chicago School were influential for a relatively short time. The market-based model, in varying forms, has been adopted in over 100 countries, worldwide, with some convergence of ideological thought between traditionally conflicting political parties. The market-based model supplanted communism in Eastern Europe in the late 1980s as a guiding ideology and has led to revolutionary change in organizational form in other largely totalitarian states.

Supranational governance structures, involving organizations such as the International Monetary Fund (IMF), the World Bank, the Organization for Economic Cooperation and Development (OECD), the Group of 7–8 leading trading nations, the World Trade Organization (WTO), and the European Union (EU) have all supported and promoted the market-based model. This may account for the active transfer and adoption processes in place from polity to polity. The EU, as the most comprehensive organizational form for supranational governance, operates with its own parliament and public administration system.

Several common themes based partly on public choice, and other less well defined neoclassical economic theories of rational choice are characteristic of the market-based model. They all have implications for public administration and organizational aspects. First, for example, there is a move away from a mixed economy to a firm belief in markets as the primary and more efficient economic mechanism. To this end, deregulation, relating to trade, finance, and monetary policies has taken place.

Second, if markets are to prevail, then the size of governments and their public sectors needs to be reduced. Included in the radical downsizing of public sectors is commercialization, and/or the privatization of GBEs where it is deemed inappropriate for governments to continue to operate in these areas. Major utilities and commercial businesses traditionally involving governments, in areas of activity such as telecommunications, power generation, distribution and transmission, water, banking, insurance, postal services, and airlines may be targeted.

Third, and most importantly, for the public sector, the strategic focus is on organizational efficiency, the creation of internal market-style competitive conditions and the more purposive application of private-sector business techniques to public management. These changes have involved a shift from a focus on processes of administration and policy, however flawed, to organizational management. A business, or corporation, metaphor may be used to define organizational form.

As a subset of public choice theory, principal–agent theory focuses on the relationships between governments and their top bureaucrats. Public bureaucrats are perceived to be budget-maximizing, self-interested agents, not trusted by governments. Therefore, governments need to develop executive structures where principal–agent theory can be applied directly so that governments can compel their administrative elite to act in the government's interests. The establishment of formal senior executive services, with performance incentives and fixed-term contracts may result.

Organizationally, the acceptance of the market-based model by governments has seen other major changes. These include, first, substantial machinery of government changes including the separation of policy, regulatory, commercial, and operational services and programs into discrete organizations. Second, the purchaser/provider split, where governments are the contract managers and service provision is contracted out. Third, more openly political appointments at the top of the public service are apparent in previously nonpartisan services. Frequently, this has involved the centralization of power away from independent personnel bodies, which have been abolished. These senior executive managers are expected to support publicly governments' electability. In such circumstances, the possibility of ministers accepting responsibility for departmental managerial failure, within the construct of ministerial responsibility, considerably diminishes.

The rational choice logic, which underpins the market-based model, is open to question, given the

presupposition of perfect markets and exchange processes. Such a conceptual framework is unrealistic for the political environment of public administration. Rewards for certain actions and behaviors of agents, required by the principals, can be achieved through both formal and informal power structures. This is an important organizational aspect of the market-based model, where parallel systems of exchange, reward, and relationships operate as a matter of course. While the formal systems may promote a constancy of action and process, the informal system can support implicitly institutionalized power relationships.

Whether the paradigm shift has resulted in better public administration and organizational aspects is still to be determined as the key economic indicators of performance, predicated within this new paradigm, suggest a mixed result. There has certainly been a shift from idiosyncratic or confined interpretations of public administration, relevant to particular nations, regions or systems, to a more universal construct. Nevertheless, there are many variations on the market-based theme depending on governments in power and individual leaderships at any given time.

From an organizational aspect, one of the weaknesses of the market-based model is that it focuses strongly on economic issues and individualism, rather than on the collective responsibility of the state to its citizens. In this sense, social and related policy fields may be diminished in their importance. Some governments are attempting to redress this imbalance with more attention to social, environmental, and other imperatives, relating to the development of human and social capital. Furthermore, the continuing growth of supranational structures in terms of greater organizational form, and advances in information technology, now provide new challenges for governments, and another likely significant paradigm shift. The organizational shape and size of public administrations into the future remains uncertain.

1.3 Virtual and Supranational Public Administration, Public Administration (Now and into the Future)

Two developing areas of interest will undoubtedly have major organizational impacts on public administration in the future. These are, first, more virtual governments, and, second, the power of supranational organizations. With developments in information technology and the related rise of the global new economy, governments are adding more virtual organizational forms to public administrations. For example, especially with customer service initiatives, knowledge management becomes increasingly important. Free information lines, supported by computer technology, provide coordinated details of government services. Similarly, one-stop shops allow citizens to attend to a range of government requirements, such as license renewals and the payment of bills, at a single location. Comprehensive information can be obtained from government Internet websites, and civil servants can be contacted, more informally and directly, by e-mail. E-commerce, relating to procurement, contracting, and the lodging of documents may now apply. Within governments, more virtual communication also takes place involving typed, aural, and visual media. Undoubtedly, this will continue to impact upon organizational aspects of public administration in the future.

Second, the continuing development of supranational governance structures, in the context of globalization, has led to both cogent and conspiratorial debate in recent years about future organizational aspects of public administration. The Parliament of the EU as a structure challenging the sovereignty of national member governments, is an example of how future organizational aspects may evolve as other trading coalitions consolidate their form. However, responses to globalization, especially negative ones, also prompt support for localism, which emphasizes the role of individual communities.

With all these complex local and global pressures, it is likely that organizational aspects of public administration will experience quite dramatic change in the first decades of the twenty-first century. Whether the more traditional structures of agencies, departments, trading enterprises, and review bodies, with the complex relationships between key actors, will remain as the primary organizational aspects is difficult to predict. Even if they do remain, the more virtual aspects of public administration are expected to endure and grow, to the point where there will be far more virtual public administration than there is at the time of writing.

See also: Administration in Organizations; Bureaucracy and Bureaucratization; Civil Service; Organization: Overview; Organizations and the Law; Organizations: Authority and Power; Organizations, Sociology of; Parliamentary Government; Planning, Administrative Organization of; Public Administration, Politics of; Public Bureaucracies; Public Sector Organizations; Weber, Max (1864–1920)

Bibliography

Aucoin P 1995 *The New Public Management CANADA in Comparative Perspective*. Institute for Research on Public Policy, Ottawa, Canada

Barberis P (ed.) *The Civil Service in an Era of Change*. Dartmouth, Aldershot, UK

Barratt Brown M 1995 *Models in Political Economy*. Penguin Books, London

Beloff M 1959 *The American Federal Government*. Oxford University Press, London

Boston J, Martin J, Pallot J, Walsh P (eds.) 1991 *Reshaping the State*. Oxford University Press, London

Campbell C, Wilson G 1995 *The End of Whitehall*. Blackwell, Oxford, UK
Clegg S 1989 *Frameworks of Power*. Sage, London
Eldridge C C 1978 *Victorian Imperialism*. Hodder & Stoughton, London
Galbraith J K 1994 *The World Economy Since the Wars*. Mandarin, London
Halligan J, Power J 1991 *Political Management in the 1990s*. Oxford University Press, Melbourne, Australia
Hood A 1998 Public officials, government and the public trust. *Australian Journal of Public Administration*. **57**(1), March: 98–114
Hughes O 1994 *Public Management and Administration*. St. Martin's Press, New York
Hutton W 1995 *The State We're In*. Jonathan Cape, London
Institut Internationale d'Administration Publique 1996 An introduction to French administration. La Documentation française. Institut Internationale d'Administration Publique, Paris
Kingdom J 1991 *Government and Politics in Britain*. Polity Press, Cambridge, UK
Koh B 1989 *Japan's Administrative Elite*. University of California Press, Berkeley, CA
Laffin M, Painter M (eds.) 1995 *Reform and Reversal: Lessons from the Coalition Government in New South Wales 1988–1995*. Macmillan, South Melbourne, Australia
Lynn L E Jr 1996 *Public Management as Art, Science and Profession*. Chatham House, NJ
Northcote–Trevelyan Report 1853 *Report on the Organisation of the Permanent Civil Service*. Appendix B in Fulton, Lord (Committee under the Chairmanship of) 1968 *The Civil Service. Vol. 1*. HMSO, London
Ormerod P 1994 *The Death of Economics*. Faber and Faber, London
Peters B G 1996 *The Future of Governing: Four Emerging Models*. University Press of Kansas, St Lawrence
Pollitt C 1993 *Managerialism and the Public Services*, 2nd edn. Blackwell, Oxford, UK
Pusey M 1991 *Economic Rationalism in Canberra*. Cambridge University Press, Cambridge
Rainey H 1991 *Understanding Public Organizations*. Jossey-Bass, San Francisco, CA
Savoie D 1994 *Thatcher, Reagan, Mulroney*. University of Toronto Press, Toronto, Canada
Self P 1993 *Government by the Market?* Macmillan, London
Shafritz J, Hyde A 1992 *Classics of Public Administration*. Brooks/Cole, Pacific Grove, CA
Spann R N 1979 *Government Administration in Australia*. Allen & Unwin, Sydney, Australia
Sturgess G 1993 Can parliament cope? Towards a new era in public accountability, *Canberra Bulletin of Public Administration*, **7**, September: 45–52
Taylor F 1912 *Scientific Management* (Reprinted in Shafritz J, Hyde A 1992 *Classics of Public Administration*). Brooks/Cole, Pacific Grove, CA
Wanna J, O'Faircheallaigh C, Weller P 1992 *Public Sector Management in Australia*. Macmillan, South Melbourne, Australia
Weber M 1946 *Bureaucracy*. Reprinted in Shafritz J, Hyde A 1992 *Classics of Public Administration*. Brooks/Cole, Pacific Grove, CA
Wilenski P 1986 *Public Power and Public Administration*. Hale and Ironmonger, Sydney, Australia
Wilson W 1887 *The Study of Administration*. Reprinted in Shafritz J, Hyde A 1992 *Classics of Public Administration*. Brooks/Cole, Pacific Grove, CA

J. Johnston

Public Administration, Politics of

From its self-conscious beginnings in the late nineteenth century until the years around World War II, the practice of public administration in America, conceived then, as now, as the organization of government and the management of its activities, was regarded as a rational, technical enterprise properly pursued entirely apart from the give and take of politics. Working against the backdrop of civil service reform, those who sought initially to define the field in both a descriptive and prescriptive sense were certain that public administration could only be accomplished by strict adherence to principles of neutral competence and efficiency rather than partisanship, compromise, and preference. In this view, representative political institutions, whose function is to weigh competing interests and make choices for the society, set the tasks of the bureaucracy. In his seminal essay of 1887, *The Study of Administration*, Woodrow Wilson wrote of the possibility of a 'science of administration,' maintaining that 'administrative questions are not political questions' (Wilson 1941). In the crucible of the Great Depression of the 1930s and the war that followed, however, challenges to the politics–administration dichotomy began to take form, and by 1950 John Gaus (1950) could declare that 'a theory of public administration means in our time a theory of politics also.' Today, there is no controversy: no one would deny that public administration is a decidedly political process.

1. How Public Administration is Political

There are two general ways in which public administration may be regarded as a political process. One is that by their actions public administrators make value choices that allocate benefits and costs authoritatively; these choices are likely therefore to be contested. The other is that public administrators, far from being passive agents of representative political institutions, act strategically, seeking to maximize the influence of their agencies by competing with other actors for resources and by building alliances.

1.1 Administration as the Pursuit of Values

The organization of government and the implementation and evaluation of public policy are carried out

in the context of certain normative standards or values that guide both management strategies as well as day-to-day actions. The origin of these standards lies normally with the legislative or judicial branches of government or the elected executive. They include, among others, injunctions to pursue the business of government with economy, efficiency, equity, responsiveness, transparency, and accountability.

Many of these standards are in conflict with one another, however: economy and equity may not be attainable simultaneously, for example, if the former stresses cost-savings at the expense of broad distribution of benefits. Nor is efficiency, which may require swift and confidential action, always compatible with transparency, which may entail open decision-making and citizen participation. Furthermore, each of these standards is susceptible to interpretation as it relates to any given concrete situation. Since representative political institutions may not provide specific guidance about priorities or offer ways to resolve ambiguities as these standards are applied in the field, public administrators must necessarily make choices and render interpretations that inevitably create advantages for some interests and deny them to others. Thus, as Carl Friedrich (quoted in Lynn 1996) once observed, 'Public policy is being formed as it is being executed … In so far as particular individuals or groups are gaining or losing power or control in a given area, there is politics.' To the extent that politics is the process of deciding who gets what, then public administration is a political activity.

1.2 Public Administrators as Strategic Actors

Through the aggrandizement of resources and the solidification of alliances, managers of bureaucratic agencies strengthen their ability to exercise influence in the shaping and implementation of public policy. Selznick's study of the Tennessee Valley Authority (1949) was among the first to suggest the importance of establishing external allies (such as client groups, interest groups, the news media, or members of Congress) as a key to the survival and prosperity of a bureaucratic agency. Subsequent observers have seen the efforts of bureaucratic agencies to build a resource base and foster external support as creating new competitors for power with legislative bodies in government (Rourke 1987). Another perspective, however, has emphasized the mutually reinforcing relationships that bureaucratic agencies at the national level tend to forge with interest groups and the legislative committees critical to their funding and authority. These relationships are variously referred to as 'iron triangles' or, to suggest the greater fluidity with which these alliances evolve, 'issue networks' (Heclo 1979).

2. The Public Management Perspective

The political nature of public administration has taken on sharper focus with the emergence in the last quarter of the twentieth century of the subdiscipline of public management. Whereas traditional public administration was concerned primarily with the organization, design, and operation of bureaucratic institutions, public management is more interested in how managers actually manage public policy (Lynn 1996). Kettl (1996) attributes the rise of public management as a field of study and practice in part to research on policy implementation that began in the 1970s. These investigations suggested that policy success and failure lay not so much with the laws that established programs, or even with the design of the organizations responsible for implementation, as with the skills and capacities of the managers. Some were more politically capable than others. Lynn (1996), for example, writes of 'the politics of management,' by which he means the ability of managers of public agencies to set a course, maintain credibility with overseers, marshal authority and resources, and position the organization advantageously in the political environment. Thus, the public management perspective focuses explicitly on strategy and interorganizational relations rather than the administration of processes internal to bureaucratic agencies.

3. The Politics of Public Administration in Action

The politics of public administration are revealed in the routine processes of government organization and operations, such as reform, personnel administration, and policy implementation.

3.1 Reform

Reformers have made repeated attempts during the twentieth century to transform American government to make it more effective and more responsive to public desires. These efforts have often focused on reorganization schemes whose effect was to redistribute power within government. For example, the creation of an executive budget at the national level in the United States in the 1920s and the establishment of an Executive Office of the President in the postwar period both served to strengthen the presidency in relation to Congress. In more recent times, the movement to 'reinvent' government has sought to shift power both from Congress to the bureaucracy and from the higher reaches of the bureaucracy to mid-level managers and line functionaries (Kettl 1998). Not only has the reinvention movement aimed at 'empowering' government functionaries at the lower rungs of the bureaucracy, but it has also attempted to

recast citizens as customers rather than supplicants. Since customers may exercise their right not to buy, to 'exit,' the successful market actor must be attentive to their desires. The reinvention movement at the end of the twentieth century is, thus, partly an effort to transform the culture of public administration by conferring on citizens seeking service from government rights and powers—respect, attentiveness, quality—similar to those enjoyed by customers in the private market (Osborne and Gaebler 1992).

3.2 Personnel Administration

Hiring public employees, defining job requirements, motivating workers to perform, and appraising their achievements are standard tasks of the public administrator. As Levine et al. (1990) have pointed out, public executives seek to achieve a diverse set of values in personnel administration, including instrumental achievement, social equity, merit, and employee rights. Often these may be in conflict with one another. This is nowhere more apparent than in the debate over affirmative action and descriptive representation in the public workforce.

Holden (1996) has observed that no theory of public administration can fail to take account of competition among ethnic groups in a plural society. Public employment has long been a prize over which ethnic and racial groups have contended, for work in government has provided not only a measure of economic security but also advancement. The need to respond to the political demands of groups underrepresented in public jobs may be seen as clashing at times with the exercise of strict merit standards. Public administrators face the dilemma of resolving this value conflict as they seek to manage their agencies, posing clear political choices. Elected officials also struggle with this issue, and those who resolve it by committing to affirmative action illustrate how political considerations may merge with and shape the public administration of personnel management.

3.3 Policy Implementation

One of the most important factors accounting for the political nature of the administrative process is the degree to which government functionaries necessarily exercise discretion in implementing laws and administrative directives. When state highway department planners choose to establish a right-of-way for a new road through a poor neighborhood rather than a middle-class area, or decide to locate an interchange near one suburban commercial development rather than another, it is clear that the implementation of highway policy is not simply a matter of the application of engineering and administrative considerations; management choices have distributive consequences. They create winners and losers.

Discretion may be more or less circumscribed. Schneider and Ingram (1997) point out that the allocation of discretion in any given policy area varies from the situation in which the laws and rules establish unambiguous goals, means, and target populations to a situation in which the law is so general as to leave all of these in great measure to the judgment of government employees in the field. At the lower levels of the organization so-called 'street-level bureaucrats' (Lipsky 1980), such as police officers, welfare case workers, health and safety inspectors, and other government employees who deal on a face-to-face basis with the public, often labor under conditions that conform more to the latter end of the spectrum than the former. The choices of these bureaucrats are fraught with political meaning: police officers who stop automobiles driven by young black men more often than those driven by whites simply in the belief that blacks are more likely to be criminals, or caseworkers whose sympathies lead them to facilitate applications by elderly clients for food stamps but not for single mothers, serve as key actors in the allocation of policy benefits and costs.

Discretion may even lead to the sabotage of laws with which bureaucratic officials do not agree. For example, in the mid-1970s the secretary of the US Department of Agriculture resisted implementing the Women, Infants and Children supplemental food program, claiming that such assistance was both duplicative of other programs and ineffective. Delays in the implementation of the program led to congressional claims that the department was breaking the law.

4. The Administration–Politics Synthesis

To reformers in the early twentieth century, for whom the business corporation provided the preferred template for the administrative state, politics was understood as the competitive struggle for partisan advantage. This struggle served its purpose in the selection of leaders and the adoption of policy, but in an ideal society, administration, the carrying out of political decisions as well as the housekeeping functions of government, such as street paving, would be separated from partisan conflict and placed in the hands of disinterested professionals, beholden neither to party nor politician, but to rational principles of management. Yet in the century since these ideas took form, students of the administrative process came to see that the operational details of organization and management were themselves deeply political.

Public administration has come to be understood as the continuation of the political process—the authoritative allocation of values—by government actors

in government settings other than the elected institutions. To the extent that public administrators make choices in organizing their bureaus, in choosing their personnel, in establishing operational priorities, and most important, in applying the laws, they distribute costs and benefits across groups and interests in the society. In that sense, public administration is a political activity.

See also: Accountability: Political; Administration in Organizations; Bureaucracy and Bureaucratization; Management Accounting; Organizations, Sociology of; Planning, Administrative Organization of; Public Administration: Organizational Aspects; Public Bureaucracies; Public Sector Organizations

Bibliography

Gaus J 1950 Trends in the theory of public administration. *Public Administration Review* **10**: 161–68

Heclo H 1979 Issue networks and the executive establishment. In: King A (ed.) *The New American Political System*. American Enterprise Institute, Washington, DC

Holden M 1996 *Continuity & Disruption*. University of Pittsburgh Press, Pittsburgh, PA

Kettl D 1996 Introduction. In: Kettl D, Milward H B (eds.) *The State of Public Management*. Johns Hopkins University Press, Baltimore, MD

Kettl D 1998 *Reinventing Government: A Fifth Year Report Card*. Brookings Institution, Washington, DC

Levine C, Peters B G, Thompson F 1990 *Public Administration: Challenges, Choices, Consequences*. Scott, Foresman, Glenview, IL

Lipsky M 1980 *Street Level Bureaucracy*. Russell Sage, New York

Lynn L 1996 *Public Management as Art, Science and Profession*. Chatham House, Chatham, NJ

Osborne D, Gaebler N 1992 *Reinventing Government*. Addison Wesley, Reading, MA

Rourke F 1987 Bureaucracy in the American constitutional order. *Political Science Quarterly* **102**: 217–32

Schneider A, Ingram H 1997 *Policy Design for Democracy*. University of Kansas Press, Lawrence, KS

Selznick P 1949 *TVA and the Grassroots*. University of California Press, Berkeley, CA

Wilson W 1941 The study of administration. *Political Science Quarterly* **56**: 481–506

P. Eisinger

Public Broadcasting

The history of broadcasting, beginning in the early 1920s, has seen three main types or models of broadcasting organizations. One is the commercial model, in which advertising is used to provide revenues, with programs seen as vehicles for providing an audience to which products can be sold. A second model is one in which the broadcasting organization is funded by the state, usually so that the state can project its political message and thus influence the public. The third model is known in the USA as public broadcasting but in most of the rest of the world as public service broadcasting. The most famous and influential example of this model is the BBC. It is this model with which this article is concerned, particularly by defining it in ways that will allow the reader to see the fundamental difference between it and the other two models.

1. The General Concept and Background of Public Broadcasting

In a public system, television and radio producers acquire money to make programs. In a commercial system they make programs to acquire money. However simple, this little epithet articulates the divergence of basic principles, the different philosophical assumptions, on which broadcasting is built. History and experience fashioned inside public broadcasting a definable canon, a set of principles and practices which constitute the core theses around which the institution has been formed and which have guided its performance.

The particular institutional structures and forms of funding may vary from place to place, but public broadcasting is above all else an ambition, a belief that the sheer presence of broadcasting within all our lives can and must be used to nurture society, to proffer the opportunity that society and its inhabitants can be better served than by systems which primarily seek consumers for advertisers or subjects for the state.

Some of the most powerful visions of the purpose of broadcasting emerged within unusual and trying circumstances. Consider, for instance, the cultural histories of the occupations of Germany and Japan in the late 1940s and the formulation of Allied policy for broadcasting in the rebuilding of those societies. Here was testament to the idea of broadcasting as primarily a social rather than an economic process, as something with moral, cultural, intellectual, and creative purpose and not just as a source of mild comment and moderate pleasure. The charters that established broadcasting in Japan and in Germany, by the Allied occupying powers, were replete with the public service ideal. If broadcasting was to comment, it should do so with a flourish. If it was to amuse, it should do so with elan. If it was to educate, it should do so with real professionalism. It was believed by the Allied leadership, including the USA, that the life of the mind of a society was far too precious and important to be left to what they took to be the vagaries of a commercial system or the oppressive force of state control.

It could be argued that such policies were creatures of the moment, as massive destruction demanded enormous reconstruction, of which communications would inevitably be part. But what was required was the restoration not just of highways, buildings, plants, but also of the shattered imaginative lives of whole populations. The architects of postwar Germany and Japan sensed correctly that healthy, diverse cultural institutions were a prerequisite to a functioning liberal democracy. Broadcasting was thus to be used as a key part of the cultural and social regeneration of those societies. In that lies the real clue to the nature and purpose of great public broadcasting: that it makes best sense when it represents a national and moral optimism within a society, when it suggests through the diversity and quality of its programs that societies be better than they are, better served, better amused, better informed, and, thus, better societies.

A period which is widely regarded within the advanced industrial societies as a high water mark of public service broadcasting, is the BBC in the early 1960s. A key figure from those years was Sir Arthur fforde, possibly the greatest of the chairmen of the BBC. In 1963 he wrote: 'By its nature broadcasting must be in a constant and sensitive relationship with the moral condition of society.' He felt that the moral establishment had failed modern society and that broadcasting was a way in which that failure could be rectified. He added that it 'is of cardinal importance that everyone in a position of responsibility should be ready to set himself or herself the duty of assuring, to those creative members of staff … that measure of freedom, independence and elan without which the arts do not flourish' (quoted in Tracey 1983, p. 235).

That idea of providing a protective layer within which the imaginative spirit might create, lay at the heart of the BBC version of public service broadcasting which flourished in the postwar years and which constituted a model that other nations sought to emulate. Ian Jacob, Director-General of the BBC from 1952 to 1959, refined the notion. In 1958, in an internal document called 'Basic Propositions', he described public service broadcasting as:

> … a compound of a system of control, an attitude of mind, and an aim, which if successfully achieved results in a service which cannot be given by any other means. The system of control is full independence, or the maximum degree of independence that Parliament will accord. The attitude of mind is an intelligent one capable of attracting to the service the highest quality of character and intellect. The aim is to give the best and the most comprehensive service of broadcasting to the public that is possible. The motive that underlies the whole operation is a vital factor; it must not be vitiated by political or commercial consideration.

This is one of the best attempts to capture in words a concept and view of broadcasting which continues to circulate in the world of cultural politics. Yet even here the vision, the articulation, is limited. Jacob's words imply that we understand the nature of public service broadcasting not by defining it, but by recognizing its results, rather as one plots the presence of a hidden planet or a subatomic particle not by 'seeing' it, but by measuring the effects of its presence. The Pilkington Committee, on the future of British broadcasting, said as much in 1962 when it observed: 'though its standards exist and are recognizable, broadcasting is more nearly an art than an exact science. It deals in tastes and values and is not precisely definable' (Home Office 1962, p.13, para. 34).

One obvious characteristic of efforts to define public service broadcasting is the elegant vagueness that is studied and deliberate. In effect what is suggested is that one thing that has to be understood about defining public service broadcasting is that in any abstract sense one cannot.

The beast that lurks in the shrubbery of those discussions is that whatever the definitional uncertainties, in that public service broadcasting can be experienced and recognized but never properly captured by language, someone has to decide on what is 'good' and 'bad,' and there should be a guiding hand, what has been referred to as 'the custodians, the caretakers' of culture. As traditionally understood by public broadcasters the whole idea of custodianship was totally unproblematic, as was understanding just who the custodians were, the characteristics they would possess and the locations in which they would be found. Such presuppositions appeared only natural throughout the history of broadcasting because that history was embedded within a social order in which hierarchy was also assumed: hierarchies of social status and cultural judgment. And in a curious kind of way the point of hierarchy is to reproduce itself, since the fundamental belief of the hierarchical system is, and has to be, that such arrangements have worth and merit. The plausibility of such a thesis becomes very much dependent on delivering evidence that attests to both worthiness and meritoriousness in such a way as to drown out the noise of any emergent countervailing thesis.

A key justification for the custodial role in most societies where public service broadcasting was established was that since the radio frequency used for transmissions was a limited natural resource, someone had to ensure that its use served the public good, the whole community. The cultural geology of this decision had, however, a deeper level to it, based on nineteenth century assumptions about the ways in which the arts and humanities could elevate the human condition. In an essay on the formation of modern literary analysis, Steiner wrote that behind the study of literature in the nineteenth century lay:

> a kind of rational and moral optimism … a large hope, a great positivism … The study of literature was assumed to carry an almost necessary implication of moral force. It was thought self-evident that the teaching and reading of the great poets and prose writers would enrich not only taste or style

> but moral feeling: that it would cultivate human judgment and act against barbarism (Steiner 1985, pp. 77–8).

Reading this account one could quite properly substitute the word 'broadcasting' for 'literature' and have a powerful explanation of what the creation of the model of public service broadcasting was all about: a relocation of a nineteenth century humanistic dream of human betterment. The fear that drove that dream was of 'the mob,' the pervasive belief among cultural, religious, and political elites that there was indeed a dark side to the human soul that was, when let loose, dangerous and devastating to the flesh as well as the spirit. Whatever the apparent elitism, who is to say that they were wrong nestling, as they did, between the first great war and a looming second. There remained, however, a residual faith, tied to the whole condition of the Enlightenment, humanism, and belief in progress, that popular culture need not be debauched but could in fact transcend itself. Consider these key passages from the Pilkington Report:

> Television does not, and cannot, merely reflect the moral standards of society. It must affect them, either by changing or by reinforcing them … Because the range of experience is not finite but constantly growing, and because the growing points are usually most significant, it is on these that challenges to existing assumptions and beliefs are made, where the claims to new knowledge and new awareness are stated. If our society is to respond to the challenges and judge the claims, they must be put before it. All broadcasting, and television especially, must be ready and anxious to experiment, to show the new and unusual, to give a hearing to dissent. Here, broadcasting must be most willing to make mistakes; for if it does not, it will make no discoveries (Home Office 1962, pp. 19–20, paras. 50–3).

The suggestion here is not that public broadcasters are all hoping and dreaming that their programs will transform people from cultural and intellectual slobs into something of which one can more readily approve, but rather that objectively some such argument must be the public broadcaster's last line of defense. The language is of standards, quality, excellence, range. The logic is of social enrichment, that in however indefinable a manner this society is 'better' for having programs produced from within the framework of those social arguments termed public service broadcasting, compared to those programs produced within an environment in which commerce or politics prevail.

One cannot, however, escape the charge that those sentiments rest on a set of ill-explored assumptions about the sociological organization of modern culture. How are we better? What are the mechanics? And, vitally, where is the evidence? Indeed, what would such evidence even look like?

2. *The Public Service Idea: Eight Principles*

What then are the definable principles of public service broadcasting? Eight principles are outlined below.

2.1 *Universality of Availability*

Public broadcasting has historically sought to ensure that its signals are available to all, that no one should be disenfranchised by distance or by accident of geography. The imperative that guides this principle is not that of maximizing customers in a market but of serving citizens in a democracy, that if one defines one's audience as the citizens of a country, then logically one has to reach them all. To a remarkable extent in country after country this principle has been made real.

2.2 *Universality of Appeal*

Public broadcasting seeks to provide programs that cater to the many different tastes and interests that constitute a society's life. The public broadcasting community understands that each of us, at different moments, is part of a majority and a minority. In seeking to provide programs for a wide range of tastes and interests, public broadcasting does so with an eye cocked to the need to ensure that whether the program is pitched at the many or the few it is done so with real quality. Public broadcasting does not expect that it can please all the people all of the time—indeed it sees in that approach precisely the kind of populism that nurtures cultural mediocrity, as quality is sacrificed on the altar of maximizing the audience size.

It is an important element of this principle that public broadcasting serves not only tastes and interests that are readily apparent, but also those that are dormant and latent—that may be part of the potential we all possess but that circumstance may not have allowed us to develop.

2.3 *Provision for Minorities, Especially Those Disadvantaged by Physical or Social Circumstance*

It is commonplace to characterize the medium of television as essentially serving 'the mass.' Certainly public broadcasting understands the vast capability of one medium to reach enormous numbers of people. It sets its face, however, against the logic of commercial systems to see people as no more than statistics in skins, with a definable value captured in the most desirable rates, demographic buys, and cost per thousand. As suggested in the second principle, public broadcasting views the public as a rich tapestry of tastes and interests, each of which, insofar as possible, should be served.

In particular it recognizes that there are whole subcultures of minority social experiences crying out for attention. People of different color, language groups, and religious preferences all have vital needs for expression in the political and social discourse of the nation. Public broadcasting is dedicated to a dual role here—on the one hand to give access to such groups, to provide them with the opportunities to

speak to one another and to voice the issues as they see them, and on the other to provide coverage of their histories, interests, and concerns for the public at large.

2.4 Serving the Public Sphere

Some television programs are successful because they get a fair-sized audience, make some money, and sometimes even exemplify the craft of popular television. Other programs are successful because they reach out and touch a small, particular, but powerful audience. Some programs are successful because the craft of the program maker is used to speak to us all. They touch us, move us, make us laugh and cry and cheer. They speak to us because they speak for us. Like all great art, they help us make sense out of life, to see and understand things with a fresh eye, and give us a burning sense of the collective, of belonging to the nation-as-community. It is an increasingly vital principle of the work of public broadcasting that it recognizes its special relationship to a sense of national identity and broad community. Any nation is a patchwork of localities and regions, but it also is a nation, heterogeneous and homogeneous to a remarkable degree at one and the same time. Public broadcasting's very nature is then to nurture the public sphere as a means of serving the public good. It understands that while within civil society individuals pursue their own private self-interests, it is within the public sphere that they function as citizens. It is a fundamental principle then that public broadcasting must motivate the viewers as citizens possessing duties as well as rights, rather than as individual consumers possessing wallets and credit cards.

2.5 A Commitment to the Education of the Public

The most outstanding example of public broadcasting's commitment to the audience-as-citizen is the long-time provision in almost all systems of educational programming at every level. Public broadcasting knows that political and social literacy, as well as of course literal literacy, is an essential prerequisite to the healthy working of a democratic order. Above all else, the commitment to this principle requires that it treat its audience as mature, rational beings capable of learning and growing in many ways. Thus much of public broadcasting has retained its commitment to institutional services. Daytime school broadcasting and formal learning services of all kinds continue to play a role in most national services.

2.6 Independence from Vested Interests

It is a simple but key principle of public broadcasting that its programs can best serve the public with excellence and diversity when they are produced from within a structure of independence. Programs funded by advertising necessarily have their character influenced in some shape or form by the demand to maximize the garnering of consumers. Programs directly funded by the government, and with no intervening structural heat shield, inevitably tend to utter the tones of their master's voice.

The whole history of public broadcasting has been dominated by the idea that it can best serve the nation when it remains distanced from any particular commitment to any particular power structure inside the nation. Of particular importance to this principle is the ability of public broadcasting to support a cadre of independent-minded program makers, who are thus well able to speak with authentic tones and to offer that singularity of vision allied to creativity and passion that has traditionally produced some of public television's finest moments. It follows that the political and economic architecture of this principle is such as to support the making of programs that are good in their own terms, whatever their intended audience. In the making of programs for public broadcasting, there should be no ulterior purpose or motive. It is axiomatic to this principle that the funding of public broadcasting should be such, in total amount and in the absence of any strings attached, as to encourage rather than negate the independence enjoyed.

2.7 Encouraging Competition in Good Programming Rather Than Competition for Numbers

This principle is central to public service broadcasting and essentially involves a commitment to making programs which, whatever their intended audience, are of high quality. The overwhelming brunt of the evidence leads to the conclusion that the most important aspect of such structuring relates to the forms of finance. Where commercial sources of revenue are dominant, or even present, or where there is direct subvention from government, the program maker's eye is almost inevitably diverted away from what should be the main focus, the inherent quality of the program he or she is making.

2.8 The Rules of Broadcasting Should Liberate Rather Than Restrict the Program Maker

While all broadcasting will inevitably be governed by certain prescriptions—'educate,' 'inform,' 'entertain,' 'balance,' 'fairness'—and certain broadly drawn restrictions—obscenity, national security—the essence of the legislative foundation by which it is empowered should sustain a liberal function for the program maker. The legislation should create secure living space, arena for action, for broadcasters with all kinds of interests in possible programs and possible varieties of audience, rather than leaving the field to those who are interested chiefly in delivering maximum audiences most of the time. The legislation should also ensure

that the higher echelons of broadcasting contain individuals who understand its potential and who themselves care for the importance of the creative work of their staff. Part of that understanding would be the need for experiment and innovation in broadcasting, the need to provide a focus for a society's quarrel with itself, the recognition that mistakes will be made but as such may signify the health of the system.

Perhaps above all else, such leadership should be helped to understand that experiment, innovation, quarrel, and mistake are likely to come from the younger program maker, without whom the system is in danger of institutional arteriosclerosis.

See also: Broadcasting: General; Broadcasting: Regulation; Mass Media, Political Economy of; Mass Media, Representations in; Media, Uses of; Public Relations in Media; Public Sphere and the Media

Bibliography

Avery R L (ed.) 1993 *Public Service Broadcasting in a Multichannel Environment: The History and Survival of an Ideal*. Longman, New York

Briggs A 1985 *The BBC: The First Fifty Years*. Oxford University Press, Oxford, UK

Bullert B J 1997 *Public Television: Politics and the Battle Over Documentary Film*. Rutgers University Press, New Brunswick, NJ

Home Office 1962 *Report of the Committee on Broadcasting* (*Pilkington Report*). HMSO, London

Jarvik L A 1997 *PBS, Behind the Screen*. Forum, Rocklin, CA

Ledbettor J 1997 *Made Possible By: The Death Of Public Broadcasting in the United States*. Verso, New York

Skene W 1993 *Fade to Black: A Requiem for the CBC*. Douglas and McIntyre, Vancouver, BC

Smith A, Paterson R (eds) 1998 *Television: An International History*. Oxford University Press, New York

Stewart D 1999 *The PBS Companion—A History of Public Television*. TV Books, New York

Stuart C (ed.) 1975 *The Reith Diaries*. Collins, London

Tracey M 1983 *A Variety of Lives: A Biography of Sir Hugh Greene*. Bodley Head, London

Tracey M 1998 *The Decline and Fall of Public Service Broadcasting*. Oxford University Press, Oxford, UK

M. Tracey

Public Bureaucracies

Throughout history public bureaucracies have served as a source of continuity and stability in the life of the state. This role has sometimes led to their becoming a significant obstacle to innovation in government policies. More recently, however, these bureaucracies have also become useful agents of change in many societies.

The power that bureaucrats exercise comes mainly from the expertise they command both as individuals and when linked together in organizations. This power has come under increasing challenge in the industrialized democracies. Other institutions—governmental and nongovernmental—have begun to acquire their own expertise, and public bureaucracies are less able to cloak their operations in secrecy. Moreover, many public agencies that formerly performed regulatory or service delivery functions have been abolished, and private or quasi-public organizations have begun to perform tasks that were once within the exclusive preserve of government agencies. By the end of the twentieth century, it had become clear that public bureaucracy was not the growth industry it had appeared to be in the aftermath of World War II, fifty years earlier.

1. Bureaucracy: the Permanent Government

Although perhaps the least venerable, public bureaucracies rank among the oldest of all existing government institutions. Such bureaucracies played a central and continuing role in the operation of two of the largest and the most effective of the empires that emerged in the early stages of human history, the Chinese and the Egyptian, as they did in the imperial systems of Rome and Byzantium. Each of these was a political system that endured for centuries.

As history evolved and as nation-states emerged as the dominant form of political community in Western society, public bureaucracies were still able to demonstrate as formidable a capacity for holding on to power as their ancient forbears. Even when other political institutions, such as elected executives and legislatures, began to emerge as new centers of government authority, these public bureaucracies still managed to maintain a leading position in crafting and carrying out state programs and decisions.

This continuity in the exercise of power enabled public bureaucracies to be a source of stability in the life of a nation even when fundamental changes were taking place in the nature of the regime they served. Alexis De Tocqueville has given us a memorable description of bureaucratic organizations serving in this way amidst the turbulence of French politics at the end of the eighteenth century: 'For though in each successive revolution the administration was, so to speak, decapitated, its body survived intact and active. The same duties were performed by the same civil servants … These men administered the country or rendered justice in the name of the King, then in that of the Republic, thereafter in the Emperor's' (De Tocqueville 1955).

2. Bureaucracy and Change

Since De Tocqueville wrote, the role that public bureaucracies can play as a stabilizing force in societies beset by a time of troubles has sometimes been seen in a less favorable light. It is argued, for example, that

their durability as government institutions has enabled bureaucracies to function as an obstacle to change in the social order, long after it has become clear that such change is urgently needed. During the twentieth century, some bureaucracies, especially in the military sector, have played this kind of obstructive role in the emerging nations of Africa and Asia. In its most extreme form, resistance to change on the part of a public bureaucracy may even lead a government agency to sabotage programs placed under its control that it opposes.

2.1 Bureaucracies as Change Agents

In the modern world, however, it is also possible to find public bureaucracies that serve not as barriers to innovation but as agents of change in their own societies. When Social Democratic political parties came to power in Europe after World War II, it was widely anticipated that the reforms they intended would be undercut by the conservative bureaucracies in their own country that would be responsible for putting these programs into effect. This kind of sabotage did not in fact occur. Such bureaucracies carried out the changes proposed by the Labour Party when it took control over the British government after the parliamentary elections of 1945 with the same diligence they had shown when administering the programs of the Conservative governments under which they had previously served.

Moreover, one of the enduring legacies of that great period of social and economic reform in the USA in the 1930s, which came to be known as the New Deal, was the creation of new bureaucratic organizations such as the Rural Electrification Administration (REA) and the Tennessee Valley Authority (TVA) that were designed to serve neglected sectors of the American population. In the 1960s a similar reform movement gave birth to agencies at all levels of government in the USA that were authorized to protect women and minority groups from discrimination in employment and other spheres of everyday life. Of course, even when a bureaucracy is created to serve as an agent of change, it can still find itself under attack from its own supporters on the grounds that it is insufficiently zealous in pursuing this objective. As a result, the stereotype of bureaucracy as an obstacle to change tends to endure, even in settings where an organization is breaking new ground.

The effort to use public bureaucracies as instruments of social reform is the product of a general movement toward democracy that has taken place throughout the Western world. As this has occurred, victory in regularly scheduled elections has become the chief source of empowerment, and the political parties competing in such electoral contests have every reason to see public bureaucracies as vehicles for promoting policy changes or carrying on programs that will ultimately provide the political support necessary to win or maintain control of the government. This convergence of interest between party and bureaucracy has energized the effort in modern democratic societies to use bureaucratic institutions as instruments for breaking with the past rather than preserving it.

3. Bureaucratic Expertise

Throughout history the fundamental source from which public bureaucracies derive their power is the possession of a body of knowledge or skill indispensable to the successful performance of the government in which they operate. In many cases, especially during the early stages in their historical development, these organizations held a monopoly of such expertise. In addition to the influence it gave them within the councils of government, it enabled them to win a certain measure of deference within the community at large. To be sure, this deference was usually coupled with hostility, since ordinary citizens commonly resent the power exercised by bureaucrats in democratic as well as authoritarian societies. Such antagonism toward bureaucracies is perhaps the chief way in which these disparate systems of government continue to resemble one another in the modern world.

Over the course of modern history societies have differed in their willingness to allow expertise to become the signature characteristic of their public bureaucracies. The UK, modeling itself after the example of ancient China, pioneered in this respect, and various segments of the British Empire followed this example even after they won independence. In the more egalitarian culture of the USA, proposals to establish such a Mandarin-style bureaucracy met with strong initial resistance in the nineteenth century as betrayal of the country's democratic traditions. A similar outcry can still be triggered today in US politics by proposals that the country create an institutional counterpart to France's '*Ecole Nationale d'Administration*,' a school for the training of the French administrative elite.

3.1 Competition among Experts

One development that has greatly diminished the advantage public bureaucracies derive from their command over bodies of expertise in specialized areas of government policy has been the gradual migration of such expertise to other institutions of government. In the USA, for example, Presidents now have a large number of experts of their own clustered in various units within the Executive Office of the President. While Prime Ministers and Cabinet members in the parliamentary systems that prevail elsewhere in democratic societies do not have similar access to such organized cadres of experts, they are often able to turn elsewhere to satisfy their need for expert assistance.

3.2 The Rise of Think-tanks

Another important source of expertise widely available to elected officials today lies within the private sector in the form of private research organizations known as 'think-tanks' that have emerged and begun to focus on current issues of government policy in the Western democracies. The research findings of these organizations sometimes attract almost as much, if not more, public attention than the reports emanating from the public bureaucracies whose work they monitor. In earlier days these think-tanks tended to provide intellectual support for programs and policies that led to an expansion in the size and power of public bureaucracies, especially in the areas of the social services and economic regulation. More recently, however, think-tanks with a conservative orientation have begun to emerge, and they have provided intellectual ammunition for groups critical of public bureaucracies and the programs and policies these organizations pursue.

4. Bureaucratic Secrecy and Open Government

The emergence and expansion of new forms of mass communication has tended to make public bureaucracies increasingly vulnerable to scrutiny and criticism from institutions and groups outside the government. Pressure from such media organizations has been chiefly responsible for the enactment in a number of modern societies of 'freedom of information' statutes which require public bureaucracies to provide information in their files to private citizens or groups upon request, except for certain specified categories of data. Max Weber long ago saw the secrecy in which the public bureaucracies of his day normally operated as a major source of the power they exercised. As a study edited by Itzhak Galnoor has shown, the scope of such secrecy has been greatly narrowed in both Europe and the USA today.

5. Alternatives to Public Bureaucracies

Critics of the role of public bureaucracies in government, or of the programs they administer, have long argued that they are in fact immortal institutions, which, once established, are impossible to terminate. Certainly it is true that such bureaucracies are tenacious organizations, sinking deep roots into the fabric of the societies they serve.

Still, recent developments in modern societies make it abundantly clear that public bureaucracies are far from immortal. During the past two decades a wide variety of regulatory agencies have been put to death in both the USA and Europe, casualties of a bipartisan consensus that they mainly serve not the public but the interests of the industries they regulate. Public bureaucracies have also been victims of the growing belief that private ownership is more efficient than state operation. A striking example of this trend occurred in Canada in 1996, when Parliament transferred the vital task of carrying on the country's air traffic control system from the government's transport department to a nonprofit firm created by the legislature.

Along with these developments, there has been a continuing expansion in both Europe and the USA of organizations in the nonprofit sector that administer programs that were once thought to lie within the exclusive domain of government. These nongovernmental agencies, or 'semi-public' bureaucracies as they might be called, supplement rather than supplant the work of public bureaucracies in such areas as health, welfare, and education, and much of their revenue actually comes from the government itself. One such organization in the international sphere—Medecins Sans Frontieres—received the Nobel prize for Peace in 1999 for its work in providing medical care and moral support for population groups in distress around the world.

See also: Administration in Organizations; Administrative Law; Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Civil Service; Public Policy Schools; Public Sector Organizations; Think Tanks

Bibliography

De Tocqueville A 1955 *The Old Regime and the French Revolution*. 1st edn., Doubleday, Garden City, NY

Galnoor I (ed.) 1977 *Government Secrecy in Democracies*. Harper and Row, New York

Gidron B, Kramer R M, Salamon L M (eds.) 1992 *Government and the Third Sector*. Jossey-Bass, Inc., San Francisco

Huddleston M W, Boyer W W 1996 *The Higher Civil Service in the United States: Quest for Reform*. University of Pittsburgh Press, Pittsburgh, PA

Kaufman H 1976 *Are Government Organizations Immortal?* Brookings Institution, Washington, DC

Rourke F E 1984 *Bureaucracy, Politics, and Public Policy*. Little, Brown, Boston

F. E. Rourke

Public Commemoration

Public commemoration is a matrix of activity, through which social groups express 'a collective shared knowledge … of the past, on which a group's sense of unity and individuality is based' (Assmann 1995) The group that organizes the commemoration inherits earlier meanings attached to the event, as well as adding new meanings. Their activity is crucial to the presentation and preservation of commemorative forms. When such groups disperse or disappear,

commemoration loses its initial force, and may fade away entirely.

Public commemoration is therefore a process with a life history. It has an initial, creative phase; followed by a period of institutionalization and routinization. Such markings of the calendar can last for decades, or be abruptly halted. In most instances, the date and place of commemoration fade away with the passing of the social groups which initiated the practice.

Commemoration operates on many levels of aggregation and touches many facets of associative life. While commemorative forms were familiar in the ancient and medieval period, they have proliferated in more recent times. Consequently, the subject has attracted much discussion in social scientific literature. We therefore concentrate on public commemoration in the epoch of the nation state, primarily in the nineteenth and twentieth centuries.

In the modern period, most commemorative events were marked distinctively and separately from the religious calendar. There has been, though, some overlap. Armistice Day, 11 November, in countries remembering the end of the 1914–18 war, is close enough to the Catholic feast of All Saints on 2 November; in some countries with a large Catholic population, the two days occupy a semisacred space of public commemoration. The day marking the end of the Second World War in Europe, 8 May, is also the Saint's day of Joan of Arc. Those engaging in commemorative acts on that day may be addressing the secular celebration or the Catholic one; some celebrate the two together.

Commemoration is an act arising out of a conviction, shared by a broad community, that the moment recalled is both significant and informed by a moral message. Moments of national humiliation are rarely commemorated, although here too there are exceptions of a hortatory kind. 'Never again' is the hallmark of public commemoration on the Israeli Day of Remembrance for victims of the Nazi persecution of the Jews. Where moral doubts persist about a war or public policy, commemorative moments are hard to fix. That is why there is no date commemorating the end of the Algerian War in France, or the end of the Vietnam War in the USA. There was no moral consensus about the nature of the conflict; hence there was no moral consensus about what was being remembered in public, and when was the appropriate time to remember it (Prost 1999).

1. *Commemoration and Political Power*

Much of the scholarly debate about commemoration concerns the extent to which it is an instrument of the dominant political elements in a society. One school of opinion emphasizes the usefulness to political elites of public events establishing the legitimacy of their rule (Nora 1984–92). Some such events are observed whoever is in power—witness Bastille Day in France or Independence Day in the USA. But other events are closely tied to the establishment of a new regime and the overthrow of an older one: 7 November was the date in the calendar marking the Bolshevik revolution and establishing the Communist regime in power in Russia. That date symbolized the new order and its challenge to its world-wide enemies. May Day similarly was a moment when labor movements, and Labor parties in power, publicly demonstrated their place in history.

This top-down approach proclaims the significance of commemoration as a grammar of national, imperial or political identity. Anzac Day, 25 April, is celebrated as the moment when the Australian nation was born. It commemorates the landing of Australian and New Zealand troops as part of the British-led expeditionary force sent to Turkey in 1915. The fact that the landing was a failure does not diminish the iconic character of the date to Australians. It is the day, they hold, when their nation came of age (Inglis 1992).

By no means are all commemorative activities associated with warfare. The birthdates of monarchs or deceased presidents are marked in similar ways. Queen Victoria's birthday, 24 May, was Empire Day in Britain; now (2000) it is celebrated as Commonwealth Day. The creation of such commemorative dates was part of a wider movement of what some scholars have termed 'the invention of tradition.' That is, at the end of the nineteenth century, new nation states and preeminent imperial powers deepened the repertoire of their ceremonial activity. Such flourishes of the majesty of power were then immediately sanctified by a spurious pedigree. To display ceremonies with a supposed link to ancient habits or forms located in a foggy and distant past created an effective cover for political innovation, instability or insecurity (MacKenzie 1986, Ranger and Hobsbawm 1986).

This functionalist interpretation of commemoration has been challenged. A second school of scholarship emphasizes the ways that public commemoration has the potential for dominated groups to contest their subordinate status. However much political leaders or their agents try to choreograph commemorative activity, there is much space for subversion or creative interpretation of the official commemorative script. Armistice Day, 11 November, was, for different groups, a moment for both the celebration and the denigration of military values. Pacifists announced their message of 'Never again' through their presence at commemorative ceremonies; military men used these moments to glorify the profession of arms, and to demonstrate the duty of citizens, if necessary, to give their lives for their country in a future war. The contradictions in these forms of expression on the same day were never resolved (Gregory 1994, Winter 1999).

This alternative interpretation of the political meaning of commemorative activity emphasizes the multivocal character of remembrance. From this point of view, there is always a chorus of voices in commemorations; some are louder than others, but they never sound alone. Decentering the history of commemoration ensures that we recognize the regional, local, and idiosyncratic character of such activities.

Very occasionally, these dissonant voices come together. Between 1919 and 1938 in Britain, there was a two-minute silence observed throughout the country. Telephonists pulled the plugs on all conversations. Traffic stopped. The normal flow of life was arrested. Then the Second World War intervened, and such disruption to war production was not in the national interest. Thereafter the two-minute silence was moved to the Sunday nearest 11 November. But in the two decades between the Wars, it was a moment of national reflection. Mass Observation, a pioneering social survey organization, asked hundreds of ordinary people in Britain what they thought about during the silence. The answer was that they thought not of the nation or of victory or of armies, but of the men who were not there. This silence was a meditation about absence. As such it moved away from political orchestration into the realm of family history. To be sure, families commemorated their own within a wider social and political framework. But the richest texture of remembrance was always within family life. This intersection of the public and the private, the macro-historical and the micro-historical, is what has given commemoration in the twenty-first century its power and its rich repertoire of forms.

2. *The Business of Remembering*

Commemoration is and always has been a business. It costs money; it requires specialists' services; it needs funding and, over time, re-funding. There are two kinds of expenditure we can trace in the history of commemoration. The first is capital expenditure; the second is recurrent expenditure.

The land for such sites must be purchased; and an appropriate symbolic form must be designed and then constructed to focus remembrance activities. The first step may require substantial sums of public money. Private land, especially in urban areas, comes at a premium. Then there are the costs of architects' fees, especially when a public competitive tender is offered inviting proposals from professionals. Finally, once the symbolic form is chosen, it must be constructed out of selected materials and finished according to the architect's or artist's designs.

When these projects are national in character, the process of production is in the public eye. National art schools and bodies of 'experts' have to have their say. Standards of 'taste' and 'decorum' are proclaimed. Professional interests and conflicts come into play. Much of this professional infighting is confined to national commemorative projects, but the same complex stepwise procedure occurs on the local level too, this time without the same level of attendant publicity. Local authorities usually take charge of these projects, and local notables can deflect plans towards their own particular visions, whatever public opinion may think about the subject.

Most of the time, public funding covers only part of the costs of commemorative objects. Public subscriptions are critical, especially in Protestant countries where the concept of utilitarian memorials is dominant. In Catholic countries, the notion of a 'useful' memorial is a contradiction in terms; symbolic language and utilitarian language are deemed mutually exclusive. But the Protestant voluntary traditions has it otherwise. In Protestant countries, commemorative projects took many forms, from the sacred to the mundane: in Britain there are memorial wards in hospitals, memorial scholarships in schools and universities, alongside memorial cricket pitches and memorial water troughs for horses. In the USA and in Australia there are memorial highways. The rule of thumb is that private citizens pick up most of the tab for these memorial forms. The state provides subsidies and occasional matching grants, but the money comes out of the pockets of ordinary people. The same is true in Britain with respect to a very widely shared form of public commemoration: the purchase of paper poppies, the symbol of the Lost Generation of the First World War. These poppies are worn on the lapel, and the proceeds of the sale go to aid disabled veterans and their families.

Recurrent expenditure for commemorative sites is almost always paid for by taxpayers. War cemeteries require masons and gardeners. The Imperial (now Commonwealth) War Graves Commission looks after hundreds of such cemeteries all over the world. The cost of their maintenance is a public charge. Private charities, in particular Christian groups, maintain German war cemeteries. Once constructed, memorial statues, cemeteries, or highways also become public property, and require public support to prevent them from decomposing. They are preserved as sites of commemorative activity.

Much of this activity is directed towards inviting the public to remember in public. This means directed the public towards particular sites of remembrance. Some of them are nearby their homes. In Britain and France there are war memorials in every city, in every town, and in every village; it is there that Armistice Day ceremonies are held annually. Churches throughout Europe of all denominations have memorial plaques to those who died in war. Special prayers were added to the Jewish prayer book to commemorate the victims of the Nazis in the Second World War, and later, those who died on active service in the Israeli army.

Remembrance in local houses of worship or at war memorials required that the public travel a short

distance from their homes to sites of remembrance. But given the scale of losses in the two World Wars, and the widely dispersed cemeteries around the world in which lie the remains of millions of such men and women, the business of remembrance also entails international travel. Such voyages start as pilgrimage; many are mixed with tourism (Lloyd 1998). But in either case, there are train and boat journeys to take; hotel rooms to reserve; guides to hire; flowers to lay at graves; trinkets and mementos to purchase. In some places, museums have arisen to tell more of the story the pilgrims have come to hear and to share. There too money is exchanged along with the narratives and the symbols of remembrance.

This mixture of the sacred and the profane is hardly an innovation. It is merely a secular form of the kind of pilgrimage, for example, that made San Juan de Compostela in Spain the destination of millions of men and women in the middle ages who came to honor the conventionally designated resting place of the remains of one of the original Apostles. Pilgrimage to war cemeteries is public commemoration over long, sometimes very long, distances. Today (2000) thousands of Australians make the journey to Gallipoli, on the border between Asia and Europe, to be there for the dawn service on 25 April (as we have already noted), the day of the Anzac landing in 1915. Where does pilgrimage stop and tourism take over? It is impossible to say, but in all cases, the business of remembrance remains just that—a business.

3. *Aesthetic Redemption*

Public commemoration is not only a set of political gestures and material tasks. It is also an art form, the art of arranging and interpreting signifying practices. This field of action can be analyzed on two different levels: the first is aesthetic; the second is semiotic. The two are intimately related.

Some national commemorative forms are distinctive. Others are shared by populations in many countries. The figure of Marianne as the national symbol of the French Republic could not be used in Germany or Britain. The German Iron Cross, on commemorative plaques, denotes the location and the tradition in which commemoration is expressed. Germany's heroes' forests or fortresses are also imbricated in Teutonic history.

At times the repertoire of one country's symbols overlap with that of others, even when they were adversaries. After the First World War, the first industrialized war fought among fully industrialized nations, many commemorative forms adopted medieval notation. Throughout Europe, the revolutionary character of warfare was marked by a notation of a backward-looking kind. Medieval images of heroic and saintly warriors recaptured a time when combat was between individuals, rather than the impersonal and unbalanced duel between artillery and human flesh. The war in the air took on the form and romance of chivalry. On the losing and the winning sides, medievalism flourished.

On war memorials, the human form survived. In some instances, classical images of male beauty were chosen to mark the 'lost generation'; others adopted more stoical and emphatically untriumphalist poses of men in uniform. In most cases, victory was either partially or totally eclipsed by a sense of overwhelming loss. Within this aesthetic landscape, traditional Christian motifs were commonplace. The form of the grieving mother—Stabat Mater—brought women into the local and national constellation of grief.

In Protestant countries, the aesthetic debate took on a quasireligious character. War memorials with crosses on them offended some Protestants, who believed that the Reformation of the sixteenth century precluded such 'Catholic' notation. Obelisks were preferable. In France, war memorials were by law restricted to public and not church grounds, although many local groups found a way around this proscription. In schools and universities, the location of such memorials touched on such issues. Some were placed in sacred space—in chapels—semisacred space—around chapels—or in secular space. Public thoroughfares and train stations also housed such lists of men who had died in war.

Twentieth-century warfare democratized bereavement. Previously armies were composed of mercenaries, volunteers, and professionals. After 1914, Everyman went to war. The social incidence of war losses was thereby transformed. In Britain, France and Germany, virtually every household had lost someone—a father, a son, a brother, a cousin, a friend. Given the nature of static warfare on the Western front, many—perhaps half—of those killed had no known grave. Consequently, commemorative forms highlighted names above all. The names of the dead were all that remained of them, and chiseled in stone or etched on plaques, these names were the foci of public commemoration on both the local and the national scale.

This essential practice of naming set the pattern for commemorative forms after the Second World War and beyond. After 1945, names were simply added to Great War memorials. This was partly in recognition of the links between the two twentieth-century conflicts; partly it was a matter of economy. After the Vietnam war, naming still mattered, and First World War forms inspired memorials, most notably Maya Lin's Vietnam Veterans' Memorial in Washington. Her work clearly drew on Sir Edwin Lutyens's memorial to the missing of the Battle of the River Somme at Thiepval, inaugurated in 1932.

By the latter decades of the twentieth century, artistic opinion and aesthetic tastes had changed sufficiently to make abstraction the key language of commemorative expression. Statues and installations

thereby escaped from specific national notation and moved away from the earlier emphasis upon the human figure. The exception to the rule is Soviet commemorative art, which resolutely stuck to the path of heroic romanticism in marking out the meaning of what they called the Great Patriotic War. In many instances in Western Europe, but by no means all, forms which suggested absence or nothingness replaced classical, religious or romantic notions in commemorative art.

This aesthetic shift had important social and political implications. Great War commemorative forms had sought out some meaning, some significance in the enormous loss of life attending that conflict. There was an implicit warning in many of these monuments. 'Never again' was there ultimate meaning. But 'never' had lasted a bare 20 years. Thus after the Second World War, the search for meaning became infinitely more complex. And the fact that more civilians died than soldiers in the Second World War made matters even more difficult to configure in art.

Finally, the extreme character of two elements of the Second World War challenged the capacity of art—any art—to express a sense of loss when it is linked to genocidal murder or thermonuclear destruction. Both Auschwitz and Hiroshima defied conventional notations of 'meaning,' although some individuals continue to try to rescue redemptive elements from them both. The construction of a national Holocaust memorial at the Brandenburg gate in Berlin, the construction of which began in 1999, is a case in point (Young 1999). The debate over this and similar projects does not abate, since such events seem to force all participants—public officials, the public, survivors, architects, artists—to the limits of representation and beyond.

4. *Ritual*

Public commemoration is an activity defined by the gestures and words of those who come together to remember the past. It is rarely the simple reflection of a fixed text, a script rigidly prepared by political leaders determined to fortify their position of power. Inevitably, commemoration overlaps with political conflicts, but it can never be reduced to a simple reflection of power relationships.

There are at least three stages in the history of rituals surrounding public commemoration. The first we have already dealt with: the construction of a commemorative form. But there are two other levels in the life history of monuments which need attention. The second is the grounding of ritual action in the calendar, and the routinization of such activities; the third is their transformation or their disappearance as active sites of memory.

One case in point may illustrate this trajectory. The date of 1 July 1916 is not a national holiday in Britain; but it marks the date of the opening of the British offensive on the river Somme, an offensive which symbolized the terrible character of industrial warfare. On that day the British army suffered the highest casualty totals in its history; on that day a volunteer army, and the society that had created it, were introduced to the full terrors of twentieth-century warfare. To this day, groups of people come to the Somme battlefields to mark this day, without national legislation to enable them to do so. Theirs are locally-defined rituals. In France, 11 November is a national holiday, but not in Britain. Legislation codifies activities the origins and force of which lie on the local level.

Public commemoration flourishes within the orbit of civil society. This is not true in countries where dictatorships rule; Stalinist Russia smashed civil society to a point that it could not sustain commemorative activity independent of the party and the state (Merridale 1999). But elsewhere, local associations matter.

And so do families. Commemorative ritual survives when it is inscribed within the rhythms of community and, in particular, family life. Public commemoration lasts when it draws about overlaps between national history and family history. Most of those who take the time to engage in the rituals of remembrance bring with them memories of family members touched by these vast events. This is what enables people born long after wars and revolutions to commemorate them as essential parts of their own lives. For example, children born in the aftermath of the First World War told the story of their family upbringing to grandchildren born 60 or 70 years later. This transmission of childhood memories over two or sometimes three generations gives family stories a power which is translated at times into activity—the activity of remembrance (Winter and Sivan 1999).

This framework of family transmission of narratives about the past is an essential part of public commemoration. It also helps us understand why some commemorative forms are changed or simply fade away. When the link between family life and public commemoration is broken, a powerful prop of remembrance is removed. Then, in a short time, remembrance atrophies and fades away. Public reinforcements may help keep alive the ritual and practice of commemoration. But the event becomes hollow when removed from the myriad small-scale social units that breathed life into it in the first place.

At that moment, commemorative sites and practices can be revived and reappropriated. The same sites used for one purpose can be used for another. But most of the time, commemorative forms live through their life cycle, and like the rest of us, inevitably fade away. This natural process of dissolution closes the circle on public commemoration. And rightly so, since it arises out of the needs of groups of people to link their lives with salient events in the past. When that

need vanishes, so does the glue that holds together the social practice of commemoration. Then collective memories fade away.

We have reached, therefore, a quixotic conclusion. Public commemoration is both irresistible and unsustainable. Time and again people have come together in public to seek meaning in vast events in the past and try to relate them to their own smaller networks of social life. These associations are bound to dissolve, to be replaced by other forms, with other needs, and other histories. At that point, the characteristic trajectory of public commemoration—creation, institutionalization, and decomposition—comes to an end.

See also: Collective Identity and Expressive Forms; Collective Memory, Anthropology of; Collective Memory, Psychology of; Media Events; Pilgrimage; Public Sphere and the Media; Ritual

Bibliography

Assmann J 1995 Collective memory and cultural identity. *New German Critique* **65**: 125–34 (esp. 132)

Gregory A 1994 *The Silence of Memory*. Berg, Leamington Spa, UK

Inglis K 1992 World War One memorials in Australia. *Guerres Mondiales et Conflits Contemporains* **167**: 51–8

Lloyd D 1998 *Battlefield Tourism: Pilgrimage and the Commemoration of the Great War in Britain, Australia and Canada, 1919–1939*. Berg, Oxford, UK

MacKenzie J (ed.) 1986 *Imperialism and Popular Culture*. Manchester University Press, Manchester, UK

Merridale C 1999 War, death and remembrance in Soviet Russia. In: Winter J, Sivan E (eds.) *War and Remembrance in the Twentieth Century*. Cambridge University Press, Cambridge, UK, pp. 61–83

Nora P (ed.) 1984–92 *Les Lieux de Mémoire*. Gallimard, Paris

Prost A 1999 The Algerian War in French collective memory. In: Winter J, Sivan E (eds.) *War and Remembrance in the Twentieth Century*. Cambridge University Press, Cambridge, UK pp. 161–76

Ranger T, Hobsbawm E (eds.) 1986 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Winter J 1999 *Sites of Memory, Sites of Mourning: the Great War in European Cultural History*. Cambridge University Press, Cambridge, UK

Winter J, Sivan E 1999 Setting the framework. In: Winter J, Sivan E (eds.) *War and Remembrance in the Twentieth Century*. Cambridge University Press, Cambridge, UK, pp. 1–40

Young J E 1999 Berlin's Holocaust Memorial: a report to the Bundestag Committee on Media and Culture. *German Politics and Society* **17**(3): 54–70

J. Winter

Public Good, The: Cultural Concerns

From the inception of political theory to twenty-first century reflections on democracy and the welfare state, the ideal of public good has consistently been linked to the sociomoral resource of civic spirit. This is so insofar as the ideal of the public good claims civic spirit from a society's citizens, while the presence of civic spirit is an indispensable premise for any orientation toward the public good.

Through such a circular linkage, the ideal has gained strategic importance as an element of political rhetoric catalyzing a process of normative self-limitation, due to conditions at play in the democratic public sphere.

1. The Public Good in the History of Political Ideas

From time immemorial, the idea of the public good has been a fixed quantity in debates concerning legal, economic, and political forms of social cooperation. For example, in ancient Egypt, Pharaoh's duty to see to the welfare of his subjects was considered an element of *Ma'at*, the world-order connecting human beings to the gods; and in Islamic legal doctrine, the normative standard for the legal-religious community, *maslaha*, can be interpreted as similar to the idea of the public good. The concept's occidental career began in ancient Greece. Plato distinguished the public good from private profit, linking the former to the idea of justice (Plato 1993, 420 b–c, 421 a–c). Aristotle followed Plato in this respect (Aristotle 1990, III./VII./1); but in contrast to Plato he did not equate the just state with an authoritarian pedagogic utopia. Rather, he understood the public good to involve a political limitation on the right to rule, demarcating a difference between the proper constitutional form, which acknowledges such a good, and the improper, serving the ruler's interests (Aristotle 1990, III./IV./7). The political thinking of Roman republicanism focused on reconciling individual interests with the public good, the legal system here taking on special meaning. But even at this early period, lamenting at a loss of orientation around the public good—this seen as the source of a growing danger to the republic—is a predominant theme.

The redemptive orientation of the early Christian fathers transcended the antique approach to the public good, depoliticizing it in the process. But within scholasticism, the dogma of the subordination of earthly existence to the divine order again received a ruler-limiting and hence political accent. In his *Summa Theologiae*, Thomas Aquinas defined a regime as unjust tyranny when the ruler does not strive for the public good but for his private benefit (1952, Pars IIa, IIae, q.42, a.2). Subjects can here thus evoke the public good, placed higher than the ruler's interests, in opposition to the ruler; and the ruler can appeal to the public good to justify his actions. In this manner, a struggle emerged regarding semantic supremacy over the concept of public good—the concept being capable of both limiting and legitimizing political rule. If the ruler could claim to be enacting the public good,

nothing stood in the way of his rule; to be sure, the latter was then limited through involvement of the estates—such involvement being inextricably bound up with the idea of public good in the medieval world of political ideas (Eberhard 1993).

Whereas into the early modern period, self-interest and the public good represented 'asymmetrical counterconcepts' in Reinhart Koselleck's sense, i.e., 'binary concepts with claims to universality … involving a mutual recognition of the parties involved' (Koselleck 1985), the Reformation saw the onset of a relativization process. Set in motion by the 1555 Peace of Augsburg, the increased individualization in estate-bound society was accompanied by a normative transformation, its import captured in Leonhard Fronsberger's initial (at the time singular) praise of self-interest. This already reflected an important paradigm shift that would unfold over three decades, replacing the old European ideal of political virtue, in the sense of the citizen's voluntary orientation toward the public good (Münkler 1991). Instead, rational self-interest now emerged as a core political–theoretical concept. Its immediate context was a market society in the course of self-differentiation, with the state now simply expected to set the conditions for maximized economic well-being of the citizenry in a *civil* society. By turning the inherited conceptual opposition of public good and self-interest upside down, Bernard Mandeville's *The Fable of the Bees* (1929; its first, pamphlet form appearing in 1795 as *The Grumbling Hive or Knaves turn'd Honest*) played a key part in this transformative process. 'Private vices, public benefits' was Mandeville's credo. The end effect of each individual pursuing his private good is the greatest possible advantage for the polity, so that one, particular maxim prevails: 'The worst of all the Multitude Did something for the Common Good' (Mandeville 1929). Admittedly, this construct is based on citizens being in a position to pursue their own interests rationally; and Mandeville remained skeptical regarding his own theory's implicit premise. In fact, he demanded state intervention to check the loss of civic spirit—particularly feared on account of short-sightedness and ignorance when it came to the needs of following generations.

In respect to this general problem, Adam Smith was essentially more optimistic. Bringing Mandeville's idea to a head, he saw a mechanism at work that produces a common benefit in direct proportion to efforts at its opposite. The citizen, he argued, 'generally, indeed, neither intends to promote the public interest, … he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention' (Smith 1981). With this argument, Adam Smith in a way achieved liberalism's semantic coup vis-à-vis the public good. Through the definition of the public good as a wondrous result of well-understood self-interest—a predictable product of the individual's egoistic maximization of advantage—the asymmetric conceptual opposition between self-interest and public good, already relativized by Mandeville, was enduringly destroyed (Bohlender 1998). Reflecting the advent of modern contractual theory, this way of thinking can lead to the sort of relativizing of the idea of public good that we find in Kant: a complete focus on formal legal principles, and the treatment of the sociomoral qualities and intentions of both rulers and the ruled as a secondary matter. Kant, however, was not the final word in the long conceptual history of the public good, for speaking up against him, Hegel had already insisted that 'the so-called "general good," the welfare of the state, i.e., the right of mind actual and concrete, is quite a different sphere, a sphere in which abstract right is a subordinate moment like particular welfare and the happiness of the individual' (Hegel 1942, Sect. 126 Remark).

In the twentieth century, the notion of public good was subject to steady abuse: first by the Nazis, who used the empty phrase 'the common interest precedes self-interest' as propaganda for its collectivist 'you are nothing; your *Volk* is everything' ideology, then by the Bolshevists. As a consequence, a refusal substantively to define the public good and how it is to be realized, once and for all, has emerged as an identifying trait of proponents of freedom and democracy. Rather, through use of democratic procedures, society needs constantly to arrive at a consensus regarding the concept—a consensus subject to revision. This does not preclude the pluralistic process of opinion formation being oriented toward normative concepts of public good with universal, i.e., ubiquitous and extra-temporal claims to validity: the natural-law claims, for instance, staked by Catholic advocates of social transformation. In the context of, among other things, a global ecological crisis, such universal claims have been further developed into a concept of a worldwide public good.

2. *Public Good and Civic Spirit*

This overview of the conceptual history of 'public good' points to two basic features. First, it represents a functional formal concept with effects that vary considerably according to the context of usage and semantic competence. It can, for instance, serve for both the legitimation and limitation of rule. The renunciation by democratic societies of an *a priori*, essentialist understanding of the public good can thus be seen as a consequence of the insight into its semantic indeterminacy having become manifest in the course of time.

Second, precisely when thematized from a democratic perspective, the concept of public good needs to be viewed in connection with another central theme of political theory, namely, civic virtue or civic spirit, which Mandeville had already addressed in his skep-

tical fable. Against the contractualistic tendency to place an economic model of order in the foreground, Montesquieu, for example, had stressed that, more than is the case with any other form of state, democracies must see to it that their citizens identify with the public good. For a republic, he indicates, needs 'a continuous preference of the public interest over one's own' (Montesquieu 1989, IV, 5)—and this without the state force that Mandeville speaks of. The emphasis on this voluntariness underscores the importance of a minimum of civic spirit (or civic virtue) as an indispensable premise for any orientation toward the public good. The ideal of public good sets the norms for the behavior of a commonwealth's free citizens; it needs, however, the preliminary presence of some degree of civic virtue as a willingness to act according to normative ideals in the first place. As such a preliminary, motivational specification for any orientation toward the public good, civic spirit emerges as a prepolitical basis for the political, only capable of being inadequately guaranteed by democratic politics, which may not control (or even guide) the citizenry's motivations (Böckenförde 1991, Goodin 1992).

In this sense, civic spirit is a *nonrenewable* sociomoral resource (as the metaphor of social capital suggests, by alluding to the economic understanding of 'capital') and hence it ought to be neither undertaxed nor overtaxed. Undertaxing occurs when, for instance, the solidarity every society needs increasingly becomes a form of institutional machinery (Gunnar Myrdal), creating an impression among individual citizens that the state is itself removing all required acts of solidarity from their own purview. Overtaxing is apparent when the political–social unit subject to the norms set by the public good exceeds a certain size, the imagination of belonging and reciprocal interdependence required for public good thus no longer being capable of production. This is all the more likely when such extension is a successive process making use of moral appeals and excessive public-good rhetoric with visible semantic development. Old European republicanism already saw the erosion of a society's sociomoral resources being threatened here; and communitarianism confirms this in referring to the need for supportive habits, the experience of personal identification, and one's own acts of solidarity (Bellah et al. 1985, 1991, Walzer 1983). But even Adam Smith's trust in the invisible hand was tied to the premise of a limited, overseeable marketplace: one guaranteeing that consumers could themselves sanction an egoistic benefit-maximization spurred forward at their own cost.

A game-theoretical analysis using postulates of rational choice backs up this assumption of a link between, on the one hand, a group subject to public-good norms having a comprehensible size, and, on the other hand, civic virtue (or civic spirit) as the readiness to follow these norms. Take, for example, a group of people who visit an expensive restaurant, having agreed together to share the costs equally regardless of the different prices of individual orders: the factors determining the choice between self-interest and cooperation are the group's size, communicative capacity, and duration. In this manner, we can perceive the expectation of cooperation as a decisive element in the calculation of individual benefit—thus rendering civic spirit an explicable quantity of political–social action (Glance and Huberman 1994).

But not only democracy-theoretical factors connect the two basic features of the political idea of the public good—the fact that it represents a functional formal concept, and its immanent ties to civic spirit. The ongoing debate over the future of the Western welfare-state model underscores an interesting variation on the historically familiar semantics of *public good* and *civic spirit*. It is evident, for instance, in an argument regarding distribution: policies misperceived as a guarantee of welfare by the state produce a situation in which egoistic benefit-maximization can no longer further the public good, but—caught up in an illiberal system of administrative care—simply results in an exploitation of social output. According to this argument, a welfare state oriented towards a false ideal of public good effects something like an amputation of the invisible hand conjured up by Adam Smith; it does so by gradually diminishing the productive readiness of its members—hence eroding its own sociomoral basis, the free and responsible engagement of its citizens. Without the sociomoral resource of civic spirit, the development of active and productive self-interest—the self-interest that promotes the public good—is paternalistically thwarted. What remains is passive and unproductive self-interest: the sort damaging the public good.

3. *The Public Good as a Political Concept*

The meaning of the public good as a political concept—one through which interest-conflicts are consistently played out as interpretive struggles—emerges not least of all in this rhetorical mode of a differing public-good semantics. With the satisfaction of one's own demands being represented as a general advantage, precisely this theme has the merit of masking a variety of self-interests. As has been shown in an empirical study of the orientation of different German industrial organizations toward the public good, participating in such interpretive struggles has now basically become compulsory. For those failing to offer a 'declaration of public good' risk finding their positions cast under the suspicion of group-egoism and irrationality. Lack of responsibility and self-inflicted damage then emerge as the discrediting political-semantic labels (Vobruba 1992, Luhmann 1990, 2000). Once the connection between public good and civic spirit is established, political rhetoric can

even have its effect as self-fulfilling prophecy. This is the case when an impression of general decline in civic spirit is awakened among the citizenry, in turn prompting a politically-induced disappointment, defined by Albert O. Hirschman as the basis for withdrawal into the self-interested, private sphere: a disappointment cyclically alternating with periods of engagement for the public good (Hirschman 1982).

We can thus discern one particular continuity between the medieval and contemporary approach to the public good: the effort at a monopoly over its definition as constituting a central element of political conflict. Now as before, the theme can be invoked both as a basis for political demands and proposals for reform and, as it were, a 'governmental concept' meant to defend individual claims of a various interest groups. The latter function involves the government's task being depicted as to protect the public good from damaging special interests—even if such a protective function in a pluralistic democracy itself requires validation. In this respect, the breadth of the concept's variations—taking the form, as well, of common good, common weal, and public interest, each with its own fine shade of meaning—offers a rich range of possibilities for political communication.

Nevertheless, accompanying these different rhetorical functions, the idea of the public good consistently retains what—broadly speaking—constitutes a normative dimension: this as a result of the self-limitation necessarily imposed on anyone claiming to orient his political behavior toward the public good. Such a limitation furnishes a check on any purely strategic exploitation of the rhetoric of public good. Those invoking the concept in public debate reveal great expectations regarding the total effect on the welfare state of the policies they are enacting or defending. (This remains the case even when the invocation serves to rebuff claims and demands tendered by various social groupings.) At the same time, they are publicly aligning themselves with a standard against which they and their policies are to be measured due to conditions at play in the democratic public sphere. In this sense, invoking the public good involves a self-limitation by society's political agents.

To this extent, its status as a functional formal concept contradicts neither the public good's normative dimension, resulting from the obligation of honesty in democratic politics, nor its political relevance. This leads once more to both the extreme attitudes toward the public good mentioned above. On the one hand, the radically individualistic view sees free-market mechanisms without bureaucratic restriction or attempts at political steering as the best way to promote the public good. Beside the fact that such a radical view must ignore the linkage between the public good and the sociomoral resource of civic spirit we have been discussing, the use of public-good rhetoric not only by politicians, but also by representatives of society's other functional systems, indicates that the ideal of the public good is an inevitable category for any public action, particularly within the political system. At the same time, the conditions of pluralistic democracy exclude any substantialistic and durably valid definition of the public good *a priori*, as characterizing totalitarian ideologies. In a pluralistic democracy, the public good is necessarily an 'open' formula, comprising a valid norm only if it can be freely accepted by all of the individuals who are affected by it. The application of this Habermasian idea to the ideal of the public good points to the limits of its validity, as well as its importance as a core concept of political thought and practice.

See also: Altruism and Prosocial Behavior, Sociology of; Citizenship, Historical Development of; Citizenship: Sociological Aspects; Civic Culture; Civil Society/Public Sphere, History of the Concept; Individualism versus Collectivism: Philosophical Aspects; Public Goods: International; Virtue Ethics

Bibliography

Aristotle 1990 Politics. In: *Aristotle in Twenty-three Volumes*. Harvard University Press, Cambridge, MA, Vol. 21

Bellah R N, Madsen R, Sullivan W M, Swindler A, Tipton S M 1985 *Habits of the Heart. Individualism and Commitment in American Life*. University of California Press, Berkeley, CA

Bellah R N, Madsen R, Sullivan W M, Swindler A, Tipton S M 1991 *The Good Society*. Knopf, New York

Böckenförde E W 1991 *State, Society and Liberty. Studies in Political Theory and Constitutional Law*. Berg, New York

Bohlender M 1998 Government, Commerce und Civil Society. Zur Genealogie der schottischen politischen Ökonomie. In: Kaelble H, Schriewer J (eds.) *Gesellschaften im Vergleich*. Peter Lang, Frankfurt, Germany, pp. 115–47

Eberhard W 1993 Herrscher und Stände. In: Politisches Denken in der Zeit der Reformation. In: Fetscher I, Münkler H (eds.) *Pipers Handbuch der politischen Ideen. Vol. 2: Mittelalter: Von den Anfängen des Islams bis zur Reformation*. Piper, München, Germany, pp. 467–551

Glance N S, Huberman B A 1994 The dynamics of social dilemmas. *Scientific American* **270**: 76–82

Goodin R E 1992 *Motivating Political Morality*. Blackwell, Cambridge, MA

Hegel G W F 1942 *Hegel's Philosophy of Right*. Clarendon, Oxford, UK

Hirschman A O 1982 *Shifting Involvements. Private Interest and Public Action*. Princeton University Press, Princeton, NJ

Koselleck R 1985 *Futures Past. On the Semantics of Historical Time*. MIT Press, Cambridge, MA

Luhmann N 1990 *Political Theory in the Welfare State*. W de Gruyter, Berlin

Luhmann N 2000 *Die Politik der Gesellschaft*. Suhrkamp, Frankfurt, Germany

Mandeville B 1929 *The Fable of the Bees: Or, Private Vices, Publick Benefits*. Clarendon, Oxford, UK

Montesquieu C S 1989 *The Spirit of Laws*. Cambridge University Press, Cambridge, UK

Münkler H 1991 Die Idee der Tugend. Ein politischer Leitbegriff im vorrevolutionären Europa. *Archiv für Kulturgeschichte* **73**(2): 379–403

Plato 1993 *Republic*. Oxford University Press, Oxford, UK
Smith A 1981 *An Inquiry into the Nature and Causes of the Wealth of Nations*. Liberty Classics, Indianapolis, IN
Thomas Aquinas 1952 *S. Thomae Aquinatis Summa Theologiae*. Marietti, Torino, Italy
Vobruba G 1992 Wirtschaftsverbände und Gemeinwohl. In: Mayntz R (ed.) *Verbände zwischen Mitgliederinteressen und Gemeinwohl*. Verlag Bertelsmann Stiftung, Gütersloh, Germany, pp. 80–121
Walzer M 1983 *Spheres of Justice. A Defense of Pluralism and Equality*. Basic Books, New York

H. Münkler and K. Fischer

Public Goods: International

International public goods are goods that are accessible to all at international level, the consumption of which by one individual does not reduce the possibility of consumption by other individuals. However, there are few public goods of this pure type. Instead, nearly all are characterized by mostly unlimited accessibility and simultaneous rival consumption, i.e., the use of the resource by one individual reduces the possibilities for others to consume this good. These types of goods are called common-pool resources. Where there is free accessibility and rival consumption, a collective action problem or social dilemma arises: how can the individual user be prevented from using the limited resource and jeopardizing the long-term utility of all as a result of maximizing their own short-term individual interest? A number of institutional possibilities exist to secure the provision of public goods and common-pool resources. However, the possibilities for collective action available at national level—such as legal prescriptions imposed by majority decision in order to force individual actors to comply with a scheme of collective action—are not all applicable at international level because of nation–state sovereignty. In what follows, public goods will first be defined, then different paths to the development of institutional answers to collective action problems will be presented, before different institutional modes of provision of public goods and common-pool resources and their possible implications will be discussed.

1. *Definition of Public Goods*

Public goods and common-pool resources—as stated above—may be defined by the analytic economic criteria of universal accessibility and (non-)rival consumption. These attributes may be properties that are inherent in the goods themselves, or they may be properties that are attributed to goods through societal consensus and/or political decision. In the older economic literature the central attributes of public goods—nonexcludability and nonrivalry (Samuelson 1954, Musgrave and Musgrave 1976)—have been regarded as sources of market failure, from which the need for state intervention has been derived. The problem with common-pool resources and public goods lies in their limited excludability. As a result, prices cannot be charged and hence private provision through the market is not feasible. More recently, economic theory (Cornes and Sandler 1996) has argued in a more differentiated way that an exception is the case where a particular actor has an incentive to provide a public good irrespective of the free-riding behavior of other beneficiaries. The same holds true when, in the case of a common-pool resource, the stakes of provision are high enough to induce cooperation among actors (Cornes and Sandler 1996). Only in cases where the private provision of common goods fails are governments called upon to remedy the market failure (Bator 1958). Another aspect from which public goods can be viewed is the phenomenon that some market activities generate external effects. Positive, as well as negative, external effects are not part of the actor's economic calculation and as a result they are under- or over-produced from an aggregate view. In the traditional perspective externalities are seen as by-products of economic activities and internalization of externalities traditionally has been viewed as a government task (Pigou 1920). However, in the 1960s a new interpretation started to dominate the economic literature. According to this view, Pareto-relevant externalities are internalized through market forces if property rights are well defined and transaction costs are sufficiently low (Coase 1960, Buchanan and Stubblebine 1962).

While economic theory views the properties of a public good as inherent attributes of such a good from which a specific institutional form of provision is deduced, political science goes beyond this notion to insist additionally on a processual view of common goods. It argues that what constitutes such a good and how it is provided for institutionally depends on specific situational and technological conditions, and, moreover, is determined in a political and social process. First, it is not clear in many cases what constitutes a public good or a common-pool resource in so far as there are no absolute and objective criteria for defining them in a general sense. Instead, what constitutes a public good depends in part on the specific situational context of provision and consumption of the good, and its technical properties.

Thus, the importance of a specific situational context for the definition of a good as a public good in terms of nonexcludability and nonrival consumption is illustrated by the example of the protection offered by the ozone layer vis-à-vis radiation for sunbathers, which is a public good in the classical sense. One sunbather enjoying the sun does not impinge upon the

protection offered to other sunbathers. In this sense, the ozone layer is a public good, accessible to all with nonrival consumption. In another sense, however, it may be perceived as a common-pool resource. Seen in terms of the relationship between producers and sunbathers, and given that scientific research has identified a problem of scarcity in the ozone layer, certain productive activities, such as the production of refrigerators with chlorofluorocarbons, diminish the protective quality of the ozone layer and jeopardize the protection offered to all. This is rival consumption. In this specific context, the ozone layer constitutes a common-pool resource which is accessible to all and is depletable (Snidal 1986).

Technological innovation may force a reconsideration of the definition of a public good in cases where a good, which had hitherto been regarded as universally accessible and nondiminishable, has been rendered potentially excludable. A case in point is broadcasting. This good used to be accessible to all and of nonexcludable nature, but as a result of technological change, access to the good can now be limited to those paying for the service via a telematic coding system. Put in different terms: the transaction costs for establishing property rights were reduced drastically by technological progress so that the gains from excluding free riders and the precise allocation of user costs now exceed the costs of establishing property rights. Second, the definition of what constitutes a public good and what does not, may not necessarily be determined by 'objective' characteristics, but instead by political and social definition. There are huge cultural variations across polities and time as to what is considered a public good or common-pool resource and what is a private good. Thus, universal and free healthcare may be formulated as a political goal and consequently be considered as a common-pool resource in one polity. But in another, the same service may be regarded as a private good that each individual has to buy on the market. In the same way, policy makers in one society may decide that its provision should be left entirely to the buying power of the individual through the market. Hence, what constitutes a public good or common-pool resource can be answered either in terms of analytic economic criteria or in terms of a process of social and political definition. Once a good has been identified as a public good or a common-pool resource, the institutional mode of provision of the good has to be determined.

2. *The Provision of Institutional Answers to a Collective Action Problem*

Different paths may lead to an institutional answer for a stable solution for providing public goods at international level. The solutions may be designed intentionally, as is the case of international agreements. They may develop without being planned, as in a case of spontaneous coordination, for example, in the case of the standardization of products. Either way, once in place, they develop an institutional dynamism of their own and over a period of time they may transform the processes in the institution and, in turn, transform the preferences of the actors involved.

The intentional formulation of an institutional solution to a public-good problem may presuppose an acute awareness of an existing social dilemma and a need for collective action where there is cross-boundary problem interdependence. For example, in the case of an imminent cross-border environmental problem, an international conference will be convened that seeks to develop a format of cooperation. After a lengthy process of bargaining, institutional answers for solving the problem will be found and agreed. All those involved in the negotiations will seek to maximize their own interests, for example, by minimizing the implicated costs. In other words, institutional solutions to collective action problems are interpreted in terms of their distributional impacts on the actors involved (Knight 1995). In the great majority of cases, the negotiators do not all enjoy the same degree of power, and in the event of an agreement not being reached, the stronger actors may take up a fall-back position. While a solution may be desirable for them, they may not consider it vital. In other words they have more resources for negotiating an institutional solution. This route to an institutional solution, i.e., interest-oriented negotiation with a view to the distributional results, generally applies where the interests at stake are clear-cut and substantial. A case in point is the negotiation of CO_2 emission limits at international climate conferences. All the states engaged in the bargaining process, to find an institutional solution for reducing emissions, view the process in terms of its distributional impacts on the market position of domestic industries, and so forth. However, there are situations where collective action problems need to be solved and the actors involved do not really appreciate the nature of the problem at hand, much less which interests are at stake. Consequently, the preferences of the actors involved are uncertain. In such situations deliberation, exchange of arguments, and persuasion will prevail as the actors strive to reach, at best, a common perspective on the problem at hand together with possible solutions; or as a minimum, to establish clarity as to the different stakes involved (Knight and Johnson 1994). Only then does interest-oriented bargaining follow and asymmetries of interests and power come into play. Cases in point are complex environmental problems or the problem of providing more transparency and certainty in financial markets.

The second avenue for developing an institutional solution to a collective action problem is that of the spontaneous emergence, rather than the intentional development of a solution in a planned social interaction at international level. Here, the strategic

uncertainty as to the preferences of other actors promotes behavior likely to resolve a social dilemma. It makes sense to shape other actors' perceptions of one's preferences and to give signals of willingness to coordinate and be trustworthy. These signals enable like-minded actors to locate others who are equally so, and who can in turn build up a reputation for responding cooperatively to trust (Bates 1988). Because actors prefer coordination as one of the outcomes, as opposed to noncoordination, they will use whatever salient information (focal point) they have in order to achieve that coordination (Schelling 1960). In time, other actors will follow suit, and a social convention is established which provides information about the future actions of others (Knight 1995). The focal point may be relatively contingent, but henceforth serves as a point of reference for future coordinating actions (Calvert 1986, Schelling 1960). Third, irrespective of the path along which an institutional solution develops, be it design or spontaneous emergence, once established, it gains a dynamism of its own. The institutional process then begins to develop its own rules beyond the formal and informal rules originally instituted, or interprets the original rules in the light of the day-to-day need to make decisions, thereby transforming and developing those rules incrementally (March and Olsen 1989). This process may change the preferences of the actors involved. As well as the path of development of an institutional solution to secure the provision of a public good or a common-pool resource, the specific context in which it is shaped and the institutional arrangements in which it is embedded have been discussed in the literature.

3. *The Context and Institutional Mode of Provision*

What is the general context and what are the more specific institutional and instrumental conditions in which the provision of a public good and a common-pool resource occurs? The overall question is: under which conditions do particular institutional arrangements apply particular instruments to produce favorable results for providing public goods at international level? With respect to the overall context, generally it is argued that the greater the number of participants involved in an attempt to resolve a social dilemma and the greater the heterogeneity of the actors involved, the greater the difficulty in arriving at a solution. In this sense, the prospects of an efficient solution to collective action problems for agreements at international level, with its heterogeneous actors and large numbers of participants, would appear to be bleak. However, as has been pointed out (Ostrom 1990, Martin 1994, Snidal 1984), this relationship is not as simple as it first appears. Heterogeneity may constitute a precondition for an actor to take the lead in organizing collective action because they have a specifically strong incentive for this action to be successful. Thus, the US has repeatedly acted as a hegemonic actor in cooperating with Western public and private actors when it comes to solving collective action problems such as sharing banking information to enable countries to uncover money-laundering (Schmidt 1995). Furthermore, the problem of coordinating an unwieldy number of actors is often alleviated by organizing the participants into groups according to commonly- shared criteria. Each group is represented by a speaker, as in the case of climate negotiations, thereby reducing the number of actors sitting at the bargaining table. Moreover, the extent to which homogeneity facilitates negotiations in collective action problems depends upon the nature of the homogeneity: if all actors are intent on having the same resource, homogeneity may render agreement difficult. If, by contrast, the preferences of actors differ over time and in intensity, they may be made complementary through issue-linkage (Martin 1994).

Beyond the general context conditions, various institutional arrangements and instruments are applied in the provision of public goods and common-pool resources at international level which determine the degree of success in solving the collective action problem. The institutional arrangement denotes the structure and the nature of the relationship among the actors involved; the instruments that are applied define how the desired behavior of the target actor is to be elicited. The basic types of institutional settings for providing public goods are hierarchy, negotiation, societal self-regulation, and markets (Mayntz 1997, Scharpf 1999). Hierarchy is an arrangement in which a central actor (government) has an electoral mandate that permits it to legislate specific provisions, such as the statutory requirement to use a fully regulated catalytic converter, with noncompliance being sanctioned in order to protect the quality of ambient air. Thus, legal obligation is backed up by a bureaucratic apparatus to ensure implementation. In an international context, hierarchy as an institutional arrangement for providing public goods and protecting common-pool resources can only be applied to a very limited extent. However, even in international negotiations among sovereign states seeking to solve a collective action problem, a certain amount of initiative needs to be taken by one or a group of actors if public goods are to be provided jointly. As a result, we find power being yielded by central actors in varying degrees in order to facilitate cooperation. Thus, a central body may consist of member-actors who only make framework decisions that subsequently are made more specific by actors at the decentral level, or the central body may delegate power to members to implement the decisions. However, it is also conceivable that the outcome of autonomous actors cooperating voluntarily is strict implementation by command-and-control once a decision is agreed. Depending on whether the decision is vaguely, as

opposed to precisely, formulated the ensuing latitude for such actors is either great or little. Power to sanction noncompliance may again be vested in a central body or, alternatively, in decentral bodies (Martin 1994).

Another important institutional setting in which solutions to collective action problems at international level are found is negotiations in bipartite, tripartite, or multipartite actor constellations. However, it can often be linked to elements of various forms of hierarchy. Thus, if negotiations are not successful in bringing about a solution to a problem, actors may move out of the arena and turn to a court or a tribunal for a ruling in order to overcome the conflict. In anticipating this possibility, actors may be more willing to compromise. Thus, in the European dispute over social policy, if social partners do not negotiate an agreement successfully, the Commission may propose legislation for submission to the Council of Ministers. The 'shadow of hierarchy' (Scharpf 1997) speeds up negotiations. Some actors are particularly apt at brokering agreements in negotiations at international level in order to bring about a common solution. In the European context, the Commission has played a pre-eminent role in hammering out such agreements by acting as a policy entrepreneur and broker in different arenas, bargaining with various actors simultaneously in order to develop solutions acceptable to all, for example, by proposing package deals, compensation payments, etc. (Grande 1995).

Yet another institutional arrangement is the self-organization and self-regulation of actors. Self-regulation in providing a public good or common-pool resource implies that the same actors who make joint decisions about how to provide the common good also implement these decisions. The actors concerned build on the motivation, expertise, and resources of the actors immediately affected in order to provide public goods (Mayntz and Scharpf 1995). These forms of self-governance are also linked frequently to elements of hierarchy, such as the possibility of appeal to a third-party tribunal (Ostrom 1990).

Finally, markets may serve as an institutional context for the provision of public goods and common-pool resources under specific conditions, for example, when a group of actors have such high stakes that they engage in the provision. It has been increasingly claimed that quasimarkets should be created to provide public goods and common-pool resources (Le Grand 1991). Thus, many Western countries have promoted the liberalization of state monopolies in the provision of network utilities. However, a pure form of market as an institutional arrangement for providing public goods and common-pool resources can rarely be found either. Rather—as, for example, in the case of the provision of network utilities—markets are regulated at two levels: in terms of market creation to guarantee competition on the one hand, and at the level of market correction as regards 'service public' obligations to guarantee the accessibility and affordability of these services. Here—as with other institutional arrangements—two or three different types of institutional arrangements for providing public goods or common-pool resources are combined rather than existing separately as pure forms.

A further, more contextual institutional mechanism for solving collective action problems is the development of 'social capital' and trust among the actors concerned (Taylor 1996, Hardin 1993). The underlying notion is that you can trust the others involved to cooperate in the provision of a public good. Trusting others is reasonable because it is in their interest to be trustworthy in a specific situation in which a social dilemma is to be overcome. One's own trust depends on the interest of the trusted (Hardin 1999).

The instruments of command-and-control, bargaining, setting incentives, persuasion, and information are often closely linked to the institutional settings in which they are applied. Thus, command-and-control is often applied in a hierarchical context where a government authority requires the regulatees to behave in a particular way and where sanctions are applied in cases of noncompliance. However, a majority decision can also be made in favor of applying incentives. Hence, hierarchy may be linked to instruments other than command-and-control. In the institutional setting of negotiation in bi-, tri-, and multipartist settings, bargaining, together with issue-linkage, compensation payments, and compromises are the main instruments used to reach agreement, while deliberation and argument also play an important part. The negotiation process may, however, also result in the application of a strict prescription of behavior linked with the levying of fines. Incentives are set in the institutional environment of the market. Trust would be linked to information and persuasion.

4. *Implementation*

Agreement on a joint institutional solution to provide a public good does not automatically guarantee its implementation. The path from decision-making to the final realization of the measure is a long one, and many obstacles may prevent it being a smooth process (Pressman and Wildavsky 1973, Mayntz 1983, Windhoff-Héritier 1980). At international level, typically we are faced with the situation that the definition of the goods and institutional solutions for providing the public good or common-pool resource occur at international level, while implementation takes place at national level. This may raise particular problems of compliance. Thus, for example, if an international environmental agreement stipulates the extent and time-span for a reduction in CO_2 emissions, this does not mean that the signatory states, in cooperation with industry, will necessarily honor their obligations.

Research on implementation concludes that solutions for providing public goods that are redistributive in nature have little chance of being realized, whereas solutions with distributive outcomes, where all stand to win, meet with fewer difficulties. This applies as long as the redistributive solutions do not rely on hierarchy/command-and-control that prevents actors who are adversely affected by the decision from trying to change the contents of the decision in the course of implementation; and as long as those who are to benefit from the potential solution do not mobilize in order to exert pressure for a speedy implementation process (Windhoff-Héritier 1980). However, since hierarchical guidance and command-and-control can only be used to a limited extent in international governance, effective implementation depends much more on setting incentives which induce actors, over whom there is no hierarchical control, to comply with the agreed solution. Alternatively, it depends on societal mobilization where the actors profiting from the agreement press those who resist implementation into compliance. The same effect may be achieved if an agreed provision is deeply embedded in existing societal value systems. Furthermore, by using reputational mechanisms parties which do not comply are 'named and shamed' into compliance. The noncompliance of a state would result in the loss of reputation.

5. *Outcomes and Impact of the Institutional Provision of Public Goods*

A final question relates to the outcomes and impacts of specific institutional provision of a public good at international level. What are the policy outcomes of such a solution and how does it affect existing governance structures? The first question addresses the problem-solving capacity of an institutional solution, i.e., the performance of the institutional setting that has been created and the instruments applied in solving the problem. Thus, the question may be raised: does an international environmental agreement help protect endangered natural resources? Since institutional solutions are largely negotiated with a view to their potential distributional outcomes, their *de facto* distributive effects must be identified. Who gains and who loses on account of a specific institutional solution, and what does this mean in terms of the stability of the chosen solution? The second aspect relates to the fact that specific institutional solutions to public good problems across boundaries affect existing governmental and governance structures. Thus, the formal decision-making powers and the *de facto* power of national actors are changed by specific institutional solutions to problems. A case in point is the attempts of the European Union to solve collective action problems in environmental policy that have challenged the competencies of subnational actors in the member states. Similarly, the *de facto* involvement of Euro-associations in policy shaping has implied a loss of political power of national associations. Finally, the question must be raised as to how the overall structure and functions of the traditional state change in an environment where public goods are provided by forms of international governance. Does it imply the demise of the traditional state (Majone 1996, Grande and Eberlein 1999) in direction of a the regulatory state?

See also: Action, Collective; Global Governance; International and Transboundary Accords, Environmental; International Organization

Bibliography

Bates R H 1988 Contra contractarianism: Some reflections on the new institutionalism. *Politics and Society* **16**: 387–401

Bator F M 1958 The anatomy of market failure. *Quarterly Journal of Economics* **72**(August): 351–79

Buchanan J M, Stubblebine W C 1962 Externality. *Economica* **29**: 371–84

Calvert R L 1986 *Models of Imperfect Information in Politics*. Harwood Academic, Chur, Switzerland

Coase R H 1960 The problem of social cost. *Journal of Law and Economics* **3**

Cornes R, Sandler T 1996 *The Theory of Externalities, Public Goods, and Club Goods*, 2nd edn. Cambridge University Press, Cambridge, UK

Grande E 1995 Regieren im verflochtenen Verhandlungssystemen. In: Mayntz R, Scharpf F W (eds.) *Gesellschaftliche Selbstregelung und politische Steuerung*. Campus, Frankfurt, Germany

Grande E, Eberlein B 1999 Der Aufstieg des Regulierungsstaates. Zur Transformation der politischen Ökonomie der Bundesrepublik Deutschland. [The Rise of the Regulatory State. The Transformation of the Political Economy of the Federal Republic of Germany]. Presentation March 1999 at the Max-Planck-Projektgruppe Recht der Gemeinschaftsgüter, Bonn, Germany

Hardin R 1993 The street-level epistemology of trust. *Politics and Society* **21**(4): 505–29

Hardin R 1999 *Liberalism, Constitutionalism, and Democracy*. Oxford University Press, Oxford, UK

Knight J 1995 Models, interpretations and theories: Constructing explanations of institutional emergence and change. In: Knight J, Sened I (eds.) *Explaining Social Institutions*. University of Michigan Press, Ann Arbor, MI

Knight J, Johnson J 1994 Aggregation and deliberation: On the possibility of democratic legitimacy. *Political Theory* **22**: 277–96

Le Grand J 1991 *Equity and Choice: An Essay in Economics and Applied Philosophy*. Harper Collins Academic, London

Majone G 1996 *Regulating Europe*. Routledge, London

March J G, Olsen J P 1989 *Rediscovering Institutions: The Organizational Basis of Politics*. Free Press, New York

Martin L L 1994 Heterogeneity, linkage and commons problems. *Journal of Theoretical Politics* **6**(4): 473–93

Mayntz R 1983 Zur Einleitung: Probleme der Theoriebildung in der Implementationsforschung. In: Mayntz R (ed.) *Implemen-*

tation politischer Programme II: Ansätze zur Theoriebildung [The Implementation of Political Programmes II: First Attempts to Develop a Theory]. Westdeutscher Verlag, Opladen, Germany
Mayntz R 1997 *Soziologie der öffentlichen Verwaltung*, 4th edn. [Sociology in Public Administration]. Mueller Verlag, Heidelberg, Germany
Mayntz R, Scharpf F W (eds.) 1995 *Gesellschaftliche Selbstregelung und politische Steuerung* [Societal Self-regulation and Political Steering]. Campus, Frankfurt, Germany
Musgrave R A, Musgrave P B 1976 *Public Finance in Theory and Practice*, 2nd edn. McGraw-Hill, New York
Ostrom E 1990 *Governing the Commons. The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK
Pressman J L, Wildavsky A B 1973 *Implementation: How Great Expectations in Washington Are Dashed in Oakland*. University of California Press, Berkeley, CA
Pigou A C 1920 *The Economics of Welfare*. Macmillan, London
Samuelson P A 1954 The pure theory of public expenditure. *Review of Economics and Statistics* **36**: 387–9
Scharpf F W 1997 *Games Real Actors Play: Actor-centered Institutionalism in Policy Research*. Westview Press, Boulder, CO
Scharpf F W 1999 *Regieren in Europa: effektiv und demokratisch?* [*Governing Europe: Effective and Democratic?* Oxford University Press, Oxford, UK]. Campus Verlag, Frankfurt, Germany
Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA
Schmidt V A 1995 The new world order, incorporated: the rise of business and the decline of the nation-state. European University Institute working paper 5
Snidal D J 1984 *Interdependence, Regimes and International Cooperation*. University Microfilms International, Anne Arbor, MI
Taylor M 1996 Good government: on hierarchy, social capital and the limitations of rational choice theory. *The Journal of Political Philosophy* **4**(1): 1–28
Windhoff-Heritier A 1980 *Politikimplementation: Ziel und Wirklichkeit politischer Entscheidungen* [Policy Implementation: Aims and Reality in Political Decision-making]. Hain, Königstein/Taunus, Germany

A. Héritier

Public Health

Calling to mind Pasteur, his students and successors, Frederic Gorham (1921) wrote, 'In fifty short years, the world has been transformed by their discoveries.' A 'blissful optimism' surrounded public health. In the following decades, the advent of DDT, sulfa drugs, penicillin, and more recent antibiotics would endow public health doctors with unshakeable self-confidence. In contrast, the rapid expansion of health care systems during the 1960s signaled a decline for public health until the 'comeback' of infectious diseases, the prevalence of chronic illnesses, the increasing number of 'complex emergencies' (famines, wars, 'ethnic cleansing,' natural catastrophes), the growth of inequality, the development of resistance against proven drugs, the proliferation of technological risks, the economic recession and the spread of poverty—all proved the need for stronger structures and political support, especially for monitoring and controlling public health.

In the nineteenth century, sanitarians saw poverty as a product of disease. In the twentieth century, social medicine reversed the relation by affirming that destitution and exploitation bred disease. Although there is hardly any doubt that malaria, to take one example, is a cause and not an effect of underdevelopment (in Kenya and Nigeria, it costs from two to six percent of the GNP), can we not see the aggressive return of poverty in industrial societies since the 1980s or the ongoing impoverishment of the Third World as substantial evidence that increasing poverty is the major factor in current health crises?

Trends in health policy closely depend on the relation between the requirements of social justice and the objectives of political authorities. Debate about the utopia of peaceful cooperation among citizens and among nations along with the political decision to provide for national security—of which 'health security' is a part—have marked public health's history. What is original at the end of the twentieth century is that public health has been drawn off the sidelines of charity, its traditional quarters, and placed center field in world politics and nation-state strategies.

1. Health as a Local Concern

Under exceptional circumstances, public health rises to the rank of a duty for centralized states. Such has been the case when the purpose was nation building or the protection of the national territory and its population. But under usual circumstances, public health has been a local matter.

Owing to these deep local roots, the philosophical grounds for public health have been the science and art of preventing illness and of furthering the health and well-being of all thanks to an organized community effort. According to C. E. A. Winslow (1923), public health involves far more than granting health authorities police powers; it is 'not a concrete intellectual discipline, but a field of social activity.' This is still an accepted notion. In 1936, certain thinkers maintained that medicine was not a branch of science but that, if a science, it was then a social science. This position was decisive not only because it sparked debate about the history of health (as part of social history or of the history of science?) but also because it foreshadowed another debate, namely, the current controversy between public health as a social science

and epidemiology as a biomedical science. Public health is a practical science halfway between the laboratory and the public administration. Public health doctors must be, all at once, experts in the field of prevention, judges in matters of community health, and leaders. More than just clinicians, they are (in Winslow's words) scientists, administrators, and statesmen.

At present, there is no consensus about public health, about its duties, jurisdiction, and qualifications. Historically, these have seldom been clearly defined in relation to medicine, bacteriology, epidemiology, and laboratories—a sensitive question in tropical medicine. Starting in 1948, and even more as of 1965 in the United States, the gap between biomedical research (National Institute of Health), epidemiology (Centers for Disease Control), and public health (US Public Health Service) has never stopped widening.

Public health has been understood as a government duty—by revolutionaries in 1848 or the architects of state medicine in Great Britain. This is still a generally accepted idea, even in the United States despite opposition in the past not just from doctors but from hygienists themselves. Despite a chronic staff shortage, public health in the USA still provides a wide array of services coordinated by federal, state, and local governments. As the Secretary of Health, Education and Welfare declared in 1979, these health services represent 'a governmental presence at the local level.' Public health's history has been marked by swings back and forth between a narrow focus on issues and a broad public duty, between preventive and curative medicine, between hygienists and doctors.

Noting the epidemiological transition underway during the 1920s, Winslow wanted public health to restore the broken link between preventive and curative medicine, and to infuse students and practitioners with a 'spirit of prevention.' At that time, medical schools in Scandinavia and, of course, the Soviet Union offered courses in social medicine. In contrast, the 1933 declaration of the Committee of Administrative Practice of the American Public Health Association (of which Winslow had been the first president) deliberately split the field of health in two. Offices of hygiene were assigned the tasks to fight against contagious diseases and protect children's health, as reflected in the new definition of the 'basic six' duties of local health institutions (statistics, sanitation, contagious diseases, laboratories, maternal and child care, health education) and in the national plan ('Local health units for the nation') published by the aforementioned committee in 1940 and 1945.

Public health still has far to go to resolve its ambiguous assignment. Basing their argument on the variety and multiplicity of health determinants, the authors of WHO's *1999 Report* stated that public health doctors should measure the effects on health of factors lying outside the field of medicine. But just as soon, they admitted that these doctors had quite limited qualifications, jurisdiction, and credibility outside their field, and that they could better concentrate their energy in the health field.

2. *The People's War for Health*

Public health cannot deny that it leaned for a long time toward social reform—specifically, toward sociology at a time when the latter was more a form of social and political analysis than the academic discipline it has since become. Even before World War I, Newsholme in Brighton, Calmette in working class areas in northern France, Goldberger and Sydenstricker in the mining towns of South Carolina and, following the war, Stampar in the Balkans, Rubinow, Sigerist, and the Milbank Memorial Fund in the USA, René Sand a little later in Belgium, a current was proclaiming itself the heir of Virchow (or rather of the threesome Frank–Virchow–Grotjahn) in Germany, social medicine's homeland. Eighty years later, the advocates of social medicine, public health, and human rights still lay claim to this heritage.

For Warren and Sydenstricker in 1918, overcrowding and insalubrious housing were the major determinants of inequality in health. Illness and health had social, economic, as well as biological causes; Sigerist vastly widened this range of health determinants by including cultural causes. Public health—a national health program—should set the goal of universal protection so as to 'equalize social classes in relation to health' in the words of the Pasteurist Étienne Burnet in 1935. More recently, an interpretation prevailing in some quarters following Alma-Ata (1978) has sought to 'put medicine where it belongs' by seeing public health as reaching far beyond illness treatment and prevention. But how to implement this 'social medical contract,' which called for redefining democracy?

The idea of social medicine was not at all unanimously approved. Typical of differing opinions was the debate about what caused malaria: an illness bred by poverty and poor hygiene, as some claimed, or a clinical and entomological problem that, as others asserted, called for eradicating the vector rather than poverty? The very phrase 'social medicine' was cause for debate. Did it not risk making people believe public health should be based on the medical profession, whereas its only true allies were local authorities and the people?

The 1929 depression had a part in this shift toward a more sociological conception of public health. Studies in nutrition or on the links between unemployment and health took sides with a movement opposed to eugenics. Social medicine was but a clinical medicine aware of the social determinants of a patient's state of health. This awareness would inevitably call for 'socializing' health systems and opening

health care to everyone. But 'social' should not be confused with 'socialist.' The three terms 'social-socialized-socialist' tended to correspond to a political geography that might, on closer examination, turn out to be little more than a series of cultural gradients: Europe tended to be more socialist as one moved from areas using horse-power to areas using horses! The Russian 'example' must be set in this light. The Webbs, Winslow, Sigerist, like delegates at the 1978 Alma-Ata Conference, were enthusiastic about Soviet medicine (or even about community medicine in Russia since Alexander II's reign). Even though Moscow had proclaimed the doctrine of the social origins of disease, this guiding Eastern light could not outshine other sources of luminosity. In Socialist Berlin and Vienna—not to mention East Harlem—health centers and dispensaries were springing up that provided maternal and child care as well as birth control.

Once 'war Communism' came to an end, social medicine in the former USSR took a radical turn into state medicine. As conveyed to the USA under Franklin D. Roosevelt, it served, for a while, as a political alternative to the mercantile spirit of the American Medical Association. In the absence of a Federal health plan, Oklahoma and Maryland, with the backing of the Farm Security Administration, farmer organizations, labor unions, and medical cooperatives, actively defended rural interests. In its radical versions, social medicine focused on zones with high rates of poverty, malnutrition, and child mortality—areas abandoned by hospitals, which, located in cities, were too far away and too expensive. In the late 1960s once again, Tanzania gave priority to rural areas as it built a decentralized system, covering the whole country, of health stations and dispensaries, and appointed village health agents. In 1972, the proportion of funds appropriated there for rural compared with urban health services was 20:80. Eight years later, it stood at 40:60. But WHO did not reorient its primary health care strategy until 1989 in order to cope with massive rural flight toward urban areas.

3. *Alma-Ata, the Peak of Convergence*

Following World War II, social medicine experienced a golden age. The coming to power of the socialist Cooperative Commonwealth Federation in Saskatchewan, Canada in 1944, and the emergence of the New Democratic Party at the national level in 1961 opened the way for setting up, between 1962 and 1971, a national health insurance based on values of solidarity and social justice. Even in the United States, a hierarchical organization of health centers existed in New York City. In contrast, social medicine was losing wind in Europe even as it was attaining official recognition by international organizations.

Social medicine—'promotive, preventive, curative, and rehabilitative,' to borrow Sigerist's words—was crowned in Alma-Ata at the joint WHO–UNICEF conference on September 12, 1978. Potentially subversive ideas had been carefully defused by replacing certain phrases (e.g., 'health centers' with 'primary health care' or 'barefoot doctors' with 'community health workers') in the 'Health for all by the year 2000' declaration already approved by WHO's 30th General Assembly in May 1977. This 1978 Alma-Ata Conference represented a peak for the theory of convergence, itself the outcome of thawed relations between the two blocs, as the former Soviet Union sought to cooperate with the West in science and technology while still exporting its model of development toward the Third World. Shortly after the conference, WHO's regional committee for Europe would refer to this theory as 'Equity in care.' Borne by the expanding social sciences and by development theory (after all, 'Health for all by the year 2000' was intended to be an indispensable part of development), this theory stipulated that differing social systems would ultimately converge into a single 'human society' for satisfying all basic needs (food, housing, health care) after eliminating poverty. Current in the 1960s and 1970s, this ideology might be seen as a new form of the functionalism that had thrived 30 years earlier at the League of Nations. Food, housing, social insurance, paid holidays, cooperatives, leisure activities, education, and especially public health, all this would shrink the field of politics as welfare expanded. Would national administrations not be integrated into a single international system that would free Geneva from the domination of sovereign states?

The Alma-Ata conference defined the public for primary health care in terms of high morbidity and mortality rates and a socioeconomic status with risks for health. Health education, a safe water and food supply, maternal and child care (including birth control), and vaccination (against tuberculosis, diphtheria, neonatal tetanus, polio, and measles) were declared priorities along with universal access to health services and the necessity for coordinating health with education and housing. Primary health-care was to be based on the most recent scientific developments, which were to be made universally available—what Aneurin Bevan in 1946 had called 'universalizing the best.' Adopted by more than 100 states in 1981, the 1978 Alma-Ata Declaration sought to reflect criticisms both of the Western medical model and of the population explosion in poor lands. In particular, it systematically demolished the 'vertical' programs that WHO had, until then, undertaken with quite inconsistent results (success against smallpox but a relative setback in combating malaria). Based on participatory democracy, 'peripheral' services were to lie at the core of the new strategy, but were also to be integrated into 'districts' around a first referral-level hospital (140 beds, internal medicine, gynecology,

obstetrics, a laboratory and blood bank, and the capacity for performing minor surgery and providing outpatient care). The aim was not to juxtapose primary care with a still centralized system but to combine the advantages of hospitals and of health centers so as to facilitate both prevention work and access to medical and social services for a poor, often scattered, population.

To its credit, primary health care has made major, though inconsistent, progress. In Colombia, the 1500 health centers set up as of 1976 were covering 82 percent of the population by the early 1980s. Between 1978 and 1982, the percentage of the vaccinated population rose from 23 percent to 41 percent for polio, 36–71 percent for tuberculosis, 21–50 percent for measles, and 22–37 percent for diphtheria, tetanus, and pertusis. If, by 1989, 90 percent of babies less than one year old had been vaccinated against polio in Cuba, Botswana, Brazil, Nicaragua, China, Tunisia, Egypt, Korea, and Saudi Arabia, in India this rate did not reach 20 percent in the two-thirds of the poorest strata of the population, where infant mortality is the highest.

True, the methods used were sometimes counterproductive. As we know, the Expanded Program on Immunization (launched in 1974 and broadened in 1987 to include polio) had better results in places with a solid health infrastructure. In continuity with the Rockefeller Foundation's campaigns against yellow fever in Latin America, and those against 'sleeping sickness' in the Belgian Congo or French Equatorial Africa, or the campaigns against smallpox organized by WHO in association with Indian and Bengali authorities in 1962 and 1967, international organizations still preferred mass vaccination programs conducted like military campaigns. Coercion was all the more willingly used insofar as vaccinators blamed the superstitious population for any opposition.

The installation of 'community health agents' elected by villagers, would, it was believed, overcome local opposition. This lightweight health auxiliary, a sort of 'barefoot doctor,' would bridge the gap between the health system and rural dwellers. These experiments were disappointing: this community health agent often turned out to be a 'transitional last resort' who fell victim to the envy aroused by power or fell into the 'voluntary help' trap. True, public health provided certain governments (in Cuba, Sandinist Nicaragua, or Nyerere's Tanzania) with means for mobilizing the population and nation building. But five years after the signing of the Alma-Ata agreement, only experimental or small-scale projects were in the works, mainly as a function of circumstances, when a local leader showed concern for his constituents' welfare.

Under the pressure of users, prevention work in health centers often deviated toward more curative purposes. Like state medicine in the Soviet Union, social medicine in the Third World represented a limited injection of 'collective vaccine' in an 'individualistic' village setting. Except for Scandinavia, where national identities are strongly imbued with concern for health and with eugenic ideas, Europe, too, has tended to resist this type of medicine.

Primary health care was originally intended to better distribute public expenditures by avoiding congestion in big urban hospitals—an objective that the Harare Conference, organized by the WHO, reaffirmed in 1987 but that has been a resounding failure. At the Bamako meeting organized by WHO and UNICEF in 1987, African ministers of Health had to admit this, but then went on to ratify a proposal for furthering decentralization and making health care users pay part of the costs. Applied in 33 countries by 1994 (28 of them in sub-Saharan Africa), the Bamako Initiative looks like an antidote to Alma-Ata.

4. *Transnational Networks and the Territory's Revanche*

Symbolized by the publication of the 1993 World Bank report, Investing in Health, the waning of WHO—which as recently as 1973 had been said to be the 'world's health conscience'—signals a turning point. In Africa, from 60 to 70 percent of health budgets come from outside sources: non-governmental organizations (NGOs), the World Bank, or the International Monetary Fund. The World Bank's centralized approach allows little room for a local or national diagnosis. Zimbabwe, where expenditures on prevention work had increased significantly from 1982 to 1990, was, in 1991, summoned to reduce its public health investments by 30 percent in line with the Bank's recommendations for shifting funds toward 'selective' and more productive primary health care activities and toward prevention work for reducing maternal and infant mortality (such as growth-monitoring, rehydration, breast-feeding, and immunization). The Bank's recommendation to start making users pay may well reduce the utilization of health services and thus worsen the prevalence of endemic diseases such as malaria or AIDS. Another drawback is as follows: the need to obtain fast results in order to please donors keeps governments who receive aid from acquiring the ability to manage in the long term.

The World Bank's growing role in financing the health field can but amplify its intellectual role and lead to neglecting WHO's abilities and experience. Unlike Geneva, whose relations with national administrations are still limited to (weak) ministries of Health, the Bank has direct access to the centers of power (ministers of National Planning or Finances). It may even, as happened in Bangladesh, take the place of the state in heading health policy. As happened with the interventions by the Rockefeller Foundation or the Milbank Memorial Fund in France during the

1920s or in several central European and Balkan lands, the organization of public health in low-income countries comes to be based on an imported state. Since the 1980s, however, the big foundations have given up their place in the field to other non NGOs specializing in humanitarian aid. The number of NGOs rose from 1,600 in 1980 to 2,970 in 1993. A trend on this scale is a sign of the times, namely: the proportion of funds reserved for emergency actions in war zones has increased constantly since 1990. In Asia, Africa, or the Balkans, public health is, once again, seen against the backdrop of warfare.

Welfare and warfare were, indeed, intimately connected between the two world wars. Interventionism thrived with World War I and the 'Spanish flu' epidemic. As a consequence of growing government interventionism between 1914 and 1918, the first health ministries sprung up in Europe, New Zealand, and South Africa. Following the Great War, there was a general awareness of the role of public health in reconstruction and the dangers of mass epidemics (such as cholera or typhus) for political stability. World War I stimulated reformers like Dawson, who can rightfully be considered to be a forerunner to the British National Health Service (1948). But by doing away with taboos surrounding the respect for life, the war also opened wide the doors for the excesses committed by a racial form of hygienics.

Throughout the inter-World War period, a tragic dialectics prevailed in the international health order (at least in Europe), a constant swinging back and forth between the triumph of networks of transnational experts and the revanche of the territory, i.e., the nation or 'race.' Hitler's regime did not hesitate to use public health for a policy genocide. Its fight against typhus in Poland led to 'the final consolidation of the notion of a quarantine' in the Warsaw ghetto (Browning 1992). This was proof of the potential lethality of the health sciences.

As the twenty-first century begins, war is, once again, menacing health. It has even been said that 'today's armed conflicts are essentially wars on public health' (Perrin 1999). Civilian populations are the first war casualties. Since the wars in Abyssinia and Spain during the 1930s (well before Rwanda and the first Chechnya war), 80 percent of the total 20 million persons killed and 60 million wounded in various conflicts have been civilians. Three out of five have been children. The most intense war zone in Africa lies at the heart of the continent, which is also the epicenter of new diseases (Ebola) and epidemics (HIV/AIDS). But civilians fall casualty not just as a result of war's toll on health, as it sets off mass migrations and disrupts food supplies. Between 1982 and 1986, fighting destroyed more than 40 percent of Mozambique's health centers. By interrupting routine activities related, in particular, to maternal and child care, war increases mortality and the number of premature births. It also reduces vaccination coverage, not to mention its long-run effects, such as outbreaks of malaria as people seek refuge in infested areas. According to a 1996 UNICEF estimate, war-related shortages of food and medical services in Africa, added onto the stress of flight from battle zones, accounted for 20 times more deaths than actual fighting. On the one hand, the renewed threat of infectious diseases (in particular, the interaction between HIV and tuberculosis in poor countries) and, on the other, the consequences of the epidemiological transition and the aging of populations in wealthy lands, this is the 'double burden' that, according to the WHO, public health will have to carry in the twenty-first century.

Prevention work now represents a very small percentage of public health expenditures: 2.9 percent in the United States, a statistic similar to that in most other industrialized lands. However, this percentage includes only prevention activities such as screening for disease or vaccinating. It is not to be confused with 'prevention education,' such as campaigns against smoking. 'Health for all by the year 2000' is still a far cry from reality, even in wealthy nations.

The somewhat vague phrase 'health promotion,' officially used in the British Health Ministry's 1944 report, refers, all at once, to health education, taxation as a health policy instrument, and the prohibition against advertising alcoholic beverages or tobacco. This phrase was revived in the 1970s by the Canadian Health Ministry to support the thesis that a population's state of health depends less on medical progress than on its standards of living and nutrition. The chapter devoted to the fight against nicotine addiction in the 'Health for all' program has been followed up on, since it does not cost the government anything and can even, as in Australia, help pay for 'health promotion' thanks to cigarette taxes. To its advantage, it also garners broad support for public health. The keynote of the 1986 Ottawa Charter, the phrase 'health promotion' indicates how closely prevention work in wealthy countries hinges on the system of medical care. The state itself participates in this trend to make individuals responsible for health. Witness the campaigns focused on tobacco, alcohol, or certain foodstuffs with the objective of changing 'lifestyles'—a far cry from the 'old' public health which depended on the dispensary and other facilities as the basic means of education.

Consequent to discussions in the late 1970s about women's role in the sociology of health, the awareness dawned that more persons are involved in health care in the often neglected 'informal' sector than in public or private health services. Everywhere around the world, family, kin, and neighbors bear the largest burden of care for the sick. Evidence of this awareness is, on the right, the defense of the family's role and, on the left, the activism of patients' associations (especially homosexual associations in the fight against AIDS). Owing to its ability to mobilize various sorts of

actors with different opinions on anything other than health issues, thus 'civil' society is partly taking the state's place in both wealthy and poor countries. As this emergence of local actors once again proves, applying decisions in prevention work is not at all a purely administrative matter.

Have governments been put to rout? But public health raises problems directly related to the state's powers. Grounding its arguments in new bacteriological concepts, New York City's Health Service had no qualms at the start of the twentieth century about locking up a 37-year-old Irish cook ('Typhoid Mary' Mallon) for having infected 32 persons. Given the tragic affairs of blood transfusions in France and 'mad cow disease' in the UK, public health now seems to have become a fully fledged national security issue. The most recent evidence of this is the 'principle of precaution,' which—the ideology of national sovereignty tempered with expertise—would entail political decisions for strengthening public health. State security increasingly means taking into account not only political and military threats but also public health and environmental risks. As the concept of security expands into a 'shared human security,' monitoring transmissible diseases is becoming a major concern for public authorities.

On the threshold of the twenty-first century, a global public health strategy would entail relating the fight against transmissible diseases and environmentally related risks to the management of emergencies. But was this not also the issue 100 years earlier? The twentieth century started with the French–English dispute about cholera, which centered around the quarantine issue. The conflict between national independence and public health has persisted during controversial affairs all the way down through the century. Seen from Europe, the century seems to be ending with a clash about 'mad cow disease' between Great Britain, on the one hand, and Germany and France, on the other. By reviving the old Churchill myth of 'Britain alone against Europe,' the beef war (1996–2000) fully illustrates that public health, far from being an ordinary administrative matter, can turn into a fully fledged political controversy and field for government actions.

See also: Health Care Delivery Services; Health Economics; Health in Developing Countries: Cultural Concerns; Health Policy; Public Health as a Social Science

Bibliography

Browning C R 1992 Genocide and public health: German doctors and Polish Jews, 1939–1941. In: *The Path to Genocide*. Harvard University Press, Cambridge, MA

Burnet E 1935 Médecine expérimentale et médecine sociale. *Review of Hygeine* **57**: 321–42

Fee E, Brown T (eds.) 1997 *Making Medical History: The Life and Times of Henry E. Sigerist*. The Johns Hopkins University Press, Baltimore, MD

Gorham F P 1921 The history of bacteriology and its contribution to public health work. In: Porcher Ravenel M (ed.) *A Half-Century of Public Health*. American Public Health Association, New York

Perrin P 1999 The risks of military participation. In: Leaning J, Briggs S M, Chen L C (eds.) *Humanitarian Crises: The Medical and Public Health Response*. Harvard University Press, Cambridge, MA

Viseltear A J 1982 Compulsory health insurance and the definition of public health. In: Numbers R L (ed.) *Compulsory Health Insurance: The Continuing American Debate*. Greenwood Press, Westport, CT

Warren B S, Sydenstricker 1918 The relation of wages to the public health. *American Journal of Public Health* **8**: 883–7 [1999 *American Journal of Public Health* **89**: 1641–3]

Webster C (ed.) 1995 *Caring for Health: History and Diversity*. The Open University Press, Buckingham, UK

Winslow C E A 1923 *The Evolution and Significance of the Modern Public Health*. Yale University Press, New Haven, CT

L. Murard and P. Zylberman

Public Health as a Social Science

Public health can be defined as:

> the combination of science, skills and beliefs that are directed to the maintenance and improvement of the health of all people through collective or social actions. The programs, services, and institutions involved emphasize the prevention of disease and the health needs of the population as a whole. Public health activities change with changing technology and values, but the goals remain the same: to reduce the amount of disease, premature death and disability in the population (Last 1995, p. 134).

Health was defined by the World Health Organization (WHO) almost 40 years ago as a state of complete mental, physical, and social well-being, and not merely the absence of disease or infirmity (WHO 1964). From this definition, health is clearly more than just a physical state and is inextricably linked to personal well-being and potential, and incorporates other dimensions of human experience, such as quality of life, psychological health, and social well-being. Public health is different from many other fields of health as it focuses on preventing disease, illness and disability, and health promotion at the *population* level. *Populations* are groups of individuals with some similar characteristic(s) that unify them in some way, such as people living in rural areas, older people,

women, an indigenous group, or the population of a specific region or country.

Contemporary public health is particularly concerned with the sections of the population that are typically under- or poorly served by traditional health services, and consequently, most disadvantaged and vulnerable in terms of health. Importantly, public health is therefore concerned with the more *distal* (i.e., distant from the person) social, physical, economic, and environmental determinants of health, in addition to the more *proximal* determinants (i.e., closer to the person), such as risk factors and health behaviors.

Another distinction is often made between primary, secondary, and tertiary levels of prevention in public health. *Primary prevention* reduces the incidence of poor health by means such as vaccinations to prevent disease. *Secondary prevention* (or early detection) involves reducing the prevalence of poor health by shortening its duration and limiting its adverse social, psychological, and physical effects, such as different types of screening to identify a condition in the early stages. *Tertiary prevention* involves reducing the complications associated with poor health and minimizing disability and suffering, such as cardiac rehabilitation programs for patients after a heart attack. The boundaries between different levels of prevention are often indistinct and blurred, and a specific intervention strategy might simultaneously address the three different levels of prevention in different individuals.

The application of behavioral and social science theories and methods to improve health and prevent disease and disability at a population level can occur at multiple levels and in many different settings throughout society. Many public health initiatives are aimed at the individual or interpersonal level, and are delivered through the primary care setting, workplaces, and schools. Such programs typically have a low *reach*, in that only a limited number of people are involved in the programs and/or have access. However, these programs have high levels of *exposure* for the individuals in that they are quite intensive. Other programs can be organized and implemented through whole organizations or systems. Finally, community-wide or societal-level interventions may involve media, policy, or legislative action at a national or governmental level, and have a very large *reach*; however, program *exposure* is typically quite low.

The practice of public health is therefore very broad and can include the delivery of many different types of programs and services throughout the whole community. It can include efforts to reduce health disparities, increase access to health services, improve childhood rates of immunization, control the spread of HIV/AIDS, reduce uptake of smoking by young people, implement legislation and policies to improve the quality of water and food supplies, reduce the incidence of motor vehicle accidents, reduce pollution, and so on. In fact, the combined health of the population is more directly related to the quality and quantity of a country's total public health effort than it is to the society's direct investment in health services *per se*.

This article discusses the extent to which theories and models from the behavioral and social sciences inform our understanding and practice of public health. The article provides a brief overview of the major public health problems and challenges confronting developed and developing countries; the contribution made by social and behavioral factors to many of these health problems; and how social and behavioral theories and models can be usefully applied to the development, implementation, and evaluation of public health programs.

1. *Global Patterns of Health and Disease: The Major Public Health Challenges*

Disease burden includes mortality, disability, impairment, illness, and injury arising from diseases, injuries, and risk factors. It is commonly measured by calculating the combination of years of life lost due to premature mortality and equivalent 'healthy' years of life loss due to disability. The past 100 years have seen dramatic changes in health and the patterns of disease burden. Most countries have experienced an increase in *life expectancy*, with greater numbers of people expected to reach late adulthood. In the last 40 years, life expectancy has improved more than during the entire previous span of human history, increasing from 40 years to 63 years during 1950 to 1990 in developing countries (World Bank 1993). The number of children dying before five years has similarly reduced by 50 percent since 1960 (World Bank 1993).

Developed countries in particular have seen a decline in the proportion of deaths due to *communicable and infectious diseases* such as smallpox and cholera. However, this decline has occurred simultaneously with a marked increase in *noncommunicable diseases*, particularly the *chronic diseases*. In the early 1990s, the chronic diseases of ischemic heart disease and cerebrovascular disease were identified as the leading causes of mortality, together accounting for 21 percent of all deaths in the world (WHO 1999). In 1998, 81 percent of the disease burden in high-income countries was attributable to noncommunicable diseases, in particular cardiovascular disease, mental health conditions such as depression, and malignant neoplasms, with alcohol use being one of the leading identified causes of disability (WHO 1999). Other major noncommunicable causes of death include road traffic accidents, self-inflicted injuries, diabetes, and drowning.

In developing countries, current major public health problems continue to include infectious diseases, undernutrition, and complications of childbirth, as well as the increasing prevalence of chronic diseases. Maternal conditions, HIV/AIDS, and tuberculosis

Table 1
Changes in ranking for the most important causes of death from 1990 to 2020

Disorder	Ranking 1990	Ranking 2020	Change
Ischemic heart disease	1	1	0
Cerebrovascular disease	2	2	0
Lower respiratory infections	3	4	−1
Diarrheal diseases	4	11	−7
Perinatal disease	5	16	−11
Chronic obstructive pulmonary disease	6	3	+3
Tuberculosis	7	7	0
Measles	8	27	−19
Road traffic accidents	9	6	+3
Trachea, bronchus, lung cancers	10	5	−5
Malaria	11	29	−18
Self-inflicted injuries	12	10	+2
Cirrhosis of the liver	13	12	+1
Stomach cancer	14	8	+6
Diabetes mellitus	15	19	−4
Violence	16	14	–2
War injuries	20	15	+5
Liver cancer	21	13	+8
HIV	30	9	+21

Source: Murray and Lopez, *Lancet* 1997, **349** (9064): 1498–504.

are three of the major causes of disease burden for adults in developing countries (WHO 1999). Among children, diarrhea, acute respiratory infections, malaria, measles, and perinatal conditions account for 21 percent of all deaths in developing countries, compared with 1 percent in developed countries (WHO 1999). Developing regions account for 98 percent of all deaths in children younger than 15 years; the probability of death between birth and 15 years ranges from 1.1 percent in established market economies to 22 percent in sub-Saharan Africa (Murray and Lopez 1997). Importantly, many of these conditions are preventable; at least two million children die each year from disease for which vaccines are available (WHO 1999).

Accidents and injuries—both intentional and nonintentional—are a significant and neglected cause of health problems internationally, accounting for 18 percent of the global burden of disease (WHO 1999). In high-income countries, road traffic accidents and self-inflicted injuries are among the 10 leading causes of disease burden (WHO 1999). In less developed countries, road traffic accidents are the most significant cause of injuries, and war, violence, and self-inflicted injuries are among the leading 20 causes of lost years of healthy life (WHO 1999).

The Global Burden of Disease Study has projected future patterns of mortality and disease burden (Murray and Lopez 1997). Life expectancy is anticipated to increase notably in females to 90 years, with a lesser increase for males. Table 1 presents the changes in ranking for the most important causes of death. Worldwide mortality from perinatal, maternal, and nutritional conditions is expected to decline. Noncommunicable diseases are expected to account for an increasing proportion of disease burden, rising from 43 percent in 1998 to 73 percent in 2020. HIV will become one of the 10 leading causes of mortality and disability, and this epidemic represents 'an unprecedented reversal of the progress of human health' (Murray and Lopez 1997, p. 1506). The leading causes of disease burden will include ischemic heart disease, depression, road traffic accidents, cerebrovascular disease, chronic obstructive pulmonary disease, lower respiratory infections, tuberclerosis, war injuries, diarrheal disease, and HIV. Important health risk factors will include tobacco, obesity, physical inactivity, and heavy alcohol consumption. By 2020 tobacco is expected to cause more premature death and disability than any single disease, with mortality projected to increase from 3 million deaths in 1990 to 8.4 million deaths.

2. *The Behavioral and Social Epidemiology of Health and Disease*

There are a multitude of factors that can be used to explain the changing patterns of morbidity, mortality, and the spread of diseases, both globally and between regions and countries. Some of these factors are related to population growth and aging and social patterns,

while others are more related to changes in the local and global environment. An impressive amount of epidemiological evidence collected over the past 50 years has also identified the importance of a range of behavioral determinants and modifiable risk factors such as physical inactivity, smoking, and poor nutrition. The social and behavioral sciences are key disciplines underpinning public health as they have generated theories and models which provide a link between our understanding of these modifiable determinants of disease and health and strategies for disease prevention and health promotion.

The modern era for this evidence base began with the publication of the landmark study that investigated the relationship between British doctors' smoking habits and disease (Doll and Hill 1964); the first reports from the Framingham Study that identified a number of 'risk factors' for cardiovascular disease (Kannel and Gordon 1968); and the United States Surgeon General's Report on Smoking and Health (US State Surgeon General's Advisory Committee 1964). The Lalonde report (1974) and the Alameda County study (Belloc and Breslow 1972) both identified the importance of wider environmental factors such as social support, social networks, and socioeconomic factors as influences on health. In 1980, the United States Centers for Disease Control released a report estimating that 50 percent of mortality from the 10 leading causes of death in the United States could be traced to lifestyle factors (Centers for Disease Control 1980). McGinnis and Foege (1993) assessed the relative contribution of important and potentially alterable health behaviors and social factors to mortality and morbidity in the USA, and concluded that for approximately half of all deaths in 1990, the actual cause of death was attributable to health behaviors. In 1996, the report from the US Surgeon on Physical Activity and Health stated that physical activity reduced the risk of premature mortality in general, and of coronary heart disease, hypertension, colon cancer, and diabetes mellitus in particular (US Department of Health and Human Services 1996).

This rapidly expanding evidence base has been used by many countries for setting public health priorities and recommended goals and targets for addressing national public health challenges and reducing premature mortality and the burden of disease via changing lifestyle and related risk factors. Lifestyle factors targeted to influence population health include physical inactivity, diet, smoking, alcohol and other drug use, injury control, sun protective behaviors, appropriate use of medicines, immunization, sexual and reproductive health, oral hygiene, mental health, and many others.

Behavioral and social researchers have contributed to our understanding of the broad array of ecological, social, and psychological factors that influence lifestyle and health behaviors, and in turn, the etiology and pathogenesis of disease and health. In a recent edited series of monographs, Gochman (1997) provides an extensive overview of the range of environmental, cultural, family and personal factors that have been shown to influence lifestyle and health behaviors.

Moving more *upstream* from individual behaviors, there is also increasing interest in the ways in which the broader environment influences health, in particular the specific role played by social and economic factors. The Ottawa Charter (WHO 1986) recognized the importance of adopting a more *social view of health* by stressing that in order for public health programs to be effective they must incorporate the following elements: the development of personal skills; the creation of environments which are supportive of health; the refocusing of health and related services; an increase in community participation; and finally, the development of public policy which is health enhancing. The trend towards trying to understand more about the social, economic, and environmental determinants of health is probably one of the most fundamental developments to occur in mainstream public health over the past 20 years.

Recently there has been a very explicit research focus on socioeconomic *health inequalities*, social disadvantage, and poverty as major influences of health risk factors, disease, and reduced life expectancy. Poverty is undoubtedly one of the most important, single causes of preventable death, disease, and disability; the major variations and disparities in health observed both within and between countries can be explained to a large measure by social and economic disparities. Two key reports from the United Kingdom, the Black Report (Department of Health and Social Services 1980) and the Acheson Report (Acheson 1998), have emphasized the importance of these issues and ameliorative strategies.

Future social and behavioral epidemiological research into public health will need to address more adequately the multiplicity of factors contributing to health and disease. This will involve not only studying factors at the level of the individual, such as their beliefs and behavior, but also the full array of economic, social, environmental, and cultural factors as well.

3. *Theories and Models for Prevention and Change*

Social and behavioral theories and models not only help to explain health-related behavior and its determinants, they can also guide the development of interventions to influence and change health-related behavior and ultimately improve health. Theories can help understand adherence to a health regime, identify the information needed to develop an effective intervention program, provide guidelines on effective communication, and identify barriers to be overcome in implementing an intervention. While some theories

are most usefully applied to understanding the role and influence of personal-level factors, others are more useful for understanding the importance of families, social relationships, and culture, and how these can be influenced in a positive fashion to promote improvements in health.

Models of individual health behavior such as the Health Belief Model and the Theory of Reasoned Action/Planned Behavior, as well as models of interpersonal behavior such as Social Cognitive Theory, have been used widely in public health and health promotion over the past 20 years and are reviewed in various publications (e.g., Glanz et al. 1997). Staging models, such as the Transtheoretical Model, have often been used to improve the potency of health promotion programs (e.g., see Glanz et al. 1997). A combination of models or theories can often assist in the development of more individually tailored and targeted strategies (see *Health Behavior: Psychosocial Theories*).

While these theories and models focus primarily on achieving changes in health behaviors, other models are more concerned with communication and diffusion of effective programs, so as to benefit the health of a larger population. McGuire's communication–behavior change model and Rogers' diffusion of innovations model are examples of these types of theories (see Glanz et al. 1997). Health behaviors, and the broad array of environmental and social influences on behavior, are usually so complex that models drawing on a number of individual theories to help understand a specific problem in a particular setting or context are usually required to develop, implement, and evaluate effective public health responses. For example, Green and Kreuter's PRECEDE–PROCEED model, social marketing theory, and ecological planning approaches are models and approaches that are used commonly in public health and health promotion (see Glanz et al. 1997).

However, within the behavioral sciences, much less attention has been given to models or theories that understand change strategies and programs within the broader environment, groups, organizations, and whole communities. This is despite increasing research evidence indicating that social and environmental factors are critically important in understanding individual attitudes, behavior, and health. The design of programs to reach populations requires an understanding of how social systems operate, how change occurs within and among systems, and how large-scale changes influence people's health behavior and health more generally. A combination of theories and models of individual behavior can be used to shift the focus of intervention strategies beyond a clinical or individual level to a population level by their application via a range of settings in the community, such as the workplace and primary health care (Oldenburg 1994). This increases the reach of intervention programs beyond a relatively small number of individuals. Many of the social and behavioral theories or models that have been used extensively to understand and describe change at an individual or small group level can also inform our understanding of the change and diffusion process at an organizational or population-based level.

4. *Wider Implementation of the Social and Behavioral Science Evidence Base in Public Health*

With the development, implementation, and evaluation of a series of major community-based cardiovascular prevention trials during the 1970s and 1980s, widespread use was made of multilevel intervention or change strategies underpinned by behavioral and social science theories and research for the first time. Importantly, these strategies did not focus exclusively on change in the individual. Instead, preventive strategies using multiple approaches were directed at the media, legislation, and the use of restrictive policies, and involved specific settings such as schools, the workplace, health care settings, and other key groups in the community, with the aim of reducing population rates of risk behaviors, morbidity, and mortality.

The North Karelia Project in Finland (Puska et al. 1985) was initiated in response to concern that the community, at that time, had the highest risk of heart attack worldwide. Strategies included tobacco taxation and restriction, televised instruction in skills for nonsmoking and vegetable growing, and extensive organization and networking to build an education and advocacy organization. After 10 years, results indicated significant reductions in smoking, blood pressure, and cholesterol, and a 24 percent reduction in coronary heart disease mortality among middle-aged males. The Stanford Three Community Study (Farquhar and Maccoby 1977) demonstrated the feasibility and effectiveness of mass media-based educational campaigns and achieved significant reductions in cholesterol and fat intake. The Minnesota Heart Health Project (Mittelmark et al. 1986), the Pawtucket Heart Health Program (Lefebvre et al. 1987), and the Stanford Five-City Project (Farquhar et al. 1990) used interventions aimed at raising public awareness of risk factors for coronary heart disease, and changing risk behaviors through education of health professionals and environmental change programs, such as grocery store and restaurant labeling.

Worksite intervention trials, such as the Working Well Trial (Sorensen et al. 1996) and an Australian study with ambulance workers (Gomel et al. 1993), have utilized interventions aimed at favorably altering social and psychological factors known to influence health-related behaviors, including increased awareness of benefits, increased self-efficacy, increased social support, and reduced perception of barriers. Strategies in this setting can include promotion of nonsmoking policies, increasing the availability of healthy foods in

canteens and vending machines, group education sessions, and access to exercise facilities. School-based interventions provide an established setting for reaching children, adolescents, and their families to address age-pertinent health issues such as smoking, teenage pregnancy, and substance use. Strategies have targeted social reinforcement, school food service, health-related knowledge and attitudes, local ordinances, peer leaders, and social norms.

5. *Evaluation of Public Health Intervention Trials*

In evaluating the effectiveness of such trials, Rose (1992) states that it is important to acknowledge the differences between clinical and more population-based approaches. Clinical approaches, while more likely to result in greater mean changes in individuals, are less significant at a population level because fewer individuals are involved. In contrast, the small mean changes achieved by population-based approaches, if applied to a large number of people, have the potential to impact much more significantly on population rates of disease and health.

One of the difficulties in evaluating population-based approaches is the length of time required to demonstrate changes in population rates of health outcomes. While researchers can readily measure changes in a behavioral outcome—such as the rate of smoking cessation—following a widespread and well-conducted program, the time required for such behavioral changes to impact upon long-term morbidity and mortality—such as lung cancer—is much more problematic. Nevertheless, many of the above identified behavior-based trials demonstrate the value of social and behavioral theories and the feasibility of activating communities in the pursuit of health.

6. *Diffusion and Dissemination of the Evidence Base*

The ultimate value of this research must be determined by the extent to which the knowledge generated is widely disseminated, adopted, implemented, and maintained by 'users,' and ultimately, by its impact on systems and policy at a regional, state, and/or national level (i.e., *institutionalization*). While considerable effort and resources have been devoted to developing effective interventions, relatively little attention has been given to developing and researching effective methods for the *diffusion* of their use. It is important to acknowledge that the availability of relevant research findings does not in itself guarantee good practice. Such knowledge transfer involves the development of formal research policies, formalized organizational structural support, appropriate and targeted funding, formal monitoring of activity, and ongoing training.

In essence, achieving the effective diffusion of innovations, both within the general community and in organizational settings, involves change, and the change principles which underpin the diffusion process are not so different to those previously identified for understanding change at the individual, organizational, or community levels (see Oldenburg et al. 1997). At the level of the individual, family, or small group, uptake of a health promotion innovation typically involves changes in behaviors or lifestyle practices which will either reduce risk factors or promote health. At an organizational level, such as the workplace, school, or the health care setting, successful uptake of an innovation may require the introduction of particular programs or services, changes in policies or regulations, or changes in the roles and functions of particular personnel. At a broader community-wide or even societal level, the change process can involve the use of the media and changes in governmental policies and legislation, as well as coordination of a variety of other initiatives at the individual and the settings level.

However, in considering the principles that underpin the diffusion of innovations at a population level, further complexity arises from the need to consider change occurring at multiple levels, across many different settings, and resulting from the use of many different change strategies. This requires the application of multiple models and theories in order to develop frameworks with sufficient explanatory power.

7. *Looking to the Future*

It is difficult to be certain about the public health challenges that lie ahead in this new millennium. In all countries, there is a need to reduce preventable disease and premature death, reduce health inequalities, and develop sustainable public health programs.

Much of the social and behavioral science research that has been conducted with individuals is still relevant in the field of public health, as are many of the models and theories that have been developed and tested over the past 30 years. To move to a population-based focus, we need innovative means of implementing such knowledge and methodologies so as to increase the reach of such programs to more people, particularly those individuals and sections of society who are most disadvantaged.

Adopting a preventive approach to health, acknowledging and acting on the importance of social and environmental factors as major determinants of population health, and reducing health inequalities provides a substantial challenge to public health in the coming years. Furthermore, limited resources place an added emphasis on evaluating the affordability and sustainability of public health interventions. Public

health should be at the forefront of developing and researching effective intervention methods and strategies for current and future health challenges, with global sharing of information in terms of monitoring health outcomes and development of the evidence base. However, this research should not focus solely on developing strategies and testing their efficacy and effectiveness. There is an urgent need for research that examines larger-scale implementation and diffusion of these strategies to whole communities and populations, through the innovative use of a variety of psychological, social, and environmental intervention methods, including evidence from the social and behavioral sciences.

See also: Behavioral Medicine; Epidemiology, Social; Health Behavior: Psychosocial Theories; Health Behaviors; Health Education and Health Promotion; Health Interventions: Community-based; Health Psychology; Mental Health: Community Interventions; Public Health; Socioeconomic Status and Health; Socioeconomic Status and Health Care

Bibliography

Acheson D 1998 *Independent Inquiry into Inequalities in Health*. The Stationary Office, London

Belloc N B, Breslow L 1972 Relationship of physical health status and health practices. *Preventive Medicine* **1**: 409–12

Centers for Disease Control 1980 *Ten Leading Causes of Death in the United States 1977*. US Government Printing Office, Washington, DC

Department of Health and Social Services 1980 Inequalities in Health: Report of a Research Working Group Chaired by Sir Douglas Black. DHHS, London

Doll R, Hill A B 1964 Mortality in relation to smoking: Ten years' observations of British doctors. *British Medical Journal* Vol.(i): 1399–410

Farquhar J W, Wood P D, Breitrose H 1977 Community education for cardiovascular health. *Lancet* **1**: 192–5

Farquhar J W, Fortmann S P, Flora J A, Taylor C B, Haskell W L, Williams P T, Maccoby N, Wood P D 1990 Effects of community-wide education on cardiovascular risk factors: The Stanford Five-City Project. *Journal of the American Medical Association* **264**: 359–65

Glanz K, Lewis F M, Rimer B K (eds.) 1997 *Health Behavior and Health Education: Theory, Research, and Practice*. Jossey-Bass, San Francisco, CA

Gochman D S (ed.) 1997 *Handbook of Health Behavior Research*. Plenum, New York

Gomel M, Oldenburg B, Simpson J, Owen N 1993 Worksite cardiovascular risk reduction: A randomised trial of health risk assessment, education, counselling, and incentives. *American Journal of Public Health* **83**: 1231–8

Kannel W B, Gordon T 1968 *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*. US Department of Health Education and Welfare, Washington, DC

Lalonde M 1974 *A New Perspective on the Health of Canadians: A Working Document*. Government of Canada, Ottawa, ON

Last J M 1995 *A Dictionary of Epidemiology*. Oxford University Press, New York

Lefebvre R C, Lasater T M, Carleton R A, Peterson G 1987 Theory and delivery of health programming in the community: The Pawtucket Heart Health Program. *Preventive Medicine* **16**: 95

McGinnis J M, Foege W H 1993 Actual causes of death in the United States. *Journal of the American Medical Association* **270**: 2207–12

Mittelmark M B, Luepker R V, Jacobs D R 1986 Community-wide prevention of cardiovascular disease: Education strategies of the Minnesota Heart Health Program. *Preventive Medicine* **15**: 1–17

Murray C J L, Lopez A D 1997 Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *Lancet* **349**(9064): 1498–504

Oldenburg B 1994 Promotion of health: integrating the clinical and public health approaches. In: Maes S, Leventhal H, Johnston M (eds.) *International Review of Health Psychology*. Wiley, New York pp. 121–44

Oldenburg B, Hardcastle D, Kok G 1997 Diffusion of innovations. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research, and Practice*. Jossey Bass, San Francisco, CA

Puska P, Nissinen A, Tuomilehto J, Salonen J T, Koskela K, McAlister A, Kotke T E, Maccoby N, Farquhar J W 1985 The community-based strategy to prevent coronary heart disease: Conclusions from ten years of the North Karelia Project. *Annual Review of Public Health* **6**: 147–93

Rose G 1992 *The Strategy of Preventive Medicine*. Oxford University Press, Oxford, UK

Sorensen G, Thompson B, Glanz K, Feng Z, Kinne S, DiClemente C, Emmons K, Heimendinger J, Probart C, Lichtenstein E 1996 Worksite based cancer prevention: primary results from the Working Well Trial. *American Journal of Public Health* **86**: 939–47

US Department of Health and Human Services 1996 *Physical Activity and Health: A Report of the Surgeon General*. US Department of Health, Washington, DC

US State Surgeon General's Advisory Committee 1964 *Smoking and Health*. US Department of Health, Washington, DC

The World Bank 1993 *World Development Report 1993: Investing in Health*. Oxford University Press, New York

World Health Organization 1986 The Ottawa Charter for Health Promotion. *Health Promotion International* **1**(4): iii–v

World Health Organization 1964 *World Health Organization Constitution*. WHO, Geneva

World Health Organization 1999 *World Health Report: Making a Difference*. WHO, Geneva

B. Oldenburg

Public Interest

Scholars debate the question of the public interest somewhat less than they once did, but it remains an important issue. As is true of other central concepts, the passage of time does not always lead to an increase in acuity, although it does lead to changes in emphasis

(Friedrich 1962, Lewin 1991). Much that might be considered under the rubric of the public interest, moreover, continues to be discussed in terms of associated ideas such as justice, responsibility, and community. This ambiguity complicates a topic that is itself elusive.

1. *What is the Public Interest, Conceptually?*

The issue of the public interest is best organized under four headings. First is the formal or conceptual question. This includes problems such as the meaning of public as opposed to private, self or common interest, and the meaning of interests as distinguished from other goods. 'Public' sometimes is taken as anything that is not private, so that regard for the public interest means any regard for interests or groups that are not strictly or exclusively one's own (Lewin 1991). Less broadly, a public often is said to be those affected by private groups, in the sense that people who are neither labor nor management are affected by strikes. Such publics, moreover, sometimes are taken to refer to those potentially and not just actually affected as, say, any potential passenger might be inconvenienced by a strike of air traffic controllers.

(Someone can, therefore, have both a private and a public interest in the same matter, and 'self' interest need not be identical to private interest.) In this sense we are all members of many publics. Often, however, the public interest is taken to be what is in the interests of 'all' of us, presumably as citizens within some boundary, although the questions of scope and authority are often curiously undiscussed. There are not many publics in this view but, finally, one public (Friedrich 1962, Barth 1992). The public is the group that makes law. Of course, its interests may be the sum of those of the many affected publics into which it can be divided.

Many discussions of the 'public' in these terms have in mind economic matters, or the point of view of academic economics (Friedrich 1962). A public is seen as an affected or lawmaking group of individuals that seeks satisfaction in economically calculable terms. Publics want things that market failures cause them to receive inadequately. It is serving such publics that defines the responsibilities of professionals and of businesses 'affected with the public interest.' One strength of this conception is that since the 1950s it has brought to bear an increasingly sophisticated understanding of economically rational action, both strategic and straightforward.

There are, however, limits to this perspective. They are evident in the narrowness of the economic view of the public when it is compared to other notions of what is potentially common in common goods. For, what is common is not limited to what each independently and equally can use or enjoy, as we use markets. The common also can involve things of which we are only parts, sometimes equal or identical and sometimes not, as a face is common to its different parts or a team or family common to its members. The ties that characterize what is common, moreover, are not limited to what covers or blankets us universally, but include attaching, placing, fitting, ordering, attracting and other links integral to images of politics from Plato through Hegel. Practical and economically based notions of the public interest will, of course, often take these distinctions for granted, but deeper exploration of the concept of the public interest cannot.

Just as 'public' is often narrower than common, so too are 'interests' conceptually different from goods more broadly conceived. Discussions of the public interest often refer to Adam Smith and, therefore, to a view of what is good as what satisfies desire in the somewhat flat, calculable, and exchangeable manner of post-Hobbesian or Lockean bourgeois life. Interests are linked to what is merely interesting, not to goods as objects of passion, longing, pride or nobility or to the character and institutions that support and direct these. Public interests are what several of us happen to desire, or means to satisfaction that we generally want. Part of the conceptual task in understanding the public interest is to consider the different ways of experiencing goods so that we can grasp how interests are goods conceived in a limited manner that is linked to the premises of liberal individualism.

2. *What is the Public Interest, Substantively?*

A second question about the public interest is what it is substantively. This issue involves both scope and content. Some scholars seek to limit questions of the public interest only to certain kinds of public goods. Others use it to stand for national goals, purposes or goods more comprehensively. In practice, too, speaking of the public interest generally differs from considering the public's interest in assuring the provision to all of some specific good such as telephone service.

When the scope of the public interest is limited to economic concerns, so too is the content of what is said to comprise it; for some the public interest does not even include protecting rights. To limit the concept in this way risks confusion when 'public interest' retains its broad use in practice and even in theory; to use it broadly, however, seems to involve every question of human purpose and to require some other concept to take the place of the idea used narrowly.

Just as our understanding of the public interest conceived narrowly has benefited from developments in economic understanding, so too has substantive discussion about the public interest conceived broadly undergone notable developments since the 1950s. The triumph of liberal democracies in the Cold War has directed the concerns of the left away from socialism

simply to expanding the place of egalitarianism and community within liberal democracy itself. A wish among those who reject the revivified concept of natural rights to nonetheless ground the place of rights and defend them from simple utilitarian calculations has become widespread. Concern with the place of virtue and nobility in liberal democracies and, indeed, concern with these phenomena simply has been renewed (Rawls 1971, Strauss 1952). All these trends have enriched and re-energized discussions of the public interest.

3. *Can the Public Interest be Implemented?*

The third question is how the public interest might be implemented concretely and used as a guide for action. Some thinkers argue that the 'public interest' is too vague to be used other than rhetorically or that it is merely a name that sums up a country's political and economic life at some moment frozen in analytical time (Friedrich 1962). More charitably, one might concede that it means a certain level of economic satisfaction and political freedom but that these are best reached through private and political competition, that is, when no one aims at them directly.

When we conceive the public interest broadly, however, it is not always so difficult to discern how it might guide action. Constitutions, for example, cannot be formed well without explicit attention to more than private concerns: Charles Beard's worm's eye view of the framing of the American Constitution has been debunked by more thoughtfully elevated judgments (Kesler 2000, Mansfield 1991) Once constitutionally secured, individual rights and liberties must be exercised individually, but steps to protect these rights publicly must be attended to explicitly by jurists. National defense is a public interest that guides specific legislative and executive action.

To say that the public good often receives explicit attention is not to say that it goes without debate, but to call it debatable is not to call it meaningless. The deliberations and result of the US' constitutional convention prove the contrary. Similarly, if we take our private interest to include moral and intellectual excellence and not only economic opportunity and satisfaction we would not say that the disputability of these ends makes them pointless. On the contrary, their subtlety may make our attention to them all the more intense.

If we conceive the public interest more narrowly as assuring some equitable or safe provision and distribution of certain goods then it is again clear that the public interest can be a concrete guide. Are expanding access to communications and increasing public health served better by explicit government regulation with a defined public interest in mind of such and such a level of drug safety and computer ownership, or are they better served by discovery and application fueled largely by private desires for wealth, knowledge, and mastery, with little regulatory adjustment of the private application of resources? Is improved education secured better by unfunded parental choice, by funded choice, or by public provision or standards? Explicit attention to health, education and responsible choice as public interests seems necessary to help secure their sufficient provision and fair distribution, if only to ward off excessive regulation and allow private efforts to flourish.

4. *Is the Public Interest Followed?*

The fourth question is whether anyone in practice pays attention to the public interest or whether private interest in fact dominates even in public matters. It is argued that the actions of voters, representatives, and bureaucrats all can be explained as attempts to enhance private goods; this analysis can be expanded to professionals, journalists, and intellectuals who do public talking, if not public business, for a living. Even if the public interest could guide us, in this view it does not.

Some recent research in the US and Western Europe indicates that voters do not simply or usually vote their immediately selfish economic concerns, that the behavior of representatives cannot always or primarily be explained by their wish for re-election, that bureaucrats do not always want to increase their agency's size, and that groups and individuals sometimes do seek benefits for publics from which they and their members do not receive special advantages (Lewin 1991, Wilson 1989). Of course, a link between private and public interest is to be expected and largely desired in liberal democratic regimes, for the responsibility we encourage in citizens and officials is meant to benefit not just others but oneself as well (Kesler 2000).

Obviously, whether political action actually attempts to serve public interests requires continued study. Only the excessively naive would believe that unbridled devotion to a clearly conceived common good directs each and every political choice. But only the false or foolish sophisticate would believe the public interest to be as fleeting and airy as the summer snow.

See also: Communitarianism: Political Theory; Democracy; Environment and Common Property Institutions; Interest: History of the Concept; Political Culture; Political Representation; Political Sociology; Public Interest Law; Public Opinion: Political Aspects; Public Reason; Representation: History of the Problem

Bibliography

Antitrust and the Economics of the Public Interest 1997 Symposium: Economics of Antitrust Enforcement. *Antitrust Bulletin* **42**(Spring)

Barth T J 1992 The public interest and administrative discretion. *American Review of Public Administration* **22**: 289–300
Blitz M 2000 Liberal freedom and responsibility. In: Boxx W T, Quinlaven G (eds.) *Public Morality, Civic Virtue and the Problem of Liberalism*. Eerdmans, Grand Rapids, MI
Box R C 1992 The administrator as trustee of the public interest: Normative ideals and daily practice. *Administration & Society* **24**: 323–45
Dewey J 1927 *The Public and Its Problems*. Holt, New York
Friedrich C J (ed.) 1962 *The Public Interest*. Atherton, New York
Goodin R E 1996 Institutionalizing the public interest: the defense of deadlock and beyond. *American Political Science Review* **90**: 331–43
Kesler C (ed.) 2000 *Alexander Hamilton, John Jay and James Madison, The Federalist Papers*. Random House, New York
Lewin L 1991 *Self Interest and Public Interest in Western Politics* (trans. Lavery D). Oxford University Press, Oxford, UK
Mansfield Jr. H C 1991 *America's Constitutional Soul*. Johns Hopkins University Press, Baltimore, MD
Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA
Strauss L 1952 *Natural Right and History*. University of Chicago Press, Chicago
Wilson J Q 1989 *Bureaucracy*. Basic Books, New York

M. Blitz

Public Interest Law

Public interest law is a branch of legal practice committed to representing traditionally unrepresented groups and interests. It went through a period of rapid growth in the USA in the 1970s and at the beginning of the twenty-first century is an important, though small component of the legal system in the USA and many other countries. Public interest law is often advanced through test case litigation designed to not merely effect the outcome of a particular dispute, but to influence public policy. In tandem with the litigative work, public interest lawyers frequently engage in advocacy before government administrative agencies and legislative bodies.

Public interest law can be distinguished from legal services that focus on representing the personal legal needs of individual clients. The outcome of a public interest lawsuit is likely to reshape public policy or affect the interests of a substantial population. The class action device is sometimes used in such cases, although similar results can be had by representing an organization or, in appropriate cases, even by representation of a single individual on a matter where substantial issues of legal principle are involved. Moreover, a lawsuit in which a doctrinal innovation is made is likely to be followed by strategic efforts to elaborate, or expand, the new principles. For example, the principle that an involuntarily confined mental patient is entitled to receive adequate treatment, which has far-reaching consequences for mental health delivery systems, was established in the case of an individual mental patient (Rouse vs. Cameron).

Subsequently, the principle was developed in class actions in other jurisdictions, and expanded to cover mentally retarded people. The right to refuse treatment was explored in other settings. Legislative, budgetary, and administrative activities also were catalyzed by the judicial decisions.

During the last third of the twentieth century there were significant shifts in public policy in the USA in critical areas such as race relations, environmental protection, and womens rights. In each of these shifts in public policy, public interest lawyers played an important role in tandem with citizen organizing and political action. When legislatures and other agencies of government were closed to citizen advocates, courts were often available. Moreover, the courts frequently, through their decisions in public interest cases, precipitated action in administrative agencies, legislatures, and the executive branch, which would otherwise not have come about.

1. *Antecedents*

Although the name public interest law was not coined until 1969, its antecedents existed as a set of practices for much of the twentieth century. Historically, public interest law in the USA grew out of four movements, some of which were concerned with broad societal issues, others of which were more concerned with the individual's access to basic legal representation.

The first of the precursors to modern public interest law is the decades-long campaign to redress racial discrimination and racist institutional structures. In this arena, the National Association for the Advancement of Colored People (NAACP) and its Legal Defense and Education Fund (LDEF) were the paradigm. The LDEF was established in the mid-1930s, and a brilliant group of civil rights lawyers led by Thurgood Marshall (later a Supreme Court Justice) devised a strategy to bring a succession of test cases through the federal courts to the Supreme Court that would gradually undermine the legal foundations for racial segregation. The sequence of cases led to the landmark decision in 1954 of Brown vs. Board of Education, which ended formal, legally mandated segregation in American public schools. In subsequent decades, a series of lawsuits followed to apply the principles of Brown in specific situations and to carry them forward into other areas where racial discrimination was practiced such as employment and housing.

The second pillar upon which public interest law rests is the movement for freedom of expression centered around the American Civil Liberties Union, which grew out of the National Civil Liberties

Bureau's defense of anti-war advocates during the World War I years. Third, modern-day consumer advocacy, which is central to much public interest law and is embodied in the various organizations founded by Ralph Nader, traces its lineage back to Supreme Court Justice Louis Brandeis and progressive-era anti-trust legislation. Those years, at the start of the twentieth century, saw the emergence of the idea that the American government had a right, and often a duty, to regulate the marketplace and ameliorate some of the injustices thrown up by a *laissez-faire* economic structure. Brandeis developed the idea that a 'private attorney general' could go to court to enforce the law if government failed to act.

The last is the legal aid movement that began in the late nineteenth century in New York as a way of providing charity-based legal representation for people too poor to pay private attorneys. Legal aid expanded in the post World War I period in the wake of Reginald Heber Smith's 1919 book *Justice and the Poor*. In the 1960s, as part of President Lyndon Johnson's War on Poverty, a major federally funded legal services program supplemented legal aid. With the expansion of federal anti-poverty programs in the 1960s, many lawyers within legal services developed public interest law strategies and highly specialized knowledge, litigating on such issues as welfare policy, health care for the poor, and the availability of subsidized housing. They used class actions and other litigation strategies to achieve important policy shifts.

Since the 1960s, public interest law has broadened its scope to include such areas as environmental law, an area where public interest advocacy groups have been particularly successful in building a framework of effective judicial precedents, federal statutes, and regulations. There have also been public interest law efforts addressing gender-based and sexual preference-based discrimination, and legal work on behalf of prisoners. In the 1970s, public interest groups litigated suits that delayed the construction of the Trans-Alaska oil pipeline and forced re-design; this challenged, and eventually halted, the domestic use of the pesticide DDT; and that won stricter oversight of nuclear power plants. In the following decades, public interest law centers specializing in environmental issues have developed into large membership organizations with increasingly sophisticated technical expertise, and a range of coordinated strategies involving litigation, administrative advocacy, grassroots organizing, political advocacy, and public relations.

2. *1969: The 'Public Interest Law Firm'*

Sarat and Scheingold (1998) describe public interest law as using 'legal skills to further a vision of the good society.' Most legal work is, in theory, value-neutral, and the attorney representing a fee-paying client without regard to the merits of his claim. Public interest law is, at its core, a systematic effort to bring new voices into policy formulation, and to use the law to attain certain public policy objectives, such as the abolition of legally authorized segregation or the implementation of better environmental protection. Lawyers who accept lower levels of remuneration in exchange for having the freedom to pursue social goals and to represent only clients whose causes they believe have genuine merit practice it.

Most public interest lawyers in the USA work in non-profit public interest law firms, granted tax-exempt status by the Internal Revenue Service on the condition that they are paid a salary by the firm instead of fees by their clients, that they take on cases broadly defined as having implications for the public good, and that their firm operates under a Board of Trustees. The firms' efforts are supplemented by the work of lawyers in private firms who are willing to take on a limited number of *pro-bono* cases; and by law students who receive academic credit for their work in law school clinics.

The Internal Revenue Service's recognition of the charitable status of public interest law firms has allowed foundations to channel money to these centers. Philanthropic organizations such as the Ford Foundation, the Rockefeller Brothers Fund and the Edna McConnell Clark Foundation donated millions of dollars to public interest law groups. Foundations were particularly important in providing seed money to the public interest movement in the 1970s. Subsequently, many of them became reluctant to fund public interest law activities, especially litigation, because of the controversy such activities engender. Nonetheless, many progressive foundations have continued to support regional organizations to provide funds for local and regional public interest law activities.

The first of the new generation of public interest law firms, and one representative of the genus, is the Center for Law and Social Policy (CLASP) established in 1969. It was founded by four attorneys (including the author) experienced in the distinctive legal culture of Washington, DC. They had backgrounds in corporate law practice and the executive and legislative branches of the federal government. They believed that the effective legal advocacy on behalf of corporations often subverted the mechanisms of government; that the process was impoverished by the absence of effective advocacy on behalf of citizen interests consumers, environmentalists, the poor, the victims of discrimination and others. They organized a bipartisan group of eminent bar leaders, led by Arthur Goldberg, a former Supreme Court Justice and cabinet member, to serve as a Board of Trustees for CLASP. Within a year, CLASP had won landmark victories in two major environmental lawsuits. Early CLASP

clients included Ralph Nader, the Wilderness Society and the American Public Health Association.

The CLASP model inspired other public interest law firms, and over time, spun off public interest law groups which served other constituencies among them the Mental Health Law Project (renamed the David L. Bazelon Center for Law and Mental Health) and the National Womens Law Center. It also was a co-founder of the Project on Corporate Responsibility which pioneered in the use of stockholder initiatives to promote socially significant issues, such as environmental protection, consumer interests, and civil rights, and which sponsored Campaign GM in the early 1970s, an effort to elect a slate of independent directors to the Board of General Motors.

Historically, public interest law centers have focused on issues that can be broadly defined as progressive or liberal, acting as a counterbalancing force to corporate power and serving the needs of the poor and powerless. Public interest lawyers frequently provided a counterweight to corporate advocacy. As such, it was essential that they render top-flight legal services and that they be backed up by the specialized technical expertise that was vital in many complicated cases.

Progressive public interest law thrives on both publicity and broader citizen participation. Often litigation launched on behalf of individuals or groups by public interest lawyers is one element in broad campaigns involving political lobbying, sophisticated media campaigns, and sometimes direct action.

3. *The Evolution of Public Interest Law*

In 1961, Rachel Carson's book *Silent Spring* was published, putting environmental pollution firmly into American public consciousness. Five years later, Ralph Nader's book *Unsafe at Any Speed* (1966) did the same for unsafe automobiles. Coming at a time of dramatic social change and political unrest, the two books helped create two new constituencies in American politics: the environmental and the consumer. In the following decades, both issues have joined civil rights, women's rights, and poverty issues at the core of public interest law. When public interest lawyers brought these issues to the courts, they were framed in a way that required decision by judges and other policy makers.

The 1970s were a decade of growth for public interest law, and by 1975 there were more than 90 public interest law firms employing over 600 lawyers. But, although the number of progressive public interest law firms continued to increase in the 1980s, and to spread beyond their traditional bases of the Northeastern USA and California, they did so at a slower rate.

The political climate during the Reagan and Bush presidencies (1981–93) was hostile to many of the constituencies represented by such groups; and their administrations made a concerted effort to counter public interest law activity, through cutting the federal legal services program, limiting government grants to 'advocacy' organizations and appointing federal judges hostile to the interests represented by public interest lawyers.

That the public interest movement weathered these years as well as it did, and was able to enter a new period of growth in the 1990s, reflects the sophisticated fund-raising and publicity techniques it developed, as well as the institutionalization of this branch of law that had been regarded only a decade earlier as a somewhat marginal voice of conscience to the legal profession. Indeed, so successful have some Washington public interest firms become that critics argue that they have become policy 'insiders' and lost touch with the grassroots groups, which they originally set out to serve. As the courts have become less receptive to public interest litigation, public interest lawyers have turned more often to other forms to legislative advocacy, coalition building, initiatives and referenda, and public education. Litigation remains an important dimension of public interest law activity, but it is only one element in developing advocacy strategy.

3.1 *The Different Forms of Public Interest Law Practice*

Although the main vehicle for public interest law is the nonprofit public interest law firm in which salaried lawyers are employed full-time to handle public interest cases, other forms also exist. Many public interest law firms, such as the Natural Resources Defense Council, have become membership organizations and rely heavily on dues. Some began as affiliates of existing membership organizations, such as the Sierra Club Legal Defense Fund (later spun off and renamed the Earth Justice Legal Defense Fund). Some, like the Institute for Public Representation at Georgetown Law Center, function as clinics in law schools. Some, like the Environmental Defense Fund, integrate scientists and other experts into their staff operations.

Some private lawyers have established public interest practices, without tax exemption, and rely on court-ordered attorney's fee awards in successful cases. The private public interest law firm may also combine public interest law with traditional, financially more profitable cases on behalf of fee-paying clients. What is characteristic of the public interest law firm in its various permutations is the sense of social mission and the prominent role that lawyers play in establishing policy, even in areas where technical expertise is critically important.

Many large corporate firms encourage their lawyers to do a certain amount of '*pro bono* work,' although conflict of interest considerations disqualify them from some kinds of cases. Some firms, such as Skadden, Arps in New York, have set up their own public

interest fellowship program. The Skadden Fellowship Foundation has awarded 25 such fellowships each year since 1988 to newly graduating law students who are paid to work on public interest matters for a year in connection with a public interest law firm.

Federally funded legal service offices, although facing financial cuts and stringent restrictions on their advocacy strategies (e.g., no class actions), still sometimes, though increasingly rarely, litigate public interest cases.

3.2 New Forms of Financial Support

Many law schools defer or cancel the tuition loans owed by graduates who choose public interest law as a career, and have also set up public interest legal foundations, which raise money among students and alumni to support public interest law work by recent graduates and student interns. The National Association of Public Interest Law, based in Washington, DC, awards student fellowships in the field, and over 120 law schools now belong to the Public Service Law Network (PSL Net), which steers students toward work in approximately 10,000 law-related public service organizations throughout the world. In the case of large membership-based groups such as the Environmental Defense Fund and the Natural Resources Defense Council money is also raised through membership dues. Government grants are sometimes awarded for specific projects. Some regulatory laws, such as the Clean Water Act, have built-in provisions that allow public interest law firms to be awarded fees in successful cases to be paid by the losing party.

3.3 The Divergent Politics

Traditionally, public interest law has promoted progressive causes and the rights of the powerless. But a parallel development has been the development of conservative public interest law firms with an entirely different, generally pro-business agenda. From the mid-1970s a number of aggressively conservative public interest legal centers have emerged in the USA. These law groups the Pacific Legal Foundation, the National Legal Center for the Public Interest, the National Chamber Litigation Group and others represent interests, such as timber companies, that they define as being under attack in a world of environmental protections. However, because of their link to affluent business interests through leadership and financing, it is by no means accepted that these groups meet the traditional criteria for public interest law. Other groups, which represent conservative social positions, such as opposition to abortion and affirmative action, are sometimes grouped with public interest law firms.

3.4 The Internationalization of Public Interest Law

Although public interest law is most widely practiced in the USA, there has been a gradual expansion into other countries. Many groups, especially in the Third World, are funded by Western foundations and non-governmental organizations. During the apartheid era in South Africa, Geoffrey Budlender toured the USA interviewing public interest lawyers on their strategies. Then, funded by the Carnegie and Ford Foundations, he returned to South Africa to participate in the founding of the Legal Resource Center (LRC). Although fighting in a system that had appropriated to itself emergency powers, the LRC litigated over the Pass Laws and issues such as internal banishment. In post-apartheid South Africa, several lawyers who practiced public interest law during the apartheid years have risen to positions of power: Arthur Chaskelson, for example, one-time chair of the LRC board is chief judge of the countrys constitutional court.

In Bangladesh, the Madaripur Legal Aid Association represents poor rural Bangladeshis. Likewise, in Nepal, the Women's Legal Services Project works with abused women in a society that is overwhelmingly patriarchal. In India, lawyers from Action for Welfare and Awakening in Rural Environment (AWARE) have litigated against the oppression of lower caste bonded laborers in a culture still heavily influenced by the caste system.

Public interest law has also begun in Latin America; the Latin American Institute of Alternative Legal Services works with over 6,000 lawyers, paralegals and social workers throughout the continent, bringing to public attention issues such as environmental protection and the treatment of street children and indigenous peoples. In Eastern Europe, public interest law groups funded by financier George Soros' Open Society Institute, the German Marshall Fund, the Ford Foundation and others, have worked to solidify freedom of expression, and to counteract racism against groups such as the *Roma* (Gypsies).

Public interest lawyers are also active in the UK. They have established a Law Centres movement, and worked on immigration issues, trying to halt deportations and to secure refugee status for more asylum-seekers.

In recent years, one promising arena for public interest law has been on a transnational level: international human rights law and the attempts by labor unions to have labor protections introduced into treaties such as NAFTA and the WTO.

4. Summary

Public interest law is an assertion of a commitment within the justice system to fairness, open process, and diversity of viewpoints. It has significantly shaped public policy debates in the USA, even during the

Reagan and Bush presidencies. Their administrations helped to make access to the courts a less significant tool for citizen organizations. In response to these developments, many public interest lawyers developed a range of alternate strategies; working more directly in the legislative and administrative processes; filing suits in state courts; and relying more heavily on citizen actions.

The turn of the federal courts in a conservative direction was matched by increasing sophistication and resources in the advocacy effort of large corporations. No longer content to rely on the skills of legal advocates to shape public policy, corporations drew more heavily on public relations professionals. They relied more heavily on political action committees and the financing of political candidates and campaigns, and on massive advertising campaigns to shape public opinion.

Despite these challenges, public interest law in the USA continued to thrive, affecting policy decisions, and providing a measure of balance in the legal system. There are many more law school graduates seeking public interest law work than there are jobs; and there are many situations in which effective public interest advocacy, were it available, would lead to better policy outcomes. The development of public interest law in other countries holds significant promise for fair processes and broader public participation in critical policy decisions ranging from environmental protection to the protection of racial minorities. Globalization creates new needs and opportunities for public interest lawyers to function in transnational settings.

See also: Democracy; Discrimination; Equality and Inequality: Legal Aspects; Justice and Law; Justice, Access to: Legal Representation of the Poor; Law and Democracy; Law and Society: Sociolegal Studies; Law as an Instrument of Social Change; Law: Overview; Political Lawyering; Public Interest; Race and the Law

Bibliography

Aron N 1989 *Liberty and Justice for All: Public Interest Law in the 1980s and Beyond*. Westview Press

Carson R 1961 *Silent Spring*

Council for Public Interest Law 1976 *Balancing the Scales of Justice: Financing Public Interest Law in America*. Council for Public Interest Law

McClymont M, Golub S (eds.) 2000 *Many Roads to Justice: The Law Related Work of Ford Foundation Grantees Around the World*. The Ford Foundation

Nader R 1966 *Unsafe at Any Speed*

Sarat A, Scheingold S (eds.) 1998 *Cause Lawyering: Political Commitments and Professional Responsibilities*. Oxford University Press

Weisbrod B A et al. 1978 *Public Interest Law: An Economic and Institutional Analysis*. University of California Press, CA

C. Halpern

Public Management, New

A term coined in the late 1980s to denote a new (or renewed) stress on the importance of management and 'production engineering' in public service delivery, often linked to doctrines of economic rationalism (see Hood 1989, Pollitt 1993). The apparent emergence of a more managerial 'mood' in several (mainly but not exclusively English-speaking) countries at that time created a need for a new label. The new term was intended to denote public service reform programs that were not confined to the 'new right' in a narrow sense, but also came from labor and social-democratic parties and in that sense could be considered as part of what was later labeled a 'third way' agenda.

New Public Management is sometimes (understandably) confused with the 'New Public Administration' movement in the USA of the late 1960s and early 1970s (cf. Marini 1971). But though there may have been some common features, the central themes of the two movements were different. The main thrust of the New Public Administration movement was to bring academic public administration into line with a radical egalitarian agenda that was influential in US university campuses at that time. By contrast, the emphasis of the New Public Management movement a decade or so later was firmly managerial in the sense that it stressed the difference management could and should make to the quality and efficiency of public services. Its focus on public service production functions and operational issues contrasted with the focus on public accountability, 'model employer' public service values, 'due process,' and what happens inside public organizations in conventional public administration. That meant New Public Management doctrines tended to be opposed to egalitarian ideas of managing without managers, juridical doctrines of rigidly rule-bound administration and doctrines of self-government by public-service professionals like teachers and doctors.

However, like most divinities, the core of New Public Management is somewhat mystical in essence, despite or perhaps because of the amount that has been written about its central content. Different authors give various lists of its key traits (e.g., Hood 1989, Pollitt 1993). Some have identified different styles of public-sector managerialism over time (see Ferlie et al. 1996). How far the small-government economic-rationalist agenda that went together with more stress on public-sector management in the 1980s and 1990s was integral to those managerial ideas is

debatable (see Barzelay 2000). But it is hard to separate these elements historically, since the advent of a new generation of public-sector managerialism coincided with concern by numerous OECD governments to reduce the power of public service trade unions, increase regulatory transparency and tackle perceived inefficiencies of public enterprises. A commonly-cited view of New Public Management's central doctrinal content is Aucoin's (1990) argument that it comprised a mixture of ideas drawn from corporate management and from institutional economics or public choice. To the extent that Aucoin's characterization is accurate, it suggests New Public Management involves a marriage, if not exactly of opposites, at least of different outlooks, attitudes, and beliefs that are in tension. Savoie (1995) argues that the central doctrinal theme of public-sector managerialism is the idea of giving those at the head of public organizations more discretionary decision space in exchange for direct accountability for their actions.

Despite the label, many of the doctrines commonly associated with New Public Management are not new. Jeremy Bentham's voluminous philosophy of public administration developed in the late eighteenth and early nineteenth century provides the *locus classicus* for many supposedly contemporary ideas, including transparent management, pay for performance, public service provision by private organizations, and individual responsibility. The idea that more effective public services could be obtained by judicious application of private-sector management ideas is also a theme going back at least to the US city-manager movement of the late nineteenth century (cf. Downs and Larkey 1986). It was advanced early in the twentieth century by figures like Taylor (1916) and Demitriadi (1921). The idea that public services can be improved by giving some autonomy to managers operating at arms length from political standard-setters was often invoked in the nationalized public enterprise era. It was central to Beatrice and Sydney Webb's early twentieth-century Fabian idea of the proper way of organizing the 'social' tasks of government. Some have argued the contemporary doctrine of creating 'managerial space' in public services harks back to the US Progressive-era doctrine of a politics-administration dichotomy and independent regulators (cf. Overman 1984).

Like feminism or environmentalism, New Public Management is both a social movement and a subject of academic study. Indeed, during the 1990s, New Public Management has become a major academic industry across the world, filling bookshelves and websites with writings and conference proceedings using the term in their titles. It has its advocates and its critics, its analysts, morphologists and epistemologists, its evaluators and case-historians. Advocates stress the value to citizens and consumers to be gained by enlightened managers moving beyond what is claimed to be an outdated 'bureaucratic paradigm,' paying more attention to how to satisfy citizen demands and to service delivery through organizations other than traditional public bureaucracies (see Barzelay and Armajani 1992, Osborne and Gaebler 1992, Jones and Thompson 1999). A sophisticated defence of public-sector managerialism is Moore's (1995) exposition of a 'strategic triangle' (of political possibility, substantive value, and administrative feasibility) within which skillful and entrepreneurial public managers can 'add value' to public services. Critics of public-sector managerialism stress the virtues of traditional Weberian bureaucracy for rule-of-law or public accountability (cf. Goodsell 1994) or see managerialism as a 'wrong problem problem' (Frederickson 1996) diverting governments' attention from hard policy choices. Some critics of New Public Management doctrines see them as too heavily based on business-school and private-sector management perspectives, and some public-choice scholars have offered 'rent-seeking' explanations of contemporary public sector reforms (see Dunleavy 1992).

Morphologists and analysts of New Public Management have several concerns. They include exploring different forms and types of public-sector managerialism, identifying how managerialism varies cross-nationally, and explaining the observed commonalities and differences (cf. Savoie 1995, Aucoin 1995). A key debate (to date inconclusive because of poor benchmarking of historical points of origin) concerns how far or in what ways contemporary public management reforms represent convergence on some global paradigm. Some stress the international commonality of public-sector management reform themes while others stress the different political motivations that prompt reforms (Cheung 1997) or argue that New Public Management represents an 'Anglo-Saxon' reform agenda that does not resonate in other contexts (Derlien 1992, Kickert 1997). A related explanatory enterprise is represented by a search to understand why management-reform ideas became popular when and where they did (cf. Hood and Jackson 1991). Epistemological debates concern the appropriate disciplinary framework for reasoning about contemporary public sector management and the rival claims of different schools or approaches to intellectual ownership of the subject, particularly in the US context of competing schools of public policy, public administration and political science (see Lynn 1996, Barzelay 2000).

Evaluators and case-historians of public-sector managerial change track the processes of reform and explore their effects. Much of this work is pitched at the level of single agencies (such as Jones and Thompson 1999), but there have been some attempts to evaluate government-wide changes (see Schick 1996). Evaluation of the effects of new managerial techniques has been patchy, but several scholars have identified 'managerial paradoxes' of one kind or another. One of those paradoxes is Maor's (1999)

observation that the development of a more managerial approach to public service produced more, not less, politicization of the senior public service in six countries. A second is Gregory's (1995) controversial claim that orthodox managerial approaches foster a 'production' approach to public services that leads to several unintended effects, including downgrading of responsibility and what he termed 'careful incompetence.' A third is the claim, redolent of Tocqueville's paradox of administrative development in post-revolutionary France, that contemporary public management may in fact involve more rather than less 'rules-based, process-driven' bureaucracy, as a result of increasing oversight and regulation and continuing stress on compliance-based rather than result-based evaluation (see Light 1993, Jones and Thompson 1999, Hood et al. 1999, Pollitt et al. 1999).

In spite of the scale and growth of the New Public Management 'industry,' or perhaps because of it, the term New Public Management has probably outlived its analytic usefulness. The term is ambiguous because the agenda of public sector reform has moved in some respects beyond the traits identified by scholars of public management in the 1990s reflecting the various cultural cross-currents that have swept through managerial debate. The term is also too crude for the fine-grained distinctions between different sorts and themes of managerialism that academic scholars need to make as the study of public services develops and the public sector reform movement becomes professionalized. Just as Eskimos are said to have many different terms to distinguish different types of snow, we need more words to describe the cultural and technical variety of contemporary managerialism. So it is not surprising that there have been numerous attempts to proclaim a move beyond New Public Management (e.g. Minogue et al. 1998). Nevertheless, in spite of its oft-proclaimed death, the term refuses to lie down and continues to be widely used by practitioners and academics alike.

See also: Administration in Organizations; Bentham, Jeremy (1748–1832); Planning, Administrative Organization of; Public Administration: Organizational Aspects; Public Administration, Politics of; Public Bureaucracies; Public Sector Organizations

Bibliography

Aucoin P 1990 Administrative reform in public management: paradigms, principles, paradoxes and pendulums. *Governance* **3**: 115–37

Aucoin P 1995 *The New Public Management*. IRPP, Montreal, Canada

Barzelay M 2000 *The New Public Management*. University of California Press, Berkeley, CA

Barzelay M, Armajani B J 1992 *Breaking Through Bureaucracy*. University of California Press, Berkeley, CA

Cheung A B L 1997 Understanding public-sector reforms: Global trends and diverse agendas. *International Review of Administrative Sciences* **63**(4): 435–57

Demitriadi S 1921 *A Reform for the Civil Service*. Cassell, London

Derlien H-U 1992 Administrative reform in Europe—Rather comparable than comparative. *Governance* **5**(3): 279–311

Downs G W, Larkey P D 1986 *The Search for Government Efficiency*. 1st edn. Temple University Press, Philadelphia, PA

Dunleavy P 1992 *Democracy, Bureaucracy and Public Choice*. Harvester Wheatsheaf, Aldershot, UK

Ferlie E, Pettigrew A, Ashburner, Fitzgerald L 1996 *The New Public Management in Action*. Oxford University Press, Oxford, UK

Frederickson G H 1996 Comparing the reinventing government movement with the new public administration. *Public Administration Review* **56**(3): 263–70

Goodsell C T 1994 *The Case for Bureaucracy*, 3rd edn. Chatham House, Chatham, NJ

Gregory R 1995 Accountability, responsibility and corruption: Managing the 'public production process'. In: Boston J (ed.) *The State Under Contract*. Bridget Williams Books, Wellington, New Zealand

Hood C 1989 Public administration and public policy: Intellectual challenges for the 1990s. *Australian Journal of Public Administration* **48**: 346–58

Hood C, Jackson M 1991 *Administrative Argument*. Aldershot, Dartmouth, UK

Hood C, Scott C, James O, Jones G W, Travers A 1999 *Regulation Inside Government*. Oxford University Press, Oxford, UK

Jones L, Thompson F 1999 *Public Management*. JAI Press, Stamford, CA

Kickert W J 1997 Public management in the United States and Europe. In: Kickert W J M (ed.) *Public Management and Administrative Reform in Western Europe*. Edward Elgar, Cheltenham, UK, pp. 15–38

Light P Q 1993 *Monitoring Government*. Brookings Institution, Washington, DC

Lynn Jr. L E 1996 *Public Management as Art Science and Profession*. Chatham House, Chatham, NJ

Marini F 1971 *Toward a New Public Administration*. Chandler, Scranton, PA

Maor M 1999 The Paradox of Managerialism. *Public Administration Review* **59**(1): 5–18

Minogue M, Polidano C, Hulme D (eds.) 1998 *Beyond the New Public Management*. E Elgar, Cheltenham, UK

Moore M H 1995 *Creating Public Value*. Harvard University Press, Cambridge, MA

Osborne D, Gaebler T 1992 *Reinventing Government*. Addison-Wesley, Reading, MA

Overman E S 1984 Public management. *Public Administration Review* **44**: 275–8

Pollitt C 1993 *Managerialism and the Public Services*, 2nd edn. Blackwell, Oxford, UK

Pollitt C, Girre X, Lonsdale J, Mul R, Summa H, Waerness M 1999 *Performance or Compliance?* Oxford University Press, Oxford, UK

Savoie D J 1995 *Thatcher, Reagan, Mulroney*. Pittsburgh University Press, Pittsburgh, PA

Schick A 1996 *The Spirit of Reform*. State Services Commission and NZ Treasury, Wellington, New Zealand

Taylor F W 1916 Government efficiency. *Bulletin of the Taylor Society* **2**: 7–13

C. Hood

Public Opinion: Microsociological Aspects

Although the insight that public opinion is a powerful force goes back to premodern social thought, the task of modeling how public opinion evolves through the interdependent choices of individuals was not taken up until the last quarter of the twentieth century. Through the ensuing theorizing, the term has acquired a more precise meaning than it holds in everyday language. In choice-based, microsociological theory, a group's 'public opinion' now refers to the distribution of the preferences that its members express publicly. By contrast, 'private opinion' signifies the corresponding distribution of genuine preferences. Where individuals choose to misrepresent their actual desires in public settings, these two distributions may differ.

1. Theoretical Foundations

The theoretical literature on public opinion addresses diverse phenomena that have puzzled social thinkers. They include its resistance to changes in social structures and policy outcomes; its capacity for immense movement if ever this resistance is overcome; its sensitivity to the ordering of social shocks; and its imperfect predictability.

Most variants of the theory draw on a class of constructs known as 'threshold,' 'critical mass,' 'bandwagon,' or 'cascade' models. Originally designed to explain epidemics, these were subsequently applied most influentially by Schelling (1978) and Granovetter (1978) to various social phenomena that exhibit periods of stability punctuated by sudden change, for example, stock market bubbles, bank runs, cultural fashions, corporate conservatism, and academic fads. The common feature of these phenomena is that they are shaped by interdependencies among individual decisions.

Public opinion harbors two distinct interdependencies. Wherever individuals lack reliable information of their own, they look to others for clues about the reality they are seeking to grasp. Such 'free-riding' on publicly available knowledge makes people's genuine wants interdependent. Also, in articulating preferences and conveying knowledge, individuals frequently tailor their choices to what appears socially acceptable. This conformism, which is motivated by the wish to earn acceptance and respect, creates interdependencies among the wants that individuals convey publicly.

2. The Public Opinion of a Poorly Informed Group

The essential contributions of the theory may be sketched through a model drawn from Kuran (1995), which contains additional details. Certain initial assumptions will be relaxed further on.

2.1 Thresholds and Expressive Equilibria

Consider a 10-person society that must decide a certain issue, say, the government's crime-fighting budget. Two alternatives present themselves: low (L) and high (H). Each individual will support whichever option appears socially optimal. Everyone will want this budget increased if crime seems to be rising but decreased if crime appears to be falling.

By assumption, no one has perfectly reliable private information. Consequently, every individual's judgment will depend partly, if not largely, on the apparent judgments of others. This does not mean that individual beliefs must converge. Because of differences in private information, any given common signal about the distribution of preferences may make some people favor H and others L.

For any individual, the minimum share that must consider H optimal for him or her to concur constitutes what is called that individual's responsiveness threshold. This threshold is a number between 0 and 100. Listing all 10 thresholds in ascending order yields society's threshold sequence (T). Here is an example:

$$T\text{: } (0, 10, 30, 30, 40, 50, 50, 60, 70, 80)$$

These individuals may be labeled a, b, …, j, with a considered female, b male, and so on, in alternating fashion.

Given the heterogeneity reflected in this sequence, the perceived division between the supporters of H and those of L can influence dramatically the realization of public opinion. For a demonstration, suppose that initially only 10 percent of society is thought to favor H. Whatever the source of this perception, a and b will favor H, because their thresholds lie at or below 10; and the remaining eight members of the group will favor L, because their thresholds lie above 10. The resulting public opinion will thus lean heavily against H, which will enjoy the support of only 20 percent of the 10 individuals. The outcome will not impel anyone to switch positions, so it represents an expressive equilibrium. The knowledge that only two of the 10 individuals favor H will make exactly that share do so.

This particular expressive equilibrium is not the only possible self-sustaining outcome. If initially society was believed to be divided evenly between the two options, the first seven members of the sequence (all

except those with thresholds above 50) would have favored *H*. Observing that 70 percent favor *H*, the two individuals with thresholds of 60 and 70 would then have joined the majority, raising this support to 90 percent. Finally, completing the learning process, this second adjustment would have impelled the lone holdout to join in. As with the above case of a divided public opinion, the resulting unanimity would have been self-reproducing.

2.2 Properties of Public Opinion

This pair of illustrations highlights several characteristics. Public opinion is path dependent, in that changing its historical starting point might alter the outcome. It is also sensitive to small events, which are intrinsically insignificant happenings that an analyst seeking a restrospective explanation would probably overlook. Whereas an initial perception of 20 percent support for *H* would be self-confirming, a perception of 30 percent would lead to a consensus in favor of *H*. The difference in these two outcomes is vastly disproportionate to the difference in their starting conditions, whose determinants would not necessarily register on a historian's viewscreen. Finally, public opinion may be unrepresentative of the available private information. Where the outcome is a consensus, it conceals the existence of substantial private information favorable to the unanimously rejected option.

2.3 Informational Cascades

The iterative process through which a self-reproducing public opinion takes shape generally is known as an informational cascade (Bikchandani et al. 1992). Under certain conditions, an informational cascade will enhance social efficiency. For example, where the first people to articulate their views are those with the most accurate information, the consequent interdependent learning will help to dampen private ignorance. However, such a fortuitous outcome cannot be guaranteed. Instead of educating the poorly informed, the informational cascade might serve to un-educate the most knowledgeable. This finding points to yet another characteristic of public opinion. It is sensitive to the order in which people express themselves.

Insofar as the evolution of public opinion influences policy decisions, it may weaken the political system's effectiveness in addressing social concerns. For example, it might induce the government to pour resources into solving a problem widely considered serious only because the pertinent social learning was spearheaded by the ignorant. Yet, the resulting allocation would not be undemocratic. Whatever the reason why public opinion has come to favor the allocation, and however wasteful an outsider might consider the instituted policy, the members of society would only be exercising the freedom to make their own mistakes. Moreover, the government would simply be responding to the popular will.

3. Reputational Determinants of Public Opinion

Though analytically useful, the above account provides an undersocialized view of human nature. In practice, people do not express their potentially controversial views mechanically, without considering how their social standings might be affected. As social beings who wish to be liked and accepted, they may refrain from expressing beliefs or preferences that are likely to offend powerful groups. Some evidence lies in Asch's (1957) line experiment, in which subjects contradicted the evidence of their own senses to avoid isolation and criticism, and in Noelle-Neumann's (1984) pre-election surveys, in which the publicly acknowledged party preferences of respondents, though not necessarily their private sentiments, varied according to their perceptions of which party was running ahead. Additional evidence comes from Scott's (1985) ethnographic research, which finds that peasants may profess loyalty to their political regime, yet also work to undermine it in settings where they can do so with impunity.

The activists who direct society's pressure groups seek to exploit the reputational concerns that such studies highlight. They do so partly by treating people differently according to their public expressions and partly by providing incentives for criticizing, belittling, perhaps even ostracizing, overt opponents of the groups they are promoting. Insofar as such tactics are successful, individuals tailor their public positions to the prevailing social demands, and public opinion is thereby affected.

3.1 Private and Public Preferences

Depending on the nature of the information distorted to obtain social rewards and escape social punishment, such misrepresentation is called either preference falsification or knowledge falsification. On any given issue, preference falsification occurs when an individual's publicly expressed preference, or simply public preference, differs from the corresponding privately held preference, or private preference. Similarly, knowledge falsification entails a discrepancy between the communicator's private knowledge and public knowledge.

Whereas a person's private preference always reflects the relevant private knowledge, this person's public preference and public knowledge need not be

mutually consistent. In practice, however, the two choices tend to be tightly coordinated in order to make the former credible. Someone who pretends to favor one particular option, but then proceeds to give reasons for another option's superiority, will hardly come across as sincere; the person's intended audience will sense that they are concealing something.

3.2 Reputational Cascades

For further insights into how individual expressions both mold public opinion and are molded by it, reconsider the threshold sequence of Section 2.1, adding the assumption that if the 10 individuals were polled anonymously, *H* would be favored five-to-three, with two abstentions. Specifically, the five members with thresholds of up to 40 would support *H*, the two with thresholds of 50 would be indifferent, and the three with the highest thresholds would support *L*. These preferences expressed under the cover of anonymity are private preferences. Thresholds depend partly on private preferences, because submitting to others' demands causes psychological discomfort, albeit of a magnitude subject to variation across contexts and individuals. A second determinant of thresholds is sensitivity to social pressures. Under this section's more realistic scenario, then, a threshold reflects not only the nature and depth of an individual's private preference but also that individual's preparedness to pay the price of joining an underprivileged expressive minority. Put differently, a threshold captures an individual's tradeoff between the benefits of truthful self-expression and those of reputationally advantageous preference falsification. If person *c*'s threshold is 30, the sum of their expected satisfaction from all sources, including truthful self-expression and both favorable and unfavorable reputational consequences, is maximized by expressing a preference for *H* whenever the share of its supporters appears to be at least 30. Otherwise, their expected satisfaction is maximized by publicly favoring *L*.

This threshold sequence produces, as already noted, two stable equilibria that represent highly different realizations of public opinion. Under the current assumptions, the selection among these equilibria occurs through a reputational cascade. In the course of a reputational cascade, people conform to the apparently popular choice not because they see wisdom in numbers but merely to maintain a favorable reputation. In so doing they make it increasingly difficult for others to remain nonconformist.

3.3 The Possibility of Inefficiency

Like the social learning mechanism described earlier, this social adaptation mechanism driven by reputational concerns makes public opinion path dependent, unrepresentative of the available private information, and sensitive to both small events and the order of individual choices. Additionally, it may make public opinion unrepresentative of private opinion. In the illustration of Section 3.2, neither of the possible outcomes matches what an anonymous opinion survey would identify as the 'popular will.' Whereas three members privately favor it, *H* could achieve either unanimous support or the support of just a two-person minority. Either way, social policies designed to accommodate public opinion would be responding to signals distorted by preference falsification. As such, these policies could be inefficient.

3.4 Availability Cascades

There are contexts in which public opinion is shaped by either purely informational or purely reputational cascades. As a rule, however, the outlined informational and reputational mechanisms feed on each other, or at least interact. The composite mechanism whereby public opinion evolves through both social learning and conformism aimed at securing social acceptance is known as an availability cascade (Kuran and Sunstein 1999).

4. Feedback from Public Discourse to Private Opinion

Reputational cascades treat private preferences as fixed. For their part, informational and availability cascades recognize the plasticity of private preferences, but only during the transition to a self-sustaining public opinion. In reality, this plasticity is not limited to periods of disequilibrium. Whether public opinion is at rest or in flux, the private knowledge that undergirds private preferences is ordinarily shaped and reshaped by public discourse, which consists of the suppositions, observations, and arguments that are communicated publicly (Coleman 1990).

Preference falsification affects this private learning through two channels. As both informational and availability cascades posit, the state of public opinion, whatever its inconsistencies with private opinion, molds individual interpretations of reality. In addition, the knowledge falsification that accompanies preference falsification sows ignorance and confusion by corrupting information in the public domain through the substitution of facts that people know to be false for ones they consider true. Over time, knowledge falsification may even make once-common objections to a publicly popular option disappear from society's collective memory.

Under diverse conditions, population renewal will facilitate such an outcome. If public discourse excludes

criticisms of the fashionable option, its drawbacks will tend to be forgotten as society's composition changes through births and deaths, or simply entries and exits. This is because new cohorts will not be exposed to the unfiltered private knowledge of previous cohorts but only to the reconstructed knowledge that was considered safe to communicate. The previous cohorts may have disliked the option yet refrained from challenging it. The younger cohorts will preserve it not only because they are afraid to object but also, perhaps mainly, because the impoverished public discourse of the past has blunted their capacity to criticize the inherited social order and imagine better alternatives. Insofar as past preference falsification removes individual inclinations to want something different, current preference falsification will cease to be a source of public opinion's stability. People will support established arrangements genuinely rather than to make a good impression (Kuran 1995).

A complementary mechanism leading to intellectual impoverishment and the absence of dissent is driven by individual decisions to choose 'exit' over 'voice.' When people withdraw from a troubled society to avoid the costs of speaking up, the society's ability to institute reforms diminishes (Hirschman 1970).

The foregoing dissent-dampening mechanisms have universal relevance. Customs, norms, laws, and other institutions that keep societies poor, organizations ineffective, and inequalities unchallenged often owe their stability to the paucity of overt demands for reform. By the same token, other institutions that promote growth, efficiency, and equality are held in place through pressures that diminish both the desire and the willingness to work toward their overthrow. As Tocqueville (1945) observes, these mechanisms even contribute to the stability of traits that give nations and cultures their unique identities.

5. *The Imperfect Predictability of Public Opinion*

If public discourse were the only determinant of private knowledge, once a public consensus had taken shape, it would become frozen. In fact, private knowledge has other determinants, including the material consequences of prevailing policies. By sowing doubts about established structures, these other factors can induce changes in private preferences and, ultimately, also in public preferences. However, the unraveling of an established consensus need not track the escalation of private opposition to the status quo. As long as the prevailing social pressures keep preference falsification optimal for all potential dissenters, private opposition will intensify without any change in public opposition. Yet, as the undercurrent of opposition swells, a shock that affects the right people's thresholds in the right manner will stimulate switches in certain public preferences. Even without further shocks, these switches will then encourage individuals with slightly higher thresholds to join the growing chorus of dissent. Once put in motion, then, open opposition can feed on itself, with each addition generating further additions until much of society publicly identifies with the reformist camp.

Because preference falsification had concealed the opposition brewing under the surface, this revolutionary transformation will not have been anticipated. Even so, the massive change in public opinion may easily be explained with the benefit of hindsight. The revolution's success will have lowered the personal risk of acknowledging past acts of preference falsification. It will also have created incentives for people who had supported the status quo ante sincerely to pretend that they had simply been waiting for a safe opportunity to speak out for change. Examples of largely unanticipated revolutions include the French Revolution of 1789, the Iranian Revolution of 1978, and the East European Revolutions of 1989. Each of these earth-shaking transformations entailed a massive shift in public opinion that stunned its beneficiaries, victims, and observers.

6. *Methodology*

A distinction of the choice-based, microsociological analysis of public opinion is that it admits circular relationships generated by feedback effects. For one thing, it links the social pressures that determine the extent to which private and public opinion diverge to the very collective process that shapes public opinion. For another, it recognizes that public opinion shapes the private preferences that are among its own determinants. By virtue of these characteristics, the microsociological approach differs methodologically from social scientific traditions in which either culture or the motives of individuals are fixed. Another salient distinction of this approach lies in its multidimensional conception of individual motives. In recognizing that individuals make tradeoffs between their expressive and reputational needs, it stands out from analytical traditions, including most variants of the public choice school, that treat individuals as invariably sincere (Downs 1957). At the opposite extreme, it stands out from self-presentation theories that view individuals as one-dimensional agents interested only in making a good impression (Goffman 1973).

Nothing in the arguments surveyed here presuppose that certain cultures promote truthfulness and others mendacity. Because people everywhere express themselves with an eye toward balancing diverse objectives, the identified social processes may be observed in any society—economically developed or advanced, democratic or undemocratic, large or small. Where public opinion shows variations across societies—as it often

does markedly—the variations do not stem from differences in the underlying basic mechanisms. Rather, they reflect historical contingencies that generated persistent differences in institutions, social pressures, and private knowledge.

See also: Collective Beliefs: Sociological Explanation; Democracy; Influence: Social; Lazarsfeld, Paul Felix (1901–76); Media Effects; Opinion Formation; Political Discourse; Political Sociology; Polling; Popular Culture; Public Opinion: Political Aspects; Public Sphere and the Media; Rational Choice Theory in Sociology; Tocqueville, Alexis de (1805–59)

Bibliography

Asch S E 1957 Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs* **70**: 416

Bikchandani S, Hirshleifer D, Welch I 1992 A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* **100**: 992–1026

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press, Harvard University Press, Cambridge, MA

Downs A 1957 *An Economic Theory of Democracy*. Harper, New York

Goffman E 1973 *The Presentation of Self in Everday Life*. Overlook Press, Woodstock, NY

Granovetter M 1978 Threshold models of collective behavior. *Am. J. Soc.* **83**: 1420–43

Hirschman A O 1970 *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, Cambridge, MA

Kuran T 1995 *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press, Cambridge, MA

Kuran T, Sunstein C R 1999 Availability cascades and risk regulation. *Stanford Law Rev.* **51**: 683–768

Noelle-Neumann E 1984 *The Spiral of Silence: Public Opinion—Our Social Skin*. University of Chicago Press, Chicago

Schelling T C 1978 *Micromotives and Macrobehavior*. Norton, New York

Scott J C 1985 *Weapons of the Weak: Everyday Forms of Peasant Resistance*. Yale University Press, New Haven, CT

de Tocqueville A 1945 *Democracy in America*. Knopf, New York

T. Kuran

Public Opinion: Political Aspects

1. Scope and Methods

Public opinion has been usefully defined by V. O. Key Jr. as 'those opinions held by private persons which governments find it prudent to heed' (Key 1961). Scholars of public opinion attempt to describe and account for the politically relevant preferences and beliefs of ordinary citizens, and to assess the political impact of those preferences and beliefs. They investigate a wide range of subject matter, including broad ideologies, specific policy preferences, evaluations of political parties and candidates, assessments of self-interest and collective interest, attributions of political responsibility, and attitudes toward social groups and political institutions (Kinder 1998).

The study of public opinion draws upon and overlaps the study of political psychology, but tends to focus on collective preferences and beliefs in their broad political context rather than on individual mental processes (see *Opinion Formation*). For example, the most notable work in the field begins with a psychological account of 'how citizens acquire information and convert it into public opinion,' but 'pays vastly more attention to the social sources of mass attitudes—in particular, the availability of information in elite discourse—than to the largely autonomous operation of people's minds and psyches' (Zaller 1992, p. 3).

Obviously, which opinions held by private persons governments will find it prudent to heed, depends significantly upon the political setting. It should not be surprising that systematic attention to the workings of public opinion has been most common in modern liberal democracies, where the preferences and beliefs of ordinary citizens are supposed to be broadly and routinely consequential, whether through elections of public officials, voting in referenda, interest group activities, or other mechanisms. By the same logic, the proliferation of new democracies in Eastern Europe, Latin America, and Asia since the 1980s may be expected to stimulate unprecedented scholarly attention to public opinion in those parts of the world.

Contemporary studies of public opinion are profoundly shaped by insights and concerns stemming from the rise of mass politics in the nineteenth century and from the intertwined evolution of modern democratic states and societies. For example, Zaller's (1992) influential model of the dynamics of opinion change builds upon Bryce's (1888) distinction between the 'active class' of political entrepreneurs and the 'passive class' of ordinary citizens who have 'been told what to think, and why to think it,' and upon Lippmann's (1922) analysis of the role of the mass media and political elites (including government bureaucrats) in shaping 'the pictures in our head' that govern our reactions to complex and distant public affairs.

While long-standing insights and concerns are readily recognizable in modern scholarship on public opinion, empirical research in the field has been transformed since the 1940s by the development and proliferation of increasingly ambitious and sophisticated opinion surveys. Semistructured in-depth interviews, focus groups, and laboratory experiments continue to generate valuable insights, but their

relatively small scale and problems of external validity make them better suited to the study of individual political psychology than of collective public sentiment. By comparison, an opinion survey with several hundred randomly selected respondents can provide a reasonably accurate representation of the preferences and beliefs of millions of ordinary citizens—public opinion on a scale which governments may well find it prudent to heed.

2. *Pioneering Analysis*

The pioneering analyses of public opinion based upon survey data were conducted in connection with US presidential elections of the 1940s (Berelson et al. 1954) and 1950s (Campbell et al. 1960). Berelson et al. concluded that election campaigns mostly activate pre-existing political preferences rooted in ethnic, sectional, class, and family traditions and reinforced by social interactions with like-minded acquaintances (see *Campaigning: Political*; *Voting: Class*). Campbell et al. emphasized the central importance of long-standing psychological attachments to the political parties (see *Party Identification*), while also acknowledging the significance of such short-term forces as Dwight Eisenhower's personal popularity, which overcame but did not supplant the Democratic advantage in party identification dating from the New Deal era. Both of these studies portrayed public opinion as surprisingly thin and unsystematic, arguing, for example, that 'many people know the existence of few if any of the major issues of policy' (Campbell et al. 1960, p. 170) and that political preferences 'are characterized more by faith than by conviction and by wishful expectation rather than careful prediction of consequences' (Berelson et al. 1954, p. 311).

These uncomplimentary characterizations of public opinion were refined and elaborated in what remains the single most influential (and controversial) work in the field, Philip Converse's essay on 'The nature of belief systems in mass publics' (Converse 1964). Converse's analysis was based on three primary sorts of evidence. First, he showed that survey respondents answering open-ended questions only rarely made spontaneous use of systematic ideological concepts or language. Second, he showed that respondents' political views were only weakly constrained by conventional notions of 'what goes with what': liberals on one issue often took conservative positions on other issues, even within the same policy domain (see *Issue Constraint in Political Science*). Third, repeated interviews with the same respondents demonstrated a great deal of instability in their opinions, and the pattern of that instability suggested that it was caused more by random fluctuation than by systematic responses to political events. On the basis of this evidence, Converse (1964, p. 245) concluded that 'large portions of an electorate do not have meaningful beliefs, even on issues that have formed the basis for intense political controversy among elites for substantial periods of time.'

3. *Civic Competence*

The civic competence of ordinary citizens has remained a central issue of scholarly debate in the decades since Converse's devastating analysis. One line of revisionist argument accepted the accuracy of Converse's characterization of public opinion in the USA of the 1950s, but denied its relevance to other times and places. Most notably, Nie et al. (1976) and others argued that the political upheavals of the 1960s had produced a 'new American voter' more sophisticated, more ideological, and more attuned to political issues than the one portrayed by Campbell et al. (1960) and Converse (1964) (see *Voting: Issue*). While this thesis has been subjected to vigorous methodological critiques—and has come with the passage of time to seem at least as time-bound in its own right as the portrayal it aimed to supplant—it focused attention on the potential responsiveness of public opinion to such powerful contextual factors as social change and the nature of political rhetoric.

A newer wave of revisionist scholarship has acknowledged that most citizens in most circumstances are relatively inattentive to politics and uninformed about public affairs, but has interpreted these facts as evidence of 'rational ignorance' in Anthony Downs's sense of avoiding costly information gathering in a realm in which individual investment can have only negligible collective consequences. This interpretation is nicely conveyed by the titles of three notable works of the early 1990s: *The Reasoning Voter* (Popkin 1991), *Reasoning and Choice* (Sniderman et al. 1991), and *The Rational Public* (Page and Shapiro 1992). These and other works argued that ordinary citizens make sense of the political world despite their lack of detailed information about ideologies, policies, and candidates. In some cases, the emphasis is on how citizens use 'information shortcuts' and 'gut reasoning' (Popkin 1991) to deal efficiently with political complexity. In others, the emphasis is on the tendency of statistical aggregation to iron out individual errors and idiosyncrasies in mass publics, building upon an argument first set forth by Condorcet in the eighteenth century. This 'extenuationist' literature (to use a label coined by Robert Luskin) has in turn inspired attempts to measure the extent to which information shortcuts and aggregation actually do mitigate the political effects of 'rational ignorance.' For example, Bartels (1996) estimated that individual vote choices in a series of US presidential elections fell about half way between random behavior and hypothetical 'fully informed' choices (imputed to each voter by observing

the choices of better-informed voters with similar demographic and social characteristics). Deviations from 'fully informed' behavior at the individual level were mitigated but not eliminated at the aggregate level: on average, Democrats did almost two percentage points better and incumbent Presidents did almost five percentage points better than they would have if all voters were, in fact, 'fully informed.'

Normative concerns about the quality of public opinion have also led to the development by James Fishkin and others of 'deliberative opinion polls' designed to combine the efficiency and representativeness of conventional opinion surveys with the informative and deliberative aspects of a seminar or town meeting. Rather than simply reporting their existing opinions, respondents are brought together with policy experts, public officials, moderators, and each other to consider political issues in depth. The aim is to find out—and then to publicize—'what the public *would* think, if it had a more adequate chance to think about the questions at issue' (Fishkin 1991, p. 1).

4. *Political Consequences of Public Opinion*

Political scientists have been more successful in characterizing and accounting for public opinion than in specifying its role in the broader political process. The path-breaking study by Miller and Stokes (1963) of representation in the US Congress provided a model for much subsequent research examining the relationship between public opinion and elite attitudes and behavior. However, the basic research design employed in these studies—cross-sectional analysis of the impact of constituents' opinions on the opinions or behavior of individual legislators—is arguably ill-suited to answering questions about the political impact of public opinion in the aggregate, especially in legislative systems marked by strong party discipline or proportional representation, or when crucial policy decisions are made by executives or bureaucrats rather than by legislatures.

An alternative research design focuses on co-variation between aggregate public opinion and policy outcomes across political units or in a single political unit over time. The analysis of 'dynamic representation' by Stimson et al. (1995) is an ambitious example of this approach, relating variations in the liberalism of America's 'public mood' over a 35-year period to subsequent variations in the behavior of Presidents, Congresses, and courts. A more specialized literature treats fluctuations in presidential popularity not only as a product of political, economic, and international events (as in the pioneering work of Mueller (1973)), but also as a factor in accounting for the varying ability of Presidents to get things done (see *Presidential Popularity*).

Efforts to specify the broad political causes and consequences of public opinion will no doubt be facilitated by the continuing accumulation of survey data over time and space. For example, the cumulative data file of the American National Election Studies project now includes more than 40,000 individual survey respondents in more than two dozen national elections spanning half a century, while the Eurobarometer series includes biannual surveys since 1973 in each of the member states of the European Union. By merging these enormous survey data sets with aggregate-level data reflecting variations in national economic conditions, electoral laws, party platforms and campaign appeals, policy initiatives, media coverage of public affairs, and many other contextual factors, analysts should be able to illuminate important connections between mass opinion and elite behavior, as well as the role of social and political institutions in shaping those connections.

On the other hand, the availability of richly detailed survey data has probably distracted scholarly attention from crucial aspects of public opinion that are not readily operationalized in surveys. For example, Converse's (1964, pp. 245–6) 'issue publics'—the various minorities of the public whose 'activated' opinions on particular issues of interest are 'expressed in the writing of letters to the editor, the changing of votes, and the like'—have seldom been subjected to detailed scrutiny. Even more notably, the elusiveness of what Key (1961) referred to as 'latent opinion'—the preferences and beliefs politicians *expect* to confront in future elections, after today's policies have had their effects and tomorrow's political opponents have had their say—has discouraged political scientists from giving it attention commensurate with its undoubted political significance. In these cases, among others, scholars will have to be more ingenious and eclectic in collecting and analyzing relevant data if they are to place the study of public opinion more squarely in the mainstream of political science.

See also: Attitude Measurement; Attitudes and Behavior; Issue Constraint in Political Science; Issue Evolution in Political Science; Party Identification; Political Advertising; Political Communication; Polling

Bibliography

Bartels L M 1996 Uninformed votes: Information effects in presidential elections. *American Journal of Political Science* **40**: 194–230

Berelson B R, Lazarsfeld P F, McPhee W N 1954 *Voting: A Study of Opinion Formation in a Presidential Campaign*. University of Chicago Press, Chicago

Bryce J 1888 *The American Commonwealth*. Macmillan, New York

Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. Wiley, New York
Converse P E 1964 The nature of belief systems in mass publics. In: Apter D E (ed.) *Ideology and Discontent*. Free Press of Glencoe, Glencoe, IL
Fishkin J S 1991 *Democracy and Deliberation: New Directions for Democratic Reform*. Yale University Press, New Haven, CT
Key V O 1961 *Public Opinion and American Democracy*. Knopf, New York, 1st edn.
Kinder D R 1998 Opinion and action in the realm of politics. In: Gilbert D, Fiske S, Lindsey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston
Lippmann W 1922 *Public Opinion*. Macmillan, New York
Miller W E, Stokes D E 1963 Constituency influence in Congress. *American Political Science Review* **57**: 45–56
Mueller J E 1973 *War, Presidents and Public Opinion*. Wiley, New York
Nie N H, Verba S, Petrocik J 1976 *The New American Voter*. Harvard University Press, Cambridge, MA
Page B I, Shapiro R Y 1992 *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. University of Chicago Press, Chicago
Popkin S L 1991 *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. University of Chicago Press, Chicago
Sniderman P M, Brody R A, Tetlock P E 1991 *Reasoning and Choice: Explorations in Political Psychology*. Cambridge University Press, New York
Stimson J A, MacKuen M B, Erikson R S 1995 Dynamic representation. *American Political Science Review* **89**: 543–65
Zaller J R 1992 *The Nature and Origins of Mass Opinion*. Cambridge University Press, New York

L. M. Bartels

Public Policy Schools

1. *What is Policy Analysis?*

Before describing public policy schools, it is first necessary to describe the field of policy analysis. Although neither academics nor practitioners of policy analysis have been able to agree on a common formal definition, almost all would agree that policy analysis involves the use of formal reasoning to formulate advice for the solution of problems faced by governmental and nonprofit sectors. Of course, political leaders, decision-makers, managers, and administrators have relied on advisors over recorded time. Scholars and practitioners in many fields, particularly in the professions such as law, medicine, public health, social welfare, and urban planning, claim to incorporate policy analysis as part of their task. How then do public policy schools and their graduates differentiate themselves from the mass of advisors? They do so by using formal analytic techniques to make informed guesses about the outcomes of alternative choices, as they would be implemented on the ground in the real world.

The modern policy analysis movement had its genesis in the United States and has spread very slowly, if at all, to other countries. For that reason, the history of public policy as it evolved into an academic curriculum is a history of the application of policy analysis in the United States.

The staffing effort of the Second World War and the subsequent Cold War with the Soviet Union brought the policy analysis movement into existence. Although many historians of the policy analysis movement point to Daniel Lerner and Harold Lasswell's call for a policy sciences field that would apply reasoned advice to all stages of the policy process (deLeon and Steelman 1999, Lerner and Lasswell 1951), in practice the field developed with the application of what might be called the optimization sciences—particularly microeconomic analysis and the techniques of operations research—to policy choices. These techniques were first used in this way during the Second Word War when the need for war mobilization led to a planned economy significantly replacing the market economy. After the war the new techniques found a home at a newly formed home for defense analysis (the RAND Corporation). Ultimately these techniques, and many defense analysts, found a home in offices or ministers of defense and then diffused throughout government. The analytic offices that were set up became an established source of policy advice using the techniques of benefit-cost analysis, cost-effectiveness analysis, operations research, and policy evaluation. These techniques became the heart of the new policy analysis curriculum.

Since the 1970s in the United States, public policy schools have become major centers for graduate education for service in the public and nonprofit sectors in the United States. At many prestigious US universities they have replaced schools and programs of public administration. Their number has grown from under 10 in 1970 to over 60 in 2000. Many existing public administration programs, moreover, have incorporated parts of the public policy curricula, and in some cases have adopted public policy concentrations.

The growth in public policy schools and programs has been associated with the creation of an academic field complete with an association—the Association for Public Policy Analysis and Management (APPAM)—and a journal—the *Journal of Policy Analysis and Management* (*JPAM*). Of equal if not greater importance has been the spread of the intellectual approach of policy analysis as graduates of policy schools have replaced traditional public administrators in high-level staff positions at the US federal, state, and local governments.

As graduates of policy programs have replaced students of accounting and public administration in various governmental staff agencies—particularly

budget offices—the policy analysis movement and approach have been institutionalized within the governmental service. This institutionalization was reinforced by the slow economic growth and resulting budgetary constraints of the 1970s and 1980s which forced decision-makers to continually face tighter constraints and tradeoffs. New institutions, such as the Congressional Budget Office of the US Congress, were largely staffed by those with policy analysis training obtained either in schools of public policy or Ph.D. programs in economics.

These new staff organizations reflected the norms and values of their dominant staff professions—the graduates of economics departments and policy analysis schools and programs. Chief among these was a belief that, all things being equal, markets are the best way to allocate society's goods and services. That a decision to use other allocation mechanisms was justified only on grounds of equity or rights (the distribution of benefits and costs and/or the existence of some activities that society believed should not be freely exchanged), or when one could show that markets had failed and that governmental intervention would (not could) lead to a more economically efficient allocation than was the case under the failed market. This reliance on markets separated the new policy analysis movement from the traditional planning approach and built within the policy analysis framework a conservative bias when it came to justification of public sector activity.

The centrality of the norms of economics to the new policy analysis movement led to a belief that analysis of policy requires the laying out of the tradeoffs inevitable in the creation and implementation of policy. These tradeoffs are reflected in the centrality of two ideas of economics—opportunity costs and the margin. The policy analyst relies on opportunity cost to inform the decision-maker that no option is free, that one can always use the resources for other activities. The concept of the margin allows the economist or policy analyst to tell the decision-maker what she/he can obtain from adding or subtracting one unit of resources. The use of these concepts separates the economist and policy analyst from the political analyst, who often seeks to suppress tradeoffs so as to increase the membership of political coalitions.

A third effect was a conscious attempt to present decision makers with analyses of policy options that make choices concrete by clearly defining problems, setting out possible solutions, and specifying the criteria to be used to judge those solutions so that those decision-makers can evaluate the solutions in a comparative, predictable manner. Although students of policy analysis often wrote about qualitative as well as quantitative indicators, in practice practitioners of the new field found that their comparative advantage vs. other advisors was in the use and presentation of quantitative measures—particularly those in a benefit/cost format.

The reliance on analytic techniques and on policy options separated economists and policy analysts from traditional American public administrators. Since the progressive era and the early academic writings of Woodrow Wilson on public administration (Wilson 1887), students of public administration had operated under the belief that it was possible to separate the creation of policy from its efficient administration. Although by the 1960s most public administrators knew that this dichotomy between politics and policy on the one hand and administration on the other was a myth, much, if not most, of the public administration approach still centered on the efficient running of bureaus and departments. Consequently, to a great extent, the educational core of masters in public administration programs reflected this approach with required courses on the functional skills needed by a line manager—budgeting, organizational behavior, personnel management, and the role of an unelected civil servant in representative government.

2. *What Are Public Policy Schools and How Do They Differ From Other Programs?*

The first public policy schools and programs were created in the US in the late 1960s. With considerable philanthropic support, 1971 had founded nine schools or programs in the United States. In a pattern that has continued to this day, several existing programs quickly adopted the new approach and curricula of the policy school movement.

The advocates and sponsors of these new schools and programs sought to apply the new optimizing techniques of policy analysis to the problems facing nations in the late 1960s—particularly the problems facing urban areas. At this time it was still widely believed that government could solve social problems if the right policies were found, created, adopted, and implemented. Thus, the initial educational effort sought to support positive public sector activity to solve problems, a view that has much less support today. The creators of the new programs also sought to do for education towards public service what the changes in business administration education had done for business administration in the mid-1950s—replace a curriculum centered on institutional description and anecdotes with one grounded on analytics.

2.1 *Similarities Across the Schools*

Although one can point to many differences among the original nine policy schools and programs—one had a heavy emphasis on undergraduate education, a second only offered a Ph.D.—viewed from afar their similarities far outweighed their differences. With the exception of the one that only offered a Ph.D, all

focused their graduate education efforts on a Masters degree that was viewed as a terminal professional degree rather than a step toward a doctorate. In most cases they labeled this primary degree the masters in public policy (MPP).

2.2 Concentration on Analytics and Professionalism

The basic analytic public policy core set of courses has remained quite similar since the founding of the nine original schools. It centers on a combination of three yearlong sequences of analytic skill offerings—in economics, quantitative methods, and political and organizational analysis—one or more workshops in which students undertake policy analyses for real, as against hypothetical, clients, and a required summer internship. Some programs also required courses in individual and/or policy ethics. Others required courses in legal analysis. Initially the economics sequence centered on microeconomics, although several of the programs required a microeconomics–macroeconomics sequence. Although centered on courses on descriptive and inferential statistics, many of the initial required curricula also included material drawn from decision analysis, game theory, and operations research.

The political and organizational analysis sequences were included so that students trained in technical skills would also graduate from the programs with the political and managerial skills needed to enact, implement, and administer policies. Of all the required sequences in the policy analysis curriculum, political and organizational analysis has remained the most problematical. From their inception public policy programs rejected the public administration separation of politics from administration. Instead it was assumed that to be an effective analyst one had to be an effective political actor and manager. But unlike economics with its instrumental approach, the academic discipline of political science was largely descriptive and as such was less useful for a professional curriculum. This posed and still poses a great challenge to the public policy faculty charged with the political and managerial core of the policy curriculum. Attempts have been made at a number of institutions to take the findings of political science and public administration research and, through the use of decision forcing case studies, to build a prescriptive and strategic approach to management. Students were challenged to go from answering the descriptive question, 'What are the political forces or organizational tendencies which affect the kind of policy we can formulate?' to the more instrumental question, 'What do I, as analyst or manager, need to do, as I shape my policy prescriptions or implement the chosen policy, to ensure that the political forces and organizational tendencies do not succeed in frustrating or changing desirable policy in undesirable ways?' (Fleishman 1990, pp. 743–4).

Many observers have felt that this attempt has been only partially successful, however. To overcome this deficiency some scholars have focused on identifying examples of 'best practice' that might form the basis of improved political and organizational behavior (Bardach 1994, 2000). Others, criticizing the lack of scientific rigor in this approach, have urged their colleagues to create a body of knowledge in which statements of appropriate political and organizational actions are subject to formal hypothesis testing (Lynn 1994).

The professional nature of public policy programs is reflected in the fact that, unlike the traditional pattern in which professional masters students simply take graduate courses offered by the economics and political science departments, the required policy analysis courses were and are especially designed for a professional program. The goal in most cases is to recast traditional material in a policy centered, problem solving framework and to present the material in an instrumental rather than descriptive manner. As indicated, political science approaches are largely descriptive, and the new programs, therefore, had to take this material and turn it on its head by presenting it in a fashion that would allow future practitioners to use it as a lever to achieve policy and managerial goals—what might be called political and managerial engineering.

One constant across nearly all the original courses was the focus on domestic policy. Over time international material slowly entered the curricula, but this occurred largely through applications of the analytic skills of the curricula (such as benefit/cost analysis) to international situations, such as whether a water project should be built in another country. In recent years several policy schools have moved into the area of economic development, but almost all have avoided the traditional subjects of international relations and country analysis.

2.3 Differences Between Public Policy and Public Administration Programs

The new public policy programs sought to replace the existing public administration educational model of public administration. In addition to their focus on analytic skills, the new programs differed from their public administration counterparts in their large required set of core courses, the absence of required concentrations, and the greater number of economists and statisticians on the faculty.

Existing public administration programs had a small set of core required programs (four to five courses) and lasted a year to a year and a half. The new public policy programs were two-year programs with

a large number of required core courses (at least a full year's worth at a minimum and in some cases up to two-thirds of a standard two-year course load). In most cases students were required to take these courses in a set sequence. The average public administration program (MPA) offered many specialties or concentrations (with a mean of five in 1984). (Ellwood 1984). Initially at least, public policy masters (MPP) programs almost totally focused on the analytic techniques taught in their large cores. In 1971, in the United States, the average MPA program faculty member's doctorate was either in political science or public administration (21 and 20 percent in 1971). Economists represented only 9 percent. In the new public policy schools and programs economists represented a plurality (31 percent in 1971) and in some cases a majority of the faculty with political scientists being the second most represented discipline (23 percent) (Ellwood 1984).

3. *Changes and Challenges to Public Policy Schools*

Masters programs in public policy currently dominate graduate education for public and nonprofit service in the United States. Graduates from the programs are now found in key decision-making positions throughout government and the nonprofit sector. However, recent developments provide a challenge to the future of the approach and the field.

3.1 *The Decline in the Belief that the Public Sector Can Solve Problems*

The public policy school initiative was fostered by the belief that better public policies could lead to the solution of public problems—particularly those of the poor and of urban areas. Over the past three decades this belief has attenuated if not disappeared in the United States and elsewhere. To paraphrase Ronald Reagan, in the United States government is now seen as the problem, not the solution. Public service itself has become less attractive. The private sector increasingly has come to be seen as the creator and locus of heroes. Private sector remuneration has grown at a far faster pace than public or nonprofit sector pay.

The effect of all these changes has been to draw graduates of public policy programs away from careers in the public and nonprofit sectors toward jobs in the private sector. One recent study found that between 1973 and 1993, in the United States, the percentage of MPP graduates taking jobs in the public sector declined from 76 to 49 percent (Light 1999). At the largest US MPP program this percentage dropped to 28 percent for one year in the late 1990s before rebounding to the mid-40 percent range. This trend poses this question, will policy programs continue to attract the 'best and the brightest' as they have over the past three decades?

3.2 *The Difficulty of Management and Leadership Education*

As indicated above, from the beginning of the policy school movement it was understood that high technical skill was but a necessary condition for an effective policy analyst. The sufficient condition was the acquisition of political, leadership, and managerial skills. The inclusion of a required sequence of political analysis and management in most programs was meant to impart these skills. But this effort has always been the most problematic endeavor of public policy education. Academics simply know less and have a more difficult time imparting what they know to practitioners in these areas than they do when it comes to technical skills. Moreover, from the beginning, management skills were grafted on to a curriculum focusing on the skills needed by staff analysts. In contrast to masters of business administration (MBA) programs in business schools, where the entire curriculum (or at least the first year of required courses) was seen as management education, policy schools sought to condense courses in accounting, finance, marketing, and operations into one or two required offerings. The result has never been totally satisfying.

3.3 *The Tradeoff Between Breadth and Depth*

The original public policy curricula focused on imparting the skills that would be needed by policy analysts who, acting as advisors to decision-makers, would create policies that would hopefully solve society's problems. Over time policy schools have moved away from this narrow focus on the training of staff to a broader focus of training future managers and leaders. Moreover, the initial narrow focus on domestic policy has also broadened. Today the policy schools seek to serve the needs of students with a wide set of career goals. Some want to become elective officials, some political activists, some social activists, some career civil servants, some managers of nonprofit organizations, and some private sector consultants. The percentage seeking careers in technical policy staff positions has become smaller and smaller. As the programs have tried to meet the needs of this increasingly diverse set of students, they have to some extent lost the coherence and depth of the required set of skills that made up the core of their curricula.

As they have tried to serve future elected officials, political activists, public and nonprofit managers, as well as policy analysts, some policy schools have dropped such topics as decision analysis and oper-

ations research from their core set of requirements. Others have moved to the creation of tracks or concentrations, with some providing more quantitative, analytic rigor. Some have approved these changes as responding to the needs of a new set of students while others fear that public policy by trying to be relevant to all will lose its distinctive purpose.

3.4 The Intellectual Challenge

Professional education always faces the challenge that it is simply an applied or less rigorous version of the material offered by the traditional academic professions. Frequently this charge is misplaced. The need to apply theory to actual situations often leads to new breakthroughs—sometimes theoretical breakthroughs. Such was the case with the creation of the modern theory of finance in business schools.

It was hoped that implementation theory or principal-agent theory would be the public policy equivalents of the business schools' finance theory. But this has not occurred. Implementation theory has never got beyond its inductive beginnings and principle-agent theory is limited when applied to the multiple-goal, multiple-agent world of the public sector (Moe 1984). This has meant that public policy education has remained derivative and applied. For the long-term health of the public policy movement within the academy, overcoming this weakness is perhaps the greatest challenge to the field.

See also: Policy Knowledge: Advocacy Organizations; Policy Knowledge: Epistemic Communities; Policy Knowledge: Nonprofit Institutions; Policy Knowledge: Universities

Bibliography

Bardach E 1994 Comment: the problem of best practice research. *Journal of Policy Analysis and Management* **13**(2): 260–8

Bardach E 2000 *A Practical Guide for Policy Analysis: the Eightfold Path to More Effective Problem Solving*, Part III, Chatham House Publishers, New York

deLeon P, Steelman T A 1999 The once and future public policy program. *Policy Currents: Newsletter of the Public Policy Section of the American Political Science Association* **9**(2): 1–9

Ellwood J W 1984 A Morphology of Graduate Education for Public Service in the United States. Report for the Study of Trends and Innovations in Graduate Education for Public Service. National Association of Schools of Public Affairs and Administration

Fleishman J L 1990 A new framework for integration: policy analysis and public management. *American Behavioral Scientist* **33**(6): 733–54

Lerner D, Lasswell H D 1951 *The Policy Sciences*. Stanford University Press, Palo Alto, CA

Light P C 1999 *The New Public Service*. The Brookings Institution Press, Washington, DC

Lynn L E 1994 Public management research: the triumph of art over science. *Journal of Policy Analysis and Management* **13**(2): 231–59

Moe T M 1984 The new economics of organization. *American Journal of Political Science* **28**(4): 739–77

Wilson W 1887 The study of administration. *Political Science Quarterly* **2**: 197–222

J. W. Ellwood and E. Smolensky

Public Psychiatry and Public Mental Health Systems

Public psychiatry in the US encompasses several different elements: buildings, such as state mental hospitals; programs, such as community mental health centers; and insurance mechanisms, such as Medicaid and Medicare, that enable poor or elderly recipients (respectively) to receive care in private facilities. Public mental health systems are found at all levels of government: the federal government runs one of the largest health care systems in the world, the Veterans' Administration Medical System, all individual states run state psychiatric hospitals and systems, and local governments (e.g., cities and counties) run hospitals and clinics. In this article, both public psychiatry and public mental health systems will be discussed simultaneously, starting with their historical context and moving on to a more detailed examination of their elements.

1. The History of Public Psychiatry and Public Psychiatric Systems

For many years after the US was colonized, there were no services for the mentally ill. In time, local communities built jails, workhouses, and almshouses or poorhouses, intended for 'rogues and vagabonds,' the 'idle and disorderly,' and the needy, respectively, to which mentally ill persons were sent if violent, disruptive, or simply wandering. While the first facility for the mentally ill, the 'Publick Hospital' in Williamsburg, Virginia, opened in 1752, it was not until the 1840s that Dorothea Dix, a pioneering advocate for the mentally ill, sought higher governmental responsibility for this population, and as a consequence, state mental hospitals and systems flourished (Talbott 1978). The federal government entered the scene in the eighteenth century with the creation first of Public Health Service Hospitals for the Merchant Marine,

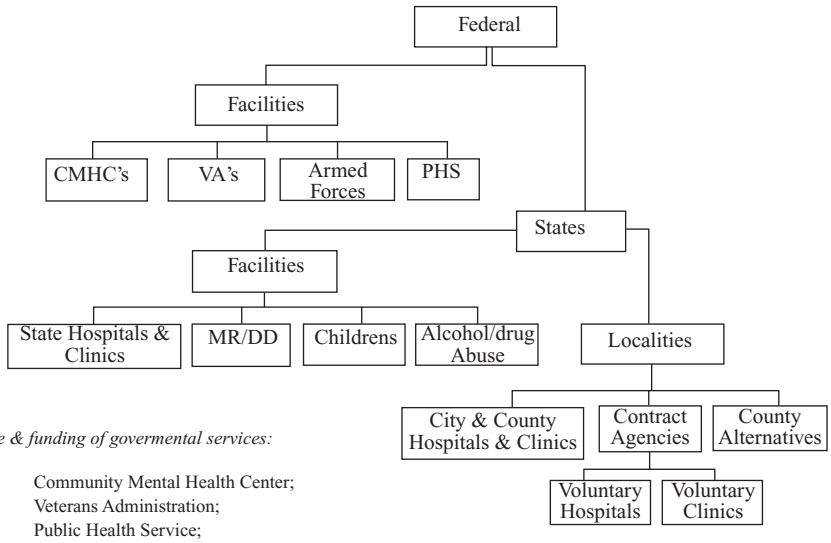

Structure & funding of govermental services:

CMHC Community Mental Health Center;
VA Veterans Administration;
PHS Public Health Service;
MR/DD Mental Retardation/Development Disability.

Figure 1
Structure and funding of governmental services

then Veterans Hospitals and finally, in 1963, Community Mental Health Centers. In some ways, the federal government's commitment to returning the treatment and care of the mentally ill back to their communities represented a coming full-circle from the jails, workhouses, and almshouses or poorhouses of colonial days, but now this responsibility was coupled with the technology, programs, and staffing necessary to provide effective care nearer to home, family, and work. Thus, by the turn of the twentieth century there was a rich variety but often confusing multiplicity of mental health services sponsored by different governmental levels and agencies extant in this country (Fig. 1).

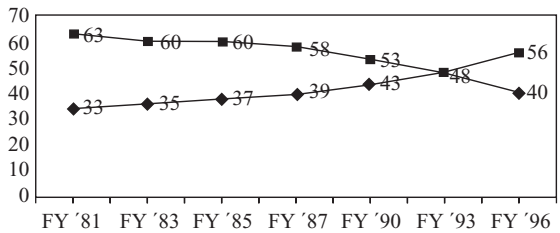

Figure 2
State Mental Health Agency controlled expenditures for state mental hospital inpatient (◆) and community-based (■) services: 1981–1996

2. *City and County Care*

As mentioned above, the very first public institutions in which the mentally ill were cared, were county or city jails, workhouses, and almshouses or poorhouses. While these 'houses' existed for a number of decades (Rothman 1971), in the 1880s the mentally ill began to be cared for and treated in separate psychiatric hospitals, established by individual states. However, to this day, some cities (e.g., New York, New Orleans, San Francisco) and some counties (e.g., Cook (Chicago), Milwaukee, San Diego) have city and county general and/or psychiatric facilities. In addition, as also was mentioned above, the federal legislation in the 1960s creating community mental health centers plus the devolution of the state's single authority to counties in the 1980s (Hogan 1999) have ensured a prominent role for local entities in taking public responsibility for the mentally ill at the close of the twentieth century. In terms of numbers, at their height in 1955, city and county mental health facilities cared for only 100,000 persons, compared to the 460,000 persons residing in state hospitals. Today there are fewer than a dozen city or county facilities remaining, caring for 10,000 patients. Recently, the Robert Wood Johnson. Foundation initiated a project in nine cities to 'devolve' mental health services from state to city 'quasi-public authorities' (Goldman et al. 1990). Because of these and other initiatives to move from (state) institutional to (local) community care, the majority of moneys allocated by each state now goes to community-directed mental health care (Fig. 2).

3. State Mental Health Services

As mentioned above, the first facility for the mentally ill, the 'Publick Hospital' in Williamsburg, Virginia, opened in 1752. Called 'public(k),' the hospital was actually private and it was not until the 1820s and 1830s that state governmentally supported and run psychiatric hospitals came into being. While Massachusetts and New York pioneered in opening new facilities to accommodate for the need for treatment and care of the mentally ill, other states lagged behind until the 1840s when Dorothea Dix, a Sunday-school teacher, sought first federal and then state initiatives to care for all the mentally ill in America. From this point to 1955, state mental hospital systems grew; first slowly, and then with the waves of new immigrants, more quickly (Talbott 1978). By 1955, about 460,000 Americans were housed in such facilities (Fig. 3).

The role of the state mental hospital throughout its first century was comprehensive and inclusive; it provided almost everyone almost all the care and treatment then imaginable and available, including housing, food, clothing, supervision, social services, medical, psychiatric and dental services, and leisure time activities. If the number of admissions grew, new hospitals were built. Alternatives to state hospitals began to appear in the US starting in 1855 beginning with the Farm of Sainte Anne's in Illinois, and eventually encompassed cottages on the grounds, 'boarding out,' 'aftercare,' or outpatient clinics, traveling clinics, crisis intervention, satellite clinics, day, night, and weekend hospitals, social and vocational rehabilitation, home care, and halfway houses. However, the creation of all these 'alternatives' did not reverse or halt the increasing population treated in state facilities—until 1955. In that year, deinstitutionalization of these large facilities began as the result of a combination of technological (a new, the first, antipsychotic medication) and philosophical (that care in the 'community,' nearer family, work, and friends, was better than that in 'asylums') initiatives, soon to be joined by legal (the rights to treatment, to refuse treatment, and to be treated in less restrictive settings) and economic (medical insurance via Medicaid and Medicare for treatment in the community and income support via Supplemental Security Income for housing in the community) forces. In the next 30 years, the population in state hospitals fell 80 percent. Thus, at present, there are fewer than 90,000 residents in state facilities. It must be noted, however, that state governments were not only responsible for state mental hospitals (Fig. 1), but for 'aftercare' and outpatient clinics, as well as facilities and clinics for children and persons suffering from alcohol and drug abuse and mental retardation/developmental disorders, and these services have continued to this day.

4. Federal Mental Health Services

4.1 Veterans' Hospitals

While pensions, cash payments, and land grants were awarded to veterans of the Revolutionary War by grateful states and the federal government, it was not until the American Civil War (1861–1865) that national veterans' homes were established (Godleski et al. 1999) and not until after World War I, in the 1920s, that health benefits were offered to all veterans.

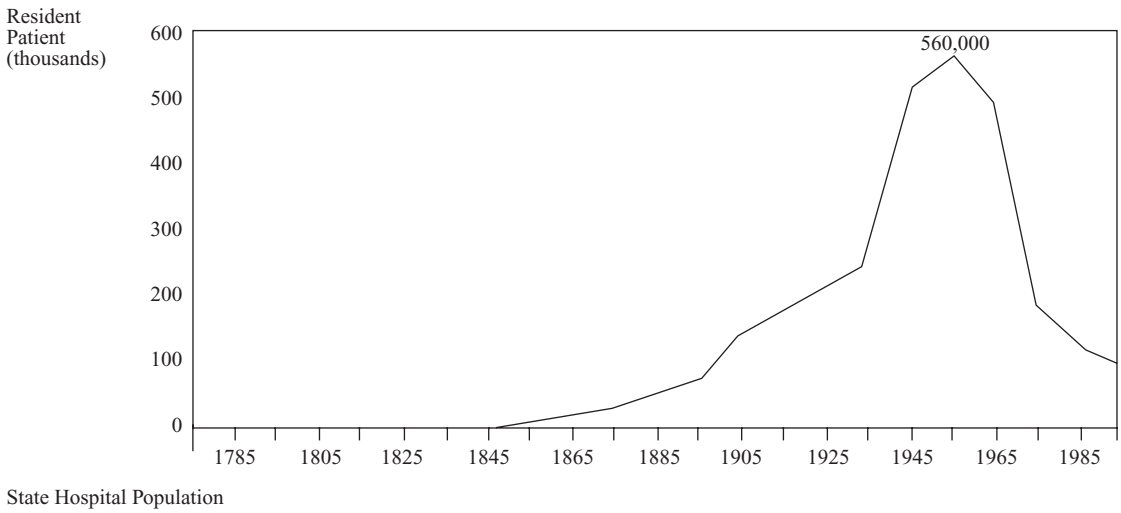

Figure 3
State hospital population

By 1930, 4.7 million veterans were being served by 54 hospitals and in 1995 the 'Veterans Administration' system of care included 172 medical centers, 375 outpatient clinics, 133 nursing homes, 39 'domiciliary' care facilities, and 202 'Vet Centers' (largely urban drop-in centers).

4.2 Military Psychiatric Services

Physicians have always accompanied American forces during war. The recognition of the role of stress in combat was first documented in America during the Civil War when 'combat heart' was often diagnosed. In World War I, psychiatrist Thomas Salmon visited the battle lines in France before America's entry and it was his observation of the difference in the practices and outcomes of the French (treatment near the lines, expectation of recovery; with low resultant disability) and English (evacuation; with high disability) in treating 'shell-shock' victims that predicated the American military's practice henceforth. In World War II and in Korea, these lessons had to be re-learned but by Vietnam, American psychiatrists were inducted into the Armed Forces in significant numbers and assigned to both (a) hospitals in combat and evacuation areas and (b) line troop divisions, where they practiced 'preventive psychiatry' (Talbott 1966). Currently, to serve a 'peacetime military' of almost 1 million men and women, there are approximately 300 psychiatrists on active duty.

4.3 Public Health Service Hospitals

In 1798, Congress established the US Marine Hospital Service—the predecessor of today's US Public Health Service—to provide health care to sick and injured merchant seamen (History of the Office of the Surgeon General 1999). Eventually, 25 hospitals were established across the country (Mullen 1989). In 1981 these hospitals began being phased out and at present the Surgeon General's Office and the Public Health Service, while active in articulating policy (on smoking, the treatment of schizophrenia, etc.) and employing 6,000 persons who serve as a 'uniformed cadre of health professionals on call 24 hours a day, 7 days a week in the event of a public health emergency,' does not run nonresearch related clinical services.

4.4 The National Institutes, Administrations, and Centers

The first element in the federal array of research laboratories and agencies to be established was the Laboratory of Hygiene in 1887; the first psychiatric component was the Division of Mental Hygiene of the US Public Health Service in the 1930s. The National Cancer Institute was the first of the now 25 national institutes now comprising the National Institutes of Health (NIH) to be established for research; the National Institute of Mental Health (NIMH) coming into being in 1946. The research mission was broadened in the 1950s and 1960s to include teaching and services (e.g., community mental health centers) as well as alcohol and substance abuse treatment and prevention. Soon there were Institutes on Drug Abuse (NIDA) and Alcohol Abuse (NIAA) and in 1968 the Health Services and Mental Health Administration was established as an umbrella organization to house all of them. In 1973, NIMH rejoined the NIH and later came under a new 'umbrella,' the Alcohol, Drug Abuse and Mental Health Administration (ADAMHA). Finally, in 1992, prevention, services, and services research were separated off into a new agency—the Substance Abuse and Mental Health Services Administration (SAMSHA), while the NIMH, NIDA, and NIAA remained in the NIH as 'research institutes.'

4.5 The Community Mental Health Center

As mentioned above, since colonial days, localities had always provided some services. But, in 1963, the shift from localities to states to federal auspices came full circle, and it was the federal government that legislated the establishment of what was intended originally by President John F. Kennedy's administration to be about 750 community mental health centers (CMHCs), each of which was to serve 75,000–150,000 persons through the provision of 10 essential services (inpatient, outpatient, partial hospital, emergency, consultation and education, diagnostic, rehabilitative, precare and aftercare, training and research, and evaluation) (Glasscote et al. 1969). While the 1963 legislation had to be amended in 1975 to include care for children and the elderly and those suffering from alcohol and drug abuse as well as transitional housing, follow-up care, and screening for courts and agencies prior to hospitalization, the model of the CMHC was a powerful one in the US for most of the last half of the twentieth century. At present, while the US cannot claim to have either comprehensive or universally available, federally initiated community mental health centers, there are 675 in the nation (Ray and Finley 1994).

4.6 Federal Entitlements for Community Care

The blind and needy aged have been provided federal 'Social Security' to enable them to make ends meet

since 1935, and in 1950 persons suffering from other disabilities, including mental disorders, were included under two additional programs: Supplemental Security Income (SSI) and Social Security Disability Income (SSDI). In 1966, two other large 'public' populations, the poor and the aged and disabled were guaranteed medical insurance through two federal entitlement programs, Medicaid and Medicare, respectively. What no one had realized at the time was the fact that with these four massive federal entitlement programs, the mentally ill could 'afford' to live and receive treatment from 'private' clinics and general hospitals rather than 'public' state hospitals. This further added to the technological, philosophical, and legal forces moving psychiatric care and treatment in America from institutional to community care; albeit publicly financed. In 1997, 13.4 percent (37,514 million) of the population was covered by Medicare and 13 percent (35,028 million) by Medicaid; by 1994, 16.5 percent (42,878 million) of the population was covered by Social Security, 10 percent (26,281 million) by SSI, and 1.5 percent (5,893 million) by SSDI (US Bureau of the Census 1998).

4.7 Community Support and Case Management

Starting in 1977–1978, the NIMH provided grants to community agencies to initiate community support programs. Such a system was to include: identification of the population, assistance in applying for entitlements, crisis-stabilization, psychosocial rehabilitation, supportive services, medical and mental health care, back-up support to families, involvement of concerned community members, protection of client rights, and case-management (Shaftstein et al. 1978). Like the CMHC, this model was compelling and moved community services to adopt such strategies and services even in the absence of federal moneys. Most widely undertaken have been case management programs for high-intensity and average utilizers. It is impossible to estimate how many community support and case management systems there are in America today, but it is in the tens of thousands.

4.8 The Indian Health Service

Members of Federally recognized Indian tribes and their descendants have been eligible for services provided by the Indian Health Service (IHS) since 1787. The IHS, part of the US Public Health Service, Department of Health and Human Services, operates a comprehensive health service delivery system for approximately 1.5 million of the nation's two million American Indians and Alaska Natives and in 1999 its annual appropriation was approximately $2.2 billion. There are more than 550 federally recognized tribes in the US whose members live mainly on reservations and in rural communities in 34 states, mostly in the western US and Alaska. IHS services are provided in three ways; by the IHS directly, through tribally contracted and operated health programs, and by purchasing care from more than 2,000 private providers. As of March 1996, the federal system consisted of 37 hospitals, 64 health centers, 50 health stations, and five school health centers. In addition, 34 urban Indian health projects provide a variety of health and referral services. The IHS clinical staff consists of approximately 840 physicians (approximately 80 of whom are psychiatrists), 100 physician assistants, and 2,580 nurses. As of March 1996, American Indian tribes and Alaska Native Corporations administered 12 hospitals, 116 health centers, three school health centers, 56 health stations, and 167 Alaska village clinics.

5. The Correctional System

The correctional system in America cuts across all governmental levels: federal, state, and local (city/county). It may well house more mentally ill persons than any other 'governmental health system.' There are about 10 million adults booked into US jails each year and more than 1.5 million persons currently incarcerated. The US Bureau of Justice estimates a rate of mental illness in local and state jails of over 16 percent and in federal prisons of over 7 percent; thus over 250,000 mentally ill persons are in prison and jail on any one day (Ditton 1999). It is not known how many psychiatrists are employed in the correctional system nor how many services are rendered.

6. Public Managed Care

In the 1930s the Kaiser Corporation, a large industrial concern, formed the first employer-run group practices in America to reach its employees working in rural areas. These tightly run group practices became known as health maintenance organizations (HMOs) and in the 1970s were promoted by the Nixon Administration. HMOs, along with other forms of tightly managed forms of service provision, have grown and expanded from their original base serving employed persons through industry or 'private' insurance to serving 'public' populations, e.g., the aged, disabled, and poor who are insured by Medicare and Medicaid. The popularity of such 'managed care' as a method of cost-control and provision is apparent from the fact that, in 1992, the percentage of Americans covered by managed-care programs exceeded those covered by 'traditional' (e.g., Blue Cross/Blue Shield) insurance

plans. In the 1990s, the federal and state governments have actively sought to enroll their publicly funded populations (e.g., those receiving Medicaid and Medicare) in such programs. Estimates of the percentage of Medicare recipients enrolled in managed care programs are at present over 10 percent nationally; regarding Medicaid, at present every state in the union has begun a 'managed Medicaid program,' has applied for federal government permission to do so, or is actively discussing it.

7. The Future

The future of public psychiatry and public mental health systems is unknown. While some systems (e.g., county and state systems) have endured for almost the entire three centuries of our young country's existence, others (such as public managed care) are growing at a rate that will surely diminish the importance and even existence of these traditional systems. What is known, however, is that until poverty and severe mental illness no longer exist there will be public psychiatrists and public mental health systems.

See also: Mental Health: Community Interventions; Mental Health Programs: Children and Adolescents

Bibliography

Ditton P M 1999 Mental health and treatment of inmates and probationers. *Bureau of Justice Statistics*, NCJ 174463

Glasscote R M, Sussex J N, Cumming E, Smith L H 1969 *Community Mental Health Center: An Interim Appraisal.* Joint Information Service of the American Psychiatric Association and the National Association for Mental Health, Washington, DC

Godleski L S, Vadnal R, Tasman A 2001 Psychiatric services in the veterans' health administration. In: Talbott J A, Hales R E (eds.) *The Textbook of Administrative Psychiatry: New Concepts for a Changing Behavioral Health System*. American Psychiatric Press, Washington, DC

Goldman H H, Morrissey J P, Ridgely M S 1990 Form and functions of mental health authorities at RWJ Foundation program sites: Preliminary observations. *Hospital and Community Psychiatry* **41**: 1222–30

History of the Office of the Surgeon General 1999 http://www.surgeongeneral.gov/osg/sghist.htm

Hogan M F 2001 State and county agencies. In: Talbott J A, Hales R E (eds.) *The Textbook of Administrative Psychiatry: New Concepts for a Changing Behavioral Health System.* American Psychiatric Press, Washington, DC

Indian Health Service Internet Home Page 1999 http://www.ihs.gov

Mullen F 1989 *Plague and Politics: The Story of the United States Public Health Service*. Basic Books, New York

Ray C G, Finley J K 1994 Did CMHCs fail or succeed? Analysis of the expectations and outcomes of the community mental health movement. *Administration and Policy in Mental Health* **21**(4): 283–93

Rothman D J 1971 *The Discovery of the Asylum: Social Order and Disorder in the New Republic*, 1st edn. Little, Brown, Boston

Shaftstein S S, Turner J E, Clark H W 1978 Financing issues in providing services for the chronically mentally ill and disabled. In: Talbot J A (ed.) *Ad Hoc Committee on the Chronic Mental Patient. The Chronic Mental Patient: Problems, Solutions, and Recommendations for a Public Policy*. American Psychiatric Association, Washington, DC

Talbott J A 1966 Community psychiatry in the army. *Journal of the American Medical Association* **210**(7): 1233–7

Talbott J A 1978 *The Death of the Asylum: A Critical Study of State Hospital Management, Services, and Care*. Grune and Stratton, New York

US Bureau of the Census 1998 *Statistical Abstract of the United States*, 18th edn. Washington, DC

J. A. Talbott

Public Reason

'Public reason' implies a contrast to 'nonpublic' or 'personal reason.' Identifying a distinctly public or shared notion of reason is only important—indeed is only intelligible—when the concept of reason is understood in ways which allow that individuals, employing good reasoning, may come to divergent conclusions. This article begins by examining the lines of inquiry that have led many to question whether all fully-informed and fully-competent inquirers would come to the same conclusions about science and morals. According to some philosophers and social scientists, the very concept of reason has broken apart, leaving different individuals and groups with different conceptions of rationality, and so of reasons for belief and action. The article then considers the different ideals of public reason that have been advanced to overcome worries about the 'fragmentation of reason.'

1. The Diversity of Reason

1.1 The Enlightenment

Before considering different conceptions of public reason, it is important to grasp what the very idea of public reason presupposes. According to the conception of reasoning that dominated the European Enlightenment of the seventeenth and eighteenth centuries, reason is inherently universal, and so shared and public. On this Enlightenment conception, reason is a shared capacity of all human beings, and norms of

good reasoning are universal. Any premise p that is true for one person is necessarily true for all others; if the inferential rule '$(p \cdot (p \rightarrow q)) \rightarrow q$' is valid for one person, it is necessarily valid for all. The true and valid results of one person's reasoning are thus necessarily true and valid for all. According to this standard Enlightenment view, the use of common human reason not only produces consensus on scientific beliefs, it also leads to convergence of moral and political opinions. To be sure, people might disagree on matters of science or ethics, but that would be due to mistaken beliefs or irrationality: some have arrived at the wrong answer. The process of Enlightenment was the only remedy for this—the increasingly better use of reason to uncover truths about the natural and the social world. The ideal or model was that of Newtonian physics: just as our common reason had uncovered the laws of matter and motion, so too could it be expected to uncover the laws of human nature, society, morals, and politics. Each field was awaiting its Newton.

1.2 Relativism and Social Science

It is a familiar story how the systematic study of cultures with radically different norms from those of Europe gave rise to cultural relativism. When confronted with the very different mores of the Zuñi and Kwakiutl Indians and Dobu Islanders of Papua New Guinea, Ruth Benedict (1934) endorsed a relativistic view—their norms were justified for them, just as Western norms were for Europeans. As anthropologists began to study non-European cultures in more depth, and became increasingly more sophisticated and self-reflective in their techniques, disputes arose about the proper method for anthropologists to interpret the belief systems of other cultures. For example, was a belief in magic or a deity to be understood as simply a false belief on which the 'natives' relied? John Stuart Mill (1974 [1843], pp. 766–7) believed that an investigator could best understand the magical beliefs of other cultures by attributing to them erroneous beliefs and invalid inferential rules. Others, adopting a *principle of charity*, have insisted that the best interpretation of a culture minimizes the number of false beliefs attributed to its members (Davidson 1973). Thus the best interpretation—which makes sense of the 'native's' metaphysical theories, religious convictions, and their beliefs about nature—might seek to show that a belief in spirits is, after all, rational given their world view. Thus the second step in the relativist attack on the Enlightenment's ideal of reason was to endorse *relativism concerning what beliefs are rational* in different cultures. The last step in the relativistic project is to apply tolerance to the idea of reason itself. Can Western anthropologists properly interpret other cultures if they apply their Western conception of reason in their interpretation? Is the very idea of reason culturally relative?

Independently, psychological studies of human reasoning led to doubts about whether everyone shares the same norms of reasoning. The work of, among others, Richard Nisbett, Lee Ross, Paul Slovic, and Amos Tversky showed that normal adults often do not employ the norms of reasoning long advocated by philosophers as correct. For example, many normal adults adhere to the 'gambler's fallacy.' According to standard probability theory, the odds of independent events occurring do not depend on whether that type of event has occurred in the past: if the odds of a 'six' being rolled by a fair dice is one-in-six, this probability in no way depends on whether no sixes have been rolled or all sixes have been rolled in the past. Yet many normal reasoners believe that, after a long run of sixes, the odds of another six are less than one-in-six, or if none recently have been rolled, many are convinced that a six is 'due.' Based on a variety of studies of such ordinary reasoners, some philosophers advocate 'normative cognitive pluralism.' According to Stephen Stitch, '*Normative cognitive pluralism* is not a claim about the cognitive processes people do use; rather it is a claim about *good processes*—the cognitive processes that people ought to use. What it asserts is that there is no unique system of cognitive processes that people should use, because various systems of cognitive processes that are very different from each other are equally good' (1990, p. 13).

1.3 Pluralism and Indeterminancy

Social science thus provided one source of the 'fragmentation of reason.' Of course there was, and still is, lively debate within social scientific and philosophical circles about whether any such relativistic view of reasoning is justified, and if so to what extent individuals employ different, but equally good, norms of reasoning. Nevertheless, some form of cognitive relativism is widely accepted. Concurrently, developments in the philosophy of science were challenging the idea of common, human, reasoning in the very citadel of the Enlightenment—science itself. Whatever doubts may be entertained about the application of Western reason to other cultures, or departures from expert norms by ordinary reasoners, surely science is still the model of rational discourse tending to convergence of opinion leading to the truth. To many, Thomas Kuhn's analysis of scientific practice demonstrated that, rather than a shared project of pursuit of the truth, judged by rules of reason held by all, scientific inquiry takes place within 'paradigms' that determine not only what constitutes research problems but what constitutes success in a research program. *Within* a group of scholars sharing a paradigm the Enlightenment's ideal of common norms of reason and consensus on the truth will be approximated; and insofar as 'normal science' is dominated by a single paradigm in each field, there often is a common mode of reasoning. But in times of scientific crisis,

when paradigms compete, some have insisted that there is no rational, impartial way to adjudicate among the competing theories.

The crux of Kuhn's analysis is that because data underdetermines theories, scientists draw on values in adjudicating among theories. 'To a greater extent than other sorts of components of the disciplinary matrix,' writes Kuhn, 'values may be shared by men who differ in their application. Judgments of accuracy are relatively, though not entirely, stable from one time to another and from one member to another in a particular group. But judgments of simplicity, consistency, plausibility, and so on often vary greatly from individual to individual … Even more important, in those situations where values must be applied, different values, taken alone, would often dictate different choices. One theory may be more accurate but less consistent or plausible than another' (Kuhn 1970, p. 185). The crucial claim is that there is no uniquely rational way to order these various desiderata—simplicity, consistency, plausibility—and different orderings endorse different scientific theories upholding competing truth claims. Equally well-informed scientists employing their reasoning in perfectly legitimate ways can arrive at different judgments about what is true. Fred D'Agostino has generalized the analysis. In any field of inquiry, when there exists both a plurality of criteria and no impartial way to order the criteria, the criteria are apt to be indeterminate in their application. Different orderings produce different outcomes—reasons for belief. According to D'Agostino, the criteria only yield a determinate result when an ordering is provided, but impartial reason cannot give us that. Much of postmodernism is based on this key idea. Reason is inherently perspectival because of the inherent plurality of relevant considerations for which there is no rational, impartial ranking.

1.4 Reasonable Pluralism

Kuhn's analysis of scientific practice and D'Agostino's generalization rely on the idea of indeterminancy: reason undetermines choice between various theories and perspectives. A more modest view maintains that our powers of reasoning are inconclusive on many complex matters of science, morality, and politics. Claim C is characterized by reasonable pluralism if some perfectly reasonable agents do, while others do not, have good reasons for accepting C. For a controversy to be characterized by reasonable pluralism it is not simply the case that people actually disagree about the merits of C, but that it is reasonable for one person to assert C and another not-C. This is a far less radical claim than was examined in the last section: it is not argued that the question 'C or not-C?' is inherently *indeterminate* but only that present beliefs about C are *inconclusively* justified. According to John Rawls, our disputes about value seem subject to reasonable disagreement because our understanding of what is good and valuable is especially subject to what he has called the 'burdens of judgment.' According to Rawls, reasonable judgments so often are at odds because:

(a) the evidence is often conflicting and difficult to evaluate;

(b) (as in Kuhn's example) even when we agree on the relevant considerations, we often weigh them differently;

(c) because our concepts are vague, we must rely on interpretations that are often controversial;

(d) the manner in which we evaluate evidence and rank considerations seems to some extent the function of our total life experiences, which of course differ;

(e) because different sides of an issue rely on different types of normative considerations, it is often hard to assess their relative merits;

(f) in conflicts between values, there often seems to be no uniquely correct answer (Rawls 1993, p. 57).

Because these matters are so complex and uncertain, different people will reach different, competing, credible, or reasonable conclusions. In the face of complexity, our powers of reasoning do not produce convergence.

2. Conceptions of Public Reason

2.1 Public Rationality and Public Reasons

Important movements in social science, the philosophy of science, and moral and political philosophy thus have combined to cast grave doubt on the Enlightenment's confidence that the free exercise of human reason leads us to agree. In the face of these challenges to the ideal of common human reason, theories of distinctly public reason have been advanced to show how people can *share reasons* (Korsgaard 1996, Chap. 10). However, this idea is ambiguous. It may constitute a response to relativistic social science (Sect.1.2), which challenges the very idea of a shared conception of human reason. Understood thus, the ideal of public reason is an ideal of *public rationality*, or a shared conception of rationality. The challenge posed by reasonable pluralism (Sect. 1.4) is more modest. Advocates of reasonable pluralism maintain that even if we apply shared standards of reason, our deliberations will not yield shared reasons for belief. In a weaker sense, then, the ideal of public reason is an ideal of *public reasons*, or how we can arrive at shared reasons to believe and act. A theory of public reason may advance a theory of public rationality and/or public reasons.

Three main conceptions of public reason stand out: (a) the epistemic, (b) the consensual, and (c) the political.

2.2 Epistemic Theories of Public Reason

In philosophy, 'epistemology' refers to the theory of knowledge or, more broadly, the theory of justified belief. Epistemic theories of public rationality, then, maintain that at least some common norms of reasoning can be justified to everyone. The Enlightenment view was that *all* valid norms of reasoning can be justified to everyone—reason was inherently public. Epistemic theories of public rationality need not, however, advance such a sweeping claim. Importantly, an epistemic theory of public rationality may adopt a modest version of the 'relativism of reasons' (Sect. 1.2), according to which in *some* cases person Alpha may have a fully justified norm of reasoning N while Beta accepts a fully justified norm of reasoning M, where it is the case that N and M are inconsistent. In cases where only N and M are relevant to deliberation, a relativism of rationality will manifest itself; Alpha may have reasons based on N that Beta will deny on the basis of M, and both will be justified. However, to establish a thoroughgoing relativism of reasoning that undermines the possibility of public reason, it is not enough to establish that this can *sometimes* occur; it must be shown that *no* common norms of rationality can be justified to all—that it is *always* (or at least usually) the case that for any justified norm of reasoning held by Alpha, some person Beta will be justified in accepting a competing norm. That is the sweeping relativistic claim rejected by epistemic theories of public rationality (Gaus 1996).

Advocates of epistemic public rationality can reply to both the anthropological argument for the relativism of reasons, and psychological studies uncovering the diversity of norms employed in reasoning (Sect. 1.2). In reply to the *anthropological argument*, advocates of public reason can embrace Martin Hollis' view that an anthropologist must build up an interpretation of an alien language by locating a 'bridgehead.' The assumption underlying the bridgehead strategy is that others generally perceive what we perceive and tend to say about it the sorts of things we would say. Relying on these assumptions, the anthropologist begins by translating everyday, basic perceptual sentences such as 'Yes, this is a brown cow' and 'No, it is not raining right now.' Now, Hollis insists, this bridgehead not only includes translations of such basic beliefs, but basic logical rules, since 'what a sentence means depends on how the beliefs which they express are connected, and that to justify a claim to have identified a belief one must show the belief is connected to others.' Consequently, in the main, the logic employed by those in different cultures must either be very similar to our own or unintelligible. Axioms such as '$p \rightarrow p$' '$\sim(p \cdot \sim p)$' and '$(p \cdot (p \rightarrow q)) \rightarrow q$' are more than simply *our* norms of reasoning. They are constituents of any system that we can identify as a way of reasoning (Hollis 1970a, pp. 214–15, 1970b, p. 232.) If 'natives' reason logically at all, then they reason as we do. The point, then, is that the context that sets the stage for the possibility of relativism—that we confront others, different from us but of whom we can make some sense—presupposes widespread shared norms of inference as well as beliefs. Thus the very possibility of mutual intelligibility sets a limit on the extent to which we can understand others as employing cognitive processes different from our own.

In reply to psychological studies showing the diversity of norms employed by normal reasoners, advocates of public reason can rely on evidence supporting the hypothesis that people employ a mental logic—a system of mental inference rules—in their actual reasoning. Indeed, most psychologists studying actual reasoning suppose some version of the mental logic theory. To be sure, mental logic may depart from the standard propositional calculus. Psychologists studying English reasoners, for instance, have found that English speakers do not interpret connectives in a way consistent with standard logic. Particularly troubling are interpretations of 'if, then' statements (as well as 'or' connectives), which are not typically used in ways that correspond to standard analyses in logic. In response to these findings, psychologists such as Lance J. Rips and Martin D. Braine have developed sets of natural inferential rules that correspond to the way in which speakers of English employ connectives. Rips has gone beyond this to develop rules modeling causal reasoning as well.

Mental logic views strongly incline toward the hypothesis that these logics are in some way natural or innate, though to precisely what extent is open to investigation. There may well be a complicated interplay between language and logic, such that English speakers tend to employ some inferential rules not utilized by others. A plausible hypothesis is that the capacity to develop a mental logic is innate, though the precise developmental path taken depends on a variety of factors, including one's native language. If so, however, this is sufficient to undermine the plausibility of a radical relativism of norms of rationality. The set of possible mental logics—laws governing thought—is apt to be pretty severely constrained.

2.3 Consensus Theories of Public Reason

Epistemic accounts of public reason uphold our ability to reason together because they insist that we do, indeed must, share basic rules of inference. In contrast, consensual accounts of public reason conceive it as arising out of a shared social life and/or shared discourse (see *Shared Belief*). A number of otherwise diverse philosophers such as Ludwig Wittgenstein, Jürgen Habermas, Richard Rorty, Susan Hurley, and Philip Pettit all point to the idea that reasoning is a social phenomenon. In order to be a deliberator, one must be a member of a community in which one's beliefs and rules are intelligible to others. One comes

to understand what is a justified or unjustified belief, and what is a good or a bad way of reasoning, by sharing the norms and standards of communities, including communities of inquirers. Thus some insist that 'ultimately, there is only one criterion by which beliefs can be judged valid, and that is that they are based on agreement reached by argumentation'—we converge on them (this is Habermas's 1991, p. 14 gloss on Rorty's 'particular version of discourse theory'). Habermas draws an intimate connection between mutual understanding (*Verständingung*) and agreement (*Einverständnis*). It would seem that to understand others we must in some way arrive at intersubjective agreement about norms of rationality and good reasons. Consequently, the phrase of 'public reason' becomes almost redundant: something can only be identified as reason because it is the object of public, interpersonal, agreement.

Wittgenstein and his contemporary followers such as Hurley (1989) and Pettit (1993) suggest a broadly similar view. Reasoning itself presupposes a 'community'—common rules of thought. Wittgenstein (1958) argued that one could not have a private rule that one only followed once—for in such a case we could not know whether the rule was actually followed. More generally, it has been argued that one cannot have a purely private rule, for one needs to be able to distinguish when one has correctly followed the rule and when one has gotten it wrong. But, assert Wittgensteinians, the difference between correct and incorrect application of a rule cannot be located in individual belief systems, but only by the individual comparing her application of the rule to those of her fellows: each must look to others to check and correct her rule-based performances. Since thinking is necessarily rule driven, the analysis applies to reason itself. Thus it seems that the very idea of rule-following—and, so, of reasoning—supposes convergence with others. If so, the search for reason is a search for consensus.

In one respect this view, somewhat surprisingly perhaps, returns to the basic idea of the Enlightenment—all reason is public reason. But whereas the Enlightenment was confident in asserting this because of the widespread faith in a universal, common, human faculty of reason that could accurately represent reality, consensual theorists see reason as arising out of a shared public life. To put it crudely, rather than reason producing consensus, consensus generates reason. Thus the publicness of such reason, unlike that of the Enlightenment's, is consistent with a cultural relativistic conception of rationality. The public, intersubjective agreement, reached by one culture—which is constitutive of their very conception of rationality—may be very different from that upheld by another.

Because consensual accounts of public reason stress the constitutive nature of public consensus to the very idea of reason, and because public consensus may be promoted by political institutions, more radical consensual theorists such as D'Agostino give political institutions a constitutive role in determining *what reason is* (1996, Chap. 9). This is a radical view indeed—however, it is not entirely novel, perhaps going back as far as Hobbes. For Hobbes ([1651] 1948), the conflict in the state of nature arises from conflicting private judgments; people's private reasoning yields conflicting judgments of right and wrong, as well as matters of fact, and this leads to the less intellectual conflict that characterizes the state of nature. Hobbes's solution is to appoint an 'arbitrator.' This 'judge,' says Hobbes, provides 'public reason' to which private reason 'must submit.' The arbitrator proclaims what each has reason to do, and so defines a single, coherent conception of reason. Because, in Hobbes' view, we have authorized the judge (i.e., sovereign) to define public reason for us, the sovereign's pronouncements constitute a shared public reason on which there is consensus.

2.4 Public Reason qua Political Reason

D'Agostino's radical view allows the political to shape the rational. A more modest proposal with which this radical view is easily confused maintains that the notion of public reason is a purely political idea, in the sense that it concerns only how we reason together *about politics*. This understanding of public reasons lies at the heart of what is commonly known as 'political liberalism.'

Rawls's political liberalism is a response to the problem posed by reasonable pluralism (Sect. 1.4). For Rawls the crucial problem is that citizens in democratic societies entertain a plurality of what he calls reasonable 'comprehensive doctrines'—overall philosophies of life centered on religious, philosophical, or moral beliefs. This pluralism seems a permanent feature of modern societies. 'Political liberalism assumes that, for political purposes, a plurality of reasonable yet incompatible comprehensive doctrines is the normal result of the exercise of human reason within the framework of free institutions of a constitutional regime' (Rawls 1993, p. xvi). The reasonable pluralism of comprehensive views renders such views unacceptable as bases for the justification of political power. Rawls endorses the 'liberal principle of legitimacy,' i.e., that 'our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may be reasonably expected to endorse in the light of principles and ideals acceptable to their common human reason' (p. 137). Thus, because there exists a reasonable plurality of comprehensive doctrines, basing the justification of political power on any of them violates the liberal principle of legitimacy. This leads Rawls to seek a political conception that 'all affirm' and that is 'shared

by everyone.' Such a conception would be supported by the diverse reasonable comprehensive doctrines that characterize our democratic societies. It would generate public reasons that would fit into our many reasonable, but irreconcilable, comprehensive views. Some version of the principle of political legitimacy motivates the search for public reasons in contemporary liberal political philosophy: if the exercise of coercive political power relies on nonpublic reasoning, it is oppressive and illegitimate in relation to those who do not share those reasons.

Rawls's core argument—and so his entire political liberalism—apparently depends on the contrast between comprehensive views and the shared, political, conception. Whereas comprehensive conceptions produce diverse reasons that conflict, the political point of view provides public reasons—reasons that all reasonable comprehensive views can endorse.

Habermas (1995) suggests that Rawls posits an '*a priori*' distinction between the political and nonpolitical spheres. At least at times Rawls seems to suggest that the political point of view can be conceptually distinguished from moral, religious, and philosophical matters; whereas the former identifies a common point of view that can be affirmed by all, when human reason is applied to moral, religious, and philosophical issues it leads to reasonable disagreement. It is difficult, however, to accept the claim that some single conception of the political is shared by all reasonable persons in modern democratic societies. Political theorists have often pointed to the 'political' as an 'essentially contested' concept—one that is open to divergent, reasonable, interpretations (Connolly 1983, Chap. 1). If citizens of modern democratic societies disagree about the concept of the political, it cannot serve the function of identifying public reasons that overcome reasonable pluralism. Like the moral and philosophical, the nature of the political itself seems a matter of reasonable disagreement.

An alternative interpretation of Rawls' view is to take the political conception as *constructed out of* that which is shared. On this reading the *non*political is, by definition, those matters on which the free use of reason by citizens in democratic societies leads to different, reasonable conclusions. It is, by its very nature, the realm of reasonable pluralism. In contrast, one might *define* the political as those matters on which reason converges in democratic societies, and so necessarily generates constitutional principles that satisfy the principle of liberal legitimacy. The political is thus characterized as the overlapping consensus of those diverse comprehensive views characterizing modern societies. The problem for this interpretation is that Rawls himself indicates that the free use of human reason leads us to reasonable disagreement about conceptions of justice and constitutional essentials. The 'political' qua 'shared perspective' is limited to the abstract concept of a 'liberal political order' (Rawls 1993, p. 241). 'By this [Rawls says] I mean three things: first, it specifies certain basic rights, liberties, and opportunities (of the kind familiar from constitutional democratic regimes); second, it assigns a special priority to these rights, liberties and opportunities, especially with respect to claims of the general good and of perfectionist values; and third, it affirms measures assuring all citizens' adequate all-purpose means to make effective use of their basic liberties and opportunities' (p. 223). Thus stated, the content of public reason seems sparse indeed. Unfortunately, if a more detailed articulation of this idea is provided, Rawls indicates that the use of human reason again leads to disagreement. Apparently modern democratic societies agree on only the most abstract of public—political—reasons.

3. *Conclusion: Reasoning in a Plural Society*

Puzzles about the possibility of shared, public rationality and reasons derive from a variety of sources: social science, the philosophy of science, and the moral pluralism of modern societies. Although developments in all these fields serve to cast doubt on the Enlightenment's ideal of a common universal reason, they pose different challenges. The findings of anthropology call into question universal norms of rationality; psychology questions whether so-called universal norms are actually employed in human reasoning. Kuhnian philosophy of science suggests an irremediable indeterminancy of the application of our general standards, even in science itself. Worries about reasonable pluralism focus on the possibility of shared moral and political norms in societies that disagree deeply on questions of value and the nature of the good life.

Just as the challenges arise from multiple sources, so do theories of public reason. *Epistemic theories* dispute that modern social science casts serious doubt on the necessity of shared norms of rationality in the interpretation of other cultures and in individual deliberation. Although some divergence in justified norms of reasoning no doubt occurs—and can be made sense of—radical divergence is implausible. *Consensus theories* of public reason maintain that to make sense to each other we must think alike—we must employ common rules of thought. The application of such common rules, say neo-Wittgensteinians, only makes sense in the context of shared social practices. Reason arises out of shared social life; reason is inherently public because it is necessarily a product of social consensus. Finally, *political liberalism* seeks to establish that, amid the diversity of moral and philosophical views that characterize modern societies, a common political point of view can generate shared public reasons to regulate coercion.

The ideal of public reason, then, is as diverse as the sources of our doubts about the possibility of shared ways of reasoning in contemporary societies. In all its manifestations, however, the ideal of public reason

seeks to maintain the heritage of the Enlightenment—the free use of human reason will lead us to converge on the best justified beliefs.

See also: Bioethics: Examples from the Life Sciences; Economics and Ethics; Ethics and Values; Prisoner's Dilemma, One-shot and Iterated; Utilitarian Social Thought, History of; Utilitarianism: Contemporary Applications; Welfare

Bibliography

Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston, MA

Braine M D S 1978 On the relation between the natural logic of reasoning and standard logic. *Psychological Review* **85**: 1–21

Connolly W E 1983 *The Terms of Political Discourse*, 2nd edn. Princeton University Press, Princeton, NJ

Davidson D 1973 Radical interpretation. *Dialictica* **27**: 314–28

D'Agostino F 1996 *Free Public Reason*. Oxford University Press, New York

D'Agostino F, Gaus G F 1998 *Public Reason*. Ashgate, Aldershot, UK

Gaus G F 1996 *Justificatory Liberalism: An Essay on Epistemology and Political Theory*. Oxford University Press, New York

Habermas J 1990 *Moral Consciousness and Communicative Action*. MIT Press, Cambridge, MA

Habermas J 1995 Reconciliation through public reason: remarks on John Rawls's political liberalism. *Journal of Philosophy* **92**: 109–31

Hobbes T 1948 [1651] *Leviathan*. Basil Blackwell, Oxford, UK

Hollis M 1970a The limits of irrationality. In: Wilson B R (ed.) *Rationality*. Blackwell, Oxford, UK, pp. 214–20

Hollis M 1970b Reason and ritual. In: Wilson B R (ed.) *Rationality*. Blackwell, Oxford, UK, pp. 221–39

Hurley S 1989 *Natural Reasons: Personality and Polity*. Oxford University Press, New York

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgments Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK

Korsgaard C M 1996 *Creating the Kingdom of Ends*. Cambridge University Press, Cambridge, UK

Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

MacIntyre A C 1988 *Whose Justice? Which Rationality?* University of Notre Dame Press, Notre Dame, IN

Mill J S 1974 [1843] A System of Logic. In: Robson J M (ed.) *The Collected Works of John Stuart Mill*. University of Toronto Press, Toronto, ON, Vols. 7 and 8

Nisbett R, Ross L 1980 *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall, Englewood Cliffs, NJ

Pettit P 1993 *The Common Mind: An Essay on Psychology, Society and Politics*. Oxford University Press, New York

Rawls J 1993 *Political Liberalism*. Columbia University Press, New York

Rorty R 1991 *Objectivity, Relativism and Truth*. Cambridge University Press, Cambridge, UK

Rips L J 1983 Cognitive processes and propositional reasoning. *Psychological Review* **90**: 38–71

Stich S P 1990 *The Fragmentation of Reason*. MIT Press, Cambridge, MA

Wittgenstein L 1958 *Philosophical Investigations*, 3rd edn. Macmillan, London

G. F. Gaus

Public Relations in Media

Of all the many factors that shape the news appearing in the media, the rise of source professionals or public relations experts is one of the most dramatic. Known under a variety of titles, from public relations practitioners, communications managers, and political consultants, to pollsters, spin doctors, and lobbyists, their job is to manage the news on behalf of their employers, ensuring their visibility, boosting their image, or repairing their reputation. These cultural intermediaries also help to feed the growing appetite for stories generated by the twenty-first century's proliferating media and, as such, have a significant influence on both the agenda and content of current journalistic output (for further references see *Political Advertising*; *Political Communication*).

It was the establishment of universal suffrage that gave the biggest fillip to the spread of modern day public relations activity in the public sphere. For, just as the newly enfranchised citizens began to seek information about their country's power bases—over which they could now exercise some control—their opinions began to assume an impact that they did not have before. Public opinion became worth swaying by those who needed popular support. This coincided with increasing literacy that initially promoted the spread of newspapers and subsequently the development of the broadcast media, all of which were hungry for news items.

In early definitions of the role of the professional source, Bernays—one of the profession's founding fathers—commented that the job of those brought in to court this newly influential public opinion was to isolate ideas and develop them into events that could be used as news items by the media (1923, p. 171). Writing in 1922, Lippmann, observed that a new age of opinion management had arrived and the creation of consent among the governed had begun to develop beyond anything seen before in the political organization of human affairs. Public relations thus emerged in the early decades of the twentieth century particularly to link the political actor as a source of news with the media who would produce it, in a relationship that met the needs of both (McNair 1998, p. 146).

The role of public relations in swaying public opinion has since been regarded as crucial to those

seeking to gain or maintain political power. Consequently, as an instrument in manipulating a powerful public opinion consensus, the professional source operating within the political arenas of liberal democratic states has attracted increasing attention.

1. *Source–Media Relations*

News that influences public opinion is heavily affected by the nature of the relationship between source and media. Throughout the 1970s and for much of the 1980s, two contrasting sociopolitical paradigms prevailed. One was the Marxist view, which argued that media content was governed by elite sources representing the dominant classes of society with the result that their ideology permeated the media to the virtual exclusion of alternative views. This dominance was achieved with the compliance of journalists who readily accepted, and often sought, their input because these elite sources were perceived as authoritative and even objective. Furthermore, the dominant class learnt to deliver stories that conformed to the criteria of newsworthiness and the regularity of their news supply could be relied upon by the news media. In contrast, the pluralist view maintained that there was a wealth of groups and interests all competing for influence. The media were thus acting as a marketplace of ideas, none predominating to the exclusion of others.

However, recent research on public relations has delved into the question of access and the degree of openness and closure of the communications system, revealing a more complex picture than either of these early paradigms suggested. To what extent does the dominant class, characterized by the corporate and state elite, control the media agenda through direct access to the detriment of non-institutional and poorly resourced groups? This new research on news sources and promotional culture (see Wernick 1991), concentrates on the public relations activities of corporate and institutional sources in the public sphere and also on the strategies of opposition and resource-poor groups as they seek to gain access to the media.

The political economy perspective, with its emphasis on economic power and media ownership, argues that corporate and institutional/state sources have huge resource advantages that enable them to dictate the current of news by restricting or enabling information flows (Glasgow Media Group 1976, Herman and Chomsky 1988, Gandy 1982, for further reference see *Mass Media, Political Economy of*). Similar to this position is Hall et al.'s (1978) radical structural culturalist approach which concentrates on identifying the legitimacy conferred on state and corporate sources by journalists. Media statements are grounded in objective and authoritative statements from sources accredited by virtue of their institutional power and position. News production inevitably leads to a systematically structured over accessing to the media of those in powerful and privileged institutional positions. The result of structured preference conferred by journalists to the views of these elites is that they become the primary definers of topics and the primary interpreters of all subsequent debate on the issues.

Challenges to this account of source-journalist behavior as being overly deterministic come from the recently termed radical pluralist. Whilst acknowledging the structuralist view that elite sources hold strategic advantages and wield substantial economic and political power over the news agenda, Schlesinger and Tumber point out that 'because the conception of primary definition resolves the question of source power on the basis of structuralist assumptions, it closes off any engagement with the dynamic processes of contestation in a given field of discourse' (1994, p. 21). The concept of primary definition ignores: the conflict between elite sources therefore presenting the media with more than one message; the role of 'leaks' and off-the record briefings by non attributable sources; the non-retention of access over time as new forces and their representatives emerge; the vulnerability and possible discreditation of 'accredited' sources; and finally, journalists' challenges to official sources (1994, pp. 18–20). In this formulation primary definers are therefore not structurally predetermined, but earn this role by providing journalists with regular, reliable, and authoritative story leads.

These considerations have rethought the behavior of sources and led to the introduction of new methodologies in conducting empirical studies. Instead of relying solely on internalist accounts that rely on either journalists' descriptions of their relationships with sources or a reading of media content, recent research has expanded externalist or source centered approaches by examining the media strategies and tactics of news sources (see Ericson et al. 1989, Deacon and Golding 1994, Hallin 1994, Schlesinger and Tumber 1994, Miller 1994, Anderson 1997). All these studies with their various approaches emphasize the way all groups, official and nonofficial, whether state, corporate, voluntary, or pressure need to foster favorable public opinion through the establishment and protection of identity and image. New pluralist accounts offer the view that 'public relations is less capital dependent than other forms of public communication and non-official sources can employ strategies that require little institutional legitimacy' (Davis 2000, p. 51) thereby conferring on them the potential for gaining access to the media that traditional radical accounts did not acknowledge.

2. *The Growth of Public Relations in the Media*

As a business activity, public relations began to gain in importance from the end of World War I particularly in the USA and later in Europe. Corporate identity work grew up in the early 1960s and, as a professional practice, gained wider acceptance in the following decade during which there was an upsurge in

acquisitions, mergers, and divestitures. In commercial climates such as these, both predatory companies and their corporate prey have found that it has paid to heighten their image. And, as they have merged, both nationally and internationally, there has been a need to establish and maintain a single identity by which the corporate whole can be recognized (Meech 1996, pp. 65–6). Privatization programs through which previously state-owned companies seek shareholders from among the general public, also boosts the need for the companies involved to seek regular, favorable coverage in the media. This is generated through public relations professionals, both in-house and from hired agencies, whose specialisms now include media relations, public affairs, issues management, and lobbying. The lead shown by the corporate world has been enthusiastically embraced by both political parties and politicians.

2.1 Proactive Public Relations

The work of these public relations professionals can be described in two broad categories. First, there are proactive public relations through which, in particular, 'pseudo events' are used to create news that will be attractive to journalists (Boorstin 1962). These artificially created 'happenings'—such as press conferences and photo opportunities—are staged purely to present journalists with material that is tailored to their needs. But their lack of authenticity and rationality has been criticized for promoting style over substance, and performance over policy, thereby turning news into a promotional form for politics and business.

Public relations, particularly lobbying on behalf of corporate and other sectional interests, has risen in importance. The size and development of the public relations industry in the United States and the United Kingdom mean that British and American firms dominate the market in most member states of the European Union, and their actions are increasing in scope and intensity. Public relations specialists and lobbyists are also diversifying by targeting not just the national and regional governments in the member states, but also the EU institutions, along with trade associations, corporations, and interest groups (Miller and Schlesinger 1999).

2.2 Reactive Public Relations

The second category can be described as reactive public relations. Sometimes called crisis management, this form of news manipulation springs into action to counter real news events that are potentially damaging. For politicians the rise in scandal news has led to a preoccupation with new strategies and tactics to counter negative images. Within the corporate world, as globalization has developed, multinational companies, face problems arising in one small part of their global operation turning into an international crisis within minutes thanks to ever-faster and more ubiquitous electronic links and the consequent speed with which media agencies can flash news around the world. During the 1980s and 1990s, a spate of corporate public relations offensives was launched to refute criticisms over safety procedures and corporate responsibility in response to a number of well-publicized disasters and accidents. Swift action was also deemed necessary in the wake of takeover and other financial scandals, and to allay public fears about the real or potential dangers of contaminated or faulty products. High profile cases illustrate that an efficient crisis management plan, with an emphasis on skilful news management, can be effective in lessening the damage such catastrophes can wreak on a company. Indeed, those that have managed crises well in public relations terms have even managed to benefit from them. They have achieved this by attracting public sympathy in the immediate aftermath of a disaster, and then by winning people's trust and respect for the way in which it has been subsequently handled. In contrast, a corporate crisis that is badly handled in public relations terms can cause irreparable damage leading to board resignations, a plunge in share value, and even corporate takeover (for further reference see *Scandal: Political*).

The message that even the largest companies ignore at their peril is that no organization or individual is immune to negative media coverage, especially now that effective public relations techniques are being increasingly understood and adopted by campaigning organizations. These relatively small bodies have found that lack of resources need not prevent them, too, from manipulating the news and, as a result, they have scored a number of public relations successes at the expense of government and big business. Since the 1960s, there has been a significant increase in organized lobbying with the emergence of several single-issue pressure groups particularly within the green and consumer movements. These organizations have pioneered new ways to harness media coverage to coincide with campaigns in a way that has compelled politicians and other powerful groups to take them seriously. Thus, through the creation of pseudo events, the targeting of specialist correspondents, and other techniques to meet journalists' need for an endless flow of new material, these pressure groups indicate that non-official and resource poor organizations can also make significant interventions in media discourses despite lacking definitional power and cultural capital (see Bourdieu 1979). The ability of resource poor groups to gain access to the media, though, requires favorable conditions and the requirement to establish credibility and maintain reliability. Occasional entry into the public arena is possible but not secure in the long term (Goldenberg 1975).

3. Impact on Democracy

Public relations has emerged as a key specialism in the conduct and management of national and international political affairs and also corporate image. But its impact on democracy has often been viewed with alarm. For instance, the rise of the spin-doctor on the political scene has led to accusations of distorting news to an undesirable extent. The argument is that if the electorate requires objective information, then the aim of professional sources to manage this information represents a threat to democracy. This concern has been heightened by changes in the working conditions of journalists, which have increased their reliance on the input of sources. As many media organizations have evolved into businesses whose prime concern is the bottom line, cost cutting has been a focus. Thus, at the same time as external organizations are investing growing amounts in public relations activity, media companies are reducing editorial budgets, compelling journalists to produce more for less. With less time and money to produce their own stories, journalists are obliged to rely more heavily on information subsidies supplied to them by sources. Consequently, now that journalists frequently have little time to check facts or to mount any independent investigation, professional sources have assumed more power in placing their stories in the media, complete with their particular spin. And as a result, the journalist's role, both as gatekeeper and as watchdog, is diminished.

Despite a reformulation of source media relations and public relations activity there is general acknowledgment that source access is still far from equal. Although non-institutional and resource poor organizations have become skilled at manipulating the news media, they are still at a disadvantages lacking the cultural capital of corporate and state sources in the ability to limit or enable access to information and, neither possessing the political, legal, and financial power to apply pressure to journalists. Furthermore, despite conflict between elite sources, these still have an automatic legitimacy that other sources have to earn.

Powerful institutions, 'driven by the promotional dynamic to attain new heights of sophistication in trying to manage the news' (Schlesinger and Tumber 1994, p. 275) are, therefore, likely to remain the most common suppliers of material to journalists, with elite sources continuing to invest heavily in activities to sway public opinion in their favor. And, whatever concerns it might provoke, public relations is now seen as a necessary element of both the modern political process and of corporate management.

See also: Advertising Agencies; Advertising: Effects; Mass Media: Introduction and Schools of Thought; Mass Media, Representations in; Media Ethics; Media Imperialism; Media, Uses of; Scandal: Political

Bibliography

Anderson A 1997 *Media, Culture and the Environment*. UCL Press, London

Bernays E 1923 *Crystallizing Public Opinion*. Boni & Liveright, New York

Boorstin D 1962 *The Image or What Happened to the American Dream*. Weidenfeld and Nicholson, London

Bourdieu P 1979 *Distinction*. Routledge, London

Ericson R V, Baranek P M, Chan J B L 1989 *Negotiating Control: A Study of News Sources*. Open University Press, University of Toronto Press, Toronto

Davis A 2000 Public relations, new production and changing patterns of source access in the British national media. *Media Culture and Society* **22**: 39–60

Deacon D, Golding P 1994 *Taxation and Representation*. John Libby, London

Gandy O H Jr 1982 *Beyond Agenda Setting: Information Subsidies and Public Policy*. Ablex, Norwood, NJ

Glasgow University Media Group 1976 *Bad News*. Routledge and Kegan Paul, London

Goldenberg E 1975 *Making the Papers: The Access of Resource-poor Groups to the Metropolitan Press*. Lexington Press, Lexington, MA

Hall S, Crichter C, Jefferson T, Clarke J, Roberts B 1978 *Policing the Crisis: Mugging, the State, and Law and Order*. Macmillan, London

Hallin D C 1994 *We Keep America on Top of the World: Television Journalism and the Public Sphere*. Routledge, London

Herman E S, Chomsky N 1988 *Manufacturing Consent: The Political Economy of the Mass Media*. Pantheon, New York

Lippmann W 1922 *Public Opinion*. Harcourt, Brace and Co., New York

McNair B 1998 *The Sociology of Journalism*. Arnold, London

Meech P 1996 Corporate identity and corporate image. In: L'Etang J, Pieczka M (eds.) *Critical Perspectives in Public Relations*. International Thomson Business Press, London

Miller D 1994 *Don't Mention the War: Northern Ireland, Propaganda and the Media*. Pluto Press, London

Miller D, Schlesinger P 1999 The changing shape of public relations in the European Union. In: Heath R (ed.) *The Handbook of Public Relations*. Sage, London

Schlesinger P, Tumber H 1994 *Reporting Crime: The Media Politics of Criminal Justice*. Clarendon Press, Oxford, UK

Wernick A 1991 *Promotional Culture: Advertising, Ideology and Symbolic Impression*. Sage, London

H. Tumber

Public Sector Organizations

'Public sector organizations' are entities that have been formed to manage the policy and operating requirements that enable a government to achieve its goals of public governance. The term 'public governance' has been defined as the management of a nation through the use of political power (Rainey 1991), or more simply put, answers the question: 'How should government govern, and what should it do' (Peters

1996, p. 19)? Such a broad definition of public governance permits a wide range of activities to be undertaken by public sector organizations as the administrative arm of government.

The role and structure of public sector organizations depend on the choices made by governments over time and the traditions widely recognized within the system of government operating in a country. Public sector organizations can be structured as central agencies of government, service providers, government business enterprises (GBEs), regulatory bodies, watchdog agencies, and as quasi-government bodies. While many of these institutions have remained comparatively unchanged, the emergence of Next Steps agencies in the United Kingdom represents a significant example of a new approach to the structure of public sector organizations. New Steps agencies operate under a form of contract that sets out the scope of their responsibilities and accountability to both government and clients, while related policy issues remain under the control of government.

In the 1930s, Mary Parker Follett suggested that there were significant 'similarities between the business of government and the government of business' (Corkery 1989, p. 193) However, it took until the last two decades of the twentieth century for reform of public sector organizations to explore this notion through the application of private sector practices to the public sector. This reform process coincided with a shift away from a politico-economic ideology involving the pursuit of full employment, a focus on demand management within an economy, and intense government intervention, characterized by the principles expounded by John Maynard Keynes. In its place emerged the so-called, neoclassical model of economic management. Pioneered by Milton Friedman, this economic ideology centers upon the capacity of an economy to meet potential demand (supply-side economics), and the regulation of economic activity through the acceptance of market forces in place of government intervention. The neoclassical model is closely associated with the principles of smaller government pursued at first instance, by national leaders Thatcher, Reagan and Mulroney of the United Kingdom, the USA, and Canada respectively.

By the early 1990s, the concept of reform of public sector organizations had gained a significant international following. Often referred to as the New Public Management (NPM), the drive for reform was accompanied by emergence of the range of public sector organizations, an example of which is outlined above. The NPM was accompanied by a wave of privatizations in a number of jurisdictions, and by a rhetoric that espoused a new approach to the management of public sector organizations in terms of 'doing more with less,' 'letting the managers manage,' and 'pursuing greater efficiency and effectiveness.' This rhetoric, often described as 'managerialism,' supported the application of private sector principles to the management of public sector organizations and translated into a variety of radical reforms in public organizations designed to support the philosophy of smaller government. These reforms included: reducing the number of public employees through 'downsizing,' outsourcing a variety of activities previously undertaken by public sector organizations, introducing budget reductions to enforce financial restraint within public sector organizations, and selling government assets and entities to reduce the budgetary problems experienced by many governments.

Over this period, research into public sector organizations has examined the consequences of these reforms. The research has divided along lines that either support market-oriented reforms or seek to justify a return to more traditional models of public sector organizations. However, some recent research highlights the importance of strong citizen support for the public sector and attempts to put to rest the view that the 'public-sector service is inherently inferior to private-sector service' (Kernaghan 2000, p. 7). Organizational theory research in this area has also explored the consequences of adopting private sector organizational practices in the public sector. It is likely that the divergences of opinion about the relative merits of organizational performance in each sector will remain a source of intense debate and a matter of personal opinion.

1. *The Rise of the Public Sector Organization*

A theory of a public sector organization might be first found in the development of the democratic city-state described by Plato in *The Republic*. His concept arose from the notion that people are 'not independent of one another, but need the aid and co-operation of others in the production of the necessities of life' (Copleston 1962, p. 251). Livy (15 BC) also described a public sector that was involved in sophisticated risk management negotiations with the private sector, whose help was needed to save part of the Roman army stationed in Spain. However, evidence from Chinese scholars suggests that public sector organizations existed in that country in the Xia Dynasty, many centuries earlier than in Europe.

The origins of public sector organizations in the English-speaking world can be traced to the arrival of William the Conqueror on British shores in 1066. His administration led to the development of public institutions that supported the power and financial strength of the monarch, through the development of the Common Law and the creation of the Doomsday Book, which recorded the ownership of land throughout the kingdom which then encompassed England. At this time, the accountability of these institutions was to the monarch, rather than the polity, demonstrating the nature of the public sector that existed during this period of feudal society.

The emerging role and structure of public sector organizations can be seen in their support of empires in the late Middle Ages and the transfer of traditional public sector roles and structures from the dominant power to a subordinate country. This transfer led to some commonality of ideals in public sector organizations that can be traced back to the influence of, or reaction against, periods of colonial rule of a nation. Bureaucratic traditions also became well-established features of public sector organizations, a factor that has been a source of criticism of public sector organizations and the subject of considerable theoretical and practical evaluation.

In 1776, Adam Smith had suggested the traditional role of government should be limited to the protection of a nation from internal and external violence, the administration of a fair system of justice, and the development and maintenance of major public works. These principles found favor in the early nineteenth century and, more recently have attracted approval from those supporting contemporary neoclassical views. The Northcote–Trevelyan report in 1853 eventually led to a radical and influential restructure of the British civil service and provided the foundation of modes of public sector organization in the twentieth century.

It could be argued that in many jurisdictions public sector organizations enjoyed the peak of their influence in the economic boom years of the 1960s and 1970s. However, by the 1980s, the pressure for reform had begun as the changing economic ideology of governments and emerging budgetary pressures forced governments to re-examine the role of their public sector institutions.

2. *The Nature of a Public Sector Organization*

It is the complex relationship between public sector organizations and government, which distinguishes this type of organization from those in the private and nonprofit sectors, regardless of the recent attempts to merge public with private sector management practices. Public sector organizations can be distinguished from other sectors by their accountability to the principles of public governance adopted by a government. It can be claimed that this accountability gives rise to an overriding management (including ministerial) culture that punishes mistakes and guarantees that few risks will be taken. In a democratic state, there is an expectation that the public sector is also accountable to the government, which is itself answerable to the electorate. In a more authoritarian state, the public sector can be expected to be accountable to the government alone, although in this environment, public opinion may remain an important countervailing force.

The issue of accountability to government and in turn, civil society, highlights another distinctive feature of public sector organizations: the decision-making complexity arising from the need to cater for the diversity of demands faced by public sector organizations. The challenges to the decision-making process emerge, first, from the comprehensive legal requirements that bind the form and operation of public sector organizations. In addition, the demands of civil society, the intensity of competing interests within an economy, and the unpopular role of a number of public entities, all contribute to the decision-making complexity experienced by public sector organizations. For example, as societies have become more diverse and interest groups more influential, their competing demands make the role of both government and public sector organizations, more difficult (Olson 1982). In the face of these tensions, there are suggestions that the emerging challenges of globalization require 'the emergence of new forms of government' (Metcalfe 2000, p. 122) able to conceptualize and respond to the increasingly varied and competing demands of public sector management.

Public scrutiny of the policies and operations of public sector organizations is also widespread. Formal scrutiny on behalf of government may occur through a systematic process of audit as well as periodic review by politicians and the judiciary. However, informal review by the media represents a major tension for public sector organizations whose performance may be the subject of intense media review at any time, for political purposes and 'in the public interest.' There are suggestions that this pressure makes the performance and decision processes of public sector organizations slower and more conservative as the managers of these organizations work to avoid public criticism of their organization and the government.

The decision-making processes used in the public sector may be seen as slower and more dilatory than other sectors because of the bureaucratic and accountability structures of government organizations. This perception was supported by Drucker who, in the late 1960s suggested that 'any attempt to combine governing with 'doing' on a large scale paralyses the decision-making capacity' (1969, p. 233). While this criticism may arise from a frustration with the speed with which public sector organizations move towards many decisions, it occurs because of a failure to recognize the impact of bureaucratic and political demands on the decision processes within government and public sector organizations.

The comparative absence of competition between public sector organizations and between public and private sector organizations has also been a distinguishing feature of public sector organizations, although there are exceptions to this rule, especially in the case of some GBEs that operate in a competitive business environment. As Osborne has commented: 'Businesses typically exist in fiercely competitive markets, most public organizations don't' (1994, p. 137). Business performance is also built upon comparatively

concrete targets such as overall profitability, return on assets, return on owners' funds, inventory turnover, and the maintenance of cash flow.

Non-traditional performance measures such as assessing customer loyalty, the level of customer satisfaction, and relationship management have also been developed by some organizations seeking to enhance their market standing. However, there are a number of other performance criteria, such as timeliness, ability to operate within budget expectations, and perceived quality of implementing government policy, which might be substituted for profit in public sector organizations. A customer service focus has been a feature of the reform of public sector organizations in the 1980s and 1990s, and remains high on the agenda of the US and UK governments.

The operating environment of a public sector organization has often been regarded as being in some way inferior to private sector organizations. This view, which has long been promoted in literature (Parkinson 1958) and the media, may have influenced public opinion against members of public sector organizations, a view that several governments are actively attempting to redress. The changing nature of the service requirements of many public sector institutions has also created pressure to move away from the generalist model of management. There is an increasing expectation that managers of these entities, in addition to their existing management skills, may find they now need some of the skills possessed by many private sector managers. These capabilities include a working knowledge of sophisticated aspects of historical cost accounting, information technology, national and international finance, contract negotiation, sales and marketing, human resources management, and assessment of customer satisfaction.

3. *Public Sector Organizations and the New Public Management*

The causes of the drive to reform public sector organizations are multiple. They can be traced to the political pressures faced by governments in the early 1980s, which began to find that budgetary constraints emerged, full employment had become an apparently unattainable goal, inflation was more difficult to manage than expected, and the ideology of economic management had begun to change. The model of demand management enunciated by Keynes came to be replaced by the more rational, market forces model, or supply-side economics, an ideology that also found favor in developing new policies for defining the relationship between public sector organizations and their respective governments and polities. However, reform can also be attributed to the need for politicians to be seen to be doing something constructive to please both the polity generally as well as the financially powerful private sector.

As noted earlier, the most contentious reform of public sector organizations focuses on the reengineering of the public sector within the philosophies espoused by those promoting the concept of the NPM. At its extreme, the emergence of the NPM has seen a shift in the focus of government from seeking policy advice and implementation, to a drive for a more 'efficient and effective' public sector based on performance principles more commonly associated with the business sector. The accompanying rhetoric includes the goals of 'economic competitiveness, doing more with less, building faith in government and strengthening community and civil society' (Gore 1999, p. 1). Across the Atlantic, the goals of the British Government have been expressed as targets in five key areas of public service, namely:

forward looking policy making;
public services focused on needs of users;
quality public services;
the use of technology to support a modernized public service; and
a move to value, rather than denigrate, public service (Her Majesty's Government 1999).

From an organizational point of view, these emerging trends have resulted in a number of significant changes to many public sector organizations. Because of the range of ways in which public sectors are structured, these trends vary in their importance. In many public sector organizations, attaching private sector salary levels to senior executives who are appointed under contract for a specific term has tended to increase the politicization of public sector organizations and has changed their traditional relationship with the government.

The contracting out of activities previously undertaken by public sector organizations has had the potential to reduce their knowledge of particular markets and affect their ability to enter contractual arrangements. One of the main areas of risk is ensuring that the legal and operational sophistication possessed by private sector organizations, is matched by those acting on behalf of the public sector. Contracts are required which take account of possible contingencies arising during the life of a contract. Because these contracts exist between the public sector and private sector, in some jurisdictions there are increasing numbers of cases where the transparency usually associated with public contracts is increasingly hidden behind the barrier of 'commercial in-confidence,' perhaps to the detriment of the polity through the diminution of public review.

The outcomes of the reform process have been mixed and are subject to widely varying interpretations. The reform process is not an event; it represents a continuum of change in public sectors in most countries, where the pressures of globalization, technological change, reduction in government debt, and emergence of aspects of the neoclassical approach have taken hold. As a result, it is not possible to

evaluate the precise cost-benefit of public sector reforms presently, nor is such an evaluation likely to be very reliable in the future. Some public sector organizations have reduced their regulatory control within the economy, while others have found it necessary to consider a re-regulation process. The contracting-out of activities within some public institutions has provided both benefits and costs to government and civil society, an experience that is likely to be mirrored within private sector organizations that have undergone similar reengineering processes. Major experiments in social change such as that attempted in New Zealand have led to complex outcomes for both public and private sector organizations, the merits of which remain widely debated.

4. *Future Directions*

As the twenty-first century begins, it is clear that public sector organizations will continue to experience the tensions that arise between the demands of government, business, and civil society. The basic role of public sector organizations may not change dramatically, although governments may find themselves under intense pressure to meet the demands of rapidly aging societies, to reduce the inequities apparent from contemporary versions of globalization and to manage the consequences of emerging e-business trends.

Current trends to downsize public sector organizations as part of a dedication to the goals of smaller government will force these entities to work towards further financial and operating efficiencies, to achieve higher service quality standards, and to emulate private sector organization structures. However, it is unlikely that these organizations will gain any relief from the responsibilities of public governance.

The tensions between organizational theory and public sector practice can be expected to continue to be a source of academic review and public scrutiny as specific events cause governments and public sector organizations to reexamine their role. For example, re-regulation of certain markets, a requirement for public investment in previously privatized businesses, and a range of potential responses to the combined effects of e-business and globalization are examples of major challenges for public sector organizations. These demands are likely to increase as the structure of these organizations continues to be affected by the combined effects of downsizing, outsourcing, budgetary restrictions, and the decentralization of organization structures inherent in the moves towards electronic government.

See also: Bureaucratization and Bureaucracy, History of; Organizations and the Law; Organizations, Sociology of; People in Organizations; Public Goods: International; Public Interest; Public Management, New; Quangos in Politics;

Bibliography

Barney J, Hesterley W 1996 Organizational economics: Understanding the relationship between organizations and economic analysis. In: Clegg S, Hardy C, Nord W (eds.) *Handbook of Organization Studies*. Sage, London

Copleston F 1962 *A History of Philosophy: Greece and Rome*. Image Books, New York, Vol. 1, p. 1

Corkery J 1989 Challenges of the next twenty years: Future directions for training. *Australian Journal of Public Administration* **48**: 185–95

Drucker P 1969 *The Age of Discontinuity*. Heinemann, London

Friedman M 1962 *Capitalism and Freedom*. The University of Chicago Press, Chicago

Gore A 1999 Transforming governments in the 21st century. Global forum on reinventing government. (http://www.21stcentury.gov/conf/gore1.htm) Washington, DC

Her Majesty's Government 1999 *Modernising Government*. The White Paper. (http://www.official-documents.co.uk/document/cm43/4310/4310.htm) Her Majesty's Stationery Office, London

Kernaghan K 2000 *Rediscovering Public Service: Recognizing the Value of an Essential Institution*. Institute of Public Administration of Canada, Toronto, Canada

Keynes J M 1964 *The General Theory of Employment Interest and Money*. Macmillan, London

Mauk T 1999 The death of bureaucracy. *Public Management* (*US*) **81**(7): 4–8

Metcalfe L 2000 Linking levels of government: European integration and globalization. *International Review of Administrative Sciences* **66**(1): 119–42

Nutt P 1999 Public-private differences and the assessment of alternatives for decision making. *Public Administration Research and Theory* **9**(2): 305–50

Olson M 1982 *The Rise and Decline of Nations*. Yale University Press, New Haven, CT

Osborne D 1994 Reinventing the business of government: An interview with change catalyst David Osborne. *Harvard Business Review* May–June: 132–43

Parkinson C N 1958 *Parkinson's Law or the Pursuit of Progress*. Murray, London

Peters B G 1996 *The Future of Governing: Four Emerging Models*. University Press of Kansas, Lawrence, KS

Rainey H 1991 *Understanding and Managing Public Organizations*. Jossey-Bass, San Francisco, CA

World Bank 1992 *Governance and Development*. World Bank, Washington, DC

G. Callender

Public Sphere and the Media

To analyze the media from the perspective of the public sphere is to focus upon the media's relationship with democracy. The concept of the public sphere has been used to describe both the historical development of modern, post-Enlightenment forms of representative democracy and the role of the media and public

opinion within them, and has also served as a normative concept with which to pass judgement on the current performance of the media from the perspective of the creation and maintenance of democracy.

In recent years, the public sphere perspective has become dominant in thinking about the media and their political role, by providing, in a return to the central questions of enlightenment political thought, a bridge between Marxist theories of ideology and the liberal 'free press' tradition. It provides a critique of Marxist theories of ideology by revalidating the specificity of the political, by giving due weight to the emancipatory potential of liberal bourgeois concepts of free assembly and debate, and by shifting attention from worker to citizen. On the other hand, it provides a critique of an abstract, property-based 'free press' model by stressing both the systemic social barriers to entry and the manipulative distortions to which its discourse is subject.

This return has focused on concepts of the public sphere, civil society, citizenship, and identity. It has in particular pitted those, such as Habermas, who see the emancipatory project of the enlightenment as unfinished business and wish, in a movement of immanent critique, to hold liberalism to its emancipatory ideals, against a range of communitarians and postmoderns. The latter, in the name of various versions of identity politics, see liberal, rights-based political theories as inherently repressive and their emancipatory and universalizing tendencies as bogus (Gray 1995).

The public sphere approach has provided, against the background of the turn away from Marxism associated with the collapse of 'actual existing socialism,' an alternative to theories of dominant ideology or hegemony as an explanation for the coincidence, within what used to be called 'bourgeois democracies,' of growing social inequality and political apathy on the one hand with the relative stability and increasingly consensual, nonideological nature of representative party politics on the other. Its emphasis on discursive practices and on communicative action as central to democratic practice and legitimization fitted well with the wider 'linguistic turn' in the human and social sciences. At the same time, its use of the spatial metaphor of sphere, and its stress on the necessary institutional foundations for the realization of those citizen rights to free expression and debate central to democratic theory, addressed the questions arising from both the perceived empirical reality in the mature democracies of an increasingly media-saturated, image-driven, public relations-oriented politics and the task of constructing democracy in both state and civil society in newly democratized countries with little or no historical traditions of democracy to call upon.

The concept of the public sphere was introduced into media analysis by Jurgen Habermas (1989) in *The Structural Transformation of the Public Sphere*. In Habermas's original formulation, the term was designed to describe a historical process whereby in the shift from premodern feudal, absolutist rule to forms of modern representative democracy public opinion, based upon publicly available information and debate, developed as both a check on, and as a source for the legitimacy of, government.

The public sphere was both a set of institutional spaces—newspapers, lecture halls, coffee houses, etc.—and a set of discursive rules. Drawing upon a Kantian heritage that links freedom to a personal autonomy grounded in the exercise of public reason, Habermas and his followers have stressed the role of the public sphere as a site within which the formation of public opinion, and the political will stemming from and legitimized by such opinion, is subject to the disciplines of a discourse, or communicative ethics, by which all views are subjected to the critical reasoning of others. At the same time, a democratically legitimate public sphere requires that access to it is open to both all citizens and all views equally, provided only that all participants are governed by the search for general agreement. This general model is then applied as a normative test against which the performance of contemporary media, in terms of political effects and democratic potential, can be judged in terms of either the critical rationality of their discourse or in terms of the range of views or speakers granted access.

The public sphere is added as a fourth sphere to Hegel's tripartite division of society into family, civil society, and state. The public sphere is distinguished from the family, the sphere of intimate privacy, although it is the creation of this sphere, particularly under the influence of the literary public sphere that creates both the norms of noncoercive, literate, interpersonal discussion and the associated subjectivities that form the foundation of the public sphere. The public sphere is distinguished from civil society, and especially the market, where, following Hegel, citizens pursue their private interests, since although the private economy provided the resources for the public sphere, and citizens participated in the public sphere as private persons (i.e., access was not governed by socially conferred status or the possession of public office), the purpose of debate within the public sphere was common agreement on matters of general interest. On the other hand, it is distinguished from the state since it was there that citizens held the state to account and guided its actions.

Habermas then argued, following in part the general Frankfurt School argument of the Dialectic of Enlightenment, that there was a pervasive crisis of democratic legitimacy in the modern world because public spheres had been squeezed almost out of existence by both market and state. On the one hand, the market had increasingly come to dominate the institutions and practices of public communication through the commodification of information and opinion for sale on a mass consumer market and through the associated manipulative use of advertising. On the other hand, the state, ironically in

response to political pressures from below, had become increasingly both economically interventionist and manipulative of public opinion. The result was a public opinion and related politics increasingly dominated by advertising and public relations, what Habermas termed the refeudalization of the public sphere—a shift back from a politics of rational, participatory debate to a politics of display.

The public sphere approach has been criticized on three main grounds. First that its procedural rules are too rationalist—that the persuasive use of rhetoric can never, and indeed should never, be excluded from political communication—and that to impose such discursive norms in practice excludes from the public sphere as illegitimate not just a range of culturally specific discursive forms but those who do not possess the cultural capital required to mobilize those discursive forms. Here we find the criticism that the model of procedural rationality being deployed is in fact a model of a certain intellectual practice and thus excludes, in a movement of symbolic violence, those who are not members of the social group who are the carriers of that practice, whether conceived as intellectuals or as the white male bourgeoisie, and that it leads to a privileging of certain genres of media. These rationalist assumptions may take news and overt political coverage to be 'serious' while taking entertainment-oriented programs to be a sign of 'dumbing down' or 'refeudalization.' This neglects the role that entertainment plays in the formation of publics and as a site for the development of an understanding of issues of public importance.

Second, the public sphere model of procedural rationality is criticized for drawing the distinction between public and private in such a way as to exclude both key groups of citizens, e.g., women, or key matters of potential public political concern, e.g., the regulation of domestic, intrafamilial, or sexual relations. It is in this sense that the famous slogan 'the personal is political' mounts a challenge to the way in which public sphere theory has drawn the distinction between the private and public spheres. And this in turn will affect the ways in which we think about the ethical issues relating to invasions of 'privacy' by the media and to what is or is not an appropriate matter for public discussion in the media.

Third, it is criticized on the grounds that it has valued general agreement around a set of universal values or norms, however rationally discursively arrived at, which, its critics argue, derive in turn from a liberal model of proceduralism, abstract individual rights, and ethical neutrality which prioritizes the just over the good and thus denies difference and thus the inevitability of perpetual normative conflict within modern societies. This in its turn, so critics argue, leads to an overcentralized model of the public sphere incompatible with the politics of identity in multicultural societies, whereas what is needed is a decentralized model of multiple public spheres expressive of each distinct collective identity, culture or form of life.

This debate concerning the proper relationship between social communication and politics can be traced back to the Enlightenment and the basic paradox that modernity posed; the paradox of what Kant called the 'unsocial sociability' of human beings. We might think of this as the inherent tension between liberty and fraternity. In essence, modern democracy is about how we handle the relationship between individual freedom and moral agency on the one hand and the necessary and unavoidable social norms and structures within which alone such freedom can be exercised on the other. As Rousseau put it in *The Social Contract*, 'the problem is to find a form of association … in which each, while uniting himself with all, may still obey himself alone, and remain as free as before.'

The political problem to which the public sphere tradition of analysis addresses itself is then two-sided and much of the problems with current uses of the term public sphere, and criticisms of it, derive from a confusion about which of the two sides of the dilemma are being addressed.

Here we are brought back to the Hegelian concern for Sittlichkeit and to critiques of Habermas's public sphere approach from a communitarian and neo-Aristotelian perspective. Here the debate over the public sphere has to be seen as part of the wider debate between advocates and defenders of the theory and practice of rights-based liberalism and the various proponents of communitarianism and identity politics (Benhabib 1992, Guttman 1994).

The communitarian critique leads to the opposite solution to that of social contract theory and the liberalism based upon it. Now political values are social before they are individual. Politics is embedded in, and ideally expresses, a set of pre-existing social values, or a way of life, and the role and legitimacy of the state, or the public realm, is then to foster and uphold those communal values and defend that way of life. The citizens find their identity in, and give their loyalty to, not a set of abstract rights, but to a way of life, an ethos, that embodies a set of moral values. From this perspective liberalism, and the particular forms of democracy that it supports, is but one way of life among many possible alternatives. From this perspective, the role of the media is then to foster and defend this shared way of life, sometimes described, in the nationalist model of communitarianism, as national identity or culture, rather than to serve as a value neutral space for the critical rational debate over values.

The communitarian position on the public/private divide, and the approaches to both politics and the media that stem from it, is in a sense contradictory. The core of identity politics is the call for the recognition in the public realm of values hitherto deemed private, both in the sense of being excluded

from the public gaze and from public debate, but also in the sense of stemming from private group interests and identities rather than from a generally shared interest and identity, and thus for the acquisition of rights that recognition of these values as public entails, while at the same time drawing for its evaluative arguments upon a range of sources which must exclude the very concept of the public, and its liberal valuation of rights and the recognition and equal treatment of the diversity of private interests, that they are demanding.

Following Benhabib, we can distinguish between the integrationist and participatory versions of communitarianism. The integrationist strand seeks Sittlichkeit in an attempted return to the moral certainties and social unities of traditional societies. Here everything is in a sense public; the values that motivate people and give them their social anchorage are derived from and shape a whole shared way of life. In a theocracy there is no room for the individual or the private. Indeed, this strand in communitarianism criticizes the liberal tradition precisely for creating, and philosophically justifying as the highest good, social arrangements that separate political or public spheres of action and value on the one hand from private spheres of action and value on the other, to the impoverishment, in their view, of both (Macintyre 1998).

On the other hand, the participationist strand wishes to build its way out of the alienation and formalism of liberalism by both accepting the conditions of modernity which are liberalism's starting point—namely posttraditional societies and reflexive individuals—while at the same time arguing for a refounding of political communities on the universalization of the discourse ethic and its practices. The more postmodern end of communitarianism, which links to so-called new social movements and the politics of identity, makes this attempt by, in effect, advocating the fragmentation of societies into small-scale or specialized communities of interest or identity, each with its own public sphere. They tend to see more universalistic and unified definitions of the public sphere as repressive of difference and thus antidemocratic. What is at issue here is both whether within a polity we should be looking at one or multiple public spheres and how we think of the relation between public and private. An integrationist communitarianism and the politics of recognition that stems from it demand a unified public sphere and place continuous pressure on the existence of a meaningful private sphere. Its aim is a one-to-one fit between a unitary set of values, a single public sphere (if this term any longer has meaning here) and a single polity. If diverse communities exist within a single territory or polity, the aim becomes one of political and cultural fragmentation, not coexistence and the toleration of diversity. It is clearly incompatible with the exercise of public reason in the Kantian sense and with media practices and institutional forms in harmony with such an ideal.

The case of participatory communitarianism and the identity politics that stems from it is more complex. Here the problem arises from the ambivalence of the liberal value of tolerance vis-à-vis the public/private divide. On the one hand toleration can be taken to mean the acceptance by public authorities of a range of practices and beliefs by defining them as private and outside the realm of public regulation, e.g., religious observance. Such toleration rests, as the communitarian critics rightly point out, on a prior judgement that the practice or belief in question is not in any sense a threat to the public weal or interest. This form of toleration can be rejected by certain advocates of the politics of recognition because, it is argued, it implicitly downgrades, and thus in some sense fails to recognize fully, the importance or centrality of the practice or beliefs in question. On the other hand, we can also understand toleration as giving public recognition to and bringing into the public realm people, practices, or beliefs to which by so doing we signal that we give equal value to those already so recognized and with which we are prepared to live and argue within a shared culture and polity. Thus we need to distinguish between two kinds of demands for recognition and for multiple public spheres. One set of demands seeks to extend the private, so that their group identities remain unthreatened by the risk of corrosion that participation in the critical discourse and compromise of the public sphere carries.

The second set demands a widening of the definition of what is public to let them into what remains a common arena for critical public debate and decision making with the acceptance of the duties and risks that such entry carries with it. In the former case, what is being demanded in the name of difference is the dissolution of any shared culture, polity, and set of rights and associated obligations. Here the question is how much of such private fragmentation can any society or polity sustain while remaining viable, and whether in fact we will only regard such demands as democratically legitimate where the group identity claimed is itself subject to the disciplines of the discourse ethic. In the latter case, the liberal polity is being asked, rightly, to live up to its ideals and to accept that those values that it considers central to its way of life must always be held provisionally and be subject to the review of critical discourse. Here we need to distinguish between a truly private realm, the irreducible site of individual autonomy, and the multiple public spheres within which we necessarily live and have our being, defined by Benhabib as any occasion or situation where people share views on matters of common interest, multiple public spheres that all contribute to the formation of our identities and which may be more or less democratic in the sense of meeting the requirements of the discourse ethic. The media, ranging from the small-scale, special-interest

media of magazines, newsletters, Internet bulletin boards up to the space for public debate and exposure of views offered by the mainstream mass media, are integral to these multiple public spheres and should be judged in each case on the basis of the identities and practices that they foster. However, these multiple public spheres then need to be distinguished from the political public sphere where discussion is necessarily aimed at the common agreement necessary for concerted action within a unified polity. Research and debate within and around the public sphere perspective now focuses on three issues. The first is the institutional question of the relationship between the political public sphere and the nation-state. Both the actual institutional development of the modern mass media, the public sphere, and representative democracy, and thinking about them, have taken place within a nation-state framework. It is argued that current globalizing trends, both within society at large and within the media, are undermining this nation-state foundation of both democratic politics and its associated political public sphere. The question then is whether this is a process of cultural and discursive emancipation and the beginnings of the construction of a global public sphere or, on the contrary, is a further tightening of the screw of refeudalization which places the key centers of power outside the structures of democratic accountability and rational critical debate. On the one hand, the trends towards globalization within the media can be exaggerated. For reasons of audience taste and interest, most media output, and the organizations that provide it, remain obstinately linked to local, often national, markets, cultures, languages, and political structures. There are few examples of successful transnational newspapers or broadcasting news services. Even in the field of entertainment their spread is limited. This is itself, as the example of the European Union shows, a major barrier to the construction of a transnational public sphere. On the other hand, it can be argued that current developments on the World Wide Web, particularly its use for mobilization, debate, and advocacy by NGOs, are a welcome start in the construction of a global public sphere. The question here is how representative and how public these new Net-based interest groups and associated debates are. What certainly is the case is that the development of global media has begun to create a situation where nation-states, both within their own borders and in their inter-State relations, are having increasingly to take account of global public opinion (e.g., the cases of Tienanmen Square or the Balkans).

Second is the discursive question of how we can both analyze and normatively judge from the perspective of democracy the wide range of rhetorical forms used and identities appealed to within the range of modern media. Here the debate around the public sphere reprises an older debate about mass culture. The question at issue is whether current developments in media forms, genres, and their associated audiences/publics is a liberatory process of demassification and cultural pluralism on a global scale or, on the contrary, whether it is a process of 'dumbing down' which, in its appeals to irrationality and social fragmentation, undermines the rational/universalistic norms upon which citizenship, and thus democracy, rest. In the press, attention is focused on tabloidization, i.e., the stress on the scandalous and the sensational and the coverage of politics according to the norms of show business; in broadcasting on the shift in the balance of the programming mix in favor of entertainment at the expense of news, current affairs, and political coverage and debate, a shift associated, at least in Europe, with the retreat of public service broadcasting in the face of increasingly intense commercial competition for audiences and revenues. Analysis and debate have focused in particular on the spread of audience participation shows of the Jerry Springer type, of 'docu-soaps' and of a general sound-bite coverage of politics and public issues. In this debate there are no easy answers. First, empirically the trends are by no means as clear cut as many critics claim. It is as easy to argue for a rise in standards. Second, it is all too easy to fall into familiar elitist or populist stances, neither of which take due account of the complexity, both actual and historical, of the construction of the identity of citizens or of the rhetorical problems of communicating with the generality of such citizens. There is also often a danger of stretching beyond breaking point the proper and necessary boundaries of politics and the public sphere.

Third, there is the question of the role of the mediators. Public sphere theory finds it difficult to break free from a participatory model of direct democracy. Representation always introduces a potentially distorting element into the communication process between the private citizen and political whole. The public sphere is conceived as a structure for the aggregation of the opinions of individual private citizens through a public debate to which all have equal access. The ideal model is the Athenian agora. The scale, social complexity, and specialization of the modern world make such a model unrealistic. It has clearly never fitted the actual forms and practices of the mass media. It therefore leads to unrealistic demands being made upon them and to an excessively dismissive critique when these demands are not met. It has in particular led to exaggerated claims being made for the democratic potential of the Internet on the grounds precisely that cyberspace is a new agora without mediators or mediation. The challenge, therefore, is to think of a legitimate role for mediators within the public sphere—for journalists, TV producers, bulletin board moderators, etc.—which is neither one of pure transparent 'objective' mediation nor one of the mere pursuit of instrumental pro-

fessional or economic self-interest. This in its turn links to the wider question of the social and political role of intellectuals and experts within modern democratic polities.

See also: Critical Theory: Contemporary; Critical Theory: Frankfurt School; Legal Issues: Public Opinion; Mass Communication: Normative Frameworks; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Mass Society: History of the Concept; Public Opinion: Microsociological Aspects; Public Opinion: Political Aspects

Bibliography

Benhabib S 1992 *Situating the Self: Gender, Community and Post-modernism in Contemporary Ethics*. Polity, Cambridge, UK

Calhoun C (ed.) 1992 *Habermas and the Public Sphere*. MIT Press, Cambridge, MA

Gray J 1995 *Enlightenment's Wake: Politics and Culture at the Close of the Modern Age*. Routledge, London

Guttman A (ed.) 1994 *Multiculturalism*. Princeton University Press, Princeton, NJ

Habermas J 1989 *The Structural Transformation of the Public Sphere*. Polity, Cambridge, UK

Macintyre A 1998 *Whose Justice? Whose Rationality?* Duckworth, London

Peters J 1993 Distrust of representation: Habermas on the public sphere. *Media, Culture and Society* **15** No. 4: 541–72

Schudsen M 1998 *The Good Citizen: a History of US Civic Life*. New York

Splichal S 1999 *Public Opinion*. Rowman and Littlefield, Lantion, MD

N. Garnham

Public Sphere: Eighteenth-century History

Any reflection trying to understand how the notion of public sphere was constructed in the eighteenth century is necessarily referred to the classical book published by Jürgen Habermas in 1962 (Habermas 1962). According to Habermas, the 'political public sphere' was construed in the eighteenth century (in some places sooner, in others later) as a space for discussion and exchange removed from the control of the state. The new 'public sphere in the political realm,' founded on the exercise of public criticism was thus opposed to the 'sphere of public authority' identified with the exercise of state power. For Habermas, this new intellectual and political 'space' was distinct from the court, which belonged within the domain of state power, and from the common opinions of the people who had no access to enlightened critical debates. This is why it is possible for him to qualify it as 'bourgeois.'

1. Definitions of the Public

There were different ways of conceiving the construction of the new political public sphere of the eighteenth century. Was it issued from the literary controversies of the seventeenth century and rendered possible by the diffusion of new forms of intellectual sociability? Or was it the extension of the modalities of print communication proper to the Republic of Letters to a greater number of individuals which made possible the transformation of dispersed private persons into a unified public community?

1.1 Habermas's Political Public Sphere

For Habermas such a 'public' supposed, on the one hand, that the various participants were not acting as agents or subjects of the ruler and, on the other hand, that they were considered as equal. The political public sphere thus ignored both the obedience required by the exercise of state power and the distinctions of 'orders' and 'estates' imposed by the traditional social hierarchies.

No limit could subject the critical exercise of reason, and no domain was to be forbidden to it. Overcoming the division instituted by the cartesian methodical doubt between opinions which could legitimately be submitted to examination and obligatory credences and obediences, the members of the new public sphere considered that no domain of thought, including the political and religious ones, was to be removed from critical and reasonable judgement.

This judgment was exercised within the different forms of sociability which had been previously erected as a tribunal of aesthetic criticism: the salons, the cafés, the clubs. In the seventeenth century, this new public challenged the traditional monopoly of the court, the official academies, or the connoisseurs in such matters. It was enlarged during the eighteenth century thanks to the multiplication and diffusion of the periodicals, and the success of new forms of association: the book clubs, the reading societies, the masonic lodges. But it supposed also the exclusion of all those who lacked the competence which made possible the participation into a critical community welcoming only the persons who had sufficient wealth and culture.

1.2 Kant's Public Use of Reason

Habermas's perspective was based on one of the most famous text of the eighteenth century, Kant's *What Is*

Enlightenment? published as an article in the periodical *Berlinische Monatsschrift* in 1784. But it also distorted it. In his text Kant proposed a new way of conceptualizing the relation of the public to the private. He equated the public exercise of reason with judgment produced and communicated by private individuals acting 'as scholars before the reading public' and he defined the public as the sphere of the universal in opposition to the private considered as the domain of particular and domestic interests (which may even be those of an army, a church, or a state).

He inverted, thus, the accepted meanings of the two terms by associating the private use of reason with the exercise of an office or with the duties toward the state and by locating the public use of reason within a community without limitations or boundaries. If the private use of reason can be restrained legitimately in the name of public ends without prejudice to the progress of enlightenment, the liberty of the public use of the reason was absolute and has to be respected by the prince himself.

In this same text, Kant shifted the way in which the legitimate limits put on critical judgment should be conceived. Such limits were no longer defined by domains forbidden to methodical doubt; they laid in the position of the subject who exercises his reason. He could be constrained legitimately when he was executing the duties of his charge or of his status, but he was necessarily free when he acted as a member of the 'society of world citizens.'

That universal society was unified by the circulation of written works that authorized the communication and discussion of thoughts without the necessity of physical presence of the different interlocutors. Kant systematically associated the public use of reason with the production or reading of written matter by private individuals. The public was not construed on the basis of new forms of intellectual sociability such as clubs, cafés, literary societies, or lodges, because those associations were modalities of particular, circumscribed, and domestic entities. For him, only written communication was admissible as a figure for the universal.

Kant's conception of the public use of reason was drawn for the notion of the Republic of Letters as it was constructed through correspondence and the circulation of manuscript or printed texts from the late seventeenth century onwards (see Goldgar 1995). Founded on the free engagement of its members, on equality among its interlocutors, and on reciprocity and disinterestedness, the Republic of Letters which united the learned and the scholars provided a model and a support for thinking of the new public sphere. At the same time, it showed the distance separating the universality of the 'society of world citizens' and the restricted social composition of the 'reading public.' In Kant's time, this reading public was not the whole of society by any means, and those capable of writing were even less numerous. It is only when everyone will be able to exercise critically one's reason that humanity will experience 'an enlightened age.'

1.3 The French Philosophers: Public Opinion vs. Popular Opinions

Kant held the distinction between the public and the people to be temporary, transitory, and characteristic of a century which was a 'age of enlightenment' but not yet an 'enlightened age.' But for many thinkers of the eighteenth century, particularly within the framework of the French Enlightenment, the two notions constituted an irreconcilable dichotomy (see Ozouf 1989, pp. 21–35). Public opinion was defined precisely in contrast to the opinions of the multitude. Public opinion, set up as a sovereign authority and a final arbiter, was necessarily stable, unified, and founded on reason. The universality of its judgments and the constraining self-evidence of its decrees derived from that unvarying and dispassionate constancy. On the other hand, popular opinions were multiple, versatile, and inhabited by prejudice and passion. Such a qualification reveals a strong persistence of older representations of the people considered as subject to extremes, inconstant, and foolish.

When the concept of public opinion did emerge in France around 1750 as the superior authority to which all particular opinions must bow, the distinction between the public and the popular became essential. A new political culture thus took shape which transferred the seat of authority from the will of the king alone, who decided without appeal and in secret, to the judgment of an abstract entity embodied in no institution, which debated publicly and was more sovereign than the sovereign who was obliged to seek to win its approval and support.

This gave acuity and urgency to a fundamental question: how could the people, eliminated from the exercise of critical judgment by their lack of competence, be 'represented' in the new political space? A first answer to this question assigned this function to one or another of the theories of political representation confronted after the mid-eighteenth century. Following Keith Baker, three of the theories were fundamental: the absolutist theory, which made the person of the king the only possible representative of a kingdom divided into orders, estates, and bodies; the judiciary theory which instituted the Parliaments as interpreters of the consent or remonstrances of the nation; and the administrative theory, which attributed the rational representation of social interests to muncipal or provincial assemblies founded not on privilege but on property (see Baker 1990).

An alternative model of representation that removed the notion institutional setting—monarchical, parliamentary, or administrative—was offered by the category of public opinion. Substituting the self-

evidence of unanimity to the uncertainties of particular and popular opinions, detached from any form of exercise of governmental and state authority, the concept delegated the representation of the whole nation to those who were able to express its reasonable decrees: the men of letters.

2. Public Opinion: Voice or Tribunal?

Although everyone recognized the existence of public opinion and postulated its unity, there was no unanimous consent on two issues: first, who were its true spokesmen and, second, how was one to evaluate the self-evidence of its judgments? In eighteenth-century France, the answers to such questions were expressed through the metaphors used to designate this new entity: public opinion as a voice to be heard; public opinion as a tribunal which had to be persuaded.

2.1 From Theater to Tribunal

In 1775, in his *discours de réception* before the *Académie française*, Malesherbes forcefully expresses the idea, by then commonly accepted, that public opinion was to be considered a court of justice more imperious than any other. It is a tribunal 'independent of all powers and that all powers respect,' composed by the men of letters who are 'amid the public dispersed, what the orators of Rome and Athens were in the middle of the public assembled.'

There are several arguments contained in this comparison. First, it invested the new 'judges' with an authority that ordinary judges did not have. Their competence knew no bounds and their jurisdiction no limits; their freedom of judgment was guaranteed because they were in no way dependent upon the power of the ruler; their decrees had the force of self-evident propositions. With such a definition of public opinion, the men of letters were invested with the judiciary legitimacy of all traditional powers and with veritable public office.

Reference to this dispersed judiciary function had another meaning, however. It aimed at establishing a connection between the universality of judgments and the dispersal of persons, and at constructing a uniform opinion which, unlike that of the ancients, had no physical location in which it could construe and experience its unity. As for Kant later, it was the circulation of written matter and particularly the printed one which made it possible to envisage the constitution of a unified public in a nation in which people were necessarily separated from each other and formed their ideas individually. By associating the public nature of the written word, vastly increased by the invention of the printing press, with the supreme authority of the judgments pronounced in the new political sphere, Malesherbes (as many others) converted the plurality of reflections and criticisms that emerged from solitary reading into a collective and anonymous conceptual entity.

From the seventeenth century to the eighteenth century there had been a radical shift in the manner of conceiving the public. In the age of baroque politics the traits that defined the public were the same as those which characterized theatrical audiences: heterogeneous, hierarchized, and formed into a public only by the spectacle in which the most powerful effects were produced by hidden machines and secret maneuvers.

When the concept of public opinion did emerge, it effected a dual rupture. It opposed the art of dissimulation and secrecy by appealing to transparency. Before the tribunal of opinion all causes were to be argued without duplicity. But all citizens were not (or not yet) adept at joining together to form enlightened opinion. If the audience which mingled in the theaters potentially was composed of men and women from all social levels, the public which served as the tribunal to judge literary merits, religious matters, or political issues was more selected and homogeneous. When opinion was thought of more as actor than as acted upon, it became public, and by that token it excluded many people who lacked the competence to establish the decrees which it proclaimed.

2.2 From Audience to Public

Constituting the public as an entity the decrees of which had more force than those of the established authorities supposed several operations. They can be exemplified by the memoirs published in great number by both French lawyers and litigants from 1770 onward. They took the judicial comparison literally and exposed before opinion affairs which were examined by a court of justice. The cases subjected to the secret procedures of justice were thus transformed into public debates, shifting the place in which judgment had to be pronounced (see Maza 1993).

The most fundamental strategy required by such an operation consisted in endowing particular cause with general and exemplary value. The debt that a court noble refused to pay to his bourgeois creditors became an ideal occasion for denouncing unjust privilege, just as the arbitrary imprisonment of a young gentleman was an opportunity to criticize the *lettre de cachet*. To achieve such goals supposed, first, that the secrecy of judicial procedure had to be broken by the circulation of printed memoirs on the largest possible scale, and, second, that a dramatic style or a first-person narrative had to replace the customary legal prose. Universalizing the particular, making public what had been secret, and 'fictionalizing' discourse were the techniques that lawyers used to appeal to opinion and, in doing so, to proclaim themselves the authorized and legitimate interpreters of that opinion.

The traditional, discrete, and exclusive relationship that bound individuals to the king, the guarantor and guardian of domestic secrets, gave way to a totally

different situation in the public exposition of private conflicts. From that point of view, judicial memoirs are the exact inverse of the *lettres de cachet* accorded by the sovereign in response to requests from families interested in concealing familial disorders which sullied their honor. The memoirs displayed what the *lettres* concealed; they expected from the judgment of opinion what the *lettres* hoped to gain from the omnipotence of the monarch; they converted into public issues the scandals that the *lettres* were charged with burying.

It is in this sense that the new public space was built on conflicts proper to the private sphere, assigning political significance to simple familial or conjugal strife. In a parallel manner, the clandestine pamphlets stigmatized the aristocracy, the courtiers, the queen, and finally the king by revealing their corrupt mores. People became accustomed to seeing individual affairs transformed into general causes. This procedure was not restricted to pamphlet literature but also lay behind the lawsuits instituted by rural communities against theirs landlords or by journeymen workers against their masters. It was the process of 'privatization' in which individuals conquered autonomy and freedom removed from state authority which permitted the very existence of a new public sphere.

Even if, or because, it was defined as a conceptual entity, and not in sociological terms, the notion of public opinion which invaded the discourse of all segments of society—political, administrative, judicial—in the second half of the eighteenth century operated as a powerful instrument for social division and legitimization. The mobilization of the category founded the authority of all who, by affirming that they recognized its decrees alone, set themselves up as mandated to pronounce its judgments. Universal in its philosophical essence, the unified, enlightened, and sovereign public was far from universal in its social composition. The public sphere, emancipated from the domain in which the ruler held sway, thus had nothing in common with the shifting opinions and blind emotions of the multitude. Between the people and the public there was a fundamental difference.

3. *Prohibited Books, Subversive Words*

According to Darnton, this new public was detached from the traditional authorities by a wider and wider dissemination of philosophical ideas. In this process the role played by the great classics of the Enlightenment was perhaps less important than the texts that the eighteenth-century book trade called 'philosophical books': critical pamphlets, scandalmongering chronicles, anticlerical satires, pornographic works. The circulation of this clandestine and seditious literature, which depicted the despotic corruption of the monarchy or the depraved mores of the court, engendered a change in thought and in collective representations, henceforth detached the subjects from their old loyalties: 'Sedition was hatching. It was instilled in people's minds [...] We know for certain that it was communicated by a formidable instrument: the book' (see Darnton 1991, 1995).

But other historians like Farge have challenged this interpretation which sees the increase in seditious writings in the last two or three decades of the *ancien régime* as the matrix of a desacralization of the monarchy. She affirms: 'Popular opinion did not emerge from the cumulative reading of pamphlets and placards; it was not unilinear and did not base arguments on the sum of what is read' (see Farge 1994, p. 35).

Statements of disapproval or hatred of the sovereign, attacks on his acts, and expressions of a desire to kill him did not arise from the flourishing clandestine literature. Nor did they start with Damiens's attempt to assassinate Louis XV, which, rather than unleashing a proliferation of regicide discourse, persuaded the authorities of the reality of Jansenist and Jesuit plots and set them in pursuit of such literature. For Farge, 'the failed murder of Louis XV had a ready-made public opinion; it tells us more about the monarchy's reactions than about any new and original turn of popular thought' (p. 174).

The symbolic and affective disinvestment which transformed relations with authority when it was deprived of all transcendence was manifested but not caused by the wide diffusion of 'philosophical books.' The erosion of founding myths of the monarchy, the desacralization of royal symbols, and the distancing of the person of the king formed a body of representations which was 'already there,' 'ready-made' when the public sphere arose conceptually and sociologically.

4. *Women and the Public Sphere*

What was the role played by women in this new public sphere? Joan Landes' thesis opposes radically the 'symbolic politics of the emerging bourgeois public sphere (which) was framed from the outset by masculinist interests and assumptions' with the forms of sociability dominated by women in the *ancien régime*, for example, the salon society conceived as 'an alternative sphere of cultural production inside absolutism' (Landes 1988, p. 40). This is the reason why the exclusion of women from political rights during the French Revolution is inscribed by Landes within the logics of the enlightened condemnation of the role that women played within the *absolutist* public sphere: 'the structures of modern republican politics can be construed as part of an elaborate defense against women's power and public presence' (p. 203).

Against such a perspective Goodman argues that women played a central role both in the collective

project of the Enlightenment and in the construction of a new political and critical public space. For her, 'the French Enlightenment was grounded in a female-centered mixed-gender sociability that gendered French culture, the Enlightenment, and civilization itself as feminine' (Goodman 1994, p. 6). In this perspective the salon society did not belong to the old regime absolutist public sphere but had a fundamental role in shaping a 'discursive space' opened to public and political criticism. Between 1750 and 1775 women were essential in the government of such society because they kept within the limits of civility and politeness the tensions which necessarily arose from the free confrontations of opinions and the competitions for intellectual leadership: 'Enlightenment salons were places where male egos were brought into harmony through the agency of female selflessness' (p. 104).

According to Goodman (1994, p. 280), it is during the 1770s that women began to be excluded from the new public sphere. On the one hand, the harmony of Republic of Letters was then destroyed by a series a ferocious discords starting with controversies about Physiocracy. On the other hand, new forms of association (among them, literary societies called *lycées* or *musées* and masonic lodges) either excluded women or gave them a subordinate position. The return to 'masculine self-governance' prepared women's expulsion from the revolutionary political space: 'When the literary public sphere was transformed into the political public sphere in 1789, it had already become masculine; the 'democratic' republic of 1792 would reflect the limitations and exclusions of the Republic of Letters of the 1780s.'

It is possible to discuss such an interpretation. The Freemasonry did not ignore women's presence and gave them an important role in the 'lodges of adoption.' In these lodges many women were involved in the new values coined by the late Enlightenment: the philanthropic activities, the interest for political issues, the friendship between men and women (Burke and Jacob 1996, pp. 513–49). Besides that, it is clear that even between 1750 and 1775, the salon society was always very minoritary and limited to a restricted number of participants. They cannot be considered as the center of the Republic of Letters.

The collective construction of the new public sphere cannot be reduced either to a unique process or to a sole place. It arose from the intertwining between conceptual novelties and social practices, the alliance between speech and print, and the common (but dissimilar) participation of men and women.

See also: Absolutism, History of; Civil Society, Concept and History of; Enlightenment; Intellectual History; Kant, Immanuel (1724–1804); Public Sphere: Nineteenth- and Twentieth-century History; Space and Social Theory in Geography

Bibliography

Baker K M 1990 *Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century*. Cambridge University Press, Cambridge, UK

Burke J M, Jacob M C 1996 French freemasonry, women, and feminist scholarship. *Journal of Modern History* **68**(3): 513–49

Chartier R 1991 *The Cultural Origins of the French Revolution*. Duke University Press, Durham and London

Crow T E 1985 *Painters and Public Life in Eighteenth-century Paris*. Yale University Press, New Haven and London

Darnton R 1991 *Edition et sédition. L'univers de la littérature clandestine au XVIIIe siècle*. Gallimard, Paris

Darnton R 1995 *The Forbidden Best-sellers of Pre-revolutionary France*. Norton, New York

Farge A 1994 *Subversive Words: Public Opinion in Eighteenth-century France*. Polity Press, Cambridge, MA

Goldgar A 1995 *Impolite Learning: Conduct and Community in the Republic of Letters 1680–1750*. Yale University Press, New Haven

Goodman D 1994 *The Republic of Letters: A Cultural History of the French Enlightenment*. Cornell University Press, Ithaca

Habermas J 1962 *Strukturwandel der Öffentlichkeit. Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. Luchterhand, Neuwied (English trans 1989 *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Polity Press, Cambridge, MA)

Kant E I 1975 *Foundations of the Metaphysics of Morals and What Is Enlightenment (1784)*. Bobbs-Merrill, Indianapolis

Landes J B 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca

Maza S 1993 *Private Lives and Public Affairs: The Causes Célèbres of Prerevolutionary France*. University of California Press, Berkeley, CA

Ozouf M 1989 *L'homme régénéré. Essais sur la Révolution française*. Gallimard, Paris

Roche D 1988 *Les Républicains des Lettres. Gens de culture et Lumières au XVIIIe siècle*. Fayard, Paris

R. Chartier

Public Sphere: Nineteenth- and Twentieth-century History

The closely related concepts of public sphere and public space are rooted in ancient Greek distinctions of public and private and developed in early modern Europe through changing ideas of citizenship, civil society, and republican virtue (see *Civil Society/Public Sphere, History of the Concept*). They refer first and foremost to settings for political discussion among citizens about the order of their lives together, and secondarily to a much wider range of public culture including art, literature, and popular discourse on a host of subjects.

1. Publicness as a Modern Concern

New ideas about public discourse were complemented by development of new communications media, especially those dependent on print; rising literacy and education levels; growth of the state; and

expansion of popular political participation. In this process, the distinction of public and private took on new importance and complexity. On the one hand, the realm of public interaction expanded; cities were the primary setting for this, especially cosmopolitan trading and capital cities. Public spaces appeared literally with coffee houses, parks, theaters, and other places where people who were not bound by private relations gathered and communicated. They also grew metaphorically with printed sermons, pamphlets, newspapers, books in vernacular languages, journals that reviewed them, and other media of public communication. On the other hand, the state also expanded and with it the range of *res publica*, public things which included property held in common and matters of concern to the whole polity. Publicness took on a dual sense, referring both to openness of access and interaction and to collective affairs as managed by the state. The public referred both to the collective subject of democracy—the people organized as a discursive and decision-making public—and as its object—the public good.

The two dimensions were linked in the notion that political debate among responsible citizens was a way to arrive at sound understanding of common affairs. This depended on notions of participation in debate that developed in the realms of science and religion (Ezrahi 1990, Zaret 1999) and literature (Habermas 1962, Hohendahl 1982) as well as politics. Processes of rational-critical debate were held to form educated public opinion as distinct from other forms such as the 'representative publicity' of monarchs appearing before their subjects or the 'mere opinion' of uneducated masses. Interest in such public opinion grew alongside civil society as a self-organizing realm of social relations, and especially with the rise of democracy.

Complementing the growth of the public realm were new senses of the private. The 'privacy' of the family was more sharply affirmed. Women and children were increasingly sequestered in private homes, especially among the bourgoisie; the modern idea of the individual included the notion of nurturance and maturation in private as preparation for action in public, but this applied in the first instance to men. At the same time, economic activity was increasingly moved out of the household. In one sense, then, going to work meant going out into public, being exposed to the public gaze. In another sense, however, property relations continued to be understood as private in the sense that they were to be managed by individual persons and not the state. The eventual rise of business corporations further complicated the distinction, since these held property as artificial private persons but operated as collective, public actors, especially when shares of ownership were openly available on the market rather than closely held within families.

As the example of corporations and private property suggests, the distinction of public and private was sometimes difficult to sustain. This undermined the classical notion of the public sphere: 'The model of the bourgeois public sphere presupposed strict separation of the public from the private realm in such a way that the public sphere, made up of private people gathered together as a public and articulating the needs of society with the state, was itself considered part of the private realm' (Habermas 1962, pp. 175–6). As corporations became public actors while still claiming the status of private property, states began to intervene ever more into civil society and even intimate life. These trends joined with the rise of mass media, and especially professions and institutions devoted to the manipulation of public opinion through mass media (advertising, public relations) to undermine the conditions for the effective operation of the public sphere as a source for educated public opinion.

2. *History as Decline*

For many later analysts, the eighteenth century was a sort of 'golden age' of modern public life (see *Public Sphere: Eighteenth-century History*). Hannah Arendt (1958) theorized 'public' in terms of creative action, the making of a world shared among citizens, and saw the founding of the USA as a crucial example. Jürgen Habermas (1962) idealized eighteenth century English parliamentarianism, newspapers, and coffee house conversation. He presented the public sphere as a realm of civil society in which private citizens could communicate openly about matters of public concern, transcending their particular statuses and addressing the state without becoming part of it. Others have focused on Parisian clubs in the Revolution of 1789 (Amann 1975) or the literary journalism of the German Enlightenment (Hohendahl 1982).

Idealization of the eighteenth and sometimes the early nineteenth century public sphere underwrites explicit or implicit narratives of decline. In Habermas's classic *Structural Transformation of the Public Sphere*, for example, nineteenth and twentieth century public discourse is analyzed in terms of the loss of rational-critical capacity attendant on expansion of scale and the rise of public relations management that incorporated the public into the realm of administered society (see also Horkheimer and Adorno 1944). Such accounts reveal that the notion of public is not only basic to democracy, but linked to an apparent contradiction in democratic ideals. At least in its most prominent modern, liberal forms, democracy seems to promise both a politics based on reason rather than power, and the inclusion of all adults in the political process. The latter, the classic accounts of the public sphere suggest, undermines the former. The nineteenth and twentieth century history of democratic politics is taken to demonstrate this. Public opinion is increasingly 'managed' on the basis of research and manipulation rather than developed in open, rational-critical discourse.

Already in the nineteenth century itself, political theory of many persuasions emphasized the fragility and limitations of the liberal democratic conception of the public. Tocqueville, most famously, argued as early as the 1830s that the democratization of society tended to eliminate the *intermediary* public bodies which had traditionally refined opinion and furnished the individual a collective social identity outside the state. Engaged, politicized publics composed of distinct views and interests could be reshaped over time into mass publics—passive, conformist, and atomized before the state. Tocqueville's fear of the unmediated state would resonate with generations of critics of mass society.

Central to the analyses of historical decline in the public sphere is the idea that as its scope expanded, it was unable to sustain a high level of rational-critical debate. This happened with the rise of various sorts of opinion managers, from pollsters to public relations specialists, and with the institutionalization of political parties as products of negotiations among interest groups rather than arenas for debate about the public good in general. A key phase in this, accordingly to the critics, was the institutionalization of class politics during the late nineteenth and early twentieth centuries. Habermas treated this in terms of negotiation among interests, as distinct from reasoned debate about the public good. Negt and Kluge ([1964] 1993), however, famously emphasized the existence of a proletarian public sphere, rooted in distinctive experiences, and productive of reasoned internal discourse. Indeed, the German Social Democratic Party, like successful workers organizations elsewhere, created a parallel public sphere to that dominated by the bourgeoisie (Eley 1992, see also Katznelson and Zolberg 1987). This faltered principally because of a nationalization of popular politics, not class politics *per se* (Mosse 1975, 1987). Such nationalization was often accompanied by mass spectacle and reliance on mass media. Both did frequently collapse the distinction of public and private, especially in emotive appeals to national unity. Equally important was the rise of two-party electoral systems. Rather than being able to participate activity in public life, late twentieth century citizens of democratic societies were reduced to being invited to respond to plebiscites (Lowi 1985, Ginsburg 1986).

Historical research on the public sphere has questioned both the accuracy of the late eighteenth century ideal and the narrative of subsequent decline. Schudson (1998), for example, has argued that active participation of citizens in public discourse and politics has ebbed and flowed without linear trend. The ideal of the good citizen as an active participant in the public sphere has long been contrasted with the failings of actual citizens. A substantial line of research on 'civic culture' shifted the emphasis away from the problematic of public debate and towards a more general idea of participation and incorporation into civic life, for example, in voluntary associations (see, in particular, Almond and Verba 1966 and critical analysis in Somers 1993, 1995). This, most researchers argued, grew through the nineteenth and early twentieth centuries. More recently, however, critics of late twentieth and early twenty-first century democracies have emphasized a decline in civic participation (e.g., Putnam 2000).

3. *Participation and Exclusion*

Although openness was basic to the ideology and theory of the public sphere, various forms of exclusion were basic to actually existing publics in the nineteenth and twentieth centuries. Women, for example, had been prominent in late eighteenth century salon society and elite public spheres, but were widely excluded from nineteenth and twentieth century public space; this was mirrored by a commensurate bias in political thought (Landes 1988, Elshtain 1993). Feminist struggles made gender a basic political issue in new ways (Ryan 1990, 1992). In addition, women were prominent in other movements, such as abolitionism in the USA, and may have figured less in subsequent historical accounts than in actual nineteenth century social movements. This may be partly because of different styles of address in the public sphere, including a greater emphasis on narrative and less on argumentation (Lara 1999). Similarly, Blacks were largely excluded from the US public sphere, even after abolition. This encouraged a similar duality: struggles for inclusion and development of parallel public discourse marked by distinctive communicative styles (Diawara 1995).

This has occasioned more general inquiry into 'counterpublics' which contest the apparently neutral dominance of more mainstream publics (Fraser 1992, Warner 2001). This argument is rooted in Negt and Kluge's ([1964] 1993) account of the proletarian public sphere. Members of religious communities, ethnic groups, and the women's movement have been among the many who have constructed counterpublics. During the classic heyday of the late eighteenth and early nineteenth century British public sphere, thus, many artisans and workers were denied participation in the public sphere. They were not simply and unambivalently members of a proletarian public sphere, though they did develop their own media and organizations and to some extent constitute a counterpublic (Gilmartin 1999). They claimed the right to participate in the dominant, unmarked public sphere and challenged those who introduced restrictive measures to make it a specifically 'bourgeois' (or more generally, propertied) public sphere (Jones 1983). The same people who excluded those with less wealth from the public sphere nonetheless claimed it in unmarked form as simply *the* British public.

Accounts of the public sphere by both contemporaries and historians have been tacitly linked to the idea of nation, since nation defined the range of citizens entitled to participate in public affairs and the basis for legitimacy of modern, especially democratic, states (Calhoun 1997). In liberalism's triumphant early phase, experiments in republicanism and constitutionalism placed these issues squarely on the table. They asked how new political structures were to acquire legitimacy and how, within those structures, the plurality of opinion was to be represented. Defenses of free speech and the press, too, signaled an appreciation of the importance of venues in which views could be widely and, in principle, freely circulated. At the same time, the rhetoric of nationalism typically presented the nation not as creature of public discourse but as pre-existing unity, commonly on ethnic or cultural grounds. This is true even, if ironically, in the USA, the nation quintessentially created through the public action of its citizens. A tension running through American history has pitted the capacity for creation of a new political order exemplified by the founders against claims that the defining national essence is already in place and needs only defense from the loyal.

Notions of a vital republican public sphere have been influential especially in American historical studies, visible in the work of J. G. A. Pocock, Bernard Bailyn, Gordon Wood, and more recently, Michael Warner. Here the focus, in Hannah Arendt's terms, is on the capacity of action in public not only to define individual identity but to create the world citizens share in common. Republicanism, implied several things: (a) a concerted effort to reach agreement about the public good, as distinct from the attempt to find the net balance of private, individual goods—the essence of utilitarianism; (b) a sense of politics as a work in progress regarding the appropriate structures and limits of the political, as opposed to a hallowed set of procedures for achieving particular ends—such as the cult of the constitution; and (c) a belief that individuality and identity are only fully expressed and enjoyed through engagement in public life, rather than constituted beforehand in the private realm. These writers emphasize the ephemeral nature of republicanism and the rapid emergence in the nineteenth century of a variously liberal, national and representative (rather than participatory) model of the public sphere. In American Studies, as in Frankfurt School critical theory, this decline is commonly associated with the rise of mass culture—and the blurring of divisions between reasoned political discourse and other social processes of opinion formation.

4. *Cities and Media*

Transformations of public life were also shaped by the ways in which communications media and the physical public spaces of cities changed through the nineteenth and twentieth centuries. Richard Sennets (1977) *The Fall of Public Man*, for example, contrasts a vibrant and highly varied eighteenth and nineteenth century urban life (especially among the bourgeoisie) to a reduction in both the occasions for interaction across lines of difference and the psychological ability to manage open relations with strangers well. Citizens retreated, Sennett suggests, into both privacy and the conformity of mass culture.

Urban public spaces anchor face-to-face interaction, which was basic to the classical image of the public sphere, conceived of largely in terms of direct interaction. This is important for a sense of serendipitous contact among people of diverse backgrounds. Parks, cafes, and theaters all inform this. Many of Europe's cities, especially older ones, were distinctive in their pedestrian character and their scale. From the mid-nineteenth century, however, this began to shift. Hausmann's post-1848 construction of the Paris boulevards and destruction of some of the local *quartiers* is taken as emblematic (e.g., by Harvey 1985). Both suburbanization and less pedestrian-friendly, larger-scale designs for newer cities have changed the character of public interaction in Europe, largely in ways familiar earlier in the US Commentators throughout the nineteenth and twentieth centuries have been concerned for the implications of this for democracy. As Mumford (1938, p. 483) wrote, 'One of the difficulties in the way of political association is that we have not provided it with the necessary physical organs of existence: we have failed to provide the necessary sites, the necessary buildings, the necessary halls, rooms, meeting places … '

This has been linked to a huge increase in scale. One crucial result is that directly interpersonal relations organize proportionately less of public life; mediations of various kinds have become increasingly important (Calhoun 1988, Garnham 1992, Keane 1991, Thompson 1995). This is partly a matter of communications media *per se*; it is also a matter of growing importance for bureaucracies and other formal organizations. The eighteenth and early nineteenth century public sphere was closely tied to the spread of print media. The nineteenth and early twentieth centuries were the heyday of great urban newspapers; after this, media that transcended locality became increasingly important. First radio and then television fundamentally altered the public sphere. They contributed to a shift in what was publicly visible as well as in how public discourse was organized (Meyrowitz 1985, Thompson 1995). New media shared both information and emotionally powerful images widely. Broadcast media are often charged with furthering the debasement of reason by substituting powerful images for sustained analysis, by appealing to a lowest common denominator in audiences, by blurring the lines between entertainment and critical discourse, and by centralizing control over messages in the hands of a few corporations. Some caution is in order about

generalizations from medium, though, as newspapers produced their own mass forms, with broadly similar if perhaps less extreme effects.

Crucially, the changes in scale and media forms of the twentieth century challenge the classical eighteenth century model of the public sphere, which was grounded in readership on the one hand and directly interpersonal discourse on the other. Any account of contemporary public life must deal with a dramatic increase in indirect, mediated relationships. Transformations of media continue, of course, and the implications of the Internet and computer-mediated communication are important topics of contemporary investigation (Rasmussen 2000). Relatedly, the dominant approaches to the public sphere have emphasized domestic politics within nation-states. Especially during the last years of the twentieth century, however, communications media and the rise of transnational organizations have facilitated growth of an international public sphere (Thompson 1995, Held 1995, Calhoun 2002). To what extent transnational public opinion can steer global society directly, and to what extent only through the agency of state structures, remains a crucial question.

Finally, though this article concentrates on the public sphere in Western European and American history, it is important to note that the nineteenth and twentieth century also saw dramatic and transformative growth in public discourses elsewhere. In late Imperial and early Republican China, for example, the category of 'public' became influential and both directly interpersonal and mediated public spheres were created (Wakeman 1993, Rankin 1993). This was basic to development of the idea of citizen, as distinct from subject, to constitution of notions of public welfare, and to cultural transformations such as the introduction of new vernacular writing. It also contributed to the growth of both nationalist and class politics in the twentieth century. Further transformations followed, attendant on changes in media, politics, and culture, and creating important linkages within the Chinese diaspora as well as with the West.

An unusually vital print-mediated public sphere developed in India. The early history of this is closely interwoven with the history of British colonialism and opposition to it. The history of Indian public life has been an important setting for examining the role of both international power relations and issues of domestic diversity. Not least, questions like the claims of nonreaders in a public sphere dominated by print media have surfaced and shaped not only Indian but international discussions. Broadcast media have played an important role recently in the constitution of a new Hindu nationalism (Rajogopal 2001). In many settings, including in the European countries and the USA as well as both India and China, the development of a public sphere has been closely related to the development of national consciousness, identity, and political contestation.

5. *Conclusion*

Growth and transformation of public spheres has been a major theme in nineteenth and twentieth century history. It is crucial to the history of democratic politics, including not only electoral and parliamentary regimes, but also broader struggles in civil society and beyond national borders. A dominant historical theme is the inclusion of a broader range of participants in public discourse. This has been coupled with analyses of the implications of growing scale, including mismatch between local—especially urban—public spaces and enlarged range of potential participants, dynamics of inclusion and exclusion, and changing roles of media. A prominent line of analysis suggests a concomitant decline in the rational-critical capacity of public discourse. This is challenged by both research on the limits of earlier public spheres and the strengths of later counterpublics. The importance of the public sphere is in any case not limited to argumentation but includes also the formation and spread of culture, as in the development of national identity.

See also: Civic Culture; Civil Society, Concept and History of; Civil Society/Public Sphere, History of the Concept; Communication and Democracy; Mass Society: History of the Concept; Media, Uses of; Participation: Political; Policy History: State and Economy; Political Communication; Political Discourse; Privacy: Legal Aspects; Public Sphere and the Media; Public Sphere: Eighteenth-century History; State and Society

Bibliography

Almond G, Verba S 1966 *The Civic Culture*

Amann P 1975 *Revolution and Mass Democracy: The Paris Club Movement in 1848*. Princeton University Press, Princeton, NJ

Arendt H 1958 *The Human Condition*. University of Chicago Press, Chicago

Bellah R 1985 *Habits of the Heart: Individualism and Commitment in American Life*. University of California Press, Berkeley, CA

Calhoun C 1988 Populist politics, communications media, and large scale social integration. *Sociological Theory* **6**(2): 219–41

Calhoun C 1997 Nationalism and the public sphere. In: Weintraub J, Kumar K (eds.) *Public and Private in Thought and Practice*. University of Chicago Press, Chicago, pp. 75–102

Calhoun C 2002 Constitutional patriotism and the public sphere: Interests, identity, and solidarity in the integration of Europe. In: DeGreiff P, Cronin C (eds.) *Transnational Politics*. MIT Press, Cambridge, MA

Diawara M (ed.) 1995 *The Black Public Sphere*. University of Chicago Press, Chicago

Eley G 1992 Nations, publics and political cultures: Placing Habermas in the nineteenth century. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA, pp. 289–339

Elshtain J B 1993 *Private Man, Public Woman*, rev. edn. Princeton University Press, Princeton, NJ

Fraser N 1992 Rethinking the public sphere: A contribution to the critique of actually existing democracy. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA, pp. 109–42
Garnham N 1992 The media and the public sphere. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA
Gilmartin K 1999 *Print Politics: The Press and Radical Opposition in Early Nineteenth-Century England*. Cambridge University Press, Cambridge, UK
Ginsburg B 1986 *The Captive Public: How Mass Opinion Promotes State Power*
Habermas J 1984 1988 *Theory of Communicative Action*. Beacon Press, Boston, Vols. 1 and 2
Habermas J 1991 *The Structural Transformation of the Bourgeois Public Sphere: An Inquiry into a Category of Bourgeois Society* [trans. Burger T]. MIT Press, Cambridge, MA
Harvey D 1985 *Consciousness and the Urban Experience*. Johns Hopkins University Press, Baltimore, MD
Held D 1995 *Democracy and the Global Order*. Polity, Cambridge, UK
Hohendahl P U 1982 *The Institution of Criticism*. Cornell University Press, Ithaca, NY
Horkheimer M, Adorno T 1944 *Dialectic of Enlightenment*. Continuum, New York
Jones G S 1983 *Languages of Class: Studies in English Working Class History, 1832–1982*. Cambridge University Press, Cambridge, UK
Katznelson I, Zolberg A (eds.) 1987 *Working-Class Formation: Nineteenth-Century Patterns in Western Europe and the United States*. Princeton University Press, Princeton, NJ
Keane J 1991 *The Media and Democracy*. Polity, Cambridge, UK
Landes J 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca, NY
Lara M P 1999 *Moral Textures: Feminist Narratives in the Public Sphere*. University of California Press, Berkeley, CA
Lowi T 1985 *The Personal President*. Cornell University Press, Ithaca, NY
Meyrowitz J 1985 *No Sense of Place*. Oxford University Press, New York
Mosse G L 1975 *The Nationalization of the Masses*. Cornell University Press, Ithaca, NY
Mosse G L 1987 *Masses and Man: Nationalist and Fascist Perceptions of Reality*. Wayne State University Press, Detroit, MI
Mumford L 1938 *The Culture of Cities*. Secker and Warburg, London
Negt O, Kluge A [1964] 1993 *The Public Sphere and Experience*. University of Minnesota Press, Minneapolis, MN
Putnam R 2000 *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster, New York
Rajogopal A 2001 *Politics after Television: Religious Nationalism and the Reshaping of the Indian Public*. Cambridge University Press, Cambridge, UK
Rankin M 1993 Some observations on a Chinese public sphere. *Modern China* **19**(2): 158–82
Rasmussen T 2000 *Social Theory and Communication Technology*. Ashgate, London
Ryan M 1990 *Women in Public: Between Banners and Ballots, 1825–1880*. Johns Hopkins University Press, Baltimore, MD
Schudson M 1998 *The Good Citizen. A History of American Civic Life*. Free Press, New York
Sennet R 1977 *The Fall of Public Man*. Knopf, New York
Somers M 1993 Citizenship and the place of the public sphere: Law, community, and political culture in the transition to democracy. *American Sociological Review* **58**: 587–620
Somers M 1995 What's political or cultural about political culture and the public sphere? Toward an historical sociology of concept formation. *Sociological Theory* **13**: 113–44
Thompson J 1995 *Media and Modernity*. Stanford University Press, Stanford, CA
Tocqueville A de [1840] 1844 *Democracy in America*. Scribners, New York
Wakeman F 1993 The civil society and public sphere debate: Western reflections on Chinese political culture. *Modern China* **19**(2): 102–38
Warner M 2001 *Publics and Counterpublics*. Zone Books, Cambridge, MA
Yeo E J 1998 *Radical Femininity: Women's Self-Representation in the Public Sphere*. Manchester University Press, Manchester, UK
Zaret D 1999 *Origins of Democratic Culture*. Princeton University Press, Princeton, NJ

C. Calhoun

Publishing as Medium

Publishing, in its most general sense, refers to the act of making a work of words or images known to the public. Book publishing encompasses a range of activities through which an author's manuscript is turned into a printed and bound book which can be made available to an audience of readers. These activities are carried out by networks of organizations and individuals who coordinate production, and the promotion and distribution of books to their potential publics.

1. *History and Description*

In the ancient world, there were a variety of ways in which the reproduction and circulation of texts occurred. Evidence from China, India, Greece, and Rome suggests that publishing was carried out under the auspices of religious, political, and commercial bodies; in some places, individuals copied and traded books among themselves. During the Middle Ages, publishing in Europe was centered in the monasteries where manuscripts were reproduced primarily for religious purposes. Following the twelfth century, scribes associated with European universities took over this task, serving the clergy as well as selling books to new readers among officials, the nobility, professionals, and merchants. But with the appearance of the printing press in 1450, book publishing and bookselling became organized into an industry, its members striving to sell books for profit to a growing and increasingly far-flung audience.

Before the seventeenth century in Europe, and the nineteenth century in North America, the book trade tended to remain undifferentiated and unstandardized. While conditions varied in different regions, the functions of publisher, printer, and bookseller were often melded into a single individual who printed and sold books at the request (and payment) of authors. Nevertheless, the trade in books would rarely have been enough to support the printer who often published a newspaper as well as books, and who derived a large portion of his income from printing bills, stationery, and other jobs. As the market for books grew, a more complex division of labor gradually emerged, with different segments of the industry responsible for publishing, printing, and selling. Along with this, many publishers and booksellers began to specialize in specific types of books, such as texts or medical works.

Within the modern book industry, the publisher is central to the entire process of getting books from author to reader. Publishing firms select, edit, and prepare manuscripts for production, and also assume the financial risks involved in bringing a book into print. It is the publisher who interacts and contracts with the author (or the author's representative, the agent), and it is also usually the publisher who determines a sales strategy and how a book will be marketed. Following the pattern of other industries, a separate wholesaling sector has taken over from publishers some of the functions of warehousing and filling orders. A more differentiated retail sector has also developed, as books are sold through both exclusive bookstores and retailers whose primary business is not in books, through educational, employment, and voluntary organizations, and through mail order and electronic venues.

Since the early twentieth century, the publishing industry has been characterized by a degree of specialization that corresponds in large part to the market for particular kinds of books as well as to channel of distribution. The most visible form of publishing is in trade books, those titles directed at a general audience. However, both textbooks and professional titles compose a significant part of the overall book market, and in many developing countries textbooks make up the single largest segment. Other important segments include scholarly and religious books. As these various divisions between book sectors developed during the twentieth century, different organizational and sales practices also appeared within their corresponding publishing houses. These divisions now appear to be breaking down as merger activity has brought together dissimilar publishing houses under a single roof.

2. *Publishing as a Field for Analysis*

The world of book publishing has been viewed in rather contradictory ways: as a rarefied environment of literary connoisseurs, and as a profane site where expression is reduced to a commodity; as an industrial backwater, and as a central player in the spread of capitalism. Perhaps these contrasting images are related to the way in which publishing, as a field for analysis, belongs to no single group. Those who work in the industry, journalists, and scholars from several disciplines have all been influential in shaping ideas about publishing's significance.

2.1 *Historical Perspectives*

Befitting lives spent among writers and books, numerous publishers, both prominent and obscure, have penned reminiscences of their activities. To some degree, this set the tone for accounts of the book business since the nineteenth century. Many such accounts are really more histories of publishers than of publishing, describing the rise and fall of various publishing houses and the personalities who directed them.

A different kind of scholarship emerged with the tradition known as 'the history of the book,' which many date to the publication of Febvre and Martin's (1958) *L'Apparition du Livre*. Febvre, who was a prominent member of the *Annales* school, and his co-author described the early publishing and bookselling scene in Europe, while also attempting to account for the book's impact on the spread of various ideas and cultural movements. For some, such as Darnton (1982), the field of research that has evolved since Febvre and Martin's book is equivalent to the social and cultural history of print communication. In this view, the object is to understand the role books have played in society at various times and in various places, and to understand how books have interacted with other social developments and processes. Darnton's proposal is also methodological in nature, suggesting that the book be studied in terms of a communication circuit that runs from author to publisher to printer to shipper to bookseller and finally, to reader.

Although the program outlined by Darnton has been influential among book historians, other historians have taken a different approach by emphasizing the history of printing and the subject of bibliography, that is, the study of books as physical objects. Bibliography is concerned with describing books and identifying their attributes, such as paper, type, and bindings, so as to be able to distinguish different editions or printings. This kind of examination can also be used to provide information about particular publishers and manufacturers, or about authorial intention (see, for example, McKenzie 1986). Frequently, the underlying purpose of such studies also differs from that emphasized by the history of the book. The intent is to make the book itself the central focus of attention rather than using the book as a means for discovering other social patterns.

2.2 Social Science Perspectives

While historians have long recognized the book as an important area of study, social science research on book publishing has been intermittent and surprisingly sparse considering social scientists' interest in other forms of mass media. Instead, some of the most detailed descriptions and critiques of publishing have been written by people directly associated with the publishing industry. Of course, books and book publishing tend to be central to the academic's work, and so, despite the relative lack of empirical research, have not escaped academic scrutiny altogether. The state of the publishing world is taken to be a matter of public concern, and scholars and other public intellectuals frequently address a general audience when discussing the significance of developments in publishing for print culture. Throughout the twentieth century, such discussions have often reflected the assumptions of mass culture theory. Commentators question whether publishing practices that cater to a mass market leave room for the quality books that sell in moderate numbers, and whether increasing commercialism and centralization will bring about a standardization of the available literary fare (see, for example, Dutscher 1954).

As this suggests, the curiosity of publishing tends to lie in its dual nature as a commercial enterprise on the one hand, and as a producer of literary objects associated with artistic, educational, and religious values on the other. Indeed, the perceived difficulty of reconciling these two dimensions of publishing—often described as the conflict between culture and commerce—is probably the central problem analysts of the book industry have confronted.

Beginning with work by Escarpit (1958), a relatively small group of sociologists has tried to dissect the workings of the publishing industry. These studies draw on organizational theory (sometimes called the production of culture perspective) which, especially during the 1970s and 1980s, was being applied to various media firms and industries. This perspective begins with the assumption that publishers are similar to other kinds of bureaucratic organizations, facing the same problems of how to reduce uncertainty and risk (Hirsch 1972). Scholars in this tradition have provided specific accounts of how books are acquired and distributed, and how publishing strategies vary according to size and type of firm. Common to much of this work is a stress on the gatekeeping role that publishers play, meaning the ability to determine what ideas and literary forms will make it into print and thus be able to enter the marketplace of ideas.

Congruent with trends in mass media and popular culture studies, more recent researchers have reacted against theories of mass culture and hegemony which posit audiences as easily manipulated and degraded by mass marketed literature. In this view, scholarship needs to grant the reader a more privileged place in analyses of publishing, in terms of both acknowledging the reader's active role in choosing and finding meaning in reading material, and in understanding how conceptions of the audience shape the work of book professionals. Moreover, this theoretical stance questions the very distinction between popular and quality books. Instead, some suggest that the fear of commercialism has been primarily a cover for the discomfort of cultural elites who see their authority over cultural matters challenged (Long 1985–86, Radway 1997). Similarly, it has been argued that publishers play an important role in the production of literary value (Verdaasdonk and van Rees 1991).

3. Themes in Publishing Research

Changes occurring within the publishing industry during the last few decades of the twentieth century shaped many of the questions posed in publishing research. Renewed interest in publishing has been concurrent with its integration into large, multinational, entertainment conglomerates. Researchers have been especially attentive to the consequences of concentration, the significance of publishing as a global phenomenon, and the convergence between books and other entertainment and information media.

3.1 The Dynamics of Concentration

The American situation perhaps best illustrates certain trends in the industry. During the post-World War II era, most major New York publishing houses were transformed from private or family-owned companies into public corporations. Various social changes contributed to this development. With the post-War education boom, many publishers desired to meet a new demand for printed matter by expanding their existing educational divisions or founding new ones. This coincided with the aging of heads of publishing houses who wished to protect their heirs from inheritance taxes. Issuing public stock appeared to be an attractive proposition, a way to both raise capital and guarantee the long-term future of the firm.

Along with the move towards corporatization, publishing houses became favored acquisition targets by electronics firms in the 1960s. Although these buyers later divested their publishing properties, merger and acquisition activity was renewed in the mid-1970s. These years saw the formation of conglomerates with holdings in film, broadcasting, music, newspapers, and magazine publishing as well as books. Book publishers would not add much in terms of revenues, but entertainment conglomerates hoped that book content could be reproduced across a variety of media, and cross-promotional strategies could be

employed. Similar waves of merger and acquisition activity were seen in the next two decades. By this time, foreign-owned companies were often doing the acquiring, including media giants such as Bertelsmann, News Corporation, and Pearson. The result has been the consolidation of the nation's largest publishers into a handful of companies which also include American-owned Viacom, AOL Time Warner, and Disney. Concentration at the level of distribution happened somewhat later. While chain booksellers began to grow in the 1960s with the establishment of standardized outlets in shopping malls, their decisive dominance of the bookstore market did not occur until the 1990s with the appearance of the superstore.

The effects of concentration have been a constant source of debate. Along with longstanding concerns about declining quality and the homogenization of reading material, these developments raise questions about changes in the organizational culture of the book business. Attempts to rationalize the publishing process, invest heavily in promotion, and exploit new markets have altered the balance of power within the industry, diminishing the autonomy of editors while giving increased clout to accounting, marketing, and subsidiary rights divisions. At the same time, publishers have grown dependent on a few large bookseller customers who subsequently play a more authoritative role in the industry.

Complicating the issue of concentration is that while a declining number of publishers account for a large share of book sales, the absolute number of publishers, along with titles published, has steadily risen. Some argue that these various trends result in much diversity at the margins of the industry, and a less elitist conception of books and publishing at the center. For others, the concentration of control of the book business means that a small number of for-profit companies have undue power over what individuals can and will read.

3.2 Globalization and National Identities

The role that the circulation of books plays in maintaining or undermining specific identities has long been an important theme in publishing research. The relatively new concept of globalization is readily applied to the publishing industry. On the one hand, trends in ownership mean that multinational companies with far-flung holdings are now common among the major players. But the integration of international markets is the primary way in which publishing can be conceived as a global phenomenon. The selling of rights to publish or distribute books in foreign territories has become one of the most important sources of revenue for major publishers. The trade in books goes in many directions, but disproportionately travels from Europe and especially the USA to the rest of the world.

This raises both theoretical and policy questions about national sovereignty and the maintenance of national identities. In Canada, for instance, the overwhelming presence of first British and then American books has hampered the growth of a national publishing industry. While government policy has helped to preserve Canadian-controlled firms, foreign-controlled publishers in Canada remain more profitable. Also, many Canadian publishers can survive only because they have the exclusive right to sell American books, the popularity of which is attested to by their dominance of best-seller lists in the English-language market. Even in Quebec, where French publishers play a strong role, American translations are plentiful. Many argue that this situation has helped to further Americanize Canadian culture, diffusing American tastes, styles, and values through the medium of popular books.

Economic agreements between nations to eliminate free trade barriers have in the past often excluded cultural products, with books in particular singled out for exception. There are significant obstacles to maintaining a viable publishing industry in less developed countries, and as the Canadian situation demonstrates, even highly developed nations find it difficult to make the economics of publishing work without protective measures. With the USA and the European Commission trying to eliminate the cultural exception in trade negotiations, these issues will probably move more to the forefront of publishing research.

3.3 Unyoking Content from Print

The international trade in books has been greatly dependent on the state of copyright law, and a significant body of research has examined how copyright and its breach have affected the publishing industry. However, the development of new technologies along with publishing's integration into entertainment conglomerates have raised new issues concerning the control of literary property. Content is now easily transportable across media, and entertainment firms attempt to create synergies, as it has been termed, between book, film, television, music, magazine, and electronic divisions. This introduces a new level of instability into the publishing process. On the one hand, traditional arrangements for the control of intellectual property are challenged when content is so mobile, with publishers and authors in conflict over who will own the new forms that emerge. On the other hand, new technologies, including electronic books and on-demand publishing, where single copies are printed as needed, are utilized by distributors, authors, and software companies as well as publishers. Such developments suggest that what it means to engage in publishing may be more difficult to define in the future.

See also: Audiences; Internet: Psychological Perspectives; Literacy and Illiteracy, History of; Literacy and Orality; Literary Texts: Comprehension and Memory; Mass Communication: Empirical Research; Mass Communication: Normative Frameworks; Mass Communication: Technology; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Textbooks

Bibliography

Altbach P G, Hoshino E S (eds.) 1995 *International Book Publishing: an Encyclopedia*. Garland Publishing, New York

Coser L A, Kadushin C, Powell W W 1982 *Books: the Culture and Commerce of Publishing*. Basic Books, New York

Darnton R 1982 What is the history of books? *Daedalus* **111**: 65–83

Dutscher A 1954 The book business in America. *Contemporary Issues* **5**: 38–58

Escarpit R 1958 *Sociologie de la Litterature*. Presses Universitaires de France, Paris [1971 *Sociology of Literature*. Frank Cass, London]

Febvre L P V, Martin H-J 1958 *L'Apparition du Livre*. Éditions Albin Michel, Paris [1990 *The Coming of the Book: the Impact of Printing 1450–1800*. Verso, London]

Hirsch P M 1972 Processing fads and fashions: an organization-set analysis of cultural industry systems. *American Journal of Sociology* **77**: 639–59

Horowitz I L 1991 *Communicating Ideas: The Politics of Scholarly Publishing*, 2nd edn. Transaction, New Brunswick, NJ

Lane M, Booth J 1980 *Books and Publishers: Commerce Against Culture in Postwar Britain*. Lexington Books, Lexington, MA

Long E 1985–86 The cultural meaning of concentration in publishing. *Book Research Quarterly* **1**: 3–27

McKenzie D F 1986 *Bibliography and the Sociology of Texts*. British Library, London

Powell W W 1985 *Getting Into Print: The Decision-making Process in Scholarly Publishing*. University of Chicago Press, Chicago

Radway J A 1997 *A Feeling for Books: The Book-of-the-month Club, Literary Taste, and Middle-class Desire*. University of North Carolina Press, Chapel Hill, NC

Verdaasdonk H, van Rees C J 1991 The dynamics of choice behavior towards books: the role of cultural products in cultural behavior. *Poetics* **20**: 421–37

L. J. Miller

Punishment, Comparative Politics of

1. The Minimally Punishing Democratic State

The most striking development in twentieth-century penal politics was the emergence in the advanced democracies of the minimally punishing state. The development was striking because it occurred during one of the largest crime waves in recorded history—a tidal wave ignored by all democratic political systems for one or even several decades. The modern democratic state was able to dismiss the demand for a state response to the crime wave constituted by the annually mounting numbers of thousands, even millions of crime reports. Most democratic governments developed surprising capacities *not* to punish in spite of the victim's entitlement to punishment implied by penal codes. (No penal code contains a warning label to victims that the punishments it threatens will be applied at most to three percent of those arrested and one percent of those who violate its terms.) Attrition rates for criminal cases became extremely high in all democratic systems, from the apparently most punitive to the least. A decision seems to have been taken in most democracies that the state will never punish (or at least imprison) much more than 100 per 100,000 people at any time, no matter which or how many crimes are committed. It has been possible for democratic states *not* to enforce the criminal law on a massive scale, although enforcing the penal code against violence and predation is popular among voters.

2. The Convergence on Minimal Punishment

2.1 The Convergence

The facts of convergence on minimal punishment levels are clear. In 1940, most of the currently advanced democracies had capital punishment; by 2000, only two of them still inflicted the death penalty and only on a much reduced scale. In 1960, there were in the advanced democracies on average 11 prisoners (for all offenses) per reported robbery; in 1990, only three, and, without Japan, only one. In 1960, the range in punishment lay between New Zealand with 30 prisoners per robbery and Norway with 24, on the one hand, and the USA with only two; by 1990, the range (excluding Japan) lay between three prisoners per robbery in California and Austria, and fewer than 0.5 in Italy and The Netherlands. Between 1960 and 1990, the democratic state largely relinquished the threat of death punishment and reduced to nearly token levels the threat of prison punishment.

2.2 Elite Conversion to Leniency

The decision to ration punishment rested on what had become, by the 1960s, a gap between the lenient mentalities of political elites and the more traditionally punitive, if also ambivalent, mass public. This gap was best explained by Norbert Elias, who traced what he

called 'the civilizing process'—the diffusion of anti-violence mentalities among and by the middle and upper classes since the Middle Ages, especially since the monopolization of legitimate violent punishment by the state. (Elias unfortunately did not theorize the strong decivilizing processes on full display in the twentieth century, including the decivilizing results of overcivilizing the state: Elias 1982.)

2.3 The Emergence of Penal Sentimentalism

The medieval elite–mass gap widened with the top-down spread of the ideas of the eighteenth-century Enlightenment and of nineteenth-century Romantic, liberal, and socialist movements–humanitarianism, human perfectibility, and the social responsibility for crime. Gradually created was a powerful group of aristocratic and bourgeois reformers who changed elite attitudes toward the lower classes, the poor, the disorderly, and eventually even minorities. The revision of (elite) attitudes toward the poor, the sick, the aged (as well as vagrants, drunks, addicts, juveniles) which eventually brought about the welfare state was accompanied by a closely related revision of attitudes toward criminals—which brought about the minimally punishing state. Both the poor and criminals were seen as irresponsible victims of society—to be helped, not chastened. And the penal sentimentalism of enlightened seculars soon also came to characterize the elites still faithful to Christianity but less and less faithful to traditional hard-line Pauline concepts of punishment, either in this world or the next.

2.4 Leniency in Spite of Marxian Capitalism

The growing leniency of elites in nineteenth- and twentieth-century capitalist democracies was directly contrary to Marxist predictions; indeed, Marx had to be 'turned on his head' to explain the diminution of punishment under democratic capitalism and its exasperation under communism. Ironically, one of the few serious Marxist works on penality, Rusche and Kirchheimer's *Punishment and Social Structure* (1968), was also one of the first and few accounts of 'the general tendency toward leniency' under capitalism; it purported to use Marxist concepts (the relationship between capitalism, labor supply, and punishment) to explain penal evolution generally, and the trend toward penal leniency, in particular. It argued grandly that 'Every system of production tends to discover punishments which correspond to its productive relationships…The use or avoidance of specific punishments, and the intensity of penal practices (are) determined by social forces, above all by economic and fiscal forces' (Rusche and Kirchheimer 1968). But to account for the trend toward leniency under capitalism, Rusche and Kirchheimer were forced to be eclectic, rather than dialectic, to explain leniency in terms only incidentally related to evolving productive relationships. They invoked instead the rising standards of living for the lower classes; a 'consequent' decline in criminality; the development of 'penal liberalism' and social insurance ('welfare statism'); declining conviction rates; and successful penal reform campaigns by bourgeois professionals. (This particular package of explanations would not, of course, explain why, a decade after World War II, when crime became rampant, crime-based punishment rates under capitalism continued to decline (Rusche and Kirchheimer 1968).)

2.5 Leniency in Spite of Social Diversification

In contrast to Marx, but with the advantage of greater hindsight, Durkheim correctly predicted (in 1902) growing leniency under democratic capitalism, but he attributed leniency to a supposed fracturing of the collective conscience due to the division of labor (Durkheim 1902/1978). Actually, social diversification has not destroyed consensus on the evils of violence and predation. The factor most pertinent to the decline of punishment, in fact, was not social and economic but political: the ability of elites (as theorized by Michels, Mosca, and other 'democratic elitist' theorists) to ignore rank and file opinion in otherwise democratic structures.

2.6 Leniency in Spite of Democratic Competition

Convergence on leniency was indeed paradoxical—given democratic competition and the persisting, even growing, crime-induced popularity of punishment among voters. The paradox is possibly explained by:

(a) party elites' ability to manipulate electoral and political agendas, to collude in order to suppress conflict over penal policy;

(b) the ability of the Left for decades to ignore constituent demands for security;

(c) conservative complicity in the antipunishment project;

(d) antipunishment bias in the ideological-group lobby and the ideology production system (ideological or advocacy groups vastly outweighed interest groups);

(e) stratification of the media, with elite media reinforcing antipunishment ideological hegemony and rewarding elites for antipunishment action;

(f) the ability of elites to freeze and even to reduce prison capacity during an historic crime wave;

(g) declining detection rates (though more and more suspects were being produced by the police);

(h) the formation of a powerful international antipunishment movement enabling lenient domestic elites to invoke international norms, praise, and interventions on behalf of their domestic agendas;

(i) the ability of lenient political elites sharply to reduce prosecution and prison-sentencing rates, even in compulsory-prosecution systems and in inquisitorial justice systems, with low evidentiary barriers to conviction; and

(j) the stigmatization within elite circles of responses to demands for punishment as 'populist', or worse.

Thus penal policy has often become one of many policy areas in which public influence, especially in Europe, has often been weak—in spite of the public's interest, known preferences, and voting power.

Convergence on leniency also occurred in the advanced democracies in spite of periods of high unemployment, increasing ethnic diversity, a drugs pandemic, periods of unremitting terrorism, the flourishing of organized crime, and a sevenfold increase in serious crime (using robbery as an indicator).

3. *The Differential Development of State Capacity Not to Punish*

But if all democracies have become minimally punishing states, some have become less minimally punishing than others. The token-punishment state has been politically most viable in Europe and least viable in the USA. The factors behind the convergence on leniency operated with differential effect in Europe and the USA, resulting in somewhat differential state capacity *not* to punish. Among 90 national and subnational jurisdictions (all in advanced democracies), the strongest predictor (in cross-sectional analysis) of per capita incarceration rates is the serious crime rate (using murder or robbery rates). The USA has had higher per capita prison rates because it has had much serious crime to be punished: it is (marginally) harder not to punish in high-crime jurisdictions than in low-crime jurisdictions.

3.1 *Does Party Count?*

One of the best predictors of per capita prison rates, after crime, is the strength of the Left since 1960. Although its historical record for leniency in penal matters is quite checkered (Robespierre, Lenin, Stalin, Mao, Pol Pot, and Castro come to mind), the Left—at least after World War II and in the democracies—became the major political force behind the antipunishment movement. The Left became solidly against capital punishment, sympathetic to criminals, defensive of prisoner rights, and hostile to spending on prisons, especially after the myth of therapeutic prisons was shaken (but far from destroyed). Ironically, the Left was slower than the Right to heed the demands of its own constituents—the predominantly lower-class and minority victims of crime—but the Left had major stakes in trying to prove that the welfare state could replace the punitive state in preventing crime. As it became clearer in the 1970s and 1980s that criminals thrived in even the most perfected welfare states (Sweden and The Netherlands, for example), that working class voters could be even angrier about crime (especially minority crime) than conservative voters, that the crime wave denied for so long by criminologists was real, and that left-voting middle-class penal liberals were not likely to desert to the Right, the rhetoric of the Left and eventually some of its policy stands began to correspond more closely to widespread public sentiments of insecurity and ineffective justice (Tham 1999).

Moreover, in spite of the correlation between Leftism and low per capita punishment rates, the creation of the minimally punishing state has required, as noted above, the complicity of conservatives. The lowest imprisonment levels in their respective postwar state histories occurred under Reagan in California, Agnew in Maryland, and Wallace in Alabama; the lowest levels of federal imprisonment occurred under Nixon; the lowest imprisonment levels occurred under religious party hegemony in The Netherlands and Italy, and under Catholic auspices in Ireland; the leading historical examples of decarceration (Rutherford 1984) occurred under Liberal and Tory auspices in England and Liberal Democratic Party auspices in Japan. The role of conservatives in the antipunishment movement is, of course, an old story. Capital punishment abolition, after all, was carried out by enlightened despots in the eighteenth century; by the czar and other monarchs in the nineteenth century; and by nineteenth-century bourgeois liberal oligarchs in The Netherlands, Italy, and Portugal. Abolition after World War II was proposed in West Germany by the extreme Right and supported by the Christian Democrats in West Germany and Italy. Abolition in France in 1981 was supported by Gaullist and Giscardian elites. Restoration of capital punishment in the UK and Canada has repeatedly been blocked by Tory cabinet ministers.

4. *Punishment and State Formation*

We should not, however, exaggerate the differences among the advanced democracies. Even in the USA public tolerance for penal sentimentalism lasted well into the 1970s and early 1980s. And the subsequent backlash that doubled and tripled US prison populations only brought the prisoner–robbery ratio (1.8 in 1970, 1.6 in 1990, 3.5 in 1997) into line with European ratios—about the same as Sweden in 1970 or Austria in 1990, and nowhere near the 11 prisoners per robbery average for the advanced democracies in 1960.

The modern state was built by substituting state for private punishment. Bentham could write without fear of contradiction: 'No punishment, no government; no government, no political society' (Bentham 1831, p. 7). Many democracies of the twentieth century set out to disprove Bentham by showing that a state could protect its citizens and preserve order with few sanctions involving life or liberty. They succeeded in preserving the state's monopoly of *legitimate* violence—and proving the electoral viability of the minimally punishing state—but often at the cost of allowing the level of *illegitimate* violence to return to near medieval (i.e., prestate) levels.

See also: Crime and Class; Crime and Ethnicity (Including Race); Crime: Sociological Aspects; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Drugs: Illicit Use and Prevention; Imprisonment: Sociological Aspects; Prison Life, Sociology of; Prisons and Imprisonment; Punishment: Social and Legal Aspects

Bibliography

Bentham J 1831 *Jeremy Bentham to his Fellow Citizens of France on Death Punishment*. Heward, London

Cusson M 1990 *Croissance et decroissance du crime*. 1st edn. Presses Universitaires de France, Paris

Durkheim E 1902/1978 Two laws of penal evolution. In: Traugott M (ed.) *Emile Durkheim on Institutional Analysis*. University of Chicago Press, Chicago pp. 153–80

Elias N 1978–82 *The Civilizing Process*. Blackwell, Oxford, UK

Enguéléguélé S 1998 *Les Politiques Pénales*. L'Harmattan, Paris

Garland D 1990 *Punishment and Modern Society*. Clarendon Press, Oxford, UK

Hood R G 1996 *The Death Penalty: A Worldwide Perspective*. 2nd rev. and updated edn. Clarendon Press, Oxford, UK

Jankovic I 1984 Socialism and the death penalty. *Research in Law, Deviance and Social Control* **6**: 109–37

Jescheck H-H 1983–4 *Die Freiheitsstrafe und ihre Surrogate im deutschen und ausländischen Recht*. Nomos, Baden-Baden, Germany

Lynch J P 1987 *Imprisonment in Four Countries*. Bureau of Justice Statistics, Washington, DC

Rusche G, Kirchheimer O 1968 *Punishment and Social Structure*. Russell and Russell, New York

Rutherford A 1984 *Prisons and the Process of Justice: The Reductionist Challenge*. Heinemann, London

Tham H 1999 Lag & ordning som vänster projekt? Socialdemokratin och kriminalpolitiken. In: *8 Reflektionen om Kriminalpolitik*. Swedish Council for Crime Prevention, Stockholm

Törnudd P 1993 *Fifteen Years of Decreasing Prisoner Rates in Finland*. Oikeuspolittinem tutkimuslaitos, Helsinki, Finland

Young W, Brown M 1993 Cross-national comparisons of imprisonment. In: Tonry M (ed.) *Crime and Justice: A Review of Research*. University of Chicago Press, Chicago, Vol. 17, pp. 1–43

Zimring F, Hawkins G 1991 *The Scale of Imprisonment*. University of Chicago Press, Chicago

R. C. Fried

Punishment: Social and Legal Aspects

1. Punishment in Modern and Postmodern Society

As recently as the 1970s, this article would have almost certainly situated punishment in modern society as the culmination of an evolution toward greater rationality and humanity. This vision of penal progress influenced both reformers and scholars of punishment.

The past was represented in Enlightenment tracts against bloody scaffolds and dark dungeons. The present was embodied in reform oriented prisons rooted in scientific studies of human behavior. The prison represented a space of public control (as against the tradition of private 'dungeons') in which the state could be held accountable for its treatment of offenders. It offered a penalty, the deprivation of freedom, uniquely suited to the core values of modernizing societies (mobility, responsibility, personal gain). Imprisonment was also a punishment that could be set precisely to establish a scale of deterrence, or varied in practice to allow individualized treatment.

The prison did not altogether replace earlier modes of punishment (Spierenberg 1984). Corporal punishment, fines, and the death penalty, the mainstay of penal technique before the nineteenth century, remained, although in diminished and modernized form. In the twentieth century, new forms of community sanction, probation and parole, were introduced, designed to elaborate the prison's capacity for surveillance while minimizing its tendency to isolate the offender from the normalizing influences of the community at large.

The limitations on twentieth century penology were long seen in the context of a future that saw the elimination of residues of private vengeance seeking, and the perfection of scientific techniques for resocializing 'delinquent' subjects. This program, known as 'correctionalism' or 'rehabilitation', grew from nineteenth-century approaches to moral reform anchored in religion, especially Protestant strands of Christianity. Beginning in the late nineteenth century, this was secularized (if only in some respects) into a social scientific penology that promised a range of interventions from psychological therapy to community action to secure the normalization of criminal offenders. Even dystopian visions, like Anthony Burgess's *A Clockwork Orange*, assumed that authorities would acquire the capacity for transforming the behaviors and motivations of deviant individuals.

Social theorists from Durkheim (1964) to Foucault (1977) saw penal progress as signaling a fundamental break in social organization (Garland 1990). Durkheim suggested that more 'primitive' societies were unified by themes of shared identity expressed in strict penal laws and reinforced in rituals of severe punishments. More modern societies were unified by a complex division of labor expressed in civil law and reinforced in varied individual relationships. Durkheim's theory suggested that the proportion of penal to civil law in a particular society formed a rough and ready index of its modernization. Max Weber, writing in the same period, saw in the evolution of prisons and penal law the process of rationalization that he had identified as the crucial pattern of social change in modern societies (Garland 1990). At the other end of the twentieth century, Foucault (1977) saw the prison as an exemplar of a new technology of power that became more generalized in the nineteenth century, that is the disciplinary organization of space for the purpose of governing potentially unruly masses of people. Although rejecting the humanitarian impulse as the primary motivation behind the rise of the prison, Foucault saw punishment as moving toward forms of exercising power that created fewer economic costs and political resistances. Where Durkheim emphasized the normative integration produced by the division of labor, Foucault saw modern society as evolving subtle forms of surveillance and coercion to establish conformity. Thus, while Foucault rejected the substantive normative optimism regarding 'progress' in punishment and society, he recognized the modern approach as distinctive and representative of an increase in the capacity and security of power.

The narrative of penal progress began to unwind in the 1970s (Rotman 1995). Ideas of rehabilitation and deterrence that dominated modern thinking about punishment for more than two centuries have not disappeared, but rather are viewed more skeptically and compete with a broad array of conceptions, some openly directed at vengeance. These developments have not been unique to punishment, but resonate with a growing skepticism about science, rationality, and reformist government that some have identified with the idea of postmodernity (Beck 1992). While modern punishment evidenced a persistent aspiration to progress beyond itself and to become a form of education, therapy, or pastoral care, postmodern punishment has embraced punishment as a good in itself and as a model for governing in other fields like education, healthcare, and welfare (Simon 1997).

These changing cultural sentiments are reflected in an increase in penal severity throughout many societies since the 1970s (Stern 1998). The United States has been far and away the leader in this new pattern. At the end of the twentieth century, some three percent of adult residents of the United States were in some form of correctional custody, up from one percent in 1980 (Maguire and Pastore 1998). The post-World War II trend toward abolishing the death penalty has also become much more complicated since the 1970s. The post-1989 collapse of Communist governments in Europe and Asia, and the end of Apartheid in South Africa unleashed a wave of abolitions. In the same time period, however, a number of countries that had temporarily ceased executions resumed them (Hood 1996). The United States is again the leading example. After a brief judicially imposed moratorium on executions in the early 1970s (Zimring and Hawkins 1986) the death penalty has returned to use in 38 of 50 states and overall numbers of those executed have returned to the levels of the 1950s.

2. *The Origins, Meanings, and Functions of Punishment*

Punishment, as that term is used here, refers to all legally regulated practices intended to denounce, harm, control, or reform specific individuals convicted of violating a criminal law or comparable norm. This is a broader definition than that used by many courts (for instance, in the United States), where many kinds of disadvantages or burdens imposed on those convicted of crimes are not considered punishments, including forfeiture of property linked to crime and requirements for certain types of offenders to register their address with police. At the same time, it is a more limited definition than might follow from popular understandings of punishment which can include illicit acts of vengeance and self imposed forms of discipline (Hudson 1996).

At the end of the twentieth century a wide variety of penal practices were in use, including the death penalty, imprisonment, fines, and various forms of supervised release under conditions of limited freedom in the community. Societies vary enormously with respect to the relative proportion of these punishments. Popular sentiment has been growing for a return to penal practices of the past, such as flogging and symbolic markings aimed at shaming the offender.

There is also a growing set of penal practices outside the formal limits of the criminal law. These include forfeiture of property linked to criminal activity, confinement of those deemed mentally abnormal and dangerous (e.g., sex offenders), deportation of aliens deemed a threat to security, and mechanisms to notify private individual and community organizations of the presence of convicted sex offenders in their proximate vicinity. Forms of civil punishment date back to the Middle Ages but have become more common in recent years.

For interpreting trends in penal practices it is useful to distinguish two further terms: 'penology' and 'penality.' Penology is often defined as the 'science of punishment' and can refer to the rigorous study of penal techniques by professionals in both scholarly and institutional settings. While this reflects the

prestige of science in modern society, it is too narrow to capture the range of nonscientific discourses about punishment, both earlier and today. Penology here refers to any organized and systematic discourse about punishment. 'Penality' is a term that has been applied usefully to thinking about the entire constellation of penal practices and penologies in operation in a society at a particular time (Garland and Young 1983).

Any particular practice of punishment or penology has a genealogy that can in principle be traced back to an origin. Likewise, we can speak of the meaning or function intended for a specific penal practice or projected by a particular penology. It is safe to assume, however, that any society's penality will reflect a wide range of different penal practices, institutions, and penologies with quite different origins, meanings, and functions. As the philosopher Nietzsche pointed out at the turn of the nineteenth century, 'the concept of ''punishment'' presents, at a very late stage of culture … not just one meaning but a whole synthesis of ''meanings''' (Nietzsche 1994, p. 57). This is surely even more true a century later (O'Malley 1999).

Since the nineteenth century, philosophical, legal, and social science writers have 'cannonized' four 'appropriate' functions of punishment: deterrence, prevention (or incapacitation as it is more commonly referred to today), rehabilitation or reform, and retribution. Deterrence as a systematic theory of punishment or penology originates with the influential Enlightenment penal thinker Beccaria (1996) and was elaborated by Jeremy Bentham (Hudson 1996). Deterrence views punishments as evils that can only be justified by their effect on the rational calculus of the free and responsible individuals. The ideal of deterrence in this perspective is to threaten just enough 'pain' or 'disutility' in response to the proscribed act that a rational chooser who considers both the 'pleasure' or 'utility' of the crime and the likelihood of being caught and convicted will choose to forbear. Deterrence presumes a common perception of the disutility of punishment and what might be thought of analogically as an acoustically consistent social space in which subjects, as it were, receive deterrent signals from the state. Despite considerable criticism over the decades of both of these core presumptions, deterrence has proved one of the most enduring penologies whose logic can be found in recent trends like mandatory sentences.

Reform or rehabilitation means using the punishment process to identify and to transform pathological features in the character or motivation of an offender that have led to the offense, with the goal of reintegrating the offender into the community as a productive (or at least nonoffending) participant. Like deterrence, rehabilitation is committed to the possibility of reducing crime by inducing the offender to change his or her pattern of behavior. But while deterrence presumes that offenders share with nonoffenders a capacity for rational calculation of utility that the penalty seeks to invoke, rehabilitation assumes that offenders are deviant in their subjective features and seeks to normalize them.

Beginning in the early nineteenth century, reform was associated first with religious penitence and later with rigorous military-like discipline, surveillance and control (Rothman 1990, Foncault 1977). The penitentiary style prisons constructed from the late eighteenth century on were designed to support reform through isolation, compulsory contemplation, and intense surveillance tied to rewards for cooperation and sanctions for lack of cooperation. The cellular structure helped frame the individual as the crucial focus of penal art and lent itself to the comparative evaluation upon which modern criminological science has been based.

Twentieth-century rehabilitation took its cue from the rise of scientific criminology in the same period that emphasized the roots of crime in pathological individual anomalies. While much of the initial scientific criminology focused on seemingly irreparable biological anomalies (which lent itself to the eugenic aspect of prevention), its ethos and later much of its substance fostered the quasiexperimental approach that guided twentieth-century rehabilitation with its emphasis on individual diagnosis and treatment.

Prevention or incapacitation received systematic treatment in the late nineteenth and early twentieth century in response to the rise of scientific criminology. Punishment, on this theory, works by making it impossible for a given offender to commit crimes while he or she is confined. In this sense prevention achieves its crime control effects by removing the offender from the community either temporarily or permanently. Earlier in the twentieth century, this was framed primarily in eugenic terms. Imprisonment or execution prevented offenders from passing on criminal traits to future generations. Eugenically oriented preventive penology suffered the same cultural decline that eugenics did generally during the second half of the twentieth century, due to its perceived link to the Holocaust and the generally genocidal strategies of the Nazis (Duster 1990). In the 1980s, however, in response to the rise of crime rates that started in the 1960s and the seeming failure of rehabilitation methods, penologists became interested in the incapacitative effect of prisons. On this theory, the crimes an offender would have committed had he or she been free constitute the benefits of prevention, regardless of whether the offender resumes criminal behavior upon release (assuming that imprisonment does not actually lead to an increase in the frequency or severity of crime).

The first three of the four common legitimate functions of punishment are often referred to as 'utilitarian' because they are justified primarily by their hopes for reducing crime. A fourth function, retribution, generally is associated with the idea that punishment serves a moral function by inflicting

suffering on those who deserve to suffer for having violated the moral rights of another. The core issue in retribution from that perspective is that of proportionality, that is, making sure that the punishment is proportionate to the moral wrongfulness of the criminal act. Retribution is often confused with revenge since both emphasize the suffering of the offender and do not emphasize the goal of reducing crime. But while revenge highlights the individual right of the victim to carry out a reprisal, retribution highlights the moral values of society as a whole and views crime as an affront to those values.

Retribution received its most famous and influential statement by the philosopher Immanuel Kant. Kant was so adamant that punishments be targeted at (and only at) moral desert that he suggested it would be morally necessary to carry out all punishments even if organized society was about to dissolve and had no future concern for the criminal behavior of offenders (Kant 1974). During much of the nineteenth and twentieth centuries retribution was taken to be a powerful sentiment in society, but it received little elaboration from penologists more interested in pursuing utilitarian leads. Only in the last third of the twentieth century, with the collapse of enthusiasm for rehabilitation, did a new penological discourse on retribution revive under the slogan 'just deserts,' (Von Hirsch 1976). With severe punishment politically popular in many countries there has been a renewed effort by philosophers and legal theorists to provide new foundations for retribution.

These four oft-repeated functions or justifications of punishment refer necessarily to the idea of legitimate or appropriate punishment. Many other functions of punishment have been articulated by critics and by sociological observers of penality. For example, sociologists in the Marxist tradition have suggested that the key function of punishment is to help regulate labor markets for the stability of capitalist enterprise (Rusche and Kirchheimer 1968), or to produce obedience to and even loyalty toward exploitive and authoritarian governments (Thompson 1975, Hay et al. 1975). Sociologists in the Durkheimian tradition suggest that punishment has the function of increasing the sense of solidarity among members of community. Others have even suggested that the function of punishment is to produce a class of delinquents convenient for manipulating society from below through the control of informants and provocateurs (Foucault 1977). Such functions are necessarily latent, since making them an explicit purpose of punishment would delegitimize penal institutions. A third kind of function also excluded from modern penal thought is not quite so hidden: punishment should serve as an instrument for private vengeance for the victim or catharsis for the public. As shall be addressed later, 'populist punitiveness' (Bottom 1995) has become a major feature of contemporary, postmodern penality.

It is tempting to treat the history of modern punishment as a shift in dominance among the four justifications. Deterrence seemed to prevail for much of the nineteenth century, followed by rehabilitation, then a resurgence of retribution, and by the end of the twentieth century, by prevention. While helpful in capturing moments of shifting power within the overall effort to coordinate and control penality at any one time, this approach is likely to disguise the significant practical overlap between the established functions and to cut off observation and analysis of different logics entering the contemporary field of penality. Likewise, the general discussion of the four functions treats them as eternal verities rather than as a cluster of responses to a specifically modern framing of the problem of punishment since the eighteenth century. Enlightenment critics attacked monarchical penality as both excessive and ineffective. They set out to develop penal techniques that were simultaneously appropriate to the new status of individuals as moral if not political equals while effectively checking crime. Each of the four major functions of punishment offers a vision of how punishment can provide crime control effectively or at least in a way that is consistent with the political status of the individual as a free subject.

3. *The Triumph of Modernist Penology*

In the aftermath of the Enlightenment and the age of Revolution, penology was largely a debate between proponents of deterrence, primarily lawyers and politicians who saw the whole population as the primary subject of penal sanctions, and proponents of normalization who saw the individual offender as the subject. By the 1920s establishment opinion was well set in favor of increasing the rehabilitative capacity of the penal system, even at the expense of deterrence. The ubiquity of three particular penal practices reflects the spread of this approach. Juvenile justice was the first great step in a global pattern toward reform and rehabilitation followed by the most advanced societies in the late nineteenth and early twentieth centuries (Rothman 1980, Garland 1985). In creating a distinct penal system for youth, these societies reflected both the influence of scientific ideas on penal practice (the recognition of adolescence as a distinct phase of life was being promoted by the rising field of psychology), and the belief that at least some offenders should be handled in a nonharmful way. Juvenile courts, the heart of these juvenile systems, in turn showcased two other features that would become synonymous with modernist penology, probation and the indeterminate sentence.

The crucial legal invention of probation was the 'probation officer,' an agent of the court who presumably gathered knowledge of the offender's individual circumstances by conducting close research into the offender's biological, psychological, and

sociological circumstances. As a punishment for both minor and serious offenses, probation subjected offenders to the supervision of a probation officer, charged to direct their rehabilitation and empowered to enter their homes, to search them, to interview their neighbors and co-workers, and to arrest them for violations. The indeterminate sentence meant that most of those convicted of serious crimes like robbery or assault were given a formally quite lengthy sentence with the expectation that an administrative agency, usually known as the 'parole board,' would exercise discretion to release the inmate from prison and allow him or her to serve some remaining years under probation-like supervision in the community (by parole agents) (Simon 1993).

The establishment of probation and indeterminate sentencing programs reflected a willingness to de-emphasize both deterrence and retribution as penal values. An indeterminate sentence (e.g., 5 years to life for a burglary) left the potential offender free to dream that he or she would get out in five. Juvenile justice, probation, and parole all lowered deterrence by diminishing the severity of the expected sanction, since all three involved reducing the amount of time that would otherwise be served in confinement. Moreover, in all three institutions, the legal authority to punish was grafted onto an elaborate decision-making mechanism incorporating not only legal but social and psychological facts.

By World War II most industrial societies operated some version of this modernist penology. The Soviet Union under dictator Joseph Stalin adopted a particularly humanist version of this penology (albeit undercut by the penal violence of the state security organizations). Of the three systems engaged in World War II only the Axis powers repudiated modernist penology, adopting penological ideas consistent with fascism central concerns with racial unity and populist excess (Evans 1996). The victory of the Allies in World War II secured a place for modernist penology in the legal and political structure of the post war world. For the next 25 years the rehabilitative logic of modernist penology enjoyed unprecedented ideological support and investment. As the European economies struggled to rebuild and the United States enjoyed surging affluence, the underlying presumption of modernist penology, that society had a significant stake in the adequate adjustment of even its most recalcitrant members, had broad currency.

Modernist penology also enjoyed unprecedented legal support after World War II due to the genocidal use of penal-like sanctions by the Axis powers. As the victorious Allies began to construct an international legal order to assure that such actions would never be repeated (an aspiration that remains unfulfilled), modernist penology became a constitutive element. With its focus on the whole circumstance of the individual and on rehabilitating rather than eliminating those the state defines as enemies, modernist penology reaffirmed core elements of human rights ideology and offered a normative check within the penal organs of the state against that organ becoming an adjunct to genocidal crimes of the state. For example opposition groups powerfully resisted the death penalty in the post-World War II international legal order. Abolition and limitation of capital punishment has been a central issue of concern since the Enlightenment, but the Nazi atrocities underlined the potential for execution to become mass state policy.

4. *The Crisis of Modernist Penology*

The dominance of modernist penology began to falter across Europe, Australia and North America during the 1970s. Skepticism about the capacities of scientific penology to control crime, combined with greater demands for government accountability, shattered the consensus that punishment was a kind of technical state service aimed at resocializing offenders (Garland 2001). Punishment as a painful harm to the guilty subject, an embarrassing residue to modern penology, returned as a positive product of the system.

The delegitimation of modernist penology spread initially by the political left that had traditionally supported it. This was an outgrowth of the same human rights ideology that also enshrined modernist penology. The emphasis on equal treatment, especially in the framework set by racial discrimination in the United States, made suspect the tremendous discretion enjoyed by administrative decision makers under modernist penology.

In the 1970s these kinds of considerations led to a wave of critical discourse on modern punishment comparable in quality and consistent in ethos with the Enlightenment writers on punishment. The United States was the primary center for the resurgence of retributive thought that came to be known as the 'just deserts' movement. Widely read books by Von Hirsch (1976), the American Friends Service Committee (1972), Frankel (1973), and Fogel (1975), among many others, deeply criticized modernist penology and offered in its place a more transparent and presumably more limited state use of coercion to punish or to incapacitate those who act clumsily and dangerously. The central principles of just deserts were sentences proportionate to the seriousness of the offense, equal punishment of those committing similar offenses, and transparent decision making that would prevent the possibility of the total loss of liberty faced by minor criminals.

By the 1980s and 1990s, the philosophy of just deserts largely had been superceded by two quite different penologies that are postmodernist and post-Enlightenment. One is a discourse of punishment as risk management; it represents some continuity with the modernist quest for effective crime control, but in terms that reject the key values of modernism and the

Enlightenment. The other is a populist punitiveness (Bottom 1995) or authoritarian populism (Hall et al. 1978) that valorizes victims and the therapeutic potential of vengeance and cruelty (Scheingold 1984, 1991, Sarat 2001).

5. *The New Penology and Populist Punitiveness*

As modernist penology began to unravel in the 1970s, scholars have begun to describe the emergence of an inchoate successor penology, a 'new penology' (Cohen 1985, Feeley and Simon 1992), or 'managerialism' (Bottom 1995). This new penology is most apparent in its contrast with the central tendencies of modernist penology. Rejecting the historic modern commitment to the individual as the target of penal justice, the new penology deals primarily in aggregates, dangerous classes, high risk populations and the like. The new penology replaces the objective of transforming criminal deviance with that of managing it. The new penology drops the focus on the social context and interest in penality in favor of the system itself as the primary source of normative order.

These elements can be seen in such recent penal practices as preventive detention of sex offenders, mandatory minimum sentences (Tonry 1995), and super max prisons (King 1999). These practices combine a discourse of risk management and a skepticism about the potential for transforming offenders with a systematic preference for decision-making criteria rooted in forms of knowledge internal to the system itself.

While the 'new penology' describes the way professionals within the penal system understand their mission, there is a large gap between the managerialism of this discourse and increasingly emotive populist discourse of punishment which emphasizes vengeance, shaming, and the therapeutic benefit of punishment for victims. Throughout the era of modernist penology popular consciousness about punishment was presumed to lag behind the progressive logic of professionals. With the exception of aberrant regimes like the National Socialists in Germany in the 1930s and 1940s, most modern governments treated penal policy as something that should be shaped by professional concerns with only grudging concessions to popular demands. What is distinctive about the turn in penal policy since 1980, especially in the United States, is the open embrace of populism as an acceptable logic of punishment. This is exemplified by the embrace of the death penalty by virtually all political leaders in the United States and the enactment of expressly emotional laws like 'Three-Strikes' (Sechrest and Sichor 1996).

See also: Criminal Law and Crime Policy; Death Penalty; Deterrence; Deterrence: Legal Perspectives; Deterrence Theory: Crime; Juvenile Justice: International Law Perspectives; Prisons and Imprisonment; Punishment, Comparative Politics of; Sanctions in Political Science

Bibliography

American Friends Service Committee 1972 *Struggle for Justice*. Hill and Wang, New York

Beccaria C 1996 *Of Crime and Punishment*. Trans. Grigson J. Marsilio Publishers, New York

Beck U 1992 *The Risk Society: Towards a New Modernity*. Trans. Ritter M. Sage Publications, London

Bottom A 1995 The philosophy and politics of punishment and sentencing. In: Clarkson C, Morgan R (eds.) *The Politics of Sentencing Reform*. Clarendon Press, Oxford, UK

Cohen S 1985 *Visions of Social Control*. Polity Press, Cambridge, UK

Cummins E 1994 *The Rise and Fall of California's Radical Prison Movement*. Stanford University Press, Stanford, CA

Durkheim E 1964 *The Division of Labor in Society*. Trans. Simpson G. Free Press of Glencoe, New York

Duster T 1990 *Backdoor to Eugenics*. Routledge, New York

Evans R J 1996 *Rituals of Retribution: Capital Punishment in Germany, 1600–1987*. Oxford University Press, Oxford, UK

Feeley M, Simon J 1992 The new penology: Notes on the emerging strategy of corrections and its implications. *Criminology* **30**: 449–74

Fogel D 1975 *We are the Living Proof: The Justice Model for Corrections*. Anderson, Cincinnati

Foucault M 1977 *Discipline and Punish: The Birth of the Prison*. Trans. Sheridan A. Pantheon, New York

Frankel M E 1973 *Criminal Sentences*; *Law Without Order*. Hill and Wang, New York

Garland D 1985 *Punishment and Welfare*. Gower, Brookfield, VT

Garland D 1990 *Punishment and Modern Society*: *a Study in Social Theory*. University of Chicago Press, Chicago

Garland D, Young P 1983 Toward a social analysis of penality. In: Garland D, Young P (eds.) *The Power to Punish: Contemporary Penality and Social Analysis*. Heinemann, London

Garland D 2001 *The Culture of Control*: *Crime and Social Order in Contemporary Society*. University of Chicago Press, Chicago

Hall S, Critcher C, Jefferson T, Clarke J, Roberts B 1978 *Policing the Crisis: Mugging, the State, and Law and Order*. MacMillan, London

Hay D, Linebaugh P, Rule J G, Thompson E P, Winslow C 1975 *Albion's Fatal Tree: Crime and Society in Eighteenth-Century England*. Pantheon, New York

Hood R G 1996 *The Death Penalty: a World Wide Perspective*. Oxford University Press, New York

Hudson B A 1996 *Understanding Justice: An Introduction to Ideas, Perspectives and Controversies in Modern Penal Theory*. Open University Press, Philadelphia PA

Kant I 1974 *The Philosophy of Law* (trans. Hastie W 1877). Reprinted 1974, A M Kelley, Clifton, NY

King R 1999 The rise and rise of the Supermax: an American solution in search of a problem. *Punishment and Society* **1**: 163–86

Maguire K, Pastore A L 1998 *Sourcebook of Criminal Justice Statistics 1997*. Bureau of Justice Statistics, Washington, DC

Nietzsche F W 1994 *On the Genealogy of Morals*. Trans. Diethe C. Cambridge: Cambridge University Press, UK
O'Malley P 1999 Volatile punishments: Contemporary penality and neo-liberal government. *Theoretical Criminology* **3**: 175–96
Rothman D J 1980 *Conscience and Convenience*: *The Asylum and its Alternatives in Progressive America*. Little Brown, Boston
Rothman D J 1990 *The Discovery of the Asylum*: *Social Order and Disorder in the New Republic*, rev. edn. Little Brown, Boston
Rotman E 1995 The failure of reform: United States, 1865–1965. In: Marris N, Rothman D (eds) *The Oxford History of the Prison*: *The Practice of Punishment in Western Society*. Oxford University Press, New York
Rusche G, Kirchheimer O 1968 *Punishment and Social Structure*. Russell and Russell, New York
Sarat A 2001 *When the State Kills*: *Capital Punishment and the American Condition*. Princeton University Press, Princeton, NJ
Scheingold S 1984 *The Politics of Law and Order: Street Crime and Public Policy*. Longman, New York
Scheingold S 1991 *The Politics of Street Crime*: *Criminal Process and Cultural Obsession*. Temple University Press, Philadelphia
Sechrest D, Sichor D (eds.) 1996 *'Three Strikes' and Public Policy*. Sage, London
Simon J 1993 *Poor Discipline: Parole and the Social Control of the Urban Underclass, 1890–1990*. University of Chicago Press, Chicago
Simon J 1997 Governing through crime. In: Friedman L M, Fisher G (eds.) *The Crime Conundrum: Essays on Criminal Justice*. Westview, Boulder, CO
Spierenburg P 1984 *The Spectacle of Suffering*: *Executions and the Evolution of Repression from a Pre-Industrial Metropolis to the European Experience*. Cambridge University Press, Cambridge, UK
Stem V 1998 *A Sin Against the Future*: *Imprisonment in the World*. Northeastern University Press, Boston
Thompson E P 1975 *Whigs and Hunters: The Origin of the Black Act*. Pantheon, New York
Tonry M H 1995 *Malign Neglect—Race, Crime, and Punishment in America*. Oxford University Press, New York
Von Hirsch 1976 *Doing Justice: The Choice of Punishments*. Hill and Wang, New York
Zimring F E, Hawkins G 1986 *Capital Punishment and the American Agenda*. Cambridge University Press, New York

J. Simon

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Q

Qualitative Methods, History of

This brief history traces the development of sociohistorical methodologies employed in relation to data that are either archival or based on direct social observation and interaction, insofar as these methods do not depend on quantitative analysis. The history of qualitative methods can be explored in three respects: (a) institutionalization; (b) how objects of inquiry have been construed; and (c) logics and analytic strategies. Recently, debates in the wake of the 'postmodern' turn have created new possibilities for understanding qualitative research, and research more broadly, as a terrain of inter-discursive communication with properties of 'integrated disparity.'

1. Institutional Emergences of Qualitative Methods

Sociohistorical inquiry emerged long before the rise of the modern academic disciplines. In China, statecraft early on was informed by typologies of social circumstances and transitional possibilities of action specified by the *I Ching*. In the fifth century before the modern era, ancient Greece had Herodotus, as much an observer of the contemporary as a historian. And during the fourteenth century, in the Arab world, Ibn Khaldun wrote in a way that anticipated comparative and historical sociology. These and other relatively sophisticated practices of inquiry coexisted with much more basic efforts of social groups to document their existences through the development of archives, king lists, annals, and chronicles. Thus, the arrangements through which sociohistorical knowledge has been produced are longstanding and diverse.

However, contemporary qualitative inquiry is especially a product of modernity, and of critical and postmodern reactions to the research practices spawned of modernity. Initially, qualitative research was inextricably linked with the gradual emergence of modern national societies and their colonial expansion, most concertedly during the nineteenth and twentieth centuries (Vermeulan and Roldán 1995). Intellectually, the rise of social and humanistic disciplines can be traced to the European Enlightenment of the eighteenth century. Thinkers such as Giambattista Vico and Johnann Gottfried Herder increasingly began to regard social and historical processes as subject to understanding in their own terms rather than products of the will of God alone, and Edward Gibbon developed analytic history by use of concrete social explanations (i.e., in his *Decline and Fall of the Roman Empire*).

In the nineteenth century, history developed under two aligned tendencies: efforts to differentiate the professional discipline from popular historical writing, and attempts to harness sociohistorical inquiry to the 'liberal' project of constituting the State, citizens, and the economy as a viable assemblage for social development. Thus, Leopold von Ranke sought to establish accounts of 'what really happened' on the basis of 'scientific' validation of facts drawn from archival sources, while simultaneously positing the history of political and religious elites as the subject matter that would reveal the meaning of history as a whole, that is, as a manifestation of the will of God. Historians in the United States for the most part embraced a modernist quest for 'objectivity' inspired by Ranke (Novick 1988), while leaving the ideological underpinnings of Ranke's scientific historicism largely uninspected. By contrast, historical sociology developed through the middle of the twentieth century mostly in relation to grand evolutionary social theories, at points aligned with anthropological ethnology, which was focused on constructing racialized world evolutionary sequences of social-group development. Only after World War II, and especially in the 1960s, did historical sociology develop a methodologically sophisticated array of comparative and theoretical strategies for the analysis of history (Barnes 1948, Smith 1991).

As for field research, it emerged in relation to modern colonial expansion and capitalist industrial development. Among Europeans, especially from the sixteenth century onward, knowledge gained from travel became a basis of colonial power sustained by exploratory expeditions, practical sciences, missionary work, and territorial penetration of colonial civil services. Sometimes authors combined traveling for a trading company or state with writing for a popular audience that wished to visit other lands vicariously. But epistemological principles of social knowledge were also at stake. Samuel Johnson's latter eighteenth-century trip to western Scotland implicitly demonstrated that cultural differences would require forms of knowledge different from the then emerging rationalized social quantification (Poovey 1998).

'Exploration' journals became increasingly systematic not only in describing the flora and fauna of the 'discovered' world, but also in inventorying the lan-

guages and practices of its peoples. Organizations such as the Smithsonian Institute (founded in 1846) coordinated efforts to gather data and circulated inventory questionnaires that could be filled out by non-professional travelers. Yet travelers' accounts—often saturated with racist and ethnocentric assumptions—largely served as grist for evolutionary ethnology and philosophical ('armchair') anthropology, for example, in Emile Durkheim's work on religion and James Frazer's *The Golden Bough*.

By the late nineteenth century, ethnology and philosophical anthropology still dominated the disciplinary space where modern ethnography was beginning to emerge. Lewis Henry Morgan's field study of the Iroquois, published in 1851, was largely a springboard to his ethnology. Only with the work of Franz Boas and Bronislas Malinowski did inquiry begin to abandon the ethnological search for trans-group anthropological principles of development, and emphasize analysis of social groups in their contemporaneous existences (a shift much bemoaned by the theorists). Boas is a crucial transitional figure. His late nineteenth-century observation and interviewing of indigenous peoples such as the Kwakiutl was initially intended to advance his studies of perception and speech, to the point that he sometimes became impatient with ceremonial potlatches that wasted research time! Nevertheless, Boas pioneered efforts to create valid information based on direct, systematic field research, and he trained students such as Edward Sapir and Margaret Mead.

During and after World War I, Malinowski even more strongly embraced sustained participant observation oriented toward characterizing a group's present way of life. His Trobriand Islands fieldwork, published as *Argonauts of the Western Pacific*, proposed a methodology borrowed from sociology, placing empirical observations within a functionalist framework, yet it also consolidated the modern ethnographic role as expert offering an authoritative account aimed at an emic understanding of the 'other's' point of view. This tension—between an objective theoretical framework and emic analysis—became a hallmark of modern anthropology, for example, in the French Durkheimian tradition, carried forward by Claude Lévi-Strauss (and in Britain by Alfred R. Radcliffe-Brown), and even in the way Clifford Geertz sometimes combined 'thick description' with social-systems rhetoric.

Ethnographic study of the exotic 'other' was hardly the only impetus to institutionalization of field research. Famously, Alexis de Tocqueville voyaged to the United States in the 1830s to learn more about the American prison system. During the late nineteenth and early twentieth centuries, sociology was influenced by similar practical quests for knowledge about social conditions, for example, in studies of social issues sponsored by the *Verein für Sozialpolitik* in Germany, by socialists in England, and among social workers such as those at Chicago's Hull House in the United States.

Building on the latter tradition, during the early twentieth century, sociology's strongest institutional development of qualitative methodology came at the 'Chicago school' founded by Robert Park, who took inspiration from Boas and Malinowski to encourage field research. The city of Chicago became an urban laboratory for the direct observational study of social phenomena. Chicago faculty and students produced a stream of monographs such as *The Hobo* (1923) and *The Gold Coast and the Slum* (1929), and their approach found common cause with observational and in-depth interviewing field studies of communities, such as *Middletown* (1929) and the *Yankee City* studies (1941–59), work itself influenced by anthropological ethnographies (and sometimes carried out by anthropologists). In the 1950s, Herbert Hyman formalized interviewing as a methodology, and Chicago students such as Howard S. Becker, Jr., continued the pragmatic and symbolic interactionist Chicago tradition that had been initiated by George Herbert Mead and developed by Everett Hughes. During the 1960s, phenomenological sociology, inspired by Alfred Schutz, and ethnomethodology, promoted by Harold Garfinkle, emerged as alternative approaches that problematized objectivity, realism, and the social activities through which qualitative knowledge is created. Added to this mix during the same decade were the constructivism of Peter Berger and Thomas Luckmann, the dramaturgical approach of Erving Goffman (himself trained at Chicago), and a nascent bridge between field research and history via the emerging interest in oral history. These developments, together with the 1960s retooling of historical sociology, set the stage for an efflorescence of qualitative research in the last decades of the twentieth century.

Yet, just when qualitative research flourished, it also was beset by differentiation and methodological crisis. Some symbolic interactionists, ethnomethodologists, and comparative and historical sociologists sought to affirm the scientific legitimacy of their enterprises by formalizing research methods and strategies, while other researchers took seriously the rejection of scientism consolidated from the 1960s onwards by critical theorists such as Jürgen Habermas, hermeneuticists such as Paul Ricoeur and Charles Taylor, philosophers of history (most notably Hayden White), feminist epistemologists such as Dorothy Smith and Sandra Harding, anthropologists like Marshall Sahlins and James Clifford, and the wave of poststructuralist and postmodern 'turns' heralded by Michel Foucault and Jacques Derrida. Ethical considerations about privacy, asymmetry of field research, and exploitation of subjects—raised especially by feminists and within anthropology—cast long shadows on the high modernist quest for objectivity and validity in qualitative analysis. During the last two decades of the century, these developments, along with

the rise of world systems and globalization theories, led to the formulation of new hybrid qualitative methodologies that linked the local and the global, ethnographic and historical analysis, by such researchers as Michael Burawoy and Philip McMichael. The old dream of the ethnologists, of theorizing case studies in relation to general social processes, was reborn, but in a dispensation that eschewed ethnological ethnocentrism.

2. *The Object of Inquiry, a Brief History*

One enduring problem of qualitative research concerns how to construe the object of inquiry, that is, what is studied. At the end of the eighteenth century, Immanuel Kant raised the philosophical problem of selection, namely, which of myriad events revealed the 'guiding thread' of history. In the nineteenth century, under the influence of teleologies such as those of Hegel and Ranke, historians were generally content to regard their studies as centered on 'natural' objects that had their own coherences in the world itself. However, as Hayden White (1973) has shown, the nineteenth-century European historians established rhetorical coherence through one or another genre of emplotment that offered a metanarratological structure into which to fit events. Outside of history (and increasingly within it), an alternative tendency developed the object of analysis via theory. Thus, Karl Marx and Frederick Engels theorized about capitalism as a dynamic totality changing through its own contradictions over time, and used their perspective to thematize social events and processes in places as diverse as India, the United States, and Manchester, England. Within academia, the Marxian approach found its counterpart in the program of German historian Karl Lamprecht, who sought to displace Ranke's elites as the object of historical analysis with the study of diverse social groups and their interaction.

The Kantian problem of the relation between object and selection was more directly addressed during the German *Methodenstreit* toward the end of the nineteenth century, especially through the effort of Wilhelm Rickert to posit the objective value of truth in a way that would yield a shared way of looking at the world, and by extension, a shared object. Among mainstream historians during the twentieth century, a sort of ontological historicism emerged, in which shifts in styles (for example, the emergence of social history) depended on shifting the object of historical analysis, rather than its methods. The major alternative to historicism, which Max Weber spelled out at the beginning of the twentieth century, depended on two central assertions:

(a) first, that because values could not be warranted scientifically, the objects of sociohistorical inquiry could not be scientifically determined, even though, once determined, scientific methods could be applied to their study; and

(b) that the objects of sociohistorical inquiry are not natural kinds, but rather, infinitely varying sociohistorical phenomena that become objects of inquiry through cultural interest in them.

Alternatives to Weber's approach can be traced in history to the Annales School and Marxism, and in the social sciences, methodologically to the objectivism of Emile Durkheim, and analytically to Georg Simmel. Durkheim had provided the general warrant for objectivist approaches in his 1895 *Rules of Sociological Method*. And, in a way that is not widely recognized, Durkheim's structuralism had its equivalent in Simmel's early twentieth-century identification of social 'forms' (such as 'sociability'), a central inspiration for a number of qualitative symbolic interactionists (Hall 1999, Chap. 4). Within structural Marxism, the perspective of the proletariat was assumed to offer an objective standpoint for the analysis of capitalism, though during the first half of the twentieth century this position became increasingly problematic at the hands of Georg Lukács and critical theorists Max Horkheimer and Theodor Adorno. If the object of inquiry for structuralist Marxism was a totality, for the Annales school, in the formulation of Fernand Braudel, the totality was the grid of objective temporality, onto which multiple levels of change—ecological, structural, *histoire événementielle*—could all be mapped (Hall 1980, Burke 1990).

Ethnographic field researchers during the first half of the twentieth century generally treated their cases as naturally occurring 'primitive' societies subject to 'salvage anthropology.' Yet, the supposed isolation of the natural case became increasingly problematic with the post-World War II emergence of first modernization and then world-system theory approaches, both of which drew into question the autonomy of indigenous social groups. A parallel tension developed within sociological domains of field research. At the moment of high modernism, symbolic interactionists like Herbert Blumer would posit the empirical world 'out there' as the object of 'naturalistic research.' Yet, interactionists shared with phenomenologically oriented researchers and ethnomethodologists the sense that the empirical object of research was 'constructed,' i.e., that the meaningful world was constituted by its participants' practices.

By the beginning of the twenty-first century, the objectivist-relativist divide in constituting objects of qualitative inquiry persisted, but the cultural and Foucaldian turns in history, the linguistic turn in anthropology, and increased understanding of how ethical choices shape data from field research all created conditions in which answers to the question 'what is a case?' (Ragin and Becker 1992) had become considerably more nuanced and sophisticated. In particular, by 1978, Arthur Stinchcombe argued against comparing cases as total entities one against

another, and in favor of searching out the deeply analogous social processes that might be at work in cases that, on the face of it, might not even be considered as members of the same set. Overall, in today's post-positivist climate, reflexivity rules. Even objectivists recognize that theoretical orientations and concepts have consequences for how cases are construed, while dialogic relativists, some of whom doubt the existence of any straightforwardly accessible empirical world to be studied, have had to come to grips with the problem of how adequately to 'represent' the (constructed) realities of their encounters with the people about whom they write.

3. *Developments in Logics of Qualitative Analysis*

Philosophers have debated the logics of human reasoning for millennia. But the most compelling specification directly relevant to modern sociohistorical inquiry came in John Stuart Mill's (1843/1950) philosophy of scientific method, which specified two methods of non-experimental research. The method of agreement entailed isolating crucial similarities in a set of diverse cases, yielding the claim that if two aspects co-occur under otherwise diverse circumstances, they are likely to be somehow connected with each other. On the other hand, the indirect method of difference extended the logic of control-group experiments to nonexperimental settings, by searching for contingent co-related variations. In effect, two (or more) findings derived from parallel uses of the method of agreement could be placed side by side. If a particular propensity were found only in the presence of another particular propensity, and never when the second propensity was absent, Mill reasoned that one was 'the effect, or the cause, or an indispensable part of the cause' of the other.

The application of Mill's logic to sociohistorical inquiry met with opposing visions of qualitative analysis that became crystallized by issues raised during the late nineteenth-century *Methodenstreit*. Whereas the logic of Mill yielded a nomothetic search for general principles, an alternative, idiographic view suggested that the study of human affairs necessarily dictates emphasis on the individually distinctive, indeed unique, aspects of social phenomena. Whatever the resolution to this question, a second issue concerned whether the special character of sociohistorical phenomena made them necessarily the subjects of a different kind of knowledge than the explanatory search for universal causal laws, namely, a cultural science based on understanding social meanings and actions. Three broad logics of qualitative methodology—interpretive strategies, ideal-type analysis, and formal analysis—can be traced in relation to these contestations.

First, as the twentieth century dawned, Wilhelm Dilthey proposed a historicist hermeneutics that would seek to understand history as embedded in the life meanings of the people who participate in it. This basic approach has its parallels in both ethnography and sociological field research. Elaborations of analytic logics were especially concentrated in sociology. Alfred R. Lindesmith (1947) developed the method of 'analytic induction' as a procedure for successively refining hypotheses in relation to cycles of field research, an approach that Glaser and Strauss (1967) subsequently reinterpreted as 'grounded theory.' By the 1960s, 'participant observation' (Platt 1983) and in-depth interviewing became increasingly institutionalized as legitimate approaches to inquiry, with their own epistemological rationales, challenges (notably how to construe and know subjectivity of the other person), research procedures, and handbooks, such as those of Buford Junker, Severyn Bruyn, and John and Lyn Lofland.

Second, working at the same time as Dilthey, Max Weber consolidated a *Verstehende* approach that was both comparative and interpretive. He did not deny the possibility that sociohistorical inquiry might discover general regularities of the sort Mill sought to identify, but he regarded individualizing knowledge as the central value interest of sociohistorical inquiry. Therefore, he combined explanation and understanding, by using generalized ideal types that would be 'adequate on the level of meaning' and thus serve as meaningfully coherent analogs to the patterned character of empirical cases, yet nevertheless allow for comparison beyond individual particularities. Weber's methodology became central to the renaissance of historical sociology consolidated in the 1960s (Bendix and Roth 1971). In the 1990s, Geoffrey Hawthorn and David William Cohen, among others, developed Weber's logic of the 'mental experiment' into 'counterfactual analysis.'

Third, Mill's methodological effort became the touchstone for subsequent formalizations of qualitative logic. The most important elaboration has been the Qualitative Comparative Analysis pioneered by Charles Ragin (2000), which builds on Mill's principles by using set theory and Boolean algebra to identify multiple configurational patterns within an array of systematically described cases. In a different development, with the late twentieth-century turn toward the study of discourse (partly inspired both by ethnomethodology and post-structuralism), scholars such as Larry Griffin theorized ways in which narrative accounts might become subjected to formal analysis that would yield descriptions of discrete social processes. These and related developments of formalization—as well as techniques of 'triangulation' that compared findings gained from diverse methodologies—increasingly blurred the lines dividing qualitative from quantitative analysis. Yet with logical development also came critique, such as that of Stanley Lieberson (in Ragin and Becker 1992), who questioned whether formalization alone could resolve logical

difficulties of comparative studies with small numbers of cases.

4. Postmodern Reflexivity in the Shared Domain of Qualitative Methods

By the end of the twentieth century, there had been substantial differentiation and development of qualitative methods, as well as experimentation in both research role relationships in field research and the form and rhetoric of texts reporting qualitative research more generally. Yet, these methodological shifts typically transpired within relatively autonomous schools where one or another approach was alternatively embraced, reinvented under a new flag, or syncretically revised in relation to alternatives. The differences between inductive and deductive historical sociologists, or between mainstream symbolic interactionist field researchers and those who had taken the postmodern turn, often loomed as important as the disjuncture between quantitative and qualitative research more generally.

Under these conditions, a fruitful development has been the specification of typologies that describe alternative practices of inquiry in terms of the shared or contrasting cultural logics of their research methods, rather than assuming an equivalence between particular qualitative methodologies and various schools or genres (e.g., symbolic interactionism, cultural history). Theda Skocpol and Margaret Somers, and Charles Tilly developed typologies of historical and comparative sociology, and John R. Hall (1999) used discursive analysis to elaborate a general typology describing eight core cultural logics of research practice. In these terms, the three methodologies most favored in field research and historical sociology are historicism, contrast-oriented comparison, and analytic generalization. These methodologies are part of a wider array of possibilities, including practices such as 'configurational history' that are often employed without benefit of formalization, and other practices, such as 'universal history,' that persist despite postmodern condemnation of 'metanarratives.' Typologies of qualitative analysis might seem to harden the boundaries between alternatives. However, the outcome of discursive analysis is the opposite, for it reveals a condition of 'integrated disparity' among research practices that share cultural logics in ways not previously understood, and it clarifies the conditions and potential for translation of research knowledge produced through radically alternative research practices (Hall 1999, Chap. 9). Thus, although qualitative research is shot through with contemporary tensions between science, relativism, and constructivism, qualitative researchers may find ripe conditions for a new era of sophistication about the character of qualitative research that will have implications for research practices more widely. But to consider those possibilities moves beyond history into the realm of as-yet contingent possibility.

See also: Analytic Induction; Ethnomethodology: General; Field Observational Research in Anthropology and Sociology; Grounded Theory: Methodology and Theory Construction; Health Research, Qualitative; Interactionism: Symbolic; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Qualitative Methods in Geography; Science and Technology Studies: Ethnomethodology; Symbolic Interaction: Methodology

Bibliography

Barnes H E 1948 *Historical Sociology*. Philosophical Library, New York

Bendix R, Roth G 1971 *Scholarship and Partisanship: Essays on Max Weber*. University of California Press, Berkeley, CA

Burawoy M et al. 1992 *Ethnography Unbound*. University of California Press, Berkeley, CA

Burke P 1990 *The French Historical Revolution: The Annales School, 1929–89*. Polity, Cambridge, UK

Glaser B G, Strauss A L 1967 *The Discovery of Grounded Theory*. Aldine, Chicago, IL

Hall J R 1980 The time of history and the history of times. *History and Theory* **19**: 113–31

Hall J R 1999 *Cultures of Inquiry: From epistemology to discourse in sociohistorical research*. Cambridge University Press, Cambridge, UK

Lindesmith A R 1947 *Opiate Addiction*. Principia Press, Bloomington, IN

Mill J S 1843 1950 *John Stuart Mill's Philosophy of Scientific Method*. Hafner, New York

Novick P 1988 *That Noble Dream: The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK

Platt J 1983 The development of the 'participant observation' method in sociology: origin myth and history. *Journal of the History of the Behavioral Sciences* **19**: 379–93

Poovey M 1998 *A History of the Modern Fact*. University of Chicago Press, Chicago, IL

Ragin C C 2000 *Fuzzy-Set Social Science*. University of Chicago Press, Chicago

Ragin C C, Becker H S (eds.) 1992 *What is a Case?* Cambridge University Press, Cambridge, UK

Smith D 1991 *The Rise of Historical Sociology*. Temple University Press, Philadelphia, PA

Stinchcombe A L 1978 *Theoretical Methods in Social History*. Academic Press, New York

Vermeulan H F, Roldán A A (eds.) 1995 *Fieldwork and Footnotes: Studies in the History of European Anthropology*. Routledge, London

White H 1973 *Metahistory: The Historical Imagination in Nineteenth-century Europe*. Johns Hopkins University Press, Baltimore, MD

J. R. Hall

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Qualitative Methods in Geography

Contemporary human geography is characterized by a diversity of topic areas, theoretical questions, and methodological plurality. A central concern with spatiality is constant, but a diversity of geographical imaginations (Gregory 1994) suggests a growing number of approaches appropriate to investigating how 'geography matters' in how social, economic, and political life unfolds. Among these, qualitative methodology provides an alternative paradigm of knowledge creation to that of quantitative techniques. With its admission of subjective experience and focus on meaning, qualitative methodology provides a variety of ways through which to analyze and interpret landscape, region, location, space, place, and the complex intersections of social and spatial relations. It has been critical to the re-imagining of the discipline as contemporary philosophical and theoretical debates have been integrated into empirical inquiry.

1. *A Paradigm of Inquiry and Collection of Methods*

Rather than accepting the possibility of a unitary, objective explanation of observed phenomena, qualitative research approaches acknowledge and seek the multiple and partial versions of reality that are constituted in the course of social life. Embracing a variety of approaches and field methods, a common interest in qualitative research is to enhance understanding through description and concepts grounded in the perspective of those studied. The meanings brought to bear by individuals and groups in ordering and making sense of their everyday social life and its spatiality are privileged in analysis. Methods of inquiry include various ethnographic approaches, including participation in and observation of the worlds of those studied and nonstandardized forms of interviewing with individuals or groups; text and visual analysis; and a growing repertoire of other nonnumerical ways of observing, describing, and interpreting the world.

The shifting understandings of relationships between people and places and society and space are important in circumscribing how specific qualitative research strategies are employed. Interest in epistemological issues, particularly through the critical arm of the discipline, has fostered debate about the production of geographical knowledge and further shaped the choice of methods. There is a close association between theoretical developments, philosophies underlying inquiry, societal change, and the increasing use of qualitative methods in geography. As geographers seek to understand complex socio–spatial relations from different subdisciplinary and theoretical perspectives, such methods are used in various ways according to the questions and purpose of research. They can provide descriptive data in mixed strategy designs to enhance findings from quantitative data, but more commonly are associated with postpositivist analyses that center the interpretation of meanings in explanation of narrative and visual data. Inductive analyses of narrative data, informed by concepts of theory, focus on generating concepts grounded in categories emerging from the data. Interpretation of various types of visual data, such as paintings, film, photographs, maps, and historical documents, similarly are sensitized by the theoretical and philosophical lenses of the 'reader.'

2. *Qualitative Methodology in a Context of Philosophical and Theoretical Turns*

Since the challenge of interpretative understanding to the dominance of quantitative methods in the 1960s, intellectual turns have brought new questions and new topics to the agenda of geographical inquiry, which have required a rethinking of appropriate methods. Geographers from a number of subdisciplines have explored qualitative approaches in their research as the influences of, first, humanism and then postmodernism, feminist scholarship, and cultural studies have problematized conventional methods of inquiry and categories of analysis.

Humanistic geography, with its philosophical inspirations from idealism, phenomenology, and existentialism, first challenged the strong quantitative turn of the 1960s (Ley and Samuels 1978). Interest in human agency, the inclusion of people's intentionality, a focus on lived experience, and the quest of understanding how meanings were produced in the context of interacting social and geographical worlds, marked the value of qualitative research strategies. Participant observation studies produced ethnographic accounts that permitted a linking of the social with the geographical in ways that opened up inquiry to numerous topic areas that privileged the perspective of the human subject. Interpretation of landscapes included concerns with sense of place, while case study approaches informed by a theoretical framework of social constructionism explored the embeddedness of human action and perceptions in place and provided models of inquiry for the 1980s. The qualitative methods text of Eyles and Smith (1988) signaled the establishment of qualitative research as an appropriate and useful paradigm of knowledge production within the discipline of geography.

Paralleling these developments, the influence of feminist scholarship resulted in further embracing of qualitative strategies in research. Noting the absence of women in geographical representations of the world, feminist geographers set out to include the neglected 'half' of the population in the mappings and

narratives of the discipline. In making the lives and geographies of women visible, qualitative methods, as in other disciplines, were privileged in their ability to uncover the everyday lives of women and the spatialities of their domestic and paid work (McDowell 1992). Bringing subjective realities of women into geographical explanation exposed the masculinist partiality of geographical inquiry and permitted the construction of concepts reflecting the geographies of women's lives. Linked with principles of research with an emancipatory aim, qualitative approaches remain important in feminist methodology as a vehicle, together with the contribution of quantitative methods, in exploring the spatiality of gendered social relations (Jones et al. 1997).

Postmodernism, the ongoing development of feminist theorizing and the later influence of cultural studies brought a 'cultural turn' to human geography (Massey et al. 1999). The importance of discourse in constructing human subjectivity and representing difference, whether reflected in the organization of space, social life, or in visual representations, was an important aspect of a complex literature on culture that influenced methodological approaches. Throughout human geography qualitative methods enabled inquiry to reveal the lifeworlds of marginalized 'others,' to engage critically with various media as cultural representations, to expose the multiple, partial meanings of landscapes and texts, and to disrupt many taken-for-granted categories of analysis. Immigrants, the ill, the disabled, the colonized, and other marginalized groups emerge as others whose subjectivities and geographies are being included in a broadening geographical knowledge. Ethnographies of workplaces and interviews with elites, analysis of contemporary visual representations of social diversity, and the inclusion of historical texts that include women's experiential accounts are examples that further reflect a concern with competing knowledges and representations of the world. These journeys into the lives of others has led to expansion and innovation in field techniques, including the use of life stories, focus groups, autobiographical writing, photography, film, and the adaptation of interviewing techniques and formats to meet the needs of particular research projects.

3. *Constructing Knowledge Through Qualitative Research: Methodological Issues*

With the increased use of qualitative research strategies, geographers have engaged with a methodological literature in the social sciences discussing the difficulties that emerge when conducting research with the 'other' of studies. Reflection centers on the complexity of constructing knowledge through reliance on the intersubjective relationship between researcher and researched and the interpretation of data through the self as a socially positioned researcher who is part of the world they study. Issues of rigor in research design and practice have been addressed with attempts made to translate the concepts of quantitative research into those more amenable to qualitative methodology, but most discussion has focused on the close association between practical and theoretical dilemmas in research, and their location in epistemological concerns.

The particular attentiveness of feminist scholars to the micropolitics of research encounters, together with the acknowledgment of the situatedness and partiality of all experience and knowledge, have become the concerns of many qualitative researchers in the discipline. Gaining access to previously subjugated knowledges is shown to be a process located within a layering of power-laden social relations that situate the face-to-face research encounter within the wider politics and processes of 'the field,' theoretical and philosophical tensions, and the institutional values and processes of universities and funding bodies. Specificity of place is shown to mediate how this intricate layering plays-out in the microdynamics of particular fieldwork situations.

Particular issues vary in different types of inquiry, but the exercise of reflexivity by researchers has problematized the whole enterprise of knowledge production. Whether in the context of reading landscape as text, rewriting colonial geographies, the demedicalization of medical geography, the shifting concerns of feminist inquiry, interviewing elites in economic geography, or exposing processes of exclusion of marginalized groups in society, researchers consider the different ways their own identities and contextuality influence their analyses.

That the findings of research are theory laden and interpreted through the researcher's subjectivity has brought attention to the presentation and writing up of research. The representation of the multiple voices of the researched is seen as a politically charged and theoretically tricky enterprise. Both in relation to narrative and visual data interpretation, the text produced and its claims to knowledge are problematized. The authority of the author is challenged, in that a particular representation is one of several possibilities. Following from this, the reading and consumption of research writing are viewed as crucial dimensions of the research process and the development of knowledge, rather than separate activity.

Attempts to disrupt hierarchical research relations, including issues concerning the politics and ethics of qualitative research have moved increasingly to center stage in methodological debate. For whom research is conducted, how findings are used and related to theory production, and concern over the possible exploitative nature of the relationship with research participants have become important considerations in research conduct. Participatory and emancipatory models of research are being explored, accompanied by thinking

through how academia and activism are and may be connected. The ethical issues involved in such attempts to 'empower,' whether research is conducted at home, or in developing countries, have become an important focus of discussion.

4. Continuing Contributions

Expansion of the use of qualitative methods in geography is reflected in special issues of geographical journals and new texts on research methods that focus on qualitative methods or include qualitative approaches as important, rather than subordinate research strategies. Their value is recognized as theoretical developments have signaled the need to find new ways to discover and analyze data appropriate to current topics and questions. The various methods of qualitative research have opened up different ways of 'reading' and 'seeing' landscape, space, place, and the relationships between people and environments, so continually reshaping the nature and scope of geographical inquiry. The ongoing theorization of space, that has emphasized its relational, social, and recursive dimensions, has been accompanied by qualitative studies that are inclusive of human experience, perceptions, and meaning. Such studies have brought into view the diversity of local, everyday geographies, while at the same time locating individual and group experience within complexly linked local, national, and global relations.

The use of qualitative and quantitative methods is sometimes presented in terms of competing paradigms, sitting uncomfortably with each other. Theoretical turns and epistemological concerns may suggest that one be privileged over the other in the quest for valid knowledge in geographical inquiry, but different questions require different approaches. Both quantitative and qualitative work can and do contribute to furthering the understanding of spaces, places, locations, and how variously positioned individuals and groups live within and experience them. Qualitative geographical inquiry, however, has exposed the complexity of gaining access to and representing the situated knowledges of research participants, that precludes the development of rigid protocols or 'ways of doing' qualitative research. Different ways of telling, listening and knowing require flexibility in research design and approaches. Ways of doing qualitative research need to be situated within and be sensitive to the many situations that comprise the academic geographer's field. It may be anticipated that further innovation of methods and flexible models will remain the hallmark of research concerned with the interconnections of people, place, and space.

See also: Humanistic Geography; Phenomenology: Philosophical Aspects; Qualitative Methods, History of; Reflexivity: Method and Evidence

Bibliography

Area 1999 **30** (whole issue)
Barnes T J, Duncan J S 1992 *Writing Worlds: Discourse, Text and the Representation of Landscape*. Routledge, London, UK
Environment and Planning D: Society and space 1992 **10**(5) (whole issue)
Eyles J, Smith D M 1988 *Qualitative Methods in Human Geography*. Blackwell, Oxford, UK
Geoforum 1999 **30**(4) (whole issue)
Gregory D 1994 *Geographical Imaginations*. Blackwell, Cambridge, Oxford, UK
Hay I 2000 *Qualitative Research Methods in Geography*. Oxford University Press, Melbourne, Vic
Jones III J P, Nast H J, Roberts S M (eds.) 1997 *Thresholds in Feminist Geography: Difference, Methodology, Representation*. Rowman and Littlefield, Lanham, MD
Ley D, Samuels M S 1978 *Humanistic Geography*. Maaroufa Press, Chicago
Massey D, Allen J, Sarre P (eds.) 1999 *Human Geography Today*. Polity Press, Cambridge, UK
McDowell L 1992 Doing gender: feminism, feminists and research methods in human geography. *Transactions of the Institute of British Geographers* **17**: 399–416
The Professional Geographer 1994 **46**(1) (whole issue)
The Professional Geographer 1999 **51**(2) (whole issue)

I. Dyck

Quality Control, Statistical: Methods

1. Introduction

Modern quality improvement by statistical methods began in the 1920s through efforts of W. A. Shewhart (see Shewhart 1931) of the Bell Telephone Laboratories. Rather than wait until manufactured items arrived at an inspection center at the end of the production line, he conceived of the idea of occasionally sampling during the manufacturing process and making corrections immediately, if needed. To monitor the need for possible adjustments, certain characteristics—called statistics—of each sample were recorded. If any of these statistics had unusual values that indicated something was wrong with the process, corrective action was taken before too many more defective items were produced. Accordingly, the quality of the final output was improved.

While Shewhart started this modern quality improvement by trying to reduce the number of defects in manufacturing, many organizations soon recognized that there was much waste throughout their operations. There were simply many activities that added nothing to satisfactory outcomes. This is not only true in manufacturing but in most service areas as well, including hospitals, banks, and educational institutions. Estimates of how much waste and worthless activities exist range from 25 to 50 percent of the total effort, although there are some organizations that

probably exceed that bound of 50 percent. In any case, most industries and businesses could make huge improvements in quality products and services if they only recognized how poorly they are doing. Often such recognition is found by taking measurements, collecting data, and turning these data into information through statistics.

There are many philosophies and techniques about improving quality of products and services. Some of these have been formalized by considering the ISO 9000 or Baldrige awards (see Ledolter and Burrill 1999). Many companies have earned these awards and have had continued success; but some have then gone on to fail because they have not had a continuous commitment by top management to maintaining quality products and services. Bob Galvin, former CEO of Motorola, truly believed that 'quality improvement is a daily, personal, priority obligation,' and his commitment made Motorola what it is today.

At the end of the twentieth century it is worth making note of one quality guru who almost lived that entire twentieth century. W. Edwards Deming was a statistician who was born October 14, 1900 and died December 20, 1993 and was giving short courses in quality improvement ten days before his death. Deming went to Japan after World War II and taught the Japanese how to make quality products. For his work there, he was awarded the Emperor's Medal, and the Japanese established the Deming Prize to be awarded each year to a company or individual contributing the most to quality improvement.

There are others too; certainly Joe Juran and Japan's Kaoru Ishikawa (see Ishikawa 1982) must be mentioned. However, it was Deming's approach to optimizing the entire system that was a major point. In addition, his thoughts on 'profound knowledge' are extremely valuable, particularly the important part of it about understanding variation and statistical theory.

Understanding variation is a key to quality improvement. Many of Deming's 14 points (Deming 1986), while not mentioning statistics, centered on reducing variation. Deming believed that barriers between departments, between management and workers, and among workers, must be broken down to improve communication and the ability to work as a team.

The lines of communication, all the way from suppliers to customers, must be open to help reduce variation and improve products and services. For example, he argued that a company should not only buy on price tag but have a few reliable suppliers (possibly one) for a single part because that will reduce variation. That is, many different suppliers would obviously increase variation. Moreover, he argued that you should become partners, friends if you like, with your suppliers. You learn to trust the other; in that way, you can use methods like 'just in time,' in which your inventory can be kept reasonably small, to keep costs down.

Deming also preached constancy of purpose. If management ideas tend to change too much, employees really do not know what to do and cannot do their best. That is, they are mixed up, increasing variation. It is better for them to receive a constant signal from their employer, a signal that only changes if research dictates ways of improving. More training and education for everyone associated with a company also decreases variation by teaching how to make the product more uniform. Workers must know that they should not be afraid to make suggestions to improve the process.

Being team members will make it easier for workers to speak up without fear of reprisal. Deming also noted that requiring quotas does not help quality. A foreman who has a quota of 100 per day will often ship out 90 good and 10 bad items just to make the quota. Clearly, it would reduce the variation and satisfy the customer better if only 90 good ones were shipped.

This leads to the point that a final inspection of products does not really improve the quality. With such a procedure, you can only eliminate the bad ones and send on the good ones. Improvements in the design of the products and manufacturing processes are needed. If these are done well, often with the help of statistical methods, that final inspection can be eliminated. That is, improvements should be made early in the process rather than try to correct things at an end inspection by weeding out the bad items.

Deming's points are still extremely important in the quality movement today. Understanding variation and reducing it still 'makes the doors fit better.' However, most of Deming's ideas also apply to service activities, ranging from market surveys to financial institutions (see Latzko 1986). Since so many more persons are employed in the service areas than in manufacturing, there are many more opportunities for quality improvement there. Why are some restaurants, clothing stores, law offices, banks, accounting firms, and hospitals better than others? Of course, those involved must know the technical aspects of their work, but the really excellent managers know that quality begins by treating their employees and customers right. The cost of a few dissatisfied persons is difficult to assess, but it is usually much greater than most think because these people will tell many others about their unpleasant experiences. Most of this quality improvement is common sense but it is not commonly used.

2. *Statistical Methods for Quality Improvement*

Let us begin our statistical methods by starting with one of the Shewhart charts, the mean chart. We sample from the output of a process. The frequency of sampling of the output of the process is really determined by the number of items being produced. If

there are a large number produced in a given day, the sampling might occur every hour. On the other hand, few in number might suggest sampling once per day is enough. Often these samples are small, and a typical size is n = 5, particularly when the measurement taken is something like length, weight, or strength. For example, in the manufacturing of paper, there is interest in its 'tear strength' because most often the manufacturer does not want paper to tear easily.

All of the underlying process have some mean μ and standard deviation σ. Hopefully the level μ is acceptable; and, in most cases, it is desirable to reduce the spread, as measured by σ, to make the output as consistent as possible. Hence, from each sample of $n = 5$ items, two statistics are computed: one to measure the level (middle) and one to measure the spread. We first concentrate on measure of the middle and that statistic is denoted by $\bar{x}$ and called the sample mean. Theory shows that once μ and σ of measurements of the underlying process are known (possibly estimated), the statistic will be between $\mu - 3\,\sigma/\sqrt{n}$ and $\mu + 3\,\sigma/\sqrt{n}$ unless there has been some change in the process producing these items. That is, if is not within the control limits $\mu \pm 3\ \sigma/\sqrt{n}$, then the process is investigated and usually some corrective action taken. The usual measures of spread are the sample range R and the sample standard deviation σ, and they have control charts similar to that of $\bar{x}$, and are called R- and σ- charts.

There are two other Shewhart control charts that must be described. The first of these is when we observe the number of flaws, say c, on a certain unit; for illustration, count the number of deaths on a certain highway each week. Here c can equal one of the nonnegative integers: 0, 1, 2, 3, 4, and so on. A Poisson probability model is used for the statistic c, and these probabilities can be determined from the mean μ of the number of flaws. Again μ is often estimated using past data. However, once it is determined, then, for the Poisson model, the standard deviation of c equals. Thus, $\mu \pm 3\sqrt{\mu}$ gives the control limits for the c-chart. Future values of a c are plotted sequentially on this chart, taking some action if one is outside the control limits: if above $\mu + 3\sqrt{\mu}$, find out why and hopefully learn how to improve the process; if below $\mu - 3\sqrt{\mu}$, again find out why and hopefully learn how to continue this gain in quality.

The last control chart arises in a situation in which the items are measured as being 'good' or 'bad.' If n of these items are observed, let D equal the number of defective items in the n items. The binomial probability model is appropriate for the distribution of D. Here n is much larger than 5, often equal to as much as 100. Of course, $\hat{p} = D/n$ is an estimate of p, the fraction of defective items being produced. The binomial theory states that $\hat{p} = D/n$ has mean p and standard deviation $\sqrt{p(1-p)/n}$.

Once some estimate of the fraction defective, p, is found from past data, the control limits $p \pm 3\sqrt{p(1-p)/n}$ are calculated for the p-chart. Future values of $\hat{p} = D/n$ are plotted on this control chart, and appropriate action is taken if one value is outside the control limits. Again, very small values of D/n would indicate some sort of improvement, and we would try to maintain that gain. When the statistics are between the control limits, which are usually estimated from the data, the process is said to be in statistical control.

The manufacturing team must make the decision whether or not the actual items from an 'in control' process are satisfactory. That is, they might find that the items are not satisfactory even though the process is in statistical control; then some adjustments must be made. However, once this is done, Shewhart control charts have proved to be extremely valuable in the past and have helped maintain the gain in the quality improvement of manufactured products. In addition, these control charts are now being used in many service areas, for example, studying errors made in recording the amounts on checks by banks.

Since most major manufacturing companies are requiring their suppliers to use statistical methods that help improve their products, the need for acceptance sampling to check upon the quality of incoming products has diminished. That is, partnerships with suppliers have been established, and hence the trust among the various parties is usually high.

Nevertheless, there are enough cases in which the incoming lot of items is poor; accordingly a large number of companies still use acceptance sampling. To describe acceptance sampling, suppose a lot has $N = 5{,}000$ items. Say that these materials are acceptable if two percent or less is defective.

This number, 2 percent = 0.02 is called the acceptable quality level. Understand that 0.02 is used only as an example, and some might insist that the incoming lot is only acceptable if there are almost zero defectives. In the latter case, this requires 100 percent inspection, which has some disadvantages, particularly if the item must be destroyed to inspect it, like 'blowing a fuse.' So most manufacturers are willing to accept lots with a small percentage of defects; needless to say, this percentage varies depending upon the importance of the item under consideration. To continue with the example, say a sample of $n = 200$ is taken from the lot of $N = 5000$, and it is decided to accept the lot if the number of defectives, D, in the sample is less than or equal to $Ac = 7$, the acceptance number.

Now clearly, as with any sampling acceptance plan, there are two types of errors that can be made. Say p is the fraction defective in the lot. If $p = 0.02$ and $D > 7$, then the lot is rejected when it should be accepted. This is an error of the 'first type.' The 'second type' error occurs if p is large, say $p = 0.06$ and $D = 7$, so that the lot is accepted when it should be rejected. It is desirable to have the probabilities of these errors as small as possible. Using a Poisson

probability approximation, it is found in this illustration that the probabilities of the two errors are $a = 0.051$ and $b = 0.090$, respectively. If these probabilities are not small enough, then the sample size n must be increased and the acceptance number Ac changed.

Typically, the manufacturer and the customer require certain specifications associated with the product. These are usually given in terms of a target value, and the lower specification limit, LSL, and the upper specification limit, USL. These 'specs' are set, often by the engineer or customer, so that if the product is within specs, it will work satisfactorily. One of the best ways to see if the products are within specs is to construct a histogram concerning the measurement of interest. If the product is satisfactory, the lower and upper values associated with the histogram should be within specifications. If several items fall outside of specs, then seemingly the process, without proper adjustments, is not capable of producing satisfactory products. If adequate improvements cannot be made with the current machinery, possibly 100 percent inspection must be used to eliminate the bad items, or else new machinery should be installed.

Clearly, economic characteristics of the situation are needed to make a final decision on the appropriate action. While a histogram seems to be a satisfactory way to measure the capability of the process, a large number of companies are resorting to various measures called capability indices. Possibly the most popular one is Cpk. Say μ and σ are the mean and the standard deviation, respectively, of the measurements associated with a process which is in statistical control.

The index Cpk is equal to distance of μ to the closest specification divided by 3σ. For example, suppose $\text{LSL} = 48.1$ and $\text{USL} = 49.9$ and $\mu = 49.1$ and $\sigma = 0.2$, then $\text{Cpk} = 0.8/3(0.2) = 1.33$, because $\text{USL} = 49.9$ is the closest spec to $\mu = 49.1$ and the distance between them is 0.8.

The reason that such a measure seems good is that if the underlying distribution is normal, then 99.73 percent of the items are within $\mu \pm 3\sigma$. If $\text{Cpk} > 1$, then the closest spec is more than 3σ away from μ. Most companies using Cpk require that the Cpk be greater than 1.33 or even 1.5, so that μ is 4σ or even 4.5σ away from the nearest spec.

However, there are some difficulties associated with the use of Cpk, or any of the capability indices. One is that it is assumed that the underlying distribution of measurements of an item is normally distributed, when often it is skewed with possibly a very long tail. But even worse, usually m and s have been estimated using some sample mean and standard deviation.

Knowing these are only estimates with some error structure, the computed Cpk using these estimates may be in error as much as 0.2 or 0.3, even with samples of sizes as large as $n = 50$ or 100. For illustration, the computed Cpk using estimates might be 1.19 while the real Cpk might be 1.39 or 0.99.

Frequently, the management might be very satisfied with a computed Cpk of 1.39 and not one of 1.28, if it is using the 1.33 cutoff; and yet there may not be any 'real' difference between the two values. Management should be aware of this possibility in making certain judgments and decisions.

Only some basic, but important, statistical methods for quality improvement have been touched upon. There are many more quantitative methods that can help, and these are considered in the references. These range from simple ideas like flowcharts, Pareto diagrams, and cause-and-effect diagrams, through CUSUMS and exponential weighted moving average charts to more complicated techniques in the design of experiments, including Taguchi's methods. Most of these techniques are explained in such books as those of Montgomery (1996) or Ledolter and Burrill (1999). Accordingly only a few remarks are made about the simple ones, and the references should be used for the more complicated statistical techniques.

A flowchart of the total process shows the steps that are needed to complete the process. Often by constructing a flowchart, the persons involved can see where certain steps can be eliminated or worthwhile shortcuts taken. A Pareto diagram is named for an Italian economist; it is essentially a bar graph in which the height of each bar represents the percentage of defects caused by a certain source. The possible sources are ordered so that the corresponding bars are in order of heights, with the tallest being on the left. Often one or two (possibly three) sources account for more than 85 percent of the defects; hence the Pareto diagram indicates clearly where the most effort should be placed to achieve improvement. The 'cause and effect' diagram is often referred to as an Ishikawa 'fish bone' diagram, because it looks like one. The main horizontal line points to the 'trouble', which is labeled to the right side of the horizontal line. Four to six diagonal lines leave the horizontal line, slanting to the 'northwest' or 'southwest' with labels at their ends that might be major sources of the trouble, like people, materials, machines, environment, etc. Then little 'bones' are attached to each of these with appropriate labels. It is amazing how completely these can be filled in with four or five knowledgeable persons 'brainstorming' for 30 minutes or so. All labels on the bones should then be considered for possible sources of the trouble, and often such a diagram calls attention to sources that might otherwise have been overlooked.

More advanced techniques like CUSUMS and exponential weighted moving average charts that can be used in place of or along side of Shewhart charts have proved most worthwhile. However, the designing of new products, or troubleshooting to find factors that are causing defects, are best served through statistical design of experiments. Some of these methods, in particular 2^k and 2^{k-p} designs, have been exceptional tools; they are considered in the book by Box et al. 1978. Overall, quality improvement has been greatly

enhanced by statistical theory and methods. The major contributors, including Shewhart, Deming, Juran, Ishikawa, Taguchi, and many more, deserve the appreciation of many in the quality movement. Clearly, from various experiences that have not been successful, much more remains to be done, and statisticians should play a major role in future efforts in improving the quality of products and services.

See also: Quality Control, Statistical: Reliability and Life Testing

Bibliography

Box G E P, Hunter W G, Hunter J S 1978 *Statistics for Experimenters*. Wiley, New York
Deming W E 1986 *Out of the Crisis*. Massachusetts Institute of Technology, Cambridge, MA
Hogg R V, Ledolter J 1992 *Applied Statistics for Engineers and Physical Scientists*. Macmillan, New York
Ishikawa K 1982 *Guide to Quality Control*. Asian Productivity Organization, Tokyo
Latzko W J 1986 *Quality and Productivity for Bankers and Financial Managers*. Marcel Dekker, New York
Ledolter J, Burrill C W 1999 *Statistical Quality Control*. Wiley, New York
Montgomery D C 1996 *Introduction to Statistical Quality Control*. Wiley, New York
Shewhart W A 1931 *Economic Control of Quality of Manufactured Product*. Nostrand, New York

R. V. Hogg

Quality Control, Statistical: Reliability and Life Testing

1. Reliability: What Does it Mean? Its Relation to Quality

The term reliability has become a part of our daily vocabulary. For example, the word unreliable describes the untrustworthy behavior of an individual, an item, or a procedure. In most situations the terms reliable and unreliable are conversational, and convey an intuitive feel. However, in dealing with situations involving modern machines and technologies, the need to be more scientific is on the increase. This has resulted in efforts to come up with sharper definitions of reliability, definitions that lend to quantification. But why must one quantify, and how should the quantification be done?

An answer to the first part of this question lies in the thesis that the notion of reliability is useful for certain kinds of decision making under uncertainty, and that the normative approach to decisions is based on quantifiable inputs (see Lindley 1985). Thus, in a broad sense, quantified measures of reliability are inputs to normative decision making. The answer to the second part of the question depends on how reliability is defined, and this depends on how it is to be used. Thus, in psychological testing, reliability reflects consistency of results provided by a test instrument, and so quantification is based on a notion of consistency. By contrast, reliability used in the engineering sciences pertains to an attribute of functional quality, and this indeed is its connection with quality. Functional quality is an ability to satisfy a set of requirements, and reliability comes into play when the requirement is either survival, or an ability to perform. It is in this context that reliability is of interest. Here, the quantification of reliability becomes a subtle issue, because the approach used to quantify leads one to a philosophical debate, a debate that has plagued generations of thinkers. More about this will be said later. For now it suffices to say that reliability is a quantified measure of uncertainty about a unit's survival for a prescribed time, or its ability to perform as desired. The unit can be biological or physical; when biological, reliability takes the label 'survival analysis,' a topic that has become the backbone of biostatistics.

2. Background: Interest in Reliability—Past, Present, and Future

Reliability is a quantification of a certain type of uncertainty. From a historical perspective, the quantification of this type of uncertainty dates back to Huygens, who in 1669 constructed what is now known as a 'survival curve.' Huygens' work was based on Graunt's data on vital statistics. Since the first 'survival table' was published by Edmund Halley in 1693, one may claim that statistical methodology, and statistical data analysis, have their origins in reliability. During the early 1930s, the problems of commerce and insurance sustained an interest in this topic, whereas between the 1950s and the 1970s, quantified measures of reliability were needed for engineering design, for government acquisitions in aerospace and defense, for enforcing regulation in drug approval, and in matters of public policy, such as the safety of nuclear power generators.

Landmark achievements during this period are the papers by Davis (1952) on the analysis of equipment failure, by Epstein and Sobel (1953) on life testing for the exponential distribution, by Kao (1956) for acceptance sampling based on the Weibull distribution, and by Grenander (1956) on the analysis of mortality measurement. The 1960s also heralded the birth of the mathematical theory of reliability, with Barlow and Proschan's (1965) classic book summarizing the knowledge to date. This work essentially defined the subject and provided avenues of research for years to come. Also noteworthy are the book by Gnedenko et

al. (1969) which summarized the Soviet work in the area, and Marshall and Olkin's (1967) fascinating paper on modeling dependencies.

During the 1980s and the early 1990s, pressures of consumerism and competitiveness forced manufacturers to use quantified measures of reliability for specifying warranties particularly with regards to the credibility of software (see, for example, Singpurwalla and Wilson 1999). The late 1990s also began to see concepts of reliability used in business, economics, and finance, and in matters of political science, particularly disarmament and national security.

The continued importance of reliability is anticipated in the twenty-first century because of a growing reliance on the infrastructure of communication systems, transportation networks, distributed power grids, and civil engineering structures at earthquake prone zones. Major initiatives by government and industry are underway under the label of 'infrastructure protection,' and quantified measures of reliability are a key factor of these. Finally, an area of growing concern wherein reliability is poised to play a signal role pertains to the credibility of nuclear stockpiles in the face of test ban treaties (cf. Armageddon Moves Inside The Computer, *The Washington Post*, November 28, 1998).

3. *Foundational Issues in Reliability: Is a Paradigm Shift Needed?*

Reliability, as a quantification of a certain uncertainty, leads to philosophical issues that have been the subject of debate. These issues are prompted by several questions that arise naturally, the first of which is: whose uncertainty? Is it that of an individual or a group of individuals, or is uncertainty inherent and therefore universal? The second question is if uncertainty is to be quantified, what should the quantification be based upon? Should it be subject matter knowledge, collective judgement, hard data alone, or a combination of all the above? The third issue pertains to the strategy for quantification. Should it be probability and the calculus of probability or should it be alternatives to probability like possibility theory, belief functions, confidence limits, tests of hypotheses, etc.?

Much of the literature on reliability fails to acknowledge these issues; rather, the focus has been on methodology—not principles. All the same, a major portion of the written material on reliability regards uncertainty as being inherent or universal, and subscribes to the principle that uncertainty can be quantified only by means of hard data. Furthermore, it takes the position that probability and its calculus need not be the only way to quantify uncertainty. Consequently, the mathematical theory (cf. Barlow and Proschan 1965, 1975) treats all probabilities as being known, whereas the statistical literature emphasizes 'life-testing' (a way to generate hard data), confidence limits, tests of hypotheses, etc. Indeed, statistical reliability theory's major export, the Military Standard 781C, wholeheartedly subscribes to the above approaches for describing uncertainty and for decision making. A more recent example is the book by Meeker and Escobar (1998). Whereas a use of this strategy may be meaningful in the context of mass phenomena, like the output of a continuous assembly line, its use for one of a kind items, like computer software or limited size populations, is ill-founded. Furthermore, approaches that rely solely on hard data tend to be ignored by the users. This in turn has undermined the role of statistics and calls for a paradigm shift for addressing the modern problems of reliability that have been raised by the engineering community. Does this mean that one should forsake the calculus of probability in favor of alternatives like possibility theory and fuzzy-logic that many are attracted to?

The answer to the above question is a cautious no! Despite some shortcomings of the probability calculus, there does not appear to be any need to forsake it. However, a paradigm shift is indeed necessary, but this has to do with the interpretation of probability, not an alternative to probability! To address the modern problems of reliability, problems involving one of a kind items for which there is little or no failure data but a lot of informed testimony, it is the interpretation of probability that matters. Specifically, one needs to subscribe to the subjective or the personal view of probability in its purest form, as advocated by de Finetti (1964) and by Savage (1972). In addition to facilitating reliability assessment of one of a kind items, the subjective view offers other advantages. First, it helps explain several results in reliability theory that have been obtained via purely mathematical derivations (see, for example, Lynn and Singpurwalla 1997, in Block and Savits (1997), and Kotz and Singpurwalla (1999). Second, it facilitates the development of normative decision making procedures for acceptance sampling and drug testing (see, for example, Lindley and Singpurwalla 1991 and Etzioni and Kadane 1993). Most important, the subjective view of probability enables the treatment of dependencies in system reliability assessment in a natural manner (cf. Lindley and Singpurwalla 1986). The matter of dependence has plagued reliability analysts for quite some time, because, it is often the case that independence is assumed in place of dependence, with the consequence that assessed reliabilities do not agree with actual experiences (cf. US Nuclear Regulatory Commission 1975, National Research Council 1998).

Thus, to summarize, the view is held (cf. Singpurwalla 1988) that the twenty-first century will see a paradigm shift in the methods for assessing reliability. This paradigm shift will entail the interpretation of reliability as a personal probability, and not the adoption of a new set of rules for describing and combining uncertainties.

4. Key Notions, Terminology, and Overview of Developments

In what follows, the position that reliability is a personal probability about the occurrence of certain types of events is adopted. Since the most common event of interest is survival of an item, under specified conditions, for a duration of time τ, $\tau \geqslant 0$, the reliability of the item is defined as

$$P(T \geqslant \tau) \stackrel{def}{=} R_T(\tau)$$

where P denotes probability, and $T \geqslant 0$, stands for the item's life-length. The interval $[0, \tau]$ is called the mission time, this terminology reflecting reliability's connections to aerospace. The quantity $R_T(\tau)$, as a function of $\tau \geqslant 0$, is called the reliability function, and if the item is a biological unit, then this function is called the survival function, denoted by $S_T(\tau)$. Clearly, $R_T(\tau)$ decreases in τ, going from one at $\tau = 0$, to zero, as τ increases to infinity. The life-length T could be continuous, as is usually assumed, or discrete when survival is measured in terms of units of performance, like miles traveled or rounds fired.

Let $f_T(\tau)$ be the derivative of $-R_T(\tau)$ with respect to $\tau \geqslant 0$, if it exists; the quantity

$$f_T(\tau) = -\frac{\mathrm{d}}{\mathrm{d}\tau} R_T(\tau)$$

is the probability density of $R_T(\tau)$ at τ. When $f_T(\tau)$ exists for (almost) all values of $\tau \geqslant 0$, then $R_T(\tau)$ is absolutely continuous, and $f_T(\tau)$ is called the failure model. Much literature in reliability pertains to ways of specifying failure models. Classic examples are the exponential with a scale parameter $\lambda > 0$, i.e.,

$$f_T(\tau \mid \lambda) = \lambda \mathrm{e}^{-\lambda\tau}$$

and the Weibull with a scale (shape) parameter λ $(\alpha) > 0$, i.e.,

$$R_T(\tau \mid \lambda, \alpha) = \mathrm{e}^{-\lambda\tau^{\alpha}}$$

Other examples of failure models are surveyed in Singpurwalla (1995). The expected value of T

$$E(T) \stackrel{def}{=} \int_0^{\infty} \tau f_T(\tau)\,\mathrm{d}\tau < \infty$$

is called the *mean time to failure* (MTTF). For the exponential model

$$E(T \mid \lambda) = \lambda^{-1}$$

A concept that is specific and unique to reliability is the failure rate function or the hazard function. For an absolutely continuous $R_T(\tau)$, the failure rate function $h_T(\tau)$, $\tau \geqslant 0$, is

$$h_T(\tau) \stackrel{def}{=} \frac{f_T(\tau)}{R_T(\tau)}.$$

The failure rate function derives its importance from two features, one interpretative and the other, technical. Specifically, since

$$h_T(\tau)\,\mathrm{d}\tau \approx P(\tau \leqslant T \leqslant \tau + \mathrm{d}\tau \mid T \geqslant \tau)$$

the failure rate at τ is (approximately) the probability of an item's failure in $[\tau, \tau + \mathrm{d}\tau)$, were the item surviving at τ. Clearly, for items that age with time, $h_T(\tau)$ will increase with τ, and vice versa for those that do not. This is the interpretative feature. The technical feature pertains to the fact that if

$$H_T(\tau) \stackrel{def}{=} \int_0^{\tau} h_T(s)\mathrm{d}s$$

then

$$R_T(\tau) = \exp(-H_T(\tau))$$

thus, knowing $h_T(\tau)$ is equivalent to knowing $R_T(\tau)$ and vice versa. The quantity $H_T(\tau)$ is known as the cumulative hazard at τ, and $H_T(\tau)$ as a function of τ is known as the cumulative hazard function. Biostatisticians like Kalbfleisch and Prentice (1980) have used a continuously increasing stochastic process, like the gamma process, to describe $H_T(\tau)$ for items operating in a random environment. The connection between $H_T(\tau)$ and the Lorenz curve of econometric theory (see Gastwirth 1971) has been noted by Chandra and Singpurwalla (1981). This connection suggests that concepts of reliability have relevance to econometrics vis-à-vis measures of income inequality and wealth concentration.

It is easy to verify that when

$$R_T(\tau \mid \lambda) = \mathrm{e}^{-\lambda\tau}$$

$\lambda > 0$, $\tau \geqslant 0$, $h_T(\tau \mid \lambda) = \lambda$, and vice versa. Thus, under an exponential model, the failure rate is the reciprocal of the MTTF and vice versa. However, this reciprocal relationship holds only for the exponential, and not for all other distributions as practitioners often assume. An attractive feature of the Weibull distribution is that by choosing α to be greater (smaller) than one, the failure rate function can be made to be increasing (decreasing) with τ. When $\alpha = 1$, the Weibull becomes an exponential. Because of this flexibility, the Weibull distribution has become one of the most widely used models in engineering reliability and survival analysis.

The failure rate function has become a cornerstone of the mathematical theory of reliability. Specifically,

all models whose failure rate increases (decreases) monotonically have been classified into one group called the IFR (DFR) class (for increasing (decreasing) failure rate), and properties of this class have been studied. All properties are in relation to the exponential which is both IFR and DFR, since increasing (decreasing) in IFR (DFR) is taken to be nondecreasing (nonincreasing). The famous 'bath-tub curve' of reliability engineering pertains to a distribution whose failure rate is initially decreasing, then becomes a constant, and finally increases, just like an old fashioned bath-tub. This functional form is appropriate for describing the life-length of humans, and large systems of many components.

The mathematical theory of reliability has many interesting results, several of which are intuitive, but some not. One such unintuitive result states that a mixture of exponential distributions (which have a constant failure rate) has a decreasing failure rate. Another counterintuitive result states that the time to failure distribution of a parallel redundant system of components having exponentially distributed life-lengths, has an increasing failure rate, but is not necessarily monotonic. Both these results appear in Barlow and Proschan (1975), but the arguments used to prove them are purely technical. Their intuitive import is apparent only when we adopt the subjective view of probability; Barlow (1985) makes this point clear.

The reliability assessment of systems of components is relatively straightforward if the component's life-lengths are assumed to be independent, the component's reliabilities are assumed, and the system is not a network. When the component reliabilities are unknown, the life-lengths are dependent. This situation becomes even more complicated when the system is a network. Lynn et al. (1998) proposed a Monte Carlo approach for treating such problems. Dependence in system reliability can be treated by multivariate failure models of the type introduced by Marshall and Olkin (1967). However, the number of parameters of such models grows exponentially with the size of the system, so that even for moderate size systems a use of multivariate models becomes an onerous task. One way of overcoming this difficulty has been to assume that the component life-times are associated and then to argue that reliability assessments assuming independence provide bounds for the assessed reliabilities if dependence was incorporated. This strategy may be suitable for small systems, but with large systems the lower (upper) bound tends to zero (one), so that the bounding is effectively meaningless. There is a pressing need for new multivariate models with a small number of parameters; an example is in Singpurwalla and Youngren (1993).

Thus far, the discussion has been restricted to the case of a single index of measurement, namely time or some other unit of performance, such as miles. In many applications, both engineering and biomedical, the survival of an item is indexed by two (or more) scales. For example, automobiles under warranty are indexed by both time and miles. In the biomedical scenario, the onset of disease is recorded with respect to age and also the amount of exposure to a hazardous element. Failure models involving more than one scale are, therefore, germane and initial progress on this topic is currently underway. Such models are known as failure models with multiple scales; it is important not to confuse these models with multivariate failure models. The former pertain to a single unit, whereas the latter to multiple units.

5. *Life Testing: Its Role in Reliability Assessment*

Life testing is a scientific and controlled way of obtaining data. For those whose do not subscribe to the personalistic view of probability, data are mandatory for reliability assessment, and this includes the selection of a model. The data could be from a life test or from the field. For those who subscribe to the personalistic view, the actual data is desirable but not essential. Desirable because the role of the data is to update ones initial assessment of reliability. This is achieved by using Bayes's law. The initial assessment of reliability is done by subjectively specifying a model and a prior distribution for parameters of the model, the prior also being subjective. The term subjective specification includes the use of expert testimonies (see, for example, Kadane and Wolfson 1998). The role played by the data, be it from a life test or the field, is to update the prior distribution to the posterior distribution.

Most life testing experiments are statistically designed experiments wherein several identical copies of the item whose reliability is being assessed are observed under conditions that replicate those of actual use. The life test is conducted until all items under observation fail or until a prespecified number fail. When the latter is done, the life test is said to be censored. The test can also be such that it is terminated after a fixed amount of time, irrespective of how many items fail in that time; such a life test is said to be a truncated test. A truncated test is terminated as soon as the last surviving item on the test fails. A life test can also be a combination of censoring and truncation, i.e., the test terminates at a fixed time or after a fixed number of failures, whichever occurs first.

When life tests are time consuming, it is common to make the conditions of the test harsher than those encountered in actual use. Such tests are called accelerated life tests, and tests with continuously increasing stress are those in which acceleration is done continuously (see De Groot and Goel 1979). Whereas accelerated life tests produce early failures, their disadvantage is that acceleration may induce changes in the material properties of the item, so that

the failure behavior is different. Accelerated life testing has a parallel in the biological context; it goes under the name of dose–response experiments.

Whereas censoring and truncation can be well controlled in laboratory tests, actual data may have arisen from either scenario, and the analyst may not know from which. This is known as the stopping rule problem (see, for example, Roberts 1967). Interestingly, for a subjectivist, the stopping rule is immaterial (provided that it is noninformative). However, for those who do not subscribe to the Bayesian point of view, knowledge of the stopping rule is fundamental. In the absence of such knowledge they cannot proceed. A very readable account of the stopping rule problem is in Lindley and Phillips (1976). It is because of these and the other arguments mentioned before that reliability in the twenty-first century has to be not only Bayesian, but also subjective Bayesian!

See also: Causal Inference and Statistical Fallacies; Quality Control, Statistical: Methods; Reliability: Measurement; Survival Analysis: Overview

Bibliography

Barlow R E 1985 A Bayes explanation of an apparent failure rate paradox. *IEEE Transactions Reliability* **R-34**: 107–8

Barlow R E, Proschan F 1965 *Mathematical Theory of Reliability*. Wiley, New York

Barlow R E, Proschan F 1975 *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart and Winston, New York

Block H W, Savits T H 1997 Burn-in. *Statistical Science* **12**(1): 1–13

Chandra M, Singpurwalla N D 1981 Relationships between some notions which are common to reliability theory and economics. *Mathematics of Operations Research* **6**(1): 113–21

Davis D J 1952 An analysis of some failure data. *Journal of the American Statistical Association* **47**(258): 113–50

Epstein B, Sobel M 1953 Life testing. *Journal of the American Statistical Association* **48**(263): 486–502

Etzioni R, Kadane J B 1993 Optimal experimental design for another's analysis. *Journal of the American Statistical Association* **88**(424): 1404–11

de Finetti B 1964 Foresight: Its logical laws, its subjective sources. In: Kyburg Jr. H E, Smokler H E (eds.) *Studies in Subjective Probability*. Wiley, New York

Gastwirth J L 1971 A general definition of the Lorenz curve. *Econometrica* **39**(6): 1037–9

Gnedenko B V, Belyayev Y K, Solovyev A D 1969 *Mathematical Methods of Reliability Theory*. Academic Press, New York

Grenander U 1956 On the theory of mortality measurements, Part II. *Skand Akt* **39**: 125–53

De Groot M H, Goel P K 1979 Baysian estimation and optimal designs in partially accelerated lift testing. *Naval Research Logistics Quarterly* **26**: 223–35

Kadane J B, Wolfson L J 1998 Experiences in Elicitation. *Journal of the Royal Statistical Society, Series D (The Statistician)* **47**(1): 3–19

Kalbfleisch J D, Prentice R L 1980 *The Statistical Analysis of Failure Time Data*. Wiley, New York

Kao J H K 1956 A new life-quality measure for electron tubes. *IRE Transactions Reliability and Quality Control* **PGRQC-7**

Kotz S, Singpurwalla N D 1999 On a bivariate distribution with exponential marginals. *Scandinavian Journal of Statistics* **26**: 451–64

Lindley D V 1985 *Making Decisions*, 2nd edn. Wiley, London

Lindley D V, Phillips L D 1976 Inference for a Bernoulli process (a Bayesian view). *American Statistics* **30**: 112–9

Lindley D V, Singpurwalla N D 1986 Multivariate distributions for the life lengths of components of a system sharing a common environment. *Journal of Applied Probability* **23**(2): 418–31

Lindley D V, Singpurwalla N D 1991 On the evidence needed to reach agreed action between adversaries with application to acceptance sampling. *Journal of the American Statistical Association* **86**(416): 933–7

Lynn N J, Singpurwalla N D 1997 Comment: 'Burn-in' makes us feel good. *Statistical Science* **12**(1): 13–9

Lynn N J, Singpurwalla N D, Smith A 1998 Bayesian assessment of network reliability. *SIAM Review* **40**(2): 202–27

Marshall A W, Olkin I 1967 A multivariate exponential distribution. *Journal of the American Statistical Association* **62**: 30–44

Meeker W Q, Escobar L A 1998 *Statistical Methods For Reliability Data*. Wiley, New York

National Research Council 1998 *Improving the Continued Airworthiness of Civil Aircraft. A Strategy for the FAA's Aircraft Certification Service*. National Academy Press, Washington, DC

Roberts H V 1967 Informative stopping rules and inferences about population size. *Journal of the American Statistical Association* **62**(319): 763–75

Savage L J 1972 *The Foundations of Statistics*, 2nd rev. edn. Dover Publications, New York

Singpurwalla N D 1988 Foundational issues in reliability and risk analysis. *SIAM Review* **30**(2): 264–82

Singpurwalla N D 1995 Survival in dynamic environments. *Statistical Science* **10**(1): 86–103

Singpurwalla N D, Wilson S P 1999 *Statistical Methods in Software Engineering*. Springer, New York

Singpurwalla N D, Youngren M A 1993 Multivariate distributions induced by dynamic environments. *Scandinavian Journal of Statistics* **20**: 251–61

US Nuclear Regulatory Commission 1975 Reactor Safety Study—An Assessment of Accident Risks in US Commercial Nuclear Power Plants. WASH-1400 (NUREG-75/014) October

N. D. Singpurwalla

Quality of Life: Assessment in Health Settings

1. Quality of Life

Since the 1980s or so there has been an increasing realization that traditional biologically based endpoints such as morbidity and mortality alone do not represent adequately the potential outcomes of medical interventions. Health status measurement has evolved to allow insight into patients' experiences in such areas of function as mobility, mood, life satisfaction, sexuality, cognition, and ability to fulfil

occupational, social, and family roles. Quality of life (QoL) has emerged as a broad term to describe this domain of measurement. The QoL construct may be viewed as a paradigm shift since it shifts the focus of attention from symptoms to functioning and establishes the primacy, or at least the legitimacy, of the patient perspective. QoL measures have many applications in medicine and healthcare. They are used to describe the subjectively perceived health and social status of given populations, to compare interventions and to assess the costs and benefits of treatments and health policies (Spilker 1996). Despite the increasing popularity of QoL, there is dissent about the meaning of the term, how it should be measured, and indeed whether it should be measured at all (Hunt 1997). In a critical review, Gill and Feinstein (1994) found that investigators defined what they meant by the term QoL in only 15 percent of the 75 articles reviewed. As Humpty Dumpty said to Alice: 'when I use a word, it means exactly what I want it to mean, nothing more and nothing less!'

2. *Health-related QoL*

Because of the amorphous and multidimensional nature of QoL, most researchers in medicine and healthcare concern themselves with a subcomponent of QoL called health-related QoL (HRQoL). This is distinguished from QoL as a whole, which would also include adequacy of education, housing, income, and perceptions of the immediate environment. Health status and HRQoL measures increasingly are used in studies addressing the costs and benefits of services and treatments (Patrick and Erickson 1993). The US Congress established the Agency for Health Care Policy and Research to undertake research on the effectiveness of medical care in terms of its impact on patient outcomes including HRQoL. The US Food and Drug Administration encourages the collection of HRQoL data for new drugs and, in the UK, health status and the experiences of patients and carers are included explicitly in outcome assessments.

3. *Definition of Health-related QoL*

Patrick and Erickson (1993) defined HRQoL as:

> the value assigned to the duration of life as modified by the social opportunities, perceptions, functional states and impairments that are influenced by disease, injuries, treatments or policy.

Assessment of HRQoL usually focuses on physical function, psychological state, social function, somatic symptoms, sexual function, occupational function, and, occasionally, on financial state. Many existing measures frequently are criticized for including only concepts at the negative or illness end of the continuum while neglecting concepts at the more positive end.

4. *Assessment of Health-related QoL*

There are a number of approaches to the assessment of HRQoL. Most of these involve some form of subjective multidimensional assessment preferably completed by the respondent whose QoL is under scrutiny since levels of agreement among doctors and between doctors' and patients' judgments are usually low (Sprangers and Aaranson 1992). Measures include generic instruments such as health profiles and utility measures, disease-specific questionnaires, and individually focused measures.

4.1 *Generic Measures*

Generic measures, such as the SF-36, The Nottingham Health Questionnaire, the McMaster Health Index Questionnaire, and The Sickness Impact Profile are broadly-based questionnaires and can be applied in a wide range of conditions and populations. The main limitation of generic questionnaires is that they are broad measures of health status. Consequently, they may not provide sufficient information on specific aspects of a given disease which are likely to be important in assessing disease impact and the outcome of treatment. It is common practice to use a generic measure together with a disease-specific measure. It is likely that the use of generic instruments will diminish as more sophisticated disease-specific measures are developed.

4.2 *The World Health Organization QoL Instrument (WHOQOL)*

The concept of HRQoL owes much to the original World Health Organization (WHO) definition of health as a state of complete physical, mental, and social well-being and not merely the absence of disease. The WHO established a multinational group to develop a measure that would allow for cross-cultural comparisons of QoL (WHOQOL 1993). The definition underlying the measure was:

> QoL is defined as the individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept affected in a complex way by a person's physical health, psychological state, level of independence and their relationships to salient features of their environment.

The WHOQOL is a somewhat unusual measure of HRQoL because, in addition to measuring physical and psychological health and social relationships, it also measures spirituality and it deals with facets of the environment such as financial resources and home environment. In developing the WHOQOL, the authors unexpectedly found that the basic factors inherent in QoL did not differ substantially across cultures (Power et al. 1999).

4.3 Utility Measures

Utility measures are generic measures derived from economics and decision theory and represent patients' preferences for different health states in the form of a single summary score. Their key characteristics are that they are based on preference judgments and usually are measured along a continuum from death (0.0) to perfect health (1.0), although scores less than 0 representing states worse than death are possible. Utility scores represent both health status and the value of that health status to the patient. They are used in cost-utility studies in which the cost of an intervention is related to the utility value afforded by the intervention. Various approaches to measuring utilities have been developed. In the Kaplan Index of Well-being Scale, case descriptions were compiled to illustrate combinations of functional levels, symptoms or problems and these cases were rated by random samples of the public who gave their preference ratings which were then used to determine utility values. In the somewhat controversial QALY (quality-adjusted life years) approach, improvements in the length and quality of life are amalgamated into a single index. In the standard-gamble approach, patients are asked to assess the risks they would tolerate for certain medical interventions. In the time-trade-off method, respondents are asked to estimate their relative preference for quality vs. quantity of life.

The utility approach assumes that individuals are capable of reporting their preferences and that they can predict accurately these preferences in relation to some future putative health scenario. It is further assumed that the preferences would not change if the patient were actually to experience such a health scenario. An additional problem relates to who should provide the utility values—the general public, healthcare providers, and/or patients and their families. One response to these problems has been the development of the Euroqol (Euroqol Group 1990). The aim of the Euroqol is to provide a standardized, nondisease specific survey instrument for describing HRQoL and to generate a single health index for each health state. The measure, however, suffers from a number of logistical and methodological limitations (Bowling 1995).

4.4 Disease-specific Measures

The second approach to QoL assessment focuses on aspects of health status that are specific to the area of primary interest. It is assumed that responsiveness will be increased by including only important aspects of HRQoL that are relevant to the patients being studied. The measure may be specific to a disease (such as cancer or arthritis), to a population of patients (frail elderly), to a certain function (sexual function or sleep), or to a particular problem (such as pain). Specific measures focus on areas that are likely to be explored routinely by clinicians and are likely to be more responsive to small but clinically significant changes. The modular approach adopted by the European Organization for Research on Treatment of Cancer (EORTC) provides a core generic measure for use in a variety of cancers supplemented by disease-specific modules for use in particular types of cancer (see *Cancer: Psychosocial Aspects*).

4.5 Limitations of Health-related QoL

The restriction of QoL assessment to HRQoL and the use of health-status measures to assess HRQoL is not entirely satisfactory. Most of the questionnaires used are essentially measures of health status in a new guise and these place an overwhelming reliance on the assessment of functional capacity. As such, while purporting to incorporate patients' perspectives, they represent implicitly a medical model that stresses the ability to perform everyday tasks and fulfil social and occupational roles. Such measures ignore the meaning and importance of such tasks and roles for the individual and often preserve the supremacy of professional judgements leading to the suppression of what is supposed to be under scrutiny (Joyce et al.1999, Hunt 1997).

5. Individual QoL Measures

A number of researchers have argued that QoL depends on the unique interpretation and perceptions of the individual and that questionnaires represent an oversimplification of what is a complex, multidimensional, subjective phenomenon. QoL is seen as a uniquely personal perception, denoting the way that individuals feel about their health status and/or nonmedical aspects of their lives. It can be measured suitably only by determining the opinions of patients and by supplementing (or replacing) the instruments developed by 'experts.' This hermeneutic approach would seek to understand individuals as self-reflective beings who are responsible for their own actions and who are the best judges of their QoL. This approach has resulted in the development of a number of individualized measures of QoL such as the Schedule for the Evaluation of Individual QoL or SEIQoL (Joyce et al. 1999).

6. Future Trends and the Need for Theoretical Models

Assessment of patient QoL facilitates improved clinical intervention, assists in treatment comparisons, and should prove increasingly important in the identification of services and facilities and in resource allocation. However, much of the research, to date,

has been underpinned by somewhat tautological operational definitions of QoL. There is a growing need for theoretical models that capture the psychological reality and complexity of QoL. For example, most current measures fail to address adequately the dynamic nature of QoL. Patients confronted with a life-threatening or chronic disease are faced with the necessity to accommodate to their illness. An important mediator of this adaptation process is 'response shift' which has been defined by Schwartz and Sprangers (2000) as changing one's internal standards, values, or one's conceptualization of QoL. Response shift is but one example of a psychological process that underpins the essentially dynamic and phenomenological nature of QoL. Research in this area must incorporate psychological variables into theory construction in order to produce a body of research that more validly reflects the true nature of QoL.

QoL has, heretofore, been considered solely as an outcome measure in health research. However, exciting developments in psychoneuroimmunology and evidence from a number of clinical studies raise the intriguing possibility that QoL might influence pathological processes such as tumor progression. If this were proven, then interventions aimed at maximizing patient QoL might also influence patient recovery.

See also: Cancer: Psychosocial Aspects; Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Health Outcomes, Assessment of

Bibliography

Bowling A 1995 *Measuring Disease: A Review of Disease Specific Quality Of Life Measurement Scales*. Open University Press, Buckingham, UK

Euroqol Group (Buxton M, O'Hanlon M, Pekurinen M ym) 1990 Euroqol: A new facility for the measurement of health-related quality of life. *Health Policy* **16**: 199–208

Gill T M, Feinstein A R 1994 A critical appraisal of the quality of QoL measurements. *Journal of the American Medical Association* **272**: 619–26

Hunt S M 1997 The problem of QoL. *Quality of Life Research* **6**: 205–12

Joyce C R B, O'Boyle C A, McGee H M 1999 *Individual QoL: Approaches to Conceptualization and Assessment*. Harwood, The Netherlands

Patrick D L, Erickson P 1993 *Health Status and Health Policy*. Oxford University Press, Oxford, UK

Power M, Bullinger M, Harper A, WHOQOL Group 1999 The World Health Organization WHOQOL-100: Tests of the universality of life in 15 different cultural groups worldwide. *Health Psychology* **18**: 495–505

Schwartz C E, Sprangers M A G 2000 *Adaptation to Changing Health: Response Shift in QoL Research*. American Psychological Association, Washington, DC

Spilker B (ed.) 1996 *Quality-of-life and Pharmacoeconomics in Clinical Trials*, 2nd edn. Lippincott-Raven, Hagerstown, MD

Sprangers M A G, Aaranson N K 1992 The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: A review. *Journal of Clinical Epidemiology* **45**: 743–60

WHOQOL Group 1993 *Measuring the Quality of Life: The Development of the World Health Organization Quality of Life Instrument*. WHO, Geneva, Switzerland

C. A. O'Boyle

Quangos in Politics

Quasi-governmental (q-g) bodies exist in their thousands, particularly in the developed countries. Their degree of quasi-official standing varies greatly. A simple typology, with examples, would offer (a) direct public services by national government ('g') itself (e.g., US Post Office); (b) indirect service by a state industry ('q-g') (e.g., the UK's Royal Mail); (c) a more indirect service by a wide variety of 'agencies,' 'quangos' (the loose British term) or 'public bodies.'

This wider q-g class extends to commercial firms and other private bodies which contract, in some sense, with government to provide a regulated service, perhaps in a privileged market position (telephone, bus or train companies in many countries). Given their official legal status as companies, foundations, charities, etc., these latter deliverers of public services are more 'quasi-private' than quasi-governmental.

When the state creates or sponsors a body to act as its agent, problems of democratic accountability are raised and can only be managed politically rather than solved by any particular method of appointment or regulatory regime. These relationships do not involve full privatization (sale of public assets into the private sector) but may involve contracting public sector work to the private sector or to this third sector of q-g bodies.

1. The Quasi-governmental Sectors

In the more advanced and elaborate political systems, quasi-governmental (q-g) bodies have increasingly come to deliver or even devise public policies and services on a noncivil service basis. Managerial assertions that the q-g approach to public services offers better quality have claimed intellectual support from organization theory. Political assertions that the state should itself perform fewer tasks in general and avoid some particular types of task altogether have arisen from antistatist, procommercial or procivil society values. Separating practical issues of how to improve public policy design and delivery from ideological and normative issues of what the state should be doing with its direct budgets and civil service cadres is an essential step in understanding the q-g phenomenon.

Any body which is not part of central or local government may be intended (or may come) to act in an q-g role. Its formal standing may be as an official body ('agency' or, in the UK, 'quango' – quasi-autonomous national government organization), perhaps specifically created by law ('statutory') with its own controlling board. It may be a corporation (company, firm) whether 'for-profit' (commercial) or 'not-for-profit' (philanthropic, charitable). It may be a voluntary membership organization made up of individuals, corporations (a trade association) or other types of body such as schools or colleges. Whatever it once was or may appear to still be, any body which stands outside the formal state bureaucracy may come to behave as a quasi-official extension of state authority. This process commonly has a legal basis (the body undertakes government contracts to provide public services) or a financial one (receiving official grants for its charitable, artistic, or equivalent work). Occasionally, a private body becomes, in effect, quasi-governmental without contracts or other money passing to it but on a simply political and psychological level: the body helps the government by providing information and access to a problem area or target group and is rewarded with consultative status which draws it close to official policies and thinking.

2. *Government-appointed Bodies*

The rise of a quasi-governmental sector in advanced political systems displays two broad types of body within great variety of detail: 'public bodies' created by governments to do a policy job on a q-g basis, and existing private bodies (commercial, philanthropic, representative, or fraternal) which become significantly q-g in their attitude and behaviour. The former group is always more prominent and significant for public service delivery and therefore also for the problem of 'independence versus accountability' which arises from unelected bodies disposing of public functions and budgets. These bodies are distinguishable from central or local government's own bureaucratic cadres by their separate titles, budgets, governing boards and (latterly) development and 'business' plans, performance targets and other fashionable banners (even though mainstream civil services now often also display them). It is the separate, government-appointed and paid board which most clearly distinguishes these q-g bodies from the government's own direct operations because this board is a legal entity and employer, doing its work through its own staff and contracted or cooperative links with other bodies, including commercial suppliers.

Various terms have been applied since the q-g sector first attracted systematic academic attention some 30 years ago (Smith and Hague 1971, Smith 1975). The term 'the contract state' reflected the US scene with its easier acceptance of business-type contracts for public programs, notably during and after the War on Poverty period. British assumptions (and therefore language) about public services eschewed talk of contracts, although a few had long existed, in public education, for example (Hague et al. 1975). Different theories of the essential exclusivity of state functions were encouraging or discouraging new forms of 'agency' or 'intermediary body' in different European systems, notably in the more pluralist Netherlands or Scandinavia and the more statist France or Italy, respectively.

The British tradition combined a strong state theory, stewarded by a very powerful senior civil service, within a quite rich civil society context—including, for example, the senior professions and leading charities at national level. Nominated public bodies with specialist remits and only generalized 'upwards' accountability to government ministers (and none 'downwards' to users or recipients of their services) were quite familiar devices. In this soil, the British 'quango' form multiplied (Barker 1982). The flexibility and precise control offered by a single-purpose public body was generally irresistible to both parties in government and to their senior civil servants who continually steer and supervise these bodies. By 1996, a properly functional independent analysis of British appointed public bodies found some 6,400 being funded (and therefore in some way controlled) by central government. The 5,750 executive (cf. purely advisory) bodies were judged to have spent some £60B of public money in 1994–5, an increase of 50 percent at constant prices in 16 years, following continual transfers of expenditure functions from central and local governments to new nominated bodies and the creation of new programs under these 'quango' auspices (Hall and Weir 1996).

3. *The Politics of Quasi-governmental Bodies*

The most ideological and most partisan of political debates about q-g bodies deliberately created by the state seems to have arisen in the UK: the 'quango issue.' Superficially (and tediously) the two main political parties have, for 25 years, accused each other of abuse of the patronage involved. A more serious criticism of all appointees is their lack of accountability. Continental Europe seems to have been spared much of this British conflict, perhaps because business, labor unions, the churches and other civil society formations accept each other and expect to share in delivering public services through special agencies or boards, under broadly consensual political oversight. The Dutch have their three 'pillars' of social service delivery (Catholic, Protestant, and secular), while in Germany it is officially polite for organized capital and organized labor to call themselves 'social partners' (Barker 1997). The USA no more follows that custom than does the UK, but the vast proliferation of

appointed and joint bodies which constitute 'America's quangos' seems not to have sparked either union–business rivalry as to patronage or specific partisan hostility. The long-term right wing rumblings against 'big government' subsumes these many 'agencies' within broader issues of overall federal or state civil service numbers and public expenditure levels. Academic critiques have often also taken a broad view of problems and pathologies (Burns 1994, Carter 1994, Kettl 1988).

4. *The Rationale of Quasi-governmental Bodies*

The principal practical rationale for creating a new public body to tackle either a novel or a familiar policy task is its specialist character. The new single-purpose agency or quango has been expected to succeed whereas the old, multipurpose government department would wearily add this extra burden to its already overstretched bureaucracy. If the new body's budget and staff are also outside the official civil service total budget and numbers, and if a small, 'dedicated' (specialist) office can bring sensitivity rather than standardized bureaucracy to some novel public policy challenge (such as encouraging ethnic or gender rights, offering drug or HIV health care, making grants to artists and theaters, or promoting community development) then there is no contesting this way forward. The same applies where an existing private body is to be a vehicle for such novel tasks by becoming a government contractor or grantee in return for its experienced and expert work on policy design or delivery.

A further rationale may well apply when professional expertise is crucial to a policy or program. Senior scientists and other experts are very unlikely to become civil servants, much less join the military, and can therefore be employed by the state only on their own ground and terms. The US Air Force and US Army could not get the high-quality operations researchers or mathematicians they needed except by founding the RAND Corporation (initially to analyze bombing targets in the USSR) and an applied mathematics centre on the University of Wisconsin's campus, respectively. Even in the crisis of war, the assorted, mainly British, academic intellectuals who were recruited to break the Enigma and other enemy codes at Bletchley Park ('Station X') in 1940 or the scientists and mathematicians organized into the Manhattan Project to build the US atom bomb would perform their best, or even perform at all, only in a carefully managed academic and civilian setting, as their governments realized. These civilian–military contrasts were merely one example of the lessons of organization theory which have declared the 'flat' (not hierarchical) 'open' (not sectionalized) 'organic' or flexible and small agency or company to be the best setting for creative, innovative, interactive or sensitive tasks. As modern governments have been pressed into policy work which is either technically demanding (requiring expert staff) or socially sensitive (requiring nontraditional official styles and methods) so the appeal of specialist, flexible, and suitably qualified staff working in a small single-issue body has steadily increased. With obvious differences in funding and controls, these basic characteristics make either a specifically created new public agency ('quango') or an existing private body drawn in to q-g status the attractive option.

5. *The Dilemma of Independence vs. Accountability*

The essence of the q-g problem is that practical reasoning favors a democratically responsive government trying to obtain good policy or policy delivery by creating or recruiting a specialist agency ('quango') to concentrate on that job. The trickier the task, in political, technical, or social terms, the greater the appeal of this method, in many cases at least. But this very difficulty and sensitivity prompts critics to claim that democratic accountability to elected representatives (national or local) is particularly needed. Having an unelected body of government appointees running (say) the local mass transit, or even regular public grants to theaters or art galleries may be tolerable but life and death decisions on (say) drug budget allocations to hospitals or decisions affecting employment must be taken directly by elected representatives. To produce the advantages of flexibility, innovation, and effective informality and to attract the most talented and expert staffs or consultants, a q-g body does need to be independent of day-to-day control.

This dilemma of operational independence (in search of creative quality of service) versus public accountability (in support of democratic legitimacy) can only be softened, not solved. A q-g body's initial pioneering independence could be time-limited and fuller reporting and answering of questions and debate then imposed upon it. Turnover of board members could be formally encouraged or even required. In general, an 'explanatory accountability' regime is needed for the army of q-g bodies which advanced political systems now maintain (AMA 1995). Although specific cases or issues are not normally removable or reversible, board members can be made to explain their views and decisions. Rules requiring openness, transparency, and full reporting would all illuminate this explanatory accountability. Even this working culture would inhibit some valuable creative independence in some policy fields. A statutory duty to explain itself to a possibly critical public would be a major challenge for any q-g body, particularly if the courts are also open to calls for their judicial review of any such explanations of actual decisions which appear to have been based on mistakes or illogical

argument. The standard work on the British position (Skelcher 1998) explores these ideas.

6. A New Sector of Government and Society

So many tasks of policy formation and delivery; of regulation of private activity 'affected with the public interest' (to use the traditional US phrase); and of social innovation are now required in advanced democratic societies that the rise of this middle sector, standing between state and private sectors can be no surprise. Its lack of much political controversy in these systems suggests a declining popular interest in democratic principles in favor of practical, consumerist attitudes to public services.

See also: Policy Knowledge: Nonprofit Institutions; Public Administration, Politics of

Bibliography

AMA (Association of Metropolitan Authorities) 1995 *Changing the Face of Quangos: Proposals for the Reform of Local Appointed Bodies*. AMA, London

Barker A (ed.) 1982 *Quangos in Britain: The Making of Public Policy Networks*. Macmillan, London

Barker A 1997 Other people's quangos: Quasi-government in seven countries. In: Flinders M (ed.) *How to Make Quangos More Democratic*. Charter 88 Publications, London

Burns N 1994 *The Formation of American Local Governments: Private Values in Public Institutions*. Oxford University Press, Oxford

Carter S L 1994 *The Confirmation Mess: Cleaning-Up the Federal Appointment Process*. Basic Books, New York

Hague D C, Mackenzie W J M, Barker A (eds.) 1975 *Public Policy and Private Interests: The Institutions of Compromise*. Macmillan, London

Hall W, Weir S 1996 The untouchables: Power and accountability in the Quango state. Democratic Audit Paper #8 The Scarman Trust, London

Kettl D F 1988 *Government by Proxy: Managing Federal Programs*. Congressional Quarterly Press, Washington, DC

Select Committee on Public Administration (1999) Quangos Sixth Report, 1998–99, HC 209 – 1 and 2, 9 November 1999. The Stationery Office, London, Vols 1 and 2I

Skelcher C 1998 *The Appointed State: Quasi-Governmental Organizations and Democracy*. Open University Press, Buckingham, UK

Smith B L R (ed.) 1975 *The New Political Economy: The Public Use of the Private Sector*. Macmillan, London

Smith B L R, Hague D C (eds.) 1971 *The Dilemma of Accountability in Modern Government: Independence Versus Control*. Macmillan, London

A. Barker

Quantification in History

'Quantification' refers to a specific method of historical research that departs from the prevalent practice of textual analysis by actually seeking to measure changes in the past. Due to the mass of data to be handled and the complexity of the calculations involved, the application of quantitative methods usually has to be *computer-assisted*. In contrast to a mere illustration of certain tendencies by tables or graphs, quantification tends to use *analytical statistics* in order to probe relationships between certain variables. Such statistical analysis also requires a *high degree of formalization* in the research design that formulates hypotheses to be tested and tries to combine them into models. The final aim of the quantitative approach is not just to explain a particular set of developments but to make a *theoretical contribution* to some facet of historical understanding.

1. The Rise of Quantification

Within historical research, the development of quantitative methods was a product of a particular constellation during the 1960s. Already during the Enlightenment some scholars had sought to describe socioeconomic developments by compiling time-series, and by the middle of the nineteenth century national governments took on the task of determining population trends through a series of censuses. But it took the shift to the 'new social history' a century later to direct historical interests away from affairs of state to mass politics, societal changes, demographic transformations, or economic growth. At the same time the arrival of mainframe computers, followed eventually by powerful PCs, made it possible to handle large amounts of data and to use complicated statistical procedures in customized form via statistical packages such as SPSS or SAS. In order to explore these exciting new subjects with novel techniques, historians became interested in the theoretical constructs of the social sciences, borrowing eclectically from political science, sociology, or demography and somewhat less so from economics. The result of this interaction was the rapid development of a new hybrid, combining elements of historical research with social scientific analysis.

During the 1970s, this 'social science history' or 'historical social science' was successfully institutionalized. In the United States the multitude of initiatives in history and the neighboring disciplines combined in founding the Social Science History Association in 1975, and made its annual meeting a showcase of the most innovative work. In France many of the long-term community studies, produced by the *Annales School*, turned quantitative, compiling detailed time-series of population, prices, or wages and probing their inter-relationship. In Germany, the quantifiers linked with the theory-oriented project of a *Historische Sozialwissenschaft* to create a vigorous organization, called QUANTUM. Within the decade, a spate of new journals emerged to provide publication

outlets for the new work, such as the *Journal of Social History*, *Journal of Interdisciplinary History*, *Historical Methods Newsletter*, *Social Science History*, *Histoire et Mésure*, and *Historical Social Research*. New Textbooks (Dollar and Jensen 1971, Floud 1979) as well as influential anthologies (Aydelotte et al. 1972, Lorwin and Price 1972) provided a powerful platform for the propagation of quantitative research.

2. *Research Contributions*

The first field transformed by quantitatively-oriented work was the *new political history* that differed from its predecessor by its attention to collective phenomena (Bogue 1983). For instance, Michael Kater investigated the rise of the Nazi movement in Germany by analyzing its membership files, indicating precisely which social stratum, religious group, gender, region, etc. joined the bandwagon at what time. For democratic societies like the US, Lee Benson's or Alan Bogue's analysis of electoral behavior proved quite productive, since comparing returns over time made it possible to establish which polls were the 'critical elections' when party realignments were transformed by 'ethno-cultural' or other factors. For parliaments without party discipline, Thomas B. Alexander's roll-call analysis also offered a powerful instrument for establishing who actually voted with whom, indicating what networks dominated legislative behavior. This retro-projection of political science methods gradually provided a firmer empirical basis for an analysis of the transformation of mass politics in the last two centuries.

The second area that was deeply affected by quantitative methods and conceptual borrowing from sociology was the *new social history* (Stearns 1985). As 'history from the bottom up,' this perspective examined previously untouched serial records, such as police files, employment registers, or draft books in order to infer from the behavior of ordinary people what they might have felt and thought. The appeal of this effort was the diffusion of its topics, since the labor history of the Tillys included subjects as different as proletarian living conditions and their collective efforts at protesting through strikes. Another question, treated e.g., by Hartmut Kaelble concerned the pattern of social mobility, comparing the relative openness of societies to upward or downward movement over time or across borders. A related issue was the societal impact of education that required an analysis of the enrollment and composition of schools or universities in order to get at changes in educational opportunities. These quantitative probes produced much new and sometimes conflicting information about the lower orders, the make-up of institutions, and basic trends in societal transformations.

A third area propelled by computer-assisted source collection and statistical analysis was the field of *historical demography* (Hareven and Vinovskis 1978). For time-periods before national censuses, the demographers faced the challenge of reaggregating vital statistics from the ground up by converting parish registers into reliable measures of natality, nuptiality, or mortality. Under the direction of Anthony Wrigley the Cambridge population project established the outlines of a population history for England, while the demography center in Umea undertook a national project of quantifying church records so as to explain demographic changes in Sweden. American demographers like John Demos and Tamara Hareven instead combined individual life-event data to scrutinize the development of family structures from extended via stem to nuclear families and the like. Other scholars such as Roderick Floud and Richard Steckel studied changes in the height of military recruits or slaves in order to develop more accurate measures of social deprivation. This historical demography amassed a source base for the prestatistical area and produced interesting interpretative concepts for explaining the population transition.

The final area in which quantification transformed research priorities and methods was the *new economic history* (Davis and Engerman 1987). Redirecting attention away from entrepreneurship, 'cliometrics' as it was also called, inspired efforts to establish indicators of national production and income for the prestatistical age as well as to transform information from manuscript censuses into published form. In the US, a heated debate revolved around the profitability of slavery, with Robert Fogel's and Stanley Engerman's claims producing much controversy, due to their combination of economic and moral argumentation. Much attention also focused on the key factors of economic development, such as railroads, with some scholars actually arguing that their impact had been exaggerated. In Europe discussions revolved around the pattern of industrialization instead, with some specialists like Sidney Pollard proposing a more complex regional diffusion model. On the whole, the quantitative impulse made economic history more rigorous and theoretical, but removed it at the same time further from the mainstream of historical debate.

By the 1980s these signal contributions had managed to overcome traditionalist skepticism and to establish quantification as a new and growing subfield within the historical discipline. Though they were sometimes difficult to read for laymen, high-quality monographs convinced reluctant granting agencies to fund major collective initiatives, such as the Philadelphia Social History Project, directed by Theodore Hershberg. The number of articles with graphs and tables in respected historical journals began to reach almost one-quarter (Kousser 1989). Opinion surveys of the profession suggested that most scholars expected quantification to be here to stay. Even if some critics still bemoaned the spottiness of graduate training in this area, the rapid rise of quantitative methods in

historical research had been impressive and their future prospects seemed secure.

3. The Culturalist Challenge

During the decade of the 1980s, however, the seemingly unstoppable progress of the quantitative history 'project' began to stall. By the middle of the decade, the number of quantitative journal articles and the frequency of graduate courses in quantitative techniques suddenly began to drop (Reynolds 1998). It seems that over the past 10 years (plus or minus five!) quantitative history has become increasingly marginalized, since a variety of other methods have risen to prominence, making so-called 'quants' seem rather quaint. Ironically, the demise of quantitative history has occurred simultaneously with the development of enhanced technology, more complete datasets, and increasingly sophisticated models and methods, which, taken together, have rendered quantitative history *potentially* more useful than ever before.

The factors responsible for this loss of attention are still being debated vigorously. Some scholars attribute the problems afflicting quantitative history to the initial overselling of the method by insufferable practitioners who promised more insights than their research was able to deliver. These critics, comprised largely of traditionalists, have always been suspicious of the proposition that history had anything to learn from the social sciences and have long bristled at heady sentiments such as that expressed by Emmanuel Le Roy Ladurie who predicted in 1967 that in the future 'the historian will be a programmer or he will be nothing' (Rabb 1983). No wonder that such skeptics gladly seized upon Bernard Bailyn's and Lawrence Stone's calls for a revival of the narrative.

Another reason was the cultural turn of many younger historians who rejected both the materialist premises and positivist vision of most practitioners of quantitative history. To these critics, quantitative representations of the past were at once too cold, too remote, too impersonal, and too abstract. Large datasets were said to dehumanize, and 'soulless' quantifiers were assumed to be ready, like Marx's bourgeoisie, to drown all human sentiment 'in the icy water of egotistical calculation' (Coclanis 1992). Moreover, in making fetishes of rigor and precision, quantitative historians purportedly attempted to exact more from the vast, vague, largely unknowable past than historians had any right to claim. Instead, the past was reconfigured as a 'foreign country,' to invoke the title of historian David Lowenthal's influential 1985 book.

The changing demographics of the historical profession, particularly in the United States, also served to reinforce the trend towards qualitative approaches. As history opened up to women and minorities—groups, generally speaking, less inclined to pursue quantitative approaches to the past—quantitative history could be expected to decline in a relative sense. Just as women and minorities have been under-represented in the *hard* sciences over the years, they have proven less interested in the more social scientific sides of history. Though both groups have embraced social history as a vehicle for their identity politics, relatively few of their members can be found in fields such as economic history, historical demography, or quantitative political history.

During the 1980s, many historians turned instead to cultural anthropology for methodological inspiration. Rejecting the rigorous, formal and remote, data-driven 'truths' sought by quantifiers, scholars interested in experience and meaning preferred in-depth microstudies, whether of single villages, small groups, or even well-documented individuals. Gradually, 'thick description' came to trump ecological regression, crude death rates, and multiple R-squares, no matter how robust! The philosophical basis of this shift to anthropology was the rise of postmodernism, which deprecated the rationalism, universalism, and empiricism of historical social science. By the late 1990s the ascendancy of cultural studies left quantifiers off in a corner silently counting all by themselves.

4. The Continued Importance of Quantification

The current disregard of quantitative methods in the discipline is unfortunate, since quantification continues to be an essential technique for historical research. Some questions can best be answered through recourse to what Sir William Petty referred to in the seventeenth century as 'Number, Weight, or Measure,' and it is more than 'physics envy' that has drawn scholars to quantitative analyses of the past. Even after conceding that many early practitioners were a bit overly zealous in their claims, it seems fair to suggest that quantitative history can provide the methodological discipline, intellectual rigor, and moral purpose that subjectivist forms of history often lack. With all due respect to such influential postmodernist thinkers such as Baudrillard and Lyotard not every aspect of the past is positional, free-floating, and contextual.

The utility, indeed, the necessity of quantitative history seems more compelling in light of important new developments in the field. Technological constraints have eased considerably, for example, as miniaturization and digitalization in the electronics industry have advanced. Today's personal computers, programmable calculators, and other platforms offer tremendous power and, thus, computational possibilities to quantitative historians. Similarly new developments in statistical packages, spreadsheets, CD-ROM

technology, GIS (geographic information systems), and the like have vastly expanded the analytical reach of software that can be applied to historical questions.

The increasing availability of new and/or enhanced sources and datasets also presents unprecedented research opportunities. Materials readily retrievable from established institutions such as the ICPSR (Inter-university Consortium for Political and Social Research) in Ann Arbor, Michigan, and data available on-line from the World Bank, the IMF, and other governmental, quasigovernmental, and nongovernmental organizations and agencies are transforming the nature (and sites) of scholarly production. The new CD-ROM version of the *Historical Statistics of the United States*, datasets assembled by Yale's Cambodia Genocide Project, the University of Minnesota's IPUMS (Integrated Public Use Micro Sample) Project, drawn from US census records, and the East German central cadre files—to cite but four examples—are of similar scholarly importance, and have tremendous potential as teaching tools. Various research projects undertaken under the auspices of, or promoted by the Association for History and Computing, founded in England in 1987, have proven extremely useful as well.

At the same time, quantitative historians have in recent years developed more sophisticated methods, allowing them to analyze and interpret historical questions with greater sensitivity and precision. Economic historians, for example, have utilized both game-theoretical approaches and advanced econometric techniques to excellent effect, and anthropometric historians have employed important new procedures from auxology and biostatistics to facilitate their reconstruction of our biological past (Steckel 1995). Economic and social historians, along with historical sociologists, have also begun to employ regression analysis in more subtle and refined ways, using temporally recursive regression procedures, when appropriate, instead of standard time-series regression techniques.

Other quantitative historians have begun to 'push the envelope' in different ways, in so doing, often trespassing on the territory of scholars in other social sciences or even in the humanities. Thus, as procedures such as social network analysis (SNA), event-structure analysis, and quantitative language analysis are successfully employed in answering historical questions—quantitative analyses of major 'texts' in order to produce so-called semantic grammars, for example—non- and anti-quantifiers ignore the results at their peril. In addition, quantitative historians' recent incorporation of spatial relationships into historical analysis through spatial-effects regression models and the like has produced similarly impressive results (Griffin and van der Linden 1998).

At the beginning of the twenty-first century, the future prospects of quantification in history are therefore indistinct. Clearly, individual quantitative historians, among the profession's most avid proponents of scientific rationalism, universalism, and empiricism, are continuing to expand the frontiers of their research and are still making important contributions. But whether or not the interest of mainstream historians will return to quantification and the subfield as such will ever again recapture the excitement and energy it possessed in the 1970s remains an open question. The relative popularity of various approaches does not just depend on the quality of their monographic work but on the larger tides of intellectual fashion. However, one unexpected development offers some cause for optimism: though quantitative *historians* hang on for their professional lives, historically-inclined political scientists, economists, and sociologists are winning kudos in their respective disciplines by bringing quantitative methods to bear on the past.

See also: Economic History; Economics, History of; Historical Demography; Historiography and Historical Thought: Current Trends; History and the Social Sciences; History: Theories and Methods; Measurement Theory: History and Philosophy; Positivism, History of; Postmodernism: Methodology; Prosopography (Collective Biography); Quantification in the History of the Social Sciences; Social History

Bibliography

Aydelotte W O, Bogue A G, Fogel R W (eds.) 1972 *The Dimensions of Quantitative Research in History*. Princeton University Press, Princeton, NJ

Bogue A G 1983 *Clio and the Bitch Goddess*. Sage Publications, Beverly Hills, CA

Coclanis P A 1992 History by the numbers: why counting matters. *OAH Magazine of History* **7**: 5–8

Davis L E, Engerman S 1987 Cliometrics: the state of the science. *Historical Methods* **20**: 97–106

Dollar C M, Jensen R J 1971 *Historian's Guide to Statistics*. Holt, Rinehard and Winston, New York

Floud R 1979 *An Introduction to Quantitative Methods for Historians*, 2nd edn. Methuen, London

Griffin L J, van der Linden M (eds.) 1998 New methods for social history. *International Review of Social History* **43**: 3–8 (Supp. 6)

Grossbart S R 1992 Quantitative and social science methods for historians. *Historical Methods* **25**: 100–20

Hareven T K, Vinovskis M A (eds.) 1978 *Family and Population in Nineteenth Century America*. Princeton University Press, Princeton, NJ

Haskins L, Jeffrey K 1990 *Understanding Quantitative History*. MIT Press, Cambridge, MA

Hudson P 2000 History by numbers: An introduction to quantitative approaches. Oxford University Press, New York

Jarausch K H, Hardy K 1991 *Quantitative Methods for Historians*. University of North Carolina Press, Chapel Hill, NC

Kocka J 1984 Theories and quantification in history. *Social Science History* **8**: 169–78
Kousser J M 1989 The state of social science history in the late 1980s. *Historical Methods* **22**: 13–20
Lorwin V R, Price J M (eds.) 1972 *The Dimensions of the Past*. Yale University Press, New Haven, CT
Rabb T K 1983 The development of quantification in historical research. *Journal of Interdisciplinary History* **13**: 591–601
Reynolds J F 1998 Do historians count anymore? *Historical Methods* **31**: 141–8
Stearns P N 1985 Social history and history: a progress report. *Journal of Social History* **19**: 319–34
Steckel R H 1995 Stature and the standard of living. *Journal of Economic Literature* **33**: 1903–40

K. H. Jarausch and P. A. Coclanis

Quantification in the History of the Social Sciences

Quantification means, very broadly, the use of numbers to comprehend objects and events. While quantification is often identified with science, it has been equally significant for business, administration, and the professions. Quantitative methods had a crucial role in shaping the social and behavioral sciences, and remain central to them. The purpose of this article is not to supply a practical guide to the use of quantitative tools in social science, but to examine their backgrounds and assess their cultural significance from the standpoint of science studies and the history of science. In modern times, quantitative methods and the numbers they produce have assumed a powerful role in most societies as intersecting discourses of science, administration, and the public sphere. This article emphasizes the more public uses of quantification.

The broad category of quantification refers to several allied but distinct human activities. Mathematization means the characterization of objects and events in terms of mathematical relationships. While it is not the main topic of this article, it is of course relevant. The expression of scientific theories in mathematical terms has provided one important reason for relying on quantitative or numerical data. There have been other reasons of a technical, administrative, or moral kind. Among the more empirical forms of quantification, measurement grew out of such practical activities as surveying lands or buying and selling. Counting, the basic tool for assigning numbers to sets of discrete objects was deployed for such tasks as inventorying one's possessions or recording human populations. *Calculation* refers loosely to methods for combining and manipulating numerical information into more readily useable forms. Statistics, a body of methods for analyzing quantitative information, is discussed historically in other essays in these volumes, and receives also some attention here.

1. Theories and Measures

As early as the 1660s, and more systematically from about 1830, social and economic investigators held up quantification as the proper method of science. The growth of university-based social disciplines in the twentieth century, and especially after about 1930, was attended by a widespread faith that through numbers, social research could be made genuinely scientific. A generation or more of positivistically-inclined researchers looked to systematic measurement as the proper foundation for solid, mathematical theory. Some compared its progress with the development of physics in the seventeenth century, and anticipated a breakthrough in the near future.

The most influential historical and philosophical writers on science in the postwar period took almost the opposite view. Koyré (1968), whose philosophical studies of seventeenth-century science inspired the first generation of professional historians of science in the anglophone world, seriously doubted that mere measurements could offer much to a new or unformed science. Science could never have become mathematical, he argued, through the mere processing of empirical measurements. He doubted that Galileo had learned much from rolling balls down inclined planes. The new kinematics applied not to real objects, but to abstract bodies in geometrical space, and Galileo must have derived his law of falling bodies mainly by working out logically the consequences of what he already knew. This mathematical and metaphysical work was, for Koyré, the most important achievement of the new science. Indeed, only through theory had the measures attained scientific meaning. Kuhn 1961, 1976) reaffirmed Koyré's broad argument and provided a historical perspective that reached forward to the nineteenth century. The most successful sciences of the seventeenth-century 'scientific revolution,' including mechanics, astronomy, and geometrical optics, had already been mathematical in ancient times. The more Baconian, experimental studies of the early modern period were comparatively ineffective until a 'second scientific revolution,' beginning around 1800, when new scientific objects such as electricity and heat began to be comprehended mathematically.

Scholars in science studies since about 1980 have tended to dethrone theory in favor of a more respectful attention to scientific practices. Accordingly, they take more seriously the quantification of phenomena that have not been grasped by exact laws. Just as the life of experiment has never been strictly subordinated to theory, so numbers have often been independent of higher mathematics. They are linked not just to

theoretical science, but also, and perhaps even more consequentially, to economic and administrative changes. The history of ideas alone cannot convey their full significance.

2. Quantities and Standards

Quantity is not first of all the language of science, imported as something alien into everyday affairs, but an indispensable part of the fabric of life. Yet the aspiration to science has gone a long way toward reshaping the role of numbers and measures in the world. In this—as Max Weber suggested—we find much in common between bureaucratic and scientific rationality. Mathematics has often, and plausibly, been held up as an exemplar of universalized knowledge. The rules of arithmetic, or constants like π, are the same everywhere, at least in principle. Yet the universalism of measures and numbers, particularly when these have empirical content, has been achieved only in a limited way, by overcoming great obstacles. Even now, even in the industrialized West, quantities and prices are managed in everyday activities such as shopping by techniques quite different from those taught in schools (Lave 1986). Before the rise of modern science, in the pre-industrial world, quantification had very different meanings. Numbers were pervasive, but for most purposes their use and meanings were thoroughly local. The process through which they gained general validity is one of the crucial transformations of modernity.

2.1 A Plethora of Units

As Kula (1986) has shown, measures remained immensely variable in Europe as late as the eighteenth century. This was partly by design, since small cities and lesser states often claimed the right to define their own weights and measures as a mark of autonomy and an obstacle to rule by higher authorities. A diversity of measures, however, is the natural state of things wherever the pressures of commerce and administration have been insufficient to bring about unification. At a local level, authorities took some pains to reduce the ambiguity of measurement. A bushel measure, valid for the region, would often be fixed to the town hall, and could be used to settle disputes. In the next town, however, measures might well be different, and over greater distances they varied considerably. The subtle geography of measures, moreover, was not defined exclusively by distance. Milk, oil, and wine came in different units; measures of cloth depended on the width of looms, and also on the costliness of fabric and weave. The practices of measurement, too, were complexly nuanced. A bushel measure, even if the volume was fixed, could be tall and narrow or short and wide; the wheat might or might not be compacted by being poured in from a considerable height; and a complex set of conditions governed whether to heap the bushel, and by how much. Parties to a transaction often negotiated about measures rather than about prices. Add to this the consideration that almost nothing was decimalized, not even money, and it is evident that arithmetical calculation provided no fixed routine for managing affairs or settling disputes (Kula 1986, Frängsmyr et al. 1990). The conversions required by trade provided an important source of employment for mathematicians.

The simplification and standardization of measures was encouraged by the expansion of long-distance trade, but was achieved primarily by governments in alliance with scientists. Among Europeans, it was the British who came nearest to creating uniformity of measures in the eighteenth century. A more systematic, almost utopian, scheme, the metric system, was initiated in the 1790s under the French Revolution. In its most ambitious form this included a new calendar, with the declaration of the Republic defining the Year One, using months of 30 days divided into weeks of 10 days, and even a few decimalized clocks, beating out 100 seconds per minute. The meter was set so that 10 million of them would reach from pole to equator, in the vain hope of integrating geodetic and astronomical measures with those on a human scale (Gillispie 1997). Hence, bold expeditions were sent out into a disordered countryside to survey and measure a meridian. Although the meter was justified in part as a response to pleas in the *cahiers de doléance* for fixed measures, against the social and economic power implied by seigniorial discretion, the highly rationalized metric system of decimalized conversions and strange Greek names—of milliliters, kilometers, and centigrams—went well beyond this. Not until the July Monarchy, after 1830, did the metric system begin to succeed even in France (Alder 1995). In time, however, it has come to provide a common set of measures for most of the world, with the United States a notable, though partial, holdout.

A precise and maximally rigorous definition of units is a basic element in the infrastructure of modern life. The role of standardized measures in systems of manufacturing based on interchangeable parts is widely recognized. They are indispensable also for large-scale networks such as electric power distribution, telephony, and, by now, computers and the Internet. A collaboration of scientists and bureaucrats was required to create them. In some cases, as with the electrical units defined in the latter nineteenth century, eminent scientists such as James Clerk Maxwell, William Thomson (Kelvin), and Hermann von Helmholtz were deeply involved in the fixing of standards.

Other units, far from the domain of scientific theory, play an indispensable role in the regulation of modern life. Measures of effluent concentrations, for example, are deployed by environmental agencies to regulate

pollutants. They depend on meticulous specification not only of instruments, but also of sampling methods, training of technicians, custody of materials, and so on. Since penalties and taxes often depend on them, they are always vulnerable to challenge. In this domain, as in many others, an immense apparatus of instrumentation and control lies behind what scientists and citizens have the luxury (usually) of regarding as 'raw,' or merely 'descriptive,' data.

2.2 Producing Uniformity

Units and instruments are not by themselves sufficient to quantify effectively. A quantifiable world contains reasonably homogeneous objects, objects whose measures can be compared, or that can be grouped together and counted. Nature and unregulated human labor produce a few of these, though real uniformity occurs mainly on a scale—for example, the molecular one—that is not readily accessible to the human senses. Physics was quantified partly in terms of abstractly 'mathematical' entities, such as forces, and partly through increasingly rigorous experimental controls, involving an expanding apparatus of instruments.

In the social world, where quantification is also a tool of administration, these same regulatory processes have helped to produce new quantifiable entities. The investigation and compensation of unemployment, for example, solidified the category of 'the unemployed.' To collect unemployment insurance requires that a person meet certain standards, which are more or less closely monitored, and to which persons who lack work are encouraged to conform. They also must enroll themselves, thereby providing basic data for the statistics. The point here is not that government offices cause people to be without work, though mechanisms of compensation will certainly affect the rates. It is that official categories sharpen definitions, and so help to create countable entities. Most human and social types are variable, or at least fuzzy at the margins (Hacking 1995). Quantification is allied to processes that make them more uniform, and hence countable. A similar story can be told about ethnicity or mental illness. Economic categories such as profits, investment, and trade, are similarly dependent on the processes through which they are recorded and regulated.

3. The Authority of Calculation

Numbers and measures are not merely tools of description and analysis. They work also as guides to action, especially public action. An explicit and formal analysis, often largely quantitative in form, is by now required for a variety of public decisions involving investment or regulation, especially in the United States and Britain. Business organizations, too, rely increasingly on measures and projections in order to make decisions. These analyses are often expressed in money terms, and their modern expansion attests to a triumph of capitalism. But public reliance on economic quantification is by no means an imitation of its commercial uses. The tools of quantitative analysis were developed for as much for government as for business purposes, and have become increasingly dependent on academic researchers. In the background is an evolving political and administrative culture, reflecting changes in the status and composition of elites.

3.1 Commensuration

It is characteristically modern to suppose that a conscientious decision should involve explicit consideration of the available data. Decision by calculation goes beyond this, in expecting a course of action to be dictated objectively by facts. The prototype of this ideal of public reason is cost-benefit analysis. In practice, cost-benefit often involves predictions in many domains, including weather patterns, health, hydrology, ecology, and commerce, but its most distinctive aspect is the attempt to convert a great diversity of considerations into a common financial metric. The difficulties of commensuration are most acute in regard to such considerations as human lives lost or saved, the disappearance of biological populations, or the degradation of a beautiful landscape. To the lay public, this appears equivalent to sacrificing a life for a sum of money, and of very doubtful morality (Sagoff 1988).

Economics provides a rationale for equations of this kind. Economists argue that people routinely accept increased risks for better wages. Neoclassical economic theories presuppose an underlying variable, utility, which individuals seek to maximize. Even if it cannot be directly measured, personal choices therefore involve some implicit commensuration of pleasures and responsibilities, acquisition and leisure, friendship and personal ambition. Cost-benefit analysis did not, however, develop out of theoretical economics, but from specific political and administrative problems of the mid-twentieth century. Engineers, not economists, initially faced the immense difficulties of incorporating life, health, recreation, and scenic beauty along with economic development as they proffered measures of the costs and benefits of water projects. Despite political opposition to some of these measures, a political logic of rationalized choice made them necessary none the less. The task was to devise methods that led to reasonably consistent answers, and where necessary to put aside what seemed their implausibility or even immorality. Beginning in the 1950s, economists and social scientists were recruited into agencies like the US Army Corps of Engineers and the Bureau of Reclamation to provide more consistent or more acceptable solutions to these thorny problems (Espeland 1998, Porter 1995).

3.2 Trust and Impersonality

Academic problems of social measurement and commensuration are, thus, embedded in social and political conditions that put a premium on scientific objectivity. In important ways, of course, science and philosophy have always been concerned with objectivity, in the sense of truth. Since the seventeenth century, the relation of fact and values has become more problematical, and for some purposes the credibility of science has been enhanced by its separation from the 'subjective' domain of values. The ambition to provide a scientific, possibly quantitative, basis for public decisions remained quixotic or utopian until at least the late eighteenth century. It remained unrealistic wherever an aristocratic political class retained power. Not everyone would have assented to Frédéric Le Play's assessment that numbers have become necessary only in democracies, where leaders no longer control the wisdom of birth and experience. But the aspiration to replace judgment with objective, quantitative rules grew up in situations where elite authority had become suspect.

The move by American state engineers to an increasingly rigorous cost-benefit methodology was a response to systematic challenge, and exemplifies this point (Porter 1995). So also do other exemplary tools of 'mechanical objectivity,' such as the accountant's bottom line and the test of statistical significance of the social scientist or medical researcher. Accountants, for example, did not abandon willingly what many saw as their professional responsibility to offer an expert judgment of the financial condition of firms. They did so, rather, in the context of regulatory pressure and public distrust. This must be understood in the context of a broad cultural and political history of the twentieth-century, and especially of the United States. Some basic features of that history would include the growth of government, frequent interagency conflict, increasingly effective opposition to administrative secrecy, and public distrust of 'bureaucracy.' The unprecedented quantitative enthusiasm of postwar social science involved, along with its optimism about science, a renunciation of the personal and the subjective that was nicely adapted to contemporary political circumstances. Allied to it was the institutionalized reliance on quantification taught in management and policy schools.

4. A Historical Perspective on Social Science Quantification

Reflective accounts about the social and behavioral sciences have often supposed that their fields have short histories, and that the pursuit of quantification has meant following the track of natural science triumphant. But the rise of a quantitative ethos in natural science can be dated no earlier than the mid-eighteenth century. Systematic measurement until then was almost confined to astronomy, for which a host of 'mathematical instruments' had been developed in China, India, the Near East, and Europe. Many were geometrical and not numerical. Their purposes were practical as well as scientific: to survey land, navigate ships, calculate calendars and horoscopes. Barometers and thermometers helped to extend the domain of measurement into meteorology. Measurement in early modern Europe was associated as closely (or more closely) with natural history as with experimental physics. By 1750, instruments were deployed systematically to measure mountains and other natural wonders and, thus, to lend precision to the appreciation of nature. The development of an exact science of electricity in the later eighteenth century marked the triumph of experimental quantification (Heilbron 1979). Measurement had also become routine in technological endeavors, such as mineral assaying, decades before Lavoisier and his contemporaries began to insist on its centrality to chemistry.

From this perspective, it would be difficult to argue that political and economic studies lagged behind natural science in their reliance on numbers. Among the early members of the Royal Society of London, founded in 1660, were William Petty and John Graunt. Petty announced a new science of 'political arithmetic,' which in his hands involved speculative estimates of population and of wealth. Graunt is noted for his more strictly empirical compilations from the London bills of mortality. Near the same time, Jan de Wit, the Dutch Stadholder, used rudimentary probability theory to assign rates for the sale of annuities, and the astronomer Edmond Halley calculated the first mortality table. Political arithmetic, which entailed the compilation as well as the analysis of demographic and medical records, flourished in the eighteenth century. At the same time, probability theory was pursued as a mathematical theory of belief and credibility (Hacking 1975, Daston 1988). The *philosophe* and mathematician Condorcet, a leader of the French Revolution who then was hunted down by it, hoped to develop probability into the theoretical basis for elections and judicial decisions (Baker 1975).

4.1 The Nineteenth Century and the Rise of Statistics

By the early nineteenth century, as Kuhn observed, the natural-historical investigation of heat, electricity, and the physics of light had given way to sciences that were both mathematical and quantitative. Possibly the analysis of mortality records for insurance can be regarded as comparable in certain respects. But the explosive growth of quantification beginning in the 1820s, what Hacking (1990) has called an 'avalanche of printed numbers,' was not concentrated in the

experimental sciences. Rather, it involved systematic surveys of stellar positions, terrestrial magnetism, weather phenomena, tides, biogeographical distributions, and, most prominently, social statistics.

The collection and investigation of social numbers had expanded gradually over the course of the eighteenth century. Few Enlightenment states had the bureaucratic means to conduct a full census, and most were disinclined to release such sensitive information to their enemies or even their subjects. In many places, too, the people would have viewed census official as harbingers of new taxes or military conscription, and in Britain the census was blocked by a gentry class that refused to be counted. In contrast to piecemeal efforts at home, there were some systematic surveys abroad, the machinery of the information state having been turned loose first of all in colonial situations. Among European states, Sweden pioneered the regular census, in 1749; its results remained secret until 1764. The French used sampling, of a sort, and enlisted mathematicians such as Pierre Simon Laplace to estimate the population. The US census of 1790 heralded a new era of regular, public censuses, which were encouraged in Europe by French military conquest and by the pressure of wartime mobilization. The British decennial census, for example, was instituted in 1801. Only in the1830s did bureaucratic professionals begin to take over the collection of social and economic numbers in the most advanced European states.

At the same time, an increasing pace of quantitative study by voluntary organizations made 'statistics,' pursued by 'statists,' a plausible 'science of society.' One model for it was developed by the Belgian Adolphe Quetelet, whose quantitative pursuits ranged from observational astronomy and meteorology to the bureaucratic organization of official statistics. He argued for an alliance of data gathering with mathematical probability, in the quest to uncover 'social laws.' Apart from life insurance, however, most statistical collection was in the control of reformers and administrators with little mathematical knowledge. Their enthusiasm for counting was perhaps indiscriminate, yet it was nurtured by specific anxieties about poverty, sanitation, epidemic disease, economic dislocation, and urban unrest. Social investigation, still more than natural science, was an object of contest, and there were active efforts to organize a social science to speak for working people (Yeo 1996). Karl Marx's immense statistical labors for his *Capital*, drawing mainly from official British inquiries whose results he praised as honest and reliable, testify impressively to the credibility of social numbers. Middle-class organizations had more staying power than working-class ones, and better standing to claim the mantle of 'social science.' In Britain, the Statistical Society of London (founded 1834) and the National Association for the Promotion of Social Science (1857) joined bureaucrats with professionals and even aristocratic political leaders. In Germany, statistics became a university discipline, which, from about 1860 to 1900, combined empirical quantification with a historicized 'national economics,' in a reformist campaign for state action on behalf of workers and peasants. Some of the founders of social science in America were educated in this tradition, as was Max Weber. Emile Durkheim, too, engaged with it, as his *Suicide* of 1897 plainly attests (Porter 1986).

The nineteenth-century science of statistics embraced much of the subject matter of the modern social sciences. It overlapped, and sometimes competed, with political economy, political science, geography, anthropology, ethnology, sociology, and demography. These stood for genres, more or less distinct, though not for anything so definite as modern professional disciplines. Their quantitative methods were not identical, but neither were they clearly differentiated from each other or from measurement in biology. There was much uncertainty and debate from 1830 through the end of the century as to whether statistics was a science, with its own subject matter, or only a method. It was resolutely empirical, even to the point that sampling seemed too speculative. Far from being allied to any mathematical social science, statistics was more often at war with deduction and abstract theory. Some, like the economist William Stanley Jevons, would claim that political economy should be mathematical because it was inherently quantitative, but he could not attach the mathematical formulations he introduced to empirical numbers.

4.2 Quantification in the Modern Social and Behavioral Sciences

Toward the end of the nineteenth century, a more mathematical field of statistics began to develop. The most influential version of the new statistics was developed within a British biometric tradition by Francis Galton and Karl Pearson. At almost the same time, however, distinct social and behavioral sciences began to crystallize, in part around increasingly distinct quantitative methods. All were conceived as ways to enhance the scientific standing of their fields, and each was engaged also with issues of politics, policy, and the management of economies and populations. The typology to follow is intended to be suggestive, and certainly cannot be exhaustive.

4.2.1 Physical anthropology and eugenics. The measurement of humans was carried out for bureaucratic as well as scientific purposes throughout the nineteenth century. The increasing concern with race in anthropology was linked to a preoccupation with skulls, whose measures were used to distinguish distinct 'races' of Europeans. What Francis Galton called 'correlation' was developed to measure the

interconnections among physical human characteristics, their perpetuation from one generation to another, and their relation to measures of merit or achievement.

4.2.2 Psychometrics and psychology. Correlation provided also a solution to a problem of contemporary psychophysics, related to education. Gustav Theodor Fechner began applying the methods of error theory to the measurement of sensory discrimination in the mid-nineteenth century. Psychophysics was plagued, however, by the problem that a task—for example, of distinguishing two slightly different weights—could not be repeated indefinitely, because the subject learned over time to do better. Indeed, the left hand acquired some of the fine discrimination developed through practice with the right hand. By the end of the century, this obstacle had become a topic of investigation in its own right, as an approach to the psychology of learning. Systematic use of experimental randomization was developed for studies of this kind (Dehue 1997). They provided also a wide field of application for correlational methods. Did childhood study of dead languages really train the mind for speedier or more thorough learning of other subjects, such as science, history, or literature? The growth of mental testing in the early twentieth century opened another question of correlations. To what extent did students who proved themselves adept in some particular study excel also in others? Charles Spearman's *g*, or general intelligence, and the 'intelligence quotient' or IQ, which flourished in America, presupposed a basic unity of the mental faculties. Psychometricians and educational psychologists developed a body of statistical methods to measure this intelligence, or to decompose it, and to assess the relations of its components (Danziger 1990).

In the 1930s and 1940s, as psychology turned resolutely experimental, quantification became practically mandatory. Once again, educational psychology and parapsychology took the lead in the assimilation of novel methods of experimental design and analysis. The new paradigm of a proper experiment involved a randomized design of control and experimental populations, yielding results that would, in most cases, be subjected to analysis of variance (ANOVA). A worthwhile result was defined in part by a test of significance: the 'null hypothesis' that the treatment had no effect should be rejected at some specified level, commonly 0.05 (Gigerenzer et al. 1989).

4.2.3 Econometrics and economics. The new mathematics of marginalism was developed in the 1870s and self-consciously reconciled to the classical tradition in the 1890s by Alfred Marshall. His work included abundant graphical illustrations, but these involved quantities that could not easily be measured. The origins of econometrics are largely distinct. From origins in the 1870s to the establishment of a society and a journal in the 1930s, the field focused on the study of business cycles (Morgan 1990). Econometricians took their numbers where they could get them, often from public sources. They were highly dependent on national accounts, prepared by governments, and indeed participated in structuring these accounts. The paradigmatic statistical tool of econometrics, adapted from biometrics mainly after 1950, was regression analysis.

4.2.4 Surveys. From the founding of statistical societies in the 1830s, social science has relied on surveys to learn about how lives and opinions. Charles Booth's monumental studies of East London beginning in the 1880s inspired a movement of surveys of towns and regions in Britain and America. At first, researchers were expected to examine everyone, but after 1900, sampling began to gain acceptance. Arthur Bowley and then Jerzy Neyman developed sampling methods that permitted generalization using probability theory. The techniques of survey sampling have linked sociology and political science to marketing and political campaigns (Bulmer et al. 1991, Desrosières 1998).

4.2.5 Modeling. By the late twentieth century, the use of models was pervasive in natural and social sciences alike. Modeling methods are far too diverse to think of summing them up, but the modeling outlook deserves comment. It is perhaps under this rubric that mathematical methods are most often integrated with the data of tallies and measurements. A quantitative model, very broadly, means working out the consequences of certain assumptions for a particular situation or problem. Often, models are used for predicting, and they can equally well be purely academic or highly pragmatic, for very specific applications. In the twentieth century they have come to pervade the sciences. Among the social disciplines, modeling is especially central to economics.

5. Conclusion

Quantification by now is virtually as important for social as for natural science. Its expansion since the 1700s reveals some key characteristics of the social and behavioral sciences: among them, the drive for the status of science, the role of applied studies in forming

academic ones, and the often problematic relationship of theoretical to empirical work. The quantification of social phenomena goes back several centuries, and is by no means a mere product of the effort to attain scientific standing. Indeed, the advance of measurement and statistical analysis in social science is not distinctively academic, but has been greatly stimulated by administrative and political demands. Yet quantification in social science has, since the 1800s, been surrounded by a certain mystique, as an indication of rigor or at least of scientific maturity. It should be regarded not as the key to all mysteries, but as a powerful set of tools and concepts, to be integrated as much as possible with theoretical understanding and with other ways of comprehending social phenomena.

See also: Econometrics, History of; Psychometrics; Scaling and Classification in Social Measurement; Scientific Revolution: History and Sociology; Statistics, History of; Statistics: The Field

Bibliography

Alder K 1995 A revolution to measure: The political economy of the metric system in France. In: Wise N (ed.) *The Values of Precision*. Princeton University Press, Princeton, NJ pp. 39–71

Baker K 1975 *Condorcet: From Natural Philosophy to Social Mathematics*. University of Chicago Press, Chicago

Bulmer M, Bales K, Sklar K (eds.) 1991 *The Social Survey in Historical Perspective, 1880–1940*. Cambridge University Press, Cambridge, UK

Daston L 1988 *Classical Probability in the Enlightenment*. Princeton University Press, Princeton, NJ

Danziger K 1990 *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge UniversiPress, Cambridge, UK

Dehue T 1997 Deception, efficiency, and random groups: Psychology and the gradual origination of the random group design. *Isis* **88**: 653–73

Desrosières A 1998 *The Politics of Large Numbers*. Harvard University Press, Cambridge, MA

Espeland W 1998 *The Struggle for Water*. University of Chicago Press, Chicago

Frängsmyr T, Heilbron J, Rider R (eds.) 1990 *The Quantifying Spirit in the Eighteenth Century*. University of California Press, Berkeley, CA

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L 1989 *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, Cambridge, UK

Gillispie C 1997 *Pierre Simon Laplace, 1749–1827: A Life in Exact Science*. Princeton University Press, Princeton, NJ

Hacking I 1975 *The Emergence of Probability*. Cambridge University Press, Cambridge, UK

Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK

Hacking I 1995 *Rewriting the Soul*. Princeton University Press, Princeton, NJ

Heilbron J 1979 *Electricity in the 17th and 18th Centuries: A Study in Early Modern Physics*. University of California Press, Berkeley, CA

Koyré A 1968 *Metaphysics and Measurement*. Harvard University Press, Cambridge, MA

Kuhn T 1961 The function of measurement in modern physical science. In: Woolf H (ed.) *Measurement*. Bobbs-Merrill, Indianapolis, IN pp. 31–63

Kuhn T 1976 Mathematical versus experimental traditions in the development of physical science. *Journal of Interdisciplinary History* **7**: 1–31 (reprinted in Kuhn T 1977 *The Essential Tension*. University of Chicago Press, Chicago, pp. 31–65)

Kula W 1986 *Measures and Men* (Szreter R trans.). Princeton University Press, Princeton, NJ

Lave J 1986 The values of quantification. In: Law J (ed.) *Power, Action, and Belief: A New Sociology of Knowledge*? Routledge, London pp. 88–111

Morgan M 1990 *The History of Econometric Ideas*. Cambridge University Press, Cambridge, UK

Porter T 1986 *The Rise of Statistical Thinking, 1820–1900*. Princeton University Press, Princeton, NJ

Porter T 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ

Sagoff M 1988 *The Economy of the Earth: Philosophy, Law, and the Environment*. Cambridge University Press, Cambridge, UK

Yeo E 1996 *The Contest for Social Science*. Rivers Oram Press, London

T. M. Porter

Quantifiers, in Linguistics

1. Introduction

Quantifiers are free-standing expressions whose meanings involve the notion of quantity, such as English three, several, numerous, most, every, one hundred and twenty three, all but seventeen, and so forth.

The basic semantic structure of quantification is bipartite, consisting of the quantifier itself plus the expression that it quantifies. For example, in a sentence such as John saw three cats, *three* is the quantifier, and *cats* the quantified expression.

Quantification has traditionally been of great interest to semanticists, logicians and philosophers of language, due at least in part to the perceived 'logical' or 'mathematical' nature of the meanings involved. As such, it is striking to observe how such basic and seemingly immutable meanings may, in different languages, be expressed with very different morphosyntactic strategies, exhibiting a great degree of cross-linguistic variation.

Quantifiers may be classified internally, with respect to their morphosyntactic structure and basic semantic

properties. In addition, they may be classified externally, with regard to the syntactic and semantic relationships between the quantifiers and the quantified expressions and/or the remainder of the sentences in which they occur.

2. Internal Classifications

Internal classifications of quantification make reference to a variety of fundamental morphosyntactic and semantic properties.

2.1 Mass and Count Quantifiers

One of the most basic distinctions is between mass and count quantifiers (Pelletier 1979, Allan 1980). Mass quantifiers constitute expressions which denote an undifferentiated homogeneous mass; for example, English *much* is a mass quantifier, because it forms expressions such as much milk, much cake. In contrast, count quantifiers constitute expressions which refer to one or more countable units of characteristic size and shape; for example, English *many* is a count quantifier, because it forms expressions such as many cats, many cakes. Whereas many quantifiers belong to just one of the two types, mass or count, some quantifiers are undifferentiated with respect to the mass/count distinction, and may appear in expressions of either kind.

An interesting case is that of the so-called 'classifier languages,' many of which are located in the East Asian linguistic region (Craig 1986; see example (4) below). In such languages, many or all of the count quantifiers cannot occur in immediate construction with the noun that they quantify; instead, the quantifier must occur in construction with a classifier, and the quantifier-plus-classifier constituent may then occur in construction with the noun. Typically, the classifier makes reference to various characteristics—size, shape, function, and so forth—of the quantified noun. Accordingly, different nouns often require different classifiers. However, in some cases, different classifiers may occur with the same noun, sometimes resulting in subtle differences in meaning.

2.2 Existential and Universal Quantifiers

Of particular interest to logicians are the existential and universal quantifiers (Vendler 1967, Heim 1982, Gil 1995). Existential quantifiers, such as English some and a(n), form expressions denoting at least a minimal, non-zero amount or number of the quantified expression, for example some water, a boy. Universal quantifiers, such as English *all* and *every*, form expressions referring to an exhaustive amount or number of the quantified expression, for example all the water, every boy. Existential and universal quantifiers are semantically related to each other through negation by means of the classical 'square of opposition': thus, in English, Not a boy came is equivalent to All the boys did not come, and Not all the boys came is equivalent to A boy did not come.

Existential quantification, in many languages, is inextricably intertwined with singular number and/or indefiniteness: in some cases they are expressed identically, while in other cases they are expressed by means of the same marker occurring in different constructions, or by means of similar though not identical markers. Universal quantification, too, is expressed in a variety of different ways across languages. Some languages, such as English, do have 'dedicated' universal quantifiers, that is to say, words such as all and every whose primary or exclusive function is the expression of universal quantification. However, many other languages lack such words, instead making use of more complex circumlocutions if and when the expressive need arises.

Given the importance of existential and universal quantifiers in mathematical logic, it is worthy of note how relatively unimportant these quantifiers are within the grammatical systems of languages. It is indeed not all that common to find a single form, in any language, whose interpretation corresponds precisely and unproblematically to either the existential or the universal quantifier. Moreover, there is probably no language in which expressions of existential and universal quantification constitute a natural grammatical class to the exclusion of other expressions. The diversity of morphosyntactic expression of existential and universal quantifiers across the worlds languages suggests that, although greatly valued by logicians, they may not be endowed with privileged status vis à vis the structure of language.

2.3 Mid-range Quantifiers

'In between' the existential and the universal quantifiers lie a variety of mid-range quantifiers, such as, in English, few, several, many, most, and so forth. Among the mid-range quantifiers, a particularly privileged category is that of cardinal numerals—quantifiers which refer to natural numbers, for example, one, two, thirteen, one hundred and twenty seven (Menninger 1969, Hurford 1975). Complex phrasal numerals often make use of syntactic patterns available elsewhere in the language; however, in addition, they sometimes exhibit unique structural features not found in other grammatical domains

2.4 Other Internal Classifications

Quantifiers are sometimes classified as either *strong* (alternatively referred to as definite) or *weak* (indefinite) (Milsark 1977). Strong quantifiers, such as

English all and most, are inherently definite in their meaning, whereas weak quantifiers, such as English some and three, are not. Strong and weak quantifiers often exhibit different syntactic behaviour; for example, strong quantifiers are typically awkward or ungrammatical in existential sentences, e.g.,? There is all the water on the floor.

Some further semantic distinctions are captured by the categories of *increasing* (alternatively referred to as upward entailing) and *decreasing* (or downward entailing) (Keenan and Stavi 1986). A quantifier is increasing if a sentence containing it would be logically true in a situation in which the amount or number of quantified expressions involved is larger than that explicitly stipulated by the quantifier, for example, at least three. Conversely, a quantifier is decreasing if a sentence containing it would be logically true in a situation in which the amount or number of quantified expressions involved is smaller than that explicitly stipulated by the quantifier, for example, at most three. Other quantifiers are neither increasing nor decreasing, for example exactly three.

An interesting problem is posed by simple monomorphemic quantifiers such as three. At first blush, such quantifiers appear to be neither increasing nor decreasing: if somebody says There are three cats on the bed, the mental picture that is conjured up is, of course, one of three cats, not two, or four, or any other number. However, a moment's reflection will reveal an important asymmetry. If in fact there are exactly two cats on the bed, then the above sentence is clearly false. In contrast, if there are exactly four cats on the bed, then, in most contexts, the above sentence would obviously be misleading and inappropriate—but it would not, strictly speaking, be false. Thus, while the pragmatic appropriateness conditions of three resemble those of exactly three, the actual truth conditions of three are identical to those of at least three. Hence, three is, in fact, increasing.

3. *External Classifications*

External classifications of quantification make reference to the morphosyntactic and semantic relationships which obtain between the quantifiers and the quantified expressions and/or the remainder of the sentences in which they occur.

3.1 *Nominal and Verbal Quantifiers*

One basic semantic classification relates to the category of the quantified expression, specifically whether it is of a nominal or verbal nature (Gil 1993). For example, in the English sentence John saw three cats, the numeral three quantifies a noun; accordingly this sentence illustrates nominal quantification. In contrast, in the sentence John sang three times, the numeral phrase three times quantifies a verb; hence this sentence instantiates verbal quantification.

Up to a point, nominal and verbal quantification exhibit parallel grammatical properties. Obvious, but still worth noting, is the fact that in both of the preceding sentences, the same numeral, three, is used. However, nominal and verbal quantification also differ in many ways; typically, verbal quantification is more highly marked than its nominal counterpart.

3.2 *Continuous and Discontinuous Quantifiers*

A distinct albeit related syntactic classification pertains to the position of the quantifier with respect to the quantified expression, specifically, whether or not the two form a continuous constituent (Gil 1993, Bach et al. 1995). Sentences (1)–(4) below all instantiate nominal quantification, in which a numeral quantifies an expression referring to an animal; however, each of the examples exhibits a different syntactic configuration.

(1) HEBREW
raiti [*šaloš* arnavot]
see-PAST: 1: SG three-F rabbit-PL: F
'I saw three rabbits'

(2) TAGALOG
Tatlo [*ang* *nakita*
three TOP PAT.TOP: PFV-see
kong *baboy*]
DIR: 1: SG-LIG pig
'I saw three pigs'
('The pigs seen by me were three')

(3) KUTENAI (Matthew Dryer, p.c.)
hu [*qaˇsaˇ* *wu:kati*] *cupqa*
1: SG three-PRV see-IND deer
'I saw three deer' ('I three-saw deer')

(4) JAPANESE
Watasi *wa* *inu* o
1: SG TOP dog ACC
sanbiki *mita*
three-CLF see-PFV
'I saw three dogs' ('I saw dogs threely')

In Modern Hebrew, (1), as in its English gloss, the quantifier groups with the quantified expression, *šaloš arnavot*, to the exclusion of the remainder of the sentence; in Tagalog, (2), the quantifier *tatlo* stands alone as a sister to the remainder of the sentence, which forms a constituent; in Kutenai, (3), the quantifier occurs in construction with the verb, *qaˇsaˇ wu:kati*, to the exclusion of the remainder of the sentence, while in Japanese, (4), the quantifier *sanbiki* is a sentential adverb not forming an obvious constituent with any part of the remainder of the sentence.

Constructions (1) and (2) exemplify continuous quantification, in which the quantifier and the quantified expression form a constituent. In contrast, constructions (3) and (4) exemplify discontinuous quanti-

fication, in which the quantifier occurs apart from the quantified expression. Examples (1)–(4) above illustrate the distinction between continuous and discontinuous quantification with respect to nominal quantification. In principle, a similar distinction may also be drawn with respect to verbal quantification, though examples of discontinuous verbal quantification are more hard to come by.

3.3 Scope and Scopal Quantifiers

When two quantified expressions are present in the same construction, a variety of semantic relations may obtain between the two expressions (Jackendoff 1971, Gil 1982b). The following English sentence, containing two numerically quantified NPs, may, at least potentially, be interpreted in any or all of the following four ways.

(5) Three boys saw two girls
  (a) 'Each of three boys saw each of two girls'
    not scope differentiated: strong symmetric.
  (b) 'Three boys saw two girls between them'
    not scope differentiated: weak symmetric.
  (c) 'Three boys saw two girls each'
    scope differentiated: wide scope for *three* boys
    distributive key: *three boys*; distributive
    share: *two girls*.
  (d) 'Two girls were seen by three boys each'
    scope differentiated: wide scope for *two girls*
    distributive key: *two girls*; distributive
    share: *three boys*.

Interpretations (5a) and (5b) both involve a single set of three boys and a single set of two girls: each of the two NPs has independent reference. These interpretations are accordingly characterized as non-scope-differentiated, or symmetric. Interpretations (5c) and (5d) present a more complex picture. In (5c) there is a single set of three boys; however, each of the three boys is associated with a different set of two girls. Thus, while the subject NP has independent reference, the direct-object NP is referentially dependent on the subject NP. In this interpretation, then, the subject NP has scope over the direct-object NP. More precisely, the two NPs are in a relationship of distributivity, with the subject NP as distributive key, and the direct-object NP as *distributive share*. In (5d), a mirror-image situation obtains. Here, there is a single set of two girls; however, each of the two girls is associated with a different set of three boys. Accordingly, while the direct-object NP has independent reference, the subject NP is referentially dependent on the direct-object NP; the direct-object NP has scope over the subject NP. Here too then, a relation of distributivity obtains; this time, though, it is the direct object NP which is the distributive key and the subject NP which is the distributive share.

The above four interpretations are not equally readily available for speakers of English. Specifically, the symmetric interpretations (5a) and (5b) are more easily available than the asymmetric interpretations (5c) and (5d). Moreover, between the latter two interpretations, (5c) with wide scope for the subject NP, is more readily available than (5d) with wide scope for the direct-object NP. These preferences are not accidental; rather, they are particular consequences of more general principles governing the assignment of quantifier scope across languages.

Given the general preference for symmetric interpretations, languages typically have at their disposal various lexical and morphosyntactic devices whose function is to induce the more marked asymmetric interpretations, when these are required. Consider the following English sentences:

(6) (a) *All* the men carried three suitcases.
  (b) *Every* man carried three suitcases.

Sentences (6a) and (6b) differ in the lexical choice of quantifier within the subject NP. Whereas in (6a), with *all*, a variety of scope relationships may obtain between the two NPs, and the number of suitcases may vary between three in total and three per man, in (6b), with *every*, the subject NP has scope over the direct object NP, and the suitcases number three per man. Whereas *all*, like *three*, is a simple quantifier, *every* is a *scopal* quantifier, that is to say, a quantifier endowed with an additional denotational component, one that forces a particular scope relationship to obtain between the NP in which it occurs, and some other expression. In particular, it is a distributive-key quantifier, inducing a relationship of distributivity, in which the NP containing the quantifier in question is the distributive key, or element with wide scope (Gil 1995).

A different kind of scopal quantifier is exemplified by the following pair of Maricopa sentences:

(7) MARICOPA
  (a) *ʔipač* *ʔii* *xmokm*
    man-PL:NOM stick 3-three-SG-DS
    upaavk
    3-carry-PL-REAL
    'The men carried three sticks'
  (b) *ʔipač* *ʔii* *xmokxperm*
    man-PL:NOM stick 3-three-SG-DIST.SHARE-DS
    upaavk
    3-carry-PL-REAL
    (i) 'The men carried three sticks each'
    (ii) 'The men carried the sticks in threes'

Sentences (7a) and (7b) differ only in the form of the numeral within the direct-object expression. Sentence (7a) contains a simple numeral, *xmok-* 'three,' and is ambiguous in roughly the same ways as its English gloss. In contrast, sentence (7b) contains the same numeral plus the suffix -xper, whose effect is to mark the numeral *xmok-* and the expression containing it as distributive share. The ambiguity exhibited by sentence (7b) results from different expressions being selected as the distributive key counterpart. Under interpretation (7b(i)), the subject NP *ʔipač* 'men' is

chosen as distributive key: the sticks accordingly number three per man. Under interpretation (7b(ii)), the verb upaavk 'carried' is chosen as distributive key: here the sticks accordingly number three per carrying. Cross-linguistically, distributive-share numerals are formed by a variety of formal processes, including affixation (as in the above example), periphrasis, reduplication, and suppletion (Gil 1982a).

3.4 Quantifier as a Grammatical Category

Quantifier is primarily a semantic category: an expression is a quantifier only if it bears certain meaning properties. One of the most striking characteristics of quantifiers is the absence of any distinctive shared formal morphosyntactic features, which might justify the positing of a corresponding grammatical category of quantifier. Within languages, different quantifiers often display different grammatical properties. And across languages, the 'same' quantifier may exhibit different morphosyntactic behavior as well. Thus, in terms of their formal properties, quantifiers constitute a heterogeneous collection, exhibiting a great amount of morphological and syntactic diversity.

Looking within individual languages, there is probably no language within which there is a formal grammatical category consisting exactly of all quantifiers but no other expressions. Rather, in most or all languages, different quantifiers exhibit different arrays of grammatical properties, some grouping together with expressions belonging to one category, others patterning together with expressions belonging to some other category, yet others exhibiting idiosyncratic morphological or syntactic behaviour.

Looking across languages, quantifiers with the same meaning often exhibit quite different morphological and syntactic behaviour. Examples (1)–(4) above showed how the numeral 'three,' as a nominal quantifier, may occur, in different languages, in a variety of different constructions, exhibiting continuous and discontinuous quantification of various kinds. Even if attention is restricted to the most well-known type, that of continuous NP-internal quantification, as in (l), a great amount of cross-linguistic variation may be observed. In particular, in different languages, nominal quantifiers in attributive NP internal position may resemble, to various degrees, nouns, adjectives, and/or verbs.

4. Delimitation of Quantification

The discussion in the preceding sections was based on a bipartite definition of quantifiers as 'free-standing expressions' whose meanings 'involve the notion of quantity.' Like many similar definitions, this one is useful, but not unproblematical. Specifically, in addition to prototypical quantifiers such as those discussed above, there are numerous expressions and construction types for which it is difficult to adjudicate whether or not they may appropriately be characterized as quantificational.

The first part of the definition is formal: in order for an expression to be considered as a quantifier, it must be free standing—a word or a phrase, rather than an affix or some other bound morphological unit. Accordingly, various kinds of clitics, particles, and other intermediate units smaller than words but larger than affixes may provide borderline cases straddling the boundary between what is a quantifier and what is not. The distinction between free and bound forms is what underlies the distinction between quantification and the semantically-related grammatical category of *number*. Specifically, whereas quantifiers are free standing, number markers are bound forms. However, in some instances it may be difficult to adjudicate whether a form associated with a quantificational meaning is free or bound, and hence whether it constitutes an instance of quantification or of number. In order to determine whether forms such as these are quantifiers or number markers, it may be necessary to make use of additional diagnostic criteria distinguishing quantifiers from number markers.

The second part of the definition of quantifier is the obvious semantic condition: its meaning must involve the notion of quantity. While in many instances this is quite straightforward, in other cases, quantificational meanings may shade off gradually into other kinds of meanings not generally considered to be quantificational. One class of expressions related in meaning to nominal quantifiers are those denoting size and other scalar properties. An important class of expressions related to verbal quantifiers are those marking aspectual categories, such as perfect and progressive. More generally, different quantifiers in different languages may be related to an extremely broad range of non-quantificational concepts, synchronically and/or diachronically.

See also: Lexical Semantics; Logic and Linguistics; Numeral Systems; Semantics; Word Classes and Parts of Speech

Bibliography

Allan K 1980 Nouns and countability. *Language* **56**: 541–67

Bach E, Jeline K E, Kratzer A, Partee B H (eds.) 1995 *Cross-Linguistic Quantification*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Craig C (ed.) 1986 Noun classes and categorization. *Typological Studies in Language* **7**

Gil D 1982a Distributive numerals. PhD. dissertation, University of California, Los Angeles, CA

Gil D 1982b Quantifier scope, linguistic variation, and natural language semantics. *Linguistics and Philosophy* **5**: 421–72

Gil D 1993 Nominal and verbal quantification. *Sprachtypologie und Universalienforschung* **46**: 275–317

Gil D 1995 Universal quantifiers and distributivity. In: Bach E, Jeline K E, Kratzer A, Partee B H (eds.) *Quantification in Natural Languages*. Kluwer, Dordrecht, The Netherlands, pp. 321–62
Heim I 1982 The semantics of definite and indefinite noun phrases. PhD. dissertation, University of Massachusetts, Amherst, MA
Hurford J R 1975 *The Linguistic Study of Numerals*. Cambridge University Press, Cambridge, UK
Jackendoff R S 1971 *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA
Keenan E L, Stavi J 1986 A semantic characterization of natural language determiners. *Linguistics and Philosophy* **9**: 253–326
Menninger K 1969 *Number Words and Number Symbols* [trans. Broneer P.]. MIT Press, Cambridge, MA
Milsark G 1977 Toward an explanation of certain peculiarities in the existential construction in English. *Linguistic Analysis* **3**: 1–30
Pelletier F J (ed.) 1979 *Mass Terms: Some Philosophical Problems*. Synthese Language Library, Reidel, Dordrecht, The Netherlands
Vendler Z 1967 *Linguistics and Philosophy*. Cornell University Press, Ithaca, NY

D. Gil

Quantitative Cross-national Research Methods

1. The Main Challenges of Quantitative Nation Comparisons

Quantitative cross-national comparisons have become an important feature of all the social sciences. Their popularity has been spurred, in part, by the increasing availability of good data for a growing number of countries and, in part, by the diffusion of statistical skills among researchers. There are many constraints and potential pitfalls involved in cross-national statistical analyses, and these will be the main theme of this article. But this research strategy has, regardless, proven itself of great scientific value. It has, above all, helped highlight how fundamental many international variations are, thus, providing a powerful check against parochialism and unwaranted generalizations. Quantitative macrocomparisons have also proven themselves eminently useful in testing the validity of core theoretical assumptions, such as the convergence thesis that has long dominated development and modernization theory in political science and sociology especially.

Quantitative nation comparisons will, however, inevitably face a number of tradeoffs; some more serious than others. One is that the capacity for broader generalization that quantification permits is bought at the price of less attention to the contextual reality of individual nations. The uniqueness of a nation's culture, historical heritage, and endemic logic is lost easily when all we actually observe are a few indicators. Moreover, these indicators may mean somewhat different things in the unique context of specific countries (Ragin 1987). Scholars are frequently quite sensitive to this issue and so attempt to embed their statistical analyses within the histories and social uniqueness of the countries being studied. Also, as Ragin (1994) argues and empirically demonstrates, Boolean analysis can help overcome this dilemma by enabling us to build conjunctural models with very few cases and to analyze nonevents. But the approach is somewhat circumscribed. It needs to be guided by strong theory and background knowledge, its applicability is limited to relatively few cases, and it may be biased in favor of corroborating nonadditive, conjunctural models.

A second constraint has to do with the often limited number of observations available. This is especially true of studies of advanced (OECD) democracies where the N rarely exceeds 25. In broader worldwide comparisons, however, the N approaches 200. Many researchers attempt to supplement the paucity of nations by collecting longitudinal data on each nation, as with pooled cross-sectional and time series analyses. A 20-year time series for 20 countries yields 400 observations; a 20-year time series for 200 nations yields 4,000. As ever, longer data series for individual countries become available, the small N problem will gradually diminish.

Small samples do not necessarily pose problems of statistical inference, especially when the substantive theory is strong. They do, however, limit what statistical tools can be brought to bear. Fearon (1991) argues that the smaller the sample size, the greater is the need to make counterfactuals explicit—in other words, to tighten the theoretical formulation about the precise conditions that will (or will not) produce an outcome. Vague theory implies uncertainty about the number of contending explanations (which can grow very large) and about their mutual relationship within a causal order. The consequences of this include causal mispecification, multi-colinearity and large error terms. The last two of these can only be managed by having more observations. Without them, estimations are likely to yield results that are not robust. The less ambiguous the theory, the less the need for large N's. Small N's are, therefore, an especially acute problem in disciplines without a firm theoretical architecture, like sociology or political science. Western (1998a) argues that Bayesian estimation, by allowing for greater uncertainty, produces more robust results when N's are few, theory vague, and explanations are many.

Small N's can even hold advantages, because a limited nation sample implies that the researcher can gain more easily maximum advantage from diagnostic scrutiny of the residual plots. A set of systematic outliers, for example, can be easily identified in terms

of their nationhood, and any good comparativist should be able to pin down what variable(s) drives their deviance from the fitted regression line. Diagnostics on small samples can approach the advantages of in-depth case studies.

The constraints imposed by small samples are generally well known to researchers. There are additional problems inherent in quantitative national comparisons that are both more generic and potentially more serious, but also less recognized. Chief among these are issues related to: qualitative and limited dependent variables, the potential for selection bias, possible lack of independence between observations and, finally, the question of endogeneity of variables.

2. *Qualitative and Limited Dependent Variables*

Many dependent variables in cross-national research are either qualitative (assuming discrete values) or limited (they assume continuous values within some range). The choice whether to measure a variable in a qualitative or continuous way is often controversial. Bollen and Jackman (1989), for example, argue that difficulties in classifying some political regimes speak in favor of using continuous scales because 'Dichotomizing democracy blurs distinctions between borderline cases.' In contrast, Przeworski et al. (in press) prefer to treat regimes dichotomously or multinomially.

When the dependent variable is either qualitative or limited, we need to use nonlinear models. Whether we give political regimes the values of 0–1 (as do Przeworski et al. in press) or 1–100 (as does Bollen 1980), it remains that the value on the dependent variable cannot exceed its maximum when the independent variable(s) tend to infinity, and it cannot fall below its minimum when these variables tend to infinity. Linear models, at best, can provide an approximation within some range of the independent variables.

The standard model when the dependent variable is multinomial is

$$\Pr(Y = j,\ X = x) = F(x) \qquad (1)$$

where $j = 0, 1, \ldots, J-1$ and F is the cumulative distribution function. Since such models are covered by any standard textbook we need not present them here. Such models can be applied to panel data unless the number of repeated observations is large.

'Event history analysis' is one particular class of nonlinear models applied in cross-national research. In such models the dependent variable is an event, such as a revolution, regime transition, or a policy adoption. The general model is

$$\Pr(Y(t+\mathrm{d}t) = j) = f(y(t),\ x(t)),\ \mathrm{d}t \quad 0 \qquad (2)$$

Most often, such models can be conveniently estimated as

$$\log S(t) = \log(1 - F(t)) \qquad (3)$$

where $S(t)$ is the survival function, or the probability that an event lasts beyond time t, and $F(t)$ is the cdf.

For dichotomous dependent variables, logit and probit give very similar results. For multinomial variables, it is often assumed that the errors are independent across the values of the dependent variable, which leads to a logit specification. But this implies a strong assumption, namely the irrelevance of independent alternatives that rarely holds in practice. Multinomial probit, in turn, requires computing multiple integrals, which was until recently computationally expensive. Alternatives would be semi- and nonparametric methods.

The distributions that are commonly used in estimating survival models include the exponential, Weibul, logistic, and Poisson distributions. But, except for the Weibul and exponential, it is very difficult to distinguish statistically between such distributions.

Methods for studying qualitative dependent variables are now standard textbook fare, but what warrants emphasis is that the traditional distinction between 'qualitative' and 'quantitative' research is becoming increasingly obsolute. Phenomena such as social revolutions may be rare, but this just means they occur with a low probability. They can, nevertheless, be studied systematically using maximum likelihood methods.

3. *The Problem of Selection Bias*

In comparative research we are often interested in the effect of some systemic feature, institution, or policy on some outcome. Examples include the effect of labor market institutions on unemployment, the effect of political regimes on economic growth, or the effect of electoral rules on the number of political parties.

In such cases, the question is how some feature of X affects some outcomes Y in the presence of conditions Z, or $PR\ (Y, X, Z)$, where the hypothesis is that $y = f(x)$. The generic problem is how to isolate the effects of X and Z on Y.

If sampling is exogenous (i.e., the probability of observing any y is independent of z), the $pr(Y = yz) = pr(y)$, and standard statistical methods can be used. But if $pr(Y = yz)$ is different under different conditions Z, sampling is endogenous and the assumptions of the standard statistical model are violated.

To exemplify the problem: suppose we want to know the impact of political regimes on economic growth. We observe Chile in 1985 (Z), which was a dictatorship (X), in which case *per capita* income declined at the rate of $Y = -2.26$. To assess the effect of political regimes, we need to know what would have

been the rate of growth (Y) in Chile had it been a democracy. This cannot be answered with the available observations because Chile in 1985 was indeed a dictatorship. The comparativist in such a case would proceed quasi-experimentally, and look for a case that matches Chile in all aspects except authoritarianism. In this way one would compare an authoritarian with a democratic 'Chile.'

But what if Chile's status as a dictatorship in 1985 was due to some factors that also affected its economic performance? This could be because growth itself influences the survival of regimes; a country's level of development may be associated both with regime selection and growth; some unobservable factors (say 'enlightened leadership' or, perhaps, measurement error) may be common to both variables. Whatever the underlying selection mechanism, the basic consequence is that there will be cases without a match, and this implies that our inferences will be biased.

The problem raised by nonrandom selection is how to make inferences from what we observe to what we do not. It is a problem of identification (Manski 1995). For what do we know when we observe Chile as a dictatorship in 1985? We observe the fact that, given conditions Z, the Chilean regime (X) is a dictatorship. We also know its rate of economic growth, given that it is a dictatorship under conditions Z. We do not observe, however, its rate of growth as a democracy under conditions Z. The issue is one of counterfactual observations: the rate of growth that an observation with $Z = zi$, observed under dictatorship, would have had under democracy. We do not know what this value is and, hence, we face under-identification.

What we need to know are two distributions, $P(Y = y,\ x = j,\ Z)$ and $P(Y = y,\ x = k,\ Z)$, and their expected values, where we now think of the possible values of X more generally as $j, k = 0, 1, 2, \ldots, N$. The first is the distribution of Y in the entire population if all cases were observed as dictatorships; the second is the distribution of Y in the entire population if all cases were observed as democracies. Clearly, if we want to know the impact on y of being in states x, we need to determine the difference between these two distributions, conditional on z.

It may seem strange to refer to the 'population' as including unobserved cases. But in order to compare the effects of states x on the performance of an individual case characterized by a zi, we must allow the case to be potentially observable under the full range of X. We observed this case as $X = j$, but must also imagine the possibility that it would have been Xj. We must think in terms of a 'superpopulation' consisting of a continuum of potential cases. The actual sample is then regarded as having been drawn by 'nature' from this superpopulation.

If the actual population is an exogenous sample of the potential one, any multivariate combination (xi, zi) can be drawn with a positive probability. Such exogenous sampling allows us to isolate the effect of x on y given z, that is, to 'control' for z the effects of x on y. If the sample is exogenous, we can 'match' pairs ($xi = j$, zi) and ($xi = k$, zi). But if sampling is endogenous, there will be some cases without a 'match': some of the xi's will not have support. Where we have nonrandom selection, quasi-experimental comparisons fail regardless of the number of observations.

The methods to correct for selection bias consist of constructing counterfactuals, that is, of filling the unobserved supports of the distribution Y for all X. Following Heckman (1979), when sampling is not random, the regressions $y = f(x, z)$ suffer from omitted variable bias. When such regressions are estimated on the basis of the observed sample, the variable that is omitted is the expected value of the error in the equation that specifies how the observations are selected into the sample

$$\Pr(x = j) = F(Z) + u \tag{4}$$

If

$$E(uj\ x = j)\ 0 \tag{5}$$

then

$$E(uj\ x) = juE(u\ x) \tag{6}$$

where ju is a regression coefficient of uj on u.

Finally, $E(ux = j) = j$, where the j's are the inverse Mill ratios, or the hazard rates. Since we know that Yj is observed when $x = j$, we can write the expected values in the observed sample as

$$E(\,yj\ x = j,\ z) = zj + ju\ j \tag{7}$$

Note that if this equation is estimated on the basis of the observed sample, the variable j is omitted from the specification.

We can now see why controlling for the variables that enter both into selection and outcome equations may indeed worsen the selection bias. Following Heckman (1988), we distinguish first between selection on observables and on unobservables. Selection on observables occurs when the expected covariance $E(uju\ z)$ 0, but once the observed variables Z are controlled for it vanishes, so that $E(uju\ z) = 0$. Selection is on unobservables when $E(uju)$ 0 and $E(uju\ z)$ 0, which means that controlling for the factors observed by the investigator does not remove the covariance between the errors in the outcome and the selection equations. Now note that the regression coefficient $ju = \mathrm{cov}(uju)/\mathrm{var}(uj)$. If selection is on unobservables, controlling for some variable x in the outcome equation may reduce the error variance uj without equally reducing the covariance uju. Hence, the coefficient on the omitted variable will be larger and the bias will be exacerbated.

In short, the expected values of the observed cases will be biased because they co-vary with the variable which determines which cases are observed. If selection is exclusively on observables, this bias can be corrected by traditional controlling techniques. But if it is on unobservables, such controls only worsen the bias.

Correcting for selection bias is not uncontroversial. It has been found that corrections for selection are not robust, but are highly sensitive to relatively minor changes in assumptions about distributions (Goldberger 1983). Others have found that some estimation methods fail to correct for this bias and may even exacerbate it (Stolzenberg and Relles 1990). As Heckman (1988) argues, the quandary we face is that different methods of correcting for selection bias are robust if there is no bias to begin with; if there is, there is no guarantee that the methods are robust.

The logic of the problem is similar whether we study large or small N's. When de Toqueville concluded that revolutions do not bring about social change, the reason might be that they occur only in countries where it is difficult to change society. Even when $N = 1$, the issue of selection bias does not disappear: The French revolution in 1789 may have been caused by the same conditions that made social change so difficult. It is possible that a revolution in a country where social relations are easier to change would have provoked change. But then a revolution would not have been necessary.

Comparativists who conduct case studies cannot benefit from statistical distributions to generate the counterfactuals. Yet, as in all cases of comparison the problem remains and it should, therefore, be standard practice to ask counterfactual questions of one's case or cases.

4. *The Problem of Independence*

Statistical inference must assume that the observations on a variable are independent one of the other. Is country A's performance truly independent of what happens in country B? Is what happens at $t+1$ independent of events in t? Usually not, and this implies the need for corrective procedures. In most cases, however, rigorous correction will entail that the *de facto N* (nations or years) diminishes; in some instances, statistical dependency cannot be resolved at all.

Cross-sectional analysis almost invariably assumes that nations and their properties (say budgets or institutions) are independent one of the other. This should not be assumed. We know that the Scandinavian countries have a shared history, deliberately learning from each other through centuries, thus creating similar institutions and path dependencies. The same goes for Austria and Germany, for Belgium and The Netherlands and, arguably, for all the Anglo-saxon nations. World samples have a similar problem: Japan's long hegemony in East Asia will have influenced Korean society; Confucianism has had a pervasive influence throughout the region. Similar stories are easily told for Latin America and Africa. If the World is a set of nation clusters, the real N is not 20-odd OECD countries or 150-odd World nations.

When nations form families, but are treated as if they were all unique and independent, we are likely to get biased coefficients and, very probably, unequal error variance (heteroskadicity). Sweden alone will drive the regression line in just about any welfare state analysis, and when also Denmark and Norway are treated as discrete observations, the bias is multiplied in so far as all three in reality form part of the same political family ('Scandinavia'). Diffusion effects that operate between members of a nation–cluster can also result in heteroskadistic disturbance in the cross-section. Such can be corrected by, for example, adding a variable that captures the common underlying property that drives the disturbance (say, a dummy for being 'Scandinavia') but, again, this correction absorbs precious degrees of freedom in a small N study and, substantively, amounts to reducing the three nations to one observation.

Lack of independence in a time-series is normally taken for granted, since this year's budget or election outcome is almost inevitably related to last year's budget or the previous election. The standard assumption is a first-order (AR1) serial correlation. In comparative research virtually all time-series applications are pooled with cross-sections. But, where N's are very small, one may as well simply compare across individual time-series estimations, as do Esping-Andersen and Sonnberger (1991). Time-series are meant to capture historical process. Yet as Isaac and Griffin (1989) argue, they easily end up being a-historical.

Pooling cross-sectional with time-series data (panel regressions) has become very widespread, especially in studies of the limited group of advanced (OECD) societies. In many cases, the panel design is chiefly cross-sectional (more nations than years); others are temporally dominated (for a discussion, see Stimson 1985). Panel models are especially problematic because they can contain simultaneous diachronic and spatial interdependence and, worse, the two may interact. The standard method for correcting contemporaneous error correlation (GLS) applies only where the t's well exceed nations (which is rare). The consequence is that t-statistics are overestimated, errors underestimated, and the results may therefore not be robust (Beck and Katz 1995). The Beck and Katz (1995) procedure can correct for temporal and cross-sectional dependency one at a time, but if the two interact, no solution exists. See also Beck et al. (1998) for an application to maximum likelihood estimation.

Panel models can be based on two types of theoretical justification. One is that events or shocks occur over time that affect the cross-sectional variance.

There is, for example, a huge recent literature on the impact of labor market 'rigidities' on unemployment: regulations vary across nations but also across time because of deregulatory legislation (see, for example, Nickell 1997). Deregulation in a country should produce a break in its time series, and the autocorrelation element will be split into the years preceding and following the break. The second justification, not often exploited, is to interpret autocorrelation as an expression of institutional or policy path dependency. In this instant, the rho must be treated as a variable. The problem, of course, is that the rho is likely to combine theoretically relevant information as well as unknown residual autocorr elation. The researcher can accordingly not avoid including a variable that explicitly measures path dependency.

If we insist on faithful adherence to the real world, panel regressions may require so much correction against dependency that the hard-won additional degrees of freedom that come with a time-series are easily eaten up. And how many can truthfully claim that time and country dependencies do not interact? Indeed, most sensible comparativists would assume they do: if nations form part of families it should also be the case that the timing of their shocks, events, or policies is interdependent. Such intractable problems are certainly much more severe in small-N comparisons. Here, a marginal difference in measurement, the inclusion or exclusion of one country, the addition or subtraction of a year here or there, or the substitution of one variable for another, can change the entire model.

One alternative is to construct multilevel models which explicitly take into account the possibility that nations may 'cluster' (for an overview, see Goldstein 1987). Nieuwbeerta and Ultee (1999) have, for example, estimated a three level (nation, time, and individual) model of the impact of class on party choice within the context of nations' social mobility structure. A second alternative, in particular when the dependent variable is categorical, is to exploit the advantages of event history analysis. But, here the time series needs to be quite long considering that theoretically interesting events, such as revolutions, democratization, or even welfare reforms, are far between. In the event history context, analytical priority usually is given to temporal change, which brings it much closer to traditional time series analysis. But rather than having to manipulate autocorrelation, time sequencing (states and events) is actively modeled and thus gains analytic status. For an application to nation comparisons, see, for example, Western (1998b), which also can stand as an exemplar of how to minimize the interdependency problem.

There are two particular cases where the lack of independence among observations simply prohibits adequate estimation. The first, noted above, occurs when time and nation dependencies interact. The second is when 'globalization' penetrates all nations and when many nations (such as the European Union) become subsumed under identical constraints. In this instance, existing cross-national correlations will strengthen and we may, indeed be moving towards an $N = 1$. Of course, global shocks or European Union membership do not necessarily produce similar effects on the dependent variable across nations or time. If nations' institutional filters differ, so will most likely the impact of a global shock on, say, national unemployment rates. Here we would specify interaction effects, but that would be impossible in a pure cross-section, and extremely difficult in a time series, unless we already know how the lag structure will differ according to institutional variation.

5. *The Endogeneity Problem*

All probabilistic statistics require conditional independence, namely that the values of the predictor variables are assigned independently of the dependent variable. The basic problem of endogeneity occurs when the explanans (X) may be influenced by the explanandum (Y) or both may be jointly influenced by an unmeasured third. The endogeneity problem is one aspect of the broader question of selection bias discussed earlier.

The endogeneity issue has been debated intensely within the economic growth literature in terms of the causal relationship between technology and growth. But it applies equally to many fields. For example, comparativists often argue that left power explains welfare state development. But are we certain that left power, itself, is not a function of strong welfare states? Or, equally likely, are both large welfare states and left power just two faces of the same coin, different manifestations of one underlying, yet undefined, phenomenon? Would Sweden have had the same welfare state even without its legendary social democratic tradition? Perhaps, if Sweden's cultural past overdetermines its unique kind of social democracy and social policy. If this kind of endogeneity exists, the true X for Sweden is not left power but a full list of all that is 'Sweden.' The vector of the X's becomes a list of all that is nationally unique.

The endogeneity problem becomes easily intractable in quantitative cross-national research because we observe variables (Y's and X's) that represent part of the reality of the nations we sample. Our variables are in effect a partial reflection of the society under study, and the meaning of a variable score for one nation may not be metrically equivalent to that of another—a weighted left cabinet score of 35 for Denmark and 40 for Sweden probably misrepresents the Danish–Swedish difference if, that is, the two social democracies are two different beasts.

A related, and equally problematic, issue arises in the interpretation of coefficient estimations. In the

simple cross-sectional regression, the coefficient for left power, economic development or for coordinated bargaining can only be interpreted in one way: as a constant, across-the-board effect irrespective of national context. If Denmark had 5 points more left power, its welfare state should match the Swedish (all else held constant). In fixed-effects panel regressions, the X's are assumed to have an identical impact on Y, irrespective of country. Often we cannot accept such assumptions.

The reason that we cannot is that it is difficult to assume that variables, say a bargaining institution or party power balance, will produce homogenous, monotonically identical, effects across nations for the simple reason that they are embedded in a more complex reality (called nation) which has also given rise to its version of the dependent variable. In this case, the bias remains equally problematic whether we study few or many N's; it will not disappear if we add more nations. Quantitative national comparisons rarely address such endogeneity problems. Studies of labor market or economic performance routinely presume that labor market regulations or bargaining centralization are truly exogenous variables, whose effects on employment is conditionally identical whether it is Germany, Norway, or the United States. The same goes for welfare state comparisons with their belief that demography and left power are fully exogenous and conditionally unitary.

There exist several, not necessarily efficient, ways of correcting for the endogeneity bias. If we have some inkling that the bias comes from variable omission, the obvious correction entails the inclusion of additional controls. An example of this kind was Jackman's (1986) argument that Norway's North Sea Oil was over-determining the results in the Lange and Garrett (1985) study. Small N studies with strong endogeneity have little capacity to extend the number of potentially necessary controls. A second approach is to limit endogeneity in X by reconceptualizing and, most likely, narrowing Y. The welfare state literature provides a prototypical example: aggregate social expenditure increasingly was replaced by measures of specific welfare state traits.

Controls, no matter how many, will however not resolve the problem under conditions of strong sampling bias. As discussed above, the best solution in such a situation is to concentrate more on the theoretical elaboration of causal relations between variables. If we can assume that our estimations are biased because Y affects the values on X, or because both are jointly attributable to a third underlying force, thinking in counterfactual terms ('would Sweden's welfare state be the same without Sweden's social democracy') will force the researcher to identify more precisely the direct or derived causal connections (Fearon 1991). If bias can be assumed to come from the assumption of monotonically homogeneous effects of the X across all nations, the researcher's attention should concentrate on identifying more precisely the conditional mechanisms that are involved in the causal passage from an X to a Y (why and how will a 5-point rise in left power make the Danish welfare state converge with the Swedish?).

6. Conclusion

The difficulties and tradeoffs that we have synthesized above are intrinsic to all quantitative research, but are often more difficult to deal with in national comparisons because of the inevitability of small samples of countries and of the ways in which nations have developed together in interaction and will continue to do so even more in the future, given globalization. However, the difficulties should not be seen as paralyzing research but rather as stimulating attempts to deal with them. Indeed, in most practical situations today, each of these problems can be dealt with, albeit not perfectly. So, today's partial vices may be the impetus to tomorrow's virtues. Although we may never be able to completely resolve all problems, ongoing trends suggest that our capacity to conduct good quality research in this field is improving rapidly. For one, the constraints of small N's are diminishing because of rapid progress in data quality and quantity, covering both more years and nations. Second, the field has made great strides in terms of researchers' substantial comparative knowledge of the nations and cultures under study. Hence, our ability to interpret and contextualize statistical findings has greatly strengthened. And this, in turn, is a major precondition for what we believe is the single most effective tool with which to manage the challenges of quantitative comparison: namely, to systematically think in terms of counterfactuals.

Bibliography

Amemyia T 1985 *Advanced Econometrics*. Harvard University Press, Cambridge, MA

Beck N, Katz J N 1995 What to do (and not to do) with times-series, cross-section data. *American Political Science Review* **89**: 634–47

Beck N, Katz J N, Tucker R 1998 Taking time seriously: Time series-cross-section analysis with a binary dependent variable. *American Journal of Political Science* **42**: 1260–88

Blossfeld H P, Roehwer G 1995 Techniques of Event History Modelling: New Approaches to Causal Analysis. L. Erlbaum, Mahwah, NJ

Bollen K A 1980 Issues in the comparative measurement of political democracy. *American Sociological Review* **45**: 370–90

Bollen K A, Jackman R W 1989 Democracy, stability, and dichotomies. *American Sociological Review* **54**: 612–21

Esping-Andersen G, Sonnberger H 1991 The demographics of age in labor market management. In: Myles J, Quadagno J (eds.) *States, Labor Markets, and the Future of Old-Age Policy*. Temple University Press, Philadelphia, PA, pp. 227–49

Fearon J D 1991 Counterfactuals and hypothesis testing in political science. *World Politics* **43**: 169–95
Goldberger A 1983 Abnormal selection bias. In: Karlin S, Amemyia T, Goodman L A (eds.) *Studies in Econometrics, Time Series, and Multivariate Statistics*. Academic Press, New York
Goldstein H 1987 *Multilevel Models in Educational and Social Research*. Oxford University Press, New York
Heckman J J 1988 The microeconomic evaluation of social programs and economic institutions. In: *Chung-Hua Series of Lectures by Invited Eminent Economists no. 14*. The Institute of Economics, Academia Sinica, Teipei, China
Hsiao C 1986 *Analysis of Panel Data*. Cambridge University Press, Cambridge, UK
Isaac L W, Griffin L J 1989 A historicism in time-series analyses of historical process. *American Sociological Review* **54**: 873–90
Jackman R 1986 The politics of economic growth in industrial democracies, 1974–1980: Leftist strength or North Sea oil? *Journal of Politics* **49**: 242–56
Janoski T, Hicks A (eds.) 1994 *The Comparative Political Economy of the Welfare State*. Cambridge University Press, Cambridge, UK
Lange P, Jarrett G 1985 The politics of growth. *Journal of Politics* **47**: 792–827
Maddala G S 1983 *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge, UK
Nickell S 1997 Unemployment and labor market rigidities: Europe versus North America. *Journal of Economic Perspectives* **11**(3): 55–74
Nieuwbeerta P, Ultee W 1999 Class voting in Western industrialized countries, 1945–1990: systematizing and testing explanations. *European Journal of Political Research* **35**: 123–60
Przeworski A, Alvarez M R, Cheibub J A, Limongi F in press*Democracy and Development: Political Regimes and Material Welfare in the World, 1950–1990*. Cambridge University Press, New York
Ragin C 1987 *The Comparative Method*. University of California Press, Berkeley, CA
Ragin C 1994 A qualitative comparative analysis of pension systems. In: Janoski T, Hicks A M (eds.) *The Comparative Political Economy of the Welfare State*. Cambridge University Press, Cambridge, UK, pp. 320–45
Stimson J A 1985 Regression in time and space: A statistical essay. *American Journal of Political Science* **29**: 914–47
Western B 1998a Causal heterogeneity in comparative research: A Bayesian hierarchical modeling approach. *American Journal of Political Science* **42**: 1233–59
Western B 1998b *Between Class and Market: Postwar Unionization in the Capitalist Democracies*. Princeton University Press, Princeton, NJ

G. Esping-Andersen and A. Przeworski

Quasi-Experimental Designs

The theory of quasi-experimentation addresses one of the great unanswered questions in the social and behavioral sciences: how can researchers make a valid causal inference when units cannot be randomly assigned to conditions? Like all experiments, quasi-experiments deliberately manipulate presumed causes to discover their effects. However, quasi-experiments are also defined by what they do not do: the researcher does *not* assign units to conditions randomly. Instead, quasi-experiments use a combination of design features, practical logic, and statistical analysis to show that the presumed cause is likely to be responsible for the observed effect, and other causes are not. The term nonrandomized experiment is synonymous with quasi-experiment; and the terms observational study and nonexperimental design often include quasi-experiments as a subset. This article discusses the need for quasi-experimentation, describes the kinds of designs that fall into this class of methods, reviews the intellectual and practical history of these designs, and notes important current developments.

1. *The Need for Quasi-experimentation*

Given the desirable properties of randomized experiments, one might question why quasi-experiments are needed. When properly implemented, randomized experiments yield unbiased estimates of treatment effects, accompanied by known probabilities of error in identifying effect size. Quasi-experimental designs do not have these properties. Yet quasi-experiments are necessary in the arsenal of science because it is not always possible to randomize. Ethical constraints may preclude withholding treatment from needy people based on chance, those who administer treatment may refuse to honor randomization, or questions about program effects may arise after a treatment was already implemented so that randomization is impossible. Consequently, the use of quasi-experimental designs is frequent and inevitable in practice.

2. *Kinds of Quasi-experimental Designs*

The range of quasi-experimental designs is large, including but not limited to: (a) Nonequivalent control group designs in which the outcomes of two or more treatment or comparison conditions are studied but the experimenter does not control assignment to conditions; (b) Interrupted time series designs in which many consecutive observations over time (prototypically 100) are available on an outcome, and treatment is introduced in the midst of those observations to demonstrate its impact on the outcome through a discontinuity in the time series after treatment; (c) Regression discontinuity designs in which the experimenter uses a cutoff score on a measured variable to determine eligibility for treatment, and

an effect is observed if the regression line (of the assignment variable on outcome) for the treatment group is discontinuous from that of the comparison group at the cutoff score; (d) Single-case designs in which one participant is repeatedly observed over time (usually on fewer occasions than in time series) while the scheduling and dose of treatment are manipulated to demonstrate that treatment controls outcome.

In the preceding designs, treatment is manipulated, and outcome is then observed. Two other classes of designs are sometimes included as quasi-experiments, even though the presumed cause is not manipulated (and often not even manipulable) prior to observing the outcome. In (e) case–control designs, a group with an outcome of interest is compared to a group without that outcome to see if they differ retrospectively in exposure to possible causes in the past; and in (f) correlational designs, observations on possible treatments and outcomes are observed simultaneously, often with a survey, to see if they are related. Because these designs do not ensure that cause precedes effect, as it must logically do, they usually yield more equivocal causal inferences.

3. *The History of Quasi-experimental Designs*

Quasi-experimental designs have an even longer history than randomized experiments. For example, around 1,850 epidemiologists used case–control methods to identify contaminated water supplies as the cause of cholera in London (Schlesselman 1982), and in 1898, Triplett used a nonequivalent control group design to show that the presence of audience and competitors improved the performance of bicyclists. In fact, nearly all experiments conducted prior to Fisher's work were quasi-experiments.

However, it was not until 1963 that the term quasi-experiment was coined by Campbell and Stanley (1963) to describe this class of designs. Campbell and his colleagues (Cook and Campbell 1979, Shadish et al. in press) extended the theory and practice of these designs in three ways. First, they described a large number of these designs, including variations of the designs described above. For example, some quasi-experimental designs are inherently longitudinal (e.g., time series, single case designs), observing participants over time, but other designs can be made longitudinal by adding more observations before or after treatment. Similarly, more than one treatment or control group can be used, and the designs can be combined, as when adding a nonequivalent control group to a time series.

Second, Campbell developed a method to evaluate the quality of causal inferences resulting from quasi-experimental designs—a validity typology that was elaborated in Cook and Campbell (1979). The typology includes four validity types and threats to validity for each type. Threats are common reasons why researchers may be wrong about the causal inferences they draw. Statistical conclusion validity concerns inferences about whether and how much presumed cause and effect co-vary; examples of threats to statistical conclusion validity include low statistical power, violated assumptions of statistical tests, and inaccurate effect size estimates. Internal validity concerns inferences that observed co-variation is due to the presumed treatment causing the presumed outcome; examples include history (extraneous events that could also cause the effect), maturation (natural growth processes that could cause an observed change), and selection (differences between groups before treatment that may cause differences after treatment). Construct validity concerns inferences about higher-order constructs that research operations represent; threats include experimenter expectancy effects whereby participants react to what they believe the experimenter wants to observe rather than to the intended treatment, and mono-operation bias in which researchers use only one measure that reflects a construct imperfectly or incorrectly. External validity concerns inferences about generalizing a causal relationship over variations in units, treatments, observations, settings, and times; threats include interactions of the treatment with other features of the design that produce unique effects that would not otherwise be observed.

Third, Campbell's theory emphasized addressing threats to validity using design features—things that a researcher can manipulate to prevent a threat from occurring or to diagnose its presence and potential impact on study results (see Table 1). For example, suppose maturation (normal development) is an anticipated threat to validity because it could cause a pretest–post-test change like that attributed to the treatment. The inclusion of several consecutive pretests before treatment can indicate whether the rate of maturation before treatment is similar to the rate of change from during and after treatment. If it is similar, maturation is a threat. All quasi-experiments are combinations of these design features, thoughtfully chosen to diagnose or rule out threats to validity in a particular context. Conversely, Campbell was skeptical about the more difficult task of trying to adjust threats statistically after they have already occurred. The reason is that statistical adjustments require making assumptions, the validity of which are usually impossible to test, and some of which are dubious (e.g., that the selection model is known fully, or that the functional form of errors is known).

Other scholars during this time were also interested in causal inferences in quasi-experiments, such as Cochran (1965) in statistics, Heckman (1979) in economics, and Hill (1953) in epidemiology. However, Campbell's work was unique for its extensive emphasis on design rather than statistical analysis, for its theory of how to evaluate causal inferences, and for its sustained development of quasi-experimental theory

Table 1
Design elements used in constructing quasi-experiments

Assignment (Control of assignment strategies to increase group comparability)
- *Cutoff-based assignment*. Controlled assignment to conditions based solely on one or more fully measured covariates. This can yield an unbiased effect estimate
- Other *nonrandom assignment*. Various forms of 'haphazard' assignment that sometimes approximate randomization (e.g., alternating assignment in a two condition quasi-experiment whereby every other unit is assigned to one condition, etc.)
- *Matching and stratifying*. Efforts to create groups equivalent on observed covariates in ways that are stable, do not lead to regression artifacts, and are correlated with the outcome. Preference is for pretreatment measures of the outcome itself

Measurement (Use of measures to learn whether threats to causal inference actually operate)
- Post-test observations
  - *Nonequivalent dependent variables*. Measures that are not sensitive to the causal forces of the treatment, but are sensitive to all or most of the confounding causal forces that might lead to false conclusions about treatment effects (if such measures show no effect, but the outcome measures do show an effect, the causal inference is bolstered because it is less likely due to the confounds)
  - *Multiple substantive post-tests*. Used to assess whether the treatment affects a complex pattern of theoretically predicted outcomes
- Pretest observations
  - *Single pretest*. A pretreatment measure on the outcome variable, useful to help diagnose selection bias
  - *Retrospective pretest*. Reconstructed pretests when actual pretests are not feasible—by itself, a very weak design feature, but sometimes better than nothing
  - *Proxy pretest*. When a true pretest is not feasible, a pretest on a variable correlated with the outcome—also often weak by itself
  - *Multiple pretest time points on the outcome*. Helps reveal pretreatment trends or regression artifacts that might complicate causal inference
  - *Pretests on independent samples*. When a pretest is not feasible on the treated sample, one is obtained from a randomly equivalent sample
- *Complex predictions* such as *predicted interactions*. Successfully predicted interactions lend support to causal inference because alternative explanations become less plausible
- *Measurement of threats to internal validity*. Help diagnose the presence of specific threats to the inference that A caused B such as whether units actively sought out additional treatments outside the experiment

Comparison groups (Selecting comparisons that are 'less nonequivalent' or that bracket the treatment group at the pretest(s))
- *Single nonequivalent groups*. Compared to studies without control groups, using a nonequivalent control group helps identify many plausible threats to validity
- *Multiple nonequivalent groups*. Serve several functions. For instance, groups are selected that are as similar as possible to the treated group but at least one outperforms it initially and at least one underperforms it, thus bracketing the treated group
- *Cohorts*. Comparison groups chosen from the same institution in a different cycle (e.g., sibling controls in families or last year's students in schools)
- *Internal* (*vs. external*) *controls*. Plausibly chosen from within the same population (e.g., within the same school rather than from a different school)

Treatment (Manipulations of the treatment to demonstrate that treatment variability affects outcome variability)
- *Removed treatments*. Showing an effect diminishes if treatment is removed
- *Repeated treatments*. Reintroduces treatments after they have been removed from some group—common in laboratory sciences or where treatments have short-term effects
- *Switching replications*. Reverses treatment and control group roles so that one group is the control while the other receives treatment, but the controls receive treatment later while the original treatment group receives no further treatment or has treatment removed
- *Reversed treatments*. Provides a conceptually similar treatment that reverses an effect—e.g., reducing access to a computer for some students but increasing access for others
- *Dosage variation*. Demonstrates that outcome responds systematically to different levels of treatment

and method over four decades. Both the theory and the methods he outlined were widely adopted in practice during the last half of the twentieth century, and his terms like internal and external validity became so much a part of the scientific lexicon that today they are often used without reference to Campbell.

4. Contemporary Research about Quasi-experimental Design

4.1 Statistics and Quasi-experimental Design

Although work on the statistical analysis of quasi-experimental designs deserves separate treatment, several contemporary developments deserve mention here. One is the work of statisticians such as Paul Holland, Paul Rosenbaum, and Donald Rubin on statistical models for quasi-experimental designs (e.g., Rosenbaum 1995). They emphasize the need to measure what would have happened to treatment participants without treatment (the counterfactual), and focus on statistics that can improve estimates of the counterfactual without randomization. A central method uses propensity scores, a predicted probability of group membership obtained from logistic regression of actual group membership on predictors of outcome or of how participants got into treatment. Matching, stratifying, or co-varying on the propensity score can balance nonequivalent groups on those predictors, but they cannot balance groups for unobserved variables, so hidden bias may remain. Hence these statisticians have developed sensitivity analyses to measure how much hidden bias would be necessary to change an effect in important ways. Both propensity scores and sensitivity analysis are promising developments warranting wider exploration in quasi-experimental designs.

A second statistical development has been pursued mostly by economists, especially James Heckman and his colleagues, called selection bias modeling (Winship and Mare 1992). The aim is to remove hidden bias in effect estimates from quasi-experiments by modeling the selection process. In principle the statistical models are exciting, but in practice they have been less successful. A series of studies in the 1980s and 1990s found that effect estimates from selection bias models did not match results from randomized experiments. Economists responded with various adjustments to these models, and proposed tests for their appropriate application, but so far results remain discouraging. Most recently, some economists have improved results by combining selection bias models with propensity scores. Although selection bias models cannot yet be recommended for widespread adoption, this topic continues to develop rapidly and serious scholars must attend to it. For example, other economists have developed useful econometrically based sensitivity analyses. Along with the incorporation of propensity scores, this may promise a future convergence in statistical and econometric literatures.

A third development is the use of structural equation modeling (SEM) to study causal relationships in quasi-experiments, but this effort has also been only partly successful (Bollen 1989). The capacity of SEM to model latent variables can sometimes reduce problems of bias caused by unreliability of measurement, but its capacity to generate unbiased effect estimates is hamstrung by the same lack of knowledge of selection that thwarts selection bias models.

4.2 The Empirical Program of Quasi-experimental Design

Many features of quasi-experimentation pertain to matters of empirical fact that cannot be resolved by statistical theory or logical analysis. For these features, a theory of quasi-experimental design benefits from empirical research about these facts. Shadish (2000) has presented an extended discussion of what such an empirical program of quasi-experimentation might look like, including studies of questions like the following:

(a) Can quasi-experiments yield accurate effect estimates, and if so, under what conditions?

(b) Which threats to validity actually occur in practice (e.g., pretest sensitization, experimenter expectancy effects), and if so, under what conditions?

(c) Do the design features in Table 1 improve causal inference when applied to quasi-experimental design, and if so, under what conditions?

Some of this empirical research has already been conducted (see Shadish 2000, for examples). The methodologies used to investigate such questions are eclectic, including case studies, surveys, literature reviews (quantitative and qualitative), and experiments themselves. Until very recently, however, these studies have generally not been systematically used to critique and improve quasi-experimental theory.

5. Conclusion

Three important factors have converged at the end of the twentieth century to create the conditions under which the development of better quasi-experimentation may be possible. First, over 30 years of practical experience with quasi-experimental designs have provided a database from which we can conduct empirical studies of the theory. Second, after decades of focus on randomized designs, statisticians and economists have turned their attention to improving quasi-experimental designs. Third, the computer revolution provided both theorists and practitioners with increased capacity to invent and use more sophisticated and computationally intense methods for improving quasi-experiments. Each in their own way, these three factors have taken us several steps closer to answering that great unanswered question with which this article began.

See also: Campbell, Donald Thomas (1916–96); Comparative Studies: Method and Design; Experimental Design: Overview; Experimental Design:

Randomization and Social Experiments; Experimentation in Psychology, History of; Panel Surveys: Uses and Applications

Bibliography

Bollen K A 1989 *Structural Equations with Latent Variables*. Wiley, New York

Campbell D T, Stanley J C C 1963/1966 *Experimental and Quasi-experimental Designs for Research*. Rand-McNally, Chicago

Cochran W G 1965 The planning of observational studies in human populations. *Journal of the Royal Statistical Society, Series A General* **128**: 234–66

Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand-McNally, Chicago

Heckman J J 1979 Sample selection bias as a specification error. *Econometrica* **47**: 153–61

Hill A B 1953 Observation and experiment. *The New England Journal of Medicine* **248**: 995–1001

Rosenbaum P R 1995 *Observational Studies*. Springer-Verlag, New York

Schlesselman J J 1982 *Case–Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York

Shadish W R 2000 The empirical program of quasi-experimentation. In: Bickman L (ed.) *Validity and Social Experimentation: Donald Campbell's Legacy*. Sage Publications, Thousand Oaks, CA

Shadish W R, Cook T D, Campbell D T in press *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston

Triplett N 1898 The dynamogenic factors in pacemaking and competition. *American Journal of Psychology* **9**: 507–33

Winship C, Mare R D 1992 Models for sample selection bias. *Annual Review of Sociology* **18**: 327–50

W. R. Shadish

Queer Theory

'Queer theory' is a notoriously unstable phrase, and one much in contention. As a new theoretical movement with equally new political counterparts, it is in constant flux and development, and is characterized more by what it challenges and contests than by what it offers in the shape of a unified social theory. Drawing on the work of theorists such as Eve Sedgwick and Judith Butler, queer theory 'describes those gestures or analytical models which dramatise incoherencies in the allegedly stable relations between chromosomal sex, gender and sexual desire (Jagose 1996). In this sense, queer theory is a challenge to the 'obvious categories (man, woman, *latina*, jew, butch, *femme*), oppositions (man vs. woman, heterosexual vs. homosexual), or equations (gender = sex) upon which conventional notions of sexuality and identity rely' (Hennessy 1993). Queer theory argues instead that sexual desire and sexual practices are not reducible or explicable solely in terms of identity categories, such as gender, race, class, or sexual orientation. It is radically anti-essentialist, in that it challenges a notion of homosexuality as intrinsic, fixed, innate, and universally present across time and space.

Queer theorists reject any mode of thought that relies on a conception of identity as unified and self-evident (e.g., I have sex with people of the opposite sex, therefore I must be heterosexual), and instead demonstrate that desires, sexual practices, and gendered identities are performances and enactments, rather than expressions of 'true' subjectivity. Heterosexuality is therefore challenged by queer theory not simply as a 'hegemonic' mode of identity, but as a false claim to unity and coherence that is constantly undermined by the incoherencies of sex and gender, incoherencies that the queer analytic hopes to expose and celebrate.

1. Intellectual Origins

In the broadest sense, queer theory emerged in what might be called the postmodern moment, when intellectual unease with unitary and cohesive frameworks of knowing reached a fever pitch. While impossible to summarize here, queer theory's allegiance to postmodern and/or poststructural modes of thought can be traced in its challenge to the notion of unitary identity (as in 'gay' or 'straight'), its refusal to understand sexuality through a singular and unified lens (homosexual desire, feminist theory, gender), a rejection of binary models (gay/straight, man/woman, biological/social, real/constructed), and a more generic critique of identity-based theories and politics that, according to poststructuralist accounts, invariably reproduce the very conditions of repression they desire to challenge. For example, the term 'gay' or 'homosexual' might be critiqued as a (fictional) category that shores up the binary opposition between 'gay' and 'straight' that is itself part of the repressive logic of identity. To claim 'gayness' is therefore not simply or solely an act of self-revelation but is also a way of corralling sexuality within the framework of a category that only appears coherent but that, when opened up, reveals its instability (e.g., Am I still gay if I sleep with a person of the opposite sex? Or if I sleep with those of the same sex but only in certain conditions and in certain ways? Or if my self-understanding is of myself as 'straight?'). Queer theory, in that sense, has developed within and through the deconstructive impulse of poststructuralism, challenging assertions of unitary identity and necessary linkages (between, say, sexual desire and gender orientation) and arguing instead for a more provisional, contingent, and fluid conception of the 'queer' in contemporary culture.

While queer theory emerges as coterminous with postmodern impulses, it also traces its intellectual origins in lesbian/gay studies and feminist theory even

as it challenges those very paradigms. While for some the term queer theory is simply a synonym for 'lesbian and gay studies' or even 'feminist theory,' for many it signifies a real shift away from what is understood to be the 'essentialist' underpinnings of those more traditionally defined fields. In that sense, queer theory positions itself as a theoretical framework that rejects the typical binarisms (man/woman, gay/straight) in favor of a more universalist and overarching critique of the whole edifice of 'normalization,' a set of rules, regulations, and normative commands that create the rubric of 'good' vs. 'bad' sexualities. Queer theory takes issue with the ways in which feminist theory has tended to analyze sexuality within the framework of gender, and contests the ways in which 'gay' assumes a lining up of identity with acts, behavior with being. Drawing on the pioneering work of postmodern thinkers such as Michel Foucault and Eve Sedgwick, as well as hard-working historians such as Jonathan Ned Katz, queer theory joins in the disarticulating of sexual identities, sexual practices, and gender that characterizes contemporary postmodern theorizing. Indeed, queer theory takes all the mainstreaming, normalizing impulses to task—be it the scientistic move of disciplines or the ethnic minority model of mainstream gay organizations. In general, then, queer theory can be understood as part of the larger intellectual ferment of the late twentieth century that has challenged the very core concepts of Western humanism: identity, science, truth, and authenticity.

2. *Social Context*

If queer theory develops out of the intellectual movements of poststructuralism, feminist theory, and lesbian and gay studies, it also emerged in the context of recent political alliances and strategies. As the gay community faced AIDS, and governmental reluctance to address this disease was slow and insufficient, gay activist groups developed new, 'in your face' strategies, exemplified by the work of groups such as ACT-UP and Queer Nation. While not always or necessarily drawing on queer theory explicitly, these groups helped to define the terrain of the intellectual discourse by emphasizing a nonnormative, anti-assimilationist alternative to traditional identity politics and minority group formations. As Warner (1991) argues,

> The insistence on 'queer'—a term defined against 'normal' and generated precisely in the context of terror—has the effect of pointing out a wide field of normalization, rather than simple intolerance, as the site of violence. Its brilliance as a naming strategy lies in combining resistance on the broad social terrain of the normal with more specific resistance on the terrains of phobia and queer-bashing, on one hand, or of pleasure on the other. 'Queer,' therefore, also suggests the difficulty in defining the population whose interests are at stake in queer politics. And as a partial replacement for 'lesbian and gay' it attempts partially to separate questions of sexuality from those of gender.

Queer theory therefore has a strong affinity with radical political praxis and thus brings to the academy an intellectual movement informed substantively by political crises and interventions.

3. *Core Challenges*

Queer theory takes as its primary theoretical and analytic focus the destabilization of 'gay and lesbian' and the disruption of the category of gender, particularly in terms of how gendered analysis is often understood to be the appropriate intellectual framework for the study of sexuality. Queer theory critiques the conceptualization of gender as a self-evident and unified category, and argues that this framework leads inexorably to another kind of polarizing and oppressing logic in which categories of identity become ways to 'police' populations whose behaviors, desires, and modes of living seem somehow discordant with the identity in question.

Because queer theory understands sexuality as fluid, mobile, and plural, the concept of 'performance' has been a central one, insisting on the situational and provisional nature of identity. Thus, 'gay' or 'lesbian' or even 'woman' are not understood as self-evident and essential categories, but rather are understood to be enactments of subjectivity. In this queer reading of identity, gender is understood as process, acquisition, enactment, creation. Thus, queer theorists often take as their analytic site the spaces where these performances occur and are highlighted—cross-dressing and drag, for example. Rather than assert a universal humanism (gays are essentially the same as heterosexuals, women are 'as good as' men), queer theory embraces the vilified, denigrated, ostracized identities (much as it embraces the epithet 'queer') and sees in these hybrid spaces (the transsexual, the drag queen) possibilities for a more thorough-going critique of normative culture and its repressive apparatuses.

Therefore, for queer theorists, the problem is not primarily one of bigotry, or lack of tolerance, or even homophobia *per se*, but rather the edifice of categorization and normalization that forces even those considered 'deviant' to tie behaviors and identities to narrow lines of power, lines that always imply repression and limitation. Queer theory celebrates sexual 'otherness' outside of identity categories (even outside those putatively radical categories such as 'gay' and 'lesbian') because it believes fervently that the power of domination is to be found primarily in the production and enforcement of identity itself. Identities, even alternative ones, carry with them normative assumptions about correct and incorrect ways of living and loving. And while identities give meaning to our lives, they also curtail possibilities by promoting a

false sense of unity (this is what a homosexual is) to a human life that is invariably riven with contradictions and inconsistencies. Thus queer theorists search for and celebrate moments of sexual subjectivity that appear outside of, or resistant to, firm categories of unified identity.

4. *Queer Theory and the Academy*

Queer theory cannot, by any stretch of the imagination, be located in any singular field or characterized by any unified methodology. While it is undoubtedly true that queer theory has found a more hospitable home in the humanities (particularly in departments of comparative literature and English), increasing numbers of social scientists (particularly those working within the broad framework of cultural studies) are engaging with queer theory and adapting it to their own research agendas. However, part of the uniqueness of queer theory comes in its deployment of multiple strategies to analyze a myriad of social and cultural phenomenon. In that sense, queer theory itself becomes a model of a new sort of trans- or even antidisciplinarity in the academic world. That is why most of the interventions in queer theory emerge from specific and located microstudies, rather than from large-scale, grand theoretical ventures. In that way, queer theory evinces much similarity with the growing 'field' of cultural studies, itself marked by a traversing of typical academic and disciplinary boundaries. And, given its emphasis on performativity and on non-normativity, queer theory presents a challenge to academic discourses that seek assimilation within mainstream intellectual life. It is precisely that mainstreaming, that incorporation, that queer theory sets its sights on.

5. *Some Critical Notes*

Queer theory has been critiqued from any number of locations, but none so persuasively as feminist theorists and theorists of color. Both the feminist and race critiques of queer focus on the ways in which queer theory often assumes implicitly a white, gay maleness as the prototype of queer subjectivity. In addition, both feminists and radical theorists of color have been suspicious of any theoretical framework that vitiates the need for an aggressive identity politics (a politics based on a notion of a unified and 'real' identity—gay, woman, etc.), given that identity politics have been for many the route to social change. For many feminist theorists, queer theory is seen as naive in its claim to transcend gender as a meaningful category of analysis. For them, it is seen as having eviscerated the centrality of gender to social identity, social power, sexuality and sexual, identity by dethroning gender critique as the cornerstone of the critique of both homophobia and sexism. Queer theory, it is argued, sees patriarchal sexuality and sexist culture as almost tangential to the analysis of the repression of sexualities, while feminists have long argued that, for example, homophobia emerges from the larger and more pervasive rubric of sexist and patriarchal ideology and practice.

The point that gender is not the only significant marker of difference is an important one, and one that deserves development and reiteration. The regimes of sexuality and gender are not identical, either historically or theoretically, but many remain skeptical of their premature separation. In much of queer writing, it appears that feminism emerges as the enemy, the agent of a repressive regime of identity categorization, rather than heterosexism and patriarchal violence. Feminism is often depicted in queer theoretical writings as an outdated and essentialist relic of a humanist past, and queer theory, in its transcendence of gender, race, and class, as the postmodern mode of the future. Martin (1994) has argued that, for Sedgwick and others, race and gender often assume a fixity, a stability, a ground, whereas sexuality (typically thematized as male) becomes the 'means of crossing' and the figure of mobility. In the process of making the female body the 'drag' on the (male) play of sexuality, 'the female body appears to become its own trap, and the operations of misogyny disappear from view.'

Queer theory's relation to the politics and theorizing of racialized identities is no less fraught with difficulties than its relation to feminism and feminist identities. The generalizing tendencies of queerness (queer as everything not normatively heterosexual) has the potential to evacuate the specificities of racialized identities in favor of a queer universalism that claims multiracial status without ever developing seriously a race-based critique of heteronormativity, and without ever considering seriously the ways in which heterosexual dominance depends on certain racial ideologies. Queer can 'de-race' the homosexual of color in much the same way 'old-time' gay studies often has, effectively erasing the specificity of 'raced' gay existence under a queer rubric in which whiteness is not problematized. Dhairyam (1994) critiques the implicit whiteness of queerness while still attempting to activate the category 'queer women of color': 'Queer theory comes increasingly to be reckoned with as critical discourse, but concomitantly writes a queer whiteness over raced queerness, it domesticates race in its elaboration of sexual difference.' Anzaldua (1991) also accuses white academics of co-opting the term 'queer' and using it to construct 'a false unifying umbrella which all "queers" of all races, ethnicities and classes are shoved under.'

Many have argued that queer theory lives in a realm of almost utopian idealism and abstraction, where the material conditions of marginalized populations are 'performed' away by theoretical sleights of hand, as if oppression is solely a matter of sexuality, and its representation and regulation. This inattention to

material social relations (commodification, the fluctuations of international capital, shifting forms of familial life, rise in antigay activism, regressive social legislation, increasing disenfranchisement of people of color, etc.) and the academicism of much of queer writing might be problems for a lesbian/gay praxis that is both class and race conscious.

6. Future Directions

Many self-defined queer theorists have taken these criticisms to heart and are attempting to reframe queer theory to pay more attention to structural social conditions, bodily realities, and the lived vicissitudes of race, class, and gender violence. Hennessy's (1994) piece is an exemplary attempt to hold on to the insights of queer theory while insisting that we pay attention to the material social conditions (such as class) that help to determine who can or cannot have access to the brave new queer world. She demands forcefully that queer theorists pay more attention to the processes of commodification and avoid valorizing a politics of the outrageous at the expense of attending to the realities of structured social relations, relations not reducible to the discursive or cultural, although certainly not determinative of them either. Queer may hold out some possibilities for a politics and a theory that challenges the fixity and clarity of identity, that speaks to the fractured (non)self of postmodern subjectivity. The queer challenge to the notion of sexual identity as monolithic, obvious, and dichotomous is a healthy corrective to our vexing inability to see beyond the limitations of the homo/hetero opposition. In addition, the openness of the term 'queer' provides the possibility of theorizing 'beyond the hyphen,' beyond the additive models (race, class, gender, sexual orientation = oppressed identity) that have so often seemed to set up new hierarchies, or instead retreated into an empty recitation of 'difference.' In this sense, queer theory can make theoretical and political space for more substantive notions of multiplicity and the intersecting and often contradictory lines of desire, identity, and sexuality.

See also: Androgyny; Feminist Ethics; Feminist Movements; Feminist Theology; Feminist Theory; Feminist Theory: Postmodern; Feminist Theory: Radical Lesbian; Foucault, Michel (1926–84); Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender and Feminist Studies; Gender, Class, Race, and Ethnicity, Social Construction of; Gender Ideology: Cross-cultural Aspects; Heterosexism and Homophobia; Homosexuality and Psychiatry; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Male Dominance; Masculinities and Femininities; Nontraditional Families and Child Development; Rationality and Feminist Thought; Regulation: Sexual Behavior; Sexual Behavior and Maternal Functions, Neurobiology of; Sexual Behavior: Sociological Perspective; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Preference: Genetic Aspects; Sexual Risk Behaviors; Sexuality and Gender; Sexually Transmitted Diseases: Psychosocial Aspects; Transsexuality, Transvestism, and Transgender

Bibliography

Anzaldua G 1991 To(o) queer the writer—loca, escritora y chicana. In: Warland B (ed.) *Inversions: Writing by Dykes, Queers, and Lesbians*. Press Gang Publishers, Vancouver, BC, pp. 249–63

Beemyn B, Eliason M (eds.) 1996 *Queer Studies: A Lesbian, Gay, Bisexual, and Transgender Anthology*. New York University Press, New York

Bergman D 1993 Disciplining gay and lesbian studies. Andrew R Hilen Lecture at the University of Washington, Seattle, WA

Butler J 1990 *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York

Butler J 1991 Imitation and gender insubordination. In: Fuss D (ed.) *Inside/Out*. Routledge, New York, pp. 13–31

Butler J 1993a Critically queer. *GLQ-A Journal of Lesbian and Gay Studies* **1**: 17–32

Butler J 1993b *Bodies That Matter*. Routledge, New York

de Lauretis T 1991 Queer theory: Lesbian and gay sexualities, an introduction. *Differences* **3**(2): iii–xviii

D'Emilio J 1983 *Sexual Politics, Sexual Communities: The Making of a Homosexual Minority in the United States, 1940–1970*. University of Chicago Press, Chicago

Dhairyam S 1994 Racing the lesbian, dodging white critics. In: Doan L (ed.) *The Lesbian Postmodern*. Columbia University Press, New York pp. 25–46

Doty A 1993 *Making Things Perfectly Queer: Interpreting Mass Culture*. University of Minnesota Press, Minneapolis, MN

Duggan L 1992 Making it perfectly queer. *Socialist Review* **22**(1): 11–31

Duggan L 1994 Queering the state. *Social Text* **39**: 1–14

Editors 1990/1993 Introduction. *Radical America* **24**(4): 2–13

Fuss D (ed.) 1991 *Inside/Out: Lesbian Theories, Gay Theories*. Routledge, New York

Harper P B, White E F, Cerullo M 1990/1993 Multi/Queer/Culture. *Radical America* **24**(4): 27–37

Hennessy R 1993 Queer theory: A review of the *Differences* special issue and Wittig's *The Straight Mind*. *Signs* **18**(4): 964–73

Hennessy R 1995 Queer visibility in commodity culture. *Cultural Critique* **29**: 31–76

Jagose A 1996 *Queer Theory: An Introduction*. New York University Press, New York

Kader C, Piontek T 1992 Introduction. *Discourse* **15**(1): 5–10

Katz J N 1995 *The Invention of Heterosexuality*. Dutton, New York

Kitzinger C, Wilkinson S 1994 Virgins and queers: Rehabilitating heterosexuality? *Gender & Society* **8**(3): 444–63

Martin B 1993 Lesbian identity and autobiographical difference(s). In: Abelove H, Barale M A, Halperin D M (eds.) *The Lesbian and Gay Studies Reader*. Routledge, New York pp. 274–93

Martin B 1994 Sexualities without genders and other queer Utopias and gay and lesbian studies. *Diacritics* **24**(2–3): 104–21

Moraga C, Hollibaugh A 1983 What we're rollin around in bed with: Sexual silences in feminism. In: Snitow A, Stansell C, Thompson S (eds.) *Powers of Desire: The Politics of Sexuality*. Monthly Review Press, New York, pp. 394–405
Morton D (ed.) 1996 *The Material Queer: A LesBiGay Cultural Studies Reader*. Westview Press, Boulder, CO
Nunokawa J 1992 Identity crisis: Queer politics in the age of possibilities. Interview by Alisa Solomon. *Village Voice* **30** (June): 27
Phelan S 1993 (Be)Coming out: Lesbian identity and politics. *Signs* **18**(4): 773–90
Reid-Pharr R 1990/1993 The spectacle of blackness. *Radical America* **24**(4): 57–65
Rubin G S 1993 Thinking sex: notes for a radical theory of the politics of sexuality. In: Abelove H, Barale M A, Halperin D M (eds.) *The Lesbian and Gay Studies Reader*. Routledge, New York, pp. 3–44
Sedgwick E K 1989 Across gender, across sexuality: Willa Cather and others. *South Atlantic Quarterly* **88**(1): 53–72
Sedgwick E K 1990 *Epistemology of the Closet*. University of California Press, Berkeley, CA
Sedgwick E K 1991 How to bring your kids up gay. *Social Text* **29**: 18–27
Sedgwick E K 1993 *Tendencies*. Duke University Press, Durham, NC
Seidman S 1994 Queer pedagogy/queer-ing sociology. *Critical Sociology* **20**(3): 167–76
Seidman S (ed.) 1996 *Queer Theory/Sociology*. Blackwell, Cambridge, MA
Seidman S 1997 *Difference Troubles: Queering Social Theory and Sexual Politics*. Cambridge University Press, Cambridge, UK
Smith B 1993 Queer politics: Where's the revolution? *The Nation* **5** (July): 12–16
Stein A 1992 Sisters and queers: The decentering of lesbian feminism. *Socialist Review* **22**(1): 33–55
Walters S 1995 *Material Girls: Making Sense of Feminist Cultural Theory*. University of California Press, Berkeley, CA
Walters S 1996 From here to queer: Radical feminism, postmodernism, and the lesbian menace (or, Why can't a woman be more like a fag?) *Signs* **21**(4) (Summer): 830–69
Warner M 1991 Fear of a queer planet. *Social Text* **9**(4): 3–17
Warner M 1992 From queer to eternity: An army of theorists cannot fail. *Voice Literary Supplement* June: 18–19
Warner M (ed.) 1993 *Fear of a Queer Planet: Queer Politics and Social Theory*. University of Minnesota Press, Minneapolis, MN
Whisman V 1993 Identity crises: Who is a lesbian, anyway? In: Stein A (ed.) *Sisters, Sexperts, Queers*. Plume, New York, pp. 47–60

S. D. Walters

Questionnaires: Assessing Subjective Expectations

Questionnaires are used to elicit from individuals a broad array of objective information about their experiences and environments, as well as subjective information about their thoughts and perceptions. An important objective of social and behavioral science research is to make questionnaires more informative. In practical terms, how should the questionnaire be administered and what should be its content? Aspects of questionnaire administration include when and where the data are collected, who does the data collection, and the mode of administration (e.g., face-to-face, telephone, or Internet). Aspects of content begin with the basic matter of determining the purposes of the questionnaire and end with specific issues of item wording, the sequencing of questions, and the formatting of responses.

This article focuses on a single topic about questionnaires—how to elicit respondents' subjective expectations of future events. For example, demographers ask women to predict their future fertility and political scientists ask citizens to predict their voting behavior in upcoming elections. Researchers use the responses to expectation questions to predict fertility, voting behavior, and other future events. They also use the responses to assess how individuals' perceptions of the future are related to the decisions they make today. Subjective expectations of future events seem to be measured with increasing frequency. Yet the social and behavioral sciences have not formed a common view on how subjective expectations should be conceptualized and measured. A particularly intriguing controversy, described in this article, concerns the relative merits of verbal and probabilistic forms of questioning.

1. Subjective Concepts and Measures

A US National Research Council (NRC) panel study on *Surveying Subjective Phenomena* has distinguished subjective from objective phenomena as follows (Turner and Martin 1984, p. 8):

> Subjective phenomena are those that, in principle, can be directly known, if at all, only by persons themselves, although a person's intimate associates or a skilled observer may be able to surmise from indirect evidence what is going on 'inside.' Objective phenomena are those that can be known by evidence that is, in principle, directly accessible to an external observer.

The NRC panel acknowledged that the sharp dichotomy just drawn is not always tenable, but the essential idea clearly has content: subjective phenomena are not amenable to external verification.

The social and behavioral sciences have a common interest in understanding subjective phenomena, but the various disciplines have not been able to agree on concepts or measures. The conceptual problem is to characterize what is going on 'inside.' A reader of social and behavioral science research encounters an enormous variety of subjective concepts in use, with concepts that are central to one discipline often playing no role in others. A modest sense of this variety can be obtained by scanning the introductory chapter of Turner and Martin (1984), which mentions affect, aspirations, attitudes, disappointments, dispositions,

emotions, frustrations, fulfillment, happiness, hopes, intentions, motives, opinions, plans, preferences, quality of life, satisfaction, and values.

The measurement problem is to obtain empirical evidence on what is going on 'inside.' A reader of empirical research in sociology, social psychology, or political science finds much analysis of subjective data elicited from survey respondents through questionnaires. Yet a reader of anthropology or economics finds little reliance on questionnaire responses. Anthropologists, who tend to recoil from the tightly scripted structure of questionnaires, have preferred to elicit information through conversations. Economists, who tend to be skeptical of subjective data of any kind, have sought to infer subjective phenomena from observations of behavior and experiences.

To illuminate the diversity of perspectives on subjective phenomena, this article examines how different disciplines have used questionnaires to study subjective expectations about potential future events. How persons perceive the future is a shared concern of all social and behavioral science disciplines. Yet there is a wide spectrum of views on how expectations should be conceptualized and measured.

Section 2 examines the prevailing practice in economics of inferring expectations from realization data, and recalls early efforts to collect data on economic expectations through questionnaires. Section 3 describes qualitative attitudinal research as practiced by sociologists and social psychologists. Section 4 describes the growing body of research in economics and cognitive psychology eliciting subjective probabilities of future events. Section 5 considers the relative merits of qualitative and probabilistic modes of questioning. Section 6 poses directions for future research on subjective expectations. See Dominitz and Manski (1999) for a more detailed exposition of the subjects discussed below.

2. *Inference on Expectations in Economics*

Early in their careers, economics students are taught that a good economist believes only what people do, not what they say. Skeptical of subjective data, economists have sought to infer expectations of future events by combining data on past experiences with assumptions about processes of expectation formation. This is a daunting task. To infer expectations from data on past experiences, economists must speculate about what information persons possess and how they use the available information to form their expectations.

Consider, for example, the matter of income expectations. Economists often assume that persons use their own past incomes to forecast their future incomes. Perhaps so, but how do persons form expectations of future income conditional on past income? Economists often assume that persons have *rational expectations*; that is, they know the actual stochastic process generating their income streams. Perhaps so, but what is this stochastic process? The practice has been to specify the income process up to some parameters and use available data on income realizations to estimate the parameters. Authors differ, however, in their specifications. So the literature displays competing models of expectations formation, with no clear way of distinguishing among them. See Dominitz and Manski (1997a).

If inference on expectations from data on experiences is so challenging, why do economists not collect and analyze data on expectations? The answer lies at least partly in the history of a scientific controversy that began in the 1940s and persisted until the 1960s. In the 1940s, the US Federal Reserve Board began to fund an annual Survey of Consumer Finances, conducted by the University of Michigan Survey Research Center (SRC), that elicited qualitative assessments of expected household finances. A typical question took this form:

> How about a year from now—do you think you people will be making more money or less money than you are now, or what do you expect?

The usefulness of responses to such qualitative questions was controversial and the Federal Reserve Board appointed a committee to assess their value. The Federal Reserve Consultant Committee on Consumer Survey Statistics (1955) issued findings that questioned the predictive power of the SRC expectations data. The negative findings of the Committee were challenged by SRC researchers, notably George Katona, a leading proponent of research on consumer attitudes and expectations (Katona 1957). A contentious conference on expectations data followed (National Bureau of Economic Research 1960). Then Juster (1964) reported an intensive study, drawing largely negative conclusions, of the usefulness of qualitative expectations data in predicting individual behavior.

The Federal Reserve–SRC controversy only concerned the usefulness of a certain type of qualitative expectations data in predicting consumer behavior. Academic economists, however, appear to have drawn the broader conclusion that all expectations data are suspect. Economists also failed to take note when Juster (1966) proposed elicitation of probabilistic expectations of consumer durable purchases and reported favorable empirical evidence. Exploration of this idea had to wait a quarter century, as described in Sect. 4.

3. *Qualitative Attitudinal Research*

The Federal Reserve–SRC controversy may have soured most economists on the usefulness of qualitative expectations data, but the collection and analysis of such data have long been important activities in

other behavioral and social science disciplines. Many questions in common use have a structure similar to that of this General Social Survey (GSS) question eliciting expectations of job loss (Davis and Smith 1994):

> Thinking about the next twelve months, how likely do you think it is that you will lose your job or be laid off—very likely, fairly likely, not too likely, or not at all likely?

This GSS question illustrates well a basic persistent problem that researchers face in interpreting qualitative expectations data—assessment of the inter- and intrapersonal comparability of responses. Respondents are often asked to report whether a specified event is 'expected,' 'likely,' or 'probable' to occur. Sometimes they are asked to report their strength of belief in the event on a numerical scale with undefined units. Do different respondents interpret these verbal phrases and numerical scales in the same way? Does a given respondent interpret them in the same way when asked about different events? Cognitive research does not give reason to think that responses should be or are comparable. Indeed, the available empirical evidence indicates that interpretations of qualitative expectations questions vary substantially between persons (see Sect. 5).

Concern about interpersonal comparability comes into even sharper focus when considering the events about which respondents are asked. Attitudinal researchers often ask respondents about vaguely defined future outcomes, as in this SRC question (Curtin 1982):

> Now looking ahead—do you think that a year from now you (and your family living there) will be better off financially, or worse off, or just about the same as now?

How should a respondent interpret the phrase 'better off financially?' Do different respondents interpret it the same way? The data provide no way of knowing.

A second persistent problem in qualitative attitudinal research is that the coarseness of the response options limits the information contained in the responses. Consider, for example, the fertility question asked female respondents in the annual June Supplement to the Current Population Survey of the US Bureau of the Census:

> Looking ahead, do you expect to have any (more) children? Yes No Uncertain

The three response options can express little of the richness of the uncertainty that women may perceive about their future childbearing. Savage (1971, p. 795) called attention to the problem when he observed that 'Yes, No, or Maybe is Not Enough.' Juster (1966) hypothesized that a person facing a Yes/No expectations question responds as would a statistician asked to make a best point prediction of a future random event. Manski (1990) formalized this hypothesis and derived an upper bound on the information about expectations a researcher can extract from responses to Yes/No questions.

4. *Elicitation of Probabilistic Expectations*

A growing body of research in economics and cognitive psychology elicits expectations in the form of subjective probabilities. For example, the Survey of Economic Expectations uses this question to elicit expectations of job loss (Dominitz and Manski 1997b):

> What do you think is the percent chance that you will lose your job during the next 12 months?

The National Longitudinal Study of Youth: 1997 Cohort of the US Bureau of Labor Statistics elicits the fertility expectations of teenage [girls/boys] as follows:

> What is the percent chance that you will become the [mom/dad] of a baby sometime between now and when you turn 20?

Probabilistic questions of similar structure are posed in other national surveys, such as the Health and Retirement Study (Juster and Suzman 1995) and the Bank of Italy's Survey of Household Income and Wealth, as well as in small-scale surveys such as the one described in Quadrel et al. (1993).

The idea that probabilistic elicitation of expectations might improve on the traditional qualitative approaches of attitudinal research developed independently in economics and in cognitive psychology. In economics, the idea appears to have originated with Juster (1966), who recommended elicitation of purchase probabilities for consumer durable goods. Market researchers were immediately attracted to Juster's proposal, and elicitation of purchase probabilities has since become a common practice in market research (e.g., Jamieson and Bass 1989). However, the idea that expectations might be elicited probabilistically from survey respondents did not draw serious attention within economics until the early 1990s, when economists began to re-assess the discipline's self-imposed prohibition on the collection and analysis of subjective data.

Elicitation of subjective probabilities has a more continuous history of research in cognitive psychology. Psychologists have long conducted calibration studies assessing the accuracy of elicited expectations reports when the outcome is known. Lichtenstein et al. (1982) reviewed findings from calibration studies dating back to 1906. McClelland and Bolger (1994) updated the review with findings from 1980 through 1994. Whereas the older studies mostly examined the accuracy of experts (e.g., weather forecasters' reported probabilities of precipitation), the more recent re-

search typically analyzes the expectations of non-experts, especially students situated in a cognitive laboratory.

5. Qualitative vs. Probabilistic Expectations

As this article is written, a consensus has yet to emerge on the relative merits in practice of qualitative and probabilistic elicitation of subjective expectations. However, there is reasonably widespread agreement that elicitation of probabilistic expectations has several *a priori* desirable features. Perhaps the most basic attraction is that probability provides a well-defined absolute numerical scale for responses. Hence there is some reason to think that responses may be inter- and intrapersonally comparable.

A second attraction is that some empirical assessment of the internal consistency and external accuracy of respondents' expectations is possible. A researcher can use the algebra of probability (Bayes Theorem, the law of total probability, etc.) to examine the internal consistency of a respondent's elicited expectations about different events. In those cases where probability has a frequentist interpretation, a researcher can compare elicited subjective probabilities with known event frequencies, as is done in calibration studies, and reach conclusions about the correspondence between subjective beliefs and frequentist realities.

A third consideration is the usefulness of elicited expectations in predicting prospective outcomes. Suppose that respondents have reasonably accurate expectations about the likelihood of future events. Then, as argued by Juster (1966), Savage (1971), and Manski (1990), numerical responses to probability questions should have more predictive power than categorical responses to qualitative expectations questions.

These *a priori* desirable features of probabilistic expectations data must be weighed against a set of unresolved issues. Perhaps the most basic concern is that there is no direct way to assess how well elicited subjective probabilities reflect respondents' thinking. Of course this concern is not specific to elicitation of probabilities. Interpretation of responses to subjective questions of any type runs up against the generic problem that a researcher cannot directly observe a respondent's thinking.

A specific concern in the elicitation of probabilities is that respondents may not think probabilistically about uncertain events. A large and diverse literature suggests that persons may think about uncertain events using less than the full structure of modern probability theory (e.g., Camerer and Weber 1992). If so, then questions eliciting precise probabilities force respondents to give point responses when only interval responses may be meaningful.

Within cognitive psychology, there has long been controversy about the way in which humans internally represent their beliefs, and their ability and willingness to express their internal beliefs as numerical probabilities. Ferrell and McGoey (1980) posed models in which individuals may have some difficulty expressing beliefs as numerical probabilities, but nevertheless concluded that elicitation of numerical subjective probabilities is feasible. However, Zimmer (1983), who argued that humans process information using verbal rather than numerical modes of thinking, concluded that expectations should be elicited in verbal rather than numerical forms.

Wallsten et al. (1993) has reported that a majority of respondents prefer to communicate their own beliefs verbally and to receive the beliefs of others in the form of numerical probabilities. This asymmetry is intriguing but only marginally relevant to the design of expectations questions. The relevant question is not what communication mode respondents prefer to use, but rather what modes they are willing and able to use. Wallsten et al. (1993) report that virtually all of their respondents were willing to communicate their beliefs numerically, should the situation warrant it. Certainly the accumulating experience of economists using questionnaires to elicit probabilistic expectations has been that respondents are willing and able to communicate their beliefs in numerical probabilistic terms (e.g., Dominitz and Manski 1997a, 1997b, Hurd and McGarry 1995).

A related theme of cognitive research has been to learn how individuals associate numerical probabilities and verbal phrases such as 'very likely,' 'probable,' or 'rarely' (e.g., Wallsten et al. 1986). A central finding is that individuals vary considerably in the numerical probabilities that they associate with given verbal phrases. There is, however, controversy about the proper conclusion to be drawn. Should one conclude that individuals differ in their interpretation of verbal phrases, in their interpretation of numerical probabilities, or in their interpretation of both? The empirical findings are compatible with all of these conclusions.

6. Expectations Formation and Behavior

Empirical research on subjective expectations is still in its infancy. Basic questions about the conceptualization and measurement of expectations are not yet resolved. When sufficiently satisfactory answers to these questions are achieved, research will be able to move to yet more challenging and interesting matters.

Future research should seek to understand how persons form their expectations. Attitudinal researchers have given little systematic thought to this basic matter. Researchers who conceptualize expectations in probabilistic terms have not been able to agree on how persons update their subjective probabilities in light of new evidence. Economists and statistical decision analysts have commonly assumed that persons apply Bayes rule to update their expectations.

Psychologists have reported intriguing empirical evidence to the contrary (e.g., Kahneman et al. 1982), but the practical import of this evidence remains unsettled. Many social scientists hypothesize that persons form expectations by observation of the behavior and experiences of others in their environments. However, we know precious little about the nature of social learning processes.

Research on subjective expectations should aim to understand how a person's perception of the future affects behavior. Economists suppose that youth make schooling choices based on their expectations of the returns to schooling relative to labor market activity. They also hypothesize that uncertainty about future income induces individuals to increase savings, thereby decreasing consumption. Criminologists suggest that perceived risks of crime victimization may cause individuals to take precautions in their daily activities. Public health researchers speculate on how perceptions of health risks affect smoking, drug use, and unprotected sexual behavior. To shed light on these and similar questions of practical import will require much more intensive collection and analysis of expectations and behavioral data than the social and behavioral sciences have yet undertaken.

See also: Classical (Psychometric) Test Theory; Test Administration: Methodology

Bibliography

Camerer C, Weber M 1992 Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty* **5**: 325–70

Curtin R 1982 Indicators of consumer behavior: The University of Michigan Surveys of Consumers. *Public Opinion Quarterly* **46**: 340–52

Davis J, Smith T 1994 *The General Social Surveys, 1972–1994, Cumulative File*. National Opinion Research Center, Chicago

Dominitz J, Manski C 1997a Using expectations data to study subjective income expectations. *Journal of the American Statistical Association* **92**: 855–67

Dominitz J, Manski C 1997b Perceptions of economic insecurity: Evidence from the survey of economic expectations. *Public Opinion Quarterly* **61**: 261–87

Dominitz J, Manski C 1999 The several cultures of research on subjective expectations. In: Smith J, Willis R (eds.) *Wealth, Work, and Health: Essays in Honor of F. Thomas Juster*. University of Michigan Press, Ann Arbor, MI

Federal Reserve Consultant Committee on Consumer Survey Statistics 1955 Smithies Committee report in *Reports of the Federal Reserve Consultant Committees on Economic Statistics*. Hearings of the Subcommittee on Economic Statistics of the Joint Committee on the Economic Report, 84th US Congress

Ferrell W, McGoey P 1980 A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance* **26**: 32–53

Fishbein M, Ajzen I 1975 *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA

Hurd M, McGarry K 1995 Evaluation of the subjective probabilities of survival in the health and retirement study. *Journal of Human Resources* **30**: S268–92

Jamieson L, Bass F 1989 Adjusting stated intentions measures to predict trial purchase of new products: A comparison of models and methods. *Journal of Marketing Research* **26**: 336–45

Juster F T 1964 *Anticipations and Purchases*. Princeton University Press, Princeton, NJ

Juster F T 1966 Consumer buying intentions and purchase probability: An experiment in survey design. *Journal of the American Statistical Association* **61**: 658–96

Juster F T, Suzman R 1995 An overview of the health and retirement study. *Journal of Human Resources* **30**: S7–56

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York

Katona G 1957 Federal Reserve Board Committee Reports on Consumer Expectations and Savings Statistics. *Review of Economics and Statistics* **39**: 40–6

Lichtenstein S, Fischoff B, Phillips L 1982 Calibration of probabilities: The state of the art to 1980. In: Kahneman D, Slovic P, Tversky A (eds.) *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York

Manski C 1990 The use of intentions data to predict behavior: A best case analysis. *Journal of the American Statistical Association* **85**: 934–40

McClelland A, Bolger F 1994 The calibration of subjective probabilities: Theories and models 1980–94. In: Wright G, Ayton P (eds.) *Subjective Probability*. Wiley, New York

National Bureau of Economic Research 1960 *The Quality and Economic Significance of Anticipations Data*. Princeton University Press, Princeton, NJ

Quadrel M, Fischhoff B, Davis W 1993 Adolescent (in)vulnerability. *American Psychologist* **48**: 102–16

Savage L 1971 Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66**: 783–801

Schuman H, Johnson M 1976 Attitudes and behavior. *Annual Review of Sociology* **2**: 161–207

Turner C, Martin E (eds.) 1984 *Surveying Subjective Phenomena*. Russell Sage Foundation, New York

Wallsten T, Budescu D, Rapoport A, Zwick R, Forsyth B 1986 Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General* **115**: 348–65

Wallsten T, Budescu D, Zwick R, Kemp S 1993 Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society* **31**: 135–8

Zimmer A 1983 Verbal vs. numerical processing of subjective probabilities. In: Scholz R (ed.) *Decision Making Under Uncertainty*. North-Holland, Amsterdam

C. F. Manski

Questionnaires: Cognitive Approaches

Respondents' answers to standardized questions presented in self-administered questionnaires or face-to-face and telephone interviews are a major source of social science data. These answers, however, are

profoundly influenced by characteristics of the question asked, including the wording of the question, the format of the response alternatives, and the context in which it is presented. The cognitive and communicative processes underlying these influences are increasingly well understood and this article reviews their implications for question wording and questionnaire construction.

1. Respondents' Tasks

Answering a question requires that respondents (a) interpret the question to understand what is meant and (b) retrieve relevant information from memory to form an answer. In most cases, they cannot provide their answer in their own words but (c) need to map it onto a set of response alternatives provided by the researcher. Finally, (d) respondents may wish to edit their answer before they communicate it to the interviewer for reasons of social desirability and self-presentation. Respondents' performance at each of these steps is context dependent and profoundly influenced by characteristics of the questionnaire (Sudman et al. 1996, Chap. 3), including its administration mode (self-administered, or face-to-face, or telephone interview; Schwarz et al. 1991).

1.1 Question Comprehension

The key issue at the question comprehension stage is whether respondents' understanding of the question matches the meaning the researcher had in mind. Not surprisingly, methodology textbooks urge researchers to write simple questions and to avoid unfamiliar or ambiguous terms (see Sudman and Bradburn 1983, for good practical advice). However, understanding the literal meaning of a question is not sufficient to provide an informative answer. When asked, 'What have you done today?' respondents are likely to understand the words—yet, they still need to determine what kind of activities the researcher is interested in. Should they report, for example, that they took a shower, or not? Providing an informative answer requires inferences about the questioner's intention to determine the pragmatic meaning of the question.

To infer the pragmatic meaning, respondents draw on contextual information, including the content of adjacent questions and the nature of the response alternatives. Their use of this information is licensed by the tacit assumptions that govern the conduct of conversation in daily life, which respondents apply to the research interview (Clark and Schober 1992, Schwarz 1996). These pragmatic inferences have important implications for questionnaire construction.

1.1.1 Question context. Respondents interpret questions in the context of the overall interview and a term like 'drugs' acquires different meanings when presented in survey pertaining to respondents' medical history rather than to crime in the neighborhood. Moreover, they assume that adjacent questions are meaningfully related to one another, unless otherwise indicated. They may infer, for example, that an ambiguous 'educational contribution' refers either to fellowships that students receive, or tuition they have to pay, depending on the content of adjacent questions. Finally, they hesitate to reiterate information they have already provided, consistent with conversational norms that discourage redundancy. For example, questions pertaining either to happiness with one's life or to life-satisfaction typically elicit highly similar answers, unless both questions are presented together. In the latter case, respondents assume the researcher wouldn't ask for the same thing twice and differentiate between both questions, resulting in less similar answers. In self-administered questionnaires, question interpretation is influenced by preceding as well as following questions, whereas only preceding questions can exert an influence in face-to-face and telephone interviews (see Clark and Schober 1992, Sudman et al. 1996, Schwarz 1996 for extensive reviews).

1.1.2 Response alternatives. Most questions are presented with a fixed set of response alternatives, which facilitates administration and data analysis. Far from being 'neutral' measurement devices, however, response alternatives influence question interpretation (Schwarz and Hippler 1991). First, consider the differences between open and closed response formats. When asked in an open response format 'What have you done today?' respondents are likely to omit activities that the researcher is obviously aware of (e.g., 'I gave a survey interview') or may take for granted anyway (e.g., 'I took a shower'). Yet, if these activities were included in a closed list of response alternatives, most respondents would endorse them. At the same time, such a list would reduce reports of activities that are not represented on the list, even if an 'other' option were provided, which respondents rarely use. Both of these question form effects reflect that response alternatives can clarify the intended meaning of a question by specifying what the researcher is interested in. When using a closed response format, it is therefore important to ensure that the list of response alternatives covers the full range of behaviors or opinions.

Similarly, suppose respondents are asked how frequently they felt 'really irritated' recently. To infer if 'really irritated' refers to major or to minor annoyances, respondents may draw on an apparently formal feature of the question, namely the numeric values of the frequency scale. When the scale presents low (high) frequency alternatives, respondents infer that the

researcher is interested in rare (frequent) events. Hence, they report on major annoyances (which are relatively rare) when given a low frequency scale, but on minor annoyances when given a high frequency scale. As a result, the same question elicits reports of substantively different experiences, depending on the frequency scale used (Schwarz 1996, Chap. 5, Schwarz and Hippler 1991).

In sum, apparently formal differences in the response alternatives can profoundly influence respondents' interpretation of the question. Far from reflecting superficial responding, such effects indicate that respondents do their best to make sense of the question asked by drawing on the full range of information available to them—including information the researcher never intended to provide. To safeguard against unintended question interpretations, researchers have developed a number of different pretesting procedures, reviewed in Sect. 3.

1.2 Recalling or Computing a Judgment

Once respondents determined what the researcher is interested in, they need to recall relevant information from memory. In some cases, respondents may have direct access to a previously formed relevant judgment that they can offer as an answer. In most cases, however, they will not find an appropriate answer readily stored in memory and will need to compute a judgment on the spot. The different judgmental processes pertaining to behavioral questions and opinion questions are discussed in Sect. 2 and *Attitude Measurement*, respectively.

1.3 Formatting the Response

Unless the question is asked in an open response format, respondents need to format their answer to fit the response alternatives provided by the researcher. Respondents observe these question constraints and avoid answers that are not explicitly offered, as already seen in Sect. 1.1.2. For example, many more respondents answer 'don't know,' or choose a middle alternative, when these options are explicitly offered than when they are not (see Schwarz and Hippler 1991, for a review). Under which conditions the addition or omission of these response options results in different substantive conclusions, and whether some respondents are more likely to be affected by the response alternatives than others, is the topic of some debate (see Krosnick and Fabrigar in press).

Response formatting processes are particularly relevant when rating scales are used (Sudman et al. 1996, Chap. 4). Specifically, respondents use the most extreme stimuli to anchor the endpoints of the scale. Hence, a given stimulus will be rated as less extreme if presented in the context of a more extreme one, than if presented in the context of a less extreme one. In addition, if the number of stimuli to be rated is large, respondents attempt to use all categories of the rating scale about equally often. Hence, two similar stimuli may receive different ratings when only a few stimuli are presented, but may receive the same rating when a large number of stimuli has to be located along the scale. Accordingly, ratings of the same object cannot be directly compared when they were collected in different contexts, rendering comparisons over time or between studies difficult. The psychometric properties of rating scales have been the topic of extensive investigation (see Krosnick and Fabrigar in press). Seven point scales seem to be best in terms of reliability, percentage of undecided respondents, and respondents' ability to discriminate between scale points. Moreover, scales that provide verbal labels for each scale value seem more reliable than scales with verbal endpoints only.

Finally, respondents' choice of a given response alternative is often influenced by the order in which the alternatives are presented. Under the auditory presentation conditions of telephone interviews, a given alternative is more likely to be endorsed when it is presented last rather than first (referred to as a recency effect). Under the visual presentation conditions of a self-administered questionnaire, a given alternative is more likely to be endorsed when it is presented first rather than last (referred to as a primacy effect; see Sudman et al. 1996, Chap. 6, for a discussion of the underlying processes).

1.4 Editing the Response

Finally, respondents may want to edit their response before they communicate it, reflecting considerations of social desirability and self-presentation (see DeMaio 1984 for a review). Not surprisingly, the impact of self-presentation concerns is more pronounced in face-to-face interviews than in self-administered questionnaires. It is important to emphasize, however, that influences of social desirability are limited to potentially threatening questions and are typically modest in size. These influences can be reduced through techniques that ensure the confidentiality of the answer (see Sudman and Bradburn 1983 for detailed advice).

2. Asking Questions about Behavior

Most questions about respondents' behavior are frequency questions, pertaining, for example, to how often the respondent has bought something or has missed a day at work during some specified period of time. Researchers would ideally like respondents to identify the behavior of interest; to scan the reference period; to retrieve all instances that match the target

behavior; and to count these instances to determine the overall frequency of the behavior. This, however, is the route that respondents are least likely to take.

Except for rare and very important behaviors, respondents are unlikely to have detailed representations of numerous individual instances of a behavior stored in memory. Instead, the details of various instances of closely related behaviors blend into one global representation, rendering it difficult to distinguish and retrieve individual episodes. Accordingly, a 'recall-and-count' model does not capture how respondents answer questions about frequent behaviors or experiences. Rather, their answers are likely to be based on some fragmented recall and the application of inference rules to compute a frequency estimate. Accordingly, the best we can hope for under most conditions is a reasonable estimate, unless the behavior is rare and of considerable importance to respondents.

Conway (1990) provides a useful introduction to psychological research into autobiographical memory (i.e., individuals' memory for their experiences and behaviors) and Bradburn et al. (1987), Sudman et al. (1996, Chaps. 7–9), Tourangeau et al. (2000), and the contributions in Schwarz and Sudman (1994) review the implications for questionnaire construction. Section 2.1 summarizes what researchers can do to facilitate recall and Sect. 2.2 addresses respondents' estimation strategies.

2.1 Facilitating Recall

2.1.1 Recall cues and reference periods. If researchers are interested in obtaining reports that are based on recalled episodes, they may simplify respondents' task by providing appropriate recall cues and by restricting the recall task to a short and recent reference period. There are, however, important drawbacks to both of these strategies. Although the quality of recall will generally improve as the retrieval cues become more specific, respondents are likely to restrict their memory search to the particular cues presented to them, reflecting that the cues constrain the meaning of the question. As a result, respondents will omit instances that do not match the specific cues, resulting in under-reports if the list is not exhaustive. Moreover, using a short reference period may result in many 'zero' answers from respondents who rarely engage in the behavior, thus limiting later analyses to respondents with a high behavioral frequency. Finally, reference periods of different length may result in different question interpretations. Long reference periods ('How often have you been angry last year?') suggest that the researcher is interested in relatively rare events, whereas short reference periods ('How often have you been angry last week?') indicate an interest in frequent events. Hence, respondents may report on substantively different behaviors, e.g., rare and intense vs. frequent but minor episodes of anger.

Not surprisingly, different recall cues are differentially effective. The date of an event is the poorest cue, whereas cues pertaining to what happened, where it happened, and who was involved have been found to be very effective. In addition, recall will improve when respondents are given sufficient time to search memory. Recalling specific events may take up to several seconds and repeated attempts to recall may result in the retrieval of additional material, even after a considerable number of previous trials. Unfortunately, respondents are unlikely to have sufficient time to engage in repeated retrieval attempts in most research situations, and may often not be motivated to do so even if they had the time. This is particularly crucial in the context of survey research, where the available time per question is usually less than one minute.

Moreover, the direction in which respondents search memory may influence the quality of recall. Specifically, better recall is achieved when respondents begin with the most recent occurrence of a behavior and work backward in time than when they begin at the beginning of the reference period. This presumably occurs because memory for recent occurrences is richer and the recalled instances may serve as cues for recalling previous ones. Given free choice, however, respondents tend to prefer the less efficient strategy of forward recall.

Yet, even under optimal conditions, respondents will frequently be unable to recall an event or some of its critical details, even if they believed they would 'certainly' remember it at the time it occurred. In general, the available evidence suggests that respondents are likely to under-report behaviors and events, which has led many researchers to assume that higher reports of mundane behaviors are likely to be more valid. Accordingly, a 'the-more-the-better' rule frequently is substituted for external validity checks.

2.1.2 Dating recalled instances. After recalling or reconstructing a specific instance of the behavior under study, respondents have to determine if this instance occurred during the reference period. This requires that they understand the extension of the reference period and that they can accurately date the instance with regard to that period. Reference periods that are defined in terms of several weeks or months are highly susceptible to misinterpretations. For example, the term 'during the last 12 months' can be construed as a reference to the last calendar year, as including or excluding the current month, and so on. Similarly, anchoring the reference period with a specific date ('Since March 1, how often …?') is not very helpful because respondents will usually not be able to relate an abstract date to meaningful memories.

The most efficient way to anchor a reference period is the use of salient personal or public events, often referred to as landmarks. In addition to improving respondents' understanding of the reference period, the use of landmarks facilitates the dating of recalled instances. Given that the calendar date of an event will usually not be among its encoded features, respondents were found to relate recalled events to other, more outstanding events in order to reconstruct the exact time and day. Accordingly, using public events, important personal memories or outstanding dates (such as New Year's Eve) as landmarks was found to reduce dating biases (see Sudman et al. 1996, Tourangeau et al. 2000 for reviews).

Without a chance to relate a recalled event to a well-dated landmark, time dating is likely to reflect both forward and backward telescoping, that is, distant events are assumed to have happened more recently than they did, whereas recent events are assumed to be more distant than they are (Sudman et al. 1996, Chap. 8).

2.2 Estimation Strategies

In most cases, respondents need to rely on estimation strategies to arrive at an answer (Sudman et al. 1996, Chap. 9). Which strategy they use is often a function of the research instrument and hence to some extent under researchers' control.

2.2.1 Decomposition strategies. Many recall problems become easier when the recall task is decomposed into several subtasks. To estimate how often she has been eating out during the last three months, for example, a respondent may determine that she eats out about every weekend and had dinner at a restaurant this Wednesday, but apparently not the week before. Based on such partial recall she may arrive at an estimate of '18 times during the last 3 months.' Such estimates are likely to be accurate if the respondent's inference rule is adequate, and if exceptions to the usual behavior are rare.

In the absence of these fortunate conditions, however, decomposition strategies are likely to result in overestimates because people usually overestimate the occurrence of low frequency events and underestimate the occurrence of high frequency events. Hence, asking for estimates of a global, and hence frequent, category (e.g., 'eating out') is likely to elicit an underestimate, whereas asking for estimates of a narrow, and hence rare, category (e.g., 'eating at a Mexican restaurant') is likely to elicit an overestimate. The observation that decomposition usually results in higher estimates does therefore not necessarily reflect better recall.

2.2.2 Subjective theories. A particularly important inference strategy is based on subjective theories of stability and change (Ross 1989). In answering retrospective questions, respondents often use their current behavior or opinion as a bench-mark and invoke an implicit theory of self to assess whether their past behavior or opinion was similar to, or different from, their present behavior or opinion. The resulting estimates are correct to the extent that the subjective theory is correct. This, however, is often not the case.

In many domains, individuals assume a rather high degree of stability, resulting in retrospective reports that are too close to the current state of affairs, thus underestimating the degree of change that has occurred over time. For example, retrospective estimates of income or of tobacco, marijuana, and alcohol consumption were found to be heavily influenced by respondents' income or consumption habits at the time of interview. On the other hand, when respondents have reason to believe in change, they will detect change, even though none has occurred. Assuming, for example, that therapy is helpful, respondents may infer that their problem behavior prior to therapy was more frequent than was actually the case. Relying on such theory-driven retrospective reports, researchers may infer therapeutic success where none has occurred (see Ross 1989 for examples).

2.2.3 Response alternatives. In many studies, respondents are asked to report their behavioral frequencies by checking the appropriate alternative on a frequency scale of the type shown in Table 1. Assuming that the researcher constructed a meaningful scale, respondents commonly believe that the values in the middle range of the scale reflect the 'average' or 'usual' behavioral frequency, whereas the extremes of the scale correspond to the extremes of the distribution. Given this assumption, respondents can use the range of the response alternatives as a frame of reference in estimating their own behavioral frequency. This strategy results in higher estimates along scales that present high rather than low frequency response alternatives, as shown in Table 1. As may be expected, the impact of frequency scale is most pronounced for frequent and mundane behaviors, which are poorly represented in memory (see Schwarz 1996 for a review).

In addition to influencing respondents' behavioral reports, response alternatives may also affect subsequent comparative judgments. For example, a frequency of '$2\frac{1}{2}$h a day' constitutes a high response on the low frequency scale, but a low response on the high frequency scale shown in Table 1. Checking this alternative, a respondent may infer that their own TV consumption is above average in the former case, but below average in the latter. Hence, respondents in this study reported lower satisfaction with the variety of things they do in their leisure time when the low frequency scale suggested that they watch more TV than most other people.

Table 1
Reported daily TV consumption as a function of response alternatives

Low frequency alternatives		High frequency alternatives	
Up to $\frac{1}{2}$h	7.4 percent	Up to $2\frac{1}{2}$h	62.5 percent
$\frac{1}{2}$h to 1h	17.7 percent	$2\frac{1}{2}$h to 3h	23.4 percent
1h to $1\frac{1}{2}$h	26.5 percent	3h to $3\frac{1}{2}$h	7.8 percent
$1\frac{1}{2}$h to 2h	14.7 percent	$3\frac{1}{2}$h to 4h	4.7 percent
2h to $2\frac{1}{2}$h	17.7 percent	4h to $4\frac{1}{2}$h	1.6 percent
More than $2\frac{1}{2}$h	16.2 percent	More than $4\frac{1}{2}$h	0.0 percent

Note $N = 132$. Adapted from Schwarz et al. (1985). Response scales: Effects of category range on reported behavior and comparative judgments: Reprinted by permission from *Public Opinion Quarterly* **49**: 388–95

To avoid these systematic influences of response alternatives, it is advisable to ask frequency questions in an open response format, such as, 'How many hours a day do you watch TV? … hours per day.' Note that such an open format needs to specify the relevant units of measurement, for example, 'hours per day' to avoid answers like 'a few.'

As another alternative, researchers are often tempted to use vague quantifiers, such as 'sometimes,' 'frequently,' and so on. This, however, is the worst possible choice (see Pepper 1981 for a review). Most importantly, the same expression denotes different frequencies in different content domains, for example, 'frequently' suffering from headaches reflects higher absolute frequencies than 'frequently' suffering from heart attacks. Moreover, different respondents use the same term to denote different objective frequencies of the same behavior. For example, suffering from headaches 'occasionally' denotes a higher frequency for respondents with a medical history of migraine than for respondents without that migraine history. Accordingly, the use of vague quantifiers reflects the objective frequency relative to respondents' subjective standard, rendering vague quantifiers inadequate for the assessment of objective frequencies, despite the popularity of their use.

3. *Pretesting*

Given the complexity of the question answering process, extensive pretesting is of paramount importance. Traditionally, questions were often assumed to 'work' if respondents could provide an answer. Recent psychological research, however, facilitated the development of more sensitive pretest procedures (Sudman et al. 1996, Chap. 2, Schwarz and Sudman 1996). These procedures include the extensive use of probes and think-aloud protocols (summarily referred to as cognitive interviewing), detailed codings of interview transcripts, and the use of expert systems that alert researchers to likely problems. Drawing on these procedures, researchers can increase the odds that respondents understand the question as intended and use cognitive strategies that are likely to produce a meaningful answer.

See also: Content Validity; Questionnaires: Assessing Subjective Expectations; Sample Surveys, History of; Sample Surveys: Methods; Social Survey, History of

Bibliography

Bradburn N M, Rips L J, Shevell S K 1987 Answering autobiographical questions: The impact of memory and inference on surveys. *Science* **236**: 157–61

Clark H H, Schober M F 1992 Asking questions and influencing answers. In: Tanur J M (ed.) *Questions about Questions*. Sage, New York, pp. 15–48

Conway M A 1990 *Autobiographical Memory: An Introduction*. Open University Press, Milton Keynes, UK

DeMaio T J 1984 Social desirability and survey measurement: A review. In: Turner C F, Martin E (eds.) *Surveying Subjective Phenomena*, Vol. 2. Russell Sage Foundation, New York, pp. 257–281

Krosnick J A, Fabrigar L in press *Handbook of Attitude Questionnaires*. Oxford University Press, Oxford, UK

Pepper S C 1981 Problems in the quantification of frequency expressions. In: Fiske D W (ed.) *Problems with Language Imprecision* (*New Directions for Methodology of Social and Behavioral Science*), Vol. 9. Jossey-Bass, San Francisco

Ross M 1989 The relation of implicit theories to the construction of personal histories. *Psychological Review* **96**: 341–57

Schuman H, Presser S 1981 *Questions and Answers in Attitude Surveys*. Academic Press, New York

Schwarz N 1996 *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation*. Erlbaum Associates, Mahwah, NJ

Schwarz N, Hippler H J 1991 Response alternatives: The impact of their choice and ordering. In: Biemer P, Groves R, Mathiowetz N, Sudman S (eds.) *Measurement Error in Surveys*. Wiley, Chichester, UK, pp. 41–56

Schwarz N, Strack F, Hippler H J, Bishop G 1991 The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology* **5**: 193–212

Schwarz N, Sudman S (eds.) 1994 *Autobiographical Memory and the Validity of Retrospective Reports*. Springer Verlag, New York

Schwarz N, Sudman S (eds.) 1996 *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. Jossey-Bass, San Francisco

Sudman S, Bradburn N M 1983 *Asking Questions*. Jossey-Bass, San Francisco
Sudman S, Bradburn N M, Schwarz N 1996 *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco
Tanur J M (ed.) 1992 *Questions about Questions*. Sage, New York
Tourangeau R, Rips L J, Rasinski K 2000 *The Psychology of Survey Responses*. Cambridge University Press, Cambridge, UK

N. Schwarz

Quetelet, Adolphe (1796–1874)

1. Life and Career

Born in Ghent (Belgium) to a French father and Flemish mother, the young Quetelet started his career as a teacher, and besides his interest in mathematics, he also had artistic talents.

Quetelet was the first student to obtain a doctorate in 1819 at the newly founded State University of Ghent. In his dissertation he presented a new geometric function called the focal curve. In 1819 he was appointed mathematics professor at the Atheneum in Brussels and in 1820 he was elected to the *Académie royale des sciences et des belles-lettres*. The academy became the central place from which Quetelet directed all his other activities. In 1823 his scientific orientation took a decisive turn: he proposed the foundation of a new observatory to the Dutch government, and in the same year he was sent to Paris for further preparation of this plan. At the Paris observatory, contacts were established with Laplace, Poisson, and Fourier, and Quetelet's interests in probability theory and astronomy stem from this short period in France. The construction of the Brussels observatory was completed only in 1832, but in the meantime Quetelet had also become involved in the planning of the 1830 population census.

Unfortunately this census was never completed as a result of the Belgium's secession from The Netherlands in 1830. But, elected as secretary for life at the Academy in 1834, Quetelet remained in control not only of the activities in meteorology, astronomy, and physical geography located at the observatory, but also of the activities in demography and statistics at the *Bureau de Statistique* and, from 1841 on, at the *Commission centrale de Statistique*. The procedures to be used in the new census were tried out first in Brussels in 1842, and the nationwide census took place in October 1846. Both the method of census taking and the subsequent analyses with detailed breakdowns by age and marital status set new standards for similar undertakings in most other European countries. In the meantime Quetelet had gained an international reputation, first by the publication of his *Physique Sociale* in 1835, in which he brought together his earlier statistical work on demographic subjects, anthropometry, and criminology, and then by his relentless efforts to establish statistics as a new internationally oriented and empirically grounded scientific discipline. The latter was accomplished mainly through organizing a steady succession of international statistics conferences and by inspiring the foundation of statistical societies in other countries.

In 1855, however, Quetelet suffered a major stroke, and from then on his own scientific contribution was seriously reduced. But, until his death in 1874, he continued to inspire statistical applications in a wide variety of fields, and above all, to promote international comparability of statistical information. In Desrosières' words (1993, p. 95), 'Quetelet was the orchestra conductor of 19th Century statistics.'

2. Quetelet's Contribution to Statistics and Demography

Quetelet's statistical work was profoundly influenced by early probability theory—particularly the binomial distribution of events with equal odds—and by the use that Laplace had made of the Gaussian bell curve in astronomy. From Laplace he had gained the insight that the binomial distribution could also represent measurement error in astronomical observations, and Quetelet was convinced that this would also hold for measurements in the physical, social, and 'moral' domains. In the 1820s, measurements and frequency distributions pertaining to these domains and based on large numbers were still very rare. Quetelet had to use the frequency distribution of the chest circumferences of over 5,000 Scottish recruits to verify his hypothesis that these would conform to the binomial distribution (Hacking 1990, pp. 108–11). In a series of heights of French recruits, however, he detected a deviation from the expected distribution and therefore suspected the presence of systematic errors associated with attempts to avoid conscription. This illustrates how Quetelet used theoretical distributions to assess measurement reliability.

His use of the bell curve led him also into far deeper water: he proposed the notion of *l'homme moyen* or the average man, of which all members of a given population would be imperfect copies. For Quetelet, and for many after him, the emergence of a bell curve is a strong indication that there must be a link between the individuals producing this particular frequency distribution. Individuals would be drawn as balls from the same urn, and populations would correspond to different urns. The average men, each reconstructed from a different population, would therefore facilitate comparisons between populations, which themselves would be sufficiently homogeneous. Quetelet's 'aver-

age man' has often been misunderstood and misused. First, Quetelet was never preoccupied exclusively by averages. Whenever possible he presented complete distributions, and one of his contributions to demography is precisely his systematic presentation of *age-specific distributions* of vital events or of other occurrences (e.g., crime). This is the beginning of the statistical study of the life cycle. Second, Quetelet would be quite shocked to see his average man being associated with national prejudice, e.g., the stingy Scot, the thrifty Frenchman, ..., for the simple reason that these would all be *unmeasured* attributes. In fact, he perceived his average man as an antidote against statements based on prejudice, anecdote, and impression. It is for this reason that Quetelet can be considered as one of the founding fathers of modern empirical sociology: adequate statistical measurement presupposes operationalization with satisfactory reliability and validity.

The most fundamental critique of the average man came from A. Lexis who discovered that the variances of observed distributions are generally too large to uphold the thesis of balls drawn from the same urn. The hypothesis of a homogeneous population ceased to be tenable, and this would have major consequences for the advancement of statistics and for theories in the biological and social sciences (cf. Desrosières 1993, 111ff). In present day multivariate analyses we now routinely calculate 'little average men' in terms of subgroup means, odds ratios for different subpopulations at risk, and for different combinations of categories of co-variates. Furthermore, these numerous 'little average men' are compared and the results are tested for the presence or absence of significant differences.

The use of the bell curve was taken one step further by Quetelet's younger colleague and professor of mathematics at his alma mater in Ghent. Pierre-François Verhulst (1804–49) imagined that the growth rate r of a population would evolve according to a normal distribution. The size of a population, or of any other stock, would then follow a logistic curve, i.e. the elongated S-curve, with 50 percent of the total growth corresponding to the maximum of r, and then leveling off toward a saturation plateau. The logistic curve proved to have many applications: later demographic transitions produced population evolutions that very closely resemble the logistic curve, and diffusion processes modeled along the principle of contagion (of rumors, knowledge, disease, technology, etc.) or of competition commonly led to growth curves in accordance with Verhulst's logistic.

It is rather ironic that Verhulst's view of this particular form of population change did not feed back into Quetelet's own demographic work. Quetelet continued to think within the framework of a stationary population. Like several observers before him (e.g., Vauban, Süssmilch), Quetelet had been impressed by the observation that both numbers and distributions of vital events (deaths, births, marriages, ages at marriage, and age differences between spouses) showed a remarkable stability over time. His early work with Smits on the demography of the low countries had convinced him even more. Only major disturbances, such as a revolution, were capable of producing a temporary distortion (cf. his *causes constantes* et *causes accidentelles* in the *Physique sociale*). Quetelet kept thinking in terms of a homeostatic model, in the same way as Malthus had before him.

In his construction of the Belgian life table of 1841–50, centered around the census of 1846 and in his 'population tables' (population by age, sex, and marital status simultaneously) of 1850, Quetelet explicitly discusses the properties of a stationary population and shows that the actual age composition ought to be the same as the l_x or ${}_nL_x$ functions of the life table (l_x = number of survivors at each exact age x; ${}_nL_x$ = number of person-years lived in the age interval x to $x+n$). Furthermore, Quetelet goes on to show that the hypothesis of constant mortality could be relaxed (1835 p. 310, 1850 p. 16): 'the necessary condition for deducing a population table (i.e., an age structure) from a life table is that the deaths by age annually preserve the same ratios [in French: *rapports*] between them' (present author's translation and clarifications).

Quetelet was clearly on the way to show that there exists a neutral mortality decline which does not affect the shape of the age distribution (for the proof see A. J. Coale 1972 pp. 33–6). What Quetelet never developed—probably because there was not yet a need for it in pre-transition Belgium—was the model of a stable population with a growth rate different from zero. Yet, his demographic work shows that he can be considered as an early founding father of the branch of contemporary demography called 'indirect estimation.' In this branch extensive use is made of population models (stationary, stable, quasistable), of model life tables, and of the mathematical properties of particular model distributions for detecting and correcting errors in recorded data, and especially for estimating basic demographic parameters of mortality, fertility, and nuptiality on the basis of fragmentary information. Modern historical demography and the demography of developing countries progressed rapidly thanks to these techniques and ways of thinking.

3. *Quetelet's Contribution to Sociology*

In contrast to Auguste Comte, Quetelet never developed a general plan for a new social science, and by 1900 many of his ideas had already become outdated (cf. J. Lottin 1912, pp. 391–515). He is, however, highly typical of the beginnings of sociology: there is this new entity, called 'society,' that can be studied and ana-

lyzed from the outside with objective methods. This entity follows its own 'laws' (i.e., patterning, regularities) and these can be detected via statistical methods. At first, Quetelet's sociology had a deterministic ring, and his distinction between *causes constantes* and *causes accidentelles* led straight into the debate pitting collective determinism against individual freedom of action. Quetelet is clearly more impressed by the *causes constantes*, but in various places in his work he also proposed a solution to the conundrum. In an early letter to Villermé of 1832, for instance, Quetelet writes the following: 'As a member of society [in French: *corps social*], the individual experiences at each moment the social requirements [*la nécessité des causes*] and pays them regular tribute; but as an individual he masters these influences [*causes*] to some extent, modifies their effect and can seek to reach a superior state' (quoted in Desrosières 1993, pp. 103–4; present author's translation). In other words, through socialization the actor becomes an integral part of society, but individual and institutional agency (for Quetelet: the advancement of the sciences) play decisive roles in social change.

For many of his contemporaries the deterministic ring of the *Physique Sociale* of 1835 continued to be a source of steady criticism, particularly when Quetelet went on with his assertion that the *causes accidentelles* merely compensate each other. Later on, however, Emile Durkheim viewed the social influence on individuals and the feedback stemming from individual and collective action in very much the same way as Quetelet. Durkheim's equally holistic view of society, as in his *Formes élémentaires de la vie religieuse*, was directly influenced by Quetelet's work.

In actual practice, however, Quetelet remains essentially a master of 'comparative statics' rather than of 'social dynamics.' He paid little attention to the effects of early industrialization (definitely a *cause constante*) or to the impact of the economic, social, and health crises of the 1840s. This is very surprising for a chief statistician and demographer. The social reform movements, the rise in real wages, and the decline in adult mortality in Belgium since the 1860s also remain without comment. His successors in sociology were much more alert to analyzing the dynamics of the major societal transformations they were witnessing.

4. Contributions to Psychology

Quetelet's 1835 work on the development of human faculties has also had a major impact on the field of developmental psychology (Baltes 1983, Lindenberger and Baltes 1999). His work is the first effort to consider, based on empirical evidence, the nature of intellectual development from childhood into old age.

Treating human-ontogenetic development as a lifelong process is a fairly recent advent in psychology. Researchers in this field have been drawn to Quetelet's early efforts in this regard. They have also emphasized that Quetelet in his discussion of methodological issues was able to prefigure some of the issues that are of high significance for modern researchers. Among these are the role of period effects and selective survival.

See also: Censuses: Demographic Issues; Censuses: History and Methods; Demography, History of; Distributions, Statistical: Approximations; Distributions, Statistical: Special and Continuous; Distributions, Statistical: Special and Discrete; Elicitation of Probabilities and Probability Distributions; Individual/Society: History of the Concept; Life Table; Measurement Theory: History and Philosophy; Measurement Theory: Probabilistic; Population Size Estimation and Capture–Recapture Methods; Simultaneous Equation Estimates (Exact and Approximate), Distribution of; Statistical Methods, History of: Pre-1900; Statistical Systems: Censuses of Population; Statistics, History of

Bibliography

Académie Royale de Belgique 1997 *Actualité et universalité de la pensée scientifique d'Adolphe Quetelet*. Classe des sciences, actes du colloque 24–25.10.96, Brussels, Belgium

Baltes P B 1983 Life-span developmental psychology: Observations on history and theory revisited. In: Lerner R M (ed.) *Developmental Psychology: Historical and Philosophical Perspectives*. Erlbaum, Hillsdale, NJ, pp. 79–111

Coale A J 1972 *The Growth and Structure of Human Populations—A Mathematical Investigation*. Princeton University Press, Princeton, NJ

Desrosières A 1993 *La politique des grands nombres—Histoire de la raison statistique*. Editions la Découverte, Paris

Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK

Lindenberger U, Baltes P B 1999 Die Entwicklungspsychologie der Lebensspanne (Lifespan-Psychologie): Johann Nicolaus Tetens (1736–1807) zu Ehren. *Zeitschrift für Psychologie* **207**: 299–323

Lottin J 1912 *Quetelet, statisticien et sociologue*. Institut Supérieur de Philosophie, Louvain, France

Mailly E 1875 Essai sur la vie et les ouvrages de L-A-J. Quetelet. *Annuaire de l'Académie Royale de Belgique* **41**: 5–191

Porter T M 1986 *The Rise of Statistical Thinking*. Princeton University Press, Princeton, NJ

Quetelet A 1832 Sur la possibilité de mesurer l'influence des causes qui modifient les éléments sociaux—Lettre à M. De Villermé. *Correspondance Mathématique et Publique* **7**: 321–46

Quetelet A 1833 *Recherches sur le penchant au crimes aux différents âges*. M. Hayez, Brussels, Belgium

Quetelet A 1835 *Sur l'homme et le développement de ses facultés, ou essai de physique sociale*. Editions Bachelier, Paris [2nd edn. 1869 eds C. Muquart, Brussels]

Quetelet A 1847 De l'influence du libre arbitre de l'homme sur les faits sociaux, et particulièrement sur le nombre des marriages. *Bulletin de la Commission Centrale de Statistique Brussels* **2**: 135–55

Quetelet A 1846 *Lettres à S.A.R. le Duc Regnant de Saxe Cobourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politiques*. M. Hayez, Académie Royale, Brussels, Belgium
Quetelet A 1848 *Du système social et des lois qui le régissent*. Guillaumin et Cie, Paris
Quetelet A 1849 Nouvelles tables de mortalité pour la Belgique. *Bulletin de la Commission Centrale de Statistique*, *Brussels* **4**: 1–22
Quetelet A 1850 Nouvelles tables de population pour la Belgique. *Bulletin de la Commission Centrale de Statistique*, *Brussels* **4**: 1–24
Stigler S M 1986 *The History of Statistics—The Measurement of Uncertainty Before 1900*. Belknap Press of Harvard University Press, Cambridge, MA

R. Lesthaeghe

Queues

The intuitive notion that everybody has of the so familiar 'queues' or waiting lines closely parallels its mathematical meaning. Queueing systems are simplified mathematical models to explain congestion. This is a field of impressive activity and growth, which has mainly been concerned with the building of the models and the study of its probability properties. Relatively little effort has been put into their statistical analysis. In this article, the main concepts underlying simple queuing systems are discussed; some Bayesian inferential tools are also given.

1. *Queueing Systems*

Broadly speaking, a queueing system occurs any time 'customers' demand 'service' from some facility; usually both the arrival of the customers and the service times are assumed to be random. If all of the 'servers' are busy when new customers arrive, these will generally wait in line for the next available server. In some examples, the terms 'customer,' 'service' and 'server' have the usual, everyday meaning. For instance, queueing systems occur when customers check out in a grocery store, or require the service of a teller in a bank; here, the queueing term 'customer' refers to the actual customers at the store or the bank, the required service is to be checked out or to resolve the bank inquiry they have, and if the server (the check-out person or the teller) is busy when a customer require his/her service, then the customer has to wait in line, that is, queue up for service. In general, the terms 'customers,' 'service,' and 'servers' are very broad and can refer to almost anything; for instance, telephone calls ('customers') arriving at a switchboard (the service time being the time each call needs to be completed), machine failures (the 'customer' is the machine, and the server the repair person), demands of CPU time (server is the CPU, and the service the completion of the task), etc. Sometimes, a 'waiting line' (real or conceptual) never occurs. For instance, buses following their route can be considered a queueing system in which the customers are the buses, and the service is the completion of the route; here, customers are immediately 'served' and no waiting is necessary (technically, infinite servers, or as many servers as the number of buses, are assumed). In many interesting applications, customers require more than one service from different servers, and upon completion of service in one node, they have to proceed to another node. If the new server is busy, the customer has to queue up again for service. This is a typical situation modeled by the so called networks of queues, which are not treated in this article. An example occurs when patients enter an emergency department in a hospital: they might have to 'queue up' for primary attention, x-rays, lab service, surgery, and so on.

Queues (and queueing networks) have become very useful tools for evaluating (at least approximately) the performance of complex interacting stochastic systems in which congestion occurs. Obvious applications occur in airport terminals, health-care centers, and check-out lines at different counters. They have also been traditionally applied to biology (migration, population models), electrical engineering (flow models) and chemistry (polymerization, clustering) among others. The most prevalent applications of queueing systems and networks occur in the fields of manufacturing, computer networking, and telecommunications and broadcasting. But there are also many applications in the social sciences:

(a) Several types of airport queues: aircrafts waiting to take off or to land, queues of passengers at terminals (they have to go through ticket counters, check-in counters, security and passport controls, baggage carousels, customs, etc.).

(b) Police, fire, and emergency medical systems can all be regarded as queueing systems. Here the main concern is with responding to the calls for assistance as quickly as possible, important questions being to decide the number of units and which unit or units to dispatch to a specific call.

(c) Health-care systems provide many examples of analyses well suited for queueing systems, as modeling the movement of coronary patients in a hospital, determining the staffing of a department, modeling waiting queues for transplants, etc.

(d) In justice scenarios, cases queue up for trial (and the length of the trial depends on the actions of both the 'customer' (that is, the lawyers), and the judge.

(e) Analyzing and evaluating the delays in the mail, studying queues at voting polls to optimize the number (and type) of voting machines, and so on.

(f) Latent phase of diseases (such as Aids) in which the times of infection are arrivals to the queuing

system, the incubation is the 'service,' and the onset (or diagnosis) of the disease is the departure from the queuing system. Questions of interest are the number of people in the system and the time that 'customers' stay in the system (that is, from infection to onset or diagnosis).

Since the economical and social impact of congestion can be considerable, the importance of analyzing queueing systems has been clear ever since Erlang's pioneering work on telecommunications at the beginning of the twentieth century. For mathematicians, queues offer an added bonus, namely that the mathematics involved can be complex and intriguing; they also have an enormous potential for varying the conditions and the assumptions of the system. It does not come as a surprise that queues developed quickly into the huge area of study that it is today. Thousands of papers, and many journals, meetings, and books are entirely devoted to the subject. It is impossible to list here even the key references, and only a small selection will be given, from which other references can be obtained. Among the numerous books, it is worth mentioning the general books of Cox and Smith (1961), Takács (1962), Cooper (1981), Gross and Harris (1985), Medhi (1991), and Nelson (1995); Kleinrock (1975, 1976) and Newell (1982) are good sources of applications; Cooper (1990) is a nice survey with an extensive bibliography. Very useful reviews of many different aspects of queueing theory (including history, statistical analysis, simulation methods, applications, etc.) can be found in Dshalalow (1995, 1997).

Simple queueing systems are defined by specifying the following:

(a) *The arrival pattern*, which among other things can be deterministic or stochastic, individual or in batches (like travelers arriving at an airport passport control), etc. Also customers can abandon the line before being served, or can interact in some other ways with the rest of the customers.

(b) *The service mechanism*, which describes the number of servers, how many customers at the time the server can handle (usually one, but bulk servers such as buses, elevators, etc. can serve more than one), and the distribution of service times (which can be independent among servers, time-dependent, dependent on queue length, etc.).

(c) *Queue discipline*, usually FIFO (first-in-first-out), but others are possible, such as last-in-first-out (LIFO) (as the subroutines in a computer program), priority customers (as express check-out lines, giving priority to shorter service times), customers switching lines, etc. Besides these, other aspects can also be relevant, such as the capacity of the waiting room, or the size of the population of customers.

In the early 1950s, Kendall introduced a useful shorthand notation for simple queueing systems. It consists on a series of symbols separated by slashes, as $A/B/c$. 'A' refers to the distribution of the inter-arrival times between customers; thus Er stands for Erlang, G for General, M for Exponential (Markovian), etc. 'B' characterizes the service distribution for each server, with the same symbols as before, and 'c' refers to the number of servers, who are usually assumed to be equally efficient. More symbols are added when needed, to impose restrictions on the system capacity, and/or on the size of the population that feeds the system, and/or to specify the queue discipline if other than FIFO. One of the simplest and most studied queues is the $M/M/c$ queue, especially the $M/M/1$ and $M/M/\infty$; the latter are particularly useful for modeling self-service facilities or as approximations to systems with a very large number of servers (such as emergency services).

2. *Equilibrium: Measures of Performance*

From the probabilistic point of view, properties of queues are usually derived from the properties of stochastic processes associated with them. Let $N(t)$ denote the number of customers in the system at time $t \geqslant 0$. The *state probabilities* are defined as $P_j(t|\theta) = Pr(N(t) = j|\theta)$, for $j = 0, 1, 2, \ldots$, where θ denotes all parameters governing the arrival and service processes. The derivation of $P_j(t|\theta)$ is crucial in the probabilistic analysis of queueing systems. For many simple systems, the stochastic process $\{N(t), t \geqslant 0\}$ is a birth–death process (a 'birth' occurring with every arrival to the system, and a 'death' with every departure), greatly simplifying their study. (A brief introduction to birth–death, and other, stochastic processes with their connections to queues can be found, for example, in Ross (1993), as well as in the general references for queues given in Sect. 1). More complex systems sometimes resort to studying other *embedded* stochastic (hopefully Markovian) processes by contemplating the system as it would be seen by an arriving customer or by a departing customer.

In all but the simplest queues, determination of these probabilities is, unfortunately, extremely difficult. Often, however, it is possible to determine their large time limit, the so-called *equilibrium* or *steady-state* distribution

$$P_j(\theta) = \lim_{t \to \infty} P_j(t|\theta)$$

without actually having to compute $P_j(t|\theta)$. The equilibrium distribution does not depend on the initial conditions of the system and it is stationary (does not depend on t once the statistical equilibrium is reached). The ergodic conditions give the restrictions on θ under which the system will eventually reach the equilibrium (with probability one). Thus, for instance, in an $M/M/1$ queue, the ergodic condition establishes that the arrival rate (λ) should be smaller than the service rate (μ) (that is, the average number of customers arriving in a unit time is smaller than the average

number that the server can handle, a most intuitive condition).

For the most part, queueing theory deals with computations of the steady-state probabilities $P_j(\theta)$, and their use in computing other (steady-state) measures of performance of the queue, such as the expected number of customers in the system and/or in the waiting line, the distribution of the time spent queueing or in the system or their expected values, or the mean (or distribution) of the time the servers are busy (or idle). These steady-state distributions are also often very difficult to compute, and many require solving systems of differential equations, inversion of transforms (generating probability functions, Laplace transforms,…), etc. When only the expected values are required, an extremely useful formula for systems in equilibrium is *Little's law* which establishes that the expected number of customers in the system (L) is equal to the average time that a customer spends in the system (W) times the rate (λ) at which customers enter the system.

3. Statistical Analysis: A Bayesian Approach

Most of the vast effort in queueing theory has been devoted to the probabilistic development of queueing models and to the study of its mathematical properties; that is, the parameters governing the models are, for the most part, assumed given. Statistical analyses, in which uncertainty is introduced, are comparatively very scarce. A good review is Bhat et al. (1996). It could be argued that the general results of the (well developed) area of inference for stochastic processes (see for instance, Basawa and Prakasa Rao 1980, Lehoczky 1990) could be applied to queues. These general results do not take advantage of the special structure of queueing systems, however, nor do they address its specific questions.

Inference in queueing systems is not easy: From a frequentist perspective, development of the needed sampling distributions can be very involved and often the analysis is restricted to asymptotic results. Also, estimates can be different, for example, if the observational process differ, if the initial queue length is or is not included in the analysis. The statistical analysis is simpler if approached from the Bayesian perspective, and we take this approach below (see *Bayesian Statistics*). More arguments in favor of Bayesian analyses of queueing systems are given in Armero and Bayarri (1999), where a comprehensive list of references with Bayesian analyses of queues is also given. As an illustration of the main issues, the analysis of the much studied $M/M/1$ queue is outlined here.

According to the assumptions of the $M/M/1$ queue, the interarrival times, Y, are i.i.d. exponential with arrival rate λ, that is

$$f(y \mid \lambda) = \lambda e^{-\lambda y}, \quad y > 0. \tag{1}$$

Similarly, the service times X are also i.i.d., $X \sim \mathrm{Ex}(\mu)$, independently of Y, where μ is the service rate. That is

$$f(x \mid \mu) = \mu e^{-\mu x}, \; x > 0. \tag{2}$$

Steady-state distributions exist if and only if the ergodic condition

$$\rho = \lambda / \mu < 1 \tag{3}$$

holds, where ρ is the called traffic intensity and plays an important role in the study of $M/M/1$ queues.

A queue can be observed in a variety of ways and different quantities can be recorded. The most traditional experiments (providing complete information) in queues observe the system over a continuous period of time $(0, T]$. T can either be a value fixed in advance, or be determined by a suitable stopping rule. For instance, the system can be observed until the busy time reaches some preassigned fixed value, until a fixed number of customers have departed from the system, until a fixed number of transitions have been recorded, and so on (references can be found in Armero and Bayarri, 1997). Different quantities can be recorded (instants of arrivals, of departures, of commencing of service, number of arrivals and/or departures, interarrival times, service times, number of transitions, busy times, etc). If the queue is assumed to be in equilibrium, the observation of the number of customers at time 0 can also be added, providing more informative but rather cumbersome likelihood functions.

Suppose that an experiment providing n_a interarrival times, $Y_1, \ldots, Y_{n_a}$ and n_s service completions, $X_1, \ldots, X_{n_s}$ has been performed. Then, from (1) and (2), a likelihood function for (λ, μ) based on the observed $\mathbf{x} = (x_1, x_2, \ldots, x_{n_a})$ and $\mathbf{y} = (y_1, y_2, \ldots, y_{n_a})$ is

$$\ell(\lambda, \mu) \propto f(\mathbf{x}, \mathbf{y} \mid \lambda, \mu) \propto \lambda^{n_a} e^{-\lambda t_a} \mu^{n_s} e^{-\mu t_s} \tag{4}$$

where $t_a = \sum_{i=1}^{n_a} y_i$, $t_s = \sum_{j=1}^{n_s} x_j$ are the totals of observed interarrival times and service times, respectively. Notice that this likelihood is valid whether n_a and/or n_s are fixed or random (as long as the likelihood function is proportional to (4)). Also, the observation of the arrival and the service processes does not need to be simultaneous and the X's and Y's does not need to correspond to successive customers. (Of course, many of the observational procedures mentioned above result in likelihood functions proportional to (4).) Notice also that (4) applies whether or not the system is in equilibrium and whether or not equilibrium can eventually be reached.

Maximum likelihood estimators are immediate from (4), and given by

$$\hat{\lambda} = \frac{n_a}{t_a}, \quad \hat{\mu} = \frac{n_s}{t_s}, \quad \hat{\rho} = \frac{\hat{\lambda}}{\hat{\mu}} \tag{5}$$

However, standard errors of confidence intervals do require careful consideration of the specific stopping rule used, and different experimental designs will in general produce different sampling distributions for these estimators, and hence different standard errors and confidence intervals.

Bayesian analyses are insensitive to (noninformative) stopping rules, and all that it is required from the data is the likelihood function (4). A prior distribution has also to be assessed quantifying whatever is known about the system before the data is collected. Usually, there is plenty of information *a priori* about the queue, especially if it is assumed to be in equilibrium (which implies, at the very least, that the system has been working for an extended period of time). A prior distribution incorporating all this information should then be assessed. However, to keep the parallelism with the likelihood analysis and also to avoid the extra work of quantifying subjective beliefs, we use a Bayesian analysis usually called 'objective' because the prior distribution used is of the 'non-informative' or 'objective' type. Since both λ and μ are scale parameters, the invariant measures for each will be used, and they will be assumed independent *a priori*, that is

$$\pi(\lambda, \mu) \propto \frac{1}{\lambda}\frac{1}{\mu}. \tag{6}$$

Note that no assumptions about the stationarity of the queue are made in (6). In fact, they are not needed to produce posterior distributions, although they could be incorporated if desired (see (9) below). If it is wished to incorporate prior information about the values of λ and μ, a very simple prior distribution to use is the conjugate $\pi(\lambda, \mu) \propto Ga\,(\lambda \,|\, a, b)\; Ga\,(\mu \,|\, c, d)$, a product of two independent gamma distributions, with the hyperparameters a, b, c, d chosen to reflect the prior information. (Again, this prior makes no assumptions about stationarity). In fact, the prior (6) can be seen as a limiting case of this conjugate prior when all hyperparameters tend to 0 ('vague' prior information).

Combining (4) with (6) with Bayes theorem, we get the joint posterior distribution

$$\begin{aligned}\pi(\lambda, \mu \,|\, \mathbf{x}, \mathbf{y}) &\propto \pi(\lambda, \mu)\, f\,(\mathbf{x}, \mathbf{y} \,|\, \lambda, \mu)\\ &\propto Ga(\lambda \,|\, n_a, t_a)\; Ga(\mu \,|\, n_s, t_s)\end{aligned} \tag{7}$$

so that λ and μ are independent *a posteriori* with scaled chi-squared distributions $2t_\alpha\lambda \sim \chi^2_{(2n_a)}$, $2t_s\mu \sim \chi^2_{(2n_s)}$, and since they are independent, $(t_a/t_s)\rho \sim F_{2n_a, 2n_s}$, an F distribution. These distributions will be *exactly* valid for all experiments producing the likelihood function (4), irrespective of the stopping rule. Usual estimates of λ, μ and ρ are the means (or the modes) of their (marginal) posterior distributions, and 'standard errors' associated to these estimates are most naturally given by the standard deviations of those distributions. Intervals with a given probability content can similarly be given. All these will numerically reproduce their non-Bayesian counterparts when the design has n_a and n_s fixed in advance.

An important issue when studying queues is how likely it is that steady-state can be reached. From the marginal distribution of ρ, this and other questions of direct interest can be answered. In particular, the probability that the ergodic condition holds is computed from the F distribution above as

$$\begin{aligned}\pi_e = Pr(\rho < 1 \,|\, \mathbf{x}, \mathbf{y}) &= \int_0^1 \pi(\rho \,|\, \mathbf{x}, \mathbf{y})\mathrm{d}\rho\\ &= \Psi_{(2n_a, 2n_s)}(t_a/t_s)\end{aligned} \tag{8}$$

where $\Psi_{(2n_a, 2n_s)}$ is the distribution function of an $F_{2n_a, 2n_s}$ distribution.

Most often, queues are assumed to be in equilibrium. Under this assumption, the restriction $\rho < 1$ imposed by the ergodic conditions should be incorporated into the analysis. This is easily done, and in fact amounts to truncating the unrestricted posterior distribution for ρ giving

$$\pi(\rho|\mathbf{x}, \mathbf{y}, \rho < 1) = \frac{\pi(\rho \,|\, \mathbf{x}, \mathbf{y})}{\pi_e} \quad \text{for}\, \rho < 1, \tag{9}$$

and $\pi(\rho \,|\, \mathbf{x}, \mathbf{y}, \rho < 1) = 0$ for $\rho \geqslant 1$. From (9), estimates, standard errors, credible intervals, etc. that take into account the restriction can be produced. (In this simple case, the probability π_e in (8), the proportionality constant in (9) and the moments of the truncated posterior can be given close form expressions in terms of Gauss' hypergeometric functions.) In Armero and Bayarri (1999), an explicit expression for a factor that can be used for adjusting the usual MLE $\hat{\rho}$ is given, as well as (adjusted) 'standard errors.' It can there be seen that incorporating the restriction $\rho < 1$ has the effect of 'shrinking' the usual MLE, the shrinkage being larger the larger $\hat{\rho}$ and the smaller the sample sizes, a most intuitive result.

When statistically analyzing queueing systems, inference about the parameters is rarely (if ever) of direct interest. Usually the main goal is prediction of observable quantities that describe the utilization and degree of congestion of the system. From the Bayesian perspective, it is also very easy to produce predictive distributions incorporating the uncertainty in the estimation of the unknown parameters. The predictive distribution of the steady-state number, N, of customers in the system is computed as

$$\begin{aligned}&p(N = n \,|\, \mathbf{x}, \mathbf{y}, \rho < 1)\\ &\quad = \int_0^1 Pr(N = n \,|\, \rho)\pi(\rho \,|\, \mathbf{x}, \mathbf{y}, \rho < 1)\, \mathrm{d}\rho\end{aligned} \tag{10}$$

where $Pr(N = n \,|\, \rho)$ is given by the Geometric $Ge(1$

$-\rho$) distribution and $\pi(\rho \mid \mathbf{x}, \mathbf{y}, \rho < 1)$ is given in (9). In this case, again, (10) can be given a close form expression (see Armero and Bayarri 1994, for details).

Similar arguments apply for prediction of the number of the (steady-state) number of customers in the waiting line, time spent by a customer in the system (or waiting in line), length of busy (idle) periods, etc. All these predictive distributions are a most natural tool for deciding future interventions in the queue (see Armero and Bayarri 1997).

4. Conclusion

Queues systems are useful to stochastically model systems in which subjects interact and congestion occurs. They can describe a wide variety of such situations, in virtually all areas of application. In probabilistic terms, queues have been extensively studied under the assumption that the governing parameters are known. Inferences about these parameters can be very challenging when approached from a frequentist perspective, and the common approach of simply replacing the unknown parameters by some estimates when predicting measures of performance of the queues can produce deceiving results. Bayesian methods appear to be ideally suited to handle inferential aims in queuing systems, as incorporating restrictions in the parameter space (imposed by the ergodic conditions), predicting measures of performance, designing the system, etc. Moreover, its independence of the stopping rule makes the inferential task much simpler.

Bibliography

Armero C, Bayarri M J 1994 Bayesian prediction in $M/M/1$ queues. *Queueing Systems* **15**: 401–17

Armero C, Bayarri M J 1997 Bayesian questions and answers in queues. In: Bernardo J M, Berger J O, Dawid A P, Smith A F M (eds.) *Bayesian Statistics 5*. Oxford University Press, Oxford, UK, pp. 3–23

Armero C, Bayarri M J 1999 Dealing with uncertainties in queues and networks of queues: a Bayesian approach. In: Ghosh S (ed.) *Multivariate Analysis, Design of Experiments and Survey Sampling*. Marcel Dekker, New York, pp. 579–608

Basawa I V, Prakasa Rao B L S 1980 *Statistical Inference for Stochastic Processes*. Academic Press, New York

Bhat U N, Miller G K, Rao S S 1996 Statistical analysis of queueing systems. In: Dshalalow J H (ed.) *Frontiers in Queueing: Models and Applications in Science and Engineering*. CRC
Press, Boca Raton, FL, pp. 351–94

Cooper R B 1981 *Introduction to Queueing Theory*. 2nd edn. North-Holland, New York

Cooper R B 1990 Queueing Theory. In: Heyman D P, Sobel M J (eds.) *Hand-books in OR and MS, Vol 2*. Elsevier, Amsterdam, pp. 469–518

Cox D R, Smith W L 1961 *Queues*. Wiley, London

Dshalalow J H (ed.) 1995 *Advances in Queueing: Theory, Methods, and Open Problems*. CRC Press, Boca Raton, FL

Dshalalow J H (ed.) 1997 *Frontiers in Queuing. Models and Applications in Science and Engineering*. CRC Press, Boca Raton, FL

Gross D, Harris C M 1985 *Fundamentals of Queueing Theory*. 2nd edn. Wiley, New York

Kleinrock L 1975 *Queueing Systems, Vol. 1: Theory*. Wiley, New York

Kleinrock L 1976 *Queueing Systems, Vol. 2: Computer Applications*. Wiley, New York

Lehoczky J 1990 Statistical Methods. In: Heyman D P, Sobel M J (eds.) *Handbooks in OR and MS, Vol 2*. Elsevier, Amsterdam, pp. 469–518

Medhi J 1991 *Stochastic Models in Queueing Theory*. Academic Press, Boston

Nelson R 1995 *Probability, Stochastic Processes and Queueing Theory*. Springer Verlag, New York

Newell G F 1982 *Applications of Queueing Theory*. 2nd edn. Chapman and Hall, London

Ross S M 1993 *Introduction to Probability Models*. 5th edn. Academic Press, New York

Takács L 1962 *Introduction to the Theory of Queues*. Oxford University Press, New York

C. Armero and M. J. Bayarri

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

R

Race and Gender Intersections

In the 1980s and 1990s there has been a proliferation of scholarship by and about women of color that challenges heretofore accepted theories of society. Yet, this writing is even more marginalized than thinking that centers on white women. While a great deal of feminist work has been written to challenge existing sociological theory, there has been only limited impact on new theory formation. Currently, in most social theory texts, gender generally is segregated into a section on feminist theory, or ignored completely. Moreover, few articles in the *American Journal of Sociology* and the *American Sociological Review* contain discussions of feminist theory.

Thus, while 'feminist theory' remains marginalized, or ignored, the theorizing by and about women of color remains marginalized within the margins. Lest we think that this is simply a byproduct of white male hegemony, this marginalization is also present in feminist theorizing. To be sure, most edited volumes now include articles by and about women of color, the question remains to what extent have these writings influenced mainstream feminist theory?

Theorizing by and about women of color has had a bifurcated influence on the development of feminist theory. As Hooks (2000) notes, there is a 'high' and 'low' theory, and the former still continues to exclude much of the experientially based theorizing of women of color. Caraway (1991) notes that 'in academic practice, some white feminists have exhibited a tendency to ghettoize the creative and expressive writings of Black women as "empirical" supplements to the "formal" theory texts of white feminists.' For example, in a brief overview of some of the main feminist theory texts, women of color generally are segregated into a section on difference, multiculturalism, or women of color. Or, the text is peppered with reprints of 1980s writings by prominent women of color such as Bell Hooks, Pat Hill Collins, Adrienne Rich, Patricia Williams, Audre Lorde, Gloria Anzaldua, and Angela Davis. But often more recent scholarship on women of color is absent. In an overview of the literature, Glenn (1999) concludes that, '… despite increased recognition of the interconnectedness of gender and race, race remained undertheorized even in the writings of women of color. In the absence of a "theory" of race comparable to a "theory" of gender, building a comprehensive theory has proven elusive.' Following the work of Liu (1991), Higginbotham (1992), Kaminsky (1994), and Stoler (1996), Glenn calls for a synthesis through social constructionism, arguing that race and gender are 'relational concepts whose construction involves both representation and social structural processes in which power is a constitutive element' (1999).

1. *Intersectionality Theory*

Theorizing about women of color has had a major impact on postmodernist feminist theory. Postmodernism is committed to the deconstruction of absolute truths and views knowledge as socially constructed. Women of color have developed 'Intersectionality Theory' that subscribes to many postmodern tenets, and is committed to what McLaren (1994) calls resistance postmodernism. As Cordova (1998) states, 'Along with the critiques of modernism comes the notion that social "facts" are social constructions; social reality is mediated by one's positionality; that identity is tied to location (subjectivity); and that meaning is unstable and connected to the advanced capitalist construction of consumerism.'

'Intersectionality Theory' objects to the notion that all truths are equal and to the notion that individual narratives should not become a basis for shared resistance to oppressive power relations. So, while postmodernism emphasizes deconstructionism, 'Intersectionality Theory' insists upon an equal attention to constructionism. Equally important is that 'Intersectionality Theory' is written in a language more readily accessible to the general public than is that of French Postmodernism. It is often written in the language of experience.

Many women of color have contributed to the development of 'Intersectionality Theory,' which is interdisiplinary (Hooks 1984, Anzaldua 1987, 1990, Collins 1990, 1998, Crenshaw 1989, 1991, 1997, de la Torre and Pesquera 1993, Glenn 1985, Lorde 1984, Williams 1991, Zinn and Dill 1993, Brewer 1990, King 1988, Espiritu 1997). Collins (1990) argued for a more complex conceptualization of gender relations; one that included an understanding of the 'matrix of domination,' that is the intersection of gender, class, race, sexual preference, and other relevant systems of oppression. For Collins and other feminists of color,

these intersections are not simply separate cumulative variables by which we can assess one's experience of oppression. Rather it is the way these positions interlock in interactive or unique ways at their intersections that provides the experience of oppression. She writes: 'replacing additive models of oppression with interlocking ones creates possibilities for new paradigms. The significance of seeing race, class, and gender as interlocking systems of oppression is that such an approach fosters a paradigmatic shift of thinking inclusively about other oppressions, such as age, sexual orientation, religion and ethnicity.'

Hooks (1989) offers a similar analysis with her notion of a 'politic of domination,' arguing that, all of these axes are imbued with 'notions of superior and inferior … component.' 'Intersectionality Theory' simultaneously challenges multiple levels of oppression, while viewing race and gender as something more than the sum of their parts (King 1988). This theoretical approach arose in parallel with mainstream feminist theory as evidenced by the statement of the Combahee River Collective (1983), which was formed in the mid-1970s by a group of Black lesbian feminists. They wrote: 'The most general statement of our politics at the present time would be that we are actively committed to struggling against racial, sexual, heterosexual, and class oppression, and see as our particular task the development of integrated analysis and practice based upon the fact that the major systems of oppression are interlocking. The synthesis of these oppressions creates the conditions of our lives. As Black women we see Black feminism as the logical political movement to combat the manifold and simultaneous oppressions that all women of color face.'

'Intersectionality Theory' is more than the notion that the whole does not equal the sum of its parts. It also encompasses some creative, novel, methodological approaches to the study of social phenomena. For example, Sandoval (1998) writes about the extent to which Anzaldua's (1987) book influences feminists of color: 'This "borderlands" feminism, many argue, calls up a syncretic form of consciousness made up of transversions and crossings; its recognition makes possible another kind of critical apparatus and political operation in which *mestiza* feminism comes to function as a working chiasmus (a mobile crossing) between races, genders, sexes, culture, languages, and nations. Thus conceived, *La conciencia de la mestiza* makes visible the operation of another metaform of consciousness that insists upon polymodal forms of poetics, ethics, identities, and politics not only for Chicanas/os but for any constituency resisting the old and new hierarchies of the coming millennium.'

In this view, the study of any social problem entails a holistic approach, one that includes analyses of race, class, gender, sexuality, language, culture, and nation.

2. *Intersectionality Theory and Mainstream Sociology*

'Intersectionality Theory' encompasses a methodological approach critical to the study of social phenomena, and it has been employed by many women of color social scientists in such areas as the study of family, social movements, race and ethnicity, social stratification, legal institutions and crime, organizations, work and welfare, and politics. The following discussion provides examples of ways in which analysis of the intersectionality of race and gender enrich our theoretical understanding in several areas of study.

Let us begin with family studies. Wilson's (1987) work on the American Black underclass, and its implications for the Black family, has gained national attention. By contrast, the research conducted by Black women scholars, many of whom contextualize his findings, has received less attention. Dickerson (1995) examines predominant approaches to the study of African-American families that often judge them by their similarity to white families. She finds these approaches less useful than a form of 'centrism,' an Afrocentric paradigm that 'centers' groups in terms of their own cultural meanings, attitudes, and values, and gathers data from the *standpoint* of the group being studied. This approach is very much in keeping with 'Intersectionality Theory.'

Using this approach, Sudarkasa (1993) and Jewell (1988) both argue against the claim that 'female-headed households are the root cause of the deplorable conditions in which many Blacks find themselves in urban ghettos' (Sudarkasa 1993). Rather, they argue that the most critical factors for family survival and success are economic resources and proximity to kin. Harriet McAdoo's research indicates that kin systems are critical for the survival of black single mothers and the elderly. Wilson (1987) seems to drop the importance of kin and translates the need for economic resources into a need for a marriageable pool of Black males. Black male unemployment, he argues, is a critical intervention in the fight to uplift the underclass. While this is certainly of central importance, Black women's employment is equally important. After all, many upper-class white women are choosing to bear children out of wedlock, but there has not been the corresponding outcry, because they are economically capable of sustaining themselves and their offspring. But it is most difficult for Black women to support their children. Black women are disadvantaged by both racial and gender differences in earnings, which persist even after levels of educational attainment are held constant (Herring 1999).

A perspective that emphasizes the intersectionality of race, class, and gender are absent in most mainstream theories of racial/ethnic families, and this omission leaves such theories inadequate.

This pattern of omission is also visible in social movement theory. While feminists such as Verta

Taylor, Myra Marx Ferree, Pam Oliver, Nancy Whittier, and several others have made significant contributions to social movement theory, scholarly contributions by and/or about women of color are largely absent from major theoretical texts in this area, ignoring the contributions of the few women of color who are writing in this area, such as Patricia Zavella, Belinda Robnett, and Bernice McNair Barnett. Zavella's (1987) work on union organizing has shown that recruitment strategies must specifically address issues of race/class/gender as they affect the reality of potential recruits. Both Barnett (1993) and Robnett (1996, 1997) have shown the importance of Black women's contributions to the civil rights movement. Robnett (1997) argues that the intersectionality of race/class/gender shaped the organization of social movement organizations and leadership in social movements. Robnett (1997) shows that the experiences and contributions of women in the civil rights movement are varied by race and class. African-American women were critical bridge leaders in the civil rights movement, forging ties between movement organizations and the grass roots participants. Without their participation, the movement would not have succeeded. Her intersectionality approach lends insight into the decline of that movement, as the clash between white politicians, Black male elites, and rural Black women led to the disillusionment and disengagement of grass roots leadership. Scholars such as Dill (1988), Gilkes (1988), and Sacks (1988), have written about the redefinition of what we define as 'political,' suggesting that the boundaries are much too narrow, which has important implications for social movement theory, including what counts as a political identity or political activity.

Indeed, the theorizing by and about women of color has been excluded in a number of areas including work, the state, and the sociology of knowledge. For example, while feminist scholars have analyzed social reproduction, that is, childbearing and childcare, its racialization has been noted less. Glenn (1999) points out that while scholars theorize about how gender leads women to be assigned to child care in the context of a sex division of labor, they often ignore the fact that women of color are the ones who end up in the situation of leaving their own children at home or with kin, while they take care of other women's children to make the money to support their own children. Class, race, and gender affect the life chances of women of color, making them the group often concentrated in low paid childcare because of a lack of better employment options.

Glenn notes a similar omission in studies of citizenship. Whereas many feminists such as Pateman (1988), Young (1989), and Okin (1979) have challenged conceptualizations of the 'universal citizen,' short shrift has been paid to race/gender. Glenn (1999) asserts, 'The problem with looking at citizenship as only gendered or only raced is the familiar one: women of color fall through the cracks.'

Collins's (1999) critique of feminist analyses of science is equally instructive. She suggests that, 'feminist analyses of gender and scientific knowledge might benefit from closer association with emerging scholarship on intersectionality.' Collins convincingly illustrates the ways in which 'selecting a specific social location, social practice, group history, or topic, and subjecting it to an intersectionality analysis,' challenges the notion that Western science is an objective, apolitical, ahistorical endeavor.

In short, the social location (in the margin in the margin) of the scholarship by and about women of color, impedes the development of our understanding of social phenomena.

See also: Altruism and Prosocial Behavior, Sociology of; Civil Rights Movement, The; Conflict Sociology; Critical Race Theory; Feminist Theory and Women of Color; Gender and Feminist Studies; Gender and Feminist Studies in Sociology; Gender, Class, Race, and Ethnicity, Social Construction of; Gender History; Multicultural Feminism: Cultural Concerns; Multiculturalism: Sociological Aspects; Poverty and Gender in Affluent Nations; Race: History of the Concept; Race Identity; Racial Relations; Racism, History of; Racism, Sociology of

Bibliography

Anzaldua G 1987 *Borderlands/La Frontera: The New Mestiza*. Aunt Lute Press, San Francisco

Anzaldua G (ed.) 1990 *Making Face, Making Soul: Haciendo Caras: Creative and Critical Perspectives by Women of Color*. Aunt Lute Press, San Francisco

Barnett B 1993 Invisible southern black women leaders in the civil rights movement: The triple constraints of gender, race, and class. *Gender and Society* **7**(2): 162–82

Brewer R 1990 In: Albrecht L, Brewer R M (eds.) *Bridges of Power: Women's Multicultural Alliances*. New Society Publishers, Philadelphia, PA

Brewer R 1993 Theorizing race, class and gender: The new scholarship of Black feminist intellectuals and Black women's labor. In: James S M, Busia A P A (eds.) *Theorizing Black Feminisms: The Visionary Pragmatism of Black Women*. Routledge, New York, pp. 13–30

Caraway N 1991 *Segregated Sisterhood: Racism and the Politics of American Feminism*. University of Tennessee Press, Knoxville, TN

Collins P H 1990 *Black Feminist Thought: Knowledge, Consciousness, and Empowerment*. Unwin Hyman, Boston

Collins P H 1998 *Fighting Words: Black Women and the Search for Justice*. University of Minnesota Press, Minneapolis, MN

Collins P H 1999 Moving beyond gender: Intersectionality and scientific knowledge. In: Ferree M M, Lorber J, Hess B (eds.) *Revisioning Gender*. Sage, Thousand Oaks, CA, pp. 261–84

Combahee River Collective 1983 Combahee river collective statement. In: Smith B (ed.) *Home Girls: A Black Feminist Anthology*. Kitchen Table—Women of Color Press, New York, pp. 272–82

Cordova T 1998 Power and knowledge: Colonialism in the academy. In: Trujillo C (ed.) *Living Chicana Theory*. Third Woman Press, Berkeley, CA, pp. 17–45
Crenshaw K 1989 Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. *University of Chicago Legal Forum* **7**: 139–67
Crenshaw K 1991 Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review* **43**: 6
Crenshaw K 1997 Intersectionality and identity politics: Learning from violence against women of color. In: Shanley M, Narayan U (eds.) *Reconstructing Political Theory: Feminists Perspectives*. Pennsylvania State University Press, University Park, PA
De la Torre A, Pesquera B (eds.) 1993 *Building with Our Hands: New Directions in Chicana Studies*. University of California Press, Berkeley, CA
Dickerson B 1995 (ed.) *African American Single Mothers: Understanding their Lives and Families*. Sage, Thousand Oaks, CA
Dill B T 1988 Making your job good yourself: Domestic service and the construction of personal dignity. In: Bookman A, Morgen S (eds.) *Women and the Politics of Empowerment*. Temple University Press, Philadelphia, PA, pp. 33–52
Espiritu Y L 1997 *Asian American Women and Men*. Sage, Thousand Oaks, CA
Gilkes C T 1988 Building in many places: Multiple commitments and ideologies in black women's community work. In: Bookman A, Morgen S (eds.) *Women and the Politics of Empowerment*. Temple University Press, Philadelphia, PA, pp. 53–76
Glenn E N 1985 Racial ethnic women's labor: The intersection of race, gender, and class oppression. *Review of Radical Political Economics* **17**: 86–108
Glenn E N 1999 The social construction and institutionalization of gender and race: An integrative framework. In: Ferree M M, Lorber J, Hess B B (eds.) *Revisioning Gender*. Sage, Thousand Oaks, CA, pp. 3–41
Herring C 1999 African Americans in contemporary America: Progress and retrenchment. In: Dworkin A G, Dworkin R J (eds.) *The Minority Report*. Harcourt Brace, Orlando, FL, pp. 181–208
Higginbotham E 1992 African American women's history and the metalanguage of race. *Signs* **17**: 251–74
Hooks B 1984 *Feminist Theory from Margin to Center*. South End Press, Boston
Hooks B 1989 *Talking Back: Thinking Feminist, Thinking Black*. South End Press, Boston
Hooks B 2000 Theory as liberatory practice (from *Teaching to Transgress* (1994)). In: Kolmar W, Bartkowski F (eds.) *Feminist Theory: A Reader*. Mayfield Publishing, Mountain View, CA, pp. 28–33
Jewell K S 1988 *From Mammy to Miss America and Beyond: Cultural Images and the Shaping of US Social Policy*. Routledge, New York
Kaminsky A 1994 Gender, race, raza. *Feminist Studies* **20**: 3–32
King D K 1988 Multiple jeopardy, multiple consciousness: The context of a black feminist ideology. *Signs* **14**(1): 42–72
Liu T 1991 Teaching differences among women from a historical perspective: Rethinking race and gender as social categories. *Women's Studies International Forum* **14**(4): 265–76
Lorde A 1984 *Sister Outsider: Essays and Speeches*. Crossings Press, Trumansburg, NY
McLaren P 1994 *Critical Pedagogy and Predatory Culture: Oppositional Politics in a Postmodern Era*. Routledge, New York
Okin S 1979 *Women in Western Political Thought*. Princeton University Press, Princeton, NJ
Pateman C 1988 *The Sexual Contract*. Polity, Cambridge, MA
Robnett B 1996 African American women in the civil rights movement: Gender, leadership, and micromobilization. *American Journal of Sociology* **101**(6): 1661–93
Robnett B 1997 *How Long? How Long? African American Women in the Struggle for Civil Rights*. Oxford University Press, New York
Sacks K 1988 *Caring by the Hour*. University of Illinois Press, Chicago
Sandoval C 1998 Mestizaje as method: Feminists-of-color challenge the canon. In: Trujillo C (ed.) *Living Chicana Theory*. Third Woman Press, Berkeley, CA, pp. 352–70
Stoler A 1996 Carnal knowledge and imperial power: Gender, race and morality in colonial Asia. In: Scott J W (ed.) *Feminism and History*. Oxford University Press, New York, pp. 209–66
Sudarkasa N 1993 Female-headed African American households: Some neglected dimensions. In: McAdoo H P (ed.) *Family Ethnicity: Strength in Diversity*. Sage, Newbury Park, CA, pp. 81–9
Williams P 1991 *The Alchemy of Race and Rights: Diary of a Law Professor*. Harvard University Press, Cambridge, MA
Wilson W J 1987 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago
Young I 1989 Polity and group difference: A critique of the ideal of universal citizenship. *Ethics* **99**: 250–74
Zavella P 1987 *Women's Work and Chicana Families: Cannery Workers of the Santa Clara Valley*. Cornell University Press, Ithaca, NY
Zinn M B, Dill B T (eds.) 1993 *Women of Color in United States Society*. Temple University Press, Philadelphia, PA

B. Robnett

Race and the Law

1. Definition of the Field

The study of race and the law is most comprehensively developed and best defined as a genre of scholarship in the US. There the field is primarily, but far from exclusively, occupied by scholars trained in law and working in law schools. Significant contributions to the field have also been made in Canada and Europe, where the field is more evenly divided among lawyers and social scientists trained in anthropology, sociology, and cultural studies.

1.1 Mainstream vs. Critical Scholarship

Scholarship addressing race and the law is split between mainstream approaches and critical ap-

proaches. Mainstream approaches generally conceive of racial discrimination as a minor and isolated exception to the otherwise liberal and enlightened societies of the industrialized West. They understand racial injustice as largely unconnected to other institutions in the societies in which it operates and are therefore optimistic that racism can be eliminated 'surgically,' without requiring significant or widespread changes in those societies. Critical scholarship by contrast understands racism as deeply implicated in the institutions and traditions of societies and is therefore accordingly both more radical in its prescriptions and more circumspect in its predictions of success (Bell 1989, Crenshaw 1995, Williams 1991).

Because the definition of race is malleable, fluctuating, and contested, the study of race and the law encompasses not only 'race' as it is colloquially understood but also issues of ethnicity, national origin, caste, and religious and cultural difference. The field could therefore be understood to include the study of ethnic and nationalist divisions in Asian countries, caste hierarchy in Hindu societies, tribal conflict in central Africa, and religious conflict worldwide. It implicates the study of nationalism and pan-nationalist movements, communities of diaspora and indigenous peoples. Moreover, the field is deeply concerned with the intersection of race with issues of gender, sexual orientation, and class (Crenshaw 1995, A Harris 1990, Thomas 1995, Williams 1991).

Generally speaking, the study of race and the law encompasses two major projects. The first is a critique that demonstrates how legal institutions and formal laws produce and perpetuate racism and racial hierarchy. The second is the analysis of legal reform that might eliminate or minimize the effects of racial hierarchy. This article will address each in turn.

1.2 The Critique of Law as a Cause of Racism

Legal scholarship addresses the various ways in which law has produced and continues to produce racial hierarchy. Law contributes to racism and racial hierarchy in at least three distinct ways. First, law codifies racial categories, which are then employed by both public and private actors to erect and maintain racial hierarchies. Second, law institutionalizes racial hierarchy by creating formal rules requiring or encouraging racial hierarchy: by segregating the races, by excluding the members of certain races from participation in social, economic, and political activities and institutions, or by legally codifying the subordinate status of one or more races. These first two types of legal action are commonly known as *de jure* racial discrimination because they explicitly employ or create racial categories by force of law. Third, law allows, encourages, and requires the maintenance of institutional practices and social customs that perpetuate racial subordination by entrenching the racial hierarchies of the past, even when no formal legal rule requires or promotes racial subordination on its face. This latter type of legal action is known as *de facto* racial discrimination because it in fact produces and enforces racial subordination, although it does not do so explicitly.

2. Racial Categorization

Many scholars have studied the development and use of racial classificatory schemes (Davis 1991, Haney-Lopez 1994, Omi and Wynant 1994). Virtually all scholars agree that, historically, schemes of racial classification were not based on sound scientific theories and therefore do not reflect any 'real' biological differences between the races. Given the dubious scientific basis for any racial classification of human beings, it is not an exaggeration to assert that law not only enforced racial hierarchy, it also (along with other social institutions and practices) in fact produced a specific type of racial difference.

2.1 The Ontological Status of Racial Categories

There is disagreement about the ontological status of racial categories and the centrality of law in producing racial categories. Some scholars believe that racial categories have a substance that predates or transcends legally enforced and encoded racial categorization. While few scholars would argue that races are well understood in terms of biological or genetic difference, many argue for or assume that cultural differences distinguish the races (Gotanda 1995, Peller 1995). The belief that racial categories describe an essence or positive (in the sense of 'positive science' not necessarily in the sense of 'good') attribute of human beings—whether biological or cultural—is known as racial essentialism.

Other scholars argue that social practices, including legal codification, produced racial differences where there were no such differences before. These scholars contend race is not an attribute of individuals or groups but instead the effect of a social discourse that ascribes difference to individuals and groups (Appiah 1996, Ford 1999). They insist that without the discourse of race, there would be no races. The belief that race is not an attribute or essence of individuals or groups but is instead the effect of the social discourses that describe it and the social practices that make it relevant is known as anti-essentialism or social constructivism.

2.2 The Legal Codification of Racial Categories

Legal historians have examined the development and transformation of racial categories and definitions. A

vast literature details the variety of racial classificatory schemes employed in the USA, Europe, and South Africa. The contemporary literature is overwhelmingly critical of most racial categorization, both because of its opportunistic use of scientific data (or its use of questionable science to justify racial categories) and because of the oppressive use to which racial categories have been put.

There is, however, some dissension on the question of remedial or so-called benign racial classification. A widespread understanding of constitutional law and a pervasive popular attitude holds that the legal codification of *any* racial classification is socially deleterious and morally indefensible (Thernstrom and Thernstrom 1999). This view asserts that the law should be *race blind*. Many scholars who may not insist on strict color blindness nevertheless have serious misgivings about the use of racial classifications, even for remedial purposes (Carter 1992). Others believe that racial classification is not *per se* undesirable and argue that remedial racial classification is the necessary antidote to centuries of oppressive racial classification. They assert that the law should be affirmatively *race conscious* (Crenshaw 1995, Gotanda 1995, Matsuda and Lawrence 1997, Peller 1995).

2.3 The Legal Enforcement of Racial Hierarchy

Along with the study of racial categorization, a core concern of the study of race and the law is the legal enforcement of racial hierarchy. Of course the two are closely related: racial categories developed hand in glove with policies of racial exclusion, segregation, disenfranchisement, and involuntary servitude. The study of racial hierarchy has traditionally distinguished between formally enforced racial distinctions, hierarchies, and exclusions—*de jure* racial practices—and the formal requirement and enforcement of practices that have the effect of establishing or perpetuating racial hierarchy—*de facto* racial practices.

2.3.1 De jure *racial subordination*. Legal historians have detailed the legal mechanisms that historically have enforced and perpetuated racial hierarchy. Historians of the USA study the institution of racial hierarchy through chattel slavery (Tushnet 1981) and the post-emancipation development of 'Jim Crow' laws, which required racial segregation in public institutions and accommodations such as school rest rooms, drinking fountains, eating facilities, railroad coaches, and theater or stadium seating (Woodward 1989). Scholars of South Africa study the emergence, transformation, and eventual dismantling of racial *apartheid*, a policy that codified a belief in black inferiority as explicit policy and required segregation of the races in virtually every aspect of political and social life (Goldberg 1994, McClintock and Nixon 1991). Comparative scholars analyze the difference and similarities of South African apartheid and American Jim Crow laws (Ford 1999, Goldberg 1994).

Much of the historical study of European racial policy takes place as part of the study of European colonialism (Appiah 1993, Bhabha 1994, Gilroy 1991). Explicit colonial racial policy and the treatment of non-whites in Europe are understood to implicate the ideology and political economy of European colonial rule.

2.3.2 De facto *racial subordination*. Formal racial exclusion, segregation, and hierarchical classification are rare in the post-industrial mass democracies of today. They have been rejected by mass opinion, eliminated by legislation, and struck down by courts applying constitutional law. Nevertheless, racial hierarchy persists as a significant feature of every society with a history of racial differentiation. Accordingly, the study of race and the law focuses a significant amount of attention on informal or implicit practices that effectively perpetuate racial hierarchy.

The focus on *de facto* racial subordination almost certainly began as a response to racial practices in the post civil war USA. Federal constitutional guarantees of emancipation from slavery, equal protection and the right to vote prohibited certain explicit racial barriers (see *Civil Rights*; *Fundamental Rights and Constitutional Guarantees*). But many states, especially those of the former confederacy, found mechanisms to effectively re-establish these racial barriers without explicit reference to race. For instance, although blacks had a formal right to vote, many states effectively eliminated the black franchise through poll taxes and literacy requirements that were informally waived or ignored for whites but strictly enforced against blacks. And despite the formal emancipation of black slaves, many blacks were effectively forced into perpetual indentured servitude through the notorious sharecropping system of Southern agriculture.

Informal and *de facto* forms of racial subordination emerged again as the response to more comprehensive civil rights reforms of the Civil Rights Act of 1964. Significantly, the Civil Rights Act prohibited racial discrimination in employment and accommodation in private institutions such as hotels, restaurants, and theaters, and in the sale and rental of housing (see *Civil Rights*; *Fundamental Rights and Constitutional Guarantees*). Many of the private institutions covered by the Act sought effectively to reproduce racial exclusion or discrimination through less explicit means. Lawyers, activists, and scholars, armed with a knowledge of post civil war history, were able to identify and critique these practices. The attack on *de facto* racial subordination and discrimination became a staple of civil rights litigation as scholars and lawyers worked to expose the variety of practices that effectively per-

petuated racial subordination. Antiracist scholarship and legal practice has attacked a host of formally colorblind practices as effectively racially discriminatory, including electoral gerrymandering (Guinier 1995), standardized tests (Matsuda and Lawrence 1997), neighborhood schools in a context of racially segregated neighborhoods (Ford 1995), restrictive land use regulations (Ford 1995, C Harris 1995), informal racial steering in real estate sales (Ford 1995), and informal racial discrimination in hiring and promotions (Crenshaw 1995, Williams 1991).

3. *Antiracist Legal Reform*

Scholarship that examines, analyzes, and advocates the development of antiracist legal reform encompasses a vast range of proposals and is based on a number of methodologies and ideological commitments. Many approaches to antiracist reform build on civil rights reforms of the 1960s in the USA, advocating the expansion of such reforms and/or the adaptation of such reforms in other national contexts (see *Civil Rights*).

A significant split has emerged between mainstream and critical approaches. Critical approaches question the efficacy and the underlying ideologies of mainstream civil rights and liberal antiracism (see *Critical Race Theory*). Beginning from an analysis that sees the problems of racism as deeply imbedded in the structure of social practices and institutions—including law itself—critical approaches are skeptical of reform proposals that attempt 'surgically' to correct or ameliorate racial bias while leaving most institutions and social practices more or less unchanged (Bell 1989, Crenshaw 1995). Nevertheless, many critical approaches often build on mainstream reforms, arguing for their radical expansion (Matsuda and Lawrence 1997, Williams 1991). Others suggest that mainstream reforms do more harm than good and should be abandoned altogether in favor of more comprehensive changes (Tushnet 1982).

3.1 *Integrationism vs. Separatism*

Crudely put, antiracist legal reform is characterized by a split between two strategies and/or normative commitments, one which favors the integration of the races and one which favors the development of distinct, autonomous and separate communities (see Ford 1995, Peller 1995).

Historically, racial hierarchy has been legally enforced through spatial quarantine: American Jim Crow laws enforced the physical separation of the races in public accommodations and state institutions; racial zoning and restrictive covenants have divided the metropolitan landscape into racial neighborhoods; South African apartheid enforced racial segregation on a comprehensive regional scale and imposed racially specific curfews in white-only jurisdictions. Spatial segregation not only had potent symbolic consequences but also facilitated the inequitable distribution of resources: white-only institutions were consistently better situated, better funded, and socially and politically privileged compared to those assigned to racial minorities.

Given this history, a significant approach to remediation has been to encourage the integration of the races in public and private institutions and residential environments. Many American civil rights reformers believed that antiracist legal reform should embody a positive preference for integration *qua* integration. Others believed that integration was desirable only because it would undo or stymie the inequitable distribution of resources in a society. These two rationales for integration produced a powerful consensus that has dominated American civil rights discourse since the 1960s and which has been influential in many other national contexts.

At the same time, however, a competing view has held that the development of distinctive and empowered minority communities is essential to racial justice and equality (Peller 1995). This view focuses on the need for political solidarity, social support, and cultural autonomy. Racial 'separatism' has grown in popularity as many racial minorities have become disillusioned with the limitations of mainstream civil rights and the difficulties and compromises entailed in integration.

3.2 *Anti-discrimination*

A dominant conceptual and practical approach to antiracist legal reform has been that of anti-discrimination. The anti-discrimination concept presupposes that racial hierarchy is the result of irrational prejudice or bigotry based on race that results in racial discrimination. Therefore, legal reform prohibits intentional racial discrimination and practices and some policies that disparately impact members of racial minority groups. The anti-discrimination approach to antiracism focuses on specific wrongful actions by identifiable decision makers: if no wrongful act can be identified, this approach recognizes no legally cognizable or remediable harm (see Freeman 1995; *Discrimination*; *Equality and Inequality: Legal Aspects*).

3.3 *Anti-subordination*

An alternative approach to antiracist legal reform is that of anti-subordination. The anti-subordination approach views racism as a set of practices that have the effect of systematically reproducing racial hierarchy, even if no identifiable decision maker can be said to be racially biased. Anti-subordination approaches therefore find it uncontroversial that practices that adversely effect racial minorities should

come under legal scrutiny, even in the absence of evidence of intent or bias. More importantly anti-subordination approaches look for patterns and systems in which several distinct practices combine to produce regular, predictable, and systematic racial subordination even when each individual practice, viewed in isolation, might seem unproblematic (Crenshaw 1995).

3.4 Identity Politics and Multiculturalism

One significant strand of antiracist politics seeks to advance a project of racial recognition or multiculturalism, emphasizing the distinctiveness of racial minority communities in terms of culture. This strand seeks legal reform prohibiting discrimination against the cultural practices of racial minority groups as well discrimination on the basis of race *qua* status (Gotanda 1995, Peller 1995, Kymlicka 2000, Taylor 1994).

This approach can be understood as a logical extension of the anti-subordination approach to antiracist legal reform: discrimination against cultural practices effectively harms members of racial minority groups and can be seen as a component of systematic racial subordination.

One necessary consequence of racial multiculturalism is that it requires a definition or widely shared understanding of the content of racial identity or a racial culture. Such understandings or definitions are often controversial, both in society at large and within the given racial group itself. Moreover, the idea that racial identity or culture exists in a static or stable form is itself controversial. Some scholars would reject the very idea of a static or widely shared understanding of racial identity or culture as mistakenly essentialist (see Sect. 2.1.1).

3.5 Anti-essentialist and Cosmopolitan Conceptions of Racial Identity

An emerging strand of antiracist scholarship first articulated within postcolonial African, Caribbean, and Indian communities approaches racial identity as fluid and supple, yet also the source of rich political, social, and cultural affiliations. These approaches define racial identity not in terms of a biological or cultural essence, but instead in terms of identification (Appiah 1996, Bhabha 1994, Gilroy 1991). According to one strand of this scholarship, members of racial minority groups share common 'passions' based on shared or analogous experiences of racial subordination (Gilroy 1991). This approach dispenses altogether with any remnant of biological or inheritance notions of racial identity: members of what might traditionally have been several different 'races' can and do share a cosmopolitan racial identity which can transcend national boundaries, cultural traditions, skin color, and phenotype. This scholarship adopts an explicitly anti-subordination approach to antiracism, but rejects the multiculturalist focus on cultural traditions within distinctive racial or ethnic communities as traditionally defined. Instead it embraces a cosmopolitan or diasporic culture that joins racial minorities of various ethnicities and cultural backgrounds.

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on its Effectiveness; Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Civil Liberties and Human Rights; Conflict Sociology; Critical Race Theory; Discrimination; Gender, Class, Race, and Ethnicity, Social Construction of; Race: Genetic Aspects; Race: History of the Concept; Race Relations in the United States, Politics of; Racism, History of; Racism, Sociology of; Segregation Indices

Bibliography

Appiah K A 1993 *In My Father's House, Africa in the Philosophy of Culture*. Oxford University Press, New York

Appiah K A 1996 Race, culture, identity. In: Appiah K A, Gutman A (eds.) *Color Conscious*. Princeton University Press, Princeton, NJ

Bell D 1987 *And We Are Not Saved: The Illusive Quest for Racial Justice*. Basic Books, New York

Bell D 1992 *Race, Racism and American Law*. Little Brown, Boston

Bhabha H K 1994 *The Location of Culture*. Routledge, New York

Carter S L 1991 *Reflections of an Affirmative Action Baby*. Basic Books, New York

Crenshaw K 1995 Race, reform and retrenchment: Transformation and legitimation in anti-discrimination law. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) 1995 *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York, pp. 103–26

Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) 1995 *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York

Davis F J 1991 *Who is Black? One Nation's Definition*. Penn St. Press, PA

Delgado R, Stephanic J (eds.) 1999 *Critical Race Theory: The Cutting Edge*, 2nd edn. Temple University Press, PA

Denton N A, Massey D S 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA

Ford R T 1995 The boundaries of race. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) 1995 *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York, pp. 449–64

Ford R T 1999 Law's territory (a history of jurisdiction). *Michigan Law Review* **97**: 843–930

Freeman A D 1995 Legitimizing racial discrimination through anti-discrimination law: A critical review of supreme court doctrine. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York, pp. 29–45

Gates H L Jr. (ed.) 1986 *Race, Writing and Difference*. University of Chicago Press, Chicago

Gilroy P 1991 *There Ain't No Black in the Union Jack: The Cultural Politics of Race and Nation*. University of Chicago Press, Chicago
Goldberg D 1994 *Racist Culture*. Blackwell, London
Gotanda N 1995 Our constitution is colorblind. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) *Critical Race Theory: The Key Writings that Formed the Movement*. New Press, New York, pp. 257–75
Guinier L 1994 *The Tyranny of the Majority: Fundamental Fairness in Representative Democracy*. Free Press, New York
Haney-Lopez I F 1994 *White by Law*. New York University Press, New York
Harris A P 1990 Race and essentialism in feminist legal theory. *Stanford Law Review* **42**: 581–616
Harris C 1995 Whiteness as property. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) *Critical Race Theory: The Key Writings that Formed the Movement*. The New Press, New York, pp. 276–91
Kennedy R 1998 *Race, Crime and the Law*. Vintage Books, New York
Kymlicka W 2001 *Politics in the Vernacular: Nationalism, Multiculturalism, and Citizenship*. Oxford University Press, Oxford, UK
Matsuda M J, Lawrence C R III 1997 *We Won't Go Back: Making the Case for Affirmative Action*. Houghton Mifflin, Boston
McClintock A, Nixon R 1991 No names apart: The separation of word and history in Derrida's 'Le dernier mot du racisme.' In: Gates H (ed.) 1986. *Race, Writing and Difference*. University of Chicago Press, Chicago, pp. 339–53
Omi M, Wynant H 1994 *Racial Formation in the United States: From the 1960s to the 1990s*, 2nd edn. Routledge, New York
Peller G 1995 Race consciousness. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) *Critical Race Theory: The Key Writings that Formed the Movement*. The New Press, New York, pp. 127–58
Taylor C 1994 The politics of recognition. In: Gutman A, Taylor C (eds.) *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton
Thernstrom S, Thernstrom A 1999 *America in Black and White: One Nation, Indivisible*. Touchstone Books, New York
Thomas K 1995 Rouge et Noir reread. In: Crenshaw K, Gotanda N, Peller G, Thomas K (eds.) 1995 *Critical Race Theory: The Key Writings that Formed the Movement*. The New Press, New York, pp. 465–94
Tushnet M V 1981 *The American Law of Slavery, 1810–1860: Considerations of Humanity and Interest*. Princeton University Press, Princeton, NJ
Tushnet M 1984 An essay on rights. *Texas Law Review* **62**: 1363–403
Williams P J 1991 *The Alchemy of Race and Rights*. Harvard University Press, Cambridge, MA
Woodward C V 1989 *The Strange Career of Jim Crow*, 3rd edn. Oxford University Press, New York

R. T. Ford

Race and Urban Planning

In general, the term 'race and urban planning' refers to those aspects of urban planning which affect or are affected by the racial variety of the human condition, as this is expressed in multiracial societies. 'Urban planning,' also known as urban and regional planning, or as town and country planning, depending on the country of practice, helps to develop and enhance various aspects of life in human settlements. These include the location and arrangement of housing, community services, and commercial facilities, or the broader management of land use. The word 'race' typically refers to one of a few broad categories of humanity, such as Caucasian, Asian, or African. Although urban planning should ideally be race-neutral, in fact the racial characteristics of various populations have often affected their receipt of services, their spatial location or lack of spatial mobility, and their interaction with or participation in planning decisions. The lessons learned about how multiracial contexts affect urban planning can often be applied to contexts diversified by some other dimension, such as ethnicity, tribe, religion, or nationality.

1. *Background of the Issue*

The linkage between urban planning and race arose in several modern multiracial societies because of the obvious ways in which decision makers were using urban planning as a tool for racial oppression. The context in these early cases was one of at least one dominant and one subordinate racial group. For example, in the USA during the early part of the twentieth century, some of the first tools which professional urban planners used to regulate land use helped enforce the segregation of African American households from white households (Silver 1997). In societies characterized by colonialism, professional designers of new cities controlled by European powers made sure that the spatial arrangement of urban settlements sent at least two clear signals. The first, symbolized by grand approaches to isolated government buildings, was that the European colonial government was to be respected and feared. The second was that those sections of the city which housed Europeans were favored with higher standards of municipal services and housing than other sections set aside for 'natives' (Hall 1996).

As multiracial urban societies evolved during the twentieth century, many governments abandoned the purposeful isolation of supposedly subordinate racial groups through such means. Other manifestations of the problem arose, however. In the USA during the 1930s through the 1960s, some localities used programs allied with urban planning and redevelopment to isolate racial minorities, either through constructing racially segregated public housing or through selectively moving people out of and bulldozing minority neighborhoods by using federal 'urban renewal' funds. These activities brought mixed reactions: in only a few cases during the World War II era did civil rights organizations protest the blatantly segregationist public housing decisions, for example, which were some of

the first manifestations of racism in modern public policy. Outrage over the abuses of the federal 'urban renewal' program, on the other hand, brought spirited protests. The public outcry over 'urban renewal,' which became nicknamed 'Negro removal,' led to significant reforms in federal policy and in local urban planning activities after the 1960s. Yet racial segregation was so entrenched in US society that it proved difficult to eliminate its many manifestations simply by changing federal and local policy making. Similarly, as colonialism withered and died, former colonies arose to govern and plan their own cities and to eliminate the most pervasive abuses of colonial urban planning. Yet in urban places they found themselves dealing with significant social divisions nonetheless, because of caste, class, ethnicity, and any number of other dimensions. Once set in place, city structure proved hard to change.

Although in most cases the purposeful use of urban planning tools to enforce racial subordination, isolation, and segregation has waned as time has passed, the problem is by no means irrelevant to the present. As we will see in the sections below, subtle exclusion still characterizes seemingly innocuous urban planning decisions in modern countries. One society that has continued, until very recently, to use overt techniques of racial oppression through urban planning is South Africa. Other examples of cities with continuing conflict over related issues include Jerusalem, Israel, and Belfast, Northern Ireland. Yet the nature of the issues has changed markedly as broader societal conditions have changed. In Johannesburg, South Africa, for example, the end of apartheid has brought the end of legally enforced spatial segregation by race, but not the end of racial segregation and uneven development. In such a case, it may be harder to address present conditions and problems with meaningful solutions than it was to operate under the pernicious but forthright apartheid.

2. *Areas of Scholarly Inquiry*

Contemporary scholars of urban planning have focused on several general areas of connection between race and urban planning. Some scholarly works are connected with other more traditional social science endeavors in that they emphasize the empirical analysis of such phenomena as racial segregation. Other scholars, sometimes using qualitative or holistic methodologies such as history or case studies, document the existence of problems of racial estrangement or oppression in the context of private or public decisions concerning planning and development. Particularly valuable are those discussions which deal with the distribution of power and decision making, when urban planning is being practiced in a context of racial or other inequality.

In addition, several authors focus on specific dimensions of the problems. Certain findings which relate to land use, zoning, housing, and neighborhoods, as well as those related to services such as transportation, or to general urban conditions such as the environment, are now briefly summarized.

2.1 *General Connections Between Race and Urban Planning*

Several general works about urban history or urban planning history have addressed the use of planning for racial purposes through case studies or 'stories' about injustice or oppression. Such works have found that poor race relations have strongly influenced local development and planning policy as well as federal policies. A typical story in this tradition, for example, might trace the deleterious effects of racial conflict upon local efforts to improve one city, perhaps covering all or part of a period extending from the early twentieth century until the 1990s. Topics covered for a US scholar might include decisions about public housing and urban renewal, effects of civil rebellion or racial disorder upon inner-city redevelopment, housing discrimination, or failed attempts to stabilize neighborhoods in the context of racial change. In more recent years, issues covered might include how racial conflict bars attempts to bridge the gap between cities and suburbs in metropolitan areas, or weakens coalitions which might otherwise unify in order to revitalize neighborhoods or central business districts (Thomas 1997, Catlin 1993).

Other books in this general category document the continued existence of racial segregation, or suggest policy actions designed to lessen the effects of such segregation. More connected with the disciplinary areas of the social sciences than with urban planning, these works have nonetheless continued to have a major influence on thinking about the problems of racial separation and isolation. In some cases, recent works documenting conditions of racial and income segregation and of metropolitan fragmentation have helped heighten the dialogue about how to undertake fundamental society reforms, many of which could profoundly affect the work of urban planners. Examples include possibilities of merging racially segregated municipalities or heightening their cooperation, sharing tax bases, dispersing low-income housing, or in other ways halting or minimizing the effects of society's tendency to disperse and separate in racially exclusionary ways (Rusk 1995, Orfield 1998).

Within this same category of 'general' scholarship concerning urban planning and race also belong those writings connecting issues of race and empowerment, particularly as this concerns participation in making decisions that affect the communities in which racial minorities live. In addition to the extensive literature on urban renewal which deals with this issue, a few additional works have addressed directly the problem of how racial minority communities have tried to wrestle control of their destinies from the hands of city

officials. In some cases in the USA, it turns out, such minority communities have had outstanding success in doing so; one case study documents this evolution in Birmingham, Alabama, for example (Connerly and Wilson 1997). In other cases, it was federal programs from the 1960s and 1970s such as the Office of Economic Opportunity or Model Cities that led to increases in power and influence, since these programs required extensive citizen participation by low-income populations in the areas they served.

Sometimes planners in strongly segregated societies have made valiant attempts to overcome historic misuse of planning activities for purposes of subjugation, but have discovered that the issues are not so easily resolved. As apartheid has ended in South Africa, for example, attempts have been made to develop a more egalitarian approach to municipal decision making concerning planning. As a few observers have begun to document, however, the lingering effects of unequal power and unequal resources continue. The 'deracialization' of urban space in Cape Town led to laudable efforts to promote racial mixture in previously segregated neighborhoods. In addition, local planners sponsored public forums which tried to develop an inclusionary model for guiding decisions which would affect the future layout of residential and commercial areas. Yet the 'forum' approach to egalitarian decision making wrongfully assumed that all stakeholders would be included, would participate, and would have equal voice. In a society beset by profound disadvantages for people of color, such assumptions have sometimes proved premature (Watson 1998).

Where the history of racial segregation and oppression is less debilitating than in South Africa, planners in multiracial and multicultural contexts are beginning to make creative strides. Urban planners in several Canadian cities, for example, are being forced to examine the cultural and political biases inherent in a wide range of their activities. They are becoming more attentive to the need to invite participation in decision making by a wide range of racial, ethnic, and national groups, and to listen to these communities when they protest that various regulations hamper their free cultural expression. As a simple example, Chinese communities expect that their 'shopping centers' will meet different functions and have a different scale than would the mainstream population, which has set regulations according to European standards. As another example, many regulations concerning religious facilities assume that the building to be sited will be a church. An immigrant-rich context in which Middle Eastern Muslims are increasing in numbers has required creative flexibility concerning the need for progressive congregational growth. Small Muslim communities often start with meetings in houses which cannot therefore be strictly limited to residential use. These groups may then move to rented stores or other buildings, and then construct mosques with unique architectural features not allowed by regulations designed only for Christian churches (Qadeer 1997).

2.2 *Land Use and Zoning*

Relevant research in this area documents the fact that land use and zoning decision making in cities, towns, and suburban areas routinely excludes people on the basis of socioeconomic status or race. While in decades past exclusion specifically focused on race, more modern and subtle versions keep people out through informal means, or through socioeconomic exclusion. An example would be zoning ordinances which exclude low-income or multifamily housing from well-to-do suburbs. The general term for such discriminatory action is exclusionary zoning. In metropolitan areas in which low-income racial minorities are isolated and segregated, such action may have much the same effect as would zoning exclusions based on racial characteristics, which are now, in most places, illegal.

A related phenomenon is expulsive zoning. This refers to the practice of using minority neighborhoods as dumping grounds for Locally Unwanted Land Use (known in planning circles by the acronym LULU, or as a NIMBY, which stands for 'Not In My Back Yard'). Examples of such uses are landfills, transfer stations, and inappropriately large concentrations of group homes for people with chemical dependencies. Localities often accomplish unfair concentrations of LULUs by granting inappropriate variances, conditional use permits, or spot zonings in residentially zoned low-income minority neighborhoods. These actions effectively 'expel' residential use and residents from an area and lower the property values of the remaining residences (Rabin 1989).

In many cities, residential minority neighborhoods (often stable living environments for most of their history) are also externally controlled by the fact that, years ago, they were zoned for commercial or industrial use. When in order to maintain stability they attempt to seek a less intense, residential classification, which is known as downzoning, such neighborhoods may be rebuffed because of their history, minority status, or lack of power within the political arena. Zoning also controls aspects of the accessibility and availability of childcare, elder care, and other services needed to maintain neighborhood livability.

In the USA, one drawback to effective action is that the federal Supreme Court, in a long series of cases, has made it clear that socioeconomic exclusion through zoning will not be substantively challenged from the courts. Yet seemingly innocuous opposition to subsidized or low- or moderate-income housing in towns or suburban communities may indeed have enormous racial implications. The failure to grant standing (the ability to be heard as a substantive party in court cases) to several low-income plaintiffs who

wished to press for approval of affordable housing in suburban communities has effectively shut off an important route to access to suburban housing for inner-city minorities. Extensive surveys of zoning ordinances show that many local communities discriminate, through seemingly innocuous means, against 'nontraditional' families, such as single female-headed families or extended families which choose to live together to reduce expenses. Such means may include strict zoning definitions of family, barring nonrelated family groupings or extended families from residential areas zoned for single-family housing, or placing limitations on home-based occupations, meaning that low-income residents cannot legally support themselves or supplement their incomes by opening up home businesses (Ritzdorf 1997).

2.3 Housing, Neighborhoods, Mortgage Lending

Yet another area of concern focuses on the effects of continued racial segregation in residential markets, and of discrimination among lending institutions, both of which can have a marked effect on the nature, stability, and viability of urban neighborhoods. To some extent it has been necessary to document the problem, since it has not always been accepted that discrimination among lenders existed, or that segregation in housing was anything but good business practice. The problem of segregation itself is largely the purview of sociologists and geographers, and three particularly influential urban scholars in the contemporary USA are Douglas Massey, Nancy Denton, and William Julius Wilson. They have clearly documented the fact that racial segregation has continued to be a problem in the last half of the twentieth century, and has had profound effects in reducing opportunity, shaping city growth, and crippling several generations of black youth (Wilson 1996, Massey and Denton 1993). Several good works have also documented continuing problems over equitable treatment for racial minorities in obtaining money for buying, rehabilitating, or insuring their homes and their other property (Squires 1997, Yinger 1995).

These issues are particularly important to planners because, with racially skewed housing markets, it becomes very difficult for planners or other local decision makers to trust private housing markets to supplement public investments. Yet no country has sufficient public investment money to reshape urban areas without significant support by the private sector. Neither can planners undertake lasting reforms via neighborhood revitalization if the population is in constant turmoil because of racial change. This is particularly a problem in the USA, which is characterized by extensive decentralization of middle-class populations to racially homogeneous suburbs. When neighborhoods in multiracial societies become identified exclusively with the subordinate class, it becomes very difficult to market them as viable places to live for a wide variety of people. Only in a few cases have US scholars been able to discover means of stabilizing such communities, such as using fair housing initiatives—that is, initiatives designed to promote residential decisions based on the principles of free access for racial minorities—and purposeful efforts to convince home buyers of one race to move to residential neighborhoods populated by another race. Cleveland, Ohio, is one US metropolitan area that has attempted to promote a proactive agenda of this nature. Observers have found, however, that the private housing market will resist policies designed to remedy racial imbalances unless these are supported by home buyers, renters, politicians, realtors, financial institutions, and insurers. Only in a few metropolitan areas have prointegrative efforts by several local organizations and governments resulted in neighborhoods characterized by long-term racial diversity, however (Keating 1994).

The other side of coin is that in some national communities, diversity of neighborhoods is a sign of healthy diversity of cultural choice, rather than of a history of exclusionary discrimination. Again, the example of Canadian cities offers some insight into the implications of such variety for the activities which urban planners undertake (Qadeer 1997). In such cases, the challenges are to remove the last vestiges of policies based on the unconscious assumption that dominant cultural values rule the world, and to begin to allow for a multitude of life-styles and cultural expressions.

A related concept concerns community development, which refers to the purposeful improvement of residential communities using, in part, the participatory involvement of community residents. Given the previous discussions on segregation and on empowerment, it is fitting to note that in some cases isolated minority communities have been able to use tools of urban development and of urban planning to improve their own conditions. Several observers have offered inspiring case studies of successful attempts by African American communities to overcome the handicaps of race and class in order to create viable living spaces (Leavitt and Saegert 1989, McDougall 1993).

2.4 Transportation and Environment

Several other aspects of urban life can also relate to issues of race. Two are briefly mentioned here: transportation and the environment. Examples related to transportation, would be the negative effects of highway construction policies upon the viability of minority neighborhoods, or the problem of lack of accessibility to means of transportation by racial minorities (Bullard and Johnson 1997). This problem is also evident in several historical case studies of African American community life, since federal highway legislation functioned much as redevelopment

legislation did in the physical clearance of African American neighborhoods.

A major race-related issue of concern to planners is the disproportionately negative impact that environmental pollutants have upon racial minorities. Reference was made earlier to expulsionary zoning (see Sect. 2.2), indicating some of the problems that can arise. Examples of such problem areas include the siting of regular and hazardous waste landfills in minority neighborhoods, the prevalence of lead poisoning in minority housing, and the consumption by indigenous hunters and fishers of fish and wildlife contaminated by harmful substances. Social scientists have helped to document the extent of the problem, and have joined with civil rights and other groups to examine the racial implications of such cases (Bullard 1994). One of the claims among activists in the USA is that environmental pollution disproportionately affects people of color, and negatively impacts their health and well-being. Another key concern is that the mainstream environmental movement has largely ignored these racial issues. Recent environmental justice initiatives, however, have forced the environmental movement to become more inclusive. Environmental racism is an important phenomenon for urban planners because of their role in helping to determine land uses, including the location of landfills and heavy industry, and the reclamation of environmentally contaminated land for urban redevelopment projects.

3. *Responses to Racial Inequality and Injustice*

Urban planners and urban planning scholars have reacted in several ways to the continued problems related to the connections between racial injustice and urban planning. Scott Bollens has offered a typology of responses that is useful as a framework for discussion (Bollens 1988).

The first possible response by an urban planner is to side with the status quo in the name of objectivity, even in the face of blatant inequities. The urban planner or urban planning policy is supposedly 'neutral,' but in fact reinforces the dominant race's position. Such a 'neutral' planner may refuse to acknowledge or address issues of segregation, estrangement, or polarization. This approach is sometimes associated with so-called 'technical' planners, who prefer to focus on the technical, non-social aspects of their work without dealing with social problems. The shield, under such circumstances, is that urban planning is a professional exercise not involved with politics or with issues of social justice. This particular attitude is not uncommon among professionals, and may indeed be relatively harmless. It is not so harmless when exemplified by planners in racially conflicted situations, particularly when those planners are members of the dominant race or ethnic group, or support that dominant group to the exclusion of their own. The claim that the planner is 'color-blind,' in such cases, may be only a poor excuse for collaboration with systematic injustice.

A second set of responses might be called 'partisan.' Under such a situation, racial inequality and injustice are obvious, but the planner rejects the concerns of the powerless group and sides with the powerful one. The city's government leaders view neighborhoods and policies through a lens colored by race, and the interests of the dominant group—to exclude, isolate, and perhaps even oppress the subordinate group—become the norm for professional urban planners. As the political leaders set the policies, the urban planners carry them out, even in the face of blatant favoritism or discrimination. Bollens offers several examples from South Africa, but also notes the same syndrome in Jerusalem, where the systematic and purposeful isolation of Arab neighborhoods has gained the force of righteousness, even though it is, in many cases, patently unjust. As Bollens notes, somehow many local urban planners have been able to function without any noticeable discomfort about the obvious partisanship of their actions.

A third category attempts to reduce inequalities between groups, sometimes through remedial action for the oppressed group. Many US urban planners consider the equity planning movement to be one of their most important responses to the racial crisis in America's cities. According to its advocates, equity planning offers a reasoned and practical response to the crisis of local policy caused by racial segregation and poverty in America's cities. That is, urban planners (and by implication other urban professionals) are not bound by the confines of their municipal jobs to an endless cycle of repression or inattention toward the disadvantaged. Instead, urban planners can and do arise above the limitations of government bureaucracy, and pursue redevelopment and other local policies to channel direct benefits to people hindered by race or socioeconomic status. Several key authors explicitly address the important role of race in helping to create the unjust conditions to which equity planners must respond (Krumholz and Forester 1990).

Finally, Bollens refers to 'a resolver strategy,' which he indicates extends beyond equity planning. A resolver approach attempts to address the root causes of power imbalances, competition among racial and ethnic groups, and disempowerment. It then aims to confront the status quo and profoundly transform the system. Although this approach is in some ways simply a more impassioned form of equity planning, it also has broader implications in countries such as South Africa. There, for example, 'resolvers' have come forward who have begun to carry out development planning that differs markedly from traditional urban planning activities, which favored white areas and white concepts even after apartheid. Development planning, in contrast, more fully integrates spatial

planning with social and economic planning than did traditional urban planning; it restructures policy directions by cutting across traditional departments and governments; and, perhaps most importantly, it includes 'a participatory process aimed at empowering the poor and marginalized.' It also happens to involve more members of the victim class as professional planners than was previously the case.

That the urban planning profession maintains an ongoing dialogue about this and other approaches to resolving racial injustice is a healthy sign. Although many societies remain severely fragmented and alienated by race, the situation is open to proactive efforts to improve existing conditions. As larger societies become more comfortable with multiracialism and with multiculturalism, urban planning as a field can and will adjust.

See also: Community Economic Development; Income Distribution: Demographic Aspects; Neighborhood: General; Neighborhood Revitalization and Community Development; Poverty: Measurement and Analysis; Poverty, Sociology of; Racism, Sociology of; Social Inequality in History (Stratification and Classes); Social Stratification; Urban Geography; Urban Growth Models; Urban Poverty in Neighborhoods; Urban Sociology

Bibliography

Bollens S 1988 Urban planning amidst ethnic conflict: Jerusalem and Johannesburg. *Urban Studies* **35**: 729–50

Bullard R (ed.) 1994 *Unequal Protection: Environmental Justice and Communities of Color*. Sierra Club, San Francisco, CA

Bullard R, Johnson G (eds.) 1997 *Just Transportation: Dismantling Race and Class Barriers to Mobility*. New Society Press, Branford, CT

Catlin R 1993 *Racial Politics and Urban Planning: Gary, Indiana 1980–1989*. University Press of Kentucky, Lexington, KY

Connerly C, Wilson B 1997 The roots and origins of African American planning in Birmingham, Alabama. In: Thomas J M, Ritzdorf M (eds.) *Urban Planning and the African American Community: In the Shadows*. Sage, Thousand Oaks, CA

Hall P 1996 *Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century*, 2nd edn. Blackwell, Oxford, UK

Keating W D 1994 *The Suburban Racial Dilemma: Housing and Neighborhoods*. Temple University Press, Philadelphia, PA

Krumholz N, Forester J 1990 *Making Equity Planning Work: Leadership in the Public Sector*. Temple University Press, Philadelphia, PA

Leavitt J, Saegert S 1989 *From Abandonment to Hope: Community Households in Harlem*. Columbia University Press, New York

Massey D, Denton N 1993 *American Apartheid*. Harvard University Press, Cambridge, MA

McDougall H 1993 *Black Baltimore: A New Theory of Community*. Temple University Press, Philadelphia, PA

Orfield M 1998 *Metropolitics: A Regional Agenda for Community and Stability*. Brookings Institution Press, Washington, DC

Qadeer M A 1997 Pluralistic planning for multicultural cities. *Journal of the American Planning Association* **63**: 481–94

Rabin Y 1989 Expulsive zoning: the inequitable legacy of Euclid. In: Harr C, Kayden J (eds.) *Zoning and the American Dream: Promises Still to Keep*. American Planning Association Press, Washington, DC, pp. 101–20

Ritzdorf M 1977 Locked out of paradise: Contemporary exclusionary zoning, the Supreme Court, and African Americans, 1970 to the present. In: Thomas J M, Ritzdorf M (eds.) *Urban Planning and the African American Community: In the Shadows*. Sage, Thousand Oaks, CA, pp. 43–57

Rusk D 1995 *Cities Without Suburbs*. Woodrow Wilson Press, Washington, DC

Silver C 1997 The racial origins of zoning in American cities. In: Thomas J M, Ritzdorf M (eds.) *Urban Planning and the African American Community: In the Shadows*. Sage, Thousand Oaks, CA, pp. 23–42

Squires G D (ed.) 1997 *Insurance Redlining: Disinvestment, Reinvestment, and the Evolving Role of Financial Institutions*. Urban Institute Press, Washington, DC

Thomas J M 1997 *Redevelopment and Race: Planning a Finer City in Postwar Detroit*. Johns Hopkins University Press, Baltimore, MD

Thomas J M, Ritzdorf M (eds.) 1997 *Urban Planning and the African American Community: In the Shadows*. Sage, Thousand Oaks, CA

Watson V 1998 Planning under political transition—lessons from Cape Town's metropolitan planning forum. *International Planning Studies* **3**: 335–50

Wilson W J 1996 *When Work Disappears*. Knopf, New York

Yinger J 1995 *Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination*. Russell Sage Foundation, New York

J. M. Thomas

Race: Genetic Aspects

Natural species are genetically polymorphic, in two related, but distinct ways. Different individuals may carry different alleles of the same gene (polymorphism within populations), and different populations may be characterized by different alleles, or by different frequencies of the same alleles (polymorphism among populations). Differences among populations depend on various evolutionary factors, but populations that exchange migrants tend to resemble each other, whereas mutually isolated groups and communities tend to diverge genetically.

The relative importance of polymorphism within and among populations varies across species. In many plant and animal species, the largest share of genetic diversity occurs among populations or groups thereof. When that is the case, most individuals of unknown origin can be assigned to one group on the basis of their genetic characteristics, with a high degree of statistical confidence. If distinct groups occupy different geographical areas, it is customary to define them as subspecies or races. The question is how

faithfully this concept also describes the levels and patterns of genetic diversity observed among humans.

One of the latest definitions of race is as follows: 'A large population of individuals who have a significant fraction of their genes in common, and can be distinguished from another race by their common gene pool' (Vogel and Motulsky 1986). But the concept of race is older than genetics. Long before Mendel the term referred to people sharing the same 'blood,' that is, to groups of genealogically-related individuals, especially those of noble families (Furetiére 1690, cited in Cohen 1991). In the eighteenth and nineteenth centuries, the term race came to designate constant human types, and came to be used not only to describe, but also to explain, human diversity. People who look different belong to different races, and they are different because they belong to different races. That view is still widespread; it seems based on an empirical observation, and on an assumption. The observation is that biologically related people look alike, and the assumption is that our species is naturally subdivided in clusters of individuals who look alike because they are genealogically related, or races.

Is that assumption based on solid scientific grounds? Or are human races arbitrary groups of populations, which do not share a common evolutionary history more than other populations classified into different races? Only in the second half of the twentieth century has it become possible to address this question using genetic data. For the purposes of this article, the hypothesis being tested is whether the observed patterns of genetic diversity suggest that humankind is subdivided in nearly discrete, and internally homogeneous, groups. The results of those tests, and their evolutionary implications, will be summarized in this article.

1. Listing Human Races

Prior to measuring genetic differences within and among human groups, one has to define these groups. The many lists of races which have appeared in the course of history may be regarded as successive attempts to identify the set of physical traits which leads to the best discrimination of groups, that is, traits for which the among-group component of variation is maximum. Ideally, once such 'best' physical traits are identified, most individuals and populations should be classified unambiguously into the appropriate group, except perhaps for a minority of them, traditionally regarded as derived from interracial crosses.

Table 1 summarizes some lists of human races proposed in the last three centuries. Although this selection is necessarily incomplete (the interested reader is referred to Cohen 1991 and Molnar 1998) a problem, and a trend, are evident. The problem is, any group defined on the basis of morphological traits harbors some degree of internal variation, and different criteria, say skin color and skull shape, lead to different classifications. This is clearly shown by the fact that, although the existence of human races has been challenged only recently (Livingstone 1964), little agreemement has evidently been reached on how many races exist, and which. The trend is a consequence of the problem. As more and more groups came in contact with the European scientists who were developing these classifications, the number of proposed races increased steadily.

There is an exception to that trend. In 1963, Coon proposed that there are only five subspecies among humans. In that context, subspecies was synonymous to race, and it was defined as a regional population 'which meets two tests: (1) it occupies a distinct geographical territory; (2) it differs from other subspecies of the same species in measurable characteristics to a considerable degree' (1962, p. 19). Coon's definition overlaps with the one given by Vogel and Motulsky (1986), and both can be tested using genetic data.

2. Apportioning Human Diversity

Because there is no agreement on the best system of human racial classification, the groups or putative races being compared in the studies here reviewed were seldom the same. In general, when the groups con-

Table 1
Some lists of proposed human races

Author	No. of races	Races proposed
Linnaeus (1735)	6	Europaeus, Asiaticus, Afer, Americanus, Ferus, Monstruosus
Buffon (1749)	6	Laplander, Tartar, South Asiatic, European, Ethiopian, American
Blumenbach (1795)	5	Caucasian, Mongolian, Ethiopian, American, Malay
Cuvier (1828)	3	Caucasoid, Negroid, Mongoloid
Deniker (1900)	29	
Weinert (1935)	17	
Von Eickstedt (1937)	38	
Biasutti (1959)	53	
Coon (1962)	5	Congoid, Capoid, Caucasoid, Mongoloid, Australoid

Source: Coon 1962, Cohen 1991, and Molnar 1998

Table 2
Percentages of the overall genetic variance at three levels of population subdivision, as estimated from protein (P), RFLP (R), microsatellite (M), or single-nucleotide (S) polymorphisms; G are multilocus genotypes, that is, sets of loci jointly analyzed which, (for mitochondrial DNA and the Y chromosome) are also jointly inherited, as a single locus. Genome location refers to the fact that the loci of interest are autosomal (A), mitochondrial (M), or Y-linked (Y)

Authors	N Loci	Type	Genome location	Within populations	Between populations within groups	Between groups
Lewontin (1972)	17	P	A	85.4	8.3	6.3
Latter (1980)	18	P	A	83.8–87.0	5.5–6.6	7.5–10.4
Ryman et al. (1983)	25	P	A	86.0	2.8	11.2
Barbujani et al. (1997)	79	R	A	84.5	3.9	11.7
Barbujani et al. (1997)	30	M	A	84.5	5.5	10.0
Barbujani et al. (1997)	16	G	A	83.6	8.4	8.0
Barbujani et al. (1997)	109	M + R	A	84.4	4.7	10.8
Seielstad et al. (1998)	n.a.	S	M	81.4	6.1	12.5
Seielstad et al. (1998)	n.a.	S	Y	35.5	11.8	52.7
Santos et al. (1999)	30	S	Y	59.0	25.0	16.0
Poloni et al. (1997)	9	R	Y	70.1	6.9	23.0
Barbujani and Excoffier (1999)	100	R	M	65.5	7.1	27.5
Barbujani and Excoffier (1999)	5	G	M	77.4	7.3	15.3

Note: Latter 1980 gave three sets of values, based on three different statistical approaches. n.a.: not available

sidered are many, the variance among them is inflated. At the same time, on the other hand, as the number of groups increases, the boundaries between them tend to be ill-defined, which may lead to overestimating variation within groups.

The statistical methods used to apportion human diversity vary too, when one considers different studies. Nevertheless, the conceptual framework is the same, and is as follows. Several groups are defined on the basis of independent evidence. Each individual is compared to all other individuals studied, and a number representing the amount of genetic differentiation between them is calculated. Three components of genetic variance are then estimated from those numbers, namely: (a) between individuals of the same population; (b) among populations of the same group; (c) among major groups (see Lewontin 1972, Latter 1980, Barbujani et al. 1997). In this way, the overall genetic diversity is broken down in three hierarchical components.

2.1 Genetic Diversity at the Protein Level

In 1972, Lewontin first quantified the amount of genetic variation within and among major human groups. He chose a seven-race system, comprising Caucasians, Black Africans, Mongoloids, Amerinds, Oceanians (the five groups of Blumenbach's classification; see Table 1), plus South Asian and Australian aborigines. Analysis of 17 independent genetic loci, for which information on all available populations was collected, gave somewhat different results, but the average proportion of the overall variance that fell within populations was 85.4 percent (with a maximum of 99.7 percent for the *Xm* locus and a minimum of 63.6 percent for the *Duffy* locus). Additional differences between populations of the same race represented, on average, 8.3 percent of the total, and differences between races accounted for the remaining 6.3 percent of the total. At the protein level, therefore, variation among seven races seems to represent less than one tenth of the overall genetic diversity of our species.

Lewontin's conclusion that the human racial classification has virtually no genetic or taxonomic significance was variously criticised, not always on scientific grounds. One serious objection was that only a fraction of the total DNA is actually translated into proteins. As a consequence, protein polymorphism may not fully reflect the underlying DNA diversity. Although successive studies of proteins confirmed that more than four fifths of the overall genetic diversity of our species is accounted for by differences among members of the same community (Table 2), skepticism remained widspreaad.

2.2 Genetic Diversity at the DNA Level

The first global study of DNA diversity, based on 109 genetic loci in 16 populations, appeared only 25 years later (Barbujani et al. 1997). The samples were chosen so as to form well-separated groups over the map of the world, whereas geographically intermediate populations were not considered. Therefore, if sampling affected the genetic variances estimated, it did so by increasing the weight of the between-continent components. In addition, this made the samples of each continent fall within one of the racial groups of

Blumenbach's classification, without ambiguity. The loci chosen were 30 short-tandem repeats (STR or microsatellite loci, whose alleles differ in length, that is, in the number of nucleotides they comprise), and 79 biallelic DNA polymorphisms (restriction-fragment length polymorphisms, or RFLP). For 16 RFLPs, compound genotypes referring to subsamples of individuals were available, and variance components were also estimated from them. By jointly analyzing many loci in this way, one may hope to identify significant differences better than by separately considering each gene.

The within-population variance component was significantly greater than 0 at all loci, representing between 54.4 percent and 94.6 percent of the total. Approximately one-third of the loci studied showed significant variance also between populations or between continents. Very similar results were obtained in the analysis of single RFLP polymorphisms, of the multilocus RFLP genotypes, and of the STR loci (Table 2).

Three recent papers provided additional data (Table 2) about genes inherited only from one parent. Every couple possesses four potentially transmissible copies of most chromosomes, two maternal and two paternal. But there is only one (paternal) copy of the Y chromosome; and there are several, presumably identical, copies of the mitochondrial DNA, which are only transmitted by the mother. Population genetics theory predicts that genes on these chromosomes should diverge four times as fast as the biparentally-transmitted (or autosomal) genes considered in the previous studies, all other demographic and evolutionary factors being equal. There is no simple model quantifying the effect of such a fast divergence on components of variance, but the differences among populations and continents (jointly estimated around 15 percent for autosomic genes) are expected to represent a substantially larger fraction of the genetic variance estimated from Y-chromosome and mitochondrial data.

Contrary to those expectations, mitochondrial variation appeared distributed much like autosomic variation. Such a lower-than-expected population diversity raises evolutionary problems that will not be discussed here, related with possible effects of natural selection on mitochondrial DNA, or with a high prehistoric female mobility. On the other hand, two analyses of the Y chromosome did show an increased differentiation among continents and populations, but confirmed that the largest component of human diversity is within populations (Poloni et al. 1997, Santos et al. 1999). However, in a third study of the Y chromosome, Seielstad et al. (1998) estimated even higher among-continent variances, exceeding 50 percent of the total. Note that the values estimated by Santos et al. (1999), which are closer to the other figures listed in Table 2, derive from the analysis of a different set of the same genetic markers, single-nucleotide polymorphisms, that is, variable sites in the Y-chromosome sequence. The fact that such large differences have only been found in the male-transmitted portion of the genome, and only for one set of data, is not easy to interpret.

In synthesis, it seems that the within-population component of the overall genetic diversity of our species is close to 85 percent, regardless of the genetic markers and of the populations studied. Despite the limited polymorphism of the protein markers considered, and despite the choice of different sets of populations, Lewontin's results essentially are confirmed at the DNA level, for more than 300 genes analyzed so far. Greater among-continent diversities are estimated, as predicted by theory, on the basis of uniparentally-transmitted genes, and the allelic variants of the Y chromosome are more restricted geographically than the alleles of autosomal and mitochondrial loci. Among the plausible explanations of such a high Y-chromosome diversity are, again, selection differences in the males' and females' prehistorical migration rates, and a high variability of the males' reproductive success (discussed in Hammer et al. 1997). The conclusions of the other studies seem rather coherent, and they can be summarized as follows: if we arbitrarily set to 100 the average genetic difference between two very distant members of our species, the expected difference between members of the same community (who are not relatives, of course) seems to be approximately 85, and not 10 or 40, as previously illustrated definitions of race appear to imply.

3. *Are We All Equal, Then?*

The fact that genetic differences among major human groups are relatively small does not mean that all humans are genetically identical. In fact, the among-population and among-group components of variance are significantly greater than zero, not only for the Y chromosome, but also at several autosomic loci (Barbujani et al. 1997). That fraction of diversity, albeit small in relative terms, contains all the information that population geneticists use to reconstruct human evolutionary past. From its study, we know that both allele frequencies and DNA sequences vary nonrandomly in space, often forming gradients over entire continents (reviewed by Cavalli-Sforza et al. 1994). We also know that zones of sharp genetic change, or genetic boundaries, exist, and often overlap with various kinds of migrational barriers, both physical (mountain ranges, arms of sea) and cultural (especially language boundaries) (Sokal et al. 1990). However, very seldom do genetic boundaries form complete rings around single populations, or groups thereof. Over much of the planet, continuous variation is the rule, with the genetic characteristics of the various regions changing more or less smoothly in the

geographical space. Locating on the map the place at which one racial group ends and another one begins proved, so far, an impossible task. But nearly all populations studied are somewhat different from their neighbors.

4. Comparisons with Other Species

The subspecies concept is widely used in zoology and botany, nd the suspicion may be raised that moral and political, rather than scientific, considerations may influence our willingness to use the same concept for humans. Although only a few authors have addressed this problem so far, their results suggest that humans are actually a comparatively uniform species. To cite just one example, the fruit fly *Drosophila pseudoobscura* shows a 10-fold greater nucleotide diversity than our species (Li and Sadler 1991).

Templeton (1999) compared Barbujani et al.'s (1997) estimates of genetic variances with those referring to 12 large-bodied, highly mobile mammals. Two such species (East African Buffalo and Kenyan Waterbuck) show smaller between-population diversity than humans. However, these are not cosmopolitan species like ours, but dwell in restricted geographical zones. All other species compared, including Elephant, Gazelle, and gray Wolf, show higher, or much higher, differentiation among populations and groups, with values of those variance components reaching 80 percent of the total.

On a smaller taxonomic scale, Ruvolo et al. (1994) compared mitochondrial diversity in the five genera of hominoids, that is, humans and apes. Variation appeared minimal within humans and pygmy Chimpanzees; the latter, once again, occupy a much smaller territory than humans. On the other hand, common Chimpanzee, Orangutan, and especially Gorilla, show much higher genetic differentiation. In particular, the most mitochondrially diverse humans studied appear to be less different than random pairs of lowland gorillas from West Africa.

5. Evolution of Human Genetic Diversity

Understanding human diversity also means to understand how that diversity came about. The low genetic variation among humans, as compared to other species, points to two, not incompatible, evolutionary processes, namely a high dispersal ability leading to limited isolation among populations, and/or a recent origin, too recent indeed for extensive genetic variation to have accumulated ever since.

Various evolutionary scenarios have been proposed to explain the available fossil, archaeological, and genetic evidence on human origins. In Coon's (1962) view, some 700,000 years ago, the population of our ancestor, *Homo erectus*, was subdivided geographically in five groups. Coon's five subspecies of *Homo sapiens* would differ because they are descended from those groups, with very little interbreeding, if any. This is no longer considered a valid scientific hypothesis, but the idea that the differences among contemporary humans started accumulating before the appearence of anatomically modern humans is still part of the so-called multiregional theory of human evolution (Wolpoff and Caspari 1997). This theory differs from Coon's in that the major human lineages are here regarded as only partly isolated, with extensive genetic exchange even between distant populations.

The basis of the multiregional theory of human evolution is the observation that certain traits persist in some regions of the earth from *H. erectus* times through *H. sapiens* times, and are unique to those regions. However, the statistical significance of that observation has been called into question (Lahr 1996), and it also seems that implausibly high levels of gene flow would be required, for hundreds of millennia, to maintain the unity of our species under a multiregional model (Rouhani 1989). The alternative hypothesis, the so-called out-of-Africa model (Stringer and Andrews 1988), regards the contemporary population as derived from ancestors who lived in Africa until recently. When those people expanded into Eurasia and Australia, perhaps 100,000 years ago, they replaced without admixture previous local populations of *Homo erectus* and (in Europe) *Homo sapiens neandertalensis*.

The multiregional and the out-of-Africa models have several implications. If the main human groups separated before the appearance of anatomically modern *Homo sapiens*, deep differences are to be expected among them, and quite distinct evolutionary lineages should exist, in agreement with Vogel and Motulsky's (1986) definition of race. On the contrary, if all continental populations derived recently from the same ancestors, even distant human groups should share many alleles, and the differences among groups should be relatively small. The studies reviewed in this article seem to clearly support the latter model, although Templeton (1998) argued that high levels of migration could lead to low variances between continents, even under a multiregional model. Be that as it may, most investigators, including Templeton, agree that the available genetic evidence does not suggest the existence of distinct clusters of humans.

Support to this view also comes from studies which, although not explicitly addressing the issue of racial subdivision, are in evident agreement with a recent African origin of humankind. First, genetic diversity is greater in Africa than in any other continent (Armour et al. 1996, Hammer et al. 1997, Jorde et al. 1997), an observation that a multiregional model can hardly explain. Second, genetic diversity within populations decreases as one moves away from Africa, with near populations showing subsets of African genes, and more distant populations showing smaller subsets of

those genes (Tishkoff et al. 1996). Third, Neandertal man is considered a direct ancestor of the Europeans, and only of them, under the multiregional model; but the only Neandertalian DNA sequences obtained so far fall out of the range of current human variation, and bear no special resemblance to the European sequences (Krings et al. 1997). All these findings, and other paleontological data, emphasize the crucial role of Africa in human evolution (Foley 1998), and suggest that, during their dispersal in Eurasia, Australia, and the Americas, human groups spent too little time in isolation to significantly diverge genetically and form separate races.

6. *Future Prospects*

Several problems remain open. The first one has to do with the exact amount of genetic diversity that can be attributed to the various levels of population subdivision. Figures around 10 percent of the total diversity occurring among continents have been estimated by analyzing samples that were well separated in space, so that no ambiguity existed as for the continent each of them belonged to. However, in nature, populations are not so discretely distributed. May that sampling structure have led to overestimating the among-group component of genetic variance? And may the choice of the DNA markers studied (most of them falling in noncoding regions of the DNA) have affected the variance estimates too, perhaps in the opposite direction?

A second question is whether the apparently continuous distribution of genetic variation also implies that recognizable geographical groups do not exist, as has been argued by Templeton (1998). In fact, as previously mentioned, genetic variances among groups, although quantitatively small, are significantly greater than zero at many loci; this may have a meaning.

Some DNA alleles occur only in one specific population or continent. Individuals of unknown origin who carry those alleles can be assigned with high statistical confidence to the continent they belong to. However, it is not clear yet, and this is a third question, whether also broad human groups can be defined based on the presence of those alleles (rather than on the basis of most loci of the genome, as implied by traditional definitions of race); and it is unclear whether such groups, if any, could be called races. Most studies reviewed in this article do not support the idea that *the majority* of the members of our species can be put into a specific continental or racial group, on the basis of their genotype. This is a statistically tractable problem, and investigation on it is currently in progress. Good candidate genes for discriminating among groups could be the Y-chromosome single-nucleotide polymorphisms which, as mentioned, vary substantially across continents (Seielstad et al. 1998). However, population groups defined on the basis of Y-chromosome diversity only could hardly be called races. A statement such as 'Human races exist, but only among males' would sound puzzling.

The genes determining morphological traits such as skull and eye shape, hair texture, or skin color, have not been identified yet, and some have argued that, once they will be known, a clearer identification of human groups will be possible. There are three reasons to doubt it, though. First, as shown in Table 1, centuries of studies of the phenotypes determined by those genes have not led to a consensus on the existing human races; whether and how a better knowledge of the underlying genes may clarify this picture is not obvious. Second, groups sharing traits typically used for racial classification, such as Melanesians and Africans, have been shown to be very distant genetically, both at the protein and at the DNA level (Cavalli-Sforza et al. 1994). Third, morphological traits are determined by the interaction between several genes and environmental factors. Theory shows that this polygenic and multifactorial inheritance results in a continuous distributions of traits, that is, the distributions which are less likely to show the sharp discontinuities necessary to discriminate among groups. Probably, morphological variation reflects differential selection by climate (Owens and King 1999), or the effects of sexual selection, both sometimes leading genetically distant groups to resemble each other.

Grouping individuals in rough categories based on morphology may still be useful for some practical purposes. In countries with extensive recent immigration, different groups may have different disease risks. Despite the low among-continent genetic variances discussed so far, associating a different risk to different, sometimes ill defined, groups (e.g., 'Hispanics') may simplify the clinicians' work. Similar problems also exist in forensic science, for DNA-based personal identification; and similar solutions, based on the definition of somewhat arbitrary racial or ethnic groups, have been put forward to reduce the risk of a wrong decision against a suspect (Nichols and Balding 1991). Despite these examples, genetic studies have shown that humans do not neatly fall within few well-distinct racial groups. Any population differs from its neighbors for several genetic features and, with appropriate sample sizes, these differences may achieve statistical significance. But if the existence of at least one significant genetic difference were the criterion for defining races, every village would be occupied by a different race.

In conclusion, newly discovered polymorphisms, especially those of the Y chromosome, and larger sampling efforts, may modify our ideas on the levels and patterns of human genetic diversity. But at present, clear discontinuities among major groups have not been detected, and there is no clear evidence of largely independent evolution of distinct human

lineages. In addition, nobody has yet identified sets of genes whose variation is correlated, so that a classification based on one of them will not be contradicted by the classification based on any other of them. Judging from the available genetic information, therefore, Vogel and Motulsky's (1986) definition of race does not seem to plausibly describe the structure of the human population. New defintions may be proposed, and, if so, they will have to stand the comparison with the data. At present, the burden of the proof seems on those who maintain that our species is subdivided into races that can be objectively recognized by studying their genes.

See also: Ethical Issues in the 'New' Genetics; Eugenics as the Basis of Population Policy; Gender, Class, Race, and Ethnicity, Social Construction of; Human Evolutionary Genetics; Human Genome Diversity and its Value; Population Policy: International; Y-chromosomes and Evolution

Bibliography

Armour J A L, Anttinen T, May C A, Vega E E, Sajantila A, Kidd J R, Kidd K K, Bertranpetit J, Pääbo S, Jeffreys A 1996 Minisatellite diversity supports a recent African origin for modern humans. *Nature Genetics* **13**: 154–60

Barbujani G, Excoffier L 1999 The history and geography of human genetic diversity. In: Stearns S (ed.) *Evolution in Health and Disease*. Oxford University Press, Oxford, UK, pp. 27–40

Barbujani G, Magagni A, Minch E, Cavalli-Sforza L L 1997 An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences USA* **94**: 4516–9

Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

Cohen C 1991 Les races humaines en histoire des sciences. In: Hublin J J, Tillier A M (eds.) *Aux Origines d'Homo sapiens*. Presses Universitaires de France, Paris, pp. 9–47

Coon C 1962 *The Origin of Races*. Knopf, New York

Foley R 1998 The context of human genetic evolution. *Genome Research* **8**: 339–47

Hammer M F, Spurdle A B, Karafet T, Bonner M R, Wood E T, Novelletto A, Malaspina P, Mitchell R J, Horai S, Jenkins T, Zegura S L 1997 The geographic distribution of human Y chromosome variation. *Genetics* **145**: 787–805

Krings M, Stone A, Schmitz R W, Krainitzki H, Stoneking M, Pääbo S 1997 Neandertal DNA sequences and the origin of modern humans. *Cell* **90**: 19–30

Jorde L B, Rogers A R, Bamshad M, Watkins W S, Krakowiak P, Sung S, Kere J, Harpending H 1997 Microsatellite diversity and the demographic history of modern humans. *Proceedings of the National Academy of Sciences USA* **94**: 3100–3

Lahr M M 1996 *The Evolution of Human Diversity*. Cambridge University Press, Cambridge, UK

Latter B D H 1980 Genetic differences within and between populations of the major human subgroups. *American Naturalist* **116**: 220–37

Lewontin C 1972 The apportionment of human diversity. *Evolutionary Biology* **6**: 381–98

Li W H, Sadler A 1991 Low nucleotide diversity in man. *Genetics* **129**: 513–23

Livingstone F B 1964 On the nonexistence of human races. In: Montagu A (ed.) *The Concept of Race*. Free Press, Glencoe, NY, pp. 46–60

Molnar S 1998 *Human Variation. Races, Types, and Ethnic Groups*. Prentice Hall, Upper Saddle River, NJ, p. 19

Nichols R A, Balding D J 1991 Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity* **66**: 297–302

Owens K, King M C 1999 Genomic views of human history. *Science* **286**: 451–3

Poloni E S, Semino O, Passarino G, Santachiara-Benerecetti A S, Dupanloup I, Langaney A, Excoffier L 1997 Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *American Journal of Human Genetics* **61**: 1015–35

Rouhani S 1989 Molecular genetics and the pattern of human evolution: Plausible and implausible models. In: Mellars P, Stringer C B (eds.) *The Human Revolution: Behavioural and Biological perspectives on the Origins of Modern Humans*. Princeton University Press, Princeton, NJ, pp. 47–61

Ruvolo M, Pan D, Zehr S, Goldberg T, Disotell T R, von Dornum M 1994 Gene trees and hominoid phylogeny. *Proceedings of the National Academy of Sciences USA* **91**: 8900–4

Ryman N, Chakraborty R, Nei M 1983 Differences in the relative distribution of human gene diversity between electrophoretic and red and white cell antigen loci. *Human Heredity* **33**: 93–102

Santos F R, Pandya A, Tyler-Smith C, Pena S D J, Schanfield M, Leonard W R, Osipova L, Crawford M H, Mitchell R J 1999 The Central Siberian origin for native American Y chromosomes. *American Journal of Human Genetics* **64**: 619–28

Seielstad M T, Minch E, Cavalli-Sforza L L 1998 Genetic evidence for a higher female migration rate in humans. *Nature Genetics* **20**: 278–80

Sokal R R, Oden N L, Legendre P, Fortin M J, Kim J Y, Thomson B A, Vaudor A, Harding R M, Barbujani G 1990 Genetics and language in European populations. *American Naturalist* **135**: 157–75

Stringer C B, Andrews P 1988 Genetic and fossil evidence for the origin of modern humans. *Science* **239**: 1263–8

Templeton A R 1998 Human races: A genetic and evolutionary perspective. *American Anthropologist* **100**: 632–50

Tishkoff S A, Dietzsch E, Speed W, Pakstis A J, Kidd J R, Cheung K, Bonné-Tamir B, Santachiara-Benerecetti A S, Moral P, Krings M, Pääbo S, Watson E, Risch N, Jenkins T, Kidd K K 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**: 1380–7

Vogel F, Motulsky A G 1986 *Human Genetics: Problems and Approaches*, 2nd edn. Springer-Verlag, Berlin, p. 534

Wolpoff M H, Caspari R 1997 *Race and Human Evolution*. Simon & Schuster, New York

G. Barbujani

Race: History of the Concept

All societies resort to the use of genealogy. The genealogical myth is the first form of historical thought and metaphysical enquiry. The concept of race dom-

inated modes of thought when research into the origins of man developed in scientific terms. From this point of view, the classification of all living beings by the Swedish scientist Karl Linné (1707–78) was a significant event. Man, or *homo sapiens*, was classified according to species, gender, and class, as were other living beings. From this point onwards, interpretation of the history of humanity was considered in terms of race until the 1930s.

Since the 1930s, most biologists have criticized the notion of race, arguing that, even if phenotypic variations are genetically determined, there is no general classification of human beings according to the category of race. 'The concept of race has lost all its value' (Jacob 1981).

Biologists and social scientists employ the term in different ways. Biologists inquire whether humans are divided into racial categories and whether belonging to a particular race has effects on behavior. Social scientists are less interested in whether the idea of race is well founded or not and instead focus on the social of consequences of the fact that man believes that racial groups exist. The main themes of their research are as follows: social relationships which are thought of in terms of race; the consequences of the belief that races exist on social relations and even on political organization as a whole. Although today race no longer has any meaning for biologists, the belief that there is a white race and a black race has had and still continues to have social consequences. Sociologists argue, according to Thomas' theorem, that if men refer to a situation or phenomenon as real, its consequences become real.

1. *The Notion of Race Before the Science of the Races*

The term race predates the emergence of racial theories claiming to be scientific. From the sixteenth century onwards, the British interpreted their history in these terms. They claimed that their political genius was due to their Saxon blood ties and the superiority of the Germanic race. So, according to them, William the Conqueror was seen to have imposed the Norman yoke upon free Anglo-Saxon peoples. If the parliament had asserted its rights against the Stuart Monarchy during the Glorious Revolution in 1688, then this was because the British were Germanic, in the way that Tacite had described them: free, asserting their liberty by limiting sovereign power (Banton 1977).

Thought in terms of race took up old representations and themes, in the name of science. The stigmatization of blacks already was current during the Antiquity period. According to a sixth-century statement attributed to Jesus: 'the children of Israel are among those people on the same level as Negroes-….' The 'Purity of Blood-ties Statutes,' introduced in Spain after the *Reconquitsa* and the expulsion of Jews and Muslims in 1492, established social categories based on 'blood' by distinguishing between pure-blooded Old Christians and the New Christians: Jews and Muslims who had converted to Catholicism, yet whose blood still remained 'impure.' A biological defect, unerasable even by baptism, was therefore passed on from generation to generation. This system of stigmatization did not, however, make any claims to scientific truth, which would become the case from the end of the eighteenth century onwards.

Racist thought developed more in English and German-speaking countries in particular. However, although thought in terms of race had less effect in France, the quarrel between two races, or in other words, the myth of two French peoples developed from the sixteenth century onwards. The noble classes claimed that the fact that they were of Germanic blood justified the freedom and privilege they enjoyed, the masses, being made up of descendents of the Gauls who had been defeated by the German invaders. Even Montesquieu, subscribing to this explanation, referred to the 'former Germans' as *our fathers*, and praised their admirable simplicity and traditions of liberty and independence.

2. *The Century of Racial Ideas*

Genealogical myth took on a new meaning during the modern period which saw the development of scientific ambition in fields such as physical anthropology and biology (Guillaumin 1972). Feelings and attitudes which had always been expressed towards certain stigmatized groups in the past were now reinterpreted and stamped with the scientifically legitimate trademark. Classic racist thought became fully established from the end of the eighteenth century onwards, as was the classification of living beings, which culminated in erasing the philosophical or religious distinctions between man, gifted with reason, responsibility, and thus liberty and other living beings.

This approach was revived with the development of social Darwinism. For the Darwin theorists, the decisive influence was biological and biology was defined by race. They believed that hostility between different races was natural and that the mixing of races led to the decline of the species concerned. They were obsessed by the need to preserve racial purity, fearing the dangers of intermarriage and its consequences, namely decline (Gobineau 1853). In order to define races, it became common to take measurements of the cranium and the angle of the nose, etc. It was scientifically claimed that blond *dolichocephalic* races were superior to brown-haired *bracchycephalic* peoples.

Race gradually appeared to be the factor that explained all historical phenomena. According to Robert Knox (1791–1862), 'Human character, in-

dividual and national, is traceable solely to the nature of that race to which the individual or nation belongs' (Banton 1977). In France, Gustave Le Bon (1841–1931) wrote in 1894 that 'different races are unable to feel, think and act in the same manner, and subsequently are also unable to understand one another.' The revolutions of 1848, which can be seen as political events, *par excellence*, were explained by Gobineau in terms of race. The Aryan myth was used by the British in order to legitimise the colonisation of India. Walter Scott explained the conflict between the Scottish and the English by their different 'blood.' The concept of race was also used as an argument for all nationalist demands. These interpretations conformed to the scientism of the times and satisfied the human need for one sole explanation of the diversity and complexity encountered in the social world.

The word racism was first used at the end of the 1930s to refer to the doctrine which claimed that race determined culture. However, the theory predated the emergence of the term itself. This theory can be summarized by the two following claims: first, that human races exist; they are biologically different and therefore will always be unequal, and second, that there is a necessary link between biological characteristics and social behavior. In other words, the biological determines the social. There is an analytical distinction between race and racism. It is possible to argue that different races exist, without recourse to the biological determinism and essentializing ideas which characterize racist thought. Central to racist thought is the notion that races exist but above all that history and human behavior are determined by race. Some thinkers who believed that races did exist still criticized racist ideas. On the other hand, those with anti-Semitic views did not necessarily refer to the 'Jewish race' in justification of their hostility towards Jews. There are also those who do not believe in the existence of races but still conceive of social relations in a deterministic manner by explaining relations between people in terms of their culture, their class, or their gender. Nevertheless, it is equally true that thinking in terms of race has, more often than not, led to racist ideas.

The interpretation of historical events in terms of race has often been linked to the need to justify European domination over the rest of the world (Park and Burgess 1924). This is the case of Marxist thinkers in particular. Cox, for example, claims that race relations first emerged in 1493 when the borders showing the limits of Spanish and Portuguese control in the Americas were first marked out (Cox 1948). Yet the negative representations of black-skinned peoples in relation to whites by far predated the age of European expansionism. The development of racist thought cannot be linked to one sole cause but must be seen as the result of a number of converging factors. What is remarkable is the capacity of deterministic racist thought to renew and reinvent itself throughout the different periods of history.

3. *Development of the Social Science Critique*

Although the first American sociologists justified slavery, social scientists have, at different stages, criticized the concept of racial determinism. Max Weber clearly denounced racial explanations of historical events. 'That there exists today, only one precise and concrete factor, relevant to Sociology, which clearly and permanently reduces any given sociological category down to innate or hereditary characteristics, whilst claiming that another category cannot ever not possess those same qualities, I categorically deny, and I will continue to deny this until it is proven right before my eyes' (Weber 1974, p. 120).

The work of the Chicago school sociologists (such as Charles S. Johnson, John Dollard, Allison Davies, E. Franklin, and even Gunar Myrdal whose work was based on the Chicago school's research) marked another stage in the social science critique. Through study of the 'assimilation' of various populations in American society, they replaced the notion of race with that of *relations* between races or *race relations*. Race, in the biological sense of the term, was replaced by a social definition: race became a social category which contributed to defining an individual's position in society, according to his origins which were inferred by certain biological characteristics. However, it is clear that this category is only acknowledged if society accepts that 'race' is meaningful. In this way, it was henceforth taken as given by all sociologists that it is racism that creates race, not vice versa.

The rediscovery of the weight of ethnicity by American sociologists in the 1960s led to a redefinition of different groups in terms of 'ethnic groups': an historical definition thus replaced a biological definition. However, these sociologists did not all unanimously agree that Afro-Americans or the native-Americans constituted a group like any other. It was argued that the term 'ethnic group' was perhaps no more than just a euphemism or hypocrisy as it tended to obscure the stigmatization, segregation, and discrimination which Afro-Americans and native-Americans had suffered as a result of their race. Wasn't 'ethnic' anything else than 'racial'? Shouldn't race relations be distinguished from ethnic relations, as individuals could abandon their ethnic group but could not escape their racial group?

The sociologists of *race relations* continue, however, to use the concept of race as though races do actually exist, and the sociologists of ethnicity as though ethnic groups are static, fixed entities. When, from the 1960s onwards, the term *ethnic group* took over from *race*, it continued to imply a certain hereditary link. Furthermore, both terms are closely linked to the American experience, a society made up of immigrants (ethnic groups), autochthonous peoples, and those brought over as slaves (racial groups). Should we not take into account the consequences of the critique of the sociology of race relations and stop claiming that

the term race is meaningless, only to then use it continuously, even in the titles of scholarly journals? Is it not possible to reach a new stage by adopting the concept of *historical collectivity*? If, as Weber had already observed, those who we refer to as Afro-Americans in the United States make up a particular group, this is not due to the color of their skin, but rather because they consider themselves and are considered by others to be the descendents of Africans who were subjected to a certain historical experience, which makes them a particular *historical collectivity* (Schnapper 1998).

Societies use natural differences and relations as a way to organize relationships between people. This is the case for gender and relationships between parents and children. In the same way, human beings use the physical differences, thought of in terms of race, in order to establish and identify social roles. Sociological thought has progressively criticized the notion of race. The relations between the 'races' are now analyzed as part of the processes of inclusion and exclusion of the different social groups in a more general analysis of social structures and political legitimacy.

See also: Critical Race Theory; Genetics and Mate Choice; Race and the Law; Race Identity

Bibliography

Banton M P 1977 *The Idea of Race*. Tavistock, London

Cox O C 1948 *Caste, Class & Race. A Study in social Dynamics*. Doubleday, New York

Gobineau J A 1853 *Essai sur l'inégalité des races*. Gallimard, La Pléiade, Paris

Guillaumin C 1972 *L'idéologie raciste. Genèse et langage actuel*. Mouton, Paris and La Haye

Jacob A 1981 La science face au racisme. In: Olender M (eds.) *Le racisme, Mythes et sciences*. Complexe, Brussels

Park R E, Burgess E W 1924 *Introduction to the Science of Sociology*. University of Chicago Press, Chicago

Schnapper D 1998 *La Relation à l'autre. Au coeur de la pensée sociologique*. Gallimard, Paris

Weber M 1974 Max Weber et les théories bioraciales du Xxème siècle. *Cahiers Internationaux de Sociologie* **1**: 115–26

D. Schnapper

Race Identity

The concept of 'racial identity' has generated many and different bodies of literature that bear directly upon its multiple deployments in the social and behavioral sciences. Given that every individual is universally and simultaneously also part of a larger group, from family to community to language group to (for most of the last three centuries) nation–state, the salient question for the social sciences is: under what conditions do individuals come to think of themselves as 'racial'? The corollary, but by no means symmetrical question is: under what conditions do larger social units, from nation–states to communities, come to see, classify, and identify the individual as 'racial'?

1. Racial Identity and Individual Identity

There is a vast literature on the psychology of personal identity that provides a particular angle for approach to the topic. Identity formation as a developmental process through adolescence is one significant track, and a subset of this work is early-stage racial identity formation. This research can be traced back more than a half-century to the studies of the preferences of preschool children for white or black dolls (Clark and Clark 1947). A parallel track of scholarship emphasizes the power of 'group identity' to shape behavior through the imposition of cultural and national identities (Burman and Harrell Bond 1979, Shibutani and Kwan 1969). From political mobilization to choice of a marital partner, each of these renderings of racial identity has substantially penetrated the disciplines of sociology, political science, anthropology, and social psychology. In the USA, the courts became the site of the connection between the vastly different worlds of research on racial identity formation and social and political institutions. The Supreme Court's 1954 decision on school desegregation was based partly upon research claiming that lowered self esteem of the subordinated group was 'inherently a function of racial segregation' (Brown vs. Board of Education), and explicitly referred to the doll studies noted above. This generated a research trajectory involving studies on individual responses to racial identification that would extend through the remainder of the twentieth century (Montalvo 1999).

While the concept of individual identity has no clearly locatable beginning, it has been traced, in the West, to 'the discovery of the individual' (Morris 1972, Doi 1985). Individuals typically identified with their families, then perhaps with the local village or community (Linton 1955). As communities grew and members became mobile, region, language group, religion, and the nation-state would each claim some measure of allegiance via identity. Multiplicity of identities for a single individual can also be understood in terms of the simultaneity of those identities. That is, a single individual can sense and affirm a complex set of identities, being simultaneously a grandfather and being Jewish, being French and being working-class, being an alcoholic and a chess player. No single identity need claim any essentialist primacy nor be privileged as 'fundamental' across time, space, and culture. The social context frames which identity will

be called up and acted upon, or be the reference point of further action for the individual.

The mobilization of constituencies around a perceived collective interest is the essence of political and social movements, and in this sense all political mobilization is 'identity politics.' In a work that would set the frame for later research on racial identity and political consciousness, E. P. Thompson's (1963) monumental social history of the 'making of the English working class' detailed how class identity is forged from a particular set of economic and social circumstances. In a different and parallel kind of scholarship, other researchers have sought to explain the emergence of a racial identity. For example, the US civil rights movement (1956–72) produced a number of scholarly analyses that demonstrated how different sets of circumstances explain the emergence and sustenance of racial identity (Hampton and Fayer 1990, Bloom 1987, Morris 1984). However, the most direct descendant of Thompson's (1963) demonstration of the historical specificity of political mobilization around identity, relevant to race, is David Roediger's *The Wages of Whiteness: Race and the Making of the American Working Class* (1991). Roediger mines the archives of the US labor movement of the nineteenth century to provide rich historical material showing how the racial identity of whites in general, and the Irish in particular, were formed in relationship to the enslavement of blacks.

2. *Politics and Identity Politics*

Thus racial identity was generated by massive historical, economic, and political forces that far transcend the level of individual choice (Banton 1977, Miles 1989). Indeed, there is a substantial body of research indicating that in the nineteenth century European and US academics expended much energy trying to demonstrate the intellectual and moral superiority of whites (Gould 1981, Bernal 1987). The biological and medical sciences of the late nineteenth century assumed a Darwinian model of racial hierarchy of adaptation to civilization (Frazer 1900). Thus, in order to counter the assertion of white superiority, from Europe to South Africa to the North American continent to Australia, groups that wished to mobilize to change their position in a racial hierarchy often found that racial identity was a useful and powerful basis for achieving group solidarity (Kelley 1997, Fredrickson 1995, Crenshaw et al. 1995). Obviously, these groups would encounter competing group identities, such as tribal, regional, ethnic, linguistic, and class. This list would permit the white colonialists the advantage of playing one off against the other, and the strategic use of ethnic identity to trump 'race' could be deployed. For an elaboration of the distinction between race and ethnicity, see Fenton (1999).

If we take even a casual excursion through the last few centuries of racial classification, there is overwhelming evidence on the side of those who have argued that race is arbitrary, shifting, and often biologically and sociologically inconsistent and contradictory. A consortium of leading scientists across the disciplines from biology to physical anthropology issued a 'Revised UNESCO Statement on Race' in 1995—a definitive declaration that summarizes 11 central issues, and concludes that in terms of 'scientific' discourse, there is no such thing as a 'race' that has any scientific utility:' … the same scientific groups that developed the biological concept over the last century have now concluded that its use for characterizing human populations is so flawed that it is no longer a scientifically valid concept. In fact, the statement makes clear that the biological concept of race as applied to humans has no legitimate place in biological science' (Katz 1995, pp. 4, 5).

The rule that one drop of black blood makes one black is the easy mark along a full continuum of contradictory and incoherent taxonomies. 'Passing' and 'miscegenation laws' and slaveowner/slave offspring do more than simply dot the landscape of this topic (Harris 1995, Haney-López 1996). This continuum extends well into the present period where we find more and more people asserting a mixed-race identity. Since the classification of race is arbitrary and often whimsical (e.g., one drop of blood), accepting the idea that race is something identifiable with fixed borders that could be crossed and thus 'mixed' is just another indication of the power of social forces shaping racial identity. At the biochemical level of blood types and hematology, at the neurological level of neurotransmission patterns, at the level of cell function—at all these levels, we are all 'mixed' by any taxonomy or measure of allele frequencies in large population groups (Molnar 1992).

Yet a number of scholars of race have noted the increasing use of the category of 'mixed race' as the self-identity of a younger generation, mainly of relatively privileged and middle-class persons. Paul Spickard (1989) and Werner Sollers (1997) have each produced important studies of 'mixed race' identities that cross boundaries of Europe and the Americas. Maria Root (1996) has culled a collection of essays that deal with identity construction along ethnic and racial lines, and further explores these issues well beyond the black and white focus that has riveted much of this literature. The difference between 'individual choice' of a racial identity as opposed to having a racial identity 'imposed' by the state and/or the obdurate empirical reality of a group consensus (based primarily upon phenotype-stereotype) is profound, yet variable over time and place. There are concrete economic, political, social, and sometimes psychological benefits to choosing certain identities. That is why scores of thousands of light-skinned 'blacks' under slavery in Brazil, Jamaica, and the USA

chose to 'pass' as white (Small 1994). For parallel reasons, thousands of persons in the USA have attempted to reclassify themselves as Native American or Latino or African American to benefit from affirmative action laws.

For an extensive theoretical treatment of the way in which racial taxonomies shift over time, see Omi and Winant (1986). In addition, for specific cases of this phenomenon, see Harris (1995), Haney-López (1996) and Takaki (1998). From the other side, namely, individuals who come to collectively redefine their own situation, and thereby reveal how 'race' is sometimes in flux due as much to agency, interpretation, and circumstance, see Blumer (1958), Lal (1990), and Perry (1998). For an 'interest-driven' version of the collective will to create, sustain, or maintain a particular racial identity, see Lipsitz (1998) and Wellman (1993).

3. *White Identity*

In recent years, the concept of white racial identity has come under increasing focus as a basis for a research agenda. Blumer's (1958) classic article reframed race as a structural relationship between competing group interests. This position was adopted and developed by an early entry into this field, David Wellman's *Portraits of White Racism* (1993). Roediger (1991, 1994), Frankenberg (1993), Allen (1994), Jacobson (1998), Lipsitz (1998), and Perry (1998) have each examined the ways in which white privilege is embedded in a faceless and taken-for-granted normalness of white identity. Twine and Warren (2000) have assembled an anthology in which they bring to greater consciousness the ways in which researchers in racially stratified societies ignore at their peril the subject world of race as an integral feature of methodology, largely through the racial identity of subjects and researchers.

4. *Which Identity Gets Triggered?*

Given that there are multiple and simultaneous identities, the task for the social analyst is to try to understand the social circumstances in which a particular set of identities gets 'triggered' or brought to the fore. That is, rather than arguing for the primacy or 'essentialist' character of an identity, the analytic task is to understand why identity A, B, or C emerges as primary. War between nation-states is an especially propitious time for a 'national identity' to be forged, or come to the fore. Thus, it may be heuristically useful to think of a series of latent multiple identities that get turned on or off based upon social and political circumstances—and determine 'which button gets pushed.' In wartime, the national identity button gets pushed frequently, and lights up. In courtship, the gender button is pushed, and so forth. In these terms, the empirical question of when and why the 'racial identity' is triggered is no less compelling when posed as a theoretical question.

We can expect the next period to generate many more investigations of racial identities deployed as one among multiple identities—moving away from static conceptions of fixed boundaries of racial identity, with richer explorations of its contingent and sometimes fluid character. An example of this kind of work at the vanguard would be Gloria Anzaldua's (1987) *Borderlands/La Frontera: The New Mestiza* and George Lipsitz' (1991) *Collective Memory and American Popular Culture*. These authors identify 'border identities,' which are emergent identities crafted and recrafted continually by movements across different borders. As the new millennium witnesses the increasing thousands and even millions of border crossings (and recrossings) with the movement of capital and labor, studies of 'borderland identities' around issues of race and ethnicity will come to play a significant role in this field of inquiry.

See also: Civil Rights Movement, The; Identity: Social; Integration: Social; Racial Relations; Residential Segregation: Sociological Aspects

Bibliography

Allen T 1994 *The Invention of the White Race*. Verso, New York
Anzaldua G 1987 *Borderlands/La Frontera: The New Mestiza*. Spinsters/Aunte Lute Press, San Francisco
Banton M 1977 *The Idea of Race*. Tavistock, London
Bernal M 1987 *Black Athena: The Afroasiatic Roots of Classical Civilization*. Rutgers University Press, New Brunswick, NJ
Bloom J M 1987 *Class, Race, and the Civil Rights Movement*. Indiana University Press, Bloomington, IN
Blumer H 1958 Race prejudice as a sense of group position. *Pacific Sociological Review* **1**(Spring): 3–6
Burman S B, Harrell-Bond B E 1979 *The Imposition of Law*. Academic Press, New York
Clark K B, Clark M P 1947 Racial identification and preference in Negro children. In: Newcomb T M, Hartley E (eds.) *Readings in Social Psychology*. Henry Holt, New York, 602–11
Crenshaw K, Gotanda N, Peller G, Thomas K 1995 *Critical Race Theory*. The Free Press, New York. pp. 609–611
Doi T 1985 *The Anatomy of Self: The Individual versus Society*. Kodansha International, Tokyo
Fenton S 1999 *Ethnicity: Racism, Class and Culture*. Rowman & Littlefield, London
Frankenberg R 1993 *White Women, Race Matters: The Social Construction of Whiteness*. University of Minnesota Press, Minneapolis, MN
Frazer J 1900 *The Golden Bough: A Study in Magic and Religion*, 2nd edn. Macmillan, New York
Fredrickson G M 1995 *Black Liberation: A Comparative History of Black Ideologies in the United States and South Africa*. Oxford University Press, New York
Gould S J 1981 *The Mismeasure of Man*. W. W. Norton, New York

Hall S 1991 Ethnicity: Identity and difference. *Radical America* **23**(4): 9–19
Hampton H, Fayer S (eds.) 1990 *Voices of Freedom: An Oral History of the Civil Rights Movement from the 1950s Through the 1980s*. Bantam Books, New York
Haney-López I 1996 *White By Law: The Legal Construction of Race*. New York University Press, New York
Harris C I 1995 Whiteness as property. In: Crenshaw K, Gotanda N, Peller G, Thomas K 1995 (eds.) *Critical Race Theory*. The Free Press, New York, pp. 276–91
Ifekwunigwe J O 1999 *Scattered Belongings. Cultural Paradoxes of 'Race,' Nation and Gender*. Routledge, New York
Jacobson M F 1998 *Whiteness of a Different Color: European Immigrants and the Alchemy of Race*. Harvard University Press, Cambridge, MA
Jordan W D 1968 *White Over Black, White Over Black: American Attitudes Toward the Negro, 1550–1912*. University of North Carolina Press, Chapel Hill, NC
Katz S H 1995 Is race a legitimate concept for science? *The AAPA Revised Statement on Race: A Brief Analysis and Commentary*. University of Pennsylvania, February
Kelley R D G 1997 Identity politics and class struggle. *New Politics* **VI**(2): 84–97
Lal B B 1990 *The Romance of Culture in an Urban Civilization: Robert E. Park on Race and Ethnic Relations in Cities*. Routledge, New York
Linton R 1955 *The Tree of Culture*. Alfred A. Knopf, New York
Lipsitz G 1991 *Time Passages: Collective Memory and American Popular Culture*. University of Minnesota Press, Minneapolis, MN
Lipsitz G 1998 *The Possessive Investment in Whiteness: How White People Profit from Identity Politics*. Temple University Press, Philadelphia, PA
Miles R 1989 *Racism*. Routledge, London
Molnar S 1992 *Human Variation: Races, Types, and Ethnic Groups*, 3rd edn. Prentice Hall, Englewood Cliffs, NJ
Montalvo F F 1999 The critical incident interview and ethno-racial identity. *Journal of Multicultural Social Work* **7**: 1–43
Morris A D 1984 *The Origins of The Civil Rights Movement: Black Communities Organizing for Change*. Free Press, Collier Macmillan, New York
Morris C 1972 *The Discovery of the Individual, 1050–1200*. Harper & Row, New York
Omi M, Winant H 1986 *Racial Formation in the United States: From the 1960s to the 1980s*. Routledge & Kegan Paul, New York
Perry P 1998 Beginning to see the white. Unpublished PhD thesis, University of California, Berkeley
Roediger D R 1991 *The Wages of Whiteness, Race and the Making of the American Working Class*. Verso, New York
Roediger D R 1994 *Towards the Abolition of Whiteness: Essays on Race, Politics, and Working Class*. Verso, New York
Root M P P 1996 *The Multiracial Experience: Racial Borders as the New Frontier*. Sage, Thousand Oaks, CA
Shibutani T, Kwan K W 1969 *Ethnic Stratification: A Comparative Approach*. Macmillan, London
Small S 1994 *Racialised Barriers: The Black Experience in England and the United States in the 1980s*. Routledge, New York
Smedley A 1993 *Race in North America: Origin and Evolution of a Worldview*. Westview Press, Boulder, CO
Sollers W 1997 *Neither Black Nor White, Yet Both: Thematic Explorations of Interracial Literature*. Oxford University Press, New York
Spickard P 1989 *Mixed Blood: Intermarriage and Ethnic Identity in Twentieth-century America*. University of Wisconsin Press, Madison, WI
Takaki R T 1998 *Strangers From a Different Shore: A History of Asian Americans*. Little Brown, Boston
Thompson E P 1963 *The Making of the English Working Class*. Vintage Books, New York
Twine F W, Warren J (eds.) 2000 *Racing Research, Researching Race*. New York University Press, New York
Wellman D T 1993 *Portraits of White Racism*, 2nd edn. Cambridge University Press, New York

T. Duster

Race Relations in the United States, Politics of

The problem of race has been very badly handled in political science. Except for Marxists who see racism and capitalism as intertwined, postmodern skeptics, and those like V. O. Key Jr., who studied the South, democratic theorists have not systematically addressed the racial factor in American politics (McClain and Garcia 1993, Dawson and Wilson 1991). Yet although the problems of racism, racial discrimination, and inequality have scarcely been resolved, many scholars of this current post-civil rights period have continued to ignore the profound and dynamic relationship between race and American politics. A large part of their failure to address the problem of race in America stems from the pervasiveness of a 'gradualist' liberal bias that views the race problem as ultimately resolving itself through integration and assimilation of racial minorities as well as through the disappearance of White racism. From Myrdal (1944) to Thernstrom and Thernstrom (1997), widely disseminated and important treatments of race relations in the USA remain rooted in this gradualist bias. Some scholars, however, recognize that race and ethnicity remain central features of a rational and modernizing society, and have begun to explore the political foundations of racial conflict and racism. The latest research is expanding the question of race to include a focus on groups other than the politics of American Blacks and Whites.

1. *Black Politics*

The dominant theories put forth by leading American democratic theorists following World War II obviously excluded African-Americans in broadly characterizing the American political process as one at its core, where no one group is excluded and all, although hardly equals, can be heard. Even after such theories were later revised to portray the inclusion of Blacks as

an inevitability, because as Robert Dahl explained, 'democratic cultures have considerable capacity for correcting their own failures' (1989, p. 180), what politics would look like after their inclusion was poorly foreshadowed and misunderstood by such scholars. Democratic theorists like Dahl (1961) and David Truman (1971) generally assumed that the political incorporation of Blacks would 'normalize' American politics, with all the various groups and subgroupings able to engage in pressure politics. Browning et al. (1984) in an important book, would argue that eventually minorities were integrated and empowered in the American pluralist system (Eisenger 1982, Sigelman and Welch 1982, Sonenshein 1993). These same authors and others, would later critically re-evaluate their claim of 'empowerment' for American minorities (Hero 1992, Browning et al. 1991). But others scholars, including Adolph Reed Jr. (1988) would seriously question whether the pluralist model could ever be refashioned to fit American Blacks, since to do so would be to ignore the 'racial element' in American politics (see also Holden 1973, Jones 1972, McLemore 1972, Morris 1975, Pinderhughes 1987).

Given its exclusion in the discipline, research on the politics of Blacks would constitute a separate inquiry, apart from the mainstream (McClain and Garcia 1993). Research on Black politics would not only establish the continuing significance of race in American politics (Carmichael and Hamilton 1967, Jones 1972), but also the fact that politics as practiced, encompasses many more activities than recognized by mainstream scholars. Social movement scholars established how protest is not unlike electoral politics (McAdam 1982, Morris 1984) and the continuity between the two in the postcivil rights era (Holden 1973, Smith 1981, Tate 1993, Walton 1973). Jesse Jackson's 1984 and 1988 presidential bids clearly reflected the continuity (Marable 1985, 1990). Religion and social context, important to the emergence of the civil right movement, remain central to Black electoral politics (Cohen and Dawson 1993, Harris 1994). The racial identities of Blacks, a factor in the movement, continues to play a critical role in their political participation and attitudes today (Dawson 1994, Miller et al. 1981, Tate 1993, Shingles 1981, Verba and Nie 1972).

Nevertheless, William J. Wilson's (1981) controversial thesis that race is receding in importance relative to class in the Black community, lay the seeds for a new controversy in Black politics over the degree to which race matters. Carol M. Swain's (1993) award-winning book would question the relevance of race in the political representation of Blacks; others would contest it (Cannon 1999, Whitby 1997) or offer qualifications of it (Lublin 1996). While this debate continues, the study of the politics of Blacks which was generally the only way political scientists addressed the question of race has actually reshaped the contours and direction of mainstream work in political science. New attention is now paid to the nonelectoral antecedents of politics (e.g., Verba et al. 1995) and on the relationship between the institution and political behavior (e.g., Rosenstone and Hansen 1993) as others insisted was critical in the study of Black politics (e.g., Barker et al. 1998, Walton 1985).

2. *The Debate over Racial Prejudice*

Although the bulk of political science has ignored race, there exists nevertheless a voluminous literature devoted to the empirical study of how racial attitudes influence the policy attitudes of Americans. Nearly all of this empirical work has focused on Whites, although some scholars have addressed the racial and policy attitudes of Blacks (Hochschild 1995, Sigelman and Welch 1991). Broadly, two lines of argument have emerged in the study of White racial attitudes. Those on one side of the debate argue that racial prejudice remains the fundamental factor behind Whites' opposition to racial policies such as busing school children to achieve racially balanced public schools, and affirmative action. White racism has not disappeared, as Whites' responses to the general principle of racial equality items imply. For some scholars, education has merely given more White Americans the ability to appear more racially tolerant than they really are (Jackman and Muha 1984). Rather, racial intolerance along with fundamental American values about the work ethic and individualism blended together, explains the negative attitudes that Whites have toward government efforts to eradicate racial inequality (Kinder 1986, Kinder and Sanders 1996, Kinder and Sears 1981, McConahay 1986, McConahay and Hough 1976, Sears 1988, Sears et al. 2000). Using their defense of traditional values as a cover for their resentment and fears of Blacks, Whites strongly object to policies that would change the racial status quo.

Those on the other side of the debate have sought to account for White opposition to racial policies differently, choosing to focus on factors other than racial prejudice, and not only because they provide a more flattering portrayal of White Americans. For those seeking alternative explanations, ideology and moralistic values and beliefs are posited to be the major reasons underlying White opposition to racial implementation policies. Politically conservative Whites object to such policies because of their dislike of 'big government.' Kuklinski and Parent 1981, Margolis and Hague 1981, Sniderman et al. 1984, 1991). In addition, these scholars point to the 'natural contradiction' between the White majority's commitment to individual rights and freedoms, its preference for a limited government, and support for new government programs that would favor Blacks (Carmines and Merriman 1993, Sniderman and Hagen 1985). Pro-

grams that are designed to promote racially equal results at once contradict the very values, including color-blindness, that most Whites now espouse (Sniderman and Piazza 1993).

Along these lines, other scholars have sought to link White racial attitudes to their beliefs about the causes of racial inequality and racial stratification. Several scholars have shown that Whites will not support programs that assist Blacks if they believe the system to be fair and that Blacks themselves are not working hard enough to get ahead (Kluegel and Smith 1986, 1986, Sniderman and Hagen 1985). While civil rights organizations have asserted that Blacks are denied equal opportunities to Whites to get ahead, most White Americans express full faith in the abundance of opportunity in American society. Their faith contrasts greatly with that of Blacks (Kluegel and Bobo 1993, Kluegel and Smith 1986, pp. 67–68, Sigelman and Welch 1991). In the 1990 General Social Survey, 72 percent of Blacks, but only 37 percent of Whites, thought that racial discrimination was a main factor in the unequal outcomes for Blacks in jobs, income, and housing. Given that most Whites believe that Blacks currently have equal opportunities to Whites and that the economic system works well, most oppose programs that demand equality of outcomes.

As one of the more radical theoretical shifts, the meaning of racial prejudice has changed (see Sniderman et al. 1993, for a similar discussion). Once considered to be an archaic, irrational need located deep in a person's heart to scapegoat others, research has consistently favored another view, that racial bias is ubiquitous, forming naturally in a person's mind through social group interactions (Nisbett and Ross 1980). The new definition of prejudice as an instinctive and widespread phenomenon can be interpreted in two ways. On the one hand, the very ubiquitousness of racial stereotypes undermines their relevance in the study of White racial attitudes. Because stereotypes are so widespread, Whites' opposition to racial policies must more critically depend on other factors, such as politics, ideologies, and the framing of the issue. On the other hand, the ubiquitous nature of racial stereotypes exposes its foundation in the social stratification system or a racial hierarchy and in history. Scholars working in sociology and social psychology argue that racism, new or old, persists because it justifies and extends one group's domination over another (Sidanious and Pratto 1993, Sidanius et al. 1996).

3. *The New Institutionalism in Racial Politics*

There is a new and different approach to the study of race and politics that I classify under the heading of the 'new institutionalism in racial politics.' The core thesis is that politics is as much an underlying cause of America's racial cleavages as it gives expression to such cleavages. Although consistent with a growing branch of political science that argues that institutions structure political behavior, not only political outcomes (March and Olsen 1984), to argue that politics is structurally implicated in America's racial divisions still represents an important break from the intellectual tradition in political science.

American federalism was a pragmatic response to the contentious issue of slavery, but greatly prolonged the institution of slavery and the subjugation of American Blacks. States would use their constitutional powers to deny citizenship to Mexicans and Asian immigrants, as well as deny Blacks their new right to vote under the Fourteenth Amendment. An entire subfield exists that examines the ways that electoral arrangements and practices negate the Black vote (e.g., Davidson and Grofman 1994, Grofman et al., 1992, Kousser, 1999).

New research on racial politics and voting has continued to identify factors other than the racial attitudes of Whites in their rejection of Black candidates and their defection from the Democratic party (see Giles and Buckner 1993, Giles and Evans 1986, Giles and Hertz 1994, Glaser 1994, Huckfeldt and Kohfeld 1989, Reeves 1997) Racial prejudice is not deemed unimportant or irrelevant in these new studies, but the racial behavior of Whites is conceptualized as more situational-based than attitudinally-based. Candidates, political parties, and the media can make race the focal point of the election generally, but not always, to the detriment of the Black candidate (Lublin and Tate 1995, Metz and Tate 1995, Reeves 1997).

Finally, while sociologists and social psychologists contend that racial categories are socially constructed, some political scientists (Smith 1997) have begun to show how in the USA racial identities have been politically constructed through US laws covering immigration, state laws that defined race, the US census, and affirmative action guidelines along with legal and political discourse. Comparative race studies have emerged, underscoring the role that state policies play in defining race (Gilroy 1991, Marx 1998).

4. *Beyond Black and White*

Finally, most of the research on racial prejudice and racial politics has been conceptualized and tested on Whites and Blacks. Researchers have begun to recognize that even race relations defined as narrowly as Black–White relations is critically related to status and politics of other American minorities as well, notably America's fastest-growing groups, Asian-Americans and Latinos (Bedolla 2000, Cain et al. 1991, De la Garza et al. 1992, DeSipio 1996, Espiritu 1992, Garcia 1997, Hardy-Fanta 1993, Hero 1992, Jennings 1994, Jennings and Rivera 1984, Saito 1998, Uhlaner et al. 1989, Wrinkle et al. 1996). Whether these groups suffer discrimination on the basis of their nationality, and

the degree to which national origin affects their politics is often the unstated subtext of in these body in work (but see Kim, 2000). In response to recent highly publicized conflicts between Blacks, Latinos, and Koreans in cities such as Los Angeles and New York, much of this work addresses interminority conflict (Abelmann and Lie 1995, Kim, in press, McClain 1993, McClain and Karnig 1990) Finally, new work has explored not just interracial or interethnic politics, but intraminority group politics. The limits of treating African American and Latino politics in 'gender-and-sex neutral' terms, for example, have been exposed in Cathy Cohen's (1999) study of politics of AIDS in the Black community and Carol Hardy-Fanta's (1993) work on community organizing among Latinos. Moving beyond the Black–White dichotomy is critical for reaching a better understanding of not only the politics of the neglected 'other' minorities, but for Black politics and US politics more broadly.

While research on race exists and grows, it still remains a neglected focus in political science. The 1990s scholarship I have presented clearly establishes that race does not lie on a progressive-regressive continuum as the gradualist assumption posits. This idea that we are headed in the direction of a color-blind society remains extremely popular to scientists and nonscientists alike. Nevertheless, ample social science research in social and cognitive psychology shows that people routinely discriminate, that racial stereotypes persist (Nisbett and Ross 1980). Race therefore remains a dynamic feature of American politics, whose complexities we have only just begun to understood and appreciate. At best, one can talk about minimizing its significance. It is vital that we understand how politics is implicated in the persistence of racial divisions if we are to have a true democracy organized in a way that is truly inclusive, working for the common good of us all, not just for those at the top of the racial order. To do this, political scientists must pay greater and careful attention to the problem of race than they have in the past.

See also: Civil Rights Movement, The; Ethnic and Racial Social Movements; Ethnic Conflict, Geography of; Ethnic Conflicts; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns; Political Protest and Civil Disobedience; Race Identity; Racial Relations; Racism, History of; Racism, Sociology of; Residential Segregation: Geographic Aspects; Residential Segregation: Sociological Aspects; Urban Politics: United States; Urban Poverty in Neighborhoods

Bibliography

Abelmann N, Lie J 1955 *Blue Dreams: Korean Americans and the Los Angeles Riots*. Harvard University Press, Cambridge, MA

Barker L J, Jones M H, Tate K 1998 *African Americans and the American Political System*, 4th edn. Prentice Hall, Englewood Cliffs, NJ

Bedolla L G 2000 They and we: Identity, gender, and politics among Latino youth in Los Angeles. *Social Science Quarterly* **81**: 106–22

Bobo L 1983 Whites' opposition to busing: Symbolic racism or realistic group conflict? *Journal of Personality and Social Psychology* **45**: 1196–210

Bobo L 1986 Group conflict, prejudice, and the paradox of contemporary racial attitudes. In: Katz P A, Taylor D A (eds.) *Eliminating Racism: Means and Controversies*. Plenum Press, New York

Browning R P, Marshall D R, Tabb D H 1984 *Protest is Not Enough*. University of California Press, Los Angeles and Berkeley

Browning R P, Marshall D R, Tabb D H (eds.) 1991 *Racial Politics in American Cities*, 2nd edn. Longman, New York

Cain B E, Kiewiet D R, Uhlaner C J 1991 The acquisition of partisanship by Latinos and Asian Americans. *American Journal of Political Science* **35**: 390–442

Cannon D T 1999 *Race, Redistricting, and Representation*. University of Chicago Press, Chicago

Carmichael S, Hamilton C V 1967 *Black Power*. Vintage Press, New York

Carmines E G, Merriman W R Jr. 1993 The changing American dilemma: Liberal values and racial policies. In: Sniderman P M, Tetlock P E, Carmines G E (eds.) *Prejudice, Politics, and the American Dilemma*. Stanford University Press, Stanford, CA, pp. 237–55

Cohen C J 1999 *The Boundaries of Blackness: AIDS and the Breakdown of Black Politics*. University of Chicago Press, Chicago

Cohen C J, Dawson M C 1993 Neighborhood poverty and African-American politics. *American Political Science Review* **87**: 286–302

Dahl R 1961 *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT

Dahl R 1989 *Democracy and Its Critics*. Yale University Press, New Haven, CT

Davidson C, Grofman B (eds.) 1994 *Quiet Revolution in the South: The Impact of the Voting Rights Act of 1965–1990*. Princeton University Press, Princeton, NJ

Dawson M C 1994 *Behind the Mule: Race and Class in African-American Politics*. Princeton University Press, Princeton, NJ

Dawson M C, Wilson E J III 1991 Paradigms and paradoxes: Political science and African American politics. In: Crotty W (ed.) *Theory and Practice of Political Science*. Northwestern University Press, Evanston, IL

De la Garza R O, DeSipio L, Garcia F C, Garcia J A, Falcon A 1992 *Latino Voices: Mexican, Puerto Rican, and Cuban Perspectives on American Politics*. Westview Press, Boulder, CO

DeSipio L 1996 *Counting on the Latino Vote*. University of Virginia Press, Charlottesville, VA

Espiritu Y L 1992 *Asian American Panethnicity: Bridging Institutions and Identities*. Temple University Press, Philadelphia

Eisenger P 1982 Black empowerment in municipal jobs: the impact of Black political power. *American Political Science Review* **76**: 380–92

Fanta-Hardy C 1993 *Latina Politics, Latino Politics*. Temple University Press, Philadelphia

Garcia F C 1997 *Pursuing Power: Latinos and the Political System*. University of Notre Dame Press, Notre Dame, IN

Giles M W 1977 Percent black and racial hostility: An old assumption reexamined. *Social Science Quarterly* **58**: 412–17
Giles M W, Buckner M 1993 David Duke and black threat: An old hypothesis revisited. *Journal of Politics* **55**: 702–13
Giles M W, Evans A 1986 The power approach to intergroup hostility. *Journal of Conflict Resolution* **30**: 469–85
Giles M W, Hertz K 1994 Racial threat and partisan identification. *American Political Science Review* **88**: 317–26
Gilroy P 1991 *There Ain't No Black in the Union Jack*. University of Chicago Press, Chicago
Glaser J M 1994 Back to the black belt: Racial environment and white racial attitudes in the South. *Journal of Politics* **56**: 21–41
Grofman B, Handley L, Niemi R G 1992 *Minority Representation and the Quest for Voting Equality*. Cambridge University Press, New York
Harris F 1994 Something from within: Religion as a mobilizer of African-American political activism. *Journal of Politics* **56**: 42–68
Hero R 1992 *Latinos and the Political System: Two-tiered Pluralism*. Temple University Press, Philadelphia
Hochschild J 1995 *Facing Up to the American Dream: Race, Class, and the Soul of the Nation*. Princeton University Press, Princeton, NJ
Holden M 1973 *The Politics of the Black Nation*. Chandler Publishing, New York
Huckfeldt R, Kohfeld C W 1989 *Race and the Decline of Class in American Politics*. University of Illinois Press, Urbana and Chicago
Jackman M, Muha M J 1984 Education and inter-group attitudes: Moral enlightenment, superficial democratic commitment, or ideological refinement. *American Sociological Review* **49**: 93–102
Jennings J (ed.) 1994 *Blacks, Latinos, and Asians in Urban America: Status and Prospects for Politics and Activism*. Praeger, Westport, CT
Jennings J, Rivera M (eds.) 1984 *Puerto Rican Politics in Urban America*. Greenwood Press, Westport, CT
Jones M H 1972 A frame of reference for Black politics. In: Henderson L J (ed.) *Black Political Life in the United States*. Chandler Publishing Company, San Francisco
Kim C J In Press. 1997 *Bitter Fruit: The Politics of Black-Korean Conflict in New York City*. Yale University Press, New Haven, CT
Kim C J 1999 The racial triangulation of Asian-Americans. *Politics and Society* **27**: 105–38
Kinder D R 1986 The continuing American dilemma: White resistance to racial change 40 years after Myrdal. *Journal of Social Issues* **42**: 151–71
Kinder D R, Sanders L M 1996 *Divided by Color: Racial Politics and Democratics Ideals*. University of Chicago Press, Chicago
Kinder D R, Sears D O 1981 Prejudice and politics: Symbolic racism versus racial threats to the good life. *Journal of Personality and Social Psychology* **40**: 414–31
Kouser J M 1999 Colorblind injustice: Minority voting rights and the undoing of the second reconstruction. University of North Carolina Press, Chapel, NC
Kluegel J R, Bobo L 1993 Dimensions of whites' beliefs about the black–white socioeconomic gap. In: Sniderman P M, Tetlock P E, Carmines E G (eds.) *Prejudice, Politics, and the American Dilemma*. Stanford University Press, Stanford, California, pp. 127–47
Kluegel J R, Smith E R 1986 *Beliefs about Inequality*. Aldine de Gruyter, Hawthorne, NY
Kuklinski J H, Parent T W 1981 Race and big government: Contamination in measuring racial attitudes. *Political Methodology* **7**: 131–59
Lublin D 1997 *The Paradox of Representation: Racial Gerrymandering and Minority Interests in Congress*. Princeton University Press, Princeton, NJ
Lublin D, Tate K 1995 Racial group competition in US mayoral elections. In: Peterson P E (ed.) *Classifying by Race*. Princeton University Press, Princeton, NJ
Marable M 1985 *Black American Politics: From the Washington Marches to Jesse Jackson*. Thetford Press, London
Marable M 1990 *Race, Reform, and Rebellion*, 2nd edn. University Press of Mississippi, Jackson, MS
March J G, Olsen J P 1984 The new institutionalism: Organizational factors in political life. *American Political Science Review* **78**: 734–49
Margolis M, Hague K 1981 Applied tolerance or fear of government? An alternative interpretation of Jackman's finding. *American Journal of Political Science* **25**: 241–55
Marx A W 1998 *Making Race and Nation: A Comparison of South Africa, the United States, and Brazil*. Cambridge University Press, New York
McAdam D 1982 *Political Process and Development of Black Insurgency, 1930–1970*. University of Chicago Press, Chicago
McClain P D 1993 The changing dynamics of urban politics—Black and Hispanic municipal employment: Is there competition? *Journal of Politics* **55**: 399–414
McClain P D, Garcia J A 1993 Expanding disciplinary boundaries: Black, Latino, and racial minority group politics in political science. In: Finifter A W (ed.) *Political Science: The State of the Discipline II*. The American Political Science Association, Washington, DC, pp. 247–79
McClain P D, Karnig A K 1990 Black and Hispanic socioeconomic and political competition. *American Political Science Review* **84**: 535–45
McConahay J B 1986 Modern racism, ambivalence, and the modern racism scale. In: Dovidio J, Gaertner S L (eds.) *Prejudice, Discrimination, and Racism*. Academic Press, Orlando, FL, pp. 91–125
McConahay J B, Hough J C 1976 Symbolic racism. *Journal of Social Issues* **32**: 23–45
McLemore L B 1972 Toward a theory of Black politics—The Black and ethnic models revisited. *Journal of Black Studies* **2**: 323–31
Metz D H, Tate K 1995 The color of the campaign. In: Peterson P (ed.) *Classifying by Race*. Princeton University Press, Princeton, NJ
Miller A, Gurin P, Malunchuk O 1981 Group consciousness and political participation. *American Journal of Political Science*. **25**: 494–511
Morris A 1984 *The Origins of the Civil Rights Movement*. Free Press, New York
Morris M D 1975 *The Politics of Black America*. Harper and Row, New York
Myrdal G 1944 *An American Dilemma: The Negro Problem and Modern Democracy*. Harper and Brothers, New York
Nisbett R, Ross L 1980 *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall, Inc, Englewood Cliffs, NJ
Pinderhughes D 1987 *Race and Ethnicity in Chicago Politics*. University of Illinois Press, Urbana and Chicago
Reed A Jr. 1988 The Black urban regime: Structural origins and constraints. In: Smith M P (ed.) *Power, Community, and the City*. Transaction Books, New Brunswick, NJ, Vol. 1

Reeves K 1997 *Voting Hopes or Fears*? Oxford University Press, New York
Rosenstone S J, Hansen J M 1993 *Mobilization, Participation, and Democracy in America*. Macmillan Publishing Co., New York
Saito L T 1998 *Race and Politics: Asian Americans, Latinos, and Whites in a Los Angeles Suburb*. University of Illinois Press, Urbana and Chicago
Sears D O 1988 Symbolic Racism. In: Katz P A, Taylor D A (eds.) *Eliminating Racism: Profiles in Controversies*. Plenum Press, New York
Sears D O, Kinder D R 1985 Whites' opposition to busing: On conceptualizing and operationalizing conflict. *Journal of Personality and Social Psychology* **48**: 1141–7
Sears D O, Sidanius J, Bobo L 2000 *Racialized Politics: The Debate About Racism in America*. University of Chicago Press, Chicago
Shingles R 1981 Black consciousness and political participation. *American Political Science Review* **75**: 76–91
Sidanius J, Pratto F 1993 The inevitability of oppression and the dynamics of social dominance. In: Sniderman P M, Tetlock P E, Carmines E G (eds.) *Prejudice, Politics, and the American Dilemma*. Stanford University Press, Stanford, CA, pp. 173–211
Sidanis J 1996 Racism, conservation, affirmative action and intellectual sophistication: a matter of principled conservatism or group dominance? *Journal of Personality and Social Psychology*. **70**: 476–90
Sigelman L, Welch S 1991 *Black Americans' Views of Racial Inequality: The Dream Deferred*. Cambridge University Press, New York
Smith R C 1981 Black power and the transformation from protest to politics. *Political Science Quarterly* **96**: 431–43
Smith, R M 1997 *Civic Ideals: Conflicting Visions of Citizenship in US History*. Yale University Press, New Haven, CT
Sniderman P M, Hagen M G 1985 *Race and Inequality: A Study in American Values*. Chatham House, Chatham, NJ
Sniderman P M, Piazza T 1993 *The Scar of Race*. Harvard University Press, Cambridge, MA
Sniderman P M, Tetlock P E 1986 Symbolic racism: Problems of motive attribution in political analysis. *Journal of Social Issues* **42**: 129–50
Sniderman P M, Brody R A, Kuklinski J 1984 Policy reasoning and political values: The problem of racial equality. *American Journal of Political Science* **28**: 75–94
Sniderman P M, Piazza T, Tetlock P E, Kendrick A 1991 The new racism. *American Journal of Political Science* **35**: 423–47
Sniderman P M, Tetlock P E, Carmines E G 1993 Prejudice and politics: An introduction. In: Sniderman P M, Tetlock P E, Carmines E G (eds.) *Prejudice, Politics, and the American Dilemma*. Stanford University Press, Stanford, CA, pp. 1–31
Sonenshein R 1993 *Politics in Black and White: Race and Power in Los Angeles*. Princeton University Press, Princeton, NJ
Swain C M 1993 *Black Faces, Black Interest: The Representation of African Americans in Congress*. Harvard University Press, Cambridge, MA
Tate K 1993 *From Protest to Politics: The New Black voters in American Elections*. Harvard University Press and the Russell Sage Foundation, Cambridge, MA
Thernstrom S, Thernstrom A 1997 *America in Black and White*. Simon and Schuster, New York
Truman D B 1971 *The Governmental Process: Political Interests and Public Opinion*, 2nd edn. Knopf, New York
Uhlaner C J, Cain B, Kiewiet D R 1989 Political participation of ethnic minorities in the 1980s. *Political Behavior* **11**: 195–231
Verba S, Nie N H 1972 *Participation in America: Political Democracy and Social Equality*. Harper and Row, New York
Verba S, Schlozman K L, Brady H E 1995 *Voice and Equality*. Harvard University Press, Cambridge, MA
Walton H 1973 *The Study and Analysis of Black Politics*. Scarecrow Press, Metuchen, NJ
Walton H 1985 *Invisible Politics*. State University of New York Press, Albany, NY
Whitby K J 1997 *The Color of Representation: Congressional Behavior and Black Interests*. University of Michigan Press, Ann Arbor, MI
Wrinkle R D, Stewart J Jr, Polinard L, Meier K J, Arvizu J R 1996 Ethnicity and nonelectoral political participation. *Hispanic Journal of Behavioral Sciences* **18**: 142–53

K. Tate

Racial Relations

The term *racial relations* denotes the relationships between groups socially defined as 'races.' In common usage this term is ambiguous and is taken by many popular analysts to mean that racial relations are not primarily matters of power and oppression, but rather mostly involve relationships between groups with relatively equal resources. However, in many cases racial relations are more accurately termed *racist relations*. For example, in areas of the globe where there was significant European colonization—eventually most areas of the globe—systems of racist oppression were created by European colonists who enriched themselves substantially at the expense of indigenous peoples' land, resources, and labor. This often brutally executed enrichment became part of the economic foundation for Europe's and North America's long-term economic prosperity (Du Bois 1965 [1946]).

1. *The Idea of 'Race'*

Prior to the late eighteenth century the word 'race' was used for descendants of a common ancestor, emphasizing kinship rather than skin color or other physical characteristics. By the late eighteenth century the word had evolved into its modern sense of a category of human beings with distinctive physical characteristics transmitted by descent. In the 1770s Immanuel Kant was the first prominent scholar to use the term 'races of mankind' (in German) in the sense of physically and biologically distinct categories. Soon thereafter, the German anatomist Johann Blumenbach aid out a ladder of races that became influential (Hannaford 1996, pp. 205–7). At the top were the 'Caucasians,' a term Blumenbach coined for Europeans. Groups lower on the ladder were, in order, the

Mongolians (Asians), the Ethiopians (Africans), the Americans (Native Americans), and the Malays (Polynesians).

In the late eighteenth and early nineteenth centuries Kant, Blumenbach, and other leading European intellectuals increasingly proclaimed the virtues and privileges of whiteness. Educated Europeans and European Americans crystallized the distinction between the superior 'white race' and 'inferior races' and disseminated these views to the general public. This social construction of races (see *Conflict Sociology*) developed as Europeans expanded their colonial adventures overseas. Colonizers developed racist rationalizations for their destruction of indigenous societies, enslavement of Africans, and other colonial pursuits. This congealed in an *ideological racism* that legitimated oppression and exploitation, especially as the racist ideas became integrated into the general ideology dominant in Western societies (Balibar and Wallerstein 1991).

2. Racial and Ethnic Groups

Taking into account this history of racial relations and racial oppression, most social scientists define a *racial group* as a social group that people inside or outside that group have decided is important to single out as inferior or superior, typically on the basis of physical characteristics subjectively selected. For marking off racial groups certain characteristics like skin color are chosen, and others like hair color are rejected, solely for social reasons. Similarly, an *ethnic group* is a group that is set apart by insiders or outsiders primarily on the basis of cultural or national-origin characteristics subjectively selected (Feagin and Feagin 1999).

Racial and enthnic groups are not fixed groups or essences that endure forever but are temporary inventions shaped in sociopolitical struggles at historical times and places. They have significance only in group interaction. While the two types of groups can be distinguished on theoretical grounds, in the everyday operation of societies, racial and ethnic markers, identities, and oppressions often intertwine.

3. Conceptual Approaches to Racial Relations

Most conceptual approaches to intergroup relations, including racial relations, can be roughly classified as (a) order-assimilation theories or (b) power-conflict theories. Order-assimilation theories tend to accent the gradual incorporation of immigrant groups into a dominant culture and stability over time in intergroup relations. In contrast, power-conflict theories give central attention to the outcomes of racial contacts that result in oppression, labor exploitation, and genocide.

3.1 Assimilation Approaches

In Europe and the United States much theorizing about racial-ethnic relations has placed an emphasis on assimilation, on the more or less orderly adaptation of a migrating group such as Italian immigrants to the ways of an established host group, such as an Anglo-Protestant core culture in the United States. Related theories accent individual or group competition seen as resolving itself in the direction of eventual accommodation (see, for example, Banton 1983). In this order-assimilation theorizing little attention is paid to outcomes of intergroup contact taking the form of long-term oppression. One early theorist, Robert E. Park (1950), argued that intergroup relations move through stages of a *race relations cycle*—contacts. competition, accommodation, and eventual assimilation—that are more or less progressive and irreversible. Scholars in this tradition argue that there is a long-term trend toward orderly adaptation of immigrant groups in Western societies and that racial-ethnic conflict will diminish with the eventual assimilation of these incoming groups.

Giving primary attention to immigrants to the United States, Milton Gordon (1964) views the typical trend of adaptation to be in the direction of immigrants giving up much of their heritage for the pre-existing Anglo-Protestant core culture. This adaptation takes place in seven different arenas of assimilation: cultural assimilation, structural (primary group) assimilation, marital assimilation, attitude-receptional assimilation (breakdown of prejudices), behavior-receptional assimilation (decline of discrimination), identificational assimilation, and civic assimilation. Classical assimilation theorists have had in mind as their main examples of racial-ethnic adaptation the European groups migrating more or less voluntarily to the United States or European nations. Until relatively recently, theorists who accent assimilation gave less attention to the adaptation and assimilation of non-European groups to the core culture. When they did, they often viewed non-Europeans as assimilating like other groups. One prominent European writer on US racial relations, Gunnar Myrdal (1964 [1944], Vol. 2, p. 929), argued that as a practical matter it is to the advantage of black Americans 'as individuals and as a group to become assimilated into American culture, to acquire the traits held in esteem by the dominant white Americans.' Even today, many assimilation analysts expect the importance of racial-ethnic stratification to decline as the powerful universalistic forces of industrialization and urbanization wipe out the vestiges of earlier particularistic value systems in various societies.

3.2 Power-conflict Approaches

Generally speaking, assimilation theorists are most concerned with cultural (such as language) or primary-

group network (such as intermarriage) features of the adaptation of migrating groups to a host society, while power-conflict theorists are most concerned with racial-ethnic oppression, exploitation, and subordination. An early US advocate of this latter perspective, Oliver C. Cox (1948), emphasized the role of the capitalist class in creating racial conflict by exploiting the labor of groups like African Americans and by propagating racist ideologies to divide the working class in Western societies. Racial (white-on-black) stratification grew out of the expansion of profit-oriented capitalism.

Some scholars have adopted the concept of *internal colonialism* to describe what happens to many non-European groups coming into Western societies. An emphasis on great power inequalities between whites and people of color is at the heart of this model. Thus, people of color, such as Native Americans and African Americans, were brought into the emerging political-economic system of North America by force. Subsequent adaptation was characterized by dominant group coercion and oppression, as well as by subordinate group resistance. Racial stratification has a material base, a firm underpinning rooted in a specific political-economic history. Internal-colonialism theorists are not primarily concerned with white immigrant groups but rather with the persistence of systemic oppression and the control processes that maintain white dominance (Blauner 1972).

Power-conflict analysts look at the immigration of non-European peoples as involving much more than assimilation. John Rex (1973) has analyzed the immigration of such groups from the former European colonies to the metropolitan cities of the colonizing countries. In the latter countries the colonial immigrants get the most inferior and marginal jobs and suffer the highest unemployment—thereby becoming structurally distinct from the native-born, white working class and, thus, an outcast group of alien workers.

The concept of *institutional racism* accents the importance of patterns of racism built into a society's major institutions and patterns. Racism is much more than whites' racist attitudes, for it encompasses a complex array of inegalitarian relationships developed over generations and imbedded in all institutions. In racist systems most whites become stakeholders in a hierarchical structure of opportunities, wealth, and privileges linked to a long history of racial exploitation and oppression. Joe R. Feagin (2000) has shown that the concept of *systemic racism* encompasses not only the construction of racist ideologies but also the maintenance of white privilege, wealth, and power. People do not experience racial relations in the abstract but in concrete and recurring social situations. Individuals, whether they are the perpetrators or recipients of discrimination, are caught in a complex web of alienating racist-relations. A racist system categorizes and divides human beings from each other and severely impedes the development of a common consciousness. At the macrolevel, large-scale institutions routinely perpetuate racial subordination and inequalities. These institutions are created and recreated by the routine discriminatory practices of dominant group individuals at the microlevel of everyday interaction. Individual acts of discrimination regularly activate whites' group power position and authority, a point underscored by the European scholar (Philomena Essed 1991).

In much theorizing about contemporary racial conflict, there is an accent on class or on socioeconomic status. Some analysts argue that systemic racism is of declining importance in Western societies like the United States and is more or less consigned to the fringes, while class divisions are now increasing (Wilson 1978). Yet the evidence (see below) suggests that such a conclusion is premature. Critics (for example, Omi and Winant 1994) of this view of the declining significance of racism have shown that, for the United States at least, race cannot be reduced to class or socioeconomic status, but remains an autonomous field of intergroup conflict, political organization, and ideological meaning.

4. *Continuing Racial Oppression: The US Case*

Today, people of African descent are still the globe's largest racially oppressed group; they are now resident in a number of countries as a result of the overseas slave trade carried out by Europeans. One of the largest groups in this African diaspora is in the United States, where there remains much evidence of continuing white-on-black oppression.

Recent field studies and opinion surveys show that among white Americans racist attitudes remain widespread. In one survey three-quarters of white respondents openly agreed with one of more of the eight antiblack stereotypes presented to them, while just over half agreed with two or more (Anti-Defamation League 1993). Discrimination against African Americans and other Americans of color also remains commonplace. One Los Angeles survey (Bobo and Suh 1995) found that within the past year six in ten black workers had encountered workplace discrimination; that proportion increased with workers' education. This study also found that a majority of highly educated Asian and Latino American workers faced workplace discrimination. More generally, researchers in the United States have found that, after decades of affirmative action programs attempting to address discrimination in employment, at least 95 percent of the top positions in major economic, political, and educational organizations are still held by white men. In the United States housing patterns are still highly segregated, largely as a result of active discrimination by white landlords, homeowners, and real estate

agents. In a national survey using 3,800 test audits in 25 metropolitan areas, black renters and homebuyers were found to face discrimination about half the time in seeking housing (Turner et al. 1991).

In the United States, as in other Western nations, whites also direct much racist hostility and discrimination at other groups of color. A number of research studies have shown that many Asian Americans are moving into the mainstream of the US economy and suburban neighborhoods, yet even as they do so they still encounter much racial discrimination. White-collar workers, including administrators and professionals in historically white institutions, report ingrained patterns of anti-Asian discrimination (Cheng 1997). Similarly, Latino Americans also report much continuing discrimination in employment and housing. Political discrimination also targets Latino and Asian Americans. During the 1990s California voters, especially white voters, supported several ballot propositions designed to abolish government services for Latino and Asian immigrants and eliminate affirmative action and bilingual programs. In addition, hostility to Spanish speaking Americans persists in widespread mocking of the Spanish language and Latino cultures.

5. *Racism in Europe: The Case of France*

The United States is not the only Western nation pervaded by racist relations. The reality of racial oppression can be seen in the outbreak of hate crimes across the European continent since the 1980s, as well as in routine discrimination in housing and employment in many European countries. For example, in France immigrants from Algeria and other African countries often face racist practices much like those in the United States. France imposed colonial rule on Algeria in the nineteenth century; later, beginning in the 1960s, large numbers of Algerians migrated to France. Defining immigrant groups as undesirable, whites there publicly accented the lack of assimilation by African immigrants into French culture as the major issue, playing down or ignoring the way in which such realities as housing segregation result from institutionalized, race-linked discrimination against Africans. In France racist-right political parties have grown in influence in recent years, and the national government has responded periodically with repressive attacks on African immigrants (Batur-VanderLippe 1999).

For numerous European countries, overseas colonialism involved the use of violence to exploit the resources of the colonies and oppress their peoples —practices generally legitimated with racist ideologies. More recently, the descendants of the formerly colonized face racist oppression in the colonizers' home countries. In a number of European countries the legacy of colonialism can be seen graphically in the often violent expressions of racialized hatred for the identity and culture of African, Asian, and West Indian immigrants.

6. *Racism in Former Empires: The South African Case*

In many ways South Africa is the symbol of racist relations as they developed between European and non-European peoples in the ever spreading European empires. In the South African case, two colonizing powers, England and Holland, invaded the lands of indigenous African peoples, and their descendants became the economically and politically dominant group. In the first decade of the twentieth century a new country, called the Union of South Africa, was formed, combining residents of British and Dutch origin indigenais as the ruling group over a number of African groups. Black Africans were gradually forced into a system of segregation (apartheid) where the relations between the white and black groups, as well as between these groups and middle groups like the 'Asians' and the 'coloreds' (those of mixed racial backgrounds), were determined by an extensive system of segregation laws. Over time the system of apartheid came to encompass every aspect of the relations between whites and non-whites, including relations in the institutional areas of employment, education, and housing (Batur-VanderLippe 1999, Feagin and Feagin 1999).

As in the case of African Americans, this intensive system of racial oppression generated resistance in the form of major organizations like the African National Congress (ANC). In the 1990s the black African majority finally took back political control. The ANC, working with other anti-apartheid groups, eventually forced an end to apartheid and brought new hope for the oppressed majority. Racial relations dramatically changed. A formerly total-racist society, the Union of South Africa moved from white to black political control and began to change the patterns of its social and economic apartheid with relatively little bloodshed. Still this change has so far marked only a beginning, for the political changes have not yet ended economic domination by whites, and racist practices are still found in housing and employment.

Late twentieth century events in South Africa have linked black peoples around the globe in a common struggle against the descendants of European colonists, who long maintained economic or political control in a number of nations where people of color are the numerical majority. Since the early 1900s black American efforts for liberation from racism (such as the recurring civil rights movements) have provided inspiration for the struggle in South Africa and other nations (for example, Brazil), and in a similar fashion the recent political successes by black Africans in South Africa have encouraged African Americans in

their struggle against oppression in the United States. Black Americans have long drawn on the history, spirituality, and symbolism of Africa as part of their individual and group strategies for coping with the white-dominated society. This is yet another example of the way in which racial relations in one nation are set within a larger global context.

7. *The Future of the Global Racial Order*

In many countries across the globe challenges to white domination and racism are arising from population and sociopolitical changes. Europeans and European Americans are a decreasing fifth of the world's population and a decreasing proportion of US population. In the United States, at the beginning of the twenty-first century, whites are a statistical minority in many of the largest cities and in the states of New Mexico and Hawaii. By the early 2000s whites will be a minority in the states of California and Texas, with other major states to follow in the next few decades. Demographic projections indicate that shortly after the year 2050 white Americans will become a minority of the US population.

As whites become the minority in the United States and in many international organizations where they once dominated, the traditional patterns of racial relations will be under even greater pressure to change. For the US case, by the 2050s, the majority of the labor force will no longer be white, and the population and labor force will be older. This will likely give people of color ever more political and economic clout. Tensions will likely increase between a retired population that is mostly white and workers who are mostly not white. In several Western nations, white politicians opposed to legal immigration by non-Europeans and to remedial programs for racial discrimination will not likely be elected when the majority of their political constituencies become citizens of color. As constituencies change, juries and justice systems, educational systems, and government policies will also be likely to change.

Across the globe the relations of colonialism and racism are slowly but dramatically changing. The deep-lying contradictions of the global racial order were set into motion by the colonial adventures of numerous European nations. This colonialism generated sociopolitical structures that have long embedded the practices, ideologies, and institutions of racist subordination in numerous nations. Today, international relations, global markets, the global mass media, and multinational corporations are all racialized, with white European perspectives and executives often in control. Half a millennium of European-dominated national and international institutions is slowly coming to an end. As the twenty-first century develops, racial relations in Europe, the United States, and the rest of the world are likely to change. Many once oppressed groups and societies will move farther out from under the dominance of Europeans and European Americans. Over the last few decades strong antiracism movements have arisen in South Africa, Brazil, and the United States. The possibility of a global democratic order rid of the oppressive relations of racism remains only a dream, but the South African revolution shows that it is a powerful dream for many people.

See also: Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Colonization and Colonialism, History of; Conflict Sociology; Discrimination: Racial; Environmental Justice; Ethnic Conflicts; Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity, Sociology of; Hate Crimes; Policing; Race and Gender Intersections; Race and the Law; Race and Urban Planning; Race: History of the Concept; Race Identity; Race Relations in the United States, Politics of; Racism, History of; Racism, Sociology of; Residential Segregation: Sociological Aspects; Slavery as Social Institution; Slaves/Slavery, History of

Bibliography

Anti-Defamation League 1993 *Highlights from an Anti-Defamation League Survey on Racial Attitudes in America*. Anti-Defamation League, New York

Balibar E, Wallerstein I 1991 *Race, Nation, Class: Ambiguous Identities*. Verso, London

Banton M P 1983 *Racial and Ethnic Competition*. Cambridge University Press, Cambridge, UK

Batur-VanderLippe P 1999 Colonialism and post-colonialism: the global expansion of racism. In: Feagin J R, Feagin C B (eds.) *Racial and Ethnic Relations*. Prentice-Hall, Upper Saddle River, NJ

Blauner R 1972 *Racial Oppression in America*. Harper & Row, New York

Bobo L, Suh S A 1995 Surveying Racial Discrimination: Analyses from a Multiethnic Labor Market. Research Report. Department of Sociology, University of California, Los Angeles

Cheng C 1997 Are Asian American employees a model minority or just a minority? *Journal of Applied Behavioral Science* **33**: 277–90

Cox O C 1948 *Caste, Class, and Race*. Doubleday, Garden City, NY

Du Bois W E B 1965 [1946] *The World and Africa*. International Publishers, New York

Essed P 1991 *Understanding Everyday Racism*. Sage, Newbury Park, CA

Feagin J R 2000 *Racist America: Old Roots, Current Realities, and Future Reparations*. Routledge, New York

Feagin J R, Feagin C B 1999 *Racial and Ethnic Relations,* 6th edn., Prentice-Hall, Upper Saddle River, NJ

Gordon M M 1964 *Assimilation in American Life*. Oxford University Press, New York

Hannaford I 1996 *Race: The History of an Idea in the West*. Johns Hopkins University Press, Baltimore, MD

Myrdal G 1964 [1944] *An American Dilemma*. McGraw-Hill, New York
Omi M, Winant H 1994 *Racial Formation in the United States: From the 1960s to the 1990s,* 2nd edn. Routledge, New York
Park R E 1950 *Race and Culture*. Free Press, Glencoe, IL
Rex J 1973 *Race, Colonialism and the City*. Routledge and Kegan Paul, London
Turner M A, Struyk R J, Yinger J 1991 *Housing Discrimination Study: Synthesis*. US Dept. of Housing and Urban Development, office of Policy Development and Research, Washington, DC
Wilson W J 1978 *The Declining Significance of Race: Blacks and Changing American Institutions*. University of Chicago Press, Chicago, IL

J. R. Feagin

Racism, History of

In common usage the term racism is often employed in a loose and unreflective way to describe the hostile or negative feelings of one ethnic group or historical collectivity toward another and the actions resulting from such attitudes. But historians and social scientists normally require a more precise and limited definition, although they may disagree on what it should be. It is useful, therefore, to review the historiographic discourse on the meaning of racism before attempting a brief survey of what contemporary historians might accept as its principal manifestations.

1. *Historiography of Racism*

The term racism was coined in the 1930s mainly in reference to the theories on which the Nazis based their persecution of the Jews. The first historian to focus directly on the subject was Barzun. His book *Race: A Study in Superstition* (originally published in 1937 and reprinted in 1965) set two important precedents for most future historians of racism: it presumed that claims of the innate inferiority of one 'race' to another were false or at least unproven, and its main concern was with the history of ideas rather than with the social and political applications of prejudiced beliefs and attitudes. Barzun's concern with the intellectual origins of Nazi antisemitism was shared by the many scholars who wrote on the subject after World War II in response to what the Holocaust had revealed about the horrifying consequences of racist ideas. A notable example was Poliakov's *Le Myth Aryen*, translated into English as *The Aryan Myth* (1974). But disagreements developed on whether biological racism was continuous with earlier antisemitic attitudes that were, ostensibly at least, based on religion or whether it was a radical new departure, a *sine qua non* for something like the Holocaust.

In the immediate postwar decades, the term racism was also applied with increasing frequency to the relationships between whites or Europeans and people of color, especially Africans or those of African descent, in colonial, postcolonial, and former slave societies. The rising concern with black–white relations in the US, South Africa, and (to a somewhat lesser extent) in Great Britain, beginning in the 1950s and early 1960s, gave rise to an enormous literature on the origins and development of white supremacy as an ideology. Historians of ideas did most of the early work, but gradually the emphasis shifted somewhat from the racist ideas themselves to the patterns or racial discrimination and domination for which they served as a rationale. Many historians of black–white relations used the term casually and without reflecting on its exact meaning. At its most imprecise, it could mean anything that whites did or thought that worked to the disadvantage of blacks.

There have been relatively few attempts to write general histories of racism, which encompass both the antisemitic and color-coded varieties. For the most part, the two historiographies have not addressed each other, and the definitions employed have sometimes been incompatible. Some historians of antisemitic racism have found the term applicable only if the aim was the elimination of the stigmatized group. Some historians of white supremacy, on the contrary, would limit the term to ideologies associated with patterns of domination or hierarchical differentiation. The first attempt at a comprehensive history of American racism, Gossett's *Race: The History of an Idea in America* (1963) traced consciousness of race to the ancient world. Its treatment of specifically American manifestations cast its net quite wide to include representation and treatment of American Indians and immigrants from southern and eastern Europe, including Jews, as well as blacks. But most subsequent work on the history of American racism has been group specific and has concentrated most heavily on attitudes toward African Americans. Mosse's general history of European racism, on the other hand, pays most of its attention to the growth of antisemitism between the Enlightenment and the Holocaust (Mosse 1978). But there appear to be only two significant attempts to cover Western attitudes toward race comprehensively: Hannaford's *Race: The History of an Idea in the West* (1996) and Geiss' *Geschichte des Rassismus* (1988). Hannaford's study, as its title indicates, is strictly an intellectual history and considers race as a concept more than racism as an ideology. It argues strenuously that no clear concept of race existed before the seventeenth century, thus raising the issue of whether anything that existed before the invention of race in the modern sense can be legitimately labeled racism. Geiss, on the contrary, sees racism as anticipated in most respects by the ethnocentrism or xenophobia that developed in the ancient world, as reflected, for example, in the Old Testament.

2. Toward a Definition of Racism for Historical Purposes

Somewhere between the view that race is a peculiar modern idea without much historical precedent and the notion that it is simply an extension of the ancient phenomena of ethnocentrism and xenophobia may lie a working definition that is neither too broad for historical specificity or too narrow to cover more than the limited span of Western history during which a racism based on scientific theories of human variation was widely accepted. If racism is defined as an ideology rather than as a theory, links can be established between belief and practice that the history of ideas may obscure. But ideologies have content, and it is necessary to distinguish racist ideologies from other belief systems that emphasize human differences and can be used as rationalizations of inequality. The classic sociological distinction between racism and ethnocentrism is helpful, but not perhaps in the usual sense, in which the key variable is whether differences are described in cultural or physical terms. It is actually quite difficult in specific historical cases to say whether appearance or 'culture' is the source of the salient differences, because culture can be reified and essentialized to the point where it has the same deterministic effect as skin color. But it would stretch the concept of racism much too far to make it cover the pride and loyalty that may result from ethnic identity. Such group-centeredness may engender prejudice and discrimination against those outside the group, but two additional elements would seem to be required before the categorization of racism is justified. One is a belief that the differences between the ethnic groups involved are permanent and ineradicable. If conversion or assimilation is genuinely on offer, we have religious or cultural intolerance but not racism. The second is the social side of the ideology—its linkage to patterns of domination or exclusion. To attempt a short formulation, we might say that racism exists when one ethnic group or historical collectivity dominates, excludes, or seeks to eliminate another on the basis of differences that it believes are hereditary and unalterable.

3. The Emergence of Racism

Historians who can accept this middle-of-the-road definition can agree with the scholars who have examined conceptions of human diversity in classical antiquity and the early Middle Ages without finding clear evidence of color prejudice or racism. Neither the classical conception that one becomes civilized through being able to engage in politics nor the Christian belief that everyone regardless of color or ancestry is a potential convert with a soul to be saved could readily sustain a racist view of the world. The period between the fourteenth and the seventeenth centuries is much more problematic. The literal demonization of the Jews in the late middle ages, the belief that they were in league with the devil and plotting the destruction of Christianity, could lead unsophisticated folk to believe that they were outside the bounds of humanity, even though church authorities persisted in advocating their conversion. Folk antisemitism resulted in massacres and expulsions of Jews in virtually every country in Western Europe. A degree of elite or official sanction for such attitudes came in sixteenth century Spain when Jews who had converted to Christianity and their descendants became the victims of a pattern of discrimination and exclusion. When purity of blood, or *limpieza de sangre*, became a qualification for certain offices and honors it signified that what people could accomplish or achieve no longer depended on what they did but on who they were. Precedents for the notion that blood will tell could be found in the inheritability of royal or noble status but the designation of an entire ethnic group as incapable of a reliable conversion to Christianity constituted, at the very least, a transitional stage between religious bigotry and biological racism.

The period of the Renaissance and Reformation was also the time when Europeans were coming into increasing contact with people of darker pigmentation in Africa, Asia, and the Americas and were making judgments about them as they traded with them, fought with them over territory, or enslaved them. A belief that all these non-Europeans were children of God with souls that missionaries could save did not prevent a brutal expropriation of their land and labor, but it did inhibit the articulation of a racial justification for how they were being treated. When Africans were carried by force to the New World to labor on plantations, slave traders and slave owners sometimes invoked an obscure passage in the book of Genesis to explain the color of their human property. Ham had committed a sin against his father Noah that resulted in a divine curse on his descendents to be 'servants unto servants.' When Africans were enslaved, first by Arabs, and then by Europeans, it became convenient to believe that the punishment of God had included a blackening of the skin. But the passage had other possible meanings, and no orthodox religious authorities ever endorsed the interpretation of the curse popular among slaveholders, because they feared that it would impede conversion of blacks to Christianity. The standard and official justification for enslaving Africans was that they were heathens and that enslavement made heaven accessible to them.

An implicit racism became evident, however, when converted slaves were kept in bondage, not because they were actual heathens but because they had heathen ancestry. As with the doctrine of purity of blood in Spain, descent rather than performance became the basis for determining the qualifications for membership in a community that was still theoretically based on a shared Christian faith. Beginning in the late

seventeenth century laws were passed in English North America forbidding marriage between whites and blacks and discriminating against the mixed offspring of informal liaisons. Without clearly saying so, such laws implied that blacks were unalterably alien and inferior.

4. The Rise of Modern Racism

A modern naturalistic form of racism could not be articulated fully until the Enlightenment. A secular or scientific theory of race required a new way of looking at human diversity, one that moved the subject away from the Bible, with its insistence on the essential unity and homogeneity of the human race and its collective elevation above the animal kingdom. Eighteenth-century ethnologists began to think of human beings as part of the natural world and subdivided them into three to five races, usually considered as varieties or subspecies. Most eighteenth-century ethnological theorists believed that whiteness was the original color of humanity and that blackness or brownness resulted from a process of degeneration caused by climate and conditions of life. Although the white or Caucasian race was generally considered more beautiful and civilized than the Mongolian and especially the Negro, the belief that human beings constituted a single species with a common origin remained dominant. In the late eighteenth and early nineteenth centuries, however, an increasing number of ethnologists asserted that the races were separately created and constituted distinct species. As debates over the morality of slavery and the slave trade erupted in Western Europe and the US in the wake of democratic revolutions and the rise of humanitarian movements, some of the most militant defenders of black bondage adopted these polygenetic arguments as a justification for enslaving Africans and holding them in permanent servitude. But the fact that polygenesis seemed to conflict with scripture limited the usage of this extreme version of scientific racism. Many defenders of slavery in the American South, where evangelical religion enforced biblical literalism, reverted to the Curse on Ham or made the degeneracy hypothesis of Enlightenment ethnology into an argument for the permanent or irreversible divergence of the races from their common origin.

The nineteenth century was an age of emancipation, nationalism, and imperialism—all of which contributed to the growth and intensification of ideological racism in Europe and in the US. Although the emancipation of blacks from slavery and Jews from the ghettoes, which seemed a logical consequence of the democratic revolutions of the late eighteenth century, received most of its support from religious or secular believers in an essential human equality, the consequences of these reforms was to intensify rather than diminish racism. Race relations became less paternalistic and more competitive. Poor or working class whites in the US, for example, feared competition with emancipated blacks for work, land, and social status. Middle class (especially lower middle class) Germans viewed the increasing involvement of Jews in finance, commerce, the professions, journalism, and the arts during the late nineteenth century as a potential threat to their own economic security, social status, and traditional values. Of course, it was the legacy of inherited stereotypes about 'the Other,' in addition to the immediate economic and social circumstances, that gave edge and intensity to these feelings of insecurity and hostility. The idiom of Darwinism, with its emphasis on 'the struggle for existence' and concern for 'the survival of the fittest' was conducive to the development of a new and more credible scientific racism in an era that increasingly viewed race relations as an arena for conflict rather than as a stable hierarchy.

The growth of nationalism, especially romantic cultural nationalism, encouraged the growth of a more mystical variant of racist thought. When it could be assumed that one had to be of German ancestry to participate in the German *Volkgeist*, Jews could be discriminated against or excluded just as effectively as if their deficiencies were clearly defined as genetic or biological. For most of the nineteenth century the question of whether German Jews could or should be converted to Christianity was an open one. Beginning in the late 1870s and early 1880s, however, the coiners of the term 'antisemetism' made explicit what some cultural nationalists had previously implied—that to be Jewish was not simply to adhere to a set of religious beliefs or cultural practices but meant belonging to a race that was the antithesis of the race to which true Germans belonged. The latter was variously designated as Teutonic, Aryan, Nordic, or simply Germanic. In Great Britain and the US romantic nationalism took the form of a veneration for Anglo-Saxon ancestors and legacies that could turn racist, as it did for a time in the US, when confronted with large-scale immigration from eastern and southern Europe. In France, an ethnocentric cultural nationalism fed the fury against the Jews that surfaced in the Dreyfus affair at the turn of century. It did not triumph against the universalistic civic nationalism inherited from the revolutionary past, but it did make antisemitic racism a strain of right-wing French thought that would resurface during the 1930s and especially under the German occupation.

The climax of Western imperialism in the late nineteenth century 'scramble for Africa' and parts of Asia and the Pacific represented an assertion of the competitive ethnic nationalism that existed among European nations (and which, as a result of the Spanish–American War came to include the US). A belief that different races, subraces, or racial mixtures inhabited each nation led some, especially the Germans and the English, to suppose that the quality of each nation's racial stock was being tested for its

fitness. But imperialism also required the massive subjugation and colonial domination of non-European populations. An ideology proclaiming that whites were superior to 'lesser breeds' and responsible for ruling over them and tutoring them in the rudiments of civilization was a prime rationale for this new burst of European expansionism, even if it was not a cause of it. The relation of racism to imperialism is more problematic than one might suppose, however, because the most consistent and extreme racists were often anti-imperialists. They believed that nothing useful could be done with people whose inferiority was so profound and permanent that they were incapable of being civilized. For them the taming or domestication of such savages was not worth the trouble. Within colonies that attracted substantial numbers of European settlers, such as French Algeria or some of the British colonies of South and East Africa, a more extreme and explicit racism than the one usually professed by government officials or missionaries could be found in the discourse of white colonists about their 'native' servants or farm laborers.

5. *The Climax: Racism in the Twentieth Century*

The climax of the history of racism came in the twentieth century in the rise and fall of what might be called overtly racist regimes. In the American South, the passage of racial segregation laws and restrictions on black voting rights reduced African Americans to lower caste status, despite the Constitutional Amendments that had made them equal citizens. Extreme racist propaganda, which represented black males as ravening beasts lusting after white women, served to rationalize the practice of lynching. These extra-legal executions increasingly were reserved for blacks accused of offenses against the color line, and they became more brutal and sadistic as time went on; by the early twentieth century victims were likely to be tortured to death rather than simply being killed. A key feature of the racist regime maintained by state law in the South was a fear of sexual contamination through rape or intermarriage, which led to efforts to prevent the conjugal union of whites with those with any known or discernable African ancestry. The effort to guarantee 'race purity' in the American South anticipated aspects of the official Nazi persecution of Jews in the 1930s. The Nuremberg laws of 1935 prohibited intermarriage or sexual relations between Jews and Gentiles, and the propaganda surrounding the legislation emphasized the sexual threat that predatory Jewish males presented to German womanhood and the purity of German blood. Racist ideology was eventually of course carried to a more extreme point in Nazi Germany than in the American South of the Jim Crow era. Individual blacks had been hanged or burned to death by the lynch mobs to serve as examples to insure that the mass of southern African Americans would scrupulously respect the color line. But it took Hitler and the Nazis to attempt the extermination of an entire ethnic group on the basis of a racist ideology.

Hitler, it has been said, gave racism a bad name. The moral revulsion of people throughout the world against what the Nazis did, reinforced by scientific studies undermining racist genetics (or eugenics), served to discredit the scientific racism that had been respectable and influential in the US and Europe before World War II. But explicit racism also came under devastating attack from the new nations resulting from the decolonization of Africa and Asia and their representatives in the United Nations. The Civil Rights movement in the US, which succeeded in outlawing legalized racial segregation and discrimination in the 1960s, was the beneficiary of revulsion against the Holocaust as a logical outcome of racism. But it also drew crucial support from the growing sense that national interests were threatened when blacks in the US were mistreated and abused. In the competition with the Soviet Union for 'the hearts and minds' of independent Africans and Asians, Jim Crow and the ideology that sustained it became a national embarrassment with possible strategic consequences.

The one overtly racist regime that survived World War II and the Cold War was the South African, which did not in fact come to fruition until the advent of Apartheid in 1948. The laws passed banning all marriage and sexual relations between different 'population groups' and requiring separate residential areas for people of mixed race ('Coloureds'), as well as for Africans, signified the same obsession with 'race purity' that characterized the other racist regimes. However, the climate of world opinion in the wake of the Holocaust induced apologists for apartheid to avoid straightforward biological racism and rest their case for 'separate development' mainly on cultural rather than physical differences. The extent to which Afrikaner nationalism was inspired by nineteenth-century European cultural nationalism also contributed to this avoidance of a pseudo-scientific rationale. No better example can be found of how a 'cultural essentialism' based on nationality can do the work of a racism based squarely on skin color or other physical characteristics. The South African government also tried to accommodate itself to the age of decolonization. It offered a dubious independence to the overcrowded 'homelands,' from which African migrants went forth to work for limited periods in the mines and factories of the nine-tenths of the country reserved for a white minority that constituted less than a sixth of the total population.

6. *The End of Racism?*

The defeat of Nazi Germany, the desegregation of the American South in the 1960s, and the establishment of

majority rule in South Africa suggest that regimes based on biological racism or its cultural essentialist equivalent are a thing of the past. But racism does not require the full and explicit support of the state and the law. Nor does it require an ideology centered on the concept of biological inequality. Discrimination by institutions and individuals against those perceived as racially different can long persist and even flourish under the illusion of nonracism, as historians of Brazil have discovered recently. The use of allegedly deep-seated cultural differences as a justification for hostility and discrimination against newcomers from the Third World in several European countries has led to allegations of a new 'cultural racism.' Similarly, those sympathetic to the plight of poor African Americans and Latinos in the US have described as 'racist' the view of some whites that many denizens of the ghettos and barrios can be written off as incurably infected by cultural pathologies. From the historian's perspective such recent examples of cultural determinism are not in fact unprecedented. They rather represent a reversion to the way that the differences between ethnoracial groups could be made to seem indelible and unbridgeable before the articulation of a scientific or naturalistic conception of race in the eighteenth century.

See also: Anti-Semitism; Discrimination; Discrimination, Economics of; Discrimination: Racial; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psychology of; Ethnicity: Anthropological Aspects; Holocaust, The; Prejudice in Society; Race and the Law; Race: History of the Concept; Race Identity; Race Relations in the United States, Politics of; Racial Relations; Racism, Sociology of

Bibliography

Banton M 1998 *Racial Theories*. Cambridge University Press, Cambridge, UK

Barzun J 1965/1937 *Race: A Study in Superstition*. Harper and Row, New York

Fredrickson G M 1981 *White Supremacy: A Comparative Study in American and South African History*. Oxford University Press, New York

Geiss I 1988 *Geschichte des Rassismus*. Suhrkamp, Frankfurt, Germany

Gossett T F 1963 *Race: The History of an Idea in America*. Southern Methodist University Press, Dallas, TX

Hannaford I 1996 *Race: The History of an Idea in the West*. Johns Hopkins University Press, Baltimore, MD

Langmuir G I 1990 *Toward a Definition of Antisemitism*. University of California Press, Berkeley, CA

Lauren P L 1988 *Power and Prejudice: The Politics and Diplomacy of Racial Discrimination*. Westview Press, Boulder, CO

Mosse G L 1978 *Toward the Final Solution: A History of European Racism*. H Fertig, New York

Pieterse J N 1992 *White on Black: Images of Africa and Blacks in Western Popular Culture*. Yale University Press, New Haven, CT

Poliakov L 1974 *The Aryan Myth: A History of Racist and Nationalist Ideas in Europe*. Chatto for Essex University Press, London

Smedley A 1993 *Race in North America: Origin and Evolution of a World View*. Westview Press, Boulder, CO

Taguieff P 1997 *Le Racisme*. Flammarion, Paris

G. M. Fredrickson

Racism, Sociology of

Few concepts in social science have generated more heat and less light than 'race' and its derivative 'racism,' the former because of its multiplicity of meanings, and the latter because of its dual career as a concept and as an ideological bludgeon.

1. Definitions

'Race' has been used with four distinct meanings:

(a) As a synonym for a biological species, as in 'the human race.'

(b) As a synonym for a biological subspecies as in 'the Caucasoid race.'

(c) As a synonym for a nation or ethnic group, as in 'the French race.'

(d) As a term to designate a group of diverse origins thought to be a biological subspecies in a given society, as in 'the Jewish race' in Nazi Germany. To reduce the confusion and distinguish this usage from the second, the term 'social race' has sometimes been used, but generally the qualifier is omitted.

The first usage is largely metaphorical and, thus, unproblematic. However, the other three, are hopelessly confusing, either because they have no clear empirical referent (as in the second), or because they overlap empirically with one another (as in the second, third, and fourth), or both. The result is a conceptual muddle that has plagued the social science literature since the turn of the twentieth century.

'Racism,' i.e., the ideology or attitudes associated with one or more of the listed definitions of 'race,' is in an even more parlous state of conceptual confusion. The following list includes some but not all of the current usages:

(a) As an ideology or set of attitudes accompanying the second meaning of 'race,' that is, a belief in the relative inferiority or superiority in moral, intellectual, physical, and other attributes of biological subspecies of *Homo sapiens*.

(b) As an ideology or set of attitudes corresponding to the third meaning of 'race,' that is, as a synonym for ethnocentrism, as in the belief that one's own ethnic group is superior to others.

(c) As an ideology or set of attitudes linked to the fourth meaning of 'race,' namely 'social race.'

(d) As a set of practices or institutions, divorced from individual beliefs or attitudes, that, by design or otherwise, have a differential impact on various racial or ethnic groups. This is usually labeled 'institutional racism,' and is a synonym for discrimination, witting or unwitting. Thus, an ostensibly neutral test on which some groups consistently perform better than others is said to be 'institutionally racist.'

(e) As a general term of opprobrium hurled at any person, policy, or practice one disapproves of. Examples would be the 'Zionism is racism' resolution passed by the United Nations, or accusations leveled at opponents of race-based affirmative action in the USA. In the latter case, and indeed in much contemporary usage in the USA, race-blindness is said to be racist, while race discrimination is defined as liberal, thus turning the concept on its head.

2. *Evolution in the Use of the Term 'Racism'*

Clearly, racism as a social concept has outlived whatever use it may have had until the 1960s, before its usage became broadened beyond recognition. At mid-twentieth century, a clear distinction was maintained in most social literature between attitudes and behavior, between prejudice and discrimination, and between race and ethnicity. The concept of racism was largely restricted to prejudicial 'attitudes' applied to 'social races.' Discrimination was treated as varying in partial independence of prejudice, and discrimination was seen as a possible (but not necessary) consequence of racist attitudes. Prejudicial attitudes applied to ethnic groups were termed 'ethnocentrism,' as distinct from racism. Driven by an ideological agenda of stigmatizing any beliefs or practices thought to sustain any kind of group differential in society (e.g., group differences in performance in tests, in admission to universities, in crime or incarceration rates, in school drop-out rates, in mortality or morbidity rates, in representation in various occupations, and so on), social scientists since the 1960s have gradually blurred or eliminated this double distinction between prejudice and discrimination, and between race and ethnicity. The result has been to universalize the concept of racism. Everybody can now be called a racist, including those who steadfastly refuse to make racial distinctions.

To blur the distinction between prejudice and discrimination, as the concept of 'institutional racism' does, achieves at least two aims. First, it dissociates 'racism' from individuals, and makes it an inherent property of social institutions, or indeed of entire societies. Second, it dissociates racism from consciousness and intentionality. Both individuals and institutions can be labeled 'racist' irrespective of beliefs, attitudes, or intentions. A classical liberal, for instance, who believes in philosophy of individual rights and equality of opportunity, without any consideration of race or ethnicity, can easily be labeled a racist if, for instance, she opposes 'race norming' of tests in employment or education. Thus, we find opponents on the two sides of a clear debate on whether to take race into account in social policy hurl the epithet of 'racist' at each other. 'Racism' is now void of meaning and reduced to pure invective.

The confounding of race and ethnicity has also served a variety of political agenda. The analytical use of the distinction between race and ethnicity is that racial (e.g., physical) categorization of groups leads to much more rigid, enduring, invidious, and stigmatizing social cleavages than ethnic (e.g., cultural) criteria of group membership. For example, the categorically different position of blacks in the USA compared to other minority groups, such as Hispanics or Asian Americans, can only be understood because they are overwhelmingly defined by the double stigma of racial slavery.

African-Americans were essentially treated as a stigmatized pariah group in American society. However, redefining American blacks, as merely one ethnic minority in a 'multicultural' society may, it is hoped, destigmatize their uniquely invidious position, but it does nothing to explain the stubborn resilience of their disabilities and predicaments in American society. The confounding of race and ethnicity in this cited example cuts both ways ideologically: it may on the one hand promote pride in 'black culture,' but it also underpins the racist argument of innate black inferiority. If blacks are but one ethnic group among many, how come they fail where others (notably Asian Americans) succeed?

3. *Recent Developments*

The unintended consequences of this broadening of the meaning of racism are three-fold. First, it robbed social science of a useful tool that permitted complex analysis of complicated social phenomena. Examples of such analyses are the relationship between prejudice and discrimination (one can be for instance, an unprejudiced discriminator or a prejudiced non-discriminator, depending on a social context which rewards or punishes discrimination); the difference between profoundly racist societies such as the Ante Bellum US and apartheid South Africa, and less racist ones such as Mexico or Brazil, where ethnic and class distinction often overshadow racial ones; and, the social consequence of stressing physical phenotypes such as skin color, as distinguished from cultural characteristics, such as language or religion, for the dynamics of group conflicts. Broadening the meaning of racism not only transformed a concept into an

epithet; it simultaneously impoverished social analysis. It also made the teaching of race relations in race-polarized societies a minefield of political correctness liberally sprinkled with euphemisms and impervious to rational discourse.

Second, the broadening usage of racism and the confounding of race and ethnicity have noticeably weakened the effectiveness of remedial policies by facilitating and legitimating a shift from universalistic, race-blind policies of equal opportunity and individual rights, to one of groups rights, proportional representation, quota systems, and collective grievances and guilt. Policies such as race-based affirmative action in the USA are not only ineffective but even retrogressive because they are based on bad social science. The inherent contradiction involved in race-based affirmative action is that the remedy of 'positive discrimination' uses as a selection criterion the very stigma the policy seeks to eliminate. Surely, the way to reduce the social significance of race is not to classify the population into official race groups, to use these categories in official statistics and reports, and to request individuals to identify themselves by race for the purpose of school attendance, employment, promotion, college admission, faculty recruitment, etc.

4. *Racism and Sociobiology*

Third, the extension of the concept of race and its confounding with ethnicity are congruent with the sociobiological approach to the study of race and ethnicity that has been slowly gaining ground in the social sciences over the last 20 years of the twentieth century. The motor of natural selection is differential reproduction. Organisms are selected to behave in ways that are reproductively successful. Reproductive success can be direct or indirect. An organism can have offspring, or it can help other related organisms (that share a proportion of their genes) to reproduce. In short, nepotism—the favoring of kin—is a common strategy for maximizing inclusive fitness. This clearly operates in humans as in other social species, and is particularly evident in nuclear and extended families. However, members of ethnic and racial groups also define themselves as linked by common biological descent, that is, as large extended families. Racism and ethnocentrism can thus be seen as extended forms of nepotism. Members of such groups are typically more closely related to one another on average, than to members of other groups.

Racial and ethnic groups, in short, share the fundamental property of real, or at least putative, common descent, although they only constitute diluted forms of nepotism. The distinction between racial and ethnic groups is in the criteria used to determine membership: either physical traits in the case of races, or cultural characteristics in the case of ethnics. Preference for other individuals who resemble one physically is common where certain visible traits, such as skin color, enable one to discriminate easily between groups. However, in the case of neighboring groups that closely resemble one another (such as Swedes and Norwegians, or Zulu and Swazi), cultural traits, notably language and dialect, are more useful than biological features in establishing group membership.

The choice of racial or cultural criteria of group membership is thus opportunistically dictated by which does the best job of cheaply, quickly, easily, and reliably discriminating between co-existing groups. Sometimes both racial and cultural traits are used, but purely physical traits are often unreliable because neighboring groups frequently interbreed. Ethnocentrism, therefore, is far more general a phenomenon than racism, even though racism is generally present where it 'works,' e.g., in situations of recent long-distance migration rapidly putting in close contact large numbers of phenotypically distinct populations. Military conquest, colonialism, slavery, and labor migration are examples of such situations where racism thrives. Conversely, after several generations of contact and interbreeding, racism tends to wane, unless forcibly maintained as in the USA and South Africa, by a discriminatory political and legal system.

Seen in evolutionary perspective, racism becomes a natural propensity to favor individuals to whom one assesses to be more closely biologically related to one, and thus an expression of the nepotism that drives natural selection through inclusive fitness. This general preference for 'look-alikes' operates both at the individual level (e.g., in assortative mating between members of the same group), and between groups where group differences in the distribution of certain phenotypes are pronounced enough to make reliable and easy discrimination possible.

According to this evolutionary perspective, racism is both primordial, that is rooted in the biology of nepotism, and circumstantial or instrumental, varying with the degree to which co-existing groups look alike or different. It is neither inevitable nor unchangeable, but it is likely to occur whenever physical phenotypes, especially highly viable ones, such as skin pigmentation, sharply differentiate between groups. Extensive interbreeding, however, can quickly undermine it. Racism is also responsive in both directions, to state-imposed policies that either reward, or punish it. A state determined to maintain a racial caste society can do so for centuries, as US and South African history amply demonstrate. Conversely, the state can create an environment in which racial tolerance can thrive. The persistence or waning of racism is also a function of the degree of overlap between class and race. Racism is more likely to endure when large socio-economic differences exist between racially defined groups, and to subside where those differences are reduced.

In conclusion, the double side of racism is that while, in all likelihood, it is indeed 'in our genes' as an

expression of the biology of nepotism, it takes definite environmental conditions to become expressed. In this respect, it is no different from all our genetic propensities. The phenotype is always the product of interaction between the genotype and the environment. But that environment is not only changeable; it can also be manipulated for malign or benign ends. However, the main lesson of multiracial societies is that any state policy based on race, whether purportedly malign or benign, consolidates and perpetuates racial distinctions.

See also: Colonialism: Political Aspects; Colonization and Colonialism, History of; Discrimination: Racial; Ethnic Conflicts; Ethnicity, Sociology of; Eugenics, History of; Gender, Class, Race, and Ethnicity, Social Construction of; Genotype and Phenotype; Hate Crimes; Poverty, Sociology of; Prejudice in Society; Race and Gender Intersections; Race and the Law; Race Identity; Racial Relations; Racism, History of; Social Stratification; Urban Poverty in Neighborhoods; Xenophobia

Bibliography

Allport G W 1954 *The Nature of Prejudice*. Addison-Wesley, Cambridge, MA

Banton, M 1977 *The Idea of Race*. Tavistock, London

Francis E K 1976 *Interethnic Relations*. Elsevier, New York

Hutchinson J, Smith A D (eds.) 1996 *Ethnicity*. Oxford University Press, Oxford, UK

Rex J, Mason D (eds.) 1986 *Theories of Race and Ethnic Relations*. Cambridge University Press, Cambridge, UK

Schermerhorn R A 1970 *Comparative Ethnic Relations*. Random House, New York

Sowell T 1994 *Race and Culture*. Basic Books, New York

van den Berghe P L 1981 *The Ethnic Phenomenon*. Elsevier, New York

P. L. van den Berghe

Radicalism

The word 'radicalism' is derived from the Latin 'radix' meaning 'root.' Radicalism is any stance, practical, intellectual, or both, that goes to the root of existing practices, beliefs, or values. By transposition the term has shifted from meaning things related to the root or foundation of an institution or set of arrangements, to describing the view that existing arrangements should be transformed on the basis of foundations, roots, which are either to be rediscovered and re-asserted, or to be transplanted from the ideal to the actual. It is normally associated with the left, with liberalism (see *Liberalism*) and socialism (see *Socialism*), but since the term is relative, any fundamental criticism of or assault on existing practices can reasonably be termed radical.

1. Radicalism is a Relationship Term not a Content Term

Radicalism is, like conservatism (see *Conservatism: Theory and Contemporary Political Ideology*), a relationship term not a content term. It describes not any particular set of aspirations and aversions, but a distance between what exists and what is desired, and the way in which proposed changes are justified. Its particular character is therefore dependent on the historical circumstances in which it exists. Views put forward by radicals in Communist China might be shared by conservatives in Connecticut; the radicalism of California might be the traditionalism of Catalunya. Hannah Arendt's observation that the 'most radical revolutionary will become a conservative on the day after the revolution' may thus describe not only a change of heart following the acquisition of power, but a change in the relationship between a substantive policy which remains consistent, but which moves from opposition to implementation.

2. Some General Features of Radicalism as a Form Can Be Identified

Despite the contingency of radicalism however, some general features of its form can be identified. Radicalism normally involves the detection or description of principles that are considered to lie at the root of both existing and desired arrangements or ways of thinking. This concentration on a single and logical hierarchy of cause and effect or structure and foundation, is accompanied by a goal and a program that similarly concentrates on a single policy or set of policies the pursuit of which is intended or expected to transform a broad swathe of social life. The transformative keys are as various as are the forms of radicalism: democracy (see *Democratic Theory*), the market, religious faith.

3. Attempts to Describe Radicalism as a Feature of Personality or Social Situation

Various attempts have been made to account for radicalism in terms of the psychological disposition of radicals. This generally has been done by those who do not welcome whatever project radicals may currently be pursuing. These attempts are vitiated by the manner in which a change in the political location of an actor or group can alter the descriptive appropriateness of the term. Stalinists can move from radicalism to conservatism simply by successfully seizing and ex-

ercising power. Their critical stance towards established authority comes to an end once they are themselves the embodiment of that authority.

4. *Radicalism, Conservatism, Pragmatism, and Traditionalism*

Since it goes to the roots, radicalism tends to ignore the complexity of the branches. It has, therefore, been contrasted with pragmatism, traditionalism, and conservatism. This contrast can be misleading. Pragmatism, traditionalism, and conservatism can be terms almost as contentless and historically contingent as radicalism. On the other hand, each of them operates as a matter of both historical fact and logical necessity with principles of selection, and none of them accepts unconditionally all aspects of existing arrangements and practices. Even the most stubborn conservative will have criteria for preferring Renoir to rats. Conversely, radicalism is a part of the tradition of many if not all societies, and the title of the English socialist and Christian Tawney's collection of essays, *The Radical Tradition* (1964) was only superficially paradoxical. It is, in other words, a conservative fallacy that tradition is homogeneous. Radicalism is one of its historically contingent components in most if not all societies. It is conversely a radical fallacy that radicalism seeks a transformation of all aspects of a society's life. A radical program seldom lacks foundations in some existing practices, institutions, or values, however peripheral or subordinated those may be at the time the program is advanced. And if radicals believe their own particular principles are deeply embedded in their society's traditions, they can recommend a return to roots rather than a departure from them. Franklin D. Roosevelt's remark, therefore, that the radical is someone with both feet firmly planted in the clouds, misses the point. The radical, rather, is someone who stands on their own patch of solid ground and wants the small cloud which they can see from that real but partial vantage point to encompass the whole sky.

5. *Radicalism and Reason*

The term 'radical' has been used most frequently to describe those who, whilst their criticism may be fundamental, do not reject the democratic and liberal methods of reason and persuasion in favour of insurrection or coup. Radicalism normally implies an appeal to reason, and therefore a dependence on persuasion, reasoned assent, and free individual choice. But appeals to other roots, instinct, faith, national myth are possible and have been made. Radicalism, as a matter of historical nomenclature, is more likely to be used to describe the ambitious advocacy of democracy or markets than of totalitarianism (see *Totalitarianism*) or authoritarianism. But it can describe Fascists who appeal to racial myth or religious extremists who demand that social life be determined, and law applied, on the basis of their interpretation of Christianity, Islam, or Hinduism.

6. *Radicalism and Rulers*

Were radicalism simply an abstract disposition, then it might be found to characterize equally rulers and their critics. But because of its association with challenges to existing arrangements, if it does characterize rulers, it is presented frequently by them as if they were Davids facing the Goliaths of custom or entrenched interest. Radicalism as a characteristic of government has usually laid great stress on the allegedly tremendous, and unfairly insuperable, obstacles in its path. Radicalism out of power has the real adversary of incumbent elites. Radicalism in power must to an extent invent enemies to maintain its radical credentials. It is not an accident that those most eager to use the powers of government to shape economy, society, or religion to their own favored roots are most likely to detect enemies within.

7. *Approbatory and Critical Uses of the Term*

'Radical' and 'radicalism' have been used with approbatory, pejorative, critical, or neutral tone according to the different historical locations in which they have been employed. Whereas in the 1960s and 1970s 'radical' was often used by conservatives to dismiss the proposals of their opponents, by the end of the twentieth century the term could be employed by left and right alike with entirely positive purpose. In the United Kingdom, Prime Minister Tony Blair, after the Labour Party's election victory of 1997, used radical to mean progressive, optimistic, forward looking, and subsumed socialism and liberalism under the title. The leader of the Conservative Party, William Hague, was equally able to recommend what he termed radical measures, for the term can mean thorough, imaginative, unhidebound.

8. *Historical Instances of Radicalism*

As an abstract term, therefore, radicalism lacks content, though it may have a distinguishable form, and can gain substance only in context. Its substantial meanings have, therefore, been historical and particular, proper names rather than taxonomic categories. 'Radical' as an adjective qualifying a noun has been more common than 'radicalism' as a noun by itself: radical liberalism in eighteenth century Europe; radical philosophy, the radical right, and radical feminism in the twentieth century.

Because any set of arrangements can be subjected to radical appraisal, the terms 'radical' and 'radicalism' have been applied to an extremely wide range of ideas and movements, from Lollards in fourteenth century Europe, Diggers and Levellers during the English revolution of the seventeenth century, to neo-Nazis in twenty-first century Europe and North American. The earliest substantial use was to describe democrats and liberals in the eighteenth century, those inspired by human rights, and by the examples of the American and French Revolutions. Radical liberalism and radical democracy were movements of active opinion that challenged the old order in Europe with demands based on individual social, civil, and political rights. In the nineteenth century, radicalism concentrated on the extension of political rights. The various fascist movements of the mid-twentieth century have been described as the radical right, in that they were in part an accentuation of aims and aversions already present in conservative and right wing politics. In the 1960s and 1970s, the New Left in Europe and North American was an instance of radicalism with a broad range of targets, but a narrower range of solutions from within which each segment chose its favorite: the abolition of capitalism, the overthrow of gerontocracy, the transformation of consciousness with drugs, or the decentralization and de-industrialization of western society. A different sort of radicalism, though one which also stressed individual liberation, was found within the New Right of the 1980s and 1990s. Here, the character of the radical demand was made more complex by a double track radicalism, on the one hand calling for a root and branch destruction of what were seen as restraints on the market and hence on individual utility maximization and self-fulfilment through choice about the use of property, but on the other calling for a radical imposition of cultural and moral controls, if necessary by government. By the end of the twentieth century radicalism was an aspect of environmental or green theory (see *Green Parties*), and of feminism. Radicalism had, too, religious avatars in the New Christian Right in the US, Islamism in Asia and the Middle East (see *Islamic Fundamentalism*), and political Hinduism in India. But in each case, within the variety of the movement, the breadth of targets was given coherence by the unity of ascribed causes for identified malaises.

9. *Conclusion*

Not only does the term radicalism lack any firm historical content, it possesses an equal normative flexibility, and is used neutrally and descriptively, and as a term of both approval and condemnation. For those who believe that language is the only reality we have, it provides a congenial illustration.

A radical may be a person with a vision, but it is usually a vision derived from experience, an extension of the radical's own values to a whole society. There is a conservative foundation for radical politics, which is why, once radicals achieve power, they can appear to shift from change to conservation. Radicals are not for change as such, but only, like conservatives, for particular sets of arrangements, for the cultivation of which change may appear necessary, and for the achievement of which they believe they have identified a single cause or cluster of causes.

See also: Communism; Communist Parties; National Socialism and Fascism; Parties/Movements: Extreme Right; Revolutions, History of; Revolutions, Sociology of; Revolutions, Theories of; Social Movements, Sociology of

Bibliography

Belchem J 1996 *Popular Radicalism in Nineteenth-century Britain*. Macmillan, Basingstoke, UK

Biagini E F 1996 *Citizenship and Community: Liberals, Radicals and Collective Identities in the British Isles, 1865–1931*. Cambridge University Press, Cambridge, UK

Blinkhorn M (ed.) 1990 *Fascists and Conservatives: The Radical Right and the Establishment in Twentieth-century Europe*. Unwin Hyman, London

Inglis F 1982 *Radical Earnestness: English Social Theory 1880–1980*. Robertson, Oxford, UK

Kaplan J, Weinberg L 1999 *The Emergence of a Euro-American Radical Right*. Rutgers University Press, New Brunswick, NJ

Kramnick I 1990 *Republicanism and Bourgeois Radicalism: Political Ideology in Late Eighteenth Century England and America*. Cornell University Press, Ithica, NY

Tawney R H 1964 *The Radical Tradition: Twelve Essays on Politics, Education and Literature*. George Allen & Unwin, London

Wilks S 1993 *Talking About Tomorrow: A New Radical Politics*. Pluto/New Times, London

R. Barker

Radio as Medium

Radio, the transmission of sound via electromagnetic waves, is the most widely used electronic means of communication on the globe. It is second only to spoken language and nonverbal interaction as a medium of human communication. Despite the broad reach of television and the growing role of cable and Internet, radio remains an important arena of cultural expression, commerce, politics, and education in advanced capitalist countries. In many developing countries, including most of Africa, much of Asia, and the Middle East, and parts of Latin America, Eastern Europe, and Oceania, radio continues to be the primary means of mass communication. Due to its mobility, low cost, and flexibility, radio is one of the fastest growing media of communication worldwide.

Although radio has a variety of scientific and commercial applications as a technology of radio wave transmission (e.g., radio astronomy, radio telegraphy, radio telephony), as a social medium radio has been configured principally as a means of point-to-mass communication or broadcasting. Two broad models of radio broadcasting emerged during the first half of the twentieth century: commercial broadcasting and public monopoly broadcasting. While both models defined the radio wave spectrum as a public resource or good, they differed on whether the spectrum should be exploited by competing private entities for profit or developed by a single public entity as a nonprofit service. During the last quarter of the twentieth century both commercial broadcasters and community-based nonprofit broadcasters challenged the public monopoly model and launched new and hybrid broadcasting forms. This article examines the cultural, technological, economic, and political factors that shaped radio as a social medium of broadcasting.

1. Technology and Cultural Form

Radio is comprised of a dynamic sound technology wedded to a flexible yet resilient set of cultural practices. This combination of technological and cultural factors both enables and constrains radio as a medium. Radio production, for example, is shaped by the capacities of the technology (real-time, sound-only) as well as the forms of expression that have been adapted to the medium over time (program genres, formats, and presentation styles). Radio reception is likewise constrained. For example, a Walkman radio differs from a console radio in both its technological configuration and in the cultural assumptions that shape its use (mobile-individualized listening vs. situated-shared listening). This interaction of technology and culture shapes not only the form and content of the radio medium, but radio's relationship to time, space, community, and identity.

In this sense, a defining feature of radio as a technology and cultural form is its conflicted relationship to modernity. On the one hand, radio broadcasting represents a heightened sense of movement towards modernity; that is, movement toward a condition in which human social relations are increasingly abstract, individualized, and future-directed. To the extent that radio is understood to speed up processes of social progress and development (through, for example, the diffusion of innovations) the medium evokes a sense of 'hypermodernity.' On the other hand, radio provides a means of resisting or allaying the impact of increased mobility and rationalization in social life. To the extent that radio is used as a tool for building intimacy and community in an individuated and changeable modern world, broadcasting practices are 'antimodern.'

Radio broadcasting both reflects and constitutes uniquely modern forms of social organization. It works to integrate and coordinate an increasingly mobile, dispersed, and complex society, at the same time that it promotes the privatization of social life. The simultaneous access of vast numbers of isolated listeners to a single radio content creates a virtual common space that tends to become coterminous with the nation (due, in part, to government regulation and commercial market formation). In this way, radio, like print and cinema, acts as a medium for the 'community in anonymity which is the hallmark of modern nations' (Anderson 1991 p. 36).

At the same time, however, radio voices have a materiality that reproduces the immediate, tactile quality of primary orality in a number of ways. Roland Barthes's concept of the 'grain' of the voice is helpful in characterizing the somatic presence of radio voices. Barthes (1990) describes the grain as 'the body in the voice as it sings;' that is, the listener's sense of the physical presence of the performer in the sound of his or her voice (p. 299). Walter Ong (1995) argues that radio talk has the same ability as oral speech to envelope and immerse listeners, uniting the audience 'with themselves and with the speaker' (p. 74). Indeed, broadcasters actively work to recreate the sociability and interactivity of oral settings by developing intimate, chatty forms of radio talk. In this way, radio gives listeners a sense of continuity with older forms of intimacy and community at the same time that it enables distinctly modern forms of social life.

2. History and Political Economy

Radio has historically been deployed by a range of political and economic institutions largely within the bounds of the nation-state. During the last two decades of the twentieth century radio changed from a predominantly public service medium to an eminently commercial one. To understand the meaning of this global transformation it is necessary to investigate the social forces that guided and shaped radio broadcasting beginning in the 1920s. Among the most important influences on the medium have been national and transnational communication corporations, the state (nationalist and expansionist) and community groups.

2.1 Beginnings: The 1920s and 1930s

Radio expanded from a means of point-to-point communication to a medium of point-to-mass communication, or broadcasting, during the 1920s and 1930s. The rise of broadcasting was both shaped by and reflective of larger political and economic changes in the first quarter of the twentieth century, including the rise of transnational communication corporations following World War I (e.g., Philips, RCA, Tele-

funken), the growth of product marketing, increased trade in cultural goods (e.g., film and recorded music), and the escalation of both nationalism and international expansion. Technological developments including voice broadcasting (replacing Morse code) and high-power radio transmissions were also necessary ingredients for broadcasting.

Approaching the medium as a strategic communication technology, scarce resource, and tool for national integration, central governments began asserting control over the radio spectrum in the 1910s and 1920s. In part, this followed the pattern established by state intervention in telegraph and telephone development. As a US Commerce Department publication lamented, 'When the social possibilities of broadcasting became apparent many countries determined upon maintaining it as a public utility, particularly those where government ownership of communication services had previously been the rule' (Batson 1932, p. 2). Even in the USA, however, the central government claimed radio as a public resource and used regulatory powers to promote a national broadcasting system with citizenship and public service requirements for broadcasters. These regulations, codified in the 1927 Radio Act and again in the Communications Act of 1934, granted concessions or licenses for commercial broadcasting and actively encouraged the development of commercial networks on a national scale.

Radio was shaped by the climate of intense nationalism and international expansion in which it was born. In industrialized countries radio was viewed as a means of consolidating the nation as both market and polity, strengthening ties with colonies and dependencies and opening foreign markets. For example, the British Broadcasting Corporation (BBC) began its Empire Service in 1932, while the Japanese government established broadcasting stations in occupied territories in both Korea and China. US media corporations such as the Radio Corporation of America, General Electric, and Westinghouse financed broadcasting stations in foreign markets in order to promote the sale of radio receivers (e.g., Latin America). In less industrialized countries and colonial territories broadcasting rarely reached beyond a few major cities and was often limited to the interests of colonial governments (e.g., Hong Kong, Kenya).

In post-revolutionary societies such as Mexico and Soviet Russia, as well as fascist countries like Germany and Italy, the state turned to radio as a tool for building political solidarity and cultural identity through official broadcasting projects and the distribution of radio sets for both public and in-home listening. In Nazi Germany inexpensive radio sets known as the 'People's Wireless' or 'Goebbel's Lips' allowed listeners to tune in to Adolf Hitler's speeches but blocked the reception of foreign stations (Euromedia Research Group 1997, p. 79). In almost all independent countries radio developed within a nationalistic framework: Broadcasters were required to be citizens transmitting from national soil, the state was given special access to the medium on the grounds of national interest and security, and content was frequently regulated to promote and 'improve' the national culture (e.g., language and music requirements).

Although radio broadcasting took many different forms, two broad types of radio organization—commercial broadcasting and public monopoly broadcasting—emerged during this period. In the commercial model, radio programs were designed to attract large audiences that could be sold to consumer product advertisers through program sponsorship or 'spot' advertisements. This model predominated only in the USA and Latin America, although Canada, Australia, and New Zealand had significant commercial radio sectors and small pockets of commercial radio existed elsewhere. In France and Luxembourg, for example, private broadcasters transmitted commercial programs in direct competition with the BBC (e.g., Radio Luxembourg, Radio-Normandie). In the vast majority of countries and occupied territories radio was a public monopoly (or modified monopoly) financed by government subsidy or receiver license fees and subject to varying degrees of government control. These two general models yielded different geographies of broadcasting: Whereas public monopolies tended to produce a small number of stations serving the national community (with limited regional or local service), commercial systems tended to generate a large number of stations serving the local community. With the rise of national networks, however, local commercial affiliates developed a strong national orientation.

Only countries in North America and Western Europe approached universal radio access by the end of the 1930s when, for example, about 75 percent of the British population had radios. One of the most influential broadcasts of the era, Franklin Roosevelt's first 'Fireside Chat' of March 18, 1933, was heard by an estimated audience of 60 million (out of a total population of about 120 million). Urban and industrializing areas of Latin America, Oceania, and Asia, including the Soviet Union, had small but growing numbers of radio listeners. Rural parts of these regions, along with most of Africa and the Middle East, had negligible radio listenership.

2.2 Take-Offs and Transitions: The 1940s and 1950s

World War II had a major impact on radio organization, content, and listenership. Timely broadcasts of news and information greatly increased the utility of the medium, promoted frequent listening and raised demand for radio sets both during and after the war. Radio listening grew even in countries where most people could not afford radio sets (e.g., Egypt) 'because merchants or owners of coffee shops would

have a set to attract customers' (Dabbous 1994, p. 65). The role of the central state in broadcasting also increased significantly on both the domestic and international fronts. At home, the state took an interest in the security and efficiency of the radio system. Official reports and government-sponsored programs proliferated on both public monopoly and commercial radio stations. In some cases (e.g., Vichy France) the state used the wartime emergency to prohibit private broadcasting and institute a public monopoly system.

International broadcasting was exploited by all of the belligerents during the war, largely in response to Italy and Germany who invested heavily in short-wave broadcasting. The BBC, for example, became the medium for Charles de Gaulle's appeals to the French Resistance, beginning with his celebrated broadcast of June 18, 1940. The US government developed international broadcasting projects aimed at Asia, Europe, and Latin America. In the case of Latin America, the US Coordinator of Inter-American Affairs developed a wartime broadcasting offensive that worked to solidify US hegemony in the region while strengthening the cultural influence of US advertisers and a small number of Latin American broadcasters. In 1948 the Coordinator's radio unit was transferred to the US State Department where it became the Voice of America (VOA) broadcasting service. By the end of the war, international broadcasting became the provision of the central state.

Radio broadcasting became one of the most important ideological and political tools of the Cold War. Along with international broadcasting, the USA, Soviet Union, and their allies sponsored clandestine radio stations in countries around the world (e.g., Guatemala, Turkey, Yugoslavia). The mobility and low cost of radio, however, also made it a powerful tool for grassroots revolutionary movements in the postwar period. Communist insurgents in China in the late1940s and revolutionaries in Cuba in the late 1950s used clandestine broadcasting stations to maneuver against their enemies and mobilize popular support. Fidel Castro's Radio Rebelde, for example, became a model of effective clandestine broadcasting for decades to come.

Most of the countries that fought for independence at the end of World War II (e.g., China, India, Indonesia) followed a path similar to earlier postrevolutionary societies: They established state-directed broadcasting systems that controlled language and content and used the medium for nationalist ends. New governments struggled to improve broadcasting coverage and increase listenership quickly and efficiently. Consequently, broadcasting services were often highly centralized and limited to one or two national services (e.g., the Central People's Broadcasting Station in Beijing, China).

The history of radio in the postwar period was also shaped by the rise of a second broadcasting medium: television. Although television fundamentally transformed radio broadcasting in the USA during this period, television's impact was much more gradual worldwide. Radio remained a public monopoly in most countries and a strong competitor for the national audience. In parts of Europe (e.g., Italy, Norway, Sweden) television failed to reach a majority of the population until the early 1960s and in Latin America the high cost of television guaranteed that radio would continue as the primary mass medium. Indeed, this period can be characterized as a golden age of radio in countries with developed broadcasting systems. In Latin America, for example, popular music (e.g., mambos, rancheras, sambas) provided the primary material for broadcasting while the *radionovela* (drama serial) reached the height of its popularity. In both commercial and public monopoly systems this was an important period of genre formation, with the live production of musical variety programs, comedies, soap operas, novelas, news, and game shows filling the airwaves.

In the USA radio changed radically beginning in the early 1950s. In order to develop national television networks, broadcasting corporations like the National Broadcasting Company stripped capital, talent, and program genres from radio and ultimately dismantled their national radio networks. Without national network programming, radio stations were forced to reorient their activities to the local and regional level and cultivate new markets (e.g., teenagers, ethnic minorities, and out-of-home listeners). Radio content became increasingly diversified and localized as radio shifted to a repertoire of recorded music supervised by disk jockeys (DJs) who advertised products and promoted the radio station. By the late 1950s stations were catering to narrow market segments with increasingly standardized program formats (e.g., 'beautiful music', and 'top-40').

Overall the early postwar period was marked by two divergent trends in radio broadcasting. While richer Western countries moved away from radio and toward television as the primary broadcast medium, many newly independent, postcolonial countries looked to radio as the primary tool for national communication and unification projects. Although television came to symbolize development, it remained impractical for the vast majority of countries due to high costs and infrastructure limitations. Beginning in the postwar decades, then, a significant divide emerged and grew between countries that saw high levels of television penetration beginning in the second half of the 1950s and those in which radio remained the primary mass medium through the late 1970s and beyond.

2.3 Change and Competition: The 1960s and 1970s

Public monopoly broadcasting remained dominant during this period despite technological changes and

increasing competition from television. In the countries that emerged from the last major wave of decolonization (e.g., Algeria, Nigeria, Uganda) public monopoly broadcasting with strong government influence was the norm. In their study, *Broadcasting in the Third World*, Elihu Katz and George Wedell (1977) noted the tendency towards government control of radio broadcasting in many developing countries. In the case of Nigeria, for example, they observed that: 'The demands of a government needing to assert its authority over a country of 50 million people of diverse ethnic, linguistic, and cultural characteristics, between 70 and 80 percent of whom were illiterate, made closer control of broadcasting inevitable' (p. 83).

At the same time, however, public monopoly broadcasters faced growing criticism for controlling and limiting public information. In India, for example, a 1964 government report noted that 'Suspicion of official information has deepened in India because of an incorrect, even improper use of media for personalized publicity and an undue accent on [government] achievements … ' (Ninan 1998, p. 4). In France, during the political crisis of 1968, listeners turned to commercial radio stations outside of France (the peripheral stations) to hear balanced coverage of events that French radio covered with a progovernment bias. In many countries protests against the public monopoly system were registered in the form of unlicensed, pirate stations operating within the national territory. While some countries managed to stave off reforms, significant liberalization began during this period. For example, Italy legalized private broadcasting in 1976 and Sweden permitted noncommercial 'neighborhood radio' as early as 1978.

During the same period commercial broadcasting (both television and radio) came under attack as a wasteland of low-quality programming that failed to meet basic public needs for information, education, and culture. In the USA, legislators responded with the Public Broadcasting Act of 1967 which created a public corporation that helped to finance educational television as well as National Public Radio, a public service national radio network. In Latin America, where radio became accessible to the majority of the population during this period, most governments either initiated or greatly expanded the use of radio in national education projects. Overall, this period witnessed a sizable growth in educational and nonprofit broadcasting by diverse civic groups, including religious organizations, universities, and other nonprofit groups.

Although radio listening increased worldwide during this period, the divide between rich and poor countries continued. Radio's growth in the developed world was driven, in part, by innovations such as frequency modulation broadcasting (FM) and transistor radios. While FM radio made high-fidelity broadcasting possible, inexpensive transistor sets increased radio's portability outside of the home and facilitated the expansion of car radio. In the USA, for example, both innovations helped to push the number of radio sets to 1,813 per 1,000 population by 1975. In the developing world, however, FM expanded much more slowly due to the costs of conversion from amplitude modulation (AM) receivers. In addition, although inexpensive transistor sets made radio much more accessible to communities without electricity, the price of batteries remained prohibitively high for many listeners. Thus radio penetration remained relatively low in poorer regions of the world at the end of the decade. Asia and the Middle East, for example, had 62 sets per 1,000 population while Africa had 70 sets per 1,000 population in 1977 (Katz and Wedell 1977, p. 62).

2.4 Liberalization, Deregulation, and Consolidation: The 1980s and 1990s

The move to reform public monopoly broadcasting that began in the 1970s grew enormously during the 1980s. Liberalization was a response to a variety of pressures: social pressure to increase political and cultural pluralism over the air, political pressure to reduce public spending (especially during a time of economic crisis), and ideological pressure in favor of free market capitalism. This process was particularly evident in Western Europe, where local noncommercial stations were legalized in the early 1980s, followed by commercial stations in the mid- to late-1980s (e.g., Belgium, Denmark, FRG). In many cases, the radio system was opened to allow competition at the local level while the public monopoly on national radio broadcasting remained in place (e.g., the UK).

After the break-up of the Soviet Union and the Eastern Bloc in the late 1980s, most of the world's radio systems faced liberalization of some kind. Much of Eastern Europe became an unregulated market open to competition from North American and European companies (although, in fact, development remained limited due to severe economic dislocation in the region). In Africa most countries opened their radio systems to commercial broadcasting, although commercial development was also highly circumscribed by economic conditions. South Africa, for example, began to liberalize its broadcasting system in the mid-1990s with the privatization of some government-owned stations, licensing of new commercial stations, and the establishment of community broadcasting. Only in the Arab world did monopoly systems remain largely unscathed.

What began as liberalization in the 1980s became a move toward market regulation in the 1990s. As commercial broadcasters grew and faced local competition, they moved toward networking, syndication, and consolidation at a national level. In Sweden, for example, although the Radio Act of 1996 prohibited multiple ownership 'all but a few of the more than 90

commercial stations [were] closely integrated into four more or less national networks' (Euromedia Research Group 1997, p. 218). A similar process took place in France where a political movement in opposition to the public broadcasting monopoly was ultimately exploited by commercial interests. Whereas the 1980s had been dominated by small, activist stations, the 1990s witnessed the 'oligopolization of the private radio sector in France' at the hands of professional broadcasting corporations (Kuhn 1995, p. 107).

The move towards market regulation also produced oligopoly in the US radio industry. Beginning in the early 1980s, deregulation removed federal content regulations and reduced the limits on cross-ownership of media. With the Telecommunications Act of 1996 all national limits on station ownership were removed, opening the door to a series of mergers and consolidations. By the year 2000 radio broadcasting had moved from monopolistic competition to oligopoly: Many major cities had one station owner with 40 percent of the market and some had two owners with 70 percent of the market. According to Charles Fairchild (1999) the 1996 deregulation aimed, in part, to make radio competitive with other consolidated media industries, thus preparing the radio system for exploitation by integrated media conglomerates. At the same time, developments such as Internet radio and digital audio broadcasting (DAB) provided ideal media platforms for the convergence of recorded music, text, and image through a real-time broadcasting-style format.

Radio reached almost every person on the globe by the end of the twentieth century. Radio listening continued to grow in the world's expanding urban areas, primarily in the form of in-car and portable radio listening. Radio listening also grew in rural areas due to electrification projects as well as the development of low-tech, wind-up radio receivers that finally made radio affordable in areas without electricity. Radio's popularity also increased during this period as broadcasters exploited the potential interactivity and intimacy of the medium. Listener-participation talk shows reinvigorated the medium and promoted public discussion of personal topics (especially personal relationships) as well as political issues in countries as different as China, Mexico, and the USA.

In the growing commercial sector, radio stations delivered increasingly standardized program formats to targeted market segments. Format-specific radio chains developed a 'brand value' or 'lifestyle statement' that could be 'applied profitably to other aspects of their listeners' lives,' including other forms of media consumption (Hendy 2000, p. 47). Convergence and conglomeration in radio and related media industries (recorded music, television, Internet), as well as the growth of satellite delivery systems, set the stage for transnational deployment of increasingly 'branded' radio services. Although linguistic diversity and national regulatory policies continued to pose obstacles, regional radio integration was feasible in a few linguistically and economically unified regions (e.g., the Southern Cone of Latin America and the European Economic Community).

In contrast, the last decade of the twentieth century was also a boom period for decentralized community-based radio broadcasting. Community broadcasting, which emphasized community control and participation, was nurtured by democratization movements around the globe. The World Association of Community Radio Broadcasters (AMARC) provided a venue for the organization and promotion of community radio and lobbied for 'social enterprise' broadcasting as a means of encouraging economically viable community stations. Just such a model of community radio was initiated in South Africa in the second half of the 1990s, where nonprofit community broadcasters were allowed to support themselves through advertising. These community stations, some of which served populations of over 50,000, provided local service where none had existed before and covered events such as the local Truth and Reconciliation Hearings. By 2000, despite the vocal protests of the commercial broadcasting industry, even the United States Federal Communication Commission had approved noncommercial low-power FM broadcasting at the community level.

3. *The Future of Radio Broadcasting*

Radio's future is likely to unfold along two quite different paths in the early twenty-first century. First, radio's flexibility as a real-time, sound-only medium will ensure its survival as a social medium of broadcasting. Radio will continue to accompany people while they drive cars, work, and do other activities that require visual attention and mobility. Radio's flexibility and low cost will continue to drive the proliferation of small, decentralized stations at the local level. In less developed regions of the world this form of traditional radio broadcasting will continue to be a viable medium for development by commercial, governmental, and nonprofit groups.

Second, radio will tend to merge with other digital media. Both industrial conglomeration and technological convergence will work to collapse the distinction between radio, television, and Internet. This will occur primarily in advanced capitalist countries where radio will be integrated into multimedia formats that can be delivered via satellite, cable, and mobile radio telephony to privatized consumers. At the same time, a real-time 'radio' style of mass intimacy and anonymous interactivity will continue to shape electronic media and their contradictory relationship to modern life.

See also: Broadcasting: General; Broadcasting: Regulation; International Communication: History; International Communication: Regulation; Media Effects; Media Imperialism; Media, Uses of; Talk Show in Media; Television: History; Television: Industry

Bibliography

Anderson B R O'G 1991 *Imagined Communities*. Verso, London

Barthes R 1990 The grain of the voice. In: Firth S, Goodwin A (eds.) *On Record: Rock, Pop, and the Written Word*. Pantheon Books, New York, pp. 293–300

Batson L D 1932 *Radio Markets of the World*. Government Printing Office, Washington, DC

Dabbous S 1994 Egypt. In: Kamalipour Y R, Mowlana H (eds.) *Mass Media in the Middle East*. Greenwood Press, Westport, CT, pp. 64–8

Douglas S J 1999 *Listening In: Radio and the American Imagination: from Amos 'n Andy and Edward R. Murrow to Wolfman Jack and Howard Stern*. Times Books, New York

Euromedia Research Group (ed.) B. S. Ostergaard 1992 *The Media in Western Europe*. Sage, London

Fairchild C 1999 Deterritorizing radio: deregulation and the continuing triumph of the corporatist perspective in the USA. *Media Culture and Society* **21**: 549–61

Hayes J E 2000 Radio. In: Balderston D, González M (eds.) *Encyclopedia of Contemporary Latin American and Caribbean Cultures*. Routledge, London, Vol. 1, pp. 1234–8

Hayes J E 2000 *Radio Nation: Communication, Popular Culture and Nationalism in Mexico, 1920–1950*. University of Arizona Press, Tucson, AZ

Hendy D 2000 *Radio in the Global Age*. Blackwell Publishers, Malden, MA

Hilmes M 1996 *Radio Voices: American Broadcasting, 1922–1952*. University of Minnesota Press, Minneapolis, MN

Hong L 1994 Radio Revival in China. *Beijing Review* **37**: 15–18

Kamalipour Y R, Mowlana H (eds.) 1994 *Mass Media in the Middle East*. Greenwood Press, Westport, CT

Katz E, Wedell G 1977 *Broadcasting in the Third World*. Harvard University Press, Cambridge, MA

Kuhn R 1995 Radio. In: *The Media in France*. Routledge, London, pp. 77–108

Lent J A (ed.) 1978 *Broadcasting in Asia and the Pacific*. Temple University Press, Philadelphia, PA

McDaniel D O 1994 *Broadcasting in the Malay World*. Ablex, Norwood, NJ

Miller T 1997 Radio. In: Cunningham S, Turner G (eds.) *The Media in Australia*. Allen & Unwin, St. Leonards, NSW, pp. 47–69

Ninan S 1998 History of Indian broadcasting reform. In: Price M E, Verhulst S (eds.) *Broadcasting Reform in India*. Oxford University Press, Delhi, India, pp. 1–21

Ong W 1995 *Orality and Literacy*. Routledge, London

Ritchie S E 1999 The impact of community radio broadcasting on rural development in South Africa. In: *Radio Africa '99: Proceedings of the Third Regional Workshop on Radio Communication in Africa*. University of Botswana, Gaborone, Botswana, pp. 134–7

Scannell P, Cardiff D 1991 *A Social History of British Broadcasting*. Blackwell, Oxford

Strauss N (ed.) 1993 *Radiotext(e). Semiotext(e)*. Columbia University, New York

Williams R 1974 *Television: Technology and Cultural Form*. Schocken Books, New York

J. E. Hayes

Random Assignment: Implementation in Complex Field Settings

Random assignment of units to treatment and control groups allows researchers to make the strongest possible causal connection between the treatment and observed outcomes. This article considers the way in which random assignment is implemented in educational and social field settings.

1. Widespread Acceptance of Random Assignment

The last three decades of the twentieth century have seen widespread acceptance of random assignment as an important aspect of high-quality research studies designed to test the effectiveness of interventions. In the 1970s and 1980s, proponents of randomized designs for studying program impacts began to assemble a literature attesting to their feasibility and frequent use (Boruch 1974, Boruch et al. 1978, Boruch and Wothke 1985). By the 1990s, randomized experiments in the social sciences had become so numerous that Greenberg and Shroder (1997) prepared a 500-page compendium of recent experiments in various content areas. The US Congress now routinely specifies that random assignment be used in studies of the effectiveness of federally funded programs, and recently called for a random assignment study of Head Start, America's flagship social program for disadvantaged children (Advisory Committee on Head Start Research and Evaluation 1999).

Randomized experiments have limitations as research vehicles (see Cook and Campbell 1979 for a discussion of the strengths and weaknesses of randomized experiments), and they are not universally accepted as the best way to learn about program effectiveness. One sector of the evaluation community espouses the virtues of qualitative research in place of randomized experiments and other quantitative approaches (e.g., Guba and Lincoln 1982), while another sector is more interested in issues of generalizability than internal validity (e.g., Cronbach 1982). Cook (1999) has written about the anti-random assignment culture that exists in American schools of education. Still, there is no denying that American society relies heavily on randomized experiments as one source of information about the effectiveness of social programs (St. Pierre and Puma 2000).

2. The Purpose of Random Assignment

As long as researchers take the time to explain the strengths and weaknesses of random assignment studies, those who commonly question experiments are easily able to understand the rationale underlying random assignment. Perhaps most helpful is the argument that random assignment is a special purpose tool, and while it is good only for one purpose, for making causal inferences, it is far better for that purpose than any alternative. As a special purpose tool, random assignment has many limitations. For example, it does not assure group comparability over time (attrition from the evaluation sample may result in nonrandom analytic groups), it cannot ensure that subjects remain in the intervention being studied (with or without random assignment, interventions may incur high dropout rates), and it does not help with issues of generalizability (a one-site random assignment study is no more generalizable than a one-site quasi-experiment). Cook and Campbell (1979) stated it clearly: 'Though random assignment is germane to the research goal of assessing whether the treatment may have caused any observed effects, it is conceptually irrelevant to all other research goals' (p. 343).

3. Common Objections to Random Assignment

Most of the common objections to random assignment in studies of program effectiveness do not dispute its utility as a tool for understanding causal connections. Instead, objections take other forms. Some are made in an attempt to avoid participation in a research study. Take the oft-made claim that 'random assignment is unethical' or that 'random assignment means denying services to some subjects.' These are political statements, either made without understanding what random assignment means, or made in an attempt to retain control over which subjects are to be served by a given social program.

When there is excess demand for a social intervention, program operators are in the position of denying services to needy families every day as they make decisions about which families are the neediest or the most deserving. As long as a randomized experimental study of such an in-demand program does not reduce the total number of subjects served, then there is no denial of services—the same aggregate number of families are served with or without an experiment. What an experiment does change is the mechanism for selecting which families receive the services; random assignment is used to decide who participates in the program instead of the judgment of program staff. Thus, when program operators say that random assignment is unethical or involves denying services, they usually are saying that they don't want to give up the power to determine which families are served.

But what about denying services to families? Is this always unethical? Suppose a funding agency wants to learn about the effectiveness of one of its social programs and in the absence of evidence of effectiveness, the funding agency plans to withdraw its support. Suppose further that there is no waiting list and hence that a randomized experiment would make it impossible to maintain aggregate service levels since some families would be assigned to a control group. In this case, the ethical problem of doing an experiment which reduces the number of families serviced by a social program has to be weighed against the possibly greater ethical problem of not doing an experiment and risking reduced funding for the entire program.

An underlying, and typically unvoiced, objection to random assignment is that program staff or funders assume the effectiveness of social programs instead of worrying about whether scarce resources are being spent in the most appropriate manner. This is understandable, since scientific skepticism seems to counteract the ability of program operators to do their best work. But again, this is a political issue—that it is appropriate to assume the effectiveness of a social program instead of subjecting it to a rigorous test.

Another objection to random assignment is that assignment to a control group 'feels bad' to program staff and to the families they serve. The argument that randomization is like a lottery doesn't work in many social program settings. Why? Because a lottery usually has only one winner, and the losers can take solace in the fact that they are in the great majority. On the other hand, those assigned to the control group in a randomized study do not view their assignment the same way they view a lottery loss, since the odds of winning the lottery are miniscule, whereas the odds of being randomly assigned to the desired social program are usually 50/50.

Some other arguments are that randomized experiments are expensive (Levitan 1992) or that they take a long time. The response to this sort of complaint is, 'Compared to what?' A randomized experiment is expensive and time-consuming when compared with the alternative of doing no research at all. But this is rarely a realistic alternative. Many quasi-experimental alternatives are just as expensive (any design in which data are collected from a treatment group and a nonequivalent comparison group) and time-consuming (the common pre–post study) as randomized studies. Some alternative approaches are quick and inexpensive, for example, relying on data collected for other purposes. However, the fact is that none of these alternatives are capable of doing the one thing that randomized studies excel at—providing strong evidence about causal connections. The problem with using alternatives to random assignment to study program effectiveness is that even if they are less expensive, even if they require less time, even if they are easier to implement, they rarely are worth the time and money spent on them. Why? Because the results

are not helpful for their intended purpose—for understanding program effectiveness. To paraphrase the common aphorism about education, 'If you think randomized studies are expensive, try ignorance.'

4. Conditions Making Random Assignment Easier

There are several circumstances that reduce the difficulty of implementing random assignment studies. Suppose an intervention program is in short supply and there is a list of subjects waiting to participate. This often happens with Head Start (the largest federal preschool program in the US) projects in large metropolitan areas. If being placed on a waiting list means that a family is likely to miss a substantial portion of the Head Start year, then randomly assigning families from the waiting list to be in the program or in a control group might well be preferable to being on the waiting list with certainty. The absence of a waiting list does not mean that subjects are uninterested in a program. It often means that program staff do not want to spent time and energy on recruiting until they have openings. Hence, it is possible in short-supply situations for researchers to generate a waiting list by helping to recruit subjects and informing them about the planned study.

Suppose a treatment can be delayed with no ill effect on the participants. This is likely the situation in a school that wishes to test the short-term effectiveness of a one-semester school health curriculum. In this case, it would be possible to randomly assign students to take the school health course in the fall semester or in the spring semester.

Suppose a program's proponents are skeptical and are searching for the best solution to a social problem instead of trying to promote a specific intervention. Or, suppose a demand for evidence of program effectiveness cannot be ignored, for example, a congressional mandate. In these cases, political objections to random assignment are eased because program implementers are supportive of or at least are willing to cooperate with the research.

Suppose that a social program has many local projects, for example, the Head Start program has about 1,600 local projects and the Even Start program (a family literacy program) has about 800 projects. An experimental evaluation of these programs could be done be selecting a relatively large number of projects and randomly assigning only a few subjects within each project. This approach minimizes the unpleasant effects of random assignment for any given project, making it more palatable to local program staff.

Assigning higher-level units is another way to ease the pain of random assignment. Suppose that there is an interest in testing three different solutions to a given social problem. One way to construct this test is to randomly assign individuals to one of three interventions implementing the three different solutions. A different approach which avoids many of the common objections to random assignment is to randomize groups of individuals. For example, a federal agency could hold a grant competition to test three different approaches to preschool education. School districts would not be allowed to receive a grant unless they indicated their willingness to implement any of the three approaches. School districts that receive grants would then be randomly assigned to implement one of the three preschool approaches.

5. Working Towards Correct Random Assignment

Random assignment is more likely to be successfully implemented if done by researchers and research organizations with prior experience in conducting experimental studies. Certain research organizations have built the capacity to and reputation for conducting random assignment studies. Greenberg et al. (1999) report that almost half of the social experiments started in the US since 1983 have been done by three large research organizations.

In the same vein, random assignment is more likely to be successful if it is under the control of researchers rather than program implementers (Conner 1977). The latter are likely to make exceptions or may misunderstand random assignment rules, even if those rules are carefully prepared by researchers. The research team does not have to be physically present in order to control the random assignment. One approach is to have program staff be responsible for recruiting subjects, explaining the experimental alternatives, and transmitting lists of study subjects to the research team via fax, email, or other means. The research team then does the random assignment and sends back listings of research subjects, sorted into treatment and control groups. A related approach is for the research team to prepare a random assignment computer program for use by program staff who recruit subjects and enter basic data on a microcomputer which does the random assignment on the spot.

The way in which random assignment is implemented depends on the way in which applicants are recruited for an intervention. Sometimes a large pool of applicants are recruited. In these cases, information on the entire pool can be sent to the research team at the same time, simplifying the random assignment. The process is more complicated when applicants enter a program on a rolling basis. In these cases, a random assignment algorithm has to be developed so that applicants are told their treatment/control status soon after they apply. However, simple assignment rules such as assigning the first applicant to the

treatment, the second to control, the third to treatment, etc. are easily manipulated by program staff who often have control over the flow of applicants. More complex assignment systems are less subject to local control. For example, random assignment of families to an intervention could be done in blocks of four, where the order of treatment and control assignment is random within each block.

6. *Preserving Randomization*

Implementing random assignment is difficult. Once achieved, how can it be preserved throughout the life of an evaluation? One issue has to do with when randomization takes place. This involves tradeoffs between generalizability and internal validity. Families interested in participating in the Even Start family literacy program could be randomized as soon as they are first recruited for the program, after they have had the evaluation explained to them and have indicated their willingness to accept any of the alternatives, or after they have completed a one-month 'tryout period' during which they can try Even Start and see whether they truly are interested in the program. Early randomization draws on the largest possible pool of families (increased generalizability) but risks losing many families (decreased internal validity) who find out that they are not assigned to the desired group or who leave the program after a tryout period. Late randomization winnows the pool of families (decreased generalizability) to those who are most likely to participate in the study (increased internal validity). Gueron and Pauly (1991) and Rieken et al. (1974) give additional examples.

Whether assigned to an intervention or a control group, subjects may refuse to cooperate with research measurements once an experiment has been started. Such attrition has the potential to destroy the comparability of the intervention and control groups. Clearly, it is best to use all means possible to avoid attrition of subjects. In some circumstances (e.g., studies of welfare programs), subjects can be sanctioned if they do not participate in research; in other cases, incentives for participation can be offered. Attrition is all but unavoidable in most social experiments since families move, are ill or otherwise unavailable for measurement activities, or simply refuse to participate in the research. Methods for determining whether attrition is related to the treatment and ways of dealing with this problem are discussed by Cook and Campbell (1979).

Are there ways to reduce the likelihood that subjects will refuse to go along with the randomization, or refuse to participate in the study? One approach is to plan the study so that subjects who agree to participate in the randomization have a better chance of getting the desired intervention than they would have if the study had not existed. This works best when there is a waiting list and when the intervention has a large number of openings relative to the number of subjects needed for the study. A second approach is to offer a competing intervention to members of the control group. This could work in several different ways. One competing intervention could be a financial or in-kind incentive for participating in the study. Another competing intervention could be a reduced or modified version of the treatment being studied. Yet a third alternative could be to offer control group subjects the opportunity to enroll in the desired intervention at a later point in time.

Bibliography

Advisory Committee on Head Start Research and Evaluation 1999 *Evaluating Head Start: A Recommended Framework for Studying the Impact of the Head Start Program*. US Department of Health and Human Services, Washington, DC

Boruch R 1974 Bibliography: Illustrative randomized field experiments for program planning and evaluation. *Evaluation* **2**: 83–7

Boruch R, McSweeny A J, Soderstrom E J 1978 Randomized field experiments for program planning, development and evaluation. *Evaluation Quarterly* **2**: 655–95

Boruch R, Wothke W 1985 Randomization and field experimentation. *New Directions for Program Evaluation* **28**

Conner R F 1977 Selecting a control group: An analysis of the randomization process in twelve social reform programs. *Evaluation Quarterly* **1**: 195–244

Cronbach L J 1982 *Designing Evaluations of Educational and Social Programs*, 1st edn. Jossey-Bass, San Francisco

Cook T D 1999 *Considering the major arguments against random assignment: An analysis of the intellectual culture surrounding evaluation in American schools of education*. Paper presented at the Harvard Faculty Seminar on Experiments in Evaluation, Cambridge, MA

Cook T D, Campbell D T 1979 *Quasi-experimentation: Design & Analysis Issues for Field Settings*. Rand McNally College Pub. Co., Chicago

Greenberg D, Shroder M 1997 *The Digest of Social Experiments*, 2nd edn. The Urban Institute Press, Washington, DC

Greenberg D, Shroder M, Onstott M 1999 The social experiment market. *Journal of Economic Perspectives* **13**: 157–72

Guba E G, Lincoln Y S 1982 *Effective Evaluation*, 1st edn. Jossey-Bass, San Francisco

Gueron J M, Pauly E 1991 *From Welfare to Work*. Russell Sage Foundation, New York

Levitan S A 1992 *Evaluation of Federal Social Programs: An Uncertain Impact*. George Washington University, Center for Social Policy Studies, Washington, DC

Rieken H W, Boruch R F, Campbell D T, Coplan W, Glennan T K, Pratt J, Rees A, Williams W 1974 *Social Experimentation: A Method for Planning and Evaluating Social Innovations*. Academic Press, New York

St. Pierre R G, Puma M J 2000 Toward the dream of the experimenting society. In: Bickman L (ed.) *Advances in Social Research Methods: The Heritage of Donald Campbell*. Sage Publications, New York

R. G. St. Pierre

Random Numbers

Random number generators (RNGs) are available from many computer software libraries. Their purpose is to produce sequences of numbers that appear as if they were generated randomly from a specified probability distribution. These *pseudorandom numbers*, often called *random numbers* for short (with a slight abuse of language), are crucial ingredients for a whole range of computer usages, such as statistical experiments, simulation of stochastic systems, numerical analysis with Monte Carlo methods, probabilistic algorithms, computer games, cryptography, and gambling machines, to name just a few. These RNGs are small computer programs implementing carefully crafted algorithms, whose design is (or should be) based on solid mathematical analysis. Usually, there is a basic uniform RNG whose aim is to produce numbers that imitate independent random variables from the uniform distribution over the interval [0,1] (i.i.d. $U[0,1]$), and the random variables from other distributions (normal, chi-square, exponential, Poisson, etc.) are simulated by applying appropriate transformations to the uniform random numbers.

1. Random Number Generators

The words *random number* actually refer to a poorly defined concept. Formally speaking, one can argue that no number is more random than any other. If a fair dice is thrown eight times, for example, the sequences of outcomes 22222222 and 26142363 have exactly the same chance of occurring, on what basis could the second one be declared more random than the first?

A more precisely defined concept is that of independent random variables. This is a pure mathematical abstraction, which is nevertheless convenient for modeling real-life situations. The outcomes of successive throws of a fair dice, for instance, can be modeled as a sequence of independent random variables uniformly distributed over the set of integers from 1 to 6, i.e., where each of these integers has the probability 1/6 for each outcome, independently of the other outcomes. This is an abstract concept because no absolutely perfect dice exists in the real world.

An RNG is a computer program whose aim is to imitate or simulate the typical behavior of a sequence of independent random variables. Such a program is usually purely deterministic: if started at different times or on different computers with the same initial state of its internal data, i.e., with the same *seed*, the sequence of numbers produced by the RNG will be exactly the same.

When constructing an RNG to simulate a dice, for example, the aim is that the statistical properties of the sequence of outcomes be representative of what should be expected from an idealized dice. In particular, each number should appear with a frequency near 1/6 in the long run, each pair of successive numbers should appear with a frequency near 1/36, each triplet should appear with a frequency near 1/216, and so on. This would guarantee not only that the virtual dice is unbiased, but also that the successive outcomes are uncorrelated.

However, practical RNGs use a finite amount of memory. The number of states that they can take is finite, and so is the total number of possible streams of successive output values they can produce, from all possible initial states.

Mathematically, an RNG can be defined (see L'Ecuyer 1994) as a structure (S, μ, f, U, g), where S is a finite set of states; μ is a probability distribution on S used to select the initial state (or seed) s_0; the mapping $f: S \to S$ is the transition function; U is a finite set of output symbols; and $g: S \to U$ is the output function.

The state evolves according to the recurrence $s_n = f(s_{n-1})$, for $n \geqslant 1$. The output at step n is $u_n = g(s_n) \in U$. These u_n are the so-called random numbers produced by the RNG. Because S is finite, the generator will eventually return to a state already visited (i.e., $s_{i+j} = s_i$ for some $i \geqslant 0$ and $j > 0$). Then, $s_{n+j} = s_n$ and $u_{n+j} = u_n$ for all $n \geqslant i$. The smallest $j > 0$ for which this happens is called the period length ρ. It cannot exceed the cardinality of S. In particular, if b bits are used to represent the state, then $\rho \leqslant 2^b$. Good RNGs are designed so that their period length is close to that upper bound.

Choosing s_0 from a uniform initial distribution μ can be crudely approximated by using external randomness such as picking balls from a box or throwing a real dice. The RNG can then be viewed as an extensor of randomness, stretching a short random seed into a long stream of random-looking numbers.

2. Randomness from Physical Devices

Why use a deterministic computer algorithm instead of a truly random mechanism for generating random numbers? Simply calling up a system's function from a program is certainly more convenient than throwing dice or picking balls from a box and entering the corresponding numbers on a computer's keyboard, especially when thousands (often millions or even billions) of random numbers are needed for a computer experiment.

Attempts have been made at constructing RNGs from physical devices such as noise diodes, gamma-ray counters, and so on, but these remain largely impractical and unreliable, because they are cumbersome, and because it is generally not true that the successive numbers they produce are independent and uniformly distributed. If one insists on getting away

from a purely deterministic algorithm, a sensible compromise is to combine the output of a well-designed RNG with some physical noise, e.g., via bitwise exclusive-or.

3. Quality Criteria

Assume, for the following, that the generator's purpose is to simulate i.i.d. $U[0,1]$ random variables. The goal is that one cannot easily distinguish between the output sequence of the RNG and a sequence of i.i.d. $U[0,1]$ random variables. That is, the RNG should behave according to the null hypothesis $\mathbb{H}_0$: 'The u_n are i.i.d. $U[0,1]$,' which means that for each t, the vector $(u_0, \ldots, u_{t-1})$ is uniformally distributed over the t-dimensional unit cube $[0,1]^t$. Clearly, $\mathbb{H}_0$ cannot be exactly true, because these vectors always take their values only from the finite set

$$\Psi_t = \{(u_0, \ldots, u_{t-1}) : s_0 \in S\}.$$

If s_0 is random, Ψ_t can be viewed as the sample space from which the output vectors are taken. The idea, then, is to require that Ψ_t be very evenly distributed over the unit cube, so that $\mathbb{H}_0$ be *approximately true for practical purposes*. A huge state set S and a very large period length ρ are necessary requirements, but they are not sufficient; good uniformity of the Ψ_t is also needed. When several t-dimensional vectors are produced by an RNG by taking non-overlapping blocks of t output values, this can be viewed in a way as picking points at random from Ψ_t, without replacement.

It remains to agree on a computable figure of merit that measures the evenness of the distribution, and to construct specific RNGs having a good figure of merit in all dimensions t up to some preset number t_1 (because computing the figure of merit for all t up to infinity is infeasible in general). Several measures of discrepancy between the empirical distribution of a point set Ψ_t and the uniform distribution over $[0, 1]^t$ were proposed and studied in the last few decades of the twentieth century (Niederreiter 1992, Hellekalek and Larcher 1998). These measures can be used to define figures of merit for RNGs. A very important criterion in choosing such a measure is the ability to compute it efficiently, and this depends on the mathematical structure of Ψ_t. For this reason, it is customary to use different figures of merit (i.e., different measures of discrepancy) for analyzing different classes of RNGs. Some would argue that Ψ_t should look like a typical set of random points over the unit cube instead of being too evenly distributed. That it should have a chaotic structure, not a regular one. However, chaotic structures are hard to analyze mathematically. It is probably safer to select RNG classes for which the structure of Ψ_t can be analyzed and understood, even if this implies more regularity, rather than selecting an RNG with a chaotic but poorly understood structure.

For a concrete illustration, consider the two linear recurrences

$$x_{1,n} = (a_{1,1}x_{1,n-1} + \ldots + a_{1,k}x_{1,n-k}) \bmod m_1$$
$$x_{2,n} = (a_{2,1}x_{2,n-1} + \ldots + a_{2,k}x_{2,n-k}) \bmod m_2$$

where k, the m_j's, and the $a_{i,j}$'s are fixed integers, 'mod m_j' means the remainder of the integer division by m_j, and the output function

$$u_n = (x_{1,n}/m_1 - x_{2,n}/m_2) \bmod 1$$

This RNG is a *combined linear multiple recursive generator* (L'Ecuyer 1999a). Its state at step n is the $2k$-dimensional vector

$$(x_{1,n}, \ldots, x_{1,n-k+1}, x_{2,n}, \ldots, x_{2,n-k+1}),$$

whose first k components are in $\{0, 1, \ldots, m_1-1\}$ and last k components are in $\{0, 1, \ldots, m_2-1\}$. There are thus $(m_1m_2)^3$ different states and it can be proved that Ψ_k is the set of all k-dimensional vectors with coordinates in $\{0, 1/m, \ldots, (m-1)/m\}$, where $m = m_1m_2$. This implies, if m is large, that for each $t \leqslant k$, the t-dimensional cube $[0, 1]^t$ is evenly covered by Ψ_t. For $t > k$, it turns out that in this case, Ψ_t has a regular lattice structure (Knuth 1998, L'Ecuyer 1998), so that all its points are contained in a series of equidistant parallel hyperplanes, say at a distance d_t apart. To obtain a good coverage of the unit cube, d_t must be as small as possible, and this value can be used as a selection criterion.

L'Ecuyer (1999a) recommends a concrete implementation of this RNG with the parameter values: $k = 3$, $m_1 = 2^{32}-209$, $(a_{11}, a_{12}, a_{13}) = (0, 1403580, -810728)$, $m_2 = 2^{32}-22853$, and $(a_{11}, a_{12}, a_{13}) = (527612, 0, -1370589)$. This RNG has two main cycles of length $\rho \approx 2^{191}$ each (the total number of states is approximately 2^{192}). Its lattice structure has been analyzed in up to $t = 48$ dimensions and was found to be excellent, in the sense that d_t is not far from the smallest possible value that can be achieved for an arbitrary lattice with $(m_1m_2)^3$ points per unit of volume. These parameters were found by an extensive computer search.

For a second example, consider the *linear congruential generator* (LCG) defined by the recurrence $x_n = ax_{n-1} \bmod m$ and output function $u_n = x_n/m$. This type of RNG is widely used and is still the default generator in several (perhaps the majority of) software systems. With the popular choice of $m = 2^{31}-1$ and $a = 16807$ for the parameters, the period length is $2^{31}-2$, which is now deemed much too small for serious applications (Law and Kelton 2000, L'Ecuyer 1998). This RNG also has a lattice structure similar to the previous one, except that the distances d_t are much larger, and this can easily show up in simulation

results (Hellekalek and Larcher 1998, L'Ecuyer and Simard 2000). Small LCGs of this type should be discarded.

Another way of assessing the uniformity is in terms of the leading bits of the successive output values, as follows. Suppose that S (and Ψ_t) has cardinality 2^e. There are tl possibilities for the l most significant bits of t successive output values $u_n, \ldots, u_{n+t-1}$. If each of these possibilities occurs exactly 2^{e-tl} times in Ψ_t, for all l and t such that $tl \leqslant e$, the RNG is called *maximally equidistributed* (ME). Explicit implementations of ME or nearly ME generators can be constructed by using linear recurrences modulo 2 (L'Ecuyer 1999b, Tezuka 1995, Matsumoto and Nishimura 1998).

Besides having a long period and good uniformity of Ψ_t, RNGs must be efficient (run fast and use little memory), repeatable (the ability of repeating exactly the same sequence of numbers is a major advantage of RNGs over physical devices, e.g., for program verification and variance reduction in simulation; Law and Kelton 2000, and portable (work the same way in different software/hardware environments). The availability of efficient methods for jumping ahead in the sequence by a large number of steps is also an important asset of certain RNGs. It permits the sequence to be partitioned into long, disjoint substreams and to create an arbitrary number of virtual generators (one per substream) from a single backbone RNG (Law and Kelton 2000).

Cryptologists use different quality criteria for RNGs (Knuth 1998, Lagarias 1993). Their main concern is the unpredictability of the forthcoming numbers. Their analysis of RNGs is usually based on asymptotic analysis, in the framework of computational complexity theory. This analysis is for families of generators, not for particular instances. They study what happens when an RNG of size k (i.e., whose state is represented over k bits of memory) is selected randomly from the family, for $k \to \infty$. They seek RNGs for which the work needed to guess the next bit of output significantly better than by flipping a fair coin increases faster, asymptotically, than any polynominal function of k. This means that an intelligent guess quickly becomes impractical when k is increased. So far, generator families conjectured to have these properties have been constructed, but proving even the existence of families with these properties remain an open problem.

4. Statistical Testing

RNGs should be constructed based on a sound mathematical analysis of their structural properties. Once they are constructed and implemented, they are usually submitted to empirical statistical tests that try to detect statistical deficiencies by looking for empirical evidence against the hypothesis $\mathbb{H}_0$ defined previously. A test is defined by a test statistic T, function of a fixed set of u_n's, whose distribution under $\mathbb{H}_0$ is known. The number of different tests that can be defined in infinite and these different tests detect different problems with the RNGs. There is no universal test or battery of tests that can guarantee, when passed, that a given generator is fully reliable for all kinds of simulations. Passing a lot of tests may improve one's confidence in the RNG, but never proves that the RNG is foolproof. In fact, no RNG can pass all statistical tests. One could say that a bad RNG is one that fails simple tests, and a good RNG is one that fails only complicated tests that are very hard to find and run. In the end, there is no definitive answer to the question 'What are the good tests to apply?'

Ideally, T should mimic the random variable of practical interest in a way that a bad structural interference between the RNG and the problem will show up in the test. But this is rarely practical. This cannot be done, for example, for testing RNGs for general-purpose software packages. For a sensitive application, if one cannot test the RNG specifically for the problem at hand, it is a good idea to try RNGs from totally different classes and compare the results.

Specific statistical tests for RNGs are described in Knuth (1998), Hellekalek and Larcher (1998), Marsaglia (1985), and other references given there. Experience with empirical testing tells us that RNGs with very long periods, good structure of their set Ψ_t, and based on recurrences that are not too simplistic, pass most reasonable tests, whereas RNGs with short periods or bad structures are usually easy to crack by statistical tests.

5. Non-uniform Random Variates

Random variates from distributions other than the uniform over [0, 1] are generated by applying further transformations to the output values u_n of the uniform RNG. For example, $y_n = a + (b-a)u_n$ produces a random variate uniformly distributed over the interval $[a, b]$. To generate a random variate y_n that takes the values -1 and 1 with probability 0.2 each, and 0 with probability 0.6, one can define $y_n = -1$ if $u_n \leqslant 0.2$, $y_n = 1$ if $u_n > 0.8$, and $y_n = 0$ otherwise.

Things are not always so easy, however. For certain distributions, compromises must be made between simplicity of the algorithm, quality of the approximation, robustness with respect to the distribution parameter, and efficiency (generation speed, memory requirements, and setup time). Generally speaking, simplicity should be favored over small speed gains.

Conceptually, the simplest method for generating a random variate y_n from distribution F is *inversion*: define $y_n = F^{-1}(u_n) \stackrel{\text{def}}{=} \min\{y | F(y) \geqslant u_n\}$. Then, if we view u_n as a uniform random variable, $P[y_n \leqslant y] = P[F^{-1}(u_n) \leqslant y] = P[u_n \leqslant F(y)] = F(y)$, so y_n has distribution F. Computing this y_n requires a formula or a good numerical approximation for F^{-1}. As an example, if y_n is to have the geometric distribution with

parameter p, i.e., $F(y) = 1-(1-p)^y$ for $y = 1, 2, \ldots$, the inversion method gives $y_n = F^{-1}(u_n) = 1+\ln(1-u_n)/\ln(1-p)$. The normal, student, and chi-square are examples of distributions for which there are no close form expression for F^{-1}, but good numerical approximations are available (Bratley et al. 1987).

In most simulation applications, inversion should be the method of choice, because it is a monotone transformation of u_n into y_n, which makes it compatible with major variance reductions techniques (Bratley et al. 1987). But there are also situations where speed is the real issue and where monotonicity is no real concern. Then, it might be more appropriate to use fast non-inversion methods such as, for example, the *alias method* for discrete distributions or its continuous *acceptance-complement* version, the *acceptance/rejection* method and its several variants, *composition* and *convolution* methods, and several other specialized and fancy techniques often designed for specific distributions (Bratley et al. 1987, Devroye 1986, Gentle 1998).

Recently, there has been an effort in developing so-called automatic or black box algorithms for generating variates from an arbitrary (known) density, based on acceptance/rejection with a transformed density (Hörmann and Leydold 2000). Another important class of general non-parametric methods is those that sample directly from a smoothed version of the empirical distribution of a given data set (Law and Kelton 2000, Hörmann and Leydold 2000). These methods shortcut the fitting a specific type of distribution to the data.

6. *A Consumer's Guide*

No uniform RNG can be guaranteed against all possible defects, but the following ones have fairly good theoretical support, have been extensively tested, and are easy to use: the Mersenne twister of Matsumoto and Nishimura (1988), the combined MRGs of L'Ecuyer (1999a), the combined LCGs of L'Ecuyer and Andres (1997), and the combined Tausworthe generators of L'Ecuyer (1999b). Further discussion and up-to-date developments on RNGs can be found in Knuth (1998), L'Ecuyer (1998) and from the following web pages:

http://www.iro.umontreal.ca/~lecuyer
http://random.mat.sbg.ac.at
http://cgm.cs.mcgill.ca/~luc/
http://www.webnz.com/robert/rng_links.htm

For non-uniform random variate generation, software and related information can be found at:

http://cgm.cs.mcgill.ca/~luc/
http://statistik.wu-wien.ac.at/src/unuran/

See also: Monte Carlo Methods and Bayesian Computation: Overview; Random Assignment: Implementation in Complex Field Settings

Bibliography

Bratley P, Fox B L, Schrage L E 1987 *A Guide to Simultation*, 2nd edn. Springer, New York

Devroye L 1986 *Non-uniform Random Variate Generation*. Springer, New York

Gentle J E 1998 *Random Number Generation and Monte Carlo Methods*. Springer, New York

Hellekalek P, Larcher G (eds.) 1998 *Random and Quasi-random Point Sets, Vol. 138, Lecture Notes in Statistics*. Springer, New York

Hörmann W, Leydold J 2000 Automatic random variate generation for simulation input. In: Joines J A, Barton R R, Kang K, Fishwick P A (eds.) *Proceedings of the 2000 Winter Simulation Conference*. IEEE Press, Pistacaway, NJ, pp. 675–82

Knuth D E 1998 *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*, 3rd edn. Addison-Wesley, Reading, MA

Lagarias J C 1993 Pseudorandom numbers. *Statistical Science* **8**(1): 31–9

Law A M, Kelton W D 2000 *Simulation Modeling and Analysis*, 3rd edn. McGraw-Hill, New York

L'Ecuyer P 1994 Uniform random number generation. *Annals of Operations Research* **53**: 77–120

L'Ecuyer P 1998 Random number generation. In: Banks J (ed.) *Handbook of Simulation*. Wiley, New York, pp. 93–137

L'Ecuyer P 1999a Good parameters and implementations for combined multiple recursive random number generators. *Operations Research* **47**(1): 159–64

L'Ecuyer P 1999b Tables of maximally equidistributed combined LFSR generators. *Mathematics of Computation* **68**(225): 261–9

L'Ecuyer P, Andres T H 1997 A random number generator based on the combination of four LCGs. *Mathematics and Computers in Simulation* **44**: 99–107

L'Ecuyer P, Simard R 2000 On the performance of birthday spacings tests for certain families of random number generators. *Mathematics and Computers in Simulation* **55**(1–3): 131–7

Marsaglia G 1985 A current view of random number generators. In: *Computer Science and Statistics, Sixteenth Symposium on the Interface*. North-Holland, Amsterdam, pp. 3–10

Matsumoto M, Nishimura T 1998 Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation* **8**(1): 3–30

Niederreiter H 1992 *Random Number Generation and Quasi-Monte Carlo Methods, Vol. 63, SIAM CBMS–NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, PA

Tezuka S 1995 *Uniform Random Numbers: Theory and Practice*. Kluwer, Norwell, MA

P. L'Ecuyer

Ranke, Leopold von (1795–1886)

Leopold Ranke was born in Wiehe an der Unstrut in Thuringia on the December 20, 1795 as the son of the lawyer Gottlob Israel Ranke. His family had long served as evangelical pastors in the area of Saxony

between the rivers Saale and Unstrut. From 1809 to 1814 he attended the renowned Pforta school at Naumburg in Saxony, and from the spring of 1814 to the summer of 1818 he studied theology and classical literature in Leipzig. He completed his doctorate in 1817 with a (lost) dissertation on Thukydides. In 1818 Ranke started teaching history at the Fridericianum Gymnasium in Frankfurt an der Oder. In 1825 he was appointed associate professor of history at the University of Berlin, and in 1834 he was promoted to a full professor. In 1841 he was appointed official historian to the Prussian state by Friedrich Wilhelm IV, and in 1854 to the renewed parliament. Ranke was ennobled in 1865, becoming Leopold von Ranke, and in 1867 he was named chancellor of the order *pour le merité*. He rejected countless offers to teach at other universities, at Munich among others, and remained in Berlin, where he died on the May 23, 1886. From 1880 onwards he was almost blind, and was forced to dictate his later works to assistants.

Ranke's career as a historian was influenced less by the historical science of the time, but rather by his interest in classical literature. His earliest existing studies are on the subject of translation and interpretation of classical Greek texts. Recent research has shown that the German literature of his time, Goethe, Schiller, and Friedrich Schlegel, also had an important influence on his conception of historical writing. It is also known that Ranke admired the literary form of contemporary British and French historiography. Ranke's early years were during the era of French hegemony over Germany and Europe, and the time of the rising German nationalist movement from 1806 onwards. These experiences of contemporary history left lasting impressions on Ranke's thinking. His political sympathies lay with Prussian state conservatism. On behalf of the Prussian government, but with independent control, Ranke published the *Historisch-politische Zeitschrift* from 1832 to 1836, through which the government aimed to exert a moderate conservative influence on the public. Ranke had to give up the journal after a short time because of lack of success; however, he did use the opportunity to publish a number of historical-political essays, in which he developed important methodological principles, and also began to sketch out his later, larger works. The most important of these articles are for example *Über die Zeit Ferdinands I. und Maximilians II.* (On the Time of Ferdinand I. and Maximilian II.), *Frankreich und Deutschland* (France and Germany), *Vom Einfluß der Theorie* (On the Influence of Theory) (all 1832), *Rom 1815–1823* (Rome from 1815 to 1823), *Die großen Mächte* (*The Great Powers*) (both 1833), and *Das politische Gespräch* (*The Political Conversation*) (1836). Ranke placed importance on his good relations with the Prussian court, but was also held in high esteem by the Bavarian King Maximilian II, who had attended his lectures as a student at the University of Berlin, and for whom Ranke held a series of free lectures in Berchtesgaden in the September and October 1854 which were taken down in shorthand. The king had wished for 'a short and condensed overview of the motivating ideas of the various centuries of history from the Christian era on,' and Ranke developed basic characteristics of his historical perception in eight lectures which were then published in 1888 under the title *Über die Epochen der neueren Geschichte* (On the Epochs of Recent History).

Ranke's works total 63 volumes, including nine volumes of a world history which he dictated late in his life but was unable to complete. His first work *Geschichte der romanischen und germanischen Völker von 1494–1514* (History of the Latin and Teutonic Nations from 1494 to 1514) (1824) was a great success. It already showed some of the basic characteristics of his conception of Europe, and was of historiographical importance particularly because Ranke made an exemplary critical analysis of his sources in a separate volume, *Zur Kritik neuerer Geschichtsschreiber* (On the Critical Methods of Recent Historians). In this work he raised the method of textual criticism used in the late eighteenth century, particularly in classical philology to the standard method of scientific historical writing. In 1827 he visited the archives of Vienna, Venice, Rome, and Florence and various German libraries with a stipend of the Prussian government and discovered the since famous Venetian relations, legation reports from the sixteenth and seventeenth centuries, which became an important basic source for the series of his great historical works beginning in 1832. The experiences of the July Revolution in France in 1830 led him to interpret the modern history of the European states more strongly in perspective of the relation between the present and the past. From 1832 to 1836 he published the three volumes of *Die römischen Päpste, ihre Kirche und ihr Staat im 16 und 17. Jahrhundert* (Ecclesiastical and Political History of the Popes during the 16th and 17th Centuries), and from 1839 to 1847 the six volumes of *Die deutsche Geschichte im Zeitalter der Reformation* (History of the Reformation in Germany). These two works were developed in a single process and complement one another. Further main works followed in 1847–1848, *Preußische Geschichte* (History of Prussia), 1852–1861, *Französische Geschichte* (History of France), and 1859–1868, *Englische Geschichte* (History of England), as well as the incompleted world history of 1881–1888. After the foundation of the German Reich in 1870 and 1871, Ranke partly revised his *Preußische Geschichte* by emphasizing the 'German' role of Prussia in the old Reich more strongly in a second edition. Numerous studies concentrating for the most part on Prussian and Austrian history complete his work. It must be noted that the focal point of these works, with editions and investigations on the Prussian state chancellor von Hardenberg and on Friedrich Wilhelm IV, was in the area of contemporary history.

From 1830 onwards Ranke's thought centers on the history of the European states of the sixteenth and seventeenth centuries, based on his experiences of the time. However, depending on his estimation of the problems and their continuations, he did look back into the late and even the early middle ages, and on the other hand pursued the configurations of the European states not only up to 1789–1795, but also into the beginning of the nineteenth century. We can see from the focal point of the sixteenth and seventeenth centuries, the age of the secularized Renaissance papacy, the reformation and religious wars, the modern process of building states and the crystallization of the European powers system and its functioning between repeated hegemonic claims and equilibrium, that Ranke viewed the European historical process as essentially grounded in the tension between state and religion. The state and the church are buttresses of the system, at the same time however also buttresses of change, as they tend to broaden and reinforce their power so widely that they cause opposition and opposite tendencies, just as do the acting personalities themselves. In Ranke's historical perception however, each state and church, as well as their representatives, have a specific 'task,' with which they contribute to the overall human task of cultural development. The task of historians is to recognise this 'task' and the 'leading ideas,' and to portray at which points the protagonists deviate from them, causing turbulences and even historical catastrophes, and where they return to realizing these leading ideas and thereby to a harmonious equilibrium of the forces. Multiplicity is essential for the workings of this self-regulating system of states and religion, or churches. 'States are "Thoughts of God"' (*Politisches Gespräch*, ed. U. Muhlack, p. 95), and historical research pursues their development over the centuries. Ranke's great historical writings begin at the point when the medieval universalism of the church and the 'Sacrum Imperium' of the Germans started to break down and the specific European culture developed, in which the pretention of religious or state universality repeatedly appeared, but was crushed by the resistance of the state and national individualities. Multiplicity and unity are connected in this history of the great European nations and their summing together in the European powers system in such a way as to enable human energies to be realized more productively than ever before.

The recognition of this historical process has for Ranke, a value in itself. Ranke formulated this thought, which is a plea against liberal progressive thinking, repeatedly; most markedly in the sentence from the lectures '*Über die Epochen der neueren Geschichte*': 'Each age is immediate to God, and its value is not based on that which it brings forth, but in its own existence, in its own self, which however does not rule out that something else emerged from it' (von Ranke et al. 1971, p. 60). Ranke hereby opposes every functionalization of historical events and situations for contemporary interests, yet makes it clear that the present cannot be understood without the historical process and that contemporary action may not ignore the way things have come to pass. For example one can learn from the past that is not possible to re-establish the absolute monarchy or the old importance of the church in the 'Age of Revolution.' Therefore Ranke did not limit himself to a purely contemplative ideal of historical insight beyond all practical applications. In his inaugural lecture in Berlin in 1836, *Über die Verwandtschaft und den Unterschied der Historie und der Politik* (On the Relation and the Difference of History and Politics), he declared that an insight into the past as free from prejudice as possible, was a condition for the possibility of reasonable politics in the present. Historical insight explains the conditions under which contemporaries act. However, these must be drawn into the leading ideas of their predecessors, and may not criticize them from the standpoint of their contemporary interests. Ranke also described this concept of historical objectivity in various works, the most well known of which being the formulation: 'We have given history the task of judging the past, of teaching the world for the use of future years: the present attempt does not forbid itself such high tasks: it merely aims to show how it really was' (SW 33, VII).

Ranke founded his ideal of objectivity for the most part on historical theology, not on logical insights. The basis of historical insight was for him a certain knowledge of God: 'History is religion' … 'At least there is the closest connection between them' (letter to Georg Waitz, December 30, 1841, Briefwerk 313). As Ranke saw God as manifested in all historical appearances, and especially in their inner relations to one another, historical insight represented for him the specific modern and up to date search for the recognition of God. The religious conviction that all historical action is carried by providence meant on one hand a limitation of the possibilities of insight, because the historian can at best guess at God's will. On the other hand this conviction lends Ranke's historical perception a harmonistical character, which has earned sharp criticism in the past decades of historism-critical historical science. His ideal of objectivism was rejected by his Berlin colleague and methodological opponent Johann Gustav Droysen but also by his pupil Heinrich von Sybel and by Heinrich von Treitschke, the protagonist of the politically active, Prussia-oriented liberal-conservative historiography in late nineteenth century Germany.

In the field of German history of the late nineteenth and early twentieth century, however, Ranke was very highly respected. His rather too uncritical successors unfortunately produced a perception of history which did not take sufficient account of the specific forces of the modern age. As mentioned previously, Ranke himself had always related history to the present and

had included the French Revolution as well as the re-establishment of the European system of powers at the Congress of Vienna in his continuity-oriented thinking. He saw in the French Revolution in particular, a dynamic element which gave the stagnating European powers system new energies and ability to integrate, so that European state system could regenerate and reinforce itself through increased nationalization. There was no longer room for industrialization, the modern competitive economy and capitalism, liberalism, even socialism and communism in this historical perception.

Leopold Ranke is counted along with Johann Gustav Droysen (1808–1884), Barthold Georg Neibuhr (1776–1831), and Theodor Mommsen (1817–1903) as one of the most important founders of modern history in Germany and as the most influential international representative of classical German historism. His own research work and his teaching at the University of Berlin made an essential contribution to the circulation and final establishment of modern source of criticism. He pushed ahead the autonomization of history as a science, by distancing himself both from the rhetorical tradition of history as an 'art,' and also from the contemporary historical philosophy, particularly that of Georg Wilhelm Friedrich Hegel. With his systematic work from sources he emphasized the importance of empiricism, and simultaneously made it repeatedly clear in theoretical comments and in his historical practice than the individual event or fact can only be interpreted in its surrounding context. His art of portrayal was essential to his success and to his, at times, world recognition in the developing modern international field of history, with which style he has gone down in history as one of the founders of modern historiography. Ranke expressly aimed to make history into a 'pleasure.' His mastery of specifically narrative forms contributed to bridging the gap between the empirical and specialized historian and the bourgeois readership of the nineteenth century. As mentioned previously, one consequences of this esthetization could also be a harmonizing view of events and situations. However, Ranke's narrative mastership did allow him to lead from the apparently coincidental individual event on to the contemplation of the most general insights, in a way that is very plausible for the historical-political consciousness of the nineteenth century. His stories deal with the responsibility of individuals for their fate, without suggesting that they could control this fate. That is the meaning of the sentence, that the historian is to deal with humanity, 'as it is, explainable or unexplainable: the life of the individual, the generations, the peoples, at times the hand of God above them' (SW 33/34, VIII).

See also: Historicism; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); History: Overview; Objectivity, History of; Religion, History of

Bibliography

Berding H 1971 *Leopold von Ranke*. In: Wehler H-U (ed.) *Deutsche Historiker*, Göttingen, vol. 1, pp. 7–24

Fulda D 1996 *Wissenschaft aus Kunst. Die Entstehung der modernen deutschen Geschichtsschreibung 1760–1860*. Berlin, pp. 296–410

Hardtwig W 1991 Geschichtsreligion—Wissenschaft als Arbeit—Objektivität. Der Historismus in neuer Sicht. *Historische Zeitschrift* **252**: 1–32

Hardtwig W 1997 Historismus als ästhetische Geschichtsschreibung. *Geschichte und Gesellschaft* **23**: 99–114

Hinrichs C 1754 *Ranke und die Geschichtstheologie der Goethezeit*. Göttingen/Frankfurt Main/Berlin

Iggers G 1971 *Deutsche Geschichtswissenschaft*. Munich, pp. 86–119

Iggers G, von Moltke K (eds.) 1973 *The Theory and Practice of History, Leopold Ranke*. Indianapolis, New York

Krieger L 1977 *Ranke. The Meaning of History*. Chicago, London

Metz K-H 1979 *Grundformen historiographischen Denkens. Wissenschaftsgeschichte als Methodologie*. Dargestellt an Ranke, Treitschke und Lamprecht, Munich

Muhlack U 1988 Leopold von Ranke. In: Hammerstein N (ed.) *Deutsche Geschichtswissenschaft um 1900*. Stuttgart, pp. 11–36

Vierhaus R 1957 *Ranke und die soziale Welt*. Münster

von Ranke L 1897–1890 *Sämtliche Werke*, 54 vols. [Ed. Dove A] Leipzig, Germany

von Ranke L 1926 *Werke Gesamtausgabe der Deutschen Akademie* [ed. Joachimsen P] Munich

von Ranke L 1949 *Das Briefwerk* [ed. Fuchs W P] Hamburg

von Ranke L 1953 *Die römischen Päpste in den letzten vier Jahrhunderten*, with an introduction by V. Friedrich. Baethgen, Stuttgart

von Ranke L 1957 *12 Bücher preußischer Geschichte* [ed. Andreas W] Hamburg, 4 vols.

von Ranke L 1963 *Die großen Mächte, politisches Gespräch*. Göttingen

von Ranke L 1964–75 *Aus Werk und Nachlaß* [eds. Fuchs W P and Schieder T]. Munich/Vienna, 4 vols.

von Ranke L 1988 *Und die moderne Geschichtswissenschaft*. Stuttgart

von Ranke L 1989 *Über die Epochen der neueren Geschichte*. Darmstadt

von Ranke L 1995 *Die großen Mächte, politisches Gespräch*. Frankfurt Main/Leipzig

W. Hardtwig

Ranking Models, Mathematics of

Suppose that k items are ranked from best (1) to worst (k) by a nondeterministic process. For example, the observed ranking could represent an arbitrarily selected voter's preferences for the $k = 5$ candidates in an election, or a judge's ranking of $k = 4$ salad dressings by their perceived tartness. A probability model on rankings—or, more simply, a *ranking model*—specifies

the probability, $P(\pi)$, of observing each of the $k!$ rankings π. This article discusses many of the ranking models that have been proposed in the statistical and psychological literature, and interrelates their basic properties. The focus will be on three widely used methods for constructing ranking models, based on: (a) order statistics, (b) paired comparisons, and (c) distances between permutations.

Sect. 1 describes Thurstone order statistics models. In Sect. 2, several thematic properties of ranking models are introduced, and then these general properties are discussed in the particular context of Thurstone order statistics models. Sects. 3 and 4 consider ranking models induced by paired comparisons and distance-based ranking models, respectively. For each model, its underlying motivation is studied, as well as its relationship to the basic properties of Sect. 2.

Much of this material is from the survey article by Critchlow et al. (1991). Other useful general references include the comprehensive monograph by Marden (1995) and the volume edited by Fligner and Verducci (1993).

Some basic notation is as follows. Once the k items are labeled arbitrarily as item 1 to item k, a *ranking* of the k items corresponds to a permutation function π from $\{1,\ldots, k\}$ onto $\{1,\ldots, k\}$, where $\pi(i)$ is the rank assigned to item i, $i = 1,\ldots, k$, and where $(\pi(1),\ldots, \pi(k))$ is the associated vector of ranks. Alternatively, any ranking π can be described by the corresponding *ordering* of the k items, denoted by the bracketed vector $\langle \pi^{-1}(1),\ldots, \pi^{-1}(k)\rangle$, where $\pi^{-1}(r)$ is the item assigned rank r, $r = 1,\ldots, k$. Often, the simpler notation $\langle i_1,\ldots, i_k\rangle$ will be used. For example, $\langle 3, 1, 2\rangle$ denotes the ordering in which item 3 is ranked as best, item 1 is ranked as second best, and item 2 is ranked as worst.

1. Order Statistics Models

In his research on psychophysical phenomena, Thurstone (1927) hypothesizes that the actual perception of each item is a random variable. The relative ordering of these random variables then determines the judge's ordering of the items. Thurstone's original method considers only paired comparisons of the items, but Daniels (1950) extends the method to full orderings of the k items.

Formally, let X_i be a continuous random variable associated with item i, for $i = 1,\ldots, k$. A random ranking π is defined by setting $\pi(i)$ equal to the rank, from smallest to largest, of X_i among $\{X_1,\ldots, X_k\}$. In other words, the ranking π with ordering $\langle i_1,\ldots, i_k\rangle$ is assigned the probability

$$P(\pi) = P\{X_{i_1} < \cdots < X_{i_k}\} \tag{1}$$

If the random variables X_i are allowed to have arbitrary dependencies, then *any* probability distribution on rankings can be obtained from this construction. To restrict the model, it is often assumed that the X_i are independent, and that their cumulative distribution functions $F_i(x) = F(x-\mu_i)$ constitute a location family, where F is fixed and where the μ_i ($i = 1,\ldots, k$) are shift parameters. The resulting ranking model is referred to as a *Thurstone order statistics model* (see, e.g., Yellott 1977).

Two well-studied cases of the Thurstone model are the Thurstone (1927)–Mosteller (1951)–Daniels (1950) model (*TMD model*) where F is the standard normal distribution, and the Luce (1959)–Plackett (1975) model where $F(x) = 1-\exp(-\exp x)$ is the Gumbel distribution. Brook and Upton (1974) use the TMD model to analyze voters' rankings of candidates for public office, while Henery (1981) uses the Luce–Plackett model to study outcomes of horse races. As discussed in the next section, the Luce–Plackett model affords a simple expression for the ranking probabilities given in Eqn. (1). Also, this model coincides with the first order model in Plackett's (1975) system of logistic models. Further development of order statistics models can be found, for example, in Yellott (1977), Henery (1983), Stern (1990), and Bockenholt (1993).

2. Properties of Ranking Models

This section introduces the study of ranking models with respect to certain thematic properties. In particular, the properties of reversibility, complete consensus, and L-decomposability—which are shared by many diverse types of ranking models—are motivated and defined. Each of these properties is then discussed in the context of Thurstone order statistics models. In later sections, the same thematic properties will be reconsidered in relation to additional classes of ranking models.

2.1 Reversibility

Let $\mathscr{P}$ denotes a *class* of ranking models $\{P(\pi): P \in \mathscr{P}\}$, usually indexed by a set of parameters. Assume that a particular strategy of ranking items from best (rank 1) to worst (rank k) corresponds to a ranking model $P \in \mathscr{P}$. Suppose one now reverses the meaning of the ranks, so that rank 1 is given to the worst item and rank k is given to the best item. Is the resulting ranking model $P^*(\pi)$ still in the same class $\mathscr{P}$—in other words, does P^* correspond to another choice of the parameters? If so, then the class $\mathscr{P}$ of ranking models is said to be *reversible* (Luce 1959 pp. 69–70, Marley 1968, 1982).

For fixed F, consider the class $\mathscr{P}_F$ of Thurstone order statistics models, with shift parameters $\mu_1,\ldots, \mu_k$. Without loss of generality, assume that F has median zero. If the distribution F is symmetric, then the class $\mathscr{P}_F$ is reversible. Indeed, if particular parameter values

$\mu_1,\ldots,\mu_k$ correspond to ranking from best to worst, then the values $-\mu_1,\ldots,-\mu_k$ correspond to ranking from worst to best. Consequently, the class $\mathscr{P}_F$ of TMD ranking models (F standard normal) is reversible. On the other hand, Pendergrass and Bradley (1960) show that the class $\mathscr{P}_F$ of Luce–Plackett ranking models—with F the asymmetric Gumbel distribution—is not reversible (for recent work, see Yellott 1997).

2.2 Complete Consensus

An important problem is to identify a single ranking that best summarizes a probability distribution or sample of rankings (see, e.g., Marley 1993). By definition, a ranking model $P(\pi)$ is *unimodal*, with modal ranking π_0, if π_0 uniquely maximizes P. Any ranking model satisfies this simple condition for unimodality, provided there is not a tie for the most likely ranking. However, for many ranking models, the items may be ordered in a much stronger sense than that implied by the mere existence of a modal ranking π_0. For any two items i and j, suppose that $\pi_0(i) < \pi_0(j)$, and that $P[\pi(i) < \pi(j)]$, the probability that item i is preferred to item j, exceeds 0.5. If this probability continues to exceed 0.5 given any fixed assignment of ranks to the other $k-2$ items, then item i is *strongly preferred* to item j. If every pair of items can be related by strong preference, then the ranking model $P(\pi)$ is said to have *complete consensus* with respect to the modal ranking π_0. In brief, the condition is that if item a is ranked ahead of item b in π_0, then

$$P\{\langle i_1,\ldots, a,\ldots, b,\ldots, i_{k-2}\rangle\} > P\{\langle i_1,\ldots, b,\ldots, a,\ldots, i_{k-2}\rangle\}$$

for *any* fixed assignment of ranks to the other $k-2$ items $i_1,\ldots, i_{k-2}$. The notion of complete consensus is discussed by Henery (1981), Fligner and Verducci (1988), and Critchlow et al. (1991).

One can construct distributions F for which the resulting Thurstone order statistics model does not even have a unique modal ranking (see Savage 1956). However, if the μ_i are distinct and if the density F' has monotone likelihood ratio, that is, if $F'(x-\mu_i)/F'(x-\mu_j)$ is decreasing in x for $\mu_i < \mu_j$, then the Thurstone order statistics model has the complete consensus property (Savage 1956, 1957), and the ordering of the items in the modal ranking coincides with the ordering of their associated parameters μ_i from least to greatest. Thus, for example, the TMD and Luce–Plackett models have complete consensus.

2.3 L-decomposability

The property of L-decomposability is a generalization of Luce's (1959) ranking postulate, defined in Eqn. (2) below. If L-decomposability holds for a ranking model, then the ranking process can be viewed as a sequence of 'independent' stages: at stage 1 the most preferred item is chosen, at stage 2 the most preferred of the remaining items is chosen, and so forth until a full ranking is produced. The probability of the resulting ranking is assumed to be the product of the choice probabilities across the various stages.

Formally, for any subset B of the item labels $\{1,\ldots, k\}$, and for each $i \in B$, let the *choice probability* $P_B(i)$ be the probability that item i is most preferred among those in B. Note that by definition, the choice probability $P_B(i)$ can depend only on i and B, and not on the relative ordering of any items not in B. Luce's ranking postulate states that for every ranking π with associated ordering $\langle i_1,\ldots, i_k\rangle$

$$P(\pi) = P_{\{i_1,\ldots, i_k\}}(i_1)\cdot P_{\{i_2,\ldots,i_k\}}(i_2)\cdot\ldots\cdot P_{\{i_{k-1}, i_k\}}(i_{k-1}). \quad (2)$$

Any ranking model $P(\pi)$ that can be expressed in the form (2) for *some* collection of choice probabilities $P_B(i)$ is said to be *L-decomposable* (Critchlow et al. 1991, Fisburn 1994).

In formulating the ranking postulate, Luce had a *particular* family of choice probabilities $P_B(i)$ in mind—discussed in Eqn. (4) below—and from these he obtained a particular ranking model $P(\pi)$. On the other hand, the notion of L-decomposability is a slight generalization: given the ranking model $P(\pi)$, does there exist some collection of choice probabilities—not necessarily Luce's—that give rise to $P(\pi)$ via Eqn. (2)?

Critchlow et al. (1991) prove that a necessary and sufficient condition for L-decomposability is that for each rank $r = 3,\ldots, k$ and for each $i_1,\ldots, i_r$, the conditional probability

$$P[\pi^{-1}(r) = i_r \mid \pi^{-1}(1) = i_1,\ldots, \pi^{-1}(r-1) = i_{r-1}] \quad (3)$$

is invariant under permuting the $i_1,\ldots, i_{r-1}$. If a ranking model is L-decomposable, then it is easy to generate random rankings from the model, simply by generating the ranking stagewise, and using the choice probabilities at each stage.

The Luce–Plackett model is an important example of a Thurstone order statistics model that satisfies L-decomposability, and indeed this property is closely linked with Luce's (1959) original formulation of the model. Luce's approach is to construct choice probabilities $P_B(i)$ on the basis of his Choice Axiom, which implies that for each item i there is a value p_i—the 'merit' of item i—such that

$$P_B(i) = p_i \Big/ \sum_{j\in B} p_j \quad (4)$$

for any subset B of $\{1,\ldots, k\}$. Without loss of generality, the p_i can be scaled to sum to 1, so that p_i is interpreted as the probability that item i is most preferred among the full set of items. In conjunction

with the ranking postulate in Eqn. (2), the choice probabilities in Eqn. (4) yield the ranking model

$$P\{\langle i_1,\ldots, i_k\rangle\} = \prod_{r=1}^{k-1} \left(p_{i_r} \Big/ \sum_{j\in B_r} p_j \right) \tag{5}$$

where $B_r = \{i_r,\ldots, i_k\}$ is the set of items remaining at stage r. Moreover, it can be proved that this is indeed an alternative description of the Luce–Plackett model, since Eqn. (5) coincides with the Thurstone order statistics model with Gumbel F and with $\mu_i = -\log p_i$. See, for example, Yellott (1977), who attributes this result to Holman and Marley see also *Luce's Choice Axiom*).

3. Ranking Models Induced by Paired Comparisons

Babington Smith (1950) uses a conditioning argument for inducing a probability model on rankings from a set of arbitrary paired comparison probabilities. For each pair of items $i < j$, let p_{ij} be the probability that item i is preferred to item j $(i \neq j)$ in a paired comparison of these two items. Imagine a tournament in which all of the $C(k; 2) \equiv k(k-1)/2$ possible paired comparisons are made independently. If the results of this tournament contain no circular triads $(h \neq i \neq j \neq h)$, then the tournament corresponds to a unique ranking π of the items; otherwise the entire tournament is repeated until a unique ranking is obtained. The probability of any resulting π is thus given by

$$P(\pi) = C \prod_{\{(i,\, j):\pi(i) < \pi(j)\}} p_{ij} \tag{6}$$

where p_{ij} for $i > j$ is defined by $p_{ij} \equiv 1 - p_{ji}$, and C is a constant chosen to make the probabilities sum to 1.

With regard to the properties of Sect. 2, the class of all *Babington Smith models* in Eqn. (6) is reversible, since if the parameter values p_{ij} (for $i < j$) correspond to ranking from best to worst, then the values $p^*_{ij} = 1 - p_{ij}$ correspond to ranking from worst to best. For a Babington Smith model to have complete consensus, Critchlow et al. (1991) prove that sufficient conditions are that no $p_{ij} = 0.5$ and that the $\{p_{ij}\}$ are strongly stochastically transitive (i.e., $p_{ij} \geqslant 0.5$ and $p_{jm} \geqslant 0.5$ imply $p_{im} \geqslant \max\{p_{ij}, p_{jm}\}$; see David 1988)). Finally, to verify L-decomposability, it is straightforward to check that Eqn. (3) is invariant under permuting the $i_1,\ldots, i_{r-1}$, for all models of the form Eqn. (6).

Although the Babington Smith model is indexed by $k(k-1)/2$ parameters, constraints on the $\{p_{ij}\}$ proposed by Mallows (1957) lead to three important subclasses of the general Babington Smith model, having significantly fewer parameters. These subclasses are called the Mallows–Bradley–Terry (MBT) model, the rho-based model, and the ϕ-model, respectively.

To describe the MBT model, suppose the paired comparison probabilities have the Bradley–Terry (Bradley and Terry 1952) form

$$p_{ij} = v_i/(v_i + v_j)$$

for some nonnegative parameters $v_1,\ldots, v_k$. Mallows suggests substituting the Bradley–Terry paired comparison probabilities into Eqn. (6), which leads to the *MBT ranking model*: for any ranking π with associated ordering $<i_1,\ldots, i_k>$,

$$P(\pi) = C(\boldsymbol{v}) \prod_{r=1}^{k-1} (V_{i_r})^{k-r} \tag{7}$$

where $\boldsymbol{v} = (v_1,\ldots, v_k)$ and the constant $C(\boldsymbol{v})$ is chosen to make the probabilities sum to 1. Since the Bradley–Terry paired comparison probabilities are strongly stochastically transitive, any MBT model has complete consensus, provided the v_i are distinct. Moreover, the ordering of the v_i from largest to smallest corresponds to the ordering of the items from best to worst in the modal ranking.

A further simplification of the MBT model has just one continuous parameter $\theta \in (0, 1)$ and a vector of discrete parameters $(\pi_0(1),\ldots, \pi_0(k))$, where $\pi_0(i)$ is the rank assigned to item i by the modal ranking π_0. Assuming that $v_i = \theta^{2\pi_0(i)}$ in Eqn. (7) gives Mallows's *rho-based model*

$$P(\pi \mid \theta, \pi_0) = C(\theta)\, \theta^{R^2(\pi,\, \pi_0)} \tag{8}$$

where

$$R^2(\pi, \pi_0) = \sum_{i=1}^{k} [\pi(i) - \pi_0(i)]^2$$

is the Spearman's rho-distance between the rankings π and π_0, and is related to Spearman's (1904) correlation coefficient. The rho-based model has the interpretation that the probability $P(\pi)$ decreases geometrically, as the R^2 distance from π to π_0 increases. θ is a dispersion parameter: as $\theta \neq 1$, the rho-based model approaches the uniform distribution on all $k!$ possible rankings, whereas when $\theta \neq 0$, the model becomes increasingly concentrated about the modal ranking π_0.

The final simplification of the general Babington Smith model is the well-known *ϕ-model* (Mallows 1957, Mann 1945). Unlike the rho-based model, the ϕ-model is not an MBT model. To describe the ϕ-model, suppose that in Eqn. (6) the corresponding paired

comparison probabilities have the simple form $p_{ij} = (\phi+1)^{-1} \geqslant 0.5$ for $\pi_0(i) < \pi_0(j)$, so that the probability that item i is preferred to item j only depends upon their relative order in the modal ranking. Then the resulting ranking model is

$$P(\pi \mid \phi, \pi_0) = C(\phi)\phi^{T(\pi, \pi_0)} \tag{9}$$

where

$$T(\pi, \pi_0) = \sum_{i<j} I\{[\pi(i)-\pi(j)][\pi_0(i)-\pi_0(j)] < 0\}$$

is the Kendall's tau-distance between the rankings π and π_0, and $I(\cdot)$ is the indicator function: $I(A) = 1$ if the event A occurs, and $= 0$ otherwise. Thus, the ϕ-model and its parameters have interpretations analogous to the rho-based model, except the tau-distance is used instead of Spearman's rho-distance. The ϕ-model has been developed, applied, and extended subsequently by many authors, for example Feigin and Cohen (1978), Schulman (1979), Smith and Evans (1991), and Huh et al. (1995), to name just a few.

4. *Distance-based Ranking Models*

By letting $\lambda = -\log\phi$ in (8), the ϕ-model can be written in the form of an exponential family as

$$P(\pi \mid \lambda, \pi_0) = C(\lambda)\, exp\{-\lambda T(\pi, \pi_0)\}$$

where the constant $C(\lambda) = [\sum_{\pi \in S_k} \exp\{-\lambda T(\pi, \pi_0)\}]^{-1}$. Diaconis (1988 p. 104) uses this representation to motivate a general class of metric-based ranking models

$$P(\pi \mid \lambda, \pi_0) = C(\lambda) \exp\{-\lambda d(\pi, \pi_0)\}$$

where d is an arbitrary metric on the group S_k of all permutations.

Diaconis (1988 pp. 112–9) considers the following metrics on S_k to be among the most widely used in applied scientific and statistical problems:

$$T(\pi, v) = \sum_{i<j} I\{[\pi(i)-\pi(j)][v(i)-v(j)] < 0\}$$

is the Kendall's tau metric:

$$R(\pi, v) = \left(\sum_i [\pi(i)-v(i)]^2 \right)^{1/2}$$

is the Spearman's rho metric:

$$F(\pi, v) = \sum_i |\pi(i)-v(i)|$$

is the Spearman's footrule metric:

$$H(\pi, v) = \sum_i I\{\pi(i) \neq v(i)\}$$

is the Hamming metric:

$$C(\pi, v) = \text{minimal number of transpositions needed to transform } \pi \text{ into } v$$

is Cayley's metric and

$$U(\pi, v) = k - (\text{the maximal number of items ranked in the same relative order by } \pi \text{ and } v)$$

is Ulam's metric. All of these satisfy the usual metric axioms of positivity, symmetry, and the triangle inequality, and thus give rise to metric-based ranking models. Additionally, the class of metric-based models may be enlarged to include Mallows's rho-based model, Eqn. (8), by dropping the triangle inequality requirement for d. The resulting function d is called a distance rather than a true metric, and the resulting model is a distance-based ranking model (Fligner and Verducci 1986). Not much is lost in this generalization, although the triangle inequality property has a desirable interpretation with respect to the behavior of likelihood ratio tests for π_0 (Critchlow 1985 pp. 101–2).

Properties of distance-based ranking models are related to analogous properties of the distance d. Let $\mathscr{P}_d$ denote the class of all ranking models associated with a fixed distance d. Then $\mathscr{P}_d$ is reversible if the distance d satisfies

$$d(\pi^*, v^*) = d(\pi, v) \quad \text{for all } \pi, v \in S_k \tag{10}$$

where $\pi^*(i) \equiv k+1-\pi(i)$ and $v^*(i) \equiv k+1-v(i)$ are the permutations that rank the items in exactly the reverse order as π and v, respectively. All of the distances T, R, R^2, F, H, C, and U satisfy Eqn. (10), and hence give rise to reversible classes.

A distance-based model has complete consensus if and only if d possesses the transposition property (Critchlow 1985 pp.101–2): $d(\pi, v) \leqslant d(\pi', v)$ whenever π, π', and v are permutations such that π and π' assign identical ranks to all items except for i and j, and v agrees with π on the relative ordering of i and j. Since the distances T, R, R^2, and F satisfy this transposition property, the models based on these distances exhibit complete consensus.

Finally, Critchlow et al. (1991) prove that a distance-based ranking model is L-decomposable if and only if the distance d is additively decomposable, in the sense that for each $r = 2,\ldots, k$, there exist functions f_r and g_r such that

$$d(\pi, e) = f_r[\pi^{-1}(1),\ldots, \pi^{-1}(r-1)] + g_r[\pi^{-1}(r),\ldots, \pi^{-1}(k)]$$

where e is the identity permutation $e(i) \equiv i$. It follows that the models based on T, R^2, F, and H are L-decomposable.

4.1 Extensions

Some important multiparameter extensions of the distance-based and ϕ-models include the ϕ-component models and multistage models of Fligner and Verducci (1986, 1988), as well as the orthogonal contrast models (Chung and Marden 1991, 1993, Marden 1992).

Critchlow (1985) develops distance-based models for partially ranked data, such as rankings of only the q best items out of k, where $q < k-1$. The set of all partial rankings of the best q items is identified with the quotient space S_k/S_{k-q} of the permutation group, consisting of all right cosets of the subgroup $S_{k-q} = \{\pi \in S_k \mid \pi(i) = i,\ i = 1,\ldots, q\}$. Any metric d on S_k induces a Hausdorff metric d^* on the quotient space S_k/S_{k-q}, which serves as the metric for these models. Critchlow gives the Hausdorff extensions of T, R, F, H, C, and U, and applies the resulting models to partially ranked data sets.

See also: Consensus Panels: Methodology; Nonparametric Statistics: Rank-based Methods; Order Statistics; Ordered Relational Structures; Scaling and Classification in Social Measurement; Unfolding and Vector Models

Bibliography

Babington Smith B 1950 Discussion on Professor Ross's paper. *Journal of the Royal Statistical Society B* **12**: 153–62

Bradley R A, Terry M A 1952 Rank analysis of incomplete block designs. I. *Biometrika* **39**(3): 324–45

Bockenholt U 1993 Applications of Thurstonian models to ranking data. In: Fligner M A, Verducci J S (eds.) *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, New York, pp. 157–72

Brook D, Upton G J G 1974 Biases in local government elections due to position on the ballot paper. *Applied Statistics* **23**: 414–9

Chung L, Marden J I 1991 Use of nonnull models for rank statistics in bivariate, two-sample, and analysis-of-variance problems. *Journal of the American Statistical Association* **86**: 188–200

Chung L, Marden J I 1993 Extensions of Mallows' ϕ model. In: Fligner M A, Verducci J S (eds.) *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, New York, pp. 108–39

Critchlow D E 1985 *Metric Methods for Analyzing Partially Ranked Data. Lecture Notes in Statistics, Vol. 34*. Springer-Verlag, New York

Critchlow D E, Fligner M A, Verducci J S 1991 Probability models on rankings. *Journal of Mathematical Psychology* **35**: 294–318

Daniels H E 1950 Rank correlation and population models. *Biometrika* **33**: 129–35

David H A 1988 *The Method of Paired Comparisons*, 2nd edn. Oxford University Press, New York

Diaconis P 1988 *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics, Hayward

Feigin P, Cohen A 1978 On a model for concordance between judges. *Journal of the Royal Statistical Society B* **40**: 203–13

Fishburn P C 1994 On 'choice' probabilities derived from ranking distributions. *Journal of Mathematical Psychology* **38**: 274–85

Fligner M A, Verducci J S 1986 Distance-based ranking models. *Journal of the Royal Statistical Society B* **83**: 859–69

Fligner M A, Verducci J S 1988 Multi-stage ranking models. *Journal of the American Statistical Association* **83**: 892–901

Fligner M A, Verducci J S (eds.) 1993 *Probability Models and Statistical Analyses for Ranking Data. Lecture Notes in Statistics, Volume 80*. Springer-Verlag, New York

Henery R J 1981 Permutation probabilities as models for horse races. *Journal of the Royal Statistical Society B* **43**: 86–91

Henery R J 1983 Permutation probabilities for gamma random variables. *Journal of Applied Probability* **20**: 822–34

Huh M-H, Critchlow D E, Verducci J S, Kiecolt-Glaser J, Glaser R, Malarkey W B 1995 A symmetric analysis of paired rankings with application to temporal patterns of hormonal concentration. *Biometrics* **51**: 1361–71

Luce R D 1959 *Individual Choice Behavior*. Wiley, New York

Mallows C L 1957 Non-null ranking models. I. *Biometrika* **44**: 114–30

Mann H B 1945 Non-parametric tests against trends. *Econometrica* **13**: 245–59

Marden J I 1995 *Analyzing and Modeling Rank Data*. Chapman and Hall, London

Marley A A J 1968 Some probabilistic models of simple choice and ranking. *Journal of Mathematical Psychology* **5**: 311–32

Marley A A J 1982 Random utility models with all choice probabilities expressible as 'functions' of the binary choice probabilities. *Mathematical Social Sciences* **3**: 39–56

Marley A A J 1993 Aggregation theorems and the combination of probabilistic rank orders. In: Fligner M A, Verducci J S (eds.) *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, New York, pp. 216–40

Mosteller F 1951 Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika* **16**: 3–9

Pendergrass R N, Bradley R A 1960 Ranking in triple comparisons. In: Olkin I, Ghurye S G, Hoeffding W, Madow W G, Mann H B (eds.) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Palo Alto, CA, pp. 331–51

Plackett R L 1975 The analysis of permutations. *Applied Statistics* **24**: 193–202

Savage I R 1956 Contributions to the theory of rank order statistics—the two-sample case. *Annals of Mathematical Statistics* **27**: 590–615

Savage I R 1957 Contributions to the theory of rank order statistics—the 'trend' case. *Annals of Mathematical Statistics* **28**: 968–77

Schulman R R 1979 Ordinal data: An alternative distribution. *Psychometrika* **44**: 3–20

Smith L A, Evans D A 1991 Estimating a population distribution of sequences of k items from cross-sectional data. *Applied Statistics* **40**: 31–42

Spearman C 1904 The proof and measurement of association between two things. *American Journal of Psychology* **15**: 72–101

Stern H S 1990 Models for distributions on permutations. *Journal of the American Statistical Association* **85**: 558–64

Thurstone L L 1927 A law of comparative judgement. *Psychological Reviews* **34**: 273–86
Yellott J 1977 The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology* **15**: 109–44
Yellott J 1997 Preference models and irreversibility. In: Marley A A J (ed.) *Choice, Decision, and Measurement: Essays in Honor of R. D. Luce*. Erlbaum, Hillsdale, NJ, pp. 133–53

D. E. Critchlow

Rape and Sexual Coercion

The following summarizes the biological, and cultural approaches to the understanding of rape in human societies and evaluates these approaches by reference to studies demonstrating the incidence and social correlates of sexual coercion in the United States and cross-culturally.

1. The Biological View

The biological explanation for rape and sexual coercion is grounded in the argument that sexual aggression is an evolved adaptation in human males. The argument builds on Charles Darwin's theory of natural and sexual selection. Natural selection is the doctrine that in the struggle for existence evolutionary progress is achieved by the inheritance of advantageous characteristics which prosper at the expense of less advantageous ones. Sexual selection is a specific case of natural selection applied to the evolution of sex differences. Darwin claimed that the greater size and strength of human males was due 'in chief part to inheritance from his half-human male ancestors.' According to Darwin such characteristics 'would have been preserved or even augmented during the long ages of man's savagery, by the success of the strongest and boldest men, both in the general struggle for life, and in their contest for wives.' He measured success by the number of children some men left as compared with their 'less favoured brethen' (Darwin 1936, pp. 872–3).

In the nineteenth and early twentieth centuries, the Darwinian doctrine of sexual selection was applied to treatises on human sexual behavior produced by the famous sexologists R. F. von Krafft-Ebing and Havelock Ellis. In his magnum opus, *Psychopathia Sexualis*, first published in German in 1886, Krafft-Ebing argued that 'gratification of the sexual instinct [is] the primary motive in man as well as in beast' (quoted by Sanday 1996 p. 125). In the third volume of his *Studies in the Psychology of Sex*, published in Philadelphia between 1897 and 1910, Ellis comes very close to making rape the *sine qua non* of sexual arousal. Ellis conceives of human sexual behavior as a game of combat. Playing the role of hunted animal the female conceals her sexual passion by adopting a demeanor of modesty in order that the male may be more ardent and forceful. Ellis claims that as the hunt becomes more sexually charged, 'an element of real violence, of undisguised cruelty' is introduced. Accepting the Darwinian theory of natural selection with its emphasis on competition and brute strength, Ellis claimed that a woman who resisted 'the assaults of the male' aided natural selection 'by putting to the test man's most important quality, force' (quoted by Sanday 1996, pp. 127–8) (see *Male Dominance* for more on the social meaning of sex differences in Western thought).

At the end of the twentieth century, the Darwinian approach was again applied to the subject of the 'biological bases of sexual coercion' by Randy Thornhill and Craig T. Palmer in their book *A Natural History of Rape* (2000). Like Krafft-Ebing and Ellis they make a distinction between sexual desire in human males and females. Citing the work of anthropologist Donald Symons (1979, pp. 264–7), these authors claim that rape is due to certain adaptations making males more easily aroused, with a greater sex drive, a reduced ability to abstain from sexual activity, in need of greater sexual variety, and less discriminating in choice of sexual partner (Thornhill and Palmer 2000, p. 62). However, while Symons suggests that rape is a by-product of such adaptations, Thornhill and Palmer suggest that rape is itself an adaptation, which they call 'the human rape-adaptation' (Thornhill and Palmer 2000, p. 62). By this they mean that rape has a direct biological basis rather than being a behavioral consequence of other evolved behaviors.

To support their argument for the evolved nature of such an adaptation, Thornhill and Palmer (2000, p. 62) say it is necessary to identify 'mechanisms involved in rape that were designed by selection in the past specifically for reproduction by means of rape'. Because such mechanisms exist in the morphology of certain insects, they seem to suggest, they must also exist in humans. To make their case, the authors draw parallels with scorpion flies, which they point out have a physical adaptation for rape. Male scorpion flies have an organ, 'a pair of clamp-like structures, one on either side of the penis,' that serves to keep unwilling females in a mating position (2000, p. 63).

Admitting that there is no 'conspicuous morphology that might be a rape adaptation' in human males, Thornhill and Palmer (2000, pp. 64–5) look 'to the male psyche for candidates for rape adaptations.' They propose (2000, pp. 65–6) a number of 'psychological mechanisms' as 'adaptations' that can be viewed as 'analogous' to the rape adaptations observed in 'male insects'.

In a review of the argument put forward by Thornhill and Palmer, Frans B. M. de Waal, a prominent biologist, points out that they have not demonstrated

that rape in humans is a product of natural selection. According to Waal, 'for natural selection to favor rape, rapists would have to differ genetically from nonrapists and need to sow their seed more successfully, so to speak, causing more pregnancies than nonrapists, or at least more than they would without raping' (Waal 2000, p. 24). 'Not a shred of evidence for these two requirements is present' Waal concludes (2000, p. 24). Waal also points out that the psychological mechanisms Thornhill and Palmer propose do not demonstrate natural selection but have to do 'with judgment of people and situations, a multipurpose capacity also present in women' (2000, p. 24). Waal's emphasis on judgement leads us to the cultural argument.

2. *The Cultural Approach*

Against a strictly biological explanation of human behavior anthropologist Clifford Geertz (1973, p. 46) argues that without culture patterns (i.e. 'organized systems of significant symbols') human behavior would be shapeless and without direction, 'a mere chaos of pointless acts and exploding emotions.' Applied to sexual behavior, the cultural approach, accepted by many anthropologists and sociologists, suggests that sexual behavior is part of a broader cultural system which includes notions of masculinity and femininity, ideas about appropriate sexual relations, cues for sexual arousal, and messages about the consequences of sexual expression. As anthropologist Malinowski said long ago, '[s]ex in its widest meaning … is rather a sociological and cultural force than a mere bodily relation of two individuals' (Malinowski 1929, p. xxiii).

Peggy Reeves Sanday (1981) puts Malinowski's claim to the test in a study of the sociocultural context of rape cross-culturally. Using a standard sample of band and tribal societies, Sanday found that 47 percent of the societies for which there was adequate information showed a low frequency of rape, compared with 18 percent which were unambiguously 'rape prone.' Commenting on the pattern of correlations revealed by her study, Sanday (1981, p. 5) concludes that rape by males is part of 'a cultural configuration' which includes interpersonal violence among men as well as between men and women and male social dominance. Rape-prone societies are more likely to be characterized by male segregation in separate houses, an ideology of male toughness, emphasis on competition, low respect for women as citizens, and the absence of women in the public domain of economic and political affairs. In the more rape-free societies, there is an ethos of cooperation and consensus in human affairs. The social separation of the sexes is less marked and both sexes are more likely to hold exalted positions in public decision-making and to be integrated and equal in the affairs of everyday life.

Anthropologist Carole Vance (1984, pp. 7–8) also cites the diversity in human sexual practices to suggest that biology does not 'directly or simply determine the configuration or meaning of sexuality'. If this were the case, she argues, we would encounter uniformity cross-culturally rather than the startling diversity that actually exists in which 'activities condemned in one society are encouraged in another, and ideas about what is attractive or erotic, or sexually satisfying or even sexually possible vary a great deal.' Vance concludes that 'the body and its actions [must be] understood according to prevailing codes of meaning.' She suggests that an important question for research must be 'What is the nature of the relationship between the arbitrariness of social construction and the immediacy of … bodily sensations and functions?' (Vance 1984, pp. 7–9; see also Caplan 1987, pp. 1–27 on the cultural construction of sexuality).

In a study of the sociocultural and historical contexts of rape in the United States, Sanday (1996, p. 26) coins the term 'sexual culture' to describe the system of meanings, standards, and expected behaviors by which male and female sexuality is judged and understood. She argues that historically theories like those of Darwin, Krafft-Ebing, and Ellis functioned as cultural templates for popular sexual stereotypes by means of which males and females judged appropriate sexual behavior and acted accordingly. Because these stereotypes condoned sexual aggression by making it 'natural,' such as seen in the idea that 'boys will be boys,' US rape laws applied primarily to the use of force by a stranger. This changed after feminist activism in the 1970s resulted in more stringent legal codes. The same activism initiated studies on the incidence and social correlates of rape in America showing that acquaintance rape was much more common than stranger rape (Sanday 1996, pp. 161–207).

3. *Rape Statistics in the United States*

US statistics come from numerous studies utilizing national, community, or college samples. In general, these studies define rape as nonconsensual sexual intercourse irrespective of whether the parties know one another. Nonconsent is defined using the modern legal definition of rape. In most states, rape is defined as nonconsensual sex either due to the use of force or taking advantage of a person's inability to consent because of intimidation or because of alcohol or drugs. Although most studies focus on heterosexual rape, some authors note that rape also affects same-sex couples.

The first study designed to measure the percentage of those claiming they had been raped in a specified sample population was conducted in 1957 by sociologist Eugene Kanin on a college campus. This survey revealed that 55 percent of the 291 college

women interviewed said they had experienced offensive episodes 'at some level of erotic intimacy.' Twenty-one percent said they were offended 'by forceful attempts at intercourse' and 6.2 percent by 'aggressively forceful attempts at sex intercourse in the course of which menacing threats or coercive infliction of physical pain were employed.' Of the latter group 48 percent told no one and none reported to anyone in authority (Kirkpatrick and Kanin 1957, pp. 53, 56).

Another well-known path breaking study was conducted by Diane Russell in 1978 in San Francisco. Russell interviewed 930 women ranging in age from 18 to 80 of diverse social classes and racial/ethnic groups. Twenty-four percent of the women interviewed reported at least one completed rape, and 31 percent reported at least one attempted rape. The majority experienced rape by acquaintances and a much smaller percentage by strangers (Russell 1984, pp. 34–8).

Two large-scale national studies yielded similar results. In 1985 Mary Koss joined with *Ms.* magazine to survey a national sample of 6,159 students on 32 college campuses. This study found that many more women and men were having sex in the 1980s compared to the figures reported by Alfred Kinsey in his landmark studies of male and female sexual behavior published in the 1940s and 1950s. One-quarter of the men interviewed in the Koss–*Ms.* study reported involvement in some form of sexual aggression, ranging from unwanted touching to rape. A high percentage of the males did not name their use of force as rape. Eighty-eight percent said it was definitely not rape. Forty-seven percent said they would do the same thing again (summarized by Sanday 1996, pp. 191–2; see also Koss 1988, 1992, Warshaw 1988).

One in four of the women surveyed by Koss said yes to questions asking about behaviors that in most states would be legally defined as rape or attempted rape (for this definition see above). The power of popular stereotypes for defining acceptable sexual behavior and the adversarial nature of these stereotypes is reflected in the fact that most of the women reporting incidents that would legally qualify as rape did not call it rape. The term 'date rape' evolved from this study because many of the women surveyed said that the behavior often occurred on dates. Although most did not call it rape, Koss reported that 'the great majority of rape victims conceptualized their experience in highly negative terms and felt victimized whether or not they realized that legal standards for rape had been met' (Koss 1992, pp. 122–6).

By the end of the 1990s a number of scientifically designed research studies revealed similar statistics (for a summary of studies see Koss 1993 and Sanday 1996, pp. 184–206). The findings reported in these studies are corroborated by a national study of sex in America conducted in 1992 employing a representative probability sample of 3,432 Americans (see Michael et al. 1994). The findings fall within the range of many other studies: '22 percent of women were forced to do something sexually at some time,' while 'just 2 percent of men were forced.' The authors of this study concluded that 'as many as one in five women' experience some form of sexual coercion, including rape, in the United States (Michael et al. 1994, pp. 33, 221–2).

4. *The Social Correlates of Rape in US Studies*

A number of studies conducted late in the twentieth century indicate that the adversarial view of male–female sexual interaction described by the early sexologists is frequently held by sexually coercive males. In a study of a large southeastern university, Boeringer (1996, pp. 137–9) reports that 56 percent of the males he interviewed admitted to obtaining sex by verbal harassment (i.e., 'threatening to end a relationship unless the victim consents to sex, falsely professing love, or telling the victim lies to render her more sexually receptive'). One-quarter of the males reported using drugs or alcohol to obtain sex and 9 percent reported at least one use of force or threatened force to obtain sex. Boeringer concludes that such verbally coercive tactics suggest an 'adversarial view of sexuality in which one should use deceit and guile to win favors from a woman' (p. 140).

Sexually aggressive men, from convicted rapists to college males answering questions on social surveys, share a remarkably similar set of attitudes. Most believe that sexual aggression is normal, that sexual relationships involve game playing, that men should dominate women, that women are responsible for rape, and that relations between the sexes are adversarial and manipulative on both sides (Koss and Leonard 1984, pp. 221, 223). Reanalyzing Koss' data to pinpoint attitudes held by the self-admitted sexually aggressive men of her study (i.e., those who admitted to forcing a woman to have sex in Koss's 1985 questionnaire), Sanday (1996, pp. 196–7) found that these men often expressed adversarial-like beliefs. For example, many stated that they believe that a woman's No means Yes and that women say no to intercourse because they don't want to seem loose but really hope the man will force her. Compared with men who do not admit to forcing a woman, more of these men also think that being roughed up by a man is sexually stimulating to women and that women have an unconscious wish to be raped. Few of the women in Koss' sample hold such attitudes.

To conclude, there is considerable evidence suggesting that whatever the biological basis of male sexual aggression might be, family background and cultural context make a difference. The incidence of rape is far lower among men raised in stable, peaceful environments be it a family context in a complex society like the United States or the intimate communities of small-scale traditional societies in which the ethos is one of mutual respect between the sexes.

To say that some men will rape in all societies and to use this fact to make generalizations about the sexual predilections of all men may bring about the very sexual culture one is trying to avoid by creating a lore that makes sexual aggression a 'natural' part of masculinity. More important for the education of young males is the conclusion based on studies in the US and cross-culturally that most men in most societies do not rape, and those who do face social rejection or incarceration.

5. *Prevention*

Examples of prevention strategies based on the biological approach are offered by Thornhill and Palmer (2000, pp. 169–88). Believing that rape occurs 'in all the environments in which humans societies [sic] have been known to exist' (p. 171), they are less interested in changing the environment by challenging popular stereotypes than with focusing on the 'exact nature of the psychological mechanisms that guide male sexual behavior' (p. 172). They offer a number of suggestions including certain types of educational programs. Citing the view of rape proposed by Camille Paglia, claim Thornhill and Palmer (2000, p. 183) that since men 'have evolved sexual preferences for young and healthy women and are attracted to women who signal potential availability by means of dress and behavior,' women should be informed of the risk factors and use the information to lower their risk of rape by paying attention, for example, to dress and appearance. They also suggest more direct social interventions such as separating vulnerable females from males.

The cultural approach argues for a change in rape laws and sexual stereotypes so that nonconsensual sex is treated unambiguously as a crime. Increasingly, this approach defines rape not just in terms of the use of force or drugs, but more broadly as failure to obtain verbal consent. The operative words of the new American sexual ideology are 'affirmative verbal consent,' 'freely given agreement,' and 'sexual communication.' Such modern adaptations would be acceptable to Darwin's view of natural selection because they promote sexual health and well being in an era of sexually transmitted diseases. One point left out of discussions of the biological bases of rape is the fact that as much as Darwin stressed competition, he also stressed cooperation or acting on behalf of what he called 'the general good or welfare of the community' (quoted by Sanday 1996, p. 288). Called 'altruism' by students of human evolution, some suggest that the primary selective pressure in human evolution was not for violent men but for those who were able to cooperate with women and other males in organized food gathering. Some anthropologists believe that this ability to cooperate was responsible for the evolution of human culture (Tanner 1981). In light of the ever increasing incidence of sexually transmitted diseases together with the evolution of nuclear weapons, one can also note that today there is a fine line between aggressive behavior that results in reproductive success and behavior that leads human populations instead to the brink of extinction.

See also: Domestic Violence: Sociological Perspectives; Family and Gender; Family Bargaining; Female Genital Mutilation; Feminist Legal Theory; Feminist Theory; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Reproductive Health; Gender and the Law; Gender Ideology: Cross-cultural Aspects; Heterosexism and Homophobia; Male Dominance; Masculinities and Femininities; Pornography; Prostitution; Rape and Sexual Coercion; Regulation: Sexual Behavior; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sex Offenders, Clinical Psychology of; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Perversions (Paraphilias); Sexual Risk Behaviors; Sexuality and Gender; Teen Sexuality; Treatment of the Repetitive Criminal Sex Offender: United States; Women's Health

Bibliography

Boeringer S B 1996 Influences of fraternity membership, athletics, and male living arrangements on sexual aggression. *Violence Against Women* **2**(2): 134–47

Caplan P (ed.) 1987 *The Cultural Construction of Sexuality*. Tavistock, London

Darwin C 1936 *The Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life and the Descent of Man and Selection in Relation to Sex*. The Modern Library, New York

Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York

Kirkpatrick C, Kanin E J 1957 Male sex aggression on a university campus. *American Sociological Review* **22**: 52–8

Koss M P 1988 Hidden rape: Sexual aggression and victimization in a national sample of students in higher education. In: Burgess A W (ed.) *Rape and Sexual Assault II*. Garland Publishing, New York, pp. 3–25

Koss M P 1992 Defending date rape. *Journal of Interpersonal Violence* **7**(1): 122–6

Koss M P 1993 Rape: Scope, impact, interventions, and public policy responses. *American Psychologist* **48**(10): 1062–9

Koss M P, Leonard K E 1984 Sexually aggressive men: Empirical findings and theoretical implications. In: Malamuth N M, Donnerstein E (eds.) *Pornography and Sexual Aggression*. Academic Press, New York, pp. 213–32

Malinowski B 1929 *The Sexual Life of Savages in North-Western Melanesia: An Ethnographic Account of Courtship, Marriage, and Family Life among the Natives of the Trobriand Islands, British New Guinea*. G. Routledge & Sons, London

Michael, R T, Gagnon J H, Laumann E O, Kolata G 1994 *Sex in America: A Definitive Survey*. Little, Brown and Company, Boston

Russell D E H 1984 *Sexual Exploitation: Rape, Child Sexual Abuse, and Workplace Harassment*. Sage, Beverly Hills, CA
Sanday P R 1981 The socio-cultural context of rape: A cross-cultural study. *Journal of Social Issues* **37**(4): 5–27
Sanday P R 1996 *A Woman Scorned: Acquaintance Rape on Trial*. University of California Press, Berkeley, CA
Symons D 1979 *The Evolution of Human Sexuality*. Oxford University Press, New York
Tanner N M 1981 *On Becoming Human*. Cambridge University Press, Cambridge, UK
Thornhill R, Palmer C T 2000 *A Natural History of Rape: Biological Bases of Sexual Coercion*. MIT Press, Cambridge, MA
Vance C S 1984 Pleasure and danger. In: Vance C (ed.) *Pleasure and Danger: Exploring Female Sexuality*. Routledge and Kegan Paul, Boston, pp. 1–28
Warshaw R 1988 *I Never Called It Rape: The Ms Report on Recognizing, Fighting and Surviving Date and Acquaintance Rape*. Harper & Row, New York
de Waal F B M 2000 Survival of the rapist. *NY Times Book Review* **April 2**: 24–25

P. R. Sanday

Rational Choice and Organization Theory

Rational Choice Theory (RCT) and Organizational Theory (OT) are two different, yet—in the works of some neoclassical economists, rational choice sociologists, and economic organizational theorists—closely related theories. Rational choice theory is used by economic organizational theorists as a method of aggregation. Specifically, it is used when it is necessary to link the change in actions of individuals with change in characteristics of organizations. However, organizational analysts generally take a broader view in which organizational context, organizational structure, and individual actions interact and move together to change organizational functioning. Such a perspective does not rest on the assumption that aggregated individual choices explain organizational behavior. In some cases, organizational theory explains how individual actions articulate with structure. In other cases, organizational theory does not deal with individual actions. Rather, it deals with articulating relationships between and among organizational and contextual characteristics.

The purpose of this analysis is to compare RCT with OT. Rational choice theory is based on the premise of self-interested utility maximizing actions/choices of individuals. Organizational theory is based on the premise of organizational rationality, the efficient functioning of organizations through means/ends relationships. Organizational economic theory, which consists of transactions cost economics and agency theory, forms an indirect link between the RCT and OT. First, differences by defining the origin and premises of RCT and premises of Organizational Theory are elaborated. This is followed by demonstrating the link between the two, as developed in the premises of transaction cost analysis and agency theory. Although transaction theory holds to the premises of organizational efficiency, it finds the premise of self-interest utility maximization problematic. On the other hand, while agency theory holds the premise of self-interested, utility maximizing individual choice of RCT, and finds the guiding premise of organizational theory, the efficient functioning of firms, to be problematic. Thus, even the economic organizational theories, which purport to link RCT and OT are able to do so only if they accept either the RCT premises or the OT premises. As one set of premises are assumed, the other becomes problematic and becomes the questions to be examined in the analysis. No organizational theory has successfully integrated RCT and OT.

1. *Rational Choice Theory: Origin and Premises*

Rational Choice Theory originates in the British utilitarian tradition, via neo-classical economics. It is the core of classical economic theory. Although often qualified beyond recognition to meet conditional constraints of organizational analysts, it has been adopted in sociology (social choice) and in political science (public choice, game theory). In organizational theory and analysis, rational choice is most often used for the development of economic models of organizations. Transaction cost economics (TCE) and agency theory (AT) are the dominant economic organizational perspectives.

The premises at the basis of RCT as derived from neoclassical economic theory are: (a) self-maximizing individual utility driven actors, unconstrained by others and social norms (Coleman 1990, p. 503); (b) actors possess complete information and knowledge of their tastes, existing resources, the resources of others and market characteristics of the products and services they seek; (c) actors calculate and recalculate the changing conditions and act rationally; (d) however, preferences do not change, they are stable, and they are givens often implied from the very behavior they are to predict; (e) actors are not emotional, irrational, impulsive, habitual in their choices, they are rational; (f) interaction is defined as an exchange between buyer and seller achieved when utility and cost converge in a market transaction of supply and demand equilibrium.

Rational choice theory (RCT) is an axiomatic system that provides a basis for making individual decisions/choices. The axioms of RCT define rational choice on the basis of choices between pairs of alternatives, which meet the axiomatic requirements of: consistency, transitivity, independence, continuity,

and monotonicity. Consistency demands that it be possible for all the decision maker's options to be ranked. The assumption of consistency is also called the assumption of connectedness. Preferences must be either equal or unequal. Unequal preferences can be rank ordered for comparison across the decision maker's entire set of preferences. Transitivity is the assumption that if A is preferred to B, and B is preferred to C, then consistency requires that A is preferred to C (Green and Shapiro 1994, pp. 14–15). The distance between preferences or magnitude of preferences does not have to be known to the subject or the analyst. Independence assumes that all preferences are independent of all other preferences. The continuity assumption is that preferences hold across time and space. Rational choice theorists assume monotonicity—all decision rules and preferences are stable across individuals and time. Allowing for assumptions of interpersonal variation produces enormous tractability problems (Storm 1990, p. 126). Another assumption of RCT is that individuals 'maximize the expected value of their payoff measured on some utility scale' and that, usually, 'numerical probabilities can be attached to different eventualities' (Luce and Raiffa 1957, p. 50). Eventuality is calculated as the value of the outcome 'weighted by the probability of achieving it' (Elster 1986, Harsanyi 1986).

2. *Organizational Theory: Origin and Premises*

Rational choice theory and organizational theory are brought together by the term rational. However, the term has different meanings (Zey-Ferrell 1981) for RCT and OT. RCT is based on the idea that an individual seeks to maximize his/her utilities as they conduct organization business. The central tenant of rational choice is that individuals at all levels of organizations act selfishly based on their own preferences to maximize their utility. This is rational individual behavior. OT defines rational as the way in which economic organizations are designed—the most efficient method of organizing. This is rational organizational behavior.

Rationality in OT originates in Weber's ([1922] 1968) concept of 'rational bourgeois capitalism,' which connotes the systematic organization of ideas, people, and resources in the interests of instrumental efficiency and effectiveness. Weber wrote that rationality and its accompanying bureaucratic organizational form, based on formalization of the rule of law and differentiation of functions, pervaded not only the modern economy, but public administration, the legal system, and religions. The rationalization of bureaucratic structures led to increased efficiency. In Weber's model, organizational rationality and rationalization are essentially the same. This is not a (micro-level) individual process as it is in RCT. It is an organizational process or specification of mean/ends relationships based on systematic calculation and recalculation to achieve the most efficient organizational functioning. Subsequent analyses of organizations have assumed this rationality is the basis of organizations and developed theories around this premise.

3. *Divergences of Rational Choice and Organizational Theories*

In addition to definitional divergence of the term rational, the most apparent divergence is the difference in the unit of analysis. RCT attempts to explain individual choice behavior. RCT is based on the utility maximizing behavior of individual choice, while OT is based on the assumption of efficiency seeking organizational design. OT explains behavior of organizations. The utility maximizing behavior of individuals and the assumption of efficiency-seeking organizational design are similar, but are at different levels of analysis. If the unit is the individual, the utility being maximized is individual owner or employee self-interest. If the unit is the organization, the utility being maximized is profitability. Profitability of the firm does not always translate into income or other self-interests of firm members. Likewise, rational efficient organizational functioning does not always maximize the self-interest of organizational members.

Economic organizational theorist define organizations as firms doing business in markets. Neoclassical economists begin with market assumptions and work backward to firms and individuals assuming that actors operating in firms act in ways consistent with the theory of the market. However, assumptions that hold for RCT do not hold for organizational actors. Let us take just one assumption, that of assuming that all actors have full information about costs. Because each of them knows the cost and adjusts the quantity of their purchase, it is reasonable to assume that the market will settle on an equilibrium. In organizations these assumptions are false. Many types of decisions are not afforded to each actor. The choices available to each actor are not the same. Some actors' choices impose constraints on the choices available to others. Some actors do not know that a particular choice exists. For example, information about pending mergers and acquisitions. During mergers and acquisitions, parts of corporations maintain secrecy, not only from other firms, but from other units of their own firm and individual within their own units (Zey 1993, 1999). For example, securities firms are required by law to construct walls of secrecy, 'Chinese Walls,' between M&A units and traders to prevent insider trading. Information is also kept secret to prevent oppositional reactions to present and new allocations of organizational resources. Information is not equally distributed among or within organizational units for deliberate power-based reasons. Finally, the norms of

Table 1
Rational Choice Theory (RCT) and Organizational Theory (OT): Ideal type contrasts

	Rational choice theory	Organizational theory
Intellectual Origins	Neumann and Morgensterns Milton Freideman	Max Weber and Herbert Simon
	Oliver Williamson Michael Jensen	
Unit of Analysis	Individual	Organization
Content of Analysis	Preferences	Characteristics of organization and relationships to context
Human Nature	Self-interested	Self- and other interested
Rationality	Individual, self-interested Utility maximizer	Means/ends efficiency of functioning
Utility Maximization	Self-interests	Profits in firms
Human Preferences	Fixed hierarchy	Variable over time and place
Basis for Interaction	Assumed trust	Trust is problematic
Constraints	Economic actions are constrained by tastes and scarcity of resources	Organization actions are constrained by scaity but also by social structure, and historical context
Method	Predictive and explanation Not descriptive. Formal, mathematical model building no empirical data (clean hands)	Many methodologies comparative, historical based on data from empirical observations (dirty hands)

organizational rationality, efficiency dictates that various units within an organization should not receive the same information, but only what is necessary for their organizational entity. For all units to receive all information would be inefficient.

4. *The Link Between Rational Choice Theory and Organizational Theory*

Although rational choice theory and organizational theory are based on different premises and units of analysis, one type of theory, organizational economic theory, has attempted to link the two sets of premises and units of analysis. There are two dominant perspectives in organizational economic theory transaction cost economics and agency theory.

Economic organizational theory, both agency theory (AT) and transaction cost economics (TCE), begins with the relationships between individuals and organizations that are often conceptualized as events or transactions in which goods and services are exchanged (transaction cost analysis) and as contracts, which are agreements to conduct specific transactions (agency analysis). One is immediately aware that the unit of analysis is aggregated individuals. Transactions and contracts are aggregated either cross-sectionally or over time. In contrast, OT deals with the properties and characteristics of organizations, not individuals.

Coase (1937) from whom Williamson elaborates his TCE, explains the origin of firms as alternatives to inefficient markets, while organizational theorists observe that people create organizations to generate income and power, to create larger empires with which they are identified, for themselves or groups to which they belong. Mergers to create new organizations often take place to take advantage of new technologies, to take advantage of tax changes, or to reposition the organization in a growing market. These changes may have little to do with rational economic organization or inefficient market functioning.

5. *Convergence and Divergence of Economic Theories of the Firm and Organizational Theory*

Transaction cost economics (TCE) and agency theory (AT), generally labeled economic organizational theories, have attempted to link RCT and OT and their different units of analysis and methods of theorizing through a set of assumptions that aggregate utility-maximizing individual behavior so that their behavior promotes organizational efficiency (profitability). However, survey research has demonstrated that not everyone has the same (or even individually

consistent) utilities. Thus, the aggregation, specifically non-additive aggregation functions, of individual choices and action into organizational actions is not clearcut, because the complexity of the aggregation process often violates RCT assumptions.

Furthermore, the two economic organizational theories, AT and TCE, do not operate under the same premises. AT is based on the utility-maximizing individual of RCT (Jensen 1983, 1993, Jensen and Meckling 1976), while transaction cost economics assumes that humans are boundedly rational at best. AT attempts to explain why organizations are more or less efficient while TCE assumes that organizations are an efficient alternative to inefficient markets (Williamson 1975, 1985, 1996). The utility-maximizing behavior of organizations is efficiency-seeking organizational design. If the unit of analysis is the organization, and the utility being maximized is profit, then efficiency maximization is implied when income seeking is not compromised. Organizational theorists question these assumptions because organizations, even firms, are never constraint free and individuals within organizations do not operate under the same constraints.

Agency theory assumes that individuals are utility maximizers but does not assume that organizations are structured in the way they are because it is the most efficient method of organizing. This is demonstrated, not only in the assumption that managers and workers act as opportunistic agents—that is, not in the interest of the owners—but also by the preoccupations of agency theorists with governance and methods of controlling nonowners to maximize profits (returns to shareholders) to the owners (Jensen and Ruback 1983). Agency theorists and economic organizational theorists in general do not integrate into their analyses historical changes, which provide opportunities and constraints on individual and organizational actions as organizational theorists do. Jensen's presidential address to the American Finance Association demonstrates that he had not until recently discovered Chandler's work (1962, 1977, 1990) and that organizational forms which are most efficient change over time with each historical context in which they develop.

Transaction cost economics assumes that organizations are structured in the way they are because it is the most efficient method of organizing, but TCE does not assume that individuals within organizations are utility maximizers. For example, in Williamson's work, *Markets and Hierarchies* (1975), he assumes that organizations more efficiently eliminate (by management fiat) opportunistic ('self-interested with guile') individual behavior than market mechanisms. TCE theorists analyze uncertainty and asset specificity of transactions. Individual utility maximization is problematic because of the obvious tendency of rational actors to act opportunistically under conditions of uncertain and asset specificity. TCE, like AT ignores historical changes and processes which constrain and provide opportunities for organizations. Both contextual and organizational changes are ignored. Organizational change, especially transformational change cannot be explained by economic theories of the firm (see Zey and Swenson 1998, 1999). This is problematic because of the major transformations of firms at the turn into the twenty-first century.

In both Williamson's transactions cost economics (1975) and in Jensen's agency theory (Jensen and Meckling 1976), the idea of self-interest in contractual arrangements between and within organizations is problematic. Ultimately, questions of reconciliation of differential self-interest in the absence of trust, is impossible. Trust, rules of fairness, norms of reciprocity, and conflict, are non-existent and irrelevant in RCT and economic theories of the firm. They are essential to organizational formation and perpetuation.

Neither TCE or AT holds both the basic premise of RCT, self-interested, utility maximizing individuals, and the basic premise of OT, that firms are designed in the most efficient form of organizational design. Agency theory holds the former, while TCE holds the latter. Each finds the others basic premises problematic and grist for the mill of organizational analysis. Thus, economic organizational theory has not yet integrated rational choice theory and organizational theory. This author does not think it is possible to do so because of conflicting premises and different units of analysis.

See also: Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision Making: Nonrational Theories; Firm Behavior; Game Theory; Hierarchies and Markets; Institutional Economic Thought; Market Structure and Performance; Rational Choice Theory: Cultural Concerns; Transaction Costs and Property Rights; Utilitarian Social Thought, History of; Utilitarianism: Contemporary Applications; Weber, Max (1864–1920)

Bibliography

Chandler A D Jr. 1962 *Strategy and Structure: Chapters in the History of the Industrial Enterprise*. MIT Press, Cambridge, MA

Chandler A D Jr 1977 *The Visible Hand: The Managerial Revolution in American Business*. Belknap Press, Harvard University Press, Cambridge, MA

Chandler A D Jr. 1990 *Scale and Scope: The Dynamics of Industrial Capitalism*. Belknap Press, Harvard University Press, Cambridge, MA

Coase R 1937 The nature of the firm. *Economica* **4**: 386–405

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press, Harvard University Press, Cambridge, MA

Elster J (ed.) 1986 *Rational Choice*. New York University Press, New York

Green D P, Shapiro I 1994 *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. Yale University Press, New Haven, CT

Harsanyi J C 1986 Advances in understanding rational behavior. In: Elster J (ed.) *Rational Choice*. New York University Press, New York
Jensen M 1983 Organizational theory and methodology. *Accounting Review* **55**(2): 319–39
Jensen M 1993 Presidential address: The modern industrial revolution, exit, and the failure of internal control systems. *Journal of Finance* **48**(3): 225–64
Jensen M, Meckling W 1976 Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* **3**: 305–60
Jensen M, Ruback R S 1983 The market for corporate control. *Journal of Financial Economics* **11**: 5–50
Luce D R, Raiffa H 1957 *Games and Decisions: Introduction and Critical Survey. A Study of the Behavioral Models Project* Bureau of Applied Social Research, Columbia University. Wiley, New York
Storm G S 1990 *The Logic of Lawmaking: A Spatial Theory Approach*. Johns Hopkins University Press, Baltimore, MD
Weber M [1922] 1968 *Economy and Society*. Roth G, Wittich C (eds.). Bedminster Press, New York
Williamson O E 1975 *Markets and Hierarchies: Analysis and Antitrust Implications: A study in the Economics of Internal Organization*. The Free Press, New York
Williamson O E 1985 *The Economic Institutions of Capitalism: Firms, Market, Relational Contracting*. Free Press, New York
Williamson O E 1996 *The Mechanisms of Governance*. Oxford University Press, New York
Zey-Ferrell M 1981 Criticisms of the dominant perspectives in organizations. *The Sociological Quarterly* **22**: 181–205
Zey M 1993 *Banking on Fraud: Drexel, Junk Bonds, and Buyouts*. Aldine de Gruyter, Hawthorne, NY
Zey M 1998 *Rational Choice Theory and Organizational Theory: A Critique*. Sage, Thousand Oaks, CA
Zey M 1999 The subsidiarization of the securities industry and the organization of securities fraud networks to return profits in the 1980s. *Work and Occupations* **26**: 50–76
Zey M, Swenson T 1998 Corporate tax laws, corporate restructuring, and subsidiarization of corporate form, 1981–95. *The Sociology Quarterly* **39**: 555–82
Zey M, Swenson T 1999 The transformation of the dominant corporate form from multidivisional to multisubsidiary: The role of the 1986 Tax Reform Act. *The Sociological Quarterly* **40**: 241–67

M. Zey

Rational Choice Explanation: Philosophical Aspects

Rational choice explanation is the progeny of the Scottish Enlightenment and neoclassical economics. It is essentially reasoning from economic assumptions of what actors value to reach conclusions about both individual actions and aggregate or collective-level outcomes. It is applied to virtually any significant behavior or social outcome, from political participation, crime, and drug addiction to institutional structures, legislative and interest group politics, law, norms, failures of collective action, and revolution. The most extensively developed parts of such theory are in the theory of elections and voting, the analysis of law under the rubric of law and economics, the study of group behavior, and game theory. Rational choice explanations gain their power from the attempt to reduce particular motivations to some narrower set of motivations, often simply to interests, which can be moderately well defined and measured across a large variety of contexts and actions.

The most common criticism of rational choice explanation is that it is false to the facts of actual motivations of people. This is a by-inspection claim. Because very much of actual behavior surely is motivated by interests, the dispute must actually be whether rational choice explanation is likely to help understand certain activities but not others. This is a question that cannot be answered in the abstract. Sometimes the criticism of rational choice trades on the fact that we commonly resolve many apparent choices by acting from habit or routine. As A. N. Whitehead quipped, civilization advances by the reduction of more and more choices to habit. A similar criticism of utilitarianism is that, if we spend as much time in calculating costs and benefits of our choices as would seem to be necessary to get our choices right, we will have no time left for life. J. S. Mill (1969) responded, rightly, that a sailor need not work out the positions of the stars anew each trip to sea but can simply rely on an astrolabe.

To illustrate the nature and power of rational choice explanation, as well as its plausible limits, I will address the aggregate-level phenomena of collective or group action, majority rule, and democracy. I will then address game theory and the value theory of rational choice. Finally, I will address the application of rational choice explanation to norms and social order.

1. *Fallacies of Composition*

A common criticism of rational choice explanation is that it cannot handle aggregate-level phenomena because it is fundamentally grounded in individual-level interests, choices, and actions. On the contrary, the chief successes of such theory have been to relate aggregate-level phenomena to individual-level actions. Several of the most important findings have been the—typically negative—demonstration that various aggregate-level claims are fallacies of composition from individual-level actions, choices, or values. It is a fallacy of composition to suppose that a collection of items has itself the characteristics of those items. This can be true, but it must be shown and not merely assumed. Consider three of these demonstrations.

The logic of collective action suggests that, although individuals may share a general interest in supplying themselves some collective benefit, they individually may have no interest in contributing effort or resources to the project. My contribution will buy a small part of

our good, of which an even smaller part will go to me, so that my cost of contribution will be larger than my benefit from it. All members of a group may therefore be individually worse off because they failed to take some action (Olson 1965, Hardin 1982). This is a problem that is pervasive in societies, ranging from very small societies with limited resources to very large and complex industrial societies.

The Arrow theorem (Arrow 1963, Barry and Hardin 1982) concludes that a collective utility or welfare function cannot be aggregated from individual utility functions if both are ordinal (see discussion of ordinal utility theory below). In essence, the theorem says that there cannot be a collective preference that has the simple features of an individual preference function. Among the implications of this theorem is that every standard choice procedure is potentially incoherent. Majority rule, for example, can produce cyclic preferences such that candidate A would defeat B by a majority, B would defeat C, and C would defeat A. Hence, simple majority rule is potentially indeterminate. It often works well enough—or seems to work—in practice, however, because there might not be cyclic preferences or because there are devices that block such preferences from being effective and even from being revealed. For example, plurality elections or the use of run-offs between the top two candidates in a multicandidate election can hide any cyclicities because there cannot be cycles over only two candidates.

The economic theory of democracy explains why candidates might not be substantially differentiated in certain common contexts and why voters often have little interest in voting. It also makes sense of the otherwise surprising level of ignorance of typical voters. In this latter case, however, there may be no fallacy of composition, because individual-level ignorance often translates into collective ignorance. Consider the problem of voting, as analyzed by Anthony Downs (1957). The chance that my vote would make a difference in an election decided by a large electorate is very nearly nil. In a 1974 statewide contest for senator in the very small state of New Hampshire, with somewhat more voters than Iceland, the vote was so close that errors in counting made it practicably undecidable who had won. No-one's single vote mattered. Suppose that the trouble to me of voting is such that I would sooner spend several dollars to vote by telephone to avoid going to the polls. The value to me of electing one candidate rather than another would have to be in the many thousands of dollars to offset my costs of voting, because the benefit would have to be discounted by the very low likelihood that my vote would make a difference. Yet many people do vote. Much behavior seems not to fit the rational choice theory and must be explained in other ways.

If it would not be worth my very slight effort to vote, Joseph Schumpeter (1950) concludes, then it would be even less worth the much more substantial effort it would take for me to be well informed about the candidates and issues at stake. Evidence from many national contexts suggests, in fact, that citizens are abysmally ignorant when they vote. Here rational choice explanations seem to fit well. Editorial campaigns against voter ignorance are often misframed as being about the irrationality of citizens, who are acting all too rationally.

2. *Game Theory*

Rational choice explanations are commonly based in game theory, neoclassical economics, or commonsense arguments such as those of Smith and Hume. Game theory is perhaps the framework that best characterizes the nature of most of the problems in rational choice theory, although many of the most important insights in such theory predate game theory, which originated roughly during World War II (Von Neumann and Morgenstern 1953). Although it is called a theory, game theory is little more than a descriptive framework on which many theories about particular or quite general phenomena are founded. That framework takes many forms, the most transparent of which is the matrix form in which each of the players or participants faces a set of possible strategy choices. Outcomes of a play of a game are jointly determined by the strategy choices of all the players.

This characterization of our individual choices—of strategies, not of outcomes—makes game theory virtually antideontological. Judging action types as right or wrong makes no sense in interactive contexts, which are the principal contexts for most of moral theory. An action theory that takes what are essentially individual choices of outcomes as central is almost irrelevant to choice in social contexts in which, typically, important outcomes necessarily are not under the control of individual deciders (Hardin 1988, pp. 68–70). An ethic of rules makes nonsense of choices of strategies, which are chosen because of their effect on outcomes, as these are influenced by the strategy choices of others. Game theory can readily be fitted to consequentialism and especially utilitarianism. Utilitarianism essentially generalizes the principle of rational choice to choosing the welfare of all instead of only the welfare of the individual chooser.

To fit utilitarianism to game theory is to revamp its central concern with outcomes as valued by their benefits and costs. In interactive social choice, we typically cannot say that a particular action is right or wrong in producing a good overall outcome. We must, rather, judge strategy choices and we must judge them contingently. A strategy choice is right only to the extent that, in interaction with the choices of others, it produces a good overall outcome. If the others with whom I am interacting are sometimes utilitarian, sometimes self-seeking, and sometimes rule-guided in their choices, I may have difficulty determining my

best strategy choice for producing a utilitarian outcome. This is not different from the difficulty that self-interested players have in determining the best strategy choices for themselves. Game theory is indeterminate in many contexts, and therefore both pragmatic and utilitarian choices in a real-world setting are apt to be indeterminate as well. Merely seeing the nature of interactive choice in descriptive game theory changes our understanding of the possibilities of rational choice theory and of utilitarianism.

3. *Value Theory: Interests, Consumptions, and Welfare*

Rational choice accounts are commonly grounded in interests. Sometimes, interests are the whole story of someone's motivations in a particular context. But typically, my interest in having more resources, such as money, is itself grounded in my desire to consume various things. Resources, capabilities, and opportunities matter to us only because they enable us to consume or experience various things, and consuming or experiencing these things is what constitutes our welfare. Resources are worthless otherwise. Focusing only on interests is therefore a shorthand for what the resources can enable us to do. Moreover, we could relate our consumptions to the overall welfare that they bring us. It would be a mistake to suppose interests translate smoothly into welfare or consumptions. Indeed, consumptions typically trade off against each other and against interests. If I invest much of my resources in a holiday, I may not have enough left for other things, such as a new refrigerator. Rational choice explanations involve all three of these categories—interests, consumptions, and welfare.

It would also be a mistake to suppose that my welfare is wholly self-interested. One of the things I might wish to consume or experience might be your happiness. One may say this is against my interests because it means using resources that I could use for my own happiness in some other, perhaps more direct, way. But this is a confusion in at least two ways.

First, the only point in having resources is to be able to consume in order to contribute to one's welfare, as is evident to everyone but a Scrooge for the resource of money. True, there is a conflict between using my resources for one consumption and saving them for another, but this is properly seen not as a conflict between consumptions and interests but as a conflict between different consumptions, or between different paths to welfare.

Second, there is no reason to suppose that, say, enjoying an opera, a good meal, or a lovely journey is a more direct or more successful route to enhancing my welfare with some bit of my resources than providing for my son's or lover's, or friend's happiness with that same bit of resources. More generally, my welfare is enhanced best by a mixture of consumptions and experiences, including contributions to the welfare of others, especially certain others. With those certain others, my welfare will depend on our sharing intentions to do various things for and with each other.

If we could suppose that economic interests as measured in resources were the primary issue for people, we could fairly easily explain actions as though people were all alike except for their circumstances. Rational choice explanations typically reduce the focus to interests, rather than other demands, such as demands for moral positions (ranging from deontological to utilitarian). In part they do so for methodological reasons. Analysis is more productive if we keep it clean and focused, and simplification is required for virtually all social analyses. But in part, the focus on interests is often basically right, especially for aggregate or collective issues. Other concerns matter to collective groups and societies, but they matter the way they do in life more generally. They often do not conflict with interests although, of course, occasionally they do and they may even trump interests in many contexts. For example, voters may vote against their personal interests in support of welfare programs and policies favoring ill-represented ethnic groups. But those who voted against the British Raj in India, French control of Algeria, or the US involvement in Vietnam may generally have voted with their interests as well as with moral concern. Still, many social and individual phenomena cannot readily be explained as the product of interest-seeking by individuals.

4. *Value Theory: Cardinal, Ordinal, Ramsey–Neumann Cardinal*

Apart from the structure or content of the value theory discussed above, there has been extensive debate over the form of such theory. Jeremy Bentham supposed that pleasures and pains during a day of someone's life could be defined in such a way that we could add the value of pleasures and subtract the value of pains for an overall measure of benefit to the person for that day. Today, we would speak of such a welfare or utility measure as cardinal in the sense that each experience one has could be assigned a utility number that could be added straightforwardly to other utilities.

Bentham's terminology was confused in that he spoke of the utility one gets from consuming some object and also of the utility inherent in the object. The latter notion underlies the labor theory of value that corrupted much of economic thinking throughout the nineteenth century and often still corrupts it today, in part under the doleful influence—on this matter—of Karl Marx. Bentham contributed to two opposing streams of value theory: that of the labor theory of value and that of the subjective theory that holds value to be strictly what a person gets out of an experience or consumption. The value to me of consuming some-

thing such as, say, beluga caviar might be far below the value to me of consuming other things that take equal expenditures of my resources. The value to you, even with the same overall level of resources, might be radically higher. Hence, you might consume a lot of beluga caviar and I might consume none after an initial taste. The subjective theory obviously has the disadvantage that we cannot always generalize from one person to others, although often we can generalize enough not to need to do in-depth studies of the tastes and preferences of everyone in order to explain overall behavior.

Bentham supposed further that the utility for one person could be added to that for another. This supposition would allow us to add up the utilities to all under one set of circumstances and to compare the total to that under another set of circumstances. Supposing we could do the measures, we could then say that, for example, one form of government is superior to another.

The so-called ordinal revolution in demand theory in economics overthrew all of the cardinal assumptions and set in their place a utility theory that is entirely subjective and that cannot be given interpersonal comparisons. The remarkable fact for economics is that such a limited theory clears up many prior problems and makes coherent sense of demand and supply relationships and, hence, prices in market exchange. Unfortunately, this revolution came only in the 1930s when moral theorists had ceased also to be economic theorists, as Smith, Hume, Mill, and Henry Sidgwick had been. It has therefore not had as much influence in general social theory as in economics, although it clearly should be brought to bear on the former as much as on the latter. A clear implication of interpersonally noncomparable value theory is that many traditional questions of interaction between people cannot be readily settled, so that moral and social theory must often be open on many issues.

In what might count as a minor sideline, Frank Ramsey (1931) and John von Neumann (Von Neumann and Morgenstern 1953) independently proposed a way to cardinalize preferences that are merely ordinal. Suppose you face a list of mutually exclusive outcomes, and suppose you can rank order any two outcomes. Assign a value of 0 to the least preferred, Z, and a value of 100 to the most preferred, A, outcomes in your list. For some outcome, K, between these two in your ranking, you can rank it against any lottery over A and Z. For example, suppose you would be indifferent to getting K and getting the combination of a p percent chance of getting A and a $100–p$ percent chance of getting Z ($0 < p < 100$). On the scale of 0 to 100, K has a cardinal value to you of p.

The Ramsey–Von Neumann cardinalization of utility (RN) clearly applies to personal decisions over outcomes. But we cannot use it to construct a Benthamite cardinal utility that is additive across persons, because there is no sense in which your 0 and my 0 are the same or your scale and mine, because each of them is arbitrary. If I assign a value of p and you assign a value of q to outcome K, we cannot meaningfully say that K has a joint value of $p+q$. Indeed, we can perform a linear transformation on your utility scale without affecting the results that follow from it. For example, we could divide the value of each outcome by 10 and add 37 to it, so that your scale would then run from 37 to 47, and this transformed scale would get all the results that your original scale from 0 to 100 would get. In addition, we can assign utility values to outcomes that fall outside the original ranking from A to Z, so that some outcomes might have negative utility and others might have extremely large positive values. This is suggested by the fact that the original normalization was performed over the arbitrary set of outcomes from A to Z. Hence, there is no meaning to any claim that your RN utility can be added to mine to yield an overall joint utility. The RN theory is therefore primarily of interest in individual decision making under risk, and is seldom of interest for more general social theory. It is irrelevant to Benthamite hopes for a cardinally additive utilitarianism.

5. *Norms*

There is a growing literature on the rational choice explication of norms. Edna Ullmann-Margalit (1977) gave rational reconstructions of norms for dealing with strategic contexts that are modeled by prisoner's dilemma, coordination, and partiality or unequal coordination games. In this categorization the descriptive apparatus of game theory is used to establish and characterize types of norms according to the strategic problems they resolve. The first of these norms essentially handle collective action problems because, without the force of the norms, many individuals would free ride, and any effort to provide a collective good would fail. The second handle such coordinations as the driving convention. That convention surely requires no norm to keep drivers in compliance with it, because the threat of grievous harm from violating the convention is sufficient. But many conventions are less regulated by such clear incentives and some of these might be more forceful if backed by norms. Norms in Ullmann-Margalit's third category handle unequal coordinations such as the division of labor by gender roles or social order in any hierarchical society in which obedience to norms might keep even those at the bottom of the hierarchy loyal and orderly. Again, these conventions are grounded in coordination incentives and they might be fully self-enforcing, but norms that moralize the inequalities might make those in the less well off positions less sullen in acquiescing in the arrangements.

More generally we may say that there are two broad categories of norms: norms of exclusion and universa-

listic norms (Hardin 1995, Chaps. 4 and 5). Norms of exclusion are specific to particular groups, which can enforce the norms by sanctioning violators within the group; the sanctions can go so far as to shun and even exclude members. The sanctioning is grounded in the interests of individual sanctioners, so that the norms are commonly interest based, although group members might claim they are inherently moral or right. A norm of exclusion can work to help enforce cooperativeness in a group or community, or to enforce a coordination that might serve the interests of some but not all. For example, a minority culture might enforce upon its members the use of a minority language or a particular dress code. Using that language or following that dress code might reduce the possibilities for some members of the community to gain employment outside the community. The community might protect its hegemony over deviant members by shunning those who violate the language or dress code.

Universalistic norms can be dyadic, as in promising or truth telling, or large-number in the sense of applying to actions toward the whole society, as in not littering. Dyadic norms are commonly enforced through the sanctions inherent in the iterated prisoner's dilemma. If you lie to me or break your promise to me, I can simply avoid further interaction with you, to your likely cost. If you break a norm that affects everyone, such as a norm against littering or in favor of voting, no particular individual is likely to have an interest in sanctioning you. Hence, such norms are likely to be relatively weak in their force and in our adherence to them. Oddly, therefore, it is the often odious norms of exclusion, such as racist and ethnic norms, that have strongest force, while universalistic norms are relatively weak unless they are played out essentially in dyadic interactions. Those who insist that, morally, norms should commonly trump interests, should be bothered that it is interests that trump any plausible moral content of many norms, and that norms not backed by interests are likely to be weak. Large-number universalistic norms may be backed primarily by moral or psychologically inculcated commitments. If so, then it is of interest whether they are, in fact, very effective in motivating people.

6. *Social Order*

There are three grand schools of social theory. One, associated with the French sociologist Émile Durkheim and many others, holds that society is held together by shared values. The second, associated with Karl Marx and Plato's *Thrasymachus*, holds that conflict is pervasive and that social order must therefore be maintained through the imposition of power to coerce when necessary. The third, associated with Adam Smith, holds that society is organized by and for exchange, which involves a combination of coordination and conflict. Each of these schools has many strong contemporary advocates. A fourth plausible view is that much of social life is simply organized by coordination (Hardin 1999, Chap. 1). We need not share particular values in order to share interests in avoiding conflicts with each other. The ideal type of this last form of social order is coordination on traffic rules, especially on the rule to drive right (or to drive left). We can be hostile to each other in many ways and still find it sensible to avoid needless conflict over many things, so that we would willingly be coordinated. A shared value theory is also a coordination theory, but its advocates ground it in value commitments and are generally hostile to the inclusion of interests among those value commitments. The conflict and exchange schools are commonly grounded in interests, although the socially destructive conflict that sparked much of early liberal thought was over commitment to divergent religious views.

The pre-eminent coordination theorists among classical political philosophers were Thomas Hobbes and David Hume. They were both protogame theorists in their grasp of the fundamental nature of various social interactions. While Hobbes is commonly thought to be a major conflict theorist, his actual resolution of the problem of social order is to arrange for coordination on an all-powerful sovereign, whose task is essentially to coordinate us on order so that we may construct our own separate lives and engage with one another in productive pursuits (Hobbes 1968). Hume's (1978, Book 3) articulation of a remarkable array of coordination problems qualifies him to be seen as the greatest coordination theorist, although his complex insights were largely neglected for two centuries, in part because he lacked the technical vocabulary of game theory and did not himself bother to work out the general principles of it. He did, however, work out a general moral and political theory that was substantially grounded in strategic understanding (Hardin 1988, Chap. 2).

In complex modern societies, coordination may be the central, dominant mode of social order. It does not require consensus on values—pluralists can coordinate on an order for the society in which they seek their diverse values. It does perhaps fit especially well with modern individualist liberalism that allows individuals to live by their own lights. Indeed, coordination is a simple version of Mill's principle that we should be free to pursue our own ends without interference so long as our doing so does not interfere with others (Mill 1977). When there is such interference, of course, other principles may be required to regulate our interactions.

7. *Concluding Remarks*

Rational choice theory is a diverse set of approaches to the study of society that are based in assumptions of individual rationality. Indeterminacies in such theory often mirror indeterminacies in social relations and

individual understandings of these. Indeed, many rational choice explanations have demonstrated the indeterminacy of social choice. Among the most important contributions of such theorizing is the descriptive clarity it often adds to our understanding of issues that have been poorly characterized before. Rational choice explanation is inherently about states of affairs, although it can include in a state of affairs concerns for anything that motivates people. It is therefore easy to see how it fits with utilitarianism.

Rational choice value theories, when introduced into the study of law, have led to the burgeoning movement of law and economics, which has produced the most systematic analysis of the broad sweep of law and legal rules that legal scholarship and practice have hitherto seen. The quick march of that movement—roughly beginning with the Coase theorem (Coase 1988)—through virtually all areas of law has provoked counterattacks, but none of these is itself grounded in a value theory systematic enough to permit more than piecemeal objections. The Coase theorem says, with audacious generality and simplicity, that if there are no transaction costs to impede economic cooperation between owners of various resources, then ownership will not affect production, although it presumably will affect how the profits from the production are divided. That theorem, along with earlier work of Ronald Coase and many others, has led to the massive study of transaction costs.

We have a rich understanding of the problems of rational choice value theories just because those theories have been a major focus in the development and articulation of the entire discipline of economics. The theories are a product of centuries of determined and often brilliant debate with dozens of major contributions. Alternative value theories have had far too few advocates and critics to yield much understanding at all. Indeed, value claims commonly seem to be *ad hoc* and not systematically generalizable. Such considerations are no argument for the superiority or rightness of any value theory. But they may be a tonic for those who wonder how the only well-articulated class of value theories is so easy to criticize: It has enough content to be subject to extensive and varied criticism.

See also: Decision Theory: Bayesian; Economics and Ethics; Economics, Philosophy of; Game Theory: Noncooperative Games; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Irrationality: Philosophical Aspects; Practical Reasoning: Philosophical Aspects; Prisoner's Dilemma, One-shot and Iterated

Bibliography

Arrow K J 1963 *Social Choice and Individual Values*, 2nd edn. Wiley, New York

Barry B, Hardin R (eds.) 1982 *Rational Man and Irrational Society?* Sage, Beverly Hills, CA

Coase R H 1988 The problem of social cost. In: Coase R H *The Firm, the Market and the Law*. University of Chicago Press, Chicago, pp. 95–156

Downs A 1957 *An Economic Theory of Democracy*. Harper, New York

Hardin R 1982 *Collective Action*. Johns Hopkins University Press, Baltimore, MD

Hardin R 1988 *Morality within the Limits of Reason*. University of Chicago Press, Chicago

Hardin R 1995 *One for All: The Logic of Group Conflict*. Princeton University Press, Princeton, NJ

Hardin R 1999 *Liberalism, Constitutionalism, and Democracy*. Oxford University Press, Oxford, UK

Hobbes T 1968 *Leviathan*. Penguin, London

Hume D 1978 *A Treatise of Human Nature*, 2nd edn. Oxford University Press, Oxford, UK

Mill J S 1977 On Liberty. In: Robson J M (ed.) *Collected Works of John Stuart Mill*. University of Toronto Press, Toronto, Vol. 18, pp. 213–310

Mill J S 1969 Utilitarianism. In: Robson J M (ed.) *Collected Works of John Stuart Mill*. University of Toronto Press, Toronto, Vol. 10, pp. 203–59

Olson M Jr 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA

Ramsey F P 1931 Truth and probability. In: Braithwaite R B (ed.) *The Foundations of Mathematics and Other Logical Essays by F. P. Ramsey*. Harcourt Brace, New York

Schumpeter J A 1950 *Capitalism, Socialism and Democracy*, 3rd edn. Harper, New York

Ullmann-Margalit E 1977 *The Emergence of Norms*. Clarendon, Oxford, UK

Von Neumann J, Morgenstern O 1953 *Theory of Games and Economic Behavior*, 3rd edn. Princeton University Press, Princeton, NJ

R. Hardin

Rational Choice in Politics

'Rational choice' (henceforth RC) is one of the principal approaches to political science as practiced in the USA. The approach has generated important insights relevant to a wide range of political phenomena, but it has also generated controversy as it moved into each of the various subfields. Much of the controversy arises from misunderstandings, as well as a failure to appreciate that RC is not a monolithic enterprise.

1. The Rational Choice Approach

The central premise uniting the various schools of RC is that behavior is purposive. Political behavior is not the product of psychological drives, childhood socialization, organizational norms, or other a-rational influences stressed by the various schools of behaviorists. Such influences certainly shape indi-

vidual preferences, but rational choice further posits that individuals pursue their goals as rationally as their knowledge, their resources, and the context permit. For example, much congressional scholarship of the 1950s and 1960s treated members of Congress as acting primarily in response to role expectations and norms established and enforced by other members. Rational choice scholars in the 1970s suggested that a more fundamental starting point would be the presumption that, first and foremost, members of Congress wanted to be re-elected—and acted accordingly. That premise would not describe all members all the time of course, but it would fit more members more of the time than any other assumption, and thus was the most appropriate starting point.

1.1 Origins of RC

Although RC shares with economics the maintained hypothesis of purposive behavior, RC is not a direct import from economics, as implied by critics who level charges of 'economic imperialism.' Some schools of RC owe much to economics, but others have independent roots. Many of the modeling tools first utilized in RC research were borrowed from operations research and other branches of applied mathematics. In particular, political scientists saw potential in game theory earlier than did economists, who continued to use mechanistic maximizing models for some years after RC practitioners in political science had gravitated toward strategic models. The noncooperative game theoretical revolution brought the two groups closer together, but present convergence should not mask the eclectic origins of the approach. Consider that the membership of the Committee on Nonmarket Decision Making (the forerunner of the Public Choice Society) included William Riker, William Mitchell, and Vincent Ostrom (political science), Kenneth Arrow, James Buchanan, and John Harsanyi, (economics), Herbert Simon (then public administration), Gordon Tullock (law), John Rawls (philosophy), and James Coleman (sociology).

1.2 Schools of RC

There are at least four recognizable schools of RC, although as the approach has spread into the mainstream of political science the distinctions have eroded. The most influential school within political science is the Rochester school, founded by Riker. A product of one of the country's most traditional political science programs (Harvard); Riker had both dissatisfaction with the state of political science in the middle of the twentieth century and a clear vision of the direction in which it should go. His program at the University of Rochester began producing graduates in the late 1960s and within a decade they had established thriving groups at Carnegie-Mellon, Washington University, and Caltech. Within another decade Rochester graduates and their students had accepted positions at the country's top political science departments.

In contrast, the Virginia School is based on economics. Led by Buchanan and Tullock this school is most closely identified with the Public Choice Society in particular, and the term 'public choice' in general. From the beginning the Virginians emphasized empirical work and especially policy relevant work. Virginians tend to be less abstract in their methods and more focused on outcomes than their Rochester counterparts. And more than Rochester the Virginia School is associated with distinct normative point-of-view, advocating markets as a better means of allocating resources than governments.

The Chicago School is associated most closely with George Stigler and Gary Becker. Reflecting their roots in economics, Chicagoans generally assume that actors' goals are egoistic and material. This view of humans as selfish wealth maximizers strikes many political scientists as too narrow and has colored nonrational choice political scientists' views of the larger RC enterprise. In addition, the Chicago School discounts the importance of institutions, believing that rational actors easily find ways to get around institutional constraints. Virginians and Rochesterians, in contrast, take institutions seriously. Virginians focus heavily on constitutional choice and Rochesterians are at the forefront of the 'new institutionalizm.' Thus Americanists and comparativists working in rich institutional contexts have found Chicago's emphasis on freewheeling interest groups less useful than international relations scholars who study an anarchic arena populated by weak institutions.

Finally, the Indiana School is smaller and more specialized. Founded by Vincent and Elinor Ostrom, the Indiana School takes as its special focus collective action problems in their various manifestations. Like Rochesterians, Indianans are more of a homegrown school of thought that attaches considerable importance to institutions. They are also more comparative in their orientation, tracking the operation of general principles across space and time.

2. Evolution of the RC Approach in Political Science

The RC approach first became widespread in the subfield of American politics. Probably this reflects the interests of several of its founders in democratic politics, as well as the fact that the American subfield was the most methodologically advanced and thus the most receptive to the sophisticated modeling tools that RC practitioners brought to the table. The main source of criticism within the subfield was theoretically and conceptually based. In the 1960s the subfield was dominated by behaviorists, who had tried to construct a scientific political science by borrowing concepts and

theories from sociology and psychology. From their studies of presidential elections, for example, behaviorists had concluded that the most important determinant of vote choice was party identification, then viewed as largely apolitical—a product of childhood socialization. The RC emphasis on rational, goal-oriented behavior seemed highly inconsistent with much of the behavioral literature.

The disruptions of the late 1960s undercut that judgment. As demonstrations raged outside, as an insurgent third candidate—George Wallace—contended for the Presidency, and as long-time Democratic identifiers defected from their party, it became increasingly untenable to assert that people voted on the basis of immutable, apolitical party identifications. RC scholars joined others in refocusing attention on the importance of issues and government performance and re-examined party identification, finding a political component within it.

By the 1980s RC had moved into the subfield of international relations. This was a natural development for the portion of the subfield focused on international economic relations, but natural as well for the international security side of the subfield, for whom coalitions, strategy, rational deterrence, and similar concepts had long been part of their stock in trade. Indeed, much of the early work on game theory and other operations research methods was done at the Rand Corporation with funding from the Department of Defense.

Most recently, RC has been moving into the subfield of comparative politics where its reception has perhaps been more controversial than in the other subfields. Some possible reasons for this contentious debate are offered.

2.1 Criticisms of RC

A common criticism of RC is that it is a hegemonic research program, not content to coexist with other programs but intent on replacing them. There is some truth to this charge, but it does not apply uniquely to RC—traditionalists leveled the same charge at behaviorists. Emphasizing the volitional element of politics does not require that the non-rational and emotional be ignored. But applying an approach to every possible research question helps scholars to determine its limits.

For example, it seems clear that RC is most directly applicable to the behavior of the elite, whose individual decisions make a difference, than to the behavior of the masses, whose individual decisions are submerged in those of millions of others. Attempts to behave rationally are most likely where the stakes are relatively high and the number of players relatively low—making a careful decision is not worth the effort if the consequences are trivial or your decision makes no difference to the outcome, or both. Thus, RC fails to provide a compelling explanation of why millions of ordinary citizens vote in a presidential elections, but that does not imply that the two presidential candidates do not make every effort to behave rationally!

Others quarrel with the RC approach because it is assertively social scientific. Formalization and generalization are social science ideals, and RC is closely associated with both, although nothing requires that all RC research be expressed in terms of mathematical symbols. Understandably, those who prize detailed narratives of particular cases feel uncomfortable with RC, as do those who believe that a country's unique history or culture determines the operation of its institutions and political processes. In both cases the critics' quarrel is as much with social science as with RC.

Of course, even some card-carrying social scientists have qualms about the RC approach. Not everyone views human behavior as purposive. Moreover, RC is more self-consciously theoretical than other approaches. Practitioners engage in model-building, making simplifying assumptions and drawing logical consequences using the classic deductive method. Scholars whose research commitments are inductive and empirical sometimes feel that RC models omit much of the rich detail they have painstakingly compiled.

Finally, some part of the criticism of RC is undoubtedly ideological. Scholars have political commitments, and it is natural, although regrettable, that these commitments sometimes influence their work. As noted, rational choice is associated with a free-market viewpoint. There is nothing inherent in RC that logically requires such a point-of-view. Indeed, there are rational choice scholars who are Democrats, Liberal Democrats even! But the perception exists, and scholars who position themselves more on the social democratic side of the spectrum may criticize the approach because of its perceived ideological associations.

The acrimonious nature of the debates that have accompanied the movement of RC into comparative politics probably reflects the fact that many comparative politics scholars are on the same side of several, if not all, of the preceding divisions. Some comparativists believe in thick description, not parsimonious formal models, and many of these same scholars believe that unique cultural and historical conditions are the key elements of political systems, not timeless underlying principles common to all systems. Some comparativists emphasize the non-rational nature of cultural and historical practices, not rational calculations about means to ends. And finally, some comparativists, especially those who specialize in the developing areas, are personally farther to the left on the political spectrum than Americanists steeped in the liberal tradition and international relations scholars comfortable with discussions of realism and the national interest.

2.2 The Evolution of RC

In the years to come the controversies surrounding the RC enterprise are likely to diminish. The principal reason is the convergence generated by attempts to explain real world behavior. As noted, RC forced Americanists to reconsider the attempt to explain politics primarily in sociological or psychological terms. But just as clearly, the attempt to explain mass behavior such as political participation indicates that recourse to sociology and psychology will be necessary. Similarly, RC researchers recognize that there are phenomena underlying labels such as 'cultures' and 'norms,' and provocative current research seeks to explain such phenomena as 'equilibria' or 'focal points' of human interaction that evolve over time through processes of trial and error, learning, selection, or what-not, and persist if their consequences continue to benefit the community. Trends suggest movement toward a richer RC approach, but also toward political sociology and political psychology with a firmer grounding in first principles of human behavior.

See also: Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory: Cultural Concerns; Rational Choice Theory in Sociology

Bibliography

Amadae S M, Bueno de Mesquita B 1999 The Rochester school: The origins of positive political theory. *Annual Review of Political Science* **2**: 269–95

Fiorina M P 1995 Rational choice, empirical contributions, and the scientific enterprise. *Critical Review* **9**: 85–94

Hardin R 1987 Rational choice theories. In: Ball T (eds.) *Idioms of Inquiry*. State University of New York Press, Albany, NY, pp. 67–91.

Mitchell W C 1988 Virginia, Rochester, and Bloomington: Twenty-five years of public choice and political science. *Public Choice* **56**: 101–19

Mueller D C 1997 *Perspectives on Public Choice: A Handbook*. Cambridge University Press, Cambridge, UK

Stigler G (ed.) 1988 *Chicago Studies in Political Economy*. University of Chicago Press, Chicago

M. P. Fiorina

Rational Choice Theory: Cultural Concerns

The notions of rationality, rational choice, and rational beliefs have a central place in many of the social sciences. Although originally deployed mainly in economics, they are increasingly being incorporated into sociology and political science. (They have not gained the same importance in social anthropology.) Some scholars argue that by virtue of its simplicity, universality and explanatory power the theory of rational choice is, or should be, the unifying framework for all the social sciences. As an explanatory theory it harbors formal and empirical problems that suggest a need for alternative approaches. At the same time, whatever its explanatory flaws, the notion of rationality retains a normative privilege, because we want to be rational (Føllesdal 1982). Rational choice theory is primarily prescriptive: it tells agents what to do to achieve their ends as well as possible. Social scientists may then try to explain their behavior by assuming that they do what the theory tells them to do.

1. The Development of Rational Choice Theory

The modern notion of rationality did not emerge until this century. Earlier references to rationality were to *ratio*, or 'reason,' understood as the faculty of calm and deliberative choice, as opposed to emotion. In the words of David Hume (1960, p. 415), 'Reason is, and ought only to be, the slave of the passions.' Some writers opposed reason both to interest and emotion: 'Nothing is easier for passion than to overcome reason; its greatest triumph is to conquer interest' (De La Bruyère 1990, IV.77). In this tradition, reason implies not only a dispassionate but also a disinterested and impartial attitude.

1.1 The Economic Meaning of Rationality

The modern concept of rationally as, essentially, means–end efficiency owes much to the marginalist revolution in economics in the last third of the nineteenth century. By conceptualizing the consumer's choice and the producer's choice as involving trade-offs among different consumption goods or different factors of production, the marginalists made it possible to think more systematically about the optimal choice of means for a given end. A profit-maximizing employer, for instance, would react to a wage increase by using more nonlabor inputs and less labor.

1.2 Max Weber

In *Economy and Society*, Weber (1968, p. 24) offered an influential definition of instrumentally rational (*zweckrational*) action: it is 'determined by expectations as to the behavior of objects in the environment and of other human beings; these expectations are used as "conditions" or "means" for the attainment of the actor's own rationally pursued and calculated ends.' Although Weber's professed aim was to offer a purely subjective notion of rationality as coherence among action, ends, and beliefs, some of his analyses

point to a more objective definition in terms of successful adaptation to the environment (Elster 2000). The closing pages of the *Protestant Ethic* (Weber 1958) suggest that he was influenced by the fact that market competition generates objectively successful behavior by eliminating those who fail.

1.3 The Emergence of Game Theory

Weber (1968) writes that 'in attempting to explain the campaign of 1866, it is indispensable in the case of [the two main generals] to attempt to construct imaginatively how each, given fully adequate knowledge both of his own situation and of that of his opponent, would have acted.' Weber does not seem to be aware of the need to 'construct imaginatively' how each general would try to 'construct imaginatively' what the other would be doing, including their further constructive imaginings, and so on. The modern theory of interdependent decisions, or (in somewhat misleading language) 'game theory,' emerged only after Weber's death. Neumann and Morgenstern (1944) and Nash (1950) showed how one can avoid the infinite regress lurking in the phrase 'and so on' by the notion of 'equilibrium', defined as a set of individual choices each of which is optimal, given the others.

1.4 The Emergence of Decision Theory

In much of rational choice theory, the intuitive notions of desire and belief are replaced by those of preference (or utility) and subjective probability. In a pioneering work, Ramsey (1931) showed how one can elicit utilities and subjective probabilities by asking subjects to respond to various hypothetical bets. Later, Savage (1954) offered an alternative demonstration. The key conclusion of these and later works in the same tradition is that rational actors maximize expected utility. The agent is assumed to face a well-defined set of feasible courses of action, each of which can with definite probabilities result in various outcomes, each of which has a definite utility for the agent. A rational agent will then choose the action that maximizes the average (or expected) utility of the outcomes associated with it.

2. The Structure of Rational Choice

In defining rationality, one can emphasize rational choice or rational action. In the latter approach, one can define rational action as action that results from rational choice or more objectively as success or adaptation to the environment. Even if social agents do not make rational choices, they may be constrained by their environment to act rationally. As noted above, this outcome may be produced by competition. In a population of firms, we may observe profit-maximizing behavior even if none of the decision-makers have profit-maximizing intentions, simply because firms that do not maximize profit are eliminated (Alchian 1950). As noted by Nelson and Winter (1982), the argument assumes that the environment is sufficiently stable for the process of elimination to have time to do its work. With a rapidly changing environment, the selection process is tracking a moving target and may not bring about optimal behavior.

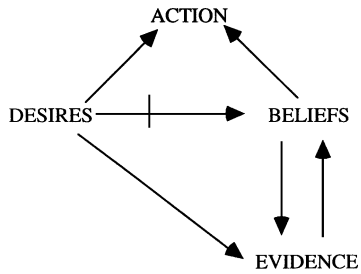

Figure 1
Rational action

2.1 A Basic Model

A widely used model of rational choice (and rational action) is captured in Fig. 1. In this diagram, the arrows represent three optimality relations. To be rational, an action has to be the best means to realize the agent's desire, given his or her beliefs. The beliefs must be well-grounded in the evidence available to the agent. The information must reflect an optimal investment in the acquisition of new information. (Because investing in information-gathering is an action, we may think of it as a shadow action that accompanies the primary action.) A rational choice is the choice of a rational action because it is the rational thing to do, excluding cases in which the agent makes the rational choice by fluke.

2.2 The Model is Subjective Through and Through

A choice and the subsequent action are rational only by virtue of the optimality relations among elements of the agent's mental set. A rational action may fail, if the agent has false (but well-grounded) beliefs. The relation between beliefs and evidence must not be confused with the relation between beliefs and the world. In inferring beliefs from the evidence a rational agent would use the procedures which in the long run yield the largest number of true beliefs, weighed by their subjective importance to the agent. Yet on any given occasion this procedure may lead to the formation of a false belief.

In the model, the desire is the 'unmoved mover.' The agent's desires allow us to assess the optimality of actions, beliefs, and information-gathering, but they are not themselves subject to rationality or optimality assessments. In some cases, this may seem counter-

intuitive. In the case of the agent's time preferences (a subset of the desires), we might want to say that an agent who discounts the future very heavily is irrational. The life of somebody who pays very little attention to the future consequences of present behavior is likely to be nasty, brutish, and short. Yet for the purposes of explaining the agent's behavior, this fact is irrelevant. Similarly, it may be tempting to say that a belief which (to us) is obviously false is irrational. Yet if it is part of the belief that the costs of testing it are very high, the agent may be caught in a belief trap (Mackie 1996) in which it is rational to hold on to the false belief. In applying rational choice theory, one has to adopt the point of view of the agent and recognize that the eye cannot see beyond its horizon.

2.3 Belief–Desire Interactions

As shown in Fig. 1 the model explicitly excludes a direct influence of desires on beliefs, as in wishful thinking. The converse possibility is that desires might be shaped by the beliefs, as when the belief that something is inaccessible stifles the desire for it ('sour grapes') or enhances it ('the grass is greener'). The general principle that desires cannot be assessed as more or less rational precludes us from qualifying desires shaped in these ways as irrational, however undesirable they might be on other grounds.

Rationality is, however, compatible with an indirect influence of desires on beliefs. As indicated in Fig. 1, the optimal amount of investment in evidence-gathering is shaped not only by beliefs (about the expected costs and benefits of gathering information), but also by desires. As beliefs are rationally shaped by evidence, it follows that beliefs may be rationally sensitive to desires. Consider two doctors who have to diagnose two identical patients with severe injuries. One is a GP who wants to save the patient's life and who knows that spending too much time on the correct diagnosis might undermine that goal. The other is a Nazi doctor in a concentration camp, who is unconcerned about the patient's survival and thus of the opportunity costs of information-gathering. Because of the different desires, the latter is more likely to form a correct belief.

2.4 Some Common Misunderstandings

Rational choice theory is often criticized, sometimes with good arguments, and sometimes with bad. Although some of the bad arguments may apply to bad versions of the theory, critics ought to address the best versions. The most common misunderstanding is that the theory assumes agents to have selfish motivations. Rationality is consistent with selfishness, but also with altruism, as when someone is trying to select the charity where a donation can do the most good.

Rational choice theory is also sometimes confused with the principle of methodological individualism. True, the theory presupposes that principle. Talk about beliefs and desires of supra-individual entities, such as a class or a nation, is in general meaningless. The converse does not hold, however. Some of the alternatives to rational choice theory also presuppose methodological individualism.

Furthermore, it is sometimes asserted that the theory is atomistic and that it ignores social interactions. Almost the exact opposite is true. Game theory is superbly equipped to handle three important interdependencies: (a) the welfare of each depends on the decisions of all; (b) the welfare of each depends on the welfare of all; and (c) the decision of each depends on the decisions of all. It is also sometimes asserted that the theory assumes human beings to be like powerful computers, which can instantly work out the most complex ramifications of all conceivable options. In principle, rational choice theory can incorporate cognitive constraints on a par with physical or financial constraints. In practice, the critics often have a valid point.

Finally, it is sometimes asserted that the theory is culturally biased, reflecting (and perhaps describing) modern, Western societies or their subcultures. However, the minimal model set out above is transhistorically and transculturally valid. This statement does not imply that people always and everywhere act rationally, nor that they have the same desires and beliefs. It means that the normative ideal of rationality embodied in the model is one that is explicitly or implicitly shared by all human beings. People are instrumentally rational because they adopt the principle of least effort. 'Do not cross the river to fetch water,' says a Norwegian proverb. Also, as people know that acting on false beliefs undermines the pursuit of their ends, they want to use cognitive procedures that reduce the risk of getting it wrong when getting it right matters. That they often fail to adopt the right procedures does not undermine the normative ideal.

3. Some Problems

As an explanatory approach, rational choice theory faces two objections. The first is formal: under certain circumstances the notions of rationality or optimality are not uniquely defined or not defined at all. This is the problem of indeterminacy. The other is empirical: sometimes the social agents do not behave as the theory predicts they will. This is the problem of irrationality.

3.1 Indeterminacy

The problems of non-unicity or non-existence of an optimal solution arises at each of the three levels in Fig. 1

At the level of action, indeterminacy can arise if the agent's preference ordering is incomplete, i.e., for a

pair of options **a** and **b**, the agent neither prefers **a** to **b**, nor **b** to **a**, nor is indifferent between them. This case can arise if the options are very different from each other and from what the agent has experienced in the past, as in choosing a career or a spouse.

At the level of beliefs, indeterminacy—meaning that the agent has no reliable grounds for assigning probabilities to the possible outcomes—can take the form either of raw uncertainty or of strategic uncertainty. The former is illustrated by natural disasters or by the process of technical innovation. The latter is illustrated by games in which there are several equilibria none of which can be unambiguously singled out as 'superior' or 'natural' (in the sense of Schelling 1960). An example involving both forms of uncertainty involves investment by firms in research and development (Dasgupta and Stiglitz 1980).

At the level of information-gathering, indeterminacy arises through an infinite regress. To decide how much information to gather, one must first form beliefs about the expected costs and benefits of information. The formation of these beliefs also requires the acquisition of information, and so on (Winter 1964). Although the regress can be avoided in repeated and stereotyped situations such as many medical diagnoses, it is unavoidable when, as in warfare, there is both an urgent need to act and no precedent to rely on.

3.2 Irrationality

In this case, too, irrationality arises at all three levels. The problems are of two kinds. 'Cold' problems are caused by cognitive malfunctioning without any motivational bias. 'Hot' problems arise through such bias.

At the level of action, cold problems arise in many ways. Choices are sensitive to how questions are framed—whether the glass is described as half empty or as half full (Tversky and Kahneman 1981). Many people are prone to the 'sunk cost fallacy,' i.e., to pursue a project simply because they have already invested much in it (Arkes and Blumer 1985). Hot problems can arise when people act under the influence of emotion, e.g., by wanting to take action immediately instead of taking time to maximize the chance of success. Under the influence of addictive cravings one may fail to carry out a decision to quit. (See Elster 1999 for both issues.)

At the level of belief formation, cold problems can arise because people fail to understand the nature of statistical inference. During World War II Londoners were persuaded that the Germans systematically concentrated their bombing in certain parts of their city, because the bombs fell in clusters. This invalid inference reflected a lack of understanding of the statistical principle that random processes usually generate clustering (Feller 1968, p. 161). Flight instructors who noticed that praise for an unusually smooth landing is typically followed by a poorer landing on the next try while criticism after a rough landing is followed by an improved performance, concluded that rewards are less effective than punishments. This invalid inference reflected a lack of understanding of the statistical principle of regression to the mean (Tversky and Kahneman 1974). Hot problems arise when 'the wish is father to the thought,' as in self-deception and wishful thinking. Coffee drinkers are less inclined to believe evidence of negative effects of caffeine (Kunda 1987). Those who work in risky occupations tend to believe the risks are less than they actually are (Akerlof and Dickens 1982).

Less is known about cold errors in information gathering. One might conjecture that people may invest too much in information because they underestimate opportunity costs compared to out-of-pocket expenses (Thaler 1980). A common hot mistake also involves overinvestment, when people have a strong desire to have sufficient reasons for their decisions even when cost–benefit analysis suggests that flipping a coin might be superior (Elster 1989a). Motivational biases can also, however, lead to underinvestment: 'Marry in haste, repent at leisure.'

4. Alternatives

Problems such as those discussed above have stimulated the search for alternative theories. Some theories address the issue of indeterminacy as well as 'cold' deviations from rationality, others the 'hot' deviations.

4.1 Bounded Rationality and Behavioral Economics

Some alternatives try to identify the cognitive mechanisms that produce deviations from what rational choice theory predicts.

The research program on 'bounded rationality' stems from Herbert Simon (1954). Nelson and Winter (1982) is an outstanding application. The core assumption is that agents 'satisfice,' by choosing an option that is 'good enough,' rather than maximize. Agents do not look for better alternatives if the status quo is satisfactory. If it falls below the agent's aspiration level, a search for new options is triggered until a new, satisfactory outcome is found. The theory has a good match to common sense ('Never change a winning team') and captures the heavy dependence of many choices on search procedures. By eschewing the maximizing assumption, it avoids problems of indeterminacy.

The research program on 'behavioral economics' or 'quasi-rational economics' (Thaler 1991) stems from the work of Kahneman and Tversky (Kahneman et al. 1982, Kahneman and Tversky 2000). In addition to identifying many of the 'cold' anomalies discussed above, these authors tried to account for them by

means of 'prospect theory.' It deviates from rational choice theory by assuming that agents evaluate gains and losses by how much they deviate from a reference point rather than in absolute terms, and that they are more risk-averse for losses than for gains. These two assumptions explain, for instance, that people who refuse to use credit cards when they are said to carry a surcharge accept them when the charge is redescribed as a cash discount.

4.2 Emotions and Social Norms

Other theories try to identify the motivational mechanisms that generate anomalies.

Frijda (1986, pp. 118–21) argues that emotions cause deviations from rationality in many ways: (a) they affect 'probability and credibility estimates' concerning events outside one's control; (b) they 'cause some measure of belief in the efficacy of actions one would not believe in under other conditions'; (c) they induce fantasy behavior, as when a widow for many years after the death of her husband sets the table for two each day; (d) they induce various forms of 'painful fantasies that, however painful, are yet sought by the subject himself.'—Othello's jealousy is cited as an example; and (e) they cause irrational behavior, such as 'anger at some deed that cannot be undone by the angry aggression, nor its recurrence prevented.' In addition, as mentioned earlier, they induce suboptimal investment in information. Loewenstein (1996) generalizes these ideas, by appealing to a broad range of 'visceral factors' that interfere with rational choice and rational cognition. In addition to emotion, these include hunger, thirst, pain, and the euphoria or dysphoria created by addictive drugs or abstinence from them.

Rational choice theory is consequentialist, in that it enjoins people to act in certain ways because of the expected outcomes. Social norms typically prescribe acts for their own sake, not for their consequences. In their simplest form, they take the form of unconditional imperatives: 'always wear black at funerals.' They may be contrasted with conditional, outcome-oriented imperatives: 'always wear black in strong sunshine' (as do people in Mediterranean countries to maintain circulation of air between the clothes and the body). Social norms regulate the use of money, how hard people work at their jobs, the occasions for and modalities of revenge, sexual behavior, and many other matters (Elster 1989b). They are sustained largely by the emotions of contempt and shame.

5. Conclusion

Many rational choice theorists will respond either that some of the alleged problems are empirically insignificant or that the alternative explanations can be restated within the rational choice framework. These are ongoing debates with many unresolved questions. Overall, however, it is increasingly hard to deny that rational choice theory has serious limitations. Empirical research increasingly shows that human beings are severely limited in their cognitive functioning, and, when under the influence of emotion and other strong feelings, liable to act in ways that undermine their ends. Paraphrasing Veyne (1976, p. 676), beliefs born of passion serve passion badly. At the same time, there is no single rival or contender that could replace rational choice theory. For one thing, most people *are* rational much of the time. For another, the various alternatives to the theory have not been integrated with one another, and are unlikely to be so. There is no other theory that explains as much of human behavior, although the jury is still out on how much it explains.

See also: Bounded and Costly Rationality; Bounded Rationality; Intentionality and Rationality: A Continental-European Perspective; Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory in Sociology; Rational Theory of Cognition in Psychology; Rationality and Feminist Thought; Rationality in Society

Bibliography

Akerlof G A, Dickens W T 1982 The economic consequences of cognitive dissonance. *American Economic Review* **72**: 307–19

Alchian H 1950 Uncertainty, evolution, and economic theory. *Journal of Political Economy* **58**: 211–21

Arkes H R, Blumer C 1985 The psychology of sunk costs. *Organization Behavior and Human Decision Processes* **35**: 124–40

Dasgupta P, Stiglitz J 1980 Uncertainty, industrial structure, and the speed of R&D. *Bell Journal of Economics* **11**: 1–28

De la Bruyère J 1990 *Les Caractères*. Garnier-Flammarion, Paris

Elster J 1989a *Solomonic Judgements*. Cambridge University Press, Cambridge, UK

Elster J 1989b *The Cement of Society*. Cambridge University Press, Cambridge, UK

Elster J 1999 *Strong Feelings: Emotion, Addiction and Human Behavior*. MIT Press, Cambridge, MA

Elster J 2000 Rationality, economy and society. In: Turner S (ed.) *The Cambridge Companion to Weber*. Cambridge University Press, New York

Feller W 1968 *An Introduction to Probability Theory and its Applications*, 3rd. edn. Wiley, New York, Vol. 1

Frijda N H 1986 *The Emotions*. Cambridge University Press, Cambridge, UK

Føllesdal D 1982 The status of rationality assumptions in interpretation and in the explanation of action. *Dialectica* **36**: 301–16

Hume D 1960 *A Treatise of Human Nature*. Oxford University Press, Oxford, UK

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment Under Uncertainty*. Cambridge University Press, Cambridge, UK
Kahneman D, Tversky A (eds.) 2000 *Choices, Values, and Frames*. Cambridge University Press, Cambridge, UK
Kunda Z 1987 Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology* **53**: 636–47
Loewenstein G 1996 Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes* **65**: 272–92
Mackie G 1996 Ending footbinding and infibulation: A convention account. *American Sociological Review* **61**: 999–1017
Nash J 1950 Equilibrium points in *N*-person games. *Proceedings of the National Academy of Sciences* **36**: 48–9
Nelson R R, Winter G 1982 *An Evolutionary Theory of Economic Change*. Belknap Press of Harvard University Press, Cambridge, MA
Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ
Ramsey F P 1931 Truth and probability. In: Ramsey F P (ed.) *The Foundations of Mathematics*. Routledge and Kegan Paul, London
Savage L J 1954 *The Foundations of Statistics*. Wiley, New York
Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA
Simon H 1954 A behavioral theory of rational choice. *Quarterly Journal of Economics* **69**: 99–118
Thaler R H 1980 Towards a positive theory of consumer choice. *Journal of Economic Behavior and Organization* **1**: 39–60
Thaler R 1991 *Quasi-rational Economics*. Russell Sage, New York
Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* **185**: 1124–31
Tversky A, Kahneman D 1981 The framing of decisions and the psychology of choice. *Science* **211**: 453–8
Veyne P 1976 *Le pain et le cirque*. Seuil, Paris
Weber M 1958 *The Protestant Ethic and the Spirit of Capitalism*. Scribner, New York
Weber M 1968 *Economy and Society*. Bedminster, New York
Winter S 1964 Economic 'natural selection' and the theory of the firm. *Yale Economic Essays* **4**: 225–72

J. Elster

Rational Choice Theory in Sociology

Outside the confines of economics and closely related disciplines such as the decision sciences, where it is adhered to almost universally, rational choice theory (RCT) has not gained wide acceptance. Due to its relative success in economics and the limited achievements of other theoretical approaches adopted elsewhere, particularly in sociology and political science, rational choice theory gained significant support in the late twentieth century.

Use of RCT in sociology ranges from formulating mathematically deductive models assuming fully informed and flawlessly calculating actors, to a much broader approach, whence actors are seen as 'doing their best' (i.e., subjectively optimizing), given certain preferences/desires/affects/utilities, with only a limited understanding of their environment and partial reasoning or calculative ability. It is this flexibility which recommends RCT to many of its proponents, ranging from those for whom it comprises the hallmark of any acceptable social science to those for whom it is pragmatically the best framework currently available.

RCT may be used in two complementary ways. First, in a positive manner by proposing explanations of individual and collective actions (decisions) and consequently responding to the question—how (in context) *do* actors behave? Second, in a normative manner, responding to the question of how they (again in context) ought, given their preferences, to behave. While strong in economics, the normative approach is not widely embraced in sociology.

Rigorous applications of RCT are almost invariably based upon an assumption of the maximization of (often subjective (expected) utility). Actors are thus conceived as (a) capable, at any instant, of taking one of a series of feasible actions, (b) each of which leads to an estimable set of outcomes, with (c) a known (subjective/objective) probability. Further, each outcome carries affect for the actor. The actor can then form a preference ordering over the actions in terms of affects/utilities. Actions (decisions) under certainty, risk and uncertainty are thus special cases of the general model. Various axiomatic systems are available, though the one attributed to Savage is closest to sociological concerns. It is now recognized that actors do not always abide by the precepts of expected utility theory. Prospect theory and developments thereof, which introduce a wide range of 'biases,' are widely accepted as falling under the broad ambit of RCT, as are models based upon limited rationality (like satisficing).

Three related types of RCT modeling frameworks can be distinguished. First, independent decision-making (actions), where the actor's environment may be treated as fixed (i.e., parametric); second, where the environment involves other actors and where consequently decisions or actions are interactive (i.e., strategic), where one shot and (finitely and infinitely) repeated game theory and their associated Nash equilibria are appropriate. Third, evolutionary game theory, where strategies are selected by imitation of either the relatively frequent or relatively successful strategies of others. While the first two frameworks, analytically speaking, usually commence with the full assumption of well informed and rationally calculating individuals, the latter requires no such initial assumptions. It does, though, usually assume that evolutionary process is driven towards an evolutionary stable equilibrium (or some refinement thereof).

The assumptions of expected utility maximization (more generally RCT) in complex environments (parametric or strategic) prove demanding on both the

reasoning power and knowledge (information) of the actors concerned, and are sometimes deemed by critics to be entirely unrealistic in a *positive* context. RCT is, however, strongly wedded to the idea of equilibrium, and in off-equilibrium adjustments all manner of trial and error may be allowed for. Equilibrium, in this view, is not necessarily arrived at (learned) because actors are rational, but rather because actors appear rational (i.e., discover rationality) once an equilibrium has been achieved. Evolutionary game-theoretic models derive the same conclusion by myopic processes of selection (emulation) eventuating in a evolutionary stable equilibrium.

All manner of RCT models may be said to rely initially upon three assumptions, each of which may be regarded as controversial in some sociological circles. They are:

Individualism. Only individual persons ultimately take actions.

Optimality. Individual actions are chosen optimally (at least at equilibrium).

Self-regard (or egoism). Individuals' actions are ultimately concerned with their own welfare (and nobody else's).

Though any one or more of these assumptions may be adjusted in a perspicacious manner (see below), rational choice theory may nevertheless be said to retain its paradigmatic privilege in the sense that it (with its standard assumptions) provides an analytical point of departure.

Despite its explicit individualism, system level or collective/macro relationships between variables are a perfectly acceptable currency for RCT as long as they are construed ultimately as 'shorthand' for the conjunction of relationships that ultimately run through individuals. Indeed, a theory about the connection between two system-level variables must ultimately imply a mechanism involving individual actions. The theoretical question is precisely this: How has the system level relationship been generated by the actions of individuals? The assumption of individualism in RCT implies no more than this about the explanation of system or collective level empirical generalizations.

No contradiction exists between RCT and a view which pictures individuals' actions as 'shaped' by their membership in a group (more generally a collectivity). If, as some 'structuralists' opine, individuals are often so circumscribed by internalized standards and external constraints upon their actions that they barely have any choice at all, then this may call into question optimality and self-regard, but it does not undermine the assumption of individualism in the sense that it is still the motor power behind individual actions (albeit 'programmed ones') that drive the social world.

Things happen in the social world because individuals do and do not (i.e., forbear) to do things, and they are the only things that do and do not do things. All statements that attribute 'doing' to other entities can, in principle if not in practice, be translated, without loss, into statements about individuals doing things.

Individuals choose optimally if no other action exists whose consequences are preferred to the chosen action. An actor's preference-driven action will in turn depend upon that person's (a) beliefs or reasoning about the consequences of the available actions, and (b) affect (desire for/against) the consequences (and perhaps for the action itself). It is thus possible by derivation to speak of optimal beliefs and affects. It is this broad assumption of optimality that gives RCT its paradigmatic privilege.

The major alternative assumptions to this viewpoint are:

(a) That individuals frequently do not act optimally in terms of their preferences.

(b) That although individuals may act optimally in terms of their preferences, senses exist in which their preferences are themselves not formed optimally (nonoptimal beliefs and affects).

(c) That individuals do not act upon choices but are driven by forces entirely beyond their control (structuralism).

Cases are well documented in which 'weakness of will' may be said to prevent people from acting in the way they would 'really prefer.' Preferences may also be imprecise or unclear, actions entirely experimental or based upon wishful thinking, and more than one optimal choice may be available. All of this is acknowledged by RCT, although in practice most of the problems reduce to examples of nonoptimal preference formation.

Some take the view (particularly many economists) that preferences are not only stable (compared to constraints) but also are formed in ways that entirely resist the attentions of RCT. Their genesis is either uninteresting (changed constraints accounting for changed action) or beyond rationality. Sociologists reject the first view almost unanimously, but find some sympathy for the second. Not infrequently it is also claimed that the idea of optimality is so weak, when taken by itself, as to constitute an affront to serious understanding. It is observed that, if we are to be permitted any sort of utility arguments, then actions may (certainly *ex-post facto*) always be made to look as if they are optimal.

The accusation is that, under such circumstances, the assumption of optimality can be protected entirely against potential refutation. In this context, sociologist think in terms of a series of derivative optimality questions:

(a) Are preferences formed optimally with respect to beliefs and affects?

(b) Are beliefs formed optimally with respect to the information available?

(c) Is available information (collected/processed) optimal with respect to affects?

(d) Are affects optimal with respect to individual autonomy?

An action is objectively optimal when conditions (a) to (d) hold true. Such actions may take place under certainty, risk, or uncertainty, according to standard assumptions about the nature of beliefs. These assumptions are collectively very strong, and many social theorists maintain that it is only rarely, if ever, that actions will conform to the ideal pattern.

The issue at stake is whether we may assume that affective dispositions in some sense flow from a balanced or autonomous personality. No doubt forms of behaviour/actions exist that flow from affective dispositions that are 'irrational' or pathological. When this occurs, it usually implies that individuals resist all occasions to consider the opportunity costs of maintaining the disposition (e.g., constantly striving for what is beyond their grasp or manifestly damages them).

It is points (a), (b) and (c) that have attracted the most attention. It is often proposed that preferences cannot be formed optimally for a variety of reasons, prominent among which are the following:

Satisficing. Beliefs are not performed optimally with respect to information available.

Search. No adequate theory of optimal search exists; therefore, once again, beliefs are not optimal with respect to information, nor information with respect to affects.

Decision biases. Preferences are not optimal with respect to beliefs and affects.

Exclusions. Not all possible actions are considered.

Limited deliberation. Not all consequences of all actions are thought through.

Received styles of reasoning. These prevent optimal belief formation.

Ritual/routine/habit/copying. These systematically lead to exclusions and do not evolve optimally in an evolutionary sense.

In each case a conceptual understanding of the apparent limitations of RCT is established in terms of the departure from full optimality. This is one significant sense in which RCT can make a claim for paradigmatic privilege: It still serves, even in default, as a benchmark. Furthermore, many of the above factors are modelled in terms of limited information environments.

The standard version of RCT usually starts with the assumption that individuals act (optimally) in order ultimately to satisfy their self-regarding preferences (utilities). That is, optimal actions are taken in a way that is indifferent to the welfare/utility/consumption of others, considered either individually or collectively. Although nothing intrinsic to the theory mandates this assumption, it is frequently regarded as a natural analytical starting point, to be discarded only cautiously if and when the facts speak otherwise.

Furthermore, if other-regarding preferences are invoked eventually, not infrequently it is maintained that they must in some way be explained or accounted for, whereas self-regarding sentiments do not invite explanation, being in some way natural or self-explanatory. Thus many theorists adopt the practice of assuming that, if no evidence exists to the contrary, then postulate self-regarding utilities. If, however, independent evidence exists for other-regarding sentiments, it would be a mistake to ignore it and go on to ask for an explanation of the source. However, many RCT models interpret other regarding sentiments as instrumental to self-regard (enlightened self-interest). Such models are usually excluded before other regard is introduced into the utility function (affective orientation).

Evolutionary biologists (and evolutionary game theorists) continue to debate the issue as to whether altruistic motives can evolve at an evolutioning stable equilibrium. Sociologists are required increasingly to relate their own theories to those of colleagues working in these adjacent disciplines. This is particularly so when they adopt the view whereby it is the *strategic interaction* between two or more actors which constitutes much social interaction. These developments are bringing sociological theory increasingly into contact with game theory and evolutionary game theory. Although games of incomplete information currently play a limited role in sociology, we can expect this to change in the near future. Perfect Bayesian equilibra will thus come centre stage.

The paradigm of sociological investigation which is emerging from an RCT perspective takes the following form:

(a) Initially, assume the actors who are responsible for generating a collective outcome (which may not be intended but only a consequence of what is intentionally brought about) to be rationally self-interested, fully informed and endowed with calculative ability given their (exogenous) objective resources, and (exogenous) preferences/utilities.

(b) Model the structure of the actors interdependencies (social interaction) in the sense that the outcome depends jointly upon what each does. Thus each actor has to take account of what others do (will do, are likely to do, and, if appropriate, have done in the past). Such modeling will usually define a feasible set of actions strategies for each actor.

(c) In the light of (a) and (b), determine the course of action each actor will (rationally) pursue. If these are not unique, then specify their ranges of actions (i.e., multiple equilibria). If the interactions are repeated, adopt the appropriate repeated game theoretical framework.

(d) Thus predict the outcome(s) and test these against descriptions of the outcomes one wishes to explain. Of course, if the model possesses no unique outcome, then all one can require is that the event in question is among the set of possible outcomes predicted by the model (an envelope rather than point prediction). Specific cultural/historical occurrences may select one equilibrium (i.e., focal points) and suggests path-dependence on a particular outcome.

(e) If this simple model fails, construct further models which might involve one or more of the following possibilities:

(i) Relax assumptions of full information, probably following recognized procedure in game theory.

(ii) Relax the assumption of self-interest, introducing richer utility/preference functions (e.g., altruism, malice, indifference, envy, etc.).

(iii) Relax the assumption of objectively calculated resources introducing subjective beliefs about resources.

(iv) Relax the assumption of a set of feasible actions, calculated objectively, introducing subjective beliefs about what is feasible.

(v) Re-compute the structure of interdependencies.

When either (i), (ii), (iii), or (iv) are seen to hold, this will inevitably prompt a further question as to why it should be so. Answers to these questions might also be couched in terms of some of the RCT precepts outlined above (e.g., strategic limitations of information). It is only if this very general framework fails to provide the answers we seek that we should then reach for an alternative theoretical framework.

See also: Action, Theories of Social; Altruism and Self-interest; Bounded Rationality; Coleman, James Samuel (1926–95); Game Theory; Interests, Sociological Analysis of; Macrosociology–Microsociology; Methodological Individualism in Sociology; Rational Choice Explanation: Philosophical Aspects; Rational Choice in Politics; Rational Choice Theory: Cultural Concerns; Sociology: Overview; Theory: Sociological; Traditions in Sociology

Bibliography

Abell P (ed.) 1991 *Rational Choice Theory*. Elgar, Aldershot, UK

Abell P 1996 Sociological theory and rational choice theory. In: Turner B S (ed.) *Social Theory*. Blackwell, Oxford, UK

Boudon R 1989 Subjective rationality and the exploration of social behaviour. *Rationality and Society* **12**: 173–96

Coleman J S, Fararo T J 1992 *Rational Choice Theory: Advocacy and Critique*. Sage, Newbury Park, CA

Moser P K (ed.) 1990 *Rationality in Action*: *Contemporary Approaches*. Cambridge University Press, Cambridge, UK

P. Abell

Rational Theory of Cognition in Psychology

1. People Appear Irrational in Cognitive Experiments

Michael Watkins has said that a cognitive theory 'is a bit like someone else's toothbrush—it is fine for that individual's use, but for the rest of us … well, we would just rather not, thank you' (Watkins 1984, p. 86). To say whether people are behaving rationally requires a definition of what it means to behave rationally, and like a toothbrush everyone has their own. For the purposes of this article, rational behavior will be defined as follows: to behave rationally in some context is to display behavior that corresponds to a normative standard of behavior for that context. The choice of the normative standard determines what constitutes rational behavior. For the economist, behaving rationally involves maximizing utility; for the logician it is following the deductive rules of logic; and for the (Bayesian) statistician it is acting according to Bayes's rule. One need not look far to find that people do not reason rationally with respect to a variety of normative standards. There are entire literatures on how and why people violate the rules of deductive logic and statistical inference. In the Wason (1968) card task, central to one such literature, people are given a rule of the form 'if p, then q,' such as 'if there is a vowel printed on one side of a card, then an even number is printed on the other.' The participants are next presented with a set of cards, such as A, K, 2, 7. Their task is to choose only those cards that need to be flipped to check whether the rule holds. In this example, only the A-card (p) and the 7-card ($\sim$ q) need to be checked. An odd number on the opposite side of the A-card would clearly violate the rule as would a vowel on the other side of the 7-card ($\sim$ q). The rule says nothing about what is on the opposite side of a consonant, so flipping the K-card ($\sim$ p) does not test the rule. Also, flipping the 2-card (q) cannot disconfirm the rule, because the rule does not restrict what is on the opposite side of an even-numbered card. In general people perform terribly on this task, or at least their performance does not appear to be rational with respect to the rules of deduction. Oaksford and Chater (1998) surveyed 13 studies, covering a variety of Wason tasks. The proportions of people who flip the p(A), $\sim$ p(K), q(2), $\sim$ q(7) cards were 89, 16, 62, and 25 percent respectively. That is, people rarely flip the $\sim$ q card, which they should flip, and frequently flip the q card, which they need not do.

People are similarly deficient in drawing inferences from statistical information. Consider the following example that requires people to interpret the results of a positive colon cancer test (Gigerenzer 1998). Suppose that 0.3 percent of people have colon cancer, there is a 50 percent chance that a colon cancer test will detect a cancer (a hit), and a 3 percent chance that it will indicate that there is cancer when there is none (a false positive). What is the probability that a person with a positive test has cancer? When Gigerenzer posed this problem to physicians their median estimate was 47 percent. The normative answer is more like 4.8 percent (see Table 1). This answer is obtained by applying Bayes's rule to the statistics given. Bayes's rule shows the correct way to make statistical inferences. Bayes's

Table 1
Bayes's rule applied to Gigerenzer's (1998) cancer example

This shows the probability of some hypothesis, H_j being true given some data, D. Bayes's rule takes the following form:

$$p(H_j|D) = \frac{p(H_j)\,p(D|H_j)}{\sum_i p(H_i)\,p(D|H_i)}$$

where $p(H_j)$ is the belief that the hypothesis H_j is true prior to observing the data, D, and $p(D|H_j)$ is the probability of observing the data, given the hypothesis is true. The denominator is the probability of observing the data. It assumes that there are i different hypotheses, one of which is true. We can apply Bayes's rule to Gigerenzer's (1998) cancer example. The two hypotheses are that the patient has cancer, or that he or she does not ($\sim$cancer). The base rate, or prior, probability of having colon cancer is 0.3 percent. The hit rate of the test is 50 percent, and the false alarm rate is 3 percent. Applying Bayes's rule to this example yields an estimate of having cancer to be 4.8 percent, a much more reassuring estimate than given by the doctors.

$$p(\text{cancer}|+\text{test}) = \frac{p(\text{cancer})*p(+\text{test}|\text{cancer})}{p(\text{cancer})*p(+\text{test}|\text{cancer})+p(\sim\text{cancer})*p(+\text{test}|\sim\text{cancer})} = \frac{.003*.5}{.003*.5+.997*.03} = .0477$$

rule implies that given the rarity of cancer, a positive test is more likely to follow a false positive than the actual detection of cancer. As with the Wason (1968) task, subjects' performance on statistical tasks like this clearly shows that they are behaving irrationally. The standard explanation for this is that people tend to ignore base rates; in this example they fail to take into account that cancer is uncommon. That is, only 0.3 percent of people have cancer.

2. *People Employ Ecologically Valid Strategies in the Laboratory*

Performance on Wason (1968) card tasks, Gigerenzer's (1998) statistical tasks, and other experiments demonstrate that people are, in fact, irrational, when measured against accepted standard inferential rules. There has been a number of proposals to describe what people are doing (e.g., Tversky and Kahneman 1974). The idea behind these approaches is that people are applying heuristics, rules of thumb that tend to be successful in their daily lives, to these laboratory situations where they do not apply. Anderson (1990) has taken this idea one step further, arguing that people are not just bringing real-world heuristics into the laboratory, but rather are employing optimal solutions to certain problems faced in their natural environments. However, behavior that is rational in natural environments may not necessarily be rational in the peculiar environments that experimental psychologists concoct. Some of the roots of this idea are discussed next.

3. *David Marr's Levels of Explanation*

Anderson's motivations closely follow the arguments that David Marr laid out in his influential book *Vision*. Marr (1982) argues that fully understanding an information processing system requires considering the system from multiple levels. Marr was interested in the visual system, which he argues evolved to 'tell(s) us about shape and space and spatial arrangement' (Marr 1982, p. 36). The distinction that Marr makes among the levels of explanation can be more readily understood by referring to a far simpler information processing system than vision—a clock. What does a clock do? It indexes time. How does it do it? It achieves this goal by measuring the passage of time by incrementing counters at fixed intervals. Marr says that descriptions like this are at the computational level.

The next level down, called the representation and algorithm level, describes how the goals should be achieved. In the case of a clock it would map the cycles of an oscillator into seconds, minutes, and hours. Descriptions at this level of analysis require a specification of the representations and the algorithms that operate on them to achieve the goals specified at the computational level. Many combinations of representation and algorithms can achieve these goals; a 12 or 24 hour clock can index time. The choice of representation does constrain the choice of algorithm, such as how seconds are rolled over into minutes, and minutes into hours (e.g., what happens at 12:59). Further, not all representation and algorithms are equivalent; some computations may be simpler with one representation than another. Calculating the duration of a trip is simpler when the train leaves at 10:00 and arrives at 14:00 than when it leaves at 10:00 a.m. and arrives at 2:00 p.m.

The lowest level of description in Marr's hierarchy, the hardware implementation level, is concerned with describing the physical entities that carry out the computations of the representation and algorithm level. Here, the current time could be represented by the configuration of hands on a clock face, the coordination of seconds and minutes handled by brass

wheels, the oscillator could be realized as a pendulum. Marr's point is that if you stumbled onto a morass of gears, you would be better able to make sense of these gears if you knew that they were part of a clock as opposed to a cash register, or a sewing machine. Similarly, one is going to be better able to understand people as information processing systems, if one understands what those systems are trying to accomplish. As Marr puts it, 'trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: it just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structures of feathers and the different shapes of birds' wings make sense' (Marr 1982, p. 27).

4. *John Anderson's Principle of Rationality*

Marr (1982) demonstrated the utility of approaching the problem of vision from multiple levels, and particularly from the computational level. Based on Marr's success with vision, Anderson (1990) hypothesized that the approach might work well for higher-level cognitive processes. Anderson argued that most theorizing in cognitive psychology concerns representations and the processes that act on them, corresponding to Marr's representation and algorithm level. Cognitive neuroscience and neuroscience focus on the hardware implementation level. Relatively little theorizing, he points out, occurs at Marr's computational level. Anderson (1990) explored whether it would be useful to think about higher level cognition, such as categorization (Anderson and Matessa 1992) and memory (Anderson and Milson 1989, Anderson and Schooler 2000) at the rational level, analogous to Marr's computational level. His principle of rationality is that 'the cognitive system operates at all times to optimize the adaptation of the behavior of the organism' (Anderson 1990, p. 28). Anderson renamed Marr's computational level the rational level, because the computational level sounds like it should be describing the algorithm and representation level. His choice of the term rational also serves as an allusion to economic theorizing that often takes place at Marr's computational level. Economists focus more on the decisions agents make, rather than on the processes involved in coming to those decisions. In the economic realm, it is easy to appreciate that it is rational for firms to maximize profits (or, as in the case of Ben and Jerry's, the famous American premium ice-cream maker, to maximize a combination of profits and social good). For the cognitive system, the currency is less clear. Thus, the critical step in what Anderson calls a 'rational analysis' is to figure out what quantity the cognitive system might be optimizing, and to make predictions based on this about how people will behave in particular experimental tasks.

5. *Oaksford and Chater's Rational Analysis of the Wason Card Task*

There would seem to be a weak case for the rationality of the cognitive system in light of people's irrational behavior on the Wason (1968) card task and Gigerenzer's (1998) cancer problem. However, Oaksford and Chater's (1996) rational analysis of the Wason card task shows that typical performance on the Wason, while violating traditional notions of rationality, is indeed quite rational when seen in a broader context of how people seek information in the world.

Consider the following situations that parents may face. In one situation, hearing an utterance from the baby he was caring for, a young father might ask 'was that just babbling?' or if the baby says 'baba,' then the baby wants a bottle. If the mother found her husband giving the baby a bottle, and if she wanted to know whether 'baba' was serving as the baby's signal it wanted a bottle, would it be rational for the mother to ask whether the baby had said 'baba' before the feeding? Now imagine 16 years have passed, and the mother is teaching her son to drive. She wonders whether he knows that when the oil light is on, he should stop the car. Would it be rational for the mother to ask whether the oil light is on? It would seem quite natural for the young mother to ask about what the baby had said, and for the middle-aged mother to remain silent. Asking in the first case would be like flipping the 2-card, irrational with respect to deductive inference. Asking cannot disconfirm the hypothesis that the baby knows how to use the word, since logically many circumstances (e.g., fussing) can suggest an infant is hungry. In contrast, asking about the oil light would be like flipping the 7-card, rational with respect to the rules of logical inference. If the mother found that the oil light was on, this would be a violation of the rule. Our intuitions are at odds with logic.

Perhaps our intuitions are correct that the young mother is rational to ask about what the baby said, and the older mother is rational in remaining silent. The essential difference between the two situations is the amount of information the answer to the question is likely to yield. Though asking whether the baby said 'baba' could not provide a definitive answer, knowing what the baby said tells something about whether the baby knows the word. In contrast, asking about the oil light, though potentially highly informative in the unlikely event that the oil light is on, will tell nothing in the more likely event that the oil light is off.

Oaksford and Chater's (1996) analysis of the Wason task formalizes these intuitions. They assume that people apply experimental, information-seeking strategies to deductive tasks like the Wason. Though the details of their mathematical analysis are beyond the scope of this article, the flavor of it can be given here. As is true for nearly all rational analyses, they assume

that people are behaving as if they are following Bayes's rule. They suggest that people are not treating the task as a test of logic, but rather are attempting to gauge the causal relation between two events. More specifically, they assume what people are really trying to do in deductive tasks is decide between two hypotheses: when p occurs (e.g., baby says 'baba'), q must follow (e.g., baby wants a bottle), or the alternative hypothesis that event p is independent of event q. Sometimes the baby says 'baba,' and sometimes the baby is hungry, and it is only by chance that the baby says 'baba' when it is hungry. In the case of the Wason task described earlier, the competing hypotheses are that an even number depends on a vowel or the alternative that evens and vowels are independent of each other. The question, then, is which card will provide the most evidence in terms of discriminating between these two rival hypotheses. Since people do not know in advance how an experiment (i.e., flipping a card) is going to turn out, they make their decisions based on how much information they expect to gain.

A critical difference between the parenting examples and the Wason (1968) task is that people have experience with how children learn words, but relatively little experience with numbered and lettered cards. Lacking any relevant experience about the cards, Oaksford and Chater (1996) assume that people treat them as if they are typical of causal relations they have seen in the past. When their model of information-seeking is combined with the assumption that causally related events are relatively rare, it predicts the observed subject preferences for flipping cards, namely p (e.g., A-card) is chosen more than q (e.g., 2-card), q more than $\sim$ p (e.g., K-card), and $\sim$ p more than $\sim$ q (e.g., 7-card). So in the Wason task we have a situation where the behavior is irrational with respect to the laws of deduction, but can be understood to be rational in the context of how people seek information.

6. *People are Rational When Ecologically Valid Strategies are Appropriate*

Gigerenzer's (1998) colon cancer experiment demonstrated that people are irrational with respect to proper Bayesian inference. This result seems to be at odds with Anderson's rational hypothesis, as Bayesian analyses underlie nearly all rational analyses. Since the problem was laid out perfectly for Bayesian inference, we would expect people to perform well. Gigerenzer argues that people perform poorly, because the problem format is all wrong. People did not evolve, he argues, to process probabilities, but rather to draw inferences based on their direct experience. His prediction is that people should do well on statistical tasks when the data are presented more naturally in terms of raw frequencies. Consider again the colon cancer problem, but this time in raw frequencies. Thirty out of every 10,000 people have colon cancer. Of these 30, 15 will test positive. Of the remaining 9,970 people without cancer, 300 will still test positive. What proportion of people who test positive will actually have cancer? Now it is clear that only 15 of the 315 (or 4.45 percent) of those who test positive will have cancer. When the problem was presented this way to another group of doctors, 67 percent answered correctly, as compared to 4 percent when the data were presented in terms of probabilities. When the problem presentation is consistent with how we experience events in the world, people's behavior corresponds to the prescriptions of a normative standard, a Bayesian inference. As with the Wason task, people appear to be behaving irrationally, when the experimental task and conditions conflict with the natural environment.

7. *In Practice, Rational Analyses are Bounded*

The discussions of the Wason (1968) task and Gigerenzer's (1998) task were couched strictly at the rational (or computational) level. In practice, a rational analysis cannot focus strictly on the rational level, but must also consider the algorithm and representation level. The reason for this is that sometimes the rational solution requires calculations that would be physically impossible for any system to perform. In such circumstances a mapping needs to be made from the results of the rational level into algorithms and representations that approximate the computations called for by the rational analysis. It is at the algorithm and representation level that compromises have to be made. In particular, assumptions have to be made about processing limitations. This ties in with Simon's (1956) notion of bounded rationality. That is, people are rational within the constraints of their ability to process information. For example, Simon (1990) pointed out that ignoring processing limitations suggests that knowledge of the rules of chess should lead to perfect play. If a person or machine had an infinite amount of time to contemplate a chess move, and the potential countermoves of the opponent to that move, and the countermoves to all the opponent's potential countermoves, *ad infinitum*, then the person could select an initial move that would inevitably lead them to check mate. The problem is that more board positions would have to be considered than there are molecules in the universe (Simon 1990). Instead, people and chess programs use heuristics to mimic the behavior of this rational strategy. Similarly, because of processing limitations, most rational analyses are approximations to the optimal solution.

There are some games, however, for which it is possible to know the game completely. For example, many adults know the moves and countermoves in tic-tac-toe to guarantee that they will at least not lose. Like tic-tac-toe, the Wason (1968) task is atypical. Thus the number of potential hypotheses and experi-

ments raised by Oaksford and Chater's (1996) analysis of the task is relatively small compared to the number raised by problems people often face in their daily and work lives. For example, consider again the problem of interpreting what the baby means when it says 'baba.' The potential number of hypotheses for what 'baba' means certainly exceeds the number of configurations of a chessboard. If Oaksford and Chater's information-seeking strategy were applied to the problem of vocabulary acquisition, where large numbers of hypotheses and experiments are potentially relevant, they would need to employ heuristics and build processing bounds into their system as well.

8. *Whether People are Rational Depends on Your Perspective*

This article demonstrated that there is no definitive answer to the question of whether people behave rationally. For example, the Wason (1968) task clearly demonstrates that people behave irrationally with respect to logic, while Oaksford and Chater's (1996) analysis shows that the behavior is rational when the context is broadened to encompass a person's '"normal" life conditions' (Brunswick 1943, p. 259). In short, people can be seen to behave rationally with respect to the environment, but appear to be operating irrationally with respect to a particular task, especially when the task that subjects are performing differs from the one intended by the experimenter.

See also: Functionalism, History of

Bibliography

Anderson J R, Matessa M 1992 Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning* **9**: 275–308

Anderson J R, Milson R 1989 Human memory: An adaptive perspective. *Psychological Review* **96**: 703–19

Anderson J R, Schooler L J 2000 The adaptive nature of memory. In: Tulving E, Craik F I M (eds.) *Handbook of Memory*. Oxford University Press, New York

Brunswick E 1943 Organanismic achievement and environmental probability. *Psychological Review* **50**: 255–72

Gigerenzer G 1998 Ecological intelligence: An adaptation for frequencies. In: Cummins D D, Allen C (eds.) *The Evolution of Mind*. Academic Press, San Diego, CA

Gigerenzer G, Todd P M, the ABC Research Group 1999 *Simple Heuristics that Make Us Smart*. Oxford University Press, New York

Manktelow K 1999 *Reasoning and Thinking*. Psychology Press, Hove, UK

Marr D 1982 *Vision*. Freeman, San Francisco

Oaksford M, Chater N 1996 Rational explanation of the selection task. *Psychological Review* **103**: 381–91

Oaksford M, Chater N 1998a *Rationality in an Uncertain World*. Psychology Press, Hove, UK

Oaksford M, Chater N (eds.) 1998b *Rational Models of Cognition*. Oxford University Press, New York

Simon H A 1956 Rational choice and the structure of the environment. *Psychological Review* **63**: 129–38

Simon H A 1990 Invariants of human behavior. In: Rozenweig M R, Porter L W (eds.) *Annual Review of Psychology*. Annual Reviews, Palo Alto, CA

Tversky A, Kahneman D 1974 Judgement under uncertainty: Heuristics and biases. *Science* **185**: 1124–31

Wason P C 1968 Reasoning about a rule. *Quarterly Journal of Experimental Psychology* **20**: 273–81

Watkins M J 1984 Models as toothbrushes. *The Behavioral and Brain Science* **7**: 86

L. Schooler

Rationalism

When philosophy consisted of set piece battles between grand schools of thought (such as realism, scepticism, or monism), 'rationalism' referred to the belief that human beings disposed of a faculty called 'reason' which gave them access to the structure of reality. Man, wrote Aristotle (384–322 BC), is a rational animal, but Aristotle recognized that reason had to be trained. In most human beings, the understanding was distorted by passion. Aristotle's teacher Plato (427–347 BC) had argued in *The Republic* and other dialogues that the world we experience was a confused copy of a world of forms or ideas which could be discovered by philosophical inquiry. His parable of the cave (*Republic*, Book VII) is the founding image of rationalism. In its classical Greek version, rationalism assumed that we might understand the structure of the universe by the power of reason, an assumption connected with the fact that rationalism began its philosophical career as a generalization of the procedures of mathematics. In his dialogue the *Meno*. Plato had argued that knowledge is accessible to rational inquiry independently of experience. Later, the Stoics argued that moral knowledge of the laws of nature is available to any rational creature who looks into himself.

Medieval philosophy revived Greek ideas and cross-fertilized them with Christian doctrine. St. Thomas Aquinas (1225–1274 AD) combined reason and revelation by taking his account of nature from Aristotle, and his account of higher things (or 'supernature') from Christian revelation. This synthesis, however, soon began to fall apart, and in modern times rationalism has commonly been taken to refer to rejection of Christian revelation because of rational criticism.

1. *Modern Rationalism*

This was not, however, the view taken by the great rationalist philosophers of the seventeenth century, for whom science was the model of understanding. Descartes invoked God's veracity as a foundation of his system, while Spinoza was a monist who argued

that there was only one substance in the world, which he called 'God or nature'—Deus sive natura. Reason and logic alone could generate certain knowledge. Yet even at this point the idea that nature was to be understood as a result of observation was emerging in thinkers such as Francis Bacon who sought 'to put nature to the torture' of experiment. The reason in which seventeenth century thinkers put their faith was less and less the human faculty supposedly giving us access to ends and values than what later came to be called 'instrumental reason' which revealed causes and linked ends and means. Confusingly, reason and rationalism have commonly been used to cover both ideas.

Philosophers had seldom entirely lost sight of the fact that understanding of the world depended upon observation and the seventeenth century growth of science brought this idea to the fore. Its germ lay in the medieval doctrine that we can only fully understand what we can make ourselves—a class of things thought to include not only artefacts but also society and language. After Bacon, the observational basis of knowledge was developed in the *Essay Concerning Human Understanding* by John Locke (1632–1704). Locke was sceptical of syllogistic reasoning, and denied in particular the Cartesian doctrine that reason was innate in man. Locke described himself as an 'underlaborer' in a generation of giants, by which he particularly referred to Isaac Newton the great physicist. Although the founder of empiricism as a coherent alternative to rationalism, Locke's philosophy did not fully free itself from rationalism, particularly in his account of 'powers of the mind.' Empiricism was refined in the work of Bishop Berkeley (1685–1753) and particularly David Hume (1711–76), a line of philosophical development that for some purposes culminated in the synthesis of Immanuel Kant (1724–1804), who argued that we can understand the phenomenal world only in terms of the set of categories (substance, causality, etc.) built into the human mind.

Thomas Hobbes (1588–1679) had used instrumental reason to generate the principles on which the state might be based. In *Leviathan* (1651), he attempted to demonstrate that civil order could be based only on obedience to a sovereign power. The Sovereign guaranteed laws of nature which are rationally derivable as axioms of prudence from the human propensity to avoid death. Hobbes's work was philosophical in that it analyzed the concept of a Commonwealth, but like most political philosophy, it contained some programmatic elements for practical life—positions clearly revealed in his *Behemoth* (written in 1668). In the next century, many thinkers of the French Enlightenment turned reason into an instrument for criticizing society. Writers such as Condillac and Helvetius adopted the empiricism of Locke, but by their emphasis on ideas, and their hostility to Catholicism and the *ancient regime*, they turned it into a form of rationalism. Reason led these *philosophers to* reject Christian revelation because of its conflict with logic (in the case of the Trinity) or science (in the case of miracles for example). Reason was also invoked against monarchy in favor of a revival of the classical idea of citizenship. These powerful movements of thought exploded in 1789 into the French Revolution, in which reason as the generator of natural rights was pitted not only against experience in epistemology, and against revelation in theology, but against tradition in politics. Reason was now a revolutionary principle seeking to reform everything in social and political life.

2. *Rationalism in History*

In the nineteenth century, rationalism invaded history by schematizing, as the doctrine of progress, the evolution of mankind from stone age life to the advanced condition of modern Europe. In rationalist terms, the progress of mankind led from superstition and prejudice towards forms of mutual understanding. Each such theory generally located the present as the hour before the dawn of a new era. Auguste Comte the positivist (1798–1857) thought mankind had moved from a theological stage through a metaphysical period and was now moving into a positivist era. The most famous of these adventures of ideas was that of Karl Marx (1818–88), who incorporated his account of human progress into a lossely connected narrative which was passed off as history. The drive to new rationalist accounts of society was inspired by technological optimism. Given that human beings had learned how to control nature, ought they not also to learn how to control society in the interests of human happiness? Rationalism thus became the vehicle of active programs of social reform, while those who took a more pessimistic view of human possibility were stigmatized as reactionary, or even irrational.

Such activist movements often collided with each other, and liberal and socialist programs for the inclusive improvement of the human condition competed for support with exclusivist movements such as those of the Fascists and the Nazis. In terms of the progress of human beings towards universal betterment, the oppression of totalitarian movements seemed not merely irrational, but actually irrationalist. Such irrationalism was often taken to be a form of nihilism, symbolised by the Francoist general whose cry was *Viva la muerte* ('Long live death'). Rationalism thus developed a contrast with irrationalism, understood as any general rejection of the liberal program of rights and the liberal search for universal happiness.

Rational understanding seeks to grasp the universal structure of whatever it studies, and is therefore more at home with theory and universal principles—with science and philosophy—than with history. Science and philosophy are thus pre-eminently rational in a sense in which history, as explanation in terms of

event and circumstance, is not. Further, philosophy assembles in a coherent manner whatever is known about an activity. If knowledge is power, then theoretical knowledge ought to give us power over the world. The most powerful rationalist impulse in recent centuries has been to create an understanding of human life which will allow us to build what is conceived to be a better world.

The problem is that theory and practice seem not to be related in this way. Plato's Republic is a masterpiece of political philosophy, but Plato's attempt to improve the politics of Syracuse was a failure. Nevertheless, the belief that successful practice could be encapsulated in a doctrine of a more or less philosophical kind has powerfully influenced the character of modern Western societies. The fusion of a philosophical method with practical ambitions is one of the central meanings of the elusive idea of ideology. Marx along with Friedrich Engels (1820–95), for example, directed their theory towards a new class of political actor, which they called the 'proletariat,' which was instructed in how to take over political power. Niccolo Machiavelli (1469–1527) was a pioneering exponent of this kind of rationalism, and Locke's *Two Treatises of Government* and Jeremy Bentham's (1748–1832) doctrine of utilitarianism might also be understood in this way as handbooks of political skill designed for new actors on the political stage.

3. *Critics of Rationalism*

The central principle of rationalism in politics, as analyzed by Michael Oakeshott (1901–90), is the sovereignty of technique. The skill of politics (which Lenin (1870–1924) and others thought remarkably easy once mystifications had been shed) can be learned from a book containing basic principles. Politics in this sense is thus no longer the conduct of public business by a small set of rulers, but a universal activity, not basically affected by the particularities of any given political culture. This is why for example Bentham's ignorance of Latin American politics was thought an actual qualification for the project of writing constitutions for that area.

The most common rationalist image of politics is that of a building program, in which abstract foundations logically generate a good society. Such projects are often based on abstractions such as democracy, justice, rights, or community. Rationalist programs focus on only one such desirable foundation, for the danger (as Plato had already recognized) is that two principles might come into conflict. The aim is to solve the problem of conflict, and both rulers and ruled need instruction from theoreticians (called 'education') in the philosophical foundations of the proposed good society.

Contemporary normative political philosophy is rationalist in its concern to analyze some basic norm of current liberal societies, commonly with an eye to improving practice. In the analysis of a just society advanced by John Rawls in *A Theory of Justice*, the future members of the society guarantee the rationality of the rules of justice because the rules have been agreed behind the famous 'veil of ignorance' in order to avoid the contamination of any particular knowledge. The equality of the Rawlesian citizens is thus matched by their vacuity.

The practical aspects of this form of rationalism are driven by instrumental reason in the sense that thinking takes off from whatever is construed as a problem (usually, indeed, a 'social problem') and the aim is to achieve a solution. Rationalism is sometimes criticized as assuming that all human beings are basically good, and only made bad by bad institutions, but the more fundamental assumption is that human beings are plastic entities who can be made into whatever type of social being is desired through the power of ideas. Politics thus becomes not so much social as attitudinal engineering; it requires removing the wrong ideas (racism, sexism, etc.) from society and restocking it with the right ideas. The general project of improving the world by 'modernizing' it is one of the most powerful of current rationalist projects.

In the second half of the twentieth century, rationalism as the search for the right foundations of society was challenged by people who said that the search was futile. There were no rational foundations, merely the changing responses of different sets of people. The derivation of this form of rationalism was from Heidegger (1889–1976) and Wittgenstein (1889–1951) but it was popularized in the United States by disciples of John Dewey such as Richard Rorty in *Philosophy and the Mirror of Nature* (1980). Radical critics charged that the whole rationalist project was designed to establish the hegemony of some specific group—capitalists, white males, commissars, and the like. Ethical opinions such as 'political correctness' and political projects such as multiculturalism were juxtaposed against rationalism. These revived forms of scepticism look like what rationalists would see as an irrationalist challenge to Western life, but, as is common in forms of scepticism, there are discernible traces of rationalist dogmatism in the egalitarian morality and communitarian politics taken for granted in the practices of the challengers.

See also: Aristotle (384–322 BC); Comte, Auguste (1798–1857); Hobbes, Thomas (1588–1679); Justice: Philosophical Aspects; Justice: Political; Locke, John (1632–1704); Marx, Karl (1818–89); Progress: History of the Concept; Rational Choice Explanation: Philosophical Aspects; Rationality in Society; Social Justice

Bibliography

Berlin I 1998 *The Proper Study of Mankind*. Farrar, Straus and Giroux, New York

Gellner F 1992 *Reason and Culture*. Basil Blackwell, Oxford, UK
Oakeshott M 1991 *Rationalism in Politics and Other Essays*. Liberty Press, Indianapolis, IN
Popper K R 1966 *The Open Society and its Enemies*, 5th edn., Routledge, London
Rorty R 1980 *Philosophy and the Mirror of Nature*. Basil Blackwell, Oxford, UK

K. Minogue

Rationality and Feminist Thought

One of the most profound reorganizations of the twentieth century has been a thorough re-evaluation of concepts of rationality as they have evolved in Western culture for the past three or so millennia (Gigerenzer et al. 1999, Labouvie-Vief 1994). This reorganization has been of profound interest to women, who have actively participated in the many lively debates about the nature of rationality. Models of rationality are, in large part, aimed at specifying the nature of humans and their desirable and less desirable characteristics, and, since these characteristics often were thought to find their most prototypical expression in men rather than in women, the issue of gender occupies a central place in this profound metatheoretical shift.

In response to traditional dualistic notions of rationality and the implied devaluation of women's rationality, some women have pointed to the fact that women traditionally were prevented from cultivating rationality (Hollingsworth 1926). Others have suggested that the traditional masculinist concept of rationality does not apply to women who utilize uniquely 'feminine' ways of knowing (Belenky et al. 1986, Gilligan 1992). Neither of these alternatives, however, transcends the traditional dualistic concept of rationality. Thus, a further suggestion has been that we move away from dualistic concepts of rationality altogether. Accordingly, concepts of reason are being transformed and new models are emerging that acknowledge that rational and nonrational processes mutually enhance each other and contribute in the process of knowing (Labouvie-Vief 1994).

1. *Rationality and Women's Mind in History*

Concepts of rationality have been central to definitions of human nature. Often couched explicitly in terms of the mind and its relationship to emotions, social context, and culture, rationality was defined dualistically—as that which transcends these categories, i.e., as conscious thought, abstract thinking, and universal laws of truth (Labouvie-Vief 1994). Its relation to gender often was more implicit, but the rational being was a 'man of reason' (Lloyd 1993). This association of reason with masculinity—and of femininity as a lack of reason—widely pervaded Western intellectual history and is visible not only in philosophy and science, but also in art and literature, myth and religion, and ultimately, the social structures and social regulation processes that have supported these representations.

1.1 *The Gendered Mind in Philosophy*

The dualistic theory of rationality and the mind became institutionalized when Greek philosophers began to speculate about human nature. These efforts culminated in the work of Plato, who proposed a two-layer view of human nature and reality to supplant more ancient views (Labouvie-Vief 1994). For more ancient views of the self as a collection of motor actions, bodily processes, or mythic–divine injunctions, Plato substituted a language of a self no longer identified primarily with its bodily processes and concrete actions, but of a mental agent different from its bodily manifestations. The new language was that of a self who was the author of those actions, of a psychological causal agent who was at their center and who was responsible for them.

Plato's dualistic, two-layered conception of the individual contrasted a layer of mind, abstract thought, soul, spirit, or ideals to a layer of body, matter, and enslavement to concrete sensory and organismic happenings. These layers were arranged hierarchically, with mind-spirit forming the superior pole, body, and senses the inferior one. That hierarchical worldview also became associated with gender. Thus, in the *Symposium*, Plato argued that the spiritual love between men who create ideas is more valuable than the physical and material love with women, since women's creativity is merely one of bodily procreation rather than the creation of universal ideas in which men participate (Labouvie-Vief 1994).

1.2 *Women's Mind in Western Science*

The association of the masculine with mind and spirit, and of the feminine with body and matter, widely came to characterize thinking about and representations of women. It has characterized not only the view of most philosophers but has also has remained basic to other theoretical accounts of human nature. For example, Aristotle (*De Generatione Animalium*) thought that women's contribution to fertilization consisted of a passive and material principle rather than an active and spiritual masculine one (Lerner 1986). Freud proposed that women's lack of phallic sexuality made their sexual and psychological orientations more passive and masochistic (Freud 1925). Others taught that the anatomy of women's brains made them less capable of conscious, rational thought and of extraordinary creativity or achievement (Bem

1993, Labouvie-Vief 1994). Even in the twentieth century, the notion became popularized by de Beauvoir (1952), who argued that woman is closer to nature because physiology makes her so: Her mode of creating is by submitting to a natural bodily function, by remaining enslaved to her biological fate, while the man creates by 'acts that transcend his animal nature.' By creating culture, the man is thought to be involved in a new, freer form of creation, one that transcends what is given organismically and shapes the future rather than passively suffering it.

This gendered dualistic notion of mind has been influential in shaping theoretical notions of development. Traditionally, development has been construed as an onward rush, a heroic and energetic striving for mental ideals and control, a triumph of conscious, abstract forms of thinking, a victory of rationality over emotions and the body. Since these features of the mind also are identified with masculinity, femininity was represented as an inferior and less developed state of being—a form, as it were, of mental and physical castration. Freud (1925) was most explicit in equating development of mind and identity with a progression from the domain of the castrated mother to the potent and genital father, but most major theories of development of the past followed this same pattern of the devaluation of the emotive, imaginative, and sensory aspects of the self (Labouvie-Vief 1994).

1.3 Women's Mind in Myth, Art, and Religion

The notion of the masculine as identified with mind, spirit, and transcendence is also pervasive in mythology and religion, literature, and art (Labouvie-Vief 1994, Ortner and Whitehead 1981, Rosaldo and Lamphere 1974). In all of those, one polarity of human functioning tends to be idealized and even inflated as 'masculine': it is based on an imagery of rise, light, and sun, and it symbolizes such psychological qualities as leadership, rational self-control, self-assertiveness, individuality, and willfulness. The other polarity, in turn, is degraded and devalued: its imagery is based on falling, being conquered and surrendering, and engulfment in dark spaces, and it symbolizes suffering, powerlessness, and entrapment in unconscious processes. In mythology and religion, such imagery often occurs as the theme of the overthrow of original female divine figures by the male godheads or heroes who came to be the original creators, creating the world not organically but conceptually, by pronouncing a word (Lerner 1986). A well known example of this process of the replacement of the feminine principle is offered by the great Greek tragedies by Aeschylus and Euripides centering on the myths of Agamemnon and Orestes. These myths address the question of which of two murders was a more heinous act. Was it Klytemnestra's killing her husband Agamemnon to revenge his sacrificing their daughter Iphigenia to the gods so as to secure a good wind for his war ships. Or was it their son Orestes's killing of his mother to avenge the murder of his father? The answer was that Orestes could be forgiven because he represented a new masculine order that superseded an earlier feminine order—a new paradigm in which fatherhood was said to have replaced motherhood as a principle of creation (Labouvie-Vief 1994).

2. Explanations of Dualistic Theories of Mind

How universal is the equation of masculinity with mind and rationality, and of femininity with an 'older' emotional and organic order of being? Much writing has suggested that such patriarchal ideologies in actuality came to replace more women-focused ones of previous times. Originally proposed by Bachofen in 1861 (Duby and Perrot 1992/93), this 'matriarchy' thesis holds that there must have been, in prehistoric times and before the rise of the major patriarchal societies, the reign of a maternal principle. However, the historical accuracy of this thesis has recently been doubted. Instead, a number of anthropological, sociological, and psychological explanations have been proposed.

2.1 Anthropological Evidence

Most historical and anthropological evidence (Duby and Perrot 1992/94, Ortner and Whitehead 1981, Rosaldo and Lamphere 1974) does not support the view that earlier and/or less civilized societies granted women more power universally. On the contrary, strong patriarchal systems in which women are devalued along a mind–nature dimension exist worldwide in preliterate societies. Often such societies also have myths of early matriarchies that eventually were superseded by the rule of man. However, such matriarchy myths can be part of an ideology that devalues women (Rosaldo and Lamphere 1974). Around the world, such myths often tell of the past bad behavior and evil of women that eventually was ended only when men overturned the power of women and took control. In this way, matriarchy myths can be used to justify the male order. They form part of a cultural code distinguishing men from women in terms of morality and authority, they incorporate the values that grant men more power in social life. Thus, even though they talk about a time before the current social order, they fix that order as legitimate and invariant.

2.2 Sociological Mechanisms

Although there is considerable variation in women's behavior and activities across cultures and historical times, there are nearly universal differences in what is believed to constitute the nature of women (Brown

and Gilligan 1992, Labouvie-Vief 1994, Rosaldo and Lamphere 1974)—and, correspondingly, the nature of men (Gilmore 1990). In many cultures, women are held to be nurturers, to be gentle and concerned with caring for others rather than self-expression, they are considered passive and unaggressive and to readily subordinate themselves to those in power. They are felt to be closer to children—more emotional and expressive, less capable of rational decision-making and abstract thought. They are also believed to be poorly suited for positions of leadership and power (Bem 1993, Gutmann 1987, Ortner and Whithead 1981, Rosaldo and Lamphere 1974). Sometimes these beliefs are attributed to biological bases of gender, but many think they root in the social roles and realities of women's lives.

In most societies women are subordinated socially or, at least, they do not hold publicly recognized power and authority surpassing that of men (Ortner and Whitehead 1981, Rosaldo and Lamphere 1974). In some societies, individual women may have achieved considerable power and recognition, but in general all societies tend to exclude women from certain crucial economic, political, and intellectual activities. Even in societies where women exert considerable power by doing much productive work and controlling resources, their activities are not associated with the same social status as those of men. Their roles as wives and mothers are associated with fewer powers and prerogatives than are the roles of men (Bernard 1987, Duby and Perrot 1992/93, Ortner and Whithead 1981, Rosaldo and Lamphere 1974). In turn, this social devaluation is often attributed to some major biological deficiencies (Bem 1993, Labouvie-Vief 1994). These same 'lacks' also were thought to make women eminently suitable to take on positions subservient to men.

2.3 *Psychological Mechanisms*

The near universality of the devaluation of women's mind also has suggested that the construal of women as mentally inferior and nonrational may be rooted in psychological processes through which gender is defined. Psychoanalyst Stoller (Herdt and Stoller 1990) and sociologist Chodorow (1978) suggested that these processes render the gender development of boys particularly vulnerable. This vulnerability results from the mother's role in early development. For both boys and girls, the mother defines an early core of identity, which in the case of the boy's development becomes problematic. He, unlike the girl, needs to change his core identity from 'feminine' to 'masculine', a change that leaves a life-long vulnerability of self. As a result, men may defensively devaluate women as shown in cultural practices (usually, related to initiation rites) aimed at defining the unique nature of the 'masculine' as a way of creation unique to men and 'better' than the 'feminine.' Complementary to the definition of manhood, womanhood in many cultures is defined by practices aimed at creating a sexual and social object to the action of men—whether it is direct genital mutilation or isolation from participation in public life through withholding education and high status work (de Beauvoir 1952, Labouvie-Vief 1994, Ortner and Whithead 1981).

However, such mechanisms may be themselves relative to the family structure that operates in a society strongly divided along gender lines and devaluing women (Bem 1993, Labouvie-Vief 1994). But with family structure and gender roles changing to create more shared and equitable activities among men and women, these domains are becoming less clearly related to gender, *per se*.

3. *Rationality and Women's Minds*

Traditional dualistic theories of rationality, with their devaluation of femininity and inflation of masculinity, have influenced deeply women's process of growing up. Whereas masculine development is associated with increases in abstract thought, in individuality, independence, and self-confidence, women often have experienced their development as renunciation, surrender, and victimization—a loss of voice, a blunting of mind, a process of silencing (Bem 1993, Brown and Gilligan 1992, Labouvie-Vief 1994). Such struggles of women with their own sense of creativity is expressed very poignantly in the lives of many writers of the twentieth century, such as Sylvia Plath or Virginia Woolf.

As educational and employment opportunities have opened up for women and as women have begun to participate in defining themselves, research has upheld few objective differences in intellectual status between men and women. Although women tend to score somewhat higher on verbal and men on mathematical abilities, these differences are overall small (Linn and Hyde 1989). Moreover, since Hollingsworth's (1926) pioneering work showing that gender differences in intellectual can be explained usually in terms of differences in educational opportunities, research has documented that over time, as disparities of opportunity are declining, gender differences in intellectual functioning are becoming smaller, too (Linn and Hyde 1989).

Yet despite evidence that female intellectual endowment does not lag behind that of men, women on average continue to lag in achievement behind men. Moreover, this gender lag increases in the course of development. Indeed, in early childhood, girls surpass boys on most measures of school achievement. However, as they grow older, girls often are found to retreat from achievement-related challenges; standing out for women becomes a source of conflict and shame, and they attempt to hide their excellence (Brown and Gilligan 1992, Labouvie-Vief 1994). Thus, women's feelings of shame and discounting of their achievement

persist well into the present day. However, in many cultures women's transition into later life is often characterized by a degree of freeing from the cultural constraints of 'womanhood.' No longer defined as sexual objects, women often move from a world of confinement to one of daring openness, of a release from their earlier interiority. Thus, from young adulthood to midlife, women have been found to become less traditionally 'feminine' but more dominant, independent, objective, and intellectually oriented. In many societies this figure of the older woman emerges as the woman of power who holds authority over the family and who can control healing, magic, witchcraft, and other powerful (and often dangerous) resources (Gutmann 1987, Rosaldo and Lamphere 1974).

4. *Transforming Concepts of Mind and Gender*

This review has shown that traditional concepts of rationality have implied a strong gender bias, one that has corresponded with the stereotyping of women as nonrational and more closely bound to nature and their bodies. Such a feminist critique ultimately does more than restore to women the claim to be rational, or to claim for them a unique form of knowing that is different from the rationality of men. Rather, the intricate association of traditional concepts of gender and rationality reveals that the very claim that women lack rationality is problematic, since it was on a narrow concept of rationality in the first place. Thus, the association of mind and gender indicates the need to reorganize the very meaning of 'rationality' in theories of mind and knowledge. A transformation of the concept of rationality is not just a feminist enterprise. Criticisms of traditional views of rationality have been a mainstay of the intellectual history of this century, which systematically has eroded beliefs in dualistic concepts of rationality (Baynes et al. 1987). The mind and knowledge are no longer believed to stand secure above nature and human foible. Instead, we are witnessing a growing belief that the dimensions of the mind and self that were abandoned in the classical view must in fact be included in a viable theory of mind and knowledge. Emotion and intuition, time and process, variability and disorder—all the elements previously devalued as feminine have taken a forefront place in philosophical discussions, and are being incorporated into new theories of the mind. Thus, reason also has an immanent pole and must be explained through the emotional propensities, cultural scenarios, linguistic practices, and narrative and symbolic structures of meaning.

Freud's (1925) original proposal that rational processes are built upon nonrational processes was one early voice warning that rationality ultimately must include a balance between our biological nature and our strivings for transcendent realities. It was he who pointed out the paradox that 'rationality,' when one-sided and unbalanced, in fact is a sign of the defensive break through of 'irrational' processes. However, Freud's very concept of 'irrational' processes still implied remnants of a dualistic and oppositional view, since it relegated these processes to a problematic status. Instead of such dualism, notions of balance and exchange have been carried forward into modern theories of the mind. In those theories, a tremendous part of knowledge is, in fact, based on processes that are profoundly adaptive, yet nonrational. Such theories follow Darwin's insight that biology provides the basis for the behavioral patterns—such as emotions—that are of the most urgent adaptive significance. More generally, we are realizing that an important contributor to knowledge are other nonrational decision systems or 'heuristics' that allow us to deal with information processing in situations where time and resources are limited (Gigerenzer et al. 1999, Labouvie-Vief 1994).

Sometimes the revaluation of the nonrational is taken as a dethroning and destructive critique of rationality. Instead, calling for a transformation of the concept of rationality implies that it merely constitutes one important pole or function of the mind. Thus body, nature, emotions, and concrete existence vs. mind, culture, reason, and necessary, deductive law no longer are considered in hierarchical terms. Instead, the mind becomes a cooperative interplay of these two sets of processes, rational and nonrational. Such dual processing theories of the mind are, in fact, proliferating and becoming the standard in theories of the mind (Chaiken and Torpe 1999, Gigerenzer et al. 1999, Labouvie-Vief 1994).

Such a reorganization of theories of the mind inevitably also involves a reexamination of the role of gender in the order of things. Indeed, such a reevaluation is visible in many strata of contemporary thought and life, visible not only philosophy and scientific theory, but also religion, literature, and art. These efforts allow us to acknowledge that rationality and non-rationality are processes of a mind that is, at last—and at least theoretically—part of nature and independent of gender.

See also: Androgyny; Child Care and Child Development; Deconstruction: Cultural Concerns; Feminist Economics; Feminist Epistemology; Feminist Ethics; Feminist Legal Theory; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Foucault, Michel (1926–84); Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender and School Learning: Mathematics and Science; Gender and the Law; Gender, Feminism, and Sexuality in Archaeological Studies; Gender-related

Development; Heterosexism and Homophobia; Homosexuality and Psychiatry; Irrationality: Philosophical Aspects; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Love and Intimacy, Psychology of; Mind–Body Dualism; Postmodernism: Philosophical Aspects; Queer Theory; Rationalism; Rationality in Society; Regulation: Sexual Behavior; Sex-role Development and Education; Sexual Behavior: Sociological Perspective; Sexual Orientation and the Law; Sexuality and Gender; Theory: Sociological; Transsexuality, Transvestism, and Transgender

Bibliography

Baynes K, Bohman J, McCarthy T T (eds.) 1987 *Philosophy: End or Transformation*. MIT Press, Cambridge, MA

de Beauvoir S 1952 *The Second Sex*. Knopf, New York

Belenky M F, Clinchy B M, Goldberger N R, Tarule J M 1986 *Women's Ways of Knowing*. Basic Books, New York

Bem S L 1993 *The Lenses of Gender*. Yale University Press, New Haven, CT

Bernard J S 1987 *The Female World from a Global Perspective*. Indiana University Press, Bloomington, IN

Bettelheim B 1962 *Symbolic Wounds: Puberty Rites and the Envious Male* (rev. edn.). Collier Books, New York

Brown L M, Gilligan C 1992 *Meeting at the Crossroads: Women's Psychology and Girls' Development*. Harvard University Press, Cambridge, MA

Chaiken S, Trope Y (eds.) 1999 *Dual-process Theories in Social Psychology*. Guilford Press, New York

Chodorow N 1978 *The Reproduction of Mothering*. University of California Press, Berkeley, CA

Duby G, Perrot M (eds.) 1992/94 *A History of Women in the West*, Vols. 1–4. Harvard University Press, Cambridge, MA

Linn M C, Hyde J S 1989 Gender, Mathematics, and Science. *Educational Researcher* **18**(8): 17–27

Freud S 1925 Some psychological consequences of the anatomical differences between the sexes. In: Rieff P (ed.) 1963 *Sexuality and the Psychology of Love*. Macmillan, New York pp. 183–93

Gigerenzer G, Todd P M, ABC Research Group 1999 *Simple Heuristics that Make Us Smart*. Oxford University Press, New York

Gilmore D D 1990 *Manhood in the Making: Cultural Concepts of Masculinity*. Yale University Press, New Haven, CT

Gutmann D 1987 *Reclaimed Powers: Toward a New Psychology of Men and Women in Later Life*. Basic Books, New York

Herdt G H, Stoller R J 1990 *Intimate Communications: Erotics and the Study of Culture*. Columbia University Press, New York

Labouvie-Vief G 1994 *Psyche and Eros: Mind and Gender in the Life Course*. Cambridge University Press, New York

Lerner G 1986 *The Creation of Patriarchy*. Oxford University Press, New York

Lloyd G 1993 *The Man of Reason: 'Male' and 'Female' in Western Philosophy* 2nd edn. University of Minnesota Press, Minneapolis, MN

Ortner S B, Whitehead H (eds.) 1981 *Sexual Meanings: The Cultural Construction of Gender and Sexuality*. Cambridge University Press, New York

Rosaldo M Z, Lamphere L 1974 *Women Culture and Society*. Stanford University Press, Stanford, CA

Schott R M (ed.) 1997 *Feminist Interpretations of Immanuel Kant*. Pennsylvania State University Press, University Park, PA

G. Labouvie-Vief

Rationality in Society

Human behavior is rational when appropriate to its goals. The brain provides the biological foundation for human rationality; it accepts sensory stimuli from both the natural and social environments and stores knowledge, beliefs, and values that guide the thought processes leading to action. This article focuses upon the social environment of human rationality.

Rationality has been an object of scholarly analysis since Greek times, but largely as part of ethics, in its concern with goals, sound reasoning and logic. Empirical study of human rational behavior is mostly quite recent, although we find such significant precursors as Machiavelli. Even Adam Smith provides no explicit theory of rationality in his *Wealth of Nations*, and such theories begin to appear only in the nineteenth century (e.g., Cournot 1838), initially in economics and, much later, in sociology (Parsons 1937, see also *Action, Theories of Social*). In the twentieth century, utility theory took shape in economics, while more realistic but less formal theories of rationality developed in the other social sciences.

1. Theories of Rationality

The social context molds behavior in three main ways. First, it largely determines what actions are available. Second, choice must try to anticipate the reactions of others to contemplated actions. Third, people's society is the major source and shaper of the knowledge, skills, and goals they use in choosing actions.

Hence, although a large part of today's understanding of rationality derives from psychological experiments with individuals, most of the contents of individuals' knowledge and beliefs, and their cognitive processes derive from social sources. Moreover, experimental stimuli themselves often are stated in language that has rich social content. Thus, even 'individual' psychology is the study of highly socialized beings (Simon 1992). The entrepreneur and consumer of economics, the voter and legislator of political science, the husband and delinquent, and soldier of sociology and anthropology are deeply social.

1.1 The Rationality of Experts

Cognitive psychology has acquired substantial knowledge about the decision-making of experts and novices

in such professional domains as medical diagnosis, chess-playing, and scientific discovery, and has successfully simulated these behaviors on computers (Anderson 1999, Newell and Simon 1962, Shrager and Langley 1990; see also *Expert Memory, Psychology of*; *Scientific Reasoning and Discovery, Cognitive Psychology of*; *Chess Expertise, Cognitive Psychology of*; *Medical Expertise, Cognitive Psychology of*). Within the professional domain, an expert employs two main procedures: (a) recognizing familiar situations and patterns that give access to knowledge in memory about handling these situations; and (b) searching selectively for satisfactory problem solutions, guided by learned rules of thumb, or heuristics. The same procedures are used to find better problem definitions and representations, instruments and tools for observing or reasoning, experimental designs, and search strategies. They even account for the problem solving that is called 'intuitive,' 'insightful,' or 'creative.'

1.2 Rationality as Utility Maximization

Utility maximization, the best developed formal theory of rationality, which forms the core of neoclassical economics, does not refer to the social context of action (see also *Decision Theory: Classical*). It postulates a utility function, which measures the degree to which an individual's (aggregate) goals are achieved as a result of their actions. The rational actor chooses the action, from among those given, which maximizes utility. If the actor's goals are food and sleep, then rationality calls for choosing the attainable combination of food and sleep that yields the greatest utility. The core theory does not specify the content of the utility function (Simon 1983).

Without a concrete meaning for utility, the theory has few refutable consequences; its relevance to real situations depends mainly on adding auxiliary assumptions. This is not wholly true, for the theory does assume constancy of the utility function over sequences of choices. Many experiments have shown that this implication is testable—and frequently false (Tversky and Kahneman 1974). Human actors often make sequences of choices that are incompatible with any consistent utility function.

In order to predict more than consistency in any particular situation, additional strong assumptions must be added. Economic applications usually add the vastly oversimplified assumption that utility is measured by total income or wealth, which, in turn, may be expressed in terms of goods possessed. Additional assumptions introduce markets, in which goods can be exchanged (for each other or for money), and firms, which consume goods to produce other goods.

Without auxiliary assumptions, the theory is also silent about the range of available alternatives and the processes used to find the optimal action. It says nothing about how to select auxiliary assumptions, or what empirical support they should require. As a result, assumptions are often introduced in order to preserve or enhance mathematical tractability.

To produce an empirically sound theory, a much richer set of empirically based auxiliary assumptions must be provided, to take account of the actual limits on human thinking, and the influences of its social context (Green and Shapiro 1994).

1.3 Forms of Rationality

An action may be regarded as rational because of the way it was chosen (procedural rationality), or because of its outcomes (substantive rationality) (see also *Decision Research: Behavioral*; *Bounded Rationality*; *Decision and Choice: Economic Psychology*; *Heuristics for Decision and Choice*).

Attributions of rationality may or may not respect the limits of actors' abilities to adapt actions to goals. Utility-maximizing behavior, as discussed previously, selects an action that is among those objectively available, attains a real-world optimum, and is of a comprehensive function of wants and needs (utility function).

As an alternative, bounded rational behavior selects some action among those that memory, observation, and thought bring to attention. It is aimed at attaining outcomes that satisfice (attain aspirations) for the wants and needs that have been evoked (Conlisk 1996). The expert behavior described in Sect. 1.1 is an example of bounded rationality.

There is a close affinity between optimizing and substantive rationality, and an affinity between bounded and procedural rationality. The utility-maximizing optimum is independent of the choice process; the outcome is all that counts. In bounded rationality, the choice depends upon the process leading up to it. The social sciences, other than neoclassical economics, usually explicitly incorporate human cognitive limitations and social influences into their theories, using some version (seldom highly formalized) of bounded rationality. Bounded and procedural rationality require generating courses of action, assessing their possible outcomes under different futures and estimating how much satisfaction they will attain. They postulate procedures that respect human limits on knowledge of present and future, on abilities to calculate the implications of knowledge, and on abilities to evoke relevant goals (Simon 1983).

1.4 Habits, Emotion, and Motivation

Although dictionaries often define rational action narrowly, as the action chosen by reasoning, numerous actions that are rational (i.e., appropriate to goals) are taken without conscious deliberation or any deliberation at all. Rather, these rely on heredity responses,

or prior knowledge and belief (e.g., see also *Perception without Awareness, Psychology of*; *Emotion in Cognition*). Evolution has seen to it that eyes blink when something comes flying at them. Similarly, but as a result of learning and experience, when a car suddenly stops, the driver in the car behind applies brakes without deliberation. These goal-oriented, but unreflective actions make up much of the activity of everyday life, and greatly expand the capacity for (boundedly) rational response. Rationality is inextricably interwoven with emotion and motivation. Action is directed toward goals by such motivators as hunger, sexual drive, anger, fear, pride, and others. These are represented by the aspirations of bounded rationality. Evoked into consciousness, or activating 'automatic' responses, they provide the goals for rational choice. Additionally, cognitive processes cannot handle large quantities of information simultaneously. Deliberative and conscious thoughts have to pass through the narrow straits of short-term memory, which hold only a few symbols (approximately six), and can attend to only one thing at a time (or perhaps two or three, by alternation of attention over intervals of 1–2 seconds). Recent research using functional molecular resonance imaging to record neuronal activity has shown that even simple acts (reading a short sentence) employ a sequence of neural processes (see also *Short-term Memory, Cognitive Psychology of*; *Working Memory, Psychology of*).

Because human beings must satisfy many drives, rationality fixes attention on particular goals, usually until these are satisfied, but shifts attention to other goals when they become more pressing. Interruption is triggered both by sensory processes, such as peripheral vision, and by emotions and inner drives, such as fear and hunger. So rationality, is not just a cognitive function, but depends on motivation and emotion to control attention.

2. *Rationality in Social Context*

The social context of action that plays such an enormous role in determining choices will now to be examined more closely.

2.1 *The Sources of Knowledge*

Most of the knowledge that enters into either habitual responses or deliberate choice is received from other persons, beginning at birth. Direct experience of the nonhuman world contributes only a small fraction of what people know and believe. Language, a major tool of thought, is a social artifact (Simon 1992).

Of course, what people know is also strongly influenced by the natural environment, some of it sensed directly. Nonetheless, much of one's knowledge of this environment is acquired by social processes, not direct sensation; sensory stimuli are interpreted by social information. People believe that cholesterol is bad for them, but their belief derives from doctors, the media, and friends, not their own laboratory experiments. The same rains fall on humankind around the world, wetting anyone it strikes. It enters into each person's beliefs and knowledge in many ways, and social processes have created a wide range of different expedients for mitigating its discomforts and using its waters. The range of available actions and the criteria and procedures for choosing among them are largely circumscribed by the actions, criteria, and procedures discovered and propagated by each society.

2.2 *The Social Sciences are Historical Sciences*

Because of the social origin of knowledge, the social and behavioral sciences are, essentially, historical, for the behavior of each person changes with the enveloping society's states of belief and knowledge. Functionalism in anthropology and sociology has pointed out the common functions shared by all societies, but also the variety of ways in which these functions are performed in particular societies, depending on their values, beliefs, and knowledge.

Of course, dependence on history is not peculiar to the social sciences. Historical laws are found throughout natural science: in modern astrophysics, the Big Bang and its sequelae; in geology and oceanography, mountain formation and continental drift; evolutionary biology is wholly historical.

Today, the social sciences differ greatly in the extent to which their theories attend to history. Historical change in basic laws has played only a small role in utility-maximizing economic theory. The central theoretical structure in economics today, as in Adam Smith's day, is an idealized set of markets with vestigial business firms (each characterized by a production function). Perhaps these roughly match Adam Smith's world, but they are a far cry from the economy of large international business corporations throughout today's world; and this model does not explain the growth (or even the possibility) of these large organizations.

On recognizing its historical dependence, a science must reexamine the meaning of natural law. It must seek to find invariant underlying relations that account for continuing change as well as equilibrium in dynamic systems. In the social sciences, temporal structural change is still absent from most formal theory.

2.3 *Social Determinants of Motives*

Social modification of motives and actions is usually dealt with in contemporary social science by introducing evolutionary mechanisms (Nelson and Winter 1982). In most current versions of Darwinism, the sole

effective motive is to increase fitness (essentially rate of population growth); altruism will not compete. This has often been interpreted as implying that only selfish human motives survive, and that, consequently, maximization of utility means maximization of income or wealth, profit of the business firm, or power (the latter in Public Choice Theory in political science and sociology). In this view, only the special genetics of ants and bees, and some other social insects, allow nonpropagating offspring to exhibit altruism to close kin.

However, it has been shown (Simon 1997) that, in the presence of bounded rationality, altruism may increase the fitness of altruists, and, hence, may survive and grow within a society. If bounded rationality admits a large capability for social learning, persons willing and able to learn (docile), will have a large fitness advantage over others. By natural selection the society will help individuals acquire knowledge and beliefs that usually enhance their individual fitness, but are sometimes advantageous to the society's fitness at the individual's expense (e.g., volunteering for military service). As long as social influence, on balance, increases individual fitness, altruism can also thrive, contributing to the fitness and growth of the society (see also *Altruism and Self-interest*).

2.4 Altruism and Group Identification

Perhaps the most important human expression of altruism is identification with, and loyalty to, groups ranging from the family, through ethnic or religious groups, to whole societies or nations. Group identification is a major determinant both of goals (defining the 'we') and of knowledge and beliefs, for both are formed in the groups that people associate with in work and leisure. Identification is the primary basis for the intense within-group collaboration and between-group conflict that is an enduring characteristic of human interaction. It is also one of the bases for the ability of large organizations to coordinate human efforts efficiently (see also *Identity: Social*).

2.5 Organizations

In a modern society, formal work organizations—companies, government agencies, schools—provide environments for a large fraction of human waking hours (Simon 1997). This has presumably come about because organizations can, through appropriate specialization and coordination of activities, enormously increase individual productivity. Effective specialization requires dividing up activities in such a way that each specialty calls on only a small part of the knowledge and skill that is required for the whole, and each component is mainly concerned with the initial inputs and final outputs of the others, and need not be involved with their internal processes.

At the same time, it is rarely possible to obtain complete independence among components, so that some provision, often involving the exercise of authority, must be made for securing coordination where it is required (the axles that one department makes must fit the wheels made by another). The theory of organizations, in business and public administration, is concerned with the architectural arrangements that secure a good balance of specialization and coordination, and the means of using authority, organizational identification and rewards to secure acceptance of the structure by those who inhabit it. Because of the difficulty of measuring the precise contributions of individuals' work efforts to an organization's product and success, economic incentives cannot account for the high level of support for organization goals that is often observed (for example, in the behavior of employees who do not share in profits or bonuses). Much of this support can be attributed to organizational identification, and it is one of the novelties of the modern world (with perhaps some precedents in armies, even of the distant past) that strong loyalties can be evoked by formal organizations as well as by more traditional social groups. Human attachment to organizational goals has been a major force in converting the market economy of Adam Smith's time into the very different organizations-cum-markets society of the present (Simon 1997).

Recent developments in the formal theory of complex organizations also suggest that a delicate balance between specialization and coordination is essential if large systems are to innovate and continue to evolve successfully. Excessive need for coordination prevents one unit from evolving without damaging the efficiency of others; excessive independence of components limits the sophistication of the products of the whole organization.

2.6 Social Structure

Although more nebulous than formal organizations, the framework of laws and customs of a whole society also provides an important object of identification for members, and in particular a stable cognitive framework within which a creature of bounded rationality can operate with some probability of predicting correctly the consequences of action. An important component of the structure of most societies is a system of markets and social practices that maintain them. Every modern industrial society is a composite system of organizations-cum-markets, where the markets permit a considerable degree of decentralization and independence of action of the individual organizations. The division of labor in the social sciences between organization theorists, on the one hand and economists on the other has left something of a lacuna in our understanding of the interactions between these two prominent kinds of social structure.

2.7 Mutual Expectations: Game Theory

What concept of rationality suits situations where several or many actors have competitive but highly interdependent goals? Cournot raised the problem in 1838 of two or more firms competing in an imperfectly competitive market, and threatening market instability by trying to outguess each other. Although von Neumann and Morgenstern (1944), in their *Theory of Games and Evolution*, proposed a solution to the problem, they defined rationality in a rather implausible way that did not guarantee unique choices or equilibria (Aumann and Sergiu 1992). Although their work has had a large impact, especially in economics and international relations theory, its most important result has been to show the near-impossibility of framing a univocal 'logical' definition of rationality in conditions of mutual uncertainty and interdependence. The ambiguities can sometimes be resolved by retreating from utility maximizing to bounded rationality, and by determining empirically which of their many options humans in fact employ in these situations (Axelrod 1980).

Political revolutions provide excellent demonstrations of the importance of stability of mutual expectations to the operation of a society. Revolutions dissolve expectations about what reactions actions will evoke from opposing groups. This 'gaseous' dissolution of social structure is ultimately reversed by the ascendance to power of new groups with internal identifications strong enough to remove the uncertainties (see also *Revolutions, Sociology of*).

3. Conclusion

Today, considerable effort is going into reexamining the concept of rationality, questioning the assumptions of utility maximization theory and seeking foundations for an empirically grounded theory of bounded rationality that gives proper attention to the social environment of choice. There is considerable new experimental work, particularly experiments with markets and games where 'outguessing' is important, but much more empirical study, in the field as well as the laboratory, will be needed to arrive at a theory that takes full account of the limits on human cognitive capabilities, and the social context which informs human reasoning.

Bibliography

Anderson J R 1999 *Cognitive Psychology and its Implications*, 5th edn. WH Freeman, New York

Aumann R J, Sergiu H (eds.) 1992 *Handbook of Game Theory*. North-Holland, Amsterdam

Axelrod R 1980 *The Evolution of Cooperation*. Basic Books, New York

Conlisk J 1996 Why bounded rationality? *Journal of Economic Literature* **34**: 669–700

Cournot A 1838 *Researches into the Mathematical Principles of the Theory of Wealth*. Reprinted 1960, Augustus M. Kelley, New York

Green D P, Shapiro I 1994 *Pathologies of Rational Choice Theory*. Yale University Press, New Haven, CT

Nelson R R, Winter S G 1982 *Evolutionary Theory of Economic Change*. Harvard University Press, Cambridge, MA

Newell A, Simon H A 1962 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Parsons T 1937 *The Structure of Social Action*. McGraw-Hill, New York

Rapoport A, Chammah A M 1965 *Prisoner's Dilemma*. University of Michigan Press, Ann Arbor, MI

Shrager J, Langley P 1990 *Computer Models of Scientific Discovery*. Morgan Kaufman, San Mateo, CA

Simon H A 1947 *Administrative Behavior*, 4th edn. The Free Press, New York

Simon H A 1983 *Reason in Human Affairs*. Stanford University Press, Stanford, CA

Simon H A 1992 What is an 'explanation' of behavior? *Psychological Science* **3**: 150–161

Simon H A 1997 *Models of Bounded Rationality Vol 3: Empirically Grounded Economic Reason*. MIT Press, Cambridge, MA

Tversky A, Kahnemann D 1974 Judgment under uncertainty: heuristics and biases. *Science* **185**: 1124–31

von Neumann J, Morgenstern O 1944 *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

H. A. Simon

Reactance, Psychology of

The theory of psychological reactance (Brehm 1966) pertains to the reactive human tendency to protect or defend threatened freedoms. Unlike most conceptions about motivation, reactance theory does not speak of the human as arriving at a particular goal, achieving clarity in ambiguous stimulus fields, reducing conflict, or preserving self-worth. On the contrary, the reactance process is one of remaining open, of preserving one's freedom of movement, in the face of pressures to firm up or narrow the direction of movement.

1. Freedom: The Starting Point

The theory can best be introduced by considering single behavior realms in which freedom for the individual is conceivable. For instance, while shopping a person expects the freedom to choose from a vast array of breakfast cereals or brands of coffee. In more important matters, the person has a choice among numerous kinds of automobiles. In commerce with others, the person can potentially choose between being overtly friendly vs. taking a more distanced stance toward others. Even in such affective realms as emotions there is the potential—as in a setting that would normally set off sadness or grief—to react appropriately or to begin laughing.

In short, the theory addresses the freedoms associated with segments of human existence. These usually entail concrete behavioral scenarios in which a decision is eventually possible, or in which the person senses that several different solutions or directions can be chosen or tried out. In the context of the theory one does not speak of 'the free person' vs. 'the unfree person.' Freedom here is defined by the concrete endeavor of interest, such as finding oneself in a coffee bar and having to choose between espresso and normal coffee.

By no means does the theory assure that people automatically feel free across all such settings. On the contrary, Brehm (1966, 1968, 1993) postulates that, in each behavioral setting, the person will have a strong (or less strong, ranging down to zero) sense of freedom as a function of one of two factors. (a) Prior experience in the setting will be a guide as to expectations of freedom, as in, for instance, the case of a child who has always been entrusted with the freedom to select any of the breakfast cereals in the grocery store. This case my be contrasted with the child who has always been instructed to pick only the one brand of cereal that contains no sugar, preservatives, or coloring. (b) The other factor underlying the person's sense of being free, or less free, in a particular setting has to do with culturally-defined expectations. For example, pedestrians may (or may not) expect the freedom to cross against the red light, depending on local customs.

In practice, this all-important starting point for the theory—the person's expecting to be free or not—has been handled in an experimental manner. In some studies research subjects are told that they can expect to select freely from among several available consumer alternatives. For other participants, the expectation is created that they will receive one or more of the consumer items on a random basis, or at the whim of the person who controls the situation. People in this latter case thus find themselves expecting to gain something, such as a CD, pen, or article of clothing, but there is no person-freedom associated with their gaining that end.

With this illustration the reader can see that the theory is highly specific in its treatment of freedom, choice, and the free person. Setting-by-setting, individuals expect to be able to move about and choose freely, or not. One child expects to choose freely among cereals; another does not. Students in one culture might expect to choose which textbook to read for an exam; students in another culture may have no such expectation.

2. *Reactance Arousal*

Given that a person expects freedom of movement in a setting, infringements on that freedom will instigate a motivational force of reactance. The state of reactance, going into gear as a function of the strength of threat to freedom and the importance of the freedom, brings the person to resist and counter the shrinking of the pertinent freedom. This is manifested in a direct behavioral fashion and also in terms of affective reactions.

Take the instance of the child who expects freedom of choice among breakfast cereals. If the child is standing before the array of breakfast products, pondering what to choose, reactance will come about by another person's picking a box of cornflakes from the top shelf and dropping it into the child's grocery basket. This unequivocal social influence will set up reactance in the child, resulting in the tendency to remove the cornflakes from the basket and to select a different variety. Further, the motivational state will show up in terms of affective changes; as an outcome of the social interference with freedom, the child will become relatively unattracted to the cornflakes and increasingly attracted to other salient, interesting brands.

3. *Frustration*

Reactance is not about attaining the goal that is top-ranked in the person's need or preference hierarchy. If the child had initially ranked Swiss muesli as No. 1, and cornflakes as No. 4, then the adult's placing the cornflakes in the child's grocery basket would have been directly frustrating. Perhaps obviously, the child would have reacted negatively to the frustration of a simple need to select the top-ranked item. However, the deciding quality of reactance theory is in its addressing freedom *per se*, and not the issue of the person's attaining the most-preferred goal object. The implications of this statement are far reaching; as long as the child's freedom to choose any of the considered alternatives is interfered with, reactance will be engendered. Quite aside from whether the muesli is No. 1 and the cornflakes No. 4, or vice versa, the adult's constraining the choice will produce reactance and the consequent behavioral/affective effects.

The point regarding frustration is similar when reactance is brought about in areas such as freedom of values, attitudes, or emotions. There is a sense in which it is frustrating to be pressured to take a certain opinion position, particularly one at variance with one's usual opinions. When a member of the American Civil Liberties Union is confronted with others who authoritatively argue in favor of restricting certain rights, the felt pressure to succumb to such pressure toward change is in a sense frustrating. But for reactance theory, the actual direction of such social pressure is not a terribly critical factor. If the member of the ACLU feels free to hold any of several positions along the human rights continuum, then this person will also experience reactance when a forceful communicator insists that the only correct opinion is a totally liberal one. Reactance, in the case of pressure

to change an opinion, value, or emotion, is manifested in terms of 'boomerang' effects, i.e. switching the opinion to a direction opposing the communication.

4. Areas of Application

Given the breadth of the theory, it seems reasonable that it should be applied to such diverse realms as choices among products, among people (e.g., choosing a partner), and to numerous social realms such as whether or not to help another person and whether or not to abide by another's advice or efforts to influence. The research evidence addressing these topics is largely experimental, and follows the formula given by the theory. First the investigator must know whether or not the person to be studied expects freedom of choice in the given context, and second, the strength of the threat to that freedom is varied. Within these general methodological guidelines, research participants have been shown to resist persuasive attempts, refuse to help, and frequently to upgrade their attraction to potential choice objects or people whose availability has been threatened. Reviews of these many effects can be found in Brehm (1966, 1968, 1993), Brehm and Brehm (1981), Wicklund (1974, 1997), and Wortman and Brehm (1975).

5. The Longer-range Dynamics

These dynamics can best be depicted in three phases, with the assistance of an expanded theory by Wortman and Brehm (1975). The first two phases, as described above, are (a) the instigation of an expectation of acting freely in a domain, and (b) infringements on that freedom, producing reactance. The issue raised by Wortman and Brehm—implying a third phase—has to do with the effect of repeated restrictions on freedom. A person may be accustomed to choosing freely between espresso and normal coffee in a certain coffee bar. When the management announces one day that normal coffee is not available, the reactance-produced response is one of increased attraction to normal coffee and a corresponding downgrading of espresso. But what transpires upon repeated, continued instances of this restriction? Will the person continue to express reactance on the second, third, fourth, and fifth days of the freedom restriction?

In alternative theoretical language, such repeated restrictions are called helplessness inducing (Seligman 1975), and in the Wortman and Brehm model, the idea is that repeated violation of the freedom will eventually bring the person to no longer expect freedom of action within that domain, as in the particular coffee shop. The expectation of free choice is altered into the expectation of espresso, and nothing more. This adaptation to the repeated restriction is thus the third phase of the longer-term reactance process, and as such, describes the ebb and flow of reactance in numerous situations over the course of a child's development, as well as the ebb and flow of reactance in diverse areas as a result of cultural change.

6. Individual Differences

Brehm (1993) differentiates sharply between the chronic need to expand areas of freedom and control, and the defense of a specific area of freedom or control. The former, normally associated with such concepts as control need or competence striving, addresses a human who is, across time and place, chronically interested in enlarging the scope of freedom, thus in maximizing personal control in a general sense. Such a chronic personal striving can be operationalized in terms of certain individual differences, such as Rotter's Internal/external control scale and, to be sure, researchers have introduced these very individual differences into a variety of reactance paradigms. The research suggests that those with a high chronic need for control appear to evidence stronger responses in reactance-creating situations (see Brehm and Brehm 1981).

On the other hand, there is little theoretical reason to expect that general human tendencies will be pertinent to the reactance process as described in detail in the numerous experiments. This is because the question of whether a person will evidence the counter-behaviors, the boomerang effect attitude changes, and other freedom-restoring maneuvers, is a highly concrete issue. The parameters of each behavioral scenario, combined with the individual's expectation of freedom within those confines, determines the momentary predilection to protect and regain freedom.

7. Future Directions

A salient direction to take in future development of the theory involves the social circumstances which give rise to expectations of freedom. While the original theory refers to such expectations as stemming from (a) direct, personal experience and (b) societal conventions regarding specific classes of actions, there is a further possibility: a person who is raised in a community in which a variety of points of view are acknowledged will stand to gain a sense of freedom to express those many points of view. This should apply no matter whether they are expressed in discussions of opinions and values, or even in solutions to problems. In contrast, a person coming from an environment in which there is 'one correct opinion,' or solution, will be less likely to incorporate the idea that multiple opinions or solutions are possible or desirable (Wicklund 1999). Accordingly, we may be able to speak of a person's more general sense of freedom of opinions and values, as long as the person's socializ-

ation experiences are in a setting that fosters such multiple values systems, or even multiple personality traits. The direction for corresponding research lies both in developmental psychology and in the detailed descriptions of cultural settings allowed by cultural anthropology.

See also: Control Behavior: Psychological Perspectives; Control: Social; Freedom/Liberty: Impact on the Social Sciences

Bibliography

Brehm J W 1966 *A Theory of Psychological Reactance*. Academic Press, New York

Brehm J W 1968 Attitude change from threat to attitudinal freedom. In: Greenwald A G, Brock T C, Ostrom T M (eds.) *Psychological Foundations of Attitudes*. Academic Press, New York

Brehm J W 1993 Control, its loss, and psychological reactance. In: Weary G, Gleicher F, Marsh K L (eds.) *Control Motivation and Social Cognition*. Springer, New York, pp. 3–30

Brehm S S, Brehm J W 1981 *Psychological Reactance: A Theory of Freedom and Control*. Academic Press, New York

Seligman M E P 1975 *Helplessness: On Depression, Development, and Death*. Freeman, San Francisco

Wicklund R A 1974 *Freedom and Reactance*. Erlbaum, Hillsdale, NJ

Wicklund R A 1997 The theory of psychological reactance: a reliably revolutionary perspective. *Contemporary Psychology* **42**: 679–81

Wicklund R A 1999 Multiple perspectives in person perception. *Theory and Psychology* **9**: 667–78

Wortman C B, Brehm J W 1975 Responses to uncontrollable outcomes: An integration of reactance theory and the learned helplessness model. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 8, pp. 277–336

R. A. Wicklund

Readability, Applied Psychology of

Almost since the beginning of research into the process of text comprehension, researchers and educators have attempted to find accurate, meaningful, and simple methods to assess readability. Early researchers such as Quantz, Cattell, Huey, and Dearborn—most of their research was performed around the turn of the twentieth century—focused on differences among readers' abilities, offering what might now be considered 'process models,' albeit simple ones. The factors they studied, however, were more perceptual in nature, such as the effects of saccadic eye movements and eye–voice spans in reading aloud. Despite some important insights, their work waned in influence during the period from 1910 until 1970, or so. Not coincidentally, this decline in the interest of process models mirrored the rise in influence of behaviorism in American psychology. Behaviorists tended to focus on observable, easily measured factors. Reading researchers directed their attention to properties of the texts, rather than how a text might induce changes in the processing of the text by the reader.

During the middle portion of the twentieth century, the first readability formulas were developed. Consistent with the underlying philosophy of behaviorism, these formulas considered readability to be a property of the stimulus—that is, the text itself. Most of the commonly used readability formulas, such as the Gunning Fog index and the Flesch Readability scales, were developed during this time.

The so-called cognitive revolution of the 1960s renewed interest in natural language. In the last 25 years, especially, the psychological process of language comprehension has received considerable attention. A number of wide-reaching theories of text comprehension have been developed (Gernsbacher 1990, Just and Carpenter 1980, Kintsch 1988, Kintsch and van Dijk 1978). These models do not assume the text to be a static entity; instead, they emphasize the process of reading, or more specifically, the interaction between the text and the reader. Ironically, contemporary researchers have essentially returned to the same philosophical positions that drove researchers such as Huey and Dearborn a century before.

1. Early Readability Research

The challenge facing educational psychologists in the 1930s and 1940s was rather pragmatic, and is still one of the goals of contemporary researchers and educators: to evaluate and compare readability of texts used in a variety of school settings. In theory, this would allow teachers to select age- (or ability-) appropriate materials for use in the classroom. In addition, educators were looking for a way to assess the readability of various books students might read on their own.

Understandably, these formulas tended to follow from behaviorally oriented research: identify the objective (although generally superficial), easily measured properties in the stimuli—the texts. Early work tended to focus primarily on such simple ideas as vocabulary difficulty, word frequency, and sentence length (e.g., Gray and Leary 1935). It is important to note that the question of *why* these variables were related to readability was considered largely irrelevant. An understanding that sentence length was important because it provided an estimate of the demands placed on working memory was decades away.

Within a few years, readability formulas were developed that could be applied virtually to any text (Dale and Chall 1948a, 1948b, Flesch 1948, Klare 1963, 1974/5, 1984, Vogel and Washburne 1928).

Though each of these differed in some respect, in truth they were remarkably similar. Most of these scales used some combination of word length (presumably related to word frequency or word difficulty), number of sentences per paragraph (a crude measure of complexity), sentence length, and/or number of syllables per word. They soon became widely adopted in academic, educational, and other applied settings. Standardized testing procedures also made wide use of these formulas. Some governmental agencies even specified readability levels for things such as insurance policies; in order to be 'approved,' for example, an insurance policy must have a readability grade level of 9 or below.

Despite the sophistication of contemporary reading research, most of the readability formulas in wide use today continue to use the same key variables. For example, recent versions of Microsoft Word provide an assessment of readability using the Flesch Reading Ease scale, as well as a normed score using the Flesch-Kincaid Grade Level scale (which scales readability relative to US public school grade levels). Cassady and Greene's program Grammarian also provides an assessment of readability using the Gunning Fog index. (The Gunning Fog index uses the number of 'long words': those of three or more syllables.) These formulas are somewhat successful, at least to a point, and provide a reference from which texts can be compared.

Perhaps such widespread use of these simple formulas is not warranted. After all, these formulas analyze only a few variables, and cases where the formulas fail are abundant. However, the limitations of the formulas (the fact that they analyzed only a few components) were also their most attractive feature—they were easy to use. Thus, these criticisms, even when recognized by the researchers, have not attenuated the popularity of readability formulas.

Do revisions intended to improve the readability of a text actually produce more easily understood texts? One of the most thorough investigations in rewriting and readability was performed by Britton and Gulgoz (1991). They used a text of roughly a thousand words long, detailing the success and failure of the air bombing 'Operation Rolling Thunder' offensive in the Vietnam war (Greene 1985). They compared the original text (which had a grade level equivalent of about 13.5 years, as measured by the Flesch-Kincaid Index) to one of three revisions: a readability revision, which used shorter sentences and more frequently used words, improving the readability level to 12 years; a principle revision, which modified the text according to the comprehension model of Kintsch and van Dijk (1978) and had a readability level the same as the original text (13.5 years); and a heuristic revision, performed by an expert in text revision (Britton himself), designed to make the text as understandable as possible. The reading level of the latter version was lowered to grade 12. The only revision which actually improved comprehension was this final revision. Lowering the readability score itself produced no significant improvement—simply using shorter sentences and more frequent words was not sufficient to make a text more understandable.

Improving readability has one other benefit—texts written at a level consistent with a reader's ability are monitored more accurately while reading. Simply put, comprehension monitoring refers to the accuracy with which individuals can assess their comprehension. After reading a passage, individuals are asked to predict later performance on a test. The correlation between predicted and actual performance provides a measure of comprehension monitoring accuracy.

Weaver and Bryant (1995) examined the effect of readability on comprehension monitoring. They had college students read texts that were either easy (reading level below grade 8), standard (reading level around grade 12), or difficult (reading level around grade 16). Comprehension performance was related to readability: recognition performance increased as reading material became easier (more readable). Comprehension monitoring, in contrast, showed an inverted U-shaped function. The greatest metamemory accuracy occurred not for the easiest materials, but for those materials best matching the readers' ability (standard readability, approximately 12th grade).

2. *Summary*

The theoretical case against readability formulas has been argued many times (e.g., from the present perspective by Kintsch and Vipond 1979). Their merits are clear, and center on their ease of use. But the drawbacks are equally clear. At best these formulas provide an oversimplification, too often a serious distortion. Nevertheless, they continue to be used, and will be as long as comprehension research does not provide viable alternatives.

Great progress into the reading process has been made in the past 30 years. But why have these advances not led to the development of viable alternatives to readability formulas?

The answer to this question has three parts. First of all, most of the research discussed above has dealt with the construction of internal representations, often called textbases. A great deal is now known about how to construct comprehensible, recallable, and summarizable texts for given populations. However, while this knowledge is quite adequate for laboratory research on discourse processing, simple, easy-to-use procedures for practical applications have not yet been developed.

The second part of the answer involves a fundamental change in the nature of readability. Readability has typically been considered as a property of the text alone. Clearly, true readability is more

complicated than that and is probably best considered as an interaction between the text itself and the readers' *ability* and *knowledge*. Unfortunately, a simple readability formula incorporating readers' abilities has yet to be developed.

The final part of the answer is purely pragmatic. When using a text in an experiment—often a relatively brief text—it does not matter whether the procedures for text analysis are time-consuming, require expertise, and so on. For the applied settings in which most readability formulas are used, though, ease of use matters a great deal. What is needed is a large-scale effort directed toward the development of practical methods of text analysis that allow non-experts to do with relatively little effort what only experts can do today with considerable effort. While such an enterprise would be nontrivial, it is certainly feasible. For the moment, however, readability measures have little connection to the cognitive science research on text comprehension.

In light of this fact, Bruce et al. (1981; see also Weaver and Kintsch 1991) have proposed that readability formulas be used only when the following criteria are met:

(a) The material will be freely read, and reading time will not be determined by 'external factors' such as experimental control or (in their example) captions that are presented along with television programs (for the hearing impaired).

(b) The text is 'honestly written': that is, not directly written in such a way to satisfy the criteria of readability formulas.

(c) The higher-level features of a text, such as organization and intention, correlate with the sentence and word features. As current formulas make little use of the top-level information, such material is otherwise ignored.

(d) The purpose of the reading is similar to that of the readers in the validation studies. If the two groups are reading with different strategies, then the formula may not be valid across the situations.

(e) The formulas are applied across a large number of texts, and individual texts are not singled out. Bruce et al. (1981) emphasize that readability scores are composite averages and may not apply to specific texts.

(f) The population of readers is the same in the 'real-world' setting as it was in the validation studies. If a text is to be used as a remedial text, it should not be validated on a group of average readers.

See also: Comprehension, Cognitive Psychology of; Eye Movements in Reading; Inferences in Discourse, Psychology of; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Learning from Text; Literary Texts: Comprehension and Memory; Macrostructure in Discourse Comprehension, Psychology of; Memory for Text; Narrative Comprehension, Psychology of; Reading, Neural Basis of; Reading Skills; Sentence Comprehension, Psychology of; Speed Reading; Text Comprehension: Models in Psychology; Textbooks

Bibliography

Britton B K, Gulgoz S 1991 Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology* **83**: 329–45

Bruce B, Rubin A, Starr K 1981 *Why Readability Formulas Fail*. Center for the Study of Reading, Champaign, IL

Dale E, Chall J S 1948a A formula for predicting readability. *Educational Research Bulletin* **27**(January): 11–12

Dale E, Chall J S 1948b A formula for predicting readability. *Educational Research Bulletin* **27**(February): 37–54

Flesch R F 1948 A new readability yardstick. *Journal of Applied Psychology* **32**: 221–33

Gernsbacher M A 1990 *Language Comprehension as Structure Building*. Erlbaum, Hillsdale, NJ

Gray W S, Leary B E 1935 *What Makes a Book Readable*? University of Chicago Press, Chicago

Greene M P 1985 *US Air Power: Key to Deterrence*. Maxwell Air Force Base

Just M A, Carpenter P A 1980 A theory of reading: From eye fixations to comprehension. *Psychological Review* **87**: 329–54

Kintsch W 1988 The role of knowledge in discourse comprehension: A constructive-integration model. *Psychological Review* **95**: 163–82

Kintsch W, van Dijk T A 1978 Toward a model of text comprehension and production. *Psychological Review* **85**: 363–94

Kintsch W, Vipond D 1979 Reading comprehension and readability in educational practice and psychological theory. In: Nilsson L G (ed.) *Perspectives on Memory Research*. L. Erlbaum Associates, Hillsdale, NJ, pp. 329–65

Klare G R 1963 *The Measurement of Readability*. Iowa State University Press, Ames, IA

Klare G R 1974/5 Assessing readability. *Reading Research Quarterly* **10**: 62–102

Klare G R 1976 A second look at the validity of readability formulas. *Journal of Reading Behavior* **8**: 129–52

Klare G R 1984 Readability. In: Pearson P D, Barr R, Kamil M L, Mosenthal P (eds.) *Handbook of Reading Research. Vol. I*. Longman, New York, pp. 681–744

Vogel M, Washburne C 1928 An objective method of determining grade placement of children's reading material. *Elementary School Journal* **28**

Weaver C A III, Bryant D S 1995 Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition* **23**: 12–22

Weaver C A, III, Kintsch W 1991 Expository text. In: Barr R, Kamil M, Mosenthal P, Pearson P D (eds.) *The Handbook of Reading Research*. Longman, White Plains, NY, Vol. 2, pp. 230–45

C. A. Weaver, III and A. Renken

Reading, Neural Basis of

Psychologists and brain researchers who investigate reading usually focus on the basic cognitive processes

and brain regions that are necessary to recognize and pronounce a printed word. More complex aspects of reading, such as the ability to understand the grammatical roles played by each word in a written or spoken sentence, typically are assumed to rely upon general language functions that are not specific to reading (see *Syntactic Aspects of Language, Neural Basis of*). This chapter will focus on three components of skilled reading, and the brain regions that support them: a) the representation of visual word forms (orthography), b) the use of orthographic information to access information about word sounds (phonology), and c) the use of orthographic information to access information about meaning (semantics).

Claims about the neural bases of component processes associated with reading are based upon three primary methods. Studies of subjects with brain damage (neuropsychological studies) are used to understand how different aspects of reading performance are impaired and preserved following brain injury or degeneration (see *Dyslexia (Acquired) and Agraphia*). Studies of brain blood-flow changes (functional neuroimaging studies) use either PET (Positron emission tomography) or fMRI (functional magnetic resonance imaging) to localize and understand the functional contributions of brain regions associated with reading. ERP (see *Event-related/Evoked Potentials*) and MEG (see *Magnetoencephalography*) studies measure brain electromagnetic changes, and provide information about the temporal sequence of brain processing during reading, as well as some information about the location of brain regions associated with reading.

1. An Overview of the Brain Regions Involved in Reading

An overview of the brain regions that support reading can be obtained by using neuroimaging to compare the pattern of brain activation associated with reading aloud a single word, versus merely looking at a small cross or dot. Some of the areas identified in this task comparison support relatively general sensory and motor functions. For instance, the primary motor cortex, which is located in the precentral gyrus (Fig. 1), is usually active during any task that incorporates movement. The identification of the motor cortex in the reading versus fixation comparison reflects the fact that a task that involves mouth movements is being compared to a task that does not. One approach for identifying regions that are more specifically involved in reading is to compare reading aloud to a control task that has similar sensory and motor components, but which minimizes orthographic, phonological, or semantic processing. For example, reading aloud has been compared to a task in which subjects say a simple utterance in response to a random string of letters. This type of contrast subtracts out basic visual, auditory, and motor areas, and reveals a more restricted set of regions that support higher-order aspects of reading (Fiez and Petersen 1998).

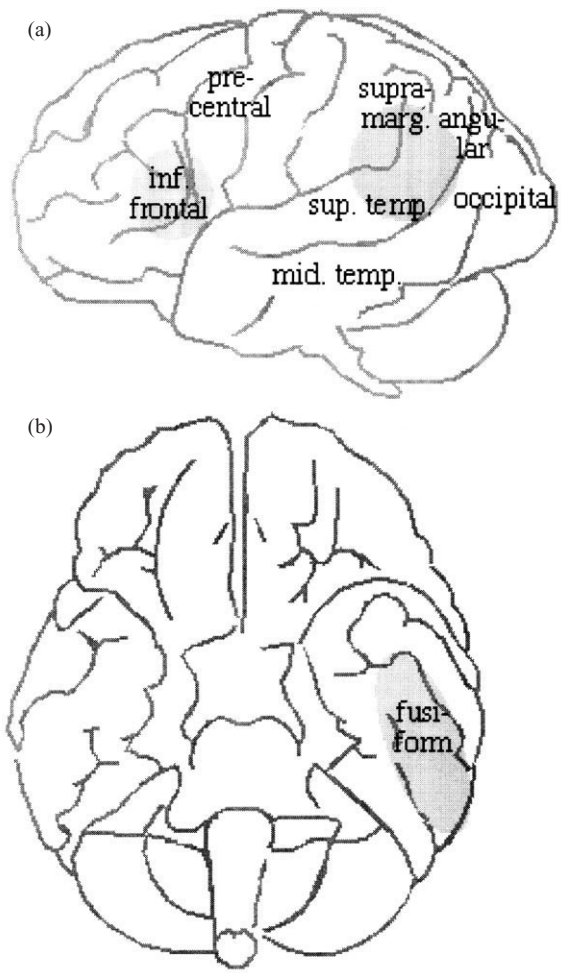

Figure 1
Brain areas associated with reading. Brain regions involved in higher-level aspects of reading can be identified through comparisons of reading to simpler tasks that contain similar sensory and motor components. Three broad areas identified through such task comparisons have attracted a significant amount of attention. A left inferior frontal area (1a) is centered on the inferior frontal gyrus. A left temporoparietal area (1a) is centered on the junction of the posterior superior temporal cortex and the supramarginal and angular gyri of the inferior parietal lobe. A left basal temporal area is centered within the fusiform gyrus, which lies between temporal and occipital cortex. This area is best viewed from the underside of the brain (Fig. 1b)

The exact number and location of the brain regions that support higher-order aspects of reading is a matter

of active investigation, but three broad areas have attracted the most attention (Fiez and Petersen 1998; Price 1998). As illustrated in Fig. 1, these areas are centered in the left inferior frontal cortex, the left temporoparietal cortex, and the left basal temporal cortex (primarily the fusiform gyrus). Most of the remainder of this article will focus on the potential relationships between these brain areas and the basic components of skilled reading outlined in the introduction. It is important to note, as will become clearer in the discussion of the basal temporal area, that a more precise mapping of the brain regions that contribute to reading is still needed. It is likely that each of the broad areas focused on in this article contains multiple functionally distinct subregions. Furthermore, it is likely that other areas that are not discussed, such as the left middle temporal gyrus (Fig. 1), make some contribution to reading.

2. *The Representation of Orthography*

The organization of the visual system has been a topic of longstanding interest in the neuroscience and cognitive neuroscience research communities. One question that is highly relevant for studies of reading is whether different regions of the visual object recognition system (located in the ventral occipital and temporal cortex) are specialized for processing different types of stimuli, such as faces or text. In the domain of reading, the search for a region that selectively processes text has generally involved two types of comparisons. The first has been comparisons of words and nameable pictures. This allows regions that are specifically involved in orthographic processing to be teased apart from regions that are more generally involved in the naming of meaningful stimuli. The second approach has been comparisons between different types of orthographic stimuli. Most often, words and pseudowords (pronounceable nonwords, such as 'floop') have been compared to nonword stimuli that are visually similar to words (e.g., stimuli composed of artificial letters, or 'false fonts,' or random consonant strings). This allows regions that are specifically involved in the representation of words to be dissociated from regions that are involved in lower-level processes, such as letter recognition and the detection of line segments.

Several regions that may mediate orthographically specific processing have been proposed. The most promising 'word form area' is in the left middle fusiform gyrus, near the border between the temporal and occipital cortex. Activation in the fusiform gyrus is found reliably when subjects read words aloud (Fig. 1), and one reason to suspect that it may be involved in orthographic processing is the fact that it lies near areas associated with simple visual processing and visual object recognition. More direct evidence comes from studies that have found responses in the fusiform to words and pseudowords, but not visually similar nonword stimuli, such as random letter strings (Fiez and Petersen 1998). Evidence that this area is influenced by the degree to which an orthographic stimulus functions as a whole word (lexical) unit comes from studies in Japanese readers. The Japanese writing system contains two types of characters: Kana characters represent syllables, while Kanji characters represent entire words. Significant activation in the fusiform is found when subjects read words aloud written with Kanji versus Kana characters, and when subjects are given a task that requires them to access the orthographic form of a Kanji character, such as writing the Kanji equivalent of a word written in Kana (Nakamura et al. 2000, Sakurai et al. 2000).

Data from neuropsychology also point toward the important role of occipitotemporal areas, such as the fusiform gyrus, in orthographic processing. Subjects with brain damage in this area can exhibit the syndrome of pure alexia (Patterson and Lambon-Ralph 1999). In pure alexia, the ability rapidly to access or represent the orthographic form of an entire word, or the ability to use orthographic information to access phonological and semantic information, is lost. As a result, subjects use a laborious letter-by-letter strategy to identify and read aloud a word. Subjects with pure alexia generally are able to perform simple visual perceptual tasks, recognize and name visual objects, and write words normally. Japanese subjects can also exhibit alexia following damage to the fusiform, but long-lasting impairments are typically limited to the reading and writing of Kanji characters (Iwata 1984).

Results from ERP and MEG studies indicate that there are differences in the electromagnetic current evoked by words and wordlike stimuli, as compared to nonword stimuli such as letter strings (Cohen et al. 2000; Tarkiainen et al. 2000). The temporal information gained from these studies indicates that orthographically specific processing at the word level begins 100–200 milliseconds (ms) after the onset of a stimulus. Attempts to localize the source of the waveform differences have generally implicated the left basal temporal cortex. More specific localization has come from a study in which recordings were done using an electrode placed in the fusiform gyrus of patients prior to brain surgery. This study found neurons that selectively responded to orthographic stimuli (Nobre et al. 1994).

While there is a significant amount of accumulated evidence across multiple methodologies that points towards a role for the fusiform gyrus in orthographic processing, the precise function of this area remains a point of debate. One issue is whether domain-specific regions—'face areas' or 'word form areas'—exist in the visual system. Since reading is a skill acquired very recently on the evolutionary time-scale, it is unlikely that any brain region would be innately specified to represent word forms. Rather, any specialization is

most likely to occur as a result of extensive experience with a relatively unique class of visual stimuli. An alternative point of view is that the analysis of orthography places heavy demands on particular types of visual analysis, such as the ability to resolve input with high-frequency spatial information, or to represent a set of distinct objects that are perceptually very similar. Portions of the fusiform may be particularly suited for these types of analysis (either innately or through experience), and thus may be used to a greater extent for processing orthographic versus non-orthographic stimuli.

A second issue is the level of information that may be represented in the fusiform. Recent evidence indicates that letter-by-letter reading is typically accompanied by visual-perceptual difficulties below the level of the whole word. Deficits in lower-level processing could impair the access to or generation of higher-level orthographic representations (Behrmann et al. 1998). Conversely, imaging evidence reviewed more extensively below indicates that portions of the fusiform are active during linguistic tasks that are not thought to require access to orthographic information. Activation associated with orthographic processing may thus reflect more abstract lexical (whole word) or semantic information that can be accessed through either printed words or speech.

3. *The Transformation Between Orthography and Phonology*

In order to read a word aloud, it is essential that information about the printed form of the word is transformed into information about the corresponding sound, or pronunciation, of the word. Two general brain regions have been identified which appear to play a major role in phonological (word sound) representation and processing. One region is located in the left inferior frontal lobe, and the other is in the temporoparietal cortex.

Much of the neuropsychological evidence that these regions are important for reading stems from studies of patients with damage to either or both of these regions. These patients often exhibit the syndrome of phonological dyslexia, which is characterized by great difficulty reading aloud pseudowords (Fiez and Petersen 1998; Patterson and Lambon-Ralph 1999). The ability to read aloud pseudowords is considered to be a marker of phonological processing, because the pronunciation of pseudowords cannot be specified by stored lexical (whole word) or semantic knowledge. Instead, pronunciation is thought to rely upon some type of sublexical procedure for deriving a phonological representation. Subjects with phonological dyslexia can read real words relatively well. This is thought to reflect their ability to take advantage of stored lexical and semantic information using brain regions that are not damaged. Interestingly, when direct electrical stimulation is applied to the temporoparietal region during a reading task, subjects appear to exhibit a temporary phonological dyslexia: their ability to pronounce nonwords is temporarily disrupted, but their reading of real words is intact (Simos et al. 2000; see also *Electrical Stimulation of the Brain*).

Conclusions based upon neuropsychological data are supported by converging evidence from neuroimaging. Both the left frontal and temporoparietal regions are reliably active during normal word reading (Fig. 1), and during a wide variety of tasks that encourage phonological processing (Demonet et al. 1996). ERP and MEG evidence indicates that the activation of these areas occurs at least 300 ms after the onset of a stimulus, which is after the temporal differences observed in the basal temporal cortex (Simos et al. 2000). The importance of these regions in reading is further demonstrated by comparisons of normal adults to adults who experienced difficulties learning to read as children (developmental dyslexics). Many investigators believe that subjects with developmental dyslexia have a core deficit in phonological processing that persists into adulthood. When subjects with developmental dyslexia read words aloud, neuroimaging and MEG studies have revealed abnormally low levels of activity in the temporoparietal area (Shaywitz et al. 1998, Rumsey et al. 1999, Simos et al. 2000). The relative reductions in temporoparietal activation can be paralleled by above-normal levels of activity in the left frontal region. The left frontal increases may reflect use of a compensatory strategy for reading that normally plays a secondary role in pronunciation for skilled readers (except under more challenging conditions, such as when a stimulus is a nonword).

Several broad schemes that might distinguish the contributions of the frontal and temporoparietal regions to reading and phonological processing exist. One potential distinction is whether the phonological information and processes supported by a particular brain region are related to the acoustics of speech sounds (input phonology), or the articulation of speech sounds (output phonology). The proximity of the temporoparietal region to basic auditory areas, coupled with the fact that lesions to this area produce deficits in speech comprehension, indicates that it may support input phonology. Conversely, the proximity of the left frontal region to general motor-related areas, coupled with the fact that lesions to this area produce deficits in speech production, indicates that it may support output phonology.

A second potential distinction comes from broad differences between the roles of the frontal and posterior brain regions in cognition. Frontal regions are recruited most often by tasks that require conscious and effortful processing, or controlled processing. When a task is well learned and relatively automatic, activation in the frontal cortex is generally

minimal. The frontal cortex may thus represent an area used for controlled phonological processing, whereas the temporoparietal region may represent an area used for automatic phonological processing. This could explain why subjects with developmental dyslexia, who do not develop the normal level of reading skills and automaticity, show less activation in the temporoparietal area and more in more frontal areas. It is also possible that different subregions within the inferior frontal cortex are responsible for different aspects of speech production (e.g., output phonology, grammatical analysis, etc.).

A third potential distinction comes from theoretical models of reading, which have outlined different types of mechanisms that could support orthographic to phonological transformation. Two broad classes of models have been proposed, dual-route (Coltheart et al. 1993) and connectionist models of reading (Plaut et al. 1996). In both types of models, access to the pronunciation of a word is mediated by two different 'routes,' though the routes function in different ways. In the dual-route models, a lexical route functions to associate orthographic information with stored representational information about the pronunciation of the whole word. A sublexical route functions to break a word or pseudoword into smaller parts, such as individual letters or letter clusters (graphemes), that can be associated with corresponding sounds (phonemes); the sounds can then be put together to yield a pronunciation for the entire item.

In connectionist models, direct mappings between orthography and phonology are used to capture statistical regularities between orthography and phonology at both the lexical (or whole word) and the sublexical (e.g., the 'b' in 'back' and 'bear' has the same sound) levels. The pronunciation of a written stimulus is represented as a distributed pattern of activity across a set of interconnected phonological units. Indirect mappings between orthography and phonology via semantics can also be used to support the pronunciation of a word. That is, you could represent the meaning of the word 'dog,' and also represent the fact that this meaning is associated with the pronunciation/dog/. As in the direct mappings between orthography and phonology, these mappings are represented as a distributed pattern of activity, but unlike the direct mappings the mappings should be relatively insensitive to sublexical information (since there is little relationship between parts of words and word meanings, at least for simple 'mono-morphomic' words like 'dog' and 'table').

Simple assumptions have been used to suggest potential mappings between theoretical components in models of reading and specific brain regions (Fiez and Petersen 1998, Price 1998). For instance, it has been suggested that the left frontal area contributes to a sublexical phonological process. This mapping is based upon the fact that in models of reading the pronunciation of pseudowords is primarily based upon output from a sublexical phonological procedure, and in the brain greater activation is seen in the left frontal area when subjects read pseudowords as compared to words. Even this potential correspondence, however, falls short of identifying whether the left frontal region functions according to dual-route or connectionist principles (or some alternative process), and it fails to acknowledge puzzling differences between words and pseudowords in other brain regions. In order to reach more complete levels of description, it will be essential to develop more rigorous methods for relating computational modeling and neuroimaging results.

4. The Transformation Between Orthography and Lexical Semantics

In addition to lexical procedures that facilitate direct associations between an orthographic and a phonological form of a word, it is also possible that other types of stored representational information can be used to support pronunciation. This could include lexical information about the word itself, the meaning of the word, or the concept denoted by the word. Once such lexical-semantic information has been accessed, it could be used to access information about the pronunciation of the word or concept, as suggested in connectionist models of reading. Importantly, access to most lexical-semantic information is expected to be modality and task independent. Either a written word or a picture, for instance, can be used to represent a word name or a semantic concept, and individuals can produce a spoken word in response to a written word, picture, or their own internal thoughts.

Of the three general areas that are commonly active during word reading (Fig. 1), the basal temporal region is the best candidate for an area that contributes to lexical-semantic processing. This is the same general region that has been implicated in orthographic representation, and the potential overlap between orthographic and lexical-semantic processing is discussed at the end of this section. Assertions that the basal temporal region contributes to lexical-semantic processing are based upon evidence that the region responds to a variety of stimuli and task conditions that emphasize the retrieval of word names or concepts. For instance, neuroimaging studies have found overlapping activation during picture naming and word naming tasks (Moore and Price 1999). Outside the domain of reading, basal temporal activation is commonly associated with tasks that emphasize semantic processing or word retrieval, such as deciding whether a pair of words belongs to the same semantic category (Price 1998; see also *Semantic Knowledge: Neural Basis of*).

Neuropsychological studies also point towards the role of the inferior temporal area in lexical-semantic processing. With respect to reading, the most critical

information comes from studies of patients with acquired surface dyslexia. In the purest forms of surface dyslexia, subjects tend to mispronounce inconsistent words, which are words with unusual spelling-to-sound correspondences. For instance, the word 'pint' is not pronounced like 'mint,' 'lint,' or 'hint.' Theoretical models of reading have postulated that the pronunciation of inconsistent words, especially those that occur infrequently in print, is heavily dependent upon access to stored lexical-semantic information. Most reported subjects with surface dyslexia have large regions of damage or degeneration affecting the left middle and inferior temporal lobe, and the deficits of these subjects typically extend beyond the domain of reading (Patterson and Lambon-Ralph 1999). For instance, they often have difficulty accurately naming pictures of objects. Surface dyslexics are generally able to read aloud most words with consistent spelling-to-sound correspondences, and pseudowords. This preserved ability may reflect the use of sublexical orthographic to phonological procedures that are supported by intact brain regions, such as the left frontal area.

An issue that is not well understood is whether the reading-related area discussed in Sect. 2 is the same as the lexical-semantic area discussed in this section. One reason to believe that they may be the same is the fact that in many neuroimaging studies of reading, only a single region of activation is identified in the fusiform gyrus. If they do overlap, future research will need to reconcile the orthographic and lexical-semantic points of view to provide a unified interpretation of the role of the inferior temporal area in reading. However, there is some evidence that there may be more than one functional region within the basal temporal area. A neuroimaging study of word versus picture naming, for example, identified multiple regions within the fusiform that had different patterns of activation (Moore and Price 1999). This is consistent with converging data from animal and human studies of visual object recognition. Such studies indicate that modality and domain specific visual processing, such as that involved in orthographic representation of lexical information, may be located in the posterior fusiform. In contrast, the anterior portions of the fusiform may be associated with increasingly conceptual and modality-independent processing. This could include the conceptual representation of objects and ideas that are denoted by pictures, and by written and spoken words (Ricci et al. 1999).

5. *The 'Tuning' of Brain Regions that Support Reading*

Neuropsychological and neuroimaging studies of Chinese, Japanese, Italian, German, and English-speaking subjects have reached similar conclusions about the basic cognitive processes and brain regions involved in reading. This consistency reinforces the conclusion that the network of brain areas summarized in Fig. 1 is somehow specialized to support reading and other language processes. However, it should not be taken as evidence that the neural network that underlies reading is inflexible. Quite the contrary, there is growing evidence that a number of factors can have a profound impact on how the network is tuned to most effectively support reading.

Some of the tuning of the reading system may reflect differences in experience or training. For instance, in Italian the mapping between how words look and how they sound is much more consistent than it is in English. Italian-speaking subjects take advantage of this consistency to read aloud words and pseudowords more quickly than English-speaking subjects. Furthermore, when they read words aloud, they produce more brain activation in the left temporoparietal region than do English-speaking subjects; the converse is true in the left inferior temporal region (Paulesu et al. 2000).

Other differences may reflect the use of compensatory or alternative strategies. For instance, as noted above, subjects with developmental dyslexia show a hyperactivation of the left frontal region, and a hypoactivation of the left temporoparietal region. More subtle differences in the activation of brain areas associated with phonology are even apparent in readers at different ends of the normal distribution of reading skill (Rumsey et al. 1999). In subjects with acquired brain damage, neuroimaging studies indicate that right hemisphere brain regions, typically similar in location to sites of damage in the left hemisphere, may begin to play a more substantial role in language tasks (Buckner et al. 1996).

See also: Lexical Processes (Word Knowledge): Psychological and Neural Aspects

Bibliography

Behrmann M, Plaut D C, Nelson J 1998 A literature review and new data supporting an interactive account of postlexical effects in letter-by-letter reading. *Cognitive Neuropsychology* **15**: 7–52

Buckner R L, Corbetta M, Schatz J, Raichle M E, Petersen S E 1996 Preserved speech abilities and compensation following prefrontal damage. *Proceedings of the National Academy of Science USA* **93**: 1249–53

Cohen L, Dehaene S, Naccache L, Lehericy S, Dehaene-Lamberz G, Henaff M A, Michel F 2000 The visual word form area. Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain* **123**: 291–307

Coltheart M, Curtis B, Atkins P, Haller M 1993 Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review* **100**: 589–608

Demonet J F, Fiez J A, Paulesu E, Petersen S E, Zatorre R 1996 PET studies of phonological processing: A critical reply to Poeppel. *Brain and Language* **55**: 352–79

Fiez J A, Petersen S E 1998 Neuroimaging studies of word reading. *Proceedings of the National Academy of Science USA* **95**: 914–21
Iwata M 1984 Kanji versus Kana: Neuropsychological correlates of the Japanese writing system. *Trends in Neuroscience* **7**: 290–3
Moore C J, Price C J 1999 Three distinct ventral occipito-temporal regions for reading and object naming. *Neuroimage* **10**: 181–92
Nakamura K, Honda M, Okada T 2000 Participation of the left posterior inferior temporal cortex in writing and mental recall of kanji orthography. A functional MRI study. *Brain* **123**: 954–67
Nobre A C, Allison T, McCarthy G 1994 Word recognition in the human inferior temporal lobe. *Nature* **372**: 260–3
Patterson K, Lambon-Ralph M A 1999 Selective disorders of reading? *Current Opinion in Neurobiology* **9**: 235–9
Paulesu E, McCrory E, Fazio F, Menoncello L, Brunswick N, Cappa S F, Cotelli M, Cossu G, Corte F, Lorusso M, Pesenti S, Gallagher A, Perani D, Price C, Frith C D, Frith U 2000 A cultural effect on brain function. *Nature Neuroscience* **3**: 91–6
Plaut D C, McClelland J F, Seidenberg M S, Patterson K 1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115
Price C J 1998 The functional anatomy of word comprehension and production. *Trends in Cognitive Sciences* **2**: 281–8
Ricci P T, Zelkowicz B J, Nebes R D, Meltzer C C, Mintun M A, Becker J T 1999 Functional neuroanatomy of semantic memory: Recognition of semantic associates. *Neuroimage* **9**: 88–96
Rumsey J M, Horwitz B, Donohue B C, Nace K L, Maisog J M, Andreason P 1999 A functional lesion in developmental dyslexia: Left angular gyral blood flow predicts severity. *Brain and Language* **70**: 187–204
Sakurai Y, Momose T, Iwata M, Sudo Y, Ohtomo K, Kanazawa I 2000 Different cortical activity in reading of Kanji words, Kana words and Kana nonwords. *Cognitive Brain Research* **9**: 111–15
Shaywitz S E, Shaywitz B A, Pugh K R, Fulbright R K, Constable R T, Mencl W E, Shankweiler D P, Liberman A M, Skudlarski P, Fletcher J M, Katz L, Marchione K E, Lacadie C, Gatennby C, Gore J C 1998 Functional disruption in the organization of the brain for reading in dyslexia. *Proceedings of the National Academy of Sciences USA* **95**: 2636–41
Simos P G, Papanicolaou A C, Breier J I, Fletcher J M, Wheless J W, Maggio W W, Gormley W, Constantinou J E C, Kramer L 2000 Insights into brain function and neural plasticity using magnetic source imaging. *Journal of Clinical Neurophysiology* **17**: 143–62
Tarkiainen A, Helenius P, Hansen P C, Cornelissen P L, Salmelin R 2000 Dynamics of letter string perception in the human occipitotemporal cortex. *Brain* **122**: 2119–213

J. A. Fiez

Reading Nonalphabetic Scripts

1. *The Japanese Language*

Linguistically, Japanese does not relate to Chinese, Korean, or English, though it is similar to Korean in syntax, but not in speech sounds or vocabulary. However, Japanese has borrowed many words and characters, mainly from Chinese. Almost all indigenous people living in Japan speak Japanese, and ethnic Japanese living outside of Japan also do so. There are about 130 million Japanese speakers, and the language ranks sixth among the languages of the world after Chinese, English, Russian, Hindi, and Spanish.

The Japanese language employs five vowels and 16 consonants. Compared to other languages such as English, the Japanese sound system contains few sounds. The Japanese sound system uses a unit called a 'mora,' which is a short beat, the time taken to pronounce a short syllable. The mora is the main unit in the Japanese language, and is also the unit of Japanese phonetic script—hiragana (or hirakana) and katakana. Each Japanese syllabary sign is based on a mora. There are 110 different moras, or syllables, in Japanese. It is one of the smallest inventories of any major languages. For example, there are 400 syllables in Chinese and several thousand in English. The small inventory of Japanese sounds gives rise to many homophones, but it is easy to represent the language by a simple syllabary.

The Japanese vocabulary consists of four word types: Japanese origin (47.5 percent), Sino-Japanese (36.7 percent; long ago adopted from China), foreign loan words (9.8 percent; mostly European adopted in modern times), and hybrids (6 per cent; Sino-Japanese or foreign word stems with Japanese endings).

2. *The Japanese Writing System*

The Japanese writing system consists of five kinds of script—kanji, hiragana, katakana, roman alphabet, and arabic numerals. Among these, the first three script types are employed most frequently in writing Japanese. Japanese text is therefore written using these five kinds of script, and it is also true that both the traditional vertical (tategaki) from top to bottom, and the horizontal (yokogaki), from left to right, styles are used depending on the kind of being written. For example, newspapers are printed in takegaki style, whereas academic journals are printed in yokogaki style.

3. *Kanji Script and Reading*

As a result of waves of Chinese and Korean emigration to Japan during the fourth and seventh centuries AD, Chinese-based culture, mainly in the form of Confucian classics, medicine, and Buddhism, came to Japan. A writing system with Chinese characters was, therefore, the only script until two forms of syllabary, hiragana and katakana were created from kanji during the ninth century AD. 'Kan-on' kanji readings were borrowings from Confucian scholarship. Chinese 'hanzi,' associated with the Han Dynasty period, became Japanese 'kanji,' which remains the backbone of the Japanese writing system. As Japanese kanji

borrowed Chinese hanzi in distinct historical periods, different pronunciations were adopted for these kanji, and makes reading complicated in modern writing systems. This feature of having multiple readings for many kanji characters in Japanese is different from those of Chinese and Korean, which have only a single reading.

The Japanese language uses not only Chinese and native readings of the logographic characters, but also several varieties of each type of reading, which makes the oral reading of characters complicated. For example, Mandarin Chinese read the simple character for 'ten' 十 as one sound 'shi,' while the Korean reading is 'sip' (no tone or soundless). Both languages use only one reading, but Japanese speakers read the character as 'to,' 'toh,' 'so,' 'jyu,' 'jutt-,' all without tone (Taylor and Taylor 1995). The first three sounds are examples of Kun readings (Japanese native readings), while the others are 'On' readings (Chinese readings). The Kun reading is quite different from the Chinese because the two languages are unrelated. Besides the Kun reading, most kanji have one or more On readings which are not identical with Chinese sounds. Depending upon when and where the Chinese characters were borrowed, four different types of On reading (go-on, kan-on, toh-soh-on, and the habitual one) were agreed. Most Japanese kanji are given only one of the four types of On reading.

The number of kanji is enormous. Kanji dictionaries contain between 12,000 and 50,000 entries. In 1981, the Japanese Ministry of Education attempted to restrict the number of kanji as 'joyo kanji' to be used in the educational system. This contains 1945 kanji, but because almost all kanji possess both On and Kun readings, Japanese people therefore learn a total of 4084 kanji readings during their first 9 years of education. Generally, each kanji consists of a combination of radicals. There are 250 kinds of radical, each consisting of between one and 17 components (strokes). Most radicals consist of two to nine strokes (Ueda et al. 1993). As the number of kanji is so huge, this reduces any difficulty in understanding homophones. However, there is a clear asymmetry between reading and writing ability, and it is true that some adult Japanese fail to read even joyo kanji.

Cognitive psychologists once suggested that the processing route of reading depended on the type of orthography—that is, kanji has a single route from orthography to semantics, bypassing the contribution of phonological information, while the kana travel a phonological route to access semantics. But this is too simple an explanation. Recent reviews of kanji studies suggest that it is reasonable to conclude that, depending upon the task settings, familiarity, frequency, and complexity of kanji, one of the most efficient processing routes of phonology-based or grapheme-based routes (or both routes) is employed to access semantics (Kess and Miyamoto 1999). However, in the case of high-frequency and high-familiarity kanji, semantic access is achieved only through the grapheme-based route, and this is not specific to Japanese. There are examples in alphabetic systems where the cognitive route is also direct from grapheme to semantics.

Table 1
Samples of radicals

Radicals	Examples	
Hen	波(wave)	河(river)
Tsukuri	清(clear water)	晴(clear weather)
Kanmuri	草(grass)	花(flower)
Ashi	熊(bear)	馬(horse)
Kamae	国(state)	囲(to enclose)
Tare	庭(garden)	店(shop)
Nyo	返(to return)	逃(to escape)

As shown in Table 1, complex kanji consist of semantic and phonetic radicals, hen (left hand), tsukuri (right hand), kanmuri (top), ashi (bottom), kamae (enclosure), tare (top-left), and nyo (bottom-left). Each acts differentially in reading. That is, generally, semantic radicals give clues to meanings, whereas phonetic radicals give clues to pronunciation. Research findings also suggest that the right hand radical is more informative than the left.

Many words in Japanese are represented not only by individual kanji but also by a compound which is a polysyllabic composite of between two and four kanji, usually carrying an On reading. More than 50 percent of the words in most Japanese dictionaries are compound kanji. Research findings also suggest that the reading processes of kanji compound words constitute another grouping. In addition to the frequency of the compound word itself, the first kanji of a compound is important, because the first kanji defines the range of possible collocations in the cohort of potential kanji compounds.

There is strong evidence that the individual kanji are processed as a recognition unit and kanji compounds themselves are also recognition units. A case has been that one agraphic patient was able to read the single kanji 手 'hand' and 紙 'paper' but failed to read their coupling as 手紙 'letter,' while another patient had less difficulty in reading a compound word, but could not read the single kanji which constituted the compound word (Hatta et al. 1998). This double dissociation suggests that single and compound kanji words are stored with different addresses in the mental lexicon.

4. *Kana Scripts and Reading*

In addition to kanji, there are two syllabary types of script in Japanese—hiragana and katakana (see Table 2). The kana unit represents a sound unit, which corresponds roughly to a syllable. Both scripts contain

Table 2
Hiragana, katakana, and kanji descriptions of the words 'firefly' and 'dreams'

	Firefly	Dream
Hiragana	ほたる	ゆめ
Katakana	ホタル	ユメ
Kanji	蛍	夢

46 basic symbols and 25 additional symbols with diacritic marks. Hiragana and katakana share the same syllabic reference points, therefore the same syllable can be written by either system. Hiragana is cursive, as shown in Table 2, and used mainly for some content words, morphological endings, function words, and the rest of the grammatical scaffoldings of sentences. In contrast, katakana is angular in shape and is used for writing foreign loan words.

Kana developed gradually out of Chinese characters in the following order—kanji, kanji as phonetic signs, simplified kanji shapes, and kana. During this development, mainly female authors used hiragana to write letters, poems, diaries, and stories; this is because hiragana is a graceful, cursive form of its original kanji. For a long time, it has been the convention that a cursively simplified hiragana is used for grammatical morphemes, and the angular simplified katakana is used for foreign loan words and onomatopoeia. However, an explanation for the use of both types of script because of differing roles in syntax cannot be the only reason why the Japanese have kept three kinds of script usage irrespective of high cost. Japanese people might have found it useful or convenient to express semantic emotional information by employing different types of script.

Conventionally, there is a connection between orthography type and the words used. This has a strong effect on Japanese word readings. Japanese people possess a strong, emotional semantic image for each type of script—kanji is seen as masculine, hiragana feminine, and katakana exotic. Japanese people sometimes employ this emotional image of a script when writing, irrespective of the conventional usage of scripts and words. For example, reading experiments suggested that when Japanese people see the kanji word for chair 椅子, an image of an old-fashioned, strongly-built chair is generated; when they see the katakana word for chair イス, they imagine a modern and elegant chair; and in the case of the hiragana word for chair いす, they may imagine a simple wooden chair (Ukita et al. 1996).

5. *Summary of Research Findings in Cognitive and Neuropsychological Studies*

The results of earlier cognitive psychological experiments suggest that orthographic configurations can be used to access directly phonological form, meaning, or both, depending on the processing task requirement. Although the role and the importance of these factors may differ for each orthography in respect to the processing strategy favored and the processing times required, Japanese seems not to be an exceptional writing system, but possesses idiosyncratic requirements of the world's writing systems.

However, neuropsychological studies have suggested that kanji and kana are processed differentially and stress the importance of orthography difference in reading (Iwata 1984, Yamadori 1975). In alexia with agraphia due to angular gyrus lesions, kana reading is impaired more severely than kanji, whereas left posterior–inferior temporal lesions impair kanji reading (Iwata 1984, Kawamura et al. 1987). These findings support the hypothesis that there are two pathways of reading in Japanese: one is the ventral pathway from the visual area in the occipital cortex via the posterior part of the inferior and middle temporal gyli of the left hemisphere, which is used in semantic reading of kanji; the other is the dorsal route from the visual area via the angular gyrus to Wernicke's area, and is used for the phonological processing of kana reading.

Regarding the visual half field, several studies have revealed that the right hemisphere contributes more than the left in kanji recognition. For example, a tendency toward a right hemisphere advantage was reported in the case of single kanji (Hatta 1977), and in the case of two-kanji non-word pairs (Sasanuma et al. 1977). On the other hand, a left hemisphere advantage was reported in the case of two-kanji compound words (Hatta 1978). By employing a very recent new research method (MEG), Koyama et al. (1998) revealed that not only the left but also the right hemisphere is activated in reading kanji words, and support the view of the right hemisphere engagement in kanji reading. These studies suggest that hemisphere function in kanji reading is not the same as in the reading of English and kana words.

The discrepancy between cognitive psychological and neuropsychological research findings in Japanese reading leaves us to conclude that there is a language non-specific cognitive processing in reading. Further studies concerning non-alphabetical scripts are necessary to answer this question.

See also: Dyslexia (Acquired) and Agraphia; Dyslexia, Developmental; Letter and Character Recognition, Cognitive Psychology of; Reading, Neural Basis of; Reading Skills; Writing, Evolution of; Writing Systems

Bibliography

Hatta T 1977 Recognition of Japanese kanji in the left and right visual fields. *Neuropsychology* **15**: 685–88

Hatta T 1978 Recognition of Japanese kanji and Hiragana in the left and right visual fields. *Japanese Psychological Research* **20**: 51–9

Hatta T, Kawakami A, Tamaoka K 1998 Writing errors in Japanese kanji. *Reading and Writing* **10**: 457–70
Iwata M 1984 Kanji versus kana: Neuropsychological correlates of the Japanese writing system. *Trends in Neuroscience* **7**: 290–93
Kawamura M, Hirayama K, Hasegawa K, Takahashi N, Yamaura A 1987 Alexia with agraphia of kanji (Japanese morphograms). *Journal of Neurology, Neurosurgery and Psychiatry* **50**: 1125–29
Kess J F, Miyamoto T 1999 *The Japanese Mental Lexicon*. Benjamins, Amsterdam
Koyama S, Kakigi R, Hoshiyama M, Kitamura Y 1998 Reading of Japanese kanji (morphograms) and kana (syllabograms): A magneto-encephalographic study. *Neuropsychology* **36**: 83–98
Sasanuma S, Itoh K, Mori K, Kobayashi Y 1977 Tachistoscopic recognition of kana and kanji words. *Neuropsychology* **15**: 547–53
Taylor I, Taylor M M 1995 *Writing and Literacy in Chinese, Korean, and Japanese*. Benjamins, Amsterdam
Ueda M, Okada M, Iijima T, Sakaeda T, Iida D 1993 *Shin Daijiten*. Kodansha, Tokyo
Ukita J, Sugishima M, Minagawa M, Inoue M, Kashu K 1996 *Psychological Research on Orthographic Forms in Japanese*. Psycholology Monographs 25, Japanese Psychological Association
Yamadori A 1975 Ideogram reading in alexia. *Brain* **98**: 231–8

T. Hatta

Reading Skills

The question of reading skill includes definitional and substantive components. The definitional question is answered as follows: reading skill is an individual's standing on some reading assessment. Skilled readers are those who score above some standard on this assessment; readers of low skill are those who score below some standard. The substantive question is this: What are the processes of reading that produce variation in assessed reading skill? This question is the focus here: given that two individuals differ in some global assessment of their reading, what differences in reading processes are candidates to explain this difference?

1. *Normal and Special Variation in Reading Skill*

Is a single analysis of reading skill sufficient to characterize both 'garden variety' variation and severe reading difficulties? Severe difficulties in reading in the absence of general intellectual problems is the standard definition of *dyslexia*, or specific reading disability. The discrepancy between achievement in reading and achievement in other domains is what sets dyslexia apart. However, there are reasons to blur the distinction between specific and nonspecific reading problems. Individuals in the two categories may differ only in their achievements in some other nonreading area. The processes that go wrong in a specific disability may not be much different from those that go wrong for an individual who also has a problem in some other area (Stanovich and Siegel 1994). For both groups, difficulties in reading must be understood in terms of the processes of reading. What are the processes that can go wrong?

2. *A Process Analysis*

Reading processes depend on the language of the reader and the writing system that encodes that language. The units of the writing system are converted into mental representations that include the units of the language system. Specifically important are (a) the identification of words and (b) the engagement of language and general cognitive mechanisms that assemble these words into messages.

It is visual word identification that is the process most distinctive to reading. Beginning with a visual input—a string of letters—perceptual processes produce the activation of the *grapheme units* (individual and multiple letters) that constitute words. In traditional models of human cognition, the words are represented in a *lexicon*, the reader's mental representation of word forms and meanings. Successful word reading occurs when there is a match between the input letter string and a word representation. As part of this process, *phonological units*, including individual phonemes associated with individual letters, are also activated. The joint contribution of graphemic and phonological units brings about the identification of a word. It is common to refer to the phonological contribution to this process as *phonological mediation*. It is also common to represent two pathways, one from graphemic units to meaning directly, and one from graphemic units to phonological units, and then to meaning (the mediation pathway). The issues of mediation and one-or-two routes are central to alternative theoretical models of the reading process.

The alternative models do not assume a representation system that stores words. In one class of nonrepresentational models, words emerge from patterns of parallel and distributed activation (Plaut et al. 1996). In resonance models, word identification results from the stabilization of dynamic patterns that are continuously modified by interactions among inputs and various dynamic states resulting from prior experience (Van Orden and Goldinger 1994). An interesting feature of this model is that patterns of graphic-phonological activation stabilize more rapidly than do patterns of graphic-semantic activation. In effect, a word form becomes identified primarily through the convergence of orthography and phonology. Meaning is slower to exert an influence on the identification process.

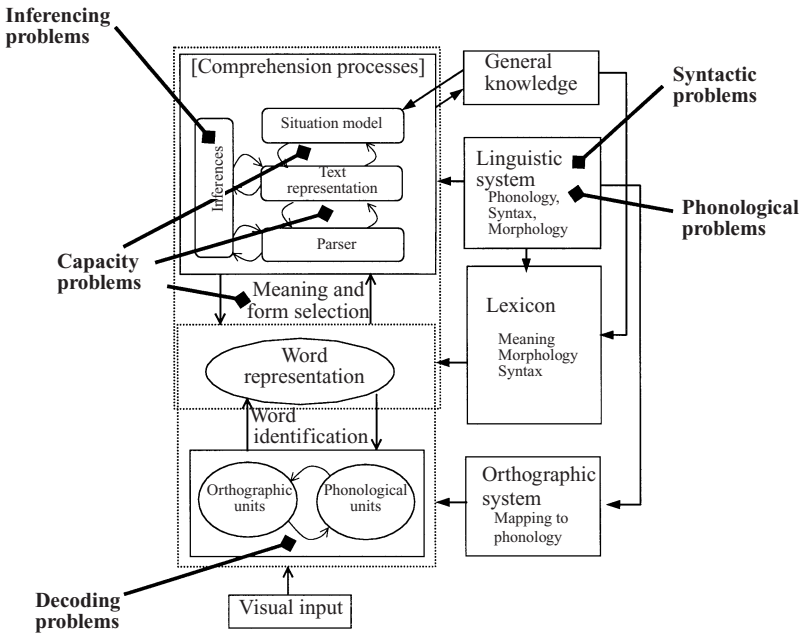

Figure 1
A schematic view of reading processes with candidate sources of reading problems. Adapted from Perfetti (1999)

In considering reading skill, it is possible to ignore differences among theoretical models to a certain extent. However, the theoretical models of reading processes actually make some commitments about the sources of reading problems. For example, a dual route model allows two very different sources of word reading difficulties: Either the direct (print-to-meaning) route or the indirect (print-to-phonology-to-meaning) route can be impaired (Coltheart et al. 1993). This provides a model for both developmental and acquired dyslexia. In acquired dyslexia, surface dyslexics are assumed to have selective damage to the direct route; phonological dyslexics are assumed to have selective damage to the phonological route. For developmental dyslexia, children may have a phonological deficit or an 'orthographic' (direct route) deficit. However, single mechanism models with learning procedures can give an alternative account of developmental dyslexia: only one type, phonological dyslexia, is the result of a processing defect; surface or orthographic dyslexia becomes a delay in the acquisition of word-specific knowledge (Harm and Seidenberg 1999). Such an example illustrates that understanding individual differences in reading is part of the same problem as theoretical understanding of reading processes.

Figure 1 schematizes an organization of the cognitive architecture for reading. Reading begins with (a) a visual input that, with immediate use of phonology, leads to word identification that (b) yields semantic information connected to the word as constrained by the current context. A word immediately is (c) integrated syntactically with an ongoing sentence parse and (d) semantically with an ongoing message interpretation (proposition). As sentences are read (e), an integrated text representation is developed, consisting of interrelated propositions. To establish a reasonably specific understanding of a text, (f) inferences establish a coherent referential model of what is being read.

Potential reading difficulties are seen in each component. At the word level, a reader could have trouble with processing letter strings or in selecting the required meaning of a word. A reader might have defective phonological processes, limiting word identification and memory for words. On the comprehension end, a reader could have syntactic difficulties or fail to generate inferences or to monitor comprehension. Because these processes can demand mental resources from a limited pool, a very general constraint is working memory limitations. Thus reading difficulty can arise from a general limit in functional working memory capacity or from the failure of some processes to escape this limit. The following sections consider some of these possibilities.

2.1 Lexical Orthographic–Phonological Processes

Readers who fail to read words accurately fail to comprehend. Thus, word-level problems are potentially the most important in creating reading failures,

because they lead both to word reading problems themselves and to derivative comprehension problems. For example, verbal efficiency theory (Perfetti 1985) assumes that readers who lack efficient word identification procedures are at risk for comprehension failure.

A word identification problem is a failure of phonological and orthographic information to cohere onto a word identity (Fig. 1). Knowledge of letters and knowledge of phonemes that connect with letters in specific letter environments are necessary. Less skilled readers tend to have weaker knowledge of word spellings (orthography) and of word sounds (phonology), and especially weaker knowledge of orthography–phonology connections (decoding). Various measures assess phonological processes, but the ability to read legal nonwords (pseudowords) has become a benchmark for decoding ability. Both the phonological and orthographic components of word reading appear to have a significant heritable component (Olson et al. 1999).

The relationship between orthographic and phonological knowledge may be complex. As indicated in Fig. 1, an orthographic processing event is also a phonological processing event, perhaps implying that the two knowledge sources develop in tandem. However, the effects of reading experience may be a bit different for these two components. Early encounters with words may be crucial for establishing (phonological) decoding knowledge with orthographic knowledge acquired more gradually, as individual words are 'practiced' in reading situations. Stanovich and West (1989) found that reading experience accounts for variance in word reading skill beyond that accounted for by decoding. However, this does not mean that word knowledge gained through experience is independent of phonology. Indeed, early in reading experience, decoding unfamiliar words may be the primary mechanism for establishing orthographic representations of specific words (Share and Stanovich 1995). More generally, phonological processes that link graphic inputs to word identities may enable gains from further experience in reading, increasing the quality and quantity of word representations (Perfetti 1992).

Referring to 'orthographic' and 'phonological' as separate sources of knowledge does not entail a commitment to two separate routes in word identification. Whether as defects in the phonological component of single mechanism or as impairments of a distinct processing route, phonological problems may be the central deficiency for both specific reading disability and nonspecific low reading skill (Stanovich 1988). They can arise at many levels, including speech processing mechanisms and sensitivity to phonological structures, as well as written word identification. Becoming aware that spoken words consist of meaningless sound segments (phonemes) is important in learning to read, and children who do not become aware are at risk for failure (see Liberman and Shankweiler 1991). Moreover, phonological processing problems may have consequences throughout the reading process: learning to read words, remembering just-read words, and understanding phrases and sentences are all processes operating on partly phonological representations.

2.2 *Processing Lexical Meaning*

The ability to get context-appropriate meaning from words is central to reading skill. At one level, this is a question of vocabulary. The more words a reader knows, the better the comprehension. At another level, it is the ability to select the right meaning of a word in a given context. Although the description of lexical meaning selection has become complex, a still widely (not universally) shared conclusion is this: the selection of word meaning proceeds in two-stages: (a) a general activation stage in which a word is accessed and its associated meanings nonselectively activated, and (b) a selection stage in which the meaning appropriate for context is selected while meanings inappropriate for context are suppressed.

One hypothesis is that less skilled readers are less effective in selecting a contextually appropriate meaning. According to the structure building framework (Gernsbacher 1990), readers build a coherent framework for a text by enhancing concepts required by the text while suppressing those that are irrelevant. The suppression hypothesis is that less skilled readers have deficient suppression mechanisms. To illustrate, in the sentence, 'He dug with the spade,' the final word has two meanings, but only one fits the context of the sentence. However, when adult readers are immediately asked to decide whether a following word is related to the meaning of the sentence, their decisions are initially slow for the word 'ace' (related to the inappropriate meaning of spade). Both appropriate and inappropriate meanings may be activated at first. With more time before the appearance of 'ace,' skilled readers show no delay in rejecting it; i.e., they 'suppress' the irrelevant meaning. However, less skilled readers continue to react slowly to 'ace,' as if they have not completely suppressed the irrelevant meaning of 'spade.'

Whether ineffective use of context is a source of reading problems has become a complex issue. Prescientific beliefs on this question seemed to be that poor readers failed to use context in reading words. However, research on children's word identification led to the opposite result: less skilled readers use context in word identification at least as much and perhaps more than do skilled readers (Perfetti 1985, Stanovich 1980). Less skilled comprehenders are good users of context, which they use in compensation for weaker word identification skill to identify words. The suppression proposal adds to the poor reader's basic

word identification problem a specific suppression problem, the effect of which is a less productive use of context.

2.3 Processing Syntax

Less skilled readers often show a wide range of problems in syntax and morphology. The question is whether such problems, which are found across a wide age range, arise from some deficit in processing syntax or from some other source that affects performance on syntactic tasks. One possibility is that syntactic problems reflect a lag in the development of linguistic structures. An alternative hypothesis is that the syntactic problems reflect constraints of working memory or lexical processing limitations.

To illustrate one class of syntactic problem, consider two sentences with relative clauses below. (a) is the easier subject relative; (b) is the more difficult object relative.

(a) The girl that the boy believed understood the problem.

(b) The girl that believed the boy understood the problem.

The greater difficulty of (b) compared with (a) can arise from different degrees of interference they produce in the attempt to assigning a subject for 'understood.' For syntactic reasons, the interference produced by the object relative (b) is greater than the interference produced by the subject relative (a). Details aside, if syntactic deficits rest on lack of specific syntactic knowledge, then one might expect some qualitative differences between skilled and less skilled readers. However, in the relative clause example and in other syntactic examples, patterns of difficulty for skilled and less skilled young readers tend to be the same, suggesting quantitative rather than qualitative differences (Crain and Shankweiler 1988).

Research with adults suggests that both reading problems and spoken language problems (aphasia) may arise from processing limitations rather than structural deficits (Carpenter et al. 1994). In a study of subject and object relatives similar to those in (a) and (b), King and Just (1991) found that readers with low working memory have problems with object-relative sentences. More interesting, these problems (as seen in reading times for words) were most severe where the processing load was hypothesized to be the greatest—at the second verb in the object relative, i.e., 'understood' in (b). Comprehension difficulties may be localized at points of high processing demands—whether from syntax or something else. If this analysis is correct, then the problem is not intrinsic deficits in syntax, but the processing capacity to handle complexity.

Another perspective on this issue is the opportunity for practice. Because some syntactic structures are more typical of written language than spoken language, the opportunity for practice is limited by the ability to read. Thus, continuing development of reading skill as a result of initial success at reading—and the parallel increasing failure as a result of initial failure—is undoubtedly a major contributor to individual differences in reading.

2.4 Processing Text

The source of skill differences in text comprehension can lie in the lexical and working memory processes considered above or in higher-level processes specific to reading texts. In addition to identifying words, parsing sentences, and encoding context-sensitive meanings, readers must use knowledge from outside of the text and apply procedures that derive intended meanings. The overall complexity of text comprehension implies several possibilities for processing failure, any resulting in a less integrated or less coherent representation of the text.

One example is the processes of coreference, by which a reader maps a pronoun to an already established referent, as illustrated in (c).

(c) Jane saw Margaret shopping in the grocery store. She was buying bread.

When there are two possible referents for a pronoun, less skilled readers may take longer to assign the pronoun to the intended referent (Margaret in this case), at least when the referent is not the first noun in the preceding sentence (Frederiksen 1981). If this simple pronoun case is general, less skilled readers may be less adept at integrating referential information across sentences.

The text processing component that has received the most attention as a source of individual differences is making inferences. Oakhill and Garnham (1988) summarize evidence suggesting that less skilled readers fail to make a range inferences in comprehension. When skill differences are observed, an important question is whether they occur in the absence of problems in lexical skills, working memory, or other lower-level factors. The answer to this question remains unclear in much of the research. However, there has been some success in identifying a group of children whose problems can be considered comprehension-specific, although highly general across reading and spoken language (Stothard and Hulme 1996).

Another example is comprehension monitoring, the reader's implicit attempts to assure a consistent and meaningful understanding. Skilled readers can use the detection of a comprehension breakdown (e.g., an apparent inconsistency) as a signal for rereading and repair. Less skilled readers may fail to engage this monitoring process (Baker 1984, Garner 1980). Low skilled comprehenders seem especially poor at detecting higher level text inconsistencies that negate a coherent model of the text content—for example, whether successive paragraphs are on unrelated topics.

However, it is not clear whether these higher level differences are independent of the reader's ability to construct a simple understanding of the text. Some evidence suggests that less skilled comprehenders fail accurately to represent or weight the propositions in a text (Otero and Kintsch 1992). The general interpretive problem here is that comprehension monitoring, like inference making, both contributes to and results from reader's text representation. This makes it difficult to attribute comprehension problems uniquely to failures to monitor comprehension, as opposed to more basic comprehension failures.

Because no texts are fully explicit, the reader must import knowledge from outside the text. Thus, a powerful source of comprehension differences is the extent to which a reader has such knowledge for a given text (Anderson et al. 1977). However, readers of high skill can compensate for lack of knowledge to some extent (Adams et al. 1995). It is the reader who lacks both knowledge and reading skill who is assured failure. The deleterious effect of low reading skill (and its motivational consequences) on learning through reading creates readers who lack knowledge of all sorts.

3. Summary

With a focus on the component processes, individual differences in reading skill become a matter of understanding how these processes and their interactions contribute to successful reading outcomes. Where the successful outcome is reading individual words, the processes are localized in knowledge of word forms—both general and word-specific phonological and orthographic knowledge—and word meanings. Inadequate knowledge of word forms is the central obstacle to acquiring high levels of skill. Severe problems in word reading reflect severe problems in phonological knowledge. Where the successful outcome is comprehension, the critical processes continue to include word processes, and problems in comprehension are associated with problems in word processing. In addition, processes that contribute to basic sentence understanding and sentence integration become critical. Processes that provide basic propositional meaning, including word meaning selection and parsing, and those that establish coherent text representations (integration processes, inferences, monitoring, conceptual knowledge) become critical to success. Less skilled readers, as assessed by comprehension tests, often show difficulties in one or more of these processes. Less clear is how to understand the causes of observed failures. A processing model helps to see the relationships among component processes and to guide studies of skill differences. The candidate causes of skill variation cannot be equally probable, when the output of lower-level processes are needed by higher level processes. The identification of basic causes, as opposed to derivative symptoms that reflect the costs of low reading skill, requires research that can provide strong causal tests.

See also: Dyslexia (Acquired) and Agraphia; Dyslexia, Developmental; Dyslexia: Diagnosis and Training; Dyslexia, Genetics of; Eye Movements in Reading; Letter and Character Recognition, Cognitive Psychology of; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Reading, Neural Basis of; Reading Nonalphabetic Scripts; Speed Reading; Word Recognition, Cognitive Psychology of

Bibliography

Adams B C, Bell L C, Perfetti C A 1995 A trading relationship between reading skill and domain knowledge in children's text comprehension. *Discourse Processes* **20**: 807–23

Anderson R C, Reynolds R E, Shallert D L, Goetz E T 1977 Frameworks for comprehending discourse. *American Educational Research Journal* **4**: 367–81

Baker L 1984 Spontaneous versus instructed use of multiple standards for evaluating comprehension: Effects of age, reading proficiency and type of standard. *Journal of Experimental Child Psychology* **38**: 289–311

Carpenter P A, Miyake A, Just M A 1994 Working memory constraints in comprehension: Evidence from individual differences, aphasia, and aging. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA, pp. 1075–122

Coltheart M, Curtis B, Atkins P, Haller M 1993 Models of reading aloud: Dualroute and parallel-distributed-processing approaches. *Psychological Review* **100**: 589–608

Crain S, Shankweiler D 1988 Syntactic complexity and reading acquisition. In: Davison A, Green G M (eds.) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*. Erlbaum, Hillsdale, NJ, pp. 167–92

Frederiksen J R 1981 Sources of process interactions in reading. In: Lesgold A M, Perfetti C A (eds.) *Interactive Processes in Reading*. Erlbaum, Hillsdale, NJ, pp. 361–86

Garner R 1980 Monitoring of understanding: An investigation of good and poor readers' awareness of induced miscomprehension of text. *Journal of Reading Behaviour* **12**: 55–63

Gernsbacher M A 1990 *Language Comprehension as Structure Building*. Erlbaum, Hillsdale, NJ

Harm M W, Seidenberg M S 1999 Phonology, reading and dyslexia: Insights from connectionist models. *Psychological Review* **106**: 491–528

King J, Just M A 1991 Individual differences in syntactic processing. *Journal of Memory and Language* **30**: 580–602

Liberman I Y, Shankweiler D 1991 Phonology and beginning reading: A tutorial. In: Rieben L, Perfetti C A (eds.) *Learning to Read: Basic Research and its Implications*. Erlbaum, Hillsdale, NJ, pp. 3–17

Oakhill J, Garnham A 1988 *Becoming a Skilled Reader*. Blackwell, New York

Olson R K, Forsberg H, Gayan J, Defries J C 1999 A behavioral-genetic analysis of reading disabilities and component processes. In: Klein R M, McMullen P A (eds.) *Converging Methods for Understanding Dyslexia*. MIT Press, Cambridge, MA, pp. 133–53

Otero J, Kintsch W 1992 Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science* **3**(4): 229–35
Perfetti C A 1985 *Reading Ability*. Oxford University Press, New York
Perfetti C A 1992 The representation problem in reading acquisition. In: Gough P B, Ehri L C, Treiman R (eds.) *Reading Acquisition*. Erlbaum, Hillsdale, NJ, pp. 145–74
Perfetti C A 1999 Comprehending written language. A blue print of the Re. In: Hagoort P, Brown C (eds.) *Neurocognition of Language Processing*. Oxford University Press, Oxford, pp. 167–208
Plaut D C, McClelland J L, Seidenberg M S, Patterson K 1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115
Share D L, Stanovich K E 1995 Cognitive processes in early reading development: Accommodating individual differences into a model of acquisition. *Issues in Education* **1**: 1–57
Stanovich K E 1980 Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly* **16**: 32–71
Stanovich K E 1988 Explaining the differences between the dyslexic and the garden-variety poor reader: The phonological-care variable-difference model. *Journal of Learning Disabilities* **21**: 590–604
Stanovich K E, Siegel L S 1994 Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology* **86**(1): 24–53
Stanovich K E, West R F 1989 Exposure to pring and orthographic processing. *Reading Research Quarterly* **24**: 402–33
Stothard S, Hulme C 1996 A comparison of reading comprehension and decoding difficulties in children. In: Cornoldi C, Oakhill J (eds.) *Reading Comprehension Difficulties: Processes and Intervention*. Erlbaum, Hillsdale, NJ, pp. 93–112
Van Orden G C, Goldinger S D 1994 The interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception and Performance* **20**: 1269–1291.

C. A. Perfetti

Real Estate Development

1. What is Development?

Real estate development is the process of *adding value* to real estate. While most people think of development in terms of new buildings and subdivisions, development refers to any situation where value is added to a property. Developers take 'development risk' which ranges from leasing-risk associated with buying buildings with high vacancy to buying raw land with no utilities or zoning, and undertaking all the risks associated with obtaining public approvals, financing, construction, leasing, and operations of a new development project.

Developers initiate, plan, and coordinate the entire development process—buying the land, designing the product, arranging financing, obtaining public approvals, constructing the building, leasing, managing, and ultimately selling the completed project. Development is distinguished from owning and managing property because it includes at least some element of adding value. While most developers also own and manage their buildings, many real estate investors and owners buy buildings that are completed and fully leased and which do not involve adding value through some form of development.

1.1 What Distinguishes Good Development From Bad?

While there is no absolute standard to distinguish good development from bad, the test is a simple one: Does the development enhance the social welfare of the community? (Peiser 1990). Providing new housing and office space enhances the community, but if it does so at the expense of environmentally precious resources, then on balance it may reduce the social welfare. Unfortunately social welfare is difficult to measure—especially the indirect costs and benefits of externalities associated with the development. Obtaining approvals for development is a political process that often is long and costly. One of the biggest challenges facing city planners and council members is to create an approval process that is unbiased and gives the various stakeholders appropriate voice in the outcome. At the same time, they must do their best to streamline the process and prevent unnecessary delays and costs.

1.2 Academic and Professional Antecedents

Real estate development relates to a number of academic fields including economics, urban planning, architecture, finance, law, political science, civil engineering, and geography. Urban economics provides the foundation for understanding real estate value and how cities change over time (Mills 1972). Real estate finance, public finance, land use, planning, and environmental law and economics, infrastructure, market analysis, demography, architecture, and landscape architecture all have significant bodies of research that contribute to our understanding of urban development and its impact on cities and countryside (see Brueggeman and Fisher 1997, DiPasquale and Wheaton 1996, Case and Shiller 1989).

Developers are generalists. They need to be conversant with many different disciplines. Because they are responsible for conceiving, planning, and executing the entire process, developers typically are strongest in finance and marketing. Because they determine what is built and how the urban landscape appears, they should be intimately familiar with cities—what works and what does not, both from a

private development perspective and from a government perspective (see Garvin 1996).

Successful development requires knowledge of all of the real estate disciplines—brokerage, finance, law, political approvals, design, construction, and management. Developers usually have several years of real estate experience before they undertake their own projects, most often in brokerage or real estate finance and investment. Legal, architectural, planning, and construction backgrounds also provide excellent training.

2. Industry Structure

2.1 Ownership

The real estate development industry is an industry in transition. Historically, it consisted primarily of private entrepreneurs who operated in a single market. Today, large publicly held national companies are emerging as significant owners of property, although their role and competitive position as developers is still being defined.

The different segments of the real estate industry comprise almost 20 percent of the nation's gross domestic product (Urban Land Institute 1997). The real estate industry is composed of many different players (see Fig. 1). Developers are the entrepreneurs who create and produce real estate. They may continue to own and manage the completed buildings or they may sell them to other parties. While developers have traditionally come from the private sector in the USA, redevelopment agencies and public housing departments act as public developers, often through public–private partnerships. Many corporations and public agencies including the armed forces also develop property for their own use, which is later recycled into the inventory of space available for anyone.

Each of the players is involved in a different phase of producing or managing real estate. Developers, architects, lenders, planners, and city council workers are more involved at the front end. Owners, property managers, building departments, police and fire departments, and brokers tend to be more involved with the real estate after it is built.

Developers
Contractors
Lenders
Brokers
Property managers
Architects/Engineers
Planners/City council
Government agencies

Figure 1
Real estate industry players

Location	**Property type**
CBD	*Residential*
Inner city	*Office*
Suburbs—Wealthy	*Industrial*
Graying	*Retail*
Exurbs	*Hospitality*
Rural	*Land*

Figure 2
Real estate markets are defined by location and property type

Owners of property include the following:
Corporations
REITS (Real Estate Invest Trusts)
Banks—REO (Real Estate Owned)
Partnerships/Joint Ventures (JVs)
Individuals

2.2 Property Type

Figure 2 provides a breakdown of the real estate industry by location and property type. Each location represents a different and distinct market. All property types may be found in all locations. Certain property types, however, such as major office buildings and hotels are more likely to be developed in the Central Business District (CBD) and wealthy suburban nodes—typically located at the intersections of major freeways whereas industrial buildings, apartments, and shopping centers are more ubiquitous.

The real estate industry is divided into six main property types: residential, office, commercial, industrial, hospitality, and land (see Fig. 2). Each property type is virtually a separate industry, serving different purposes, markets, and customers. Although industrial warehouses, for example, may have completely different development requirements from apartments, they share a common institutional framework with respect to land, financing, legal aspects, and even construction. Still, each product is distinct, and value is created from different features. Apartments and industrial both benefit from locations near major highways. Industrial developers, however, prefer to avoid residential areas where homeowners complain about truck traffic. And apartment developers are attracted to areas with good schools, good retail services, recreation, and other amenities that are much less important to industrial tenants.

Developers, lenders, brokers, professional service providers, and most suppliers tend to specialize in one property type. While there is considerable cross-over

between property types, such as air conditioning subcontractors who work on both office buildings and apartments, professionals develop a network of relationships over the years that favor those who concentrate on a single property type.

2.3 For-Sale vs. Income-Producing Property

A critical distinction between property types is whether they are 'for-sale' or are 'income-producing.'

For-sale property	*Income-producing* property
Land	Apartments
Single-family homes	Office
Condominiums	Retail
	Industrial
	Hotels

For-sale property is sold to a buyer after it is developed whereas income-producing property is held for the cash flow it generates. For-sale property is valued based on how much buyers are willing to pay for enjoying the benefits of ownership—such as the consumption value of living in a home in a good school district near shopping and recreation.

Income-producing property includes five main property types. Values for income properties are based on the cash flow they generate from leases. Office buildings, for example, have tenants who sign 3 to 10 year leases. The developer uses the lease income to obtain long-term mortgage financing. When the office building is sold, investors pay a premium for leases signed by major corporations who have excellent credit ratings. A lease signed by a major corporation like IBM with a AAA credit rating is like purchasing a bond.

Income properties are valued by capitalizing their Net Operating Income (NOI). Capitalization rates are inversely proportional to prices. For example, if a building produces an NOI of $100,000 per year, it is worth $1.25 million on an 8 percent cap-rate ($100,000 divided by 0.08) vs. $1.0 million on a 10 percent cap-rate ($100,000 divided by 0.10). Investors may be willing to purchase a building occupied by a major corporation such as IBM using an 8 percent capitalization rate whereas they would use a much higher cap-rate, say a 10 percent, for the same property occupied by a small non-credit tenant. The rental payments of a major corporation are like a bond—carrying the same credit rating as the corporation itself. In fact, the owner can sue the corporation for lost rent if the tenant defaults on the lease payment.

2.4 Design and Density

In addition to for-sale vs. income property, the development industry may be subdivided by design and density—expressed in units per acre for residential and Floor Area Ratio (FAR) for commercial buildings. Densities for residential property range from one unit/2+ acres for single-family detached homes to 80–200 units/acre for a high-rise apartment building. Residential developers tend to focus either on for-sale single-family homes or income-producing apartments. Figure 3 shows the range in densities for different types of residential construction.

Design	**Density**
Hi rise	*80 + DUs/acre*
Low rise/garden	*30 – 80 DUs/acre*
	18 – 30 DUs/acre
Single family attached	*8 – 16 DUs/acre*
Single family detached	*2 + acres/DU–*
	12 DUs/acre

Figure 3
Residential design and density

3. The Development Process

Figure 4 shows the time line of development for income-producing property. The two main periods—development and operations—involve dramatically different risks, types of financing, and activities (see Peiser 1992).

The 'Development period' runs from the inception of the deal through stabilized occupancy (90–95 percent occupancy). This period is financed by a construction loan, which covers the construction and lease-up of the building. The building is technically placed in operation when the 'Certificate of occupancy' is received. Since lease-up represents a major activity of development and risk, however, it is shown in Fig. 4 as part of the development period. The 'operating period' runs from the point in time when the building reaches stabilized occupancy to the time of sale. Once stabilized occupancy is achieved, the developer is able to obtain a long-term permanent mortgage which is used to pay off ('take out') the construction loan (see Sect. 3.5).

3.1 Stages of Development

The development period in Fig. 4 includes predevelopment, construction, and leasing activities. The predevelopment period runs from project initiation up to the start of construction. It includes feasibility studies, site acquisition, design, and finance.

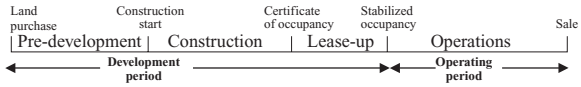

Figure 4
Time line of development

3.2 Feasibility Studies

Many of the critical development decisions are made even before a developer acquires a site. Market feasibility studies help the developer identify the best location and site characteristics for a particular development (see DiPasquale and Wheaton 1996). Market feasibility differs for each product type. It includes demographic projections, income, employment, retail sales, and growth projections. The market study should conclude with estimates of demand for the product expressed in units or square feet/year in a particular location for a particular product-product type, as well as estimates of current and projected supply. The market study tells the developer how much space he can build, what rent to charge, and how quickly it can be absorbed by the market place. The market study identifies competitive properties, and helps the developer determine the best location in which to develop the product. It also tells the developer what the design features and amenities should be in order to attract the greatest number of customers.

For site selection, developers will often consider as many as 50–100 different sites in order to find five sites which they examine in detail. From the five they may pick one site to purchase. Development is said to be a local business because experienced developers will know information about most of the potential development sites, competition, rents, and costs in their market area.

3.3 Site Acquisition

Property acquisition occurs in three main stages: (a) putting the property under contract ('opening escrow'), (b) going hard on the earnest money deposit, and (c) closing. Purchase contracts typically run from 60 to 120 days or longer. When the buyer makes an offer, he places an 'earnest money deposit' of 3–5 percent of the total purchase price. Purchase contracts have a number of contingencies, which the developer wants to satisfy before he consummates the purchase. 'Closing' is the transfer of title to the buyer and the payment of all money to the seller.

After a site is placed under contract, the developer has a 15 to 30 day contingency period to do feasibility studies. The developer will hire consultants to determine the site's suitability for development. The analysis should cover site characteristics, environmental status, topography, drainage, utilities (water, sewer, storm, electricity, phone, gas, and cable), as well as neighboring properties, and the chance of obtaining necessary zoning and building approvals. Once the developer is satisfied that these conditions are suitable for development, he will go hard on the contract. 'Going hard' means that the earnest money becomes non-refundable—when the major contingencies such as good title and finding no environmental problems are satisfied. Since it may take longer than the typical 15–30 day contingency period for the developer to perform the full range of studies necessary for development, the developer will often need to ask for an extension. The developer may also do much of the investigation before the property is even placed under contract or while the purchase contract is being negotiated. In a hot market, sellers are usually unwilling to extend the contingency period because other buyers are waiting in the wings.

For a large project, the feasibility period can extend several years and cost millions of dollars before the developer receives all of the necessary approvals. In a large project, the developers almost always has to purchase the property before many of the feasibility studies are done and long before the public approvals are in place. Such projects carry full 'entitlement' risk and represent the most risky (and potentially the most profitable) form of development.

3.4 Design

Design includes land planning, architecture, engineering, and consulting services that go into producing completed working drawings from which the building is built. The developer will often perform considerable preliminary design work on a property even before 'going hard' in order to determine how much space can be built, and what it will cost based on preliminary construction bids. Design drawings are used by the developer to obtain financing and political approvals. The architect prepares working drawings after the developer approves the final design. The general contractor and the subcontractors use the working drawings to bid the project and to build it.

3.5 Finance

If developers have a distinct expertise, it is in their ability to raise equity and debt money to finance their development activities. Development financing has two main pieces—construction financing to build the project, and permanent financing for operating it. Developers obtain construction financing primarily from commercial banks. Construction loans have short terms—1–2 years—and higher interest rates—2–3 percent above 'prime'—than permanent mortgages. Permanent mortgages are long term—30 years, and have lower interest rates than construction loans because the completed property, especially if it has high-quality tenants, offers better security to the lender than one under construction. Historically, pension funds, insurance companies, and savings and loan associations provided most of the permanent mortgage financing, but today Wall Street is heavily involved in it through securitization, commercial mortgage-backed securities, and conduit lending. As a result, permanent financing is highly competitive in the market place and developers and owners of

Development cost

Land	$ 2,000,000
Building	$ 8,000,000
Total	$10,000,000
Financing	
Equity	$ 2,500,000
Mortgage	$ 7,500,000
Total	$10,000,000

Annual operating income

Gross rent		$ 1,500,000
Vacancy	5 percent	$ 75,000
Effecting gross income		$ 1,425,000
Less operating expenses		$ 425,000
NOI		$ 1,000,000
Less debt service		$ 660,388
Cash throw-off		$ 339,612

***Mortgage assumptions**

Mortgage amount	$7,500,000
Interest rate	8 percent
Term	30 years
Annual debt service	$660,388

Figure 5
Development cost and operating income

property have many more sources for permanent debt than they used to.

Equity financing is the most difficult and expensive money that developers raise. It represents the developer's own cash investment in a deal, or the money raised from 'equity investors.' It is much riskier than debt financing because equity investors receive cash flow only after the mortgage payments are made during operations and after the debt is paid off when the property is sold.

For example, if a building costs $10 million altogether, the developer would typically invest 25–30 percent cash equity and the rest would be covered by debt. Suppose the building at stabilized occupancy (90–5 percent occupancy) generates $1,000,000 in Net Operating Income (revenues minus expenses—$1,000,000 NOI divided by $10,000,000 purchase price equates to a 10 percent capitalization rate). Figure 5 shows the Development Cost, Financing, Annual Operating Income, and Mortgage assumptions for the sample building.

In Fig. 5 the simple return measures are the overall capitalization rate and the cash on cash return. The overall capitalization rate equals the NOI/Total Cost = $1,000,000/$10,000,000 = 10 percent. The *cash on cash return* equals the Cash Throw Off/Equity = $339,612/$2,500,000 = 13.58 percent.

Sophisticated investors focus on before-tax and after-tax internal rates of return (IRR) to determine whether or not they should invest in a particular real estate deal. IRRs represent annual rates of return on equity that are comparable to returns from investing in stocks or bonds. IRR requirements depend on risk. Investors require IRRs on equity that are higher than 20 percent for new properties under development whereas they may require returns as low as 10–12 percent for investing in a fully leased office building with high-quality tenants (see Brueggeman and Fisher (1997) for a detailed discussion of IRRs).

3.6 Construction

Developers hire general contractors to build most nonresidential buildings. Homebuilders and many apartment developers, however, often do their own construction. The general contractor (or homebuilder) hires subcontractors who specialize in constructing various parts of the building such as foundations, framing, glass, granite, electrical, plumbing, heating, ventilating, and air conditioning (HVAC). Legal arrangements between the developer and the contractor may take several forms depending on who bears the risk for delivering the completed building within budget and on time. Common contract arrangements include fixed price, negotiated bid, design-build, fixed-price with a savings clause (where savings are split between the general contractor and the developer), and time and materials. Larger contractors are often unionized while small contractors and

especially those who work on residential projects and in certain 'right-to-work' states such as Texas are typically not unionized and tend to be lower cost. During the construction phase, the developer is most concerned about delivering the building on time and under budget. Since working drawings are often incomplete or have mistakes, a nonconfrontational relationship between the developer and contractor is key to success.

3.7 Leasing and Sales

Marketing for a new project begins even before the developer purchases the land. In order to obtain financing for office and retail projects, the developer often needs to demonstrate that tenants are interested in the property by obtaining letters of intent or signed leases on 30–50 percent or more of the property.

Developers may have in-house marketing personnel or use outside brokers to lease and sell their property. Exclusive brokerage arrangements are common, whereby a brokerage firm is the developer's primary agent. Brokers work on commission which ranges from 4–6 percent of the lease or sale value for smaller deals down to 1–2 percent for larger transactions in the millions of dollars. For leasing office or warehouse space, the developer will typically pay the broker who he has hired as the exclusive agent for his property a 3–4 percent commission for signing a lease if there are no other brokers involved. The tenant often has his own broker, in which case, a higher commission will be split between the two brokers.

Successful leasing involves more than having the lowest rent for the nicest space. The developer typically gives the tenant an allowance for tenant improvements (TIs) which covers interior finish. For law offices in which tenant finish can run \$70–\$80 or more per square foot (compared to standard finish of \$15–\$25 per square foot), the TI allowance can make the difference in getting the law firm to sign a lease.

3.8 Operations

Larger developers usually have their own property management departments to handle the ongoing leasing and management of a building after it is completed. Smaller developers may hire outside firms that specialize in property management. Property management activities include repair and maintenance of the property, leasing, relations with tenants, security, building and payment of all capital and operating expenses. Good developers give careful attention to property management because it sets the tone of relations with tenants and determines their firm's image in the marketplace.

The operating period ends with the sale of the property to a third party. Most real estate deals generate a significant part of their financial returns from appreciation in the value of the property relative to its development or purchase cost. IRRs are highly sensitive to the length of the holding period. The shorter the holding period, say 3 years or less, the higher the IRR. It is standard practice when preparing a financial proforma to assume a 7 or 10 year holding period. The actual holding period depends on the owner's personal or institutional cash flow, tax, and life-cycle issues.

4. Current Issues

The real estate industry is highly cyclical. Because development can take from 1 to 5 or more years to complete a new project, depending on the difficulty of obtaining public approvals, buildings are often started when the market is hot, and completed when it is not. The real estate industry collapsed in the late 1980s–early 1990s in the deepest real estate recession since the Great Depression. The collapse was caused by a number of factors including over-building, tax policy that contributed to over-investment, too much careless money, and a major recession in the general economy due to downsizing of many corporations and national defense. Developers and investors who bought property at bargain prices in the early 1990s have made fortunes as rents and prices have recovered and surpassed pre-recession peaks. At the beginning of the twenty-first century, real estate markets in the USA are relatively 'in balance'—the supply of new space is in balance with demand for new space. Cycles, however, are endemic to real estate (Grenadier 1995). Those who learn how to play them well are the ones who will be most profitable.

Sects. 4.1–4.5 detail some of the most important issues in real estate development at the beginning of the twenty-first century.

4.1 Institutionalization of the Industry

During the recession of the late 1980s–early 1990s, many developers went bankrupt, along with the Savings and Loan Industry and many other financial institutions. Two powerful forces emerged from the recession—a vast expansion in the presence of Wall Street in real estate financing, primarily through securitization and Commercial Mortgage-Backed Securities, and the re-emergence of Real Estate Investment Trusts (REITs) as a major owner of property (Sanders 1998). REITs grew to more than \$140 billion in stock market capitalization—exceeding the 1 percent barrier of total market capitalization that represents the point at which Wall Street analysts pay serious attention to an industry. While REITs and other publicly financed companies, especially national homebuilders, will play a more important role in real

estate development and ownership than before, entrepreneurial developers will continue to play a significant if less dominant role in the creation of new product.

4.2 Future of Cities

Central Business Districts and inner-ring suburbs have gone through ups and downs over the last 50 years as many companies and higher-income residents have moved to outer-ring suburbs and smaller towns at the urban fringe. The constant outward migration has created enormous expansion of rapidly growing cities such as Atlanta, Phoenix, Dallas, Denver, Los Angeles, San Francisco, and Seattle. The outward growth has put pressure on open space and natural resources and placed demand on the public sector to build highways and other infrastructure to serve the new development.

At the same time, central cities and inner suburbs have suffered under declining municipal tax revenues and increasing demands for services from a poorer population, often augmented by new immigrants. While many cities such as Manhattan and Los Angeles are seeing signs of a return to the CBD by investors and residents, outward migration is likely to continue. Urban containment boundaries are a popular though controversial, technique for containing outward expansion. Developers have historically tended to prefer green-field suburban sites because they are easier to build, easier to finance, and more attractive to tenants and homeowners. Whether central cities and inner suburbs will be able to attract new investment and redevelopment is one of their most important challenges.

4.3 Smart Growth/Urban Sprawl

The rapid expansion of suburban growth in the last half of the twentieth century has fueled a backlash against development and *NIMBYism* (not in my backyard). 'Smart growth' is the latest approach to solving the problems associated with urban sprawl—congestion, loss of farmland, lack of functional open space, monotonous and boring development, and inadequate infrastructure (see Gordon et al. 1989, Peiser 1989).

4.4 Public–Private Development

As the regulatory process for major development projects has become more time consuming and costly, more developers are joining forces with cities, redevelopment agencies, school districts, armed forces, and other government agencies to develop property. The public–private partnerships help developers to obtain lower cost financing and necessary public approvals while giving public agencies a share in the profits and the ability to deliver more public benefits such as parks, plazas, roads, sewer plants, and other infrastructure (Kotin and Peiser 1997).

4.5 Technology

The internet and other new technologies are having dramatic impacts on all aspects of real estate, from office building construction to shopping center functionality resulting from Internet commerce. As old shopping centers and industrial buildings become obsolete faster, new industrial distribution centers and 'entertainment' shopping centers are replacing them. While home offices and telecommuting have caused alarm for future office space demand and spawned new work patterns such as 'hot desks' (where employees share office space, plugging their lap top into the company computer network as they move from office to office) workers still appear to want the 'social' aspects of coming to work. Suburban corporate campuses with high tech communications, on-site recreation and sporting facilities, and high security are being built to respond to new workplace needs and employee preferences.

5. Social Responsibility

Real estate development touches on every aspect of people's lives. It determines not only the quality of the built-form environment, but also affects the life-style and social wellbeing of residents and workers in every community. What developers build may last a lifetime or longer. Because of their far-reaching role, developers bear a special responsibility. Most developers care deeply about the quality of their projects. High quality development depends as much on public sector support and services as it does on the private activities and financial resources of the developer. Since real estate involves very long-term investment, the long-term stability of the community and continuing attractiveness to new residents and firms is essential to long-term profitability for the developer. While developers are often vilified for destroying the environment and hastening neighborhood change, good development is responsive to neighborhood concerns and is sensitive to maintaining and improving environmental conditions. As public funds have been reduced for open space infrastructure improvements, development is often the instrument through which open space is preserved and new infrastructure is built. As their role expands, developers need to be ever more knowledgeable about the urban context in which they build and about the multitude of forces with which they must deal. The Urban Land Institute's mission statement sums up the developer's charge: 'to provide responsible leadership in the use of land in order to enhance the total environment.'

See also: Development and Urbanization; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Property Rights; Urban Geography; Urban Growth Models; Urban Planning: Central City Revitalization; Urban Sprawl

Bibliography

Brueggeman W, Fisher J 1997 *Real Estate Finance and Investments*. Irwin/McGraw-Hill, New York

Case K E, Shiller, R J 1989 The efficiency of the market for single-family homes. *American Economic Review* **79**: 125–37

Corgel J, Smith H C, Ling D 1998 *Real Estate Perspectives*. Irwin/McGraw-Hill, New York

DiPasquale D, Wheaton W 1996 *Economics and Real Estate Markets*. Prentice Hall, London

Garvin A 1996 *The American City: What Works, What Doesn't*. McGraw-Hill, New York

Gordon P, Kumar A, Richardson H 1989 Influence of metropolitan structure on commuting time. *Journal of Urban Economics* **26**(2): 138–51

Grenadier S R 1995 The persistence of real estate cycles. *Journal of Real Estate Finance & Economics* **10**(2): 95–119

Kotin A, Peiser R 1997 Public-private joint ventures for high volume retailers: who benefits? *Urban Studies* **34**(12): 1971–86

Mills E S 1972 *Studies in the Structure of the Urban Economy, Resources for the Future*. The Johns Hopkins Press, Baltimore

Peiser R 1989 Density and urban sprawl. *Land Economics* **65**(3): 194–204

Peiser R 1990 Who plans America: Planners or developers? *Journal of the American Planning Association* **Autumn:** 496–503

Peiser R 1992 *Professional Real Estate Development: The ULI Guide to the Business*. The Urban Land Institute, Washington, DC and Dearborn, Chicago

Sanders A B 1998 The historical behavior of REIT returns: A capital markets perspective. Garrigan R T, Parsons F C (eds.) *Real Estate Investment Trusts*. pp. 277–305

Urban Land Institute 1997 *America's Real Estate*. Washington, DC

R. Peiser

Realism/Neorealism

Realism is a tenet that views the world as it is, and not in terms of an unrealized ideal. It emphasizes the constraints of the real world and the need for pragmatism and prudence. Applied to international politics, realism sees a conflict-ridden world of states concerned pre-eminently with their security and pursuing power as the means to assure their survival. Realists therefore see the distribution of power as the major determinant of state behavior and international stability. Within this broad rubric, realists have developed alternative formulations, emphases, and distinctions.

1. The Realist Vision

The realist vision of international politics holds that states interact in an anarchical international environment in which survival is uncertain. Because power is the key requisite for survival, states define their interests in terms of power and constantly assess their strength relative to that of others. What emerges from their interaction is a balance of power that is the sole basis for order, regularity, and stability in international politics.

Realism emphasizes the competitive and conflictual nature of international politics. States either constantly prepare for war or are at war with one another. Fearing adverse shifts in the balance of power, they respond with internal mobilization and, should that prove inadequate, with external alignment. Self-reliance is preferred in a system of self-help, but allying with others to confront a common threat may be temporarily necessary. Cooperation is thus short-lived, lasting only as long as the common threat, and is a function of common interests and not common values or ideology. Realism can therefore readily explain alliances, such as the Nazi–Soviet pact of 1939, between wholly antithetical political systems. The result, mobilization and alignment in response to adverse shifts in the balance of power, returns the system to equilibrium. Hence, balances of power recur and reemerge in international politics.

1.1 Realist Assumptions

Realism is a statist perspective, for it assumes states to be the key actors in the international system. It excludes multinational corporations, nongovernmental organizations, and a variety of transnational and subnational entities that matter only to the extent that they affect states.

The theory assumes that states face recurrent threats and challenges in the international system. Power, especially military power, is the basis of both threats to a state and to its ability to assure its survival.

Realism presumes survival and its requisites to be the highest priority of states, indeed, their very reason for being (*raison d'état*). States value security considerations above all and respond to the imperatives of the international system. The primacy of security also implies that other policies, including foreign economic ones, reflect considerations of international power politics.

1.2 Realism Deemphasizes

For realists, international law and international organizations are largely irrelevant to international politics. Both reflect the interests of powerful states, for in international politics might makes right, and international institutions do not constrain great powers but merely provide additional venues for the

exercise of power. At most, both play a role only in matters of low politics, in regulating transportation, communication, and economic exchange, for example, and not in matters of high politics, especially military security.

Realists also consider the requisites of survival to dominate any moral considerations. Indeed, realists transform necessity into a moral imperative.

Realism downplays the role of domestic politics. It treats states as unitary actors that are functionally equivalent to one another, and the constraints imposed by the international system generate similar behavior even by dissimilar states. The international system once consisted largely of monarchies and then of authoritarian and military governments, and is now increasingly composed of democratic states. For realists, however, it continues to function in the same way.

Self-reliance precludes states from allowing the development of any division of labor or any interdependence regarding the requisites of state survival. Small states may have no choice but to rely on others, but great powers do have such choices and thus do not allow the emergence of any meaningful interdependence.

1.3 Realism as Equilibrium Theory

The balance of power is an equilibrium theory. The security policies of nations result in a balance of power among states. The equilibrium can be disturbed, but only by exogenous shocks, ones outside the model. The disturbance then sets in motion forces that return the system to a new equilibrium. The implosion of the Soviet Union, for example, disturbed the balance of international power, but was an event outside the model and inexplicable by it. But realism does purport to explain the set of forces that then would come into play to return the system to equilibrium.

Put this way, balance of power theory functions like Adam Smith's invisible hand in creating an unintended outcome. For Smith, the self-interested behavior of producers in a competitive market generated the best price for consumers. In international politics, the pursuit of national interest and the struggle for power in an anarchical system create a peaceful stable order. It is not the search for peace but the search for security that creates this stability.

1.4 Realism—its Historical Roots

Realism traces its lineage to Thucydides, who argued that his *History of the Peloponnesian War* (1996) provided the one and true explanation of the war: the growth of Athenian power and the fear this struck in Sparta. Thucydides argued that preceding incidents and specific complaints notwithstanding, the true cause was a shift in the balance of power.

As important to current realist formulations is the work of Thomas Hobbes, who translated Thucydides into English and argued (in *Leviathan*) that the state of nature, which he characterized as a war of all against all, is one in which war and preparation for war are a constant. Conflict is limited within societies by the creation of the Leviathan, a coercive state. For modern realists, who distinguish international from domestic politics, international relations remains an anarchic realm; there is no Leviathan and the war of all against all endures.

Realism dominated the study of international relations in the middle of the twentieth century as the world came to grips with the very occurrence of the Second World War and the failure of both international law and international organizations to prevent it. Hans Morgenthau's (1948) influential formulation of realism drew upon psychology and rooted realism in a human nature presumed to be inherently aggressive and in a view of individuals as power seeking.

Beginning in the 1950s and increasingly in the 1970s, realists (led by Waltz 1959, 1979) linked their arguments to systems theory and structuralism. This variant rejected basing realism in any theory of human nature or any internalized commitment on a balance of power. This neorealist formulation did not posit any individual or collective propensity toward power maximization *per se*. Rather, it posited a minimalist concern with survival in an anarchic setting as the only necessary conditions. Even good people and good states caught in an anarchic structure have no choice but to arm themselves in preparation for the possibility of war and must assiduously guard their relative power. Neorealism or structural realism firmly rooted the problem of international politics in the international system and not in the nature of individuals or states.

2. Disagreements

Despite the consensus in emphasizing states as the central actors, security as their prime concern, and power as their key instrumentality, there is some confusion and disagreement in the core of realist thought. First, there is confusion whether the theory is positive or normative. Realists typically argue that theirs is a positive theory of how the system actually works (an iron law of international politics), yet many have also used realist precepts to criticize the foreign policies of specific countries. Realists have argued, for example, that states do only what is in their national interest but have criticized specific actions, such as US intervention in Vietnam, as not being in the national interest.

That politicians adopted the argument has also led to confusion. First, the term balance of power has come to be used loosely and applied to virtually every configuration of power. Second, there is a disagree-

ment between scholars as to whether realism operates through intention (states must aim for a balance of power) or irrespective of it (as an unintended consequence).

Despite agreement that the distribution of power determines stability, scholars disagree about the relative stability of different configurations of power. Within the balance of power tradition, some scholars find bipolarity more stable, others find multipolar systems more stable. Although most agree that the international system tends towards balance (though more predictably and more quickly in some settings than others) and that hegemony and unipolarity are not possible or sustainable, some hold that the system tends towards power preponderance rather than balance, and that hegemony is stable but that a balance of power is not.

3. *Problems*

Realism is criticized for leaving too much out. A pristine realism that only says that balances recur cannot explain most foreign policies of most states most of the time. Moreover, the theory is static and cannot explain change over time. Important transformations of recent centuries, such as the rise of nationalism, the industrial revolution, the rise of ideological politics, the spread of democracy, the revolutions in communication, transportation, and warfare, realists judge irrelevant to international politics, which they treat as unchanged as long as anarchy continues to characterize international politics. Realists deem these factors to matter only to the extent that they affect the relative power of states (whose alteration is the only source of change in international relations) and are viewed as exogenous shocks. The only transformation that some realists accept as affecting world politics is the nuclear revolution, which they consider to undercut fundamentally the importance of relative power in a classical military sense.

The theory is also vague and imprecise. It is not clear what factors enter into a calculation of the balance of power, nor how balance is ascertained. It is difficult to test assertions of the theory when there can be wide disagreement over whether a particular distribution of power is balanced or not. Scholars dispute whether stability during the Cold War derived from US military superiority, a balance of power between the US and USSR, or the existence of nuclear weapons.

More important, the theory is indeterminate. It cannot predict which particular distribution of power will emerge, only that balance will. Nor can it predict the process by which balance is restored. US containment of the Soviet Union during the Cold War can be explained as the working of the balance of power, but so can the US roll back of German and Japanese power during World War II. A US decision simply to contain German and Japanese power in the 1940s would have been as consistent with realism as the actual US decision to wage war and undo the consequences of German and Japanese aggression.

The theory makes precise predictions only when the constraints of the international system are presumed to be tight. Although the theory holds that it is great powers that structure global politics, the international system is most constraining of small powers and least constraining of large ones. Hence, the theory is typically indeterminate regarding the behavior of the great powers, which are presumed to matter most for international politics.

Formal studies have shown that realist premises are consistent with international cooperation, the construction of international organizations, and even the emergence of international norms. These works challenge the deductive logic of realism, arguing that the premises are insufficient to lead to realist conclusions.

Realists have had some difficulty in recent years in explaining significant international developments. The disjuncture between military and economic power has been sustained for an extended period of time. During the later stages of the Cold War, the world remained militarily bipolar, dominated by the might of the United States and the Soviet Union. Yet economically the world was multipolar, with Germany (and the EU more broadly) and Japan as significant economic powers that did not translate their economic prowess into military strength. Such sustained disjuncture between the distribution of economic and military power is problematic for realism.

4. *Modifications*

Realist thought continues to develop, in part to deal with such anomalies. One variant now posits that it is the perception of power rather than the actual distribution of power that matters. Another decouples changes in the distribution of power from the perception of threat, and holds that states respond to threats and not just shifts in power. Some see realism as characterizing particular eras rather than international politics generally. Eras in which territorial acquisition brings greater rewards than does commercial exchange are ones to which the logic of realism is more applicable; eras in which the nature of warfare gives the offense an edge are ones in which a realist logic is more likely to prevail.

Realism has a core strength: its emphasis on power and interest and an unregulated international system. Yet its most pristine formulation is rejected by many, if not most, scholars who articulate their positions relative to a core realist argument.

See also: Balance of Power, History of; Balance of Power: Political; Cold War, The; Conflict and War,

Archaeology of; Conflict Sociology; Geopolitics; Globalization: Political Aspects; National Security Studies and War Potential of Nations; Political Science: Overview; State, History of

Bibliography

Baldwin D A (ed.) 1993 *Neorealism and Neoliberalism: The Contemporary Debate*. Columbia University Press, New York

Brown M E, Lynn-Jones S M, Miller S E (eds.) 1995 *The Perils of Anarchy: Contemporary Realism and International Security*. MIT Press, Cambridge, MA

Doyle M W 1997 *Ways of War and Peace: Realism, Liberalism, and Socialism*. Norton, New York

Frankel B (ed.) 1996 *Realism: Restatements and Renewal*. Cass, London

Jervis R 1978 Cooperation under the security dilemma. *World Politics* **30**: 167–214

Jervis R 1997 *System Effects: Complexity in Political and Social Life*. Princeton University Press, Princeton, NJ

Keohane R O (ed.) 1986 *Neorealism and its Critics*. Columbia University Press, New York

Morgenthau H J 1948 *Politics Among Nations: The Struggle for Power and Peace*. Knopf, New York

Stein A A 1990 *Why Nations Cooperate: Circumstance and Choice in International Relations*. Cornell University Press, Ithaca, NY

Thucydides 1996 *The Landmark Thucydides: A Comprehensive Guide to the Peloponnesian War*. [trans. Crawley G; Strassler R B (ed.)]. Free Press, New York

Waltz K N 1959 *Man, the State and War: A Theoretical Analysis*. Columbia University Press, New York

Waltz K N 1979 *Theory of International Politics*. Addison-Wesley, Menlo Park, CA

A. Stein

Realisms and their Opponents: Philosophical Aspects

In everyday usage, 'realism' is often used as a name for a practically or epistemically low-ambition attitude, while 'idealism' is often taken to denote a high-ambition—if not utopian—attitude. In philosophcal usage, mostly, it is the other way around: those who are called realists tend to claim more than their opponents—they are the philosophical optimists.

Within philosophy itself, 'realism' adopts a variety of interrelated and contested meanings. It is used as the name for doctrines about issues such as perceptual access to reality, the existence of universals, the goals and achievements of science, the nature of truth, the objectivity of morality, and many other things. Given this variety, no single shorthand definition of the term 'realism' can be provided.

One manifestation of this variety is that the opponents of realism about these issues are not called uniformly by a single label (other than the rather uninformative 'antirealism'): the labels used include idealism, phenomenalism, instrumentalism, conventionalism, fictionalism, noncognitivism, constructivism, relativism, irrealism, and others. Sometimes such labels vary from one dispute or domain to another even when similar ideas are being expressed. Sometimes they are used to make claims that are specific to the domain at hand and distinct from other antirealist theses.

In what follows a brief tour will be taken along some of the representative philosophical highways and lanes, from ontology through semantics to epistemology. Along the way, realism will be confronted with some of its various opponents and internal conflicts. The philosophical landscape in the neighborhood of realism is nowadays much broader than it used to be: in a typical older encyclopedia, 'realism' was taken to name doctrines about universals or about perception, or both, to the exclusion of much of its present coverage, and many of the current opponents were missing from view altogether.

1. Ontology

As an ontological doctrine, realism is a claim about the existence of something. Versions of ontological realism differ from one another in regard to what this something is—the domain of existents (things that exist)—and what kind of claim is being made about items in this domain. If we take ontological realism to have the form of the statement '*X* exists,' then in order to identify different versions of realism and its denial we need to examine the various ways in which its two components are interpreted. We need to consider the two questions: 'what exists?' and 'what is existence?'

If we just claim that the world exists, without providing further specifications of the constituents and structure of the world, our realism will remain rather uninteresting. Philosophers of the nominalist persuasion add that particulars, and nothing but particulars in time and space, exist. The world consists of objects such as you and me and the particular copy of this encyclopedia that you are reading now. A radical nominalist will claim that only 'bare' or propertyless particulars exist, while properties do not. Whereas nominalists can be realists of sorts, they are opposed by those who have been called the realists in the debate over the existence of universals. In this debate, realists claim that universals—kinds, properties, relations—exist, that they are, or are among, the constituents of the world. Not only does this particular encylopedia exist, but among the furniture of the world are books as a cultural kind, properties such as weight and colour, and the relation of being owned by someone. Such universals can be multiply instantiated in particular objects in space and time. Platonic realism takes universals to constitute reality: redness, 'book-

ness' and ownership inhabit a separate realm of ideas, independent of the particulars that exemplify them. Aristotelean realism takes both universals and particulars to be real, and claims that universals only exist when exemplified in particulars. In the philosophy of mathematics, the view that numbers, real-valued functions, and sets exist as 'abstract entities' is called Platonism, while the opponents denying their existence are known as intuitionists, formalists, and logicists.

Common sense takes the world to be furnished with medium-sized material objects that are the possible objects of sense perception. Considering these objects to exist leads to realist doctrines about the ontology of perception. The opponents of realism in the theory of perception include the idealists (to whom it is 'ideas' that constitute the world) and phenomenalists (who try to construe the world out of 'sense data'). *Direct realism* says that perception is a direct awareness of material objects that exist in reality and that nothing else exists between perception and perceptible objects. Naive realism, entertained by philosophers as a target of criticism, is a version of direct realism that takes sensible qualities to be the intrinsic properties of material objects and these objects to have all the properties they are perceived as having. *Indirect realism* states that perception is directly about mental representations (such as bodily sensations and afterimages) and only indirectly about the external world and that both the direct and indirect objects of perception exist.

An encompassing commonsense realism takes objects in the commonsense domain, whether material or mental, to exist. Its list of existents includes my bathroom and my belief that it is vacant, as well as ducks and the desire to eat one. Anyone ontologically committed to folk psychology, holding the commonsense view that beliefs, desires, hopes, and fears exist, is a realist about the mental. Common sense also seems to suggest that there are moral facts about the world. To say that something is morally good or morally bad is to describe aspects of objective reality. Moral realists deny a sharp distinction between facts and values, holding that moral values are among the facts of commonsense experience. In contrast to these views, an eliminative materialist denies that mental and moral facts exist and takes only matter or physical objects to exist, thus counting as a realist about the material but not about the mental or the moral.

Scientific theories characteristically discuss unobservable entities such as quarks and electromagnetic fields. If we take these to exist, we hold the ontological position of *scientific realism*. The opponents of realism about the ontology of scientific theories are the phenomenalists and fictionalists. For them, unobservables do not exist—only observables do. Rather than being among the fully authorized inhabitants of the real world, the unobservables of science are nothing but fictions of imagination in the scientists' minds.

One may combine scientific realism with a commonsense realism that takes both scientific objects and the objects of folk physics, folk psychology, morality, and folk sociology to exist: the full set of existents includes plans and plaintiveness, plantations and plebeians, plants and planets, platelets and plastids. This is disputed by a radical scientific realism that holds that only scientific objects exist; thus, trees do not exist, but the atomic and cellular structures that constitute them do. A radical commonsense realist will deny existence to scientific objects while granting existence to commonsense objects: plants exist while plastids do not. One can also be more discriminate among the various classes of commonsense objects, granting or denying existence to material, mental, and moral facts in different permutations, for example, holding that only the material exists; or that the material and the mental exist; or that all three exist.

One's notion of scientific realism has to be qualified in relation to specific scientific disciplines. The standard formulations are in terms of unobservable entities that lie far beyond the commonsense domain, with physics serving as the paradigm science. But suppose one wants to be a scientific realist about the explanatory posits of a sociological theory that discusses community and custom, role and authority, motive and reciprocity? If one takes these commonsense items to exist, the scientific realism that ensues will not be strictly modeled after the realism about photons and neutrinos and other noncommonsense posits of physical theories.

One can also choose to be a realist about modality. Modal realists argue that science and common sense deal with possibilities and necessities as aspects of the world. In decision theory, the decision-maker is viewed as considering possible choices and possible outcomes with probability values. In experimental science, the experimentalist considers various possible designs and possible outcomes with estimated probabilities. A modal realist takes such possibilities as real. An actualist denies the existence of possibilities and only regards actual things in a realist fashion. Yet, among the actual existents (in contrast to merely possible), one may include dispositions, propensities, capacities, and causal powers: they are potentialities that, under certain conditions, manifest themselves.

Having thus listed some possible domains of existents, and having made a selection from among them, the realist might leave it there. Of some X, the realist about X just says that X exists. Yet, pressed by antirealist arguments, the realist may feel forced to take a few more steps. This is because antirealists can go along with the realist to accept that X exists, but just interpret the meaning of 'X exists' differently. Thus the realist has to raise the further question of what is claimed about the existence of the existents, or what is meant by 'X exists'.

This question has a number of variants. One is about whether one reads existence claims literally or

reductively. A phenomenalist may say that electrons and elm trees exist, and add that what she means to say is that certain constellations of sense data exist; claims about electrons and elm trees and other such entities can be reduced to claims about sensory experiences. A realist would say that electrons and elm trees exist, period—signaling that the claim is to be understood literally. An eliminative materialist may say that a given mental property exists, and then add that this is just another way of saying that a certain neurophysiological configuration exists. A mental realist would resist such a reduction. In general, when considering the existence of *X*, the issue is whether the statement '*X* exists' is to be taken literally or to be reduced to another claim: '*Y* exists.'

Once we have fixed a literal answer to the question of what is claimed about the existence of the existents, another version of the question can be raised. This concerns the possibility and actuality of existence. The weakest realist claim is that it is *possible* for *X* to exist: the claim that *X* exists is intelligible and coherent. One is a realist about *X* not by claiming that *X* actually exists, but by claiming that *X* is the sort of thing that might exist. This is consistent with having doubts about its actual existence or a later discovery that it does not exist after all.

Many realists make stronger claims than just possible existence. They include statements about *actual* existence in their definition of realism. There are different versions of this, depending on the quantifier that is used in relation to the sets of items that are claimed to exist. No realist would like to claim that *all* alleged universals, particulars, commonsense objects, and/or scientific objects exist (this would imply that unicorns and ether are as real as magnolia trees and DNA molecules). Many realists would say that *some* of them, or *most* of them, exist.

Another version of the question (of what is claimed about the existence of existents) arises as we recognize that such claims are customarily formulated in terms of independence. Such claims take on the form, '*X* exists independently of *Z*'. Two further questions arise from such a claim: 'what is *Z*?' and 'what is independence?'

The ordinary formulation is that something exists mind-independently or independently of the mental, with 'mind' and 'mental' designating facts about the human mind. The paradigm meaning of 'exists' is based on an image of physical reality that existed before there were any humans around. Physical particles and their systems, such as molecules and galaxies as well as natural laws, existed independently of human perception and thought, and now that humans are around, these things continue to so exist—and were human life to disappear, they would still exist. One issue is whether 'exists' in this paradigm meaning applies to anything at all. Another issue is whether it is appropriately general to be extended to cover all important cases of existence claims, including those about things that have emerged along with or since the arrival of humankind into the world. But what is independence?

If we take it to mean logical independence, we may want to say that physical objects are logically mind-independent since statements about them do not imply the existence of any human minds. This will require reading '*X*' literally as denoting physical objects: otherwise it would be open to a nonrealist to argue that *X*'s existence is logically mind-dependent for the simple reason that '*X*' is just another name for a bundle of sense data or some such. But if '*X*' denotes a human mind or a mental property itself, we cannot say that its existence is logically independent of the mental, because it entails the existence of the mental.

We may also say that something exists in the sense that it is causally independent of the mental: the human mind does not causally produce or reproduce it. Were this to apply to physical objects, 'causal production' itself would have to be literally understood as denoting a real connection in the world rather than a function of our minds. Whether it applies to human minds themselves depends on whether we take the reproduction of the species of sentient creatures like us to be causally mind-dependent activity. But it certainly does not apply to the many things in the world that are causally mind-dependent, such as material artifacts and social institutions.

Other notions of independence may be more successful in taking care of logically or causally mind-dependent things. We may choose to say that *X* exists independently of its recognition or any evidence we may have for its existence, or that it exists independently of any particular inquiry addressing it, or that it exists independently of any particular linguistic or other representation of it. Such formulations may enable realism about mental matters as well as about artifacts and many social entities that are causally dependent on the mental. The question of whether a given mental property or social entity exists is an objective question in the sense that it is separate from the question of whether anybody has thoughts, beliefs, or representations about it, has inquired into it, or has evidence for it.

Many realists hold the view that the world has a definite structure independently of our conceptual accounts of it. One may think that the world consists of propertied and related objects, such as natural kinds that are linked with one another by causal relations, and that these causal relations have a specific character, such as involving causal powers, causal transmission, and natural necessity. A nominalist would object to this by arguing that the world itself has no intrinsic structure, even though it may exist as an unstructured collection of propertyless particulars. Whatever structure the world is taken to have, it is rather our conceptualizations that impose it upon the world. The world is conceptually constructed so as to grant it a structure, and the origins of that structure

are in our ideas and conversations. This gives us an idealist and constructivist view on this matter. Thinkers are capable of 'worldmaking' as Nelson Goodman's 'irrealism' puts it, or of the 'social construction of reality' as the radical wing of social constructivism alleges. Insofar as our ideas are permitted to legitimately vary, this results in a form of ontological relativism as well: to each different communally-held or justified conceptualization corresponds a different structure of the world. Thomas Kuhn's suggestion that after a scientific revolution from one paradigm to another, scientists 'live in a different world,' can be—contestably—read as implying this idea.

The social world poses a special challenge to the realist: the social world *is* constructed and *is* dependent on people's ideas. There can be no communities and conventions, authorities and audiences without people acting on their ideas. Yet the realist wants to say they exist. These things may be taken to exist independently of, say, scientific theories about them, in the sense that the construction of a theory about a social entity is not a matter of constructing the entity.

2. *Semantics*

Ontological realism is about the existence of *X*. One may also raise questions about linguistic expressions '*X*' and '*X* is such-and-such' in relation to *X*. Indeed, 'realism' has become increasingly defined, especially by many antirealists, as a name for certain semantic views concerning such things as reference, meaning, and truth. It is characteristic of many antirealists to take semantic issues pertaining to language as primary, whereas realists often give priority to ontology and view semantic theses as derivative of, or motivated by, ontological positions.

The claim that scientific theories and the terms they include refer to real existents is part of the semantic thesis of scientific realism. Another part is the view that at least some of the major sentences contained in scientific theories can be used to make genuine statements about the real world, and as such, they may be true or may be false. More generally, semantic realism about some subject matter or domain maintains that statements about that domain are capable of being true or false—that they are capable of describing facts about the domain truly or falsely.

Some philosophers, such as Michael Dummett, have taken bivalence as a defining notion of realism: one is a realist about a statement or theory if one takes it to be either true or false. The ontological ramification of this is as follows: if there were gaps between these truth-values, this would indicate that the world itself is somehow indeterminate. If 'N is bald' is not clearly true or false but somewhere in between, then it would seem there is no determinate fact of the matter about N's baldness. This may be taken to motivate antirealism. The realist should not reason along these lines. If bivalence has a difficulty with statements like 'N is bald,' this may be nothing but an indication of the vagueness of the language of baldness, not a matter of some indeterminacy about N's head. For a realist, the properties of language have no implications for the properties of reality in such cases.

Realist theses about possible reference and possible truth in scientific theories are often complemented with claims about actual properties of actual science. One such claim—made by Richard Boyd, Michael Devitt and others—is that the theoretical terms of most current (or the best of 'mature') scientific theories typically refer and that their lawlike statements are at least approximately true. Another related claim is the convergence thesis: as science develops, its theories get progressively closer to the truth. Both of these are empirical claims about actual science and should not be made part of the concept of scientific realism. Their truth is dependent on contingent matters such as the institutional structure and other resources of scientific research as these happen to be in any given society and time period. This is not to deny that a realist may need a concept of approximate truth, truth-likeness, or closeness to the truth in order to defend the possibility of truth acquisition in science. Nor is it to deny the obvious idea that the notion of convergence towards the truth may be used to give content to the notion of scientific progress.

As for the concept of truth itself, a realist will hold the view that truth transcends any conceivable evidence. In this sense, truth is characterized negatively as nonepistemic, as being independent of whether, and how, we may have epistemic access to it. This runs counter to neopragmatist suggestions to define truth as provability (as does Dummett) or ideal acceptability (as does Hilary Putnam). For a realist, truth is one thing, and it is quite another to prove it, discover it, confirm our belief in it, or find evidential warrants for our beliefs. Since there is no necessary connection between truth and these things, the realist thinks that evidence-transcendent truth is not necessarily attainable. This implies that even if the epistemic warrants for a statement or theory were as good as possible, or in some sense ideal, our theory might be mistaken.

'Realism' is also used in connection with–but is not defined by—a specific view of what truth is: the correspondence account of truth. This view allows for many alternative formulations. A generic formulation is to say that a sentence (or proposition, or utterance, or whatever the appropriate truth-bearer is taken to be) is true in virtue of its correspondence with certain facts about the world, the truth-makers. This can be linked with ontological matters. On the one hand, truth presupposes an ontology within which truth-makers find a place. On the other hand, there is the route from an account of truth to ontology, from claims of truth to claims of existence. If it is up to the truth-makers to make the truth-bearers true, and if the truth-makers are objective facts about the world, then

the truth-makers of a true truth-bearer must exist. Thus, if mathematical statements can be true, then there must be mathematical objects of which they are true. If moral statements can be true, then there must be objective moral facts that make them so. Some realists have severed the connection between realism and the correspondence account of truth: they want to endorse the former without endorsing the latter. Most of these realists hold a deflationary or minimalist account of truth.

Nonrealists can also choose from different options. An instrumentalist about scientific theories can insist that no truth-values can be ascribed to theories as they make no claims about the world. Theories are no more than instruments for organizing the data, predicting future events, or helping to manipulate the world. They are tools of management or tickets of inference, and it is in the nature of tools that one should not ascribe truth or falsehood to them. Tools are characterized as useful or useless, convenient or inconvenient, relevant or irrelevant, powerful or powerless, and so on—all these attributions being relative to the purpose to which a given theory is put. This is the traditional version of instrumentalism that has been inspired by the familiar feature of physical theories that they appear to postulate unobservable entities. In case one has empiricist suspicions about such unobservables, it is tempting to construe theories as neither referring to them nor making any claims about them, whether true or false.

Likewise, in case one is ontologically suspicious about intentional properties, there is still an option to use the intentional vocabulary of folk psychology instrumentally: even though there really are no intentional properties on top of the physical ones, folk psychology can be usefully employed for predictive purposes provided it is construed nonreferentially. Daniel Dennett's 'intentional stance' serves as an example. One may adopt this stance as a perspective from which to treat systems *as* intentional without asserting that they really *are* intentional. No truth claims are being made about a system *really* possessing beliefs and intentions, but the predictive benefits created or lost by treating any given system as intentional is an objective matter.

Another instrumentalist possibility is to view a theory as a false tool, rather than as a truth-valueless tool. This is in line with Hans Vaihinger's fictionalism: falsehood is a characteristic of scientific theories regarded as fictions. Even though Vaihinger did not discriminate between natural and social sciences in applying his fictionalism—Isaac Newton's gravity and Adam Smith's economic man served as his examples—this view may be able to accommodate a feature of many social scientific theories: they do not seem to postulate unobservables in the same sense in which one finds them in physical sciences. Many social theories, and especially economic theories, involve idealizations and simplifications about familiar commonsense objects: immensely informed consumers with transitive preferences over all options of choice, and international economies with two countries with identical production functions, two goods, two fully-employed and internationally-immobile factors of production. Since many such idealizations are wide of the mark, one may be tempted to construe theories involving them as false, but nevertheless useful, instruments.

A popular view among the opponents of moral realism is to claim that apparent statements about some domain are not really genuine statements at all, but that they are rather used to express emotions or commitments or to persuade audiences. Thus, one may argue that moral claims of the form, '*X* is morally good' are not statements of some objective moral facts capable of being true or false, but express the utterer's emotions or persuasive intentions. This nonrealist position is called a noncognitivist position in contrast to its realist opponent, the cognitivist view, which takes moral claims to be truth-valued. This is an example of a reductive, nonliteral account: a descriptive claim that '*X* is good' is not supposed to be taken literally, but rather to be reduced to recommendations or approvals.

Yet another nonrealist possibility is to retain the vocabulary of truth and refute the correspondence account. Truth can be viewed as essentially involving pragmatic features, such as instrumental efficiency, ideal consensus, or warranted assertability. Whatever works for this or that purpose and meets certain constraints is true. These views make truth dependent on things such as our interests and goals, our thoughts and recognitional capabilities, our linguistic and other social conventions. Realists tend to think of truth as independent of such things.

Radical relativism regards all truths as indexical: '*P* is true in culture or community *C* but not in *D*,' '*P* is true for me but not for her,' and so on. Some formulations of constructivist sociology of scientific knowledge result in relativism of this sort by equating truth with acceptability within a scientific community. Equating truth with *ideal* acceptability is supposed to avoid the pitfalls of such relativism: the ideal conditions are supposed to fix truth uniquely. Realists claim to avoid them by not equating truth with anything epistemic.

3. *Epistemology*

Many of the issues above can be turned into issues of belief, justification, and knowledge. Some formulations of scientific realism refer to epistemic attitudes: scientists accepting a theory believe it to be true; and it is the goal of science to pursue true theories. Both formulations are problematic. A scientist with ontologically and semantically realist convictions may, with no inconsistency, accept for some purposes a

theory not believed to be true or not yet so believed. And whether or not truth is actually pursued as a goal is not implied by the notion that scientific theories include sentences that are capable of being true or false—or that their posits are capable of existence. It is also questionable whether science is the sort of entity to which goals can be attributed. Whether truth is among the goals of a particular body of scientists, and whether science as a collective endeavor will actually deliver truths as an outcome, depends on the actual prevailing institutional organization of science. Claims about scientists' actual beliefs, goals, and achievements are empirical claims whose truth depends on various contingencies; they should not be included in one's core definition of 'realism'. Of course, to accommodate such ideas, one may stipulate complex labels such as 'axiological veristic scientific realism,' which says that scientists had better pursue true accounts of the world.

Forms of epistemological realism are best formulated in terms of 'knowability'—of believability and justifiability. Supposing the world, and these or those of its constituents exist, can we acquire knowledge about them? Realist theories of perception share the view that there is no veil of perception, that perception is able to provide us with (perhaps very complex and indirect) access to at least some parts of reality. More generally, an epistemological realist holds that the world is knowable; for example, the theoretical claims of science may be justifiably believed to be (approximately) true.

With respect to scientific theories and the entities and relations they postulate, one may hold the view that the ontological question of their existence is irrelevant. They may exist, and theories may be true or false about them, but what is crucial is that we *cannot know* that they exist, nor that claims about them are true or are false. Therefore all that matters for science—insofar as its goals and criteria of success are concerned—is empirical adequacy, that is, the truth of the claims science makes about observable phenomena. This empiricist combination of ontological realism and epistemological antirealism has a long tradition, including Auguste Comte's positivism and, presently, Bas van Fraassen's constructive empiricism. People like Ernst Mach have been closer to taking the more radical step from epistemological to ontological antirealism, from denying epistemic access to the unobservable posits of science to denying their existence.

Many contemporary versions of antirealism, including Putnam's 'internal realism,' hold that insofar as truth is concerned, the semantic and epistemological issues cannot be kept distinct: truth is epistemic. Truth consists in warranted assertability, or in ideal acceptability. This means that a theory that is epistemically ideal cannot be false. An epistemically ideal theory is a fully justified theory: it is a product of our very best efforts in the most favorable circumstances, one that is perfectly supported by all available evidence and meets all relevant theoretical constraints. While the realist will distinguish between the notions of being true and passing for truth in any circumstances, the antirealist will not acknowledge a difference between truth and passing for truth in ideal circumstances. The antirealist view makes truth conditions identical to the conditions of the recognition of truth, thus there is no recognition-independent or evidence-transcendent truth to be sought. Truth becomes epistemic; semantics and epistemology become fused. This may be taken to have ontological import. If an ideal theory suggests that quarks exist, then the statement 'quarks exist' is true, and therefore quarks exist. In this picture, ontic matters are derivative from epistemic matters.

This antirealist view is often presented as providing protection against skepticism about truth acquisition: making truth attainable by definition is supposed to help avoid the misery of truth possibly escaping us in all circumstances. Semantic realism decouples truth and justification by viewing truth as evidence-transcendent and otherwise nonepistemic. This may be taken to create room for skepticism: since even epistemically ideal theories may be false, we may be doomed to be mistaken forever, even though most realists would say that, as a matter of fact, we are not. Indeed, epistemological realists believe that we are able to acquire justified true beliefs about the world, thus they deny skepticism. Whether one formulates scientific realism as suggesting that most (or some) current scientific theories are recognizably and justifiably (approximately) true, or as the implied weaker view that it is in the character of scientific theories that they can be recognized as being true or false, one thereby subscribes to epistemological realism. Yet, given that truth and justification have been conceptually decoupled, the realist has to face the difficult problem of showing how these two things are related empirically so as to ensure that our cognitive efforts are somehow conducive to truth-acquisition.

The antirealist who holds truth to be epistemic is not necessarily much better equipped against skepticism. While actual justification will not guarantee truth, ideal justification does, or so some antirealists suggest. The challenge is to give a detailed account of what ideal justification consists of, and to do this in purely epistemic terms without invoking a prior grasp of the notion of truth itself. This may be difficult or impossible. If we say that ideal justification takes place in circumstances where all sources of error have been removed, and take this to mean circumstances where there are no obstacles to the attainment of truth, we will have failed to meet the challenge.

Bibliography

Armstrong D M 1978 *Universals and Scientific Realism*, Vols. I-II. Cambridge University Press, Cambridge, UK

Cohen R (ed.) 1996 *Realism and Anti-realism in the Philosophy of Science*. Kluwer, Dordrecht, The Netherlands
Cornman J W 1975 *Perception, Common Sense, and Science*. Yale University Press, New Haven, CT
Devitt M 1984 *Realism and Truth*. 2nd edn. Princeton University Press, Princeton, NJ
Hacking I 1983 *Representing and Intervening*. Cambridge University Press, Cambridge, UK
Haldane J, Wright C (eds.) 1993 *Reality, Representation, and Projection*. Oxford University Press, Oxford, UK
Hooker C A 1987 *A Realistic Theory of Science*. SUNY Press, Albany, NY
Leplin J (ed.) 1984 *Scientific Realism*. University of California Press, Berkeley, CA
Maddy P 1990 *Realism in Mathematics*. Clarendon Press, Oxford, UK
McCormick P J 1996 *Starmaking. Realism, Anti-realism, and Irrealism*. MIT Press, Cambridge, MA
Niiniluoto I 1999 *Critical Scientific Realism*. Clarendon Press, Oxford, UK
Psillos S 1999 *Scientific Realism. How Science Tracks Truth*. Routledge, London
Putnam H 1981 *Reason, Truth and History*. Cambridge University Press, Cambridge, UK
Sayre-McCord G 1988 *Essays on Moral Realism*. Cornell University Press, Ithaca, NY
Sellars W 1963 *Science, Perception, and Reality*. Routledge, London
van Fraassen B 1980 *The Scientific Image*. Clarendon Press, Oxford, UK

U. Mäki

Reasoning with Mental Models

Reasoning is the mental process of drawing a conclusion from a set of premises. The premises may be statements, perceptions, beliefs, or items of general knowledge. The conclusion may be a statement or a thought that guides action. Reasoning is a central component of human intelligence, and without it, there would be no laws, civilization, or science (see *Scientific Reasoning and Discovery, Cognitive Psychology of*). Laws would have no application because individuals would be unable to make the following sort of inference:

Pat has a license or else Pat is disqualified from driving;
Pat does not have a license;
∴ Pat is disqualified from driving.

This inference is a valid deduction, that is, its conclusion must be true given that its premises are true. Logic is the science of valid deductions, but not a theory of how people reason. Such theories are the province of psychology, and psychologists have proposed a variety of theories (see *Logic and Cognition, Psychology of*). Human beings can reason about topics for which they have no general knowledge, and theories of this ability fall into two main categories. Some theorists postulate that reasoners rely unconsciously on formal rules of inference akin to those of a logical calculus (see *Natural Concepts, Psychology of*). The preceding inference depends on a rule of the form: A or else B, not A, therefore B. Other theorists propose that reasoners rely instead on their grasp of meaning and of principles akin to those for the semantics of a logical calculus. These theories rely on mental models, which are internal representations mirroring the structure of the external world. This article describes their role in reasoning and evidence corroborating it.

1. Mental Models

The idea that humans construct models of the external world goes back to the Scottish psychologist, Craik (see also *Mental Models, Psychology of*). He wrote:

> If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it (Craik 1943, Chap. 5).

Another antecedent is Wittgenstein's (1922) thesis that propositions represent reality in a similar way to pictures. What the modern theory of mental models adds to these programmatic proposals are three main assumptions:

(a) Each model represents a possibility. Thus, the assertion: either Pat has a license or else Pat is disqualified from driving calls for two models to represent the two possibilities (shown here in simplified form on separate horizontal lines):

License
Disqualified

where 'License' denotes a model of the possibility in which Pat has a license, and 'Disqualified' denotes a model of the possibility in which Pat is disqualified from driving.

(b) Models have a rich internal structure (not shown in the preceding diagram). Like an architect's plan, the parts of a model correspond to the parts of what it represents, and so the structure of a model corresponds to the structure of the world (Wittgenstein 1922, Propositions 2.13–2.17). Visual images are based on models, though many models are not visualizable (see *Imagery versus Propositional Reasoning*).

(c) The principle of truth: models normally represent what is true according to the premises, but not what is false. Hence, the preceding models of the disjunction represent only the possibilities that are true. Likewise, for true possibilities, models represent clauses in premises only when they are true in the possibility. For instance, the first model in the set above represents the possibility that Pat has a license,

but it does not represent explicitly that in this case it is false that Pat is disqualified from driving; likewise, the second model does not represent explicitly that in this case it is false that Pat has a license. Individuals make 'mental footnotes' about these matters, but normally they soon forget them. The principle of truth makes for parsimonious representations: reasoners do not have to bother with what is false.

Models are constructed as a result of perception, the comprehension of discourse, and imagination (see *Vision, Psychology of; Situation Model: Psychological*). They can represent real, hypothetical, or imaginary situations. They may also reside in long-term memory as a representation of knowledge (see *Mental Models, Psychology of*). They are accordingly a form of mental representation advocated by many psychologists as the way in which the mind represents reality, conceives alternative possibilities, and simulates the workings of the world. In principle, they could also underlie reasoning.

2. *Deduction with Mental Models*

The model theory of reasoning is due originally to the present author and his colleagues, notably the Irish psychologist Ruth Byrne (see Johnson-Laird and Byrne 1991, and for a recent alternative, Polk and Newell 1995). The theory postulates that models are based on the meaning of each premise, and that reasoners formulate a conclusion by describing something in the models that was not explicit in the premises. A conclusion that holds in all the models is necessary given the premises. A conclusion that holds in at least one model is possible given the premises. If models represent equiprobable possibilities—as reasoners often assume—the probability of a conclusion depends on the proportion of models in which it holds, as shown by the Italian psychologists Paolo Legrenzi, Maria Legrenzi, and Vittorio Girotto, and the French psychologist Jean-Paul Caverni (see Johnson-Laird et al. 1999). The model theory accordingly unifies logical reasoning, reasoning about possibilities, and probabilistic reasoning.

To illustrate the theory, consider again the inference:

Pat has a license or else Pat is disqualified from driving.

Pat does not have a license.

Therefore Pat is disqualified from driving.

The disjunctive premise calls for the models above. The second premise eliminates the model representing the possibility that Pat has a license. The remaining model is:

Disqualified

It yields the conclusion that Pat is disqualified from driving. This conclusion is a valid deduction because it holds in all the models—in this case, the single model—of the premises.

If individuals reason with models, then the more models that they have to construct—the more possibilities they have to envisage—the harder the task should be. Many experiments have corroborated this prediction (for a review including the studies described in this section, see Johnson-Laird 1999). For example, consider the following problem in spatial reasoning:

The cup is on the right of the knife.

The plate is on the left of knife.

The fork is in front of the plate.

The saucer is in front of the knife.

What is the relation between the fork and the saucer?

The theory predicts that reasoners should construct a two-dimensional model of the spatial layout of the items of the sort shown in the following plan:

plate	knife	cup
fork	saucer	

This model yields the conclusion: the fork is on the left of the saucer. In contrast, consider a problem in which the first premise is instead:

The knife is on the right of the cup

and the remaining premises are the same as above. This problem calls for two distinct models because the premises do not fix the relation between the plate and the cup:

plate	cup	knife	cup	plate	knife
fork		saucer		fork	saucer

The two models yield the same conclusion as before, but the problem should be harder because reasoners have to envisage two models. They do indeed draw a smaller percentage of correct conclusions to such multiple-model problems than to the one-model problems (Johnson-Laird and Byrne 1991).

Schaeken et al. have demonstrated similar differences between one-model and multiple-model problems about the temporal relations between events, using such premises as: John takes a shower before he drinks his coffee. Another group of Belgian psychologists (Vandierendonck et al.) have confirmed the difference in both temporal and spatial problems, and also shown that a task which preoccupies working memory for visual and spatial relations interferes with these inferences. Madruga et al. in Madrid have shown that inferences based on a conjunction, which calls for a single model, become harder when that premise is replaced by a disjunction, which calls for multiple models. Likewise, the Italian psychologists, Girotto et al. have shown that an inference based on a conditional premise of the form, If A then B, which calls for at least two models, becomes easier when a premise that eliminates one of these models is presented first rather than after the conditional premise. Two other Italian psychologists at Turin University, Bara and Bucciarelli, have shown that the difficulty of inferences known as syllogisms can also be predicted from the number of models that they require. Thus, the following syllogism is easy:

Some of the artists are beekeepers.

All of the beekeepers are chemists.
What follows?

Seven year-old children can draw the valid conclusion: some of the artists are chemists. In contrast, the next syllogism is hard even for adults:

None of the athletes is a banker.
All the bankers are chefs.
What follows?

Most reasoners construct a model in which the athletes are distinct from the chefs, and conclude that none of the athletes is a chef. To reach the correct conclusion, they need to realize that there could be chefs who are not bankers, and that each of the athletes could be such a chef. The only valid conclusion about chefs and athletes is accordingly that at least some of the chefs (i.e., those who are bankers) are not athletes. Few people reach this conclusion.

Another corroboration of the model theory comes from the study of children's understanding of assertions. Various experimenters, notably the French psychologists Barrouillet and Lecas, have demonstrated a trend in how children develop an understanding of conditional assertions of the form, If A then B. Young children treat them as though they were compatible with just a single model (in which both *A* and *B* occur), slightly older children treat them as compatible with two models (*A* and *B*, and not-*A* and not-*B*), and adolescents correctly treat them as compatible with three models (*A* and *B*, not-*A* and *B*, and not-*A* and not-*B*). Similarly, Sloutsky et al. at Ohio State University have observed that children tend to ignore the second clause of a disjunctive premise in order to ensure that the premise calls for just one model. This tendency to minimize the number of models has been corroborated by Ormerod et al. in their studies of how people paraphrase conditionals and other assertions.

A computer program implementing the model theory made a striking prediction. It showed that if reasoners abide by the principle of truth, then they should make systematic fallacies. These errors should arise because reasoners fail to represent what is false. Several recent studies have corroborated the occurrence of these 'illusory' inferences. Here is an example from an experiment carried out by Goldvarg and Johnson-Laird. Suppose that only one of the following assertions is true about a particular hand of cards:

There is a king in the hand or there is an ace in the hand, or both.

There is a queen in the hand or there is an ace in the hand, or both.

There is a jack in the hand or there is a ten in the hand, or both.

Is it possible that there is an ace in the hand?

In the experiment, 99 percent of the participants responded 'yes,' and the answer seems obvious. Yet, it is an illusion. It is impossible for an ace to be in the hand, because both of the first two assertions would then be true, contrary to the rubric that only one assertion is true. Other experiments have corroborated the occurrence of illusions in causal and conditional reasoning, and Yang, a Chinese psychologist, has shown that they occur with syllogistic premises.

The model theory and theories based on formal rules, which were mentioned earlier, run in parallel for many inferences, but illusory inferences are a case where the two accounts diverge. The model theory predicts the illusions, but the formal rule theories rely solely on valid principles of inference and so they are unable to explain them.

3. *Models, Induction, and Reasoning in Daily Life*

Many of the inferences that you make in daily life are not deductive, that is, even granted the truth of your premises, there is no guarantee that your conclusion is true. You leap to a conclusion that goes beyond the information in your premises. The conclusion rules out some additional possibilities over and above those that the premises rule out, and so it may be false. By definition, such inferences are known as inductions.

The model theory extends naturally to inductive inferences including the informal inductions of daily life (see *Informal Reasoning, Psychology of*). Induction consists in the addition of information to models, sometimes with the consequence of eliminating a model (of a possibility). Often the process is hardly distinct from the normal business of making sense of the world. For instance, you use a voltmeter to examine an electrical circuit and you discover that the following proposition is true:

The battery is dead or the voltmeter is faulty (or both).

You then find out that:

The voltmeter is faulty.

And so you infer: the battery is not dead. Your inference is an induction, because your conclusion could be false even though your premises are true. Your disjunctive premise has three models:

```
dead
        faulty
dead    faulty
```

where 'dead' is a model of the battery as dead and 'faulty' is a model of the voltmeter as faulty. Your discovery that the voltmeter is faulty is consistent with two of these models, and in one of them, the battery is dead (contrary to your conclusion). Hence, your conclusion goes beyond your premises. Yet, it is highly plausible. You may have based it on an unstated belief that the chances are remote of both a dead battery and a faulty voltmeter. This belief leads to you to discount a model so that only the single model yielding your conclusion survives.

Reasoning in real life almost always depends on background knowledge—a knowledge of the meaning of words, of the world, and of the situation at hand.

This knowledge enables you to leap to plausible conclusions. In logic, when a conclusion follows validly from premises, no subsequent information can invalidate it. Logic is monotonic in that additional premises lead merely to additional conclusions, and never to the withdrawal of an earlier conclusion. Logic means never having to be sorry about a valid conclusion. In daily life, however, you do withdraw a conclusion when it conflicts with subsequent evidence. Sometimes, you do so because the conclusion was based on an assumption that you made by default—an assumption warranted only if no evidence exists to the contrary, for example, Quakers are pacifists. If a President is a Quaker, then you may infer that he is a pacifist, but when you learn that he is war-mongering, you withdraw this conclusion. The model theory allows for the withdrawal of the consequences of default assumptions (Johnson-Laird and Byrne 1991).

Many other sorts of inference in daily life lead to conflicts with reality. Suppose you know, for example:

If I turn on the ignition then the engine will start.

You turn on the ignition, and so it follows validly:

The engine will start.

Unfortunately, nothing happens. There is a conflict between a valid consequence of your beliefs and the facts of the matter. Something has to give. At the very least, you have to withdraw your conclusion. Researchers in artificial intelligence have developed various systems of nonmonotonic reasoning to try to deal with such cases (see Brewka et al. 1997), but no one knows what mental processes underlie human nonmonotonic reasoning. At its heart, however, there appears to be the generation of diagnostic possibilities. You try to imagine a scenario—a mental model—that explains why the engine did not start.

4. *Conclusions*

The theory of mental models postulates that human beings reason by imagining the circumstances described in the premises. They envisage mental models of these possibilities, focusing on what is true at the expense of what is false. Unlike other theories of reasoning, this principle correctly predicts the occurrence of certain systematic fallacies. The evidence accordingly confirms that human reasoners do rely on mental models. They construct them from perception, memory, and imagination, manipulate them to make inferences, and base their actions on the results.

See also: Imagery versus Propositional Reasoning; Informal Reasoning, Psychology of; Logic and Cognition, Psychology of; Mental Models, Psychology of; Practical Reasoning: Philosophical Aspects; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Brewka G, Dix J, Konolige K 1997 *Nonmonotonic Reasoning: An Overview*. CLSI Publications, Stanford University, Stanford, CA

Craik K J W 1943 *The Nature of Explanation*. Cambridge University Press, Cambridge, UK

Evans J St B T, Newstead S E, Byrne R M J 1993 *Human Reasoning: The Psychology of Deduction*. Lawrence Erlbaum, Hove, UK

Johnson-Laird P N 1983 *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press, Cambridge, UK and Harvard University Press, Cambridge, MA

Johnson-Laird P N 1999 Deductive reasoning. *Annual Review of Psychology* **50**: 109–35

Johnson-Laird P N, Byrne R M J 1991 *Deduction*. Erlbaum, Hillsdale, NJ

Johnson-Laird P N, Legrenzi P, Girotto V, Legrenzi M, Caverni J-P 1999 Naive probability: A mental model theory of extensional reasoning. *Psychological Review* **106**: 62–88

Polk T A, Newell A 1995 Deduction as verbal reasoning. *Psychological Review* **102**: 533–66

Web page on mental models and the bibliography maintained by Byrne et al. (http://www.tcd.ie/Psychology/Ruth_Byrne/mental_models/)

Wittgenstein L 1922 *Tractatus Logico-Philosophicus*. Routledge & Kegan Paul, London

P. N. Johnson-Laird

Rechtsstaat (Rule of Law: German Perspective)

'*Rechtsstaat*' is a term of the German language and is not easily translated. '*Recht*' stands for 'law' and '*Staat*' for 'state'; in English and most other languages grammar or semantics stand in the way of combining both elements into a compound. The theme evoked by the term '*Rechtsstaat*' is often referred to as 'rule of law,' although the latter expression possesses a tradition of its own and carries connotations different from the German '*Rechtsstaat*.'

1. 'Rechtsstaat': *an Introduction*

Rechtsstaat concerns the age-old question of how to achieve order and freedom within a state. It connotes a balance between creating government authority powerful enough to keep peace both internally and

externally, to protect the state from violence or intervention on the one hand, and to secure a maximum of individual freedom on the other.

Rechtsstaat is a modern coinage which gained currency in German-speaking Europe in the early nineteenth century, when political thought from the French Revolution and the struggle for independence in British North America found wide support among the middle classes in Germany. English concepts for reining in the exercise of governmental power through law thus also count among the sources which influenced the evolution of '*Rechtsstaat*.' From this point of view, a '*Rechtsstaat*' is more than mere statehood characterized by an exercise of government power based on, and limited by, law. *Rechtsstaat* is closely connected to democracy, as democracy is a mode of legitimizing the exercise of power. If a democratic system is intended to be a *Rechtsstaat*, it needs to be governed by legal rules on the sources of such legitimacy.

The German-language term *Rechtsstaat* does not represent a phenomenon unique to Germany, but is the semantic and conceptual reaction to a fundamental issue of modern statehood. In the course of efforts to establish the rule of law in Germany, *Rechtsstaat* had been a political slogan to which various contents and political demands attached. It was only later that the term became textually included in written constitutions. With its many facets and unclear limits, it was misused during the Nazi period. After the end of Nazism, *Rechtsstaat* substantially informed constitutional thinking and is now considered to be a key element characterizing the German constitution of 1949, the Basic Law (*Grundgesetz*); its significance in that text will be further explored in Sect. 2.2. For the moment, we wish to concentrate on how and with what kind of implications *Rechtsstaat* describes a type or a model of state organization (see Sect. 2.1).

Counterbalancing authority and freedom, *Rechtsstaat* is a task not only for today's states, but also for organizations transcending the state level. This is true in particular of integrative processes that have progressed quite far, as in the case of the European Community or similar processes that have been launched in South America and other regions. In a wider perspective, it applies likewise to the task of consolidating the United Nations. Evidence for this may be seen in references to a European Constitution and in increasing references to a 'constitutional order' in international law, which has been considerably shaped by the United Nations (see Sect. 3).

2. *The Modern Meaning of 'Rechtsstaat': The State Level*

The present-day meaning of *Rechtsstaat*—in general theory of the state—on the one hand stands for a category or type of state and, on the other hand, a legal principle. We may speak of '*Rechtsstaat*' as a type, if the law of a state exhibits certain characteristics; these will be presented in Sect. 2.1. They derive above all from the constitution of the state in question and may be identified therein. It is immaterial in this regard whether the constitution is embodied textually in a document, as will most often be the case, or whether it is founded wholly or partly on unwritten law. The question of whether a given state may be considered a *Rechtsstaat* is a theoretical as well as a practical matter. It is a matter of theory insofar as political and legal doctrine seeks to comprehend and to differentiate hypothetical as well as existing states through systematic concepts. It is a matter of practice in the sense that the tradition of the term *Rechtsstaat* as a slogan continues to exert its influence: states whose laws wholly or partly fail to satisfy the requirements of *Rechtsstaat* will be forced to justify these failures in domestic political controversy and even in international relations. The more consensus exists on which core elements are necessary to a reasonable organization of a state which optimally balances the contrasting factors of effective government authority and individual freedom, the more will states be inclined, in their mutual relations, to demand that the concept of *Rechtsstaat* be made reality.

Traditional international law treated such demands as an illegal interference in internal affairs, contradicting the rule of the sovereign equality of states. The principle of non-interference in internal affairs has, however, been much eroded during the latter half of the last century. This concerned initially the area of fundamental human rights. International law today no longer categorically rules out demands from one state that another respect fundamental human rights. Moreover, there is a tendency to widen the ambit of human rights that are thus internationalized and to include aspects of domestic state organization: this pertains in particular to human rights which are directly related to democracy, such as freedom of expression. For example, states may make the establishment of economic or general political relations conditional on beneficiary states satisfying, in their municipal laws, the requirements of *Rechtsstaat*. While formal international recognition of another state used to be a merely declaratory act, this might change in future: recognition of another state is increasingly being predicated on substantive criteria. In the wake of the transformation in Europe during the last decade of the twentieth century, the Member States of the European Union, for example, demanded that the newly emerged states establish governmental structures of a *Rechtsstaat* type before the EU would put relations with these states on a formal footing.

While *Rechtsstaat* as a type covers all states, 'Rechtsstaat' as a principle of constitutional law is confined to a particular state within whose constitutional law this principle may operate. *Rechtsstaat* as a principle is a legal norm. It governs state authority

and may serve as a yardstick in the jurisprudence of a constitutional court. This case is briefly described under Sect. 2.2 with reference to German constitutional law.

2.1 'Rechtsstaat' as a Type

Consensus may be assumed today that certain elements have to be present for a state to be considered a *Rechtsstaat*. Regarding aspects of state organization (so called formal *Rechtsstaat*) it is first and foremost the separation of powers, the independence of the judiciary, the need for administrative behavior to be law-based, and recourse to the courts against acts of government. These elements, however, are closely interrelated with substantive matters. A substantive *Rechtsstaat* refers to a state aiming at safeguarding justice through law-conforming actions of its legislative bodies, i.e., legislative enactments must respect higher-ranking constitutional law. More recently the notion of substantive *Rechtsstaat* has come to include the constitutional guarantee of individual rights, including fundamental and human rights. Central to all this is the need for government authority to be bound by law. Thus the distinction between formal and substantive *Rechtsstaat* is merely a rough one, and the following three aspects may better illustrate the commonly accepted constituent components of *Rechtsstaat*. *Rechtsstaat* formulates requirements for the make-up of a constitution; *Rechtsstaat* is a state whose objective is the realization of human dignity; and *Rechtsstaat* formulates requirements for the make-up of law in general (not merely constitutional law).

The idea of *Rechtsstaat* implies precedence of the constitution over other law. The legislative power is likewise bound by law, in the tradition of most states through a constitutional instrument, i.e., a written constitution, although this is not a necessary feature. Binding the legislative power through higher-ranking law does not mean that legislation to amend the constitution may not be passed. A *Rechtsstaat*, however, will be characterized by formalizing any changes to its constitution and by erecting barriers to its modification, usually by requiring especially enhanced majorities. A *Rechtsstaat*-type constitution will also be guarded against constitutional change through stipulating—in the case of written constitutions—that any amendments shall be expressly made in the text, i.e., by modification of the wording itself or appendix to the constitutional document. This avoids transgressing the bounds of the constitution and prevents confusion which may be caused by the coexistence of texts of equal legal standing. A *Rechtsstaat* does not require that a constitution purport to declare some particularly important component parts or substantive elements as immutable, thus fixing them 'in perpetuity,' although this has indeed been done in some constitutions.

The distinction between higher-ranking constitutional and other law in the concept of *Rechtsstaat* entails the precedence of the higher-ranking law in cases of conflict. Law incompatible with the higher-ranking law is void or, if possible, has to be interpreted in conformity with constitutional law. *Rechtsstaat*-type constitutions often spell out which judicial body is qualified to rule with binding effect on the constitutionality or unconstitutionality of ordinary law, in order to protect the authority of the law from those applying it (executive or judicial branches of government) who might otherwise be tempted to avoid applying the law by alleging its unconstitutionality. It is part of the logic of a *Rechtsstaat* that, for this reason among others, special courts are created for constitutional matters; however, this is not an essential element of a *Rechtsstaat* as long as other courts remain competent to safeguard the constitution and maintain its precedence over ordinary law.

The classical doctrine of separation of powers, which distinguishes between legislative, executive and judicial branches of government, forms the core of the modern concept of *Rechtsstaat*. It should be noted, however, that this distinction is only a basic one, with creation of legal norms by the executive branch or administrative bodies not being alien to the concept of *Rechtsstaat*. Such creation of legal norms does, however, need legal boundaries: laws not enacted by parliament will be acceptable in a *Rechtsstaat* only if the body creating such laws is by itself sufficiently legitimized. If such laws generate rights and duties of individuals, their creation needs to be backed up by an act of the parliamentary legislature. This in turn requires a sufficiently defined scope of the law-making power so delegated—parliament may not transfer its primary legislative responsibility summarily and without limitation to other governmental bodies, or even private norm-creating institutions. *Rechtsstaat* also demands that political parties—as associations designed to gain political power and which are legitimately also oriented towards particular group interests—are placed outside the ambit of government authority.

A *Rechtsstaat*-type constitution purports to make fundamental decisions on values regarding the organization of the state—but only of the state: it does not pretend to achieve comprehensive regulation of society as a whole. On the contrary, such a constitution will leave room for choice. Similarly, it will leave room for choice as regards ordinary domestic law, which must confirm to the constitution without, however, being predetermined by it. Legislating and applying laws does not amount to mere implementation of a constitution. The constitution of a state will not permit us to predict how political processes may unfold in terms of substance. This limits constitutional courts or other bodies entrusted with safeguarding the constitution—for them, interpreting the constitution means bearing in mind its openness. Safeguarding the constitution

thus means identifying where it draws strict limits and where it leaves room for choice and development. Modern constitutions characteristically use general terms, which may imply the prerogative of legislative choices yet to be made. Constitutional jurisprudence in a *Rechtsstaat* does bear political traits but remains confined to retracing the decisions already made in the constitution; it may not disregard the decisions of the constituent power.

Rechtsstaat includes also the guarantee of individual rights. The core of individual rights is the protection of human dignity. Rights to freedom and to equality alike are intended to afford such protection; they are defensive rights vis-à-vis the state which has a duty to prevent, or at least to make an effort to prevent, infringement of these rights in societal relations, i.e., between private persons dealing on an equal footing. Individual rights, whether manifested in international, domestic, constitutional, or ordinary law, have developed in great variety. While not all of these rights will count among the minimum of what a *Rechtsstaat* requires, a *Rechtsstaat* must include individual rights guaranteeing life and health, property and the freedom to choose one's occupation, freedom of expression, assembly, and association, of individual access to the courts, and of protection from discrimination. Access to the courts must not be understood in a formalistic sense: there needs also to be effective protection through granting a fair hearing and independence of the courts. Since *Rechtsstaat*, as has been explained, denotes a type, the concept of *Rechtsstaat* does not intend to impose a uniform level of fundamental rights. It does not prescribe, for instance, the scope of the protection to be afforded to property or the manner of safeguarding judicial independence.

Given the coexistence on an equal footing of numerous persons enjoying fundamental rights—i.e., all citizens—and given also the need to limit the latitude of individual behavior in the interests of the state as a whole, a *Rechtsstaat* needs to formulate rules on the limitability of individual rights and also to define the boundaries of such limitability. In this context, the principle of proportionality in a *Rechtsstaat* merits emphasis. Proportionality requires a balance between different rights and interests. The test of proportionality in a *Rechtsstaat* will be met only by such governmental action as is suitable to further a legitimate objective, which is necessary in the sense that it provides the least restrictive, yet equally suitable alternative available, and which, on consideration of all relevant circumstances, can be regarded as reasonable.

A third characteristic of *Rechtsstaat* is the quality of its law, encompassing the entire legal order and not merely those norms primarily concerned with state-individual relationships; the quality of civil and criminal law are also included. The whole legal system must possess certain qualities. For example, it must meet requirements as to normative clarity to provide legal certainty, because indeterminate standards are incapable of lending themselves to judicial scrutiny. Moreover, law has to satisfy certain immanent criteria in order to serve as an instrument of guidance: it has to be free of contradictions and designed to retain continuity without unduly impeding social change. It has to be capable of acceptance by those bound by it, not perhaps in every detail but in its entirety. Law whose only prospect of implementation would lie in coercive measures or the threat of such would be doomed to fail. Law in a *Rechtsstaat* needs to be founded on a basic consensus of what is just.

2.2 'Rechtsstaat' as a Constitutional Principle

As explained above, one needs to differentiate between using the term *Rechtsstaat* for denoting a type and, on the other hand, for denoting a legally operative norm of constitutional law. To qualify a state as a *Rechtsstaat* does not require such a constitutional principle; its existence is not required if the sum total of constitutional norms already suffices to safeguard the objectives of *Rechtsstaat*, as described in Sect. 2.1. Some constitutions spell out expressly a principle of *Rechtsstaat*; in others the principle is derived from interpreting the constitution. In the Basic Law of the Federal Republic of Germany, for example, *Rechtsstaat* appears as a legal norm in its own right and is frequently referred to by the Constitutional Court. While there is disagreement as to which provision of the Basic Law should serve as the basis for the principle of *Rechtsstaat*, it is by most considered to be Article 20 para. 3 of the *Basic Law*, which reads that legislation shall be subject to the constitutional order and that the executive and the judiciary shall be bound by law and justice. This covers only a segment of the objectives of *Rechtsstaat*. Beyond this, the provision is taken to be a directly operative precept to ensure legal certainty, such as by avoiding indeterminacy in ordinary law or retroactivity of enactments; the principle of proportionality is likewise derived from this precept. State acts not in conformity with this precept are unconstitutional, being incompatible with the principle of *Rechtsstaat*. Laws may be null and void for this reason.

A general principle of *Rechtsstaat* as a constitutional norm in its own right is not without risks, due to its own indeterminacy. *Rechtsstaat* shares this risk with all blanket clauses. Workable civil law could not be imagined if it were limited to stating that behavior in legal relations should be governed by the principle of good faith alone, and constitutional law may similarly not retreat to merely stating a principle of *Rechtsstaat*, as this would leave decisions on what has to be considered in conformity with *Rechtsstaat* entirely to the judicial branch. Law dominated by blanket clauses concedes considerable decision-making powers to the courts and could thus not be safe

from arbitrariness. This is why a blanket-clause style of the *Rechtsstaat* principle can only have secondary relevance. Translating a *Rechtsstaat* into reality is better achieved by detailed norms that specify relevant objectives and that furnish a safer foundation for applying the law.

3. 'Rechtsstaat' and International Organizations

The notion of *Rechtsstaat* could be a useful concept also in the sphere of structures transcending individual states, as international organizations need to strike a balance between overarching interests affecting the whole organization and those regarding individual members. In the case of traditional international organizations, the issue has hardly ever been put forward. These organizations group together equal subjects of international law, i.e., states, in order to pursue jointly certain aims which as a rule remain individual objectives; the sovereignty of participating states remains unaffected. The European integrative community had its starting point in such a structure, but from the beginning it was designed to include transfer of sovereignty, and this community has increasingly acquired legal powers that justify looking at it as a new type of organization. Here, as in other organizations which initially aimed at economic integration but which have the potential to expand beyond this particular sphere, the question of arrangements addressing requirements of *Rechtsstaat* needs to be raised. The basic distinction between higher-ranking constitutional law and other law would apply in such a case. Matters of the quality of the law, such as the absence of systemic logical inconsistency or its indeterminacy, would also need to be addressed, as would issues of safeguarding compliance with the law and judicial control. Derived from the municipal laws of its member states, the objective of building a *Rechtsstaat* legal environment has already been realized in the European Union to a considerable degree, especially in the judicial branch. The European Court, while not a constitutional court, may be considered as a guardian of the fundamentals of Community law. Preserving the municipal *Rechtsstaat* heritage in supranational structures is an important task for the future that will grow more urgent as creation, enforcement, and control in the area of law is transferred to the integrated, supranational level.

Experiences gained in connection with *Rechtsstaat* are of interest also to international organizations to which member states have not transferred sovereign rights, insofar as such organizations are competent to issue legally binding decisions, to define legal norms, or to impose sanctions. Special organizations such as the World Trade Organization (WTO) or the United Nations (UN) (the global international organization designed to secure peace, to renounce the use of force, and, beyond that, to secure basic human needs), come to mind in this regard as their charters comprise normative potentiality. The UN Security Council has powers that would enable it to fashion international relations in a legally binding manner far more than it has to date. The UN Security Council may be viewed as an organ of law enforcement. Apart from the right of self-defence of individual states, the Council enjoys a kind of monopoly of the use of force. Measuring this position against the yardstick of *Rechtsstaat* will reveal numerous shortcomings. The Council acts on the strength of norms whose determinacy may be open to question. It enjoys a monopoly on the use of force without legal safeguards, which would ensure that it acts to protect subjects of law who are threatened with violence by individual states. Being dominated by a privileged group of states, its legitimacy is in doubt. Judicial control is not ensured either, as the International Court of Justice does not exercise the role of a constitutional court in respect of the UN Charter as an international constitutional order. Any step in the direction of transforming the United Nations into a global state-like structure therefore imperiously calls for expansion of *Rechtsstaat* elements. This naturally cannot be achieved by simply imposing regulatory patterns tried and tested in *Rechtsstaat*-type national jurisdictions, but the tradition of *Rechtsstaat* offers a rich lode of lessons that will have to be heeded in the efforts to find an adequate legal framework for the modern world system and its segments.

See also: Constitutional Courts; Constitutionalism; Constitutionalism, Comparative; Democracy; European Union Law; Fundamental Rights and Constitutional Guarantees; International Organization; Rule of Law; State, History of; State, Sociology of the

Bibliography

Bähr O 1864 *Der Rechtsstaat*. Kassel, Göttingen, Germany

Franck Th M 1999 *Democracy, Legitimacy and the Rule of Law: Linkages*. New York University Law School, Public and Legal Theory Working Paper No. 2, New York, Social Science Research Network Electronic Library (SSRN), Working Paper Series. http://papers.ssrn.com

Hofmann H 1995 Geschichtlichkeit und Universalitätsanspruch des Rechtsstaats. *Der Staat* **34**: 1–32

Kunig Ph 1986 *Das Rechtsstaatsprinzip*. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany

Kunig Ph 1995 Völkerrecht als Öffentliches Recht—Ein Glasperlenspiel? In: Randelzhofer A, Scholz R, Wilke D (eds.) *Gedächtnisschrift für Eberhard Grabitz*. C. H. Beck'sche Verlagsbuchhandlung, Münich, Germany, pp. 325–46

Neuman G L 1999 U.S. Constitutional conception of the rule of law and the Rechtsstaatsprinzip of the Grundgesetz. In: Battis U, Kunig Ph, Pernice I, Randelzhofer A (eds.) *Das Grundgesetz im Prozess europäischer und globaler Verfassungsentwicklung*. Nomos-Verlagsgesellschaft, Baden-Baden, Germany, pp. 253–68

Schmidt-Aßmann E 1987 Der Rechtsstaat. In: Isensee J, Kirchhof P (eds.) *Handbuch des Staatsrechts der Bundesrepublik Deutschland*. C. F. Müller Juristischer Verlag, Heidelberg, Germany, p. 24

Shapiro I (ed.) 1994 *The Rule of Law*. New York University Press, New York
Sobota K 1997 *Das Prinzip Rechtsstaat, verfassungs- und verwaltungsrechtliche Aspekte*. J. C. B. Mohr (Paul Siebeck), Tübingen, Germany
Voigt St 1998 *Making Constitutions Work—Conditions for Maintaining the Rule of Law*. Social Science Research Network Electronic Library (SSRN), Working Paper Series. http://papers.ssrn.com

P. Kunig

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Recognition Memory (in Primates), Neural Basis of

The term *recognition* refers to the knowledge or feeling that an object has been encountered previously. Two different mental processes may contribute to recognition judgments: a general feeling of 'familiarity' and a specific 'recollection.' These two aspects of recognition may depend on different brain systems, but are treated together here.

1. Neural Systems Mediating Visual Recognition

For most sensory modalities (e.g., vision and touch), receptors convert sensory information into neural signals, and modality-specific cortical fields, such as the primary visual cortex ('V1' in Figure 1), process that information further. These early sensory processing stages are a prerequisite for recognition memory. However, the core recognition system must function more generally. If a blind person fails to recognize a visually presented object, we say that he or she has a visual impairment, not a failure of recognition memory. Therefore, this article will focus on general aspects of recognition and its neural basis, neglecting modality-specific ones.

1.1 Methods for Assessing Recognition Memory in Nonhuman Primates

Because both human and nonhuman primates rely heavily on visual information in their daily activities, recognition is best understood for visible objects. Visual recognition memory in monkeys has been studied by making selective brain lesions and by disconnecting one brain region from another. These studies identify whether a given brain structure is critical for recognition. Monkeys, of course, cannot tell us whether they remember seeing an object. For this reason, methods have been developed that allow them to reveal that information. One commonly used method is the *delayed nonmatching-to-sample task* (DNMS). In DNMS, each trial has two parts: a sample presentation followed by a choice test. During the sample presentation, the monkey sees a single object, which covers the central well of a three-well test tray. The monkey then displaces the object to obtain a small piece of food hidden in the well underneath. A few seconds later the monkey sees the same, now-familiar object plus a novel one, with one object over the left well of the test tray and the other over the right well. The monkey can obtain more food by displacing the novel object but not by pushing aside the familiar one. Thus, on the choice test, the monkey solves the problem by applying a 'nonmatching' rule, i.e., by choosing the object that does not match the sample. If the monkey consistently chooses the novel object on the choice test, one can infer that the monkey *recognizes* the sample. In the laboratory, there are many variations of this basic method. For example, monkeys can be trained using a 'matching' rule (the *delayed matching-to-sample task*, or DMS) rather than the nonmatching one. It is also common to use a computerized apparatus that presents two-dimensional pictures on a touch-sensitive video screen rather than using actual objects on a test tray.

1.2 Role of the Medial Temporal Lobe in Monkeys

It has been known for many years that a limited set of brain structures, including components of the medial temporal lobe (MTL), are necessary for storing information about facts, objects, and events from all sensory modalities (Squire 1987). Early investigations into the neural basis of recognition memory pointed to two structures buried deep in the MTL: the amygdala and hippocampus (Mishkin 1982, see Figure 1). However, those experiments involved the mechanical removal of brain tissue, a technique that produced unintended damage to other parts of the MTL. More recent studies have used stereotaxic injection of neurotoxins to make more selective tests. These experiments show that, within the MTL, the perirhinal cortex and the entorhinal cortex, rather than the amygdala and hippocampus, are essential for visual recognition (Murray 2000). Monkeys with an intact amygdala and hippocampus, but without the perirhinal and entorhinal cortex, cannot recognize objects if they have to remember them for more than a few seconds.

Experts accept a central role for perirhinal and entorhinal cortex in object recognition and also agree that the amygdala is not important for this type of memory. Controversy remains, however, about the potential contribution of the hippocampus. Damage to the hippocampus leads to impairments in recognition memory in some studies but not in others. For example, one test of recognition relies on the tendency of monkeys to spend more time looking at novel objects than at familiar ones. This innate tendency is called 'preferential viewing.' Although hippocampal damage fails to produce consistent and systematic deficits on DNMS, it does disrupt pref-

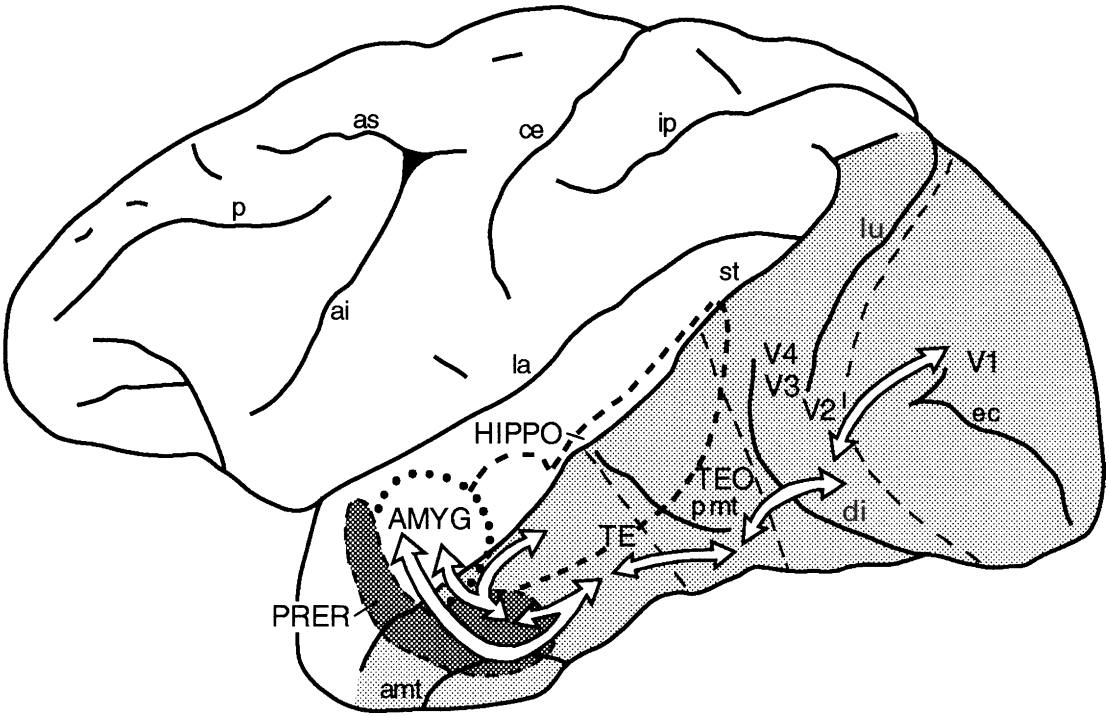

Figure 1
Diagram of the lateral view of a monkey brain showing some of the brain structures mediating visual recognition memory. The perirhinal and entorhinal cortex (PRER) are located on the ventromedial surface of the temporal lobe, and so are only visible by looking through a 'transparent' temporal lobe. The amygdala (AMYG) and hippocampus (HIPPO) are buried deep in the medial temporal lobe, and the hippocampus is shown slightly higher than its real location, for ease in viewing the lateral temporal cortex. Arrows indicate two-way information flow from early (V1–V4) to late (TEO and TE) cortical visual processing areas. These regions then transfer information into the amygdala and into the perirhinal and entorhinal cortex, which provide the main route by which information reaches the hippocampus. Abbreviations: ai, inferior arcuate sulcus; amt, anterior middle temporal sulcus; as, superior arcuate sulcus; ce, central sulcus; ec, external calcarine sulcus; io, inferior occipital sulcus; ip, intraparietal sulcus; lu, lunate sulcus; la, lateral fissure; p, principal sulcus; pmt, posterior middle temporal sulcus; st, superior temporal sulcus

erential viewing. This could be because the preferential viewing test, unlike DNMS, does not rely on learned rules, or because the stimuli have no association with food. Interestingly, on DNMS, larger perirhinal and entorhinal cortex lesions cause greater memory deficits than smaller ones, which is what one would predict if a structure were playing a direct role in a particular function. The reverse relationship holds for the hippocampus. That is, larger hippocampal lesions are associated with better recognition memory than smaller lesions, suggesting an indirect relationship of the damage to recognition memory deficits. The fact that only the perirhinal and entorhinal cortex have the predicted relationship to recognition memory supports the view that they are the parts of the MTL mediating that function.

The same parts of the MTL that are essential for visual recognition memory subserve tactual recognition memory as well. They are also critical for many types of stimulus–stimulus association memory. Accordingly, their function may best be characterized as one of object identification. Through this process, the visual, tactual, gustatory, olfactory, and auditory properties of objects are somehow linked or 'bound.' Thus, the perirhinal and entorhinal cortex are essential not only for visual recognition, but also for storage of information about objects generally (Murray 2000).

1.3 Role of the Thalamus and Frontal Lobe in Monkeys

The perirhinal and entorhinal cortex do not operate alone, but instead work in concert with other anatomically connected brain structures that, together, form a neural network mediating visual recognition

memory. At least two other brain regions participate: the medial thalamus and the ventromedial part of the prefrontal cortex (Mishkin and Murray 1994). Damage to either of these structures produces essentially the same effect on visual recognition memory as does damage to the perirhinal and entorhinal cortex. In addition, the thalamic and frontal regions noted above are connected with each other, and each is connected to the perirhinal and entorhinal cortex. These structures need to interact to mediate visual recognition; surgical manipulations that prevent information from one structure from reaching a second produce a deficit in recognition memory much like that seen after lesions to each individual site (Parker and Gaffan 1998). Precisely how these structures interact to produce recognition memory remains uncertain; it is possible that each structure underlies some unique aspect of recognition memory.

1.4 The Recognition Network in the Human Brain

The same neural structures that are critical for recognition memory in monkeys appear to be essential in humans as well. For example, memory impairments are found in humans following damage to the MTL as a result of either brain infections (e.g., Herpes encephalitis) or temporary interruption of the blood supply. Damage to the medial thalamus, as a result of Korsakoff's disease, and to portions of the ventromedial prefrontal cortex, as a result of aneurysms of the anterior communicating artery, yield similar effects. It must be kept in mind, however, that in clinical cases brain damage rarely respects the boundaries of individual structures, so that the kind of precision outlined above for monkeys is not often possible for the human brain. Nevertheless, there is evidence that in humans, as in monkeys, the perirhinal and entorhinal cortex are important for recognizing objects. And, as in monkeys, there is controversy about the role of the hippocampus. People with damage to the hippocampus and adjacent perirhinal and entorhinal cortex are much more impaired in stimulus recognition than those with damage restricted to the hippocampus. Whereas some studies have shown that humans with damage limited to the hippocampus and closely related structures have relatively intact recognition abilities (Aggleton and Brown 1999), others have reported that such subjects are, on average, impaired in recognition memory (Squire and Knowlton 2000). In humans, the hippocampal contribution to recognition memory appears to reflect 'recollection' as opposed to 'familiarity' judgments.

2. Neurophysiology of Visual Recognition

Although the neuronal mechanisms mediating visual recognition memory are not known, there is some information about the way in which neurons signal that a visual stimulus has been seen previously (Brown 1996). As shown in Figure 1, the temporal lobe consists of several parts. The MTL has already been discussed. On the outer, lateral surface of the temporal lobe are located visual areas TE and TEO. These regions are called the *inferior temporal cortex* (IT). The perirhinal and entorhinal cortex may be considered part of either IT or MTL. Many neurons in IT, including the perirhinal and entorhinal cortex, show changes in sensory responses with stimulus repetition. Two main kinds of recognition-related responses predominate.

Some neurons respond less to a visual stimulus on a monkey's second or subsequent viewing of an object than they do on the first. Other neurons respond more. Similar phenomena occur in the medial thalamus and the prefrontal cortex, but not in the hippocampus. Thus, these types of activity signaling stimulus repetition occur in precisely those areas thought to be essential for performance on DNMS. In addition, there is a third type of neuronal activity related to recognition. In monkeys performing DMS (see Sect. 1.1), some neurons in both IT and the prefrontal cortex reflect information about the sample object during the interval between sample presentation and choice test, as though the cell were maintaining a memory of the sample (Desimone et al. 1995).

Furthermore, for prefrontal neurons, this type of activity persists even when other objects appear during a memory period. This kind of activity could be very useful for the short-term storage of information about an object and for directing attention to items of interest.

3. Cognition and Recognition

Recognition is only one among many kinds of memory and cognition. It means, literally, to 'know again.' In people, recognition can be normal, or nearly so, in the face of devastating memory impairments of other types—for example, an inability to remember personal daily activities. Conversely, recognition can be extremely impaired in people who are capable of learning and applying complex rules and strategies. Accordingly, recognition memory can be viewed as an aid to more complex cognitive operations, but is, itself, a modular function with adaptive advantages of its own. The nature of these advantages is uncertain, but probably involves the ability to direct attention to novelties in the environment (perhaps by reducing the sensory responses to recognized objects) or to familiar objects that are of importance to the behavior of the moment. It remains to be seen whether the mechanisms that underlie the recognition of objects, as discussed in this article, also subserve the recognition of thoughts and ideas, logical truth and falsity, esthetic beauty and ugliness, or other abstract aspects of human recognition that distinguish us from our fellow primates.

See also: Amygdala (Amygdaloid Complex); Declarative Memory, Neural Basis of; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Recognition Memory, Psychology of

Bibliography

Aggleton J P, Brown M W 1999 Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral and Brain Science* **22**: 425–89

Brown M W 1996 Neuronal responses and recognition memory. *Seminars in Neuroscience* **8**: 23–32

Desimone R, Miller E K, Chelazzi L, Lueschow A 1995 Multiple memory systems in the visual cortex. In: Gazzaniga M S (ed.) *The Cognitive Neurosciences*. The MIT Press, Cambridge, MA

Mishkin M 1982 A memory system in the monkey. *Philosophical Transactions: Royal Society London B* **298**: 85–95

Mishkin M, Murray E A 1994 Stimulus recognition. *Current Opinion in Neurobiology* **4**: 200–206

Murray E A 2000 Memory for objects in nonhuman primates. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. The MIT Press, Cambridge, MA

Parker A, Gaffan D 1998 Interaction of frontal and perirhinal cortices in visual object recognition memory in monkeys. *European Journal of Neuroscience* **10**: 3044–57

Squire L R 1987 *Memory and Brain*. Oxford University Press, New York

Squire L R, Knowlton B J 2000 The medial temporal lobe, the hippocampus, and the memory systems of the brain. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. The MIT Press, Cambridge, MA

E. A. Murray

Recognition Memory, Psychology of

Recognition memory is usually used to refer to both a memory measurement and a memory process that occurs in this measure of memory. When discussing recognition memory as a memory measurement, the difference between a recognition memory test and other memory tests such as recall tests can be easily distinguished. Recognition tests also produce certain effects that are particular to the processes underlying recognition memory. Even though the nature of these processes is very complicated, some models of recognition memory have nonetheless been proposed (see, for example, Raaijmakers and Shiffrin 1992). In this article, these two characteristics of recognition memory are examined in more detail.

1. Recognition Memory as a Memory Test

1.1 The Nature of a Recognition Memory Test

In a recognition memory experiment, to-be-remembered items (also called target items) are presented to subjects. An item may refer to a nonsense syllable, a word, a number, a picture, and so on. In such tasks, subjects are given a recognition memory test after exposure to such items.

There are basically two types of recognition memory tests. In a yes/no recognition memory test, a subject is presented a series of items one at a time. The subject is then asked to say 'yes' or 'old' if the subject thinks the current item is a target item, and 'no' or 'new' if the subject thinks it does not qualify as a target one (nontarget items are called distractor items).

This type of a recognition memory test is similar to the sort of true/false test often encountered in school examinations. Another type of recognition memory test is a forced-choice test. In a forced-choice test, the subject is shown two or more items at a time. One of the items is usually a target item and the others are distractor items. In this type of recognition memory test, the subject is asked to pick out the target item from a set of items. If the subject is shown three items at a time, for example, it is called a three-alternative forced-choice test.

As such, the forced-choice test is similar to multiple-choice tests often encountered in school examinations. Of course, many additional variations on recognition memory tests may be considered. For instance, in addition to judging whether an item given on a recognition memory test is a target item or a distractor item, a subject may be asked to give a numerical confidence rating consisting of the range of, for example, 1 (least certain) to 5 (most certain), to indicate how certain the subject is that the judgment is correct.

In order to develop a better understanding of recognition memory tests, it is worthwhile contrasting them with other popular forms of memory test, such as the aforementioned recall tests. In recall tests of memory, a subject is asked to reproduce or generate all the items that had been presented during an earlier study phase. Thus, the basic difference between recognition and recall test procedures is that in a recognition memory test, subjects are given the target and distractor items during the test, whereas in a recall test, the subjects are given instructions to reproduce or generate the items without displaying them.

One of the important differences is that, in general, the subjects can recognize items in a study list much better than they can recall them. In fact, if the subjects are first given a recall test and then given a recognition memory test, it will be usually found that they can recognize many items that they were unable to recall.

1.2 Signal Detection Theory and Performance on a Recognition Memory Test

In a recognition memory test, instead of being assumed that subjects either memorize target items or don't memorize them, it is assumed that the subjects judge as

the target items on something like level of familiarity which varies in memory strength. When an item can be retrieved more readily than the other items, it will be said that the item is at full strength.

The memory strength of a test item is thought of as the degree of its familiarity in long-term memory (LTM). The stronger the test item is in LTM, the more familiar it will be.

To illustrate the assumption, one recognition memory experiment will be considered. Assume that a list of items is given to two groups of subjects, followed by a yes/no recognition memory test. The subjects are tested with a mixed sequence of target and distractor items one at a time. Prior to the recognition memory test, one group of subjects (called a guessing group) is instructed to say 'yes' if they think an item was on the list, and 'no' if they think it is a distractor item. Guessing is permitted in this group when the subjects receive a recognition memory test. The other group of subjects (called a strict group) receives the same instruction as the guessing group except the guessing instruction. The subjects in this group are instructed not to make a guess when they judge whether an item given on a recognition memory test is a target item or a distractor item. The subjects in the guessing group will of course use guessing strategies. So the results may show that the subjects in the guessing group will pick up more target and distractor items than those in the strict group. The results mean that the subjects in the guessing group judge the target and distractor items on a higher level of familiarity than those in the strict group. However, as can be seen, performance on a recognition memory test in the guessing group may be very similar to that in the strict group.

The signal detection theory which originates with the study of auditory detection tasks (see Green and Swets 1966) and has been applied to recognition memory explains the results of this experiment clearly. The signal detection theory enables researchers to estimate the amount of information stored in LTM on which subjects' judgments on a recognition memory test are based.

The signal detection theory assumes that the subjects estimate the probabilities that a test item came from the distributions for target and distractor items, respectively, on the basis of its memory strength (or familiarity).

In a typical yes/no recognition memory test, there are four cells that can occur in combination with a subject's 'yes' or 'no' responses and target-distractor items. First, the item is a target item and the subject says 'yes'; this is the correct response and is called a *hit*. Second, the item is a target item but the subject says 'no' to it; this is not the correct response and is called a *miss*. Third, the item is a distractor item and the subject says 'no'; this is the correct response and is called a *correct rejection*. Fourth, the item is a distractor item but the subject say 'yes' to it; this is not the correct response and is called a *false alarm*.

These four cells are not independent. As shown above, the hit rate and the miss rate mean the response rates to the target items and the false alarm rate and the correct rejection rate mean the response rates to the distractor items. When the hit rate is known, the miss rate can be understood automatically. Similarly, when the correct rejection rate is known, the false alarm rate can be understood. Generally, the hit rate and the false alarm rate are referred to on recognition memory.

Figure 1 shows both a relationship between the distributions of memory strength for target and distractor items and four cells in a yes/no recognition memory test. Figure 1 shows that the distributions for target and distractor items on the basis of their memory strength represent normal curves and that the mean strength of the target items is usually higher than that of the distractor items.

The distance between the means of the distributions for target and distractor items is called d' (called *d-prime*). More precisely, the d' value is obtained by dividing the difference between the two means by the common standard deviation of the distributions. We can easily get the d' value from some publications (e.g., Elliot 1964, Hochhaus 1972). When the d' value increases, the subject will discriminate between target and distractor items easily. When the d' value is small, the subject will find it difficult to discriminate between two sets of items.

In addition to the d' value, there is another value called β (*beta*). The β value means the level of the criterion strength on which the subject bases the decision. When the d' value remains constant and the β value goes up (e.g., the β value moves to the target distribution), both hit rate and false alarm rate go down. When the d' value remains constant and the β value goes down (e.g., the β value moves to the distractor distribution), both hit rate and false alarm rate go up. When the d' values are assumed to be the same, the β value in the guessing group stated earlier may be very low and the β value in the strict group may be very high. Figure 1 shows one example of d' and β values on recognition memory. As can be seen, the d' value is referred to recognition accuracy which means the strength difference between target and distractor items in LTM and the β value is referred to as decision criterion. The subject evaluates the memory strength difference between target and distractor items on the basis of decision criterion.

1.3 Some Effects in Recognition Memory

Some effects have been demonstrated in recognition memory, which are referred to as the mirror effect and the context effect. These effects in recognition memory are explained in brief.

First, there is a strong regularity of recognition memory called the mirror effect (see Glanzer and

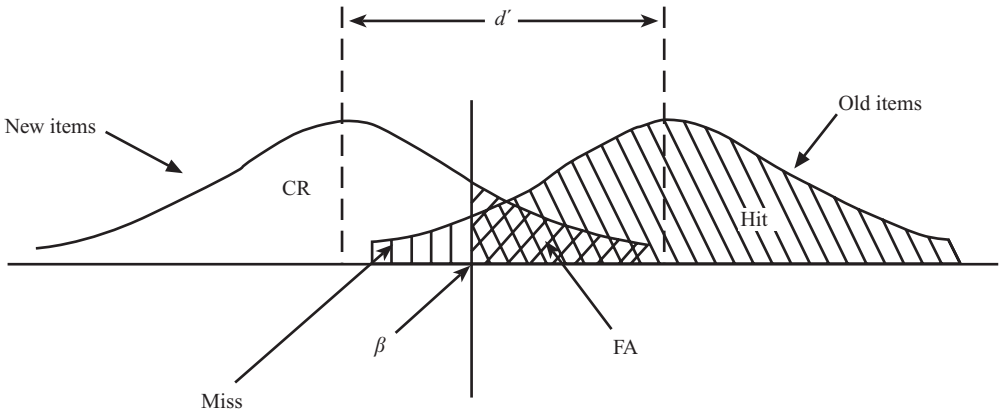

Figure 1
Four types of response results (hit, miss, false alarm (FA), and correct rejection (CR)), d', and β in recognition memory

Adams 1985). The mirror effect refers to the regularity of recognition memory in which the type of items that is relatively easy to recognize as old when old is also relatively easy to recognize as new when new. Glanzer and Adams (1985) analyzed a total of 80 experimental findings with several different classes of items and found that performance on new items from each class is correlated with performance on the corresponding classes of old items.

Second, subjects can usually recognize target items better than they can recall them. However, in such test circumstances as recognition memory is measured under the changes in the context between study and test, it will be found that the subjects fail to recognize target items. It is known as the context effect in recognition memory. The context effect in recognition memory shows that even recognition memory depends on the relation between the context at the time of the study and at the time of its test. For detailed treatment of the context effect see *Encoding Specificity in Memory*.

2. *Retrieval Processes in Recognition Memory*

As stated earlier, recognition and recall are usually used to refer to not only memory measurements, but also memory (or retrieval) processes which occur in these memory measurements. There have been hot controversies about how retrieval processes in recognition and recall are similar and different. That question is very difficult to answer and is still controversial (Brown 1976, Tulving 1983).

Generally speaking, recognition memory seems to be easier than recall. Subjects who receive a recognition memory test only discriminate between target and distractor items, even though they do not know the target items clearly. There are some reasons why recognition memory seems to be easier than recall. According to a classical threshold model of recognition memory, the level of the discriminative threshold in recognition is thought to be higher than that in recall. According to the two-process model of recall (Kintsch 1970), recognition memory and recall involve different retrieval processes. Recall involves an extra search (or generate) process in addition to a recognition (or decision) process. For more detailed treatment of retrieval processes in recall see *Memory: Organization and Recall*; *Memory Retrieval*; *Encoding Specificity in Memory*.

2.1 *One-process Models or Two-process Models in Recognition Memory*

How about retrieval processes in recognition memory? Recognition memory can usually be referred to by two types of models, one known as one-process models and the other known as two-process models.

One-process models include the Search of Associative Memory (SAM) model proposed by Gillund and Shiffrin (1984) and the Hintzman (1988) model. They assume that recognition judgments are based on the target item's familiarity involving a single retrieval (search-like) step. Two-process models include the Atkinson and Juola (1974) model and the Mandler (1980) model.

They assume that the retrieval processes in recognition memory are the result of two processes: familiarity and search. For example, Atkinson and Juola (1974) carried out response time experiments from well-learned lists. They obtained the results that items with high or low familiarity values are associated with a fast response of a target item or a fast response

of a distractor item, respectively, whereas items with intermediate familiarity values are associated with a slower search process to judge whether an item is a target item or a distractor item. This search process is generally supposed to be slower than the familiarity-based process in the two-process models of recognition memory. The controversies about whether there is a search process in recognition memory are still continuing.

2.2 Recognition Memory With and Without Recollective Experiences

In recent years, there has been considerable interest in comparing performance on a task that does not require conscious or intentional recollection of previous experiences with performance on a task that requires it (see *Implicit Memory, Cognitive Psychology of*). A recognition memory test is known as one of the tests that requires conscious recollection of previous experiences. However, Gardiner (1988) assumes that recognition memory may entail two processes concerning recollective experiences: one is a process with recollective experiences and the other is a process without recollective experiences. These two processes are related to differences in subjective experiences.

In a recognition memory test, when recognizing each item, subjects are instructed to indicate whether they are able consciously to recollect its prior occurrence in the study list (called 'remember') or whether they recognize it on some other basis (called 'know'). The results show that a few independent variables, such as levels of processing (deep vs. shallow processing), generating vs. reading target items, and divided vs. undivided attention, have been found to influence 'remember' recognition and to have little effect on 'know' recognition. For example, performance on semantic processing was better than that on shallow processing in 'remember' recognition. However, performance on semantic processing was the same as that on shallow processing in 'know' recognition.

The measures of a 'remember' judgment and of a 'know' judgment are introduced by Tulving (1985), according to which conscious recollection or 'remembering' is a by-product of retrieval from the episodic memory system. Feelings of familiarity or 'knowing' are characteristic of retrieval from the semantic memory system. Using 'remember'-'know' judgments in recognition memory, it may be possible to examine retrieval processes in recognition memory and memory systems.

See also: Learning and Memory, Neural Basis of; Memory Models: Quantitative; Memory Retrieval; Microsimulation in Demographic Research; Recognition Memory (in Primates), Neural Basis of; Short-term Memory, Cognitive Psychology of; Short-term Memory: Psychological and Neural Aspects; Signal Detection Theory; Signal Detection Theory, History of; Working Memory, Neural Basis of; Working Memory, Psychology of

Bibliography

Atkinson R C, Juola J F 1974 Search and decision processes in recognition memory. In: Krantz D H, Atkinson R C, Luce R D, Suppes P (eds.) *Contemporary Developments in Mathematical Psychology*. Freeman, San Francisco, CA, Vol. 1

Brown J (ed.) 1976 *Recall and Recognition*. John Wiley & Sons, London

Elliot P B 1964 Tables of d'. In: Swets J A (ed.) *Signal Detection and Recognition by Human Observers*. Wiley, New York

Gardiner J M 1988 Functional aspects of recollective experience. *Memory and Cognition* **16**: 309–13

Gillund G, Shiffrin R M 1984 A retrieval model for both recognition and recall. *Psychological Review* **91**: 1–67

Glanzer M, Adams J K 1985 The mirror effect in recognition memory. *Memory and Cognition* **13**: 8–20

Green D M, Swers J A 1966 *Signal Direction Theory and Psychophsics*. John Wiley and Sons, New York

Hintzman D L 1988 Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review* **95**: 528–51

Hochhaus L 1972 A table for the calculation of d' and β. *Psychological Bulletin* **77**: 375–6

Kintsch W 1970 Models for free recall and recognition. In: Norman D A (ed.) *Models of Human Memory*. Academic Press, New York

Mandler G 1980 Recognizing: The judgment of previous occurrence. *Psychological Review* **87**: 252–71

Raaijmakers J G W, Shiffrin R M 1992 Models for recall and recognition. *Annual Review of Psychology* **43**: 205–34

Tulving E 1983 *Elements of Episodic Memory*. Oxford University Press, New York

Tulving E 1985 Memory and conscionsness. *Canadian Psychology* **26**: 1–12

H. Tajika

Reconstruction/Disaster Planning: Germany

In 1945, with World War II over and the National Socialist dictatorship under siege by Allied troops, great parts of German cities lay in rubble, their populations often reduced to only a few tens of thousands. Despite great hardship, life reemerged unexpectedly fast and soon discussions began on rebuilding the cities. Reconstruction plans had been in preparation since 1943. There was therefore no 'zero hour' for urban planning. The following article will underline the continual development in planning

theory and practice, the inspirations for the concepts, and the debate about urban design before and after 1945.

1. Preliminaries

After World War I, town and country planning in Germany in general changed its focus from the policy of improvement—slowly introduced since the late-nineteenth century—to a strategic approach. This phase was hampered by the economic crisis in 1929. When the National Socialists came to power in 1933, they had in fact no specific policies for urban planning but were able to select from existing scholarship those strategies which were useful to the new dictatorship:

1933/34: Improvement of both traffic and housing conditions in the historic urban cores of German cities had been widely neglected since World War I in favor of building new housing schemes at the urban fringe. The first concepts for the renewal of historically valuable and attractive housing stock from pre-industrial periods date from around 1930. This new approach then was to maintain the pretty street layout by clearing the blocks' interiors (as occurred in Brunswick, Frankfurt, and Kassel). Lack of money due to armaments production led to such schemes being deferred. However, it was useful to early National Socialist policy to support them as work providing measures which brought beneficial improvements to local economies, as well as for ideological reasons (see von Petz 1990).

1937: Albert Speer's plan to redesign Berlin as a future 'Germania' became the model for the redesign of German provincial capitals. Showing conventional architecture of monumental style with axes appropriate to the parading party and public buildings within central urban areas, these plans were nevertheless designed according to existing planning knowledge which had already been developed widely in the preceding decade.

1940: The policy of armament and the war of conquest led to a stronger functional approach to planning. Growing individual mobility and the change from the detached settlements of the early 'blood and soil' period to compact housing concepts provoked a change to planning paradigms even in wartime. The advance to the East led to initiatives for redesigning the settlement structure of the conquered territories. Walter Christaller's 'central places' theory of 1933 was easily adaptable to the required hierarchical structural system. Political goals combined with military ambitions also led to marked improvements in planning methods and techniques, such as statistics, cartography, and aerial photography.

1943: The disastrous bombing of cities forced a return to the debate dating back to the CIAM discussion of 1933 on rebuilding towns. A decree by the Führer regarding the reconstruction of German cities after the anticipated victory empowered Albert Speer to establish a team of reconstruction planners, with each member given responsibility for one or more cities. The plans were quite heavily influenced by comparisons with advanced design methods to international standard and some, indeed, served as first concepts for reconstruction after 1945 (Durth 1986).

2. Political Reconstruction

From 1871, national state pride and the desire for international reputation dominated the politics of the Imperial Reich. The Morgenthau doctrine aimed to prevent Germany from returning to such a powerful state. In 1945, therefore, the Allies promoted the decentralization of Germany, basing their strategy on a tradition dating from the Middle Ages to the nineteenth century of strong feudal states with attractive provincial capitals. The city of Bonn, a modest city among those former feudal seats, became the capital of the Federal Republic of Germany. However, many of the new West German states retained their former seats as capitals. Berlin was divided and the German Democratic Republic kept East Berlin as its capital. In the East, the former state capitals were promoted less. Thus two systems existed simultaneously between 1961 and 1990: East Berlin as the capital and center of power in the socialist state, with other cities as only regional centers, and in the western federal state a more modest capital with 11 quite strong state capitals and, under them, a tier of regional towns.

3. Rebuilding the Cities: The Western Approach

Before the war the modernization process in German cities was badly impeded by their historic heritage. West German postwar planners therefore used the opportunity of rebuilding the bombed cities according to modern international planning standards that, before 1933, had been developed continuously and which Germany, with its dreams of a thousand-year Reich, had for a while gambled away. Nevertheless a very few international links were maintained (Fischer 1990). The founding of the Federal Republic in 1948, the monetary reform urged by the Marshall Plan, and the Korean War, produced economic growth which helped to speed reconstruction. This focused on two goals: first, to bring sun, air, and light into the urban cores which, to most minds, had suffered so much from urban concentration and industrialization and, second, to provide motor car access to the cities.

But West Germany adopted different approaches for rebuilding the demolished core areas in the postwar period (for statistics on rubble, see Diefendorf 1993, pp. 3–17). The schemes ranged from radical modernization to a moderate reconstruction or tra-

ditional adaptation (also Beyme 1987, pp. 175–7). Munster in Westfalia or the Bavarian capital of Munich, for example, employed quite traditional lines, based on the historic layout and scale and, in cases of new constructions, a moderate adaptation of modern design for the facades. Monuments were rebuilt or restored. Cities such as Kassel, Dortmund, Cologne or Hanover, on the other hand, were proud of their rather radical approach to the modern town plan.

Recently some cities have given greater emphasis to re-evaluating historic structures and inner city areas. In Frankfurt, rebuilding from scratch the eastern row of buildings along Romerberg in medieval style in 1984 provoked an intense controversy in West Germany about reconstruction and the 'truth' of history. This is one of several examples of postmodern city design in Germany.

4. Urban Development Policy

The Athens Charter was published in Germany only in 1957 but its ideas were already 25 years old and, during these years, had more or less been the basis for urban development policies. Organizing the city according to four main functions (work, housing, leisure, transport) was from this point the planner's core strategy. The treatment of the neighborhood as the basic planning unit was introduced into German planning from 1940. News of Abercrombie's Greater London Plan of 1943/44 with its green belt strategy spread quickly, as did the reconstruction plan for Rotterdam. Thus, after 1948, a flourishing economy

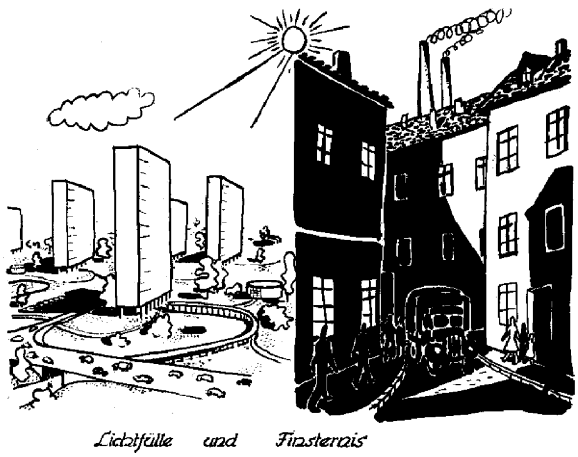

Figure 1
Marcel Lods, Werbung für die Forderungen der 'Charta von Athen.' Section du Plan (Mainz) 1947: Lichtfülle—Finsternis (Marcel Lods, promotion for the demands of the Athens Charter, (Mainz) 1947: richness of light—darkness)

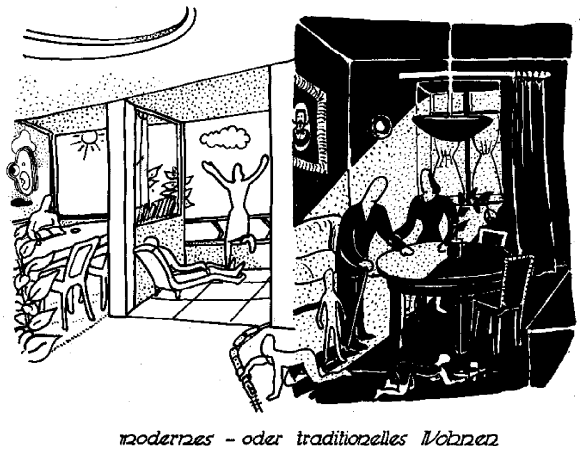

Figure 2
Marcel Lods, Werbung für die Forderungen der 'Charta von Athen.' Section du Plan (Mainz) 1947: modernes—traditionelles Wohnen (Marcel Lods, promotion for the demands of the Athens Charter, (Mainz) 1947: modern—traditional housing)

Figure 3
Marcel Lods, Werbung für die Forderungen der 'Charta von Athen.' Section du Plan (Mainz) 1947: Die Naturgegebenheiten—Sonne, Weiträumigkeit, Grünflächen (Marcel Lods, promotion for the demands of the Athens Charter, (Mainz) 1947: natural conditions—sun, space, green)

helped urban growth to develop to modern standards, based until the mid-sixties on the main concept of the 'organic art of town planning' (Reichow 1948), the 'composite and disaggregated city' (Göderitz, Rainer, Hoffmann 1957) or 'the car city' (Reichow 1959). The tenement house (*Mietskaserne*) and the noisy, crowded and dusty nineteenth-century industrial city were seen in a negative light (Figs. 1–3). The vision was of a green

city spread across the landscape, designed on a basic human scale as neighborhoods each with its own facilities, located in zones, suburbs or quarters. Motor car access was universal, but separated from walkways and rail systems, with a rich variety of open green spaces for leisure pursuits. The city center and central business district (CBD) were the dominant feature, incorporating reconstructed historic monuments and some high-rise buildings as the crown of the city.

The comprehensive urban renewal of nineteenth-century housing stock meanwhile aggravated the need for new housing, necessitating a change to a new policy of high-density housing in the periphery. Density outside the city also needed to be increased substantially to make public transport schemes efficient, and dreary suburban neighborhoods should, based on the perceptions of Jane Jacobs, or Hans Paul Bahrdt in Germany, be amended by 'urbanity,' by the transfer of urban design and dimensions to the periphery, without its lamentable deficiencies and defects.

Around 1975, severe economic changes stemming from the oil crisis and new technologies demanded an urgent review of planning aims and procedures. Several new arguments had already influenced planning—participation, ecology, the rediscovery of the nineteenth-century city and the attractiveness of its tenement house, and the ineffectiveness and inflexibility of large-scale planning. Planning was put under the microscope, necessitating reflection on, and alteration to, its theories, strategies, methods, and procedures. The new planning culture had to work with new technologies and economies, greater social mobility and the breakup of society and as a result new urban models developed. This discussion is still in progress.

5. *Rebuilding the Cities: The Eastern Approach*

Without a doubt, East Germany did things differently. Apart from Berlin, Dresden, Magdeburg, Leipzig and Chemnitz and a few others, many cities in the East were not touched by air raids. The building stock and inner areas of these cities therefore still date more or less from Hanseatic or court tradition as well as from the nineteenth and early twentieth centuries.

The German Democratic Republic (GDR) after its founding in 1948 abolished the private market, nationalized land and built structures, while the government took responsibility for their regulation. But the priority of the new socialist state was the improvement of its own economy to provide the population with basic goods and to make the productive sector independent from the Western market. Money for housing was scarce. Projects of the mid-1950s, like the *Stalinallee* in East Berlin (riposted in West Berlin with the International Building Exhibition, the 'Interbau' at Hansaviertel, Tiergarten with buildings by internationally renowned architects such as Aalto or le Corbusier), or the construction of the new industrial town of Stalinstadt (later Eisenhuttenstadt) (May in press), were outstanding and proved to be excellent propaganda. As early as 1950, however, East Berlin developed a list of 16 principles of town planning as a socialist riposte to the Athens Charter, at the insistence of Moscow. According to these principles, a city had to be compact and of a traditional national, or regional style.

Only after Stalin's death was the GDR able to adopt a functional town plan and a modern language in architecture. This led to industrialization in the building sector. Lack of building materials and a dramatic shortage of housing forced the government to rationalize the building process radically. Nationally organized housing production units with regional competence, were from the early 1960s responsible for planning and construction. The planning competencies of local authorities were limited. Urban reconstruction under these circumstances meant erecting new prefabricated housing schemes in rows of multistorey blocks on the urban fringe and moving people to them. The existing building stock was not renewed as no building material or a sufficiently skilled labor force were available. Private initiative was therefore more or less impossible. It was intended to replace the existing but decaying late-nineteenth century urban housing stock at a later date with new prefabricated blocks. Inspired by West Germany, a few test projects started around 1980 and some modified concrete slabs were produced to integrate better into the urban fabric. This proved too expensive. The policy to replace old housing stock as quickly as possible with blocks of prefabricated concrete units also met with limited success. The reprehensible treatment of the urban environment and the appalling decay of the cities was at the root of the people's movement in 1989.

See also: Accidents, Normal; Disasters, Sociology of; National Socialism and Fascism; Reconstruction/Disaster Planning: Japan; Reconstruction/Disaster Planning: United States; Transportation: Supply and Congestion

Bibliography

Beyme von K 1987 *Der Wiederaufbau. Architektur und Städtebaupolitik in beiden deutschen Staaten*. Pieper, Munich, Germany

Diefendorf J M 1993 *In the Wake of War: The Reconstruction of German Cities After World War II*. Oxford University Press, New York, pp. 3–17

Durth W 1986 *Deutsche Architekten. Biographische Verflechtungen 1900–1970* (German Architects. Biographical Interlacings 1900–1970). Vieweg, Braunschweig/Wiesbaden, Germany

Durth W, Gutschow N 1988 *Träume in Trümmern. Planungen zum Wiederaufbau zerstörter Städte im Westen Deutschland 1940–1950* (Dreams in rubble. Plans for the reconstruction of destroyed cities in West Germany 1940–1950) 2 vols. Vieweg, Braunschweig/Wiesbaden, Germany
Durth W, Düwel J, Gutschow N 1998 *Ostkreuz, Aufbau, Architectur und Städtebau der DDR* (Architecture and town planning of the GDR) 2 vols. Campus, Frankfurt, Germany
Fischer F K 1990 German reconstruction as an international activity. In: Diefendorf J M (ed.) *Rebuilding Europe's Bombed Cities*. Macmillan, Basingstoke, UK, pp. 131–44
Göderitz J, Rainer R, Hoffmann H 1957 *Die gegliederte und aufgelockerte Stadt*. Wasuruth, Tübingen, Germany
May R in press Planned city Stalinstadt: A manifesto of the early German Democratic Republic. *Planning Perspectives*
Petz von U 1990 Urban renewal under National Socialism: Practical policy and political objectives in Hitler's Germany. *Planning Perspectives* **5**(2): 169–87
Petz von U 1997 Ruhr Basin 1920 In: Bosma K, Hellinga H (eds.) *Mastering the City*, 2 vols. NAI, Rotterdam, The Netherlands, vol. I, pp. 56–65, vol. II, pp. 184–91
Reichow H B 1948 *Organische Stadtbaukunst, von der Großstadt zur Stadtlandschaft*. Westermann, Braunschweig, Germany
Reichow H B 1959 *Die autogerechte Stadt. Ein Weg aus dem Verkehrschaos*. Otto Maier, Ravensburg, Germany

U. von Petz

Reconstruction/Disaster Planning: Japan

Japan has been prone to natural disasters, especially as it lies in the Pacific volcano zone, but among the most destructive have been frequent earthquakes. Moreover, its traditional urban dwellings, typically wooden buildings made with thatched roofs, are easily inflammable, making cities liable to major conflagrations. But it was not until the Meiji Restoration in 1868 and the subsequent era of rapid industrialization and urban growth that serious thought was given to constructing urban structures to survive such disasters. Since then, the development of reconstruction and disaster planning has been interrelated closely with that of modern urban planning, which in turn has been influenced by changes in the political, economic, and social environment.

1. Reconstruction/Disaster Planning in the Meiji Period

One of the earliest examples of reconstruction and disaster planning after the Meiji Restoration was the *Ginza district project* following the large 1872 fire in Tokyo that burnt out some 3,000 dwellings. A radical proposal was put forward to rebuild the entire district with Western-style buildings made of bricks. It was an attempt to set a precedent in tranforming the traditional urban structure of the capital into a modern one, reflecting the Meiji government's strenuous efforts to swiftly Westernize the country in order to facilitate the revision of the unequal treaties with the Western powers. Individual buildings were designed by expert foreign advisers employed by the government, and the project was completed in 1877. Although the financial difficulties allowed just half the planned area to be rebuilt with bricks, the newly reborn Ginza district quickly became popular as the most fashionable shopping center in the country.

However, there was a gradual realization that a more comprehensive and yet realistic approach than this kind of extravagant but essentially superficial Westernization of the capital was necessary. While this recognition led to the urban improvement projects for Tokyo, mainly consisting of road, water supply, and drainage works, that started in 1888 and lasted for 30 years, there were important prior developments in the form of reconstruction and disaster planning. In the wake of four big fires within a few months during the winter of 1880–81, the governor of Tokyo proclaimed regulations mandating fire-prevention measures in the central wards of Tokyo. Twenty-two streets and canals were designated as areas along which buildings had to be constructed of either bricks or stone, or as godowns made of clay, with roofs of incombustible materials. Extensive slum clearance was also carried out in the Kandahashimotocho district following one of the four big fires in 1880–81, though it meant the merciless uprooting of inhabitants from their old community, admittedly one of the major slums in Tokyo at that time. Meanwhile, the demand for national planning legislation intensified, and the *Urban Planning Act* was enacted in 1919. In practical terms, its prime concern was to check indiscriminate urban expansion by controlling future development through such measures as land readjustment projects, a principal feature of Japanese urban planning based on street widening and realignments. But an opportunity for its application to built-up areas soon came about in a tragic way.

2. Reconstruction after the Great Kanto Earthquake

On September 1, 1923, a severe earthquake of magnitude 7.9 hit the southern Kanto region including Tokyo and Yokohama. As it occured just before noon, when people were cooking meals, the Great Kanto Earthquake caused massive fires that claimed more than 100,000 lives and destroyed 460,000 dwellings. The total damage amounted to approximately six billion yen, four times the amount of the national budget. The government set to reconstruction planning immediately, with Shinpei Goto, the Home

Minister, as the man in charge. He tackled the task enthusiastically, not least because he saw an opportunity to restore his reputation. The renovation plan for Tokyo made public five months earlier when he was the mayor of Tokyo had been criticized heavily as too extravagant, and he had been derided as a braggart. Within a week, he established basic principles for reconstruction, which included proposals for a powerful Ministry for Reconstruction and public purchase by the state of all land affected at an estimated cost of 3 billion yen, reflecting his determination to rebuild in a bold and comprehensive way. However, neither of these proposals was put into practice: opposition from other ministers and ministries to a mighty reconstruction ministry led to the abandonment of the idea. Public purchase of land was out of the question not only on financial grounds but also on political ones, for the landed interests had a strong say within the government as well as in the Diet. Moreover, the budget was reduced to 468 million yen at the 47th extraordinary Diet session in December, three months after the disaster.

Nonetheless, reconstruction projects with actual working expenses of 847 million yen over seven years between 1924 and 1930 meant grand engineering on an unprecedented scale. This becomes clear when they are compared with the earlier urban improvement projects for Tokyo which took 30 years but cost only 42 million yen. Approximately 80 percent of the afflicted area in Tokyo, 3,600 hectares, were covered by reconstruction land readjustment projects under the *Ad Hoc Special Urban Planning Act*, which made possible large-scale land readjustment in built-up areas. As a result, Tokyo's urban structure, of which the core had been Edo-era construction modified by the Meiji-era urban improvement projects, was transformed. This change laid the foundation for the infrastructure of the capital's central area until the 1960s. The reconstruction land readjustment projects gave rise to 52 principal roads (114 kilometers in total length), 122 subsidiary principal roads (139 kilometers), 55 parks (42 hectares), and a number of structures built of reinforced concrete to render them earthquake- and fire-proof. The latter included 121 schools and 424 bridges. Of particular interest were the activities of *Dojunkai*, an extra-departmental body to the Home Ministry set up in 1924. It provided, among other things, 2,501 apartments of reinforced concrete with the most up-to-date communal facilities, built on 15 estates between 1926 and 1933. These became the forerunner of the public apartments built to alleviate major urban housing shortages in the postwar years.

It should be also noted that reconstruction planning after the Kanto earthquake brought important advances in planning techniques and the training of technocrats and other planning personnel. This contributed to the substantial progress made in the development of Japanese urban planning in the 1930s and 1940s. However, the main cause of this progress, the intensification of military preparations, was soon to lead to a war that would bring massive destruction to the country.

3. *Air Defense Planning and Postwar Urban Reconstruction*

During the 1930s, air defense became the prime concern of Japanese urban planning, as was duly shown in the *revised Urban Planning Act* of 1940, which made air defense an integral part of urban planning. Emphasis was placed upon the provision of open spaces and wider streets for fire-preventive purposes. Admittedly, urban planning benefited from this. For instance, the state purchase of an area of 13,730 hectares was decided in 1939 to provide a number of parks as air defense measures. Out of these, six large parks in Tokyo with a total of 640 hectares were established as part of the ceremony commemorating the 2,600th anniversary of the founding of the Japanese Empire. Evacuation of buildings to make room for air defense roads also led to the building of wide thoroughfares in cities such as Nagoya, Hiroshima, and Kyoto.

However, the sheer intensity of air raids during the closing months of WWII was far more devastating than these measures could handle, and Japan found its urban areas in ruins when it surrendered in August 1945. One hundred and fifteen cities and towns throughout the four main islands in Japan were officially designated in November 1945 by the government as war-damaged cities. Approximately 2,300,000 houses or 20 percent of the stock in those cities had been destroyed or seriously damaged, and 63,000 hectares of land burnt out. The damage was particularly extensive in the two largest cities, Tokyo and Osaka, and these accounted for nearly 40 percent of the total of land (26.8 percent for Tokyo and 12.1 percent for Osaka) and more than 45 percent of houses (32.9 percent for Tokyo and 13.8 percent for Osaka) affected. The area affected in Tokyo was more than four times as great as that afflicted in the Great Earthquake in 1923.

Despite the extent of damage and the fact that Japan was under American occupation, official policies and plans for postwar reconstruction were formulated speedily and implemented with determination. Evidently, such destruction was regarded as having provided an opportunity for drastic urban renovation. In November 1945, the War damage Rehabilitation Board (WRB; *Sensaifukkoin*) was set up as a responsible central body. Ichizo Kobayashi, a well-known businessman, was appointed as its first president and given ministerial status. The intention underlying his appointment was to introduce the spirit of private enterprise and thus to get rid of the bureaucracy that had dominated Japanese urban planning before the war. This reflected the wave of

democratization of Japan's political, economic, and social systems. The following month, the cabinet adopted the Basic Policy Principles for the Planned Reconstruction of War damaged Areas (*Sensaichi-fukkokeikaku Kihonhoshin*). It featured nine principles in planning, according to which the local authorities drew up plans very quickly. It was recommended that the damaged areas should be covered by land readjustment projects, and that principal roads should be quite wide, even as much as 100 meters for large cities.

However, the planning process continued to be dominated by civil servants. For one thing, the WRB officials simply ignored Kobayashi, their superior, when he argued for compact rather than grandiose reconstruction. The officials insisted that plans had to be bold and comprehensive. This was perhaps in line with what cities like Tokyo wished, but many other cities thought quite differently, only to be forced to accept the officials' arguments. Osaka, for instance, had no choice but to add a couple of 100-meter-wide thoroughfares to its original plan when it was told by the WRB officials that any approval of their plan would otherwise be out of the question. Osaka officials believed that such proposals were too extravagant given the country's economic situation. In fact, the plans were subject to constant cutbacks due to the economic difficulties, culminating in the Revision (*Saikento*) in 1949 of the basic principles for reconstruction drafted in 1945. The revision was made following the *Dodge Line*, a draconian anti-inflation program implemented by the occupation authorities in 1949. The *Dodge Line* forced the government to cut spending drastically. As a result, the area covered by land readjustment projects was reduced by approximately 30,000 hectares, or 50 percent over the country as a whole, and many ambitious road and park projects were eliminated. The cuts were especially severe for big cities. In the Tokyo plan, all seven of the proposed 100-meter-wide roads were simply abandoned, and the area for land readjustment projects was reduced by three-fourths. In the Osaka plan, proposals for 100-meter-wide roads were also dropped and the extent of land readjustment projects reduced by nearly 45 percent. Moreover, there was widespread public disapproval of reconstruction planning: in large cities, the open space proposals were often simply ignored by those who erected illegal houses on the designated sites, which thus became prejudicial to planned development in the future. In many provincial cities, there emerged organized opposition by citizens to the official plans, usually reflecting discontent that their opinions were not heard in the planning process.

4. Conclusions

Nevertheless, planning for disaster-proof urban structures made some progress from the 1950s onwards, thanks to a series of legislative measures aimed at enhancing public safety, such as the *Fireproof Building Promotion Act* of 1952 (covering the total area of 64 hectares in some 80 cities), and the *Disaster-proof Building Blocks Construction Act* of 1961 (covering 253 hectares in 101 cities). Further, the *Urban Renewal Act* of 1969 led to the redevelopment of six disaster-prevention districts covering 526 hectares in Koto Ward, a heavily populated downtown area in Tokyo. But the country is still vulnerable to disasters, as shown in the Hanshin/Awaji Earthquake in 1995. Regrettably, there was considerable anger among local residents in Kobe, the worst-hit city, about the lack of consultation in the reconstruction planning and the manner in which plans were imposed from above. As Japan cannot be immune from earthquakes, there is still an enormous need to plan for disaster-proof urban structures throughout the country, and to promote greater public participation in the process.

See also: East Asia: Environmental Issues; East Asian Studies: Economics; East Asian Studies: Politics; Japan: Sociocultural Aspects; Reconstruction/Disaster Planning: Germany; Reconstruction/Disaster Planning: United States

Bibliography

Fujimori T 1982 *Meiji no Tokyokeikaku*. Iwanamishoten, Tokyo
Fukuoka S 1991 *Tokyo no Fukkokeikaku*. Nihonhyoronsha, Tokyo
Hasegawa J, Mason T, Matsumura T, Tiratsoo N 2001 *Urban Reconstruction in Britain and Japan, 1945–1955*. Luton University Press, Luton, UK
Ishida Y 1987 *Nihonkindaitoshikeikaku no Hyakunen*. Jichitaikenkyusha, Tokyo
Koshizawa A 1991 *Tokyo no Toshikeikaku*. Iwanamishoten, Tokyo
Mikuriya T 1984 *Shutokeikaku no Seiji*. Yamakawashuppan, Tokyo
Zenkokushigaichisaikaihatsukyokai 1991 *Nihon no Toshisaikaihatsushi*. Jutakushinposha, Tokyo

J. Hasegawa

Reconstruction/Disaster Planning: United States

Reconstruction/disaster planning includes plans prepared specifically to guide reconstruction following disasters, plans prepared to mitigate losses from hazards through actions taken prior to the onset of disasters, and hazard-mitigation elements of more general urban and regional comprehensive plans. Although the content of each type of plan differs, the critical components and key features of the planning process are similar. In this article, the basic elements of reconstruction/disaster plans and the processes involved in preparing them are described.

1. The Problem

World-wide, the risks from natural and related technological hazards has been increasing at an accelerating pace. Population growth and migration toward at-risk areas, increased development and urbanization of hazard zones, long-term climactic trends, and other factors have combined to expose populations to losses of life and property. Although improved forecasting and warning systems have generally reduced the number of fatalities associated with disasters, particularly in developed countries, annual monetary losses everywhere are staggering.

This problem is particularly severe in the United States, where losses of property and income attributed to natural hazards are estimated at $500 million per week (Mileti 1999). Large disasters, such as Hurricane Andrew in 1992, the Midwest floods of 1993, and the Northridge earthquake of 1994 have resulted in damages ranging from $15 to $44 billion (Board on Natural Disasters 1999). Worst case scenarios suggest that much larger disasters are probable. A $200 billion earthquake in California and a $50 billion hurricane in Florida are distinct possibilities. In the face of this large and growing exposure to economic loss, policy makers are slowly coming to realize that urban planning may be an effective way to foster greater resilience to natural disasters.

2. Evolution of Planning for Resilience

Throughout history, communities have been located so as to limit the potential for disaster and after disasters, governments have sought to rebuild in ways to reduce the likelihood of future catastrophes. Italian hill towns, for example, were located on high ground to avoid flooding and diseases such as malaria that were prevalent on the valley floors. After the Great Fire of London in 1666, Charles II issued a proclamation that called for rebuilding in ways that would reduce the potential for future disasters. In more recent times, however, the lessons of the past were forgotten. After the San Francisco earthquake and fire of 1906, for example, the city was quickly rebuilt with little attention to hazard mitigation. In fact, some of the largest losses in the Loma Prieta earthquake of 1989 occurred in the Marina District, which is located on debris pushed into the bay after the 1906 disaster.

Modern thinking about planning for resilience to hazards can be traced to the geographer Gilbert White, who first proposed the concept in 1936 in an article in the *Planners Journal* (1936). White's ideas were not acted upon until twenty years later, when the Tennessee Valley Authority prepared a flood-plain management plan for Bristol, Tennessee-Virginia in 1956. However, because few communities followed the TVA lead, state governments, beginning with California in 1972, began to require that local governments incorporate hazard mitigation provisions in their comprehensive planning. At last count, half of the states mention natural hazards in their planning enabling statutes, and eleven states, like California, actually mandate local government planning for the mitigation of natural hazards (Schwab et al. 1998).

Plans that address natural hazards take two forms. Post-disaster reconstruction and recovery plans focus on policies and actions to bring about rapid recovery from a disaster and to guide reconstruction of damaged structures and infrastructure to increase resilience to future disasters. Hazard mitigation plans focus on policies and actions that can prevent damage and loss of life before disasters occur. Each type of plan may be formulated as a stand-alone plan or as an element of a broader comprehensive plan, but the latter approach is probably best, since it embeds hazard mitigation in regular planning processes that have legal standing.

3. Post-disaster Reconstruction and Recovery Plans

The post-disaster period creates a window of opportunity to make changes that can improve resilience to future disasters and accomplish a number of other community goals. But, these opportunities are often lost due to the need to make decisions quickly in an atmosphere of confusion and strong political pressures to return quickly to normalcy. Post-disaster reconstruction and recovery plans provide a means to ensure that reconstruction occurs in ways that contribute to long-term sustainability. These plans focus on decisions, made in the short-term recovery period after a disaster, that have long-term repercussions. Schwab and his associates (1998) call attention to four groups of decisions that are particularly important to address in plans: (1) sites for temporary housing, relocation of damaged businesses, and dumping of debris; (2) closure of roads and bridges; (3) restoration or relocation of critical infrastructure; and (4) reconstruction or relocation of dwelling units and nonconforming uses that have suffered substantial damage. In addition, the plan should clarify roles and responsibilities in recovery, identify potential sources of financing, and minimize wasteful and conflicting efforts (Mileti 1999).

Experts such as Mileti (1999) and Schwab (1998) argue that postdisaster recovery and reconstruction plans will be most effective if they are prepared in accordance with the following principles. *Consensus building* through community involvement should be sought to avoid community conflict over reconstruction strategies. *Information* should be provided about the hazard, exposed populations, buildings, and infrastructure, resources available for recovery after the disaster, and governmental powers and programs to aid recovery. *Organization* should be addressed through identification of topic-specific groups that can

be put in place immediately after a disaster (in particular, designation of an official rebuilding/restoration group is critical). *Procedures* should vary as little as possible from those followed during pre-disaster periods, although permitting and code-review procedures should be streamlined in order to expedite rebuilding. *Damage evaluation* protocols should be in place, such as protocols for damage assessment by building inspectors and standards for post-disaster repair and construction. *Finances* for reconstruction should be identified including funding from both governmental and private sources. Finally, provision should be made to adjust the local government's *planning style* so that more flexible and cooperative procedures replace the formal, deliberative rule-oriented procedures that characterize pre-disaster decision-making.

4. Hazard Mitigation Plans

The American Planning Association recommends in its *Growing Smart Legislative Guidebook* (1998) that local government comprehensive plans include an element on hazard mitigation. A number of states (e.g., California, Florida, North Carolina for coastal areas) not only recommend this, they require it of every community (see Burby et al. 1997). Through a comprehensive plan's safety element, hazard mitigation is addressed directly and hazard considerations are also included in portions of the plan dealing with housing, transportation, protection of environmentally sensitive areas, natural resource protection, and capital improvements. Hazard mitigation plans, as stand-alone documents, also have been prepared to address a variety of hazards. In either case, the preparation of a plan involves similar steps.

The first step is to secure a resolution by the local governing body to put someone in charge of the planning effort and to establish a planning committee or task force. Official sanction for the planning process is essential to obtain the participation of other agencies and to tap into their resources and expertise. Participating agencies should include planning/community development, engineering/public works, finance, public safety (police, fire, emergency management), code enforcement, parks and recreation, public information, and a representative of the governing board.

One of the planning group's first jobs is to build a constituency for hazard mitigation and consensus about the need to take action. The lack of such a constituency, coupled with local officials' own lack of concern about hazards, has resulted in minimal attention to hazards in past planning (e.g., Burby 1998, Mileti 1999, Platt 1999). This is best accomplished by involving key stakeholders in the planning process from the beginning. The public involvement effort should target people and organizations exposed to hazards, neighborhood and civic organizations, the business community, managers of critical facilities, land developers and others whose decisions affect urban development. In addition to gaining support for the planning effort, citizen involvement helps secure local knowledge about hazards, insures that information gathered in the planning process is relevant to issues citizens view as important, and helps to prevent misunderstandings and gain consensus for the plan's recommendations.

The next step in the planning process is to develop information about the problems hazards pose and the capacity of the community to build resilience to hazards. Hazard assessment includes identifying hazards (magnitude, severity, frequency, and causative factors) and mapping the areas of a community that are exposed to them. Losses experienced in past disasters are identified, and vulnerability to loss in future disasters is estimated by examining potential property damages in hazard events of various magnitudes.

Analysis of capacity helps in identifying hazard reduction measures that are feasible for a community and actions that are needed to increase capacity, such as adding staff with hazards expertise or developing dedicated revenue sources to fund mitigation. Other topics for capacity analysis include funding sources, legal authority for hazard mitigation actions, and evaluation of hazard mitigation measures that are already in use. The hazard assessment and capacity analysis lead to a statement of the problem facing the community and formulation of goals and objectives. It can be combined with a visioning process to develop, in collaboration with citizens and policy makers, an image of what a hazard resilient community would look like in the future. Visioning can help develop consensus about the goals of hazard mitigation and can be useful in tying hazard mitigation to other community concerns such as economic development and environmental protection.

Once the problem is understood and a vision of a hazard resilient community is agreed upon, preparation of the hazard mitigation plan involves making choices among possible strategies to build resilience. Six sets of possible measures should be reviewed.

(a) *Preventive* policies and actions are designed to limit the exposure of new development to losses from hazards (e.g., building code standards, zoning, and land acquisition).

(b) *Property protection* policies and actions limit the exposure of existing development to hazards on a parcel-by-parcel basis (e.g., retrofitting, relocation, acquisition and clearance, and insurance).

(c) *Structural protection* policies and actions limit the exposure of existing development to hazards (principally flooding) on an area-wide basis (e.g., flood control dams, levees, sea walls, and channel modifications).

(d) *Emergency services* policies and actions lessen the impact of a hazard after its onset (e.g., warning,

evacuation, flood fighting, fire fighting, and emergency sheltering).

(e) *Natural resource protection* policies and actions preserve and restore natural areas that also help lessen the magnitude of hazards (e.g., wetland protection, erosion and sedimentation control, and sand dune protection).

(f) *Public information* policies and actions build awareness of hazards and knowledge of individual actions to mitigate them (e.g., hazard maps, library resources, and technical assistance).

The strategies that have been identified as useful in creating a sustainable community that is resilient to hazards are evaluated in terms of their effectiveness, efficiency, equity, and feasibility (administrative, technical, financial, legal, and political). The goal is to formulate a program that draws wide support, addresses existing problems from hazards, and also prevents new ones from developing. The final plan includes, at a minimum, five components. These include

(a) a description of how the plan was prepared and who participated in its preparation,

(b) an assessment of hazards facing the community,

(c) the vision of a hazard-resilient community,

(d) recommendations of policies and actions, and

(e) identification when they will be undertaken, who is to undertake them, how they will be financed, and who will monitor performance (see Wetmore and Jamieson 1999).

5. Conclusion

The preparation of a post-disaster-reconstruction/hazard-mitigation plan provides the technical analysis, coordination among governmental agencies, and community involvement and consensus building necessary to increase resilience to natural and related technological hazards. Communities with coherent plans are able to build human settlements that are resistant to disasters and able to recover from disasters quickly. The plan-making processes and principles discussed in this article are proven to be effective. But state and local governments have been slow to prepare and adopt the needed plans. The challenge for governments at every level is to act on the knowledge that is available to reduce or even reverse the steady rise in destruction and human suffering from disasters.

See also: Accidents, Normal; Disasters, Sociology of; National Socialism and Fascism; Reconstruction/Disaster Planning: Japan; Reconstruction/Disaster Planning: United States; Transportation: Supply and Congestion

Bibliography

American Planning Association 1998 *Growing Smart Legislative Guidebook*. Chapter 7. Local Comprehensive Plans. American Planning Association, Chicago

Board on Natural Disasters, National Research Council 1999 Mitigation emerges as major strategy for reducing losses caused by natural disasters. *Science* **284**: 1943–7

Burby R J (ed.) 1998 *Cooperating with Nature: Confronting Natural Hazards with Land-use Planning for Sustainable Communities*. Joseph Henry Press, Washington, DC

Burby R J, Beatley T, Berke P R, Deyle R E, French S P, Godschalk D R, Kaiser E J, Kartez J D, May P J, Olshansky R, Paterson R G, Platt R H 1999 Unleashing the power of planning to create disaster resistant communities. *Journal of the American Planning Association* **65**(3): 247–58

Burby R J, May P J, Berke P R, Dalton L C, French S P, Kaiser E J 1997 *Making Governments Plan: State Experiments in Managing Land Use*. Johns Hopkins University Press, Baltimore, MD

Godschalk D R, Kaiser E J, Berke P R 1998 Integrating hazard mitigation and local land use planning. In: *Modernizing State Planning Statutes: The Growing Smart Working Papers*, Vol. 2. PAS Report 480–1, pp. 57–81. American Planning Association, Chicago

Mileti D 1999 *Disasters by Design: A Reassessment of Natural Hazards in the United States*. Joseph Henry/National Academy Press, Washington, DC

Platt R H 1999 *Disasters and Democracy: The Politics of Extreme Natural Events*. Island Press, Washington, DC

Schwab J, Topping K C, Eadie C C, Deyle R E, Smith R A 1998 *Planning for Post-disaster Recovery and Reconstruction*. PAS Report 483–484. American Planning Association, Chicago

Wetmore F, Jamieson G 1999 Flood mitigation planning: The CRS approach. *Natural Hazards Informer* **1 (July)**: 1–15

White G F 1936 Notes on flood protection and land use planning. *Planners Journal* **3**: 57–61

R. J. Burby

Reconstructive Memory, Psychology of

Reconstructive memory refers to the idea that retrieval of memories does not occur in some completely accurate form, as a video-recorder might replay a scene, but rather that recollection of memories is a process of trying to reconstruct (rather than replay) past events. Although the reconstruction can be quite accurate, the processes responsible can also introduce errors during retrieval. In fact, systematic errors in memory are the primary evidence for its reconstructive nature. This article provides an overview of the evidence that remembering is reconstructive (see also *Eyewitness Memory: Psychological Aspects*; *False Memories, Psychology of*; *Memory for Meaning and Surface Memory*).

1. Background

The hypothesis that remembering should be viewed as reconstructive dates to an important book by Sir Frederic Bartlett (1932). Bartlett contrasted reproductive memory (veridical, rote forms of memory, such as reproducing a telephone number) with recon-

structive memory and argued that the latter was more typical of our uses of memory outside laboratory and educational circumstances. That is, we are rarely faced with the task of remembering something exactly the way it happened, but more typically need only to get the essence of the event right. If a friend asks you, 'What did Kathleen tell you last night?' the request is not for a literal rendering of last night's conversation, but rather for the gist of what was said.

Bartlett argued that perceiving and comprehending events do not simply happen automatically, but that every event of comprehension involves the mental construction of one's understanding of the event in the world. We all struggle with the 'effort after meaning' in comprehending the events in the world around us. If encoding or perceiving is a construction, then when one wants to recall the events later, the attempt is to reconstruct the event.

Bartlett argued that recollection is guided by schemas, or general organizing structures, which aid encoding and retrieval. If you see a scene at the beach and are asked to recall it later, you might recall seeing a beach umbrella even if none was present in the actual scene itself, because it is consistent with the general schema of items that belong in a beach scene. Therefore, although schema can aid encoding and retrieval of information, they can also lead to errors.

A large amount of research is consistent with the idea that remembering is reconstructive. The standard textbook account holds that certain forms of remembering are reconstructive whereas others are reproductive. However, a strong case can be made that all remembering is reconstructive. What appears to be reproductive memory occurs in situations in which the reconstruction is quite accurate (Roediger and McDermott 1995). This article considers various forms of memory as they are experimentally studied and discusses evidence for reconstructive processes at work.

2. Sensory and Short-term Memory

The greatest challenge to the claim in the previous paragraph perhaps comes from studies of sensory memory, the borderline between perceiving and remembering. However, many theorists have argued that perceiving is itself a constructive activity, with fascinating phenomena of visual and auditory illusions used as evidence for this claim (see Gregory 1998, among many others). Sensory memory refers to the temporary persistence of information that has struck the senses, which lingers briefly as it is being comprehended. Visual persistence is called iconic memory and auditory persistence is labeled echoic memory. It would seem that iconic memory—essentially a fleeting after-image of the scene from the outside world—would surely be a form of reproductive memory. Yet even in these situations errors arise, showing that at least the retrieval processes from this type of memory involve reconstruction. For example, in Sperling's studies of iconic storage, in which people had to report letters that they had briefly seen on a screen, a common error when people missed a letter was to report another letter than either looked like or sounded like the original letter (Sperling 1967). This type of error indicates that people may code even such simple items as single letters into visual patterns and associated sounds. When people miss a *b* in these experiments, they may substitute a *v* (which sounds like *b*) or a *p* or *d*, which share both similar appearance (a long line and a curve) and similar sound (the letters rhyme with *b*). Even reports from iconic memory may show reconstructive tendencies.

Short-term memory or working memory (see related entries on these topics) hold information for longer than sensory memories, but people are still ordinarily accurate in retrieving information from short-term stores, if no interference occurs. Does this accuracy reflect a rote, reproductive process? The answer seems to be no, because when the short-term memory system is challenged by having people operate under, for example, fast rates of presentation, errors occur. Errors are often (but not always) phonological in nature. That is, if someone tries to recall letter strings and misses a letter, similar sounding letters are confused (Conrad 1964). In the case of words, those that share visual and phonemic (sound-based) features are confused (Crowder 1976, Chap. 4). Therefore, even though short-term memory processes are often considered quite accurate (and they can be), recall in these situations typically occurs under conditions (short unfilled delays between study and test) that make for accurate reconstructions. Stress the system by presenting material fast, or by creating interference, and the characteristic error pattern indicative of a reconstructive process appears.

3. Long-term Memory

By far the greatest amount of research on reconstructive memory has occurred in the domain of long-term, episodic memory. Many different kinds of experiments have been conducted and we will consider some of the main types here, which have used a wide variety of materials: word lists, pictures and scenes, sentences, prose passages, and videotape. Research with all these materials reveals remembrance to be quite prone to error and to filling in the gaps.

3.1 Word Lists

Roediger and McDermott (1995) developed a paradigm in which people heard lists of 15 words that were generated as associates to another word, which was not itself presented. So, for example, subjects heard

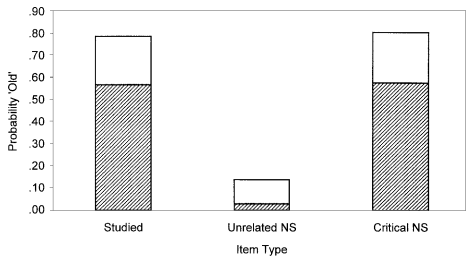

Figure 1
False recognition and false remembering. Mean probability of recognizing (total bar) and remembering (cross-hatched bar) studied words and lures. Unrelated lures were rarely recognized, but critical lures were recognized and remembered at about the same levels as studied words. People remember the occurrence of the critical items at the same level as words that were actually studied. Data are from Roediger and McDermott (1995, Experiment 2)

the list 'hot, snow, warm, winter, ice, wet, frigid, chilly, heat, weather, freeze, air, shiver, Arctic, frost.' All these words are associated to the word 'cold,' which was the generator of the list, but 'cold' was not presented. Immediately after hearing the list, subjects were asked to recall the items and were urged to be sure that the words they recalled had actually been presented in the list. Despite this warning not to guess, Roediger and McDermott (1995) found that their subjects recalled missing words like 'cold' at very high levels; in fact, probability of recalling the critical item often equals probability of recall of list items in this paradigm. After recall of many lists, subjects received a recognition test in which they were instructed to pick out only items that had actually appeared in the list. The test included words from the list ('winter,' 'chilly'), completely unrelated words ('automobile'), and the critical words such as 'cold.' Subjects examined each word and, if they judged it to be old (or studied), they were asked to make a further remember/know judgment. This judgment (Tulving 1985) asks people to report if they can remember the moment of the item's occurrence in the study list (a remember judgment) or if they just know it was on the list but cannot remember its moment of occurrence (a know judgment).

The recognition results are shown in Fig. 1, where it can be seen that subjects were good at distinguishing studied words from unrelated lures or distracters. However, the critical lures such as 'cold' were recognized at the same level as were the words in the list, about 80 percent. People recognize the related but nonstudied word at the same level as they do items in the list. In addition, subjects claimed to remember the moment of occurrence of the critical words (that, of course, were not actually presented) at the same level as for the words that were actually studied). In short, in this paradigm subjects recall, recognize, and claim to remember the occurrence of words that were never presented at the same level as of the words that really were presented, which provides a compelling demonstration of illusory memory. The implication is that subjects reconstruct the list using their general knowledge and yet insert a quite specific memory—they remember 'cold' as having occurred—because it fits with the general theme or schema of the list.

3.2 Pictures and Scenes

Miller and Gazzaniga (1998) developed a similar procedure with common pictorial scenes, ones that had been covers of the *Saturday Evening Post* magazine. They showed people the scenes with a critical object from the scene removed. To use our earlier example, people might view a beach scene with a beach umbrella either present in the scene or absent from it. Miller and Gazzaniga (1998) found that when people were given recognition tests on objects in the scene, they were very likely to recognize the object as having been present when in fact it was not. Again, in reconstructing the scene, the object is inserted and seems part of the memory.

3.3 The Misinformation Paradigm

One of the most popular paradigms for studying reconstructive processes in memory was developed by Elizabeth Loftus and her colleagues (e.g., Loftus et al. 1978) and is called the misinformation paradigm. It is modeled after the events that an eyewitness to a crime might experience. In the lab, subjects view a set of slides or a videotape, then later they receive questions or a narrative. In Loftus et al. (1978), subjects witnessed an automobile accident that occurred at an intersection where a car failed to stop at a stop sign. In the later series of questions, subjects would be asked either one of three types of question: 'Did the car stop at the traffic sign?' (control condition), 'Did the car stop at the stop sign?' (consistent information condition) or 'Did the car stop at the yield sign?' (inconsistent or misleading information condition). This question served as the experimental manipulation for a question asked at a later point in time: 'What kind of sign was at the intersection, stop or yield?'

Results of this experiment are shown in Fig. 2, where it can be seen that, relative to the control condition in which the question was posed in a neutral manner, wording of the question that was consistent with the original information improved recognition whereas the conflicting information reduced recognition. This last finding is the misinformation effect —wrong information received after an event is often incorporated into one's memory for the event. This misinformation effect has been demonstrated in

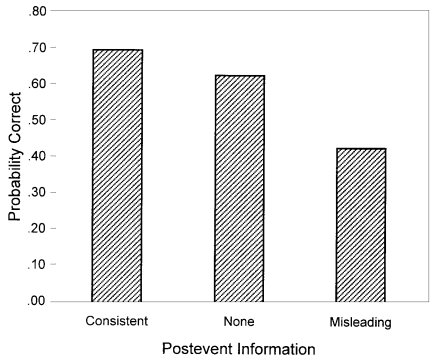

Figure 2
Probability correct (on a two-alternative forced choice recognition test) as a function of whether subjects had received information consistent with what they had previously seen, information inconsistent with what had been seen, or no specific postevent information. Misleading information after the event can reshape memory for the event. Data from Loftus et al. (1978, Experiment 2)

hundreds of experiments, although the magnitude of the effect depends on the exact nature of the test.

The general principle to arise from experiments on the misinformation effect is that memory for an event is not encapsulated in time the way the event itself is. Rather, information provided after the event can modify our memories for the event itself. Consistent information improves our later reconstruction, whereas conflicting or misleading information is harmful. When we are trying to reconstruct events from our past, we draw on all the information and resources at our disposal and the reported memory may include 'facts' that were not really part of the event, but incorporated into it later. Practically, if police or other people ask suggestive questions or embed wrong information into questions asked of witnesses after the event, then the witnesses' later recollections and testimony might be tainted (see *Eyewitness Memory: Psychological Aspects*).

3.4 Prose Recall

Perhaps the greatest body of research demonstrating reconstructive processes in retrieval has been carried out in prose recall. Brewer (1977) examined pragmatic implications in retention of sentences. A pragmatic implication is made when the person reading or hearing a sentence infers something that is neither explicitly stated nor logically implied by the sentence (see *Inferences in Discourse, Psychology of*). For example, 'The karate champion hit the cinder block' implies that the block was broken. However, this is a pragmatic implication; it is perfectly possible that the block was struck but not broken. After people studied sentences of this type mixed with other sentences, they were given the first parts of the sentences as cues to recall the entire sentence. Brewer showed that in recall of sentences with pragmatic inferences, people were more likely to recall the implied verb than the verb actually used in the sentence. When reconstructing the action, they remember the champion as breaking the cinder block, not just hitting it. This result reflects a kind of inferential momentum, with people going beyond the literal truth of a statement to infer the events that probably occurred next. Often such inferences may be correct, but they are not necessarily so. When the inference is strongly made, but false, it can generate a false memory.

Sulin and Dooling (1974) also studied people's remembrance of implications as facts. In one case, they presented subjects with short paragraphs about a troubled girl and later tested for recognition both of ideas from the paragraph and of nonpresented ideas that were either consistent or inconsistent with the theme of the passage. The paragraphs studied by two groups of subjects were identical, with the exception of the name above the story. In one case it was Helen Keller and in the other case it was Carol Harris. One item on the recognition test asked subjects if the following sentence had appeared in the narrative: 'She was deaf, dumb, and blind.' Students who had read the story about Helen Keller frequently made the error of saying that this sentence had appeared in the passage, whereas those reading about Carol Harris almost never made this error. In trying to reconstruct whether the sentence may have been in the story, the knowledge about Helen Keller caused subjects to fill in what 'must have been' in the story and led subjects erroneously to conclude that the sentence was there.

Sometimes people remember the sentences that were in a prose passage no better than sentences they have never read before, so long as the new sentences are consistent with the material that was actually presented. Bransford and Franks (1971) gave subjects short sentences such as 'The rock rolled down the hill' or 'The rock crushed the hut.' There were four simple ideas like this that, when put together, would make up the complex sentence: 'The rock rolled down the hill and crushed the tiny hut.' The experimenters gave their subjects various sentences to build up the complex idea unit. They later tested subjects by asking them to pick out sentences they had actually studied from among other sentences. Two types of distractor sentences were used: some were consistent with the idea unit but had not been studied, whereas other sentences were unrelated to the idea unit. The test sentences themselves could vary in the number of idea units present, from one to four. Subjects rated their belief that the sentences had been previously studied on a 10-point scale, from -5 (sure the sentence was new or nonstudied) to $+5$ (sure the sentence was old or studied). The sentences that were not consistent

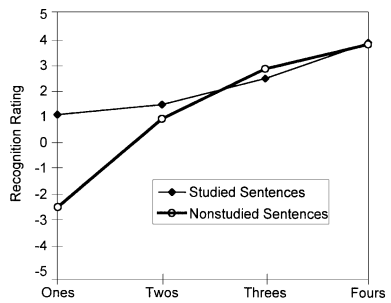

Figure 3
Mean recognition ratings for studied and nonstudied sentences as a function of number of propositions per sentence. Subjects could not distinguish sentences actually presented from those that had not been presented but that were consistent with the ideas. Data from Bransford and Franks (1971)

with the idea units were accurately rejected as not having been studied.

The results for the other two types of items are shown in Fig. 3 and reveal that the more idea units embedded in the test sentence, the higher the confidence rating that the sentence had been studied. Interestingly, this relation held both for sentences that had been studied and for new, nonstudied, sentences that were consistent with the overarching idea unit. Remarkably, except for sentences expressing only one idea, subjects were equally confident in the sentences that had not been presented in the study phase as for those that had been presented! Therefore, consistent with Bartlett's ideas, people seemed to retain the meaning of the sentences, but did not retain the literal wording of the sentences (see *Memory for Meaning and Surface Memory*).

Bartlett (1932), who began this tradition of work, was interested in repeated recollection of material. He presented subjects with a native American folktale, '*The War of the Ghosts*,' which he had them read twice. The story is a bit disjointed, has supernatural elements, and is rather difficult to understand. He found that subjects frequently made errors in recalling the story and that these errors increased over repeated retellings of the story. The stories got shorter, reflecting forgetting, but more importantly for present purposes, the story changed in systematic ways by subjects introducing errors that served to rationalize its bizarre elements. For example, people might 'remember' elements that had not been present to explain some of the unusual features in the story. Interestingly, this tendency to rationalization—to reconstruct the story more in line with the rememberers' own schema—seemed to grow over time and repeated retellings. Although Bartlett's original studies were rather informal, the work has been replicated under more systematic conditions (Bergman and Roediger 1999).

4. Conclusion

This article has reviewed selectively some key evidence that permits us to draw the conclusion that remembering the past should be viewed as reconstructing it. To draw on a metaphor from Hebb (1949), we can think of the process of remembering the past as we can conceive of paleontologists' reconstruction of a dinosaur from bone fragments and chips. The archeologist recovers a partial skeleton, but the finished product in a museum is shown as complete, with new bones added, old ones refinished or enhanced, and the entire skeleton constructed based on knowledge of what the animal must have looked like. Similarly, Hebb (1949) argued, remembering the past involves recollection of specific facts and details (the bones) that are woven together into a complete story of the event (like the skeleton). However, the story about the event might involve considerable constructive activity on the rememberer's part and the rememberer's present knowledge and goals may shape and determine how he or she remembers the past. A quote from Bartlett captures the essence of this reconstructive approach:

> … the one overwhelming impression produced by this more "realistic" type of memory experiment is that human remembering is normally exceedingly subject to error. It looks as if what is said to be reproduced is, far more generally than is commonly admitted, really a construction, serving to justify whatever impression may have been left by the original. It is this "impression," rarely defined as with much exactitude, which most readily persists. So long as the details which can be built up around it are such that they would give it a "reasonable" setting, most of us are fairly content, and we are apt to think that what we have built we have literally retained (Bartlett 1932, pp. 175–6).

See also: Episodic and Autobiographical Memory: Psychological and Neural Aspects; Eyewitness Memory: Psychological Aspects; False Memories, Psychology of; Memory: Autobiographical; Memory Retrieval; Metamemory, Psychology of; Short-term Memory, Cognitive Psychology of; Working Memory, Psychology of

Bibliography

Bartlett F C 1932 *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, Cambridge, UK

Bergman E T, Roediger H L 1999 Can Bartlett's repeated reproduction experiments be replicated? *Memory & Cognition* **27**: 937–47

Bransford J D, Franks J J 1971 The abstraction of linguistic ideas. *Cognitive Psychology* **2**: 331–50

Brewer W F 1977 Memory for the pragmatic implications of sentences. *Memory and Cognition* **5**: 673–8

Conrad R 1964 Acoustic confusions in immediate memory. *British Journal of Psychology* **55**: 75–84
Crowder R G 1976 *Principles of Learning and Memory*. Erlbaum, Hillsdale, NJ
Gregory R L 1998 *Eye and Brain*. Oxford University Press, Oxford, UK
Hebb D O 1949 *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York
Loftus E F, Miller D G, Burns H J 1978 Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory* **4**: 19–31
Miller M B, Gazzaniga M S 1998 Creating false memories for visual scenes. *Neuropsychologia* **36**: 513–20
Roediger H L, McDermott K B 1995 Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**: 803–14
Sperling G 1967 Successive approximations to a model for short-term memory. *Acta Psychologica* **27**: 285–92
Sulin R A, Dooling D J 1974 Intrusion of a thematic idea in the retention of prose. *Journal of Experimental Psychology* **103**: 255–62
Tulving E 1985 Memory and consciousness. *Canadian Psychology* **26**: 1–12

H. L. Roediger III

Recovery of Function: Dependency on Age

When the brain is injured in adulthood there is a loss of function followed by a prolonged period during which there is partial recovery of some types of cognitive and motor functions. The extent of functional recovery varies with age and although injury to the infant brain can have less severe consequences than similar injury in adulthood, there are times when injury is especially devastating. These differences are related to the details of anatomical changes at different stages of brain development. The difference in functional outcome after injury at different ages is related to the injury-induced changes in neuronal structure, changes that are referred to as plastic changes. These plastic changes are similar to the neural and glial changes that are associated with learning in the normal brain. Various factors, such as experience, hormones, and drugs, modulate these plastic changes. Even these factors cannot completely restore lost functions because there are limits to brain plasticity.

1. *The Nature of 'Recovery' after Brain Injury*

There are many ways to damage the brain (e.g., stroke, disease, trauma) and it is rare indeed for a family not to contain a member who is affected by one of them. For example, it is estimated that about 400,000 people with new head injuries are admitted to hospitals each year in the United States. Similarly, it is estimated that there are about 400,000 new cases of stroke every year in the United States and most surviving stroke victims experience significant cognitive loss. These figures do not include neurological disorders such as Parkinson's and Alzheimer's diseases, which together affect more than 1 percent of the population. The estimate in the United States is that about 48 million people have some sort of neurological disorder, excluding mental diseases such as schizophrenia, which are clearly a result of neural dysfunction.

The high incidence of brain injury and disease thus poses perhaps the most perplexing clinical neurological problem, which is the issue of how to repair the damaged brain. It is tempting to assume that if behavior improves after injury, then there is recovery of the original behavior. However, just because a behavior appears to recover does not mean that the lost behavior itself has returned. The problem is to determine when recovery has actually occurred. In fact, true recovery is a rare occurrence after brain injury.

We can identify at least three outcomes that could occur after brain injury. First, there may be compensation resulting from an adaptation to the lost function. Compensation could reflect a change in strategy or it may represent a substitution of a new behavior for the lost one. For example, a person who has a deficit in moving the eyes can be expected to have difficulty in reading. This difficulty can be solved, however, by moving the head, which allows restitution of function (reading) by substitution of one behavior for another. Similarly, if, following a stroke, a person can walk only with the aid of a cane, the ability to walk has returned but the original walking ability has not returned, just an adaptation to the behavioral difficulty.

A second outcome might be partial restitution of the original behavior. This could reflect recovery from some sort of nonspecific effect of the injury, such as swelling, or it may reflect a genuine return of function. Behaviors that return over a period of months are likely to reflect a partial restitution of function. This is most clearly seen in people with significant language disturbances (i.e., aphasias) in which the severe deficits in making the movements of the tongue and mouth necessary to speak may diminish leaving the person with more subtle language disturbances, such as in finding words (anomia).

Third, there is the possibility that there could be a complete restitution of the original behavior. Although it is at least theoretically possible for lost functions to return completely in a plastic brain, careful behavioral analysis shows that this is rarely, if ever, the case in either adult laboratory animals or human patients. In principle, however, it ought to be possible to reorganize the brain to allow at least some functions to completely recover. This process is most likely in the developing brain and it has been docu-

mented for over 100 years that children with damage to language areas can show an almost complete recovery of lost language functions.

2. Age-dependent Recovery

Broca first noticed in the late 1800s that damage to the speech zones of the left hemisphere during infancy did not interfere with the later development of language (for a review, see Finger and Almli 1988). Rasmussen and Milner (1977) later showed that the sparing of language functions was due to the shifting of the speech zones either to the opposite hemisphere or, in some circumstances, within the left hemisphere. In these later studies the critical factor was age: lesions of the left hemisphere language zones prior to 5 years of age changed the speech representation so that the affected language zone moved to the right hemisphere.

Similar lesions from the ages of about 6 to 12 still allowed recovery of language functions but this was accomplished by intrahemispheric reorganization (see also Krashen 1973). The studies of the recovery of language functions after infant brain injury led to the idea that damage to the infant brain led to better functional outcomes than damage after similar injury in adulthood. This conclusion was supported by the studies of Margaret Kennard in the late 1930s in which she showed that unilateral motor cortex impairments are less severe after motor cortex injury if the lesion occurs in infancy rather than adulthood (e.g., Kennard 1940). One difficulty with the studies of both Kennard and the language reorganization in children is that there are also many examples in which early brain damage results in rather severe functional loss, such as in cerebral palsy or hemiplegia resulting from unilateral cerebral injury at birth. In fact, Hebb (1949) noted that children with frontal lobe injuries in infancy showed little functional recovery in adulthood and for some behavior, such as social behavior, these children might actually have more severe deficits than adults with similar injuries.

The apparent contradiction between the recovery of language and some motor functions on the one hand, and the poor recovery of some executive, social, and other motor functions on the other hand, has been resolved by laboratory studies in which precise age at injury and the nature of the behavioral deficits have been systematically examined. A key element of these studies is that the age of the brain is measured not in relation to the birth day, but rather it is measured by its stage of anatomical development. For example, compared to humans, newborn rats are approximately equivalent to about 6-month old fetuses when they are born. Thus, if the newborn rat brain is damaged, we would anticipate functional symptoms similar to those of human infants whose brain is injured at the beginning of the third trimester. In order to compare across species we can separate the periods of brain development into five broad stages: neurogenesis, cell migration, cell differentiation, synaptogenesis, and gliogenesis (see *Prenatal and Infant Development: Overview*).

Studies on both laboratory rats and cats have shown that if the cortex is injured bilaterally during neurogenesis, when neurons are formed, there is substantial recovery and, if the injury is not too extensive, there can be virtually complete functional recovery. In contrast, however, if the cortex is injured in the first few days after birth in the rat, which is a time of neural migration and cell differentiation, the effect is functionally devastating as animals show far more severe effects of injury than would be expected even in aging animals. This poor outcome is not a function of extent of injury, nor is it localized to particular cortical areas (see Kolb 1995).

Rather, there is something about the rat cortex shortly after birth that makes it especially vulnerable. For example, damage during this time may disturb the subsequent process of synapse formation or may even alter stem cell activity in some way. Once this phase of development is over, however, and brain development moves to the production of connections and glia, the brain shows a remarkable ability to compensate for injury. Rats or cats with cortical injuries during this phase show far better behavioral capacities in adulthood than animals with similar lesions at any other time. In fact, on some behavioral tests these animals show recovery that is virtually complete.

Importantly, this recovery is far more extensive in cognitive tasks, such as the learning of various spatial navigation problems, than it is on tests of species-typical or motor behavior (e.g., Bachevalier and Malkova 2000, Kolb and Whishaw 1989). In humans, spatial behavior appears to show less recovery than verbal behavior (Stiles 2000). This behavior-specific recovery likely reflects the relative ease of reorganizing cortical circuitry in supporting different types of behavior. The neural substrates for less plastic behavior are presumably more highly specified and constrained during development than the neural substrates for more plastic behavior. It also reflects a difference in the way the brain produces cognitive processes versus basic sensory and motor processing. Thus, the most extensive recovery is associated with lesions of the frontal areas and the least extensive recovery is associated with damage to the primary sensory areas.

3. Mechanisms Underlying Recovery of Function

The fundamental assumption here is that recovery of function can occur only if there is some type of modification of the structure of the remaining brain. This structural change is necessary if the remaining brain cells are to be recruited in some manner to control behaviors that preinjury were controlled by the lost neurons. In principle, there are three ways that the brain could make changes that might support

recovery. First, there could be changes in the organization of the remaining, intact, circuits in the brain. The general idea is that the brain could reorganize in some way to do 'more with less.' It is unlikely that a complexly integrated structure like the cerebral cortex could undergo a wholesale reorganization of cortical connectivity, but rather, recovery from cortical injury would be most likely to result from a change in the intrinsic organization of local cortical circuits in regions directly or indirectly disrupted by the cortical injury. Although it might be possible to produce significant reorganization of cortical connections in the young brain, the overwhelming evidence in experimental animals is that this is rare and it is just as likely to be associated with abnormal functioning as with recovery.

Second, there could be a generation of new circuitry. Cerebral reorganization can be stimulated either by some sort of behavioral therapy or the application of some sort of pharmacological treatment. In either event, the stimulus could influence reparative processes in the remaining brain or could enhance the production of new circuitry. Once again, it seems most likely that the induced neuronal changes would be in the intrinsic organization of the cortex. One might predict that induced neuronal changes might be more extensive than in the case of endogenous change, in part because the treatment can act upon the whole brain rather than just on affected regions.

Third, there could be a generation of new neurons and glia to replace at least some lost cells. Stem cells that give rise to the neurons and glia of the brain remain active in the brain throughout life (see *Cognitive Development in Infancy: Neural Mechanisms*; *Cortical Development: Transplantation and Rewiring Studies*; *Hypnosis, Neural Basis of*).It is thus possible that neurogenesis could be stimulated after injury, especially during development, and that these new neurons could replace those lost to injury or disease (e.g., Weiss et al. 1996). All three of these possibilities for cerebral plasticity do occur, but the likelihood of each type of change varies with age at injury.

4. *Structural Correlates of Functional Recovery*

When the brain is damaged, there is a loss of neurons in the injured area but, in addition, many neurons that are not directly injured lose their normal synaptic inputs. The immediate effect of this synaptic loss is (a) neuronal death, (b) atrophy of the dendritic and/or axonal fields that have lost their connections, and/or subsequent reorganization of the connections. This reorganization can be inferred by using a Golgi stain to visualize the dendritic fields of the remaining neurons. Because the dendritic branches of a neuron provide the substrate for most of a cell's synapses, a decrease (or increase) in the amount of dendritic space available is correlated with a decrease (or increase) in the number of synapses.

The hippocampal formation of adult rats has provided a good model to study synaptic replacement. When input to the hippocampus is reduced by damage to the major input pathway via the entorhinal cortex, there is a massive denervation of the hippocampal formation, which can be seen in the atrophy of the dendritic fields of the hippocampal cells in the week after injury. Beginning at about 7–10 days after the injury, there is then a slow regrowth of the dendritic fields, although the pattern of regrowth is different from the original pattern. This regrowth reflects a reorganization of the circuitry and it is correlated with a partial recovery of function. The restructuring of the synaptic organization of the hippocampus occurs primarily because the remaining inputs to the hippocampus expand to occupy the regions of lost inputs.

A similar process appears to occur after restricted neocortical injury in adulthood as well, although the changes are less dramatic than in the hippocampus and there is much less functional recovery (for a review, see Kolb 1995). In contrast, however, there is extensive reorganization after cortical injury during development, and this reorganization is correlated with functional outcome. Golgi analyses of cortical neurons of rats with perinatal lesions consistently show a general atrophy of dendritic fields and a drop in spine density across the cortical mantle. In contrast, rats with cortical lesions around 10 days of age show an increase in dendritic fields and an increase in spine density relative to normal control littermates. Thus, animals with the best functional outcome show the greatest synaptic increase whereas animals with the worst functional outcome have a decrease in synapses relative to control animals.

But infant animals not only show changes in local cortical circuits but also in more distant cortical connectivity. Perhaps the most extensive studies of changes in cortical connectivity are those showing that after unilateral motor cortex lesions in infant rats or cats there is a major expansion of the ipsilateral corticospinal pathway from the undamaged hemisphere, which is correlated with partial recovery of skilled forelimb use (e.g., Castro 1990). These expanded connections do not reflect the growth of new pathways, however, but rather reflect the failure of superfluous pathways to die off during development. That is, in development there is normally a period during which many pathways are eliminated but after some types of brain injury these pathways are maintained (for a review, see O'Leary 1992). The presence of these abnormal pathways may support some types of functional recovery but it may also interfere with the normal functioning of remaining cortical functions. This possibility has been termed 'crowding' to reflect the idea that the normal functions of a cortical region can be crowded out by the development of abnormal connections (e.g., Teuber 1975). Thus, there may be recovery of some types of behavior, such as

motor functions, at the expense of other functions (e.g., Kolb et al. 2000). This appears to be especially likely if the injury occurs during the period of neuronal migration and unlikely if it occurs during the period of maximal synaptogenesis.

As noted earlier, synaptic change may also occur after injury because the brain generates new neurons. Although this is likely to be rare in the adult, it does occur during development, but the generation of these neurons varies with precise embryonic age and location of injury. If it is damaged during the time of neuronal generation, the brain can compensate by generating rather large numbers of new cells that appear to support at least partial restitution of function (e.g., Hicks et al. 1984). It is not known if this regenerative process occurs after injury to specific regions of the brain but it is likely a general phenomenon. Thus, if the brain is injured while it is still producing neurons, it appears capable of generating replacements (e.g., Kolb et al. 1998a). In contrast, if the brain is injured during neuronal migration there appears to be little generation of new neurons. If the same injury occurs during maximal glial formation, there is neuronal generation. This generation is selective, however, and occurs only if the midline telencephalon is damaged, namely the olfactory bulb or midline cortex (e.g., Kolb et al. 1998b). The newly generated tissue is functional and animals regain at least partial function.

See also: Adolescence, Sociology of; Adolescent Behavior: Demographic; Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Adolescent Vulnerability and Psychological Interventions; Adolescent Work and Unemployment; Adolescents: Leisure-time Activities; Adult Development, Psychology of; Antisocial Behavior in Childhood and Adolescence; Brain Damage: Neuropsychological Rehabilitation; Brain Implants and Transplants; Cerebral Cortex: Organization and Function; Childhood and Adolescence: Developmental Assets; Cognitive Development in Childhood and Adolescence; Friendship: Development in Childhood and Adolescence; Identity in Childhood and Adolescence; Neural Plasticity; Social Competence: Childhood and Adolescence; Socialization in Adolescence; Youth Culture, Anthropology of; Youth to Adulthood across Cultures, Transition from

Bibliography

Castro A J 1990 Plasticity in the motor system. In: Kolb B, Tees R C (eds.) *The Cerebral Cortex of the Rat*. MIT Press, Cambridge, MA, pp. 563–88

Bachevalier J, Malkova L 2000 Neuropsychological indices of early medial temporal lobe dysfunction in primates. In: Levin H S, Grafman J (eds.) *Cerebral Reorganization of Function after Brain Damage*. Oxford University Press, New York, pp. 27–48

Finger S, Almli C R 1988 Margaret Kennard and her 'principle' in historical perspective. In: Finger S, Levere T E, Almli C R, Stein D G (eds.) *Brain Injury and Recovery*. Plenum, New York, pp. 117–32

Hebb D O 1949 *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York

Hicks S P, D'Amato C J, Glover R A 1984 Recovery or malformation after fetal radiation and other injuries. In: Finger S, Almli C R (eds.) *Early Brain Damage*. Academic Press, Orlando, FL, pp. 127–47

Kennard M A 1940 Relation of age to motor impairment in man and in subhuman primates. *Archives of Neurol. Psychiatry* **44**: 377–97

Kolb B 1995 *Brain Plasticity and Behavior*. Erlbaum, Mahwah, NJ

Kolb B, Cioe J, Muirhead D 1998a Cerebral morphology and functional sparing after prenatal frontal cortex lesions in rats. *Behavioral Brain Research* **91**: 143–55

Kolb B, Cioe J, Whishaw I Q 2000 Is there an optimal age for recovery from motor cortex lesions? I. Behavioral and anatomical sequelae of bilateral motor cortex lesions in rats on postnatal days 1, 10, and in adulthood. *Brain Research* in press

Kolb B, Gibb R, Gorny G, Whishaw I Q 1998b Possible brain regrowth after cortical lesions in rats. *Behavioral Brain Research* **91**: 127–41

Kolb B, Whishaw I Q 1989 Plasticity in the neocortex: Mechanisms underlying recovery from early brain damage. *Prog Neurobiol* **32**: 235–76

Krashen S D 1973 Lateralization, language learning, and the critical period: Some new evidence. *Language Learning* **23**: 63–74

Levin H S, Eisenberg H M, Benton A L 1989 *Mild Head Injury*. Oxford University Press, New York

O'Leary D D 1992 Development of connectional diversity and specificity in the mammalian brain by the pruning of collateral projections. *Current Opinion in Neuroscience* **2**: 70–7

Rasmussen T, Milner B 1977 The role of early left-brain injury in determining lateralization of cerebral speech functions. *Annals of the New York Academy of Sciences* **299**: 355–69

Steward O 1991 Synapse replacement on cortical neurons following denervation. In: Peters A, Jones E G (eds.) *Cerebral Cortex*. Plenum, New York, Vol. 9, pp. 81–131

Stiles J 2000 Spatial cognitive development following prenatal or perinatal focal brain injury. In: Levin H S, Grafman J (eds.) *Cerebral Reorganization of Function after Brain Damage*. Oxford University Press, New York, pp. 201–17

Teuber H-L 1975 Recovery of function after brain injury in man. In: *Outcome of Severe Damage to the Nervous System*. CIBA Foundation Symposium 34. Elsevier, North-Holland, Amsterdam, The Netherlands

Weiss S, Reynolds B A, Vescovi A L, Morshead C, Craig C G, van der Kooy D 1996 Is there a neural stem cell in the mammalian forebrain? *Trends in Neuroscience* **19**: 387–93

B. Kolb

Recovery of Function, Neural Basis of

Everyone has had the experience of having had an injury in one's life, from burning a finger to scraping an elbow to stubbing a toe, perhaps cutting oneself or

breaking a bone. The universal feature of this experience is that, with care and treatment and time, the injury heals and usually nothing more than a scar and a faded memory remain.

Such recovery is not common with an injury to the nervous system. One of the most important goals in the field of neuroscience is the issue of recovery of function following nerve, spinal cord, and brain injury. It is an unfortunate fact that injuries of the nervous system can produce profound deficits in the quality of life. Sometimes these deficits are obvious, as when an injured nerve affects motor or sensory functioning. Sometimes there are profound deficits without obvious signs of injury, as when someone suffers from memory loss from injury or developmental/aging effects, for example, autism or Alzheimer's disease. In either case we would like to be able to restore functioning.

Sometimes, however, there are instances of considerable recovery of function following brain injury. Stroke patients who initially have profound language disorders, for example, may in the long term have considerable restoration of function. Until recently, in the animal experimental literature, there has been difficulty in localizing neural correlates of functioning because the effects of damage were only temporarily observed with significant amounts of recovery of function over time. Ideally, if we had enough understanding of the nervous system then we might be able to promote recovery of function. The study of brain-behavior relationships, of localization of function, and of recovery of function, are intimately associated themes.

1. Theories

The four major theories that account for instances of recovery of function following brain injury are examined in this article. No one theory fully accounts for the spectrum of observed recovery, nor are the theories mutually exclusive—more than one factor may be involved in recovery. To a certain extent, these theories provide a heuristic outline of the factors that affect recovery. It is the hope that a better understanding of these factors will lead to more effective treatment of the injured.

1.1 Vicariation

Vicariation means second-handed. Recovery by vicariation occurs when an injury damages the part of the nervous system that is actually responsible for a function, and a second neural structure that had never been involved in that function now takes over for the injured structure and performs that previously lost function. A key feature of vicariation is that the structure that takes over the function was never involved in the original behavior. This distinction is lost in some treatments.

The best example of vicariation is the observation that injuries in young individuals are said to be less deleterious than similar damage when it occurs in adults. For example, damage to the left hemisphere in a stroke can cause profound loss of language functions in the adult. Children with left hemisphere damage, however, often do not show severe language deficits. This observation fits our concept of the plasticity of the developing organism.

This concept—that youngsters are less affected by brain injury—has been applied to laboratory experiments where it has generally been found that damage is less intrusive in young individuals. For example, damage to the frontal lobes does not produce delayed response deficits in monkeys, and damage to the septal region does not produce social cohesiveness in rats. However, it is not always true that an early lesion is beneficial. In these two examples, when the individual becomes an adult then delayed response deficits and increased social cohesiveness appear. Thus, the effects of the damage may not be obvious until the individual matures.

Of all the theories of recovery of function, the theory of vicariation is the least likely one to be true. This is because much of the recovery that has been attributed to vicariation is actually due to a region that already has some functioning. As a result, the following theories are more likely explanations, and thus more likely to cause recovery.

1.2 Diaschesis

Diaschesis refers to the effects of shock. Recovery by diaschesis occurs when a nervous system injury does not actually damage the tissue that is responsible for the lost function, but rather that the injury causes a shock effect which indirectly and deleteriously affects normal functioning, and function returns when the shock subsides. Shock can occur when there is temporary loss of the blood supply (loss of nutrients including glucose and oxygen, or loss of functions to remove waste metabolic products) or temporary loss of necessary modulating neural activity. The key feature is that when functioning returns it is mediated by the very same tissue that was responsible for the function in the first place, so that the recovered behavior is exactly the same as the original one. Thus, the idea of sparing is essential in the context of diaschesis.

Time appears to be the major manipulation to effect diaschesis. However, time itself is not a very profound or necessarily a helpful explanation. The real question is what happens that takes time to dissipate the effects of shock, or more to the point, what can be done to diminish the effects of shock. Steroid drugs have been given to prevent the effects of swelling, for example, and have some effect in mitigating loss of function from some spinal cord injuries. To be effective the treatment must occur soon after the injury. Since the

brain and spinal cord are enclosed within the cranium and spinal column, respectively, any swelling of the nervous system in these confined spaces can cause substantial damage that is secondary to the original injury. Thus, besides indirect effects of shock any injury can cause additional damage.

One example of diaschesis from the laboratory is known as the Sprague effect. Damage to one superior colliculus in cats causes a visual neglect in the opposite visual field; that is, objects are ignored if they are presented in that visual field. Subsequent damage to the visual cortex on the same side as the visual neglect immediately restores reactions in the affected visual field. This example represents diaschesis by loss of normal neural modulatory influences. Here subcortical visual regions in the tectal area receive three visual inputs, one from the eye through the optic nerve to the superior colliculus, one from the opposite superior collicus, and the other through the visual cortex. By damaging the superior colliculus there is an imbalance in the surviving tectal region. By subsequently damaging the visual cortex on the same side as the visual neglect, the balance in the surviving tectal region is restored. This illustrates one of the major ways of thinking about diaschesis, which is to think of it as a balance of modulation of neural sources.

Another example from the laboratory is the loss of a simple visual discrimination for brightness. Rats trained on a black–white discrimination will appear to have lost that memory when their visual cortices are removed. With subsequent training the rats reacquire the ability to discriminate black from white. Since it takes about as many trials to relearn as it did to originally learn, this appears to be a case of vicariation where another part of the brain (presumably the remaining cortex) takes over functioning. If that is the case, however, that subsequent training is creating a new memory, then it should not matter whether training after the brain injury is made with the black or white stimulus as the correct choice. But reversing the correct choice profoundly interferes with subsequent training. This shows that although it appears that the brain injury has removed the original memory that in fact the memory remains to interfere with reversal training.

A number of manipulations have been found to affect recovery through diaschesis, each associated with diminishing the effects of shock and enhancing the expression of the spared tissue: extensive preoperative training, interoperative retraining (serial lesions), extensive postoperative retraining, cross-modal training on similar problems (i.e., training using other sensory modalities), drugs that diminish shock (e.g., steroids), drugs that enhance access to the spared tissue. As with strategies to improve performance on exams, a practical strategy to lessen the effects of future brain injury or the inevitable effects of aging would be to overlearn with multiple cues and learning methods.

Parkinson's disease provides an example from the clinical literature where the diminishing effects of diaschesis prevents the initial appearance of symptoms. Parkinson's disease is a movement and thought disorder caused by progressive degeneration of cells in the substantia nigra that project dopamine to the basal ganglia. Its symptoms only appear when so many nigra cells die that only about 20 percent of the normal amount of dopamine remains in the basal ganglia. Until that point, it appears that the basal ganglia can compensate for the loss of dopamine (see below). It is only when the dopamine projection to the basal ganglia falls below a critical point that the symptoms of Parkinson's show themselves. Presumably, slow progressive disorder does not produce the effects of shock that would be seen with one larger lesion.

Of the factors that account for recovery of function from brain injury, diaschesis is one of the most important.

1.3 Neural Reorganization

Neural reorganization generally refers to the body's own mechanisms for recognizing loss and for healing itself. However, man-made interventions can also be classified as reorganization. By far, neural reorganization receives the most attention of the four theories, although its practical usefulness has yet to be realized. Several means of effecting neural reorganization have been described. Importantly, these mechanisms are adaptations of means that the nervous system has for normal development and growth rather than specialized adaptations that might have evolved solely for the purpose of a complex brain being able to repair itself. That is, one questionable idea about brain organization in general is the idea that more complex brains, being more susceptible to injury of its delicate mechanisms, somehow developed a new method of dealing with brain injury. That seems not to be the case. Here the known mechanisms of reorganization are reviewed.

1.3.1 Reactive synaptogenesis. Synaptogenesis refers to new synapse formation. This mechanism is essential developmentally for correctly hooking up the nervous system (along with overproduction and selective cell death, as in programmed cell death or apoptosis). Reactive refers to an action in response to injury. New connections can be made by a number of activities, including new axonal growth (sprouting), new pre- and post-synaptic growth and terminal and dendrite formation, new receptor formation and/or distribution, for example. Furthermore, injury to the visual cortex causes degeneration of its projections to subcortical visual areas which leaves vacated synaptic space that is filled by sprouting of the optic nerve. In the example above (see Sect.

1.2) where rats were trained on a black–white discrimination, sprouting in subcortical visual areas could account for the recovery of function. It is generally thought that sprouting should be beneficial and attempts have been made to promote sprouting. However, in the case of brightness discrimination this idea is questionable since subsequent damage to the areas of sprouting has no permanent effect in preventing learning differences in brightness. In general, since rats with visual cortical damage do not recover perceptions of form and pattern, neural reorganization is not a significant factor for those functions.

Two examples illustrate where sprouting is actually less than helpful. First, spinal cord injuries cause sprouting that is accompanied by tremors. The tremors are alleviated when the sprouted connections are destroyed. Second, injury of one superior colliculus in hamsters causes sprouting to the opposite superior colliculus. The result is that the animals visually orient to the side away from the visual stimulus. Cutting the sprouted optic nerve to the remaining superior colliculus resolves the problem. Thus, in these two examples, blindly enhancing sprouting by generally giving nerve growth factors, for example, may not be such a good idea. The problem with sprouting is to have control over where and what type of new growth takes place.

1.3.2 Neurogenesis. We learn that we are born with the total number of neurons that we are ever going to have, and that we continually lose neurons as we grow older. The one celebrated exception to this rule is that certain birds which learn a new song each mating season show cell division and new neuron development as adults. Adult rats and recently adult humans have been shown to create new neurons in the hippocampus, a limbic system structure associated with certain kinds of memories. It remains to be determined whether there are sufficient numbers of new neurons and whether there are functional connections of these newly available neurons. Similarly, it is unknown whether these neurons have a role in repairing brain injury. However, on the face of it, this is an exciting development.

1.3.3 Transplants. Related to the idea of natural neural replacement (neurogenesis) is the idea that we can purposely implant neurons to replace the damaged ones. Transplants have been tried most often in the case of patients with Parkinson's disease, where the transplant donor is either the patient (dopamine producing cells from the patient's own adrenal glands) or an aborted fetus (dopamine producing cells from the substantia nigra; this source is controversial). Initial results of such operations were mixed at best, with more recent limited successes. The ameliorating effects may not be due to the effects of dopamine or to neural connections formed by the implant, however. Nerve growth factors that are secreted by the implant are more likely candidates. Transplants themselves have been recently overshadowed by a return to surgical intervention, a procedure known as a pallidotomy which is a lesion of the globus pallidus.

1.3.4 Supersensitivity. Synapses that are deprived of their normal expected amount of input activity can become supersensitive whereby the synaptic transmission becomes unusually sensitive to small amounts of neurotransmitter. In Parkinson's disease the deleterious symptoms are delayed presumably because of the compensating supersensitivity of cells in the basal ganglia that are deprived of their usual amount of dopamine from degenerating neurons in the substantia nigra. It is only when there is too much loss of dopamine that the basal ganglia neurons cannot compensate and symptoms of Parkinson's appear. The actual mechanism of supersensitivity can involve increased sensitivity of the postsynaptic receptors, increased reactivity of postsynaptic receptors, increased number of postsynaptic receptors, changes in dendritic spines that reduce impedance (i.e., enhance trasmission), increased transmitter release, and changes in intracellular events that enhance the signal.

1.3.5 Prostheses. Man–machine implants may be used to overcome deficits. For sensory loss, these involve a transduction device to convert an external stimulus into a neural-like signal with a neural stimulator. Currently, auditory implants into the cochlear nerve or nuclei are being tried to restore hearing. Visual implants using a primitive camera and stimulation of visual cortex have a history of limited success in effecting primitive sensations. Not all prostheses are surgically placed into the brain. One alternative for visual problems is to place the pattern of visual stimuli onto a skin surface where the visual stimulus is now interpreted as a tactile stimulus. Other approaches have not involved sensory transduction but rather replace the damaged element responsible for higher level processing. In this regard a recent attempt has been made to implement functioning in a silicon chip that is used to replace a damaged hippocampus. For motor loss, electrodes have been implanted to stimulate muscles, coordinated by a computer, to effect walking. This requires sensors for proprioception and balance.

1.3.6 Unmasking. It is now well appreciated that the nervous system has multiple representations for each

of the senses and motor functions. Moreover, it is becoming clear that representations may overlap each other, and that under normal circumstances some of the underlying representations may be suppressed. When a part of the main representation is damaged the underlying map may now express itself, either preventing any deficit or causing an apparent recovery. Unmasking is generally found in instances where the recovery occurs too quickly for sprouting or other types of reorganization to have been made.

1.3.7 Regeneration. Certain amphibians and lizards will grow a new limb to replace one that has been lost to a preditor. Amphibians and lizards also have the ability to re-establish connections when a nerve is severed. This regeneration is also seen to varying degrees in the peripheral nervous system of mammals, aided by the supporting Schwann cells which guide the reconnection. Although the capacity for regeneration persists in the central nervous system of mammals, regeneration is not a significant factor in recovery of function because central glial cells create scar tissue which becomes a significant barrier to reconnection.

1.4 Substitution

Substitution refers to using an alternative means to achieve the same goal. If damage has actually removed that part of the brain that is responsible for a function (as opposed to Diaschesis, above, for example) then substitution is the most likely means of recovery. Within the animal literature one needs to be cognizant of factors like substitution because otherwise false conclusions can be made about the nature of the recovery. Yet substitution is often ignored because it does not involve surgical or chemical intervention. There are three significant means of substitution. Recovery by substitution occurs when another sense, motor act or strategy can be used.

1.4.1 Sensory. A sensory stimulus or property of a stimulus can be used as a substitute for the original sensation. Damage to visual cortex in rats makes them blind to form and pattern perception yet they are still capable of detecting primitive visual experiences such as brightness and contour. As a result, such rats appear to be able to solve discriminations for orientation (e. g., horizontal vs. vertical black and white stripes) because they can make use of local brightness and contour cues to solve the problem. In extreme conditions where the rats are prevented from using these local cues, they can no longer 'solve' the orientation problem, and thus they are blind. The goal for a therapist would be to design conditions where a problem can be solved in multiple ways. For example, persons who are red–green color blind can learn to drive by noting the position of the active light on a traffic signal.

1.4.2 Motor. One motor act can substitute for the original motor act. Cats with dorsal rhizotomies (denervation of the sensory roots of the spinal cord) for the hind limb initially have difficulty walking a thin railing. After a period of time, however, the cats negotiate the task very well and appear to be completely normal. However, upon close examination with high-speed photography, it has been shown that these cats adopt a motor strategy of shifting their weight to properly place the hind foot. Clearly, functional recovery does not imply neural recovery.

Another example from the older literature would be the instance where a rat has been trained to run a maze. The same rat that has been trained to run can solve the very same maze when it is flooded and it has to swim. The same idea applies to conditions where brain damage prevents the original behavior.

1.4.3 Strategy. After ruling out that the senses and motor functions are not affected, it might be found that there may be a problem with the implementation of a sequence of perceptions and actions. This higher order activity can be referred to as a strategy. For example, a woman with global (diffuse) brain damage got easily distracted, so much so that a telephone ring would be enough of an interruption that she would forget she was cooking. Or she would spend her whole day watching TV. Her strategy to overcome this deficit was to write a list of things to do, checking them off as she did them, and to have a timer set to go off every 15 mins. At every alarm she would check her list to see what it was she was supposed to be doing.

From the animal literature comes another example. Rats with hippocampal brain damage have difficulty in alternately pressing one of two bars for a food reward. This task requires that the animals remember which bar was pressed last so that they can press the other bar the next time. Some clever rats figure out that if they hold onto the bar that they just pressed then they immediately know to press the other bar the next time.

2. Factors

The discussion above identifies the major theories that can account for instances of recovery of function. Within this discussion we have identified a number of factors that can be used clinically and within the laboratory to promote recovery of function. It is

reasonable at this time to simply list those factors, because although they were discussed in terms of theories above in truth many of these factors apply to more than one theory. Here is a listing of some factors that might enhance recovery of function: age at time of the injury, reduction in secondary effects of the injury (shock), the proficiency of preinjury training, training on alternative stimuli or strategies or acts, training between injuries (serial lesion effect), time after injury, amount of postinjury training on the same or similar tasks, substitutions. It would be misleading to suggest that any or all of these factors will lead to recovery in all instances. It is a sad fact that injuries to the nervous system produce profound and lasting deficits. The hope for further research is to find new applications of these factors and to develop new methods of treatment.

3. Implications

Clinically the importance of finding methods that enhance recovery of function is obvious. What is not so obvious is that the clinical and experimental literature about brain injury also give insight into the fundamental organization and operation of the brain and our mental life. The mind–body problem—the relationship of our physical being with our mental experience—continues to be a fundamental interest in psychology. Both the clinical and experimental literature make fundamental contributions to each other. Newer methods for imaging the brain during cognitive activity (fMRI, PET, neuromagnetic methods, along with the old standby EEG) have confirmed much of our understanding of the organization of the brain and what happens during injury, on the one hand, and given insights and further directions, on the other. The optimism of our times holds out great promise for sophisticated advances that may someday allow us to repair the damaged mind.

Importantly, we learn from the field of recovery of function the following lessons. First, localization of function is the norm in the nervous system. Second, there is a hierarchical organization to the nervous system, with basic functions organized at low levels. Third, even low levels of the nervous system are capable of seemingly sophisticated functions like learning and memory. Fourth, any one function is probably represented in only one place in the nervous system. However, fifth, similar functions can often accomplish the same goals.

See also: Behavioral Neuroscience; Brain Damage: Neuropsychological Rehabilitation; Brain Implants and Transplants; Memory Problems and Rehabilitation; Neural Plasticity; Neural Plasticity in Auditory Cortex; Neural Plasticity in Visual Cortex; Neural Plasticity of Spinal Reflexes; Recovery of Function: Dependency on Age

Bibliography

Bach-y-Rita P (ed.) 1980 *Recovery of Function: Theoretical Considerations for Brain Injury Rehabilitation*. University Park Press, Baltimore, MD

Finger S (ed.) 1978 *Recovery from Brain Damage*. Plenum Press, New York

Finger S, Levere T E, Almli C R, Stein D (eds.) 1988 *Brain Injury and Recovery: Theoretical and Controversial Issues*. Plenum Press, New York

Johnson J A (ed.) 1990 *Occupational Therapy Approaches to Traumatic Brain Injury*. Haworth Press, Binghamtom, NY

Levin H S, Benton A L, Grossman R G 1982 *Neurobehavioral Consequences of Closed Head Injury*. Oxford University Press, New York

Montgomery J (ed.) 1994 *Physical Therapy for Traumatic Brain Injury (Clinics in Physical Therapy)*. Churchill Livingstone, New York

Scheff S W (ed.) 1984 *Aging and Recovery of Function in the Central Nervous System*. Plenum Press, New York

Stein D G, Brailwsky S, Will B 1995 *Brain Repair*. Oxford University Press, New York

Stein D G, Rosen J J, Butters N (eds.) 1974 *Plasticity and Recovery of Function in the Central Nervous System*. Academic Press, New York

Uzzell B P, Stonnington H H (eds.) 1996 *Recovery after Traumatic Brain Injury*. Erlbaum, Hillsdale, NJ

D. Lavond

Reduction, Varieties of

Fundamental issues in the ontology of the sciences can depend on whether the elements of one domain turn out to be 'nothing over and above' a congeries of elements of some other, perhaps less mysterious or more basic, domain. It has been argued that natural numbers, for example, are nothing over and above certain sets, and their properties just the familiar properties of those sets; that temperature is nothing over and above the mean kinetic energy of molecules; that genes are nothing over and above strings of DNA molecules; that the mental is nothing over and above the physical; and that the social is nothing over and above an aggregate of individuals, their antics, and attitudes. Each such claim, whether correct or not, postulates a dependency of one domain on another, a dependency strong enough to justify regarding the dependent domain as ontologically derivative, the other as ontologically more basic. Given the more basic domain, the dependent domain can plausibly be regarded as an ontological 'free lunch.' The elements of the dependent domain are not denied existence or reality, but neither are they granted an independent reality.

Such dependency claims have been thought to have far-reaching ramifications not just for ontology but also for methodology. For example, the social sciences investigate social phenomena, involving distinctively

social entities, properties, facts, and laws. The individualism–holism debate concerns, in the first instance, the status of these items—whether they are anything over and above the totality of individual entities, properties, laws, and facts—but also the methodological autonomy of the associated sciences. (For an even finer-grained analysis which disentangles other conflated but independent components of this debate, see Pettit 1993.) According to Watkins (writing in the 1950s), individualism holds that social institutions 'do not possess a quasi-divine autonomy' (in O'Neill 1973, p. 149), and that 'every complex social situation or event is the result of a particular configuration of individuals, their dispositions, situations, beliefs, and physical resources and environments' (in O'Neill 1973, p. 168). He argued that, as a consequence, explanations of social phenomena should be couched in terms of the 'situations, dispositions, and beliefs of *individuals*' (in O'Neill 1973, p. 179). His adversary, Mandelbaum, states *holism* thus: 'Societal facts are as ultimate as are psychological facts … those concepts which are used to refer to the form or organization of a society cannot be reduced without remainder to concepts which only refer to the thoughts and actions of specific individuals' (in O'Neill 1973, p. 223).

As Mandelbaum's claim here illustrates, definability was once the Holy Grail of the reductionist enterprise. Bertrand Russell famously maintained that postulating entities, in contrast to defining (or 'constructing') them, 'has many advantages; they are the same as the advantages of theft over honest toil' (Russell 1919, p. 71). But just as when the quest for the Holy Grail looked misguided, some more realistic venture might engage a knight's energies, so when definability failed to capture the interesting dependency claims, philosophers looked elsewhere. And they have discovered a plethora of dependency relations, some of which hold out the promise of prizing apart the methodological and ontological issues, enabling, for example, both the metaphysical strictures of the individualist and the methodological dispositions of the holist to be accommodated.

1. Explicit Definability

Explicit definition is the clearest way of demonstrating that *A*s are nothing over and above *B*s, guaranteeing that all truths about *A*s are expressible by talking overtly only about *B*s. A full specification of the *B*s yields *ipso facto* a full specification of the *A*s. An explicit definition of A in terms of $B_1, \ldots, B_n$ has the following general form:

$$A =_{\text{defn.}} F(B_1, \ldots, B_n)$$

The right-hand side, $F(B_1, \ldots, B_n)$, may involve, in addition to $B_1, \ldots, B_n$, only certain *permissible* operations. In the strictest case these are the purely logical connectives (*and*, *not*, *all*, and *identity*). More liberally one might include the concept of *set* and *set membership*, or perhaps the whole apparatus of mathematics. At least as traditionally understood, the definition should be knowable *a priori*, through pure analysis of the concepts involved, and it should hold by logical *necessity*.

A striking illustration of the power of definition is *logicism*, the program to reduce mathematics to logic (Russell 1919). In the nineteenth century, the basic concepts of analysis had been successively defined in terms of the concepts of natural numbers and sets. It began with the elimination of infinitesimals via the definition of a limit. Irrationals were defined as sequences of rationals, and rationals as sets of pairs of natural numbers, leaving the natural numbers alone as primitive, undefined entities. The Russell–Frege definition of natural numbers is remarkable in that it reaches right outside the category of number. Numbers are defined using only the apparatus of logic—including the logic of sets. For example:

zero = the set of all empty sets

one = the set of all singleton sets

For this to be adequate, truths about numbers must follow from the definitions, together with definitions of the mathematical operations, and the logic of sets. This hope was almost vindicated—but for the discovery of troublesome paradoxes in the foundations of set theory. Despite the paradoxes, the program remained a paradigm of ontological reduction, inspiring much subsequent philosophical activity (Carnap 1928) and providing the framework for most debates about ontological commitments.

2. Reducibility

Definability proved too strong a requirement for typical claims of ontological dependence for reasons involving the constraints that definitions be known *a priori* and that they hold true by logical necessity.

As is well known, our understanding of the phenomena of heat, codified in thermodynamics, was transformed by the reduction of thermodynamics to statistical mechanics (Nagel 1961, Chap. 11). The reduction involves deducing the laws of thermodynamics from laws of statistical mechanics together with bridge assumptions connecting the concepts of the two domains (*temperature* and *heat* with *molecule*, *mass*, and *motion*). For example, the Boyle–Charles law for ideal gases (pressure times volume is proportional to temperature) can be shown to follow from the assumptions of statistical mechanics and the following assumption:

(1) *Temperature is mean kinetic energy of component molecules.*

Assumption (1) may look like a definition, and is sometimes touted as such, but it is not the result of *a*

priori analysis of its concepts, and it is not logically necessary. (Suppose everything were continuous stuff all the way down and temperature were, say, proportional to density of calorific fluid.) If true, (1) is not logically necessary, and if known, not known *a priori*.

Some maintain that (1), even if not true by definition or by logical necessity, is nevertheless a matter of identity, vindicating the reductive claim that temperature is nothing but motion of molecules. However, unlike definitional identities, it is a *contingent* identity, analogous to the contingent identity of *the morning star* and *the evening star* (Smart 1959). Analysis of concepts would not yield the fact that the brightest celestial body in the morning sky is identical with the brightest celestial body in the evening sky, and we can easily imagine a world in which they are not. But the morning and evening stars are nevertheless one and the same—Venus. So perhaps (1) states a contingent identity known *a posteriori*.

The idea that *identity* (the relation which each entity bears to itself necessarily and to no other thing) could be a contingent matter is still baffling. One way to reduce the baffle is to distinguish between intension (concept) and extension (item singled out). The intension is a certain way of singling out the extension. The extension of both *the morning star* and *the evening star* is a chunk of matter, but the intensions are different ways of determining that chunk. The intensions are *coextensive* in fact but not of necessity.

Property terms also come with an extensional and an intensional component. The extension of *black* is the class of things that are black. If all black things happened to be chunks of coal, and all chunks of coal happened to be black, then *black* and *coal* would be coextensive. They would single out the same class of objects. But is coextensiveness enough for reduction? The two terms are associated with different intensions, they pick out distinct features of the objects, and even if coextensive they would not be coextensive of necessity. Further, we would not be inclined to say that *black* reduces to *coal* (or vice versa) just because they were coextensive. So there must be more to reduction of properties than coextensiveness.

Temperature and *mean kinetic energy* are not properties but *magnitudes*. Unlike a property, the extension of temperature is not a class of objects. Rather, the *temperature* of an item can be any one of a number of different possible numerical values—relative to a unit of measure and an origin. The extension of *temperature*, like *mean kinetic energy*, thus associates an item with a number. We could construe (1) as the simple claim that the magnitudes *temperature* and *mean kinetic energy* are coextensive. However, as the argument concerning *black* and *coal* illustrates, a reduction requires a tighter connection. The extensional equivalence reported by (1) has to hold by *some kind of necessity* (not logical necessity).

A possible world is a complete way the totality of things might be. A state is logically necessary if it holds in all possible worlds. Laws of nature rule out some but not all possibilities (some laws might have been different from what they are in fact) so they are not logically necessary. But what the laws of nature entail involves a kind of necessity—*nomological* necessity. A state is nomologically necessary if it holds in all worlds in which the actual laws of nature hold. For example, if the theory of relativity is true then in some sense it is impossible for an object to accelerate through the light barrier. This is not just an accidental fact, like the decay of some radium atom, but it is not *logically* necessary either. It is nomologically necessary.

For (1) to constitute a reduction it must be understood as the claim that *temperature* is coextensive with *mean kinetic energy* by nomological necessity. We can now exhibit the difference between definability and reducibility:

A is *definable* in terms of the *B*s, just in case *A* is coextensive with $F(B_1, \ldots, B_n)$ by *logical* necessity.

(Or: *A* is logically equivalent to $F(B_1, \ldots, B_n)$.)

A is *reducible* to the *B*s, just in case *A* is coextensive with $F(B_1, \ldots, B_n)$ by *nomological* necessity (in the domain of the *B*s).

(Or: *A* is nomologically equivalent to $F(B_1, \ldots, B_n)$.)

It follows that *A* is reducible to the *B*s if *A* is nomologically equivalent to something definable in terms of the *B*s. Definability entails reducibility (since logical necessity entails nomological necessity) but not vice versa.

Just as (1) transformed our understanding of temperature, our understanding of a mental state like *pain* may be transformed by the following claim:

(2) *Pain is the stimulation of C-fibers in the brain.*

For this to be a reductive claim on the model of (1), *pain* must be nomologically equivalent to *C-fiber stimulation*. What play the roles here of reducing and reduced theory? The reducing theory would be neurophysiology, and the reduced theory could be our ordinary everyday conception of pain—collections of platitudes like:

Pain is typically brought about by bodily damage, and results in wincing, groaning, and avoidance behavior.

Such platitudes were central to the *behaviorist* hope that pain might be characterized, or *defined*, in purely behavioral terms (Ryle 1949). When this hope of definability was abandoned the platitudes were central to the *functionalist* hope that pain could be suitably *reduced* to physical states of the brain (Lewis 1993, essay 7, Armstrong 1968). Assumption (2), together with a suitably rich scientific theory, might well yield such platitudes. There is a puzzle here to do with the status of the bridging principles like (1) and (2). If we want to reduce the mental to the physical, for example, by exhibiting purely physical (nomological) equivalents of mental features, we would not want to have to acknowledge irreducible laws governing the psychological realm, or irreducible connections between the two realisms. So for reductionism to follow, the laws underpinning nomological equivalents stated in the

bridging principles would have to be laws of the lower-level domain—in this case purely physical laws. The puzzle is how purely physical laws could entail principles straddling the two domains—an answer will be presented in Sect. 5.

If analogous reductions were to hold for all mental features then *physicalism*—the thesis that the mental is nothing over and above the physical—would be vindicated. Undergirding the reduction would be a multitude of nomologically necessary extensional equivalences between mental terms on the one hand and logical products of purely physical terms on the other. A successful reduction would not, of course, entail the elimination of mental entities or of the laws within the mental realm. (The reducibility of thermodynamic laws to statistical mechanics is rather a *vindication* of the former.) And far from being the denial of systematic and lawlike mental–physical connections it is the very affirmation of them—in the form of bridge principles.

3. *The Problem of Networking*

For a successful reduction we must have appropriate logical or nomological equivalents. There is, however, an important objection to the existence of such equivalents in two central cases—the relation between the mental and the physical, and the relation between the social and the individual—namely, the heavily networked character of the concepts characteristic of different domains. Mental concepts are heavily networked not just with behavioral and physical concepts but also with other mental concepts. Social concepts are heavily networked not only with each other but also with psychological concepts via those very same social concepts.

Consider *pain* again. Pain is tied not only to physical inputs and behavioral outputs, but also to other psychological states. Those in pain tend to believe they are in pain, and desire to get out of it. These are platitudes about pain which any reductive account of pain should accommodate. So it looks as though there may be no simple way of giving logical or nomological equivalents of pain in purely physical terms.

The case of social concepts is even more acute. Consider the statement, 'Jones cashes a check at the bank.' The predicate involves a social property—making essential use of the social concept of a *bank*—and the claim itself entails the existence of a bank. According to Mandelbaum, we cannot eliminate the concept *bank* in explicating the content of this claim. An analysis in terms of psychological states (of Jones, the teller, the bank manager, etc.) would have to include beliefs and desires the contents of which will essentially involve the concept of bank.

> … concepts involving the notion of status and role cannot themselves be reduced to a conjunction of statements in which these or other societal concepts do not appear (O'Neill 1973, p. 225).

So here is an obstacle to finding a purely individualistic equivalent of the social property. While Mandelbaum is thinking in terms of definability, a similar argument might be mounted against reducibility.

4. *Implicit Definability*

The networking problem was solved by an elegant theory of implicit definition, initially discovered by Carnap in the course of his search for a reduction of the theoretical to the observable (Carnap 1966) and fully developed by Lewis into a general reductionist framework (Lewis 1993, essay 6).

Quine famously argued, against positivism, that theoretical terms gain their meaning from being heavily networked with each other and with observable terms only in the context of a rich and testable theory (Quine 1961). Carnap proposed an essentially reductionist account of such terms which nevertheless accommodated this holistic thesis. The essential ingredient was provided by a then little known result of Ramsey (1931). Ramsey showed how a theory about, for example, electrons and protons, T(*electron*, *proton*, …) could be stripped of explicit mention of unobservable entities and replaced by a lean purely observational surrogate, its so-called *ramsification*. Theoretical terms—*electron*, *proton*—are first replaced with variables, E, P, (or 'gaps') to yield a 'gappy' sentence $T(E, P, \ldots)$ which may or may not be true depending on what the gaps are filled with. Now consider the definite claim that the gaps can be indeed truly filled by something or other:

Ramsification of T(*electron*, *proton*)

$$= \textit{there are items } E, P \textit{ satisfying } T(E, P)$$

This stripped-down version of the theory entails nothing nontrivial about electrons or protons (because it does not mention them at all), but does entail all the purely observational consequences of the full-blown theory. Since the point of the theory (according to the positivists) is to generate observational consequences, and the ramsification does this without reference to theoretical entities, Ramsey concluded that the theoretical ontology is redundant (Ramsey 1931).

Carnap drew a different lesson. He agreed with Ramsey that the factual component of the theory is delivered by its ramsification, but he noted that for the factual claim to be true there must be some items satisfying the gappy statement $T(E, P, \ldots)$—for that is exactly what the ramsification says. In that case, we may as well regard those items as the intended

extensions of the concepts *electron* and *proton*. The whole theory thus serves as a partial constraint on the meanings of its own theoretical terms. It is partial, both because it places no constraints in the case in which the theory is not observationally adequate, and because even when it is there may be a range of legitimate satisfiers of the gappy claim. However, in certain cases we may be lucky and the theory might be strong enough to single out a unique set of satisfiers. In that case we can use the theory to give an explicit definition (Lewis 1993, essay 6).

Something analogous can be sketched for mental properties. We begin with our ordinary everyday collection of platitudes about the mental ('folk psychology') in which, as noted, the mental concepts such as pain and desire are heavily networked with each other and with behavioral and physical concepts. Now ramsify this theory, eliminating the mental concepts and replacing them with variables (*P* for pain and *D* for desire). The ramsified theory will look something like this: there is a property *P* typically caused by bodily damage and which causes wincing, and if a something has *P* it typically bears *D* to getting rid of *P*.

This ramsification is stated purely in physicalistic and behavioral terms and might be true of a system which consisted only of physical constituents with purely physical properties. Neurophysiologists might discover that C-fiber stimulation is a feature typically caused by bodily damage and causes wincing; that C-fiber stimulation initiates D-fiber stimulation, which in turn causes C-fiber stimulation-ridding behavior, and so on. Enough of these would support the claim that pain is nomologically equivalent to C-fiber stimulation; the desire to get rid of pain nomologically equivalent to D-fiber stimulation, and so on.

If this is right, then the holistic and networked character of the mental is not itself an insurmountable obstacle to reduction in the case of the mental. Hopefully something similar might be achieved for the social domain. The reduction in the latter case would be to purely individual properties and relations—psychological, behavioral, and physical—which may in turn be reduced to the purely physical.

5. *Multiple Realizability*

Not long after functionalism had provided this systematic framework for reduction, philosophers began to worry that there may be no appropriate, lawlike extensional equivalences in the case of the mind (Putnam 1967, Davidson 1970). Suppose in a typical human being pain is 'realized' by C-fiber stimulation, whereas in some untypical humans it is D-fiber stimulation, and in Fido the dog it is E-fiber stimulation. Suppose, further, that there is no end to the number of different states which realize pain in different beings. Suppose in some distant galaxy silicon-based aliens have evolved, or could evolve, with a very different physical structure. Would not they be capable of experiencing pain too? There may be no limit to the ways in which pain can be thus realized in physical beings. So there may be no single, purely physical, nomological equivalent of pain. Pain would not be reducible to the physical.

Or think of the economic institution of cash. Lots of different kinds of stuff in fact play the role of cash, and almost any suitably portable object could play the role of cash given an appropriate combination of beliefs, desires, and conventions. Now reducibility is a transitive relation: if *A* is reducible to the *B*s and the *B*s are reducible to the *C*s, then *A* is thereby reducible to the *C*s. So if the social is reducible to the physical and the psychological, and the psychological in turn to the physical, then the social is reducible to the physical. But, given that almost any physical item at all could be cash, is it plausible that there be a single, purely physicalistic, nomological equivalent of *cash*? Hardly.

Does multiple realizability render physicalism and individualism false? Alternatively, is reducibility as characterized necessary for ontological dependence? Suppose every property which realizes pain (or cash) in all nomologically possible circumstances is itself a purely physical property. Would that not be compatible with physicalism? This suggests the possibility of a relation between the mental and the physical weaker than full reducibility but strong enough to vindicate the idea that the mental is nothing over and above the physical.

6. *Supervenience*

The favored term for the desired relation, one which guarantees determination without entailing reducibility, is supervenience—a term first used to characterize the relation between the moral and the natural.

It is widely held that moral differences are undergirded by natural or nonmoral differences. Suppose Francis of Assisi is good in virtue of kissing a leper with putrefying flesh (Francis of Assisi 1903). Then if Clare kisses an exactly similar leper and she is not good then there must be some nonmoral difference between the two—perhaps, unlike Francis, Clare is suppressing revulsion. Sameness of nonmoral features guarantees sameness of moral features. A difference at the moral level requires a difference at the nonmoral level. The moral supervenes on the nonmoral.

In the social realm, the individualist's battle cry might be: no difference at the social level without some difference at the individual level. Sameness at the individual level guarantees sameness of social features. That kind of supervenience claim seems eminently reasonable, one which even a holist might be embarrassed to deny. If Jones cashes a check in situation A but not in situation B, then there must be *some* difference at the individual level between the two situations—either between Jones in A and B, or

between some other individuals in A and B. Kim generalizes the supervenience condition thus:

> One family of properties is 'supervenient' upon another family of properties in the sense that two things alike with respect to the second must be alike with respect to the first (Kim 1978, p. 149).

So stated the principle is open to interpretation. The necessity involved ('*must* be alike') may be logical or nomological. Some supervenience claims—like the supervenience of the moral on the non-moral—are a matter of logical necessity (although see Oddie 1991). Other supervenience claims are a contingent affair. Most physicalists think that the supervenience of the mental on the physical is a contingent matter—for we can imagine worlds consisting of unembodied minds differing in their mental attributes but (lacking bodies) being physically indistinguishable. The supervenience of the social on the individual is disputed—some hold it to be a logical affair (Currie 1984) others a contingent affair (Pettit 1993).

Kim's statement of the supervenience principle requires individual by individual comparison: any two *B*-indiscernible individuals must be *A*-indiscernible. It has therefore been dubbed a kind of local supervenience. This may be too stringent for characterizing the supervenience of the social on the individual (Currie 1984) and even for the supervenience of the mental on the physical (Burge 1979). In both cases, it is claimed, the supervening properties involve not only the local properties of the individual, but the relations of the individual to its environment as well as to other individuals. (Whether Clinton is the President depends not just on Clinton's psychological and physical state but on psychological and physical states nationwide. Whether Clinton thinks that *that* woman is sexy requires a real woman, not just an hallucinated woman, to be there.) Furthermore, in the social case, some of the supervening entities (First Bank of Boulder, the Stock Market, France) are in any case of a different logical type from the individuals on which they are claimed to supervene, so we need to compare more than just the social properties of individuals. Finally, a number of arguments have shown that, given reasonable assumptions, local supervenience of the logical variety entails definability, and local supervenience of the nomological variety entails reducibility (Kim 1978, 1989, Bacon 1986, Oddie 1991), thereby failing the nonreductionist desideratum.

In order to avoid these problems global supervenience has been put forward as the salient relation. Call two worlds *B-indiscernible* if the *B*s have the same extensions in both (that is, the *B*s are identically distributed in the two worlds). Then *A* globally supervenes on the *B*s if any two (nomologically) possible worlds which are *B*-indiscernible (the *B*s have the same extensions in both) are also *A*-indiscernible (*A* has the same extension in both). This principle is a desideratum of any concept of determination, and so it is the weakest relation that might serve the supervenience theorist's purposes.

Definability entails global supervenience of both logical and nomological varieties, and reducibility entails global supervenience of the nomological variety. Since global supervenience is the weakest principle guaranteeing determination, the crucial remaining question is whether it avoids reducibility. To illustrate this possibility let us define a very simple, albeit artificial, property, *G*:

(3) *G = being both poor and identical either to Clare or to Francis.*

Only Clare and Francis can have *G*, and they have it just in case they are poor. So any two worlds which are identical in the distribution of poverty will be identical in the distribution of *G*. That is to say, *G* globally supervenes on poverty.

Is *G* reducible to poverty? Let us suppose all four logically possible distributions of *G* are also nomologically possible. Is *G* nomologically equivalent to something definable solely in terms of poverty? (3) is not such a definition. On the right-hand side, in addition to *poor*, the logical connectives (*and* and *or*) and the identity relation, we have reference to two individuals. Identity, like the logical connectives, is typically counted part of the logical apparatus and so a permissible component of a definition. But the individuals Francis and Clare are not. If we disallow direct reference to Francis and Clare, or to any entity like the property of *being identical to either Francis or Clare* which could perform the same logical role, then *G* cannot be defined in terms of *poverty* alone. Given this, there is no nomological equivalent of *G* in terms of *poverty*. *G* is neither definable in terms of, nor reducible to, *poverty*.

Like nearly every philosophical argument, this one involves some disputed assumptions, but if it is right then it illustrates the possibility of global supervenience without reducibility (a substantial simplification of the point first made by Hellman and Thompson (1975)).

7. *Conclusion*

If global supervenience yields determination without reducibility then this is good news indeed for nonreductive physicalists and non-reductive individualists. The determination component articulates a clear sense in which the social and the mental are nothing over and above the physical. In creating the actual world God would not first have to create physical stuff and the laws governing it, then add a whole layer of extra mental entities, properties, and laws, and finish it up with social entities and laws. The physical by itself would suffice, guaranteeing its mental and social superstructure. The nonreductive component, how-

ever, would guarantee that mental and social concepts would still be essential for discourse and theorizing. The social sciences could enjoy methodological autonomy without having to apologize for a bloated ontology.

See also: Atomism and Holism: Philosophical Aspects; Causes and Laws: Philosophical Aspects; Functional Explanation: Philosophical Aspects; History of Science; History of Science: Constructivist Perspectives; Individualism versus Collectivism: Philosophical Aspects; Physicalism and Alternatives; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Armstrong D M 1968 *A Materialist Theory of the Mind*. Routledge, London

Bacon J 1986 Supervenience, necessary coextension, and reducibility. *Philosophical Studies* **49**: 163–76

Burge T 1979 Individualism and the mental. *Midwest Studies in Philosophy* **4**: 73–121

Carnap R 1928 *Der Logische Aufbau der Welt*. Weltkreisverlag, Berlin

Carnap R *Philosophical Foundations of Physics*. Basic Books, New York

Currie G 1984 Individualism and global supervenience. *British Journal for the Philosophy of Science* **35**: 345–58

Davidson D 1970 Mental events. In: Foster L, Swanson J W (eds.) *Experience and Theory*. University of Massachusetts Press, Amherst, MA, pp. 79–101

Francis of Assisi St 1903 *The Little Flowers of Saint Francis* [trans. Arnold T W]. Dent, London

Hellman G, Thompson F 1975 Physicalism: ontology, determination, reduction. *Journal of Philosophy* **72**: 551–64

Kim J 1978 Supervenience and nomological incommensurables. *American Philosophical Quarterly* **15**: 149–56

Kim J 1989 The myth of non-reductive materialism. *Proceedings and Addresses of the American Philosophical Association* **63**: 31–47

Lewis D 1993 *Philosophical Papers*. Oxford University Press, New York, Vol. 1

Nagel E 1961 *The Structure of Science*. Harcourt, Brace & World, New York

Oddie G 1991 Supervenience, goodness, and higher-order universals. *Australasian Journal of Philosophy* **69**: 20–47

Oddie G, Tichȳ P 1990 Resplicing properties in the supervenience base. *Philosophical Studies* **58**: 259–69

O'Neill J 1973 *Modes of Individualism and Collectivism*. Heinemann Educational, London

Pettit P 1993 *The Common Mind*. Oxford University Press, New York

Putnam H 1967 Psychological predicates. In: Capitan W H, Merrill D D (eds.) *Art, Mind and Religion*. Pittsburgh University Press, Pittsburgh, PA, pp. 37–48

Quine W V O 1961 Two dogmas of empiricism. *From a Logical Point of View*. Harvard University Press, Cambridge, MA

Ramsey F P 1931 *The Foundations of Mathematics*. Kegan Paul Tranch, Trubner, London

Russell B 1919 *Introduction to Mathematical Philosophy*. Macmillan, New York

Ryle G 1949 *The Concept of Mind*. Barnes & Noble, New York

Smart J J C 1959 Sensations and brain processes. *Philosophical Review* **68**: 141–56

G. Oddie

Reference and Representation: Philosophical Aspects

Paintings, barometers, sentences, and minds are examples of things that represent—they represent things as being one way or another. In this sense, they can be thought of as being about, or as referring to, certain things: the putative subject of the painting, air pressure, or approaching weather, etc. This article addresses the connections between representation and reference; the notion of representational content; representation in language and mental representation; discusses some theories of reference for names; and discusses some issues about the representational content of mental states.

1. Representation: the Most General Notion

What is common to all cases of representation is a correspondence between various states of what does the representing and various states of what is represented. Thus, the number of tree rings represents the age of a tree by virtue of the fact of the possibility of matching different tree ring states with the different ages. In this case, the correspondence is a very simple one: n tree rings corresponds to an age of n years. Likewise, there is a simple correspondence between various states of petrol gages and various levels of gas in the tank to which the gage is connected. But there are highly complex cases as well. Inscriptions in secure codes of credit card numbers represent those numbers, but via extraordinarily complex correspondences. When there are pairwise correspondences between differing states, there is the possibility of carrying information. This is how the number of tree rings can give information about the age of a tree. But, of course, the nature of the correspondence needs to be known for this possibility to be realized.

In many cases, the relevant correspondence relation has a causal underpinning that runs from the various states being represented to the various states doing the representing. For example, the level of gas in the tank causes the reading on the gage, and the appearance of the subject causes the array of paint on the canvas (via the actions of the painter). But structures can represent how things *will* be; a drawing may represent where a playground will be located but is, obviously, not

caused by where it will be. There is representation in the wide sense when there are nonaccidental correspondences between states of affairs. There will be a causal explanation of these nonaccidental correspondences but its precise nature varies from case to case. (See Dretske 1995, Chap. 1 for more on the variety, but note that he ties representation to *purpose*.)

1.1 Representational Content and Reference in General

Things that represent are said to have representational content, that content being how things are being represented as being. The usual way of capturing representational content is set theoretically: content is given by a division of all the complete ways things might be—the event space in the language of statistics—into those that accord with how things are being represented as being and those that do not. Complete ways things might be are ways things might be in which *everything* is settled one way or the other. They are often called possible worlds. The content is then the set of possible worlds that accords with how things are being represented as being. For example, if a petrol gage represents a gas tank as half-full, its being half-full is the content, and that is identified with the set of possible worlds where the tank is half-full. There are, though, considerable differences of opinion about the nature of the possible worlds appealed to in this style of account. This article touches briefly on these in the discussion of representation in mind and language. (For an accessible discussion of the main views, see Bigelow and Pargetter 1990, Chap. 4.)

One advantage of the set-theoretic approach to content is that it allows a simple account of when things are as they are being represented as being. Things are as they are being represented as being—the representation is true or correct or accurate—if and only if the way things actually are is a member of the content. Or, to say it in terms of possible worlds, the representation is accurate if and only if the actual world is a member of the set of worlds which is the content of the representation. In still other words, states that represent say that things are a certain way in the sense of delimiting the possibilities for how they might be, and count as accurate when how things actually are is inside the relevant region of 'logical space.'

A second advantage is that it allows a simple account of the widest sense in which states that represent are about, or refer to, this, that, or the next thing. Suppose that a barometer's reading such and such represents that the surrounding air pressure is so and so. Intuitively, the reading is about the air pressure being so and so. This can be spelt out as follows. In every possible world where the reading is correct, in every world where things are as the barometer represents them, the air pressure is so and so. States which represent are about or refer to what is the case in all the worlds where things are as they are being represented as being. (For more on 'aboutness' and the difference between being about an object and a property, see Dretske 1995, Chap. 1, Sect. 5.)

1.2 Mental Representation

A footprint can be thought of as representing the shape of the foot that made it an hour ago. But equally it can be thought of as representing the weight of the person who made it, the shape of the footprint itself half an hour ago, the shape of the foot that made it one day before the foot print was made, and so on. Almost any state that represents will have many representational contents, and it is a nonsense question to ask which is its real content, to ask what the state really represents.

Among the representational structures of especial interest to humans are their brains, and, especially, those parts of our brains that constitute our powers of mental representation. In particular, humans *believe* that things are thus and so, and do so by virtue of the processes in their brains which constitute their belief representing that things are thus and so. These processes *qua* states of brains will themselves have many contents including, for example, that the temperature of the surrounding brain fluid is thus and so, but only one will be the content of what is believed. To suppose otherwise would make what is believed absurdly indeterminate. Therefore, a major question in the philosophy of mind is what determines, out of all the contents of the brain structures that are beliefs, the one which is the content of those beliefs.

One popular answer is: the functional role the structure plays, where this is understood as the role the structure plays in mediating between: inputs to subjects, typically via their sense organs, outputs understood in terms of behavioral responses, and other internal states. However, there are many different ways of fleshing out this answer. (For a recent survey, see Braddon-Mitchell and Jackson 1996, Chaps. 3, 5.)

2. Linguistic Representation and Reference

2.1 Introduction

Language is an especially powerful representational device. With it, highly detailed claims are possible concerning how things are that far outrun, in terms of detail and precision, what can be done with barometers, petrol gages, etc. There is a syntactical distinction between sentences and parts of sentences that roughly corresponds to an important semantic or representational distinction.

A sentence makes a claim about how things are. This is how it gets to be true or false: true when the way the sentence represents things is the way they are, false otherwise. This explains the self-evident status of the biconditional 'Snow is white' is true if and only if snow is white. The left-hand side is true if and only if things are as 'Snow is white' represents them as being. Moreover, as a good way of stating how 'Snow is white' represents things as being is to use the sentence itself, the biconditional comes out true.

By contrast, parts of sentences like 'Snow' and 'is white' contribute in a systematic way to how whole sentences in which they appear represent things as being, but do not themselves represent how things are. This is why they are not themselves either true or false. They can, however, be thought of as referring to, respectively, snow and white things. This is because the contribution that the presence of the word 'snow' makes to the conditions under which 'Snow is white' is true, that is, to how the sentence represents things as being, is such that the sentence cannot be true unless there is some snow. Similarly, the contribution that the presence of 'is white' makes is such that the sentence cannot be true unless there are some white things.

Many recent discussions have focused on the referential properties of names, and many authors restrict the use of the term 'refer' to names. They prefer to talk of the relation between predicates (and, sometimes, common nouns) and things in terms of whether the predicates are *true of* the things or whether the things *satisfy* the predicates.

The next section considers the three main theories of reference for names.

2.2 *Theories of Reference for Names*

A word like 'London' is used for something taken to have certain properties. This is why it is not used for water or Paris—it is known that they do not have the right properties. On the other hand, the large city that is the capital of England does have the right properties. In short, certain properties or descriptions are associated with the word 'London' and this guides its use. This suggests a simple theory of reference for names in general: 'N' refers to that which satisfies the descriptions associated with 'N.' If we use 'D' for those descriptions, another way of stating this theory is by saying that 'N' is short for the definite description 'the D.'

A famous advantage of the *description theory of reference* is that it can explain why certain identity sentences are informative. Everything is, trivially, self-identical. How then can it be informative to be told that Hesperus is identical to Phosphorus? The answer is that the first name is associated with being the star first seen in the evening and the second name is associated with being the last star to disappear from the night sky as morning approaches, and it is informative to be told that one thing has these two properties. (The classic source for claims of this kind is Frege 1952.)

Recent discussions in the philosophy of language have raised a number of objections to the description theory of reference. (A good account of the major objections can be found in Devitt and Sterelny 1987. A classic source is Kripke 1980.)

First, if 'N' is an abbreviation of 'the D,' it should make no sense to suppose that N might have failed to be D. But, in fact, it typically makes perfect sense. For example, the descriptions associated with 'Shakespeare' include as a major part that he was the author of certain plays. (Surprisingly little is known about Shakespeare's life.) But it makes perfect sense to suppose that Shakespeare did not write the plays associated with his name. This, indeed, is what those who think that Bacon wrote them do suppose. A similar point can be made for names of natural kinds. The description associated with 'water' is something like potable, clear liquid that fills lakes and falls as rain from the sky—this is roughly what is in a dictionary. However, if on some remote planet—traditionally called 'Twin Earth' by philosophers—there is found clear potable stuff in its lakes which falls from its sky, etc. that is not H_2O, it seems that the right thing to say would be that this stuff—XYZ, as it is often dubbed—is not water. As it would naturally be said, 'It has been discovered that water is not the only clear potable liquid that fills lakes, falls from skies etc. in the universe.' Conversely, if H_2O on Twin Earth were black and rather tar-like, it would still be water. As it would naturally be said, 'It has been discovered that water comes in a quite different form on Twin Earth from its form at room temperature on Earth.' (The classic source for Twin Earth is Putnam 1975.)

Second, if names are abbreviated definite descriptions, then their reference should in general be variable. By changing what is supposed to be the case, it should be possible to shift what they refer to. This is because the reference of definite descriptions is in general shiftable. 'The President of the USA in 1998' refers to Bill Clinton, but had he lost to George Bush, it would refer to Bush and not Clinton. But it seems that the reference of names cannot be shifted; they are, as it is said, rigid designators. 'Clinton' always refers to the very same person.

Description theorists have proposed a number of replies to these (and other) objections. For example, they have urged that it is wrong to conflate the description that settles the reference of a name with the description (s) most famously associated with a name. Although Shakespeare is most famous for the plays, the key description that settles the reference of the name is something like 'the person who is a certain kind of causal origin of the use of the term "Shakespeare." ' The debate over whether or not Shakespeare wrote the plays is the debate over whether the person who is a certain kind of causal origin of the occurrences

of the name on the covers of plays etc. did indeed write the plays. (For a recent catalogue of replies that the description theorists might make to various objections, see Jackson 1998.)

The causal theory of reference is the best known alternative to the description theory. The fundamental idea is that 'Aristotle' in our mouths refers to the person it does because of an initial act of baptism that tags that person with a name which then got passed on from speaker to speaker as the name to use for that tagged person until it ended up in our mouths. Of course, the name itself underwent various changes as it passed from the ancient Greek to Modern English. A similar account is suggested to explain why 'water' refers to H_2O. H_2O was baptized with the word 'water' (in ignorance of the fact that it was H_2O that was being baptized) and then the word was passed on as the word to use for that which had been baptized. This account explains why the clear potable stuff that fills the lakes etc. on Twin Earth is not water, is not what is referred to when the word 'water' is used, by appealing to the fact that it was not the natural kind baptized.

There are a number of questions to ask about the causal theory. Is it superior to a description theory that gives a prominent role to a description like 'was baptized ... and subsequently used for what had been so baptized'? Is it supposed to apply to the naming of as yet unborn children? Remembering that causal interactions are invariably with a very large number of factors, should a descriptive element be added to ensure the 'right' reference? (Some of these questions are addressed in Devitt and Sterelny 1987.)

Finally, there is the direct reference theory of reference. On this view, names like 'Aristotle' refer directly to Aristotle. The only semantic property of a name is its bearer. It is hard to discuss this view because it steadfastly refuses to address what seems a perfectly legitimate question, namely, what makes it true that 'Aristotle' refers to Aristotle and not to Plato or my dog? It is also hard to see what it can say about empty names. It has sometimes been speculated that Socrates never existed. 'He' was merely an expository device of Plato's. Although this is almost certainly false, it makes sense to suppose that it is true. But is this possible if 'Socrates' has no meaning apart from its bearer? (For references and more discussion and dissent, see Devitt 1996, Chap. 4.)

3. *Mental Representation and Linguistic Representation*

3.1 *How are they Connected?*

As noted above, our minds represent; in particular, to believe is to represent that things are thus and so. Usually, the easiest way for humans to express how they believe things are is to use a sentence. For example, the sentence 'Clinton will not be impeached' was used in February 1999 by many commentators on American politics to express what they believed. It is obvious that there is an intimate connection between mental and linguistic representation. What is controversial is the extent to which mental representation requires linguistic representation.

One view is that the phenomenon of mental representation in general, and belief in particular, is independent of language; dogs, for example, really do have thoughts and beliefs. Language is simply a device for telling about how things are taken to be. This view can though allow that there are some beliefs one could not have without having a language. It is, for instance, hard to make sense of a creature believing that there are a prime number of apples in the bowl in front of it in the absence of that creature having mastery of a complex coding system, i.e., a language.

However, some hold that belief in general requires language, where language means something substantially more sophisticated than the kind of signaling systems that, e.g., dogs have. On this view, although contentful utterance requires *mental* representation—telephone answering machines do not *mean* anything by the sentences they produce—in order to be a believer, you must have a language. Sometimes, this seems to be, at bottom, a verbal stipulation about the use of the term 'belief': dogs have quasibelief which is very like belief except that it is not expressed in a public language. But sometimes it is clear that the view is that dogs do not have anything like belief at all. A major question for this view is the very close connection between perception and belief. Perceptual experience seems to be shot through with belief-like representational content: the experience of it's raining seems to be a state that by its very nature points to believing that it is raining, though it may be overridden by collateral information, for example, about the operation of a nearby sprinkler system. It is, accordingly, hard for advocates of the more radical view to avoid the implausible view that animals like dogs do not see or hear. (For a defense of the view that belief requires language, see Davidson 1984, Chap. 11, see also McDowell 1994.)

3.2 *Capturing the Content of Mental and Linguistic Representation*

It was noted earlier that the standard way of capturing representational content is in terms of divisions among complete ways things might be, possible worlds. The problem of the content of logically equivalent sentences and beliefs raises an important question for how to think of these possible worlds. Intuitively, the sentences 'X is an equilateral triangle' and 'X is an equiangular triangle' differ in meaning. Likewise, intuitively, the belief that X is an equilateral triangle and the belief that X is an equiangular triangle are different beliefs. In both cases, the intuition rests on

the fact that it was a *discovery* that, necessarily, every equilateral triangle was equiangular, and conversely. But, on standard conceptions of possible worlds, 'X is an equilateral triangle' and 'X is an equiangular triangle' are true in exactly the same possible worlds precisely because they are logically equivalent.

There have been, broadly speaking, two reactions to this problem. One has been to bite the bullet and insist that the contents are the same. (The most worked out version of this response can be found in Stalnaker 1984.) The other has been to argue that, for the purposes of capturing content, a more fine-grained conception of the entities among which the division is being made is needed. For example, possible worlds might be thought of as constructions out of certain concepts so that, for example, the constructions at which 'X is an equilateral triangle' is true differ from those at which 'X is an equiangular triangle' is true by virtue of the former having the concept *side*, where the latter have the concept *angle*. (See Peacocke 1983, for an overview of the issues see Braddon-Mitchell and Jackson 1996.)

3.3 Mental Representation and the Language of Thought

The functionalist answer to how a belief gets to have the belief content that it does—and, more generally, how a mental state gets to be the mental state that it is—was described earlier, and it was noted that there are many views about how to spell out the key functional roles. However, a popular answer holds that an important part of any functionalist account is the way belief and desire work together to produce behavior: roughly, subjects behave in a way such that what they desire is satisfied if what they believe is true (Stalnaker 1984, Braddon-Mitchell and Jackson 1996).

An alternative approach—the language of thought hypothesis—appeals to the idea that creatures capable of representational thought, creatures capable of having intentional mental states, have a language of thought. When subjects believe that snow is white, they have in their heads a sentence in 'mentalese' that says that snow is white. (This approach allows that dogs may have beliefs, because it is possible to have an internal language of thought, while lacking a public language in which to express thoughts.) This approach gives functional role a place in determining the difference, for example, between belief and desire, but content is treated via a 'semantics for mentalese.' Consider, for example, subjects who believe that snow is white. They have in their heads a token that stands for snow (the mentalese word for snow), a token that stands for whiteness (the mentalese predicate for whiteness), put together in a structure that makes up the mentalese sentence *snow is white*. What makes it true that they believe rather than desire that snow is white is the functional role this structure plays: it serves more of a 'responding to the world' role and less of a 'changing the world' role. (A classic source is Fodor 1988.)

This approach raises many important issues. One is how we should spell out the crucial relation between the words of mentalese and what they stand for? What makes it true that a structure in the brain is the mentalese for snow and not water, for example? One proffered answer is some kind of causal cum informational link between brain token and what is tokened, but the details are controversial. A second issue is whether neurological investigations might refute the language of thought hypothesis. Some cognitive scientists have argued that the way the brain represents the environment is more like the way maps represent than the way sentences represents (see Braddon-Mitchell and Jackson 1996).

4. Representation and Context

Representational content is often thought of as context dependent. For example, what a petrol gage represents about the level of gas in the tank depends on the nature of the connection between gage and tank. Connected in one of the more usual ways, the gage having its pointer in the far left position indicates that the tank is empty, but connected differently that pointer position might represent that the tank is full. That is to say, identical gage configurations can have very different representational contents depending on the gage's 'context.' Or, as it is often put, representational content with respect to the amount of gas in the tank does not *supervene* on a petrol gage's configuration. As it is said, the content is broad or wide or context dependent.

It is equally possible, however, to think of any given reading as saying: either the gage is connected to the tank in way W_1 and the gas level is L_1, or the gage is connected in way W_2 and the level is L_2, or ... . In this way of thinking, a way naturally adopted when the manner of connection between gage and tank is unknown, the content is not dependent on the manner of connection.

In other words, and speaking more generally, there are two essentially equivalent ways of thinking of the interplay of representational content and context. In the first, it looks like this:

In C_1, S represents that P_1
…
In C_i, S represents that P_i

In the second, it looks like this:

S represents that if C_1 then P_1
…
S represents that if C_i then P_i

In the first way of thinking, content comes out as context dependent or broad; in the second, it comes out context independent or, as it is said, narrow.

A pressing, currently much debated question about both mental and linguistic content is whether we should think of it in the context dependent or in the context independent way. Or, equivalently, is what a person believes and what a person claims about how things are when they speak and write a broad or narrow property of them?

The traditional view is that mental and linguistic content is narrow. A principal source for this view is a powerful intuition often exploited by science fiction writers. The idea is that if the stimuli actually impinging on subjects via their environments is duplicated *exactly*, say, by a computer, the subjects will not be able to tell the difference. This is the core idea behind flight simulators and virtual reality machines, of course. Although these devices fall short of perfectly duplicating the impacts of their target environments on their subjects, their makers take it for granted that were perfect duplication achieved, subjects would be unable to tell whether, for example, they were really flying or were in a simulator. This suggests that subjects' mental lives, including how their mental states represent things as being, supervene on their nature from the skin in (or, perhaps, from their central processors in). How their mental states represent things as being is, therefore, a narrow property of subjects: two subjects which are identical from the skin in are identical in how they take, desire etc. their worlds to be. Moreover, if the content of sentences is given by the content of the mental states they serve to express, it will follow that linguistic content is narrow.

However, many philosophers, often influenced by the Twin Earth example discussed earlier, repudiate this intuition. Consider, runs their argument, Mary and Mary's duplicate from the skin in, Twin Mary, on Twin Earth. They both use the word 'water' for the clear, potable etc. stuff around them, but the conditions under which what is said is true, and equally those under which what is thought is true, differ. What Mary says and thinks when she uses the sentence 'There is water nearby' is true if and only if there is H_2O nearby, but what Twin Mary says and believes will be true if and only if there is XYZ nearby. Hence, runs the argument, the contents of the words and thoughts of Mary and Twin Mary differ; ergo, content is context dependent.

The issue is complicated, however, by the fact that it is plausible that there is an indexical element to the use of a great many words including 'water.' It is plausible that the word 'water' is used in part to mean the stuff that the users of the word have come across which is clear, potable etc. If this is right, a plausible explanation of why Mary and Twin Mary refer to different stuff in thought and talk when they use the word 'water' is that they and their language communities interact with different stuff when the word 'water' is used. That is, Mary refers to the clear, potable stuff that s*he* and *her* community interact with, whereas Twin Mary refers to the stuff s*he* and *her* (different) community interact with, and the reason they refer differently is that the stuff they interact with is different. This means that there is *a* sense in which what they say and think is the same when they say and think 'There is water nearby.' They both say and think that some of the clear potable etc. stuff that she herself interacts with is nearby. This does not affect the point that the truth conditions of what they say and think differ, but they differ in the same way that the truth conditions of 'I am bald' differ depending on who is speaking. There is a good sense in which each person who says or thinks 'I am bald' says or thinks the same thing, namely, that I myself am bald. (There is a large literature on this important question; see Pessin and Goldberg 1996 and Braddon-Mitchell and Jackson 1996.)

See also: Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Meaning and Rule-following: Philosophical Aspects Cognitive Science: Overview; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective; Meaning and Rule-following: Philosophical Aspects; Methodological Individualism: Philosophical Aspects; Perception: Philosophical Aspects

Bibliography

Bigelow J, Pargetter R 1990 *Science and Necessity*. Cambridge University Press, Cambridge, UK

Braddon-Mitchell D, Jackson F 1996 *The Philosophy of Mind and Cognition*. Blackwell, Oxford, UK

Davidson D 1984 *Inquiries into Truth and Interpretation*. Oxford University Press, Oxford, UK

Devitt M 1996 *Coming to our Senses: A Naturalistic Program for Semantic Localism*. Cambridge University Press, Cambridge, UK

Devitt M, Sterelny K 1987 *Language and Reality*. MIT Press, Cambridge, MA

Dretske F 1995 *Naturalizing the Mind*. MIT Press, Cambridge, MA

Fodor J 1988 *Psychosemantics*. MIT Press, Cambridge, MA

Frege G 1952 Translations from the philosphical writing of Gottlob Frege. In: Geach P, Black M (eds.) *Translations from the Philosophical Writings of Gottlob Frege*. Philosophical Library, New York

Jackson F 1998 Reference and description revisited. *Philosophical Perspectives* **12**: 201–18

Kripke S 1980 *Naming and Necessity*. Harvard University Press, Cambridge, MA

McDowell J 1994 *Mind and World*. Harvard University Press, Cambridge MA

Peacocke C 1983 *Sense and Content*. Oxford University Press, Oxford, UK

Pessin A, Goldberg S (eds.) 1996 *The Twin Earth Chronicles*. M E Sharpe, Armonk, NY

Putnam H 1975 The meaning of 'meaning'. In: Putnam H (ed.) *Mind, Language and Reality*. Cambridge University Press, Cambridge, UK
Stalnaker R 1984 *Inquiry*. MIT Press, Cambridge, MA

F. Jackson

Referendums

The referendum allows citizens to vote on a public issue in an election at the local, subnational, or national level. The vote may be advisory or mandatory, statutory or constitutional. Most democracies have voted on a referendum, and the incidence of referendums increased in the last quarter of the twentieth century. The referendum agenda does not have a distinctive ideological slant (Butler and Ranney 1994). National and subnational referendums on matters like entry into international agreements, new constitutional arrangements, and changes in monetary systems have been frequent topics put to voters. Recently in the USA conservative issues have driven the process, but liberal concerns dominated in other decades (Magleby 1994).

Several notable referendums from the last quarter of the twentieth century include the vote in the UK on entry into the European Economic community, the separatist votes in Quebec, the devolution votes in Scotland and Wales, the votes in several European countries on a common currency, and a growing interest in the process in the new democracies of Africa, South America, or Eastern Europe. The USA is one of the few democracies that never experimented with a national referendum, although the process has occasionally been proposed. One notable proposal for a national referendum in the USA was the failed Ludlow Amendment to the US Constitution, which would have required a national referendum before war could be declared (Magleby 1984). More recently, forms of national referendums have been proposed by former Representative Jack Kemp, among others (Kemp 1979).

Referendums can be proposed by the elected representatives and put before the voters in a special or regular election. When such referendums are optional they are called facultative (Bogdanor 1994). The popular referendum permits citizens to petition to place a law or constitutional change before the voters that has previously been enacted by the legislature.

Referendums are sometimes called plebiscites, a reflection of the early use of the process in the Greek city-states. More recently, critics of the process point to plebiscites called by Hitler and other totalitarian leaders as examples of how the process can be used for undemocratic ends. Referendums were found to be consistent with representative democracy in the USA by the Supreme Court (*Pacific States Telephone and Telegraph Company vs. Oregon* 1912).

In their expansive review of referendums around the world, David Butler and Austin Ranney (1994) find that the subject matter of referendums falls into four main categories: constitutional issues especially after a revolution, territorial breakup, or when changes in the rules are needed; territorial issues; moral issues on matters that reflect deep divisions and which may best be resolved by a popular vote (e.g., alcoholic beverage regulation, divorce, or abortion); and finally a range of miscellaneous issues.

A more common form of citizen petitioning is the initiative process. Under the initiative, citizens may petition to place before the voters statutes (statutory initiative) or constitutional changes (constitutional initiative). As with the popular referendum, access to the ballot is based on the proponents obtaining a specified number of petition signatures. Some governments permit the direct initiative whereby petitioners go directly to the ballot, while some require the proposal to be considered by the legislature before it may go to the voters (the indirect initiative).

At the national level, the initiative process is more identified with Switzerland. Swiss cantons also use the initiative with as many as four elections per year where voters decide initiatives or referendums (Kobach 1994).

Imitating the Swiss, several American states adopted the initiative and popular referendum process—sometimes called direct legislation—shortly after the turn of the last century. As with other Progressive reforms like the direct election of US Senators and the direct primary, reformers hoped to expand the role of voters and limit the role of elected officials and intermediary institutions like political parties, legislatures, or elected executives. The United States experienced a second wave of reform during the 1970s and since; these reforms again expanded suffrage, pressed for more expansive use of direct primaries especially in presidential elections, and renewed interest in and use of initiatives and popular referendums.

In the United States, the initiative and popular referendum occur most frequently in western states (see Fig. 1). These states were in their formative political years during the peak of the movement, and this region was a stronghold for the Progressives (Price 1975). Only five states have adopted the process since the 1920s: Alaska (1956), Wyoming (1968), Illinois (1970), and Florida (1972). Since the mid-1970s, several states have considered adopting the process but only Mississippi has done so. Direct legislation is even more common at the local level in the United States, although precise counts of the numbers of jurisdictions with the process nor the extent of use do not exist.

As in Switzerland, the most common form of initiative and popular referendum in the USA requires

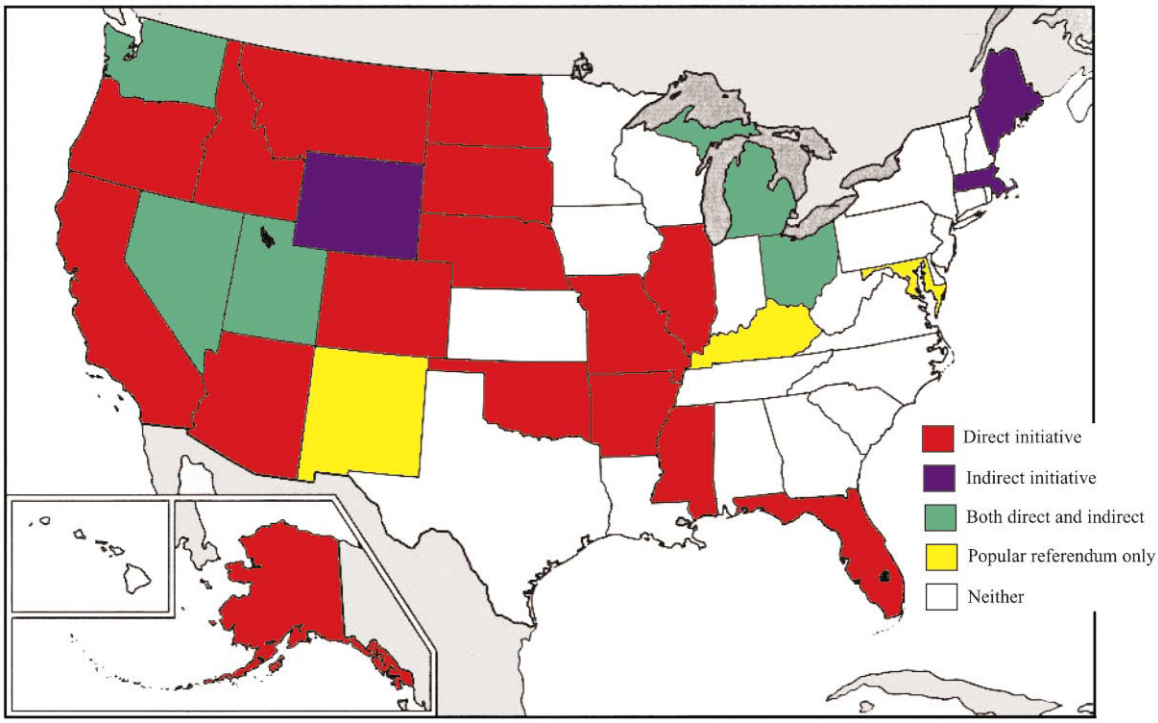

Figure 1
Provisions for initiative and popular referendum in the USA

voters to vote for or against specific statutes or constitutional amendments. As a result, ballot measures are lengthy and written in technical and legal language. Official summaries provided by state election officials are also written at a sophisticated level, making it difficult for voters to often obtain a simple neutral summary of the proposition (Magleby 1984). The length and complexity of the process in the USA makes the process less accessible to less educated voters, having a policy bias against the poor and racial minorities. Initiatives are preferred over popular referendums because petitions may be in circulation longer, and initiatives permit their sponsors to write the language of the measure rather than the legislature. An example of an initiative that was in effect a referendum is the 1964 vote on open housing in California (Wolfinger and Greenstein 1969).

Referendums put on the ballot by state legislatures or changes to the state constitution occur frequently. Every state except Delaware requires that changes in the state constitution be approved by voters. Voters are much more likely to vote affirmatively on these legislative referendums than they are on initiatives. More than 60 percent of legislative referendums pass while the same percent of initiatives fail (Magleby 1994).

Since the mid-1970s, the initiative has experienced renewed usage (see Fig. 2). The growth in initiative usage includes an even more dramatic increase in the number of proposed initiatives in the petition circulation process that fail to reach the ballot. Why the increase? Initiative activists have learned that the process is a powerful agenda-setting device. A vote in a single state can push an issue onto the national agenda, as votes on tax reduction, elimination of affirmative action, a nuclear weapons freeze, and legislative term limitation has demonstrated. The news media give substantial attention to contentious initiatives, calling attention to the sponsors of the measures and, in some instances, advancing political careers. California gubernatorial candidates make frequent use of the process for self-promotion and governors in noninitiative states advocate the process to establish populist credentials. The large sums of money spent on initiatives have also called attention to the renewed use of the process. Not surprisingly, an initiative industry has developed that helps with petition circulation, legal appeals, campaign management, and media advertising. The process has become an integral part of the governance of several states, most notably California (Schrag 1998).

Absent from the initiative process are the normal checks and balances of the US political system. There is no bicameral check except for Nevada's requirement

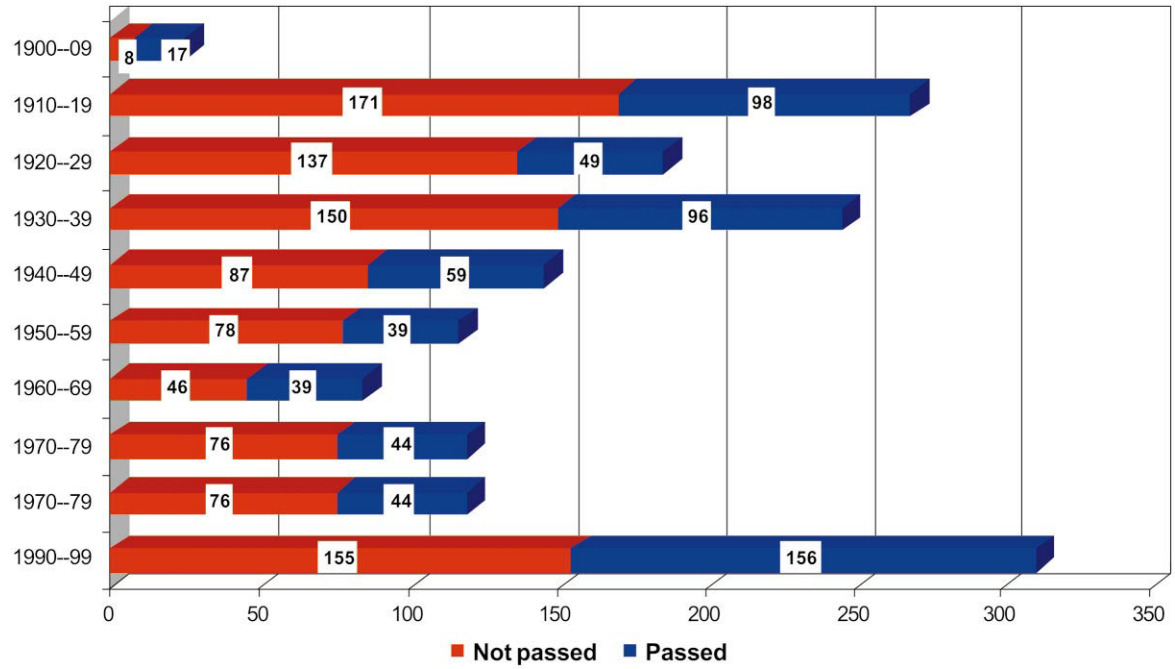

Figure 2
History of initiatives and referendums USA 1900–97. Source: Initiatives and Referendums Institute

that constitutional initiatives must pass in identical form in two consecutive elections. There is no gubernatorial veto, no hearing process, and no ability to amend or modify. The only institutional check on the initiative is the courts. Judicial review of initiatives occurs often before the election on matters having to do with signature collection, ballot title, and summary, and whether the initiative contains more than one subject. Postelection judicial review is more concerned with constitutionality, and federal courts have not been timid about overturning the people's votes. State courts, like the California Supreme Court, faced a voter backlash in the 1980s because of the frequency with which that court had declared initiatives unconstitutional.

With the growth in democracy around the world, there has followed renewed interest in referendum devices as a mode of popular consultation. Indeed, some analysts point to a coming clash between direct and representative democracy in the USA, in part precipitated by technology (Broder 2000). States like Oregon have experimented with statewide referendums entirely with mail balloting and some see the Internet as a way to facilitate frequent referendum voting (Becker and Slaton 2000).

As the process has grown in importance, so has research on it. Much of the research focuses on voter decision making in the absence of party cues. Ideology and elite cues are important to voting in this context (Magleby 1984, Lupia 1994). Other research has focused on the role of consultants in the process, finding that these campaign professionals find the process more lucrative than candidate elections, and prefer it because it gives them greater creative license and means they can avoid the problems associated with candidate scandals and intrusive spouses (Magleby and Patterson 2000).

Another expanding research emphasis is the policy consequences of initiatives. Scores of books have been written on the impact of California's Proposition 13 property tax reduction and similar measures in other states (Sears and Citrin 1985, O'Sullivan et al. 1995). Other books examine the policy and institutional implication of the process (Gerber et al. 2001).

One major methodological challenge to researchers is the lack of a centralized database on local initiatives. Administration of local elections is left to counties and rarely are records kept for more than a few years. Given the large number of counties and the frequency of referendums at the local level, the task of assembling this database has proven difficult.

Given the likely ease of electronic referendum voting, research on how voters behave in this environment will be important. Do electronic voting schemes limit some voters' ability to participate? Would voting electronically further disenfranchise poor and less educated voters from a process that they already find difficult to navigate?

See also: Constitutionalism; Democracy; Democracy, History of; Electoral Systems; Legislatures: United States; Participation: Political; Political Representation

Bibliography

Becker T, Slaton C D 2000 *The Future of Teledemocracy*. Praeger, Westport, CT

Bogdanor V 1994 Western Europe. In: Butler D, Ranney A (eds.) *Referendums Around the World: The Growing Use of Direct Democracy*. AEI Press, Washington, DC

Broder D S 2000 *Democracy Derailed: Initiative Campaigns and the Power of Money*. Harcourt, New York

Butler D, Ranney A 1994 Practice. In: Butler D, Ranney A (eds.) *Referendums Around the World: The Growing Use of Direct Democracy*. AEI Press, Washington, DC

Gerber E, Lupia A, McCubbins M, Kieweit D 2001 *Stealing the Initiative: How State Government Responds to Direct Democracy*. Prentice-Hall, Upper Saddle River, NJ

Kemp J 1979 *An American Renaissance: A Strategy for the 1980s*. Harper and Row, New York

Kobach K 1994 Switzerland. In: Butler D, Ranney A (eds.) *Referendums Around the World: The Growing Use of Direct Democracy*. AEI Press, Washington, DC

Lupia A 1994 Shortcuts versus encyclopedias: Information and voting behavior in California insurance reform elections. *American Political Science Review* **88**: 63–76

Magleby D B 1984 *Direct Legislation: Voting on Ballot Propositions in the United States*. Johns Hopkins University Press, Baltimore

Magleby D 1994 Direct legislation in the American States. In: Butler D, Ranney A (eds.) *Referendums Around the World: The Growing Use of Direct Democracy*. AEI Press, Washington, DC

Magleby D B, Patterson K D 2000 Campaign consultants and direct democracy: Politics of citizen control. In: Thurber J E, Nelson C J (eds.) *Campaign Warriors: The Role of Political Consultants in Elections*. Brookings Institution Press, Washington, DC

O'Sullivan A, Sexton T, Shefrin S 1995 *Property Taxes and Tax Revolts: The Legacy of Proposition 13*. Cambridge University Press, Cambridge, UK

Pacific States Telephone and Telegraph Company vs. Oregon 233 US 118 (1912)

Price C M 1975 The initiative: A comparative state analysis and reassessment of a western phenomenon. *Western Political Quarterly* **28**: 243–62

Schrag P 1998 *Paradise Lost: California's Experience, America's Future*. University of California Press, Berkeley, CA

Sears D, Citrin J 1985 *Tax Revolt: Something for Nothing in California*. Harvard University Press, Cambridge, MA

Wolfinger R E, Greenstein F I 1969 The repeal of fair housing in California: An analysis of referendum voting. *American Political Science Review* **63**: 755

D. B. Magleby

Reflective Equilibrium

A 'reflective equilibrium' is the end point of a deliberative process in which we reflect on and revise our beliefs about an area of inquiry. The inquiry might be as specific as the question, 'What is the right thing to do in this case?' or 'Is this the correct inference to make?' Alternatively, the inquiry might be much more general, asking which theory or account of justice or right action we should accept, or which principles of inductive reasoning are justifiable. We can also refer to the process or method itself as the 'method of reflective equilibrium.' After briefly describing the method, we shall describe its history, the controversy surrounding it, and some implications it has for work in ethics.

1. *The Method of Reflective Equilibrium*

The method of reflective equilibrium consists in working back and forth among our considered judgments (some say our 'intuitions') about particular instances or cases, the principles or rules that we believe govern them, and the theoretical considerations that we believe bear on accepting these considered judgments, principles, or rules, revising any of these elements wherever necessary in order to achieve an acceptable coherence among them. The method succeeds and we achieve reflective equilibrium when we arrive at an acceptable coherence among these beliefs. An acceptable coherence requires that our beliefs not only be consistent with each other (a weak requirement), but that some of these beliefs provide support or provide a best explanation for others. We arrive at an optimal equilibrium when the component judgments, principles, and theories are ones we are uninclined to revise any further.

The method of reflective equilibrium has been advocated as a coherence account of justification in several areas of inquiry, including inductive and deductive logic as well as ethics and political philosophy. The key idea underlying this view of justification is that we 'test' various parts of our system of beliefs against the other beliefs we hold, looking for ways in which some of these beliefs support others, seeking coherence among the widest set of beliefs, and revising and refining them at all levels when challenges to some arise from others. For example, a moral principle or moral judgment about a particular case (or, alternatively, a rule of inductive or deductive inference or a particular inference) would be justified if it cohered with the rest of our beliefs about right action (or correct inferences) on due reflection and after appropriate revisions throughout our system of beliefs. By extension of this account, a person who holds a principle or judgment in reflective equilibrium with other relevant beliefs can be said to be justified in believing that principle or judgment.

Because there is revisability of our beliefs at all levels, this coherence view contrasts sharply with foundationalist approaches to justification. A foundationalist approach would take some of these beliefs—some theoretical claim, or moral principles or judgments—as fixed. As we shall see, however, the point of

controversy, especially in ethics, is not that reflective equilibrium allows for the revision of moral judgments, but rather that it involves giving some justificatory weight to them.

2. Recent History

This approach to the justification of rules of inductive logic was proposed by Nelson Goodman in his classic *Fact, Fiction, and Forecast* (1955). Goodman's idea was that we justify rules of inference in inductive or deductive logic by bringing them into reflective equilibrium with what we judge to be acceptable inferences in a broad range of particular cases. No rule of inference would be acceptable as a logical principle if it was not compatible with what we take to be acceptable instances of inferential reasoning. In this sense, our beliefs about acceptable rules of inference are constrained by the 'evidence' provided by what we believe to be good or correct examples or instances of inferential reasoning. At the same time, we should correct or revise our views about particular inferences we initially might think are acceptable if we come to see them as incompatible with rules that we generally accept and refuse to reject because they, in turn, best account for a broad range of other acceptable inferences.

Despite the fact that the origins of reflective equilibrium lie in mid-twentieth century discussions about the justification of inductive logic, its principal development through the rest of the century lies primarily in ethics and political philosophy. Specifically, the method was given prominence by John Rawls's description and use of it in *A Theory of Justice* (1971). Most of our discussion will concentrate on its use in ethics and political philosophy. (Although accounts of the justification of empirical knowledge were developed that share with reflective equilibrium its coherentist approach, they generally do not make explicit use of the terminology of reflective equilibrium.)

Rawls argues that the goal of a theory of justice is to establish the terms of fair cooperation that should govern free and equal moral agents. On this view, the appropriate perspective from which to choose among competing conceptions or principles of justice is a hypothetical social contract or choice situation in which contractors are constrained in their knowledge, motivations, and tasks in specific ways. Because this choice situation is fair to all participants, Rawls calls the conception of justice that emerges from this choice 'justice as fairness.' Under these constraints, he argues, rational contractors would choose principles guaranteeing equal basic liberties and equality of opportunity, and a principle that permitted inequalities only if they made the worst-off groups in society as well off as possible.

Instead of simply accepting whatever principles contractors would choose under these constraints on choice, however, Rawls imposed a further condition of adequacy on them. The chosen principles must also match our considered judgments about justice in reflective equilibrium. If they do not, then we are to revise the constraints on choice in the contract situation until we arrive at a contract that yields principles that are in reflective equilibrium with our considered judgments about justice. In effect, the device of the contract must itself be in reflective equilibrium with the rest of our beliefs about justice. The contract helps us determine what principles we should choose from among competing views, but the justification for using it and designing it so that it serves that purpose must itself derive from the reflective equilibrium that it helps us achieve.

The method of reflective equilibrium thus plays a role in both the construction and justification of Rawls's theory of justice. Critics of Rawls's theory and his method of reflective equilibrium, especially utilitarians, challenge the prominence the method gives to moral judgments or intuitions. Indeed, some argue that there is no justificatory force to the contract itself if we 'rig' the contract so that it yields principles that match our intuitions (Hare 1973). Despite various criticisms, defenders of the method have elaborated it and extended its use in broad areas of ethics. Later in this entry we shall consider some of the criticisms of this method and some of its extensions in more detail. Before doing so, however, it is necessary to describe it more fully.

3. Narrow and Wide Reflective Equilibrium

3.1 Narrow Reflective Equilibrium

A reflective equilibrium may be *narrow* or *wide* (Rawls 1974). All of us are familiar with a process in moral deliberation in which we work back and forth between a judgment we are inclined to make about right action in a particular case, and the reasons or principles we offer for that judgment. Often we consider variations on the particular case, 'testing' the principle against them, and then refining and specifying it to accommodate our judgments about these variations. We might also revise what we say about certain cases if our initial views do not fit with the principles we grow inclined to accept.

Such a revision may constitute a moral surprise or discovery (Daniels 1996). Suppose, for example, that we are considering whether we should ignore age in the distribution of medical treatments. Many initially believe that age is a 'morally irrelevant trait,' just like race, and they would insist that rationing medical services by age is just as unacceptable as rationing by race. On considering a variety of cases, however, it might become apparent that we all age, but that we do not change race, and that this means that the different treatment of people at different ages, if systematically applied over the life span, does not create inequalities

between persons, as it would in the case of race. We might be led by this realization to think that age rationing might be acceptable under some conditions, when race rationing would never be, and this would be a moral surprise for many who changed their view.

To the extent that we focus solely on particular cases and a group of principles that apply to them, we are seeking only *narrow* reflective equilibrium. Presumably, the principles we arrive at best 'account for' the cases examined. Others, however, may arrive at different narrow reflective equilibria, containing different principles and judgments about justice. As a result, we still face an important question about justification unanswered by the method of narrow reflective equilibrium: which set of beliefs about justice should we accept?

Because narrow reflective equilibrium does not answer this question, it may seem to be a *descriptive* method appropriate to moral anthropology, not a *normative* account of justification in ethics. In fact, Rawls (1971) at one point suggested that arriving at the principles that match our moral judgments in reflective equilibrium might reveal our 'moral grammar' in a way that is analogous to uncovering the grammar that underlies our syntactic ability as native speakers of a language to make judgments about grammatical form. In support of the analogy, some contemporary theorists who systematically examine our moral intuitions, often through hypothetical as opposed to real cases, believe they are uncovering a moral structure of principles that is *a priori*.

Uncovering a syntax, however, is a descriptive and not a justificatory task. Once we can identify the grammar or rules that best account for a person's syntactic competency, we do not ask the question, 'Should that person have this grammar?' We are satisfied to have captured the grammar underlying a person's idiolect. In ethics and political philosophy, in contrast, we must answer that justificatory question, especially since there is often disagreement among people about what is right, disagreement that is not resolved simply by pursuing narrow reflective equilibrium.

3.2 Wide Reflective Equilibrium

Rawls's proposal is that we can achieve real justification in the choice of principles of justice by broadening the circle of beliefs that must cohere. In a *wide* reflective equilibrium, for example, we broaden the field of relevant moral and nonmoral beliefs to include an account of the conditions under which it would be fair for reasonable people to choose among competing principles, as well as evidence that the resulting principles constitute a feasible or stable conception of justice, that is that people could sustain their commitment to such principles.

For example, the constraints on choice in Rawls's contract situation, such as the veil of ignorance that keeps us from knowing facts about ourselves and our specific preferences, require justification. Rawls needs to show that these constraints are 'fair' to all contractors (thus 'justice as fairness' is his label for this procedural account of how we find out what is just). To provide such justification, Rawls appeals to beliefs about the fundamental 'moral powers' of 'free and equal' agents (they can form and revise their conceptions of what is good and they have a sense of justice). He also appeals to the ideal of a well-ordered society in which principles of justice play a particular role as public principles for reconciling disputes. Rawls must even provide us with an account of 'primary social goods,' necessary if agents who do not know their own actual preferences are to decide what principles would be better for them to choose. The device of the contract is thus in reflective equilibrium with certain background theories that themselves contain moral beliefs. The contract is not simply based on uncontroversial assumptions about human rationality (although it describes a rational choice problem within the constraints it imposes). Without acceptance of this wider circle of moral beliefs, Rawls's construction lacks support.

Our beliefs about justice are justified (and, by extension, we are justified in holding them) if they cohere in such a wide reflective equilibrium. Obviously, the method of wide reflective equilibrium is here only illustrated by appeal to the detail of Rawls's use of it. If we abstract from that detail, we see that there is a complex structure and interaction of beliefs, at many levels of generality, that bear on the construction of an account of justice. We shall later return to this point when we talk about implications of this method for work in ethics.

In *A Theory of Justice*, Rawls seemed to think that all people might converge on a common or *shared* wide reflective equilibrium that included 'justice as fairness,' the conception of justice for which he argues. Wide reflective equilibrium thus played a role in the construction of the theory, helping 'us' to articulate its key features in ways that led to principles that matched 'our' considered judgments. At the same time, wide reflective equilibrium constituted an account of justification. Shared agreement on that wide equilibrium would produce a well-ordered society governed by principles guaranteeing equal basic liberties, fair equality of opportunity, and the requirement that inequalities be arranged to make the worst-off groups as well off as possible.

In his later work, *Poltical Liberalism* (1993), Rawls abandons the suggestion that all people might converge in the same, shared wide reflective equilibrium that contains his conception of justice. This change comes about because Rawls believes that justification must accommodate the unavoidable 'fact' that human reason, exercised under conditions of liberty, will yield a pluralism of reasonable comprehensive moral views. Because of 'reasonable pluralism,' Rawls recasts jus-

tice as fairness as a 'free standing' political conception of justice on which people with different comprehensive views may agree in an 'overlapping consensus.' The *public* justification of such a political conception involves no appeal to the philosophical or religious views that appear in the comprehensive doctrines that form this overlapping consensus. Instead, we might think of this process of working back and forth among the key shared ideas in the public, democratic culture and the articulated features of the political conception of justice as a *political* reflective equilibrium (Daniels 1996).

Although there is no convergence on a shared wide reflective equilibrium that contains the political conception of justice, wide reflective equilibrium still plays a critical role in justification. For individuals to be *fully justified* in adopting the political conception of justice, the conception articulated in the political reflective equilibrium, they must incorporate it within a wide reflective equilibrium that includes their own comprehensive moral or religious doctrine. It will count as a reasonable view for them only if it coheres with their other (religious and philosophical) beliefs in reflective equilibrium. Over time, people modify their comprehensive views so that they may accommodate the overlapping consensus, a process that involves both philosophical reflection and the moderating influence of living under institutions that are governed by the shared conception of justice.

4. *Criticisms of Reflective Equilibrium*

4.1 *Challenges to Moral Intuitions*

Central to the method of reflective equilibrium in ethics and political philosophy is the claim that our considered moral judgments about particular cases carry weight, if only initial weight, in seeking justification. This claim is controversial. Some of the most vigorous criticism of it has come from utilitarians, and it is instructive to see why.

A traditional criticism of utilitarianism is that it leads us to moral judgments about what is right which conflict with our 'ordinary' moral judgments. In response, some utilitarians accept the relevance of some of these judgments and argue that utilitarianism is compatible with them. Thus Mill argued for a utilitarian foundation for our beliefs about the importance of individual liberty. Some utilitarians have even argued that key features of our 'common sense morality' approximate utilitarian requirements, and that we have acquired these beliefs just because they do, unconsciously, reflect what promotes utility—they reflect the wisdom of a heritage.

An alternative utilitarian response to the claim that utilitarianism conflicts with certain ordinary moral judgments is to dismiss these judgments as pretheoretical 'intuitions.' As such, they are likely to result from cultural indoctrination and thus reflect superstition, bias, and mere historical accident. On this view, moral intuitions or judgments should have no evidentiary credentials and should play no role in moral theory construction or justification. Indeed, as the prominent utilitarians Richard Brandt and Richard Hare argued against Rawls, simply making 'coherent' a set of beliefs that have no 'initial credibility' cannot produce justification, since coherent fictions are still only fictions. Indeed, when Rawls describes the conditions under which we might solicit considered moral judgments, namely that people be calm and have adequate information about the cases, they do not by themselves do anything to assuage the utilitarian worries.

This criticism has some force, since two standard ways of supplying credentials for initial judgments are not available. One traditional way to support the reliability of these judgments or intuitions is to claim, as eighteenth-century theorists did, that they are the result of a special moral faculty that allows us to grasp particular moral facts or universal principles. Modern proponents of reflective equilibrium reject such mysterious faculties. Indeed, they claim moral judgments are revisable, not foundational.

A second way to support the initial credibility of considered judgments is to draw an analogy between them and observations in science or everyday life. For example, what counts as observational evidence in science depends on theory, and theory may give us reasons to reject some observations as not constituting counterevidence to a scientific law or theory. In this way we might see an analogy between the revisability of moral judgments and observations.

Developing this analogy, however, seems to require that we also tell some story about why moral judgments are reliable 'observations' about what is right. Perhaps we might need something like the causal story that some theorists of knowledge offer to explain the reliability of observations. Since no such story is forthcoming, opponents argue, proponents of reflective equilibrium must reject the requirement or give up the analogy. Proponents of reflective equilibrium might reject the requirement by suggesting it is premature to ask for such a story in ethics or by claiming that we can provide no analogous causal story for credible judgments we make in other areas, including mathematics or logic.

It might seem that the burden of argument has shifted to advocates of reflective equilibrium to show why 'initial credibility' should be ascribed to moral judgments or intuitions. Defenders of reflective equilibrium may nevertheless reject this burden, arguing that critics, especially utilitarian critics, actually face the same problem. For example, Richard Brandt argues that 'facts and logic' alone, and not moral intuitions, should play a role in moral theory construction and justification. On Brandt's view, we should choose moral principles when they are based on desires that have been subjected to maximal criticism by facts and logic alone (he calls it 'cognitive

psychotherapy'). We should avoid any appeal to moral intuitions that might infect this critique.

The desires that Brandt appeals to, however, are themselves shaped and influenced by the very same social structures that utilitarians complain have biased and corrupted our moral judgments. Nothing in the process of critique by facts and logic alone can eliminate this source of bias. If this claim is right, it undercuts the suggestion that we can step outside our beliefs to arrive at some more objective form of justification. Instead, we may be better off recognizing that our process of critique—in the method of wide reflective equilibrium—is explicit about exposing the sources of bias and historical accident to criticism and revision, rather than fooling ourselves into thinking that desires are some sort of morally uninfluenced layer of 'facts.'

4.2 Other Criticisms of Reflective Equilibrium

Other criticisms of reflective equilibrium in ethics have focused on points familiar from discussions in epistemology more generally. One important complaint concerns the vagueness of the concept of coherence. If we simply take logical consistency as the criterion for coherence, we have much too weak a constraint. What account can we give of a stronger notion?

More has to be said about how some parts of the system of beliefs 'support or explain' others than is provided in the account above. Critics of Rawls's requirement that the contract situation be adjusted, if necessary, to yield principles that are in reflective equilibrium with our judgments about justice complained that this was a 'rigged' contract and that it did no justificatory work beyond reflective equilibrium. If, however, the moral judgments that play a role in discussions of fair process, the well-ordered society, and the moral powers of agents are somewhat independent of the considered judgments about justice, then we get the kind of independent support for the principles that add justificatory force. Little work has actually been done, however, to flesh out a stronger account of coherence in ethics as opposed to epistemology more generally, where coherentism has been defended and elaborated.

A second line of objection derives from anticoherentist accounts of justification rooted in the theory of knowledge. Some insist, for example, that coherence accounts of justification cannot be divorced from coherence accounts of truth. On the other hand, Rawls's view fits with the claim by some contemporary theorists of knowledge that a coherence account of justification is distinguishable from a coherence account of truth, and defensible when so separated.

This separability is important to Rawls, even more so in his later work than in *A Theory of Justice*. Rawls never took a stand on the question of whether there are 'truths' about justice that correspond to a conception of justice that we justify through reflective equilibrium. In his early work, the method of reflective equilibrium suggested ways in which convergence might be achieved among those who began with disagreements about justice. By drawing attention to the many features of a comprehensive theory on which argument and evidence could be brought to bear, the method of wide reflective equilibrium held out promise that more convergence might result than if people had only considered judgments and principles to agree and disagree about. Convergence did not imply that truth was reached, but under some conditions, it could be taken as evidence for truth, just as it might be in nonnormative areas of inquiry.

Once Rawls took the 'burdens of judgment' and reasonable pluralism seriously, that is once he 'politicized' justice in his later work, it became much less plausible to talk about convergence on a shared wide equilibrium that might contain moral truth. Instead, even if we converged in an 'overlapping consensus' on a conception of justice, it would have to be justified from the perspective of distinct wide reflective equilibria. It became more important for Rawls to suggest how claims about justice could be 'objective' without presupposing moral truth. These developments in Rawls's account thus appear to move him farther away from those who would seek to give a realist account of moral truth.

One further group of criticisms of reflective equilibrium derives largely from claims about human rationality. A starting point for this line of argument is the complaint that people who have different starting points as their initial set of beliefs, say with different degrees of credence given to their beliefs, may arrive at different reflective points of equilibrium. This possibility undercuts the suggestion that critical pressures alone can produce convergence. A more general form of this worry is that the model of reflection involved in reflective equilibrium, in which we are to produce coherence among all our beliefs, overemphasizes and idealizes human rationality. The suggestion is that we would be better off presupposing a more minimal form of rationality. A specific version of this complaint is that the method involves an information burden that cannot be met. Yet another version of this criticism is that we should allow for much less 'rationalist' forms of modification of our views, recognizing that 'conversion' experiences such as those involved in 'paradigm shifts' may affect any account of coherence.

A full defense of reflective equilibrium as a method would require a more developed response to many of these lines of criticism than exists in the literature.

5. Applications and Implications of Reflective Equilibrium

Despite these criticisms, some philosophers have argued for a broader understanding of the relevance of the method of reflective equilibrium to practical ethics.

In thinking about the course of right action in a particular case, we often appeal to reasons and principles that are notoriously general and lack the kind of specificity that make them suitable to govern the case at hand without committing us to implications we cannot accept in other contexts. This requires that we refine or specify the reasons and principles if we are to provide appropriate justifications for what we do and appropriate guidance for related cases. Philosophers who have focused attention on the importance of specification have drawn on the method of reflective equilibrium for their insights into the problem.

In practical ethics, especially bioethics, there has been a vigorous debate about methodology. Some argue that we must root all claims about specific cases in specific ethical theories, and that the hard work is showing how these theories apply to specific cases. Others argue that we may disagree about many aspects of general theory but still agree on principles, and that the hard work of practical ethics consists in fitting sometimes conflicting principles to particular cases. Still others argue that we must begin our philosophical work with detailed understanding of the texture and specificity of a case, avoiding the temptation to intrude general principles or theories into the analysis.

A grasp of the method of wide reflective equilibrium suggests a way around the exclusionary nature of this debate. Wide reflective equilibrium shows us the complex structure of justification in ethics and political philosophy, revealing many connections among our component beliefs. At the same time, there are many different types of ethical analysis and normative inquiry.

This suggests a more eclectic view of the debate about method. Work in ethics requires all levels of inquiry proposed by the disputants, not always all at once or in each case, but at some time or other. Sometimes it is true that we cannot resolve disputes about how to weigh conflicts among principles unless we bring more theoretical considerations to bear (these considerations need not involve comprehensive ethical theories). Sometimes we can agree on relevant principles and agree to disagree on other theoretical issues, still arriving at agreement about the rightness of particular policies or actions and their justification in light of relevant reasons and principles. Sometimes we must see what is distinctive about particular cases and revise or refine our reasons and principles before we can arrive at understanding of what to do. Finally, and of the greatest significance, the method of wide reflective equilibrium should make it clear that work in ethical theory cannot be divorced from work in practical ethics. We must test and revise theory in light of our considered judgments about moral practice. The condescending attitude of many who work in ethical theory toward work in practical ethics thus is incompatible with what wide reflective equilibrium establishes about the relationship between these areas of ethical inquiry.

The method of wide reflective equilibrium makes it plausible that all such 'methods' should be seen as appropriate to some tasks in ethics and are but parts of a more encompassing method.

Bibliography

Daniels N 1979 Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy* **76**: 256–82

Daniels N 1996 *Justice and Justification: Reflective Equilibrium in Theory and Practice*. Cambridge University Press, New York

DePaul M R 1993 *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry*. Routledge, New York

Goodman N 1955 *Fact, Fiction, and Forecast*. Harvard University Press, Cambridge, MA

Hare R M 1973 Rawls' theory of justice. *Philosophical Quarterly* **23**: 144–55, 241–51

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Rawls J 1974 The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association* **47**: 5–22

Rawls J 1993 *Political Liberalism*. Columbia University Press, New York

Sencerz S 1986 Moral intuitions and justification in ethics. *Philosophical Studies* **50**: 77–95

Stich S P 1990 *The Fragmentation of Reason*. MIT Press, Cambridge, MA

Van der Burg W, Van Willigenburg T (eds.) 1998 *Reflective Equilibrium: Essays in Honour of Robert Heeger*. Kluwer Academic Publishers, Dordrecht, The Netherlands

N. Daniels

Reflexivity in Anthropology

Critiques of the social sciences as practiced in the United States emerged in the 1980s in which reflexivity became a keyword, standing for both the means of the critiques as well as the form that alternative practices might take. The consequences for disciplinary practices of reflexivity as means of critique and alternative form of inquiry were nowhere as great among the social sciences as in social and cultural anthropology.

According to one of the early statements in this critical trend, reflexivity is '… the capacity of any system of signification to turn back upon itself, to make itself its own object by referring to itself: subject and object fuse' (Myerhoff and Ruby 1982). Myerhoff and Ruby go on to stress that 'significant distinctions exist between reflexiveness and related attitudes such as self-regard, self-absorption, solipsism, self-reference, self-consciousness, and autobiography. Reflexiveness does not leave the subject lost in its own concerns; it pulls one toward the other and away from isolated attentiveness toward oneself.'

Reflexivity as a practice in the social sciences is, thus, both a distinctive mode of thinking and a strategy

of writing that concerns itself with the process of inquiry as an integral part of the claims to knowledge that the inquiry might produce. As a concept, it is related to the idea of feedback in cybernetic models, of metacommunication, and of such functions as deixis and indexicality inherent in the faculty of language use. However, it is the more ideological uses of reflexivity in debates about the revision of anthropology's distinctive methods of fieldwork and ethnography that are of concern here. Debates about the role and purpose of reflexivity in the process of research as well as in the changing form of the ethnographic text became the major specifiable way of assessing the impact upon anthropology of the broader intellectual movement of the 1970s and 1980s known loosely as postmodernism.

In the broad sweep of European philosophy, reflexivity as a category of interest in social theory begins with Kant's distinction between analysis and an analytic in which an analytic is the kind of inquiry where the subject of inquiry is also the object of inquiry. One genealogical stream from this originary idea of reflexive inquiry leads through Hegel to phenomenology and eventually to Heidegger and existentialism in France. The other, and for us, more relevant genealogical stream leads to hermeneutics in the work of Hans-Georg Gadamer and Paul Ricoeur by way of the neo-Kantian writings of Dilthey and the work of Max Weber, in which reflexivity, while implicit, is a strong element of his methodological considerations. In the work of Gadamer and Ricoeur, reflexivity becomes an explicit and crucial intellectual operation in the activity of interpretation whereby the individual ethical consciousness expands reflectively through encounters with the foreign or the different. Thus, while long a potential existing in a sort of minor key to the predominant positivist tendency of social theory which oversaw the post-World War II expansion of sociology and anthropology in the United States, philosophical works of hermeneutics, in which the operation of reflexivity was highlighted, only developed explicitly a strong appeal among sociologists and anthropologists from the early 1980s onward.

Of all of the social sciences, the importance of reflexivity as a style of inquiry has been most natural to precisely those methods of anthropology which have distinguished it as a discipline. Pioneered by Bronislaw Malinowski and Franz Boas, the pursuit of fieldwork investigation through long-term participant observation, working directly in native languages, and the constant interpretation and translation of behaviors and speech encountered on a daily basis, inherently involves on the part of the investigator a keen consciousness of oneself and one's relations with specific others as the medium of all claims to knowledge from this research practice.

For a long time, anthropologists only came to terms with the complexities and problems of this distinctive form of knowledge production informally as a pastime of professional culture and folklore, as something to be communicated between mentors and students. It was only at the end of the 1960s, and then again, and more definitively in the 1980s, that anthropologists examined and developed explicitly the varieties of meanings and implications of reflexivity as a major style of knowledge production for the future practice of their discipline.

Since the 1980s, reflexivity has become a 'keyword' indicating both a self-conscious, mainstream style of work in anthropology and part of its own enduring disciplinary politics of knowledge. Here, it is important to note that reflexivity has only served as such a central category, standing in turn for the most concrete aspect of the controversial impact of postmodernism on anthropological practice, in the recent evolution of the discipline in the United States. It has been discussed and debated from the 1980s on in other anthropological establishments around the world, but mainly through the stimulus of the prestige and diffusion of trends in United States anthropology.

1. *Reflexivity in Two Moments in the Critique of Anthropology*

During the late 1960s, US anthropologists became preoccupied openly with a long simmering critique of their discipline.This critique had theoretical and methodological dimensions, but it was primarily political in nature, stimulated by the ferment of domestic resistance to the Vietnam War, renewed discussions of anthropology's historic development within the expansion of European colonialisms, and the interest in new work in Marxist theory. At that time as well, Geertz was developing a broad and deep theoretical foundation for cultural anthropology in hermeneutic philosophy, work that was eventually to serve as both the primary foil and influence for later discussions of reflexivity in anthropological method. (See Geertz 1968 for his early position on reflexivity in fieldwork.) But in the *de facto* reflexive attempt to undertake a critical anthropology of anthropology in the politicized atmosphere of the 1960s, the issue of reflexivity itself as the major dynamic of what it was to practice critical analysis was little addressed. An important exception is the paper by Scholte (1969) which summarized this period of critique. Important as a figure who introduced currents of European social theory into US anthropology, Scholte foreshadowed what sort of theoretical work and changes in practice would be necessary to transform anthropology into a discipline with a sustained critical purpose beyond that moment of heightened political critique.

In the interval between the two periods of internal disciplinary critique, perhaps the most important work that served as the raw material on which the eventual surfacing of explicit considerations of reflexivity in

anthropology depended was the burgeoning of a literature on fieldwork experience, written alongside the production of more formal ethnographies from which such reflections about the process of inquiry were largely absent. This literature, fueled by the publication of Malinowski's diaries in 1967, incorporated the spirit of self-critique of the late 1960s and was produced by anthropologists who were then students and young professors. There emerged a more theoretical strain of this literature in which the reflexive character of fieldwork became the major preoccupation (e.g., Rabinow 1979) and that led eventually to experiments with the form of the ethnography itself (e.g., Crapanzano 1980). These works, combined with an interdisciplinary movement, variously characterized as postmodernism and poststructuralism and led by scholars of literature who were experiencing their own disciplinary crisis, stimulated the theoretically deeper, more wideranging, and sustained internal critique of anthropology of the 1980s.

Incorporating all of the dimensions of the earlier critique, the 1980s critique primarily examined the central practices of the discipline, especially in its character as a hermeneutic enterprise, a view for which Geertz had laid the foundation. In particular, it did so by a focus on the rhetoric and forms of the signature textual product of anthropology, the ethnography. (See Clifford and Marcus 1986, the volume that has stood for this critique in the way that the Hymes volume did for the earlier one.) This critique was, thus, the efflorescence of a comprehensive discussion of all of the issues relating to reflexivity that had long been so essential to the practice of anthropology. Influenced by the circulation of much more diverse and sophisticated theoretical resources, anthropologists looked with an unprecedented intensity and thoroughness at the relationships in their work between reflexive modes of knowing, critical and ethical awareness, the political contexts of inquiry, and alternative possibilities for practice.

The basic act of reflexivity which initiated this critique were the probing discussions of the relationships of anthropologists to informants—of the self to other—and the dependence of the production of anthropological knowledge on these relationships which had so often been effaced in the ways that ethnographies had in fact been written. Evoking alternative modes of representing others, questioning the act of representation itself, suggesting the ways ethnographic knowledge emerges from collaboration and dialogue, and probing the political conditions which have inhibited such possibilities, this moment of critique served both to stimulate degrees of experiment with the ethnographic form and to encourage ramifying discussions about the critical purpose of anthropological analysis (e.g., Marcus and Fischer 1999). This undertaking was itself an umbrella act of reflexivity, and reflexive practice became the common frame for thinking about past ethnographies, future possibilities, and the form and purpose of anthropological knowledge itself. This field of discussion was highly contested: some felt that the critique by its nature was too hermetic, too literary, and not sufficiently committed to anthropology's study of social struggles in the contemporary world; some thought the attention to reflexive issues was an aesthetic distraction and a major blow to the production of objective knowledge that could be shared by a scholarly community and on which disciplinary progress depended; and contrast-ingly, for some, the kinds and levels of reflexivity encouraged by the critique were neither reflexive nor critical enough, leading to a spiralling array of discussions about insufficiencies and exclusions that had even more ideal reflexive practice in view.

The immediate effect of this period of critique were the changes in the ways that ethnographies were both written and discussed. But while the critique was developed through attention to texts and the way that anthropologists rhetorically established authority for them, the larger stake was the effort finally to justify them theoretically and ethically and expand anthropology as a critically reflexive field of inquiry. There can be little doubt that this much has been established as at least a long-term mainstream tendency in the discipline. This tendency has placed anthropology in much closer relation to such parallel interdisciplinary fields, themselves rooted in varieties of reflexive inquiry, as media studies, feminist studies, postcolonial studies, and most broadly, cultural studies, than to its historic partners in the other social sciences.

2. *The Incorporation of Reflexivity in the Mainstream of Anthropological Research*

Discussions about reflexivity, or at least the evaluation of ethnographic claims to knowledge in terms of the specific conditions and relations of inquiry that produced them from fieldwork, are not only highly visible and commonplace in anthropological discourse of the 1990s, they are *de rigueur*. Further, some form and degree of reflexive commentary is worked prominently into virtually every ethnography written these days. If not enacted in experimental text design, there is the presumption that all ethnography is produced collaboratively, and there are far richer and more probing discussions than ever before about the kinds of contexts which influence ethnographic representations of subjects and shape arguments from ethnographic data (see Marcus 1994). Concern for reflexivity has had a special appeal for ethnographic filmmakers, and some of the most cogent writing on the importance of reflexivity in the representation of other forms of life have come from this arena (e.g., MacDougall 1998). Current productions of ethnographic film and media demonstrate most clearly that the practice of reflex-

ivity crucially redefines and adapts the classic conditions of ethnographic inquiry to new conditions of fieldwork where anthropological research projects must negotiate different identities and roles in relation to the highly articulated projects of their subjects in the overlapping contexts of a globally interconnected world.

2.1 Variant Ideologies in the Practice and Form of Reflexivity

In sociology, discussion of reflexivity has tended to be integrated formally as a methodological tool or technique (e.g., see Bourdieu 1990), limiting its ability to affect fundamentally the forms and purposes of sociological research. In anthropology, the outcome has been quite different, generating a number of changes in the ways that ethnographies are written and fieldwork is conceived. This in turn has obviously had considerable implications for changes in both the form and content of the knowledge that anthropology's core research process generates.

In engagement with the terms of the 'Writing Culture' critique of the 1980s, the largest body of literature stimulated explicitly by reflexive inquiry examines in great detail the self-other relation in classic conditions and specific sites of fieldwork. While some have characterized this literature as narcissistic or merely self quests, in its most accomplished examples (e.g., Behar and Gordon 1995, Reed-Danahay 1997, Taylor 1999, Tsing 1993), it has addressed cogently issues of contemporary changes in concepts of self, personhood, and identity among many of the peoples who had been the subject of study by previous generations of ethnographers. At the microlevel of fieldwork, it has also clarified aspects of the history of anthropology as an academic enterprise and in relation to the broad range of effects that inquiry itself has on the peoples and subjects studied (e.g., Rosaldo 1989). While somewhat unruly and less governed by genre rules of form in social science writing, this literature has, in sum, been very useful in deepening and critiquing the complexities of constructs of subjectivity and otherness within their conventional usages in Western intellectual discourse—a long-term, continuing objective of anthropology as a discipline. It has also probed critically ethnographers' positions in relation to initial objects of study so as to transform the objects in novel ways and reconfigure the compass of research in response to radical contemporary changes in both the worlds of the academic and who is identified as subjects of study.

There has also been a strain of reflexive inquiry in anthropology for which reflexivity has been more distinctly a means by which to define an object of study or to circumscribe new temporal and spatial boundaries for a fieldwork project (Marcus 1998). Key to these works is clarity of the positioning or location of ethnographic inquiry, which reflexivity enables, in relation to other historic and geographical agents in the same social field. Reflexivity as positioning not only serves the creation of new forms of objectivity in anthropology, sensitive to the issues of critique that discussions of reflexivity have raised, but redefines both the scope of ethnography and its arguments. In these works the limits and reach of once very locally focussed ethnographic arguments expand in unpredictable and intriguing ways (for a now classic experiment of this sort, see Taussig 1987; for more recent examples, see Martin 1994, Fortun 2000).

Finally, there is the presence of reflexivity in ethnography because of the evolution of certain realities on the ground over the course of career-long scholarly involvements with certain peoples and places. In recent decades, literally self-conscious cultural awareness has increased markedly and even been mobilized politically in many locales and regions such that the conditions of ethnographic research require a rethinking and reorientation toward one's subjects, now organized into social movements, nongovernmental organizations, and other groups involved in objectifying and representing themselves. Here, reflexive inquiry is the means to a complex and necessary politics of research which not only shapes fieldwork but makes it continuation possible at all (see Warren 1999 for an exemplar of this type of increasing variant of reflexive inquiry).

2.2 Reflexive Modernization and Reflexivity in the Changing Conditions of Anthropological Fieldwork

Broader changes in contemporary social conditions worldwide that stimulate the trend of reflexive inquiry in anthropology is encompassed in the concept of reflexive modernization (Beck et al. 1994), a development in social theory of the 1990s that responds to the postulation of a present era of postmodernity by trying to specify the characteristics of such an era for the pursuit of empirical inquiry. This construct proposes that the modernity characteristic of industrial society is succeeded by a modernity that involves living with irreducible contingency, living in a more complex and less controllable world. Most importantly, reflexivity as a self-monitoring process on the level of both institutions and individual persons becomes not only a pervasive aspect of society, but also a key focus of social thought itself.

This version of social theory suggests for the proliferation of styles of reflexive inquiry in anthropology that these are not driven only as a result of the internal critique of anthropology, as a seminal countertrend to a preceding era of positivist tendencies, but as a more generally shared condition with the societies and cultures which anthropology makes its subjects of study. This suggestion in turn has a potentially profound and enduring effect on the research proto-

cols of anthropology in recognizing certain affinities and identifications between the reflexive predicaments of ethnographers and those whom they study. Reflexivity is relevant not because it provides truer or more powerful explanatory models with which anthropologists can treat social facts among their subjects, but rather because reflexivity is also practiced differentially 'out there' on a widespread basis (see Marcus 1999). This fact suggests the breaking of the frame of conventional scholarly work altogether, at least in the anthropological tradition, rather than the demise and replacement of specific concepts and characteristics of this academic frame. Thus, reflexivity is no longer only a moment in the critique of disciplines but becomes more generally the object and medium of social and cultural analysis as well.

See also: Cultural Critique: Anthropological; Ethnography; Feminist Theory: Postmodern; Hermeneutics, History of; Hermeneutics, Including Critical Theory; Interpretation in Anthropology; Postmodernism: Methodology

Bibliography

Beck U, Giddens A, Lash S 1994 *Reflexive Modernization: Politics, Tradition and Aesthetics in the Modern Social Order*. Stanford University Press, Stanford, CA

Behar R, Gordon D A (eds.) 1995 *Women Writing Culture*. University of California Press, Berkeley, CA

Bourdieu P 1990 *The Logic of Practice*. Stanford University Press, Stanford, CA

Clifford J, Marcus G E 1986 *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA

Crapanzano V 1980 *Tuhami: Portrait of a Moroccan*. University of Chicago Press, Chicago

Fortun K 2000 *Advocacy After Bhopal: Environmentalism, Disaster, New Global Orders*. University of Chicago Press, Chicago

Geertz C 1968 Thinking as a moral act: Ethical dimensions of anthropological fieldwork in the new states. *Antioch Review* **28**: 139–58

Hymes D (ed.) 1972 *Reinventing Anthropology* 1st edn Pantheon Books, New York

MacDougall D 1998 *Transcultural Cinema*. Princeton University Press, Princeton, NJ

Marcus G E 1994 On ideologies of reflexivity in contemporary efforts to remake the human sciences. *Poetics Today* **15**: 383–404

Marcus G E 1998 *Ethnography Through Thick and Thin*. Princeton University Press, Princeton, NJ

Marcus G E (ed.) 1999 *Critical Anthropology Now: Unexpected Contexts, Shifting Constituencies, Changing Agendas*. School of American Research Press, Santa Fé, NM

Marcus G E, Fischer M J 1999 [1986] *Anthropology as Cultural Critique: An Experimental moment in the Human Sciences*. University of Chicago Press, Chicago

Martin E 1994 *Flexible Bodies: Tracking Immunity in American Culture from the Days of Polio to the Age of AIDS*. Beacon Press, Boston

Myerhoff B, Ruby J 1982 Introduction. In: Ruby J (ed.) *A Crack in the Mirror: Reflexive Perspectives in Anthropology*. University of Pennsylvania Press, Philadelphia

Rabinow P 1977 *Reflections on Fieldwork in Morocco*. University of California Press, Berkeley, CA

Reed-Danahay D E (ed.) 1997 *Auto/Ethnography: Rewriting of the Self and the Social*. Berg, Oxford, UK

Rosaldo R 1989 Imperialist nostalgia. In: *Culture & Truth: The Remaking of Social Analysis*. Beacon Press, Boston, pp. 68–87

Ruby J (ed.) 1982 *A Crack in the Mirror: Reflexive Perspectives in Anthropology*. University of Pennsylvania Press, Philadelphia

Scholte B 1969 Toward a reflexive and critical anthropology. In: Hymes D (ed.) *Reinventing Anthropology*. Random House, New York, pp. 430–57

Taussig M 1987 *Shamanism,Colonialism and the Wild Man: A Study in Terror and Healing*. University of Chicago Press, Chicago

Taylor J 1999 *Paper Tangoes*. Duke University Press, Durham, NC

Tsing A L 1993 *In the Realm of the Diamond Queen*. Princeton University Press, Princeton, NJ

Warren K B 1999 *Indigenous Movements and Their Critics: Pan-Maya Activism in Guatemala*. Princeton University Press, Princeton, NJ

G. E. Marcus

Reflexivity, in Science and Technology Studies

Considerations of reflexivity have salience in diverse intellectual practices and traditions: literature, philosophical logic, the visual arts, cybernetics, as well as social theory and sociology. Indeed, it is an issue with potential relevance for all modes of representational and performative practice. In the most general terms, considered as a feature or style of inquiry, reflexivity is a concept that expresses a concern to turn 'subject' into 'object,' to turn the activity of inquiry into a topic for (its own) inquiry. To 'be reflexive' is to bring the process of doing an activity into the purview of that activity as a feature of it. Thus, it implies a certain self-consciousness, and a certain paradoxical, even Zen-like, mode of operation: bootstrapping; attempting to see the flashlight that illuminates the scene as a part of the scene it illuminates; taking account of this moment as one writes of 'this moment.' This article details the ways in which reflexivity has featured in science and technology studies (STS), and speculates on the current and future status of reflexive study in STS.

1. Alternative Introduction

This heading is not to declare, exactly, that there is anything 'wrong' with that first introductory paragraph. It does the job, and it does it correctly; that is,

in accord with the 'Instructions for Manuscript Preparation' for this volume. It is just that in doing 'just that,' it is unable to acknowledge the conditions under which it does so. Thus, it leaves all those things out of (its) account; which is normal, and normally unobjectionable. Perhaps no other article in this *Encyclopedia* makes any reference to such things, preferring instead, perfectly 'obviously,' to concentrate on the primary task of speaking authoritatively (and, importantly, briefly) about its topic. Thus, the conditions of production of these particular texts are constructed as irrelevancies for, and obstacles to, the proper work. Social science inquiry that focuses on conditions of production has always acted (intentionally or not) to 'make trouble' of this kind: Marxist economics for capitalism; sociology of education for teaching; STS for science. Reflexivity adds a twist to this dynamic: in collapsing the external/internal, outsider/insider, on/in relation, the source of trouble, being as much 'here' as it is 'there,' is no longer clearly located or easily attacked. This, then, is the primary role for reflexive inquiry: to make things difficult; to enrich the process of inquiry; to go round and round in circles; to fill out the picture; to frustrate all action; to act authentically and without self-contradiction; to change the subject.

2. *Reflexivity as a Problem*

In STS, the impetus for explicitly reflexive studies has been felt most strongly in the important sub-field of STS known as the sociology of scientific knowledge (SSK) (see *Sociology, Epistemology of*; *Technology, Social Construction of*). Here, the reflexive tie between its subject (knowledge about scientific knowledge) and its object (scientific knowledge) is very marked. That is, SSK exhibits a strong 'X of X' structure. Because of this, and because SSK's epistemology is explicitly relativist, reflexivity in SSK has a 'bite' akin to classic philosophical skepticism.

Extremely crudely, and for illustrative purposes only, SSK's relativism with respect to scientific knowledge can be, and indeed has been, taken as equivalent to a statement of this kind: 'Scientific knowledge does not constitute truth because it is socially constructed.' Read this way, SSK then becomes vulnerable to the standard anti-relativist critique, the argument from self-refutation: 'If this statement is asserted as a (scientific) truth, then it suffers from performative self-contradiction, and thus refutes itself' (see, for example, Hollis and Lukes 1982). Note that the self-refutation argument uses reflexivity: it is through the process of turning its target onto itself that the demonstration of self-refutation is effected.

This is indeed one of the commonest modes in which reflexivity operates: as a weapon in the agonistic game of criticism. Thus, for those threatened with attack-by-reflexivity, it becomes a source of vulnerability to be abjured and defended against. Reflexivity becomes, that is to say, a 'problem' to be 'managed.' The various ways that practitioners of SSK have done this have been one of the preoccupations of 'reflexivist' studies (see Sect. 2.3).

2.1 *Logical Consistency: Bloor's Strong Programme*

David Bloor's *Knowledge and Social Imagery* (1991), is one of the earliest texts to set out the elements of what has become known as the SSK. It contains an endorsement of reflexivity as the fourth and last of the 'tenets' for the 'strong programme in the sociology of knowledge' (see *Strong Program, in Sociology of Scientific Knowledge*):

> In principle [the strong programme's] patterns of explanation would have to be applicable to sociology itself. Like the requirement of symmetry, this is a response to the need to seek for general explanations. It is an obvious requirement of principle because otherwise sociology would be a standing refutation of its own theories (1991, p. 7).

Here, Bloor's positive support for reflexivity is a matter of logical consistency and a creative way of managing the self-refutation charge. Rather than deny the reflexive applicability of the strong programme, he emphasizes it, secure in his understanding that the socially constructed character of all knowledge, including his own, does not count against its validity. Bloor's reflexivity, however, remains at this 'in principle' level. His negative attitude to any empirical instantiation of his reflexive programmatics is documented in a text that claims to be precisely that (Ashmore 1989). Thus for Bloor (as well as for his former 'Edinburgh school' colleague Barry Barnes 1974, Epilogue), reflexivity is an argument used to defend a position; it is certainly not a practice.

2.2 *Harry Collins: Special Relativism and the Ban on Reflexivity*

Not so for, perhaps, the most influential writer through the heyday of SSK (ca. 1975–90), Harry Collins (e.g., 1985). Collins's doctrine of 'special relativism' separates the natural world from the social world, and the distinctive tasks and 'natural attitudes' of the natural scientist from those of the social scientist. The sociologist of scientific knowledge should treat the natural world as socially constructed, while treating the social world as 'real' and a source of sound data. However, this sense of 'realism' is taken from a reading of how natural scientists act towards their realm of inquiry, namely, the natural world. In effect, what Collins is doing here is to deny the practical relevance of the purported reflexive similarity between subject and object in SSK for purposes of the conduct

of research. The result is the short-circuiting of attempts to apply Collins's conclusions about, for example, the replication of experiments to those conclusions themselves (nevertheless see Mulkay 1985, Chaps. 4 and 5, Ashmore 1989, Chap. 4). In short, Collins's 'management strategy' for reflexivity is to ban it altogether. This radical solution to the problem of reflexivity has laid Collins open to charges of inconsistency from both wings of the 'realist–relativist' continuum. The recommended moves for repairing this perceived inconsistency differ, however, with the 'realist' suggesting he abandon his problematic relativism about nature, and the 'more-radical-relativist' proposing the dropping of his social realism (see *Technology, Social Construction of*).

2.3 Relativist Consistency, Radical Reflexivity: The Reflexivists

In the late 1980s, a small number of 'full-blown' reflexive texts were published (Ashmore 1989, Mulkay 1985, Woolgar 1988). This group of writers became known as 'reflexivists.' Ashmore's *The Reflexive Thesis* is an SSK study of SSK, a general exploration of reflexivity, and an experiment in 'new literary forms.' Mulkay's *The Word and the World* is a series of 'explorations in the form of sociological analysis' that draws upon analyses of biochemists' discourse. Woolgar's *Knowledge and Reflexivity* is a collection of papers from a small group of regular contributors (including Ashmore, Mulkay, and Woolgar) to the Discourse and Reflexivity Workshops held in the UK in the early 1980s. What links all these texts together, apart from their authors' close socio-cognitive relation, is the common attempt to treat reflexivity as an opportunity; even, for celebration. What happens when one stops evading reflexivity, when one looks 'the monster' in the eye, when one gives it a place at the heart of the research process? Self-destruction? Massive irrelevance? Third apocalyptic simile?

Bruno Latour's contribution to *Knowledge and Reflexivity* (Latour 1988) is its exception. Latour argues that the 'meta-reflexive' writing strategies of the reflexivists is a solution to a false problem: the naiveté of readers who believe too much and too easily. On the contrary, the real problem that (scientific, factual) writers have is being believed, or indeed, read at all. Latour further claims that the reflexivists are searching, despite their own gestures at weakening their authorial privilege, for the final meta-level that leaves no remainder and from which vantage point, everything may be taken into (their) account. As an antidote to all this, Latour recommends a strategy of 'infra-reflexivity': just to tell stories, drawing upon the richness and variety of all the 'ordinary' resources available. Latour's *Aramis* (1992, 1996)—despite its appearance as an experiment in 'new literary forms'—is, perhaps, an example.

3. Criticisms: Pitfalls and Pratfalls

Reflexivity gets you nowhere. It is not a useful thing to do and its results cannot be used. It is impractical. As an activity, it immobilizes you. It leads to idle navel-gazing and foolish self-absorption. It is 'all pots and pans and no pudding' (Bloor, cited in Ashmore 1989, p. 20). Logically, it leads to the abyss of the infinite regress or the never-ending search for the perfect meta-level. Reflexivity is politically and morally irresponsible games playing, fiddling while Rome burns. It has no soul (to bare) and no wisdom (to distribute), just an overdose of clever-clever irony. It disables criticism by taking that role on for itself; in its attempt to play all the roles of the academic community at once, and all by itself, it betrays itself as deeply, and ironically, anti-social. Indeed, it appears to seek nothing from its readers but a kind of 'aesthetic' admiration marked by a knowing, and smug, and boyish, amusement.

However, for the larger STS community, reflexivity has had positive functions and effects. The reflexivist fulfils the important role of the jester at the court of STS. Moreover, the presence of reflexive work as part of the 'canon' of STS acts as a corrective to a kind of methodological overconfidence resulting from the recent success of the field. Reflexivity reminds us that the fundamental epistemic questions and quandaries that originally animated SSK are permanent features of the practice of 'inquiry into inquiry' epitomized by STS scholarship. The text that aspires to 'transparency,' that is blithely unconcerned with the conditions of its own production, is now, at the very least, 'visible,' a member of a marked, rather than unmarked category.

4. The Current Situation

If the work of the reflexivists is taken as the model for 'adequately' reflexive work in STS, then it has to be said that this style has been taken up on very few, if any, occasions between 1990 and 2000, Latour's *Aramis* (1992, 1996) being a possible exception (see Sect. 2.3). Nevertheless, interest in reflexivity has not vanished. Explicit discussions still appear (e.g., Pels 2000, Lynch 2000). Indeed, it can be argued that some new currents of work in STS appear animated by reflexive concerns, though couched in a new register. For example, the emphasis on the performativity of social objects, on the multiplicity of ontologies, and on the essential situatedness of (their and our) knowledge production, found in the work of John Law, Annemarie Mol, and others attempting to move beyond actor network theory (Law and Hassard 1999) is implicitly deeply reflexive.

See also: Actor Network Theory; Sociology, Epistemology of; Strong Program, in Sociology of Scientific Knowledge; Technology, Social Construction of

Bibliography

Ashmore M 1989 *The Reflexive Thesis: Wrighting Sociology of Scientific Knowledge*. University of Chicago Press, Chicago
Barnes B 1974 *Scientific Knowledge and Sociological Theory*. Routledge and Kegan Paul, London
Bloor D 1991 *Knowledge and Social Imagery*, 2nd edn. University of Chicago Press, Chicago
Collins H M 1985 *Changing Order: Replication and Induction in Scientific Practice*. Sage, London
Hollis M, Lukes S (eds.) 1982 *Rationality and Relativism*. Blackwell, Oxford
Latour B 1988 The politics of explanation: An alternative. In: Woolgar S (ed.) *Knowledge and Reflexivity*. Sage, London
Latour B 1992 *Aramis ou l'amour des techniques*. La Decouverte, Paris
Latour B 1996 *Aramis or The Love of Technology*. Harvard University Press, Cambridge, MA
Law J, Hassard J (eds.) 1999 *Actor Network Theory and After*. Blackwell, Oxford, UK
Lynch M 2000 Against reflexivity as an academic virtue and source of privileged knowledge. *Theory, Culture and Society* **17**(3): 26–54
Mulkay M 1985 *The Word and the World: Explorations in the Form of Sociological Analysis*. George Allen and Unwin, London
Pels D 2000 Reflexivity: One step up. *Theory, Culture and Society* **17**(3): 1–25
Woolgar S (ed.) 1988 *Knowledge and Reflexivity: New Frontiers in the Sociology of Knowledge*. Sage, London

M. Ashmore

Reflexivity: Method and Evidence

Unity of Science, sometimes referred to as logical positivism, is the dominant paradigm in social science. It assumes a fundamental unity between the physical and social world, and the applicability of the same model of science and scientific inquiry. This unity derives from the belief that reality is 'out there' waiting to be described. The goal of science—physical and social—is the discovery of universal laws that allow us to predict and explain according to causes that are contained in the laws or some less demanding 'mechanism.' Reflexivism is a catchall term for a variety of approaches that deny that the main task of a social science is prediction or the discovery of universal laws. The world, and especially the social world, is not a container of things that have to be captured by our concepts, but is to a large degree constituted by our descriptions and theories. For this reason, most reflexivists see fundamental differences between the physical and social worlds and believe that our task in the latter is understanding (*verstehen*), and only secondarily, or derivatively, explanation (*erklären*).

According to Hans-Georg Gadamer, the principal difference between the natural and human sciences, is not one of method but of the aims of knowledge (Gadamer 1987, Schutz, 1973a). Despite this fundamental difference in purpose, reflexivism and the 'mainstream'—our shorthand for social scientists who embrace the unity of science position—share a significant number of methodological practices. These include a clear differentiation of premises from conclusions, respect for the canons of inference, recognition that some standards of validation must be established for data and source materials, differentiation of correlation from cause; the recognition of spuriousness problems when we rely on correlations, reliance on logic in establishing warrants for our assertions, and the belief that all findings can be contested. The deeper division about method is *within* the reflexivist community, between phenomenological, interpretevist, and hermeneutic approaches on the one hand and some postmodern or critical approaches on the other which do not subscribe to the above criteria. Jean-François Lyotard (1987), a representative of the latter tradition, asserts that 'narrative knowledge certifies itself in the pragmatics of its own transmission without having recourse to argumentation and proof.

Because reflexivists are concerned with the social order and its reproduction through the behavior of actors, they have to be attentive to questions of context and meaning within which these actions take place. Context and meaning are principally shaped by language, and reflexivists accordingly recognize the importance of norms and problems of interpretation. This recognition has important implications for the constitution of facts and their evidentiary value, the problem of truth and its role in 'theory- building.'

1. First: What is Evidence

Alfred Schutz (1973a) observed that all facts are created by cognitive processes. John Searle (1995) distinguished between 'brute,' or observable facts (e.g., a mountain), and 'social,' or intentional and institutional facts (e.g., an encyclopedia article). For the mainstream, facts are created by interpretation because they need to be detached from their context by intellectual abstraction. For reflexivists, they need to be considered in their particular setting. Neither the mainstream nor reflexivists can describe, let alone explain and theorize, without importing meaning. Reflexivists accept this as an epistemological given, not a corrupting intrusion to be ignored, controlled for, or explained away. Mainstream approaches, especially quantitative ones, often seem oblivious to the complications that arise from the impossibility of a pure empiricism. James Coleman has nevertheless shown that every measurement procedure that assigns a numerical value to a phenomenon has to be preceded by a qualitative comparison. While the assignment of

Table 1
What is evidence?

A	Mainstream—the observed, including the effects of social structures on behavior
B	Phenomenologists—subject's observations of self and structure
C	Interpretivists—B + observations of structure and subjects through each
D	Hermeneuticians—C + observer's relationship with subject and structure

numbers may permit powerful mathematical transformations, it is illicit to make such assignments if the antecedent qualitative comparison has not or cannot be completed (Coleman 1964). It is here that reflexivist and mainstreamers clash. The latter trying—illegitimately in the eyes of reflexivists—to apply the traditional measurement techniques and scientific operations to the social world, while the mainstream assert that even if the preconditions for successful measurements or causal modeling are not present, the 'scientific method' should serve as a regulative idea (see Table 1).

For the mainstream, the reality that matters most is usually that of the theorist who looks at the facts from an 'objective' point of view. Phenomenology, interpretivism and hermeneutics emphasize the subject's story, but treat this evidence differently. Interpretivists may or may not situate themselves in their own account of the evidence, whereas hermeneuticans are consistently reflexivist. Hermeneuticans may treat the social structure as a more or less autonomous evidence, and may or may not become evidence themselves as a result of changes they experience through the act of interpreting the subject.

The first move of the interpretivist is phenomenological. It is to gather inductively evidence that allows [the researcher] to evaluate the problem as understood by the actor (Ricoeur 1984, Berger and Luckmann 1966). This is not significantly different from applying a garden variety decision making approach to a social science problem. But the reconstruction of an actor's perspective and understanding may be only the first step. Treatments of evidence acquire validity when the subjects themselves recognize that understanding as adequate (Schutz 1973b, p. 64). Researchers must read back their interpretations to those who have been interpreted. This standard of validation is not uncontested, as it too is based on imputation. Paul Ricoeur insists that such interpretations as 'configurational acts'—might *not* be intelligible to the actors themselves, since they represent constructions that no actor or witness could have put together when the events occurred (Ricoeur 1984, p. 157).

Why not just ask people what their statements and behavior mean? Interpretivists and hermeneuticans insist that unobservable social structures exist in all social contexts, and make the meaning of a peoples actions and words not theirs alone to interpret or control (Ricoeur 1971). The researcher's task is to reproduce both the subject's actions *and* the encompassing frames and social structures within which this action took place. Imagine asking a white woman why she moved her handbag from one side of her body to the other when a group of black men were approaching on the other side of the sidewalk. She would probably deny her action was racist. But irrespective of her actual intentions, her behavior may have reproduced a particular racist structure by typecasting those young black men as potentially dangerous felons.

Hermeneuticans go a step further and insist that the role of researchers must be included in accounts of actions. (That this is a big step away from and beyond phenomenology is obvious in Schutz's (1973a, 1973b) uncritical treatment.) Researchers must comprehend the evidence through their own conceptions, while simultaneously realizing the perspectives of actors (1973a, pp. 104–11). Gadamer describes this process as placing oneself in the position of the other not to generate empathy or to better apply one's own criteria, but to attain a higher universality that overcomes, not only our own particularity, but also that of the other. This is the fusion of horizons of observer and subject (Habermas 1994). The hermaneutic assumption of intertexuality calls for an understanding of an actor's story in its complete relationship to all other stories available in a site, as well as in its relationship to the observer.

2. *Second: Where is Evidence Found?*

In a certain sense it is unfair to compare main stream and interpretivist evidence selection strategies, because they have different objectives. The mainstream wishes to test theories, interpretivism intends to generate understandings. Reflexivism wants to expand the field of inquiry. Instead of focusing only on 'why' questions that automatically privilege causal accounts, reflexivists insist that 'how possible' and 'what questions are also legitimate objects of inquiry. 'Why' questions direct attention to the antecedent causes in a phenomenon, whereas 'how possible' and 'what' questions are what makes a phenomenon what it is. (For a fundamental discussion of the differences between 'what' and 'why' questions, see Alexander Wendt 1999.) Reflexivists who study sovereignty are not interested, at least not primarily, in what brought it about but rather in the generative logic of the concept, its various instantiations in institutions, and the constraining and enabling force of its norms that allow actors to reproduce domestic and international structures.

To the extent that mainstream rules are designed to promote the assessment of alternative theoretical

accounts of a phenomenon, whether in a single case, or across some universe, they do not address interpretivist concerns. Because reflexivists do not usually engage in the competitive testing of theories, they are on the whole uninterested in the method of difference, crucial cases, and statistical significance. While the mainstream generally treats the knowledge of what constitutes a case as self-evident, reflexivists contend that case selection pre-judges 'facts' not yet in evidence.

This difference in orientation has implications for the choice of methods. While the mainstream stresses the deductive rigor in the models they employ, or the somewhat more problematic justifications for inductive inference, reflexivists stress the role of analogies and of conceptual frames that are antecedent to any causal argument or inductive inference. Given its choice of methods, the mainstream must consider counterfactual alternatives to which it rarely assigns 'weight' to the several causes considered relevant and apply the standard statistical controls to large enough samples whose cases are independent of one another. In practice, the mainstream decision rules that establish a warrant for a particular interpretation resemble more a procedure of sequential exclusion of alternatives than a 'deductive' demonstration or a generalizing inference.

Reflexivists do not believe it possible to establish complete analytical or theoretical control over a subject and consequently regard the 'bivalence' principle of logic (either something is so, or not so, and no third possibility exists) as generally inapplicable. The best our explanations can do is offer plausible accounts, even though some accounts might be more persuasive than others. Here a variety of criteria might interact and any proposed solution is not only 'preliminary' but contestable from the outset. By reversing the default drive of the mainstream theorists, who believe that a case or subject is theoretically identical until shown to be meaningfully different, reflexivists *assume* such differences among subjects until they can be persuaded otherwise. Reflexivists expect to find heterogeneity in their evidence; this is not a problem to be solved, but justification for emphasizing interpretative methods. They do not expect to find evidence that would allow predictions about future behavior within a case, let alone about a class of subjects across the historical universe. They argue that their stance allows for a more appropriate heuristics and thereby for a stronger warrant for their assertions. Reflexivist also believe that the traditional trade-off between rigor and detail misstates the underlying epistemological problem.

To elaborate on this last point, let us consider the problem of predictive success as a warrant for 'science.' Some mainstream approaches assert that the only warrant for a theory is predictive success, and that concern for the appropriateness of characterization is misguided. Reflexivists—scientific realists as well as logical positivists—counter this claim by pointing out that accurate predictions are sometimes made by demonstrably false theories, and that making predictions the only criterion for warranted knowledge discourages *systematic* search for mistakes when predictions fail. To that extent we are in danger of substituting *ad hocery* for disciplined inquiry. In short, reflexivists recognize control as a problem that is very hard to solve, and so reduce their claims about their theoretical constructs. The mainstream tends to see control as a technical problem, and so does not adjust its aspirations and expectations about evidence. (See, for example, the high level of attention 'omitted variable bias' receives in King et al. 1994, pp. 168–82, but how easily it is expected to succumb to simple techniques. On p. 172, we are reassured that it is possible to account for omitted variables because 'fortunately, in most cases, researchers have considerable information about variables outside their analysis.' Interpretivists recognize that such information is often impossible to obtain.)

3. *Third: Evidence and Theory-building*

The secret to mainstream confidence in using evidence to *test theories* and develop generalizations for use in other places and times is the implicit assumption of uniformity. This assumes that things, concepts, and people are meaningfully identical across time and space. Votes are votes, wars are wars, and power is power—everywhere and for all time. (This problem has long been recognized in political science. In 1970, Giovanni Sartori pointed out the perils of doing comparative research. He warned against 'stretching' the meanings of variables across contexts in order to achieve some contrived generalization.) Some mainstream scholars acknowledge that unit homogeneity is often impossible to attain, but argue that understanding the degree of unavoidable heterogeneity helps to estimate the degree of uncertainty or likely biases of inferences (King et al. 1994, pp. 93–4). They allow the theory being tested to determine if a feature is meaningful. But what if the theory—and in social science this is virtually always the case—is not perfectly specified so as to avoid omitting a meaningful source of variance?

For reflexivists, this misstates the problem and so offers a misguided solution. If we knew which untheorized features of the field were relevant to understanding social outcomes, presumably we would have accounted for them in the *a priori* theory. Reflexivists see no solution to this problem because it is impossible, logically, empirically, and theoretically, to specify a causal model that could possibly account for everything that is excluded outside its boundaries, and everything that is purportedly causally effective within those boundaries. Nevertheless, reflexivism is subject to the same tension between homogenization and

uniqueness. It objects to mainstream homogenization without often explicitly acknowledging that it also uses forms of homogenization to make sense of the social world. *Intersubjectivity* is really a way of homogenizing various parts of society or of historical developments through common webs of meaning. This is why common understandings and the practices based on institutional rules are so important. If people understand the same social practices in the same way, then they offer an example of sameness, of homogeneity or at least of comparability. It should be clear that we are not dealing with 'constant conjunctions' à la Hume, or antecedent (efficient) causes as in mainstream analysis. Common understandings or norms that are constitutive of practices (institutions) do not cause 'behavior.' They may provide weighty reasons, guide or inspire behavior, rationalize or justify it, or they may be ignored. They do not effect cause in the sense that a bullet through the heart causes death or an uncontrolled surge in the money supply causes price inflation (Kratochwil and Ruggie 1986, p. 767).

Reflexivists need a methodological technique to deal with both difference and similarity. (The difference between positivist and reflexivist homogenization of social reality is that positivism simply assumes homogeneity or uses prior theoretical categories to impose it on difference, while reflexivism theorizes about and uncovers homogeneity, treating it as a social phenomenon to be discovered and understood, not as one to be assumed and inflicted; see Habermas 1994, p. 108.) The image of a text, requiring both interpretation and an audience to which it is addressed and which brings to it certain pre-understandings that frame it, provides here their heuristic metaphor. Instead of viewing actors as simple 'throughputs' for norms and institutional rules, reflexivists point to the constraining and enabling functions of norms that cannot be reduced to a structure that determines behavior. Here the term 'actor' who has to act out or perform something rather than serve as a 'mechanism' or conveyor belt of forces is an alternative metaphor with considerable heuristic power. Reflexivist interrogation of mainstream homogenization is not always successful, but encourages us to guard against adducing identity where meaningful difference exists and vice versa.

The differences between the mainstream and reflexivism on this issue are easily observed when we consider how each handles the anomaly, a piece of evidence that confounds its theory or understanding of some phenomenon. The mainstream response is similar to how nineteenth century naturalists and anatomists dealt with physical 'monstrosities.' Rather than indicating systematic chaos or contingency, an interpretation with which reflexivists would feel comfortable, these anomalies instead provided opportunities to reaffirm the natural order and to define it with more precision. They were incorporated into more elaborate and powerful taxonomies (Ritvo 1997, p. 137). (Based on an empirical evaluation of the history of science, Pierre Duhem 1962) pointed out the power of prior theories, methodological commitments, and other elements of the scientific project to ensure the rejection of anomalous findings. Lakatos, Kuhn, and Feyerabend have also offered different accounts of why theories are so sticky in the face of countervailing evidence. As if these were not sufficient, there are excellent motivated and cognitive reasons to restore balance and persevere in ones prior beliefs.) While interpretivists try to place any given historical event within a narrative, within a contextualized bounded social structure, the mainstream places that same event within the framework of an opposition between the particular and the universal (Ricoeur 1984, p. 112). It treats the historical event according to whether it is part of a more general story to tell about the world, or an idiosyncratic 'outlier' that needs to be explained away.

Mainstream theories are sticky, in the Lakatosian or Kuhnian sense. Single pieces of contrary evidence do not and must not be used to change the observer's account of the situation (Kuhn 1970). But, ironically, and perhaps only in this regard, reflexivism is truer to Popperian (logical) positivism than the mainstream which adopts the practical and sociological defenses of Lakatos and Kuhn. Popper offers an account of falsification that resembles that of an open market or Darwinian system. Explanations fail when they are no longer competitive, according to established falsification criteria, relative to the other available explanations. In other words, one's understandings are far more responsive to new evidence, no matter how anomalous, than would be the case in either a Lakatosian or Kuhnian world. The product could be called a relative working truth, or, as Popper called it, 'situational certainty' (Popper 1979, pp. 78–81). It is a relative truth in that its validity exists only in comparison to other possible accounts. It is a 'working truth' in the sense that its validity is acknowledged to be a pragmatic convention, rather than an absolute fact that is simply part of the world.

Since we never test our theories against a reality, but only against other theoretical interpretations of the world, two additional problems arise. The first is that truth is no longer a property of 'the world out there,' but of theoretically informed assertions for which we need warrants. Consequently, the bivalence principle that links data and theories is no longer applicable, since most questions in science remain undecidable. This makes all scientific statements subject to revisions and introduces a second important element of pragmatics into theorizing. A statement, or 'fact,' is true only in the sense that can be accepted in daily practice until something more satisfactory comes along. It is a 'truth' in the sense that it is believed to be contextually valid, within the prior two constraints.

Reflexivism is doubly Popperian. It resists both theoretical priors, sociological and institutional con-

sensus, and motivated and cognitive impulsion toward premature closure. It treats every anomaly as if it is, at least potentially, meaningful. (Of course, Popper was never the naïve falsificationist who argued that a single disconfirmation meant the abandonment of the imperfect theory. Disconfirmation had to be accompanied by an alternative theory that could account for all that the failed theory had accomplished, and then some.) The mainstream, on the contrary, rejects treating every disconfirming instance as a serious challenge, instead appearing to don the protective belts offered by a Lakatosian research program, or the Kuhnian mantle of normal science toiling on within a settled paradigm.

Bibliography

Berger P L, Luckmann T 1966 *The Social Construction of Reality. A Treatise in the Sociology of Knowledge*, 1st edn. Doubleday, Garden City, NY

Coleman J S 1964 *Introduction to Mathematical Sociology*. Free Press of Glencoe, New York

Duhem P 1962 *The Aim and Structure of Physical Theory*. Atheneum, New York

Gadamer H-G 1987 Foreword. In: Baynes, Bohman, McCarthy (eds.) *Truth and Method, After Philosophy*, 2nd edn. MIT Press, Cambridge, MA, p. 340

Habermas J 1994 *On the Logic of the Social Sciences*. MIT Press, Cambridge, MA, pp. 108, 151

Lyotard J-F 1987 The postmodern condition. In: Baynes K, Bohman J, McCarthy T (eds.) *After Philosophy. End or Transformation?* The MIT Press, Cambridge, MA, p. 80

King G, Keohane R O, Verba S 1994 *Designing Social Inquiry: Scientific Inference and Qualitative Research*. Princeton University Press, Princeton, NJ, pp. 93–4; 168–82

Kratochwil F V, Ruggie J G 1986 International organization: a state of the art on the art of the state. *International Organization* **40**: 753–76

Kuhn T 1970 *The Structure of Scientific Revolutions*. 2nd. edn. University of Chicago Press, Chicago

Popper K 1979 *Objective Knowledge: an Evolutionary Approach*. Oxford University Press, Oxford, UK, pp. 78–81

Ricoeur P 1971 The model of the text: meaningful action considered as a text. *Social Research* **38**(4): 534

Ricoeur P 1984 *Time and Narrative*. University of Chicago Press, Chicago, Vol. 1, pp. 129–130, 157

Ritvo H 1997 *Platypus and the Mermaid, and Other Figments of the Classifying Imagination*. Harvard University Press, Cambridge, MA, p. 137

Sartori G 1970 Concept misformation in comparative politics. *American Political Science Review* **64**(4): 1033–53

Schutz A 1973a Common-sense and scientific interpretation of human action. In: Natanson M (ed.) *Collected Papers, The Problem of Social Reality*. Martinus Nijhoff, The Hague, Vol. I, pp. 6; 36–8; 104–11

Schutz A 1973b Concept and theory formation. In: Natanson M (ed.) *Collected Papers, The Problem of Social Reality*. Martinus Nijhoff, The Hague, p. 64

Searle J R 1995 *The Construction of Social Reality*. The Free Press, New York

Wendt A 1999 *Social Theory of International Politics*. Cambridge University Press, Cambridge, UK

T. Hopf, F. V. Kratochwil, and R. N. Lebow

Reform: Political

In its most ordinary sense, political reform is any departure from the status quo that a proponent believes or represents as an improvement. But if reform is defined as those changes that democratic political systems have adopted in order to improve the legitimacy, accountability, and efficiency of modern government, it is possible to identify some common core values and reform actions in advanced democratic states. A recent survey of the 29 member states of the Organization for Economic Cooperation and Development (OECD) found that the most frequently stated public service values were (in order): impartiality, legality, integrity, transparency, efficiency, equality, responsibility, and justice (OECD 2000). Since to date less than half the OECD countries have explicitly adopted the last five of these core goals, they constitute the political reform frontier.

Democracies incorporate core values into their legal frameworks in the form of statutes, constitutions, civil service regulations, and public service acts. Reformers seek to change the legal framework in order to better realize one or more of the core values. The three main sources of these core values are general democratic principles (i.e., values consonant with systems in which citizens ultimately control governments), professional norms (i.e., standards for those who assume roles in the government), and prevailing ethical systems (i.e., societal conceptions of appropriate behavior).

The OECD survey discovered that impartiality is the most common core value held by its member states. This is defined as acting in a neutral or objective manner when dealing with various citizens and groups. The central idea embodied by this value is that states should not be partial towards specific persons in the administration of their laws. Examples of modern reforms that aim to increase impartiality are the creation of a neutral civil service, merit reviews for public employees, rules for bidding and winning public contracts, and judicial independence. It is not surprising that this is the most commonly stated value in advanced democracies. A government that administered its laws in a biased or partial fashion could not possibly act in the public interest. Partiality undermines democracy (by favoring the interests of the few over the majority's will), administrative professionalism (by violating the trust that citizens have in the government to act on their behalf), and conventional ethical norms (fair treatment as practiced in civil society).

Granting the importance of impartiality to modern governments, there is a serious debate as to how far impartiality should extend in public life. Elected officials and their top policy advisors are expected to represent the groups and individuals who put them in office so some forms of 'partiality' are routinely accepted in modern democracies. The key reform questions occur over when partiality is permissible and at what levels of government. In general, those who make policy are allowed to be partial to parties, constituents, and key support groups, while those who implement (i.e., public administrators below the highest levels) and those who interpret the laws (i.e., the judiciary or regulators) are expected to be more strictly impartial. But there is considerable cross-national variation in this rule as exemplified by the differences between the American and British Civil Service traditions (Kaufman 1956).

Closely related to the first value is a second: legality (a stated value in 22 of the 29 OECD countries). Legality means that decisions are incorporated into published laws and codes, and that there is due process protection for individuals when governments make decisions. It would be rare to find unpublished laws in advanced democracies although many, for instance, provide for relatively unlimited emergency powers in crises. More commonly, the reform focus is on due process—i.e., what constitutes a fair hearing for individuals who might be affected by government decisions. The problem with due process is that it can create highly complex and slow proceedings, greatly adding to the expense and untimeliness of government decisions. In polities like the USA where judicial due process is highly developed, reform means finding ways to adjudicate matters more informally and expeditiously (Kagan 1991), and in polities where due process is lacking, it might mean adopting an administrative procedures act (i.e., Greece and Portugal).

Integrity, the third most commonly stated value among OECD nations, is the subject of the most heated reform debates. It means that public office holders should not place their own interests above the public interest in the performance of official duties. When they do, this is usually considered corrupt behavior (Rose-Ackerman 1999). The most common forms of corruption are bribery, conflict of interest, extortion, and personal enrichment. Bribery is the exchange of something of personal value for the performance of an official act (Lowenstein 1985). It is almost universally proscribed in advanced democratic societies, but can take increasingly subtle forms. Hence, some bans on gifts, honoraria, and post-public service employment are attempts to prevent potential quid pro quo exchanges that might distort the judgment of public officials.

In the USA, this has led to increasing conflict of interest regulation, the purpose of which is to remove interests that might potentially seem to be corrupting (McAlevy 1988). Hence, officials have been required to recuse themselves from participating in decisions that might directly and materially affect them. Alternatively, they might be asked to place their assets in a blind trust or even sell off potentially problematic holdings to prevent potential corrupt motives. Extortion and personal enrichment occur when public officials exploit the monopoly that the state has over some decision or service in order to extract rents. Some reforms that deal with this type of abuse include prohibitions against the unauthorized use of official or personal information, criminal penalties for those who misuse their office resources, anti-nepotism laws, limitations on accepting gifts, penalties for the failure to perform prescribed duties, bans against illicit enrichment, and laws against impeding or interfering with the public procurement process.

Corruption is a critical area of political reform. In Mexico, it has been a key issue for decades, and ultimately played a role in the *Partido Revolucionario Institucional*'s demise in the 2000 national Presidential election. In the USA, anti-bribery logic has been widely incorporated into the campaign finance reform debate. As the modern state has assumed a wider regulatory and service provider role, the potential to use the state for corrupt purposes has grown. Nations that seek to enter into closer trading and political associations with OECD nations are under considerable pressure to address their unresolved corruption problems.

Another hotly contested area of political reform is transparency. Transparency (meaning the openness of the government to citizen observation and participation) is valued both for its own sake and for its instrumental value in promoting other core values. Full disclosure of government information helps to produce a more informed citizenry, which is a good in itself if it improves the quality of electoral decision-making. At the same time, full disclosure fosters greater accountability in the sense that it makes it harder for public officials to act in particularistic and corrupt ways without the public finding out and voting them out of office. Similarly, public participation can be beneficial to citizens, but also to the agency and decision making bodies that must make choices affecting real lives.

The most common transparency reforms are open meeting laws and freedom of information requirements (Thompson 1987, pp.116–122, Pupillo 1993). In the case of the former, so-called 'sunshine' laws require that the public and press be notified of impending meetings in which decisions might be taken. Citizens are given time to comment on the contemplated actions, and votes and discussions related to the meeting must be held in the open. 'Freedom of information' statutes require the government to release documents and information upon public request. Almost all advanced democracies have some restrictions on the release of confidential information that might compromise national security, police

actions, or personal privacy. As before, one can find considerable variation among advanced democracies. The Official Secrets Act in Great Britain gives government ministers much more ability to resist the disclosure of information that would be potentially damaging or embarrassing to the government. At the other extreme, some states in the USA have placed the right to request government information into their state constitutions.

The fifth value is efficiency. It is fair to say that this is a more contemporary reform goal. In an earlier period, public sector behavior was focused on the impartial administration of rules and programs, while private sector behavior was driven by market forces and profit. In the latter half of the twentieth century, private sector norms have been imported into public sector thinking. Government has been 'reinvented' to be more consumer-oriented and efficient. Efficiencies can be achieved by fostering competition through the contracting out of service activities to private firms, the promotion of inter-agency competition, and the benchmarking of government performance linked to outcome based rewards. One of the large issues this trend raises is whether accountability can be retained when activities are given over to private sector agents (OECD 1999).

Accountability is the ability to hold public officials responsible for their actions and decisions. Transparency is essential to a system of accountability in the sense that the public needs to know what officials are doing in order to hold them responsible for what the government does. While necessary, however, transparency is not sufficient for accountability; i.e., the public needs to be able to reward and punish officials for the consequences and quality of their actions and decisions. In democracies, this means that voters must have the ability to vote out incumbents who fail them (i.e., hence the need for truly competitive elections) and elected officials must be able to control the actions of those who administer the laws.

The problem that arises in advanced democracies is that incumbent officials and administrative officials often enjoy electoral advantages and legal protections that insulate them from the need to be highly responsive to public demands. In the USA, term limitations for legislators and campaign finance reform are seen as means to increase turnover and responsiveness. Critics charge that this greater responsiveness is purchased at the price of lower expertise and weaker judgment (Cain and Levin 1999).

The last two core values are equity and justice. Equity can mean either procedural or substantive equality. In the US for instance, the equity principle led to the 'one person, one vote' doctrine of the Supreme court—essentially requiring that all districts that elect officials for general governmental bodies must be equally populated to avoid diluting the vote of one individual vs. another. It also informs the desire to place limits on campaign contributions so that the voice of one constituent is not unequal to that of another because the former has more wealth and resources than the latter (Cain 1995). The preoccupation with equity is stronger in the US than in most other advanced democracies.

Justice, on the other hand, refers to the substantive fairness of decisions that the government makes. Reform at this level begins to conflate with ideology, focusing on the government's obligation to make laws that provide positive opportunities for its citizens. Parties of the left and right will see these issues differently, and reform for one party can mean undoing the reforms of the other. The most general formulation that can be given to justice is that the laws accord with the common ethical and moral norms in civil society. Hence, where there is consensus in civil society about these norms, justice can be elevated beyond the merely political to broadly accepted goals.

1. *Perspectives on Reform*

This review of core values among OECD nations and the reforms that track those values is certainly not comprehensive. It merely illustrates the most common concerns. Apart from the debate over each of them, there are general differences among advanced democracies in reform orientations. They relate to the three sources of core goals: democratic, professional, and private ethical values.

With respect to the first, democratic values, there are important differences between those who interpret more democratic to be more majoritarian and those who resist such a formulation. American populists for instance prefer reforms that put more power directly in the hands of voters and less in the hands of elected and unelected officials. Neo-Madisonians in the US see popular control as lacking in the requisite checks against majority tyranny and in the expertise that representative government affords. Advocates of parliamentary systems prefer leaders to be chosen by (and therefore tied more directly to) other elected officials, and that the public show more deference to those in public office. Clearly, the position a society takes on this dimension strongly influences reforms that affect accountability, transparency, and equity.

Perspectives on professional values also shape modern governmental reforms. Systems that rely heavily on the socialization and training of public officials will tend to rely more on codes and internal monitoring (e.g., the British Civil Service). Polities that view professionalism as elitist or that are skeptical that people can be educated to high levels of public service, will rely on external checks and legal regulations (e.g., the American system).

There will also be important differences in the perspectives of those who believe that public ethics derives directly from private ethics vs. those who believe that the two differ in important ways. In the

view of the latter group, public officials need to be held to a higher standard and to ever growing scrutiny. In the view of the former, the two sectors can be analyzed within the same framework, and lessons learned from one sector can be imported to another. The shape of reforms to come will be determined by how future reformers situate themselves on these three continuums.

See also: Accountability: Political; Corruption: Political and Public Aspects; Democracy; Democracy, History of; Democratic Theory; Equity; Interest Groups; Justice and Law; Justice: Political; Social Justice

Bibliography

Cain B E 1995 Moralism and realism in campaign finance reform. *The Legal Forum*, pp. 111–40

Cain B E, Levin M 1999 Term limits. *Annual Review of Political Science* **2**: 163–88

Kagan R 1991 Adversarial legalism and American government. *Journal of Policy Analysis and Management* **10**(3): 369–406

Kaufman H 1956 Emerging conflicts in the doctrines of public administration. *The American Political Science Review* **50**(4): 1057–73

Lowenstein D W 1985 Political bribery and the intermediate theory of politics. *UCLA Law Review* **32**(4): 784–851

McAlevy K 1988 Conflicts of interest and state legislatures. *The Journal of Law and Politics* **5**(1): 209–37

OECD 1999 *Managing Accountability in Intergovernmental Partnerships*. OECD Publications, Paris

OECD 2000 *Trust in Government: Ethics Measures in OECD Countries*. OECD Publications, Paris

Pupillo T D 1993 The changing weather forecast: Government in the sunshine in the 1990s—An analysis of state sunshine laws. *Washington University Law Quarterly* **71**(4): 1165–87

Rose-Ackerman S 1999 Political corruption and democracy. *Connecticut Journal of International Law* **14**(2): 363–78

Thompson D F 1987 *Political Ethics and Public Office*. Harvard University Press, Cambridge, MA

B. E. Cain

Reformation and Confessionalization

1. The Reformation or 'Temps des Réformes'

The term 'Reformation' traditionally refers to those events and changes which in sixteenth-century Europe originated in church and theology and then took hold of the social, political, and cultural order, reorienting the thinking and behavior of individuals as well as social groups. From this point of view, the originator of all these changes was the German Augustinian friar Martin Luther (1483–1546), who published his 95 theses against the system of indulgences on October 31, 1517. This date was for the first time celebrated by Protestants as a secular holiday in 1617, and has since the nineteenth century been celebrated yearly as 'Reformation day' in countries all over the world. This chronological, geographical, and theological restriction to Luther and 1517 corresponds neither to the historical facts nor is it confessionally or historiographically undisputed. Rather, it is the result of a double interpretation: on the one hand, the contemporary heroization of Luther by his adherents or his demonization as arch-heretic by Rome and Roman Catholic countries like Spain (the Reformation as *pestis Germaniae*), and on the other hand, the monopoly on interpretation by the Lesser German (*kleindeutsch*) Lutheran historiography, which gained worldwide acceptance, particularly in the United States. This interpretation continued to influence the so-called Luther Renaissance of the 1920s and 1930s.

After World War II Reformation historians became less deeply rooted in confessional traditions, and a pluralism of methods and theories gained acceptance in historiography. As a consequence, the traditional 'Lutheran restriction' (*lutherische Engführung*, Bernd Moeller 1998) lost influence, and the Reformation increasingly appeared as a chronologically and geographically wide-ranging phenomenon, as a 'temps des Réformes' (Jean Delumeau 1997), a period of change from the medieval *societas christiana* to the differentiated, multi-confessional society of the early modern period. It increasingly appeared as a period of several centuries, in which the Reformations of Luther, Zwingli, Calvin, and all the other sixteenth-century reformers represented an important phase which accelerated a long-term development. Under the specific historical and historiographical conditions of English history, there was recently also developed a revisionist historiography, which on the one hand proceeds from the assumption that there were a number of 'Reformations' in sixteenth-century England and which on the other hand stresses the continuity of late medieval conditions in church and religion, particularly in popular religion (Eamon Duffy 1992, Christopher Haigh 1993).

To describe the meaning of 'Reformation' in current scholarly discussions as well as from a broader cultural perspective, three main dimensions should be distinguished: first, the historical events; second, the myth of the Reformation in Protestant and 'Western' self-understanding and self-assurance; and third, the historiographical concepts and the main issues of current research.

2. The Historical Context

On the one hand, there were new beginnings in theology, piety, spirituality, and the social manifestations of religion in the highly urbanized zones of late medieval Europe, above all in Italy, in the

Netherlands, in central Europe including Bohemia, in England, France, and Spain (the reform crisis around archbishop Cisneros, 1436–1517). In the context of the Humanism of the *Devotio moderna* and similar currents of thought, the laity acquired a more important role in the church, rational forms of church service (preaching) and religious behavior were strengthened, and a 'normative centeredness' (*normative Zentrierung*, Berndt Hamm) as well as an organizational and spiritual formation set in.

On the other hand, decisive changes occurred in the organization and administration of the church as well as in the relationship between state and church. The development of early modern national and territorial states led to a regionalization—but not yet to a division—of the universal church into 'national', territorial, or town churches, in which the secular governments gained control over the church of their respective territories either through concordats with the Pope or through a revolution (as the Hussites in Bohemia).

The Roman curia had also changed dramatically since the fourteenth century, because of the formation and bureaucratization of its administration and because of the triumph of monarchical papalism over conciliarism after the Council of Basel. Recent research on the papacy therefore considers Rome to have been more advanced on the path of modernization, and sees Luther's Wittenberg Reformation as the result of a crisis of modernization in a remote, backward area of Christendom. Through Luther and the other reformers the religious and ecclesiastical changes received a new direction, legitimacy, and dynamism, which since the middle of the sixteenth century also took hold of the Roman Catholic church and changed it further. In particular the sola formulas—only through faith, grace, and the reading of the Bible could human beings be justified before God (*sola fide*, *sola gratia*, *sola scriptura*)—focused the different reform attempts in a new and unique way. The widespread church criticism of the late Middle Ages was put on a new foundation. Medieval, accidental anticlericalism developed into early modern fundamental anticlericalism. The popes and the clergy were no longer criticized only for the neglect of their duties, but it was claimed that the priests and the pope were superfluous as intermediaries between God and human beings: indeed, that they were harmful to the salvation of the individual, the community, and society as a whole. In this way, the pope and the Roman church were—in the eyes of their opponents—for the first time completely de-legitimized and unmasked as the Antichrist.

This also meant that the state or other secular governments (town councils, estates, etc.) gained more power over the church. When there was no longer a difference between clergy and laity, the prince could claim to be *praecipuum membrum ecclesiae* (privileged member of the church) and administer and govern the church as *summus episcopus* or 'supreme head of the church.' In the late Middle Ages, there existed territorial or national churches within the universal church. With the Reformation, Protestant territorial or state churches which no longer belonged to a universal church came into being.

Luther's appearance (above all at the Imperial Diet of Worms in 1521), Zwingli's appearance in Zurich, and Calvin's appearance in Geneva in the 1540s, but also the appearance of several other reformers who had their own ideas about theology and church policy, thoroughly changed the intellectual and religious climate and opened up new perspectives in state and society. The printing press became decisive for the quick dissemination of the new teachings: not only were the writings of Luther and other reformers published in print runs of hundreds of thousands, but the mostly illustrated broadsheets constituted a new medium of mass communication. In this way, not only the intellectuals, the political elites, and the princes were seized by the new visions and ideas, but also the lower nobility, the burghers, and the peasants. In the Holy Roman Empire, the Reformation was a mass movement which climaxed in the rebellion of the imperial knights (Sickingen, Hutten), the Peasant's War of 1525/26 and—until the 1540s—in burgher movements in the imperial and territorial towns.

Because of the close connection between religion and society in old Europe, the Reformation of the church inevitably led to a reformation of the political and social order. In this context, the idea of a communal church, decisively influenced by the laity, had a strong appeal. In the secular sphere, this corresponded with corporative (estate), communal, or republican (urban) concepts of order defended by the knights or the urban and rural communities (*Gemeindereformation*, communal Reformation). Later, comparable reform movements occurred in the Netherlands, France (nobility and towns, above all in the south), Poland (nobility), with considerable delay also in England (in the seventeenth century, above all in London and among the gentry). These communal movements only had a lasting success in regions with favorable social, political, and constitutional preconditions, above all in Switzerland and the northern Netherlands.

In most of the states where the Protestant Reformation took hold, including the German territories, the princes and the officials of the central state defeated these Reformation movements of the nobility and the common man 'from below.' They took the Reformation into their hands and utilized it for their own interests, especially the formation of the early modern state and of an early modern society of disciplined subjects (*Fürstenreformation*, princely Reformation). As a consequence, early modern territorial churches (as in the Empire) or national churches (as in Scandinavia, England, etc.) came into being. These were increasingly staffed by a university educated and

married clergy, whose area of responsibility was identical with the territory of the respective state and who in numerous ways supported early modern state-building.

In this context, important developments were: First, the increase in funds and property assets through secularization; second, the extension of the activities of the state upon the so-called *res mixtae*, i.e., education, marriage and family, poor relief, and social welfare; third, the mutual strengthening of religious and political identities among the people; fourth, the increased legitimacy of the prince who as supreme head of the church or *summus episcopus* had a quasi-sacral quality; and fifth, the instruments of the state church (church discipline, visitations, sermons, and catechesis) which aimed at influencing the behavior and thinking of the people and which effectively supplemented secular forms of disciplining.

However, the climax of this connection between religious and political change was only brought about by the process of confessionalization between ca. 1550 and ca. 1650, which was a fundamental process of society on the basis of explicitly formulated confessions of faith. On the Protestant side these were the *Confessio Augustana* of 1530 and especially the *Book of Concord* of 1580 for Lutheranism, the *Helvetic Confessions* of 1536 and 1566 as well as the other 'national' confessions, but especially the *Canons of Dordrecht* of 1619 for Calvinism, the *Book of Common Prayer* of 1549, and the *Thirty-nine Articles* of 1563 for Anglicanism. By committing the subjects to a fixed set of norms concerning dogma, belief, and behavior, the different confessional churches became clearly delimitated and this corresponded to a high degree of integration on the 'inside.' At the same time, the processes of formation in state and society became more dynamic and were given a new legitimacy.

As the Catholic church also went through such a process of confessionalization (based on the doctrinal canons of the Council of Trent, especially the *Professio Fidei Tridentina*), the Catholic and re-Catholicized countries of Europe (besides the German territories, e.g., Austria, Bavaria, and the prince bishoprics, above all Spain, Poland, France, and the Italian states) participated in the above-mentioned renewal of state and society in the wake of the reformation of the church. In contrast to Max Weber and the sociology of religion influenced by him, the historiography on confessionalization no longer assumes that only one of the early modern confessional churches—the Calvinist church—was the driving force of modernization, but that modernization resulted from the Calvinist, the Lutheran, and the Catholic confessionalization: in this respect, all three confessionalizations were functionally equivalent (Wolfgang Reinhard 1977).

However, in Catholic societies in particular there were often extra-confessional factors like geographical location, structure of the social elites, etc. that hampered the potential for modernization of the Catholic confessionalization. In general, the state authorities, which in the course of the Reformation and the processes of confessionalization aimed at social, political, and cultural concentration and standardization as well as at the formation of an early modern society of disciplined subjects, were again and again confronted with the opposition of these subjects. Popular religion, including the belief in magic, remained widespread as did older forms of behavior which contradicted the norms of state and church. Changes in society more often than not lagged behind the normative aims of state and church. Moreover, it should not be forgotten that the differentiation of Latin Christendom had from the beginning overshot the Protestant Reformations accepted by the state. Already during the sixteenth century Baptists, Spiritualists, Anti-Trinitarians, etc. formed a big group of '*Chrétiens sans église*' (Kollakowski 1969) outside of the established confessional churches, and there were also some atheists. After the age of confessionalization in the second half of the seventeenth century, such trends became more and more important.

3. *The 'Reformation' as Myth and Concept in Universal History*

Just like the French Revolution, the Reformation also became a *lieu de memoir* and a myth: first of all for the Germans, but eventually with a universal claim. Luther was already made a hero by his contemporaries, for example as *Hercules Germanicus*, who fights against the clerical Hydra with a mace (woodcut by Hans Holbein, 1520). Especially, the nailing of the 95 theses to the door of the Castle Church in Wittenberg became an icon of modern times. However, this event appears only one generation later in one of Melanchthon's writings, and therefore probably never happened. But for generations of history teachers and the educated bourgeoisie the blows of Luther's hammer on the church door were a signal for the beginning of the modern period.

A critical distance to the religious meaning of the Reformation notwithstanding, the Enlightenment prized Luther and his rebellion against Rome as an important steppingstone towards liberty of conscience and the emancipation of the individual from authority, an interpretation further elaborated by Hegel and the liberals of the nineteenth century. Hegel identified the Reformation with the rise of modernity and saw it as '*die alles verklärende Sonne*,' which follows upon '*die Morgenröte des Mittelalters*.'

Leopold von Ranke's interpretation of the Reformation was even more influential in historiography. In his *Deutsche Geschichte im Zeitalter der Reformation*, which was first published in 1839, widely read and translated into several languages, Ranke interpreted the German Reformation of 1517 as a signifi-

cant turning point in universal history. Hegel's and Ranke's interpretations of course included the Calvinists and Western European Protestants, who called themselves '*églises réformées*' or '*ceux de la réligion réformée*.' In a secularized form, this interpretation became one of the kernels of the concept of 'Western civilization,' particularly influential in the USA, as well as of the Marxist concept of human progress by revolutions (see Sect. 4, *frühbürgerliche Revolution*, early bourgeois revolution). In current historiography the term 'Reformation' remains relevant especially in the English-speaking world, where handbooks and encyclopedias are given the title 'of the Reformation,' even if they are equally concerned with the Renaissance and the period of confessionalization and cover much more than the history of religion and the churches.

4. Historiography and Issues of Current Research

In the research and interpretation of the Reformation and the period of confessionalization four main currents can be identified at present: first, church history in the stricter sense of the word; second, societal history (*Gesellschaftsgeschichte*), which modifies the classic sociology of religion in the tradition of Max Weber and Ernst Troeltsch and adapts their findings to the historiography of the early modern period; third, neo-Marxist interpretations; and fourth, post-modern approaches like cultural, gender, and ethno-history.

Its commitment to modern scholarly standards notwithstanding, church historians in theological seminaries are still oriented towards the meaning of the Reformation for the history of salvation as well as the historical background of the respective scholar's own confessional existence. In contrast to this, the other approaches emphasize the secular preconditions and consequences of the Reformation. Ranke's interpretation of the Reformation as a turning point in universal history pointed the way and was further developed by Max Weber, who brought the potential of the modern period for rationalization, discipline, and modernization in connection with the Reformation (*Die calvinistische Ethik und der Geist des Kapitalismus*).

The Marxist interpretation of the Reformation as an 'early bourgeois revolution' (*frühbürgerliche Revolution*, F. Engels, M. M. Smirin, M. Steinmetz) interpreted the role of the Reformation in universal history in a revolutionary, progressive sense. Moreover, in the historiography of the German Democratic Republic, the German nationalist perspective of the nineteenth century was retained: Reformation and 'early bourgeois revolution' were seen as compensations for the lack of a secular bourgeois revolution in Germany. At present these macro-historical interpretations continue to make themselves felt in two research paradigms developed by German historians: in the thesis of 'the Revolution of 1525' and the interpretation of the Reformation as communalism by Peter Blickle and Thomas A. Brady on the one hand, and in the paradigm of confessionalization by Wolfgang Reinhard and Heinz Schilling on the other hand. The paradigm of confessionalization neutralizes Weber's fixation on Calvinism and instead sees the dynamism of change in the overarching process of confessionalization be it Lutheran, Catholic, or Calvinist. In France and in other Romance countries the history of the Reformation was and is not at the center of interest.

In the countries of Eastern central Europe a consciousness for the general importance of the early modern history of church and religion is only gradually developing. English historiography follows its pragmatic traditions also in this respect, and—apart from the revisionism of the English Reformation mostly lead by Catholic historians—it is little interested in the theoretical and methodological discussions about the macro-historical role of the Reformation. By contrast, the approaches of micro-history and cultural history, which claim to supplement or replace the traditional macro-historical interpretations and which have recently become the focus of attention, have their roots especially in Anglo-American historiography. On the one hand, the advocates of these approaches stress the continuous meaning of magic and popular religion (Bob Scribner 1981), which also in Protestant countries resisted the rationalization of the interpretation of the world as well as the disciplining of belief and behavior according to the norms of the respective confessional churches. On the other hand, they stress the long-term change in the meaning of rites and images, in the concepts of order, thinking, and behavior, etc., from the late Middle Ages until the eighteenth century, and they thereby open up new sources and further knowledge to the discussion about the character of the Reformation and its historical meaning.

See also: Catholicism; Cultural History; European Religion; Fundamentalism (New Christian Right); Media and History: Cultural Concerns; Political Parties, History of; Poverty in History; Protestantism and Gender; Renaissance

Bibliography

Archiv für Reformationsgeschichte: Internationale Zeitschrift zur Erforschung der Reformation und ihrer Weltwirkungen [Archive for Reformation History: An International Journal concerned with the history of the Reformation and its significance in world affairs] 1903/1904

Blickle P 1992 *The Communal Reformation*. Humanities Press, Atlantic Highlands, NJ

Chaunu P 1975 *Le Temps des réformes: Histoire religieuse et systeme de civilisation: La Crise de la chrétienté. L'Éclatement, 1250–1555*. Fayard, Paris
Delumeau J, Wanegffelen T 1997 *Naissance et affirmation de la Réforme*. P.U.F., Paris
Duffy E 1992 *The Stripping of the Altars: Traditional Religion in England c. 1400–c. 1580*. Yale University Press, New Haven, CT
Guggisberg H R, Krodel G G (eds.) 1993 *Die Reformation in Deutschland und Europa: Interpretationen und Debatten [The Reformation in Germany and Europe: Interpretations and Issues]*. Gütersloher Verlagshaus, Gütersloh, Germany
Haigh C 1993 *English Reformations: Religion, Politics, and Society under the Tudors*. Clarendon Press, Oxford, UK
Hillerbrand H J (ed.) 1996 *The Oxford Encyclopedia of the Reformation*, 4 vols. Oxford University Press, Oxford, UK
Kolakowski L 1969 *Chrétiens sans église: La Conscience religieuse et le lien confessionnel au 17e siècle*. Gallimard, Paris
Moeller B (ed.) 1998 *Die frühe Reformation in Deutschland als Umbruch*. Gütersloher Verlagshaus, Gütersloh, Germany
Oberman H A 1994 *The Reformation: Roots and Ramifications*. W B Eerdmans, Grand Rapids, MI
Reinhard W 1977 Gegenreformation als Modernisierung? Proleg zu einer Theorie des konfessionellen Zeitalters. *Archiv für Reformationsgeschichte* **68**: 226–52
Reinhard W 1989 Reformation, Counter-Reformation, and the early modern state: A reassessment. *The Catholic Historical Review* **75**: 383–404
Schilling H 1992 *Religion, Political Culture and the Emergence of Early Modern Society: Essays in German and Dutch History*. E J Brill, Leiden, The Netherlands
Schilling H 1995 Confessional Europe: Bureaucrats, la bonne police, civilizations. In: Brady Th A Jr, Oberman H A, Tracy J D (eds.) *Handbook of European History 1400–1600. Late Middle Ages, Renaissance and Reformation*. Vol. II, E J Brill, Leiden, The Netherlands, pp. 641–81
Scribner R W 1981 *For the Sake of Simple Folk: Popular Propaganda for the German Reformation*. Cambridge University Press, Cambridge, UK

H. Schilling

Refugees: Geographical Aspects

Migrants are commonly differentiated based on their motivation for moving. The term 'refugee' is used to denote an individual who is forced flee in search of safety in contrast to someone who chooses of their own accord to relocate, i.e., a voluntary migrant. This entry examines geographical aspects of forced migration, a condition that applies to more than 34 million of the estimated 100 million permanent migrants worldwide.

1. *The Term 'Refugee'*

Legal distinctions regarding refugees can be traced to the United Nations Conference on the Status of Refugees and Stateless Persons. Article 1 of the 1951 Geneva Convention Relating to the Status of Refugees applied the term to

> [Any person who] … owing to a well-founded fear of persecution for reasons of race, religion, nationality, membership of a particular social group, or political opinion is forced to migrate involuntarily across an international boundary and remain outside his country of nationality.

Cause, motive, and political boundary thus became the classic benchmark determiners of refugee status. The 1951 Refugee Convention specifically addressed the status of Europeans displaced by WWII.

An amended 1967 Protocol established legal standards for refugees in all world regions at any time. Although the 133 signatories are not required to provide asylum to refugees, they are prohibited against *refoulement*—expelling or returning refugees to countries where their life or freedom would be threatened (Hamilton 1999).

Under international law, final authority to grant refugee status rests in the hands of a given host government. Practically speaking, however, a branch of the United Nations bears the primary task of protecting refugees. In 1951, the UN General Assembly created the Office of the United Nations High Commissioner for Refugees with a brief three-year life span in mind. In 1954, however, the Assembly extended UNHCR's mandate for another five years. Successive five-year renewals have followed ever since.

Headquartered in Geneva, UNHCR is charged with refugee protection and the promotion of lasting solutions to their displacement. According to its statute, UNHCR is competent to assist:

> [Any person who] … owing to well founded fear of being persecuted for reasons of race, religion, nationality, or political opinion, is outside the country of his nationality and is unable or, owing to such fear or for reasons other than personal convenience, is unwilling to avail himself of the protection of that country; or who, not having a nationality and being outside the country of his former habitual residence, is unable or, owing to such fear or for reasons other than personal convenience, is unwilling to return to it (UNHCR 1993).

Protection remains UNHCR's *raison d'être*, but the complexities of post-Cold War conflicts have forced the agency to broaden its scope considerably. Whereas the Office of the United Nations High Commissioner for Refugees once restricted its concern to people outside their country of origin, it has become increasingly involved in assisting and protecting returnees who repatriate to their home countries as well as groups of people who face refugee-like situations within their country of origin (UNHCR 1993). As the dimensions and dynamics of forced migration change so the term 'refugee' has come to apply to other forms

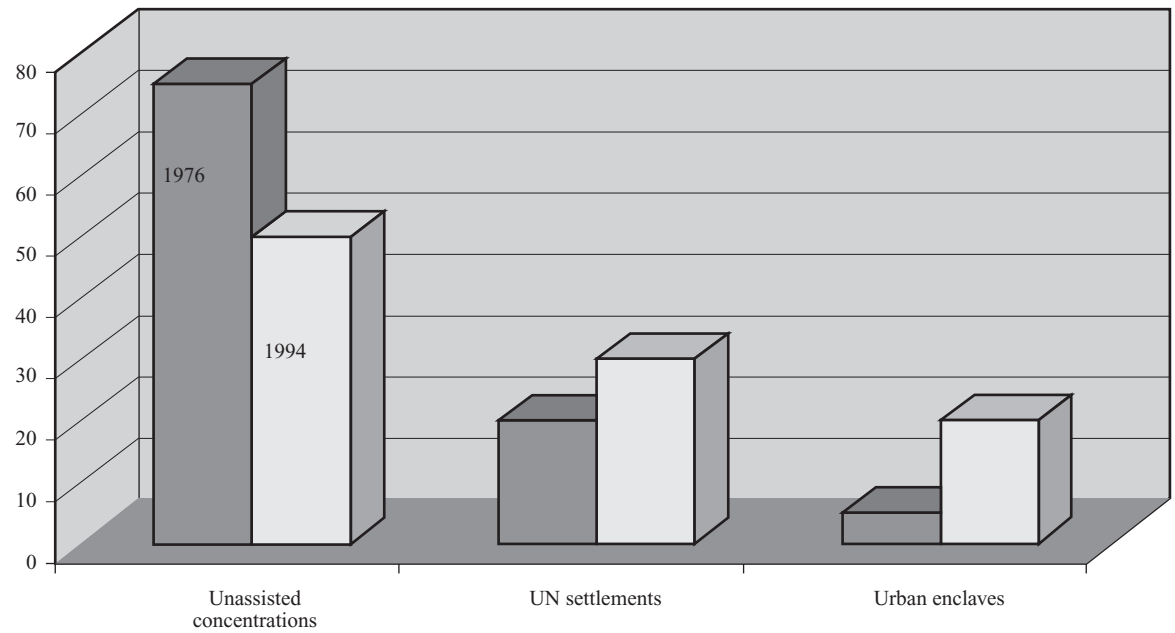

Figure 1
Principal forms of refugee settlement in Africa, 1976 vs. 1994

of displacement. The discussion to follow examines 'classic' international refugees who cross a political boundary and then, several types of 'internal' refugees.

1.1 International Refugees

Although geographers, starting with Ravenstein's work in the late 1800s, have long studied voluntary migration, their analysis of forced migration lagged far behind. At the 1950 meetings of the Association of American Geographers, Proudfoot introduced a nine-fold categorization of refugee population. That paper constitutes an early example of geographic work on the specific topic of refugees. Geographers thus began documenting the flight and distribution of specific refugee populations, and, they have continued to do so ever since.

Kunz (1973) offered a set of conceptual postulates structured around the dimensions of space and time. That a sociologist provided the model of refugee flight patterns reflected that fact that geographers still gave limited attention to forced migration. In 1977, an article published in *The Professional Geographer* summarized the state of the refugee literature. Rogge concluded that geographers had produced a considerable volume of literature on human mobility, but that '… one form of mobility, however, namely the forced or impelled migration of refugees, has received only peripheral attention by the discipline.'

Refugee research, including that of geographers, has grown substantially during the last 25 years (Hyndman 1999). Refugee study programs now exist at universities in Canada, Egypt, Great Britain, Kenya, Sweden, and the United States. Academic interest has correlated with the growing magnitude of refugee crises worldwide wherein the global total grew from less than 3.5 million in 1977 to a peak point of 18.2 million in 1992 (UNHCR 1993). Although media coverage tends to reinforce the perception that all refugees are in camps, most refugees—at least in the case of Africa—are still more likely to reside outside UN settlements (see Fig. 1).

During the 1980s, refugee research largely focused on the first and second stages of the refugee experience —the processes of flight and resettlement. Resettled Central Americans and Indochinese in the United States were the focus of study in the context of more developed countries. That East Africa had become the epicenter of refugee crisis in the developing world was reflected in books on Eritrean and Ugandan resettlement (see Rogge 1985, Harrell-Bond 1986, and Bulcha 1988). A basic distinction was frequently made between assisted refugees housed in UN-sanctioned camps or settlements and unassisted refugees who either self-settled in rural areas or in urban centers (Wijbrandi 1990).

In 1990, the locus of the global refugee population was predominantly in the global South. Developing countries hosted nearly 97 percent of all refugees. Conflicts in Afghanistan, Ethiopia, and Mozambique produced populations that each numbered in the millions apiece. The Palestinian community (2.48

million) had long since become permanent stateless exiles dispersed throughout the Middle East. Together, these four groups accounted for more than 13 million of the 16.6 million refugees worldwide. Yet, refugees could be found among more than one hundred different countries. And exile periods were extending longer and longer.

During the decade of the 1990s, geographers diversified their approach to match the growing complexities of the refugee picture worldwide (Black 1991 and Black and Robinson 1993). Refugee settlement was treated less and less often as if it were a self-contained process in isolation, thus moving beyond a 'refugee-centric' perspective (Chambers 1986). Bascom (1993), for example, related the larger processes of rural transformation and agrarian change to that of refugee settlement. Although media images give the impression that refugees are mass 'flows,' it became abundantly clear that ethnic, age, gender, and class differentiation all exist among uprooted populations (Bascom 1998).

Repatriation became a 'lead' research theme in the 1990s (Wood 1989 and Allen and Morsink 1994). As early as 1984, UNHCR had identified repatriation as one of three 'durable' solutions for refugees. The *de facto* solution is long-term integration of refugees in the country of first asylum. The exceptional solution, reserved for only a small percentage of asylum applicants, is resettlement to a third country of asylum. The optimum solution of the three is repatriation—the return migration of refugees from exile to their homeland. The UN High Commissioner for Refugees proclaimed the 1990s as the 'decade of repatriation.'

Repatriation research focused first on legal or logistical matters, then shifted to the socioeconomic and psychological dimensions of repatriations (Rogge 1991). Mozambique, Eritrea, and Afghanistan were some of the principal countries of concern. In theory, the repatriation decision is meant to be voluntary, fully informed, and only to occur when the conditions that prompted flight no longer exist. Practically, however, repatriations can involve different scenarios such as refugees choosing to return despite ongoing conflict or being forced to return home by host governments. Forced deportation, or *refoulement*, of refugees is the worst possible violation of the 1951 Refugee Convention.

The face of refugee studies changed dramatically during the last decade of the twentieth century. It began with the dissolution of the Soviet Union, whose fragmentation prefigured a pattern to come; states under siege amid competing national interests and thus producing complicated flows and counter-flows of displaced people. Wars in the Persian Gulf and Yugoslavia spiked the refugee total up to its record high. The year of 1994 marked the largest and fastest exodus in the modern history of refugee migration. More than 250,000 Rwandans fled the country in only 24 hours. Yet the era of mass movement across discrete political boundaries was already giving way to more 'complex humanitarian emergencies' (Hyndman 1999).

The painful saga of intervention in Somalia followed by the Rwandan debacle have made the international community increasingly aware of the relationship of refugees to peacekeeping, peacemaking, political stability, and humanitarian action. The UN began emphasizing the need to:

(a) incite and expedite conflict resolution in source countries so refugees can return,

(b) take vigorous action—including emergency response to crises and the provision of international protection on site—to eradicate the causes of flight before refugees flee, and

(c) initiate more proactive intervention to prevent the development of conditions that might impel people to leave in the first place (UNHCR 1993).

The United Nations forged cross-mandate approaches to refugee crises. In the case of Somalia and Mozambique, for example, UNHCR sought collaboration with governments, the UN Department of Humanitarian Affairs (DHA), other concerned United Nations agencies, and NGOs to establish means of protection and assistance for all affected populations, i.e., returnees, internally displaced persons and other impoverished nationals. UNHCR has also increased assistance for refugees outside formal camps and settlements as well as to local populations in resettlement areas.

Meanwhile, the balance of the globe's refugee population began to swing toward Europe. People had come to think of refugees as a Third World problem, but the shattering of Yugoslavia forced Europeans and North Americans to engage with refugees much more directly. First, warfare sites moved closer. Then refugees turned to neighboring countries for protection, assistance, and asylum. And the need for political stability in southern Europe as well as ethnic links to Bosnians, Croatians, Slovenians, and Serbs could not be ignored. Whereas Europeans and North Americans enjoyed the luxury of modest refugee resettlement programs during the 1980s, they were inundated with asylum requests during the 1990s. The United States, Canada, and 17 European nations all received refugees from Bosnia. Meanwhile, the plight of Kurds entrapped within Iraq initiated a new debate about 'internal refugees.'

1.2 Internal Refugees

During the 1980s and 1990s, the character of war shifted in balance from conflict between sovereign states to civil conflict. International refugees now represent the smaller portion of the globe's displaced population. Internally displaced persons (20 million) outrank the number of 'classic' refugees—asylum-seekers who have involuntarily crossed an interna-

Table 1
Significant populations of internally displaced civilians, as of early 1999

Country	Total number
Sudan	4,000,000
Angola	1,000,000–1,500,000
Columbia	1,400,000
Iraq	1,000,000
Afghanistan	540,000–1,000,000
Burma	500,000–1,000,000
Turkey	400,000–1,000,000
Bosnia–Herzegovina	836,000
Azerbaijan	576,000
Sri Lanka	560,000
India	520,000
Burundi	500,000
Rwanda	500,000
Lebanon	400,000–500,000
Uganda	400,000
Russian Federation	350,000
Peru	340,000
Congo-Kinshasa	300,000
Sierra Leone	300,000
Georgia	280,000
Cyprus	265,000
Yugoslavia	257,000
Congo-Brazzaville	250,000
Somalia	250,000
Guinea-Bissau	200,000
Kenya	200,000
Ethiopia	150,000
Syria	125,000
Philippines	122,000
Algeria	100,000–200,000
Eritrea	100,000
Liberia	75,000
Croatia	61,000
Armenia	60,000
Bangladesh	50,000

tional boundary and remain outside their country of nationality. Sudan (4 million) tops the list of countries with large numbers of civilians displaced by persecution or armed conflict (see Table 1).

'Internal refugees,' more commonly known as internally displaced persons (IDPs), became a central focus of refugee research (see Bennett 1998, Cohen and Deng 1998, and Hampton 1998). In 1999, the IDP table in the *World Refugee Survey* included 41 countries as compared to only 25 in 1990. IDP growth reflects the pandemic of civil conflict as well as paramilitary activity in settings of chronic social anarchy like Somalia and Columbia. Most IDPs are refugees in every respect except that they have not, or, cannot cross an international boundary. Although civilians are more frequently targeted as a deliberate tactic of war, they have no right to protection or humanitarian assistance as long as they remain inside their countries. Estimates of the number of internally displaced persons are often fragmentary and unreliable because they remain largely inaccessible to outside monitors and non-governmental relief organizations (Hamilton 1999). The total number of IDPs may be much higher than 17 million.

'Internal refugees' have become a central policy concern. In 1992, the secretary-general of the UN appointed a Representative on Internally Displaced Persons. Representative Francis Deng continues to parlay the status associated with his title into a limited protection framework for IDPs. The current working definition of IDP is that of a:

> Person or group of people who have been forced to flee or to leave their homes or places of habitual residence as a result of, or in order to avoid, in particular, the effects of armed conflict, situations of generalized violence, violations of human rights or natural or human-made disasters, and who have not crossed an internationally recognized border (Bennett 1998).

Clearly, the locus of concern is limited to those forcibly uprooted within a given national territory. The working definition is broadly inclusive, combining several forms of internal migration, which can also be are treated as separate entities.

Most IDPs are forced to move due to explicit forms of persecution or armed conflict, but others migrate because of natural or man-made disasters. Victims of clear-cut natural catastrophes such as a hurricane in Honduras, massive mud slides in Venezuela, or repeated earthquakes in Turkey are sometimes referred to as 'environmental refugees.' Yet, famine-induced migrations often involve man-made factors. Drought, for example, was the immediate, observable cause of large population movements in Ethiopia and Sudan during 1984 and 1985. Poor rainfall may have prompted flight, but displacement within Ethiopia was amplified by the deliberate obstruction of food aid into stricken areas. In Sudan, more than 100,000 people left their homes. Largely at fault was the structure of the Sudanese political economy that set the stage for migration out of rural regions. For many who left, migration was but the culmination of an impoverishment process associated with the appropriation of labor and land by mechanized agricultural schemes, obstructionist tactics employed by merchants, and the state's long-term neglect of the rural sector. The flight of 'environmental refugees' like these is a symptomatic manifestation of oppressive economic and political conditions.

Two other forms of forced migration bear mention —migrants in refugee-like conditions and development-induced displacement. In its highly regarded annual *World Refugee Survey*, the US Committee for Refugees, a public information program of Immigration and Refugee Services of America, reserves a separate table entitled 'For people in refugee-like

Table 2
Significant groups of people in refugee-like situations, as of early 1999

Host country	Place of origin	Total number
Russian Federation	Former USSR	1,000,000
Jordan	Palestinians	800,000
Thailand	Burma	350,000
Bangladesh	Pakistan (Biharis)	238,000
Jordan	Iraq	150,000
Syria	Stateless Kurds	142,000
Kuwait	Kuwait (Bidoon)	117,000
Iraq	Kuwait (Bidoon)	100,000
Tanzania	Burundi	95,000
Ukraine	Stateless Tartars	91,000
Syria	Palestinians	64,000
Venezuela	Colombia	60,000
Eritrea	Ethiopia	35,000
Cameroon	Chad	30,000
Congo-Kinshasa	Rwanda	30,000
Ecuador	Columbia	30,000
Kyrgyzstan	Tajikistan	26,000
Syria	Iraq	23,000
Lebanon	Palestinians	22,000
Yemen	Iraq	20,000
Yemen	Somalia	15,500
Turkmenistan	Tajikistan	11,000
Slovenia	Former Yugoslavia	8,000
Kenya	Rwanda	7,000
Panama	Columbia	7,000
Congo-Brazzaville	Rwanda	5,000
Tanzania	Rwanda	5,000
Yemen	Sudan	5,000

conditions' (see Hamilton 1999 and Table 2). The 1999 report identifies more than 3.5 million such persons found within 28 countries. Some of these refugee-like people are stateless, denied the protection afforded by citizenship, e.g., Kurds in Syria or Tartars in Ukraine. Others are regarded by host governments as illegal aliens, or simply tolerated, or ignored (Hamilton 1999). Development-induced displacement is a final form of forced migration. Revolutionary governments in China and Ethiopia launched ambitious development programs that required resettlement of hundreds of thousands of people, some voluntarily, many not. Today in China, work continues on the largest dam resettlement project in history. Thus far, more than 1.2 million persons have been displaced in the wake of the Three Gorges Dam project.

2. *Future Directions*

At the end of the 1990s, the locus of the world's refugee population remained predominantly in the South, but 2.38 million of the world's 13.4 million refugees (18 percent) are settled in the First World. Although the total number of refugees has decreased since 1992, the number of countries that host asylum-seekers increased to 125. Many countries host refugees from several countries simultaneously; thus 161 different populations of exiled refugees exist worldwide. Palestinians (3.8 million) and Afghans (2.6 million) still constitute the two largest refugee communities (see Table 3). Iraq, Sierra Leone, Bosnia and Herzegovina, Somalia, Sudan, Eritrea, Liberia, Croatia, Angola, Burundi, Vietnam, and El Salvador are also important source countries. At least 250,000 citizens from each one have fled into exile. Meanwhile, migration policies are tightening in Europe and North America. Geo-policing of migration poses new challenges to refugees who desire asylum and permanent residence in fortress Europe. In addition, the United States has placed some of the most severe restrictions on asylum procedures in its history.

Contemporary refugee flows are subject to far more complexity than prevailed in the earlier European context of World War II. The case of the Kurdish population entrapped within Iraq initiated a new debate about human rights, national sovereignty, and new modalities for international intervention. Massive upheaval in Kosovo and NATO's unprecedented level

Table 3
Principal locations of international refugees and asylum seekers, as of early 1999

Country of asylum	Place of origin	Total number
Jordan	Palestinians	1,463,000
Iran	Afghanistan	1,400,000
Pakistan	Afghanistan	1,200,000
Russian Federation	Former USSR	1,000,000
Jordan	Palestinians	800,000
Gaza Strip	Palestinians	773,000
West Bank	Palestinians	555,000
Syria	Palestinians	365,800
Guinea	Sierra Leone	350,000
Iran	Iraq	531,000
Lebanon	Palestinians	364,600
Thailand	Burma	350,000
China	Vietnam	281,000
Bangladesh	Pakistan (Biharis)	238,000
Armenia	Azerbaijan	218,000
Sudan	Eritrea	320,000
Yugoslavia	Croatia	279,000
Tanzania	Burundi	260,000
United States	El Salvador	242,500
Yugoslavia	Bosnia–Herzegovina	201,000
Azerbaijan	Armenia	180,000
Ethiopia	Somalia	180,000
Uganda	Sudan	170,000
Guinea	Liberia	160,000
United States	Guatemala	145,500
Congo-Kinshasa	Angola	140,000
Thailand	Burma	140,000
Zambia	Angola	140,000
Russian Federation	Former USSR	129,500
Côte d'Ivoire	Liberia	125,000
Saudi Arabia	Palestinians	122,500
Liberia	Sierra Leone	120,000
India	China (Tibet)	110,000
India	Sri Lanka	110,000
Nepal	Bhutan	100,000
Germany	Bosnia–Herzegovina	100,000

of intervention intensified the debate further. 'Safe havens' for Iraqi Kurds, 'zones of tranquility' for returning Afghan refugees, 'open relief centers' for would-be Sri Lankan refugees, and 'safe corridors' to Muslim enclaves in Bosnia exemplify 'new' spaces and discourses associated with refugees (Hyndman 1999).

Social scientists face major challenges and opportunities on many fronts. Longitudinal studies are needed to trace the refugee experience and associated processes through the 'full circle' of flight, resettlement, repatriation, and reintegration. The fourth stage of the refugee cycle—reintegration of former refugees after their return home – is especially in need of analysis. The use of remote sensing is a promising new research technique that geographers have begun employing to document displacement and to monitor environmental degradation around refugee sites (see Lodhi et al. 1998 and Bjorgo 2000). 'Refugeeism,' diaspora experiences, ethnic identity, de-territorialization, post-colonial expressions of place, and sense of home are important refugee related themes that social theorists in geography are well equipped to address.

The era of discrete refugee definitions is over. The traditional distinction between the refugee and voluntary migrant continues to be eroded in the face of new human dramas and political controversies. One need look no farther than the recent case of a Cuban six-year old rescued after his mother died in search of asylum or that of Chinese stowaways in container ships bound for America.

See also: Diaspora; Ethnic Groups/Ethnicity: Historical Aspects; International Migration by Ethnic Germans; International Migration by Ethnic South and Southeast Asians; Migration into the Americas and between American Countries; Migration: Sociological Aspects; Migrations, Colonizations, and Diasporas in Archaeology; Refugees in Anthropology

Bibliography

Allen T, Morsink H (eds.) 1994 *When Refugees Go Home*. Africa World Press, Trenton, NJ

Bascom J 1993 The peasant economy of refugee resettlement in eastern Sudan. *Annals of the Association of American Geographers* **83**: 320–46

Bascom J 1998 *Losing Place: Refugee Populations and Rural Transformations in East Africa*. Berghahn Press, New York

Bennett J 1998 Forced migration within national borders: The IDP agenda. *Forced Migration Review* **1**: 4–6

Bjorgo E 2000 Using very high resolution multispectral sensor imagery to monitor refugee camps. *International Journal of Remote Sensing* **21**: 611–16

Black R 1991 Refugees and displaced persons: Geographical perspectives and research directions. *Progress in Human Geography* **15**: 281–98

Black R, Robinson V (eds.) 1993 *Geography and Refugees: Patterns and Processes of Change*. Belhaven Press, London

Bulcha M 1988 *Flight and Integration: Causes of Mass Exodus from Ethiopia and Problems of Integration in the Sudan*. Scandinavian Institute of African Studies, Uppsala, Sweden

Chambers R 1986 Hidden losers? The impact of rural refugee programs on poorer hosts. *International Migration Review* **20**: 245–63

Cohen R, Deng F (eds.) 1998 *The Forsaken People: Case Studies of the Internally Displaced*. The Brookings Institution, Washington, DC

Hamilton V 1999 *The 1999 World Refugee Survey*. US Committee for Refugees, Washington, DC

Hampton J (ed.) 1998 *Internally Displaced People: A Global Survey*. Earthscan, London

Harrell-Bond B 1986 *Imposing Aid: Emergency Assistance to Refugees*. Oxford University Press, London

Hyndman J 1999 A post-cold war geography of forced migration in Kenya and Somalia. *The Professional Geographer* **51**: 104–14

Kunz E 1973 The refugee in flight: Kinetic models and forms of displacement. *International Migration Review* **7**: 125–46

Lodhi M, Echavarria F, Keithley C 1998 Using remote sensing data to monitor land cover changes near Afghan refugee camps in northern Pakistan. *GeoCarto International* **13**: 33–40

Rogge J 1977 A geography of refugees: Some illustrations from Africa. *Professional Geographer* **29**: 186–9

Rogge J 1985 *Too Many, Too Long: Sudan's Twenty-year Refugee Dilemma*. Rowman and Littlefield, Totowa, NJ

Rogge J 1991 Repatriation of refugees: A not-so-simple 'optimum' solution. In: Allen T, Morsink H (eds.) *When Refugees Go Home*. Africa World Press, Trenton, NJ

UNHCR 1993 *The State of the World's Refugees 1993: The Challenge of Protection*. Penguin, New York

Wijbrandi J 1990 Organized and spontaneous settlement. In: Khulman T, Tieleman H (eds.) *Enduring Crisis: Refugee Problems in Eastern Sudan*. African Studies Centre, Lieden, The Netherlands

Wood W 1989 Long time coming: The repatriation of Afghan refugees. *Annals of the Association of American Geographers* **79**: 345–69

J. Bascom

Refugees in Anthropology

People being forced to leave their homes to settle in new environments is nothing new. However, 'refugees' as a legal category and as the object of academic research are more recent phenomena. This article will outline the development of refugee studies, in response to global refugee crises of the twentieth century. It will also present some key concepts and theoretical influences and discuss some of the challenges to anthropology of refugee studies.

1. The Emergence of 'Refugees'

The historical context in which refugees emerged as an internationally recognized category was provided by the massive displacements in Europe following two world wars. The Geneva Convention of 1951 is the critical landmark of the establishment of institutions and legal instruments to protect and assist such populations. A convention refugee is an individual who has crossed a national boundary and has a well-founded fear of persecution in the country of origin for reasons of race, religion, nationality, membership of a particular social group, or political opinion. Subsequently, other, regionally based legal conventions, notably in Africa and Latin America, have adopted more inclusive definitions of refugees which better correspond to the manifold types of mass displacements that often follow war and social disruptions in their regions today. Categories outside of legal provisions but equally of concern to agencies of protection are internally displaced people, people who flee within their own countries. Others are humanitarian refugees—like those offered temporary protection, these are persons who are not personally persecuted but are victims of war or war-related conditions. Another category of displaced people with important parallels to the experience of refugees, but without a global agency to protect their interests, are those forcibly relocated by their governments for the purpose of development. Such development-induced displacement may result, for instance, from the acquisition of land to make way for commercial forestry or the construction of a dam.

Refugee studies developed in response to the changing realities of the global refugee crisis in the early 1980s. The escalation of violent conflicts and mass disruptions in many areas of the world had then, within three years, dramatically doubled the global

refugee population. New increases followed the political changes at the ending of the Cold War. At the end of the twentieth century, large numbers of refugees from national conflicts and upheavals in postcolonial states and from crumbling multiethnic federations in Europe still dominated the asylum claims. According to estimates of the United Nations Commission for Refugees (UNHCR), there were 12 million recognized refugees in the world in 1999. Adding the millions of internally displaced people and asylum seekers (persons whose petitions for asylum are still pending), the figure was 19 million. The majority of refugees are found in the regions of the southern hemisphere, in particular in Africa and Asia, where many seek refuge in neighboring states. Only a trickle of displaced people make their way to the more affluent but more restrictive host countries of the north, in Europe or in North America. Even fewer are allowed to settle there.

Until the early 1980s, when refugee studies emerged as an academic field of study, with academic degrees and international journals, 'international migration' had been the main academic context in which research in the social sciences on refugees was usually found. The field is largely interdisciplinary, drawing together a range of research from the behavioral and social sciences, but with a strong presence since its inception of international law and anthropology. The interdisciplinary nature of the field has also meant the fruitful confrontation of anthropology with other theoretical and methodological approaches and the need to clarify the nature of its specific contribution. Furthermore, with its clear relevance for the acute issues of many refugee situations, much anthropological research in the field has a strong focus on practical application and advocacy. However, depicting the severing of a community and culture from their place, refugees have also implied a challenge to some basic theoretical assumptions in anthropology.

1.1 Who is a Refugee? Moving Beyond Legal Definitions

The answer to the question 'Who is a refugee?' is a crucial one in today's world. For those being classified, it may be a matter of life or death, because of the exclusive rights and entitlements of humanitarian concern and international protection that come with the status. In the academic field, the question is a continuing concern as a matter of theoretical and conceptual delineation of its object of study. The definition of 'refugees' as given in the Geneva Convention is too narrow, both as a protection instrument (many would argue), and as a theoretical concept. There is an agreement that, as a useful academic tool, the concept needs to take account of the many different situations which force people to leave their homes, and the proliferation of other legal statuses that have developed in different countries for people moving in 'refugee-like circumstances' (Escalona and Black 1995). Another basis of definition on more theoretical grounds has to do with the cause of flight. Today, causes are often complex, involving multiple factors and the blurring of distinctions, such as that between economic and political motives (for instance, economic collapse may often follow from political change). It is clear that, given the complexity, monocausal explanations are not very useful, because one status may blend into another. For instance, the Palestinian guest workers in Kuwait, once ousted from their homeland, were again displaced during the Gulf War in the early 1990s. Such explanations tend to be simplistic and obscure important similarities in the experiences of the victims of these forces (Shami 1994). A broader definition, well accepted in the field today and better attuned to current realities of displacement in the world, is to consider as refugees those people who are forced to abandon their countries of origin due to violent conflict or insupportable living conditions (Zolberg et al. 1989). These new realities have also questioned the relevance of differentiating between refugees and economic migrants. Rather than being seen as dichotomous, categories can be defined along a continuum, based on the degree of force involved. This would range from voluntary, such as labor migrants moving to better their situation, to involuntary, those forcibly expelled from their home areas (Hansen and Oliver-Smith 1982). While comparative approaches to kinds of migrations are increasingly emphasized by researchers in the field, so is the importance of maintaining that which is specific to refugees and other forced migrants, in terms of human rights and other humanitarian provisions made in their name (Shami 1994). From a somewhat different position, Elizabeth Colson makes a similar plea: 'Despite their similarities with other migrants, refugees have a unique status and force themselves upon the world's attention, because they emerge from and provoke acute crisis' (Colson 1987, p. 4).

The burning question of who and what a refugee is, is also a matter of 'the politics of space and otherness' (Gupta and Ferguson 1992). Answers are sought by examining the powerful institutional contexts and discourses, national and international, in which the category of 'refugee' is construed and public policy shaped, often in ways that misrepresent and exclude refugees. Further, eschewing overgeneralizations of 'the refugee experience' that have emerged out of such contexts (Malkki 1995a), anthropologists today would rather ask what it means to be a refugee, given the specific history of displaced groups and the circumstances in which they find themselves.

2. The Dynamics of Displacement

As refugees, people and their communities are not only forcibly removed, they are also remade. Forced

migration is an important aspect of social change, and any theoretical approach to refugees necessarily engages with process. While the study of refugees shares some common ground with international migration, it must also respond to the realities of sudden, involuntary change and the different forms of power which act on refugees' lives. For those displaced, change is often multifaceted and compressed in time, thus also providing a theoretical challenge to anthropologists, more used in the past to study change as slow and cumulative. However, emerging theoretical frameworks can draw insights from a growing body of research in anthropology on war and violent conflicts, social collapse and reconstruction. As displacement affects most aspects of people's lives, it addresses core issues in the anthropological study of communities, such as patterns of subsistence and livelihood, kinship and family, social and political organization, as well as the systems of cultural meaning which produce a sense of coherence and predictability in everyday life.

These issues are embedded in the wider dynamics of social and political change which form important contexts for the analysis of refugee communities. Within these contexts, anthropological case studies typically focus on uncovering the perspectives and strategies through which displaced groups deal with their predicament, renegotiating livelihoods, relations, and identities in interaction with their hosts. Anthropologists also examine the making of refugee communities, exile cultures, and identities in the places of settlement, including the dilemmas of those refugees whose attachments to the homeland and to fellow countrymen in other parts of the world remain stronger than to the society in which they reside (Fuglerud 1999). For a political movement such as the Chilean left, exiled in the 1970s, this may be manifested as a conflict between political ideology and the realities of everyday life, between 'life as it is' and 'life as it should be' (Eastmond 1997).

3. *The Diversity of Refugee Situations and Research Emphases*

The great diversity of conditions, in particular the marked contrast between southern and northern host states, has meant the development of different emphases in the concerns and theoretical approaches applied to the problem of refugees in different parts of the world.

3.1 *Northern Perspectives: Refugees in Resettlement*

Issues of resettlement and integration are common in studies of refugees in North America or Western Europe, reflecting the legal provisions for permanent residence of most refugees into these societies. Much research on resettlement in the USA, for instance, seems to focus on sociocultural adaptation, which is a conventional framework in the study of the immigrant experience. However, the analyses also draw fruitfully on current theories of ethnicity and identity, emphasizing agency and cultural construction in the process (e.g., DeVoe 1992). These concerns reveal that many exile communities are concerned with continuity but also that, despite their claims, exile cultures typically represent dynamic forms. In these fields, American anthropologists, like other Westerners in the profession, increasingly turn to study the institutions of their own societies and their interaction with refugees. Being a refugee often means something different in Western Europe, where large influxes of newcomers are relatively recent and where integration is often preceded by a long asylum-seeking process. An emphasis in much of the current research is the vicissitudes of integration in relation to policies and processes of social exclusion. Importantly, perspectives from the inside of diasporas, groups which maintain strong ties to the home country and to members in other parts of the world, indicate that full integration into the host country may not be a primary concern (Fuglerud 1999). However, more systematic and comparative studies of the experience of different groups in different countries and circumstances of reception in northern countries are called for (Escalona and Black 1995).

3.2 *Southern Frameworks: Refugees and Development*

In southern countries, particularly in Africa, displacement and refugees have frequently been studied as a problem of development. A general framework for analyzing displacement and involuntary resettlement with a focus on the responses of dislocated people and the role of the state in resettlement is provided by Hansen and Oliver-Smith (1982). This has been an important starting point for some of the comparative work in southern countries.

One theme is that of development schemes and modernization displacing large numbers of people within their countries. An early contribution from anthropology is Colson's (1971) study of the social consequences of the relocation of the Gwembe Tonga due to the building of the Kariba Dam in 1958. More recent analyses take account of power relations and local forms of resistance by those ousted from their home areas (McDowell 1996). Another theme is the critical evaluation of humanitarian assistance and the potential of more long-term development that also benefits the host population. Harrell-Bond's (1986) illuminating study analyzed the implications of relief operations on refugees and their hosts in Uganda. Neglecting refugees' own concerns and resources, the interventions of relief agencies were found to create conflicts and fragmentation both within refugee collectivities and with the local host population. Other critical studies have focused on camps as organized

solutions to refugee crises. Comparing the experience of camp-settled to self-settled Angolan refugees in Zambia over many years, Hansen (1992) found a preference for staying with kin even in meager and less secure circumstances, rather than submitting to the rigidly controlled but well-provisioned camps. While camp refugees remained strangers in the new country and wanted to return, most of the others did not. These studies highlight the problem of defining assistance in narrow economic terms without taking account of the perspectives and resources of refugees themselves. Malkki's ethnography (1995b) of displaced Hutu refugees from Burundi settled in Tanzania compared identity and historical memory of refugees in camps with those who had settled in a nearby town and relied on their own creative strategies for survival. Whereas the latter were pragmatically juggling social identities, in the diversity typical of many African towns, camp refugees were passionately concerned with creating their mythico-history as members of the Hutu nation.

4. *Refugees as People Out of Place: Challenges to Anthropology*

Camps, as a form of containment, also reveal something about the workings of power within the international refugee regime and the deeply cultural yet global notions of order which underpin it. In a world where nation-states are seen as the 'natural' basis of belonging, according to Malkki, refugees occupy an in-between or liminal position and as such challenge the dominant 'national order of things.' Drawing on theories of social classification and of the cultural constructions of nation-ness, such analyses of displacement have opened up new theoretical spaces of anthropological inquiry into the relationship between people and place, nationalism and history. Inquiries into the significance of place and the creative ways in which people come to see different places as 'home' also provide important theoretical frameworks for the study of repatriation of refugees. This applies both to international policy, in which going home is construed as the naturally desired end of 'the refugee cycle,' and to return as lived experience. Images of home, which are a central premise in much exile ideology and upheld in memories and claims to the homeland, cannot always be reconciled with the realities of life upon return (Black and Koser 1999).

Clearly, refugees present a challenge to nation-states, but they do so also to the social sciences and some of their cherished concepts. In anthropology, until recently, refugees and displacements were rather marginal as research themes, in interesting ways reflecting the position of displaced populations themselves. This neglect may be attributed to a theoretical bias in anthropology that has tended to downplay violence and suffering (Harrell-Bond and Voutira 1992). Other commentators, like Malkki (1995a) refer to a 'sedentary bias' in anthropology, meaning that communities and cultures have been assumed to be naturally located in a particular place. Studies of displacement thus require anthropologists to re-examine their concepts of culture and community as bounded and territorialized units. Also, with the growing theoretical interest in anthropology in globalization and transnational phenomena, nationalism and conflict, violence and social memory, the study of refugees and displacement is gaining new relevance and fresh theoretical insights. Taken together, these new challenges will contribute to the coming of age of the study of refugees as an academic field in anthropology with a more solid theoretical foundation. If, as it is often claimed, refugees are symptomatic of the state of the world, with global mobility, fluidity, and fragmentation, but also claims and allegiances to particular lands and cultural identities, then they may well occupy a central position in future anthropology.

See also: Diaspora; Globalization, Anthropology of; Human Rights, Anthropology of; Human Rights, History of; Human Rights: Political Aspects; Refugees: Geographical Aspects

Bibliography

Black R, Koser K 1999 *The End of the Refugee Cycle? Refugee Repatriation and Reconstruction*. Berghahn Books, Oxford, UK

Colson E 1971 *The Social Consequences of Resettlement*. Manchester University Press, Manchester, UK

Colson E 1987 Introduction: Migrants and their hosts. In: Colson E, Morgan S M (eds.) *People in Upheaval*. Center for Migration Studies, New York, pp. 11–16

DeVoe P 1992 *Selected Papers on Refugee Issues*. Committee on Refugee Issues (CORI), American Anthropological Association, Washington, DC

Eastmond M 1997 *The Dilemmas of Exile: Chilean Refugees in the United States*. Gothenburg Studies in Social Anthropology **13**, Gothenburg, Sweden

Escalona A, Black R 1995 Refugees in Western Europe: Bibliographic review and state of the art. *Journal of Refugee Studies* **8**(4): 364–89

Fuglerud O 1999 *Life on the Outside. The Tamil Diaspora and Long Distance Nationalism*. Pluto, London

Gupta A, Ferguson J 1992 Beyond 'culture': Space, identity and the politics of difference. *Cultural Anthropology* **7**(1): 6–23

Hansen A 1992 Some insights on African refugees. In: DeVoe P (ed.) *Selected Papers on Refugee Issues*. Committee on Refugee Issues (CORI), American Anthropological Association, Washington, DC, pp. 100–9

Hansen A, Oliver-Smith A (eds.) 1982 *Involuntary Migration and Resettlement: The Problems and Responses of Dislocated Peoples*. Westview Press, Boulder, CO

Harrell-Bond B E 1986 *Imposing Aid: Emergency Assistance to Refugees*. Oxford University Press, Oxford, UK

Harrell-Bond B E, Voutira E 1992 Anthropology and the study of refugees. *Anthropology Today* **8**(4): 6–10

Malkki L 1995a Refugees and exile: From 'refugee studies' to the national order of things. *Annual Review of Anthropology* **24**: 495–523
Malkki L 1995b *The Purity of Exile. Violence, Memory, and National Cosmology Among Hutu Refugees in Tanzania.* University of Chicago Press, Chicago
McDowell C 1996 *Understanding Impoverishment: The Consequences of Development-induced Displacement*. Berghahn Books, Oxford, UK
Shami S 1994 *Population Displacement and Resettlement: Development and Conflict in the Middle East*. Center for Migration Studies, New York
Zolberg A, Zuhrke A, Aguayo S 1989 *Escape from Violence, Conflict and the Refugee Crisis in the Developing World.* Oxford University Press, Oxford, UK

M. Eastmond

Region: Geographical

Region is a concept of great significance that has a long and storied history in geography. When the discipline was becoming an organized field of study during the late nineteenth and early twentieth centuries in Europe and North America, regions were its principal subject matter. Two conceptions of the geographical region vied for legitimacy. The region's importance was eclipsed during the 1960s by new systematic and quantitative approaches. Regional synthesis did not disappear, but ceased to provide geography's unifying framework. By the end of the twentieth century, geography lost confidence in spatial science and unifying concepts and frameworks of most sorts. The region was rehabilitated and enjoyed something of a vogue as geographers and others refocused attention on the nature and meaning of space and place.

1. *Regions for Everyone*

Region is a concept both familiar and useful to fields of knowledge ranging across the arts and sciences. Distinctive resource endowments, climatic activity, and plant and animal life, along with widespread patterns of human occupance all differentiate the earth's surface. That differentiation is the basis for identifying and naming regions: alpine forest; coal fields; wine country. Areas of the human body, too, are commonly referred to as regions; one of the most frequent physiological uses of region is in reference to the human brain, with distinct regions said to control discrete neurological functions. The essence of regions is the areal continuity (whether on the earth, the human body, or some other object of study, such as a painting) of some specified facts of interest to the observer (whether it be agricultural practices, digestion, or brush stroke). Regions of all sorts are identified on the basis of areally associated common characteristics.

Despite its widespread parlance, region enjoys a special relevance in geography because of the concept's dictionary definition of and popular association with the earth's surface. Regions, *ipso facto*, are geographical regions. As a field of knowledge widely perceived to study the earth's natural and human variation, geography is similarly understood to rightly 'own' regions as its logical and meaningful object of study. Geography certainly has devoted itself to the continuing transformation and differentiation of global nature and society. Yet the discipline has had a problematic relationship with the region: an on-again, off-again love affair with its putative object of study.

2. *The Region in Geography*

The region has been such a troubled concept owing to geography's hybrid status as, at once, a natural and social science. When the great disciplines were borne during the late nineteenth century, geography came into being with its shoulders rubbing kindred disciplines of geology and anthropology. Geology laid claim to the scientific study of the origin, history, and structure of the earth, while anthropology claimed an equally scientific status for its study of the origin, cultural development, and behavior of humans.

Geography was the study of the earth and life on earth, especially human life and its capacity to change the face of the earth. More than a little geology, a little anthropology, geography on the continent, British Isles and in North America identified itself as the study of the relationships and interactions between nature and culture. Halford Mackinder, a founder of British academic geography and great champion of the regional approach, lamented in 1887 'one of the greatest of all gaps … between the natural sciences and the study of humanity.' He went on to insist that it was 'the duty of the geographer to build one bridge over [this] abyss …' (cited in Livingstone 1992, p. 190).

This early history of the discipline bore direct consequences for the conceptualization of regions, for what better place, quite literally, to examine such connections than in areas ostensibly characterized by common natural and human features—which is to say, than in regions? Geography neither invented the region nor wrested it away from a rival discipline. Rather, the region fell into the discipline's lap as the obvious focus for its pursuits.

The study of regions seemed a natural fit with the new discipline of geography not merely by default, however. Geographers attracted to the study of what were then called 'man–land' relations tended to focus, if not lavish, their attention on rural areas in an often deliberate rural bias. Well into the twentieth century, most people around the world lived in rural areas or small towns and to some extent, surely, geography was

simply following demography in its preoccupation with the countryside. But more willfull considerations were also at play. The countryside offered an ostensibly more transparent scrim for observing interactions between nature and culture: it was easier to parse a relationship between soils and climate on the one hand, and cultivation, settlement, and even religious practices on the other in the countryside than to try to make analogous sense of the nature–culture nexus in burgeoning cities of the day. Livingstone (1992, p. 264–6) further suggests that geographers seeking a moral compass for their work deliberately sought out rural areas because of the perceived superiority of rural values in a rapidly changing world. Anti-urbanism was a conscious strategy, often fraught with political overtones, as was the case with regional geography in Germany and France. From the start, regional and rural geography were synonymous.

Early in the twentieth century, when geography was being established as a university discipline in Western Europe and the United States, the study of regions took place in service to geography's larger purpose of bridging the natural and human sciences. Regions were thought to disclose nature–culture relations with quintessential clarity and were a basis for the most hotly contested debates of the age, namely over the relative weights of nature and culture in determining the character of people and places. Regional geography supported a variety of positions about the degree and direction of what was somewhat delicately called environmental influence. Taking cues from a stringent reading of Darwin's principles, Huntington (1924) found evidence of a determining environmental influence on the fate of world regions through the instrumentality of race and intelligence, a determinist position staked out earlier in the influential work of German geographer Ratzel (1896–8). Elsewhere, a more nuanced approach was developed in France by Vidal de la Blache (1926). 'To him, the regional articulation of life-style consisted in the material expressions of human–land relationships' (Livingstone 1992, p. 267) which he explored in a series of essays about French *pays*. Vidal emphasized society's variable role in modifying nature and was concerned with investigating how different natural realms gave rise to what he called *genres de vie*, or different ways of living.

2.1 The Region is Geography

A subtle shift took place in the early twentieth century in the status accorded geographical regions. Debates over environmental influence waned with the inevitable rejection of crass determinism and the emergence of a possibilist consensus built around *Lamarkian* notions of human adaptation as an active rather than passive response to nature. Under these changing conditions, regions were no longer the mere handmaidens of research but came to be the discipline's *raison d'etre*. Regions held a newly intrinsic interest. In Germany, Alfred Hettner promulgated '… the principle of *chorology*—the explanatory investigation of terrestrial reality divided into a series of component regions—' as the only way 'geography could retain both its physical and human components because it was precisely in their dialectic that particular places acquired regional character' (Livingstone 1992, p. 263).

In the United States, two very different regional concepts competed for acceptance and legitimacy. The first was contained in the work of Sauer and the so-called Berkeley school of (cultural) geography. Borrowing from Germanic anthropology, Sauer translated the anthropologist (and Berkeley colleague) Alfred Kroeber's 'culture area' into the geographers' 'cultural landscape'—synonymous with then-contemporary meanings of region. In Sauer's conception, regions existed as organic wholes, awaiting discovery and interpretation through intensive geographical field work. Also presumed was the inherent uniqueness of regions in terms of the presence and combination of natural and cultural features. The Sauerian tradition rejected grand theorizing in favor of a particularism that found perfect expression in richly textured and seemingly exhaustive cultural landscape studies (Sauer 1963).

The rival regional concept was associated with areal differentiation as enunciated by Hartshorne in his exegetical *The Nature of Geography*, published in 1939. *Contra*-Sauerian particularism, Hartshorne rejected the heretofore common notion that regions existed as organic wholes. Regions might be observed and perhaps even experienced in a harmonious, holistic fashion, but otherwise such holism was 'not inherent in the world which the geographer studies — neither in the world of nature nor in the actual world which nature and man together have made' (Hartshorne 1961, p. 362). While denying regions an independent ontological status, Hartshorne sought to guide their study on to firmer ground. '(F)or him, the regional project consisted in determining how "particular elements and complexes of elements are related to those in others"' (Livingstone 1992, p. 310)—such particulars including relief, vegetation, climate, land use, etc. Hartshorne called for a formal process of *areal differentiation* to define geographical regions based on one or another consciously chosen geographical particular. Such a scheme would lay the systematic basis for areal classification (e.g., the agricultural/climatic/religious/botanical regions of the world), comparison, and mapping.

Until the mid-twentieth century, these two regional concepts uncomfortably co-existed with each other. 'In the first (Hartshornian) view, regions were a form of areal classification, and in the second (Sauerian), regions were relatively unique areas of the earth that possessed a geographic individuality' (En-

trikin 1991, p. 16). Hartshornian regions enjoyed a wider purchase in the discipline because they insinuated themselves easily in the discipline's burgeoning systematic subfields of cultural, social, and economic geography.

By all accounts, during the 1940s and 1930s regional geography turned formulaic, with material synthesized and presented in a numbingly predictable format, beginning with the physical environment (landforms and climate) and then moving through several layers, as it were, of human occupance (settlement patterns, house types, farming practices, barn and fence types, etc.). While some scholarship rose above this narrow groove, sadly much did not. '(A)s clearly illustrated by many of the regional texts that appeared during those decades, the result was a rather tedious catalogue of factual material within an uninspiring framework' (Johnson 1991, p. 40). Like an overly tended hothouse flower, regional geography withered at the very moment it enjoyed the discipline's central place.

2.2 *The Region is Dead! Long Live the Region!*

By the 1960s, many geographers openly questioned whether the region could any longer provide an integrative framework of study. The criticism leveled at regional geography was often upsparing, as when Peter Gould characterized it as 'shabby, parochial, and unintelligent' (cited in Johnston 1991, p. 40) and easily caricatured. Gould's comments contributed to a wider assault on regional geography by a new generation of geographers seeking to redefine the discipline as a spatial science. Through the very real weaknesses of the traditional regional approach, and the sheer hubris of geography's so-called young turks (such as Gould), spatial analysis decisively eclipsed regional synthesis as the discipline's heart and soul. Suddenly the strenuous disagreements about particularism and areal classification sounded oddly antiquarian compared with the clarion and oh-so-modern call of the quantitative revolution.

Amidst this upheaval, regional geography did not vanish. Quantitative methods lent the Hartshornian region something of a new lease on life. Those working in the Sauerian tradition simply ignored the spatial science gold rush and continued advocating their cultural landscape approach in what became something of a lively, though curious, intellectual backwater. But the region generated little in the way of theoretical or methodological debate until the late 1970s. By that time, the euphoria surrounding spatial science had run its course and quantification was finally understood as a tool, albeit a powerful one, rather than geography's *raison d'être*. A new methodological pluralism overtook geography and the region once again began inspiring fierce discussion and debate—the news of its death having been vastly overstated.

The 1980s witnessed a resurgence of interest in regional geography emanating from two disparate sources (Pudup 1988). The first was practitioners of traditional regional geography who shared an unwavering belief that geography's *modus vivendi* remained regional synthesis. Emblematic of the revivified tradition was the election of two of its chief contemporary practitioners, John Fraser Hart and Peirce Lewis, to the Presidency of the Association of American Geographers (AAG) during the 1980s. Both used their AAG presidential addresses to celebrate traditional regional approaches. Hart called regional geography 'the highest form of the geographer's art' and the discipline's long neglected unifying theme: 'I know of no other theme that is even remotely so satisfying as the idea of the region. All of the different strands of geography converge when we try to understand regions, and we need the concept of the region in order to understand why we need the diverse and variegated systematic sub-fields of geography.' (Hart 1982, p. 18). Three years later, Peirce Lewis echoed Hart's sentiment in another spirited defense of synthetic description which he considered 'a kind of rough and ready definition of geography: describing the earth's surface and trying to make sense of it.' (Lewis 1985, p. 47).

Such 'back to basics' calls could be dismissed as pure nostalgia were it not for the equally spirited discussions about regions arising from a very different quarter—geographers who sought to rehabilitate the concept and create a 'reconstructed regional geography.' The new regional scholarship was pioneered, perhaps unsurprisingly, by a generational changing of the guard. Skeptical about both spatial science and regional synthesis, geographers coming of age in the 1980s brought to the study of regions a deep engagement with theoretical debates and philosophical discussions ranging across the social sciences (Thrift 1983). They also brought to the study of regions a consciousness borne from being living witnesses to dramatic changes wrought by globalization (Massey 1995, Allen et al. 1998, Storper 1997).

The study of regions underwent a profound transformation. Regions ceased to be a topic interesting chiefly to geographers and historians and became an object of fascination in mainstream social sciences. Economists, political scientists, sociologists and, of course, geographers began asserting 'that the region might be a fundamental basis of economic and social life "after mass production."' That is, since new successful forms of production—different from the canonical mass production systems of the postwar period—were emerging in some regions and not others, and since they seemed to involve both localization and regional differences and specificities (institutional, technological), it followed that there might be something fundamental that linked late 20th century capitalism to regionalism and regionalization' (Storper 1997, p. 3). The apparent *fin de siècle* tendency

toward regionalization of economic and social life, in places like 'the third Italy,' Baden-Wurttemburg, and Silicon Valley, was wildly at odds with post-World War II era expectations of global integration and homogeneity. The global village seemed to be giving way to a world of regions.

3. The Regional World

Reconstructed regional geography radically departs from traditional regional synthesis. Gone are assumptions of organic holism and naturalism, or that geographers take unmediated cues from the physical environment in discovering neatly packaged regions. Rural bias has given way to a plurality of spatial possibilities, including rural and urban but also sub- and ex-urban regions, as in the case of 'edge cities' like Orange County, California. Economic processes of trade and investment are accorded far greater importance in regional transformation, if not central importance, in contrast to traditional regional geography's emphasis on culture and its tendency to either ignore or collapse most aspects of social life into the material expressions of culture. Indeed social relations, understood as 'always inevitably spatialized' (Allen et al. 1998, p. 138) take on an unprecedented significance in the new regional geography. Regions are socially constructed and open to deconstruction through social relations both in- and outside the region. The brave new regional world is an unstable one: '… that process of construction is constantly evolving. The spatialities of our lives are the product of continual negotiation, the outcome of the articulation of differentially powerful social relations…. Any settlement of social relations into a spatial form is likely to be temporary. Some settlements will be longer lasting than others' (Allen et al. 1998, p. 138). Finally, regions are also discursively produced through cultural representations that imbue meanings differentiating one region from another, e.g., a 'core' vs. a 'peripheral' region. 'What is said' about a region can powerfully shape the form and content of its social relations and what literally takes place there. Another obvious change is the new regional geography's nomenclature. Perhaps because 'region' carries such bruised and battered baggage, concepts like space, place, locality, and even territory are typically used interchangeably with region. The long-standing question of the region's appropriate geographical scale appears moot.

But not all the hoary questions have gone away. As the twentieth century turns into the twenty-first, geographers continue to wrestle with how to understand regional difference and uniqueness. That places differ from one another—and not just in their landforms and climate—can no longer be assumed, as was the earlier luxury, but has to be explained in a world governed by recognizably systemic processes of global capitalism. In this regard, regions continue to be human places where geographers explore relations between the general and the particular. And because such questions by definition defy agreement, the regional enterprise is guaranteed a long and storied future.

See also: Cultural Geography; Geography; Globalization: Geographical Aspects; Local Economic Development; Local–Global Linkages: Environmental Aspects; Regional Geography; Rural Geography; Urban Geography

Bibliography

Allen J, Massey D 1998 *Rethinking the Region*. Routledge, London

Entrikin J N 1991 *The Betweenness of Place*. Johns Hopkins University Press, Baltimore, MD

Hart J F 1982 The highest form of the geographer's art. Presidential address. *Annals of the Association of American Geographers* **72**: 1–29

Hartshorne R 1939 The nature of geography. *Annals of the Association of American Geographers* **29**: 1–469 (1961 *The Nature of Geography*. Edwards, Ann Arbor MI)

Huntington E 1924 *The Character of Races, as Influenced by Physical Environment. Natural Selection and Historical Development*. C Scribners Sons, New York

Johnson R J 1991 *A Question of Place*. Blackwell, Oxford, UK

Lewis P F 1985 Beyond description. *Annals of the Association of American Geographers* **75**: 465–78

Livingstone D N 1993 *The Geographical Tradition*. Blackwell, Oxford, UK

Massey D 1995 *Spatial Divisions of Labour: Social Structures and the Geography of Production*, 2nd edn. Routledge, New York

Pudup M 1988 Arguments within regional geography. *Progress in Human Geography* **12**: 369–90

Ratzel F 1896–8 *The History of Mankind*. Macmillan, London

Sauer C O 1963 *Land and Life: A Selection from the Writings of Carl Ortwin Sauer*. University of California Press, Berkeley, CA

Storper M 1997 *The Regional World*. Guilford Press, New York

Thrift N 1983 On the determination of social action in space and time. *Society and Space* **1**: 23–57

Vidal de la Blache P 1926 *Principles of Human Geography*. Lowe and Brydon, London

M. B. Pudup

Regional Geography

1. Introduction

Regional geography is a field of study which strongly contrasts with systematic geography. While the latter is concerned with the distribution of certain characteristics, such as landforms, climate, vegetation, and

human activities, over the surface of the earth, regional geography looks into the relationships these phenomena have within the different regions. The prime interest in the systematic fields of geography is to describe the patterns certain phenomena display over the entire globe and to elucidate the (general) rules that govern these patterns. Regional geography, on the other hand—at least in the mainstream thinking of the nineteenth and early twentieth centuries—is not interested in general rules but, on the contrary, in the singularity or the specificity of regions. Thus, traditional regional geography is characterized by a synthetic view of the world, i.e., it tries to put things together in a regional context, and by the (epistemological) conviction that it needs to view the world as a mosaic of idiosyncratic, singular entities—the regions.

This understanding is deeply rooted in most people's everyday consciousness of the ways in which the world is structured. Thus, when we think of the American Midwest we conjure up images of commercial agriculture, endless cornfields, livestock production, and a predominance of white middle-class Protestants, whereas when we think of the Sahara or Central Australia we probably imagine little vegetation, heat, and dryness, and an environment hostile to most life-forms. The fact that people tend to be regional geographers in their everyday consciousness may be attributable to the fact that most of them were taught geography at school. It can also be seen as a strategy of the individual to simplify a complex world by structuring it along the lines of regional entities. Whichever the reason, the fact that we do tend to think in idiosyncratic ways helps to explain the lasting success of regional geography—at least in the form of regional monographs. In our age of extensive travel over the entire globe these monographs are useful backdrops for our encounters with strange environments, since they allow us quickly to attune to (rapidly) varying circumstances.

As an academic endeavor, determining geographers' shared conviction as to what constitutes the essence of the discipline, regional geography was not a lasting success story. However, after a period of strong rejection in the 1960s and 1970s there were signs of a renaissance in the last decades of the twentieth century. The following sections of this entry deal with the mixed press regional geography encountered in the past, with recent developments, and with the—possible—future(s).

2. *Historical Development*

Although regional geography can be traced back to ancient Greece, the 'classical' period was that of the late nineteenth and early twentieth centuries. The external factors explaining the success of geography as a discipline and of regional geography as its center at the time were the rapid exploration (and exploitation) of the interior of continents, the setting up of a regular or more efficient administration in many countries, and the stronger ties that were forged between countries and places. All this necessitated the generation and dissemination of (geographical) knowledge, much of which was of a place-specific nature. Before geography was set up as a university discipline later in the nineteenth century, naturalists such as Alexander von Humboldt (together with Aimé Bonpland) ventured to study hitherto unexplored parts of the world. Von Humboldt supported a holistic understanding of (his) academic explorations: in his extensive research papers on South America, which covered more than 9,000 pages, for instance, natural and human phenomena were seen in context.

Thus, regional geography is associated with the early beginnings of geography as a university discipline and it was, in the case of von Humboldt, informed by the Enlightenment, as von Humboldt concerned himself critically with the situation in the South American colonies. However, later in the nineteenth century a holistic understanding of the (natural) sciences became increasingly marginal because they were undergoing a process of continuing specialization. In geography, on the other hand, the holistic approach experienced a sustained success as it influenced much of the mainstream methodological thinking well into the twentieth century.

In the German-speaking countries the Austrian Viktor Kraft (1929) offered his very influential epistemological analysis of the object, the aims, and the methods of geography. Following a neo-Kantian line of argument Kraft subdivided the discipline into 'general,' i.e., systematic, geography (as a nomothetic science), and regional geography ('Länderkunde'—as an idiographic science). In line with this view, it was maintained that the study of regions is the prime object of geography (an object truly exclusive to the discipline), that therefore regional geography is at the core of the discipline, and that, consequently, systematic geography and its branches are ancillary sciences. Moreover, regional geography had an integrating role to play as it incorporated physical as well as human geography under one epistemological roof. The fact that regional geography thus gave unity and an exclusive object to geography explains why it remained the hegemonic methodological discourse well into the twentieth century.

In Germany, Hettner (1927), amongst others, had successfully codified regional geography along the lines of Kraft's neo-Kantian epistemological reflections so that it occupied a pivotal position within geography from the 1920s onwards. In the USA, Richard Hartshorne's *The Nature of Geography* (1939) gave a similar importance to regional geography. The view that regional geography had an overarching, integrating role to play, however, came under in-

creasing attack in the 1950s and 1960s when the spread of an analytical-pragmatic and nomothetic paradigm reached the community of geographers.

In essence, practitioners of regional geography were accused of unscientific working practices, of gathering—and not analyzing—facts, and of emphasizing the 'individuality' of spatial units, the regions. The critics whose discourse on the philosophy of science was informed by logical positivism (cf. Harvey 1969) and critical rationalism pointed out that regional geography was lacking a theoretical foundation on which an explanation of empirical evidence could be grounded. Part of the criticism was directed at what the critics regarded to be a premodern frame of mind of regional geographers, who saw the world as a well-ordered mosaic (of regions) which the regional geographer's 'art' had to put together, at the implicit geo-determinism and at the exclusion of social practices from academic work.

By the 1960s and early 1970s a nomothetic discourse had asserted itself in geography which was to form the mainstream methodological underpinning of the discipline for the next decades (Hanson 1999). As a consequence, from then on regional geography only played a minor role until the last decades of the twentieth century. This minor role, however, mainly expressed itself in the low esteem that regional geography enjoyed in the eyes of those geographers whose frame of mind was shaped by an understanding of academic work deeply rooted in logical positivism and critical rationalism. This state of affairs was sharply contrasted with the fact that at the same time many regional studies were being published (Gregory 1988). This is why it is not possible to infer from the (then) dominant paradigmatic understanding of the discipline the performance and the relevance of regional geography in the wider society.

The peripheral position of regional geography in academia during the 1970s did not change, despite the attempts of individuals to put it on a (new) epistemological foundation. In Germany, Wirth (1978), for example, in his endeavor to resurrect regional geography takes up certain key elements of critical rationalism as developed by Popper. According to Wirth, regional geography should be treated as a historical science which deduces singular events from general laws (which in the case of regional geography can be taken from the systematic fields of geography) and special antecedent conditions (which result from the unique situation under investigation). This approach will not only render a description but also provide the academic with an *explanation* of empirical phenomena. This view challenged other academics, notably Bahrenberg (1979). The dispute that evolved between the two antagonists did not, however, have a widespread impact. The fact that a broad discussion on the significance of regional geography did not take place at the time may have its roots in the repercussions of the 'quantitative and theoretical revolution' of the 1960s and 1970s, as well as in the growing inner diversity of the discipline as a whole.

3. *Recent Trends: The Emergence of a 'New Regional Geography'*

In the 1980s there was a development, primarily taking place in the French- and English-speaking countries and in The Netherlands, which can be characterized as a trend back to the study of the specific, a renaissance of regional (human) geography (Gilbert 1988, Pudup 1988). This tendency can be seen as a reflection of a number of developments.

3.1 *The Background*

To begin with, a positivist mode of explanation in general and spatial analysis in particular were seen to be insufficient tools for geographical studies. The criticism leveled at the mainstream geography of the 1960s and 1970s was that it neglected humans as intentionally acting beings ('structure–agency–debate') and that it tended to reduce space to an entirely social construct ('spatial exorcism'). Second, 'space' and 'place'—key geographical concepts—gained in importance in the neighboring disciplines. This had repercussions on geography. Now geographers were able to point out that distinguished scholars, such as the sociologist A. Giddens, were borrowing from geography. And third, in the light of the crisis of the 'grand narratives' of humankind, such as religion, communism, science, etc. (and of 'grand theories' in the academic world), in the age of postmodernity there was an overall social trend which gave more importance to the special/specific (idiographic), and which acknowledged a greater diversity in regional cultures, ways of life, etc. Hence, a totalizing view as to the ways in which societies, economies, cultures, etc. should be seen and understood came under increasing pressure.

Regional specificity had reinstated itself on the methodological agenda because it was recognized by a growing number of academics that, as Massey (1984, p. 120) points out,

> just as there are no purely spatial processes, neither are there any non-spatial social processes. Nothing much happens, bar angels dancing, on the head of a pin. … The fact that processes take place over space, the facts of distance and closeness, of geographical variation between areas, of the individual character and meaning of specific places and regions—all these are essentials to the operation of social processes themselves.

3.2 *The Different Approaches*

According to Gilbert (1988) there are at least three different approaches dealing with regional specificity

that can be distinguished in the recent endeavor to promote a 'new regional geography': the political–economic, the humanistic–phenomenological, and the structurationist.

The political–economic approach is very much inspired by the tradition of dialectical materialism in Marxist geography. It focuses on the ways in which (global) mechanisms of capitalism take effect on the ground, in particular on the ways in which the division of labor is spatially structured in capitalist societies. This understanding of the spatial operation of capitalism is in essence based on two main premises: it rejects the notion of a linear realization of systemic processes in different spatial contexts ('regions') and it assumes a dialectical relationship between the systemic framework (of economic development) and the locally and regionally specific development trajectories. Hence, regions are not just spatial backdrops against which overall developments take place, they are the specific sociospatial, historically contingent context ('layer of history,' Massey 1984) which determines the specific pathways of capitalist development in the varying particular and the overall contexts.

According to Gilbert (1988, p. 219) the term 'region' is given a substantially different meaning and therefore a 'new regional geography' differs markedly from the traditional understanding: 'The traditional concern with the people/nature relation has widened to include society as the prime agent in region formation. Thus, the substance of regional geography has become the triangular relations between people, society and nature.'

The humanistic–phenomenological approach can best be understood as a multilevel change of perspective. It is a break with the established positivist methodology and it marks a different approach of the academic toward social reality. In their critique, Van der Laan and Piersma (1982, p. 416f.) emphasize the reductionism of quantitative geography which in essence meant that human beings were treated as 'things, as non-intentional materialistic research objects.' As a consequence—and in direct opposition to the positivist notion of *explanation*—the proponents of a humanistic approach place the intentionally acting human being at the center of academic inquiry. Furthermore, academics are to discard their 'classical' perspective as outsiders and to assume an insider's view of social reality in the sense of comprehending, *understanding* the meaning of social acts and the intentions of agents (Buttimer 1979). Only through this shift will social practice be rendered intelligible.

The humanistic approach looks into the cultural meaning of the 'region' which is being conceived of as the expression of a specific—culturally embedded—relationship between people and their spatial environment. The distinguishing quality of this relationship is the awareness of a shared culture among the population, and the associated distinct identity of that population as it is noticeable to outsiders. This relationship also encompasses the resultant feeling within the population of belonging to 'their' region.

According to the programmatic considerations of its proponents, a humanistic–phenomenological approach has as its focal point individuals, their everyday view of the world, and the meaning which they attribute to their environment. However, the individual is not seen as a solitary agent but as a carrier of a shared cultural identity.

The most striking characteristic of the structurationist perspective, which is mainly based on Giddens's (1988) theory of structuration, is the dialectical understanding of the 'region' as the outcome and the producer of social practice. The 'region' is regarded as a specific spatial and temporal constellation of social relationships, and at the same time as a medium for these social relations. Pohl (1993) very aptly speaks of a 'co-reacting catalyst.' Behind this view is a comprehensive definition of the 'cultural region' in which elements of the political–economic approach are also incorporated.

3.3 Their Conceptual Connections

These different conceptional approaches do not in any way form coherent edifices of knowledge, nor are they as clearly demarcated as these few observations may suggest. Additionally, they are not at all canonized by the scientific community of geographers, particularly outside the French- and English-speaking countries. Notwithstanding these qualifications, there are a number of issues which link the different approaches on a conceptional level: it is the understanding of the term 'region' and of the central epistemological dimensions 'individual' and 'history.'

One of the outstanding arguments in the discussion is the ontological status being ascribed to the 'region.' According to a number of proponents of a new regional geography, regions are not only empirically important—as analytically relevant constructions of the academic—they are also the outcome of human history and therefore of an ontological nature (Pickles 1986). This is why the individual and history are at the core of a new, reconstituted regional geography. Another point which many advocates of a new regional geography agree with is the assumption that 'regions' are a persistent feature of societies. Contrary to the modernist conviction that regional differences will be eroded in the process of social modernization, these differences are assumed to remain, as spatial differentiation is seen as a central element of society (in the sense of social relations being spatially and temporally contingent phenomena). Such a notion of the 'region,' however, has little in common with the chorological view according to which the 'region' is an expression of the opposition against (or the inertia inherent in) the homogenizing trends of modernization. The third notion which runs through the debate

is the structural analogy of the three approaches. What is meant here is the mentioned dialectical concept of the regions being both the result and the 'medium' of social relations. As a co-reacting catalyst, the region has emergent powers as 'social relations within the region develop because of the specific way individuals and groups relate within the particular regional space' (Gilbert 1988, p. 215).

One distinctive feature of the debate about a new regional geography is the fact that it has a heavy theoretical and conceptual slant, whereas empirical studies are scarce. The existent empirical work does not exhibit a uniform approach toward method, but rather a certain methodical pluralism or voluntarism. There are basically two empirical approaches. Whereas one group of academics works historically by reinterpreting regional developments which have taken place (cf. for instance Gregory's outstanding work *Regional Transformation and Industrial Revolution: A Geography of the Yorkshire Woollen Industry*, 1982), the focus of the other group ('locality research'; cf. Cooke 1989, Harloe et al. 1990) lies on the present and on the examination of the consequences of (global) economic restructuring on 'localities.'

3.4 Unresolved Issues

After a fairly intensive (conceptual) debate in the 1980s and early 1990s the interest in a reconstituted regional geography waned somewhat in the late 1990s. The reasons behind this trend may have been a certain fatigue after a period of intensive debate, the fact that many geographers who have migrated to other fields of interest in the discipline do have strong reservations about (prejudices against) 'regional geography,' and the problematic or unresolved issues of the debate. All three are inter-related in various ways, and they need to be taken seriously if regional geography is to have a future as a stimulating, academically rewarding pursuit.

The interest of geographers—like that of other academics—usually is arrested by methodological debates which have the potential to renew the cognitive and the social system of which academics are members. However, fatigue sets in if these debates fail to live up to the expectations of the scientific community. In the case of a new regional geography this may be due to the fact that the promise of modernization was not convincing enough, both to those who took part and to the grant-giving bodies (who are operating on a system of academic self-control; thus colleagues with reservations about regional geography decide on submitted research proposals). As the new methodological direction is characterized not by one but by several approaches, it is hard to find sufficient collective support within the scientific community of geographers. What is more, despite their innovative and stimulating new ideas, these approaches have inherent shortcomings which weaken the case for a new regional geography.

To begin with, there is, typical for a 'young' paradigm, the need to clarify the usage and meaning of key concepts and terms. This is particularly so in the case of the 'locality' debate, in which the usage of the term 'locality' leaves much to be desired. Most importantly, the usage of that term reproduces the casual handling of concepts that was typical of the 'old school' and that had received so much—deserved—criticism. Terminological constructs are taken as starting points without confronting these with the terms relevant for both the life world ('Lebenswelt') and the social systems. It was not the process and the outcome of the socially constructed spatial coding which was on the agenda but rather the examination of social interaction within the boundaries of the previously demarcated study areas ('localities'). This, however, bears the danger of reifying (predetermined) spatial concepts and thus of wasting a chance for epistemological progress.

This problem refers to another unresolved issue: the inadequate interaction between theory and empirical evidence. In the case of the locality studies, this becomes apparent due to the exclusive focus on the level of the locality and the exclusion of national or global economic development trajectories and their analytical framework. Because a new regional geography—rightly—assumes that phenomena cannot be explained by themselves, a stronger effort needs to be made to interrelate empirical description and theoretical abstraction.

The last point to be discussed here is as fundamental as the previous issues. What are the implications of a regional geography which is conceived of as a social science, thus relinquishing the former claim of giving unity to geography (by integrating physical and human geography)? Will a social scientific understanding of regional geography therefore lead to a loss of disciplinary unity? There can be little doubt that this issue is an important one as geographers have to compete with academics from other disciplines for funds, etc. Showing disciplinary unity can help to underscore the importance of the discipline and thereby demonstrate the importance of academic subjects as social systems which can effectively organize funds, support, and the like. However, as a cognitive system, geography has to approach the question of disciplinary unity via a different route. Unity is no value in itself, of course, so the issue at stake is whether there are social developments which suggest a holistic approach. One key area to be dealt with in such a way is the crisis of late modern societies effected by ecological and other human-made problems. Because of the complexity of these issues a comprehensive approach has to be taken in order to resolve the problems involved. Serious as these matters may be, they offer a chance for academics to cooperate

in various ways: amongst colleagues of the same subject, on an interdisciplinary level, and with the wider public. Only a concerted effort can help to come to grips with the numerous dangers inherent in the modern world we live in and which we helped to create.

4. *The Prospects of Regional Geography*

The future of geography as a whole and of regional geography as one of its branches will very much hinge on the ability of geographers to contribute to these and other socially relevant issues. Such an argument is based on pragmatic and normative considerations and not, as was the case with traditional regional geography, on the methodological neo-Kantian assumptions outlined above (the latter, however, having the normative implication as to how the subject, the methods, etc. of regional geography should be seen).

After the paradigmatic shifts in the twentieth century which first brought traditional regional geography into disrepute, and then questioned the hegemony of a positivist mode of explanation ('spatial analysis'), we now live in an era where the grand, all-encompassing theories have lost their authority and where different analytical frameworks (methodologies) coexist. It has become clear that there is not *one* way of coming to grips with the complexity academics (have to) deal with. It is understandable that the proponents of a positivist mode of explanation, in their wish to be accepted as equals by their colleagues from other (positivist) disciplines, broke with the regional tradition in the 1950s and 1960s. However, from the vantage point of the early twenty-first century, this development can now be seen as a certain stage in the development of the discipline which has evolved as the disciplinary discourse progressed. As Hanson (1999, p. 136) points out, there is now

> a greater willingness to acknowledge complexity and to engage multiple approaches to understand that complexity. There is now a greater acceptance of the notion that people create knowledge rather than discovering objective truths … and that taking advantage of what is useful in each of a number of approaches can yield greater understanding than does adhering only to one.

What are the implications of these considerations for the future of regional geography?

Taking up the line of argument presented by Hanson, an enlightened methodological understanding of a modern regional geography cannot afford to dismiss lightly any conceptual framework, as different perspectives may yield different facets of knowledge. If regional geography is to be taken seriously as a competitive academic subject, it needs to engage in the methodically controlled generation of knowledge, and it needs to do that by probing the relevance of different methodological approaches. This is no easy task as different understandings of what science is (or should be) about rest upon different epistemological and methodological assumptions, for instance on contrasting concepts of knowledge or causality. But the options are much less attractive. Additionally, it needs to be pointed out that along with such a postmodern methodological understanding comes the inherent danger of warranting an 'anything goes' attitude (including a regional geography without or with implicit theory). A competitive regional geography, however, needs to take on board an explicit reference to theory, and it needs to take seriously the reciprocal relationship between theory and empirical evidence, as this is an important mechanism for expanding the theoretical foundation of the discipline.

At this point the above-mentioned pragmatic and normative considerations need to be reconsidered again. This is so because any definition of competitive research would be incomplete if it did not reflect the social contexts in which research agendas are being drawn up or in which research results are being applied. It is important that the value-free stance of traditional regional geography is given up because it would be a self-deception to believe that regional geography can do without explicitly referring to or incorporating normative issues. As academic disciplines are an intrinsic part of society, they need to reflect the origin and the legitimacy of research topics and the use of research results. But which questions are to be raised and which problems are to be tackled by regional geographers? A systematic discussion of this issue has not yet taken place. However, it needs to be addressed in the context of a concerted effort to bring to life a 'new' new regional geography as a theoretically informed empirical project (Thrift 1998). The debate of the 1980s and 1990s broke new ground and generated a number of stimulating (conceptional) ideas from where the debate could be taken.

See also: Globalization: Geographical Aspects; Region: Geographical; Regional Integration

Bibliography

Bahrenberg G 1979 Anmerkungen zu E. Wirths vergeblichem Versuch einer wissenschaftstheoretischen Begründung der Länderkunde. [Some remarks on E. Wirths failed attempt to justify Lauderkunde Epistemologically.] *Geographische Zeitschrift* **67**: 147–57

Bassand M, Guindani S 1983 Maldéveloppement régional et luttes identitaires. *Espaces et Sociétés* **42**: 13–26

Buttimer A 1979 'Insiders', 'outsiders' and the geography of regional life. In: Kuklinski A, Kultalahti O, Koskiaho B (eds.) *Regional Dynamics of Socioeconomic Change*. Finnpublishers, Tampere, Finland, pp. 157–78

Claval P 1998 *An Introduction to Regional Geography*. Blackwell, Oxford, UK

Cooke P (ed.) 1989 *Localities. The Changing Face of Urban Britain*. Unwin Hyman, London

Giddens A 1988 *Die Konstitution der Gesellschaft. Grundzüge einer Theorie der Strukturierung*. Campus, Frankfurt, Germany
Giddens A 1991 *Modernity and Self-identity: Self and Society in the Late Modern Age*. Stanford University Press, Stanford, CA
Gilbert A 1988 The new regional geography in English and French-speaking countries. *Progress in Human Geography* **12**: 208–28
Gregory D 1982 *Regional Transformation and Industrial Revolution: A Geography of the Yorkshire Woollen Industry*. MacMillan, Basingstoke, UK
Gregory D 1988 Regional geography. In: Johnston R J, Gregory D, Smith D M (eds.) *The Dictionary of Human Geography*. Blackwell, Oxford, UK, p. 293
Hanson S 1999 Isms and schisms: Healing the rift between the nature-society and space-society traditions in human geography. *Annals of the Association of American Geographers* **89**: 133–43
Harloe M, Pickvance C G, Urry J (eds.) 1990 *Place, Policy and Politics: Do Localities Matter?* Unwin Hyman, London
Hartshorne R 1939 *The Nature of Geography: A Critical Survey of Current Thought in the Light of the Past*. Association of American Geographers, Lancaster, PA
Harvey D 1969 *Explanation in Geography*. Edward Arnold, London
Hauer J, Hoekveld G 1993 Introduction: Why regional human geography? In: Hauer J, Hoekveld G (eds.) *Moving Regions*. Koninklijk Nederlands Aardrijks-Kundig Genootschap, Utrecht, The Netherlands, pp. 13–24
Hettner A 1927 *Die Geographie. Ihre Geschichte, ihr Wesen und ihre Methoden*. F Hirt, Breslau, Poland
Holmen H 1995 What's new and what's regional in the 'new regional geography'? *Geografiska Annaler, Series B* **77**: 47–63
Johnston R J, Hoekveld G A, Hauer J 1990 *Regional Geography. Current Developments and Future Prospects*. Routledge, London
Kraft V 1929 Die Geographie als Wissenschaft. In: Kraft V, Lampe F (eds.) *Methodenlehre der Geographie*. Franz Deuticke, Leipzig, Germany, pp. 1–22
Massey D 1984 *Spatial Divisions of Labor. Social Structures and the Geography of Production*. MacMillan, Basingstoke, UK
Pickles J 1986 Geographical theory and educating for democracy. *Antipode* **18**: 136–54
Pohl J 1993 Kann es eine Geographie ohne Raum geben? *Erdkunde* **47**: 255–66
Pudup M B 1988 Arguments within regional geography. *Progress in Human Geography* **12**: 369–90
Sayer A 1989 The 'new' regional geography and the problems of narrative. *Environment and Planning D: Society and Space* **7**: 253–76
Smith N 1984 *Uneven Development: Nature, Capital and the Production of Space*. Blackwell, New York
Swyngedouw E A 1989 The heart of the place: the resurrection of locality in an age of hyperspace. *Geografiska Annaler, B* **71**: 31–42
Thrift N 1998 Towards a new new regional geography. *Berichte zur deutschen Landeskunde* 37–46
Van der Laan L, Piersma A 1982 The image of man: paradigmatic cornerstone in human geography. *Annals of the Association of American Geographers* **72**: 411–26
Villeneuve P 1986 Les théories sur la région: Polarisation et capillarité. In: De Koninck R, Landry L (eds.) *Les genres de vie urbains: essais exploratoires*. Université Laval, PQ, Canada, pp. 17–33
Wirth E 1978 Zur wissenschaftstheoretischen Problematik der Länderkunde. [Towards an Epistemological foundation of regional geography.] *Geographische Zeitschrift* **66**: 241–61

G. Wood

Regional Government

Regional government analysis has traditionally been restricted to state politics in federal systems or to center-periphery relations in the integration of the nation-state. We focus in this entry on Western Europe and North America to show: (a) that the political and legal status of regional institutions is extremely diverse; (b) that the rationales for regionalization have to be referred to state traditions; (c) that these rationales converge to develop both regional politics and public policies; and (d) that transnational integration gives a new relevance and meaning to regionalism.

1. The Ambiguous Status of Regional Government

Regional government is among the topics in comparative politics that clearly reveal the contingency of political institutions and the diversity of political processes behind them. Some countries exhibit very strong subnational governments, such as states in federal political systems (e.g., the USA or Switzerland), while others do not have any regional institutions at all (in centralized states such as Greece or France until the Fifth Republic). When such institutions do exist they may range from weak administrative units without much social significance to real representative institutions with powerful political capacities. To enhance the complexity of the picture, regions often coexist with other subregional units (counties, provinces, and *départements*), and display very different legal and fiscal authorities over the lower levels of territorial governments.

This variance suggests that the analysis of regional government is anything but simple, and that the idea of a universal form or of a common feature of regional institutions, even in the limited area of Western democracies, is misleading. Regional government cannot be analyzed independently of the political context within which it is taking place. It has to be referred to the general process of state building and state tradition, and understood as the product of different trends of regionalization.

In a restrictive meaning, regional government can be defined as the activity of political authorities specifically designed as an intermediary level between the state and local governments, and grounded in legitimacy vested in regional forms of political rep-

resentation. Therefore, it involves the scale of political authority (the level where sovereignty is defined and bounded), the aims of government (whether it takes care of regional issues or not), and the tools used to fulfil these goals (centralized vs. decentralized means of authority). It is conceivable to observe a government of regional issues without corresponding mechanisms of political representation at the regional level (e.g., with the Scottish National Office before the devolution reforms in the United Kingdom or the regional *Préfets* before the decentralization reform in France). Nevertheless, most Western democracies have evolved since the 1980s towards a greater empowerment of regional levels of political representation (most dramatically in Spain, France, Belgium, Italy, Canada, and more recently the UK), while federal systems turned to more interdependent patterns of relations between states and federal governments (notably in Germany, the USA, and Switzerland). In a more extensive definition, it is therefore more appropriate to conceive regional government as a set of complex institutional interactions and social mobilizations around political representation and public policies defined at an intermediary level between the state and local governments. If all Western democracies do not display distinctive political institutions at the regional level, all of them nevertheless show strong tendencies towards social and political mobilizations, and adjustments in institutional settings at this mesolevel of government (Sharpe 1993). Thus, this extensive definition is more appropriate to capture the actual issues of regional government in its contemporary context.

2. *State Traditions and Rationales for Regionalization*

From this definition we can infer that different state traditions, that is, differences in historical paths of state building and related administrative cultures, will matter in defining regional government. In exploring this topic, Loughlin and Peters have identified four main state traditions. (i) The Anglo-Saxon tradition (including countries such as the USA, the UK until 1998, Ireland, and Canada with the exception of Quebec), where there is no (in the UK) or a limited (in the USA) legal basis for the state, and where regional government can be either the 'state power' of the federal system (the USA), or shadowed by the strength of local government (in the UK). (ii) The Germanic tradition (covering countries such as Germany, Austria, Spain after 1978, and Belgium after 1988), rooted in a strong legal status of a Weberian state and a co-operative federalism giving large powers to regional government. (iii) The French State tradition, which is Jacobean and centralist, and where regions appear as technocratic creations progressively building a regionalized unitary state (for example, France, Spain until 1988, Portugal, Greece, Belgium until 1988, and Quebec). (iv) Finally, the Scandinavian state tradition of Sweden, Norway, and Denmark, which, like the Germanic one, is rooted in a strong legal basis of the state, but in which the territorial organization is based more on a strong local autonomy than on the development of regional government (Loughlin and Peters 1997).

State tradition conceptualized in this way does not explain everything regarding regional government. For instance, there are strong discrepancies in the Anglo-Saxon model between European and American countries, and The Netherlands appears as an exception in the Germanic tradition. But it helps nevertheless to understand the different territorial policy paradigms (Balme et al. 1994) and political rationales for regionalization, understood as the forces driving the creation and development of regional government.

The first of these rationales is strictly political and linked to the integration of the nation-state. Regionalism here is close to nationalism, and regional government can be seen either as a tool to sustain the integrity of the state or as the 'slippery slope' leading first to larger claims to regional autonomy and ultimately to separatism and independence. This issue is dependent on social and territorial contiguity, on the degree of economic integration of regions to the country, and on the use of political violence between protagonists. However, the meaning of institutional development at the regional level will be different in an Anglo-Saxon context, where it appears as explicitly political with devolution (Elcock and Keating 1998), from a continental one, where regional decentralization can also be justified by administrative arguments and in this respect be more easily acceptable.

The second form of regionalization is federalism. Here the rationale is not so much political as constitutional–legal. It refers not to its integrity or to the coexistence between different cultural communities as in the previous case, but to the structure of the state, to its territorial organization normatively defined. Again, the strengthening of regional government has different meanings in the Germanic and in the Anglo-Saxon tradition. When states gain in power in the co-operative federalism of Germany, it is understood as more political interlocking (*Politikverflechtung*) between federal and state levels of government. This process leads to complex patterns of interdependence, often sub-optimal and sometimes unable to avoid 'the joint decision-trap', that is, the impossibility to make any decision other than the smallest core for agreement (Scharpf et al. 1976). The governing capacity of the system is in this context limited by vetoing powers of territorial governments, but the general picture is nevertheless a context of 'big government', increasing public expenditure, and governmental overload. It is noticeable that this kind of analysis has been expanded by this author to European governance as a whole

(Scharpf 1988). In the European context, federalism has been transformed by European integration (Jeffery and Sturm 1993, Jeffery 1997).

The increase in regional government in an Anglo-Saxon 'disjoined' federal system would take a very different meaning, and does not appear likely in the contemporary context. In such a political system, states do not have an institutional access to federal government other than through the Senate. Federal and state policies are supposed to be distinct areas of government, and there is in principle no co-operation between the two levels of government in defining public policies, other than through judicial review. In fact, the 'marble cake' concept of federalism has long been violated by numerous overlaps, redundancies, and competition between federal and state administrations through the development of public policies (Gray et al. 1990, Wright 1988). Here the state tradition is liberal and rests on the idea of a minimal central government. Consequently, the empowerment of state governments would mean resisting the continuous development of federal policies, and saving the political system both from centralization and from 'big government'. Empirically, however, despite a lower level of public spending than in Europe and more limited welfare policies, the US political system has gradually evolved towards higher levels of public intervention and more overlap and interdependence between levels of government.

Finally, the third rationale for regionalization lies in public policies. Even in centralized states such as France, the development of the public sector during the decades after World War II, as well as changes in the economy associated with increased mobility, have often led to recognition of the need for a regional scale of decision making. In France, regions were first created as administrative bodies aimed at improving the territorial fitness of public investments and only gradually became political institutions, with the first election of their councils in 1986. Again, in a centralized state, the reform is mainly administrative (at least at the beginning), whereas in a federal system it is openly political with the transfer of the burden of public policy from federal to state governments.

3. *Regional Politics and Policies*

These three rationales have converged since the 1970s to strengthen in one way or another the political salience of regional government. This in no way means more autonomy at the regional level. It means, on the contrary, more interdependence in complex institutional settings, where regional specificities, be they economic, social, cultural, or political, have come to be recognized. This has also led to altogether fewer conflictual and hierarchical modes of integration between center and peripheries, defining new forms of public policies and political authority.

Regarding regional politics, political mobilization around regionalist parties was noticeably sustained in Quebec, Catalonia and the Basque Country, Corsica and Overseas French regions, Wallony and Flanders, in Italy with the Northern League, and to a lesser extent in Wales and Scotland. More generally speaking, independently of quasinationalist movements, politics at the regional level gained in saliency with the development of public policies at the regional level. Regions more specifically develop public programs in education and professional training, culture, economic development, transportation, and regional policy. They answer in a way to a functional logic of public policy, where sectoral policies are selected by the relevance of the scale of government, but also and mainly by the opportunity structure defined by the existing institutional design and potential political support. Regional public policies are thus linked to mobilizations of social interests, sometimes through neo-corporatist modes of organization, and in any case to some forms of political exchange structuring political leadership (Koller-Koch 1997, Négrier and Jouve 1998). More controversially, the performance of regional institutions also refers to social capital and civic traditions (Putnam 1993). Without overestimating such a causality, the relation between regional sociality and identity on the one hand and regional governance on the other hand offers stimulating perspectives for analysis. Generally, societal mobilizations at the regional level tend to correspond to inter-institutional modes of regionalization of public policy.

4. *Regional Government and Transnational Integration*

Finally, regional government is affected by transnational integration in two different ways. First, economic globalization increasing the mobility of economic factors defines an opportunity for the development of regional policies aimed at managing the territorial consequences of these changes, both in attracting investments and in defending regional firms, industries, and social sectors. The creation of transnational markets such as the European community, NAFTA, and MERCOSUR transforms socioeconomic conditions and needs, most dramatically in the case of border territories and monoindustrial regions affected by international competition. Second, political integration at a supranational level, as it is best exemplified with the European Union, defines new issues for regions to get access to regional policy programs, and more generally to preserve their interests in the new political process. Getting involved in new public policies, regions also invent new channels of influence progressively designing a new political game made of transnational arenas and forums. The creation in 1992 by the Maastricht Treaty of the Comity of the Regions

reveals this logic of recognition of regions by European institutions and their inclusion in this political process. Even if their influence is extremely diverse (Keating and Jones 1985) and needs not to be understood to be an alternative to the power of nation-states (Le Galès and Lequesne 1998), it is nevertheless part of the more polycentric and multilevel forms of governance (Hooghe 1996) characteristic of the European Union. In the same way, the development of foreign relations (Keating and Aldecoa 1999) and of transnational co-operations between regions (Balme 1996) contributes to the bypassing of the Westphalian and exclusively interstate system of international relations, and defines regional collective action as an answer to globalization. This context defines a new relevance for regionalism, but also transforms its meaning (Letamendia 1999), as this new regionalism turns away from parochialism and tradition, and combines economic openness and cultural identity to sustain socioeconomic development in a more cosmopolitan way.

See also: European Integration; Federalism; Interest Groups; Regional Integration

Bibliography

Balme R (ed.) 1996 *Les Politiques du Néo-régionalisme*. Economica, Paris

Balme R, Garraud Ph, Hoffmann-Martinot V, Ritaine E 1994 *Le Territoire pour Politiques: Variations Européennes*. L'Harmattan, Paris

Elcock H, Keating M 1998 *Remaking the Union. Devolution and British Politics in the 1990s*. Cass, London

Gray V, Jacob H, Albritton R B (eds.) 1990 *Politics in the American States. A Comparative Analysis*. 5th edn. Scott, Forsman, Glenview, IL

Hooghe L (ed.) 1996 *Cohesion Policy and European Integration. Building Multi-level Governance*. Oxford University Press, Oxford

Jeffery C (ed.) 1997 *The Regional Dimension of the European Union. Towards a Third Level in Europe*. Cass, London

Jeffery C, Sturm R 1993 *Federalism, Unification and European Integration*. Cass, London

Keating M, Aldecoa F 1999 Paradiplomacy in action. The foreign relations of subnational governments. *Regional and Federal Studies* **9**(1): special issue

Keating M, Jones B (eds.) 1985 *Regions in the European Community*. Oxford University Press, Oxford

Koller-Koch B (ed.) 1997 *Interactive Politik in Europa: Regionen im Netzwerk der Integration*. Leske und Budrich, Opladen, Germany

Le Galès P, Lequesne C (eds.) 1998 *Regions in Europe*. Routledge, London

Letamendia F 1999 *Nacionalidades y Regiones en la Union Europea*. Editorial Fundamentos, Madrid

Loughlin J, Peters B G 1997 State traditions, administrative reforms and regionalization. In: Keating M, Loughlin J (eds.) *The Political Economy of Regionalism*. Cass, London, pp. 41–62

Négrier E, Jouve B 1998 *Que Gouvernent les Régions d'Europe? Echanges Politiques et Mobilisations Régionales*. L'Harmattan, Paris

Putnam R D, Leonardi R 1993 *Making Democracy Work. Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ

Scharpf F W 1988 The joint-decision trap. Lessons from German federalism and European integration. *Public Administration* **66**: 239–78

Scharpf F W, Reissert B, Schnabel F 1976 *Politikverflechtung*. Scriptor-Verlag, Kronberg

Sharpe L J (ed.) 1993 *The Rise of Meso Government in Europe*. Sage, London

Wright D S 1988 *Understanding Intergovernmental Relations*. Brooks, Cole, Pacific Grove, CA

R. Balme

Regional Identity

1. Introduction

The concept 'identity' is concerned with the answer to the question: 'Who am I?' 'Place identity' and 'regional identity' are answers to the question: 'Where am I?' Sometimes this question is more differentiated: 'Where do I come from' and 'Where do I belong to?' The phenomenon regional identity has an enormous range: it stretches from the personal home to territorial politics (geopolitics), from self-consciousness up to the problem, how to cope with the impacts of globalization.

2. Definition

(a) Regional identity is part of the self-identity of the human being. Regional identity results from the close ties of an individual to a region. It may consist of different intensities: it reaches from (i) a vague sense of belonging to a particular region, (ii) a close attachment to a region, (iii) a deliberate confession to and identification with a region, up to (iv) an active engagement for the region. In each of these cases areas, which are defined emotionally or rationally, serve as an objective, relatively stable source of orientation for the behavior of the individual.

Regional identity is very close to personal identity. Personal identity does not exist for itself but is built—and permanently rebuilt—in social interactions. So there is a strong relation to the (social) environment of a person. In the same way regional identity is not only a personal feeling towards a place but is part—or counterpart—of the collective. There is no self-identity without embeddedness into a community and there is no identity of the collective without individuals.

Therefore regional identity is not only part of the personality which is bounded to the region, but also and particularly consists of the knowledge of others having similar feelings for the region as oneself.

(b) Regional identity is not always regarded as an

attitude or a characteristic of a person (or a group of persons). Sometimes it means a particular region as a whole. Then 'region' is regarded as an entity with special characteristics so far different from other regions. Although it obviously bases on individual attitudes, regional identity is regarded as a relatively autonomous social structural formation, which neither can be seen as the total of individual attitudes nor can be dissolved into society.

To this point there are different positions according to the underlying epistemological background. To declare regional identity as a structure of its own derives from the tradition of the 'physique sociale' founded by Emile Durkheim (1893) and—in a lower degree—from communitarism. Following the position of action theory this second definition of regional identity is seen as an improper naturalism or essentialism. But nevertheless it can be pointed out that a specific characteristic of regional identity obviously is the oscillation between the level of action of the individual and the level of structural formation of the society.

3. Special Aspects of Regional Identity

3.1 Characteristics of Regional Identity

Regional identity is not an isolated status quo of an individual, but is bounded to the development of social interactions as well as to the individual biography. Regional identity develops from different sources, from individual experiences during childhood, from the perception of the personal space, and other experiences. However, even more important are socially mediated aspects, because identity does not exist in individuals for itself, but develops, grows, and changes in social interactions. Therefore culture, tradition, landscape, and history of the region form regional identity and are part of it.

A decisive element of regional identity is the close relation between the region and the historical dimension. The common history is not only referred to as the history of a political territory, it may also refer to the affiliation to an unified confessional territory, a linguistic boundary or similar aspects. Normally there is a bundle of cultural elements important for the identity of a region and the identification of individuals with the region in question. Such elements can be:

(a) the economic basis, e.g. the production of coal and steel in the Ruhr area in West Germany. The popular name of the region 'Ruhrpott' (melting pot at the river Ruhr) reminds one of the process of melting ore with coal;

(b) the language or specific dialect, e.g. 'Languedoc' or 'Occitanie' in the south of France (where people say 'hoc' instead of 'oui' in the north of France), 'Catalunia' in the east of Spain, 'Friul' in the northeast of Italy;

(c) a particular landscape like an island, a big river, or a mountain area (Corsica, The Rhine Valley, the 'Padania' around the river Po, The Highlands);

(d) even specific regional dishes (Cornish pastry and Cornish cream, Tirolian 'Knoedel') can be part of the shared frame of reference. Of course regional identity rarely develops from that kind of particularity alone.

Abstract phenomena as well as objective elements can stand as objects of identity. Normally it is not a single element alone but several characteristics of a specific region which contribute to regional identity.

These shared symbols are normally grounded in the history and development of the region. But the historical dimension also may be an artificial 'myth' or an imagined community in the sense of Anderson (1996). What is really important is the belief in the common history and the existence of a collective. Following Weber (1922) this belief is called 'Gemeinsamkeitsglaube' (for details see Sect. 4). If these shared symbols become an important element of communication in the lifeworld or even in political discourse then there is a switch from regional identity (a) to (b).

3.2 Regional Identity and Scale

Regional identity is a kind of spatial identity on a certain scale called the meso-level. Regional identity is located between the national and the local level. To illustrate regional identity spatially the term 'county' should most likely fit.

Scale is helpful to distinguish regional identity from place identity. Although in some cases place identity is used as the overall term for the different levels of scale, place identity is normally associated with the directly perceived space.

An aspect which distinguishes regional identity from place identity is a higher degree of abstraction from the concrete space. To a lesser extent regional identity is bounded to the concrete perceived space, the personal perception and experience, and the co-presence of individuals. 'Local identity' is bounded to the network of friends and relatives, to the personal socialization of the individual and therefore people are deeply attached to particular places. Regional identity, however, is more abstract: it does not need an actual or the remembrance of a former co-presence of people at the same place. The person with close ties to the region normally knows the region, the physical space, and even more its inhabitants, only in part by personal experience. Particularly against this background symbols which are known and shared by a collective are quite important to establish a sense of community.

3.3 Regional Identity and Regionalism

Regional identity is different from localism as well as from nationalism or patriotism. Above all regional identity is a cultural, social, and psychological phenomenon and not a political one, but this depends on the

degree of the intensity described above. As far as regional identity includes a political dimension, it tends to keep cultural traditions or to blame economic disparities.

If the interest for regional issues leads to a strong political engagement the term 'regionalism' is used to describe this attitude which leads to a larger political movement. Regionalism is the political movement with the aim to preserve or to gain cultural or economic self-reliance (or both). Political autonomy is not the first aim but tends to develop during the history of a political movement. Regionalism often is marked by clear aims and strategies. If the intention is political autonomy one even has to talk about nationalism at least under the perspective of its supporters. These regional political movements can be organized like political parties. Examples are the 'Lega Lombarda' in Northern Italy, the 'Partis Quebecois' in Canada in a more traditional political sense on the one hand, and the liberation movements in Corsica or the Basque provinces in the north of Spain on the other hand.

The line between vague regional identity and active political regionalism is not sharp and the motifs are mixed together. See for instance the Occitanian case: on the one hand in the late 1960s the inhabitants of this region felt disadvantaged. The starting point of the protest was the bad situation of a region depending economically on the production of wine. Most people in that region were affected by this and it had little to do with regional culture and so forth. On the other hand, the regional protest led to much solidarity in the heart of France, especially in the intellectual milieu around the politically left students and academics. The engagement of these people for Occitanian, drinking Herault wine and spending their holidays at the Causse de Larzac, had little to do with their place of origin or the feeling-at-home in that region. The Occitanian identity of these intellectuals expressed their (general) political engagement in favor of disadvantaged regions and the utopia of an intact world. Regionalism unites individual interests and focuses people on common ends. Therefore, regional identity is to be seen as an element of the social system in the sense of the second definition given above.

3.4 Regional Identity and 'Real' Region

Talking about the Occitanian identity of the intellectuals at Paris has already expressed that the region as an area on the earth's surface and the region of the identity are not always congruent. Of course there must be a relation of a particular regional identity to a specific region. But this does not mean identical maps. For example, an administrative region like 'Bretagne' in France or 'Andalusia' in Spain is not congruent with the area of people calling themselves 'Bretons' or 'Andalusians' and is even less congruent with their regional identity. If people draw their mental maps of the region in favor, the results differ considerably. Normally it can be assumed that the regional identity is more distinctive in the heartland than in the border area. In the center the regional language or dialect prevails and the specific traditions and rules of the lifeworld are cultivated. But even the contrary is possible: if the neighboring region seems to be strange or even hostile regional identity will be strongest at the border, e.g., this phenomenon can be found in the case of South Tyrol.

Another important point in this context is the personal absence from or presence in the region in question. Regional identity is not attached to the physical presence in that particular region. On the contrary, regional identity is sometimes stronger in foreign surroundings than at home: Friulians in Canada, people from Sicily in Torino, Bavarians in Hamburg, and so forth often feel and think more Friulian, Sicilian, and Bavarian than people who see their region as the natural place of the everyday life.

Finally, the Occitanian example shows that regional identity is not always a particular regional identity. Regional identity is an universal value too and not attached to a specific region. To prefer federalism to centralism in a political system expresses this attitude.

4. The Formation of Regional Identity

How does regional identity develop? According to the two aspects of the notion mentioned above in the definition, at least the answer to this question has to make a difference between two levels: the first is the anthropological aspect (in close relation to self-identity), the second is the social theory with special respect to social structural formation.

4.1 Regional Identity as a Basic Need of Human Beings

Regional identity is part of self-identity and therefore stretches from the unconsciousness to the rational consciousness. Especially an active political engagement for the reason in favor is already part of the discursive consciousness.

Above all regional identity is part of the practical consciousness in the sense of Giddens (1984). Among others practical consciousness means living and working in everyday life by routines. Therefore the surroundings and experiences in the places where people live form a stable frame for individual decisions and behavior. One can say that regional identity is built and permanently rebuilt in the social and spatial context of the individuals. The well-known region and its people give trust and security for the behavior of the individuals.

From a phenomenological point of view region has a similar function: at the heart of the lifeworld is the home world in contrast to the strange world outside.

Regional identity as a social process is constructed on different levels: in kindergartens, schools, clubs, and other institutions. Communication and experiences in everyday life form the attitude towards the region.

4.2 Regional Identity as an Element of Social Cohesion

In social theory the topic 'regional identity' is part of the discussion about inclusion and exclusion starting with the book of Ferdinand Tönnies *Gemeinschaft und Gesellschaft* (*Community and Society*) at the end of the nineteenth century (1887). Max Weber says that the formation of community (*Vergemeinschaftung*) and the formation of society (*Vergesellschaftung*) are two of the essential sociological notions. If people believe that there is a traditional or emotional feeling of togetherness then community can arise. To understand regional identity two points are of special interest. First, there must be a real starting point of a regional community (e.g. language, religion, origin) but it is just the belief in the force of this shared characteristic that can lead to the formation of a community. Therefore, even regional identity is socially constructed. Second, there is no need for the co-presence of individuals. The feeling of belonging together can be a powerful motor for the formation of inclusion, especially if there are no other characteristics easy to recognize and to manipulate. So the notion regional identity is not very different from the modern notion of ethnicity. The modern theory of ethnology too regards 'ethnicity' as a construct, while the traditional (primordial) notion of ethnicity is essentialistic and has a strong affinity to race.

From a systematic point of view social cohesion needs rationality as well as emotion. In a similar way to Weber, Emile Durkheim (1893) illustrates a difference between 'mechanical solidarity' and 'organic solidarity' in human societies. In the traditional 'Physique sociale' as well as in the modern 'New Social Physics' social integration is to be seen as a result of collective consciousness (on the basis of emotion, religion, and so forth) in the first case and of treaties between agents in the second. It helps people to avoid making arduous decisions if they follow the idea of 'natural cohesion'. Or as B. Anderson points out: 'In this way, nationness is assimilated to skin-colour, gender, parentage and birth-era—all those things one can not help. And in these "natural ties" one senses what one might call "the beauty of *gemeinschaft*"' (Anderson 1996, p. 143).

5. Regional Identity Between Modernism and Postmodernism

Is regional identity a relic of premodern societies disappearing during the process of modernization? Following the ideas of Giddens (1984), people are more and more disembedded and this means that their activities and their thinking, that social relations in general, are no longer bound to the space–time-context. Time and space get more and more divided. The boundaries of the region and the self coincide no longer. The space of reference for the individual actor in the time of globalization finally is the whole earth. Therefore it is declared that regional identity loses its importance during the process of modernization. A similar position can be derived from the General Systems Theory of N. Luhmann (1994). According to Luhmann, the whole modern world is inter-related and regional worlds of their own no longer exist. Differentiation in the world is based on functions. Normally the same person belongs to different systems in different places depending on their role as a consumer, employee, or lover. M. Weber had a strong historical interest and regarded social change and the process of modernization but there is no clear statement that he regarded community as a—more and more disappearing—historical phase during the formation of (modern) society.

On the one hand one can be sure that there are no longer closed up regional worlds like in premodern times. On the other hand 'region' is a very simple and therefore very practical code to produce cohesion and identification. Even more: the more complicated and differentiated the modern world becomes the more we need simple possibilities to discern people, political subsystems, and so forth. Therefore regional identity is one of the powerful social bases for political movements, starting from regional protest based on environmental issues up to most of the wars in the late twentieth century. 'Belonging to'—not only same region, but also to the same religion, the same 'race' and so forth—is much more relevant than class differences or other social aspects to include or to exclude people. Such kinds of inclusions—and the corresponding exclusions—are responsible for most of the recent (violent) political developments especially in the east of Europe, but also in Africa or in the Near East.

Regional identity is embedded into the more general debate on the role of identities in relation to the universal values (of the Enlightenment). It is part of the debate which of these two basic values is of more importance: to respect individual historically grown traditions or to assert the general human rights all over the world. There is a strong political dimension connected to this subject: it stretches from the apparently simple question if Islamic girls are allowed to join school lessons with a scarf around their head in laizistic French schools up to the question if western politicians are allowed to claim the recognition of human rights in the People's Republic of China. The popular considerations of Finkielkraut (1989) in France and Germany and of Said (1978) and Bauman (1991) in the English speaking countries give new importance to this subject. The position of the—now

as occidental recognized—values of the Enlightenment has diminished and pluralism had to be put on the agenda. There are different 'language games' (in the words of Ludwig Wittgenstein) which are incomparable.

So it has to be pronounced that in the end regional identity is part of the postmodern debate. In a postmodern context regional identity can be submitted to the leading category of difference (and unity). In (post-)modern guise questions come back which were important already at the end of the eighteenth century: Immanual Kaut defended the achievements of the (French) Enlightenment, while John Gottried Herder so preparing romanticism and historism declared that these values are not valid for the rest of the world: in different regions develop different cultures which cannot be compared and ranked.

6. *Relevance of Regional Identity Today*

Following the theoretical considerations, regional identity as 'a sentiment of belonging to' always exists to a higher or lower degree at the individual level. Sometimes it is just folklore, but very often it gets at the heart of social and political issues. Regional identity is part of the general tendency in societies to construct social cohesion. Some examples are ecological and political movements, emigration colonies, and economic developments. For instance, political movements focused on ecological topics or resistance against nuclear power plants often affirm to defend 'their' region. As already shown in the example of Occitania there is often a mixture of political interest and traditional emotions: without doubt regional identity serves as symbolic arm in political discussions but without regional identity such political movements would not exist—or only for a short time.

People who are forced to emigrate try to keep their former identity in Little Sicily, in the Arabian 'banlieue' of Paris, in the Inner City Ghettos of the 'Guest workers' in Western Germany. Home is what you feel when you have lost it, so people try to keep their identity by staying together. In globalization, regional identity has a new dimension. On the one hand, people lose control to their surroundings more and more, so they try to keep regional traditions, on the other hand, it is much easier to get involved in the process of globalization and to keep its regional identity within a transnational space at the same time. Even more, jobs are more and more insecure, the firm is no longer the only center of thinking of the employees, while place gains new importance (Sennett 1997). So regional identity becomes more important.

In a similar way regional identity works in regional economics. The moral economy precedes the market economy and is still of importance. The success of industrial districts is sometimes an outcome of regional embeddedness of the entrepreneurs as well as the employees of the region where it is located and where the network works. Shared rules and common values produce trust, confidence, and reciprocity which are regarded as the surplus making such specialized regions more successful than others. On the other hand it is also possible that there is an economic decline because of too much regional identity and following lock-in-effects: the Ruhr area in the 1970s and 1980s or—in a broader sense—Sicily and its whole Mafia problem may be seen as examples. Regional identity is therefore not a sentimental relic of premodern times. Regional identity is one of the geographical dimensions of the construction of modern society.

See also: Identity in Anthropology; Identity Movements; Identity: Social; Language and Ethnicity; Region: Geographical

Bibliography

Agnew J 1993 Representing space: Space, scale and culture in social science. In: Duncan J, Ley D (eds.) *Place, Culture, Representation*. Routledge, London, pp. 251–71

Anderson B 1996 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, rev. and extended 2nd edn. Verso, London

Bauman Z 1991 *Modernity and Ambivalence*. Polity Press, Cambridge, UK

Durkheim E 1893 *De La Division du Travail Social*. Alcan, Paris

Entrikin J N 1994 Place and region. *Progress in Human Geography* **18**: 227–33

Entrikin J N 1996 Place and region 2. *Progress in Human Geography* **20**: 215–21

Entrikin J N 1997 Place and region 3. *Progress in Human Geography* **21**: 263–8

Finkielkraut A 1989 *Die Niederlage des Denkens*. Rowohlt Taschenbuch Verlag, Reinbek [*La Défaite de la Pensée*]

Giddens A 1984 *The Constitution of Society, Outline of the Theory of Structuration*. Polity Press, Cambridge, UK

Hard G 1987 'Bewußtseinsräume'. Interpretationen zu geographischen Versuchen, regionales Bewußtsein zu erforschen. *Geographische Zeitschrift* **75**: 127–48 ['Regions of regional consciousness'. Geographical approaches to regional identity]

Luhmann N 1994 Inklusion und Exklusion. In: Berding H (ed.) *Nationales Bewußtsein und kollektive Identität*. Suhrkamp, Frankfurt a. Main, pp. 15–45, [*National Consciousness and Collective Identity*]

Pohl J 1993 *Regionalbewußtsein als Thema der Sozialgeographie*. Münchener Geographische Hefte 70. Laßleben, Kallmünz [*Regional Consciousness as Subject of Social Geography*]

Said E W 1978 *Orientalism*. Pantheon Books, New York

Sennett R 1997 The new Capitalism. *Social Research* **64**: 161–80

Short J R 1991 *Imagined Country: Environment, Culture and Society*. Routledge, London

Tönnies F 1887 *Gemeinschaft und Gesellschaft. Abhandlung des Communismus und des Socialismus als Empirischer Culturmasse*. Fues, Leipzig, Germany

Weber M 1922 *Wirtschaft und Gesellschaft. Mohr, Tübingen* [Weber M 1978 In: Roth G, Wittich C (eds.) *Economy and Society*. University of California Press, Berkeley, CA]

J. Pohl

Regional Integration

1. Definition

The term 'regional integration' refers to the process whereby sovereign nation-states voluntarily establish common institutions for collective governance. This process is multifaceted. At the political level, regional integration involves the creation of supranational institutions, with executive, legislative and judicial powers, and the delegation of policy competences from domestic governance institutions to these institutions. On a legal level, regional integration involves the establishment of a new hierarchy of legal norms, where supranational law is not only an integral part of domestic law (the doctrine of 'direct effect') but is also sovereign over domestic law (the doctrine of 'supremacy'). At the economic level, regional integration involves the development from a common market (in parallel with existing national markets), through a single market (with the removal of barriers to the free circulation of goods, services, capital and labor), to a common or single currency (in parallel with, or replacing, national currencies), and eventually to full economic union (with common macroeconomic and fiscal policies). At the sociological level, regional integration involves the creation of new supranational identities, loyalties, interest groups, and social movements, in addition to and sometimes in opposition to existing national identities, interests, and movements.

2. The Early Study of Regional Integration

The first steps towards regional integration occurred in Europe in the immediate aftermath of the Second World War. This led to the establishment of the European Coal and Steel Community (ESRC) in 1952 by six states (Germany, France, Italy, and the Benelux countries), who then created the European Economic Community and the European Atomic Energy Community in 1958. Hence by the end of the 1950s, regional integration was more advanced in Europe than in any other region in the world. Nevertheless, modeled on the European process, inter-state agreements and supranational economic regimes began to emerge in South East Asia, Latin America and North America.

As a result, in the 1950s, the process of regional integration was regarded by most social scientists as a generalizable phenomenon. For example, Karl Deutsch and his colleagues (Deutsch et al. 1957) developed a general 'transactions theory' of integration: as cross-border economic and social transactions increase, new regional identities and communities are created, which facilitate new institutions, and so on. The zenith of the early study was the 1970 special issue of the journal *International Organization*, which tried to develop and test theories that could explain the varying stages of regional integration throughout the world (Lindberg and Scheingold 1971).

3. European Integration: A Sui Generis Phenomenon

However, with the global economic crises in the early 1970s, it became clear that the experiments in regional integration outside Europe were not as deeply entrenched as the process within Western Europe. Ernst Haas, the founding father of 'regional integration theory,' consequently proclaimed that the generalizable study of regional integration was now 'obsolete' (Haas 1976). From the launch of European integration in the 1950s to the relaunch in the mid-1980s, there have been two dominant explanations of European integration.

3.1 Neofunctionalism

Neofunctionalism was first proposed by Haas (1958) to explain the establishment of the ECSC Treaty, and was developed by his disciples to explain subsequent integration steps. The basic argument of neofunctionalism is that European integration is a deterministic process, whereby 'a given action, related to a specific goal, creates a situation in which the original goal can be assured only by taking further actions, which in turn create a further condition and a need for more, and so forth' (Lindberg 1963, p. 9).

This 'spillover' process is promoted by two sets of actors. First, at the domestic level, societal and sectional interests promote integration of different policy sectors to achieve political advantage over their opponents or secure benefits already accrued by the integration of other policy sectors. For example, in the 1980s, multinational business interests lobbied for the establishment of a European-wide single market, to secure greater economies of scale and lower transactions costs in cross-border trade. Second, at the European level, the supranational institutions promoted further integration to increase their ability to shape policy outcomes. For example, in the 1960s the European Court of Justice consciously made rulings and established doctrines to promote legal integration in Europe, and in the mid-1980s the European Commission was instrumental in persuading governments to adopt the program for completing the single market.

3.2 Intergovernmentalism

Neofunctionalism's failure to explain the slowdown of European integration in the 1970s, and the subsequent strengthening of the 'intergovernmental' elements of the European Communities, led to the emergence of a starkly opposing theory of European integration, known as intergovernmentalism. In direct contrast with neofunctionalism's emphasis on nonstate interests at the domestic and supranational levels, intergovernmentalism argues that European integration is driven by the interests and actions of the European nation-states. In this interpretation, the main aim of nation-states is to protect their geopolitical interests, such as national security and 'sovereignty.' Also, interstate bargaining is a zero-sum game, where 'losses are not compensated by gains on other issues: nobody wants to be fooled' (Hoffmann 1966, p. 882). Consequently, against the neo-functionalist 'logic of integration,' intergovernmentalists see a 'logic of diversity [which] suggests that, in areas of key importance to the national interest, nations prefer the certainty, or the self-controlled uncertainty, of national self-reliance, to the uncontrolled uncertainty of the untested blunder' (Hoffmann 1966, p. 88). From the intergovernmentalist perspective, then, European integration was relaunched in the mid-1980s because the governments in the large member states shared the same policy preference for a single market and the governments in the periphery states were bought off through 'side payments,' such as regional transfers (e.g., Moravcsik 1991).

4. Reintegrating Regional Integration with General Social Science

Ironically, as European integration gathered pace at the end of the 1980s, regional integration began again to be understood as a general phenomenon in social science. There were three main reasons for this shift. First, with the Single European Act in 1987, the establishment of the EU in 1993, and the launch of European economic and monetary union in 1999, European integration began to have a significant impact on domestic political systems and on global political and economic relations. European integration and the EU hence became a major focus of research in a number of social science disciplines.

Second, in the 1990s, the success of the European single market encouraged regional integration in other parts of the world—where governments breathed new life into moribund regional organizations (such as the Association of South East Asian Nations) or created new regional governance regimes (such as the North American Free Trade Area). Once again, regional integration was a general rather than a *sui generis* phenomenon.

Third, and related to these two aspects, with this renewed interest in the European process came specialization, where scholarship focused on specific aspects of the integration process or the EU rather than on the broad historical development of regional integration. This specialization enabled scholars of European integration to apply general social science theories, e.g., of interstate bargaining, interest group behavior, legislative–judicial interactions, and monetary integration. Specialization consequently increased the subtlety and sophistication of our explanations of regional integration. It also enabled scholars of regional integration to be part of the mainstream study of these cross-systemic social science phenomena. This integration of regional integration and social science occurred in political science, legal studies, economics, and sociology.

4.1 Regional Integration in Political Science

In the International Relations subdiscipline of political science, regional integration is a particular type of supranational regime building between nation-states. For example, in his 'liberal intergovernmentalist' (LI) theory of regional integration, Andrew Moravcsik (Moravcsik 1998) deliberately places regional integration at the intersection between the two main approaches in international relations: liberalism and realism (see *International Organization*). From liberalism, LI argues that state preferences are formed through a process of domestic competition between rival economic interests in society. From realism, LI argues that once state preferences have been formed, states bargain to protect and promote their private interests, and cooperate to design institutions to reduce transactions costs, overcome collective action problems, and establish credible commitments (for example, by delegating agenda-setting and enforcement powers to supranational agents).

In the subdiscipline Comparative Politics, the EU, despite its peculiarity, possesses all the main attributes and processes of a political system. This enables the different elements of regional integration to be explained using general theories from comparative politics (Hix 1999). For example, the EU governments have delegated executive powers to the European Commission, and have designed institutions to control these powers, in the same way that legislative principals delegate to and control the powers of executive agencies in the domestic arena. Similarly, on the 'inputs' side of the political system, mass attitudes towards regional integration are shaped by the same utilitarian calculations and affective beliefs as the public show towards domestic institutions. Further, on the 'outputs' side, regulation (rather than redistribution) emerged as the central policy instrument of regional integration in parallel with the expansion of regulation and independent regulatory agencies in the domestic arena.

4.2 Regional Integration in Legal Studies

In legal studies, with the establishment of the doctrines of direct effect and supremacy, and the hierarchical relationship between the European Court of Justice and national courts (where national courts apply EU law in the domestic arena), the EU legal system is akin to a federal legal system (Weiler 1991). However, since the German Constitutional Court judgement on the legality of the 1992 Treaty on European Union (the Maastricht Treaty), there is not a clear hierarchy of norms in the EU system. EU law, in its primary form in the EU Treaty and in secondary legislation, is supreme over conflicting domestic statutes and legislation. However, according to the German Constitutional Court and similar constitutional authorities in other member states, EU law remains subservient to basic fundamental rights, as set out in domestic constitutions and the European Convention on Human Rights.

In the intersection between political science and legal studies, the interaction between the EU governments and the European Court of Justice is similar to the interaction between domestic legislative and judicial authorities. On one side, EU governments have delegated judicial power to the European Court of Justice to establish a rule of law, without which there would be little incentive for governments to implement EU legislation. On the other side, protected by the constitutional guarantees in the EU Treaty, the European Court of Justice has pursued the political strategies of many domestic constitutional courts, by reinforcing its legal authority versus lower courts and shaping political outcomes through the policy adjudication process.

4.3 Regional Integration in Economics

In economics, regional economic integration is a particular case of international economics and monetary economics. In international economics, states have liberalized trade to promote comparative advantage and economies of scale. From this perspective, the United States (in the ninteenth century), the European Union, the North American Free Trade Area and the World Trade Organization are all examples of the same phenomenon. In all these cases, the removal of barriers to the free circulation of factors of production was then followed, to a greater or lesser extent, by the creation of common regulations to address market failures, such as unfair competition (through merger rules), asymmetric information in market transactions (through rules on consumer protection), and negative externalities (through rules on environmental protection).

In monetary economics, the process of economic and monetary union in Europe, with the creation of the single currency, has led to the resurrection of the old 'Optimum Currency Area' (OCA) theory of monetary unions (Eichengreen and Frieden 1994). In the OCA theory, nations decide to form a monetary union when the benefits of keeping independent exchange rates and interest rates (to address asymmetric demand shocks) are lower than the benefits of establishing a common currency (of lower transactions costs and greater certainty). However, through the empirical and theoretical study of monetary integration in Europe, this theory has been refined to take account of both political and economic calculations, how costs and benefits change with different levels of trade integration, and how different assumptions about the trade-off between inflation and unemployment change these calculations.

4.4 Regional Integration in Sociology

Although to a lesser extent than these other social science disciplines, regional integration has also become an important topic in sociology. For example, Soysal (1994) argues that supranational human rights norms—through the European Convention on Human Rights and the policies of the European Union—were as important as domestic factors in transforming the conception and operation of citizenship norms in Germany and France in the 1980s.

See also: Cultural Assimilation; Cultural Geography; International and Transboundary Accords, Environmental; International Law and Treaties; International Trade: Economic Integration; Regional Government; Regional Identity; Regional Science; Regionalism and Cultural Expression; Sovereignty: Political

Bibliography

Deutsch K W, Barrell S A, Kann R A, Lee M Jr, Lichterman M, Lindgren R E, Lowenheim F L, Van Wagenen R W 1957 *Political Community in the North Atlantic Area*. Princeton University Press, Princeton, NJ

Eichengreen B, Frieden J (eds.) 1994 *The Political Economy of European Monetary Unification*. Westview Press, Boulder, CO

Haas E B 1958 *The Uniting of Europe: Political, Social and Economic Forces, 1950–57*. Stevens & Sons, London

Haas E B 1976 *The Obsolescence of Regional Integration Theory*. Institute of International Studies, Berkeley, CA

Hix S 1999 *The Political System of the European Union*. Macmillan, London

Hoffmann S 1966 Obstinate or obsolete? The fate of the nation state and the case of Western Europe. *Daedalus* **95**: 862–915

Lindberg L N 1963 *The Political Dynamics of European Economic Integration*. Oxford University Press, Oxford, UK

Lindberg L N, Scheingold S A (eds.) 1971 *Regional Integration: Theory and Research*. Harvard University Press, Cambridge, MA

Moravcsik A 1991 Negotiating the Single European Act: national interests and conventional statecraft in the European Community. *International Organization* **45**: 19–56

Moravcsik A 1998 *The Choice for Europe: Social Purpose and State Power from Messina to Maastricht*. Cornell University Press, Ithaca, NY
Soysal Y N 1994 *Limits of Citizenship: Migrants and Post-national Membership in Europe*. University of Chicago, Chicago
Weiler J H H 1991 The transformation of Europe. *Yale Law Journal* **100**: 2403–83

S. J. Hix

Regional Planning, History of

1. The Term 'Regional Planning'

'*Gallia est omnia in partes tres*.' All Gaul is divided into three parts. This command, from the reign of Julius Caesar in the first century BC, no doubt is one of the earliest example of regional planning. A region is a district or area defined by generally recognizable attributes, physical cultural, economic, or political. The word 'region' derives from the Latin '*regio*,' meaning direction or boundary, and from '*regere*,' to direct. Thus the concept of region embodies the interaction among place, people, and work or economy. A region may or may not be formally recognized as a political entity or sub-division. More likely a region is simply a shared mental construct of its constituents.

Regional planning then is the effort of the society or elements in the society to intervene in the integrated developmental evolution of the area in question. In ancient kingdoms, the wishes and wisdom of the ruler constituted the regional planning conception for the society. Values and objectives underlying regional planning and foresight particularly included protection from external threats, but also the provision of infrastructure—roads, water, the enhancement and protection of productive fields and forests, provision for population growth and economic development activities such as industry and trade, cultural enrichment, and provision for societal continuity in succeeding generations. These values and practices were often codified in religious beliefs and scripts. These values still underlie and are used to justify contemporary efforts to define and carry out regional planning.

At this point it will be helpful to discuss a few related words or concepts. We need first to distinguish regional from city planning. The concepts are related but city planning implies a *polis* or civic entity governed by elected or other recognized official bodies. A region does not necessarily have such official status. Indeed, often it does not.

Regional analysis is the study of the interaction of key variables—physical, social and economic—in a regional context. Regional science, a relatively recent evolution founded upon economic location theory, seeks to describe and predict systemic regional phenomena.

We need also to distinguish types of regions. The European Community is a region comprised of nation states seeking to develop a commonality of purpose and policy, which presumably will result in superior outcomes for its constituent members. The European Union thus practices a kind of multi-state regional planning, much as the federal government in the United States does for the aggregate of its fifty constituent states. Parts of nation-states also may constitute regions. One thinks of Scotland and Wales in the United Kingdom or the Appalachian Region or the Great Plains of the United States.

Another distinction among regions is between mono-nodal, multi-nodal, and non-nodal regions. Metropolitan regions are examples of mono- and sometimes multi-nodal regions. Natural resource regions are often non-nodal. River basin regions are another category: linear regions.

Regions can also be distinguished as growth regions or as lagging regions. Regions may lag in development behind national averages, due to lack of natural resources, limited accessibility, or cultural or educational factors. Much attention has been given to assisting lagging regions to 'catch up,' with mixed results.

Today, we hear much about 'integrated regional resource management.' and 'sustainable development.' These concepts, while imprecise and difficult to measure, are increasingly shaping the directions towards which regional policy planning is moving.

We should also distinguish between 'regional planning,' and 'regional plan.' A regional plan is a specific document which lays out in physical detail a future desired end-state. It may also contain social and economic policy statements. The regional plan is typically based on a regional survey of social, economic, and physical factors. 'Regional planning' is the process by which regional plans are made. 'Regional policy' is a related concept. Regional policies are, in effect, guidelines to realize or implement regional plans.

2. Origins and Practice of Regional Planning

The origins of modern regional planning doctrine can be traced to reactions to the oppressive conditions created by the Industrial Revolution in nineteenth century England and France. Robert Owen and Charles Fourier proposed utopian schemes for alternative workers' housing and communities. Piotr Kropotkin's *Fields, Factories and Workshops* (1901) offered insights into the connections of people, land and labor. Charles Booth's *Life and Labor in London* (1892) systematically surveyed the conditions of work-

ing class people. Ebenezer Howard's 1902 *Garden Cities of Tomorrow* presented an influential alternative to unplanned metropolitan expansion. Patrick Geddes, a Scottish biologist, published his *Cities in Evolution* in 1915. Geddes' synthesis of the salient connections among people, land, economic activity, and natural systems greatly influenced the American essayist and critic, Lewis Mumford. Mumford, in turn promoted Geddes' integrative approach to social and public policy through such books as *The Culture of Cities* (1938). Through Mumford, Geddes' ideas influenced the Regional Planning Association of America (RPAA), a progressive private group concerned with development which was active in the 1920s and 1930s.

2.1 The Regional Planning Association of America

The RPAA embraced Ebenezer Howard's idea to siphon off and channel new growth at the suburban edges of existing large cities into planned garden cities. The group also espoused the notion that economic growth should be decentralized to smaller cities and towns and diverted from large metropolitan areas. The RPAA was especially influential in New York State, where Governor Al Smith appointed Stein and his fellow RPAA colleague, Henry Wright, to a newly-created Commission on Housing and Regional Planning. In 1926, the Commission issued a regional plan for the state that called for a concentrated grid of lines of transportation and communication along major corridors, separation of cities and highways, and the creation of independent communities away from the centralizing tendencies of its major cities. Protection of forest, farmlands and recreation areas were also proposed. The plan had considerable influence on policy and some of its proposals were ultimately carried out. What is notable is that these policies and issues are still under serious debate some 75 years later (Spann 1996).

The RPAA's conception of regional planning inevitably brought the group into conflict with the well-funded effort to create a Regional Plan for the New York metropolitan area begun in 1921 under the direction of Thomas Adams, a British town planner who had worked on plans for garden cities of Welwyn and Letchworth, and the Greater London region.

2.2 The Regional Plan of New York and Its Environs

The monumental Regional Plan of New York and Its Environs, begun in 1921 and published in 1929 (Committee on the Regional Plan of New York and Its Environs 1929). The plan brought together many of the development trends and proposals of the time and was underwritten by generous grants from the Russell Sage Foundation totaling some $1.2 million. In 1990, this would be equivalent to roughly $20 million. The Regional Plan and Survey was and probably remains the best funded metropolitan regional plan ever undertaken (Hays 1965).

The plan, encompassing 22 counties surrounding the Port of New York and targeted on the year 1965, incorporated proposals of the Port Authority rail plans, including those for belt line freight rail lines. Schemes for rapid-transit improvements, highway construction, zoning rationalization, and parks and parkways were assembled from a variety of sources and agencies and synthesized into a unified conception. The plan called for a 're-centralization' of industry and business, involving, among other proposals, a shift of manufacturing outward from Manhattan and the core of the region. Whatever its flaws, the 1929 Regional Plan of New York and Its Environs left a significant imprint on Metropolitan New York. The plan's basic physical planning conceptions were reflected in the subsequent development of the region.

The Regional Plan of 1929 clearly had a significant effect on the physical development of the region. But it also brought important changes in the institutional and public bodies concerned with regional development. Numerous planning and zoning agencies were established at the county and municipal levels as a direct result of field work by the staff of the Regional Plan. State policies for regional development were also stimulated by the plan and its advocates. The plan also influenced the Depression-era New Deal policies of the Roosevelt Administration. President Roosevelt was well-aware of the ideas of both the RPAA and the Regional Plan of New York.

Roosevelt spoke at a Roundtable on Regionalism held at the University of Virginia on July 11, 1931 on the need for regional planning and development, aided by the initiatives of the federal government to lead the country out of the economic stagnation it had found itself. Thus Roosevelt was instrumental in getting the thought of Southern Regionalists, such as Howard Odum and Rupert Vance of the University of North Carolina, and the ideas of the RPAA group onto the agenda of the New Deal. This may have been the origin of the idea for a Tennessee Valley Authority (TVA) in Roosevelt's thinking. On a national scale, the New Deal also created a National Resources Committee to survey the needs and resources of the country.

2.3 The Tennessee Valley Authority

The Tennessee Valley Authority (TVA) was the first of the river basin or other large regional agencies to attempt to bring together the then-new concepts of integrated resource development, which was central to the conservation ideas of the time, and regional planning which was a part of the new city and area planning movement.

Roosevelt's early speeches and papers clearly reveal a deep-seated and long-standing concern with farm problems, rural poverty, and the inadequate methods of providing services to rural areas. He also saw these problems as coupled with those of overcrowding and poor living conditions for the working people in the big cities, which he believed were fueled in part by the migration from farms to cities. The emerging concepts of regional planning seemed to Roosevelt to provide the linkage between the two problems. It is within this context that the regional planning ideas of Roosevelt were appended to earlier ideas for a multi-purpose plan for the Tennessee River, which had focused on flood control, navigation, and power generation. The main body of the TVA Act, as introduced on March 9, 1933, consisted of a comprehensive river improvement program plus two sections which provided for surveys of regional development problems and for recommendations to Congress on necessary legislation to solve these problems. The river improvement sections of the TVA Act were very explicit and provided for a navigation channel from Knoxville to the Ohio River, flood control to protect valley cities and to help control floods on the Mississippi River, and power generation. In contrast, the regional planning sections were very general and thus susceptible to either broad or narrow interpretation. TVA's mandate to conduct and implement regional planning was never clear or unambiguous.

The National Resources Committee (later the National Resources Planning Board) envisioned TVA as the first of a nationwide array of river basin planning regions. TVA was viewed as a model for the rest and as a testing ground for new communities as originally proposed by the RPAA group. The first of these was Norris, Tennessee, built near the first hydroelectric dam, Norris Dam, originally to house construction workers, and then to be converted to housing for the local populations dislocated by the rising waters of Norris Lake. Roosevelt's grand vision for river basin regional planning initiatives was thwarted by the interventionist scale of the vision, in a country dominated by unregulated market forces, and by the outbreak of World War Two. TVA nevertheless became a model for numerous grand multi-purpose hydro-electric and irrigation projects around the world, some misconceived and contrary to sound, integrated regional resource planning and management principles. Sadly, TVA itself has relinquished any serious commitment to regional planning leadership and has become, in effect, primarily a publicly-owned wholesale electricity producer.

2.4 Appalachian Regional Commission

Besides TVA, America's other important multi-state regional planning effort has been the Appalachian Regional Commission (ARC). Like TVA, the ARC was created to deal with an economically lagging region, the mountainous, relatively inaccessible chain of counties stretching from Alabama and Georgia to the Southern Tier of Upstate New York. The ARC began as an idea of President John Kennedy, who toured West Virginia's poverty-stricken mining towns in his campaign for the White House. The ARC was established in 1965 in the Johnson Administration. The ARC developed a plan for improving access and social services for isolated communities in parts of thirteen states. Highway improvement and infrastructure programs were central elements of ARC regional development. Thirty years later, assessments of the effectiveness of federal investments in the region are mixed. But according to one comprehensive study, the federal investment infusion seems to have been effective. Counties in the ARC program were noted on average to have developed faster economically than comparable counties not in the program (Bradshaw 1992, Isserman and Rephann 1995). ARC was still functioning in the year 2000 but other federal regional development efforts such as the eight Department of Commerce Title V regional commissions setup in other lagging regions of the US were not so fortunate. All were closed down in 1980.

2.5 Metropolitan Regional Planning

The post-war era in the United States, with its spreading suburbs and growing automobile dependency, raised new issues for regional planners, notably in the large metropolitan areas. Federal aid for metropolitan planning became available with the enactment of Section 701 of the Housing Act of 1954, stimulating the creation of planning commissions for the nation's metropolitan areas. By 1960, two-thirds of US metropolitan areas had some form of area-wide planning and, by 1971, thanks to federal grants and statutory requirements, all of the nation's 247 metropolitan areas had adopted official regional planning, mostly through elected, voluntary, Councils of Government (COGs). The A-95 review process, which required regional review and coordination to receive funding for large scale, federally-assisted projects, and Section 208 of the federal Clean Water Act of 1972 further stimulated the creation of regional planning bodies in metropolitan areas.

In the New York Metropolitan Region, the privately-supported Regional Plan Association produced a Second Regional Plan for the New York metropolitan area, released in November 1968. The Second Regional Plan emphasized the creation of regional subcenters linked by fast rail to Manhattan to expand job opportunities for those trapped in the central cities. For the suburbs, these centers would provide an organizing focus for social and civic life, employment opportunities, and easy access to and from Manhattan. The centers would be composed of 'regional activities' which would have to be built in any event to

accommodate the expected additional twelve million people who would join the 18 million already in the 31-county region (expanded from the original 22-county definition) by the year 2000.

Forecasts for a large shift in employment away from manufacturing to service and white-collar jobs suggested the potential for significant office growth, some of which could go to centers outside of Manhattan. The plan proposed massive new office development for Manhattan, both in downtown and midtown but the plan also proposed that regional centers be built, both as redeveloped inner-city sub-centers such as Downtown Brooklyn, Jamaica and Newark, as older city centers in suburban cities, or as entirely new centers on open land. The centers idea took hold in a number of places, notably Jamaica, Queens, Downtown Brooklyn, and Fordham Road in the Bronx. Downtown Newark also experienced new central place development and expanded centers were developed in the outlying areas of White Plains, Stamford, and New Brunswick.

The Second Regional Plan draft addressed other issues as well. Its housing policies called for increased densities, a more compact region, better related to accessibility to jobs and activities, and more housing opportunities for lower-income and minority families. The draft also called attention to the continuing need to capture more open space and to deal with the increasingly difficult problem of waste management. The plan called for completion of the region's share of the interstate highway network but it also made a strong case for the preservation and enlargement of existing rail transit systems.

In the 1980s, regional planning in the United States suffered a severe setback when the federal government withdrew virtually all its support for regional planning agencies, resulting in the demise of many programs. Only the transportation planning function survived in most metropolitan areas. TVA and the Appalachian programs managed to continue, though their planning functions were greatly reduced. Some states replaced federal assistance for regional planning with state funds. Other regional agencies survived by becoming more entrepreneurial, focussing on service delivery. A few states such as Florida, Georgia and Massachusetts, concerned about the negative aspects of growth, have developed growth management or 'Smart Growth' initiatives, which require regional planning as a framework for development policies, in some cases.

Regional cooperation in politically fragmented metropolitan areas has generally proved difficult to realize in the United States. A number of metro areas, however, have managed to achieve success in facilitating regional planning consensus. The Twin Cities region of Minneapolis/St. Paul, and the Portland Oregon region are often cited as examples of successful regional initiatives. The Minnesota Twin Cities area created in 1967 through state legislation a seven-county Metropolitan Council. The Council was created to address specific region-wide problems such as sewers and transportation. The Council has adopted a development guide for regional and local decision-making that identifies planned urban areas and growth centers. The Twin Cities regional initiative has enjoyed some major success, though its planning function has been scaled back in recent years.

The Portland Oregon three-county region in 1979 authorized by referendum a metropolitan service district. The district provides technical assistance and reviews local plans. In 1992, the Metro, as it is called, was given approval by charter to adopt a regional framework plan to guide regional development. The framework plan will address land use, development densities, housing, transportation, natural areas, and water supply. The Portland process is generally viewed as a successful example of the application of effective regional planning to control the negative aspects of rapid growth, though some concern has been raised about the impact of controls on the housing costs.

Metropolitan regional planning has experienced a resurgence in the 1990s in a number of American urban areas such as San Diego, California and Chattanooga, Tennessee, while, in the New York metropolitan region, the Regional Plan Association released in 1995 a Third Regional Plan, successor to the 1929 Regional Plan of New York and Its Environs, and to its Second Regional Plan of 1968 (Yaro and Hiss 1996). The Third Plan has focused on ways to help the region recover its momentum so as to maintain its faltering position in the global marketplace. To address the problems facing the region, the Third Regional Plan proposes five ambitious campaigns focussed on open and green space, transportation improvements, strengthening urban centers, improving workforce skills, and streamlining governance structures. It remains to be seen whether the Third Regional Plan will have as much impact as its two predecessors.

2.6 British and European Regional Planning

The trajectory of regional planning in Britain and Europe has somewhat paralleled that of the United States. Regional planning in post-World War II Britain found new support and momentum. The New Towns idea was re-embraced as a means to rebuild a Britain exhausted by war. Regional plans for the utilization of the old, lagging industrial regions of the North of England, Wales, and Scotland were adopted. These provided for decanting of populations into a new generation of expanded and new towns. Older urban areas would be contained within protected greenbelts.

In the 1960s, regional planning was further spurred by commission reports concerned about fast growing areas, notably in the Southeast of England. Metro-

politan regional planning, mainly advisory, was undertaken by county and regional councils. Their mandate was to:

(a) work out broad objectives for each region and provide a framework for detailed sectoral decision-making;

(b) advise on national policy-making where it could significant affect regions;

(c) advise on the application of national policy in their region; and,

(d) help create a consensus within a region as to its problems and possible solutions.

Regional planning tended to be more favored by the Labour Party, less so by the Conservatives. The Conservatives tended to view taxation for social schemes and infrastructure as a burden on business and regulation and control of development as disincentives to market responses to problems of growth. The Thatcher government in the 1980s reconfigured regional planning bodies and responsibilities: the Greater London Council and other regional councils were abolished and regional planning was transformed into a mechanism for assisting large-scale private redevelopment schemes.

With the return of a more liberal-left government in the 1990s, there has been a modest return to regional planning initiatives. These initiatives, however, have continued the private/public partnership approach promoted by the Thatcher government. The central government places more emphasis on local government bodies. Grand regional planning schemes are not in vogue, as pragmatism and incremental and market approaches are more the order of the day. Whether this approach will enable the UK to meet the challenges of the first decades of the new century remains to be seen. With continuing population growth, a deteriorating environment, limited local resources, and continued un- and under-employment in some regions, together with a need for further integration into the European system, Britain faces hard choices. Robust, effective regional planning and implementation intermediate between the central government and local governments would appear to be a prerequisite for meeting these challenges. Whether the political climate will permit and support an articulated interventionist approach remains to be seen.

On the European continent, regional planning has been practiced with considerable success in both France and the Netherlands. France, with its strong provincial history and identities, has been able to focus its regional planning efforts on channeling excessive growth away from the Paris region towards second- and third-order city-regions. Meanwhile, the Paris region has benefited from metropolitan region planning and development that has focused on transit-based development and suburban centers. Regional planning in the Netherlands has emphasized a '*Randstadt*' plan which focuses development on the ring of cities surrounding a protected green center in the middle of the country. The European Union has fostered regional cooperation among regions that share common characteristics. An example is the Atlantic Arc Region, the grouping of the regions of EU member country fronting the Atlantic Ocean, which share technologies and cooperative approaches to managing planning and development.

3. *Future Prospects for Regional Planning*

To summarize, it is reasonable to say that regional planning in western industrialized countries has historically had a checkered history of meeting the challenges of growth and decay at the sub-national level. But it is also accurate to point out that those areas that adopted and tried to carry out regional planning schemes have often achieved better results than those which simply drifted into incremental, piecemeal development. It is almost always an advantage to have some systematic framework and political consensus on which to build, even if all relevant variables cannot be predicted or controlled.

There is reason to believe that regional planning will enjoy new vitality and acceptance. Powerful new forces, needs, and demands are compelling a renewed look at the desirability of regional planning approaches. Globalization trends are fostering the integration of economies on an international basis. Metropolitan economies are increasing seen as the building blocks of a world trading system. The communication and air travel revolutions have reduced the accessibility advantages enjoyed by a few large cities. Competitiveness will require regions to make the most of their assets, and grow new ones. This will take conscious, deliberate, smart planning. Environmental pressures and the deterioration of air, water, and land have reached serious proportions in many countries. Regional approaches will be needed to address the pollution of air and water sheds, which rarely conform to the boundaries of local jurisdictions. Global warming and the need to reduce carbon emissions will inevitably require regional approaches to the transportation and land-use interface. Sustainability is becoming the norm by which regional development decisions increasingly are being judged. More attention is being paid to the protection of natural and cultural amenities, natural habitat, green areas, wetlands and waterways, which, by their scale, mandate regional approaches. At the same time, the digital revolution has provided policy-makers and the public with new tools to collect, manage, and display vast amounts of relevant data, in ways only dreamed of in the early years of regional planning. Powerful computer-based geographic information systems (GIS) now permit mapping and analysis of trends, projection of impacts, and improved understanding of alternatives by regional bodies. This new capacity will surely enhance regional planning possibilities and effectiveness in many areas and countries.

See also: Planning Issues and Sustainable Development; Planning Theory: Interaction with Institutional Contexts; Regional Geography; Regional Integration; Resource Geography; Urban Planning: Central City Revitalization; Urban System in Geography

Bibliography

Boardman P 1978 *The Worlds of Patrick Geddes: Biologist, Town Planner, Re-educator, Peace-warrior*. Routledge & Kegan Paul, London

Booth C (ed.) 1892 *Life and Labour of the People of London*. Macmillan, London

Bradshaw M 1992 *The Appalachian Regional Commission: Twenty-five Years of Government Policy*. The University Press of Kentucky, Lexington, KY

Committee on the Regional Plan of New York and Its Environs 1929 *Regional Plan of New York and Its Environs*. The Committee on the Regional Plan of New York and Its Environs, New York

Cullingworth B 1999 *British Planning: 50 Years of Urban and Regional Policy*. Athlone Press, Somerset, NJ

Downs A 1994 *New Visions for Metropolitan America*. Brookings Institution and the Lincoln Institute of Land Studies, Washington, DC

Friedmann J, Weaver C 1979 *Territory and Function: The Evolution of Regional Planning*. University of California Press, Berkeley and Los Angeles

Geddes P 1915 *Cities in Evolution*. Williams and Norgate, London

Hansen N 1968 *French Regional Planning*. Indiana University Press, Bloomington, IN

Hays F 1965 *Community Leadership*. Columbia University Press, New York

Isard W 1960 *Methods of Regional Analysis*. John Wiley & Sons, New York

Isserman A, Rephann T 1995 The economic effects of the appalachian regional commission: An empirical assessment of 26 years of regional development planning. *Journal of the American Planning Association* **61**(3): 345–64

John D et al. 1998 *Building Stronger Communities and Regions: Can the Federal Government Help?* National Academy of Public Administration, Washington, DC

Johnson D A (ed.) 1996 *Planning the Great Metropolis: The 1929 Regional Plan of New York and its Environs*. E & F Spon, London

Katz B (ed.) 2000 *Reflections on Regionalism*. Brookings Institution, Washington, DC

Kropotkin P 1901 *Fields, Factories and Workshops*. S. Sonnenschein, London

MacFadyen D 1970 *Sir Ebenezer Howard and the Town Planning Movement*. MIT Press, Cambridge, MA

MacKaye B 1962 *The New Exploration: A Philosophy of Regional Planning*. University of Illinois Press, Urbana, IL

Meck S (ed.) 1999 *Planning Communities for the 21st Century*. American Planning Association, Chicago

Mumford L 1938 *The Culture of Cities*. Harcourt Brace Jovanovich, New York

So F, Hand I, McDowell B D (eds.) 1986 *The Practice of State and Regional Planning*. International City Managers Association and the American Planning Association, Chicago

Spann E K 1996 *Designing Modern America: The Regional Planning Association of America and its Members*. Ohio State University Press, Columbus, OH

Wannop U A 1995 *The Regional Imperative: Regional Planning and Governance in Britain, Europe, and the United States*. Jessica Kingsley Publishers, London

Yaro R, Hiss T 1996 *A Region at Risk: The Third Regional Plan for the New York, New Jersey, Connecticut Metropolitan Area*. Island Press, Washington, DC

D. A. Johnson

Regional Science

Economists pay too little attention to the location of economic activity and the effects of distance on economic systems. Geographers focus on the spatial dimension but lack rigorous tools to study the effects of location and distance on human behavior or explain why activities are located where they are. Urban and regional planners need a stronger analytical foundation to go beyond the physical and aesthetic aspects of urban form to consider social and economic dimensions of city and region. With such premises, 60 economists, geographers, political scientists, sociologists, and planners gathered at the annual meeting of the Allied Social Science Associations in 1954 and founded the Regional Science Association. Its objective was 'to promote studies focusing on the region.' Envisioned were the use of theoretical frameworks and methods of the social and other sciences and the development of new ones specifically for regional analysis.

This interest in the region as a subject of study had important and immediate political and intellectual antecedents in the United States. In 1927 the regional plan for New York considered economic factors in metropolitan growth and spatial form, largely through the work of Robert Haig. In 1933 the Tennessee Valley Authority became the nation's first major regional development program, arguably since the western expansion. In 1935, asserting that policies of national development depend on regional factors, the US National Resources Committee asked 'modern students of regional science' (10 regional geographers and two regional sociologists) to identify the best type of region to use for regional planning and development. In 1942, Alvin Hansen, perhaps the leading American Keynesian economist and a Harvard professor, advocated a comprehensive postwar expansionary program of regional development on the premise that the effective utilization of resources in any region makes the whole nation richer.

Two organized, interdisciplinary efforts with very different intellectual roots had focused on regions. In the 1920s Lewis Mumford and his colleagues in the Regional Planning Association of America saw cities

as congested and unmanageable. They advocated a mode of thinking and procedural methods aimed at seeing the region as a whole—integrating city and region, and utilizing, not destroying, the natural environment. Closer to the regional scientists because of their emphasis on both science and the region and their scholarly, university base were the southern regionalists. Centered at the University of North Carolina, they were well established by 1931 in terms of research achievements, agendas, and infrastructure. Howard Odum, the acknowledged leader of the effort and a sociologist, argued in 1945 that the nation owes the fullest development of the South not only to the South but also to the nation itself. Similar in ambition but predating the Regional Science Association by a quarter century, Odum advocated a regional approach to the scientific study of human society, arguing that it would entail the integration of method and theory, require the joint effort of the traditional social science disciplines, and create fundamental new understanding of all society, not just its regions.

Members of the Regional Science Association from the outset debated whether regional science would become a new discipline or an interdisciplinary activity. Walter Isard, widely recognized as the father of regional science, was the leading proponent of the disciplinary view. In 1956, he became an economics professor at the University of Pennsylvania, where he formed the Department of Regional Science in 1958. By then he had published 23 papers in the leading economics journals (acknowledging the help of Alvin Hansen in the first five). Yet the Regional Science program would go far beyond economics in its scope and impact. In 1960 the first Ph.D. in Regional Science was awarded to William Alonso, later regional planning professor at Berkeley and Saltonstall Professor of Population Studies at Harvard. Subsequent doctoral recipients took positions in economics, engineering, geography, planning, and other university faculties as well as public agencies and consulting firms.

Isard's prediction that regional science programs would emerge from geography departments and achieve wide recognition before 2000 was not realized. In the 1970s Cornell University established the second Regional Science Ph.D. program, and the University of Illinois initiated a multidepartmental program requiring completion of the Ph.D. in a participating department. By 1976, regional science degree programs or specializations existed at 14 universities internationally. The lone department devoted exclusively to regional science remained the one at Pennsylvania until it was closed during university retrenchment in 1994.

Among the early challengers of the disciplinary view was Lloyd Rodwin, a planning professor at the Massachusetts Institute of Technology. He envisioned the social sciences developing more systematic spatial studies, geography becoming more analytical, and regional science becoming 'a generic name for regional studies.' This vision has become reality, but not fully. Regional science has become a generic name for a certain kind of regional studies, one whose origin is in mathematical abstraction and quantitative modeling.

Today regional science is an international, interdisciplinary endeavor. It brings together economists, geographers, planners, and others whose research has a regional dimension. It is the confluence of many streams, but arguably also the origin of a few. Each discipline brings different theoretical and methodological approaches to regional science, and each typically has different substantive concerns within a given topic. Migration, for example, interests several disciplines. An economist might focus on migration behavior as a way of measuring quality of life in cities, a demographer on regularities in the age pattern of migration, a geographer on turnarounds in core–periphery migration, a planner in forecasting migration for land use and transportation modeling, a sociologist on the effect of kinship links on migration destinations, and a real estate specialist on the relationship between housing prices and migration.

The numerous scholarly regional science journals and annual conferences create effective forums for exchanging ideas among disciplines. Although dominated by the abstraction and modeling inherited from economics, these forums are eclectic and open. They have been helpful incubators for scholarly groups who went on to create their own forums, be they input–output analysts, spatial social theorists, urban economists, or transportation modelers. Collectively regional scientists actively borrow ideas from a wide range of sources and combine, adapt, and extend them to regional research.

The recognition Isard anticipated has materialized in many ways other than through a proliferation of university departments. The Regional Science Association became the Regional Science Association International (RSAI), an umbrella organization overseeing three major supraregional organizations, the North American, European, and Pacific coordinating councils. Most countries in Asia, Europe, and North America have active regional science associations, and there are several multinational ones, including the British–Irish, French-language, German-speaking, and Australia–New Zealand organizations. Most sponsor annual meetings, and several have journals.

A widely accepted notion is that almost all regional scientists have dual identities, economist and regional scientist, geographer and regional scientist, and so on. For many, the regional science meetings and journals provide the best audience for honing and distributing their research. Others' participation ebbs and flows with changes in their immediate research agendas. No complete census of this fluid regional science community exists. The associations, including those in China, India, and Brazil, may have over 10,000 members combined. RSAI itself has about 2,000 dues-paying members annually.

Regional science journals have become numerous and well established. Five international ones with regional science in their titles are now published by major commercial presses: *Annals of Regional Science*, *International Regional Science Review*, *Journal of Regional Science*, *Papers in Regional Science*, and *Regional Science and Urban Economics*. Echoing the debate a half century ago about the name 'regional science,' some regional science associations leave it out of their journal titles, for example, *Journal of Regional Analysis and Policy*, *Review of Urban and Regional Development Studies*, or *Review of Regional Studies*.

Book series, position announcements, honorary degrees, and external research prizes are other marks of regional science's role within the social sciences. Noteworthy, too, are the National Science Foundations program in Geography and Regional Science in the US and similar ones in other countries. Perhaps the most spectacular recognition of regional science occurred when the Emperor and Empress of Japan personally welcomed participants to the RSAI World Congress in 1996.

1. Subject Matter

The international regional science community studies a wide range of topics, sometimes ambitiously defined as those social, economic, political, and behavioral phenomena that have a spatial dimension. A common, narrower view groups regional science research under three themes: location theory, methods of regional and spatial analysis, and regional development and policy.

Location theory is the study of what social and economic activities are located where and why. Traditionally, emphasis has been on the location of firms and households. The latter considers inter-regional and international migration and labor markets, as well as intrametropolitan residential mobility and housing markets. Early contributions focused on the role of transportation costs and proximity to suppliers and markets. Recognizing interdependencies of behavior focuses attention on agglomerations or clusters in space, as well as questions of land use, urban form, housing markets, city size, hierarchies among and systems of cities, market thresholds, and market boundaries. Recognizing interactions among activities and individuals distributed over space raises questions about commuting, trade flows, congestion, transportation, other infrastructure, facility location, taxes, user fees, local public finance, and diffusion of innovations and growth.

The phrase, methods of regional analysis, describes a growing set of techniques used to identify, analyze, or forecast key regional characteristics or changes. Among the methods are three invented to study national economies but adapted to regions. The inventors received Nobel Prizes in Economic Sciences: Wassily Leontief (1973) 'for the development of the input–output method and for its application to important economic problems,' Lawrence Klein (1980) 'for the creation of econometric models and their application to the analysis of economic fluctuations and economic policies,' and Sir Richard Stone (1984) 'for having made fundamental contributions to the development of systems of national accounts.' Regional scientists have adapted and extended techniques from demography. The cohort-component model focuses on the components of population change and their probabilities (births, deaths, and migration), grouping the population into cohorts by age, sex, and often race or ethnicity. Regional scientists borrowed the gravity and entropy models directly from physics and found that they described certain aggregate human behavior well. Migration, trade, transportation flows, and other interaction between two places tend to increase with the product of their masses (e.g., populations) and decrease with distance between them. Regional scientists have also adapted and applied optimization techniques to contexts in which spatial interaction is important, such as transportation networks, location of facilities, and land use.

Regional scientists use statistical methods extensively. Yet here, too, important adaptations and extensions permit the study of spatial and regional phenomena. Geographers, in particular, have extended standard statistical concepts to situations where observations have locations, for example, to draw random samples over space or to determine the likelihood that an observed locational pattern could have occurred by chance. Spatial statistics, spatial econometrics, geographical information systems, and other methods of spatial statistical analysis have become major parts of regional science—for economists, planners, engineers, and others as well as geographers. In the case of statistical methods widely used to study travel demand, residential location, migration, and other behavior with a spatial dimension, the pioneer was an economist, Daniel McFadden, who won the Nobel Prize in 2000 'for his development of theory and methods for analyzing discrete choice.'

Regional development and policy focuses on why some regions grow and prosper while others lag and even decline. Research ranges from theories of growth and development to evaluation of specific policies. Persistent questions deal with the divergence and convergence of regional growth rates and incomes, the relationships between core regions and peripheral regions, and the geography and timing of the spread of growth. Other long-standing questions deal with regional labor markets, including the relationship between growth and unemployment and the persistence of wage differences among regions. Key policy issues concern people prosperity vs. place prosperity (should the focus be on helping people directly, for example, through education or out-migration incen-

tives or on helping people indirectly by making the region more competitive through infrastructure, tax incentives, or other actions), ways of making lagging regions more competitive and prosperous, the rationales for regional policy, and the roles of different levels of government.

Policy evaluation has attracted much attention as well. Commonly the objective is to measure the effects of various (nonregional) national policies on regions, for example, free trade treaties, defense spending, or tax reductions. Other times the objective is to evaluate regional policies, for example, by estimating what would have happened to lagging regions in the absence of the regional support programs. Sometimes the emphasis is on comparing alternative regional plans or policy instruments before they are implemented, generally using cost-benefit, cost-effectiveness, or goals achievement methods augmented to include policy objectives, distributional issues, and hard to quantify outcomes.

2. *Intellectual Progress*

Two easily contested, unavoidable decisions must be made before discussing the intellectual progress of regional science. The first is whose contributions should be counted. The open, fluid nature of the regional science community makes this decision difficult, and the early claims to disciplinary status make it controversial. The perspective here is that the idea itself is more important than the primary tribal self-affiliation of its originator. The key criterion is whether the idea is part of the scholarly conversation in regional science meetings and journals. Mention of an individual's contribution does not imply that the person considers him or herself a regional scientist, although each person mentioned has participated in regional science meetings or published in regional science journals. Debates about whether particular intellectual contributions were initiated by regional scientists, borrowed by regional scientists, or adapted by regional scientists, and who is really a regional scientist are fruitless. The gravity model provides an early case in point. One observer described it as work that may clearly be termed regional science, while another argued that it had been pioneered by outsiders and applied in many disciplines. By far the most interesting fact here is how understanding and use of an idea have evolved.

The second decision concerns the scope of regional science. Something narrower than human activities and phenomena with an important spatial dimension must bound this discussion, but many offer too incomplete a view. Some describe regional science as a toolbox for practical analysis that lacks rigorous theory, while others decry it as a theoretical never-never land with too little realism. The path chosen here recognizes Isard's deliberate efforts to nurture the field of regional science. He edited several of its journals, commissioned papers, supported dissertations, organized meetings, and shaped an influential book series. Three of those books frame a useful way of discussing regional science: Isard's *Location and Space-Economy* (1956), the Isard-led collaborative effort, *Methods of Regional Analysis* (1960), and a reader edited by John Friedmann and William Alonso and dedicated to Isard, *Regional Planning and Development* (1964). The books' foci—theory, method, and policy—frame the discussion of intellectual progress well.

2.1 *Location and the Spatial Economy*

The quest here is to model mathematically and show graphically both the determinants and consequences of individual location decisions. Regularities in spatial patterns are readily observable, among them, the emergence and persistence of large cities, clusters of related industries, relatively homogenous residential areas within cities, and regularities in city size distributions. The goal, chiefly of economists who work in this tradition, is to show how these agglomerations materialize from the interaction of self-serving decisions of individuals and firms. Isard's book drew on the pioneering German contributions by Johann Heinrich von Thünen, Alfred Weber, August Lösch, and Walter Christaller.

Two works from the birth period of modern regional science were extraordinarily seminal. Alonso's dissertation, *Location and Land Use*, focuses on transportation costs and its role in residential decisions and the resultant land use patterns within a metropolitan area. Concentric rings of land use emerge from the behavior of utility maximizing individuals. Alonso's starting point like von Thünen's is the existence of an agglomeration, in this case the central business district of a monocentric city. Charles Tiebout, a founding member of the Regional Science Association, argued that 'people vote with their feet' by choosing locations within the metropolitan area that offer the combination of revenue and expenditure patterns that appeal most to them. Thus, spatial mobility among diverse communities within an urban area provides a solution to the collective consumption problem posed by Samuelson and another mechanism by which individual preferences create spatial variation. Much of today's field of urban economics builds on these two contributions.

A contemporary contribution of this genre is *The Spatial Economy* (1999) by Masahisa Fujita, Paul Krugman, and Anthony Venables, which they describe as a continuation and possible validation of Isard's project. It derives agglomerations instead of positing them as a starting point (e.g., Alonso's central business district around which commuters organize themselves). In the 'new economic geography,' as Krugman

calls it, agglomerations result from the tension between centripetal forces that promote concentration of economic activities and centrifugal forces that oppose it. Of key interest is when a spatial concentration of economic activity is sustainable, and when a system without spatial concentration breaks down and concentration emerges. Combining formal theoretical models that include increasing returns to scale and computer simulations to study nonlinear dynamic properties, this work links urban, regional, and trade economics in a way that exceeds the dreams of the formal modelers of a half-century ago. As was the case then, however, many geographers and others within the regional science community do not embrace this formal mathematical, theoretical mainstream economics approach. They are far more interested in the social, cultural, and institutional factors involved in spatial development and in the study of real places and real regional histories in contrast to abstract ones. The new *Journal of Economic Geography* hopes to encourage dialogue among the groups.

2.2 Methods of Regional Analysis

A very different early quest of modern regional science was the creation of practical tools for the study of real places, and here regional science has had spectacular success. The input–output model attracted early attention. A series of regional input–output articles appeared in the *Review of Economics and Statistics* between 1951 and 1971. Their foci ranged from the theoretical specifications of the models and the definitions of concepts to studies of the nature of metropolitan economies, the location of a major steel complex, and the effects of arms reduction. This initial period culminated in the creation of two landmark regional input–output models constructed from survey-based data, the Philadelphia model of Isard and Thomas Langford and the West Virginia model of William Miernyk. The next three decades focused on building and using models, with attention paid to the treatment of households and consumption, economic–demographic linkages, and strategies for building the models in a cost-effective way. Here too progress exceeded the dreams of the 1950s, with regional input–output models now widely used for impact studies in the private and public sectors in many countries of the world. In fact, this method of regional analysis has become an industry with data and software readily available from commercial firms, most notably the IMPLAN system within the United States.

Econometric modeling similarly saw astonishing progress from the efforts of pioneering scholars to the commonplace of professional practice. Lawrence Klein sketched out how regional econometric models might be developed as satellites of national models, and his student Norman Glickman developed a fully operational landmark model of the Philadelphia economy. For roughly a decade, many others built regional econometric models as doctoral dissertations, one becoming the blueprint for the National Regional Impact evaluation system of the US Department of Commerce. Today most states and many companies maintain such forecasting models. This method, too, has become an industry with many companies selling forecasts to private subscribers. A particular noteworthy case is the eclectic REMI model of George Treyz, which grew from a design at a regional science meeting in 1978 into a widely used commercial product.

Modeling land use and transportation flows within a metropolitan region proved to be arguably the most ambitious search for practical methods, and the most spatially explicit. Its intellectual scope is daunting: understanding the determinants of the location of industrial, commercial, residential, and other activities and of the movement of people among those locations. It combined five elements: the theoretical urban economic framework of Alonso and Lowdon Wingo, mathematical programming methods to make them operational starting with the Herbert–Stevens model of land use, gravity models of spatial interaction eventually enriched by the entropy formulations of Alan Wilson, discrete choice statistical methods eventually incorporating McFadden's work, and the transportation forecasting framework established by Robert Mitchell and Chester Rapkin. For four decades, university research groups, notably those led by Britton Harris, Alan Wilson, and David Boyce, have produced a sustained flow of dissertations and other contributions. Yet, unlike the cases of input–output and econometric models, the pioneering research phase did not lead to widespread adoption in routine public and private practice. Dominant practice remains a set of incremental steps—predict population and employment change, decide (usually unaided by models) where those activities will locate, and then use transportation modeling software to forecast the flows among the locations. The research quest continues, large-scale operational models continue to be built (about a dozen were in use in the early 1990s), and the research leads to improvements in the transportation modeling software and other elements of the incremental approach.

2.3 Regional Development

The basic quest here is to understand differences in regional growth and development. During the early years of modern regional science it attracted the attention of two more Nobel Laureates in Economics. Gunnar Myrdal was honored in 1974 for his 'penetrating analysis of the interdependence of economic, social and institutional phenomena.' His work on economic theory and underdeveloped regions centered

on the principle of cumulative causation, a 'vicious circle' of economic and noneconomic factors that led activities to cluster in certain regions and caused regional inequalities. Douglass North was honored in 1993 for 'for having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change.' He studied regional economic growth using exportable commodities or the export base as the key concept, pointing out the incentives regions have to reduce their processing and transportation costs to better the competitive position of their exports. An influential book by Michael Porter, *The Competitive Advantage of Nations* (1990), builds on these notions of clusters of activities and competitiveness of export industries. He argues that national advantage derives from successful firms concentrating geographically within a country, their competitive success stimulated by their local environment, including government policy at the regional and local level. An earlier formulation remains popular in regional policy. Based on the growth pole research initiated by Francois Perroux, governments concentrate investment in selected urban centers to stimulate the economies of their surrounding areas or draw population away from the crowded, larger metropolitan areas.

2.4 Other

Brief mention of some other work completes a broader picture of regional science. RSAI has awarded its Founder's Medal to Martin Beckmann for research in transportation network theory, location theory, and spatial economics; William Alonso for urban economics, spatial demography, and regional development; Jean Paelinck for methods of regional analysis, spatial econometrics, and regional policy; and David Boyce for transportation network models and urban transportation planning methods. The three most cited articles in the *International Regional Science Review* are on population dispersal from major metropolitan regions internationally (D. Vining 1978), intra-urban residential mobility (J. Quigley and D. Weinberg 1977), and the shift of the US population to nonmetropolitan areas (C. Beale 1977). In the *Journal of Regional Science* they are on the distribution of residential activity in urban areas (J. Herbert and B. Stevens 1960), determinants and consequences of migration (M. Greenwood 1985), and an efficient algorithm to solve the Weber location problem (H. Kuhn and R. Kuenne 1962). In *Regional Science and Urban Economics*, they are on the baby boom and the housing market (N. Mankiw and D. Weil 1989), the demand for housing characteristics (J. Follain and E. Jimenez 1985), and publicly provided inputs and economic growth (T. Garcia-Mila and T. McGuire 1992). The supplementary Bibliography includes well-cited books and recent overviews of aspects of regional science.

See also: Cities, Internal Organization of; Ecological Economics; Location Theory; Migration, Economics of; Region: Geographical; Regional Geography; Space and Social Theory in Geography; Spatial Interaction Models; Spatial Pattern, Analysis of; Transportation Geography; Transportation Planning; Transportation: Supply and Congestion; Urban System in Geography

Bibliography

Anselin L 1988 *Spatial Econometrics*. Kluwer Academic Publishers, Boston

Fujita M 1989 *Urban Economic Theory*. Cambridge University Press, Cambridge, UK

Friedmann J, Weaver C 1979 *Territory and Function*. University of California Press, Berkeley, CA

Hansen N, Higgins B, Savoie D J 1990 *Regional Policy in a Changing World*. Plenum Press, New York

Henderson J V 1985 *Economic Theory and the Cities*, 2nd edn. Academic Press, Orlando, FL

International Regional Science Review 1995 **17**(3) and **18**(2) (Regional Science at 40 issues)

Isard W, Azis I J, Drennan M P, Miller R E, Saltzman S, Thorbecke E 1998 *Methods of Interregional and Regional Analysis*. Ashgate Publishing, Aldershot, UK

Journal of the American Planning Association 1994 **60**(1): 3–44 (Large Scale Models symposium)

Journal of Regional Science 1985 **24**(4) (Silver Anniversary issue)

Krugman P 1991 *Geography and Trade*. MIT Press, Cambridge, MA

Lahr M, Miller R (eds.) 2001 *Regional Science Perspectives in Economic Analysis*. Elsevier, Amsterdam

Markusen A 1985 *Profit Cycles, Oligopoly, and Regional Development*. MIT Press, Cambridge, MA

Miller R E, Blair P D 1985 *Input–Output Analysis Foundations and Extensions*. Prentice-Hall, Englewood Cliffs, NJ

Nijkamp P (ed.) 1986 *Handbook of Regional and Urban Economics*. North Holland, Amsterdam

Plane D A, Rogerson P A 1994 *The Geographical Analysis of Population*. Prentice-Hall, Englewood Cliffs, NJ

Wilson A G 1970 *Entropy in Urban and Regional Modeling*. Pion Press, London

A. Isserman

Regionalism and Cultural Expression

A region is a geographic concept that denotes a set of places understood as having something in common. While this common ground is typically geographic, political, and/or economic, it gives rise to shared forms of cultural expression as well; conversely, shared culture may encourage political or economic linkage, even perceived geographic unity. Regional culture refers to these shared cultural patterns most generally, while regionalism is reserved for a highly favorable evaluation of the culture of place. Research has

addressed questions about regionalism's political implications, its persistence under conditions of modernity, and its capacity to ground identities under conditions of postmodernity.

1. Definitions

When people refer to a region, they are indicating a set of places that they view as having some common characteristic. Regions can be specific to a small group or organization; for example a regional sales manager might cover a specific territory that no one outside the firm would think of as a region. They can be individual, as in 'the region where I grew up.' And they may involve an exquisite local knowledge. This essay concerns regions having a wider application than these.

Usually the places included in a region are contiguous; if they are not, the plural form gets used, as in the 'polar regions.' The basis for defining a region may be purely geographic (the Mediterranean region) or historical (the Spanish empire), but the places within any single region usually come to share more than geography or history. Regions are or tend to become economic (the Ruhr (*Ruhrgebiet*)), ethnic (Yorubaland in West Africa), or cultural (the Bible Belt of the American South). Political boundaries may or may not contain a region, which can include all or parts of different nation states. Bound by their common characteristics, regions are often at odds with the political unit(s) that claims sovereignty over them (e.g., the Kurdish region of Turkey and Iraq).

Even in the case of regions defined strictly by geography or political necessity, some degree of common culture emerges. Climate, topography, and relationships to the sea will influence agricultural and commercial activities. Thus, in the most obvious illustration of this point, political antagonists like the Palestinians and Jews of the Middle East prepare similar food in similar ways. For this reason scholarship on such regional forms of cultural expression once took them as natural, the straightforward products of material circumstances and history. Such shared characteristics are weak versions of regional culture.

Cultural regions typically derive from a preexisting geographic, economic, or political region, but in some cases the cultural common ground can come first. Nations that share the same language become trading partners, for example, and elaborate other ties; even here, however, there is usually the prior political connection of colonialism. Religion would offer another example of this potential, for example in the Organization of the Islamic Conference. Stretching from Indonesia to Nigeria, this is hardly a territorial region but is sometimes envisioned as such, and it constitutes the potential for economic and political links.

Regional culture comes in a strong version as well. Geographer D. W. Meinig (1986) defined regional culture as, 'that which is characteristic of a group of people who are deep-rooted and dominant in a particular territory, who are conscious of their identity as deriving from a common heritage, and who share a common language and basic patterns of life' (p. 80). He contrasts the seventeenth-century Welsh region, which shares geography, ethnicity, language, and a sense of self-identity, with the West Country, a much vaguer region with its own traditions, dialect, and ancestral families but no unmistakable 'ethnic' distinctions. The West Country was a 'relatively subdued form of regional culture, one which its members are conscious of but not normally militant about, though strong feelings of identity may be aroused under special circumstances' (p. 81).

In light of this distinction between strong and weak regional cultures, manifest versus latent, regionalism is a strong and self-aware form of regional culture, and sectionalism is an intolerant and divisive form of regionalism. The term regionalism generally has a positive connotation, referring to things that win local and often external approval; other strong but destructive manifestations of regional culture, such as the vendetta tradition in Corsica, are not usually designated as regionalism. Regionalism aggressively promotes its own symbolic expression, which can reinforce, or even generate, social and political bonds.

2. Real Regionalism, Debatable Impact

Considerations of regionalism used to begin from a bimodal model: urban culture versus folk culture; cosmopolitan versus provincial; center versus periphery. In this view regionalism was often conflated with localism, the residual cultural forms that hegemonic centralized culture had not yet obliterated. Regional culture, although it made meaningful the lives of people in the region and perhaps enriched the larger cultural pool, was one of those things that was likely to fade away under industrial capitalism and centralized political authority. Research mapped (and sometimes celebrated) regional differences, and it also addressed the impact of these cultural differences on the larger political units.

For example, in the 1930s, sociologist Howard W. Odum argued that the American South remained backward because it lacked social institutions that could facilitate social change while protecting traditional culture; the pro-business New South could not reconcile change with continuity. Odum concluded that the South must reinvent its debilitating local culture by blending it with national culture into regionalism, whereby (in his somewhat idiosyncratic definition) the local was part of the larger unity. In Odum's view, sectionalism was a pathological form of localism, the result of cultural inbreeding (Odum and

Moore 1938, p. 43), although many contended that this regional/sectional distinction was hard to maintain. Odum saw literary and aesthetic regionalism as an undeniable fact—he catalogued over two thousand regionalist titles from two decades of *Publishers' Weekly*—but rather than local color provincialism, which would smack of sectionalism, he maintained that 'regional portraitures are only fundamental units in the great American literary fabric.' True regionalism, in other words, contributes to the health of both region and nation.

As the New South has become both established and even normative on a national level, the regionalism/sectionalism distinction has faded. Social scientists like John Shelton Reed (1986) have continued to assess the degree of Southern regional variation that persists. Analyzing the economic and political differences among a nation's regions is a standard activity on the part of both social scientists and the popular press.

A contemporary political scientist who has made a strong case for the political impact of regional culture is Robert Putnam (1993). Using data from Italy's twenty administrative regions, Putnam shows how 'institutional performance'—effectiveness of and satisfaction with regional government—is high in Italy's north and low in the south. He connects this with presence or absence of civic traditions, which he traces back to medieval period; 'the astonishing tensile strength of civic traditions testifies to the power of the past' (p. 162). Why is the past so powerful? Putnam draws on theories of social capital to show how norms, whether of 'never cooperate' or of trust and reciprocity, produce stable cultures, with only the latter—'civic equilibrium'—able to solve the collective action problem. Culture and structure reinforce each other.

Sociology and urban studies have renewed attention to place as a variable, as in Douglas Massey and Nancy Denton's study of the devastating impact that inner-city spatial isolation—'American apartheid'—has on African–Americans (see also Gieryn 2000). Ethnographers study the symbolic and aesthetic worlds such conditions produce, as when Elijah Anderson analyses the intricate set of signs and behaviors that black youth must master to steer between trouble with the police and with their peers (E. Anderson 1999). While this type of research is often at the level of neighborhood, some employ a broader, regional canvas.

A good example of regional cultural analysis is Francis Violich's 'urban readings,' in which he offers a phenomenological account of the intersection of ecology, urban functions, and cultural history on the Dalmatian coast, now in Croatia (1998). Ecologically, the region possesses rich sea resources along with a poor, mountainous land where the limestone foundation makes water collection a societal preoccupation. Isolation and the need to manage their water supply and harbors has produced strong urban cultures, independent of one another and of whichever regime—in the past thousand years these have been Hungarian, Venetian, French, Austrian, Marxist, and Croatian—was currently in power. A regional culture of autonomy, tradition, and resistance to change has been the result, making Dalmatian cities like Dubrovnik both elegant and backward. Violich is not just describing regional culture but is also making the case for regionalism, for he is concerned about the human cost of losing our 'rootedness to place.' Older towns that reflect regional aesthetics and local building materials become 'treasures of the past,' standing in sharp contrast with the anesthesia of today's cities, 'hodge-podges of synthetic eclecticism and tiresome look-alikes, grown too large to allow the common citizen to claim a sense of identity with them or pride of belonging' (p. 51). This association of identity with territory, the spiritual nourishment that comes from being 'at home' somewhere, is characteristic of regionalism. It is surprisingly parallel to the inner-city gang members' desire to be with their 'homies' in a secure and familiar territory.

3. *Debatable Regionalism, Real Impact*

Scholars of regionalism like Putnum or Violich assume that regions are both naturally based and historically configured. They are materialist in that they see the natural features of the land as encouraging certain economic formations, and these in turn give rise to political and cultural patterns distinctive to the region. Yet even these scholars recognize that human ties to any particular region have become far more optional than in the past. In his discussion of identity, Violich points out that people seek environments compatible with their desired self-images in a way that was not necessary in earlier, more homogeneous cultures that provided a secure sense of belonging. 'Today, in contrast, with great individual opportunities for regional mobility, one can choose among diverse physical and cultural environments in various parts of a large city or metropolitan area as a way of individualizing one's cultural expression' (p. 59).

This is not something new, some scholars have asserted. Spatial boundaries and their cultural representations have always been subject to decisions made in the pursuit of interests. Beginning in the 1970s the view emerged that politics and the power of dominant groups, not geography, has always shaped humanity's mental and physical maps. Much of this literature centered on the origins of nationalism, as in the theses of 'the imagined community' (B. Anderson 1991) or 'the invention of tradition' (Hobsbaum and Ranger 1983), although smaller communities were similarly fictitious (Laitin 1986, Griswold 1992). Once in place, the arbitrary or fanciful boundaries could have powerful and lasting consequences, and indeed some like 'the nation' became unavoidable, as the 'world culture'

thesis pointed out. Cultural expression played a role in such inventions, both encouraging them, as in Benedict Anderson's discussion of the role of newspapers, and in sustaining them, as in critiques of heritage movements.

Revisionist historical and sociological research, such as that of the New Western Historians, has focused on the manipulation of collective memory (Ayers et al. 1996). Regional representations, in this view, are highly suspect. For example, a universally recognized regional symbol is the New England village with its white steeples and town commons. Stephen Nissenbaum (1996) contends that New England towns were originally not different that other towns in British North America, for there was no reason for farmers to live in villages. The compact village arrangement was a consequence of the commercial revolution in the late eighteenth and early nineteenth centuries, when it made sense for merchants and professionals both to set up business in and live in villages. And the 'pastoralization' of town commons—previously they had been bare patches, rather like parking lots—didn't take place until after the Civil War where local Village Improvement Societies, 'usually dominated by businessmen or newcomers to the community,' began to plant grass and elm trees. Litchfield, Connecticut actually hired Frederick Law Olmstead, designer of Central Park, to 'restore' their common.

Once region became seen as subject to human agency, the door was open to a lively debate on the relationship between regionalism and modernism, no longer assuming that the latter simply came along and eclipsed the former. Yatsuka (1998) describes how internationalist architects believed modernism to be 'globally shared beyond a regional understanding.' In architecture, unlike in politics, the international (modernist) preceded the regional: 'contrary to the simple notion of ''regional things,'' the idea of ''regionalism'' arose due to the appearance of ''internationalism''' (p. 167). Regional reaction—'a romantic nationalism'—arose in response to modernism, as in the case of the Finnish modernist Alvar Aalto. Yatsuka cites 'critical regionalism'—Kenneth Frampton's term for resistance against corporate modernism on the one hand and the postmodern reaction (compensatory façade) on the other—critical regionalism couldn't have been conceived of without modernism. He looks at Japanese architects such as Kenzo Tange, who both used and altered traditional (regional) forms such as the *sukiya* (sixteenth-century teahouse) and *shinden-zukuri* (nobleman's palace), grafting elements onto modernist layouts; by asserting 'we inherit spirit, not individual idioms' (p. 177), Tange avoided nostalgia just as he avoided placeless internationalism. Along similar lines James Dennis (1998) shows how the regionalist painters of the thirties, despite being given to bouts of antimodernist rhetoric, demonstrated strong modernist influences. The dichotomy between regionalist and modernist, like that between provincial and cosmopolitan, has been not just overdrawn but misconceived.

During the 1970s and 1980s some theorists of postmodernity argued that in the world of electronic communications and unrestricted cultural flows, place-based notions like 'region' and 'regional culture' were irrelevancies, at best productions for a tourist market intent on collecting scenic memorabilia (MacCannell 1976). The 'no sense of place' thesis maintained that electronic media has severed the link between social place and physical place. '[A]s place and information access become disconnected, place-specific behaviors and activities begin to fade' (Meyrowitz 1985, p. 148). This was something of a return to the bi-model model, only instead of industrialization, capitalism and imperialism, now electronic media and globalization were bleaching regional cultures out of the postmodern fabric.

More recent research has drawn back from this position, in part due to the undeniable assertiveness of place-based social formations, traditional or new-mint (Logan and Molotch 1987). Indeed, globalization itself, to the degree that it makes state boundaries less significant, gives rise to place-specific formations like the much discussed 'Europe of the regions.' Proliferation of communications channels encourages formation of new possibilities for the assertion and communication of cultural identities, and in some cases regionalism seems to find new life through these media (Griswold and Engelstad 1998).

4. *Future Research Directions*

Research at the end of the twentieth century accepts the constructed nature of regional cultures but takes a W. I. Thomas view that such regional cultures are real in their consequences, regardless of their origins. Making space into place, 'mapping the invisible landscape' through words and symbols (Ryden 1993), is a cultural process that, while by no means innocent of politics and economics, seems to fulfill a human need for meaningful spatial orientation. Regionalism and the culture of place represent, following this line of thought, a psychologically fundamental dimension of cultural expression. Thus, the regionalist impulse is not likely to atrophy, regardless of changes in mobility or communications technology.

Scholars have been addressing the lines along which regionalism gets configured (Ayers et al. 1996), the circumstances under which regionalism expands (Griswold and Engelstad 1998), the ability of regional culture to offer a viable alternative to class and racial hierarchies (Stack 1996), and the degree to which regionalism influences social and political behavior (Putnam 1993). Particular attention is being paid to how a regional culture is apt to be less the result of shared local experience, as celebrators of regionalism have maintained, and more the consequence of elite

conflicts brought about by external social changes. Anthony Cosgrove (1993) describes how the sixteenth-century Venetian elite still had its fortune tied to maritime power, and felt increasingly insecure due to French pressure. One response was a rash of estate construction in the rural interior, 'terra firma.' While these were working estates, they were ideological as well, representations of the harmony of various spheres: macrocosm, microcosm, divine will, world, theatre, human body, buildings, and lands. In an era when professionals were rising in status, architects like Andrea Palladio (1508–1580) created this regional culture of harmony and control. Palladio's use of classical forms and his emphasis on harmonizing buildings with landscape combined 'Roman imperial grandeur [with] Venetian sensitivity to atmosphere, site, and situation' (p. 20). Thus, the Palladian rural landscape both expressed the unease of the urban regional elite and reassured them of their centrality within an orderly cosmos.

This type of analysis, while recognizing the fundamental human desire to connect to place, looks for how political and economic struggles shape the way this connection gets articulated as regional cultural expression. This is one of the liveliest lines of research at the beginning of the twenty-first century and seems likely to continue. A second direction for the future is a renewed interest in the relationship between technology and regionalism. Studies of how elite conflicts and the pursuit of economic interests mold regional culture have tended to downplay the impact of new communications media, but the question of technological change is being revived by the Internet. The impact of the Internet on regionalism is unclear, with possibilities ranging from the theory that it will constitute the final blow to 'place' and other material forms of identity to the theory that it will encourage new 'tribal' groupings, including those of region. A new generation of researchers, unconstrained by former assumptions about how modernization would devastate regional culture but also largely free from a romanticist reverence for the traditional, will undoubtedly explore these possibilities.

See also: Ethnonationalism: Cultural Concerns; Locational Conflict (NIMBY); Neighborhood: General; Neighborhood Revitalization and Community Development; Place in Geography; Postmodernism in Geography; Regional Geography; Regional Identity; Residential Segregation: Sociological Aspects; Space: Linguistic Expression

Bibliography

Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, rev. edn. Verso, London

Anderson E 1999 *Code of the Street: Decency, Violence, and the Moral Life of the Inner City*, 1st edn. W. W. Norton, New York

Ayers E L, Limerick P N, Nissenbaum S, Onuf P S 1996 *All Over the Map: Rethinking American Regions*. Johns Hopkins University Press, Baltimore

Cosgrove D 1993 *The Palladian Landscape: Geographical Change and Its Cultural Representation in Sixteenth-century Italy*. Pennsylvania State University Press, University Park, PA

Dennis J M 1998 *Renegade Regionalists: The Modern Independence of Grant Wood, Thomas Hart Benton, and John Steuart Curry*. University of Wisconsin Press, Madison, WI

Gieryn T F 2000 A space for place in sociology. *Annual Review of Sociology* **26**: 463–96

Griswold W 1992 The writing on the mud wall: Nigerian novels and the imaginary village. *American Sociological Review* **57**: 709–24

Griswold W, Engelstad F 1998 Does the center imagine the periphery?: State support and literary regionalism in Norway and the United States. *Comparative Social Research* **17**: 129–75

Hobsbaum E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Laitin D 1986 *Hegemony and Culture: Politics and Religious Change Among the Yoruba*. University of Chicago Press, Chicago

Logan J R, Molotch H L 1987 *Urban Fortunes:The Political Economy of Place*. University of California Press, Berkeley, CA

Maffesoli M 1996 *The Time of the Tribes: The Decline of Individualism in Mass Society*, trans. Smith D. Sage, London

Massey D S, Denton N A 1994 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA

MacCannell D 1976 *The Tourist: A New Theory of the Leisure Class*. Schocken Books, New York

Meinig D W 1986 *The Shaping of America: A Geographical Perspective on 500 Years of History. Volume 1 Atlantic America, 1492–1800*. Yale University Press, New Haven, CT

Meyrowitz J 1985 *No Sense of Place: The Impact of Electronic Media on Social Behavior*. Oxford University Press, New York

Nissenbaum S 1996 New England as region and nation. In: Ayers E L, Limerick P N, Nissenbaum S, Onuf P S (eds.) *All Over the Map: Rethinking American Regions*. Johns Hopkins University Press, Baltimore, Chap. 2

Odum H W, Moore H E 1938 *American Regionalism: A Cultural–Historical Approach to National Integration*. Henry Holt, New York

Putnam R D 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ

Reed J S 1986 *The Enduring South: Subcultural Persistence in Mass Society*. University of North Carolina Press, Chapel Hill, NC

Ryden K C 1993 *Mapping the Invisible Landscape*. University of Iowa Press, Iowa City, IA

Stack C 1996 *Call to Home: African Americans Reclaim the Rural South*. Basic Books, New York

Violich F 1998 *The Bridge to Dalmatia: A Search for the Meaning of Place*. Johns Hopkins University Press, Baltimore

Yatsuka H 1998 Internationalism versus regionalism. In: Ferguson R (ed.) *At the End of the Century: One Hundred Years of Architecture*. Museum of Contemporary Art, Los Angeles

W. Griswold

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Regression Discontinuity Design

1. Definition

The regression discontinuity (RD) research design is a quasi-experimental design that can be used to assess the effects of a treatment or intervention. Unique to the RD design is that participants are assigned to groups solely on the basis of a pretreatment cutoff score. The name 'regression discontinuity' indicates that a treatment effect appears as a 'jump' or *discontinuity* at the cutoff point in the *regression* function linking the assignment variable to the outcome. In its simplest form, the design has two groups (those scoring above and below the cutoff), a pretest (the assignment variable) and a post-test (the outcome). More complex variations are also possible.

To many, the RD design seems counterintuitive. Using a cutoff to assign participants to the treatment and comparison groups creates a pretreatment group nonequivalence that seems like it should lead to selection bias. But while the design does induce initial nonequivalence, it does not necessarily lead to biased treatment effect estimates. Instead of assuming pretreatment equivalence in measured and unmeasured means and variances—as with the randomized experiment—in the absence of a treatment effect the RD design requires the two groups to be equivalent in their pre–post regression functions. This assumption can be tested.

The major attraction of RD designs is that they can be used to estimate the effects of treatments given to those who most need or deserve them, provided that need or merit is determined as a qualification score on the assignment variable (i.e., the cutoff) and nothing else. Because the design does not require some needy individuals being assigned to a no-treatment group, it has ethical advantages over experiments for assessing treatment effects.

2. History

The RD design was first used in psychology by Thistlethwaite and Campbell (1960) and was first discussed in detail in Campbell and Stanley (1963). It has been repeatedly reinvented—by the economist Goldberger (1972), the statistician Rubin (1977), the educators Tallmadge and Horst (1976), and by the public health researchers Finkelstein et al. (1996). Despite such independent reinvention, the RD design has not been used much outside of compensatory education where students scoring below some predetermined cutoff value on an achievement test are then assigned to remedial training (Trochim 1984). But it has also been used to study how unemployment benefits influence recently released prisoners (Berk and Rauma 1983) and how different levels of incarceration affect those in prison (Berk and de Leeuw 1999). A variation was also used to study how class size affects test scores (Angrist and Lavy 1999).

The low frequency of published examples may be due to several factors. One is its recency. Another is that RD designs require administrators to assign participants to conditions *solely* on the basis of quantitative indicators, restricting the role of judgment, discretion, or favoritism. Finally, understanding the design depends on some familiarity with regression analysis, making the strategy difficult to convey to nonstatistical audiences.

3. The Basic RD Design Structure

Simulated data for an RD design are shown in Fig. 1. It depicts the example of a compensatory treatment—perhaps a reading program designed to help students who initially score poorly (defined as below the cutoff) on a standardized reading measure. The figure depicts the standard bivariate plot of post-test and pretest for 500 simulated participants. Data points for the treatment recipients are represented with '+' while the comparison cases are 'o.' The cutoff value in this example occurs at the middle of the pretest distribution (pretest = 50) and is indicated by the vertical line. All students scoring below 50 are defined as needing compensatory training and get the treatment, while the remainder (i.e., those to the right of the cutoff) are assigned to the comparison group.

The solid lines represent the straight-line regression of post-test on pretest for each group. In the absence of a treatment effect, the critical assumption is that the regression line in the comparison group would continue to the left of the cutoff into the treatment group region, as indicated by the heavy dashed line. That is,

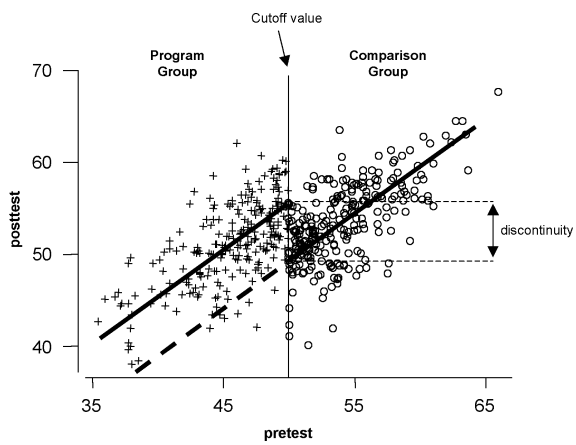

Figure 1
Basic regression discontinuity design for a compensatory treatment showing a positive effect

if the reading program does not work, then the observed treatment group regression line would be where the dashed line is. The fact that the observed treatment group line is displaced from this 'expected' line suggests a treatment effect—treatment group participants scored higher on the post-test (i.e., vertically) than would have been predicted from the comparison group pre–post relationship. The size of the treatment main effect is estimated as the vertical jump or discontinuity between the regression lines at the cutoff. Assuming that higher scores on both measures indicate better performance, one can conclude from the figure that the treatment improved participant performance by the amount of the vertical discontinuity.

In experimental designs one assures through random assignment or otherwise that the treatment and comparison groups are initially equivalent. Otherwise, post-treatment differences can be attributed to the intervention. In the RD design one obviously does not expect this kind of equivalence. Instead, it is assumed that the *relationship* between the assignment variable and the outcome is equivalent for the two groups—that is, the same continuous regression line describes both groups. Thus, interpretation of the RD design depends mostly on two factors: (a) that no alternative cause operates at or near the cutoff; and (b) that we can perfectly model the pre–post relationship. These issues constitute the major problems in the statistical analysis of the RD design, particularly the second one.

4. Design Considerations and Variations

4.1 Selection of the Cutoff

The choice of cutoff value can be made solely on the basis of available resources. If a program can only handle 25 persons and 70 apply, then a cutoff point can be selected that distinguishes the 25 most 'needy' persons from the rest. The cutoff can also be chosen on substantive grounds. If the assignment measure is an indication of severity of illness measured on a 1–7 scale, relevant experts might contend that all those scoring 5 or more are ill enough to deserve treatment. So, a theoretically justified cutoff of 5 is used.

4.2 Assignment Variations

While the use of a pretreatment cutoff value distinguishes the RD design, it is often difficult to implement. In many situations it does not appear to allow room for professional judgment or discretion. In fact, the design does not preclude incorporating such judgment so long as it can be quantified or explicitly accounted for when specifying the cutoff value.

Violation of the assignment rule leads to 'mis-assignment.' It seems to be most prevalent at scores around the cutoff and is often labeled as the 'fuzzy' RD design. Standard analyses will yield biased estimates of treatment effect in the fuzzy RD case. While statistical procedures for dealing with mis-assignment have been suggested (Trochim 1984), mis-assignment is better avoided if possible.

4.2.1 Multiple cutoff points. RD designs are not limited to a single cutoff value. In the absence of treatment the assumption is that the pre–post relationship can be described as a single continuous regression line extending over the entire range of the assignment scores. Even when multiple cutoff points are used, the comparison group's pre–post relationship is still used as the counterfactual for projecting where other groups' regression lines should be if there is no treatment effect.

The simplest multiple cutoff case involves two cutoff scorees. In this variant, persons scoring above the higher cutoff might be assigned to one treatment, those scoring below the lower cutoff to another, and those scoring within the cutoff interval might be controls. The assignment of treatments to each part of the assignment distribution can be made in many ways, including by random assignment.

Considerable work has been done on another way of 'coupling' the RD design to the randomized experiment. This is when individuals scoring in a range just above or below an originally selected cutoff value are randomly assigned to conditions, while those scoring further below and further above the cutoff are analyzed as from an RD design (Boruch 1973). Trochim and Cappelleri (1992) discuss four other ways of coupling random assignment to RD design features.

Multiple cutoff points are also involved in designs where the treatment is implemented a series of times (i.e., in phases), perhaps beginning with the most needy before successively giving it to those less needy. In the first phase, one uses a single cutoff value, assigning those most in need to the treatment and then measuring post-treatment performance. Then the cutoff value would be moved to include the next most needy group on the assignment measure. Eventually, everybody could receive the treatment, but in groups ordered by need and phase. This is particularly useful in organizations seeking to 'roll-out' a treatment over time, but eventually serving all.

4.2.2 Multiple assignment measures. The requirement of strict adherence to the cutoff criterion entails a single quantitative indicator. When this does not capture the degree of pretreatment need well, Trochim (1990) discusses two multivariate strategies.

The first involves the use of several separate measures, each having its own cutoff value. This is most feasible when a prospective participant must meet the different cutoff criteria on measure 1 'and' measure 2 'and' measure 3, etc. Less feasible is an 'or' rule when assignment depends on meeting a fixed number of criteria from a larger set—say, 5 out of 10. The second involves assignment according to a new variable that is a composite of several individual measures. Using multiple measures this way can help incorporate more judgment into a quantitative assignment strategy.

4.3 Treatment Variations

When treatment comparisons are absolute, the control group receives no formal treatment or a placebo treatment. When they are relative, the comparison group receives an alternative treatment, very often the current treatment to which a novelty is to be compared.

An advantage of the RD design is that it enables assignment of progressively riskier treatments to those in greater need. Assume three treatment programs: the standard, plus two increasingly riskier experimental ones. It is then possible to have two cutoff points and to assign the riskiest treatment to the neediest patients, the least risky (i.e., standard) treatment assigned to the least needy, and the moderately risky treatment assigned to those falling between the two cutoffs. (If no *a priori* basis exists for making risk judgments, an alternative would be to use a single cutoff and to assign those least in need to standard treatment while randomly assigning those on the other side of the cutoff into one of the two experimental protocols.) RD designs are flexible for examining treatment variations.

5. Post-treatment Measurement Variations

In RD designs, pre- and postmeasures do not have to be on the same or equivalent measures. For instance, one might assign persons to a health education treatment on the basis of household income but then examine its effects on attitudes towards health. The assignment variable is an economic indicator and the outcome is a cognitive one.

The RD design is not restricted to a single outcome, and for each post-treatment measure there can be a separate RD design and analysis. But it will often be useful to create an aggregate outcome or outcomes. For instance, we may have many knowledge items that are scaled to give a total and subtest scores for mathematical reasoning, computational skills, verbal reasoning, analogies, and so on. We might then conduct separate analyses for both the total score and each subtest.

Outcome measures need not be continuous normal variables. While such distributions facilitate RD analysis, they are not necessary for it. Thus, Berk and Rauma (1983) evaluated the effects of a California law that extended unemployment benefits to released prisoners who had not previously been eligible. In order to qualify, prisoners had to earn at least $1,500 working in the prison over the 12 months prior to release. Would extending these benefits reduce subsequent recidivism—a dichotomous outcome where 1 indicated recidivism and 0 staying out of prison? To analyze this outcome, Berk and Rauma (1983) relied on a linear random utility model related to the binary logit model. When continuous normal outcomes are not feasible, the analyst must specify a statistical model to account for the specified distributional form of the outcome.

5.1 Internal Validity

Internal validity refers to the degree to which causal inference is reasonable because all alternative explanations for an apparent treatment effect can be ruled out. The RD design is very strong with respect to internal validity because the process of selection into treatments is fully known and only factors that would serendipitously induce a discontinuity in the pre–post relationship at the cutoff can be considered threats. However, validity also depends on how well the analyst can model the true pre–post relationship so that underlying nonlinear relationships do not masquerade as discontinuities at the cutoff point.

5.2 Measurement Error

Stanley (1991) has argued that the RD design yields biased estimates of the treatment effect when the assignment variable is measured with error. Cappelleri et al. (1991) refuted this for situations with no interaction effects and Trochim et al. (1991) for the interaction case. The reason such error does not induce bias here although it does in quasiexperiments is that assignment to treatment is by the fallible assignment variable and not by some latent underlying selection process, as happens in quasiexperiments.

5.3 Statistical Power

It takes approximately 2.75 times as many participants in an RD design to achieve statistical power comparable to a simple randomized experiment (Goldberger 1972). When RD is combined with random assignment within an interval around this cutoff, 2.48 times the number of respondents are needed if 20 percent of participants are randomized; 1.96 times with 40 percent randomization; 1.46 with 60 percent; and 1.14 with 80 percent (Cappelleri and Trochim 1994, 1995).

6. Analysis

6.1 Assumptions of RD Analysis

Five central assumptions must be made in order for responsible analysis of data from an RD design:

(a) *The cutoff criterion*. The cutoff criterion must be followed without exception. Otherwise, a selection threat arises and estimates of the treatment effect are likely to be biased.

(b) *The pre–post distribution*. The true pre–post distribution must be known and correctly specified as linear, polynomial, logarithmic, or the like. Reichardt et al. (1995) point out the difficulties in specifying the appropriate model, especially when some curvilinearity exists in the pre–post relationship.

(c) *Comparison group pretest variance*. There must be a sufficient range of pretest values in the comparison group to enable adequate estimation of the pre–post regression line for that group. Variability in the treatment group is also desirable, though not strictly required because one can project the comparison group line even to a single treatment group point.

(d) *Continuous pretest distribution*. Both groups must come from a single continuous pretest distribution. In some cases one might find intact groups that serendipitously divide on a measure so as to imply some cutoff (e.g., pents from two different geographic locations). But such naturally discontinuous groups must be used with caution.

(e) *Treatment implementation*. It is assumed that the treatment is uniformly delivered to all recipients—that is, they all receive the same dosage, length of stay, amount of training, or whatever. If not, it is necessary to model explicitly the treatment as implemented, thus complicating the analysis considerably.

6.2 A Model for the Basic RD Design

The model presented here is for the basic RD design. Given a pretest assignment measure, x_i, and a post-treatment measure, y_i, the model can be stated as follows:

$$y_i = \beta_0 + \beta_1 \tilde{x}_i + \beta_2 z_i + \beta_3 \tilde{x}_i z_i + \ldots + \beta_{n-1} \tilde{x}_i^s z_i + e_i$$

Where $\tilde{x}_i$ = pretreatment measure for individual i minus the value of the cutoff, x_0 (i.e., $\tilde{x}_i = x_i - x_0$); y_i = post-treatment measure for individual i; z_i = assignment variable (1 if treatment participant; 0 if comparison participant); s = the degree of the polynomial for the associated $\tilde{x}_i$; β_0 = parameter for comparison group intercept at cutoff x_0; β_1 = linear slope parameter; β_2 = treatment effect estimate; β_n = parameter for the sth polynomial or interaction terms if paired with z; e_1 = random error.

The major hypothesis of interest is: H_0: $\beta_2 = 0$ tested against the alternative: H_1: $\beta_2 \neq 0$. This model estimates both main and interaction effects at the cutoff point in order to examine changes in means and slopes. It postulates a pre–post relationship that is polynomial at the cutoff. It also requires subtracting the cutoff score from each pretest score. The term $\tilde{x}_i$ has a superscript tilde to indicate this transformation of the pretest x_i. Finally, the model allows for any order of polynomial function (although certain restrictions are made in specifying the function as described below). Thus, the true pre–post relationship can in theory be linear, quadratic, cubic, quartic, and so on, or any combinations of these.

6.3 Model Specification

The key analytic problem is correctly specifying the model for the data—in this example, as a polynomial model. No simple or mechanical way exists to determine definitively the appropriate model. As with any statistical modeling, the RD analysis requires judgment and discretion and conducting multiple analyses based on different assumptions about the true pre–post relationship.

6.4 Steps in the Analysis

Recall what data the basic RD design provides on each unit. There is a pretest assignment value x_i in the model. Knowing this and the cutoff value allows creation of a new variable, z_i, which is equal to 1 in the treatment group and 0 if not. Finally, there is a post-test score labeled y_i in the model. Given the variables x_i, z_i, and y_i, the steps to be followed in the polynomial analysis are:

(a) *Transform the pretest*. The analysis begins by subtracting the cutoff value from each pretest score, thus creating the term as in the model. This sets the intercept equal to the cutoff so that estimates of effect are made at the cutoff rather than at $x_i = 0$.

(b) *Examine the relationship visually*. It is important to determine whether there is any visually discernable discontinuity at the cutoff. It could be a change in level, in slope, or both. If a discontinuity is visually clear at the cutoff one should not be satisfied with analytic results that indicate no effect. However, if no discontinuity is apparent, variability in the data may be masking an effect and one must attend carefully to the analytic results.

The second thing to look for is the degree of polynomial that may be required as indicated by the overall slope of the distribution, but particularly in the comparison group part. A good approach is to count the number of flexion points (how often the distribution 'flexes' or 'bends'). A linear distribution implies no flexion points, while a single flexion point could indicate a quadratic function. This information is used to specify the initial model.

(c) *Create higher-order terms and interactions*. Depending on the number of flexion points, create transformations of the transformed assignment variable, $\tilde{x}_i$. The rule of thumb is that one goes two orders

of polynomial higher than indicated by the number of flexion points. Thus, if the bivariate relationship appeared linear (i.e., no flexion points), one would want to create transformations up to a second-order (0+2) polynomial. The first-order polynomial already exists in the model ($\tilde{x}_i$) and so one would only have to create the second-order polynomial by squaring $\tilde{x}_i$ to obtain $\tilde{x}_i^2$. For each transformation of $\tilde{x}_i$ one also creates the interaction term by multiplying the polynomial by z_i. In this example there would be two interaction terms: $\tilde{x}_i z_i$ and $\tilde{x}_i^2 z_i$. If there seem to be two flexion points in the bivariate distribution, one would then create transformations up to the fourth (2+2) power and their interactions. This rule of thumb errs towards overestimating the true polynomial function needed, for reasons outlined in Trochim (1984).

(d) *Estimate the initial model*. The true analysis can now begin. One simply regresses the post-test scores, y_i, on $\tilde{x}_i$, z_i and all higher-order transformations and interactions created in step (c) above. The regression coefficient associated with the z_i term (i.e., the group membership variable) is the estimate of the main effect of the treatment. If there is a vertical discontinuity at the cutoff it will be estimated by this coefficient. One can test the significance of the coefficient (or any other) by constructing a standard *t*-test. If the analyst correctly overestimated the polynomial function required to model the distribution at step (c), then the treatment effect estimate will be unbiased. However, by initially including terms that may not be needed in the true model, the estimate is likely to be inefficient—that is, standard error terms will be inflated and statistical significance underestimated. But if the coefficient for the effect is highly significant, it would be reasonable to conclude there is an effect. Interaction effects can also be examined. A linear interaction is implied by a significant coefficient for the $\tilde{x}_i z_i$ term.

(e) *Refining the model*. The procedure described thus far is conservative, designed to reduce the chances of a biased treatment effect estimate even at the risk of increasing the error of the estimate (Trochim 1984). So, on the basis of the results of step (d) one might wish to attempt to remove apparently unnecessary terms and re-estimate the treatment effect with greater efficiency. This is tricky and should be approached with caution lest it introduces bias. One should certainly examine the regression results from step (d), noting the degree to which the overall model fit the data, the presence of any insignificant coefficients, and the pattern of residuals. One might examine the highest-order term in the current model and its interaction. If both coefficients are nonsignificant, and the goodness-of-fit measures and pattern of residuals indicate a good fit, one might then drop these two terms and re-estimate the model. One would repeat this procedure until either: (i) all of the coefficients are significant; (ii) the goodness-of-fit measure drops appreciably; or (iii) the pattern of residuals indicates a poor fit.

6.5 *Analyses with Multiple Cutoff Points*

The basic model described above can be applied directly when multiple cutoff points are used and random assignment is used within cutoff intervals. If assignment within intervals is nonrandom, one must also address the potential for selection bias in the analysis. When multiple cutoffs are used to distinguish separate treatments (i.e., multiple treatments are not assigned within the same cutoff interval), one would have to construct multiple treatment assignment variables for the analytic model (e.g., z_1, z_2, z_3) and all necessary interaction terms. Clearly, as more cutoffs and groups are added, model specification becomes more complex.

6.6 *Analyses with Multiple Assignment Measures*

Imagine having multiple measures, each with its own cutoff. Each assignment measure is then transformed by having its own cutoff value subtracted from it to create $\tilde{x}_i$. The analysis would then include all transformed assignment measures, group membership, higher-order terms, and interactions. For the simple first-order (or linear) case, one could use the model:

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 \tilde{x}_{1i} + \beta_3 \tilde{x}_{2i} + \beta_4 \tilde{x}_{3i} + \beta_5 \tilde{x}_{1i} z_i + \beta_6 \tilde{x}_{2i} z_i + \beta_7 \tilde{x}_{3i} z_i + e_i$$

This model does not include any two-way assignment variable interactions terms (e.g., $\beta_i \tilde{x}_{1i} x_{2i}$ or $\beta_i \tilde{x}_{1i} x_{2i} z_i$) or any three-way terms (e.g., $\beta_i \tilde{x}_{1i} x_{2i} x_{3i}$ and $\beta_i \tilde{x}_{1i} x_{2i} x_{3i} z_i$) but such an assumption may be reasonable where primary interest is in estimating β_1, the treatment effect, and not interactions. Clearly, though, the use of multiple assignment variables with higher-order polynomial models will quickly lead to an unwieldy analysis.

The situation is simpler when multiple assignment variables are rescaled to reflect 'number of assignment variables on which a person meets the criterion' (e.g., on 5 of 10). This number then becomes the x_i variable and one would conduct the basic analysis described earlier. When multiple assignment variables are combined into a single index, the analysis is also a straightforward application of the basic procedures described initially.

7. *Conclusions*

From a methodological point of view, inferences drawn from a well-implemented RD design are considered comparable in internal validity to conclusions from randomized experiments. However, the lower statistical efficiency of RD designs, and the resulting sample size demands, may limit the design's utility. From an ethical perspective, RD designs are com-

patible with the goal of getting the treatment to those most in need. It is not necessary to deny the treatment from potentially deserving recipients simply for the sake of a scientific test. From an administrative viewpoint, the RD design may be directly usable when allocation formulas are the basis for assigning treatments or programs.

With all of these considerations in mind, the RD design must be used judiciously. In general, the randomized experiment is still the first method of choice when assessing causal hypotheses (Cappelleri and Trochim 2000). However, where they are ruled out as impractical, the RD design should be considered as a practical alternative.

See also: Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Linear Hypothesis: Regression (Quantile); Quasi-Experimental Designs

Bibliography

Angrist J D, Levy V 1999 Using Maimonides' Rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* **114**: 533–75

Berk R A, de Leeuw J 1999 An evaluation of California's Inmate Classification System using a generalized regression discontinuity design. *Journal of the American Statistical Association* **94**: 1045–52

Berk R A, Rauma D 1983 Capitalizing on nonrandom assignment to treatments A regression discontinuity evaluation of a crime control program. *Journal of the American Statistical Association* **78**: 21–8

Boruch R F 1973 Regression-discontinuity designs revisited. Unpublished manuscript. Northwestern University

Campbell D T, Stanley J C 1963 Experimental and quasi-experimental designs for research on teaching. In: Gage N L (ed.) *Handbook of Research on Teaching*. Rand McNally, Chicago (Also published as *Experimental and Quasi-experimental Designs for Research*. Rand McNally, Chicago, 1966)

Cappelleri J C, Trochim W M K 1994 An illustrative statistical analysis of cutoff-based randomized clinical trials. *Journal of Clinical Epidemiology* **47**(3): 261–70

Cappelleri J C, Trochim W M K 1995 Ethical and scientific features of cutoff-based designs of clinical trials: A simulation study. *Medical Decision Making* **15**(4): 387–94

Capelleri J C, Trochin W 2000 Cutoff designs. In: Chow S-C (ed.) *Encyclopedia of Biopharmaceutical Statistics*. Marcel Dekker, New York, pp. 149–56

Cappelleri J C, Trochim W M K, Stanley T-D, Reichardt C S 1991 Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Evaluation Review* **15**(4): 395–419

Finkelstein M O, Levin B, Robbins H 1996 Clinical and prophylactic trials with assured new treatment for those at greater risk. 2. Examples. *American Journal of Public Health* **86**: 696–705

Goldberger A S 1972 Selection bias in evaluating treatment effects: Some formal illustrations. Discussion Papers, 123–72. University of Wisconsin, Madison, WI

Reichardt C S, Trochim W M K, Cappelleri J C 1995 Reports of the death of regression-discontinuity analysis are greatly exaggerated. *Evaluation Review* **19**(1): 39–63

Rubin D B 1977 Assignment to treatment on the basis of a covariate. *Journal of Educational Statistics* **2**: 1–26

Stanley T D 1991 'Regression-discontinuity design' by any other name might be less problematic. *Evaluation Review* **15**(5): 605–24

Tallmadge G K, Horst D P 1976 A Procedural Guide for Validating Achievement Gains in Educational Projects. Office of Education, Washington, DC

Thistlethwaite D L, Campbell D T 1960 Regression-discontinuity analysis: An alternative to the *ex post facto* experiment. *Journal of Educational Psychology* **51**: 309–17

Trochim W M K 1984 *Research Design for Treatment Evaluation: The Regression-Discontinuity Approach Program*. Sage, Beverly Hills, CA

Trochim W 1990 Regression-discontinuity design in health evaluation. In: Sechrest L, Perrin E, Bunker J (eds.) *Research Methodology: Strengthening Causal Interpretations of Non-experimental Data*. US Dept. of HHS, Agency for Health Care Policy and Research, Rockville, MD, pp. 119–39

Trochim W M K, Cappelleri J C 1992 Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials* **13**: 190–212

Trochim W M K, Cappelleri J C, Reichardt C S 1991 Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review* **15**(4): 571–604

W. M. Trochim

Regulation and Administration

Regulation is the effort to shape significant institutions—industry, universities, immigration—to serve some social concerns that might not be taken into account in the ordinary process of doing business. According to one definition, it is the imposition of legal orders on those engaged in economic activity (Kagan 1978). States regulate capitalist enterprises within capitalist or mixed economies.

1. Emergence and Justification

Bureaucratic regulation of industries emerged in the nineteenth century with the rise of industrial capitalism and the bureaucratic division of labor. Britain established a factory inspectorate for the textile industry as early as 1832. The use of inspectorates to address the more dangerous conditions at work—in match factories, in tenement labor—expanded dramatically throughout the Western world in the 1890s (Rodgers 1998, pp. 254–9).

Regulation has proven to be at least partly reversible: central trends in capitalist economies during the 1970s and 1980s included both privatization of

state-owned industries (see e.g., Graham and Prosser 1988, Pistula and Rasmussen 1990) and political pressure toward deregulating already regulated industries (Derthick and Quirk 1985, Harris and Milkis 1996).

Economists have thought it important to explain why regulation is necessary within a capitalist economy, given that neoclassical economic theory holds that social purposes are best accomplished by markets. The importance of neoclassical theory has meant that advocates and critics of regulation have also expended effort understanding the justifications (see e.g., Breyer 1982). Justification has sometimes centered on the concept of 'externalities.' A market requires that business pays for some components, for example, steel in manufacturing a car. However, the manufacturer does not have to buy the air that it pollutes in the process of making cars, and in turn, neither do consumers who drive the cars. In the United States in the 1960s and 1970s in particular, concern about the inability of companies to take into account the social costs of doing business was used to justify regulating products and processes in manufacturing.

In addition to regulating products and processes, governments regulate price and entry: the price that an industry may charge and the requirements to enter into the business. The justification of price and entry regulation does not rely upon the concept of externalities. Rather, the argument is that some industries take on the characteristics of monopoly, whether as a natural monopoly or because there are very high barriers to entry. For example, it could be inefficient to have multiple railways competing against each other since the costs of laying railroad track are very high. What is more, because it is expensive to build a railway, it is extremely difficult to enter the business, which means that the competition that is supposed to ensure that prices and services are responsive to consumer demand does not exist. Without competition, railroads could charge what they would for services. Thus railroads, the first industry subject to national bureaucratic regulation, in the United States would charge more for short-haul than for long-haul trips because there was much less competition in short haul (Skowronek 1982, pp. 124–5, 141–50). In air transport, too, the barriers to entry are high because airspace is limited and in any one airport runways and gates are limited. Someone must allocate the space and ensure that once the airline has the space it does not ignore what customers want. Economists also note the importance of 'network effects' in monopoly. If having compatible computer systems facilitates the exchange of information, each person buying a computer system has an incentive to buy one compatible with those people already have, whether or not that is the best system. Advocates assert that professions like medicine and law require regulation because expertise is required to assess whether the profession is being practiced correctly. Consumers cannot assess whether a surgeon knows what she is doing until it might be too late.

These justifications for regulation have been subject to serious critique from those who believe that markets should be unregulated and from those who believe that regulation is ineffective because industries have special access to the state not shared by their workers or customers (Derthick and Quirk 1985). In this view regulatory bureaucracies, designed to incorporate public concerns into capitalist economies, become 'captured' by the industries they are to regulate. The regulatory bureaucracies then act largely to insulate the industries from competition and protect their interests. Sometimes that has happened by design: railroad regulators meant to maintain and steady the railroad industry. In the late nineteenth century railroads in the United States wanted to be regulated to moderate the incentives to cut prices that were putting railroads out of business (Skowronek 1982). In France *notaires* welcome regulation to preserve their monopoly on drawing up legal contracts; opening that monopoly would have meant decreasing income as a result of competition (Suleiman 1987).

Economists call the effort to gain protection from the market by legislative or bureaucratic intervention 'rent-seeking behavior'—that is, an effort to gain a return beyond what is required to keep one in the business—and many have argued that rent-seeking is the heart of regulation. For example, lawyers benefit from the licensing of lawyers; it restricts the numbers of lawyers available and thereby maintains a higher price than would be the case if anyone could practice law. The claim that regulation protected inefficient industries from competition and used state power to keep prices artificially high provided a substantial part of the support for deregulation in the United States. In the late 1970s and 1980s (Derthick and Quirk 1985, Harris and Milkis 1996). In other nations, critiques of state ownership of industries led to privatization, which in turn entailed regulation (Graham and Prosser 1988). Across the industrialized Second World, industries differed in their ability to maintain monopolies or other advantages regulation provides (Derthick and Quirk 1985, Suleiman 1987).

In the United States, deregulation seems to have promoted consolidation in the airline industry and the trucking industry just as understandings of 'natural monopolies' might have predicted. Perhaps surprisingly, the strength of the pro-environmental regulation lobby meant that environmental deregulation did not go far (Harris and Milkis 1996). Those advocating a more market-based approach in pollution rights gained recognition in Clean Air Act Amendments, but government officials were still deeply involved in regulation (Harris and Milkis 1996). Presidents elected on antiregulation platforms do not completely control the political environment in which they work. The American federal system also provides access in the states when people have lost at the federal level. For

example, when the federal government tried to limit the regulation of advertising, activists turned to the more sympathetic state governments.

States approach regulation relying on different institutions that have, in turn, different styles. Courts, legislatures, local governments, marketing boards, and bureaucratic agencies all can (and have) imposed orders on capitalist enterprise. In the United States before the Civil War, rather than relying on negotiations between business and bureaucrats, or on rules from bureaucrats that were enforced in the courts, the United States relied on courts—largely state courts—to enforce statutes or common law duties. In addition, state legislatures granted charters to businesses and in return exacted some form of regulation: limits on what entrepreneurs could charge to cross a bridge or use a road (Hartz 1948), or requirements that those who owned docks contribute toward keeping harbors dredged and open (Hartog 1983). Such regulation is not a matter of general rules implemented through a bureaucracy, but it was an effort to require businesses to take into account public concerns not readily addressed in the market. By the late nineteenth century most states had instituted some bewildering array of boards, commissions, and special courts to instead ensure that a broader public interest was taken into account.

2. *Institutions that Regulate*

The first American federal regulatory agency, the Interstate Commerce Commission, was created in 1887. It was unusual in that it established an agency independent of immediate control by ministers, responsible for making a framework of policy through trial-type procedures, with decisions that could be appealed to general jurisdiction courts (Baldwin 1985, p. 7). The United States' official regulatory procedure is more rule-oriented than those in other Western industrialized states, and courts are more deeply involved in making policy. The American system also allows for much more public participation and opportunity to argue against regulations both within an administrative agency and through courts after administrative agencies have decided.

After World War Two regulation in the United States became open to wider public participation. Administrative agencies must hold public hearings after announcing the topic of a rule they are considering. They must then justify the rule they issue according to the evidence they have. Often an organization concerned with the rule will bring the regulatory agency to court. Federal courts usually review the processes by which the rules were promulgated and decide whether the rule or decision is justified according to the evidence. What lawyers have won in the United States has been observed sometimes with horror, sometimes with envy in other nations.

Other nations do not rely so extensively on the equivalent of generalist courts to assess the procedures by which rules are enacted. The French have a specialized court that hears challenges to administrative decisionmaking, the *Conseil d'Etat*, or Council of State. The *Conseil d'Etat* is staffed with people who have built careers within administration; the purpose is to have a court that will take into account public purposes, or the interests of the state, in deciding cases brought by individual challengers. In Britain, practitioners in regulation were frustrated enough with how their general jurisdiction courts supervised administrative decisions that the state created a version of a specialized administrative court in the 1970s (Sterett 1997). Particular judges specialize in hearing complaints about administrators. The political justification for both courts is that specialization in administrative complaints will lead to sensitivity to public purpose and an awareness of what can go wrong in administration. Germany, also, relies on a specialized administrative court staffed by a trained civil service judiciary that lies above a system of specialized tribunals.

A further contribution to the complexity of regulation in the United States is the extent of interest group litigation. Interest group challenges to the general rules under which administrators work are far more numerous than in any other nation. However, social welfare groups have in recent years organized to challenge rules concerning immigration and environmental regulation in Western Europe. The expansion of standards via the European Court of Justice and directives via the European Union have made it more possible to challenge domestic regulation, or refusals to regulate, in Europe. Finally, cross national treaties or other agreements also shape domestic regulation, for example in ozone regulation (Braithwaite and Drahos 2000; Canan and Reichman 1993)

In most countries there are specialized courts between generalist courts or the upper level administrative court and front line administrators. These specialized courts usually include representatives of relevant sectors (e.g., employers and union representatives in labor tribunals) in their staff. Members of specialized courts are not always legally trained. Their procedures are often less formal than those of generalist courts, and those before them do not always have legal representation. The informality and discretion allowed most tribunals provides substantial opportunity for a slip between formal statement of government policy and its meaning in particular cases.

3. *What Does it Cost?*

Line items in government budgets for regulatory agencies do not actually tell what regulation costs. To comply with regulatory requirements, businesses incur the costs of improving safety or documenting that they

are trying to. To the extent that regulation requires businesses to internalize their externalities, to pay for pollution abatements, for example, there is no one thing to measure. Business representatives have argued that the processes of American regulation are more difficult than those in other Western postindustrial states: compliance requires more documentation, inspectors more readily impose fines, businesses can expect more public interest group confrontation, hearings, and possibly court battles (Kagan and Axelarad 2000) The outcomes in Western states—the regulatory standards that are imposed—tend to converge. (Kagan and Axelarad 2000). It is unclear whether this means that other states achieve comparable standards without the process costs incurred in the United States or that the United States through its extensive procedures sets standards that are copied elsewhere.

Regulation involves trade-offs between effectiveness and other considerations. For example, we might want nursing homes to provide good care for people, but that is hard to measure. To the extent that we try to enforce it, we might direct effort in an organization toward documenting practices that do not quite get at caring, such as providing meals on time (Braithwaite 1993). A more substantive mandate to care might better capture what we want, but it would be unenforceable. In addition, different substantive regulations are acceptable to businesses and broader publics in different countries. For example, American commitments to nondiscrimination, which critics see as more extensive than those in most other countries, impose costs through requirements of documentation of fair procedures. Such provisions do not require any particular substantive outcome. Other countries, however, employ approaches that would be unacceptable in the United States. For example, Sweden tried to implement employment equality in a way that could keep administrative costs low while resulting in real change, by setting employment quotas for women in jobs in which they were underrepresented (Brown and Wilcher 1987). Britain, on the other hand, has relied on loose and open-ended definitions of nondiscrimination implemented through tribunals whose members are not legally trained and whose primary jurisdiction is not nondiscrimination law. The result seems to be a relatively inexpensive system that has done little to challenge discriminatory practices in the workplace (Kenney 1992).

Globalization has increased the importance of legal rules and lawyers in regulation in national states outside the United States, raising questions about the costs of more formal styles of regulation (Kagan 1997, Kagan 2000). Despite concerns about the expense of regulation, enterprises sometimes choose to pursue the goals of regulation beyond what governments require, because compliance offices sometimes employ people who take the goals of regulation to be their mission, or because large corporations concern themselves with their public image as good citizens (Edelman 1990; Rees 1988). Compliance is not always a matter of direct orders.

4. *Explaining Regulation: Political Institutions and Political Culture*

Explanations for the different approaches to regulation—contentious, and emphasizing public procedures and rules as against consensual, and without formal rules—across postindustrial states are of two different types. The first explanation emphasizes the historical emergence of governance in different countries. The second explanation emphasizes cultural predilections such as individualism as against faith in authority.

A historical-institutional approach holds that political institutions have shaped the emergence of different regulatory regimes. Those political institutions are in turn the result of large-scale historical processes. In particular, in much of Western Europe, a career civil service insulated from politics was created in the nineteenth century (Kelman 1981). Administrative arms of government often started as the personal servants of the monarch; the political environment within which they worked did not include mass electorates. Governments could develop bureaucracies insulated from political pressure, setting in place systems difficult to disrupt, no matter what cultural preferences there might be for a broader or more accessible or more aggressive system of regulation. Alternatively, the incorporation of political fragmentation also has a history necessary, for example, to understand affirmitive action policy in India (Parikh 1997)

Parliamentary systems, in which the government controls the legislative agenda, provide the government substantial control over how a regulator bureaucracy will be organized and the forms accountability will take. Should a government believe it is best to include only responsible-seeming or moderate interest groups in their discussion of regulation, they will find little institutionalized opposition (Vogel 1986). Parliaments see little reason to institute checks on administrators that can be invoked directly by interest groups, such as a right to sue. From the point of view of government ministers, both in charge of government departments and sitting in parliament, such a capability only makes regulation more difficult.

In addition to administrators and their ministerial supervisors, courts participate in regulation, to a different extent in different national states. In civil law systems, judging is a career to which people are trained. The highest level administrative court in France, the *Conseil d'Etat*, is staffed with top-level civil servants. Their charge is quite clearly to attend to complaints against administrators with an eye to maintaining the

effectiveness of a regulatory regime; it is not a system designed to allow confrontation over the fundamentals of public policy. Even within British-style common law systems, judges in the general legal system are selected from the elite of the bar, who are narrowly trained and disinclined to confront administrators. The insulation from the electoral politics that so profoundly shapes the civil services includes the judiciary.

The historical-institutional explanation points us toward the very different institutional development in the United States in order to understand differences in the contentiousness of regulation. In the American presidential and federal system, the legislatures—Congress and the states—are not wholly within the control of the executive. Those who wish to insist on visions of regulation different from that of the executive can gain leverage in the legislature to do so, not least through mobilizing popular mistrust of closed procedures. The American system of formal rulemaking procedures and some openness in administration resulted first from the American Bar Association's mistrust of regulation and then from the politically distinct mistrust emergent in the 1960s and 1970s. Both gained leverage through Congress and through the courts, which had more generous standards for allowing people to sue than did courts in other nations.

In contrast, cultural explanation for differences in regulation among states focuses more on shared meanings and less on the institutional sources of those meanings. American culture is individualist and assertive according to this interpretation, while cultures in other Western countries are more deferential, in part based on a monarchical past and strong social hierarchies (Vogel 1986). Deference, however, hardly characterizes the upheavals of the 1960s and 1970s, pervasive throughout Western industrialized states. Demands for participation were institutionalized in law in the United States but not in other nation-states in that period. In other nation-states attempts to make the government more responsive included instituting ombudsmen, state institutions to listen to complaints about administration in particular cases. Responses did not include a fundamental change in how rules were instituted, though they did in the United States.

Further evidence raises questions concerning the completeness of the cultural explanation. In Britain, for example, a country cited as having informal regulatory practices in comparison to the United States (Vogel 1986), land-use planning has been contentious, rule-bound, and full of litigation. A cultural explanation arguing that the British are more deferential than Americans cannot account for that contentiousness. An alternative institutional and historical explanation is relatively simple. The legal profession long made most of its income in land. By the time bureaucratic land-use regulation emerged, the government had to take into account the interests of an already organized legal profession accustomed to litigating land disputes. In attending to national cultures as a basis for explanation, tensions within cultures require scholars to attempt to explain why some cultural preferences have been enacted into law and some have not.

Supranational courts shape regulation within the European Union. Both the European Court of Human Rights and the European Court of Justice make decisions concerning the rules that nation-states use in admitting immigrants, limiting sex discrimination, and determining the right to a family life or the right to travel. The European Court of Justice also decides cases concerning environmental regulation, based on directives from the Council of Ministers. The context for regulatory bodies therefore includes not only the businesses and practices that they regulate, alongside the attentive publics following regulation and their domestic political institutions. Increasingly, it includes supranational institutions harmonizing requirements cross-nationally. A perspective that takes nation-states as monolithic entities would see the supranational courts as an imposition on national autonomy. Since, however, not everyone within a nation-state agrees—some political activists want stricter regulation of pollution, or more openness to the families of settled immigrants—the introduction of supranational institutions makes for more complex alliances. State officials themselves are often not united, so supranational decisions can strengthen the hand of one administrative office over another.

5. *Constituting Regulation: Rules and Discretion*

National legislatures and executives may state general rules for regulation: that sex discrimination in employment is illegal, that businesses must reduce air pollution. Regulation is a complex task in which goals are often in tension—for example, to clean the air while ensuring that businesses that pollute do not have to close, or without severely curtailing driving. Multiple and conflicting goals for complex organizations mean that what counts as compliance requires interpretation. Even in the United States, where the courts are in the business of continually clarifying the rules, the facts of a situation can be complicated enough that compliance with a rule is something that will entail negotiation between the regulator and regulated, which a court then only ratifies. An order to emit less smoke can easily become subject to negotiations regarding the kind of scrubbers to install (Melnick 1983). Distinguishing between general rules and decision making in individual cases is difficult. In regulation, general rules often emerge out of particular decisions.

Where officials have substantial discretion in determining compliance, the meaning of regulation emerges

in negotiations with the regulated company. Case-level administrators interpret their instructions in light of what regulated businesses ordinarily do. For example, in Britain, specialized courts (or tribunals) that were responsible for hearing labor complaints interpreted discrimination law so to prohibit *unreasonable* unfavorable treatment of women, even though the sex discrimination statute did not rely upon a standard of reasonableness. 'Reasonableness' was a standard in unfair dismissal cases outside of sex discrimination. Common sense told tribunal members that reasonableness must be part of the sex discrimination statute; in practice the tribunals ratified most of what businesses had been doing (Kenney 1992, pp. 104–8).

Changing the ordinary practices of an organization is difficult. Where there are applicable rules or direct court orders against an organization, the organization must construe what it means to comply. Organizations have every reason to interpret decisions so that they fit with organizational priorities. For example, a court order to pay women equally with men in comparable jobs still leaves open what count as equivalent jobs, how equal the pay must be, and over how long a period it must be equalized, all of which will be up to an employer to determine (McCann 1994). Administrative agency decisions or court interpretations of them gain meaning in their application.

The professional training of decision-makers shapes what regulations mean. In employment discrimination law in the United States, personnel administrators interpret the law so as to require what they see as good management practices, which requires that people who work together get along. They do not often interpret the law as a matter of legal right, that is, as an enforceable claim that someone can make against his or her employer. Rather, personnel administrators are more likely to interpret problems as matters of personality conflicts (Edelman 1993). Thus the presence of enforceable rights does not necessarily make regulation very different from, for example, enforcement of water pollution laws in Britain, where there are no clear rules and enforcement officicals negotiate and bluff (Hawkins 1984, see also Graboski and Braithwaite 1986). A substantial difference is whether officials can enlist courts in enforcement efforts.

Recent scholarship focuses on both business organizations and administrative officials working in complex environments in which they mutually shape meanings of the law (Edelman 1993, Harris and Milkis 1996). Charting variation across sectors requires attending to that environment in which regulation works, including the variation in attentive publics as well as the organization of the state. While national states sometimes do have national approaches shared across regulatory sectors, the emergent supranational context is shaping regulation so as to require analysts to attend to reasons for convergence as much as differences between nation-states. Differences could result less from differences between postindustrial states; the global context of operation might lead to convergence among First World states (Drahos and Braithwaite 2000), while Third World countries produce goods, often in conditions much worse for workers. Regulation there could depend as much on global political forces as on internal political dynamics.

See also: Broadcasting: Regulation; Bureaucratization and Bureaucracy, History of; Employment and Labor, Regulation of; Environment Regulation: Legal Aspects; Environmental Policy: Protection and Regulation; International Communication: Regulation; Occupational Health and Safety, Regulation of; Public Administration, Politics of; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulation: Family and Gender; Regulation: Sexual Behavior; Regulation Theory in Geography; Regulation: Working Conditions; Regulatory Agencies

Bibliography

Baldwin R 1985 *Regulating the Airlines: Administrative Justice and Agency Discretion*. Clarendon Press, Oxford, UK

Boyle E 1998 Political frames and legal activity: The case of nuclear power in four countries. *Law and Society Review* **32**(1): 141–74

Braithwaite J 1993 The nursing home industry. In: Tonry M, Reiss A J (eds) *Beyond the Law: Crime in Complex Organizations*. *Crime and Justice* **18**: 11–54

Braithwaite J, Drahos P 2000 *Global Business Regulation*. Cambridge University Press, Cambridge, UK

Breyer S 1982 *Regulation and its Reform*. Harvard University Press, Cambridge, MA

Brown C, Wilcher S J 1987 Sex-based employment quotas in Sweden. In: Brown C, Pechman J A (eds.) *Gender in the Workplace*. Brookings Institution, Washington, DC

Canan P, Reichman N 1993 Ozone partnerships, the construction of regulatory communities, and the future of global regulatory power. *Law and Policy* **15**(1): 61–74

Derthick M, Quirk P J 1985 *The Politics of Deregulation*. Brookings Institution, Washington, DC

Edelman L B 1990 Legal environments on organizational governance: the expansion of due process in the American workplace. *American Journal of Sociology* **95**(6): 11401–440

Edelman L B 1993 Internal dispute resolution: The transformation of civil rights in the workplace. *Law and Society Review* **27**(3): 497–514

Edelman L B, Suchman M C 1997 Legal rational myths: The new institutionalism and the law and society tradition. *Law and Social Inquiry* **22**: 906–41

Einhorn R 1991 *Property Rules*. University of Chicago Press, Chicago

Graboski P, Braithwaite J 1986 *Of Manners Gentle: Enforcement Strategies of Australian Business Regulatory Agencies*. Oxford University Press, Melbourne, Australia

Graham C, Prosser T 1988 *Waiving the Rules*. Open University Press, Philadelphia

Harris R A, Milkis S M 1996 *The Politics of Regulatory Change: A Tale of Two Agencies*. Oxford University Press, New York
Hartog H 1983 *Public Property and Private Power: The Corporation of the City of New York in American Law, 1730–1870*. University of North Carolina Press, Chapel Hill, NC
Hartz L 1948 *Economic Policy and Democratic Thought: Pennsylvania: 1776–1860*. Harvard University Press, Cambridge, MA
Hawkins K 1984 *Environment and Enforcement: Regulation and the Social Definition of Pollution*. Clarendon Press, Oxford, UK
Kagan R A 1978 *Regulatory Justice: Implementing a Wage-price Freeze*. Russell Sage Foundation, New York
Kagan R A 1997 Should Europe worry about adversarial legalism? *Oxford Journal of Legal Studies* **17**(2): 165
Kagan R A 2000 The consequences of adversarial legalism. In: Kagan R A, Axelarad L (eds.) *Regulatory Encounters: Multinational Corporations and American Adversarial Legalism*. University of California Press, Berkeley, CA
Kelman S 1981 *Regulating America, Regulating Sweden: A Comparative Study of Occupational Safety and Health Policy*. MIT Press, Cambridge, MA
Kenney S J 1992 *For Whose Protection? Fetal Protection Policies in the United States and Great Britain*. University of Michigan Press, Ann Arbor, MI
McCann M 1994 *Rights at Work*. University of Chicago Press, Chicago
Melnick R S 1983 *Regulation and the Courts*. Brookings Institution, Washington, DC
Parikh S 1997 *The Politics of Preferences*. University of Michigan Press, Ann Arbor, MI
Pistula J M, Rasmussen K 1990 *Privatizing a Province: The New Right in Saskatchewan*. New Star Books, Vancouver, Canada
Rees J 1988 *Reforming the Workplace: A Study of Self-regulation in Occupational Safety*. University of Pennsylvania Press, Philadelphia
Rodgers D 1998 *Atlantic Crossings: Social Politics in a Progressive Age*. Harvard University Press, Cambridge, MA
Ruggie M 1984 *The State and Working Women*. Princeton University Press, Princeton, NJ
Skowronek S 1982 *Building a New American State*. Cambridge University Press, New York
Sterett S 1997 *Creating Constitutionalism?* University of Michigan Press, Ann Arbor, MI
Suleiman E 1987 *Private Power and Centralization in France: The Notaires and the State*. Princeton University Press, Princeton, NJ
Vogel D 1986 *National Styles of Regulation: Environmental Policy in Great Britain and the United States*. Cornell University Press, Ithaca, NY

S. Sterett

Regulation, Economic Theory of

Regulation is defined as 'the act of controlling or directing by a rule, principle, [or] method' (Webster 1989). This definition encompasses many forms of control by a variety of actors. The ensuing discussion will focus on the important form of economic regulation in which the government controls the activities of producers in order to enhance the well being of the customers they serve.

1. Regulating with Accurate Information

When the average cost that each firm incurs in serving customers (i.e., the firm's cost per customer) declines as the firm serves more customers, industry costs will be minimized when only one producer serves all customers. For example, the cost of providing local telephone service to a community is often lower when a single telecommunications network is constructed and operated than when multiple networks are operated. Similarly, because of declining average costs, the local distribution of electricity, water, and gas is often provided at minimum cost by a single firm. But if a single profit-maximizing firm faces no competition, consumers will have no alternative source of supply, and so an unregulated sole producer may have considerable freedom to charge high prices for its services. Regulation of the prices charged by a single (monopoly) supplier can help to protect consumers while permitting the industry configuration that minimizes the costs of serving customers.

Consumer protection can take many forms, particularly when the regulated supplier provides more than one service. To illustrate this fact, consider the mandate of the US Postal Rate Commission. When it sets the prices that the US Postal Service charges for the services it delivers, the Commission is required to ensure that the pricing structure is simple and that prices are fair and equitable. The Commission must also ensure that established prices reflect the value of the service rendered and cover the costs of providing each service (Postal Reorganization Act 1970, § 3622(b)).

To determine how a regulatory body (hereafter, a regulator) might attempt to satisfy such mandates, consider first, the case of an omniscient regulator who knows both how highly consumers value the services produced by the regulated firm and the minimum possible costs that the firm incurs in serving customers.

1.1 Linear Pricing with a Single Output

Initially suppose that the regulated firm produces only one product. Also suppose that the regulator desires to set the price for this product so as to maximize the aggregate welfare of all consumers, while delivering sufficient profit to the firm to guarantee that it will voluntarily serve customers at the established price. Consumer welfare measures the total value that consumers derive from their purchase and consumption of the product. A firm's profit is the difference

between the revenues it receives and the costs it incurs. Profit is a measure of producer welfare.

To maximize consumer welfare while delivering at least the minimum required level of producer welfare, the regulator must set the price that consumers pay for each unit of the product they purchase so as to deliver the largest possible sum of consumer and producer welfare. This welfare-maximizing price is the one that equals the extra (or marginal) cost the firm incurs in producing the last unit of the product it sells. This is because consumers will continue to buy more of the product as long as the extra value they receive from doing so exceeds the price they must pay for an additional unit of the product. Therefore, when the price reflects the producer's marginal cost of production, consumers will purchase additional units of the product whenever the value they derive from consumption exceeds the cost of allowing that consumption. Consequently, a regulated price that reflects the producer's marginal cost of production will ensure the level of output that maximizes the combined welfare of consumers and the firm.

However, when unit production costs decline as the firm produces more output, a price set equal to the marginal cost of producing the last unit of output will cause the firm to suffer a financial loss. For all but the last unit produced, the compensation the firm receives (i.e., the price) will be less than the incremental production cost it incurs. To ensure that the total revenue the firm receives does not fall below the total production cost it incurs, the regulator must raise the authorized price above the level of marginal cost. A price equal to the firm's average cost of production at the prevailing level of production will deliver the highest possible level of welfare to consumers while ensuring that the regulated firm's costs do not exceed its revenues.

1.2 Nonlinear Pricing

A regulator might contemplate implementing a pricing structure in which customers do not simply pay the same unit price for each unit of the product they purchase. For example, each customer might be charged both a fixed participation fee that does not vary with the amount of the product purchased and a constant unit price for each unit that is purchased. Under such a pricing structure, the unit price might be set equal to the regulated supplier's marginal cost of producing the last unit of output it sells and the participation fee might be set to recover exactly the deficit the firm incurs (i.e., the difference between the total revenue it receives from sales and the total production costs it incurs) when it sells its product at the specified unit price. This pricing structure would deliver the maximum possible level of consumer welfare while ensuring non-negative profit for the firm, provided the participation fee is not so onerous as to compel some customers to terminate all purchases of the regulated product (Wilson 1993). In practice, large participation fees are often viewed as unduly onerous for consumers with limited resources. Consequently, particular concern for the well being of these consumers or general concerns with equity and fairness can limit the use of participation fees and other alternatives to constant unit prices.

1.3 Linear Pricing with Multiple Outputs

Now consider the more common setting in which the regulated firm produces multiple products (for example, both rural and urban telephone services). Also suppose the regulator chooses to implement only a constant unit price for each regulated product. Since the prices charged for different products can vary, there are many different ways in which the regulator can raise prices above marginal production costs while ensuring that revenues match total operating costs exactly. To maximize the combined welfare of all consumers, the regulator will implement proportionate mark-ups ($[p_i - c_i]/p_i$) of price (p_i) above marginal cost (c_i) that vary inversely with the sensitivity of consumer demand to price (Ramsey 1927, Baumol and Bradford 1970). The regulator will set the price of a product close to its marginal cost of production when a larger price increase would cause consumers to severely curtail their purchases of the product. In contrast, the regulator will raise the price of a product well above its marginal cost of production when doing so does not cause consumers to reduce their purchases of the product significantly. This pricing policy secures the greatest combined welfare for consumers by limiting the aggregate reduction in consumption that arises as prices are raised sufficiently far above marginal production costs to generate revenues that match production costs.

1.4 Quality Deterioration and Cream Skimming

As noted above, regulators do not always seek solely to maximize the aggregate welfare of consumers. Consequently, in practice, regulated prices are not always raised above marginal costs by greater amounts the less sensitive consumer demand is to price. In fact, some prices (e.g., those for rural, residential telephone service) may be set below marginal production costs while other prices (e.g., those for telephone service to urban business customers) may be set well above marginal production costs. Such prices may arise from fairness and equity concerns, for example. Pricing patterns of this sort can introduce two important problems: quality deterioration and cream skimming.

When the price of a product is set below its marginal cost of production, each extra unit of the product that the firm sells causes its profit to decline. Consequently,

the firm would secure greater profit if consumers purchased fewer units of the product. One way to induce consumers to purchase fewer units of a product is to reduce the quality of the product. Therefore, when prices are set below marginal production costs, the regulator may need to impose penalties on the firm if the level of service quality it delivers is found to be inadequate.

When prices are set above cost for some products and below cost for other products, competitors may find it profitable to supply only those products for which price exceeds cost and leave the incumbent regulated firm to supply those services for which price is set below cost. For example, unless they are prohibited from doing so, competitors may provide telecommunications service only to large businesses in urban areas. This practice is known as cream skimming. A regulator can prevent cream skimming by barring potential competitors of the incumbent regulated firm from providing services in the regulated industry. Such entry prohibitions, though, can hinder industry performance by precluding the operation of firms that have lower operating costs than the incumbent supplier and by removing the discipline that the threat of entry can impose on the incumbent firm. Furthermore, competitors are sometimes able to circumvent prohibitions on entry by developing and marketing substitute products. A regulator can limit incentives for cream skimming by setting prices that more closely reflect the incumbent firm's operating costs. When prices reflect costs in this manner, entry will typically be encouraged only when a competitor can satisfy consumers' needs at lower cost than the incumbent supplier can.

2. *Regulating With Limited Information: Incentive Regulation*

When, as presumed above, the regulator is well informed about the firm's operating capabilities, the regulator can set prices to generate revenues that exactly match production costs when the firm operates efficiently (i.e., when it incurs only the minimum possible cost required to serve customers). When the regulator is not perfectly informed about the firm's capabilities, the regulator cannot simply mandate that the firm operates efficiently. Instead, the regulator must induce the firm to minimize its operating costs. There are a variety of techniques the regulator can employ to do so. These techniques constitute elements of what is termed incentive regulation.

2.1 *Regulatory Lag*

When prices are set to match realized production costs only periodically (e.g., once every three or four years) rather than continually, the firm derives some financial benefit from reducing its operating costs. The larger the period of time before prices are reduced to match lower realized operating costs, the more pronounced is the firm's incentive to reduce its costs. On the other hand, the longer prices are permitted to remain above operating costs, the longer consumer welfare is held below the level that could be secured given the firm's realized costs. This tradeoff determines the optimal length of time between price revisions, i.e., the optimal regulatory lag.

2.2 *Price Cap Regulation*

Price cap regulation has enjoyed increasing popularity in recent years. It incorporates a form of regulatory lag while delegating some pricing discretion to the regulated firm. The typical price cap regulation plan specifies a price cap period (often four or five years) and a maximum rate at which the regulated firm's inflation-adjusted prices can rise annually, on average, during this period (Sappington in press). The firm's realized costs and earnings are usually reviewed at the end of the specified price cap period, and often employed to update estimates of the pricing restriction that would allow the regulated firm to secure a reasonable level of earnings in the next price cap period. When the firm's realized performance is employed to set future standards, the firm's incentives for exceptional performance are dulled. However, price cap regulation provides some incentives for cost reduction by instituting an explicit regulatory lag.

By controlling the rate at which regulated prices can increase on average, price cap regulation affords the firm some discretion in setting prices for individual services. This discretion can help the regulated firm to limit undesirable cream skimming by realigning prices to more closely match production costs. However, this discretion can also enable the firm to raise certain prices that the regulator prefers to remain low (e.g., the price of local telephone service to rural residential customers). To limit the firm's ability to implement undesired pricing structures, price cap regulation plans often prohibit the firm from raising significantly the prices of certain specified products.

2.3 *Earnings Sharing Regulation*

By providing incentives to reduce operating costs, price cap regulation can simultaneously provide incentives to reduce service quality, since superior quality is often more costly to provide. By allowing prices to diverge from realized costs for considerable periods of time, price cap regulation can also admit earnings that depart significantly from expected levels. To mitigate these potential problems, an earnings sharing plan is sometimes appended to price cap regulation. An earnings sharing plan requires the firm to share with its customers a fraction of realized earnings above (and

often below) specified thresholds. Such sharing helps to insure the firm and the regulator against extremely high (and low) earnings, but at the expense of reduced incentives for exceptional performance (Lyon 1996).

2.4 Yardstick Regulation

In certain settings, yardstick regulation can be employed to provide strong incentives for superior performance while limiting the risk of extreme earnings. Under yardstick regulation, the standards to which one firm is held are based upon the performance levels achieved by other firms. To illustrate, in the British water industry, the prices and service quality standards that are imposed on the monopoly provider in one jurisdiction are linked to the costs and service quality levels achieved by monopoly producers in other geographic regions (Cubbin and Tzanidakis 1998).

When firms face similar operating circumstances, yardstick regulation can enable the regulator to impose a reasonable but challenging performance standard on the regulated firm without basing this standard on the firm's own historic performance. Consequently, yardstick regulation can ensure reasonable earnings for the regulated firm while providing the firm with strong incentives for superior performance. However, these desirable outcomes can only be secured via yardstick regulation if the firms are known to face similar operating circumstances or if all relevant differences in their operating circumstances are well understood (Shleifer 1985). In the absence of this knowledge, the standards imposed under yardstick regulation can result in extreme earnings, much like under price cap regulation. Yardstick regulation may also encourage collusion among the firms that are implicitly being compared. By agreeing to reduce their performance below achievable levels, the firms can effectively relax the constraints imposed by yardstick regulation.

2.5 Franchise Bidding

Even when there is a single regulatory jurisdiction and a single producer in the industry, a regulator may still be able to harness the discipline of competition by structuring competition for the industry. The regulator may allow potential producers to bid for the right to serve as the sole producer in the industry. When the potential producers have similar abilities and similar information about operating circumstances, intense bidding competition for the market can secure for consumers the low prices and high service quality that competition among producers in the market might otherwise provide. However, successful franchise bidding requires that all relevant performance requirements be specified completely in advance, and that the regulator and firm both adhere precisely to the terms of the franchise contract once a producer has been selected. If the regulated firm can renege on its responsibilities after it is installed as the sole producer, or if the regulator can demand more services from the selected producer than it originally agreed to provide, franchise bidding can fail to deliver the maximum possible attainable level of consumer welfare (Williamson 1976).

2.6 Regulatory Options

When the regulated firm has better information than the regulator about its operating capabilities, consumers are often best served when the regulator allows the firm a choice among regulatory plans. A careful structuring of the options afforded the firm can induce it to employ its superior information to choose a plan that best serves consumers (Laffont and Tirole 1993, Armstrong and Sappington in press).

To illustrate this point, suppose that a regulator decides to impose price cap regulation on a firm rather than allow the firm to choose between price cap regulation and cost-plus regulation, whereby prices are re-set frequently to match realized production costs. To limit the likelihood of financial distress, the regulator may not insist that prices decline rapidly. In contrast, when the firm has the option to choose cost-plus regulation, the regulator can impose more stringent limits on price increases under price cap regulation without the fear of causing financial distress. If the firm believes that it will be unable to realize adequate earnings under the proposed price cap regime, it will choose to operate under cost-plus regulation instead. But if the regulator has not overestimated the firm's capabilities in setting the challenging requirements of the price cap regime, the firm will choose this regime and deliver its pronounced benefits to consumers.

3. Regulation in Vertically Related Industries

Sometimes, a regulated firm may have the ability to supply both inputs and final products. Inputs are products that firms employ to make other products. Final products are those that customers consume, and that are not used to make other products. For example, a local telecommunications network operator may sell to long-distance telecommunications firms the local network access (an input) they must have to complete the long-distance messages they transport. In addition, the local network operator may have the ability to deliver long-distance telephone calls (a final product) itself. Two fundamental questions arise in such settings: (a) should the monopoly supplier of an essential

input be permitted to compete downstream, i.e., against the firms to which it sells the input? (b) how should the price of the essential input be regulated?

3.1 Regulating Industry Structure

One potential benefit of allowing an input supplier to operate downstream is the additional competitive pressure that the input supplier may impose on other downstream producers. The additional competitive pressure can serve to reduce industry prices, and thereby enhance consumer welfare. An input supplier may be a particularly strong competitor when economies of scope are present. Economies of scope exist if multiple products can be supplied at lower cost when they are produced simultaneously by a single firm rather than when they are produced in isolation by distinct firms (Baumol et al. 1982 p.71).

A potential cost of authorizing downstream participation is that it may encourage the input supplier to disadvantage its downstream rivals unfairly. For example, the input supplier may intentionally reduce the quality of the essential input that it sells to its downstream competitors in order to make it more difficult for the competitors to provide high quality service at low cost. However, the input supplier may not have an unambiguous incentive to disadvantage its downstream rivals, because their demand for the essential input typically declines as the quality of the input declines. This reduced demand can reduce the profit derived from supplying the essential input (Weisman in press). The magnitude of the profit the input supplier foregoes by reducing the demand for the input it supplies is greater the higher the price the firm is permitted to charge for the input, *ceteris paribus*. Consequently, raising the price of an essential input above its marginal cost of production can reduce the incentive of an input supplier to disadvantage its downstream rivals.

3.2 Regulating Input Prices

Raising the price of the essential input above its marginal cost of production can also help to limit cream skimming and ensure that downstream production is carried out by the least-cost suppliers. To illustrate this point, suppose that if the input supplier operates downstream, the prices it charges there are regulated. Further suppose that the price of one downstream product (called product A) is set at a level (p^A) that exceeds the supplier's marginal cost of production (perhaps to facilitate a price below marginal cost for another product). If the price of the essential input is set equal to its marginal cost of production in this setting, a downstream producer may be able to serve all consumer demand for product A profitably even though its marginal cost of production ($\tilde{c}^A$) exceeds the corresponding cost (c^A) of the input supplier. The higher-cost producer can set a price below the regulated price for product A and still secure positive profit if its cost disadvantage ($\tilde{c}^A - c^A$) is smaller than the regulated profit margin on product A ($p^A - c^A$). To preclude operation by higher-cost competitors, the price of the essential input can be raised above its marginal cost of production by the amount of profit ($p^A - c^A$) the input supplier foregoes when it provides the input to competitors (Laffont and Tirole 1994, Baumol et al. 1997, Armstrong in press). One potential drawback to this pricing policy, though, is that it provides little incentive for the input supplier to reduce its operating costs.

4. Additional Forms of Regulation

Regulation of producers' activities can help to enhance consumer welfare even when industry cost minimization does not require operation by a single producer. To illustrate this fact, consider the potential value of environmental regulation and health and safety regulation.

4.1 Environmental Regulation

In certain instances, production activity can impose costs on society in addition to the direct physical costs of the inputs used in production. For example, some production processes may expel harmful pollutants into the surrounding land, water, or air. If no restrictions are placed on producers' rights to expel pollutants into the environment, producers may produce more than the surplus-maximizing level of pollution. This is the level of pollution at which the social damage that would be caused by an additional unit of pollution is exactly equal to the social cost that must be incurred to eliminate that additional unit of pollution. Pollution in excess of the surplus-maximizing level imposes social damage that exceeds the cost of eliminating the extra pollution. In contrast, the cost of reducing pollution below the surplus-maximizing level exceeds the social benefits derived from the reduction in pollution.

Various regulatory policies can be employed to control environmental pollution. For instance, direct taxes (effluent fees) can be imposed on pollution when it is readily observed and measured. By setting the tax on pollution equal to the marginal social damage caused by the pollution, the producer can be induced to realize the surplus-maximizing level of pollution. When the level of pollution produced by a firm is difficult to measure, taxes can be imposed instead on the products whose production causes pollution, or on inputs (such as gasoline) to the activity (automobile driving) that causes pollution. The use of specific

pollution abatement technologies can also be mandated.

Recently, some governments have introduced tradable pollution permits. Each permit entitles its bearer to emit one unit of pollution (e.g., one ton of sulfur dioxide). Polluting firms are able to buy and sell permits at prices determined by the available supply of and demand for these permits. A key potential benefit of tradable permits is that they promote pollution abatement at minimum cost. Producers that are able to reduce the pollutants they expel at a cost below the prevailing price of a permit will find it profitable to do so and sell any permits that they own. In contrast, producers for whom pollution abatement is more costly than the price of a permit will continue to expel pollutants but will pay for the right to do so (Schmalensee et al. 1998).

4.2 Health and Safety Regulation

Pollution abatement regulation is often intended to improve the health and well being of citizens. Other forms of health and safety regulation can serve this same purpose. The regulations come in many forms, including information requirements, standards, and prohibitions.

Regulations that require manufacturers to provide relevant product information to consumers can help consumers choose among available products. These regulations include requirements that product contents be fully specified and that the potential health risks associated with consumption be stated clearly. In contrast, direct regulation of product standards and outright prohibitions on the sale of certain products can substitute information collection and decision-making by a regulatory body for decision-making by consumers. For example, when the government imposes minimum home construction standards or when it bans certain drugs from the marketplace, the government eliminates the option a consumer might otherwise have to purchase below-standard homes or banned drugs. One rationale for such regulations is that they can implement the decisions that consumers would make if they were fully informed about all relevant benefits and costs of consumption, without requiring consumers to develop the expertise and obtain and evaluate the information required to assess the benefits and costs (Viscusi 1992).

5. Competing for Regulatory Benefits

The preceding discussion illustrates some of the ways in which government regulation of producers can be designed to benefit consumers. It is important to note, however, that regulation need not only serve this role in practice. Since regulators typically have the power to deliver benefits to some parties and impose costs on others, the affected parties will have an incentive to try to secure favorable treatment from regulators. When industry producers are better organized and have more at stake than industry consumers, regulation that benefits producers at the expense of consumers can arise (Stigler 1971, Peltzman 1989). For example, new competitors may be barred from an industry even though their entry and operation would be in the best interest of consumers. These realistic possibilities merit careful consideration when evaluating the benefits and costs of existing and proposed regulations.

See also: Business Law; Environmental Health and Safety: Social Aspects; Environmental Planning; Policy Process: Business Participation; Regulation and Administration; Regulation: Empirical Analysis; Regulation Theory in Geography; Regulatory Agencies

Bibliography

Armstrong M (in press) The theory of access pricing and interconnection. In: Cave M, Sumit M, Vogelsang I (eds.) *International Handbook of Telecommunications Economics*. North-Holland Publishers, Amsterdam

Armstrong M, Doyle C, Vickers J 1996 The access pricing problem: a synthesis. *Journal of Industrial Economics* **44**(2): 131–50

Armstrong M, Sappington D (in press) Recent developments in the theory of regulation. In: Armstrong M, Porter R (eds.) *Handbook of Industrial Organization*. Elsevier Science Publishers, Amsterdam

Baumol W, Bradford D 1970 Optimal departures from marginal cost pricing. *American Economic Review* **60**(2): 265–83

Baumol W, Ordover J, Willig R 1997 Parity pricing and its critics: a necessary condition for efficiency in the provision of bottleneck services to competitors. *Yale Journal on Regulation* **14**(1): 145–64

Baumol W, Panzar P, Willig R 1982 *Contestable Markets and the Theory of Industry Structure*. Harcourt Brace Jovanovich, Inc., New York

Cubbin J, Tzanidakis G 1998 Regression versus data envelopment analysis for efficiency measurement: an application to the England and Wales regulated water industry. *Utilities Policy* **7**(1): 75–85

Laffont J, Tirole J 1993 *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA

Laffont J, Tirole J 1994 Access pricing and competition. *European Economic Review* **38**(9): 1673–710

Lyon T P 1996 A model of sliding-scale regulation. *Journal of Regulatory Economics* **9**(3): 227–47

Peltzman S 1989 The economic theory of regulation after a decade of deregulation. *Brookings Papers on Economic Activity: Microeconomics*. The Brookings Institution, Washington, DC, pp. 1–59

Postal Reorganization Act 1970 *United States Public Law Number 91-375, 84 Statute 719*

Ramsey F 1927 A contribution to the theory of taxation. *Economic Journal* **37**: 47–61

Sappington D (in press) Price regulation and incentives. In: Cave M, Majumdar S, Vogelsang I (eds.) *International Handbook of Telecommunications Economics*. North-Holland Publishers, Amsterdam

Schmalensee R, Joskow P L, Ellerman A D, Montero J, Bailey E 1998 An Interim evaluation of sulfur dioxide emissions trading. *Journal of Economic Perspectives* **12**(3): 53–68
Shleifer A 1985 A theory of yardstick competition. *Rand Journal of Economics* **16**(3): 319–27
Stigler G 1971 The theory of economic regulation. *Bell Journal of Economics and Management Science* **2**(1): 3–21
Viscusi K 1992 *Fatal Tradeoffs*. Oxford University Press, New York
Webster's Encyclopedic Unabridged Dictionary of the English Language 1989 Portland House, New York
Weisman D (in press) Vertical integration. In: Madden G, Savage S (eds.) *Handbook of Telecommunications Economics*. Edward Elger Publishers, Aldershot, UK
Williamson O 1976 Franchise bidding for natural monopolies—in general and with respect to CATV. *Bell Journal of Economics* **7**(1): 73–104
Wilson R 1993 *Nonlinear Pricing*. Oxford University Press, Oxford

D. E. M. Sappington

Regulation: Empirical Analysis

The economic regulation of firms and industries constitutes an extreme form of government intervention in markets. Government antitrust, competition, tax, and trade policies, as well as most environmental, health, safety, and information regulations, shape but do not replace the market, although the most severe of these policies may blur that distinction. Economic regulation stands in sharp contrast. In industries subject to economic regulation, government agencies or departments exercise considerable control over firms' entry, pricing, investment, and product choice decisions, and administrative decisions replace market outcomes. Intervention may go as far as replacing private ownership with government-owned and -operated enterprises.

The modern field of empirical regulatory economics developed during the 1960s and early 1970s, as economists began 'to apply the familiar principles of microeconomics in comprehensive fashion to the subject' (Kahn 1988, p. xv). George Stigler, whose contributions to the field were recognized in his 1982 Economics Nobel prize citation, sparked much of this growth. Through the 1980s, economists focused on quantifying the effect of regulation on firm behavior, industry outcomes, and consumer welfare. The diversity of regulatory forms and industry characteristics across the transportation, energy, utility, financial, and communication sectors provided a rich set of institutions for investigation. Significant policy reforms, even deregulation, in structurally competitive sectors during the 1970s and 1980s provided new sources of variation in the regulatory environment and invigorated empirical regulatory research. More recent attention has turned to the traditional 'public utility' sector, particularly the telecommunication and electric utility industries. Market-based institutions have begun to replace state-ownership and regulation, as technological change called into question the natural monopoly assumption in many of these industries and policy interest in privatization raised questions of how best to restructure and sell-off vertically-integrated monopolies. Analyzing the operation of these redesigned markets has become a main focus of empirical regulatory research.

This article focuses on empirical analysis of the effects and reform of economic regulation. Section 1 describes the dominant empirical methodologies. Section 2 sketches the evidence on the effects of traditional economic regulation. Section 3 discusses recent analyses of market design and regulatory scope. To keep the reference list manageable, readers are referred to surveys and collective volumes wherever possible. These guide readers interested in delving more deeply into particular issues to a wealth of primary studies.

1. *Empirical Methodologies*

Measuring the effect of economic regulation requires the establishment of a counterfactual benchmark: what would firm or industry performance look like in the absence of regulation? A variety of approaches to this have been used. The most compelling lesson of the existing literature is that the impact of economic regulation depends critically upon its particular institutional form and the characteristics of the industry under study. Measuring the impact of regulation requires as much attention to the details of how regulators operate in practice as to the prescribed legal form of the regulation. Paul Joskow's (1974) study of electric utility regulation in the US provides a persuasive exposition of this argument.

The most common methodology used to measure regulatory effects compares outcomes in regulated markets with those in unregulated markets. Time series variation may facilitate this comparison; the wave of US deregulation in the late 1970s and early 1980s generated numerous 'before-and-after' studies of regulation. Differences across political jurisdictions, e.g., across states within a country or across countries, may provide another source of variation in regulatory regimes. Empirical studies also may use differences in the intensity or form of regulation to infer its effects, particularly if regulation is likely to bind only under certain well-defined circumstances.

Two conditions limit the usefulness of this approach. First, unregulated comparisons may not exist. Industries that are regulated in one political jurisdiction tend to be regulated or government-owned in other political jurisdictions. Regulation may be so

long-lived that comparing current regulated performance with historical unregulated performance is meaningless. Second, this method implicitly assumes that differences in regulatory regimes are exogenous to firm or market performance. This is often false. If, for example, jurisdictions impose electricity regulation in response to high or rapidly increasing local electricity rates, comparing price levels in regulated and unregulated locales might suggest no effect even if regulation does, in fact, constrain prices. Using panel data to estimate effects from 'differences in differences' econometric models may be useful in some cases, as may developing more complete models of the determinants of regulation.

A second methodology estimates a structural model of the market, forms assumptions about behavior in the absence of regulation, and compares simulated unregulated outcomes with those observed under regulation. This technique may be useful even in the absence of regulatory variation. It is most appropriate when regulatory constraints can be clearly specified. Suppose regulators constrain prices but do not influence investment or factor use decisions. Structural estimates of underlying demand and cost functions permit one to predict unregulated profit-maximizing prices and compare these with current regulated prices. If regulated firms do not minimize costs, however, or if regulatory intervention alters investment decisions in a systematic way, then the regulated cost function may be a poor estimate of the unregulated long-run cost function, and the implied price comparison will be misleading.

A third approach uses experimental data to measure regulatory effects. Field experiments, such as the time-of-use electricity pricing studies conducted in the US during the early 1970s, can reveal responses to alternative regulatory or market designs. In the best of circumstances, these can provide information on how market outcomes respond to exogenous variation in regulatory mechanisms, holding other economic conditions constant. Their cost limits their widespread use, however, and their temporary nature may preclude an assessment of long-run equilibrium responses to alternative policies. Laboratory experiments, which present volunteers (often students) with a set of objectives, operating guidelines, and payoffs, also have been used in a limited way to simulate responses to alternative regulatory designs.

2. *The Effects of Traditional Economic Regulation*

It is impossible to describe succinctly the full effects of economic regulation across industries, over time, and through space. Regulation may influence prices, costs, and the pace of technological change in an industry. This may lead to the creation or suppression of economic rents, and change the distribution of rents across various stakeholders or interests in the industry. These effects are likely to vary considerably across markets and regulatory forms, and preclude simple generalizations. This article highlights some illustrative results. Further detail is provided in the many reviews referenced at the end of this article and the primary sources they cite.

2.1 *Empirical Analysis of Natural Monopoly Regulation*

Industries such as electricity supply, telecommunications, natural gas transmission and distribution, water and sewer service, cable television service, and to some extent railroads have historically been considered 'natural monopolies.' This applies when total production costs are minimized by providing service to all customers in a given market through a single firm. As economists since Adam Smith have recognized, minimizing production cost may not minimize price, since monopolies use their market power to mark up price above competitive levels and earn monopoly rents. This historically engendered one of two policy responses. Government ownership in natural monopoly sectors was typical outside the USA, and was used frequently in the USA for water and sewer service, but less so for electric utilities and cable television service. Government regulation of privately-owned firms was the dominant US response in most other cases. In this scheme, regulators typically grant a legal, 'franchise' monopoly to a single provider in a geographic market, but retain control over price to mitigate the monopoly mark-ups that an unconstrained legal monopolist would choose. Prices are determined generally by commissions that review costs, and often investment decisions, and set prices to cover allowable costs and a 'fair rate of return' on 'prudent' capital investments.

Early economic research on the regulated public utility sector was generally theoretical and focused on how to perfect regulatory pricing, treating as given the desirability and general effectiveness of government regulation. Optimal pricing questions also arose for state-owned utility enterprises, and French and British economists were in the vanguard of research in this area (Kahn 1988, Laffont 1994, and *Poverty: Measurement and Analysis*). Stigler and Friedland's (1962) study of the early electricity industry was one of the first empirical analyses of the actual effect of utility regulation, and concluded that regulation did little to reduce prices below unregulated monopoly levels. While the particulars of this research generated controversy at the time, and subsequent work suggests that electricity prices in the Friedland–Stigler sample were, in fact, lower for regulated firms (Peltzman 1993), the notion that regulation might not improve

upon market outcomes even in those sectors with the strongest public interest rationale stimulated considerable additional research.

The literature subsequent to Stigler and Friedland (1962) suggests generally that price regulation, particularly in the electricity and cable television industries, has constrained average prices below the level an unconstrained legal monopolist would choose. Analysis of average price levels masks the effect of regulation on price structures, however. Price structure analyses in electricity and telephone industries suggest redistributive rather than Ramsey-like efficient pricing objectives, often with considerable transfers from customers with relatively elastic demands to those with relatively inelastic demands. For example, regulatory cross-subsidies from long-distance telephone rates to basic residential telephone service and from urban to rural telephone customers were a major force driving new entry into the US long-distance sector, leading ultimately to the 1984 divestiture of AT&T and the regulatory reform of the telephone industry. The tenacity of these cross-subsidies more than 15 years later testifies to the strength of political interests supporting these policies. Some economists have estimated the social welfare cost of the residential service cross-subsidy to be as high as $7.0 billion per year (in Peltzman and Winston 2000, p. 78).

The effect of natural monopoly regulation on static production costs is less well-measured. One of the most developed theoretical and empirical literatures on incentives for excessive capital investment in rate-of-return regulated industries proved largely inconclusive (Averch and Johnson 1962, Joskow and Rose 1989). There is some evidence of higher wages in firms that expect regulators to flow increased labor costs through to higher consumer prices, though wages do not tend to be higher overall in regulated sectors (Hendricks 1975, Peoples 1998). In the railroad industry, low regulated profit rates led to underinvestment in track maintenance and repair on many lines, reducing efficiency and increasing average transit time over the track.

Of particular concern is the effect of regulation on the pace of technological innovation. This may occur from direct regulatory constraints on new process or product introductions, such as the Federal Communications Commission's delay in allocating radio spectrum to wireless telephone systems in the USA and its restrictions on enhanced telephone service offerings. Indirect regulatory effects may also influence the pace of technological change by altering expected payoffs to investment in new technology. While often difficult to measure, these costs can be enormous. For example, Hausman estimated lost surplus from the delayed introduction of wireless phone service in the USA at more than $30 billion annually (in Peltzman and Winston 2000, p. 80).

The ambiguity of cost and innovation effects may arise in part from the difficulty of establishing an appropriate empirical counterfactual benchmark to regulation, an essential prerequisite to measuring either static or dynamic efficiency effects. Nevertheless, ongoing policy concerns about the efficiency of production, investment incentives, and the pace of technological change in natural monopoly sectors have fueled much of the recent movement toward market-based reforms, discussed in Sect. 3.

2.2 Empirical Analysis of Regulation in Multifirm Industries

A substantial portion of the regulatory apparatus built during the twentieth century involved multifirm industries, such as trucking, airlines, natural gas and oil production, property and liability insurance, and banking. Regulation in many of these was rationalized by the deleterious effects of excessive, rather than insufficient, competition, particularly during the depression of the 1930s. In others, such as airlines, regulation was advocated to promote the orderly development of an emerging commercial industry. Many of these regulatory systems were dismantled or reformed substantially between 1975 and 1985, reflecting a broad consensus on its social costs.

Multifirm price regulation usually requires all firms serving a given market to adhere to a standard tariff. In some industries, this tariff was set by the regulatory commission, as the Civil Aeronautics Board did for the US airline industry. In others, tariffs were set collectively by the firms and ratified by the regulatory commission, as the Interstate Commerce Commission did for trucking rate bureau filings. Regulation generally discourages entry or market extension. By eliminating the ability of existing firms to discount rates and the ability of new firms to enter the market, regulation limits competition and tends to raise prices significantly. For example, economists estimate that regulated prices in the late 1970s were at least 15—20 percent above competitive levels in the trucking industry and 25—30 percent above unregulated prices in the airline industry.

High prices generate significant potential rents. The value of these rents is embodied frequently in the market value of regulatory assets, such as taxicab medallions, operating certificates for regulated trucking firms, or state liquor licenses. Asset values may reflect only a portion of the rents generated by regulation, however. In some sectors, organized labor was able to garnish a substantial share of regulatory rents (Rose 1987, Peoples 1998). Members of the Teamsters Union, for example, were a major beneficiary of trucking regulation, with wages 50 percent above those for nonunion truck drivers and drivers outside the regulated industry. In other industries, regulatory rents were dissipated through various rent-seeking behaviors. This seems to have been particularly

acute in the airline industry, where nonprice competition repeatedly eroded load factors and profit margins despite increasing regulated prices. In the rare situations in which regulatory policy effectively constrains maximum prices and exit from markets, rents may be negative. In the short-run, this can transfer quasi-rents from owners of regulated capital to consumers of their products. Over the longer-run, its adverse effect on investment tends to erode both the quantity and quality of service provided. The rash of bankruptcies in the US railroad industry during the 1970s and early 1980s, and natural gas shortages during periods of below-market ceiling prices in the 1960s and 1970s, were consequences of this type of policy.

As in the natural monopoly sectors, regulation influences the structure of prices as well as the average price level. Economic regulation often leads to a more uniform price structure than in unregulated markets. Prices tend to be less responsive to cost differences across consumers or over time, due in part to cross-subsidies for favored higher-cost consumer groups and to aversion to congestion-based or peak-load pricing. This distorts both capacity utilization and firms' investment incentives. These effects have been analyzed in the context of traffic mix distortions between railroads and trucking firms, provision of air service to small communities, and excessive pooling in insurance markets.

Postderegulation experience suggests that regulation often suppressed variation due to price discrimination, relative to the unregulated market. Airlines are perhaps the best example of this. It is important to recognize that price discrimination, while controversial, may enhance overall welfare relative to a uniform price structure; see the literature on Ramsey–Boiteaux pricing (Kahn 1988). In airline markets, for example, the pervasive use of low discount fares to fill seats that would otherwise depart vacant increases substantially the population for whom air travel is viable, and raises airline revenue per available seat-mile. The net welfare effect of current levels of price discrimination is difficult to measure, and has not been directly calculated for any deregulated market.

Economic regulation in multifirm industries has been associated with both direct and indirect cost effects. Regulatory restrictions and rigidities may prevent firms from optimizing production in both static and dynamic contexts. For example, entry restrictions in the trucking industry led to high rates of partial or empty backhauls and circuitous routings, raising the cost of transporting a given commodity flow. Unit banking regulations, which restrict a firm to a single physical location in one state, both limit realizations of economies of scale in provision of service and insulate less efficient banks from takeovers by more efficient competitors. Inefficiencies such as these generate potentially huge regulatory deadweight losses (Winston 1998).

The welfare effects of cost increases may be difficult to quantify when associated with offsetting changes in the product provided to consumers. For example, the suppression of price competition may encourage nonprice competition, increasing both costs and quality relative to outcomes in unregulated markets. In the regulated airline industry, increased flight frequency and rapid adoption of new aircraft increased both costs and value to passengers. Measuring cost or productivity effects without controlling for the value of higher service quality leads one to overstate the adverse welfare effects of regulation in situations like this. Recent advances in structural demand estimation may help to quantify the value of particular quality attributes, though both data and modeling limitations continue to make this one of the weaker empirical links in most assessments of the overall effects of regulation.

3. *The Effects of Restructuring Regulated Natural Monopolies*

The focus of regulatory policy and research changed sharply in the late 1980s and 1990s, with the ascendancy of market-based institutions over state-ownership and regulation in even traditional 'natural monopoly' sectors. Restructuring, requiring firms to divest along vertical segment lines, has become common. Segments viewed as structurally competitive, such as electricity generation, have been increasingly deregulated. Network segments that continue to operate under natural monopoly cost conditions, such as electricity transmission grids or distribution systems, generally are required to provide access to competitive suppliers, often at terms subject to continued regulatory oversight. Where economic regulation has been retained, it has become more market-oriented. This reflects both dissatisfaction with recent performance under regulation and a response to theoretical research emphasizing attractive properties of incentive-based regulatory schemes (see *Poverty: Measurement and Analysis*)

These changes raise new questions about market design and regulatory interfaces. The emerging empirical literature has focused on three main issues: the impact of privatization, the performance of restructured sectors, and the effects of incentive-based regulation. Policy in this area is very much in flux, as the initial wave of enthusiasm for market solutions confronts unanticipated consequences from their implementation.

3.1 *The Impact of Privatization*

Privatization has been motivated by a range of concerns including the cost of service, pace of technological innovation, and impact of continuing subsidies on government budget deficits. A review of the

voluminous literature on privatization is beyond the scope of the present article; see Vickers and Yarrow (1991), La Porta and López-de-Silanes (1999), and their references. Because privatization has been so closely linked with restructuring and regulatory reform initiatives over the past quarter-century, however, it is useful to highlight key empirical results.

One of the most robust empirical results from this literature is the dramatic impact privatization typically has on cost reduction. This is particularly evident in labor costs. Substantial labor shedding has accompanied privatization almost everywhere, with labor force reductions of 50 percent or more a common result. Where comparisons with privately-owned firms are feasible, studies suggest that postprivatization labor rates approach those at firms that were never governmentowned. This suggests considerable excess labor in state-owned enterprises. This may reflect the use of state-owned enterprise to promote certain social policies as well as direct operating inefficiencies, suggesting that excess labor costs may include both transfers as well as social dead-weight loss.

Privatization often leads to both price and output increases. Rising output despite higher prices suggests the elimination of artificial constraints, such as rationing, or higher quality service for which consumers are willing to pay more. In numerous studies of utility sectors, waiting time for service falls and investment and service quality increase following privatization. This seems most common in sectors that were capital-starved or in which service was underpriced under state-ownership. Even in countries where state-owned utilities were perceived to be functioning relatively well, however, increased investment and innovation may be important benefits of privatization.

One caveat is in order: the frequent concurrence of privatization with changes in the competitive environment can make it difficult to empirically disentangle the relative contributions of ownership change, hard budget constraints, and increased competition to improved performance. Moreover, it appears that the benefits of privatization may be greater when accompanied by increased competition or deregulation. Studies vary considerably in their recognition and successful treatment of this problem.

3.2 Performance in Restructured Natural Monopoly Sectors

Vertical restructuring has become common in natural monopoly industries worldwide. In telecommunications, equipment provision and long-distance service were among the first to be carved out of vertically-integrated telecommunications firms and opened to competitive suppliers. Power generation has been divested from transmission and distribution functions and opened to competition in many electricity systems. Much of this policy movement appears stimulated by concerns that prices charged by monopoly providers were too high, relative to those attainable by privately-owned firms operating in competitive sectors. In some cases, this reflects real operating inefficiencies and reduced innovation under monopoly. In other cases, apparently high prices may in part reflect perverse regulatory pricing schemes. Action on these concerns was facilitated by technological innovations that reduced scale economies at certain production stages and enabled decentralized operations to achieve many of the economies of scope previously attained by vertically integrated firms.

Economists have been especially interested in assessing the competitiveness of restructured markets. Empirical research suggests these markets have yet to converge to competitive ideals. In the US long-distance telephone market, for example, AT&T lost its sanctioned monopoly provider status in 1984, following the earlier technological erosion of its 'natural monopoly.' While long-distance prices have declined, the convergence to competitive norms has been relatively slow, with average per-minute prices more than double estimated long-run incremental costs as late as 1998 (Peltzman and Winston 2000, p. 77). Some economists have suggested that this reflects relatively weak competition among the small and decreasing number of existing long-distance providers.

Analyses of electricity markets echo this theme. In most restructured electricity markets, unregulated generators bid to supply electricity to the system operator, which determines the dispatch order of generators to supply the electricity demanded at each moment in time. There is considerable evidence of generator market power in many markets, with generators bidding above their marginal generation costs and strategically withholding capacity from the market to increase prices (see Wolfram 1999, Borenstein et al. 2000, Joskow in Peltzman and Winston 2000, and Joskow and Kahn 2001).

While many restructured markets have been prone to supracompetitive prices, the extent to which price exceeds cost varies tremendously across different settings. There is growing evidence that performance variations are heavily dependent upon the institutional details of the market's design. These include the way the competitive sector is structured, the terms on which suppliers interface with their customers and the network, and the management of network investments. In electricity markets, for example, the relative sizes of different suppliers, their inframarginal capacity, the use of long-term hedging contracts between generators and their customers, and the degree to which customers see and can respond to real-time price variation have all been demonstrated to influence the exercise of market power.

While competition may enhance cost efficiency in the deregulated sector, restructuring may increase other costs. For example, substitution of market mechanisms for internal decision-making or 'com-

mand and control' by a vertically integrated firm may increase coordination costs in electricity markets (Joskow in Peltman and Winston 2000). In California, for example, reserve margins and hence system operating costs have increased as a result of moving to a dispatch market. Local distribution companies may face difficulties when their requirement to serve, which was imposed in exchange for franchise monopoly in many utility sectors, is maintained, but restructuring eliminates their direct control over electricity supplies, as evidenced by California utilities in 2000–2001.

In telecommunications, access issues focus on pricing and interconnection requirements for the 'local loop,' the set of wires connecting individual subscribers to the local network. These problems are magnified by the potential for local exchange carriers to compete with other firms for services beyond the local loop. The determination of access terms has substantial effects for investment incentives by incumbent exchange companies and potential competitors, entry into local, long-distance, and enhanced services, and the performance of competitive markets where entry occurs. This is the focus of much current regulatory policy and some empirical analysis (Crandall and Hausman in Peltzman and Winston 2000).

3.3 *Effects of Incentive-based Regulation*

While regulatory reform over the past two decades has been substantial, economic regulation continues to play an important role in many sectors. Economic regulation in the new millennium is more market-based than its counterpart of 1970, but less market-driven than current rhetoric would suggest. Policy innovation has been particularly notable outside the USA, perhaps because these systems were being designed *de novo* rather than patched on to existing regulatory practices. Regulation has moved away from traditional rate-of-return regulation, which determined prices through cost-based mark-up rules (Kahn 1988). More common are higher-powered incentive schemes such as price caps, which fix price levels or rates of change, and allow firms to keep the surplus generated from unanticipated improvements in efficiency, at least until the next review of the cap or its adjustment factor. The models behind these schemes are discussed elsewhere (see *Poverty: Measurement and Analysis*). It is important to recognize that implementation, however, generally deviates considerably from the theoretical construct.

Many researchers have attempted to measure the impact of incentive-based regulation, particularly in the telecommunications sector. This literature broadly suggests that incentive regulation is accompanied by lower prices. Kridel et al. (1996) suggest that these effects may reflect coincident policies that increase competition more than the incentive effects of incentive regulation itself, however. This interpretation is particularly suggested by the difficulty in finding consistent effects of incentive regulation on costs or productivity, although the short time span of many studies may impede detection of such effects. As earlier empirical work in regulation suggests, one should exercise great caution in affirming the superiority of incentive-based policies based on the theoretical properties of idealized schemes.

4. *Conclusion*

The reforms of the 1990s have created a vast new set of institutions and regulatory policies. This is an area in which empirical analysis greatly lags behind the theoretical literature, and the returns to additional careful work are likely to be significant. Empirical research in regulatory economics, argued by some in the 1980s to belong to the domain of economic history, remains vital to guiding, understanding, and evaluating policy initiatives today.

See also: Cost–Benefit Analysis; Law and Development; Policy Knowledge: Advocacy Organizations; Poverty: Measurement and Analysis; Regulation, Economic Theory of; Regulation, Political Economy of; Socialism

Bibliography

Averch H, Johnson L L 1962 Behavior of the firm under regulatory constraint. *American Economic Review* **52**: 1052–69

Borenstein S, Bushnell J, Wolak F 2000 *Diagnosing Market Power in California's Restructured Wholesale Electricity Market*. National Bureau of Economic Research Working Paper 7868, Cambridge, MA

Hendricks W 1975 The effect of regulation on collective bargaining in electric utilities. *Bell Journal of Economics and Management* **6**: 451–65

Joskow P L 1974 Inflation and environmental concern: Structural change in the process of public utility price regulation. *Journal of Law and Economics* **17**: 291–327

Joskow P L, Kahn E P 2001 *A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market During Summer 2000* National Bureau of Economic Research, Cambridge, MA

Joskow P L, Noll R G 1994 Economic regulation during the 1980s. In: Feldstein M (ed.) *Economic Policy During the 1980s* University of Chicago Press, Chicago

Joskow P L, Rose N L 1989 The effects of economic regulation. In: Schmalensee R, Willig R (eds.) *Handbook of Industrial Organization*. North-Holland, Amsterdam, Vol. 2

Kahn A E 1988 *The Economics of Regulation: Principles and Institutions*. MIT Press, Cambridge, MA

Kridel D J, Sappington D E M, Weisman D L 1996 The effects of incentive regulation in the telecommunications industry: A survey. *Journal of Regulatory Economics* **9**: 269–306

La Porta R, López-de-Silanes F 1999 The benefits of privatization: Evidence from Mexico. *Quarterly Journal of Economics* **114**: 1193–242

Laffont J-J 1994 The new economics of regulation ten years after. *Econometrica* **62**: 507–37

Peltzman S 1993 George Stigler's contribution to the economic analysis of regulation. *Journal of Political Economy* **101**: 818–32
Peltzman S, Winston C (eds.) 2000 *Deregulation of Network Industries: What's Next?* AEI-Brookings Joint Center for Regulatory Studies, Washington, DC
Peoples J 1998 Deregulation and the labor market. *Journal of Economic Perspectives* **12**: 111–30
Rose N L 1987 Labor rent-sharing and regulation: Evidence from the trucking industry. *Journal of Political Economy* **95**: 1146–78
Stigler G J 1971 The theory of economic regulation. *Bell Journal of Economics and Management* **2**: 3–21
Stigler G J, Friedland C 1962 What can regulators regulate? The case of electricity. *Journal of Law and Economics* **5**: 1–16
Vickers J, Yarrow G 1991 Economic perspectives on privatization. *Journal of Economic Perspectives* **5**: 111–32
Winston C 1998 US industry adjustment to economic deregulation. *Journal of Economic Perspectives* **12**: 89–110
Wolak F 2001 *An Empirical Analysis of the Impact of Hedge Contracts on Bidding Behavior in a Competitive Electricity Market*. National Bureau of Economic Research, Cambridge, MA
Wolfram C D 1999 Measuring duopoly power in the British electricity spot market. *American Economic Review* **89**: 805–26

N. L. Rose

Regulation: Family and Gender

There are few more politically sensitive areas than the role the state should play in regard to 'the family.' It is not easy for liberal democratic states to justify intervention in the private sphere of personal relationships and family life. In the first place memories of extreme examples of state intervention during the twentieth century, most notably in Nazi Germany, where the state dictated who married and who had the right to reproduce, make this difficult territory. Second, even when governments have little by way of explicit family policies, they usually make implicit assumptions about what families should look like, which in pluralist societies are bound be at odds with the way in which some people organize their affairs. Not only have modern Western states tended to assume that the family will consist of two parents and children, but historically both government policies and the professional experts employed by the state have also made gendered assumptions about the part played by men and women in family life, particularly in respect of responsibilities for earning and for caring, which are increasingly at odds with the social reality.

Feminists have been one of the most vocal groups criticizing government policies on the family, but the relationship between parents, children, and the state has been one of the most hotly debated issues of the twentieth century, as important as the more common touchstone of individualism vs. collectivism. Gender, understood as a fundamental organizing principle in society (Scott 1988), is crucial to the analysis of the relationship between the state and the family. The concept of 'gendered divisions,' particularly in respect of resources, labor, and investments, is key to understanding the development and impact of the public policies and family law that regulate the family.

1. *Ideas about the Nature of State Regulation and the Family*

The debate about state regulation, the family, and gender in the last quarter of the twentieth century has revolved around the extent to which states have propped up a particular family form, but the positions taken are complicated and do not follow a simple political division between Left and Right.

Feminists have argued that for governments, family policy all too often meant ways of bolstering the traditional two-parent family, with a male breadwinner and an economically dependent female carer, and that policies rarely impacted on different family members in the same way. For example, in the 1970s, Scandinavian feminist writers proposed that the modern welfare state amounted to a new form of 'public patriarchy,' the argument being that the dramatic increase in women's employment, which began in Sweden and Denmark in the 1970s, involved for the most part women leaving their homes to engage in similar caring work, often for children, in the service of the state. In the UK, Wilson (1977) suggested that the postwar welfare state amounted to nothing less than the state organization of domestic life, by which she meant that state policies, for example in respect of cash benefit entitlements, sought to reinforce the male breadwinner family model. A few years later, a similar kind of criticism was made from an entirely different quarter.

Arguing from a 1980s New Right position, Mount (1982) suggested that the family had been in 'permanent revolution' against the unwarranted intrusions of the state. In his view, even apparently helpful public visitors, such as home nurses, were essentially seeking to impose standardized modes of care on families and were able to invoke a wide range of powers if they deemed the family to have 'failed' in some respect. Mount's view keyed into the long-standing suspicions of working people, especially in the English-speaking countries, of officialdom. However, this does not mean that nonintrusive, nonstigmatizing state intervention has not been welcomed. The panoply of Swedish collective social provision has been referred to by the citizens of that country as 'the people's home.'

The idea that state officials practice some kind of surveillance over the family also found favor with followers of Foucault, although Donzelot (1979) stressed the extent to which the professionals in emerging welfare states formed alliances with women

in order to do so. In her history of domestic violence, Gordon (1990) also stressed the extent to which social workers had sided with the weaker family members against the more powerful, thus questioning the extent to which state regulation worked to reinforce traditional power relationships within the family. Indeed, most feminists, while criticizing the kinds of assumptions about the family inherent in social policies, have also looked to the state to work for gender justice (Okin 1989). Few have followed Elshtain (1981) in defending the family as a separate, private sphere, outside the reach of the state.

In recent years the very different idea that the state has played a part in deregulating family life has gained in strength. The argument is that rather than working to uphold traditional family patterns, the state has effectively underpinned a trend towards individualistic behavior that is essentially selfish. This view has been formulated in the context of the dramatic postwar statistics on family change. Marriage rates have fallen in most Western countries; people are marrying later; fertility rates (especially in some Western European countries) have fallen well below replacement rates; divorce rates have increased since the 1970s and remain high and stable in Northern Europe and the US; rates of extramarital childbearing have increased; and behind these changes lies the huge increase in cohabitation. In addition, all Western states have seen an increase in women's employment, albeit predominantly part-time in many countries. These changes in behavior appear to signal greater sexual and economic autonomy on the part of women and, in the view of some, more irresponsible behavior on the part of men. Together with a shift in attitudes towards individualism, which has been documented for a large number of Western countries (Inglehart 1997), the changes in behavior have been widely interpreted as signaling that men and women are less prepared to commit to, and to invest in, family life.

The optimism of academics in the 1960s and early 1970s about the fundamental stability of the traditional two-parent family has given way to profound pessimism at the end of the twentieth century. Bellah et al.'s (1985, p.85) influential study of middle America concluded that 'if love and marriage are seen primarily in terms of psychological gratification, they may fail to fulfil their older social function of providing people with stable, committed relationships that tie them into the larger society.' The causes of behavioral and attitudinal change are extremely complex, but the fear of many commentators is that both public policy and family law have encouraged rather than countered them.

In respect of public policy, Murray (1984) proved particularly influential in the English-speaking world in arguing that the provision of welfare benefits, far from sustaining the male breadwinner model family, allowed women a measure of financial autonomy and thus promoted family breakdown. In respect of family law, Glendon (1981) argued strongly that legal ties among family members in the postwar period were loosened. In particular, the move that took place in most Western countries towards the implementation of no-fault divorce in the 1960s and 1970s has been interpreted in terms of the withdrawal of government from the attempt to impose an external moral code in favor of allowing the parties to a marriage to work out their own destiny. A similar view has been taken by leading commentators in non-English speaking countries with very different legal traditions. Thus Thery (1994) has suggested that marriage has become a private choice and that marriage as an institution has become disengaged from wider social structures, the law included.

In whatever way the role of the state in regulating the family is interpreted, the solutions are far from easy to determine. Those who are suspicious of the state, whether because they feel that it has perpetuated traditional family forms and gender roles, or because they feel that these have been effectively undermined by the state, have difficulty in trusting government to do better, and in any case will demand radically different kinds of action. However, the picture of a homogenous state acting either to support or to undermine traditional family patterns is in all probability too simple. In the first place, the relationship between law and behavior is very difficult to determine. For example, Phillips' (1988) comparative study of divorce has shown that levels of marital breakdown are not dependent on legal change, and may be high in the absence of a liberal divorce law. Nor should people's resistance in the face of legal change be underestimated. Mount's account of the family resisting state interference is supported by findings from comparative studies that show legal change to have less effect on family behavior than on other areas of human activity. Second, the various departments of Western liberal governments rarely act coherently. Different departments may pursue very different agendas in respect of an issue such as childcare, for example.

Indeed, it may be more helpful to think of government regulation in the area of the family as layered. In all probability, Western governments in the twentieth century have tended to be Janus-faced in regard to family policy and law. While most have had a clear idea of the (traditional, two-parent) family form they wish to promote, policies have often pulled in a different direction. Thus, for example, while most income maintenance systems have a cohabitation rule, which prohibits the payment of welfare benefits to lone mothers who have a 'man in the house' (in accordance with the assumptions of the male breadwinner model family), they have also served to facilitate and legitimate, if not actually cause, the formation of lone mother families (Kiernan et al. 1998). The means of regulating the family have been many and various and certainly include the huge and

vexed area of birth control, abortion, and reproduction (see *Regulation: Sexual Behavior*). Historically, public policy and family law have sought to impose two powerful prescriptive frameworks on the family: that of the male breadwinner model, which has had a considerable impact on the gendered division of labor, and a strict moral code, which resulted in the stigmatizing of divorce and extramarital births. These frameworks were embodied in legislation but were also internalized by ordinary men and women. Their erosion in the late twentieth century has meant that increasingly people must negotiate their own solutions.

2. *Public Policy*

Given that in modern societies independence derives primarily from wage-earning, the assumption that women are located mainly in the private sphere as wives and mothers, supported by breadwinning husbands, has also meant that women have only been partially individualized for the purposes of public policy making (Pateman 1989). The liberal dilemma first described by Okin (1979), whereby for the purposes of determining political citizenship 'individual' effectively meant 'male head of household,' persisted. Thus most Western states have subscribed to some degree to the idea of a male breadwinner model family and have therefore treated women as dependants of men, although this set of assumptions has been modified to very different degrees in different countries (Lewis 1993). In its pure form, the male breadwinner model assumes that married women will not engage in paid work but will rather undertake the work of caring for children and other dependants at home without the support of collectively provided services, and that they will be subservient to their husbands for the purposes of social security entitlements and tax. In reality, such a model never existed; working class women have always gone out to work as and when the family economy demanded it. However, the model matched the social reality of middle class women fairly closely at the turn of the century. For the late nineteenth century social theorist Herbert Spencer, society was 'progressing' towards a position whereby all women would be able to stay at home in their 'natural sphere.'

It is no coincidence that the foundations of most Western European and Australasian welfare states were also laid down at the turn of the century, when the ideology of 'separate spheres' for men and women was at its height. Skocpol (1992) has argued that the core program of modern welfare states—social insurance against sickness and unemployment—is fundamentally 'paternalist.' The program sought to protect the full-time, respectable male worker. Women were rarely covered as workers in their own right and were usually eligible for benefits only as the wives of male employees, the justification being a division of labor that was perceived to follow 'naturally' from their capacity for motherhood. Women have thus tended to make (smaller) contributions and draw (smaller) benefits via their husbands. Furthermore, in welfare regimes such as that of the US and early and mid-twentieth century Britain, where the social security systems developed dual social insurance and social assistance benefits, this provision was gendered, with first class (insurance) benefits going mainly to men as workers, and second class (assistance) benefits, such as aid to families with dependent children in the US, to women, usually in their capacity as mothers.

Nevertheless the male breadwinner logic that informed most early twentieth century social protection was broadly accepted, not just by men and by policy makers who regarded the responsibility for dependants as a guarantor of male work incentives, but by women themselves. In the context of hard household labor and frequent pregnancy the male breadwinner model was a shared ideal. This changed in the postwar period as fertility rates fell and housework became less onerous. Assumptions regarding the existence of a male breadwinner model were eroded to different degrees in different countries, but the assumptions informing public policy usually lagged behind behavioral change. In France, because of the very different occupational structure (of family farms and firms), women tended to gain entitlements both as wives and mothers on the one hand, and as paid workers on the other.

Given the additional importance of pronatalism in France, the former tended to be more generous, but abundant public child care provision made possible a long tradition of full-time work for women. The Scandinavian countries moved furthest away from the male breadwinner logic in terms of public policy provision, deliberately pulling women into paid employment from the 1970s by the introduction of separate taxation and parental leaves and by increasing child care provision, to the point where the dual-breadwinner family became the norm. The contrast with Britain and Germany is striking. The latter has no independent taxation, which works against female employment when there are high marginal rates of tax, and a strong social insurance welfare system, in which women's eligibility remains tied to that of their husbands. Until very recently neither country made a significant commitment to collectively provided care services. In Germany female labor market participation rates have been low in the postwar period, while British women have tended to work short, part-time hours. In the US, where collective provision for able bodied adults did not develop beyond assistance benefits for lone mothers, state policy has played little part in determining how male and female members of the household decide the division of paid and unpaid work. However, the absence of explicit policies does not mean the absence

of a regulatory framework. The policy logic operates to constrain choices: adult women who are also mothers must depend on either men or on their own efforts in the labor market.

Policies towards lone mothers provide a particularly striking illustration of the extent to which public policy frameworks do have an impact on behavior. In the context of assumptions regarding the traditional two-parent family, the problem of lone motherhood becomes that of women with children and without men. In the absence of a male breadwinner, the state has to decide to what extent and under what conditions it will step in to fill the place of the missing husband; how far it will exert pressure on an able-bodied absent father to support his child and possibly its mother; and how far it will expect the lone mother to support the child by herself becoming a breadwinner. Policies in respect of each of these have varied enormously between countries and over time. But broadly speaking there has been a pendulum effect.

In the nineteenth century, lone mothers were treated as workers and were expected to support as many children as they could; by the mid-twentieth century they were more likely to be treated as mothers and permitted to draw benefits for as long as they had children under a particular age, and by the end of the century policy in most Western countries has veered again towards treating them as workers, although only in the US has the right to draw benefit been removed (see *Lone Mothers in Affluent Nations* and *Lone Mothers in Nations of the South*).

3. *Family Law*

The family has been seen both as bedrock and yet as fragile. Historically, the stability of marriage and the family has been seen as fundamental to the stability of the *polis*. Sociologists have suggested that marriage has moved from institution to relationship. In fact it remains both, but the regulation of the institution in accordance with an externally prescribed moral code has been substantially eroded. In the last quarter of the twentieth century Western states have largely ceased to regulate the role of men and women as husbands and wives. In respect of family law, the attention of the state has turned to regulating their roles as fathers and mothers.

Whereas the erosion of the male breadwinner model lagged behind social change, many would argue that changes in family law have run ahead of the social reality, certainly in terms of mentalities if not behavior. The reasons for abandoning the attempt to impose a strict moral code have been attributed to changing moral beliefs with an increased tolerance for pluralism, the rise of a psychological view of personal affairs, and the increased emphasis on individual rights. Nevertheless, most of those campaigning for reform of the divorce laws in the 1960s were convinced that they were proposing a higher form of morality that would be based on love and that would come from within, rather than being imposed from outside. The removal of fault from divorce law made it impossible to allocate blame and then, as a matter of course, to award long-term financial support to 'innocent' female spouses. Reform of family law meant that the state ceased to regulate the dissolution of marriage and concentrated on adjustment between the parties. Under a no-fault regime, a marriage without children is effectively terminable at will. The import of the process of liberalization has been to let men and women as husbands and wives order their own affairs, in other words to assume the existence of a degree of individualization, while also putting in place measures that seek to regulate them as fathers and mothers. The problem is that it has been difficult to make a neat separation between these two sets of roles, not least because of the need of children for caretakers and the tendency for women to perform more of this unpaid work.

The idea that divorcing couples should be treated in exactly the same way, thereby ignoring the different nature of their contributions to the marriage and the inequalities resulting from women's greater investment in childbearing and childrearing, followed the introduction of no-fault divorce. By the 1980s, in the English-speaking countries, the presumption of female dependence had been replaced in family law by the equally problematic assumption that husbands and wives were equal members of a partnership. This ignored the substantive inequalities which meant both that the view of the more powerful party as to the desirability of continuing the relationship might prevail, and that the outcomes for the economically weaker partner were likely to be poor. As academic research sounded the alarm about the latter (Weitzman 1985), and about the effects of divorce on children (Wallerstein and Kelly 1980), so the state made greater efforts to regulate the roles of men and women as fathers and mothers, focusing particularly on fathers.

In the face of rapid family change, the growing numbers of lone mother families and the resulting rise in the burden on public expenditure, governments sought to enforce the responsibility of men for their biological children. In the US, Australia, and the UK in the 1980s and 1990s new child support legislation was passed to this effect. Continental European countries had a much longer history of efforts to secure maintenance from absent fathers. A government-appointed commission in the UK examined these in the early 1970s, but rejected them as too intrusive, which indicates how far opinion moved in the last quarter of the century. If husbands and wives were to be left alone, parents were not. Fathers were also increasingly expected to have contact with their children, in part because the psychological evidence showed this was important for children, in part

because policy analysis showed that fathers who maintained contact were also more likely to pay for their children, and in part as a response to fathers' rights campaigners.

However, it has not been easy to deal with children separately from their carers. No-fault divorce removed the link between the grounds for divorce and maintenance without substituting any alternative principle to justify the latter. The relaxation of the divorce law took place in Western liberal democracies which felt that it was neither possible nor desirable to regulate the marriage system such that men's right to remarry and further reproduce was curtailed. This made it difficult to enforce the responsibilities of men in relation to their biological children, and by extension, those children's carers.

4. *Future Prospects*

The assumptions lying behind traditional marriage constructed husbands and wives as equal-but-different partners, something public policy accepted when it assumed the existence of the male breadwinner model family. In so doing, it constructed women as dependants and ran the risk of failing to recognize either the nature of their contribution as both paid and unpaid workers or the entitlements that were their due. Family law, on the other hand, moved towards a construction of marriage as a partnership of formal equals after the 1960s, which ran the risk of ignoring the substantive inequalities between men and women. Thus the problems of the relationship between the state and the family in regard to gender hinge on the larger problem of equality and difference.

Faced with family change that has been both dramatic and rapid in the last quarter of the twentieth century, Western governments have two basic options: to recognize the social reality and work with it or to seek to put the clock back. Just as it is possible to see the state acting both to bolster the traditional family and to transform it, so it is possible to identify ways in which the state is recognizing and resisting change. To move towards an emphasis on the responsibilities of men and women as parents is to recognize the changes that have taken place in the marriage system, but as Smart and Neale (1999) have argued, at the same time public policy and family law have attempted to perpetuate the traditional responsibilities of men for maintenance and women for care beyond the life of their intimate relationships.

See also: Family and Gender; Family, Anthropology of; Family as Institution; Family Law; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Fatherhood; Fertility Control: Overview; Gender and the Law; Reproductive Rights in Affluent Nations

Bibliography

Bellah R, Madsen R, Sullivan W, Swidler A, Tipton S M 1985 *Habits of the Heart: Individualism and Commitment in American Life*. University of California Press, Berkeley, CA

Donzelot J 1979 *The Policing of Families* (trans. by Robert Hurley). Pantheon, New York

Elshtain J B 1981 *Public Man Private Woman: Women in Social and Political Thought*. University of Princeton Press, Princeton, NJ

Glendon M A 1981 *The New Family and the New Property*. Butterworths, Toronto, ON

Gordon L (ed.) 1990 *Women, the State and Welfare*. University of Wisconsin Press, Madison, WI

Inglehart R 1997 *Modernization and Post-modernization: Cultural, Economic and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ

Kiernan K, Land H, Lewis J 1998 *Lone Motherhood in Twentieth Century Britain: From Footnote to Front Page*. Oxford University Press, Oxford, UK

Lewis J (ed.) 1993 *Women and Social Policies in Europe: Work, Family and the State*. Edward Elgar, Aldershot, UK

Mount F 1982 *The Subversive Family: An Alternative History of Love and Marriage*. Allen and Unwin, London

Murray C A 1984 *Losing Ground. American Social Policy 1950–1980*. Basic Books, New York

Okin S M 1979 *Women in Western Political Thought*. Princeton University Press, Princeton, NJ

Okin S M 1989 *Justice, Gender and the Family*. Basic Books, New York

Pateman C 1989 *The Disorder of Women: Democracy, Feminism, and Political Theory*. Stanford University Press, Stanford, CA

Phillips R 1988 *Putting Asunder. A History of Divorce in Western Society*. Cambridge University Press, Cambridge, UK

Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York

Skocpol T 1992 *Protecting Soldiers and Mothers. The Political Origins of Social Policy in the United States*. Belknap Press of the Harvard University Press, Cambridge, MA

Smart C, Neale B 1999 *Family Fragments*. Polity Press, Malden, MA

Wallerstein J S, Kelly J B 1980 *Surviving the Breakup. How Children and Parents Cope with Divorce*. Grant McIntyre, London

Weitzman L J 1985 *The Divorce Revolution: The Unexpected Social and Economic Consequences for Women and Children in America*. Free Press, New York

Wilson E 1977 *Women and the Welfare State*. Tavistock, London

J. Lewis

Regulation, Political Economy of

Government regulation of firms uses the 'coercive power' of the state to alter firms' pricing, entry, production, investment, or product choice decisions. An extensive empirical literature analyzes the effects of

'economic regulation' of price and entry as well as environmental, health, safety, and information regulation. These analyses amassed substantial evidence of high prices, excess profits, and production inefficiencies under economic regulation (see Joskow and Rose 1989, Joskow and Noll 1994; and *Regulation: Empirical Analysis*). Studies of environmental, health, and safety regulation revealed substantial deviations from least-cost solutions and dramatic variation in measures such as expected cost per life saved across regulatory interventions (see Viscusi 1994, Gruenspecht and Lave 1989). These outcomes, so systematic and pervasive, were difficult to reconcile with 'public interest' models of regulatory intervention. Examination of regulatory systems further suggested that they often deviated from cost minimization or efficiency objectives by design. This begged for an explanation beyond bureaucratic incompetence or malfeasance, common to earlier Marxian or 'capture' theories of government. Stigler's (1971) *Economic Theory of Regulation* provided the conceptual underpinnings of what has become a vast enterprise among economists and political scientists to understand the origin, objectives, maintenance, and reversal of regulatory policy in the context of rational choice models of political behavior. These simple insights stimulated a substantial theoretical and empirical literature (Noll 1989, Peltzman 1993).

Theories of legislative and bureaucratic behavior have become increasingly sophisticated since the 1980s. The literature has moved well beyond Stigler's simple model to incorporate the role of competing interests, policy entrepreneurs, majoritarian political institutions, legislative oversight of regulatory agencies and commissions, and asymmetric information (see Noll 1989, Joskow and Noll 1994, Laffont 1999, and *Regulatory Agencies*). The empirical literature has struggled to keep pace with these developments.

Common to virtually all empirical studies in this area is the importance, and difficulty, of defining well-specified alternatives to the political economy model under investigation. To test a given model of political behavior, or to interpret results as consistent with a given model, one must understand how predicted outcomes differ between competing hypotheses of behavior and the model being tested. For example, there is a natural tendency to identify rent conveyance with rent-seeking activity. In some cases, however, increased rents for regulated firms may be consistent with both successful rent-seeking behavior by those firms and elimination of some market inefficiency. It is important, though often quite difficult, for empirical analyses to identify and attempt to distinguish alternatives such as these. Studies vary considerably in the degree to which they meet this standard.

This article introduces the main empirical methodologies used to explore the political economy of economic, environmental, health and safety regulation. Examples of their applications are briefly discussed. Interested readers are urged to pursue the surveys listed in the references for more complete discussions of this literature.

1. *Measuring Distributional Effects*

Empirical studies of political economy generally begin with estimates of the distributional effects of a given policy. While many empirical methods may be used (see Joskow and Rose 1989 and *Regulation: Empirical Analysis*), asset market 'event studies' have been particularly popular. These measure price movements in capital assets, such as company share prices, at the time policy changes are first announced. If the announcement is not previously anticipated, and asset markets efficiently process information, these price changes provide estimates of the policy's effect on the capitalized value of expected rent streams. This method may be used to measure the effects of the policy change on firms within a regulated industry, corporate customers and suppliers, and possible competitors. The distributional implications for individual or small business customers, employee groups, and other noncorporate interests usually must be measured using alternative methods.

The simplest political economy studies use measured distributional effects to infer the relative influence of different interest groups in the political process. These implicitly assume that the observed changes in rent distribution are anticipated and intentional, and that larger rent gains reflect greater political power. Consider, for example, the extensive debate over the origin of the first federal economic regulatory commission in the US, the Interstate Commerce Commission created in 1887 to regulate railroads. Was this legislation intended to solve an inefficiency in an industry characterized by large sunk costs and natural monopoly conditions? If the legislation was a response to interest group politics, did it respond to agrarian interests and other shippers' demands to constrain monopoly pricing by railroads, or was it a railroad-sponsored effort to establish a government-enforced cartel? Did legislation intend one outcome, only to be thwarted by the regulatory agency's subsequent behavior? There have been numerous empirical studies attempting to answer these questions, many using railroad share price responses to various regulatory 'events' (e.g., Prager 1989, Mullin 2000). In these, railroad share price gains in response to the Interstate Commerce Act, for example, have been interpreted as evidence that the regulation was intended to benefit railroads, supporting Stiglerian models of regulatory politics. While it may be reasonable to assume that *ex post* gains identify *ex ante* intended beneficiaries, this should be made explicit and considered against alternative explanations. To move convincingly from a positive

statement about the distributional consequences of a policy to an inference about its political motivation requires an understanding of distributional effects under leading alternative models.

2. Estimating Political Support for Regulation

A second empirical methodology estimates the determinants of regulatory incidence or political support as a function of underlying economic interests. These studies generally involve three steps. First, *ex ante* winners and losers of a regulatory policy are identified, using either prior estimates or a model of the regulation's effect. For example, Stigler's (1971) analysis of truck weight laws modeled state truck weight and size limits as raising the cost of trucking service, particularly over longer hauls. Truckers therefore lose from lower limits, as do their current customers, businesses using their own trucks, and potential customers for whom trucking rates constrain alternative transportation rates. Competitors to trucks, such as railroads, gain from tighter restrictions. There may be diffuse 'public interests' as well, for example, in limiting damage to roadbeds and reducing congestion, though these may well be endogenous.

Second, the importance of the identified economic interests within a given jurisdiction or at a particular point in time is measured. There is considerable variation in how this is done. For example, the number of trucks, scaled or unscaled, the dollar volume of activity in trucking, and the fraction of freight hauls susceptible to potential truck competition are all possible measures of the importance of trucking interests in a jurisdiction.

Finally, the incidence or intensity of regulation, or the political support for regulation, is estimated to be a function of the measured economic interests and any other relevant factors. Stigler used this method to explain differences in the intensity of truck weight limitations across states, for example. Other studies apply this technique to model variation in political support, such as voting records, as a function of constituent economic interests in the policy, *inter alia*.

Most studies in this area find that constituent interests are significant determinants of regulatory policy, though interests must be defined broadly to include constituent preferences (such as conservation or environmental preferences) as well as direct economic interests. Detailed studies of voting behavior suggest that the underlying models may be quite complex, however; see, for example, Joskow and Schmalensee (1998) on the US acid rain program. The significant impact of legislators' ideology on their voting behavior (e.g., Kalt and Zupan 1984) suggests possible slack in the agency relationship between constituents and their legislators, as well support for more sophisticated models of voting in majoritarian systems across issues and over time.

3. Modeling Regulatory Agency Behavior

A final strand of empirical work in political economy begins with the recognition that most regulatory legislation leaves considerable discretion over its implementation to regulatory agencies or commissions. Early work on regulatory 'capture' assumed that industry groups could influence regulators to use this discretion to benefit the industry in direct contravention to legislative intent. As economists and political scientists began to explore the extent to which apparent regulatory capture was instead a reflection of legislative intent, they confronted the questions of how much discretion regulators retain and how regulators use that discretion to form policy (e.g., McCubbins et al. 1987). This led to an extensive literature on the operation and oversight of the administrative process, most of it within political science. This area remains one of the least-developed empirical literatures in regulatory economics.

While some early analyses explored the responsiveness of regulatory agency decision-making to congressional preferences, most economics research in this area analyzes the responsiveness of agency decision-making to interest groups and various measures of efficiency or social welfare. This is methodologically quite similar to the legislative studies described in Sect. 2, with individual regulatory decisions as the unit of observation. As in those studies, the results suggest a role for concentrated economic interests to influence regulatory outcomes. This does not itself demonstrate that regulators respond to interest groups in contravention with broader social objectives, however. For example, studies have found that despite significant overall delays in the process by which the Food and Drug Administration (FDA) controls entry of new pharmaceutical products, the FDA responds more quickly to drugs with greater market potential (Dranove and Meltzer 1994). In this case, both producer and social interests are aligned with reducing delay of an innovation.

4. Conclusion

While research on the political economy of regulation has made significant progress since the 1970s, fundamental questions remain unanswered by either theoretical or empirical work (Peltzman 1993, Joskow and Noll 1994). One of the most basic is why some industries are regulated (or deregulated) and others are not. Another is why regulatory policy seems characterized by 'waves' or cycles: changes tend to occur across broad ranges of industries and sectors within fairly short periods, separated by longer periods of relative inactivity. Considerable scope remains for additional work in this area.

See also: Antitrust Policy; Broadcasting: Regulation; Environmental Economics; Industrial Relations and

Collective Bargaining; International Communication: Regulation; Occupational Health and Safety, Regulation of; Policy History: State and Economy; Policy Process: Business Participation; Privatization of the Economy; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulatory Agencies; Safety, Economics of

Bibliography

Dranove D, Meltzer D 1994 Do important drugs reach the market sooner? *RAND Journal of Economics* **25**: 402–23

Gruenspecht H K, Lave L B 1989 The economics of health, safety, and environmental regulation. In: Schmalensee R, Willig R (eds.) *Handbook of Industrial Organization*. North-Holland, Amsterdam, Vol. 2

Joskow P L, Noll R G 1994 Economic regulation during the 1980s. In: Feldstein M (ed.) *Economic Policy During the 1980s*. University of Chicago Press, Chicago

Joskow P L, Rose N L 1989 The effects of economic regulation. In: Schmalensee R, Willig R (eds.) *Handbook of Industrial Organization*. North-Holland, Amsterdam, Vol. 2

Joskow P L, Schmalensee R L 1998 The political economy of market-based environmental policy: The US acid rain program. *Journal of Law and Economics* **41**: 37–83

Kalt J P, Zupan M A 1984 Capture and ideology in the economic theory of politics. *American Economic Review* **74**: 279–300

Laffont J-J 1999 Political economy, information and incentives. *European Economic Review* **43**: 649–69

McCubbins M, Noll R, Weingast B 1987 Administrative procedures as instruments of political control. *Journal of Law, Economics and Organizations* **3**: 243–77

Mullin W 2000 Railroad revisionists revisited: Stock market evidence from the progressive era. *Journal of Regulatory Economics* **17**: 25–47

Noll R G 1989 Economic perspectives on the politics of regulation. In: Schmalensee R, Willig R (eds.) *Handbook of Industrial Organization*. North-Holland, Amsterdam, Vol. 2

Peltzman S 1993 George Stigler's contribution to the economic analysis of regulation. *Journal of Political Economy* **101**: 818–32

Prager, R 1989 Using stock price data to measure the effects of regulation: The Interstate Commerce Act and the Railroad Industry. *RAND Journal of Economics* **20**: 280–90

Stigler G J 1971 The theory of economic regulation. *Bell Journal of Economics and Management* **2**: 3–21

Viscusi K 1994 Health and safety regulation. In: Feldstein M (ed.) *Economic Policy During the 1980s*. University of Chicago Press, Chicago

N. L. Rose

Regulation: Sexual Behavior

The regulation of sexual behavior refers to the enactment of laws and legal codes at the federal, state, or local level whose intent is to define and restrict specific sexual conduct. Sexual conduct encompasses a range of behaviors and activities, only some of which are regulated by law depending on the age, marital status, and legal residence of the individual. Sexual behaviors include those practiced between individuals—as in any intimate bodily contact such as kissing, touching, intercourse, oral and/or anal sex—and those practices, usually solitary, such as masturbation and nude public exposure. While all societies have enacted some form of a regulatory mechanism to prohibit sexual behaviors considered socially offensive or physically and/or psychologically harmful to its members, the regulation of sexual behaviors typically reflects 'a society's particular vision of sex, marriage, and the family' (Francoeur 1991a, p. 583). In the case of the USA, sexual regulations have historically served to reinforce traditional moral values in the form of social and religious taboos proscribing non-procreative sexual behaviors, including those practiced within and outside of marriage. Changes in social relations and in the expanding configurations of coupling, marriage, and family structure since around 1950, have resulted in more tolerant and permissive attitudes about sexuality and more diverse sexual lifestyles (D'Emilio and Freedman 1988). Many of the existing legal regulations are either not enforced, in the process of reform or repeal, or are confronting considerable resistance. Moreover, cultural forces, such as changing gender roles and norms of sexual development, influences of mass media, popular culture, information technology, and Internet access, as well as alterations in the workplace have provided new issues with regard to the regulation of sexual behavior, which continue to be intensely contested in the political and legal arenas in the USA. Since around 1980, such issues as sexual harassment, the sexual abuse of women and girls, the role of sexually explicit materials in public and private life, government policy regarding sexual orientation and military personnel, and the prevention and treatment of HIV/AIDS have placed this issue at the center of social controversy (Gagnon 1991).

1. *The Legal Sphere: Protection vs. Restriction*

The multifaceted role of our legal system is particularly evident in this complex process of enacting, interpreting, and/or disputing the laws regulating sexual behavior. In this process, the legal system is intermittently supported or challenged by a myriad of cultural, social and political forces that view this system either as an appropriate and necessary force enacting restrictive laws in the interest of individual safety, family cohesion, and societal order, a system that arbitrarily applies laws to certain individuals or groups of individuals, or conversely, as an ameliorating one that can redress the loss of individual right to privacy and consensual sexual expression.

In the USA, sexual behaviors are primarily regulated by way of a legal process that restricts what it

defines as either harmful or socially unacceptable sexual conduct, both in private and public. More specifically, it is the general consensus that certain sexual behaviors merit prohibiting. In these instances, the laws are typically viewed as necessary and justifiable because they either provide parameters for sexual relations, as with regulations dictating the age of majority, or they prohibit coerced sexual contact between individuals, unsolicited sexual behavior on the part of the individual, and sexual exploitation, such as laws prohibiting rape, coerced nudity, sexual exploitation of an incapacitated individual, sexual contact between family members (incest), sexual contact between adults and children (pedophilia), sexual exposure to another individual (exhibitionist/indecent exposure), or the unsolicited sexual attention and/or pursuit by another individual (sexual harassment) (Leonard 1993). Other regulations are controversial because they restrict consensual sexual behavior and/or private sexual activity that is subjectively defined as socially offensive, inappropriate or immoral, as in sexual relations between: individuals married to other partners (laws prohibiting adultery); sexual relations between homosexual partners (laws targeting homosexual sodomy); oral or anal sex between heterosexual individuals, married or unmarried (those targeting heterosexual sodomy); and sexual relations exchanged for payment (prostitution). Resistance to these regulations and increasing controversy regarding their enactment has emerged from a growing awareness of one's right to privacy and the constitutional requirement that any restriction must be justified by compelling social need and benefit.

A related set of controversial regulations pertaining to sexuality but not directly involving sexual behaviors are those whose objective is to limit the production of and access to, sexual 'entertainment' and sexually explicit material. Operating in both the private and public sphere, these regulations focus on the right of the individual to access or 'consume' sexual entertainment and/or sexually explicit material in a variety of venues and mass cultural mediums—including various entertainment spheres, TV, movies, video, and via the Internet. Also restricting private, consensual behavior, these regulations range from censoring sexually explicit material in print to banishing live entertainment shows to peripheral locations in urban centers to confining sexually explicit content on cable TV to specific channels and hours (Communications Decency Act of 1995, Telecommunications Act of 1996).

2. *Sexual Regulations that 'Protect'*

Even regulations whose objective is to protect the individual from harm have a history of vague legal definitions and arbitrary enforcement. A case in point is the discrepancies in the penalties attached to violations of sexual regulations and other criminal laws such as rape vs. murder, the disproportionate penalties given to minority vs. non-minority individuals charged with rape, and the increasingly subjective determination of incidents of sexual harassment.

More often, regulations pertaining to sexual behavior are applied to some but not others, and in many instances the determination is made on the basis of sexual orientation, sexual preferences or to behaviors considered outside of normative heterosexual activity. Individuals who fall into any one of these categories are then part of a 'sexual minority' with the unequal application of laws to these persons often questioned as to its constitutional legality.

3. *Regulations that Restrict: A Normative View of Sexuality*

In the legal enactment and interpretation of these regulations, rarely does the body of scientific and accurate knowledge about sexuality inform this process, a body of knowledge best represented by the repository of scholarly contributions on the subject from the social, biological and physical sciences, and from the humanities. As a result, knowledge of sexuality has had limited influence on developing a meaningful public and legislative discussion and cogent public policy regarding the regulation of sexual behaviors. Instead, the regulations of behaviors, especially those conducted in private between consenting adults, are typically based on a normative, procreative view of sexuality originating in English common law. This view approaches human sexuality as a potentially uncontrollable, inherently deviant and socially disruptive force in need of restraint, one rightfully subject to stigma, penalty, and criminalization, especially outside of marriage. In the USA it has resulted in very clear demarcations of appropriate and inappropriate sexual behavior in which fornication between married, heterosexual partners is considered a sexual behavior worthy of protection under the law. However, at the same time other sexual behaviors, such as those between homosexual partners or oral/anal sex between married or unmarried partners, are determined to be immoral and illegal. Some laws do not specify 'non-reproductive sex, but use laws against disorderly conduct, carnal abuse, lewdness, lascivious behavior, public indecency, loitering, and indecent exposure to prosecute crimes of oral and anal sex' (Franceour 1991a, p. 588).

Regulations based on a normative, procreative view of sexuality typically target adult behavior inherently private and consensual. As they become increasingly at odds with the sexual practices of the population and are enacted against an increasingly vocal sexual minority (more typically evident in the western world), their application and enactment becomes not

only controversial, but often arbitrary. Two examples stand out as particularly incongruent with actual sexual practices—the state sodomy laws and the military law regarding homosexual behavior among its members—and their application has become increasingly both controversial and unworkable. Forces supporting their enactment have confronted and continue to confront those seeking their revocation; a conflict played out in the courts, in the media, and in strategy meetings by opposing teams of legal experts and community activists. In both examples, the point of contention is whether their continued enforcement either ensures the 'common good' and protects someone from harm, or whether it promotes intolerance and violence against a sexual minority who are the target of their enactment. To illustrate this point, a brief history of these two regulations follows.

4. *State Sodomy Laws*

State sodomy laws criminalize certain sexual behaviors within heterosexual marriage and all sexual behaviors occurring between individuals of the same gender. More specifically, sodomy is often defined as deviate sexual intercourse, including genital–anal, genital–mouth, manual–anal, or manual–genital contact. Sodomy laws are any one of the many laws, which criminalize non-reproductive, non-commercial sexual behavior between adults in private. Since their enactment, restrictions against sodomy between unmarried individuals have typically served the purpose of dictating 'normative heterosexuality.' This is clear when reviewing their application, as sodomy laws are rarely enforced in the case of married, heterosexual couples engaging in oral or anal sex. Not only are they primarily enacted in those cases when sexual acts occur between members of the same gender, but they have been used to justify legal discrimination against homosexual people (Cain 1993, Copelon 1990) and to deny lesbians and gay men a range of other rights. This occurred when the laws were introduced as relevant evidence in adoption and custody cases and they have justified the taking of children away from gay and lesbian parents.

5. *The History of Sodomy Law Repeal*

The original 13 colonies continued the English tradition in which sodomy carried the penalty of death and every state adopted some form of a sodomy law as it joined the USA. By 1960, all states had some form of sodomy law in place, although their terms were often vague and they rarely stated clearly what the law prohibited. Throughout the long period of their enactment state by state, sodomy laws covered a variety of sexual behaviors, including oral sexual relations between two men, two women or between a man and a woman; anal intercourse was included either under sodomy or buggery laws, without distinction as to whether the two persons are heterosexual or gay, married or unmarried (Franceour 1991a, p. 587).

It was not until 1961 that the first state repealed its law when the Illinois legislature revised their criminal code without prohibiting sodomy. Other states followed soon afterward, although this movement significantly tapered off in the 1970s and 1980s during which time a spate of 'homosexual-only' sodomy laws were introduced. These laws were quickly, and in a number of cases successfully, challenged for their constitutionality under equal protection and privacy rights at the state level. The brief movement toward repeal came to a standstill in the 1980s as it coincided with the early stages of the HIV/AIDS crisis and courts and lawmakers withdrew from [such] efforts in light of the potential criticism of promoting homosexual sex at a time that the HIV transmission modes were becoming known.

In 1986, the US Supreme Court upheld a Georgia anti-sodomy law in the case of Bowers vs. Hardwick, ruling that the constitution 'does not protect homosexual relations between adults even in the privacy of their own homes' claiming that (a) homosexual couples did not have a constitutional right to engage in acts of consensual sexual acts, and (b) oral and anal sex is a criminal offense 'regardless of whether the parties who engage in it are married or unmarried, or of the same or different sexes' (Franceour 1991a, p. 587). It was not until the 1990s that the HIV/AIDS transmission was more clearly understood and the changing political climate allowed for repeal legislation to move forward.

Repeal legislation has continued to slowly dismantle state sodomy laws one by one, and in the process often unveils the weak foundation upon which the pro-sodomy camp stands. A telling example is the recent case argued by the state of Montana in defending its 'Deviate Sexual Conduct Law' to the Montana Supreme Court. Here, the state argued that there is no right to privacy when it comes to sexual conduct; some of the justices, not persuaded by this argument asked, 'what could citizens in the state expect to be covered under the privacy clause if it was not sexual activity?' (ACLU Newswire, 1997).

At the beginning of the twenty-first century, 16 states plus Puerto Rico have sodomy laws; 11 states plus Puerto Rico have sodomy laws which apply to both heterosexuals and homosexuals, and five states have laws, which apply only to homosexuals. The growing movement for repeal has been successful in 34 states: 27 states have repealed the law through legislative action, and seven states have had their laws invalidated by courts. Sodomy laws are primarily being challenged in state courts on the basis that state constitutions can provide greater rights than does the

federal constitution. One brief in support of repeal states said: 'When considering a public morality justification for (sodomy laws), it is important for the Court to recall that prohibitions on interracial marriage, and the segregation of races generally, were values deeply imbedded in social morality and were values that invoked both God and nature.' But appeals to natural or theological ethics cannot constitutionally be used to legitimate laws whose sole function is to give effect to private citizens' prejudice or conviction (ACLU Newswire 2000).

6. *Military Law Regulating Sexual Behavior*

The most significant military law enacted to regulate the sexual behavior of military personnel in active service is the Pentagon policy of 'don't ask, don't tell' regarding the sexual behavior of homosexual service personnel. Enacted on July 19, 1993, this rule specifically forbade US military commanders from asking, pursuing, or harassing homosexual service members regarding their orientation. More significantly, it codified the same restrictions for homosexual service members as had existed before 1994 under Pentagon policy, requiring secrecy and celibacy as a condition of their military service, under penalty of being fired or imprisoned. It also openly barred homosexual individuals from serving in the military. Five years later, the 1998 discharge figures have shown a record high of 1145 individuals discharged under the current policy, a 92 percent increase in homosexual discharges since the law was implemented (Service members Legal Defense Network).

When challenged on the figures, Pentagon spokespersons indicated that the high number of discharges was due to an increase in 'voluntary' statements made by service members; what they failed to state was that the overwhelming majority of such statements were made to psychotherapists, in personal diaries, contained in anonymous online profiles, or, in direct violation of the law, constituted responses to direct questions about one's orientation. During that time, service members have indicated that they were coerced to make such statements due to hostile command climates characterized by fear, intimidation, and threat of assault, death and/or criminal prosecution. This was evidenced by the documented cases of physical assault and the brutal murder of Private Barry Winchell at Fort in Campbell, KY, which marked a turning point in the debate over the Pentagon's policy.

Coming under considerable pressure, the Pentagon initiated a series of measures in 1998 designed to 'improve' policy implementation by incorporating such measures as: prohibiting 'plea bargains' to obtain evidence of consensual homosexual conduct; protecting service members from investigations in order to encourage the reporting of anti-gay harassment or lesbian-baiting without fear of reprisal; limiting the scope of investigations to prevent fishing expeditions; requiring commanders to consult with higher headquarters before initiating investigations into alleged homosexuality, and providing necessary and effective training on investigation parameters under the policy. Moreover, in December 1998, the Pentagon implemented a first-of-its-kind survey to determine the views of its service members about anti-gay discrimination evident in the military. It additionally required universal training on the 'don't ask, don't tell' policy of all military personnel with the intent to promote full compliance with military policy that anti-gay threats and harassment will not be tolerated and to educate all service members on policy's parameters.

In issuing the results of its survey, the Pentagon indicated that disparaging, anti-gay remarks, hostile attitudes and harassment are commonplace in the American military, with a majority of the service members stating that they believed that superiors and colleagues tolerated it. Such behavior has been confirmed by independent investigatory reports indicating those incidents of verbal abuse and physical assault in 1999 more than doubled from the year before. Placing the discharge figures in realistic terms, the report specified that on average, the policy has resulted in three persons per day being discharged after declaring they are gay, with a 142 percent increase in harassment reports for 2000 (SMDN, March 10, 2000).

The regulation of sexual behaviors is ultimately an issue of privacy, consent, and choice. In the US 'there is no shortage of opinions on what constitutes acceptable or unacceptable sexual behavior or what justifies regulating or prohibiting a particular behavior or activity' (Francoeur 1991a, p. 583). Yet the legal process of regulation enactment or revocation inevitably provides crucial opportunities for engaging public discourse, refining strategies to promote social change, and questioning ideological assumptions about sexuality and morality. A continuing challenge is to effectuate the appropriate balance between the common good and individual rights, privacy, and the social/personal cost exacted for laws that unnecessarily restrict sexual behavior.

See also: Child Abuse; Gay/Lesbian Movements; Heterosexism and Homophobia; Homosexuality and Psychiatry; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Prostitution; Queer Theory; Rape and Sexual Coercion; Rationality and Feminist Thought; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Orientation and the Law; Sexual Perversions (Paraphilias); Sexuality and Gender; Slavery as Social Institution; Transsexuality, Transvestism, and Transgender

Bibliography

ACLU Newswire 1997 *Montana's Deviate Sex Law Challenged*. May 20, 1997

ACLU Newswire 2000 *Louisiana May be Latest in Overwhelming Trend of State Courts Finding Sodomy Laws Unconstitutional*. April 10, 2000

ACLU on Censorship: http://www.aclu.org/library/pbp14.html

Bancroft J, Rosen R, Davis C, Weinstein D, Ruppel H 1991–95 The Society for the Scientific Study of Sexuality. *Annual Review of Sex Research* **1–6**

Cain P A 1993 Litigating for lesbians and gay rights: a legal history. *Virginia Law Review* **79**: 1551–641

Copelon R 1990 A crime not fit to be named: Sex, lies and the constitution. In: Kairys D (ed.) *The Politics of Law: A Progressive Critique*, rev. edn. Pantheon Books, NY, pp. 177–94

D'Emilio J, Freedman E 1988 *Intimate Matters: A History of Sexuality in America*. Harper and Row, New York

Federal Communications Commission, The Telecommunications Act of 1996: www.fcc.gov/telecom.html

Francoeur R 1991a Sex and the law. In: Miller R (ed.) *Becoming a Sexual Person*. Macmillan Publishing, New York

Francoeur R (ed.) 1991b *A Descriptive Dictionary and Atlas of Sexology*. Greenwood Press, New York

Friedman S 1990 *Sex Law: A Legal Sourcebook on Critical Sexual Issues for the Non-lawyer*. McFarland and Co., Jefferson, NC

Gagnon J H 1991 Sexual conduct and the life course. In: Turner C F, Gagnon J, Miller H, Moses L (eds.) *AIDS, Sexual Behavior and IV Drug Use*. National Academy Press, Washington, DC, pp. 500–53

Lauman E, Gagnon J, Michael R, Michaels S 1994 *The Social Organization of Sexuality: Sexual Practices in the USA*. University of Chicago Press, Chicago

Leonard A S (ed.) 1993 *Sexuality and the Law: An Encyclopedia of Major Law Cases*. Garland Publishing, New York

Lichtenberg K E 1994 *A Research Guide to Human Sexuality*. Garland Publishing, Inc., New York

Miller vs. California 1973 413 US 15

NACLA 1998 Report on the Americas. *Sexual Politics in Latin America*, Vol. **XXXI**, January/February

National Coalition Against Censorship: http://www.ncac.org/cen_news/cn71sex_censorship.html

Parker R, Gagnon J (eds.) 1995 *Conceiving Sexuality*. Routledge, New York

Portelli C 1998 *SIECUS Report*, Vol. 26, No. 5, p. 2

Reno vs. ACLU 117 S. Ct. 2329, 138L. ed. 2nd 874 (1997)

Roth vs. US 354 US 476 (1957)

Sobell S, Wescott K, Benecke M H, Dixon Osburn C 2000 Don't ask, don't tell, don't pursue, don't harass. SLDN 6th Annual Report. March 9 2000. www.sldn.org/Templates/Law/Record.html? Record = 22

Sodomy Laws State by State: http://www.sodomylaws.org/; http://www.aclu.org/issues/gay/sodomy.html

Service Members Legal Defense Network (SLDN) is comprised of a nationwide network of attorneys offers legal advice and assistance to serve US service members at risk of discrimination, persecution and dismissal on the basis of sexual orientation. This information has been culled from its 6th Annual Report.

Sex and Censorship: Dangers to Minors and Others? Background paper of National Coalition Against Censorship

Bob Summersgill, History of Sodomy Laws, in Creating Change: Public Policy, Civil Rights and Sexuality (to be provided.)

1998 *State of Sodomy Laws*. American Civil Liberties Union, New York

2001 (Spring) *Advocates Report: Vol 6, No 2* Sexuality Information and Education Council of the US, New York

D. Di Mauro

Regulation Theory in Geography

1. Introduction and General Argument

The regulation theory aims to understand the economic and extra-economic mechanisms which, relatively, stabilize capitalistic societies despite the fundamental contradictions generating crisis and disintegration. The regulation approach (RA) originated in French academic debates about the economic and social changes of the 1970s. Thirty years after World War II, a first serious structural crisis affected western societies indicating the end of the dream of a never-ending prosperity ('the golden age'). New forms of conflict configuration and new balances of power arose—without undermining the fundamental principles of capitalist societies altogether.

To understand the restructuring of postwar capitalism, neither the equilibrium models of orthodox economies nor the traditional concepts of Marxism seemed to be satisfactory. In this context an integrative approach has been developed—combining the analysis of political economy with that of civil society and the state, sometimes looking like a kind of institutional economies. But the term 'regulation' within RA may not be understood as an authoritative direction in order to govern the economy. Instead a process of self-regulating is addressed to conceptualize the evolution and temporary stability of a crisis-driven capitalist society. Four core concepts help one to understand processes of regulation (Boyer 1990, Boyer and Saillard 1995, Jessop 1997):

(a) an 'industrial paradigm' as a model governing the technical and social divisions of labor;

(b) an 'accumulation regime' as a complementary pattern of production and consumption which is reproducible over a long period;

(c) a 'mode of regulation' as an emergent ensemble of rules, norms, conventions, patterns of conduct, social networks, organizational forms and institutions which can stabilize an accumulation regime; and

(d) a 'model of development' as the result when an industrial paradigm, an accumulation regime and a mode of regulation complement each other sufficiently.

These four aspects of capitalist societies are differently configured in successive stages. As the most

discussed and inquired, the term 'Fordism' refers to a new production concept, introduced by Henry Ford between 1908 and 1925. The Fordist accumulation regime is characterized by a specific combination of mass production and mass consumption. Purchasing power has to be sustained by appropriate wages to ensure demand. This process depends on continuous productivity growth. The corresponding type of state is an intervening welfare state ('Keynesian welfare state') that aims to adjust demand to the supply-driven needs of mass production.

Fordism entered into structural crisis in the 1970s. The regulationists discussed several arguments to understand this phenomenon: differentiation of consumer behavior; rising wages, caused by the bargaining power of strong trade unions, but without a corresponding increase of productivity; the internationalization of economy, especially an increasing low cost production in developing countries; and a lack of a corresponding internationalization of politics and policy.

The stage after Fordism is characterized by productivity gains, ideologies and constraints of 'flexibilization.' Fordist mass production seems to disappear whereas flexible networks of decentralized production units, partly hierarchically controlled, are significant for the new stage using new technologies and new forms of production organization which enable 'flexible specialization.' At the same time there are strong attempts to deregulate the labor market while collective wage bargaining, rigid work relations, and intervening legal forms as institutionalized compromise between capital and labor have been serious constraints for the flexibilization of economy and society. Further, objectives of state acting have changed in order to strengthen the international competitiveness of the national economy. At the same time redistributive welfare rights have been subordinated to a productionist reordering of social policy.

This rather rough sketch of 'post-Fordism' is differentiated in the regulationist debate. Thus, it is not possible to identify a definite post-Fordist model of development. Some authors have distinguished two main directions of post-Fordist development (Leborgne and Lipietz 1994): 'neo-Taylorism' (deregulation and flexibilization without major constraints; for example, in the UK and USA) and 'Toyotism' (decentralization and flexibilization while keeping a high degree of social cohesion; for example, in Japan and Northern Europe).

Spatial patterns are varying between the different stages. The Fordist stage is characterized by strong spatial divisions of labor, homogenization, and core–periphery differentials on all spatial scales (for example, city–suburb, north–south divide). Post-Fordist spatial development is characterized by heterogeneity on all scales: different types of spatial structures (industrial districts, global cities, high-tech regions, regions with functional mixtures, etc.) built up 'patchworks.' Definite trends of spatial structures are hard to recognize.

2. *Origins and Development of the Approach*

Today there are several RA directions. In the beginning at least two schools of regulationist thinking evolved in France; the 'Groupe de Recherche sur la Régulation de l'Economie Capitaliste' (GRREC) based in Grenoble, and the 'Centre d'étude Prospective d'Economie Mathématique Appliquées à la Planificacion' (CEPREMAP) in Paris. An important input to both approaches has been the work of François Perroux. Whereas the work of the GRREC is characterized by an elaboration within a more traditional Marxism, the Parisian RA, especially Michel Aglietta, Robert Boyer, and Alain Lipietz, have been basically involved in a renewal of Marxist thought and Althusserian structuralism with respect to evolutionary and institutional economics. Their contributions mark the beginning of the globalization of the RA (see reviews by Jessop 1997, MacLeod 1997). Since the mid-1970s and especially in the 1980s, their ideas spread, leading to new schools of innovation within the RA. A closer look at these discussion groups reveals a broad heterogeneity, bound together by common understandings of basic elements of radical political economy and the ability of social groups to act in order to transform society and economy.

Up to now, three periods of the RA with changing topics are observable: the first one may be labeled as the formative period that is characterized by theoretical and empirical work by the French regulationists. Among others, Aglietta's outline of the theory of regulation based on an analysis of the evolution and crisis of US capitalism in the twentieth century is usually seen as a well recognized starting point (published in French in 1976; in English in 1979, reprinted 2000). In general the formative period delivered the basic concepts of the RA and a deep understanding of Fordism including its crisis in the 1960s and 1970s.

A second period may be characterized by an intensive search and controversial debate around the term 'post Fordism.' Though often rejected as a misleading concept, post-Fordism served as a focus for those who looked after new economic and social principles of capitalist development in the late 1980s and 1990s. Examples are:

(a) the search for the microfoundation of regulation using convention theory or Bourdieu's concept of habitus (e.g. Painter 1997, Jenson 1993);

(b) debates about new organizational patterns of enterprises and the internationalization of production (e.g. Waringo 1998);

(c) attempts to overcome structuralist shortcomings and a reflection of the structure–agency debate;

(d) comparative perspectives on modes of regulation in order to understand crisis and changes of the welfare state more profoundly (e.g., the concept of the workfare state by Jessop (1995) or the 'Wettbewerbsstaat' (competition state) by Hirsch 1995); and

(e) the question of scale related to the diminishing power of the nation state and the growing importance of substates (regional, urban) and supranational levels (international regimes) (e.g., Boyer and Hollingsworth 1997).

As outlined above, these attempts to understand the new (flexible) regime of accumulation and new (deconcentrated) mode of regulation indicate a thematic and regional expansion of the original French RA—a globalization of a theoretical program. However, the critical notes about the term post-Fordism—whether it is the guiding concept of a new development model as temporal spatial fix of capitalism evolved at the end of the twentieth century or it is an overloaded catchword demonstrating the inability of the RA to deliver a consistent explanation of changes—illustrate severe difficulties to open and use the RA for new topics and problems.

A regional perspective within RA has been introduced by different authors like Leborgne and Lipietz (1988), Esser and Hirsch (1994), or (Moulaert 1996). More recently Jessop and colleagues have announced what they call a 'third generation' of regulationists. They emphasize the ability of RA especially to answer questions of the changing form of the state and the relations between scale, space, and time. In their recent reviews (see Jessop 1997, MacLeod 1997), the RA is located between political science, geography, and sociology leaving the traditional macroeconomic perspectives of RA behind. However, in the new edition of Aglietta, the start-up analysis indicates an ongoing interest in macroeconomic regulation (Aglietta 2000).

We may conclude that the RA has experienced a remarkable differentiation in the last two decades of the twentieth century. This diffusion process did not favor a consistent body of thinking—one theory or school—but has led to a range of different concepts which today form the components of the RA. Though these concepts are far away from being parts of a whole, they are proofs of an ongoing interest of academics to organize their thinking around this approach. But there is also a tendency of the spatial and thematic diffusion of RA to a dissolution of the specialties of this approach into the broad stream of critical social science.

3. *Examples of Special Topics*

The diffusion process of the RA makes it difficult to focus on special topics. The following selection tries to outline those we think are probable pathways for future work, especially in geography. The common question of the first body of literature that we would like to emphasize here is addressing problems and changes of scale-dependant institutional arrangements. One of the major changes between the period of Fordism and today is the dissolution of the former regulatory fix that is grounded on the temporal compromise between capital and labor (based on rising wages related to productivity growth), and on macroeconomic regulation (based on the defined territoriality and relatively autonomous power of the nation state). Owing to the loss of importance of the concept of the welfare state as the democratic form of the 'golden age' of Fordism, a debate about new forms of institutional arrangements has evolved. Deregulation, privatization, and the evolution of new political re-gimes at local as well as at the international level have resulted in the debate about the concept of governance, which includes new public, private interrelations beyond and within the old concept of nation state.

Within this context, the work of Jessop is of particular interest. The shift from—what he calls—'the Keynesian welfare national state' to an emergent 'Schumpeterian workfare postnational regime' forms the center of his analysis. Three changes are emphasized:

(a) New state activities evolve which should strengthen the competitiveness of state spaces by stimulating and promoting permanent innovation processes. Besides direct incentives to create national systems of technological innovation, new institutional forms, resources, and values are invented in order to create a 'Schumpeterian' mode of regulation.

(b) System competitiveness and labor market flexibility are accompanied by a subordination of social policy to imperatives of the new economic orientation. The term workfare is used to underline a general tendency of post-Fordism to cut back welfare expenditures and to redirect social assistance in order to promote individualistic responsibilities and risk acceptance.

(c) The state itself changes its form substantially. The 'hollowing out' of the national state apparatus favors subnational as well as supranational levels. Though the nation state remains important, new state activities emerge on different scales. This deconcentration of state power is accompanied by a shift 'from government to governance' which reflects a growing importance of semi-public and private actors, illustrated by new urban and international regimes. At the same time, trends towards de-statization or decentralization are counteracted by new state functions (metagovernance), addressing activities such as self-organization of partnerships, networks, and governance regimes.

The concept of the 'Schumpeterian workfare postnational regime' is a promising and far reaching concept of regulation under post-Fordist conditions. Yet, though well elaborated in relation towards RA

and state theory (especially Poulantzas position), the concept has to be grounded in empirical work. In particular, the reconfiguration of scales and state activities need further analysis in order to improve our understanding of new organizational principles beyond the nation state.

A second promising body of literature has drawn attention to local and regional modes of social regulation (Tickell and Peck 1992). There are approaches which address urban management and politics (Mayer 1994), and urban social movements (Hitz 1995), local labor markets (Peck 1996) and regional development in a postsocialist environment (Krätke et al. 1997). In particular, in reaction to challenges of urban regime theory, a new theory has set in (a) to conceptualize the relationships between accumulation and regulation on urban and regional scale; (b) to find adequate solutions for structure–agency problems which are inherent in the RA; and (c) to direct concrete research (Goodwin and Painter 1997). According to Goodwin and Painter, the concept of regulation should not be viewed as a static and stable, but as a process which is constituted by material and discursive practices. Because these practices are temporally and spatially embedded, there are many possible patterns of regulation as contingent combinations that might operate to support accumulation. This variety of modes of regulation is enlarged by the possibility that quality and intensity of regulation might differ due to spatial differentiation and uneven development. Therefore, the process of regulation is constituted geographically as result 'of the contingent interaction of disparate practices' (Goodwin and Painter 1997, p. 21).

To view regulation as a process, rather than as a fixed mode, and to conceptualize the process as contingent practices is an important contribution to reduce missing links in RA. But new problems arise when these suggestions are followed. For example, in applying the concept of Goodwin and Painter, it becomes difficult to decide what kind of social practices might be relevant in order to reveal processes of regulation. If it is assumed that regulation might be a result of intended and unintended consequences of social practices, we are trapped in the universe of possible practices. Though the introduction of the habitus concept of Bourdieu may serve as a microfoundation of the RA and the identification of structures beyond the variety of practices (Painter 1997), the relationships between habitus and fields, on the one hand, and the concept of regulation as process, on the other, remain undefined. Therefore, if we give up the idea that the mode of regulation is like a development model beyond Bourdieu's fields of political contest, e.g., as a spatially and temporally defined class compromises, which the subjects are reflecting, the term regulation receives a technical connotation of a self-organized system that creates order out of nowhere.

4. *Future Directions*

There are several missing links and open questions within RA often partly addressed as a general critique. Some of the important deficits are as follows.

The theoretical status of the terms Fordism and especially post-Fordism has become unclear. Some of the regulationist thinkers seem to give up the idea of development models serving as a background necessary for the evolution of structural coherence in time and space.

The ecological question is still neglected in the majority of contributions to the RA and a further concept is needed within RA, addressing an economic–ecological compromise. The evidence that the Western mode of production is self-neglecting in terms of sustainability of ecological resources and is partly elaborated in the work of Alain Lipietz, who has become one of the major political leaders of the red–green movement in France and Europe (Lipietz 1995).

The immanent functionalism of the RA has to be overcome in order to refuse reductionist explanations of modes of regulation for stabilizing capitalist development.

Though the listed deficits are already challenges for future work within RA, some directions are addressed here because they should receive major attention. The first perspective for future research reflects the growing integration of RA, evolutionary economics and institutional approaches in the social sciences. Within this broad field, the term regulation as a mode or as a process has to be worked out in a more detailed way in relation to state theory and the growing body of literature dealing with 'governance' approaches. Whereas RA usually puts more emphasis on structure, macroeconomic forces and large-scale institutional mediations, the post-Fordist variety requires differentiated concepts (a) of the enabling and restricting qualities of institutionalized systems of norms and rules, (b) of forms of leaderships and regimes on different scales, and (c) of forms of social network integration through conventions, coercion, and strategic behavior. As we have shown, there are some promising starting points in RA, but a comprehensive approach on regulation is missing. Further, it is rather unclear what a geographical perspective might look like. Interpretations like reterritorialization and politics of scale (Brenner 1999) as a dominant feature of current restructuring of capitalist society, as well as the notion that 'all regulation is generated in temporally and spatially embedded social contexts' (Goodwin and Painter 1997, p. 13), are still beyond a clear understanding of the role of space and scale.

A second promising field of future research evolves within the debate about poststructuralism and the societal paradigm of the mode of regulation (e.g., Scherrer 1995, Jenson 1993). Though the more recent literature is very aware of the problem of economic

determinism and the latent functionalism of the RA, it seems to be promising to review the concepts of the RA using discourse analysis (e.g., Laclau and Mouffe 1985). This methodological inquiry may not only serve as a reformulation of basic approaches within RA (like the refusal of a mode of regulation in favor of a process, or a more cautious usage of the term Fordism), but also may lead to (new) answers of the question of political practices deriving from RA analysis.

See also: Capitalism; Finance, Geography of; Globalization, Anthropology of; Globalization: Legal Aspects; Globalization: Political Aspects; International Business; International Law and Treaties; Regulation, Economic Theory of; Regulatory Agencies; World Trade Organization

Bibliography

Aglietta M 2000 *A Theory of Capitalist Regulation*. Verso, London [first French edn. 1976; first English edn. 1979]

Boyer R 1990 The regulation school: a critical introduction. Columbia University Press, New York

Boyer R, Saillard Y 1995 Théorie de la régulation, etat des savoir. La Décoùverte, Paris

Boyer R, Hollingsworth J R 1997 From national embeddedness to spatial and institutional nestedness. In: Boyer R, Hollingworth J R (eds.) *Contemporary Capitalism. The Embeddedness of Institutions*. Cambridge University Press, Cambridge, UK, pp. 433–84

Brenner N 1999 Globalisation as reterritorialisation: the rescaling of urban governance in the European Union. *Urban Studies* **36**: 431–51

Esser J, Hirsch J 1994 The crisis of Fordism and the dimension of a 'post-Fordist' regional and urban structure. In: Amin A (ed.) *Post-Fordism. A Reader*. Blackwell, Oxford, UK

Goodwin M, Painter J 1997 Concrete research, urban regimes, and regulation theory. In: Lauria M (ed.) *Reconstructing Urban Politics in a Global Economy*. Sage, London, pp. 13–29

Hirsch J 1995 *Der nationale Wettbewerbsstaat: Staat, Demokratie, und Politik im globalen Kapitalismus*. ID-Archiv, Berlin

Hitz H 1995 *Capitales Fatales: Urbanisierung und Politik in den Finanzmetropolen Frankfurt und Zürich*. Rotpunktverlag, Zürich

Jenson J 1993 Ideas, spaces, and times in Canadian political economy. In: Jenson J, Mahon R, Brenefeld M A (eds.) *Production, Space, Identity: Political Economy Faces the 21st Century*. Canadian Scholar's Press, Toronto, ON, pp. 143–69

Jessop B 1995 The regulation approach, governance, and post-Fordism. Alternative perspectives on economic and political change? *Economy and Society* **24**: 307–33

Jessop B 1997 Survey article: the regulation approach. *Journal of Political Philosophy*. **5**: 287–326

Krätke S, Heeg S, Stein R 1997 *Regionen im Umbruch. Probleme der Regionalentwicklung an den Grenzen zwischen 'Ost' und 'West'*. Campus, Frankfurt/Main, Germany

Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy: Towards a Radical Democratic Politics*. Verso, London

Le Galès P 1998 Regulation and governance in European cities. *International Journal of Urban and Regional Research* **22**: 482–506

Leborgne D, Lipietz A 1988 New technologies, new modes of regulations: some spatial implications. *Society and Space* **6**: 263–80

Leborgne D, Lipietz A 1994 *Nach dem Fordismus*. In: Noller P, Prigge W, Ronneberger K (eds.) *Stadt-Welt. Über die Globalisierung städtischer Milieus*. Campus, Frankfurt am Main, Germany, New York, pp. 94–111

Lipietz A 1995 *Green Hopes: The Future of Political Ecology*. Polity Press, Cambridge, UK

MacLeod G 1997 Globalizing Parisian thought-waves: recent advances in the study of social regulation, politics, discourse, and space. *Progress in Human Geography* **21**: 530–53

Mayer M 1994 Post-Fordist city politics. In: Amin A (ed.) *Post-Fordism. A Reader*. Blackwell, Oxford, UK

Moulaert F 1996 Rediscovering spatial inequality in Europe: Building blocks for an appropriate 'regulationist' analytical framework. *Environment and Planning D: Society and Space* **14**: 155–80

Painter J 1997 Regulation, regime, and practice in urban politics. In: Lauria M (ed.) *Reconstructing Urban Politics in a Global Economy*. Sage, London, pp. 122–43

Peck J 1996 *Work-place. The Social Regulation of Labor Markets*. Guilford, London

Scherrer C 1995 Eine diskursanalytische Kritik der Regulationstheorie. *Prokla* **25**: S457–82

Stoker G 1995 Regime theory and urban politics. In: Judge D, Stoker G, Wolman H (eds.) *Theories of Urban Politics*. Sage, London, pp. 54–71

Tickell A, Peck J 1992 Accumulation regimes and the geographies of post-Fordism: missing links in regulationist theory. *Progress in Human Geography* **16**: 190–218

Waringo K 1998 *Die Internationalisierung der Produktion in der französischen Regulationstheorie*. Campus, Frankfurt/Main, Germany

R. Danielzyk and J. Ossenbrügge

Regulation: Working Conditions

Working conditions encompass the aspects of work ranging from the terms under which workers enter the employment relationship to the environment in which they perform their tasks. Historically, working conditions have varied widely across countries and across time within each country in terms of their quality—from the workers' vantage point. Working conditions have been subject to regulation to varied degrees, a situation which accounts for variability in their quality.

Working conditions are part of employment systems that have been shaped by actions, rules, and regulation emanating from several spheres of society. These include the legal, judiciary, administrative, and collective bargaining spheres as well as personnel policies and informal social norms. All of these spheres shape

the 'way of doing business' (Piore 1990) as well as working conditions for workers. This article reviews how working conditions are shaped by regulation. In industrialized societies, the regulation of working conditions has primarily affected people in waged or salaried employment, that is, those who are in a relationship, recognized legally as economically dependent, with an employer who usually controls the means of production.

1. Working Conditions, Definition

Working conditions consist of the following elements:

(a) *The terms of the employment relationship*. They include whether the employment relationship is durable or short-term, formal or informal. The terms of employment may be explicit or implicit.

(b) *Access to employmen*t. This access may be equitable or discriminatory. In this latter case, groups of workers may be excluded from employment or have differential access based on their race/ethnicity, gender, age, or other characteristic.

(c) *The workplace itself*. Work may be performed in a single site, multiple sites, or from the worker's home.

(d) *The wage level*.

(e) *Time at work and away from work*. This includes the daily, weekly, and yearly number of work hours as well as their scheduling. It also includes provisions for time away from work due to illness, vacation, or leave for family or medical reasons or for training.

(f) *The supervisory structure and mechanisms*. This structure may or may not include procedures for due process in disciplining, discharges, and layoffs. It may entail varied degrees of protection from supervisory arbitrariness.

(g) *Health and safety conditions in the workplace*. These include occupational health, that is, the ways in which the performance of work tasks affects worker physical health and well-being, and environmental health, that is, whether the environment in which the work is performed has negative effects on physical health.

(h) *Voice at work*. This encompasses access to worker representation and the shape of representation structures. These structures may include representative organizations such as labor/trade unions or worker associations as well as participatory structures such as workplace councils.

(i) *Social protection/social insurance benefits for the individual and their dependents that have been tied to waged employment and are administered through payroll systems*. These include, for example, health insurance and pension plans.

Some of the elements of working conditions listed above have a direct relationship to employment and the performance of work, e.g., health and safety conditions at work. Others have been historically tied to employment with a particular employer in the USA and to labor market participation in Western European countries but do not have to be. For example, health insurance coverage can also be provided as a universal benefit regardless of work activity.

2. Systems of Employment and Working Conditions

National systems of employment, of which the US system is one example, can be viewed as a whole which provides degrees of regulation and of protection to workers and employers in a variety of ways (Jacoby 1982, Boyer 1986, Piore 1986). Systems of employment can be seen as a conjunction of 'spheres' which contribute to varied degrees toward shaping working conditions over time. These 'spheres' are legal, executive/administrative, judiciary, collective bargaining, and employer personnel policies to the extent that they embody 'social protection' norms, as well as informal social norms. Different countries have a varied mix of reliance upon these spheres to affect working conditions. Some rely on collective bargaining approaches whereas others favor the action of the legislative and executive spheres combined. The regulation of working conditions can therefore be achieved through these spheres in a different manner across countries as well as across time in one country.

3. The Purpose of Regulation on Working Conditions

The regulation of working conditions has several purposes. Most commonly, it establishes labor standards, that is, set minima for wage, hours, or health and safety conditions that employers are mandated to respect and that are enforced through prohibition and penalties. This corresponds to the narrowest role of regulation.

Regulation can permit the formation of structures for negotiation, mediation, or problem solving that contribute to shaping working conditions in varied ways and with diverse outcomes across industrial sectors and occupations (Weiler 1990). One example of such regulation is the federal and state statutes governing worker representation and collective bargaining (e.g., National Labor Relations Act of 1935).

4. The Regulation of Working Conditions

Most narrowly and conventionally, regulation is understood as emanating from government (national, state, local) through laws and statutes (legislative branch), as well as administrative rules (executive). In the USA, separate agencies have been established over

time to develop enforcement mechanisms. For example, the Wages and Hours Division in the US Department of Labor enforces the Fair Labor Standards Act of 1938 and related legislation on minimum wages and overtime pay for hourly workers. Through inspectors and administrative courts, violations are verified and penalties issued for responsible employers. Enforcement mechanisms and their effectiveness play a key role in the shaping of working conditions. Conversely, the work setting affects the effectiveness of enforcement. For example, if workplaces are small, scattered, and numerous, it is difficult for an enforcement agency that relies on visits by inspectors to monitor work conditions and enforce a specific law.

The court system is another feature of the regulation of working conditions. Courts contribute to enforcing existing laws and statutes. For example, an employer responsible for hazardous working conditions having resulted in worker death may be sued. Sometimes, in response to cases initiated by workers, courts issue decisions in areas of working conditions where government legislation and executive enforcement have not acted yet. In this case, court actions precede legislative action.

Working conditions are also affected by the relative bargaining power of workers and their organizations. In turn, this bargaining power is shaped partly by the provisions made in legislation for worker voice and representation. For example, the National Labor Relations Act of 1935 established a mechanism for representation for many US workers in the post-World War II years and also shaped the form that labor unions took. The relative bargaining power of labor unions and worker organizations affects the ability of workers to bargain for improvements, the thoroughness of enforcement of existing laws, and the ability to obtain new laws.

Lastly, social norms and customs along with worker relative bargaining power are reflected in personnel policies and employer practices; these also affect working conditions. In the USA, many aspects of working conditions are shaped by personnel policy rather than law or collective bargaining and thus vary greatly across employers.

In sum, working conditions are shaped by the totality of action in these respective spheres and by the interaction of actions from these spheres. For example, in the USA, provisions in collective bargaining agreements or in innovative personnel policies that affect an aspect of working conditions that is not the subject of government regulation may serve as a model for state or federal legislation.

In other national systems, such as France, national sectoral bargaining in lead industrial sectors on a particular aspect of working conditions often precedes national legislation. National sectoral bargaining is a negotiation of large employers and their federations with national unions that is sometimes mediated by government and whose outcome is extended to an entire industry. In some countries, national sectoral bargaining on a particular issue (e.g., implementing shift work) is used sometimes as an alternative to government regulation.

5. *The Long-standing Debate on the Regulation of Working Conditions*

The role and benefits of the regulation of working conditions have been the subject of long-standing debate. One view sees regulation, particularly regulation on labor standards, as an externally imposed cost and/or as having a distorting effect on the operation of the labor market. This view stems from the neoclassical school of economics whose core perspective—presented here in stylized fashion—is that markets are self-regulating and that prices must be allowed to vary freely in order to be efficient allocators of human and capital resources. This view also tends to assume there is no significant conflict or power differential between workers as a group and employers as a group. Rather, workers and employers are individual decision makers who make choices unhampered by power differentials. Critics of federal and state minimum wage laws have argued that the laws push wages above the 'market clearing' level and thus cause unemployment. Other analysts have taken a more benign view of their impacts on firm decisions and employment levels. The research literature has centered on the empirical measurement of the relative magnitude of effects on employer costs and employment levels (Card and Krueger 1995, Houseman 1998).

Another view in this debate sees the regulation of working conditions and its effects as a special case in economic analysis because labor markets are deemed the least self-regulating among all markets. In this view, the regulation of labor markets cannot easily be analyzed with the conventional tools of economic analysis.

Consistent with this view, another perspective sees regulation as a reflection of customs and social norms and of power relations between employers and workers—all of which vary across countries. In this view, regulation may permit, hinder, add costs to, or facilitate greater efficiency in production. In other words, the effects of the regulation of working conditions are not conceived as unidirectional. They are expected to vary widely depending on the contents of regulation, the structure of particular industries, and their production organization. For example, minimal labor standard on wages, hours, or health and safety have been argued to discourage 'sweatshops' and favor other business strategies that are less onerous to workers (Piore 1990). In sweatshops, workers are unprotected because employers are insulated from the economic impact of variations in productivity among workers and over time thanks to the piece-rate system (workers are paid for measurable units of output). A

minimum wage rate which is high enough to create employer concern with productivity may compel a change in a firm's practice and working conditions; it may deter the use of low-productivity workers such as young children.

In the USA, recent debates on the regulation of working conditions have been colored by the general policy trend toward deregulation of markets. Laws on hours and scheduling, and those on worker representation structures have been questioned. Advocates of deregulation have argued that existing regulations fit an old order of production organization, and market structure and must be removed. Others have called for their replacement with different standards and representation systems that provide equivalent protections for workers but through new means that are more appropriate for contemporary forms of production and work organization.

6. *Future Trends and Research Issues*

6.1 *New Forms of Employment*

In the post-World War II period, the contents and enforcement of the regulation of working conditions in late industrialized countries have been constructed, albeit in different ways, around a certain notion of the employment relationship as a relatively long-term, stable attachment to a single employer with the work tasks being performed on the employer's work site. While there were always exceptions to this notion—for example, farm labor, construction, expert consulting, or domestic labor—the bulk of post-war regulatory attention to working conditions has been structured around this employment arrangement.

In the future, most observers expect less employer commitment to long-term employment relationships, less worker attachment to particular employers, and greater lateral mobility (across firms, industries, and possibly occupation). This is due to changes in the organization of production and employment structures within firms; increasingly, firms recruit workers from the outside labor market rather than rely on internal training and promotion. Career paths for workers will become less predictable and may span several companies and industries (Arthur and Rousseau 1996, Cappelli et al. 1997).

The manifestations of these transformations include the emergence and establishment of varied forms of employment which are explicitly short term (limited-duration hiring, temporary assignment), may be brokered by an intermediary (a for-profit or nonprofit temporary/staffing service), or may be the result of contracting arrangements between two businesses (e.g., workforce of a janitorial subcontractor), thus creating a triangulated employment relationship. Independent contracting has also appeared—a situation where individual workers are treated as self-employed under the law, that is, not wage/salary employees, while they have a sole purchaser of their service and are economically dependent on the relationship.

Short-term employment relationships may render existing regulations and their enforcement ineffective; workers may not stay long enough in a particular job to bring a complaint about a violation of a labor standard. Triangulated employment relationships entail ambiguities as to employer responsibility for compliance with particular laws. Independent contracting means that there is no employment relationship and therefore no option for regulation under most labor and employment law. In practice, the financing and implementation of social protection benefits such as health insurance and pension through employment relationships are weakened if not rendered ineffectual.

As employment relationships change in these ways, the contents and means of the regulation of working conditions will likely change. The regulation of working conditions will require a reconceptualization and analysis.

Although employment relationships change in these ways, they still remain the primary means through which core social protection benefits continue to be provided. In the USA, which has relied on social protection benefits being provided by employers either voluntarily (fostered by tax deduction incentives) or through rare mandates, regulatory attention has focused on mechanisms to enable workers better to meet their responsibilities to family. For example, family leave—time off for caring for a child or sick relative—has been the subject of US legislation (in 1993), of continued activism at the state and federal level, and of innovative benefit policies by some large employers. These developments reflect the increased presence of members of dual earning couples as well as that of single parents (mostly female) in the workforce (Blau and Ehrenberg 1997).

6.2 *The Regulation of Working Conditions and Globalization of Production*

International economic integration has accelerated in recent years and is expected to continue. Its main features include the reduction or removal of trade tariffs and of restrictions on capital flows either by policy change or simply through the use of new communication technologies (which put distant production areas directly in competition). It has resulted in greater international capital mobility, the freer movement of goods, and the development of production processes ('production chains') that span national boundaries. This pattern holds particularly for manufacturing and, increasingly, for service activities (Whitman 1999).

In this context, the issue of the regulation of working conditions will take on new facets requiring further consideration and research. First, the adaptation and strategic responses of firms in industrialized countries to minimal labor standards were developed during the post-World War II period when mass production was in expansion, and domestic markets were relatively sheltered from foreign competition. Growing product competition from countries with dramatically lower wages and fewer regulations of working conditions, as well as the increased opportunities for firms to expand their production processes cross-nationally, are changing the parameters that affect working conditions for workers in both industrialized and developing countries. These developments challenge the ability of national and local governments to regulate working conditions in ways that provide for effective enforcement.

Second, in developing countries, the development of activities linked to the production processes of cross-national firms has given rise to forms of employment (e.g., electronic parts assembly at home) which are not subject to existing national regulation of working conditions, or which take place in settings and on terms in which monitoring and enforcement are difficult (see debates on the 'informal economy').

Additionally, the ability of countries enforce the regulation of working conditions through labor standards in those developing countries with low wages and low-quality working conditions is weakened by the ability of cross-national firms to shift production to another low-wage country with a workforce of similar skill level.

Three groups of issues will need to be considered in future research. First, the parameters of firms' choices will need to be documented in the context of particular countries and for specific activities, contrasting capital-intensive and labor-intensive manufacturing as well as comparing manufacturing with service activities.

Second, future research on regulation will increasingly examine the sphere of international organizations (e.g., World Trade Organization, Organization for Economic Cooperation and Development, International Labor Organization), the process of international negotiation of conventions on working conditions, and the impacts of these conventions, first, on national regulation efforts and, second, on firm strategies and practices regarding working conditions.

Lastly, for policy analysis and discussion, the core question will be what *should* be the complex of regulation of working conditions that is most compatible with new forms of cross-national production, the productive use of new production and communication technologies, and the well-being of workers in industrialized and developing countries?

See also: Environmental Health and Safety: Social Aspects; Industrial Policy; Industrial Relations and Collective Bargaining; Industrial Sociology; Injuries and Accidents: Psychosocial Aspects; Safety, Economics of; Social Insurance: Legal Aspects; Work, History of; Work, Sociology of; Workplace Safety and Health

Bibliography

Arthur M B, Rousseau D M (eds.) 1996 *The Boundaryless Career: A New Employment Principle for a New Organizational Era*. Oxford University Press, New York

Blau F D, Ehrenberg R G 1997 *Gender and Family Issues in the Workplace*. Russell Sage Foundation, New York

Boyer R 1986 *La Flexibilité du Travail en Europe*. Editions La Découverte, Paris

Cappelli P, Bassi L, Katz H C, Knoke D, Osterman P, Useem M 1997 *Change at Work*. Oxford University Press, New York

Card D, Krueger A B 1995 *Myth and Measurement: The New Economies of the Minimum Wage*. Princeton University Press, Princeton, NJ

Doeringer P B, Piore M J 1971 *Internal Labor Markets and Manpower Analysis*. D C Heath, Lexington, MA

Houseman S 1998 The effects of employer mandates. In: Freeman R B, Gottschalk P (eds.) *Generating Jobs: How to Increase Demand for Less-skilled Workers*. Russell Sage Foundation, New York

Jacoby S M 1982 The duration of indefinite employment contracts in the United States and England: An historical approach. *Comparative Labor Law* **5**: 85–128

Jacoby S M 1997 *Modern Manors: Welfare Capitalism Since the New Deal*. Princeton University Press, Princeton, NJ

Piore M J 1986 Perspectives on labor market flexibility. *Industrial Relations* **25**: 146–66

Piore M J 1990 Labor standards and business strategies. In: Herzenberg S, Perez-Lopez J F (eds.) *Labor Standards and Development in the Global Economy*. Government Printing Office, Washington, DC pp. 35–49

Standing G 1999 *Global Labor Flexibility: Seeking Distributive Justice*. St. Martin's Press, New York

Weiler P C 1990 *Governing the Workplace: The Future of Labor and Employment Law*. Harvard University Press, Cambridge, MA

Whitman M N 1999 *New World, New Rules: The Changing World of the American Corporation*. Harvard Business School Press, Boston, MA

F. J. Carré

Regulatory Agencies

Broadly defined, regulation includes any governmental effort to control behavior by other entities, including business firms, subordinate levels of government, or individuals. Sovereign governments may cooperate under regimes of mutually-recognized restraint, the rules or decisions of which might also be characterized as regulations. Campaigns for elective office may likewise be subject to controls on spending

and advertising. However, in most public discourse and social science inquiry, the term refers more narrowly to formal government programs (or to specific implementing decisions) intended to constrain economic activity in accord with some public-interest justification.

Four primary concerns of regulation in this sense may be distinguished (Breyer 1982, Stone 1982). First, regulation may control (as in the case of occupational licensing) market entry. Second, it may (as in the case of electricity utilities and other 'natural monopolies') establish directly the prices that firms can charge customers. Third, it may limit the competitive practices available to firms, especially as regards price-setting (e.g., the law of antitrust), advertising, and disclosure. Fourth, regulation may aim to improve the quality of specific work practices or products, often in the interest of protecting human health and safety, or the environment. These activities are sometimes subsumed under two more general hearings—'economic' (i.e., price, entry, and competitive practices), and 'social' (meaning generally health, safety, and environmental) regulation.

Social scientists have posed three primary empirical questions about regulatory programs. First, what forces lead to the creation of such programs? Second, and most commonly, they have asked what constellation of factors influences program behavior? Third, under what conditions may regulatory programs be significantly scaled-back or terminated? Normative questions also proliferate. Some varieties of regulation, especially in the United States, have been prone to intense controversy. Both business representatives and policy-focused academics have pursued long-running debates about the merits of various regulatory programs, the quality of their specific decisions, and possible institutional alternatives to the status quo.

1. *Empirical Inquiry*

Students of regulation have long found the formally-stated public goals of regulation inadequate to explain program creation and maintenance. From the 1950s to the 1970s, working within disciplines as diverse as economics, political science, and historical analysis, commentators viewed regulatory programs frequently as being prone to 'capture' by regulated interests (Huntington 1952, Bernstein 1955, Kolko 1965, Stigler 1971). By some accounts, regulation was said to have arisen because a diffuse majority favoring regulation (and motivated by a 'public interest' rationale, such as the elimination of predatory pricing or consumer fraud) succeeded, at least briefly, in obtaining legislative consensus, often in the wake of well-publicized scandal. Over time, however, according to this 'life cycle theory' of regulation posited by Bernstein (1955), waning public attention and media scrutiny was said to have allowed resourceful regulated interests, whose attention to regulatory program activity did not wane, to exercise a strong (even dominant) influence over the program, thus thwarting its original intent. A different version of capture analysis suggested that such programs were sometimes created, at least partially, at the behest of economic interests. Agencies might thus, in effect, be 'born captured' due to business incentives to seek stable and reliably profitable operating environments. The capture model of regulation is in some respects akin to a traditional political science metaphor, widely employed during the same period, that described the politics of narrowly distributive public policies in terms of mutually-reinforcing 'iron triangles' or subgovernments composed of interest groups, legislative committees, and agencies (Freeman 1965). The similarity is striking, even though an analytic scheme widely appropriated among political scientists (Lowi 1964) highlighted a distinction between 'distributive' and 'regulatory' politics.

During the 1970s and thereafter, however, scholars challenged increasingly the descriptive accuracy of both the capture theory and the 'iron triangle' metaphor. In the United States especially, this challenge stemmed importantly from the creation of regulatory programs (e.g., in environmental, consumer, and worker protection) where lack of media scrutiny, public indifference, and business dominance plainly did not apply over the long run (Wilson 1980). Indeed, the creation of these programs often represented sustained triumphs by idea-driven policy entrepreneurs over business interests that had to adjust to the presence of long-term political competitors. In one prominent instance, the regulation of nuclear power generation, a policy milieu changed dramatically from one to which the term 'capture' could arguably have applied in the 1950s. By 1980, the politics of nuclear power in the United States had evolved into a far more complicated set of power relations, in which environmental advocates and citizen groups competed effectively with the nuclear industry (Temples 1980).

Throughout the 1980s, political science commented increasingly about the inadequacy of broad generalizations regarding regulatory politics as a whole. The concentration of costs and benefits prevailing in regulation were found to differ importantly among agencies, and across issues within a single agency (Wilson 1980). The politics of broadcast regulation, pursued within a markedly vague standard of 'public interest, convenience, and necessity,' clearly varied between technical industry structure matters and issues of program content (Krasnow et al. 1982). Within public utility regulation, not only were issues determined to differ in complexity and conflict, with significant implications for the power of different kinds of advocates (Gormley 1983), but the characteristics of commission personnel were also found to vary in ways important to generating distinctive regulatory outputs (Berry 1984).

Regulatory policy also became the staging ground for scholarly argument over the relative influence of various political institutions. A conventional position among journalists and political scientists had for many years held that Congress does little to oversee bureaucratic behavior, largely because of the modest political incentives said to prevail (Scher 1963). In arguing against what he called 'interest-group liberalism,' Lowi (1969) bemoaned the vagueness that had crept into legislative delegations for regulatory and other programs.

A later principal–agent perspective on the legislative–executive relationship argued for the existence of a more robust oversight mechanism grounded in a rational legislative approach to structure and procedure that appeared only superficially to be weak (McCubbins and Schwartz 1984). Careful sifting of agency case histories determined that widely-trumpeted accusations of 'runaway bureaucracy' often directed at regulation were misplaced (Foreman 1988).

Because regulation is politically contentious and generates substantial numbers of comparable implementation and enforcement decisions over time, political scientists (often working within the tradition of rational-choice models and principal–agent theory) have gravitated to regulation as a window on relative institutional influence more generally. Some analysts have emphasized the power of legislative principals (activated by the material stakes of affected parties) over agency behavior (Weingast and Moran 1983, Weingast 1984). Meanwhile, Moe (1982), examining time series data drawn from three US regulatory commissions, concluded that presidential power deserved emphasis. On the other hand, some analysts have looked inside regulatory bureaucracies, to the professional worldviews and incentives of staff members, as key drivers of agency behavior. In antitrust enforcement, economists and attorneys are among those professionals whose views have been found to have a profound impact on agencies, both in case studies (Katzmann 1980, Weaver 1977) and in quantitative analysis (Eisner and Meier 1990). On the other hand, the theory of overhead democracy embraced by Wood and Anderson (1993) ascribed variation in both the level and substance of antitrust regulation to a complex set of institutional forces; changing presidential administrations, congressional committee preferences, and court decisions all helped to mold antitrust activity.

Deregulation is rarer, and politically more problematic, than regulatory initiatives. As regards social regulation, such goals as a clean environment, and safe workplaces and pharmaceutical products, are popular with the mass of the public. In economic regulation, organized stakeholders would be expected to resist unfavorable change (i.e., alteration of profitable regulated environments). Given such forces, it is worth emphasizing that deregulation has occurred on occasion, and asking how this has come about. Political scientists Derthick and Quirk (1985) point to the power of both ideas and effective leadership. They observe that, by the late 1970s, economists had mounted a sustained, coherent, and nearly unanimous attack on the underlying justifications for price and entry regulation in the transportation sector. This pro-competition critique gained added political impetus from the increasing visibility of economists in national policy debates, from the infiltration of economists into leadership of the regulatory apparatus, and from the transformation of deregulation into an anti-inflation and consumerist cause that both liberals and conservatives could support. Strong presidential and legislative leadership also proved to be pivotal.

2. *Normative Assessments*

Nearly all regulatory programs will find supporters and detractors in the social sciences. Unlike price and entry regulation, which regularly has found advocates of abolition in the economics profession, other types of regulation enjoy strong academic support on principle (e.g., recognition of 'market failure') though often not in specific detail.

Leading the charge for varieties of regulatory reform have been economists and economically-oriented analysts associated with policy research institutions or 'think tanks.' Criticisms of environmental policy stress varieties of putative inefficiency or irrationality. Costs are said to exceed benefits, to be higher than necessary to achieve a given level of environmental benefit, and perhaps to be masked altogether by inadequate economic analysis (Kraft 1996, pp. 165–6, Litan and Nordhaus 1983, Lave 1981). Inefficiency is sometimes argued to have been caused by the logic of political compromise (Ackerman and Hassler 1981), or by unwillingness to impose overtly visible losses on large numbers of voters (Nivola and Crandall 1995). Inadequate monitoring has long been said to bedevil assessment of pollution control efficacy (Crandall 1983, Davies and Mazurek 1998).

A broader, efficiency-based critique encompasses additional concerns, including the discordant, meandering, and arguably even panic-stricken character of regulatory politics and policy (Breyer 1993). Critics point to a chronic inability to set risk priorities, to confront risk–benefit and risk–risk tradeoffs (Graham and Wiener 1995), and to reconsider regulatory statutes in the light of such concerns.

3. *International Perspectives*

All developed nations regulate, but they vary substantially in both regulatory structure and regulatory political culture. The United States is often portrayed as having a markedly more adversarial and legalistic system than other countries. One study concluded that

while 'regulatory power has been directed at very similar ends [i.e., control of carcinogens]' in Britain, France, Germany, and the United States, 'the openness, complexity and divisiveness of American decision making have no parallel in Europe' (Brickman et al. 1985, pp. 301–2). It is not clear, however, that the American style of regulation is necessarily more stringent than that of other nations. Conceding the more open regulatory style of the United States in comparison with Canada, Harrison and Hoberg (1994) nevertheless do not find the latter consistently less stringent than the former. Although institutions matter, their effects are 'contingent on a number of other variables that influence policy, such as the balance of power between competing interest groups and the interests of the party in power' (p. 181). Nevertheless there can be little doubt of the generally greater levels of hierarchy, and the more tightly-woven informal networks of personal relations, that characterize regulatory regimes in other advanced democracies in comparison with the United States (Badaracco 1985).

See also: Business and Society: Social Accounting; Policy History: State and Economy; Policy Process: Business Participation; Public Interest; Regulation and Administration; Regulation, Economic Theory of; Regulation: Empirical Analysis; Regulation: Working Conditions

Bibliography

Ackerman B A, Hassler W T 1981 *Clean Coal/Dirty Air: Or How the Clean Air Act Became a Multibillion Dollar Bail-out for High-sulfur Coal Producers and What Should be Done About It*. Yale University Press, New Haven, CT

Badaracco J L 1985 *Loading the Dice: A Five-country Study of Vinyl Chloride Regulation*. Harvard Business School Press, Boston, MA

Bernstein M H 1955 *Regulating Business by Independent Commission*. Princeton University Press, Princeton, NJ

Berry W D 1984 An alternative to the capture theory of regulation: The case of state public utility commissions. *American Journal of Political Science* **28**: 524–58

Breyer S 1982 *Regulation and its Reform*. Harvard University Press, Cambridge, MA

Breyer S 1993 *Breaking the Vicious Circle: Toward Effective Risk Regulation*. Harvard University Press, Cambridge, MA

Brickman R, Jasanoff S, Ilgen T 1985 *Controlling Chemicals: The Politics of Regulation in Europe and the United States*. Cornell University Press, Ithaca, NY

Crandall R W 1983 *Controlling Industrial Pollution: The Economics and Politics of Clean Air*. Brookings Institution, Washington, DC

Davies J C, Mazurek J 1998 *Pollution Control in the United States: Evaluating the System*. Resources for the future, Washington, DC

Derthick M, Quirk P J 1985 *The Politics of Deregulation*. Brookings Institution, Washington, DC

Eisner M A, Meier K J 1990 Presidential control versus bureaucratic power: Explaining the Reagan revolution in antitrust. *American Journal of Political Science* **34**: 269–87

Foreman C H 1988 *Signals from the Hill: Congressional Oversight and the Challenge of Social Regulation*. Yale University Press, New Haven, CT

Freeman J L 1965 *The Political Process: Executive Bureau—Legislative Committee Relations* (rev. edn.). Random House, New York

Gormley W T 1983 Policy, politics, and public utility regulation. *American Journal of Political Science* **27**: 86–105

Graham J G, Wiener J B (eds.) 1995 *Risk vs. Risk: Tradeoffs in Protecting Health and the Environment*. Harvard University Press, Cambridge, MA

Harrison K, Hoberg G 1994 *Risk, Science, and Politics: Regulating Toxic Substances in Canada and the United States*. McGill-Queen's University Press, Montreal, Canada

Huntington S P 1952 The Marasmus of the ICC. *Yale Law Journal* **61**: 467–509

Katzmann R A 1980 *Regulatory Bureaucracy: The FTC and Antitrust Policy*. MIT Press, Cambridge, MA

Kolko G 1965 *Railroads and Regulation, 1877–1918*. Norton, New York

Kraft M E 1996 *Environmental Policy and Politics: Toward the Twenty-first Century*. HarperCollins, New York

Krasnow E, Longley L, Terry H 1982 *The Politics of Broadcast Regulation*. St Martin's Press, New York

Lave L B 1981 *The Strategy of Social Regulation: Decision Frameworks for Policy*. Brookings Institution, Washington, DC

Litan R E, Nordhaus W D 1983 *Reforming Federal Regulation*. Yale University Press, New Haven, CT

Lowi T J 1964 American business, public policy, case studies and political theory. *World Politics* **16**: 677–715

Lowi T J 1969 *The End of Liberalism: Ideology, Policy and the Crisis of Public Authority*. Norton, New York

McCubbins M, Schwartz T 1984 Congressional oversight overlooked: Police patrols vs. fire alarms. *American Journal of Political Science* **28**: 165–79

Moe T M 1982 Regulatory performance and presidential administration. *American Journal of Political Science* **26**: 197–224

Nivola P S, Crandall R W 1995 *The Extra Mile: Rethinking Energy Policy for Automotive Transportation*. Brookings Institution, Washington, DC

Scher S 1963 Conditions for legislative control. *Journal of Politics* **25**: 526–51

Stigler G 1971 The theory of economic regulation. *Bell Journal of Economics and Management Science* **2**: 3–21

Stone A 1982 *Regulation and Its Alternatives*. Congressional Quarterly Press, Washington, DC

Temples J R 1980 The politics of nuclear power: A subgovernment in transition. *Political Science Quarterly* **95**: 239–60

Weaver S 1977 *Decision to Prosecute: Organization and Public Policy in the Antitrust Division*. MIT Press, Cambridge, MA

Weingast B R 1984 The congressional bureaucratic system: A principal–agent perspective (with applications to the SEC). *Public Choice* **44**: 147–91

Weingast B R, Moran M 1983 Bureaucratic discretion of congressional control: Regulatory agency policy making at the FTC. *Journal of Political Economy* **91**: 765–800

Wilson J Q 1980 (ed.) *The Politics of Regulation*. Basic Books, New York

Wood B D, Anderson J E 1993 The politics of US antitrust regulation. *American Journal of Political Science* **37**: 1–39

C. H. Foreman Jr.

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Rehabilitation (Cardiac): Psychosocial Aspects

The primary aim of rehabilitation of patients with physical conditions such as skeletal, spinal, or brain injuries, or following a cardiac event such as a myocardial infarction, is to enable the patient to return to as near their level of pre-morbid functioning as their physical condition allows. Since much heart disease is a product of an unhealthy life-style, an important secondary aim in cardiac rehabilitation is to enable the patient to change their behavior to reduce the risk of further arterial damage and recurrent cardiac events.

Cardiovascular disease is a major problem in most industrialized countries. Currently in the UK approximately 300,000 people per year have myocardial infarction,of whom 50 percent survive for at least 28 days. There are also large numbers with the disabling symptoms of coronary disease such as the 1.4 million individuals in Britain with angina. While there has been progress in the prevention of coronary artery disease in most counties with a high incidence of heart disease, an aging population and advances in medical treatments mean the number of people with coronary disease and its consequences is in fact increasing. A sizeable proportion of these people would benefit from cardiac rehabilitation. While prolonged inpatient rehabilitation—often in sanatoriums in country environments thought to be healthy—has been common in some countries for many years, the development of intensive outpatient programs is a product of the last quarter of the twentieth century. Founded with great enthusiasm, the early results from cardiac rehabilitation programs were disappointing but developments in therapeutic and data analytic techniques have led to a much more positive picture emerging since the mid-1990s. Many now hold that that comprehensive cardiac rehabilitation can improve the quality of life and physical health of patients with coronary heart disease and most sophisticated systems of healthcare aim to offer rehabilitation to patients suffering the effects of cardiovascular disease (see *Chronic Illness: Quality of Life*).

1. *Main Features of Cardiac Rehabilitation*

The main group of patients who receive cardiac rehabilitation comprises survivors of acute myocardial infarction, although patients who have undergone coronary bypass surgery or angioplasty may also receive it. Since the latter are often elective procedures, rather than the unpredicted disaster of a heart attack, the patients' problems may differ and less is known about the most appropriate forms of rehabilitation for them. The first cardiac rehabilitation programs were primarily based around exercise regimes and this remains the core of many programs. However with the increasing understanding of the psychological effects of myocardial infarction and the psychological and behavioral determinants of coronary heart disease (Krantz and Lundgren 1998), programs have widened to include many psychosocial components. There is no agreed standard program of cardiac rehabilitation, and programs vary widely in practice, but a modal program would be offered in an outpatient setting, starting six to eight weeks after discharge from hospital, would run for two or three months and would include regular exercise, an educational component about life-style change, and some procedures—such as relaxation training—to reduce distress. Exercise is usually conducted in specialist hospital gymnasiums, often with careful monitoring of the heart rate to diminish the risk to the patient. Rehabilitation programs are usually offered to middle-aged patients but might not be offered to the elderly, although they are the main sufferers from heart disease. Uptake of rehabilitation can be poor, especially in older patients. Such rehabilitation programs are expensive and difficult to deliver to the infirm, the socially disadvantaged, or those who live far from a specialist centre. Fortunately, effective versions of some aspects of rehabilitation can also be delivered by manual and supplementary telephone contact (Lewin et al. 1992), or by brief educational or counselling interventions while the patient is still in hospital (Johnston et al. 1999).

2. *The Effects of Cardiac Rehabilitation*

2.1 *Psychosocial Outcomes*

It would appear self-evident that patients will be distressed immediately after an unexpected, painful, and life-threatening event such as an acute myocardial infarction, and that this distress will continue for some time. While this is true of perhaps 15 to 30 percent of patients, a number of studies have shown that anxiety and depression are only slightly raised in most patients while still in hospital, but rise several weeks later and may remain elevated for at least a year (Johnston et al. 1999). Perhaps as significant is the finding that anxiety and depression are very high in the patients' partners while the patients are in hospital and these remain high—higher than that of the patients—for at least the subsequent year. There are many studies using rehabilitative techniques designed to reduce the distress associated with an infarction. Issues covered in such interventions will typically include an explanation of the attack and the healing process, including medical treatments and investigations, information on resumption of activities, and material on the emotional effects of a heart attack and the management of stress. Many recent studies of such interventions are positive and until recently there was a clear consensus, which had been confirmed by meta-analysis (Linden et al.

1996), that anxiety and distress could be reduced by appropriate rehabilitation programs, including programs that focused primarily on exercise (Kugler et al. 1994). However, an analysis by Dusseldorp et al. (1999) failed to confirm this finding. This is almost certainly because of the inclusion of two major studies with substantial samples that failed to find a reliable effect of psychosocial rehabilitation on distress (Jones and West 1996, Frasure-Smith et al. 1997). These studies are aberrant, and while their impressive size (which came about because they were primarily designed to detect effects on reinfarction rate) means that they have a substantial effect on summary analyses, few in the field accept that distress cannot be reduced after a myocardial infarction. Other studies have shown highly significant effects of simple interventions (Lewin et al. 1992, Johnston et al. 1999) on distress and on the resumption of the normal activities of daily life. Johnston et al. also showed that anxiety and depression was greatly reduced in the partners of patients if they were included in the rehabilitation program.

2.2 *Risk Factor Reduction*

A substantial proportion of heart disease is predicted by a few well-recognized and modifiable risk factors, all of which relate to some extent to behavior. These are elevated blood pressure, some lipids (including cholesterol), and cigarette smoking. Dusseldorp et al. (1999) have provided the most complete account of the effects of psychoeducational programs on the reduction of these conventional risk factors. All were reliably reduced by the interventions.

As well as these well-accepted risk factors there are a number of psychological risk factors that, although not necessarily accepted by the whole medical community, have substantial epidemiological support. Currently the most prominent of these are Type A behavior and depression (see Krantz and Lundgren 1998), and there is increasing interest in social support. Type A behavior, a pattern of behavior characterized by competitiveness, time pressure, and hostility, has been recognized as a risk factor for myocardial infarction for some time. However, possibly because of continuing controversy over which component of this complex behavior pattern is actually harmful, there have been few studies that both modified the behavior and were large enough to detect any effect on physical health. By far the largest of such studies is the Recurrent Coronary Prevention Project (Friedman et al. 1986) in which nearly 900 patients were randomized to either a prolonged program of Type A modification using cognitive behavioral techniques (see *Cognitive Behavioral Therapy*) or an alternative treatment emphasizing cardiological care and information. The behavioral intervention led to substantial reductions in Type A behavior and in recurrent myocardial infarction. There have been a number of successful small-scale replications of this study, but it does not seem to have been incorporated into clinical practice to a significant extent. Future studies may concentrate more on hostility than the complete range of Type A behavior, since it is generally held that hostility is the component of Type A most reliably related to cardiovascular disease. Depression has emerged much more recently than Type A as a risk factor for heart disease (and indeed other physical conditions) and appears to be finding widespread acceptance. While studies are ongoing, there are no completed studies of the effects of treating depression on recurrent heart disease. Frasure-Smith, one of the leading advocates of depression as a risk factor in recurrent heart disease, has described interventions for related states with disappointingly mixed effects. In an early study she found that distress (which could include both anxiety and depression) could be successfully treated in patients after a myocardial infarction and that this was associated with reductions in morbidity and mortality over the subsequent seven years. However in a substantial replication (Frasure-Smith et al. 1997) she was unable to either reduce distress or affect the recurrence rate.

3. *Morbidity and Mortality*

Few studies of rehabilitation have large enough samples to demonstrate reliable changes in morbidity or mortality, but meta-analytic techniques enable independent studies to be combined. Oldridge et al. (1988), in a highly influential analysis of exercise and life-style change programs, showed a reduction in mortality, but not reinfarction rate, following rehabilitation. Subsequent meta-analyses incorporating more studies of exercise have largely confirmed these findings and also shown effects on reinfarction, as have the meta-analyses of life-style change and stress management programs (Dusseldorp et al. 1999). They show that such programs were associated with a 34 percent reduction in mortality and a 29 percent reduction in reinfarction rate. The findings were most persuasive in studies in which there was evidence of change in life-style-related risk factors such as blood pressure, cholesterol, and cigarette smoking.

4. *The Future*

The outstanding issues for the future are effectiveness, efficiency, and access. While it is clear that effective programs exist, the results from ineffective trials are still reported and the critical components have yet to be isolated. In addition we can anticipate increasing efforts to incorporate modifications of the newer psychological risk factors such as depression, hostility and, very probably, social support. Many programs are expensive to implement, they require gymnasiums, physiotherapists and other health professionals, and

time—all scarce resources. Increasingly complex programs will have to be delivered in ever more efficient ways. This will surely involve close targeting of interventions to patients' particular problems and risks as well as more efficient methods of delivery such as patient manuals, both printed and computerized, and the use of the internet. Perhaps the greatest advances need to be made in increasing access. In many counties with otherwise sophisticated healthcare systems, too few people are offered cardiac rehabilitation, and many find it difficult to accept because of factors such as other commitments, poverty, and, no doubt, a failure to realize or accept the possible benefits. In England and Wales 40 percent of centers restrict rehabilitation to those aged below 60, and when it is offered to the elderly it is not geared to their special needs. The situation may be a little different in the USA, where Richardson et al. (2000) report an increase in the uptake of rehabilitation in those aged 65 and over, from 28.3 percent in 1986 to 52.1 percent in 1996 for one particular service. Cardiac rehabilitation in the elderly appears to be safe and of value (Foreman and Farquhar 2000), but Oldridge (1998) has pointed to the lack of good, controlled studies. Most studies, like most rehabilitation programs, have focused on the middle-aged. We can expect to see in the near future a change in both access and a strengthening of the scientific basis of cardiac rehabilitation for the elderly.

See also: Cardiovascular Conditioning: Neural Substrates; Chronic Illness: Quality of Life; Coronary Heart Disease (CHD), Coping with; Coronary Heart Disease (CHD): Psychosocial Aspects; Coronary-prone Behavior, Type A; Gender and Cardiovascular Health; Stress and Cardiac Response; Stress and Health Research

Bibliography

Dusseldorp E, van Elderen T, Maes S, Meulman J, Kraaij V 1999 A meta-analysis of psychoeductional programs for coronary heart disease patients. *Health Psychology* **18**: 506–19

Foreman D E, Farquhar W 2000 Cardiac rehabilitation and secondary prevention programs for elderly cardiac patients. *Clinics in Geriatric Medicine* **16**: 619–29

Frasure-Smith N, Lesperance F, Prince R, Verrier P, Garber R A, Juneau M, Wolfson C, Bourassa M G 1997 Randomised trial of home-based psychosocial nursing intervention for patients recovering from myocardial infarction. *Lancet* **360**: 473–9

Friedman M, Thoresen E, Gill J J, Ulmer O, Powell L H, Price V A, Brown B, Thompson L, Rabin D D, Breall W S, Bourg E, Levy E, Dixon T 1986 Alterations of Type A behavior and its effect on cardiac recurrences in post myocardial infraction patients: summary results of the recurrent coronary prevention project. *American Heart Journal* **112**: 653–65

Jones D A, West R R 1996 Psychological rehabilitation after myocardial infarction: multicentre randomised controlled trial. *British Medical Journal* **313**: 1517–21

Johnston M, Foulkes J, Johnston D W, Pollard B, Gudmunsdottir H 1999 Impact of patients and partners of inpatient and extended cardiac counseling and rehabilitation: a controlled trial. *Psychosomatic Medicine* **61**: 225–33

Krantz D S, Lundgren N R 1998 Cardiovascular disorders. In: Johnston D W, Johnston M (eds.) *Health Psychology, Volume 8 of Comprehensive Clinical Psychology*. Pergamon, Oxford, UK, pp. 217–36

Kugler J, Seelbach H, Krueskemper G M 1994 Effects of rehabilitation exercise programmes on anxiety and depression in coronary patients. *British Journal of Clinical Psychology* **33**: 401–10

Lewin B, Robertson I H, Cay E L, Irving J B, Campbell M 1992 Effects of self-help postmyocardial infarction training program on psychological adjustment and use of health services. *Lancet* **339**: 1036–40

Linden W, Stossel C, Maurice J 1996 Psychosocial interventions for patients with coronary artery disease—a meta-analysis. *Archives of Internal Medicine* **156**: 745–52

Oldridge N 1998 Cardiac rehabilitation in the elderly. *Aging* **10**: 273–83

Oldridge N B, Guyatt G H, Fischer M E, Rimm A A 1988 Cardiac rehabilitation after myocardial infraction: combined experience of randomized clinical trials. *Journal of the American Medical Association* **260**: 945–50

Richardson L A, Buckenmeyer P J, Bauman B D, Rosneck J S, Newman I, J 2000 Contemporary cardiac rehabilitation: patient characteristics and temporal trends over the past decade. *Journal of Cardiopulmonary Rehabilitation* **20**: 57–64

D. W. Johnston

Rehabilitation of the Offender

The ways in which societies react to those who break criminal laws show major historical changes. In most civilized countries, corporal punishment and the death penalty have been replaced by prison sentences, fines, and a broad range of diversion programs, alternative sanctions, and community treatment (Tonry 1998). The latter include, for example, restitution, offender–victim mediation, social training courses, community service, intensive supervision, electronic monitoring, boot camps, day treatment centers, and other alternatives to institutionalization and routine probation.

Although the scope of criminal justice interventions has broadened, the 'right way' forward for crime policy is still controversial in many countries. The respective opinions are related to general political attitudes. They emphasize different aims in reactions to criminal behavior, such as compensation by the guilty, safety of the public, deterrence of other people from similar crimes, and prevention of future law breaking in the offender. The last aim is central to offender rehabilitation and relates to similar concepts such as resocialization or correctional treatment. These concepts refer to educational, psychological, and medical measures that are delivered in a systematic

way to reduce the risk of recidivism and contribute to (re-)integration in society.

During the 1960s and early 1970s, this idea became a centerpiece of crime policy in many countries. However, the concept of rehabilitation was also criticized from different perspectives that emphasized the other aims mentioned above. In the mid-1970s, a comprehensive review of offender rehabilitation programs concluded that the majority of available studies were methodologically too weak to prove treatment success (Lipton et al. 1975). Although this and other reviews revealed various positive effects, the slogan of 'nothing works' in offender rehabilitation became widespread. However, the 1990s saw a re-emergence of the issue of offender rehabilitation. Among other reasons, this was due to systematic research integrations via meta-analyses that led to a more differentiated discussion in research and practice (McGuire 1995, Palmer 1992).

1. General Efficacy

There are now hundreds of evaluations of offender treatment that include an untreated comparison group. The outcomes differ somewhat depending on the age of offenders, previous criminality, offense types, modes of treatment, criteria of success, length of follow-up, and other issues. For example, various studies of sexual offender treatment revealed sexual reoffending rates between 3 percent and 44 percent with a mean of 18.8 percent (Hall 1995). The respective data for untreated control groups were 6 percent to 75 percent with a mean of 26.8 percent. Therefore, a single study cannot be generalized, and it is necessary to base research syntheses on common measures of effect size (ES). Such an indicator is the correlation coefficient *phi* or *r*. Table 1 summarizes the results of a number of recent meta-analyses on offender rehabilitation.

Although effects differ according to study samples and methods of computation, all mean ESs are positive. However, they are generally small. A realistic estimate of overall effectiveness is only approximately 0.10 or a little less. Taking the example of a control group with 50 percent reoffending, such an effect size of 0.10 would indicate a reduction to 40 percent recidivism. This is a 10 percentage points or 20 percent difference in comparison with the untreated group. Such a small effect may well be relevant in terms of cost utility. A realistic evaluation of correctional programs also has to take into account that the

Table 1
Selected meta-analyses on offender rehabilitation

Authors and topics of meta-analyses	Studies	ES (r)
Andrews et al. 1990		
Community/institutional programs for juvenile/adult offenders	154	0.10
Gendreau and Goggin 1996		
Community/institutional programs for juvenile/adult offenders	215	0.13
Gottschalk et al. 1987		
Community programs for juvenile offenders	101	0.12
Hall 1995		
Community/institutional programs for sexual offenders	12	0.12
Lipsey 1992		
Community/institutional programs for juvenile offenders	440	0.05/0.08
Lipsey and Wilson 1998		
Comm./institutional. programs for serious and violent juvenile offenders	200	0.06/0.09
Lösel 1995		
German therapeutic prisons for serious adult offenders	18	0.11
Redondo et al. 1999		
European comm./inst. programs for juvenile/adult offenders	32	0.12
Pearson et al. 1996		
Community/institutional programs for juvenile offenders	293	0.09
Pearson et al. 1996		
Community/institutional programs for adult offenders	211	0.07
Petrosino 1997		
Randomized studies on programs for juveuile/adult offenders	150	0.02
Polizzi et al. 1999		
Prison- and nonprison-based programs for sex offenders	13	0.09

optimum ES is not 1. Depending on the recidivism rate in the untreated group, 0.30 may already represent the maximum possible efficacy.

2. *Different Types of Program*

Most research reviews reveal significant differences between kinds of programs: the largest effects come from theoretically well-founded, multimodal, cognitive-behavioral, and skill-oriented programs. Also promising are family-oriented programs in ambulatory treatment for serious juvenile offenders and clearly structured therapeutic communities in the institutional treatment of adult offenders. Successful interventions fit the level of offenders' risk, address criminogenic needs precisely (instead of vague personality changes), and take into account the specific learning styles of the offenders (responsivity principle). Such programs are founded on empirical knowledge about the causes and development of the respective criminal behavior and are not just derived unspecifically from psychotherapeutic 'schools.' They address, for example, motivation for change, self-control, crime-related beliefs and attitudes, social skills, interpersonal problem-solving, self-critical thinking, moral reasoning, anger management, victim awareness, and coping with risk situations for relapse (e.g., Ross and Ross 1995). Cognitive-behavioral and relapse prevention approaches are not only most appropriate for the treatment of 'ordinary,' violent, and personality-disordered offenders but also for sexual offenders (Marshall et al. 1998). Here, combinations with hormonal measures seem to be relatively successful as well.

Effect sizes in the most appropriate programs can be twice as high as the average. In contrast, unstructured case work, traditional psychodynamic, and non-directive approaches show smaller effects than the overall mean. Relatively low-structured, self-governing, permissive therapeutic communities and milieu therapy also seem to range at the lower end of effectivity. The same holds for merely formal variations in punishment or probation; diversion without any educational or psychosocial component; deterrence, boot camps, and other measures of smarter punishment or intermediate sanctions. Aversive behavioral methods for sex offenders are also not successful. In some studies, the effects of inappropriate treatment are even negative.

3. *Methodological Issues*

Methodological features explain a substantial part of outcome variance. For example, studies with random assignment tend to produce smaller effects than quasi-experimental designs. There are also smaller effects in studies with larger samples, higher statistical power, and low rates of attrition between subject assignment and outcome measurement. Juridically-proven recidivism shows somewhat smaller effects than measures of institutional adaptation or attitude change. Outcome criteria with good reliability and longer follow-up periods also tend to be associated with smaller effects.

Although the number of evaluations with reasonable methodological quality has increased, there are still very few well-controlled studies on specific modes of treatment and specific offender groups. This is particularly the case for the institutional treatment of serious adult offenders, psychopaths, and sex offenders. In these cases, it can be rather difficult to form equivalent control groups that do not receive any kind of treatment. So far, evaluations often address different degrees or modes of treatment and not an all-or-none dichotomy.

4. *Program Integrity and Intensity*

A certain treatment label may be misleading with respect to what actually happens in a program. However, high integrity (reliability) of program delivery is necessary for good outcomes. Naturally, this holds only for programs that are successful in principle. Low integrity may result from, for example, weakly-structured programs with no treatment manuals, organizational barriers and staff resistance, or insufficient staff training and supervision. Although psychosocial interventions are basically an individual process and not a standardized technique, it is important to assess and monitor program implementation systematically. Doing this may be one reason why studies in which researchers were involved in program delivery show better effects than other evaluations.

Treatment intensity and dosage also relate to outcome. Lipsey and Wilson (1998), for example, found that the duration of treatment correlated positively with effect size, whereas the mean number of hours a week correlated negatively with outcome. The latter result may be due to problems of implementation and some satiation in very compact programs.

5. *Offender Characteristics*

The efficacy of rehabilitation programs depends not least on the characteristics of the offenders. For example, child molesters seem to benefit more from sex offender treatment than rapists. In juvenile offenders, Lipsey and Wilson (1998) found that samples with mixed prior offense types (including violent offenses) had better outcomes than those containing primarily property offenders. Effects were also larger when all offenders had prior sentences. These and other data suggest that serious and violent offenders may benefit more from treatment than less serious ones. This is plausible when considered in light of the risk principle: offenders with various prior offenses

are more often at risk of life-long persistent antisocial behavior. In contrast, offenders with fewer prior offenses more often belong to the type of adolescence-limited antisocial behavior. In these cases, natural protective factors in everyday life counteract recidivism in the untreated control groups as well.

However, the finding that high-risk groups benefit most from rehabilitation programs cannot be generalized to very high-risk cases such as psychopaths. These offenders lack the interpersonal and motivational characteristics that are essential for successful treatment, and thus rarely show a good outcome (Lösel 1998). This would indicate that a nonlinear, U-shaped function may be an adequate description of the relationship between risk and treatment efficacy. It is in particular the low-risk and very high-risk cases that show the weakest effects.

The need principle and responsivity principle can be applied adequately only when a systematic assessment of the offender is performed. This should be based not only on static variables such as offender age or criminal history but also on dynamic characteristics that are accessible to change. A thorough assessment is necessary for adequate placement decisions, for example, in the case of dangerous offenders. This should also reduce the number of treatment dropouts. Most studies have shown that the latter group is particularly at risk for reoffending.

6. *Setting Characteristics*

Overall, there seem to be larger effects in community programs than in institutional treatment. As a result, custody should be the last resort (for financial reasons as well). However, placement in a closed setting is normally confounded with a higher risk of reoffending. Sweeping assumptions that successful treatment is not possible within prison or other institutional settings are not supported by empirical research. Negative incarceration effects may depend on personal, situational, and organizational moderators that can be addressed at least partially by adequate programs. In cases of drug use and other dangerous lifestyles, institutions may even have some protective effects. The alternative therefore of 'community vs. institution' is too undifferentiated. The systematic assessment of risk can contribute to a stepwise reduction of custody to accompany positive developments in the offender (e.g., Quinsey et al. 1998).

7. *Organizational Climate and Regime*

Perhaps, it is not institutionalization by itself but the quality of regime and social climate that accounts for positive or negative effects. Although institutional features vary widely even under similar formal conditions, they have not yet been investigated to any great extent. A regime that is emotionally and socially responsive as well as structured, norm-oriented, and controlling is most promising. From process research on general psychotherapy, it is known that the relations between clients and therapists are very important for efficacy. Similarly, correctional programming requires well-selected, specifically trained, highly motivated, and continuously supervised staff. Attitudes and competencies that are not in line with the aims and content of a program may not only lower treatment integrity but even counterbalance potential efficacy.

8. *Aftercare and Relapse Prevention*

Even intensive rehabilitation programs make up only a small part of the factors that impact on the clients' thinking and behavior. It is therefore necessary to deliver services that consolidate the competencies acquired during treatment, to provide control over risks, and to cope with crises that may arise afterwards. Without such measures, treatment efficacy often can be demonstrated only during short follow-up periods. Although systematic programs of through-care, aftercare, and relapse prevention are particularly emphasized for the treatment of drug-addicted and sexual delinquents (Annis 1986, Laws 1999), they are relevant for other violent and personality-disordered offenders as well. Thorough planning and monitoring of relapse prevention refers to type and intensity of services, cooperation routines between the institutions, probation officers, and other staff involved. These multiagency networks are very important for preventing recidivism (Spencer 1999).

9. *Natural Protective Factors*

Even without any rehabilitation program, a subgroup of high-risk offenders desists from criminal behavior. In these cases, natural protective mechanisms make delinquent behavior less attractive. Although our knowledge about such processes is still rudimentary, various personal and social resources have repeatedly shown a protective function against criminality (see *Religiosity, Sociology of*; *Non-violence: Protective Factors*): for example, bonds to nondeviant family members or other reference persons; cognitive competencies, school achievement, and planning for the future; nondeviant peer groups; work motivation, vocational skills, and employment; or a supportive and prosocial intimate partner. Such protective factors are relevant for positive turning points in criminal careers. Therefore they need to be assessed and integrated into program planning and delivery.

10. *Community Features*

Individual and microsocial risk and protective factors also interact with the broader social context. Therefore community features such as the concentration of

poverty, unemployment, multiproblem families, violence, drug use, and other deficits in 'social capital' must be taken into account in offender rehabilitation. If, for example, a program promotes vocational skills, it may be particularly successful in communities with low unemployment rates (Sherman et al. 1997). In other contexts, the skills acquired will be less useful. Counterproductive influences may result from neighborhoods with high criminality and violence. In a community in which peers and neighbors model and reinforce deviant behavior, it is more difficult to preserve program effects than in a social context with less negative influences. Offender treatment and relapse prevention should therefore be integrated into broader community programs of crime prevention.

11. Conclusions and Perspectives

During the last decade, there has been growing evidence that some programs of offender rehabilitation work and others are promising. However, there is still a lack of methodologically sound evaluation, particularly of complex programs and multiple services for specific adult and serious offender groups and of the service delivery in everyday practice. More process data are needed on the content of program delivery, staff characteristics, and dimensions of the institutional regime. It is also necessary to assess control group conditions in a similar way. The dropout problem must be tackled more effectively. Adequate assessment procedures, program modules on motivation, and flexible procedures of case management are suggested here. Outcome evaluation should not rely primarily on one-shot studies but refer to a theoretically-based chain of various change measures, including longer follow-up periods. Last but not least, practice-oriented evaluations must place offender rehabilitation programs in their wider social context.

Although more systematic research is needed, empirical program evaluations and experiences from practice already suggest guidelines for effective rehabilitation. These include, for example:

(a) Realistic expectancies regarding possible efficacy.

(b) Theoretically sound concept (social learning, multimodal approach).

(c) Systematic assessment of offenders' risk and matching to service level.

(d) Appropriate targeting of specific criminogenic needs and not unspecific personality characteristics.

(e) Teaching of self-control, nondeviant thinking, and social skills.

(f) Consequence in applying reinforcement.

(g) Matching of program type, offender, and staff.

(h) Thorough selection, training, and supervision of staff.

(i) Development of a supportive institutional climate and regime.

(j) Neutralization of criminogenic social networks.

(k) Strengthening of 'natural' protective factors.

(l) Attainment of high program integrity and adequate intensity.

(m) Dynamic assessment and monitoring of offender change.

(n) Provision of systematic through-care, aftercare, and relapse prevention.

(o) Continous process and outcome evaluation of the program.

Countries such as Canada, England, and Scotland use similar guidelines for program accreditation. Naturally, we cannot expect that one specific type of program will emerge as the gold standard for successful offender rehabiliation. However, the current empirically-based knowledge on 'what works' is promising in securing service quality for both the society and the offender.

See also: Crime and Delinquency, Prevention of; Criminal Defenses; Criminal Justice, Ethics of: Aspects of Human Dignity; Criminal Law and Crime Policy; Deterrence; Deterrence: Legal Perspectives; Deterrence Theory: Crime; Program Evaluation; Relapse Prevention Training; Treatment of the Repetitive Criminal Sex Offender: United States

Bibliography

Andrews D A, Zinger I, Hoge R D, Bonta J, Gendreau P, Cullen F T 1990 Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology* **28**: 369–404

Bernfeld G, Farrington D P, Leschied A (eds.) 2001 *Offender Rehabilitation in Action: Implementing and Evaluating Effective Programs*. Wiley, Chichester, UK

Cullen E, Jones L, Woodward R (eds.) 1997 *Therapeutic Communties for Offenders*. Wiley, Chichester, UK

Gendreau P 1995 The principles of effective intervention with offenders. In: Harland A J (ed.) *Choosing Correctional Options that Work: Defining the Demand and Evaluating the Supply*. Sage, Thousand Oaks, CA

Gendreau P, Goggin C, Smith P 1999 The forgotton issue in effective correctional treatment: Program implementation. *International Journal of Offender Therapy and Comparative Criminology* **43**: 180–187

Hall G C N 1995 Sexual offender recidivism revisited: A meta-analysis of recent treatment studies. *Journal of Consulting and Clinical Psychology* **63**: 802–809

Hollin C (ed.) 2000 *Handbook of Offender Treatment and Assessment*. Wiley, Chichester, UK

Laws D R 1999 Relapse prevention: The state of the art. *Journal of Interpersonal Violence* **14**: 285–302

Lipsey M W 1992 Juvenile delinquency treatment: A meta-analytic inquiry into variability of effects. In: Cook T D, Cooper H, Cordray D S, Hartmann H, Hedges L V, Light L R, Louis T A, Mosteller F (eds.) *Meta-analysis for Explanation*. Russell Sage Foundation, New York

Lipsey M W, Wilson D B 1998 Effective intervention for serious juvenile offenders. In: Loeber R, Farrington D P (eds.) *Serious and Violent Juvenile Offenders*. Sage, Thousand Oaks, CA
Lipton D, Martinson R, Wilks J 1975 *The Effectiveness of Correctional Treatment*. Praeger, New York
Lösel F 1995 The efficacy of correctional treatment: A review and synthesis of meta-evaluations. In: McGuire J (ed.) *What Works: Reducing Reoffending*. Wiley, Chichester, UK
Lösel F 1998 Treatment and management of psychopaths. In: Cooke D J, Forth A E, Hare R D (eds.) *Psychopathy: Theory, Research and Implications for Society*. Kluwer, Dordrecht, NL
Marshall W L, Fernandez Y M, Hudson S M, Ward T (eds.) 1998 *Sourcebook of Treatment Programs for Sexual Offenders*. Plenum Press, New York
McGuire J (ed.) 1995 *What Works: Reducing Reoffending. Guidelines From Research and Practice*. Wiley, Chichester, UK
Palmer T 1992 *The Re-emergence of Correctional Intervention*. Sage, Newbury Park, CA
Quinsey V L, Harris G T, Rice M E, Cormier C A 1998 *Violent Offenders: Appraising and Managing Risk*. American Psychological Association, Washington, DC
Redondo S, Sánchez-Meca J, Garrido V 1999 The influence of treatment programmes on the recidivism of juvenile and adult offenders: An European meta-analytic review. *Psychology, Crime and Law* **5**: 251–278
Ross R R, Ross B (eds.) 1995 *Thinking Straight*. Cognitive Centre, Ottawa, Canada
Sherman L, Gottfredson D, MacKenzie D, Eck J, Reuter P, Bushway S 1997 *Preventing Crime: What Works, What Doesn't, What's Promising*. Report to the United States Congress: University of Maryland, US Department of Justice, Washington, DC
Spencer A P 1999 *Working With Sex Offenders in Prisons and Through Release to the Community*. Jessica Kingsley, London
Tonry M (ed.) 1998 *The Handbook of Crime and Punishment*. Oxford University Press, New York

F. Lösel

Reification: History of the Concept

Notwithstanding its esoteric overtones and philosophical intricacies, reification (*Verdinglichung*) is a central concept of social theory in general and of critical theory in particular. Although reification has received the greatest attention in Western Marxism, and above all in Lukács' *History and Class Consciousness* (1923), it is important not to restrict the use of the concept to that tradition but to see that it can also and already be found in the work of Kant, Hegel, Nietzsche, Dilthey, Husserl, Heidegger, Simmel, and Max Weber to criticize the dehumanizing, rationalizing, and alienating tendencies of modernity.

As a technical term, the term reification emerged in the English language in the 1860s out of the contraction of the verb *facere* (to make) and the substantive *res* (thing), which can refer both to concrete and empirically observable things (*ens*) and to abstract, indeterminate things (*aliquid*). As a synonym of 'thingification,' the inverse of personification, reification metaphorically refers to the transformation of human properties, relations, processes, actions, concepts, etc. into *res*, into things that act as pseudopersons, endowed with a life of their own. This material transmogrification of persons, relations, processes, concepts, etc. into thing-like entities that act like pseudopersons can operate both on a methodological and a social level. In both cases, however, the concept is usually used as a *Kampfwort* to denounce the 'violence of abstractions,' either of conceptual abstractions (Sohn-Rethel's *Denkabstraktionen*) that suppress the reflexive embeddedness of concepts into their social context, treat social facts as things, and transform metasubjects into megasubjects, or of real abstractions (Marx's *Realabstraktionen*) that strip individuals of their autonomy and reduce them to cogs of an abstract social machinery. This reference to the Kantian notion of autonomy is important. Outside of the normative context of an 'enlightened critique of the Enlightenment,' the notion of reification hardly makes sense, as can be gathered from the fact that conservative, skeptical, and ultraliberal theorists such as Gehlen, Luhmann, and Hayek have a theory of the autonomization of social structures but no theory of reification to connect it to the alienation of individuals in modern capitalist industrial societies.

1. Methodological Reification—or the Critique of Conceptual Abstractions

In the philosophy of the social sciences, the concept of reification is used (a) to denounce the hypostasis or substantialization of concepts (nominalist critique of reism), (b) the naturalization of the subject and the life-world (humanist critique of naturalism), and (c) the ideological justification of the status quo (dialectical critique of fetishism).

1.1 Nominalist Critique of Reism

In the case of the critique of reism (or naive conceptual realism), the notion of reification of concepts is used to denounce, from a nominalist, vitalist, or criticist perspective, the categorical error of transforming abstractions (notions, representations, concepts) into a material reality, in a concrete object 'out there.' Reification is here understood as a synonym of the 'fallacy of misplaced concreteness' (Whitehead). What is criticized is the hypostasis of concepts, analytical constructs, and ideal types, the sliding from the substantive to the substance, which involves a subreption of the categorical thing with the 'thing in itself.' This is, for instance, the case with macrosociologists who transform their own conceptual constructs or those of

the actors ('the State,' the *Bourgeoisie*, the 'Proletariat') into historical subjects capable of agency and of determining their own ends ('the State decides,' 'the Anglican Church resists,' 'the glorious Proletariat triumphs,' etc.). It should, however, be noted that due to the absence of a consensus on the ultimate referents of reality and the fact that one can always submit the concepts of the scientist to a neo-Kantian critique of ontology, the charge of reification is almost inevitable. Given that one's typification is another's reification, the critique of 'false conceptual realism' (Weber) is endemic in sociology.

1.2 Humanist Critique of Naturalism

The critique of positivist naturalism in terms of reification of the subject and the life-world is linked to the series of methodological disputes (*Methodenstreit*) which, since the double foundation of sociology by Comte and Dilthey in the nineteenth century, have opposed the partisans of the method of causal explanation (*Erklären*) to the partisans of the interpretative methods (*Verstehen*). Drawing on Vico's principle of the *verum factum* (*verum et factum convertuntur*), according to which we can understand the sociohistorical reality because it is a human product, but not nature which is a divine product, humanists claim that the appropriate method of sociology is interpretative in that it aims to understand, by means of a phenomenological and hermeneutic reconstitution of the meaning of action, the social-historical world (Hegel's objective spirit) as an objectivation of subjective actions. Social facts thus have a meaning and cannot be treated 'as if they were things' (Durkheim). The naturalistic elimination of the meaningfulness of action through statistical observation is reifying in that it transforms psychic acts into pseudophysical facts and reduces culture to (second) nature. Against Durkheim and his fellow 'factists' who 'change the subject' of the human sciences by substituting factors for actors, humanists thus argue that social facts are not things but that things are social facts whose meaning can be understood and which can be interpreted as an 'ongoing accomplishment of the concerted activities of daily life' (Garfinkel).

1.3 Dialectical Critique of Fetishism

The dialectical critique of fetishism offers a metatheoretical critique of the ideological implications of *bourgeois* theories and methodologies of the social that, due to a lack of reflexivity on their context of genesis and application, legitimize the status quo. Dialecticians accept the limits of 'hermeneutic idealism.' When social relations have crystallized into a 'second nature' and social subsystems follow their own pseudonatural laws, 'dehumanizing' theories (e.g., functionalism, structuralism, systems theory) and methods (e.g., linear modeling, statistical regression) can and have to be applied. But if one does not want to fall prey to a 'reification of the second order' and give a 'reified perception of the reifying' (Adorno) that willy-nilly endorses the reality it registers, the observed facts have to be 'mediated by the totality' (Lukács 1923) and defetishized in such a way that the tension between the real and the possible, between what is and what could or should be, becomes perceptible within the facts themselves.

2. Social Reification—or the Critique of Real Abstractions

In the tradition of Western Marxism, Marx's theories of alienation and the fetishism of commodities are combined with Hegel's dialectical phenomenology of spirit, Simmel's theory of the tragedy of culture, and Max Weber's theory of formal rationalization to form a critical theory of society. The concept of reification is used to refer to the relatively autonomous, alienating and alienated functioning of the social and cultural (sub-) systems of modern capitalist societies that impose their constraints from without on individuals, limit their freedom and tend to reduce them to powerless 'carriers' or passive 'executioners' of the system. As products of praxis, institutions and organizations are human objectivations, but in the course of their development, the cultural and social (sub)systems have been complexified, formalized, rationalized, and depersonalized to such an extent that eventually they have been transmuted into self-referentially closed systems that function independently of the will and the consciousness of individuals, thwart their plans, threaten their autonomy, and perhaps even their existence. The critique of reification is dialectical and thus somewhat paradoxical: the insistence on the alienating autonomy of the system, which is an objectivation of action, aims to reactivate the autonomy of the individuals and to overcome their alienation.

Although the concept of reification (*Verdinglichung*) can already be found in Hegel's *Philosophy of Right*, the real history of the concept begins with Marx and with Lukács's Hegelian interpretation of Marx. The origins of the theory of reification are usually found just where the word itself is absent, namely in the famous section on the fetishism of commodities (Marx 1869, Chap. 1, Sect. 4). Analyzing capitalism as a system of generalized exchange, Marx notes that the commodity has become the universal form of the product of labor, with the result that the exchange value of the commodity supplants the use value. Consequently, the exchange value appears to those

who exchange goods as a property of the commodity itself, whereas in reality it is the result of the labor that is incorporated in the commodity and that expresses itself as a quantitative relationship between the exchanged goods. 'It is nothing but the definite social relation between men themselves which assumes here, for them, the phantasmagoric form of a relation between things' (Marx 1869, pp. 23, 86). This inversion of humans and things is not simply an illusion, however, but an expression of the real nature of social relations in a competitive market environment. In the absence of a central organism that regulates both the production and the distribution of the products of labor, the social integration of humans is imposed from without by the systemic interconnection of things.

In 'Reification and the Consciousness of the Proletariat,' the central chapter of *History and Class Consciousness*, Lukács, a Hegelian Marxist who was once a student of Simmel and Max Weber, presents the classic formulation of the theory of reification. Synthesizing Weber's theory of formal rationalization with Marx's theory of commodity fetishism, Lukács generalizes the theory of commodity fetishism beyond the sphere of circulation. In the problem of fetishism, which he immediately identifies with the phenomenon of reification, he discovers the 'central, structural problem of capitalist societies in all its aspects' (Lukács 1923/1968, p. 257). The universality of the commodity form, conceived as the prototype of all the forms of objectivity that follow their own laws and dissimulate the traces of human relations that subtend them, affects the life of everybody, both in its objective and subjective manifestations. Objectively, individuals are confronted with a second nature of pseudo-things against which they are powerless; subjectively, they are estranged from their own activity, apprehending the products of their own activity in an alienated mode—'as if they were something else than human products' (Berger and Luckmann). Moving from the sphere of circulation to the sphere of production, Lukács rediscovers the theory of the alienation of labor which the young Marx had developed but not published in the *Parisian Manuscripts* of 1844 (see *Alienation, Sociology of*). In the sphere of material production, reification expresses itself most clearly in the reduction of labor power to a commodity and of the laborer to an appendix of the machine. In capitalism, reification is generalized and the fate of the worker becomes paradigmatic of the fate of everyone. Expressing the Messianism of the oppressed, Lukács eventually reintroduces the Proletariat as the 'identical subject–object' of history whose revolutionary actions overcome alienation and reification and thus realize the Hegelian dream of the restoration of the 'beautiful totality.'

The development of the so-called Critical Theory of the Frankfurt School can best be understood as the result of a progressive disillusion with revolutionary expectations. Eliminating the Hegelian-Marxist dialectic of consciousness, which they replace by a Freudian account of sublimation and repression, Horkheimer, Marcuse, and especially Adorno, who also gave a Nietzschean twist to the concept of reification, radicalize the Weberian-Marxist strand in Lukács' theory. Universalizing and totalizing reification to the point that it appears as an ontological feature of human civilisation, they almost end up indicting Reason as such. Indeed, to explain totalitarianism, Horkheimer and Adorno develop a negative philosophy of history which uncovers in the first protohistorical attempts to dominate nature the origin of the fatal unfolding of a diabolic logic of increasing reification that will find its culmination (but not its endpoint) in the death camps. In his *Theory of Communicative Action*, Jürgen Habermas (1981), the main representative of the second generation of critical theory, reformulates the theory of reification in terms of the paradigm of language. In this perspective, reification is no longer associated with rationalization as such, as was the case with Max Weber and the Frankfurt School, but reconceptualized in terms of the 'colonization of the life-world' by the subsystems of the economy and the administration. When the mechanisms of systemic integration (money and power) force back the forms of social integration from those domains that can only be integrated through language, a reification ensues that leads to a pathological deformation of the life-world.

Bibliography

Arato A 1974 The neo-idealist defense of subjectivity. *Telos* **21**: 108–61

Arato A, Breines P 1979 *The Young Lukács and the Origins of Western Marxism*. Seabury Press, New York

Berger P, Pullberg S 1965 Reification and the sociological critique of consciousness. *History and Theory* **4**(2): 196–211

Gabel J 1962 *La fausse conscience. Essai sur la réification*. Editions de Minuit, Paris

Habermas J 1981 *Theorie des kommunikativen Handelns*, Vols. 1 and 2. Suhrkamp, Frankfurt, Germany (1984 *Theory of Communicative Action*, Vol. 1. Heinemann, London, 1987, Vol. 2. Polity Press, Cambridge, UK)

Lukács G 1923/1968 Geschichte und klassenbewusstsein. In: (eds.) *Werke*, Band 2. Neuwied, Luchterhand (1971 *History and Class Consciousness*. Merlin, London)

Marx K 1869/1972 Das Kapital, Vol. I. In: (eds.) *Marx Engels Werke*, Vol. 23. Dietz Verlag, Berlin (1974 *Capital*, Vol. 1, Lawrence & Wishart, London)

Pitkin H 1987 Rethinking reification. *Theory and Society* **16**(2): 263–93

Rose G 1978 *The Melancholy Science: An Introduction to the Thought of Theodor W. Adorno*. Macmillan, London

Thomason B C 1982 *Making Sense of Reification: Alfred Schutz and Constructionist Theory*. Humanities Press, Atlantic Highlands, NJ

Vandenberghe F 1997–1998 *Une histoire critique de la sociologie allemande. Aliénation et réification.* 2 vols. Editions la Découverte, Paris (2001 *Critique of Reification.* University of Minnesota Press, Minneapolis, MN)

F. Vandenberghe

Reinforcement: Neurochemical Substrates

1. The Operant–Respondent Distinction

The neurochemical mechanisms that mediate reinforced or operant behavior may differ in a fundamental way from those underlying reflexes or respondent behavior. This is because environmental stimuli appear to control the two classes of behavior in fundamentally different ways. In reflexes, whether conditioned or unconditioned, the controlling stimulus precedes the response and elicits it. In operant conditioning, the controlling stimulus follows the response and elevates its subsequent probability. When the controlling stimulus precedes the response, information flow in the brain is afferent to efferent, as in the conventional reflex arc. On the other hand, when the controlling stimulus follows the response, as in reinforced behavior, the underlying brain organization seems to require an unconventional circuitry in which efferents are activated before afferents. However, the mechanisms for reinforced behavior do not require circuits that directly link efferent to afferent elements. This is because operant behaviors do not directly activate the goal-detecting afferent systems. Rather, the correct response operates on the environment to produce the goal object and it is this environmental change that activates the goal-detecting systems. Thus, although the reinforcement mechanism does not require efferent-to-afferent circuitry, it must recognize efferent–afferent contingencies and must be operated selectively by them, i.e., it must cause behavioral reinforcement only when the neuronal substrates of the correct response and goal object, in that order, are activated sequentially.

2. Characterizing the Brain Reinforcement Mechanism

The problem of characterizing the brain reinforcement mechanism has two main parts: the first is to identify the neuronal substrate that performs the reinforcing function (reinforcing substrate) and the second is to identify the neuronal substrate that is modified by the reinforcement process (target substrate). Since any goal object can reinforce any behavior in an animal's repertoire, it seems likely that there is (a) convergence of goal-object input to the reinforcing substrate, and (b) divergence of reinforcing output from this substrate to the target substrate. Here we briefly summarize the substantial published literature that has been devoted to the identification of the reinforcing substrate. Until very recently, little consideration has been given to the question of the target substrate, but it may be useful to review some initial findings and to consider the directions that research on this problem may take.

2.1 Neurochemical Specialization

The idea that reinforcing functions are specialized neurochemically has guided research in this field for more than 30 years. The hypothesis that certain dopamine and opioid peptide brain cells may serve as reinforcing neurons is supported most directly by evidence from brain self-stimulation and drug self-administration experiments.

2.2 Electrical Self-stimulation of the Brain

In self-stimulation experiments, animals work to deliver electrical stimulation to their own brains through permanently indwelling electrodes. In the absence of other sources of reward, the reinforcement for self-stimulation behavior must arise from the neuronal activity that is excited by the electrical stimulus. If so, it would be logical to assume that some of the neurons under the electrode tip actually are the reinforcing neurons that mediate the effects of natural reinforcers or at least are neurons that directly excite them.

High self-stimulation rates are observed when electrodes are implanted in regions containing dopamine (or opioid peptide) cell bodies or pathways. In particular, self-stimulation tightly overlaps the distribution of dopamine cells in the ventral tegmentum and substantia nigra. Self-stimulation closely follows the anteriorly projecting dopamine fibers through the hypothalamus, but it correlates somewhat less closely with the dopamine terminal fields in the forebrain. The involvement of norepinephrine neurons in self-stimulation is more controversial. Although many laboratories report self-stimulation from sites in the vicinity of the locus coeruleus, it has not been possible to establish convincingly that the noradrenergic neurons that make up this nucleus are responsible for the reinforcing effect. Mapping of opioid peptide sites for self-stimulation is consistent with the idea that certain beta-endorphin and dynorphin neurons are involved in self-stimulation, but these studies are still in an early stage. The dopamine-opioid peptide reinforcement hypothesis also is supported by pharmacologic experiments. Antagonists of dopamine and opioid peptides, such as haloperidol and naloxone respectively, should block chemical transmission of reinforcement messages. In support of the model, there are many

reports that these drugs selectively block self-stimulation.

Interestingly, the pharmacology of avoidance behavior (a prototypical example of negative reinforcement) closely overlaps that of self-stimulation (exemplifying positive reinforcement). In particular, avoidance and self-stimulation behaviors are selectively facilitated by dopamine agonists, such as amphetamine, and both are selectively inhibited by dopamine antagonists, such as chlorpromazine. Such evidence suggests that a common dopamine substrate may underlie both positive and negative reinforcement processes.

2.3 Intravenous Drug Self-administration

In self-administration experiments, behavior is reinforced by systemic or central injections of neurotransmitters or drugs. Although thousands of chemical substances are available, animals and humans avidly self-administer only a few. These self-administered substances may properly be termed pharmacologic reinforcers and it is interesting to ask why these chemicals are selectively associated with behavioral reinforcement. It is no coincidence that most powerful pharmacologic reinforcers have the ability to mimic or release the hypothesized natural reinforcement transmitters. Thus, many dopamine and opioid receptor activators are known to support self-administration behavior. In particular, selective agonists acting at μ and δ opioid receptors function as positive reinforcers, as do the naturally occurring opioid peptides, β-endorphin and dynorphin. Similarly, agonists acting at dopamine D_1-like and D_2-like receptors, such as apomorphine (D_1/D_2), piribedil (D_2), and SKF 82958 (D_1) are self-administered. The patterns of self-administration for many of these receptor agonists (particularly, D_1 and μ agonists) resemble the patterns observed when highly addictive stimulant drugs (e.g., cocaine and amphetamine) or opiate drugs (e.g., morphine and heroin) are self-administered. Hence it is reasonable to suppose that addiction to stimulant or opiate drugs may depend largely on their ability to activate or to enhance the activity of natural dopamine or opioid peptide reinforcement systems respectively. Finally, dopamine and opioid peptide antagonists, such as chlorpromazine and naloxone, not only appropriately block the reinforcing effects of agents that activate dopamine or endorphin receptors, but also appropriately block the reinforcing effects of stimulant and opiate drugs. Serotonergic agents generally are not self-administered but serotonin agonists and antagonists have been shown to alter self-administration of psychostimulants such as cocaine.

The self-administration method also has been used in attempts to identify the sites of action of stimulants and opiates. In some studies, animals are trained to self-inject cocaine or morphine directly into the brain in an attempt to find preferred sites for reinforcing effects. In other studies, dopamine or opioid receptor antagonists are injected into different brain regions in attempts to block intravenous self-administration of drugs of abuse. Results from several laboratories suggest that important sites of action for both stimulant and opiate drugs are concentrated in the 'extended amygdala' including brain regions on an axis from the nucleus accumbens to the central amygdala. Reinforcing actions also are reported after direct injections of cocaine in the prefrontal cortex, morphine in the ventral tegmental area, and dynorphin in the hippocampus. Evidently, target sites for the reinforcing actions of drugs are widely distributed in the brain.

A closer analysis of brain transmitter involvement in reinforcement has been accomplished using *in vivo* microdialysis to monitor neurotransmitter levels in various brain areas during drug self-administration. Each cocaine self-injection caused a large increase in dopamine levels in the nucleus accumbens and these dopamine levels usually fell substantially to a 'trigger' concentration before the animal delivered the next self-injection. The addition of heroin to the cocaine solution caused synergistic increases in nucleus accumbens dopamine levels that are consistent with the higher abuse potential of such 'speed ball' combinations.

3. Methodological Considerations

Although these anatomical and pharmacological findings are generally consistent with the idea that reinforcing functions are performed by specialized systems of dopamine and opioid peptide neurons, some important problems remain. Administration of pure substances in the self-administration experiments strictly controls the chemical nature of the reinforcing injections but because the distribution of injected transmitters to active sites cannot exactly duplicate that of naturally released transmitters, the ensuing pattern of receptor activation could be artifactual or misleading. On the other hand, while electrical activation of transmitter pathways during self-stimulation presumably releases the reinforcing chemical messenger in a relatively natural distribution at appropriate postsynaptic sites, the electrical stimulus also must cause the simultaneous release of many irrelevant transmitters and neurohormones, including some that are still unknown. Thus, in self-stimulation, identification of which cells and which transmitters are relevant to the reinforcement process is largely a matter of inference. Solutions based on mapping studies demonstrate self-stimulation in anatomically coherent systems and those based on pharmacological studies implicate specific neurotransmitters. However, the anatomical mapping data from most self-stimulation experiments are insufficiently detailed. At

the same time, many of the pharmacological reports can be criticized because they often fail to distinguish specific drug effects on reinforcement from nonspecific effects on motor performance or attention. In addition, some self-stimulation findings seem to contradict the dopamine theory of reinforcement. For example, lesions of the nigrostriatal bundle do not permanently eliminate substantia nigra self-stimulation and, if only motorically simple responses are required, hypothalamic self-stimulation survives extirpation of all forebrain dopamine terminals.

4. *The Dopamine-opioid Peptide Theory of Reinforcement*

Despite such difficulties, the dopamine-opioid peptide theory of reinforcement currently enjoys a widespread acceptance. First, many investigators are impressed by the powerful and reliable reinforcing actions that highly localized electrical brain stimulation and specific chemicals can produce. Second, investigators are impressed by the consistent picture of dopamine-opioid peptide involvement in reinforcement processes that has emerged from the self-stimulation and self-administration data. Other neurochemical systems, in addition to dopamine and opioid peptides, also are likely to participate in reinforcement functions. For example, brain cannabinoid and nicotine receptors are probably involved in the reinforcing actions of marijuana and nicotine respectively. As additional receptors are identified as candidates for reinforcing actions, questions arise regarding their primacy and specific roles in reinforcement. Evidence now exists, for example, that several transmitter systems (e.g., cannabinoid, nicotine) have receptors localized on dopaminergic neurons that in turn may represent the final common pathway for reward. Another indication of a final common pathway utilizing dopamine is that agents that directly stimulate dopamine systems (e.g., amphetamine or cocaine) are more robustly self-administered than are agents that indirectly stimulate dopamine neurons (e.g., marijuana or nicotine).

Although a final common dopamine reward pathway remains a possibility, the crucial importance of reinforcement for survival makes it more likely that redundant neurotransmitter systems are involved. Most research on reinforcement has focused on the familiar transmitter systems mentioned above but complete understanding of reinforcement neurochemistry must await the likely discovery of novel transmitters. For example, several hundred G-protein coupled receptors have been cloned but their endogenous transmitters have not yet been identified. Some of these 'orphan' receptors may well play a role in reinforcement and their endogenous ligands (when elucidated) will serve as novel reinforcement transmitters. Such discoveries will lead to a more subtle and comprehensive understanding of reinforcement.

5. *Identification of the Target Substrate*

As noted, little theoretical or experimental effort has been devoted to the identification of the substrate that is modified by the reinforcement process. What are the neuronal targets of the reinforcing neurons? Because it is operant behavior that is reinforced, it is plausible to assume that behavioral substrates are major targets. It is commonly believed that operant conditioning involves the strengthening or reorganization of complex (but yet to be identified) neuronal circuits, which presumably form the substrates of reinforceable responses. According to this conventional view, the whole response is the unit of reinforcement. Where in the rat's brain, for example, would one find the circuit for a lever-press response? How might this circuitry be reorganized when lever pressing is reinforced? These anatomical and physiological considerations, as well as difficulties at the behavioral level associated with response definition, have led some scientists to question the idea that elaborate neuronal circuits are the functional units for reinforcement. According to B. F. Skinner, a more useful conceptual scheme recognizes the ultimately continuous nature of operant behavior. In this scheme, behavior is construed—not as a series of discrete acts—but rather as a continuous flow of activity and interaction with the environment, which arises from the largely spontaneous activation of infinitesimal behavioral elements. It is these elements, or 'behavioral atoms,' and not whole responses, that are the units strengthened by reinforcement. If so, and if atoms of behavior can be represented by the activity of individual neurons, then it could be argued that it is the behavior of individual neurons, and not of neuronal circuits, that is directly modified by reinforcing signals.

The cellular reinforcement hypothesis is currently under test in our laboratories by use of a novel '*in vitro* reinforcement' paradigm. This method employs training procedures closely analogous to those of behavioral operant conditioning to answer the question: Can the activity of relatively isolated brain cells in hippocampal slices be reinforced by local micropressure injections of dopamine or other reward-related transmitters or drugs? It was found that the bursting responses of individual pyramidal neurons in the CA I layer were in fact progressively increased by burst-contingent applications of dopaminergic and cannabinoid receptor agonists. The bursting responses of CA3 units were similarly increased by opioid receptor agonists. The same microinjections, administered independently of cellular activity, failed to facilitate and frequently suppressed cellular bursting. This observation, and the fact that glutamate (an excitatory transmitter that is not associated with the behaviorally-reinforcing effects of drugs) failed to increase bursting, indicated that general stimulation of cellular activity is an unlikely explanation of the facilitatory action of the burst-contingent injections.

Reinforcement delays exceeding 200 ms largely eliminated the reinforcing action of dopaminergic agonists. These findings support the conclusion that the behaviorally-reinforcing effects of dopaminergic, opioid, and cannabinoid drugs can be modeled *in vitro* at the cellular level.

See also: Amygdala (Amygdaloid Complex); Anxiety and Fear, Neural Basis of; Autonomic Classical and Operant Conditioning; Avoidance Learning and Escape Learning; Brain Stimulation Reward; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Drug Addiction; Drug Addiction: Sociological Aspects; Drugs and Behavior, Psychiatry of; Memory: Synaptic Mechanisms; Motivation, Neural Basis of; Reinforcement, Principle of

Bibliography

Gardner E L 1997 Brain Reward Mechanisms. In: Lowinson J H, Ruiz P, Millman R B, Langrod J G (eds.) *Substance Abuse: A Comprehensive Textbook*. Williams & Wilkins, Baltimore, MD, pp. 51–85

Klopf A H 1982 *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*. Hemisphere Press, New York

Self D W, Stein L 1992 Receptor subtypes in opioid and stimulant reward. *Pharmacology Toxicology* **70**: 87–94

Skinner B F 1953 *Science and Human Behavior*. The Free Press, New York

Stein L 1978 Reward transmitters: Catecholamines and opioid peptides. In: Lipton M A, DiMascio A, Killam K I (eds.) *Psychopharmacology: A Generation of Progress*. Raven Press, New York, pp. 569–81

Stein L, Xue B C, Belluzzi J D 1994 *In vitro* reinforcement of hippocampal bursting: A search for Skinner's atoms of behavior. *Journal of the Experimental Analysis of Behavior* **61**: 155–68

Wise R A 1996 Addictive drugs and brain stimulation reward. *Annual Review of Neuroscience* **19**: 319–40

J. D. Belluzzi and L. Stein

Reinforcement, Principle of

Reinforcement refers to an increase in the probability of a behavior when the behavior results in a positive consequence. That behavior is determined by its consequences is one of the most important principles within behavior therapy (see *Behavior Therapy: Psychological Perspectives*) for understanding the origins, maintenance, and change of clinically-relevant activities. 'Reinforcement' can denote (a) a principle of learning or behavior regulation; (b) a procedure designed to modify the frequency of an approved behavior (the verbal form 'to reinforce' referring to the contingent delivery of a positive stimulus); or (c) the consequence itself (also 'a reinforcer'). The term connotes 'strengthening' and what is strengthened is the likelihood of a behavior reoccurring in a specific context.

There is circularity in such a statement, as a stimulus event is judged to be a positive reinforcement to the extent that it actually increases the likelihood of the behavior that produced it. Thus, reinforcement is not quite the same as reward, since a reinforcer is defined in terms of what 'works' or functions to strengthen behavior, whereas a reward is identified by what an individual expresses a desire for, or enjoys. In practice, however, one selects reinforcements on the basis of prior experience, observation, and intuition that the planned consequence will be desired or enjoyable. When an organism will engage in a behavior that produces a consequence (food, praise, stimulation, or even apparently negative outcomes such as fatigue or annoying someone else) it is referred to as positive reinforcement; when an organism works actively to reduce or eliminate events currently in existence (such as cold, teacher demands, loneliness, boring activities), it is known as negative reinforcement.

Any previously neutral stimulus that signals the onset or further probability of a reinforcer comes to have reinforcing properties of its own, and is described as a secondary or conditioned reinforcer. Socially, money is the classic example. In applied contexts, a stimulus that has no value other than a designated rate of exchange with something desired is called a token reinforcement—often these are symbols such as poker chips or points that can later be exchanged for privileges.

1. History of the Concept

Reinforcement is closely linked to evolutionary principles, in that it represents adaptive functioning with obvious survival implications, and applies to all species capable of benefitting from experience. Its origin as a formal principle of modern psychology derives from Thorndike's (1927) Law of Effect: the occurrence of a 'satisfying' event 'stamps in' the behavior that was instrumental in producing it. Hull's (1943) influential learning theory proposed that the increase is in the strength of the association between a prior stimulus (cue or context) and the response, with the mechanism underlying reinforcement being the reduction in a basic biological drive. However, many stimuli can serve as reinforcers when they have only incentive value (such as a sweet but non-nutritive substance).

Much past research has been devoted to the question of whether reinforcement is necessary for learning to occur; the present consensus is that it is not. The question, however, may be a spurious one, since it is difficult to imagine a behavior that has no conse-

quences, and the positive nature of the outcome might be as simple as reducing constraints on the individual, or predicting future pleasures.

In the classical conditioning paradigm (see *Classical Conditioning and Clinical Psychology*), Pavlov also used the term to refer to the 'unconditioned stimulus' (UCS: the stimulus that evokes the reflexive response). When this stimulus is regularly paired with (predicted by) a neutral stimulus, the latter comes to evoke a somewhat similar response. The UCS is thus considered reinforcement, since its occurrence increases the tendency to emit a response in the presence of a previously neutral stimulus. Of considerable theoretical and practical importance is the phenomenon that when the UCS is removed and the previously neutral stimulus presented repeatedly, the conditioned response eventually ceases to occur. This phenomenon, as well as the procedure of removing the UCS, is known as extinction. It has an exact parallel in instrumental conditioning: if the overt behavior no longer produces the positive outcome, its probability decreases. This process is not simply the opposite of reinforcement; it is not unlearning or forgetting, but learning to do something else.

Although central to both conventional instrumental and classical learning situations, the concept of reinforcement has acquired much of its practical, applied implication from Skinner's (1969) paradigm of operant conditioning (see *Operant Conditioning and Clinical Psychology*). The typical laboratory procedure is for a specific 'operant' response, such as the pressing of a lever or push button, to result in the automatic delivery of a small amount of food for a hungry animal. The operants studied have been discrete, simple responses that can be repeated easily by the subject. A vast body of experimental research has demonstrated that the form, intensity, and temporal pattern (rate) of responding is a function of the rule specifying the relationship between the operant response, its consequence, and the context in which the consequence occurs (referred to as the 'three-term contingency').

2. *Contingency Principles*

The effectiveness of this procedure for developing or maintaining a specified operant is influenced by the immediacy and predictability with which the reinforcement is delivered. Although it will slow the acquisition of new responses, the reinforcement need not be delivered after every response, an arrangement called a 'partial schedule of reinforcement.' The most common are ratio schedules, in which reinforcement is delivered after a specified number of responses; and interval schedules, in which reinforcement is delivered to the first response occurring after a predetermined period of time has elapsed. In the laboratory, these different schedules of reinforcement produce characteristic patterns of response rate. Analogies are often made with workplace contingencies; for example, if one is paid on a piecework basis it is a ratio schedule, whereas being paid every month is an interval schedule of reinforcement. Reinforcement can also be delivered after a variable number of responses or periods of time—gambling on a slot machine represents a variable ratio schedule.

Of greater importance for application is that partial schedules of reinforcement result in a behavior being highly resistant to extinction, such that the subject will persist in the behavior for long periods of time, even when the reinforcement has been withdrawn (extinction). In clinical situations one might think that a behavior is maladaptive as it is apparently not being reinforced and should therefore extinguish. If the individual's history of reinforcement were known, however, it would be recognized that this behavior may have been reinforced sporadically in the past.

The precise form of any behavior can be determined by differential reinforcement, in which the criteria for exactly what response will result in reinforcement are carefully specified. By changing or narrowing these criteria gradually, a behavior can be 'shaped' into a specified configuration. Similarly, it is quite common in practice to reinforce differentially decreases in an undesirable behavior, or reinforce differentially alternative, more appropriate behaviors.

3. *Limitations in Generalizing from the Laboratory*

Despite the substantial body of knowledge derived from studies of animal behavior, the analogy between the operant laboratory and real life has limitations. Laboratory contingencies are contrived—unlike foraging in the wild, there is no intrinsic relationship between a lever press and obtaining food. There are many comparable artificial contingencies operating in everyday life, particularly work settings, in which a person is reinforced with money, food, or favors, for activities desired by the individual controlling these reinforcers. But more commonly our activities are intrinsically as well as socially rewarding, for instance, reading a book is reinforced by enjoyment of the story as well as by teacher approval, or climbing a mountain may be reinforced by a sense of accomplishment in addition to the admiration of friends.

It has long been debated that if an activity which is usually reinforced intrinsically is placed under an artificial external contingency (such as a monetary payment), this will later interfere with the intrinsic reward, such that once the artificial reward is withdrawn, the frequency of the behavior will drop below its previously (intrinsically) reinforced rate (see Carton and Nowicki 1998 for a discussion). There is also

considerable overlap in practice between the concept of reinforcement and that of feedback—providing people with information regarding their task performance. Feedback too can impair performance, regardless of whether it provides information about success or failure, by directing attention to self rather than the task at hand.

Another major difference between reinforcement studied in the laboratory and the principles affecting human social behavior is that the verbal specification of a contingency can function much like an actual contingency. Thus, for example, people can be promised a reward for an activity and this will change their behavior. Skinner referred to this as 'rule governed behavior,' whereas more cognitively oriented theorists prefer to talk of expectancies. Bandura (1977), for example, argued that performance on a task depends partly on the expectation that a certain reward will be forthcoming (outcome expectancy). Another way an individual can learn about a potential contingency between a behavior and a reinforcer is to observe someone else being reinforced for an activity; this is known as vicarious reinforcement.

4. *Motivation and Individual Differences*

Most general behavior theories argue there is a link between reinforcement and emotion, for example, Staats' (1996) A-R-D model, which proposes that stimuli serving as reinforcers (R) have simultaneous attitudinal (A) value (can elicit emotion) and discriminative stimulus (D) functions (can serve as a cue for approach). As active approach behavior and reward systems are linked in terms of neural processes, there are plausible biological mechanism for personality differences in the degree to which a given individual is responsive to positive outcomes. Relative sensitivity to reward over punishment may also be a characteristic of individuals exhibiting certain clinical syndromes, such as Attention Deficit Hyperactivity Disorder (ADHD), conduct disorder, and recent onset schizophrenia. There are also individual differences in preference for a small but immediate reinforcer as opposed to a larger, but more delayed reinforcer, a pattern that is generally considered to be maladaptive and detrimental to individual success. Humans are more influenced by reinforcers they have chosen than those arbitrarily selected by someone else. When environmental conditions deprive an organism of the opportunity to perform a particular behavior that would be engaged in a free choice situation, it will perform other activities in order to gain access to that behavior, a phenomenon sometimes referred to as the Premack (1965) principle.

Where reinforcement is artificial, such as a socially determined prize for a job well done, humans learn from prior experience the standard of performance that deserves to be rewarded. Even young children are unlikely to appreciate a reward given for performance that is below their own standard. Thus, although self-reinforcement is often included as a treatment component for enhancing self-control, it is unlikely to be influential over and above the intrinsic satisfaction of accomplishing one's own goals. Furthermore, if the outcomes achieved after much effortful behavior are less positive than expected, people tend to reevaluate the outcome in an upward direction (Lawrence and Festinger 1962).

5. *Reinforcement as Behavioral Regulation*

Much of current reinforcement theory in the operant tradition is concerned not with understanding the motivational features of reinforcement, but with predicting the effect on the distribution of available activities of different conditions of reinforcement. Neobehavioral theories that relate reinforcement to motivation (e.g., need reduction), have given way to economic-type theories that consider the sum total of potential behaviors in a situation as well as the sum total of reinforcements. One of the most important observations underlying this approach is that if there is more than one operant being reinforced (concurrent operant schedules), animals will distribute their responses more or less in proportion to the amount of reinforcement available to each one. This is known as Herrnstein's matching law (see Noll 1995 for a discussion).

In answer to the apparent anomalies discussed in Section 3, reinforcement is depicted as a homeostatic principle, referring to how an organism must adjust its actions to meet the demands of the environment. An organism will normally distribute its activities so as to minimize the demands (e.g., effortfulness of the response) and maximize opportunities, and this optimum—when there is unconstrained choice—is known as the behavioral 'bliss point.' Once reinforcement is seen in this light, it gives credence to applied approaches in which there is an emphasis on enhancing choices and opportunities for individuals in otherwise deprived environments, or those with limited responses in their repertoire.

Theories emphasizing behavioral regulation propose that contingencies serve to constrain the organism's free flow of behavior. Given the wide range of behavioral choices available to individuals in natural situations, it is logical that removing a reinforcement for one behavior will not be successful in reducing this behavior unless another, more socially desirable, behavior is able to be reinforced. Clinical research has repeatedly demonstrated the value of reinforcing more appropriate alternatives. Since the systems or economic model emphasizes that increases in one behavior must inevitably be accompanied by decreases in

others, extinguishing undesirable behavior and reinforcing appropriate responses may be two sides of the same coin.

6. Conclusions

While context/behavior/outcome contingencies are essential elements in the shaping and development of new skills, the reinforcing event itself has complex influences on human behavior and often serves as a cue initiating other activities. The effects of reinforcement are relative—related to the levels of reinforcement for other possible behaviors—which has important implications for designing environments that will support socially desirable activities. That there are motivational and emotional properties of reinforcement should not be underestimated, such as experiencing success, the relief of reduction in anxiety, or the satisfaction of a craving. There are also cognitive properties, with the history of availability being judged by the individual and influencing expectancies regarding the relative outcomes of different actions. Past experience with certain types of contingencies and deprivation experiences influence the way an individual will respond to a given reinforcement contingency. The relatively simple idea that all behavior has a function—achieves some outcome for an organism—and that function regulates or controls the behavior, is a critical insight for analyzing behaviors that one wishes to change. However, the reverse of this principle—that we can artificially manipulate behavior by managing its consequences—has proved more complex than it appears, partly perhaps because humans at least, resist the experience that they are being deliberately controlled.

See also: Classical Conditioning and Clinical Psychology; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Operant Conditioning and Clinical Psychology; Reinforcement: Neurochemical Substrates

Bibliography

Bandura A 1977 Self-efficacy: Toward a unifying theory of behavior change. *Psychology Review* **84**: 191–215

Carton J S, Nowicki Jr S 1998 Should behavior therapists stop using reinforcement? A reexamination of the undermining effect of reinforcement on intrinsic motivation. *Behavior Therapy* **29**: 65–86

Hull C L 1943 *Principles of Behavior: An Introduction to Behavior Theory*. Appleton-Century-Crofts, New York

Lawrence D H, Festinger L 1962 *Deterrents and Reinforcement: The Psychology of Insufficient Reward*. Stanford University Press, Stanford, CA

Noll J P 1995 The matching law as a theory of choice in behavior therapy. In: O'Donohue W, Krasner L (eds.) *Theories of Behavior Therapy: Exploring Behavior Change*. American Psychological Association, Washington, DC

Premack D 1965 Reinforcement theory. In: Levine D (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, Vol. 13, pp. 123–80

Skinner B F 1969 *Contingencies of Reinforcement: A Theoretical Analysis*. Appleton-Century-Crofts, New York

Staats A W 1996 *Behavior and Personality: Psychological Behaviorism*. Springer, New York

Thorndike E L 1927 The law of effect. *American Journal of Psychology* **39**: 212–22

I. M. Evans

Relapse Prevention Training

Many psychological and psychiatric problems are characterized by a temporal pattern in which symptoms are either chronic or vary in severity over a long period of time, with oscillations between remissions and return of symptoms. Substance-related disorders such as alcohol dependence are often relatively enduring, and mood disorders typically follow the second pattern of alternating symptom presence and remission. The empirical demonstration of the efficacy of psychotherapeutic interventions requires substantiation of the long-term benefits of therapies for such disorders (see *Psychological Treatments, Empirically Supported*). Accordingly, there has been a growth of interest in the development of therapeutic approaches that are designed to minimize relapse (or, conversely, maximize the maintenance of therapeutic gains) over extended periods of time. Relapse prevention refers to a set of procedures that designed to obtain long-lasting benefits. Marlatt and Gordon (1985) published the seminal work in the area of relapse prevention, in which they described a set of relapse prevention strategies for excessive alcohol consumption and other addictive disorders. This approach applies a cognitive–behavioral perspective to the problem of relapse and involves the employment of a specific set of treatment techniques. Other researchers have focused on relapse prevention for a diverse range of problems, adapting the original approach accordingly. Empirical support for the approach has been mixed, but there are some encouraging findings in relation to some specific problems. Methodological issues in the research place limitations on the conclusions that can be reached at this stage.

1. Conceptual Basis of Relapse Prevention

The term Relapse Prevention (RP) has been associated with a particular set of techniques as described by Marlatt and Gordon (1985). It can be argued that any approach which has as its clear goal the minimization of relapse (or enhancement of maintenance) might be construed as RP. Few approaches other than Marlatt and Gordon's RP, however, have been embedded in a

coherent theoretical framework that is intended to be generalizable across a variety of psychological or psychiatric problems. Cognitive and behavioral theories provide both the framework for the development of RP techniques and the conceptual basis for understanding why a particular person relapses and how best the problem might be ameliorated (see *Behavior Therapy: Psychological Perspectives*; *Cognitive Therapy*). Specifically, problem behaviors are viewed as arising from the influence of both learning principles (classical and operant conditioning) and cognitive appraisal processes (attributions, predictions, and other judgments). For example, a person's alcohol consumption is regarded as being (a) elicited by environmental cues (presence of stimuli associated with drinking such as a glass, bottle, certain companions, time of day) and (b) reinforced by its consequences (stress reduction or mood enhancement). Apart from environmental cues, a considerable amount of attention is given the role of cognitive factors in RP. Cognitive factors may include, amongst others, causal attributions or outcome expectations. Marlatt and Gordon's RP approach draws heavily upon the concept of self-efficacy (Bandura 1977) in its theoretical foundation (see *Self-efficacy*). Thus, methods which enhance a person's confidence that they will be able to perform the necessary behaviors and the belief that their performance will lead to an effective outcome, are thought to be the ones which are most likely to prevent relapse. A good deal of the research literature on predictors of relapse has focused on self-efficacy. For example, low self-efficacy about the performance of particular behaviors has been found to be predictive of relapse across several problem areas such as alcohol use, obesity, depression, and smoking. The underlying philosophy of RP emphasizes the notion that the maintenance period is an opportunity for the practice of newly acquired self-control strategies. Any 'lapses' that occur are viewed as being useful to provide information about the ways in which people identify, contribute to the occurrence of, or respond to, high risk situations. A 'lapse' may involve a return to the problem behavior on a single, specific occasion (or a small number of occasions), but a 'relapse' involves a repeated accumulation of lapses that approximate the level of severity of the original problem. From this standpoint, 'lapses' are seen as providing positive experience from which to draw important information that might assist in preventing future lapses, and, by implication, future relapse.

2. *Relapse Prevention Strategies*

In principle, RP procedures may be implemented at one of three different points in time, either as: (a) an integral part of the initial treatment program; (b) a discrete set of sessions that occur towards the end of treatment; or (c) a set of sessions that are spaced across a period of time after the end of initial treatment. In the case of the latter approach, sessions may sometimes be more properly regarded as 'booster sessions,' especially if the predominant approach involves a continuation of the basic components of the original treatment or the use of prompts for the person to employ the techniques used in the original treatment. Most commonly, Marlatt and Gordon's RP is viewed as an integrated treatment which focuses on the concept of prevention of 'lapses' from the outset, and continues this emphasis beyond the normal period of treatment into a 'maintenance phase.' In practice, studies of the efficacy of RP differ in the precise details of the timing of the delivery of the specific components. These components include (but are not limited to) the identification of high-risk situations, preparation for any high-risk situations that have been identified, explicit rehearsal of coping skills, relapse-crisis debriefing, education (e.g., about the effects of drugs), and lifestyle modification.

The identification of high-risk situations represents an essential component of the RP approach. For example, it might be found that resumption of alcohol consumption occurs when the person has had a clash of opinion with a colleague or manager at work. A list of idiosyncratic high-risk situations can be obtained, based either on knowledge of the person's previous history or on the responses to self-report inventories of typical situations in which the person is asked to indicate (a) how often these situations occur, and (b) how difficult it is to engage in the desired behavior or not engage in the undesired behavior when such a situation arises. In the case of cigarette smoking, typical high-risk situations could be (a) having a cup of coffee; (b) sitting down to study at the desk; or (c) seeing someone else smoke. Learning to recognize that a high-risk situation is occurring or is imminent is a crucial part of the RP approach. People are taught to recognize the importance of 'apparently irrelevant behaviors,' exemplified by the gambler who chooses to drive home from work via a more scenic route (justified by the better view) when the route just happens to pass the casino. On the other hand, other high-risk situations may be more obvious, even to the individual, such as the person who travels business class by plane and is, therefore, likely to be offered free alcoholic drinks. RP is also connected with the concept of the 'abstinence violation effect.' This effect is thought to underlie the occurrence of a relapse in a person who has merely lapsed by breaking a preplanned limit on daily alcohol or cigarette consumption. It is argued that once a person has broken the limit, they may fail to exert any further control over their behavior. RP decreases this emphasis on 'rule violation' by focusing on self-control and deriving useful information from the occurrence of any lapses.

RP involves teaching people not only to identify such situations as risky, but also to prepare themselves for reducing the likelihood of a lapse. Simulations of

high-risk situations can be used to provide opportunities to practice behavioral responses under supervision of the therapist. Many relapses are thought to occur following the occurrence of negative emotional states such as anger, depressive mood, and anxiety. Thus, RP involves the identification of such emotion–behavior linkages and the use of cognitive–behavioral strategies for the management of these negative states. Other lapses may be triggered by interpersonal conflict and may necessitate a heavier therapeutic focus on helping people to develop better communication and interpersonal problem-solving skills. In the case of addictive behaviors, there may be social pressure to engage in smoking, drinking or other drug taking. Preparation for dealing with such social pressure through assertiveness training may be incorporated into the treatment. Lifestyle modification is also employed in order to increase the frequency of engagement in pleasant, relaxing activities that may also be incompatible with the addictive behaviors.

3. *Efficacy of RP*

RP has been the subject of a number of experimental evaluations since it was first proposed as a treatment for alcohol and other addictive problems. Wilson (1992) reviewed the existing RP literature in the areas of alcohol problems, smoking, and obesity. Carroll (1996) also published a review of RP to a wide range of problems. Twenty-six studies, on substance use only (covering the years 1978 to 1995), were the subject of a meta-analysis conducted by Irvin et al. (1999). Even before considering the results of the empirical literature, the point needs to be made that there is surprisingly little research published on RP despite its presence as a concept in the treatment-outcome literature for about two decades. The drawing of firm conclusions is hampered by this relatively small body of literature, especially given the diverse types of psychological problems to which RP has been applied and the variety of procedural modifications that have been made in the published studies. A second cause for concern is the relatively short length of follow-up periods, especially given the explicit focus on relapse as an empirical issue. While it is likely that maintenance diminishes as time increases, and this appears to be the case (Irvin et al. 1999), it is not clear what type and level of resources need to be put into RP-based therapeutic endeavors to bring about meaningful results in the public health sense. With these, and other limitations in mind, the following tentative conclusions can be drawn.

The earliest controlled studies of RP were conducted on alcohol problems and it continues to be the main focus of attention for RP researchers. Varying degrees of success have been reported for RP in these studies. A number of methodological explanations may be offered for some failures to find differences between RP and control groups in some studies, including low statistical power, ceiling effects associated with the success of the initial intervention, the brevity of the follow-up, and the sharing of common elements in the treatments. Irvin et al. (1999) conclude that, amongst the substance use disorders, the strongest positive effects for RP have been found for problem drinking, although even within this group of studies, the outcome variable that is associated with the best response is psychosocial adjustment rather than drinking behavior *per se*. Like problem drinking, relapse rates following smoking cessation programs are notoriously high. While some success has been achieved using booster sessions, these effects frequently are lost after the eventual termination of the booster sessions. Studies of relapse prevention strategies for smoking cessation have also yielded inconsistent results. In their meta-analysis, Irvin et al. (1999) draw attention to the relatively poor outcomes with RP in the area of smoking cessation. It should be noted that a few studies have been conducted on RP for illegal drug use. Cocaine use has been the subject of a small amount of research. As with smoking, evaluations of RP for these addictions have not been noted for their successful outcomes. Irvin et al. (1999) also conclude that the effects for these problems are weaker than those for alcohol problems.

The maintenance of weight loss following cognitive–behavioral treatment was the subject of a series of studies reported by Perri and co-workers in the 1980s (see Perri et al. 1992). In this research, behavior therapy plus relapse prevention training and contact after treatment produced the largest effects. Despite the sometimes positive outcome of longer-term treatment for obesity using adaptations of RP approaches, it is clear that the overall gains associated with treatments for obesity are often modest and that many people regain the weight loss during an initial treatment period.

Other psychological problems for which an attempt has been made to develop and evaluate RP approaches include depression (Wilson 1992) and sexual offending (Laws 1989). Although it is well established that relapse following an episode of depression is as high as 50 percent over a two-year period, there has been little systematic exploration of specific, psychologically informed RP methods for this problem. It is possible that focus on RP for some types of problems such as depression is misdirected because the major impact on maintenance may be derived from the potency of the initial cognitive–behavioral intervention.

4. *Methodological Issues*

There are several methodological problems in the study of RP and its effectiveness. RP clearly is concerned with long-term maintenance, but the relatively brief length of follow-up periods in many studies

places some limitations on the conclusions that can be drawn. The RP approach contains a number of specific components, but these components are not necessarily employed in all studies and it is not clear which components are necessary for the best outcome. In many studies, RP procedures have been mixed into a broader treatment, making it difficult to separate the effects of RP from those of other components. With some problems, such as depression, it would be preferable for research on RP to be conducted using samples of people who are at high risk for a subsequent episode rather than for all people who complete an initial intervention. In some studies, especially when the RP is being delivered following the completion of an initial broad-based treatment approach (e.g., cognitive–behavior therapy for depression or anxiety), there is a need to separate initial treatment responders from nonresponders. There is little point in examining the impact of a post-treatment RP on nonresponders to an initial treatment program because they have yet to demonstrate responsiveness to treatment. These methodological issues pose particularly difficult problems for which the solutions are generally costly and time consuming to implement.

5. *Conclusion*

Despite nearly two decades of research on RP, the research produced has been quite meager and many questions remain to be addressed or resolved. The studies of RP for any given problem are so few in number that it is difficult to draw any firm conclusions at this stage. Scattered through the literature, there is some encouraging support for RP in relation to alcohol-related problems, but strong conclusions are limited due to methodological shortcomings and the relative scarcity of studies. There is a need for a large-scale evaluation of RP with addictive disorders. There has also been some progress on the development of techniques for other problems such as obesity, depression, sexual offending, and health problems.

The field of RP research may benefit from greater attention to improving our understanding of the specific ways in which RP affects basic psychological processes (i.e., investigations of the mechanism of change). Further research is also needed on both general and problem-specific predictors of relapse for particular types of symptoms so that the RP approaches can be tailored more specifically to suit the features of specific disorders. The timing of the implementation of RP procedures also needs to be considered in relation to knowledge about the temporal patterns of symptom presence and remission that are characteristic of different disorders. Continued efforts to refine and develop novel variation on RP approaches to chronic mental health and physical problems should remain a high priority for research.

See also: Behavior Therapy: Psychiatric Aspects; Behavior Therapy: Psychological Perspectives; Cognitive Therapy; Psychological Treatment, Effectiveness of; Psychotherapy Integration; Psychotherapy Process Research

Bibliography

Bandura A 1977 Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* **84**: 191–215

Carroll K M 1996 Relapse prevention as a psychosocial treatment: A review of controlled clinical trials. *Experimental and Clinical Psychopharmacology* **4**: 46–54

Irvin J E, Bowers C A, Dunn M E, Wang M C 1999 Efficacy of relapse prevention: A meta-analytic review. *Journal of Consulting and Clinical Psychology* **67**: 563–70

Laws D R (ed.) 1989 *Relapse Prevention with Sex Offenders*. Guilford Press, New York

Marlatt G A, Gordon J R 1985 *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors*. Guilford Press, New York

Perri M G, Nezu A M, Viegener B J 1992 *Improving the Long-term Management of Obesity: Theory, Research, and Clinical Guidelines*. Wiley, New York

Wilson P H (ed.) 1992 *Principles and Practice of Relapse Prevention*. Guilford Press, New York

P. H. Wilson

Relative Clauses

A relative clause (RC) can be defined as a subordinate clause which 'limits the potential reference' of a noun phrase (NP) in its containing clause, by describing a situation in which the referent of the NP participates. In the following examples, the RCs are enclosed in square brackets:

(1) (a) The woman [who we met yesterday] owns the local supermarket.
(b) We saw a lizard [which had a bright blue tongue].
(c) The developer sent an update to every user [who registered the program].

The phrase 'limits the potential reference' needs some further discussion. In (a), the subject of the sentence is a definite NP which is presumed to refer to a single individual, who is being said to own the supermarket. The RC provides a substantial portion of the information needed to identify that individual (that we met this person yesterday), the 'head' noun *woman* providing the rest (that the person is an adult female). In (b) on the other hand there is no presumption of unique identification, but the content of the relative clause, 'having a bright blue tongue,' must be true of the object NP, so is limiting its potential reference. Example (c) is the most complex. Here, the NP that the

RC is delimiting doesn't refer to any specific individual, but we can still say that it is delimiting the reference of the NP, because the quantifier *every* indicates that if an entity satisfies the conditions indicated by the RC and the noun *user* (is a potential referent of the NP), then the developer sent that entity an update (see *Quantifiers, in Linguistics*).

A problem for defining RCs is provided by so-called 'Nonrestrictive relatives' (NRCs) such as this:

(2) Lois was introduced to Clark, [who said that he had read all her books].

Here the bracketed clause is not really limiting the reference of the NP of *Clark*, since this is a proper name and presumed to identify an individual on its own. In some languages, such as English, NRCs show significant differences from ordinary RCs, while in others, such as Japanese, they don't (Kuno 1973, p. 233, Andrews 1985, pp. 28–9). It is not entirely clear whether NRCs should be regarded as a kind of RC or as a closely related construction.

1. *Essential Structure*

Since an RC delimits the reference of an NP by specifying the role taken by the referent of the NP in a situation, it needs to contain an internal position corresponding to the role of that NP, which can be called the NP_{rel} position. Many but not all contemporary syntactic theories postulate that the NP_{rel} position is occupied by an NP, which may be unpronounced, or moved from the normal location for an NP with that semantic role. In the RC (1a), for example, the NP_{rel} would be described as having the direct object role, and identified as the word *who* at the beginning of the RC, preposed from its usual postverbal location.

From the viewpoint of formal semantics (see *Montague Grammar*), one can propose that RCs have the logical form of λ-abstracts on a variable in the clause, occupying the NP_{rel} position. For example, the logical form of *who we met yesterday* would be along the lines of:

$$(\lambda_x)\ (\text{we met } x \text{ yesterday}).$$

Such an expression is taken to denote the set of all entities such that if x refers to that entity, the sentence 'we met x yesterday' is true. If RCs are defined as clauses with this logical form, then NRCs can be treated comfortably as a type of RC.

2. *Typology*

The typology of RCs across languages can be organized in terms of two major parameters involving the position of the relative clause, and the treatment of position of NP_{rel} within it, a variety of minor parameters involving other aspects of the form of RCs, and the extremely important topic of Accessibility, which will be deferred to the section on theoretical significance below.

2.1 *Major Parameters*

The first major parameter is the relationship of the RC to the NP whose reference it delimits. There appear to be five possibilities, falling into two major groups:

(3) (a) The clause forms a constituent with the NP it modifies:
  (i) positioned after the head N of that NP,
  (ii) positioned before the head N of that NP, and
  (iii) appearing in an NP with no head N (alternatively formulated by saying that the head appears internally to the RC).
(b) The clause doesn't form a constituent with the modified NP, but is a separate constituent:
  (i) at the front of the clause containing that NP, and
  (ii) at the end of the clause containing that NP.

Types 3(a)(i) and 3(b)(i) can be illustrated clearly from English; all the examples above are 3(a)(i), while 3(b)(i) is illustrated by so-called 'extraposed' relatives like this:

(4) A person was standing on the corner [who had a cockatoo on his shoulder].

Japanese provides perhaps the best-known examples of 3(a)(ii), which is widely found in verb-final languages, and sometimes in verb-medial languages such as Chinese

(5) (a) Japanese (Andrews 1985, p. 27):
Kore wa [ano hito ga kai-ta]
this THEME that person NOM write-PAST
hon desu.
book is
This is a/the book which that person has written.
(b) Chinese (Li and Thomson 1981, p. 116):
[Żhāngsān māí de] qìchē hēn guì.
Zhangsan buy NOM car very expensive
The car that Zhangsan bought was very expensive.

Example 3(c)(iii) is more exotic, but is commonly found in Native American languages. 6(a) is an example from Ancash Quechua (Cole 1987 p. 279), while 6(b) below is the corresponding 3(a)(ii) construction, also available in Ancash:

(6) (a) [Nuna bestya-ta ranti-shqa-n]
man horse-ACC buy-PERFECT-3
alli bestya-m ka-rqo-n.
good horse-EVID be-PAST-3

(b) [Nuna ranti-shqa-n] bestya
man buy-PERFECT-3 horse (NOM)
alli bestya-m ka-rqo-n.
good horse-EVID be-PAST-3
The horse the man bought was a good horse.

In 6(a), *bestya-ta* 'horse' is playing the same basic function as the head N in an English RC, or in the 6(b) example, but both its position and its accusative case-marking are those appropriate to its function as NP_{rel}. Such constructions are often called 'head internal' RCs.

Finally, type 3(b)(i) is especially characteristic of (Indo Aryan) South Asian languages such as Hindi and Marathi (Srivastav 1991), although it is also found elsewhere, such as in Aboriginal Australia (Hale 1976):

(7) (a) [Jo laRkii khaRii hai] vo lambii hai.
REL girl standing is DEM tall is
The girl who is standing is tall.

(b) [Nyuntulu-rlu kutja-npa wawirri
you-ERG RC-you(SUBJ) kangaroo
pantu-rnu], ngatyulu-rlu
spear-PAST, I-ERG
kapi-rna purra-mi.
FUT-I(SUBJ) cook-NONPAST
I will cook the kangaroo that you speared.

These structures are often called 'correlative clauses,' and one can question whether they should be seen as a type of RC or a closely related construction.

Srivastav, for example, argues that they are in fact quantifying expressions which include reference delimitation in their function, rather than mere reference delimiters.

The second major parameter of variation in RCs lies in the treatment of NP_{rel}. There are three important and partially independent parameters of variation:

(8) (a) Reduction of NP_{rel}:
(i) no reduction (NP_{rel} remains a full NP),
(ii) pronominalization (NP_{rel} is expressed as a pronoun), and
(iii) omission (NP_{rel} is not expressed).
(b) Marking of NP_{rel}:
(i) no special marking (NP_{rel} looks like an ordinary NP of a form that also appears in main clauses), or
(ii) a special 'relative' form is used for pronoun or determine of NP_{rel}.
(c) Change of position of NP_{rel}:
(i) remain *in situ*, or
(ii) move (usually to the front).

In both the internally headed RC of 6(a) and the correlatives of (7), NP_{rel} is unreduced. In standard English, NP_{rel} is either deleted or expressed as a special pronoun, which is furthermore moved to the front, perhaps accompanied by some surrounding material

(9) (a) The people [we talked with *e*] were interesting.
(deletion from position marked with *e*)
(b) The people [who(m) we talked with *e*] were interesting.
(movement of relative pronoun *who*(*m*) from *e*-position), and
(c) the people [with whom we talked *e*] were interesting.
(movement of relative pronoun together with containing PP).

In modern Hebrew (Borer 1984, p. 220), NP_{rel} can be omitted or reduced to an ordinary pronoun, which can either remain in *situ* or move to the front:

(10) (a) Ra?iti ?et ha-yeled she- rina ?ohevet (?oto).
saw-I ACC the-boy that Rina loves (him)
(b) Ra?iti ?et ha-yeled (she-) ?oto rina
saw-I ACC the-boy that him Rina
?ohevet.
loves
I saw the boy that Rina loves.

Pronominal NP_{rel} that remain *in situ* are often called 'resumptive pronouns.'

2.2 Minor Parameters

These are numerous; only two will be discussed here, clause introductory particles, and nominalization.

There may or may not be a clause introductory particle, which can appear instead of or in addition to a preposed NP_{rel}. In Modern Hebrew as illustrated in (10) above, the introductory particle is either *she-* (cliticized onto the next word) or *?asher*, and it is optional if NP_{rel} is preposed. In modern English, there is an RC-introducing word *that* which is optional, but can't co-occur with proposed NP_{rel}:

(11) (a) The person (that) I saw was tall.
(b) The person who I saw was tall.
(c) *The person who that I saw was tall.
(d) *The person that who I saw was tall.

A somewhat more subtle form of variation is in clause-reduction. In the examples we have seen so far, the RCs are full-fledged finite clauses, but it is possible for them to become reduced. In Turkish, for example, when NP_{rel} is neither the subject nor contained in the subject (this is an approximation; see Underhill 1972 for further discussion), the subject of the relative clause is put into the genitive case, and the verb of the relative clause takes a nominalizing affix followed by cross-referencing as is normally applied to possessed nouns. The relative clause of 12(a) thus bears a considerable resemblance to the possessed noun phrase of 12(b):

(12) (a) [oğlan-in git-tiğ-i] mektep
boy-GEN go-NOM-3SG school
The school that the boy goes to
(b) Ahmed-in šapka-si
Ahmed-GEN hat-3SG
Ahmed's hat.

Lehmann (1986) observes that such reduction is especially characteristic of structures of type 3(a)(ii) (RC preceding head with NP), and is apparently absent for correlatives (type 3(b)(i)).

Many further potential features of relative clauses are discussed in typological surveys such as Downing (1978), Keenan (1985), Andrews (1985), and Lehmann (1986), and in the papers in Peranteau et al. (1972).

3. *Theoretical Significance*

The theoretical significance of RCs has derived largely from the study of 'accessibility,' the possibilities for the role of 'underlying position' of NP_{rel} within the RC. Historically, this seems to have been divided into two largely independent investigations. Ross (1967) (following on from earlier work by Chomsky) studied systematically what might be called the 'long distance' aspect of the function of NP_{rel}, the capacity of NP_{rel} to be located in a subordinate clause structure within the relative clause itself. Later, Keenan and Comrie (1977) investigated the question of what kinds of functions NP_{rel} can bear in a monoclausal RC.

The long-distance aspect of accessibility revolves around the ability of NP_{rel} to occur, and be moved, specially marked, or deleted in a wide range of environments, including from arbitrarily deep subclauses of the RC (NP_{rel} position marked with *e*):

(13) (a) the book (that) Mary bought *e*,
(b) the book (that) Mary thinks Jane discussed *e*,
(c) the person (that/who) Mary thinks Jane spoke with *e*,
(d) the book that Mary thinks that Susan says that Sam borrowed *e*.

This carries implications about the kind of grammatical theory required to describe languages; simple phrase-structure rules in particular are clearly not adequate, some kind of augmentation being required (in early generative grammar, it was assumed that grammatical transformations were required; subsequently, a wide range of alternatives was found).

Most important is the fact that in spite of the long distance possibilities for NP_{rel}, its positioning is, nonetheless, limited by principles which tend to be similar across languages. In most languages that allow movement or deletion out of subordinate clauses, for example, this cannot happen from clauses inside NPs (the complex NP constraint):

(14) *the book that Mary believes the claim that Jane discussed *e*.

These constraints have been especially important for linguistic theory because:

(15) (a) They apply to more than just relativization, but rather a wide range of apparent deletion and movement phenomena.
(b) They appear in similar forms in many different languages (in both Irish and Hebrew, for example, the Complex NP constraint is observed when NP_{rel} is moved or deleted, but not when it is expressed as a resumptive pronoun).
(c) It is not clear how they could be learned on the basis of direct evidence of the kind presumably available to children learning their native language.

They therefore raise fundamental questions about universal grammar and what it is that makes human languages learnable by children (see *Movement Theory and Constraints in Syntax*).

One of the more striking discoveries is that long-distance relationships appear to be composed of shorter ones; this was first proposed on theoretical grounds by Chomsky (1973), direct empirical evidence appearing in McCloskey (1977), where it was shown that in Modern Irish, clause introductory particles assumed a special form if they were in an RC, and contained that RC's NP_{rel} (McCloskey 1977, p.17):

(16) (a) Meas mé gur thuig mé an t-úrscéal.
thought I that understood I the novel
I thought I understood the novel.
(b) an t-úrscéal a mheas mé a
the novel REL thought I REL
thuig mé
understood I
the novel I thought I understood.

In 16(a) we see the normal form *gur* of the particle introducing a complement of the verb *think*, but in 16(b), when the object of this complement is a (deleted) NP_{rel}, the form *a* is used instead. Similar 'intermediate position' effects were discovered in other languages and collected in Zaenen (1983).

The other aspect of accessibility is the function of NP_{rel} in monoclausal RCs. Keenan and Comrie (1977) proposed that there was an accessibility hierarchy:

(17) Accessibility hierarchy
SU > DO > IO > OBL > GEN > OCOMP

such that any RC strategy would apply to NP_{rel} whose functions were a continuous segment of the hierarchy, and languages could always relativize subjects. In Modern Hebrew, for example, SU (subject) and DO (direct object) can be relativized by the deletion strategy, while pronouns (*in situ* or moved) can only be used when NP_{rel} is DO or below.

The validity of the hierarchy corroborates the universal significance of the grammatical relation concepts used in formulating it (see *Grammatical Relations*), and raises significant issues about their nature. For example, Fox (1987) observed that, in spoken discourse, it was relatively rare for transitive subjects to be NP_{rel} (10 percent), most NP_{rel} being either objects (45 percent) or intransitive subjects (45 percent). The high frequency of relativing objects seems discrepant with a tendency for relativization to be restricted to subjects.

A solution is afforded by a reinterpretation of grammatical relations, in which the traditional 'subject' concept is split into one concept, associated with the semantic role of 'Agent,' and another with the semantic role of 'Topic.' This idea goes back to early work by Schachter (1997), in which relative clauses in

Tagalog constitute some of the evidence, and has been incorporated into many contemporary syntactic frameworks (see Manning 1996, Van Valin and LaPolla 1997 for recent discussion). Fox's observations then indicate that the grammatical relation at the top of the accessibility hierarchy should not be identified with the traditional concept of subject, but with the topic-oriented subject of this more recent work.

Bibliography

Andrews A D 1985 *Studies in the Syntax of Relative and Comparative Clauses*. Garland Press, New York

Borer H 1984 Restrictive relatives in Modern Hebrew. *Natural Language and Linguistic Theory* **2**: 219–60

Chomsky N 1973 Conditions on transformations. In: Anderson S R, Kiparsky P (eds.) *A Festschrift for Morris Halle*. Holt, Rinehart and Winston, New York, pp. 232–86

Cole P 1987 The structure of internally headed relative clauses. *Natural Language and Linguistic Theory* **277**: 302

Downing B T 1978 Some universals of relative clause structure. In: Greenberg J H (ed.) *Universals of Human Language*. Stanford University Press, Stanford, CA, Vol. 4, pp. 357–418

Fox B A 1987 The noun phrase accessibility hierarchy reinterpreted: Subject primary or the absolutive hypothesis. *Learning* **63**: 856–70

Hale K L 1976 The adjusted relative clause in Australia. In: Dixon R M W (ed.) *Grammatical Categories in Australian Languages*. Australian Institute of Aboriginal Studies, Canberra, pp. 78–105

Keenan E L 1985 Relative clauses. In: Shopen T (ed.) *Language Typology and Syntactic Description, Vol. II: Complex Constructions*. Cambridge University Press, Cambridge, UK, pp. 141–70

Keenan E L, Comrie B 1977 Noun phrase accessibility and universal grammar. *Linguistic Inquiry* **8**: 63–99

Kuno S 1973 *The Structure of the Japanese Language*. MIT Press, Cambridge, MA

Lehmann C 1986 On the typology of relative clauses. *Linguistics* **24**: 663–81

Li C N, Thompson S A 1981 *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, Berkeley, CA

Manning C D 1996 *Ergativity: Argument Structure and Grammatical Relations*. CSLI Publications, Stanford, CA

McCloskey J 1977 *Transformational Syntax and Model Theoretic Semantics*. Reidel, Dordrecht

Peranteau P M, Levi J N, Phares G C (eds.) 1972 *The Chicago Witch Hunt: Papers from the Relative Clause Festival*. Chicago Linguistic Society, Chicago

Ross J R 1967 *Constraints on Variables in Syntax*. Ph.D. thesis, MIT

Schachter P 1977 The subject in Philippine languages: Topic, actor, actor-topic, or none of the above. In: Li C N (ed.) *Subject and Topic*. Academic Press, New York, pp. 491–518

Srivastav V 1991 The syntax and semantics of correlatives. *Natural Language and Linguistic Theory* **9**: 637–86

Underhill R 1972 Turkish particles. *Linguistic Inquiry* **2**: 87–100

Van Valin R D, LaPolla R J 1997 *Syntax: Structure, Meaning and Function*. Cambridge University Press, Cambridge, UK

Zaenen A 1983 On syntactic binding. *Linguistic Inquiry* **14**: 469–504

A. D. Andrews

Relativism: Cognitive

Cognitive (or epistemological) relativism is the view that every belief is bound to some individual, group, place, or time—whence there are no objective, cross-cultural, and universal truths. According to it, what is true for us may be false for them. Hence, all beliefs would be equivalent. Whence there would be no point in searching for objective truths. Consequently, there would be no such thing as objective advancement of knowledge. Under cover of the search for truth there would only be struggle for power. Therefore, the task of the relativist would be to unmask ('deconstruct') scientific discourse and expose the powers lurking behind it. How this task can be accomplished without finding out what is really, truly the case, is not being explained.

1. Relations to Skepticism, Constructivism, and Postmodernism

1.1 Relativism is an Aspect of Skepticism

Cognitive relativism is an aspect of systematic (as opposed to methodological) skepticism. This is the view that truth is unattainable, whence 'anything goes.' It is part of anthropological relativism, and a frequent concomitant of axiological and moral relativism ('All value systems and moral codes are equivalent'). Relativism opposes absolutism, the view that all verities are both universal and timeless, hence uncontroversial. It also opposes scientific realism, which holds that there are universal truths, such as 'Agency and structure are mutually dependent.' But realism contends that, unlike mathematical theorems, propositions about matters of fact may be only approximately true, yet corrigible in the light of further discussion and new evidence. Thus, scientific realism is antidogmatic and moderate, whereas relativism is neither. In small doses, skepticism whets the appetite for research; in large doses, it extinguishes such appetite.

1.2 Constructivism Entails Relativism

Constructivism can be cognitive or ontological. The former holds that concepts and hypotheses are human constructions rather than either innate ideas or the product of revelation, perception, or intuition. On-

tological constructivism claims that the knower makes the world. Cognitive psychology has confirmed the hypothesis that all concepts and hypotheses are constructed rather than given. By contrast, ontological constructivism strains credibility. Moreover, it is destructive, because it dispenses with any tests aiming at checking whether a hypothesis matches the facts it refers to. In particular, it denies the difference between reality and appearance, which may be deceiving: it bans recurrent questions of the forms 'What is really the case?' and 'Which of the two hypotheses is the truer?'

Whether ontological or epistemological, constructivism comes in two varieties: individualist and social. Whereas according to the former the individual is the source of everything, according to the latter the social group, in particular the scientific community, constructs whatever it studies, even natural things such as stars. Whether individualist or collectivist, ontological constructivism implies cognitive relativism. However, cognitive relativism can be held, and is usually held, independently of ontological constructivism. After all, it is logically possible to admit the reality of the external world while claiming that it is impossible to know it objectively, hence intersubjectively. Therefore, relativism must be evaluated separately from constructivism.

1.3 Relativism is a Component of Postmodernism

What is usually called 'postmodernism' in social studies and in philosophy is neither a worldview nor a method: it just boils down to the rejection of the ideals of rationality and objectivity. Cognitive relativism fits in with postmodernism because it denies the superiority of rigorous reasoning over 'weak thought,' as well as the possibility and desirability of attaining objective knowledge. It is unfazed by hard social statistics, or even by the findings of natural science. It ignores the difference between 'I feel cold' and 'The room temperature is below the freezing point,' and *a fortiori* between 'That government is generally perceived as corrupt,' and 'The audits found that the government officials are regularly bribed.'

2. The Alleged Accomplishments of Cognitive Relativism

2.1 Relativistic Epistemology

Cognitive relativism is suggested by the multiplicity of simultaneous or successive rival views about many a domain of facts. Actually, such variety only goes to prove that scientific research does not guarantee instant, complete, and definitive truth, let alone instant worldwide diffusion. Experience shows abundantly that we often obtain approximately true data and hypotheses, which can usually be improved upon. For instance, water was first regarded as simple and continuous, later as constituted by HO molecules, and nowadays as made up of H_2O molecules—and moreover as a mixture of light and heavy water. Scientific realism accounts for such successive approximations to the exact objective truth.

The denial of the possibility of objective truth has several consequences. One is that it discourages the very search for truth. Another is that it renders the idea of error unintelligible since, by definition, error is discrepancy from truth. A third consequence of relativism is that it regards all controversies as unending unless stopped by nonscientific means, such as authority or negotiation—which invites censorship. A fourth is that, by placing magical thinking on the same footing with critical thinking, it debases a key accomplishment of modernity. By the same token, it deprives people of a key tool to diagnose deceit and resist arbitrary authority.

2.2 Relativistic Sociology of Knowledge

Constructivist–relativist sociologists of science have been known to claim to study the sociology of peptides, rather than that of peptide-research laboratories. Moreover, they contend that scientific research is a kind of struggle for power, not truth. They miss the two main generally acknowledged motivations of basic scientists, namely curiosity and peer recognition. This explains why they have not helped explain the occurrence of any of the scientific breakthroughs in recent decades, such as the birth of tectonic plate geology, molecular biology, the neurobiological study of the mind, or socioeconomics. Understanding the search for truth requires realizing that this is what scientific research is all about.

Worse yet, although the constructivist–relativists talk much about power, they overlook the main loci of social power: politics, business, and organized religion. This explains why they have not studied such important facts as the campaigns against evolutionary biology; the budget cuts in pure mathematics, sociology, and the humanities under conservative governments; the persistent overlooking of naturalist (materialist) philosophies in the philosophy curricula; or the current toleration of charlatanism in the faculties of arts alongside the continuing demand for rigor in the faculties of science, engineering, medicine, and law.

2.3 Relativistic Pedagogy

If there is no objective truth, but only opinion, the teacher's mission is not to teach but to promote the confrontation of rival views. In this view, the teacher is not a master craftsman but a moderator, and the student is not an apprentice but a discussant. This is in

fact how many university classrooms have functioned in the US and in Europe since the mid-1960s: not as learning centers but as debating clubs. Such transformation has had two effects, one positive and the other negative. The former consists in a lessening of the dogmatism, authoritarianism, and stodginess of traditional education. However, this liberation has also deprived students of the motivation and discipline required to learn difficult ideas and procedures—among them the art of informed rational discussion. Consequently, the graduates of relativistic pedagogy have few skills to sell. The most they can hope for is a university post where they can turn students away from serious learning.

3. Intellectual and Social Roots

Skepticism has waxed and waned in different periods since antiquity. It is likely that it prospers in times of social crisis, when the established order of things is challenged, and a number of individuals lose their old faith but are not yet ready to embrace any new belief system. In this connection, it has been claimed that relativism is a perverse effect of egalitarianism and liberalism. However, the Enlightenment and its heirs endorsed reason and science, which they regarded as universal. Their targets were despotism and religious dogma. Much the same holds for the Renaissance skeptics, such as Francisco Sanches and the early modern skeptics, such as Bayle, Montaigne, and Hume. But this does not hold for the ancient skeptics, such as Sextus Empiricus, who only criticized the ignorance and petulance of the professors of their time.

The postmodern skeptics, from the constructivist–relativists to the academic feminists and the 'deep ecologists,' are of a different color. Unlike the skeptics of old, they seldom argue cogently: typically, they just pile assertion upon assertion. Some of them attack reason and science out of horror of the nuclear bomb and other misuses of science. Others seem just alienated from the more rigorous, demanding, and rewarding disciplines. Some are suspicious of science because they confuse it with positivism. Others, because they mistake physical relativity (dependence of the values of certain physical magnitudes upon the reference frame) for cognitive relativism. Still others invoke either the Marxist thesis that ideas are produced by social classes, or the subjectivism inherent in phenomenology and hermeneutics.

The current vogue of cognitive relativism is further congruent with two contemporary social and intellectual currents and attitudes: individualism and uncritical nonconformism. The radical individualists, such as the members of the so-called 'me-generation,' claim that their own views are not inferior to any others. And the emotional non-conformists do not use social science to diagnose or cure social ills: they just revolt against all the powers that be—political, economic, and cultural. They mistrust or even reject the entire culture that comes with the social order they reject. In particular, they reject mathematics, science, and technology—some, for regarding them as tools of 'late capitalism.'

4. Conclusion

Cognitive relativism, like skepticism in general, has a grain of truth, namely, the thesis that all propositions about matters of fact are in principle fallible. But, to be realistic and fertile, fallibilism must be combined with meliorism, or the thesis that error can be corrected. What counts as truth in science and technology, even as partial truth, is universalizable. If a view is only acceptable to some people, then it is not scientific but aesthetic or ideological. Beauty may be in the eye of the beholder: truth is not. Unless this be admitted, it becomes impossible to distinguish science from pseudoscience and ideology, to tell science-based technology from quackery, and humanity from tribe. Worse yet, relativism undermines democracy and the rule of law, because both require rational debate and agreement about the truth or falsity of certain claims, such as, for example, whether the family is in decline, and whether 'wars on crime' work. In short, relativism paralyzes.

See also: Collective Beliefs: Sociological Explanation; Constructivism/Constructionism: Methodology; Knowledge, Sociology of; Relativism: Philosophical Aspects; Sociology, Epistemology of; Theory: Sociological

Bibliography

Archer M 1987 Resisting the revival of relativism. *International Sociology* **2**: 235–50

Bottomore T 1956 Some reflections on the sociology of knowledge. *British Journal of Sociology* **7**: 52–8

Boudon R 1995 *Le Juste et le Vrai*. Fayard, Paris

Boudon R, Clavelin M (eds.) 1994 *Le Relativism est-il Résistible?* Presses Universitaires de France, Paris

Bunge M 1996 *Finding Philosophy in Social Science*. Yale University Press, New Haven, CT

Bunge M 1999 *The Sociology–Philosophy Connection*. Transaction Publishers, New Brunswick, NJ

Collins R 1998 *The Sociology of Philosophies: A Global Theory of Intellectual Change*. Harvard University Press, Cambridge, MA

Feyerabend P K 1975 *Against Method*. Verso, London

Gellner E 1985 *Relativism and the Social Sciences*. Cambridge University Press, Cambridge, UK

Gross P R, Levitt N 1994 *Higher Superstition: The Academic Left and its Quarrels with Science*. Johns Hopkins University Press, Baltimore, MD

Knorr-Cetina K, Mulkay M (eds.) 1983 *Science Observed: Perspectives on the Social Study of Science*. Sage, London

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Latour B, Woolgar S 1986 *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, NJ
Laudan H 1990 *Science and Relativism*. University of Chicago Press, Chicago
Ross A 1992 *Strange Weather: Culture, Science, and Technology in the Age of Limits*. Verso, London
Siegel H 1987 *Relativism Refuted*. Reidel, Dordrecht, Germany
Sokal A, Bricmont J 1998 *Fashionable Nonsense: Postmodern Intellectuals' Abuse of Science*. Picador, New York
Trigger B G 1998 *Sociocultural Evolution: Calculation and Contingency*. Blackwell, Oxford

M. Bunge

Relativism: Philosophical Aspects

1. Definition

Relativism is the view that cognitive, moral, or aesthetic norms are dependent on the social or conceptual systems that underpin them and consequently a neutral standpoint is not available to us. Relativists claim that standards of truth, rationality, and ethical right and wrong vary greatly between cultures and historical epochs and that there are no universal criteria for adjudicating between them. Relativism hence stands in opposition to universalism and absolutism. Although the first expression of relativism as a philosophical thesis is found in ancient Greece, it was only in the second half of the twentieth century that relativism became a prominent feature of various intellectual fields, including philosophy, social sciences, and literary theory. In recent years, postmodernism, in particular, has become identified with relativism.

2. Motivations

The appeal of relativism arises out of certain features of our experiences of the world as well as a number of intellectual currents.

2.1 The Indexicality of our Judgements

Many of our judgements and assertions refer to events that happen at a particular time, place, and are experienced by a particular individual. Such assertions seem to be true only at the time, place, or context of their utterance. For instance, the sentence 'It's raining today' is true at the time of writing it but may be false when uttered at a different time or place. This feature of our judgements has led some to argue that the truth of all our judgements depends on, and in that sense is relative to, their time and place, and context.

This argument for relativism, despite its popularity, is not very convincing. Sentences containing indexical expressions such as 'this,' 'here,' 'now,' 'I,' etc., are abbreviations of statements or sentences containing references to specific objects, dates, geographic locations, and named persons. The content of an indexical statement can be made explicit in such a way that its truth or falsity would arise from specifiable events occurring at a particular time, place, or context. For instance, 'It's raining here today' is equivalent to 'There is rain in Dublin on September 12, at 5:45 local time, 1999' and that sentence is true or false irrespective of when or where or by whom it was uttered.

2.2 Perspectivalism

The perspectivalism of all our reports and experiences, the impossibility of having a view from nowhere, has been used to support relativism. All our judgements, beliefs, or assertions, it is claimed, are made from a point of view, a perspective. A God's eye view or a view from nowhere is not available to us. Perspectivalism goes beyond the mere indexicality of our judgements in that even a nonindexical judgement, for instance, 'there are nine planets in the Solar system,' is a statement made from a human perspective as we cannot step out of our skin, so to speak, and experience reality as it is. In addition, relativists assume, explicitly or implicitly, that different perspectives could not converge and therefore what is true or right from one perspective may not be so from a different one and consequently the truth of a judgement is relative to a perspective or point of view we have or we adopt.

2.3 Cultural Diversity

Increasing awareness of the extent of diversity of beliefs, practices, and customs of different cultures and different historical epochs has cast doubt on the universality and objectivity of our judgements. Most significantly, the reports by anthropologists about the thinking habits of remote peoples have led to the suggestion that all normative judgements, whether cognitive or ethical, may have only a local validity. Opponents of relativism argue that the extent and scope of diversity between different cultures and individuals is often exaggerated. Beyond the apparent dissimilarities, they argue, there are many core similarities which unify all human cultures and systems of beliefs. Relativists, on the other hand, believe that Western ethnocentrism and the intellectual legacy of the Enlightenment prevent us from appreciating or even seeing the uniqueness of different cultures and modes of thought.

2.4 The Disappearance of Old Certainties

The disappearance of old certainties, both in the religious and scientific arena, has been instrumental in the popularity of relativism in recent years. The collapse of a religiously motivated cosmology which

fixed the position of individual human beings within a larger and immutable framework and provided firm foundations for their ethical outlook helped to bolster a climate that was conducive to relativistic outlooks.

In science, the discovery of the possibility of non-Euclidean geometries followed by the scientific developments at the beginning of the twentieth century eroded the confidence placed in classical geometry and physics. Einstein's theory of relativity cast doubt on Newtonian physics and deprived it from the unassailable status it had enjoyed. The new theories in the area of quantum mechanics and particle physics, Heisenberg's Principle of Uncertainty in particular, often misrepresented and misunderstood, added to the belief that universal, absolute laws are rarer than it was once imagined.

2.5 *Influence of Early Philosophers*

In addition to the above intellectual and socio-historical influences, contemporary relativism is heir to the legacies left by earlier philosophers. The earliest documented source of relativism in the Western intellectual tradition is Plato's account of the Sophist philosopher Protagoras and his famous dictum, 'Man is the measure of all things; of the things that are, that they are; and of the things that are not, that they are not' (Plato 1961). As in twentieth century, the relativistic tendencies in ancient Greece were fostered by the increasing contact between the Greeks and other people; the realization that other societies may arrange their social and ethical beliefs and customs in radically different ways had a great impact on Greek thinkers. Relativism was subsumed under scepticism by the Pyrrhonian sceptics who in turn influenced Michel De Montaigne (1533–92) one of the most important figures at the outset of modern philosophy. Giambattista Vico (1668–1744) and Johann Gottfried von Herder (1744–1803) paved the way for historicism, which in turn has shaped twentieth century relativism. Contemporary relativists are influenced directly by the perspectivism of Friedrich Nietzsche (1844–1900) and the historicism of Wilhelm Dilthey (1833–1911).

According to Nietzsche, all claims of knowledge are representations of reality only from a certain standpoint. Reports of so-called 'facts' are statements of interpretation which could always be supplemented or replaced by other interpretations. Language is not a means of describing what there is, Nietzsche claims, rather it imposes its own interpretation on our thinking. Thus, the very idea of truth and reality is called into question by Nietzsche. Whether Nietzsche's perspectivism should be equated with relativism is open to debate. However, his writings have become a source of inspiration for postmodernist and relativist thinkers.

Dilthey also has influenced a relativistic understanding of social and historical conditions in contemporary hermeneutics and social theory through his argument that each nation can be seen as a self-contained unit with its own conception of reality and system of values. Different historical periods, Dilthey maintains, produce different norms and values, each of which is taken to be unconditional or universal. We cannot gain a correct understanding of social and intellectual conditions if we do not understand or take into account the inevitable historicity of all thought.

3. *Varieties of Relativism*

Relativism takes many shapes and forms. One way to distinguish between different types of relativism is to consider the questions:

(a) What is it that is being relativized?

and

(b) What is it being relativized to?

A reply to the latter question allows us to distinguish between subjective relativism (or subjectivism) on the one hand and social (or historical) relativism on the other. Subjectivists or subjective relativists argue that truth and falsity of all judgements and the right and wrong of actions are relative to the beliefs and opinions of the individual thinkers and actors. The view entails that the truth and falsity of our judgements come down only to what we believe or like. This type of relativism, although quite popular, is too naive and unsophisticated to be worthy of serious philosophical consideration. Subjective relativism rules out the possibility of disagreement and the very distinction between correct and incorrect judgements, for it turns all our judgements, as long as we believe in them, into correct or true ones.

Social (historical) relativism is the claim that the truth and falsity of all judgements and the right or wrong of actions are relative to their social, historical, and cultural contexts. According to this view, we are indeed in a position to distinguish between true and false beliefs and right and wrong actions and judgements, but only within the parameters of socially given norms and beliefs.

The most influential form of social relativism in twentieth century is known as 'cultural relativism.' Cultural relativism was inspired by the work of social anthropologists who conducted their fieldwork among tribal people. Ethnographic evidence seems to suggest that various groups of people hold systems of belief which according to Western standards are false, irrational, and even self-contradictory. The Nuer's belief that twins are birds or the Azande belief in witchcraft and oracles are some of the frequently cited examples.

Anthropologists distinguish between the methodological and the normative components of cultural relativism. The methodological principle enjoins the practicing anthropologist to refrain from taking judgmental positions on the people under study in

order to attain a greater degree of objectivity. The normative consequence of cultural relativism is respect for other cultures and value systems—which can manifest itself as an adherence to multiculturalism and avoidance of 'ethnocentrism' in contemporary Western societies.

One important stimulus for the adoption of cultural relativism has come from the study of languages of non-Western people. The work of Benjamin Whorf—known as 'linguistic relativity'—is the locus classicus of this approach. Whorf's study of American Indian languages, such as that of the Hopi, led him to claim that 'we dissect nature along lines laid down by our native languages. The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organised by our minds—and this means largely by the linguistic systems in our minds. We cut nature up, organise it into concepts, and ascribe significance as we do, largely because we are parties to an agreement to organise in this way—an agreement that holds throughout our speech community and is codified in the patterns of our language' (Whorf 1956). According to Whorf, the Hopi articulate and hence perceive the world in terms of events rather than objects, and see time not in terms of duration as English speakers do, but as relations between events. Whorf's principle of relativity states that 'all observers are not led by the same physical evidence to the same picture of the universe, unless their linguistic backgrounds are similar, or can in some way be calibrated' (p. 214). We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation. In this sense, the worlds in which different societies live, he argues, are distinct worlds and not the same world with different labels attached to them.

Further ethnographic and linguistic studies have not supported Whorf's hypothesis of linguistic relativism. Nonetheless his views have had great impact on the resurgence of interest in relativism in recent decades.

The attempt to answer the question 'what is it that is being relativized?' gives us a different set of subdivisions of relativism. Depending on whether cognitive or ethical norms are being relativized, philosophers distinguish between cognitive and moral relativism.

3.1 Cognitive Relativism

Cognitive relativism is the view that what is true or false, rational or valid can vary from society to society and culture to culture and we have no transcultural or ahistorical criterion for adjudicating between conflicting cognitive norms. Some main subdivisions of cognitive relativism are relativism about truth, relativism about rationality and logic, and conceptual relativism.

3.1.1 Relativism about truth. Relativism about truth is the claim that what is true for one individual, or one social group, may not be true for another and that every effort to adjudicate between context-dependent standards of truth and falsity is bound to be futile. The truth of an assertion, so the claim goes, is relative to the beliefs, attitudes, and other psychological idiosyncrasies of the individuals, or more generally and commonly to the conceptual, or cultural, background of the persons uttering them.

Relativism about truth is the core of all relativistic positions since other subdivisions of cognitive relativism, and even moral relativism, are reducible to it. For instance, relativism about rationality can be expressed in terms of whether there are true (universal) standards of rationality. Relativism about logic may be restated as the question whether logical truths are relative to specific cultures or cognitive schemes or whether they are in any sense universal. Ethical relativism can be seen as the claim that the truth of ethical judgements is relative to the context or the culture in which these particular judgements are being made. In this sense relativism about truth is the most general and strongest form of relativism and the target of most criticisms of relativism (see Sect. 5).

3.1.2 Relativism about rationality. The relativist about rationality argues that various societies or cultures have different standards and criteria of rationality and we are not in a position to evaluate, or choose, between these divergent standards. For instance, Peter Winch has argued that since standards of rationality in different societies do not always coincide, we should use only contextually and internally given criteria of rationality in our assessment of the systems of belief of an alien culture. Winch, who based his arguments on Evans-Pritchard's account of Azande's beliefs in witchcraft and magic and who was greatly influenced by the work of the philosopher Ludwig Wittgenstein, argues that we can only use contextually and internally given criteria of reasoning and validity in our assessment of the systems of belief of other cultures, and hence we cannot impute irrationality or illogicality to the Azande. It does not make sense to speak of a universal standard of tionality because rationality means conforming to internal norms given by the language and society, i.e., the form of life in which the language is being used.

In more recent years, the discussion of the Azande material, coupled with a strongly relativistic interpretation of Wittgenstein's views have led Barry Barnes and David Bloor to argue that not only standards of

rationality but also laws of logic, often seen as the necessary conditions for rational thought, are defined by and hence are relative to their social context. According to them the very distinction between valid and invalid modes of argument is relative to the practices prevalent in a given linguistic community or society and cannot in any sense be seen as universal. Bloor argues that there are no culture-transcendent rules of validity and inference. Different societies may have incompatible, but internally coherent, systems of logic and the norms of logical reasoning are relative to a background of cultural conditioning. According to Bloor, 'The Azande have the same psychology as us but radically different institutions. If we relate logic to the psychology of reasoning we shall be inclined to say that they have the same logic; if we relate logic more closely to the institutional framework of thought then we shall incline to the view that the two cultures have different logics' (Bloor 1976, pp. 129–130).

3.1.3 Conceptual relativism. Conceptual relativism is motivated by philosophical considerations about the relationship between human thought and the world rather than mere observation of cultural and historical differences, and for that reason it is both quite distinct from and much more sophisticated than other varieties of cognitive relativism. The roots of conceptual relativism rest with Immanuel Kant's (1724–1804) distinction between the data of our sense experiences and the principles of organization or categories we use to organize them. Kant believed that all thinkers necessarily apply the same categories or conceptual schemes to the undifferentiated flow of their experiences. However, once the distinction between a conceptual scheme and the content of that scheme was introduced, then it was easy to accept that there may be more than one system or scheme of organization and the idea of conceptual relativism was born.

One of the most influential versions of conceptual relativism in twentieth century was proposed by W. V. O. Quine (1908–2000). According to Quine, to be able to talk about the world and to cope with our stream of sensory experience, we must impose upon the world a conceptual or linguistic scheme or theory. However, it is possible to envisage a plurality of conceptual or theoretical frameworks all of which explain and predict our experiences of the world equally well. There are no determinate facts to enable us to choose between these different frameworks, hence what Quine has called 'ontological relativity' ensues.

The influence of Quine's views is widespread. Philosophers working in differing areas have explicitly, or implicitly, made use of his arguments to support their relativistic positions. We can see the impact of Quine's ontological relativism on the writings of Nelson Goodman and Hilary Putnam who have proposed very sophisticated forms of conceptual relativism. Putnam, for instance, has denied that there is a fact of the matter as to even what counts as an object outside the standards set by a conceptual scheme. One and the same situation, according to Putnam can be described 'as involving entirely different numbers and kinds of objects' (1992, p.120). Different conceptual schemes will contain bits that are 'true' (or 'right') and bits that will turn out to be 'wrong' by the standards appropriate to the scheme itself—but we do not have an account of truth which would be external to, and independent of, all conceptual schemes.

It is important to note that most conceptual relativists, Quine and Putnam in particular, wish to distance themselves from cultural relativism which they think is fundamentally incoherent (see Sect. 5). Their brand of relativism is based on their theories of meaning and truth rather than the purported differences between cultural traditions.

Conceptual relativism has also influenced the philosophers of science, Thomas Kuhn and Paul Feyerabend, who have emphasized the role of different modes of reasoning at various periods in the history of science and have argued that scientific theories belonging to these divergent styles of reasoning are incommensurable. Kuhn has claimed that scientists working with different paradigms live in different worlds, Feyerabend has argued that a change of universal principles brings about a change of the entire world. Speaking in this manner we no longer assume that there is an objective world that remains unaffected by our epistemic activities, except when moving within the confines of a particular point of view. Kuhn's thesis of incommensurability rested on his assumption that scientific theories cannot be compared or assessed independently of the paradigms in which they are embedded. For Kuhn, each paradigm acts as a unique and untranslatable language or conceptual framework and since rival paradigms do not have access to a common criterion for adjudication, they cannot be compared and hence remain self-contained and beyond the critical reach of other paradigms. According to this understanding, incommensurability implies that evaluation of different conceptual schemes, cultures, or ways of life is impossible.

Subsequent to the first publication of *The Structure of Scientific Revolutions* in 1962, Kuhn, unlike Feyerabend, attempted to distance himself from relativism but his views remain an important source of inspiration for relativistically inclined thinkers in many fields.

3.2 Moral Relativism

Moral inquiry is concerned with how life should be lived and how human beings should behave towards each other and their world. Moral relativism is the claim that there exist diverse, incompatible moral

systems and there are no overarching criteria to decide between them. Moral relativists start from the premise that moral values are grounded on societal conventions, historical conditions, metaphysical beliefs, etc., which vary from one society or social grouping to another and argue that there are no neutral standards available to us for adjudicating between such competing claims. The truth or falsity, the appropriateness or inappropriateness of ethical beliefs, are relative to their sociohistorical backgrounds and moral beliefs cannot be assessed independently of their social framework.

Moral relativism is usually subdivided into descriptive and ethical (or meta-ethical) relativism. Descriptive relativism purports to be an empirical doctrine based on observation of diversity in moral judgements and moral values in different societies and at different times. There is widespread disagreement among individuals and cultures on issues ranging from capital punishment to abortion, from human rights to animal rights, and so on. Descriptive relativists go so far as accepting that there may be broad similarities in the core beliefs of different societies, for instance, injunctions against the killing of innocents, but argue that such shared values are given very different and even incompatible interpretations by different cultures.

Ethical relativists claim that descriptive relativism entails that all ethical judgements are relative to the culture or context in which they are made and that there is no unique criterion for adjudicating between them. Of course moral diversity is compatible with the possibility that there are right and wrong answers to ethical questions. But this would depend on having a dependable path towards the discovery of ethical truths or the establishment of a universal framework for arbitrating between incompatible moral claims. Ethical relativists deny this possibility because they claim that any rational decision making procedure will have deeply seated normative historical and cultural presuppositions embedded in it. Reason and rationality are themselves evaluative concepts, constrained in turn by cultural and historical influences and hence are incapable of leading us to a set of neutral universally applicable standards of evaluation. Thus, ethical relativists deny the possibility of convergence between conflicting ethical frameworks.

Ethical relativism can be embraced independently of cognitive relativism. Ethical relativists may agree that convergence between diverse world views and theoretical frameworks is possible in the natural sciences but deny that this is the case in the realm of ethics. According to this view, moral precepts and judgements are not part of and do not pertain to the natural furniture of the universe. They are man-made and would not exist independently of human actions, beliefs, and customs, hence, there exists a fundamental difference between scientific investigations and moral enquiry.

Ethical relativism is often associated with tolerance and open-mindedness. It is sometimes argued that ethical relativism is itself a normative doctrine insofar as it prohibits interference with the ways of life and moral practices of other cultures. But this degree of tolerance can also lead to moral indifference and even nihilism. A relativist does not have the certitude to condemn genocide, torture, child abuse, etc., so long as it is argued that such actions are part of the fabric of the social life of some cultures and peoples. Furthermore, if ethical relativism is true then even tolerance and noninterference could not be seen as a universal ethical value and could not be recommended to those who do not already embrace it.

4. *New Developments*

Relativism has gained support from two recent philosophical developments. A new variety of relativism has been introduced through the work of the neo-pragmatist philosophers, the most prominent among whom is Richard Rorty. Although Rorty denies the charge of relativism he does believe that truth, rationality, and objectivity are all defined only in terms of the practices of a community. Furthermore, he argues that no external questions can be asked about the relationship between the practices of a community and a reality that exists independently of it. 'Objective truth,' then, for Rorty, is simply the best idea we currently have about how to explain things. The argument supports relativism insofar as it identifies truth with (best) views current in a society at a given time.

The work of postmodernist cultural and literary theorists has been hugely influential in establishing relativism as one of the strongest intellectual currencies of recent times. Jacques Derrida's proposed deconstruction of Western ideas of logic, truth, and rationality has been seen as a harbinger of a new intellectual era free of the logocentrism and phallocentrism of Western rational intellectual traditions. Derrida, like Rorty, has denied that he is a relativist but his own and his fellow postmodernists' writings lend themselves easily to relativistic interpretations.

5. *Arguments Against Relativism*

Relativism has been challenged on several grounds.

5.1 *The Self-refutation Argument*

The most common criticism leveled against relativism is the charge of inconsistency. The argument goes back to Plato and the main claim is that relativism, at least in its stronger form of relativism about truth, is either

inconsistent or incoherent because of the dubious status of the claim that 'truth is relative.' It is suggested that if 'truth is relative' is itself true unconditionally, then there is at least one truth which is not relative, and hence relativism is not true. If, on the other hand, 'truth is relative' is true only relative to the relativist's framework, then the claim is devoid of significance. To avoid inconsistency a relativist should accept that 'truth is relative' is itself only relatively true, i.e. true only for people who suggest it, but then she or he would not be in a position to try and convince a non-relativist of the veracity of the claim. The relativist is condemned to silence.

5.2 Argument from Translation

A new type of criticism of relativism has surfaced in recent years directed at the intelligibility of the claim that one and the same statements may be true for one linguistic community and false for another. Donald Davidson, one of the main advocates of this position, argues that all types of relativism rely on the assumption that there can be radically divergent conceptual schemes or systems of belief. In order to ascertain that a system of belief or cognitive scheme diverges from ours in a significant manner, he argues, we first need to identify such a scheme. In order to identify something as a belief we must be able to interpret it (or translate the sentence expressing that belief into our language). But, to translate a sentence or to assign meaning to it, we will have to regard that sentence, or the belief system underpinning it, as true or in agreement with us. A correct understanding of the beliefs and other propositional attitudes of a person, Davidson maintains, leads to the conclusion that most of her beliefs must be true. If this argument is right, then relativism cannot even get started, because arguments for relativism are built on the premise that there are, or can be, fundamental cultural and conceptual differences between human beings. If this assumption is untenable, then so is relativism.

5.3 Evolutionary Arguments

The origins of this approach can be traced to Quine and his infamous dictum that natural selection guarantees that organisms would know many truths and abide by the laws of logic and rational cognitive procedures or become extinct. The philosopher of mind, Daniel Dennett, following Quine, has argued that natural selections guarantee that most of an organism's beliefs will be true and most of its strategies rational. 'True' and 'rational' have to be interpreted univocally, therefore we have empirical grounds for rejecting the possibility of relativism about truth and rationality.

Relativism has survived 2,000 years of attempts to bury it. Its perennial appeal and periodic fashionability are due to its ability to capture and express an essential insight, be it in an exaggerated fashion, about the human condition. The insight that we live in a world of pluralities and do not have access to univocal answers to our diverse problems.

See also: Ethics and Values; Evans-Pritchard, Sir Edward E (1902–73); Fieldwork in Social and Cultural Anthropology; Intentionality and Rationality: An Analytic Perspective; Irrationality: Philosophical Aspects; Kant, Immanuel (1724–1804); Kuhn, Thomas S (1922–96); Nietzsche, Friedrich (1844–1900); Paradigms in the Social Sciences; Pluralism; Realisms and their Opponents: Philosophical Aspects; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects; Wittgenstein, Ludwig (1889–1951)

Bibliography

Bernstein R J 1983 *Beyond Objectivism and Relativism: Science, Hermeneutics, and Praxis*. University of Pennsylvania Press, Philadelphia

Bloor D 1976 *Knowledge and Social Imagery*. Routledge & Kegan Paul, London

Davidson D 1984 *Inquiries into Truth and Interpretation*. Oxford University Press, Oxford, UK

Feyerabend P 1987 *Farewell to Reason*. Verso, London

Harman G, Thomson J J 1996 *Moral Relativism and Moral Objectivity*. Blackwell, Oxford, UK

Harre R, Krausz M 1996 *Varieties of Relativism*. Blackwell, Oxford, UK

Hollis M, Lukes S (eds.) 1982 *Rationality and Relativism*. MIT Press, Cambridge, MA

Jarvies I C 1984 *Rationality and Relativism: In Search of a Philosophy and History of Anthropology*. Routledge, London

Krausz M (ed.) 1989 *Relativism: Interpretation and Confrontation*. University of Notre Dame Press, Notre Dame, IN

Krausz M, Meiland J W (eds.) 1982 *Relativism: Cognitive and Moral*. University of Notre Dame Press, Notre Dame, IN

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, (2nd edn., enlarged, 1970)

Laudan L 1990 *Science and Relativism: Some Key Controversies in the Philosophy of Science*. University of Chicago Press, Chicago

Margolis J 1991 *The Truth About Relativism*. Blackwell, Oxford, UK

Norris C 1997 *Against Relativism: Philosophy of Science, Deconstruction and Critical Theory*. Blackwell, Oxford, UK

Plato 1961 *Theaetetus* [trans. Cornford F M]. In: Hamilton E, Cairns H (eds.) *The Collected Dialogues of Plato*, Bollingen Series. Pantheon Books, New York, pp. 845–919

Putnam H 1981 *Reason, Truth and History*. Cambridge University Press, Cambridge, UK

Putnam H 1992 *Renewing Philosophy*. Harvard University Press, Cambridge, MA

Quine W V O 1984 Relativism and Absolutism. *Monist* **67**: 293–6

Rorty R 1991 *Objectivity, Relativism, and Truth: Philosophical Papers*. Cambridge University Press, Cambridge, UK, Vol. 1

Siegel H 1987 *Relativism Refuted: A Critique of Contemporary Epistemological Relativism*. D. Reidel Publishing Company, Dordrecht, The Netherlands
Whorf B L 1956 *Language, Thought and Reality*. MIT Press, Cambridge, MA
Williams B 1985 *Ethics and the Limits of Philosophy*. Harvard University Press, Cambridge, MA
Winch P 1964 Understanding a primitive society. *American Philosophical Quarterly* **1**: 307–24
Winch P 1958 *The Idea of a Social Science and its Relation to Philosophy*. Routledge & Kegan Paul, London
Wong D 1991 Relativism. In: Singer P (ed.) *A Companion to Ethics*. Blackwell, Oxford, UK

M. Baghramian

Relativism, Pluralism, and Skepticism (Philosophical Perspective): Cultural Concerns

Few terms in our critical language are as ambiguous as pluralism and relativism. Under diverse definitions, their relationship could be said to range all the way from twins to deadly enemies. Both have been accused by some monists as furthering rank skepticism or nihilism. Meanwhile, defenders have vigorously denied that either term necessarily entails utter skepticism, or Pyrrhonism, or what Bernard Williams has called vulgar relativism. This article pursues pluralism as the best escape route from naive dogmatism and naive (hyperskeptical) relativism. (For a useful, though somewhat superficial, history of skepticism, especially in the Renaissance, see Popkin 1979.)

Almost every serious critic would agree that some extreme forms of either term can lapse into the kind of skepticism or nihilism that threatens genuine intellectual inquiry or achievement. Genuine inquiry is doomed if no claim to 'truth' has any validity except for the claimer, and no one thus has any real reason to listen to the arguments of anyone else. Fortunately, more and more postmodernists, some of whom for a while edged towards such utter skepticism, have attended to diverse cogent attacks on the 'vulgar' or 'total' forms of doubt. They have acknowledged that all arguments for total nihilism are paradoxically self-destructive; no one can offer a coherent defense of utter skepticism without violating that defense by claiming to offer, unskeptically, this one coherent defense. What is more, the would-be utter skeptics usually reveal many other deep commitments that contradict their skeptical claims. Meanwhile, many have argued that the term relativism need not be identified with nihilism (Pepper 1942, Williams 1972, Shweder 1989, Goodman 1978, Rorty 1989, Krausz 1989.

Cultural relativism—'though truths exist, they are always relative to some cultural placement'—has been much more successful in resisting attacks from defenders of hard truth of various stripes. Its defenders have ranged over a vast diversity of arguments. Some have moved dangerously close to utter skepticism by implying, but never stating, that whatever any given culture claims is true must be true, even if it entails mass genocide or the ritual torture and murder of the innocent. At the opposite extreme, some have labeled our 'scientific,' 'rational,' 'antisuperstitious,' 'intellectual' culture as inherently, self-evidently superior to all the naive, primitive, 'believers' in claims that violate scientific convictions. As Shweder describes the way too many rationalists, including professed cultural relativists, have viewed the world:

> Nature is mindless, objective, and visible. It is empty of such unseen, unobservable, meta-physical things as god, sin, obligation, value, morality, and so on. If such things exist at all, they exist only as reality-posits in the mind. And if they exist only in the mind, they are not objective and thus ought not be (sic) allowed to be constraining. If human beings feel constrained by such things, it is only because, not yet realizing their essential nature (self-determination), they do not distinguish between truth and poetry, confusing, quite irrationally, external reality with what exists nowhere else but in their minds (Shweder 1989, p. 121).

Even when we follow Shweder and rightly reject such extremes and decide that truths do exist, though in multiple forms, we are still faced with the complex problem of how to deal with that multiplicity. For cultural pluralists who refuse to dodge the problem, the best possible method is to search for common ground that underlies surface differences. (Sometimes they—we—stumble upon another route, after the probing: 'By studying my opponent carefully I have learned that I was wrong and he/she was right; now my pluralism includes memory of my previous error.') When for example we discover that every culture, including the most rationalistic versions of our own, exhibits some form of embrace of an unprovable 'transcendent' 'world' or 'reality' or God, we no longer have good reason to consider religious commitments absurd: they are shared by all cultures. Again to quote Shweder 1989:

> In that (our) post-Nietzschian world God is not dead; only positivism and monotheism are dead. Polytheism is alive and well. Its doctrine is the relativistic idea of multiple objective worlds [including the world of scientific rationalists], and its commandment is participation in the never-ending process of overcoming partial views (pp. 132–3).

1. *Plurality of Thought-modes*

Once we embrace that form of cultural pluralism, value-terms like 'dead' and 'partial' turn our problem in the direction of philosophical inquiry: how to deal with plurality not just of culture systems but of

thought systems, including value judgments. Even those 'rationalist' philosophers who are totally anti-relativist, antiskeptical, exhibit a plurality of conflicting thought-modes. Too many have faced that conflict with either passionate commitment to 'the one true system' (such as the positivism that most of us, including most professional philosophers, by now question) or the philosophical skepticism that Richard Rorty and Stanley Fish and many others have at times edged toward (even while claiming not to).

What has been too often ignored by those who justifiably attack the first of these paths, dogmatic monism, is the problem of how we are to deal with that plurality of 'contexts' or 'belief systems,' or 'structures of thought' or 'thought-modes' to which all of our assertions are 'related.' Every thinker has to admit that all thinking takes place in a conceptual context in which all assertions are interrelated, interdependent for their full understanding; even claims to factuality have different meanings in different thought-modes. Yet only careless extremists can deny that some 'contexts' are better than others: truer, or 'fuller,' or less vulnerable to refutation.

The pluralistic problem can thus be summarized like this: a devoted, honest Nazi who asserts that the commitment to the one superior race with its superior values justifies genocide, employing as evidence some 'facts' about past misbehavior of Jews and amazing achievements of Aryans like Beethoven, is asserting facts that though in one sense are context relevant, cannot be claimed by an anti-Nazi to be just plain nonfactual. But how, then, can one assert, as we all would, that the entire context of such Nazi thinking is just plain wrong, without falling into dogmatic assertions that are finally totally dependent on an equally unprovable cultural context? How finally do we avoid a form of relativism that grants the Nazis equal intellectual status?

2. *Truth Exists—in Multiple Forms*

The only possible answer is to develop some form of 'systems pluralism' that allows for distinctions among a plurality of defensible systems and a plurality of inferior, less defensible systems. While there is a genuine plurality of 'truths,' genuine truths that are of equal validity and systems that are universally true in the sense of being irrefutable, no one of them is universally true in the sense of being easily either translatable to other 'systems' or immune to seeming refutation from other 'truths.' But that collection of irrefutable context-systems or thought-modes does not mean that other systems or modes cannot be shown to be inferior. The Nazi illustration reveals that one particular belief system can be finally rejected, first because it will be condemned by every other major mode of thought, and second because its disastrous consequences are revealed by history, as they finally were to Heidegger when he faced the consequences of his early Fascist thinking.

The history of philosophy, like the history of various cultural relativisms, can be seen as an implicit demonstration both that genuine truth exists, in a variety of defensible forms, and that the diverse forms are never easily intertranslated in intelligible form. Indeed, in most versions, 'believers' believe that they can, in using their own version of truth, refute the other versions (see Russell 1948, as an example of careless refutation of all modes but his own).

The clearest support for a plurality of legitimate thought-modes (along with a larger plurality of illegitimate ones) is found in the ways in which the major, great, surviving philosophies conflict, at least on the surface, and yet reveal immense capacity both to seem to refute other systems and to genuinely resist refutation from the others. In a short discussion, it will be impossible to demonstrate this case fully (for extended treatment, see Booth 1979, Garver and Buchanan 2000, Booth 2000, McKeon 1951).

3. *McKeon, the Greatest Pluralist*

To me by far the most penetrating of the many efforts to face plurality is that of the much-neglected Richard McKeon. As McKeon probed major thinkers throughout history, he more and more aggressively asserted that there are a small number of absolutely irrefutable (yet seemingly conflicting) thought modes, while there are many totally refutable modes and many other half-defensible but truncated modes. Having pursued every major thinker to the depths, he developed the notion of what his best defender, Walter Watson, has called 'reciprocal priority'—what might be called 'reciprocal irrefutability' (see Watson 1993, the best single introduction to reciprocal seeming refutation and actual nonrefutation; it is based on the lifetime work of Richard McKeon).

McKeon's classification of defensible modes became increasingly complex throughout his life, and it cannot be summarized here, as is evident in his final, most complicated chart, which was not published until after his death (see Fig. 1).

Such complexities may well obscure the validity of his basic project, which was to reveal how first-class philosophers, irrefutable philosophers, have approached multiple truths with valid thought-modes that differ in many conflicting ways, especially in four major matters: (a) what is the proper subject matter, or definition of subject matter ('Modes of fact') of what we should think about; (b) what 'unprovable' first principles, or notions of the 'modes of being,' should one proceed from, taken as unquestionable; (c) what methods are the proper methods ('modes of thinking') for solving all problems; and (d) what is the point of it all, what are the true goals of human thinking. In my view his philosophical multiplicities could well serve as

Modes of philosophic inquiry

Modes of Being *Being*	*Modes of Thought* *That which is*	*Modes of Fact* *Existence*	*Modes of Simplicity* *Experience*
Being and Becoming	Assimilation and Exemplification (models)	Reality and Approximation	Categories of Thought (Ideas and presentations)
Phenomena and Projections	Discrimination and Postulation	Process and Frame	Categories of Language and Action (Symbols and rules)
Elements and Composites	Construction and Decomposition (constituents)	Object and Impression	Categories of Things (Cognition and Emotion)
Actuality and Potentiality	Resolution and Question (causes)	Substance and Accident	Categories of Terms

SCHEMA OF PHILOSOPHIC SEMANTICS

Principles	*Methods*	*Interpretations*	*Selections*
Holoscopic	*Universal*	*Ontic*	
Comprehensive	Dialectical	Ontological	Hierarchies (transcendental)
Reflexive	Operational	Entitative	Matters (reductive)
Meroscopic	*Particular*	*Phenomenal*	
Simple	Logistic	Existentialist	Types (perspective)
Actional	Problematic	Essentialist	Kinds (functional)

BASIC DIVISIONS OF PHILOSOPHY

Theoretic	Physics	Philosophy	Logic
Practical	Ethics	Poetry	Rhetoric
Poetic	Logic	History	Grammar

BASIC PROBLEMS

Whole	Universal	Reality	One
Part	Particular	Process	Many

Figure 1
McKeon's final classification of defensible modes (Printed posthumously in *Selected Writings of Richard McKeon*, Vol 1, 1998, p. 218, University of Chicago Press)

a rescuer of those who fear that cultural relativism will inherently destroy such terms as truth or objectivity or reality.

3.1 Definition of the Subject Matter: 'Things,' 'Thoughts,' 'Words,' 'Actions'

3.1.1 Things. The most famous of the classical philosophers agreed that what we should pursue is reality, 'what really is,' substance, being, actuality: whether air, water, fire, atoms, or ideas. Plato always operates with a strong conviction that there are some general truths, or principles, or ideas awaiting our clumsy efforts to discover them. His definition of the subject matter (the world, the universe, the nature of things, the One) is always implicitly a world in which Truth, Goodness, and Beauty are interrelated, harmonizable: a One that can, in theory, finally harmonize it all, even though no one thinker can capture it. Many modern cosmologists begin with a surprisingly similar pursuit of a supreme truth that is 'really there'—nature and its laws: this or that final theory will explain it all. For all of these, we should begin with some picture of the real, and then move from it to whatever questions arise about the rest of life.

3.1.2 Thoughts. Again and again, as McKeon traces brilliantly, philosophers have concluded that you cannot really begin there, with Being, struggling with ontological questions, because the picture of real reality always depends on how we think about it, and that leads to a take-over by one or another form of epistemology. Follow Kant and begin your thinking by thinking about thought, not about things or substance.

3.1.3 Words. That shift from subject matter to mind has again and again broken down as philosophers concluded that how we think depends on the

language we think with, so we should begin not with thought but with language. Especially in the twentieth century, more and more thinkers took their beginnings in language. For just about everybody thought had to begin with language. So-called analytical philosophy took over, as it had done for some 'skeptics' and 'relativists' throughout the ages, for a while, and later the deconstructionists produced a seeming revolution by claiming that language itself has no real substance.

3.1.4 Actions. Finally, the linguistic revolution has repeatedly been revolted against, by those who see action as the proper subject, the place to begin. Pragmatism in various forms has been seen as the true subject of true thinkers; inevitably they work back up from language to questions about abstract reality, ontology, epistemology, and linguistic analysis, but they always place action—with success and failure of action—at the top of the definition of subject.

3.2 Principles about 'Reality'

3.2.1 Comprehensive whole. Many philosophers, like Plato and some modern cosmologists, put as top priority a conception that there is some whole, some totality, some supreme collection or unity of all truths. Reality consists of that comprehensive unity or totality.

3.2.2 Multiple problems produced by the incomprehensible whole. Others like Aristotle while still in one sense pursuing the whole, see the philosopher's basic principle as to 'pursue truth about this or that particular aspect, or problem,' concentrating on solving the problems. Many such pluralists see no hoped-for 'top' but a collection of diverse subjects and problems; there may be a 'one' uniting them, but for us the world is multiple, created by and consisting of many causes and possible solutions.

3.2.3 Least, parts: 'atoms,' 'sensations,' etc. In contrast, philosophers like David Hume and many modern scientists put as the 'top' of their reality a kind of bottom: elements or particles or epistemological certainties from which they can reason upward to larger truths. Like some particle physicists, they hope to discover, by probing the depths, truths never even dreamed of by those who pursue the 'heights,' and those truths—whether about our most minute sensations or the most minute physical particles, will be totally irrefutable from the other perspectives.

3.2.4 Operational, pragmatic problems: truth is what works. Finally, philosophers who, like Cicero and other rhetoricians, put as top priority the successful operation of the practical world, beginning with questions of just what our pragmatic goals are, can build from their pursuit truths about life that may conflict irrefutably with many of the truths that the other three pursuers have proclaimed.

To say all this does not mean that the four choices of first principles will not reveal, down the line, agreed-upon truths. Many of the undeniable truths that each mode uncovers will be considered undeniable by each of the others. For example, it is always wrong, in principle, in all of the defensible systems, to violate the moral principles that you think your system has revealed, and many of them overlap those revealed in other systems. It is always wrong, in all systems, to deliberately, dishonestly, fabricate evidence to support your system. And there are many other overlapping moral principles, such as the relation of Cicero's 'Civility' to Aristotle's 'Ethics' and Plato's 'goodness' and Hume's moral standards, and the moral commands issued by John Stuart Mill's utilitarianism.

3.3 Methods

Major thinkers have always disagreed about just what methods are effective in yielding proof. Depending in part on their first principles and their definitions of the subject matter, they have ranged from the absolutely deductive synthesizers to the absolutely inductive analysers, with some pragmatists and linguists urging some middle ground of cooperative thinking and testing, not induction or deduction but what could be called co-duction. Plato always falls back on dialectic, even when exploring rhetoric and logic. Aristotle insists on a plurality of methods to meet a plurality of subjects and problems. Consider the contrast between that and the method of Hume, or of other thinkers who begin with the assumption that the true method of inquiry is to work analytically with the indivisible parts of the universe, not some ideological whole, and then to reason analytically 'upward' from those elementary parts. Whether or not such methods lead ultimately to this or that ontology or metaphysics, it is the right way to work.

Thus, without troubling too much about whether one will ultimately find a total view of everything, analytical reductionists know that they are dealing with reality, tough hard reality of the multiple parts. The more defensible among them generally have accorded a place for diverse methods, but some partial or flawed thought-modes have tried to reduce it all to one grand method, as modern scientists too often reduce everything to empirical demonstration of only one of Aristotle's four causes: the efficient, or mechanistic, or 'horizontal' bump-and-grind reality.

3.4 Goals

The diversity of goals is to a large degree implicit in all these other varieties. The goal of thought is to improve truth, for the sake of truth. No, the goal of thought is to improve conditions in the world, for the sake of what Plato called goodness. No, the goal of thought is to pursue beauty, proper kinds of eros, religious ecstasy, some transcendent escape from shitty reality, salvation.

For the monist or dogmatist, the goal is to demonstrate that there is only one right view of principles, subject choice, methods, and goals. For the pluralist who rejects vulgar relativism or skepticism, the goal is both to pursue some one or another of the major, irrefutable modes or, even better, to pursue what is the common ground among the defensible modes, thus demonstrating their defensibility.

The suggestion, in this cursory summary of just one brand of philosophical pluralism, is that all thinkers in all subjects—philosophical, religious, anthropological, political, literary—could reduce pointless debates about truth and objectivity if they thought harder about just which definitions, principles, methods, and goals they are depending on.

4. The Plurality of Pluralisms

One unavoidable problem for anyone who chooses the pluralist route is the immense complexity of modes and the rivalry of pluralisms in dealing with the modes. The rivalries often lead many of us pluralists to retreat, as we sometimes observed McKeon doing, to some version of 3.2.4, a loose-jointed 'operationalist,' 'rhetorical,' or 'practical' (pragmatic) corner of the chart, with an emphasis on how to improve the world by improving thought about it. But even as we do so we have to recognize that other thinkers, in contrasting modes, will be discovering truths that we ignore.

We thus end with a plurality of pluralisms, some more defensible than others, just as cultural relativists end with a plurality, some more defensible than others. All of the pluralisms are 'relativistic,' in the sense that within each system every claim, even a claim of fact, is relatable and to some degree dependent on the total thought-mode. Conclusions within a given mode are testable within that mode, and will often even prove to be testable and acceptable within other modes. But no one system is totally refutable by any other system, even when all other systems agree that one system is flawed.

Thus, serious explorers of the issues 'relevant' here have two main choices, whenever they find that no one controversy seems resolvable: retreat to the study of some one cultural system or thought-mode, ignoring the others, or pursue the mission of finding common ground among the conflicting systems. Only this second choice can avoid dogmatism, on the one hand, and on the other skepticism, reductionism, and nihilism, both directions producing absurd destruction of meaningful inquiry.

See also: Cultural Psychology; Cultural Relativism, Anthropology of; Pluralism; Pluralism and Toleration; Relativism: Cognitive; Relativism: Philosophical Aspects; Value Pluralism; Wisdom, Psychology of

Bibliography

Bakhtin M M 1981 *The Dialogic Imagination: Four Essays* (ed. Holquist M. Trans. Holquist M, Emerson C). University of Texas Press, Austin, TX

Booth W C 1979 *Critical Understanding: The Powers and Limits of Pluralism*. University of Chicago Press, Chicago

Booth W C 2000 Richard McKeon's pluralism: The path between dogmatism and relativism. In: Garver E, Buchanan R (eds.) *Pluralism in Theory and Practice*, 1st edn. Vanderbilt University Press, Nashville, TN, pp. 212–30

Dewey J 1929 *The Quest for Certainty* (Gifford Lectures, 1928–29). Minton, Balch, & Co., New York

Fish S 1999 *The Trouble with Principle*. Harvard University Press, Cambridge, MA

Garver E, Buchanan R (eds) 2000 *Pluralism in Theory and Practice*, 1st edn. Vanderbilt University Press, Nashville, TN

Goodman N 1978 *Ways of World Making*. Hackett Publishing, Indianapolis, IN

Herskovits M J 1972 *Cultural Relativism: Perspectives in Cultural Pluralism*, 1st edn. Random House, New York

James W 1907 *Pragmatism: A New Name for Some Old Ways of Thinking*. Longmans Green, New York

Krausz M (ed.) 1989 *Relativism: Interpretation and Confrontation*. University of Notre Dame Press, Notre Dame, IN

McKeon R 1943–44 The philosophic bases of art and criticism. *Modern Philology* **41**: 65–87, 129–71

McKeon R 1951 Philosophy and method. *Journal of Philosophy* **48**: 653–82

McKeon R 1998 *Selected Writings of Richard McKeon* (McKeon Z and Swenson W eds.). University of Chicago Press, Chicago, Vol. 1

Paul E F, Miller F D Jr, Paul J (eds.) 1994 *Cultural Pluralism and Moral Knowledge*. Cambridge University Press, Cambridge, UK

Pepper S C 1942 *World Hypotheses: A Study in Evidence*. University of California Press, Berkeley, CA

Popkin R H 1979 *The History of Scepticism, from Erasmus to Spinoza*. University of California Press, Berkeley, CA

Rorty A O 1989 Relativism, persons, and practices. In: Krausz M (ed.) *Relativism: Interpretation and Confrontation*. University of Notre Dame Press, Notre Dame, IN, pp. 418–40

Rorty R 1991 *Objectivity, Relativism, and Truth: Philosophical Papers*. Cambridge University Press, New York

Russell B 1948 *The History of Western Philosophy, and its Connections with Political and Social Circumstances, from the Earliest Times to the Present Day*. Allen & Unwin, London

Shweder R 1989 Post-Nietzschean anthropology: The idea of multiple objective worlds. In: Krausz M (ed.) *Relativism: Interpretation and Confrontation*. University of Notre Dame Press, Notre Dame, IN, pp. 99–139

Tracy D 1975 *Blessed Rage for Order: The New Pluralism in Theology*. Seabury Press, New York
Wahl J A 1920 *Les Philosophies Pluralistes d'Angleterre et d'Amerique*. Alcon, Paris
Watson W 1993 *The Architectonics of Meaning: Foundations of the New Pluralism*, 2nd edn. University of Chicago Press, Chicago
Williams B 1972 *Morality: An Introduction to Ethics*. Harper & Row, New York

W. C. Booth

Reliability: Measurement

Whenever an aspect of human behavior is measured, the question always arises: how good is the measurement process? In the behavioral sciences, conceptual and statistical models for answering that question are generally divided into two major categories: reliability and validity. As a first approximation, one may think of validity as concerned with whether the right construct is being measured. Reliability as concerned with the issue of how well we are measuring whatever it is that is being measured (regardless of whether or not it is the right quantity to measure). When choosing a measurement method, one must consider both of these issues; the total error of measurement comes from both the unreliability and invalidity of a measure. While this article primarily concentrates on reliability, it also discusses the measurement process as a whole, and the relationship between reliability and validity issues.

As with many terms, the scientific and psychometric uses of the word 'reliability' derive from, but are different than, the everyday use of the term. In everyday life, a car would be considered reliable if it started every time. If the word were used in the technical sense, one would also say that a car was reliable if it never started; it would only be unreliable if it sometimes started, and sometimes did not. The technical term, in other words, is more concerned with consistency of behavior, regardless of where the behavior is on a scale. The main question is whether one gets the same answer every time a measurement of an individual is made. Whether that measurement is then useful is a question of validity; a measurement procedure could be perfectly reliable but useless.

Reliability theory generally begins with a conceptual framework, which is then translated more formally into a statistical model. Most reliability theory is concerned with a particular type of measurement instrument, multiple-item tests, and some theory is applied in an even more limited setting of items that are scored dichotomously as true or false. The three most commonly used models are (a) classical test theory, based on factor analysis; (b) generalizability theory, based on analysis of variance (ANOVA); and (c) item response theory, based on 'latent trait models.' Other statistical methods are usually needed for dealing with measurements that are either not tests, such as observations of behavior, or are not multiple-item tests. While much of measurement theory (and terminology) originally derived from consideration of ability tests, the theory applies equally well to many other areas of behavior, such as personality assessment.

1. *Classical Test Theory*

Classical test theory (CTT) is based on the assumption that a single underlying dimension (a trait or skill) is being measured, and that every person has a single true score, T, on that dimension. The underlying dimension or test domain may be more or less 'pure'; for example one might think of skill in (a) doing multiplication of 3-digit numbers, (b) solving algebra word problems, or (c) solving college mathematics problems. The important consideration is that all items on the test are measuring this skill or trait, and that the skill or trait is well defined. Each person's observed test score, X, is assumed to be determined by his or her true score T plus some error of measurement, e, which would either raise or lower the observed score relative to the person's true score. The basic equation of classical test theory is therefore:

$$X = T + e \tag{1}$$

in which the quantities are as defined above, and the errors are assumed to have a mean of zero and to be uncorrelated with (unrelated to) T.

What exactly does T represent? Conceptually, the definition requires a bit of imagination, and tolerance for ambiguity: imagine a person being tested over and over again, using different sets of items selected from the collection of all items designed to measure the underlying construct. (A test may consist of any number of items, including the special case of a single item, such as an essay.) The average of all these scores (more technically, the expected value) is the person's true score T.

The first major consequence of this model is that the variation among observed test scores (σ_X^2) is the sum of the variation in true scores (σ_T^2) and the variation in the errors of measurement (σ_e^2); that is,

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \tag{2}$$

The optimal case is when the error for each person is zero; then there is no variation in errors, so $\sigma_e^2 = 0$. Consideration of this equation leads to the definition of reliability as the proportion of total variance in

scores that is attributable to true variation rather than error variation; that is,

$$\text{Reliability} = \sigma_T^2/(\sigma_T^2+\sigma_e^2) = \sigma_T^2/\sigma_X^2 \qquad (3)$$

For example, if the true score variance were 8, and the error variance 2, then the total variance would be 10, and the reliability would be $8/10 = 0.8$. One useful interpretation of this number is that it is the square of the correlation between true scores (T) and observed scores (X).

Among the interesting implications of the CTT model is that a test can be made more reliable by making it longer; adding items increases the true score variance at a faster rate than the error variance. One aspect of test design is determining how long a test must be to have a reliability great enough for the intended uses of the test.

Another implication of the model is that the reliability of a test is not a single number, but depends on how individuals were sampled to take the test. If a broad range of people are sampled, so that the true score variance is large, then the reliability will be greater than if people with a more limited range of true scores are sampled. (This is because the error variance would be the same in each case.) For example, if a test were given to all high school seniors, it would have a higher reliability estimate than if it were only given to seniors who planned to attend a highly selective college or university.

1.1 Estimation of Reliability

Unfortunately, since true score and error variation are not directly observable, the definitional equation (3) above cannot be applied directly. However, this conceptualization and the accompanying statistical model lead to several possible methods for estimating reliability. First, one could randomly divide any multiple-item test into two halves, and see how the scores on one half related to scores on the other half. If a test were perfectly reliable, then each person would get the same score on both halves of the test. Of course, the correlation is never perfect; the observed correlation allows the calculation of an estimate of the reliability of the whole test. Other methods use the correlation among the items to estimate reliability. These methods, including Cronbach's alpha (developed from an idea of Louis Guttman), are based on the same conceptual and statistical principles as the split-half coefficient.

Because each person is measured at one time point, any variability due to changes in that person across time is ignored by these methods. Here is where the definition of true score becomes critical. One can think of 'true' changes in true score (due to learning or forgetting in between testing times); one can also think of more temporary changes, due to illness, noise or distractions in the testing room, and so on. In addition, some characteristics are expected to be variable over time; these include measures of mood. In personality psychology the more labile characteristics are called *states*, while the more stable characteristics are called *traits*. In classical test theory, the temporary effects are included in errors of measurement, while changes in traits are generally not. In spite of this conceptual uncertainty, this suggests another commonly used strategy for assessing reliability: Give the same test to a group of people at two different times, and see whether their scores are closely related. For obvious reasons this is called test-retest reliability. Of course, one assumes that the first measurement does not affect the second measurement, which may or may not be the case, and which must be considered. To minimize the effects of 'true' changes in true scores over time, the time period between testings must be kept relatively short.

A related method uses two or more forms of a test, with each form being constructed in an identical manner, so that the forms will be as nearly alike as possible. If in fact they have the same means, variances, and reliabilities, then they are said to be *parallel tests*. If we administer two such forms to the same group of people (e.g., on two consecutive days), the correlation between scores on the forms is called a parallel test estimate of the reliability.

In addition to measuring the overall performance of the test, we are also interested in how well we have estimated the score of each individual. The *standard error* of the test score is such a number; if small, then the score is a more accurate estimate of ability than if the standard error is large. As an example, suppose a person has a score of 120 on an IQ test, and the test has a standard error of 5 points. Then we can construct a 95 percent *confidence interval* around the estimate by adding and subtracting two standard errors (10 points, in this case); here the interval is from 110 to 130. This means, roughly speaking that there is a 95 percent chance that the interval will contain the person's true score.

Besides considering the reliability of the test as a whole, one must also consider how good each item is in the test. The most commonly used method to assess this in CTT is the item-total correlation, which measures how well the item discriminates among people of varying skill or ability levels. To calculate this, one correlates scores on the total test (usually with the particular item omitted) with scores on the item (with 0 used for an incorrect response and 1 for a correct response). In other words, an item is good if people who score high on the test tend to get the item right more often than do people who score at lower levels on the test. An item would be considered bad if people with low scores on the test were just as likely to get the item right as people with high scores on the test.

In the test construction process, items that have low item-total correlation are generally omitted or rewritten.

Another common item statistic is the proportion of test-takers that get the item right; this is usually called the *difficulty* of the item (though *facility* would be more appropriate). While it is commonly reported, it is of little relevance for reliability; item difficulties are more important when trying to tailor a test for different ability levels, or to compare the characteristics of tests that are supposed to be equivalent (parallel, in technical terms).

1.2 Weaknesses of Classical Test Theory

CTT has several weaknesses that have led to the development of other models for test scores. First, the concept of reliability is dependent on the group used to develop the test. If the group has a wide range of skill or abilities, then the reliability will be higher than if the group has a narrow range of skill or abilities. Thus reliability is not invariant with respect to the sample of test-takers, and is therefore not a characteristic of the test itself; in addition, neither are the common measures of item discrimination (such as the item-total correlation) or item difficulty (percent getting the item correct). As if this were not bad enough, the usual assumption that the standard error is the same for test-takers at all ability levels is usually incorrect. (In some extensions of CTT, this assumption is dropped, but these extensions are not well-known or widely used.) CTT also does not adequately account for observed test score distributions that have floor and ceiling effects, where a large proportion of test-takers score at either the low or high end of the test score range.

CTT has difficulties in handling some typical test development problems, horizontal and vertical *equating*. The problem of *horizontal equating* arises when one wishes to develop another test with the same properties as (or at least with a known relationship to) an existing test. For example, students who take college admissions tests such as the ACT or SAT should get the same score regardless of which version of the test they take. *Vertical equating* involves developing a series of tests that measure a wide range of ability or skill. For example, while we could develop completely independent tests of arithmetic for each elementary school grade, we might instead want to link them to have one continuous scale of mathematical skill for grades 1 through 6. While vertical and horizontal equating is not impossible within CTT, they are much more straightforward with item response theory (IRT).

These problems of CTT are partly due to some fuzziness in the theory (the population to be sampled is usually not considered in any detail in the theory). But they are also due to the failure of most data collection to be carried out on a random sample from any population, let alone a population considered appropriate for the test being investigated. In practice, convenience samples are used; these generally fit some criteria that the investigator specifies, but are otherwise taken as they are available to the researcher. Finally, CTT as originally conceptualized was never intended to address some of the practical testing problems described above.

2. Generalizability Theory

While classical test theory provided a framework that has proved useful in a number of situations, psychometricians realized many of its limitations even as they were developing the theory. Most importantly, they realized that a wide variety of factors could influence observed test scores, including how the items were generated, tests were administered, and items were scored (especially if they were free responses scored by raters). In classical test theory, these factors were all included in the (one and only) error term in the statistical model. To include these factors in their statistical models, psychometricians developed what is known as generalizability theory (GT), which included more than one source of error in test scores.

GT is based on the statistical methods of analysis of variance. Factors thought to have a major influence on test scores are included in a design, and the analysis allows the researcher to determine how much influence each factor has on variation in test scores. GT allows the computation of different estimates of reliability, depending on which factors are considered important for a particular use of the test.

As an example, consider a test of writing, in which each student writes three brief essays, each of which is scored by two raters. GT allows the estimation of how much variation in test scores is due to variation in essay topics, and how much is due to variation in raters.

Like classical test theory, GT is primarily concerned with the behavior of the test as a whole, rather than the performance of components, such as subscores or items. For further details on GT, see *Generalizability Theory*. The classic source on GT is the book by Cronbach et al. (1972). Shavelson et al. (1989) provides a less technical overview.

3. Item Response Theory

Since 1980 there has been a steady rise in the use of item response theory, or IRT, to solve some of the problems that CTT could not adequately deal with. IRT, like CTT, is based on an assumption that a single underlying dimension is being measured. But the basic conceptual idea behind IRT is the relationship be-

tween the score (usually right/wrong) on an item and the underlying dimension, whereas CTT is based on the relationship between total test score and the underlying dimension. This relationship represented by an item characteristic curve, or ICC, shows how the probability of getting the item right (for dichotomous items) changes with the underlying trait. As the probability cannot be below zero or greater than one, the ICC generally has an elongated S shape. In three-parameter models, the items may differ in whether the curve starts at zero for people of very low ability (no guessing) or above zero (if guessing correct is likely); how fast the curve rises (discrimination); and whether the curve is to the left (easy), in the middle, or to the right (hard). In two-parameter models, only discrimination and difficulty are used; guessing is fixed at zero. And for one-parameter (Rasch) models, only difficulty is used; guessing is fixed at zero and discrimination at one for all items.

IRT models solve many problems that were difficult for CTT, including the ability to construct a different test for each test-taker, so that students at all levels of ability can be assessed with items appropriate to their status. Such 'adaptive tests,' generally administered by computers, are becoming standard in many large-scale testing situations in some countries, including the Graduate Record Examination (GRE) and licensing examinations in many fields.

In IRT, the traditional concept of reliability has no place. Instead, the emphasis is on the more basic concept of standard error of measurement; that is, how accurately has a test-taker's true score been estimated? For a fixed test given to a group of examinees, the test characteristic curve shows the standard error for each level of ability. Unlike the simplifying assumption of CTT, the standard error can vary with ability. For example, if a test is relatively easy, the standard error will be smaller for students of low ability levels, because the test is better matched to their ability; higher ability students will not be well-measured by an easy test.

4. *Reliability in Criterion-referenced Measurement*

In some measurement situations the primary interest is in determining whether or not a person meets a predetermined standard or criterion. These judgments are sometimes made for individuals (e.g., is my student Karol reading at grade level?), and at other times for groups (e.g., what proportion of eighth-graders are reading at or above grade level?) In everyday life, drivers' license test provides a familiar example of a criterion-referenced test. In education, the theory of criterion-referenced measurement was developed to provide suitable technical scaffolding for making such judgments.

Again using the concept of consistency as a basis for conceptualization, reliability of criterion-referenced measures is concerned with issues such as whether an individual would have received the same classification if tested by different raters, on a different form of the test, or on different days. It is clear that the consistency will depend on how close the individual's true score is to the cut-point determining mastery. If a person is well above or below the cut-point, decisions will tend to be more consistent than for people near the cut-point. In the extreme, if a person's true score is exactly at the cut-point, then that person will pass half the time and fail the other half, no matter how well the test is constructed.

Because of the 'yes/no' nature of scoring, criterion-referenced test theory is somewhat different than the usual classical test theory, but only because some different statistical methods are appropriate for this type of data. Criterion-referenced measurement is also more obviously concerned with decisions made on the basis of a test; other test theory models are only implicitly concerned with such decisions. For more information about criterion-referenced measurement, see Hambleton et al. (1978), Harris et al. (1974), and Popham (1978).

5. *Issues for Observational Measures*

While many characteristics can be assessed through tests and questionnaires, others are better assessed through observation of behavior. Behavior can be assessed through direct observation, or it can be recorded for later coding into categories. The recording can be more or less direct, ranging from video and audio devices to the memory of observers about past behavior (where the recording is in the mind of the observer).

Behavioral observations tend to take place in 'real world' settings, and generally (but not always) under uncontrolled conditions. Unlike tests, where those taking them know and expect that their behavior will be observed, in natural settings one can imagine a range of possibilities for reactivity to the act of being observed. The more unobtrusive the observation, the more confidence one has in the quality of the measurement (though this generally would be classified as affecting validity, rather than reliability). The classic reference on unobtrusive measurement is Webb et al. (1966).

Some issues tend to arise much more frequently in the case of observational measures than with tests and questionnaires. One such issue is the decision of whether to consider a characteristic being assessed as categorical or continuous. For example, one might want to say that a person is either depressed or not, rather than being given a rating on a (nearly) continuous scale from not at all depressed all the way

through to extremely depressed. In such a case, the underlying trait is categorical, and different statistical methods are required. The most commonly applied method in this case is latent class analysis, which has been used in settings as diverse as determining stage of development of children, psychiatric diagnosis, and medical diagnosis.

One also tends to see different types of observed measures with observational data. Sometimes a variable has unordered categories, as would be the case if one classified a student in a class as sitting quietly, sitting but making noise, or being out of his or her seat. Other times a variable may be a count or a rate, such as the number of times a student is out of his or her seat, or the number of such times divided by the amount of time during which he or she was observed.

Typically there are multiple observers of behavior, or multiple coders of recorded behavior, or both. (This can, of course, occur for certain types of tests, not just in behavioral observations. Essay tests are a common example, because multiple raters are often used. Similarly, portfolio assessments of students' work are often based on the judgment of multiple raters.)

These considerations present additional problems that affect the reliability of observational measures. While consistency remains the primary concern, and the reliability (and validity) of behavioral observations depends on many of the same factors as for tests, additional factors for observational methods include length of the period of observation, the conditions under which observations are made (including irrelevant stimuli), and the observers, raters, and coders used.

As one might infer from the discussion of generalizability theory (GT), observational measures usually have more than one possible source of error, unlike what is assumed in classical test theory. In addition to inter-observer agreement problems, observers can make their observations at different times, in different settings, and under different conditions; all of these contribute to the error of measurement. GT allows researchers to estimate how much variability in the observations is due to each of these potential sources of error.

These differences from typical test data have additional implications for the statistical theory and data analysis methods generally used with observational data. When data come from observers, the issue of reliability is primarily a matter of whether two or more observers would come to the same conclusion (or, if a rating, the degree to which they agree), and whether the conclusion depends on when and for how long the observation occurs. For variables that fall into categories, one apparently simple method is to calculate the proportion of instances in which two raters or observers agree with each other about the category into which a particular observation should be placed. Unfortunately, the simplicity is only apparent, because for behaviors that are either rare or very common, a high percentage of agreement can occur by chance. For example, if a student is classified as out of seat on 10 percent of occasions by each observer (and in seat the other 90 percent), then the observers would agree 82 percent of the time even if they were responding independently of the actual behavior. Therefore, such measures are typically corrected for chance agreement. Additional complications arise when there are three or more observers, and when observers do not rate or observe for the same period of time.

Details on statistical methods suitable for analyzing observational data are found in Suen and Ary (1989).

6. *Relationship of Reliability to Validity*

In classical test theory, one cannot have a validity of a test that is higher than the square root of the reliability. This has led some to conclude that reliability is more important than validity of a measure, because a test cannot be valid if it is not reliable. While a test cannot be valid if its reliability is zero or extremely low, it is possible for one test to have a higher validity coefficient than another, even though the first is less reliable. Suppose that two tests (or other measures) are not actually measuring the same true score, because they are measuring different traits. Then the one with a lower reliability can have a higher validity if the trait it measures is more closely related to the outcome that the tests are meant to predict. For example, a paper-and-pencil test of rock-climbing skill might have greater reliability than a performance test done on a climbing wall in a gym, but the latter might still have greater validity as a predictor of performance might in an actual rock-climbing situation.

The (invalid) conclusion about the relationship of reliability and validity has often caused a strain between quantitative researchers and qualitative researchers. Qualitative researchers often use measurement procedures that have lower reliability than the procedures devised by quantitative researchers, but the qualitative researchers prefer these less-reliable measures because they are measuring what they want to measure. It is quite possible that in many cases they are right, and their less reliable measures have greater validity than certain measures that are more reliable.

7. *Suggestions for Further Reading*

Martin and Bateson (1986) is a short introductory work on measuring behavior. The two most popular undergraduate texts on measurement are those of Anastasi (1988) and Cronbach (1990). At a higher level are texts by Crocker and Algina (1986) and Allen and Yen (1979), and the general reference work edited by Linn (1989). McDonald (1999) is a very new text that provides a unified treatment of many issues in test theory. At a very advanced level, the classic text on

both CTT and IRT is Lord and Novick (1968). A more modern text on IRT is Hambleton and Swaminathan (1984); a simpler version is Hambleton et al. (1991).

See also: Factor Analysis and Latent Structure: IRT and Rasch Models; Generalizability Theory; Measurement Error Models; Measurement Theory: Probabilistic; Scaling and Classification in Social Measurement; Unobtrusive Measures.

Bibliography

Allen M, Yen W 1979 *Introduction to Measurement Theory*. Brooks/Cole, Monterey, CA

Anastasi A 1988 *Psychological Testing*, 6th edn. Macmillan, New York

Cronbach L J 1990 *Essentials of Psychological Testing*, 5th edn. Harper and Row, New York

Crocker L M, Algina J 1986 *Introduction to Classical and Modern Test Theory*. Holt, Rinehart, and Winston, New York

Cronbach L J, Gleser G C, Nanda H, Rajaratnam N 1972 *The Dependability of Behavioral Measurements*. Wiley, New York

Hambleton R K, Swaminathan H 1984 *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff, Boston, MA

Hambleton R K, Swaminathan H, Rogers H J 1991 *Fundamentals of Item Response Theory*. Sage, Newbury Park, CA

Hambleton R K, Swaminathan H, Algina J, Coulson D B 1978 Criterion-referenced testing and measurement: a review of technical issues and developments. *Review of Educational Research* **48**: 1–47

Harris C W, Alkin M C, Popham W J (eds.) 1974 *Problems in Criterion-referenced Measurement*. Center for the Study of Evaluation, UCLA, Los Angeles, CA

Linn R L (ed.) 1989 *Educational Measurement*, 3rd edn. Macmillan, New York

Lord F M, Novick M R 1968 *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA

Martin P, Bateson P 1986 *Measuring Behaviour: An Introductory Guide*. Cambridge University Press, Cambridge, UK

McDonald R P 1999 *Test Theory: a Unified Treatment*. Lawrence Erlbaum, Mahwah, NJ

Popham W J 1978 *Criterion-referenced Measurement*. Prentice-Hall, Englewood Cliffs, NJ

Shavelson R J, Webb N M, Rowley G L 1989 Generalizability theory. *American Psychologist* **44**: 922–32

Suen H K, Ary D 1989 *Analyzing Quantitative Behavioral Observation Data*. Lawrence Erlbaum, Mahwah NJ

Webb E J, Campbell D T, Schwartz R D, Sechrest L 1966 *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Rand McNally, Chicago

D. Rindskopf

Religion and Economic Life

Investigation of the relation between religion and economic life has long been a central concern of sociological reflection, but the topic has been largely defined through debates on the relationship of religion and capitalism, the latter being the specific and dominant expression of economic life in modernity. The work of Karl Marx on the origins and world historical significance of capitalism, and of Max Weber on the Protestant ethic and the 'spirit of capitalism' and the economic ethics of the world religions, inevitably overshadow the discussion of both the history and the contemporary forms of the ambiguous and changing relation between religion and economic life. These classical discussions do not, however, exhaust the topic, which can be divided as follows. In Sect. 1, precapitalist configurations of religion and economic life are reviewed both as regards the sociological, or socioeconomic, study of Western history and, very briefly, the anthropological study of primarily non-Western cultures. In Sect. 2, the emergence of modern capitalism is considered in relation to political economy, Marx and Marxism, and Weber's account of the elective affinities between religion and economic life. Following this, in Sect. 3, some contrasting reflective religious responses to nineteenth and twentieth century capitalism are outlined. In Sect. 4, contemporary capitalism, now unfettered by ideological confrontation with socialism, is represented as the driving force informing globalization, commodification, and managerialism, and as possessing such comprehensive power as to provoke and sustain a wide range of contrasting religious responses. In the concluding section, possible futures and research options are considered in relation to the capacity of economic life to destroy and recreate 'nature,' to reforge the parameters of human self-understanding, and to virtualize the condition of humankind. Economic life would seem increasingly to have become 'life' itself.

1. Precapitalist Configurations of Religion and Economic Life

The relation of religion to economic life is a function of the totality of any given society and its cultures; besides this, the way in which the interfusion of intrinsic, use, and exchange values may be understood depends on the theoretical assumptions of the inquirer. The most relevant founding fathers of sociology, Marx, Weber, and Simmel, sought to develop grand theory. In general terms, sociologists of religion have devised and reinforced relatively homogenous accounts of the modernization process and attendant phenomena of secularization, and they have used particular economic theories in order to inform and justify their approaches. By contrast, under the influence of Marcel Mauss' account of the gift and using exchange theory (Gregory 1994), anthropologists have generally paid more attention than sociologists to specific human ecologies and reciprocal community economies as originating in 'stone age economics' (Sahlins 1974). Some anthropologists have corre-

spondingly defended distinctive and reciprocal human ecologies and cultures against the ravages of economic modernization in ways inadmissable to orthodox sociologists.

Archeological and textual records indicate that ancient Egypt, Assyria, Babylon, and Israel all had contextually relevant, practical financial institutions, but there is little or no trace of analytical effort. The history of Western economic thought is normally traced back to the primarily agrarian world of the fourth and fifth century BC which knew the small trading firm and had a well-defined merchant class. The representative unit of the economy was the household and commodities were exchanged in free markets. In the Graeco-Roman world early 'economic' reflection was based on the '*nomos*' or law and rule of the household '*oikos*,' and the practical wisdom of household management (*oeconomicos*) was the source from which early economic conceptions were derived. In the setting of the city-state (*polis*), the universal existence of slavery ensured that labor was relatively fixed and immobile. Religion and economic life as part of the cultural superstructure were conditioned by the dependence of ancient Greek and Roman life upon a slave economy; correspondingly, manual labor was regarded as menial. Later, the Classical tradition provided some of the core values of Western culture for two millennia.

Central to Greek thought were the good life and reflection on the nature of virtue: seen thus, economic activity amounted to a mere external factor. Plato and Aristotle were, however, aware of such concepts as the distinction between use value and exchange value, the role of money, and the division of labor. In *The Republic*, the latter corresponded with the human types of ruler, soldier, and worker, and each class pursued its proper, specialized goals. The ideal rulers, the philosopher kings, would ostensibly be free from vulgarity of wealth. Aristotle, as did Plato, assumed the existence of slavery, but he envisaged a wider participation of citizens and reflected upon the distinction between use and exchange value, favoring value based upon use. The arts of commerce and the practice of usury were condemned, above all the latter, which makes gain out of money itself and not from the natural object. Money was to be understood as a medium of exchange, a unit of account, and a store of value; but in itself money was barren, and usury the oppressive exploitation of misfortune. The classical Christian doctrine of the Trinity, couched in terms of the 'economy' of God's self-revelation of his 'essential' being, employed a metaphor with an enduring importance. Later, the evolving theology of the atonement repeatedly exploited certain economic metaphors, not least in the conceptuality of payment of debt and redemption from sin.

Little new was added to Greek economics by Roman thought. The denigration of manual labor was continued by such writers as Cato and Cicero. Pragmatic management of ever larger estates (*latifundia*) and slaves was required, but the system based upon slavery was first questioned in principle by early Christianity. As with other aspects of life in the Roman Empire (for example, military service), the early Christian Church introduced subversive ideas which underwent rapid accommodation. While debates took place in the early church concerned (e.g.) alms-giving, according to Schumpeter there was little substantial interest in what was to become the dismal science of economics, for 'the Christian Church did not aim at social reform in any sense other than that of moral reform of individual behavior,' nor, moreover, 'did the Church attempt a frontal attack on the existing social system or on any of its more important institutions.' Indeed, 'It never promised economic paradise or, for that matter, any paradise this side of the grave' (Schumpeter 1954, p. 72). This is a view endorsed by Michael Mann, who argues that 'Christianity was a form of ideological power … . Only by becoming a Christian could one live a truly meaningful life' and that its 'power resides originally in the fit between the Christian message and the motivations and needs of the converted' (Mann 1986, p. 302). The question as to the 'fit' or affinity, elective or otherwise, between the impact of individual Christian belief and spirituality upon individual motivation and economic life has remained a matter of central importance above all in the thought of Max Weber. In a more uncompromising way, Marxist analysis implied that the religious ideology that emerged in the theological debates of the patristic theologians of the first four centuries in the West was but a reflex of the material base; this, however, does less than adequate justice to the complexity of the factors involved.

After the fading of the eschatological expectation of the earliest Christian communities and abandonment of the primitive communism, monasticism did, initially and in principle, preserve the radical exemplar of the economic ethic of the Synoptic Gospels (Troeltsch 1931). The religious orders that emerged from the practice of the Desert Fathers provided both the basis for the continuity of a religious culture and an economic model compatible in key respects with the warrior culture of feudalism. Monastic communities and the religious orders came to be centers of rational economic organization in ways that prefigure the business corporation. Monastic organization created wealth which, in turn, led to inner contradictions tackled by attempts at institutional reform—and which later became the target of acquisitive envy during the Protestant Reformation.

At the high point of scholastic theology and in the context of feudalism, a crucial debate took place which connects the early economic ideas of Aristotle (and of his wide philosophical concerns) with the early modern loosening of the constraints upon economic activity that facilitated the growth of international trade, and the justification of such expansion in

mercantilism. In the Middle Ages, Europe was characterized by a relatively stable agrarian, largely subsistence economy with slow urban growth. Central to urban productive activity was the guild, a form of association which ensured corporate monopolies, the protection of members, standards of quality, and price stability. The presence of the Church in relation to all social institutions ensured an unparalleled degree of cultural interfusion between religion and economic life. Social status and social function implied definite religious duties in a divinely ordered governance that extended throughout the body politic. Nineteenth-century guild socialism attempted to recover something of this integration of meaning into 'menial' work.

St Thomas Aquinas, the leading theologian of scholasticism, defended the just price and the prohibition of usury, the core principles of the attempted control by the Church of economic life. Estimating the just price based on moderate gain was a difficult task: the merchant was justified in trading at a price which ensured the life of his household in its accustomed style. Following Aristotle's account of money as barren, St Thomas condemned usury, the lending of money at interest, arguing that given the true role of money solely as a medium of exchange, it did not exist in the way that a rented house continued to exist when occupied by a tenant. St Thomas' counter-realist account of money applied more to the ethics of the personal loan rather than to large-scale financial transactions. During the course of the fourteenth and fifteenth centuries the increasing complexity of major financial transactions made it increasingly difficult to fit new practices within the constraints of Church law and this was reflected in modifications of St Thomas' rigorist views.

The prohibition of usury was circumvented through the creation of an international system identified with the Jews that was capable of supporting the accumulation of the major loans required for effective warfare. A similar role was also undertaken for a limited period by the Templars. The pariah status of Jews and their indispensable role in facilitating financial transactions and more widely in commerce came to reinforce negative attitudes which provided a long-enduring basis for anti-Semitism (Stern 1977, Chap. 18). The role of such Jewish thinkers as Marx and Sombart in the theorization of capitalism in the nineteenth century and beyond reinforced the impression that Jews and modern capitalism were intimately and fatefully interrelated (Paul Morris in Roberts 1995, Chap. 5). Freudian psychoanalytic accounts of the power of money as supreme 'reality principle' have continued the decoding of economic life. Present-day Islam continues the formal prohibition of the practice of usury, but Islamic banking and investment systems have nonetheless devised ways and means of adapting this restraint to the global economic system (Choudhury and Malik 1992).

In the early modern period, the growth of international trade, the breakdown of comprehensive ideological unity within Christendom following the Reformation, and the destruction of stable communities unified by shared religious values were factors associated with the rise of mercantilism. During the course of the sixteenth, seventeenth, and eighteenth centuries, economic analysis gradually gained increasing autonomy, throwing off the remaining constraints of scholasticism. Mercantilism and then physiocrat economic theory developed as distinct schools, which reflected the conditions of their countries of origin, England and France, where trading and rent of land represented prime sources of value, respectively.

2. *The Emergence of Modern Capitalism and its Critique*

The advent of early industrialization and the development of the factory system brought about unprecedented social disruption, breaking many ancestral ties: this process required both explanation and legitimation. In the *locus classicus* of political economy, *The Wealth of Nations* (1776), Adam Smith articulated a world view informed by the image of the pin factory in which virtue is equated with profit; his analysis of the rational division of labor and its necessary place in the making of profit grounded in the accumulation of surplus value of labor provided the justification for the removal of remaining constraints. The quasi-providential 'Invisible Hand' purportedly provided a spontaneous connection between individual wealth creation and the common good of society. Political economy attained a high status among an ascendent entrepreneurial class as it emancipated economic practice and provided the means for the elimination of embedded social, ethical, and cultural inhibitions, including those traditionally endorsed by religion in pre-industrial agrarian societies.

Marx regarded the intellectual destruction of religion by the eighteenth century Enlightenment as the axiom underpinning the true emancipation of humankind through subsequent social, political, and economic critique: there was, therefore, in principle no basis for any mutual reflective encounter between religion and economic life. Seen in full historical perspective, the theory of alienation adumbrated in the posthumously published *Paris Manuscripts* of 1844, the theory of society as composed of material basis and ideological superstructure in the *German Ideology*, the quasi-Hegelian ontology of capitalism expounded in the *Grundrisse* of 1857–8, the 'fetishism of commodities' in volume I of *Das Kapital*, and the later explorations in *Theories of Value* were all components in a theoretical scheme to which an adequate theological response has never been forthcoming. The representation of the proletariat in the

Communist Manifesto as the class excluded from all classes crystallized the fears of churches already confronted with their loss of influence over a growing urban working class and its anti-clericalism.

The uprooting of agrarian workers from an ancestral status sanctioned by divine order and their regimentation into a disciplined, wage-earning working class differentiated by a new division of labor created rather a dilemma for the churches. On the one hand, the Anglican clergyman Thomas Malthus endorsed population regulation in accordance with economic needs. On the other, the living conditions of the working class were an affront to those sensitive to the historic Christian teaching on wealth and poverty. Such figures as the Scot, Thomas Chalmers, English Christian socialists, and continental religious socialists struggled to find ways to redefine the role of religion under industrialized conditions. An unchurched, brutalized proletariat also, however, represented an inner threat to the role of Christian elites, clerical and intellectual across Europe. In the English religious establishment, serious critical engagement with 'scientific socialism,' or, moreover, with political economy was minimal. Methodism was exceptional in that it successfully harnessed religion and industrial work in ways which, according to the Halèvy thesis, spared England a revolution.

Max Weber was fully aware of the threat which Marx's thought and Marxist socialism created for Liberal Protestant European bourgeois culture and society. According to Weber, the *mentalité* associated with the capitalist mode of production had historical connections and elective affinities with 'this-worldly' Protestant asceticism. It is therefore customary (but not undisputed) to link the nascent 'spirit of capitalism' with the much disputed 'Protestant ethic,' above all within the Reformed Christian tradition. Weber explored the affinities between the single-minded pursuit of God and an equally focused desire for profit, but he also applied his analysis to the major world religions, classically, for example, to the religion of India (D. Gellner in Roberts 1995, Chap. 1). Marx and Weber agreed on the centrality of acquisitiveness to the bourgeois mind: political economy both explained and justified the alienative emancipation of single-minded acquisitiveness from contextually embedded human needs. Weber used the Latin scholastic ascription of God as supreme good when he analyses this distinctive mind-set and hinted at its wider implications:

> In fact the *summum bonum* of this ethic, the earning of more and more money, combined with strict avoidance of all spontaneous enjoyment of life, is above all completely devoid of any eudaemonistic, not to say hedonistic, admixture. It is thought of so purely as an end in itself, that from the point of view of the happiness of, or utility to, the single individual, it appears entirely transcendental and absolutely irrational. Man is dominated by the making of money, by acquisition as the ultimate purpose of his life. Economic acquisition is no longer subordinated to man as the means for the satisfaction of his material needs (Weber 1930/1976, p. 53).

In *Religion and the Rise of Capitalism*, R. H. Tawney applied Weber's insights to seventeenth-century England, and provided an indirect critique of the *laissez-faire* liberal capitalism which underwent economic and social crisis during the period between the First and Second World Wars. Later, further British explorations of the processual metaphor of the 'rise' of capitalism were extended by V. A. Demant to its 'decline' and its 'persistence' by R. H. Preston. Christian thinkers such as the Anglican William Temple relied upon the emerging Keynesian standpoint in advocating 'middle axioms' as the basis for socio-ethical intervention. Following the Second World War, the Keynesian program of partly state-managed mixed economy (and declining) capitalism underpinned the ideology and implementation of the 'welfare state' for 30 years until the 'triumph of capitalism' associated with the Reagan–Thatcher era. The re-assertion of neo-classical economics (Friedman 1962) and the ideology of the New Right (Green 1987) stripped away both tacit belief in the 'hidden hand' and all ascetic hesitance with regard to hedonism (Roberts 1992). Mainline religions in Anglo-American Western countries were split between localized variants of the 'option for the poor' and outright endorsement of the 'enterprise culture.' Those who advocated the option for the poor were depicted as naive and ill-informed practitioners of the 'kindness that kills.' Conversely, New Right ideologues proclaimed the virtues of greed, the adequacy of the 'trickle down' theory of welfare, and the necessity of a docile, 'flexible' labor force willing to undergo repeated social deracination.

3. *Reflective Religious Critiques of the Relation of Religion and Economic Life*

The most significant mainline religious response to industrialization and to the growth of capitalism and Marxist socialism is to be found in Roman Catholic social teaching. After lengthy consultation provoked by the changing socioeconomic conditions of the nineteenth century, the papal encyclical *Rerum Novarum* was published in 1891. Aware of the fate of the Catholic Church in revolutionary France (and then later of the Orthodox Church in Soviet Russia), papal teaching represented the historic struggle between capital and labor as conflict between the competing rights and obligations of owners and of workers, in relation to which the Church was to exercise a moderating role. Both parties were to recognize the mutuality of their rights and obligations, and to relativize these through acknowledgment of the

Beatific Vision as the sole and legitimate goal of humankind, thus hopefully resisting the Weberian slippage (under conditions of secularization) from the spiritual worship of God to the practical service of Mammon. For a century the Church developed this doctrine through a series of encyclicals culminating in Pope John-Paul II's *Laborem exercens* (*On Human Work*) (1981). Following the collapse of Marxist socialist states in 1989/90 the changing nature of capitalism was recognized in the encyclical *Centesimus Annus* (1991). The latter contains a cultural rather than a politico-legal analysis and critique of both socialism and capitalism in which the integrity of the human agent is asserted against the contradictions of a destructive, secularizing modernity in need of comprehensive re-evangelization.

Direct intellectual contact between Christians and Marxist socialists did not take place on a significant scale until after the Second Vatican Council (1962–5) and took the form of Christian–Marxist dialogue during the following decade, but even this discussion of the nature of 'true humanism' would seem to have had minimal long-term influence. At the same time 'political theology' embodying a synthesis between Protestant theology and various strands of revisionist Marxism also developed in Germany. Insights from political theology were transplanted to Central and Latin America, where, after the Second Vatican Council and the Medéllin Bishops' Conference of 1968, 'liberation theology' emerged in the context of rapid industrialization and urbanization, a situation similar in important respects to the conditions the young Marx and Engels had observed in the Rhineland in the 1830s. This theology drew critical perspectives from Marxist analysis of society but side-stepped the atheistic aspects of Marxism, as was to be emphasized in the critical response of the Sacred Congregation in Rome. In Central and Latin America liberation theology's stress upon the 'option for the poor' has tended to exacerbate a divergence between Roman Catholicism and Pentecostal Protestantism, where the latter has performed a role analogous to that of Methodism during the European industrial revolution (Martin 1990, pp. 27ff.). The most radical theological critique of capitalism through the representation of the imbalanced distribution of wealth in the global economy was a confessional issue (Duchrow 1987).

The Roman Catholic apologist and critic of liberation theology Michael Novak has written of a 'magic capitalism.' Humankind has to abandon the naive and infantile 'dream' (the 'false consciousness') of socialism, and to recognize the universal necessity of alienation at the heart of a vital economic reality. Every individual must pass through a rite of passage in which each encounters the 'empty shrine' of 'democratic capitalism': this is the contemporary 'dark night of the soul' (Novak 1982/1991, p. 53). Humankind may then draw upon spirituality, theology, and religious values (above all, those taken from Christianity and Judaism) for the individual strength to survive and compete. Others, like Peter Berger, deny that capitalism has such mythicizing power (Berger 1987). Other accounts of religion and economic life are to be found in Protestant Christian theologies of 'God' the Economist. (Meeks 1989).

Recently, the globally influential 'Third Way' politics of the center left advocated by such prominent figures as President William Clinton, the British 'New Labour' Prime Minister Tony Blair, and the British sociologist Anthony Giddens has been accompanied by a revival in Christian socialism and continuing campaigns to remoralize the individual. This post-Keynesian and postcommunist active capitalism seeks to combine the ideology of entrepreneurial zeal with managerialism and relentless sociocultural 'modernization.' The 'Third Way' contributes to the increasing salience of the religious factor in economic life, albeit with paradoxical consequences.

4. *Contemporary Capitalism: Globalization, Commodification and Managerialism*

Contemporary capitalism has both diversified into a plurality of capitalisms, each with its contextually relevant elective cultural affinities between religion and economic life, and has consolidated itself through globalization into an infinitely complex singularity. The representation of globalization as a new paradigm for the sociological interpretation of religion (Robertson 1992) has encouraged investigation of the struggle both between and within religions for power and influence in the context of the asymmetrically distributed economic opportunities and outcomes in an ever more competitive core–periphery world system (Beyer 1995). Moreover, the problematic interface between the global and the local has become the focal point of social identity formation (Castells 1996, Vol. II, Chap. 1). Commodification, the ascription of exchange and monetary value to all aspects of the world and human life, has advanced inexorably with the result that the economic function is central, and natural or intrinsic values become ever more marginal and are reduced to mere market values. Global managerialism (Entemann 1994), mediatized marketing techniques, and the application of 'neo-Fordist' organization to mind-work (Drucker 1993) have accelerated the uniformitarian 'production' of human beings through rational 'quality' education and training. These are now processes replicated within some influential strands of both mainline religion and alternative spiritualities.

Information technology has increased the velocity of circulation of all forms of 'capital' to the point that space–time compression is having pervasive cultural consequences in new 'economies of signs and space' (Lash and Urry 1994). For those at the heart of a

core–periphery economic system, capitalism has become 'soft' and infused with transformative cultural power (Thrift 1997). Under these conditions, religious language and theological vocabulary have become resources for the expression of economico-cultural power. Thus, terms like 'omnipresence,' 'omnipotence,' 'omniscience,' and 'instantaneity' formerly attributed to the 'God' who subsisted at the core of Western consciousness are perhaps the only adequate conceptual categories for such transmutations. Spiritual practices are now recruited into management training in order to assist through the provision of peak experiences the strongest possible integration of individual motivation with corporate goals (Roberts in Roberts 1995, Chap. 9). The religious field may be remapped in relation to these ongoing transformations of economic life. The application of rational-choice theory to religious behavior, a syncretic global marketplace of New Age 'self-religions' and spiritual body practices, the managerialization and marketization of mainline Christianity, Christian fundamentalist 'prosperity theology,' neo-pagan 'nature religion,' and ecological piety are all indications that religion and economic life are in active and sometimes mutually contradictory synergies. Consequently, there is no one 'spirit of capitalism' in advanced modernity—but a multiplicity of affinities and collusions in context marked by growing differentiation between 'religion' and 'spirituality.'

5. *Conclusion: Future Prospect—Research Options*

In the modern world, the last *residua* of hunter–gatherer societies and subsistence agriculture struggle with globalized economic power: the modern and the postmodern conditions sometimes co-exist in close, yet problematic contiguity (Harvey 1989). From being marginal to the order of things, conscious economic rationality is now for many 'life' itself. 'Economic life' tends to absorb, transmute, and meld 'reality' into seamless economies of virtualized simulacra, a globalized *maya*, outside of which there would appear to be no fulcrum for a critical ideology or politics of resistance (Baudrillard 1981). In a world in which the virtual exerts hegemony over the residual real the cosmos is silent; yet ambiguous opportunities abound, not least for women seeking freedom from both biological and social construction, as cyborg, prosthetic, and virtual cultures are created and colonized. After the so-called 'End of History,' leading sociologists of religion such as Steven Bruce and Bryan Wilson continue to represent religion as subject to terminal attrition by an inevitable process of secularization. Regarded thus, religion can no longer be a resource for the culturally significant critique of total human submission to the dictates of 'economic life.' Little would seem left for the researcher into the relation of religion and economic life other than to report on the final stages of religious decline.

Given the comprehensive economization of human life and the deteriorating global condition, the putatively normative pursuit of socio-scientific research on the assumption that it is value free and objectively disengaged is ethically untenable. Moreover, the comprehensive absorption by 'economic life' of the physical and biological basis of culture and society—and even of 'nature' itself—is ecologically unsustainable. Under conditions of totalizing inclusion in which the integrity of the ecosystem is reduced to a sociological and economic construction, deviant, marginal, and engaged research is essential. It is perhaps through an interdisciplinary and ecologically informed anthropology and in 'new economics' that the tools will be found with which to develop research into religion(s) and spiritual practices that may provide a vital but neglected resource for a sustainable human ecology and world survival.

See also: Catholicism; Economic Anthropology; Economic Sociology; Marx, Karl (1818–89); Religion and Science; Weber, Max (1864–1920)

Bibliography

Baudrillard J 1981 *For a Critique of the Political Economy of the Sign*. Telos Press, St. Louis, MO

Berger P L 1987 *The Capitalist Revolution. Fifty Propositions about Prosperity, Equality, and Liberty*. Wildwood House, London

Beyer P 1995 *Religion and Globalisation*. Sage, London

Castells M 1996 *The Power of Identity, The Information Age*. Blackwell, Oxford, UK, Vol. II

Choudhury M A, Malik U A 1992 *The Foundations of Islamic Political Economy*. Macmillan, London

Drucker P F 1993 *Post-capitalist Society*. Butterworth-Heinemann, London

Duchrow U 1987 *Global Economy: A Confessional Issue for the Churches?* WCC, Geneva

Entemann W F 1994 *Managerialism: The Emergence of a New Ideology*. University of Wisconsin Press, Madison, WI

Friedman M 1962 *Capitalism and Freedom*. University of Chicago Press, Chicago

Green D G 1987 *The New Right: The Counterrevolution in Political, Economic and Social Thought*. Harvester Wheatsheaf, London

Gregory C A 1994 Exchange and reciprocity. In: Ingold T (ed.) *Companion Encyclopaedia of Anthropology*. Routledge, London, pp. 911–37

Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK

John Paul II Pope 1991 *Centesimus Annus*. Origins, Washington, DC

Lash S, Urry J 1994 *Economies of Signs and Space*. Sage, London

Mann M 1986 *The Sources of Social Power History of Power from the Beginning to AD 1760*. Cambridge University Press, Cambridge, UK, Vol. I

Martin D 1990 *Tongues of Fire: The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK
Meeks M D 1989 *God the Economist: The Doctrine of God and Political Economy*. Fortress Press, Philadelphia, PA
Novak M 1982/1991 *The Spirit of Democratic Capitalism*. Institute for Economic Affairs, London
Roberts R H 1992 Religion and the 'Enterprise Culture': the British experience in the Thatcher era (1979–1990). *Social Compass* **39**(1): 5–33
Roberts R H (ed.) 1995 *Religion and the Transformations of Capitalism: Comparative Perspectives*. Routledge, London
Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London
Sahlins M 1974 *Stone Age Economics*. Tavistock, London
Schumpeter J A 1954 *History of Economic Analysis*. Oxford University Press, New York; Allen & Unwin, London
Simmel G 1978 *The Philosophy of Money*. Routledge and Kegan Paul, London
Stern F 1977 *Gold and Iron: Bismarck, Bleichröder, and the Building of the German Empire*. Penguin, Harmondsworth, UK
Thrift N 1997 *Soft Capitalism*. Sage, London
Troeltsch E 1931 *The Social Teachings of the Christian Churches*. Allen and Unwin, London
Weber M 1930/1976 *The Protestant Ethic and the Spirit of Capitalism*. Allen and Unwin, London

R. H. Roberts

Religion and Gender

According to Social Sciences religions are social and cultural features, resulting of specific historical constructions. Religious beliefs and symbols as well as church organizations are the result of social processes in which social agents are inscribed. Every religious action is limited and directed by the social situation of religious agents, official leaders, or simple believers. Practices, representations, symbols, and discourses of religious field are accomplished by socially placed believers. Even when they express their faith supposed to be absolutely private, social relations take place. Classical approaches of religions deal with the question of how social class influence religious believers and behaviors. Nevertheless, the ways gender and race influence beliefs and religious institutions have rarely been taken in account.

In the context of feminist critique to the female exclusion of the possibilities of formulating knowledge, the efforts to alter the ways of thinking are an integral part of the historical struggle of women. In the seventeenth century, Sor Juana Inés de la Cruz, a Mexican Catholic nun, invoked the right of women to participate in knowledge in the same manner as men. The development of feminist studies, especially from the 1970s on, led to the formulation of sophisticated feminist criticism of the traditional and modern forms of Western thinking. Thus, practically all areas of knowledge have been questioned by feminist theoreticians who developed alternative proposals to hegemonic forms of knowledge.

Women have also brought the field of religious studies into question. Feminist theories and methodologies have been applied to the field of religion especially by using the category of gender or social relations with respect to sex, showing how a significant part of traditional studies in this area treat religious agents without distinction of gender and race. This generalization can be found in the use of homogenizing categories of a complex reality. Observe the categories entitled 'clergy,' 'laity,' 'hierarchy,' etc.; they are employed without any reference to the sex or race of their components. The consequences for analysis have been shown by empirical research attesting to the manner in which the field of religion is molded not only by class, but also by gender and race. The approach in terms of gender, therefore, questions the pretentious neutrality of the social sciences in their aim to examine religions.

Among the main critical concerns of these issues is the uncertainty that the sexism inherent in historical religions can be overcome. In the opinion of a few of those who study the issue, these religious are inevitably androcentric. These scholars propose the creation of alternative religious spaces where women can produce new forms of relating with the sacred, which is also thought of in an innovative way. Other scholars believe that the problem of historical religions is that men appropriated them. It is necessary to recuperate the original traditions and foundations for women to find their place in these religions.

Another area of criticism is religious organizations, churches, and others in which women as a whole are excluded. From Judaism to Taoism (Sharma 1987, Gross 1996), religious organizational forms are analyzed as social institutions that guide and control women's lives and bodies. Special reference is made to three aspects that exclude the female population in these institutions: the sphere of decision-making and power, symbology and the predominance of a monotheism in which the divinity is male, and the formulation of theological and moral discourse.

1. Feminist Criticism of the Classics of the Sociology of Religion

The specific field of sociology of religion has also been submitted to the scrutiny of feminist critique. Scholars point to sexism in the analysis of religion found in the work of classical authors in this area, such as Durkheim and Weber (Erickson 1995, Bologh 1990). The first author places a rigid division between the sacred and the profane to explain social life and religion. The power generated by rites and beliefs establishes this separation. Men, holy beings *par*

excellence and having been gifted with a soul, are the protagonists of beliefs and rites whereby new relationships and society itself are created. Women, lacking a soul, are relegated to the sphere of the profane, repetitive day-to-day life and the absence of creative capacity. Only power and violence are able to keep these two spheres separate. In modern, secularized societies, rites and beliefs are meant to help keep the ideal world of men separate from the world of women by force.

Weber's sociology of religion also bases itself on dichotomy: religions founded on both rational and ascetic ethics combat magical and ecstatic religions that incorporate eroticism. The first religions are aimed at action in the world and allow one to attain the greatness of a hero, a charismatic leader or a prophet. The second religions lead to love and only serve to 'comfort the warrior,' and are negative in relation to action in the world. Men are associated with the religions of active asceticism and rationality and are responsible for progress since they have been equipped with heroism. Social order is founded on the empire of reason and greatness, at the cost of the repression of the sentiment, eroticism and magic attributed to women.

The two classical authors defend the dominant social order in relation to the subordinated positioning of women, thus legitimizing the dominance of the male elite both in society and in religion sustained by the female population and the masses. After them, other sociologists addressing the religions institutionalized in churches rarely realize that the nature and functions of religious power are based on the sexual division of the work. The separation of the cleric from the layperson is, at once, a separation of men and women in which the power is attributed exclusively to men. One of the problems with the sociology of social and religious organizations is, as feminist sociologists demonstrate, the abstract treatment of the categories used as supposedly neuter in terms of gender, which prevents the perception of fundamental elements of reality. Even if certain studies include considerations on women, they are generally descriptive and exclude any analysis or explanation of male dominance.

2. *Feminist Contributions to an Understanding of Religion*

Feminist critics envision creative, new directions for classical and current studies of the social studies of religion, challenging their pretentious neutrality and objectivity. For example, Gross (1996) proposes the necessity of academic objectivity and, at the same time, empathy with respect to the subject matter. Empathy would be the most appropriate methodology to deal with polemic themes arousing intense emotions. It allows one to understand the believer's point of view and creates an important point of contact between academic religious studies and religious practices from a perspective of personal faith. The effect of religious studies therefore is that which radically refutes absolutism.

In the field of theology, scholars bring traditional representations of divinities into question, in an attempt to overthrow the monotheistic, male conception of the Christian God, for example. According to a significant group of feminists, Christian women theologians reject the traditional understanding of 'human nature,' which quarters women into the specific area of maternity by destiny and not by choice. This rejection of the biological resource to explain social and religious order of the sexes leads women to radically criticize the current hierarchical organization of ecclesial institutions, especially the Catholic Church. An affirmation of an individual autonomy that is not opposed to the construction of collective ideas but rather supports them further leads women to claim recognition as moral agents capable of making ethical choices in all fields of life, including those related to the control of sexuality and the capacity to conceive new human beings.

An analysis of the consequences of exclusions that fall on women in the area of religion shows that they are not limited to the fact that women are prevented from being included in the decision-making process in churches and the formulation of authorized knowledge. These consequences also extend to the symbolic field, since they keep women within a closed circle of impurity. The impossibility of presiding over rituals considered fundamental to the development of faith and the delimitation of distinct physical spaces according to sex, with secondary importance being attributed to women, among other examples, add to the religious image of the female population as sinners and polluters of the sacred. In the case of Catholicism, this happens in a radical way. In order for a man to reestablish the relationship of a community of believers with its God, it is not only necessary that he be a man, but also that he be celibate, i.e., that he not touch any women.

Another area in which there are significant feminist contributions is the history of religion. By interrogating existing historiography and the silence with regard to female proaction, and by confronting the difficulty of sources, women scholars have been seeking to write a different history in which facts, figures, and processes are shaped by the social relations established between the sexes. Catholic religious nuns rigorously trained not to have a 'personal or collective history' have been the subject matter of historical and sociological study.

In all fields of religious study, feminist scholars affirm that the fundamental issue is to understand the way in which symbolic activities—beliefs, rites, and religious discourse—that seem to escape sexual differentiation, similar to the ordering of the internal

power structure of religious institutions, truly conform to such differentiation. Feminists are therefore concerned with showing how religions are presented in social and historical reality to be molded and interwoven with gender relations.

3. Womens' Investment in Religion

Feminists commonly affirm that religions historically have been spaces for guidance and control of the behavior, thoughts, practices, and representations of women. Nevertheless, many scholars believe that religious space is contradictory and ambiguous for allowing women to have certain gains in concrete historical situations. For this reason, women invest in religion to find possibilities for development and autonomy to counteract the institutional investment in control. Religions can function as mobilizing forces, allowing women to resist disciplining religious and social power.

Feminist scholars of religious history asked historiographical sources demonstrating religious women's protagonism in different religions and diverse societies and historical periods. For instance, some Buddhist traditions became attractive for women because of the large number of famous female practitioners and teachers found in their ranks. In Taoism we can find important women alchemists, monks with their own monasteries and accomplishing high roles as shamanistic mediums. In Christianity, women monks were the chiefs of mixed monasteries, with male and female monks, in medieval Europe. They managed high levels of ecclesiastical power.

Research on nineteenth century Europe and Latin America shows that under certain circumstances religions historically have become unique social spaces for the female population. These spaces further act as a refuge and create conditions for women to escape male control, including by opening up the possibility of not marrying in order to dedicate themselves to spiritual life in convents and similar places. Recent studies have reported alterations in the religious field that are partially favorable to women. This is the case of Christian Ecclesiastical Base Communities, which offered many women from poorer communities the possibility of removing themselves from the domestic environment and unbalancing sex relations as well as an important learning in exercizing leadership (Drogus 1997) These communitarian spaces enabled the assimilation of feminist concepts allowing women to criticize the Catholic ecclesiastical institution as male-dominated and patriarchal. It is interesting that apparently opposite fields of discourse—feminism and Christianity—were blended to empower women. New religions were also dealt with in this manner, with research showing their ambiguities in relation to the gains they offered the female population.

Another strategy women use in their favor is the reinterpretation of sacred texts. Like many Christian theologians who developed new readings of the Bible, Islamic female scholars reinterpreted the Koran, showing that male domination of the female population, as well as the idea that women were created for men have no basis on the sacred books. Equally they deny women of being guilty, based on the proposition of the guilty figure of Eve. Buddhist scholars affirm the equalitarian original character of Buddhism, They argue that Buddha taught women as well as men and had both as his followers.

Hence, womens' investment in religion can be explained not only by the social control exerted over them through religion, but also by their taking advantage of their religious participation.

4. Future Directions of Research

The affirmation of religious experience as a gendered experience has led specialists in religious studies and theologians of various religions to reflect with ever-increasing depth on the different ways in which men and women deal with their religious beliefs and practices. In this vein, many fields have lent themselves to feminist research in order to understand and explain womens' relationship with religion and their place in historical religions. One of the most useful fields seems to be that which analyzes the effects of the contemporary re-arrangements occurring in religion both in the West and the East on the lives of women and, as a consequence, men. Another area of research, in terms of claiming sexual and reproductive rights, is that which aims to discover the ways in which different religions appropriate the female body and seek to control its sexuality and reproductive capacity. One area open to empirical research still remains: the conditions for women to produce new religious practices, discourse, and symbols and the effects on religions and the entirety of believers, whether men or women.

In sum, the growth of empirical data prepared taking gender as theoretical tool should offer new clues for understanding and explaining the reality of religions in societies of the third millennium.

See also: Islam and Gender

Bibliography

Bologh R W 1990 *Love or Greatness. Max Weber and Masculine Thinking. A Feminist Inquiry*. Unwin Hyman, London

Drogus C A 1997 *Women, Religion, and Social Change in Brazil's Popular Church*. University of Notre Dame Press, Notre Dame, IN

Erickson V L 1995 *Feminism, Social Theory, and Religion*. Fortress Press, Minneapolis, MN

Gross R M 1996 *Feminism and Religion. An Introduction.* Beacon Press, Boston
Sharma A (ed.) 1987 *Women in World Religions.* State University of New York Press, New York

M. J. F. Rosado Nunes

Religion and Health

There are two distinct ways in which religion may be related to health. Religions—complicated systems of symbols, beliefs, practices, and the social institutions associated with them—may be classed with the psychosocial factors with apparent influence on health status. In this view, religiousness may be a protective or preventive factor in reducing mortality or morbidity. A second view reverses this causal order, in that religious beliefs and practices are often turned to by individuals and families faced with serious illness, or at the end of life, and religiousness can be affected by changes in aging or health status.

1. *Religion as a Factor Affecting Health*

> The earliest examples of practices designed to promote the public health are to be found, among primitive peoples, inextricably mingled with the ritual of religion.

These are the opening words of C-E. A. Winslow in the entry for 'Public Health' in the original 1935 *Encyclopedia of the Social Sciences*. Winslow, professor of public health at the Yale School of Medicine, known for his studies of the decline in infectious disease mortality, draws attention to early quarantine efforts in ancient Persia and among the Biblical Hebrews as effective public health practices, motivated and given form by religious beliefs. The burial of the dead, the separation of lepers from the community, or the ritualistic preparation of meat for consumption are but a few of the early examples of practices carried out for religious reasons that had unintended public health consequences. In general, this characteristic of being unintended is typical of the work discussed under this heading—religious practices or beliefs seen to have associations with health are not undertaken for the purpose of improving individual or community health, but rather, are ends in themselves.

The social scientific origins of both theoretical and empirical work on religion and health can be found in Emile Durkheim's (1951) *Suicide*. Durkheim demonstrated, with the use of European national and regional statistical records, that suicide rates varied dramatically and predictably with certain social characteristics, among them the rates of Protestants, Catholics, and Jews in a population. His demonstration that suicide rates were lower among regions with larger Catholic and Jewish populations led him to denote types of suicides (*anomic* and *egoistic*) in modern societies. Suicides were prevented in religious and other social groups that regulate or constrain the behavior of individual members, and also provide them with integrative and socially supportive relationships.

Despite this foundation in late nineteenth century classical sociological thought, research on religion was largely absent from social scientific work in health until the latter half of the twentieth century. When such work did appear, it was limited to studies of small, peripheral subpopulations, or to suicide as the only cause of death. Beginning in the 1960s, studies finding lower rates of mortality and morbidity (particularly hypertension) for clergy, members of religious orders, and certain sectarian groups including Seventh-Day Adventists, Mormons, and Amish, appeared in US medical and public health literature (Jarvis and Northcott 1987). Work on religiousness and suicide rates has also appeared steadily, with generally consistent support for Durkheim's original findings (Pescosolido and Georgianna 1989). More generalizable findings came in 1979, when the well-known Alameda County (California, USA) analysis of social networks and mortality appeared (Berkman and Syme 1979), quickly followed by similar findings from large population studies in Tecumseh County, Michigan (USA), and Evans County, Georgia (USA). In each of these longitudinal studies, social ties, including ties to religious congregations, were associated with lower mortality risks.

These studies were important for many reasons. Until this time, the social environment, long a member of the agent–host–environment triad in multifactorial models of disease causation, had primarily been thought of as having negative influence on health. Social environments were considered sources of stress; social conditions were frequently operationalized as poverty, urban crowding, time-pressured working conditions, or the stress of modern life. The social network studies opened for consideration the possibility that the social environment also provided positive or protective factors in health. The first social network studies were also important because they included some measure of affiliation with a religious group or attendance at religious services.

In the 1980s and 1990s, studies focusing specifically on religion as a mortality risk factor appeared, primarily from the US and Israel, most showing a moderately large and significant protective effect, even after adjustments for possible selection factors, confounders, and mediators (Kark et al. 1996). For example, Hummer et al. (1999) found that frequent attendance at religious services reduced mortality risk such that seven years was added to life expectancy at age 20. These findings were different from the earlier studies of homogeneous religious groups or clergy, in that they represented the general population, with its natural heterogeneity in religious behaviors and beliefs. Like the social network studies, they were

longitudinal, and thus were able to adjust for the health status of respondents at the start of the study. Data from prospective observational studies with well-designed health status assessment at baseline is critical, particularly when measurement of the religion variable is limited to the indicator of attendance at religious services. The potential selection effects of healthier members of the population attending religious services more frequently, and being at simultaneously lower risk of mortality, are obvious. Additional important selectivity variables included in these analyses include age, sex, socioeconomic status, education, ethnicity, and marital status, as religious participation is generally higher for women, married couples, blacks, older people, and those with higher education and income.

However, a chief criticism of the studies in this emerging field is that most contain only one or two indicators of religiousness, inadequate conceptualization and measurement for such a highly complex set of practices, beliefs, and institutions. Most frequently, attendance at religious services has been the only available measure. When some measure of private religious feelings or practices has been included and compared, attendance is usually the measure with the largest protective effects. Better measurement of the independent variable is clearly necessary before a larger understanding of explanatory mechanisms can be gained. At the present stage of research, however, several mechanisms through which religiousness might affect mortality have been described.

1.1 Reduction of Behavioral Risk Factors

The health economist Victor Fuchs observed in the early 1970s that religious teachings regulate and constrain human behaviors in numerous ways that have known health consequences (Fuchs 1974 pp. 52–5). He noted the enormous differences in infant, child, adolescent, and adult mortality rates in the two adjacent states of Utah and Nevada. Population density, climate, health care, income, and education levels in the two states were rather similar; any advantage went to Nevada. Yet death rates in Nevada for males aged 40–49 were 54 percent higher, and for females 69 percent higher than in Utah. When the causes of death were limited to cirrhosis of the liver and cancer of the respiratory system, Fuchs found that the corresponding figures were 111 percent and 296 percent. He attributed the dramatic differences to the 'different lifestyles' of the residents of the two states, Utah being primarily inhabited by Mormons who do not drink alcohol or smoke tobacco and 'in general lead stable, quiet lives' with high marriage and fertility rates and low divorce rates. Even within the Mormon church, levels of adherence to church doctrine have been shown to be associated with reductions in smoking and alcohol-related deaths (Gardner and Lyon 1982).

Most religious groups, however, do not have specific, doctrinal beliefs regarding these behavioral health risk factors, and yet rates for cigarette smoking (Koenig et al. 1998a) and substance use (Kendler et al. 1997) are significantly lower among those who attend religious services regularly, even in studies of religiously heterogeneous populations. A US national study of high school seniors found that the more frequent attenders, and those who thought religion was important in their lives, had lower rates of smoking, alcohol and marijuana use, carrying weapons, getting into fights, and driving while drinking. They also had higher rates of positive health practices such as seat belt use, and eating fruits, vegetables, and breakfast (Wallace and Forman 1998).

Religious involvement among adolescents that lowers many behavioral health risks is especially important because it represents a lowering of lifetime risks for behaviors such as smoking, which are most likely to begin in adolescence. Moreover, an analysis of data from the Alameda County study shows that religious attenders who do smoke are more likely than nonattenders to quit smoking during the course of a follow-up study, and hence further reduce their mortality risk (Strawbridge et al. 1997). Thus one of the mechanisms through which religion influences health consists of the constraints placed on behaviors that are sensation-seeking, self-stimulating, and increase risk to the self and others. The reduction in lifestyle risk factors results in lowered rates of lifestyle-related causes of death, and in at least one study a temporal pattern of religious practices preceding risk reduction behaviors has been established.

1.2 Enhancement of Social Support Networks

A second mechanism by which the effects of religious involvement may operate is through the expansion and maintenance of social support networks. The study of religion and health is in some ways derivative of the study of social networks and health, and indeed the latter studies provided the earliest modern findings, in Alameda County and elsewhere, as noted above. This mechanism represents the parallel strain in the Durkheimian theoretical tradition, that religious institutions provide both regulation for the behavior of members, as well as integration, or feelings of belonging in a social group. Religious congregants report larger social networks of both friends and relatives, and also more social support from them, compared with nonattenders (Ellison and George 1994). There are several reasons why religious congregations would serve as centers of effective social support for members of all ages, but particularly for the elderly. Congregations offer a mix of strong and weak ties, and can thus offer a broad range of emotional and instrumental supports.

Congregations of all major faiths share values and beliefs about the importance of helping others who are less fortunate and also provide organizational structure for mounting such efforts. Religious institutions provide both support and opportunities to support others, and reciprocal or bidirectional social support may be the most beneficial. Finally, as a corollary of this, members of religious groups who have served others in the past, and who have seen others receiving help, have reason to anticipate support in the future in the event of need.

Population studies of religiousness and health outcomes, particularly mortality, have generally included some assessment of these two explanatory mechanisms (e.g., Hummer et al. 1999, Strawbridge et al. 1997). Usual findings have been that: (a) religious involvement, particularly public participation in services, is associated both with better health practices and with higher levels of social support, and (b) these known behavioral and psychosocial risk factors explain some proportion (but not all) of the effect of religiousness on later health outcomes. In studies of healthy populations, then, religion (conventionally operationalized as attendance at religious services) appears to be related to a lower risk of mortality after adjustment for known confounders and selectivity variables, and the effect is partly explained by existing explanatory mechanisms and partly independent of them. This is just an emerging field, however; the number of extant studies is relatively few, the measurement of the religion variable is undeveloped, the research has been largely done in the US in regional samples with Judeo-Christian beliefs, and other explanatory mechanisms should be tested.

2. *Illness and the Search for Meaning*

The first part of this article has been concerned the unintended health consequences of religious practices, or measurable changes in health that result from religious practices undertaken for religious and not health-related purposes. Equally important for a social scientific study of religion and health, however, are religious beliefs and practices that are engaged in as a direct result of illness. Religious groups can offer both spiritual and tangible support to the sick, including prayers and comforting rituals, but also meals, visitation, childcare, and even financial help. A US survey of the use of unconventional medicine found that 25 percent of their national sample reported that they had used prayer as a treatment for illness in the past 12 months; only exercise, with 26 percent, was used more often (Eisenberg et al. 1993).

Research on the effects of religious coping (including but not limited to prayer) has frequently taken indicators of mental health and well-being as outcomes; studies of recent widowers, parents who had lost a child, women with hip fractures, or elderly men with multiple chronic illnesses have shown lower rates of depression among those with religious beliefs or participation compared with those who have none. For example, hospitalized patients diagnosed with major depressive illness were found to have a shorter time to recovery from depression if they felt themselves to be more intrinsically religious (Koenig et al. 1998b). In another study, heart patients who felt that they received strength and comfort from their religious beliefs had a lower risk of death during the six months following surgery than those who did not (Oxman et al. 1995). At the same time, not all forms or styles of religious coping may be benign, for example, when the cause of an illness is attributed to moral failure; research is just beginning on maladaptive religious coping (Pargament 1997).

Classical sociological theories of religion again anticipate these functions of religion in health. Weber's concept of 'theodicy' provides a historical model of the articulation of the problems of individuals with religious language. Religious individuals can draw on the experiences of fellow believers or the stories of their faith tradition to put their own lives into a larger context, to learn lessons from others who have faced similar circumstances, to reduce fear, or to gain hope for the future. Religious ritual practices for mourning or for healing can assist individuals in making transitions to new life stages in which adaptation to loss is required. Religious coping strategies appear prominently in the coping repertoires of aged individuals faced with their own serious illness, or the illness or death of a loved one, situations over which the individual has little direct control, and which may lack clear explanations or challenge beliefs that the world is just (Pargament 1997). In contrast with the population studies described above, in which attendance at worship services was most predictive of later health outcomes, studies of religious coping frequently find that subjective religious beliefs or nonorganizational forms of religious practice (e.g., private prayer) play an important role.

Religious beliefs and practices are especially prominent in the care for and concerns of individuals at the end of life. The modern hospice, begun in England in the 1960s, was modeled after medieval Christian institutions that cared for the dying, and the spiritual concerns of patients are still a central aspect of the new field of palliative medicine. There is also some evidence that the observance of religious holidays may alter the timing of death for elderly individuals; multiple studies in both Jewish and Christian populations have observed short delays before death that permit the celebration of a last sacred ritual (Idler and Kasl 1992).

Despite a background in the Durkheimian and Weberian traditions of the sociology of religion, the field of research in religion and health remained mostly unexplored through the century, but has begun grow-

ing quite rapidly in the 1990s, especially in the US, where levels of interest in religion are generally higher than in Europe. One may anticipate further growth in this field, as the measurement and conceptualization of the relevant domains of religiousness and spirituality improve, and as the population ages and the prevalence of chronic illness increases.

See also: Adolescent Health and Health Behaviors; Aging and Health in Old Age; Bereavement; Chronic Illness, Psychosocial Coping with; Depression, Hopelessness, Optimism, and Health; Emotions and Health; Epidemiology, Social; Health Behavior: Psychosocial Theories; Psychotherapy and Religion; Religion: Definition and Explanation; Religion, Phenomenology of; Religion: Psychiatric Aspects; Religion, Psychology of; Religion, Sociology of; Social Integration, Social Networks, and Health; Social Support and Health

Bibliography

Berkman L F, Syme S L 1979 Social networks, host resistance, and mortality. *American Journal of Epidemiology* **109**: 186–204

Durkheim E 1951 *Suicide*. Free Press, Glencoe, IL

Eisenberg D M, Kessler R C, Foster C, Norlock F E, Calkins D R, DelBanco T L 1993 Unconventional medicine in the United States: Prevalence, costs, and patterns of use. *New England Journal of Medicine* **328**: 246–52

Ellison C G, George L K 1994 Religious involvement, social ties, and social support in a southeastern community. *Journal of the Scientific Study of Religion* **33**: 46–61

Fuchs V R 1974 *Who Shall Live? Health, Economics, and Social Choice*. Basic Books, New York

Gardner J W, Lyon J L 1982 Cancer in Utah Mormon men by lay priesthood level. *American Journal of Epidemiology* **116**: 243–57

Hummer R A, Rogers R G, Nam C B, Ellison C G 1999 Religious involvement and US adult mortality. *Demography* **36**: 273–85

Idler E L, Kasl S V 1992 Religion, disability, depression, and the timing of death. *American Journal of Sociology* **97**: 1052–79

Jarvis G K, Northcott H C 1987 Religion and differences in morbidity and mortality. *Social Science & Medicine* **25**: 813–24

Kark J D, Shemi G, Friedlander Y, Martin O, Manor O, Blondheim S H 1996 Does religious observance promote health? Mortality in secular vs. religious kibbutzim in Israel. *American Journal of Public Health* **86**: 341–6

Kendler K S, Gardner C O, Prescott C A 1997 Religion, psychopathology, and substance use and abuse. *American Journal of Psychiatry* **154**: 322–9

Koenig H G, George L K, Cohen H J, Hays J C, Larson D B, Blazer D G 1998a The relationship between religious activities and cigarette smoking in older adults. *Journal of Gerontology Series A: Biological Sciences and Medical Sciences* **53A**: M426–34

Koenig H G, George L K, Peterson B L 1998b Religiosity and remission of depression in medically ill older patients. *American Journal of Psychiatry* **155**: 536–42

Oxman T E, Freeman D H, Manheimer E D 1995 Lack of social participation or religions strength and comfort as risk factors for death after cardiac surgery in the elderly. *Psychosomatic Medicine* **57**: 5–15

Pargament K I 1997 *The Psychology of Religion and Coping*. Guilford, New York

Pescosolido B A, Georgianna S 1989 Durkheim, suicide and religion: toward a network theory of suicide. *American Sociological Review* **54**: 33–48

Strawbridge W J, Cohen R D, Shema S J 1997 Frequent attendance at religious services and mortality over 28 years. *American Journal of Public Health* **87**: 957–61

Wallace J M, Forman T 1998 Religion's role in promoting health and reducing risk among American youth. *Health Education and Behavior* **25**: 721–41

E. L. Idler

Religion and Politics: United States

Contrasted with other developed Western societies, high levels of religious affiliation and practice persist in the United States. Although the USA has maintained a constitutional separation of church and state, its citizens' religious beliefs and practices have often had political consequences. This article reviews theoretical and empirical literature about the mechanisms that link religion to American politics.

1. Definitions of Central Concepts

According to Durkheim, 'religion' is a form of collective memory for a society that projects order, stability, and predictability in social interactions on to the Divine will. Leege argues that 'People (who) share a sacred community…develop the self-confidence to act in concert. Religion specifies what action to take, and religious beliefs (and religious institutions) create the obligation to act' (Leege and Kellstedt 1993, p. 10). For the adherent, the political consequences of religious beliefs are that if life is not genuine unless lived in a religious way, so politics is not just unless its policies measure up to God's law.

Culture and religion have a deep symbiotic relationship. 'Culture' is a template—in Geertz's term 'a set of control mechanisms' (Geertz 1973, p. 49) for social relationships. When undergirded by religious values, culture establishes a moral order (Wuthnow 1987). Culture performs three central functions for a society: (a) it offers identity, (b) it prescribes norms for behavior, and (c) it maintains boundaries on relationships (Wildavsky 1987).

'Politics' refers to collective action. Politics resolves conflicts over social goals and the means for attaining them through collectively binding decisions (Easton

1953, Lasswell 1958). 'Power,' the ability to get others to do one's bidding (Dahl 1956), is the central relationship in politics. Culture and religion are important to politics because they help to legitimate social hierarchies, set goals, and specify boundaries on political institutions.

In democratic societies, religious institutions are normally located in the civil society. 'Civil society' is that collection of cultural norms, social and economic institutions that compose a way of life but cannot legitimately employ the means of coercion. They range from freely chosen activities (voluntary associations, market transactions) to activities resulting from one's station in society (class obligations). Since 'church' is in civil society and the 'state' potentially possesses the legitimate monopoly over the means of coercion (Max Weber, in Gerth and Mills 1946), Americans often stress the need for constitutional separation. A theocracy merging the two would be dangerous, but a public square devoid of humane religious standards for justice would lack transcendent perspectives and meaning (de Tocqueville 1958).

2. *Role of Religion in American Social and Political Life*

Max Weber (in Gerth and Mills 1946) argued that to understand a nation's religious world views, one should ask what its people are saved from and saved for. The American social system and its mythology reinforced a deep sense of equality and enterprise; one could become whatever one wanted through individual initiative. But how could some legitimately rule or become wealthy in a nation without a natural aristocracy? After all, the USA was settled heavily by those who had run foul of the law, escaped from conscription, would not submit to ecclesiastical order, or wanted nothing of the poorhouse. Creating a nation of such people, Leege et al. (in press) argue, would be a recipe for anarchy. However, a founding myth joining the Pilgrim covenant of popular sovereignty under God to the righteous life of Puritans would provide a legitimating moral order. Out of the welter of humanity, one could know whom to trust in economic relations or whom to invest with political power by their public righteousness. Evidence of religious commitment afforded social standing. From Thomas Jefferson to William Jefferson Clinton the test of religious commitment has been a test of fitness to rule.

Beyond its cultural function, religion persisted for three reasons, and each gave it a political character. First, contrary to the lazy monopoly that most European churches exercised (see Finke and Stark (1992) for a discussion of this thesis), American churches had to compete for adherents. In addition to dispensing the sacraments, churches developed a wide range of ministries, programs, and social services to attract and hold the loyalties of their members. Political interests evolved from efforts to maintain and extend such programs. Second, as economic development created an ever more complex society, and individualism became the dominant social philosophy, people needed an institution to provide meaning in everyday life. The church became that solace and gathering place for 'people like us' (see Fowler 1989). Yet urbanization and mass communication thrust people with very different cultural and religious values into proximity with each other (see Engels, in Benson (1979, p. 205) for a realization of how this process undercut a Marxian interpretation of historical stages). Often people felt compelled to seek state action to protect or extend their parochial values (see Hertzke (1993) for two illustrations of cultural populism in recent American politics). Third, a transformationist impulse ran deeply through American piety. Calvinism, the dominant cultural form of American religion, had specified that righteousness would be evident in diligent work, the acquisition of wealth, and charity for the less fortunate (Weber 1946). Nevertheless, righteousness was difficult in an unjust society. Society needed to be transformed from time to time to be a fitting vessel for God's people. The Declaration of Independence, abolition of slavery, Civil War, prohibition, social work and welfare reform, labor rights, anticommunism, the Civil Rights movement, and current antiabortion politics—all were argued in the name of God. The myths of the righteous citizen and 'the city on a hill' not only fueled progressive reforms throughout American political history but also legitimated less-than-righteous foreign policies.

Finally, theorists argue that religion is relevant to American politics because of the permeability of its political structures. Its systems of federalism, division of powers, checks and balances, multitudes of interest groups, etc., all invite citizens to find the appropriate venue for success with their policy or issue (summarized in Wald 1997). For example, if religious groups cannot get an anti-abortion amendment out of Capitol Hill, and if the White House abandons them, they can turn to Supreme Court appointments, or they can argue for a shift in most abortion decisions at the state level (e.g., *Webster vs. Reproductive Health Services 1989*), where it is easier to mobilize a successful coalition locally. Most political careers in the USA depend on support at the micro level, and that is also where organized religion has its strength.

3. *Mechanisms that Enhance or Diminish the Linkage between Religion and Politics*

Much research in this field has addressed the political consequences of different ways of being religious (see Stark and Glock (1968) and Lenski (1963) for pioneering work on the dimensions of religiosity, and Leege and Kellstedt (1993) for more recent methodological

work on American National Election Studies datasets). The most persistent theme is that the creedal dimension, i.e., what people believe, has powerful effects on politics. Through a complex model, Glock and Stark (1966) showed that the more orthodox the Christian beliefs, the more likely an American is anti-Semitic. The findings remained controversial, however, because of the measures for orthodoxy: The items tap beliefs common to the evangelical Protestant tradition but overlook beliefs that derive from sacramental/incarnational Christianity.

To avoid over-reliance on the beliefs dimension, many scholars, following the lead of Kellstedt (Leege and Kellstedt 1993, pp. 273–304) have developed indices from 'minimal' Christian beliefs/practices across several dimensions. Regarding the partisan and ideological consequences of religion, they conclude that the higher the level of religiosity, the greater the Republican loyalty and degree of conservatism—regardless of denomination (Green et al. 1996, especially Chaps. 14 and 15). Leege (1996), however, has shown that among Catholics generational and gender differences override the effects of either beliefs alone or multiple-measure indices of religiosity. Further, higher levels of religiosity among African-American Christians and among Jews are typically associated with some aspects of political liberalism and Democratic loyalty (see recent studies in Smidt and Penning (1997)). Scholars recognize a large unfinished agenda on the measurement of beliefs and practices, especially in the over-reporting of church attendance.

In some respects, greater progress is shown in the political consequences of religion as a group phenomenon. Churches and religious traditions are often reference groups, waxing and waning in salience for their adherents. They develop interests, seek either to protect their values from threat or to evangelize their values, or become the target of persecution. The last is particularly important to political mobilization. Converse's (1966) study of Catholics in the 1960 election showed that Protestant Democrats not only left their party's Catholic nominee, but that Catholic Republicans crossed over to Kennedy, and all Catholics were more highly mobilized. Those most likely to show the mobilization effects and to be drawn to Catholic Kennedy were not those who attended mass regularly but those who identified most closely with being Catholic. Social identification trumped social cohesion, i.e., regular interaction with fellow group members (see Turner in Tajfel (1982) for this distinction). Further, reinforcing group loyalties enhanced the attraction of Kennedy, e.g., Irish urban Catholics were drawn more to him than Polish nonurban Catholics. Thus, religious effects partake also of other cultural group identifications.

Nevertheless, frequency of interaction with fellow religionists can have profound political effects. In carefully controlled studies, group norms about political values and candidate selection are shown to be propagated through behavioral contagion among church attenders (Wald et al. 1988). Where fellow church members are also informal discussion partners, political views are most likely to be reinforced and turned into political actions (Gilbert 1993, Huckfeldt et al. 1993). Even if religious congregations are composed of like-minded people, Jelen (1991) has argued, religiously based political mobilization is unlikely to occur until a negative outgroup, an external enemy, is identified.

Many leaders, both religious and political, attempt to define political agendas, issues, and choices in ways that will mobilize or demobilize religious adherents. This may or may not have political consequences. Earlier, Hadden (1969) warned that liberal activists among the Protestant clergy involved in civil rights and antiwar efforts were driving a wedge between their churches and their congregants. Quinley (1974) argued that the division was not universal, but that there was something of a two-party system emerging among the clergy: Liberals committed to social change addressed largely justice and peace issues, whereas conservatives addressed those moral issues having to do with sexuality, reproduction, family life, and national moral decline; still, the latter were less overtly involved in politics. An important study of the politics of clergy from eight Protestant denominations (excluding the catholic churches such as Anglican, Lutheran, and Roman Catholic) shows that the two-party division is very evident, the gap in political involvement has closed, and the values of the clergy can be traced directly to theological differences (Guth et al. 1997).

Do the laity follow the beat of these ordained drummers? That is a more complex matter. Most studies are limited to comparing the values and actions of clergy samples with the values and actions of laity samples from their denominations. Already, we observed that most of the action in American religion and politics is local; therefore, an appropriate design for assessing clerical cue-giving effects would use the parish or congregation as the unit of analysis. While the possibilities of controlled studies can be seen (e.g., Roozen et al. (1984), Leege et al. (1984–9), Jelen (1991), Olson (2000)), they are still largely at a typological development stage and offer samples of such small sizes that generalizations are difficult.

Promising hypotheses include the following: (a) the more consummatory the loyalty to a local church, the more influence its pastor has over the religious and political life of its people; (b) the deeper the theological understanding of each member, the more likely that moral values shared by clergy and congregants will guide political choices; (c) the greater the sense of oppression, threat, or status loss felt by church members in modern society—where the subjects of concern are the norms by which we should live—the more the receptivity to cues from religious leaders; (d) the smaller and more homogeneous or selective the religious body, the more dominant its leader's political

views; (e) the greater the legitimacy of corporate political involvement by the church, the more influential its leadership on a course of political action; (f) the more the issue concerns activities that individuals or families can influence (e.g., consequences of sexuality) rather than the product of impersonal economic forces (e.g. plant closings), the more religious leaders can define the political agenda; and (g) the greater the individual's intrinsic level of interest in politics, the more he/she will hear political cues when given by a religious leader. Impediments to political leadership by clergy abound: (a) separation of church and state illegitimatizes the more overt political actions by religious groups; (b) religious groups specialize in salvation, not public policy; (c) religious groups can lose tax exemptions and other privileges when they become too political; (d) within the pluralistic civil society, the church is only one of many groups competing for loyalty and influence; and (e) within the democratic political process, compromise is more characteristic than success on issues defined by churches as zero-sum.

Beyond the local level, churches engage vigorously in politics. Noting a leadership vacuum in many Republican county and state party organizations, Religious Right groups such as the Christian Coalition were able to consolidate dominance over about half of the state party central committees (Rozell and Wilcox 1997). On the policy advocacy front, Wuthnow (1988) has noted a great increase in the number of religious interest groups that are arranged across denominational divides. A number of scholars (see Hertzke (1988) as representative) have examined the organization and tactics of denominational and ecumenical interest groups in Washington. Although their official lobbies are small, they use the full range of inside-the-Beltway tactics. Given the decentralization of policy-making, the larger church bodies or coalitions have similar lobbies at the state level. Outside the Beltway, church interest groups seek to mobilize home district pressure through direct mail and fund-raising, letter writing and contacting, candidate ratings and thinly disguised endorsements, get-out-the-vote drives, and media dramaturgy.

A few studies have shown the impact of the public officials' own religious beliefs and practices on their policy preferences and actions. The most comprehensive of these (Benson and Williams 1982) shows that substantial segments of both political conservatives and political liberals act from religious predispositions. The former affirm an individualism-preserving religion and politics (emphasis on the ways laws create order and protect individuals from their own worst impulses), whereas the latter embrace a community-building religion and politics (emphasis on social transformation, justice, and peace). This study inspired later work by Leege and Welch (1989) where the same types were found in mass samples of Catholic laity, and by Guth et al.'s (1997) clergy study. Thus, religiously inspired public officials, clergy, and laity appear to have in common certain foundational religious beliefs that divide them from each other both theologically and politically. At the same time it is important to note that (a) a substantial proportion of public officials are only nominally religious, and (b) representational roles (read: the imperative to be re-elected) constrain translating religious beliefs into public policy.

Finally, political elites in the USA need religious groups if they are to achieve their goals of election/re-election. Religious group membership provides a useful basis for political market segmentation. Strategic politicians can mobilize religious group members by advocating policy interests, stressing group affinities (e.g., Kennedy as a modern Catholic, Reagan as an apostate Irish Catholic), or symbolically pillorying the 'enemies' of a religious group. Effective as mobilization is, demobilization of a religious group which composes part of a rival party's majority coalition has the same impact on chances of victory. This is done by stressing how the rival party has deserted the religious group's values (Leege et al. (in press) apply social attribution theory to understand this strategy). Over time, crystallized party images come to include the policy positions of religious groups and even the group itself. For example, the public associated evangelicals and fundamentalists with the Republican Party as early as 1964 and 1980, but they became central to the party image from 1988 to 1996.

Decisions to rely on churches for political mobilization are well advised. Using a resource theory of political participation, Verba et al. (1995) show that churches nurture skills that easily transfer to the political sector. Thus, minorities such as African-Americans associated with black churches participate more than a socioeconomic model would predict, while many Latinos associated with Catholic parishes not developing such skills participate less. Wald (1997) argues that the following skills are germane to politics derive from church participation in the USA: (a) social skills in listening, mediating, and leading; (b) awareness of public issues from a moral perspective; (c) encouragement to join other civic and community betterment activities; (d) a conviction that there is a sacred character to social obligations that transcends self-interest; and (e) the self-esteem that derives from practice with public assignments.

Empirical evidence, then, sustains the argument that religion has measurable effects on American politics, at the same time that certain styles of influence attempts are circumscribed. Basic problems of conceptualization and measurement, underspecified models, endogeneity, failure to anchor puzzles in strong social theory, and inappropriate samples continue to characterize this developing field of inquiry.

See also: Church and State: Political Science Aspects; Religion: Mobilization and Power; Religion, So-

ciology of; Religiosity, Sociology of; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State: Cultural Concerns; Social Movements: Resource Mobilization Theory; Weber, Max (1864–1920)

Bibliography

Benson L 1979 Marx's general and middle-range theory of social conflict. In: Merton R K, Coleman J S, Rossi P H (eds.) *Qualitative and Quantitative Research: Papers in Honor of Paul F. Lazarsfeld*. Free Press, New York

Benson P L, Williams D L 1982 *Religion on Capitol Hill: Myths and Realities* 1st edn. Harper & Row, San Francisco

Converse P E 1966 Religion and politics in the 1960 election. In: Campbell A, Converse P E, Miller W E, Stokes D T (eds.) *Elections and the Political Order*. Wiley, New York, pp. 96–124

Dahl R A 1956 *A Preface to Democratic Theory*. University of Chicago Press, Chicago

de Tocqueville A 1958 (1835/1840) *Democracy in America* (trans. Bradley P). Vintage, New York

Easton D 1953 *The Political System: An Inquiry into the State of Political Science* 1st edn. Knopf, New York

Finke R, Stark R 1992 *The Churching of America, 1776–1990: Winners and Losers in Our Religious Economy*. Rutgers University Press, New Brunswick, NJ

Fowler R B 1989 *Unconventional Partners: Religion and Liberal Culture in the United States*. W. B Eerdman's, Grand Rapids, MI

Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York

Gerth H H, Mills C W 1946 *From Max Weber: Essays in Sociology*. Oxford University Press, New York

Gilbert C P 1993 *The Impact of Churches on Political Behavior: An Empirical Report*. Greenwood, Westport, CT

Glock C Y, Stark R 1966 *Christian Beliefs and Anti-Semitism* 1st edn. Harper & Row, New York

Green J C, Guth J L, Smidt C E, Kellstedt L A 1996 *Religion and the Culture Wars: Dispatches from the Front*. Rowman and Littlefield, Lanham, MD

Guth J L, Green J C, Smidt C E, Kellstedt L A, Paloma M M 1997 *The Bully Pulpit: The Politics of the Protestant Clergy*. University Press of Kansas, Lawrence, KS

Hadden J C 1969 *The Gathering Storm in the Churches*. 1st edn. Doubleday, New York

Hertzke A D 1988 *Representing God in Washington*. University of Tennessee Press, Knoxville, TN

Hertzke A D 1993 *Echoes of Discontent: Jesse Jackson, Pat Robertson, and the Resurgence of Populism*. CQ Press, Washington, DC

Huckfeldt R, Plutzer E, Sprague J 1993 Alternative contexts of political behavior: Churches, neighborhoods, and individuals. *Journal of Politics* **55**: 365–81

Jelen T W 1991 *The Political Mobilization of the Religious Beliefs*. Praeger, New York

Lasswell H D 1958 *Politics: Who Gets What, When, How?* Meridian Books, New York

Leege D C 1996 The Catholic vote in '96: Can it be found in church?'*Commonweal* **27**(September): 11–8

Leege D C (ed.) 1984–9 *The Notre Dame Study of Catholic Parish Life, Reports 1–15*. Institute for Church Life, Notre Dame, IN

Leege D C, Kellstedt L A 1993 *Rediscovering the Religious Factor in American Politics*. M E Sharpe, Armonk, NY

Leege D C, Wald K D, Krueger B R, Mueller P D in press *The Politics of Cultural Differences: Social Change and Voter Mobilization Strategies in the Post-New Deal Period*. Princeton University Press, Princeton, NJ

Leege D C, Welch M R 1989 Religious roots of political orientations: Variations among American Catholic parishoners. *Journal of Politics* **51**: 137–62

Lenski G 1963 *The Religious Factor*. Doubleday, Garden City, NY

Olson L R 2000 *Filled with Spirit and Power: Protestant Clergy in Politics*. State University of New York Press, Albany, NY

Quinley H E 1974 *The Prophetic Clergy: Social Activism Among Protestant Ministers*. Wiley, New York

Roozen D A, McKinney W, Carroll J W 1984 *Varieties of Religious Presence: Mission in Public Life*. Pilgrim Press, New York

Rozell M J, Wilcox C (eds.) 1997 *God at the Grass Roots 1996: The Christian Right in the American Election*. Rowman and Littlefield, Lanham, MD

Smidt C E Penning J M (eds.) 1997 *Sojourners in the Wilderness: The Christian Right in Comparative Perspective*. Rowman and Littlefield, Lanham, MD

Stark R, Glock C Y 1968 *American Piety: The Nature of Religious Commitment*. University of California Press, Berkeley, CA

Tajfeh H 1982 *Social Identity and Intergroup Relations*. Cambridge University Press, New York

Verba S, Schlozman K L, Brady H E 1995 *Voice and Equality: Civic Voluntarism in American Politics*. Harvard University Press, Cambridge, MA

Wald K D 1997 *Religion and Politics in the United States*, 3rd edn. CQ Press, Washington, DC

Wald K D, Owen D E, Hill S S 1988 Churches as political communities. *American Political Science Review* **82**: 531–48

Weber M 1946 *The Protestant Ethic and the Spirit of Capitalism* (trans. Parson T). Scribner's, New York

Wildavsky A 1987 Choosing preferences by constructing institutions: A cultural theory of preference formation. *American Political Science Review* **81**: 3–21

Wuthnow R 1987 *Meaning and Moral Order: Explorations in Cultural Analysis*. University of California Press, Berkeley, CA

Wuthnow R 1988 *The Restructuring of American Religion*. Princeton University Press, Princeton, NJ

D. C. Leege

Religion and Science

The intersection of religion and sciences has been both more complex and less conflictive than is frequently depicted by those who cite the trial of Galileo, the dispute over the theological implications of Charles Darwin's theory of natural selection, and the Scopes Trial as evidence of the historical incompatibility of religious beliefs and scientific endeavor. This article will first examine the conflict thesis, then survey the history of the relations between religion and science in the Western tradition from the fifth century BC to the

end of the twentieth century with special reference to Christianity.

1. The Historiography of Religion and Sciences

The thesis that Christianity has been historically hostile to science was popularized by two influential works that appeared in the late nineteenth century: John William Draper's (1811–1882) *History of the Conflict Between Religion and Science* (1874) and Andrew Dickson White's (1832–1918) *A History of the Warfare of Science with Theology in Christendom* (1896). White, the first president of Cornell University, argued that Christianity had a long history of opposing scientific progress in the interest of dogmatic theology. Wedding a triumphalist view of science to a dismissive view of theology, his thesis struck a responsive chord in American thought, which was at the turn of the century increasingly committed both to a secular outlook and to recognizing the central role that the sciences played in modern society. The Draper–White thesis became enormously influential in America both among academics and in the popular mind. It came to dominate the historical interpretation of the relations of religion to science during much of the twentieth century. Reflecting a positivist outlook, it viewed science as continually progressing and overcoming the entrenched antagonism of religious opinions, which invariably retreated before its advance.

Some historians regarded the Draper–White thesis as oversimplifying and distorting a complicated relationship. Thus, Burtt, in *The Metaphysical Foundations of Modern Physical Science* (1932), suggested that, far from being hindered by religion, science was in fact indebted to theological ideas. Alfred North Whitehead, in *Science and the Modern World* (1926), rejected the view that in controversies between science and religion, religion was wrong, and … science was always right. The true facts of the case are very much more complex, and refuse to be summarized in these simple terms. A number of important studies, by Robert K. Merton, Herbert Butterfield, and Alexandre Koyré, among others, also challenged the conflict thesis.

A smaller number of historians of science went even further to argue that specific components of Christian theology were responsible for the scientific revolution. Reijer Hooykaas contended that Calvinism's accommodationist understanding of Scripture and its voluntaristic theology encouraged scientific endeavor. Similarly Stanley Jaki, in *Science and Creation* (1974), maintained that Christian theism, the doctrine of Creation, and the belief that the universe is governed by a rational and benevolent deity were necessary preconditions of the development of modern science.

In the last three decades of the twentieth century several scholars have provided a more systematic reevaluation of the conflict thesis. Two major contributions to the literature came from Lindberg and Numbers (1986) and Brooke (1991). In the introduction to their composite volume, Lindberg and Numbers argued for an historical picture that recognized 'a complex and diverse interaction that defies reduction to simple "conflict" or "harmony"' to replace the largely polemical one drawn by Draper and White. Brooke objected to imparting essentialist understandings to words like 'science' and 'religion,' which had many different meanings over time and could not be reduced to simple static concepts. 'Natural philosophy,' for example, denoted something very different in the seventeenth century from 'science' in the twentieth. The growing acknowledgment that in their relations religion and the sciences have exhibited a multiplicity of attitudes, reflecting local conditions and particular historical circumstances, led Brooke to suggest that a 'complexity thesis' is a more accurate model than the familiar conflict thesis. The views of Brooke, and Lindberg and Numbers have gained increasing acceptance among professional historians of science.

2. Greece and Rome

Western science as an explanatory enterprise began with the Greeks. It was in the sixth century BC that the Pre-Socratic philosophers like Thales (fl. 585 BC), Anaxagoras (c. 500–428 BC), and Anaximander (c. 610–546 BC) challenged the mythopoeic picture of the natural world. Most Pre-Socratics understood the world in nonmythological terms, viewing it as the product of natural forces. In the late fifth century BC they experienced some popular opposition in Athens. In c. 437–436 BC Anaxagoras was prosecuted on a charge of impiety and his books publicly burned.

By the end of the fifth century BC natural philosophy (physical science) had come to be recognized by the Greeks as offering the most suitable means of understanding the world of nature and religious issues and hereafter were to play a limited role in either hindering or encouraging scientific endeavor in the classical world. This does not mean that religious ideas were of no account in formulating world views for the Greeks or for the Romans, who took over the philosophical and scientific culture of the Greeks. The Epicureans explicitly repudiated a religious view of the universe while for the Stoics, science (especially cosmology) and religious thought were related intimately. Moreover, the Greek view of the world was predominantly an organicist or vitalist one. Philosophical naturalism, which was espoused by the Pre-Socratic philosophers, the sophists, Aristotle (384–322 BC), and the Epicureans, always remained a minority opinion in Graeco–Roman thought. In general the Greeks conceived of nature as a divinity. Plato (c. 427–347 BC) believed that a world soul (psyche)

animated the universe, in which the sun, the moon, the stars, and the planets were gods. A vitalist world picture, stripped of its pagan elements, was transmitted by the Romans to Western Europeans, who continued to hold it well into the Middle Ages.

3. *Late Antiquity and the Middle Ages*

Christianity, founded by Jesus of Nazareth (c. 4 BC to 29 AD), spread rapidly throughout the Roman Empire owing to its extensive missionary activity. In the second century AD Christian apologists (theologians who defended their faith philosophically against pagan critics) began the process of harmonizing Christian revelation with Graeco–Roman philosophy. Apologists like Justin Martyr (c. 100–165), Tertullian (c. 160–220), and Irenaeus (c. 130–200) were intellectuals who had received their education in classical culture and were influenced strongly by it. Their opinion of natural philosophy, which was a component of Greek philosophy and inseparable from it, was ambiguous. While they often denounced it for its pagan religious or anti-Christian elements, they employed its methods, thought forms, and vocabulary—and drew on it when it was useful in the rational defense of Christianity. These church Fathers hellenized Christianity by taking over elements of classical culture and incorporating them into the Christian world view. Tertullian, for example, who asked, 'What indeed has Athens to do with Jerusalem? What harmony is there between the Academy and the church?' is sometimes cited as an example of early Christian anti-intellectualism. Yet he often borrowed from Greek natural science and medicine to buttress his theological arguments. Not all Christian apologists accepted Greek science, however. A few, like Tatian (second century), who were influenced by ascetic ideals or concerns about potential demonic influences, were either guarded in their acceptance or hostile in their approach to Greek pagan natural philosophy.

The most influential of the church Fathers was Augustine of Hippo in North Africa (354–430). Augustine shared the reservations of some of the Fathers about the value of natural philosophy but he recognized its utility. Like other Fathers he rejected *curiositas*, the desire to pursue knowledge for the sake of knowing, which he regarded as a characteristic feature of pagan learning. But Augustine believed that the study of natural philosophy was legitimated when it served higher purposes, such as aiding in the understanding of Scripture or lending support to Christian theology. By becoming the handmaiden of Christian theology natural philosophy was assured a place in the emerging Christian culture, one that was to last through the Middle Ages and well into the early-modern period.

The early Middle Ages (c. 500–1000) was a period of disintegration that followed the collapse of the Roman Empire in the West under repeated attacks and occupation by Germanic barbarians. As towns became villages, public services ceased to exist, and roads became impassable, the cultural continuity with the classical world was nearly severed. Educational institutions disappeared and learning remained alive only in Christian monasteries. A few distinguished figures stand out from the intellectual gloom of the period for their interest in natural philosophy, much of it derived from Neoplatonic sources. They include Cassiodorus (c. 485–580), Isidore, bishop of Seville (c. 560–636), the Venerable Bede (c. 673–735), and John Scotus Erigena (c. 810–877). These men could do little more than preserve as much as possible of the accumulated classical heritage, including natural philosophy, but they did so as clerics, within a monastic framework, and with the confidence that scientific knowledge, as a handmaiden to theology (*philosophia ancilla theologiae*), was worthy of being transmitted to future generations.

The eleventh and twelfth centuries experienced the revival of trade between Western Europe and the eastern Mediterranean, together with the concomitant renewal of urban life, the growth of manufacturing, and the beginning of the gradual transformation of Europe from a rural to an urban economy. Cathedral schools sprang up in cities and gradually replaced monastic schools. In the twelfth century natural philosophers avidly studied Plato's *Timaeus* and attempted to harmonize its cosmology with the Creation account in Genesis. The Arabs had preserved Greek learning, and works by Greek authors on natural philosophy were translated into Latin from Arabic, thus making available much new scientific material. The most important were the extant works of Aristotle, which by the thirteenth century had been translated in their entirety, replacing the influence of Plato and creating a major intellectual transformation of Western European universities. Muslim contributions (for example, in optics) were significant as well.

There was much in Aristotle's writings that was useful to European Christians and that could be harmonized with Christian theology. But there were also areas of potential conflict, such as Aristotle's belief that the world is eternal, that the soul is not immortal, and that the universe can be accounted for by naturalistic and even deterministic explanations apart from divine providence. These and other incompatible doctrines caused the teaching of Aristotle's natural philosophy to be banned from the University of Paris (site of the preeminent theological faculty in Europe) in 1210 and 1215, but by 1255 his writings had gained official sanction and they came to be widely studied. A number of distinguished scholars, including Roger Bacon (1213–1291), Albert the Great (1200–1280), and his student Thomas Aquinas (c. 1224–1274), attempted to create a synthesis of Aristotelian philosophy and Christian theology that would rec-

oncile the Christian faith and classical reason. The existence of a more radical faction at Paris, headed by Siger of Brabant (c. 1240–1284), led the bishop of Paris, Etienne Tempier, to condemn, at the request of traditional theologians, 13 philosophical propositions espoused by Siger. In 1277 he increased the number of propositions condemned to 219. The condemnations represented a conservative backlash that placed limits on speculation in natural philosophy. In the long run, however, they provided arts masters at Paris with incentives to explore other scientific options that lay outside Aristotelian natural philosophy, particularly by way of 'thought experiments' that entertained hypothetical possibilities, such as the existence of other worlds or of a vacuum. In fact, freedom of scientific speculation was relatively broad in late medieval universities, which were supported by the church.

4. *The Copernican Revolution and Galileo*

When joined with Aristotelian physics, the geocentric theory of the solar system, as refined by Ptolemy of Alexandria (second century AD), was almost universally accepted in the West and taken to be compatible with their theologies by Jews, Muslims, and Christians until the seventeenth century. The first major challenge to the Ptolemaic theory came with the publication of Nicolaus Copernicus's (1473–1543) *De revolutionibus orbium coelestium* (*On the Revolutions of the Heavenly Bodies*), which appeared after the author's death in 1543. Copernicus advocated a heliocentric model of the solar system. How he came to espouse it is unknown but he might never have published it had it not been for the encouragement of a young Lutheran astronomer, Georg Rheticus (1514–1574) of Wittenberg. Although they initially rejected its heliocentric cosmology, Lutherans at the University of Wittenberg, including the Protestant reformer Philip Melanchthon (1497–1560), were attracted to the mathematical convenience of the new theory. Elsewhere, however, European natural philosophers rejected it for a variety of reasons, many of which were scientific rather than theological. In the year 1600 there were probably no more than 10 Copernicans in all Europe.

One scientist who was quick to embrace the new theory was Galileo Galilei (1564–1642), a young professor of mathematics at the University of Padua. Although secretly a Copernican by 1597, Galileo did not publicize his views until 1613, when in *Letters on Sunspots* he declared his acceptance of the heliocentric system. It was, however, his *Letter to the Grand Duchess Christina* (1615), in which Galileo argued that Scripture did not intend to impart scientific truth but accommodated itself to popular language in describing the natural world, that provoked the Catholic Church's response. Although a layman, Galileo had undertaken to interpret the Bible. In the following year (1616) he appeared before the Congregation of the Holy Office (the Inquisition). Cardinal Bellarmine stated his willingness to accept the Copernican theory if it could be proven (which it could not be at the time); otherwise Scripture ought to be read literally as teaching geocentricity. The Church declared the theory 'false and contrary to Holy Scripture' and recommended that *De revolutionibus* be 'suspended until corrected.' Galileo was warned not to teach or defend heliocentrism except as a hypothesis until it could be proven. In 1632, after maintaining silence for 16 years, Galileo published his *Dialogue on the Two Principal World Systems*, an ostensibly even-handed debate between proponents of the Ptolemaic and Copernican theories in which the arguments for the Ptolemaic theory were made to look feeble. Galileo was summoned to Rome, where he stood trial in 1633. Compelled to abjure his belief that the earth moved, he was confined to a country estate near Florence, where he continued writing until his death in 1642.

The traditional picture of Galileo as a martyr to intellectual freedom and a victim of the Catholic Church's hostility to science is little more than a caricature. The trial of Galileo must be understood against the backdrop of his own times. A vain and ambitious man, he did not suffer fools gladly and he had a penchant for making enemies in high places. Among those he alienated was Pope Urban VIII (1623–1644), who had earlier encouraged Galileo's researches on the motion of the earth, but came to believe that Galileo had betrayed him in deliberately violating an injunction not to teach Copernicanism. Moreover, strong reaction against Copernicanism (and Galileo's espousal of it) came from the Aristotelians of the scientific establishment, some of whom were determined to entangle Galileo with the Church. In particular, Galileo aroused the ire of a loosely-organized movement of traditional scientists called the *Liga*, which succeeded in introducing theological issues into scientific discourse and attacked Galileo for his views.

In contrast with the relative freedom of intellectual debate that natural philosophers in European universities had enjoyed in the late Middle Ages, the Protestant Reformation had made the Catholic Church more cautious. In the question of whether Scripture ought to be read as teaching a geocentric universe, the Church was anxious to prove its good faith in interpreting Scripture literally in the face of Protestant attacks on its allegorizing tendencies in biblical exegesis. Aristotelianism became, in its Thomistic form, a rigid dogma and the papal court became, after the Council of Trent (1545–1563), more authoritarian. The Inquisition and the Index of Prohibited Books, introduced into Italy by forces of the Counter Reformation, were employed against Galileo. The prosecution of Galileo represents not the *cause célèbre* in the alleged warfare of science with theology, but a dispute that involved multiple issues, including

powerful personalities and vested professional and ecclesiastical interests, which complicated Galileo's championing of a revolutionary change in astronomical models.

5. *The Mechanical Philosophy and Natural Theology*

By the mid-seventeenth century, the Copernican theory had come to be accepted widely throughout Europe, owing largely to Galileo's promotion of heliocentrism. The seventeenth century saw, too, the formulation of a mechanical model of the universe, which attempted to account for all natural phenomena in terms of matter and motion without reference to action at a distance. The mechanical philosophy had its roots in corpuscular materialism, which explained physical phenomena by the interaction of tiny particles of matter called atoms. While elements of the model had been adopted by several of the leading figures of the scientific revolution, its most influential advocates were René Descartes (1596–1650) and Pierre Gassendi (1592–1655), who sought to devise a new philosophy that would replace Aristotelianism. While Descartes and Gassendi grounded their mechanical philosophies in theological considerations, critics feared that their views would lead to materialism, as they did in the case of Thomas Hobbes (1588–1679), or to deism, as they did in the case of Voltaire (1694–1778).

The mechanical philosophy received its greatest synthesis at the hands of Isaac Newton (1642–1727) in his *Principia Mathematica Naturalis Philosophiae* (*Mathematical Principles of Natural Philosophy* (1687)). By his discovery of the principle of universal gravitation, Newton succeeded in harmonizing celestial and terrestrial mechanics, bringing for the first time the whole of nature into a precise rational interpretation. Newton's picture of a universe governed by uniform laws that could be described without recourse to primary causes resulted in the widespread belief that nature was self-governing and aided the 'natural religion' of deism and its idea of a 'watchmaker' God. It is ironic that Newton, himself a Christian (although an Arian), inadvertently laid the groundwork for the rationalism of the eighteenth-century Enlightenment. On the basis of his synthesis of the mechanical philosophy David Hume (1711–1776) developed a materialistic philosophy in England, while in France philosophers like Julien Offray de La Mettrie (1709–1751) and Baron d'Holbach (1723–1789) embraced atheistic materialism.

The popularity of the mechanical philosophy led in the latter half of the seventeenth century to the creation of a natural theology that continued into the nineteenth century. Natural theology describes the existence and attributes of God by means of knowledge derived from natural reason rather than from special revelation. There were several grounds for its attraction: the wonders of scientific discoveries, many of which could be seen for the first time with the invention of the microscope; the desire to secure an adequate foundation for a theology of divine providence that would confront the perceived dangers of materialistic atheism; and the enormous influence of Newton's mechanical picture of the world. The argument from design became the cornerstone of a natural theology, in which God was described as a craftsman, a geometer, or an architect. Not all exponents of natural theology were deists; many theists, especially those who were repelled by religious enthusiasm, were involved in the enterprise of demonstrating the majesty of the Creator from the intricacies embedded in his creation. In William Paley's (1743–1805) influential *Natural Theology, or Evidences of the Existence and Attributes of the Deity* (1802) natural theology reached its apex.

6. *Geology and Darwinian Evolution*

The influence of the Enlightenment, with its exaltation of reason and rejection of special revelation, can be observed in the developments that were occurring in science, particularly in the increasing tendency to seek naturalistic explanations and to divorce science altogether from theological considerations. In the seventeenth and eighteenth centuries theories of the earth had sought to weave together geology and theology in what was termed 'physico-theology.' Many of these theories calculated the age of the earth by using biblical chronologies in a way that could be harmonized with physical principles. Invariably they postulated a young earth. While most attributed organic fossils to the Noachian Flood, by the eighteenth century it was becoming increasingly difficult to explain the global distribution of fossils by a single event (the Deluge) that had taken place a few thousand years ago. The French naturalist Georges Louis Leclerc, Comte de Buffon (1707–1788), a deist, proposed a much older earth (publicly of 75,000 years and privately perhaps as old as three million years) and suggested that each of the six biblical days of Creation might correspond to an immense length of time. Georges Cuvier (1769–1832), a devout Protestant, demonstrated the existence of fossil mammals unlike any in his own day and without any apparent intermediate forms. He argued that a series of geologic catastrophes had eliminated species in past ages. Cuvier's catastrophes were not global and in his case did not derive from his religious convictions. But they were later read as such and were appropriated in English natural theology to support catastrophism, which became a popular form of physico-theology that attributed geological changes to great catastrophes of the past, the most recent being the biblical Flood.

The Scottish geologist Charles Lyell (1797–1875) rejected altogether providential interventions in geologic history. In his *Principles of Geology* (3 vols. 1830–1833) he succeeded in severing geology from theology. Lyell rejected diluvialism (catastrophism) for what became known as uniformitarianism: the theory that geological changes on the earth's surface have occurred slowly over long periods of time and that in explaining change we refer to processes that we currently observe. Previous proponents of diluvialism, like William Buckland (1784–1856) of Oxford and Adam Sedgwick (1785–1873) of Cambridge, quickly abandoned catastrophism and adopted the new geology, which Christians harmonized with the Genesis account of Creation in a relatively painless fashion. A group of laymen, who called themselves Mosaic or Scriptural geologists, continued to take their geology from Genesis but they were marginalized since nearly all professional geologists abandoned the Noachian Flood as an organizing idea in accounting for changes in the earth's history.

Charles Darwin (1809–82) had, as a young man, been influenced by Paley's natural theology. During his voyage aboard the H.M.S. *Beagle*, however, he read Lyell's work and began to think that slow changes had produced the geological phenomena that he observed. His theory of evolution by natural selection, published in *Origin of Species* (1859), made use of Lyell's theories that the earth was very old and the fossil record imperfect. Evolutionary ideas had preceded Darwin. Originating in the Enlightenment, they had already been proposed in the eighteenth century by Charles's grandfather, Erasmus Darwin (1731–1802), and by Jean-Baptiste Lamarck (1744–1829), the latter proposing the theory that acquired characteristics were inherited. By the mid-nineteenth century the belief that transmutationism (evolution) represented a divine plan was widespread. In advancing natural selection as a mechanism for evolution, Darwin argued that in the 'struggle for existence' those individuals with even a slight advantage would survive and pass their advantage on to their offspring. Herbert Spencer (1820–1903) referred to it as the principle of the 'survival of the fittest.'

Darwin's theory met with a mixed response that cut across conventional professional and religious boundaries. Some scientists initially rejected it, while a number of clergymen accepted it. Conservative anti-Darwinists, including several leading naturalists like Louis Agassiz of Harvard (1807–1873), geologists like Canadian William Dawson (1820–1899) of McGill, and theologians like Charles Hodge (1797–1878) of Princeton Seminary rejected it, fearing the break with final causes. A second group of Christians accepted evolutionary theory as compatible with a biblical faith. They included the Scot James McCosh (1811–1894), president of Princeton University, and the American botanist Asa Gray (1810–1888) of Harvard. A third group came to believe that evolution had rendered supernatural religion untenable and opted (most often) for some form of philosophical materialism.

By the 1870s, however, most scientists and a large body of the educated public had accepted some form of evolution (though not necessarily natural selection). Full acceptance of Darwin's views on human nature, as set forth in his *Descent of Man* (1871), was more difficult for Christians. Darwin maintained that moral values were rationalizations of social instincts that had developed by natural selection, thus asserting a full-blown materialism. Alfred Russell Wallace (1823–1913), the codiscoverer of the theory of evolution by natural selection, rejected materialism, maintaining that the creation of the human mind required non-material forces.

In the 1920s evolution was one of several theological issues that divided the major Protestant denominations of North America in the modernist–fundamentalist controversy. While many Protestants (as well as some Roman Catholics and Jews) had come to believe in an old earth and to harmonize belief in evolution with a nonliteral interpretation of the Genesis account of Creation, conservatives emphasized in their rejection the atheistic and materialistic dangers inherent in a naturalistic theory of human origins. But the Scopes trial, held in Dayton, TN in 1925, symbolized the inability of biblical literalists to withstand the force of Darwinism, which was taken for granted by the intellectual elites of Europe and America. Anti-evolutionism was pilloried in the American press as a rearguard reaction to scientific progress and it was associated with nativism and rural and uneducated religious fundamentalism. Although biblical catastrophism was revived in the 1920s, it enjoyed favor only in conservative and sectarian Protestant circles in North America before it exploded into the popular creationist movement (called 'Flood Geology') in the 1960s.

7. *Twentieth-Century Physics and Cosmology*

While religious issues had been increasingly excluded from scientific discourse in geology and natural history (the study of natural objects, plants, minerals, and animals) in the nineteenth century, they had continued to be discussed in natural philosophy. From about 1895 modern physics began to deal with issues that went beyond those of classical physics, including quantum mechanics and special and general relativity. Several discoveries in modern physics challenged the determinism of the Newtonian mechanical model. Foremost was the indeterminacy principle of Werner Heisenberg (1901–1976), which demonstrated that one could not know perfectly both the position and momentum of a subatomic particle. Heisenberg's principle gave rise to much discussion of its wider

philosophical and theological implications, particularly relating to the possibility of intelligent design of the universe and human free will.

In cosmology too discoveries were made that suggested theological implications. Edwin Hubble's (1889–1953) discovery that the universe was expanding at a tremendous rate led the Roman Catholic priest and physicist Georges Lemaître (1894–1966) to propose what became known as the big-bang theory: that the universe had begun as a dense 'primeval atom' that had exploded in the distant past. It was challenged by the steady-state theory, developed in England and championed by Fred Hoyle (1915–), according to which the universe is eternal but matter is being continuously created in space. In the late 1960s, several observational discoveries led to the almost universal acceptance of the big-bang theory, which was argued by some researchers, like Paul Davies (1946–) and Frank Tipler (1947–), to strengthen belief in God's creation of the universe.

Another cosmological theory that engendered a good deal of theological discussion was the Anthropic Principle (a term coined by Brandon Carter in 1974), which asserts that the emergence of human beings in the universe is not an accidental by product of purposeless cosmic evolution. According to this theory the universe was so finely tuned during its earliest moments that if any one of several physical parameters had been slightly different it would have been unsuitable for the survival of human life. Hence, it is argued that the inhabited universe is anthropocentric, having been designed for human beings. The Anthropic Principle reintroduced into cosmological speculation at the end of the twentieth century issues that were long thought to have been laid to rest, including questions of teleology, intelligent design, and divine providence.

8. Conclusion

By the 1940s the modern synthesis of genetics and Darwin's theory of natural selection, which accounted for random variation by genetic mutation, made it possible for biological scientists to adopt a fully naturalistic position in both method and metaphysics. Naturalistic evolution, it was alleged, was able offer a complete explanation for the origin and development of the universe and for all forms of life, as well as for social factors like language, religion, and ethics. While scientists like George Gaylord Simpson (1902–1984) considered materialism to be the only philosophical position capable of scientific support, it was far from universally accepted within the scientific community. The last decade of the twentieth century, moreover, saw a resurgence of interest in the relationship of religion to the sciences. This interest arose from a variety of sources both within and outside the scientific community, including those of Christian and other theistic traditions (who most often asserted the complementarity of religion and the sciences), process philosophers, proponents of New Age and pantheistic spiritualism, and non-Western religious traditions. Perhaps it reflected the widespread recognition that, for all its achievements, science does not provide ultimate answers to questions regarding the meaning of the universe or of life itself.

See also: Darwin, Charles Robert (1809–82); Darwinism: Social; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionism, Including Social Darwinism; Historiography and Historical Thought: Christian Tradition; Historiography and Historical Thought: Islamic Tradition; Religion, History of; Religion, Phenomenology of

Bibliography

Barrow J D, Tipler F J 1986 *The Anthropic Cosmological Principle*. Oxford University Press, Oxford, UK

Barbour I G 1997 *Religion and Science: Historical and Contemporary Issues*. 1st rev. edn. Harper, SanFrancisco

Biagoli M 1993 *Galileo Courtier: The Practice of Science in the Culture of Absolutism*. University of Chicago Press, Chicago

Brooke J H 1991 *Science and Religion: Some Historical Perspectives*. Cambridge University Press, Cambridge, UK

Brooke J H, Cantor G 2000 *Reconstructing Nature: The Engagement of Science and Religion*. Oxford University Press, New York

Burtt E A 1932 *The Metaphysical Foundations of Modern Physical Science*, rev. edn. Harcourt Brace, New York

Durant J (ed.) 1985 *Darwinism and Divinity*. Blackwell, Oxford, UK

Fantoli A 1996 *Galileo: For Copernicanism and For the Church*. Vatican Observatory Publications, Vatican City, Italy

Ferngren G (ed.) 2000 *The History of Science and Religion in the Western Tradition: An Encyclopedia*. Garland, New York

Harrison P 1998 *The Bible, Protestantism, and the Rise of Natural Science*. Cambridge University Press, Cambridge, UK

Jaki S L 1974 *Science and Creation: From Eternal Cycles to an Oscillating Universe*. Science History Publications, New York

Koyré A 1957 *From the Closed World to the Infinite Universe*. Johns Hopkins Press, Baltimore, MA

Larson E J 1997 *Summer for the Gods: The Scopes Trial and America's Continuing Debate over Science and Religion*. Basic Books, New York

Lindberg D C, Numbers R L (eds.) 1986 *God and Nature: Historical Essays on the Encounter Between Christianity and Science*. University of California Press, Berkeley, CA

Lindberg D C, Numbers R L 2001 *Science and the Christian Tradition: Twelve Case Histories*. University of Chicago Press, Chicago

Moore J R 1979 *The Post-Darwinian Controversies: A Study of the Protestant Struggle to Come to Terms with Darwin in Great Britain and America, 1870–1900*. Cambridge University Press, Cambridge, UK

Russell R J, Stoeger W R, Coyne G V (eds.) 1988 *Physics, Philosophy, and Theology: A Common Quest for Understanding*. Vatican Observatory Publications, Vatican City, Italy
White A D 1896 *A History of the Warfare of Science with Theology in Christendom*. Appleton, New York

G. B. Ferngren

Religion and Youth

Religion has traditionally been an important aspect of personal and social identity. Socially, it provides a 'language'—a set of beliefs, values, and practices—which form an ultimate framework of meaning and order, binding people into one community rather than another. Individually, religion locates a person within a social context, conferring a sense of belonging, a way to understand the self in relation to the world, and a purpose to life. Western society is changing rapidly, however, bringing new uncertainties to the social context of identity formation (Giddens 1991). The last 30 years have seen economic restructuring, expansion of educational provision, new developments in communication and information technology, increased social and geographical mobility, changes in personal relationship and family norms, challenges to gender roles, and the generation of new attitudes and accepted ways of being. This raises the question as to what part religion plays in the formation of young people's identity today?

1. *Who are the 'Young People'?*

Before considering the relationship between religion and youth identity, it is important to note that the category 'young people' and the concept of 'youth' are problematic. There is no consensus on an upper age for 'young people' and surveys are reported using different age bands for the youngest groups. 'Youth' signifies a period of transition between childhood dependency and full adult responsibility but significant transitional markers, such as the school leaving age and access to employment and financial independence, vary between countries. There can also be age disparities within societies regarding different types of adult responsibility such as marriage and voting rights. Indeed, 'youth' is increasingly being separated from age altogether and becoming a valued personal quality to be retained throughout life. Furthermore, the experience of youth varies greatly according to gender, ethnicity, class, etc. Bearing these conceptual problems in mind, the picture of youth presented here is necessarily a generalized one and the discussion focuses on Christianity which remains the dominant religious tradition in the West (60 percent of Europeans and 76 percent of North Americans claim a Christian identity, Brierley 1997).

2. *Young People's Religious Beliefs and Practices*

The data on young people's religious beliefs and practices are complex and sometimes contradictory. Throughout the industrial West young people are less likely than older people to actively engage with institutional religion (Halman and de Moor 1993, Clark et al. 1995). This may reflect a stage in the life cycle, but the evidence suggests long-term change as well.

Table 1 illustrates declining religiosity among youth in Europe on a number of traditional variables and Table 2 indicates how religiosity varies among young people cross-nationally.

From Tables 1 and 2 it can be seen that while Church association is declining, private prayer and belief in God remains quite high. Belief in the soul and life after death is also apparent. Half of European youth and 79 percent of young people in the United States (Inglehart et al. 1998, Vol. 151) describe themselves as 'a religious person.' More locally, there is evidence to suggest that young people take an interest in supernatural phenomenon, nontraditional beliefs, and 'folk' religion. Table 2 indicates the extent of

Table 1
Data from the European Values Survey[a] for 18–24 year olds

	1981	1990
Church attendance—percentage who attend weekly or more	43	18
I am a religious person—percentage who agree	53	51
I take moments of prayer—percentage who agree	55	52
Spiritual beliefs—percentage who believe in:		
God	77	67
Life after death	54	50
The devil	26	23
Hell	23	20
Heaven	43	39
Sin	57	51
Reincarnation	27	26
A soul	63	65
I think often about meaning of life—percentage who agree	26	29

Source: Ashford and Timms 1992, pp. 124–8.
a The European Values Survey samples adults from Great Britain, Northern Ireland, Republic of Ireland, Western Germany, The Netherlands, Belgium, France, Italy, Spain, and Portugal.

Table 2
International comparison of religious beliefs and practices among 16–29 year olds based on the 1990–93 World Values Survey

Country (predominant denomination)[a]	Religion very important in your life (%)	Sometimes or often pray to God outside of religious services (%)	Believes in God (%)	God is very important in your life (%)	Believes in Reincarnation (%)
Ireland (RC)	22	65	95	56	18
Northern Ireland (M)	28	56	96	49	24
Spain (RC)	09	39	78	33	28
East Germany (P)	08	17	24	15	10
Britain (P)	06	24	63	16	30
Italy (RC)	25	58	89	58	28
The Netherlands (M)	09	33	52	20	23
Belgium (RC)	06	23	51	24	20
Austria (RC)	13	44	78	42	36
France (RC)	08	26	51	15	32
Canada (RC)	21	50	85	51	33
United States (P)	46	70	95	68	16
West Germany (M)	05	26	62	25	26
Denmark (P)	03	14	50	08	21
Finland (P)	11	38	57	25	31
Norway (P)	09	28	53	16	22
Sweden (P)	08	20	34	13	27
Switzerland (M)	16	na	na	47	36

Source: Inglehart et al. 1998.
a RC = Roman Catholic; P = Protestant; M = Mixed Protestant and Roman Catholic.

belief in reincarnation. In England and Wales approximately one-third of young teenagers believe in their horoscopes, ghosts, and the possibility of contacting the dead (Francis and Kay 1995, p. 152).

It is also important to recognize that while established mainline church membership is declining, there is some growth among smaller nonestablished churches, especially those which adopt an evangelical and charismatic form of spirituality. In this context the nature of belief and practice tends to be more emotive and personal. For example, young people are more likely to think of God as an intimate friend than as a righteous judge or creator. Various non-Christian new religious movements also attract a minority of youth, although these groups have not counteracted the overall drift away from the Church (Bruce 1996, p. 188).

Young people therefore retain a degree of private belief and practice which could offer an ultimate basis for identity formation. However, Table 2 indicates this is not necessarily the case. While many young people believe in God, God and religion are not very important for a large number of them. Gender studies further reveal that religion is less important to men and boys than it is to women and girls. Women in paid employment, however, tend to be more like men in their religiosity, so as more young women expect to have working careers, this gender difference is likely to decrease in favor of the perceived irrelevance of religion. What then is the significance of religion for young people's identity?

3. *Young People's Religion and Identity*

The data suggest three broad patterns of relationship between religion and identity among young people in modern Western society.

3.1 *Secular Identity*

Some young people reject religion altogether (according to the 1981 European Values Survey, 27 percent of people under 35 claim to be atheists, Harding et al. 1986, p. 48), but a large number have remnants of traditional belief, echos of an early but incomplete socialization into the Christian framework of their culture. These young people tend not to be hostile to Christianity and may even 'believe' to some extent, but generally they are ignorant and indifferent to its teachings. Christian plausibility has been undermined by increasing contact with the competing truth

claims of other faiths and philosophies, and by the perception that science offers more convincing explanations of life than does religion. Consequently, Christianity has little, if any, significance for these young people's self-understanding. For the most part, they orient their lives according to a utilitarian rational world view gaining their identity through material consumption and through immediate intimate relationships with family and friends. Links between religion and self-understanding are only salient on special occasions (such as the celebration of a birth or a marriage), or at times of crisis (such as during illness or death) when an implicit Christian identity is unearthed. Even then the significance of religion is fleeting. Its main function is to provide a culturally acceptable background for the renegotiation of an essentially secular identity which is largely achieved through the gathered family and friends. Committed religious belief or understanding is unnecessary in this context.

3.2 Spiritual Identity

Some young people are more active in their approach to religion and identity. They are aware of a wide range of religious symbols from a number of traditions and approach them as a cultural resource which can be subjectively used in a reflexive process of identity formation. Religious symbols provide a means of self-exploration, a source of empowerment and personal growth, around which a sense of self can be constructed. Without recognizing an external authority, these young people are free to decontextualize, reinterpret, and juxtapose apparently contradictory symbols without any sense of conflict. Religious truth and certainty is judged according to subjective experience and utility—if it feels right, or meets an individual's needs, then it is 'true.' This type of religiosity requires little commitment to any particular tradition and the individual can easily move from one set of beliefs and practices to another.

Religion of this type is heavily bound up with the commercial world, especially as presented in New Age philosophies. It generally appeals to older youth in their late twenties rather than to teenagers who cannot afford the more expensive esoteric practices on offer. Where younger age groups do take an interest in these nontraditional forms of religion, it often overlaps with entertainment, is not taken seriously, and competes with other leisure pursuits. As such it does not form an ultimate basis of identity.

3.3 Religious Identity

The remaining young people choose to belong to a particular religious group or church and find their identity within a community of believers. This may be accomplished through a conversion experience to a new religious tradition, or it may be a reaffirmation of the religious tradition of childhood (sometimes this is conflated with broader issues of ethnic identity by way of 'cultural transition' or 'cultural defence', see Bruce 1996). Either way, the emphasis is on a conscious, autonomous decision to be a part of the group. Consequently, religious identity tends to be strong and salient.

In this respect, the churches which attract young people are often evangelical and charismatic, and are well suited to fostering religious identity. The conservative theology, strong social networks, and high degree of commitment these churches usually require encourages a dominant sense of group belonging around religious themes. A charismatic orientation to the group adds a further dimension in terms of religious experience which helps to authenticate the young person's beliefs, and allows for self-expression and a sense of intimacy with God and other believers. In this context a strong sense of religious identity is established.

4 Summary

In conclusion, over the last 30 years established institutional Christianity has generally become less influential in the formation of young people's identity. However, the rate of Church decline is beginning to slow down and religious ideas have not disappeared from society. Consequently, Christianity, independently or as part of a heterodox collection of beliefs and practices, does continue to inform youth identity to a certain extent. It remains to be seen whether the latter two patterns will overturn the general trend towards a secular identity among Western youth in the future.

See also: Adolescent Vulnerability and Psychological Interventions; Education: Cultural and Religious Concepts; Generations, Sociology of; Religion: Family and Kinship; Religion: Morality and Social Control; Religion, Sociology of; Religiosity: Modern; Socialization in Adolescence; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements

Bibliography

Ashford S, Timms N 1992 *What Europe Thinks: A Study of Western European Values*. Dartmouth Ltd, Aldershot, UK

Brierley P 1997 *UK Christian Handbook Religious Trends No. 1*, 1998/99 edn. Christian Research, London

Bruce S 1996 *Religion in the Modern World: From Cathedral to Cults*. Oxford University Press, Oxford, UK

Clark W, Carroll J W, Roozen D A (eds.) 1995 *The Post-War Generation & Establishment Religion: Cross-Cultural Perspectives*. Westview Press, Boulder, CO
Davie G 1998 God and Caesar: religion in a rapidly changing Europe. In: Bailey J (ed.) *Social Europe*, 2nd edn. Longman, London, pp. 231–53
Francis L J, Kay W K 1995 *Teenage Religion and Values*. Gracewing, Leominster, UK
Giddens A 1991 *Modernity and Self-Identity*. Polity Press, Cambridge, UK
Halman L, de Moor R 1993 Religion, churches and moral values. In: Ester P, Halman L, de Moor R (eds.) *The Individualizing Society: Value Change in Europe and North America*. Tilburg University Press, Tilburg, The Netherlands pp. 37–66
Harding S, Phillips D, Fogarty M 1986 *Contrasting Values in Western Europe: Unity, Diversity and Change*. Macmillan Press, Basingstoke, UK
Inglehart R, Basañex M, Moreno A 1998 *Human Values and Beliefs: A Cross-Cultural Sourcebook*. University of Michigan Press, Ann Arbor, MI

S. Collins

Religion: Culture Contact

1. A Model of Encounter

In many studies of culture contact there is a perceived dichotomy between the acceptance and the rejection of Christianity. This model of religious change, often advanced by those who have a stake in defending religious boundaries, stems from false assumptions about the nature of the religious encounter. Its premise is of an essential or textual Christianity confronting an essential or ideal traditional (or folk) religion—the religion of the invader confronting the religion of the autochthon. But, Christianity has never been propagated as a monolithic entity encountering a homogenous local society. A more appropriate model involves a number of modes interaction. First, there was the encounter of a variety of mission Christianities, often mediated by local agents, with a traditional religion which was open, dynamic, and plural in nature. And, this religious encounter cannot be divorced from its political and socioeconomic context. It needs to be situated within a second set of conjunctions between the forces of capitalist imperialism and local institutions. Moreover, those encountering Christianity never adopted it wholesale. Specific ethnic groups, classes, and social categories appropriated particular ideas and practices and new forms of popular or vernacular religion, both traditional and Christian emerged out of processes of gleaning, borrowing, and recoding. This article considers these various modes of interaction with specific reference to Africa but with a Latin American aside.

2. Scholarly and Popular Misconceptions about the Encounter of Religions

2.1 Syncretism

In order to arrive at the above model of religious encounter it is first necessary to deal with a number of scholarly and popular misconceptions about the process. First, there is the issue of syncretism. This is a contentious term, often used to denote inauthenticity, or the pollution of a notionally pure tradition by symbols and meanings taken from alien traditions (Shaw and Stewart 1994). In African studies, missiologists and theologians have used syncretism as 'a term of disapprobation denoting the confused mixing of religions' to explain the rise of movements of Christian Independency, such as Zionism in Southern Africa and Aladura in West Africa. In Latin American studies, syncretism has far less of a pejorative connotation. Here scholarship on culture contact has been coloured by the optimism born in the heyday of the American melting pot ideology and portrayed African derived cults such as Vodun or Santera as a celebration of indigenous agency in the face of colonialism and other forms of alien domination (Shaw and Stewart 1994; see also Rostas and Droogers 1993). Recently, anthropologists and social historians have discredited the idea of essential, untainted traditions, exploring instead the processes by which they have been invented or imagined. Where the study of syncretism has most value is when it is understood as a form of cultural politics in which religious practitioners engage in heated polemic about religious hybridity while seeking to claim the moral highground of authenticity.

2.2 African Traditional/Folk Religion

A second set of misconceptions surround the nature of African traditional religion which has been portrayed consistently as ahistorical in contrast to the dynamic religions of the book, Christianity and Islam. If missionaries prior to the twentieth century admitted Africans had a religion at all, most, but not all, saw it as no more than primitive superstition blocking the advance of Western civilization. While early twentieth century anthropologists such as Edward Evans-Pritchard and Meyer Fortes were better informed about African traditional religion they viewed it no less ahistorically. Critical of missionaries whom they argued were destroying ancient and intricate mechanisms of social cohesion, they wrote detailed studies on the function of traditional religion in maintaining the organic stability of African societies. Some of the next generation of missionaries such as Bruno Gutmann or Mabel Shaw (see below), inspired by this functionalist anthropology came to appreciate the supposed communality of African traditional religion and sought to engineer these values in an 'adapted' Christianity.

Historical models of African religion also found their way into scholarship beyond social anthropology. Scholars of comparative religion sought to respect all religious traditions and make points of connection between them. Thus, eminent comparative religionists described African traditional religion completely in experiential terms (Parrinder 1954). Leading African theologians have also asserted an historical model of African religion. Mbiti has contended that African religion has no future tense and possesses no founders, prophets, or converts. Africans shared the same underlying philosophy but each 'tribe' had its own religion into which its members were born. While celebrating this integration of religion and society, Mbiti argues that the boundedness of African traditional religion means that it is incapable of adapting to the increasing scale of interaction that colonialism has engendered, and will inevitably decline (Mbiti 1969).

The rise of African history as a discipline from the late 1950s onwards confronted ahistorical notions of African traditional religion. Particularly important has been the emergence of oral history. Some African religions such as the Mbona cult in southern Malawi have shrines maintained by priesthoods who retain extensive traditions referring back to founding figures and cycles of revival, cleansing and prophecy (Schoffeleers 1992). Moreover, the notion of the boundedness of African religious and social identity has been challenged by research on ethnicity. Many of the 'tribal' identities claimed by contemporary Africans are in fact colonial and postcolonial creations. Hence, the concept precolonial Kikuyu or Shona religion makes no historical sense.

African historians have now replaced the old model African religions as a total organic collectivity with a more dynamic and complex one. African cults symbolized and articulated social and economic relationships and thus their adherents were involved in a range of religious interactions that began at the level of family and faction but could extend across their polity and beyond. Guilds of men propitiating hunting spirits united disparate groups of hunters throughout a region. Cults of the land, sometimes operating across a number of polities, enforced ecological rules concerning clearing, planting and reaping. And cults of alien or a-social spirits established the rules of communication between traders and the communities with whom they interacted (Ranger 1987). This more open and pluralist model of African traditional religion helps explain its resilience and adaptivity through the twentieth century.

2.3 Missionaries and Local Agents in Africa (and Latin America)

Neither was African traditional religion 'crushed' by a destructive monolithic mission Christianity acting as the handmaid of colonialism. Such a wholly negative view of missionaries not only overlooks the contribution of local agents to the foundation of the African church but it also ignores the variety of mission Christianities and complex relationships they had with colonial states and African societies. Missionaries were influenced by different theologies and they were as diverse as the historical contexts and social sources from which their movements arose.

The contrast between the earliest Catholic mission enterprises and those of the first nonconformist missionaries could not be greater. The first wave of Catholic missionaries to sixteenth century Congo were controlled and sponsored by the Portuguese state. Mission took the form of a single religious enterprise conceived of in terms of territory. Co-opting state machinery in a union of faith and politics, Portuguese Catholics sought to create Christian kingdoms through the conversion of elite (Gray 1990). The early nineteenth century nonconformist missions were an extension of the culture of voluntary association and private enterprise that underpinned industrialization. Nonconformists like the London Missionary Society (LMS) in southern Africa were not especially interested in converting kings and aristocrats but rather turning ordinary Africans into sober and industrious Christians (Hastings 1994).

Missionaries also differed on the meaning of conversion. High Church Anglicans from the UMCA, German Lutherans, and Catholics believed that salvation was received through the sacraments and the liturgy, and sought to Christianize entire communities. By contrast, those Protestants influenced by revivalism placed a greater emphasis on salvation acquired in an intense and personal way. Their emphasis on moral conduct and lifestyle led to greater separation of the convert from their former community. While many missionaries were ethnocentric and culturally arrogant others did make a sincere and concerted attempt to come to terms with African culture and religion. As early as 1855 Bishop Colenso of Natal wrote a cogent argument for toleration of polygamy in his *Remarks on the Proper Treatment of Polygamy*. By the twentieth century a whole spectrum of missionary responses to their local settings had emerged. There were imaginative attempts at the translation of Christian ideas into an African idiom. Mabel Shaw of the LMS made extensive use of image of God/Jesus as 'Chief' and the idea of the Christian community as a 'tribe' in her work amongst the Bemba. Another approach was the Christianization of traditional rituals so that they might find force in a Christian community. The most common example of this was the adaptation of traditional rituals of initiation. Here the endeavors to Christianize the traditional rite of passage by UMCA Bishop to Tanzania, Vincent Lucas of Msasi, have been well documented (Ranger 1972).

Indeed, while many missionaries to Africa have been criticized for their cultural imperialism, others

may be taken to task for their conservatism. The German missionary and anthropologist, Bruno Gutmann, believed that the growth of African personality and morality was connected to the essential forms of community life: clan, neighborhood, and age grade. Thus, he developed an 'indirect rule' approach to mission, highly attractive to the Tanzanian colonial state, which sought to create Christian communities through traditional institutions and leadership (Kaplan 1995). When the radical black American educationalist W. E. B. du Bois visited West Africa in 1923, he criticized the fashionable 'emphasis on the preservation of traditional culture.' This du Bois thought, formed part of a white man's plot to maintain the African as a willing but useful servant of the imperial cause.

Perhaps unexpectedly, the last great new wave of mission Christianity, following earlier Catholic and nonconformist waves, has been the most adversarial towards traditional culture and religion. This is the Pentecostal movement that arose in America and northern Europe, 1900–10, and collided with Africa and Latin America almost immediately after its inception, and in subsequent surges throughout this century (Maxwell 1999a, Freston 1995). In part a reaction against the liberalism and modernism of established church theology, Pentecostalism actively believed in the supernatural. But while taking African and Latin American belief systems seriously Pentecostalism nevertheless demonized them (Maxwell 1999b, Lehmann 1996). Traditional spirits, whether good or bad in local cosmological terms, were collapsed into the category of the demonic. Thus, Pentecostals have been characterized by their aggressive stance towards traditional practices, and the destruction of charms and fetishes as polluted objects. Yet, even this Pentecostal attack on things 'traditional' can be interpreted as a form of Africanization: a perpetuation of a tradition of cyclical societal cleansing.

Finally, those scholars who continue to insist on the alien nature of African Christianity fail to grasp the role of local agents in the Christianizing of the continent. Pioneer missionaries were responsible for vast tracts of territory. Few and far between, their priorities were first translation work and then the construction of mission infrastructure. Many simply struggled to survive disease and learn the local language. The work of conversion was left to local agents, freed slaves, and converts from already missionized areas. Because overstretched missionaries were unable to monitor their activities these Catholic catechists and Protestant evangelists had much room for creativity. There were also movements of Christians happening completely independently of missionaries such as the Christian refugees who fled the expansion of the Zulu Kingdom in the 1830s and carried Christianity with them into Lake Malawi region. Another example was labor gangs, such as groups of Pedi in the Transvaal, South Africa in the 1840s–1860s, who encountered Christianity in their work place and spread it through the process of labor migration. Often missionaries would arrive in a village to find Africans already there teaching the faith.

3. *A Latin American Comparison*

Latin America did encounter a monolithic mission Christianity. For over three hundred years the subjects of Spanish and Portuguese empires were exposed to state-sponsored Catholicism. Yet, from the very beginning of these Iberian empires there was a debate within Catholicism about the nature of traditional belief. While Spanish imperialists exterminated Indian peoples on the pretext that they were idolaters or devil worshippers Bartholomé de Las Casas described their pagan practices as their own way of approaching the Christian God. Las Casas, a sixteenth century Dominican, also argued for a clearer separation of the church from the imperial project, while the Franciscans and later the Jesuits established local regimes which while patriarchal did offer more humane conditions than their secular coevals.

Latin American Catholicism could also embody regional and national identities. The cult of the dark faced Virgin of Guadalupe, which dates from the mid-sixteenth century, offered a range of protective and curative functions. The Virgin appeared on the banners of those who sought independence from Spain in the early nineteenth century (led by village priests) and was subsequently made universal patron for the whole of Mexico (Brading 1991). Official Catholicism, initially opposed to nationalist movements, took a generation to come to an accommodation with the new nations before becoming embroiled in a century of warfare between conservatives and anticlerical liberals or radicals.

The strength of the sacred canopy of Catholicism varied across the continent. Some of the Amazonian peoples remained 'unreached' until the arrival of twentieth-century Catholic and Protestant missionaries. In other instances the Catholic overlay remained thin. The Maya of Guatemala, who resisted for centuries, retain numerous traditional beliefs, praying to the ancestors as well as the saints, visiting the shaman alongside the priest. In Brazil, Catholicism is often cross-referenced with belief in African Gods imported along with African slaves. In contrast to the literature on African religion, scholarship on Latin America more often equates 'folk' or 'traditional' religion, not with non-Christian religion, but with a 'syncretistic' mix of Catholicism and surrounding religions.

The degree to which Latin American Catholicism drew from other religions rather than retaining a more distinctive identity is explained by combination of two

factors: first, its close relation to the institutions of the state; second, the absence of alternative and competing Christianities. Catholicism was the established religion. Even at local level the church was often a major landowner and the priest involved in local politics. Catholic ideology legitimated social relations of production on the *hacienda*. The local level Catholic representative was usually not a catechist but a foreign priest. In the absence of other Christianities and a strongly independent Catholic dynamic Latin Americans were far more selective in their response to Catholicism and that selectivity came to be accepted by the authorities. Thus, for instance, Ecuadorian Quichua used the cult of the Saints and different visions of the virgin to reach those extrahuman forces they believed dominated them. Prayers, ritual drinking, dancing, and the wearing of lavish costumes took place to influence God's intermediaries. The same process of selection, of course, occurred within African Christianity, but in an atmosphere of Christian pluralism, African Christian identities took on a sharper definition.

Latin American Catholicism, in its different forms, remained monopolistic until the late nineteenth century when liberal elites opened doors to north Atlantic Protestantism and migration. The first wave of missionaries were predominantly older Protestant denominations such as Methodists and Baptists, offering welfare and education, but throughout the twentieth century Pentecostal and conservative evangelical waves came to dominate. The opening up of the religious field was accompanied by the break-up of Catholic sociopolitical hegemony. Agrarian reforms, anticlerical regimes, the rise of the mega-city and the partial liberalization of Catholicism made alternative religious adherence possible.

The persistence of pre-Christian cults is supplemented by their re-invention by middle class intellectuals in search of an authentic religion. One such group is the Association of Concheros in Central Mexico who have fused Catholicism with a blend of notional religious traditions from peoples such as the Aztecs and Chichemecas (Rostas 1993). In the African context such invented religious traditions are not so much sources of spirituality as markers for ethnicities.

Doubtless, if Spain and Portugal had retained and expanded their mercantile empires in Africa then African Catholicism may well have more closely resembled the early Latin American type. But Spain and, to a lesser extent Portugal, did not figure highly in the 'Scramble(s)' and the accompanying Catholic impulse was generally of a non-Iberian variety, involving a plurality of Catholic missionary orders working in competition with Protestantism. Given the open and plural nature of African traditional religion, the varieties of mission Christianity and the contribution of local agents, the African encounter took a profoundly different form from the Latin American one.

4. *Resilience and Innovation in African Traditional/Folk Religion*

Throughout the nineteenth century traditional religion innovated in response to the array of changes that accompanied the slow penetration of first, mercantile, and later capitalist, relations into the African interior. The increasing scale of social interactions that these changes brought led local intellectuals to innovate to make sense of this new world breaking in. In some cases they developed the idea of the High God with a macrocosmic reach to match Africans' expanding horizons. In other contexts, cults of possession became more widespread linking people together over wide regions through their shared affliction by the same amoral spirit. In many places the colonial occupations of the 1870s–1890s coincided with and precipitated ecological crises. In what appeared to many as a millennial moment, great prophetic figures, such as the Shona Chaminuka medium, arose to interpret colonialism and capitalist imperialism in a way that offered their followers conceptual control over what was happening. Thus, the great transformations were not an alien imposition but part of an unfolding local narrative.

In a similar manner, the new religion of Christianity was understood in terms of the old categories of traditional religion. Some missionaries were seen as witchcraft eradicators, whose presence inaugurated movements destroying charms and fetishes (Maxwell 1999b). Other missionary pioneers, with their bags of simple medical cures were seen as traditional healers (Hastings 1994). Others still were welcomed as sources of revelation about the High God.

While specific forms of traditional religion, such as hunting guilds, declined under colonialism, other forms of traditional religion moved to the center of the stage. African religious leaders, such as the famous Kinjikitile in Tanganyika, offered their cults as means of uniting acephalous societies against German occupation. And during the era of decolonization, religious leaders once again helped coordinate resistance against colonial states. In Zimbabwe's war of liberation Shona spirit mediums moved popular understanding of myth and taboo to enable stranger/guerrillas to be legitimated in local terms as hunters and warriors (Lan 1985). Although the labor migration led to decline of some cults of the land, in other instances traditional religion became more ecologically orientated helping Africans to become peasants and articulating their objection to state intervention in their agriculture.

5. *The Foundation of Vernacular Christianities in Africa and Latin America*

Nevertheless, Christianity has made tremendous headway in Africa and Latin America. In Africa great movements of Christianity in Buganda, the Ivory

Coast, Igboland, and Manicaland occurred during the Scramble(s) and in the first decades of the twentieth century. But these grassroots initiatives often happened in a haphazard and unplanned manner far from missionary influence or control, and through this indigenous agency local or vernacular Christianities were founded.

Besides the significant contribution of African agents, vernacular Christianities had two other key defining features. The first was the seizing of the landscape by symbolically sensitive missionaries and African Christians. Mountains, rocks, caves, trees, and pools, which often had traditional religious associations, were resacralized in Christian fashion. Catholics turned them into shrines and grottos—places of pilgrimage. The Anglicans made holy ground through creation of cemeteries.

The second key feature of these vernacular Christianities was the adoption of Christian symbols and powers and Christian literacy because they had utility in helping Africans prosper and relate to the colonial economy. While traditional religion could become more ecological in orientation to help Africans become peasants, Christianity could equally fulfill that role. Catholic prayers, invocations, and rituals were perceived as efficacious in protecting crops and providing soil fertility. American Methodism with its strong emphasis on discipline, an intense personal experience, and gospel of the plough provided the perfect ideology for peasant entrepreneurs (Ranger 1987).

Christian approaches to healing were also important. Pentecostal and revivalist emphasis on the satanic causes of illness resonated with traditional personifications of evil and disease. Likewise, divine healing through the laying-on of hands resembled the practices of traditional healers, though, of course, Pentecostals claimed a different source of empowerment. Miraculous cures through veneration of the saints provided a Catholic counterpart to Pentecostal spiritual healing. St Thérèse of Lisieux was perceived as particularly powerful in Africa. Seminaries of African priests were placed under her protection and she was appointed 'regent' of at least one African country. Finally, African Christian movements were often animated by a grass roots adoption of literacy. For these Africans, literacy, schooling, and church were synonymous in representing progress (Ranger 1987).

Specific social categories were often responsible for the localization of Christianity. In the nineteenth century Africa, kings and chiefs often converted to extend their ideological legitimation and to benefit from the technological and diplomatic advice missionaries had to offer. But these traditional leaders were slow to change their ideas and practices. In both continents younger men and women often converted and changed the very culture these male gerontocrats wished to preserve. The conventional language of Christians: 'coming forward,' 'setting themselves apart,' and declaring for a 'completely changed life' took on a new meaning when merged with Christianity's individualist conscience (Fields 1985). It offered younger men, returning from the workplace, legitimate reasons for ignoring the burdensome demands of traditional commensality, the Latin American *fiesta* (Earle 1992), or African fertility and harvest rituals. Likewise, it enabled more individualist peasants to separate themselves from the 'wasteful' and time consuming demands of their kin, and in the Latin American context, fictive kin: God children (Goldin and Metz 1991). Faith in Christ gave women recourse to another male authority other than that of the household head, and directly undermined elder male notions of patriarchy by demonizing the ancestor spirits which legitimated it. Likewise, the prayer band and the fellowship meeting offered a vital source of solidarity for women otherwise divided by exogamy, polygamy, and fear of witchcraft accusation, and an engine of resocialization for the wayward male. Pentecostalism offered its own unique source of empowerment. Gifts of prophecy, tongues, healing, and exorcism turned a marginal woman or man into someone of consequence (Gill 1990). Catholicism was frequently an intensely feminine religion. Its Marian emphasis offered women a double-edged model of liberation and resignation.

Material benefits also accompanied conversion. The church offered networks of contacts and solidarity to labor migrants entering the cities, mines, or farms, and it created new alliances of Christian kin within the village. Although Pentecostalism's doctrine of male headship legitimated gender inequalities, its emphasis on industry, sobriety, and sexual propriety diminished aspects of male sex roles that are harmful to women (Gill 1990). Throughout the townships, shantytowns, and villages of Africa and Latin America the puritan ethics advanced by Pentecostals, other nonconformists, and groups like the Watch Tower have pulled families and small communities out of cycles of poverty, or at least stopped them from slipping over the edge.

Where Christianity has not eroded African traditional religion, it has nevertheless often changed it. Christian ideas and practices have found their way into non-Christian religions. Contemporary African genealogies tracing the origins of significant territorial spirits may include references to Hebrew patriarchs such as Noah or Abraham as useful embellishments (Bourdillon 1988). And traditional concepts of God have likewise been sharpened by Judaeo–Christian ideas about the divine.

None of the above should for one moment suggest that missionaries have not been important. It is rather that their importance came in ways they did not expect. Missionary teachings on domesticity shifted gender relations and created new social classes out of missionized elites. And work on languages such as the creation of vernacular alphabets and grammars has

had an immense impact. Not only has it created written languages where none previously existed, but these in turn have helped create new ethnic identities in both continents. In Latin America, the influence of one organization in particular, the Wycliffe Bible Translators, cannot be underestimated. In Africa, literacy and schooling enabled Africans to form their own cultural movements and provided access to political ideas other than those peddled by colonialists.

6. *Christianity in Contemporary Africa and Latin America*

The process of culture contact of course continues. Despite the fact that the center of Christendom no longer lies in the post-Christian West but in the South, the Western impulse to mission has not abated. It has, however, changed in nature. First, the majority of contemporary missionaries are North American in origin, reflecting the rise of the US as the new imperial power. Second, these missionaries are usually evangelical, charismatic, or Pentecostal and believe that conversion entails an individualized born-again experience. Finally, this new strand of born-again Christianity has a strong affinity with the electronic media.

The arrival of this new type of conservative Protestantism, which exploded across Latin America in the 1960s, and across Africa a decade and a half later, raises the inevitable question of whether it is just another American cultural export (Gifford 1998). While America is the major cultural and missionary dynamo for the born-again movement there are other epicenters in Southern and West Africa and Brazil, each contributing their own 'prophets' and media packages to increase the heterogeneity of the movement. But more importantly, the born-again movement has always been conducive to indigenization. Its lack of tradition, its scorn of formal theological education, its reliance on lay initiative and management, its family and neighborhood cell structure, and its acceptance of the supernatural make it particularly responsive to local agendas. While its stress on releasing converts from the demands of kin, both alive and dead, and its emphasis on building nuclear families may smack of western individualism, these tendencies are useful to African (and Latin American) peasants facing the decline or break-up of the communal system and migrant workers coming to terms with the rise of the mega-city (Martin 1990).

Finally, culture flows do not just move from the West to the South, they are multidirectional. While African and Latin American diaspora have for centuries transported their religions to the West, contemporary Latin American and African Christianities impact on Europe and America in ever more profound ways. In the 1970s and 1980s Latin American liberation theologians shook up Western theological establishments, while the spirituality of Latin American base communities inspired many Westerners working for justice and peace. Today, mission partnerships bring African Anglicans to British parishes and returning missionaries import Christian song and liturgy from the South into their home churches. Added to this is the growing appeal of African Pentecostal and charismatic leaders in Western preaching and healing circuits. Some of these leaders' churches also finance their own mission initiatives. The arrival of African missionaries in London's Oxford Street, or Brazilians in Portugal, brings mission Christianity full circle.

See also: African Studies: Religion; Christianity Origins: Primitive and 'Western' History; Colonialism, Anthropology of; Cultural Expression and Action; Education: Cultural and Religious Concepts; Globalization, Anthropology of; Religion and Economic Life; Religion: Definition and Explanation; Religion: Evolution and Development; Religion, History of; Religion: Morality and Social Control; Religion: Nationalism and Identity; Religion: Peace, War, and Violence; Religion, Sociology of; Religious Nationalism and the Secular State: Cultural Concerns

Bibliography

Bourdillon 1988 Gleaning: Shona selections from Biblical myth. In: James W, Johnson D H (eds.) *Vernacular Christianity. Essays in the Social Anthropology of Religion*. Lilian Barber Press, New York

Brading D A 1991 *The First America. The Spanish Monarchy, Creole Patriots, and the Liberal State, 1492–1867*. Cambridge University Press, Cambridge, UK

Earle D 1992 Authority, social conflict and the rise of protestantism: religious conversion in a Mayan village. *Social Compass* **39**: 377–88

Fields K E 1985 *Revival and Rebellion in Colonial Central Africa*. Princeton University, Princeton, NJ

Freston P 1995 Pentecostalism in Brazil: A brief history. *Religion* **25**: 119

Gifford P 1998 *African Christianity: Its Public Role*. Indiana University Press, Bloomington, IN

Gill L 1990 Like a veil to cover them: Women and Pentecostal movements in Le Paz. *American Ethnologist* **17**: 708

Goldin L R, Metz B 1991 An expression of cultural change: Invisible converts to Protestantism among Highland Guatemala Mayas. *Ethnology* **30**: 325–38

Gray R 1990 *Black Christians and White Missionaries*. Yale University Press, New Haven, CT

Hastings A 1994 *The Church in Africa 1450–1950*. Clarendon Press, Oxford, UK

Kaplan S 1995 The Africanization of missionary Christianity: history and typology. In: Kaplan S (ed.) *Indigenous Responses to Western Christianity*. New York University Press, New York

Lan D 1985 *Guns and Rain. Guerrillas and Spirit Mediums in Zimbabwe*. James Currey, London
Lehmann D 1996 *Struggle for the Spirit: Religious Transformation and Popular Culture in Brazil and Latin America*. Polity Press, Oxford, UK
Martin D 1990 *Tongues of Fire. The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK
Maxwell D 1999a Historicising Christian independency: The Southern African Pentecostal movement. *Journal of African History* **40**: 2
Maxwell D 1999b *Christians and Chiefs in Zimbabwe: A Social History of the Hwesa People c. 1870s–1990s*. Praeger, Westport, CT
Mbiti J S 1969 *African Religions and Philosophy*. Heinemann, London
Parrinder E G 1954 *African Traditional Religion*. Hutchinson's University Library, London
Ranger T O 1972 Missionary adaption of African religious institutions. In: Ranger T O, Kimambo I N (eds.) *The Historical Study of African Religion*. Heinemann, London
Ranger T O 1987 Concluding summary. In: Peterson K H (ed.) *Religion, Development and African Identity*. Scandinavian Institute of African Studies, Uppsala, Sweden
Rostas S 1993 The Mexica's reformulation of the Conchero's dance: the popular use of autochthonous religion in Mexico City. In: Rostas S, Droogers A (eds.) *The Popular Use of Popular Religion in Latin America*. CEDLA, Amsterdam
Rostas S, Droogers A 1993 Introduction. In: Rostas S, Droogers A (eds.) *The Popular Use of Popular Religion in Latin America*. CEDLA, Amsterdam
Schoffeleers J M 1992 *River of Blood. The Genesis of A Martyr Cult in Southern Malawi, c. AD 1600*. The University of Wisconsin Press, Madison, WI
Shaw R, Stewart C 1994 Introduction. In: Shaw R, Stewart C (eds.) *Syncretism/Anti-Syncretism: The Politics of Religious Synthesis*. Routledge, London

D. Maxwell

Religion: Definition and Explanation

This article sets out the main forms explanations of religion take by reference to their scope and the extent to which they preserve or subvert the self-understanding of religious believers. At the maximum level of scope explanations of religion amount to comprehensive theories of religion. Typically, such theories also seek to subvert the way in which religious individuals understand their own beliefs and actions. The need for comprehensive theories of religion is shown to arise out of the methodological demands of social and behavioral science itself. Theories of religion need to demarcate their subject matter and hence depend on and give rise to definitions of religion. The article sets out the differences between the main styles of definition of religion (substantive, functionalist, experiential, and family resemblance) and explores the possibilities of finding common ground between them.

1. Explaining Religion

1.1 The Nature of Explanation

Explanations are answers to requests for understanding. To explain something may be to: (a) make it known in detail; (b) render it intelligible; or (c) account for it. Disciplines such as the anthropology of religion may explain a religious phenomenon by citing further facts about it, thus explaining it in sense (a). Or they may explain a religious phenomenon by setting out its connections with other, more familiar or easily understood social facts, thus explaining it in sense (b). These forms of explanation need not overturn, but may merely supplement, the motives and reasons the religious themselves acknowledge as behind their beliefs and behavior. However, the social and behavioral scientist may seek to account for a religious phenomenon (type (c) explanation) by giving its cause and in the citing of that cause may set aside the self-understanding of religious actors. For example, the anthropologist may seek to account for ecstatic religious experiences in terms of the social status, particularly social power, such experiences give to their possessors (Lewis 1989). If this factor is a sufficient explanation, then the agents' belief that they are genuinely possessed by gods or spirits is an illusion.

1.2 Radical Explanations in the Science of Religion

It is characteristic of the social sciences in the nineteenth and twentieth centuries that they have gone beyond seeking to explain particular religious phenomena. They have produced explanatory theories seeking to account for the existence of religion as such. Such theories hold that the tendency of human beings to engage in religion cannot be taken as a given; there is something problematic in such engagement. For example, an evolutionary theorist may see religious phenomena as 'survivals,' that is dysfunctional judged against the circumstances of modern society and thus as a remnant of a once adaptive means of organizing social life (Tylor 1903). A psychological theorist may see a striking correlation between religious symbolism and psychic needs, suggesting that religion is motivated by persistent and universal wish-fulfillment (Freud 1927/1978).

The explanations of religion thus produced are radical, comprehensive, and external. They embody the 'hermeneutics of suspicion' (Ricoeur 1970). Religious consciousness is treated as a form of false consciousness. Religious symbolism is a code. Ostensibly it speaks of gods and spirits. In reality its components refer to aspects of the human, social world. A psychological, sociological or anthropological theory of religion enables us to reinterpret and decode religious symbols by correlating them with

mundane forces and facts hidden from the conscious mind of the religious actor.

1.3 Methodological Skepticism

The assumption that religious thought and action cannot be taken at face value but must be explained from the ground up is in part a product of the philosophical skepticism about the gods and the limits of human knowledge engendered by the Enlightenment. Some argue that such skepticism, or naturalism, is required by any sound methodology for the human sciences (Preus 1987). The only explanatory factors the human sciences can know of and thus recognize are mundane ones. They can only produce explanations of religion that deal in natural causes. Hence, they must adopt an attitude of suspicion toward religion's own 'memory' of humankind's contact with supramundane realities, and reinterpret religious symbols accordingly.

1.4 The Cultural–Symbolic Paradigm

The arguments for methodological skepticism are crucial in supporting radical, comprehensive theories of religion in the social and behavioral sciences. Without methodological skepticism, the production of such theories is not mandatory for those unconvinced of the Enlightenment's case against religion. The case against methodological skepticism can be made from within a contrasting approach to the explanation of human, social phenomena: the cultural–symbolic (Byrne 1997). This has its ancestry in the stress on *Verstehen* as the appropriate and unique stance to take toward the human world advocated by German social theorists in the idealist tradition, such as Dilthey. On a cultural–symbolic approach, there is something that marks out the human world as different from the natural. Its objects—human beliefs, actions and institutions—are endowed with meaning independent of any sense social scientists impute to them by their theories. The meaning they have is a result of the fact that they are informed by concepts. A human act, for example, has the sense it has by virtue of it being done for a reason and being directed to an end. That sense is dependent on agent's own beliefs and concepts. A standard way of explaining a human act is to bring out the reason/end the agent had in performing it. Such an explanation will fail unless it uncovers the agent's reason/end. Human action is thus informed by concepts. We explain human acts by bringing out their background in the agent's beliefs and intentions and by making those beliefs and intentions intelligible by reference to the rule-governed forms of social life in which they are embedded. Crudely put: to explain religious phenomena in this way is to account for them in terms of how the world appears to religious actors themselves and in terms of the forms of individual and social meaning these actors create in the institutions which they fashion (Winch 1958).

The cultural–symbolic approach preaches methodological agnosticism. Religious actors' beliefs relate to suprahuman and supramundane realities. These beliefs need neither be endorsed nor denied by the social scientist in bringing out their sense or relying on their explanatory power. Both their sense and explanatory power relies on the beliefs making action and thought intelligible in terms of how the world appears to these agents. Whether the world is as it appears to these agents is irrelevant. By the same token, the question of whether the referring expressions used in reporting religious actors' beliefs really do refer is irrelevant. It is enough that actors intend to refer to gods and like things by these expressions. Since methodological agnosticism in the explanation of religion is not committed to the truth of actors' beliefs or to the real existence of things corresponding to the referring expressions in reports of those beliefs, it is a thoroughly naturalistic approach to the interpretation of religion (Clarke and Byrne 1993).

The cultural–symbolic approach can admit that there are forms of religious life which function ideologically or are the product of illusion. Investigation may show that individual religious phenomena exhibit forms of human meaning-making which fail of a minimal rationality and are corrupt. In such cases the hermeneutics of suspicion may be appropriately applied (MacIntyre 1971).

2. Defining Religion

2.1 The Purposes of Definitions of Religion

The social scientific study of religion needs a definition of religion in order to demarcate the subject matter of its inquiries. Theories of religion require definitions of religion at this initial stage for a parallel reason. Definitions may also serve the purposes of theories of religion by suggesting an initial interpretation of religion on which the theory may build. A complete theory of religion will also give rise to a definition of religion by way of summing up what it has concluded about the essential nature or origins of religion. Definitions of religion offered at the first stage are operational ones; definitions of religion offered at the end of theorizing can be styled 'theoretical.'

2.2 Types of Definition

Religion may be defined in terms of the typical content of its beliefs. Substantive definitions are illustrated by Tylor's definition of religion as 'belief in spiritual beings' (Tylor 1903). Religion may also be defined by reference to the role religion plays in personal and social life or the structure of religious thought and

action. Such structural–functionalist definitions are illustrated by Durkheim's definition of religion in terms of a symbols system enforcing a distinction between sacred and profane reality and uniting members of society into a common moral community (Durkheim 1912/1967). Experiential definitions demarcate religion by reference to a putative common, or core, experience religious actors participate in, as illustrated by Müller's claim that religion consists in the ability to experience the infinite in the finite (Müller 1893). These three modes of defining religion are united in seeking necessary and sufficient conditions for the classification of an institution as religious. Family resemblance (or polythetic) definitions reject this search. Their advocates maintain that the things we call 'religions' form a loose class. There is no attribute or set of attributes which unites them all: there is merely a network of overlapping similarities (Southwold 1978).

2.3 Issues Dividing Definitions of Religion

Family resemblance definitions are divided from the rest about the degree of unity to be expected among the class of things the Western academy has chosen to call 'religions.' The remaining styles of definition illustrate the conflicting pull of issues of scope and clarity. A substantive definition of religion gives the concept of religion clear boundaries and increases its utility in the demarcation of social phenomena. However, the cost of that clarity is to restrict the scope of the class thus created, threatening thereby to rule out by definition the possibility of nontheistic religion. A structural–functionalist definition gives the concept of religion broad scope, defending the thought that religion can be seen as a universal human phenomenon, but at the cost, its critics will charge, of making the concept too vague.

A means of harmonizing the two dominant approaches (substantive and structural–functionalist may be found in a moral definition of religion. A religion is a symbolic system (with associated beliefs and practices) which articulates the thought that there is a source of moral order behind the world, that is the thought that the realms of value and fact are ultimately united. There are important social and personal needs met by religion so defined and its symbols have a typical content. They must purport to refer to transcendent realities which function as the ground of the natural order and the source or guarantors of morality. These realities are typically, though not inevitably, modeled on human agents (Byrne 1999).

See also: Belief, Anthropology of; Durkheim, Emile (1858–1917); Freud, Sigmund (1856–1939); Functionalism in Sociology; Religion, History of; Religion: Nationalism and Identity; Religion, Phenomenology of; Religion, Psychology of; Religion, Sociology of; Religiosity, Sociology of; Weber, Max (1864–1920)

Bibliography

Byrne P 1997 The study of religion: Neutral? Scientific? Or neither? *Method and Theory in the Study of Religion* **9**: 340–51

Byrne P 1999 The definition of religion: squaring the circle. In: Platovoet J, Molendijk S (eds.) *The Pragmatics of Defining Religion*. Brill, Leiden, The Netherlands

Clarke P B, Byrne P 1993 *Religion Defined and Explained*. St. Martin's Press, New York

Durkheim E 1912 *Les formes élémentaires de la vie religieuse*. Alcan. Paris (1967 *The Elementary Forms of the Religious Life*. Swain J N (trans.). Allen and Unwin, London)

Freud S 1927 *Die Zukunft einer Illusion*. Internationaler Psychoanalytischer Verlag, Leipzig (1978 *The Future of an Illusion*. Robson-Scott W D (trans.). Hogarth Press and the Institute of Psychoanalysis, London)

Lewis I M 1989 *Ecstatic Religion*, 2nd edn. Routledge, London

MacIntyre A 1971 Rationality and the explanation of action. In: MacIntyre A *Against the Self-images of the Age*. Duckworth, London

Müller F M 1893 *An Introduction to the Science of Religion*. Longman, London

Preus J S 1987 *Explaining Religion*. Yale University Press, New Haven, CT

Ricoeur P 1970 *Freud and Philosophy*. Savage D (trans.). Yale University Press, New Haven, CT

Southwold M 1978 Buddhism and the definition of religion. *Man* **13**: 362–79

Spiro M E 1966 Religion: problems of definition and explanation. In: Banton M (ed.) *Anthropological Approaches to the Study of Religion*. Tavistock Press, London

Tylor E B 1903 *Primitive Culture*, 4th edn. Murray, London

Winch P 1958 *The Idea of a Social Science*. Routledge, London

P. Byrne

Religion: Evolution and Development

Both evolution and development are ways of looking at the history of religion. Development implies some kind of directional change in the history of a particular religion, whereas evolution implies comparable directional changes in a number of religious traditions. Since histories usually describe directional change of some sort, the idea of religious development merges into the much larger field of the history of religions. Therefore this article will focus on religious evolution.

1. The Term 'Religious Evolution'

The idea of religious evolution, like the idea of social evolution, has been much criticized in the twentieth century, although neither idea has been entirely abandoned. If the notion of religious evolution is to be salvaged it must be dissociated from several preconceptions of its nineteenth century proponents. These preconceptions include the idea that all religions must be fitted into a single (unilineal) pattern, that progression through a fixed series of stages is inevitable,

and that more advanced stages are 'higher' (ethically and spiritually better) than earlier ones (Comte 1830–1842, Tylor 1871, Spencer 1876–1997 , Marett 1914, etc.). It is the notion of inevitable progress that has been rejected by twentieth century writers; the idea that religious change correlates with social change in comparable ways across traditions can still be defended as 'religious evolution.' Contrary to common belief, it is to the writings of Max Weber (1920–1921, 1921–1922/1978) that a defensible conception of religious evolution owes the most.

2. *The Protestant Ethic*

Weber began his study of religion with his famous *The Protestant Ethic and the Spirit of Caplitalism* (1904–1905/1930). At first glance, it might seem that this book is simply a developmental history focusing on a dramatic change within a single religious tradition. However, it is clear from the place of the Protestant ethic argument in *Economy and Society* (1921–1922/1978) and in the *Gesammelte Aufsätze zur Religionssoziologie* (1920–1921) that Weber believed that ascetic Protestantism was an indispensable catalyst for the emergence of a new form of society, which he called modern capitalism. Never before the Reformation in the West was the hold of agrarian society, organized through patrimonial (including patrimonial bureaucratic) and feudal political forms, broken through. Once established, however, capitalism became a worldwide phenomenon, even though taking different forms in different civilizational areas. Thus Protestantism, though occurring in only one tradition, played, indirectly but crucially, a developmental role cross culturally, and thus can legitimately be considered an instance of religious evolution.

Since Weber's time there has been a search for functional equivalents of Protestantism in other traditions that prepared some societies better than others and some groups in some societies better than others for the transition to capitalism, or, perhaps better, to modernity (Bellah 1957, Eisenstadt 1968). Ernest Gellner made the interesting suggestion that there were elements in all the great traditions, Islam in particular, that might have led to a Protestant-like breakthrough, but that a combination of fortuitous circumstances in Western Europe made it possible for it to occur there (1988). This raises the interesting question of what gave the great religious traditions a rationalizing potentiality even if it was only fully realized in one case.

3. *The Axial Age*

In his comparative studies in the sociology of religion (1920–1921) and in the sociology of religion section of *Economy and Society* (1921–1922/1978) Weber pointed out that in each of the world religions, sometime in the first millennium BC, there emerged prophets or saviors who radically rationalized previous forms of what he tended to call magical religion. In each case the emergent figure (Confucius, the Buddha, the Hebrew prophets, Socrates, Jesus) preached a systematic form of ethical conduct quite different from the diffuse ritual and sacramental practices that preceded them. By calling these new symbolic forms 'rationalized' Weber was pointing to the fact that they were more coherent, more cognitively and ethically universalizing, more potentially self-critical (reflexive), and more disengaged from the existing society than what preceded them.

Weber's friend, Karl Jaspers, spoke of this period as the Axial Age (1948), and S. N. Eisenstadt has contributed significantly to the further study of what he calls Axial civilizations (1986 and elsewhere). Following Weber one can characterize each of the Axial breakthroughs (first millennium BC China, India, Israel, Greece) as involving, to different degrees and in different ways, an element of world rejection (Weber 1920/1946). Religious renouncers (Dumont 1980) in each case called the empirical world into question and, by invoking a transcendental realm, provided a significant degree of tension between worldly interests and religious demands that enhanced the possibility of social change. Although Weber himself contributed to the stereotype of static traditional societies, in his studies of particular traditions he discerned significant possibilities of social transformation (Weber 1920–1921, 1921–1922/1978).

Subsequent scholars have pointed out the dynamic potentialities in all the great traditions. What scholars such as Gellner have emphasized is that a radically rationalized ethic of conduct, such as that found in Protestantism, was not in itself enough. As long as the existing military, political, and religious hierarchies of agrarian societies could contain these rationalizing potentialities within monastic orders or private conventicles, a new type of society could not emerge. What made Protestantism successful was not just its unique ethic but the possibility of gaining sufficient political influence (in Geneva, the Netherlands, England, and New England) to institutionalize that ethic in the major spheres of social life.

4. *The Shape of Religious Evolution*

If Weber had at least two stages of religious evolution, namely an Axial stage correlating with the emergence of historic agrarian societies, and a modern stage for which ascetic Protestantism was an indispensable precondition, did he have any more stages? In looking at his historical sociology as a whole it is clear that he did. The world religions were preceded by societies organized primarily in terms of kinship and neighborhood and in which religion was primarily magical, by

which Weber meant, one in which ritual means were used to attain utilitarian ends such as rain, good harvests, human fertility, etc. While such a three-stage typology is for many purposes adequate, it does seem possible to refine it further.

Bellah, in his essay *Religious Evolution* (1964), suggested a further differentiation of the early stages. He distinguished between a 'primitive,' now better, 'tribal', phase, and an archaic phase. Myth and ritual characterize tribal religion—indeed they may be co-terminous with hominization itself—but they are myth and ritual of a particular type. The spiritual beings whose activities are narrated in the myths are not gods and are not worshipped. These half-human, half-animal, often ancestral, beings express in archetypal fashion the social and ecological life of the people. Ritual involves the reenactment of the primordial events such that mythical time becomes present, not the worship of distanced deities. Tribal religion is most characteristic of hunter-gatherers, but is also common in horticultural societies. With agriculture and the emergence of complex chiefdoms, and particularly with the early state, archaic religion comes into being. While in tribal societies every clan can be traced to an ancestral being, in archaic societies only the aristocratic lineages claim divine descent. A hierarchic pantheon of gods appears, quite different from the mythical beings of tribal religion. The gods require worship, and, significantly, sacrifice, in many cases human sacrifice, and a differentiated class of priests who supervise the worship. In archaic societies the ruler often claims descent from the highest god, even the claim to be a living god himself. In general, archaic religion is characterized by a fusion of political and religious hierarchy largely missing in tribal society. Examples of archaic societies, often identified as Bronze Age societies, would include third and second millennium BC Mesopotamia, Egypt, the Indus Valley, and North China, but also early first millennium AD Japan, and first millennium AD Meso-America and Peru.

It was in the midst of or on the peripheries of such archaic societies that the axial religions of the first millennium BC emerged. They took the hierarchical tendencies of archaic religion in a radically transcendental direction. All value was seen to reside in a realm beyond this world and thus this world and all the worldly hierarchies within it were relativized and potentially delegitimized. But in most cases the new Axial religions either remained marginal to the centers of power or, if they were recognized by the ruling center, they made compromises with it so that they largely legitimated that center rather than challenged it, thus allowing a quasi-archaic situation to continue. This is not to minimize, however, the importance of Axial religions in the elite education and culture of each of the great traditions. Also, the critical potential of the Axial religions was continually reasserted in the form of heresies, uprisings, and rebellions (Eisenstadt 1986, 1999), but these were normally either destroyed, marginalized or, if successful, coopted by the existing structures of power.

A simplified way of looking at the shape of religious evolution was proposed by Bellah (1964). He argued that whereas tribal and archaic religions were primarily this-worldly in orientation, which is what Weber, perhaps unwisely, meant by the word magical, the Axial religions were world-rejecting (and thus, for Weber, very importantly, magic rejecting), although they differed as to whether this rejection was to be worked out within the world (ethically), or as far as possible outside the world (mystically). It should be remembered that in its radical consistency Axial religion was never more than the religion of a minority; the majority continued to entertain beliefs and practices continuous with archaic or even tribal religion, which is what Weber meant by the return to the garden of magic (1920–1921).

With the Protestant Reformation, however, the belief in a radically transcendent God had dramatic this-worldly consequences: the consistent demands of an Axial ethic were to characterize every sphere of daily life. But in the subsequent development of modernity, though the this-worldly dimension remained dominant, its transcendental basis became transformed into immanentism, thus returning the modern world in a much different way to this-worldly immanentism of pre-Axial times. Weber clearly observed this transition, but viewed it almost entirely negatively. The modern world of rationalization would run on its own bureaucratic and economic energies without any transcendental sanction, would become an iron cage (1904–1905/1930). Charles Taylor has described the same transformation in less somber hues as 'the affirmation of ordinary life' and has stressed growing individuation (1989). In any case, though the world of the Protestant Reformation and the Western world since the eighteenth century (and, more recently, much of the rest of the world, as well) are clearly different, it is doubtful whether the Protestant phase needs to be designated as a separate 'early modern stage' as Bellah did in 1964, or whether it should be seen as merely the opening transition toward modern religion, the contours of which, though it has been developing for several centuries, are still far from clear.

The shape of religious evolution should not obscure its many irregularities, its lack of unilineal consistency. Regression was always a possibility: Axial Greek civilization did not emerge from the womb of archaic Minoan and Mycenean societies, but only after the breakdown of those societies and the return to tribalism. Stages were sometimes skipped, especially on the peripheries of advanced societies: the Israelite Axial breakthrough began in a tribal confederation on the boundaries of archaic empires; Muhammad grew up in the tribal city of Mecca but founded an Axial civilization, to be sure under Jewish and Christian

influence. Not only was every religion from tribal to modern full of internal tensions and divisions but all of them were open to continuous external influence.

5. *The Causes of Religious Evolution*

To move beyond description and consider causation is always difficult in historical sociology; nevertheless, how one answers the question of causation, even tentatively, will determine how one understands the entire process. Religious evolution is certainly embedded in social evolution and cannot be understood apart from it. Long-term tendencies, such as the increasing division of labor, the increasing capacity of states to mobilize social power, the development of media such as money and writing, the differentiation of elites of various types as carriers of distinct cultural traditions, the tendency of pre-modern societies to oscillate between political centralization (patrimonialism) and decentralization (feudalism) all influenced and were influenced by religious beliefs and practices.

There has been a tendency among supporters of the idea of social evolution, Marxist and non-Marxist alike, to emphasize economic factors as causal, and religious developments as epiphenomenal. It has been common to believe that Max Weber reversed this understanding and held that religion was primary in social causation. A reading of Weber's entire work, which puts *The Protestant Ethic and the Spirit of Capitalism* in its proper perspective, makes it clear that such was not his belief. Rather, Weber argued for the relative independence of the religious dimension as a factor in social life, but one that always interacts with other factors, notably economic, political, and legal. S. N. Eisenstadt rephrases Weber by showing that it is not just a matter of emphasizing multi-causality: while never sufficient as a casual factor in itself, religion provides the indispensable element of *cultural codes*, not only for the institutionalization of new social forms but, when internalized in personality systems, for new forms of individual motivation as well (Eisenstadt 1999).

Religious elements have played a critical role in all the great social transitions. Terrence Deacon has even argued that the origin of language among the early hominids was probably facilitated by what he specifically calls ritual (1997). Some primates today can learn words, but only under the most intense pressure of human teaching. Deacon hypothesizes that the bodily and repetitive nature of ritual may have been a necessary precondition for the counterintuitive, from an early primate point of view, shift to symbolic representation that is the basis of language. In any case, Paleolithic evidence, difficult to interpret though it is, suggests that ritual played a central role in creating the social solidarity and respect for the normative order of kinship relations that differentiate humans from other primates.

The transition to archaic society is associated almost everywhere with monumental temple architecture, large-scale sacrifice, and divine kingship. It would seem that the move from the basic egalitarianism of tribal society to the intensely hierarchical and centralized order of archaic societies was only possible with the institutionalization of new religious codes, in association with an unprecedented concentration of social power.

While causal linkages are far from clear, the first millennium BC did see a number of developments that might have prepared the ground for the rise of the Axial religions. These include the beginnings of a market economy, an increase in bureaucratic and military power, and increasing differentiations of social classes, all of which could have given rise to Weber's 'problems of meaning' (1921–1922/1978). But an indispensable precondition for these radical new religious developments was the existence of cultural elites at least to some degree independent of central political power. These include the Hebrew prophets, the Greek philosophers, the Chinese sages, and the Indian ascetics and mystics. It was among these groups that archaic symbol systems were radically reinterpreted in a transcendental ethical and spiritual direction.

The influence of the Protestant Reformation on the emergence of modern society has been the subject of a vast literature. A significant recent development, however, has been the recognition of the importance of a 'disciplinary revolution,' particularly associated with the Calvinist wing of the Reformation (Gorski 1993). The disciplinary revolution operated at a number of levels: a new code of systematic ethical action expected from believers, new institutions capable of enforcing the code, and new solidary organizations strong enough to influence political and economic developments. The transformation of these new codes under a variety of influences and pressures into such tendencies as utilitarianism, liberalism, Jacobinism, and socialism, and more recently, individualism and fundamentalism, has significantly affected the shape of and the tensions within modern society.

Thus, viewing the process as a whole, it appears that religious evolution is not only a subject of great interest in itself, but is a critical dimension of social evolution.

See also: Civil Religion; Comte, Auguste (1798–1857); Reformation and Confessionalization; Religion: Definition and Explanation; Religion: Nationalism and Identity; Religion, Phenomenology of; Religion, Sociology of; Religious Fundamentalism: Cultural Concerns; Secularization; Social Evolution: Overview; Spencer, Herbert (1820–1903); Weber, Max (1864–1920)

Bibliography

Bellah R N 1957 *Tokugawa Religion*. Free Press, Glencoe Il
Bellah R N 1964 Religious evolution. *American Sociological Review* **29**: 358–74
Comte A 1830–42 *Cours de Philosophie Positive*. Bachelier, Paris
Deacon T W 1997 *The Symbolic Species: The Co-evolution of Language and the Brain*, 1st edn. Norton, New York
Dumont L 1980 *Homo Hierarchicus*. University of Chicago Press, Chicago
Eisenstadt S N 1968 *The Protestant Ethic and Modernization: A Comparative View*. Basic Book, New York
Eisenstadt S N (ed.) 1986 *The Origins and Diversity of Axial Age Civilizations*. State University of New York Press, Albany, NY
Eisenstadt S N 1999 Axial age sectarianism and the antinomies of modernity. In: McKnight S A, Pierce C I (eds.) *Politics, Order and History*. Sheffield Academic Press, Sheffield, UK
Gellner E 1988 *Plough, Sword and Book*. University of Chicago Press, Chicago
Gorski P S 1993 The Protestant ethic revisited: Disciplinary revolution and state formation in Holland and Prussia. *American Journal of Sociology* **99**: 265–316
Jaspers K 1948 *The Origin and Goal of History*. Routledge and Kegan Paul, London
Marett R R 1914 *The Threshold of Religion*, 2nd edn. Methuen, London
Spencer H 1876–1897 *The Principles of Sociology*. D. Appleton, New York
Taylor C 1989 *Sources of the Self*. Harvard University Press, Cambridge, MA
Tylor E B 1871/1974 *Primitive Culture*. Gordon, New York
Weber M 1904–1905/1930 *The Protestant Ethic and the Spirit of Capitalism* [trans. Talcott Parsons]. Scribner, New York
Weber M 1920–1921 *Gesammelte Aufsätze zur Religionssoziologie*. J. C. B. Mohr (Paul Siebeck), 3 Vols. Tübingen, Germany
Weber M 1920/1946 Religious rejections of the world and their directions. In: *From Max Weber: Essays in Sociology* [trans. Gerth H H, Wright Mills C]. Oxford University Press, New York
Weber M 1921–1922/1978 The sociology of religion. In: Roth G, Wittich C M (eds.) *Economy and Society*. University of California Press, Berkeley, CA

R. N. Bellah

Religion: Family and Kinship

This article will consider how these institutions, with their associated ideas and practices, intersect with one another. Family and kinship are closely related concepts, which together incorporate a number of issues, such as marriage, reproduction, family roles—particularly mother, father, children, and grandparents—descent and alliance, inheritance and taboo. Religion refers to the sum of beliefs and practices, ideas and attitudes, movements and institutions, symbols and experiences by which individuals and groups give meaning and order to their world, often with reference to a supernatural being or power. How has religion been involved in the social organization of kinship and the family, and how have issues surrounding kinship and the family affected religious beliefs, attitudes, and practices? In what ways have kinship arrangements and the family served the interests of religions? A range of different religions, family types, and kinship arrangements will be referred to, and contemporary debates about the relevance and ethics of the family, to which religious voices have contributed, will be discussed.

Most of those who have written about the institution of the family comment upon its resilience. Although a variety of family types have been in evidence in different historical periods and societies, families have always existed in order to create and strengthen alliances between groups and to provide a suitable location for reproduction and the nurture of children. A small family unit may incorporate a single parent and child, with larger units incorporating several generations of adults including siblings, and their children. In some societies the idea of family has referred to all those who live in a household, including pseudokin and servants. Kinship, though some have limited its definition to ties of blood, is generally held to refer to systems of both descent and alliance, and to both biological and social relationships. Anthropologists who have researched kinship have been interested, for example, in how certain groups have sought to maintain their blood-line through the generations, or have organized the family to pass down wealth through the male line, in how some groups have developed kinship arrangements to ensure rights over children through the control of women, and how others have used marriage as a means of creating powerful new kinship bonds. Such studies have often involved understanding how religious ideas or ritual practices have been used to reinforce or symbolize such arrangements. Arnold van Gennep, for example, in his famous study of rites of passage published in 1960 [1908], showed how a Turko-Mongolian group called the Bashkir used rituals to mark the important economic and social aspects of marriage. He described the extended visiting and eating rituals associated with the payment by the groom's family of the bride price, and showed how a lengthy game of hide and seek on the marriage day signified the loss of a daughter to one family and her gain by another.

Social scientists have also shown how religious ideas and practices have reinforced common conceptions of kinship and family life. Taking the example of the Christian Right in America in the 1980s, they have shown how religious groups were able to contribute to the formation of a popular family ideal. They have shown, however, that this ideal—of a nuclear family comprised of a heterosexual married couple with children—is not the traditional institution its supporters believed it to be so much as a postmodern

fantasy. 'The family is dead, long live our families,' a phrase used by Judith Stacey, a feminist sociologist, points to the redundancy of this popular ideal in the face of the complex reality of historical and living families in different social settings. How other religious groups have contributed, not only to reinforcing, but also to challenging ideas about the family, will be the subject of section 1.5.

1. *The Role of Religion in the Organization of Kinship and the Family*

1.1 *Kinship and Inheritance*

Religious ideas, myths, and practices have had a role in many societies in unholding social institutions, and many social scientists have sought to understand this. One example of this is the Indian caste system, which captured the attention of Max Weber and later anthropologists such as Louis Dumont. Weber held that there was a necessary relationship between caste and religion in India, that without caste there was no Hindu. He drew this conclusion because he saw that the religious rites and duties of Hindus were related to their social rank through a system of closed status groups or castes. These were different to tribes and guilds by virtue of their rules regarding marriage and eating. In particular, Weber noted the importance of the dominant group, the Brahmins, a priestly caste, around whom the entire system was organized. He recognized, however, that the modern caste system had grown to incorporate castes comprised of religious groups other than Hindus, and also suggested that it was severely shaken if not undermined by the challenge of modernism.

Later scholars considered aspects of the caste system in more detail. Dumont focused on hierarchy, examining the use of rules and practices concerning purity and pollution in distinguishing and ranking different castes. Others challenged his view, suggesting that those able to wield power in the system were not necessarily those who were most ritually pure. Most scholars have accepted some kind of relationship between religion and caste, however, acknowledging the role of scripture (especially the *Laws of Manu*), the function of the Brahmins in authorizing the system, the role of ritual in purification, and the part played by religious groups in both manipulating caste for advantage and challenging it.

Just as Indian kinship arrangements were legitimized by religious texts and commentators, so where those in other societies What came to be understood as the Confucian family ideal—*hsiao*, the virtue governing family relationships—was first prescribed in the classical texts of ancient China such as the *Book of Changes* and the *Book of Rites*. Within Islam, the family was privileged above other social institutions in the *Qur'an* and in later Islamic legal writings.

That European Catholicism reinforced ideas of feudal kinship and the family has been shown by the sociologist Bryan Turner among others. Developing the historical materialist ideas of Friedrich Engels on the family, he stated that the teachings of the Catholic Church in the Middle Ages on sexuality were important for reinforcing the norms of monogamous marriage and the control of women's sexual behavior. These norms, along with laws concerning inheritance through the male line, enabled dominant feudal families to ensure the stable transmission of property from one generation to the next and to avoid the dilution of the estate's wealth through illegitimacy. The Church further served this system through its Religious Orders which were often a destination for those excluded from inheritance, such as younger sons and unwanted daughters.

Religious teachings have played a part in social control at many times and in many societies. Women have often been the focus for this because of the importance of their reproductive role. For example, within Islam, polyandry (a wife taking more than one husband at a time) is prohibited, whereas polygyny is accepted, with the proviso that a husband should be able to provide equally for all his wives (a maximum of four) and children. This may be explained biologically because a woman's reproductive capacity is limited by the gestation period of the embryo, whereas a man may father multiple children simultaneously, thus reproducing his blood-line several times over. Socially, it may be explained by the fact that polygyny was the norm in Arabia at the time of Muhammad. Prevailing social practice was endorsed by the *Qur'an*, and, since that time, some Muslim marriages have been polygamous, particularly in African countries where polygyny was prevalent before the arrival of Islam. Many voices, including religious ones, have spoken out against this practice, but its place in Islamic scripture is unalterable given that the *Qur'an* is Allah's revealed word.

1.2 *Marriage*

The teachings and rituals associated with marriage are indicative of the value attributed by religions to the family as a significant institution. The family is given legal privilege in Islam, and in Judaism, and Sikhism it is of central importance. In Christianity, the family competes with the Church for priority, and, in Buddhism, with monastic alternatives. The presence of the ascetic life—lived alone or in community—in many religions challenges the adequacy of the family as the best place for spiritual progress, liberation, or salvation. In two modern, global sectarian movements, the Church of the Latter Day Saints and the Unification Church, the family is not only prioritized socially, but is of central theological importance in defining the nature of both the cosmos and divine–human relationships.

The social meaning of marriage is the alliance between two individuals and often between the families or clans of which they are a part. To avoid incestuous relationships, all societies have rules about who may marry whom. In many countries marriage alliances are legitimized in law. Religious understandings of marriage differ, although they all build upon these basic social requirements. In Judaism and Islam, marriage is understood as a contract divinely sanctioned by God, while in Christianity it is a sacrament. In Hinduism, it marks the beginning of a stage of life. Rituals of marriage often symbolize the social and religious ideas embodied by tradition. For example, the exchange of rings, the use of a marriage canopy, the act of circumambulation—together around a fire or of one partner around the other—and the sharing of garlands or food are symbolic of the nature of the marriage relationship, the intention of the partners, and, in the Jewish case of the breaking of a glass, of the fragility of the institution.

Religions have gone beyond these definitions and symbols to reinforce prevalent social ideas about marriage in their teachings, as the case of polygyny in Islam showed. Cousin marriage too is endorsed by Shar'iah or Islamic law, and is also accepted in Australian traditional religions, in order to maintain close, stable relationships and to give sanction to the retention of property within the group. Hindus, though expected to marry within their caste, are prohibited from marrying someone with a common grandparent, thus ruling out close cousin marriages. Arranged marriages, where parents or guardians select marriage partners for sons and daughters, are not obligatory in Hinduism or Sikhism, but they have become a social norm recognized by the majority of Hindus and Sikhs if not always adhered to by them.

Sexual partnerships outside marriage, both heterosexual and homosexual, have frequently been the subject of religious injunctions, often constituting a taboo or ritual prohibition. The breaking of marriage bonds or divorce, similarly, has been an issue where religious teachings have contributed to social control. The use of the *talaq* or marriage repudiation in Islam and of the *get* or bill of divorce in Judaism, both practices which endorse the rights of the male party over the female, have caused consternation among activists campaigning either for equal rights or simply for fairness in situations of relationship breakdown. However, both of these religions have made some provision for the property rights of women during marriage and after divorce, or the death of their husbands, in Judaism through the *ketubah* or marriage contract, and Islam through the *mahr* or dowry provided by the husband. In contrast to these two religions which, owing to their contractual view of marriage, have endorsed divorce, Catholic Christianity has eschewed it, seeing it as the breaking of a sacrament. Christian denominations in the West in the twentieth century, in the face of the increasing divorce rate, sought to reassert the value of marriage and to encourage reconciliation in the case of marriage breakdown. They were often divided on the question of whether second marriages should be sanctioned.

1.3 Roles and Relationships

The biological functions normally associated with the social unit of the family are reproduction, nurture of children, and preservation and control of fertile adults, particularly women. At different times and places, a variety of social arrangements have been in operation to facilitate these. Responsibilities for children, whether the provision of food and other resources, or of nurture, have been variously divided among men and women, parents and grandparents, though women have had a mothering role in most societies. Perhaps the most commonly recognized family roles have been those of mother, father, children, grandparents, siblings, and ancestors. Religious beliefs and practices have played an important part in endorsing such family roles and the duties associated with them. Roles have often been conferred through rituals, affirmed by religious teachings, and sometimes mirrored and authorized by mythic accounts of divine family roles.

The use of ritual in celebrating entry into different life-stages, conferring appropriate roles, and ensuring group membership, continuity, and loyalty has been the subject of anthropological debate. Functionalists, such as Durkheim and Malinowski, and later, Rappaport, stressed the importance of ritual for social integration, whereas Edmund Leach held the view that rituals reflected conflicting interpretations and often contained the seeds of social instability while legitimizing a particular power structure. This ambiguity was also acknowledged by Victor Turner who held that the symbolic complexity of rituals made space for competing interpretations, thus enabling solidarity to be achieved without compromise. These scholarly perspectives were significant for understanding many types of ritual, including those related to the family. For example, in his study of the Ndembu of Zambia, Turner described the various symbolic meanings of the milk tree, for girls at the time of initiation, for their mothers, and for the society of which they were a part. It symbolized breast milk and the mother–child relationship, the exclusion of men from women's experience, but also the separation of the growing daughter from her mother. Turner suggested that this tree constituted a dominant symbol which enabled the expression both of dissonant interests and of the solidarity of the group as a whole.

Religious rituals and symbols, then, may sometimes affirm divergent viewpoints about family roles and relationships for the social good. Religious teachings, however, are often less ambiguous, constituting normative statements about family life and the roles and duties of mothers, fathers, and children. Householder

life has been prescribed by religious and moral texts in many societies. A nineteenth century example of this may be found in the schoolbooks produced during the Meiji period in Japan when the imperial ideology of State Shinto was gaining its fullest expression. The emperor was held to be the supreme father or head of house, guiding his family, with each individual family being a branch of the same tree, following the same structure. This patriarchal ideal, which came to apply also to corporate families, was authorized by the Shinto myth of the sun goddess, Amaterasu-o-mi-kami, of whom the emperor was held to be a direct descendent.

A more recent example, which builds on earlier Confucian family ideals, may be witnessed in the teachings of the Unification Church, a movement which began in Korea in the 1950s and was popularized in the West in the 1970s. According to the *Divine Principle*, the movement's founder, Sun Myung Moon, and his wife are the true parents of the world. They seek to heal the broken family, bringing people together regardless of social or religious background. With God as the father of all creation, Jesus as the true spiritual father, and the holy spirit as the true spiritual mother, the movement mirrors this divine family in the human family, with God as the focal point, the male as head of the family, and the mother subject to her husband. With the essential fourth component, their children, they are expected to prepare, through active missionary service, for the second advent. The Unification Church is an example of a religious group for whom the family ideal is central in its theological, ethical, and social teachings, and for whom family relationships are the model for all other relationships.

1.4 The Significance of the Family for Religions

Having seen how religious ideas, practices, and institutions have contributed to forming and legitimizing kinship arrangements, marriage practices, and family roles and relationships, the significance of the family for the maintenance and continuity of religions will now be considered. As was seen in the case of the Unification Church, the family motif has, at times, been an organizing principle for ordering and giving meaning to both divine–human relationships and social life. Other late-modern religious movements, taking their inspiration from Christian rather than Confucian models, have likewise focused on the family as the central expression of community life. The Church of the Latter Day Saints, or Mormons, has seen the eternal family unit (including ancestors) as the primary institution for guiding the individual in his or her spiritual journey. Parents are believed to have duties, not only to care for children, but also to ensure their religious nurture and to confer priesthood upon them. As spiritual progress depends on membership of an eternal family, sex and procreation outside marriage are prohibited, and an active interest is taken in family history and genealogy. The Church's theological and ethical view of the family has led to particular family patterns developing among its members, e.g., a low marital age, large families, support of elders within families, a low divorce rate, and conservative sexual values.

Some Christian denominations and sects depend upon commitment to the family to maintain the group and pass on its beliefs and practices from one generation to the next. In the nineteenth century, Protestant sects in the New World such as the Hutterites, the Mennonites, and the Amish ensured their continuity through the generations primarily by means of the social control of families, marriage within the sect, ethical conservatism, and the religious nurture of children within the family. For some peoples, religious transmission across the generations has been related to ethnicity, itself determined by kinship arrangements, intergroup marriage, and close family ties. Despite having differing myths of origin and notions of identity, Jews, Hindus, Sikhs, and Zoroastrians have been defined by their relationship to a land or a people, and have carried their religious ancestry as a mark of birth. The family has been important for these groups in preserving biological and social ties with others in the group and with ancestors, and in maintaining the group for the future. Current concerns in Jewish and Zoroastrian communities about out-marriage are an example of the importance of family continuity. Additionally, debates within all these groups about the validity of the conversion of outsiders are indicative of the stress they have placed upon ethnic exclusivity. Within Orthodox Judaism, for example, Jewish identity and culture are only transmittable by the mother. If a Jewish woman marries a non-Jew, their offspring may return in the future to the Orthodox Jewish fold; if a Jewish man marries a non-Jew, their offspring may not. This represents an end to the Jewish line. Liberal, Reform, and Conservative Jews have a range of views on this and on the subject of conversion to Judaism.

The family is a major bearer of religious tradition. A new generation of young people may be brought into the life of a faith community by parents and grandparents, thus ensuring the survival of the religion. However, this process of religious reproduction introduces mutation and adaptation. New generations and changing times often require the reinterpretation of traditions, and the introduction of innovation, even if this is within existing religious frameworks. Many scholars of religions have been keen to understand intergenerational religious change, and to examine the methods of religious socialization used within various religions at different times. It can be instructive, for example, to compare the way in which one religious tradition is transmitted differently by families in different locations through variants on the same story, the use of local languages and rituals, and the

involvement of different family members. Traditions surrounding the birth of Jesus in Christianity, for example, are celebrated and passed on in quite different ways by Christian families across the globe.

1.5 Contemporary Religious Views of the Family

Nineteen ninety four was the United Nations International Year of the Family, and the institution of the family and the values associated with it were high on the political agenda in many countries. There was little agreement, however, about these issues. In America and Europe, some supported the nuclear family ideal outlined earlier, of two married parents with several children; others hailed what they called the 'postmodern family,' which was variously constituted of one or more adults of either gender and sexual orientation in a parental role with children. Realistically, in global terms, family types were still more open, including joint and multigenerational families, war-torn and other divided families, families with foster children, and childless families. Religious institutions were often asked for their views on the nature of the family and family values. What were their views on divorce, broken families, remarriage, and the postmodern family? What about birth control within the family, including fertility treatment? Did they have a view on gay and lesbian parenting? What did they think was the effect of women working on the family unit, especially on the socialization of children? A variety of views were offered, some traditional, some liberal, depending on the religious institution or movement and the denomination or group within it. The Catholic Church, for example, under the leadership of Pope John Paul II, continued to articulate a conservative position on marriage, the nature of the family, and sexual ethics. However, in practice, many Catholics contravened Church teachings. Some Reform and Progressive Jewish and liberal Christian groups made provisions to accept cohabitation, second marriages, and gay and lesbian unions, stressing the need for mutual support and care within partnerships and family groupings rather than traditional prescriptions about the inviolability of marriage and sexual ethics.

From the 1970s, the major collective voice for change regarding the family within religious institutions worldwide was that of women. Informed by the research and critiques of feminists, religious women from all over the world called for family-friendly teachings and practices, and respect for women's gifts and roles. In many countries, religious women challenged the law on family issues. In India, feminists fought for women's right to work outside the family, women's property and inheritance rights, and campaigned against dowry and female infanticide, all issues influenced by traditional Hindu and Muslim norms regarding women's roles and expectations.

Women were also able to challenge traditional views of the family by forging new religious roles for themselves or founding alternative religious movements. In Africa, in some traditional religions and new independent churches, women, through the acquisition of spiritual power, were able to circumvent the normal social requirements of wife and mother. Susan Starr Sered, in her research on religions dominated by women, gave examples of independent women who had broken with tradition and established new religious families and communities. She also showed that it was not only contemporary Western spiritual feminists who cried, 'No Father in heaven,' but also Thai, Burmese, Japanese, Brazilian, and Korean women who turned equally to female and male spirits for answers to problems, often seeing them as extensions to their own human families.

See also: Families and Households, Behavioral Demography of; Family and Gender; Family, Anthropology of; Family as Institution; Family Theory: Role of Changing Values; Kinship in Anthropology; Marriage; Motherhood: Social and Cultural Aspects; Religion and Economic Life; Religion and Gender; Religion and Youth; Religion: Morality and Social Control; Religion, Sociology of

Bibliography

Baker A 1993 *The Jewish Woman in Contemporary Society: Transitions and Traditions*. Macmillan, London

Burguière A, Klapisch-Zuber C, Segalen M, Zonabend F 1996 *A History of the Family* [*Histoire de la famille*]. Polity Press, Cambridge, UK, 2 Vols [trans. S H Tenison]

D'Antonio W V, Aldous J (eds.) 1983 *Families and Religions: Conflict and Change in Modern Society*. Sage, Beverly Hills, CA

Eriksen T H 1995 *Small Places, Large Issues: An Introduction to Social and Cultural Anthropology*. Pluto Press, London

Esposito J L 1982 *Women in Muslim Family Law*. 1st edn. Syracuse University Press, Syracuse, NY

van Gennep A 1960 [1908] *The Rites of Passage*. University of Chicago Press, Chicago

Gerth H H, Wright Mills C 1970 [1948] *From Max Weber: Essays in Sociology*. Routledge and Kegan Paul, London

Jackson R, Killingley D 1991 *Moral Issues in the Hindu Tradition*. Trentham Books, Stoke-on-Trent, UK

Kayongo-Male D, Onyango P 1984 *The Sociology of the African Family*. Longman, London

Pyper H S (ed.) 1996 *The Christian Family: A Concept in Crisis*. The Canterbury Press, Norwich, UK

Rabuzzi K A 1982 Family. In: Eliade M (ed.) *Encyclopedia of Religion*. Macmillan, New York

Sered S S 1994 *Priestess, Mother, Sacred Sister: Religions Dominated by Women*. Oxford University Press, New York

Stacey J 1996 *In the Name of the Family: Rethinking Family Values in the Postmodern Age*. Beacon Press, Boston

Turner B S 1983 *Religion and Social Theory*. Heinemann Educational Books, London
Young S (ed.) 1993 *An Anthology of Sacred Texts By and About Women*. Pandora, London

K. Knott

Religion, History of

1. Historical Background to the Term 'History of Religion'

The Latin term *historia religionis* first appeared in the sixteenth century and was employed alongside related terms such as *historia ecclesiastica* to entitle chronicles of important events in church history, and the life histories of influential theologians and witnesses of the Christian faith. The term was also applied to accounts on the origin and development of Judaism and portrayals of foreign religions. In his *Historia Religionum* (1697), Gebhard Theodor Meier used the concept in the sense of a comparative study, aimed at portraying a history of the three monotheistic religions. Compared with the more descriptive treatises on religion, comparative *historiae religionis* drew on theological distinctions, such as that between God's chosen people and heathens, to the end of justifying the exclusive truth claims of Christianity. In his *Abris einer Geschichte der Religionspartheyen* (1755), the Protestant theologian and student of Christian Wolff, Sigmund Jacob Baumgarten (1706–1757), examined these different methods and aims of religious historiography, explicitly distinguishing between a *usus historicus* and a *usus polemicus*.

From the 1880s onwards we find the introduction of new compound terms of religious historiography which competed with *history of religion(s)*, *Religionsgeschichte* or *histoire des religions* and reflected a heightened interest in scientifically recording and interpreting religious phenomena. The appearance of the terms 'philosophy of religion', 'science of religion', 'religious studies,' and—later—'sociology of religion', 'psychology of religion,' and 'phenomenology of religion' demonstrated the increasing differentiation within the scholarship dedicated to scientifically interpreting religious culture. However, none of these new academic subjects succeeded in establishing itself as a dominant or authoritative discipline for investigating and interpreting religious systems of belief. Academic research into religion is characterized, rather, by a high plurality of divergent and competitive disciplines that have been institutionalized quite differently, mainly in the countries of Europe and in the USA. This rivalry between disciplines also applies to the historical inquiry into religions: besides professional historians, who sought to render their discipline the new leading human science of the nineteenth century, other cultural scientists and theologians also participated in the historical study of religion and contributed significantly to its advances. As a result, most of the 'classical' works in nineteenth century religious historiography were not compiled by historians but by theologians (or former theologians), philosophers, ethnologists, philologists, sociologists, and professors in the nascent university disciplines, Sinology, Indology, and Egyptology.

In the course of these developments, the term 'history of religion' remained controversial. Friedrich Max Müller (1823–1900), who is regarded as the founder of the 'comparative science of religion' (*vergleichende Religionswissenschaft*), abandoned the term 'history of religion' in favor of the 'science of religion' in order to emphasize the scientific aims of an inquiry directed towards all historically given religious phenomena and traditions, and which strives to emancipate itself from the prejudiced approach of both theology and philosophy. Other researchers of religion, such as Cornelius Petrus Tiele (1830–1902), preferred the term 'comparative religion'. Nevertheless, common to all these different terms and their respective definitions of approach was the motivation to dispense with theologically orientated normative constructions in the study of religion. In Anglo-Saxon countries, this movement found its most important source of support from the so-called 'Lectures', an institution backed by the private endowments of wealthy patrons (in 1878 Friedrich Max Müller inaugurated the 'Hibbert lectures' and, in 1888, the first 'Gifford lectures' were given). The first professorship for the 'General History of Religion' was established in Geneva, in 1873; in 1877 the universities of Leiden and Amsterdam followed suit. In 1879 France's first chair in *Histoire des religions* was founded at the *Collège de France* and in 1886 the Graduate School of Sorbonne, *Ecole Pratique des Hautes Etudes*, instituted its *Section des sciences religieuses*. However, this appearance of several chairs in the 'history of religion' and 'religious studies' did not achieve the full emancipation repeatedly postulated by Christian theologians. In Germany, for instance, most religio-historical chairs were set up within evangelical–theological faculties; and even where this was not the case and provision made for institutional independence, as in The Netherlands, the majority of scholars appointed to such chairs were theologians.

2. History of Religious Historiography

The intellectual reflection on religion(s) can be traced back to scholars in Greek and Roman antiquity. Here we find the origination of several important topoi that continue to influence the study of religion (such as the critique of religion, theories on the origin of religion and its social function, the comparison between polytheism and monotheism). Notwithstanding this

degree of continuity, the ancient world's approach to the subject of religion cannot be termed 'historical'. Beyond the context of Graeco-Roman traditions, we can also find evidence of the examination of religious life in early Arabic, Persian, and South-East Asian scholarship. However, the influence of these traditions on European historiography of religion has been minimal. One central impetus for broadening Europe's religio-historical horizon was provided by the overseas conquests of the fifteenth and sixteenth centuries. What conquerors and travelers had to report on the foreign religious ideas and practices of the far-off lands they had visited led scholars to question the cultural self-evidence of Christianity, particularly after its confessional pluralization following the Reformation. In an attempt to intellectually overcome religious pluralism as well as the denominational conflicts within Christianity, E. Herbert of Cherbury (1581–1648) focused on a 'natural religion of reason' which he believed to be the true essence of all the different guises of positive religion. In doing so, this British deist introduced a distinction that would later be appropriated by Enlightenment theologians.

However, strictly speaking, it was the seventeenth century that gave birth to the historiography of religion proper. In the wake of the ecclesiastical schism and resultant religious wars, the Christian 'culture of unity' (Troeltsch) began to crumble and was soon replaced by an absolutist state which succeeded in politically neutralizing the vying religious parties by way of an absolute leader ruling over all religious parties. At the same time, the religious conflicts of the seventeenth century gave rise to the interest in a history of religion which was not conceived theologically, but had recourse to an impartial 'third' standpoint beyond confessional polemics. This was found in the 'historico-philological method' based on the principles formulated by Jean Mabillon (1632–1707) in his *De re diplomatica*, which was one of the first comprehensive treatises on methods of documentation and the study of hand-written manuscripts. The publication of this work in 1681 can be seen as the birth of textual criticism that played such a central role in the development of religious historiography. At about the same time, Richard Simon (1638–1712) applied literary-critical methods to the study of the Pentateuch in his *Histoire critique de l'Ancien Testament* (1678) and, in doing so, was one of the first to expose the religious history of the Judaeo-Christian tradition to the process of historicization. This rigorous distinction between the theological and historical historiography of religion also formed the guiding precept of Johann von Gro(e)ing's *Historie der heutigen Religionen* which appeared in Hamburg in 1702. In dismissing every form of *theologia polemica*, this author declared his commitment to a historical approach orientated towards the facts.

That the historian had not irrevocably displaced Christian apologetics, however, is evidenced by Jacques-Bénigne Bossuet's (1627–1704) work *Discours sur l'Histoire universelle* (1681) which appeared in the same year as Mabillon's important treatise and adhered to an interpretation of world history stressing God's saving grace. In opposition to Bossuet, Montesquieu (1689–1755) banished the guiding hand of divine providence from history and declared mankind to be history's focal point. The human being, expounded Montesquieu in *Esprit des lois* (1748), is determined by lots of factors—climate, laws, customs, and traditions as well as religion—the combination of which gives rise to the particular intellectual characteristic of a nation and its underlying 'spirit'. The work of the Göttingen exegete Johann David Michaelis (1717–1791) illustrates how Montesquieu's point of departure could be used to widen the religio-historical perspective. According to his own statements, Michaelis sought to view the Old and New Testament 'through the eyes of Montesquieu'. To this end he reconstructed in his six-volume *magnum opus*, *Mosaisches Recht* (1770–1775), the structure and contents of the Mosaic codex from the climatic, economical, political, and socio-cultural preconditions of its natural environment, and thereby offered an interpretation of this codex in the spirit of Montesquieu and in reply to the demands of his day.

Alongside this *historico-philological* examination of religion, we also find the appearance of the *psychological* approach to religion, as represented by David Hume (1711–1776) in his treatise *The Natural History of Religion*, published in 1757. Hume was not primarily concerned with explaining religion historically, but rather with accounting for religious ideas and practices as phenomena of the human psyche under the dictates of irrational emotions such as fear and hope. Because Hume saw reason as having gradually overcome these irrational forces in the course of history, however, this Scottish philosopher had also furnished the historiography of religion with an evolutionist perspective that would then feature as a key concept in the second half of the nineteenth century.

With the conception of world history as a developmental process from lower to higher, that is to say, from uncivilized to rational forms of life, religion came to represent not an exception to, but a manifestation of these historical dynamics. The evaluation of religion's status within this process of advance was divided: either one adopted a religio-critical stance, and assumed that reason would eventually replace religion (Voltaire, 1694–1778), or one took up a more positive, religio-historical position, and integrated religion into historical progress. One leading exponent of the latter position was Charles de Brosses (1707–1777). In his study *Du culte des dieux fétiches* (published in 1760) he employed a comparison between ancient and contemporary African cult ritual to support his claim that fetishism—understood as the worship of arbitrarily chosen natural objects or living creatures—is to be identified as the original form of all

religions. However, in the course of mankind's historical advance this has been successively superseded by the worship of celestial bodies, by polytheism, and then finally by monotheistic religions, he contended.

In his *Grundriß der Geschichte aller Religionen* (1787), Christoph Meiners (1747–1810) also construed the history of religion as part of a general historical process of cultural advance. As a result of this pronounced tendency in the eighteenth century to assimilate the portrayal of religion into an all-encompassing cultural historiography, the traditional anchor of religious history in the theological field of church history began to corrode, causing a telling shift of emphasis. Johann Christoph Gatterer (1727–1799), for instance, dealt with church history within the framework of a universal history of religion. In a similar vein, Christian Wilhelm Flügge approached church history as an aspect of religious history that he conceived as a branch of the general cultural history of the human spirit.

Resistance against this linear and optimistic idea of historical progress had already begun to emerge in the eighteenth century. For Jean-Jacques Rousseau (1712–1778), scientific and technical progress amounted to a moral step backwards behind the original and good 'natural state.' Even religion, he argued, has not escaped this cultural decline, but degenerated from an honest 'affair of the heart' into an institutionally depraved ecclesiasticism no longer capable of integrating the community. Nonetheless, because Rousseau considered religious feeling to be the strongest social bond, in *Contrat Social* (1762) he spoke up for a religious 'occupation' of the values which hold together the political community, viz. the establishment of a 'civil religion.'

Similar ideas were to be voiced by Alexis de Tocqueville (1805–1859) almost a century later who, in his famous work *De la démocratie en Amérique* (1834/40), prognosticated that the a politico-moral consensus within democratic societies would adopt the socially integrative role which religion had played in the social system of the Middle Ages. In eighteenth-century Germany, Johann Gottfried Herder (1744–1803) likewise objected to the belief in progress of Enlightenment thinkers, supporting the Romantic conception of history as a process of organic growth. In doing so, Herder influenced the organo-historical ideas of classical German historicism. This school of thought also rejected the Enlightenment's linear concept of history and belief in universal-historical progress, viewing history as the unfolding of individual cultural units.

The turn of the century into the 1900s was also a transitionary period for European historical consciousness which had begun to move away from universal generality and attach greater importance to that which is particular and individual. This change of perspective not only enhanced the status of the historical religions, which had been neglected by the concepts of historical progress of the eighteenth century, but also paved the way for the empirical research into the history of religion, especially the intensive analysis of ancient religious traditions.

Among others in Germany, Otto Jahn (1813–1869) and Karl Otfried Müller (1797–1840) devoted their scholarly activities to investigating ancient mythology. In his influential study *La cité antique* (1864), the French historian N. D. Fustel de Coulanges (1830–1889) examined the stabilizing function of religion for the Greek polis and Roman system of rule. Furthermore, this new interest in empirical research into religion led to what Edgar Quinet described in his work *Génie des religions* (1842) as an 'Oriental Renaissance': the first translations of India's holy scriptures had already begun to appear at the end of the eighteenth century.

In 1808, Friedrich Schlegel published his writings on the language and wisdom of India (*Über die Sprache und Weisheit der Inder*) whose conception of religious history diametrically opposed that of the Enlightenment. In India's ancient wisdom, Schlegel saw the realization of a unity between nature and spirit (Geist) which 'enlightened Europe' had forfeited and India itself submerged, and which he strove to reconstruct from the oldest of India's religio-historical sources. Drawing on Schlegel, Georg Friedrich Creuzer (1771–1858) and Joseph Görres (1776–1848) called into question the origins of antiquity's mythical tradition, claiming that Greek myths originally derive from the wisdom of Indian priests. With this Romantic enthusiasm for oriental languages and literature, the availability of translations of the latter and, not least, the advances in the field of 'historical and comparative linguistics since the turn of the nineteenth century,' the historiography of religion in the nineteenth century was afforded a wide range of new stimuli.

The endeavors of Franz Bopp (1791–1867) to distill a common Indo-European proto-language from a grammatical comparison between Sanskrit, Greek, Latin, Persian, and German, for instance, gave the comparative study of religions new impetus, temporarily pushing universal interpretations of history into the background. Aside from the work of Eugène Burnouf (1801–1852), a prominent figure in the historiography of Iranian religion, it was above all the services of his, and Bopp's student, Friedrich Max Müller, which rendered the philological approach beneficial to religious historiography.

Analogous to the Romantics' interest in the origins of religious history, Müller linguistically classified the Indo-European languages, particularly with reference to their respective names for God, to the end of reconstructing a primal religion common to all these peoples, which he saw as having once flourished in the advanced civilization of ancient India. Hermann Usener's (1834–1905) investigations into ancient religious history were also based on a combination of philologico-comparative studies and theoretical-

explanatory ambitions. With his programmatic work on an interdisciplinary ancient religious history—which his students Eduard Norden (1868–1941) and Ulrich von Wilamowitz-Moellendorff (1848–1931) advanced—Usener not only influenced the study of ancient religion, but also classical philology, art history and oriental studies, not least cultural theory until well into the twentieth century.

Besides the philological exploration of religious traditions, which chiefly focused on antiquity and the Orient, the second half of the nineteenth century also witnessed the proliferation of 'ethnological' research into contemporary, non-European religions which had not been taken into consideration by the romantically inspired religious historiography of such scholars as F. M. Müller. That these religions had become the focal point of attention within nineteenth-century religio-historiographical discourse was due to a revival of the evolutionist theory, aimed at assimilating the individual, historical, and divergent forms of religion into a universally applicable model of cultural advance.

In his major work *Primitive Culture* (1871), Edward Burnett Tylor (1832–1917) held the view that the religions of contemporary tribal societies represented the beginnings of religious history and therefore attached great importance to the investigation of such communities. However, Tylor's claim that animism (the attribution of a living soul to plants, inanimate objects, and natural phenomena) constituted the primal form of religion was soon superseded by other theories on the origins of religion, the most influential of which was William Robertson Smith's (1846–1894) totemism theory tracing religion back to the worship of ancestors symbolized by totemic emblems of plants and animals. Emile Durkheim's (1858–1917) writings on the elementary forms of religious life (*Les formes élémentaires de la vie religieuse*, 1912) can be seen as the most prominent adaptation of Smith's propositions. Religio-historical evolutionism proved to be an extremely resilient paradigm, providing the working hypothesis for James George Frazer's (1854–1941) reconstruction of religious history on the basis of contemporary ethnological material in his famous twelve-volume opus *The Golden Bough* (1890–1915).

In the latter half of the nineteenth century, ethnological and philological investigations into the history of religion found enormous support within academia and aroused ever greater interest among the general public. This trend was evidenced in the UK by the positive response to lectures in this field of study (cf. Sect. 1) and the huge popularity of Ernest Renan's (1823–1892) *La vie de Jésus* (1863) in France and David Friedrich Strauss' (1808–1974) works *Das Leben Jesu* (1835/36) and *Der alte und der neue Glaube* (1872), as well as Hermann Oldenberg's book *Buddha* (1890), in Germany. *Historical* scientists, by contrast, showed very little interest in religio-historical phenomena; the majority of professional historians adhered to a narrow politico-national approach to their subject in which religio-historical questions remained peripheral.

The few exceptions, held in high esteem today, were then decisive outsiders, such as Jules Michelet (1798–1874), who in his 17 volumes on the history of France (*Histoire de France*, 1833–1867) drew attention to the religious folklore of the Middle Ages and, in doing so, became a predecessor of the *histoire des mentalités* (cf. Sect. 3), or Jacob Burkhardt (1818–1897) who in *Weltgeschichtliche Betrachtungen* (1868–73) and *Die Kultur der Renaissance in Italien* reflected on religion as a third historical 'power' (*Potenz*) alongside culture and politics.

The cultural-historical revival at the turn of the twentieth century effectively placed religion at the forefront of scholarly discourse throughout Europe and the USA. The rapid erosion of time-honored cultural traditions, the growing insight into the social costs of capitalist modernization, and the new disintegrative tendencies in many Western societies provoked heated debates on the relationship between the religious history of European countries (as well as that of the USA) and their individual formation processes as modern societies. In defiance of traditional disciplinary borders, the historical research into religion became a central medium for the self-reflection of Western societies. From a systematic perspective, three closely connected issues came to play a decisive role in this intensive exploration of religious history around 1900. These issues included: the participation of religious traditions in the genesis of modern occidental rationalism; the socially integrative potential of religion in functionally differentiated and politically fragmented societies; and the cultural relevance of differences between Christianity's denominations on the one hand, and non-Christian religions, especially those of the East, on the other.

First and foremost, this discourse applied to Protestant theology which had largely viewed itself as a historico-cultural science of Christianity since the mid eighteenth century and, since the 1880s, became increasingly aware of the interrelations between the Christian religion and its non-Christian environment. Above all, it was the proponents of the so-called 'History-of-Religions School' (*Religionsgeschichtliche Schule*) within German Protestant theology who, in classical exegetic and religio-historical treatises, analyzed the interconnections between Jewish antiquity and the religious worlds of the ancient Orient, and the syncretistic character of Christianity. Although religio-historically minded scholars of the Old and New Testament, such as Hermann Gunkel (1862–1932) and Wilhelm Bousset (1865–1920), as well as the systematic theologian and cultural philosopher Ernst Troeltsch (1865–1923)—named the 'systematist of the History-of-Religions School'—viewed 'religion' as a relatively autonomous cultural power *sui generis*, they nevertheless gave greater priority to investigating the interactions between religious and other spheres of

culture. Ernst Troeltsch's 'The Social Teaching of the Christian Churches' (*Die Soziallehren der christlichen Kirchen und Gruppen*, 1912) is regarded as the most important work of this religio-historical group. In this publication, Troeltsch embarked on a cultural-historical reconstruction of the formative influence of Christianity's denominational churches and sects on social ideas and the effects of social structures on the formation of religious ideas and communities. Appropriating the typological distinction between 'church' and 'sects', developed by his friend Max Weber (1864–1920), Troeltsch differentiated between 'ecclesiasticism', 'sectarianism,' and 'mysticism' as specific forms of religious socialization. With Weber, Troeltsch shared the ambition to demonstrate the significance of religious ideas and practices for the emergence of the modern world. In *Die Bedeutung des Protestantismus für die Entstehung der modernen Welt* (1906), he focused on Protestantism and Calvinism, in particular.

While Troeltsch transformed church history into the cultural and social history of Christianity, Weber expanded his pioneering study 'The Protestant Ethic and the Spirit of Capitalism' (*Die protestantische Ethik und der Geist des Kapitalismus*, 1904/05) into an extremely elaborate enterprise of categorically comparing the economic ethics of the world religions (*Die Wirtschaftsethik der Weltreligionen*, 1915). In 'The Protestant Ethic', Weber had inquired into the effects of a religiously shaped *habitus* in areas of life and culture far removed from those commonly associated with religion in the narrower sense of the term—above all, in the economic sphere. In his comparative analyses of the economic ethos of the world religions, however, he primarily concentrated on occidental rationalism and sought to grasp its particularity and formative power on cultural life. Today, scholars are more skeptical of the details of Weber's ideas. Nonetheless, it cannot be denied that the worldwide reception of his theories and derivations in the study of religion has scarcely been matched by any other oeuvre.

3. *Overview of Recent Research*

Since the 1960s, the renaissance of religious movements in many societies, the upward trend of 'fundamentalism' and Protestant sectarianism, the spread of charismatic Christianity, and the manifold processes of a rapidly changing religious landscape (conversions, new syncretism) have attracted the attention of historians and led to a critical revision of the 'secularization theory', which equated social modernization with a general decline in religion. Religion has been rediscovered as a relatively autonomous system of communication and symbolism which shapes the *habitus* of individuals or identity constructions of social groups, and exerts a positive or negative influence on the development of societies. As a result, the old debates on the definition of religion and master narratives on modernization de-Christianization or secularization are being steadily replaced by micro-studies into the diverse religious milieus of the modern age. While religion has become a central topic in the historical discourse of all countries generally, several national-cultural particularities deserve mention.

In France, the work of Lucien Febvre (1878–1956) and Marc Bloch (1886–1944), together with the journal *Annales*, *Histoire*, *Sciences Sociales* (originally entitled *Annales d'histoire économique et sociale*) they founded in 1929, supplied the historiography of religion with an innovative approach which broke with the politico-historiographical tradition of the nineteenth century. Their objective was to uncover a temporal stratum of *longue durée* (Fernand Braudel, 1902–1985), that is to say, extensive periods of time which formatively shape not only political and socio-economical history, but also the concrete experiential world of the general population. Concepts stemming from *Annales*-historiography, such as 'structure' and '*longue durée*,' were then applied heuristically to the broader discussion of collective 'mentalities,' from which the historiography of religion profited in particular.

Although the term *histoire des mentalités* soon became a key concept of *Annales*-historiography, no agreement had been reached on the question as to what 'mentality' exactly means. These definitional problems notwithstanding, *histoire des mentalités* can be generally described as an historical discourse directed towards the collective consciousness of an epoch, i.e. its cognitive, ethical, and affective dispositions, which form the unquestioned cultural basis for the daily life of individuals and social groups, thus constituting their identity. In its application to the history of religion, this approach initiated a change in perspective away from the narrow argumentation of ecclesiastical historiography: religious history came to be seen as a history of religious consciousness whose carriers remain for the most part anonymous. Following this, historians of religion in France began to attach less importance to major events, personalities, and institutions, and focus on 'structural-historical' and, in particular, 'anthropological' lines of questioning.

Here their investigations have been concerned with historically varying attitudes to birth and childhood, mortality and death, corporeality and sexuality, nature and the environment, God and the Church, heaven and hell (cf. Ph. Ariès, M. Vovelle, J. Le Goff, J. Delumeau). To date, however, exponents of *histoire des mentalités* have scarcely researched the religious culture of more recent history, concentrating rather on the Middle Ages and early Modern Age, and—in co-operation with experts in ancient history such as J.-P. Vernant, M. Detienne and P. Vidal-Naquet since the

1970s—on the religious history of antiquity. Further drawbacks of applying this historical approach to religious historiography arise from its emphasis on the alleged durability of mental dispositions, its disregard of the processes of change within religious mentalities, and its neglect of the conflicts caused by differing religious backgrounds (resulting from the acculturation of foreign ideas and practices or due to inner-social differences).

In Germany, the 'socio-historical turn' of the 1960s and 1970s as well as the institutional separation of church history (pursued within theological faculties) and general history encouraged scholars to view religion and confession as nothing other than marginal factors in the history of the modern age. Since the 1970s, however, several attempts have been made to integrate religio-historical perspectives into social historiography (cf. W. Schieder, R. van Dülmen, R. von Thadden) and to open up historical research in church history and theology to the methodological debates of cultural and social scientists (cf. M. Greschat, K. Nowak, F. W. Graf). Numerous younger historians are now conducting research into a range of religious milieus, such as ecclesiastical organizations and festivities, theological controversies and quarrels over religious symbolism as well as the tense coexistence of Catholics, Protestants, and Jews. The collapse of East Germany, moreover, has generated a new sensibility for the political role of churches and their partly antagonistic, partly affirmative relationship to the 'political religions' of the twentieth century. And last, but by no means least, the holocaust initiated worldwide intensive research into the cultural climate of German Judaism and the conflict-ridden history that finally led to the destruction of the 'German–Jewish synthesis.'

In the UK, discussions continue to be determined by the weight of social anthropology (cf. B. Malinowski, A. R. Radcliffe-Brown). Here, from the outset—unlike in Germany—'society' and 'culture' were seen as one, and consideration given to the class-specific religion of the workforce and the processes of religious change in this country's urban centers (cf. H. McLeod). Particularly pioneering were the studies on the connection between religious radicalism and revolution (cf. Ch. Hill) and between Methodism and the emerging labor movement (cf. E. P. Thomson). Research into the processes of religio-theological differentiation in the Anglican churches has also drawn on the interpretative approach of *histoire des mentalités*.

In the USA, historians of religion since the 1980s have been strongly influenced by the 'interpretative' ethnology (stressing the concept of 'understanding' rather than that of 'explanation') and cultural anthropology of C. Geertz. Beyond the classical self-reflective discourse on the religion of native Americans, the faith of the Pilgrim Fathers, pluralism within Protestant denominationalism, the emergence of a specifically American Catholic self-understanding in the late nineteenth century and a 'civil religion' (N. Bellah) capable of integrating religious pluralism, in the most recent American historiography of religion themes such as gender, race, language, and the religious forms of minority groups have played an important role. In addition, the ethnological debate on 'Writing Culture' (viz. the constructive character inherent in self-portrayal and the description of that which is foreign; cf. F. Clifford/G. E. Marcus), the comparative analysis of fundamentalist movements in the world religions (M. E. Marty), the interpretation of Latin America's religious developments (D. Martin) and the interrelation between law and religion (W. Fallers Sullivan) have featured as central concerns of recent research. Current North American religious historiography has also profited greatly from the application of economico-theoretical approaches (R. Stark, Bainbridge W S (1985), Stark R, Finke R (2000)).

4. Outlook

Religion has gained in importance worldwide and this situation has accordingly afforded the research into religion greater significance. At the forefront of this discourse we find the questions concerning the interrelatedness of religion and ethos, the political relevance of religion, the construction of religious identities of minority and the conflicts between members of different religions and confessions. The debates on migration, religious conversion and new syncretism have shown, however, that historians of religion find it difficult to precisely grasp the long-term changes of religious mentalities. The correlation of conflict-ridden modernization, on the one hand, and the renaissance of strictly binding religion on the other, promises to become an issue of particular concern in the future: for the rapidly advancing globalization of capitalism, together with its constraints on heterogeneity, will only fuel those religious movements intent on communicating a more fulfilled form of human life which transcends an existence solely defined by the utilitarianism of economic rationality.

See also: African Studies: Religion; American Studies: Religion; Christianity Origins: Primitive and 'Western' History; East Asia, Religions of; Historiography and Historical Thought: Jewish Tradition; Islam: Asia; Islam: Middle East; Islam: Sub-Saharan Africa; Judaism; Latin American Studies: Religion; Near Middle East/North African Studies: Religion; Protestantism and Gender; Religion: Evolution and Development; Southeast Asian Studies: Religion; Western European Studies: Religion

Bibliography

Despland M 1979 *La Religion en Occident. Evolution des Idées et du Vécu.* Fides, Paris

Dubuisson D 1998 *L'Occident et la Religion. Mythes, Science et Idéologie*. Editions Complexe Brussels, Belgium
Feil E 1986 1997 *Religio*, 2 Vols. Vandenhoeck & Ruprecht, Göttingen, Germany
Hjelde S 1994 *Die Religionswissenschaft und das Christentum. Eine historische Untersuchung über das Verhältnis von Religionswissenschaft und Theologie*. Brill, Leiden, Germany
Idinopulos T A, Wilson B C (eds.) 1998 *What is Religion? Origins, Definitions, and Explanations*. Brill, Leiden, Germany
Jastrow M 1981 *The Study of Religion*. Reprint. Scholars Press, Atlanta, GA
Jordan L H 1986 *Comparative Religion. Its Genesis and Growth*. Reprint. Scholars Press, Atlanta, GA
Kemper D 1992 Religionsgeschichte I. In: Ritter J, Gründer K (eds.) *Historisches Wörterbuch der Philosophie*. Schwabe, Basel
Kippenberg H G 1997 *Die Entdeckung der Religionsgeschichte. Religionswissenschaft und Moderne*. Beck, Munich
Lehmann E 1925 Zur Geschichte der Religionsgeschichte. In: Bertholet A, Lehmann E (eds.) *Lehrbuch der Religionsgeschichte*, 4th edn. Mohr (Siebeck), Tübingen [1887–1889], Germany
McCutcheon R T 1997 *Manufacturing Religion. The Discourse on Sui Generis Religion and the Politics of Nostalgia*. Oxford University Press, Oxford, New York
Molendijk A L, Peter P (eds.) 1998 *Religion in the Making: The Emergence of the Sciences of Religion*. Brill, Leiden, Germany
Pailin D A 1984 *Attitudes to Other Religions. Comparative Religion in 17th and 18th Century Britain*. Manchester University Press, Manchester, UK
Pinard de la Boullaye H 1922–1925 *L'Étude Comparée des Religions. Essai Critique*. G. Beauchesne, Paris
Platvoet J G, Molendijk A L (eds.) 1999 *The Pragmatics of Defining Religion. Contexts, Concepts and Contests*. Brill, Leiden, Germany
Puech H C (ed.) 1970 1972 1976 *Histoire des Religions*, 3 Vols. Encyclopédie de la Pléiade, Gallimard, Paris
Rudolph K 1992 *Geschichte und Probleme der Religionswissenschaft*. Brill, Leiden, Germany
Saler B 1993 *Conceptualizing Religion. Immanent Anthropologists, Transcendent Natives, and Unbounded Categories*. Brill, Leiden, Germany
Schlesier R 1994 *Kulte, Mythen und Gelehrte. Anthropologie der Antike seit 1800*. Fischer, Frankfurt, Germany
Sharpe E J 1997 *Comparative Religion. A History*, 5th edn. Open Court, Chicago and La Salle, IL [1975]
Smart N, Clayton J, Sherry P, Katz S T 1985 *Nineteenth Century Religious Thought in the West*, 3 Vols. Cambridge University Press, Cambridge, London, New York
Stark R, Bainbridge W S 1985 *The Future of Religion: Secularization, Revival, and Cult Formation*. University of California Press, Berkeley, CA
Stark R, Finke R 2000 *Acts of Faith. Explaining the Human Side of Religion*. University of California Press, Berkeley, CA

F. W. Graf and A. Reuter

Religion: Mobilization and Power

In what ways does religion mobilize believers? How are they empowered by their beliefs, and to what end? Religion, as a phenomenon, has always contributed to both social conflict and social cohesion. It has the power to mobilize individuals, communities, and nations. Within religious institutions, power is as inevitable and ubiquitous as anywhere else. It can be the power of good, or morally ambiguous potency. Equally, it can come vested or masked as unction or inspiration. To speak of religion is always to speak of some kind of notion of power, whether it be attributed to divine, human, elemental, or cosmic sources. Relations of power are amongst the most significant aspects of any religion, especially when considering its context and cultural intercourse.

The study of power and mobilization in relation to religion must be rooted in theory, history, and praxis. For example, any consideration of church–state relations inevitably is concerned with control, legitimacy, and authority. It is impossible to conceive of a discussion concerning the origins of the Church of England that does not involve power–relations: those between the Pope and King Henry VIII; between the Reformers and Roman Catholicism; between competing powers at court; between competing ideological and political convictions on the nature of individual salvation and the centrality of the church; vernacular liturgies and scriptures, and their overall hermeneutical control and its centralization; between individual states and their subjects, and their right to any kind of religious choice, including that of dissent. The Church of England—'by law established'—is a form and exercise of power in nearly every era of its history and at every level of society, from government to parish, from individual to state. To study the church is to study power, in periodic, dispositional, episodic, and latent forms.

Globally, religion operates as a form of power and mobilization almost everywhere. In the USA, although no religion is established (the Constitution separates the churches and State), the outcome of increasing pluralism and secularisation has seen, amongst other phenomena, the mobilization of the New Right, and the movement of Roman Catholicism from a private to 'public' denomination, both on a select range of moral issues. As a number of scholars (Haynes, Casanova, etc.), have shown, this is a form of counterpower: the secularization theories (so prevalent in the 1960s and 1970s) have had to capitulate to the rise of a new religious consciousness in industrialized nations. They see modernity not as the harbinger of secularization, but rather of privatization, a by-product of late capitalism. Religion has fought against being squeezed into the private and subjective realm by recovering and reconstituting a public role. It is now a potent force for mobilization on a range of public issues in the political and social spheres. The example of the USA shows that the forces of pluralism, industrialization, and secularization can be combated by religious power at many levels.

In the last quarter of the twentieth century, the competition and association of powers between religion and society (e.g., Bosnia), church, and state

(e.g., Ulster) has operated differently but no less dynamically in Europe. In Eastern Europe, religion (mostly Christianity and Islam) played an important role in the diffusion of the communist monopoly. The vociferous activity of some Muslims in Western Europe has questioned the limits of liberalism in politics and education; in Northern Ireland, religious polarization has attempted to hold out against political reconciliation; the conflicts in Bosnia suggest that the purity of religious and ethnic enclaves is prized above any peaceable settlement. Added to this, there is a discernible rise in new religions, sectarian movements, and a patchy but popular inculcation of 'ordinary' spiritual ideology, indicating that religion remains a perennial public force.

In Africa, the Middle East, South America, Central Asia, the Indian subcontinent and South East Asia, religion has found itself harnessed to many nationalist causes. The symbiosis has often been so intricate as to make the political power play indistinguishable from that of the religious: Lebanon, Israel, Sikh nationalism in India, and Buddhism in Burma are perhaps the most notable examples. The origins of Pakistan lie in religious conflict and differentiation, as does civil war in Rwanda. Equally, it is also important to note that the resurgence of religious identity threatens existing religious powers, especially those allied to the State. Islamic movements in Turkey, Algeria, and Egypt demonstrate that religious fundamentalism is a powerful and popular force for social mobilization. In South Africa, religion has played a vital role in securing peace and democracy following the end of Apartheid; the 'Truth and Reconciliation Commission' was birthed in Christian theology, and chaired by an Archbishop (Desmond Tutu).

1. *Religion and Power*

Sociologically, distinctions between authority and force, and between violence and resistance, will and desire often cluster around the study of power. In political science, power usually is configured in terms of the ability of its holders to carry out their will, exact compliance, exert force, and compel obedience. To touch on authority, as part of power, is to study of relations, rules, and regulations. To focus on force, as another arm of power, is to consider compulsion, threats, and violence. In any consideration of politics, religion, or society, the subject of power (and its legitimacy) is never far from the surface.

The concept of power in social sciences interfaced with any confessional theology has no unity of discourse. Social scientists assess the 'power of God' through its social and observable outcomes: constructions of reality, rhetoric, sociality, and activity. On the other hand, believers claim to know and experience this same power personally and communally, in ways that appear to be either inaccessible to or immeasurable for social scientists. Part of the problem is this: 'power' is what we call 'something': it is primarily a noun, not a verb (Morriss 1987).

Although the verb does exist in the sense that we may speak of empowering, the fact that it is mainly a noun has implications for the way in which 'power' is described and studied. All too commonly, power is run together with verbs that are deemed to be its associates. For example, from a sociological perspective, a theoretician like Dahl (1957) describes power by contrasting it with near-synonyms like 'influence,' so that power becomes like influence. This becomes more obvious in May's *Power and Innocence: A Search for the Sources of Violence*, in which the noun again assumes the characters of the nearest relevant verbs. For May, power is exploitative, manipulative, competitive, nutrient, or integrative (1976). Each of these types of power represents a different level of threat, force, or coercion, ranging from the mutually empowering (integrative) to the potentially violent (exploitative). Thus, 'power'—at least so far as religious believers are concerned—loses its reality in some sense here, precisely because the verbs are allowed to project on to a passive noun. Perhaps this is why deconstructors of power such as Foucault (1984) have looked at power and seen 'nothing'—only reciprocal relations. On the other hand, 'realists' such as Weber (1947) have seen the concept of power as a 'highly comprehensive' schema within sociology for the analysis of circumstances and situations.

Van der Leeuw's (1938) phenomenology of power also recognises that one cannot prove there is such a 'thing.' He suggests that sensitivity to power has been largely forgotten in 'modern' society, but in primitive society, power developed into religious monism, which reached its climax in monotheism. In contemporary society, power is not encountered in nature or personality, but in the dispersed forms of social organization. However, the essence of religion is power, driven towards monism, which has a tendency to conflate the psychological with the cosmological, the sociological with the theological: religion is a description of concepts of power and their use.

More recently, and from an anthropological viewpoint, Lewis (1996) has also argued that although religious power assumes many different forms (which may appear to be unrelated or mutually exclusive), it is an appropriate 'theme' through which to approach religion holistically. For Lewis, religious power is best understood as a negative force that opposes malign spiritual powers. Thus, the shaman assumes the charismatic role par excellence—being able to exorcize the evil and prepare the way for blessing. In effect, the shaman stands between negative and positive power as a broker, a fixed point of passage and determinacy that both creates and obviates power. Like Van der Leeuw, Lewis cannot 'see' power, except in the way that it is related to, and in the social and ritual structures that are based on its assumption.

Thus, and sociologically, the conceptions of power can be divided into two very broad categories. On the one hand, there are those which are asymmetrical, stressing conflict, will, resistance, and the like: social relations are assumed to be competitive, conflictual, and dialectical. A *Jihad* ('fighting against unbelievers') may be understood in this way (Peters 1996). On the other hand, there are conceptions that imply that all may make some gain through power. Here, power is a collective capacity or achievement, born not out of conflict, but communal welfare (Lukes 1978). As McGuire (1992), many forms of Christian nationalism offer 'social cohesion.'

Although it is helpful to be mindful of these two broad categories, it should be noted that they are somewhat inadequate in the analysis of religion. To an extent, all religious movements have elements of both categories within them. This must be because the religious–social system itself is one of exchange alongside being one of power, which therefore allows for conflictual and communal action together. Beyond this, however, there is no account of subliminal power, which is so critical for an understanding of religious behavior. Power is not just attached to structures and positional personality, but also to nonstructural interaction and roles. Parsons therefore eschews 'canonically correct' definitions of power: it actually covers more than it can define. It must take account of magnitude, distribution, scope, domain, and outcomes. Moreover, it must consider resources, skills, and motivations. In considering power and mobilization in religion, the investigator is always close to the analyses of causality, which in turn refers to power; its measurement, exercise, distribution, and impact.

Morriss (1987) suggests that careful attention to the etymology of power reveals its own diversity, prior to any linkage to verbs. Power is 'ability to do or effect something or anything … energy, force, effect … an active property, capacity of producing … possession of control or command over others … dominion, rule, legal ability, commission, faculty … political or national strength.' That there is an overlap with verbs about power is clear: but that there is not complete synonymity is also clear. Morriss argues for a dispositional understanding of power. That is to say, power is seen as a relatively enduring capacity within any given object. Dispositional views of power are held in contrast (but not necessarily in opposition) to episodic ideas, that see occurrences or events as specific exercises of power. Morriss (1987) does point out that this power is not something that can be observed. Power is known only by its reification, which is ultimately an indirect but material way of inferring what power might be at work. This conclusion underlines his plea for power not to be studied in isolation: the subjects and objects of power require a methodological tolerance in the spheres of theory and evidence.

Clegg (1989) syncretizes the work of the previous theorists discussed, and effectively constructs an 'open' theory of power based on the concept of circuits, nodal points, and agency: it has political, organizational, and theological resonance. Clegg begins his work with an overview of earlier theoreticians such as Hobbes and Machiavelli. This is followed by discussion and critiques of Dahl, Russell, Bachrach, Baratz, Weber, and Wrong, revolving around the 'two faces' of power, namely the relationship between 'intention' and 'structure,' and which enjoys supremacy in society. According to Clegg, these debates, naturally enough, were eclipsed by more substantial models. Lukes proposed a three-dimensional model of power by identifying the centrality of 'interests'; Habermas focused on ideal speech situations, ideology, and hegemony; Giddens on the relationship between agency and structure in terms of power. Each of these speaks to the religion–mobilization axis.

But Clegg moves beyond each of these explications, working in Parsons, Foucault, and Mann (among others) into an overall integration that respects the dispositional and facilitative approaches to power, besides stressing the vital role of strategic agency. As he states, 'a theory of power must examine how the field of force in which power is arranged has been fixed, coupled and constituted in such a way that, intentionally or not, certain "nodal points" of practice are privileged in this unstable and shifting terrain' (1989). This is a critical observation in the field of mobilization, religion, and power, for change is neither predictable or necessary, yet its occurrence is beyond question.

Any discussion about the true nature of power and where it lies will be at the heart of all religious belief. Liberation theology, feminist theology, and other new 'disciplines' within religion could be read by some as being mainly in terms of reconstituting the balance of power in ecclesastical and social structures. The Nation of Islam has done the same within a North American context, although this is still developing. These theologies are effectively a form of revolution and resistance. The effect of the stress on power has led to more self-conscious questioning by some writers, who are beginning to question the nature of the power, that would generally be uncritically appealed to.

With respect to Clegg and Morriss, one of the difficulties of studying 'pure' power is that it is not one 'thing' to be observed. For example, a nationalist uprising in India may have a religious dimension functioning at many different levels: provision of identity, ideology, legitimization, coupled with notions of sacrifice, taboo, and salvation. Equally, an examination on the origins of Methodism would need to consider the power of mass meetings, the appeal of hymns, testimonies, and religious experience, not to mention the mobilization of displaced rural and agrarian populations in their search for meaning and community. Discerning power in such scenarios is a complex task.

Given the myriad of studies on power and attendant methods, the most practical way forward, especially given the multifarious nature of 'power' that occurs within religious movements, is to return to the agents or 'nodal points' of power identified by Clegg, and focus on their relationality. The advantage of Clegg's conceptualizing of power in terms of circuits, flow, and agency is that one can often identify the delimited 'field' of power (such as an organization), the directionality of power within the structure ('flow' or hegemony), and the specific nodal points of power such as a guru, infallible text or law, or episode (e.g., 'flashpoint'), as well as the latent forms of power embedded in culture, ethnic identity, and political situations.

This theory of circuits of power does not attempt to marry the 'nonrealist' tendencies of Foucault with the 'realist' or 'identification' trends of someone like Weber or Wright-Mills (although a focus on relations steers a middle course between realist and nonrealist accounts of power). Instead, it describes power in terms of circuits, which are organized through agencies. However, it does allow that the agencies cannot be seen as 'effortlessly rational or powerful,' since their 'carrying capacity is itself opened up for scrutiny in power terms.' This view of power is more contingent than monolithic. It has space for dispositional, facilitative, and episodic power within the same or overlapping frameworks. This has implications for the study of mobilization.

Dispositional power, as its description suggests, is the tendency or habit of an individual or group: in many religious movements, this is reflected in the 'grammar of power as a concept' that helps form the social and religious community, possibly even the State (its identity and constitution). Appeals to an 'almighty' God, or the 'monarche' of God, or to a 'Lord,' or to more specified concepts of omnipotence, have direct social and political consequences. Kings, Queens, dictators, and rulers in many faiths, past and present, have depended on such conflations. Any present sociopolitical governance can be a divine instrument and referrent that is used in the furtherance of subjugation and legitimization.

Facilitative power describes the points of access through which power can be reached, reified, and exchanged. Episodic power can be used to describe the 'surges' or 'events' of power that may alter the shape, perceptions, or behavior of individuals and groups. Central to these three 'circuits' is organization, and many a religious movement is able to demonstrate systematic configuration in each field. For example, the ideology present in postmodern charismatic worship could be said to be dispositional; the charismatic leaders are facilitative; the 'invocation' of the Holy Spirit a 'cue' for episodic manifestations of power to be unleashed (Percy 1996, 1998).

Episodic power is less easy to identity as a factor within mobilization. Any seminal moment can turn an ordinary crowd into a passionate army. A martyrdom, miracle, or the raising up of a messiah can provide a single moment from which power flows. The declaration of *Jihad*, the proclamation of a fast, or the threat of heresy to authority can do the same. The very genesis of religious mobilization can be traced back by believers to an event that revealed, defined, and motivated a new community to hold, channel, and disperse power in new and particular forms.

Charisma is also a factor within mobilization. One way of understanding the place of agencies in circuits is to see them as process of power-exchange, a mechanism where power is given up or received, or raw material transformed into a type of 'power.' In other words, John Wesley may have been a charismatic religious leader, but he also functioned within a charismatic situation, which he could not entirely control, although he might have directed.

An account for the resurrection of the charismatic religious leader in postmodernity lies, in part, in Weber's own writings (which in turn is indebted to Nietzsche's notion of 'will to power'). Weber differentiated between 'religion' and 'magic' in the way that he differentiated between ancient and modern. Religion was rational, organized, and functional. Magic was primitive, a legitimate form of domination, and placed power outside ordinary temporal spheres (1947). Within magic, there were prophets, magicians, and shamans. Weber saw the prophet as the most significant bearer of charisma, since their claims were based on personal revelations that ultimately developed a personal following (1965). Yet these distinctions are to be understood as fluid: priests and prophets practiced magic, and there was routinization in the prophetic as much as in the religious. As nearly every preacher knows, propaganda is the highest form of political authority: aims and objectives are seldom met by organization or negotiation, but rather by power and mobilization.

2. *Summary: Power and Mobilization*

Recent civil strife in Northern Ireland illustrates religion's capacity to promote both social cohesion (e.g., Corrymeela Community) and conflict (e.g., Ian Paisley's anti-Roman Catholic rhetoric). Behind this aspect of sociality lies a complex history of the struggle for power and factors in mobilization—both for war and peace. In terms of power, the divisions between Protestant and Catholic are simultaneously elementary and complex. The 'two traditions' do not just represent different religions, but rather different politics, histories, and ideologies. In terms of mobilization, the preservation and flourishing of Protestant and Catholic communities has often been cultivated through episodic and dispositional conflict, which has seen many casualties and fatalities. Power and mobilization are deeply intertwined through armed

hostility, marches of protest or ritual (Orange Day Parades), sectarian division, and political aspiration. Two 'civil religions' effectively have promoted civil division.

Yet the hopelessness of this situation, in which two forms of 'civil religion' (in terms of their power and mobilization) have failed to cancel each other out, has given rise to a third option: a peace process. Here, religion has been no less involved, although the forces for power and mobilization have recognized that a political solution should bring a degree of peaceable co-existence between competing religious–sectarian convictions. The emerging and fragile détente serves to remind that the goal for many who first engage in hostilities is actually to end conflict; whether they succeed or not, power and mobilization often end in some kind of settlement.

At the beginning of the twenty-first century, it still remains the case that some of the bloodiest conflicts known to humanity have been inspired by or utilized powerful religious symbols and ideas. Keeping the faith 'pure' and protecting the sacredness of the *Koran* has given rise to *Jihad* and *fatwah*. Promised lands and sacred sites continue to call communities and nations to arm, from Judaism to Hinduism, and from Ulster to Amritsar. Equally, faith and fear in the shadow of the Millennium remain a source of violence and violation, from Jonestown to Waco, from Aum Shinrikyo to the Manson murders, and from the Solar Temple and Cargo Cults through to Heaven's Gate. In the midst of this, 'established' religions continue to war in regions such as the Balkans, at great human cost. The history of power and mobilization in religion is, all too often, a litany of tragic events, a requiem for the human race.

Yet to see religion like this is only to view it from the dark side. Religion remains a vital force in much of the world. Millions if not billions of the world's population have a 'religious impulse'; this involves a quest for meaning 'that goes beyond the restricted empirical existence of the here and now' (Haynes 1998). Religion inspires; it binds together, creating communities of affinity and affectivity. It offers revival, revolution, and surprise; and then again respect for the past, remembrance and ritual. Gandhi, Pope John Paul II, and Martin Luther King have been instrumental in turning back tides of nationalism, communism, and racism (even though they have all used a reconfigured form of nationalism and religion to achieve political and social ends). Religion can be an irresistible force that can move the apparently unmoveable. For many, religion remains a viable form of liberation; some hope in the midst of despair.

In spite of the attempts of some sociologists of religion to retain substantial elements of a secularisation thesis (e.g., Wilson, Bruce), religion stands, in many parts of the world, as a protest against and an alternative to 'secular' life. Will American Christian fundamentalists, Egyptian and Algerian Islamists, or Hindu nationalists someday become nonreligious? It seems unlikely: religion retains its power, and its capacity to mobilize, evangelize, and scandalize.

Any discussion of power and mobilization in religion must take account of the following dynamics. First, religious groups and communities will perceive themselves as vessels or bodies that somehow hold, manifest, or reify divine power in a particular way, distinguishing themselves from 'society,' or perhaps their enemies, or other groups deemed to pose a threat or competition. Second, the structure of the group or community will most likely 'mirror' a proposed divine order; there will invariably be a conflation between the offices of leadership and the concepts of deity. Leaders of religious communities are supposed to resemble their gods. Many social communities attempt to correspond to a higher order that has theistic resonance. And yet, no democracy is immune from theocratic comparison; no theocracy can escape the process of routinization, as well as a degree of secularization. Third, there will be some sort of theology of power that will account for how divine power is to be reified and passed on: this is a key determinant in mobilization. Here, leadership, sacred texts, rituals, symbols, or stories play a vital part in moving believers from passivity to passion, and then to power. Every form of transcendence, no matter how 'other,' is received individually and communally, and worked out in social and political spheres. Fourth, the power that is witnessed to will be antisocial or suprasocial, requiring believers to be mobilized for change and social transformation whenever they perceive their own sense of divine order to be under threat, or require a new order to be imposed. It is against this background that an Islamic nationalist uprising, or Catholic and Evangelical picketing of abortion clinics, or the work of reconciliation and peace in South Africa and Northern Ireland, are to be ultimately understood.

See also: Religion: Definition and Explanation; Religion: Evolution and Development; Religion: Morality and Social Control; Religion: Nationalism and Identity; Religion: Peace, War, and Violence; Religion, Psychology of; Religion, Sociology of; Religious Nationalism and the Secular State: Cultural Concerns

Bibliography

Beckford J 1987 The restoration of power and the sociology of religion. In: Robbins T, Robertson R (eds.) *Church–State Relations*. Transaction Books, New Brunswick, NJ

Bettis J (ed.) 1969 *The Phenomenology of Religion*. SCM, London

Casanova J 1994 *Public Religions in the Modern World*. Chicago University Press, Chicago

Clegg S 1989 *Frameworks of Power*. Sage, London, pp. 210–15

Dahl R 1957 The concept of power. *Behavioral Science* **2**

Davis K 1942 A conceptual analysis of stratification. *American Sociology Review* **7**(3): 309
Dowding K M 1996 *Power: Concepts in the Social Sciences*. University of Minnesota Press, Minneapolis, MN
Foucault M 1984 Space, knowledge and power. In: Rabinow P (ed.) *The Foucault Reader*. Pantheon Books, New York
Haynes J 1998 *Religion in Global Politics*. Longman, London
Hemmerle K 1975 Power. In: Rahner K (ed.) *Encyclopaedia of Theology: A Concise Sacramentum Mundi*. Burns and Oates, London
Lewis I M 1996 *Religion in Context: Cults and Charisma*. Cambridge University Press, Cambridge, UK
Lukes S 1978 Power and authority. In: Bottomore T, Nisbet R (eds.) *A History of Sociological Analysis*. Basic Books, New York
May R 1976 *Power and Innocence*. Fontana/Collins, Glasgow, UK
McGuire M B 1992 *Religion: The Social Context*. Wadsworth, Belmont, CA
Morriss P 1987 *Power: A Philosophical Analysis*. St. Martin's Press, New York
Percy M 1996 *Words, Wonders and Power: Understanding Contemporary Christian Fundamentalism and Revivalism*. SPCK, London
Percy M 1998 *Power and the Church: Ecclesiology in an Age of Transition*. Cassell, London
Peters R 1996 *Jihad in Classical and Modern Islam*. Marcus Wiener, Princeton, NJ
Rahner K 1966 The theology of power. In: Rahner K (ed.) *Theological Investigations*. DLT, London, p. 391
Parsons T 1960 *Structure and Process in Modern Societies*. Free Press, Glencoe, IL
Van der Leeuw G 1938 *Religion in Essence and Manifestation*. Allen & Unwin, London
Weber M 1946 The social psychology of the World religions. In: Gerth H H, Wright-Mills C (eds.) *From Max Weber*. Oxford University Press, New York
Weber M 1947 *The Theory of Social and Economic Organisation*. Oxford University Press, New York
Weber M 1963 *The Sociology of Religion*. Beacon Press, Boston
Weber M 1965 *The Sociology of Religion*. Methuen, London
Weber M 1968 *Economy and Society*. Bedminster, New York, Vol. 1
Wright-Mills C 1956 *The Power Elite*. Oxford University Press, New York
Wright-Mills C 1958 The structure of power in American society. *British Journal of Sociolology* **9**(1)

M. Percy

Religion: Morality and Social Control

In most societies at most times religion has been the key institution devoted to answering three of the fundamental questions human beings have posed, namely, 'How and why were the universe, life and humankind created?,' 'What moral rules should we follow in order to lead good lives?,' and 'What becomes of us after we die?.' Most of the great world religions have answered the three questions together as part of a single coherent narrative. With some religions a just and righteous God was said to have created the world and its inhabitants, communicated moral decrees to them by revelation, and would after their deaths sit in judgment on them and assign them to paradise or to hell according to whether or not they had obeyed these precepts. Other religions held that the sum of an individual's good and bad deeds in this and previous lives would determine the nature of a reincarnation after their death according to the law of karma (Davies 1999). Clearly, beliefs and institutions of such fundamental significance are of great importance for social control. As Professor A G N Flew (1987, p. 71), the distinguished philosopher of religion, has noted:

> Can we not understand the hopes of the warriors of Allah who expect if they die in Holy Wars to go straight to the arms of the black-eyed houris in paradise? Can we not understand the fears of the slum mother kept from the contraceptive clinic by her priest's warning of penalties for those who die in mortal sin? Of course we can: they both expect—and what could be more intelligible than this?—that, if they do certain things, then they will in consequence enjoy or suffer in the future certain rewards or punishments.

The nature of the social control thus exercised varies greatly from one society to another, as does the degree to which emphasis is placed (a) on social conformity as part of the natural order of things, (b) on establishing, monitoring, and controlling the nature and content of an individual's conscience, or (c) on claiming to be able to affect or at least to predict a person's fate after death.

Sometimes these forms of social control have simply reinforced other forms of power available to the secular rulers of a society, as when these rulers have also been the key religious authorities or the two groups have been closely allied. Likewise, resistance to the ruler might well be reinforced by religious dissent (Martin 1978) but with the caveat that the degree of social control that the dissenting group, whether the seventeenth century English upholders of the Cromwellian Commonwealth, the Roman Catholic exercisers of a moral monopoly in Ireland (Inglis 1998), or the Sikhs of the Punjab, exerts over its members may be exceptionally high.

Religion can, of course, be an independent source of power and social control, sometimes exercised over matters that are of little interest to the secular authorities. The medieval Roman Catholic Church's forbidding of marriages between cousins up to the sixth degree or the refusal of the Greek Orthodox Church to permit a person to marry a fourth time after the deaths of three spouses were forms of social control that were at times a nuisance to the rulers of the relevant societies, who were concerned mainly with forming dynastic alliances and producing male heirs. They were part of a set of independent social controls

Table 1
Trends over time in England and Wales

	Reforming Britain 1860–1905	Respectable Britain 1905–60	Permissive Britain 1960–2000
Religious adherence and membership	Rising	Falling slowly and unevenly	Falling rapidly
Crimes of violence and dishonesty	Falling	Rising slowly and unevenly	Rising rapidly
Illegitimacy	Falling	Steady at a low level except in wartime	Rising rapidly
Drug and alcohol abuse	Falling unevenly	Falling to a low level	Rising

exercised through these churches' monopoly over the creation and dissolution of marriages, and hence over sexuality, reproduction, legitimacy, and inheritance, i.e., over crucial aspects of not only personal but also material and political life.

This form of social control may be seen as what Polanyi (1951) has termed deliberate or corporate order comparable to, say, a centrally planned economy, statute law, or the internal workings of a private bureaucracy. In each case there are, of course, divisions of interest, contested areas of authority, and competition between individuals and sectors using the exploitation and interpretation of rules to enhance power, but in principle there is a hierarchical system that may not be contested beyond a certain point. Rome has spoken, the matter is closed and the Brahmins alone can determine a caste's pedigree and thus influence its position in a hierarchy of purity of which they are the apex (Dumont 1970). Yet in all cases in traditional societies, most social control would have been local. Violence, theft, or sex with other men's wives would be kept in check by local opinion in one's village, *mir*, *mura*, *jati*, or kinship group rather than the direct use of priestly authority, the threat of divine wrath, or the dread of hellfire or of reincarnation as a scavenging dog. A mixture of respect for and fear of local chiefs, landowners, or state officials will also have played a part. Religion was just one part, though an important part, of local social control. Beneath the designed order lay a myriad local communities held together by an alternative form of order, the interactive order of face-to-face communities, the essence of *Gemeinschaft*.

Polanyi (1951) contrasts deliberate or planned order with spontaneous order which can be observed wherever stable arrangements evolve as a result of a large number of independent but interconnected decisions, as with the operation of markets through the price mechanism, systems of common law based on cumulative precedence, or scientific progress through conjecture, experiment, refutation, and provisional confirmation. No one directs or designs these large, widely dispersed, flexible, and responsive instances of spontaneous order, yet they are much superior in their efficacy to most attempts to create a centrally controlled designed order. Their equivalent in religious terms and one with a crucial importance for social control may be seen in the flourishing of Protestant sects and denominations, and even of factions within the established churches in Britain during the nineteenth century, as the world's first urban, commercial, and industrial society was established and consolidated. This religious spontaneous order enabled Britain to shift from being a society of small, rural villages and towns based on local social controls to one of large, modern, urban conurbations while experiencing not crime and chaos but steadily declining rates of crime (Gattrell and Hadden 1980), illegitimacy (Hartley 1975), and drink and drug excesses (Wilson 1940). Britain was a more crime free and less deviant society in 1900 than it is today and than it had been 50 years before (Davies 1994).

Over time the pattern of crime and deviance in Britain follows a ∪-curve, falling markedly in the later half of the nineteenth century (Gattrell and Hadden 1980), reasonably stable in the first half of the twentieth, and rising very rapidly from the mid-1950s (Davies 1994). These changes are inversely shadowed by the growth and decline in church adherence in England and Wales, which grew rapidly in relation to population in the latter half of the nineteenth century and peaked in about 1904–5 (Brown 1992). There then followed an irregular decline with occasional periods of recovery until the late 1950s, after which there has been a steep and unreversed decline. There was a ∩-curve in piety which was an exact inverse of the ∪-curve in crime and deviance. No other variable fits this pattern with any such degree of exactitude. It is very

Table 2
Sunday School Enrollment

Year	Percentage of population aged less than 15 years enrolled in Sunday Schools
1818	12
1851	38
1891	52
1901	53
1911	51
1931	46
1961	20
1989	14

strong proof indeed of the effectiveness of spontaneous religious order as a means of social control. The relationship is summed up in Table 1 (Davies 1994, Hartley 1975, Lynn and Hampson 1977) showing the main trends between 1860 and 2000, trends that divide the period clearly into three phases 'Reforming Britain,' 'Respectable Britain,' and 'Permissive Britain' (Davies 1975).

Religious adherence and membership, although at all times involving a minority of the adult population, were a good index of the strength of the values of respectability, and thus of spontaneous self- and social control in British society. The incidence of deviant behavior clearly fell and then rose in Britain as this form of social control rooted in religion first became stronger and then later went into decline. The churches and chapels were the crucial institution promoting respectability in the nineteenth and early twentieth centuries, and membership and attendance were one way in which respectability could be acquired and proclaimed. Rising levels of membership and attendance gave those who were committed to their religion the confidence to promote and foster the virtues of self-control and respectability among the non-attending majority whose attachment to a church or denomination was at best nominal. To a substantial extent, though, they were receptive to the morality of respectability, and in particular they wanted it instilled in their children. The rise and the later fall of Sunday School enrollment and attendance shown in Table 2 (Gill 1992, Laqueur 1976, Martin 1967, Wilkinson 1978) is perhaps an even better measure of the changing importance over time of religion as a means of social control. It has been suggested reliably that between two-thirds and three-quarters of those enrolled were generally present in school on a Sunday (Gill 1992, p. 96).

The importance of popular Protestantism as a source of spontaneous social control and moral order in Britain during a time of massive internal migration to the new industrial and commercial towns and conurbations (Munson 1991) can be seen at work today in contemporary Latin America, a continent currently undergoing a similar transformation (Martin 1990). Migrants leaving small, agricultural communities to work in the industrial and services sectors in societies as diverse as Chile and Guatemala have, like their British predecessors, sought to create for themselves through religion a spontaneous moral order, a social control based on respectability (Martin 1998, Martin 1990).

Protestantism has 'once again created a critical mass of persons for whom self-discipline, a work ethic and the values of domestic rectitude have become second nature' (Martin 1998, p. 28. See also Weber 1930). The truth has set them free from what otherwise would have been a morally degrading existence. Whether they will be able to sustain these virtues in the long run remains to be seen.

The importance of religion as an autonomous and potent means of social control has often been neglected by sociologists and political scientists, who are blinkered by their own obsession with economic and political power, and unable to appreciate the importance of religion. When they do turn their attention to religion as a means of social control, it is by choosing selectively examples such as the Inquisition, the Ayatollah Khomeini's regime in Iran, or human sacrifice by the Aztecs, which enable them to caricature religion-based social control as backward, oppressive, and irrational. On the contrary, social control rooted in religion is as varied as any other form of social control, and the distinctive spontaneous forms of social control that religion has provided in some countries during the difficult transition from a traditional to an urban and industrial society have been remarkable for their peaceable nature and civilizing influence.

See also: Religion and Economic Life; Religion and Gender; Religion and Politics: United States; Religion and Youth; Religion: Family and Kinship; Religion, Sociology of; Stigma, Social Psychology of

Bibliography

Brown C G 1992 A revisionist approach to religious change. In: Bruce S (ed.) *Religion and Modernization*. Clarendon Press, Oxford, UK, pp. 31–58

Bruce S (ed.) 1992 *Religion and Modernization*. Clarendon Press, Oxford, UK

Davies C 1975 *Permissive Britain*. Pitman, London

Davies C 1994 Does religion prevent crime? *Informationes Theologiae Europae*, *Internationales ökumenisches Jahrbuch für Theologie*, Peter Lang, Cologne, Germany, pp. 76–93

Davies C 1999 The fragmentation of the religious tradition of the creation, after-life and morality: Modernity not post-modernity. *Journal of Contemporary Religion* **14**(3): 339–60

Dumont L 1970 *Homo Hierarchicus: The Caste System and Its Implications*. Weidenfeld and Nicolson, London

Flew A G N 1987 The logic of mortality. In: Badham P, Badham L (eds.) *Death and Mortality in the Religions of the World*, Paragon House, New York, pp. 171–87
Gattrell V A C, Hadden T B 1980 The decline of theft and violence in Victorian and Edwardian England. In: Gattrell V A C (ed.) *Crime and the Law, The Social History of Crime in Western Europe since 1500*, Europa, London
Gill R 1992 Secularization and census data. In: Bruce S (ed.) *Religion and Modernization*. Clarendon Press, Oxford, UK, pp. 90–117
Hartley S F 1975 *Illegitimacy*. University of California Press, Berkeley, CA
Inglis T 1998 *Moral Monopoly, The Rise and Fall of the Catholic Church in Modern Ireland*. University College Dublin Press, Dublin, Ireland
Laqueur T W 1976 *Religion and Respectability, Sunday Schools and Working Class Culture*. Yale University Press, New Haven, CT
Lynn R, Hampson S L 1977 Fluctuations in national levels of neuroticism and extroversion 1935–1970. *British Journal of Social and Clinical Psychology* **16**: 131–8
Martin B 1998 From pre-modernity to post-modernity in Latin America: The case of Pentecostalism. In: Heelas P (ed.) *Religion, Modernity and Post-modernity*. Blackwell, Oxford, UK, pp. 101–46
Martin D 1967 *A Sociology of English Religion*. Heinemann, London
Martin D 1978 *A General Theory of Secularization*. Blackwell, Oxford, UK
Martin D 1990 *Tongues of Fire, The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK
Munson J 1991 *The Nonconformists, In Search of a Lost Culture*. SPCK, London
Polanyi M 1951 *The Logic of Liberty*. Routledge & Kegan Paul, London
Weber M 1930 *The Protestant Ethic and the Spirit of Capitalism*. Allen & Unwin, London
Wilkinson A 1978 *The Church of England and the First World War*. SPCK, London
Wilson G B 1940 *Alcohol and the Nation*. Nicholson and Watson, London

C. Davies

Religion: Nationalism and Identity

Nationalism and religion share the ambiguous status of being the most potent and inspiring sources of collective identity, but also the most destructive forces for mobilizing large numbers of people and generating social conflict and change. The conventional view of their relationship is conflictual and evolutionary: the nation is opposed to the religious community, as nationalism is to religion, and national identities must, and will, in time supersede obsolete religious allegiances.

For 'modernists' like Ernest Gellner, Tom Nairn and John Breuilly, nationalism is a purely modern, secular ideology; it appears only when the age of tradition and religion gives way to that of modernity and the nation-state. This was very much the view of Elie Kedourie, for whom nationalism was a doctrine invented in Europe at the beginning of the nineteenth century. As a secular ideology of collective Will and as a fruitless quest for unattainable perfection on earth, nationalism embodies a godless, antinomian chiliasm which ultimately can be traced back to medieval Christian millennial doctrines. Its effect is to subvert duly constituted authority and destroy the normal barriers between the public and private domains in the interests of a pure, acosmic brotherly love—in Africa and Asia, as well as in its European seedbeds. With equal regret and ire, but from the other end of the ideological spectrum, Marxists like Eric Hobsbawm can only see in nationalism, particularly of the 'divisive' post-1870 ethnolinguistic variety, a new set of 'invented traditions' in the hands of ruling capitalist elites and a despairing popular reaction, like fundamentalist religion, to the massive forces of social change and the failure to realize the project of socialist internationalism. Nationalism, in short, like religion, is increasingly a diversion and obstruction to the great globalizing 'movement of history.'

This view has not gone unchallenged. Impressed by the evidence of 'religious nationalisms' in Asia, Africa, the Americas, and Europe, scholars like Mark Juergensmeyer have argued that 'nationalism' and 'religion' are often natural allies, and that religious movements like Shi'ism are seeking to reclaim the nation from the embrace of the secular state. For others such as Conor Cruse O'Brien, this affinity of religion and nationalism, or 'holy nationalism,' can be traced back to the Old Testament. It is also possible to argue that such well-known cases of secular nationalism as Jacobin France, Kemalist Turkey, Stalinist Russia, and Maoist China did not attack religion as such, and were in any cases transient phases of revolutionary politics which always sees the return of a society's 'central values.' From the standpoint of collective identity, religion and nationalism can often be mutually reinforcing, and hence classic 'weapons of the bourgeoisie,' in Leninist parlance.

A third view sees in nationalism a species of religion in its own right, competing with traditional world or local religions for mass allegiance. Already in the 1960s, functionalists like David Apter had shown how nationalism performed the functions of a mass-mobilizing 'political religion,' a characterization recently developed by George Mosse in his penetrating studies of the liturgies of the civic religion of European, mainly German, nationalisms, and by Bruce Kapferer who, in an illuminating comparative study of Sri Lankan and Australian ideological politics, has demonstrated the ways in which a nationalist religion copes with problems of ontological insecurity and ethnic conflict. Benedict Anderson, too, though he fails to develop the point, sees nationalism more closely aligned to religion than political ideology, and the imagined community of the nation as emerging

out of, and against, the earlier world religious communities.

Each of these views has much to commend it, and can point to considerable supporting evidence. Equally, it is not difficult to find counterexamples. For every 'secular' nationalism drawing on ancient classical or pagan motifs and myths such as the Finnish Kalevala or the Homeric Iliad, we can find equally fervent religious nationalisms inspired by their respective religious ideals and appropriating their ethno-religious heritages, as the modern Hindu Indian movement has drawn on the Ramayana and Mexican nationalism on the cult of the Virgin of Gaudeloupe. The influence of traditional religions may prove equally ambivalent. If Buddhism underpins and is reinforced by Burman and Sinhalese national identities, the conflict between Protestant and Catholic Christianities only served to undermine the growing sense of German or Swiss national identities. As for Islam, it can buttress an all-Arab nationalism, but can equally undermine the Arab circle of allegiance in favor of the wider Muslim umma, or, in the hands of radical populist regimes like the Ba'ath Party, support the narrower circle of the Syrian or Iraqui nation-state.

This suggests that there can be no simple characterization of the relationship between the two modes of culture, identity, and belief. We have only to recall the contradictions that bedeviled the French revolution, often regarded as the first full-blown case of secular nationalism. From the very first oath sworn by the deputies on the tennis court in June 1789 to the last Jacobin journée, the *Fête de l'Etre Suprème* in June 1794, classically-inspired secular and Christian religious motifs and rituals were combined, inspiring in the multitudes a powerful emotional blend of patriotic, fraternal political activism. Thus, as Simon Schama points out, at the *Fête de la Fédération* held on the Champs de Mars on July 14, 1790, Lafayette administered the oath to country and king and Talleyrand provided the mass and benediction, intoning: 'Sing and weep tears of joy, for on this day France has been made anew'; while over at Notre Dame, they performed a half sacred, half profane cantata entitled *Prise de la Bastille*, set to passages from the apocryphal, patriotic Book of Judith.

Similar religious motifs pervaded other nationalist celebrations; for example, the patriotic German festivals of the Wartburg in 1817 and at Hambach in 1834, described by Mosse, as well as the six-hundredth anniversary of the Swiss Confederation and the Oath of the Rütli, held in Schwyz and on the Rütli meadow in 189l. In the twentieth century, nationalist celebrations continued to employ religious imagery, most notably in the cult of 'fallen soldiers' and patriotic sacrifice, whether in the two world wars or the many more regional wars fought in the name of the nation, or the nation-to-be.

These considerations point to a more complex interplay and layering of religious and national identities and experience. Much of the confusion surrounding this problem stems from divergent definitions of both 'religion' and 'nationalism.' The evolutionary modernist 'secular replacement' perspective employs a traditional definition of 'religion' in which our everyday world is believed to be controlled by a supraempirical order (and/or beings) which is the source of salvation—a theocentric and cosmic belief-system; while its definition of nationalism as the aspiration of a people and culture to be autonomous and united in its homeland emphasizes its secular, linear, and anthropocentric character. It is easy to see how, on this schema, nationalism and religion become opposed: the nation replaces the deity, the nationalist movement the church, and posterity the after-life. For the alternative view which sees nationalism and religion as 'natural allies,' or even conflated as 'holy nationalisms,' nationalisms can take as many forms as there are peoples and nations; similarly with religions. So nationalism and religion are really continuous and mutually reinforcing, or superimposed; nationalisms are expressions of belief, or 'faith-communities,' just as churches or religious communities are really 'nations.' As John Armstrong suggests, religious communities easily slide into ethnic groups, and vice versa. But this is, once again, to muddy the waters, and preclude the possibility of a more discriminating analysis. After all, are all nationalisms equally 'holy'? Does religion always reinforce national identities? Isn't the 'faith' of nationalists quite different to that of a traditional believer in a local or world religion? That, at least, is how it has often appeared to many participants who faced difficult choices about the claims of God and state, religion and nation, particularly where the religion was transterritorial, even universal in its claims. Hence also the vital differences between secularizing nationalisms like the republicans in France or the Westernizers in Tsarist Russia, and the religio-monarchical nationalists in France and the religious narodniks in Russia.

This is where the third, or 'political religion,' perspective may help. We do not have to subscribe to its belief in the substitution of 'religion' by 'nationalism,' but we can with profit employ its wider, Durkheimian approach to religion—as a system of beliefs and practices relative to sacred things—to understand how and why modern nationalisms, defined as ideological movements for the attainment and maintenance of autonomy, unity, and identity for populations deemed to constitute 'nations,' can fulfill many of the identity-creating functions of traditional religions. This makes nationalism both a competitor to, and an ally of, traditional religions, depending on historical and social circumstances.

In practice, this tension and symbiosis between 'religion,' so defined, and 'nationalism,' produces a series of familiar mass phenomena. These include:

(a) The politicization of religion, in which tra-

ditional religious motifs are transmuted into secular nationalist features. Thus prophets, saints, and religious lawgivers like Moses, Muhammad, and Confucius are metamorphosed into national heroes, religious revelations are turned into national shrines such as Yasna Gora and Guadeloupe, scriptures like the Bible and Bhagavad-Gita are re-read as national epics, and miracles like the eight-day lamp of Chanukah become national feats of the Maccabees. Even commemorations like the death of Hussein at Karbala, or the Exodus across the Red Sea, are re-enacted as nationalist festivals in Iran and Israel.

(b) The messianization of politics, whereby the nation-state is sanctified and exalted, and its leaders are endowed with religious charisma. This was how their followers came to regard nationalist leaders like Nkrumah, Nasser, Herzl, and Soekarno, investing them with superhuman qualities as messianic deliverers and saviors. But the messianization of politics and the sanctification of the nation-state does not stop with the leaders; it includes the sacralization of ethnohistory, such as the Gaelic history of Ireland, the Christian mission of Poland, and the Hinduization of ancient India; the belief in millennial destiny such as we find in the United States or communist Russia; and the 'invention of traditions' such as the rites and ceremonies of Bastille Day, Independence Day, the Rutli Oath, and Armistice Day at the British Cenotaph.

(c) Religious populism, or the exaltation, even deification, of the 'people,' the ordinary people, as opposed to the elites. This may include the quest for purification of the people and their culture, in the tradition of Herder's cultural populism, sought by so many Central and East European nationalists, or by Indian, Arab, and West African nationalists like Edward Blyden; the return to 'authentic' folk traditions and their elevation to national status, as with the Finnish regard for the Kalevala and the Estonian for the Kalevipoeg, the Gaelic revival in Ireland, and the Mexican and Peruvian return to pre-Columbian Indian culture; and the glorification of the simple or, common people,' a political tradition that can be traced back to St. Joan, through the narodniks and Mussourgsky's depiction of the Simpleton-Prophet in Boris Godunov and Levin's discovery of truth among the peasants in Tolstoy's Anna Karenina, to Peronism, Poujadism, and the modern Russian nationalists.

Each of these mass phenomena mixes sacred and secular elements, or more accurately, sanctifies the worldly in unexpected, yet, given the background, quite logical ways. On this reading, nationalism is, as Kedourie affirmed, a secular, this-worldly, ideological movement, with clear terrestrial and anthropocentric goals. It does not seek a world beyond this earth, nor does it posit an unseen world guiding and making sense of the world of our senses. At the same time, nationalisms draw, in varying degrees, on the religious background and tradition of the populations whom they deem to constitute nations. This is not, or not usually, a case of manipulation of 'atavistic mass emotion,' as Kedourie claims, nor of a need to control and channel the unleashed energies of the newly mobilized and enfranchised masses, as Hobsbawm argues. It is quite simply that most nationalists are steeped in the religious traditions of their forbears, even where they have rejected them, and that to make sense of their new political message, they have no alternative but to couch it in language and images that are familiar and comforting to their audience. Rarely do nationalists seek to destroy an old, religious identity, in order to build up a totally new national one, although they may attack the guardians of religious tradition who refuse to reinterpret the tenets of the faith in a 'politically relevant,' i.e., nationalist, manner. Hence, of any nationalism we must ask, not whether it is 'religious' or 'secular,' but how and to what extent it mingles sacred and worldly elements, and with what consequences. Similarly, we should enquire not whether the nation in question is seen as an 'ethnic' or 'civic' unit, and a religious or secular community, but how and to what degree its national identity is infused with sacred motifs and religious images, and with what kind of effects.

It may be useful to grasp the complexity of this relationship by distinguishing three 'levels of analysis.' The first is phenomenal: it describes and analyzes the kind of nationalism that is in the public domain, the official version proposed by the elites. Thus, we may use a shorthand to describe French nationalism during the Revolution, or Turkish nationalism under Ataturk, as avowedly 'secular,' even irreligious. And we may range official nationalisms and their political symbolism along a continuum of religious to secular poles. Here, of course, we are using the traditional definition of 'religion' as a faith that looks beyond this world for salvation.

A second level of analysis seeks to determine the content of unofficial and more popular kinds of nationalism and national identity. Here we can expect to find considerable variations in the extent and kind of religious motifs and sacred imagery and symbolism in different national identities and nationalisms, and in different religious traditions. On the one hand, are cases where classical, secular myths and imagery predominate; this was the case in France and the United States during and after their Revolutions, and in Italy from the Risorgimento to fascism. On the other hand, we find examples where religious images and motifs pervade the sense of national identity; this was the case among the Greek peasantry in 1821, the Russian peasantry under Tsarism, or the Ethiopian peasantry under the Emperors, as it is in Pakistan and Sri Lanka to this day. In all these cases, the respective traditional religions had played a large part in molding the sense of ethnic identity prior to the modern epoch; and each of these traditional world religions contained a large repertoire of sacred elements—scriptures,

liturgies, symbols, rituals, scripts, ceremonies, and the like. Some religions, indeed, like Christianity and Judaism, as Adrian Hastings argues, have provided particularly fertile terrain for the reproduction and proliferation of vernacular nations based on the Biblical prototype of the ancient Jews in their land.

Moreover, even in more advanced industrial societies, sacred imagery and religious symbolism may persist and underpin a sense of national identity, even where the majority of the population are, by conventional measurements, no longer practicing believers. In the United States, for example, for all its official civic secularism, large numbers of citizens espouse religious forms, images and rituals of the American nation, and continue to envisage its destiny in providential terms. The same is true, though in quite different ways, in France and Israel, despite the avowed secularism of the majority of their populations; the historic rationale and destiny of the nation remains part of the sacred realm, and the various religious symbols, images and motifs have done their work of forging the *raison d'etre* of the French and Israeli nations.

This brings us to the third level of analysis. It requires us to adopt a wider, Durkheimian approach to and definition of religion, and a different perspective on the significance of the nation. On this general level, the nation can be considered as a 'sacred communion' that unites the dead, the living, and the yet unborn. This is not just a matter of imagination, as Benedict Anderson claims; it is even more a matter of will and feeling. The nation is, in the first place, a willed and a felt community, a moral community of collective purpose whose members are bound by powerful sentiments arising from a sense of presumptive ancestral affinity. (Presumptive, because myths of origin rarely correspond to the historical record of origins, let alone actual biological descent.)

This points to the dual origin of the nation: on the one hand, as a community based on presumed ethnic origins and descent, tracing their genealogy back to a specific place, time, and ancestor, and on the other hand, as a union of all those espousing common moral purposes and assumptions, a community of destiny who trace their ideological and moral lineage to the sacred properties of beliefs and practices of a community of the faithful. The fusion of these two kinds of community over the *longue durée* creates the basis for the modern nation.

What are the sacred properties of the beliefs and practices of the faithful? And how are they transformed into the national 'religion of the people'? Four kinds of sacred property form the main 'pillars' of the nation: belief in the chosen people; attachments to a sacred territory; memories of the golden age; and rituals of commemoration of the glorious dead.

(a) By the concept of the chosen people, I mean not simply any community that exalts itself and carries ethnocentric sentiments of superiority and exclusiveness, such as we find throughout history. The meaning here is more specific. On the one hand, there is a broader sense in which elites of nations feel that they are entrusted with a sacred mission by the deity—to convert the heathen, protect the church, or defeat the infidel. By this divine trust and mission, they are sanctified and, by extension, their people become the elect; they derive their rationale as a community from this trust; and they feel exalted by the sense of their divine calling and the success of their leaders. It is not always clear what the effects of failure are, or indeed what constitutes failure; one has the sense that this kind of chosenness is flexible and capable of considerable reinterpretation.

So indeed is the second, more precise 'covenantal' meaning. Here chosenness is conditional and popular. The whole community is chosen by the deity for a specific purpose, but only for so long as it fulfills its task and mission. Failure to do so leads to the loss of its status as the elect. This is the essence of the 'covenant.' Its purpose is to sanctify the community and to separate its members from their profane surroundings. In the Old Testament model, the Israelites were set apart in order that they might fulfill God's commandments and hasten his universal salvation. They were to become a 'holy people' and a 'nation of priests.' In their footsteps followed Armenians and Orthodox Greeks, Amharic Ethiopians and Orthodox Russians, Afrikaners and Ulster-Scots, and many others in the monotheistic traditions.

The sense of ethnic election has been widespread in all ages and continents, particularly in the looser, missionary form. Its origins are obscure, and it takes many forms, from the conquering zeal of Muslim Arab nomads and Catalan or Hungarian knights to the Celtic-Christian monasticism of the Catholic Irish and the pioneering providential faith of Protestant settlers in America. In each case, chosenness crystallizes the emergent sense of community and confirms its sense of special destiny. As much as any rule of endogamy, it helps to keep the community intact and guard its borders. And, more than any birth rate or territory or conquest, it ensures the long-term survival and persistence of the ethnic community that espouses it. This is especially true of the stricter, more popular, covenantal forms of ethnic election.

The sense of chosenness has carried into the modern world, albeit often in more secular versions. De Gaulle's vision of France's cultural and political mission, the Great Society proclaimed by Johnson for the United States, even the 'New Britain' of Blair and New Labour, echo the missionary kind of election myth. Communist ideals, pared down to 'socialism in one country,' embodied the old myths more shrilly, as did various radical nationalisms in Asia and Africa, from the Meiji reformers in Japan to U Nu's Burma and Nkrumah's African Christianity in Ghana, drawing as they did on Islamic, Hindu, Buddhist, or Christian imagery, but ultimately attempting to fuse

the genealogy of ethnic community with the sacred properties of a moral union, into the sacred communion of the nation.

(b) The origins of ideas of sacred territory are equally obscure, but we may attempt a sociological reconstruction. Territory has been valued not only for security and, of course, as 'hearth and home,' but also as the arena in which the members of a community live, work, and die. And are buried. Reverence for, and protection of, the last resting places of our forefathers and foremothers is one of the main ways in which stretches of territory are invested with special feelings and meaning.

Allied to this is the special collective reverence felt for the tombs and the sites of miracles and exploits of holy men, sages and heroes or heroines. Pilgrimages to shrines like Yasna Gora, Jerusalem, Qom, or Mecca endow these locations with reiterated, and so more lasting, veneration and awe. By extension, any site where heroes or saints displayed the qualities and values central to the traditions of the community is invested with significance and helps to bind the community to its sacred land.

Once again, we are dealing with the confluence of two processes. On the one hand, the sacralization of territory, sketched above; on the other hand, the territorialization of memory. The latter implies two subprocesses: on the one hand, the naturalization of history, turning the community's history into an integral part of its environment and treating it as 'natural,' part of the landscape which it inhabits; on the other hand, the historicization of nature, transforming the community's habitat and landscape into an essential component of its historical development. The effect of these two subprocesses is to bind together community and land into what we may term an 'ethnoscape'—in which the *ethnie* (ethnic community) belongs to, and is an intrinsic part of its landscape, and the landscape belongs to, and becomes an integral part of, its community.

Once ethnoscapes develop, binding together *ethnies* and their poetic landscapes, it is a short step to attaching special sacred qualities to the land. The land is not only 'ours,' it is sanctified as the place of 'our' providential history and the chosen arena for the unfolding of the divine plan through the exploits of His, and our, saints, warriors, and sages. That, of course, is how the Old Testament concept transformed the 'promised' into the 'holy' land of Israel, and how the Swiss shepherds and herdsmen began to interpret the reasons for their remarkable victories over superior Habsburg oppressors, as if the purity of their Alpine mountain fastnesses reflected their simple virtues and steadfast faith in God. Today, ethnoscapes and sanctified nature have often become flashpoints of bitter conflict—in Ulster, Kosovo, the West Bank (Judea and Samaria), Amritsar, and Tamilnad.

(c) The concept of a golden age can only be grasped in the context of the broader 'ethno-history' of the community. This is the set of traditional, often sacred, histories recounted by the members of the ethnic community, as opposed to the more dispassionate attempts by professional historians to reconstruct the history of the community. The line between ethnohistory and ethnic mythology is often blurred, if we interpret 'myth' not as fiction, but simply as a widely-believed tale told about the heroic or sacred past to serve present needs and goals. This is particularly true of those periods and episodes which later generations of the community have come to remember as high or canonical epochs in the unfolding story of the community. These we may call the 'golden ages,' because the memory of these periods is invested with a specific nostalgia and reverence, because they come to serve as models or guides to future action.

There are various kinds of golden age—religious, political, economic, military, artistic—and there may be several in the shared memories of those ethnohistories that are rich in events and relics and are amply documented. Modern Greeks, for example, can look back to the golden age of Athens or to that of imperial Byzantium, Indians to the epoch of the classical post-Vedic states or of the Maurya or Gupta empires, and the English to the kingdom of Alfred and the Anglo-Saxons or to the Elisabethan age. Each of the ages forms part of a wider ethnohistory and its memory can act as the point of reference for a salvation drama of the nation. In such cases, the epoch in question becomes a model, an exemplification of 'our true nature' and a repository of 'our virtues,' spurring the community to emulation. Shared memories of golden ages create a sense of the extraordinary, exalt the community, and confer on their members a sense of collective dignity and the promise of national regeneration.

(d) Among the shrines of the sacred territory, the Tombs of Unknown Soldiers and the monuments to the fallen soldiers and individual warrior-heroes occupy a special place. These are the memorials to the glorious dead. In all ages, from the fallen at Thermopylae or Avarayr to the mass war graves of the two World Wars, the exploits and self-sacrifice of fallen patriot-soldiers and heroes have always evoked a special sombre reverence, and elicited a special determination not to forget their heroism and live up to their example.

In the modern age, the commemoration of the glorious dead has become central to the celebration of the nation. In this respect, the French Revolution with its cult of martyrs like Le Peletier, Bara, and Marat and its commemoration of famous patriots in the Pantheon, *Aux Grands Hommes*, *La Patrie Reconnaissante*, set the tone; while Napoleon's Arc de Triomphe collectivised the memory of the triumphant heroism of La Grande Armée. As George Mosse has documented so vividly, this idea became central to the emergent German nation, with its memorials to ancient heroes like Arminius and to modern heroes

like the Völkerschlachtdenkmal commemorating the fallen at the battle of Leipzig in 1813. In Britain, Lutyens' Cenotaph in Whitehall embodied the grief and pride of the nation; the annual Armistice Day ceremonies there and across the land, to the strains of Elgar and Chopin and the reveille bugle calls, evoke the terrible sacrifices in two World Wars and the collective resolve of the British in an atmosphere of sombre reverence.

There are few modern nations that do not have their memorials to the glorious dead, their ceremonies of commemoration, and their shrines, icons, and symbols of martyrdom and heroism. In this respect, as in the other sacred 'pillars,' nations and their artistic representations have taken over so much of the symbolism and emotion of earlier religious traditions, and infused their this-worldly beliefs in posterity and the judgement of history with earlier sacred longings for collective immortality.

In these ways, nations and their nationalisms have emerged on a religious foundation and are infused with many of their sacred motifs and properties—'religion' being understood here in its wider Durkheimian sense as a collective bond of obligation in respect of sacred things, continually rehearsed in ceremonies and rituals. At the same time, individual historical expressions of a nation's identity may cleave closer to this religious base, or depart from it to embrace a more secular ideal and formulation. However, even in these latter cases, the nation continues to draw upon some of the sacred properties of earlier beliefs and practices and so remains fundamentally a form of sacred communion, uniting its members in a moral and ideological community of intimate belonging and collective moral destiny.

See also: Ethnic Cleansing, History of; Ethnic Conflicts; Ethnicity, Sociology of; Identity in Anthropology; Identity: Social; National Character; Nationalism and Expressive Forms; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: East Asia; Nationalism, Historical Aspects of: South Asia; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Reformation and Confessionalization; Religious Nationalism and the Secular State: Cultural Concerns; Rokkan, Stein (1921–79)

Bibliography

Akenson D H 1992 *God's Peoples: Covenant and Land in South Africa, Israel and Ulster*. Cornell University Press, Ithaca

Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Armstrong J A 1982 *Nations before Nationalism*. University of North Carolina Press, Chapel Hill, NC

Breuilly J 1993 *Nationalism and the State*, 2nd edn. Manchester University Press, Manchester, UK

Connor W 1994 *Ethnonationalism: The Quest for Understanding*. Princeton University Press, Princeton, NJ

Gellner E 1983 *Nations and Nationalism*. Blackwell, Oxford, UK

Hobsbawm E 1990 *Nations and Nationalism since 1780*. Cambridge University Press, Cambridge, UK

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cornell University Press, Ithaca, NY

Juergensmeyer M 1993 *The New Cold War: Religious Nationalism Confronts the Secular State*. University of California Press, Berkeley, CA

Kapferer B 1988 *Legends of People, Myths of State*. Smithsonian Institution Press, Washington, DC

Kedourie E 1960 *Nationalism*. Hutchinson, London

Kedourie E (ed.) 1971 *Nationalism in Asia and Africa*. World Publishing Company, New York

Mosse G L 1975 *The Nationalization of the Masses*. Fertig, New York

Nairn T 1977 *The Break-up of Britain*. New Left Books, London

O'Brien C C 1988 *God-land*. Harvard University Press, Cambridge, MA

Schama S 1989 *Citizens*, 1st edn. Knopf, New York

Smith A D 1986 *The Ethnic Origins of Nations*. Blackwell, Oxford, UK

Smith A D 1991 *National Identity*. University of Nevada Press, Reno, NV

Smith A D 1998 *Nationalism and Modernism*. Routledge, London

A. D. Smith

Religion: Peace, War, and Violence

The relation of religion to peace, war, and violence (or nonviolence) is bound up in the various visions of right relationships and harmony within human beings and between collectivites, including nation-states and ultimately global society. Within Eastern traditions, for example, a Buddhist espousal of nonattachment and rejection of anger implies a peaceable attitude, and a Confucian or Taoist espousal of harmony implies reluctance to engage in depradation and offensive war. Within Hinduism a major strain, at least as interpreted by Gandhians, recommends a nonviolent approach and the kind of exemplary spiritual suasion embodied in *satyagraha*.

In all the Abrahamic traditions there is a positive evaluation of peace, so that 'Peace be with you' and 'Grant us thy peace' are written into the most sacred parts of the Christian liturgy, while *shalom* and *salaam* are peaceable greetings in Judaism and Islam respectively. The Hebrew Scriptures themselves are contradictory. There is a tradition of the trickster deploying deception, such as befits ethnic underdogs, and an ideology of legitimate subjugation. At the same time there is a moral reserve about the taking of blood and a positive evaluation of Israel's helplessness (and

dependence on God). This mutates into Isaiah's vision of universal peace and of the righteous rule of God so central to the teaching of Jesus.

In the teaching of Jesus there is a radical reversal of values which includes a rejection of the negative reciprocity embodied in 'an eye for an eye' and the feud. The Kingdom of God inaugurates a peaceful world empire of love in a harmonious universal city, New Jerusalem, which has its foundation in the self-offering of the Savior as he confronts the powers of violence and injustice. Versions of this approach characterize Christianity in the earliest centuries, resulting in conscientious objection to military action, reinforced by refusal to take an idolatrous oath and expectation of a divine foreclosure of human history.

1. *Christianity and War After Constantine*

However, with Constantine's embrace of Christianity as guarantee of victory, the 'City of God' has to be firmly distinguished from the 'City of Man,' though many Christians came to see the empire as the divinely ordained vehicle of the Christian religion and its triumph. Christianity was now the faith of the majority, including the power holders and senators, not the hope of a minority mostly without access to power. It was implicated in the dynamics of power, even though monks and small groups maintained a devotion to peaceable ideals. Among these small groups were evangelical sects like the Waldensians and dualistic heretics like the Cathars. The church itself iconographically still embodied its original tension with 'the world' of power and violence, set over against icons which celebrate the proximity of the powers that be to the mandates of the Almighty. The church also attempted to adjudicate disputes, and even restrict weaponry, more particularly in the feudal period. It sought to enforce a 'truce of God,' in part to further its own geopolitical ends in confronting the Holy Roman Emperor and reversing the advance of Islam.

The movements of the late Middle Ages, eventually erupting in the Reformation, disseminated the Christian repertoire of themes beyond effective ecclesiastical supervision. There were intimations of collective peaceability in the Lutheran and Calvinist reformations, though these were rapidly checked by the fusion of reformed faith with the autonomy of monarchs and with national sovereignty. There were similar intimations of peace within the natural law traditions of Catholicism. However, the espousal of peace was inevitably a minority option, for example, among Christian humanists such as the Swiss Conrad Grebel, and the anarchopacifist wing of the Radical Reformation, for example, among Anabaptists and Quakers. These minorities came to represent a sectarian encapsulation of Christian radicalism over a wide range of issues, including corporal and capital punishment, which leaked into the wider society over the next two or three centuries. It was, for example, chiefly among such groups that supporters emerged for the nascent peace societies of the UK and the USA in the early nineteenth century.

2. *Proposals for Universal Peace*

Ideas of universal peace burgeoned from the fourteenth century on, partly on the model of an imperial Pax Romana, centered on this or that universal monarchy, but also based on projects for an international order. These latter picked up the Christian humanism of Erasmus in his *Complaint of Peace* (1517) and the notions of natural law articulated by Grotius in his *De jure belli ac pacis* (1625). One of the most complex of these projects was the Abbé Saint-Pierre's *Projet pour rendre la Paix perpetuelle* (1713). The last two great peace plans of this period were Kant's *Perpetual Peace* (1795) and Bentham's *Plan for a Universal and Perpetual Peace* (1789).

Here one needs to distinguish types of approach and indicate a partial secularization of the religious vision of peaceable social and international order. A notable secularization was embodied in Machiavelli's statement of Renaissance political realism in *The Prince*, while from rather different viewpoints, Kant fused rational with Christian motifs and Bentham appealed to criteria of human happiness. This mixture of contradictory religious and secular discourses developed into the liberal internationalism of the nineteenth century. According to this tradition, especially as developed in the USA and the UK, there was a potential harmony of interests through the reciprocities established by free trade.

A parallel fusion of religious and political discourses also occurred in the democratic socialist tradition, by contrast with the Marxist variant predominant on the European continent which consigned religion to a realm of fantastic hopes devoid of scientific understanding. For Marxist-inspired movements, as also for anarchosyndicalism, violence was a midwife of a new revolutionary order. Out of the currents of liberalism and socialism emerged hopes of universal human and/or working-class solidarity which were articulated in theories of the social sources of war and the structural alterations needed for peace. However, the wars of 1914–18 and 1939–45 suggested these hopes were chimerical and the theories at best flawed, though capitalism was still widely blamed for international disorder and imperialism.

Out of this iron period grew a new realism about the role of national interest in the genesis of war, which was articulated both in religious and in secular form. The major realist theologian was Reinhold Niebuhr in his *Moral Man and Immoral Society* (1934) and the major exponent of secular realism was the American

Hans Morgenthau. At the same time there also emerged fully fledged empirical enquiries into conflict, including 'peace research.'

3. Forms of the Peace Sentiment

Thus over the last two centuries or so the 'peace sentiment' has been part of overlapping causes of a humanitarian and internationalist kind, simultaneously fed by religious and enlightenment sources. There causes have conjoined the gospels to reason, commonsense, and the quest for human rights in a radical critique of war and militarism. The nineteenth century sociologist Herbert Spencer, for example, had roots both in rational and evangelical nonconformity, and identified the alliance of state, priest, and military as a major nexus of conflict. A leftward leaning liberalism attacked the nation state and established religion in the name of a pacific internationalism. This was often referred to as 'pacificism,' meaning it presumed the onus of proof was on the proposers of war and held that the interests of those who stood to gain, for example, the armaments manufacturers, ran contrary to the interests of the majority. Liberal internationalists were not strict 'pacifists' however, because there were times when one had to crusade for liberal principles and for liberation—including that other liberal principle, national liberation.

Other humane causes loosely associated with the peace sentiment were reforms of penal practice, above all corporal and capital punishment and a new attitude towards the animal world, which might include vegetarianism, with its objection to meat, machismo, and blood. To these causes might be conjoined a preference for wholesome food to nourish wholesome bodies and lifestyles. All of these had religious roots, even though one cause need not strictly imply another, so that cornflakes were an outcrop of Adventist teaching without requiring pacifism, whereas the Quaker production of chocolate was clearly linked to it by association. Humanitarians opposed all kinds of brutalization, including male violence in the home and the army. A feminization of the psyche occurred, bred in the Sunday School and in progressive education alongside movements for factory reform and the provision of life-enhancing leisure pursuits, preferably in the context of 'nature.'

The concept of individual conscience took off from religious prescriptions like personal conversion and witness and then mutated into a new sensitivity, including sensitivity towards the effects on other peoples of conquest, imperialism, and exploitation. Thus alongside the nationalistic fervor infiltrating and informing much mass religiosity there ran a counterpoint of moral and political conscientiousness. There were movements for democratic scrutiny of foreign policy and the sale of armaments, for the arbitration of international disputes, the enforcement of international law, and eventually the League of Nations and the United Nations. Alongside these ran mostly nonviolent agitations for civil rights and/or national liberation such as those led by Ghandi, Martin Luther King, Nelson Mandela, and Archbishop Tutu.

Of course, given the schooling of populations in patriotic sentiment, dissenting attitudes and conscientiousness usually characterized only a minority in the churches and even more so among people at large. This was above all true in wartime, when dissent from all war or from the claimed objectives of a particular war, could be defined as cowardice or disobedience to lawful authority. Yet conscientious objection still made headway and, at least in the countries of the North Atlantic, Protestant world governments reluctantly accepted that national claims to loyalty were not absolute. This partial tolerance applied to the 'historic peace churches,' such as the Quakers and Mennonites, but also included sizeable groups within the larger Protestant denominations, and it just about extended to secular objections rooted in socialism or humanitarian ideals such as were promoted by Leo Tolstoy.

The social ecology of the 'peace sentiment' over the last quarter millennium is clear enough. It began in the transatlantic world through a combination of religious and rational liberal motifs—Manchester allied to Boston—above all among Protestants, and by degrees affected established churches and the Roman Catholic Church. Given that the Roman Catholic Church was rooted in more organic and communal forms of society, especially in Latin Europe, its defence of the rights of individual conscience was originally weaker. Yet its transnational character always included a sense of the law of nations as nourished by the traditional concept of a just war, such as due proportion of means to ends.

With the development of weapons of mass destruction most mainstream churches expressed reservations about the doctrine of deterrence as it related to the 'just war.' On the one hand there were religious groups allied to secular middle-class radicalism in campaigns against nuclear armament and colonial involvements such as Vietnam, and on the other hand there were specialist commissions into the contemporary nature of war and peacemaking. Thus the Church of England produced a report *The Church and the Bomb* (1982) and the Roman Catholic Church in the USA published an extended critique in 1983.

In summary one may say that there has always been a dialectic relation between the idea of an all-embracing empire (the Roman Pax, Chinese imperial civilization) and the vision of a transnational unity, such as the Islamic 'Ummah' or the Christian *Oekumene*. In times of religious establishment, whether imperial, monarchical, or national, the former will inevitably tend to colonize the latter, so that the wider vision is maintained underground in liturgy, iconography, and movements of protest. However, as

religious groups became separate from the state the visionary repertoire gains more purchase, and is increasingly reinforced by the interdependence of global society. The resources on which the search for peace may draw become increasingly international, so that the message of the gospels, of Isaiah, and Micah mingles with Eastern and especially Gandhian, notions of nonviolent protest and moral suasion. Indeed, as Western protest has drawn on Eastern resources, so have Eastern movements been cross-fertilized by Western ones, not only Christianity but the works of Tolstoy and Thoreau.

See also: Appeasement: Political; Conscription, Policy of; Ethnic Cleansing, History of; Ethnic Conflicts; Genocide: Anthropological Aspects; Genocide: Historical Aspects; Internal Warfare: Civil War, Insurgency, and Regional Conflict; Liberalism and War; Peace; Peace and Nonviolence: Anthropological Aspects; Peace Movements; Political Protest and Civil Disobedience; Religion, History of; Religion: Nationalism and Identity; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State: Cultural Concerns; War: Causes and Patterns

Bibliography

Bainton R H 1960 *Christian Attitudes Toward War and Peace*. Hodder and Stoughton, London

Beales A C F 1931 *The History of Peace*. G. Bell, London

Brock P , Socknat T P (eds.) 1999 *Challenge to Mars: Essays on Pacifism from 1918 to 1945*. University of Toronto Press, Toronto

Burns J P (ed.) 1996 *War and its Discontents: Pacifism and Quietism in the Abrahamic Traditions*. Georgetown University Press, Washington, DC

Ceadel M 1987 *Thinking About Peace and War*. Oxford University Press, Oxford, UK

Haleem H A (ed.) 1999 *The Crescent and the Cross*. Macmillan, London

Martin D 1997 *Does Christianity Cause War?* Oxford University Press, Oxford, UK

Mayer P (ed.) 1966 *The Pacifist Conscience*. Rupert Hart-Davis, London

Niditch S 1993 *War in the Hebrew Bible*. Oxford University Press, Oxford, UK

Sibley M Q, Jacob P E 1952 *Conscription of Conscience: The American State and the Conscientious Objector, 1940–1947*. Cornell University Press, Ithaca, NY

D. Martin

Religion, Phenomenology of

Phenomenology of religion is an academic approach to analyze religion predominantly within religious studies. It is documented in a series of monographs and handbooks, but it is also defined by a 'scholarly method' that utilizes principles of phenomenological philosophy. For the purpose of this paper, we shall distinguish: (a) the descriptive phenomenology of religion which refers to the classification and systematization of religious phenomena and the creation of typologies which account for different types of religion; and (b) the analytical phenomenology of religion which, in addition to the goals of descriptive phenomenology, is based on some explicit understanding of the philosophical background and methods of phenomenology and hermeneutics developed in the tradition of Edmund Husserl, Wilhelm Dilthey, Martin Heidegger, and Paul Ricoeur.

1. Historical Development, Institutions, and Representants

Whereas the notion of phenomenology had been proposed by philosophers also preoccupied with religion, such as J. H. Lambert (1721–71), I. Kant (1724–1804), and G. W. F. Hegel (1770–1831), the contemporary notion of phenomenology of religion refers to a specialized field within the academic study of religion. The origins of this field go back to the Belgian historian of religion P. D. Chantepie de La Saussaye (1848–1920). Referring to Hegel's work (which includes the *Phenomenology of Mind*), he divided the science of religion into the history of religion and the philosophy of religion. With respect to the latter, he wanted the phenomenology of religion to not only avoid dogmatics, but to systematically order the main groups of religious phenomena according to the feature of the historical material. The most decisive breakthrough in the field has been due to the work of the Dutch professor of religious history, Gerardus Van der Leeuw who, in 1924, suggested phenomenology of religion to be a special method for the history of religion. Since then, the phenomenology of religion has produced, on the one hand, a series of 'monumental' contributions which provide systematic overviews on a broad variety of religions, or, on the other hand, monographs on typically basic features of single religious phenomena. Therefore, topics of the studies include general issues, such as the distinction between the 'sacred,' and the 'profane,' the 'numinous' or the experience of power, and on the other hand comparative studies on more specific religious phenomena, such as prayer, piety, or religious ceremonies. In the 1950s, phenomenology of religion became increasingly established in the academic studies of the history of religions, so that it became institutionalized as a specialized branch within *Religionswissenschaft*, Comparative Religion or Religious Studies. Although one should not ignore Britain (James 1938) and France (the Rumanian scholar's Eliade's (1949) first important book has been published in French), the main centers of development have been The Netherlands,

Germany, Scandinavia, Italy, and North America (cf. Pettersson and Åkerberg 1981).

2. Approach and Method

Phenomenology of religion differs from theology as well as from other specialized disciplines preoccupied with religion (philosophy, sociology, psychology, etc.) by its history and its method. Historically, phenomenology of religion opposed positivist and evolutionist approaches to religion by integrating historical knowledge of the facts with phenomenological methods. Whereas the sociology of religion, which developed almost simultaneously, focused on the external, public aspects of religion, the phenomenology of religion turned towards the essential aspects and started systematically to classify religious phenomena.

2.1 Descriptive Phenomenology of Religion

Phenomenology of religion draws on the impressive amount of data collected by the history of religion, but also on data collected by sociology, anthropology or the psychology of religion. Therefore, it can be said to be empirical. The basic method of descriptive phenomenology is the comparison of religious phenomena. It assumes that religious phenomena are comparable across different cultures and religions, and that the comparison does not only yield generalizable features but allows us to grasp the general features of religion. In analyzing the empirical forms of historical religions, phenomenology of religion tries to abstract from specific religions, and to then develop ideal types of religious phenomena, such as 'sacrifice,' 'fetish' or 'mysticism,' 'prophecy,' 'priest,' or 'cult.' Therefore, it can be said to be systematical. In order to pursue this task, phenomenology is considered to provide the method for finding the common denominators covering the diversity of empirical data. Religious phenomena include manifestations of the sacred, such as sacred places, mountains, buildings, active realizations of the sacred, such as rituals, prayers, holy words, religious beliefs, such as relevation, redemption, transcendence, and experiences, such as vision, ecstasy, or miracles.

The data on these phenomena are not to be observed from an external point of view, as in most sciences of religion. On the contrary, phenomenology of religion is nonreductionistic in identifying the specifically religious features of religious data. This is the reason for turning to philosophical phenomenology that is expected to analyze the meaning of religion to the person. At the basis of phenomenology of religion lies the analysis of religious experience rather than religious ideas, the beliefs, or morals. The accentuation of the lived experience may be one of the most important contributions of phenomenology to the study of religion.

Basically, religion is the opposite of phenomena: religion posits what it believes to be as existant, whereas phenomenology brackets exactly this belief in order to investigate how it may be constituted by consciousness. Thinking of others, in religion (or parapsychology), gets close to telepathy (think of the dead), in phenomenology, the others are but constructions. Almost simultaneously, J. W. Hauer developed a phenomenology of religion mainly on the basis of data on illiterate cultures. Among the major contributors to phenomenology of religion one has to mention Allen, Bleeker, Bollnow, Hultkrantz, Kristensen, Lanszkowski, Pettazzoni, Wach, Waardenburg, Widengren, etc.

2.2 Phenomenological Methods

Whereas the descriptive and typological branch of the phenomenology of religion remains silent about the methods by which it arrives at results, the phenomenological branch is characterized by additional methodological considerations that are based in philosophical phenomenology.

From the perspective of phenomenological philosophy, experiences are the processes by which consciousness is constituted. As consciousness is always intentional, i.e., consciousness of something, experiences, too, are characterized by intentionality, i.e., the fact that they are related to something else. Also religious experience is experience of something. To give an example: as Chantepie de La Saussaye has already stressed, it is only the meaning of the act that allows us to distinguish between the sacrificial priest and the butcher.

In order to arrive at the meaning of religious experience, Husserl recommended to perform what he called '*epoché*': the bracketing or suspension of all our unexamined assumptions, preconceptions, and prejudices about the phenomena under consideration. In performing the epoché, analysts are expected to turn 'to the things themselves,' i.e., to the phenomena as they are experienced. That is to say that phenomena, such as visions, auditions, or miracles, are to be understood according to the ('religious') meanings they have about the experiencing consciousness and not according to, e.g., a neutral scientist's observations. As an additional method, religious phenomena, such as 'snake symbolism,' can be made subject to 'free variation': by mentally varying aspects of this symbol, one attempts to arrive at the essential features of the '*Wesensschau*.' On the basis of these methodological operations, descriptions of the phenomena are considered to yield typical structures, essential meanings or even the 'eidos' of religious phenomenon.

Attempts to understand the essential meaning of religious phenomena, finally, presuppose hermeneutics, i.e., the method of understanding ('*Verstehen*')

the subjects of religious experience. According to Van der Leeuw (1948), understanding is achieved by empathy: the investigator tries almost to re-experiences the religious phenomenon in order to describe these experiences. Understanding, moreover, includes reflecting the context of the interpreters, accounting for the interpretive framework, and explicating their pre-interpretations. To some representants of the phenomenology of religion, such as Eliade, hermeneutics has become a leading principle of their method that allows us to unveil the deeper structure of the experiences of mankind. On these grounds, religious studies have also been equated to a hermeneutic anthropology.

Basically, one can distinguish three related approaches: a descriptive school which is preoccupied with the systematization of religious phenomena; the typological school which tries to distinguish different types of religion, and the phenomenological school in the specific sense of the word that applies methods borrowed from philosophical phenomenology.

3. *Criticism*

From its beginnings there has been a series of critical objections against the phenomenology of religion.

(a) Especially the phenomenological approach of Eliade is the target of Wagner's (1986) criticism who objects that his concept of the *homo religiosus* is guided by a notion of 'natural religion.' He argues that it presupposes some unwarranted knowledge about the religious situation in what Eliade considers to be 'archaic cultures.'

(b) With respect to religion, phenomenology of religion takes a decidedly substantialist position (Luckmann 1983). Moreover, critics charge that this substantialism is based on nonempirical, extra-phenomenological, and theological assumptions and intentions enter into the analyses of such classic representants of phenomenology of religion such as Kristensen, Eliade, or Van der Leeuw.

(c) To many critiques, this is due to the fact that methods are rarely unveiled. Despite the reference to phenomenological methods, the basic presuppositions often seem arbitrary, and also theoretical reflections are criticized to fall short of the diligent collection and classification of data.

(d) Since phenomenology provides the basic methodology for phenomenology of religion, it is particularly consequential that phenomenology of religion lost contact with the developments in philosophical phenomenology from about the 1950s. As a result, phenomenology of religion rarely is considered of importance in the analysis of religion within the tradition of phenomenological philosophy (Guerrière 1990).

(e) The phenomenology of religion has been criticized for ignoring the social and cultural contexts of religious phenomena. Moreover, as phenomenology in general has been criticized for its naive attitude towards language and cultural perspectivism, also the phenomenology of religion is subject to criticism as to the linguistic and cultural bias implicit in the analysis of 'phenomena' and 'symbols.' Thus, the results of free variations depend on the phenomenologists' cultural background (Allen 1987).

4. *Recent Developments*

As a reaction to these criticisms, there has developed what Waardenburg calls the new style of the phenomenology of religion (Waardenburg 1973–1974). According to Waardenburg, this new style of investigation shares a series of assumptions: (a) religious phenomena are characterised by a variety of meanings and intentionalities which are the result of different, often materially based perspectives on the world; (b) religions are coherent, but essentially open systems of signs and symbols, the elements of which are selected according to the corresponding relevance systems; (c) religious phenomena are considered to be human constructions which are subject to social processes (as, e.g., traditionalization, institutionalization, or canonization). At this point, the phenomenology of religion overlaps with positions which have been developed within those branches of the sociology of religion rooted in the phenomenology of Edmund Husserl and Alfred Schutz (Berger and Luckmann 1963). This 'phenomenologically orientated sociology' regards religious experiences as, first, based on the profane transcendencies of everyday life, and second, subject to processes of the social construction of multiple realities. By focusing on the communicative processes by which this social construction is accomplished, this approach tries to bridge the gap between empirical research and methodological reflections on this research. Moreover, in admitting that linguistic and cultural background knowledge enters into the phenomenologist's background knowledge, several of the new approaches to the phenomenology of religion accept an orientation towards a 'reflexive phenomenology.'

See also: Hermeneutics, History of; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Religion: Definition and Explanation; Religion, Sociology of; Religiosity, Sociology of; Schtz, Alfred (1899–1959); Semiotics

Bibliography

Allen D 1978 *Structure and Creativity in Religion. Hermeneutics in Mircea Eliade's Phenomenology and New Directions.* Mouton, The Hague, Paris, New York

Allen D 1987 Phenomenology of religion. In: Eliade M (ed.) *The Encyclopedia of Religion*. Macmillan, New York, Vol. 11, pp. 272–85
Berger P L, Luckmann T 1963 Sociology of religion and sociology of knowledge. *Sociology and Social Research* **47**: 61–73
Bianchi U, Bleeker C, Bausani J A 1972 *Problems and Methods of the History of Religions*. Brill Leiden, Germany
Bleeker C J 1972 The contribution of the phenomenology of religion to the study of the history of religions. In: Bianchi U, Bleeker C J (eds.) *Problems and Methods in the History of Religions*, Brill, Leider, pp. 39–41, 51
Colpe C 1988 Zur Neubegründung einer Phänomenologie der Religionen und der Religion. In: Zinser H (ed.) *Religionswissenschaft. Eine Einführung*. Reimer, Berlin, pp. 131–54
Eliade M 1949 *Traité d'histoire des religions*. Paris
Eliade M 1958 *Patterns in Comparative Religion*. Sheed & Ward, New York
Guerrière D (ed.) 1990 *Phenomenology of the Truth Proper to Religion*. State University of New York Press, Albany, NY
Hauer J W 1923 *Die Religionen, ihr Sinn, ihre Wahrheit*. Berlin, Stuttgart, Leipzig, Vol. 1
Heiler F 1932 *Prayer: A Study in the History and Psychology of Religion*. Oxford University Press, London
Heiler F 1961 Erscheinungsformen und wesen der religion. Stuttgart 1961. Religionsphänomenologie. In: Müller G et al. (eds.) *Theologische Realenzyklopädie*. De Gruyter, Berlin and New York, Vol. XXVIII
James E O 1938 *Comparative Religion: An Introductory and Historical Study*. Methuen, London
Honko L (ed.) 1979 *Science of Religion. Studies in Methodology*. Mouton, The Hague
Kristensen W B 1960 *The Meaning of Religion: Lectures in the Phenomenology of Religion*. Mouton, The Hague
Luckmann T 1983 Theories of religion and social change. In: Luckmann T (ed.) *Life-World and Social Realities*. Heinemann Educational Books, London, pp. 133–61
Pettersson O, Åkerberg H 1981 *Interpreting Religious Phenomena. Studies with Reference to the Phenomenology of Religion*. Almqvist & Wiksell International, Stockholm, Sweden
Saussaye de la P D 1887 *Chantepie, Lehrbuch der Religionsgeschichte*. Muhr, Tübingen, Germany
Van der Leeuw G 1948 *Inleiding tot de Phaenomenologie van den Godsdienst*, 2nd edn. Srven F. Bohn, Haarlem
Waardenburg J J 1973–1974 *Classical Approaches to the Study of Religion: Aims, Methods and Theories of Research*, 2 vols. Mouton, The Hague/Paris
Wagner F 1986 *Was ist Religion? Studien zu ihrem Begriff und Thema in Geschichte und Gegenwart*. Mohn, Gütersloh

H. Knoblauch

Religion: Psychiatric Aspects

1. Terms

According to *Webster's New World Dictionary*, *psychiatry* is defined as 'the branch of medicine concerned with the study, treatment, and prevention of disorders of the mind, including psychoses and neuroses, emotional maladjustments, etc.' (p. 1147). In contrast, *psychology* is defined as 'the science dealing with the mind and emotional processes' and 'the sum of the actions, traits, attitudes, thoughts, mental states, etc. of a person or group' (p. 1147). From these definitions, it is clear that the realms of both these fields have considerable overlap in the questions they address about human behavior, and in many cases, knowledge that is garnered in one field is frequently shared and absorbed into the ensuing theoretical development of the other discipline. Hence, for the purposes of this article, which addresses the theoretical underpinnings rather than the actual practices of each profession, the ideas from psychology and psychiatry will be used interchangeably, except where otherwise denoted as specific to their field.

2. The Question

The question about the need for relationship between psychiatry and religion has emerged as a predominantly twentieth-century concern. The movement to inter-relate the two disciplines evolved as participants who believed strongly in the importance of religion in people's lives were concerned by the lack of religious involvement in modern medicine's approach to healing. Many adherents to this movement to intertwine the two fields reject the dualism of body and soul, believing instead that mind and body are inter-related and true healing is best realized when a holistic bio-psycho-social-spiritual approach is utilized.

As such, in recent years a growing number of psychiatrists and other physicians, psychologists, clergy, theologians, and religious philosophers who have not been satisfied with mainstream traditional and institutionalized efforts at healing have sought to combine the best of modern psychiatric medicine with religion/spirituality to develop an alternate method of healing to relieve mental and emotional distress. These practitioners are joined by religiously committed persons who are seeking treatment. At the same time, a number of traditionalists, both psychiatrists and clergy, are critical of incorporating religion and psychiatry into a single treatment mode. They raise the question whether religion and psychiatry should both be considered in aiding individuals, advocating rather that the two remain separate in their approaches to problem solving and actual practice.

3. Defining the Problem

In his book *The Structure of Scientific Revolutions* (1970), Thomas Kuhn defines a paradigm as a nascent model for looking at the universe in a specific manner, and answering questions from within that particular framework's perspective. Furthermore, according to

Kuhn, for growth to occur in science and/or ideology, rather than working toward consensus between paradigms, the premise for philosophical development should be that growth comes from complementary coexistence between paradigms, as each domain converses with the other within its specific applications.

It is arguable that religion and science are two different paradigms for understanding the universe, and humanity's role within it. The difficult and sometimes troublesome relationship between these two fields emerges in part as a result of the two different paradigmatic approaches to answering the same basic questions about humankind. Such questions include, but are not limited to these: what are the fundamental components of human beings? What causes illness? What helps keep individuals healthy?

Furthermore, the more recent developments of specialized disciplines such as psychiatry and psychology in the nineteenth and twentieth centuries also raised alternate perspectives on answering the questions to humanity's interpersonal troubles—answers different from those raised in their antecedent jurisdictions of religion and philosophy (Rechberger 1998).

The question raised is: do these alternate paradigms have communal merit in helping individuals lead happier and healthier lives, or should the division between the two fields of mental health and religion remain distinctly separate?

4. *History*

Since the 1700s, the period known as the Enlightenment, religion and science have experienced increasing differentiation and tension. Prior to this point in history, religion and science were interchangeable ideologies. In fact, as Murphy (1997) states, 'In the Middle Ages, theology was not only *a* science but also the queen of the sciences' (p. 155, italics in the original). Indeed, under the church's rule, when sixteenth and seventeenth century scientists such as Giordano Bruno, Tycho Brahe, and Galileo Gallilei offered data for understanding the universe different from the understood Biblical theology of the day, they were declared heretics, punished, and a few were even executed. However, the mathematical evidence presented by scientists such as Copernicus and Galileo was repeatedly validated and supported despite dispute from the church. Hence, for many, science became the primary source of information about the universe and its inhabitants. Furthermore, due to the infamy surrounding the church's declaration of heresy upon scientists such as Bruno and Galileo, religion was at least somewhat discredited as a viable source of answers to humanity's problems (Rechberger 1998).

Nevertheless, it is impossible to discount the effect religion has had on humanity throughout history. From the earliest times, religious beliefs and rituals have been a central component of the lives of humankind, frequently providing the structural framework from which societal, cultural, and moral norms developed. In the Western hemisphere, the primary religions of Judaism, Christianity, and Islam have as a core belief a universal God who acts as creator, sustainer, and protector. In the Eastern hemisphere, the predominant religions of Buddhism and Hinduism depict God as less actively involved in the day-to-day workings of humanity; nevertheless, almost all world religions have similar beliefs and practices in a supreme being or group of beings that are in some way interested in humanity's development. Furthermore, throughout most of known history, it was the religious leaders—whether priest or shaman—who were considered healers (a tradition which continues today in less Westernized cultures). Indeed, prior to the 1800s, what scant treatment there was for mental illness was provided by the church. Unfortunately, one of the negative outcomes of the overlapping domains of religion and medicine was evidenced in the development of concepts such as demon possession and illness as a result of, if not punishment for, an individual's misdeeds. These explanations were broadly accepted as disease etiology in most Western nations well into the nineteenth century.

In contrast, however, one of the potential benefits of having the same person exercising both religious and healing functions allows for the possibility of positive coordination and integration between the two domains. Reformers such as Philippe Pinel (1745–1826) in France began to see the mentally ill as sick rather than sinful or wicked. Another reformer, William Tuke (1732–1822), opened the York Retreat in England in 1792, and other pious Quakers were early leaders in the care and treatment of the mentally ill.

Beginning in the latter part of the eighteenth century medicine was developing as an independent science dealing with the maintenance of health and the prevention, alleviation, and cure of disease. This was a clear departure from the intertwining of religious practice and medical intervention that had existed from earliest human history. Until this time it was the church that regulated medical practice by the very fact that doctors needed license from the bishop to practice medicine (Foskett, 1996).

Despite this emerging demarcation between religion and medicine as separate disciplines, some religious leaders continued an interest in maintaining the overlapping interests of religion and medical practice. This was particularly notable in the United States during the mid-nineteenth century, during a *zeitgeist* that advanced an increasing emphasis on the mind–body relationship. A strong—albeit minority—view of the positive benefits of intermingling religious belief and medicine began to develop. Many wrote about it, including Mary Baker Eddy (1821–1910), the founder

of Christian Science, and Ellen G. White (1827–1915), one of the founders of the Seventh-day Adventist church.

Eddy wrote, 'If we understood the control of mind over body, we should put no faith in material means. Science not only reveals the origin of all disease as mental, but it also declares that all disease is cured by divine mind' (1875, p. 169). Nearly 100 years prior to psychiatry's recognition of the influence of brain chemistry on disease, Ellen G. White wrote that 'the mind affects health to a far greater degree than many realize and many diseases from which men suffer are the result of anxiety, remorse, guilt and mental depression' (1905, p. 241). These religious writers preceded Herbert Benson and Bernie S. Siegel's twentieth century emphasis on the mind–body connection.

5. *The Emergence of Mental Health Disciplines*

While the trend of separating science and religion was evidencing itself by the growing separation between medical practice and religious ideology, the emergence of psychology and psychiatry as new scientific disciplines also demonstrated a distinct separation from their religious and philosophical roots. Examining the writings of the earliest primogenitors within these newly emerging disciplines clearly reveals the tension between science and religion as they developed differing paradigms to resolve questions about what the components of 'good' mental health comprise.

In the United States, it was pre-Freudian William James (1842–1910) who was at the forefront development of the profession of psychology, along with his Harvard colleague Hugo Munsterberg. Interestingly, James (see article on *James*, *William* (1842–1910)) refused to treat physiology, psychology, and philosophy as distinct and separate disciplines. Furthermore, his endeavor to maintain a balanced position between these domains is clearly reflected in his 1902 book, *The Varieties of Religious Experience*, which addresses the questions related to religion as it interacts with mental health. As James wrote:

> The problem I have set myself is a hard one: *first*, to defend … 'experience' against 'philosophy' as being the real backbone of the world's religious life … and *second*, to make the hearer or reader believe what I myself invincibly do believe, that although all the special manifestations of religion may have been absurd (I mean its creeds and theories) yet the life of it as a whole is mankind's most important function' (1902, p. xix, italics in the original).

Although James acknowledged religion as a social and cultural phenomenon, he was interested primarily in the individual response of a person to religious experience. More specifically, he expressed fascination with the conversion process, the moral and mystical elements of the religious life, and the interaction between religious belief and illness—particularly healthy vs. pathological mental health responses to religion. James (see article on *James*, *William* (1842–1910)) espoused a favorable view of religion, based on his perception that 'healthy-minded' individuals respond positively to religious beliefs and are the happier for it. Yet this positive view of religion in the scientific community would shortly be denounced by the writings of psychiatrist Sigmund Freud (1856–1939).

The field of psychiatry, although historically a part of medicine, was marginalized as a profession until the appearance of Freud's writings (Alexander and Selesnick 1966). His ideas about the dynamic structure of personality and the benefits of psychoanalysis revolutionized the field of psychiatry. In his radical new perspectives about the inner processes of individuals, Freud assessed humanity's desire for God and the function of religion in everyday life as evidence of psychopathology.

In his book *The Future of an Illusion*, Freud (see article on *Freud*, *Sigmund* (1856–1939)) opined that faith is based on the illusion that an idealized father (God) can replace the lost, real father who has proven fallible to the child and is no longer a comfort and source of security. Freud's solution to his depiction of this 'infantile' response on the part of individuals to create a 'mythological' God was for them to develop a realistic view of the reality that there is no God: 'Man cannot remain a child forever; he must venture at last into the hostile world. This may be called '*education to reality*' (1934, p. 86, italics in the original).

Although the idea that God is a projected illusion did not originate with Freud, his declaration that religious belief is an infantile projective fantasy cast a long shadow over subsequent practitioners' theories about the relationship between religious belief and mental health. This perspective contributed to many decades in which psychiatry generally viewed religious belief with skepticism and antipathy. What little credit Freud gave to religion for 'restraining the asocial instincts' of humankind (p. 65), he reiterated that it had not done so successfully. Instead, Freud advocated that humans should place their trust in the rigors of the scientific method, out of which discoveries of the truth about the nature of individuals and humanity's relationship to the world around it would emerge.

This review of the origins of psychology and psychiatry illuminates the tension between the positivist assumption—that the scientific method and rigors of research would eventually answer all of humanity's questions about physical and emotional well being—and the alternate view that religious belief is a constructive addition to solving mental health concerns. Although some theorists continued to typify any religious belief as evidence of naivete or pathology, others dissented from Freud's strict atheism and

instead posited a more favorable view of religious beliefs in their psychological theories.

Examples of some of the psychologists and psychiatrists who integrated religion with their psychological theories would include such notables as Carl Jung (1875–1961), Erich Fromm (1900–80), and Viktor Frankl (1905–97). The broadest incorporation of the spiritual with the psychiatric is found in Jung, a Swiss psychiatrist and contemporary of Freud. His worldwide travels, interest in the occult, tribal religion, and other non-mainstream mystical beliefs heavily influenced his subsequent psychological theory. In stark contrast to Freud, Jung's liberal premise that all psychic products—including visions, dreams, and hallucinations—are 'facts' and should be considered as having the same basis in reality as other physical facts, opened wide the door to religious belief and behavior in psychic functioning that Freud (see article on *Freud*, *Sigmund* (1856–1939)) had firmly attempted to close. Furthermore, according to Jung, it is not the business of psychology or psychiatry to prove or disprove the existence of God: it can only address the psychic effects of the God image and its mythic antecedents on human behavior.

Psychoanalyst Erich Fromm recognized the value of religion in providing healthy feelings of stability and security for individuals. In his book *Psychoanalysis and Religion*, Fromm stated, 'Man is not free to choose between having or not having ideals but he is free to choose between different kinds of ideals, between being devoted to the worship of power and destruction and being devoted to reason and love' (1950, p. 25). Fromm further advocated a surrendering of one's individual goals and identity in favor of living selflessly in community with others, a not uncommon ideal of both Judaic and Christian rhetoric.

Finally, Viktor Frankl, the Holocaust survivor and psychiatrist who wrote *Man's Search for Meaning*, articulated in his book the premise that all humans are on an unmitigated quest for meaning—even in the midst of horror and despair. Frankl clearly credited his spiritual belief as sustaining his mental health throughout his imprisonment, and contrasted his experience to other prisoners who lost their faith and subsequently their will to live, succumbed to illness, and died. Of the three highlighted theorists who have acknowledged the positive merits of religious belief on mental health, Frankl's testimony is the most graphic and recent representation of the positive outcome of melding the two.

6. *Theology*

It is not only within the realm of science that the questions about the potential benefits of intertwining psychiatry and religious belief have been addressed. A number of theologians, religious philosophers, and clergy have also debated the merits of incorporating religious beliefs into the best scientific practices in mental health.

One of the earliest proponents of such an idea was Anton Boisen (1876–1965), a Protestant clergyman who had once experienced a psychotic illness and subsequently had been institutionalized in a mental hospital. He reported that the hospitalization was a religious renewal experience for him, and he believed that it could be that way for some other patients if the psychiatric community were sufficiently sensitive to the meaning of these experiences.

Another theologian, Paul Tillich (1886–1965), proffered a broader anthropological and philosophical context within which psychiatry could exist and still address its own, more specific concerns. Similar to Jung's placing limits on psychiatry's role in evaluating religious belief, Tillich opined that statements of ultimate concern belong to a domain beyond which psychoanalysis and the scientific method can neither disprove nor verify. Furthermore, Tillich reiterated that when psychiatry makes either negative or positive judgments about the reality of religious ideas and symbols, it has moved out of the realm of science and into the realm of metaphysics.

Current philosopher of religion Paul Ricoeur (1913–) has countered the reductive view of religion espoused by Freud, and the influence Freud's views have had over much of psychiatry and the wider culture shaped by these theories. In contrast to Freud's view that religion is basically a public expression of obsessive compulsive neurosis, Ricouer espouses belief that humans grow when confronted by sacred figures who challenge them to become more than they are.

Other contemporary writers include Karl Menninger (1893–1990), who has written *Whatever Became of Sin?*, addressing the atheistic stance of psychiatry and the cultural disregard for personal responsibility. Another, Paul Pruyser (1916–87), writes about the need for psychiatry, with its empirically-based findings, and religion with its faith basis, to be utilized separately rather than attempt their integration in understanding and treating mental illness.

7. *Practices Which Have Attempted to Merge Religion and Psychology*

The concern by some psychiatrists and theologians to interweave the religious with the psychological led to the development of new strains of practice in both disciplines. The most broadly defined psychiatric and religious blend emerged as 'transpersonal psychiatry,' the outgrowth of Jungian mysticism combined with tenets of Eastern religion (such as Buddhism, Hinduism, and Taoism). Proponents of transpersonal psychiatry use drug-free methods familiar in spiritual practices, i.e., meditation, as part of the ritual for maintaining optimal mental health.

A distinctly different approach to the integration of religion and mental health has been seen in the emergence of 'Christian Psychiatry' in the 1970s and 1980s. Led by theologically conservative psychiatrists, this approach was specifically tailored for Christian patients whose treatment mode, psychiatrists believed, could be enhanced by incorporating elements of their Christian faith in their treatment. 'Christian psychiatrists' based their treatment on Biblical principles, including Bible study and prayer.

Also within the theological realm, the development of Clinical Pastoral Education (CPE) emerged as clergy became educated about the nature of psychological problems. Originally developing with the work of Anton Boisen in 1925, CPE emphasizes ministering to the patient by utilizing both theological tenets and psychodynamic theory. CPE subsequently has become an important part of seminary training: a requirement of some, and an elective in even more, to best meet the needs of pastors and chaplains in hospitals and other settings.

8. *Psychiatry Addresses the Issue of Religion*

More than 50 years of writing on the positive interaction between religion and psychological health passed before the field of psychiatry officially reexamined its Freudian views on religious belief. In 1956, R. Finley Gale, Jr., the incoming president of the American Psychiatric Association, devoted his entire presidential address to the 'Conflict and Cooperation between Psychiatry and Religion.' In that speech, Gale called for the development of a committee on psychiatry and religion within the organization, and further suggested that mental health chaplains be invited to join and participate in that venture.

Although the next several decades indicated some gains in psychiatric acceptance of religion as nonpathological, diagnoses as late as those found in the 1987 *Diagnostic and Statistical Manual*, *Third Edition*, *Revised* indicated religious belief as a factor indicated in poorer mental health (Larson and Greenwold 1997). However, the 1990 adoption of the *Guidelines Regarding Possible Conflict between a Psychiatrist's Religious Commitments and Psychiatric Practice* indicated that 'patient religious beliefs were being professionally recognized as not only acceptable, but important treatment factors' (Larson and Greenwold 1997, p. 22).

These guidelines highlighted the usefulness of obtaining information about a patient's religious or ideological orientation and beliefs, as well as of addressing the procedure for handling any conflicts between the patient's and the psychiatrist's belief systems. Psychiatrists were reminded that they should not impose their own religious, anti-religious, or other ideological systems or beliefs on vulnerable patients. Furthermore, these guidelines indicated that interpretations concerning a patient's beliefs should be made in a context of empathetic respect for their value and meaning to the patient.

9. *Reflections on the Popular View*

Despite the defamation of religion by the scientific community, demographic studies indicate that those who define themselves as religious/spiritual in the population at large have remained relatively consistent over the past several decades. The first Gallup poll, conducted in 1937, indicated that 73 percent of Americans were members of a church or synagogue. A 1997 Gallup poll (March 24–26, 1997) reported that 96 percent of Americans believe in God or a universal spirit, with 67 percent indicating membership in a church or synagogue. Further, 61 percent indicate religion is 'very important' in their own life (with an additional 27 percent stating it is 'fairly important'), with 43 percent attending religious services weekly or nearly every week.

In addition, one cannot overlook the host of religio-secular self-help support groups in mainstream culture today, particularly Alcoholics Anonymous (AA), the most therapeutically successful of them all. Alcoholics Anonymous (along with its subsequent offspring such as Narcotics Anonymous, etc.) is considered the most useful organization in the United States to assist persons suffering from alcohol abuse and chemical dependence. AA operates much like a religious community, complete with doctrine. The 12 steps proposed as necessary to overcome alcoholism reflect a creed that shapes AA, and has strong religious overtones. A similar group, the Samaritans, operates primarily in the United Kingdom. Its 23,500 volunteers make themselves available by telephone 24 hours a day, by personal visits, and by letters.

10. *Recent Academic Research Supports the Benefits of Religion*

Finally, recent research studies on the interaction between religion and mental health indicate more positive outcomes for believers receiving mental health treatment than for agnostics or atheists. Supporting the anecdotal success stories from Alcoholics Anonymous, a study by Desmond and Maddux (1981) demonstrated that 45 percent of participants in a religious treatment program for opioid dependence were still sober a year later, in stark contrast to the mere 5 percent who had participated in a nonreligious treatment program.

More recently, Koenig and Larson (1998) indicated that in the last 15 years, over 50 epidemiological studies have been published which demonstrate that 'religious persons experience not only greater well-

being and higher life satisfaction, but less anxiety and depression, lower rates of suicide, less alcoholism, less loneliness, and better adaptation to stress than do persons without spiritual resources' (p. 385). Furthermore, a substantive body of research on the acts of religious belief, such as prayer, has demonstrated positive impact on both physical and mental health.

In the preface to his 1993 book *Healing Words*, Larry Dossey, M.D., originally a skeptic, writes about finding 'over one hundred experiments exhibiting the criteria of "good science," many conducted under stringent laboratory conditions, over half of which showed that prayer brings about significant changes in a variety of living beings' (p. xv). Research indicated that patients who became prayer targets experienced altered blood pressure, slowed growth rates of cancerous cells, dissipation of tumors, an increased rate of healing for wounds, and even altered probability for heart attacks.

Finally, in terms of practical and economic outcomes, research indicates that the spiritually committed are less likely to develop serious illness, spend less time in the hospital, and do better upon discharge than those patients who are nonbelievers or aspiritual (Koenig and Larson 1998).

11. Implications for Psychiatry Today

When research indicates that participation in religious activities correlates with a patient's strengthened immune system, a reduced risk (by 60 percent) of arteriosclerotic heart disease and pulmonary emphysema, an increased (between seven and fourteen years) life span (Koenig and Sloan 2000), as well as a decreased likelihood of depression, anxiety, and persistent mental illness, the link between religious belief and improved health both physically and mentally becomes noteworthy. However, some concern remains about the potential negatives of physicians and mental health practitioners encouraging religious belief in their patients.

Richard P. Sloan, Ph.D., director of the Behavioral Medicine program at Columbia Presbyterian Medical Center in New York, opines that when 'physicians depart from their medical expertise and enter the religious realm, there's an element of coercion because patients don't want to offend or disappoint their doctors' (Koenig and Sloan 2000, p. 159). Sloan believes that melding the two realms puts a physician 'in a terrible bind. If he believes that there is strong evidence associating religion with better health outcomes … he would be derelict in his duty as a physician if he didn't advocate religion' (p. 156). He further associates this advocating of religion to prescribing antibiotics: a physician doesn't ask whether the patient is in favor of taking them, but merely prescribes them as being the best course of treatment.

Sloan's concerns about the potential dangers from intertwining religious ideology and mental health are worth taking into account. History has clearly demonstrated problematic responses to the mingling of religion and science, and an uncritical absorption of religious ideology into psychiatric theory would be foolhardy. However, given the recent research that indicates the positive benefits of religious belief in better mental health outcomes, ignoring those data would be very poor science. It would seem that the best approach would acknowledge the inherent tension between the two paradigms of science and religion, with growth in both fields occurring as the result of conversant dialogue between the two, much as Kuhn (1970) advocated in his treatise on scientific revolution.

12. Future Considerations

Given the recent scientific evidence for positive outcomes for religious belief, combined with the statistics about religious believers in the population at large, a move toward educating psychiatrists about the interaction between religious belief and mental health practice appeared necessary. Hence, in 1995 the Accreditation Council for Graduate Medical Education (ACGME) responded with a requirement to include religion and spiritual factors in the human development curriculum. To assist program directors in meeting this requirement, a *Model Curriculum for Psychiatric Residency Training Programs* was developed by David B. Larson, Francis G. Lu, J. Swyers, and Elizabeth Bowman, among others. This model curriculum provided definitions, bibliography, and teaching options to fit individual programs.

Practical considerations aside, the intermingling of religion and science can remind psychiatry that it does not practice in a vacuum or without social, ethical, and public philosophical influences. Religion confronts psychiatry with decisions regarding the proper stance of a secular, therapeutic discipline on determining components of mental health. Because it serves a wide spectrum of humanity, psychiatry should not wed itself to a specific religious or cosmological creed. It should, however, consider a public philosophy that provides 'a broad, intellectual understanding of both its proper focus as well as how it should relate to the wider aspects of social and cultural action' (Browning et al. 1990, p. 20). Such an academic stance allows psychiatrists to take their patients' experiences and beliefs seriously and in a respectful manner.

Psychiatry is on its most solid ground when it is investigating the psychological meaning and motivation of the religious beliefs of any patient. To reiterate the concerns of James and Jung, what psychiatrists must not comment upon is whether there is an ultimate truth justification for religious faith, for 'God' is, logically speaking, an axiom and not an empirically testable scientific hypothesis. Rather, acknowledging

the commensurate questions of the two paradigms (What are the fundamental components of human beings? What causes illness? What helps keep individuals healthy?) and allowing each field to share knowledge with the other as a means of improving practice in both disciplines, appears to be the most prudent course. Complementary coexistence, while maintaining studied attention to the polarizing tendencies of both religion and science, is the most likely path to providing optimal care for patients.

See also: Freud, Sigmund (1856–1939); James, William (1842–1910)

Bibliography

Alexander F, Selesnick S T 1966 *The History of Psychiatry*. Harper & Row, New York

Bhugra J (ed.) 1996 *Psychiatry and Religion: Context, Consensus and Controversies*. Routledge, London

Blazer D 1998 *Freud vs. God: How Psychiatry Lost Its Soul & Christianity Lost Its Mind*. Intervarsity Press, Downers Grove, IL

Browning D, Jobe T, Evison I S 1990 *Religious and Ethical Factors in Psychiatric Practice*. Nelson Hall, Chicago

Desmond D P, Maddux J F 1981 Religious programs and careers of chronic heroin users. *American Journal of Drug and Alcohol Abuse* **8**(1): 71–83

Dossey L 1993 *Healing Words: The Power of Prayer and the Practice of Medicine*, 1st edn. Harper, San Francisco

Eddy M B 1875 *Science and Health with Key to the Scriptures*. Published by the Trustees under the Will of Mary Baker G. Eddy, Authorized Literature of the First Church of Christ, Scientist, Boston, MA

Foskett J 1996 Christianity and psychiatry. In: Bhugra J (ed.) *Psychiatry and Religion: Context, Consensus, and Controversies*. Routledge, London

Frankl V E 1987 *Man's Search for Meaning*, 5th edn. Hodder and Stoughton, London

Freud S 1934 *The Future of an Illusion*. In: Jones E (ed.) *The International Psycho-Analytic Library, No. 15* (Trans. W D Robson-Scott). Hogarth Press, London

Fromm E 1950 *Psychoanalysis and Religion*. Yale University Press, New Haven, CT

James W 1902 *The Varieties of Religious Experience*. Longmans, Green and Co., Random House, New York

Koenig H G, Larson D B 1998 Religion and mental health. In: Friedman H S (ed.) *Encyclopedia of Mental Health*. Academic Press, San Diego, Vol. 3, pp. 381–92

Koenig H G, Sloan R P 2000 Should doctors prescribe religion? *Medical Economics* **10**: 145–59

Kuhn T S 1970 The structure of scientific revolutions. In: Neurath O, Carnap R, Morris C (eds.) *International Encyclopedia of Unified Science*, 2nd edn. University of Chicago Press, Chicago, Vol. 2(2)

Larson D B, Greenwold M M 1997 Making the case for spiritual interventions in clinical practice. *Mind/Body Medicine* **2**(1): 20–30

Larson D B Lu F, Swyers J (eds.) 1996 *Model Curriculum for Psychiatric Residency Training Programs: Religion and Spirituality in Clinical Practice*. National Institute for Health Care Research, Rockville, MD

Mengh H, Freud E (eds.) 1963 *Psychoanalysis and Faith: The Letters of Sigmund Freud and Oskar Pfister*. Basic Books, New York

Menninger K A 1973 *Whatever Became of Sin?* Hawthorne, New York

Murphy N 1997 *Anglo-American Postmodernity: Philosophical Perspectives on Science, Religion, and Ethics*. Westview Press, Boulder, CO

Rechberger E 1998 *The Impact of Einsteinian Relativity and Quantum Physics Theories on Conceptualizations of the Self in Psychology*. Unpublished dissertation. Under publication review

Tillich P 1957 *Dynamics of Faith*, 1st edn. Harper and Row, New York

White E G 1905 *Ministry of Healing*. Pacific Press, Mountain View, CA

G. T. Harding, IV

Religion, Psychology of

The psychology of religion was a central part of psychology at its inception as an independent empirical discipline at the end of the nineteenth century. Psychology is also one of the family of human sciences that has made an important contribution to the study of religion. Paloutzian (1996) provides an accessible introduction. This article will look first at some key historical figures, then at contemporary research contributions, and finally at critical current issues.

1. History

Early work on the psychology of religion included quantitative methods, such as the questionnaire research of Edwin Starbuck on types of conversion. However, the most durable masterpiece of the early period was undoubtedly the qualitative research of William James. The psychology of religion continued to be a vigorous part of the discipline of psychology through the first quarter of the twentieth century, and the most widely read summation of that period was J. B. Pratt's, *The Religious Consciousness* (see Wulff 1997).

From about 1925, there followed a fallow period in the psychology of religion, in part reflecting the behaviorist direction taken by the discipline generally. However, with the growing emancipation of psychology that took place in the 1960s, there has been a revival in the psychology of religion. This recent period has not seen any major psychological theories of religion, indeed many would doubt whether they are appropriate. However, there has been a wealth of investigations in many aspects of religious experience and behavior.

1.1 William James

For his *Varieties of Religious Experience*, written at the turn of the century, James (1985) assembled a large quantity of autobiographical material about religious experience. Several of his conceptualizations of this material have become permanent contributions to the psychology of religion, such as his distinctions between 'once-born' and 'twice-born' conversion types, and between 'sick-soul' and 'healthy-soul' types. His description of mystical experience as having the hallmarks of ineffability, a noetic quality, transiency, and passivity has also stood the test of time.

However, James' approach has been criticized (e.g., Proudfoot 1985) for his overemphasis on the private nature of religious experience as 'the feelings, acts and experiences of individual men in their solitude.' Whereas James saw religious doctrines and church practices as arising out of individual experience, others have emphasized how much individual experiences are shaped by the religious tradition. These criticisms of James have been reinforced in recent years by the growing awareness of the extent to which experience is 'specially constructed.' There has also been concern about James' over-reliance on vivid but unusual experiences; others would want to place the emphasis on the religious experiences of ordinary religious people. James also focused on the special qualities of religious experience.

In contrast, some would want to emphasize that the hallmark of religious experience is not its unusual phenomenal quality, but simply the way in which a religious frame of interpretation is applied to experiences of an ordinary kind.

1.2 Sigmund Freud

Another rather separate strand in the psychology of religion that goes back to the early twentieth century is represented by the contribution of Freud, and the psychoanalytic tradition stemming from him. Freud published several books on religion, but his most enduring contribution has probably been *The Future of an Illusion* (1962). There, he argues that religion is a form of wish fulfillment, and that God is a projection of the human mind. These were not altogether new ideas, and a similar line had been taken by Feuerbach. However, the fact that Freud approached them from a well-articulated general theory of mental functions gave them a new authority.

Though Freud's views have been very influential, there has been a growing recognition that they are highly speculative, and less authoritative than was initially supposed. It has also become clear that an alternative, more sympathetic view of religion can be developed within the framework of psychoanalytic theory, something that Freud conceded might be possible. The main influence on recent work of this kind has been Winicott's concept of the 'transitional' world, that was neither wholly internal nor external, but an illusionistic world that plays a constructive role in adaptation (see Meissner 1984). Though Freud's own work on religion was surprisingly weak in case studies, others have now conducted rich case studies documenting the close parallelism that can often be found between relationships to parents and to God. Of course, this need not be interpreted as showing that God is nothing but a projection of the mind, just that concepts of God are often shaped by personal factors.

1.3 C. G. Jung

Jung took an unusual interest in religious symbols and doctrines. Indeed, one of the reasons for his break from Freud was that he was unhappy at what he saw as Freud's disparaging treatment of religion. Jung interpreted religious symbols in relation to the personal journey of 'individuation,' a journey from ego-centeredness towards the 'Self,' involving the integration of the unconscious with the ego. Symbols are important both in representing the goal of this journey, and also in facilitating it. Central to Jung's approach to religious symbols is his theory of the 'collective unconscious' and its archetypes; Jung does not regard religious symbols as just being features of the personal unconscious.

Jung had a strong interest in symbols of the Self, and saw God and Christ as being such symbols. However, he laments the extent to which religious symbols have now lost their psychological power and are no longer felt to be relevant to the journey of individuation. His work on Christianity included extended essays on the symbolism of the Trinity and the Mass (see Jung 1969). He also took an extensive interest in Eastern religion, where he also found symbols of individuation, though he was opposed to Westerners taking up Eastern practices such as yoga.

Jung's view of religion is markedly more sympathetic than Freud's, though his reformulation of religion is at points decidedly heterodox. He tended to broaden religion to include anything 'numinous.' Also he emphasized that as a psychologist he could concern himself, not with religious metaphysics, but only with psychological realties (see Palmer 1997). Though Jung saw 'God' and 'Christ' as psychic images and symbols of individuation, whether there was a 'metaphysical' God was something about which he had no official view. However, it would be a misreading of Jung to imagine that he thought people could work out their own salvation. He thought the ego was quite unable to secure 'individuation,' and emphasized the radical otherness of the 'Self,' of which God is a symbol.

2. The Contemporary Psychology of Religion

Much contemporary psychology of religion employs empirical research methods, and there follows a brief

review of representative research findings. The fullest guide to the empirical psychology of religion is that of Hood et al. (1996). Beit-Hallahmi and Argyle (1997) is briefer, but also helpful.

2.1 Individual Differences

The study of individual differences has long been an important part of the psychology of religion. The search for a particular kind of personality associated with religious allegiance has generally not been successful, largely because of the marked diversity of religious people. However, there is reliable evidence supporting an association between different personality types and different approaches to religion. This emerged first in connection with social prejudice, where 'intrinsic' religious people were found to be less prejudiced on average than extrinsic; the same is true of committed vs. consensual religious people (another set of terms for a related distinction). There has also recently been considerable interest in the relation of Jung's psychological typology, as measured by instruments such as the Myers–Briggs Indicator, to various aspects of to religious life such as prayer and Church leadership.

2.2 Mental Health

Mental health (at least as measured by self-report instruments) has been found to show the same relationship as social prejudice to types of religiousness, with intrinsic religious people having good mental health, but extrinsic religious people having relatively poor mental health. Even where religious people are not subdivided, religiousness tends to be associated with better than average scores on dimensions such as Eysenckian 'psychoticsm.' When more objective measures are considered, religious people are often found to have lower levels of formal mental illness and abuse of drugs and alcohol. There is also growing evidence that religion has a positive relationship to happiness, coping, and physical health. There are probably a variety of explanations of the relationship between religion and health. These include 'selection' factors, such as alcoholics being less likely to go to church. However, there may be various ways in which religion can improve health and happiness, such as the therapeutic value of prayer, and the social support provided by the churches (see Loewenthal 1995).

2.3 Developmental Psychology

Children's' intellectual understanding of religion has been found to follow the usual path, documented by Piaget for many areas of cognitive development, from concrete to abstract modes of understanding. Early research showed that this applied to the interpretation of Bible stories, but the same general pattern seems to apply to all aspects of religious understanding, such as prayer and sacraments. There have been attempts to extend such 'stages of development' models to other aspects of religion, and to take them into adulthood. The best-known work of this kind is Fowler's theory of 'faith development.' However, many critics have expressed caution about the proper limits of stage models of religious development. They seem to work best for development in the intellectual understanding of religion, less well for other aspects; they also only seem to work better for religious development in children than in adults. The key issue is whether religiousness changes in a sufficiently predictable and universal way for the concept of development to be appropriate.

2.4 Religious Experience

There has been extensive empirical research on mystical and religious experience, though it is beset by the problem of how such experience should be defined. The initial question used in the research of Hay (1987) is characteristic: 'Have you ever been aware of, or influenced by, a presence or power, whether you call it God or not, which is different from your everyday self ?' Most empirical surveys find roughly a third of the population respond positively to such questions. Interestingly, the proportion is not much reduced for atheists and agnostics. The experiences often occur at times of stress, and generate feelings of peace and happiness. One of the most debated matters is whether there is a common core to such experiences that is independent of religious tradition and culture. Psychometric research by Hood suggests that there are two groups of descriptors, one focusing on the experience of unity, the other set of descriptors more influenced by religious background.

2.5 Neuropsychology

The study of the brain processes involved in religion promises to become a lively research area. There has long been evidence that meditation alters people's levels of arousal, though probably not in any unique or distinctive way when compared to other procedures such as relaxation. There has been considerable interest in a possible link between the brain processes involved in religious experience and temporal lobe epilepsy. Though there are similarities in the two sets of experiences, there are also marked dissimilarities, and critics have pointed out that evidence that the same brain mechanisms are involved is weak (Jeeves 1997). d'Aquili and Newberg (1998) have proposed another neural theory in which two cognitive oper-

ators, the 'causal operator' and the 'holistic operator' underpin different aspects of religion. As evidence accumulates about the brain processes involved in religion, there is likely to be debate about whether or not they offer a complete explanation of religion.

2.6 Attitudes

The measurement of religiousness is an aspect of the study of attitudes and involves the vexed question of the relationship between self-report and behavioral measures. It has been claimed by some that religious believing currently exceeds belonging. Though, on some measures, this appears to be the case, it depends on exactly what measures of each are used. One of the methodological principles to have emerged in the general literature on attitude measurement is that single-item measures of expressed attitudes are a poor predictor of behavior, whereas fuller self-report measures predict better, and there is reason to think that this applies to religious attitudes. Attitude change can be studied in relation to religious conversion.

2.7 Conversion

Conversion has long been a central topic in the psychology of religion. It is a multifaceted, and by no means uniform phenomenon. There are well-established differences between sudden and gradual conversion types, with the former being more conservative in religion and in other matters. It has also long been noticed that adolescence is the peak time for conversion, but the interpretation of this needs to take account of the fact that it is probably the peak time for apostasy as well. Conversion often does not represent as complete or permanent a change as is supposed. Also, many of those experiencing a conversion are returning to previously held attitudes; conversions among those with no previous religious affiliation are relatively rare. In recent years there has been particular interest in conversions to new religious movements. In fact, these often seem to be associated with improvements in the sense of well being. There is some evidence of mental health problems in those who leave new religious movements, though this no means always occurs.

2.8 Charismatic Religion

Speaking in tongues (glossolalia) is much the best-investigated charismatic phenomenon (Malony and Lovekin 1985). Psycholinguistic research generally indicates that it is to be regarded as some kind of 'ecstatic utterance' rather than the use of an identifiable language. Social learning is apparently required before people are able to speak in tongues; this is suggested by the way in which people first watch it before joining in, and also get more proficient at it with practice. It has been suggested that it involves some kind of dissociated state of consciousness, perhaps a trance state or a regressive state, but there is as yet no clear support for such hypotheses. It has been difficult to identify personality characteristics of those who speak in tongues. This may perhaps be because it occurs in different ways in different settings, but there is as yet no generally accepted classification. However, there is quite good evidence that it is associated with positive personality changes, though it is not unique in having such beneficial effects.

3. General Issues

The psychology of religion involves a number of general issues about methodology and interpretation, including the following.

3.1 Relation to General Psychology

The psychology of religion currently lacks integrative conceptual frameworks and general theories. A psychology of religion that is no more than a fragmented study of various particular religious phenomena will never look impressive scientifically. The other pitfall to be avoided is, of course, offering a psychological theory of religion that is too simplistic to handle the relevant phenomena adequately. The kind of theory Freud offered, in terms of religion meeting psychological needs, proved neither convincing nor scientifically fruitful, and is not a good precedent to follow. However, the problems faced in studying religion psychologically are not unique. The study of emotion, for example, faces similar problems, in that emotion is as multifaceted a phenomenon as religion. Broadly conceived 'cognitive' approaches to religion may give useful coherence, as they are currently doing in the psychology of emotion. The psychology of religion interfaces with many other areas of general psychology. Indeed, it can be argued that the subject is pursued most fruitfully when its links with general psychology are strong.

3.2 Relation to Sociology

The separation between sociology and psychology is in some ways an artificial one, and affects the study of many social phenomena, including religion. Most religious phenomena have both social and psychological aspects. For example, conversion and speaking in tongues have been studied fruitfully from both perspectives. However, social and psychological

studies often proceed in isolation from each other; it seems unlikely that either alone will be able to offer an adequate explanation of any religious phenomenon though there is a strong case for the psychology of religion having stronger links with the sociology of religion. This tends to raise the question of whether religion is to be understood in social or individualistic terms. Both perspectives are no doubt needed for an adequate understanding of religion, and the psychology of religion does not exclude a social approach. Though the increasing interest in the brain processes in religion are likely to pull the psychology of religion in a biological direction, the psychology of religion, like other aspects of psychology, will need to remain both a social and a biological science, and to hold together the different methodologies involved.

3.3 Relation to Theology

Lastly, what should the relationship be between the psychology of religion and theology? Different countries have different traditions about the relationship between theology and religious studies, some maintaining a sharp distinction between the 'insiders'' perspective of theology and the 'outsiders'' perspective of religious studies (including the psychology of religion); others have tried to maintain links between them. This issue is affected by the postmodern turn in philosophy, which is casting doubt on the possibility of any wholly detached, objective study of anything, including religion. The psychology of religion has been slow to accommodate itself to these general intellectual developments, though it is hard to see that it can stand aside from them indefinitely. One particular issue is whether there is scope for insiders' and outsiders' explanations to run alongside one another. Charismatic phenomena represent an example. It is not clear that the study of the psychological processes involved in charismatic religion excludes a complementary religious account in terms of the activity of the Holy Spirit.

3.4 The Nature of Explanation

Many of these issues are brought together in questions about the nature of explanation in the psychology of religion. Explanations in psychology are seldom complete explanations, and are certainly unlikely to be such in the psychology of religion. This means that religion will not be reduced to psychology, or 'explained away.' Equally, religion is not a 'natural kind,' and all explanations in the psychology of religion will therefore be culturally contextual, and depend on interpretations of the nature of religion; the psychology of religion can never hope to provide a science of religion that is a 'view from nowhere.' Nevertheless, there is no sharp boundary between the explanation and interpretation of religion, and it is inherent in the nature of religion that it will adopt a mixed approach.

Bibliography

Beit-Hallahmi B, Argyle M 1997 *The Psychology of Religious Behaviour, Belief and Experience*. Routledge, New York

d'Aquili E G, Newberg A B 1998 The neuropsychology of religion. In: Watts F (ed.) *Science Meets Faith*. SPCK, London

Freud S 1962 *The Future of an Illusion*, Rev. edn. Norton, New York

Hay D 1987 *Exploring Inner Space: Scientists and Religious Experience*. 2nd edn. Mowbay, London

Hood R W Jr., Spilka B, Hunsberger B, Gorsuch R 1996 *The Psychology of Religion: An Empirical Approach*, 2nd edn. Guilford Press, New York

James W 1985 *The Varieties of Religious Experience*. Harvard University Press, Cambridge, MA

Jeeves M A 1997 *Human Nature at the Millennium*. Baker Books, Grand Rapids, MI

Jung C G 1969 *Psychology and Religion: West and East*. Princeton University Press, Princeton, NJ

Loewenthal K M 1995 *Mental Health and Religion*. Chapman and Hall, London

Malony H N, Lovekin A A 1985 *Glossolalia: Behavioral Science Perspectives on Speaking in Tongues*. Oxford University Press, New York

Meissner W W 1984 *Psychoanalysis and Religious Experience*. Yale University Press, New Haven, CT

Palmer M 1997 *Freud and Jung on Religion*. Routledge, London

Paloutzian R F 1996 *Invitation to the Psychology of Religion*, 2nd edn. Allyn and Bacon, Boston

Proudfoot W 1985 *Religious Experience*. University of California Press, Berkeley, CA

Wulff D M 1997 *Psychology of Religion: Classic and Contemporary Views*, 2nd edn. Wiley, New York

F. Watts

Religion, Sociology of

Though central to the concerns of the founding fathers and to such major mid-twentieth century figures as Talcott Parsons and Robert Merton, religion has, for some time, been regarded as of marginal interest within the discipline of sociology. However, the recent resurgence of ethno-religion and of militant variants of major world religions has incited mainstream reconsideration of the salience of the topic thereby directing attention toward the sociology of religion as a specialized area of research. Meanwhile, the subdiscipline has acquired a significant role in the generation of insights and perspectives among theologians, historians of religion and exponents of *Religionswissenschaft*.

1. *The Role of the Classics*

With roots in the earliest versions of sociological thought, the sociology of religion remains strongly influenced by the classics and Durkheim and Weber still supply the main axes of intellectual tension within subdisciplinary theoretical discourse (O'Toole 1984). The former's concern with religion's role in social cohesion, group stability, and the reproduction of sociocultural forms is complemented by the latter's preoccupation with its part in radical, large-scale social and cultural transformations. Their writings are constantly ransacked for insights while their ideas, often elaborated or even creatively misinterpreted, continue to permeate contemporary discussion of such topics as secularization, sectarianism, civil religion, new religious movements, ecclesiastical organizations, ritual, and religious transformations.

Through the scope of their theories and the depth of their insights, the classics inevitably breed rival readings and multiple interpretations, e.g., while Durkheim and Weber are often viewed as exponents of the secularization thesis, both offer ammunition for its adversaries. The ambiguity of Durkheim's legacy is evident in attempts to explain his fundamental theory of religion. Here, metaphorical parallelism battles symbolic realism as defenders of a materialist Durkheim confront those who proclaim him an idealist thinker.

Though Durkheim and Weber constitute the twin pillars of the subdiscipline's classical legacy, other figures have played significant, though more circumscribed roles. When articulation and development of the Weberian church–sect distinction was a central subdisciplinary obsession, Troeltsch's *The Social Teaching of the Christian Churches* (1923) universally was accorded classic status as the charter document of the study of religious organization. Today, Troeltsch's volumes attract attention, less for insights into church establishment and sectarian voluntarism than for their exploration of 'the secret religion of the educated classes.' With its cultic and mystical components, Troeltsch's treatment of this topic presciently anticipates a contemporary concern with religious individualization and privatization and thus deserves continuing scrutiny.

Rarely granted subdisciplinary recognition until the 1970s, the writings of Marx and Engels may now be said to have achieved minor classical standing. They have influenced theoretical discussion as well as inspiring a minority scholarship whose neo-Marxism is characterized by a sense of the 'relative autonomy' of religion and a belief in continuity rather than conflict between the views of Marx and Weber (O'Toole 1984).

Alexis de Tocqueville has long lurked in the background of sociological discussions of American religion but, until recently, his *Democracy in America* (1835) never achieved the subdisciplinary classic status of Niebuhr's *The Social Sources of Denominationalism* (1957). Whereas Niebuhr is now largely neglected, however, Tocqueville increasingly is commissioned in the service of a current fascination with the 'vitality' of US religion and the sources of a supposed 'American Exceptionalism' in the religious realm. In its assertion of the significance of religious disestablishment and voluntarism, his work is esteemed as the fountainhead of a perceived 'New Paradigm' in the study of American religion (Warner 1993).

Long invoked without ever approaching the subdisciplinary standing of Durkheim or Weber, Simmel, too, is the subject of revived interest attributable to the uncanny timeliness of his writings (Simmel 1997). Reflecting his resistance to that domination of the objective spirit over the subjective which characterizes modernity, his focus on a diffused religious attitude in preference to institutionalized religious forms has a decidedly contemporary resonance. Careful to distinguish between formal creeds and the subjective experience of pious individuals, Simmel explores religiosity as well as religion, portraying the latter as a co-creation of acting believers which bridges the private and the public. In a subdiscipline currently grappling with issues of religious individualization, autonomy, fragmentation, and *bricolage*, Simmel's potential contribution has perhaps never been greater.

While many sociologists of religion are content to continue a conversation with the classics, others are impatient to break the intellectual spell of what they perceive as the increasingly irrelevant products of a passing industrial era. Those intent on subdisciplinary transcendence of the classics are acutely aware, however, that scholars accustomed to invoking this legacy as new problems arise will not relinquish it lightly.

2. *Definitions*

In most definitional surveys, discussion hinges upon the merits of functional vs. substantive definitions. Within these frameworks, the relative advantages of inclusive or exclusive conceptualizations are assessed while specific formulations are examined for theoretical coherence and empirical integrity. Since its inception, the sociology of religion has been the scene of such a prolonged battle between the proponents of these strategies for defining religion that, in a sense, the debate between Durkheim and Tylor has been interminable. Contemporary substantivist formulations focus on such elusive phenomena as the sacred, the transcendent, the supernatural or the superempirical. Functional definitions portray religion as an indispensable source of social consensus and cohesion which aids human adaptation to the environment and ensures societal survival by explaining the meaning of existence and providing answers to questions of ultimate concern.

The question of how inclusive a definition may be while still retaining analytical utility is one that may be

directed against both functional and substantive formulations. On the substantivist side, the categories of the supernatural and superempirical are notoriously open-ended while, in the most inclusive functional definitions, society itself assumes a religious character. Debate over the relative merits of definitional strategies often hinges on the issue of how closely sociological usage should adhere to societal understandings. On one side, scholars express frustration that sociological usage of the term 'religion' is often remote from and counter-intuitive to anything which ordinary people would recognize as 'religious.' On the other, it is argued that 'religion' is a culturally specific term with no resonance in many societal contexts and that, in any case, no definition of general utility can be expected to conform to the specific demands of ordinary language usage.

Although subdisciplinary practitioners are accustomed to pursuing their research in a continuing context of definitional disagreement, it is an error to believe that lack of consensus in this area has no practical effect. The degree of social and cultural significance which sociologists of religion currently attribute to religion is influenced, to a great extent, by their definitional predilections.

3. *Major Themes*

In the period of its greatest growth, since the 1960s, the sociology of religion has exhibited both continuity and change. The classics remain a continuous presence although their range of application has expanded. A Judaeo–Christian emphasis is still clearly discernible but it has become self-consciously circumscribed within more globally comparative ambitions. The functionalist tone of earlier theoretical discussions has given way to a polycentric approach that critically assesses the merits and defects of rival perspectives claiming a hearing. Most notably, the variety of topics examined under the subdisciplinary rubric has widened significantly. Whereas the sociology of religion was previously directed largely toward familiar subject matter, it now increasingly ventures into mysterious territory and assumes novel theoretical and methodological challenges. It cannot be said that the sociology of religion has been unresponsive to the unfolding religious processes and transformations of the latter half of the twentieth century. On the contrary, its evolution has been largely propelled by such reactions. The present generation of subdisciplinary specialists is engaged in research on many fronts. While continuing to investigate such phenomena as mainline religion, civil religion, millenarianism, conversion, and the changing role of clergy, its representatives also are exploring energetically a wide array of new topics generated by the turmoil of the last quarter of the twentieth century. E.g., events in Eastern Europe and the former Soviet Union since 1989 have underlined religion's potency in the forging of ethnonationalism, thereby provoking numerous studies of the contemporary fate of religion in specific former Iron Curtain countries.

Although, in its strict American Protestant sense, it has claimed a longstanding subdisciplinary interest, 'fundamentalism' has also captured scholarly attention in its contemporary comparative meaning, particularly since the Iranian revolution. Adopting a blanket term coined by the mass media, sociologists of religion have participated in a number of subdisciplinary and interdisciplinary projects which examine the currently politically-charged topic of religious extremism (Christian, Jewish, Muslim, Hindu, etc.) under this heading. The effort to discern its underlying characteristics has raised the crucial question of whether the fundamentalist impulse toward dedifferentiation and totality represents an assault on modernity, a manifestation of modernism or an intriguing combination of both.

Subdisciplinary practitioners exhibit an enhanced capacity for scrutiny of specific religious developments within a large-scale comparative context and in terms of more general theoretical debate. The study of Pentecostalism, e.g., is no longer confined to the margins and backwoods of Western societies. In the works of contemporary specialists, it emerges as a movement which offers a serious challenge to mainline religions in the West, is rapidly transforming the religious maps of Latin America and Africa and is making a serious spiritual, social and economic impact on parts of Asia. Sociologists of religion have been forced by world events to situate their subject matter in an increasingly international, comparative, and global framework whether their interest is in fundamentalism, charismatic Christianity, New Religious Movements, or any other major topic of research. It is unsurprising, therefore, that they have led the way in promoting globalization as a concept of central theoretical importance, not just for their subdiscipline but for sociology as a whole.

Born in the political alienation, generational conflict, and crisis of values which characterized the era of the Vietnam War and its aftermath, the study of New Religious Movements (NRMs) revitalized a subdiscipline which appeared to have exhausted the possibilities of the church–sect approach to religious organization. Over the last quarter of the twentieth century, sociologists of religion have continued to study Unificationists, Scientologists, Rajneeshis, Transcendental-Meditationists, Hare Krishnas, and many other movements of religious innovation and experimentation. In the process, they have accumulated a mass of ethnographic material documenting their inner dynamics, recruitment practices, resource mobilization, and ideological development while simultaneously struggling to discern the broader theoretical import of this information. Originating amidst heady notions of an emerging 'new religious con-

sciousness,' another 'Great Awakening' and an imminent 'Return of the Sacred,' NRM research has long subscribed to more sober assessments of its subject matter. Nonetheless, scholars in this field reject suggestions that they are engaged in a mere pursuit of the trivial and exotic. The significance of NRMs is believed to lie in their role as indicators of broader sociocultural trends but, unfortunately, there is wide disagreement concerning their overall significance for contemporary and future religion. For some commentators, they represent a decisive blow to the secularization thesis, their very existence contradicting any notion of a 'disenchantment of the world.' For others, they constitute strong evidence in support of the secularization thesis, confirming by their tiny membership, syncretic ideology, and cultural marginality the inability of religion to restore its fortunes in any significant way.

What might be considered a 'second wave' of NRM research has focused, more recently, on such topics as Satanism, Witchcraft, neo-Paganism, and various manifestations of 'New Age' spirituality. Such studies provide intriguing comparisons and contrasts with first-generation NRMs as well as with occultic movements formed at the end of the nineteenth century. They also examine phenomena such as Wicca and neo-Paganism as indicators of such broad religious trends as individualization, privatization, and *bricolage*, thereby suggesting their essentially post-industrial or postmodern character. A growing fascination with New Age spirituality as a highly individualistic and experiential enterprise has perhaps played a part in the encouragement or revival of subdisciplinary interest in such topics as healing, the natural environment, the human body, and collective memory as they relate to religious issues. As sociologists of religion increasingly have shifted their attention in these directions, the possibility of fruitful interchange between students of NRMs and analysts of so-called 'New Social Movements' from other subdisciplines has been raised. Though the response to this challenge has been rather muted, such cooperation might facilitate a greater infusion of general theoretical ideas into the sociology of religion.

4. *Methods and Theories*

Within the sociology of religion, methodology and theory typically have been implicit rather than explicit. Perhaps partly due to a certain insulation and isolation, the kinds of self-consciously theoretical disputes which occur in other subdisciplines rarely have been witnessed. Nonetheless, since the unraveling of functionalist orthodoxy in the 1970s, the subdiscipline has experienced significant, if civil, discord which stubbornly resists efforts at securing synthesis.

The subdiscipline employs a wide range of quantitative and qualitative methodological techniques including large-scale social surveys, participant-observation, and textual analysis of contemporary and historical materials. Here, as in sociology generally, one investigator's sophisticated device is another's blunt instrument, although open debates on this theme are rare. For the most part, a wall of indifference segregates quantitative from qualitative methodologists although it has been breached occasionally. In the broadest methodological sense, the reciprocal relationship between sociological research and religious activity merits attention though it is often ignored on the basis of a glib distinction between religious sociology and sociology of religion formulated in a religious context less bewildering than the current one. To what extent sociologists of religion see their own reflection in the religious realm on which they report is a sensitive issue which surfaces in a number of ways. Thus, in asserting the impact of societal reality upon sociological method, some observers suggest that a purported contemporary transition from institutionalized to individually autonomous forms of religion necessitates a shift from qualitative fieldwork directed at groups to quantitative surveys aimed at individuals. In contrast to this, other commentators ponder the possibility that perception of a rise in privatized religion may itself be partly a product of specific, widely disseminated research techniques directed toward discernment of the dimensions of individual religiosity.

While the complexities and subtleties of the reciprocal relationship between observers and those they observe have come to the fore most notably in discussions of scholarship and partisanship in the NRM field, they are equally evident in other religious settings. Students of Liberation Theology and Base Communities, e.g., are often confronted by sociological ideas in explicit use by those under observation. More generally, widespread use of the term 'secularization' by mainline churches seems to offer an excellent example of social science rebounding upon society. It also provides a cautionary tale of the manner in which societal embrace of a sociological notion may possibly inspire a self-fulfilling prophecy.

In any general assessment of the theoretical situation within the sociology of religion, it is difficult to overcome the impression that scholars are traveling in all directions at once and that every assertion eventually evokes its exact opposite. Empirical–analytic, historical–hermenentic, and critical conceptions of science are all in evidence and the major tendencies in contemporary sociological theorizing are represented in varying degrees (e.g., functionalism, conflict theory, Marxism, critical theory, symbolic-interactionism, phenomenology, rational-utilitarianism). In such a situation, it is difficult to ascertain the main drift of theoretical innovation but recent important developments may be noted.

A mood of skepticism toward Grand Theories and Grand Narratives appears, in some quarters, to have intensified suspicions of deduction and prediction to a

point where the initial nomothetic impulse of the subdiscipline has been supplanted by an idiographic 'historical turn.' This involves a strategy of methodological individualism, a stress on individual autonomy and creativity, and an overall preoccupation with agency rather than structure. In a parallel development, longstanding inductive and interpretive tendencies have experienced reinvigoration and reinforcement as theorizing has taken a decidedly 'cultural' direction. In their efforts to take seriously the meaning of religion in the lives of believers and practitioners, proponents of this perspective employ the insights of cognitive and phenomenological theory while simultaneously drawing upon anthropological sources and material derived from the field of Cultural Studies. Fundamental to this cultural approach is a fervent antireductionism which rejects the notion of religion's epiphenomenal character and refuses to regard religious phenomena as really something else. Vigorously endorsing methodological individualism and deductive scientific logic, rational-choice theory has had an undeniably profound impact on the sociology of religion during the 1990s. Appointing Adam Smith to the subdisciplinary classical pantheon, its advocates employ economic metaphors to describe, explain, and predict religious behavior. Led by de Tocqueville's identification of a causal link between voluntarism and vitality in the American religious experience, they indicate the paramount importance of choice and competition in US religious 'exceptionalism.' Extending their analysis in a comparative direction, they strongly oppose suggestions of the inevitability of secularization, linking dynamic religious growth to a process of 'deregulation' in the religious realm. While it has won many converts, rational-choice theory has left many unconvinced by either its specific market analyses of religious profit-and-loss or its fundamental assumptions concerning religious behavior. To its most severe critics, its basic propositions appear embedded in ambiguity and tautology while its claims to universal applicability reflect an unfortunate regression into a presociological mode of analysis. Nonetheless, rational-choice analysis now represents such a significant presence that it is regarded by many of its admirers, and perhaps even some of its detractors, as the central constituent of a new subdisciplinary paradigm.

The themes of secularization and the contemporary transformations of religion have permeated subdisciplinary discussion in all its aspects since the 1960s. The concept of secularization has, for many, proven itself to be a plausible, persuasive, and resilient means for interpreting the fortunes of religion in the twentieth century. From another viewpoint it has been judged a misguided myth whose influence has retarded scholarly progress and whose persistence underlines the subdisciplinary need to transcend the classics. Although its classical provenance is tinged with ambiguity, the secularization perspective has long held a privileged position within the sociology of religion, being described frequently as a dominant 'paradigm.' Logically dependent for its meaning on the definition of 'religion,' the concept has an intrinsic inclination to ambiguity which is often reinforced by vague and careless usage. Despite the best efforts of its proponents to articulate the main elements of the secularization thesis with analytical exactitude, debates over its validity frequently are characterized by argument at cross-purpose. While the validity of empirical evidence is occasionally at issue, facts are not typically in dispute and discussions tend to hinge on the interpretation of various social patterns.

What is new about the current debate is that proponents of the secularization thesis are now decidedly on the defensive. The undermining of the secularization thesis has occurred gradually through the efforts of scholars ranging theoretically from functionalists to phenomenologists but this process has accelerated in the last decade as a result of the arguments of rational-choice theorists and a new subdisciplinary awareness of the historical record. Indeed, some sociologists of religion discern, in the current convergence of these tendencies, the emergence of nothing less than a new subdisciplinary paradigm destined to displace the secularization thesis. Whether this new outlook, which combines deductive theory with social history, may accurately be described as a paradigm is in dispute. There is little doubt, however, that it captures a certain *fin de siècle* subdisciplinary mood of stock-taking and reassessment in which the prospects of religion appear much brighter than a few decades ago. Perhaps the 'new paradigm' is best understood within the context of changing attitudes to modernity. As a by-product of current postmodernist debate, the concept of modernity has experienced revival and refurbishment, a situation with significant consequences for the study of religion. The familiar conception which assumed that as a result of differentiation and relativization, secularization was an inextricable and inevitable component of modernity has given way to a revisionist version of modernity which neither precludes the persistence of religion nor constrains its vitality. Although its significance is challenged by defenders of the secularization thesis, mounting evidence of the vigor of religion on the international scene encourages subdisciplinary embrace of a more encompassing vision of modernity.

Interestingly, this vision is complemented by an increasingly broad conception of religion. Any attempt to portray religion's changing face since the 1970s must stress the extremely wide range of phenomena currently accommodated under the religious rubric before tracing those trends which constitute the common currency of all analyses of the contemporary religious scene. These include the processes of disestablishment, individualization, privatization, fragmentation, and globalization as well as a number of

variants and counter-themes. In the process by which a primarily institutional conception of religion has given way to a far more inclusive and elusive subdisciplinary formulation, the impact of Luckmann's 'invisible religion' thesis cannot be overestimated (Luckmann 1967). Where earlier functionalist definitions had focused on such vast notions as 'ultimate concerns' and 'grounds of meaning,' Luckmann raises the stakes even higher by identifying the transcendence of biological nature by the human organism as a religious phenomenon. He thus describes a general anthropological condition that links the creation of religion inextricably with the formation of the individual and the self. In various guises, this has been a major subdisciplinary theme during the years following the promulgation of his provocative thesis.

Concerned for so long with religion's role in the creation, preservation, and perpetuation of collectivity and community, many sociologists of religion have transferred their attention to the role of the individual and the dimensions of personal religiosity. As with other major subdisciplinary developments, this shift of attention is a product of perceived changes in societal reality as much as theoretical logic. It is not necessary to subscribe to the secularization thesis in order to be aware of the profound crisis which has beset the major institutional forms of religion in Western societies since World War II. Moreover, the vastly diminished role of religion as the 'ideology of community' has become increasingly evident. Thus, the new direction taken by the subdiscipline is part of its effort to cope with the rapid transformation of religion under conditions of postindustrialism and postmodernity.

One aspect of this new reality is starkly evident in the work of Bellah et al. (1985) whose insight into the success and decline of American 'civil religion' has expanded into a concern with the collapse of community. Viewing dwindling involvement in religious organizations within the broader picture of a general decline in participation in political parties, trade unions, voluntary associations, and community groups, Bellah and co-workers perceive religion as increasingly a matter of private concern and personal idiosyncrasy.

This theme is echoed in all current investigations of institutionalized religion, particularly in its 'mainline' forms. Although there is no immediate prospect of the collapse of organized religion, its future arguably depends on its ability to accommodate new attitudes to religious authority and revised forms of religious involvement. Commentary on the contemporary religious climate demonstrates considerable consistency in detecting a number of key related processes at work. Some observers describe individualization and privatization. Others refer, in a US context, to a 'third disestablishment' involving an expansion of personal autonomy. Fragmentation frequently is explored and a growing tendency among believers to 'mix and match' elements from a variety of denominational and religious traditions is emphasized. The terms *bricolage*, religion *à la carte*, and consumerism are also utilized in portraying the extent to which individuals privately construct personally meaningful syncretic systems of belief for themselves. In examining the dilemmas of organized religion, many scholars discern a disinclination toward total acceptance of any denominational or religious belief system either through inheritance or conversion. This denotes a fundamental alteration in the popular conception of religion. No longer regarded as a source of anthoritative meaning, increasingly it is considered instrumentally as a broad, pliant, cultural resource at the disposal of autonomized individuals. Moreover, the accuracy of this assessment is underlined by researchers who document an accelerating incidence of 'belief without belonging' as a mode of religious commitment.

The future of organized religion is precarious and unpredictable but with the waning of the secularization thesis, it is less often discussed in terms of inevitability. Though it seems certain to remain a major player on the religious scene for the forseeable future, increasingly it will share the subdisciplinary limelight with more 'diffuse' and 'invisible' forms of religion. Often under the general umbrella of 'the sacred,' the attention of sociologists of religion focuses increasingly on uninstitutionalized forms of religion and their involvement in the transformation of the self. In this regard, studies of popular, 'common,' and 'implicit' religion, along with investigations of diffuse versions of spiritual and mystical experience, are now a noteworthy part of the subdisciplinary agenda. They are accompanied by efforts to discern the presence of the sacred in such secular settings as 'self-help' and therapy groups, fitness centers, weight-loss clinics, and healthfood stores which are often depicted as instruments in the pursuit of individual self-perfection.

True to form, however, there is no subdisciplinary consensus on an overall or inevitable passage into more privatized forms of religion in the postindustrial era. On the basis of evidence drawn from Poland, Spain, Brazil, and the US, Casanova (1994) perceives an intersection of religion and politics and a revival of 'public religion' in the modern world. This conception is given credence by globalization theory that regards both privatization and politicization as products of the modern global condition.

Whether religion's privatization or individualization will accelerate or whether its former capacity as a source of authoritative meaning will reassert itself and inspire novel, enduring, and significant forms of collective and public spiritual expression remains to be seen. One thing is certain: reappraisal of the past and adjustment to the present have ensured that, in the sociology of religion, the future is not what it used to be.

See also: Collective Beliefs: Sociological Explanation; Durkheim, Emile (1858–1917); Luhmann, Niklas

(1927–98); Rationalism; Religion: Definition and Explanation; Religion, History of; Religion, Psychology of; Religiosity, Sociology of; Simmel, Georg (1858–1918); Spencer, Herbert (1820–1903); Weber, Max (1864–1920)

Bibliography

Beckford J A 1990 The sociology of religion 1945–1989. *Social Compass* **37**(1): 45–64

Beckford J A 1989 *Religion and Advanced Industrial Society*. Unwin Hyman, London

Beckford J A, Luckmann T (eds.) 1989 *The Changing Face of Religion*. Sage, London

Bellah R N, Madsen R, Sullivan W M, Swidler A, Tipton S M 1985 *Habits of the Heart: Individualism and Commitment in American Life*. University of California Press, Berkeley, CA

Beyer P 1994 *Religion and Globalization*. Sage, London

Bibby R W 1987 *Fragmented Gods: The Poverty and Potential of Religion in Canada*. Irwin, Toronto

Bruce S (ed.) 1992 *Religion and Modernization: Sociologists and Historians Debate the Secularization Thesis*. Clarendon Press, Oxford, UK

Capps W H 1987 Society and religion. In: Eliade M (ed.) *The Encyclopedia of Religion*. Macmillan, New York, Vol. 13

Casanova J 1994 *Public Religions in the Modern World*. University of Chicago Press, Chicago

Clarke P B, Byrne P 1993 *Religion Defined and Explained*. St. Martin's Press, New York

Davie G 1998 Sociology of religion. In: Swatos W H (ed.) *Encyclopedia of Religion and Society*. AltaMira Press, Walnut Creek, CA

Davis W 1987 Sociology and religion. In: Eliade M (ed.) *The Encyclopedia of Religion*. Macmillan, New York, Vol. 13

Dawson L L 1998 *Comprehending Cults: The Sociology of New Religious Movements*. Oxford University Press, Toronto

Flanagan K, Jupp P C (eds.) 1996 *Postmodernity, Sociology and Religion*. St. Martin's Press, New York

Glock C Y, Hammond P E (eds.) 1973 *Beyond the Classics? Essays in the Scientific Study of Religion*. Harper & Row, San Francisco

Hadden J K 1995 Religion and the quest for meaning and order: old paradigms, new realities. *Sociological Focus* **28**(1): 83–100

Hamilton M 1998 *Sociology and the World's Religions*. St. Martin's Press, New York

Hammond P E (ed.) 1985 *The Sacred in a Secular Age: Toward Revision in the Scientific Study of Religion*. University of California Press, Berkeley, CA

Hammond P E 1992 *Religion and Personal Autonomy: The Third Disestablishment in America*. University of South Carolina Press, Columbia, SC

Heelas P 1996 *The New Age Movement: The Celebration of the Self and the Sacralization of Modernity*. Blackwell, Oxford

Hervieu-Léger D 1987 Faut-il définir la religion? *Archives de Sciences Sociales des Religions* **63**(1): 11–30

Kurtz L R 1995 *Gods in the Global Village: The World's Religions in Sociological Perspective*. Pine Forge Press, Thousand Oaks, CA

Luckmann T 1967 *The Invisible Religion: The Problem of Religion in Modern Society*. Macmillan, New York

Luhmann N 1984 *Religious Dogmatics and the Evolution of Societies*. Edwin Mellen Press, New York

Niebuhr H R 1957 *The Social Sources of Denominationalism*. Meridian Books, New York, (first published 1929)

O'Toole R 1984 *Religion: Classic Sociological Approaches*. McGraw-Hill Ryerson, Toronto

Poewe K (ed.) 1994 *Charismatic Christianity as a Global Culture*. University of South Carolina Press, Columbia, SC

Roof W C 1984 American religion in transition: a review and interpretation of recent trends. *Social Compass* **31**(2–3): 273–89

Simmel G 1997 *Essays on Religion*. Yale University Press, New Haven, CT

Swatos W H Jr (ed.) 1993 *A Future for Religion? New Paradigms for Social Analysis*. Sage, Newbury Park, CA

de Tocqueville A C 1835 *De la démocratie en Amerique*, 2 Vols. Librairie de Gosselin [Paris 1946 *Democracy in America*. Alfred A. Knopf, New York]

Troeltsch E 1923 *Die Soziallehren der christlichen Kirchen und Gruppen*. JCB Mohr, Tübingen, Germany (first published 1911) [1931 *The Social Teaching of the Christian Churches*, 2 Vols. Macmillan, New York]

Wallis R, Bruce S 1989 Religion: The British contribution. *British Journal of Sociology* **40**(3): 493–520

Warner R S 1993 Work in progress toward a new paradigm in the sociological study of religion in the United States. *American Journal of Sociology* **98**(5): 1044–93

Wuthnow R J 1988 Sociology of religion. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA

Wuthnow R J 1992 *Rediscovering the Sacred: Perspectives on Religion in Contemporary Society*. Eerdmans, Grand Rapids, MI

Young L A (ed.) 1997 *Rational Choice Theory and Religion: Summary and Assessment*. Routledge, New York

R. O'Toole

Religiosity: Modern

The *Shorter Oxford English Dictionary* defines 'religiosity' as 'Religiousness, religious feeling or sentiment.' Taking this as a cue, 'modern religiosity' is defined by contrasting it with modern forms of traditional religion. Thus modern religiosity is taken to refer to forms of religion which emphasize experience rather than the doctrines of the religious traditions of modern times; and which are relatively inchoate, indeterminate, or formless rather than being spelled out by way of authoritative belief systems. Religiosity, that is to say, is detraditionalized, the established, spelled-out, regulatory orders of the traditional not being much in evidence, if at all. Instead of living in terms of authoritative orders, religiosity is very much to do with the sphere of consciousness. And the fact that religiosity is so integrally bound up with the life of the person means that it is very much in the hands of the experiencing subject. It largely operates beyond tradition: that is to say, autonomous subjects—not traditions—are authoritative, subjects developing their own religiosities by way of the test of their own life requirements. In sum, rather than

heeding the injunctions of religious tradition, modern religiosity concerns self-based exploration of life's meaning; the autobiographical experience.

1. Contrasting Modern Religiosity and Modern Religion

Writing at the beginning of the twentieth century, Emile Durkheim drew a distinction between 'a religion handed down by tradition, formulated for a whole group and which it is obligatory to practice' and 'a free, private, optional religion, fashioned according to one's own needs and understanding' (cited in Pickering 1975, p. 96). With modifications, this contrast provides the basis for distinguishing between 'modern religion' and 'modern religiosity.'

1.1 Modern Religion

To begin with the more familiar term, modern religions provide tradition-informed, institutionalized forms of religious activity: the church, chapel, monastery, mosque, or temple. Unlike more insulated premodern religions, however, few modern religions are 'formulated for a whole group' or are 'obligatory to practice' (as Durkheim puts it). For with modernity, information and population flows have greatly increased, this meaning that globalizing and pluralizing processes have provided most people, in most cultures, with a whole range of religions from which to select. This said, however, modern religions strive (to varying degrees and with varying degrees of success) to ensure that participants remain faithful to the tradition on offer. Participants, that is to say, are instructed in a particular, clearly articulated, path to salvation. Examples include long-established forms of tradition (such as Catholicism) as well as more recent developments (such as those new religious movements dating from the nineteenth century, the Mormons providing an illustration, or those organizations that have developed during the later twentieth century, the Japanese 'new new religions' here serving to illustrate).

1.2 Modern Religiosity

Generally speaking, modern religiosity flourishes in realms beyond the institutionalized and traditionalized territories occupied by modern religions. Rather than going to church, chapel, or temple, at least to any significant degree, those involved draw on widely available cultural resources—books and magazines, films and videos—to develop their own religiosities. Or those involved resource themselves by way of 'new spiritual outlets,' turning to practitioners who provide spirituality (healings, meditations, empowerments) in a relatively informal, weakly institutionalized and nondirective fashion. Then again, there are those who exercise their imaginations to conclude that there must be 'something' more to life that the merely mundane.

Reliance on cultural resources and detraditionalized new spiritual outlets goes together with the fact that a key characteristic of modern religiosity is that those involved think of themselves as exercising their own authority to develop their own religiosities (or spiritualities). As Robert Wuthnow (1989) makes the point with regard to North America, 'the religion practiced by an increasing number of Americans may be entirely of their own manufacture' (p. 164). In order to resource their lives, individuals might draw on magical, occult, or esoteric teachings and practices; New Age themes to do with the sacralization of the Self or the importance of connecting to the spirituality of nature; the mysterious forces of '*X-file*' culture; notions such as 'spirituality,' 'vital force,' 'higher power,' 'life force' or 'soul,' and even 'angels'; or the 'inner wisdom' of religious traditions. In addition, individual religiosities can include the not-so-obviously religious in the quest for finding meaning in life, self-fulfillment or well-being: a point also made by Wuthnow when he writes of the growing popularity of meaning systems which involve, for example, an 'eclectic synthesis of Christianity, popular psychology, *Reader's Digest* folklore, and personal superstitions, all wrapped in the anecdotes of the individual's biography' (Wuthnow 1989, p. 164).

Although the term 'modern religiosity' can be taken to include vague 'there must be something more' beliefs, which have little direct bearing on the lives of those concerned, modern religiosity primarily dwells on what it is to be 'alive.' Durkheim ([1915] 1971) thought that the future lay with 'a religion which will *consist entirely in internal and subjective states*, and which would be constructed freely by each of us' (p. 47, my emphasis). What Durkheim is referring to is clearly very much to do with personal life: all those subjectivities which comprise 'being alive.' Now in the words of Wuthnow (1991), who is explaining what he means by the 'subjective' or 'lived realities' dimension of the religious: 'by "subjective", I mean an understanding of religion that privileges its connection to the individual, its location within the inner part of the individual (beliefs, outlooks, felt needs, emotions, inner experiences), and its role within the individual's personality (providing meaning, wholeness, comfort, psychological compensations, even a sense of attachment or belonging)' (p. 272).

Religiosities (or spiritualities) of life are most obviously in evidence among those who are not involved in strong—obeying the Other—traditionalized modern religion. For whereas traditionalized religious institutions serve to liberate people from their 'merely' human lives—as 'fallen,' or, in many Eastern traditions, as contaminated by 'attachments'—in favor of what lies beyond, modern re-

ligiosity concerns the autonomous subject, liberated from the demands of tradition and intent on exploring and developing what it is to be alive. This said, however, modern religiosity is by no means absent within less strongly traditionalized forms of modern religion. Peter Berger (1969), for example, in his discussion of liberal Christian teachings, notes that 'religious "realities" are increasingly "translated" from a frame of reference of facticities external to the individual consciousness to a frame of reference that locates them *within* consciousness' (p. 166). And to the extent that subjectivization or internalization is taking place within modern religions, so too does the divide between modern religion and modern religiosity become undermined.

2. Modern Religiosity and Peoples' Lives

Turning to how religiosities can serve peoples' lives, four main life focuses can be briefly summarized. The first concerns the life of the individual person. What matters in this regard lies with finding religiosity/spirituality within onself. What matters is addressing personal 'growth,' divorce, illness, ageing, by engaging with what one's own spirituality has to offer. The second concerns the life of the relationship. Religiosity is here put to work to enhance personal relationships, or to operate in therapeutic fashion. The third concerns relationships with nature. The emphasis here is on living in terms of the life of the natural; that life which informs nature from trees to species to the cosmic. And the fourth concerns what life has to do with work, increasing attention being paid in countries like the USA to exploring ways of bringing 'life back to work' by way of religiosity or spirituality.

3. The Growth of Modern Religiosity

Not enough research has been done to provide anything approaching a clear picture as to how modern religiosity is faring on a global scale. It is probably true to say, however, that in much of the world, where religious traditions—Islam, Hinduism, 'popular' Buddhism—remain strong, modern religiosity is relatively unimportant. But even in Islamic countries like Bangladesh, its role should not be neglected: many people could well combine components of Islam with folk beliefs and practices to do with improving the quality of personal life; and Tagore, the great figurehead of the nation, taught a 'religion of humanity' which is a powerfully relational form of life spiritualitity.

The one thing that is reasonably clear is that modern religiosity is well in evidence in countries where various processes of modernity—rationalization, differentiation (between the religious realm and the political, for example), democratization, pluralization, consumerization, and materialization—have weakened the hold of religious traditions. Indeed, in such countries modern religiosity could well be expanding. The reason is simple. More people are ceasing to participate in institutionalized, traditional religion than are becoming atheists or agnostics, it following that the number of those who are 'somehow' religious without being significantly involved with institutionalized religion has increased.

Northern Europe would appear to provide a major site for growth of this kind. In Britain, for example, the number of those regularly attending institutionalized religion dropped by 30 percent between 1950 and 1990, but the numbers of those becoming atheists or agnositics only increased by some 10 percent (entailing that there has been an increase of some 20 percent in those who are somehow religious without attending). It is true that many of these nonattenders might still be involved with modern religion, reading their Bibles at home, for example. But there are distinct signs that religiosity is growing: belief that 'there is some sort of spirit or vital force which controls life' has now overtaken the (traditional) 'There is a personal God' (Bruce 1996, p. 270), and New Age spiritualities of life are expanding (see Heelas, 1996). Or consider Finland: 15 percent are regular church attenders, with the figure in decline, while 63 percent believe that the 'best way to get in touch with God' is by being 'in' nature, such beliefs apparently gaining in numerical siginificance (Heino, 1997, p. 21).

Turning to the USA, research by Roof and Gesch (1995) provides convincing evidence of growth. Concentrating on the baby-boom generation, namely those 75 million (or so) people born between 1946 and 1964, by the 1990s approximately half were involved with traditional religion. Of the remainder, namely the 'religious individualists,' approximately half have turned to 'a deeply mystical conception of the Deity,' having little if anything to do with 'organized religion' (p. 141). As for elsewhere, it is highly likely that much the same picture of growth applies to countries like Canada and Australia, or to cities like Rio de Janeiro and Tokyo. Finally, on the question of growth, it is important to note that modern religiosity is also developing within modern religion: see, for example, Donald Miller's (1997) discussion of the (rapidly expanding) 'postdenominational churches' of North America with their strong emphasis on being alive in the right kind of way: 'purity of heart is more important than purity of doctrine' (pp. 127–8).

4. Explaining the Turn to Religiosity

Of the many factors that have contributed to the growth of religiosity, attention is paid here to what is called 'the cultural turn to life.' The key to the argument is that modernity has resulted in very high

value indeed being ascribed to life. In the words of Georg Simmel ([1918] 1997), 'this emotional reality—which we can only call *life*—makes itself increasingly felt in its formless strength as the true meaning or value of our existence' (p. 24). And given its value, life has quite naturally come to serve as the 'home' for religiosity for many of those who remain religious without participating in modern religion.

Looking at this argument in somewhat greater detail, Simmel and Max Weber, then Arnold Gehlen, Peter Berger, and Thomas Luckmann, have claimed that modernity has resulted in significant numbers becoming disillusioned with what the 'primary' institutions of society have to offer. From modern religion (loss of faith seen in secularization) to political ideologies, to the nuclear family, to the workplace (Weber's 'iron cage'), the claim is that the institutional order no longer succeeds in providing adequate sources of identity, purpose, or value in life. Accordingly, so the argument goes, people have been 'thrown back' on themselves as the key source of significance. With little to have faith in beyond themselves, what matters becomes a matter of what lies within: one's psychology, one's personal ethicality ('being true to oneself'), the quality of one's emotions, the importance of being authentic, the value of finding out what one truly is and capable of becoming; becoming 'whole' or healing oneself.

It is but a short step to explain the growth (in certain sociocultural contexts) of religiosities or spiritualities of life. Other than atheists and agnostics, those who have come to have 'ultimate' faith in what their interior or personal lives have to offer are going to seek out those forms of the religious which sacralize experiences of what it is to be alive.

See also: New Religious Movements; Religion: Definition and Explanation; Religion, Phenomenology of; Religiosity, Sociology of; Secular Religions; Secularization

Bibliography

Berger P 1969 *The Social Reality of Religion*. Faber and Faber, London

Bruce S 1996 Religion in Britain at the close of the 20th century: A challenge to the silver lining perspective. *Journal of Contemporary Religion* **11**: 261–75

Durkheim E [1915] 1971 *The Elementary Forms of the Religious Life*. George Allen & Unwin, London

Heelas P 1996 *The New Age Movement. The Celebration of the Self and the Sacralization of Modernity*. Blackwell, Oxford, UK

Heino H 1997 *Mihin Suomi Tänää Uskoo*. WSOY, Helsinki, Finland

Miller D 1977 *Reinventing American Protestantism. Christianity in the New Millennium*. University of California Press, Berkeley, CA

Pickering W S F 1975 *Durkheim on Religion*. Routledge & Kegan Paul, London

Roof W C, Gesch L 1995 Boomers and the culture of choice. In: Ammerman N, Roof W C (eds.) *Work, Family, and Religion in Contemporary Society*. Routledge, New York, pp. 61–79

Simmel G [1918] 1997 *Essays on Religion*. Yale University Press, New Haven, CT

Wuthnow R 1989 *The Struggle for America's Soul. Evangelicals, Liberals, and Secularism*. William B. Eerdmans, Grand Rapids, MI

Wuthnow R 1991 Religion as culture. In: Bromley D G (ed.) *Religion and the Social Order. New Developments in Theory and Research*. JAI Press, Greenwich, CT, Vol. 1 pp. 267–83

P. Heelas

Religiosity, Sociology of

1. Introduction

Religiosity is a broad and sweeping concept that unites much of the subject matter in the psychology and sociology of religion; consequently, the idea includes much more than it excludes.

1.1 Definition

In the most general sense, religiosity is a term that is used to refer to socially relevant manifestations of: (a) individual beliefs in a supernatural or transcendent reality; (b) relations of individuals to 'high-intensity values' (Roof 1979, p. 18); (c) 'beliefs about the nature, meaning, and purpose of reality' (Glock and Stark 1966, p. 3); and (d) all the behaviors that flow directly from them. Synonyms for the term in the literature of the social sciences include—among others—religiousness, religious affiliation, attachment, commitment, conviction, devotion, interest, and involvement.

2. Conceptualization and Measurement

2.1 Dimensions of Religiosity

Beginning in the 1950s, social scientists of religion directed a great deal of attention to elaborating definitions of religiosity that extended understanding of the concept beyond a mere count of people in the pews of churches and synagogues. Advances accumulated quickly and, according to one review (Roof 1979, p. 17), the 1960s became 'a pace-setting decade in this area.' In particular, sociologists had set to work crafting indicators of religiosity that took account of more than one aspect of religious commitment. After mid-century, scholars encountered an empirical religion that was, more than anything else, multidimensional.

2.1.1 Associational vs. communal religion. Some of the most basic quantitative studies of the postwar religious scene in the United States were the work of Gerhard E. Lenski (see, e.g., Lenski 1953). Lenski's major analysis of religious belief and behavior in Detroit, *The Religious Factor* (Lenski 1963 [1961]), gauged two fundamental types of religious involvement: associational (established through attendance at church services and meetings of church-related groups); and communal (the development of a religiously homogeneous circle of primary attachments). He additionally recognized two religious orientations among his respondents: one that stressed doctrinal orthodoxy, or right belief; and another, devotionalism, that located religiosity in personal contact with God (Lenski 1963, pp. 22–6; see also Burwell 1999b, Deconchy 1987, p. 35, Roof 1979, pp. 19, 20–1).

2.1.2 5-D and its critics. Also among the earliest examples of religiosity conceived theoretically and captured empirically in more than a single questionnaire item were the efforts of Charles Y. Glock and his colleagues to portray religion along up to five dimensions (Glock 1962, Glock and Stark 1966; see also Burwell 1999a, Deconchy 1987, pp. 33–43, Gorsuch 1984, pp. 232–34, Hilty et al. 1984,p. 252, Roof 1979, pp. 20–2). Initially, Glock proposed that there were at least four distinct ways in which a person could be religious. Religious commitment, he postulated, could take the form of religious emotion or 'direct knowledge of ultimate reality' (the experiential dimension); fidelity to a certain body of beliefs (the ideological dimension); participation in divine worship, prayer, and other ceremonies (the ritualistic dimension); and indications of the 'secular effects' of religion on the individual and society (the consequential dimension).

To this array he later added an intellectual dimension (Fukuyama 1961), because the believer would be expected to be 'informed and knowledgeable about the basic tenets of his faith and its sacred scriptures.' Although Glock admitted that this fifth dimension was 'clearly related' to religious ideology, he insisted that the two were separable, for religious belief demanded familiarity with something in which to believe (Glock 1962, p. S99). Indeed, each of the five dimensions represented to Glock a theoretically and empirically distinct aspect of religiosity; he explained at the outset that 'being religious on one dimension does not necessarily imply religiosity on other dimensions' (Glock 1962, p. S100).

Later studies referred to these constructs together as 'Religiosity in 5-D' (Faulkner and DeJong 1966, Roof 1979, pp. 22–9, Weigert and Thomas 1969). The results of several (see, *inter alia*, Hilty et al. 1984, King and Hunt 1990) offered new or augmented models for research, but the most prominent critics (Clayton and Gladden 1974), aided by large-scale surveys and data reduction through factor analysis, concluded that there was little empirical warrant for most of the presumed dimensions.

2.2 Meaning Systems and the Religious Imagination

Considering the manifold difficulties and relatively modest benefit in measuring religiosity as socially normative behavior within the confines of churches, sociologists—including Charles Glock himself—in the 1970s and 1980s devoted themselves to fashioning research techniques that would incorporate the broader connotations of the concept. A salient example of this approach is the innovative study by Robert Wuthnow (1976) on the meaning systems to which individuals routinely resort to lend structure and order to their experiences. Using data from the San Francisco Bay area, Wuthnow's survey could discern four different ways of making sense of causation in life (see Deconchy 1987, p. 39, Roof 1979, pp. 33–5). Respondents who attributed events to the intervention of divine agents (God or the devil) in human affairs were classified as ratifying theism. In contrast, those who believed that persons created the conditions of their own lives were said to back individualism. Believers in the explanatory power of social science traced the origins of personal situations to their surroundings in society, while people whose meanings came from mysticism relied on direct experience of the sacred and its mediation through nature, the creative arts, and fantasy.

A somewhat different stream of research has employed sets of interview questions to derive scales that purport to measure the symbolic and narrative basis of religion—what Andrew M. Greeley has termed 'the religious imagination.' Although formal theology is organized into series of logical statements, Greeley contends that personal religiosity predates and often overpowers tame instruction in dogma and doctrine because it is rooted in an imaginative appreciation of stories that renew hope in human actors. Religion is thus primarily composed of 'images that are prior to and perhaps more powerful than propositions' (Greeley 1981, p. 1).

To reach this imaginative capacity scientifically, Greeley and his collaborators have devised combinations of survey items that summarize varying images of God (see also Vergote et al. 1969), chart the perceived 'warmth' of different religious images, probe perceptions of religious figures like Jesus and the Virgin Mary, and describe the degree to which respondents regard heaven as a pleasurable place and the relationship between the divine and the human as 'grace-filled' (Greeley 1981, pp. 23–9).

2.3 Civil and Public Religiosity

Another aspect of religiosity that is sometimes overlooked because it lies just outside the institutional

boundaries of formal faith is what Robert N. Bellah (1967) has termed civil religion (see also Deconchy 1987, pp. 38–9, Roof 1979, pp. 29–31). Civil religion may be defined as a religious perspective that places the past, present, and future of the nation against the backdrop of transcendent values. Commenting on the United States, Bellah (1967, pp. 3–4) maintains that the components of civil religion in America 'provide a religious dimension for the whole fabric of American life, including the political sphere.' A special category of symbols and rituals that pertain to the identity and purpose of Americans thereby attains a sacred status in collective life.

In such a perspective nations not only have histories, they have destinies. Although a patriotic citizen may also be civilly religious, belief in the civil religion is different from simple patriotism. Despite, too, the frequent mixture of politics and piety (Glock and Stark 1966, pp. 82–6), civil religion is usually more than the casual consecration of secular aims. In the eyes of an adherent to the civil religion, a nation's official philosophies, public policies, and international statecraft all invite the intervention of Providence—at the same time that their consequences promise to come under divine judgment. A kind of religious significance thus permeates the entirety of the public realm yet is dependent on no single church or denomination for its vitality.

2.4 Types of Religious Orientations

2.4.1 Intrinsic vs. Extrinsic religion. To a larger extent than sociologists, psychologists of religion have concentrated on the questions of how and why people hold the religious beliefs that they do. Central in this tradition of inquiry has been Gordon W. Allport's distinction between Intrinsic (I) and Extrinsic (E) religious orientations (Allport and Ross 1967; see also Batson and Ventis 1982, pp. 140–9, Deconchy 1987, pp. 37, 40, Hunt and King 1971, Roof 1979, p. 20). 'No approach to religiousness has had greater impact on the empirical psychology of religion,' one scholar observes; Allport's scale to measure religious orientations, furthermore, is 'one of the most frequently used measures of religiousness' (Donahue 1985, p. 400). Another pair of commentators (Kirkpatrick and Hood 1990, p. 442) concurs that the I/E comparison 'currently represents the backbone of empirical research in the psychology of religion.'

As the labels imply, intrinsically religious persons are those for whom religion is an end and not a means. Religion is 'their master motive' that provides the guide to understanding all of life. In contrast, extrinsically religious persons engage in religion with the hope for something else, such as peace, protection, or status in the community. For them, a religious faith 'is lightly held or else selectively shaped to fit more primary needs' (Allport and Ross 1967, p. 434). Extrinsic religion, notes one psychologist (Donahue 1985, p. 400), 'is the religion of comfort and social convention, a self-serving instrumental approach.'

An enormous amount of research on the I and E scales has been conducted in nearly 50 years. Perhaps the most famous set of these studies has linked the Extrinsic religious orientation to expressions of prejudice. But one of the main complaints about the formulation reverts back to its first appearance—as a more value-laden contrast by Allport (1950) between 'mature' and 'immature' religious faith. The implication is that Intrinsic religiosity is somehow genuine and 'pure'—unencumbered by ulterior motives, while the Extrinsic brand is phony, selfish, and manipulative. In the descriptions of I and E, the Intrinsic believer is clearly preferred to the Extrinsic; as Michael J. Donahue (1985, p. 416) concedes, the E scale 'does a good job of measuring the sort of religion that gives religion a bad name.' James E. Dittes (1971, p. 375) is much more harsh: to him, I and E struggle beneath a 'heavy contraband load of value judgment that simply will not be sloughed off.'

Add to this problem of bias the repeated statistical finding that I and E are not two poles of the same continuum, but rather two variables with only a limited correlation (Allport and Ross 1967, p. 437, Donahue 1985, pp. 401–4, Kirkpatrick and Hood 1990, pp. 448–50, 454). For these and other reasons, I and E may survive in the scientific study of religion solely in an amended shape.

2.4.2 The 'Quest' variable. Dissatisfied with the dual categories of Intrinsic and Extrinsic religiosity, psychologist C. Daniel Batson has developed a third type of religious orientation that he dubbed Quest (Batson and Ventis 1982, pp. 149–68; see also Deconchy 1987, p. 40, Donahue 1985, pp. 411–14, Gorsuch 1984, pp. 234–5, Kirkpatrick and Hood 1990, p. 447). A person who conceives of religion as a quest is one who actually welcomes 'a readiness to doubt and to be self-critical' and who acknowledges the 'incompleteness and tentativeness' of religious faith at any moment (Batson and Ventis 1982, p. 149).

2.4.3 Other varieties of religious experience. One of the oldest traditions in the social sciences of religion attempts to study, with precision and objectivity, experiences of the sacred. The subjective nature and ordinarily fleeting passage of religious experiences, however, makes this goal extremely difficult to achieve. Yet the pioneering psychologist and philosopher William James tried, and his classic work, *The Varieties of Religious Experience* (James 1902), is the product.

In spite of this distinguished start, studies of mystical and ecstatic experiences have not yet culminated in a secure body of knowledge. It is nevertheless possible to specify several recurring features of religious experiences, as distilled from numerous cases. People who have experienced a mystical episode remember feeling as if it occurred outside the physical constraints of space and time. In this special state, they recall a spirit of peace and even elation enveloping them. Further, they report receiving personal illumination or enlightenment of unquestioned authority, while still assimilating seemingly contradictory or paradoxical perceptions. They apprehend a sacred presence, and commonly recount a sensation that all of reality was merging into oneness. In the single greatest barrier to researchers, though, mystics also contend, after their re-entry into normal consciousness, that the essence of their experience was indescribable (see Stace 1960, Stark 1965).

3. *A New Religiosity?*

Many of the ideas and approaches that social science devised in the last half-century to handle religiosity are slowly falling into disuse today, as both conceptions of religion in social science and evidence of its effects are changing.

More and more, scholars in disciplines such as psychology and sociology (e.g., Roozen et al. 1993, pp. 45–8) are describing a contemporary style of religious belief and practice that is at heart individualistic. Although people in plural societies have always had available to themselves a range of religious choices, seldom in the past have these decisions been so free of influence from social factors in one's background like family, ethnicity, or language. Rather, the assumption of a religious affiliation in the current circumstance is increasingly voluntary; it amounts to an expression of identity, and often a private or highly personal one at that.

Thus recent years have seen an abundance of religious pursuits, especially among the so-called 'baby boom' generation (cohorts born between 1946 and 1965), that have an eclectic quality. Popular varieties of faith combine elements of meditative practices, environmental activism, 'New Age' consciousness, and self-awareness with a therapeutic flavor. In the twenty-first century, many who no longer consider themselves conventionally religious nevertheless claim to possess a deep 'spirituality.'

4. *Conclusion*

Current research and theorizing in the scientific study of religion seeks to depart from older approaches to religiosity that anchored the topic in categorical contrasts, whether bipolar or multidimensional. Instead, sociologists and psychologists are busy tailoring strategies and tools for research on a concept of religion that is personal more than institutional, amorphous and fluid rather than historical and orthodox, and not static but changing.

In an age when religiosity assumes both a ubiquity in common concerns over truth and authenticity, and an idiosyncracy in multiplying paths of personal development, flexible ideas and methods of interpretation are definitely needed. Yet, now as before, the challenge for social scientists is to pursue analyses of the sacred into the meanings with which humans endow it and through the actions that proceed from these commitments.

See also: Belief, Anthropology of; New Religious Movements; Religion and Health; Religion: Definition and Explanation; Religion: Evolution and Development; Religion, History of; Religion, Psychology of; Religion, Sociology of; Religiosity: Modern; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State: Cultural Concerns; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns; Women's Religiosity

Bibliography

Allport G W 1950 *The Individual and His Religion: A Psychological Interpretation*. Macmillan, New York

Allport G W, Ross J M 1967 Personal religious orientation and prejudice. *Journal of Personality and Social Psychology* **5**: 432–43

Batson C D, Ventis W L 1982 *The Religious Experience: A Social-Psychological Perspective*. Oxford University Press, New York and Oxford, UK

Bellah R N 1967 Civil religion in America. *Daedalus* **96**: 1–21

Burwell R 1999a Dimensions of religious commitment (Glock and Stark, 1966). In: Hill P C, Hood Jr R W (eds.) *Measures of Religiosity*. Religious Education Press, Birmingham, AL

Burwell R 1999b Religious orientation and involvement (Lenski, 1961). In: Hill P C, Hood Jr R W (eds.) *Measures of Religiosity*. Religious Education Press, Birmingham, AL

Clayton R R, Gladden J W 1974 The five dimensions of religiosity: Toward demythologizing a sacred artifact. *Journal for the Scientific Study of Religion* **13**: 135–43

Deconchy J-P 1987 Les méthodes en psychologie de la religion: Leur évolution récente [Methods in psychology of religion: Current trends]. *Archives de sciences sociales des religions* **63**: 31–83

Dittes J E 1971 Typing the typologies: Some parallels in the career of Church-Sect and Extrinsic-Intrinsic. *Journal for the Scientific Study of Religion* **10**: 375–83

Donahue M J 1985 Intrinsic and Extrinsic religiousness: Review and meta-analysis. *Journal of Personality and Social Psychology* **48**: 400–19

Faulkner J E, DeJong G F 1966 Religiosity in 5-D: An empirical analysis. *Social Forces* **45**: 246–54

Fukuyama Y 1961 The major dimensions of church membership. *Review of Religious Research* **2**: 154–61

Glock C Y 1962 On the study of religious commitment. *Religious Education* [*Research Suppl.*] **57**: S98–110
Glock C Y, Stark R 1966 *Christian Beliefs and Anti-Semitism*. Harper and Row, New York and London
Gorsuch R L 1984 Measurement: The boon and bane of investigating religion. *American Psychologist* **39**: 228–36
Greeley A M 1981 *The Religious Imagination*. William H. Sadlier, Inc., New York and Los Angeles, CA
Hilty D M, Morgan R L, Burns J E 1984 King and Hunt revisited: Dimensions of religious involvement. *Journal for the Scientific Study of Religion* **23**: 252–66
Hunt R A, King M B 1971 The Intrinsic–Extrinsic concept: A review and evaluation. *Journal for the Scientific Study of Religion* **10**: 339–56
James W 1902 *The Varieties of Religious Experience*. Longmans, Green and Company, New York
King M B, Hunt R A 1990 Measuring the religious variable: Final comment. *Journal for the Scientific Study of Religion* **29**: 531–35
Kirkpatrick L A, Hood Jr R W 1990 Intrinsic–Extrinsic religious orientation: The boon or bane of contemporary psychology of religion? *Journal for the Scientific Study of Religion* **29**: 442–62
Lenski G E 1953 Social correlates of religious interest. *American Sociological Review* **18**: 533–44
Lenski G E 1963 [1961] *The Religious Factor: A Sociological Study of Religion's Impact on Politics, Economics, and Family Life*. Doubleday and Company, Garden City, NY
Roof W C 1979 Concepts and indicators of religious commitment: A critical review. In: Wuthnow R (ed.) *The Religious Dimension*: *New Directions in Quantitative Research*. Academic Press, New York
Roozen D A, Carroll J W, Roof W C 1993 La génération née après-guerre et la religion instituée: Un aperçu sur 50 ans de changement religieux aux États-Unis [The post-war generation and establishment religion: A window to 50 years of religious change in the United States]. *Archives de sciences sociales des religions* **83**: 25–52
Stace W T 1960 *Mysticism and Philosophy*. Lippincott, Philadelphia, PA
Stark R 1965 A taxonomy of religious experience. *Journal for the Scientific Study of Religion* **5**: 97–116
Vergote A, Tamayo A, Pasquali L, Bonami M, Pattyn A, Custers A 1969 Concept of God and parental images. *Journal for the Scientific Study of Religion* **8**: 79–87
Weigert A J, Thomas D L 1969 Religiosity in 5-D: A critical note. *Social Forces* **48**: 260–63
Wuthnow R 1976 *The Consciousness Reformation*. University of California Press, Berkeley, CA

K. J. Christiano

Religious Fundamentalism: Cultural Concerns

'Fundamentalism' is the name applied to characterize a variety of modern religious movements internationally, but its use is contested in some cases. The key feature in these movements is their reactive character. That is, they are not simply traditional, classic, conservative, or orthodox. Scholars and others usually speak of fundamentalism only in cases that find sets of religious people responding to certain kinds of challenges. The challenges may vary, but they tend to be variations on what the reactors or observers ordinarily would call 'the modern,' 'modernity,' 'modernization,' or 'modernism.'

1. The Term 'Fundamentalism'

One of the most controversial features of the movements called fundamentalist is the term most widely used to designate them. There are two main reasons for some participants in the movements, people in statecraft, and scholars to have reservations about the use. First, the name is derived from a reactive movement in American Protestantism, and for some decades after the 1920s remained chiefly related to them. To some, the transfer of this term to other religious contexts looks like one more example of a Western linguistic imperialism that needs to be resisted. Others contend that features peculiar to a specific Protestant conflict differ in so many respects from those in other reactive movements that only confusion could result from the use of the term. The second reason why many are reluctant to accept the term as their self-description or for scholarly purposes has to do with the fact that it can be used pejoratively and stigmatically. In a way, this resistance is ironic, since in the American context and, by analogy, in other religions and cultures, those who formed and originally adhered to such militant movements originally *wanted* to be stigmatized, and still may want this, but on their own terms.

Thus, in an American controversy in the 1920s, partisans in denominational warfare regularly complained that while the majority of their fellow-believers claimed to be conservative, as conservatives these did not do battle for the divine cause. Since they were too and compromising, it was argued, a new term had to be coined and applied. This term would help provide boundaries around the movement and create distance between it and conventional conservatisms—or against anything else that implied a potential for compromise and comfortable assimilation. But because the strategy of setting boundaries and deliberately seeking to offend others went so far that it often ran counter to some of the positive purposes of the movements, many became discontented with the self-chosen name. Some sought moderating terms. In the American case, 'neo-evangelicalism' and then 'evangelicalism' served this purpose.

The term will not be found in dictionaries or encyclopedias, even encyclopedias of religion in America, before the time of a fundamentalist–modernist controversy in the US, one that came to a

climax around 1925. When the word does make its way into reference books after that period, for example in German or French dictionaries, all the meanings of fundamentalism get reduced to something about a battle over the verbal inerrancy and literal interpretation of the Christian Bible—a reduction abhorrent to members and confusing to observers of Islamic, Jewish, or other religions' factions.

Given these burdens associated with the term, there is good reason to ask about its continued usefulness and appropriateness. Several reasons are cited in defense, among them the following: first, the word is 'there.' So widespread is the use among reporters, scholars, and politicians that it is not likely to disappear because of discontent over its use by many. Better to clarify what is meant, it is said, than to invent new coordinating terms that will meet similar resistance in due course. The second reason has to do with the need for some such coordinating designations. To mention that need is, of course, to open the door to the many controversies over comparative methods. James (1902) typically made a defense of the practice of typifying and classifying. Without classifying, it is hard to understand many features of phenomena. At the same time James warned against the danger of losing the individual and particular character of the objects studied.

In the case of fundamentalism, the interest in coordination is almost entirely formal, not substantive. That is, one notices aspects of certain religious movements by studying characteristic features while being well aware that these movements all differ from each other in respect to content. Indeed, a Muslim fundamentalist will be 'doctrinally' more distant from a Protestant fundamentalist than modernists in both religions would be to each other. So there is no point in using the word 'fundamentalist' to characterize the central teachings of these religious expressions.

Even here, however, the line is a bit fuzzy. Thus, through comparison one can generalize that all the fundamentalisms, apparently paradoxically because they resist that which is modern, and though electronic media result from modern invention, are adept at using radio, television, film, and the Internet to hold their groups together, seek converts, and express themselves to outsiders. That observation grows out of formal comparison. But a second generalization, that all the fundamentalisms are not only patriarchal but that in their talk about God, they stress God's maleness while moderate and modernist movements play down masculine references to deity does show that there can be certain kinds of substantive similarities.

A third reason for using the term, say its defenders, is that no clear alternative is apparent. Some reporters, trying to avoid the word, will use something like 'militants,' appropriate for one dimension of some of their life. But these words do not signal anything of cultural context: one can be militant about revenge, the search for territory, or any number of themes that have nothing to do with religious reaction.

Further defense of the term comes from those who say that all words travel, and globally people would make little sense of each other if they did not have terms that could be reappropriated in a variety of cultures. Thus words like 'nationalism,' 'liberalism,' 'colonialism,' 'conservatism,' 'radicalism,' 'democracy,' and 'republicanism' are born within Western contexts and are translated for usage elsewhere. The People's Republic of China means something very different with the term 'Republic' than do those who speak of America as a republic. Yet there are reasons for people to use the term in both contexts—and then clarify what they mean and point to differences.

Attempts to replace the term usually are specific and cannot be used for coordinative or comparative purposes. Thus, one might speak of 'radical revolutionary nationalist Islamic reform,' using other borrowed terms, and then find that they do not relate to phenomena in the Catholic or Hindu world. Mention of such instances suggests why many choose to use 'fundamentalism': it has acquired some rather specific meanings, meanings that deserve examination.

2. *The Originating Instance in America as a Case Study*

American Protestantism, having been divided into many denominations from the beginning, experienced some transdenominational conflicts in the nineteenth century. The first great breach occurred between the North and the South in the Civil War, when churches on each side inspired, monitored, provided morale for, and negatively defined the other side in religious terms. But all through the century another split was taking shape, thanks to latent tensions between theologically more liberal and more conservative parties.

These tensions increased with the arrival in the US after the middle of the nineteenth century of Darwinian evolution, the higher criticism of the Bible, and various forms of political progressivism. Many conservatives massively resisted theories of evolution, in part because these were seen as assaults on the Bible. They all rejected biblical criticism as a threat to the foundations of the faith. And progressivism did not do justice to apocalyptic messages in the Bible. Meanwhile, modernist parties began to develop toward the turn of the twentieth century. These were strongest at the prestigious theological schools such as Harvard, Yale, and Chicago—but rather dramatically not at Princeton University and Princeton Theological Seminary, from which some intellectual impetus for fundamentalism was to come.

Between 1910 and 1915 the conservative parties issued an influential set of booklets called *The Fundamentals*. From then on the accent on funda-

mentals grew. In 1919 a World's Christian Fundamentals Association was born. In 1920 during a battle for power in the Northern Baptist Convention a Baptist editor proudly chose and publicized the term 'fundamentalist' for his faction. The Presbyterians and to a lesser extent other mainstream Protestant denominations entered into the battle. The celebrated Scopes trial in Tennessee in 1925, one that featured conflict over the teaching of evolution in public schools, was the best known event in the formation and publicizing of fundamentalism.

This is not the place to trace in detail the outcomes of the struggles. Fundamentalists lost the battles for control, but inventively went their way developing their own Bible colleges, seminaries, mission boards and, thanks to schisms, denominations. They were a presence but not a vivid one during this period. In 1941 some formed an American Council of Christian Churches, which engaged in rivalry with a more moderate National Association of Evangelicals born in 1942. Still then seen by nonfundamentalists as marginal, eccentric, and withdrawn, fundamentalists remained generally aloof from politics. However, during the 1970s most of them reversed their stance and began to adopt political interventionist stances. They have been a prominent presence ever since.

To detail this case study further might well bias the case of those who charge that global and interreligious use of the term fundamentalism is too reliant on the American experience. But one could study cognate movements elsewhere—typically the Muslim Brotherhood in Egypt—and find similar reasons for Islamic conservative groups there to move into reaction. A description of elements that make up what is called fundamentalism in any religion will be more profitable, however, than would extending the number of case studies.

3. *A Phenomenological Approach to Fundamentalist Features*

By phenomenological here is meant a pattern of observation in which the analysts 'bracket' their presuppositions—in this case, for example, assumptions about the positive or negative aspects of fundamentalism—and as if naively asks: what stands out in the phenomena that have been called fundamentalist? Another way to put it is: what phenomena developed in the twentieth century that occasioned the need for fresh naming—since words like conservative and traditional had been easily available? These questions enable one to escape the charge that a priori definitions color the observations. These are some of those features. First, while fundamentalisms originally were dismissed by others as 'the old time religion' and while some parade themselves as traditional, they differ in significant respects from conservatisms. In the eyes of doctrinal, political, and sometimes even armed militant partisans, conservatives may be too passive. They are too content to let the world go about its business not reckoning with them. If in any particular instance the cause being challenged is seen as divine—and we are here dealing only with *religious* fundamentalisms—for anyone to be content with tradition is a betrayal.

Fundamentalisms rise when historically conservative religious movements experience something that challenges them to their core. In Iran and elsewhere in the Middle East after the late 1970s, the challenger might be 'the Great Satan' of Western imperialism, specifically in its American form. In hyper-Orthodox Israel the threat to Jewish identity itself appears to parties such as Gush Emunim to come from non-observant, compromising Jews who might even let some of the land of Israel slip away in diplomatic negotiations. In the US the challenge in the 1920s came in more doctrinal ways, through contests over biblical literalism. But in all these cases, the threat had to be seen as total; fundamentalists do not bother much about trivial issues.

Since the challenge was total, the response had to be drastic. Fundamentalisms characteristically are reactive. They rise among communities that have been acted upon, and they act back. Very often they use the very instruments that had presented the challenge. If a nation–state is the ominous force, they make efforts to win power, through the ballot if possible and through weaponry if necessary. (Let it be noted that not all fundamentalisms take on any form of military or terrorist guise. Some even shun the political arena if they feel that they can convert sufficient recruits from that worldly zone or win their way simply through cultural militancy, as with disputes over 'values' and media.) Again, fundamentalists tend to use some of the weapons that have brought change, whenever, for example, they take instruments of mass media, the agents that penetrated their homes and consciousness, and become sophisticated in using these to fight back.

Somehow, in every case, they react. Not to be reactive is to be not fundamentalist. George Marsden, the foremost historian of the American version, describes a fundamentalist as 'an evangelical who got mad.' In reaction, the fundamentalist seeks ammunition for defense and aggression. Here is where the fundamentals come in. Fundamentalists reach back to foundations, some in the Qur'an, some in Torah, some in the New Testament, still others in classic events of the past, and claim them for their own. In every case in their eyes there had to be some sort of golden moment or precious book to which arguing parties can resort in efforts to determine winners and losers.

Acting upon these fundamentals, members of the group then engage, as noted, in setting boundaries. They are often described as bearing a Manichaean cast. This means that the cosmos is divided into God versus anti-God, good vs. evil, our people vs. everyone else. They may not necessarily withdraw into enclaves,

but they will tend to use code-words, slogans, informal testing devices that help them screen out the most dangerous persons and forces: the moderates. The total outsider, the infidel, the Great Satan is useful for mobilizing forces, but is not a threat the way a compromising modernizer is seen to be.

With the setting of boundaries comes the claim that participants are an elect people, God's own instrument. They know something of where history is going; it has been revealed to them by sacred books in which they find the fundamentals. By the way, it is easier to be fundamentalist in 'religions of the book' such as Judaism, Christianity, and Islam, than it is in faiths such as Hindu and Buddhist, where no single divine scriptural revelation is foundational.

The elect people, called to represent God, must keep themselves pure, try to recruit others, and, to use the words from the American struggle in the 1920s, 'do battle for the Lord.'

The battles may vary in character. Thus in most cases in the Islamic world, the enemy has not been the Great Satan of the US. In Algeria, Egypt, and elsewhere, the foe is a regime seen as corrupt, secularizing, and compromising even though all its leaders are themselves practicing Muslims. In Israel, the parties called fundamentalist may engage in warfare against Arab Muslims, but they take most pains to define themselves over against compromising Jews.

Action may be hortatory, prophetic, political, military, or terrorist, though it is often only the latter two that draw the world's attention.

4. *Accounting for Fundamentalisms in Culture*

No scholar claims that modern fundamentalisms have no counterparts or analogues in the past within the various religions. All of them have times and forces of traditionalist reaction and conservative aggression. It is valuable to study these analogues. But the word fundamentalism has more appropriately been reserved for responses to what has been seen as modern. One way to put it: these are reactive movements that appeared after the modern academy thought there would be no more like it, that the religions that survived would be small, weak, and always compromising.

Psychologists have on occasion assessed persons in the movements or the movements and generalized about their mentality. Participants tend to be people who cannot tolerate ambiguity, paradox, contradiction, or doctrinal and moral 'fuzziness.' Yet it is difficult to reduce fundamentalisms to this psychological cast, given the fact that people with such predilections in one culture turn religious fundamentalist and those in others do not.

Political and social scientists perceiving fundamentalisms in culture often relate the rise of such movements to class conflict. Marxist interpretations, for example, can see them as futile resistances by people who take resort in other-worldly solace when they can do little to gain in this world. It is true that in many instances it is the poor and apparently powerless that turn to fundamentalism. But it is also true that those who react may do so from the comfortable confines of Princeton Theological Seminary and Princeton University, as the intellectual definers of early American fundamentalism did. And one will find fundamentalists of wealth in many an American city and note that some of these extravagantly fund fundamentalist projects. Class analysis does contribute to understanding the phenomena, but most scholars again resist reductionism of a sort that fundamentalists are 'nothing but' this or that.

Instead, they give evidence that they are complex responses and reactions to many kinds of cultural stimuli and as such manifest features that are best appraised with a variety of perspectives and disciplines. Among these would be humanistic efforts to bring to bear history, philosophy, and religious studies. These are, after all, self-described religious movements, and seen to be such.

One way to approach the study is to listen as anthropologists do for 'agent's description.' Someone who pays attention to, records, and coordinates the testimony of fundamentalists in many religious contexts will find that the issues that evoke response begin very close to the self of the believer. Diplomats, reporters, and scholars often begin at the furthest extensions of fundamentalisms, as when some engage in terrorism or revolutionary military action.

Instead, whoever is alert to fundamentalist witness will be indeed close to the ego's center. The believer begins by defending a world-view, including one that has been recently changed thanks to conversion. The world looks different to someone who is an initiate, 'born again,' in some versions, than it does to someone else who is distant from or quiet about religious impulses. Very close to that egocenter is the question of personal, social, and cultural identity? To whom do I belong? Whom can I trust? With whom can I check signals and from whom do I draw inspiration? Those who ask such questions in the erosive climate and culture of late-modernity find that identity becomes a major question. They also note that it is threatened by forces such as travel, modern mass communications, mass higher education, especially where the challenges of pluralism and relativism are fierce.

As close to the center of the person as are world view and identity is the issue of gender. Fundamentalists preoccupy themselves with what it means to be a man or a woman. The Manichaean differentiation is strong: there is no room for ambiguity in defining gender. The fundamentalisms, for example, all reject homosexual expression. From gender they move to defense of family. In all the cultures studied by fundamentalist scholars, 'family values' and shoring up the family, however defined, is central to the program.

From the orbits of gender and family one moves to other challenges. Thus, one's mind and habits are shaped by mass communications, and fundamentalists tend to react by trying to screen out uncongenial images, to engage in boycotts, or attempt legally to censor and exclude broadcast or published materials that they believe to be evil or subversive of their values.

Education is another zone in which fundamentalists do battle. They start their own schools to protect and nurture their own young and often engage in efforts to subvert the character of others', especially the public's schools, when their teaching and habits are offensive. Similarly, many of them have challenged modern medical practices, and make the clinic their arena of conflict. They organize against abortion, euthanasia, and many forms of treatment that violate their view of the body as God's temple.

In the sphere of economics it is harder to generalize. Some religious texts favored by fundamentalists—notably those detailing *shari'ah*, Islamic law—may be explicit was when they proscribe the taking of interest. But some fundamentalist movements promote communal, one might say communistic economic patterns, while others, notably in North America, often align themselves with *laissez-faire* economic forces. But when they perceive economic circumstances as a threat to them, they may adopt any number of strategies, and find their holy books less definining and definitive.

When it comes to efforts to change polities, for example, it is through the use of the instruments of political propaganda, the ballot, and constitutional revision, that fundamentalists find themselves reckoned with most strenuously. Where they cannot take over a regime, they may in God's name undertake revolutionary activity. It is activity of that sort that elicits headlines about fundamentalist militants around the world.

See also: Fundamentalism (New Christian Right); Islamic Fundamentalism; Religious Fundamentalism: Cultural Concerns

Bibliography

Ammerman N T 1987 *Bible Believers*. Rutgers University Press, New Brunswick, NJ

Arjomand S A (ed.) 1984 *From Nationalism to Revolutionary Islam*. State University of New York Press, Albany, NY

Arjomand S A 1988 *The Turban for the Crown: The Islamic Revolution in Iran*. Oxford University Press, New York

Caplan L (ed.) 1987 *Studies in Religious Fundamentalism*. State University of New York Press, Albany, NY

Cohen N J (ed.) 1990 *The Fundamentalist Phenomenon: A View from Within; A Response from Without*. Eerdmans, Grand Rapids, MI

Dekmejian R H 1985 *Islam in Revolution: Fundamentalism in the Arab World*, 1st edn. Syracuse University Press, Syracuse, NY

Dessouri A E H (ed.) 1982 *Islamic Resurgence in the Arab World*. Prager, New York

Entelis J P 1986 *Algeria: The Revolution Institutionalized*. Westview, Boulder, CO

Esposito J L (ed.) 1983 *Voices of Resurgent Islam*. Oxford University Press, New York

Friedman M 1992 *The Haredi (Ultra-Orthodox) Society: Sources, Trends and Processes*. Jerusalem Institute for Israel Studies, Jerusalem

Heilman S 1992 *Defenders of the Faith: Inside Ultra-Orthodox Jewry*. Schocken, New York

James W 1902 *The Varieties of Religious Experience*. Longmans, Green and Co., New York

Juergensmeyer M 1993 *The New Cold War? Religious Nationalism Confronts the Secular State*. University of California Press, Berkeley, CA

Kepel G 1985 *Muslim Extremism in Egypt: The Prophet and the Patriarch*. University of California, Berkeley, CA

Lawrence B B 1989 *Defenders of God: The Fundamentalist Revolt against the Modern Age*, 1st edn. Harper & Row, San Francisco

Lustick I S 1988 *For the Land and the Lord: Jewish Fundamentalism in Israel*. Council on Foreign Relations, New York

Marsden G M 1980 *Fundamentalism and American Culture*. Oxford University Press, New York

Martin D 1990 *Tongues of Fire: The Explosion of Protestantism in Latin America*. Blackwell, Oxford, UK

Marty M E, Appleby R S 1991 *Fundamentalisms Observed*. University of Chicago Press, Chicago

Marty M E, Appleby R S 1993a *Fundamentalisms and the State: Remaking Polities, Economies, and Militance*. University of Chicago Press, Chicago

Marty M E, Appleby R S 1993b *Fundamentalisms and Society: Reclaiming the Sciences, the Family, and Education*. University of Chicago Press, Chicago

Marty M E, Appleby R S 1994 *Accounting for Fundamentalisms: The Dynamic Character of Movements*. University of Chicago Press, Chicago

Marty M E, Appleby R S 1995 *Fundamentalisms Comprehended*. University of Chicago Press, Chicago

Mitchell R P 1969 *The Society of Muslim Brothers*. Oxford University Press, London

Piscatori J 1991 *Islamic Fundamentalisms and the Gulf Crisis*. American Academy of Arts and Sciences, Chicago

Rahman F 1982 *Islam and Modernity: Transformation of an Intellectual Tradition*. University of Chicago Press, Chicago

Sivan E 1985 *Radical Islam*. Yale University Press, New Haven, CT

Tibi B 1988 *The Crisis of Modern Islam*. University of Utah Press, Salt Lake City, UT

M. E. Marty

Religious Nationalism and the Secular State: Cultural Concerns

According to a widely held view, 'nationalism' is an aspect of the process of secularization, which limits the role of religious ideas and institutions in public life. Religious nationalism is therefore considered inimical to the 'secular state,' which is uninvolved with or neutral in religious matters. Empirical evidence how-

ever, reveals, that the relationship is complex and context sensitive. This article examines this complexity with special reference to South Asia where the two kinds of nationalism have long been in conflict and where both secular and religious states exist. Comparative perspectives from the West and the Middle East help frame the argument.

1. *European Origins of Nationalism and its Diffusion*

Although emphases differ, definitions of nationalism generally converge on the idea of a people's identification of themselves as a 'nation' that already is, or is expected to become, an autonomous state with its own territory. Moreover, the individual's emotional identification with the nation, and loyalty to the state, are of crucial importance. They find expression in the ideology of nationalism, which is a blueprint for the capture and use of political power for furthering common interests.

It has been argued that without nationalism there would be no nations or nation-states—that nationalism 'invents nations where they do not exist' (Gellner 1964, p. 169) and holds them together where they do. It is recalled that in Western Europe, the eighteenth century marked not only 'the dawn of the age of nationalism,' but also 'the dusk of religious modes of thought' (Anderson 1983). Ironically, secular nationalism everywhere, since the French Revolution, has displayed the characteristics of a religion both structurally (doctrine, myth, ritual, etc., are present) and functionally (an overarching moral order is provided). This is illustrated by the growth of 'civil religion' in America (Bellah 1970), transforming a historically developed national political system into a moral order.

In the articulation of nationalism, particular dimensions of collective identity (geographical, historical, cultural, or ethnic), or particular combinations of them, are highlighted. These emphases are variable not only from place to place, but also over time in the same place. One of the identity makers everywhere is religion, but religious nationalism is considered an anomaly, and not seen to be the most suitable cultural setting for a modern secular state in which individual identity is clothed in citizenship rights that override religious and ethnic distinctions.

One of the key questions about nationalism concerns its beginnings. Some scholars have argued for a Western European origin, whereas others maintain that England was the first modern nation ever. Still others consider North America to be its true place of efflorescence: the notion of the nation was implicit in that of 'the people' in the Declaration of Independence and is central to the American Constitution of 1789. Its inspiration is said to have been a unique combination of secularization, industrialization, capitalism, print technology, etc., that 'created the possibility of a new form of imagined community, which in its basic morphology set the stage for the modern nation' (Anderson 1983, p. 49).

The Western origins and secular character of nationalism have been questioned. Kohn (1955) argued for Hebrew and Greek origins of the concept 2000 years ago, but affirmed that, as a fully fledged ideology, it came into play in England in the seventeenth century. Hastings has written that the Bible presented 'in Israel itself a developed model of what it means to be a nation—a unity of people, language, religion, territory, and government' (Hastings 1997, p. 18). For him the nation and its ideology are 'characteristically Christian things' (Hastings 1997, p. 186).

One of the many strands of Hastings' argument is that since Christianity did not have a sacred language, the availability of the Bible for translation and the activities of the local clergy in different parts of Europe resulted in the fostering of a 'sense of shared local, provincial, or national identity' (Hastings 1997, p. 192). He maintains that nations and nationalism outside the Christian world bear witness to the 'imitation of the Christian world,' although the mimesis may be regarded as westernization (Hastings 1997, p. 186). According to Hastings, while 'Christianity has of its nature been a shaper of nations, even of nationalism; Islam has not, being on the contrary quite profoundly anti-national' (Hastings 1997, p. 187). Orthodoxy notwithstanding, Muslims have embraced nationalism in many places in the twentieth century, e.g., in Turkey, where it appeared as a secular ideology in the 1920s, and in India, where religious nationalism became the basis for the partition of the subcontinent.

2. *Secular and Religious Nationalism in South Asia*

2.1 *India, Pakistan and Bangaladesh*

For 100 years, beginning in the middle of the eighteenth century, British colonial territories in the Indian subcontinent expanded steadily through conquest and annexation. An image of Indians as a congeries of religious communities, castes, and tribes was assiduously cultivated. District gazetteers, compendiums of tribes and castes, and census reports recorded the divisions among the people. Such divisiveness was welcome from the colonial point of view as it helped strengthen the hold of the alien ruling power.

Ironically, however, British rule itself helped regenerate an Indian national consciousness, a prototype of which had been discernible in the heyday of the Mughal empire (seventeenth century). Establishment of law and order under the aegis of a centralized authority structure, in place of the chaotic conditions that characterized the disintegration of that empire in the first half of the eighteenth century, was one such

major factor. Also important were the introduction of railways, postal and telegraph services, education through the medium of English modeled on British schools and colleges, and newspapers in the nineteenth century. Resentment over the denigration of India's cultural heritage as 'an unmixed evil' (Henry Maine's phrase) by civil servants and Christian missionaries helped create a shared sense of common identity vis-à-vis the British. By the 1870s, 'a feeling of nationality,' in the words of a colonial magistrate, 'had sprung up in India' (see Mehrotra 1971, p. 144).

One of the main objectives of the Indian National Congress, established in 1885, was 'the fuller development and consolidation' of 'sentiments of national unity' (Mehrotra 1971, p. 413). The fact that it was attended by delegates from virtually all parts of the country, who included Hindus, Muslims, Parsis, and Christians, seemed credible evidence of (as a newspaper put it) 'the existence of something like a nation in India' (Mehrotra 1971, p. 419). Indeed, a year later, the highest British official in India, viceroy Dufferin, wrote of his sympathy for 'the aspirations of a nation so similarly circumstanced as my own [the Irish]' (Mehrotra 1971, p. 398).

The secular character of the Congress soon came under severe strain owing to the refusal of certain prominent Muslim leaders of northern India to join it. The position of the Congress was that 'the common benefit' of the people of India transcended the interests of any particular community, and that the two sets of interests were not necessarily in conflict. The emergence of secular nationalism had been preceded by the assertion of a sense of common nationality among Hindus, overriding significant regional and sectarian differences among them. Many of them brought residues of Hindu nationalism into the Congress strengthening Muslim fears.

Muslim separatism led to the formation of the Muslim League in 1906 with the aim of furthering the community's interests. In the course of time the viewpoints of the Congress and the League came to be designated as 'nationalism' and 'communalism,' respectively. Whereas the former considered territorial unity the crucial element of nationalism, the latter considered it irrelevant in Indian conditions and focused on religious identity.

The foregoing antagonism eventually led to the 'theory' that Hindus and Muslims were two nations, irreconcilably divided by history and culture, and each legitimately aspiring to be a nation-state. British rule ended in 1947 with the partition of the subcontinent into India and Pakistan. India created for itself a secular state in which citizenship rights stand above ascribed identities. Contrary to the stated secularist views of its founder, M. A. Jinnah (died 1948), Pakistan was proclaimed an Islamic state where traditional Islamic law would have precedence over other laws, and where Sunni Muslims emerged as the privileged community.

Pakistan was a unique state as it comprised two wings separated by 1,000 miles of Indian territory. The fragility of religious nationalism became obvious within a year of its founding in the context of a *lingua franca*. The people of East Pakistan, who accounted for more than half of the country's population, resisted the declaration of Urdu, the mother tongue of the urban Muslims of northern India, as the national language. Bengali was therefore accorded the status of the second language. Culture and language became rallying symbols for discontented East Pakistanis, who increasingly complained of economic exploitation and denial of political rights at the hands of a succession of 'central' governments that had their seat in West Pakistan and were dominated by the natives of that wing.

Eventually, Pakistan split in 1971, and a new nation-state, Bangladesh, came into existence. Bangladesh adopted a constitution that was secular in character. The original considerations that legitimized partition in 1947 survive, however, and are exploited by some political parties. Islamization of Bangaladesh's national culture, and the constitutional redefinition of the state as one that derives its legitimacy from Allah, emphasize the obstacles that confront the secular state in South Asia.

It has been argued that the enormous cultural and religious diversities of India necessitate a pluralist ethos and the secular state, and indeed guarantee them. The Constitution (Article 25) grants every citizen the fundamental right to profess, practice, and propagate the religion of her or his choice. It also prohibits discrimination between one citizen and another on the basis of religion. All major political parties that have held power at the state or union (federal) levels uphold the concept of the secular state. One of them, the Bharatiya Janata Party (BJP), is widely considered communal in view of its following, which is virtually exclusively Hindu, and its adherence to the ideology of Hindutva (lit. 'Hinduness'). Originally put forward in 1923 (BJP itself is a post-independence party), Hindutva equates Indian identity with Hindu identity, and defines the latter in terms of a combination of territorial and cultural criteria (fatherland and holy land should be one). The ideology points to political domination through cultural homogenization, and is hostile to pluralism. BJP emerged as the leading partner in the coalition union government that was formed in 1998. The exigencies of coalition politics have resulted in the de-emphasizing of Hindutva but not its abandonment by BJP. The ideology is synonymous with religious nationalism and therefore a threat to the secular state.

2.2 Sri Lanka

The Sri Lankan Constitution also is broadly secular in character. Although a 1972 constitutional amendment

recognized 'the foremost place' of Buddhism as the religion of the majority, there is no state religion, nor discrimination in respect of citizenship rights between the followers of different faiths (Buddhism, Hinduism, Islam, and Christianity). The fact that the ethnic Sinhalese constitute 70% of the population and are Buddhists has been a source of communal strife and has imposed severe pressures on the state.

Beginning in 1953, when the call for replacing the secular state by a Buddhist one was first made, politically organized Sinhalese have carried on campaigns against the Tamils of the northern and northeastern regions of the country (mainly Hindus and Christians) for their secessionist ambitions (the Tamils in turn complain of Sinhalese domination) and against a succession of governments for not being sufficiently responsive to Sinhalese aspirations. Prime Minister Solomon Bandaranaike, a convert from Christianity to Buddhism and more sympathetic to such aspirations than his predecessors, was nevertheless assassinated in 1959 by a Buddhist monk. Altogether the Sinhalese nationalists won many concessions during the 1960s and 1970s.

A watershed in the ethnic strife was the countrywide anti-Tamil upheaval of 1983 that had its epicenter in Colombo. It followed the killing of a dozen Sinhalese soldiers by Tamil militants. By official count several hundred Tamils were killed; the Tamils claimed that thousands died. Sinhalese militancy represented by Janatha Vimukhti Peramuna (JVP; People's Liberation Front) was eventually brought under control in 1989–90 by President Premadasa (himself a Buddhist) through accommodation of sections of the Buddhist leadership and ruthless suppression of JVP cadres. But he, too, paid with his life for his resistance to the secessionist movement led by the Liberation Tigers of Tamil Eelam (Nation).

3. Comparative Perspectives

The conflict between religious nationalism and the secular state is vividly illustrated by the wide range of events in South Asia in the second half of the twentieth century. Religious nationalisms are present elsewhere too in confrontation with a secular state or in association with a religious state that is modern in some of its objectives (economic development, political stability) and methods (elective government).

The Iranian revolution of 1978 established an Islamic state after overthrowing the modernist regime of the Shah. The earlier ouster of King Farooq of Egypt in 1952 had been followed by the establishment of a secular government under Gamel Nasser, which set out to suppress religious nationalist groups such as the Islamic Brotherhood. Although his successor, Anwar Sadat, made many concessions in their favor, he died at the hands of Muslim extremists. Religious fundamentalism and nationalism are a considerable combined force, but President Mubarak has so far succeeded in holding it at bay. In Turkey, too, the secular state founded by Kemal Ataturk remains intact but religious revival has made significant inroads into Turkish society and nationalism. Although the patterns vary, religious nationalism today is a force to reckon with among the Muslim nations.

Religious nationalism played a significant role in the disintegration of the former USSR and the emancipation of East European countries from totalitarian socialist regimes. It obviously is not an essentially reactionary force in all circumstances. The comparative perspective brings into question many assumptions about the western origins and secular foundations of nationalism. Religious nationalism is a mixture of diverse elements, not all of which are present everywhere. Some of them are in principle unexceptionable, such as the emphasis on the moral dimensions of political behavior. The overall record of religious nationalists, however, is marked by their opposition to pluralism, which is one of the primary conditions of humane governance, particularly in ethnically plural societies.

See also: Religion: Nationalism and Identity; Secular Religions; Secularization

Bibliography

Ahmed R (ed.) 1991 *Religion, Nationalism and Politics in Bangladesh*. South Asian Publishers, New Delhi, India

Anderson B 1983 *Imagined Communities, Reflections on the Origins and Spread of Nationalism*. Verso, London

Bellah R N 1970 *Beyond Belief: Essays on Religion in a Post-traditional World*. Harper and Row, New York

Chandra B 1984 *Communalism in Modern India*. Vikas, New Delhi, India

Gellner E 1964 *Thought and Change*. University of Chicago Press, Chicago

Gellner E 1983 *Nations and Nationalism*. Cornell University Press, , NY

Hastings A 1997 *The Construction of Nationhood, Ethnicity, Religion and Nationalism*. Cambridge University Press, New York

Hobsbawm E J 1990 *Nations and Nationalism Since 1780*. Cambridge University Press, Cambridge, UK

Jaffrelot C 1996 *The Hindu Nationalist Movement in India*. Columbia University Press, New York

Juergensmeyer M 1993 *The New Cold War? Religious Nationalism Confronts the Secular State*. University of California Press, Berkeley, CA

Kohn H 1955 *Nationalism: Its Meaning and History*. Van Nostrand, Princeton, NJ

Mehrotra S R 1971 *The Emergence of the Indian National Congress*. Vikas, New Delhi, India

Tambiah S J 1992 *Buddhism Betrayed? Religion, Politics, and Violence in Sri Lanka*. University of Chicago Press, Chicago, IL

T. N. Madan

Religious Organizations

Religious organizations seek to establish, promote, and regulate relations between human beings and divinities, supernatural orders, or supreme metaphysical principles. This article explains why religious organizations have been distinguished from nonreligious organizations and why their distinctiveness lies mainly in their varied patterns of authority and leadership.

1. Diversity

Religious organizations are immensely varied but they typically aim to promote worship, prayer, meditation, teaching, healing, and spiritual well-being in accordance with authoritative revelations, texts, codes, laws, and principles. They range from the groups of 'clients' who regularly consult self-appointed healers, gurus, and shamans to worldwide ecclesiastical bureaucracies such as the Roman Catholic Church. Informal and irregular religious practices can be nested inside highly formal religious organizations.

The distinctive 'products' of religious organizations are equally diverse: sacred knowledge, transcendental experiences, prophecies, worship and meditation, rituals with the power to cleanse or heal, rites of passage, religious ethics, recruiting missions, political, welfare, and educational outreach, local fellowship, and so on. They also have much in common with nonreligious organizations in so far as they face the need to secure adequate resources, to train and control staff, to retain their authority, to cope with dissent and conflict, to interact with other organizations in their environment, to defend their share of the religious 'market,' and to deter 'free riders,' and unwelcome takeovers. Religious organizations have therefore established schools, seminaries, hospitals, welfare agencies, publishing enterprises, and missionary orders in order to extend and protect their boundaries.

2. History

The history of studies of religious organizations can be divided into two streams. The first is concerned primarily with the distinctiveness of religious organizations and with their subtypes. The second, which has flourished since the 1950s, deals with the internal dynamics and shifting fortunes of religious organizations in their social environments.

2.1 Types and Subtypes

The intellectual rationale for the pioneering and highly influential studies written by Max Weber (1864–1920) and Ernst Troeltsch (1865–1923) had its roots in their respective interests in the interplay between modernity and religion. Having established the crucial but unintended contributions of Ancient Judaism and Christianity towards the forms of modern rationalization, Weber analyzed the ironic implications of rationalization for the differentiation of religion from other institutions and value spheres. Secularization, bureaucratization, and the disenchantment of the modern world were all shown to have reduced the cultural and social significance of religious organizations. Weber perceived clearly that church-type organizations (large, territorially based, and inclusive in membership terms) had the potential to exercise power but were constrained by their ties to the secular world from translating their Christian ethics into uncompromising programs of action. By contrast, sect-type organizations (voluntaristic, relatively small, and exclusive in membership terms) were capable of maintaining a distinctive and often radical version of Christian theology and ethics but were prevented from exercising much influence on the world around them because they distanced themselves from it.

Troeltsch was less preoccupied with relationalization as a process and more concerned with the historical evolution of the social forms of 'church,' 'sect,' and 'mysticism' in which the New Testament's ethical message had found expression. Troeltsch was particularly exercised by the question of whether liberal Protestantism was robust enough and sufficiently well adapted to the conditions of modernity to serve as an effective vehicle for Christian ethics in an increasingly secular world.

The concepts of 'church' and 'sect,' plus numerous variants and subtypes such as 'established sect' and 'denomination,' have played an important part in many attempts to explain patterns of change in Christian organizations throughout the twentieth century. Some scholars have tried to apply these concepts to the analysis of religions other than Christianity, but it remains questionable whether this is appropriate in the case of religions lacking strong notions of membership and corporate authority. Yet, refinements of the sect concept have undoubtedly helped to explain the highly varied developments of minority religious movements, not only in Western societies but also in East Asia. Bryan Wilson's (1970) scheme of seven subtypes of sects has proved particularly fruitful and influential. The concept of 'cult' has also undergone extensive refinement, coinciding initially with public interest in the new religious movements which became controversial in many Western countries and Japan in the 1970s.

A further impetus for sociological studies of 'sects' and 'cults' began to develop in the 1980s when

proponents of rational choice theory claimed that previous usage of these concepts has been largely *ad hoc* and untheorized (Iannaccone 1988). On the assumption that people rationally try to attain their ends by the least costly means, in religion as in any other area of life, the theory holds that people incur the costs of participating in religious organizations in order to derive supernatural compensators which purport to give answers to questions of ultimate meaning. In this theoretical perspective, churches are believed to offer answers which conform with prevailing norms, whereas sects and cults allegedly deviate from the norms and incur sanctions or costs which are nevertheless experienced by their followers as selective benefits unavailable to members of 'lax' churches. The historical formation of sects (by schism) and cults (by innovation) can be accounted for in rational choice terms as a never-ending process of competition between religious organizations standing in different degrees of tension with society.

By comparison, the sociological concept of 'church' underwent relatively little change in the twentieth century. Moreover, little interest has been shown in refining or even applying the kind of cognate concepts of 'ecclesiola in ecclesia,' 'fraternitas,' and 'order' that Joachim Wach (1898–1955) and others elaborated. Nevertheless, studies of Christian organizations in Africa, Asia, South America, and the former Soviet Union have brought to light the importance of such organizational forms as 'independent churches,' 'base ecclesial communities,' and 'underground churches.' The continuing significance of relations between religious organizations, nations, and states in some countries has also preserved a place for 'church' in analyses of, for example, civil society (Casanova 1994) or international relations (Johnston and Sampson 1995). And extensive research in North America has thrown light on numerous aspects of 'denominations' as distinctive organizations. Beginning with the embeddedness of denominations in social classes (Niebuhr 1929) and ethnic groups, studies have analyzed the organizational conditions of schism, the ambiguous position of theological seminaries, the differential responsiveness to female clergy, the increasingly complex webs of 'interorganizational fields' in which denominations operate, and tensions between religious hierarchies and functional agencies. Another rich vein of research has investigated the contributions of denominations towards ideologies of pluralism, pluralistic provision of social welfare, social movements, and the 'independent sector.'

The study of religious organizations has been understandably less important in relation to religious traditions which tend to lack the formality, corporate authority structures, and membership principles characteristic of Christianity. Nevertheless, common foci of scholarly interest have emerged in the study of Muslim 'orders,' brotherhoods, and charitable endowments (*waqf*); Buddhist monastic institutions (*sangha*); Hindu devotional groups (*bhakti*) and nationalist organizations; Jewish and Sikh communal organizations; and Shinto parishes.

2.2 Dynamics of Religious Organizations

The emerging concern with the organizational dynamics of religion in the 1950s was partly a reflection of the growing salience of 'organization theory' in the burgeoning social sciences, and partly a function of Christian theological interests in 'ministry' as a modern profession and in churches as complex organizations responding to rapid social change (Harrison 1959, Winter 1968, Beckford 1975). Yet the impact of these analyses of religious organizations on the development of organizational theories in general was modest. This state of affairs began to change slightly in the mid-1970s when notions such as 'organizational climate' and 'organizational culture' directed attention towards the significance of cultural factors which had always loomed large in analyses of religious organizations but which were emerging as important influences in virtually all kinds of organization. Investigations of communes gave a further impetus to research into the force of strong commitments and normative beliefs among organizational participants. Subsequent discoveries about the importance of rituals, retreats, flexibility, and networking in many organizations have only enhanced the interest that religious organizations now have for some organizational analysts (DiMaggio 1998). This minor rapprochement between the study of formal organizations and the study of religious organizations received a further boost in the 1980s when so-called New Age and quasireligious business began to attract attention for their practice of combining profit seeking with ethical business practices and of contributing towards the 'spiritual development' of their staff and clients. And fresh insights into the tensions between the 'religious' and 'agency' structure of US denominations (Chaves 1993) offer the prospect of explaining comparable tensions between the normative and functional bases for coordination and compliance in other organizations as well.

Two inter-related aspects of religious organizations stand out as the central foci for past and present research: authority and leadership. This does not mean that issues of organizational effectiveness, efficiency, and complexity are irrelevant in the case of religious organizations. It simply means that organizations purveying allegedly timeless truths and the highest normative commitments tend to be relatively more preoccupied than nonreligious organizations with establishing, reproducing, and defending their teachings and practices against corruption, subversion, or apathy.

3. Authority

Like profit-seeking organizations, religious organizations are engaged in competition with purveyors of rival versions of ultimate truths and in conflict with opponents, but they are also obliged to monitor and police their own members or followers. This is why questions of authority are central to religious organizations and why change often meets with strong resistance. In this sense, religious organizations are distinctive, if not unique, for the extent to which the connections between their founding truths and their organizational structure and everyday activities are constantly rehearsed in rituals, symbols, exegesis of sacred texts, preaching, prayers, teaching, meditation, and the training of leaders.

3.1 Sources of Authority

Christianity is virtually unique among the world's religions for the extent to which its practice has been controlled by centralized, formal organizations cutting across ethnic, tribal, and political groupings. By contrast, Hinduism, Judaism, and Shinto crystallized slowly out of predominantly tribal and ethnic cultures, giving rise to temples, synagogues, and shrines as centers of rites and ritual worship rather than as agencies of civil administration. Diasporic Judaism, lacking the authority previously vested in the Temple at Jerusalem and its hereditary priests, regrouped around rabbis as teachers of the law and around synagogues as ritual centers. Communal organizations have played a major role in preserving the ethnic and religious identity of Jews under persecution, but centralized organizations to control Jewish beliefs and rituals are relatively weak.

Buddhism and Islam emerged from more deliberate attempts to cultivate religious and philosophical systems partly at odds with prevailing culture and social arrangements, and this feature of their historical emergence may help to account for the fact that they have been closely allied with political regimes at various times. Thailand and Japan, for example, have had something approaching Buddhist parish systems. Modern Japan has also seen the rise of massive, hierarchical organizations of Buddhist laypeople such as Risshō Kōsei-kai and Sōka Gakkai. But elsewhere in Asia, Buddhism functions like a diffuse folk religion supported by monastic institutions and temples.

The organization of Islam is subtle, multistranded and variable by region of the world. As a conversion faith Islam has been interwoven with political regimes, but it does not have a single, authoritative organization. There are multiple, competing sources of theological, legal, spiritual, and moral authority in Islam, for the authority of learned experts, jurists, hereditary sheikshs, leaders of brotherhoods and clergy (among the Shi'ites) are all part of the complex organizational tapestry of Islam. This high degree of complexity makes it difficult for any single organization to claim convincingly that it represents all Muslims. Now that tens of millions of Muslims live as religious minorities in the predominantly Christian countries of Europe and North America, pressure may be increasing for would-be representative organizations to seek to represent all Muslims in their collective relations with nation states, nongovernmental organizations, and the European Union.

3.2 Forms of Polity

The variety of organizational forms and processes which translate these concerns with authority into action is bewilderingly wide. They range from loosely structured fellowships to vastly complex and formal organizations which require their members to accept centrally imposed dogmas and rituals mediated by authoritative officials. In practice, most religious organizations fall somewhere on a continuum between these two extreme positions, but their precise location may change in response to factors internal or external to the organizations. Moreover, the time-honored distinction between church- and sect-type organizations cannot be mapped on to the distinction between religious organizations with either strong or weak systems of authority.

The convention of classifying the polity of Christian churches as Episcopal, Presbyterian, or congregational relates to the distribution of authority as laid down in their respective theologies and legal constitutions. In particular, the distinction between ordained clergy and laity is given theological justification, and the differential authority accorded to different grades of religious professionals is explained in the same terms. Yet this classification can only be a starting-point for social scientific analysis, since the readiness of subordinates to comply with authoritative commands form their superiors is variable, and pressure from outside the churches may make it necessary to modify the practical implementation of the ecclesiastical principle. For example, the Episcopal polity of churches in the worldwide Anglican communion conceals a wide diversity of relations between bishops and other church members. Presbyterian and congregational polities have also come under pressure to adapt the exercise of authority to prevailing sociocultural values and conditions. Thus, critics of the patriarchalism allegedly underpinning many forms of Christian organizational polity have challenged the gendered distribution of authority in churches and tried to infuse them with feminist, antihierarchical values (Wallace 1992, Nesbitt 1997). Moreover, the ideal of congregational autonomy finds itself in tension with political or economic pressures to extend the authority of regional or national officials.

3.3 Fundamentalism

The increasing popularity and influence of so-called fundamentalist forms of religion roughly since the 1950s in many parts of the world and in most of the major faith traditions has prompted social scientists to pay greater attention to religious authority (Marty and Appleby 1991). Fundamentalism centers on faith in the authority of certain principles, usually enshrined in sacred texts, to generate unequivocal guidance for the conduct of personal and public life, and it favors assertive leadership and a strong suspicion of anything outside the group of committed believers. Yet the organizational bases of fundamentalism are diverse and subtle. There is no single, monolithic organization in any faith tradition or country which promotes fundamentalism. Instead, it has spread by personal witness, small group discussion, and the use of computerized mailing lists, terrestrial and satellite television and radio broadcasts, video cassettes, email, and the World Wide Web. Fundamentalists utilize modern technology and organizational forms in order to transform the world, and are not traditionalists, 'antimodern,' or withdrawn from the world. On the contrary, their aim is to make the world conform with their authoritative principles, and this calls for forms of organization which are flexible and responsive to local circumstances.

4. Leadership

4.1 Roles

Leadership roles, structures, and training vary widely between faith traditions, between their different 'schools' or denominations and between different parts of the world. The precise pattern of leadership at any level is shaped mainly by prevailing ideas about the source and nature of sacred power. Leaders of religions which identify sacred power exclusively with scriptures and laws derive their authority mainly from the capacity to read and interpret them. Judaism and Islam exemplify this pattern of leadership by scholars, teachers, and jurists. Priesthood, by comparison, is found in religions such as Hinduism, Shinto, Christianity, and Sikhism which combine reverence for sacred texts with belief in the capacity of priests to perform rituals allowing humans to interact in various ways with divinities and sacred power. The office of a priest enables this power to be effective through properly conducted rituals. Prophets and charismatic religious leaders may reform religion by opposing institutionalized practices and beliefs with radically new or different ones. Ministers of religion in most of the Protestant traditions offer leadership mainly through their preaching, exposition of scripture and, at least in evangelicalism, efforts to persuade individuals of their sinfulness and urgent need to seek redemption. Monks, nuns and 'holy men,' as the 'virtuosi' in some religious traditions, exercise leadership on the strength of their extraordinary discipline, devotion, and selflessness. The *sangha*, or monastic community, remains at the heart of Buddhism in many parts of the world; monastic institutions were the bulwark of Christianity throughout the European Middle Ages; Catholic religious orders were in the forefront of missionary work in Africa, Asia, and South America in the modern era; and monks continue to occupy the highest leadership positions in Orthodox churches. But religious traditions which locate authority in the spiritual freedom of individuals to cultivate their own, relatively independent path to the sacred tend to resist strong notions of leadership, preferring at best to have spiritual exemplars, teachers, or gurus.

4.2 Local Groups

It is common for religious leaders to be attached to a particular building or social center where they conduct teaching, preaching, meditation, workship, prayer, or pastoral functions. Some also have responsibility for the spiritual oversight of institutions such as schools, seminaries, orphanages, hospitals, refuges, and pilgrimage sites. Christianity went further than other religious traditions in establishing a comprehensive system for delivering religious and pastoral services to geographically demarcated areas known as parishes, which represented subdivisions of national or international churches. The idea that the residents of a locality would belong to a parish and form a congregation of regular worshipers under the leadership of local clergy was especially effective in European regions where the dominant church was closely aligned with political regimes.

4.3 Congregations

Congregations are the basic unit of most Christian organizations in the sense of being the social group with which their members in particular localities feel most closely associated. The precise status of congregations varies with the theology and authority structure of the larger religious organizations to which they belong. The congregations of hierarchical churches are local outlets of ideas and practices controlled by regional, national, and international authorities such as bishops, cardinals, and the Pope. By contrast, religious organizations which have little or no theological justification for hierarchy or central control tend to regard local congregations as relatively autonomous units. This is common among, for example, independent fellowships of conservative Protestants, the Reform Synagogues of Great Britain, and the Christadelphians. The implications of the balance that Christian organizations achieve between centrali-

zation and local autonomy have been extensively researched because it reveals differences in the extent to which congregations can generate resources, mobilize members in campaigns, respond to internal problems, and influence public life (Hong and Roozen 1979).

As Hinduism, Islam, Judaism, and Sikhism are not structured on the basis of inclusive, unitary organizations at national or international levels they have not generated congregations or systems of parishes. Yet, temples, mosques, synagogues, and gurdwaras do function as ritual and pastoral centers for people living in their vicinity or choosing to visit them. Lacking strong integration into higher level organizations, however, these local centers are accountable either to their private owners or to their management committees. Consequently, their religious leaders are retained as professionals or qualified volunteers by local groups rather than as employees of ecclesiastical organizations at a higher level.

4.4 Recruitment and Training

In advanced industrial societies the attractions of work in positions of religious leadership are weakening, and the rate of applications for training as religious professionals has been declining since the middle of the twentieth century. At the same time, the characteristics of recruits to religious leadership in Christian churches have also changed. They now include a much greater proportion of women, of people over the age of 30, of people retiring early from other occupations, and of people from a wider range of ethnic and social class backgrounds. The 'feminization' of religious leadership is occurring in many churches in so far as the proportion of women entering theological training and Christian ministry has been rising steadily since the 1960s (Wallace 1992). Yet the upper echelons of most churches are still controlled by men, and there is evidence that various mechanisms still prevent women from enjoying genuinely equal opportunities for professional advancement. Women are even more strongly excluded from leadership roles in the other world religions, although they have been admitted to the rabbinate in some traditions of Judaism.

It is only in developing countries that there is still a relatively strong supply of young men and women for theological training and preparation for the ministry or monastic vows. The disproportionately large numbers of trainee priests and ministers originating in developing countries may mean that the organizational 'center of gravity' of some large Christian organizations is in process of shifting away from Europe and North America. In effect, the new international division of labor is affecting religious organizations as well as transnational businesses.

5. Questions About the Future

Three emerging trends in religious organizations call for special comment. The first concern settlers in countries of large-scale immigration who seek to establish their own faith communities or congregations (see Warner and Wittner 1998). How far will these communities conform with prevailing forms of religious organization in their new setting? How far will their forms of organization come under pressure to meet the criteria laid down by stage agencies if they wish to deliver health, welfare, or educational services? And to what extent will their forms of organization influence longer established religious groups?

The second concerns the effect of new information technologies on the capacity of religious organizations to extend their influence to virtually all parts of the world. Will satellite broadcasting, video recordings, and the Internet increase competition between religious organizations? And will they cultivate a globally standardized form of these organizations?

The third emerging trend is for the appeal of 'special purpose' agencies to cut across the boundaries between Christian congregations and between denominations in the USA, thereby making ideological differences *within* religious organizations more challenging than the differences *between* these organizations (see Wuthnow 1988). Is there evidence of this kind of restructuring in other world religions? Does this repositioning of organizational 'fault lines' facilitate or hinder transnational religious developments?

See also: Civil Religion; Education: Cultural and Religious Concepts; Religion: Definition and Explanation; Religion: Evolution and Development; Religion: Morality and Social Control; Religion, Sociology of

Bibliography

Beckford J A 1975 *Religious Organization*. Mouton, The Hague

Casanova J 1994 *Public Religions in the Modern World*. University of Chicago Press, Chicago

Chaves M 1993 Intraorganization power and internal secularization in Protestant denominations. *American Journal of Sociology* **99**: 1–48

DiMaggio P 1998 The relevance of organization theory to the study of religion. In: Demerath N J III, Hall P D, Schmitt T, Williams R H (eds.) *Sacred Companies. Organizational Aspects of Religion and Religious Aspects of Organizations*. Oxford University Press, New York

Harrison P M 1959 *Authority and Power in the Free Church Tradition*. Princeton University Press, Princeton, NJ

Hoge D R, Roozen D A (eds.) 1979 *Understanding Church Growth and Decline 1950–1978*. Pilgrim Press, New York

Iannaccone L R 1988 A formal model of church and sect. *American Journal of Sociology* **94**: 241–68

Johnston D, Sampson C (eds.) 1995 *Religion, the Missing Dimension of Statecraft*. Oxford University Press, New York

Marty M E, Appleby R S (eds.) 1991 *Fundamentalisms Observed*. University of Chicago Press, Chicago
Nesbitt P D 1997 *The Feminization of the Clergy in America*. Oxford University Press, New York
Niebuhr H R 1929 *The Social Sources of Denominationalism*. Holt, New York
Wallace R A 1992 *They Call Her Pastor*. SUNY Press, Albany, NY
Warner R S, Wittner J (eds.) 1998 *Gatherings in Diaspora. Religious Communities and the New Immigration*. Temple University Press, Philadelphia, PA
Wilson B R 1970 *Religious Sects*. Weidenfeld and Nicolson, London
Winter G 1968 *Religious Identity*. Macmillan, New York
Wuthnow R 1988 *The Restructuring of American Religion*. Princeton University Press, Princeton, NJ

J. A. Beckford

Religious Specialists

'Religious specialists' is a generic term referring to select individuals or a category of individuals within different world faiths who undertake a particular responsibility for religious functions such as teaching, contemplation, transmission of scriptures, leadership, pastoral care, or ritual, on behalf of the religious community they serve. The nomenclature used to describe religious specialists differs according to the various religious traditions, but the idea of prophets, teachers, a priesthood and/or contemplative monastic orders is common to many world religions. Most religious specialists either will have inherited their role or will have responded to a sense of Divine calling to a particular and often strict way of life and adherence to religious disciplines.

1. *The Scope of Different Religious Specialists*

A community of people who share a particular ontological understanding of the world have often been inspired by prophets or religious teachers who have acted as intermediaries between the Divine and earthly realms. The prophetic role has often involved receiving revelations, and acting as an exemplar of these revelations. As religious communities grow in size numerically and geographically, and become increasingly complex, a need arises for individuals with particular knowledge and authority to oversee the religious belief and practice of the group. The need for a range of different religious specialists may become apparent, e.g., to regulate religious activities, to preserve and recite scriptures, or to lead worship, and in time, a hierarchy of different specialists with a specific authority and leadership may emerge. Access to particularly significant holy sites and shrines, or sacred texts, may be limited in some traditions to religious specialists only. In many ways, prophets are the direct opposite of priests, since the former are often a force for change in society, challenging existing practices and beliefs, while the latter serve as representatives of a religious community and seek to maintain existing practices. Charismatic religious specialists in some ways stand between prophets and priests, since they are often the source of new ideas and practices, but within existing frameworks of community and belief.

In the major monotheistic religions of the world today, Judaism, Christianity, and Islam, there is a range of religious functionaries. Despite the common link between these three faiths, there are some significant differences in the various types of religious specialists that have emerged in each. In Judaism, a rabbi does not, strictly speaking, have a priestly role, but rather a responsibility for teaching and spiritual guidance. Historically, alongside the Hebrew prophets, were sages or wise men whose writings comprise, among others, the Book of Proverbs.

In Christianity, there are two principal avenues for professional religious activity, namely pietistic contemplative, and priestly. Amid the diversity of Christian traditions, the Orthodox, Catholic, and Anglican denominations in particular have a history of contemplative monastic orders for both men and women. These orders have often been inspired by the examples of saints such as St. Francis of Assisi, and members of these religious communities aspire to follow the exceptional qualities of their founders. Other religious specialists in Christianity include priests or clergy whose role often combines a range of functions, such as pastoral care, leading worship, and teaching, generally focused around a church or place of worship.

In Islam, there is no such priestly equivalent, but there is nevertheless a range of religious specialists. An imam may be regarded primarily as a ritual specialist with particular responsibility for leading congregational prayers in a mosque. The title also has honorific connotations, perhaps in denoting the leader of a particular community or group. Among Shi'ite Muslims, the term imam has yet another meaning as an intercessor between Allah and humanity. The framework for Islamic life and conduct, individually and communally, is shaped by the Shari'ah, or Islamic law. Scholars with a particular knowledge of Islamic law and teachings, known collectively as ulema (singular alim) assist with the development of Islamic thought, while muftis are legal functionaries empowered to make fatwas or legal rulings. In the Sufi or mystical tradition of Islam, the spiritual leader of a community is often known as a sheikh.

Even between the three major monotheistic traditions, there is a clear diversity in the type and remit of different religious specialists.

Indian religions, especially Hinduism and Sikhism, share the concept of a guru or teacher. Sikhism is

founded upon the teachings of 10 successive Punjabi gurus during the sixteenth and seventeenth centuries, while in Hinduism, the term guru has a more generic connotation as a religious teacher who gives personal instruction to devotees. Brahmans are at the top of the Hindu caste system, and as religious professionals their role principally involves the transmission of sacred Sanskrit texts and the performance of priestly sacrificial rituals.

The monastic life is perhaps the defining feature of religious specialists in the many traditions of Buddhism, whether Tibetan, Zen, Theravadin, or Japanese. Not surprisingly, therefore, many of the titles associated with Buddhist religious functionaries relate to the hierarchy of the monastic order. With Buddhism becoming increasingly popular in the West, different titles such as Ajahn, Bhikkhu, Roshi, Lama, and Rinpoche are becoming more widely known and understood.

In traditional societies and communities, shamans are religious specialists with the supernatural power to contact the spirit world during ecstatic trance states, often induced through dance, intoxicants, drumming, or chanting. They are not priests, nor medicine men, but are said to have the ability to foresee and predict the future, to quell the forces of the spirit world, and undertake extreme privations without feeling pain.

Despite the wide variations among religions in terms of religious functionaries and their various roles, there are some common sociological features of professional or specialist religious activity.

2. *Becoming a Religious Specialist*

Those who dedicate either all or part of their lives to religious work are often regarded as having a vocation rather than a job. Social scientists are not so concerned with the theological or supernatural implications of vocation, but rather with visible and measurable aspects of professional religious life, such as recruitment and training. In his study of American Catholic clergy, Fichter notes that 'the ecclesiastical vocation is not only a supernatural calling, a state of life and social status. It is also a specialised social role through which the manifold functions of the Church are performed. Obviously then, he (the priest) is engaged in a full-time profession or occupation which can be submitted to sociological comparison and analysis' (Fichter 1961).

2.1 *Recruitment and Training*

Personal connections and example often have a bearing on recruitment to professional religious life. This partially explains why vocations often cluster in families and kinship groups. In some world religions, religious roles are inherited through a lineage of teachers or leaders. Entering professional religious service is a lifelong commitment in many instances, while in others an individual may through tradition or personal choice undertake to become a religious specialist for only a short period of time. A good example of the latter can be found in the Theravadin Buddhist tradition where it is common for young boys of school age to be ordained as monks for a limited period during their youth.

Most religious traditions have well-established and well-known expectations and qualifications for people interested in becoming religious functionaries, either in terms of knowledge or ethical conduct. As a step towards learning to become a religious specialist, it is common to undergo a period of isolation, training, and/or renouncing of worldly attachments. In both Buddhism and Christianity, those wishing to become monks or nuns often undertake a trial period of training as novices prior to ordination. In religious traditions where scholarly knowledge is a particular feature of professional religious activity, new recruits will often undertake an extensive period of specialist training to acquire the necessary knowledge and skills. Therefore, theological colleges or seminaries are common to the monotheistic traditions. Training institutions may be geographically isolated, or may be fully integrated into local communities, but the shift towards the latter as the locus for professional religious development is becoming more evident. This is partly due to the closing of the gap between lay communities and religious professionals, and the concern that communities have for specialists who are relevant and effective in modern societies.

2.2 *Ordination and Other Rites into Professional Religious Life*

Having successfully completed the necessary training period, whatever form that may take, some form of ceremony, initiation, consecration, vow-taking, or ordination into the religious profession is common. A number of symbolic actions or practices may take place, most of which have the function of setting apart the new religious specialist from the lay community. For example, the wearing of religious garb, such as a monastic habit or robe, or the marking of the body in some way, serves both ecological and structural purposes in 'locating' the individual within a community of fellow specialists. At the same time, a new name or religious title may be taken, and in some religious traditions, kinship terminology may be prevalent, such as 'Brother' or 'Sister.' These actions and rituals are part of the process of transformation into a socially recognized religious specialist.

3. *Religious Specialists—Some Sociological Characteristics*

Despite becoming socially recognized religious adepts, in many traditions it is common for religious specialists

to have a spiritual guide or teacher. This is a reflection of the life-long commitment that individuals make to their vocation, and the pursuit of spiritual development and perfection. The balance between, or combination of, self-centered and socially-centered spirituality and action will depend upon the characteristics of the professional role. Often there is no such thing as 'retirement' for religious professionals; they may retire from active functions certainly, but the nature of their calling entails a continuation of the religious role until death.

3.1 Status and Prestige

Most religious professionals share the social status of the group to which they belong. For example, the prestige and social standing of religious specialists in immigrant minority groups tends to rise or fall as a reflection of the upward or downward mobility of the migrant group. However, religious specialists who have subsidiary secular functions, such as social work or healthcare, may acquire recognition and status through these roles, rather than in their primary religious capacity. In some senses, religious specialists are classless individuals since generally they do not acquire status through earning nor, in the case of celibates, through favourable marriages. This does not mean, however, that they lack the possibility of status, since they can acquire other forms of recognition and prestige such as authority and other attributes of stature and power.

3.2 Professional Qualities and Hierarchies

Religious specialists can be considered as 'professionals' on a number of grounds. They are involved in a nonstandard product—religion. They undertake tasks and services that often cannot be repeated exactly, their work involves personal commitment and dedication, and they are required to have a broad knowledge of a specialized field and have authority both inside and outside the religious community. As specialists, they are often aspiring to personal development and improvement, and generally they share solidarity with other religious professionals. Like most professional groups, religious specialists, particularly in community contexts, are part of a hierarchical structure. Age, length of service, or the responsibilities particular to a role or office often determine hierarchy. The specialists who also have a leadership function face particular challenges in securing conformity, maintaining authority, and religious management of internal and external goals. The most senior members of the hierarchy generally have responsibilities that extend beyond the immediate community to a wide geographical area.

3.3 The Proliferation of Religious Specialists

In modern secular societies there has been a proliferation of work specialization. The study of careers in themselves has become a further branch of knowledge and expertise. The religious professional is yet another specialized career, and within it, there are further specialist activities, such as teaching, missionary work, or care of the sick. For example, some clergy within the Christian church feel called to chaplaincy roles often requiring a particular concern for groups of people in some way set apart from the rest of society, e.g., through imprisonment or hospitalization. The world of chaplaincy has in turn become more complex as society has become more fragmented and 'full-time personnel of the Church have been carried along on a wave of extra-ecclesiastical professionalism' (Fichter 1961).

In most cases, religious specialists are regarded as having a positive influence upon religious communities and society, in as much as their example and ministry inspires action and behavior that is socially beneficial. Most religious specialists are associated with having outstanding personal qualities of kindness, patience, tolerance, concern, holiness, devotion, sensitivity, intelligence, and wisdom. However, amid the proliferation of new religious groups and sects—most requiring leadership—some religious specialists have been so idealized, sometimes through their deliberate flouting of conventional norms and the employment of bizarre rituals, that their humanity and fallibility may be denied, and they are followed with unquestioning loyalty by groups of believers. The dangers of such an attitude have been manifest through the ages in collective acts of destructive or unconventional behavior inspired by the religious specialist or leader. There is sometimes a fine line between being a holy fool and a saint.

3.4 Religious Specialists and Gender

In most world religions, authority and power has been the preserve of men and most religious specialists in history have been male. In some world religions, nearly all formally recognized religious professionals are still men, and in these cases the situation does not seem likely to change in the immediate future. However, in other religious traditions and denominations, women increasingly are taking on roles as public religious professionals, through ordination, particularly in the lower strata of the religious hierarchy. Some of the exceptions to this historical bias towards male religious specialists are the monastic traditions of Buddhism and Christianity where separate orders for women have long been in existence and women have gained credit for their virtuoso religiosity. In various religious traditions over the course of history, individual women have been respected for their wisdom and devotion, such as the Muslim saint and mystic

Rabi'a of Basra (d. 801), and in some instances they have, informally, been the teachers of men. But rarely has this led to their admission into the select community of male religious specialists or scholars. Women have through time often done a great deal of unseen religious work, particularly in prayer and in the care of the sick. The increasingly professional and public role of women in religious life is a reflection of changing social attitudes and the senior roles that some women have achieved in other realms of public life.

4. *Religious Specialists in the Contemporary World*

The role of religious specialists today is carried out in social settings marked by a greater global awareness, increasing religious diversity, and pressures from within and outside the religious community. The impact of these pressures are manifold and have affected different religious traditions and religious professionals, collectively and individually, in various ways. It is beyond the scope of this article to examine these in detail, but some distinctive features of being a religious professional in modern society can be highlighted.

4.1 *Religious Leadership in Minority Contexts*

Historically, religious communities and faiths were often separated from each other through geographical distance, but the advent of modern communications has brought traditions together on a new and unprecedented scale. Most large cities around the globe will have a community of believers associated with the major religious traditions. In these contexts, religious specialists face the challenge of maintaining authority in a situation where religious professionals from other faith traditions also claim authenticity and authority.

There is evidence to show that over time religious specialists serving diaspora communities are often influenced by the dominant professional roles of the religious majority. For example, many of the rabbis serving Britain's well-established Jewish community have long-since expanded their primary teaching and education role, and they have a far greater involvement in pastoral care than historically and traditionally they once might. The idea of a 'Jewish chaplain' serving on a prison or hospital chaplaincy team is now commonplace. Even amongst religious communities with a much shorter history in Britain, there is evidence that the roles of religious professionals are changing. The largely ritual responsibilities of Muslim imams, based in mosques, is beginning to change as they too become more involved in public religious affairs, community support, and pastoral care in public institutions. In this respect, the gradual combining of pastoral care with teaching and leadership of rituals indicates the influence of the dominant professional roles of the religious majority, namely the established Church of England.

4.2 *Religious vs. Secular Professionals*

Through an ongoing process of social differentiation, many of the functions of the religious professional have been usurped by other, nonreligious professionals or by the state. The marginality that religious professionals may face as a consequence, particularly in modern industrial societies, can and does have an impact on future recruitment and vocation. One of the consequences of the decline in the numbers of religious professionals (in some societies and religious traditions), has been an inevitable need for ordinary, lay members of religious communities to become more actively involved in tasks that were formally part of the priestly role.

A good example of how the remit of religious professionals has changed can be seen in the post-Second World War period, particularly in the West, with the dramatic expansion of interest in holistic and alternative therapies with counseling and psychotherapy part of a new industry of self-development. As Davie (1994) notes this has had two effects, namely, 'in usurping some of the traditional functions of a pastor and, secondly, in obliging that pastor—if he wishes to keep on terms—to acquire skills (including an extensive vocabulary) of a novel and essentially unreligious nature.' The curriculum for those undertaking theological training will now almost certainly require study of counseling, bereavement care, and other forms of pastoral care in order to be able to keep abreast of secular professionals working in the fields of therapy and self-development.

See also: Charisma and Charismatic; Charisma: Social Aspects of; Religion and Gender; Religion: Definition and Explanation; Religion: Evolution and Development; Religious Organizations; Ritual; Weberian Social Thought, History Of

Bibliography

Beckford J A, Gilliat S 1998 *Religion in Prison: Equal Rites in a Multi-Faith Society*. Cambridge University Press, Cambridge, UK

Davie G 1994 *Religion in Britain since 1945: Believing Without Belonging*. Blackwell, Oxford, UK

Feuerstein G 1990 *Holy Madness: The Shock Tactics and Radical Teachings of Crazy-Wise Adepts, Holy Fools, and Rascal Gurus*. Arkana, London

Fichter J H 1961 *Religion as an Occupation: A Study in the Sociology of Professions*. University of Notre Dame Press, London and Notre Dame, IN

Gombrich R F 1988 *Theravada Buddhism: A Social History from Ancient Benares to Modern Colombo*. Routledge and Kegan Paul, London
McGuire M B 1987 *Religion: The Social Context*. Wadsworth Publishing, Belmont, CA

S. Gilliat-Ray

Religious Stratification

1. Introduction: Two Forms of Stratification

We can distinguish usefully between two separate sociological issues, namely the hierarchical stratification of religious roles in terms of their authority, and the stratification of religious communities within a general system of social stratification. The first line of inquiry considers the hierarchical organization of charismatic power within religious traditions; the second involves the historical and sociological study of the stratification of religious communities, especially in terms of social class and ethnicity. Both forms of religious stratification are analyzed from the sociological perspective of Max Weber (1864–1920).

For further treatment of these topics see *Religion, Sociology of*; *Religious Specialists*; *Charisma and Charismatic*; *Religion and Economic Life*.

2. Virtuoso and Mass Religion

The sociology of social stratification is concerned with the distribution of scarce resources and with the formation of social hierarchies by the processes of social closure and exclusion. The most common illustration of social stratification would be the unequal distribution of economic resources, resulting in a simple formation of economic classes. However, stratification also involves cultural distinctions such as education, lifestyle, and cultural consumption. Weber made an original contribution to the sociology of religion by treating charisma (a 'gift of grace') as a religious value that is also characterized by scarcity. Because these religious values are unequally and unevenly distributed in human societies, there is religious stratification. In *The Social Psychology of the World Religions*, Weber (1991, p. 287) asserted 'that men are *differently qualified* in a religious way stands at the beginning of the history of religion ... the sacred values that have been most cherished, the ecstatic and visionary capacities of shamans, sorcerers, ascetics, and pneumatics of all sorts, could not be attained by everyone. The possession of such faculties as "charisma," which, to be sure, might be awakened in some but not in all. It follows from this that all intensive religiosity has a tendency towards *status stratification*, in accordance with differences in the charismatic qualifications.'

Religious gifts (charisma) are in demand because they bring healing and salvation, but they are inevitably in short supply. The resulting hierarchy of spiritual gifts creates a stratification system that differentiates between the virtuoso and the mass, where the latter are, in Weber's terminology, religiously 'unmusical.' This hierarchy of intense religious values stands in contrast to and frequently in opposition to the formal hierarchy of the ecclesiastical organizations of Christianity that attempted to regulate and to control the flow of charisma.

This model of virtuoso–mass religion is connected with the specific forms of salvational drive in the world religions, namely with their soteriologies. Religious systems of soteriology may be directed either toward inner-worldly asceticism or toward other-worldly mysticism. Now all religions that place a special emphasis on rebirth and spiritual renewal will create a special category of religious aristocracy, because charisma cannot be institutionalized in terms of an egalitarian distribution. This virtuoso pattern emerged as a general characteristic of religious communities, for example, among the Protestant sects, the *perushim* in Judaism, the Sufi dervishes of Islam, and the Buddhist monks (see Weber 1965, pp. 151–65). The existence of distinctive forms of internal religious stratification has often evolved in sharp contrast to the official doctrines of egalitarianism and universalism in, for example, Christianity and Islam.

There is demand from popular religion to enjoy the benefits of virtuoso religion such as healing, and thus there develops a distinctive pattern of exchange between the virtuosi and the mass. In return for acts of charity and direct economic support from their followers, the virtuosi offer healing, guidance, and other charismatic gifts. The layman who is completely immersed in the everyday world needs the blessing of charisma, but in return the virtuosi depend on lay tributes merely to exist. This mutual dependency and reciprocity create the conditions whereby virtuoso religion can always be diluted or manipulated to some extent by the laity who demand tangible evidence of charismatic gifts in the form of magical displays. In fact, with the possible exceptions of Judaism and Protestantism, all religions and religious ethics have had to (re)introduce cults of saints, heroes, or functional gods in order to accommodate themselves to the needs of the masses (Weber 1965, p. 103). In Asia the religious teachers or *gurus* function as living saviors, while in Catholicism the veneration of the saints was the real basis of popular religion. In Hasidic Judaism, despite its popular appeal to the ordinary people through, for example, its use of dance as an expression of religious joy, there was also a clear division between the spiritual elite and their disciples. In these examples, Weber recognized a paradoxical tension between the spiritual quest for perfection in the virtuosi, the

pragmatic needs of the laity for healing and material sustenance, and the mutual reciprocity between elite and mass. Weber's sociology of charismatic roles was an application of philosophical critique of Christian morals by the German philosopher Friedrich Nietzsche (1844–1900), who had condemned Christianity as a system of slave morality and as a product of psychological weakness. Slave morals are created out of resentment by a subservient social class, and hence the Christian doctrine of love was analyzed critically as a subconscious doctrine of a resentful social group. Guilt and bad conscience played a pivotal role in Christian ethics. Against this moral system, Nietzsche argued in favor of a revaluation of values from a position of psychological strength. Nietzsche's genealogy of (Christian) morals in terms of a study of the stratification of values between slave and master moralities was translated directly into Weber's social psychology of the world religions (see Stauth and Turner 1988). In Weber's treatment of the psychology of religion, he argued that certain social carriers were crucial in the development of the cultural systems of the various religions. In particular, he claimed, for example, that Islam had been carried by warriors, Buddhism by mendicant monks, Confucianism by court officials and the literati, and Christianity by artisans and craftsmen. These occupational groups were particularly important historically in providing the world religions with a specific soteriology. Thus, the world view of Christianity was shaped by the practical rationalism of artisans and craftsmen, and in the long run this rationalism proved decisive in the connection between the Protestant capitalists and economic rationalism. This thesis has been challenged and various historians have attempted to find similar connections between Jewish culture and the rise of capitalism (see Sombart 1951) or between Islam and capitalism (see Turner 1998). The core of the Weber thesis has survived these criticisms insofar as it attempts to identify important cultural connections between religious discipline, class position, and economic rationalism.

3. *Religion and Social Stratification*

Religious communities or groups can also be analyzed in relation to the general system of secular stratification in terms of social class or caste. There is an extensive and complex debate about the relationship between caste, occupational groupings, notions of pollution, and Hinduism. The conventional view was that concepts of purity in Hinduism functioned to sustain a social division of labor between manual and nonmanual occupations that over many centuries evolved into a complete hierarchy of caste. Although the Hindu caste system was organized in terms of four principal castes (Brahmin, Kshatriya, Vaisya, and Sudra), there was considerable diversity at the village level, where these castes were typically subdivided into subcastes or *jati*. Castes were ordered hierarchically in terms of ritual separation around a notion of contagious pollution. This traditional pattern broke down in the twentieth century and there has since been some mobility whereby through the adoption of upper-caste cultural practices lower castes can move up the social scale, and the incorporation of marginal cultural traditions into the Brahmanical system. This process is referred to as 'sanskritization.' There is thus a debate about the differences between Asian and Western systems of stratification and the legitimation of such patterns of inequality by religion (see Dumont 1970). Some sociologists argue that the term caste can be applied to any system of rigid stratification where membership is permanent, hereditary, and guarded or legitimized by religious notions of purity. For example, in the southern states of anti-bellum North America, there is a continuing system of racial orders that is reinforced by the stratification of Christian churches along a division between blacks and whites. The National Baptist Convention, the African Methodist Episcopal, the African Methodist Episcopal Zion, and the Colored Methodist Episcopal churches emerged as exclusively black institutions and illustrate the racial stratification of American denominationalism (see Dollard 1998).

Although religion has played a significant role in the legitimation of social stratification, the relationship between religion and class has to be seen in a dynamic and historical framework. There is a considerable collection of historical and sociological research findings that explore the social mobility of individuals and groups as a consequence of their conversion to evangelical Christian sects. H. Richard Niebuhr's *The Social Sources of Denominationalism* (1957) is the classic illustration of this tradition in the sociology of religion. The basic argument is that 'the churches of the disinherited' have traditionally recruited from the poor and dispossessed through active evangelism. As a consequence of religious asceticism and education, individuals from the lower social classes move up the social scale. The same argument applies to the evangelical sects themselves which, as they become dominated by the lower-middle and middle classes, become more respectable. The process of denominationalization whereby sects acquire the characteristics of middle-class denominations in the United States has been paralleled to the social mobility of their individual members. The history of the Anabaptists, Quakers, Methodists, and Salvation Army exhibit this pattern of social change whereby sects become denominations (see Troeltsch 1931).

The churches of the middle class emphasize personal responsibility and individualism in theology, and restraint and formality in ritual. Their social teaching is concerned with individual behavior rather than with social redemption. The sects of the lower classes appeal to the deprivation of the poor through emotional

preaching, healing practices, and a collective eschatology of redemption. However, because these evangelical sects typically socialize the disinherited into the dominant values of a society, the poor rise through the social scale as they acquire the habits of temperence and personal discipline (Johnson 1961). The education of the laity slowly replaces their conversion. There is some support from the history of the evangelical sects for Max Weber's *The Protestant Ethic and the Spirit of Capitalism* because the personal discipline of the Protestant sects results in the social mobility of the disinherited into the lower and middle classes. Habits of saving, frugality, and temperance produced a reliable workforce that was compatible with the needs of capitalist discipline. In particular, the owners of industrial enterprises recognized that factory discipline could be maintained effectively through the employment of sober Christian workers (see Pope 1942).

4. *Religion and Class Conflict*

There is a close relationship between class membership and religious affiliation in which religious participation functions as a cultural ladder of social mobility. This association between evangelical Christianity and the working class suggested to Karl Marx (1818–83) that the promise of spiritual salvation compensated for the material deprivation of the disinherited. In the famous argument about human alienation in capitalism, Marx (1974, Vol. 1, p. 83) argued that for a society based on private property, 'Christianity with its cultus of abstract man, more especially in its bourgeois developments, Protestantism, Deism etc., is the most fitting form of religion.' Within the system of capitalist exploitation, Christianity was the opium of the people, because its emotional discharge served to obscure the real nature of class inequalities. While Marx treated Christianity in a cursory fashion, Friedrich Engels (1820–95) offered a more extensive treatment of the relationship between class, religion, and political consciousness. For example, he regarded Christianity in its origins as a religion of a class of slaves and oppressed that became eventually the dominant religion of the Roman Empire. There were close historical parallels between the rise of Christianity and the development of the European working class (see Marx and Engels 1955). Engels recognized that in the Reformation the peasant revolts were an expression of real material interests against oppression, and that Christian belief provided the peasantry with a powerful ideology of protest (see F. Engels, *Peasant War in Germany*, 1965). These peasant protests which drew upon the Bible as a source of social and political criticism were important because they provided a model of the millenarian struggles of the seventeenth century and the secular conflicts of the nineteenth century, when working class movements continued to absorb and transform the biblical language of protest against capitalism. Methodism often provided the training ground for political protest, because the tradition of lay preaching provided a pedagogy of popular opposition. Popular leaders who emerged within the ranks of the Methodist chapels typically were expelled from the church because their secular radicalism was incompatible with the conservative teaching of the Methodist chapels (see Thompson 1963).

In the seventeenth century, the religious sects of the peasantry and yeomanry were important engines of social change in the English civil war. In the eighteenth century, Methodism had through its emotional preaching provided a compensation for this worldly deprivation. In the nineteenth century, there was an unstable alliance between working-class protest and the evangelical sects, whereby the idiom of protest depended more on the Bible than on Marxist ideology. In the English case, these causal connections between politics and religion were summarized in the Halévy thesis, which suggested that Wesleyan Methodism functioned as a social ladder between the nonconformist dissent of the artisan and the Anglican establishment, thereby contributing to social stability and evolutionary political change (see Halévy 1962).

5. *Conclusion: Religion, Class and Egalitarianism*

Whereas religions like Confucianism and Buddhism have not systematically provided or embraced an ideology of radical egalitarianism, the Abrahamic faiths of Judaism, Christianity, and Islam are regarded as prophetic religions of the book that generate a revolutionary utopia of salvation. Because they create a radical tension between this world and a sacred realm of justice, they have been seed-beds of revolutionary movements. The actual existence of religious stratification is in this sense an offence to the theology of justice in biblical religion. Logically speaking, this tension between existing social injustice and the prophetic promise of the restoration of the faithful and the punishment of their enemies can take two forms—either a theodicy of other-worldly justice in a future kingdom or radical millenarianism in this world (see Mannheim 1997). These images of revenge and renewal provided slaves and the disinherited classes with a profound ideology of political change. However, in the Abrahamic religions the connections between social class, religious ideology, and political protest have remained highly unstable. Periods of revolutionary fervor have been followed by periods of mystical withdrawal, and the pendulum swing between protest and quietism is well illustrated by the complex history of messianism in Judaism, Christianity, and Islam (see Lantenari 1963). In Islam, the tension between the Koranic promise of a radical egalitarianism and earthly justice has proved to be a resilient source of religious protest against the dominant classes (see Marlow 1997). Although Marx and Engels as-

sumed that religious ideologies of class protest would be replaced by secular or scientific visions of political change, religious inspiration of class protest in the twentieth century has been an important aspect of liberation theology in Latin American Catholicism and of Islamic protests against colonialism. Traditional Shi'ite notions of the Hidden Imam and justice within the Islamic community provided a powerful tool in the emergence of the Iranian revolution which was produced by the alienation of the peasantry and the decline of traditional classes such as the merchants (see Kamali 1998). Religious stratification therefore continues to be an important feature of social and political institutions into the twenty-first century, and religious ideologies provide significant cultural components of both social stability and radical change.

See also: Voting: Class

Bibliography

Dollard J 1998 *Caste and Class in a Southern Town*. Routledge/Thoemmes Press, London

Dumont L 1970 *Homo Hierarchicus. The Caste System and its Implications*. University of Chicago Press, Chicago

Halévy E 1962 *A History of the English People in the Nineteenth Century*. Benn, London, 2 Vols.

Johnson B 1961 Do holiness sects socialize in dominant values? *Social Forces* **39**: 309–16

Kamali M 1998 *Revolutionary Iran Civil Society and State in the Modernization Process*. Ashgate, Aldershot, UK

Lantenari V 1963 *The Religions of the Oppressed. A Study of Modern Messianic Cults*. MacGibbon & Kee, London

Mannheim K 1997 *Ideology and Utopia*. Routledge, London

Marlow L 1997 *Hierarchy and Egalitarianism in Islamic Thought*. Cambridge University Press, Cambridge, UK

Marx K 1974 *Capital*. Lawrence and Wishart, London, 3 Vols.

Marx K, Engels F 1955 *On Religion*. Foreign Languages Publishing House, Moscow

Niebuhr H R 1957 *The Social Sources of Denominationalism*. Meridian Books, New York

Pope L 1942 *Millhands and Preachers*. Yale University Press, New Haven, CT

Sombart W 1951 *The Jews and Modern Capitalism*. Free Press, Glencoe, IL

Stauth G, Turner B S 1988 *Nietzsche's Dance Resentment Reciprocity and Resistance in Social Life*. Basil Blackwell, Oxford, UK

Thompson E P 1963 *The Making of the English Working Class*. Penguin, London

Troeltsch E 1931 *The Social Teaching of the Christian Churches*. George Allen & Unwin, London, 2 Vols.

Turner B S 1998 *Weber and Islam*. Routledge, London

Weber M 1930 *The Protestant Ethic and the Spirit of Capitalism*. George Allen & Unwin, London

Weber M 1965 *The Sociology of Religion*. Methuen, London

Weber M 1991 The social psychology of the world religions. In: Gerth H H, Wright Mills C (eds.) *From Max Weber Essays in Sociology*. Routledge, London, pp. 267–301

B. S. Turner

Remedies and Damages: Legal Aspects

Remedies can be categorized as monetary or nonmonetary. Although much litigation leads to monetary awards, many cases end without a monetary award. Injunctive relief, orders to specifically perform contracts, divorce decrees, probate of wills, and other orders do not necessarily involve damages or monetary payments.

1. Nonmonetary Relief

Several categories of nonmonetary relief exist and their use varies across countries. In contract law, countries divide into three groups with respect to the availability of specific performance: (a) countries, such as Germany and socialist systems, in which specific performance is generally available, subject to exceptions, (b) countries, such as France, which make specific performance available for specific obligations, and (c) common law countries, including England and the United States, in which specific performance is an exceptional remedy, available in limited circumstances.

Nonmonetary relief to prevent injury can be more important than redressing injuries already done. Common law countries address this need by allowing injunctive relief in cases in which irreparable harm can be demonstrated. Other legal systems, such as Germany, Switzerland, Austria, and others influenced by German law, have a general class of actions to prevent threatened injury. Courts in such systems have less discretion to deny preventive remedial relief than do common law courts. Some systems allow remedial relief but lack a coercive component, such as the penalties for violating an injunction, to enforce it (Stoll 1983). Prejudgment nonmonetary remedies, such as attachment and preliminary injunction or its equivalent, are widely available.

2. Damages Rules across Countries

With respect to monetary relief, one cannot completely separate the question of the quantum of damages from the question of liability. Blackstone, in writing about the common law, did not regard damages as a separable ending of the legal process. The right to damages arose the instant the victim was injured. Nevertheless, one can intelligently address many damages issues independently of the question whether any damages should be awarded.

In the last quarter of the twentieth century, damages issues fueled the tort reform movement in the United States and alarmed foreigners doing business in the United States. In many respects, however, damages rules are not strikingly different between the

United States and other countries. All countries seek to compensate victims fully, though most depend more on social compensation systems than does the United States. Most countries allow consideration of the injured person's fault, a consideration similar to the United States doctrine of comparative negligence, and a few (the Netherlands and Sweden) allow for mitigation of damages in view of the parties' financial circumstances and insurance coverage. European and Japanese damages rules governing injury, though less feared than United States rules, are not necessarily simpler (Pfennigstorf 1991). They contain 'a confusing variety of items of non-pecuniary loss and an equally confusing variety of methods for measuring them.'

There are, however, differences in the size of awards. Non-United States systems produce lower and more predictable within-country awards. Courts in several countries observe written and unwritten guidelines with respect to awards. In Japan and Sweden, for example, expert commissions establish consistent guidelines for automobile accidents. Sweden extends this practice to injuries caused by medical mistreatment, by pharmaceutical drugs, and by environmental pollution. Across European countries, however, awards are less consistent. One study estimates that in a wrongful death case involving a 20-year old victim with no dependents, awards across Europe would vary from £210,122 in Italy to £1,250 in England; it also finds that pain and suffering awards vary considerably across countries (McIntosh and Holmes 1990).

3. *Social Scientific Study of Compensatory Damages*

Media reports, which understandably emphasize extreme awards, promote exaggerated perceptions of large United States damages awards. The media overreport products liability and medical malpractice cases, exaggerate plaintiff win rates, and distort impressions about the actual distribution of jury awards. Social scientific assessment of damages is critical in helping to ameliorate the effect of this reporting bias.

3.1 *General Adequacy of Damages*

Despite agreement that awards are higher in the United States, there is little evidence that injury victims are systematically overcompensated in any country. Evidence exists that United States plaintiffs with small losses tend to be overcompensated but that plaintiffs with large losses tend to be undercompensated (Vidmar 1998). Yet it is the large loss cases that tend to be the most controversial. An English study found that serious injuries may not result in substantial recoveries and that minor injuries which do lead to recoveries may have unexpectedly severe consequences (Law Commission 1994). Both findings suggest systematic undercompensation.

3.2 *Award Levels and Trends*

The most studied system of damages is the tort system in the United States. Concerns about the level of awards, time trends in awards, and the relative competence of judges and juries have generated several studies. Data gathered by the National Center for State Courts from 45 of the largest 75 United States counties, under grants from the Bureau of Justice Statistics (1995, 1999) ('BJS'), provide the least biased sample of awards in litigation. The data allow for meaningful comparisons across case categories. Since the same 45 counties were surveyed in both 1991 and 1996, the data also allow meaningful comparisons over time, discussed below. Table 1 summarizes the

Table 1
Mean and median compensatory damages award, by case category and year, in thousands of 1998 dollars

	1991			1996		
	Mean	Median	n	Mean	Median	n
Auto torts	219	34	1,102	228	21	1,384
Dangerous premises	323	71	394	350	84	441
Products liability	864	319	71	901	363	95
Intentional tort	445	47	91	438	33	112
Medical malpractice	1,424	232	181	1,525	306	172
Professional malpractice	1,303	152	44	392	76	29
Slander/libel	230	23	15	372	19	17
Other torts	447	74	188	11,477	117	186
Contract fraud	774	61	87	488	52	99
Employment relations	412	149	94	537	194	110
Lease	1,963	66	47	239	63	39
Other real estate	205	53	11	154	97	12

pattern of jury compensatory damages awards, in thousands of 1998 dollars, for these 45 counties.

Table 1 shows that damages awards vary substantially by case category. Median compensatory damages awards are highest in cases involving products liability and medical malpractice. Indeed, median awards in these case categories are substantial multiples of the median award in automobile accident cases. The levels of awards are generally modest and are consistent with other comprehensive studies of particular jurisdictions.

3.3 Reliability of Compensatory Awards

In studies that include independent assessments of the defendant's behavior, damages awards correlate with the nature and degree of the wrong. Studies of medical malpractice cases show that those who behaved the worst and caused the most harm are the most likely to pay the most damages. Experimental studies confirm these findings (for example, Wissler et al. 1999). Osborne (1999) uses a different methodology but reaches similar results. His study includes many substantive areas of law and examines the relation of court awards to the pretrial expectations of litigants and their attorneys and to measurable, economically relevant damages, He finds that court awards are highly predictable, as variance in expectations explains much of the variance in awards. The awards also significantly relate to both medical costs and property damage. He rejects the hypothesis of a highly unpredictable court system.

4. Punitive Damages

Punitive or exemplary damages are in addition to the damages awarded to make the victim whole. But normally punishment is left to the state, not to private civil litigation. Thus, many legal systems do not allow damages beyond those necessary to compensate the victim. Little authority exists for punitive damages in Roman law and thus in some of the continental European systems that derived from it. But many countries allow punitive damages or their equivalent (exemplary damages, moral damages) in some classes of cases. And the European Court of Justice has ruled that national courts must be willing to award exemplary damages pursuant to claims or actions founded on Community law, if such damages may be awarded pursuant to similar claims based on domestic law. Nevertheless, punitive damages are substantially less important outside the United States.

The most discussed questions in the recent United States literature on damages relate to punitive damages. Understandably skewed media coverage emphasizing high awards combined with advocacy by lobbying groups has generated enormous pressure on punitive damages law and helped to shape professional and public impressions. Business lobbying groups have enjoyed some success, with several states enacting provisions raising the burden of proof for punitive damages claims, placing monetary ceilings on awards, and otherwise limiting the likelihood and size of punitive awards.

Social science research on punitive damages expanded as a result of the 1994 award of $5 billion in punitive damages against Exxon Corporation (now Exxon Mobil) for the 1989 Exxon Valdez oil spill. Exxon funded several leading academics, who produced an impressive body of work that makes important social science contributions. All of this work is critical of punitive damages and supports 'reform.' But the thrust of some of the work seems inconsistent with prior theoretical discussions of punitive damages and some of it has been criticized as methodologically unsound (for example, Eisenberg 1998, Vidmar 1998) and as overstating its implications.

4.1 Rate of Punitive Awards

The social science picture that emerges from systematic studies of actual cases differs in some respects from the implications of the experimental and theoretical results of the Exxon-funded work. Three large studies of punitive damages awards, Moller (1996), Daniels and Martin (1995), and data commissioned by the BJS (1995), cover multiple United States jurisdictions, cover similar time periods, and include case categories that are sufficiently well defined to warrant comparison. Each of the three sources reveal overall punitive damages award rates of less than ten percent in successful jury trials.

In two of the areas of greatest interest to policymakers, products liability and medical malpractice litigation, punitive awards are especially rare. In medical malpractice cases, the 95 percent confidence intervals for all three sources overlap and show awards in less than five percent of jury trials in which plaintiffs are successful. In products liability cases, all three estimates again overlap at under ten per cent of successful jury trials, and rates under five percent are found in studies of published opinions. Except in case categories dominated by fraud and employment law issues, all punitive award rate levels are likely less than ten percent, and many are less than five percent. These social scientific analyses undermine the belief that punitive damages are frequently awarded.

4.2 Level of Punitive Damages Awards

Large, publicized punitive awards, such as that against Exxon, create the impression that punitive awards are on average much larger than they truly are. BJS data for 1991 and 1996 allow one to assess the level of

Table 2
Mean and median punitive damages award, by year, in thousands of 1998 dollars

	Mean	Mean excluding highest award	Median	n
1991	557	404	58	168
1996	1,887	688	52	119

punitive damages awards. Table 2 summarizes punitive damages awards, in thousands of 1998 dollars, for the years 1991 and 1996. Median punitive damages awards have been more modest than the headline-grabbing awards. In both years they are in the 50,000 dollar range. Mean punitive awards are substantially affected by extremes. Excluding the most extreme punitive award in each year, reduces the mean by over 25 percent in one year and by over 50 percent in the other. In these data, million-dollar awards constitute less than 10 percent of the punitive awards. All studies confirm that the bulk of punitive awards are modest.

4.3 Relation Between Punitive and Compensatory Awards

A criticism of punitive damages is that they bear no relation to the compensatory award—the punishment inflicted therefore is not proportional to the harm caused. And the United States Supreme Court has indicated that a punitive damages award normally should bear a reasonable relation to the compensatory award. Social scientific analysis suggests that punitive awards do relate to compensatory awards. Figure 1 shows the relation (in logs) between compensatory and punitive awards in the BJS (1995) study of 45 counties. It shows an unmistakable correlation. Other data sets reveal approximately the same relation (Eisenberg et al. 1997).

Punitive damages awards are thus rarer, smaller, and more closely related to compensatory awards than is widely believed.

5. Time Trends

5.1 Compensatory Damages Awards

Table 1 allows study of award trends over time. Popular perception has it that damages awards are ever-increasing. And some empirical study supports the claim. But these studies have been questioned on methodological grounds (Vidmar 1998). Table 1 shows little systematic evidence of increasing awards. In several case categories, median or mean awards decreased from 1991 to 1996. The only statistically

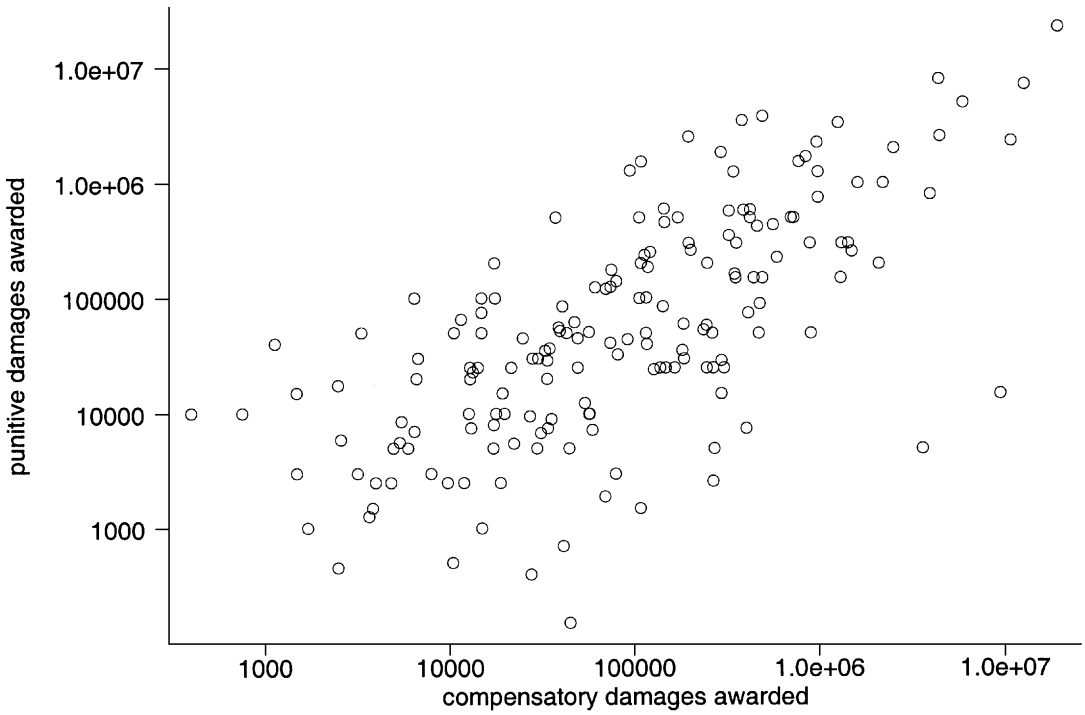

Figure 1
The relation between punitive and compensatory damages awards.

significant change in mean awards was in automobile accident cases and here the 1996 mean (in logs) was statistically significantly *lower* than the 1991 mean ($p < 0.001$). Other changes in award patterns were not statistically significant.

Table 1 is limited to state court cases. Awards in federal court cases tend to be higher, at least partially because of jurisdictional amount requirements in federal court. Median trial awards in federal courts products liability cases did increase from the early 1980s to the mid-1990s. The increase also occurred in judge-tried cases and in contracts cases.

5.2 Punitive Damages Awards

Punitive damages, like compensatory damages, are often said to be ever-increasing in number and size. The two years of BJS jury-trial data allow one to compare the rate and size of punitive awards in the 45 large counties covered by the data. Table 2 summarizes the results, in thousands of 1998 dollars, for the years 1991 and 1996. It shows that median punitive awards were substantially lower in 1996 than 1991. Mean awards were higher in 1996 but that is a function of a few large cases. The means (log) are not statistically significantly different. Regression analysis that controls for locale and case category confirms that punitive damages show no statistically significant movement from 1991 to 1996.

Table 2 also shows that the frequency of punitive damages awards decreased in these 45 counties, from 168 in 1991 to 119 in 1996. Thus, even minimal social scientific analysis provides no support for the belief that punitive damages awards are ever-increasing in size or frequency.

6. The Importance of the Adjudicator

It is widely believed that the choice of adjudicator—judge, jury, arbitrator—substantially affects the damages awarded. Higher and more variable awards in the US are attributed to the greater role of jurors as adjudicators. While real differences exist across adjudicator groups, the social scientific evidence supports skepticism about claims of massive differences across adjudicators. The classic Kalven and Zeisel study used questionnaires to presiding judges in about 4,000 actual civil jury trials in the 1950s (Kalven and Zeisel 1966). Their data showed a 78 percent agreement rate between judge and jury on liability, with no revealing pattern in the cases of disagreement. The rate of judge-jury agreement is high compared to other human endeavors. They also showed that jurors awarded on average 20 percent more than judges said they would have awarded. The modest judge-jury differences reduced interest in attacks on juries.

Renewed attacks on the jury's competence to assess damages, and especially punitive damages, re-emerged in the 1980s and 1990s. Research using interviews and experiments finds that different decisionmakers are quite similar and predictable with respect to their judgments of injury severity (Wissler et al. 1999). Such research also finds that jurors' awards lose consistency in translating severity of injury into dollar awards. Other studies suggest somewhat lesser judge-jury differences. Vidmar and Rice (1993) report experimental evidence about jurors awaiting jury duty and legal professionals bearing on pain and suffering awards. In a medical malpractice burn case, they found that six-person juries would give more variable awards than twelve-person juries but that both groups would be less variable than individual legal professionals. Their second experiment involved an automobile accident case and a medical malpractice case. They found no differences between jurors and legal professionals.

Studies of actual cases, which lack the controls of experimental settings but provide the relevant setting, suggest modest adjudicator differences. Osborne's (1999) study compares the predictability of judge and jury awards based on litigant expectations. He finds that one can reject the hypothesis that jury trials are largely random events. Litigant expectations explain more than a third of the variance in awards. And he finds that the behavior of judges is not significantly more predictable than that of juries. Helland and Tabarrok (2000) conclude that most differences between judge and jury awards can be explained by differences in the sample of cases coming before judges and juries but that some suggestive differences exist. Eisenberg et al. (2000) find little meaningful difference in the pattern of judge and jury punitive damages awards.

Exxon-funded research finds experimental evidence of vast judge-jury differences. For example, Hastie and Viscusi (1998) report mock jurors to be much more likely than judges to impose punitive sanctions in a hypothetical railroad accident case. Vast judge-jury differences on liability in these new studies, which contrast with Kalven and Zeisel's and other findings, calls into question the experimental designs' ability to detect socially meaningful punitive damages differences between judges and juries.

7. Overall Assessments and Comparisons of Damages Awards

Greater injuries lead to greater damages awards in all systems for which data are available. With respect to variability in the United States, where variability is regarded as greatest, from half to almost three-fourths of awards' variation can be accounted for by measures of injury severity, as well as by other predictors, but a

substantial random component also exists. Legal systems that rely on schedules of damages for various injuries have lower variability though not necessarily more complete compensation.

Prospects for definitive social scientific comparison of damages rules' operation are dim. Some inherent limitations apply to comparing performance of legal rules across state and national boundaries. The number of differing legal systems will tend to be small—there are, for example, only 50 states and gathering useful damages information about that many countries is an enormous task. Important differences, such as varying reliance on social compensation systems, different mixes of cases reaching court systems, different methods for compensating attorneys, and varying standards of living and purchasing power need to be considered. For the foreseeable future, reasoned analysis of data available for individual jurisdictions and informal comparison across jurisdictions will have to suffice in assessing performance of the damages component of remedial systems.

See also: Deterrence; Deterrence: Legal Perspectives; Deterrence Theory: Crime; Liability: Legal; Litigation; Parties: Litigants and Claimants; Punishment: Social and Legal Aspects

Bibliography

Bureau of Justice Statistics Special Report 1995 *Civil Justice Survey of State Courts 1992: Tort Cases in Large Counties*

Bureau of Justice Statistics Bulletin: Civil Justice Survey of State Courts 1999 *Civil Trial Cases and Verdicts in Large Counties, 1996*

Daniels S, Martin J 1995 *Civil Juries and the Politics of Reform*. Northwestern University Press, Evanston, IL

Eisenberg T 1998 Measuring the deterrent effect of punitive damages. *Georgetown Law Journal* **87**: 347–57

Eisenberg T, Goerdt J, Ostrom B, Rottman D, Wells M T 1997 The predictability of punitive damages. *Journal of Legal Studies* **26**: 623–61

Eisenberg T, LaFountain N, Ostrom B, Rottman D, Wells M T 2002 Judges, juries, and punitive damages: An empirical study. *Cornell Law Review* **87**:

Hastie R, Viscusi W K 1998 What juries can't do well: The jury's performance as a risk manager. *Arizona Law Reivew* **40**: 901–21

Helland E, Tabarrok A 2000 Runaway judges? Selection effects and the jury. *Journal of Law Economics Organization* **16**: 306–33

Kalven H, Zeisel H 1966 *The American Jury*. University of Chicago Press, Chicago, IL

Law Commission for England and Wales 1994 *Personal Injury Compensation: How Much Is Enough? A Study of the Compensation Experiences of Victims of Personal Injury*. Law Com. No. 235. HMSO, London

McIntosh D, Holmes M 1990 *Personal Injury Awards in EC Countries: An Industry Report*. Lloyd's of London Press, London

Moller E K 1996 *Trends in Civil Jury Verdicts Since 1985*. RAND, Santa Monica, CA

Osborne E 1999 Courts as casinos? An empirical investigation of randomness and efficiency in civil litigation. *Journal of Legal Studies* **28**: 187–203

Pfennigstorf W 1991 *A Comparative Study of Liability Law and Compensation Schemes in Ten Countries and the United States*. Insurance Research Council, Oak Brook, IL

Stoll H 1983 Consequences of liability: remedies. In: Tunc A (ed.) *International Encyclopedia of Comparative Law 11*, J. C. B. Mohr. Tübingen, Germany

Vidmar N 1998 The performance of the American civil jury: An empirical perspective. *Arizona Law Review* **40**: 849–99

Vidmar N, Rice J 1993 Assessments of noneconomic damage awards in medical negligence—A comparison of jurors with legal professionals. *Iowa Law Review* **78**: 883–911

Wissler R L, Hart A J, Saks M J 1999 Decisionmaking about general damages: A comparison of jurors, judges, and lawyers. *Michigan Law Review* **98**: 751–826

T. Eisenberg

Remote Sensing

Remote sensing is one of a suite of geospatial technologies that are having a growing impact in a wide variety of areas from commerce to science in public policy. While remote sensing's origins and technological advance are an outgrowth of military and intelligence applications, commercial and scientific applications of the technology have been growing. The field of remote sensing evolved from the interpretation of aerial photographs to the analysis of satellite imagery, and from local area studies to global analyses, with advances in sensor system technologies and digital computing. Today remote sensor systems can provide data from energy emitted, reflected, and/or transmitted from all parts of the electromagnetic spectrum. These data can then be transformed into information products either by manual or machine-assisted interpretation and employed in an analytical or management decision-making context. Examples of applications of these data include population and demography studies, study of archaeological sites, energy studies using hydrological models, urban planning, environmental treaty enforcement, and agricultural yields. This article provides a brief overview of the historical development of remote sensing, emphasizing the increasing complexity of platforms, systems, and tasks. The information provided here has been assembled from previously published sources to give an overview of a rapidly evolving field that has had, and will in the future have, an increasing impact on our global society.

The definition of remote sensing used here, provided by the American Society for Photogrammetry and Remote Sensing (ASPRS), is:

> In the broadest sense, the measurement or acquisition of information of some property of an object or phenomena, by a recording device that is not in physical or intimate contact

> with the object or phenomenon under study; e.g., the utilization at a distance (as from aircraft, spacecraft, or ship) of any device and its attendant display for gathering information pertinent to the environment, such as measurements of force fields, electromagnetic radiation, or acoustic energy. The technique employs such devices as the camera, lasers, and radio frequency receivers, radar systems, sonar, seismographs, magnetometers, and scintillation counters. (Reeves et al. 1975)

What follows is a brief summary of various sensor systems within the context of an historical background of remote sensing, and a short discussion of the future of remote sensing with some reference to social and behavioral sciences.

1. *Sensor Systems*

Today, a wide variety of systems are used to detect energy patterns, which provide information about the Earth's atmosphere, oceans, and land surfaces. These systems range from cameras in aircraft to advanced sensor systems in Earth-orbiting satellites. The aerial camera systems typically used for remote sensing research and applications today are highly sophisticated. Most satellite remotely sensed data are acquired by sensor systems such as multispectral scanners, linear array devices, and synthetic aperture radar systems that operate in the visible, near infrared, thermal infrared, and microwave portions of the electromagnetic spectrum.

These remote sensor systems can be broadly categorized into *passive systems* and *active systems*. A camera is an example of a passive system as its film records light energy reflected from an existing source: the sun. Add a flash attachment to the camera and it becomes an active system, since it now provides its own source of illumination. Passive systems record solar radiation reflected from the Earth's surface. Data derived from passive systems such as multispectral scanners provide information on, among other things, land use, vegetation types, distribution and condition of landforms, soils, surface waters, and river networks. Active microwave (radar) systems are used commonly in geological and hydrological applications. Because active systems do not depend on reflected energy from the sun they may be able to acquire data during the day or night. In addition, as the atmosphere does not attenuate microwave radiation as much as shorter wavelength energy, radar systems allow data acquisition in cloudy and rainy weather. This ability to operate day or night, and to create images of the Earth's surface in spite of cloud cover, makes radar sensors particularly attractive tools for improving our understanding of both tropical and polar areas.

Thermal infrared data, which represent a record of emitted energy from surfaces, have been particularly useful in monitoring fires and in improving our understanding of areas of volcanic and geothermal activity. Surface temperature of the ocean is also related to the dynamics of coastal waters and currents. Over land, plant water stress also induces changes in canopy temperatures, which can be detected by thermal sensors. Thermal imagery from satellites has been used to record the nighttime lights of the globe, giving researchers an indication of the spatial distribution of population centers. Other remote sensing systems detect the earth's magnetic and gravitational fields; these tools are used extensively in oil and mineral exploration. Employing imagery from more than one portion of the electromagnetic spectrum can increase the information derived from an analysis, and is typically referred to as multispectral image analysis. Analyses of such multispectral data can, if properly designed, increase both the quantity and quality of information for given applications.

2. *Image Interpretation*

Interpretation of remotely sensed data is part science and part art. Manual interpretation of remotely sensed data is fundamental to all image interpretation. Even with the ever-increasing sophistication of machine-assisted image interpretation software, it is still the human that ultimately must label the feature or phenomenon being identified. Not all individuals have equal interpretive skills with respect to aerial photographs or images acquired by remote sensor systems. It is axiomatic, however, that the more one knows about the elements, techniques and methods of image analysis the better an interpreter one will be. In addition, the more one knows about the systems which acquire the data one is interpreting and the subject or discipline area most closely related to the interpretation, the better an interpreter one will make. It is seldom that an interpreter uses only the information in the imagery being analyzed as the basis for their interpretation. Ancillary data in the form of interpretation keys, maps, diagrams, textual material, and other forms of site-specific literature often aid in interpretations. Fieldwork is also a key element in image interpretation and the collection of what is termed 'ground truth' should be accomplished whenever possible and practical.

2.1 *Resolution*

Resolution is at the core of the successful use of remote sensing data. Resolution, as typically applied by a user, normally involves the question 'can I identify what I need to identify on this imagery?' Often this involves the use of the term *ground resolved distance* (GRD). GRD is used to describe the smallest object that can be unambiguously identified on the ground given good object to background contrast. Object to

background contrast is important here. While it may be possible to see white lines on a black road on one set of images, it may not be possible to see the same size white lines on a concrete stretch of the same road. Resolution is thus dependent upon many things.

Indeed, the level of matching between the spatial, spectral, and temporal resolutions of the measurements and needs of the investigation will determine the usefulness of the derived data/information. Spectral range and resolution as used here refer to the portion of the electromagnetic spectrum being used in the measurement. It must be appropriate to the questions being addressed, for example, thermal range for fire detection, visible or infrared for vegetation health and conditi0on. Spatial resolution determines the level of detail that can be extracted concerning objects in a given scene. As a general rule it must correspond to the sizes of typical objects that must be separately identified. A resolution of a meter or less is needed to identify a single tree, a resolution of five meters for a single-family residence, a resolution of 50 meters might be sufficient for identification of most agricultural fields in North America, while a resolution of 500 meters to a kilometer may be enough for studying ocean surface features. A necessary compromise between area to be covered and ground resolution of the measurement must take into account the volume of data which will be generated by, say, a large area coverage—millions of km^2—using high resolution data—one to 30 meters.

The frequency of the data acquisition must also match the natural frequency of change in the landscape and the physical, social, or cultural phenomena of interest. For traffic studies, timeframes of minutes can be important. Hourly observations might be of interest for certain meteorological conditions. Daily observations may be needed for an assessment of plant evapotranspiration associated with irrigated agricultural crops, while observation once every year could be sufficient for land cover change assessment.

2.2 Geographic Information Systems

The synergy between geographic information systems (GIS) and remote sensing comes into play here. To be interpreted accurately, remotely sensed data are often supplemented with other data. Often these ancillary geospatial data can be found or included in a GIS for analysis. But to be more valuable in decision-making contexts, GIS data layers should be up-to-date as is practical. Remotely sensed data are a key technology for updating many types of GIS data. Thus when environmental planners, resource managers, and public policy decision-makers want to measure, map, monitor, or model future scenarios in order to facilitate better management decision-making, remote sensing is being employed more and more within the context of a GIS as a decision support system.

3. Historical Development

In the most general sense, the field of remote sensing begins with the invention of photography. Joseph Nicéphore Niepce took the first photograph of nature in 1827. It is a picture of the courtyard of his home in the village of Saint Loup de Varenne, France, with an exposure time of eight hours. In 1829 Niepce and Louis M. Daguerre signed a partnership agreement to work on a practical method of taking photographs. Niepce had been working on what he called Heliography, or sun drawing, and Daguerre on dioramas, which he constructed with the aid of a camera obscura. The process they pioneered was called direct positive image. The result of their collaboration, called the *daguerreotype*, was finalized in 1839. Also in 1839, an Englishman named William H. Fox-Talbot and a member of the Royal Academy invented the *calotype* (from the Greek: *kalos* for beautiful and *typos* for impression). Talbot's positive negative process is essentially the same as that used today, yet at the time the images produced by the two processes were not comparable.

In 1858, Gaspar Felix Tournachon, also known as Nadar a Parisian photographer and balloonist, took the first known aerial photograph from a captive balloon at an altitude of 1,200 feet over Paris. During the late 1870s, German foresters used aerial photography taken from balloons to both map and measure the aerial extent of tree species. This is an example of the earliest known application of both aerial photographic interpretation and photogrammetry (the science of making precise measurements from remotely sensed data).

The use of rockets to launch cameras first occurred in 1906 when Albert Maul launched a rocket propelled by compressed air that was equipped with a camera. The camera took a picture from an altitude of 2,600 feet and the camera was then returned to the ground by parachute. In 1909, Wilbur Wright took one of the first photographs from an airplane in Centrocelli, Italy. By 1915, cameras devoted to aerial photography were being produced. World War One saw an explosion in the use of aerial photographs. During the Meuse Argonne offensive some 55,000 aerial photographs were taken and delivered to Allied Expeditionary Forces over a four-day period. During this period thousands of photo interpreters were trained. These interpreters laid the groundwork for the applications of aerial photographic interpretation that began to expand after the end of World War One.

In the early 1920s the first books and articles on the applications of aerial photo interpretation began to be published. Lee's 1922 *The Face of the Earth As Seen From the Air* work 'shows familiar scenes from a new angle,' and contains chapters on mountain features, coastal mud flats, plains, and submerged landforms. Joerg's 1927 work, 'The use of the air plane in city geography,' was published in the *Annals of the*

Association of American Geographers. This work focused on the application of aerial photography for urban studies. In 1927 Bergen published his work on 'Aerial surveys for city planning' in the *Transactions of the American Society of Civil Engineers*, and in 1930 the American Geographical Society published the work by Platt and Johnson, 'Peru from the Air.' While providing a regional overview of the country this work also focuses on archeological sites. Of a more general nature is the work by Fairchild (1927) in the May issue of the *Annals of the American Academy of Political and Social Sciences* entitled: 'Aerial photography, its development and future.' This article discusses potential applications of aerial photography from highway planning to water supply management. During this period, however, by far the foremost applications of aerial photography discussed in the literature are for mapping and geological exploration, particularly on the uses of aerial photography in oil exploration, structural and engineering geological analyses, land feature mapping, and soils mapping.

It is in the late 1920s and early 1930s that the agencies of the US Government begin to make significant use of aerial photography for operational applications. At this time the Agricultural Adjustment Administration began to systematically photograph farm and ranch lands across the entire USA. This use of aerial photography for documenting agriculturally active areas has continued at five to 10-year intervals since this time. Thishistorical record has been very significant in many types of studies, but has been particularly useful in a variety of legal actions where the documentation of land use and/or land cover conditions and changes through time are in question. At the same time, the US Forest Service began to use aerial photography to document much of the nation's timber reserves. Studies such as this began in other nations as well, notably Canada. In this timeframe the US Geological Survey began to photograph many areas of the country in order to make topographical and geological maps based upon photogrammetric techniques and interpretation. The Tennessee Valley Authority and other regional agencies began to use aerial photography for planning purposes, as did state and metropolitan planning agencies. The *Master Plan of Residential Land Use*, published in 1943 by the Chicago Planning Commission, was based largely on data interpreted from aerial photographs.

World War Two led to developments in aerial photography, with a special interest in new and higher resolution film emulsions and thermal infrared imagery. Kodak patented the first false color infrared sensitive film in 1946. The 1950s saw developments in technology that allowed multispectral imagery to be taken. Thermal infrared and side-looking airborne radar systems began to be acquired by the US military. Around this time, research personnel at the US Office of Naval Research coined the term 'remote sensing' to cover the expanding range of sensor systems and technologies for acquiring images of the world around us. This was the same year that the Soviet Union launched Sputnik, the first Earth-orbiting satellite. Civilian use of remotely sensed data during this period produced some of the first studies on the application of aerial photography for population estimation.

As early as 1946 the US Government began to pursue an operational space-based imaging system when the US Air Force asked the RAND Corporation to explore the possibility of launching an object into orbit. An early experiment in satellite land remote sensing was conducted from the Explorer 6 platform in 1959. The synoptic images of Earth's weather patterns acquired by the Television and Infrared Observational Satellite (TIROS-1), launched in April of 1960, captivated the American public and demonstrated the capability and necessity of earth observation from space. Also in 1960 the first experimental spy satellite, a component of what would later become known as the Corona Program, was launched and, shortly thereafter, classified as secret. These data were withheld from the public until 1995, when the imagery and documentation were released to the public under Executive Order 12951. In 1965 NASA initiated the Earth Resources Survey Program (ERS). This early program sought to develop methods and applications for remote sensing. The Department of Agriculture researched the use of remote sensing for geological, hydrological, geographical, and cartographical applications. In 1966 the Department of Commerce began to participate in the ERS program by establishing the Environmental Sciences Group within the Environmental Sciences Administration (the predecessor to NOAA). Also in 1966, the Department of the Interior initiated the multi-agency Earth Resources Observation Satellite Program, which studied the potential of Earth-observing satellites for natural resources applications.

During the 1960s and into the mid-1970s NASA and a number of other Federal Agencies focused a significant amount of attention on applied remote sensing research. This was, in part, aimed at justifying the need for civil land satellite remote sensing systems and cataloging the variety of fundamental and applied science questions that could be addressed by such systems. It is during this period that efforts began to develop computer-assisted image interpretation techniques. Results of these activities, and the state-of-the-art practice of remote sensing during this time period, are documented in the American Society of Photogrammetry's first edition of the *Manual of Remote Sensing* (1975), as well as other sources.

The first Earth Resources Technology Satellite (ERTS-1) was designed originally as a research tool. When it was launched in July of 1972, however, this satellite ushered in a brief era of emphasis by NASA on the demonstration of operational applications using this series of satellites. Through 1974, ERTS-1

(later renamed Landsat 1) transmitted over 100,000 terrestrial images covering 75 percent of the Earth's land surface. Between 1972 and 1978, major research efforts utilizing this new civil land remote sensing data were conducted in agriculture, forestry, water resources, and geology. Other notable regional-scale applications research in the 1970s included the Large Area Crop Inventory Experiment (LACIE); Agricultural and Resource Inventory Surveys Through Aerospace Remote Sensing (AgRISTARS); and the Western Snow Melt Application System Verification Test (ASVT).

In 1975, Landsat 2 was launched and the Ford Administration approved budget appropriations for Landsat 3, which was launched in 1978. In March of 1979, President Carter announced the Administration's commitment to maintaining the continuity of civilian terrestrial satellite remotely sensed data and recommended the transition of the program to the private sector. This initiated the process of program commercialization that would come to fruition in the mid-1980s. In 1981, the Reagan Administration accelerated the pace of Landsat commercialization by essentially rejecting the Carter Administration's commitment to an operational civilian land remote sensing program and advocating the immediate end of government funding for the Landsat Program.

By the middle of the 1980s, it was clear that satellite remote sensing could provide globally consistent, spatially disaggregated, and temporally repetitive coverage of the land surface of the Earth. Landsat 4, with the first Thematic Mapper instrument, was launched in 1982. Higher spatial resolution multispectral imagery was now publicly available and could be applied by scientists and professionals to the assessment of terrestrial ecosystems and land cover. A number of researchers began to develop techniques to employ the spectral information from Landsat to measure green vegetation status, and phenology, as well as interannual variation in these parameters. These studies then led researchers to the development of vegetation measures that provide an index of condition and, in part, describe absorption of sunlight, photosynthetic capacity, and evaporation rates. With this research, the science community began to understand that space-based Earth imaging can and does provide such information concerning the spatial and temporal patterns associated with land processes. During this period of capability demonstration, researchers noted that the possibility of developing fully integrated land–ocean–atmosphere monitoring and modeling capabilities were essentially realized.

In March 1984, Landsat 5 was launched and in July of that year the US Congress provided the legislative authority to transfer the Landsat program to the private sector. Thus began the nearly decade long period of 'commercialization' of the US civil land remote sensing program. Commercialization (and the subsequent price increases for Landsat data) caused a shift in scientific research emphasis. Researchers began to move away from Landsat imagery and began to design experiments utilizing NOAA polar orbiting meteorological satellites and their Advanced Very High Resolution Radiometer (AVHRR) sensor system as a primary data source for scientific research. This shift was the result of increased costs, greater restrictions on sharing and more limited availability of Landsat data to both the basic and applied land remote sensing community.

The US Government would not regain operational control of the Landsat program until the late 1990s. Essentially, though the law changed in 1992, the governmental control of any orbital assets did not occur until Landsat 7 was launched in April 1999. The commercial pricing and data sharing restrictions prior to this transition back to government pricing had a negative effect on the developing research and application programs at the time which sought to utilize satellite imagery but could not employ it due to the high data cost. It would not be until the twenty-first century that the Landsat program would begin to come under the control of an operational land-oriented federal agency, the United States Geological Survey (USGS).

Despite these difficulties, research using Landsat data during this time included work in many disciplines and application areas. Organizations important in formulating the types of research conducted at this time include the Intergovernmental Panel on Climate Change (IPCC), the International Council of Scientific Unions (ICSU), the International Geosphere Biosphere Program (IGBP), and the Human Dimensions of Global Change (HDGC) project. Of particular significance was the research that validated the use of multitemporal Landsat and other data to monitor and assess tropical deforestation; and also unambiguously demonstrated the value of time series data to provide information on change occurring in every area of the world.

Other countries entered the space imaging arena with the launch of the French Système Probatoire de la Observation de la Terre (SPOT) satellite in 1986 and the Indian Remote Sensing Satellite (IRS-1) in 1988. In 1990, the French launched SPOT 2; a year later the European Radar Satellite (ERS-1) was launched, aimed primarily at oceanographic applications, followed shortly after by the Russian radar satellite Almaz; 1992 saw the launch by the Japanese of the first Japanese Earth Resources Satellite (JERS-1). In 1995, ERS-2 was launched and Canada launched Radarsat, with a primary objective of aiding navigation of shipping in the far north.

The most recent advances, at the end of the twentieth century, involved the launch by US agencies and organizations of Landsat 7 with the Enhanced Thematic Mapper (ETM+) system, the Shuttle Interferometeric Synthetic Aperature Radar (IFSAR) mis-

sion, Earth Observing System-1 (now named Terra), and the one-meter commercial IKONOS satellite (all in 1999). The Landsat ETM+data are already proving useful in a variety of studies from fire fuels mapping to deforestation and urban change monitoring. IFSAR data will be employed to improve the quality of topographical maps between about 60° N and 60° S of the Equator. Terra data is being employed to give continental to global-scale land cover change information with a temporal frequency never before available. IKONOS data give civil researchers the opportunity to see features and objects from space that could only be seen if aerial photography or intelligence satellite data were available to the user. These satellites will add immeasurably to our ability to measure, map, monitor, and model our Earth system as never before.

4. *Future Directions*

What is even more important is that by 2006, some 20 governments and organizations are expected to launch civilian and commercial high spatial resolution satellites in an attempt to capture a share of the growing market for remote sensing imagery. These satellites will have sensor systems with spatial resolutions that range from several kilometers to less than a meter. It is clear from developments such as these that a number of the long-standing restrictions that the military and intelligence community placed on satellite remote sensing are being put aside as technology advances and becomes more accessible, creating greater transparency among nations of the world. With this increasing transparency the opportunity for both positive and negative uses of remotely sensed data increase. On the one hand it may be easier to detect and identify violations of international agreements, while on the other it may be easier for states to gain tactically important information on their neighboring states.

There is already a growing interest in these data among social, economic, and behavioral users. This interest stems from the realization that whereas remotely sensed data cannot provide all the data/information needed to resolve or answer given questions it may, when combined with other information, allow data to be more reliably interpreted across space and through time. Indeed, an important aspect of remote sensing is its value as a historical record—applications in legal proceedings involving environmental contamination and ownership liability are one example. There is a growing diversity of applications to which remotely sensed data may be applied, in social, economic, or behavioral science, from population estimation, famine early warning and disaster assessments to inventorying industrial capacity, assessing transportation infrastructure, and monitoring of environmental laws and treaties.

There have been a great many advances in the field of remote sensing over the years since the invention of the camera. Yet, from an analysis of the materials covering the field of remote sensing, it is clearly evident that users interested in questions in the physical and biological sciences have been (and continue to be) able to use these data more than users in the social, economic, and behavioral sciences.

The reasons for this disparity are many and varied. They range from the inability of remote sensing to provide needed data/information to practitioners in the social, economic, and behavioral sciences to an unwillingness on the part of users in these areas to take the time to learn both the capabilities and limitations of the information that can be extracted from remotely sensed data. For some applications the spatial and temporal scales of data acquisition are not appropriate; for others the models to translate the data collected into meaningful information specific to a given need have yet to be developed. Remote sensing cannot see what is in the human mind. It can, however, document actions that we as human beings take in response to some stimuli. For example, it cannot tell us what crop a farmer may plant in a given year, but it can, in many instances, given appropriate dates of image acquisition and timely interpretation, tell us what has been planted before the crop is harvested. Indeed, given adequate ancillary data and models it can also give us an indication of how good the farmer's harvest may be.

As our ability to collect data at higher spatial, spectral, and temporal resolutions increases, so too does our ability to derive useful information from remote sensing data. Remote sensing will never replace other, more traditional methods of data collection, such as census and survey instruments. Yet it is a proven, powerful method of collecting data which, when properly interpreted and evaluated, can provide a wide variety of useful information. Information which, when taken alone or when augmented by other data/information collected by more traditional means, can create more useful data sets for social, economic, and behavioral scientists.

5. *Organizations that Provide Remotely Sensed Data*

- Center for International Earth Science Information Network (CIESIN)
- US National Aeronautics and Space Administration (NASA)
- US Geological Survey (USGS)
- SPOT Image
- GlobeXplorer
- US National Oceanographic and Atmospheric Administration (NOAA)
- Space Imaging
- Earthwatch

See also: Geographic Information Systems; Geo-g-raphic Information Systems: Critical Approaches; Information Society, Geography of; Landsat Imagery in Geography; Mathematical Models in Geography; Remote Sensing and Geographic Information Systems Analysis; Spatial Analysis in Geography

Bibliography

Bergen E T 1927 Aerial surveys for city planners. Transaction of the American Society of Civil Engineers **90**

Chicago Planning Commission 1943 *Master Plan of Residential Land Use*. Chicago Planning Commission, Chicago

Estes J E, Ehlers M, Malingreau J-P, Noble I, Raper J, Sellman A, Star J L, Weber J 1992 *Advanced Data Acquisition and Analysis Technologies for Sustainable Development*. MAB Digest 12, UNESCO, Paris

Fairchild S M 1927 Aerial photography, its development and future. *Annals of the American Academy of Political and Social Sciences* **131** (May)

Jensen J R 1996 *Introductory Digital Image Processing: A Remote Sensing Perspective* 2nd edn. Prentice Hall, Upper Saddle River, NJ

Jensen J R 2000 *Remote Sensing of the Environment: An Earth Resource Perspective*. Prentice Hall, Upper Saddle River, NJ

Joerg E L E 1923 The use of the air plane in city geography. *Annals of the Association of American Geographers* **13**(4)

Lee W T 1922 *The Face of the Earth As Seen from the Air: A Study in the Application of Airplane Photography to Geography*. American Geographical Society, New York

Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) 1998 *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC

National Research Council 1970 *Remote Sensing with Special Reference to Agricutlure and Forestry*. National Academy of Sciences, Washington, DC

Philipson W (ed.) 1997 *Manual of Photographic Interpretation*. 2nd edn American Society for Photogrammetry and Remote Sensing, Bethesda, MD

Platt R R, Johnson G R 1930 *Peru from the Air*. American Geographical Society, New York

Reeves R G, Anson A, Landen D (eds.) 1975 *Manual of Remote Sensing*. American Society of Photogrammetry, Falls Church, VA

J. Estes, K. Kline, and E. Collins

Remote Sensing and Geographic Information Systems Analysis

Remote sensing is the science and art of obtaining information about the Earth's surface through the analysis of data acquired by a device which is at a distance from the surface. Images from Earth observation satellites and aerial photographs are most commonly used. These instruments measure the electromagnetic radiation reflected or emitted by the surface, at different wavelengths (or frequencies). Geographic information systems (GIS) are computer-based systems that are used to capture, store, analyze, and display geographic information. These two techniques are used widely to assess natural resources and monitor environmental changes. Their main fields of application for terrestrial ecosystems are geology and mineral exploration, land-cover mapping, agricultural statistics, the management of irrigation, forest inventories, and the monitoring of natural disasters such as fires, droughts, floods, landslides, or volcanic activity. Remote sensing and GIS are also used for archeology, land-use planning, urban management, and in the transportation sector. Social scientists can gain insights into fine spatial and temporal dynamics of land-use changes and related social phenomena by analyzing time series of remote sensing data and by linking remote sensing to socioeconomic data using GIS.

1. Remote Sensing Systems

A number of satellite sensors have been acquiring multispectral data over the entire globe since the 1970s, at a spatial resolution of dozens of meters. The most widely used among these sensors are Landsat Multispectral Scanner (MSS) (resolution of 80×80 m), Landsat Thematic Mapper (TM) (resolution of 30×30 m), and SPOT High Resolution Visible (HRV) Imaging Instrument (resolution of 20×20 m for the multispectral product and 10×10 m for the panchromatic product) (Colwell et al. 1983, Lillesand and Kiefer 1994).

Radiance data measured from the Advanced Very High Resolution Radiometer (AVHRR) on the NOAA series of orbiting platforms have been used widely for monitoring changes in the attributes of the Earth's surface. These data cover the period from June 1981 to date. The daily NOAA AVHRR Local Area Coverage (LAC) data have a spatial resolution of 1.1 km at nadir, which is well suited to monitor seasonal vegetation dynamics at the scale of the landscape. Other useful sensors with a 1 km resolution are ERS's Along Track Scanning Radiometer (ATSR), SPOT VEGETATION and the Moderate Resolution Imaging Spectroradiometer (MODIS).

Several active microwave sensors are also available: the Synthetic Aperture Radar (SAR) on the ERS-1 and JERS-1 satellites, and the Advanced Synthetic Aperture Radar (ASAR) sensor as part of the forthcoming ENVISAT mission. The cloud-penetrating capability of microwaves make them an ideal tool to work in areas with a permanent cloud cover. Both active and passive microwave sensing are used. Active systems supply their own source of energy or illumination while passive systems sense energy that is naturally emitted or reflected from the Earth's surface, at extremely low levels however. Note that this

enumeration of sensors used for land applications is not exhaustive (see Colwell et al. 1983, Lillesand and Kiefer 1994).

A new generation of fine spatial resolution sensors is due to be launched around the turn of the twenty-first century. These sensors are capable of generating imagery with spatial resolutions as fine as 1 m in panchromatic mode and 4 m in multispectral mode. These systems are also characterized by high geometric precision, short revisit intervals, and rapid data supply. Such imagery will provide greater spatial details on the land surface and open new applications relevant for social sciences (e.g., for urban studies).

2. *Land Surface Attributes Measured by Remote Sensing*

The electromagnetic radiation reflected and emitted in different wavelength ranges by the surface is controlled by surface attributes, such as the nature of the material, surface moisture, state of the vegetation, canopy structure, etc. Different features can be identified on an image based on the specific 'spectral signature' of different surfaces. A good understanding of the biophysical processes controlling the radiation of a surface is necessary to interpret the remotely sensed signal in terms of surface attributes.

2.1 *In the Spectral Domain*

In the visible and near-infrared range of the electromagnetic spectrum, the land surface generally is characterized using vegetation indices, which are arithmetic combinations of spectral bands. The most widely used of these indices is the normalized difference vegetation index (NDVI). Vegetation indices derived from remote sensing data are related to several variables quantifying vegetation activity and state, such as the fraction of photosynthetically active radiation absorbed by the vegetation canopy, canopy attributes (e.g., green biomass or green leaf area index), state of the vegetation (i.e., vegetation vigor or stress) and instantaneous rates associated with the activity of the vegetation. Seasonal variations of vegetation indices are related to vegetation phenology and biome seasonality.

In the thermal range of the spectrum, land surface 'skin' brightness temperature (Ts) can be derived from the thermal channels of satellite sensors. Ts is related, through the surface energy balance equation, to surface moisture availability and evapotranspiration, as a function of latent heat flux. Empirical studies have demonstrated the usefulness of thermal data to locate dense forest boundaries and monitor the growth cycle of crops, including for irrigated systems.

In the microwave range of the spectrum, the intensity of microwave returning from a given surface depends on the geometric characteristics of the surface (i.e., shape and orientation of objects, as well as their surface roughness) and its electrical characteristics (i.e., reflectivity and conductivity of materials, as measured by their complex dielectric constant). The all-weather capacity of microwave remote sensing combined with the fine spatial resolution of the data opens promising avenues for land cover assessment and monitoring. This technology has however not yet reached the same operational level as optical remote sensing.

2.2 *In the Spatial Domain*

A major attribute of a landscape is its spatial pattern—i.e., the arrangement in space of its different elements. The concept of landscape spatial pattern covers, for example, the patch size distribution of residual forests, the location of agricultural plots, the shapes of fields, or the number and configuration of landscape elements (i.e., their spatial heterogeneity). Landscape spatial pattern is seldom static due both to natural changes in vegetation and human intervention. Remote sensing offers the possibility to analyze changes in spatial pattern at the scale of landscapes. Textural information of fine spatial resolution remote sensing images can be measured using a variety of indices such as variance, variograms, mathematical morphology, or fractals.

3. *Processing and Analysis of Remotely Sensed Data*

3.1 *Preprocessing*

Technical issues of geometric correction and rectification, radiometric correction, geometric and radiometric enhancement, and multispectral transformations of image data are described in Richards (1993). The purpose of these operations is to georeference and to transform the image data in a form that can be readily interpreted in terms of Earth surface attributes. The analysis and interpretation of remote sensing data typically is performed through image classification, to generate land-cover maps, or through change detection, to monitor land-cover changes.

3.2 *Image Classification*

Classification is a method by which labels are attached to spatial sampling units (i.e., the pixels) in view of their character. This character is generally their spectral response in different spectral ranges. It may also include their spatial attributes (i.e., texture) or temporal changes. This labeling is implemented through pattern recognition procedures, the patterns being

pixel vectors. The most commonly used classification methodology is based on the maximum likelihood, a probabilistic classification method (Richards 1993). More advanced classification techniques have been developed. For example, contextual classifiers label a pixel in the context of its neighbors in space. They exploit spatial information of an image. Neural networks are classifier networks, having a decision tree structure, where a collection of simple classifiers is brought together to solve a complex problem.

3.3 Change Detection

Time series of remote sensing data can be created to monitor changes in landscapes at local to regional scales. Classic land-cover change detection techniques are based on the comparison of sequential land cover maps derived from remote sensing data or other sources for the same area. For every sampling unit of the maps, the land cover categories at the two dates are compared. The comparison of successive maps fails to detect subtle changes within broad land-cover classes. Thus, rather than detecting changes on the basis of land-cover categories, change detection is better performed on the basis of the continuous variables defining these categories, whether these are reflectance values measured by a satellite sensor or biophysical attributes derived by model inversion. Coppin and Bauer (1996) review techniques used for this comparison. Recent studies have detected changes in land-cover between successive dates by combining different methods (e.g., Macleod and Congalton 1998).

4. Components of a GIS

A GIS is a powerful tool for handling spatial data. It is designed for the collection, storage, and analysis of objects and phenomena where geographic location is an important characteristic (Aronoff 1989, Maguire et al. 1991). The data input component of a GIS converts data from their existing form to one that can be used by the GIS. Two types of data are entered in a GIS: spatial data such as points, lines and areas that represent the geographic locations of features, and associated nonspatial attribute data that provide descriptive information. The data management component includes those functions needed to store and retrieve data from the database. In a GIS, data are maintained in a digital format. The representation of the spatial component of geographic information follows either the raster model—the homogeneous units are regular cells—or the vector model—the spatial units are the points, lines, and polygons. The data analysis and manipulation functions determine the information that can be generated by the GIS. A GIS automates certain activities and data interpretation procedures. Some of the most commonly used GIS functions are classification, overlay operations, neighborhood operations, spatial interpolations, geostatistical analyses, and connectivity functions. The output or reporting functions of a GIS produce maps, tables, or statistical results.

GIS facilitates the interpretation of remote sensing data, by linking biophysical information measured by remote sensing to maps of natural and cultural landscape attributes (e.g., soil type, topography, accessibility as measured by various distances), to field measurements (e.g., crop yield or water quality) and to socioeconomic information derived through household surveys (e.g., indicators of well-being, rate of population growth, agricultural input use). For a discussion of the ideas and social practices that have emerged with the development of new forms of data handling and spatial representation in GIS, see Pickles (1995).

5. The Contribution of Remote Sensing and GIS to Social Science Studies

The scientific literature is becoming increasingly rich in case studies using remote sensing to study social phenomena. Actually, remote sensing techniques are growing in importance for studies aimed at understanding land use dynamics, its driving forces and its impact on society (see Behrens 1994, Liverman et al. 1998). The output of remote sensing and GIS analyses are major inputs for spatial models. Remote sensing data is supporting land-use change analysis in a variety of ways.

5.1 Land-cover Mapping

Remote sensing is used widely to map land cover, land use, and agricultural patterns. The current classification algorithms succeed in identifying from 10–15 land-cover classes, including several succession stages of vegetation. Achieving this level of detail requires however a set of georeferenced field observations on a sample of locations. Current sensors do not identify very well settlements or different cropping systems. Spatially heterogeneous or fragmented classes are identified as mosaics of cover types. The further analysis of land use/land cover maps can lead to the separation of 'landscape units,' characterized by different combinations of land use and biophysical attributes.

5.2 Monitoring Land-cover Change

Remote sensing facilitates the introduction of a dynamic perspective in land use studies, by monitoring changes in the spatial patterns of landscapes caused by land use. Sequential remotely sensed data allows to detect changes in land use, such as an expansion of the area under permanent or shifting agriculture, urban growth, or increase in eroded areas. For example, quantitative estimates of rates of deforestation are

derived from time series of remote sensing data. The spatial extent and location of secondary growths on abandoned plots is also well detected by remote sensing. Remote sensing has been used in semi-arid regions to assess dryland degradation and to identify transformations of land use and cultivation. Measuring rates of dryland degradation is a complex challenge, however, given the strong interaction between erratic fluctuations in rainfall due to climatic variability and anthropogenic changes in vegetation cover.

5.3 Identification of Land-cover Change 'Hot Spots'

Land-cover change 'hot spots' are defined as areas where high rates of land-cover changes have been observed. Remote sensing-based methods for identifying land-cover change hot spots at broad spatial scales have been developed. Such methods are based on a series of indicators, such as land-cover fragmentation, occurrence of fires, opening of new roads, etc., which are detectable, by remote sensing. Deforestation, for example, is strongly related to the proximity to forested/deforested edge, the proximity to an access and the degree of fragmentation of forests. Forest fragmentation can be quantified easily from remote sensing data.

5.4 Identification of Land-use Attributes

Remote sensing allows identification and measurement of key socioeconomic and ecological characteristics of land use systems. In the Guinea Highlands, Gilruth and Hutchinson (1990) have been able to discriminate, on the basis of remote sensing techniques, between permanent agriculture (in the form of home gardens) and shifting cultivation. The spatial distribution of shifting cultivation areas can also be mapped from fine to medium resolution satellite data. Guyer and Lambin (1993) succeeded in discriminating and quantifying the total area and proportion of tractor-cleared and hand-cleared fields in a region of Nigeria by using several shape criteria derived from multispectral SPOT data. These authors also computed the crop-fallow cycle and the importance of the land reserve using a remote sensing-based land-cover map.

5.5 Spatial Stratification and Spatial Extrapolation

Remote sensing allows testing of how widely local observations apply to larger areas. Actually, a common problem of in-depth investigations of land-use dynamics is that it requires a large data collection effort in the field, at the household and plot levels. Remote sensing, through the measurement of indicators of land use, helps to define regions characterized by a similar land use, that can lead to a spatial stratification of a study area in terms of land use patterns. As a result, findings obtained over specific sites within a stratum, that is homogenous in terms of land use, can be extrapolated with more confidence to that stratum. If such a spatial stratification is performed before field surveys, it is a powerful tool to design a sampling strategy and locate villages or households that will be used for in depth data collection and investigation. This leads to an optimization of the field data collection effort.

In a study of land use intensification (Guyer and Lambin 1993), remote sensing analysis provided a means of matching the patterns from a haphazard sample of households to the population, based on a few variables that were estimated both by household survey (on the sample) and by remote sensing (on the entire population). The results tended to give greater confidence to inferences from the sample about ecological dynamics.

5.6 Input into Spatial Models

The spatial patterns, which are detectable at a variety of spatial scales on remotely-sensed data, provide the most obvious link between spatially-explicit land-use change models and remote sensing analysis. With the application of remote sensing technology to the monitoring of land-use changes, it becomes possible to integrate more explicitly spatial relationships into models of land-use change. It is also possible to apply and test these models at a finer spatial resolution and with samples comprising smaller geographic units—e.g., at the scale of individual landscape units.

Spatial, statistical models are born from the combination of remote sensing, GIS, and multivariate mathematical models. Their emphasis is on the spatial distribution of landscape elements and on changes in landscape patterns. The goal of these models is the projection and display, in a cartographic form, of future landscape patterns, which would result from the continuation of current land-management practices or the lack thereof. The approach analyses the location of land-cover changes in relation to maps of natural and cultural landscape variables in a GIS. A model is built to describe the relationship between the dependent variable—e.g., land-cover changes—and the independent landscape variables. Spatial, statistical analyses of tropical deforestation or dryland degradation have already been developed for several tropical regions (see review by Lambin 1997). Spatial models of desertification hazard have also been developed by integrating indicators of water and wind erosion, vegetation degradation, range utilization, and human settlement.

5.7 Validation of Spatially-explicit Models

Dynamic spatial simulation, actor-based models have been developed to predict spatial patterns of land-use

changes at the landscape scale—e.g., for tropical deforestation in roadside areas occupied by shifting cultivators (e.g., Wilkie and Finn 1988) or in regions of official resettlement schemes associated with road development (Southworth et al. 1991). These grid-cell models combine spatially-explicit ecological information with socio-economic factors related to land use decisions by farmers. The modeling is based on rules of behavior determined by locational attributes, land tenure system, and vegetation successions. Remote sensing offers a powerful tool to calibrate and/or validate these models by comparing actual and predicted land-use/cover patterns, and by comparing the level of fragmentation of the landscape.

6. *'Socializing the Pixel'*

By 'socializing the pixel' one means to discern information embedded within spatial imagery that is directly relevant to the core themes of the social sciences, and use it to inform the concepts and theories pertinent to those themes (Geoghean et al. 1998). Creating a direct link between spatially-explicit land cover information, as derived by remote sensing, and information on land-use change processes requires the development of new methods and models which are merging landscape data with data on human behavior.

6.1 *Linking Household-level Data with Remote Sensing Data*

A synthesis of the methods for merging socio-economic household survey data and remote sensing-based information is given in Liverman et al. (1998). The major challenges of this approach are: (a) the definition of the appropriate spatial observation units, i.e., the appropriate level of aggregation of information derived from the domains of social phenomena and natural environment, and (b) the development of the appropriate linkages between household-level and remote sensing datasets. While conceptually straightforward, these links can be difficult to implement operationally. In remote sensing, the unit of observation is the pixel that is not directly associated to any social science unit of observation, e.g., individuals, households, or villages. Establishing the correspondence between biophysical and socioeconomic variables requires taking into account the unit of decision-making, units of landscape transformation, and the spatial scale of ecological processes (e.g., a watershed).

Linking remote sensing observations to socio-economic data at the scale of the administrative units which were used for the collection of socioeconomic data leads to a loss of information as it obscures the variability within the units. More recently, research efforts have attempted to integrate remote-sensing observations and field surveys at finer levels of aggregation, i.e., at the scale of individuals, households, or villages (e.g., Entwisle et al. 1998). There are difficulties in relating remotely-sensed patterns of land-cover changes with field observations of land-use changes since, in many cases, people live in nucleated villages away from their fields, since households may cultivate multiple non-contiguous plots and since people move—especially pastoralists. For this reason, the integration of the two datasets is generally performed at the village level.

Individual household data allow for a better understanding of the land-use practices within each village, as most land-use decisions are made by individuals and households. Moran and Brondizio (1998) have investigated the linkages between remotely sensed data and traditional field methods in the social and biological sciences. Their results have led them to new research questions, including examination of the role of the developmental cycle of the household in shaping their trajectory of land-use and deforestation.

6.2 *Landscape Spatial Patterns*

The spatial pattern of a changing landscape have some information content on the processes of land-cover change. Certain categories of land-use changes tend to fragment the landscape (e.g., expansion of smallholder farming, small-scale logging, overgrazing around deep wells). Other land-use changes increase landscape homogeneity (e.g., mechanized cultivation or ranching over large areas). There is a good correlation between remotely-sensed spatial patterns and some important characteristics of farming systems. For example, Lambin (1988) found statistically that, in Burkina Faso, spatial patterns interpreted from Landsat MSS data correlate more closely with ethnic groups than with any other physical or cultural landscape variables. Other authors demonstrated that the analysis of the spatial pattern of forest-nonforest interfaces allows identification of regions affected by different deforestation processes, which are controlled by different driving forces.

The spatial patterns of fires, which are well detected from space, are also good indicators of human activities. In tropical forests, regions of fire concentration indicate frontier areas and variations in the spatial organization of fires are associated strongly with political boundaries and farming practices.

6.3 *GIS for Health and the Environment*

An emerging area of research concerns the application of remote sensing and GIS techniques to analyze the

relationships between people, their environment and their health (de Savigny and Wijeyaratne 1995). GIS provides information about the spatial distribution of and interaction between disease risk factors, patterns of morbidity and mortality, and the allocation of health resources. Vector-borne diseases (e.g., malaria, trypanosomiasis) have received most emphasis so far, as their occurrence is strongly controlled by environmental attributes which are detectable by remote sensing (Barinaga 1993). However, there is a large potential for GIS application on any environment-related health problem. A classic example is the spatial correlation between cholera deaths and contaminated water supplies.

See also: Geographic Information Systems: Critical Approaches; Geography; Globalization: Geographical Aspects; Spatial Analysis in Geography; Telecommunications and Information Policy

Bibliography

Aronoff S 1989 *Geographic Information Systems: A Management Perspective*. WDL Publications, Ottawa, Canada

Barinaga M 1993 Satellite data rocket disease control efforts into orbit. *Nature* **261**: 31–32

Behrens C A 1994 Recent advances in the regional analysis of indigenous land use and tropical deforestation: Introduction. *Human Ecology* **22**: 243–47

Colwell R N, Simonet D S, Estes J E 1983 *Manual of Remote Sensing*. American Society of Photogrammetry, Falls Church, VI

Coppin P R, Bauer M E 1996 Digital change-detection in forest ecosystems with remote sensing imagery. *Remote Sensing Review* **13**: 207–34

de Savigny D, Wijeyaratne P 1995 *GIS for Health and the Environment*. International Development Research Center, Ottawa, ON

Entwisle B, Walsh S J, Rindfuss R R, Chamratrithirong A 1998 Land-use/land-cover and population dynamics, Nang Rong, Thailand. In: Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Acadamy Press, Washington, DC

Geoghegan J, Pritchard L, Ogneva-Himmelberger Y, Chowdhury R R, Sanderson S, Turner II B L 1998 'Socializing the pixel' and 'pixelizing the social' in land-use and land-cover change. In: Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Acadamy Press, Washington, DC

Gilruth P T, Hutchinson C F 1990 Assessing deforestation in the Guinea highlands of West Africa using remote sensing. *Photogram Engineering and Remote Sensing* **56**: 1375–82

Guyer J, Lambin E F 1993 Land use in an urban hinterland: Ethnography and remote sensing in the study of African intensification. *American Anthropologist* **95**: 839–59

Lambin E F 1988 L'apport de la télédétection dans l'étude des systèmes agraires d'Afrique; l'exemple du Burkina Faso. *Africa* **58**: 337–52

Lambin E F 1997 Modelling and monitoring land-cover change processes in tropical regions. *Progress Physical Geography* **21**: 375–93

Lillesand T M, Kiefer R W 1994 *Remote Sensing and Image Interpretation*. Wiley, New York

Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) 1998 *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC

Macleod R D, Congalton R G 1998 A quantitative comparison of change-detection algorithms for monitoring eelgrass from remotely sensed data. *Photogram Engineering and Remote Sensing* **64**: 207–16

Maguire D J, Goodchild M F, Rhind D W 1991 *Geographical Information Systems: Principles and Applications*. Longman, Harlow, UK

Moran E F, Brondizio E 1998 Land-use change after deforestation in Amazonia. In: Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Acadamy Press, Washington, DC

Pickles J 1995 *Ground Truth: The Social Implications of Geographic Information Systems*. The Guildford Press, New York

Richards J A 1993 *Remote Sensing Digital Image Analysis*. Springer-Verlag, Berlin

Southworth F, Dale V H, O'Neill R V 1991 Contrasting patterns of land use in Rondonia, Brazil: simulating the effects on carbon release. *International Social Science Journal* **130**: 681–98

Wilkie D S, Finn J T 1988 A spatial model of land use and forest regeneration in the Ituri forest of Northeastern Zaire. *Ecological Modelling* **41**: 307–23

E. F. Lambin

Renaissance

Two nineteenth-century historians were instrumental in giving the term 'Renaissance' the content that it retains to this day. In his 1855 *Histoire de France*, Jules Michelet (1798–1874) chose the word for the period of the *découverte de l'homme et du monde*, the sixteenth century. But the Swiss Jacob Burckhardt (1818–97) was even more important. His famous book *Die Cultur der Renaissance in Italien. Ein Versuch* (1860) became a milestone of research. Burckhardt sketched a monumental fresco that located the origin of modernity in fifteenth century Italy—in conscious distinction from Hegel's philosophy of history, which posited that the German Reformation marked the beginning of the modern age. Burckhardt referred to the conditions in the Italian ministates as 'masterpieces of statecraft,' in which he saw the preconditions for the 'emergence of the individual' as the most important and far-reaching development of the modern age. He presented the Renaissance, with the many-faceted Humanistic movement as its intellectual substrate, as a period that forced the pace of modernization and secularization and as the beginning of the modern

world in which he himself lived and whose negative sides he analyzed perhaps more acutely than any other thinker of his time.

1. The Development of the Concept of the Renaissance

1.1 Renaissance

Burckhardt's thought-provoking image of the epoch exerted substantial influence, not only on the study of history, but also on literature and philosophy, in particular on Nietzsche's thinking. It remains the more or less explicit background of discussions of the term 'Renaissance,' though it soon roused dissension. For example, in the book *Franz von Assisi und die Anfänge der Renaissance in Italien* (1885), Henry Thode (1857–1920) attempted a romantic reinterpretation, placing the *poverello* and the religious movement he inspired at the beginning of the epoch. Thode and his followers saw the Renaissance more as a reform movement on a medieval-clerical foundation, eliminating its character as an epoch of its own. Thode rejected the idea that classicism and rationality fundamentally constituted the Renaissance and divided it from the Middle Ages. He posited an individualism of piety and sensitivity to beauty, in contrast to Burckhardt's subjects becoming aware of themselves.

But the prospect of another, religious side of the Renaissance had already existed before Thode and did not completely contradict Burckhardt's view. In his influential essay collection *The Renaissance* (1873), Walter Pater (1839–94), an admirer of Ruskin presented several artists and thinkers he considered representative of the Renaissance. The essay on Botticelli became essential to the incomparable cult of this painter, which thrived throughout Victorian society, and not solely among the Pre-Raphaelites. Botticelli's art seemed to correspond to a 'metaphysical need' in the modern, industrialized world of Victorian England.

The inventions of the historians and literary writers stimulated broad segments of the educated bourgeois class. Florence became the preferred destination for an international cultural tourism. People built in the Renaissance style, surrounded themselves with splendid interiors in the style of the fifteenth and sixteenth centuries, and purchased original works of art. Especially in the second half of the nineteenth century, collections of important Renaissance art were amassed from Boston to Berlin. The 'Renaissance' was not only a focus of scholarly interest, but also part of a lifestyle.

Around 1900, the international Renaissancism reached a zenith. Burckhardt's classic was reprinted, and novels like Heinrich Mann's 'Göttinnen' depicted the sixteenth century as a foil for the contemporary 'decadent' *fin de siècle*. The closing nineteenth century saw itself not only in a strange spiritual kinship with the Renaissance: the fiction of a 'lascivious,' 'unscrupulous' epoch paradoxically appeared to offer counterimages to the usual bourgeois existence, taking shape in princes of commerce, patrons, *condottieri*, and courtesans.

At this time, the term 'Renaissance,' was loaded with a plethora of sometimes contradictory meanings, triggering associations of 'gold and purple,' as Johan Huizinga said. The patronage of the Medici, the magnificence of the papal court, and the art of the Renaissance with its heroes Leonardo, Raphael, and Michelangelo provided new cult objects for religious worship in a demystified world. At the same time, the word recalled the art of power and *Realpolitik*, as Machiavelli had analyzed in his trailblazing work. Racist ideas of a victory of hard men and cunning women over decline and enfeeblement were inherent in this image of the epoch. A widely read literature in the succession of Carlyle, Count Gobineau, Nietzsche, and others popularized the cliché of the 'blond beast,' the amoral 'Renaissance man' up to mischief.

The myth of the Renaissance thus offered models for Caesarism and imperialism. It was no coincidence that a nationally-oriented writing of history took up Italy's *Signorie* and identified them with the precursors of the modern military and power-wielding state. But the iridescent construct of the 'Renaissance' also offered fragments for a historical foundation of modern civic societies. As early as the beginning of the nineteenth century, the Swiss Sismonde de Sismondi had described the Italian city republics in this way, and travelers from England and the USA thought they were tracing the ideal roots of their own nations when they came to Florence. It is hardly surprising that the *inglesi* were the largest group among the tourists. The Florence of the Medici was long a primarily Anglo-American myth.

Examples of the ideological interpretation of the Renaissance with an unmistakable relationship to current history are also found in the works of Hans Baron, a German Jew who emigrated to the USA. In the essay he published in the middle of World War II, he saw the crisis of Italian 'freedom' caused by the aggression of the Visconti *signoria* in the early Quattrocento as the precondition for the formation of a modern, pluralistic state system on the Italian peninsula. In Baron's view, the high degree of citizen participation in Florentine politics was a precondition for this most important result of the struggle against the tyrants: He saw it as a kind of Italian America winning the upper hand over Visconti/Hitler (Baron 1955).

1.2 The Revolution of the Medievalists

More recent research on the Renaissance has tried to free itself of mythologizing. Insight has been sought into the clientele systems of the clan that enabled

Cosimo de' Medici and above all Lorenzo the Magnificent to exercise an authoritarian power behind a republican façade (Wackernagel 1938, Thomas 1995), whereby special attention was paid to the patrons commissioning the art (Esch and Frommel 1995). But art history tended to go its own way. Wölfflin and others separated Renaissance art from the cultural-history context and treated it primarily as a problem of style, and thus of aesthetics.

The conception of the art historian Aby Warburg (1866–1929) remained fragmentary and accessible only in a very few essays. It aimed once again at the great synthesis and still provides impetus for new research. He located the epoch in a long-term process of civilization that experienced its climax in the Apollonian beauty of the art of the High Renaissance. In the triumph of the Olympic gods, Warburg saw an important step toward overcoming 'oriental,' 'medieval' irrationality. He was interested in the difficult breakthrough to the Enlightenment, humanity and reason in the history of the West. Warburg saw works of art as documents of the states of the human psyche, and he believed that, through subtle analysis, he could uncover the *Zeitgeist* preserved in them. Warburg's works addressed aspects previously not or only marginally considered. The role of the patron became a theme, along with what Michael Baxandall (1972) calls 'troc': the market of intellectual and material goods where what we call 'Renaissance art' took shape. Warburg sought to illuminate the dark side of the epoch—its interest in divination, astrology, and magic, to which Burckhardt had devoted only a few passages. Equally important was that he conceived the emergence of Renaissance culture as a transfer process and thus examined not only the Orient, but also the culture of the North—the courts of Burgundy, France, and the Netherlands. This underscored what is now the general possession of research: the cultural physiognomy of the Renaissance as an amalgamation of varying influences. Italy remained important, but it was placed in the context of a network cast across the continent (Porter and Teich 1992).

This geographical expansion, the focus on the 'occult Renaissance,' and above all his constant question of the beginnings led Warburg to avoid drawing the boundary between the 'Renaissance' and the 'Middle Ages' too sharply. Although his works had no discernible importance for it, he prepared the ground for what Wallace K. Ferguson has termed a 'revolt of the medievalists' (Ferguson 1948). This debate, characteristic of the 1920s, adopted aspects of the image of the Renaissance as sketched by Thode, Sabatier, and other researchers. It pointed out, for example, that the tradition of antiquity was never completely buried and recalled that various 'renaissances' were palpable between the ninth and thirteenth centuries, in which ancient forms clearly stand out as elements shaping culture. Some researchers, including Etienne Gilson and Johann Nordström saw the Renaissance as a mere continuation of currents existing in the twelfth century. In this famous book *The Waning of the Middle Ages* (1919), the Dutch writer Johan Huizinga drew an image of the fifteenth century diametrically opposed to Burckhardt's conception. Contemplating the courtly world of Burgundy, France, and the Netherlands, he depicted the epoch as a flowering 'autumn' of burgeoning color, rather than as the spring of a new age.

2. *A Movement of the Happy Few: The Sociohistorical Dimension*

Discussion thus ensued on the boundaries of epochs; it was the hour of the deperiodizers, who made the Renaissance their preferred target. George M. Trevelyan went as far as to locate the turn from the Middle Ages to the Modern Age in the epoch of the Industrial Revolution, the eighteenth century; and for other researchers as well (like Delio Cantimori), the Renaissance of the fourteenth and sixteenth centuries lost its meaning as a significant turning point. A more differentiated sociohistorical research underscored that Renaissance societies were more 'medieval' than widely conceived. Additionally, this culture had been borne by a tiny minority—one of the reasons Ernst H. Gombrich no longer speaks of the Renaissance as an epoch, but as a movement.

It is indeed indisputable that epochs named for styles can describe little more than the cultural stances of an often compoundly interwoven elite. Social history has provided glimpses of the world of the majority of the 'Renaissance people'; these reveal little 'Renaissance' quality (e.g., Herlihy and Klapisch-Zuber 1978, Davis 1982, Squatriti and Merideth 1999). Recently, the theme of gender has also been considered more intensely (see King 1991, with comprehensive bibliography); here, research focuses not only on the 'Renaissance women,' the princesses and courtesans who populate the mythological scenario, but also on other women who lived at the time—witches and saints, nuns, artists, authors, and women in working life in general (Ferguson 1986).

3. *'Modern Individualism' and the Concept of the Renaissance Today*

3.1 *Art Historical Findings*

Despite the still current discussion of the term 'Renaissance,' the concept has prevailed as the name of a cultural epoch (cf., e.g., Burke 1972, Delumeau 1967, Hale 1993). What arguments can be adduced for or against the concept of the Renaissance? The most conspicuous are the findings of art history, and they

were what gave the essential impetus to construe the epoch in the first place. Thus, it is beyond doubt that the painting of Giotto and Massaccio and the sculpture of the Pisani in the fourteenth and fifteenth centuries marked a fundamental transformation. What was decidedly new in painting was that dependence on the Byzantine model was overcome. Here, Flemish influences played a major role, and the Renaissance in general increasingly appears as a complex amalgamation of transfer processes embracing Europe, the Middle East, and finally the New World. The early Quattrocento brought the introduction of a central perspective formulated with geometric precision, a Florentine invention. The goal was to create the illusion of 'real' living beings, objects, and spaces.

One of the most important arguments for Burckhardt's old 'individualism hypothesis' is still that not only did artists' signatures appear even more frequently, but also that, in the fifteenth century, initially in the Netherlands and Italy, portraits were painted that appear closer to nature than anything since classical antiquity. Since the Trecento, sculpture once again clearly approached antiquity's conceptions of the body, and, increasingly frequently, portrait busts clearly strove to convey the individuality of the person portrayed. While it cannot be claimed that the 'modern individual' was invented in the fifteenth century, if we also adduce literary testimony in the broadest sense, for example Ghiberti's *Autobiography* and Giovanni Ruccelai's unique *Zibaldone*, we do note that reflection on the self gained decisively in importance (Davis 1986 is highly differentiating; see also Schulze 1996, Porter 1997, Schweikhart 1998).

In architecture, the building forms of classical antiquity reappeared first and initially exclusively on the Italian peninsula. Only at the beginning of the sixteenth century was anything remotely comparable found north of the Alps, starting with the buildings commissioned by the Fuggers in Augsburg. Regional variant forms emerged rapidly. Their common denominator was a more or less clearly articulated reference to antiquity. With Mannerism, the Renaissance flows into a typically European style (cf. Sect. 5).

3.2 The Revolution in Discourse and Renaissance Humanism

The art of the Renaissance developed in a complex content with a theoretical discourse that, on the one hand, used texts from antiquity, for example Vitruvius' theory of architecture. This was new and genuinely Italian. In the fifteenth and sixteenth centuries, no other European country produced a theory and critique of art whose quality and differentiation were even remotely comparable. Italy was also the first place with a palpable group of collectors, art agents, and patrons who, for aesthetic considerations, acquired real or putative ancient objects. These inspired an almost romantic veneration and learned discussion and correspondence. One of the earliest representatives of this modern phenomenon of collecting was the Florentine Niccolò Niccoli (1364–1437).

Interest in ancient and contemporary art was part of the humanistic culture that was closely tied to the Renaissance. The term *studia humanitatis* is of ancient origin and refers to education in the broadest sense, the acquisition of knowledge as well as of 'good' ways of living. The fifteenth and sixteenth centuries saw the emergence of the social type of the *humanista*, the scholar concerned with the canon of humanistic studies—grammar, rhetoric, poetics, history, and moral philosophy—with recourse to the authors of classical antiquity. Humanistic scholars and humanistically educated elites became the real bearers of the Renaissance. They gave art, architecture and literature essential elements of form and content (see e.g., Wind 1968, Panofsky 1962). Modern research sees the first beginnings in upper Italy in the second half of the thirteenth century. But Francesco Petrarca (1304–74) was the first thinker to unite in himself the essential aspects of the humanistic movement. The older idea that the humanists were a group of traveling scholars staging a revolution against scholastic theology and Church doctrines is no longer tenable. Social history cannot view the host of humanists as homogeneous, even as early as the fifteenth century. The majority of the humanists of Florence, the 'capital' of the Renaissance, were economically well off (Martines, 1968). Monks, prelates, princes, university scholars, urban jurists, and many others spread the *studia humanitatis* in Europe, and in many places humanism established itself as a national movement with national characteristics.

Humanism, and with it the culture of the Renaissance, were in no way pagan. The vast majority of humanists were firmly based in Christianity, and this did not change with the Reformation (Verdon and Henderson 1997). Significantly, pictorial subjects remained mostly religious during the Renaissance. In the sixteenth century, however, we note their gradual dwindling (Burke 1972), a telling indication that Renaissance humanism contained substantial potential for secularization. The great expansion of the spectrum of questions discussed and written about in the fifteenth and sixteenth centuries was of long-term importance. Theology and religion still dominated, but texts on every possible aspect of life increasingly moved to the foreground; finally, and again with recourse to ancient thinkers, a new theory of politics developed (Pocock 1975). This 'differentiation of discourse' did not depend on the Gutenberg Revolution, but began much earlier. In a certain way, the invention of the letterpress with movable type can be viewed as a response to a market ravenous for written material.

4. The Research Problem of the Genesis of the Renaissance

A question equally fascinating and controversial is why the beginnings of humanism and the Renaissance lay in Italy and what led to the formation of these movements. As Burckhardt had already underscored, there was the multiplicity of *signoria* and city republics, whose political rivalry sought artistic representation. Ancient monuments, statues, and other relics also provided models. Italy and Flanders, the regions in which the new culture developed, were probably the richest in Europe at the time, so here is where the economic prerequisites for extensive patronage were given. But the connection between art and economy should not be absolutized. There is no necessary parallel between the unfolding of Renaissance culture and a general economic upswing; Roberto S. Lopez (1963) even attempts to interpret the Renaissance as an attempt to compensate for crisis-like developments. The financial position of the patrons was decisive (see e.g., Roover 1963). The *condottieri* wars in fifteenth-century Italy resulted not merely in a destruction of resources, but also in processes of redistribution from the great centers to small seats of courts like Urbino and Bergamo, where successful military commanders spent astronomic sums on books and in patronizing the arts. It is difficult to estimate the significance of the dynamics of economic shrinkage triggered by plague epidemics and failed harvests in the first half of the fourteenth century. Along with the resulting acceleration of urbanization, the consolidation of capital in the hands of the survivors contributed to the financing of early Renaissance art—in part as an 'investment in saving one's soul,' which must have suggested itself and seemed necessary against the backdrop of the crisis of Church hierarchy and the experiences of the plague era. The Renaissance, humanism, and the Reformation thus appear as varying answers to one and the same problem, namely the determination of the human position in a fundamentally changed world. The writings of Giannozzo Manetti and Pico della Mirandola, the political theories of Machiavelli, the primary Reformation writings of Luther and the invention of the realistic portrait and of central perspectives thus all had the same preconditions.

5. The 'Mannerist' Renaissance

The question of the end of the Renaissance is no less complex and controversial than that of its beginnings. Art history sees indications of a metamorphosis to the 'Mannerist Renaissance' or to mannerism as early as the second decade of the sixteenth century (Shearman 1967, Arasse and Tönnesmann 1997). Literary history usually locates the end of the epoch around 1600. And music history, too, has its difficulties with the question of the definition and temporal boundaries of mannerism (important but controversial, see Maniates 1979). National differences in particular must be differentiated. Setting the boundaries of the epoch is also made difficult by the fact such important intellectual currents constitutive of the Renaissance as humanism and such aesthetic ideals as the orientation toward the standards of Greek classicism lived on after 1600 and experienced their own renaissances.

There has been no lack of attempts to set the stylistic transformation of the sixteenth century in direct relationship to political and societal changes (the best known is Arnold Hauser 1979). Structural changes at least in the art market and thus in the relationships between artists and patrons probably played an important role. Europe's princes, the rulers of the rising early modern states, moved increasingly into the foreground as patrons. They used the arts to buttress their legitimacy and to display their grandeur: now, in the sixteenth century, the art of the power availed itself of the power of art more than hardly ever before. Political competition also manifested itself as competition among patrons of the arts. The artists' virtuosity, or *maniera*, had to prove itself in this complicated field.

See also: Art History; Burckhardt, Jacob (1818–97); Middle Ages, The; Reformation and Confessionalization

Bibliography

Arasse D, Tönnesmann A 1997 *La Rénaissance Maniériste*. Gallimard, Paris

Baron H 1955 *The Crisis of the Early Italian Renaissance. Civic Humanism and Republican Liberty in an Age of Classicism and Tyranny*. Princeton University Press, Princeton, NJ

Baxandall M 1972 *Painting and Experience in Fifteenth Century Italy. A Primer in the Social History of Pictorial Style*. Oxford University Press, Oxford, UK

Brucker G 1977 *The Civic World of Early Renaissance Florence*. Princeton University Press, Princeton, NJ

Burke P 1972 *Culture and Society in Renaissance Italy*. B. T. Batsford, London

Burke P 1998 *The European Renaissance*. Blackwell, Oxford, UK

Davis N Z 1982 Women in the crafts in sixteenth-century Lyon. *Feminist Studies* **8**: 47–80

Davis N Z 1986 Boundaries and the sense of self in early modern France. In: Heller T C et al. (eds.) *Reconstructing Individualism*. Stanford University Press, Stanford, CA

Delumeau J 1967 *La Civilisation de la Renaissance*. Editions Arnaud, Paris

Esch A, Frommel C L 1995 (eds.) *Arte, Committenza ed Economia a Rome e nelle corti del Rinascimento*. Giulio Einaudi, Turin, Italy

Ferguson W K 1948 *The Renaissance in Historical Thought. Five Centuries of Interpretation*. Houghton Mifflin/The Riverside Press, Cambridge, MA

Ferguson W K 1986 (ed.) *Rewriting the Renaissance. The Discourses of Sexual Difference in Early Modern Europe*. University of Chicago Press, Chicago
Goldthwaite R A 1993 *Wealth and the Demand for Art in Italy, 1300–1600*. Johns Hopkins University Press, Baltimore, MD
Hale J 1993 *The Civilization of Europe in the Renaissance*. Harper Collins, London
Hauser A 1979 *Der Manierismus. Die Krise der Renaissance und der Ursprung der modernen Kunst*. Beck, Munich, Germany
Herlihy D, Klapisch-Zuber C 1978 *Tuscans and their Families: A Study of the Florentine Catasto of 1427*. Yale University Press, New Haven, CT
Holmes G 1986 *Florence, Rome and the Origins of the Renaissance*. Clarendon Press, Oxford, UK
King M L 1991 *Women of the Renaissance*. Chicago University Press, Chicago
Kristeller P O 1961, 1965 *Renaissance Thought Vols I and II*. Harper Torchbook, New York
Kristeller P O 1972 *Renaissance Concepts of Man and Other Essays*. Harper Torchbook, New York
Lopez R S 1963 Hard times and investment in culture. In: Ferguson W K (ed.) *The Renaissance. A Symposium*. Seemann, New York, pp. 29–54
Maniates M R 1979 *Mannerism in Italian Music and Culture 1530–1630*. Durham, NC
Martines L 1968 *The Social World of the Florentine Humanists, 1390–1460*. Princeton University Press, Princeton, NJ
Panofsky E 1960 *Renaissance and Renaiscences in Western Art*. Amquist & Wiksell/Gebers Förlag AB, Stockholm
Panofsky E 1962 *Studies in Iconology*. Harper Torchbook, New York
Pocock J G A 1975 *The Machiavellian Moment*. Princeton University Press, Princeton, NJ
Porter R 1997 *Rewriting the Self, Histories from the Renaissance to the Present*. Routledge, London
Porter R, Teich M 1992 *The Renaissance in National Context*. Cambridge University Press, Cambridge, UK
Roover R de 1963 *The Rise and Decline of the Medici Bank, 1397–1494*. Harvard University Press, Cambridge, MA
Rubinstein N 1997 *The Government of Florence Under the Medici (1434–1494)*. Clarendon Press, Oxford, UK
Schmitt C B, Skinner Q (eds.) 1988 *The Cambridge History of Renaissance Philosophy*. Cambridge University Press, Cambridge, UK
Schulze W (ed.) 1996 *Ego-Dokumente. Annäherungen an den Menschen in der Geschichte*. Akademie-Verlag, Berlin
Schweikhart G (ed.) 1998 *Autobiographie und Selbstporträt in der Renaissance*. Walther König, Cologne, Germany
Shearman J 1967 *Mannerism*. Penguin, London
Squatriti P, Merideth B 1999 *The Renaissance in the Fields. Family Memoirs of a Fifteenth-century Tuscan Peasant*. Penn State University Press, PA
Thomas A 1995 *The Painter's Practice in Renaissance Tuscany*. Cambridge University Press, Cambridge, UK
Verdon T, Henderson J (eds.) 1997 *Christianity and the Renaissance. Image and Religious Imagination in the Quattrocento*. Syracuse University Press, Syracuse, NY
Wackernagel M 1938 *Der Lebensraum des Künstlers in der florentinischen Renaissance. Aufgaben and Auftraggeber, Werkstatt und Kunstmarkt*. Leipzig, Germany
Wind E 1968 *Pagan Mysteries in the Renaissance*. Faber & Faber, London

B. Roeck

Repartnering and Stepchildren

1. Introduction

Sexual unions create children and hence, parenthood. Union dissolution creates single parents and absent parents. When these parents enter a new partnership, they may have a new kind of children—stepchildren—children who are not the joint concern of the couple at the start of the union, as at least one member enters the union as a step-parent. This asymmetry in parenthood is likely to shape the partner dynamics that lead to union formation and hence to affect the calculations of those with children (whether coresident or not) and any partners who might join them in a new union. This asymmetry affects their lives as a couple, particularly whether they have additional, joint children, and also affects whether their union survives. This article on children and new partnerships examines three issues: (a) the role of children in the likelihood that their parents repartner (considering different effects for men and women); (b) the role of such children on the fertility of the unions formed, and (c) the effect of stepchildren on union dissolution.

2. Background

The rapid rise in divorce in most industrialized countries together with the rise in cohabiting unions, which tend to be more unstable than marriages, is creating a new demographic situation for men and women. A union disruption places them in the partner market again, which is particularly challenging for those who are the parents of young children. Although historically, high mortality often left young parents widowed, divorce creates twice as many new prospective partners. Further, whatever wisdom had accumulated for those facing this challenge, has likely been lost. The divorce revolution that began in the 1960s and 1970s in the industrialized countries in the west followed a half century or more of rapidly declining mortality, which sharply reduced widowhood for both sexes and concentrated it at ages when children are older, or even adults. In addition, the early years of the increase in divorce rates disproportionately involved couples without children far more than has been the case in the 1980s and 1990s (Andersson 1997). Hence, the effect of existing children on new partnerships is not well understood, either in the community of scholars of the demography of the family or in the communities experiencing the new situation. Little is known about how children affect their parents' repartnering, and not much more is known about those partnerships as they develop over time in terms of their fertility and stability. The research community has been further hampered by a lack of appropriate data to study these issues.

The numbers involved are substantial. In the United States, it is estimated that 40–50 percent of all marriages dissolve and a substantial minority of all children spend some time in a single-parent family (Bumpass and Raley 1995). Half of both previously married cohabiting couples and remarried couples have children in the household and many of the rest have children living elsewhere (Bernhardt and Goldscheider 1998). Although such rates are higher in the US than in other industrialized countries, they are rising rapidly outside the US (see *Demographic Transition, Second*), while US levels appear to have reached a plateau in the last two decades of the twentieth century. At whatever level, these trends mean that a substantial proportion of adults are at risk of entering a new partnership, and many of those who do will have children, either living with them or dependent upon them financially. Those considering forming a union have to take their potential partner's children into account.

A critical question might be whether the presence of children increases the likelihood that the unions formed are cohabiting ones. There is some evidence that this is, in fact, the case. Although the rise in cohabitation has not fully offset the decline in proportions married among the never married, the offset is total among the formerly married, suggesting that remarriage is being replaced by cohabitation (e.g., see *Demographic Transition, Second*). However, most research, and most of the data on which research could be based, has made it difficult to study cohabiting partnerships that involve children. Research on 'single-parent' families has frequently studied only 'unmarried' parents. The most common control for resources in families was 'family income,' which typically excluded the income of 'non-family' members, i.e., cohabiting partners. New surveys, such as the Family and Fertility Surveys (FFS), should overcome these difficulties.

3. *Repartnering of Parents*

Nearly all research on repartnering when pre-existing children are involved has considered only women with coresident children, focusing almost entirely on remarriage and on the US. These studies have typically found that the presence and sometimes the number of children reduce women's likelihood of marrying, with effects often depending on the ages of the children and of the mother. One of the few studies that considers cohabitation suggests that children reduce the partners' commitment to the union, decreasing the risk of getting married but increasing the risk of cohabiting (Bennett et al. 1995).

There is much less evidence on how children affect *men's* likelihood of union formation. A major problem for men is distinguishing between coresidential and non-co-residential children, a distinction that has been less important for women. One study examined widowers, where presumably fewer of the children lived with other relatives (Smith et al. 1991). That analysis showed that having dependent children reduces middle-aged (< 60) men's likelihood of remarriage after widowhood, although the reduction was much less than is the case for women with children. However, more recent research on both the US and Sweden, which has included divorced and never-married men with coresident children, as well as widowed men, has found that such men enter partnerships significantly *more* quickly than men without children (Goldscheider and Sassler 2000, Bernhardt and Goldscheider 1999). This suggests that the effects of coresidential children on union formation are quite different for men than for women.

There is even less research on how non-coresidential children affect union formation, although this is a particularly important issue for men (and the women who consider entering unions with them). The studies mentioned above also examined this issue, and found that in both the US and Sweden, men with non-coresident children are significantly *more* likely than otherwise comparable men with no coresident children to enter a union than to remain outside a partnership. In Sweden, such men are also more likely to enter a union with either a woman who has co-resident children or one who does not than to remain unpartnered; in the US this effect does not appear for union entry for women without children. The few women who have non-coresident children are also more likely to enter a union than to remain outside a partnership. In this case, however, at least in Sweden, they are much more likely to choose a partner with no coresident children, although this effect may reflect the small numbers of nonpartnered men who live with their children.

Perhaps a more important question, however, is the broader one: what factors affect entry into a union with a partner who has children? Research on the US and Sweden shows that two family-related factors distinguish men by whether the partnerships they enter into include children: their parents' marital history and their own marital history (Bernhardt and Goldscheider 1999, Goldscheider and Sassler 2000). Men who experienced a nontraditional family structure in childhood are significantly more likely to form a marriage that converts them into stepfathers than one that does not, and the same general difference appears for cohabitation.

US men who have been previously married resemble those who are already parents, in being very likely to enter a union (Goldscheider and Sassler 2000). If they marry, there appears to be no difference in the likelihood that their partner has children or not. If the new union is a cohabiting one, however, they are more likely to enter one with children. In fact, divorced men are more likely either to marry or cohabit with children

than to remain unpartnered; and they are significantly less likely to cohabit without children than to remain outside a partnership. Similar results were obtained for Sweden (Bernhardt and Goldscheider 1999).

Men who enter unions with children are not simply different in terms of their family histories, however; the evidence for both Sweden and the US suggests that they are also drawn from among the less educated, but the studies do not show why. In Sweden, those with a post-gymnasium education are the most likely to enter a union with a partner without children but the least likely to enter a union where the new partner has children (Bernhardt and Goldscheider 1999). This pattern is the same for US men; college graduates are more likely to enter a union than to remain outside a partnership if the woman does not have children. They are more likely, however, to remain outside a partnership than to enter a union with a women who has children (Goldscheider and Sassler 2000).

These results suggest that there are important differences in the likelihood of repartnering for those with and without children, with patterns that differ between men and women. Children, whether coresident or not, increase repartnering for men, but only non-coresident children do so for women; for them coresident children have the opposite effect. This suggests that surveys need to go beyond providing information in their partnership histories about whether partners brought children into the union and at least also provide information on whether partners have children living elsewhere. Such children clearly affect union formation; it is likely that they would also affect partnership dynamics (fertility and union stability), though this cannot yet be assessed fully, as we shall see in the next section. It is also likely that other characteristics of the children, such as their sex, might affect their parents' repartnering. No studies to date, however, have assessed whether the gender of the child(ren) matters. This question will take on increased importance later in this review.

More fundamentally, it becomes increasingly important to understand the reasons why children reduce the likelihood of repartnering, which we now see is only the case when they are coresident children and are children of the woman and not the man (unless he is older and widowed). How important is the fact that, although in most industrialized countries stepfathers have few legal rights or responsibilities vis-à-vis their stepchildren, most men realize that the children will be claimants on their incomes, at least while they remain in the household? The effect of having coresident children appears to be less for Swedish women than for women in the US (cf. Bernhardt and Goldscheider 1999 and Bennett et al. 1995). It may be that the greater support provided to children of all kinds by the Swedish state offsets some of the perceived liability for prospective stepfathers.

A perhaps equally important problem is ambiguity over the nature of stepfather–stepchildren relationships. Stepfathers are discouraged from assuming parental responsibilities for their stepchildren's lives by the children, their mothers, and by the larger society in most western, industrialized countries. Paternal involvement is an important factor in family dynamics, particularly for union stability, and we will examine this topic below.

A further interesting question is how much of the decline in marriage formation is a result of the increase in potential partners with children? Most studies that have examined whether the quality of the partner market has declined have considered women's decisions and have examined trends in the earnings of unmarried men who would be their potential partners. It might be equally interesting to study men's decisions and examine whether trends in men's marriage likelihood are influenced by the changing composition of potential female partners, as increasing proportions have children living with them.

4. *Fertility of Couples with Children from Previous Relationships*

There have been more studies on the effects of children from a previous union on childbearing in a new union than on repartnering. Few, however, distinguish between coresident and non-coresident children, and none have taken into account the selectivity of repartnering when children are already present as revealed in the studies reviewed above. Nevertheless, these studies have assembled an impressive set of results for a wide range of industrialized countries in Europe, as well as in the US.

Not surprisingly, the general finding is that couples in later unions have lower fertility within those unions than do couples in otherwise comparable first unions, which indicates clearly that many individuals have childbearing goals they do not want to exceed. Their previous children are considered in their decisions to continue childbearing. This result has appeared in studies in Germany, Czechoslovakia, the US, France, and Sweden (cf. references in Vikat et al. 1999).

How much previous children matter and whether it makes a difference whether they are coresident or not remain open questions. They are much more difficult questions to assess for the moment, as there is rarely information on previous children for both members of a couple, particularly on their non-coresident children. The most complete study is a recent analysis of Swedish data (Vikat et al. 1999), which, although it only had data on non-co-resident children for one member of the couple, produced a number of interesting findings.

First, the study found that the effect of previous children is definitely not total, as couples in Sweden appear to have at least one shared biological child independently of previous childbearing by either

partner. The authors interpret this result as indicating that joint children are an important element of a union. Second, the authors were able to rule out some alternative reasons for childbearing in later unions. There was little evidence that the risk of having a first child in a union was higher when one of the partners was childless, suggesting that Swedes do not have a child simply as an indicator of adult status. Further, there was little evidence the couples have a child to produce a sibling for a pre-union child, although couples did appear to avoid having only one shared child. This was interpreted as suggesting that it is full siblings that are considered important by Swedish couples in new unions, not half-siblings. Research on the US indicates that while stepsiblings remain fairly close to each other in adulthood, they are much less close than full siblings (White and Riedmann 1992), although it is unclear how much this difference reflects their different histories prior to coresidence and how much it reflects problems in their relationships while sharing a home. It was not possible to assess how much fertility in repartnerships is affected by co-resident as opposed to non-coresident stepchildren, as full information on the latter was not available.

Fertility in repartnerships is clearly a complex subject, and one that might depend not only on the numbers of children each of the partners has had previously and where they live, but also the genders of those children. Strong gender effects appear in research on stepfamily relationships, and there is evidence that gender of children affects union dissolution (Brooks-Gunn 1994). No studies to date, however, have assessed this issue, as no research has addressed the selectivity of repartnering.

5. *Stepchildren and Union Dissolution*

One of the most widely observed characteristics of second partnerships is that they are less stable than first partnerships. Although many have speculated that this pattern was the result of one or more elements reducing commitment (from religious prohibitions to personal orientations toward divorce), a landmark study found that most of the difference in marital satisfaction and in stability could be attributed to the presence of stepchildren. Previously childless persons who repartnered had very little greater risk of union dissolution than those in first partnerships (White and Booth 1985). Studies of union dissolution since then have generally included indicators of the presence of non-joint children as well as joint children, and found that whatever protective effect joint children might have does not characterize stepchildren, who increase the risk of separation. Little research has been done, however, on what types of stepchildren contribute more to this effect. Are the ages of the children a factor? Is their sex? Does the involvement of the stepparent and/or of the non-coresident parent matter? An even more important element might be the parenting practices of the new partners. Two studies of paternal involvement that was restricted to first partnerships showed that greater involvement of a father with his children increased the wife's marital satisfaction (Harris and Morgan 1991) and their children's well-being (Amato 1994); stepfathers are normally less involved with their stepchildren than are biological fathers (Marsiglio 1991).

Related to this issue is a startling finding—that couples with daughters are more likely to divorce than couples with sons (Morgan et al. 1988). The key to this result appears to be paternal involvement with children. Men tend to spend more time with their sons than their daughters, both in one-on-one conversation and in terms of participation in *family* activities, sharing more meals, leisure time, and home activities with them (Cooksey and Fondell 1996). In a sense, stepchildren may be like daughters for many men, children who they wanted less, who they do not relate to as closely, and who they can leave more easily. This may also account for the commonly observed phenomenon that sons often benefit more by their mother's remarriage than daughters (Brooks-Gunn 1994). Given the increased importance of repartnering in the lives of children and their parents in western, industrialized societies, understanding this phenomenon needs to be a high priority.

See also: Child Abuse; Children and the Law; Divorce and Children's Social Development; Divorce and Gender; Divorce, Sociology of; Family Business; Family Theory: Economics of Marriage and Divorce; Fatherhood; Marriage; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Parenting: Attitudes and Beliefs

Bibliography

Amato P R 1994 Father-child relations, mother-child relations, and offspring psychological well-being in early adulthood. *Journal Marriage and the Family* **56**: 1031–42

Andersson G 1997 The impact of children on divorce risks of Swedish women. *European Journal of Population* **13**: 109–45

Bennett N G, Bloom D E, Miller C K 1995 The influence of nonmarital childbearing on the formation of first marriages. *Demography* **32**: 47–62

Bernhardt E, Goldscheider F forthcoming Men, resources and family living: The determinants of union and parental status in the US and Sweden. *Journal of Marriage and the Family*

Bernhardt E, Goldscheider F 1999 Children and Union Formation in Sweden. Paper presented at the European Population Conference, August, The Hague, The Netherlands

Brooks-Gunn J 1994 Research and policy issues in the study of stepparenting. In: Booth A, Dunn J (eds.) *Stepfamilies: Who benefits? Who does not?* L. Erlbaum Associates, Hillsdale, NJ, pp. 167–89

Bumpass L L, Raley R K 1995 Redefining single-parent families: cohabitation and changing family reality. *Demography* **32**: 97–109
Cooksey E C, Fondell M M 1996 Spending time with his kids: Effects of family structure on fathers', and children's lives. *Journal of Marriage and the Family* **58**: 693–707
Goldscheider F, Sassler S 2000 Men, children, and union formation. Unpublished paper.
Harris K M, Morgan S P 1991 Fathers, sons, and daughters: Differential paternal involvement in parenting. *Journal of Marriage and the Family* **53**: 531–44
Marsiglio W 1991 Paternal engagement activities with minor children. *Journal of Marriage and the Family* **53**: 973–86
Morgan S P, Lye D, Condran G 1988 Sons, daughters and the risk of marital disruption. *American Journal of Sociology* **94**: 110–29
Smith K, Zick C D, Duncan G J 1991 Remarriage patterns among recent widows and widowers. *Demography* **28**: 361–74
Vikat A, Thomson E, Hoem J 1999 Stepfamily fertility in contemporary Sweden: The impact of childbearing before the current union. *Population Studies* **53**: 211–25
White L K, Riedmann A 1992 When the Brady Bunch grows up: Step/half sibling and full sibling relationships in adulthood. *Journal of Marriage and the Family* **54**: 197–208
White L K, Booth A 1985 The quality and stability of remarriages: The role of stepchildren. *American Sociological Review* **50**: 689–98

F. K. Goldscheider

Replication

'Replication' refers to the repetition of the same or very similar research study. The term 'replication study' refers to a study conducted specifically to verify the results of an earlier study. The concept of replication of research studies is central to the scientific study of psychology and the social and behavioral sciences more generally, as many authors have pointed out in Neuliep (1991).

1. *The Importance of Replication*

Scientists of all disciplines have long been aware of the importance of replication to their enterprise. Now, thanks to the efforts of science writers for daily newspapers and reporters for network news, even the general public has become aware of the importance of replication in all the sciences.

The undetected equipment failure, the rare and possibly random human errors of procedure, observation, recording, computation, or report, are known well enough to make scientists wary of the unreplicated experiment. When we add to the possibility of the random 'fluke' common to all sciences, the fact of individual organismic differences and the possibility of systematic experimenter effects, the importance of replication looms larger still to the behavioral scientist.

What shall we mean by 'replication'? Clearly, the same experiment can never be repeated by a different worker. Indeed, the same experiment can never be repeated even by the same experimenter. At the very least, the participants and the experimenters themselves are different over a series of replications. The participants are usually different individuals, and the experimenter changes over time, if not necessarily dramatically. But to avoid the conclusion that there can be no replication in the social and behavioral sciences, we can speak of relative replications. We can rank order experiments on how close they are to each other in terms of participants, experimenters, tasks, and situations. We can usually agree that this experiment, more than that experiment, is like a given paradigm experiment.

Replications may be crucial, but some replications are more crucial than others. Three of the variables affecting the value, or utility, of any particular replication are:

(a) when the replication is conducted
(b) how the replication is conducted
(c) by whom the replication is conducted.

1.1 *When the Replication is Conducted*

Replications conducted early in the history of a particular question are usually more useful than replications conducted later in the history of a particular research question. Weighting all replications equally, the first replication doubles our information about the research issue, the fifth replication adds 20 percent to our information level, and the fiftieth replication adds only 2 percent to our information level. Once the number of replications grows to be substantial, subsequent investigators' felt need for further replication is likely to be due not to a real need for further replication but for a real need for the more adequate evaluation and summary of the replications already available.

1.2 *How the Replication is Conducted*

It has already been noted that replications are possible only in a relative sense. However, there is a distribution of possible replications in which the variance is generated by the degree of similarity to the original study that characterizes each possible replication. If we choose our replications to be as similar as possible to the study being replicated, we may be more true to the original idea of replication but we also pay a price; that price is external validity (see *External Validity*).

If we conduct a series of replications as exactly like

Precision of replication	Result of replication: Successful	Result of replication: Unsuccessful
Fairly precise	Supports the theory	Damages the theory
Fairly imprecise	Extends the theory	Limits the theory

Figure 1
Effects on theory of the success of replication and the precision of replication

Precision of replication	Result of replication: Successful	Result of replication: Unsuccessful
Fairly precise	Supports the investigator	Impugns the investigator
Fairly imprecise	Supports the investigator	Impugns the investigator very little

Figure 2
Effects on original investigator of the success of replication and the precision of replication

the original as we can, we have succeeded in 'replicating' but not in extending the generality of the underlying relationship investigated in the original study. The more imprecise the replications, the greater the benefit to the external validity of the tested relationship if the results support the relationship. If the results do not support the original finding, however, we cannot tell whether that lack of support stems from the instability of the original result or from the imprecision of the replications. Figure 1 summarizes the consequences for the theory that is derived from the initial study of (a) successful vs. unsuccessful replications, and (b) the precise vs. the imprecise nature of the replications.

Figure 2 summarizes the consequences for the original investigator of (a) successful vs. unsuccessful replications, and (b) the precise vs. imprecise nature of the replications. Compared to the theory tested by the replication, the investigator has much to lose if a fairly precise replication is unsuccessful since such failure is often associated with ascriptions to the original investigator of having been careless, incompetent, and in some cases, even dishonest.

1.3 By Whom the Replication is Conducted

So far in our discussion of replications we have assumed that the replications are independent of one another. But what does independence mean? The usual minimum requirement for independence is that the participants in the replication studies be different persons. But what about the independence of the replicators? Are 10 replications conducted by a single investigator as independent of one another as 10 replications each of which is conducted by a different investigator? This issue of potentially correlated replicators bears additional comment.

1.4 The Problem of Correlated Replicators

To begin with, an investigator who has devoted her life to the study of vision, or of psychological factors in somatic disorders, is less likely to carry out a study of verbal learning than the investigator whose interests have always been in the area of verbal learning. To the extent that (a) experimenters with different research interests are different kinds of people, and to the extent

	First study	
	$p>0.05$*	$p<0.05$
$p<0.05$[a]	A Failure to replicate	B Successful replication
Second study $p>0.05$	C Failure to establish effect	D Failure to replicate

Figure 3
Common model of successful replication: judgment is dichotomous and based on significance testing (*by convention 0.05 but could be any other given level, e.g., 0.01; [a] in the same tail as the results of the first study)

that (b) it has been shown that different kinds of people, experimenters, are likely to obtain different data from their subjects, we are forced to the conclusion that within any area of behavioral research, the experimenters come precorrelated by virtue of their common interests and any associated characteristics. Immediately, then, there is a limit placed on the degree of independence we may expect from workers or replications in a common vineyard. But for different areas of research interest the degree of correlation or similarity among its workers may be quite different. Certainly we all know of workers in a common area who obtain data quite opposite from that obtained by colleagues. The actual degree of correlation, then, may not be very high. It may, in fact, even be negative, as with investigators holding an area of interest in common, but holding opposite expectancies about the results of any given experiment.

A common situation in which research is conducted nowadays is within the context of a team of researchers. Sometimes these teams consist entirely of colleagues; often they are composed of one or more faculty members and postdoctoral students, and one or more predoctoral students at various stages of progress toward the Ph.D. Experimenters within a single research group may reasonably be assumed to be even more highly intercorrelated than any group of workers in the same area of interest who are not within the same research group. Perhaps students in a research group are more likely than a faculty member in the research group to be more correlated with their major professor. There are two reasons for this likelihood. The first is a selection factor. Students may elect to work in a given area with a given investigator because of their perceived and/or actual similarity of interest and associated characteristics. Colleagues are less likely to select a university, area of interest, and specific project because of a faculty member at that university. The second reason why students may be more correlated with their professor than another professor is a training factor. Students may have had a large proportion of their research experience under the direction of a single professor. Another professor, though collaborating with colleagues, has most often been trained in research elsewhere by another person. Although there may be exceptions, even frequent ones, it seems reasonable, on the whole, to assume that student researchers are more correlated with their advisor than another advisor might be.

The correlation of replicators that we have been discussing refers directly to a correlation of *attributes* and indirectly to a correlation of *data* these investigators will obtain from their subjects. The issue of correlated experimenters or observers is by no means a new one. Nearly 100 years ago Karl Pearson spoke of 'the high correlation of judgments...[suggesting] an influence of the immediate atmosphere, which may work upon two observers for a time in the same manner' in his 'On the mathematical theory of errors of judgment with special reference to the personal equation' (Pearson 1902, p. 261).

2. *The Meaning of Successful Replication*

There is a long tradition in psychology of our urging one another to replicate each other's research. But, although we have been very good at calling for replications, we have not been too good at deciding when a replication has been successful. The issue we now address is: when shall a study be deemed successfully replicated?

Successful replication is ordinarily taken to mean that a null hypothesis that has been rejected at time 1

is rejected again, and with the same direction of outcome, on the basis of a new study at time 2. The basic model of this usage can be seen in Fig. 3. The results of the first study are described dichotomously as $p < 0.05$ or $p > 0.05$ (or some other critical level, e.g., 0.01). Each of these two possible outcomes is further dichotomized as to the results of the second study as $p < 0.05$ or $p > 0.05$. Thus cells A and D of Fig. 3 are examples of failure to replicate because one study was significant and the other was not. The problem with this older view of replication is that two studies could show identical magnitudes of the effect investigated yet differ in significance level because of differences in sample sizes.

2.1 Contrasting Views of Replication Success

The traditional, not very useful view of replication modeled in Fig. 3 has two primary characteristics:

(a) It focuses on significance level as the relevant summary statistic of a study.

(b) It makes its evaluation of whether replication has been successful in dichotomous fashion. For example, replications are successful if both or neither $p < 0.05$ (or 0.01, etc.), and they are unsuccessful if one $p < 0.05$ (or 0.01, etc.) and the other $p > 0.05$ (or 0.01, etc.). Psychologists' reliance on a dichotomous decision procedure accompanied by an untenable discontinuity of credibility in results varying in p levels has been well documented, for example, by Nan Nelson, Robert Rosenthal and Ralph L. Rosnow (1986).

The newer, more useful views of replication success have two primary characteristics:

(a) A focus on effect size as the more important summary statistic of a study with only a relatively minor interest in the the statistical significance level.

(b) An evaluation of whether replication has been successful made in a continuous fashion. For example, two studies are not said to be successful or unsuccessful replicates of each other, but rather the degree of failure to replicate is specified.

See also: Experimental Design: Overview; External Validity; Generalization: Conceptions in the Social Sciences; Hypothesis Testing: Methodology and Limitations; Internal Validity

Bibliography

Campbell D T 1988 Overman E S (ed.) *Methodology and Epistemology for Social Science: Selected Papers*. University of Chicago Press, Chicago

Cooper H M 1989 *Integrating Research: A Guide For Literature Reviews*, 2nd edn. Sage, Newbury Park, CA

Hunt M 1997 *How Science Takes Stock*. Russell Sage Foundation, New York

Light R J, Pillemer D B 1984 *Summing Up: The Science of Reviewing Research*. Harvard University Press, Cambridge, MA

Nelson N, Rosenthal R, Rosnow R L 1986 Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist* **41**: 1299–301

Neuliep J W (ed.) 1991 *Replication Research in the Social Sciences*. Sage, Newbury Park, CA

Pearson K 1902 On the mathematical theory of errors of judgement with special reference to the personal equation. *Philosophical Transactions of the Royal Society of London* **198**: 235–99

Rosenthal R 1976 *Experimenter Effects in Behavioral Research*, enlarged edn. Irvington Publishers, Halsted Press Division of Wiley, New York

Rosenthal R 1991 *Meta-Analytic Procedures For Social Research*, rev. edn. Sage, Newbury Park, CA

Rosenthal R, Rosnow R L 1969 *Artifact in Behavioral Research*. Academic Press, New York

Rosnow R L, Rosenthal R 1997 *People Studying People: Artifacts and Ethics in Behavioral Research*. Freeman, New York

R. Rosenthal

Representation: History of the Problem

Philosophers and social scientists have always been concerned about the relation of their knowledge about the social world to (the rest of) this world. This concern has sometimes been phrased as the question of the representation of the social world in and through the writing of scholars, or as the *adaequatio intellectus et rei*, being the objective of scientific inquiry. From Plato and Aristotle, and then Thomas of Aquinas, this *problématique* stretches into the positivism of the social sciences of the nineteenth century as well as its critical alternative, scientific Marxism. With important qualifications, it was revived in the neopositivism of the twentieth century and the unified science program that tried to (re-) integrate the social sciences with the natural sciences.

Despite this continuity of concern, the very concept of representation of the social world has also been heavily criticized. At several points in intellectual history it appeared as if the whole idea needed to be abandoned, because there is no way to verify the relation between the world and thoughts about the world other than again by thought. Thus, knowledge of *adaequatio* is in principle unattainable. In most recent debate, during the closing decades of the twentieth century, the accumulation of such doubt has often been labelled 'crisis of representation,' implicitly and misleadingly pretending novelty of these doubts. Through all of earlier such crises, however, the *problématique* keeps reviving, often with new qualifications, and those who revive it often argue that it cannot be entirely abandoned either, since otherwise

the status of statements about the (social) world would become entirely unclear.

1. Representation in Philosophical Modernity

Even though this debate has been going on virtually through all of the history of philosophy, the break that came with the radical skepticism characteristic of what can be called the philosophy of modernity is often seen as shifting the terrain. Between the middle of the seventeenth century, Descartes's time of writing, and the end of the eighteenth century, Kant's time, the ground on which representation could be possible was systematically explored. Descartes's *cogito ergo sum* radically subjectivized knowledge claims. Nevertheless, he was able to develop on that basis a rationalism according to which mathematical representation was adequate to the world. Kant's distinction between *noumena* and phenomena has sometimes been interpreted in a similar way, that is, by considering *noumena*, although they cannot be experienced, as theoretical entities the relations between which are governed by laws. A science developing such laws would then be truthfully representing the world. Arguably, however, the more significant impulse from these one and a half centuries of debate was a deep crisis of representation, at least within the philosophical mode of reasoning. This crisis of representation stemmed from the increasing gap between the perceiving and knowing subject and the world.

The combined outcome of the Cartesian and the Kantian moves, namely, was that, on the one hand, philosophical reasoning appeared to have exhausted its possibilities, at least for the time being. At the endpoint of this reasoning, on the other hand, a space had been opened for a different form of inquiry. The gap between the subject and the world, or between the *noumena* and the phenomena, permitted, maybe even invited, attempts to fill it by positive, empirical knowledge, elaborated in systematic, conceptually guided ways (see *Positivism, History of*). It allowed the subjectivation and historicization of knowledge claims, which in turn made the social sciences possible, as psychological inquiry into human perception and experience of the world and as sociohistorical collection and systematization of the 'facts' of human social life. Much of the social sciences of the nineteenth century, often devoid of any explicated philosophy of the social sciences, indeed worked on such foundations.

2. Representation in Political Modernity

While philosophical debate opened the possibility for developing a conception of representation of the social world through social science inquiry, roughly parallel political transformations made such a representation appear as an increasingly urgent necessity. These transformations made the concept of representation increasingly problematic in political thought and practice, a culmination of this process and a turning-point being reached with the American and French revolutions.

A concept of representation had been almost completely absent from political thinking in Europe and the Mediterranean region during antiquity and the Middle Ages. First elements of such a *problématique* gradually emerged with the rivalry between various sources of religious power during the era of the schism, and between religious and political power with the contest between the pope and the emperor. Representation here, most broadly, refers to the relation between a holder of authority to that over which this authority is exercised. Employing the concept of representation for this question presupposes that authority is not held in its own right, but for something or somebody else. The disputes during the late Middle Ages and early modern times were largely cast in terms of the legitimate representation of God, as the uncontested supreme authority, on earth. William of Ockham may have been the first to conclude on the impossibility of deciding in the face of competing claims for representation and, thus, to have opened up a political version of the crisis of representation.

As in the philosophical debate about the knowledge of the world, the period between the middle of the seventeenth and the end of the eighteenth century is of crucial importance for a recasting of the issue of political representation. The first systematic attempt to solve the issue in modern form, that is, by fully acknowledging both the exigency of representation and the undecidability of the question of the sources of authority, was made by Thomas Hobbes. Faced with the religious wars and thus a violent contest over the rightful representation of God, he ruled out any religious source, but insisted on the need for unity in the polity to avoid persistent warfare. The idea of state sovereignty, prepared by Bodin, and of the unitary state represented by the monarch were the key ingredients of his solution to the problem. By making the constitution of the state itself, however, dependent on a contract between individuals, he had opened the breach for individualism, as Carl Schmitt observed. The idea of popular sovereignty, as developed by Locke and Rousseau, should indeed turn Hobbes's reasoning around and prepare the intellectual ground for the political transformations of the late eighteenth century.

Quite analogous to the situation in general philosophy, the development of political philosophy was, on the one hand, stalled with the full explication of an individualist liberal political theory, that is a political theory of modernity (see *Liberalism: Impact on Social Science*). On the other hand, however, insufficiencies of the liberal assumptions, particularly visible in the post-revolutionary situation, lent themselves to the development of the social sciences as means to deal with those insufficiencies by other means, through a

kind of 'empirical political philosophy' namely (Wagner 1998).

Liberal political philosophy was insufficient in the sense that the set of rules on which viable political institutions could be built, starting out from individuals and their potential diversity, needed propositions beyond those that could be derived from the basic assumptions of individualist liberalism. The concept of self-determination needed, for instance, procedural rules of participation in communication and in deliberation. Since such rules could not unequivocally be derived from the principles themselves, the issue of representation emerged much more forcefully than ever before. In response to the French Revolution, Burke (1993) still argued that, while there was a practical issue of representation, best handled in line with grown traditions, there was no specific legitimacy attached to representation. Paine (1993), in contrast, linked most forcefully the idea of the sovereignty of the people to the representation of the people in the governing bodies. The controversy between Burke and Paine 'marks the turning point of two ages of representation' (Sternberger 1971, p. 594). As soon as any position closer to Paine than to Burke was accepted, debates on modes of representation, such as those that had already occurred in both the American and the French constituent assemblies, became inevitable.

And it was in conjunction with this inevitability of disputes over representation that the creation of social effects of the liberal rules moved to the center of attention. The identification and the analysis of these effects called for social science inquiry and conceptualization. The observation of actual, empirical structures of representation also served the interest of enhancing some stability and certainty in political procedures that could appear to be opened to all contingencies by the abolition of any legitimacy of preordained orders (see *Freedom/Liberty: Impact on the Social Sciences*). Two main strategies for rediscovering certainties can be identified: systematic observation and reflective conceptualization. The one was more typical for American thinking, where the structure of the vote was of key interest, whereas European social thinkers focused on what came to be called social structures behind the vote. These two intellectual responses to the political *problématique* of representation inaugurate two modes of social theorizing, the behavioralist and the structural-functional one.

In sum, thus, during the nineteenth century an increasing overlap of concerns between cognitive representation, that is, the truthful representation of social reality, and political representation, that is, the need for a social map to draw up rules for electoral representation can be observed. One may speak here of the gradual transformation of a philosophical conception of representation, based on the political philosophy of the Enlightenment, into a sociological one, based on the sociology of industrial and mass society emerging at the end of the nineteenth century (d'Arcy and Saez 1985). This rephrasing of the problem of representation was brought again into crisis, however, by the turn of the nineteenth century and, yet again, in the late twentieth century. These crises, although they were also crises of philosophical representation, showed specific features of a crisis of the representation of the social world—and this was again related to political issues of representation.

3. *A Crisis in Modern Representation and its Temporary Solution*

From the late nineteenth century on, a shift in political orientations, which entailed a declining faith in the viability of liberalism, was paralleled in epistemology by a renewed skepticism about the other central tenet of the Enlightenment tradition, the intelligibility of human action and the social world. The period around the turn of the century is now known as an intellectually extremely fruitful, even a classical era in many fields of social science, most notably sociology, psychology and economics. At that time, however, much of the work was driven by a sense of crisis, a feeling that many of the assumptions of earlier social science were inadequate.

In terms of epistemology, social science saw itself forced largely to abandon the idea of representing social reality and accepted the view that conceptual constructions were dependent both on the means and forms of observation and perception and on the interest of the observer in the social world. American pragmatism is possibly the most explicit case of such a reorientation, but similar, often much more tension-ridden, discussions were led in European debates, a prominent example being Max Weber's methodological writings. As a consequence, the turn-of-the century approaches were less convinced about a determinist course of human history than earlier social science, and also less persuaded of the direct insight into any laws of the social world from empirical observation. Such uncertainty also entailed a crisis in the usability of social knowledge for purposes of action. The earlier call had been for better, 'positive' knowledge that would lead to better action; it relied on the conception of an objective social world outside the mind of the observers, but adequately represented by them in social theory and research. According to such a conception, action based on uncertain knowledge should entail uncertain outcomes.

Through the early decades of the twentieth century, however, novel conceptions of the relation between knowledge and action were proposed that were based not least on a critique of what Dewey (1984, p. 195) called the 'spectator theory of knowing.' With this term, he summarized all views, for him characteristic of most philosophy and science, that saw the world as exposed to the human gaze and assumed that it is

exactly this distance between the knowing subject and the object to be known that allows for certain knowledge. The distancing from the sensations of the world had been considered a precondition for the identification of those assumptions on which claims for valid knowledge could be erected. By the same move, the question of epistemology itself had emerged as the problem of the relation between the 'reality' of the world and the 'representations' that human beings provide of the world in their philosophy and science. The preference in the philosophy of science was for a long time, as Dewey points out, for an epistemological formalism that presupposes this distance between the knower and that which is to be known. Pragmatism precisely rejects this assumption of distance.

Similar turns were made in European philosophy during the same period. Ludwig Wittgenstein had hoped to clarify the problem of representation in the *Tractatus logicus-philosophicus*, in which he employs a picture theory of knowing, whereas in the *Philosophical Investigations* he sees our languages as being part of our forms of life, and thus not looking at life from a distance. In *Holzwege*, Martin Heidegger analyzed the modern theory of knowledge as representation comprehensively as the attempt 'to provide oneself a picture' of the world by making a world that is placed in front of oneself. Rather than making 'a picture of the world,' this move 'grasps the world as picture' (see *Imagination: History of the Concept*). In two intertwined processes, the world becomes a picture and the human being a subject. Rorty tried to summarize this reinterpretation of the history of philosophy and the sciences provided by Dewey, Heidegger, and Wittgenstein as the abdication of the attempt of seeing knowledge as 'the mirror of nature' (Rorty 1979, Friese and Wagner 1997, pp. 208–11).

The critique of representation, in this sense, was linked to the fundamental and critical epistemological presupposition that the social world is in important respects not found and discovered but made and invented. This novel combination, however, provided merely for one strand of debate in the social sciences of the first half of the twentieth century, and it did not become dominant. Partly disregarding epistemological issues, partly referring to the 'neopositivism' of the Vienna Circle, the other major strand tried to deal with the problem of representation by methodological means. At this point, the history of the problem of representation links up closely with the history of statistics and of empirical social research.

In statistics, the period between the end of the nineteenth century and the 1930s witnessed intense debates about a seemingly very specific methodological question. The problem was how to generalize observations collected from the study of a part to statements about the whole. Its solution was a number of statistical techniques summarized under the label of representativity. These techniques were developed simultaneously in debates within the statistical profession and about the design of social insurance arrangements. An institutional, a cognitive, and a political transformation were closely interlinked, 'the nationalization of social statistics, … the diffusion of the 'representative method' and of random selection, and … the beginnings of the development of the Welfare State' (Desrosières 1991, p. 228) (see *Statistics, History of*).

During the interwar period, mathematical techniques spread to other areas of empirical research, not least to the emergent fields of opinion research, of business cycle research and of research into the social structure of society. Special research institutes were created, during those years and particularly after the Second World War, that were to provide the data to which those techniques were to be applied. From the late 1920s onwards, thus, the contours of a new social science, which then should dominate the image of those sciences in the second post war era, are recognizable. This social science liberated itself from the doubts of the 'classical period' by methodologically circumventing the problem of representing society. The doubts about epistemological and conceptual issues could not be entirely removed, but they could be dealt with, it was assumed, by starting from the most secure elements one could find, that is the empirical observation and collection of data on the preferences and behaviors of individual human beings. Conclusions referring to the larger scale of society and politics were arrived at by aggregation of those data; and the questions towards the answering of which data were generated or analyzed were derived from policy needs for 'social control.' Thus, a new connection between cognitive and political representation was created. Or, in other words, a 'soft' behaviorism aligned with a similarly 'soft' pragmatism.

4. *The Re-emergence of a Crisis of Representation*

However, such solution to the problem of representation remained temporary. During the closing decades of the twentieth century, the 'interpretative turn' or, more broadly, the 'linguistic turn' in the human sciences has had strong, though uneven, effects in the sciences devoted to the study of contemporary society. Emphasis on the linguistic constitution of the social world and on the interpretative openness of social representations brings the social sciences back to a period of epistemological, ontological, and methodological reconsideration that shows many parallels to the 'classical era' at the beginning of the twentieth century.

It has been argued that this recent period of intellectual transformation occurs in parallel to an overall process of sociopolitical transformation. Crises of cognitive and of political representation then again occur in tandem (Wagner 1994). During such periods,

it becomes evident that a social science that tightly couples the concern with social order to the solidity of its own representations of the social world applies at best in periods of relative social stability, that is, only contingently. Conceptually, however, the ideas of unanimity or similarity of behavior being established by common material constraints or well mastered by shared ideas (beliefs, representations) or a diversity of actions being integrated through systematic articulation (division of functions, of roles) works with too restrictive definitions of social life (Thévenot 1993 p. 276). Attention needs instead to be directed towards 'the immense historical labor that is necessary to unite disparate beings around the same system of representation, to constitute the reality of such a heterogeneous ensemble, to inscribe it into devices (*dispositifs*) through an intensive effort of objectification and to endow it with a common interest' (Boltanski 1990 p. 70).

It is important to note that many of the authors who recently have renewed the interest in the *problématique* of representation, such as Michel Foucault and Reinhart Koselleck, have done so by emphasising the linkage between cognitive and political representation. In their own research, they have also underlined that the critical relation between these two forms of representation becomes accentuated during the political and intellectual transformations of the late eighteenth and early nineteenth century. This is at the same time the period of the emergence of the social sciences. They were created to deal with this *problématique*, but they remain nevertheless, or for that very reason, marked by it.

See also: Knowledge Representation; Marxism/Leninism; Modernity; Political Representation; Positivism, History of; Propositional Representations in Psychology; Representations and Rituals of the State; Wittgenstein, Ludwig (1889–1951)

Bibliography

d'Arcy F, Saez G 1985 De la représentation. In: d'Arcy F (ed.) *La représentation*. Economica, Paris

Boltanski L 1990 *L'amour et la justice comme compétences. Trois essais de sociologie de l'action*. Métailié, Paris

Burke E 1993/1790 *Reflections on the Revolution in France*. Oxford University Press, Oxford, UK

Desrosières A 1991 The part in relation to the whole? How to generalise. A prehistory of representative sampling. In: Bulmer M, Bales K, Sklar K K (eds.) *The Social Survey in Historical Perspective 1880–1940*. Cambridge University Press, Cambridge, UK, pp. 217–44

Dewey J 1984/1929 *The Quest for Certainty: The Later Works 1925–1953*, Vol. 4. Southern Illinois University Press, Carbondale, IL

Friese H, Wagner P 1997 Bild und Wissen. Reflexionen zu Erkenntnis und Darstellung in den Sozialwissenschaften. In: Hoffmann U, Joerges B, Severin I (eds.) *LogIcons. Bilder zwischen Theorie und Anschauung*. Edition sigma, Berlin, pp. 207–22

Hacking I 1983 *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, New York

Lynch M, Woolgar S (eds.) 1990 *Representation in Scientific Practice*. MIT Press, Cambridge, MA

Paine T 1993/1791–2 *The Rights of Man*. Dent, London

Rorty R 1979 *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ

Sternberger D 1971 A controversy of the late eighteenth century concerning representation. *Social Research* **38**

Thévenot L 1993 Agir avec d'autres. Conventions et objets dans l'action coordonnée. In: Ladrière P, Pharo P, Quéré L (eds.) *La théorie de l'action. Le sujet pratique en débat*. CNRS Editions, Paris, pp. 275–89

Wagner P 1994 *A Sociology of Modernity. Liberty and Discipline*. Routledge, London

Wagner P 1998 Certainty and order, liberty and contingency; the birth of social science as empirical political philosophy. In: Heilbron J, Magnusson L, Wittrock B (eds.) *The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850*. Kluwer, Dordrecht, Germany, pp. 241–63

P. Wagner

Representation of Serial Order, Cognitive Psychology of

Stimuli come and go. The serial order of their appearance is determined by rules in the external world as well as by the actions of the organism experiencing the stimuli. Consequently, successions of stimuli are almost never random but rather comprise some form of structure. Structure allows prediction, and predicting forthcoming events is advantageous because organisms can prepare for what will happen. Thus, sensitivity to serial order can be considered as being one of the basic mechanisms by which humans (and organisms in general) adapt behavior to constraints in the succession of events. That is why the examination of serial learning has a prolonged research history (e.g., Ebbinghaus 1902, Lashley 1951, Miller 1956).

Serial order refers to deviations from randomness in the succession of events. Consider, for instance, the three colored lights green (g), red (r), and yellow (y) as distinct events. The series g–g–r–g–y–g–r–y–y–g–r–…is random, as each of the lights was selected with equal probability in each position. In contrast, the series g–y–r–y–g–y–r–y–g…presents the highly regular sequence we experience in watching ordinary traffic lights. Knowing the serial order of traffic lights facilitates responding: if we see that the light switches from green to yellow we can expect that it will next switch to red so that we can prepare to stop before we are really forced to do so. Almost all event sequences entail some regularity: clouds are frequently followed

by rain, words follow each other according to grammatical rules, tone sequences constitute melodies, and baking a cake requires an ordered sequence of actions, etc. In this article, first two types of serial order are distinguished, then several effects of serial order on human cognition are sketched, and finally, the main theoretical approaches to account for serial order effects are mentioned.

1. *Statistical and Relational Constraints in Serial Order*

Statistical constraints refer to the conditional probabilities of events in dependence on preceding ones. Traffic lights are a good example: after 'red' and after 'green,' 'yellow' is certain to follow, whereas after 'yellow' either 'red' or 'green' follows. If one takes two preceding lights into account, however, the light following 'green–yellow' and 'red–yellow' is also safely predictable. An example of more complex statistical constraints is provided by sentences: the next word in a sentence can mostly be predicted with a probability higher than chance but almost never with certainty, regardless of how many preceding words are taken into account. *Relational constraints* refer to regularities in the relations between successive events instead of conditional probabilities. Consider, for instance, an arithmetic series such as 3–8–13–18–23–28. The regularity of the series is based on that between every pair of adjacent numbers the same relation holds. An example of a somewhat less regular relational structure is a digit sequence such as 23461938, which contains only two equally related pairs of numbers, namely 23–46 and 19–38.

2. *Serial Order Effects in Human Cognition*

2.1 *Serial Order Effects in Prediction*

In order to examine the adaptation of predictions to serial order, experiments have been performed in which items are successively presented and participants are asked to predict each forthcoming item. Typically, prediction accuracy improves corresponding to the statistical and/or relational constraints in the succession of items. Moreover, the data indicate that constraints in immediate successions are effective before constraints in the succession of items further apart from each other are taken into account. Consider, for instance, a well ordered sequence such as 1–2–1–2–2–3–2–3–6–5–6–5–5–4–5–4. Experimental data show that participants first learn to predict the respective second items of those pairs whose elements are likewise related, (12), (23), (65), and (54). They continue to learn the repetitions of pairs (1212), (2323), (6565), and (5454). Finally, participants manage to predict the order of the quadruples (Restle and Brown 1970). In other words, serial learning starts by connecting adjacent items into subunits or chunks corresponding to the given relational (as in the present case) or statistical constraints. Following this, constraints in the succession of these elementary chunks results in the formation of higher order chunks, i.e., connections between elementary chunks, which again are subject to chunking, etc. Thus, consecutive chunking may lead to a representation of the whole sequence through a hierarchy of chunks.

2.2 *Serial Order Effects in Remembering*

Ordered sequences of items are generally remembered better than unordered sequences (see *Memory: Organization and Recall*). For instance, sequences of words are remembered better if they follow the statistical constraints of word order in sentences (Miller and Selfridge 1950), a list of words is remembered better if words belonging to the same category (e.g., names of fishes, flowers, birds, or trees) are presented adjacently (Underwood et al. 1974), and a series of letters will be remembered better if it consists of familiar abbreviations as in FBICIAPHDVIP. In all these cases subsequences of the items to be remembered correspond to familiar knowledge units like word phrases, categories, or abbreviations which are used to reduce memory load. In the latter case, for instance, participants have to keep in mind the four abbreviations FBI, CIA, PHD, and VIP instead of 12 letters. Moreover, the order of the storage units or chunks may provide a useful retrieval plan which additionally helps remembering. In the categorically ordered word list mentioned above, for instance, memory will be further improved if also the categories are presented in a systematic order such as the animals (fishes and birds) first and then the plants (flowers and trees).

However, new and unfamiliar serial constraints also improve memory. Consider, for instance a so-called 'finite state grammar,' which restricts the way in which the letters T, P, V, X, and S are allowed to follow each other in letter strings. For instance, all grammatical strings may start either with the letter T or P, they may finish either with V or S, only the letters X and V may be allowed to be repeated, after letter T only the letters S, X, and V may be allowed to follow, etc. In this way the 'grammar' constitutes a statistical structure which all grammatical strings such as TXS, TSXXTVV, PTVPS, or PVPXVPS have in common. If such letter strings are to be memorized, grammatical strings are remembered increasingly better than random strings and after some training participants are even able to distinguish between grammatical and agrammatical strings fairly well (Reber 1967). The findings clearly indicate that the completely new and arbitrary stat-

istical constraints have been learned and are used to put order into the letter strings to be remembered.

2.3 Serial Order Effects in Motor Performance

In motor performance, similar effects as in predicting or remembering stimulus sequences have been reported: well ordered sequences of actions are executed faster and more correctly than unordered sequences and, even more informative, the timing of sequence execution mirrors its serial order (Rosenbaum et al. 1983). Consider, for instance, the times for pressing a series of eight distinct keys with the index fingers (I, i) and the middle fingers (M, m) of the right and left hand, respectively. The order of the keys to be pressed shall be such that a first pair of keystrokes (e.g., I–m) is to be repeated (I–m–I–m) and then a second pair is to be repeated (e.g., M–i–M–i). In performing such a regular keystroke sequence as fast as possible, execution times are highest at the beginning of the quadruples ImIm and MiMi, they are somewhat reduced at the beginning of each repeated pair ImIm and MiMi, and they are lowest for the respective second element in each pair ImIm and MiMi. The correspondence of this time pattern with the serial order of the keystrokes suggests that the time to execute a motor act in a sequence of acts is determined by hierarchically organized chunks rather in the same way as the prediction of items is: at the beginning of each new chunk execution time increases as prediction accuracy decreases and for items within chunks execution is fastest and prediction is best (see *Motor Skills, Psychology of*).

2.4 Serial Order Effects in Serial Reaction Time (SRT) Tasks

In SRT tasks, participants are required to respond to successively presented stimuli by distinct responses whereby each response triggers the presentation of the next stimulus, which in turn requires the next response, etc. Numerous studies have shown that nearly any kind of serial order in the succession of stimulus–response pairs results in a decrease in response times (Hoffmann and Koch 1998). Serial-order effects in SRT tasks deserve special attention because besides constraints in the succession of stimuli and of responses, also constraints in the succession of the responses and the following stimuli, i.e., in response–effect relations, may contribute to serial learning. Although not much research has been done yet in this respect, the available evidence clearly indicates that consistent response–effect relations do indeed contribute to the acquisition of serial order (Ziessler 1998). This is most noteworthy because consistent response–effect relations are an essential part of any ordinary behavioral sequence.

3. Intentional and Incidental Acquisition of Serial Order

Some of the mentioned cognitive effects surely result from the explicit detection and use of serial order. For instance, if one tries to predict the continuation of a stimulus sequence on the basis of regularities in the preceding series, or if one tries to reduce memory load by parsing a string of letters into abbreviations, the given serial order is explicitly used to improve performance. However, there are other cases in which it does not seem to be necessary that serial order is detected in order to become effective (Stadler and Frensch 1998). For instance, memory for letter strings generated by a 'finite state grammar' improves without the grammatical constraints being reportable. Likewise, in SRT tasks participants respond increasingly faster to statistically ordered stimulus sequences without being aware of the statistical constraints, and consistent response–effect relations have also been shown to decrease response times without being detected. Obviously, serial order can become effective in two ways, once via a strategic use of detected constraints and second via more elementary mechanisms by which serial constraints seemingly inevitably affect performance, irrespective of whether they are detected or not.

4. Theoretical Accounts

The mechanisms underlying the acquisition and use of serial order are not yet well understood. However, it is generally acknowledged that serial learning is accomplished by processes that lead to the formation of *chunks*, i.e., to an integration of successive events into coherent subsequences (Miller 1956). The theoretical accounts differ with respect to the processes which are assumed to mediate the formation of chunks as a result of experienced serial constraints. On the one hand, *explicit strategies* for searching for serial order have been analyzed (Simon and Kotovski 1963). From this perspective, the formation of chunks results from the storage and comparison of subsequences explicitly in the search for identical structures which can be generalized to the whole sequence. On the other hand, it has been acknowledged that constraints in the succession of events also become incidentally effective. For instance, an implicit learning mechanism has been assumed that incidentally associates current/preceding and forthcoming events if the former allow the prediction of the latter. Corresponding simulations in so-called *recurrent networks* have proven that adaptations to grammatical constraints may indeed result from such simple associative learning (Elman 1990, Cleeremans 1993). Other accounts have proposed that events are automatically integrated into a chunk if they are frequently and consecutively experienced in the same succession (Servan-Schreiber and Anderson

1990). Finally, the formation of response–effect associations has also been acknowledged as an autonomous process which may contribute to the formation of chunks in the sense of a series of effects which can be reliably predicted as the outcome of a certain sequence of actions (Hoffmann 1993). Thus, one can conclude that the acquisition and representation of serial order certainly is not a result of only a single process but rather the outcome of different mechanisms which all aim at improving the predictability of forthcoming events.

See also: Lashley, Karl Spencer (1890–1958); Learning and Memory, Neural Basis of; Mathematical Learning Theory; Mathematical Learning Theory, History of; Memory Models: Quantitative; Memory Retrieval; Serial Verb Constructions; Short-term Memory, Cognitive Psychology of; Short-term Memory: Psychological and Neural Aspects

Bibliography

Cleeremans A 1993 *Mechanisms of Implicit Learning*. MIT Press, Cambridge, MA

Ebbinghaus H 1902 *Grundzüge der Psychologie*. Veit, Leipzig, Germany, Vol. 1

Elman J L 1990 Finding structure in time. *Cognitive Science* **14**: 179–211

Hoffmann J 1993 *Vorhersage und Erkenntnis: Die Funktion von Antizipationen in der menschlichen Verhaltenssteuerung und Wahrnehmung*. Hogrefe, Göttingen, Germany

Hoffmann J, Koch I 1998 Implicit learning of loosely defined structures. In: Stadler M A, Frensch P A (eds.) *Handbook of Implicit Learning*, Sage, Thousand Oaks, CA, pp. 161–99

Lashley K S 1951 The problem of serial order in behavior. In: Jeffress L A (ed.) *Cerebral Mechanisms in Behavior*. Wiley, New York, pp. 112–36

Miller G A 1956 The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* **63**: 81–97

Miller G A, Selfridge J A 1950 Verbal context and the recall of meaningful material. *American Journal of Psychology* **63**: 176–85

Reber A S 1967 Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* **6**: 855–63

Restle F, Brown E R 1970 Serial pattern learning. *Journal of Experimental Psychology* **83**: 120–5

Rosenbaum D A, Kenny S B, Derr M A 1983 Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance* **9**: 86–102

Servan-Schreiber E, Anderson J R 1990 Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**: 592–608

Simon H A, Kotovski K 1963 Human acquisition of concepts for sequential patterns. *Psychological Review* **70**: 534–46

Stadler M A, Frensch P A (eds.) 1998 *Handbook of Implicit Learning*. Sage, Thousand Oaks, CA

Underwood B J, Shaughnessy J J, Zimmermann J 1974 The locus of the retention differences associated with the degree of hierarchical conceptual structure. *Journal of Experimental Psychology* **102**: 850–62

Ziessler M 1998 Response–effect learning as a major component of implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**: 962–78

J. Hoffmann

Representations and Rituals of the State

'The state,' wrote Michael Walzer (1967, p. 194), 'is invisible; it must be personified before it can be seen, symbolized before it can be loved, imagined before it can be conceived.' Like all human institutions, the state is a symbolic construction. It is not something that can be touched, nor does it exist independently of people's ability to relate individual actions to a complex and abstract set of ideas. In this article the ways in which the state is constructed through such symbolic processes and how political actors make use of a variety of symbolic means to exert control over it are examined.

1. The Role of Symbolism in Politics

This perspective on the role of symbolic processes in understanding the state is not universally shared. From the time of the Enlightenment, and through to the current day, many scholars and other analysts have identified the real stuff of politics with the play of material forces and the formation of groups by people defending their common interests. In this view, the symbolic dimension of politics—the flags, the ceremonies, the songs and insignias—is of little consequence.

Politics, however, is at its root a symbolic process, for both the creation of political groupings and the hierarchies they spawn depend on symbolic activity. Political reality does not come to us already formed. It must be created out of the symbolic framework that we have previously acquired. The selective perception and radical simplification of potentially relevant stimuli are necessary if we are to make any sense at all out of the infinite number of potential observations that we could theoretically make. The state, like other political institutions, groupings, and roles—political party, ethnic group, president, judge, king—requires symbolic work of this kind. The notion that the world is divisible into a series of mutually exclusive states, corresponding to lines on a map, and that any individual can be assigned to membership in one of them, is part of this symbolic construction of reality (Anderson 1983, Scott 1999).

1.1 Politics as Nonrational Action

Given that people's views of the state are symbolically constructed and the symbols on which they base their

views are manipulated by various political actors, it is misleading to see the state as the product of rational processes. This is all the more true because one of the most salient characteristics of the symbolic process is that it not only fashions certain views of reality, but also invests these perceptions with heavy emotional weight. 'A fine symbol,' wrote one Italian analyst, 'has greater weight than a mediocre fact' (Camillo Pellizzi, cited in Gentile 1993, p. 84).

1.2 The Nature of Political Representations

If, as beings who attempt to make sense of the world, we relate to politics not directly, but only through symbolic means, just how does this process work? Anthropologists have long been interested in such questions, although the pioneering work in this field suggested that it was peoples in small-scale, nonliterate societies who were particularly at the mercy of such symbolic activities in their political life. Meyer Fortes and E. E. Evans-Pritchard, in their 1940 volume, *African Political Systems*, which created the subfield of political anthropology, pointed to the mystical credentials of the African ruler, who served as the symbol of the people's unity. 'Members of an African society,' they wrote, 'feel their unity and perceive their common interests in symbols, and it is their attachment to these symbols which more than anything else gives their society cohesion and persistence' (1940, p. 17).

More recent anthropological investigation of state-level politics has retained the Durkheimian emphasis on the key role that symbols and rituals play in promoting social order. But it has expanded the paradigm in two ways. First, the Fortes/Evans-Pritchard formulation regarding the political role of symbolism has been generalized to apply to all societies, no matter how large or complex. Second, it is now recognized that the symbolic process is just as crucial in bringing about political change as it is in maintaining the political status quo. Clifford Geertz's work has had a major impact here. 'At the political center of any complexly organized society,' he wrote, 'there is both a governing elite and a set of symbolic forms expressing the fact that it is in truth governing … they justify their existence and order their actions in terms of a collection of stories, ceremonies, insignia, formalities, and appurtenances that they have either inherited or, in more revolutionary situations, invented' (1985, p. 15).

The views of Pierre Bourdieu, who has directed attention to the symbolic processes through which potential political leaders create followings, have also been influential. Bourdieu writes of 'the alchemy of representation … through which the representative creates the group which creates him.' Hence we have Louis XIV's magical equation '*L'état, c'est moi*,' or what Bourdieu tells us amounts to the same thing: 'the world is my representation' (1991, pp. 105–6). Of great importance is determining the conditions under which an individual is in a position to create representations that others are likely to follow, a matter dealt with by Bourdieu in terms of 'symbolic capital.' Anthropologist Simon Harrison (1995, p. 255) has picked up on this approach, viewing political conflict as a battle fought through symbolic means over symbolic capital. 'Competition for power, wealth, prestige, legitimacy or other political resources seems always to be accompanied by conflict over important symbols, by struggles to control or manipulate such symbols in some vital way.'

1.3 The Political Power of Ritual

Ritual provides one of the most powerful mechanisms through which symbolism enters the political realm. Here we adopt a broader view of ritual than the traditional equation of rites with religion. In this broader view, ritual is seen as consisting of culturally standardized, repetitive activity, which is primarily symbolic in character, yet ritual need not have any direct reference to the supernatural realm. Durkheim contributed to this perspective by linking ritual to the division of the world into two categories: the sacred and the profane. Ritual, he argued, consists of 'the rules of conduct which prescribe how a man should comport himself in the presence of these sacred objects' (1912 [1947], p. 41). The fashioning of rituals surrounding the state is in good part a matter of the creation and maintenance of sacred objects and the separation of these objects from the profane world.

Rich in symbolism, ritual derives its potency in part from its ability to arouse the emotions and the sensations through physiological stimulation. Rituals are filled with motions, gestures, colors, noises, songs, and odors. Through rituals emotions are joined to certain cognitions; that is, to certain views of the world.

Following Durkheim, scholars previously emphasized the conservative role of political ritual. Rites surrounding royalty took on a paradigmatic status. The rituals of rulers, the rites of court, and the ritualized means by which rulers interacted with their subjects—as in royal processions—were seen as means by which traditional authority was bolstered. More recently, however, a different perspective has emerged, in which ritual is viewed as a potent form of political action that serves not only power holders but also those seeking to overthrow the political status quo. In societies where the legitimacy of power is contested, we find the political battle fought, in part, through ritual. Rites (a) help develop political organization, not least through linking individuals to larger, abstract political entities (whether states or revolutionary organizations aimed at overthrowing them); (b) legitimize political power holders, would-be power holders, and political institutions; (c) create a sense of solidarity, even in the absence of consensus; and (d) foster certain views of political reality (Kertzer 1988).

The forms taken by political rituals are diverse. Individuals are linked to larger political entities through a variety of ceremonial means, from rites of membership in political parties, to rites of swearing allegiance as a new citizen (dubbed, tellingly, 'naturalization'), to participation in mass demonstrations. Leaders are legitimated through rites of inauguration, and the performance of an assortment of rituals that associate them with powerful symbols. Thus, when the president of the United States makes a public appearance, he enters to the stirring tune of 'Hail to the Chief'; all present rise, and he steps behind a podium decorated with the presidential seal, with the American flag waving from a pole at his side. Likewise, stripping a leader of his power outside regular rhythms of replacement necessitates rites of degradation, as in the case of impeachment proceedings. The death of a political leader requires special ritualization, affording a crucial opportunity for defining political reality and political values, as in the case of state funerals (Verdery 1999). These funerals have their own peculiar grammar, with the presence or absence of other political leaders sending powerful messages that can affect interstate relations.

2. *Representing the State*

Social scientists have long debated the nature of the state. Anthropologists have paid most attention to it as a stage in sociopolitical evolution, while political scientists debate whether there is anything necessarily Western about the modern state. For our purposes we view the state as consisting of 'the institutions of government and its individual representatives which together possess legitimate means of coercion and ability to tax and plan fiscal policy on a national level.' The state is charged with providing for its population's defense, security, and welfare (Biersteker 1987, pp. 50–70).

Having considered the nature and importance of symbolism to the political process, and the role of ritual in politics, we now turn specifically to how states are represented through symbolism, and the role played by ritual in the state.

2.1 *Organizing the State*

Given that people can only perceive the state through its symbolism, how do symbols do their work? The French historian, Maurice Agulhon (1985), addressing this question, distinguishes four categories of symbols through which modern states are represented. The state, first of all, must have a flag. Second, it is personified, either by a current or past head of state (e.g., a king or a dictator), or, as in the case of the French republican state, by an allegorical figure (Marianne). Third, each state has a variety of other kinds of symbolic representations of a nonanthropomorphic type: these include coats of arms, official seals, and the like. Finally, every regime, writes Agulhon, has its own Pantheon of historical notables, and typically places statues of such heroic figures (or, as the case in the Soviet Union, their embalmed bodies) in public squares.

Historians point to the French and American Revolutions as creative cauldrons for inventing new modes of representation of the modern state, with innovations ranging from the national flag to the national anthem. Historians have also shown that many of these forms of representation, while conveying an image of great antiquity, are in fact relatively recent. They are used by state elites both to solidify their own position internally and in international competition to bolster the state's position in the battle for influence (Hobsbawm and Ranger 1983).

The use of a variety of symbols and rites to link the masses of people to the state took on new importance with the rise of nationalism and democratic politics. When the state's legitimacy was vested in the leader by virtue of supernatural sanction (as in the case of divine right), the main requirement was for representations coming from above to reinforce the view that the rulers ruled by such higher authority. But when states came to rely on the notion that they existed at the behest of their citizenries, with no intermediate social body mediating between citizens and the state, the multifarious ways in which individuals identified with the state became much more important.

The heavy use made of such participatory representations of the state by twentieth-century European dictators, and especially by Hitler and Mussolini, led analysts to pay special attention to the importance of symbolism to the Nazi and Italian fascist regimes. George Mosse (1976, p. 39) has written of the ways in which in Nazi Germany the 'myths and symbols of nationalism, its rites, monuments, and festivals,' transformed an inchoate mass of people into a cohesive mass movement of citizens who saw themselves as in some sense acting as one. Mosse portrayed this process in religious terms, as 'the shaping of the crowd into a congregation.' Others have examined Mussolini's similar attempts to craft a system of symbols and rites to alter Italians' relationship with the state. Yet, in fact, such systems of representation and ritual are common to all modern states.

The importance of ritual in establishing the link between citizens and the state means that there are few more effective means of showing opposition to a regime than through contestation of the rites that bind the citizen to it. Postunification Italy offers a typical example. In the decades following the September 20, 1870 conquest of Rome, which brought with it the final construction of the Italian state, patriotic rites were organized each year to commemorate that sacred date. These annual celebrations, however, served only to inflame the passions of Catholic loyalists, who, following papal instructions, refused to recognize the

new state's legitimacy. The result was that every September 20 Catholics organized counter-rites to challenge the state-organized rites, and thereby challenged the legitimacy of the new state itself (Gentile 1996).

2.2 Legitimating the State

If states come to be known through the ways they are symbolized, any threat to the symbols with which they are identified threatens the legitimacy of the state itself. Harrison (1995, p. 258) refers to symbolic conflict between states of this sort as 'proprietary' contests, involving competition over the state's distinguishing symbols. The bitter dispute that erupted in the early 1990s between Greece and the former Yugoslav republic of Macedonia over the use of the name 'Macedonia' is a case in point.

Harrison distinguishes a variety of means through which the political representation of the group is in a process of continuous contestation and flux. This perspective applies well to the representation of states, again seen both in terms of internal competition and the competition that is waged among them. In the 'expansionary' contest, for example, one competitor attempts to displace another's symbols of identity with its own. This is found whenever revolutionary movements take power. In the case of the Russian revolution, as in revolutions everywhere, these battles were waged in part through the symbolism of space (e.g., St. Petersburg became Leningrad).

Although leaders of new states or new regimes try to replace their predecessors' symbols with their own, they rarely attempt to displace all such symbols. The reason is that, in the eyes of the population, the new regime's legitimacy is almost always tied to its identification with pre-existing symbols of state legitimacy. Hence, the Bolshevik leaders kept the Kremlin as the symbolic center of the state. Alternatively, older symbols of political legitimacy of the state are reintroduced, as in Mussolini's attempts to link his regime to ancient Rome through the symbolic elaboration of *romanità*.

Where older symbolic forms are judged unsuitable, new regimes are as likely to transform them for their own ends as they are to abandon them, for the continuing popular influence of the rites has its uses. Hence, the Nazi regime retained most of the old state holidays—from May Day to Mother's Day—but largely altered their symbolism and ritual in order to promulgate the new state ideology (Taylor 1981). In the post-World War II period, as the Polish Communist government attempted to win popular support, leaders similarly attempted to borrow legitimacy from older symbols, while changing their meaning. Accordingly, the new state authorities chose not to replace the crowned White Eagle, which had served as a national symbol of Poland for centuries. Rather, they simply removed the eagle's crown, thought to be too blatant a symbol of the old aristocracy. Instructively, on various occasions on which workers protested against the state, they displayed an image of the white eagle recrowned (Kubik 1994, pp. 50–1).

This emphasis on the struggle over representations also allows us to examine how political symbols and rituals are employed by combatants for power within states as a means of establishing their own legitimacy and carving out a particular political identity for themselves. A number of observers, in this connection, have cited the example of the Orange Order's annual parade on July 12, when men wear black bowler hats and orange collars, and march for hours holding aloft banners commemorating battles that three centuries earlier helped establish Protestant rule in Ireland. These rites have some parallels in Independence Day celebrations in the United States, but here the political context makes all the difference. In the United States, the legitimacy of the state is largely undisputed, and the July 4th rites are ceremonies that foster solidarity and reinforce the legitimacy of the state. In Northern Ireland, by contrast, the state's authority remains contested, and such rites as those of the Orangemen provide emotionally powerful means of organizing the combatants.

2.3 Establishing New Regimes

As the cases of the French and Russian revolutions show, new regimes require symbolic elaboration, and they must deal with the lingering power of the symbols of legitimacy of the old regimes that remain behind. Struggles over symbols and rites play no small part in the battle to establish such regimes and in the attempts of competitors to delegitimize them. Hence the tricolor flag appeared in France in 1789 immediately following the taking of the Bastille, then was replaced by the royal white flag on Napoleon's fall in 1814, before the July Revolution of 1830 again resulted in the banishment of the white flag and the triumph of the tricolor. Agulhon, in reviewing the history of these developments, argues that battles of this sort, fought through and over symbolism, were central to the course of political events. 'The obscure militants,' he writes, 'who, on July 28, 1830, went to raise the tricolor on the towers of the Notre-Dame and over the Hôtel-de-Ville while the outcome of the struggle was as yet uncertain, contributed to throwing the people into battle with a redoubled ardor' (1985, p. 190).

Rituals of allegiance offer particularly effective means of identifying those whose loyalty to the new regime is weak, and at the same time furnishing ammunition for the regime's claim that it enjoys widespread popular support. Because state authorities can require their performance, such rites force individuals who oppose the regime either to be publicly identifiable (and hence risk reprisal) or to undermine their own self-esteem and their sense of being opposed

to the rulers. Such was the case of the arm-stretched 'Heil Hitler' salute required in daily social encounters during the Third Reich. Similarly, in 1931, concerned about rooting out subversive university professors, and at the same time eager to offer evidence of its claim that it enjoyed overwhelming support from all segments of the population, the Italian fascist regime required that all faculty members swear allegiance to it. Only 12 (of over 1200) professors in the entire country refused to take part in the rite. Their symbolic act of resistance did not go unpunished; they were stripped of their posts (Goetz 2000).

New states formed out of former colonies continue to face dramatic problems of representation and legitimation. In sub-Saharan Africa, where many new states emerged in the 1960s, the difficulties were compounded by the fact that their boundaries were those that the colonial powers had drawn. How could such divisions of land be represented as 'natural' bases for political organization, much less objects of popular adulation? What role would continuity have in such transitions, and if the symbols of the colonial past had to be rejected, what would take their place?

Swaziland provides an intriguing example (Kuper 1978, pp. 301–4). Following decades of British colonial rule, Swaziland gained its independence in 1968, using the figure of the Swazi king as the centerpiece of the new order. The ceremonies marking independence took on great importance to the political elite, who saw them as a means both of defining the new state and building public allegiance to it.

Independence was marked by a great public ceremony, held in an open stadium, packed with thousands of enthusiastic citizens. The king made his entry in a huge new black Cadillac limousine, but as he stepped out people saw that he was wearing traditional regal garb, with high-plumed headgear. After the king took his seat, the British ambassador arrived, as the band played 'God Save the Queen.' From there the ceremonies followed Western form until, suddenly, thousands of swaying Swazi warriors began singing a traditional war song, and the king and prime minister left their seats to join them. As the song reached its crescendo, and the king along with the thousands of warriors raised his shield over his head, the crowd roared its approval. The king and prime minister returned to their seats, later getting in their stretch Cadillacs to take them to a diplomatic reception.

The Swazi case illustrates the power of symbols and rites not only to define a new political reality through the combination of symbols derived from various sources, but also their power to foster popular legitimation and a sense of identification between citizens and the new state.

2.4 The Uses of History

A common feature of representations of the state is that they contain an image of history. It is through symbolic expressions, including rituals, that leaders link the state today to its earlier incarnations, and likewise link both themselves and their enemies to particular predecessors, viewed in certain ways.

New leaders do all they can to identify themselves with popular predecessors. American presidents need not rely solely on verbal reminders, citing Washington, Jefferson, Lincoln, and Roosevelt in their speeches. By performing certain state rites performed by these predecessors—as most notably in the Inauguration ceremonies—they effectively make this connection.

In the relationships between states, the importance of representing history in certain ways, and establishing legitimacy for the present through symbolic links of a certain kind to the past, is likewise crucial. This was evident, for example, in the difficulties the German government experienced in the decades following World War II in dealing with the historical memory of the war. These came to a head in a series of state rituals held in the mid-1980s. In 1984, leaders of the former Allied nations held a commemorative rite on the anniversary of the Normandy landing. This was similar to many such rites that served to keep the memory of the war alive, and the leaders who participated in it no doubt derived some political benefit themselves. But German leaders, representing the enemy nation, were not invited.

Angered at having the memory of the war represented in this way, German Chancellor Helmut Kohl invited US President Ronald Reagan to come to Germany to celebrate with him a commemorative ceremony in honor of the war dead, to be held at the Bitburg military cemetery. By having the American president stand with him there, the German leader hoped to refashion the history of the war. Indeed, the American President cooperated in this effort, referring to the dead German soldiers as 'victims' of the Nazis, 'just as surely as the victims in the concentration camps.' Yet, history is not infinitely malleable; the symbolism of World War II in the United States could not be so rapidly and radically altered, and as a result Reagan came under sharp attack. Attempting to make partial symbolic amends, the President added a visit to a concentration camp to his itinerary that day, yet this then opened him up to the charge that he was thereby symbolically equating the concentration camp victims with their Nazi persecutors. Despite such criticisms, however, the Bitburg rites had their effect. Opinion polls following the ceremony showed a considerable rise in the proportion of Americans who saw the German soldiers as victims of Hitler rather than as responsible for Nazi atrocities (Kertzer 1988, pp. 92–5).

2.5 The Future of Representations and Rituals of the State

As long as there continue to be states, they will continue to require cultural means of representation,

and in this, ritual in its various forms will play an important role. However, as states face competition for sovereignty and for influence both from within and from without, the role and importance of such representations are likely to change, and some of their forms may alter as well. Insofar, for example, as one role of the rituals we have discussed here is linking the individual to the state, any change in the importance of the state as a basis for personal identity will be reflected in changes in participation in and popular perception of the rituals themselves.

These competing claims for people's allegiance, in turn, will require their own forms of representation, including ritual. Regionalist, ethnic, religious, and other movements opposing the state's power presuppose both the delegitimation of the state—a battle fought in good part through efforts at altering existing representations and rituals of the state—and the creation of an alternative identity. In Italy, for example, the Northern League has devoted tremendous attention to the creation of symbols of representation and rituals that link individuals to a previously nonexistent territory—Padania—and to the League itself (Destro 1997). At the same time, ritualized display of the symbols of the Italian state—such as the Italian flag—is studiously avoided.

In addition to such threats from below (in the Balkans, south Asia, Africa, and other areas), states also face threats from above. For decades theorists have been predicting the demise of the state and the emergence of larger, supranational entities. Many of these (e.g., the Arab League) have foundered on the unwillingness of national elites to cede authority. The most notable current effort, the European Union, demonstrates the link between weak popular allegiance and an underdeveloped system of representation and ritualization (although also suggesting that causation goes both ways). French anthropologist Marc Abélès (1996, pp. 125–32) refers to '*le déficit symbolique*' of the EU, writing that the supranational organization sorely lacks both the symbols and the rituals that could furnish it with some power of popular attraction. Without such symbols and rites, it is unable to mobilize popular passions. The reasons for this deficit are many, but include the inability and indeed unwillingness of national elites to agree on a clear representation of what the EU stands for. Indicative is the fact that while the EU has its own anthem (a version of Beethoven's '*Ode to Joy*'), people cannot sing it, for it has no words. The EU's rites—such as the commemoration of the fortieth anniversary of the European Parliament in 1992—have been so devoid of emotionally resonant ritual elements that they do little but reinforce the EU's image as a creature of Euro-bureaucrats.

Future research must take aim at these potential competitors to the state from below and from above, and examine the ways in which these competitors employ representations in general and ritual in particular to delegitimize the state. In all such studies, it is crucial that people's reactions to these efforts be examined carefully. It is one thing for state elites or would-be revolutionaries to employ symbolic means to try to win the allegiance of those whose 'representatives' they would like to be. It is quite another for the 'represented' to respond to these symbolic efforts in the ways the elites would like. What is needed is a mix of methods—from participant observation to the analysis of texts—and an examination of a range of social settings, from large periodic national demonstrations to quotidian interactions in rural communities. Political scientists and sociologists need to pay more attention to the symbolic dimension, while anthropologists need to take aim more directly at state-level forces.

See also: Collective Identity and Expressive Forms; Hegemony: Cultural; Legitimacy; Legitimacy: Political; Legitimacy, Sociology of; Political Advertising; Political Communication; Political Culture; Political Discourse; Ritual

Bibliography

Abélès M 1996 *En attente d'Europe*. Hachette, Paris

Abélès M 1997 La mise en représentation du politique. In: Abélès M, Jeudy H-P (eds.) *Antropologie du politique*. Armand Colin, Paris

Agulhon M 1985 Politics, images, and symbols in post-revolutionary France. In: Wilentz S (ed.) *Rites of Power: Symbolism, Ritual, and Politics Since the Middle Ages*. University of Pennsylvania Press, Philadelphia, PA

Anderson B 1983 *Imagined Communities*. Verso, London

Biersteker T 1987 *Multinationals, the State, and Control of the Nigerian Economy*. Princeton University Press, Princeton, NJ

Bourdieu P 1991 *Language and Symbolic Power* [trans. Raymond G, Adamson]. Harvard University Press, Cambridge, MA

Destro A 1997 A new era and new themes in Italian politics: The case of Padania. *Journal of Modern Italian Studies* **2**: 358–77

Durkheim E 1912 *Les formes elementaires de la vie religieuse* [1947 *The Elementary Forms of the Religious Life*. Free Press, Glencoe, IL]

Fortes M, Evans-Pritchard E E 1940 Introduction. In: Fortes M, Evans-Pritchard E E (eds.) *African Political Systems*. Oxford University Press, London

Geertz C 1985 Centers, kings, and charisma: Reflections on the symbolics of power. In: Wilentz S (ed.) *Rites of Power: Symbolism, Ritual, and Politics Since the Middle Ages*. University of Pennsylvania Press, Philadelphia, PA

Gentile E 1993 *Il culto del littorio*. Laterza, Rome [1996 *The Sacralization of Politics in Fascist Italy*. Harvard University Press, Cambridge, MA]

Goetz H 2000 *Il giuramento rifiutato: I docenti universitari e il regime fascista*. La Nuova Italia, Milan

Harrison S 1995 Four types of symbolic conflict. *Journal of the Royal Anthropological Institute* (*n.s.*) **1**: 255–72

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Kertzer D 1988 *Ritual, Politics, and Power*. Yale University Press, New Haven, CT

Kertzer D 1996 *Politics and Symbols*. Yale University Press, New Haven, CT

Kubik J 1994 *The Power of Symbols Against the Symbols of Power: The Rise of Solidarity and the Fall of State Socialism in Poland*. Pennsylvania University Press, State College, PA
Kuper H 1978 *Sobhuza II: Ngwenyama and King of Swaziland*. Africana, New York
Lukes S 1975 Political ritual and social integration. *Sociology* **9**: 289–308
Mosse G L 1976 Mass politics and the political liturgy of nationalism. In: Kamenka E (ed.) *Nationalism: The Nature and Evolution of an Idea*. St. Martin's Press, New York
Scott J C 1999 *Seeing Like a State*. Yale University Press, New Haven, CT
Taylor S 1981 Symbol and ritual under National Socialism. *British Journal of Sociology* **32**: 504–20
Verdery K 1999 *The Political Lives of Dead Bodies*. Columbia University Press, New York
Walzer M 1967 On the role of symbolism in political thought. *Political Science Quarterly* **82**: 191–205

D. I. Kertzer

Reproductive Medicine: Ethical Aspects

In reproductive medicine a variety of diagnostic methods are used to discover causes of involuntary human infertility. Diseases in this area can be treated by a number of techniques, ranging in complexity from the simple and inexpensive to the highly sophisticated and expensive (Reiss 1998). The diagnostic methods used include both the invasive, such as laparoscopy, and the noninvasive, such as ultrasonography. They may be specific, in the sense that they give the answer to a question posed, or nonspecific, providing information that has not been asked for. Some diagnostic tests are quick, while others take longer, since they have to be processed in laboratories.

The development of a variety of more or less effective approaches to involuntary human infertility, as well as the separation of human reproduction from sexuality, raises a number of ethical problems, discussed in books, journals, and the mass media (Shenfield and Sureau 1997).

1. *Ethical Controversies*

Cloning has recently been the subject of debate in many countries. There are different kinds of cloning, but the World Health Organization (WHO) and the Council of Europe have banned the cloning of human fertilized eggs for reproductive purposes. The objections to cloning, apart from safety issues, concern the identity and individuality of human beings. Nature produces identical twins, but many people argue that cloning of human beings will undermine the respect of the human being by making its identity and individuality subject to human ingenuity.

Current ethical issues include the disposal of frozen embryos (how and under what conditions this may take place), and posthumous treatment, much discussed in the UK in the late 1990s. Other controversies are raised by third party gamete donation and sperm cryopreservation, particularly concerning payment of donors, anonymity of donors, and selection of donors and recipients (discussed in, e.g., Council of Europe 1996). The different methods used in reproductive medicine raise different ethical problems, and hence need to be considered separately. The issues of oocyte cryopreservation, and reproductive tissue freezing (ovaries and testis), also raise issues of safety and consent—particularly for young people affected by cancer.

For example, assisted insemination using the semen of the husband or designated partner raises a somewhat different problem from that of assisted insemination using the semen of a third party. The need for careful counseling is important in both cases, but the status of the child, its right to know its biological father, needs to be considered in detail when a third party is involved. When semen is frozen, safety issues need to be considered, in addition to how long the semen may be stored as well as under what conditions and for what purposes it may be used. When semen banks are established, regulation is needed. If and when the semen is tested or manipulated, further problems arise.

At present, at the beginning of the twenty-first century, preimplantation diagnosis is limited to sorting out fertilized eggs carrying genes associated with severe diseases for which there is currently no cure, like Huntington's chorea and certain X-linked diseases. Concern has been expressed that the practice may be extended to diseases which are less serious. Obviously, there is a sliding scale between more or less serious and also room for subjective differences in this grading of diseases. There is thus a risk of a slippery slope in this area (Shenfield 1997).

A fundamental issue concerns the moral and philosophical status of the fertilized egg and to what sort of protection it is entitled (Steinbock 1992, Hermerén 1993). The status of the embryo is crucial in many debates of the ethical dilemmas of reproductive medicine. Other basic concerns include the partly culture-dependent reasons for reproduction.

2. *A Framework for Analysis*

The underlying values in the debates on dilemmas in reproductive ethics can be described in different ways, in terms of the rights and interests of the various stakeholders concerned, or in terms of duties towards patients and the future child. These duties may clash with each other as well as with the interests and rights of other parties. To make the underlying values and value conflicts explicit is to take a first step towards a rational discussion of these dilemmas.

An ethical problem arises in a situation where: (a) there are two or more possible courses of action; (b) one course of action satisfies or is compatible with certain interests, rights, or values, and another course of action satisfies others; and (c) all parties or all interests, rights, or values cannot be satisfied at the same time by one possible action. This is particularly the case if (c) is intensified in such a way that both options are incompatible with each other or if the values, which are included within the options, are conflicting with each other. Whatever alternative is chosen, there will be winners and losers. This is clearly illustrated by some of the well-known dilemmas in reproductive medicine, where the stakeholders (or parties involved) include the woman or the couple, present and future children, doctors and clinics, researchers, the pharmaceutical industry, taxpayers, and society at large.

3. *Terminological, Empirical, and Value Issues*

If a rational discussion of these ethical dilemmas is wanted, it is necessary to separate conceptual, empirical, and normative issues, as illustrated by the controversies over when life begins and the proposed distinction between embryo and pre-embryo (Jones and Schrader 1989).

Evaluation of criteria of success like 'take-home baby rate' also involves issues of all three kinds. Moreover, whether certain methods would harm children and parents or alter people's understanding of themselves or of the act of procreation are empirical issues that people may disagree about, even if they agree on the goals of reproductive medicine. To be able to substantiate these empirical claims it would be necessary to proceed with long-term follow-up studies of children produced with the help of the various reproductive technologies available.

The utilitarian principle of maximizing the utility or achieving the greatest good for the greatest number of people is open to many interpretations, some of which will clash with Kantian principles that each person should be treated as an end in itself and not as a means to an end. They may also clash with rights-based theories of the sort advocated by Dworkin (1977). This is particularly evident in the debate over research on embryos and the conditions for destruction of spare embryos. Here respect for the autonomy of the gametes providers, by requiring their joint consent, may yield different conclusions than a utilitarian starting point would do. According to the latter, the consent of the gametes providers may be disregarded if it can be shown that using the spare embryos for research or for donation will maximize the greatest good in society. Whereas Swedish legislation requires that sperm donors be willing to reveal their identity to the child when he or she is grown up, if the child wants to know, French legislation requires anonymity (Shenfield 1997). These different requirements may have an effect on the number and kind of people willing to donate sperm. Interestingly, both requirements can be justified by considerations based on the welfare of the future child. Thus, there is a common agreement on the goal but different views about the best way to achieve it.

These differences could clearly be solved in principle by empirical research, involving long-term longitudinal follow-up studies in both countries of children who are the result of sperm donation. But these studies themselves raise further ethical problems. As the children reach a mature age, it becomes increasingly difficult to carry out the follow-up without their informed consent. How is this to be obtained in a country which makes anonymity compulsory? And can it be obtained without risk of harming or stigmatizing the children?

4. *Goals*

It may be helpful to distinguish between questions concerning the ethical acceptability or legitimacy of certain goals of reproductive medicine, such as the improvement of success rates in *in vitro* fertilization (IVF), and the ethical acceptability of certain methods used to achieve these goals. As indicated above, it is possible to agree on the goals and differ on the methods, and vice versa.

Do the techniques mentioned above help to achieve the goals of reproductive medicine? Do they help to achieve other goals as well?

If the goal of reproductive medicine is to cure diseases, it becomes necessary to clarify what is meant by a disease to understand this goal. If a particular condition is regarded as a disease, it falls within the goals of reproductive medicine to cure it; if not, it does not.

A fundamental issue is whether involuntary infertility is a disease, the result of a disease or a socially constructed need. This is not just a matter of semantic definition, since the different ways of regarding human infertility have social and economic consequences. If it is a disease, those suffering from it are entitled to help from the national or publicly funded health care system, whose resources should be used. Moreover, if infertility is a disease or a response to a disease rather than a social need, a case for financial support of the new reproductive technologies can also be made, as well as for research on such technologies. If, however, it is a socially constructed need, the case for financial support is much weaker.

5. *Means*

The methods available to achieve the goals of reproductive medicine vary, and new methods are being

tested. Owing to breakthroughs in molecular biology, development is rapid (Fauser 1999, Harris and Holm 1998). Some methods, like surrogacy, have been criticized because they lend themselves to commercial exploitation and to seeing children as commodities, as 'made' rather than 'begotten' (O'Donovan 1984).

If and when the methods of reproductive medicine involve or presuppose preimplantatory or prenatal diagnostics, further issues are raised. The case for preimplantatory diagnostics in connection with IVF is that it can be used to avoid X-linked diseases such as hemophilia or certain forms of dystrophy. However, concerns are raised over the possibility that, for instance, indications may be expanded, and that they may be misused for eugenic purposes. Hence regulations or guidelines are necessary, and have been proposed in many countries.

Sex selection for social and economic purposes is practiced in many countries, like India and China. If there are normally about 105 girls born to 100 boys, population statistics indicate that perhaps 100 million girls are 'missing' in the world because of sex selection. The status of women in the world is subject to many cultural differences, and the most important arguments against sex selection focus on the equality of the sexes, emphasizing that gender is not a disease (Shenfield 1994). Suppose, however, that there is no social security for parents when they reach old age. Moreover, suppose that parents have to pay an expensive dowry for their girls, and that girls are not able to make a living by working outside of the home. Then the much-criticized practice of sex selection could be viewed as a culturally chosen means to achieve a certain goal (economic and social safety for the parents). This obviously further suggests that in order to abandon this practice of sex selection, alternative means of reaching this goal have to be made available. Since this may take considerable time, compromise positions have been suggested.

Further concerns have been expressed that, because some of the technologies used in reproductive medicine are expensive they will only be available for rich people in the industrialized parts of the world. They will thus help to increase the gap between the health care and quality of life available for the rich and that for the poor or, more generally, the gap between rich and poor countries.

6. *Some Important Arguments*

As already mentioned, the basic ethical dilemmas of reproductive medicine involve conflicts of interests, rights, or values. The first step in the attempt at a rational discussion of these conflicts is to make the conflicting interests, rights or values explicit, and to try to separate normative and empirical issues. This is sometimes easier said than done, because, for example, in the discussion of the importance of a genetic relation between parents and children all these issues are intertwined in a complicated way. Various arguments are used in discussing these conflicts. One rarely takes a stand on the basis of one argument or principle only—often several arguments or principles are used. Sometimes those advocating different positions use the same arguments or principles, but rank the importance of the arguments differently. The central arguments seem to be the following ones: the autonomy argument, the non-maleficence argument (an argument based on the principle that one should not harm), and justice arguments of various kinds.

6.1 *Arguments of Autonomy*

These are derived from a position that attaches great weight to individuality and liberty. The principle of autonomy is the basis of the requirement of informed consent. If the technologies used involve risks, the principle of autonomy requires careful counseling of the candidate couples. If great importance is attached to this principle, it may overrule also the concern over commodification of human beings, their bodies, and their bodily products. If people have a right to do what they want, they also have a right to serve as surrogate mothers, to earn money on their bodies, and to sell their organs or bodily products. The problem of the limits of autonomy then becomes an important issue. This is also an issue of power: who is to set the limits, and in accordance to what other principles?

6.2 *Nonmaleficence Arguments*

Arguments based on the principle that one should not harm are clearly legitimate and have a long history in medical ethics ('primum est non nocere,' a principle known already in antiquity). It has been claimed that children who grow up in nontraditional families (for example, with two parents of the same sex) or are the result of particular reproduction technologies will suffer, because they lack role models of both sexes or because their intellectual or emotional development will be impaired. At present, though, the evidence supporting this or the opposite position is missing, or at best inconclusive.

6.3 *Arguments of Justice*

Arguments based on the just distribution of scarce resources are also relevant. Clearly the resources for health care are not unlimited in any country, and there is always the problem of the fair use of these limited resources. Is the unequal access to treatment also unfair? If fewer than half of those who ask for it obtain assisted treatment by IVF, the reasons why some obtain treatment and others not take on an ethical

significance (Shenfield 1997). Is this a case of discrimination on the basis of economic, social, or ethnic grounds or not? Moreover, the practice in different counties as to how many cycles are subsidized is also relevant here, since incomes are unevenly distributed. For some, a fee of EUR.10 000 is a small sum, but to others it is an effective barrier.

Fair access is clearly also a legitimate concern. If the available resources are to be used in an equitable way, and used to benefit those who suffer most, rather than to benefit those who are willing and able to pay, it has to be demonstrated that the methods are cost-effective. Though doctors should respect the autonomy of those who come to seek their help, doctors are not under any obligation, legal or moral, to provide patients with any treatment they might request (McCullough and Chervenak 1994). The request has to be compatible with good professional standards. The proposed treatment might be not benefit the patients or it might be incompatible with the interests of the child.

A related question that has been the subject of heated discussions is who should have access to the benefits provided by some of these reproduction technologies. Should IVF be available for single people, or for homosexual or lesbian couples? Should HIV couples or couples where one of the partners is HIV positive be entitled to treatment with artificial reproduction? Are there any age limits? Are any age limits appropriate and ethically justifiable? If the answer is yes, what sort of inquiries are appropriate before proceeding? In short, on what ground, if any, should some people be denied the access to these techniques?

7. *Research Ethics*

Here one set of problems focuses on the ethical acceptability of research on embryos *in vitro*, and another on the conditions of such research, if acceptable. According to the views of the Roman Catholic Church, neither such research nor IVF is morally acceptable. Embryo research is also explicitly illegal in some countries, such as Norway, and in practice illegal in others, such as Germany.

If, however, research on embryos *in vitro* is morally accepted, a number of more specific questions have to be asked and answered. They include what kind of research should be permitted, questions about the source of the embryos and for how long may embryos be used for research (14 days is a limit in many countries, like Sweden). Moreover, conditions have been laid down concerning the length of time frozen embryos may be saved and how spare embryos may be destroyed. All these issues are controversial, as is the question of whether embryos may be created solely for the purpose of research. The latter is explicitly forbidden in French legislation and in the Council of Europe Convention on Human rights and Biomedicine (1996). The fundamental issue here concerns the philosophical and moral status of the embryo.

The human embryo deserves 'respect,' but a much-contested question is what precisely this respect allows and what it forbids. The practice of research on embryos is subjected to strict regulations, as is experimental research on human beings. Informed consent by the male and female providers of gametes is a cornerstone in these regulations. The knowledge sought must be impossible to achieve in other ways, for instance, by research on animals. Further, this knowledge has to be of vital interest to present and future patients. Others regulations concern safety issues, in particular the safety of the applications of research.

Research on new reproduction technologies (diagnostic methods or treatment options) involves the standard issues of providing patients and healthy volunteers with information they understand and which enables them to make autonomous choices. Obtaining consent from research subjects who have deficiencies in understanding poses other difficulties. In this area of medicine, specific problems include that the fact the embryo and the future child are stakeholders, or parties concerned, but for obvious reasons, they cannot be heard. Someone, therefore, has to look after their interests. Often the parents do this, but sometimes they disagree. Who is then to look after the interests of the embryo or the child-to-be? Moreover, the women or couples seeking help may be under emotional or economic pressure, which will not aid rational decision making.

Yet another problem is over the safety of the woman or the embryo. Some of the methods used in research may involve physical, psychological, or economic risks. Protection against possible harm is a requirement for research—there always has to be a reasonable balance between the anticipated benefits of research and foreseeable harm to research subjects. If unexpected harm arises, however, measures have been taken to interrupt the study and to compensate the research subjects.

When sperm or human tissues are taken from tissue banks or sperm banks for research purposes, special problems have to be dealt with. The conditions under which such research may be performed have been regulated in many countries like the US, Sweden, and the UK.

Finally, there is the problem of fair access to research results, as well as controversies over payment or compensation to research subjects.

See also: Assessment and Application of Therapeutic Effectiveness, Ethical Implications of; Bioethics: Philosophical Aspects; Ethical Dilemmas: Research and Treatment Priorities; Ethics for Biomedical Research Involving Humans: International Codes; Family Systems and the Preferred Sex of Children; Fertility and Culture: Anthropological Insights; Fertility Control:

Overview; Fertility: Institutional and Political Approaches; Fertility: Political and Political–Economic Perspectives; Fertility: Proximate Determinants; Gender and Reproductive Health; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations

Bibliography

Callahan J C (ed.) 1995 *Reproduction, Ethics and the Law: Feminist Perspectives*. Indiana University Press, Bloomington, IN

Chadwick R F (ed.) 1992 *Ethics, Reproduction and Genetic Control*, rev. edn. Routledge, London

Council of Europe 1996 *Convention on Human Rights and Biomedicine*. Council of Europe, Strasbourg, France

Dworkin R M 1977 *Taking Rights Seriously*. Duckworth, London

Fauser B C J M (ed.) 1999 *Molecular Biology in Reproductive Medicine*. Parthenon, New York

Harris J, Holm S (eds.) 1998 *The Future of Human Reproduction: Ethics, Choice and Regulation*. Oxford University Press, New York

Hermerén G 1993 *The Philosophical and Moral Status of the Embryo*. Council of Europe, Strasbourg, France

Jones H W, Schrader C 1989 And just what is a pre-embryo? *Fertility and Sterility* **52**: 189–91

McCullough L B, Chervenak F A 1994 *Ethics in Obstetrics and Gynecology*. Oxford University Press, New York

Meschede D, Horst J 1997 The molecular genetics of male infertility. *Molecular Human Reproduction* **May 3**(5): 419–30

Nuffield Council on Bioethics 1993 *Genetic Screening: Ethical Issues*. Nuffield, London

O'Donovan O 1984 *Begotten or Made?* Oxford University Press, Oxford, UK

Reiss H E (ed.) 1998 *Reproductive Medicine: From A–Z*. Oxford University Press, Oxford, UK

Shenfield F 1994 Control of human sex ratios—sex selection—A matter for fancy or for ethical debate. *Human Reproduction* **9**: 569

Shenfield F 1997 The ethics of embryo research. In: Shenfield F, Sureau C (eds.) *Ethical Dilemmas in Assisted Reproduction*. Parthenon, New York, pp. 15–21

Shenfield F, Steele S J 1995 Why gamete donors should not be paid—A gift is a gift is a gift, or why gametes donors should not be paid. *Human Reproduction* **10**: 253–5

Shenfield F, Sureau C (eds.) 1997 *Ethical Dilemmas in Assisted Reproduction*. Parthenon, New York

Statens medicinsk-etiska råd. 2000. *Om Livets Början*. Etiska vägmärken, Stockholm

Steinbock B 1992 *Life Before Birth: The Moral and Legal Status of Embryos and Fetuses*. Oxford University Press, New York

Strong C 1997 *Ethics in Reproductive and Perinatal Medicine: A New Framework*. Yale University Press, New Haven, CT

Tong R 1997 *Feminist Approaches to Bioethics: Theoretical Reflections and Practical Applications*. Westview, Boulder, CO

G. Hermerén

Reproductive Rights in Affluent Nations

Reproductive rights can be defined as the freedoms and opportunities that allow individuals to choose if, when, and how to have children. Advocates of reproductive rights have argued that this choice involves access to contraceptives, abortion, and reproductive technologies; freedom from forced sterilization; and access to health care that allows women and their partners to participate in decisions regarding prenatal care and childbirth. Many feminists argue that genuine reproductive choice requires a broad set of changes, including gender equality, employment opportunities, and available high-quality daycare.

1. Historical and Intellectual Context

The notion of reproductive rights is very much a product of ninteenth and twentieth century women's movements. Feminist perspectives on reproductive rights include both individual and social components (Petchesky 1984). The individual element, rooted in liberal thought, is the idea that women must control their bodies to be independent persons. Liberal discourse has been critical to both reproductive rights advocates, who discuss reproductive rights as essential to women's equality, and their opponents, who make claims about the competing rights of fetuses, male partners, and parents of minors (O'Connor et al. 1999, pp. 158–9). The social component, which draws on Marxist thought, is the recognition that material conditions and social relations constrain reproductive choices. As Rosalind Petchesky (1984) argues, reproductive freedom is necessarily individual, involving women's bodies, but at the same time social, involving social arrangements and class and racial divisions that affect reproduction.

In the nineteenth century, birth control movements emerged in many affluent nations. In her history of birth control in America, Linda Gordon (1976) argues that feminist efforts to gain political power through suffrage, and power within the family through birth control, were a response to rapid industrialization; by gaining power and control, women might bolster the family and help to solve social problems. Although nineteenth century advocates of 'voluntary motherhood' disapproved of contraceptives as 'unnatural,' they advocated women's right to refuse sex with their husbands. This demand for the right to refusal was part of the fight against unequal marriage laws, including legalized rape, and was considered a condition of women's independence. In the words of one voluntary motherhood activist, it was 'Woman's Natural Right to ownership and control over her own body-self—a right inseparable from Women's intelligent existence…' (Gordon 1976, pp. 103–4), a sentiment that would be echoed by twentieth-century

feminists in their continuing battles for reproductive rights.

Although nineteenth-century American feminists firmly asserted women's right to control their bodies, they never endorsed contraceptives and never rejected motherhood. Gordon explains their continued disapproval of contraception as a result of their fear that contraceptive use would lead to extramarital sex, and that this would lead to greater freedom for men, but not for women. Similarly in England, nineteenth century feminists failed to support birth control, fearing that women would be victimized by men's sexual desires (Banks and Banks 1964). Lacking economic independence, women needed men's commitment to their wives and children in marriage. The voluntary motherhood movement was an attempt to restrict sexual morality, rather than create sexual freedom, which at that time was not seen as beneficial to women. Women's movements in other Western countries were animated by similar ideologies. It was not until additional social changes made it possible for women to broaden their worlds that women's movements broadened their conceptions of reproductive rights.

By the early twentieth century, women were entering the labor force in greater numbers and sexual activity outside of marriage was on the rise. 'Birth control,' a term coined by US activist Margaret Sanger in 1915, was advocated for both economic and sexual reasons. Socialists in Europe and North America pushed for access to birth control to improve the lives of the working class. Neo-Malthusians, who were concerned about population growth, advocated contraception and were generally supportive of sexual equality (Francome 1984). Advocates of sexual liberation, influenced by European sex theorists such as Havelock Ellis, saw birth control as freeing women of their dependence on men. Early twentieth-century feminists, however, were not in complete agreement about the benefits of birth control for women. Some feared sexual promiscuity and the degradation of motherhood as a source of status for women, and some had race and class biases that influenced their views on birth control. Nevertheless, birth control movements in affluent nations expanded greatly and were linked to women's sexual freedom. When women's movements in various countries were 'reborn' in the 1960s, access to birth control was an important demand.

Movements to legalize abortion also emerged in many countries in the twentieth century. Attitudes toward sexual behavior were becoming more liberal, and various women's groups advocated liberalization of abortion laws. By the 1930s, a number of European countries passed laws that provided for some limited access to abortion. The contemporary wave of abortion reform laws in affluent countries began with the passage of the British Abortion Act in 1967, which was pushed by the Abortion Law Reform Association, an organization founded in 1936. Although the British law gave medical practitioners rather than women control over the abortion decision, it was a significant advance in women's access to abortion (Francome 1984).

Abortion rights activism increased in the 1960s as women's movements flourished in many Western countries. Feminists demanded legal abortion as part of women's right to control their bodies and developed extended analyses of reproductive rights. In North America, collectives like the Montreal Health Press and the Boston Women's Health Book Collective distributed works such as *The Birth Control Handbook* and *Our Bodies, Ourselves*, which were early sources of information to women about contraception, abortion, childbirth, and other women's health issues. Women's health activists criticized male gynecologists, obstetricians, and other practitioners for demeaning women and failing to include them in decisions about their own bodies. They questioned the safety of some contraceptives and drugs and the necessity of some surgical procedures. They criticized high-technology childbirth practices that took control over childbirth away from women. As more black women and other ethnic minorities were drawn into the women's health movement, they raised the issue of forced sterilization of poor and nonwhite women. As alternatives to established medical control over women's health care, feminists developed 'self-help gynecology' and opened feminist health clinics (Ruzek 1978).

In countries where the women's movement was strong, feminists influenced abortion reform movements to advocate abortion as a women's rights issue. In the United States, the Supreme Court legalized abortion in 1973 on the basis of a 'right to privacy,' recognizing the rights of both women and doctors to make decisions about abortion. Other Western countries also legalized abortion in the 1970s, with varying degrees of input from feminists. In Canada, for example, feminists framed abortion as a reproductive right, but the discourse of medicalization was stronger than that of feminism, and the 1969 abortion reform law focused on the concerns of doctors rather than women (Brodie et al. 1992).

Although feminists continued to be concerned with a broad range of reproductive rights, abortion became a major preoccupation after legalization of abortion provoked strong antiabortion countermovements in many countries. In the USA, the single-issue antiabortion movement expanded the conflict to new battlefields such as electoral politics, and the abortion rights movement was forced to engage in single-issue political tactics of its own. In the late 1970s and early 1980s, some feminist reproductive rights activists tried to influence single-issue abortion rights groups to broaden their approach, but they had little success at a time when the very legality of abortion was threatened (Staggenborg 1991). Nevertheless, abortion rights groups helped to maintain the larger women's movement in the 1980s and 1990s, attracting

many young women concerned about losing abortion rights. During periods when the antiabortion movement ebbs, reproductive rights groups have more freedom to pursue a broader range of concerns beyond abortion rights.

In sum, notions of reproductive rights have shifted with the larger socioeconomic and political environment. Nineteenth century voluntary motherhood activists, who lacked economic independence, demanded rights that would strengthen women's positions in families. As women gained more economic freedom with expanding industrialization, movement activists advocated birth control and later abortion rights. The rebirth of feminism in the 1960s inspired an elaborated notion of reproductive rights, which questioned existing medical authority. As black and other nonwhite feminists added their voices to the movement, consciousness of race and class issues, such as sterilization abuse, expanded. Countermovement threats to abortion rights since the 1970s limited the concerns of some movement organizations, but also helped to bring new activists into the reproductive rights movement.

2. *Research on Reproductive Rights*

Social scientific research on reproductive rights has examined reproductive rights movements and countermovements, changes in the medical profession, variations in public policies on abortion, court rulings on abortion, voting patterns of elected representatives, public opinion, demographic profiles of contraceptive and abortion users, public discourse and media coverage of reproductive rights campaigns, and a variety of other issues. Three major categories of research are described here.

2.1 *Social Movement Research*

Analyses of reproductive rights movements have focused on the conditions giving rise to such movements, their evolving organizational structures, strategies and tactics, conflicts with opponents, and outcomes. With regard to the origins of reproductive rights movements, research points to the importance of the political environment as well as available organizational bases and resources. The US abortion rights movement, for example, arose in the context of the protest cycle of the 1960s, in which the successes of the civil rights movement influenced the growth of other movements. Civil rights victories in the courts were certainly one impetus for abortion rights litigation. Other social movements, including both the population movement and the women's movement, provided organizational bases for the abortion rights movement (see Staggenborg 1991).

Beyond explaining the emergence of reproductive rights movements, analysts have examined their ongoing activities and outcomes. Studies of the women's health movement have examined the difficulties involved in establishing and maintaining feminist health clinics, changes in the movement over time, and the extent to which women's health concerns have become institutionalized in current medical practice. In the case of the abortion conflict, analysts have examined interactions between the abortion rights movement and the antiabortion countermovement. Organizational analyses of movement changes over time show that abortion rights movements in various countries have generally grown in strength in response to antiabortion threats.

To explain the intensity of the abortion conflict, many analysts have sought to connect the conflict to twentieth century shifts in values and lifestyles. For those opposed to these changes, abortion symbolizes a whole cluster of threatened values. In *Abortion and the Politics of Motherhood*, Luker (1984) argues that the abortion battle is a clash between two groups of women. Among activists in California, she found that advocates of abortion tended to be educated, career-oriented women who saw motherhood as only one part of their lives, whereas opponents had committed themselves to the traditional female role of wife and mother and felt threatened by its declining status.

Such case studies have been important in revealing the symbolic meanings of abortion for activists and in showing how the conflict has intensified over time. In Toronto, Cuneo (1989) found that the composition of the antiabortion movement changed over time. Early opponents of abortion were liberal Catholics who defended the civil liberties of the fetus and did not oppose changes in gender roles, but they were gradually pushed out of the movement by defenders of traditional family and religious values who regarded abortion as both a symbol and a cause of declining values. In Fargo, North Dakota, Ginsburg (1989) found that early activists on both sides of the abortion conflict actually shared some common ground, including a concern with community and collective responsibility. However, the conflict became more polarized over time, particularly as media coverage encouraged extremism in the antiabortion movement.

2.2 *Analyses of Public Discourse on Reproductive Rights*

Related to research on social movements are analyses of public discourse on reproductive rights, particularly abortion. Condit (1990) examines shifts in public discourse on abortion in the USA from 1960 to 1985, using sources such as newspapers and television shows. She demonstrates the persuasive force of different types of discourse, including value-laden phrases such as the 'right to choice,' narratives about women's experiences with abortion, and characterizations such as the 'back alley' of illegal abortion. Condit describes

how early 1960s narratives about women's experiences with illegal abortion, which portrayed ordinary women needing abortions, aroused public sympathy and made abortion a public concern. These narratives were limiting in that they focused on the needs of 'good' women with 'good' excuses for abortion rather than on the reproductive rights of all women. Nevertheless, they helped to change the public view of abortion as 'wrong'—and in doing so provoked 'prolife heritage tales,' which countered that abortion is always wrong. Later in the 1960s and in the 1970s, the argument broadened as abortion rights activists demanded equality for women and the 'right to choice' while antiabortionists argued for the 'right to life' of the fetus. Condit analyses in detail the social force of the different images and rhetoric used by each side and the social compromises that have been negotiated in different arenas in response to the debate.

Comparative research on abortion politics is needed to show how discursive processes vary with different political cultures and institutional structures. Myra Marx Ferree, William Gamson, Jurgen Gerhards and Dieter Rucht have been comparing the changing influences on abortion discourse in Germany and the USA. Ferree and Gamson (1999) analyse the impact of the organization and discourse of the supranational women's movement, the national women's movements in each country, and the political context of abortion policy in each country. The International Reproductive Rights Research Action Group has found both an international diffusion of feminist discourse on reproductive rights and remaining silences on the subject of female sexuality (Petchesky 1997).

2.3 Research on Public Opinion

An extensive body of literature examines public attitudes toward abortion. Analyses of public opinion surveys in the USA show that approval of abortion increased sharply between 1965 and 1973, when the US Supreme Court legalized abortion, and has remained relatively stable since then. Opposition to abortion is correlated with religiosity and other conservative positions on issues of personal morality, and is also linked, more weakly, to opposition to women's rights (see Cook et al. 1992). European data on trends in values suggests a long-term process of secularization and acceptance of individual discretion in terms of personal morality in Western industrialized countries, which may also explain trends in US attitudes on abortion (Rossi and Sitaraman 1988).

Comparative analyses of abortion attitudes in the USA and Canada reveal strikingly similar patterns in the two countries (see Jelen and Chandler 1994). In both countries, abortion attitudes are remarkably stable, with majorities of the public falling into the middle ground of continued support for abortion under certain circumstances. Nevitte and his colleagues find that support for legal abortion is predicted by similar variables in both countries, including religiosity, 'postmaterialist' values, feminism, and education. Such analyses are important in predicting the future of the abortion conflict. With economic prosperity, increases in postmaterialism and the salience of feminist issues are likely to increase support for reproductive rights, although waning materialist concerns may also increase concerns about abortion among religious fundamentalists who oppose abortion (Nevitte et al. 1993, p. 28).

Other affluent countries exhibit similar patterns. Gallup polls taken in 1981–4 and again in 1990–4 in a number of different countries show that majorities of respondents approve of legal abortion for certain reasons such as threat to the mother's health, and that these attitudes are stable over time (Simon 1998). Although opposition to feminist positions on reproductive rights remains, public opinion in affluent countries is generally supportive of women's access to abortion.

See also: Baby Booms and Baby Busts in the Twentieth Century; Children, Value of; Eugenics as the Basis of Population Policy; Family Size Preferences; Family Theory: Feminist–Economist Critique; Feminist Movements; Feminist Theory: Liberal; Fertility Control: Eastern Europe; Fertility Control: Overview; Gender and Health Care; Gender and Physical Health; Gender and Reproductive Health; Gender and Technology; Gender Role Stress and Health; Gynecological Health: Psychosocial Aspects; Human Rights in Intercultural Discourse: Cultural Concerns; Infanticide and Induced Abortion; Men's Health; Motherhood: Economic Aspects; Poverty and Gender in Affluent Nations; Regulation: Family and Gender; Regulation: Sexual Behavior; Reproductive Medicine: Ethical Aspects; Reproductive Rights in Developing Nations; Sex Preferences in Western Societies; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexuality and Gender; Sexually Transmitted Diseases: Psychosocial Aspects; Teen Sexuality; Teenage Fertility; Women's Health; Women's Suffrage

Bibliography

Banks J A, Banks O 1964 *Feminism and Family Planning in Victorian England*. Schocken Books, New York

Condit C M 1990 *Decoding Abortion Rhetoric: Communicating Social Change*. University of Illinois Press, Urbana, IL

Cook E A, Jelen T G, Wilcox C 1992 *Between Two Absolutes: Public Opinion and the Politics of Abortion*. Westview Press, Boulder, CO

Cuneo M W 1989 *Catholics Against the Church*. University of Toronto Press, Toronto, Canada

Ferree M M, Gamson W A 1999 The gendering of abortion discourse: Assessing global feminist influence in the United States and Germany. In: della Porta D, Kriesi H, Rucht D

(eds.) *Social Movements in a Globalizing World*. St. Martin's Press, New York, pp. 40–56
Francome C 1984 *Abortion Freedom: A Worldwide Movement*. Allen & Unwin, Boston
Gavigan S A M, Brodie J, Jenson J 1992 *The Politics of Abortion*. University of Toronto Press, Toronto, Canada
Ginsburg F D 1989 *Contested Lives: The Abortion Debate in an American Community*. University of California Press, Berkeley, CA
Gordon L 1976 *Woman's Body, Woman's Right: A Social History of Birth Control in America*. Grossman, New York
Jelen T G, Chandler M A (eds.) 1994 *Abortion Politics in the United States and Canada*. Praeger, Westport, CT
Luker K 1984 *Abortion and the Politics of Motherhood*. University of California Press, Berkeley, CA
Nevitte N, Brandon W P, Davis L 1993 The American abortion controversy: Lessons from cross-national evidence. *Politics and the Life Sciences* **12**(1): 19–30
O'Connor J S, Orloff A S, Shaver S 1999 *States, Markets, Families*. Cambridge University Press, Cambridge, UK
Petchesky R P 1984 *Abortion and Woman's Choice: The State, Sexuality and Reproductive Freedom*. Longman, New York
Petchesky R P 1997 Spiraling discourses of reproductive and sexual rights: A post-Beijing assessment of international feminist politics. In: Cohen C J, Jones K B, Tronto J C (eds.) *Transforming Politics*. New York University Press, New York, pp. 569–87
Rossi A S, Sitaraman B 1988 Abortion in context: Historical trends and future changes. *Family Planning Perspectives* **20**(6): 273–301
Ruzek S B 1978 *The Women's Health Movement: Feminist Alternatives to Medical Control*. Praeger, New York
Simon R J. 1998 *Abortion Statutes, Policies, and Public Attitudes the World Over*. Praeger, Westport, CT
Staggenborg S 1991 *The Pro-Choice Movement: Organization and Activism in the Abortion Conflict*. Oxford University Press, New York

S. Staggenborg

Reproductive Rights in Developing Nations

No sharp distinction can be made between affluent and developing nations with respect to the evolution of the concept of reproductive rights. The notion developed through intense dialogues involving feminist communities in Northern and Southern nations. The concept of sexual rights developed concurrently. The two concepts occasioned theoretical controversies within the feminist community and opposition from national and global conservative political forces.

The two concepts are now enshrined in United Nations documents. In 1994, the Program of Action of the International Conference of Population and Development (ICPD Cairo, 1994) defined reproductive rights as:

> [Embracing] certain human rights that are already recognized in national laws, international human rights documents and other consensus documents. These rights rest on the recognition of the basic right of all couples and individuals to decide freely and responsibly the number, spacing and timing of their children and to have the information and means to do so, and the right to attain the highest standard of sexual and reproductive health. It also includes their right to make decisions concerning reproduction free of discrimination, coercion and violence. … (Paragraph 7.3)

A year later, the Platform of Action of the IV World Women Conference (IV WOW Beijing) reaffirmed the ICPD principle and spelled out the substantive contents of women's sexual rights:

> The human rights of women include their right to have control over and decide freely and responsibly on matters related to their sexuality, including sexual and reproductive health, free of coercion, discrimination and violence. Equal relationships between women and men in matters of sexual relations and reproduction, including full respect for the integrity of the person, require mutual respect, consent and shared responsibility for sexual behavior and its consequences.

1. Historical Background

Almost a century of North–South interactions concerning birth control preceded the recent establishment of the concept of reproductive rights. From the early twentieth century, eugenic societies, neo-Malthusians, the socialist movement, and advocates of sexual hygiene had developed connections with colonial territories and Southern independent nations. Beyond these intentional efforts, ideas had been largely spread worldwide in a more incidental way, by labor migrants, colonial functionaries, and progressive political activists.

The expansion of eugenic societies and Malthusian networks has been extensively examined. The biography of Margaret Sanger, the North American birth control advocate, is a privileged subject of analysis. From the late 1920s, Ms. Sanger was involved in international advocacy in association with neo-Malthusian and eugenic networks. Research has revealed the close links between her efforts, colonial administrations, and local ideologies of class, caste, ethnic, and racial domination. These networks would play a major role in the second wave of global advocacy for fertility control in the 1960s and 1970s. By then Ms. Sanger had sponsored research leading to the invention of the contraceptive pill which, together with the first IUDs, were being disseminated in developing nations.

In contrast, the path through which Emma Goldman's socialist views on women's reproductive self-determination traveled south remains largely unexplored. The Cuban experience suggests a resonance far beyond the pioneering initiatives taken by Goldman in the early days of the Soviet Union, as abortion was legalized in 1959, immediately after the revolution. Recent research has also drawn attention

to some of the more progressive ramifications of Ms. Sanger's endeavors. One illustration is provided by Mexico where, in 1916, in Mexico, after a Feminist Congress, a booklet written by Margaret Sanger, '*Birth Regulation*,' was disseminated. In 1925, the same publication was distributed freely by the anti-clerical Calles Administration.

Having originated from contradictory theoretical sources, birth control ideas were given different meanings by diverse interpreters. In India, where the spread of neo-Malthusian ideas coincided with the intensification of the independence struggle, they sparked harsh controversies within the Gandhian movement. In Mexico, contrasting with the climate of the previous decade, a pronatalist General Population Law was approved in 1936. In the same period, mother and child care programs and family benefits were adopted in Brazil, Argentina, Chile, and Uruguay. These initiatives, which were conceived as accompaniments to early industrialization strategies, converged with the Catholic Church's resistance to all variants of birth control advocacy. This climate silenced existing progressive strands and inspired adaptations of conservative eugenic streams. In Brazil, for instance, ideologies advocating racial and ethnic miscegenation to 'whiten' the population gained legitimacy.

This background determined the immediate post-Second World War environment. In Latin America, previous such policies gained leverage from the adoption of import-substitution industrialization strategies. In other regions, the acceleration of decolonization brought the idea that large populations were a positive element in the establishment of new nation states. Rapid economic growth favored a lowering of mortality rates worldwide. However, as industrialized nations entered a second cycle of fertility decline, in developing nations, where demographic transitions were taking place, fertility rates remained high. In the 1950s, Alfred Sauvy, the French demographer, conceived the categories *First*, *Second*, *and Third World*. The central parameters for this demarcation were the demographic differentials between the three zones.

In the late 1960s, fertility control premises became a 'grand ideology,' supported by United Nations agencies, public financial institutions, bilateral donors, private foundations, and the International Planned Parenthood Federation and its national associates. Scientific evidence was collected to prove that rapid demographic growth had negative impacts on the economic development of Southern countries. Extreme positions went as far as to argue that population control would have the additional benefit of preventing revolutionary uprisings across the South.

Although pressured from outside, developing nations were not easily or immediately persuaded. Indeed, at the Population Conference of Bucharest of 1974, they counteracted by proposing a policy agenda that combined reform of the international economic order, domestic economic growth and investments in basic needs. This position became crystallized in the lemma: 'Development is the best contraceptive.' Yet, not all developing nations remained faithful to their Bucharest commitment. By the late 1970s, draconian population control was already in place in Asia, particularly in China and India. In the other regions countries resisted longer.

Feminist initiatives were also blossoming south of the Equator. Feminist groups initially supported the Bucharest critique of population control. Gradually, however, they came to perceive that women's needs and rights were not being addressed by either pronatalist or antinatalist programs. In some settings, women's fertility was targeted. In others, women's ability to freely decide on reproductive matters was curtailed by pronatalist ideologies. In a few countries the two approaches were implemented in combination, the racial and ethnic origin of the client being the determining factor for selecting which approach was used for whom. Overall, the quality of information and care provided by both MCH and family planning programs was extremely poor.

The early feminist critique of population policies was preceded and surrounded by other relevant institutional processes. The 1968 International Human Rights Conference, held in Tehran, approved a principle ensuring couples the '*basic human right to decide freely and responsibly on the number and spacing of their children and a right to adequate education and information in this respect*.' In Bucharest, it was rephrased to extend these rights to individuals (not just couples) and to ensure access to the means of exercising them. Nevertheless, until the 1990s, individual rights in regard to reproductive matters remained subordinated to the primacy of preventing the population crisis.

Bringing gender concerns into family planning frameworks evolved along with other efforts aimed at women's equality. In 1975, the First United Nations sponsored International Women's Conference was held in Mexico City. Coercive measures of population control were debated and the final document proclaimed respect for the bodily integrity of women as '*a fundamental element of human dignity and freedom*.' This was followed by the approval of the Convention on the Elimination of All Forms of Violence against Women (CEDAW) in 1979. CEDAW affirmed women's rights to family planning, their equal right (with men) to decide on the number and spacing of their children, the right to maintain their jobs while pregnant, and to have access to maternity leave and child-care. In 1981, a UN Symposium on Population and Human Rights held in Vienna also declared that '*Both the compulsory use of abortion and its unqualified prohibition would be a serious violation of human rights*.'

Despite this emerging consensus, the climate of debates began to change. In the International Population Conference of Mexico in 1984, developing

countries expressed a broader acceptance of family planning programs and reaffirmed women's right of access to contraception. However, an alliance between the Reagan Administration and the Roman Catholic Pope made the subject of abortion a major controversy. The United States cut funding for international family planning activities, saying that the programs facilitated access to abortion. (This US policy remained in place until 1993.) Some family planning organizations openly contested this US policy and disseminated information about it among their Southern partners. Abortion was openly debated at the NGO Forum of the Women's Conference of 1985 in Nairobi. In 1988, the meeting of the International Federation of Gynecology and Obstetrics in Rio de Janeiro was preceded by the Christopher Tietze International Symposium on Women's Health in the Third World, a critical event in addressing abortion rights internationally. By the end of the decade, the World Health Organization adopted the notion of reproductive health that would be rapidly incorporated by other institutions and advocacy groups becoming an important conceptual adjunct to the idea of reproductive rights.

Other contributions to the conceptualization of reproductive rights came from women's organizations in Southern nations. Though small-scale, these grassroots organizations engaged in constant dialogue with their societies to transform sexist attitudes and practices. Sometimes they influenced legislation and policy to enhance women's health. In all contexts, feminist organizations had to confront the powerful influence of international population policies or the resistance of pronatalist ideologies.

This blossoming of feminist advocacy coincided with the first ripples of globalization. The debates evolving in industrialized countries filtered into Southern societies through various communications streams. Feminist publications such as '*Our Bodies, Ourselves*' were read in English speaking circles. Political exiles brought back home the stories of feminist struggles in relation to reproduction and sexuality and adapted innovative consciousness raising strategies, as in the case of gynecological self-examination. The mainstream media provided some information.

Local groups built connections with each other and with the Committee for Abortion Rights and Against Sterilization Abuse (CARASA) in the United States, and the International Campaign on Abortion, Sterilization and Contraception (ICASC) in Europe, which had spawned the very first formulation of reproductive rights. In 1984, a first Latin American Women's Health Meeting was held in Colombia. The International Campaign on Abortion, Sterilization, and Contraception also hosted, in Amsterdam, the first International Reproductive Rights Conference.

In Amsterdam, the International Campaign was renamed the Women's Global Network for Reproductive Rights (WGNRR) and given a broad mandate to address all issues related to women's reproductive health, not only those involving abortion, sterilization, and contraception. Reproductive rights were defined as '*women's right to decide whether, when and how to have children—regardless of nationality, class, race, age, religion, disability, sexuality, or marital status—in the social, economic and political conditions that make such decisions possible*.'

The next WGNRR conference held in San Jose, Costa Rica (1987) started an international campaign on maternal mortality and inspired the creation of the Latin American Campaign for the Decriminalization of Abortion. In 1988, in a feminist meeting preceding the Tietze Symposium in Rio, the debates concentrated on ethical aspects of population control policies, abortion, and new reproductive technologies. In Manila, in 1990, during another WGRRN Women's Health Conference, the idea was raised that the international reproductive rights movement should conceptualize a feminist population policy. As globalization intensified, a clear acceleration occurred in the circulation of these ideas.

The United Nations Conferences of the 1990s were conceived as an interconnected series of global debates starting with Environment and Development (Rio 1992) and moving forward to address Human Rights (Vienna 1993), Population and Development (Cairo 1994), Social Development (Copenhagen 1995), Women (Beijing, 1995), Human Settlements (Istanbul, 1996), and Food Security (Rome, 1996). Whereas in the 1970s and 1980s, Southern feminist intellectual and political endeavors evolved largely in isolation from international policy arenas, in the 1990s, 'they involved' meant direct engagement in these negotiations. In Rio, access to reproductive health was added to conventional family planning language. Vienna's definitions with respect to women's human rights affirmed that they applied to both the public and private realms, thereby creating the basis for the further developments of Cairo and Beijing regarding reproductive and sexual rights.

2. *The Contents of Reproductive and Sexual Rights in the Southern Feminist Perspective*

The adoption of a reproductive rights framework by advocacy groups within developing nations entailed a broadening of its original focus—abortion, contraception, and sterilization—to include maternal mortality, prenatal and obstetric care, breast and cervical cancer, and adolescent and postmenopause health. The Southern perspective also underlined the need to incorporate attention to women's economically productive roles in addition to their biologically reproductive functions. This required a better understanding of the correlation between sexual and reproductive

health and endemic and chronic diseases, as well as occupational, environmental and mental health hazards, and afflictions.

In the 1980s, the scope and intensity of abortion-related initiatives varied across regions, being more visible in Latin America (particularly Brazil), the Philippines, and South Africa. Abortion was always addressed as part of a broader reproductive health and rights agenda, never as a 'single issue.' Efforts to decrease maternal mortality were prioritized by most countries and within these efforts new venues for addressing abortion have emerged. Both ICPD and IV WCW final documents recognize abortion as a major public health problem, affirming that, in circumstances where it is legal, it should be safe. The Beijing Platform of Action recommends that countries review laws containing punitive measures against women who have undergone abortion.

Contraception has always been a cross-regional concern, often involving United States and European advocacy efforts. Research has been devoted to sterilization, IUDs, injectables and implants, methods privileged by fertility control programs. Long before the feminization of the HIV-AIDS pandemic, Southern feminist organizations emphasized the disease-preventing advantages of barrier methods, such as the diaphragm.

In the 1990s, campaigns were launched against specific technologies: Norplant, the contraceptive vaccine, and Quinacrine (a method of chemical sterilization). Advocates engaged in dialogues with researchers, pressuring for better ethical standards and redefined research priorities. One critical outcome of these efforts was the investments aimed at developing microbicides to protect women against HIV-AIDS and pregnancy. Concern arose regarding the expanding use of prenatal sex-selection techniques in India and China.

In their efforts to address reproduction as a right, not merely a woman's obligation, Southern advocacy groups very early challenged gender inequalities in the sexual domain. Issues such as early marriages, rape, forced prostitution, trafficking, and female genital mutilation have been raised under the umbrella of violence against women. Sexuality, as such, however, gained enormous visibility under the impact of the HIV-AIDS pandemic. The inadequacy of contraceptive methods used in most Southern countries became clear, as they offered no protection against infections. Research in Africa showed the association between HIV and other sexually transmitted infections.

The World Health Organization coined the notion of sexual health, and initiated preventive strategies that required a better understanding of sexual behavior. Most Southern governments resisted recognizing the impact of the pandemic, partly because it was initially portrayed as affecting only certain risk groups (homosexuals, sex workers, and drug addicts). Among Southern feminists, some time elapsed before the feminization of HIV-AIDS became a priority. An important recent political development is the multiplication of gay and lesbian rights advocacy groups in developing nations. The concept of sexual rights presently encompasses nondiscrimination on the basis of sexual orientation or HIV-AIDS status. Other ideas have diffused that promote responsibility by men in sexual matters and transforming notions of masculinity.

Thus, although the emphases do vary, the content of reproductive and sexual politics within affluent and developing nations are similar. The concepts exemplify an 'ideascape' which, having broken through national and cultural boundaries, has gradually crystallized in a globally accepted discourse. These ideas are present in Southern national and regional advocacy groups and among Southern academic and research initiatives, particularly in Latin America and South and South East Asia.

3. *Conceptual Debates*

A number of inter related controversies developed as reproductive and sexual rights were conceptualized in developing nations. A first one can be retraced back to the global debates around abortion in the 1980s. In developing countries, should abortion be addressed as a single issue, as in the United States advocacy strategies, or part of a comprehensive reproductive health and rights agenda? A theoretical and political problem underlay this early controversy. In affluent societies, where women's basic needs had been largely met, restrictions in the access to abortion strikingly expresses gender inequality with respect to the individual right to decide. However, women's exercise of reproductive rights in developing regions (and among marginalized groups in affluent societies) is primarily curtailed by economic disadvantages that affect men and women.

A related controversy arose from the critique of the intrinsic individualistic and contractual nature of the Western concept of rights. Feminist thinkers recognized that any approach to sexual and reproductive politics must take into account political, economic, and cultural constraints on individuals. The full exercise of these rights is dependent upon an enabling environment including democratic conditions that allow citizenship rights, women's empowerment, and material support, such as transportation, childcare, jobs, and education. Examples of the importance of such support are easily found. In all developing nations, women die because they do not have transportation or because basic equipment is lacking in health facilities. In rural India and China, vaginal infection results from precarious sanitation. Poor women die of botched abortions, while rich women can pay for safe procedures. In India, abortion has been legal since the 1970s, but the quality of care

remains poor. Young women are infected by HIV-AIDS by having sex to get money to survive.

The emphasis on enabling conditions does not imply ignoring gender inequality and how it restricts women's ability to protect their bodily integrity and personhood. Research that explored ordinary women's perceptions of sexual and reproductive rights found that few women use these terms, but they do understand that the principles refer to being able to make decisions about procreation, to be free from harassment by men, to protect themselves against domestic violence and preventable illness, and to have access to adequate health care. When these rights are not in place, women cannot exercise civil and political rights. The framework implies both positive rights, such as reproductive health care and employment, and negative rights, such as protection against rape and discrimination, regardless of sexual orientation or marital status.

Some have criticized the emphasis on sexual and reproductive health as being too narrow when compared with the broad women's health agenda defined in the 1980s. To address this limitation requires understanding links between gender, social inequality, and health. One priority is to ensure that health reform policies improve women's access to health services. Another concern is to understand how men's violence against women affects women's health and mount policies to confront this. Exploring this topic may also lead to addressing the adverse effects on women of cultural constructs of masculinity.

4. *Policy and Political Challenges*

A human rights approach to sexual and reproductive politics still faces major obstacles to effective implementation. The definitions of related rights as enshrined in United Nations documents are not binding. Even if these new principles are translated into binding instruments, international human rights mechanisms have little power to enforce them, especially if violators are private parties rather than states.

Additionally, the effective implementation of women's reproductive and sexual rights must challenge a rigid and persistent 'public'–'private' distinction in international human rights law and national legal systems. This breaking up of boundaries is crucial for women, whose lives in many cultures are still locked within domesticity. To help women, human rights claims must penetrate the 'private sphere' within which marital rape, female genital mutilation, virginity codes, and repudiation of birth control occur. On the other hand, public actions of the state and its agents also very often violate women's sexual and reproductive self-determination.

Greater challenges may lurk in globalization. Poorly understood and poorly regulated, globalization is creating growing inequalities of income within and across countries. Women are entering labor markets everywhere. Earning more income often raises a woman's value, but can also result in even tighter controls over her life. Earnings sometimes provide women with greater physical mobility, increased personal autonomy, and the possibility of breaking through gender barriers and patriarchal or other mechanisms of male control.

Concurrently—in reaction to the negative effects of globalization—political, national, religion-based, ethnic, or other identities are being strengthened which revive traditional gender systems and controls. There is a proliferation of both antiabortion and 'profamily' groups associated with religious fundamentalists—Catholic, Protestant, Islamic—and of conservative politics that systematically oppose women's reproductive and sexual freedom on behalf of 'fetal rights,' womanly duty, and the defense of hierarchical power within the family. These forces are at play in the United Nations, and other global policy arenas, whenever gender related issues arise. Moreover, racism, nationalism, and ethnic hatreds also pose threats to women's reproductive and sexual health. Chronic violence, genocide, 'ethnic cleansing,' and military occupation have often meant mass rape, abandonment, refugee status, increased vulnerability to forced prostitution, trafficking, unwanted pregnancy, and sexually transmitted diseases.

The promoters of a globalized economy often support the breaking down of traditional gender systems and greater freedom for women. Nevertheless, they remain largely unconcerned with growing economic inequality which, in most cases, lies at the root of fundamentalist trends, patriarchal revivals, and ethnic strikes. For developing countries, globalization has led to large-scale debt and trade inequities. Fiscal austerity, imposed through structural adjustment programs, has negatively affected the availability and quality of public social services.

The fulfillment of the twentieth century promise of women's sexual and reproductive freedom requires these contradictory trends to be challenged and eventually tamed. Redefining development policy priorities to create an enabling environment that will both provide a firm basis for individual self-determination while restraining the appeal of moral conservative forces remains a daunting task for the early years of the new millennium. The historical experience in both the South and the North also suggests that such fulfillment additionally demands sustained intellectual and political efforts aimed at promoting and enhancing the acceptance of all forms of diversity and tolerance within and across societies.

See also: East Asian Studies: Gender; Economic Development and Women; Economic Globalization and Gender; Eugenics as the Basis of Population Policy; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Per-

spectives; Family Planning Programs: Quality of Care; Family Theory: Economics of Intergenerational Relations; Family Theory: Economics of Childbearing; Family Theory: Economics of Marriage and Divorce; Family Theory: Feminist–Economist Critique; Female Genital Mutilation; Feminist Movements; Fertility and Culture: Anthropological Insights; Fertility Change: Quantum and Tempo; Fertility Control: Overview; Fertility: Political and Political–Economic Perspectives; Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Middle East and North Africa; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Fertility Trends in the Formerly Socialist Countries of Europe; Gender and Health Care; Gender and Physical Health; Gender and Reproductive Health; Gender and Technology; Gender Role Stress and Health; Gynecological Health: Psychosocial Aspects; Health in Developing Countries: Cultural Concerns; HIV and Fertility; Infanticide and Induced Abortion; Labor Markets, Labor Movements, and Gender in Developing Nations; Land Rights and Gender; Latin American Studies: Gender; Lone Mothers in Nations of the South; Male Dominance; Masculinities and Femininities; Men's Health; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Nation-states, Nationalism, and Gender; Near Middle East/North African Studies: Gender; Population Policy: International; Poverty and Gender in Developing Nations; Regulation: Sexual Behavior; Reproductive Medicine: Ethical Aspects; Reproductive Rights in Affluent Nations; Rural Industrialization in Developing Nations and Gender; Science and Industry; Southeast Asian Studies: Gender; Women's Health

Bibliography

Blatliwala S L 1994 The meaning of women's empowerment: new concepts from action. In: Germain G S A, Chen L C (eds.) *Population Policies Reconsidered: Health, Empowerment and Rights*. Harvard Center for Population and Development Studies, Boston

Bunch C 1993 *Organizing for Women's Human Rights Globally, in Ours by Right: Women's Rights as Human Rights*. Zed Books, London

Chesler E 1992 *Woman of Valor: Margaret Sanger and the Birth Control Movement in America*. Simon & Schuster, New York

Cook R 1995 *Human Rights and Reproductive Self-Determination*, Washington, DC

Corrêa S, Petchesky R 1994 Reproductive Rights: A Feminist Perspective. In: Sen G, Germain A, Chen L C (eds.) *Population Policies Reconsidered: Health, Empowerment and Rights*. Harvard Center for Population and Development Studies, Boston

Corrêa S, Reichmann R 1994 *Population and Reproductive Rights: Feminist Perspectives from the South*. Zed Books, London

Dixon-Mueller R 1993 *Population Policy and Women's Rights: Transforming Reproductive Choice*. Praeger, Westport, CT

Hartmann B 1995 *Reproductive Rights and Wrongs*. South End Press, Boston

International Planned Parenthood Federation 1995 *IPPF Charter on Sexual and Reproductive Rights*. International Planned Parenthood Federation, London

Petchesky R P 1995 From population control to reproductive rights: feminist fault lines. In: *Reproductive Health Matters*

Petchesky R P, Judd K (eds.) 1998 *Negotiating Reproductive Rights: Women's Perspectives Across Countries and Cultures*. International Reproductive Rights Research Action Group, London and New York

Peters J, Wolpe A (eds.) 1995 *Women's Rights, Human Rights: International Feminist Perspectives*. Routledge, New York and London

United Nations 1994 *Programme of Action of the International Conference on Population and Development*. United Nations, New York

United Nations 1996 *Beijing Declaration and Platform for Action, adopted by the Fourth World Conference on Women*. United Nations, New York

S. Corrêa

Republic

'Republic' comes from the Latin *res publica*, the shared or common thing. It first described the political regime instituted by the Romans after they deposed their kings in 509 BC. The founding of the Republic symbolized the rejection of rulers who treated the political community as their private property. The word *publica* probably derives from the Latin *populus*, 'people.' Cicero in the first century BC calls the *res publica* a *res populi*, a thing that belongs to the people. That the political community belongs to the people rather than those who rule reverberates throughout the long history of the term as it takes on and sheds concepts drawn from the Roman experience.

1. The Roman Legacy

Polybius, in the second century BC, described (not entirely accurately) the Roman Republic as a 'mixed regime' where monarchy, aristocracy, and democracy balanced one another, providing stability and protection from the constant cycling of regimes from one form to another. Sallust, writing during the political chaos of first century BC Rome, turned away from institutional analyses to character and emphasized the foundation of the Republic on political virtue and traditional moral values, qualities whose loss would bring about the demise of the Republic. The civil wars ended with the Principate of Augustus which returned Rome to one-man rule. Tacitus, writing in the early

years of the Principate, recalled the liberty of the Republic in contrast to the tyranny of the increasingly despotic Roman emperors. All these elements, the mixed regime, political and moral virtue, and liberty as the absence of the oppressive rule of a king have followed the word 'republic,' though neither consistently nor necessarily in unison. 'Republic' recalls Rome, but the Greek city-states, different in structure and size, also came to be called republics, and Athens and Sparta, along with Rome, became models of republican regimes for later ages.

During the Middle Ages, *respublica* simply meant the political community in contrast to the church; it was the body of which the prince was the head, the wife to whom the prince was joined in matrimony. It incorporated none of the elements of liberty, institutional balance, or political virtue that connected the idea of republic back to the experiences of Rome. At the close of the medieval period, however, republic recaptured the connotations of liberty understood as hostility to tyrannical rule. The early Italian renaissance witnessed a shift from medieval praise of the Principate as a reflection of God's rule to admiration for the political liberty of the Roman Republic. A number of the Italian city-states of the Renaissance gained political independence and instituted republican regimes of self-rule. Meanwhile, humanist writers of the early Renaissance, recalling republican Rome, appealed to the liberty and political virtues of antiquity against the faith in a God-willed monarchy that dominated the medieval period. Nevertheless, tensions arose between republican virtues which fostered renown for service to the secular community and religious virtues which glorified God. With Machiavelli in the early 1500s, concern about devotion to and participation in a secularized polity disappeared.

2. *Machiavellian Republicanism*

Machiavelli's begins *The Prince* by rejecting the conventional Aristotelian typology of monarchy, aristocracy, and democracy; he boldly stated that all dominions 'now and in the past' have been 'either republics or principalities.' Machiavelli's dichotomies never persevere throughout his writings, but *The Prince* largely treats principalities and his *Discourses on the First Ten Books of Livy* (1513) discusses republics with special attention to Rome. Other republics like Sparta and Venice surface as well offering variations and qualifications on the Roman model. More nuanced than works by Renaissance writers who simply identified republics with liberty, or sought to find a republican heritage for their cities in the history of the Rome, Machiavelli identifies the tensions inherent in republican regimes, their occasional need for dictatorships, their dependence on conflict between parties as a condition for freedom, and their glory in imperialist expansion. Machiavelli, along with republican authors of the period, praised the mixed regime, but he saw it as necessary for a regime to order the conflicts that plague any free polity. Further, while praising republican virtues of devotion to and participation in the political regime, and while worrying about the effects of a corrupt self-interested citizenry, he also recognized that republican virtues entailed the conquest of others.

Despite Machiavelli's somewhat ambiguous presentation of a republican Rome that needed occasional dictators and men like Brutus who were willing to sacrifice their sons for the preservation of the regime, his *Discourses*, with its reliance on Polybius' mixed regime, became a primary source for antiroyalist, republican literature of seventeenth century England. James Harrington, for instance, begins his utopian republican tract *Oceana* praising Machiavelli as the sole retriever of ancient prudence concerning governments devoted to the public interest. In addition to praising economic equality as the foundation of a peaceful regime, Harrington continued the republican theme of the mixed government by proposing a bicameral parliament where an elected and rotating aristocratic senate would be checked by a popular assembly. As with the other republican authors of his time, Harrington relied on ancient Sparta and contemporary Venice along with Rome as models. All demonstrated the viability of the republican model with their mixed regimes, and all were frequently paraded as exemplars of popular engagement and liberty in contrast to monarchical oppression.

The participatory expectations meant that seventeenth and eighteenth century English authors often associated republicanism with the aristocratic leisure essential for an engagement in political life. The contemporaneous emergence of liberal thought based on the rational pursuit of self-interest set up a conflict between an elitist republic relying on men of public virtue with resources and leisure for public life and the egalitarian principles of a liberal society focused on rights, consent (a much weaker political act than actual engagement), and the absence of oppression. The liberal perspective did not require self-sacrifice; the liberal individual, unlike the republican, pursuing private interests could be oblivious to the public good. Liberal virtues of frugality and moderation supporting economic success were not the republican virtues that supported the moral and political life of a nation.

3. *Enlightenment Republican Thought*

The English republicans, drawing on the Polybian mixed regime, articulated a theory that allowed for the presence of a monarch, albeit one limited by assorted checks. Debates arose as to whether the checks were sufficient or too strong, but a monarchical regime that might allow for liberty contrasted starkly with the absolutism of eighteenth century France.

Montesquieu's *Spirit of the Laws* appeared in this latter context. Montesquieu replaced the still dominant Aristotelian typology with despotism, monarchy, and republic; the last he divides into the aristocratic republic, marked by the love of honor, and the democratic republic, characterized by 'political virtue' or the 'love of equality.' Drawing on themes of civic virtue from Sparta, the Roman Republic, and the *Discourses* of the 'great man' Machiavelli, Montesquieu writes that political virtue in a democratic republic means the denial of self for the sake of the whole. According to Montesquieu, the democratic republic would be possible only in a small, homogenous community where there was no attention to individual economic success nor worries about economic inequalities. The simplicity of the democratic republic would allow for none of the conflicts that arise with the emergence of the wealthy few. Montesquieu's democratic republic depended on political virtue, but it did not offer much happiness. In contrast to the democratic republic, Montesquieu wrote about a free and happy people who inhabited a large commercial nation, namely England, that with its 'mixed regime' could incorporate variety, economic activity, and inequality in a way that a democratic republic could not and that could offer liberty through the just execution and impartiality of its laws.

Montesquieu's praise of the English constitution suggested freedom would be possible in large, heterogeneous regimes, if laws were well executed and a proper balance between executive and legislative powers and within the legislative branch existed. Rousseau's *The Social Contract*, in contrast, affirmed that freedom was possible only in small homogenous republics where all citizens, as members of the sovereign body, participated in the life of the city through law making. A large city where citizens devote themselves to commerce, crafts, and comforts rather than to political life undermines attachment to the community. Free regimes, according to Rousseau, depend on the political engagement of all in self-rule; we are slaves if do not legislate for ourselves. Rousseau returned to Rome for his model of participatory republican life where, in addition to popular assemblies, there was a censorial tribunal ensuring the purity of opinions and morals. Rousseau's republic depended on a similar control of citizens' opinions and required a civil religion to instill 'sentiments of sociability.' A decline in morals would bring about political decay and the loss of the self-rule at the heart of free republics. Preservation of morals and attachment to the city for Rousseau rested on maintaining a small state, unperturbed by the bustle of economic activity or the cosmopolitanism of Parisian culture.

4. *The Issue of Size*

The question of whether republican liberty was possible in large nations or only in small ones was central to the debates about ratification of the US Constitution. *The Federalist Papers*, arguing for adoption of the constitution, drew upon the equation between a republic and liberty. Using Publius as their common pen name, the authors of *The Federalist Papers* identified with a founder of the Roman Republic who (in a history their audience would have known) helped overthrow the Roman kings. *The Federalist Papers* famously argue for a republic, not a democracy. The republican principle affirmed: 'The people are the only legitimate fountain of power,' but when a multitude of people exercise legislative functions in person, the ambitious intrigues of the executive magistrates lead to tyranny. Freedom in a republic does not require that the people themselves exercise legislative authority directly. The power that rests in the people must be filtered. Liberty may require that all power derive from the people and that those entrusted with power remain dependent on the people, but stability requires the opposite. Admitting that the petty 'republics of Greece and Italy' evoke 'sensations of horror and disgust' at the rapid succession of revolutions, the defenders of the constitution claim that it incorporates newly discovered maxims derived from political science that apply 'republican remedies' to the 'diseases most incident to republican government': namely, representation, legislative checks and balances, and the enlargement of the sphere of government. The new republic does not rely only on the virtue of its citizens, but on institutions that protect against oppression, afford liberty for its citizens, and ensure the excellence of those elected to office.

The Anti-Federalists arguing against the new constitution feared that the large centralized government necessary for a populous and heterogeneous nation would oppress its citizens, undermine any attachment they felt for one another, and stifle self-rule. Consent, critical for a free republic, they argued, cannot be represented; it must come from the individual citizen. A 'representative republic' is a contradiction in terms. Lost would be the community of like individuals; citizens would be left to nurture vices of self-interest rather than cultivate republican public virtue. A centralized government would fall into the hands of the wealthy few transforming the new republic into an aristocracy or even a monarchy. The Anti-Federalists emphasizing a small, homogenous polity of virtuous citizens were the last great defenders of the classical republic.

5. *Republic Today*

Though *The Federalist Papers* distinguish democracies with their demagogues and upheavals from orderly republics, democracy and republic have over the nineteenth and twentieth centuries converged conceptually. Both identify the body of citizens as the fount

of political authority and the overlap between the two regimes has obscured distinctive political structures that might limit recollections of Roman institutions only to republics. With the absence, indeed impossibility, of direct democracy, both democracies and republics have taken on representative forms of government and most democracies have adopted variations of the Polybian mixed regime. Nevertheless, a return to republican ideals of political virtue and engagement has surfaced in an effort to address what is portrayed as the excessive individualism of citizens in liberal democracies. Among contemporary communitarians, ancient republics are symbols of a lost political world where citizens found fulfillment in political life, not in the pursuit of economic success. This evocation of ancient republics has revived discussion about the role of republican ideals during the founding of the USA with communitarians eager to find in the early years republican rather than liberal roots. The concept of the republic to which they appeal goes well beyond the people as the fount of political authority and the mixed regime to the intangibles of political virtue and a shared life of the ancient republic.

See also: Citizen Participation; Civic Culture; Civil Society/Public Sphere, History of the Concept; Democracy; Democracy, History of; Democratic Theory; Participation: Political; Republicanism: Impact on Social Thought; Republicanism: Philosophical Aspects

Bibliography

Fink Z S 1945 *The Classical Republicans*. Northwestern University Press, Evanston, IL

Kramnick I 1990 *Republicanism and Bourgeois Radicalism*. Cornell University Press, Ithaca, NY

Machiavelli N 1998 *The Prince*. [trans. Mansfield H] Chicago University Press, Chicago

Pangle T L 1988 *The Spirit of Modern Republicanism*. University of Chicago Press, Chicago, IL

Pocock J G A 1975 *The Machiavellian Moment*. Princeton University Press, Princeton, NJ

Rahe P A 1992 *Republics Ancient and Modern*. University of North Carolina Press, Chapel Hill, NC

Wood G S 1969 *The Creation of the American Republic, 1776–1787*. University of North Carolina Press, Chapel Hill, NC

A. W. Saxonhouse

Republican Party

The Republican Party is one of the two major political parties in the United States that have dominated American politics since the mid-1850s. The Republican party (also known as the Grand Old Party—GOP) emerged amid the sectional conflicts of the pre-Civil War era. Throughout most of its lifetime the party's electoral strength was centered in northern and western states. It was not until the 1960s with the realignment of southern whites toward the Republicans that the party was able to compete effectively in all sections of the country. The Republicans are the right of center party of American politics. It has a broad, multiclass base of support, which tends to be strongest among white, middle, and upper middle-class citizens. Since the 1980s its activist base has become increasingly conservative. Republican national level organizations have become well financed and professionalized and as a result intraparty influence has shifted toward the national organization.

1. Historical Development

The Republican Party emerged amid the conflicts of the 1850s as the Democratic and Whig Parties were unable to reconcile differences within their ranks over the slavery issue and the growing economic and cultural differences between the North and South. The immediate reason for the emergence of the Republican Party was passage of the Kansas–Nebraska Act of 1854, which overturned the ban on slavery in the Northwest Territory. This action by Congress touched off antislavery protest meetings that led to the creation of the party in 1854. In its first election, the Republicans had unusual success winning approximately 40 seats in the House of Representatives (Jones 1965, Moos 1956). In 1856, the Republicans established themselves as the principal opposition party to the then dominant Democrats when the Republican presidential nominee ran a strong second in the popular vote. Abraham Lincoln, formerly a Whig, became the first Republican president after the election of 1860 as the Democratic and Whig parties splintered along a North–South axis. In the Civil War that followed Lincoln's election, the Republicans were the party of national unity and freedom for slaves.

Despite occasional losses, the Republicans were dominant in national politics from the 1860s until the 1930s. It was, however, a sectionally based party of the North and West as the legacy of the Civil War and Reconstruction made it noncompetitive in southern states. The extent of Republican electoral domination during this period is demonstrated by its control of the presidency for 56 of the 72 years between 1860 and 1932. Despite Democratic control of congressional seats in the South, the Republicans also tended to dominate the Congress—controlling the House in 23 sessions and the Senate in 31 sessions out of a total of 36 (Jones 1965).

The Great Depression of the 1930s ended Republican dominance as the party was blamed for economic hard times and the Democrats achieved electoral

hegemony. The GOP's only presidential victories between 1932 and 1964 came in 1952 and 1956 when it nominated Dwight D. Eisenhower, a World War II hero, whose appeal was above party.

2. *The Republican Party in the Post-New Deal Era*

After a disastrous defeat in 1964, Republican fortunes improved although the party continued to have fewer party identifiers among the voters than the Democrats. Benefiting from deep schisms among Democrats over the Vietnam war, civil rights, and social policy, the GOP regained the presidency with Richard Nixon's 1968 election and 1972 reelection. Evidence of the party's continuing minority status can be found in the fact that in these elections it still was unable to win control of Congress and most state governments.

The resurgence of the Republicans after 1964 was based upon fundamental electoral changes as politics became less party-centered and more candidate-centered. There were also significant electoral realignments (Aldrich 1995). Particularly significant was the movement of southern whites toward the Republicans and away from the Democrats. Beck (1999) estimates that white southerners constituted 30 percent of the GOP's base electorate in 1996. With a base of support in southern and western states, the Republicans won three presidential elections in the 1980s and in the 1994 midterm congressional elections gained control of both houses of the Congress for the first time in 40 years.

The transformation of the Republicans from a party which for most of its history drew its electoral support almost exclusively from northern and western states into one with a southern base is demonstrated by the fact that after the 1994 election, the Republicans held a majority of the seats in the House and Senate from the South. Southerners constituted the largest regional group among both House and Senate Republicans. At the same time, Republican representation from the Northeast, once a bastion of party support, declined substantially so that northeasterners were the smallest regional group among House Republicans in the 1990s.

In addition to the shift in the partisanship of white southerners, other changes in the composition of the Republican party's electoral base have occurred since 1968. Whereas women traditionally outnumbered men, by the 1990s they comprised a minority in the party and as a result the party became less sensitive to the concerns of moderate and liberal females. There was also a change in the religious composition of the party. Support among mainline and nonfundamentalist Protestants has eroded steadily since the mid-1960s when they comprised 60 percent of Republican identifiers to slightly more than one-third of party loyalists in 1996. These shifts in the demographic characteristics of Republican identifiers have been accompanied by an ideological change as the party's loyalists have become increasingly conservative. Beck's (1999) analysis revealed that whereas a majority of Republican identifiers were self-declared conservatives in 1980, two-thirds carried that designation in 1996.

Changes in the composition of the Republican electorate and the heightened conservative ideological commitment of Republican loyalists have had significant implications for the party. Conservative ideological interests have achieved greater influence within the party organization, among elected officials, most notably in the House of Representatives, and over nominations. There is also substantial internal division within the party between committed conservative ideologues, who tend to dominate the many party organizations as well as the House Republicans, and the more pragmatic and moderate Republicans exemplified by the party's governors.

3. *Republican Party Organization*

The party organization is a layered structure of national, state, and local units rather than a hierarchy controlled from the top down. The Republicans' organizational resources, however, involve more than its formal, legally constituted units. There are the thousands of personal campaign organizations created by Republican candidates, an array of partisan campaign consultants and allied interest groups which provide assistance in voter mobilization, money, staff, technical assistance, and media advertising (Bibby 1999).

3.1 *The Republican National Committee*

The Rules of the Republican Party, which are adopted at each quadrennial national convention, call for the creation of a Republican National Committee (RNC) which is charged with responsibility for the 'general management of the party, subject to direction from the national convention.' This authorization makes the RNC the official governing body of the party between conventions. Since its creation in 1856, the RNC has evolved from an organization in which functions were sketchily differentiated and performed by party notables every 4 years into a continuous operation with a large professional staff and budget (Cotter and Bibby 1980).

The RNC resisted reform efforts in the 1970s and has maintained a confederate organizational mode in which each state, plus the District of Columbia, Puerto Rico, American Samoa, Guam, and Virgin Islands have equal representation through a national committeeman and woman and a state chairman. RNC members are selected through procedures mandated by state law and state party rules. Since the full RNC

meets only twice a year, responsibility for operations of the committee rests primarily with the national chairman and senior staff. When the party controls the presidency, the national chairman is normally designated by the President and works in close conjunction with the White House staff.

Since the 1960s, the RNC has developed the capacity to raise massive amounts of money to support party activities and candidates. Thus, in the 1999–2000 election cycle, the RNC raised $205.3 million, plus $156.2 million in 'soft money' (money raised outside the restrictions of the Federal Election Campaign Act).

With this level of financial resources, the RNC has achieved a degree of autonomy that could not have been imagined in the 1940s and 1950s when it was dependent upon contributions from state affiliates. Through an extensive program of financial and technical assistance to state and local parties, it has gained intraparty leverage because RNC aid is given with requirements attached that the state units must meet. Through this process, Republican state organizations have been professionalized and gained enhanced candidate-support capabilities, but they have also become increasingly dependent upon the RNC for the resources essential for organizational maintenance and effective performance in state level campaigns.

The RNC has also used its financial resources to gain an unprecedented level of intraparty integration in implementing national campaign strategies. For example, in the 1999–2000 election cycle the RNC transferred $93.0 million in 'soft money' to state and local party organizations for spending on general party overhead, voter registration, get-out-the-vote drives, and state and local campaigns. The RNC allocated these funds among the state parties in a manner designed to implement a national strategy geared to winning critical states in the presidential race and maximizing the party's seats in the House and Senate. Through its programs of technical assistance and fund transfers the national party organization has been able to nationalize campaigns and integrate national and state party campaign strategy to an unprecedented extent (Bibby 1999).

3.2 The Senatorial and Congressional Campaign Committees

In addition to the RNC, the national organization of the party includes two autonomous units which focus their activities upon electing members of the House and Senate. Like the RNC, the National Republican Congressional Committee (NRCC) and the National Republican Senatorial Committee (NRSC) have evolved into well funded organizations with professional staffs. These campaign committees are headed respectively by an incumbent representative and senator. They engage in programs of candidate recruitment, fund raising, technical assistance to candidates, and media advertising. They also channel and mobilize nonparty money to candidates from political action committees, wealthy individuals, and incumbent members of the House and Senate. Support is doled out to candidates on the basis of electoral criteria—the competitiveness of the constituency—not on the basis of candidate ideology or party loyalty.

Since 1996, the congressional and senatorial committees have placed increasingly heavy emphasis on 'issue advocacy' advertising in support of their candidates. These are media advertisements funded with 'soft money' that do not expressly advocate the election or defeat of a candidate, but which effectively convey either favorable or unfavorable images of candidates. The campaign committees also engage in independent expenditures which do involve express advocacy of the election or defeat of candidates. Through the use of these techniques the campaign committees have expanded their influence over the national political agenda, helped shape the issue agendas of individual House and Senate campaigns, and attacked their candidates' opponents (Herrnson and Dwyre 1999).

3.3 State Parties

State level party units have also gone through a process of institutional strengthening since the 1960s. Republican state organizations tend to be stronger than their Democratic counterparts in their ability to engage in party building and candidate support activities (Cotter et al. 1984). As is true at the national level, there is a division of labor among state level organizations. Republican state central committees concentrate on state-wide races, voter mobilization, and party building, while legislative campaign committees focus their resources upon electing Republican state legislators.

4. The Republican Party Network of Issue-oriented Activists

American parties are increasingly networks of issue-based participatory activists. The sources for this trend can be found in a series of complex and interacting forces: the development of a postindustrial society in which noneconomic social/cultural issues have gained heightened saliency; sociological and economic changes that have created higher levels of educational attainment, reduced blue collar employment, and increased number of white collar workers; a decline in the availability of patronage as an incentive to participate in politics. Issue oriented activists have also been aided by changes in the rules governing presidential nominations that have diminished the role of elected officials and party leaders.

In the Republican Party these changes have had the effect of significantly shifting the locus of power toward ideologically conservative activists. In the highly participatory politics of Republican presidential nominations, candidates normally find it essential to stress their conservative bona fides if they are to have any hope of winning the nomination in a process dominated by presidential primaries and participatory party caucuses. Activists taking a leading role in these events as well as national convention delegates have been shown to be substantially to the right of the party's rank and file voters (Mayer 1996). Similarly, Bruce et al. (1991) found in a study of county level campaign workers that ideological 'true believers' outnumbered 'vote maximizers' by a ratio of 4.3:1 among Republicans.

Issue-based groups have gained an increasingly entrenched position within the Republican Party. The Christian Right, for example, has become an increasingly influential and sometimes dominant force within some state party organizations. The increasing involvement of groups and individuals whose motivation is not maximizing the vote but a commitment to policy and ideological concerns has created a problem for the Republican Party. With a conservative activist base that is needed for its campaign support and often to win party nominations, candidates often find themselves catering to the GOP's most conservative elements. However, such actions tend to alienate the more moderate elements of the party and voters without strong partisan commitments—the voters that are needed to win general elections. The party's future depends in large part on how this intraparty problem is resolved.

See also: Campaigning: Political; Congress: United States; Democratic Party; Legislatures: United States; Party Identification; Party Systems; Political Parties, History of; Presidential Government

Bibliography

Aldrich J H 1995 *Why Parties? The Origin and Transformation of Party Politics in America*. University of Chicago Press, Chicago

Beck P 1999 The changing American party coalitions. In: Green J C, Shea D M (eds.) *The State of Parties: The Changing Role of Contemporary American Parties*, 3rd edn. Rowman and Littlefield, Lanham, MD

Bibby J F 1999 Party networks: National–state integration, allied groups, and issue activists. In: Green J C, Shea D M (eds.) *The State of Parties: The Changing Role of Contemporary American Parties*, 3rd edn. Rowman and Littlefield, Lanham, MD

Bruce J M , Clark J A, Kessel J H 1991 Advocacy politics in presidential parties. *American Political Science Review* **85**: 1089–110

Cotter C P, Bibby J F 1980 Institutional development of parties and the thesis of party decline. *Political Science Quarterly* **95**(Spring): 1–28

Cotter C P, Gibson J L, Bibby J F, Huckshorn R J 1984 *Party Organizations in American Politics*. Praeger, New York

Herrnson P S, Dwyre D 1999 Party issue advocacy in congressional election campaigns. In: Green J C, Shea D M (eds.) *The State of Parties: The Changing Role of Contemporary American Parties*. J Rowman and Littlefield, Lanham, MD

Jones C O 1965 *The Republican Party in American Politics*. Macmillan, New York

Mayer W G (ed.) 1996 *In Pursuit of the White House: How we Choose our Presidential Nominees*. Chatham House, Chatham, NJ

Moos M 1956 *The Republicans: A History of their Party*. Random House, New York

J. F. Bibby

Republicanism: Impact on Social Thought

In the modern age, republicanism influences social thought by offering a distinctly political articulation of the general belief in the historical contingency of political and social order (Pocock 1975). Modern republicanism holds that political and social order is legitimate only if it serves the constitution of political freedom as nondomination. The possibility of changing forms of rule is understood as a condition of the institution of these forms (Arendt 1963). The existence of ineradicable and productive social divisions is a fundamental presupposition of the republican idea of politics, according to which social conflicts must be given their political expression and regulation (Lefort 1992).

Originating with Machiavelli's critical reception of the tradition of classical republicanism, pursued in the English and Dutch struggles against monarchic absolutism by the likes of Harrington, Sidney, Locke, and Spinoza, and systematized in the Enlightenment by Montesquieu and Rousseau, modern republican theory is directly put into practice in the American and French Revolutions by figures such as Madison, Jefferson, and Sieyès (Pocock 1975, Bock et al. 1990, Fontana 1994). In this sense, modern republicanism has been justly termed a 'revolutionary republicanism' (Wootton 1994).

1. Political Freedom as Nondomination

Modern republicanism describes political freedom as that arrangement in which 'the rule of law, resting on the power of the people, would put an end to the rule of man over man' (Arendt 1972 p. 139). Politics should be that domain of human affairs where domination

and rule can be questioned, checked, and even possibly abolished (Pettit 1997).

The modern republican ideal of freedom as nondomination or no-rule moves beyond both the negative and positive ideas of freedom (Berlin 1965). Negative freedom is an individual and private freedom from interference that requires a noncontesting relation to the commands of the political sovereign that protect and enable it from the outside. In contrast, the ideal of nondomination is political, not private, and consists of being 'subject to no man's will' (Sidney 1798, III, 16) by having the power to contest both arbitrary and legitimate domination. Freedom as nondomination also differs from ideals of self-rule or positive freedom insofar as these view some forms of rule (e.g., of the community over its members, of reason over passion) as intrinsic to the meaning of freedom, whereas from the perspective of freedom as nondomination, political activity is primarily understood in terms of its capacity to check the dynamics of domination and processes of ruling by empowering those who would otherwise fall into relations of dependence. Political freedom 'is there only when there is no abuse of power.... To prevent this abuse, it is necessary from the very nature of things that power should be a check to power' (Montesquieu 1748, XI, 4).

Modern republicanism sees the political abolition of domination as achievable through a combination of the 'rule of law' and the 'power of the people.' Although both of these elements can be found in classical republicanism as well, their meaning and relative priority undergo important changes in modern republicanism where 'all power is vested in, and consequently derived from, the People' (*Virginia Declaration of Rights* 1776, article 2) because its members are understood as the carriers of the ideal of nondomination: 'all men are by nature equally free and independent, and have certain inherent rights of which, when they enter into a state of society, they cannot, by any compact, deprive or divest their posterity' (*Virginia Declaration of Rights* 1776, article 1). As a consequence, the republican ideal requires that the rule of law find its support in the power of the people.

2. *The Power of the People*

The connection between the power of the people and the ideal of nondomination can be traced back to Machiavelli in his *Discourses on Livy* where, for the first time, 'the people' is defined as that subject position occupied by whoever acts in political ways that are motivated 'only [by the] desire not to be dominated' (Machiavelli 1517, I, 5).

The 'power' that is vested in 'the people' refers to those political practices and institutions of will and opinion formation that integrate this political subject as bearer of the desire for nondomination into political life, alongside the established practices and institutions of legitimate domination, of which traditionally the most important elements were the deliberations of the senate and the commands of the monarch (Machiavelli 1517, I, 4). Harrington does away with the monarchical element, and takes into account only the nobles, represented by the senate, and the people, in order to give ultimate say to the latter: 'as the wisdom of the commonwealth is in the aristocracy, so the interest of the commonwealth is in the whole body of the people.... Whatsoever upon debate of the senate is proposed unto the people, and resolved by them, is enacted by the authority of the fathers and the power of the people, which concurring make a law' (Harrington 1656 p. 24). As Rousseau frames the issue, 'it is only the general will which compels individuals and there can be no assurance that an individual will is in conformity with the general will until it has submitted to the free suffrage of the people' (Rousseau 1762, II, 7). It is an axiom of modern republicanism that without this kind of political power vested in the people there can be no political freedom.

Likewise, since the people is conceived as a political subject whose foremost interest is nondomination or no-rule, it follows that it should not, as a people, engage directly in the process of governing or ruling on pain of self-contradiction. For this reason Madison is perfectly consequent when he claims, in the same article of *The Federalist*, that 'the people can never willfully betray their own interests,' and that the distinguishing feature of 'American government lies in the total exclusion of the people in their collective capacity from any share in the latter' (Hamilton et al. 1787, p. 63). Modern republicanism requires the principle of representation because it understands political representation not as a condition for the pursuit of more effective ways of governing a large population, but as the primary way to limit the powers of the state as an agency of legitimate domination.

3. *Politics and the Primacy of Social Division*

The analysis of society in terms of the conflict between heterogeneous social subjects and its causality in the emergence of political forms is a fundamental contribution of modern republicanism to social thought. Machiavelli shows that the opposition between those who 'desire not to be commanded or oppressed' and those who 'desire to command and oppress' constitutes a basic social division that, depending on how it is politically interpreted, gives rise to 'one of three effects in cities: principality or freedom or license' (Machiavelli 1513, IX). Similarly, Harrington observes that every state finds its basis in the social division of property. For that reason, he argues that an 'agrarian law' must be set at the

'foundation' of the republic. In so doing, Harrington both acknowledges the fundamental role of social conflict and seeks to regulate it by giving it a political form (Harrington 1656 p. 33). Modern republicanism holds that social divisions provide the basis of politics, and that politics has the task of giving or taking away form (laws and orders) in response to such conflicts, thereby expressing and regulating rather than extirpating or denying them (Pocock 1975, Lefort 1992).

After the American and French Revolutions, and well into the twentieth century, this relative autonomy of politics with respect to 'the social' comes under criticism thanks to the emergence of the idea of society as a form of organization that, in and through its divisions, is self-constituting and self-regulating. This idea of society opens the prospect of the abolition of politics and of the state, under the appropriate social conditions. Much of the impetus for the late twentieth century retrieval of the republican tradition comes from the need to recover a positive conception of politics and, along with it, a sense of the productive tension between state and society.

4. *Republicanism and Institutionalization*

For modern republicanism, politics consists of the practice of giving an institutional expression, or interpretation, to the divisions that constitute social relations. In particular, republican politics sets out to refashion the state so as to be able to offer institutional responses to the demands for nondomination issuing from society. The state, viewed as a form of legitimate domination, can only respond to these demands by internalizing its own negation and becoming internally self-checking. This principle underlies the recurrent attacks that modern republicans wage on the very idea of a 'standing power,' against which they introduce the need for 'motion' or 'the successive revolution of authority' (Nedham 1659 *Mercurius Politicus* 4 December 1651–1 April 1652). It also accounts for the constant thematization of political change and social conflict in the very constitution of political form. The debate between Jefferson and Madison on the uses of constitutional conventions to check the power of the state is exemplary in this sense: the latter believes in designing a political form of government capable of checking itself, while the former holds that a check on government ultimately can only come from what lies outside it (Hamilton et al. 1787, pp. 49–51).

The innovation of modern republicanism with respect to thinking about institutions consists of providing both theories and examples of what may be called institutions of critique, or counterinstitutions, which take the form of inscribing a resistance to domination into the state apparatus that administers the rule of law. The most significant example of such an institution of critique is the separation of state powers when the latter is understood as *a system of checks and balances* (Hamilton et al. 1787, p. 51) whereby 'one branch of the government may be authorized to exercise some active influence on another in order to resist and counteract its power' (Manin 1994, p. 31) in view of keeping the entire state in check with respect to the people.

Modern republicanism offers other examples of such institutions, whose sole purpose is to serve as a forum for the contestation and resistance of the normal administration of state rule. Among them stands out the Tribunate of the people, interpreted by Machiavelli as a 'guard of freedom' (1517, I, 3–4), which, in the words of Rousseau, 'although it can do nothing, it can prevent anything from being done. As the defender of the laws, it is more sacred and more venerated than the prince who executes laws or the sovereign which legislates' (Rousseau 1762, IV, 5).

Another example is the Agrarian Law as discussed in Machiavelli (1517), Harrington (1656), and Hutcheson's *System of Moral Philosophy* (1755), where it serves to check the unregulated distribution of private property. Among such institutions also figures Jefferson's 'republic of wards,' designed to serve as 'a system of fundamental balances and checks for the government' (Kurland and Lerner 1987, p. 142). The same principle animating the ward system is embodied by the 'council systems' (Arendt 1963) that emerged during the republican revolutions of France in 1871, Russia in 1905 and 1917, Germany in 1918, Hungary in 1956, and Poland in 1979.

5. *Republicanism and Revolution*

Modern republicanism also recognizes the possibility of an external critique of the state-form or revolution. In political revolutions, it is the people that attempt to institute nondomination directly, and it can do so only by suspending or dissolving the state form as such. Although prefigured by Machiavelli with his theory that a free political life can be maintained only if there are periodic 'return to beginnings' in which the social and political orders are renewed (Machiavelli 1517, III, 1), one of the firmest formulations of the state-dissolving power of the people is found in Locke, for whom 'the community may be said in this respect to be always the supreme power, but not as considered under any form of government, because this power of the people can never take place till the government be dissolved' (Locke 1690, XIII, 149). Likewise, Rousseau insists that 'the moment the people is lawfully assembled as a sovereign body all jurisdiction of the government ceases' (Rousseau 1762, III, 14). Hamilton reaffirms in *The Federalist* 'that fundamental principle of republican government which admits the right of the people to alter or abolish the established Constitution whenever they find it in-

consistent with their happiness' (Hamilton et al. 1787, p. 78).

It is important to note that for modern republicanism, unlike other late modern revolutionary traditions, political revolutions are recurring events out of which all forms of legitimate domination emerge and into which they can all be revoked. By definition these revolutionary events cannot be subsumed into or become in turn a political form without denying themselves. Republican moments are in themselves evanescent, having as their fundamental purpose that 'men who live together in any order whatever [be able to] often recognize themselves…and look over the account for the men who are in that body' (Machiavelli 1517, III, 1). The purpose is the mutual recognition of the liberty and equality of all.

6. *Political Equality and the Rule of Law*

The 'rule of law' is a feature of both classical and modern republicanism. But whereas classical rule of law is substantive and intends to prescribe the common good, modern rule of law is formal and has as its essential purpose 'the restriction of each individual's freedom so that it harmonizes with the freedom of every one else' (Kant 1793, p. 73). As evidenced in the American and French Declarations of rights, the rule of law is always limited by, because it is founded upon, two elements that are absent in classical republicanism: individual natural rights and sovereign power of the people.

Following the trajectory that leads from Machiavelli through Rousseau up to the American and French Declarations, some contemporary republican theorists hold the view that unless the rule of law is grounded on the power of the people, it cannot guarantee a political life free of domination (Arendt 1963, Habermas 1992). Other republican theorists, supporting themselves on Montesquieu's definition of liberty as 'the right of doing whatever the laws permit' (Montesquieu 1748, XI, 3), believe that the relation between the rule of law and the power of the people should be reversed because it is 'the laws of a suitable state…[that] create the freedom enjoyed by citizens' (Pettit 1997 p. 36).

At stake in this discussion is the interpretation of a basic republican principle of equality before the law. On one understanding, freedom as nondomination basically requires 'an empire of laws' (Harrington 1656, p. 20) assuring that no one is above the law and thereby excluding arbitrary rule. On the other understanding, freedom as nondomination also requires equality in the making of the laws before which everyone is formally equal. So, for instance, 'No Man can be imprisoned, unless he has transgressed a Law of his own making' (Trenchard 1697, p. 2). Rousseau simply says that 'a people, since it is subject to laws, ought to be the author of them' (Rousseau 1762, pp. II, 6). Without the equality to make law, the rule of law harbors the possibility of paternalism and legal domination (Ferry and Renaut 1985, Habermas 1992).

To be politically free is to have the power to 'make' law in the sense of having the power to constitute, and therefore check the exercise of, any constituted legislative power (Sieyès 1789). Such constituent power cannot itself fall under the rule of law but rather grants it its ultimate legitimacy ground, as is made clear not only in Sieyès, but also in Rousseau and Locke. For both, the power to make law belongs to the people originally, which is constituted by a contract that everyone makes (with everyone else or with themselves) (Locke 1690, XI, 134, Rousseau 1762, I, 6). Locke then permits the transfer of the legislative power of the people to the government (X, XI); but, as an element of the state, the rule of law is always checked by the deciding power of the people (XIII, 149; XIV, 168; XIX, 240).

The question of constituent power as it is developed in modern republicanism has an impact on social thought because it highlights the existence of political but prelegal struggles for recognition (Taylor 1994), through which individuals and groups seek the constituent power of citizenship that gives access both to political and individual rights.

7. *Representation and Participation*

For modern republicanism, political representation is a response to Montesquieu's call for a separation of legislation and execution 'lest the same monarch or senate should enact tyrannical laws, to execute them in a tyrannical manner' (Montesquieu 1748, XI, 6). Such a separation of powers is echoed as much by Rousseau (1762 III, 1, III, 16) as by Kant, who defines republicanism as that 'political principle whereby the executive power (the government) is separated from the legislative power' (Kant 1795, p. 101) vested in the people as 'co-legislators' (Kant 1798, p. 184).

The modern republican discourse on political participation follows, rather than contradicts, Montesquieu's doctrine of the separation of powers. Since a republic entails the division between those who have the power to make laws and those who have the power to apply it, it follows that the activity of governing or the administration of the constituted powers of the state must be given over to representatives, and in that sense all government must be representative in order to be legitimate. But the generation of constituent power has to be kept in the hands of the people, and to that extent calls forth their participation, without which no such power would come into existence in the first place. The debate between Locke and Rousseau on whether the fundamental power of the people to make law can or cannot be represented (Locke 1690, X, 132, XIII, 149,

Rousseau 1762, III, 15) is resolved in the form of the distinction between a constituent legislative power, which is what Rousseau and Kant have in mind when they advocate its inalienability from the people, and a constituted legislative power, which is what Locke, Sieyès, and Madison have in mind when they argue for its representative character. In both cases, the goal is to assure Montesquieu's requirement of contrasting government by setting power (of the people) against power (of the state).

8. *Political Virtue and Corruption*

The discourse of 'virtue' in modern republicanism (Pocock 1975) for the most part refers to a political conception of virtue, rather than to the classical moral conception of virtue as the pursuit of the common good (Machiavelli 1513, Montesquieu 1748). The debate about political virtue is a place-holder for the necessity of a certain kind of 'government by men' in order to maintain the ideal of nondomination in the face of the potential for 'corruption' (Skinner 1998) harbored in the modern rule of law and other systems of negative liberty, of which the most significant for modern republicans is the system of commerce and money. Since the modern rule of law and modern commerce are not viewed as being morally substantive, but rather as protecting and fostering the negative liberty needed for all nonpolitical forms of self-enactment (Wootton 1994), simple adherence to them cannot replace the political activity needed to sustain both the participatory and the representative requirements of republicanism. Left to their own devices these systems of negative liberty could degenerate into forms of bureaucracy (no-man's rule), as well as deepen social and economic inequalities, thereby threatening freedom as nondomination.

Modern republicans speak of two kinds of political virtues, depending on whether they take virtue to be synonymous with the set of characteristics that make political subjects have 'power' or 'authority,' and depending on whether they are more interested in the 'freedom' or in the 'stability' of the republican form. Harrington and Madison concentrate on the need for an 'aristocratic' virtue, to be present in the representatives, for the sake of the 'wisest' and most 'authoritative' outcome in their deliberations, which is in turn the main guarantee of the stability of the republican form. By contrast, Machiavelli, Jefferson, and other proponents of 'public happiness' offer a distinctly more 'democratic' conception of political virtue, designed to foster in the people that 'power' which comes from being, in Jefferson's felicitous phrase, 'a participator in the government of affairs' (Kurland and Lerner 1987, p. 142). As a result, the rule of law requires virtuous individuals, and there is no virtue without the power to do 'strong' and perhaps 'uncivil' actions (Machiavelli 1517, I, 18, III, 1) of the kind required by engagement in practices of active nondomination, of empowerment and resistance against practices of domination and dependency.

See also: Cleavages: Political; Corruption: Political and Public Aspects; Equality and Inequality: Legal Aspects; Equality: Philosophical Aspects; Freedom/Liberty: Impact on the Social Sciences; Freedom: Political; Institutionalization; Machiavelli, Niccolo (1469–1527); Montesquieu, Charles, the Second Baron of (1689–1755); Political Representation; Power in Society; Power: Political; Republicanism: Philosophical Aspects; Revolutions, History of; Revolutions, Theories of; Rousseau, Jean-Jacques (1712–78); Rule of Law

Bibliography

Arendt H 1963 *On Revolution*. Penguin Books, New York

Arendt H 1972 *Crises of the Republic; Lying in Politics, and Disobedience on Violence, Thoughts on Politics, and Revolution*. Harcourt Brace, New York

Berlin I Sir 1965 *Four Essays on Liberty*. Oxford University Press, Oxford, UK

Bock G, Viroli M, Skinner Q (eds.) 1990 *Machiavelli and Republicanism*. Cambridge University Press, Cambridge, UK

Ferry L, Renaut A 1985 *Philosophie Politique, vol. 3: Des Droits de l'homme à l'idée républicaine*. Presses Universitaires de France, Paris [1992 *Political Philosophy 3. From the Rights of Man to the Republican Idea*. University of Chicago Press, Chicago]

Fontana B (ed.) 1994 *The Invention of the Modern Republic*. Cambridge University Press, Cambridge, UK

Habermas J 1992 *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*. Suhrkamp Verlag, Frankfurt am Main [1996 *Between Facts Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT, Cambridge, MA]

Hamilton A, Jay J, Madison J (eds.) 1987 [1787–8]. *The Federalist*. In: Kramnick I (ed.) *The Federalist Papers*. Penguin, Harmondsworth, UK

Harrington J 1992 [1656] *The Commonwealth of Oceana*. Pocock J G A (ed.), Cambridge University Press, Cambridge, UK

Hutcheson F 2000 [1755] *A System of Moral Philosophy*. Thoemmes, Bristol, UK

Kant I 1793 *Über den Gemeinspruch: 'Das mag in der Theorie richtig sein, taugt aber nicht für die Praxis'* [1991 Reiss H (ed.), Nisbet H B (trans.) *Kant: Political Writings*. Cambridge University Press, Cambridge, UK]

Kant I 1795 *Zum ewigen Frieden. Ein philosophischer Entwurf* [1991 Reiss, H (ed.) *Kant: Political Writings*. Cambridge University Press, Cambridge, UK]

Kant I 1798 *Der Streit der Facultäten* [1991 Reiss H (ed.) *Kant: Political Writings*. Cambridge University Press, Cambridge, UK]

Kurland P B, Lerner R (eds.) 1987 *The Founders' Constitution*. University of Chicago Press, Chicago

Lefort C 1992 *Écrire. A l'épreuve du politique*. Calmann-Levy, Paris
Locke J 1980 [1690] *Second Treatise of Government*. Macpherson C B (ed.), Hackett, Indianapolis, IN
Machiavelli N 1996 [1517] *Discourses on Livy*. University of Chicago Press, Chicago
Machiavelli N 1998 [1513] *The Prince*. University of Chicago Press, Chicago
Manin B 1994 Checks, balances and boundaries: the separation of powers in the constitutional debate of 1787. In: Fontana B (ed.) *The Invention of the Modern Republic*. Cambridge University Press, New York
de Secondat Montesquieu C 1989 [1748] *The Spirit of the Laws* Cohler A M, Miller B C, Ston H S (trans. and eds.). Cambridge University Press, New York
Nedham M 1659 *Mercurius Politicus*. Cornmarket Press, London
Pettit P 1997 *Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford, UK
Pocock J G A 1975 *The Machiavellian Moment. Florentine Political Thought and the Atlantic Republican Tradition*. Princeton University Press, Princeton, NJ
Rousseau J-J 1968 [1762] *The Social Contract*. Penguin, London
Sidney A 1990 [1798] *Discourses Concerning Government*. West T G (ed.), Liberty Fund, Indianapolis, IN
Sieyès E J 1970 [1789] *Qu'est-ce que le tiers état?* Droz, Geneva
Skinner Q 1998 *Liberty before Liberalism*. Cambridge University Press, Cambridge, UK
Taylor C (ed.) 1994 *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton, NJ
Trenchard J 1697 *An Argument, Shewing, that a Standing Army is Inconsistent with a Free Government, and Absolutely Destructive to the English Monarchy*. London
Virginia Declaration of Rights 1987 [1776]. In: Kurland P B, Lerner R (eds.) *The Founders' Constitution*. University of Chicago Press, Chicago
Wootton D (ed.) 1994 *Republicanism, Liberty, and Commercial Society, 1649–1776*. Stanford University Press, Stanford, CA

M. E. Vatter

Republicanism: Philosophical Aspects

Republicanism is the doctrine that laws and the state should always serve the common good (or '*res publica*') of a nation's people or citizens ('*populus*'). The basic desiderata of republican government, as articulated in the republican tradition derived from Rome, secure government for the common good through the checks and balances of a mixed constitution, comprising a sovereign people, an elected executive, a deliberative senate, and a regulated popular assembly, constrained by an independent judiciary, and subject to the rule of law (Sellers 1994). Some republicans would add representation, the separation of powers, or other new structural devices to protect public liberty ('*libertas*') and avoid Rome's eventual descent into popular tyranny and military despotism (Sellers 1998). Republican liberty signifies subjection to the law and magistrates, acting for the common good, and never to the private will or domination ('*dominatio*') of any private master (Pettit 1997).

1. The Republican Tradition

'*Res publica*' was the Romans' term for their state, its public business, all public property, and the purposes these served. The word notoriously evades translation, most often appearing in English as 'commonwealth,' or simply (more recently) 'republic,' the use preferred here. Marcus Tullius Cicero (BC 106–43) and Titus Livius (Livy) (BC 59–AD 17) constructed the first and most influential comprehensively republican ideology, to try to explain how and why the Roman republic had failed to serve the public good. Both agreed that republican institutions collapsed when military expansion and party conflict upset traditional checks and balances between the senate, the magistrates, and the people of Rome.

Cicero and Livy inaugurated a republican tradition of 'liberty' that fortified principled resistance to demagogues, emperors, and kings for the next two thousand years. Niccolò Machiavelli (1469–1527) did the most to revive this republican tradition in Italy, in his *Discorsi sopra la prima deca di Tito Livio* (1517). The resistance of the Swiss, various Italian cantons, and the United Provinces of the Netherlands to imperial control added practical models for republican liberty, as did the constitutional and theoretical writings of various English authors, in their efforts to restrain or remove their kings during the Civil War and Commonwealth (1642–60), the Glorious Revolution (1688), and the extended British controversies over American Independence (1763–83).

2. The Republican Revival

Republican political and legal philosophy provides an enduring model for philosophers and lawyers who oppose the domination of arbitrary power. The constitutional controversies of the French and American revolutions left institutional relics, which established republican ideals in the structures of Western politics, with a pervasive commitment to 'liberty' and even (in the United States), a federal 'guarantee' that every state in the union must enjoy 'a republican form of government' (*United States Constitution*, 1787 Art. IV § 4). Similar provisions requiring senates and independent judges give republican theorists a practical foundation in the constitutions of most Western democracies. Lawyers such as Frank Michelman (1936–) and Cass Sunstein (1954–) first revived republican doctrine in law schools, to justify judicial intervention for the common good, against partisan legislation and political corruption (*Symposium* 1988).

Philip Pettit inaugurated a similar revival among philosophers, embracing three central claims of the republican tradition: that government must serve the common good; that certain institutional arrangements will do so; and that liberty will be the result (Pettit 1997). Republicanism since Cicero has offered a political epistemology for finding the common good, or approximating it as closely as possible. The republican question has not been (in the first instance): 'What is the common good?'; but rather: 'Which political procedures will best find and protect the common good, given the possibilities and limitations of history and human nature?' (Sellers 1991).

3. *Liberty as Nondomination*

Republican liberty consists in not being interfered with by fellow citizens or the state, except as regulated by the common good of the people, established by republican politics. This conception of liberty as 'nondomination' contradicts a widely-held (but more recent) understanding of liberty as not being interfered with at all. Public and private actions that constrain or influence a person's activities do not restrict republican liberty unless they violate the law or republican moral principles established to serve the common good of the people as a whole. Republican liberty cannot exist without a republican state because nonrepublics will always dominate their subjects, as masters dominate their slaves.

Domination marks the end of liberty, in republican theory, because domination subjects one person to the arbitrary will of another, without regard to the common good. The unchecked power of the *dominus* (or master) makes his subordinates slaves, whether the master chooses to exercise power or not. Slaves to good masters remain slaves, and dominated, because they have no security against their master's will. States dominate citizens when their constitutions give legislatures or public officials arbitrary power to act, without regard to the public welfare. Private individuals dominate others when laws and public officials do not protect citizens against private oppression.

The common good is the measure of domination, and therefore of liberty, in republican political vocabulary, because only the common good distinguishes 'arbitrary' interference from justified state action. *Libertas* (liberty) in Rome signified the status of a free citizen in a republican state. The *res publica* (republic) represented every citizens' common interest in the public good. Restraining citizens equally with just laws in pursuit of the common good enhances their liberty by restraining arbitrary power. Neorepublican doctrine revives liberty as a worthy object of public policy by rescuing the sense in which the word first developed as a political ideal. Liberty is not, as Thomas Hobbes (1588–1679) maliciously redefined it, the unfettered license to do what one wants, but rather a public shield against uncontrolled license in others. Republican liberty consists in equal subjection to the rule of laws, made for the common good, through the carefully balanced mechanism of a republican form of government. For further references see *Freedom: Political*.

4. *The Common Good*

The fundamental republican insight that government and laws should serve the common good did not originate in Rome. Most governments have made this claim, which Cicero (Tullius Cicero *Off.* I.xxv.85) attributed to Plato (circa BC 427–347). Cicero defined the purpose of the state as seeking to create harmony from the disparate interests of all members of society (*Off.* III.vi.26). This does not require that all public interests always serve every individual's separate desires, but that laws and government exist to make every citizen's life worthwhile. Most human goods have a social component, making liberty essentially communal in character.

The first difficulty republicans face in finding the common good lies in defining their community. Ancient republics closely restricted their citizenship, excluding women, immigrants, and slaves. Most republicans would now extend citizenship and membership in the *populus* or people to all inhabitants of a given territory. This leaves open the question of boundaries. Cicero endorsed a universal republic or society of humanity, but also recognized the value of local politics. Cultures and human relationships develop and solidify locally. Therefore, the republican interest in harmony and community implies small homogenous republics, where people will be very much the same (Montesquieu 1748). But the natural diversity of human talents and interests guarantees that small communities will find their own internal minorities to dominate and oppress. Larger republics with greater diversity better protect minorities, by making dominant factions harder to assemble (Hamilton et al. 1787–8, p. X).

The republican solution to republican boundaries has been to encourage different levels of nested republics, for different purposes, with different capacities. Immanuel Kant (1724–1804) advocated a federation of national republics, assembled in perpetual peace (Kant 1795). The formula suggested by the United States and modern European Union has been to entrust the protection of individual human rights and commerce, which local majorities might threaten, to federal suprarepublican authorities, while delegating social and cultural concerns to smaller, more homogenous republics. When regional authorities consistently abuse their powers to oppress local minorities, their oppression implicitly recognizes such minorities as separate people, entitled to separate republican homelands, within the republican federa-

tion (Sellers 1996). For further references see *Public Interest*.

5. Popular Sovereignty

Republicans realized very early in Rome, and have maintained the principle ever since, that the common good will never be found or maintained without the *imperium populi* or popular sovereignty to support it (Tullius Cicero *Cat*. IV.14; *Phil*. IV.iv.8). Macchiavelli credited Cicero with the insight that although the people may be ignorant, they are capable of grasping truth, when good men place the truth before them (Macchiavelli 1517 I.iv.10). The purpose of republican popular sovereignty is not that the people should govern every day, but rather, as Benjamin Rush (1745–1813) suggested in advocating an American 'republic,' that the people should select their rulers themselves. The people exercise their power on election days, then defer to magistrates that they themselves have selected (Rush 1777).

This republican reliance on popular sovereignty distinguishes republican government from liberalism, which diverged from the republican tradition at the beginning of the nineteenth century (Sellers 1998). Early liberals rejected popular sovereignty in the wake of the French revolution, while retaining republican commitments to liberty, the rule of law, and certain specific rights against of government (Constant 1819). Liberals divorced liberty from its basis in public deliberation by distinguishing 'political liberty' from personal independence. This left liberals dependent on judges and magistrates to verify their assertions of rights, and protect individual liberty against the state. Liberalism started as a flight from politics, but found that this implied a new definition of liberty as the ability to do what one wants, without state regulation (Pettit 1997).

Contemporary liberalism has largely abandoned the fear of democracy that separated liberalism from republicanism in the first place, and most liberals now endorse the widespread political participation of citizens in elections (Rawls 1993). Republican popular sovereignty never signified democracy in any case, but rather a carefully balanced and controlled mechanism for gathering the insights of all members of society, through a process of public deliberation, in pursuit of the common good. As modern liberals have sought a method of democratic deliberation to legitimate their sense of community, and regulate conflicting private interests, they have increasingly returned to republican principles and institutions, embedded in Western constitutional democracy. For further references see *Sovereignty: Political*.

6. Checks and Balances

The primary, process-centered challenge for republicans has been to identify those institutions that find and secure the common good best. Popular sovereignty provides access to every citizen's interests and insights, but democracy would become an elective despotism, without checks and balances to restrain it (Pettit 1999). Cicero blamed the Greeks' misfortunes on the turbulence of their popular assemblies, lacking internal balance or a senate to control them (Tullius Cicero *Flacc*. 1976). Republics since Rome have maintained bicameral legislatures to prevent self-seeking in either assembly (Harrington 1656, p. 22). Democracy is one of several foundation stones of liberty, not an end in itself.

The necessary checks and balances of republican government prevent public officials from making themselves and not the *res publica* the object the state (Paine 1792, p. 168). The dispersion of power through bicameralism, federalism, and the separation of powers makes it harder for any one person or faction, including the majority, to wield arbitrary power over others (Pettit 1997, p. 177). American republicans, such as John Adams (1787–8, p. I.ii–iii) and Alexander Hamilton et al. (1787–8, IX) added representation and the life tenure of judges to bicameralism and the checks on governmental powers as basic desiderata of balanced republican institutions. If the people and judges are outside government, they can better control their government's mistakes (Hamilton et al. 1787–8, p. LXIII).

Modern republicans such as Philip Pettit differ from democrats in viewing democracy as a derivative value, in service to the broader ideal of balanced or 'contested' government (Pettit 1997, p. 187). All government decisions should be subject to challenge by institutions that prevent private inclinations from ruling public interests and ideas. Republicans have proposed bills of rights, public hearings, and many other devices designed to constrain and channel public decision making, so that ordinary citizens may influence their government's decisions, without diverting the public purposes of the state. For further references see *Constitutionalism*.

7. The Rule of Law

Republican government consists of whatever institutions 'will compel the formation of good and equal laws, an impartial execution, and faithful interpretation of them, so that citizens may constantly enjoy the benefit of them, and be sure of their continuance' (Adams 1787–8, p. I.128). Such institutions secure the '*imperia legum*' or 'empire of laws and not of men' praised by all republican authors from Livy to Pettit. Laws secure liberty against domination by delimiting and protecting the line across which private or state behavior violates individual autonomy against the common good. Cicero insisted that such laws must serve the public welfare (*populi utilitas*) not the public will (*populi voluntas*) (Tullius Cicero 1906 [*Sulla* 25])

because votes cannot alter the natural laws of justice (Tullius Cicero *Leg*. I.xvi. 44). Republics need institutions that will find justice by securing the common good.

The empire-of-law condition of republican government entails that laws should be publicly promulgated, intelligible, consistent, stable, general, and apply to everyone, including the legislators themselves. This prevents arbitrary decisions by public officials, by subjecting their will to known constraints and purposes. Republics maintain a general presumption that government action, when needed, will operate by law, not by *ad hoc* or *ex post* decisions. All agencies of government must act through principled, regulated structures, maintaining the 'due process' of laws to prevent the abuse of governmental power (Pettit 1997, pp. 174–5).

Codifying the law in every detail will not be possible or desirable. Some discretion must remain, but subject to the dispersion-of-power and popular sovereignty procedures that generate republican laws in the first place. Discretion subject to constraints against arbitrariness may sometimes secure the common good, but unlimited discretion produces mistakes about justice, through the natural partiality and limited viewpoint of any individual decision-maker. The republican rule of law protects liberty by respecting the welfare of every member of society, pooling their insights to secure the common good of the people. Legislation is the public reason of the republic, protecting citizens' liberty against private power (Harrington 1656, pp. 19–20).

8. *Neorepublican Ideals*

Contemporary philosophical republicanism reasserts the republican conception of liberty, proposing new republican institutions to support republican liberty in the modern multicultural state. Philip Pettit has given the most detailed proposal for neorepublican politics, supplanting the Benthamite liberal opposition of 'negative' and 'positive' liberty with older republican conceptions of liberty as nondomination, and protection against arbitrary power (Pettit 1997). Pettit understands arbitrary power ('*arbitrium*') to include any action against the 'interests and ideas' of those who suffer interference. This surpasses the old republican standard, which measured arbitrary power by its violation of an objective public, common, or collective good.

Neorepublican deference to private interests and ideas reflects the residual liberal commitment to moral pluralism and cultural diversity. This enduring hesitation to overrule anyone's personal ideas makes the revised standard of nondomination very strong. Interference or influence over others' choices becomes domination unless it tracks their private values. This introduces a subjective element into republican doctrine. Older republican ideals tolerated hierarchy and private influences when toleration served the common good of all citizens.

Neorepublican rhetoric describes the state as seeking to 'promote' freedom as nondomination, where older republicans would have understood republican pursuit of the common good simply to be freedom as nondomination. The difference lies in the pluralistic language of the newer theories. Some suggest that immigrant groups and minority cultures will maintain their separate identities in republics, because republics help to preserve hereditary differences within the larger culture (Pettit 1997, p. 144). This represents a significant departure from traditional doctrine. Republics exist precisely to create and maintain an overarching culture that embraces all citizens. This necessarily influences and discourages dissenting subcultures by promoting an image of national unity, which will tend to assimilate minority perspectives.

9. *Neorepublican Institutions*

The neorepublican commitment to pluralism encourages a greater emphasis on countermajoritarian aspects of republican popular sovereignty (Pettit 1997, *Symposium* 1988). Neorepublicans embrace the empire-of-law condition of republican government and the dispersion-of-power constraint, but they often seek to displace public decision-making onto independent panels of social scientists or courts (Pettit 1997, Sunstein 1993, *Symposium* 1988). This favors 'contestability' over consensus to limit the weight of unified public opinion.

Neorepublicans do not endorse interest-group pluralism in the classical amoral sense of liberal democracy, but their concept of representation favors cultural minorities. Some believe that public policy should be able to be justified according to the lights of every member of society (Pettit 1997, p. 169). This overstates the value of mistaken perceptions of the good. Republican government differs from liberal democracy precisely in that it offers a deliberative technique for overruling mistaken views, while allowing constructive dissent by those who oppose the consensus.

Traditional republics maintained geographical representation to encourage the development of local communities. Most recent proposals prefer religious, cultural, or sexual diversity. Popular elections now seem too crude to identify administrative competence, or to guarantee minimum statistical representation for every stakeholder grouping (Pettit 1997, p. 192). Neorepublicans often propose deliberation among statistically selected 'representative' officials, in place of elected representatives or public assemblies. This leads to difficulties in identifying which sub-

cultures or stakeholder 'groups' to represent. Most people have many such partial or self-defined sources of identity.

10. The People

The fundamental republican conception of the common good encourages social solidarity by seeking a community of interest among the populus or people of any given republic. Cicero described republics as the property of a people (*res populi*), in the most influential and frequently repeated passage in the republican canon. A *populus* is not just any collection of humans, but a partnership about justice, in pursuit of the common good (Tullius Cicero *Rep*. I.xxv.39). There will be no republic without a people and no people (in the fullest sense) without a republic, which is to say the mutual recognition of common citizenship (fraternity) and the equal importance (equality) of all citizens in the eyes of the state.

The concept of peoples represents the most contested element in contemporary republican doctrine, and the most likely source of disagreement between neorepublican philosophers. The United Nations Charter recognized the republican principle of self-determination of peoples (Article I (2)), implying a people's right 'freely [to] determine their political status' (International Convention on Civil and Political Rights 1966, Article 1). This gives the republican identity a practical importance in international politics, implying that recognized status as a 'people' entails some degree of political separation. The salient examples of liberated 'peoples' have been the inhabitants of colonial territories, who constitute peoples in international law for the purpose of pursuing their independence from imperial domination.

Some neorepublicans would separate the national and territorial principles to create politically separate republican peoples on the basis of ethnic, religious, or linguistic affinities within larger multinational empires. Different laws would apply to different subjects of the overarching state, according to their different internal 'citizenship' or status. This violates the republican principle of equal citizenship to achieve the republican ambition of social solidarity among state-maintained subgroups of the population. Traditional republican doctrine would insist that all citizens enjoy the same privileges and immunities in their public capacities, while permitting private initiatives to pursue elective affinities. This protects (but may subtly discourage) private diversity, because the people, as public citizens, will be the same under law.

11. Republican Philosophy

Republican philosophy since Cicero has sought to construct a harmony of interests and common sense of justice among citizens through the 'empire of laws and not of men.' Republican laws draw the line between liberty and license, in pursuit of the common good. Republican theory seeks to find and establish good laws, by discovering the principles and basic structure that serve the res publica best. So although republican philosophy begins, as Thomas Paine (1737–1809) put it, by making the '*res publica*, the public affairs, or the public good' the object of all governments, and 'republican government is no other than government established and conducted for the interest of the public' (Paine 1792, p. 168), the idea of the republic entails a constellation of political structures to secure republican legislation, embedded in two thousand years of republican tradition, derived from Rome.

The republican commitment to political institutions that protect liberty challenges the later 'liberal' flight from politics after the French revolution. Former republicans such as Benjamin Constant (1767–1830) hoped by securing certain rights or liberties to protect their own private liberty against the state. But republican philosophy recognizes the futility of 'rights' without power. Republican doctrine offers a political epistemology to discover and protect public justice and fundamental human rights. Republicans believe that without popular sovereignty, the rule of law, deliberative senators, elected executives, independent judges, a representative popular assembly, and proper checks and balances, the people cannot know or enjoy their rights and duties to each other, or to the state. The basic desiderata of republican government may appear more (or less) important in different contexts, so that even a monarchy could seem nearly 'republican' when limited by two independent branches in the legislature, and subject to the rule of law (Adams 1787–8, pp. I. xxi–xxii). Contemporary republican philosophers adapt republican principles to new circumstances to protect public liberty against improper domination ('*dominium*') by private interests, or the arbitrary government ('*imperium*') of the state (Pettit 1997). What makes such theories republican is their fundamental commitment to the common good, and to the political structures that support the *res publica* best. Republican philosophy is a theory of freedom and government, or rather of freedom through government, to secure a shared sense of justice, in pursuit of the common good. Republicans believe that there can be no justice without community, no liberty without the law.

See also: Aristotle (384–322 BC); Democracy; Justice: Philosophical Aspects; Republic; Republicanism: Impact on Social Thought

Bibliography

Adams J 1787–8 *A Defence of the Constitutions of Government of the United States of America*. 3 Vols. C. Dilly, London

Adams J 1776 Thoughts on government. 1983 In: Hyneman C S, Lutz D S (eds.) *American Political Writing During the Founding Era 1760–1805,* 2 Vols. Liberty Press, Indianapolis, IN

Baron H 1966 *The Crisis of the Early Italian Renaissance: Civic Humanism and Republican Liberty in the Age of Classicism and Tyranny*. Princeton University Press, Princeton, NJ

Bleicken J 1972 *Staatliche Ordnung und Freiheit in der römischen Republik*. Lassleben, Kallmünz, Germany

Bock G, Skinner Q, Viroli M (eds.) 1990 *Machiavelli and Republicanism*. Cambridge University Press, Cambridge, UK

Constant B 1988[1819] De la liberté des anciens comparée à celle des modernes. In: Fontana B (ed.) *Political Writings*. Cambridge University Press, Cambridge, UK

Cornelius Tacitus P 1906 Ab excessu divi Augusti annalium libri. In: Fisher C D (ed.) Oxford University Press, Oxford, UK [1974 *The Annals*. Penguin, Harmondsworth, UK]

Cornelius Tacitus P 1911 Historiarum libri. In: Fisher C D (ed.) Oxford University Press, Oxford] [1964 *The Histories*. Penguin, Harmonsworth, UK]

Fink Z S 1962 *The Classical Republicans: An Essay in the Recovery of a Pattern of Thought in Seventeenth-Century England*, 2nd edn. Northwestern University Press, Evanston, IL

Fisher H A L 1911 *The Republican Tradition in Europe*. G. P. Putnam's Sons, New York

Fontana B. (ed.) 1994 *The Invention of the Modern Republic*. Cambridge University Press, Cambridge, UK

Gordon T 1728–31 *The Works of Tacitus with Political Discourses Upon that Author,* 4 Vols. Woodward and Peele, London

Gordon T 1746 *The Works of Sallust Translated into English with Political Discourses Upon that Author*. Woodward and Peele, London

Habermas J 1992 *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*. Suhrkamp, Frankfurt, Germany [1996 *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, Cambridge, MA]

Hamilton A, Jay J Madison J ['Publius'] 1787–8 *The Federalist*. [1987 Kramnick I. (ed.) *The Federalist Papers*. Penguin, Harmondsworth, UK]

Harrington J 1656 *The Commonwealth of Oceana*. [1992 Pocock J G A (ed.) Cambridge University Press, Cambridge, UK]

Hobbes T 1651 *Leviathan, or The Matter, Form, and Power of a Commonwealth Ecclesiasticall and Civil*. [1991 Tuck R (ed.) Cambridge University Press, Cambridge, UK]

Kant I 1795 *Zum ewigen Frieden*. F. Nicolovius, Königsberg

Koenigsberger H G (ed.) 1988 *Republiken und Republikanismus im Europa der frühen Neuzeit*. R. Oldenbourg, Munich, Germany

Livius T *Ab urbe condita I–V*. 1974, Ogilvie R M (ed.) Oxford University Press, Oxford, UK [1971 *The Early History of Rome*. Penguin, Harmondsworth, UK]

Liviu, T *Ab urbe condita VI–X*. 1919, Walters C F, Conway R S (eds.) Oxford University Press, Oxford, UK [1982 *Rome and Italy*. Penguin, Harmondsworth, UK]

Macchiavelli N 1517 *Discorsi sopra la prima deca di Tito Livio* [1984 Inglese G (ed.) Rizzoli Editore, Milano] [1970 *The Discourses*. Penguin, Harmondsworth, UK]

Montesquieu C de Secondat, baron de 1734 *Considérations sur les causes de la grandeur des Romains et de leur décadence*. [1967 Truc G (ed.) Editions Garnier, Paris] [1968 *Considerations on the Causes of the Greatness of the Romans and Their Decline*. Cornell University Press, Ithaca, NY]

Montesquieu C de Secondat, baron de 1748 *De l'esprit des lois*. [1973 Derathé R (ed.) 2 Vols. Editions Garnier, Paris] [1989 *The Spirit of the Laws*. Cambridge University Press, Cambridge, UK]

Nicolet C 1982 *L'idée républicaine en France (1789–1924): Essai d'histoire critique*. Gallimard, Paris

Paine T 1792 *The Rights of Man, Part II*. [1989 In: Kuklick B (ed.) *Thomas Paine: Political Writings*. Cambridge University Press, Cambridge, UK]

Pettit P 1997 *Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford, UK

Pettit P 1999 *Postscript to the Paperback Edition of Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford, UK

Plutarchos *Bioi paralleloi*. [1949–1971 In: Lindskog C, Ziegler K (eds.) *Plutarchi Vitae Parallelae*. Teubner, Leipzig] [1914–26 *The Parallel Lives*. Harvard University Press, Cambridge, MA]

Pocock J G A 1975 *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition*. Princeton University Press, Princeton, NJ

Polybios Historia [1866–1868 Dindorf L (ed.) 4 Vols. Teubner, Leipzig] [1922–7 *The Histories*. Harvard University Press, Cambridge, MA]

Powell J G F (ed.) 1995 *Cicero the Philosopher: Twelve Papers*. Clarendon Press, Oxford, UK

Rahe P A 1992 *Republics Ancient and Modern: Classical Republicanism and the American Revolution*. University of North Carolina Press, Chapel Hill, NC

Rawls J 1993 *Political Liberalism*. Columbia University Press, New York

Robbins C 1959 *The Eighteenth-century Commonwealthman: Studies in the Transmission, Development and Circumstance of English Liberal Thought from the Restoration of Charles II until the War with the Thirteen Colonies*. Harvard University Press, Cambridge, MA

Rousseau J-J 1762 *Du contrat social*. [1973 Guillemin H (ed.) Union Générale d' Éditions, Paris] [1997 *The Social Contract*. Cambridge University Press, Cambridge, MA]

Rush B 1777 *Observations upon the Present Government of Pennsylvania in Four Letters to the People of Pennsylvania*. Steiner and Cist, Philadelphia

Sallustius Crispus C *Bellum Catilinae*. [1991 In: Reynolds L D (ed.) *Catilina; Jugurtha; Historiarum Fragmenta Selecta; Appendix Sallustiana*. Oxford University Press, Oxford, UK] [1921 *The War with Catiline*. Harvard University Press, Cambridge, MA]

Sallustius Crispus C *Bellum Iugurthinum*. [1991 In: Reynolds, L D (ed.) *Catilina; Jugurtha; Historiarum Fragmenta Selecta; Appendix Sallustiana*. Oxford University Press, Oxford, UK] [1921 *The War with Jugurtha*. Harvard University Press, Cambridge, MA]

Sandel M J 1996 *Democracy's Discontent: America in Search of a Public Philosophy*. Belknap Press of Harvard University Press, Cambridge, MA

Sellers M N S 1991 Republican impartiality. *Oxford Journal of Legal Studies* **11**: 273–82

Sellers M N S 1994 *American Republicanism: Roman Ideology in the United States Constitution*. Macmillan, Basingstoke, UK

Sellers M N S 1996 Republican principles in international law. *Connecticut Journal of International Law* **11**: 403–32

Sellers M N S 1998 *The Sacred Fire of Liberty: Republicanism, Liberalism and the Law*. Macmillan, Basingstoke, UK

Sidney A 1798 *Discourses Concerning Government*. [1990 West T G (ed.) Liberty Classics, Indianapolis, IN]
Skinner Q 1998 *Liberty Before Liberalism*. Cambridge University Press, Cambridge, UK
Spitz J-F 1995 *La liberté politique: Essai de généalogie conceptuelle*. Presses Universitaires de France, Paris
Sunstein C R 1993 *The Partial Constitution*. Harvard University Press, Cambridge, MA
Symposium: The Republican civic tradition 1988 *The Yale Law Journal* **97**(8)
Trenchard J, Gordon T 1723 *Cato's Letters or Essays on Liberty, Civil and Religious, and Other Important Subjects*, 4 Vols. [1995 Hamowy R (ed.) 2 Vols. Liberty Fund, Indianapolis, IN]
Tullius Cicero M *De legibus*. [1914 Mueller C F W (ed.) Teubner, Leipzig] [1998 *The Laws*. Oxford University Press, Oxford, UK]
Tullius Cicero M *De officiis*. [1994 Winterbottom M (ed.) Oxford University Press, Oxford, UK] [1991 *On Duties*. Cambridge University Press, Cambridge, UK]
Tullius Cicero M *De re publica*. [1969 Ziegler K (ed.) Teubner, Leipzig] [1998 *The Republic*. Oxford University Press, Oxford, UK]
Tullius Cicero M. *M. Antonium orationes Philippicae*. [1986 Fedeli P (ed.) 2nd edn. Teubner, Leipzig] [1926 *Philippics*. Harvard University Press, Cambridge, MA]
Tullius Cicero M *Orationes* in *L. Catilinam*. [1901 Clark A B (ed.) Oxford University Press, Oxford, UK] [1976 In *Catilinam*. Harvard University Press, Cambridge, MA]
Tullius Cicero M *Pro L. Flacco oratio*. [1906 In: Mueller C F W (ed.) *M. Tulli Ciceronis orationes*, *Pro P. Sulla*, *Pro A. Licinio Archia poeta*, *Pro L. Flacco*. Teubner, Leipzig] [1976 *Pro Flacco*. Harvard University Press, Cambridge, MA]
Tullius Cicero M *Pro P. Sulla oratio*. [1906] In: Mueller C F W (ed.) *M. Tulli Ciceronis orationes*, *Pro P. Sulla*, *Pro A. Licinio Archia poeta*, *Pro L. Flacco*. Teubner, Leipzig] [1976 Pro Sulla. Harvard University Press, Cambridge, MA)
Washington G 1789 The First Inaugural Speech. [1988 In: Allen W B (ed.) *George Washington: A Collection*. Liberty Classics, Indianapolis, IN]
Wirszubski C 1960 *Libertas as a Political Idea at Rome during the Late Republic and Early Principate*, 2nd edn. Cambridge University Press, Cambridge, UK
Wood G S 1969 *The Creation of the American Republic, 1776–1787*. University of North Carolina Press, Chapel Hill, NC

M. N. S. Sellers

Reputation

A reputation exists in the form of an image shared within some public. Applied to a person, it identifies 'not an intrinsic quality he possesses but rather the opinion other people have of him' (Bailey 1971). It can also refer to certain collectivities, like a company, occupational category, professional group, organization, or its leadership and, by extension, to just about anything that can somehow be rated or ranked. Thus, we talk about the reputation of a dining establishment, of a vintage year of wine, of an annual event, of a locality or city. All such reputations, whatever they be, are socially constructed. They may or may not be deserved and are always open to challenge. But, as long as they endure, they are real in their consequences.

Over 300 years ago Locke referred to the 'law of opinion and reputation' in order to highlight how much reputations were a part of the small politics of everyday life (1690/1894): 'People have a strong motive to accommodate themselves to the opinions and rules of those with whom they converse … and so they do that which keeps them in reputation with their company, with little regard for the laws of God, or of the magistrate …' Smith, a century later, argued in similar vein when he located the love of virtue in the desire for 'being what ought to be approved of,' a motive he believed to be most strongly felt among persons of middling or inferior status who, possessing none of the privileges associated with noble lineage or inherited wealth, must, in order to distinguish themselves, bring their talents into public view. Besides, they 'can never be great enough to be above the law … (Their) success … almost always depends upon the favour and good opinion of their neighbours and equals; and without a tolerably regular conduct, these can very seldom be obtained' (1804).

These ideas persist. Economists represented in a volume edited by Klein (1997) depict society as permeated by a flowing network of reputational relationships, with institutional credentials and certification substituting for reputations based on personal acquaintance.

Insofar as one can extract value from it, a reputation can be further conceptualized as a form of symbolic capital (Bourdieu 1992). It can grow or diminish yet still retain some value so long as people remember. This value is real enough for anyone who feels his or her good name to have been unjustly sullied by the careless or malicious actions of others. Redress can be sought through the courts—not always successfully, to be sure. And tax laws in some countries allow the corporate write-off of expenditures for building and protecting the reputation of a company or its brand name.

Reputations vary along three dimensions: best known is the vertical dimension that locates an agent or object on a hierarchical scale. A high reputation signifies the attribution of a level of respect, credibility, authority, or competence denied to those who do not measure up to that standard. Reputations, high as well as low, also vary along a lateral dimension, where placement is determined by the size of the public within which a name has currency. To differentiate between these two dimensions, we refer to a reputation for upright character or some talent shared by neighbors, colleagues, professional peers, and other insiders as *recognition* while reserving the term *renown* for that more universal recognition that extends beyond a particular reputational community (Lang and Lang 1990). A parallel distinction at the low

Table 1

Vertical	**Lateral (Expansion)**		**Temporal (Duration)**	
(Status)	Insiders	Public Sphere	Short	Long
High	*Recognition*	*Renown*	*Celebrity*	*Eminence*
Low	*Discredit*	*Notoriety*	*Ignominy*	*Infamy*

end of the vertical dimension is between being in *disrepute* among peers and the *notoriety* attained when disreputing information carries to the general public. The terms are diagramed in the Table 1 below.

A third *temporal* dimension has to do with duration in time. Public attention is forever being distracted by new entrants into the limelight. Reputations follow trajectories of different lengths. A celebrity is typically a person who may enjoy moments of fame but thereafter is remembered only by the few with a special interest in preserving the memory. Only the 'truly great' achieve the kind of *eminence* that accounts for their survival in the collective memory as major historical figures (Collins 1998). Alternatively, the ignominy experienced for some heinous deed may be quickly forgotten or be forever remembered in infamy. Most reputations fall somewhere between these extremes. The categories are depicted in the accompanying figure.

Sociologists, and especially historians as well as some literary scholars, have focused mainly on the lateral and temporal dimensions of high status, namely on how particular individuals attain renown and are canonized. Economists and political scientists, many of them employing game theoretical models, as well as some anthropologists have exhibited more interest in how reputations function within particular settings.

1. *Effects of Reputation*

Effects can be divided into two general categories: those involving the reactions of others to the reputation of an individual and those that stem from an individual's anticipation of reactions to their own conduct. Obviously, there is a good deal of overlap between these two.

The effect of a personal reputation on those with whom one directly interacts has been illustrated famously in the fable about the boy who cried 'Wolf' once too often. When a wolf actually showed up, as we all have learned, no one in the village heeded his warning, having come to expect just another false alarm. So it is that signals from a source deemed untrustworthy are apt to be ignored and dealings with a party of dubious reliability will, wherever possible, be avoided. But we do not always know enough about the other nor are we at liberty to choose. Where such doubts exist, the willingness to co-operate or enter into an agreement becomes contingent on external enforcement.

How reputations derived from past performances of parties engaged in repetitive play affect both the strategies adopted by competitors and their preferences for partners has been modeled by game theorists. Their highly suggestive work is commendable for its logical rigor. Still, one can question how well their propositions tested in experiments, most of them with only two players interacting within controlled settings, replicate the complexities of real life. For one thing, reputational effects are based not just on personal experience but on word that gets around. Thus, collective attributions are a cardinal factor in the power an American president can wield. Neustadt locates its source, not just in the Constitution but in what others in the Washington political community think and expect of the incumbent. Justified or not, 'it is the reality on which they act, at least until their calculations turn out wrong … (1990). What the president must demonstrate in the many different games he plays, often all at the same time, is his ability and will to make use of his bargaining advantages. Success breeds success but a pattern of repeated failure is bound to result in a loss of faith in his effectiveness that will reduce his power to persuade.

Similar self-enhancing effects show up in nonpolitical settings. Distinguished scientists get disproportionate credit for their contribution to a collaborative project (Merton 1967). The prestige of a university affects the quality of the faculty it attracts, whose members then, in turn, benefit from their position and raise the prestige of the institution; artists who are already well-known get more exhibitions; the works of writers who have already won prizes are more actively promoted. The epitomizing case is that of a celebrity, described by Boorstin (1961) as someone 'known for his being known.'

Images may be poor substitutes for more direct knowledge; yet in a world where most people are strangers to one another, there may be nothing else to govern our calculations. Images certainly affect what we buy, how we invest, and how much we are willing to pay (Fombrun 1996). The reassurances provided by an established brand name usually exacts some premium. But nowhere are the returns from a good

reputation greater than in the arts, where values are somewhat intangible, subject to fashion, highly dependent on critical and public endorsement (Moulin 1986). The name attached to a work determines not only price but also the attention it gets. The media, very much oriented toward a star system, feed this proclivity, thereby contributing to a pattern in which small differences among a number of top performers, none of them unequivocally superior to the rest, translate into large differences in acclaim (Goode 1978; Frank and Cook 1995).

Even in the corporate world the relation of reputation to measurable performance is not as close as it may seem. High publicity given the Valdez oil spill in Alaska waters was followed by a significant short-term decline in the value of Exxon stock. At least some investors, it has been found, back away from companies not so much because of any loss of faith in their future performance but doubts about their commitment to general community values (Fombrun 1996).

The fear of damage to one's reputation becomes an incentive for behavior against one's immediate self-interest. What economists, following Adam Smith, have called the discipline of repeated dealings acts as a deterrent, not always effective, against business selling shoddy goods to naive customers or against lobbyists providing clearly misleading information to legislators. But this response is not simply a matter of honesty but of strategic impression management as, for example, when a college reduces its acceptance rate as part of a marketing strategy to portray itself as highly selective. On the international scene, where to yield 'on a series of issues when matters at stake are not critical … may (make) it difficult to communicate to the other (side) just when a vital issue has been reached,' a nation may be moved to act for no other reason than to retain its 'reputation for action' (Schelling 1966). Against such an instrumental view, Chong (1992) has argued that, when a reputation becomes part of their identity, people try to live up to it for moral reasons as well, trying to turn themselves into the kind of person they want others to think they are.

Another set of reputational incentives comes into play when the expansion of a reputation into the public sphere creates what Braudy (1986) calls the 'frenzy of renown' with numerous effects, some of which are potentially deleterious to one's moral commitments, whatever these may be. Fame and the appearance of power it gives are sometimes seductive. For example, Gitlin (1980) describes how media attention undermined the connection of some leading activists of the New Left, a movement with a weak organizational structure, to their original reputational community. Competing for a place in the limelight, these 'leaders' transformed themselves into 'unattached celebrities,' no longer accountable to the rank-and-file but drawn toward increasingly 'extravagant, "incidental," expressive actions—actions which made "good copy" because they generated sensational pictures rich in symbolism.' In sports, the advent of a national public for major-league baseball, largely a product of television, has been associated with a decline in the advantage of playing on one's home turf. The present-day public, according to Leifer (1995), is less loyal to home-grown athletes and teams but looks for any winner it can celebrate. It is not averse to abandoning a team that falls short of its expectations. By interpreting what were initially random performance differences between contestants, the public casts competitive equals into stable winner and loser roles. The teams react to reputations established early in the season, not only by the public at the game but also by sportscasters and writers when in fact, Leifer contends, many outcomes might easily have gone the other way.

2. *From Recognition to Renown*

Propositions about the acquisition of fame derive largely from life histories, mostly of successful individuals who by virtue of their background, talent, or associations seemed destined for success. To gain recognition within some domain for some special talent or accomplishment depends on a number of general conditions. Endowment and ability are, without a doubt, prerequisites. But the pool of talent is large and some acclaimed figures have had to overcome what would appear to have been formidable handicaps. Albert Einstein is known to have failed his school examination and Glenn Cunningham, the champion miler for several years, had legs so severely burned as a child that his doctors were not at all sure whether he would even walk again. Regardless of any such handicap, everyone in pursuit of recognition has to do what is necessary to develop the skills that count.

There is, to be sure, a world of difference between measurable improvements of athletic performance or scores in games played according to strict rules and the evolving consensus by which the less tangible abilities prized in domains other than sports come to be validated. Nothing counts unless people take notice. The achievement of a recluse working all alone is more likely to be overlooked than comparable ones by more gregarious contemporaries. Insofar as reputations carry through networks, acceptance by peers and mentors is often the first step toward gaining recognition. Thereafter it becomes a question of how far these networks extend. Persons highly esteemed in one domain may not even be recognized outside its boundaries or can be subject to all sorts of invidious distinctions. Hence, reputations do not easily transfer, even between genotypically similar activities, especially if they are as differentially valued. In most arts, where only some genres are ordained by an arts establishment, an excess of popularity can actually be the kiss of death. Kapsis (1992) traces the career of

Alfred Hitchcock, the Hollywood director of many famous films, who transformed himself deliberately from a merely successful maker of movies into a genuine *auteur by* adopting a new and more acceptable idiom.

The general rule that one is better off not straying too far from the mainstream holds even in the sciences where the highest honors usually go to the innovators. To be working on a problem whose importance has been widely acknowledged increases the likelihood of gaining attention for work that adds only incrementally to existing knowledge. Attaching oneself to important people is a second strategy, equally appropriate to the sciences and to the arts. Yet dissidents denied their place by a complacent establishment can also collectively advance their reputations by banding together, forming their own societies, and using them as a springboard to raise their status. This is how the group of artists denigrated as 'impressionists,' due to their apparent disdain for serious subject matter, drew attention to themselves.

Whether recognition within a group converts into renown is highly contingent on distribution and publicity (Becker 1982). Visual artists depend on exhibition space for display of their accomplishments; writers on publishers and booksellers; playwrights on performers and on a suitable stage, and the performers themselves on recordings and bookings in a major medium to display their accomplishments. To be singled out for special praise by influential critics and other mediators of fame is still more important (Moulin 1986). Especially if their judgments stimulate museum purchases, give a theatrical work a permanent place in the repertory, or generate substantial interest in the writings of an author, they can make a household word out of an obscure name. Most likely to bring instant fame are the feats that receive the greatest publicity, like the breaking of an established record in one of the major sports or ventures into entirely new territories in science or in space. A few individuals are singled out as stars with very little acknowledgment of the collective effort that made their achievement possible. Only a small number of scientists know the names of the many highly competent engineers who have make space flight possible.

3. *Measuring Reputations*

In a society set on rating and ranking everything, there are essentially two approaches to the measurement of reputation: through responses to survey questions and through aggregations of indicators about the actual reactions of some public.

Judgments expressed by survey respondents are most meaningful in settings where just about everyone knows everyone else. Accordingly, performance can be predicted from measures of certain qualities based on average ratings by groups of their peers. No less ascertainable is what the residents of a village community think of their neighbors or which professionals are held in the highest regard by their colleagues. However, the content of any such consensus becomes more problematical the farther removed those queried are from whomever they are asked about. For years pollsters have obtained what legitimately can be called a reputational ranking by asking a representative sample of Americans to name the man and also the woman they most admire. Because respondents were free to name anyone alive, their choices have always been widely scattered with the president and the first lady almost always heading the lists. Their rankings, as well as that of others, has have as much to do with their prominent coverage in the news media as with the respect they enjoy within their particular reputational community.

The American president is of course one of a very small number of celebrities or infamous villains about whom just about everyone has some opinion. Thus, standardized questions about presidential performance approximate a vote and can be the basis for comparing a president's reputation to that of other actors prominent on the political scene. But there is as yet no general agreement on a standard for measuring other reputations in the public sphere. Should the reputations of corporations be based on the collective judgment of high-level executives, of customers, or of their employees? That of schools on responses from their alumni or of academics from those working in the same discipline? Who does the ranking is particularly important in the arts, governed as they are by the distinction between real 'quality' ascertained by art experts and what is merely popular.

When it comes to the past, there is no alternative to querying specialists. This method works best when evaluating a narrowly defined category, like American presidents or justices of the Supreme Court. In the first such survey, 55 experts on the presidency were asked to place each occupant of the office into one of five categories ranging from 'great' to 'failure' (Schlesinger 1997). Others have queried larger samples with more 'refined' methods only to find that the same three—Washington, Lincoln, and Franklin D. Roosevelt—consistently placed at the top, though not always in the same order. Choices of the best and worst presidents have also remained relatively stable with more fluctuation among those ranked in between. Most striking has been the steady rise in reputation of Eisenhower as new information about his presidency has come to light.

Citation counts are one of the most widely used indicators of the reputation of individuals and of academic departments. Such information is very accessible. Because scholars are obliged to cite prior publications on which their own work is built or with which they take issue, the number of references to the work of a scientist or scholar in any particular

discipline is assumed to indicate the level of recognition by colleagues. Yet strategic considerations also enter. Citing a particular predecessor can help a neophyte establish a link to one or more of the prestigious figures in their field, thereby drawing the attention of colleagues and enhancing the significance of one's own work (Camic 1987). The more a field is subject to intellectual fashion the more likely it is that credit for an idea that is very much in the air goes to the more famous.

The frequency with which a particular name appears in the more popular media indicates its role in a broader discourse but there are, at least in some fields and for some individuals, serious discrepancies between name counts for introductory texts and those for more esoteric publications aimed at narrower constituencies. In drawing conclusions from any such enumerations, one needs always to take into account the perspective of publications examined. But any set of similar publications can be useful in examining trends. Rosengren (1987) noted that in texts of Swedish literature the same names that turned up in an earlier period were still prominently reviewed four decades later.

A simple way of gauging the level of eminence (or infamy) attained by any individual is through the cumulation of literature in major libraries, most of whose holdings are now on line. After all, writers are the primary purveyors of fame (Cooley 1927). The number of items a search turns up is a measure of how much what the subjects did and the issues they raised have retained their salience as historical markers. The larger the universe searched the more finely do the findings discriminate. A search limited to full-length biographies would omit many minor figures somewhat renowned in their lifetime. Searches covering long time spans are open to another kind of bias. Although material evidence gets lost over time and some of the eminent of long ago may have left few traces, posterity has had many more opportunities to develop a consensus about those whose 'lives' have somehow survived. In terms of number of mentions and pages devoted to each in an authoritative encyclopedia of communication, a few 'greats' from the past outranked even the most distinguished contemporary scholars in this burgeoning field (Beniger 1990). It therefore makes sense to restrict such counts to differentiating between the luminaries of a particular time period, whose ideas remain of great interest, and their contemporaneous lesser lights as Collins (1998) does in his study of philosophers.

Another measure focuses less on citations than on decisions made over the years by the custodians of tradition as to what part of the past to preserve. The results are manifest in the number of memorials, shrines, plaques, exhibits, and commemorative events designed to keep alive the memory of a particular person or persons. To obtain a full record of these tributes is difficult, except perhaps for some of the arts where one can ascertain the number of times a composer's work is performed by major orchestras (Simonton 1986; Mueller 1958) or a play is performed or revived (Griswold 1987). In the visual arts, the aggregate number of works in major museums and on permanent display is an appropriate indicator of collective curatorial judgment but somewhat confounded by rarity. Because there is not enough work by some of the greatest artists to meet the demand of museums, a single indicator, based on their holdings, has to be supplemented by information about the frequency with which work is circulated in special loan exhibitions, many of which are organized precisely to commemorate the artist.

The market is the ultimate but albeit imperfect arbiter of value. Many variables, like age, rarity, and condition, enter into prices paid for works and memorabilia associated with past eminence. Still, for items comparable on these variables, consistent price differentials represent the consensus of their relative worth. They are, to this extent, valid measures of reputation.

4. *The Survival of Reputation*

Over time, as memories fade, the lives of most people end up in the dust bins of history. There is a continuing sorting out of those deemed memorable. As to likely survival in the collective memory, posthumous reputations are the combined effect of two conditions: the rhetorical force of the accomplishments of the deceased on society or some domain in their lifetime and the extent to which, even if for altogether different reasons, an accomplishment still touches the imagination of posterity. 'Greatness' can also be conferred retrospectively as validation of or correction to earlier valuations.

The renown achieved in a person's lifetime enhances the chance for reputational survival. It facilitates contact with persons who become the mediators of one's reputation across generations (Zilsel 1972). Kings, popes, presidents, generals, and executives may have archivists at their command but those without an official position have to find their place in the right networks. Attention from colleagues attracts students, who will pass their valuations of a scholar on to the next generation. For an artist, critical acclaim brings wealthy collectors, whose purchases have a good chance of ending up in a museum. It is by no means uncommon for the admirers of some creative person to form a society dedicated to preserving and promoting their reputation.

Famous people also leave more traces. Anything connected with them gains value, which provides incentives for preserving mementoes of their lives, especially any pictures, diaries, and correspondence that evidence their thoughts and character. Insofar as their paths have crossed those of other famous people,

many of the letters they have written or received are likely to survive in some archive and to attract scholars prepared to ferret out still more material.

Subjects have some, though limited, control over the record on which their reputational fate will come to depend. Those who take care to keep and organize records strengthen their claim on the memory of posterity. It has become standard for a president to initiate the setting up of a library, even though his papers would be saved in any event. Some famous people are known to have taken the further step of purging from their files items inconsistent with the image they want to project. Writers have burned manuscripts they believed detrimental to their reputation. There are painters who destroyed systematically all pictures produced during a period for which they did not want to be remembered. Theodore Roosevelt dictated letters intended for posterity. Thomas Hardy actually controlled the biography his wife published under her own name following his death (Hamilton 1994). President Richard Nixon's attempt to retain control over his secretly recorded tapes failed. The same impulse shows up in instructions to estate executors to convert the residence of the deceased into a landmark or to use remaining assets to preserve his or her reputation. To be sure, these wishes are not always followed or followed only half-heartedly.

Some factors affecting the survival of reputation remain beyond the control of the individual: first, there are the accidental circumstances of birth and death. Possible candidates for reputational immortality can be born into a time or live in a place where no one acts to preserve the record of their deeds. Thus, the Chinese predecessor of Gutenberg remains unknown and Columbus rather than Leif Ericson is hailed as the discoverer of America. As to the timing or circumstances of death, a long productive life full of honors is an obvious reputational asset but to be struck down in the prime of life by an assassin, a tragic illness, or unfortunate accident has the potential to transform the victim into an icon. No one can possibly know how John F. Kennedy would have been remembered had he served for a full one or two terms. Yet the halo put around him by his untimely death overshadows the modest achievements and all too apparent failures of his administration. A still more striking example is that of Anne Frank. Would her diary, translated into many languages, have had the same impact if she had survived the Holocaust? And had she survived, would she have tolerated the publication by her father of this very private document from which he excised passages that displeased him?

The availability of survivors, like Anne Frank's father, with a personal stake in cultivating remembrance depends, in part, on such demographic factors as marriage, fertility, and longevity. Lang and Lang (1990), studying the revival of etching as an original art form in the late-nineteenth century, found that it had attracted an unusually large number—at least for that time period—of talented women well recognized during their lifetimes, yet today much less well remembered than their male peers. Two demographic circumstances contributed to this reputational disparity: the extraordinary longevity of the women and the much smaller number ever married or, if married, with children. Hence, they outlived, often by decades, not only their spouses but also their colleagues, close friends, publishers, dealers, admiring collectors, as well as the collapse of the etching craze, so that there was often no one still around to mourn their death and write the obligatory obit. Also, because unmarried women had neither mate nor daughter to take on the tedious task of record keeping, material for an artistic biography or definite catalogue raisonné are often lacking.

All reputational communities have boundaries. The heroes celebrated in one may be largely unknown in another. Nowhere is this more obvious than in the different statesmen, military heroes, literary figures and so forth each nation celebrates. National perspectives even intrude into the allocation of credit for achievement universally recognized as breakthroughs. Obviously, the pioneers of space exploration who have received the greatest attention in America differ from those celebrated in the former Soviet Union. The inventor of the incandescent lamp, as remembered by Americans, is Thomas Edison, who was in fact preceded by Joseph Swan, who first exhibited his successful carbon-filament lamp in England, where he had greater impact. His factory near Newcastle rivaled Edison's in prosperity. But whereas Edison immediately patented everything, Swan was slow to do so. Nor did he have a wealthy sponsor like Henry Ford, who thought him the greatest man alive and started collecting Edison memorabilia as early as 1905, and ultimately erecting a shrine in Dearborn, Michigan by transporting, piece by piece, and reassembling the inventor's New Jersey laboratory. But for a reputation to cross all cultural and geographic boundaries, the achievement behind it must appear unique and be an inspiration for people all over the world.

Time creates its own boundaries. There is only so much room at the top, so that at some point what Collins (1998) calls the law of small numbers comes into play. Secondary figures recede into the background as new interests emerge. The survivors are those whose deeds and accomplishments still speak to the present. Posterity has some leeway in its view of the past. Not only does it select but it reinterprets continuously, particularly when it comes to the work of writers and artists. The qualities that contemporaries most admired in Hawthorne's oeuvre (Tompkins 1984) and in Da Vinci's Mona Lisa are not the same ones that touch us today (Boas 1940). Put another way, eminence derives from the persistence of beliefs about persons and their deeds and creations that have changed their domain in some fundamental way. The

larger the number of domains affected, the greater the likelihood that their reputations will live on.

The development, change, preservation, and decline of reputations as well as how they function have to be understood as part of an ongoing social process. Experiments and statistical studies provide useful information, mostly about outcomes, but require supplementation of case material on the process as it unfolds in a particular context. Qualitative evidence, when derived from more than one case and subjected to systematic analysis, is not to be discounted as merely anecdotal.

See also: Celebrity; Charisma and Charismatic; Charisma: Social Aspects of; Potlatch in Anthropology; Sport, Sociology of

Bibliography

Bailey F G 1971 Gifts and poison. In: *Gifts and Poison: The Politics of Reputation*. Blackwell, Oxford, UK, pp. 1–25

Becker H 1982 *Art Worlds*. University of California Press, Berkeley, CA

Beniger J R 1990 Who are the most important theorists of communication. *Communication Research* **17**: 698–715

Boorstin D J 1961 *The Image: Or, What Happened to the American Dream*. Atheneum, New York

Bourdieu P 1992 *Les règles de l'art*. Editions du Seuil, Paris

Braudy L 1986 *The Frenzy of Renown*. Oxford University Press, New York

Camic C 1987 The making of a method: A historical reinterpretation of early Parsons. *American Sociological Review* **52**: 421–39

Chong D 1992 Reputation and cooperative behavior. *Social Science Information* **31**: 683–709

Collins R 1998 *The Sociology of Philosophy: A Global Theory of Intellectual Change*. Belknap Press, Cambridge, UK

Cooley C H 1927 Fame. In: *Social Process*. Scribner's, New York, pp. 112–24

Fombrun C J 1996 *Reputation: Realizing the Value from the Corporate Image*. Harvard Business School Press, Boston

Frank R H, Cook P J 1995 *The Winner-take-all Society: How More and More Americans Compete for Ever Fewer and Bigger Prizes, Encouraging Economic Waste, Income Inequality, and an Impoverished Cultural Life*. Free Press, New York

Gitlin T 1980 *The Whole World is Watching: Mass Media in the Making & Unmaking of the New Left*. University of California Press, Berkeley, CA

Goode W J 1978 *The Celebration of Heroes: Prestige as a Social Control System*. University of California Press, Berkeley and Los Angeles, CA

Griswold W 1986 *Renaissance Revivals: City Comedy and Revenge Tragedy in the London Theatre, 1576–1980*. University of Chicago Press, Chicago

Hamilton I 1994 *Keepers of the Flame: Literary Estates and the Rise of Biography from Shakespeare to Plath*. Faber & Faber, Boston

Kapsis R E 1992 *Hitchcock: The Making of a Reputation*. University of Chicago Press, Chicago

Klein D B (ed.) 1997 *Reputation: Studies in the Voluntary Elicitation of Good Conduct*. University of Michigan Press, Ann Arbor, MI

Lang G E, Lang K 1990 *Etched in Memory: The Building and Survival of Artistic Reputation*. University of North Carolina Press, Chapel Hill, NC

Leifer E M 1995 *Making the Majors: The Transformation of Team Sports in America*. Harvard University Press, Cambridge, MA

Locke J 1690/1894 *An Essay Concerning Human Understanding*. Clarendon Press, Oxford, UK

Merton R K 1967 The Matthew effect in science. *Science* **159**(5): 56–63

Mueller J H 1958 *The American Symphony Orchestra: A Social History of Musical Taste*. Calder, London

Neustadt R E 1990 *Presidential Power and the Modern Presidents: The Politics of Leadership from Roosevelt to Reagan*. Free Press, New York

Rosengren K E 1987 Literary criticism: Future invented. *Poetics* **16**: 295–325

Schelling T C 1966 *Arms and Influence*. Yale University Press, New Haven, CT

Schlesinger A M Jr 1997 Rating the presidents: Washington to Clinton. *Political Science Quarterly* **112**: 179–90

Smith A 1804 *The Theory of Moral Sentiments*. Cadell & Davis, London

Tompkins J 1984 Masterpiece theatre: The politics of Hawthorne's literary reputation. *American Quarterly* **36**: 617–42

Zilsel E 1926/1972 *Die Entstehung des Geniebriffes: Ein Beitrag zur Ideengeschichte der Antike und des Frühkapitalismus*. Olms, Hildesheim, Germany

K. Lang and G. E. Lang

Resampling Methods of Estimation

Most scientists occasionally or frequently face problems of data analysis: what data should I collect? Having collected my data, what does it say? Having seen what it says, how far can I trust the conclusions? Statistics is the mathematical science that deals with such questions. Some statistical methods have become so familiar in the scientific literature, especially linear regression, hypothesis testing, standard errors, and confidence intervals, that they seem to date back to biblical times. In fact most of the 'classical' methods were developed between 1920 and 1950, by scientists like R. A. Fisher, Jerzy Neyman and Harold Hotelling who were senior colleagues to statisticians still active today.

The 1980s produced a rising curve of new statistical theory and methods, based on the power of electronic computation. Today's data analyst can afford to expend more computation on a single problem than the world's yearly total of statistical computation in the 1920s. How can such computational wealth be spent wisely, in a way that genuinely adds to the

classical methodology without merely elaborating it? Answering this question has become a dominant theme of modern statistical theory.

1. The Bootstrap

The following fundamental question arises in almost every statistical data analysis: on the basis of a data set $\mathbf{z}$ we calculate a statistic $s(\mathbf{z})$ for the purpose of estimating some quantity of interest. For example, $\mathbf{z}$ could be the nine cholesterol reduction scores shown in Table 1, and $s(\mathbf{z})$ their mean value $\bar{z} = 28.58$, intended as an estimate of the true mean value of the cholesterol reduction scores. (The 'true mean value' is the mean we would obtain if we observed a much larger set of scores.) How accurate is $s(\mathbf{z})$?

This question has a simple answer if $s(\mathbf{z})$ is the mean $\bar{z}$ of numbers $z_1, z_2, \ldots, z_n$. Then the *standard error* of $\bar{z}$, its root mean square error, is estimated by a formula made famous in elementary statistics courses.

$$\left[\sum_{i=1}^{n} (z_i - \bar{z})^2 / (n \cdot (n-1)) \right]^{1/2} \tag{1}$$

For the nine numbers in Table 1, Eqn. (1) gives 10.13. The estimate of the true cholesterol reduction mean usually would be expressed as 28.58 ± 10.13, or perhaps $28.58 \pm 10.13 \cdot s$, where s is some constant like 1.645 or 1.960 relating to areas under a bell-shaped curve. With $s = 1.645$, interval (2) has approximately 90 percent chance of containing the true mean value.

The bootstrap, (Efron 1979), was introduced primarily as a device for extending Eqn. (1) to estimators other than the mean. For example, suppose $s(\mathbf{z})$ is the '25 percent trimmed mean' $\bar{z}\{.25\}$, defined as the average of the middle 50 percent of the data: we order the observations $z_1, z_2, \ldots, z_n$, discard the lower and upper 25 percent of them, and take the mean of the remaining 50 percent. Interpolation is required for cases where $0.25 \cdot n$ is not an integer. For the cholesterol data, $\bar{z}\{.25\} = 27.81$.

There is no neat algebraic formula like Eqn. (1) for the standard error of a trimmed mean, or for that matter almost any estimate other than the mean. That is why the mean is so popular in statistics courses. In lieu of a formula, the bootstrap uses computational power to get a numerical estimate of the standard error. The bootstrap algorithm depends on the notion of a *bootstrap sample*, which is a sample of size n drawn *with replacement* from the original data set $\mathbf{z} = (z_1, z_2, \ldots, z_n)$. The bootstrap sample is denoted $\mathbf{z}^* = (z_1^*, z_2^*, \ldots, z_n^*)$. Each z_i^* is one of the original z values, randomly selected, perhaps $z_1^* = z_7$, $z_2^* = z_5$, $z_3^* = z_5$, $z_4^* = z_9$, $z_5^* = z_7$, etc. The name *bootstrap* refers to the use of the original data set to generate 'new' data sets $\mathbf{z}^*$.

The bootstrap estimate of standard error for $\bar{z}\{0.25\}$ is computed as follows: (a) a large number 'B' of independent bootstrap samples each of size n, is generated using a random number device; (b) the 25 percent trimmed mean is calculated for each bootstrap sample; (c) the empirical standard deviation of the B bootstrap trimmed means is the bootstrap estimate of standard error for $\bar{z}\{0.25\}$. Figure 1 shows a schematic diagram of the bootstrap algorithm, applied to a general statistic $s(\mathbf{z})$.

Trying different values of B gave these bootstrap estimates of standard error for the 25 percent trimmed mean applied to the cholesterol data

B:	25	50	100	200	400
bootstrap estimate of standard error:	12.44	9.71	11.50	10.70	10.48

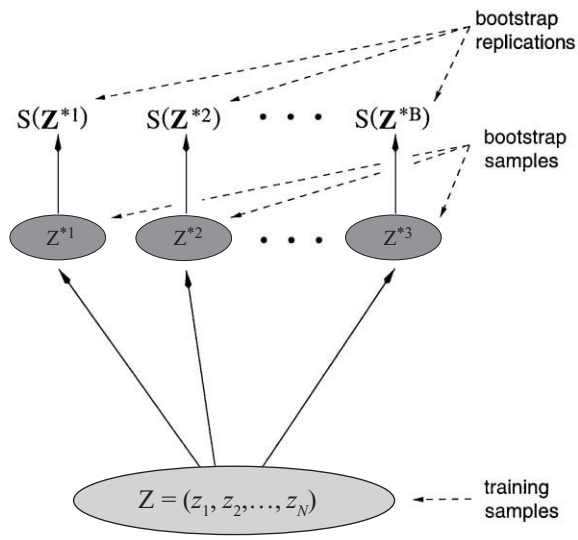

Figure 1
A diagram of the bootstrap algorithm for estimating the standard error of a statistic $s(\mathbf{z})$; each of the B bootstrap samples is a random sample of size n drawn with replacement from the original data set $(z_1, z_2, \ldots, z_n)$

Table 1
Nine men's cholesterol reduction scores after taking cholostyramine; the nine scores are an ordered random sample from the scores of 164 men (data from the Stanford arm of the LRC-CPPT experiment, courtesy of Drs D. Feldman and J. Farquhar, Stanford University (see Efron and Feldman 1991).

−21.0	3.25	10.75	13.75	32.50	39.50	41.75	56.75	80.0

Ideally we would like B to go to infinity. However, it can be shown that randomness in the bootstrap standard error that comes from using a finite value of B is usually negligible for B greater than 100. 'Negligible' here means small relative to the randomness caused by variations in the original data set $\mathbf{z}$. Even values of B as small as 25 often give satisfactory results. This can be important if the statistic $s(\mathbf{z})$ is difficult to compute, since the bootstrap algorithm requires about B times as much computation as $s(\mathbf{z})$.

The bootstrap algorithm can be applied to almost any statistical estimation problem. (a) The individual data points z_i need not be single numbers; they can be vectors, matrices, or more general quantities like maps or graphs. (b) The statistic $s(\mathbf{z})$ can be anything at all, as long as we can compute $s(\mathbf{z}^*)$ for every bootstrap data set $\mathbf{z}^*$. (c) The data set '$\mathbf{z}$' does not have to be a simple from a single distribution. Other data structures, for example, regression models, time series, or stratified samples, can be accommodated by appropriate changes in the definition of a bootstrap sample. (d) Measures of statistical accuracy other than the standard error, for instance biases, mean absolute value errors, and confidence intervals, can be calculated at the final stage of the algorithm. These points are discussed in Efron and Tibshirani (1993). The more complicated example below illustrates some of them.

There is one statistic $s(\mathbf{z})$ for which we don't need the computer to calculate the bootstrap standard error, namely the mean $\bar{z}$. In this case it can be proved that as B goes to infinity, the bootstrap standard error estimate goes to $\sqrt{(n-1)/n}$ times Eqn. (1). The factor $\sqrt{(n-1)/n}$, which equals 0.943 for $n = 9$, could be removed by redefining the last step of the bootstrap algorithm, but there is no general advantage in doing so. For the statistic $\bar{z}$, using the bootstrap algorithm is about the same as using Eqn. (1).

At a deeper level it can be shown that the logic that makes Eqn. (1) a reasonable assessment of standard error for $\bar{z}$, applies just as well to the bootstrap as an assessment of standard error for a general statistic $s(\mathbf{z})$. In both cases we are assessing the standard error of the statistic of interest by the 'true' standard error that would apply if the unknown probability distribution yielding the data exactly equalled the empirical distribution of the data. The efficacy of this simple estimation principle has been verified by a large amount of theoretical work in the statistics literature of the 1990s (see the discussion and references in Efron and Tibshirani 1993).

Why would we consider using a trimmed mean rather than $\bar{z}$ itself? The theory of *robust* statistics, developed since 1960, shows that if the data $\mathbf{z}$ comes from a long-tailed probability distribution, then the trimmed mean can be substantially more accurate than $\bar{z}$, that is, it can have substantially smaller standard error (see, e.g., Huber 1964). The trouble, of course, is that in practice we don't know *a priori* whether or not the true probability distribution is long-tailed. The bootstrap can help answer this question.

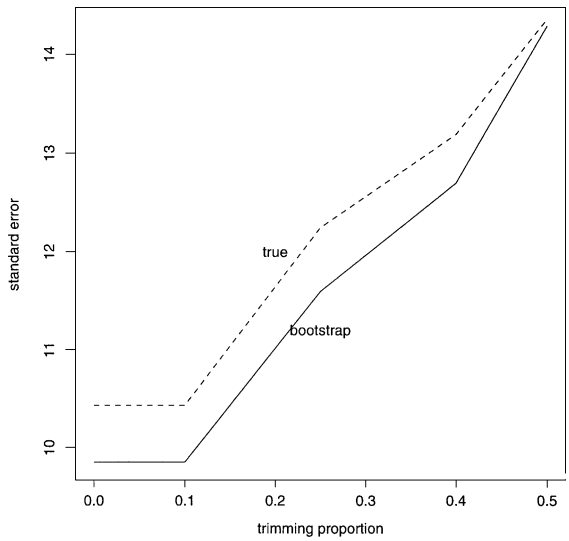

Figure 2
Bootstrap estimates of standard error for five different trimmed means $\bar{z}\{p\}$, $p = 0.\ 0.10, 0.25, 0.40, 0.5$, applied to the cholesterol data of Table 1. Based on $B = 400$ bootstrap samples. Also shown is the true standard error of $\bar{z}\{p\}$, obtained by taking random samples of size nine from the population of 164 cholesterol reduction scores. In this case the bootstrap correctly indicates that $\bar{z}\{0\}$, the ordinary mean, gives the smallest standard error

Figure 2 shows the bootstrap estimates of standard error for 5 different trimmed means $\bar{z}\{p\}$, 'p' being the proportion of the data trimmed off each end of the sample before the mean is taken. (So $\bar{z}\{0\}$ is $\bar{z}$, the usual mean, while $\bar{z}\{0.5\}$ is the median.) These were computed using the bootstrap algorithm in Fig. 1, $B = 400$, except that at step 2 five different statistics were evaluated for each bootstrap sample $\mathbf{z}^*$, namely $\bar{z}\{0\}$, $\bar{z}\{0.10\}$, $\bar{z}\{0.25\}$, $\bar{z}\{0.40\}$, and $\bar{z}\{0.50\}$.

According to the bootstrap standard errors in Fig. 2, the ordinary mean has the smallest standard error among the five trimmed means. This seems to indicate that there is no advantage to trimming for this particular data set.

In fact the nine cholesterol reduction scores in Table 1 were a random sample from a much bigger data set: 164 scores, corresponding to the 164 men in the Stanford arm of a large clinical trial designed to test the efficiency of the cholesterol-reducing drug called cholestyramine (Efron and Feldman 1991). With all of this extra data available, we can check the bootstrap standard errors. The solid line in Fig. 2 indicates the true standard errors for the five trimmed means, 'true'

meaning the standard error of random samples of size nine taken from the population of 164 scores.

We see that the true standard errors confirm the bootstrap conclusion that the ordinary mean is the estimator of choice in this case. The main point here is that the bootstrap estimates use only the data in Table 1, while the true standard errors require extra data that usually isn't available in a real data analysis problem.

2. Growth Curves

Longitudinal data are often modeled using growth curves. Our example uses educational testing data from students in North Carolina (see Williamson et al. 1991, Rogosa and Saner 1995). The data we use for the example are eight yearly observations on achievement test scores in mathematics Y for 277 females, each followed from grade 1 to grade 8. The straight-line growth curve model for each individual can be written as

$$Y_p(t) = \xi_p(0) + \theta_p t + \varepsilon_{pt} \tag{2}$$

where the observable scores Y contain error of measurement indicated by ε_{pt}. Here t denotes time and p denotes person. The constant rate of change for each person is indicated by θ_p; θ_p has a distribution over persons with mean μ_θ and variance σ^2_θ.

To illustrate bootstrap estimation we pick out two of the questions that are investigated using the growth curve models:

(a) Can individual change be measured reliably? In behavioral and social science research the reliability index for the estimated rate of change, $\rho(\hat{\theta})$, is the parameter of interest.

(b) Is there a relation between initial standing (math achievement in grade 1) and the rate of improvement? The parameter to be estimated is correlation between θ and initial status (over the population of individuals), written as $\rho_{\theta\xi(1)}$.

Maximum-likelihood estimates of $\rho(\hat{\theta})$ and $\rho_{\theta\xi(1)}$ can be obtained from standard mixed model analyses. Standard errors and interval estimates for these quantities are not available from widely-used computer packages.

Figure 3 shows histograms for 4000 bootstrap estimates of $\rho(\hat{\theta})$ and $\rho_{\theta\xi(1)}$. Bootstrap samples were constructed by resampling individuals (where the data for each individual is a vector of the eight observations). The 90 percent confidence intervals were obtained using the so-called 'percentile method': the endpoints are the 0.05 and 0.95 percentiles of the bootstrap histograms. (The percentile method is the simplest of a number of methods for confidence interval construction via the bootstrap: a generally better method is the 'BC_a'; see Efron and Tibshirani 1993, chap. 12–14. The results in Table 2 show good precision for estimation of the reliability, but relatively less precision for the correlation parameter.

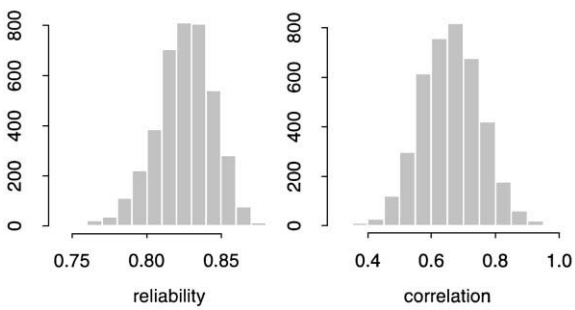

Figure 3
Histograms of 4000 bootstrap estimates of $\rho(\theta)$ (left) and $\rho_{\theta\chi(1)}$ (right).

Table 2
Results for growth curve example

Parameter	Estimate	Standard error	90% CI
Reliability $\rho(\hat{\theta})$	0.828	0.019	(0.792, 0.855)
Correlation $\rho_{\theta\xi(1)}$	0.651	0.093	(0.510, 0.814)

3. Related Methods

There are a number of important related methods, with strong similarities to the bootstrap. The *jackknife* systematically deletes one or more observations from the dataset, and applies the statistic to the remaining data. By doing this for each observation or groups of observations, one can obtain estimates of standard errors and bias that roughly approximate to the bootstrap estimates.

Cross-validation is similar to the jackknife but is used to assess the prediction error of a model. It systematically excludes one or more observations from the dataset, fits the model to the remaining data, and then predicts the data that was left out. Application of the bootstrap for estimating prediction error is not straightforward and requires some special modifications. With these, the bootstrap and cross-validation have similar performance for estimating prediction error. Details and references are given in Efron and Tibshirani (1993) and Efron & Tibshirani (1997).

There is a connection of the bootstrap to Bayesian inference. Roughly speaking, the bootstrap distribution approximates a nonparametric posterior distribution, derived using a noninformative prior. Rubin (1981) gives details.

A difficult theoretical question is: when does the bootstrap work? There is much research on this question (see, e.g., Hall 1992). In 'regular' problems where asymptotic normality can be established, first and higher order correctness of the bootstrap (and modifications) have been proven. However, the bootstrap can fail, for example, when it is used to estimate the variance of the maximum or minimum of a sample. More generally, it is difficult to prove theoretically that the bootstrap works in complicated problems: but these are the very settings in which the bootstrap is needed most. Eventually, theory will start to catch up with practice. In the meantime, the bootstrap, like any statistical tool, should be applied with care and in the company of other statistical approaches.

Some popular statistics packages have bootstrap functions, including S-PLUS and Resampling Stats. In most packages that facilitate some customization by the user, it is easy to write one's own bootstrap function.

4. Conclusions

The bootstrap is a powerful tool for data analysis, with many applications in the social sciences. It is being used more and more, as fast computational tools and convenient software become widely available. Good general references on the bootstrap include Efron and Tibshirani (1993), Davison and Hinkley (1997), Hall (1992), Shao and Tu (1995), and Westfall and Young (1993).

Bibliography

Davison A C, Hinkley D V 1997 *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, UK

Efron B 1979 Bootstrap methods: Another look at the jack-knife. *Annals of Statistics* **7**: 1–26

Efron B, Feldman D 1991 Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association* **86**: 9–26

Efron B, Tibshirani R 1993 *An Introduction to the Bootstrap*. Chapman and Hall, London

Efron B, Tibshirani R 1997 Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association* **92**: 548–60

Hall P 1992 *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, Berlin

Huber P 1964 Robust estimation of a location parameter. *Annals of Mathematics and Statistics* **53**: 73–101

Rogosa D R, Saner H M 1995 Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics* **20**: 149–70

Rubin D 1981 The Bayesian bootstrap. *Annals of Statistics:* 130–134

Shao J, Tu D 1995 *The Jack-knife and Bootstrap*. Springer-Verlag,

Westfall P H, Young S S 1993 *Resampling-based Multiple Testing: Examples and Methods for* p-*value Adjustment*. Wiley

Williamson G, Appelbaum M, Epanchin A 1991 Longitudinal analyses of academic achievement. *Journal of Educational Measurement* **1**: 61–76

B. Efron, D. Rogosa, and R. Tibshirani

Research and Development in Organizations

1. Increasing Division of Labor in Knowledge Production and Use

The growth of research and development activities (R & D) reflects the increasing division of labor in the production and use of scientific and technological knowledge—a process foreseen by Adam Smith at the beginning of the Industrial Revolution. Since then, professional education, the establishment of laboratories, and improvements in techniques of measurement and experimentation have progressively increased the efficiency of discovery, invention, and innovation (Price 1984, Mowery and Rosenberg 1989). Three complementary forms of specialization have developed in parallel, each contributing to the growth of R & D.

1.1 Specialization by Discipline: The Growth of Science and Engineering

First, progress and specialization in scientific disciplines accelerated in the nineteenth century with the development of more powerful research tools and techniques. Increasing opportunities for application also led to the emergence of the engineering disciplines: hence chemical engineering in addition to applied chemistry, and electrical engineering in addition to applied physics (Rosenberg and Nelson 1994). The differences in timing of the emergence of new opportunities for innovation reflected radical but uneven improvements in the knowledge base underlying technological change, and in particular the emergence of new technological paradigms (Freeman et al. 1982, Dosi 1982).

The mechanical paradigm was the basis of the Industrial Revolution. It did not grow out of contemporaneous scientific advances, but of improvements in the quality of metals and the precision with which they could be formed and machined. These enabled the *design*, *development*, *and testing* of prototypes of families of machines with greatly improved performance based on materials of homogeneous and predictable quality, and on shapes with compatible sizes. Similarly, the initial improvements in metals processing technologies depended less on scientific understanding than on *development*

activities, experimenting with different alloys and processing conditions in pilot plant. Even today, such development activities in prototypes and pilot plant typically account for about three quarters of the R & D expenditures of private business firms and for about two-thirds of total R & D.

1.2 Specialization by Corporate Function: The Growth of Business R & D

The science-based chemical paradigm (based increasingly on synthetic chemistry) and the electrical paradigm (based on electromagnetism and radio waves) laid the basis for the industrial R & D laboratory, initially in the chemical and electrical firms in Germany and the USA. Many of the initial applications—like the use of techniques of chemical analysis to control the quality and composition of materials in the increasingly large-scale processing industries—were relatively mundane, but very profitable in reducing costs. Management learned progressively that the science-based technological opportunities were applicable over a range of existing and new product markets, and therefore opened up opportunities for what is now called 'related' product diversification. Mowery and Rosenberg (1989) have described the spread of the establishment of R & D laboratories in the USA from the chemical industry to other sectors, and from larger to smaller firms. Similar processes have been at work in Europe, and more recently in East Asia (Caron et al. 1995, Nelson 1993).

Thus, in addition to the benefits of the cognitive division of labor into more specialized fields, the functional division of labor within business firms also augmented the rate of technical change. Corporate R & D laboratories and similar groups devoted full-time to inventive and innovative activities provided an effective method for creating, combining, and coordinating increasingly specialized knowledge. They provided improved and specialized instrumentation, which enabled firms to monitor and benefit more systematically and effectively from advances in specialized academic disciplines. They also created skills in the development and testing of laboratory concepts and prototypes, and the translation into commercialized products. Firms were able to experiment with a wider range of products and processes than had previously been possible when constrained by established products and production lines. In fields rich in technological opportunity, firms have in consequence become multiproduct as well as multitechnological.

Technological paradigms have been associated with the emergence of large dynamic firms that have been successful in exploiting the new opportunities. The largest R & D spenders today are in companies that grew with the emergence of the mechanical (and automobile), chemical, and electrical–electronic paradigms. Aerospace is a special case, having been technologically force-fed, especially since World War Two, by government R & D subsidies and procurement linked to military requirements. The fastest growing R & D spenders today are those closely associated with ICTs and software technology.

1.3 Specialization by Institution: The Growth of Academic Research

In the nineteenth century, de Tocqueville foresaw that the dynamics of capitalist competition would greatly stimulate the development of innovative activities that showed the prospect of a commercial return. He also argued that public authorities would need to support complementary public research of a more fundamental nature, in order to avoid diminishing returns, to open up new opportunities, and to provide trained researchers. This has happened too. In all advanced countries governments have become the main source of the funding of research activities, and related postgraduate training, in universities and similar institutions. Corporate R & D laboratories have come to depend increasingly on a supply of scientists and engineers aware of the latest research results and trained in the latest research techniques (Salter and Martin 2001).

1.4 R & D and Modernization

This historical pattern of specialization and growth in knowledge production has been broadly repeated in the processes of modernization of late-coming countries, and is observable in countries at different levels of development today: R & D expenditures as a percentage of Gross Domestic Product (GDP) increase with GDP per head, both in specific countries over time, and in cross-sections of countries at any point in time. In successful modernizing countries, R & D is preceded by systematic business investments in improvement-generating activities: in particular, in production engineering, quality control, and design activities, initially making minor modifications and improvements, and later becoming the basis for indigenous innovative activities (Lall 1992, Bell 1984). At this stage, business-funded R & D increases rapidly, accompanied by the growth of university-based research in underlying disciplines, with both reflected in the growth in the numbers of international patents and scientific papers. This pattern can be observed most clearly in both South Korea and Taiwan during the 1980s and 1990s.

The centrally planned Soviet model of modernization, practiced in the former USSR and imposed or adopted elsewhere, gave high priority to R & D. By the late 1950s and early 1960s, R & D expenditures in

Central and Eastern European countries were apparently higher than those in Western Europe and the USA. Certain Western observers, therefore, concluded that the Soviet system was superior in promoting R & D and technical change. However, Soviet economic performance subsequently deteriorated, and it later became clear that a very high proportion of Soviet R & D was oriented towards weapons development, and that the government-established R & D laboratories established for each industry were decoupled from the requirements of producers and consumers (Hansen and Pavitt 1987). The major reductions in R & D activities since their collapse in 1989 can be seen as painful adjustments to make R & D activities—and the underlying activities in production engineering, quality control, and design—become an integral part of a process of economic modernization (Dyker and Radosevic 1999).

The Soviet system also had inadequate linkages with technical advance in the rest of the world economy. One important feature distinguishing today's modernizing countries from those of the nineteenth century is the availability of more productive technologies in more advanced countries. Countries successful in assimilating these more advanced technologies have had two characteristics (Hobday 1995): first, strong linkages with the more advanced countries' technology whether through inward direct investment (e.g., Singapore), inward technology licensing by locally owned firms (e.g., Japan, S. Korea), or subcontracting agreements from advanced countries to local firms (e.g., Taiwan); second, the development of local, change generating activities, culminating in R & D activities, that were essential inputs to effective imitation, as well as to innovation (Cohen and Levinthal 1989).

2. *Measurement of R & D Activities*

2.1 *Patterns of R & D Activity*

By the beginning of the 1960s, initiatives by the US National Science Foundation and the Organization of Economic Co-operation and Development (OECD) had established common definitions of R & D activities, and led to the collection by governments of systematic data on R & D activities. These have since been complemented by privately funded surveys comparing the R & D activities of individual business firms. They show the following common and largely invariant features of R & D activities.

(a) In industrially advanced countries, business and government are the main sources of funds for R & D activities. Business funding is in general larger, and spent in-house mainly on *applied research and development* activities. Government funding is divided between *basic research* performed mainly in university-type institutes, and various types of R & D associated with health, the environment, defense, and the support of industry and agriculture. This R & D is performed mainly in the laboratories of government agencies and business firms.

(b) In the industrially advanced countries, the share of GDP typically spent on R & D varies between 1.5 and 3.0 percent of Gross Domestic Product (GDP), compared to 0.5 and 1.5 percent in the newly industrializing countries, and less than 0.5 percent in the rest. The share of national R & D funded and performed by business firms tends to increase along with GDP per head.

(c) More than 60 percent of all business-funded R & D is typically performed in the chemical, electrical, electronic, and automobile industries. The largest individual corporate spenders on R & D are the world's leading automobile, electrical, and ICT (information and communication technologies) firms.

2.2 *The Economic Determinants and Impact of R & D Activities*

The improved quality of R & D statistics has enabled economists and other analysts to deepen understanding of both the determinants and the economic impact of R & D activities, at the level of countries, industries, and firms. A number of analysts have been able to show that—amongst the industrially advanced countries—differences in the levels and rates of growth of national R & D activities have a significant influence on differences in national performance in exports and productivity (Fagerberg 1987).

At the industry level, considerable attention has been given to the effects of firm size and industrial structure on R & D activities. Given the dependence of specialization on scale, the proportion of firms performing R & D increases with firm size. However, there is no clear consensus on how R & D intensity varies with firm size amongst large firms: evidence can be found either way. And although R & D is found mainly in concentrated industries, there is compelling evidence that variations amongst industries in both the degree of concentration and R & D intensity are determined jointly by a third factor, namely, inter-industry variations in the extent of technological opportunities (Levin et al. 1985).

Less progress has been made so far in measuring the effects of differences in R & D expenditures on company performance. This is partly because comprehensive and comparable R & D data at the company level are only slowly becoming available. It is also because of the difficulty of defining a proper measure of corporate performance. Those used include the R & D production function, stock market evaluation, and long-term growth (Patel 2000).

2.3 Limitations

Sections 2.1 and 2.2 and show that statistics on R & D activities can be a useful proxy measure for innovative activities. But they have their limitations, the most often mentioned of which is that they measure inputs and not outputs. Equally, if not more important are the following:

(a) In large firms, R & D statistics do not measure the often considerable expenditures on complementary activities in production and marketing, that are necessary to transform the outputs of R & D into commercially successful innovations.

(b) In small firms, they do not measure part-time innovative activities, which are particularly important in the machinery and software sectors.

(c) They capture only very imperfectly the growing volume of innovative activities in software now being performed in firms in service sectors, such as finance and distribution.

Other indicators of knowledge-generating and innovative activities have therefore been developed to complement R & D indicators, the most important of which have been counts of patents, papers, citations, and numbers of innovations (van Raan 1988).

3. Some Contemporary Debates in R & D Management

Many of the characteristics and problems of managing corporate R & D have changed very little since the nineteenth century.

(a) The need to orchestrate and integrate specialized knowledge and skills across disciplines, professions, functions, and divisions.

(b) The progressive improvement in fundamental scientific understanding, accompanying the increasing systemic complexity of innovations—a process that will continue with the more widespread application of the techniques emerging from ICT.

(c) The difficulty (impossibility?) of making reliable predictions about the success or otherwise of specific innovations, especially major ones.

At the same time, new problems and challenges have emerged.

(a) With increasing specialization in both the production of knowledge, and of the components of increasingly complex systems, firms are faced with increasingly difficult choices about what products and knowledge to outsource, and what to retain in-house. An important part of the in-house R & D function is now devoted to the monitoring and co-ordination of innovative and production activities external to the firm.

(b) In addition to the well-established function of supporting foreign production, large firms are performing a growing share of their R & D outside their home country, in order to tap into the increasingly numerous international sources of leading-edge scientific and technical advance. This poses new challenges for managers in integrating skills and knowledge over long distances (Niosi 1999), and for national policy-makers in assessing the location of the benefits of public investments in academic research.

(c) There is some evidence that academic research is becoming increasingly linked to commercial R & D activities. For some, this is the consequence of unwelcome financial pressure from governments on universities to demonstrate short term 'relevance' in their research. For others, it is the consequence of fundamental changes in the locus of knowledge production. Other evidence suggests that reductions in the costs of technical experimentation through simulation software now make it easier for academic researchers to develop and test experimental prototypes.

See also: Competitive Strategies: Organizational; Corporate Culture; Design and Form: Organizational; Information and Knowledge: Organizational; Innovation and Technological Change, Economics of; Innovation, Theory of; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; Learning: Organizational; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations, Metaphors and Paradigms in; Organizations: Unintended Consequences; Strategic Intelligence; Technological Innovation; Technology and Organization; Technology Transfer

Bibliography

Bell M 1984 Learning and the accumulation of technological capacity in developing countries. In: Fransman M, King K (eds.) *Technological Capability in the Third World*. St Martin's Press, New York

Caron F, Erker P, Fischer W 1995 *Innovations in the European Economy between the Wars*. W de Gruyter, Berlin

Cohen W M, Levinthal D A 1989 Innovation and learning: The two faces of R & D. *Economic Journal* **99**: 569–96

Dosi G 1982 Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research Policy* **11**: 147–62

Dyker D, Radosevic S (eds.) 1999 *Innovation and Structural Change in Post-socialist Countries: A Quantitative Approach*. Kluwer, Dordrecht, The Netherlands

Fagerberg J 1987 A technology gap approach to why growth rates differ. In: Freeman C (ed.) *Output Measurement in Science and Technology: Essays in Honour of Y. Fabian*. North-Holland, Amsterdam

Freeman C, Clark J, Soete L 1982 *Unemployment and Technical Innovation: A Study of Long Waves and Economic Development*. Greenwood, Westport, CT

Hansen P Pavitt K 1987 The Comparative Economics of Research Innovation in East and West: A Survey. In: *Fundamentals of Pure and Applied Economics*. Harwood, Chur, Switzerland, Vol. 25

Hobday M 1995 *Innovation in East Asia*. Elgar, Brookfield, VT
Lall S 1992 Technological capabilities and industrialisation. *World Development* **20**: 165–86
Levin R, Cohen W M, Mowery D C 1985 R & D appropriability, opportunity and market structure: New evidence on some Schumpeterian hypotheses. *American Economic Review* **75**(2): 20–4
Mowery D, Rosenberg N 1989 *Technology and the Pursuit of Economic Growth*. Cambridge University Press, Cambridge, UK
Nelson R (ed.) 1993 *National Systems of Innovation: A Comparative Study*. Oxford University Press, Oxford, UK
Niosi J 1999 The internationalisation of industrial R & D: from technology transfer to the learning organisation. *Research Policy* **28**: 107–17
Patel P 2000 Technological indicators of performance. In: Tidd J (ed.) *Measuring Strategic Competencies*. Imperial College Press, London
Rosenberg N, Nelson R 1994 American universities and technical advance in industry. *Research Policy* **23**: 323–48
Salter A, Martin B 2001 The economic benefits of publicly funded basic research: A critical review. *Research Policy* **30**: 509–32
Van Raan A 1988 *Handbook of Quantitative Studies of Science and Technology*. North-Holland, Amsterdam

K. Pavitt

Research Conduct: Ethical Codes

This article will focus on the use of ethical codes in dealing with moral problems concerning research involving humans, nonhuman animals and the environment, that is, moral problems that arise mainly in the experimentally working sciences, especially biomedicine. More general issues such as truth telling, or usage of best scientific methods that if neglected may cause moral problems, will not be considered here.

1. *History of Ethical Codes of Research Conduct*

The (ab-)use of humans for research purposes has a long tradition reaching back far beyond the beginnings of modern experimental science in the Renaissance to ancient Egypt and Persia. Unfortunately, morally questionable experiments were also performed in the recent past. Well known are the atrocities perpetrated by physicians during the Nazi regime in Germany during the 1930s and 1940s. But even later, contemptible experiments were uncovered elsewhere (Beecher 1966): an inglorious example is the US Public Health Service Syphilis Study (better known as Tuskegee Syphilis Study) in the course of which a large number of black men diagnosed with syphilis had been left untreated in order to observe the 'natural' course of the disease.

Although the abuse of human experimental subjects is by no means the only moral issue in research conduct (see Sect. 3), it gave rise to the first international moral framework regulating research conduct—the Nuremberg Code of 1947. It was followed by the World Medical Association's (WMA) *Declaration of Helsinki* in 1964 (with several revised versions), and in 1993 by the *International Ethical Guidelines for Biomedical Research Involving Human Subjects*, which was drafted in cooperation by the Council of Medical Organizations of Medical Science (CIOMS) and the World Health Organization (WHO). Besides these widely adopted international guidelines there is a large number of further international and national guidelines. Although regulating research conduct was for a long time equivalent to regulating research on human subjects (for an overview see Veatch 1995), the debate on animal rights as well as on the protection of the environment has also stimulated the issuing of many guidelines covering these areas. (A large selection of relevant texts, including those named above, are reprinted in Spicer 1995.)

2. *Ethical Codes—Philosophical Remarks*

A moral, or an ethos, is understood as a set of rules of action (norms). Ethics, a scientific subdiscipline of philosophy, is the critical assessment of different morals. There are many morals (e.g., the ethos of physicians, the ethos of lawyers, etc.) but only one ethics; the latter should not, however, hide the fact that there is no agreement between ethicists as to what the correct criteria for this assessment are (see *Bioethics: Examples from the Life Sciences*).

A moral or ethos can exist in unwritten or (partly) written form. An ethical code is a written, i.e., codified, moral or ethos; an ethical code is, therefore, a moral and not, as its name might suggest, a part of ethics. Insofar as oaths (such as the Hippocratic oath), prayers, and creeds present a certain moral, they are open to an ethical assessment too.

An ethical code is an instrument of self-regulation established by a certain profession (e.g., physicians' organizations) or those groups directly involved in the actions of that profession (e.g., patients or funding bodies). In contrast, national laws and international conventions (e.g., the Convention on Human Rights and Biomedicine adopted by the European Council in 1996) are issued by government bodies. Using the principle of subsidiarity, i.e., the principle that what can be regulated on a lower level should not be regulated on a higher level, ethical codes allow an increased flexibility in view of new moral challenges when compared to statutory regulations.

An ethical code of research conduct is a codified set of rules of action (norms) that should be adopted by every researcher to whom the code is addressed. If a researcher obeys an ethical code he or she is said to

perform her research in a morally acceptable way. 'Performing morally acceptable research' is relative to the justified ethical code applied in testing a certain piece of research. It might be the case, however, that the evaluation of a certain act of research is morally acceptable relative to one ethical code but not to another. For example, a medical experimenter by performing a certain act of research (e.g., on human embryos) might obey the ethical code to which he is subject in his country, but by the very same act might disobey the ethical code obligatory in another country. In those cases where the differing evaluations of acts of research is deemed unacceptable, the philosophical evaluation of ethical codes has its place. The main areas of interest in this context are the application of results from research, the generation of knowledge through research, and the credibility of researchers. Conflicts may arise due to the incompatibility between—as already noted—different ethical codes, certain parts of a certain ethical code, or an ethical code and what might be called public morals.

The ethics of science deals with the reconstruction (and thereby critical assessment) of rules of action, whether codified or not. This reconstruction displays presuppositions concerning the scope of the research in question, the criteria for the 'scientific' generation of knowledge, and the acceptable application of this knowledge. (Motivation of research, be it longing for social acceptance or money, does not play any role in the ethics of science; only prescriptive claims of validity related to certain acts of research are at issue here.) Ethical codes of research conduct can be subjected to such a reconstructive process. The task of ethics is to make distinguishable by way of reconstructing those moral norms that can be universalized, in other words, that are valid for everybody. This assessment should be done by using a method that is, in principle, accessible to everyone. Those moral norms that have proven to be valid can then justifiably be used to test an act of research with regard to its moral acceptability.

The application of moral rules of action in research contexts—the testing of certain acts of research concerning their norm-conforming character—is easily established in small social groups only. For example, working together in a laboratory on a daily basis, the members of the peer group develop a strong intuition of whether certain acts of research do obey the relevant rules of action or not. Acts that do not conform to the rules of action are, therefore, easily identified and the committing actor is kept from doing those acts again. This habitualization of rules of action seems to be, therefore, an appropriate measure of enforcing rules of action in small social groups (Gethmann 1989).

In complex scientific communities with a highly differentiated division of labor the habitualization of moral norms does not work satisfactorily, because the peer review between researchers cannot be sufficiently secured. In this situation it is necessary to design procedures that achieve the control of obedience to moral norms and their enforcement (institutionalization): though not everyone can evaluate whether a certain experiment is morally acceptable (according to certain standards), because not everyone has the expert knowledge necessary for that evaluation, everyone—at least in principle—should be able to assess whether the procedure used by an institutional review board for evaluating this research (e.g., by a peer-reviewed journal, where the review is done by making use of certain rules of action) is correct. Or to give another example: not everyone can assess whether a certain clinical study obeys the relevant rules of action, but everyone should be able—at least in principle—to assess whether the procedure by which the responsible ethics committee evaluates this study is correct (see also *Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives*; *Ethics Committees in Science: European Perspectives*).

The codification of moral norms, that is the generation of ethical codes, runs largely in parallel with the institutionalization of these norms. Codifying norms enhances their ability to be put into operation: codified norms are less open to inter-individual interpretability of their meaning than norms that are subject to an oral tradition. This stability of meaning supports the efficiency of institutionalized procedures of control of obedience to and enforcement of those norms (cf. the legal requirement of writing for certain documents).

Connected to this is the issue of the status of obligation of ethical codes: statutory regulations must be obeyed (with the problematic exception of toleration of illegal acts or illegal acts that are exempt from punishment). In contrast to that, ethical codes can have a status of obligation varying between a mere declaration of intent and binding professional law. Depending on the act at issue, the consequences of non-obedience of an ethical code may vary from a mere moral reproof to a prohibition from practicing a profession. In addition, the status of obligation of ethical codes may vary over time. This can be illustrated by the *Declaration of Helsinki*: issued by a professional organization, the World Medical Association, it was originally seen as an ethical code binding only for the members of the WMA. Today, however, the Declaration has gained a status comparable to the United Nations' Universal Declaration of Human Rights, to the extent that even non-members of the World Medical Association cannot simply depart from the rules of action it lays down. There is a case, therefore, for transferring the ethical code into an international convention or even law issued by the United Nations.

Finally the question arises, by what procedures should ethical codes be developed and changed. Since the development of ethical codes starts from consensus statements, many codes have been blamed for being developed without a wide enough consultation of the relevant stakeholders. It is claimed, for example, that

codes with an allegedly international validity are negotiated mainly by representatives from Western countries.

Taking into account the respective opinions of the relevant stakeholders is an important step in the preparation of an ethical code. But, as said before, the task of a critical assessment of an ethical code or its draft versions is to single out those moral rules of action that can be universalized, i.e., are valid for everybody. This means, however, that the assessment and its results should be independent of the person, and their background, who is actually performing the assessment. Therefore, procedures for developing ethical codes do not make it necessary to involve all stakeholders; even involving only a very few experts can lead to an ethically justified result. But reflecting on the preconditions of a procedural development of justified norms should not lead to a neglect of the difficulties in the practical implementation of these procedures: the debate on changes to international ethical codes like the *Declaration of Helsinki* involves highly political issues such as conflicts of interest, where the involved parties persist in their claims, be they justified or not. Deciding the question of who should participate in the negotiation of ethical codes is, therefore, an important task in addition to the purely ethical assessment of the rules of action laid down in these codes (Schüklenk and Ashcroft 2000).

3. *Major Issues of Ethical Codes of Research Conduct*

Research bears chances and risks for, among others, patients, society, future generations, nonhuman animals, and the environment (for an introduction see Brody 1998). An important issue of ethical codes is to state criteria relative to which the weighing of these chances and risks can be achieved. An extensively discussed example is the concept of informed consent, which features prominently in many medical codes of research conduct. The lesson drawn from the history of medical research during the twentieth century led to a broad consensus that research without informed consent of the research subject—roughly, a decision made understandingly, intentionally, and without controlling influences—is highly problematic from the moral perspective. Taking this into account the Nuremberg Code demands that in research on humans 'the voluntary consent of the human subject is absolutely essential,' virtually making impossible research on those with diminished decision-making capacities. In contrast to that, the *Declaration of Helsinki* decreased the importance of informed consent and introduced 'best interest' criteria that are supposed to regulate research on incompetent research subjects (e.g., demented or comatose patients). Though vastly treated in the literature, the issue of whether informed consent is a necessary precondition of any research—and if yes, what informed consent actually means—is not satisfactorily solved. Especially in the context of international ethical codes that are designed to supersede national codes, codes that often rest on diverging traditions, this issue remains a controversial subject of debate (see *Research Subjects, Informed and Implied Consent of*).

Another criterion used in ethical codes for the sake of weighing chances and risks is that of benefit. But benefit for whom? Basically the question is whether those affected by research may be burdened with the risks connected to research without benefiting from this research. For example, the debate on the acceptability of nontherapeutic research focuses on this problem: the research subjects in nontherapeutic research are burdened with the risks of this research (like drug side-effects) without benefiting from it (see *Medical Experiments: Ethical Aspects*).

Delivering criteria for weighing the risks and chances of research conduct is not the only topic usually covered by ethical codes. Other topics include fraud and misconduct in research, conflicts of interest, and ownership of results of research (see *Intellectual Property Rights: Ethical Aspects*; *Research Funding: Ethical Aspects*).

An issue of increasing importance for ethical codes is posed by moral problems connected to research involving nonhuman animals and the environment. The former brings up the question as to whether the burden of such research can legitimately be left to those animals. In particular, the so called animal rights movement has done much to raise the level of attention to these problems (see *Animal Rights in Research and Research Application*).

Although the effects of research on the environment have so far not been the main threat that pushed environmental protection forward, research nonetheless can have morally questionable effects on the environment. The release of genetically modified organisms (GMOs), for example, may endanger biodiversity, in that the number of wild-type plants and animals may be reduced by suppression. Although a large number of living species is by itself not necessarily morally preferable, a reduced biodiversity may have (probably only in the future) adverse effects on the ecosystem and thereby on mankind. This would mean, however, that the potential benefits of GMOs are allocated to the living, while the potential burdens due to a damaged environment are left to future generations. Besides postulating rules for actions that demand consideration in dealing with the environment insofar as it is the 'subject' of research, ethical codes may also contain the demand that research actively seeks to improve environmental protection.

See also: Litigation: Effects on Medical Research

Bibliography

Beecher H K 1966 Ethics and clinical research. *New England Journal of Medicine* **274**(24): 1354–60

Brody B A 1998 *The Ethics of Biomedical Research. An International Perspective*. Oxford University Press, New York

Gethmann C F 1989 Proto-ethics. Towards a formal pragmatics of justificatory discourse. In: Butts R E, Brown J R (eds.) *Constructivism and Science. Essays in Recent German Philosophy*. Kluwer, Dordrecht, The Netherlands, pp. 191–220

Spicer C M (ed.) 1995 Appendix. Codes, oaths, and directives related to bioethics. In: Reich W T (ed.) *Encyclopedia of Bioethics*, 5 Vols. Macmillan, New York, pp. 2599–842

Schüklenk U, Ashcroft R 2000 International research ethics. *Bioethics* **14**(2): 158–72

Veatch R M 1995 Medical codes and oaths. History. In: Reich W T (ed.) *Encyclopedia of Bioethics*, 5 Vols. Macmillan, New York, pp. 1419–27

F. Thiele

Research: Ethical Aspects of Long-term Responsibilities

The development of the modern sciences in the twentieth century brought about unexpected consequences for humankind both of a welcome and an unwelcome nature. The ethical issue of long-term responsibility in research takes on particular importance in view of the potential threat to future generations posed by the far-reaching and unforeseeable effects of realizing the possibilities of science. Yet our experience of the unforeseeable nature of innovation processes and the sheer impossibility of foreseeing the life circumstances of future (especially of distant) generations gives us reason for serious doubts about whether it is at all reasonable to speak of long-term responsibility in reference to research. Thus affirmation of this must be founded upon an enlightened ethical analysis of long-term responsibility. If, indeed, such responsibility is to be affirmed, it remains to be clarified who shall bear this responsibility: the researcher, the user, or those who benefit from the results of research. Finally, it may be necessary to determine how long-term responsibility should be implemented within the reality of societal decision-making processes.

1. Scientific Aspects

The discipline of the ethics of science is concerned with the systematic reconstruction of those specific orientations for action that are determined by the understanding, immanent in each field of science in respect of its individual subject area, and the understanding of the scientific procedures for its description and explanation and an appreciation of the possibilities for action opened up by scientific knowledge. In this, ethics of science shall be understood as a part of philosophical ethics that refers to a specific, societal field of action determined by particular forms of knowledge. Above all, ethics of science, in respect of the task on hand, has to deal with two interrelated areas of phenomena. First, ethics of science draws on the specific ethos of the scientific community in order to reconstruct those orientations to which the scientist is committed, in the interest of seeking out the truth. This involves a smooth, unbroken transition from general rules of action (e.g., the prohibition of falsifying research results) to specific regulations for individual disciplines within the framework of their respective methodology. Second, ethics of science concerns itself with the relationship between general moral orientations and the problems of generating and applying scientific knowledge. Both in generating and applying, it is primarily a question of the practical consequences resulting from technical knowledge.

An example of the moral problems involved in the generation of scientific knowledge can be seen in the field of experiments on human beings in medical research and other medical science disciplines (above all in the field of psychology); more recent discussions target experiments on animals or genetic intervention into the genome of humans, animals, plants and microorganisms, for the purpose of getting to know more about genetics. The moral problems inherent in the application of scientific knowledge and the responsibility of the scientists concerned are now under discussion in almost every field of modern science, and have been especially since the 'Manhattan Project' (the building of the atom bomb by US physicists).

The beginning of modern science, at its core, has moved away from the concept of knowledge generated by contemplation, since it is now essentially characterized by knowledge that owes its existence to technical intervention into nature under controlled conditions. Not until a fundamental appreciation of the problems of a modern concept of knowledge was undertaken by Kant and other German proponents of idealism did it become clear that this, too, is thanks to moral presuppositions. Whereas initially it was not the technical consequences of the application of such knowledge that was in the foreground but rather the self-commitment of the scientist to the truth, in the interest of generating reliable knowledge, the discussion on the ethics of science in the twentieth century, in comparison, was first and foremost determined by the experience of the technical and practical consequences of scientific knowledge. In this it was predominantly the natural scientists, in the wake of the development of weapons of mass destruction and other large-scale technical applications of knowledge, who posed the question as to the moral principles of their actions. Above all, with the development of the modern biosciences, the question of the moral responsibility involved in the generation of knowledge

has become the focus of attention. The problems arising especially in the medical disciplines and the biosciences are proving to be highly complex issues, involving the fundamental ethical orientation of societies characterized by scientific and technical civilization.

2. *Fundamental Problems of Long-term Responsibility*

As far as ethics is concerned, the idea that human beings not only bear responsibility for their actions towards those with whom they are faced in the 'here and now' is by no means a new idea ensuing from problems arising from modern science and technology. Nevertheless, the issue of obligation beyond the immediate circumstances of our actions is dramatically brought home to us by modern science and technology in a manner hitherto unknown, in that modern technology generates for successive generations qualitative and quantitative risks—but also chances—unknown to premodern technology, and this not only when actions are undertaken but also when they are omitted.

It is an indisputable fact that present-day actors expect the acceptance of a kind of risk that earlier generations never had to face. In the sense that this is a quantitatively and qualitatively new situation, it may be stated that humankind is now confronted with a new ethical problem. Thus the question of risk is materially the core question when considering the issue of the scientist's responsibility for the future. Taking actions involving risks, however, is not in all cases conducive to conflict and thus is not always of ethical relevance. It is only of interest in those cases where others are expected to accept the taking of actions that involve risks. As far as and as long as the consequences of the actions concerned involve risks that are of a purely private nature, e.g., because (a) the consequences only effect the actors themselves; or (b) the consequences are not recognizable due to insufficient knowledge concerning the relevant causal or conditional nexus; or (c) though the consequences are foreseeable, the persons concerned, as moral subjects, are not deemed equals to the actors, so that as far and as long as this position is held, taking actions involving risks is not an ethical problem; it may be reduced to the problem of acting with predetermined consequences.

The ethical problems of taking actions involving risks are now exacerbated even more by a further dimension, i.e., when it is a matter of the consequences of technical activities for coming generations in the distant future. These are the guiding problems in the issue regarding long-term obligation. Thus, in principle, it should not be necessary to demand a new code of ethics but rather to attempt to come to terms with the 'at any time' concept of our obligation as seen in existing ethics. The question is to be asked more precisely as to what it means for human beings to be under obligation to other human beings in the distant future in view of the unforeseeable long-term consequences of their actions in applying the means of modern scientifically founded technology.

In issues of applied philosophy, like the question of responsibility for climate consequences, for the problems of the final disposal of nuclear waste, for the consequences of intervention in the human genome or for population growth, the problem of obligation and responsibility mainly concerns generations in the distant future. This entails asking a series of nontrivial questions that have to be answered if one seeks to justify the reason for and nature of such long-term responsibility. In this connection it is of primary importance to answer the following three fundamental questions:

(a) Are actors only under obligation toward those with whom they are concurrently interacting?

Only if the answer to this question is negative can there be any talk at all of obligation that extends beyond the actor's present into the future. This question is to be answered in the affirmative by proponents of such ethical concepts from which obligations arise exclusively from claims uttered by currently existing conscious beings (this primarily implying preference-oriented utilitarian approaches). Within the framework of such concepts the paradox arises that for members of a generation G_0 in relation to members of future generations Gn no obligations will ever be able to exist if the future generations are in reality not in a position actually to raise claims against the members of G_0, which after sufficient passage of time (approx. $n > 3$) is always the case. In respect of concepts that reduce obligations to the actual voicing of claims of this nature it must be pointed out that there is a distinction between having the right to make claims on the one hand and the ability in reality to assert such claims on the other.

However, if future generations have justified claims against earlier generations (even when they are not yet able to assert them), then, in appropriate cases, corresponding obligations are incumbent upon the earlier toward the later generations.

(b) Does obligation end with generation $k > i$ (e.g. where $i = 3$)?

Everyday intuition would seem to suggest what many economists profess, namely that due to the lack of knowledge concerning the needs of future generations, which increases with the growing distance in time, obligations extending into the future must end with a generation of a certain, relatively large magnitude of k. However, if one attempts to pin down this intuition precisely, it does, in fact, begin to seem implausible: on principle to allow obligations to end at certain generation markers would mean in effect that for the members of G_0 in relation to $Gi + 1$, in contrast to the members of generation Gi there would be no

obligations whatsoever, and this for the only reason that the former belong to a grade $k > i$ generation. A restriction of this kind on the scope of obligation would be incompatible with an ethical universalism, in principle presupposed both by obligation ethic as well as more recent approaches to an utilitarian ethics. If one grants recognition of moral justification to moral subjects at all, then it seems simply arbitrary to allow this to end at a particular time (or spatial) limit. To allow obligations to end with however near or distant a generation would be particularism of the present, an 'ego-preference' of those consciously concerned at the time. And in any case, such a form of particularism would be in contradiction to cognitive universalism, which is also inherent, so to speak, as a sort of regulative idea, in the search for scientific truth.

(c) Are actors under the same degree of obligation to members of the *k*th generation as they are to the first generation succeeding them?

The conclusion is often drawn from an affirmative answer to question (b) that the answer to question (c) must also be affirmative: Either there exists an obligation on the members of generation *Gk* or there exists no such obligation—but if it does exist, then it does so for the members of generation *Gk* just as for the members of generation *Gk*-1, *Gk*-2, … *G*1. This consequence, however, is not inevitable. On the contrary— irrespective of the question of temporal distance between those under obligation and those entitled to make a claim—a distinction must always be made between the existence of an obligation, on the one hand, and the degree of real liability on the other. Correspondingly the gradation of obligations must be sharply distinguished from the concept of discounting liabilities. Whereas, in fact, the idea of discounting (at any rate in most of the variants put forward) presupposes an ego-preference that is not compatible with universalism and which is therefore rightly rejected by most moral philosophers, the idea of a gradation of obligations, on the other hand, is unaffected by the presupposition of ego-preferences of those consciously involved at the time.

3. *Social Implementation of Long-term Responsibility in Research*

Obligations come into existence through discursively justified claims. In the case of long-term responsibility, however, the very element that is missing is the moral proponent. It can merely be introduced as an entity not yet asserting claims, albeit entitled to make claims. In such cases the process of justifying claims and obligations is therefore not based on real but on fictitious discourse. It must be said, of course, that fictitious discourses are not confined to long-term responsibility. Even in the case of short-term responsibility, moments of fictitious discourse are a matter of course, e.g., concerning forms of delegated, anticipated and representative obligations to act. The fictitious discourse has to be conducted as though the proponent were actually raising claims (with all the difficulties involved in the growing distance between the generations).

It is just as difficult to reconstruct the discursive arguments of the proponent as it is to reconstruct those of the opponent, i.e., the addressee of the claims introduced. Whereas the real proponents bring themselves to the notice of the real opponents, it can easily happen in the case of long-term responsibility that no one feels addressed by the issue. Thus the question must be posed as to who exactly shall be obliged to bear the long-term responsibility. Also for this question, the starting point is short-term responsibility because here, already, there is the problem of determining the addressee.

One of the first conclusions that can be drawn is that an answer like 'Everyone shall take care of their own claims' cannot constitute a satisfactory solution, since actions likely to result in conflict (distribution of resources, exercising usufruct etc.) are the very ones that do indeed concern other people. In respect of long-term responsibility, however, a special case arises in that 'the others' are not known to the actors and as yet may not even exist. Indeed, however, the 'everyone for themselves' solution already comes up against insurmountable barriers in connection with short-term responsibility.

Individuals who have come to the conclusion that they cannot take responsibility for the application of chlorinated carbohydrates as a propellant gas, that individually this conclusion is being justified, nevertheless fail to achieve the aims of acting because they are not in a position to evade the use of CFCs. Such examples can also be cited from the field of large-scale technical plant.

Examples taken from the field of short-term responsibility show that in addition to individual obligation there is also a responsibility of the kind that devolves upon the actors since 'all' the addressees constitute a plea and thus the claim which arises from it. In short it shall be referred to as 'collective responsibility.' Of course, this only indicates the direction of the solution. Collective responsibility could be regarded as a concept of escapism from obligation. It must therefore be clarified in what manner the validity of collective obligations may be asserted in real discourses.

For the further clarification of this question, it is necessary to make a terminological remark. A distinction must be made between obligation arising from an action and the responsibility for an action. Actors are subject to direct and undivided obligation for their actions, and responsibility is borne by them according to delegation of collective obligation. Responsibility arises from obligation through 'moral division of work.' Thus there exists not only the productive

division of labour (in the economy) and the cognitive division of work (in the field of science). A moral division of work is especially applicable to those moral problems that exist in connection with the weighing of risks and chances in respect of the possibilities to act and the compulsions and constraints pertaining to actions in the context of modern science and technology. Due to the complexity of these problems the obligations of all concerned towards future generations must be delegated to the responsibility of individuals.

The first solution that springs to mind in this respect is the creation of bodies of scientific experts. In this, however, the problem of (non)acceptance arises which manifests itself in mistrust towards a perceived 'expertocracy.' According to the degree in which obligations are delegated to those bearing responsibility there is naturally a growth of uncertainty as to whether those responsible are (in generalizable terms) properly assuming their responsibility.

Thus it constitutes a basic problem of modern scientifically and technically based civilization to organize the social assumption of responsibility in such a way that, on the one hand, the effect of preventing an overburdening is maintained through the process of delegating obligation, while, on the other hand, the moral competence of the decision maker remains transparent. Although not every action is relevant in respect of distant consequences, the question must be asked in those areas that are sensitive to such distant consequences as to which social agencies are in a special way capable of, and therefore should be entrusted with, bearing long-term responsibility.

In this there are many complicated issues that have to be resolved. Above all the question of who should bear the responsibility is particularly explosive at present. Is it primarily a matter for the state and its three main organs of power or is it a matter for society and its subsystems to take over responsibility for the future? The state, through the medium of its legal system, is not a good advocate of long-term responsibility for several reasons. First it is subject to the specific democratic rules governing the lifetime of a parliament and secondly, the functionality of its laws is restrictive. The democratic rules controlling the period in which a parliament stays in power have the effect that the parliamentary system is geared to ultrashort periods of obligation, i.e., the members of parliament, as mandate holders, assume responsibility for four or five years, after which they have to answer to the electorate; the voters, in their turn, have to optimize their decision as to how they should vote with regard to a wide variety of political dimensions. Since the support given by the mandate holders to the idea of long-term responsibility is only one dimension among many, it is likely that for many voters this will be the dominant dimension—as long as voters each have only one unweighted vote.

For a restructuring of the political system, which could be imagined, it would not be sufficient to introduce marginal changes like merely extending the period of parliament or establishing a system of overlapping parliamentary recruitment. In accordance with the principle of diminishing obligation, as adumbrated above, it would be ethically not permissible—and certainly not enforceable either—to abandon the foundation of parliamentary democracy in the (putative) interests of distant generations. For another reason the state is not a good agency to take over long-term responsibility, because its most important medium, the law, is not suitable to regulate subtle moral issues. An essential difference between ethics and law consists in the fact that there is no judge for morally reprehensible actions. Ethics apply rigorously—in contrast to the law; which is why in ethics the saying, 'Where there is no plaintiff, there is no judge' makes no ethical sense at all. Otherwise, there would be no such thing as long-term responsibility since persons who do not yet exist cannot stand up as plaintiffs. Thus there would be no defendants either. A further problem with regard to law in connection with future generations is the difficulty arising from the fact that the relevant occurrence that would be subject to legal punishment has not yet taken place, or may only take place with a certain (possibly very small) degree of likelihood. It can be seen clearly that principles of law are inapplicable here. Furthermore, in respect of the relationship between law and morals the principle of subsidiarity applies, i.e., if something can be accomplished by moral discourse, no attempt should be made to regulate it by law.

Without the necessity of excluding the possibility that in many issues the law may be utilized as an instrument in establishing the long-term responsibility of the members of society, these reflections show that when it comes to answering the question of who should bear responsibility, the focus of attention naturally turns to society with its moral forming and preserving groups and their respective subsystems. In this process a number of subsystems play a prominent part, such as religion, culture (in the narrow sense of the arts), education, and the economy. The findings of the sciences and their respective applications, however, are only of particular importance from the point of view of long-term responsibility.

Their future perspectives are characterized by being particularly concrete and particularly long-term and are marked by having an especially profound effect on human life as we know it. Therefore, it is a paramount task of the scientific community to promote the capability of institutions of assuming long-term responsibility.

See also: Ethical Dilemmas: Research and Treatment Priorities; Objectivity of Research: Ethical Aspects; Research Conduct: Ethical Codes; Research Ethics,

Cross-cultural Dimensions of; Research Ethics: Research; Research Funding: Ethical Aspects

Bibliography

Apel K O 1987 The problem of a macroethic of a responsibility to the future in the crisis of technological civilization: An attempt to come to terms with Hans Jonas's 'Principle of Responsibility'. *Man and World* **20**: 3–40

Birnbacher D 1988 *Verantwortung für zukünftige Generationen*. Reclam, Stuttgart, Germany

Gethmann C F 1989 Protoethics. Towards a formal pragmatics of justificatory discourse. In: Butts R E, Brown J R (eds.) *Constructivism and Science. Essays in Recent German Philosophy*. Kluwer, Dordrecht, pp. 191–220

Gethmann C F 1993 Langzeitverantwortung als ethisches Problem im Umweltstaat. In: Gethmann C F, Kloepfer M, Nutzinger H G (eds.) *Langzeitverantwortung im Umweltstaat*. Economica Bonn, Germany, pp. 1–21

Gethmann C F 1998 Reason and cultures—life-world as the common ground of ethics. In: Lange H L, Löhr A, Steinmann H (eds.) *Working Across Culture*. Kluwer, Dordrecht, Germany, pp. 213–34

Gethmann C F 2000 Die Krise des Wissenschaftsethos. Wissenschaftsethische Überlegungen. In: *Ethos der Forschung. Ringberg-Syposium Oktober 1999*. Max-Planck-Gesellschaft, Munich, Germany, pp. 25–41

Gethmann C F, Kamp G 2000 Gradierung und Diskontierung von Verbindlichkeiten bei der Langzeitverpflichtung. In: Mittelstraß J (ed.) *Die Zukunft des Wissens. XVIII. Deutscher Kongreß für Philosophie (Konstanz 4–8 Oktober 1999)*. Vorträge und Kolloquien, Berlin, pp. 281–95

Golding M 1972 Obligations to future generations. *The Monist* **1**: 85–99

Höffe O 1993 *Moral als Preis der Moderne*. Suhrkamp, Frankfurt, Germany

Jonas H 1984 *The Imperative of Responsibility. In Search of an Ethics for the Technological Age*. Chicago Universisty Press, Chicago

Lehrer K (ed.) 1987 *Science and Ethics*. Rodopi, Amsterdam

Mittelstraß J 1992 Zur Ethik der Naturforschung. In: Mittelstraß J (ed.) *Leonardo-Welt. Über Wissenschaft, Forschung und Verantwortung*. Frankfurt, Germany, pp. 120–54

Mittelstraß J 1995 Ethics in science—substance or rhetoric. In: Götschl J (ed.) *Revolutionary Changes in Understanding Man and Society. Scopes and Limits*. Kluwer, Dordrecht, Germany, pp. 269–77

Norton B G 1978 Future generations, obligations to. In: Reich W T (ed.) *Encyclopedia of Bioethics*. Macmillan, London, pp. 892–9

Norton B G 1982 Environmental ethics and the right of future generations. *Environmental Ethics* **4**: 319–37

Parfitt D 1983 Energy policy and the further future. The identity problem. In: MacLean et al. (eds.) *Energy and the Future*. Rowman and Littelfield, Totowa NJ

Passmore J 1974 *Man's Responsibility for Nature. Ecological Problems and Western Traditions*. Scribner, New York

Shea W R, Sitter B (eds.) 1989 *Scientists and their Responsibility*. Watson, Canton, MA

Sikora R I, Barry B (eds.) 1978 *Obligations to Future Generations*. Temple University Press, Philadelphia, PA

Weiss E B 1989 *In Fairness to Future Generations: International Law, Common Patrimony, and Intergenerational Equity*. United Nations University, Tokyo

Wieland W 1999 *Verantwortung—Prinzip der Ethik?* Winter, Heidelberg, Germany

C. F. Gethmann

Research Ethics, Cross-cultural Dimensions of

For the researcher working in a culturally distant setting, for whom cultural and behavioral differences impinge directly on methodological design, data collection, and everyday life in the field, maintaining ethical equilibrium can be a difficult, never-very-satisfying discipline. Anthropologists are most prone to this situation and have reflected the most about it, but to the extent that other practitioners (political scientists, demographers, geographers, etc.) utilize ethnographic approaches, they, too, expose themselves to significant ethical dilemmas. Faraway places are perhaps somewhat extreme in this regard, but studying ethnic minority groups in one's own society can present particular ethical challenges, especially when the observed practices are illegal or are more legitimately at odds with the tastes and standards of the wider community, to which the researcher typically belongs. The difficulty stems from the researcher's need to balance the demands of multiple moral codes, not only those based on personal and professional values—a requirement common to much scientific investigation, whatever the content and venue—but, in addition, those of the host culture and the laws of its land, and those derived from the interests and cultural traditions of persons to whom the researcher has become emotionally attached and obliged during the course of research. Ethical clarity is not helped when, as frequently happens, some of its elements are invisible or incomprehensible to the researcher—from whom, at least early in the project, but to some extent throughout, blunders are to be expected. Charity from the hosts usually redeems the investigator, but not always, and those are the instances of ethical infirmity which rarely survive as classroom and dinner-party anecdotes in the after years.

1. Changing Ethical Circumstances

In the early days of fieldwork among 'native peoples,' prior to about 1920, ethical issues were not so pressing. For British, French, German, and Dutch researchers of that era, the terms of engagement were largely governed by colonial authorities; in North America and South America, comparable oversight was exercised by internal government agencies, such as the

Bureau of Indian Affairs of the US Department of the Interior. Whatever retrospective moral judgment one might pass on those institutions, they were taken for granted at the time, and were the reality within which ethnographic inquiry was conducted. Considerable research was actually carried out by scientists employed or sponsored by colonial or governmental administrations, though with little effect, apparently, on the policies and procedures of rule. Access to field sites was readily achieved by bona fide researchers, who, so long as they disturbed neither village peace nor administrative purpose, were left alone. Host communities had little power to resist being studied, should they have wished to do so. In any case, such communities could not have understood very fully why the equipment-toting sojourner was there, nor could they have foreseen the consequences of the study for themselves and their posterity.

In those days, too, ethnographic research often consisted of participation in regional survey work conducted by teams of scientific specialists and government officers. Theirs was a 'collecting' mentality, with no need to stay long in one place. The expedition would record cultural inventories (word lists, kinship terms, technology surveys, basic institutional forms, various traits that were part of the standard ethnological catalogue, etc.), take anthropometric measurements and photographs, and, particularly if a museum was its sponsor, collect artifacts (see Sect. 2). Gifts and payments would be distributed—items such as steel tools, for example, which were relatively inexpensive for the expedition but highly prized by the recipients—and the party and its carriers would move on to the next group.

Beginning in the 1920s a revolutionary new fieldwork methodology expanded the ethical implications. Based on a technique loosely known as 'participant observation,' the approach involved prolonged (1–2 years or more) residence in a host community with the aim of acquiring intimate familiarity with the people and their culture. Close familiarity in turn invited the innovation of the 'case method' approach, in which specific events (often disputatious in nature) were contextually examined as indicative of underlying cultural understandings or social organization. Beyond their analytic usefulness, well-executed case studies added vividness and probity to the ethnographic account. Case studies embedded in monographs usually identified individuals by name, along with relevant genealogical histories and information of a potentially embarrassing or inflammatory nature. Risks to informants thereby entered into the research equation, but, initially, only to a slight degree. Missionaries and colonial administrators might have searched out published information and used it to the detriment of individuals or cultural traditions, and yet it appears that this was a rare occurrence. As for the people themselves, information revealing the hidden deeds and motives of neighbors and enemies might have been very useful, and damaging to the persons named but, under those remote, preliterate conditions, the idea that some explosive item from a book or journal would find its way back to the village and there make mischief would have been very farfetched.

Nowadays, such possibilities are quite thinkable. With the spread of literacy, some monographs are finding their way back to the village, or into the courts, where facts and events previously noted are being used, for example, in litigation concerning land claims. Starting in the late colonial period and continuing until the present day, educated elites in many of the countries traditionally included in cross-cultural studies comprise a new community of critics: some are the children of cultures ranking as 'classics' in the ethnographic literature, some are social and behavioral scientists, some are even anthropologists. The new intelligentsia, many of them educated in the (self-) critical Western tradition, many of them eager to forget the past and move on, are harshly critical of their colonial past and willing to condemn out of hand studies carrying the odor of those times. Ready allies are to be found from within the various critical movements that swept Western academic institutions during the last two decades of the twentieth century. Anti-orientalism, postmodernism, neomarxism, 'science studies,' the idea of ethnography as a constructed text, the distrust of studies conducted under prefeminist, 'patriarchal' assumptions—these are part of the ethical-cum-intellectual apparatus currently used to discredit the tradition of cross-cultural study.

Although a good deal of this criticism is ill-informed and irrelevant, the best of it is very good indeed, offering novel insights and perspectives, the value of which even the most diehard traditionalist cannot reasonably deny; devising new concepts and new research goals and relationships, such as befit the new parameters of cross-cultural studies in a world very different from what it was as recently as a generation ago. Gone are the reassuring isolations which formerly shielded cultures, and those who studied them, from outside scrutiny—the never-perfect isolations which lent credence to the always dubious doctrines of cultural essentialism and cultural relativism (Gellner 1985, Cook 1999). Globalization engenders recurrent issues, such as local environmental degradation, the AIDS epidemic, poverty, war, and political revolution, which pique not only the researcher's interest and intellect, but his or her deep questionings into the ways of goodness, justice, beauty, and the future of the human species. These are supremely moral and ethical questions, which, while not displacing the urgent quest for objective findings, are adding immense complexity to the task of the researcher.

2. *Cultural Property*

Reference was made to the routine practice of earlier researchers of collecting information and emblematic

artifacts from peoples they visited. Now, generations later, descendants of some of these peoples, or those speaking for them, have come to demand the return of these objects. Never mind that perishable objects, both tangible and intangible, would never have survived in their home place, their removal is retrospectively construed as a form of cultural banditry. Today, legal claims concerning cultural patrimony and indigenous peoples' exclusive rights to traditional property—material or intellectual, commercial or sentimental—circumscribe an international, cross-cultural arena of high-stakes ethical debate. The cases, to name a few, include the longstanding litigation between Greece and the United Kingdom over rights to the 'Elgin Marbles'—sculptures from the Parthenon and other buildings of the Athenian acropolis which were removed by Lord Elgin in the early nineteenth century and were later sold to the British Museum; the legal contest between scientists and local Native American groups over custody of the remains of a prehistoric individual of obscure genetic provenance known as 'Kennewick Man'; more generally, the repatriation of skeletal material to indigenous descendant groups—a process now mandated by US federal law; efforts by Australian Aboriginal groups to recover legal title to ancestral lands containing sacred sites; the question of potentially huge royalties payable to various Amazonian groups who have guided drug companies to forest products having pharmaceutical properties; and the commercial exploitation of design elements taken from among Native American traditions of the US Southwest.

Social scientists are being ever more frequently drawn into these issues as witnesses, commentators, teachers, analysts, and advocates. Increasingly, sometimes unavoidably, the role of the researcher is shifting from that of detached observer to that of social activist. What to a previous generation of investigators would have seemed an unconscionable abandonment of scientific objectivity is nowadays becoming a common, practical necessity—a commitment demanded by newly sophisticated, politically aware groups as a condition of their acceptance of a research visitor. Other researchers, decrying 'scientific objectivity and detachment' as simply one ideology among others, gladly construe their mission to be that of assisting in the struggle of powerless ('subaltern') groups to improve their lot vis-à-vis dominant state or corporate interests (D'Andrade 1995). Wherever one stands on the detached vs. committed gradient, social movements and the emergent social and political aspirations of postcolonial groups are, themselves, of compelling intellectual interest to the social scientist, further blurring the ethically consequential line between actor and observer. Gone is the age of innocence during which many ethnographic habits were formed, when ethical choices were a good deal simpler.

In less controversial terms, professional ethics now calls for the dissemination and preservation of research findings. Some host countries require, as a condition of granting research permits, that copies of field notes be deposited in their national libraries or archives, or that debriefing talks be given to academic groups. Fearing the continued neglect and eventual loss of much unpublished material, some research communities are establishing specialized archives for the preservation of field notes and other cultural records, and for their systematic repatriation to the host countries, with the aid of microfilm, microfiche, or other inexpensive, readily accessible technologies (Tuzin 1992). Not only do such practices secure priceless cultural information and benefit future generations of researchers, they also contribute immediately to goodwill and a collaborative spirit between scholars belonging to reciprocal sides of the cross-cultural divide. These trends are further assisted by the requirement of many host countries that foreign researchers be sponsored by a local academic institution or engage an assistant or co-investigator from among the local scholarly community. The importance of behaving responsibly toward host peoples and nations and toward the long-term scientific good is an explicit principle of professional ethics. Some funding agencies, such as the Wenner-Gren Foundation for Anthropological Research, are actively encouraging, and committing resources to, an approach that emphasizes custodianship rather than ownership as properly describing the researcher's relationship to cultural treasures he or she is privileged to collect.

The distinction between custodianship and ownership further points to an unresolved issue of looming ethical importance: on moral and legal grounds, what should be the tenurial relationship between a body of cultural ideas and the persons who, by accident of birth, happen to be carrying them? Or, as put by one commentator (Brown 1998), can culture be copyrighted? The alternative would be to say that, just as culture is an attribute of humanity, so all cultures, none of which are formed in perfect isolation, are manifestations of that attribute, and therefore are ultimately the property of all humanity—granting that particular groups have contingent, custodial rights over particular cultural assemblages. The former view accords with notions of cultural distinctiveness and cultural relativism, the latter with ideas of universal ethical standards and a body of interests common to the human species. The former view would seem to revere human rights—the right to their cultural traditions being the only right left to many otherwise dispossessed groups—while the latter would appear to remove the last protection against wholesale cultural trespass.

This issue is a minefield of ethical problems that will be prominent in the landscape of intercultural relations and research in the twenty-first century. The seismic convulsions of that landscape will be traumatic moments in the globalization of social and economic relations; increased human migration, possibly in

response to local population pressures and global-warming effects, bringing different cultural groups increasingly into competition with one another; the collapse of autochthony as an argument for indigenous groups in fending off migrant groups beating on the gate—groups once autochthonous themselves, driven by necessity rather than choice, with human rights of their own. In such a world, the conflation of cultures and peoples, collective memories and current circumstances, could produce dangerous vulnerabilities, generating the kind of xenophobic, stereotyping mentality that justified campaigns of ethnocide and 'ethnic cleansing' in the darker moments of the twentieth century. The better alternative, perhaps, is for universalist ethics to have a rightful place alongside the value of respecting and preserving cultural identities, and the rights that may be attached to them. Whatever shape that future takes, however, it is certain to affect the ethical circumstances of cross-cultural researchers.

3. *Regulations and Codes*

In recent decades, research ethics, once the privileged business of conscience, personal character, professional integrity, and scientific organizations, has come to the attention of lawmakers. During the 1970s federal government regulators in the USA developed rules for the protection of human subjects (Sieber 1992). With ethical reasoning rooted in the Nuremberg Code developed after World War II, the regulatory impulse was triggered by certain scandals involving biomedical research. Early in the lawmaking process, however, authorities recognized that potential risks to human subjects also existed in the social and behavioral sciences. Social and psychological hazards were recognized, as were the need for confidentiality and the importance of informed consent by subjects—not only at the outset, but through the entire course of the research. Procedures were instituted to disclose risks to peer reviewers and granting agencies, assess whether or not they are outweighed by potential benefits of the research, including to the human subject; and, in the field or laboratory, inform subjects of these risks as part of the consent process. A plan for compliance with these procedures, formally approved by university-based, faculty-run internal review boards, became a prerequisite for receiving government grants. Research centers quickly adopted these federal guidelines as their own, mandating the safeguards regardless of funding source.

These measures address ethical anomalies that are egregious in the sense of being built into the research design. No regulatory guideline can govern the many unforeseeable, quotidian dilemmas that may arise over an extended period of research in another culture—a reality acknowledged in the Code of Ethics promulgated by the American Anthropological Association (1998). Still, these measures have the salutary effect of heightening researchers' consciousness and accountability concerning the welfare of human subjects, while helping to shield institutions and innocent investigators from liability if things go wrong and unethical conduct is alleged. On the other hand, regulatory experience and input from practicing social and behavioral scientists have shown that applying the 'biomedical model' to research in their fields raises problems, some of which are peculiar to cross-cultural studies. Even if language differences *per se* are not an obstacle, other communication barriers can impede the way of genuine 'informed consent.' If the research subjects are lacking in literacy, this would preclude written consent—a circumstance for which official guidelines permit equivalent oral procedures—but there are more subtle, cultural, difficulties. Even where written consent is feasible, subjects usually would not be accustomed to direct questioning by strangers (except, ominously, by police authorities) nor would they be likely to understand the foreign culture of litigation and the bureaucratic concept of 'risk' which jointly inspired the formal consent procedure in the first place. Depending on the political environment in which such an encounter occurs, a procedure innocuously designed to identify and mitigate risk triggers, instead, reactions of suspicion, alarm, and flight. For the would-be subject, for whom the compensatory benefits of the research may seem ethereal and remote, refusal might well be the easiest course, ending the research before it begins. Where traditional communities are receptive to the researcher, an over-zealous concealment of individual identities would deprive the people of the only documentary history they might ever have. And so on. Government overseers and internal review boards, acknowledging such technical difficulties, have revised the regulatory guidelines so as to accommodate the special needs of ethically responsible research in other cultural settings.

The protections just discussed focus on the welfare of individual human subjects. What ethical imperatives pertain when not only individuals but entire cultural traditions are potentially at risk? Anthropology, by virtue of its disciplinary investment in cross-cultural research, celebration of cultural diversity, and preoccupation with the culture concept, has a special affinity for ethical issues that arise when science crosses cultural boundaries. Even if the threat stems not from the research *per se* but from other external factors (missionaries, commercial interests, etc.), is the anthropologist not duty bound to intervene vigorously in defense of the host culture? After all, in its submission to United Nations Commission on Human Rights, in 1947, the executive board of the American Anthropological Association recommended that the UN Declaration should be a statement 'of the rights of men to live in terms of their own traditions' (quoted in Cook 1999, p. 169). Should not anthropologists be willing to intervene, by word or deed, in defense of those traditions?

Critics might condemn such intervention as a veiled form of paternalism verging on racism—a denial of the ability or right of individuals to make their own decisions, to dispose of their cultural legacy in a manner they, and not well-meaning outsiders, choose. From this standpoint anthropologists are sometimes accused (sometimes by designing outsiders, with self-serving plans of their own for the group) of being culturally conservative, of wishing to see traditions preserved, of standing in the way of perceived material development and 'modernization.' Whatever their motives, the critics have a point. If the idea of indigenous rights is to mean anything, it must hold sacred the element of choice—even if the choice in question proves to be an unfortunate one in the view of contemporary observers or of future generations who must inherit the consequences. The role of the concerned outsider (anthropologist or other) is to help, if invited, to insure that the choices made are informed choices, based on the advisor's experience, expertise, and external perspective. Such advice would be offered with knowledge of authority relations occurring in that cultural setting, specifically, the mechanisms through which decisions affecting individuals or the group are made.

4. Conclusion

Social and behavioral science research has ethical implications in any setting. Though we may control the interaction to a maximum extent, we are still humans studying other humans. Lives can be affected. To a greater or lesser degree, expectations, projections, aspirations, and the reflexivities accompanying all of these, are part of the immediate research situation. They influence what we observe and how we observe it; they shape the larger lessons we draw in the long afterthought. The publication of results can further affect the original subjects or, years later, their descendants. All of these ethical implications are greatly compounded when the research engagement is intimate and prolonged, and extends across a cultural boundary separating the investigator from the persons he or she is attempting to understand (Geertz 1968). Obligations accrue. Reminiscent of the methodologically engineered interplay of transference and counter-transference in psychoanalytic practice, as the ethnographic relationship deepens, complex personal motives intervene, and the desire to belong and to assist can become so strong as to compete with the desire to know. The whole point of maintaining ethical equilibrium is to transform this competition into a synergy. This is not an easy task; it requires thoughtful, honest reflection upon the ethical dimensions of what one is doing, together with the courage not to be paralyzed by them. In cross-cultural research the effort is worth the candle, for it is this synergy which stands the best chance of producing a result that is at once truly humane and truly informed.

Bibliography

American Anthropological Association 1998 *Code of Ethics*. American Anthropological Association, Washington, DC

Brown M 1998 Can culture be copyrighted? *Current Anthropology* **39**(2): 193–222

Cook J W 1999 *Morality and Cultural Differences*. Oxford University Press, New York

D'Andrade R G 1995 Moral models in anthropology. *Current Anthropology* **36**(3): 399–408

Geertz C 1968 Thinking as a moral act: ethical dimensions of anthropological fieldwork in the New States. *Antioch Review* **28**(2): 139–58

Gellner E 1985 Relativism and universals. In: Gellner E (ed.) *Relativism and the Social Sciences*. Cambridge University Press, Cambridge, UK, pp. 83–100

Sieber J E 1992 *Planning Ethically Responsible Research: A Guide for Students and Internal Review Boards*. Applied Social Science Research Methods Series, no. 31. Sage, Newbury Park, CA

Tuzin D 1992 The Melanesian archive. In: Silverman S, Parezo N (eds.) *Preserving the Anthropological Record*. Wenner-Gren Foundation for Anthropological Research, New York, pp. 31–42

D. Tuzin

Research Ethics: Research

Research topics, approaches and contexts bring with them ethical issues—some well understood and easily resolved, some unexplored and potentially risky, and some in which one's moral outlook is inadequate to the problem. Sometimes the social or political context is such that morally wrong actions are rewarded and morally right acts are punished. Thus, ethical issues in research include conflicts between ethical principles, questions of what action will produce the most good, questions that call for a new outlook or approach, pursuit of personal gain at the expense of others (e.g., research fraud, endangerment of research participants) and research in contexts where the scientist is not free to choose his or her own scientific goals or methods.

Until issues are understood, scientists can only guess about the risks involved and how to ameliorate them. Although ethical questions are normative (What is the morally right action to take?), answers must often be grounded in scientific inquiry. For example, when investigators first undertook major field research on illegal behavior (e.g., prostitution, drug dealing, child pornography) and promised confidentiality in order to obtain honest responses from informants, perceptive researchers wondered, e.g., Can I protect

my data from subpoena, and can I gather valid data otherwise? How serious is the threat of subpoena? Would I go to jail if I were to ignore a subpoena of my data? What will happen to research careers and research in criminal behavior if protection from subpoena cannot be assured? Research by criminologists, sociologists, and political scientists has answered these questions, and led to the development of legal protection of data from subpoena. Research administrators, researchers, and students have learned to use these tools.

Empirical and methodological research coupled with social science theory is used to explicate how ethical issues arise, their possible consequences, and how they might be resolved. This article outlines a range of topics and approaches employed in research on research ethics. Other articles herein on ethics provide detailed examples.

1. History

Prior to the 1960s, most scientists considered their work value free or inherently good. Some scientific societies developed codes of ethics that were quite general and largely ignored. Some survey researchers worried about possible breaches of confidentiality; and some applied researchers recognized that knowledge is power that may be used for good or evil. However, most researchers were oblivious to the ethical dimension of their work. The field of research on research ethics was beyond the imagination of almost everyone.

With the rise of concern about ethics of human research, scientists found that social and behavioral research methods and theories could be used to explore and resolve problems in research ethics. Government funding for such research began to be offered. Increasingly, scientific societies encouraged such research, scientific journals published the findings, and new journals devoted to ethics and science came into existence.

2. Approaches

Most research on research ethics uses theory, has methodological components and involves empiricism. However, for simplicity, the discussion in this section focuses on main approaches:

2.1 Empirical Research

The entire range of empirical social/behavioral research methods may be used in pursuit of empirical answers to ethical questions. For example, survey research methods have been used to assess whether psychology departments comply with federal regulations in the operation of their subject pools. Experimental methods have been used to examine the effects of various approaches to informed consent on comprehension and decision making, and entire research funding programs have invited proposals on this topic. Psychometric methods have been used to study the scientific and ethical values of scientists in various disciplines and their relationship to kinds of problems studied and methods employed (e.g., Kimble 1984, Tarnow 1999). Participant observation methods have been used to study the nature of communication concerning informed consent. Econometric measures have examined the costs and benefits of increased institutional monitoring of human research.

2.2 Methodological Research

Methods of designing and organizing research, evoking and measuring responses, and collecting, storing, and sharing data have ethical implications. For example, does a given procedure or method evoke trust, respect privacy, assure confidentiality, or yield valid results? The design and evaluation of methods to ensure that they meet ethical criteria require methodological research. Two major areas of methodological research in the service of ethics are privacy/confidentiality, and the design of field experiments. The development of methods for protecting subject privacy and data confidentiality has focused on such diverse contexts as interviews, sampling, collecting data from internet chat rooms, and inter-file linkage of shared data. Technological developments increasingly permit unobtrusive observation and measurement of behavior which could destroy trust in major institutions in society; discovering and evaluating uses of such technology have become topics of intense interest (e.g., Agre and Rotenberg 1999). The second major area of intensive methodological research has focused on field settings where rigorous experimental methods with well designed controls are needed, but ethical considerations prohibit jeopardizing the well-being of subjects by placing them in conditions not of their choosing. This has led to development of quasi-experimental methods that are responsive to both ethical and scientific requirements.

2.3 Theoretical Research

This approach conceptualizes and synthesizes what is known about a given ethical problem in the service of ethics and good science. For example, Laufer and Wolfe (1977) developed a theory of privacy that accounts for how age, culture, and status determine persons' regulation of the access of others to themselves. Thompson (1992) synthesized what is known about youngsters' age and sensitivity to risks of research. Gunsalus (1998a, 1998b) has summarized

the destructive conflicts and institutional mishandling that often surround whistleblowing, and has synthesized best practices for institutional and individual handling of whistleblowing situations.

2.4 Legal Research

The regulatory environment that is growing around scientific endeavors inevitably brings with it ambiguities and conflicts when applied in areas that the lawmakers did not anticipate. Exact standards of research conduct may leave no room for intelligent interpretation or for resolution of conflicting ethical principles. The effort required to follow a given law may be disproportional to the expected results. As examples, Western standards of informed consent and signed consent forms are inappropriate and even disrespectful in many contexts such as fieldwork in cultural anthropology and research in third world cultures. Oral histories often involve sensitive details about living persons but cannot meaningfully be rendered anonymous. Examining and challenging laws governing research and proposing more effective laws requires ability to research and interpret laws in relation to their effect on research within specific disciplines, methods, and research topics.

3. Topics of Research on Research Ethics

Topics are organized into 17 categories, nested under five major headings:

3.1 Theory, Method and Design Issues. These Background Issues Set the Ethical Context of Research

3.1.1 Values and epistemology. Scientists' beliefs about knowledge and reality determine what are considered valid scientific problems and how data should be collected, organized and disseminated. Whether knowledge is regarded as an unlimited good or as power affects perception of research risks and benefits. Whether reality is regarded as something 'out there' or as a construction composed of interaction between investigator and subject influences which methods (e.g., deception) are acceptable.

3.1.2 Validity. To be ethical, research must be valid; otherwise it disrespects subjects and society by producing false information. Conversely, in pursuit of validity, investigators may use unethical methods (e.g., putting subjects at risk). Much research on research ethics focuses on finding ways to optimize both beneficence to participants and validity of findings.

3.1.3 Equitable treatment of participants. Procedural and distributive justice issues may arise in planning, conducting, and applying research. Who sponsors the research? Who benefits? Will the results be used in a way that may harm the subjects or their culture? How are subjects selected? How is the power of the researcher balanced against that of the subjects? What kinds of control groups are acceptable? Who sponsors and pays for the research? When is it unethical to withhold information about sponsorship from subjects?

3.2 Risk and Benefit

The perspective of scientists is influenced by their research objectives, culture, and experience with the circumstances of those studied. A narrow perspective may negate the promise of scientific research to be beneficial. Research on risk and benefit seeks to broaden this perspective and enhance the value of research to subjects, their community, researchers, their institution, funders, science, and society. Risks and benefits are generic to all ethical considerations and research topics. However, specific risks vary markedly in relation to the research context (e.g., laboratory, community, school, workplace, hospital, prison).

3.2.1 Risk, wrong, and harm. How does one weigh moral wrongs (e.g., deceiving subjects), versus risks of harm? What determines perception of risk and safety? How can this determination be reached accurately and usefully? How can risk, wrong or harm be assessed?

3.2.2 Benefit and promise. Benefit is the moral justification of research, but it is often simply taken on faith that research is, by definition, beneficial. How are benefits identified, estimated, and maximized? Who benefits? Often research has little likelihood of benefiting society; does the researcher therefore have a duty to benefit the subjects or their community directly via some services, products, training, or job opportunities connected with the research?

3.2.3 Risk/benefit assessment. The price of scientific knowledge is risk and financial cost. How is risk justified in relation to benefit? What is the risk or cost to society of research opportunities that are lost due to financial, regulatory, or political pressures? (e.g., see Rosenthal and Rosnow 1984.)

3.3 Agreements and Communication

Research participants have some understanding of what is occurring; how correct and complete is that

understanding? How informed and autonomous is the decision to participate?

3.3.1 Informed consent. There are many possible modes of communication between researcher and subject about the purpose, procedure, and intended outcome of research. What modes are most effective? What information should be included? How can comprehension and decision making ability be enhanced, especially when subject competency (e.g., of children, institutionalized persons, the mentally ill) is at issue? When must guardians and gatekeepers be included in the consent process?

3.3.2 Deception. This may include concealment, mental reservation, lying, self-deception, and distraction of subjects from the main purpose of the study in order to study spontaneous behavior. Deception raises many difficult questions, e.g., how does deception affect persons across their life span? What are methodological alternatives to deception? When is deception acceptable?

3.3.3 Relationships as a source of data. Qualitative research may involve getting to know persons then reporting on some aspect of their lives, often from a critical perspective that differs from the way the persons describe themselves. How is the research problem defined and is this disclosed to the persons? The nature of the relationship, as perceived by the participant and the researcher, may affect how the research problem is defined, how the data are collected and organized, how the research report is framed, how the results are disseminated and used, and whether the participant is wronged or harmed by this incursion into his or her private life. With the advent of the Internet, researchers are now free to observe or participate in chat room discussions of quite personal and intimate topics. What are their responsibilities to those they study? Should they reveal their presence as researchers? Will subjects' knowledge of the researchers' presence have a chilling effect on their chat room participation? What steps should be taken to disguise the identity of the chat room and its participants?

3.4 Acquisition, Presentation and Use of Data

The many facets of privacy, confidentiality, and data use have been topics of much theory, and subsequent empirical and methodological research.

3.4.1 Privacy. The need to establish boundaries between oneself and others is universal, but differs depending on learning, cultural, social, and developmental factors. People may want to disclose to researchers more than the researcher wants to know, or lie to protect privacy. Researchers' views on privacy may be inadequate for judging what privacies others want, expect, or create. How can researchers learn what people consider as private and respect that privacy? What new threats to privacy are emerging? New technology brings with it new threats to privacy, but also more powerful safeguards, and a new generation of privacy activists to establish better policies for deterring invasions of privacy.

3.4.2 Confidentiality. Part of the informed consent process typically includes an agreement about how the researcher will limit access of others to individually identifiable data. Many risks to confidentiality are not immediately obvious, e.g., subpoena of data, theft from computer files, or deductive identification. How can researchers understand sources of risk to confidentiality and prevent their occurrence? A wide range of statistical, procedural, and legal solutions continue to be developed.

3.4.3 Uses of data: data sharing and dissemination. How should scientific data be disseminated and used? Some data such as videotaped data cannot be rendered anonymous as individuals might be recognized; what safeguards are appropriate to protect the privacy of the research participants? Who may be benefited or harmed? Should harmful data be suppressed? How is misunderstanding of findings avoided? What should be included or omitted? What are effects of various modes of dissemination? What is the obligation of scientists to contribute to the scientific literacy of society? How should raw data be documented, archived and shared? Who owns data and knowledge (intellectual property)?

3.5 External Influences on Research

3.5.1 Government regulations. Studies of the regulatory process contribute to the development and improvement of research regulation.

3.5.2 Institutional review boards. Studies of IRBs focus on their processes, and on their effects on science.

3.5.3 Questionable, taboo or controversial topics of research. Some espouse prohibiting research that could cause irreparable social harm, produce misleading findings, or be used in politically harmful ways. Which is more harmful, suppression or publication of such research? Studies of the impact of research that blames the victim or otherwise seems to misdirect society's response to problems seek to answer this difficult question.

3.5.4 Scientific integrity and responsibility. Pressures to obtain grants and to publish may interfere with self-correcting scientific procedures, and motivate scientists to resort to unethical practices or outright fraud e.g., fabrication of data, plagiarism. Research has investigated conditions that inhibit or facilitate scientific responsibility.

3.5.5 Ethics and politics. Scientists who investigate controversial topics may be accused of scientific irresponsibility. Charges of scientific irresponsibility or immorality often mask profound political disagreements. Who sponsors or employs the researcher and how the data are used raise complex ethical and political questions. Because these issues usually arise in highly partisan settings, the ensuing conflict often fails to enlighten either side. Analysis of specific case studies can clarify the issues somewhat and suggest guidelines and safeguards.

3.5.6 Animal rights activism. An extraordinarily high degree of political activism has surrounded animal research. In what ways has this stimulated the advancement of theory and methodology? In what ways has it damaged scientific endeavors? Controversy surrounding specific cases of activism is usually aimed at proving the other side wrong, rather than resolving conflicts. Scholarly examination of activism and of the changes that have occurred as a consequence of conflict holds promise of guiding scientists to methods and principles that are both more acceptable morally and more beneficial scientifically.

3.5.7 Education in research ethics. In recent decades, there has been increasing concern to educate scientists and research administrators about issues of ethics. The goals have been, variously, to impart greater sensitivity to ethical issues, to teach approaches to conducting research ethically, and to examine and interpret current laws and regulations governing research. A variety of educational formats have been employed within Western societies. The rise of worldwide research programs on sensitive issues such as the epidemiology of AIDS and the intense ethical concerns this has raised, has now resulted in the development of ethics education programs in collaboration with scientists and applied ethicists in Third World countries. Despite significant expenditures on research ethics education, there is little understanding of the effects of these various education programs, of the relative effectiveness of the diverse approaches to teaching, or even of the best ways to measure their effectiveness.

4. Summary

While ethical theory and codes of ethics offer important general guidance, intelligent application requires understanding of how ethical problems may arise and of effective solutions to those problems. By its nature, social/behavioral research continues into new territory, raises new ethical issues, and requires research to explicate those issues and, where possible, to develop better methodology. Such research may be conducted as part of a project that raises an ethical issue, or as a 'stand alone' research activity. It may be the main focus or a sideline of one's research career. Whatever its origins, it deepens and broadens the ability of social and behavioral research to respect and be responsive to human (or animal) nature, to produce more valid results, and to better serve society.

See also: Animal Rights in Research and Research Application; Archiving: Ethical Aspects; Censorship in Research and Scholarship; Criminal Justice, Ethics of: Aspects of Human Dignity; Cultural Property and Ancestral Remains, Repatriation of; Deceptive Methods: Ethical Aspects; Ethical Dilemmas: Research and Treatment Priorities; Ethical Practices, Institutional Oversight, and Enforcement: United States Perspectives; Ethics Committees in Science: European Perspectives; Fieldwork: Ethical Aspects; Intellectual Property Rights: Ethical Aspects; Medical Experiments: Ethical Aspects; Mentoring: Supervision of Young Researchers; Normative Aspects of Social and Behavioral Science; Objectivity of Research: Ethical Aspects; Placebos, Use of: Ethical Aspects; Privacy of Individuals in Social Research: Confidentiality; Research Conduct: Ethical Codes; Research Ethics, Cross-cultural Dimensions of; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects; Research Subjects, Informed and Implied Consent of; Surveys and Polling: Ethical Aspects

Bibliography

Agre P E, Rotenberg M 1999 *Technology and Privacy*. MIT Press, Cambridge, MA

Gunsalus C 1998a Preventing the need for whistleblowing: Practical advice for university administrators. *Scientific and Engineering Ethics* **4**: 75–94
Gunsalus C 1998b How to blow the whistle and still have a career afterwards. *Scientific and Engineering Ethics* **4**: 51–64
Kimble G 1984 Psychology's two cultures. *American Psychologist* **39**: 833–9
Laufer R, Wolfe M 1977 Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of Social Issues* **33**: 44–87
Rosenthal R, Rosnow R 1984 Applying Hamlet's question to the ethical conduct of research: A conceptual addendum. *American Psychologist* **39**: 561–3
Stanley B, Sieber J, Melton G 1996 *Research Ethics: A Psychological Approach*. University of Nebraska, Lincoln, NE
Tarnow E 1999 The authorship list in science. *Scientific and Engineering Ethics* **5**: 73–88
Thompson R 1992 Developmental changes in research risk and benefit: A changing calculus of concerns. In: Stanley B, Sieber J (eds.) *Social Research on Children and Adolescents*. Sage, Newbury Park, CA

J. E. Sieber

Research Funding: Ethical Aspects

The issue of ethics in research funding, a feature of research policy generally, is closely connected with the issue of freedom of research, an ideal of high standing in the science community. Special focus is on relations between providers and recipients of funding (see Orlans 1967). A vitally important principle says that research must be free in the sense of being shielded against any influences that may interfere with the obligation to seek truth. Yet, research must cope with regulation and restrictions, and discussions of these issues are concerned with questions of what freedom in which areas and with what consequences (Alver and Øyen 1997). The quest for freedom of research requires that conflicting and contradictory concerns and objectives be weighed against each other. In the process, the consideration of the policies, principles, procedures, and objectives of research funding often demands ethical reflection. Some principles of ethics seem self-evident. For example, no source of funding may be used to 'buy' a researcher to produce 'right' answers. Other issues require careful consideration—and there may be no clear-cut solution.

1. Sources of Ethical Concern

A sound process of research funding depends on organization and guiding principles. Research must meet standards of quality under conditions that favor innovation and contribute to knowledge accumulation. Research ought not to be done on problems that have already been solved and where sufficient knowledge exists. But who ought to organize and govern? The science community may favor the view that the researchers themselves are in the best position to face new challenges and play key roles in deciding what research objectives to pursue. Nevertheless, responsibility for research must be shared between several partners: political authorities, research councils, universities and other research agencies, and executors of concrete research tasks. Research is cost demanding and requires highly qualified personnel resources and efficient infrastructure services. Also, research involves risks to people and environment. Priorities have to be set independently of the aspirations of researchers, and particularistic ambitions require tempering in view of universalistic standards. Coordination is needed to secure teamwork or to assure that researchers in different locations contribute their share of a totality. Funding agencies must deal with such issues, often involving rivalry among research institutions, among representatives of disciplines or fields of specialization, and among individual researchers. All of these matters may evoke ethical concern. The main point is to assure sufficient freedom and autonomy, protect the research process against unnecessary control, and guard against interference having undue effect on conclusions.

Thus, a decision to fund research is a result of a process involving several actors and partners whose goals hardly ever are identical. Provider and recipient of funding are main partners, but they cannot relate to each other without considering the interests of the context from which they act. The funding agency cannot defy interests of its hinterland or constituency, whether voters, taxpayers, or stockholders. The recipient of funding must relate to norms of scientific conduct, principles of data access and use, interests of institution of affiliation, and personal career interests and breadwinning needs. Ethical concerns become accentuated when different values and normative systems confront each another, and thus through the process of defining unified direction, seeking alignment of goals, and establishing common ground.

Traditionally, the institution of science asserts moral obligation to seek 'knowledge for its own sake,' without specific concern for applications. The ideal of basic or 'pure' research, driven by intellectual curiosity and awareness of missing links in the body of knowledge, plays a fundamental role in building and updating an integrated knowledge base. The universities have special responsibility for basic research and its integration in higher education. However, the mechanism whereby basic research is funded is going through significant changes as the role of research assumes new dimensions. Scientific discoveries accelerate change in society, but at the same time, scientific research becomes an instrument for solving problems brought about by change. Events during the

twentieth century contributed to widespread distrust in knowledge as intrinsically good, and perceptions of evil consequences of scientific findings brought public opinion to hold that some discoveries ought not to have been made—and thus not funded. These and other circumstances create a demand for more control and more pronounced goal orientation. Some funding agencies routinely demand that even basic research be justified in terms of applicability. Then, as an increasing proportion of research funding goes to applied research, the goals, and subsequent utilization of findings, become matters for ethical evaluation.

Standards of applied research bear heavily upon criteria of assessing the merits of basic research. The science system demands investment priorities, coordination for optimal utilization of funding, efficiency, and a favorable cost–benefit ratio. In academia, the freedom to pursue theories challenging scientists' imagination is under pressure from funding sources needing quick answers to urgent problems. But at the same time, large sectors of the science community are morally devoted to research on issues of environment, poverty, health, inequality, and political instability, to mention but a few critical themes.

2. *Sources of Research Funding*

Funding for research comes from a multitude of sources. Governments channel public funds through foundations or research councils to individual researchers and to research establishments. Industrial corporations engage researchers as their employees or commission researchers on contract for particular research jobs. International organizations have special budgets for research. Private endowments and voluntary organizations provide funding for specialized research areas. The pluralism of research funding provides opportunities for research across an extensive array of issues and problems. Yet, there will always be concern that funds are skewed in favor of certain research goals at the expense of others. Not surprisingly, a host of ethically loaded questions concern funding priorities and processes of assuring optimal conditions for research.

2.1 *Funding from Public Sources*

Political authorities establish guidelines and policy principles for research, usually without entering into details of strategic decision making or interfering in the design of programs or projects. However, when investments in research are perceived as vitally important in view of political or national interests, funds may become earmarked for specific purposes. In rare instances, politicians address questions of whether certain disciplines are worthy recipients of funding from public sources, as in the recurrent debates in the US Senate about the place of social sciences in the National Science Foundation (Larsen 1992).

Generally, the science community has a strong hand in strategic decision making in research funding, as much authority is delegated by politicians to science foundations, research councils, and academies of sciences, where researchers influence policies, grant funds, and resolve ethical issues. However, their performance may be subject to evaluation by political authorities in terms of criteria of political expediency and objectives. The representatives of the science community may have to accommodate to demands from above, under threats of reduced flows of public funds. Within this contest area, the flow of public funds to basic research is particularly vulnerable, as evaluation tends to follow criteria of applied research.

Increasing internationalization of research requires attention to issues of compatibility across national boundaries of legal and ethical principles bearing on the use of public funds, as for example, privacy protection and confidentiality provisions. Standards of ethical conduct have been adopted by most professional organizations, while modern versions of globalization, internationalization, and interdisciplinarity of research bring up new issues needing continued attention. For example, when research funds transgress national boundaries, sensitive issues of sponsorship and of true or hidden agendas may stir up moral and political controversy, as in the case of 'Project Camelot' (see Horowitz and Katz 1975, Homan 1991).

2.2 *Funding from Corporate Sources*

In corporate funding, a corporation enters into a contract with a researcher or a research agency engaged to produce knowledge applicable for a defined purpose. Normally, a contract is preceded by negotiations to secure mutual rights and duties. It may not be possible to foresee all problems involving ethical aspects that may come up in the course of an investigation, but it is always helpful when potential sources of problems may be dealt with at the time of negotiating a contract. In applied research, the linkage of funding and spending often involves a certain urgency and a limited time horizon. While the funding source has an interest in attaining a useful research result quickly and efficiently, the investigator is eager not to sacrifice researcher rights. The investigator must relate to ethical principles adopted by the professional association and also wishes to clarify questions of ownership to research data and results, respect for rules of confidentiality, freedom to publish—and perhaps, freedom to have second thoughts about the results of an investigation. A funding source may want provisions for secrecy or first use, while the

researcher is on guard against any clause that restricts truthful reporting. Other issues requiring clarification involve the funding agency's participation in planning and execution of research, terms of monitoring a project while in progress, publicizing or utilizing partial information before the conclusion of a project, and terms of application of findings. Both parties in a contract negotiation must be acquainted with relevant legal provisions concerning intellectual property rights. Attention to 'small print' may require qualified advice, and researchers may be well served when the institution with which they are affiliated provides backing and takes part in contract negotiations.

3. *Researcher Influence on Funding*

Research policy formation as well as decision making in matters of concrete research tasks requires input from the science community. Science policy advice may be solicited or volunteered; in either case, the decision-making process invites attention from lobbyists and special-interest groups seeking influence on policies, budgets, and the allotment of funds. Although the process is regarded as necessary and legitimate, it may generate many ethical issues that concern political authorities, funding agencies, the science community, and the general public. It is assumed that those scientists who possess the highest professional competence within highly specialized fields are best qualified to serve as judges and gatekeepers for research programs and initiatives. However, as research plays an important role in the pursuit of social, economic, and political goals, the role of scientists in research funding, as policy advisors and as participants in the decision-making process, may to some appear too closely interconnected with the political process.

Furthermore, fair disbursement of research funds requires safeguarding against unfair delimitation of turf, undue promotion of special interests, and questionable manipulation of the research opportunities of friends and foes. Yet, measures to secure competence and impartiality ought not to lead to the paradoxical situation that researchers having specialized knowledge of an area have to recuse themselves from participating in evaluation and decision making within that area. In fact, the system needs mechanisms for the protection of the integrity of the science advisor.

4. *Freedom, Market, and Control*

Research has become closely linked to the solution of pressing problems encountered through rapid social change. At the same time, the possibilities of controlling research—and the desire to control—have become stronger. With increasing control, in a situation where research functions much like a market commodity, the center of gravity is shifting from basic to applied research. Increasingly, the firm linkage of research to practical utilization determines the criteria by which decisions to fund research are made. For the research community, there may be a price to pay in terms of freedom of research and standards of research. Surely, there will be more attention paid to the ethical issues surrounding decision making and policy formation in the area of research funding.

See also: Censorship in Research and Scholarship; Intellectual Property Rights: Ethical Aspects; Objectivity of Research: Ethical Aspects; Research Ethics: Research; Research Publication: Ethical Aspects; Science Funding: Asia; Science Funding: Europe; Science Funding: United States

Bibliography

Alver B G, Øyen Ø 1997 *Forskningsetikk i forskerhverdag*. Tano Aschehoug, Oslo, Norway

Homan R 1991 *The Ethics of Social Research*. Longman, London

Horowitz I L, Katz J E 1975 *Social Science and Public Policy in the United States*. Praeger, New York

Larsen O N 1992 *Milestones and Millstones: Social Science at the National Science Foundation, 1945–1991*. Transaction, New Brunswick, NJ

Orlans H 1967 Ethical problems in the relations of research sponsors and investigators. In: Sjoberg G (ed.) *Ethics, Politics, and Social Research*. Schenkman, Cambridge, MA, Chap. 1, pp. 3–24

Ørjar Øyen

Research Publication: Ethical Aspects

Ethical aspects in research publications are all the moral issues and problems raised by any behavioral misuse or abuse of the system for communicating and disseminating scientific results and information within the international scientific community. It is intended as a social body genuinely co-operating for the cultural and material improvement of mankind regardless of any personal, political, economic, or racial interests. Misuses or abuses include plagiarism, fraudulent and repetitive publication, guest- and ghost-authorship in multi-authored papers, conflict of interests in peer reviewing, censoring, and rejecting publications.

Scientific knowledge is community property unceasingly increased by individuals or groups in mutual co-operation and fair and honest competition. Even in the professional and competitive world of today researchers are intended as primarily engaged in the disinterested pursuit of truth and in the quest of new

insights and discoveries. Honesty towards oneself and towards others is considered as a fundamental condition for co-operation and fair competition since researchers depend on each other and cannot be successful unless they are able to trust each other and their predecessors. Publications are the primary medium through which scientists give an account of their work, co-operate, contribute to the advancement of knowledge, and claim the rights to be recognized for their discoveries. Therefore, dishonesty in publication not merely throws research open to doubt but destroys the very fundamentals of science.

The publication of a fraudulent article (the very outcome of a scientific fraud consisting in the fabrication, falsification or in 'cooking' and manipulation of statistical data) remains the most serious transgression of ethics of scientific publication. It is commonly considered to be an essential part of the scientific misconduct as well as plagiarism, the deliberate presentation of another's ideas or texts as one's own. Compared to these repetitive publications (publishing the same article in different journals), sequential publication (reporting follow-up of the same study with additional subjects but without new results), 'salami slicing' (dividing up different parts of the same study for publication in different journals), and guest and ghost authorship (giving or not giving authorship credit to persons who did or didn't take part in the research or in drafting or revising the article) are all considered as minor transgressions. These result in a huge amount of unwieldy and polluted literature that distorts the direction of science. It has been roughly calculated that flawed papers published on the 5,600 scientific journals indexed by The Science Citation Index amount to about 10,000 each year. This raises the problem of the reliability of the peer-review system designed to assess the quality of scientific literature, and that of the fate of the invalid ideas and data generated by fraudulent publication (Friedman 1990).

Good scientific journals today publish original articles only after they have been examined by competent reviewers for their validity and originality. Antecedents of peer review practices go back to the seventeenth century and the first organizations of scientific societies in Italy, France, and Germany but the modern system was only fully developed after World War II and it was adopted firstly and mainly in the Anglo-Saxon cultural area. The system is credited to have greatly contributed to the efficiency and professional organization of the scientific enterprise but since the eighties it has been severely criticized not only for its enormous costs in time and money but also for the progressive loss of accuracy (Lock 1986). Reviewers have indeed some moral obligation to make every necessary effort to recognize manipulations and falsifications. However, to expect irregularities to be reliably detected would be misguiding: the original data are not available to reviewers, and if they were they would not have time to replicate experiments and observations. This is why, according to many, peer review is so vulnerable to dishonest conduct. Other flaws have been detected in negligence (it resulted from inquiries that the higher the academic status and age of the reviewer, the lower the quality of the review), in vested interests of the referees and in many forms of bias (against particular individuals, subjects, or institution), the worst of which is that against the innovators. Some critics maintain that the peer review system favors mediocre basic science and fails to recognize and promote creativity, originality, and innovation (Horrobin 1990, Campanario 1996). Even the anonymity of the referees has been questioned for its finality as contrary to the moral norms of universalism, communality, disinterestedness, and organized skepticism that are thought to characterize the ethic of science. In fact the peer review system holds some risks for the authors because ideas, research findings, and texts which are unprotected by patents or intellectual property rights are submitted to persons whose identity is unknown to the authors and who may happen to be their direct competitors.

As for the fate of invalid data generated by transgressions to the ethics of scientific communication, the scientific community has yet to agree on the level of inaccuracy required to mandate a retraction vs. an erratum or who may retract an article. Although early in 1987 the International Committee of Medical Journal Editors published guidelines for the format and terminology of retraction notices, articles known to be fraudulent may be not retracted at all, or are retracted with ambiguous retraction. In other cases, only part of the data or conclusions are retracted. Editors of journals have little leeway for action since publishing corrections is fraught with juridical risks unless all authors jointly sign them. However, omitted or ambiguous retractions are not the only reasons of invalid work not being effectively purged from the scientific literature. It resulted from inquiries on US authors (Pfeifer and Snodgrass 1990) that after retraction the invalid information they produced continues to be used since the majority of retractions are not indexed by their own journals and no reliable source exists to locate fraudulent or erroneous work.

Some ethical relevance also has the soft censoring activity by which administrations and companies tend to control and restrict the dissemination of sensitive scientific and technological knowledge on the grounds of national security and economic reasons (Relyea 1994).

Such ethical problems depend on the very structure of the scientific enterprise as shaped in the seventeenth century but they have only grown in relevance and extension in recent times. Cases of plagiarism or fraudulent publication have been noticed in the past history of science, some of which involving such outstanding figures as Ptolemy, Luca Pacioli, Ernst Haeckel and even Galilei and Newton. However, it is

generally admitted that if not the number, the relevance and the social impact of the transgressions to the ethics of scientific publication began to rise in the 1940s. This was when modern science policy was established in view of the expected contributions of the scientific community to national defense and industrial progress. The new reward and funding system enhanced the traditional competition between scientists and it was then, in 1942, that the sociologist L. Wilson apparently coined the expression 'publish or perish.' It could hardly be said however that in the first 30 years competition did in some way sensibly promote fraudulent behaviors or the pollution of scientific literature. The turning point has in fact been stated in the late 1970s when the social structure of the scientific community underwent some crucial changes due mainly to the exponential increase of the number of people involved in research and the proportional (though not absolute) decreasing rate in assigning funds to research activity. The ensuing economic constraint (upon which the resources poured in by private companies had no or little effect) led to a situation of even stronger competition among scientists for academic rewards, position, or research funds. One crucial decision was that of the adoption of publication as the main performance indicator and of measuring productivity as the number of publications per length of time. The additional tool of citation analysis which was introduced in those very years (Garfield 1972, 1979) and based on the key concept of 'impact factor', (as the ratio of the number of citations a journal receives to the number of papers published over a particular time period) and then extended to calculate the impact of the work of individuals, groups, or departments generated a race to produce, in the shortest possible time, the least publishable unit in a journal with the highest impact factor. At the same time the growth of the scientific population implied a statistical decline in ethical standards and in the quality of production. Lotka's law of productivity in fact states that in a certain field, half of the scientific literature is produced by a population of the square root ($\sqrt{N}$) of all N scientists. De Solla Price pointed out that the exponential growth of the scientific population implies a proportional increase in the number of scientists able to author professional papers, but not important and innovative articles. From a statistic compiled by the Institute for Scientific Information (ISI) 55% of the scientific papers published between 1981 and 1985 in journals indexed by the Institute received no citation at all (Hamilton 1990). This led some students to draw the conclusion that more than half, or perhaps more than three-quarters, of the scientific literature are worthless. Others reacted against the notion that most uncited papers are without value. As for the use of citations as a measure of impact and of the quality of a publication, it has been questioned whether this may not lead to abuses such as citation cartels. It was estimated by the same ISI, that self-citation accounts for between 5 percent and 20 percent of all citation.

It is from, or in any case within, those changes that according to the current views (see, for example, Petersdorf 1986) the pathological phenomenon of scientific misconduct arose to which the major ethical issues in scientific publishing are tightly related. As a consequence the analysis of these issues, the evaluation of the frequency and gravity of the transgressions, and the discussions on how to manage the problems they pose have been conducted as part of a general study on scientific misconduct.

Unfortunately, despite increasing publicity during the past 25 years, attempts to assess the quantitative dimensions of the problem of scientific misconduct did not lead to conclusive results: its magnitude is unknown and its detection largely serendipitous. Guidelines and actions adopted to manage scientific and (as a consequence) publication frauds and other abuses vary according to which of the two main different approaches they are based on. Sociologists and the vast majority of scientists share the opinion that the frequency and incidence of misconduct did not sensibly change during the twentieth century. The impression that fraud has vastly increased may be the result of the heightened self-consciousness in society coupled with strong moral expectation that scientists in the pursuit of reliable knowledge will live up to a higher standard of probity. This view seems to be substantiated both by the reports of the two most important US authorities responsible for dealing with misconduct cases, the Office of Inspector General (OIG) of National Science Foundation (NSF), and the Office of Scientific Integrity (ORI) of the Public Health Service and by that of analogous authorities in Denmark, Norway, Finland, and Australia. The ensuing suggested strategy is firstly, and mainly intended to save the autonomy right and self-regulation privilege of the scientific community from legal or government unnecessary intervention. It also recommends that inquiries primarily be made by local academic bodies, at the same time stressing the crucial role of education to improve the ethical standards of researchers.

Opposite is the view according to which, though there are no official data on the scale of the problem, the reports suggest that only a comparatively small proportion of cases of misconduct is actually detected and that what emerges is just the tip of an iceberg (Lock 1993). According to this view, self-regulation has proven to be ineffective and unsuitable. Scientists are not trained in conflict and conflict of interests resolution. Their intuitive response is usually wrong and they tend to set up shaky ad hoc procedures that do not guarantee the accused notice of all the charges; the opportunity to respond to all the charges and to the evidence; and a decision based on rigorous standards. In this view the setting up of national agencies, rather that local committees, is recom-

mended. The attention is focused on the economic and social background. It is assumed that those transgressions simply reflect the problems of a science which is today too big, too entrepreneurial, and too competitive to guarantee its very scope—the cultural and material improvement of mankind, in due and reasonable proportion to the intellectual and economic resources and at the same rate of the past three centuries. As far as scientific publication is concerned this second approach led the Committee on Publication Ethics (COPE), created in the UK by 20 editors of scientific journals, to ask for an independent agency to detect, handle, and prevent research publication abuses.

However, there is no general agreement on this view that led some of its proponents to recommend for some radical adjustment in the scientific policy first of all in connection to the ratio between scientific population and funds. While universities, scientific societies, funding agencies, and governments have opposed those proposals tacitly or patently, even their advocates acknowledge the importance of improving the ethical standards via the teaching of research ethics. This is currently considered by competent institutions and single students, besides and before, any disciplinary, economic, or legal action. Likewise, all agree on the necessity of suitable ethical and professional guidelines and measures which have in fact been proposed or adopted, though independently and according to different experiences and aims, by learning societies, agencies, journals, or individual students.

The primary importance of ethical (coupled with methodological) education emerge particularly in connection to some of the problems of scientific publishing, which seem less directly linked to the social and economic structure of the scientific system than to the peculiarity of research areas which involve cultural diversity or human and animal rights. In fact though science is largely transcultural, the human community has not yet developed a completely corresponding body of transcultural ethics. In some areas this may affect both methodology and ethics. Uncritical and unwitting reliance on the researcher's own culturally received principles and values regarding personal and physical identity, health, loving, suffering, death, etc. may generate potentially distorted initial observations which can lead to systematic divergence from truth in anthropological, psychological, sociological, or medical inquiries (where also systematic infringements of human or animal rights or the exploitation of subjugated populations may occur). Scientific publications based on such inquiries are, in principle, to be considered as poorly or improperly designed and performed studies. Current ethical and methodological standards do not include clear and effective rules for authors, editors, and peer-reviewers though the problems they raise have been debated in connection to some controversial cases.

A first general attempt to mitigate the pressure to publish that lies behind most deviant practices is, according to many, to place a ceiling on the number of publications that can be considered in evaluating a candidate for promotion or funding. However, some experts have also suggested that the number of scientific journals be drastically reduced (Bracey 1987). On the same line; although the use of quantitative indicators of scientific performance seems unavoidable, it has been recommended that impact factors shall not by themselves serve as the basis for funding decisions and that the quality of publication must be the primary consideration (Bucholz 1995). While no effective measures (other than severe sanctions) seem foreseeable against major authorship transgressions, as plagiarism or the publication of fabricated data, it has been recommended to journals to cooperate in cross-checking control activities and to research institutes and universities to store the primary data on which publications are based. All sources and methods used to obtain and analyze those data should be, when required, fully disclosed and the disappearance of primary data from a laboratory considered as an infraction of the basic principles of careful scientific practice. On the other hand, since there is no universally agreed definition of authorship, specific criteria and requirements have been invoked as an acknowledgment or footnote in every co-authored paper to state the exact contribution by each author. Or, the signing of a form where each author affirms having seen the final version and agrees to its publication. According to those proposed criteria the award of authorship should balance intellectual contributions to the conception of studies or experiments, to the generation, analysis, and interpretation of the data, and to prepare the manuscript against the collection of data and other routine work. If there is no task that can reasonably be attributed to a particular individual, then that individual should not be credited with authorship. In principle all authors must take public responsibility for the content of their paper and honorary or guest authorship is in no way acceptable. Against repetitive publication since 1969 the so-called 'Ingelfinger rule' has been proposed (Angell and Kassirer 1991) according to which papers are submitted to journals with the understanding that they (or their essential part) have been neither published nor submitted elsewhere. Reviewers are supposed to vigilate upon transgressions and it is stated in some journals' guidelines that if they suspect misconduct, they should write in confidence to the editor. As for the same reviewers it is reminded that they are bound to the duty of confidentiality in the assessment of a manuscript and asked to disclose conflict of interests and to provide speedy, accurate, courteous, unbiased, and justifiable reports. The submitted manuscript should not be retained or copied and both editors and reviewers should not make any use of the data, arguments, or interpretations, unless they have the

authors' permission. Editors should regard their function as ensuring that authors are treated fairly and their decisions to accept or reject a paper for publication should be based only on the paper's importance, originality, and clarity. To avoid the suppression of originality and creativity it has also been recommended that studies that challenge previous work should be given an especially sympathetic hearing and that independent ombudsmen for journals be appointed. Some few suggestions have also been advanced on the subject of external pressures as, for example, that editorial decisions must not be influenced by advertising revenue or reprint potential, but it seems that little can be done for the secrecy imposed by industries or governments on sensitive research fields.

Envisaged actions and sanctions rank from a letter of reprimand and warning as to future conduct, or publication of a notice of redundant publication or plagiarism, to formal withdrawal or retraction of the paper from the scientific literature, informing other editors and the indexing authorities or reporting the case to the authorities or organizations which can investigate and act with due process.

There is however no suitable insight on the impact those guidelines and measures had, or may have, in the long term.

See also: Archiving: Ethical Aspects; Ethics Committees in Science: European Perspectives; Intellectual Property Rights: Ethical Aspects; Objectivity of Research: Ethical Aspects; Research Ethics: Research

Bibliography

Angell M, Kassirer J P 1991 The Ingelfinger rule revisited. *New England Journal of Medicine* **325**: 1371–3

Bracey G W 1987 The time has come to abolish research journals: too many are writing too much about too little. *Chronicle of Higher Education* **30**: 44–5

Bucholz K 1995 Criteria for the analysis of scientific quality. *Scientometrics* **32**: 195–218

Budd J M, Sievert M-E, Schultz T 1998 Reasons for retraction and citations to the publications. *Journal of the American Medical Association* **280**: 296–7

Campanario J M 1996 Have referees rejected some of the most-cited papers of all times? *Journal of the American Society for Information Science* **47**: 302–10

CIOMS 1991 International guidelines for ethical review of epidemiological studies *Law, Medicine & Health Care* **19**(3–4) (Appendix I): 247–58

Committee on Publication Ethics 1998 *Annual report* 1998.BMJ Publishing Group London

De Solla Price J D 1963 *Little Science, Big Science*. Columbia University Press, New York

De Solla Price J D 1969 Measuring the size of science. *Proceedings of the Israel Academy of Science and Humanities* **4**: 98–111

Friedman P J 1990 Correcting the literature following fraudulent publication. *Journal of the American Medical Association* **263**: 1416–9

Garfield E 1972 Citation analysis as a tool in journal evaluation. *Science* **178**: 471–9

Garfield E 1979 *Citation Indexing: Its Theory and Application in Science, Techonolgy and Humanities*. Wiley, New York

Hamilton D P 1990 Publishing by—and for?—the numbers. *Science* **250**: 1331–2

Horrobin D F 1990 The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association* **263**: 1438–41

LaFollette M C 1992 *Stealing into Print: Fraud, Plagiarism and Misconduct in Scientific Publishing*. University of California Press, Berkeley, CA

LaFollette M 2000 The evolution of the 'scientific misconduct' issue: An historical overview. *Proceedings of the Society for Experimental Biology and Medicine* **224**: 211–5

Lock S 1986 *A Difficult Balance: Editorial Peer Review in Medicine*. ISI Press, Philadelphia, PA

Lock S 1993 Research misconduct: a résumé of recent events. In: Lock S, Wells F (eds.) *Fraud and Misconduct in Medical Research*. British Medical Journal, London pp. 5–24

Petersdorf R G 1986 The pathogenesis of fraud in medical sciences. *Annals of Internal Medicine* **104**: 252–4

Pfeifer M P, Snodgrass G L 1990 The continued use of retracted, invalid scientific literature. *Journal of the American Medical Association* **263**: 1420–3

Relyea H C 1994 *Silencing Science: National Security Controls and Scientific Communication*. Ablex, Norwood, NJ

Serebnick J 1991 Identifying unethical practices in journal publishing. *Library Trends* **40**: 357–72

Shankman P 1996 The history of Samoan sexual conduct and the Mead–Freeman controversy. *American Anthropology* **3**: 555–67

Sponholz G 2000 Teaching scientific integrity and research ethics. *Forensic Science International* **113**: 511–4

Stossel T P 1985 SPEED: An essay on biomedical communication. *New England Journal of Medicine* **313**: 123–6

Whiteley W P, Rennie D, Hafner A W 1994 The scientific community's response to evidence of fraudulent publication. *Journal of the American Medical Association* **272**: 170–3

F. Di Trocchio

Research Subjects, Informed and Implied Consent of

'Informed consent' denotes a person's autonomous agreement to a proposed course of action, usually in a treatment or research context. Consent is not considered to be informed unless the person: has received disclosure of the information that a reasonable person would consider material to the decision at hand; has the capacity to use that information in a rational process of decision making; and reaches a decision in the absence of coercive offers or threats (Appelbaum et al. 1987a). Broad agreement exists that informed consent is an indispensable requirement for the ethical conduct of research with human subjects, at least

when the research renders them liable to more than minimal risks of harm. But considerable controversy remains regarding the precise contours of the doctrine and of the circumstances that might constitute exceptions to the general rule. And since the rules have been devised largely with biomedical research in mind, issues more characteristic of social and behavioral science research have been relatively neglected.

1. *History of Informed Consent to Research*

The history of informed consent to research is usually traced back to the decision of the court in the Nazi doctors' trial at Nuremberg after World War II, which declared, 'The voluntary consent of the human subject is absolutely essential.' It is clear, however, that the Nuremberg tribunal neither originated nor established the concept of consent to research. References to requirements for consent from research subjects can be found in the writings of physician experimenters and in governmental pronouncements dating back to the late nineteenth century, when experimental medicine grew out of its infancy. Indeed, George Bernard Shaw's *Pygmalion* contains an explicit statement of the principle. Two of the leading instances, ironically, derive from Germany, where the advanced state of empirical medical science made consent an object of at least periodic concern (Grodin 1992). Nonetheless, however much the Nuremberg court desired to root the principle in contemporary practice, it appears that researchers in many countries, even after the Nazi atrocities came to light, frequently neglected to obtain consent from their subjects.

Nor did the Nuremberg decision immediately change these practices. Investigators using human subjects apparently saw the Nuremberg Code, as the ten principles embodied in the court's decision came to be known, as applying only to the anomalous situation of concentration camp inmates. Thus, the history of the postwar decades is marked by repeated revelations of questionable ethical practices, including studies in which persons were subject to clear harms without their knowledge or consent (for one of the most influential reports of problematic studies, see Beecher 1966). Pressure built for clearer rules and mechanisms of enforcement.

The World Medical Association responded to these concerns with its *Declaration of Helsinki* in 1964 (World Medical Association 1964), the first attempt to forge an international professional consensus on the ethics of research. The *Declaration* endorsed a requirement for informed consent, albeit in less sweeping terms than the Nuremberg Code. National professional organizations developed similar documents, but their nonbinding character left research subjects with few certain protections. A move to legally enforceable rules came in the USA in 1974, with the adoption of regulations governing federally funded research. Most of the industrialized countries in the world have since followed suit, by way of legislative or regulatory enactments, or binding guidelines (for a summary of the status of research regulation internationally, see Brody (1998). The general principle that research should not be conducted without the informed consent of participants has been accepted around the world.

2. *Current Contours of Informed Consent to Research*

The requirements for informed consent are rooted in the ethical principle of autonomy, which demands respect for the free, informed choice of a competent person (Faden and Beauchamp 1986). Valid informed consent requires three elements to be present: disclosure, decisionmaking capacity, and voluntariness.

2.1 *Disclosure*

The Nuremberg Code (1948) declared that the research subject 'should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision.' Brody (1998) summarizes the international consensus regarding those items that need to be disclosed to achieve this end as follows:

'(a) the purpose and the research nature of the project;

(b) the duration of the subject's participation, the treatment and procedures that the subject will undergo (with a clear specification of which are experimental), and the subject's responsibilities;

(c) the risks of harms and discomforts that might result from participating in the research;

(d) the possible benefits to the subject or to society from the research;

(e) the alternatives to participating in the research;

(f) the right not to participate and to stop participating at any time without penalty or loss of benefits to which the subject is otherwise entitled;

(g) the measures to protect confidentiality;

(h) the anticipated payments, if any, to the subjects and the anticipated expenses, if any, that will be borne by the subject;

(i) the compensation and treatment available for injuries sustained in the research;

(j) the identity of a contact person to discuss questions about the research;

(k) the circumstances under which the subject's participation will be terminated;

(l) the assurance that new information relevant to participation that becomes available during the research will be provided to the subject;

(m) the names of the sponsors of the research and the ownership of any commercial results.'

This information is generally provided in written form, though often supplemented by verbal disclosures. Frequently, the result is a lengthy form (12–14 page, single-spaced forms are not uncommon in clinical research) that is difficult for potential subjects to read and assimilate. Provision of a mass of detailed information means, moreover, that important data, such as those dealing with significant risks, are frequently neglected by subjects. A common resulting problem is the so-called 'therapeutic misconception'—a belief on the part of subjects that procedures are being undertaken for their benefit (including randomization and use of placebos), when they are actually designed solely to advance the research effort (Appelbaum et al. 1987b). Though investigators are often urged to provide additional disclosures in more understandable formats (e.g., illustrated brochures that subjects can read at their leisure), this still represents the exception rather than the rule.

2.2 Decision-making Capacity

The subject's capacity (in legal terms referred to as competence) to give a meaningful consent was considered a *sine qua non* in the Nuremberg Code. Although formulations of the requirements differ somewhat across jurisdictions, decisionmaking capacity generally involves the abilities to understand the information relevant to a decision, to appreciate its implications for one's own situation, to reason in a rational fashion, and to evidence a choice (Appelbaum and Roth 1982).

Capacity is generally presumed to be present, unless there is reason to believe otherwise. When studies involve populations at elevated risk of incapacity, however, including persons with severe mental disorders, mild dementia, head injury, and the like, particularly when the research may present significant risks to participants, investigators are increasingly assessing capacity prospectively. This can be done by means of clinical evaluations or formal screening instruments. Identification of a potential subject as having impaired capacities does not necessarily mean that the person must be excluded from research participation. Many such persons can be brought up to acceptable levels of functioning with extra educational efforts.

Were capacity to be an absolute requirement, as the Nuremberg Code suggested, no research could be conducted with children—who lack capacity as a matter of law—or with many patient groups, including potential subjects who are comatose, suffering from moderate to severe dementia or mental retardation, or severely psychotic and disorganized. The desirability of allowing research to be conducted with these groups, so that progress might be made on the disorders from which they suffer, led the drafters of the Declaration of Helsinki, and most subsequent compilers of codes, regulations, and statutes, to make some provision for the entry of incapable subjects into research projects. Rules governing their recruitment differ, but common limits include that the research be directly applicable to their conditions, not able to be performed with capable subjects, hold out the prospect of some therapeutic gain, or present no more than minimal risks of harm. Innovative proposals that would allow persons to make advance decisions while competent about whether to enter research projects should they later become incompetent have been much discussed, but rarely implemented.

2.3 Voluntariness

Overt coercion—the application of threats or unfair offers to a prospective research subject—is universally recognized as vitiating that person's decision and thus rendering consent invalid. The effects of more subtle influences on the legitimacy of subjects' consent to research is a matter of greater controversy. Some environments may be considered so intrinsically coercive—prisons, for example—that research with the persons detained therein has been severely restricted in many countries. Whether hospitals, institutions for the mentally ill or retarded, schools and colleges, places of employment, and other entities fall into that category is more difficult to say. Paradoxically, efforts to protect inmates, patients, workers, and others by prohibiting their participation in research deprive them of their residual decisionmaking power, and may increase their feelings of helplessness and their perceptions of being coerced.

Research subjects are often compensated for their participation with money or its equivalent. Fair compensation for the time expended in a study is generally considered acceptable. To the extent that payment is so substantial as to lead subjects to give insufficient attention to the risks they may run by participating, however, it may be considered an undue influence on their decisions, which renders their choices substantially involuntary.

3. Special Considerations in Behavioral and Social Science Research

Rules governing informed consent to research, from the Nuremberg Code to most contemporary governmental regulations, typically have been formulated with biomedical research in mind. Issues more characteristic of social and behavioral research have been relatively neglected. Indeed, the federal regulations in the USA categorically exempt many types of behavioral and social research from their purview, including: educational research; studies designed to evaluate governmentally provided benefits or services; and testing, surveys, interviews, or observations of

public behavior, unless information is recorded in a manner that would render subjects identifiable, and subjects would be likely to suffer harm if the information were released (Code of Federal Regulations 1991).

Yet, many behavioral and social science studies involving human subjects raise important issues related to informed consent. One of the major questions is when, if ever, it is appropriate to deceive subjects about the purpose of a research study, a frequent technique in psychological research (see *Deceptive Methods: Ethical Aspects*). Other issues concern consent for access to records, consent in observational and field work studies, and the appropriateness of seeking consent—or some equivalent—from entire communities that may be affected by an intervention that is part of a research project.

3.1 Consent in Studies Based on Review of Records

Analysis of existing records, often aggregated into large databases, is a frequent technique used by behavioral and social scientists. These records may have been collected originally with subjects' consent but for nonresearch purposes (e.g., medical records) or without subjects' consent (e.g., national birth registries, educational testing results, or census information). When individual subjects are not reasonably identifiable in the database, it is generally agreed that consent need not be obtained prior to granting researchers access. Even with identifiable data, researchers have usually been allowed to utilize such records as long as they agree to protect the confidentiality of those persons whose information they are reviewing (e.g., by publishing data only in aggregate form).

This practice has been criticized as insufficiently protective of subjects' interests in both autonomy and privacy. The autonomy argument is predicated on the belief that persons should have the right to control access to information about them, preventing its use, for example, for research purposes of which they may disapprove. Privacy concerns have been given greater weight of late with the growing ability of researchers to correlate information obtained from different databases about the same individuals, thus accumulating large amounts of information that, were it to be used or accessed for unauthorized purposes, could prove embarrassing or harmful to the people involved. Strong arguments have been made for obtaining consent from the subjects of these records prior to their release for research purposes.

Researchers object to these proposals on the grounds that they would materially increase the costs of research and, insofar as subjects' refusal of consent would bias the resulting data, decrease the validity of their findings. In many cases, it will simply be infeasible to track down and obtain consents from the persons whose data is in question. No reasonable argument can be made for precluding researchers' access in these circumstances, so long as the study is of some significant value and appropriate safeguards are in place. When smaller and more recent repositories of data are involved, and obtaining consent would not involve unreasonable expenditures of resources, it is clearly more respectful of subjects' interests to seek their agreement. When records are being compiled prospectively with the intent of using them in the future for research purposes, there is no ethical justification for failing to obtain subjects' consent at the time.

3.2 Consent in Observational Studies and Field Work

Many of the behavioral sciences have a long tradition of sending researchers into the field to study the behaviors of persons in their natural environments. This often involves requesting access from those to be observed to otherwise private interactions and territory (e.g., discussions between spouses in the subjects' home). Considerable controversy exists over whether and how truly informed consent should be obtained in such cases. Insofar as researchers may observe and/or report embarrassing of frankly criminal behavior, subjects may be placed at risk of psychological or tangible harm. Thus, the case for obtaining consent in such studies would seem strong, bolstered by the discomfort some investigators report in withholding this information from their subjects.

On the other hand, researchers have an understandable concern that informing subjects clearly of the goals of their studies may lead subjects to alter their behaviors, thus obscuring the very phenomena that are under study. Though many subjects appear remarkably prone to ignore such information, especially once they become accustomed to the observer's presence, this concern needs to be taken seriously. It should usually be possible to accommodate both parties' interests by disclosing the intent of the study in general terms, along with the likely risks to which subjects' may be exposed. If some modification of subjects' behavior occurs, this is an unavoidable price to pay for acquiring a meaningful consent.

3.3 Consent in Community Intervention Projects

Social science research increasingly involves large-scale interventions directed at particular communities or segments of a community. These may include interventions designed to avoid undesirable outcomes (e.g., juvenile delinquency, teenage pregnancy, illegal drug use) or to promote what are generally seen as salutary ends (e.g., moving persons off welfare and

into paying jobs). When such interventions occur at the level of the community as a whole—such as broad-based educational programs concerning drug abuse—it is impossible to obtain informed consent as it is usually conceptualized. Moreover, when studies are based on changes to entitlement programs that have been enacted through the political process, the consent of individual persons may be seen as irrelevant.

There may, nonetheless, be a role in such instances for a modified form of consent interactions with the relevant community. Although each person's consent cannot be obtained, disclosure of the nature and purpose of the study and its risks and benefits can still take place. Consultation with the community at large and its designated representatives may result in suggestions for modifications in the study culminating in a general consensus concerning its desirability (Trickett 1998).

4. Research on Informed Consent

The last decades of the twentieth century have seen a flowering of empirical studies on informed consent to research that promises to shape for the better a debate that has frequently been waged on theoretical grounds alone. Among the areas of investigation have been: the adequacy of investigators' disclosures and how they can be improved; the prevalence of incapacity among various groups of research subjects; subjects' understanding of the information they have received and their appreciation of its implications for their situation; and subjects' retention of information over time (Sieber 1996).

Naturalistic research on the consent process involves overcoming the reluctance of investigators to grant access to the research setting, and of subjects to consent to yet another research project. Subject recruitment can be both difficult and time-consuming. The willingness of funding agencies to sponsor such research is often related to the degree of public scrutiny of and concern about the agencies' research portfolio. But such research is likely to increase over time and to play a greater role in helping to shape investigators' practices and official policies governing informed consent.

See also: Deception, Psychology of; Deceptive Methods: Ethical Aspects; Ethics and Psychiatry; Psychiatry: Informed Consent and Care

Bibliography

Appelbaum P S, Lidz C W, Meisel A 1987a *Informed Consent: Legal Theory and Clinical Practice*. Oxford University Press, New York

Appelbaum P S, Roth L H 1982 Competency to consent to research. *Archives of General Psychiatry* **39**: 951–8

Appelbaum P S, Roth L H, Lidz C W, Benson P, Winslade W 1987b False hopes and best data: consent to research and the therapeutic misconception. *Hastings Center Report* **17**(2): 20–4

Beecher H K 1966 Ethics and clinical research. *New England Journal of Medicine* **274**: 1354–60

Brody B A 1998 *The Ethics of Biomedical Research: An International Perspective*. Oxford University Press, New York

Code of Federal Regulations (US), Title 45 Part 46. Protection of Human Subjects. Revised June 18, 1991. US Government Printing Office, Washington, DC

Faden R R, Beauchamp T L 1986 *A History and Theory of Informed Consent*. Oxford University Press, New York

Grodin M A 1992 Historical origins of the Nuremberg Code. In: Annas G J, Grodin M A (eds.) *The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation*. Oxford University Press, New York

Jones J 1981 *Bad Blood*. Free Press, New York

Levine R J 1986 *Ethics and Regulation of Clinical Research*, 2nd edn. Urban and Schwarzenberg, Baltimore, MD

National Bioethics Advisory Commission 1998 *Research Involving Persons with Mental Disorders that May Affect Decisionmaking Capacity*. National Bioethics Advisory Commission, Rockville, MD

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1978 *The Belmont Report: Ethical Guidelines for the Protection of Human Subjects of Research*. US Government Printing Office, Washington, DC

Sieber J E 1996 Empirical study of ethical issues in psychological research. In: Stanley B H, Sieber J E, Melton G B (eds.) *Research Ethics: A Psychological Approach*. University of Nebraska Press, Lincoln, NE

The Nuremberg Code 1948 In: Annas G J, Grodin M A (eds.) *The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation*. Oxford University Press, New York

Trickett E J 1998 Toward a framework for defining and resolving ethical issues in the protection of communities involved in primary prevention projects. *Ethics and Behavior* **8**: 321–37

World Medical Association 1964, last revised 1989 *Declaration of Helsinki*. In: Vanderpool H Y (ed.) *The Ethics of Research Involving Human Subjects: Facing the 21st Century*. University Publishing Group, Frederick, MD

P. S. Appelbaum

Residential Concentration/Segregation, Demographic Effects of

Segregation is the differential distribution of social groups. Most commonly, the use of the term refers to residential segregation, the differential distribution of social groups across geography. Residential segregation has served as a key indicator of social differentiation for decades. (See *Residential Segregation: Sociological Aspects* for a general review.) Segregation is usually measured by one or more summary indices, designed to reflect the differential distribution of persons or other social units across neighborhoods, occupations, or other categories (White 1986). Segregation is often seen as a multidimensional phenomenon tapped by a variety of indices (Massey et al. 1996). While the bulk of work has centered on residential segregation

among ethnic groups, virtually any trait could be considered, and research has examined segregation by income, education, occupation, nativity, age, and other demographic traits. Racial segregation across schools and occupations remains an important indicator of social distance, and residential segregation can feed school segregation and limit access to jobs, city services, and other public goods (Farley 1984, 1996).

Beyond the measurement of segregation itself, there has always been interest in the consequences of residential clustering. Theory can be invoked to suggest positive as well as negative consequences (Cutler and Glaeser 1997). This ranges from the ways in which residential concentration might generate employment networks or social support institutions, to the consequences of residential isolation for educational attainment or health outcomes. From the point of view of policy, the determinants and consequences of segregation are of particular interest. There has been heightened attention to the aftermath of living in, or being raised in, a segregated or disadvantaged neighborhood. More recent discussions of social capital are pertinent, as the concept of social capital is extended to include the influence of neighbors and nearby peers and role models on such outcomes.

Empirical work at the beginning of the twenty-first century has begun to evaluate the effects of residential segregation. The underlying concept is that residential clustering carries behavioral consequences beyond those of ethnicity, socioeconomic status, or lifestyle. Such notions have been present in demography for some time, but data and multilevel statistical techniques are now available to directly evaluate hypotheses. Social behaviors clearly vary across urban space, or equivalently other categorizations. Yet it is important to distinguish the influence of composition from that of causation. For instance the receipt of public assistance varies appreciably across urban neighborhoods. The degree to which this behavior can be linked causally to the concentration of social groups in neighborhoods (and the sociological patterns that arise through clustering) needs to be weighed against an explanation based just on composition. The key question turns on whether geographic clustering itself influences human behavior, beyond that attributable to other background traits, such as age and education.

Many demographic traits are segregated in urban neighborhoods, and there is a consistency to the ranking. Although empirical results on these matters derive preponderantly from North America, similar results have been obtained for other high-income societies. Race and ethnic groups remain very highly segregated in most urban areas. Several decades of study, from straightforward segregation measures through factorial ecology and the like, have verified the sorting of urban space along the lines of socioeconomic status. Life cycle stage is also differentiated. Segregation indices for sex and age are modest, although segregation for household size and type is more discernable. Of course, certain special concentrations of population, such as group quarters for military populations, students, and those under correctional supervision, are highly segregated and are often excluded from consideration by analysts.

While one can identify and rank the levels of segregation illustrated by various traits, the consequences of this segregation are more difficult to discern. A common causal model advanced in these cases was that persons, especially children, adapted to their disadvantaged environment, learning criminal behavior, reducing achievement motivation for school, and developing dependence on the dole. Observers cited the higher incidence of single parenthood, labor force participation, high school dropout, and the like in such disadvantaged, inner-city neighborhoods.

A demonstration of the presence or absence of the links remained for subsequent waves of investigations. Statistical estimates of neighborhood effects rely on hierarchical and multilevel models. While strategies differ, the general approach is to develop a multivariate model, in which the behavioral and demographic outcome (employment, childbearing) is taken to be first a function of unchanging (ascriptive) personal traits and a set of family and household background factors. Neighborhood and community characteristics are introduced to the model simultaneously or hierarchically. Thus, neighborhood effects are generally measured net of other characteristics. With the availability of multilevel statistical techniques and associated software, it is possible to derive more sophisticated (and econometrically appropriate) estimates of these neighborhood or community effects. Some of the same conceptual and methodological apparatus can be applied to patterns or demographic unevenness found in other social institutions, such as the workplace or schools.

1. Mechanisms Through Which Segregation May Influence Demographic Outcomes

Beyond the simple mechanism of demographic composition itself, segregation, and differential concentration might influence outcomes in several ways. These include (a) contagion effects; (b) resource allocation; and (c) institutional structures. Contagion models, also closely linked to epidemic effects and diffusion models, have a long history in social science thinking. Simply stated, individuals alter their behavior on the basis of their interpretation of those around them (Jencks and Mayer 1990, Crane 1991). The influence of peers on child behavior is one such mechanism. Such processes have been particularly prominent in thinking about school context (Alwin

1977). Role models provide a related mechanism. Much of the recent discussion of adverse inner city outcomes draws on this line of thinking. The general reasoning (see, e.g., Wilson 1987) is that low income or disadvantaged neighborhoods offer fewer mainstream role models (dual parent families, stable adult workers, school completion), and so those growing up in such areas do not acquire mainstream norms. Massey and Denton argue that the intersection of poverty and segregation creates an environment where such conditions abound (Massey and Denton 1993). Social capital is a contemporary phrase that taps some of the same mechanisms. Resources available to the individual beyond the family or household may influence a wide range of behavioral outcomes. Certainly this is of interest in schooling and neighborhoods, and recently it has become of increasing concern in the field of health.

A second line of thinking, far less readily articulated, involves resource allocation. Residents of affluent neighborhoods (or workers in advantaged workplaces) can provide superior public goods, either through direct community provision or through the political process. More than that, these public goods are likely to confer advantages that go beyond the immediate advantages of household and family. Thus, superior schools produce a benefit for all residents in the district (neighborhood) over and above the advantage of family background. Other externalities of clustering include relative advantage or deprivation in local sanitation (exposure to pathogens and toxins), public safety, and the like. Housing experiments, such as those relocating low-income individuals to middle-income communities, are premised on such mechanisms.

The third line of thinking involves the growth and maintenance of local institutions, whether formal or informal (Duncan 1994). The literature identifies both positive and negative consequences of segregation. In the realm of economic outcomes, the notion of the ethnic enclave suggests that concentration, building on ethnic networks, may provide benefits. Occupational and residential segregation with co-ethnics may generate employment opportunities unavailable in a more integrated economy. A long-standing argument is that racial residential segregation enables integration by social class. In an urban neighborhood, local institutions, such as a church or neighborhood association, might foster social ties across class boundaries, although even historical accounts of ghetto life do not always portray a comfortable coexistence of disparate groups. What is relevant in the present circumstance is that these evolving structures can serve as social actors and then influence behavioral outcomes for other members of the community. Empirical social science evidence on these points is scant, especially evidence that can tease out the influences of institutions, net of individual, family, and other community traits.

2. *Methodological Issues*

Concerns about the effects of segregation raise several methodological issues. A fundamental issue is one of definition. A 'neighborhood' might have several competing geographic boundaries. Tracts, wards, block groups, and other urban statistical subdivisions become *de facto* neighborhoods (White 1987). This has two consequences. First, such units serve as the basic geography for calculations of segregation measures. Second, these small statistical areas are the units that serve as the proxy for social conditions in multilevel models. Another methodological evolution is the shift from ecological inference to multilevel modeling. For decades of social science research, analysts were constrained to a choice of individual level analysis or 'ecological' models, which relate aggregate values of demographic characteristics to one another. While ecological analysis was the only route available for certain questions, and while some might argue for the focus on the neighborhood only, multilevel analysis has grown over the last several years as the favored approach (Diez-Roux et al. 1997, Garner and Raudenbush 1991). In multilevel models, individual behaviors are taken to be the result of both individual characteristics and traits of an aggregate. The level of segregation in the city, or the demographic composition of the neighborhood, could be one of these aggregates.

It remains a challenge to properly specify models, so that individual and contextual effects are fully distinguished statistically. The very process of residential assimilation—the achievement of neighborhood status through residential mobility—also generates correlations between community and individual. Thus, separating endogenous and exogenous elements of a process is no easy task. Even then, one must still be concerned that events are properly allocated to local administrative units.

3. *Empirical Evidence*

Interest in community effects in urban settings was rekindled with the discussion and debate over the urban 'underclass' in the 1980s social science literature. Again much of this debate was focused on US settings, but parallel issues of urban neighborhood deterioration and behavioral consequences can be raised for other high-income nations. Despite the level of concern for the effects of residing in ghettoes and inner-city disadvantaged zones, one major contemporary review argued that national social and economic conditions more powerfully determined the incidence of and concentration of poverty (Lynn and McGeary 1990). After expressing reservations about the quality and consistency of data, a related recent review of empirical studies of the effects of neighborhood SES, racial mix, for both neighborhoods and schools,

concluded that there is no general pattern of effects across a wide range of outcomes (Jencks and Mayer 1990). There is some modest evidence that growing up with affluent neighbors (for those of modest means) and in less segregated environment (for blacks) predicts higher levels of educational attainment (Jencks and Mayer 1990). One recent review of about 24 studies of neighborhood effects did tend to find that community-level (often census tract) measures of neighborhood advantage were associated with superior individual-level behavioral outcomes in several phases of childhood (Gephardt 1997). Not surprisingly, this review was critical of the state of knowledge, for the weakness of the link to the appropriate community-level empirical measures and the absence of other characteristics (school, family traits) that would be likely to exert an influence on individual outcomes. One comprehensive compendium of neighborhood level studies on poverty, a study that coordinated data and research approach, again detected neighborhood effects in several investigations (Brooks-Gunn et al. 1997). Typically family-level measures swamped neighborhood-level measures in predicting child behavioral outcomes. Perhaps more interesting is the finding that neighborhood measures could be discerned for pre-school children, despite the conventional thinking that such youngsters' social world was limited to the home.

Substantial effects of neighborhood factors, particularly neighborhood affluence and poverty, have been linked to later life achievement (IQ, teenage childbearing, and school-leaving) in US nationally representative longitudinal data (Brooks-Gunn et al. 1993). Longitudinal data only sometimes detect effects of neighborhood socioeconomic status and racial composition on later earnings, after controlling for selected background traits (Jencks and Mayer 1990). Conceptual models of the influence of neighborhood effects concentrate on the pathways through which these aggregate effects carry to the individual. At the same time qualitative approaches can suggest coping mechanisms for those in disadvantaged neighborhoods (Jarrett 1997). There is some empirical evidence that local organizational structure may mediate community disadvantage (Elliott et al. 1996). The effect of schools may outweigh neighborhoods on risky health behavior (Ennett et al. 1997).

One reasonably long-standing line of work looks at the influence of segregation itself, often introducing the value of a segregation statistic to a regression model of a behavioral outcome. The behavioral mechanism underlying the inclusion of such a covariant is not always clear, but higher level residential segregation may tap into underlying housing market constraints, a more hostile climate of local race relations, or differential access to public sector resources. Predictive power of such segregation indices is not always strong, but on balance they tend to show that individuals (usually minority group members) living in more segregated urban areas experience detrimental outcomes, such as higher levels of unemployment, or violence, or less mobility out of poverty (Iceland 1997, Peterson and Krivo 1993, South and Crowder 1997, Wilson et al. 1995). Even after including several individual and neighborhood controls in a multilevel model, residential segregation does appear to be associated with lower residential mobility on the part of both blacks and whites, supporting the housing market constriction idea (South and Dean 1993). Neighborhood concentration and metropolitan racial segregation, especially among African Americans does appear to be linked to adverse health outcomes for both adults and infants (Williams and Collins 1999, Hart et al. 1998, LaVeist 1989, LeClere et al. 1997, Polednak 1996). There is no consensus on the scale of neighborhood effects, and the magnitude of such effects would be expected to differ by behavioral outcome. The current state of knowledge is that neighborhood effects on demographic outcomes can be detected—especially when investigators use more robust data and more contemporary methodology—yet these neighborhood effects are modest by comparison to personal fixed traits and household characteristics.

See also: Neighborhood: General; Neighborhood Revitalization and Community Development; Poverty, Sociology of; Residential Segregation: Geographic Aspects; Segregation Indices; Urban Life and Health; Urban Sociology; Urban Studies: Overview.

Bibliography

Alwin D 1977 High school context effects on aspirations. *Sociology of Education* **49**: 294–303

Brewster K L 1994 Neighborhood context and the transition to sexual activity among young black women. *Demography* **31**: 603–14

Brooks-Gunn J, Duncan G J, Aber J L (eds.) 1997 *Neighborhood Poverty*. Russell Sage Foundation, NY, Vols I and II

Brooks-Gunn J, Duncan G J, Klebanov P K, Sealand N 1993 Do neighborhoods influence child and adolescent development? *American Journal of Sociology* **99**: 353–95

Crane J 1991 The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing. *American Journal of Sociology* **96**: 1226–59

Cutler D, Glaeser E L 1997 Are ghettoes good or bad? *Quarterly Journal of Economics* **112**: 827–72

Darden J T, Bagaka's J G, Ji S J 1997 Racial residential segregation and the concentration of low- and high-income households in the 45 largest US metropolitan areas. *Journal of Developing Societies* **13**: 171–94

Diez-Roux A V, Nieto F J, Muntaner C, Tyroler H A, Comstock G W, Shahar E, Cooper L S, Watson R L, Szklo M 1997. Neighborhood environments and coronary heart disease: A multilevel analysis. *American Journal of Epidemiology* **146**: 48–63

Duncan G 1994 Families and neighbors as sources of disadvantage in the schooling decisions of white and black adolescents. *American Journal of Education* **103**: 20–53
Elliott D, Wilson W J, Huizinga D, Sampson R J, Elliott A, Rankin B, 1996 The effects of neighborhood disadvantage on adolescent development. *Journal of Research in Crime and Delinquency* **33**: 389–426
Ennett S, Flewelling R L, Lindrooth R C, Norton E C 1997 School and neighborhood characteristics associated with school rates of alcohol, cigarette, and marijuana use. *Journal of Health and Social Behavior* **38**: 55–71
Farley R 1984 *Blacks and Whites*. Harvard University Press, Cambridge, MA
Farley R 1996 *The New American Reality*. Russell Sage, New York
Garner C L, Raudenbush S W 1991 Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of Education* **64**: 251–62
Gephardt M 1997 Neighborhoods and communities as contexts for development. In: Brooks-Gunn J, Duncan G J, Aber J L (eds.) 1997 *Neighborhood Poverty*. Russell Sage Foundation, NY, Vols I and II, Chap. 1, pp. 1–43
Hart K D, Kunitz S J, Sell R R, Mukamel D B 1998 Metropolitan governance, residential segregation, and mortality among African Americans. *American Journal of Public Health* **88**: 434–8
Iceland J 1997 Urban labor markets and individual transitions out of poverty. *Demography* **34**: 429–41
Jarrett R 1997 African American family and parenting strategies in impoverished neighborhoods. *Qualitative Sociology* **20**: 275–88
Jencks C, Mayer S E 1990 The social consequences of growing up in a poor neighborhood. In: Lynn L E, McGeary M G H (eds.) *Inner City Poverty in the United States*. National Academy Press, Washington, DC
LeClere F, Rogers R, Peters K D 1997 Ethnicity and mortality in the United States: Individual and community correlates. *Social Forces* **76**: 169–98
LaVeist T A 1989 Linking residential segregation to the infant-mortality race disparity in US cities. *Sociology and Social Research* **73**: 90–4
Lynn L E, McGeary M G H (eds.) 1990 *Inner City Poverty in the United States*. National Academy Press, Washington, DC
Massey D S, Denton H 1993 *American Apartheid*. Harvard University Press, Cambridge, MA
Massey D S, White M J, Phua V C 1996 The dimensions of residential segregation revisited. *Sociological Methods and Research* **25**: 172–206
Peterson R D, Krivo L J 1993 Racial segregation and black urban homicide. *Social Forces* **71**: 1001–26
Polednak A P 1996 Trends in US urban black infant mortality, by degree of residential segregation. *American Journal of Public Health* **86**: 723–6
South S, Crowder K 1997 Escaping distressed neighborhoods: Individual, community, and metropolitan influences. *American Journal of Sociology* **102**: 1040–84
South S J, Dean G D 1993 Race and residential mobility: Individual determinants and structural constraints. *Social Forces* **72**: 147–67
White M J 1986 Segregation and diversity measures in population distribution. *Population Index* **52**: 198–221
White M J 1987 *American Neighborhoods and Residential Differentiation*. Russell Sage, NY
Williams D, Collins C 1999 Segregation and mortality. *Sociological Forum* **14**: 495–523
Wilson F, Tienda M, Wu L 1995 Race and unemployment: Labor market experiences of black and white men, 1968–1988. *Work and Occupations* **22**: 245–70
Wilson W J 1987 *The Truly Disadvantaged*. University of Chicago Press, Chicago

M. J. White

Residential Environmental Psychology

Homes and residences are not merely physical and economic structures. They are centers of psychological and cultural life. Home is the most important place in many lives, which affords it great potential for facilitating or threatening human development. People think about their homes, are satisfied or not with them, plan, dream, move, use them in various ways, and often must cope with difficult social and physical conditions in their homes, if they have one.

1. *Definitions*

A house is not necessarily a home. A *residence* (house, apartment, condominium, loft, flat, tipi, or, in general, any dwelling) is a physical structure that varies tremendously in form around the world. *Home* is the rich set of evolving meanings that people often (but not always) attach to their residences (Lawrence 1987). Residences may be bought, sold, or rented; homes cannot.

1.1 *Background*

Vernacular residences in traditional societies gradually evolved to suit local building materials, climates, and social practices. In recent times, awareness of other cultures, technological advances, and prosperity in industrialized countries have permitted more choice in residential form and structure. The psychological aspects of residences became relevant, and with the growth of environmental psychology and sociology in the late 1960s, researchers began to examine the impact of residences on residents.

1.2 *Dimensions of Residences and Homes*

Residences vary in any number of physical aspects such as form, age, style, size, or cost. However, residences may be characterized in more psychological and cultural ways (Altman and Werner 1985). First, residences may be permanent or temporary. In industrialized countries, residents usually have permanent dwellings, although apparently this situation is not entirely satisfactory because many householders also maintain temporary residences such as cabins, cot-

tages, and recreational vehicles. In less-industrialized countries, residents move frequently or seasonally and often construct temporary dwellings in new places.

Second, residences range from differentiated to homogeneous, which refers to the separation, or lack of it, in the functions of rooms. A highly differentiated residence has many rooms, each of which houses a specific activity; in a homogeneous residence nearly any activity may occur in nearly any room.

Third, residences may be communal or noncommunal, in terms of the degree to which nuclear families live together or in different homes. In contrast to the typical single-family dwelling of the Western world, in many Eastern cultures several generations of a family live in one household. In some African and Native American cultures, numerous unrelated families live in one dwelling.

Fourth, identity versus communality refers to the extent to which residences are contextually unusual. Communality (the same word as above, but with a different meaning here) is the extent to which a residence is physically typical of dwellings in that culture. A tipi on a suburban lot in North America would be low on communality, but high on identity; the same tipi at a pow-wow on the open plains of central North America would be high on communality.

Fifth, residences vary in their openness to outsiders. This varies both within and between cultures. In some places, houses typically are surrounded by walls or hedges; in others they are not. Dwellings signal willingness or unwillingness to interact with outsiders.

Home, in contrast to residence, is a set of personal meanings. The word's earliest meaning probably was the village or community of one's origin; not until the sixteenth or seventeenth centuries did it refer to one's dwelling. To be a home, a residence should provide haven, order, identity, connectedness, interpersonal warmth, and physical suitability (Smith 1994). As a haven, a home must surround a person with privacy, security, and protection from the slings and arrows of life outside it.

A home provides order by helping individuals to know their place in the world. As a center from which people venture and return, home orders one's existence in the world, both spatially and temporally. Home is central to a sense of identity and connectedness because it offers a sense of kinship, belonging, and status. Through self-expression and personalization, the home comes to resemble or represent the self. When people have the means and the time to personalize their residence, home becomes a symbol of self (Cooper 1976). To the extent a residence does not provide these felt amenities, it is not a home.

1.3 Houselessness

The homeless, those who live in the streets, properly should be called houseless. Many houseless people have a home, a residence with personal meaning that—for one reason or another—they are unable to live in. Home is the most important refuge from the stresses of work, school, and public life. Unfortunately, many people live in residences that do not provide protection and rejuvenation, and therefore they are homeless. Probably one billion people have inadequate housing, and about 100 million are houseless. Institutionalized individuals often are in this situation. Houselessness is more than lack of a residence; many other social negative social consequences follow. Individuals who are both houseless *and* homeless do exist, and are extremely unfortunate.

2. Residential Preferences and Satisfaction

Residential preference and satisfaction are related but distinct concepts. Their measurement is not straightforward. Some residences would dissatisfy anyone, and there are some individuals who will not be satisfied by any residence. Between these extremes, however, lie most homes and most residents. Different homes satisfy different residents. Preference involves comparison, whereas satisfaction may be mainly based on residents' evaluations of their own dwellings.

2.1 Measurement Issues

Residential satisfaction is multidimensional. David Canter (1983) has described two cognitive processes related to residential satisfaction, *purposiveness* and *comparison*. Purposive evaluations have three facets: level, referent, and focus. Level, or specificity, refers to evaluating a single part of the residence (such as the kitchen), or a larger portion of it (such as the bedrooms as a whole, or the entire residence). Referents refer to different qualities of the residence (its beauty, or spaciousness); each suggests a certain purpose which is served well or poorly. The lighting referent, for example, is important when the resident considers tasks such as reading or sewing that require good lighting. Focus depends on the referent, and refers to how broadly the resident is asked to evaluate a referent. For example with lighting, focus might refer to the ability of a particular lamp to light a study desk, or about the ability of all the residence's lamps to light the home as a whole.

Purposiveness arises out of the resident's relationship to the residence, such as long-term vs. short-term residence, ownership, renting, or landlordship; these different purposes will make important differences in the resident's evaluation. In general, a resident's purpose in doing an evaluation is itself a crucial factor in satisfaction.

The second factor is comparison. Whether observers are explicitly asked to compare a specific residence against a series of others, or merely asked for a rating

of a particular residence without explicit comparison with others, they cannot help at least implicitly comparing the residence in question with others in their experience. Thus, different observers will be comparing the residence in question against different houses in their experience.

Residents who are unaware that their friends live in mansions might be more satisfied with their huts. Residents compare residences, and this can be an important source of satisfaction or dissatisfaction. A study of Army personnel who lived in mobile homes and apartments found that comparison with other housing accounted for much of the variance in their housing satisfaction.

Perceived physical qualities of a residence (such as its appearance) are important predictors of satisfaction, and so are preferred physical qualities. However, the difference between perceived and preferred qualities is a separate but important predictor of housing satisfaction.

2.2 Personal, Social, and Physical Influences

Residential satisfaction is influenced by many personal characteristics, including age, stage in the life cycle, socioeconomic status, personality, values, and hopes for the future. Social factors such as age and community norms, others' opinions, and neighbors also influence satisfaction. Gender and sex roles are important. In a study of Canadian households that had recently moved, for example, husbands and wives were asked to list the things that made them happy about their new homes and apartments (Michelson 1977). Among couples who moved to a downtown apartment, more than 40 percent of the wives mentioned that they were satisfied with the quality of the home, but only 5 percent of the husbands mentioned this. On the other hand, among couples that moved to suburban apartments, more than 50 percent of the husbands mentioned quality of the home as a source of their satisfaction but only 5 percent of the wives did.

As men enter the kitchen more, similarities and differences in men's and women's preferences for kitchen designs become more important. Both men and women prefer kitchens that open to dining and family room areas over kitchens that are separated from those areas. However, women who work outside the home prefer open kitchens more than women who do not work outside the home and, despite their agreement on the open-closed aspect of the kitchen, men and women differ in many other kitchen spatial preferences.

Residential satisfaction is also related to socioeconomic status. In general, of course, wealthier individuals are better able to supply themselves with homes that meet their standards, and thus are more satisfied. However, wealthier and poorer residents may be attuned to different features of their homes: One study found that wealthier individuals were more sensitive to the aesthetic qualities of their home, whereas poorer individuals were more sensitive to safety, health, and family needs.

Naturally, physical and architectural factors such as house form, architectural style, floor plan, and colors affect residential satisfaction, but these should not be considered the sole, or even the main, source of housing satisfaction. Most individuals prefer higher (than the usual eight-foot) ceilings, flat or sloping ceilings (4-in-12 slope but neither more than that nor smaller slopes), and walls that meet at 90 degrees or more (not walls that meet at smaller angles). In general, highly detailed designs are favored over plain and simple designs. In places where fear for personal security is high, enclosed and guarded enclaves are highly valued, as shown in a recent Brazilian study (Carvalho et al. 1997).

Color has been the subject of much research, but most of it has been based on preferences for small patches of color that resemble paint chips, in no particular context (that is, subjects are not usually asked if they would paint their kitchen or bedroom that color). These findings may not apply to choices in real-life situations. A study in Japan tried to overcome these limitations by showing slides of living rooms that were painted different colors (Kunishima and Yanase 1985). Preference is not merely a matter of hue (red vs. blue vs. green); it also depends on the hue's saturation (or 'denseness') and brightness (from light to dark). The study concluded that each of these three dimensions of color was related to a different aspect of preference. Hue was not strongly related to preference; instead it was primarily related to perceived warmth (reds are warmer, blues and greens cooler). Saturation (regardless of hue) was most closely related to what we usually call preference: More saturated hues were evaluated as more elegant, more comfortable, and better. Brightness or lightness (regardless of hue) was related to how 'active' the room seemed: Brighter colors were rated as fresher, lighter, and more cheerful than darker hues.

There remains a general preference for the single-family dwelling that has persisted in North America for many years. As for style, some trends are evident: In one study, farmhouse and Tudor style houses were preferred to saltbox and Mediterranean styles (Nasar 1989). However, preferences for style clearly change in part with changes in fashions.

2.3 Cultural Variations

Residences clearly vary in form around the world. These variations are not merely decorative, they are psychologically meaningful. For example, an examination of house form across 73 cultures revealed that the extent of partitioning indoor space (differentiation) is related to the degree of sociopolitical complexity in

the culture, that is, how much segmentation, specialization, and hierarchical ordering is typical of the culture. Thus, a worldwide trend is for our dwellings to mirror the complexity of the society in which it is embedded.

3. Behavior in Residences

The environmental psychology of dwelling interiors is not very developed because the study of behavior in and round homes understandably meets with a resistance on privacy grounds.

3.1 Vernacular Residential Design

Residents arrange and decorate their interiors according to certain patterns that reflect such dimensions as simple–complex, conventional–unusual, and rich–plain decor, and messy–tidy upkeep. These patterns are related to culture, social class, and marital or living arrangement differences. For example, French and Italian living room arrangements were examined (Bonnes et al. 1987). French living rooms could be described using three dimensions–decoration, functional organization, and structuring of space. For example, at one end of the decoration dimension were living rooms that were low on ambiguity and complexity; these living rooms tended to have velvet antique couches with armrests, copies of period wallpaper, and books for show (not use) in a wall-unit that also included cupboards and a bar. Italian living rooms could be described using the same three dimensions, but they tended to be combined differently.

3.2 Use of Interior Spaces

Men, women, and children use the interior dwelling spaces unequally; sex role changes have not changed the traditional use of rooms much; women still spend more time in the kitchen (Ahrentzen et al. 1989) and more time with children, and men spend more time in leisure activities. Children (until the teen years) have restricted territories: Parents generally can enter their space, but children generally cannot enter their parents' space.

In a rare study that actually monitored behavior inside homes (as opposed to asking residents to report their activities), cameras were placed (with the residents' permission) in New York homes (Ashcraft and Scheflen 1976). The study found that although definite territories exist within the home, the degree to which the boundaries of these territories were respected varied considerably. For example, parents may feel they have the right to enter a teen's bedroom, but do not think the teen should enter their bedroom. But both territories are sometimes intruded upon, which creates conflict.

When space is limited, residents use time scheduling, territoriality, and social agreements to manage household space. Half our leisure time is spent in the home, but it could be better spent. When teens spend more time sharing leisure with their parents, they report greater well-being. More TV time is associated with less time spent on intellectual and cultural pursuits.

4. Coping with Residential Shortcomings

Much housing is substandard, and there are consequences. Well-executed recent research conducted in the USA suggests that poor-quality housing adversely affects the mental health of both adults (Evans et al. 2000) and children (Evans et al. in press).

4.1 High Density

High density in the residence has been shown to have many adverse effects. High-density homes as a whole have more psychological distress and psychiatric illness and have been associated with increased substance abuse, juvenile delinquency, fertility and mortality rates, and may lead residents to believe they have less social support. Crowding at home has more serious consequences than crowding in public environments. Some of these are mitigated by social support, if the resident has any. Density appears to have negative effects apart from poverty. Very low densities (living alone) also cause problems for many people.

4.2 Housing Forms

Not every housing form is good for every resident. High rises probably are not good for families or children, and one study suggests that young children are better off if they attend a good child care center than if they stay at home. Apartment living has been associated with numerous childhood afflictions, including retarded movement skills, respiratory diseases, aggression, insomnia, nervous disorders, reduced social skills, and disrupted play. If all this is true, we may wonder how any child can survive in an apartment. However, many of these studies lack experimental rigor; their results must be questioned.

5. Residential Mobility

All of us must unwillingly adapt to new dwellings sometimes; at other times, we choose to relocate. In any given five-year period, about half of all North Americans and Australians move, as do over one-third of Britons and Japanese (Stokols et al. 1983).

People move for many reasons, including both negative factors such as stress from and dissatisfaction with the current residence, but also for positive reasons such as improved finances. Life cycle changes that add or subtract family members living at home can also spur mobility. Mobility is not necessarily pathogenic, then, but it can be, depending on one's residential history, present circumstances, and hopes for the future. For example, among those who move less often, those with less choice about moving report more illness-related symptoms than those who have greater choice about moving.

Mobility-related stress is moderated by personal characteristics; people who are highly mobile and those who were generally less inclined to explore the environment report more illness than those who are more exploratory by nature. Conversely, more illness occurs among residents who move less often but are unhappy with their present residence and feel they have little choice of residence. Thus, a lack of mobility also can be damaging—when one's options for moving are restricted.

6. *Conclusions*

Home usually is the most important physical setting. The physical building must be distinguished from the meaning structure (home). Residential satisfaction is complex and multiply determined, but life cycle stage, socioeconomic status, personality, values, and hopes for the future are important factors, as are physical features such as house form, architectural style, floorplan and colors, outdoor areas, and cultural background. People arrange residential interiors in nonrandom patterns that reflect culture and lifestyle. High indoor density is difficult for virtually anyone, but can be made less intolerable with social support. Moving is stressful or not depending on whether a person has some choice in doing so, likes to explore new settings, and likes the present residence.

See also: Environmental Psychology: Overview

Bibliography

Ahrentzen S, Levine D W, Michelson W 1989 Space, time, and activity in the home: A gender analysis. *Journal of Environmental Psychology* **9**: 89–101

Altman I, Werner C M (eds.) 1985 *Home Environments*. Plenum Press, New York

Ashcraft N, Scheflen A 1976 *People Space*. Anchor, Garden City, New York

Bonnes M, Giuliani M V, Amoni, F, Bernard Y 1987 Cross-cultural rules for the optimization of the living room. *Environment and Behavior* **19**: 204–27

Canter D 1983 The purposive evaluation of places: A facet approach. *Environment and Behavior* **15**: 659–98

Carvalho M, George R V, Anthony K H 1997 Residential satisfaction in condominiums exclusivos (gate-guarded neighborhoods) in Brazil. *Environment and Behavior* **29**: 734–68

Cooper C 1976 The house as a symbol of the self. In: Proshansky H, Ittelson W H, Rivlin L G (eds.) *Environmental Psychology*, 2nd edn. Holt, Rinehart and Winston, New York, pp. 435–48

Evans G W, Saltzman H, Cooperman J in press Housing quality and children's socioemotional health. *Environment and Behavior*

Evans G W, Wells N M, Chan H-Y E, Saltzman H 2000 Housing quality and mental health. *Journal of Consulting and Clinical Psychology* **68**: 526–30

Kunishima M, Yanase T 1985 Visual effects of wall colors in living rooms. *Ergonomics* **28**: 869–82

Lawrence R J 1987 *Housing, Dwellings, and Homes: Design Theory, Research, and Practice*. Wiley, New York

Michelson W 1977 *Environmental Choice, Human Behavior and Residential Satisfaction*. Oxford University Press, New York

Nasar J 1989 Symbolic meanings of house styles. *Environment and Behavior* **21**: 235–57

Smith S G 1994 The essential qualities of a home. *Journal of Environmental Psychology* **14**: 31–46

Stokols D, Shumaker S A, Martinez J 1983 Residential mobility and personal well-being. *Journal of Environmental Psychology* **3**: 5–19

R. Gifford

Residential Segregation: Geographic Aspects

Residential segregation refers to the practices and ideologies by which socio-economic characteristics are expressed in, and shaped through, the organization of residential space.

1. *The Division of Residential Space*

Residential space is a geographically uneven resource, and the land and housing which constitute it vary in quality, character, condition, and value. Land values vary according to demand and have traditionally been highest in city centres. Housing characteristics vary according to the circumstances of housing production and use, including the macroeconomic climate, the legislative and regulatory framework, the fortunes of the construction industry, the availability of loans and subsidies, changing architectural fashions, and so on.

A primary determinant of access to housing (and neighborhood) is cost, but the markets and institutions that make up the housing system also allow a range of social characteristics to mediate the process of inclusion within, or exclusion from, different parts of the residential environment. It is this interaction of the spatial structure of the housing stock with social structures, property values and with the operation of

commercial and public institutions that gives rise to the geography of residential segregation. These processes are, furthermore, embedded in, and legitimized through, the wider set of entitlements, obligations, and ideologies built into the constitutions of states and societies.

Housing and the residential space it comprises is both a commodity and a welfare resource. Housing is bought and sold in the market place, but housing policy can also be a means by which states intervene to meet social needs and alleviate social inequality. Both market forces and governmental priorities therefore play a part in structuring residential space; and such space is fractured in many ways. This fracturing occurs because of differences in 'need' and in 'ability to pay,' and it occurs too because the processes governing access to housing tend to discriminate on the basis of social characteristics. Direct and indirect discrimination has been documented in all sectors of the housing system. In the housing market, access to housing opportunities generally, and to housing finance in particular, may be dictated as much by who people are (or are presumed to be), by what they do, and by where they currently live, as by their actual ability to pay. In the public or social sector, the principle of allocating homes according to 'need' is always hard to operationalize and in practice may be compromised by, for example, subjective discretionary judgments about what types of people 'deserve' what kind of housing, and where.

As a consequence of all this, while class, income, wealth, and other economic indicators provide one set of fault lines in the residential landscape, ideas about 'race,' ethnicity, religion, and language provide another. Health inequalities are also expressed and reproduced in the residential environment; and although most research on segregation refers to the experience of 'households' (implicitly lumping men, women, and children together, and often referring mainly to the experience of men), the division of residential space is also fundamentally gendered and sexualized.

2. *Measurement: The 'Index Wars'*

A strong empirical tradition underpins the literature on residential segregation, and this has become the province of geographers (see *Social Geography*) and of those whom Peach (1975) terms 'spatial sociologists.' The idea implicit in all this work is that there is some kind of link between society and space; that social processes literally 'take place'; and that where people live is both an expression of who they are and a factor affecting who they can become.

Empirical studies of residential segregation are dogged by the problems of measurement and interpretation. The problem of measurement is apparent from the large literature comprising the so-called index wars. One debate concerns the appropriate index with which to measure segregation. This has revolved around a group of indices of segregation and dissimilarity, which were favored for their simplicity and apparent (though questionable) scale independence. The index of dissimilarity (ID), for example, measures the difference in the distribution of any two population groups over a given number of areal units (e.g., census districts). ID is calculated very simply as half the sum of the difference in the proportion of each group in each areal unit. The index varies from 0 (complete similarity) to 100 (complete segregation) and the value of the index refers to the proportion of either group that would have to move from one areal unit to another in order for the two distributions to be the same.

ID is a special form of the more general index of segregation (IS) which measures the residential dissimilarity of one group from the rest of the population. Several other indices can be derived from this to capture the spatial segregation and isolation of one or more population groups from one another. Perhaps the best known of these is the asymmetric index of segregation, P*, which describes the spatial isolation of a social group and provides a measure of potential interaction between that group and another (Lieberson and Carter 1982).

A second point of dispute in the measurement of residential segregation concerns whether the degree of segregation should be gauged against a norm of evenness or randomness. If populations were randomly distributed across space, some clustering would be expected as part of this random pattern. The index of dissimilarity refers to the proportion of a population that would need to move to achieve evenness, and it may therefore exaggerate the amount of segregation that is enforced (or chosen). There is still no consensus on the best way forward here. Nevertheless, all these debates are dwarfed by the problems of how to conceptualize the groups in the study, of whether to standardize (e.g., for income when measuring 'racial' segregation), and, crucially, of how to interpret the results of applying the measures.

3. *Interpretation: The 'Choice-constraint' Debate*

A major debate over the interpretation of measures of residential segregation concerns the question of causality. It hinges around the extent to which people choose to live separately from one another, and the extent to which segregation is forced upon them. This has become known as the 'choice-constraint' debate. In the UK, for example, all the major cities display enduring patterns of residential segregation between the white, African-Caribbean and (South) Asian populations. However, within the Asian populations,

Sikhs, Hindus, and Muslims tend to live in different neighborhoods to one another, while West Indians in London are segregated according to the island of origin of those who are migrants. Given these findings, it might plausibly be argued that the residential separation of Hindus and Muslims, or of Jamaicans and Trinidadians, is a matter of choice. Indeed in the USA, Clark (1992) draws on Schelling's (1971) ideas about the aggregate effects of small variations in residential preferences to argue that, in the wake of declining institutional discrimination, the accumulation of own-group residential choices is a key factor sustaining current pattterns of segregation in US cities (see *Spatial Choice Models*).

While most people do have some choice over the precise space they occupy in the residential environment, the fact that, for the poorest people, these choices are highly limited, leads many authors to focus more on sources of 'constraint' than on elements of 'choice.' Smith (1987, 1989), for example, has argued that while it might well be the case that intragroup segregation is the result of household preferences, it is hard to sustain the argument that a very substantial proportion of such a diverse group as, say, South Asians in Britain, would choose to live in neighborhoods that are on average cheaper, poorer, less well serviced, and more deprived than any others. The demonstrable links between segregation and inequality, the marked extent to which poverty overlaps with (or is constructed through) these spatial processes of marginalization and exclusion, suggests that households segregated into the least desirable residential locations have very little room for maneuver.

Despite a number of conceptual and methodological problems, the empirical tradition has established beyond question that residential segregation is a key feature of the residential landscape, and that, preferences notwithstanding, patterns of residential segregation tend more often than not to be linked systematically with patterns of social inequality (see *Spatial Equity*). Some have suggested that this means that residential segregation has the capacity to act as a catalyst in the process of class formation or as a springboard for collective action more generally. However, in one of the most influential theoretical accounts of the relationships between class structures and residential differentiation, David Harvey (1985) argues that residential segregation is part of the reproduction of social relations within capitalist societies. From this perspective the fracturing of populations into residential communities mainly serves to fragment the class consciousness which might otherwise be used to improve the quality of local life.

Some of these fragmenting, isolating effects of segregation are uncovered by Pratt and Hanson (1994) in a study of women workers in Worcester, Massachusetts. They show how residential neighborhoods are differentiated by the distinctive mix of 'race,' class, gender, and work traditions within them. This mix influences the locational decisions of employers, who work with stereotypes about the differences between men and women's work, and about the suitability of different ethnic groups for particular types of task. Employers' locational strategies thus reinforce a spatial division of labor which already emphasizes the gender divide. But because this division of labor also operates along class and ethnic lines, the potential for solidarity among women is undermined, despite their common interests (see *Gender and Feminist Studies in Geography*). Residential segregation is thus experienced as isolating and exclusionary rather than as a potential resource.

4. *Segregation and Inequality*

The literature on residential segregation explores the many social dimensions embedded in the division of residential space. While there is some interest in the location of elites, and in the continuing residential segregation of predominantly 'white' ethnic groups (for example Poles, Jews, and Italians in US cities) despite their economic integration, perhaps the bulk of writing focuses on disadvantaged and marginalized minorities. There are therefore studies of residential segregation among people with mental health problems, the old and impoverished, disabled people, single mothers, and so on. However, the main research effort has been concerned with the impact of segregation on the lives of marginalized racialized groups in Western societies (see *Ethnic Conflict, Geography of*). This perhaps reflects the magnitude of the challenge of 'otherness' that most developed countries face in a postcolonial world, where 'people of color in the historically white counties of the West are treated as second-class citizens' (Bonilla-Silva 2000, p. 194). Failure to acknowledge this—that is, failure to recognize the intensely racialized nature of much residential segregation in the Western world—has been criticized by Wacquant (1997) who (writing with specific reference to the USA) is concerned that by reducing the problem of racial segregation to one of poverty and disadvantage, analysts underestimate the power of history and ideology.

A growing literature on the racial dimension of residential segregation shows that segregation is not just an economic process and a material fact of life, but that it also has powerful ideological underpinnings. Historians George Fredrickson (1981) and John Cell (1982), for example, show that segregation is part of the changing character of racism. The idea of 'race' may have arisen from the notion that particular places (climates and environments) produced different kinds of people ('races'), but the tenacity of this idea is now secured by an unequal distribution of power which allows particular kinds of people to be restricted to particular types of place. Thus racisms become institutionally normalized in and through the spatial

configuration of the residential environment (cf. Goldberg 1993)

So, whereas many writers once assumed that racial residential segregation is a declining legacy of a racist, colonial past, Cell and Fredrickson show how segregationism emerged during the early twentieth century as a new and powerful way to manage power relations. In the US South and South Africa, this was initially legally enforced through the Jim Crow laws and apartheid, respectively. But what Cell's analysis vividly shows is that one of the key characteristics of segregationism is its tenacity, flexibility, and ability to survive as a practice even with the declining respectability of racism and even with the end of *de jure* racial separatism. Indeed, Cell goes so far as to define segregation as one of the most successful political ideologies of the twentieth century. 'In a segregated society,' he argues, 'we expect job restrictions and unequal wages; inferior housing, education, transportation, health and social services; prejudiced courts and police forces.' And to the extent that all these forms of exclusion are enacted through the practice of residential differentiation, segregation represents what he calls 'the highest stage of white supremacy.'

4.1 Residential Segregation and Racial Inequality

In the USA, the tenacity of these discriminatory and exclusionary forms of segregation has been closely documented by Massey and Denton (1993). Their analysis of the 1990 US census identified 20 metropolitan areas in which racial segregation was so complete and so expressive of the social and financial exclusion of black Americans that they merit the label 'hypersegregated.' Massey and Denton argue for the case of the USA that 'residential segregation lies beyond the ability of any individual to change: it constrains black life chances irrespective of personal traits, individual motivations or private achievements' (1993, pp. 2–3). A startling demonstration of this is given by Polednak (1997) in his study of the relationship between segregation and health which shows, among other things, that mortality rates and infant mortality rates in segregated communities in the USA can exceed those in parts of the less developed world.

In a series of related studies, Massey and colleagues suggest that not only is segregation problematic because it creates spaces of deprivation, exclusion, and neglect, but it is problematic too because the fact of residential segregation acts to concentrate poverty during times of macroeconomic change. So, while all poor people have been affected by those structural changes in the economy which increase income inequality, the most highly segregated communities—in the USA, African Americans—fare worst because of the way that residential segregation interacts with economic restructuring to concentrate urban poverty (Massey and Eggers 1990, Massey et al. 1991).

This work has been contested on both theoretical and methodological grounds by authors who place more emphasis on 'class' than 'race' in explaining recent increases in the spatial concentration of poverty among black Americans. However, Massey and Fisher (2000) have responded with a persuasive empirically-based argument that the same socio-economic changes in a given society have radically different consequences for the concentration of poverty as segregation rises. Residential segregation, they argue, 'is the key factor differentiating the experience of racial and ethnic groups in the United States, dividing them into two basic categories: those whose poor experience a high degree of class isolation in neighborhoods characterized by unusally high concentrations of poverty, and those whose poor do not' (p. 689). Because black Americans are by far the most highly segregated of the studied populations, 'the population of poor experiencing high concentrations of poverty is overwhelmingly black' (p. 689).

This kind of finding is by no means only a feature of societies with a historical legacy of *de jure* segregationism. In Britain, with no history of domestic slavery or legally enforced racial separatism, the residential segregation of racialized minorities has been a persistent theme for more than half a century (Smith 1989). There has, it is true, been a fierce empirical debate over whether British inner cities are as segregated as many US ghettos. This has centered on whether, for example, an ID of 60 calculated using British enumeration districts is somehow 'better' or less problematic than an ID of 80 for US census districts. However, Smith (1987) argues that while segregation persists at any of these levels, it is the form not the intensity of the phenomenon that is at issue. Whether or not ghettoes are the same thing in all places, that higher levels of segregation are associated with higher rates of poverty, unemployment, poor health, and so on, is indisputable. Indeed, Smith (1989) goes on to argue that segregationism is as effective in Britain as in the USA as an interlocking system of economic institutions, social practices, and political strategies designed by one group to keep another in its place. Residential segregation, then, is a material fact of the lives of marginalized, racialized groups in Britain, and it goes with poor housing, limited employment prospects, limited service provision, poorer health prospects, and a range of other inbuilt disadvantages,

A key finding in several studies of racial segregation is that, once established, residential segregation is not just an expression of the marginality of racially 'othered' groups, but that it is also a determinant of their future life chances and opportunities. So the iniquitous form that segregation takes not only expresses marginality, but also reinforces it.

This reinforcement takes place in a number of ways. In the USA, for example, it has been suggested that there is a spatial mismatch between those who are

unemployed and the jobs they need to survive (see *Spatial Mismatch*). The argument here is that the spatial restructuring of the economy has taken jobs and incomes away from the areas in which black Americans live (see *Economic Restructuring: Geographic Aspects*). Residential segregation can, in short, act as a constraint on labor market participation. In the UK, a similar process has been observed—there too, entrapment in spaces of deprivation can prevent already disadvantaged groups from seizing opportunities opened up by the restructuring of the economy. In the British example, moreover, this process is overlaid by what might be thought of as a 'welfare mismatch' as those most dependent on welfare find themselves excluded spatially and economically from the range of commodities, insurances, and financial services that have expanded with the 'rolling back' of the welfare state (Smith 1989) (see *Welfare State, Geography of*).

So residential segregation both expresses social inequalities, and materially affects the participation of spatially excluded groups in social, political, and economic life. But because of its ideological dimension (noted above) segregation is not just a material fact, it is also a political idea—and one which often bears little resemblance to the lived experience of those it refers to. Smith (1988, 1993), for example, shows how, in Britain, 'racial segregation' as a political construction has helped sustain common-sense assumptions about the natural origins of racial differentiation without tackling the problems of racism and discrimination which underpin the inequalities embedded in the structure of residential space. The importance of ideas and images (ideologies) about residential space as a factor structuring people's life chances and opportunities is further illustrated by Anderson (1991) in her work on Canadian constructions of Chinatown. She shows how that segregated space known as 'Chinatown' grew out of, and came to structure, a politically divisive system of racial discourse that justified domination over people of Chinese origin in late nineteenth-century Vancouver. In a similar vein, Wilson (1996) illuminates the power of metaphor to drive the process of uneven development between segregated segments of urban Indianapolis in the USA. His aim is to illustrate what potent political instruments metaphorical symbols can be. In the example of Indianapolis, USA, 'they injected a constellation of representations into everyday thought that rendered a form of growth as value-free and neutral that was, in fact, directly adverse to the interests of the black inner city poor' (p. 96).

Because of its ideological underpinnings and material effects, residential segregation is something that those who are marginalized and excluded by it can have little control over. However, iniquitous forms and practices of segregation can be, and are being, resisted in a number of ways. There is a policy dimension to this, of course, the relevance of which is firmly established by Bashi and Hughes (1997). In a thoughtful account of the likely implications of globalization for the character and intensity of residential segregation by race they argue that the main problems and their likely solutions are located locally (nationally) rather than globally. 'On the topic of segregation,' they note, 'the real problem is less one of sustained explanation than one of sustained action' (p. 120). This adds new significance to the work of those who argue that the best weapon with which to tackle segregation is the law (Goering 1986, McGrew 1997).

Other ways of illuminating and challenging the inequalities embedded in residential segregation range from urban unrest to artistic expression. The first of these is extensively documented. The second is less fully recognized, though a glance at the work of artists like photographer Ingrid Pollard, painter Jean-Michel Basquiat, or filmmaker Spike Lee soon illustrates its potency. Ingrid Pollard's exhibition 'Pastoral Interludes,' for example, locates a series of 'black' figures in the 'white' English countryside. This contrasts with the 'usual' positioning of black Britons within the segregated inner city, and challenges the taken-for-granted assumption that it is somehow normal for black people to be absent from rural areas. The other side of the coin—a stark vision of (and dissatisfaction with) the experiences of confinement within residentially segregated urban spaces—finds expression in the stick-like black figures which characterize so many of Basquiat's paintings and in the rise of rap and hip-hop which was used so effectively in Spike Lee's *Do the Right Thing* to speak about life in the segregated space of New York's Brooklyn.

5. *Overview*

To summarize, residential segregation is both an expression and a mediation of the ideologies and practices which structure social life. It is a product of the uneven development of residential space and of the selective operation of the markets and institutions that govern access to housing. Residential space is a gateway to services, educational, and employment opportunities, health, and social services. Where people live is therefore a reflection of who they are, but it is also an influence on who they can become. It has therefore been argued that tackling the inequalities embedded in residential segregation is a key priority for public policy and an urgent focus for social concern.

See also: Ethnic Conflict, Geography of; Gender and Feminist Studies in Geography; Population Ecology; Residential Concentration/Segregation, Demographic Effects of; Residential Segregation: Sociological Aspects; Segregation Indices; Social

Geography; Spatial Choice Models; Spatial Equity; Spatial Mismatch; Welfare State, Geography of

Bibliography

Anderson K 1991 *Vancouver's Chinatown: Racial Discourse in Canada 1865–1980*. McGill-Queen's University Press, Montreal

Bashi V I, Hughes M A 1997 Globalization and residential segregation by race. *Annals of the American Academy of Political and Social Science* **551**: 105–20

Bonilla-Silva E 2000 'This is a white country': The racial ideology of the western nations of the world system. *Sociological Inquiry* **70**: 188–214

Cell J W 1982 *The Highest Stage of White Supremacy: The Origins of Segregation in South Africa and the American South*. Cambridge University Press, Cambridge, UK

Clark W A V 1992 Residential preferences and residential choices in a multiethnic context. *Demography* **29**: 451–66

Fredrickson G 1981 *White Supremacy: A Comparative Study in American and South African History*. Oxford University Press, New York

Goering J M 1986 *Housing Desegregation and Federal Policy*. University of North Carolina Press, Chapel Hill, NC

Goldberg D T 1993 *Racist Culture: Philosophy and the Politics of Meaning*. Blackwell, Oxford, UK

Harvey D 1985 *The Urbanization of Capital: Studies in the History and Theory of Capitalist Urbanization*. Johns Hopkins University Press, Baltimore

Lieberson S, Carter D 1982 A model for inferring the voluntary and involuntary causes of residential segregation. *Demography* **19**: 511–26

Massey D S, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA

Massey D S, Eggers M L 1990 The ecology of inequality: Minorities and the concentration of poverty. *American Journal of Sociology* **95**: 1153–89

Massey D S, Eggers M L, Gross A B 1991 Segregation, the concentration of poverty, and the life chances of individuals. *Social Science Research* **20**: 397–420

Massey D S, Fischer M J 2000 How segregation concentrates poverty, 1970–1980. *Ethnic and Racial Studies* **23**: 670–91

McGrew T 1997 The history of residential segregation in the United States and Title VIII. *The Black Scholar* **27**: 22–30

Peach C (ed.) 1975 *Urban Social Segregation*. Longman, London

Polednak A 1997 *Segregation, Poverty and Mortality in Urban African Americans*. Oxford University Press, New York and London

Pratt G, Hanson S 1994 Geography and the construction of difference. *Gender, Place and Culture* **1**: 5–29

Schelling T 1971 Dynamic models of segregation. *Journal of Mathematical Sociology* **1**: 14–86

Smith S J 1987 Residential segregation: A geography of English racism?. In: Jackson P (ed.) *Race and Racism: Essays in Social Geography*. George Allen and Unwin, London, pp. 25–49

Smith S J 1988 Political interpretations of racial segregation in Britain. *Environment and Planning D: Society and Space* **6**: 423–44

Smith S J 1989 *The Politics of 'Race' and Residence: Citizenship, Segregation and White Supremacy in Britain*. Polity Press, Cambridge

Smith S J 1993 Residential segregation and the politics of racialisation. In: Cross M, Keith M (eds.) *Racism, the City and the State*. Unwin Hyman, London, pp. 128–43

Wacquant L J D 1997 Three pernicious premises in the study of the American ghetto. *International Journal of Urban and Regional Research* **21**: 341–53

Wilson D 1996 Metaphors, growth coalition discourses and black poverty neighborhoods in a US city. *Antipode* **28**: 72–96

S. J. Smith

Residential Segregation: Sociological Aspects

Segregation refers to the differential location of social groups across categories of a social structure. It may arise from voluntary forces, whereby members of one group choose positions that are different from those of other groups; it may reflect involuntary processes, whereby systematic barriers prevent certain kinds of people from occupying particular positions within the structure, or it may reflect a combination of voluntary and involuntary forces.

By far the most studied kind of segregation involves the differential location of social groups across neighborhoods of a city, a topic generally known as *residential segregation*. This field has attracted strong interest over the years because social scientists have recognized the close connection between where people live and the socioeconomic outcomes they experience. This insight originated with the Chicago School of Sociology in the early twentieth century (Park and Burgess 1925). According to Robert Park (1926, p. 8), 'it is because social relations are so frequently and so inevitably correlated with spatial relations; because physical distances so frequently are … indexes of social distances, that statistics have any significance whatever for sociology.' In other words, because social and economic resources are distributed unevenly in space, where one lives plays a significant role in determining one's prospects for attaining education, health, employment, income, and prestige (Massey and Denton 1985).

Ernest Burgess (1928) was the first to measure patterns and levels of residential segregation, but he was unable to establish standard methodology for doing so. As a result, social scientists became bogged down in a prolonged debate about the best segregation index. The issue was settled in 1955 by Duncan and Duncan (1955), who recommended using the index of dissimilarity. This straightforward measure varies between 0 (no segregation) and 100 (complete segregation) and represents the relative percentage of minority members who would have to exchange

neighborhoods with majority members to achieve an even, or integrated, residential distribution.

For the next 20 years, social scientists working in the tradition of the Chicago School employed this index to measure levels of segregation in a variety of urban settings. Duncan and Duncan (1957) undertook their classic study of the levels, causes, and consequences of black segregation in Chicago. Lieberson (1963) followed with an analysis of segregation among European-origin groups living in US industrial cities. In their ambitious study, Taeuber and Taeuber (1965) documented patterns of segregation and neighborhood change for African Americans in cities throughout the USA. As the racial and ethnic composition of other nations has been transformed by postwar immigration (Massey et al. 1998), social scientists have applied the index to study the segregation of minority groups in Britain, Germany, France, The Netherlands, Israel, Canada, and Australia (Massey 1985).

In 1976, Cortese et al. ignited a new debate about the measurement of segregation by questioning the applicability of the index of dissimilarity under certain circumstances. In response to their critique, numerous alternative indices were proposed to yield considerable confusion in the literature. Massey and Denton (1988) once again brought order to the field by demonstrating that residential segregation was a multidimensional construct characterized by five distinct axes of spatial variation—evenness, exposure, clustering, concentration, and centralization. They showed that the index of dissimilarity measured the evenness dimension fairly well, thus confirming Duncan and Duncan's earlier work. After clarifying the conceptual and empirical properties of the remaining dimensions, they went on to recommend a specific index for each one.

Massey and Denton (1989) used this multidimensional conceptualization of segregation to define the concept of *hypersegregation*, which occurs whenever a group experiences high segregation on at least four of the five dimensions at once. They showed that whereas blacks in 20 metropolitan areas displayed high levels of segregation on four or more dimensions simultaneously, Asian and European ethnic groups in US metropolitan areas never achieved a high level of segregation on more than one dimension at a time. As of 1990, more than one third of all African Americans lived under conditions of hypersegregation (Denton 1994).

Despite the widespread recognition of residential segregation as a multidimensional construct, the most commonly used measure continues to be the index of dissimilarity. The next most common index is the P^* isolation index, which Massey and Denton identify as a measure of exposure. It represents the percentage of minority members living in the neighborhood of the average minority person. An index of 80 for African Americans, for example, means that the average black person lives in a neighborhood that is 80 percent black. Unlike the index of dissimilarity, this index is strongly influenced by the relative number of minority members within an urban area, so that intergroup and intercity comparisons are best made using the index of dissimilarity.

A convenient rule of thumb for interpreting the latter index is that values under 30 are 'low,' those from 30 to 60 are 'moderate,' and those above 60 are 'high' (Kantrowitz 1973). In general, voluntary processes of segregation yield index values in the low-to-moderate range. Whenever dissimilarity indices persist at levels above 60 for any length of time, it usually signals the operation of involuntary forces of one sort or another. Studies done in the USA reveal that African Americans continue to experience high levels of discrimination in real estate, banking, and insurance markets, and that they remain objects of considerable prejudice by whites (Massey and Denton 1993). As a result, black–white residential dissimilarities in US urban areas generally range upward from 60, and unlike those of other ethnic minorities, they do not fall as socioeconomic status rises (Demon and Massey 1988).

In the USA, black residential segregation has also displayed little tendency to decline over time, except in places where the number of blacks is so small that complete desegregation would yield little interracial mixing within neighborhoods (Krivo and Kaufman 1998). Other than African Americans in the USA, the only documented case where a racial or ethnic group has experienced a prolonged, high level of segregation is that of black Africans in the Union of South Africa under apartheid (Christopher 1993). Blacks in neither Canada (Fong 1996) nor Brazil (Telles 1992) encounter the extreme levels of segregation experienced by their US counterparts.

New interest in the consequences of residential segregation was stimulated by William Julius Wilson (1987), who argued that the social isolation of blacks within neighborhoods of concentrated poverty systematically lowered their chance of success in employment, marriage, and education. Although he attributed this growing isolation to structural shifts in the economy and the changing geography of employment, Massey and Denton (1993) showed that these changes necessarily produce high concentrations of poverty when they occur under conditions of high racial segregation. Given rising income inequality and the high degree of residential segregation experienced by African Americans in the USA, the concentration of black poverty is mathematically inevitable.

Studies drawing on longitudinal data files reveal that growing up in a poor neighborhood has negative socioeconomic consequences that are independent of personal characteristics or family circumstances. Coming of age under concentrated poverty undermines cognitive development in early childhood, reduces academic achievement in later adolescence, and elevates the risk of antisocial behavior throughout

the life course (Brooks-Gunn et al. 1997). Males experience a higher likelihood of withdrawing from the labor force and increased risk of criminal involvement, whereas women are less likely to marry and more likely to become single parents (Massey and Shibuya 1995).

The link between racial segregation and stratification prevails no matter what one assumes about the extent of segregation by social class, but when high levels of class segregation occur in addition *to* racial or ethnic segregation, the concentration of poverty is exacerbated. Although the necessary data are not available for most countries, studies done in US cities reveal a marked rise in the degree of income segregation from 1970 to 1990 (Massey 1996), a trend that occurred across all regions and racial/ethnic groups (Jargowsky 1996).

Although rising income inequality and growing class segregation may undermine the socioeconomic welfare of all racial and ethnic groups, the consequences have been particularly severe for African Americans because they are also so highly segregated by race. Under these circumstances, poor blacks experience far higher concentrations of neighborhood poverty than the poor of other groups. Although few studies have explored segregation by class, race, and ethnicity simultaneously, to the extent that rising income inequality, growing class segregation, and persisting racial/ethnic segregation coincide in other nations, a similar ecology of inequality is expected to prevail.

In the new century, social scientists are working to disentangle the effects of metropolitan structure, neighborhood composition, and family circumstances on individual socioeconomic outcomes. In this effort, researchers have turned increasingly to multilevel models and longitudinal data sets. Although the questions may be similar to those posed by the Chicago School theorists in the 1920s, the measures, methods, and data are now far more complex and sophisticated. Specifying the theoretical nature and empirical strength of connections between ecological structure, neighborhood conditions and individual outcomes will remain a central preoccupation of social science in the twenty-first century.

See also: Discrimination; Discrimination, Economics of; Discrimination: Racial; Ethnic Conflicts; Ethnicity, Sociology of; Housing Economics; Race and Urban Planning; Racial Relations; Racism, History of; Racism, Sociology of; Segregation Indices; Spatial Equity; Urban Geography; Urban Poverty in Neighborhoods

Bibliography

Brooks-Gunn J, Duncan G, Aber L 1997 *Neighborhood Poverty, Volume I: Context and Consequences for Children*. Russell Sage, New York

Burgess E W 1928 Residential segregation in American cities. *Annals of the American Academy of Politics and Social Science* **140**: 105–15

Christopher A J 1993 Segregation levels in the late-apartheid city 1985–1991. *Tijdschrift voor Economische en Sociale Geografie* **83**: 15–24

Cortese C F, Falk R F, Cohen J C 1976 Further considerations on the methodological analysis of segregation indices. *American Sociological Review* **41**: 630–7

Denton N A 1994 Are African Americans still hypersegregated? In: Bullard R D, Grigsby J E, Lee C (eds.) *Residential Apartheid: The American Legacy*. CAAS Publications, Los Angeles

Denton N A, Massey D S 1988 Residential segregation of blacks, Hispanics, and Asians by socioeconomic status and generation. *Social Science Quarterly* **69**: 797–817

Duncan O D, Duncan B 1955 A methodological analysis of segregation indices. *American Sociological Review* **20**: 210–17

Duncan O D, Duncan B 1957 *The Negro Population of Chicago: A Study of Residential Succession*. University of Chicago Press, Chicago

Fong E 1996 A comparative perspective on racial residential segregation: American and Canadian experiences. *The Sociological Quarterly* **37**: 199–226

Jargowsky P A 1996 Take the money and run: Economic segregation in U.S. metropolitan areas. *American Sociological Review* **61**: 984–98

Kantrowitz N 1973 *Ethnic and Racial Segregation in the New York Metropolis*. Praeger, New York

Krivo L J, Kaufman R L 1998 How low can it go? Declining black–white segregation in a multi-ethnic context. *Demography* **36**: 93–109

Lieberson S 1963 *Ethnic Patterns in American Cities*. Free Press, New York

Massey D S 1985 Ethnic residential segregation: A theoretical synthesis and empirical review. *Sociology and Social Research* **69**: 315–50

Massey D S 1996 The age of extremes: Concentrated affluence and poverty in the 21st century. *Demography* **33**: 395–412

Massey D S, Arango J, Hugo G, Kouaouci A, Pellegrino A, Taylor J E 1998 *Worlds in Motion: Understanding International Migration at the End of the Millennium*. Oxford University Press, Oxford, UK

Massey D S, Denton N A 1985 Spatial assimilation as a socioeconomic process. *American Sociological Review* **50**: 94–105

Massey D S, Denton N A 1988 The dimensions of residential segregation. *Social Forces* **67**: 281–315

Massey D S, Denton N A 1989 Hypersegregation in US metropolitan areas: Black and Hispanic segregation along five dimensions. *Demography* **26**: 373–93

Massey D S, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA

Massey D S, Shibuya K 1995 Unravelling the tangle of pathology: The effect of spatially concentrated joblessness on the well-being of African Americans. *Social Science Research* **24**: 352–66

Park R E 1926 The urban community as a spatial pattern and a moral order. In: Burgess E W, Park R E (eds.) *The Urban Community*. University of Chicago Press, Chicago

Park R E, Burgess E W 1925 *The City*. University of Chicago Press, Chicago

Taeuber K E, Taeuber A F 1965 *Negroes in Cities: Residential Segregation and Neighborhood Change*. Aldine, Chicago
Telles E E 1992 Residential segregation by skin color in Brazil. *American Sociological Review* **57**: 186–97
Wilson W J 1987 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago

D. S. Massey

Resource Geography

Resource geography is part of a broad tradition of geographical inquiry examining the reciprocal relationships between socioeconomic systems and physical environmental processes. Within this 'nature society' tradition, resource geography is defined by its focus on how economic, political, and social institutions and practices determine patterns of resource use over space and time (see *Nature–Society in Geography*). This article begins with an explanation of the term 'resource,' outlines the scope of resource geography, and identifies four principal approaches within the field.

1. *Resources Defined*

Resources refer to those properties of the physical environment that can be exploited to meet the needs and desires of society. Resources can be understood, therefore, as culturally mediated appraisals of the physical environment. A resource does not exist *a priori*, but it produced from 'neutral stuff' through its interaction with a set of prevailing cultural norms which collectively define its social utility and economic viability: thus 'resources are not, they become' (Zimmerman 1951, p. 15). The cultural milieu consists of economic variables (markets, prices, and technology); political institutions (property rights, legislation, and regulatory codes); social attributes (perception, behavior, and practices of decision-making); and cultural belief-systems (cosmologies of nature, society, and development). For example, just as sperm whales were regarded in the early nineteenth century as a valuable resource from which to obtain high quality lighting oil, so the recent appraisal of tropical forests as resources to be 'mined' for pharmaceutical benefits derives from the confluence of a specific set of economic, legal, technological, and cultural processes.

Resources are most commonly categorized based on their renewability. Renewable (or flow) resources are those depleted by use in the short term but which can replenish themselves over the longer term (e.g., forests, soils, and fisheries). Non-renewable (or stock) resources have a finite supply: consumption today comes at the expense of future use (e.g., minerals and fossil fuels). In practice, however, this distinction is more accurately represented as a continuum since the renewability of many resources depends on when, where, and how they are consumed. For example, the experience of many communities with fisheries, forests, and groundwater demonstrates that if the rate of use exceeds the rate of renewal, even a potentially renewable resource can be depleted. In this context it is possible to speak of the 'mining' of renewable resources like soil or groundwater.

The 'resource question' has deep and enduring roots within Western intellectual history and is characterized by two distinct questions (Rees 1990). The first is the role of resource availability/resource scarcity in limiting opportunities for economic growth and development. Early explorations of this theme can be found in Malthus's (1798) thesis on the differential rates of growth of population and food supply, and Marx's (1967) account of the productive capacity of the labor process to liberate humankind from natural limits. More recently it appears in debates over the long-term availability of water, soil, and energy resources for a growing world population (Barnett and Morse 1963). The second is the sociopolitical implications of resource use and the impact of resource management on environmental quality, equity, and human welfare across a range of temporal and spatial scales. Examples of this can be found in conservationist writings of the late nineteenth century, in studies of social inequalities in exposure to environmental risk, and in current engagements with global environmental change and sustainable development.

Resource geography has tended to shun theoretical development in preference for a pragmatic engagement with concrete problems and a prescriptive approach to policy formation (see Wescoat 1987). Four principal lines of inquiry can be identified: resource inventory and evaluation; resource management institutions and decision-making; the political economy of resource use and the environment; and critiques of resource management and the question of 'nature.'

2. *Lines of Inquiry within Resource Geography*

2.1 *Resource Inventory and Evaluation*

The stock-in-trade of resource geography has traditionally been the identification and analysis of resource distribution at spatial scales from local to global (e.g., Murphy 1954). For example, descriptions of natural resource distribution conventionally form a central component of regional geography, where they provide a logical segue between the description of a region's geological, hydrological, pedological, and biological characteristics and discussion of the region's

socioeconomic geography. Using this spatial mapping approach, resource geographers have quantified natural resource endowments, evaluated the capacity of particular natural resource bases to support economic development, and determined the extent to which current strategies of resource acquisition, conservation, and environmental protection are adequate.

By providing assessments of the spatial extent, utility, and economic viability of mineral, biological, and hydrological resources, resource geography has often played the role of handmaiden to regional planning and economic development. Early examples from within US geography can be found in the watershed and irrigation survey work of geomorphologists such as Grove Karl Gilbert (1877) and John Wesley Powell (1879). Their evaluations of water resources in the arid western states in the second half of the nineteenth century helped promote the federal government's involvement in dam construction, irrigation, and inter-regional water transfer systems. In the UK, the development of the national Land Utilization Survey by Dudley Stamp (1934) is indicative of how resource geography has historically contributed to land-use planning and, more generally, to the implementation of a conservationist utilitarianism towards natural resources. Professional affiliation between resource geographers and public resource agencies continues to generate empirical resource inventories and practical approaches for implementing integrated resource and environmental management (see Mitchell 1997).

The strategic and military value of mineral and energy resources has often provided an overtly political focus to the work of resource geographers, confirming MacKinder's (1887) maxim that geography serve as an aid to statecraft. For example, resource geographers served as part of a multidisciplinary team of social scientists compiling a series of Area Handbooks commissioned by the US Department of the Army (see www.hqda.army.mil). These 'compilations of basic facts' describe economically and militarily significant aspects of selected countries including identification and inventory of natural resource endowments. Whether part of regional or applied geography, the primary objective of local, regional, and global resource appraisals is informational. As a consequence, this line of inquiry has produced maps of resource location, atlases, and descriptions of resource availability, but has contributed little in the way of theoretical or methodological development.

2.2 Resource Management Institutions and Decision-making

A second major focus of resource geography is practically orientated research on the management of natural resources and the environment. The historical and intellectual foundations of this approach lie within welfare economics with its objective of utility maximization. Adopting this welfare framework as its core paradigm, resource geography has sought to maximize the social utility of resources such as coal, oil, water, and open space by assisting in their rational and efficient management. A central objective has been to understand the influence of institutional frameworks (e.g., the distribution of property rights, the nature of environmental legislation, and the structure of commodity markets), decision-making processes, and differences in resource perception on the allocation, exploitation, and conservation of natural resources over time.

Under this paradigm, resource shortages and environmental degradation are understood to result from market failure and pervasive externalities. As a consequence, research adopts a prescriptive stance, advocating public policy approaches to intervene in existing markets to address the causes of failure (e.g., via government regulation for environmental quality or resource taxation), create markets where none had previously existed (e.g., pollution permit trading), or re-design the institutions of property ownership to internalize external costs (e.g., resource privatization, see Anderson and Leal 1992).

Prescriptions for resource management drawn from neo-classical economic models came under criticism from behavioral studies during the 1960s. Work by Gilbert White (1964) on river basin development and floodplain management, for example, argued that existing public policies for resource and hazard management over-simplified the process by which management alternatives were selected. Arguing that flood control engineering was just one of several possible responses to flood risk (others include flood insurance, disaster warning systems, and comprehensive floodplain management), White's work—and that of the Natural Hazards Research Group at the University of Chicago—focused attention on the range of choice in resource management decisions (*Environmental Adaptation and Adjustments*; *Environmental Risk and Hazards*).

In contrast to the neo-classical model which interpreted resource scarcity or deteriorating environmental quality in terms of market failure, the Chicago School argued that poor resource management decisions resulted from misperception or limited awareness of better alternatives. It rejected the assumptions of utility maximizing individuals operating in a field of perfect knowledge and instead adopted a behavioral approach in which resource managers operated under conditions of bounded rationality. Initially focused on water resources in USA, natural hazards research expanded to include other geophysical extremes in North America and in different cultural contexts. As an alternative to the unitary and utility maximizing assumptions of conventional resource management, the behavioral approach captures the

role of perception and sociocultural rules in determining the relative success of strategies such as energy and water conservation or the construction of international regimes for environmental governance (e.g., marine pollution, migratory species, ozone depletion—see Young 1999).

In an effort to contextualize the process of decision-making and explore the opportunities for co-operation as well as competition in the management of natural resources, resource geography has looked to traditional societies for alternative to market-based resource regimes. Research has focused, for example, on management regimes for common property resources—those for which control of access is difficult and joint use affects the welfare of each resource user (e.g., fisheries, river systems, wildlife, pastures, and public parks,) (see *Environmental and Resource Management*). Interest in management regimes for the commons centers on Hardin's (1968) provocative assessment that common property resources were inevitably prone to environmental degradation as a result of the cumulative decisions of economically rational individuals. Using the analogy of herders stocking a commonly held pasture, Hardin argued that the pursuit of individual self-interest will lead each herder to add to his or her stock; while the benefits of an additional animal (increased income) will be captured by the individual, the costs (declining grassland productivity) will be shared among all users. Describing this as the 'tragedy of the commons,' he concluded that either government ownership or privatization of the commons was necessary to forestall this degradation since 'freedom of the commons brings ruin to all.'

However, subsequent work by resource geographers and anthropologists has revealed historical and contemporary examples of successful common property resource management (e.g., Feeny et al. 1990). This has generated a two-pronged critique of Hardin's conclusions: first, Hardin failed to acknowledge that his herders could communicate with each other and, therefore, that they could undertake collective rather than individual action to maintain the environmental integrity of the commons; and second, Hardin conflated the nature of a common property resource with the type of property regime under which it was managed. He assumed, therefore, that his pasture—a classic common pool resource—would be managed as an open access resource regime. Historical experience illustrates that common property resources can exist under open access, communal, state, and private property regimes and that these need not be mutually exclusive but can coexist.

2.3 The Political Economy of Resource Use and the Environment

A third line of inquiry within resource geography explicitly rejects the applied management orientation of earlier approaches for failing to capture the social stratification of economic and political power. In contrast, it engages the political and social dimensions of resource use, emphasizing the ecological and socioeconomic context in which resource managers and their decisions are embedded. By embracing a broadly defined political economy, resource geography has sought to understand the nested scales of activity—from the actions of local institutions, through national public policies, to the global strategies of corporations—which constrain and influence the choices of individual resource managers.

Three distinctive subfields have emerged from this critique of conventional resource management: political ecology (e.g., Blakie and Brookfield 1987, *Feminist Political Ecology*); radical hazards research (e.g., Hewitt 1983); and environmental justice (e.g., Pulido 1996, *Environmental Justice*). Common to all three subfields is a perspective that locates the origins of resource depletion and environmental degradation in the economic, political, and social institutions of production. As a result, recognition is given to the socioeconomic—as opposed to purely physical-environmental—dynamics of vulnerability. For example, Watts' (1983) analysis of vulnerability to drought among the Hausa peasantry of northern Nigeria locates the cause of this vulnerability in the articulation between pre-capitalist modes of production and a global capitalist system.

2.4 Critiques of Resource Management and the Question of 'Nature'

Resource geography remains an evolving field influenced by both topical issues in society and theoretical trends within the academy. For example, recent public policy debates in North America and Europe have provided a topical context for research on: the effectiveness of conventional environmental policy approaches; the integration of different spatial scales in managing natural habitat; managing socioeconomic transition within natural resource communities (e.g., those based on coal mining, fishing, or timber harvesting); the environmental implications of industrialization and modernization in the developing world; and the challenges of operationalizing sustainable development.

Resource geography has traditionally shown theoretical restraint in favor of practical application (Wescoat 1987); however, the 1990s has seen an increased interest in praxis—the integration of theoretical perspectives with the day-to-day issues of producing, consuming, and managing natural resources. Indicative of this are recent engagements between resource geography and:

(a) theories of economic restructuring and rural sociology in seeking to understand the implications for communities of agricultural restructuring, resource

privatization, and the links between local and global regimes of food production;

(b) the question of 'nature' and environmental ethics, and their implications for the welfare of non-human species;

(c) discussions within critical legal theory and discourse theory on the ideologies and institutions that sustain particular actions towards nature (Proctor 1998, Wolch and Emel 1998).

Collectively these new directions offer a broad critique of the instrumentalism inherent in the categorization of parts of nature as 'resources.' A central objective is to destabilize conventional views of nature as a collection of severable, tradable resource components valued solely for their commodity value. By recognizing such common-place notions as 'resources,' 'wilderness' and 'nature' as historically and geographically specific social constructions, new directions within resource geography contribute to an understanding of the roles gender, race, and class have played in domesticating the natural world.

See also: Energy, Geography of; Resource Institutions; Water Resources

Bibliography

Anderson T, Leal D 1992 *Free Market Environmentalism*. Westview Press, Boulder, CO

Barnett H, Morse C 1963 *Scarcity and Growth: The Economics of Natural Resource Availability*. Johns Hopkins University Press, Baltimore, MD

Blakie P, Brookfield H 1987 *Land Degradation and Society*. Methuen, London

Feeny D, Berkes F, McCay B, Acheson J 1990 The tragedy of the commons: Twenty years later. *Human Ecology* **18**: 1–19

Gilbert G K 1877 *Report on the Geology of the Henry Mountains*. US Government Printing Office, Washington, DC

Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8

Hewitt K 1983 (ed.) *Interpretations of Calamity*. Allen and Unwin, Boston

MacKinder H 1887 The scope and methods of geography. *Proceedings of the Royal Geographical Society* **9**: 141–60

Malthus T 1798 (reprinted 1986) *An Essay on the Principle of Population*. Penguin Classics, Harmondsworth, UK

Mark K 1967 *Capital* (3 volumes). International Publishers, New York

Mitchell B 1997 *Resource and Environmental Management*. Longman, Harlow, UK

Murphy R 1954 The geography of mineral production. In: James P, Jones C (eds.) *American Geography: Inventory and Prospect*, Syracuse University Press, Syracuse, NY

Powell J W 1879 *Report of the Lands of the Arid Region of the United States* (*with a more detailed account of the lands of Utah*). US Government Printing Office, Washington, DC

Proctor J 1998 The social construction of nature: relativist accusations, pragmatist and critical realist responses. *Annals of the Association of American Geographers* **88**: 352–76

Pulido L 1996 *Environmentalism and Economic Justice: Two Chicano Struggles in the Southwest*. University of Arizona Press, Tucson, AZ

Rees J 1990 *Natural Resources: Allocation, Economics, and Policy*. Routledge, New York

Stamp L D 1934 Land utilization survey of Britain. *Geographical Review* **24**: 646–50

Watts M 1983 *Silent Violence: Food, Famine, and Peasantry in Northern Nigeria*. University of California Press, Berkeley, CA

Wescoat J 1987 The 'practical range of choice' in water resources geography. *Progress in Human Geography* **11**: 41–59

White G 1964 Choice of adjustment to floods. Department of Geography Research Paper No. 93, University of Chicago

Wolch J, Emel J 1998 *Animal Geographies: Place, Politics, and Identity in the Nature-Culture Borderlands*. Verso, London

Young O 1999 (ed.) *The Effectiveness of International Environmental Regimes*. MIT Press, Cambridge, MA

Zimmerman E W 1951 *World Resources and Industries*. Harper, New York

G. Bridge

Resource Institutions

Human beings use, manage, or conserve resources through institutional arrangements. The absence of institutions leads to open access to resources, a condition in which overuse and depletion may be unavoidable. Institutions exercise an influence over resource management outcomes that is independent of the forces and dynamics that created them. Institutional innovation is therefore critically important in all efforts to influence resource use and conservation. Whether one attempts to rejuvenate depleted fisheries or prevent trade in endangered wildlife, to protect forests, or prevent the erosion of biodiversity, institutional change remains a crucial part of such efforts. Given the crucial role of institutions, it is clear that an understanding of the nature of institutions and their relationship to resources is necessary if the goal is sustainable management or conservation.

Despite variations in approaches to institutional analysis, resource institutions can be defined as sets of rules that can be crafted collectively or individually, and that have the power to shape actions related to the conservation and use of resources. Resource systems and institutional arrangements are distinct entities, and it is important to pay attention to the factors that lead to a better fit between a given resource and the rules that govern the management of that resource.

However, it is only in the second half of the twentieth century that the role of institutions in resource management has found substantial recognition in the scholarly and policy world. Part of the reason institutions are recognized as important is that many resources critical to human existence and subsistence are becoming increasingly scarce, and are difficult to manage under some universalizable institutional form. Very different institutional regimes may be necessary for different types of resources. Consider,

as examples, wildlife species that need hundreds of square miles as their habitat, scattered forest patches upon which thousands of households depend for their livelihood, or irrigation systems that are too small to require government intervention, but too large to be managed by a single user. International agreements among governments may be appropriate institutional mechanisms to restrict the production of ozone depleting chemicals because they are mostly generated by a few major industrial producers. But for reduction in the levels of greenhouse gases, a far more widespread program of restraint may be necessary because of the nonpoint nature of the production of these gases. Another reason for the importance of getting institutions right are transactions costs that are associated with the allocation, processing, and exchange of resources, and that vary with institutional arrangements. Institutions that rely on external enforcement may require higher levels of expenditure to ensure that users conform to rules in comparison to those situations where it is in the interests of users to monitor each other and enforce institutional provisions directly.

The range of rules that comprise institutions is potentially infinite. But three institutional forms are possibly the most familiar: markets, states, and communities. Within each of these categories it is possible to find many variations; other forms span these categories. The precise combination of rules and how rules influence human behavior and outcomes are the basis on which it is possible to judge the performance of institutional arrangements. Typically, the objective of creating or forming institutions over resources is to enhance the efficiency, equity, or sustainability of resource use. However, a truly enormous and highly complex set of factors are at play. At the same time, the application of institutional analysis to resources is a relatively new field. It also requires collaboration across disciplinary boundaries. These are two reasons why the precise nature of the relationship between social and physical/biological systems, signified by the phrase 'resource institutions,' still requires considerable scientific research.

1. *A Brief History of Institutions and Their Place in Resource Management*

Contributions to institutional theories have flowed from economists, sociologists, and political scientists (Goodin 1996). Such contributions resonate with larger themes in social theory about the relationship between social structure, agency, and outcomes. In explicit counterpoint to the neoclassical paradigm in economics that finds little role for institutions or organizations in social life, advocates of institutions such as North (1990) and Williamson (1985) emphasize the unavoidable influence of institutional constraints on all social action. New institutionalists, property rights theorists, and scholars of the commons have been the most influential in providing insights about the nature of resource institutions. Efforts aimed at understanding institutions have drawn from rigorous mathematical approaches as well as more descriptive and interpretive orientations.

Although resource institutions may serve as constraints upon certain groups and individuals, or certain kinds of actions, they also facilitate and advantage others. In addition to facilitating or constraining actions, and advantaging some people over others in terms of power and benefits (Knight 1992), resource institutions may also help shape desires, motivations, and preferences of groups and individuals. They are rooted in past human actions, and are the result of choices made by individuals and groups. It is easy to see, therefore, that institutions are seldom stationary; they change over time because of human choices.

For a long time, scholars viewed markets and states as the appropriate institutional mechanisms to address externalities that result from the collective goods nature of a vast class of resources. Advocates of market solutions argued for privatization of collectively owned and managed resources such as fisheries and forests, and argued that without privatization collective goods are likely to be overexploited. Voluntary exchanges among private owners on the basis of pricing signals, they held, would lead to the most efficient use of resources. Advocates of state control and ownership, pointing to asymmetric allocations that often result from the operation of market mechanisms, suggested that centralized control of resources was far more likely to promote conservation and sustainable use. Without centralized enforcement, according to them, attempts to manage resources could turn out to be futile. Through much of the 1950s and 1960s, writings on resource use and management viewed community institutions as backward, inefficient, and likely to promote overexploitation.

From the late 1970s, and especially since the mid-1980s, there has been a significant shift in views about the role of community as a resource management institution. This is partly because resource management strategies that rely exclusively on exchange (market) or enforcement (coercive, state-based) have obvious deficiencies. One tends to benefit a small group or a few individuals, and the other tends to create highly asymmetric relations of domination and dependency. Both have shown a tendency to disadvantage and exclude rural residents relying on local resources, who then do not cooperate in management of these resources. Both at the level of policy initiatives, and in terms of theoretical developments, the option of collective management of resources through new types of resource institutions based in communities has found wide appeal (National Research Council 1986, Ostrom 1990).

Despite heated assertions about the superiority of one type of resource management institution over

another, in real life most resource institutions do not easily fit the mold of market, state, or community. For the operation of each of these forms of institutions, implicit or explicit understandings that shade into other forms are necessary. For example, often markets and private property arrangements work better only when the force of the state is available to enforce the security of property.

2. Resource Institutions, Rules, and Power

Since resource institutions are sets of rules and norms that shape action, it is necessary to understand the broad types of rules that are most important in making institutions effective. Effectiveness implies three categories of rules: the ability to create new rules and modify old ones pertaining to resources; the ability to implement and ensure compliance to existing and new rules; and the ability to adjudicate disputes that arise in the effort to create rules and ensure compliance. Rules pertaining to resources define who gets to access a resource, who gets to use them, who is excluded from a resource, and how and when a resource can be transferred from one individual or group to another. Institutional arrangements are likely to be ineffective when human beings can make these above rules but not enforce them, or enforce existing rules but not change them in case of deficiencies. Mechanisms that allow adjudication of disputes among resource users and managers are necessary if resource institutions are to retain their force over time. These three categories of rules correspond to a more familiar division in government: executive (making, implementing, and enforcing of decisions), legislative (creation of rules), and judicial (adjudication of disputes).

The constituent parts of all resource institutions are born as a result of human choices. Three important determinants of human institutional choices are scarcity, risk, and redistribution. Greater scarcity of resources in a group creates incentives to define institutional arrangements ever more precisely so that benefits from resources are not dissipated. Indeed, some scholars have seen this dynamic of induced institutional change to be prominent enough to argue that it is universal (Netting 1981). Institutions may also change in response to higher levels of risk in the environment. Collective action necessary for institutional emergence can be conceptualized as a response that addresses the need to smoothen intertemporal fluctuations in production (Wade 1994). These views of institutional change are insufficiently attentive to variations in the incentives of group members, and to the politics that such variations generate. A more political view holds that institutional change need not occur only in response to greater scarcity in a group, nor lead to greater efficiency (Knight 1992, North 1990). Instead, individuals and groups often act to create new resource institutions that would benefit them, rather than address problems of scarcity or risk. Particular institutions reflect existing distribution of power and wealth in a society or a group.

3. Sustainability of Resource Institutions

Perhaps the most critical question about resource institutions is how and under what conditions do they lead to sustainable management and conservation of resource systems. Group size, heterogeneity within groups, migration levels and population change, technological innovation, changing levels of demand for resources, uncertainty and sudden shocks, and cross-scale effects play an important role in determining the sustainability of institutions and resource systems (Keohane and Ostrom 1995). For certain kinds of resources, especially global commons, new institutional mechanisms have emerged to address issues of complexity and heterogenous subpopulations (Haas 1997).

Writings on resilience and adaptive management have generated important insights about the extent to which resource institutions can be sustainable in the long run (Holling 1973). Human institutions perform best when environments are predictable, yet many resource systems are characterized by potentially highly uncertain and variable processes. To date, research on institutions has not incorporated all the lessons learnt in risk-hazard studies on how sudden variations and uncertainty affect resource management needs. In part this is so because a formal theory of resource institutions requires a theory of human behavior that is mathematically rigorous and tractable, withstands empirical testing, and fits into a model of resource systems that can address fluctuations and surprise.

See also: Environmental and Resource Management; Environmental Vulnerability; Human–Environment Relationships; Institutional Economic Thought; Institutionalism; Land Degradation; Sustainable Development

Bibliography

Goodin R E (ed.) 1996 *The Theory of Institutional Design*. Cambridge University Press, Cambridge, UK

Haas P M (ed.) 1997 *Knowledge, Power, and International Policy Coordination*. University of South Carolina Press, Columbia, SC

Holling C S 1973 Resilience and stability of ecosystems. *Annual Review of Ecology and Systematics* **4**(1): 1–23

Keohane R, Ostrom E (eds.) 1995 *Local Commons, Global Interdependence: Heterogeneity and Cooperation in Two Domains*. Sage, London

Knight J 1992 *Institutions and Social Conflict*. Cambridge University Press, Cambridge, UK
National Research Council 1986 *Proceedings of the Conference on Common Property Resource Management*. National Academy Press, Washington, DC
Netting R 1981 *Balancing on an Alp: Ecological Change and Continuity in a Swiss Mountain Community*. Cambridge University Press, New York
North D C 1990 *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge, UK
Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, New York
Wade R 1994 *Village Republics: Economic Conditions for Collective Action in South India*. Institute for Contemporary Studies Press, San Francisco
Williamson O E 1985 *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. Free Press, New York

A. Agrawal

Resources and Environment: Contingent Valuation

A major impediment to performing a benefit–cost analysis involving the provision of a new public good or a change in an existing public good is that public goods are not routinely bought and sold in markets. Hence 'prices,' the economic data routinely used by economists as indicators of the economic value of goods, are not available for public goods such as environmental amenities. To overcome this obstacle, economists have developed (Freeman 1993) an indirect valuation approach that infers economic value from observations on consumer behavior with respect to marketed goods having a relationship to the public good, and a direct approach that gives consumers the opportunity to make choices with respect to the public good. The indirect approach relies upon one of two factors: that the public good can be 'bundled' in as one of the attributes of a private good (e.g., close proximity to a public park) and that it is sometimes necessary to make expenditures of money or time to use the public good (e.g., driving to a hiking trail in a national forest). The direct approach can be implemented by subjecting consumers to a constructed choice situation involving the public good, such as a vote in a referendum to provide the public good. The now common survey variant of the direct approach has come to be known as contingent valuation (CV) because the estimates of economic value obtained are 'contingent' on the features of the scenario posed in the survey.

CV surveys differ from other surveys on public policy issues in several important ways. First, the entire survey is devoted to describing the public good (or a small number of public goods) of interest. Second, they differ in that their major purpose is to obtain an estimate of the relevant Hicksian consumer surplus measure; maximum willingness-to-pay (WTP) to obtain a desired good not currently possessed, or minimum willingness-to-accept (WTA) compensation to voluntarily give up a good currently possessed. CV surveys are typically organized in the following manner which reflects current practice: (a) an introductory section identifying the sponsor and general topic, (b) a section asking questions about prior knowledge about the good and attitudes toward it, (c) the presentation of the CV scenario including what the project was designed to accomplish, how it would be implemented and paid for, and what will happen under the current status quo situation if the project were not implemented, (d) question(s) asking for information about the respondent's WTP/WTA for the good, (e) debriefing questions to help ascertain how well respondents understood the scenario, and (f) demographic questions. Mitchell and Carson (1989) provide a comprehensive overview of the issues involved in the design and analysis of CV surveys.

1. Development

Ciriacy-Wantrup (1947) put forth the first well-developed proposal on the need for CV surveys in a piece on the difficulties of measuring all the benefits of soil conservation programs. Empirical implementation of CV initiated by Davis (1963), in his Harvard dissertation, sparked considerable interest in the technique. He later compared a CV estimate with a corresponding estimate based on the travel cost method (an indirect approach then also being newly developed) and found that the two approaches produced similar estimates.

CV surveys were initially seen as having three distinct advantages. First, CV can obtain useful information where data on past consumer behavior had not been collected. Second, CV permits the creation and presentation of scenarios that provide new goods or changes in existing goods that were substantially outside the range of current consumer experience. Third, CV allows measurement of the desired Hicksian consumer surplus measure rather than its Marshallian approximation. For many economists, the major drawback to CV-based estimates was that they were based upon stated preferences rather than observed behavior.

A paper by Randall et al. (1974) greatly increased interest in CV. It valued changes in air quality necessary to maintain scenic vistas in the Southwest. The indirect valuation approach was not capable of being applied in this instance because all people in the area share the good equally (and hence, not bundled differentially into housing prices) and no expenditure of time or money is needed to enjoy it.

Randall et al. estimated what had been termed 'existence value' in an influential paper written by Krutilla (1967). The novel element in Krutilla's framework was that existence values were not generally revealed by market purchases. He argued that some people care about environmental resources, such as wildernesses areas, irrespective of their desire to visit them. Krutilla did not measure existence values but rather recommended determining how large they would have to be to tilt the decision in the other direction. Failure to include existence values in policy-making, he contended, would likely entail too great a loss of existing environmental amenities and the provision of too few new environmental amenities. Other related concepts were soon enumerated (e.g., non-use value, stewardship value, bequest value, option value) and eventually encompassed into a single term, 'passive-use value,' first used in the 1989 court decision *Ohio v. Department of Interior* (880 F.2d 432, D.C. Cir.), which held that government trustees should include passive-use values in damage claims.

What Krutilla had called existence values had often been termed intangibles. They were not well integrated into welfare economic theory and thought to be unmeasurable. The key to measuring them lies in the recognition that due to scarcity effects, any form of economic value is a relative and not an absolute concept. Monetary measures of economic value are implicitly defined by choices made subject to an income constraint. CV permits the construction of the appropriate choice scenario. Also, economic value can be expressed in terms of any constraints/tradeoffs appearing in a choice scenario including time, other public goods, or private goods.

The quantity of CV research steadily increased throughout the 1970s and early 1980s. The importance of CV in the USA was raised considerably by presidential executive orders that required an assessment of the benefits and costs of all major new government regulations and reauthorization of existing ones. Outside the USA, CV was incorporated into OECD reports on measuring the economic value of pollution impacts and into World Bank reports on providing basic infrastructure. The rapid growth in the use of CV since the mid-1980s, in the USA and elsewhere, was a response to the growing demand for more comprehensive benefit–cost assessments for policies with substantial environmental or health impacts.

By far the greatest stimulus to the current CV debate, however, was the enactment of US laws that allowed for the recovery of monetary damages for injuries to natural resources. The focal point was the Exxon Valdez oil spill where the government's case was largely based upon a monetary claim for loss of passive use (Carson et al. 1992). Potential liability for lost passive use values should increase a firm's precautionary activities and environmental restoration efforts. Not surprisingly, industry facing such liability questioned whether passive use values should be counted and whether CV estimates are reliable. A set of papers highly critical of CV from an Exxon-sponsored conference appear in Hausman (1993). In response to the industry critique, the US National Oceanic and Atmospheric Administration (NOAA) convened a Blue Ribbon Panel co-chaired by two Nobel Prize winners (Arrow et al. 1993) which concluded that passive use values should be included in damage assessments and that 'CV studies can produce estimates reliable enough to be the starting point for a judicial or administrative determination of natural resource damages—including lost passive-use value.' That report has also been influential for the guidelines that it suggested for conducting reliable CV studies for natural resource damage cases and the issues that it identified as requiring further research.

CV has been applied in more than 50 countries and there are now several thousand CV papers. Most US federal agencies with environmental, health, or natural resource responsibilities have relied upon CV studies in making policy decisions, as have many state agencies. The use of CV has dramatically increased over time in other OECD countries and international agencies regularly commission CV studies. The range of CV applications is broad. It has been used to estimate the benefits of: improving national and regional air/water quality; reducing risk from drinking and groundwater contaminants; improving outdoor recreational opportunities; protecting wetlands, wilderness, rainforests, and endangered species; providing public education; improving food and transportation safety; reducing queues for public health care services; increasing the reliability of electric and water utilities; providing cultural amenities; and improving sanitation services in developing countries.

2. *Initial Research Focus*

Two concerns first voiced during the early phase of CV development have been particularly influential: the possibility that respondents would not take seriously a 'hypothetical' survey question with no money directly changing hands and Paul Samuelson's warning that people would not truthfully reveal their preferences for public goods. Random response to a CV survey which was feared due to its 'hypothetical' nature could be ruled out by noting whether WTP responses systematically varied with particular covariates in the expected way. The tests economists paid the most attention to, however, were those comparing CV estimates with those from indirect techniques (e.g., hedonic pricing and travel cost analysis) for quasi-public goods, where both approaches are applicable. Early work suggested reasonably close comparison between the two, thus ruling out the extreme random behavior or strategic behaviour initially feared. Carson et al.'s (1996) meta-analysis of these comparative studies (616 comparisons from 83 separate

studies) found that CV estimates on average were slightly lower than their corresponding indirect technique estimates and the CV and indirect estimates were highly correlated.

For pure public goods, the indirect approach is not possible so one can only compare estimates based upon different types of constructed markets. With coercive payment (e.g., taxes), incentive theory suggests that CV responses should be similar to those obtained from observing voting in an equivalent referendum, the observed empirical result. With voluntary payment, theory suggests that survey estimates will be larger than true (and unobservable) WTP that in turn should be larger than actual contributions due to free riding. The observed empirical result is that survey WTP is substantially larger than actual contributions.

Other issues that concerned early CV researchers were whether the payment vehicle used should influence WTP estimates, how much information should be provided to respondents about the good, what fraction of WTP/WTA was due to passive use considerations, what the 'best' elicitation format was, why the difference between CV, WTP, and WTA estimates was larger than expected, and the relative merits of different modes of survey administration. Some of these issues have been resolved whereas others continue to be debated. The latter are discussed in the next section.

WTP estimates for the same good were shown to differ, often considerably, with different payment vehicles. Respondents were not indifferent to how public goods were financed. This was distressing to many who believed, or at least hoped, that the value of a good should be independent of the exact manner in which it was supplied and paid for. The CV estimates suggested that this was not the case.

As one would expect, WTP estimates were sensitive to the nature and amount of information provided. However, this sensitivity entailed the thorny issue of what information should be provided. CV researchers began undertaking extensive survey development work including focus groups to determine what information people wanted in making their decision. Many early CV surveys asked respondents to apportion their WTP estimate into different parts corresponding to various types of use and passive use values. Such decompositions were shown not to be unique and the various components had no distinct policy implications, so this practice has largely been abandoned.

3. *Research Issues*

The issue of the best elicitation format continues to be debated but is increasingly being cast in terms of a bias-variance tradeoff. The binary discrete choice format with random assignment of respondents to different amounts (Bishop and Heberlein, 1979) has gained prominence owing to its desirable incentive properties for truthful preference revelation, in many (but not all) situations. This format has two major drawbacks. First, it collects relatively little information about a respondent's value for the good (i.e., only whether this value is above or below the amount asked), thus requiring large samples for precise estimates. Second, its estimates are sensitive to the particular parametric functional form chosen. Response formats directly asking for WTP provide the maximum amount of information (the amount making the respondent indifferent between the status quo and the alternative). This format, however, invites nontruthful strategic behavior and typically results in a sizeable fraction of the sample providing either a zero WTP amount or no response. CV researchers have proposed a number of different response formats falling between these two extremes.

Early CV WTA estimates were considerably larger than corresponding WTP estimates. This was originally thought to be a defect of CV surveys but similar divergences now appear to be typical of experimental and actual markets. A number of reasons for large divergences have been suggested (Bateman and Willis, 1999), some consistent with neoclassical economic theory and some more psychological in nature. As a consequence, interest has been renewed in measuring WTA, since the substitution of a WTP measure in its place can substantially underestimate the value of a good.

The merit of different modes of survey administration also continues to be debated. Many CV studies need visual representations of the good (e.g., pictures or maps) precluding the use of telephone surveys. The in-person vs. mail survey debate has several elements, including the much greater cost of personnel interviews, the importance of controlling the order in which material is presented, the importance of being able to read material to respondents, potential sample selection bias due to interest in the topic, and the potential influence of the interviewers.

CV researchers also came to realize that they faced the standard array of survey design issues. Adequate presentation of the material in the typical CV survey is now seen as a much more difficult task than the standard public opinion survey and is one for which most economists have had little training or experience. Further, several cognitive psychologists (e.g., Kahneman and Knetsch 1992) argued that the estimates of economic value obtained in CV surveys were likely to be subject to several undesirable effects. Chief among these was that the estimates obtained in a CV survey would be insensitive to the scope of the good being valued. They argued that stated intentions and actual behavior will diverge because respondents get a 'warm glow' from 'giving' in a survey context. While there are a few study results consistent with this prediction, a fairly large number of split-sample tests

using more fully developed CV scenarios have tested this hypothesis and reject it. Nonetheless, the Kahneman and Knetsch paper stimulated considerable interest in testing the properties of CV estimates against predictions from economic theory.

4. State of the Debate

Although there is widespread general agreement among CV practitioners on many issues, they often disagree regarding the extent and applicability of various best-practice standards, such as those proposed by the NOAA Panel, when cost and complexity factors are introduced. The recent Bateman and Willis (1999) volume considers these issues.

Participants in the early years of the CV debate were mostly environmental economists. One now finds that corporate executives, ecologists, lawyers, microeconomic theorists, philosophers, psychologists, public administrators, statisticians, and survey researchers have joined in. The increasingly interdisciplinary nature of this debate attests to the complexity of the issues and importance of CV for policy decisions (Portney 1994). Opponents of CV often conclude that CV is fundamentally flawed (Diamond and Hausman 1994). Their typical recommendation is to let the 'experts' decide. Proponents of contingent valuation acknowledge that poorly designed and implemented studies have produced problematic results that should not be used. However, they believe that these problems are not inherent to the method and are not found in the better studies (Hanemann 1994). They argue that without CV it is not possible to conduct a comprehensive economic analysis of many environmental issues and that the public's preferences concerning the provision and payment for public goods cannot be adequately represented to policymakers.

See also: Consumer Economics; Consumption, Sociology of; Ecological Economics; Environmental Adaptation and Adjustments; Environmental Economics; Public Goods: International

Bibliography

Arrow K, Solow R, Portney P R, Learner E E, Radner R, Schuman H 1993 Report of the NOAA Panel on contingent valuation. *Federal Register* **58**: 4601–14

Bateman I J, Willis K G (eds.) 1999 *Valuing Environmental Preferences: Theory and Practice of the Contingent Valuation Method in the US, EU, and Developing Countries*. Oxford University Press, Oxford, UK

Bishop R C, Heberlein T A 1979 Measuring values of extra market goods. *American Journal Agricultural Economics* **61**: 926–30

Carson R T, Flores N E, Martin K M, Wright J L 1996 Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods. *Land Economics* **72**: 80–99

Carson R T, Mitchell R C, Hanemann W M, Kopp R J, Presser S, Ruud P A 1992 *A Contingent Valuation Study of Lost Passive Use Values Resulting From the Exxon Valdez Oil Spill*. Alaska Attorney General's Office, Juneau, AL

Ciriacy-Wantrup S V 1947 Capital returns from soil-conservation practices. *Journal Farm Economics* **29**: 1181–96

Davis R K 1963 Recreation planning as an economic problem. *Natural Resources Journal* **3**: 239–49

Diamond P A, Hausman J A 1994 Contingent valuation: Is some number better than no number? *Journal of Economic Perspectives* **8**(4): 45–64

Freeman A M 1993 *The Measurement of Environmental and Resource Values: Theory and Methods*. Resources for the Future, Washington, DC

Hanemann W M 1994 Valuing the environment through contingent valuation. *Journal of Economic Perspectives* **8**(4): 19–43

Hausman J A (ed.) 1993 *Contingent Valuation: A Critical Assessment*. North-Holland, Amsterdam

Kahneman D, Knetsch J L 1992 Valuing public goods: The purchase of moral satisfaction. *Journal Environmental Economics and Management* **22**: 57–70

Krutilla J V 1967 Conservation reconsidered. *American Economic Review* **57**: 777–86

Mitchell R C, Carson R T 1989 *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Johns Hopkins University Press, Baltimore, MD

Portney P R 1994 The contingent valuation debate: Why economists should care. *Journal of Economic Perspectives* **8**(4): 3–17

Randall A, Ives B, Eastman C 1974 Bidding games for the valuation of aesthetic environmental improvements. *Journal of Environmental Economics and Management* **1**: 132–49

R. T. Carson

Respiratory Disorders: Psychosocial Aspects

Man can do without food for weeks, without water for a few days. A few minutes without air, however, means death. Experiencing shortness of breath is associated with strong feelings of anxiety and panic. In this article, the three most prevalent respiratory disorders will be discussed, from a behavioral and social science perspective.

1. Asthma, Chronic Bronchitis, Pulmonary Emphysema

In order to understand the relevance of the contribution of behavioral and social sciences to the study of patients with respiratory disorders, in this paragraph definitions of asthma, chronic bronchitis, and emphysema, and risk factors, epidemiology, and medical management will be discussed briefly first.

Asthma is derived from Greek, meaning 'panting or distressed breathing.' A universally accepted definition of asthma is not available, reflecting changing views over time in the area of pulmonology and gaps in the

knowledge about the precise nature of asthma. Learned societies in the US, the UK, and in other countries proposed various definitions in an attempt to standardize epidemiological studies and diagnostic and therapeutic medical strategies. One of the most comprehensive definitions of asthma is given by the American Thoracic Society: Asthma is a clinical syndrome characterized by increased responsiveness of the tracheobronchial tree to a variety of stimuli. The major symptoms of asthma are paroxysms of dyspnea, wheezing, and cough, which may vary from mild to almost undetectable, to severe and unremitting (status asthmaticus)' (ATS 1987, http://www.ginasthma.com). Three characteristics stand out: Asthma is defined as reversible, intermittent, and variable. The episodes of airways obstruction causing dyspnea (shortness of breath) either disappear (are reversible) spontaneously or as a consequence of adequate medical, pharmacological treatment. Attacks of asthma vary in their occurrence over a certain period of time—some patients experience asthma attacks almost daily, some only once a year—and the severity of the episodes of shortness of breath varies within and between patients.

Chronic bronchitis is defined in behavioral terms: 'The presence of chronic productive cough for three months in each of two successive years in a patient in whom other causes of chronic cough have been excluded' (ATS 1995). Patients who fulfill criteria for the diagnosis of chronic bronchitis cough frequently and produce sputum over extended periods of time, and may be short of breath. In contrast to asthma, chronic bronchitis is an irreversible medical condition. Eliminating the most important risk factor—smoking tobacco—reduces the symptoms and severity of the disorder somewhat without leading to a complete reversal of the condition of the lungs and their airways.

Pulmonary emphysema is defined in terms of anatomic pathology as 'abnormal permanent enlargement of the airspaces distal to the terminal bronchioles, accompanied by destruction of their walls and without obvious fibrosis' (ATS 1995). In patients with emphysema continuous shortness of breath is the most important symptom, almost always associated with significant impairments in the ability to perform activities of daily life. Here, as is the case in patients with chronic bronchitis, smoking tobacco is the major cause of the disorder.

Over the past decades, pulmonologists, epidemiologists, researchers in related medical domains, and behavioral and social scientists viewed asthma, chronic bronchitis, and emphysema as three related disorders. Despite the differences between the three disorders, the rather similar consequences (shortness of breath, coughing, producing sputum) formed the basis for this 'lumping view.' More recently, the advocates of a 'splitting view' dominate the scientific arena. Differences in the views on causation and pathophysiological bases of the disorders are responsible for this change (Vermeire and Pride 1991). This is why asthma will be discussed separately from chronic bronchitis and emphysema. These last two disorders usually are subsumed under the heading Chronic Obstructive Pulmonary Disease (COPD), defined as 'a disease state characterized by the presence of airflow obstruction due to chronic bronchitis or emphysema' (ATS 1995, http://www.goldcopd.com).

2. *Epidemiology and Some Medical Characteristics*

2.1 *Asthma*

The prevalence of asthma varies between different studies in different parts of the world because of differences in the definition of asthma. Large differences in data on prevalence of asthma are also attributable to differences in genetic susceptibility and environmental conditions. Asthma has a clear genetic basis: Chances for a child, conceived by a man and a woman who both are asthmatics themselves, of being asthmatic are about 75 percent. Environmental stimuli appear to play a role in worsening asthma symptoms. However, air pollution as such does not appear to be a factor that causes asthma. Interestingly, a recent study in Germany demonstrated that the prevalence of asthma in the more polluted Eastern part of Germany was lower than in the Western, more highly developed, and less polluted, part of the country (Nicolai and Von Mutius 1997). This has led researchers to hypothesize about the 'hygiene effect': Having young children growing up in a relatively unpolluted environment, with a high level of sanitation (i.e., low chance of exposure to germs causing flu and colds) predisposes a child to a higher chance of being sensitive to relatively low levels of allergens and respiratory infections, increasing their chances to develop asthma.

The prevalence of asthma is about 7 percent in the US, about 15 percent in the UK, about 1 percent among Inuits, and about 30 percent among inhabitants of Tristan da Cunha, an island in the southern part of the Atlantic ocean between South America and Africa (Gergen and Mitchell 2001). The prevalence of asthma is increasing, especially in industrialized countries. The reasons for this are largely unknown. The increased attention by physicians for a possible diagnosis of asthma in a patient with respiratory problems probably is a contributing factor in the increase of asthma prevalence: Asthma used to be underdiagnosed and undertreated (Kaptein et al. 1987). In young children, boys outnumber girls in prevalence of asthma in a ratio of about 2:1. In adult patients, hardly any sex ratio differences are observable in the prevalence of asthma. Contrary to popular beliefs, one hardly ever outgrows one's asthma, 'one only outgrows one's pediatrician.'

2.2 Chronic Obstructive Pulmonary Disease (COPD)

The prevalence of COPD is about 5 percent (ATS 1995). Recent studies predict considerable increases in the prevalence of COPD (Murray and Lopez 1997), with COPD becoming the third leading cause of death in the future. Despite reductions in the number of smokers of tobacco in industrialized countries, the post-World War II generation with their smoking habits of decades, will fall victim to becoming patients with COPD. Also, as the number of women who smoke cigarettes increased substantially after World War II, COPD will no longer be an illness for men only.

Chronic bronchitis has 50 years of age as the average age of onset, emphysema sets in at around 60 years of age. COPD has a SES-gradient, with a lower socioeconomic class dominance. COPD can be labeled as a blue-collar worker disorder due to the fact that poor housing and working conditions contribute to developing COPD.

The great majority, about 80 percent, of patients with respiratory disorders are seen by the general practitioner (or primary care physician, family physician). Based on critical views and evidence on the quality of care provided by these physicians to patients with asthma in particular, considerable effort has been devoted to improving this quality. In many countries, societies of general practitioners developed guidelines on the optimal management of patients with asthma. In these guidelines early diagnosis, aggressive medical treatment, and a great emphasis on teaching patients self-management skills are the cornerstones of modern management of asthma. Similar guidelines have been developed for the management of patients with COPD in a primary care setting, with an emphasis on stopping smoking techniques.

The remaining 20 percent of patients with respiratory disorders are seen by medical specialists: pulmonary physicians, allergologists, pediatricians. Only a very small percentage of patients is being managed on an in-patient basis—for status asthmaticus in patients with asthma, for severe respiratory problems in patients with COPD. In various countries in the world, specialized long-stay respiratory clinics provide care for those patients with asthma and/or COPD in whom the intractability combines with severe psychological and social problems.

3. Asthma and Psychology

Looking back over some 100 years of behavioral and social science research on patients with asthma, one cannot help but be impressed by the gains that have been made, gauged in terms of behavioral outcomes for these patients. Three periods in this line of research and clinical care can be discerned.

3.1 Psychosomatic Views on Asthma

Between World War I and World War II, psychosomatic views on various psychosomatic disorders and illnesses represented an influential line of research of clinical care. Given the unknown etiology of a number of major diseases (e.g., hypertension, diabetes mellitus, and asthma), theories were developed in which it was postulated that psychological factors played a substantial role in causing a number of specific illnesses. With regard to asthma, a combination of three factors was supposed to be a necessary and sufficient condition: (a) a 'weak spot' in the body of the patient, (b) a characteristic psychological profile of the patient, and (c) a situation eliciting the illness.

Patients with asthma were supposed to have a respiratory system which was sensitive to disruptions. In addition, due to faulty mother–child interaction in the early years of life, a specific personality profile was assumed to be established in the patient: A combination of excessive love of the mother for the child ('smothering with kisses') and an unconscious wish of the mother to reject her child, was supposed to render the child sensitive to asthma. In the event of a stressful life event (e.g., being separated from the mother, experiencing social conflicts), an asthma attack would happen.

Empirical evidence for these psychosomatic views is lacking. No study identified the supposed mother–child psychopathology put forward by the theory. Psychosomatic therapy (i.e., psychodynamic interventions) failed to demonstrate any positive effect. It is safe to state that the highly selected sample of patients with asthma seen by the theorists (psychiatrists and specialists in internal medicine) represented patients who exhibited psychosocial problems as a consequence of having to cope with the adaptive tasks of having severe, brittle asthma.

3.2 Psychological Consequences of Asthma

The second theoretical and empirical approach to asthma from a behavioral and social science perspective focused on the consequences of asthma on the daily lives of afflicted patients: their quality of life. Using disease-specific questionnaires with adequate psychometric properties, limitations in functional, psychological, and social domains were identified. Asthma patients experience excessive absenteeism from school and work, heightened levels of anxiety and depression, and feelings of social isolation (Maillé et al. 1996).

The first studies on the effects of psychological interventions in order to reduce the limitations caused by asthma adopted a learning theory perspective: relaxation training, biofeedback, and systematic desensitization. In these studies, pulmonary function was the major outcome measure. The effects of

behavioral interventions turned out to be limited, however (Lehrer 1998).

The research performed by a group of psychologists in Denver, Colorado, USA, in the 1970s and 1980s set the scene for the third phase in psychological research on patients with asthma. Kinsman et al. (1982) demonstrated how the affective and cognitive responses of asthma patients to their illness and symptoms predicted important outcome measures such as length of hospitalization and rehospitalization, irrespective of the medical severity of the respiratory condition.

3.3 Self-management Training

A logical next step, given the evidence for the impact of the psychological response (coping behavior) of patients to their asthma, pertains to teaching asthma patients to optimize their handling, or self-management, of the disorder. Creer (1979) represents the first psychologist who studied the effects of this type of intervention in children and adults. In self-management training, patients with asthma are taught skills to perceive the first signs of periods of shortness of breath, take appropriate action to prevent or control episodes of shortness of breath, and manage bouts of asthma if they occur. Reviews of the effects of self-management training report reductions in symptom frequency and severity, use of medical resources, and fewer disruptions of daily life (e.g., absenteeism from school or work, levels of anxiety) (Clark and Gong 2000).

Current research on asthma by psychologists focuses on refining self-management training, symptom perception, and the effects of emotional expression on pulmonary function (Smyth et al. 1999).

4. COPD and Psychology

In contrast to the considerable amount of research performed by psychologists and social scientists on patients with asthma, patients with COPD have been the subject of research from this perspective only relatively recently. There are at least two reasons for this. The impact of smoking tobacco, the major cause of COPD, has become clear only after World War II; psychologists and social scientists joined medical professionals in their therapeutic pessimism for decades. The first empirical study on the psychological consequences of COPD was published only in 1961.

No psychosomatic theories were developed on patients with COPD. Studies on the neuropsychological consequences demonstrated that severe grades of COPD negatively affect cognitive functioning of patients. Depression, fatigue, and feelings of social isolation are the dominant psychological and social consequences of COPD. Restrictions in daily activities, absenteeism from work, early retirement, and invalidity are major functional consequences of COPD. The quality of life, therefore, of patients with COPD is significantly lower than in healthy persons (Kaptein and Dekker 2000).

Psychologists contribute to the care for patients with COPD in two areas: pulmonary rehabilitation programs and self-management training. Pulmonary rehabilitation pertains to 'an art of medical practice wherein an individually tailored multidisciplinary program is formulated which through accurate diagnosis, therapy, emotional support, and education stabilizes or reverses both the physio- and psychopathology of pulmonary diseases and attempts to return the patient to the highest possible functional capacity allowed by his pulmonary handicap and overall life situation' (Ries 1997). Physical therapy, breathing exercises and training, smoking cessation techniques, graded exercise aimed at improving ambulation, and coping skills training usually are parts of pulmonary rehabilitation programs. Reviews and meta-analyses of the effects of these programs indicate that use of health services, limitations in daily activities, and psychological problems tend to be reduced to some degree (Kaplan and Ries 2001). High dropout percentages, however, are the rule. Identifying the separate contributions of the various components in pulmonary rehabilitation programs to various outcome measures is not feasible, however.

The current emphasis in health psychology research in COPD focuses on teaching patients self-management skills. Cognitive behavior modification techniques are used to alter nonadaptive cognitions and behavior. In a study of patients' pessimistic self-statements ('I can't walk very far without getting short of breath, so what's the use?') were addressed and altered into more active ones ('This walking is uncomfortable, but I can handle it. Soon I will be able to walk farther.'). Results showed that levels of exercise tolerance and quality of life increased in comparison to various control conditions (Atkins et al. 1984).

Increasingly, guidelines on the medical management of patients with COPD include results of psychological research on self-management (ATS 1995). This development is positive: it reflects how pulmonary physicians acknowledge the relevance of topics such as compliance or adherence, and incorporating self-management into clinical care. Patient empowerment is instrumental for the wellbeing of patients with this chronic respiratory disorder.

Future research efforts will strengthen the position of patients with COPD. Health psychologists who work in the area of COPD face many challenges. Smoking cessation is a primary subject for health psychologists and other social scientists, given the psychological and societal determinants of smoking tobacco in particular. Simplifying self-management skills and incorporating the teaching and uptake of these skills into medical care are additional challenges.

COPD is an under-funded medical disorder (Gross et al. 1999). Future research both from a medical and a psychological perspective will help in preventing COPD to develop, and in improving quality of life in patients with COPD.

See also: Health Education and Health Promotion; Smoking Prevention and Smoking Cessation

Bibliography

Atkins C J, Kaplan R M, Timms R M, Reinsch S, Lofback K 1984 Behavioral exercise programs in the management of chronic obstructive pulmonary disease. *Journal of Consulting and Clinical Psychology* **52**: 591–603

ATS 1987 Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease (COPD) and asthma. *American Review of Respiratory Disease* **136**: 225–44

ATS 1995 Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine* **152**: S77–S120

Clark N M, Gong M 2000 Management of chronic disease by practitioners and patients: are we teaching the wrong thing? *British Medical Journal* **320**: 572–5

Creer T L 1979 *Asthma Therapy: A Behavioral Health Care System for Respiratory Disorders*. Springer, New York

Gergen P J, Mitchell H E 2001 Epidemiology of asthma. In: Kaptein A A, Creer T L (eds.) *Respiratory Medicine and Behavioral Research*. Routledge, London

Gross C P, Anderson G F, Powe N R 1999 The relation between funding by the National Institutes of Health and the burden of disease. *The New England Journal of Medicine* **340**: 1881–7

Kaplan R M, Ries A L 2001 COPD: Behavioral assessment and treatment. In: Kaptein A A, Creer T L (eds.) *Respiratory Medicine and Behavioral Research*. Routledge, London

Kaptein A A, Dekker F W 2000 Psychosocial Support. *European Respiratory Monograph* **13**: 58–69

Kaptein A A, Dekker F W, Gill K, Van der Waart M A C 1987 Undertreatment of asthma in Dutch general practice. *Family Practice* **4**: 219–25

Kinsman R A, Dirks J F, Jones N F 1982 Psychomaintenance of chronic physical illness. In: Millon T, Green C, Meagher R (eds.) *Handbook of Clinical Health Psychology*. Plenum, New York, pp. 435–66

Lehrer P M 1998 Emotionally triggered asthma: a review of research literature and some hypotheses for self-regulation therapies. *Applied Psychophysiology and Biofeedback* **23**: 13–41

Maillé A R, Kaptein A A, de Haes J C J M, Everaerd W T A M 1996 Assessing quality of life in chronic non-specific lung disease: A review of empirical studies published between 1980 and 1994. *Quality of Life Research* **5**: 287–301

Murray C J L, Lopez A D 1997 Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. *Lancet* **349**: 1498–1504

Nicolai T, Von Mutius E 1997 Pollution and the development of allergy: The East and West Germany story. *Archives of Toxicology* **19** (suppl.): 201–6

Ries A L 1997 Pulmonary rehabilitation—Joined ACCP/AACVPR evidence-based guidelines. *Chest* **112**: 1363–96

Smyth J M, Stone A A, Hurewitz A, Kaell A 1999 Effects of writing about stressful experiences on symptom reduction in patients with asthma or rheumatoid arthritis. *Journal of the American Medical Association* **281**: 1304–9

Vermeire P A, Pride N B 1991 A 'splitting' look at chronic nonspecific lung disease (CNSLD): Common features but diverse pathogenesis. *European Respiratory Journal* **4**: 490–6

A. A. Kaptein

Responsibility: Philosophical Aspects

1. *The Concept of Responsibility*

The focus of this entry is moral responsibility. Although it begins with an examination of the concepts of causal responsibility and of responsibility as duty, the discussion in these sections is intended to orient the reader to the distinctive character of the concept of moral responsibility and to attendant philosophical theories about what is meant when moral responsibility is ascribed to a human agent for some act or state of affairs. Attributing responsibility to an agent as to a cause is logically and conceptually distinct from claiming that the person is morally responsible for causing an event. Equally, discerning those tasks for which a person has a responsibility or a duty to perform is different from claiming that they bear moral responsibility for the performance of those tasks. In short, the *concept* of responsibility has a number of distinct applications, the analyses of which point to the variety of *conceptions* of responsibility.

1.1 *Causal Responsibility*

One interested in action theory may investigate *causal* responsibility. Thus, in claiming 'Alice is responsible for spilling the coffee on the floor,' we may be saying that Alice spilled the coffee as a result of some action on her part. In such a causal context, responsibility ascriptions credit a person or an entity with a role in bringing about an event or state of affairs, and nothing more. Assigning causal responsibility consists in identifying a salient factor that must actually operate in order for an event to transpire—its cause—and not in evaluating the moral quality of an event with which an actor is instrumentally associated.

Moral responsibility, and responsible agency, cannot be identified with causal responsibility. This is true for at least three reasons. First, that someone or something is causally responsible for some event or state of affairs is not sufficient for the thing being morally responsible for the event. The kind of responsibility attributed to one who is identified as

causally responsible for something is not a sufficient basis for moral evaluation. That the agent is causally responsible is not, in and of itself, a reason to judge the agent a good or bad, praiseworthy or blameworthy, person as a result. But judgments of moral responsibility are typically of this sort.

Second, a person need not be causally responsible for all those acts they are held morally responsible for—causal responsibility is not necessary for moral responsibility. For example, parents are vicariously responsible, legally and, although this is controversial, perhaps morally, for at least some of the acts of their nonadult children.

Third, that someone or something is causally responsible for some event or state of affairs is not sufficient for the cause to be an agent who can be held morally responsible. Because causal responsibility can be ascribed to events, things, nonhuman animals, and to irrational or underdeveloped humans as well as to fully mature, rational persons, it does not single out a class of things that might properly be described as moral *agents*, or *members of a moral community*. When we hold a person such as Alice morally responsible for something, it is in light of certain expectations we have of Alice, given a maturity we assume of her, given what we have reason to believe she knew herself to be doing, and given the circumstances that led her to act. But the conditions for responsible agency met by Alice, and considered more fully below, are not met by very young children, or animals, or trees. Thus, while we might claim that faulty construction is to blame for the collapse of a building, what we mean is that a salient cause of the collapse was poor construction. We might allege that the baby was causally responsible for the broken lamp. But neither the baby nor the faulty construction are morally responsible for the states they cause because neither are moral agents.

In sum, an account of agency must make sense of moral approval or disapproval. Even if the act for which we deem a person responsible is morally neutral—is neither praiseworthy nor blameworthy—we appraise the moral quality of the person who acts, and not simply the state of affairs the person initiates.

1.2 Duty and Obligation

By contrast, in describing a person as 'responsible' for some act or state of affairs, we may be expressing the belief that the person has a duty. Here, the judgment 'X is responsible for Y' is intended as a statement to the effect that X has certain responsibilities to others: tasks, duties, or obligations pertaining to Y, usually in light of X's position in society or role. Claiming that a person is responsible for Y in this sense is to credit responsibility *prospectively*; it is to look forward to what the person is duty bound or obligated to do. This idea of having responsibilities is analogous to Kurt Baier's notion of 'task responsibility,' or the 'forward-looking aspect' of agent responsibility. In its forward-looking sense, agent responsibility consists in being socially entrusted, on either customary, legal, or moral grounds, with a specified event, state of affairs, or duty.

Thus, parents are responsible for overseeing the well-being of their children, adult wage earners are responsible for paying taxes, attorneys are responsible for respecting the confidentiality of their clients, and so forth. In this forward-looking sense, the claim that Alice is responsible for giving Bernie the medicine would mean, at a minimum, that Alice has a duty to give Bernie medicine. The sentence is then ambiguous between (a) the sense in which it attributes a prospective duty to Alice; (b) the sense in which it attributes mere causal responsibility to Alice for Bernie's having been given the medicine, and (c) the sense in which it attributes moral responsibility and implies, perhaps, praise or blame.

In the latter sense, to be morally responsible is not a matter of having or owing an obligation to another. A person may be morally responsible for those things they have a moral obligation to do, such as caring for their children. But the things for which a person can be held morally responsible in the sense of being praiseworthy or blameworthy is not restricted to things they had an obligation or a responsibility to do. I might be held responsible for my slovenly habits, or for my acts of bravery, where neither of these attends the discharging of an obligation. Equally, the fact that an act is one I am obliged to perform does not decide whether I am morally accountable for performing it or for failing to do so.

Once we have set aside causal responsibility and responsibilities in the sense of duties, we are left with the idea of responsible agency and the idea of responsibility for action. What we are interested in is the idea of *being* responsible—both being a responsible person and being responsible for an act.

2. Being a Responsible Agent and Being Held Responsible

2.1 The Conditions for Responsible Agency

In order to be morally responsible for an act, or trait of character, a person must be a responsible agent. Responsible agency typically requires that an individual satisfy certain *epistemic* conditions and certain social and psychological conditions of *control*.

The epistemic conditions guarantee that the responsible agent is self-aware, that they are rational, that they are not ignorant of the circumstances in which they act, that they are cognizant of and able to act within established moral guidelines, and that they

are responsive to reasons to adjust or amend their behavior in light of these guidelines. In order to be held responsible, the moral agent must know that doing a particular act (or an act of a given type) or cultivating a particular trait of character (say, jealousy, rage, bigotry) is right or is wrong. The moral agent may be held responsible if, suffering from none of the conditions that exempt a person from responsible agency, they should have known the nature of the act or trait, and could have been motivated by the relevant moral guidelines. A person's ignorance in this case might be called culpable. By contrast, a person should not be held responsible, if they did not know, and should not be expected to know, that the act is of an undesirable moral quality.

The control conditions ensure that the agent acts freely. The morally responsible agent is not hostage to debilitating and uncontrollable neuroses of a sort that incapacitate their resolve; that is, they can control the manifestations of neuroses by medical means or by strength of will. A person who is prone to neurosis of a certain type or scope, or who finds themself the victim of coercion or manipulation, may be a responsible agent but one whom we excuse from responsibility for specific acts when it is plain that on a particular occasion coercive or manipulative elements played a decisive and ineliminable role in action.

Satisfaction of the control conditions for responsible agency ensure that the agent, rather than a mechanism foreign to them, actually guides or actually executes the agent's actions and those traits of character for which the agent is morally responsible. A mechanism of action is foreign to a person when it is one to which the person does not assent as a force by which they wish to be moved. Although the responsible person may lack the power to alter the mechanism that leads them to act—to do otherwise than they in fact do—they are moved by their own lights. At a minimum, the morally responsible agent is capable of deliberate, intentional behavior.

Absent some of the epistemic and control conditions, a person will be a less likely candidate for responsible agency. The satisfaction of some set of these will mean that the entity is a responsible agent. There is obviously a threshold that decides when a person is a responsible agent, but it is not within the scope of this discussion to explore the specifics of the threshold. Let it suffice to say that the responsible agent is someone who acts knowingly and freely, and the actions and characteristics for which they are responsible are 'their own.'

2.2 Holding Responsible and Being Responsible

Generally speaking, when we describe someone as a morally responsible agent, we mean that the person is someone whom we recognize as bearing characteristics of the sort that make them a member of the moral community. They are someone with whom morally relevant relationships are possible, and they are a fitting subject of moral appraisal. But there are different theories about the analysis of judgments that a person is morally responsible for their character, their choices, their actions or omissions. Three accounts, in particular, bear mention.

One possibility is that we might simply be signaling that an actor, having met certain conditions for responsible agency, is a suitable candidate for certain kinds of corresponding regard and treatment forthcoming on the part of others and, perhaps, on the part of himself. This approach has been coined the reactive attitudes analysis of responsibility ascriptions. The idea is that when we say a person is responsible we mean that the person is liable, through their behavior, to a response on the part of others marked by attitudes such as gratitude and resentment, and that the person is an appropriate object of such attitudes in light of certain conditions for agency that they have met. On this analysis, the claim 'X is responsible for Y' (or, for Y-ing) will amount to the conjunction of the claims 'X did (or is) Y,' 'X possesses the qualities of a responsible agent,' and 'X is rightly subject to the favorable or unfavorable responses of others as a result.' The reactive attitudes analysis offers an account of the emotional dynamics at play where decisions about responsibility occur. That a person is appropriately subject to certain attitudes and practices gives us reason to consider them a free, responsible agent, and as good or bad, praiseworthy or blameworthy, and so on.

A second possibility is that, in judging a person responsible, we may be saying the person is accountable for their behavior, in the sense that we think it appropriate that the person offer an explanation of their intentions or beliefs about their behavior. That is, we may be focusing on the phenomenon of holding an agent answerable for their actions. Construing judgments about moral responsibility as judgments about accountability is appropriate for a number of reasons. One reason is that holding persons responsible is part of a social practice of ascribing responsibility, first by making clear any moral expectations we have of persons vis-à-vis behavior and character, and then by having in place certain social instruments that ensure adequate redress where necessary. Making moral expectations clear presumes the existence of a shared set of beliefs, and a system of norms against which actions and traits of character are assessed.

A third possibility is that judging a person responsible for some act, omission, state of affairs, or trait of character is essentially a statement of attribution. The philosopher Gary Watson advocates this perspective, which he labels the *aretaic* face of responsibility, for its focus on ethical appraisal of conduct that can be attributed or imputed to an individual as its agent and thus as the author or adopter of ends. Because these ends are self-adopted,

acting upon them discloses the quality and character of the agent's life. The attributability approach does not merely assign causal responsibility, for the focus is on the behavior of a human agent, his/her character, and his/her ends—the object leading to action. Acts of attributing responsibility on this account differ from both the reactive attitudes and accountability analyses. Attribution need not involve the registration of emotions such as resentment or gratitude, nor need it be true that an explanation of the act in question is expected of the actor.

Two varieties of cases illustrate the differences among these three analyses. First consider examples of extreme and, some would contend, irremediable evil. If we consider the behavior of someone such as Hitler, it may well appear that he is morally 'beyond the pale' so that none of the standard reactive attitudes is an appropriate response to, for instance, his 'Final Solution.' Hitler, and like-minded persons, are seemingly incapacitated for ordinary adult interpersonal relationships. Such persons simply will not be responsive to the demands of membership in our moral community. At the same time, the heinous acts they have committed are properly attributed to them as perpetrators. The attributability account allows for Hitler, and similarly sociopathic or psychopathic persons, to be deemed responsible for their immoral behavior, yet permits the hesitation some might have in regarding such actors as moral interlocutors, as the accountability approach demands, or as appropriate objects of any of the ordinary moral attitudes.

A second, ostensive, merit of the attributability approach is that it enables us to credit persons with responsibility for matters that are essentially private, matters that are the affair of none but the actor themself. It seems unreasonable to reply to my slovenly garb, or foolish financial practice, with resentment or indignation, just as it seems unreasonable to expect that I account for these practices if they affect no one but myself. Where these practices reveal something about ends I have chosen for myself, they speak to my character—they express, as Watson claims, my own evaluative commitments and my identity as an agent. Thus, on the attributability account, they can be attributed to me as practices for which I bear responsibility.

3. Responsibility and Free Will

3.1 Causal Determinism and Moral Responsibility

An adequate discussion of the central problems faced by any account of moral responsibility will include the problem causal determinism poses to the idea that we have freedom of will. In brief, the problem is that if our present choices and conduct are causally determined by past events and states, then it appears that we do not have any freedom with respect to the performance or occurrence of our present choices and conduct. The past is something over which we have no control. And it is equally true that we have no control over the laws of nature operative in the world. Together, these claims suggest that we have no control over the present.

It is a matter of lively debate among philosophers whether or not the truth, or the likelihood, of causal determinism would affect our status as responsible agents. Some philosophers argue for the incompatibility of determinism and responsibility; if either obtains, the other does not. We cannot be both determined and responsible for our conduct. Others argue that the conditions for responsible agency are independent of determinism. On this view, the truth or plausibility of determinism aside, we can control our conduct in a manner sufficient for responsible agency. At the very least, to deny responsibility of persons whom we regard as participants in our moral framework would be to strain the limits of our natural practices. It is undeniable that persons treat one another as agents, as individuals answerable for their conduct, and that the alternative—that persons cultivate an attitude of detachment from the conduct of their human associates merely because of the truth of determinism—would signal an impoverished state of human interaction that few, if any, could bear.

See also: Autonomy, Philosophy of; Ethics and Values; Free Will and Action; Freedom/Liberty: Impact on the Social Sciences; Identity and Identification: Philosophical Aspects; Intentionality and Rationality: A Continental-European Perspective; Intentionality and Rationality: An Analytic Perspective

Bibliography

Baier K 1972 Types of responsibility. In: French P A (ed.) *Individual and Collective Responsibility*. Schenkman Publishing, Cambridge, MA

Frankfurt H 1988a Alternate possibilities and moral responsibility. In: *The Importance of What We Care About*. Cambridge University Press, Cambridge, UK, pp. 1–10

Fischer J M, Ravizza M 1998 *Responsibility and Control*. Cambridge University Press, Cambridge, UK

Frankfurt H 1988b Freedom of the will and the concept of a person. In: *The Importance of What We Care About*. Cambridge University Press, Cambridge, UK, pp. 11–25

Haji I 1998 *Moral Appraisability: Puzzles, Proposals, and Perplexities*. Oxford University Press, New York

Harman G 1975 Moral relativism defended. *The Philosophical Review* **84**: 3–22

Oshana M 1997 Ascriptions of responsibility. *American Philosophical Quarterly* **34**: 71–83

Pettit P, Smith M 1996 Freedom in belief and desire. *The Journal of Philosophy* **93**: 429–49

Strawson P 1993 Freedom and resentment. In: Fischer J M, Ravizza M (eds.) *Perspectives on Moral Responsibility*. Cornell University Press, Ithaca, NY, pp. 45–66
Van Inwagen P 1983 *An Essay on Free Will*. Clarendon Press, Oxford, UK
Wallace R J 1994 *Responsibility and the Moral Sentiments*. Harvard University Press, Cambridge, MA
Watson G 1996 Two faces of responsibility. *Philosophical Topics* **24**: 227–48
Wolf S 1990 *Freedom Within Reason*. Oxford University Press, New York
Zimmerman M J 1988 *An Essay on Moral Responsibility*. Rowman and Littlefield, Totowa, NJ

M. Oshana

Retail Trade

The term 'retail geography' describes the spatial structure of retailing, especially its transformation. The dominant methodology of the retail geography literature combines analyses of spatial demand patterns, producer behavior, and travel patterns, within the broader spatial patterns of economic activities. The geography of retailing is related to 'marketing geography' and the two terms are used somewhat interchangeably, although marketing geography encompasses the geographic patterns of consumption more broadly.

Explanations of the geography of retailing have traditionally been based on the conceptual frameworks of location theory (see *Location Theory*), especially the classical models of central place theory (Christaller 1933, Lösch 1941). Two themes, differentiated by the type of locational system investigated, developed early. The first studies the spatial pattern of periodic markets. The second focuses on regional systems of permanent retail outlets. A corollary of the difference between the two themes is their different regional foci. Studies of periodic markets are mostly centered in Asia, Africa, and Latin America, while the second theme has more of an Anglo-American focus. The third theme in the retail geography literature presents analytical methods of studying travel patterns to retail stores and of determining optimal locations.

1. *The Geography of Periodic Markets*

Periodic markets, which continue to be an important element of rural retail trade in many parts of the world, meet periodically at specific locations with set schedules, separated by market-less days or by days with limited trading activities. Such periodic markets are the basic component of the spatial structure of rural retailing in Asia, Africa, and Latin America. Skinner (1964) posited that periodic markets correspond to the lowest levels of the hierarchy of central places that serve the need for goods and services in an economy. The markets associated with each level possess distinctive size, functionality, and periodicity characteristics (Berry and Parr 1988). Central place theory is a leitmotif in most studies of the geography of periodic markets.

1.1 *Models of Periodic Market Systems*

Stine (1962) was the first to develop a theoretical model to explain why markets are periodic. In low-density areas the level of demand is often too low to cover the fixed cost of a permanent store. Traders can overcome this constraint by visiting several marketplaces in a regular, periodic sequence to aggregate demand from a wider geographical area. Hay (1977) and Ghosh (1982), among others, extended Stine's analyses. They outline the conditions under which a trader will adopt one of the following strategies—mobile or itinerant trading, part-time trading from a fixed location, or full-time trading from a fixed location. Traders may choose an itinerant strategy even when the expected profit at a fixed location exceeds the cost of operating a permanent store, if the incremental net earnings from a new location are greater than the sum of relocation and overhead costs. Similarly, mobility may allow traders to reduce prices and forestall competition. In general, the lower the relocation and overhead costs in relation to demand, the greater the likelihood of mobile systems. This is one reason why in many rural areas one tends to find large numbers of itinerant traders in markets that do not have a single permanent outlet.

1.2 *Spatio-temporal Organization of Markets*

An important geographic characteristic of periodic marketing systems is their integration of both the spatial and temporal dimensions of demand. Consumers perceive the adjacent marketplaces held on consecutive days—the market ring—as belonging to a particular spatio-temporal set of markets that serve their needs. According to a number of observers, the location and timing of markets have evolved to maximize access to consumers within a specified interpurchase interval. This leads to an inverse relationship between the location of two markets and the time interval between their market meetings. Markets that meet simultaneously or are separated by only a short time interval should be more geographically dispersed (Fagerlund and Smith 1970). A counter-hypothesis is that itinerant traders develop a sequence that minimizes their travel costs (Hill and Smith 1972). Empirical studies provide only partial support for the hypotheses.

2. The Morphology of Urban Retail Outlets

The organization of retail centers within cities follows a hierarchical pattern characterized by greater functional complexity of centers as one moves to higher levels of the hierarchy. The geographic reach of the center expands as one moves up the hierarchy due to its centrality, and retail businesses locate in different levels of the hierarchy depending upon the sizes of the market area required to generate adequate economic returns. The result is a hierarchical structure of retail locations similar to the tenets of central place theory. Much of the early literature on retail geography comprises empirical studies of urban retail hierarchies in North America and Europe conducted to test central place theory. Later the emphasis shifted to investigating the dispersion of retailing from urban areas by combining the analyses of spatial demand patterns with that of the changes in the social and economic conditions as well as changes in the organizational structure of the retail sector.

2.1 The Transformation of Urban Retail Geography

The dominant feature of the transformation of the post-World War II retail geography in North America and Europe is the shift of retail activities from city centers to suburbs. Before World War II, the central business district (CBD) ranked as the pre-eminent metropolitan shopping district, offering a vast array of goods and services and attracting shoppers from throughout the metropolitan region. That premier position started to fall in the 1950s and 1960s in large urban areas, and later in smaller cities too. While retail activity declined in the metropolitan core, however, it grew rapidly on the suburban periphery both in North America and Europe, reflecting the spatial redistribution of population and income in metropolitan areas (Davies and Rogers 1984). The suburb's share of retail sales increased significantly, and major shopping centers started to rival the CBD in retail sales volume.

The decentralization of sales was linked closely to the emergence of new retail formats. Shopping centers, housing diverse retail stores under one roof, represented a dramatic departure from earlier retail formats. Unlike the traditional downtown shopping district, they were planned, integrated complexes developed and marketed as a unit (Dawson and Lord 1985). Initially suburban shopping was concentrated along highways in strip malls and in neighborhood convenience centers. This was followed by the development of large, regional shopping centers, which became the focal point of the metropolitan retail structure, drawing shoppers from a wide geographical area. Thus, the new geography of retailing is much more dispersed than in the past, comprising a well-developed retail network focused on various levels of regional and neighborhood shopping centers and shopping districts.

2.2 The Context of Change

The suburbanization of population and the evolving transportation infrastructure provided the motive force for the spatial dispersion of retailing. But institutional, financial, and governmental policies also shaped this new geography. The establishment of shopping centers was greatly encouraged by suburban municipal governments, who viewed shopping centers as an important source of tax revenues. Small municipalities on the outskirts of larger towns encouraged the establishment of shopping centers within their jurisdictions to attract shoppers from the larger towns, and, thereby, increase tax revenues. The growth of shopping centers was also facilitated by easy access to financing and financial arrangements that greatly reduced the developer's risk. The consolidation of the retail industry and the growth-oriented philosophies of department and specialty store chains further fueled the expansion.

While local governments in suburban communities actively encouraged the growth of shopping centers, their counterparts in urban areas viewed them as grave threats. Retailers in older, urban shopping districts faced two competitive disadvantages. First was the lack of population, income, and employment growth. Second was the general deterioration of the shopping environment, which reduced the CBD's ability to attract middle- and upper-income shoppers. Concerned by the erosion of urban retail business and its impact on city finances, a number of governmental programs and policies were initiated to encourage retailing in CBDs. Some of these new downtown retail projects included pedestrian shopping malls, retail facilities built within large downtown office complexes, and anchorless centers housing entertainment, recreation, food, and specialty retailing. They all represented departures from the traditional hierarchical structure of retail organization.

3. Analytical Models of Store Location

3.1 Retail Forecasting Models

The rapid expansion of suburban outlets created the need for a systematic approach to determining store locations, resulting in an array of techniques for site selection and retail sales forecasting. The key to forecasting retail sales is to delineate accurately the store's trade area—the geographic area from which the store draws most of its customers. The foundation of contemporary approaches to delineating trade areas was laid by Applebaum (1968) in his 'customer spotting' method. Huff's spatial interaction model (see

Spatial Interaction Models) extended the approach by looking at the complex interactions within the total system of retail trade areas in a region (Huff 1963). Underlying spatial interaction models is the premise that the share of customers a retail outlet attracts from an area is inversely proportional to distance and directly proportional to the attraction of the store. These models develop an empirical basis to determine the relative impact of distance and attraction on store choice, and conceive trade areas as probabilistic rather than deterministic, with each store in an area having some probability of being patronized. Huff's probabilistic revealed preference approach to modeling store choice was extended with more sophisticated analytical techniques and the definition of a broader set of variables to measure store attractiveness (Ghosh and McLafferty 1987).

3.2 Location-allocation Models

Determining the optimal locations for a network of retail outlets is a complex task, requiring decisions on the number of stores to locate, their locations, and their characteristics. All the decisions must be integrated within a single decision-making framework, taking into consideration the geographical pattern of demand, transportation, and corporate objectives. Location-allocation models systematically evaluate a large number of locations to find sites that optimize objectives such as market share or accessibility, taking into consideration system-wide store-location interactions (Ghosh and McLafferty 1987). Because the models can look at the impact of each store on the entire network of stores, they are ideally suited to analyzing the location decision of retail chains operating multiple outlets in the same market area (Achabal et al. 1982). Location-allocation is a flexible analytical system in that a variety of objective functions and constraints can be incorporated (Ghosh and Harche 1993). The combination of machine-readable economic and geodemographic data, geographical information systems software, and location-allocation models create effective location decision support systems (Birkin et al. 1996).

4. Conclusion

The structure of retail trade, like most economic activities, is shaped to a large extent by geography. Indeed, it is often said that the three secrets to retailing are location, location, and location. Although this oft-repeated adage is somewhat of an exaggeration, even in the era of e-commerce and electronic retailing, the geographical pattern of retail trade is shaped by the relationship between the patterns of consumer demand and store locations, mediated by consumer travel patterns. The literature on retail trade provides frameworks for understanding how these forces, mediated by public and corporate policies, shape the observed pattern of retail trade and presents empirical studies from diverse regions of the world. It also presents analytical methods for locating retail outlets.

See also: Cities, Internal Organization of; Commodity Chains; Consumer Economics; Consumption, Sociology of; International Business; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; International Trade: Geographic Aspects; Market Areas; Market Structure and Performance; Marketing Strategies; Multinational Corporations; Urban Planning: Central City Revitalization

Bibliography

Achabal D D, Gorr W L, Mahajan V 1982 MULTILOC: A multiple store location model. *Journal of Retailing* **58**: 5–25

Applebaum W 1968 *A Guide to Store Location Research, with Emphasis on Supermarkets*. Addison-Wesley, Reading, MA

Berry B J L, Parr J B 1988 *Market Centers and Retail Location: Theory and Applications*. Prentice Hall, Englewood Cliffs, NJ

Birkin M, Clarke G, Clarke M, Wilson A 1996 *Intelligent GIS: Location Decisions and Strategic Planning*. Geoinformation, Cambridge, UK

Christaller W 1933 *Die zentralen Orte in Süddeutschland*. Fisher, Jena, Germany [trans. Baskin C W 1966 *The Central Places of Southern Germany*. Prentice Hall, Englewood Cliffs, NJ]

Davies R L, Rogers D S 1984 *Store Location and Store Assessment Research*. Wiley, Chichester, UK

Dawson J A, Lord J D 1985 *Shopping Centre Development: Policies and Prospects*. Croom Helm, London

Fagerlund V G, Smith R H T 1970 A preliminary map of market periodicities in Ghana. *Journal of Developing Areas* **4**: 333–47

Ghosh A 1982 A model of periodic marketing. *Geographical Analysis* **1**: 155–66

Ghosh A, Harche F 1993 Location-allocation models in the private sector: Progress, problems and prospects. *Location Science* **1**: 81–106

Ghosh A, McLafferty S L 1987 *Location Strategies for Retail and Service Firms*. Lexington Books, Lexington, MA

Hay A 1977 Some alternatives in the economic analysis of periodic marketing. *Geographical Analysis* **9**: 72–9

Hill P, Smith R H T 1972 The spatial and temporal synchronization of periodic markets: Evidence from four emirates in northern Nigeria. *Economic Geography* **48**: 345–55

Huff D L 1963 A probabilistic analysis of shopping center trade areas. *Land Economics* **39**: 81–90

Lösch A 1941 *Die räumliche Ordnung der Wirtschaft* Fisher, Jena, Germany [trans. into English 1944 *The Economics of Location*. Yale University Press, New Haven, CT]

Skinner G W 1964 Marketing and social structure in rural China I. *Journal of Asian Studies* **24**: 3–43

Stine J H 1962 Temporal aspects of tertiary production elements in Korea. In: Pitts F R (ed.) *Urban Systems and Economic Development: Papers and Proceedings*. University of Oregon, Eugene, OR

A. Ghosh

Retirement and Health

1. Retirement in Historical Perspective

Retirement as a structural aspect of the life course is a relatively recent phenomenon, institutionalized in industrialized societies only in the twentieth century. 'Retirement' is typically defined as later life withdrawal from the workforce, often in conjunction with public and/or employer-provided pension benefits. But growing numbers of workers retire from their career jobs and then take on employment in another job and/or unpaid community service. Retirement—in terms of eligibility for a pension—can no longer be equated with a one-way, one-time exit from the workforce or with the cessation of all productive activity. Neither is retirement occurring at one set age (such as 62 or 65). The proliferation of public and private retirement income programs have encouraged a worldwide trend toward progressively earlier retirement from career jobs (Delsen and Reday-Molvey 1996, Guillemard and Rein 1993), and greater heterogeneity in the age and nature of retirement.

2. Linking Health and Retirement

Scholars are only beginning to unravel the health correlates of both the exit from one's 'career' job and the final exit from any paid employment, as well as life after retirement. Three processes link health and retirement: social selection, role loss, and reduction of role strain.

2.1 Social Selection: Health as an Impetus to Retire

Declines in workers' health have been shown to shape their decision-making regarding retirement (Anderson and Burkhauser 1985, Mutchler et al. 1999). The best-laid plans for long-term employment or retirement timing can be destroyed by acute or chronic illness of oneself or one's spouse (or other relatives requiring care), with the onset of poor health promoting unexpected exits from the labor force. Thus, it is crucial that research on the health impacts of retirement appropriately control for preretirement health status.

2.2 Retirement as Role Loss

The second process is that of role loss. Leaving the workforce can have deleterious effects on health, given the corresponding loss of routine, relationships, identity, and purposive activity. Research has documented the importance of role occupancy for both physical and psychological well-being (Moen et al. 2000). Retirement can increase the vulnerability of those in later adulthood, removing them from the mainstream of society.

2.3 Reduction of Role Strain

Role exits such as retirement can also be viewed positively, as when workers leave a demanding or stressful job. In line with this reduction of role strain perspective, those who see their jobs as stressful or unrewarding are likely to leave the labor force early. Jobs that are demanding and provide little sense of control have been shown to be negatively related to health and well-being (e.g., Karasek and Theorell 1990). Moreover, poor health is more likely to encourage those with demanding jobs to retire (Hayward et al. 1989).

3. A Life Course, Role Context Perspective

A life course, role context approach to retirement (Moen et al. 2000, Musick et al. 1999) emphasizes the complexity of this transition. The three processes described above may well be operating at different points in the retirement transition for different subgroups or in particular circumstances. For example, health correlates may be related to the timing of retirement as well as retirees' location in the social structure (gender, race and ethnicity, socioeconomic status), their previous health, and the quality of their prior employment experience. Moreover, experiences and activities after retirement may be key to understanding health and well-being in the retirement years.

3.1 Timing of Retirement

Both policies (governmental and corporate) and cultural norms shape individual expectations and beliefs about the 'right' time to retire (Kohli et al. 1991, Settersten and Mayer 1997). The implications of 'early' or 'late' retirement or of postretirement employment for health and well-being are not clear. For example, research in the United States shows that men retiring early are not necessarily doing so for health reasons (Burkhauser et al. 1996). Some studies suggest that early retirement may have deleterious health consequences in men's lives, with early retirement pro-

ducing more negative effects than later retirement (Palmore et al. 1985), while others have found no change in health based on early retirement (Streib and Schneider 1971). However, in a survey of relatively young retired men in the United Kingdom (McGoldrick 1989) more than half reported improved health as a benefit of retirement, with a quarter seeing it as a major benefit. Health and retirement timing are obviously intertwined; if individuals retire because of health problems, they often feel better off without the physical and psychological demands of their jobs.

3.2 Gender and Linked Lives

Most studies of retirement have focused on men, but women are increasingly retiring from paid work with different experiences and consequences. For example, married women are more likely to retire at an earlier age than single or divorced women and often retire as a consequence of their husbands' retirement exit. A spouse's poor health may cause employed wives to remain in the labor force, possibly for financial reasons (O'Rand et al. 1992). And having to care for ailing relatives may become an impetus for retirement, especially for women (Moen et al. 1994, Pavalko and Smith 1999). The coordination of both spouses' retirements and its implications for each spouse's health remain topics for future research.

3.3 Conditions of Work

The psychological impact of retirement appears to depend both on gender and on previous job stress, with men who leave high-stress jobs experiencing a reduction in distress symptoms, and men who leave low-stress jobs reporting an increase in distress symptoms. There are fewer effects, in either direction, for women undergoing retirement (Wheaton 1990). As Herzog et al. (1991) found among older US workers, being able to work one's preferred number of hours (rather than more or less) is positively related to physical health and life satisfaction, and negatively related to depression, regardless of gender or occupational status. Those who stopped work and felt they had little or no choice reported lower levels of health and well-being than both the voluntarily retired and those working the amount they would like. Another study (Gallo et al. 2000) found that late-life involuntary job loss through downsizing was related to declines in both physical functioning and mental health. While particular conditions of jobs at one point in time have been associated with health and well-being on the one hand, or stress and illness on the other (Karasek and Theorell 1990), a life course, role context approach suggests the importance of considering an individual's work patterns throughout adulthood.

3.4 Employment Patterns, Health, and Mortality

The nature of jobs and career trajectories can have long-term health implications. Studies have shown that men retiring from high-status occupations and 'orderly' career paths experience better health and greater longevity than those in manual occupations or those who have held a series of unrelated jobs (Moore and Hayward 1990, Pavalko et al. 1993).

Most of the studies linking health with career patterns have focused on men. The whole process of retirement may well be a different experience for women than men, in part because of the historical difference in their attachment to the labor force. When men leave their jobs they are exiting from a role that has typically dominated their adult years. Women, on the other hand, commonly experience greater discontinuity, moving in and out of the labor force and in and out of part-time jobs in conjunction with shifting family responsibilities. Consequently, they are less likely to have the same duration of employment or accumulation of work experience as men (Han and Moen 1999), and may be more likely to already have fulfilling roles outside of employment. Whether the findings about the longevity and health benefits of orderly career paths and high-status jobs apply to women as well is an important topic for future investigation. One recent study suggests that employment for women in their 50s and 60s may slow the onset of physical limitations (Pavalko and Smith 1999).

Ethnic and racial minorities, as well as those low in socioeconomic status, are more disadvantaged in terms of their career paths, health, and retirement pensions, creating a cumulation of disadvantage in addition to their increased risk of retiring because of a disability (Shea et al. 1996).

3.5 Life After Retirement

Postretirement employment is a key, yet understudied, aspect of employment history. 'Bridge jobs' following retirement are often very different from the career jobs from which workers retired (Henretta 1992), but they may ease the potentially disruptive transition out of the labor force. It is important to examine both pre- and postretirement employment histories to fully understand the linkages between work, retirement, and health, and to consider various forms of productive engagement following retirement.

Both paid and unpaid work after retirement have been linked to longevity, health, and psychological well-being (Moen et al. 2000). This confirms the protective effects of role participation, but studies have shown that the *context* of that participation matters as well. For example, Musick et al. (1999) found that volunteering had a protective effect on

mortality for those with low levels of informal social interaction.

Another key context shaping life after retirement is marriage. Pienta et al. (2000) have shown the health benefits of marriage for those moving into the retirement years.

4. Retirement Exigencies and Possibilities

Most research shows that being retired has no deleterious effects on either physical or psychological health; most retirees say they are satisfied with their retirement and some even report better health (Atchley 1976, Quick and Moen 1998). The issue is not whether retirement influences health, or vice versa, but the pathways to health and well-being in the post-retirement years. Research evidence makes it clear that good health is an important prerequisite for the enjoyment of the retirement years, as is economic security (Szinovacz and Washo 1992).

Both increasing longevity and the aging of the baby-boom cohort suggest that the retirement passage is in considerable flux, as growing numbers of workers can expect to spend a considerable proportion of their life course retired from their primary career jobs. Moreover, retirees in the twenty-first century tend to be healthier, better educated, and more active compared to those in the middle of the twentieth century.

The retirement transition can offer an occasion for the development of a new identity and can foster the adoption of new roles and new health behaviors. Workers retire, not only because of poor health, but sometimes because of good health, wanting to do something else. And retirees with excellent health are more likely to return to the workforce (Mutchler et al. 1999). But there are few institutionalized opportunities for meaningful engagement after retirement (Riley et al. 1994). Existing structural arrangements, policies, norms, and practices fail to provide the opportunities and challenges that can fully exploit the possibilities of life—and health—after retirement.

See also: Adulthood: Developmental Tasks and Critical Life Events; Aging, Theories of; Ecology of Aging; Employment and Labor, Regulation of; Life Course in History; Life Course: Sociological Aspects; Lifespan Development, Theory of; Population Aging: Economic and Social Consequences; Retirement, Economics of; Work, History of; Work, Sociology of

Bibliography

Anderson K H, Burkhauser R V 1985 The retirement-health nexus: A new measure of an old puzzle. *The Journal of Human Resources* **20**: 315–30

Atchley R C 1976 *The Sociology of Retirement*. Schenkman, New York

Burkhauser R V, Couch K A, Phillips J W 1996 Who takes early Social Security benefits: The economic and health characteristics of early beneficiaries. *The Gerontologist* **36**: 789–99

Delsen L, Reday-Mulvey G (eds.) 1996 *Gradual Retirement in the OECD Countries: Macro and Micro Issues and Policies*. Dartmouth Publishing, Aldershot, UK

Gallo W T, Bradley E H, Siegel M, Kasl S V 2000 Health effects of involuntary job loss among older workers: Findings from the health and retirement survey. *Journal of Gerontology: Social Sciences* **55B**: S131–40

Guillemard A M, Rein M 1993 Comparative patterns of retirement: Recent trends in developed societies. *Annual Review of Sociology* **19**: 469–503

Han S-K, Moen P 1999 Clocking out: Temporal patterning of retirement. *American Journal of Sociology* **105**: 191–236

Hayward M D, Grady W R, Hardy M A, Sommers D 1989 Occupational influences in retirement, disability, and death. *Demography* **26**: 393–409

Henretta J C 1992 Uniformity and diversity: Life course institutionalization and late-life work exit. *Sociological Quarterly* **33**: 265–79

Herzog A R, House J S, Morgan J N 1991 Relation of work and retirement to health and well-being in older age. *Psychology and Aging* **6**: 202–11

Karasek R A, Theorell T 1990 *Healthy Work: Stress, Productivity, and the Reconstruction of Working Life*. Basic Books, New York

Kohli M, Rein M, Guillemard A M, van Gunsteren H (eds.) 1991 *Time for Retirement: Comparative Studies of Early Exit from the Labor Force*. Cambridge University Press

McGoldrick A E 1989 Stress, early retirement and health. In: Markides K S, Cooper C L (eds.) *Aging, Stress and Health*. Wiley, New York

Moen P, Fields V, Quick H E, Hofmeister H 2000 A life course approach to retirement and social integration. In: Pillemer K, Moen P, Wethington E, Glasgow N (eds.) *Social Integration in the Second Half of Life*. The Johns Hopkins University Press, Baltimore, MA

Moen P, Robison J, Fields V 1994 Women's work and caregiving roles: A life course approach. *Journal of Gerontology: Social Sciences* **49**: S176–86

Moore D E, Hayward M D 1990 Occupational careers and mortality of elderly men. *Demography* **27**: 31–53

Musick M A, Herzog A R, House J S 1999 Volunteering and mortality among older adults: Findings from a national sample. *Journal of Gerontology: Social Sciences* **54B**: S173–80

Mutchler J E, Burr J A, Massagli M P, Pienta A 1999 Work transitions and health in later life. *Journal of Gerontology: Social Sciences* **54B**: S252–61

O'Rand A M, Henretta J C, Krecker M 1992 Family pathways to retirement. In: Szinovacz M, Ekerdt D, Vinick B (eds.) *Families and Retirement*. Sage, Newbury Park, CA

Palmore E B, Burchett B M, Fillenbaum G G, George L K, Wallman L M 1985 *Retirement Causes and Consequences*. Springer, New York, NY

Pavalko E K, Elder G H Jr, Clipp E C 1993 Work lives and longevity: Insights from a life course perspective. *Journal of Health and Social Behavior* **34**: 363–80

Pavalko E K, Smith B 1999 The rhythm of work: Health effects of women's work dynamics. *Social Forces* **77**: 1141–62

Pienta A M, Hayward M D, Jenkins K R 2000 Health consequences of marriage for the retirement years. *Journal of Family Issues* **21**: 559–86

Quick H E, Moen P 1998 Gender, employment, and retirement quality: A life-course approach to the differential experiences of men and women. *Journal of Occupational Health Psychology* **3**: 44–64

Riley M W, Kahn R L, Foner A 1994 *Age and Structural Lag: Society's Failure to Provide Meaningful Opportunities in Work, Family, and Leisure*. Wiley, New York, NY

Settersten R A, Mayer K U 1997 The measurement of age, age structuring, and the life course. *Annual Review of Sociology* **23**: 233–61

Shea D G, Miles T, Hayward M 1996 The health-wealth connection: Racial differences. *The Gerontologist* **36**: 342–9

Streib G F, Schneider C J 1971 *Retirement in American Society*. Cornell University Press, Ithaca, NY

Szinovacz M, Washo C 1992 Gender differences in exposure to life events and adaptation to retirement. *Journal of Gerontology* **47**: S191–6

Wheaton B 1990 Life transitions, role histories, and mental health. *American Sociological Review* **55**: 209–23

P. Moen

Retirement, Economics of

When most people retire, they leave a full-time job and stop working entirely. However, for a substantial minority, retirement is not a simple transition from working full-time to not working at all. It is a complex process. Those who have been working for an extended period for a particular employer may leave that job. They may continue to work for someone else on a job that is less demanding. In some less common circumstances, the long term employer will allow the individual to phase into full retirement, by reducing the demands on the job, or hours of work. Some people may even leave the labor market and then return. This article describes the complex patterns that comprise retirement, and discusses the leading explanations for retirement outcomes.

Given this diverse set of outcomes, various researchers and government agencies define retirement differently. Because of the complexity of the retirement process, each definition of retirement oversimplifies the story. According to one definition, retirement occurs when a person stops work completely. Among the problems with this definition, many who continue to work are already in the process of retiring. Thus, this definition classifies many as not retired, even though they have left a long term job, are working in an easier job than they held for most of their work life, are employed on a part-time basis despite having normally worked full-time, are receiving a lower rate of pay than they were paid before they changed their job or work conditions, even though they may be receiving pension or social security benefits. Another definition pertains to leaving long term employment. This definition classifies a person as retired when the older individual leaves a job held for 10 or 20 years. A problem with this definition is that it classifies those who have left a long-term job but are still at work as retired people. Not only does this definition classify those who are continuing to work, but are phasing into retirement gradually as retired, but according to this definition, people who are working full time but lost or left a job in their 40s or 50s are retired, even though they are not yet in the process of leaving the labor market. From the perspective of government agencies such as the Social Security Administration, or a firm's pension plan, a person may be considered retired when benefits are accepted. But many people work after having applied for or have received benefits. None of these definitions of retirement is correct or incorrect. Each definition attempts to reduce a complex process to one or another two-dimensional outcome. Consequently, a definition of retirement that is quite useful for one purpose may be not very helpful for another.

1. *Data Describing Retirement Outcomes*

To understand how retirement varies with age, it is useful to consider data from the Health and Retirement Study. This study is the leading source of data on retirement in the USA. It is also a model for retirement surveys that are now being adopted in other countries. Funded by the National Institute on Aging, every other year since 1992 this survey interviews 12,650 respondents who were born between 1931 and 1941, extensively tracking their retirement and other behavior. In a recent article some of the basic findings on retirement outcomes from the Health and Retirement Study are presented (Gustman and Steinmeier 2000b), and these findings are summarized here. One way to measure retirement is with self-reported status, a measure that incorporates the influence of a number of factors in determining whether an individual is retired, and considers three states: not retired, partially retired, and fully retired. At age 53, about 85 percent of men and 76 percent of women consider themselves to be not retired. Starting at age 54, the fraction not retired begins to decline, falling three or four percent for each year of age until age 61. At age 62 the share not retired falls precipitously, with a fifth of men and a seventh of women leaving non-retirement in that one year. At age 65, there is another large decline, with the fraction of men who are not retired falling 11 percent, and the fraction of women falling by 5 percent, leaving a little under a fifth of men and women reporting they are not retired at age 65. Thus from the perspective of many of the definitions of retirement noted, most people are retired after age 65. At the same time, from age 53 to age 65, the share of men reporting they are partially retired increases from 6 percent to over 22 percent, with the increase for women from 4 percent to 15 percent.

About 8 percent of full-time men and 6 percent of full-time women are still working after having left a long-time job, that is a job they held at age 45 that lasted more than 10 years. Although less than a fifth of the men and women are working full-time after reaching age 65, roughly another fifth classify themselves as partially retired. The percentages at each age of men and women who report they are completely retired are similar, but women are around 15 percent more likely to report that the retirement question is not relevant, presumably because they have not worked outside the home in a number of years.

Retirement status is not static. The number who are not retired in any year is determined in part by how many of those who reported they were not retired in the previous year remain not retired. It is also determined by how many who were not retired in the previous year leave that state for partial or full retirement. Finally, some people may classify themselves as not retired in a particular year, having previously been retired or partially retired. The same is true for the number who are partially retired or fully retired in any given year: those numbers depend on how many continue in the particular state from the previous year, how many leave the state from the previous year, and how many enter that state. Thus to understand the number reporting a given retirement status by age, we need to understand the flows into and out of, various retirement states over time. The Health and Retirement Study (HRS) is a panel data set that resurveys each respondent every other year. It records changes in retirement status over time as respondents age. In the HRS panel, 77 percent of the time respondents continue in the same Retirement State between waves of the survey—retired, partially retired, and not retired. In 17 percent of the cases there is a transition from greater to lesser work, and more difficult to model, 6 percent of transitions are from lesser to greater work. These transitions cumulate over the course of the survey. Altogether, 17 percent of the sample experienced a reversal in the course of the survey, moving from a state of less work to a state of more work.

Also of interest to students of retirement are trends in retirement outcomes. The aging of the baby boomers and decline in mortality have increased the financial pressure on social security and other retirement programs. One solution would be to postpone retirement. If people worked longer, the retirement period would be shorter and less costly. Instead, despite improving health, throughout the post World War II years, and until the end of the twentieth century, there was a continuing trend toward earlier retirement. Indeed, the trend to lower labor force participation by older males is one of the strongest labor market trends witnessed in recent years. (The trends in labor force participation and retirement for older women are not nearly as sharp. They are difficult to discern as the increased participation of women in the labor market offsets a tendency to leave earlier.) According to data from the US Current Population Survey cited in Anderson et al. (1999), continuing an earlier trend, from 1947 to 1985, the fraction of married men over 65 in the labor force fell from 55 percent to less than 20 percent. For those aged 55 to 64, the trend to lower labor force participation started later. From 1957 to 1985, the share of married men in the labor force fell from 90 percent to less than 66 percent. Similar trends are found in other countries. Indeed, the decline in the labor force participation rate of males in the US from 1960 is actually relatively modest compared to other countries, with Japan and Sweden experiencing milder trends, while the UK, Germany, Belgium, Spain, the Netherlands, France, and Italy have experienced even greater withdrawal (Gruber and Wise 1999).

There is evidence, though, of a leveling off in the middle 1980s in the trend toward reduced labor force participation of older men. For example, in 1985 in the USA, men who were 55 years old in the Current Population Survey had a labor force participation rate of 79 percent; in 1991 the participation rate of 55-year-olds had increased to 81 percent. Similarly, between 1985 and 1991, the participation rate for 60-year-olds increased by one percentage point, from 66 percent to 67 percent, and for those 65 years old, participation remained the same at 26 percent. Participation rates did not resume their downward trend in the 1990s.

The trend to earlier retirement for men is apparent when retirement outcomes from the Health and Retirement Study, collected from 1992 to 1998 in the USA, are compared with retirement outcomes from the Retirement History Study, collected from 1969 to 1979. Using figures from Gustman and Steinmeier (2000b), between these two surveys, the fraction of 60-year-olds who considered themselves to be not retired fell from 80 percent to 66 percent. The fraction of 62-year-olds not retired fell from almost 66 percent in the 1970s to a little more than 33 percent in the 1990s. Only at age 65 are the fractions not retired similar in the 1970s and 1990s. A very large number of people retired in the 1970s RHS at age 65, with almost 20 percent of the sample leaving nonretirement. In the 1990s HRS, the comparable number leaving at 65 is a little more than half as large—many more have already left. These trends can be seen in the numbers who retire at any given age. In the 1990s HRS sample, 20 percent of the men left nonretirement at age 62. Although there is evidence that the trend to earlier retirement has abated in the past few years, there is a continuing debate about whether the trend will reassert itself.

Researchers have also focused on other aspects of retirement. The family is the center of decision making. Thus there is a question about whether couples coordinate their retirement decisions. It seems they do. Researchers have documented the coincidence of retirement dates for husbands and wives, despite age

differences and differences in incentives to retire (Gustman and Steinmeier 2000a, Hurd 1990).

2. Explaining Retirement Outcomes

Students of retirement behavior would like to explain what determines these patterns of retirement by age. Of special interest is the onset of substantial flows into retirement beginning at age 53 or so, the sharp increases in retirement that are observed at ages 62 and 65, and the transition for most out of non-retirement by age 65. Also of interest is the question of why a majority of people proceeds directly from full-time work to complete retirement. A substantial minority partially retires, but what determines how long a person stays partially retired? Why do most people reduce their work effort as they age while others sometimes increase work effort after having reduced it? Also, why do different people choose different ages to enter and leave full time-work, partial retirement, and full retirement? Similarly, researchers would like to know what has caused the trends to earlier retirement over time. Why has the number of people retiring at age 62 increased while the large spike in the number retiring at age 65 has been halved, and more generally what caused the changes in retirement outcomes at each age? They also would like to understand other key patterns in the data, such as the coincidence of retirement by husbands and wives, and how the new patterns of increased participation by women are affecting retirement outcomes.

To understand what determines the wide variety of retirement patterns, economists rely on a model in which forward looking decision-makers weigh the benefits and costs of alternative retirement dates. Older people decide when to retire by taking into account not only current work and leisure opportunities, but also future opportunities. These opportunities include the pay that would be offered for work on the current job, now and into the future, and the pay on alternative jobs. Involuntary layoffs affect these opportunities, reducing the offer for continuing full-time work. Incentives created by pension plans and social security play important roles, especially in shaping the future consequences of a decision to retire. These programs have features that may change sharply the incentive to work at different ages, first encouraging workers to stay in the labor market, then at a later age, reducing the reward to continue work. Against the monetary rewards to continue work, older individuals weigh the difficulty of the work. Whether these demands are physical or mental, they are likely to increase with age at different rates, depending on the job held. The opportunity cost of working is the value of leisure time. Health problems may increase the difficulty of work, and lower the reward to continued work if the individual becomes less efficient at certain tasks. Retirement decisions of spouses, and demands from dependent children and elderly parents also help to shape the retirement decision.

It is differences in the rewards to continued work, the effects of social security and their differing pension plans on these rewards, differences in health status, job conditions, as well as different evaluations of the same set of benefits and rewards, that create wide differences in decisions of when to retire. There is also an interaction with the decision to save, and accumulated assets, leaving some more able to afford retirement at a given age than others with the same earnings history. Similarly, changes over time in wages, pensions, social security, and other factors determining the reward structure to continued work, as well as changes in the valuation of leisure, may affect the trends observed in retirement outcomes.

2.1 Key Explanatory Factors Affecting Retirement Outcomes

Consider now in more detail some of the leading factors shaping retirement outcomes and how they influence retirement behavior.

2.1.1 Wage offers for full-time work. Theory does not tell us what relationship to expect between the wage a person makes and retirement age. On the one hand, a higher wage implies greater forgone earnings when one leaves work. On the other, a higher wage implies a higher lifetime income, which should lead the high wageworker to prefer more leisure, just as those with higher incomes prefer more of other luxury goods. To date, the evidence suggests these effects are roughly offsetting, so that higher wages on long term jobs do not have a strong relation to retirement outcomes. There are other possible explanations that might allow the wage to account for an increase in retirement with age; for example, if wage offers declined with age, that might reduce the reward to continued work for older individuals. But they do not. Sometimes researchers record the wage rate received by workers at different ages and find that the observed wage declines with age. This effect is illusory. As workers age, some of the older workers have left their main jobs and are holding lower paying full- or part-time jobs as they transit to retirement. To be sure, some older workers experience layoffs, and, once laid off, older individuals certainly have greater difficulty locating new jobs, but that is not a dominating factor in shaping retirement.

2.1.2 Minimum hours constraints and lower wages when partially retired. There is evidence that many jobs do not permit older workers to phase gradually

into retirement. Rather, at the job they have held for many years minimum hours constraints allow the worker to choose only whether to work full-time, or not at all. Work is available on a part-time basis, but only at much reduced wage. These minimum hours' constraints, together with a lower wage offer when one works part-time, account for the decision by many to move directly from full-time work to full retirement.

Pensions and their provisions may account for some of the diversity in retirement ages. There are two basic types of pensions. One plan, called a defined benefit plan, provides a periodic payment after retirement that is based on a formula. The amount of the benefit is typically determined by years of service, age, and pay. So is the age of eligibility for receiving a full, unreduced, benefit (normal retirement age) and the age of eligibility for a reduced benefit (early retirement age). Defined contribution plans, which at the beginning of the twenty-first century are most often in the form of 401(k) plans, are essentially tax-favored retirement saving plans. Both the covered worker and the employer may contribute to an account, often with the employer matching the worker's contribution according to a set schedule, up to a predetermined amount.

'Defined contribution plans' do not create sharp incentives that affect the reward to remaining on the job, but defined benefit plans do. Those who leave a 'defined benefit plan' before qualifying for early retirement benefits may be penalized heavily. Conversely, an individual who stays until qualifying for early retirement benefits may be heavily rewarded. Studies of workers employed at firms offering defined benefit plans indicate that workers respond to these incentives. Among studies that pioneered in investigating the relation of pension plan incentives to retirement outcomes, see Burkhauser (1979), Fields and Mitchell (1984), and Stock and Wise (1987). According to pension data for respondents to the Health and Retirement Study (Gustman and Steinmeier, 2000c), defined benefit pensions on average had an early retirement age of 54, and an average normal retirement age of 61. Only 6 percent have an early retirement age above 60. Forty percent face a normal retirement age of 65; with 14 percent having normal retirement at age 62; over 20 percent with normal retirement at age 60; and over a fifth having a normal retirement age below 60. A man who has a defined benefit (DB) pension plan and works the year before qualifying for early retirement, on average, will find benefits increased by an amount equal to about 120 percent of one year's of pay. For a woman, the increase amounts to about a third of a year's pay.

In addition to contributing to the explanation for the diversity of retirement dates, changes in pensions account for some of the trend to earlier retirement. For the 1970s and 1980s, it has been estimated that about 10 percent of the trend to earlier retirement is due to the decline in ages of early retirement in defined benefit pensions and to other changes in pensions (Anderson et al. 1999).

2.1.3 Social security. Analyses of retirement behavior suggest that social security has had a modest effect in influencing retirement, and that effect will decline for those now approaching retirement. Social security benefits are determined from past covered earnings history, where past earnings are indexed to age 60 and are averaged to a summary statistic called the Average Indexed Monthly Earnings (AIME). For those reaching age 62 after 1991, AIME is calculated using the highest 35 years of indexed earnings. If an individual has covered earnings for fewer than 35 years, then zeros are entered into the AIME calculation for the remaining years. From the AIME, the basic benefit, called the primary insurance amount (PIA), is computed. A progressive benefit formula is applied so that those who have low computed lifetime earnings have higher benefits, relative to earnings, than do those with high earnings. The earnings measure is typically expressed as a monthly amount, but when annualized, the formula for the year 2000 specifies benefits that are 90 percent of the first $6,372 of annual earnings, 32 percent of the next $32,052, and 15 percent of remaining earnings. Benefits to spouses and survivors affect the relationship between benefits and earnings, both at the level of the individual and at the level of the family. Spouses are entitled to roughly half of their partner's benefits, and survivors are entitled to an amount roughly equal to the benefits that would have been payable to the deceased spouse. The social security benefit is subject to an earnings test for those between the ages of 62 and 65. But any benefits that are lost to the earnings test are restored at age 65. Indeed, for married couples, the additional benefits from waiting until age 65 to claim benefits are worth slightly more than the benefits that were lost.

However, this was not always the case. In the past, the earnings test applied to earnings of respondents up to the age of 72. However, the age at which the earnings test no longer applies has been reduced slowly, until in the twenty-first century it only applies to those between the early and normal retirement age. Moreover, in the past, the adjustment for lost benefits was not nearly as generous as it is today. Initially, the social security benefit structure discouraged work after the age of 65. With recent changes, the program is roughly neutral in its effects on the incentive to retire. Thus social security in the past may have created incentives that accounted for some of the retirements, perhaps both before and after age 65. For example, it is estimated that changes in the benefit formula in the 1970s and 1980s may have increased early retirement by about an eighth. But social security should not have

any effect on retirements after the age of 65 in future years.

International evidence reinforces the view that pensions and social security have important effects on retirement behavior. With the aging of the baby boom, the worry in many countries is that pensions and social security discourage still productive individuals who would otherwise work to leave the labor force.

Different countries have very different pension and social security systems than the USA. Differences in design among countries create incentives to retire at different ages. The different retirement incentives created by the diverse retirement systems among countries provide another test of the influences of social security and pensions in shaping retirement behavior. Gruber and Wise (1999) have brought together researchers from a number of countries, both to describe retirement outcomes, and to see how their pension plans and social security systems affect retirement behavior. They find a close correspondence between the early retirement dates and other features of the programs, and the ages at which older workers tend to leave the labor force. Moreover, they find that when these dates of early retirement and normal retirement specified in the social security systems of these countries change, retirement behavior changes correspondingly. Compared to many other countries, the USA has a low relative tax rate on work after the early retirement age specified in the social security system, which is one reason for the higher labor force participation of older men in the USA. Countries with very high tax rates, which include Belgium, Italy, France, and The Netherlands, have lower participation rates by older workers than the USA.

Many researchers believe that social security has substituted for other forms of retirement saving. Many people enter retirement with no assets, and social security benefits as their sole support. Low-income families may have half or more of their earnings replaced by social security. Those who rely entirely on social security cannot retire before the age of 62. They do not have savings to draw on to support their consumption until they became eligible for social security benefits at 62. As a result, if the social security early retirement age were raised from 62 to 65, there would be a large increase in the number delaying their retirement.

2.1.4 Health status. Health changes are perhaps the most important of all the uncertain events affecting older workers' labor force participation. Data on the course of HRS respondents' health events from 1992–8 suggest a wide variety of health problems among workers, including limitations in activity, difficulties in carrying out mental and physical tasks, and medical problems such as high blood pressure. Both the level of health status and changes in health status are associated with different patterns of labor force participation in the HRS (Bound et al. 1998).

2.1.5 Retirement decisions in the family. Findings at the end of the twentieth suggested that many spouses retire together, even though wives typically are younger than their husbands. Available estimates suggest that each spouse, and husbands in particular, values retirement more once their spouse has retired (Gustman and Steinmeier 2000a). It does not appear that husbands and wives time their retirements to coincide simply because of their pensions or social security. Other family related considerations might help to shape retirement. Some families married or remarried late, have young children, had a spouse who never returned to the labor market once having had children, or faced great demands on time to care for older parents, or a spouse. These and other differences among families may help to account for differences in retirement.

2.1.6 Other factors affecting the retirement decision. There is also evidence supporting the influence on retirement of many other factors that differ among individuals and their jobs. Labor market demands for older workers have been investigated. It is clear from this work that many firms are reluctant to hire older workers, because they will not work over a long enough period for their hiring and training investments to be repaid, together with other reasons (Hutchens 1986). The business cycle also influences the demand for older workers. On the one hand, in times of tight labor markets, firms will have difficulty replacing older workers and will encourage them to postpone retirement. On the other hand, in times of slack demand, firms adopt policies such as one-time windows that are designed to encourage earlier retirement. Those who invested more in training for the labor market by staying in school longer may work longer to justify their investments. Unions may shape retirement conditions both to accommodate the preferences of their older workers, and to share work with younger union members. Government policies might have had profound effects, but they have been pushing against the trend. Thus the government since 1980 has outlawed mandatory retirement, forced firms to make pensions more actuarially fair so as not to penalize work after the normal retirement age, and eliminated the penalty to work after age 65 under social security. Perhaps these policies have had some small effect in mitigating the trend to earlier retirement in the 1990s. However, there is no firm evidence to support that view.

2.1.7 Imperfect information and nonmaximizing behavior. Of course, the choices we see are deter-

mined by more than the maximizing choices made by people who fully understand all their opportunities. Many people do not understand fully the incentives from their retirement programs. Some plan well, some don't plan at all. In addition, some that are well aware of the importance of preparing for retirement do not have the discipline to follow through. Most people are well prepared for retirement. But if they have not prepared, they will either have to consume less when retired, or to work longer. Economic models are now being expanded to understand better the behavior of those who may not understand fully all the incentives, or for other reasons may not be choosing their retirement date according to a model of strict maximization. Certainly this work will modify our perception of how each of the factors discussed influence retirement choices. Nevertheless, the modifications are expected to be modest, and the major influences on retirement are the forces discussed.

3. Conclusions

Most of the information we have on how pensions, social security, health status, and other factors affect retirement outcomes comes from analyses that were based on 1970s data (e.g., see Gustman and Steinmeier 1986, Rust and Phalen 1997). The data from the new Health and Retirement Study are being used to estimate comprehensive models of retirement. With these newly available data, researchers are in a position to quantify how the factors noted in Sect. 2 influence retirement outcomes for those who are just a little older than the baby boomers. The wave of research now in process should also leave us in a good position to understand what will determine the retirement behavior of the baby boomers, and how changes in government social security and pension policies will affect their retirement outcomes.

See also: Age: Anthropological Aspects; Age Policy; Age, Sociology of; Age Stratification; Age Structure; Ecology of Aging; Leisure and Cultural Consumption; Life Course in History; Retirement and Health; Sex Differences in Pay

Bibliography

Anderson A L, Gustman T, Steinmeier L 1999 Trends in male labor force participation and retirement: Some evidence on the role of pensions and Social Security in the 1970s and 1980s. *Journal of Labor Economics* **17**(4) p 1: 757–83

Bound J, Schoenbaum M, Stinebrickner T R, Waidman T 1998Thedynamiceffectsofhealthonthelaborforcetransitionsof older workers. *NBER Working Paper* **6777**

Brown C 2000 Early retirement windows. In: Olivia S, Mitchell P, Hammond B, Rappaport A M (eds.) *Forecasting Retirement Needs and Retirement Wealth*. University of Pennsylvania Press for The Pension Research Council, PA, pp. 253–73

Burkhauser R V 1979 The pension acceptance decision of older workers. *Journal of Human Resources* **14**(1): 63–75

Fields G S, Mitchell O S 1984 *Retirement, Pensions and Social Security*. MIT Press, Cambridge, MA

Gruber J, Wise D A (eds.) 1999 *Social Security and Retirement Around the World*. University of Chicago Press, Chicago

Gustman A L, Steinmeier T L 1986 A structural retirement model. *Econometrica* **54**(3): 555–84

Gustman A L, Steinmeier T L 2000a Retirement in dual career families. *Journal of Labor Economics* **18**(3): 503–45

Gustman A L, Steinmeier T L 2000b Retirement outcomes in the health and retirement study. *Social Security Bulletin* **64**(4): 3–17

Gustman A L, Steinmeier T L 2000c Employer provided pension data in the NLS mature women's survey and in the health and retirement study. *Research in Labor Economics* **19**: 215–52

Hurd M D 1990 The joint retirement decision of husbands and wives. In: Wise D A (ed.) *Issues in the Economics of Aging*. University of Chicago Press, Chicago, pp. 231–54

Hutchens R 1986 Delayed payment contracts and a firm's propensity to hire older workers. *Journal of Labor Economics* **4**(4): 439–57

Rust J, Phalen C 1997 How social security and medicare affect retirement in a world of incomplete markets. *Econometrica* **65**: 781–832

Stock J H, Wise D A 1987 Pensions, the option value of work and retirement. *Econometrica* **58**(5): 1151–80

A. L. Gustman and T. L. Steinmeier

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Rev

Reviews, Critics, and Cultural Gatekeepers

It is argued here that the expertise of cultural gatekeepers does not pertain to the identification of those intrinsic properties that make cultural products into works of art. The expertise of cultural gatekeepers rather consists in following procedures that enable them to classify cultural products according to genre and to rank them according to quality. The use of these group-bound procedures enhances the chance that proposed classifications and evalutations will meet agreement from other gatekeepers. Empirical research has clarified the nature and effects of procedures used by gatekeepers, in particular by literary critics. These procedures seem to be endogenous to the institution of criticism, thereby suggesting its autonomy from other realms of society. However, when other cultural gatekeepers are taken into account, e.g., publishers, and when questions are posed about the content and value they attribute to cultural products, the procedures involved in this process manifest a complex interplay between a specific cultural sector and other fields of society. The autonomy of the cultural sector from other social sectors is relative in nature. All labels attached to cultural products, referring to their content, genre or value, are multidimensional in nature. What is to be understood by the value various parties attach to cultural products calls for cooperation between sociologists and economists: the former can profit from their insights into the nature of the procedures followed by cultural gatekeepers; the latter have their sophisticated analyses of value ('utility') to offer.

1. The Traditional View of Cultural Gatekeepers' Expertise

Cultural gatekeepers—reviewers, publishers, gallery owners, film producers—occupy institutional positions between cultural producers and consumers. They determine which products cross the boundary between the private and the public domains, and when this boundary is crossed, they propose classifications and evaluations of the products that are included in cultural repertoires. Gatekeepers are credited with having professional expertise on what makes cultural products into works of art.

Giving or denying access to the public sphere is an eminently social task, yet the social sciences have been slow in developing a general perspective on what gatekeepers do, the conditions under which they operate, their practices, and their impact on the reputations of cultural products and their makers. One of the reasons why cultural gatekeepers have remained so long outside the scope of the social sciences is that, for centuries, cultural expertise has been believed to pertain to the *intrinsic* properties on the basis of which an individual cultural product may count as a work of art. This is an instance of 'charismatic ideology', to use a term coined by Bourdieu and Darbel (1969, p. 161). According to this ideology, works of art have the power to convince any rational being of their particular nature (i.e., that of being a work of art), and of their quality. If works of art can induce by themselves the recognition of their nature and quality, they must have specific characteristics which are lacking in other cultural products. The expertise of gatekeepers is believed to consist essentially in the ability to identify those intrinsic characteristics that induce the judgment that we are dealing with works of art.

However, as far as cultural gatekeepers are concerned, it is extremely hard to identify sets of general and firmly established procedures for interpreting or judging a poem, a movie, or a painting; similarly, there seem to be no clear-cut procedures by which works of art are classified as having a 'style' in common, or as differing from one another in this respect. In light of the charismatic ideology, the question of what procedures cultural gatekeepers use seems to lose its urgency: If one believes that works of art have persuasive powers, and if one believes that each work of art is unique, then it seems useless to suppose that the successful distinction between cultural products that belong to the respective domains of 'art' and 'non-art' requires explicit, let alone general, procedures.

In the nineteenth and twentieth centuries, the study of the arts developed into specialized academic disciplines. As a result, the question of which procedures should be observed in order to arrive at sound conclusions became preeminent. Literature offers a key example, for among the human sciences, the study of literature has produced the largest number of approaches to what counts as its central object of research, the literary text. Russian Formalism, Czech

Structuralism, New Criticism, Jakobson's linguistic poetics, reception aesthetics, and narratology are cases in point, not to mention Marxist, psychoanalytic, feminist, and deconstructivist studies of literature—all of which encompass a considerable number of widely diverging approaches. The privileged position of literature in secondary education has facilitated the proliferation of valuative procedures. Young students are initiated into how to 'appreciate' a nation's great literary works, and to a much lesser extent into appreciation for a nation's great movies, paintings, or musical compositions. The assessment of the value of the canonical works is constantly adapted to the needs of new generations.

In the recent past, two parallel, but mutually independent, developments have undercut the idea that cultural gatekeepers' expertise essentially pertains to those intrinsic properties that endow cultural products with the specific nature and quality of works of art. First, researchers have come to the view that value judgments are not statements which can be shown to be true or false. Moreover, they cannot be derived logically from descriptive or analytical statements to which the true/false distinction applies. Instead, any logical derivation of an evaluative statement from a descriptive one involves an appeal to 'institutional facts', i.e., to constitutive rules embodying norms and obligations to which one is committed (cf. Searle 1970, p. 175ff.). This has contributed to adopting a relativistic attitude towards values, in that they were seen as beliefs or views that are endorsed (or not) by specific social groups. A sociological approach to norms and values, and, in particular, to the value-laden distinction between 'art' and 'non-art' became urgent. It soon became a viable option because Bourdieu's work offered a framework within which the adjudication of 'symbolic value' to products, i.e., value deemed to be specifically cultural in nature, could be assessed empirically.

The second development, which made Bourdieu's views a viable alternative for research done in the humanities, was the extensive discussion of the methodological status of interpretations of literary texts and other forms of cultural expression. Again taking literature as the exemplary case, the outcome of this discussion showed literary criticism, as practiced in newspapers, in literary magazines, and at universities, to be linked tightly to normative conceptions of literature, i.e., to mostly implicit ideas about the characteristics a text should have in order to count as a form of literature (Verdaasdonk 1979, Van Rees 1986). In consequence of this, critics' statements about literary texts cannot be regarded as assessments of intrinsic properties these texts possess. Their statements show an inextricable mixture of classificatory and valuative aims (cf. Verdaasdonk 1994, 2001). It should be emphasized that not only literary criticism, but all studies of art and culture current in the humanities suffer from a similar lack of clear and reliable procedures for identifying properties that are supposed to make cultural products into works of art.

2. *Empirical Research on Gatekeepers' Adjudication of Value to Cultural Products*

Bourdieu (1980) offers a comprehensive view of what critics, reviewers and cultural gatekeepers do in dealing with cultural products. The questions Bourdieu (1980) addresses are: What constitutes the authority of cultural gatekeepers? What factors affect the consensus they attain on the nature and value of cultural products? How do gatekeepers interact with one another in this process? Under what conditions do gatekeepers have an impact on audiences of art and culture?

In deciding that a product counts as a work of art, gatekeepers assign 'symbolic value' to it, thereby setting it apart from the set of products to which they refuse to give artistic status. Evidently, gatekeepers act as representatives of cultural institutions, such as criticism, publishing houses, movie studios, theaters, concert halls, etc. This largely accounts for the legitimacy of their decisions: as representatives of specific institutions, gatekeepers are seen—and see themselves—as being endowed with expertise that is socially acknowledged. However, Bourdieu stresses that this form of legitimacy is not sufficient in itself: gatekeepers may criticize decisions made by their counterparts; and critics and reviewers often have sharply contrasting opinions about the nature and value of particular cultural products. Gatekeepers' authority adds strongly to the legitimacy their decisions have in the eyes of other gatekeepers. Bourdieu holds that gatekeepers' authority essentially is the credit they have earned by making previous successful decisions—decisions that were adopted by their fellow gatekeepers. Authority is a reputation for success: it is *believed* that authoritative gatekeepers will be as successful in the future as they were in the past in discovering and promoting artists and their work. In this view, authority is not thought to be based on a successful determination of those intrinsic properties that account for the artistic status of cultural products; authority is rather conceptualized as the impact of a gatekeeper's stances on those of others.

Consensus formation among reviewers and critics about the value of cultural products, and the interaction these gatekeepers engage in, have been taken up by researchers. So far, the nature of gatekeepers' expertise and their relations with audiences have received much less attention.

Many researchers have taken a closer look at the forms of institutionally determined behavior that gatekeepers manifest in attuning their selections to those of their counterparts. It has been found that they strongly agree on which products have to be regarded as being of superior quality; and, conversely, they

attain much less agreement about how to classify and to characterize individual products. New literary releases that in the majority of newspapers were hailed as 'important' stand a much better chance of becoming, after a certain period of time, objects of scholarly attention (Verdaasdonk 1983). This suggests that, in the course of time, the selections made by a first group of gatekeepers are taken over by other groups of gatekeepers—with a certain amount of downsizing of the choices that were made previously. Van Rees (1983) has argued that reviewers in newspapers, critics in literary magazines, and academic scholars act as three successive filters through which, gradually, a repertoire of 'important' literary works is established. Ultimately, this repertoire enters the literature curriculum of secondary schools. The universities seem to follow the selections made by the scientifically less prestigious institutions of newspapers and literary magazines. Such research suggests that consensus formation among reviewers and critics is a process that is strongly reproductive in nature: Previous selections shape subsequent choices.

Do gatekeepers, in composing repertoires of noteworthy products, characterize them in the same way? This seems doubtful. Critics classify books, movies. paintings, etc. as having a certain 'style' or 'content' in common. De Nooy (1993) found that such classifications do not cross the borders of one single institution. The classifications which, in the 1970s, reviewers and critics proposed of the recent Dutch literary production appeared to be unknown to other gatekeepers, such as booksellers, librarians, and teachers of literature. De Nooy performed network analyses on these classifications and came to the conclusion that the groupings of authors were based on the specific magazines and publishing houses that featured these authors—although the critics held firmly that stylistic and thematic analogies, and differences between the texts concerned, provided the basis for their clustering.

The 'production of culture' perspective, advocated by Peterson (1976, 1985), holds that 'the nature and content of symbolic products [...] are significantly shaped by the social, legal and economic milieux in which they are created, marketed, purchased and evaluated' (Peterson 1985, p. 46). This view is fully congenial to Bourdieu's approach. However, posing questions about gatekeepers' behavior other than those addressed by Bourdieu, leads to different assessments of the interaction between factors which are thought to be 'endogenic' or 'exogenous' to specific cultural institutions, and to a fresh view on the nature of such factors. In their study of the American publishing industry, Coser et al. (1982, p. 146) found that, in judging book proposals or manuscripts, there were systematic differences between editors according to the type of publishing house where they were employed: Fewer than half of the university press editors or commercial scholarly editors rated profitability or commercial prospects as the most important factor; in contrast, more than 60 percent of both textbook and trade editors listed commercial prospects as the single most important consideration. Trade and college editors are more concerned with first-year profitability, while scholarly editors pay more attention to long-term commercial success. Here, organizational differences affect the weight and the nature of judgments of quality and profitability, and of their interrelations. For these gatekeepers, 'denials of the economy', to use Bourdieu's term, are less possible than for gatekeepers deciding on a product's artistic status.

When questions about content are posed—the content of cultural products or of what gatekeepers say about them—the nature of the factors that seem 'endogenous' to the cultural sector and of those that seem 'exogenous' to it will shift. Griswold (1981) notes that after the USA recognized international copyright agreements in 1891, the themes of trade novels showed a strong convergence. American publishers did not concentrate any more on English novels about love and domestic life that could be obtained free of rights; American authors did not compete any more with these novels, by treating other themes that could also appeal to a large audience. Economic viability appears to be a major factor shaping the fare trade publishers brought out at the beginning of the twentieth century.

The cultural sector offers a huge and rapidly changing supply of products. Audiences also change—in size and composition. In many Western countries, an increase of participation in higher education has led to a growth and a refined segmentation of the audiences of art and culture. Of course, this has an impact on the behavior of gatekeepers. Cameron (1995) observes that classical music is marketed now by record companies as a mass consumption good. A number of marketing ploys has been used to reduce the entry barriers for consumers. One of these ploys is to provide excerpts of classical works. A free CD featuring these excerpts is distributed together with a magazine which reviews the CD in nontechnical language. This represents a change in the function of criticism and in its major task: The adjudication of symbolic value to cultural products. Cameron (1995, p. 328) says: 'The aesthetic dimension remains in the product but consumption is increased despite it.' In this example, the awareness of the market place takes a form that is very different from that envisioned by Bourdieu.

3. *Perspectives on Future Research*

As the above discussion shows, considerable progress has been made in pinpointing factors affecting the behavior of cultural gatekeepers. However, most models are 'static' in that they do not account for the impact of time on interactions between dependent and

independent variables. Gatekeepers' activities encompass a considerable period of time; their decisions are conditional on previous ones. Audiences change, as do the factors—and their interactions—that have impact on the production of cultural objects, and on their classification and evaluation. Many commentators, e.g., Peterson and Simkus (1982) and Griswold (1998) have pointed out that the traditional boundaries between 'high' and 'popular' culture undergo rapid and fundamental changes. This entails a focus on the variations in cultural products, their appraisal, and their consumption manifest over time. Griswold (1998, p. 34) speaks of a 'cultural swirl of mutual influence with multiple flows, no impermeable boundaries and few fundamentalists.' If cultural consumers are to be seen as 'occupying different positions in a differentiated distribution of knowledge and cultural resources' (Griswold 1998) this calls for a 'dynamic' approach to analyzing the behavior of gatekeepers, distributors, reviewers, and producers serving these audiences. As yet, longitudinal, time series, and event history data are rarely used in sociological research on culture. However, such data may be essential to capture the time-varying impact some independent variables may have on dependent variables.

Although many studies address the modus operandi of cultural gatekeepers, there has been scant attention paid to what constitutes their expertise. It has been argued here that gatekeepers are unable to identify those intrinsic properties that are deemed to account for the artistic nature and value of a cultural product. When it is acknowledged that meaning is not holistic or fixed, but is whatever works for a specific group (cf. Griswold 1998, p. 35), then the focus should be on the procedures gatekeepers actually follow in reaching agreement on the nature and value of cultural products. In the final chapter of *Distinction*, Bourdieu distinguishes between the practical mastery of classifications and the reflexive mastery that is required in order to construct a taxonomy that is coherent and adequate to social reality (1984, p. 472). This idea may also be useful to analyzes the ways in which gatekeepers come up with taxonomies and value judgments. If one admits that cultural expertise does not pertain predominantly to knowledge of the intrinsic characteristics of cultural products, gatekeepers may be seen mainly as experts in executing specific tasks with regard to cultural products. Such tasks may be: Presenting a product to connoisseurs or laymen, elaborating on its meaning or value by making comparisons with other products, assessing a product's eligibility for a subsidy or an award, etc. The use gatekeepers make of their experience with individual products may vary with such tasks; the procedures they follow in order to arrive at results that are condoned by other gatekeepers may also vary, and may lead to different groupings and evaluations of the same products.

Future research on cultural gatekeeping is likely to address the still considerable gap between economic and sociological analyses. Frey and Pommerehne (1989, p. 12f), two founding fathers of economic research on culture, remark that sociologists tend to focus on the demand side of culture, whereas economists take the interplay between supply and demand into account. Most contributions to the economics of culture follow the neoclassical approach in asking questions about the determinants of prices of works of art, the rationality of multiple art markets (Singer and Lynch 1997), or the profit-maximizing behavior of consumers of culture. In these studies, a strong similarity is assumed to exist between the symbolic and economic value of cultural products (an assumption that sociologists, as discussed above, have been at pains to attack). The finding, for instance, that successful blockbuster movies feature megastars who account for the large audience these movies attract comes very close to making the assertion that movies with a huge production budget—megastars get gigantic fees—are more likely to enter the top 20 than movies made with smaller sums of money (cf. Albert 1998).

Economics and econometrics are well equipped to analyze many types of value assignments—manifested by preferences and by choices. Gatekeepers engage in choice behavior and develop preferences, but these processes have not received much attention from sociologists. Not only the process of value attribution, but also the question of how value develops over time—under what conditions does it accumulate or diminish?—is important. Combining economic and sociological techniques, it should be possible to determine how gatekeepers and 'ordinary' consumers perceive differences between individual cultural products, and how they go about assigning increasing, decreasing, or stable values to them.

See also: Cultural Critique: Anthropological; Cultural Policy: Outsider Art; Hegemony: Cultural

Bibliography

Albert S 1998 Movie stars and the distribution of financially successful films in the motion picture industry. *Journal of Cultural Economics* **22**: 249–70

Bourdieu P 1980 The production of belief: Contribution to an economy of symbolic goods. *Media, Culture and Society* **2**: 261–93

Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA]

Bourdieu P, Darbel A 1969 *L'amour de l'art. Les musées d'art européens et leur public*. Minuit, Paris

Cameron S 1995 On the role of critics in the culture industry. *Journal of Cultural Economics* **19**: 321–31

Coser R, Kadushin C, Powell W 1982 *Books: The Culture and Commerce of Publishing*. Basic Books, New York

De Nooy W 1993 Richtingen en lichtingen. Literaire classificaties, netwerken, instituties [Movements and currents. Literary classifications, networks, institutions]. Unpublished PhD thesis, Tilburg University, The Netherlands
Frey B, Pommerehne W 1989 *Muses and Markets. Explorations in the Economics of the Arts*. Blackwell, Oxford, UK
Griswold W 1981 American character and the American novel. *American Journal of Sociology* **86**: 740–65
Griswold W 1998 The burnished steel watch. What a sample of a single year's fiction indicates. In: Janssen S, Van Dijk N (eds.) *The Empirical Study of Literature and the Media*. Bajesteh Van Waalwijk Van Doorn, Rotterdam, The Netherlands
Peterson R 1976 The production of culture: Prolegomenon. *American Behavioral Scientist* **19**: 669–84
Peterson R 1985 Six constraints on the production of literary works. *Poetics* **14**: 45–67
Peterson R, Simkus A 1992 How musical tastes mark occupational status groups. In: Lamont M, Fournier M (eds.) *Cultivating Differences. Symbolic Boundaries and the Making of Inequality*. The University of Chicago Press, Chicago, IL
Searle J 1970 *Speech Acts. An Essay on the Philosophy of Language*. Cambridge University Press, Cambridge, UK
Singer L, Lynch G 1997 Are multiple art markets rational? *Journal of Cultural Economics* **21**: 197–218
Van Rees C J 1983 How a literary work becomes a masterpiece. On the threefold selection practised by literary critics. *Poetics* **12**: 397–418
Van Rees C J 1986 Literary theory and criticism. Conceptions of literature and their application. Unpublished PhD thesis, University of Groningen, The Netherlands
Verdaasdonk H 1979 Critique littéraire et argumentation. Unpublished PhD thesis, Free University of Amsterdam, The Netherlands
Verdaasdonk H 1983 Social and economic factors in the attribution of literary quality. *Poetics* **12**: 383–95
Verdaasdonk H 1994 Analogies as tools for classifying and appraising literary texts. *Poetics* **22**: 373–88
Verdaasdonk H 2001 Quotations expressing the valuative stands of literary reviewers. In: Schram D, Steen G (eds.) *The Psychology and Sociology of Literature*. Benjamins, Amsterdam, The Netherlands

H. Verdaasdonk

Revolution

The term 'revolution' has evolved from a general notion of political 'turning-over' to refer more specifically to a rapid, fundamental transformation of the state and social structure, accompanied by mass uprising from below. Theories of revolution that emphasize structural conditions tend to focus on the revolutionary seizure of state power, while theories emphasizing actor motivations view revolution as an open-ended process of social transformation. This article examines the definitions and alternative explanatory frameworks, and considers the future of revolution in a changing global context.

1. *Conceptualizing 'Revolution'*

The earliest studies of comparative politics noted recurring conflict between elites and populace, sometimes leading to power 'revolving' from one group of rulers to another within a society. After the French Revolution of 1789, the term 'revolution' took on new meaning as a radical overthrow of the social and political order itself (Goldstone 1998). The nineteenth century European challenges to monarchic authority by the rising bourgeoisie and the class conflicts of industrial capitalism, punctuated by Marx and Engels' [1848] (1998) publication of *The Communist Manifesto*, focused attention on historical processes in analyzing revolution. After Lenin and Mao adapted Marxist theory to the Russian and Chinese revolutions (see *Marxism/Leninism*; *Maoism*), the decades following World War II saw a wave of national liberation movements against colonialism and neo-colonialism (see *Third World*; *Colonialism: Political Aspects*; and *Imperialism: Political Aspects*). Contrary to Marx's expectation that revolution would occur in the most industrialized capitalist countries, led by the proletariat, revolutions in the twentieth century occurred in the global periphery through mobilization of peasants and artisan workers.

1.1 *Common Usage*

The eclectic 'socialism' of many Third World revolts and nationalist movements, and their use of unorthodox tactics of guerrilla warfare—analyzed by practitioners such as Mao Zedong, Che Guevara, and Vo Nguyen Giap—led those characteristics to be generically associated with 'revolution,' as in Régis Debray's (1967) *Revolution in the Revolution?*. This usage was reinforced by Cold War counterinsurgency strategies aimed at forestalling radical change in national political systems, which were seen as extensions of the global East–West contest. The term 'revolution' moved away from this historically specific usage after the conservative religious turn in the 1979 Iranian revolution, the Eastern European uprisings against Communist Party states after 1989, and the end of the Cold War and other global power shifts.

Sometimes the term 'revolution' is used loosely to refer to any kind of paradigmatic shift, as in 'the Industrial Revolution,' 'the Reagan revolution,' 'the behaviorist revolution' in social sciences, or Thomas Kuhn's *The Structure of Scientific Revolutions* (1997). In another variant usage, 'revolution' is contrasted with reformism (see *Reform: Political*) to connote a radical political program, short time horizon, and willingness to use extreme (violent) means; as in Malcolm X's call for African-American liberation 'by any means necessary,' or J F Kennedy's warning that 'those who make peaceful revolution impossible

will make violent revolution inevitable' in Latin America.

1.2 Political Revolution and Social Revolution

The term 'political revolution' is employed to specify the application of the concept to a national political system. This usage does not include issue-specific social movements, nor all sudden changes of political power through *coup d'état* or civil war (see *Internal Warfare: Civil War, Insurgency, and Regional Conflict*; *Political Protest and Civil Disobedience*). However, 'political revolution' can refer to a wide range of revolts, rebellions, and mobilizational seizures of power—including those led from above such as the Japanese Meiji Restoration of 1868, Nasser's 1952 coup in Egypt, or the 1944–54 period in Guatemala—if they propose major lasting changes in the political order (Goldstone 1998).

Skocpol's influential work (1979, p. 4) proposed a more restrictive focus on 'social revolution':

> Social revolutions are rapid, basic transformations of a society's state and class structures, and they are accompanied and in part carried through by class-based revolts from below. Social revolutions are set apart from other sorts of conflicts and transformative processes above all by the combination of two coincidences: the coincidence of societal structural change with class upheaval, and the coincidence of political with social transformation.

Even accepting Wickham-Crowley's (1997) substitution of 'mass' for 'class-based' to include ethnic and populist upheavals, this defines a relatively rare phenomenon. This definition is widely used by scholars of revolution.

2. Alternative Ways of Understanding Revolution

Goldstone (1980) usefully grouped explanations of revolution into three generations of theory, with a fourth generation emerging since the late twentieth century. In the 1920s and 1930s, a first generation of 'natural histories' of revolutions used techniques of 'thick description' and inductive logic based on empirical observation of major Western revolutions. These efforts to identify key actors, factions, sequences, and patterns helped establish the phenomenon of revolution as a distinctive subject of comparative study; for example, Brinton's (1938) classic, *The Anatomy of Revolution*.

2.1 From Description to Explanation

In the 1950s and 1960s, the upheaval associated with decolonization led to a second generation of 'general theories of political violence.' Some focused on the individual psychological motivations of people experiencing frustrated expectations, as in the 'relative deprivation' approach later refined in T R Gurr's *Why Men Rebel* (1970). Frantz Fanon's *The Wretched of the Earth* (1965) interpreted political violence as redemption, healing the psychological wounds of colonization. Another branch of general theories sought explanations of political violence at the societal rather than individual level. Huntington's (1968) influential work drew on modernization theory, attributing revolutionary upheaval to economic modernization in developing countries, which raised expectations faster than they could be accommodated by the lagging development of political institutions (see *Modernization, Political: Development of the Concept*).

Critics of modernization theory in the late 1960s and 1970s pushed general theories of political violence in new directions. Tilly (1978) challenged the assumption that social systems were in natural equilibrium, arguing that conflict was normal, so explanations of political violence should focus on the relative resource mobilization capacity of power contenders. Moore (1966), in a sweeping macrohistorical comparative work, showed that there is no single, unilinear process of modernization. Extending Marx's class analysis to a more detailed classification of landlord–peasant relations, he combined this variable with distinct political forms of class struggle in agrarian societies undergoing commercial modernization, to map out different possible revolutionary outcomes.

Moore's (1966) work was the point of departure for Skocpol's (1979) ground-breaking study most associated with a third generation of 'structural theories' of revolution (see also her critique of Moore in Skocpol 1994). Shifting from the actor-centered focus that characterized general theories of violence, these theories highlighted the structural relations between units that created revolutionary opportunities. They built multicausal explanations based not only on class dynamics, but also on other aspects of the structures of society, the state, and the international system. Structural theories identified revolutionary conditions in the combination of administrative breakdown of the state, dissident elite political movements, a permissive world context, and widespread rural rebellion. Some like Skocpol and Goodwin 1994 focused on the types of state structures most vulnerable to revolutionary overthrow, including politically exclusionary authoritarian regimes (see *Authoritarianism*), neopatrimonial dictatorships, and directly ruled colonies. Goldstone (1991) examined the interaction of factors such as population growth with political institutions and social structure, developing quantitative indicators of revolutionary conditions. Structuralists defined revolution in terms of successful outcomes, distinguishing revolutionary seizure of state power from cases in which one or more necessary conditions were absent (Wickham-Crowley 1992).

2.2 Structure vs. Agency: Emerging Fourth-Generation Theory

An alternative to defining revolution as a set of structural conditions is to conceptualize it in terms of distinct systems of ideas and beliefs that subjectively mobilize actors. An emerging fourth generation of 'actor-centered' theories of revolution emphasize the defining role of ideology, culture, and identity of the subjects who make revolution. To borrow Mao's metaphorical observation that a single spark can ignite a prairie fire, these theorists shifted attention from the wind, temperature, and dryness of the grasslands to the spark itself. This contrasts with Skocpol's (1979) quotation of Wendell Phillips's statement that 'revolutions are not made; they come.'

Fourth-generation theories inspired new debates about whether revolutions can be accurately predicted on the basis of structural conditions (Keddie 1995), particularly in light of the surprising developments in the Iranian and Eastern European revolutions of the late twentieth century. Structuralists began re-examining the role of ideology in revolutions and the interaction of states with other domestic and international structures (Skocpol 1994). New insights from cultural studies and feminist theory challenged the more state-centered concepts of revolution (Foran 1997). These new perspectives defined revolution not as an outcome achieved by overthrow of the state, but as an ongoing process of transformation of society, involving both the institutionalization of a new order and the consolidation of a new ideology, guided with different degrees of success by different styles of leadership (Selbin 1999). As the Central American revolutions moved from armed conflict to a new dynamic of mobilization and negotiation, the idea of a revolutionary outcome or endpoint seemed less useful than the concept of revolution as process.

2.3 Rethinking Revolution in the Context of Globalization

The emergence of late twentieth century movements that both resembled and differed from earlier waves of 'revolution' suggested a need to rethink the concept. Some interpreted uprisings such as the Philippine 'people's power' movement and the Czech 'velvet revolution' as part of a new, democratic revolutionary wave. Others drew on world-systems theory to suggest that revolutions should be defined not as discrete phenomena of nation-states, but rather as transformations in the world system (Boswell and Chase-Dunn 2000). Thus the 'world revolution' of 1989 might be seen as an outgrowth of the shift in the capitalist world system, from the Fordist regime of capital accumulation based on industrial mass production to a regime of flexible accumulation, now regulated by global institutions governed by neoliberal ideology.

When Nicaragua's revolutionaries made history by turning over the state power they had won through insurrection in 1979 after losing elections in 1990, Sandinista leader Daniel Ortega vowed they would 'govern from below.' This vision, albeit unrealized, suggested a new way of thinking about where the essence of a revolution is found. The demise of state socialism fueled the search for radical democratic alternatives, rooted in autonomous action within civil society. As globalization undermined territorial definitions of social and political organization, shifting decision-making authority to transnational structures (see *Globalization: Political Aspects*), revolution seemed less about seizing state power. From the Internet-savvy Zapatista rebels of Chiapas, Mexico to the transnational antisweatshop movement and the 1999 Seattle protest against the World Trade Organization, there were intriguing signs of an emerging global civil society. These new forms of resistance against globalization stretched the boundaries of existing concepts of 'revolution.'

See also: American Revolution, The; Chinese Revolutions: Twentieth Century; Communism, History of; Cuban Revolution, The; English Revolution, The; French Revolution, The; Globalization: Political Aspects; Meiji Revolution, The; Mexican Revolution, The; Revolutions, History of; Revolutions of 1989–90 in Eastern Central Europe; Revolutions, Sociology of; Revolutions, Theories of; Russian Revolution, The; Social Movements, History of: General; Vietnamese Revolution, The

Bibliography

Boswell T, Chase-Dunn C 2000 *The Spiral of Capitalism and Socialism*. Lynne Rienner Publishers, Boulder, CO

Brinton C 1938 *The Anatomy of Revolution*. Vintage Books, New York

Debray R 1967 *Revolution in the Revolution*? Grove/Atlantic, New York

Fanon F 1965 *The Wretched of the Earth*. Grove Press, New York 1965

Foran J (ed.) 1997 *Theorizing Revolutions*. Routledge, London

Goldstone J A 1980 Revolution from above – military bureaucrats and development in Japan, Turkey, Egypt, and Peru-Timberger, Ke. *World Politics* **32**: 425–53

Goldstone J A 1991 *Revolution and Rebellion in the Early Modern World*. University of California Press, Berkeley, CA

Goldstone J A (ed.) 1998 *The Encyclopedia of Political Revolutions*. Congressional Quarterly, Washington, DC

Gurr T R 1970 *Why Men Rebel*. Princeton University Press, Princeton, NJ

Huntington S P 1968 *Political Order in Changing Societies*. Yale University Press, New Haven, CT

Keddie N R (ed.) 1995 *Debating Revolutions*. New York University Press, New York

Kuhn T 1997 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

Marx K, Engels F 1998 *The Communist Manifesto*, 150th anniversary edn. with essay by E Hobsbawm. Verso, London
Moore B Jr. 1966 *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon Press, Boston, MA
Selbin E 1999 *Modern Latin American Revolutions*, 2nd edn. Westview Press, Boulder, CO
Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge University Press, Cambridge, UK
Skocpol T 1994 *Social Revolutions in the Modern World*. Cambridge University Press, Cambridge, UK
Skocpol T, Goodwin J 1994 Exploring revolutions in the contemporary. Third World. In: Skocpol T (ed.) *Social Revolutions in the Modern World*. Cambridge University Press, Cambridge, UK
Tilly C 1978 *From Mobilization to Revolution*. Addison-Wesley, Reading, MA
Wickham-Crowley T P 1992 *Guerrillas and Revolution in Latin America: A Comparative Study of Insurgents and Regimes since 1956*. Princeton University Press, Princeton, NJ
Wickham-Crowley T P 1997 Structural theories of revolution. In: Foran J (ed.) *Theorizing Revolutions*. Routledge, London

R. Stahler-Sholk

Revolutions, History of

This brief article on the general problems of studying revolutions is divided into five sections. The first will discuss the term 'revolution' as an historical phenomenon which has changed its meaning over time with important consequences. The second will discuss the issue of taxonomy or classification, attempting to exclude upheavals which are not really revolutionary, and to distinguish different types of genuine revolution. The third will discuss etiology or origins, examining the attempts of historians, political scientists, and sociologists to find common causes of revolution. The fourth will discuss morphology or the possibility of finding a paradigm. This section will discuss the problems of drawing a graph which has validity for the all of the great revolutions of modern times. This subsection will question the appropriateness of trying to place the course of major revolutions on a grid showing extreme right, right, center, left, and extreme left. The fifth subsection will make some concluding remarks.

1. The Evolution of the Term Revolution

Only a handful of historians and social scientists have paid much attention to the history of the term 'revolution,' despite the fact that this history casts light on important developments since the Renaissance. The term emerged from an obscure Latin word meaning the periodic return of a moving object to its point of origin. Throughout the Middle Ages it was used in its Latin, French, and English forms to describe the supposed cyclical movement of celestial bodies around the Earth. Then it was popularized by heliocentrics following Copernicus's *On the Revolution of the Heavenly Orbs* in 1543. Galileo and his followers used the term repeatedly. Contemporary dictionaries reflected the domination of the term by scientists. In his famous French–English dictionary published in 1611, Randle Cotgrave defined the word as 'a full compassing; rounding, turning back to its first place, or point, th'accomplishing of a circular course.' Significantly, he gave no non-scientific meanings.

Gradually the meaning of the term began to expand. Poets and playwrights had already begun to appropriate the word. In *The Compleynt of Mars*, (1.03), written c. 1378 Chaucer speaks of 'hevenish revolucioun.' Then at the beginning of the seventeenth century Shakespeare pre-empted the word to describe his view of death. In *Hamlet*, when the grave-digging clown unearths a skull, the prince of Denmark reflects that it might have belonged to a politician, a courtier, or a lord: 'here's a fine revolution, an we had the trick to see't (v.1).' Shakespeare thus used the word to mean the return to one's original condition, earth to earth, ashes to ashes, dust to dust.

While literary men were using the term 'revolution' to mean a cycle, political commentators were already using it to mean a political upheaval. Several Italian authors used the word during the Renaissance to describe political turns of fortune, the replacement of one prince by another, in the city-states of Italy. Then in 1674 Alessandro Giraffi published a book entitled *Le Rivolutioni di Napoli* to describe the rebellion of the Neapolitans against the Spanish viceroy. Another Italian and a Frenchman used the term to describe the same event. Moreover, Giraffi's book was translated into English within a few years.

In the UK most writers referred to the struggle in the middle of the seventeenth century between parliament and the king as the Great Rebellion, but some authors on both sides referred to it as a revolution, for instance the radical parliamentarian Anthony Ascham and the royalist Matthew Wren. In a letter to the Earl of Dorset in January 1646, the royal historiographer James Howell used the word to mean both a political upheaval and the cycle of the years: 'for within these 12 years there have been the strangest revolutions and the horridist things happen'd not only in Europe, but all the World Over, that have befallen mankind, I dare say since Adam fell, in so short a revolution in time' (James Hollowell, 1890. *Epistolae Ho-elianae*, Bk. III, 512). His contemporary John Milton also used the word in these two senses (*Paradise Lost*, II, Lines 597–603; VIII, lines 15–31).

Although the radical parliamentarians in mid-seventeenth-century England claimed that they were restoring ancient rights, they already had some

inkling of a new beginning. After the overthrow of the monarchy they spoke of Year I of Liberty, Year II, Year III, and so on for a few years. Some radicals even had a vision of society completely transformed. 'But why may we not have our Heaven here,' Gerrard Winstanley asked, 'and Heaven thereafter to?' (Sabine, 1941, p. 409). Nevertheless the idea of reversion persisted: even Winstanley was thinking of a return to a primitive simplicity. The royalists also used the word to mean a return to a previous state. For William Temple the return of the Stuart dynasty in 1660 and the Orange dynasty in The Netherlands in 1672 were revolutionary events in the cyclical sense of the word. Although in the second of his *Two Treatises on Government*, published in 1690, Locke defended the right to revolution, he used the word only twice, each time to mean a return via the state of nature to a previously existing constitutional position (1964, Cambridge edn. Paragraphs 223 and 225, 432–3).

Oddly enough it was not the Great Rebellion of mid-seventeenth-century England which made the word an important term in political vocabulary. Only after the Glorious Revolution in 1688 did lexicographers assign the word a political meaning. The use of the word to describe the overthrow of James II and the signing of the Bill of Rights by William and Mary gave the word a wider meaning. It was not quite a restoration of a previously existing order, rather it consolidated and advanced rights for which parliament had fought for the preceding century. Nevertheless, it was a very conservative event, entrenching the power in parliament of aristocrats, landed gentry, and rich merchants. Also, the Revolution was not seen as something with a universal application; it was viewed as English, unique, unrepeatable. The Bill of Rights of 1688 was not for all humanity.

Then in the last quarter of the eighteenth century the term came to mean a profound rupture with the past and the beginning of a new era. On the Great Seal of the US we see 1776 associated with the Latin motto *Novus Ordo Seclorum*, 'a new order in the ages.' Then the French Revolution consolidated and intensified the idea of a new beginning in history. Like the Commonwealth-men earlier, they renumbered the years starting in 1789. When the Republic was established in 1792 they began to number the years Year 1 of the Republic and so on. Finally, in October 1793 the Convention introduced a new calendar numbering the years, not from the birth of Christ, but from the inauguration of the Republic on September 22 of the previous year. Even more significant was that the French revolutionaries believed that their ideals would have a universal appeal. The three Declarations of Rights—1789, 1793, and 1795—were to apply to all men, not just Frenchmen.

Because of the cluster of revolutions in the late eighteenth and early nineteenth centuries revolution became recognized as a recurrent phenomenon. This in turn led to the appearance of a new character on the world stage: the professional revolutionary. There was an apostolic succession of such characters—Filippo Buonarroti (1761–1837), Auguste Blanqui (1805–1881), Mikhail Bakunin (1814–1867), Piotr Kropotkin (1842–1921), Victor Serge (1891–1947), Sun Yat-sen (1866–1925), Mao Zedong (1893–1976), and many others. These life-long revolutionaries anticipated revolution, created organizations to bring it about, and formulated plans about what to do when it arrived. Thus was born the revolutionary political parties which changed the course of modern revolutions. Once such parties capture power, they are able to suppress opposition for very long periods. Until these parties are finally dislodged from power changes in policy can only take place by decisions within the party leadership.

2. *Taxonomy*

Many upheavals are labeled revolutions, but few actually qualify for the term. Most *coups d'etat* do not qualify since participation in them is limited and their result is often confined to a change in government personnel. So-called 'palace revolutions' usually involve struggles with factions or families within the government. Rebellions should probably be excluded, not only because they do not succeed, but also because they usually lack a clear program and are often led by nobles who do not really want to destroy the existing political and social order. Even wars of independence are dubious candidates for the term, contrary to conventional usage. Even the American Revolution scarcely deserves this title. Although the War for Independence saw individual rights entrenched, considerable exchange of property, and a large emigration, the internal institutions of the new states were not radically different from those of the former colonies. Some African struggles in the twentieth century against European imperialist powers had arguably deeper internal repercussions.

Even among upheavals which deserve the title there are many degrees of revolution. There are political revolutions such as the Revolution of 1830 in France, which produce a change in the groups enjoying power, but do not shake society up very much. The great vertical upheavals or social revolutions such as the French Revolutions of 1789, the Russian Revolution of 1917, or the Chinese Communist Revolution of 1949, have led to deep social changes. The latter two, however, are not only political and social revolutions, but socialist ones as well. The Babouvist Conspiracy of Equals in France in 1796 was probably the first case of a group, in which some members supported a primitive communist ideal, which actually tried to seize power. Genuine socialist revolutions, attempting to control and redirect the means of production, could obviously not occur until socialism developed in the nineteenth and twentieth centuries.

Revolution in the deepest sense involves not only a change in the groups exercising power, profound changes in the social structure, and in modern times an attempt to change the control over the means of production, but also an attempt to change the way citizens think, feel, and act. There are glimmerings of such an attempt in the Hussite and Puritan Revolutions, but the French Revolution of 1789–1799 was the first to try to implement it in a large way. At the peak of the Revolution in Year II the *gouvernement révolutionnaire* tried to 'regenerate' citizens through an attempted mobilization of all the existing media: art, architecture, festivals, music, newspapers, theatre, and children's books and games. Such use of the media has since become a feature of major revolutions on both the left and right, such as Nazi Germany, Bolshevik Russia, and Communist China. Comparative studies of revolutions seldom discuss this added dimension to revolution.

3. *Etiology*

There has been great attention to the question of the origins of revolution. If it were possible to establish common causes for revolutions, it would greatly simplify the study of the great revolutions since the seventeenth century, each of which has produced thousands of studies. Once common causes were identified, this knowledge would also be useful both to those anxious to promote revolutions and those intent on preventing them. The hope of applying social science for such purposes has produced such books as Michael Velli's *Manual for Revolutionary Leaders* published in 1972, and on the other side the collection edited by Harry Eckstein (1963), intent on curbing such internal violence.

Since Aristotle there has been wide discussion of the etiology of upheavals, rebellions, and revolutions. Most historians, however, do not find the literature comparing revolutions of much use in understanding the revolutions in which they specialize. Many of them point to the problem of assessing the weight to be given to various factors at work, before and during revolutions. Unlike scientists who can run experiments in which all factors except one are held constant, historians cannot rerun events to test the role of certain factors. An historian, for example, cannot demonstrate precisely how the closing decades of the eighteenth century in France would have been different if reformers had been able to solve the financial problems, if those problems had not been exacerbated by intervention in the American War for Independence, if there had been a different King from Louis XVI, if he had not called royal troops into Paris in July 1789, if he had not tried to flee the country in June 1791, if the Girondins had not led the country into war in 1792, or if Robespierre had never been born. Similar questions can be posed about all subsequent revolutions. Specialists try to weigh various factors by getting as deeply as possible into the complexities of each upheaval.

Aristotle argued that rival concepts of justice led to upheaval. Modern theoreticians such as Karl Marx, Max Weber, Pitirim Sorokin, Lyford Edwards, Jaroslav Krejci, Mark Hagopian, Peter Calvert, Chalmers Johnson, and host of others have focused on increasing tensions in societies leading to revolutions, provided there are precipitants and triggers. These tensions are described in various way including class conflict, other contradictions of various kinds, frustration of the expectations of certain groups, dissynchronization between social development and existing values, and other sources of disequilibrium. The fact that there have to be severe tensions in society in order to have an internal war seems a truism. The difficulty with such generalizations is in applying them to very different circumstances. Take, for example, the varying role of war. The American Revolution was made possible in part because the threat of invasion from French territories to the north had been removed by British victory in the Seven Years War. France was not at war when revolution broke out in 1789. In contrast, the government in Russia had been severely undermined by war by 1917. Also in China the long struggle with Japan contributed to the collapse of the Guomindang and the Communist victory in 1949.

Another important variation among revolutions is their very different objectives. As was mentioned earlier, there were no socialist revolutions before the twentieth century. Twentieth-century revolutions in Russia, China, Cuba, and elsewhere have attempted to seize and redirect the means of production. Related to this has been the attempt of countries such as Russia, China, and North Korea to force development in order to catch up with Western Europe and North America. Such a drive was not present in the Hussite, English, or French Revolutions, in fact the hope for a sudden increase in productivity could scarcely have arisen before industrialization, stream power, electrification, and the internal combustion engine. Comparing revolutions over five or six centuries seems to minimize new ideologies, in fact novelty in human history.

One way of looking at the objectives of various revolutions is to study how they were symbolized. In the French Revolution the dominant symbols were the figure of Liberty, which gradually evolved into the figure of the Republic, the Liberty Bonnet, and the Liberty Tree, none of which reappeared in the Bolshevik Revolution. Russian revolutionary symbolism featured giant figures of workers holding Hammers and Sickles, workers depicted as part of industrial machinery, electric lights in modernized villages, and of course red stars and flags. In Maoist China there were no symbols of liberty, but thousands of images of Mao like a red sun in the sky, numberless little red books of Mao's writings, and endless images of peasants using muscular power to create terraced

hillsides, build dams, canalize rivers, and irrigate the lands. In fact their terraced hillsides became a symbol of using intensive labor to 'conquer nature.' Machinery was relatively inconspicuous until after Mao's death in 1976. The figure of Liberty was absent until students made her prominent briefly in the abortive democracy movement of 1989. Comparative studies of revolution have so far largely ignored different allegorical figures and symbols.

4. *Morphology*

There are similar problems with the attempt to find common patterns in the course of revolutions. One of these is that most historians, sociologists, and political scientists give much more attention to the origins of revolutions that to the inner dynamics which drive them forward, producing the curves so dear to social scientists. A further problem is how long a period one considers a curve. Jaroslav Krejci extends the Hussite Revolution over 50 years, the English Revolution over 60 years, and the French Revolution over 100 years. Can one speak of a single French Revolution over a century which saw two empires, three forms of monarchy, five revolutions (if one considers 1792 a revolution within the revolution of 1789–99), and numerous coups? Also, if you compare his graphs of the six revolutions that he covers—the Hussite, the English, the French, the Russian Bolshevik, the Turkish, and the Chinese Communist Revolutions they vary considerably from each other (1983, pp. 35, 67, 82, 117, 149). Moreover, his curves vary considerably from those published by Mark Hagopian (1974, p. 245). Also it is strange that Hagopian shows the Chinese Revolution dipping in 1958, the precise date of the Great Leap Forward. Obviously reducing great revolutions to scientific-looking graphs is a risky endeavor.

A further problem arises from placing revolutions on a grid marked extreme right, right, left, and extreme left. These terms arose from the accidental seating of conservative deputies to the right of the speaker in the Constituent Assembly in the French Revolution of 1789. In the liberal democratic revolutions in Europe and North America the left has usually meant those who supported genuine parliamentary government, political opposition, and defense of human rights. Can one, however, put the English, the American, and the French Revolutions on the same right-left grid as the Russian, Chinese, and Cuban Revolutions? Placing modern communist revolutions on the left distorts the original meaning of the term. These twentieth-century revolutions have introduced single-party regimes, have created assemblies which approve decisions already made by the leaders, and have violated human rights. In fact the way they have operated is very similar to governments on the extreme right, '*les extrêmes qui se touchent*,' as the French put it. This is another example of the perils of applying the same grid to revolutions with very different objectives and consequences.

5. *Conclusion*

While it seems very difficult to establish a common taxonomy, etiology, and morphology covering all the major revolutions in the last five centuries, this does not mean that the comparative study of revolutions is not worthwhile. The history of the term 'revolution' demonstrates that it has taken on important new dimensions since it first became politicized in Renaissance Italy. The taxonomy of revolutions can exclude upheavals that are not full revolutions and can distinguish among political, social, and socialist revolutions, or combinations of all three. The etiology of revolutions is particularly problematical because the conditions that occur vary widely, nevertheless the idea that disequilibrium of some kind precedes revolution seems a valuable one. In many cases revolutions occur when society is changing while the regime stands still, but in the case of the Iranian Revolution of 1979 the reverse seems true, disequilibrium resulting from a westernizing shah moving more quickly than powerful conservative groups, especially religious ones, were willing to accept. On the other hand the collapse of the satellite Communist regimes in eastern Europe after 1989 seems to be the result of the failure of those governments to satisfy the needs of their citizens and the weakness of systems imposed by external force. As for establishing a common pattern or morphology, the differences in the graphs which scholars have produced warn of the dangers in such attempts. Those historians, however, who steep themselves in just one revolution, are very unlikely to recognize what is unique about the particular upheaval in which they specialize.

Since revolution has been a recurrent phenomenon, and may be so in the future, social scientists will doubtlessly attempt to reach general conclusions about them. In such future attempts more attention needs to be given to the internal dynamic of revolutions once they have broken out. The idea of attempting to create a new citizenry by mobilizing all the available media, which has marked revolutions since 1789, deserves further study. So too does the use of symbols, allegorical figures, and political images which have distinguished revolutions in modern times. Moreover, graphs based on a left-to-right scale need to be modified because the meaning of left and right has shifted drastically over time. Finally, the comparison of revolutions must bring out the unique dimensions of each revolution instead of focusing mainly on similarities.

See also: Communism, History of; Conflict: Anthropological Aspects; Conflict Sociology; Crowds in

History; Internal Warfare: Civil War, Insurgency, and Regional Conflict; Marx, Karl (1818–89); Revolution; Revolutions of 1989–90 in Eastern Central Europe; Revolutions, Sociology of; Revolutions, Theories of; Social Movements, History of: General; Social Movements, Sociology of; Social Protests, History of

Bibliography

Brinton C 1960 *The Anatomy of Revolution*. Vintage Books, New York

Brunno O, von Conze W, Koselleck R 1972 *Geschichtliche Grundebegriffe*. Klette, Stuttgart, Germany

Close D, Bridge C (eds.) 1985 *Revolution: A History of the Idea*. Croom Helm, London

DeFronzo J 1999 *Revolutions and Revolutionary Movements*, 2nd edn. Westview, Bolder, CO

Dunn J 1989 *Modern Revolutions: An Introduction to the Analysis of a Political Phenomenon*. Cambridge University Press, Cambridge, UK

Eckstein H (ed.) 1963 *Internal War. Problems and Approaches*. Free Press, Glencoe, NY

Ellul J 1971 *Autopsy of Revolution*. Knopp, New York

Goldstone A 1991 *Revolution and Rebellion in the Early Modern World*. University of California Press, Berkeley, CA

Hagopian M N 1974 *The Phenomenon of Revolution*. Dodd Mead, New York

Johnson C 1964 *Revolution and the Social System*. Hoover Institution Studies 13, Standford University, Stanford, CA

Krejci J 1987 *Great Revolutions Compared*. Harvester Press, Brighton, UK

Leith J A 1974 *Media and Revolution*. Canadian Broadcasting Corporation, Toronto, Canada

Lipsky W E 1976 Comparative approaches to the study of revolution: A historiographical essay. *Review of Politics* **38**(4): 494–509

Pettee G 1938 *The Process of Revolution*. Harper and Row, New York

Sabine H (ed.) 1941 *Works of Gerrard Winstanley*. Cornell University Press, Cornell, NY

Sorokin P 1967 *The Sociology of Revolution*. Howard Fertig, New York

Velli M 1972 *Manual for Revolutionary Leaders*. Blade and Red, Detroit, MI

Zagarin P 1973 Theories of revolution in contempory historiography. *Political Science Quarterly* **88**(1): 23–52

J. Leith

Revolutions of 1989–90 in Eastern Central Europe

1. A Deviant Type of Revolution

The revolutions in Eastern Europe were no conventional changes of the system that can be dealt with in the tradition of sequence models. There was no uniform anatomy of the changes, though never in European history have so many regimes collapsed at the same time and for similar reasons. Revolution was traditionally linked to violent change that made it different from coups d'état. Nevertheless, the depth of change made the events between 1989 and 1991 a political and a social revolution at the same time. In this respect it was comparable only to the Bolshevik revolution of 1917 and its later repercussions in the 'people's democracies' after 1945 in the realm of influence of the Red Army. This double revolution made consolidation of the new 'revolutionary regime' so difficult.

No former system's change was as burdened with the same problems of institution-building as this one. Authoritarian and fascist dictatorships may have changed the political institutions, but they hardly touched the system of stock markets and banks. There was no need in post-authoritarian regimes to create offices to administer unemployment insurance nor institutions for social security.

This process of transformation was so unique that the old notions of revolution and reform tend to be inappropriate. Perestroika in the Soviet Union was frequently compared to the politics of Peter the Great, the Meiji Revolution, or the innovations of the Young Turks in the Ottoman Empire. Reform was defined as an eclectic un-ideological process of un-dogmatic actors. The consequence was contradictory measures and even 'messiness' of reform. *Ex post facto*, perestroika proved its messiness. But was it really not ideological? This author prefers to list perestroika under the type of reforms from above, steered by the ideological thrust of a great leader.

In Gorbachev's self-perception, perestroika was a revolution. The old Soviet Union under Brezhnev started to accept rather loose talk about revolutions. Revolution lost its connotation of a process intended by political elites and by political means. It was watered down to processes that the Central Committee and the Politburo of the party were no longer able to control, thus demographic, environmental, and technological revolutions entered the discourse in communist countries. Gorbachev tried to use revolution again in a more political way. Perestroika was dubbed a 'second revolution' and Gorbachev hinted at Lenin who had already envisaged several revivals of revolutionary activities in the tradition of the great French revolution that failed to implement its goals in one coup and had to be repeated in 1830, 1848, and 1871. Western analysts, on the other hand, never used the term revolution in describing the process of perestroika.

Only at the end of perestroika had the term system's change found universal acceptance. It lacked the connotations of violence and civil war implied in the term revolution, but is nevertheless aimed at a far-reaching change of politics, society, and economics. The minimal criterion is a break-down of the old

system. This does not need, however, revolutionary force.

The revolutions in Eastern Europe were different from former revolutions because—with the exception of Romania—never have so radical changes been achieved with so little violence from below and from above. Eastern Europe proved to be a deviant case in the history of revolutions for several reasons:

Political modernization was achieved in two different phases. The system's change was a second attempt at modernization after the socialist attempts in 1917 and 1945.

The collapse of socialism resulted from a threefold dissatisfaction of East European populations: dissatisfaction with a repressive political system, with an inefficient economic system, and with the dominance of hegemonic ethnic groups within a federation (Russians, Serbs, and Czechs) or within the socialist camp (the Soviet Union vis-à-vis its satellites).

There was not the same continuity of periods as in Southern Europe in the 1970s with phases such as liberalization, democratization, and consolidation.

The new process of transition was burdened with the unique double task of transforming the political and the economic system at the same time, something which has never before happened in world history.

The new transition process was hampered by ethnic strife, which occurred to an unknown extent. Spain had experienced some of it, but managed to solve the problem by various autonomy concessions.

In former transitions, a new elite was ready to take over power. There was no real preparation for the new task in Eastern Europe. Even non-Marxist scholars sometimes believed in Marxist periodization schemes. They held that a system needed to develop an economic basis and a recruiting ground for the new political elite. The old regime developed the preconditions of change. Revolution was only the last blow for the old system. The simplicity of this concept of transitions did not even work in the great French revolution. Historians have shown that the French revolution did not abolish feudalism or establish a market society throughout the country. This process of accomplishing a true market society was not successful before the beginning of the Third Republic (1875 ff).

2. *Reasons for the Velvet Revolutions*

First, in the history of changes of regimes the leadership groups have frequently disunited. One part of the power elite started mobilizing the masses. In Eastern Europe this model did not apply. The elites remained quiet until the last moment. The technical intelligentsia did not develop into an opposition to the 'strategic clique' in power, as had been forecasted. The technical intelligentsia increasingly de-ideologized and created more room for maneuvering in its own field, independent of political leadership. It avoided conflict, however.

Reformist elites in those socialist systems where the regime eroded slowly and revolution was pacted in a corporalist mood also avoided the appeal to the masses. In Hungary, the negotiated revolution took place with complete discipline and there was no unified counter-elite. In Poland, Solidarity could have been the counterforce, but in the last stage it was pressed by the people rather than directing the masses into conflict with the 'nomenclature' leadership.

Only in Romania—in the third model of a collapse of the system, but under control of a middle-level elite of the old regime—was there conflict among elite sectors. The military, after some wavering, finally sided with the protest movement, not so much because of ideological sympathies with the revolutionaries, but because of its rivalry with the Securitate armed forces. The military played the role of a rear guard—blamed by Lenin as '*chvostizm*'—but apparently the capacity to wait and see and to grasp the right moment for action in the rear guard is not always a negative feature, as Lenin supposed.

Second, the European fascist dictatorships in 1945 collapsed because they were defeated in a war. In Southern Europe and Latin America external factors contributed to the collapse of a regime (Greece, Portugal, and Argentina). In Eastern Europe the changes of the international situation facilitated the peaceful velvet revolutions. The Soviet Union still existed as a hegemonic power, but Gorbachev had indicated that the satellites could no longer rely on Soviet help in a case of internal strife. It was less the active behavior of a superpower, than the non-decisions of the hegemonic bloc power that did not intervene, which caused the collapse of the whole alliance. Theoretically, each national elite could have followed the Chinese example of bloody suppression. Why did this not happen?

The most important reason is again the international factor: the 'Brezhnev doctrine' was meant to strengthen weak regimes within the bloc. But the bloc center at Moscow was not prepared for a situation where all the leaderships of her satellites entered into troubles at the same time. The Brezhnev doctrine proved its dysfunctional consequences. It had spoiled the satellite leadership because they were always able to rely on big brother's tanks. Only Romania was the deviant case which did not recognize the doctrine and therefore had to provide for her own security. All of the other governments were not accustomed to big actions for their own security and caved in as soon as the masses confronted them directly.

The old regime in its last defensive stage pretended that the turmoil was caused by subversion from the 'imperialist powers.' This was wrong, since the Western powers were so accustomed to détente that they avoided direct actions for destabilizing the communist systems. It was not international but rather transna-

tional policies which contributed to the destabilization of communism from outside. The protest movements in the streets of Leipzig did not receive weapons, but got information instead from the Western media about the situation in the country, the counter-reactions of their leadership, and the points of rallying for the protest movement.

The international factor could explain certain framing conditions of the development, but nobody would have dared to forecast that the Soviet Union was so occupied with her internal problems that she had no capacity for intervention abroad.

Third, the most complex explanation of regime breakdown is the loss of legitimacy. In socialist countries legitimacy could not be measured by surveys. The mass media in the West, such as *Free Europe* and *Liberty* have tried to measure legitimacy by the instrument of mail from the East, but results were unreliable. In the early 1980s, it was calculated that the communists in free elections would not get more than 3–5 percent of the votes, whereas they actually received 10–15 percent in the most developed systems and much more in the less-developed communist regimes in the southeastern parts of Europe.

Was socialism ever legitimized? Its foundation by force was not forgotten by large parts of the population in the satellite countries. Acceptance of the regime was concentrated on the cadres. With growing success and material benefits, however, legitimacy grew also under communist conditions.

Paradoxically, communism collapsed due to its partial success, especially in the sphere of producing intellectually trained elites. There was a mobilization for education, which hardly had an equivalent in the Western world. Communist systems produced a surplus of elites, whom they were unable to funnel into meaningful positions. Western democracies do the same, as periodical overproduction of elites seems to be as inevitable as the famous pig cycle, causing economic slumps in the economic and political conflicts. Western systems, though, can offer jobs for the new elites in the third sector of services when the elites cannot be integrated into the bureaucracy. This tertiary sector was chronically underdeveloped in socialist systems. Only the systems that were able to coopt the dissatisfied young elites, such as Hungary, experienced a slow transition to the new regime. Here the rebellion could be dubbed as the 'revolution of the deputy heads of administrative units.'

It would be unhistorical to look for a legitimacy of the whole life span of socialism. Legitimacy had also ideological foundations, and beliefs in legitimation are subject to change. In the early days of revolutionary socialism, the system claimed a kind of teleological legitimation because the party pretended to know the outcome of history. In the Soviet Union this was certainly combined with charismatic legitimation in the time of Lenin. As time went on, however, both withered away, charisma as well as the capacity to predict the development of history. Quite frequently, the people experienced false prognoses that undermined the belief in the ideology. As Max Weber predicted, the personal charisma of the leader did not survive the founding father. *Veralltäglichung*, the erosion of charisma, was the necessary consequence. Stalin tried to substitute for his lack of charisma through organization and ideological leadership in all questions: from literature to biology. But this kind of mixed legitimation was increasingly built on force. With diminishing capacities of the regime for repression and mobilization, new forms of legitimation had to be found. Communist repression had to be substituted by more legal security. In the early days, huge groups of defeatists and subversive elements were discovered. After Stalin, nobody believed in this kind of singling out of groups any more. Only 'parasitism' and '*chuliganstvo*,' spheres where the Soviet citizen felt endangered for his own security, was he ready to buy from the Communist propaganda.

Communism failed to establish a certain balance between the four basic principles of legitimization in Western systems. The legal state was violated by arbitrary decisions of special hierarchies in the party, in the security forces, and in the planning administration. The nation state was increasingly invoked by official propaganda in order to reinforce socialist ideology, but the hegemony of the Soviet Union prevented the development of authentic independent national states, even if more systems tried to embellish the socialist dogma by the national colors. The democratic state was proclaimed by state propaganda, but in reality, for the citizens it was only manipulated participation without alternatives in the political arena. Certain ultra-democratic requisites such as 'recall' and 'imperative mandate' were operating only with a hint of the party, never because of a free decision of the citizens. Thus, the welfare state was used to legitimize the late socialist systems. In the early days, communist propaganda thought that a social state was unnecessary, because a socialist state is socially minded *per se*. But the leveling down of all social differences was no longer accepted by the leadership. Stalin had already denounced *uravnilovka* (excessive equalization).

During the years of economic growth, Hungarian Goulash Communism or Polski Fiat Communism developed a certain legitimization of the system, but in the 1980s it became clear that the communist countries were falling further behind the West. Communication with Western countries grew, which contributed to a growing capacity of the people to compare East and West. During the 1980s, the system lost its legitimacy due to the economic crisis of socialism, which was felt in most of the countries.

Transition studies on a quantitative basis confronted us with the finding that democratization is impossible as long as the power resources of a regime are highly concentrated. This assumption was not

Table 1
Pathways to democracy

	Pragmatic piecemeal engineering	Ideological innovation
Control from above	Opening of the communist regime (Romania, Bulgaria)	Innovation of socialism (Perestroika under Gorbachev)
Pressures from below	Erosion of Socialism (Poland, Hungary)	Collapse of socialism and takeover by the opposition (GDR, CSSR)

even correct in the case of the communist countries. All of them had the power to suppress the opposition and had in the 1980s increased their service in both staff and technological equipment. Paradoxically, some of the socialist countries, such as the German Democratic Republic (GDR) invested most of their technological innovation in the social control of the population instead of investing it in production. Nevertheless, the system collapsed. The enormous security equipment did not help the system survive: as was the case with the Chinese Wall, Rome's Limes in the North of the Empire, or France's Maginot line in 1940. Unlike these cases, the fabulous equipment was not even applied in the last conflict. The index of power resources of the old regimes proved to be of minor importance in the fourth wave of democratization. On the contrary, the better equipped a communist regime was with repressive capacities, the more radical was the change.

Three models predominated in the system's change of 1989 (see Table 1):

The erosion of the regime by corporatist bargains between the power center and the opposition in a negotiated revolution. This type of transition caused a low rate of participation among the people, because the change was pacted by elites without participation of the citizen. Many voters apparently did not even realize the amount of change and the voter turnout stagnated between 50 and 60 percent (Poland, Hungary).

The implosion of the communist regime (Czechoslovakia, GDR). In this model, the control of the ancien regime was least developed. A new elite took power. Continuity of elites and institutions was lower than in the first path to democracy.

System change was achieved by the old cadres of the second rank who kept control of the transition (Romania, Bulgaria (until 1991), Albania (until 1992), Serbia, Croatia, and most of the Commonwealth of Independent States (CIS) throughout the first part of the 1990s).

Change proved most complicated in all the multiethnic states. The disintegration of the Soviet Union and Yugoslavia showed elements of a 'bargained revolution.' Bargaining, however, was done not by functional groups but by territorially old and new governments. In no case did the old federation survive and in no case was a new federation feasible after the collapse of the old regime.

Transition theories—in the tradition of theories about the developments of the 1990s—were most interested in the dichotomy between pact and dictate. This dichotomy was confronted with reform and revolution in a matrix (Karl and Schmitter 1991). Most of these matrices only offered a realistic picture of the initial phase of transition.

3. *Democratization and Consolidation of the Peaceful Revolutions*

'Transitology' as a new branch of knowledge dealing with the peaceful revolutions since the 1970s and 1980s originally used a sequence model along the lines: liberalization, democratization, and consolidation. Only in the model of transition via erosion of the communist rule did a liberalization take place. The period of democratization is normally identified with institution-building (constitutional power and secondary institutions such as parties and interest groups). This process was successful in most Eastern European countries but hardly in the successor states of the Soviet Union. Consolidation of elite attitudes (the 'only game in town' criterion developed by Juan Linz) and the citizens is incomplete in most countries. The Vicegrad states, which formed a unit of economic cooperation, and those which entered negotiations with the European Union (Poland, Czechia, Hungary, Slovenia, Estonia) are fairly consolidated even in terms of economic indicators. But also other systems—-outside the CIS—have consolidated by experiencing several free elections, by reducing the volatility of votes, by avoiding violence, and reducing the weight of extremist parties by practicing alternating governments that included even the acceptance of a comeback of the post-communists that turned in many countries to democratic socialism (Lithuania 1992, Poland 1993, Hungary 1994). Most countries have changed the composition of their governments from The Czech Republic (1998) down to Romania (1996) without turning the critical election into a crisis of the system.

Not all the processes of consolidation were successful, however. 'Consolidology' as a new branch in the study of democratization discovered various types of unconsolidated democracies that have been called 'delegated democracies' or 'defective democracies'. Most of these defective democracies have implemented

a rather democratic mode of participation, but fall short of consolidation in the sphere of the legal state and the treatment of minorities.

See also: Area and International Studies: Development in Eastern Europe; Communism; Communist Parties; Democratic Transitions; Eastern European Studies: History; Eastern European Studies: Politics; Eastern European Studies: Society; Peace Movements, History of; Socialism: Historical Aspects; Socialist Societies: Anthropological Aspects; Socialization, Sociology of; Youth Movements, History of

Bibliography

Ágh A 1998 *Emerging Democracies in East Central Europe and the Balkans*. Elgar, Cheltenham, UK

Dawisha K, Parrott B (eds.) 1997 *The Consolidation of Democracy in East-Central Europe*. Cambridge University Press, Cambridge, UK

Elster J, Offe C, Preuss U K 1998 *Institutional Design in Post-communist Societies*. Cambridge University Press, Cambridge, UK

Huntington S P 1991 *The Third Wave. Democratization in the Late 20th Century*. University of Oklahoma Press, Norman, OK

Karl T, Schmitter P 1991 *Transition to Democracy in Eastern Europe*. Macmillan, Basingstoke, UK

Merkel W, Puhle H-J 1999 *Von der Diktatur zur Demokratie*. Westdeutscher Verlag, Opladen, Germany

Linz J, Stepan A 1996 *Problems of Democratic Consolidation*. Johns Hopkins University Press, Baltimore, MD

Oksenberg M, Dickson B J 1991 The origins, processes and outcomes of great political reform. In: Rustow D A, Erickson K P (eds.) *Comparative Political Dynamics*. Harper Collins, New York

von Beyme K 1996 *Transition to Democracy in Eastern Europe*. Macmillan, Basingstoke, UK

Whitefield S 1993 *The New Institutional Architecture of Eastern Europe*. St. Martin's Press, New York

Zecchini S (ed.) 1997 *Lessons from the Economic Transition. Central and Eastern Europe in the 1990s*. Kluwer, Dordrecht, The Netherlands

K. von Beyme

Revolutions, Sociology of

The modern concept of revolution comes out of one particular experience at the end of the eighteenth century, an experience that led some to hope for and others to fear similar occurrences. Reflections on the nature, causes, unfolding, and consequences of revolution were for a long time dominated by the French Revolution. As other cases came to preoccupy social scientists in the course of the twentieth century, and particularly in the wake of the varied social upheavals after World War Two, scholarly reflection on revolution flourished as it developed in new directions.

1. An Old Word Takes on a New Meaning

'Revolution' (from Latin *revolutio*) initially had the sense of a return to a point of origin. In political discussion it might be used to characterize a dynastic restoration. By the seventeenth century it had also taken on the sense of an alteration in fortune and might be employed to describe some significant improvement or deterioration in political affairs. In this latter sense revolutions were often seen as frequent occurrences in the histories of states. For some eighteenth century writers, the installation of a new monarch in England in 1688 was a great singular revolution in the older sense of a return to an initial state of order after several tumultuous decades; for others it was but one of many revolutions that marked all of English history.

A century after this 'Glorious Revolution' in England, the geographic range and social diversity of French turbulence in 1789 and beyond (rural and urban insurrections throughout the country; military mutinies; elite challenges to royal authority; priests defying bishops; a tidal wave of legislative enactments) provided many revolutions (in the plural) for commentators to reflect upon. But as these upheavals continued to multiply and as various contenders sought to channel the dramatic events in one or another direction, participants began to speak of these many separate events as components of something singular, the French Revolution, that had a structure and causes, that was an ongoing process, and that rival contenders sought to understand and control. Some participants who were disappointed by specific events in France could even imagine that this particular process was but an instance of possible revolutions and look forward, as one radical group declared in 1794, to 'another, far greater, far more solemn revolution' of which 'the French Revolution is but the forerunner' (Furet and Ozouf 1988, p. 203). So some sought to make new (and perhaps differently constituted) revolutions and others to avert them; and theorists of either (or neither) persuasion sought to understand them.

Until 1917 the French instance was the central source of empirical evidence of what a revolution might be like. Much sociological reflection was profoundly stimulated by spectacular failures of radical causes to seize and hold power despite episodes of impressive popular mobilization such as the widespread European turbulence of 1848 or the Russian upheaval of 1905. But theorists were paying relatively little attention to social conflicts in lesser powers. Even well into the twentieth century, other revolutions, both imagined and experienced, were apt to be understood

with concepts derived from that French experience. Military terminations of popular upheaval were held to be 'Bonapartist'; Russian revolutionaries accused each other of being 'Thermidorians'; it has remained very difficult to think of political conflict without invoking notions of 'left' and 'right' (initially, spatial locations in revolutionary France's legislature).

The Russian Revolution of 1917 added some new possibilities to stimulate imaginations; the post-World War Two era, with its many anticolonial struggles, civil wars in which at least one party has claimed itself to be revolutionary, and regime transitions of many sorts, has provided even more food for thought about the causes, dynamics, and consequences of revolution. By the late twentieth century, the French dramas from 1789 on no longer stood as the central model for theoretical contemplation.

2. *An Elusive Definition*

Despite two centuries of reflection and an extremely vigorous body of comparative scholarship on revolution in the last quarter of the twentieth century, the very definition remained in dispute. First of all, to the extent that notions of revolution were abstracted from that singular French experience rather than from some sense of the commonalties of a large number of experiences, observers could readily differ on what they thought were defining features. Some of the elements that, in varying combinations, entered into various definitions might include:

(a) Illegal seizure of national political power.
(b) Extensive property transfers.
(c) Radical alteration in the form of government.
(d) Pervasive change in a wide variety of institutions.
(e) Extensiveness of popular mobilizations.
(f) Substantial violence.
(g) New formulas for the legitimation of authority.
(h) A new elite's claim to be effecting a radical break in the course of history.
(i) A new elite's claim to be advancing forward to a new level of human progress.
(j) Widespread sense of profound change.

Although scholars have never achieved consensus, this definitional debate opened up a whole host of fruitful questions about how these phenomena might be connected to one another.

Second, like many other concepts that are used by political actors, the term revolution has continued to be a term of legitimation and delegitimation rather than simply analysis. All sorts of things have been labeled (or not labeled) as revolutions in order to honor or dishonor them. This has been terribly confusing. A grammatically simple sentence, 'Revolutions entail considerable violence' might be any of at least four very different sorts of statement: part of a *definition* of revolution; a *proposition* about revolution (when defined without reference to violence); an *exhortation* to violent tactics (when uttered among committed revolutionaries); a *condemnation* of revolution (when uttered among people who abhor violence).

3. *Major Research Traditions*

Divergent emphases fueled overlapping but distinguishable traditions of research on popular collective mobilization, regime overthrow, political violence, elite dissidence, governmental legitimation and delegitimation, institutional change, and class conflict, and a good deal of reflection on how such processes affected each other. Beginning in the nineteenth century, a well-developed Marxian tradition examined ways in which long-term economic transformations created new group identities and interests while altering political structures. Revolution was seen as a traumatic readjustment of power relations, at bottom driven by the shifting economic interests of rational actors. A theoretically more diffuse body of work looked for strains of various sorts produced by industrialization, urbanization, and secularization and tried to explain collective outbursts as responses to such strains. In such interpretations, revolutions were apt to be seen as 'moments of madness,' as frenzied and doomed responses to the forces of modernity, for example, in Durkheimian notions of 'breakdown' or in the 'collective behavior' research tradition (see *Collective Behavior, Sociology of*). Tocqueville's thesis linked the growing might of a rationalizing state with a broad array of cultural transformations that undermined the legitimation of aristocratic privilege by transferring to the state activities previously carried out by the lords; indeed the lords might themselves become oriented to serving the new state bureaucracy (see *State, Sociology of the*). This attack on the problem lay relatively undeveloped in twentieth century social science until revived by a renewed sociological emphasis on the state from about the 1970s.

Although abstract definitions of revolution continued to vary widely, the field of comparative studies of revolution developed despite varying notions of what relevant instances might be. Some scholars attempted to find processual uniformities and constructed models of the 'natural history' of revolutions. Others looked for some robust causal mechanism that could serve as necessary or sufficient conditions (or both). This latter research tradition was dominated by the search for sources of increased grievance, as in a variety of theories that stressed economic downturns. A very different research tradition, impressed by the role of revolutionary parties in the twentieth century, sought to examine the capacities of relatively small but organized groups, to take action, to catalyze changes in ideas, or to mobilize broad support.

4. The Post-World War Two Era

The social upheavals of the post-World War Two era brought many new instances of collective action, political violence, and regime overthrow into play. Anticolonial movements, some defeating great powers (as in Vietnam); long-term rural warfare with divergent outcomes (as in Guatemala and Nicaragua); and Third World state breakdowns (some leading to radically new regimes, as in Iran) all added new cases.

The vastness of human suffering so evident in the twentieth century increasingly suggested attention to opportunity as well as grievance in the genesis of revolt and the success of revolution. Since revolt was so much more scarce than misery, it was evident that something other than (or at least in addition to) misery was needed to explain dramatic struggle for change, and social scientists became adept at identifying structures and moments that furnished opportunities. If one peasant community rather than another supported a revolutionary movement, social scientists now sought an understanding of what resources in the daily lives of these (but not other) peasants might account for why these (but not other) peasants revolted. If one revolutionary party in a small, poor country (for example, Nicaragua) succeeded in coming to power, while another revolutionary party in another small, poor country (for example, El Salvador) succeeded only in sustaining a bitter, stalemated war with the government, social scientists began to explore the differential vulnerabilities of different kinds of regimes. Distinguishing long-standing differences in resources available to rebels from temporary opportunities, and distinguishing stable structures of elite domination from specific policies, became very important for this research tradition. A 'structural' school of explaining revolution was emerging.

But accounts of favorable structures soon seemed incomplete as well. The Iranian Revolution of 1979, in which many leading participants spoke a religious language at radical variance with the language of those who had come to power in previous, paradigmatic cases, brought many scholars to reconsider the role of cultural elements in revolutionary processes. The sudden breakdown of communist regimes in Europe were accompanied by extensive property transfers, but the victors claimed to be undoing revolution, even to be terminating the dream of revolution, which again raised important questions of culture as well as structure (see *Postsocialist Societies*).

Many of the instances in the second half of the twentieth century were occurring in lesser powers on the world stage, in places obviously subject to military, economic, or ideological domination by more powerful states, and whose upheavals often involved attempts (some partially successful) to throw off some of these forms of domination. One would hardly try to understand anticolonial revolutions without paying attention to the policies and institutions of colonial rule, social upheavals in countries exporting tropical crops without paying attention to the world economy, and the resources available to governments and rebels in small, poor, and weak countries without paying attention to the capacities of both sides to elicit support from great powers.

Insights garnered in examination of the newer instances began to suggest different emphases in new examination of older cases. The French or Russian revolutions were looked at as cultural processes; the embeddedness of the 'great' revolutions of the past in transnational connections was noted; and conflicts beyond the few previously paradigmatic instances entered the comparativists' field of vision. The very variety of participants in the newer cases expanded the intellectual horizons of researchers. As varying combinations of participants in the newer cases became evident, comparative scholars were recognizing how multisided the conflicts were in past revolutions as well. Revolutions, both past and recent, were increasingly likely to be seen as complex affairs, in which governing authority failed in many different ways, elites divided in many different ways, and rebels mobilized in many different ways. Methodologically this meant that social scientists were likely to be looking at configurations: at complex intersections and interactions of processes that mobilized collective actions of varying kinds, produced elite conflict, and weakened governing institutions. Theoretically it means that the search for generalizable statements about neccessary or sufficient conditions for revolution needed to be rethought.

5. Emerging Directions in the Sociology of Revolution

The issues engaging comparative scholars in the late twentieth century meshed with some of the large issues engaging macrosociology generally, indeed, were suggesting revolutions as a fertile terrain for engaging those issues. Among the issues generating much discussion:

(a) *How to integrate cultural accounts with structural ones.* Cultural issues were an emerging agenda in various ways. Social historians were making enormous advances in understanding popular culture in varied times and places. A new generation of scholars was examining the overt cultural politics of revolutionary elites. The direction of regime change in such places as post-Shah Iran and post-communist Eastern Europe seemed altogether impossible to discuss without considering legitimating ideas.

(b) *How to integrate accounts of individual action with macrostructural processes.* As elsewhere in sociological theory this constituted a major challenge. Proceeding from assumptions of individual rational actors, how could one account for the forms assumed

by collective action, especially in light of the often evident individual risk assumed in participation in revolutions in combination with their uncertain and often undesired outcomes? (see *Action, Collective*).

(c) *Adequate attention to the visions of revolutionary elites, the wellsprings of popular mobilizations, and the interaction of the two.* Some structural accounts seemed to make the visions and intentions of revolutionary leadership inconsequential; some other accounts seemed to make them everything and by implication ignored the motivations, intentions, and wishes of ordinary participants. Some urged us to consider how it was that some revolutionary parties managed to win the adherence of large numbers of followers; others urged us to see popular participation in revolution as rooted in the structures and culture of the popular classes and therefore quite separate from the concerns expressed by revolutionary elites. The empirical difficulty in getting these connections right remained a challenge in studies of particular revolutions, let alone efforts at comparative statement. Poorer, anonymous participants not only often lack the means to make their own wishes widely known (peasants seizing land are less likely to hold press conferences or write memoirs than revolutionary agriculture ministers), but are, also as usual, often highly motivated to conceal their intentions (especially in highly charged situations like revolutions). It is a major methodological problem that revolutionaries who claim to speak on behalf of, say, Central American peasants, are rarely peasants themselves.

(d) *Attention to the outcomes of revolutions.* The consequences of revolution for economic well-being, state power in the international arena, the rationalization of administration, national identities, citizenship rights, and democratization remain intensely controversial matters about which relatively little systematic investigation has been done by comparativists.

(e) *How to integrate accounts of revolution with other kinds of political conflict.* Scholars of revolution were increasingly making use of tools that resembled those used by other students of politics: attending to the play of threat and opportunity, the variety of resources, the forging of identities in struggle, the irony of unintended consequence, the gaps between elites and ordinary participants (and the interaction between the two). Increasing interchange with scholars of other modes of conflictual action (urged by Doug McAdam, Sidney Tarrow, and Charles Tilly) seemed a probable future direction (McAdam et al. 1997).

(f) *How to make superior use of comparative methodologies.* With increasing recognition of the complexity and diversity of revolutionary processes, the very logic of comparison was coming under heightened scrutiny. If revolutions are complex amalgams of multiple conflicts, some comparativists suggest that rather than searching for some single general cause, or some single general process, or some set of causes whose separate contributions to revolution may be simply added together, there may be multiple configurations of causal processes which produce revolutions, which themselves unfold in a variety of ways. Some scholars sought to distinguish varieties of revolution about which valid generalization might be made. Others suggested that the way forward, theoretically, was to seek generalizations about much smaller processes, which were components of revolutions, but which combined together in so many distinct ways as to defy generalization on that concatenated level. That is to say, in this view 'revolution' is a complex made up of many components, and one revolution is made up differently from another; looking for uniformities in the class 'revolution' is hopeless, but there may be regularities in the causation or action of some of the components.

6. *Conclusion*

More than two centuries of reflection since 1789 have hardly exhausted the stimulation provided by revolution to the thinking of social scientists; if anything, the last quarter of the twentieth century saw a tremendous intellectual reinvigoration. This is likely to continue, for revolutions seem a particularly privileged terrain for many significant issues in macrosociology, among which I indicate four:

(a) the intermingling of change and persistence

(b) the interplay of long- and short-run processes

(c) the relationship of local, national, and transnational processes

(d) finally, the collision of hopes for change and recalcitrant realities.

Revolution brings to the fore all the many issues that have swirled around the interplay of 'structure' and 'agency.' There are no other social processes which so strikingly juxtapose the capacity of human beings to construct for themselves previously unrealized and unexpected patterns alongside the incapacity of human beings to construct their world as they please.

See also: Conflict Sociology; Political Sociology; Power in Society; Revolutions, Theories of; Social Movements, Sociology of; Violence: Public

Bibliography

Baker K M 1990 *Inventing the French Revolution*. Cambridge University Press, New York

Brinton C 1965 *The Anatomy of Revolution*. Vintage Books, New York

Foran J (ed.) 1997 *Theorizing Revolutions*. Routledge, London and New York

Furet F, Ozouf M 1988 *Dictionnaire critique de la Révolution française*. Flammarion, Paris

Goldstone J A (ed.) 1998 *The Encyclopedia of Political Revolutions*. Congressional Quarterly, Washington, DC
Keddie N R (ed.) 1995 *Debating Revolutions*. New York University Press, New York
Markoff J 1996 *The Abolition of Feudalism: Peasants, Lords and Legislators in the French Revolution*. Pennsylvania State University Press, University Park, PA
McAdam D, Tarrow S, Tilly C 1997 Towards an integrated perspective on social movements and revolution. In: Lichbach M I, Zuckerman A (eds.) *Ideals, Interests and Institutions: Advancing Theory in Comparative Politics*. Cambridge University Press, Cambridge, UK
Rey A 1989 *'Révolution', histoire d'un mot*. Gallimard, Paris
Selbin E 1998 *Modern Latin American Revolutions*, 2nd edn. Westview Press, Boulder, CO
Skocpol T 1979 *States and Social Revolutions*. Cambridge University Press, Cambridge, UK
Tilly C 1993 *European Revolutions, 1492–1992*. Blackwell, Malden, UK
Tocqueville A de 1952–3 *L'Ancien Régime et la Révolution*, 2 Vols. Gallimard, Paris
Wickham-Crowley T P 1992 *Guerrillas and Revolution in Latin America. A Comparative Study of Insurgents and Regimes Since 1956*. Princeton University Press, Princeton, NJ

J. Markoff

Revolutions, Theories of

Theories of revolution are (or purport to be) causal explanations of revolution—they say that if certain sufficient conditions (the causes) occur, then revolution (the effect) will occur as well. But *revolution* has different meanings, and theories of revolution do not all explain the same facts—different theories explain different facts as well as cite different causes. Before analyzing theories of revolution and the problem situation they address, therefore, one must distinguish between the different facts that those theories would explain—the different events that *revolution* refers to.

1. Definitions

Some theories explain revolution mainly in the sense of a revolutionary movement—a concerted attempt to change government, regime, or society (or all three) by means of violence. Other theories explain revolution mainly in the sense of a revolutionary outcome—a successful change of government, regime, or society (or all three) by means of violence. Still other theories explain revolution mainly in the sense of a revolutionary situation—a contest for state power by means of violence that begins when insurgents claim sovereignty and ends when they overthrow, succumb to, or make peace with the government.

A revolution in one sense can occur without a revolution in the other senses occurring as well. A revolutionary movement can occur without a revolutionary outcome occurring because revolutionaries can fail. A revolutionary outcome can occur without a revolutionary situation occurring because the government in power can make a 'revolution from above' *à la* Stalin and because revolutionaries can take power legally *à la* Hitler. And a revolutionary situation can occur without a revolutionary outcome occurring because the government under siege can win. But neither a revolutionary situation nor a revolutionary outcome can occur without a revolutionary movement occurring because they each entail revolutionary violence, which takes concerted effort to commit.

The above definitions of revolution overlap with common social science usage, which calls a change of government by means of violence a *coup d'état*, a change of regime (as well as of government) by means of violence a political revolution, a change of society (as well as of government and regime) by means of violence a social revolution, and a contest for state power by means of violence a civil war. Common social science usage can be misleading, however, especially when it suggests that the facts referred to are mutually exclusive. A *coup d'état* can become a political revolution, and a political revolution can become a social revolution—if the new government decides to make a 'revolution from above' (Colburn 1994). Indeed, a social revolution is a change of state and society by means of violence that the new government mainly commits (Walzer 1998). And a civil war that gives not only radical revolutionaries but also parochial rebels a fighting chance, and ends in victory for whichever side best combines what Machiavelli called *virtù* and *fortuna*, can start because conservatives revolt in order not to make a revolution but to prevent one.

In sum, theories of revolution that would explain a revolutionary movement or a revolutionary situation need not explain a revolutionary outcome, but theories that would explain a revolutionary situation, a revolutionary outcome, or both have to explain a revolutionary movement as well. And whatever the facts that theories of revolution would explain, those facts entail violence—the intentional infliction of physical harm. The Bolshevik theoreticians who ridiculed nonviolent revolution as a contradiction in terms conformed to social science usage: no violence, no revolution—by any definition. All theories of revolution converge on the *modus operandi*—revolutionary violence. The question they all have to answer is what motivates revolutionary violence and what intended or unintended consequences follow from committing it.

2. Theories

Theories of revolution in Western social science date back to the Greeks, who viewed revolution as the

worst kind of war. Thucydides blamed revolutionary situations (with or without revolutionary outcomes) on revolutionary movements or conspiracies, and blamed those conspiracies on hope of success. Plato agreed and blamed hope of success on weak government torn by faction, which he thought to eliminate from his ideal state by abolishing the family and private property among the rulers. Aristotle added (on the widest though not deepest comparative study) that since hope along with faction springs eternal, no regime is immune against revolution.

The theory that people attempt revolution because they have hope of success, and have hope of success because they think the government is vulnerable, lived on in classic political sociology as an explanation of what to avoid and how to avoid it. Machiavelli, who wrote in effect the first handbook on counterinsurgency, told new rulers who had just seized power how to use force and fraud in order to avoid being overthrown. Hobbes, trusting more to organization than to statecraft, designed a regime that would prevent revolutionary movements and revolutionary situations (which he called sedition and civil war) by repressing faction. *The Federalist* took Hobbes for granted, but held that checks and balances could avert faction and insurrection as well as tyranny. Burke—who excoriated the French Revolution already in its 'liberal' phase and predicted the regicide 2 years, the Terror 3 years, and the military dictatorship 9 years in advance (O'Brien 1993)—observed that the revolutionary parliament usurped a king busy making concessions. As Tocqueville summed up the whole line of thought: 'The evil that one suffers patiently as inevitable seems unbearable as soon as one conceives the idea of escaping it.'

Not all theorizing on revolution was counterrevolutionary. 'Violence is the midwife of every old society pregnant with a new one,' wrote Marx, though his instant histories of revolutionary situations autopsied abortions, and his class-conflict theory of revolutionary movements and their abortionists proved false—street fighters on both sides of the June 1848 revolutionary situation in Paris, for instance, not only had the same class background, but were a cross section of the manual work force (Traugott 1985). The theory explained history no better elsewhere—industrial workers neither launched the revolutionary movement nor achieved the revolutionary outcome that the theory predicted. In Germany they did not try, and in Russia they acquiesced as a revolutionary movement of Marxist intellectuals achieved a revolutionary outcome in their name by a military *coup d'état*—the October Revolution was so low-profile that the author of *Ten Days That Shook the World* slept through much of it—followed up with Red Terror and Civil War (Moore 1978, Pipes 1991, 1995). But for the vicarious, posthumous prestige Marx got from that much-imitated revolutionary outcome, whose official ideology-cum-state religion made him a household word, no one except social scientists might have heard of him today (Kolakowski 1978).

Whether counterrevolutionary or revolutionary, however, theories of revolution were handmaidens to political sociology and historiography before the twentieth century, when—after the revolutions attending World War 1, World War 2, and decolonization—they hived off as a cluster of answers to the why-question of revolutionary violence. The chief theorists were not activists but academics because activists (even successful ones) propounded only trite theories of revolution consisting mainly of prophecy and propaganda. Where activist theories rose above agitprop to consider strategic means and ends, they rehearsed the classic wisdom in reverse—they advised revolutionaries to play on faction in order to weaken the government and thereby raise hope of success. Some activists became revolutionary dictators, but more academics became theorists of revolution and launched a social science growth industry (Rule 1988, O'Kane 2000).

Academic theories of revolution, in the standard pilgrim's progress view, went through three stages of development since the 1920s. First there were natural histories explaining recurrent sequences of similar events in 'great' revolutions by the reasons that experts say motivated the people who made them, but not generalizing the reasons as a causal theory. Then there were general theories explaining revolutionary violence by discontent, but accounting for history ad hoc if at all. Now there are comparative-historical theories explaining social and political revolutions by 'structural conditions' of state and society cogently enough that older theories look *passé* (Skocpol 1979, 1994, Goldstone 1986, 1991, Goldstone et al. 1991).

The standard view is not far wrong, though it ignores continuities of theory and method. All theories of revolution cite a short list of sufficient conditions or explanatory causes and validate, or rather illustrate, the explanation with factual evidence consisting of case histories and statistics arranged to show by the comparative method that where the explanatory causes occur, revolution (by some definition) occurs as well, and that where the explanatory causes do not occur, revolution (by that definition) does not occur either.

Grouped by the explanatory causes they cite, theories of revolution form four families: psychological, functional, structural, and political theories. As a rule, psychological and functional theories explain revolutionary movements—they explain why people attempt to change government, regime, or society (or all three) by violence; structural theories explain revolutionary outcomes—they explain why people successfully change government, regime, or society (or all three) by violence; and political theories explain revolutionary situations—they explain why people fight civil wars for state power. Psychological and functional theories that explain revolutionary

movements mostly leave revolutionary situations and revolutionary outcomes over to historians, whereas structural theories that explain revolutionary outcomes also explain revolutionary situations and revolutionary movements along the way, and political theories that explain revolutionary situations also explain revolutionary outcomes and revolutionary movements at least in passing.

Psychological theories explain revolutionary movements by frustration (alias relative deprivation or discontent) defined as a discrepancy between what people want and expect and what they get. These theories all assume the frustration-aggression principle that people who are frustrated commit aggression against the perceived source of frustration, but they disagree about what social facts cause frustration—one blames steady improvement followed by a sudden setback that raises expectations and reduces attainments; another blames rapid social change that raises expectations beyond attainments; yet another blames whatever raises expectations, lowers attainments, or both. Few psychological theories explain revolutionary violence by frustration alone, however; most add insurgent strength and government weakness to their short list of sufficient conditions.

Functional theories explain revolutionary movements by dysfunction (alias strain or disequilibrium) defined as a discrepancy between socially prescribed ends, goals, or values, and the available institutional means or procedures for attaining them. When social change creates a gap between ideology and experience that the government omits to close with reform, people turn to revolutionaries for salvation. If the government looks weak or loses control of its armed forces, or if revolutionaries think they can overthrow it, revolutionary violence erupts.

Structural theories explain revolutionary outcomes piecemeal—they explain first why élites subvert governments, then why subalterns rebel, then why radical vanguards seize power, and then why new rulers reconstruct state and society. The first two events make up revolutionary situations; the last two make up revolutionary outcomes. Each event depends on the one that precedes it—no subversion, no rebellion; no rebellion, no power transfer; no power transfer, no reconstruction. All these events comprise collective actions and reactions, which the theories explain as rational efforts to satisfy interests under structural constraints like government weakness, subaltern community organization, vanguard politico-military organization, and geopolitical exigency. Ideology is a cover for interest—the motivating constraints are structural.

Political theories likewise explain revolutionary situations piecemeal—they explain first why insurgents claim sovereignty, then why certain people support them, and then why the government fails to repress the insurgency. Revolutionary situations do not all have revolutionary outcomes, but some do, and political theories also explain these outcomes piecemeal—they explain first why government supporters defect, then why insurgents gain armed force, then why government forces waver or desert, and then why insurgents take power. All these events comprise collective actions and reactions, which political theories explain as rational efforts to satisfy interests under political and strategic constraints like organization, armament, threat, and opportunity. Ideology again is a cover for interest—the motivating constraints are political and strategic. Whether and to what extent new rulers reconstruct state and society is another question, which political theories mostly pass over.

These family portraits of theories of revolution are not just composite sketches, they are caricatures, though caricatures accentuate real features. Thankfully, competent accounts of the theories as well as internecine debates among the theorists are legion (Rule 1988, Goldstone et al. 1991, Skocpol 1994).

Theories of revolution are contentious. Advocates of structural theory point out that psychological and functional theories explain neither revolutionary situations that occur because élite infighting paralyzes the government nor revolutionary outcomes that see new regimes more bureaucratic and tyrannical than the old (Skocpol 1979, Goldstone 1991). Advocates of political theory point out that revolutionary outcomes depend on the fortunes of war in revolutionary situations—structural *virtù* proposes, political *fortuna* disposes (Tilly 1993). And skeptics add that since revolutionary outcomes owe to actions that vanguards take because they expect these actions to yield benefits that often do not materialize, no theory that downplays ideology can explain revolution cogently (Aya 1990). How else to explain strategic decisions like Lenin's based on false expectations of imminent world revolution, the rage for socialism in poor countries ruled by revolutionaries, and the resort to mass terror by revolutionary governments in proportion to the innovation they attempt (Pipes 1991, 1995, Colburn 1994, O'Brien 1993).

Though theories of revolution are contentious, few of them disagree with Thucydides that revolutionary movements owe to hope of success or with Plato that hope feeds on weak government torn by faction. The wisdom of the ages is no less wise for being ancient.

See also: Collective Behavior, Sociology of; Conflict Sociology; Legitimacy: Political; Revolution; Revolutions, History of; Revolutions, Sociology of; Social Movements, History of: General; Social Protests, History of; Violence, History of; Warfare in History

Bibliography

Aya R 1990 *Rethinking Revolutions and Collective Violence: Studies on Concept, Theory, and Method.* Spinhuis, Amsterdam, The Netherlands

Colburn F D 1994 *The Vogue of Revolution in Poor Countries*. Princeton University Press, Princeton, NJ
Goldstone J (ed.) 1986 *Revolutions: Theoretical, Comparative, and Historical Studies*. Harcourt, Brace, Jovanovich, San Diego
Goldstone J 1991 *Revolution and Rebellion in the Early Modern World*. University of California Press, Berkeley, CA
Goldstone J A, Gurr T R, Moshiri F (eds.) 1991 *Revolutions of the Late Twentieth Century*. Westview, Boulder, CO
Kolakowski L 1978 *Main Currents of Marxism*, 3 vols. Clarendon, Oxford, UK
Moore B Jr 1978 *Injustice: The Social Bases of Obedience and Revolt*. Pantheon Books, London
O'Brien C C 1993 *The Great Melody: A Thematic Biography and Commented Anthology of Edmund Burke*. Minerva, London
O'Kane R H T (ed.) 1999 *Revolution: Critical Concepts in Political Science*, 4 vols. Routledge, London
Pipes R 1991 *The Russian Revolution*. Vintage, New York
Pipes R 1993 *Russia Under the Bolshevik Regime*. Knopf, New York
Rule J B 1988 *Theories of Civil Violence*. University of California Press, Berkeley, CA
Skocpol T 1979 *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge University Press, Cambridge, UK
Skocpol T 1994 *Social Revolutions in the Modern World*. Cambridge University Press, Cambridge, UK
Tilly C 1993 *European Revolutions, 1492–1992*. Blackwell, Oxford, UK
Traugott M 1985 *Armies of the Poor: Determinants of Working-class Participation in the Parisian Insurrection of June 1848*. Princeton University Press, Princeton, NJ
Walzer M 1998 Intellectuals, social classes, and revolutions. In: Skocpol T (ed.) *Democracy, Revolution, and History*. Cornell University Press, Ithaca, NY

R. Aya

Rhetoric

'Rhetoric' has different meanings depending on the context and whether used abstractly or concretely. The word, found with slightly different spellings in modern European languages, is derived from Greek *rhêtorikê*, which first appears in Plato's dialogue *Gorgias*, written about 385 BC. It is used there of the art (*tekhnê*) of a public speaker (*rhêtôr*), and may be defined as the art of persuasive public speaking; it can refer to the theory of discourse, to the art of orators and writers, or to speeches, writings, or sometimes other art forms (painting, sculpture, architecture, music, and dance) when these are intended to influence an audience. In scholarly contexts 'rhetoric' is usually a morally neutral term, but it has had influential detractors beginning with Plato and continuing with Descartes and Locke, with the result that in modern popular language or journalism the term often describes empty or misleading verbosity.

From the fourth century BC until the nineteenth century AD rhetoric was an academic discipline taught in most western schools and universities, following grammar and preceding dialectic in the liberal arts trivium. Its function was to teach and practice students in techniques of public speaking or written composition in prose and verse. Thousands of rhetorical handbooks, treatises, and model compositions were written from ancient to modern times; many have survived in manuscript or print, but many were ephemeral. The Romantic movement, with its rejection of rule-based composition and celebration of individual genius, and the development of science, with its assertions of objective truths, undermined study of rhetoric in the nineteenth century, so that it often came to be thought of as limited to identification of linguistic tropes, such as metaphor and metonymy. In the twentieth century, however, rhetoric experienced a renaissance in connection with new communication and media studies, new literary theories, and new forms of linguistic analysis. Confidence in objective truth was challenged, and some thinkers came to believe that all speaking and writing, even discourse about the natural and social sciences, is a rhetorical act and necessarily contains rhetorical elements. The extension of rhetoric from its traditional concern with public address to embrace discourse of all sorts is called 'globalization' (Gross and Keith 1997), a term perhaps better reserved for another recent development, the comparative study of the rhetorical traditions of western and nonwestern societies. In contrast, 'rhetoric' is sometimes the name given to introductory courses in composition at the college level. In these courses students are practiced in a variety of forms of writing, including descriptive, narrative, and argumentative compositions together with attention to spelling, word choice, grammar, arrangement, and the use of figures.

1. *The Historical Development of Rhetoric in Europe and America*

Systematic study of rhetoric in Greece resulted from the need for effective public speaking by ordinary citizens under democratic governments, especially in Athens, in the fifth and fourth centuries BC. In the Athenian democracy litigants were expected to plead their own cases before large juries of fellow citizens, thus putting a premium on rhetorical ability, and the functioning of the council and popular assembly was predicated on extensive debate among citizens. The first teachers of rhetorical techniques were the traveling lecturers known as 'sophists,' of whom the most famous are Protagoras and Gorgias. In Plato's dialogue *Gorgias*, in opposition to Gorgias and his followers, Socrates takes a very negative view of rhetoric, which he represents as a form of flattery of an audience and as lacking concern with truth, but in Plato's later dialogue, *Phaedrus*, Socrates is made to describe an ideal philosophical rhetoric that would use

knowledge of logic and psychology for legitimate ends. By the late fifth century BC rhetorical handbooks could be bought and studied. They seem to have included examples of introductions, narratives, probable arguments, and conclusions in the form of commonplaces that readers could adapt to specific situations.

1.1 Aristotle's Theory of Rhetoric

Plato's student, Aristotle, developed some earlier teaching in greater depth and added many practical observations in his lectures, embodied in the treatise *On Rhetoric* (ca. 336 BC). Although it was rather little read until modern times, through the work of his students many of Aristotle's theories became part of traditional rhetorical teaching. Major concepts and divisions of modern rhetorical theory are derived from Aristotle. He defined rhetoric abstractly as 'an ability (or faculty: *dynamis*), in each case, to see the available means of persuasion.' A speaker uses 'nonartistic' means of persuasion, such as witnesses or documents, and 'invents' artistic means of persuasion, of which there are, Aristotle said, three and only three: they are often referred to as *ethos* (presentation of the character of the speaker as trustworthy), *pathos* (the emotions of the audience as aroused by the speaker), and *logos* (rational argument). There are two kinds of logical argument: the enthymeme, or rhetorical syllogism, is a deductive, probable argument in popular form; it usually takes the form of a statement with a supporting reason or of an 'if …, then …' hypothesis. The other type of argument is the paradigm, or example; it is an inductive argument, supporting a proposition with examples from history or analogy.

Rhetoric as Aristotle understood it dealt with probable knowledge in the realm of politics (though it was found also in speeches in dramatic and narrative literature), not with scientific or philosophical knowledge. The sources of enthymemes are of two kinds: specific topics, that is, propositions drawn from other disciplines, such as politics, ethics, psychology, economics, or military strategy; and dialectical or 'common' topics, that is, techniques of reasoning, such as definition, division, argument from cause and effect, etc. Aristotle defined three species of rhetoric: judicial (sometimes called forensic) rhetoric, practiced in courts of law, is concerned with determining the justice of an action in the past; deliberative rhetoric, practiced in formal or informal political assemblies, is concerned with determining the benefits or expediency of a proposed future action; epideictic (or demonstrative) rhetoric, practiced on ceremonial occasions such as a holiday or funeral, is concerned with what is honorable or dishonorable and takes the form of praise or blame of a person or thing. Many of Aristotle's rhetorical concepts are universals, applicable to discourse in all cultures; some need modification to apply to discourse more generally. Ethos, for example, has come to refer to any moral factor or any presentation of character used in a discourse for persuasive purposes, not just the credible presentation of the speaker's veracity, and epideictic is now viewed as ceremonial or ritual discourse concerned with instilling, transmitting, or reinvigorating cultural values.

1.2 Greco-Roman Rhetorical Theory

Aristotle was primarily interested in what came to be known as 'invention,' the aspect of rhetoric concerned with the use of topics and means of persuasion, but in the third book of his treatise he began with some remarks on delivery, followed by an extended discussion of prose style and finally an account of the structural parts of a public address. Subsequent Greek and Roman rhetoricians outlined a theory of rhetoric in five parts which recapitulate the process of planning and delivering a speech: invention, or the planning of the argument; arrangement, or division of the speech into parts; style (the choice of words and composition of sentences); memory (or preparation for delivery); and delivery. Greek and Roman teachers of rhetoric were especially interested in judicial oratory and had much to say about the definition of the question at issue, which is known as the *stasis* of a case. Three main categories of stasis were recognized, with numerous subdivisions and many variants. Stasis of fact (also called conjectural stasis) deals with the situation in which a defendant denies committing the act with which he is charged (e.g., I did not kill him); in a case of legal stasis (also called stasis of definition) the defendant admits the act but claims it was legally justified (e.g., I killed him, but in self-defense); in stasis of quality the act and its definition are admitted, but extenuating circumstances are alleged (e.g., I did not intend to kill him). A topic much discussed by the rhetoricians was argument about interpretation of the letter and intent of a law or document, something of continuing interest in modern jurisprudence.

Greek and Latin rhetorical handbooks, of which a number have survived, generally describe the arrangement of a speech into a series of parts, each with its own function. The *prooemion* or introduction should make the audience attentive, receptive, and well disposed, and specific suggestions were offered to effect this. A narration should follow, giving a clear, brief, and persuasive account of the facts of the case. There is then often a statement, or proposition, of what the speaker hopes to prove, with perhaps division, or partition, into a series of headings. Next comes the proof, supported by witnesses and other evidence when possible, but often largely a more or less probable argument based on the circumstances and the character of those involved. Finally comes an epilogue, or peroration, in which the speaker summarizes the argument and seeks to stir the emotions of the audience.

The subject of style was much developed by rhetoricians after Aristotle; it is found discussed in rhetorical handbooks such as *Rhetoric for Herennius*, in Latin and dating from the early first century BC, and also in a number of monographic works, of which the most famous is the Greek treatise *On the Sublime*, probably dating from second century AD (the authorship of both works is uncertain). Discussions of rhetorical style dealt with the choice of words, the composition of sentences, both simple and periodic, appropriate rhythms for spoken prose, tropes, figures, and kinds of style. A trope is a substitution of one word or phrase for another; among the most common are metaphor, metonymy, synecdoche, and hyperbole. Figures are divided into verbal figures, sometimes also called grammatical figures, such as alliteration and anaphora, and figures of thought, such as rhetorical question and antithesis. Style was divided by many authorities into three kinds: plain, grand, and middle; others, of whom the Greek rhetorician Hermogenes (fl. ca. AD 180) is the best known, sought to describe a series of qualities such as grandeur, beauty, sincerity, sweetness, subtlety, and vehemence, that could be blended into a harmonious and effective whole. The works of the orator Demosthenes (384–322 BC) were commonly taken as the best examples of eloquence in Greek, the speeches of Cicero (106–43 BC) in Latin, and students were encouraged to analyze and imitate these models.

The rhetoricians developed mnemonic techniques to help their students memorize speeches. The commonest system was based on identifying words or ideas with visual symbols and imagining these against a familiar spatial background. Discussions of delivery dealt with control of the voice and the use of gestures.

By the first century BC the study of rhetoric had been introduced into Roman schools, although after resistance from conservatives; it always kept close to concepts taught by Greeks. The turbulent conditions of the late Roman Republic provided the context for vigorous debate in the senate, popular assembly, and lawcourts, which were the venues of the rhetorical masterpieces of Cicero. He also wrote dialogs and treatises on rhetorical theory in which the orator is celebrated as the ideal figure of civilized society, to be achieved by a man learned in philosophy, law, history, literature, and rhetoric and applied in public life for the common good. Cicero recast the Aristotlean means of persuasion into what he called 'the duties of the orator': to teach, to charm, and to move, and he identified these with the plain, the middle, and the grand style respectively.

The establishment of the Roman Empire by Augustus after 30 BC led to a decline in opportunities for political oratory, but schools of rhetoric strengthened their dominance of the educational system. In these schools much of the students' time went into the practice of 'declamation,' writing and orally delivering speeches on imaginary historical and legal themes, often expressed in highly artificial prose. The most extensive account of classical rhetorical theory and rhetorical education is found in *The Education of the Orator* (*Institutio Oratoria*) in twelve books, published by Quintilian in 95 AD. Quintilian followed Cicero in regarding the orator as the highest civic ideal, and repeatedly insisted that only a morally 'good man' (*vir bonus*) can become a 'good orator.' Early Christians, Saint Paul among them, show ambivalent attitudes toward rhetoric, but preachers often borrowed techniques from classical rhetoric. About AD 426 Saint Augustine, a teacher of rhetoric before his conversion to Christianity, completed an influential treatise, *On Christian Learning* (*De Doctrina Christiana*), which is essentially a handbook of Christian rhetoric, applying Ciceronian teaching to the training of a Christian preacher and teacher.

1.3 Medieval and Renaissance Rhetoric

Although freedom of speech and opportunities for public address were limited under autocratic governments in late antiquity and the Middle Ages, some understanding of rhetoric was transmitted in encyclopedias of the liberal arts compiled by Martianus Capella (fifth century), Cassiodorus (sixth century), and Isidore of Seville (seventh century). In the later Middle Ages rhetoric found new application in handbooks of letter writing (*ars dictaminis*), verse composition (*ars poetriae*), and thematic preaching (*ars praedicandi*). The recovery of a reading knowledge of Greek in Western Europe in the fifteenth century, and the rediscovery by Italian humanists of important manuscripts on rhetoric from the classical period were major contributions to the vigorous intellectual life of the Renaissance. Rhetoric as taught in schools and universities in the Renaissance was primarily based on writings by Cicero and Quintilian, but some dissenting voices were heard. In the sixteenth century Peter Ramus, lecturing in Paris, transferred the teaching of invention from the course in rhetoric to the course in dialectic and limited rhetoric to the study of style, primarily the use of tropes and figures of speech. His position proved especially popular with Calvinists and Puritans; rhetoric as first taught in America (at Harvard College, founded 1636) was based on Ramist teaching. Meanwhile, textbooks of rhetoric began to appear in the vernacular languages; the first in English was *The Art or Crafte of Rhethoryke* by Leonard Cox (1535). A more profound understanding of the subject is found in Francis Bacon's *Advancement of Learning* (1605), where 'the duty and office of rhetoric' is defined as 'to apply reason to imagination for the better moving of the will.' A split between a narrow view of rhetoric as a linguistic or literary phenomenon, largely limited to the use of tropes and figures in written language, and a broader view of rhetoric as a persuasive tool of political speaking and discursive writing

has persisted. The linguistic–literary approach has often characterized the teaching of rhetoric on the European continent and in English Departments in schools and universities; the view of rhetoric as concerned with political issues and with discursive writing flourished in eighteenth-century Britain and America and has its modern home in Departments of Speech, Communication, or Media Studies.

1.4 The Relationship of Rhetoric to Poetics and Dialectic in the Western Tradition

Although the purpose of rhetoric was traditionally assumed to be persuasion, and the purpose of poetry to be pleasure and enjoyment, rhetoric overlapped with poetics in that both had a concern with artistic composition: word choice, tropes, figures, and sentence structure. In addition, traditional rhetorical handbooks regularly discussed narrative techniques, also applicable to narrative poetry; and rhetorical theories of the presentation of character, arousing emotions, and the effective use of humor had some application to dramatic and other poetic genres. Rhetoricians insisted that their students study poetry as a source of diction, compositional techniques, and inspiration. The close relationship between rhetoric and poetics can be well seen in Horace's Latin didactic poem *The Art of Poetry*, written toward the end of the first century BC, describing the task of the poet as effectively mixing useful teaching with sweetness of expression, a view that reappears in the medieval *artes poetriae*, and in renaissance and early modern literary criticism such as Boileau's *L'Art poétique* (1674).

Rhetoric traditionally overlapped with dialectic as practiced in schools in the treatment of logical argument, including definition, division of a question, and the use of topics. The two disciplines differed in that rhetoric was usually regarded as embodied in a continuous discourse, whether oral or written, whereas dialectic took the form of a question and answer dialogue. Moreover, rhetoric dealt with a specific question, primarily of a political or judicial sort, while dialectic dealt with general questions or philosophical universals, and perhaps most important of all, in its effort to persuade an audience rhetoric utilized the appeal of character and emotion as well as logical argument, while in dialectic only logical argument was regarded as legitimate. A good discussion of the traditional relationship of dialectic and rhetoric is found in the fourth book of Boethius' *On Topical Differentiations* (*De Topicis Differentiis*), written in the sixth century and used as a textbook in the Middle Ages and Renaissance.

2. The 'New' Rhetorics

The thinkers of the Scottish Enlightenment of the eighteenth century were much interested in rhetoric and in their lectures and writings combined traditional elements from Aristotle, Cicero, and Quintilian with their new understanding of psychology and consideration of modern conditions. Among important figures in this movement were David Hume, Adam Smith, George Campbell, and Hugh Blair. Writings of Campbell and Blair were translated into several languages and used in the United States and elsewhere as textbooks of rhetoric well into the nineteenth century.

In the twentieth century many new studies of rhetoric were published, based on historical, linguistic, literary, philosophical, political, and psychological research and influenced by the growth of mass media, the use of propaganda and manipulation of audiences for economic or political advantage, and new research methods, both quantitative and qualitative. New professional organizations were founded for those interested in rhetoric, among them the Rhetoric Society of America and the International Society for the History of Rhetoric. *The Quarterly Journal of Speech* and other older journals published an increased number of articles on rhetoric, and new journals devoted to rhetorical subjects began to appear. These include *Philosophy and Rhetoric*, *Rhetoric Review*, *Rhetoric Society Quarterly*, *Rhetorica*, and *Rhetorik: Ein internationales Jahrbuch*. Several dictionaries and encyclopedias of rhetoric, listed below in the Bibliography, have also been published recently.

Among twentieth-century writers on rhetoric, two have had special influence. Kenneth Burke developed his theories of rhetoric in numerous writings over a long career. In *Grammar of Motives* (1945) he described his best known tool of rhetorical criticism, the pentad of act, scene, agent, agency, and purpose. In *Rhetoric of Motives* (1950) he defined the function of rhetoric as 'the use of language as a symbolic means of inducing cooperation in beings that by nature respond to symbols,' and he had much to say about 'identification' between speaker and audience as a rhetorical technique. Burke, an American, came to the study of rhetoric from an interest in language, literature, and political philosophy; Chaim Perelman, a Belgian and author with Lucie Olbrechts-Tyteca of *La Rhétorique nouvelle* (1958), approached rhetoric from study of argumentation and jurisprudence. He distinguished argumentation, which is always addressed to an audience, from formal proof, which is not, and explored topics of argument, divided into those of quantity and quality and those that help to identify and exemplify facts.

An example of a narrow linguistic approach to rhetoric is *Rhétorique générale* (1970), the work of six scholars at the University of Liège calling themselves *Groupe* μ. Here rhetoric was limited to the phenomenon of troping as seen in metaphor and metonomy. Of many modern studies of metaphor, one that is sensitive to rhetorical issues is Paul Ricoeur's *The Rule of Metaphor: Multi-disciplinary Studies of the Creation*

of Meaning in Language. Another new approach to rhetoric is included in *A Theory of Semiotics* by Umberto Eco (1979). Eco described rhetoric from the point of view of semiotics as 'over-coding,' that is, a network of associations invoked by linguistic usages or by the switching of codes.

3. *The Natural Sources of Rhetoric*

Many traditional features of human rhetoric have analogies in animal communication. Social animals in particular use a variety of oral, visual, and chemical signs to persuade others to do what they want, primarily in connection with mating and securing food. Animal communication shares elements of invention, arrangement, style, and delivery with human rhetoric. Many animals perform epideictic rituals, and honey bees engage in a kind of deliberation when swarming. Some intelligent animals learn to understand and even to use metonymy, but the use of metaphor, as well as logical arguments based on citation of examples, is a human development. Rhetoric in some form is a basic feature of human and nonhuman animal life, a tool in the survival of both species and individual. It can be viewed as a form of mental and physical energy, reacting to a challenge, need, or desire by producing and transmitting signs to a real or imagined audience. The simplest natural rhetorical techniques, exploited by animals and humans alike, are gesture, volume, pitch, and repetition of a message.

4. *Rhetoric Without Writing in Traditional Societies*

Ethnologists have collected much information about the forms and function of speech and techniques of persuasion in traditional societies in Africa, Australia, the South Pacific, and the Americas, although they generally use terms like 'political language' rather than 'rhetoric' to describe what they have observed. Formal deliberation is a regular feature of acephalous societies, where individuals gain reputations as orators, and in hierarchic societies recognized orators are employed as spokesmen for kings and chiefs. The goal of deliberation in traditional societies is usually consensus, real or imposed, and opportunity should be offered to opponents 'to save face' by seeming to compromise. Formal language is required of speakers on official occasions everywhere and is learned by imitation of earlier speakers. Such formal language often has archaic elements of diction and grammar, as well as conventions of etiquette and arrangement, and makes use of proverbs, examples, and traditional metaphors. Epideictic speech, as understood in the West, has nonwestern parallels in religious rituals, where extreme forms of formal language, often incomprehensible to the general audience, are employed, frequently accompanied by music and dance. As in official deliberation, formal ritual language seems to certify the authenticity or truth of what is said and helps to perpetuate conservative social control over the society on the part of chiefs and priests. Among many examples of oral discourse Aztec *huehuehtlahtolli* from Mexico show an unusually complex development or rhetorical artistry. Many examples of Aztec speeches were written down by Bernardino de Sahagún and others in the sixteenth century when the genre was still flourishing.

5. *Rhetoric in Nonwestern Literate Societies*

The invention of writing in Mesopotamia and in China, and its spread to Egypt, Greece, India, and elsewhere clearly had some effect on rhetoric, though definition and proof of that effect is controversial. Scribes in Mesopotamia and Egypt celebrated their craft and obtained considerable power as necessary intermediaries between illiterate rulers and the general populace. Writing made possible communication over distance, record keeping, scientific and historical research, crystallization of oral poetic texts, creation of textbooks, and standardization of grammar. It may have facilitated extended logical argument and the use of complex sentences with a variety of subordinate clauses.

6. *Comparative Rhetoric*

The study of rhetoric on a cross-cultural basis is a recent phenomenon. It attempts to identify what is universal and what culturally specific in rhetorical practice across linguistic and cultural barriers, with the aim of formulating a General Theory of Rhetoric as well as to improve cross-cultural understanding. One approach is to test the applicability of western rhetorical concepts, including those outlined earlier in this article, to discursive practices in other societies. Of particular interest to date has been study of rhetoric in China, where thousands of speeches, as well as handbooks of composition, rhetorical exercises, and critical, historical, and poetic writings, dating from antiquity to modern times, provide a rich source for comparative research. One remarkable early thinker was Han Fei-tzu (born about 280 BC), sometimes called 'the Machiavelli of China' because of his pragmatic, cynical teaching to rulers about how to employ rhetoric to solidify their power. In the early centuries of the common era Chinese scholars produced moralizing and allegorizing commentaries on the Chinese classics that employ an extensive technical terminology, sometimes resembling but often differing from rhetorical concepts in the West. Comparable developments occurred in ancient India, where the *Arthashastra* of Kautilya (about 300 BC) outlines the

rhetoric of power and provides extensive technical terminology for teaching and criticizing rhetoric. Poetic and dramatic theory was later developed in India to a high degree of sophistication.

In comparison with other traditions, western rhetorical practice, beginning with the Greeks, has been more tolerant of contention, personal invective, and flattery, less insistent on consensus, politeness, and restraint than elsewhere. The Greeks invented the counting of votes as a way of settling political and legal disputes, something unknown elsewhere until modern times. Although voting performs that function, its potential acceptance of a majority of one as determinative can sharpen disagreement. Artistic judicial rhetoric, with argument from probability and exploitation of circumstantial evidence in court, is largely a western phenomenon; elsewhere, judges have traditionally interrogated the principals and have usually demanded direct evidence or trial by ordeal. Handbooks of public speaking were composed in ancient China, India, and Egypt—the earliest known is the *Instructions* attributed to the Egyptian vizier Ptahhotep of which the earliest manuscript dates from the early second millennium BC—but outside the West rhetoric has always been taught as an aspect of politics, ethical philosophy, or literary criticism, and not as a separate discipline within a standard curriculum of liberal arts, as found in western schools and universities.

7. *The Defense of Rhetoric*

Rhetoric has often been denigrated; the traditional defense, originating with Aristotle, argues that it is in itself neither good nor bad; like the other arts, it is a morally neutral tool that can be used by good or bad purposes. This is sometimes now labeled the 'weak defense' in contrast to a modern 'strong defense' that seeks to justify rhetoric on the ground that ethical and political knowledge is not based in *a priori*, abstract truth but is culturally formed and reformed through rhetorical engagement in concrete situations. The strong defense works best in a liberal democratic society where the majority grant some measure of toleration to minorities. On the global level it seems to justify any values and beliefs in society that are rhetorically constituted and accepted by an audience. That includes fascism, anti-Semitism, misleading advertising, chauvinism, religious bigotry, and most of the ills of history. At this level of description, we then have to fall back to defend rhetoric as a neutral tool that can be used equally for beneficial or for exploitive purposes.

8. *Current Issues in the Study of Rhetoric*

Rhetoric, as practice, was affected by the invention of writing and both as practice and as theory by the invention of printing. The rhetorical implications of the electronic revolution of the late twentieth-century are only now beginning to be assessed. Rhetoric is, almost by definition, an interdisciplinary study. Throughout western history it has had a love–hate relationship with philosophy, and modern students of rhetoric draw extensively from twentieth-century philosophy in their analyses of rhetorical practices and in their efforts to show how rhetoric can be said to generate knowledge. Ideas derived from the epistemology and hermeneutics of Martin Heidegger and Hans-Georg Gadamer have been influential on some scholars, and others show indebtedness to Ludwig Wittgenstein, Michel Foucault, Jacques Derrida, Richard Rorty, Jürgen Habermas, or other recent philosophers, most of whom do not specifically discuss rhetoric. Much rhetorical research at the end of the twentieth century was concerned with issues of gender as rhetorically defined and practiced. Long overlooked evidence for speaking and writing by women has been discovered and is now being studied for the first time. The evidence from the classical period is fragmentary, but medieval, renaissance, and early modern women developed distinctive rhetorical forms, and the contemporary women's movement has employed a variety of rhetorical techniques of its own as well as analyzing the repressive rhetoric of the past.

See also: Aristotelian Social Thought; Language and Poetic Structure; Models, Metaphors, Narrative, and Rhetoric: Philosophical Aspects; Rhetorical Analysis

Bibliography

Aristotle 1991 *On Rhetoric: A Theory of Civic Discourse*, translated by G A Kennedy. Oxford University Press, New York

Bloch M (ed.) 1975 *Political Language and Oratory in Traditional Society*. Academic Press, London

Brenneis D, Myers F R (eds.) 1984 *Dangerous Words: Language and Politics in the Pacific*. Waveland, Prospect Heights, IL

Burke K 1945 *A Grammar of Motives*. Prentice-Hall, New York [Out of print, in print is the paperback edition, 1969, University of California Press, Berkeley CA]

Burke K 1950 *A Rhetoric of Motives*. Prentice-Hall, New York [Only edition now in print is 1969 University of California Press, Berkeley, CA]

Eco U 1979 *A Theory of Semiotics*. Indiana University Press, Bloomington, IN

Enos T (ed.) 1996 *Encyclopedia of Rhetoric and Composition*. Garland, New York

Fumaroli M (ed.) 1999 *Histoire de la rhétorique dans l'Europe moderne, 1450–1950*. Presses Universitaires de France, Paris

Gross A G 1990 *The Rhetoric of Science*. Harvard University Press, Cambridge, MA

Gross A G, Keith W M (eds.) 1997 *Rhetorical Hermeneutics: Invention and Interpretation in the Age of Science*. State University of New York Press, Albany, NY

Groupe μ 1982 *Rhétorique générale*. Seuil, Paris
Kennedy G A 1998 *Comparative Rhetoric: An Historical and Cross-cultural Introduction*. Oxford University Press, New York
Kennedy G A 1999 *Classical Rhetoric and its Christian and Secular Tradition from Ancient to Modern Times*, 2nd edn. University of North Carolina Press, Chapel Hill, NC
Lausberg H 1998 *Handbook of Literary Rhetoric: A Foundation for Literary Study*. Brill, Leiden, The Netherlands
Lu X 1999 *Rhetoric in Ancient China, Fifth to Third Century B.C.E: A Comparison with Classical Rhetoric*. University of South Carolina Press, Columbia, SC
Porter S E (ed.) 1997 *Handbook of Classical Rhetoric in the Hellenistic Period*. Brill, Leiden, The Netherlands
Prelli L J 1989 *A Rhetoric of Science: Inventing Scientific Discourse*. University of South Carolina Press, Columbia, SC
Ricoeur P 1970 *The Rule of Metaphor: Multi-disciplinary Studies of the Creation of Meaning in Language*. University of Toronto Press, Toronto, ON, Canada
Simons H W (ed.) 1990 *The Rhetorical Turn: Invention and Persuasion in the Conduct of Inquiry*. University of Chicago Press, Chicago
Sloane T O (ed.) 2001 *Encyclopedia of Rhetoric*. Oxford University Press, New York
Sutherland C M, Sutcliffe R (eds.) 1999 *The Changing Tradition; Women in the History of Rhetoric*. University of Calgary Press, Calgary, AB, Canada
Swearingen C J (ed.) 1999 *Rhetoric, the Polis, and the Global Village*. Lawrence Erlbaum, Mahwah, NJ
Ueding G et al. (eds.) 1992 *Historisches Wörterbuch der Rhetorik*. Niemeyer, Tübingen, Germany
Vickers B 1988 *In Defence of Rhetoric*. Clarendon Press, Oxford, UK

G. A. Kennedy

Rhetorical Analysis

Rhetoric, the *art of oratory*, was developed in ancient Greece and Rome. Both as a general communication theory and as a collection of practical advices on how to convince an audience, rhetoric exerted decisive influence on the history of Western civilization and culture. In the period from antiquity until the nineteenth century, rhetoric was the single most important subject taught in Western educational institutions. During these more than 2,000 years it changed both in scope and perspective. Corresponding with the growth in written communication, it was gradually transformed into an *art of writing*. In later centuries it became closely associated with the writing of literature while less emphasis was put upon argumentation and persuasion.

Rhetoric deals with communicative situations. Speeches and texts are always considered in their relation to the aims and intentions of the speaker/writer, the subject matter in question, the general mood and prior knowledge of the expected audience, etc. Correspondingly, any kind of analysis which sets out to describe the ways in which a given discourse is determined or affected by its immediate communicative context could be said to be rhetorical in nature. Rhetorical analyses of this kind have always played an important role in literary studies, but during the few last decades similar methods have also been applied to nonliterary texts and even to nonverbal means of expressions, for instance in connection with studies of political propaganda, advertising strategies, as well as mass media in general.

1. The History of Traditional Rhetoric

1.1 Classical Antiquity

Some ground rules of oratory were developed and taught by Sicilians in 485 BC in connection with legal disputes about real estate. This rudimentary rhetoric quickly spread to Athens where a more comprehensive body of knowledge about the art of public speaking was developed from the mid-fifth century and onwards. Greek rhetoric was introduced in Rome in the first century BC as part of the education of older children.

In the Greek period, the so-called Sophists—persons who taught rhetoric, politics, and mathematics for money—played an important role in the development of rhetoric as a formal discipline. They were attacked by Plato for putting more emphasis on style than on substance. Plato's critique raises a question which has followed rhetoric ever since: Is rhetoric a method for discovering the truth, or merely an effective technique for manipulating an audience?

There is, however, no clear answer to this question. The old Greco-Roman rhetoric—as described by Aristotle (384–22 BC; see *Rhetoric*), Cicero (106–43 BC; De Oratore), and Quintilian (c. AD 35–c. 95; Institutio Oratoria)—was a complex social system of interrelated, and in some cases even contradictory, practices (cf. Barthes 1970, Vickers 1988). First of all, rhetoric was an art in the classical, Aristotelian sense of the word, i.e., *a craft or technique*, a collection of rules and advices for speakers who want to convince or persuade an audience. These rules were taught in and transmitted through strong educational institutions. Thus, rhetoric was also a discipline, a *subject*, a body of knowledge. In addition, this body of knowledge constituted a *theory*, or at least a semitheoretical field: it described and classified some important effects of verbal language.

1.2 The Rhetorical System

The old manuals divided rhetoric into five parts: *Inventio*, *Dispositio*, *Elocutio*, *Memoria*, and *Actio*.

Each part covers a specific field of knowledge and aims at training certain skills related to the field in question.

(a) *inventio* (finding what to say) is the part dealing with *argumentation*. Speakers learn to analyze a given subject matter, to find relevant information, and to determine which types of information are the most useful for their purposes. They also learn to convince an audience with quasilogical proofs (*logos*), and to move the audience by appearing to be good and trustworthy (*ethos*) or by stirring the audience's emotions (*pathos*).

(b) *dispositio* (organizing what one has found) deals with *composition*. Speakers learn to organize their speeches in clearly separated sections.

(c) *elucutio* (choosing the proper words) is the part that deals with verbal *ornamentation*. Some elementary rules are presented together with examples of striking images and other verbal devices which can be used to color a text.

(d) *memoria* (remembering what to say) is an introduction to *mnemonics*.

(e) *actio* (performing the speech) contains practical advices on how to *deliver* a text before an audience.

1.3 The Reduction of Rhetoric

The *memoria* and *actio* parts were rather sketchy in the old manuals, and because they were directly related to the practice of oratory, they disappeared as soon as rhetoric became a theory of written discourse. During the Middle Ages the tripartite system of *inventio/dispositio/elocutio* was taught together with grammar and logic. This system collapsed during the seventeenth century, and rhetoric ended up being a part of *poetics*. From now on it was further reduced: *inventio and dispositio*, the parts most closely associated with logic and argumentation, lost in importance and in the end all that remained of the old system was *elocutio*, the part dealing with questions of stylistics.

The *elocutio* sections of the old manuals taught the reader that significant persuasive or poetic effects can be obtained by selecting the proper words and combining them in the right manner. A flat statement of facts can be transformed into a strong, colorful, and convincing message by the use of *tropes* or *figures*, i.e., by substituting single words (*tropes*) or by manipulating phrases consisting of several words (*figures*). However, when eighteenth century linguists and literary theorists began to study and classify the many varieties of discursive ornamentation, they found that most of these poetic effects could be described as variations of two semantic relations, viz. *similarity* and *contiguity*. The similarity-relation finds its clearest expression in the *metaphor* while *metonymy* and *synecdoche* are well-known examples of ornaments based on causal and spatial contiguity-relations. From the eighteenth century and onwards, metaphor gradually became a kind of supertrope representing all the other tropes and figures based on similarity-relations, while metonymy correspondingly became the representative of all the various ornaments based on contiguity-relations (Genette 1970).

2. Modern Rhetoric

2.1 Rhetoric and Semiotics

During the twentieth century several writers have sought to incorporate this reduced version of traditional rhetoric into a broader, more general framework. One of them is the Russian linguist Roman Jakobson who in 1956 presented a structuralist interpretation of *elocutio*, arguing that the metaphor/metonymy-dichotomy corresponds with a general bipolar structure in language (Jakobson 1956). All discursive production is the result of two fundamental linguistic processes based on similarity and contiguity. These metaphoric and metonymic poles in language can be observed in all normal linguistic behavior but, Jakobson argued, they are also manifest in any symbolic process, either intrapersonal or social (for a fuller discussion of the metaphor/metonymy-dichotomy, see the following section).

Jakobson thus opened the field of rhetoric towards sign systems other than language. His own examples were the dream work studied by psychoanalists, the magic rites described by anthropologists, as well as nonverbal forms of art like cubist painting, the films of Chaplin and Eisenstein, etc.

Jakobson exerted decisive influence on French structuralists and semioticians during the 1950s and 1960s, among them Roland Barthes who in a seminal article, *Rhetoric of the Image*, proposed to rethink classical rhetoric in structural terms (Barthes 1964, see also *Semiotics*). Like Jakobson he argued that the immense variety of tropes and figures categorized by traditional rhetoric are merely surface manifestations of a few formal relations which can be shown to be the structural armature in all kinds of verbal discourse as well as in several nonverbal forms of expression. He predicted that it would be possible to establish a general rhetoric which would be valid for all types of discourses, for texts as well as for sound and images.

Since the 1960s there have been several attempts at implementing Barthes' project. The most comprehensive contribution is the work of the so-called *Groupe μ*, a group of Belgian researchers, who have presented a general description of rhetorical tropes and figures on a semiotic/structuralist basis (cf. Groupe μ 1970). In recent years the group has focused on the question of visual rhetoric (Groupe μ 1992).

The semiotic-structuralist interpretation of traditional rhetoric was mainly a theoretical project, but it had important practical and analytical consequences as well. In his 1964 article Barthes argued that there is

a close connection between rhetoric and *ideology*, a view supported by the fact that the classical manuals usually were more concerned with probabilities than with the actual truth of the matter. In order to be convincing one must know one's audience, know what people will or will not accept as probable, in short one has to respect *doxa*, i.e. public opinion, the stereotypical forms of thought which members of the audience rely on in their everyday lives. This means that the rhetorical devices used in a text can be understood as discursive expressions—or exploitations—of the dominant ideologies within a given society and that such ideologies can be brought to light by means of rhetorical analysis.

That rhetorical analysis can be used as a decisive critical tool had already been proven by the American critic Kenneth Burke who in 1939 wrote an influential article on the argumentative strategy in Hitler's *Mein Kampf* (Burke 1941, see also Burke 1962). From the 1960s and onwards Barthes and many other writers followed Burke's example and made rhetorical analysis a part of a general critique of ideology. Using the insights gained by the fusion of semiotics and rhetoric they made systematic studies of the patterns and systems of belief that are presented as 'natural' or 'evident' in newspapers, fashion magazines, popular films, television programs, etc. This form of analysis was particularly efficient in studies of advertising (for an overview see Dyer 1982).

2.2 Metaphor, Metonymy, and Mental Mapping

Elocutio was the only part that was left of the old rhetorical system by the beginning of the twentieth century. There was however a growing interest in this surviving part and particularly in metaphor. Not only did metaphor become one of the most studied subjects in literary theory, the interest spread to several other areas. At present the study of metaphor is one of the most fertile fields in the humanities and the social sciences. It plays a crucial role in academic disciplines as diverse as linguistics, film and media studies, constructivist psychology, anthropology, and cognition theory.

During the last century the conception of metaphor has changed dramatically. From being a poetic device producing a specialized type of discursive effect, it is now generally regarded as a very fundamental mental tool: 'the essence of metaphor is understanding and experiencing one kind of thing in terms of another,' as George Lakoff and Mark Johnson put it in their book *Metaphors We Live By* (1980). Metaphor is the kind of understanding which is produced when a set of relations is displaced from one kind of thing to another. Or to put it another way: Metaphor is a type of *mental mapping*. When people say that 'time is money,' they establish a simple relation of similarity between two elements but this process entails a further mapping of a whole series of relations from the time-field onto the money-field: if time is money, then time must also be valuable, scarce, etc.

According to Lakoff and Johnson there are three major types of metaphor. One is the *structural metaphor*, the 'time is money' variety just mentioned: a restricted, almost local, structuration of one field in terms of another. Another one is the *orientational metaphor* which is used to give larger systems of concepts a spatial orientation. When for example happiness is described as 'being up' and sadness as 'being down,' a spatial relation (up/down) is mapped onto an emotional relation (happy/sad). Such orientational metaphors are clearly grounded in very concrete, physical experiences of the body and its position in space. According to Lakoff and Johnson all metaphors do actually have some kind of concrete experiential basis. One example is the third major type, the *ontological metaphor*, which is based on the everyday experience of handling physical objects. Ontological metaphors make it possible for an individual to understand—and thereby to deal with—events, activities, emotions, ideas, etc. as if they were objects, entities, or substances.

Metonymy, the other supertrope, does also have important cognitive functions. Its basic principle is that one entity *stands for* another. In the phrase 'sails on the water,' a part (sail) is used as a short hand reference to a totality (boat). Thus, a metonymy emphasizes or highlights certain aspects or features within a given totality. In the very act of referring, metonymy thereby serves some of the same purposes as does metaphor.

Metaphor and metonymy are mental tools which enable people to focus their everyday experiences and to understand the world they live in. They are, as Lakoff and Johnson point out, part of the ordinary way in which people think, act, and talk. Sometimes people use these tools in new and original ways, but most of the resemblances they perceive and most of the short-hand references they make in their everyday lives are the result of relatively conventional metaphors and metonymies that have become part of their conceptual system. In later years there have been a growing number of studies of such conventional metaphor/metonymy-based systems and the role they play in public debates and political communication. One striking example is George Lakoff's own analysis of the political rhetoric used by US politicians in the weeks before the war in the Gulf in 1991 (Lakoff 1991).

2.3 Rhetoric, Argumentation, and Thinking

Traditional rhetoric originated in relatively small, closed societies characterized by limited flows of information. Politicians and lawyers who should speak in public were faced with the task of convincing or

persuading uneducated and incompetent audiences. The manuals told them how to do that.

Communication today is still a question of finding and interpreting relevant information and presenting it in effective ways. Nevertheless, the general communicative situation has changed radically since the days of classical antiquity. The most important communication is mediated by radio, films, television, the Internet, etc.—media characterized by what has been called a new type of orality (Ong 1982). Information is distributed at an incredible speed, and the total amount of available information is growing rapidly. Even though the general level of education is considerably higher than anytime before in the history of the world, a modern mass audience is a rather heterogeneous phenomenon consisting of people with very varying degrees of knowledge and competence.

The fear that this complex situation with its rapid flow of information may result in massive manipulation of the audiences provides a significant background for many of the newer attempts at reviving rhetoric. One example is the so-called *New Rhetoric* which was developed in the mid-twentieth century (Perelman and Olbrects-Tyteca 1958, Toulmin 1958). The New Rhetoric was initially an attempt at re-establishing the total rhetorical system in order to be able to cope with all aspects of the new, media-driven types of communication. However, from the very beginning the strongest emphasis was put upon *inventio*, and more specifically on *argumentation*, one of the parts of traditional rhetoric that more or less disappeared when rhetoric became associated with literature and poetics.

In the philosophical tradition argumentation has usually been associated with logic and the techniques of deduction and induction. What the New Rhetoricians have in mind, however, is not this strict, formal, and logical type of argumentation, but the 'fuzzy' thinking, the jumping-to-conclusions known from everyday life. Aristotle excluded this kind of thinking from logic proper but included it in his Rhetoric and actually considered it to be one of its most important parts.

Rhetorical argumentation is a 'situated' form of argumentation, it cannot be performed or understood without taking the total communicative situation into consideration. It is dependent on the subject matter, the level of available information, the pre-existing attitudes and prejudices of the speaker and the audience, the type of media which is used, etc. In short, what is at stake here and what the New Rhetoricians as well as many other, later writers (cf. for example Billig 1987) set out to describe and analyze are the various ways in which practical, everyday reasoning is determined by its immediate context.

Viewed like this rhetoric is much more than a theory of oratory or a collection of simple rules to be used by poets—it is a discipline which studies the underlying principles governing all human discourse, the way in which people argue with themselves and with others, the way in which they make value judgments, etc. Rhetorical analyses of such arguments and of how judgments are made can provide a better understanding of how opinions and social attitudes are formed in complex modern media-driven societies.

See also: Advertising: Effects; Advertising: General; Jakobson, Roman (1896–1982); Language and Social Inferences; Mass Media, Representations in; Media Effects; Media, Uses of; Metaphor and its Role in Social Thought: History of the Concept; Organizations, Metaphors and Paradigms in; Political Communication; Political Discourse; Popular Culture; Reasoning with Mental Models; Rhetoric; Semiotics

Bibliography

Barthes R 1964 Rhétorique de l'image. *Communications* **4** [1977 Rhetoric of the image. In: *Image Music Text*. Fontana Press, London]

Barthes R 1970 L'ancienne rhétorique. Aide-mémoire. *Communications* **16** [1988 The Old Rhetoric. An aide-mémoire. In: *The Semiotic Challenge*. Basil Blackwell, Oxford, UK]

Billig M 1987 *Arguing and Thinking. A Rhetorical Approach to Social Psychology*. Cambridge University Press, Cambridge, UK

Burke K 1941 The rhetoric of Hitler's 'battle'. In: *The Philosophy of Literary Form. Studies in Symbolic Action*. Louisiana State University Press, Baton Rouge, LA

Burke K 1962 *A Grammar of Motives and a Rhetoric of Motives*. World Publishing Company, Cleveland, OH

Dyer G 1982 *Advertising as Communication*. Methuen, London

Genette G 1970 La rhétorique restreinte. *Communications* **16**: 158–71

Groupeμ (Dubois J, Edeline F, Klinkenberg J-M, Minguet P) 1970 *Rhétorique générale*. Libraire Larousse, Paris [1981 *A General Rhetoric*. Johns Hopkins University Press, Baltimore, MD]

Groupeμ (Edeline F, Klinkenberg J-M, Minguet P) 1992 *Traité du signe visuel: pour une rhétorique de l'image*. Seuil, Paris

Jakobson R 1956 Two aspects of language and two types of aphasic disturbances. In: Jakobson R, Halle M (eds.) *Fundamentals of Language*. Mouton, The Hague

Lakoff G 1991 Metaphor and war: The metaphor system used to justify war in the Gulf. In: Hallet B (ed.) *Engulfed in War: Just War and the Persian Gulf*. Matsunaga Institute for Peace, Honolulu, HI

Lakoff G, Johnson M 1980 *Metaphors We Live By*. University of Chicago Press, Chicago, IL

Ong W J 1982 *Orality and Literacy. The Technologizing of the Word*. Methuen, London

Perelman C, Olbrects-Tyteca L 1958 *Traité de l'argumentation; la nouvelle rhétorique vols. 1–2, 1st edn*. Presses universitaires de France, Paris [1969 *The New Rhetoric. A Treatise on Argumentation*. University of Notre Dame Press, Notre Dame, IN]

Toulmin S E 1958 *The Uses of Argument*. Cambridge University Press, Cambridge, UK
Vickers B 1988 *In Defence of Rhetoric*. Clarendon Press, Oxford, UK

P. Larsen

Rheumatoid Arthritis: Psychosocial Aspects

1. Introduction

Rheumatoid arthritis (RA) is a chronic, disabling inflammatory disease with pain and stiffness in the joints. RA with its often unpredictable and painful course has a considerable impact on patients' lives. The disease has physical as well as behavioral and psychosocial implications for the patient and their family. On the other hand, behavioral and psychosocial factors may be determinants of disease development and of patients' ability to adapt to the condition. Over the last two decades there has been an increasing interest in psycho-educational interventions such as patient education or cognitive-behavioral therapies to help patients deal with the consequences of their disease.

2. Disease Characteristics

RA is a chronic disease characterized by inflammation of joints and joint tissues (Wegener et al. 1996). Some patients may experience remission of the disease without any noticeable damage. As the disease progresses and becomes chronic, the inflammation may lead to joint damage and disability. Cardinal symptoms are joint pain and swelling, morning stiffness, malaise, and fatigue. The course of the disease is highly variable. Patients experience periods of remission and exacerbation of the inflammatory activity. Sometimes patients feel very good and in other periods they feel worse.

The prevalence of RA is about 0.5–1 percent of the population in most cultures. The disease can affect people of all ages, including children, but the onset is usually between the ages of 40–50 years. Women are more affected than men; the overall ratio is 3 to 1.

3. Psychosocial Factors Related to the Onset and Course of RA

There has been much research into the role of personality characteristics ('arthritic personality') and stress in the onset and course of RA. However, there is no empirical evidence for the hypothesis that certain personality characteristics predispose individuals to get RA. More recent research shows that psychological characteristics in RA patients are reactions to the physical disease process rather than causal factors.

There are indications that stressful life events are related to the disease process in RA. The mechanisms through which stress exerts its influence on the disease process are still unclear. Recent developments indicate that stress influences physiological processes. Relationships have been found between minor stresses (daily hassles, such as car breakdowns, arguing with spouse) and disease exacerbation. Recent evidence suggests that major life events (e.g., death, divorce) are associated with RA improvement (Huyser and Parker 1998). It appears that physiological processes mediate the relationships between environmental stressors and disease activity in RA. Psychological factors (e.g., coping style, social support) appear to moderate these processes. A greater understanding of the complex relationships between stress and RA requires integrative studies that simultaneously consider environmental, physiological, and psychological factors (Huyser and Parker 1998).

4. Physical, Psychological, and Social Consequences of RA

For the patient with RA it is probably less significant that stress influences the process of his or her disease than it is that RA with its often unpredictable and painful course has a stressful impact on the life of the patient.

4.1 Physical Consequences

Disability and pain are major outcomes of RA. Many patients are restricted in their mobility, both inside and outside of the house, and have problems with activities of daily living such as dressing, bathing, housekeeping, or cooking. They often need the help of others to perform these activities. Most patients see not being able to do things and dependence upon others as the main disadvantages of their disease (Cornelissen et al. 1988, Taal et al. 1993). Pain is generally regarded as one of the most important consequences of RA and a major challenge for patients as well as healthcare providers. For many patients, however, it seems that pain is not their main concern, but the restrictions in daily functioning and the lack of independence, probably for a great deal caused by the pain (Taal et al. 1993).

Another significant problem for people with RA is fatigue. Studies have found that the prevalence of fatigue in adults with RA ranges from 80 to 93 percent (Riemsma et al. 1998). Fatigue has multiple causes that include physiological, psychological, and environmental factors.

4.2 Psychological Consequences

Feelings of depression and anxiety are often mentioned as psychological consequences of RA. Studies

among RA patients have found heightened depression scores on self-report questionnaires such as the MMPI and the Beck Depression Inventory (BDI). Using a structured interview (Diagnostic Interview Schedule), Frank et al. (1988) found that 42 percent of the patients with RA in their study met criteria for some form of depression, and 17 percent met criteria for major depressive disorder. These rates of depression are higher than those found in community populations, but approximate the rates found in other chronic diseases. An important methodological problem in the assessment of the psychological impact of rheumatic disease is that many instruments such as the MMPI or the BDI were originally validated in psychiatric patients. Depression scales may overestimate the prevalence and severity of depression among persons with arthritis due to the endorsement of somatically related items such as 'I felt everything I did was an effort' or 'I am about as able to work as I ever was' (Wegener et al. 1996). Questions about physical symptoms may be accurate to assess emotional distress among psychiatric patients, but in arthritis patients these questions measure symptoms that reflect their actual physical condition.

4.3 *Family Relationships*

RA has a considerable impact on the patient's family. Patients can no longer fulfill all their family roles such as household chores, maintaining family ties, or shopping. Limitations in family role function are related to a diminished psychological well-being among people with RA. The disease affects the relationship with the spouse (see *Family Health*). This relationship does not always deteriorate, often spouses grow closer to the patient (Revenson 1993).

RA patients may have sexual problems, mainly caused by physical limitations, pain, and fatigue (Van Berlo et al. 1999). Stiff and painful knees and limited range of motion can hinder sexual intercourse. It has been found that RA patients have a decreased sexual motivation, but no decreases in frequency of sexual activity and intensity of orgasm compared to controls. Patients with RA experience greater loss of sexual satisfaction over time than controls, and greater functional disability is related to greater sexual dissatisfaction (Van Berlo et al. 1999).

4.4 *Social Contacts*

A chronic disease such as RA can lead to diminished social contacts (see *Social Support and Health*). Reduced mobility, loss of work, or the inability to be involved in hobbies or sports may cause this. It has been found that RA patients have fewer opportunities for contacts with friends and acquaintances, and less satisfaction with these relationships. Cornelissen et al. (1988) established that a reduction of social contacts was mainly found among patients who lived in cities and less in patients living in rural areas. A decrease in social activities among arthritis patients was found not to be strongly related to the severity of the disease (Cornelissen et al. 1988, Taal et al. 1993).

4.5 *Work Disability*

Many RA patients experience problems in their work and often become work disabled (Wegener et al. 1996). After 10 years of disease, between 25 percent and 50 percent of patients with RA have stopped working, and after 30 years of disease about 90 percent of patients are work disabled. The most important individual risk factors for work disability among patients with RA are poor disease status, greater job physical demand, and older age. Other risk factors are lack of control over work pace and other work activities, less education, less support from family members and co-workers, and depression.

5. *Behavioral and Psychosocial Determinants of the Adjustment to the Consequences of RA*

The impact of RA on physical and psychosocial functioning is not directly related to the severity of the disease, but there is considerable individual variation in adjustment to the disease.

Several studies have demonstrated relationships between the use of certain coping strategies and psychological adjustment (Manne and Zautra 1992). Coping can be defined as a patient's cognitive, behavioral, and emotional efforts to manage specific internal or external demands. RA patients' use of emotion-focused coping strategies, such as wish-fulfilling fantasies and self-blame, is associated with poor psychological adjustment, while the use of problem-focused coping strategies, such as information seeking and cognitive restructuring, is related to good psychological adjustment. From studies on coping with pain, it can be concluded that the use of active and problem-focused coping strategies (e.g., diverting attention from the pain, increasing activity level, ignoring pain) is related to less pain and less depression, while the use of passive and emotion focused strategies (e.g., restricting social activities, praying, catastrophizing) is related to more pain and worse psychological adjustment.

Perceptions of control over one's situation, such as self-efficacy or learned helplessness, influence the way people cope and adapt to the stresses of their disease (see *Control Beliefs: Health Perspectives*). It has been shown that self-efficacy and learned helplessness are related to pain, disability, and psychological well-being in arthritis patients (Taal et al. 1996). Self-efficacy refers to beliefs in one's capabilities to success-

fully execute the behavior required to produce a certain desired outcome. Learned helplessness refers to the experience of having no control over one's life, in general, and across different situations. The unpredictable nature of RA and the varying disease activity may cause patients to view their disease as uncontrollable. This often leads to low self-efficacy expectations about needed causes of action to cope with the consequences of the disease or to more generalized feelings of learned helplessness.

Social support, for example from family and friends, can also play a significant role in the adaptation to the disease (see *Social Support and Stress*). A distinction must be made between structural aspects of social support, such as the size of the social network, and qualitative aspects, such as perceived social support. Perceived social support refers to the perceived comfort, caring, esteem, or instrumental help a person receives from others. In several studies with RA patients, it was found that more perceived social support was related to improved psychological well-being (Revenson 1993). In most studies, however, the size of the social network did not seem to have a strong relationship with psychological well-being. An exception is a study by Evers et al. (1997) among recently diagnosed RA patients. They found that patients who have a smaller social network shortly after diagnosis showed an increase in depressed mood after one year. They concluded that an extended social network which probably offers broad sources of support seems to be more important for the psychological adaptation of a patient shortly after diagnosis, when usual patterns of life are disturbed and many life domains have to be reorganized, than at a more progressive stage of the disease.

Social interactions do not only have positive effects on well being: negative aspects of social interactions can decrease well-being. Manne and Zautra (1989) found in a study with female RA patients that critical remarks made by the husband were related to maladaptive coping by the patient, which in turn led to a worse psychological well-being. Problematic aspects of social support (for instance support that is neither desired, needed, nor requested, or when the type of support does not match the recipient's needs) were found to be related to psychological depression (Revenson 1993).

6. *Psycho-educational Interventions*

Psycho-educational programs can teach patients the knowledge and skills they need to manage the consequences of their disease (see *Health Education and Health Promotion*). This should eventually lead to an improved physical, psychological, and social functioning. A wide range of behaviors can be important in the management of RA. Among the behaviors that are thought to influence physical and psychosocial health status are adherence to medication regimens, physical exercise, relaxation exercises, joint protection measures, rest, pain coping strategies, and problem-solving (Taal et al. 1996). A well-known and successful American educational program is the Arthritis Self-Management Program (ASMP; Lorig and Holman 1993). In this program self-efficacy enhancing strategies, such as skills mastery and modeling, were used. Participation in this course resulted in improved self-efficacy, more consistent performance of health practices, reductions in pain, and improved physical functioning and psychological well-being among patients with various rheumatic diseases (see *Health: Self-regulation*). Changes in self-efficacy were related to changes in health status. A Dutch self-management program, partly based on the ASMP, showed beneficial short-term effects on knowledge, self-efficacy, performance of learned skills, physical function, and joint tenderness (Taal et al. 1996). After 14 months, effects were still found on knowledge, self-efficacy, and the practice of physical exercises, but most effects had disappeared.

Next to self-management programs, cognitive-behavioral therapy programs have been developed that are mainly aimed at the teaching of skills to cope with pain. Examples of skills that are taught in these programs are relaxation exercises, active coping strategies, such as diverting attention from the pain, and problem-solving skills (Keefe and Van Horn 1993). These programs have shown promising effects on the performance of coping strategies, pain, and physical and psychosocial functioning. Parker et al. (1995) evaluated an individual stress-management program for RA patients that included relaxation training and instruction in cognitive-behavioral coping strategies for managing stress and pain. Significant improvements were found on measures of learned helplessness, self-efficacy, coping, pain, and disability.

Long-term effects of psycho-educational interventions on pain, disability, or psychological well-being of RA patients are not often found (Keefe and Van Horn 1993, Taal et al. 1996). It is necessary to look for ways to improve the long-term benefits of psycho-educational interventions, for instance by including booster sessions in educational programs, teaching patients skills for the prevention of relapse (e.g., learning to identify high-risk situations and early warning signs, methods for coping with setbacks), or the improvement of social support through the involvement of patients' family members in educational programs (Keefe and Van Horn 1993, Taal et al. 1996). Radojevic et al. (1992) showed that cognitive-behavioral therapy with active involvement of a family member led to greater health benefits among RA patients than cognitive-behavioral therapy without family support.

It can be concluded that psycho-educational interventions that teach patients with RA self-management

skills and strategies for coping with pain and stress show promising results, but ways to improve long-term benefits have to be looked for. However, it must be emphasized that psycho-educational interventions for patients with RA can only be supplementary to medical treatment.

See also: Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Chronic Pain: Models and Treatment Approaches; Environmental Stress and Health; Pain, Management of

Bibliography

Berlo W T M van, Vennix P, Rasker J J, Rijswijk M H van, Taal E, Weijmar Schultz W C M, Wiel H B M van de 1999 Rheumatic diseases and sexuality: A review of the literature. *Rheumatology in Europe* **28**: 113–7

Cornelissen P G J, Rasker J J, Valkenburg H A 1988 The arthritis sufferer and the community: A comparison of arthritis sufferers in rural and urban areas. *Annals of Rheumatic Diseases* **47**: 150–6

Evers A W, Kraaimaat F W, Geenen R, Bijlsma J W 1997 Determinants of psychological distress and its course in the first year after diagnosis in rheumatoid arthritis patients. *Journal of Behavioral Medicine* **20**: 489–504

Frank R G, Beck N C, Parker J C, Kashani J H, Elliott T R, Haut A E, Smith E, Atwood C, Brownlee-Duffeck M, Kay D R 1988 Depression in rheumatoid arthritis. *Journal of Rheumatology* **15**: 920–5

Huyser B, Parker J C 1998 Stress and rheumatoid arthritis: An integrative review. *Arthritis Care and Research* **11**: 135–45

Keefe F J, Van Horn Y 1993 Cognitive-behavioral treatment of rheumatoid arthritis pain: Maintaining treatment gains. *Arthritis Care and Research* **6**: 213–22

Lorig K, Holman H 1993 Arthritis self-management studies: A twelve-year review. *Health Education Quarterly* **20**: 17–28

Manne S L, Zautra A J 1989 Spouse criticism and support: Their association with coping and psychological adjustment among women with rheumatoid arthritis. *Journal of Personal and Social Psychology* **56**: 608–17

Manne S L, Zautra A J 1992 Coping with arthritis. Current status and critique. *Arthritis and Rheumatism* **35**: 1273–80

Parker J C, Smarr K L, Buckelew S P, Stucky-Ropp R C, Hewett J E, Johnson J C, Wright G E, Irvin S W, Walker S E 1995 Effects of stress management on clinical outcomes in rheumatoid arthritis. *Arthritis and Rheumatism* **38**: 1807–18

Radojevic V, Nicassio P N, Weisman M H 1992 Behavioral intervention with and without family support for rheumatoid arthritis. *Behavior Therapy* **23**: 13–30

Revenson T A 1993 The role of social support with rheumatic disease. *Baillière's Clinical Rheumatology* **7**: 377–96

Riemsma R P, Rasker J J, Taal E, Griep E N, Wouters J M G W, Wiegman O 1998 Fatigue in rheumatoid arthritis: The role of self-efficacy and problematic social support. *British Journal of Rheumatology* **37**: 1042–6

Taal E, Rasker J J, Seydel E R, Wiegman O 1993 Health status, adherence with health recommendations, self-efficacy and social support in patients with rheumatoid arthritis. *Patient Education and Counseling* **20**: 63–76

Taal E, Rasker J J, Wiegman O 1996 Patient education and self-management in the rheumatic diseases: A self-efficacy approach. *Arthritis Care and Research* **9**: 229–38

Wegener S T, Belza B S, Gall E P (eds.) 1996 *Clinical Care in the Rheumatic Diseases*. American College of Rheumatology, Atlanta, GA

E. Taal

Ricardo, David (1772–1823)

Among the educated public only a few names of economists stand out: Adam Smith, the spokesman for capitalism; (T.) Robert Malthus, who earned for economics the title of 'the dismal science,' by virtue of his belief that any temporary gain in well-being would self-destruct by inducing population explosion that would entail diminishing returns and a retreat back to a subsistence ration of real wages; Karl Marx, the prophet of capitalism's doom and the exponent of allegedly scientific socialism; John Maynard Keynes, proponent of the Mixed Economy, who rescued capitalism from the post-1929 Great Depression by activistic governmental macroeconomic policies and public regulation.

Within the economics profession, it is a different story. David Ricardo was the archetypical classical economist, the economist's economist who supposedly improved upon Smith and spelled out with Euclidean logic the eternal behavior of the market system. In amiable epistolary debates with his friend Malthus, Ricardo used to be reckoned the hands-down winner: Malthusian fears of recurrent depression stemming from under-consumption spending were argued to be fallacies. Like poet Robert Frost, beloved by both highbrows and lowbrows, the self-made millionaire David Ricardo captured the respect of the Victorian devotees to *laissez faire* as well as socialistic critics of Marxian or romantic stripe.

Schooled only in the Darwinian jungle of speculative finance, the broker Ricardo wrote as an autodidact, often in an obscure style that impressed his limited number of readers as being profound. Since 1817 literally scores of articles have appeared in learned journals debating whether or not David Ricardo did believe in 'a labor theory of value.' This is the stuff of which PhD theses can be made.

Ricardo was also lucky in having the learned Piero Sraffa edit for the Royal Economic Society, over a period of 25 years, a magnificent and complete edition in 11 volumes of *The Works and Correspondence of David Ricardo* (1951–73). Although this brought Ricardo's name back into the limelight, oddly it seems to have contributed to a revisionist erosion of his analytic reputation. Whereas the late-Victorian Alfred Marshall, pillar of post-1870 neoclassicism, had idiosyncratically argued that the good modern stuff was

already in Ricardo, and whereas Keynes had written in the 1920s that David Ricardo's was 'the finest mind that had come to economics'—neither Ricardo's late-twentieth-century admirers nor detractors could agree among themselves as to what his different virtues and vices actually were. A scholar who is much misunderstood cannot be acquitted of all responsibility for ambiguity.

Ricardo's parents were Sephardic Jews who migrated to England from Amsterdam in the mid-eighteenth century. His ancestors had gone from Spain to Italy and then on to Holland. They seem to have been for the most part successful dealers in finance, an occupation that young David took up at the age of 14. When he married a Quaker, he was cut off by his father with the proverbial shilling; but within a score of years, working as a broker in the City of London, he had by his early forties amassed a comfortable fortune—the period of the Napoleonic Wars offered windows of opportunity for an alert and nimble financier. During vacation, when he picked up by chance Smith's *The Wealth of Nations* (1776), Ricardo was hooked for political economy.

Reversing the usual order of study, Ricardo first made his mark in what has come to be called *macro*economics. His unqualified verdict typifies his non-eclectic style. To finance a very long war against Napoleon, Britain expanded its money supply of paper currency in order to procure goods at home and abroad for its armies. Price-level inflation, not surprisingly, was part of the economic history in that period. Eclectic Smithian theory would no doubt have predicted as well a worsening of Britain's terms of trade as she and other belligerents bid up the relative prices of needed imports. Had gold and silver been the sole money used then in England and elsewhere, the 1750 theories of David Hume would have predicted a drain of specie from London to pay for the war-induced trade deficit of exports compared to imports.

However, even in those distant times much of the money employed for purchase and sale transactions involved use of paper-money currency. And, as is customary in times of war, whatever the supply of domestic gold or silver, there ensued a considerable rotation of the printing press to finance government needs and accommodate the desire for the enhanced cash necessary to conduct the enlarged volume of goods transactions and their higher average prices. Under these circumstances, whenever a government chooses *not* to redeem paper currency notes into specie at their previous face value, those notes will depreciate in price relative to ounces of specie. The high price of gold is what the person in the street calls this phenomenon and passionate voices debated its cause(s). Young Ricardo attributed *all* of the rise in gold prices to over-issue of paper notes.

Today one can write counterfactual history. Somehow the British government might have taxed the populace more heavily to finance in a balanced-budget fashion the enhanced war effort and the irreducible rise in relative import prices; imposed tightness in bank lending, enforceable only at higher interest rates, conceivably could have kept the stock of paper currency unchanged, at the same time almost certainly imposing unemployment and extra loss of consumption on the civilian populace. With heroic tightness, trade balances might even have been forced to cancel out to zero. (It is hard to envisage how normal peacetime unemployment levels could have somehow been made to prevail from say 1795 to 1810 in this contrafactual scenario.) An inflation-free war can thus be painfully conjured up in contrafactual history; and with no wartime increment of paper-currency issue.

In actual history, paper notes did depreciate and price levels did soar. As happened a century later during the 1920–23 German hyperinflation, economists lined up in opposing camps. Young David argued: over-issue in currency M explained virtually *all* of the rise in the price of bullion. Henry Thornton, practical banker and Bank of England authority, (at first) argued that the microeconomic phenomena of adverse trade balances and war-induced buoyant business activity explained a significant fraction of the observed appreciation of bullion.

Ricardo's 1809–10 strong assertion of what was called the 'bullionist view' won him instant celebrity, particularly with the dogmatic James Mill. But this economic debate was not really a zero-sum game in which Ricardo's gain had to equal Thornton's loss.

Both men agreed on the likely longest-run outlook. The price of gold in terms of paper would be the same in 1850 if, in the many preceding decades, the supply of paper currency was limited enough to permit free specie–currency convertibility at the old 1790 ratio. The so-called 'Ricardian vice' was his use of *long-run* true relations to characterize actual short-run patterns in economic history.

After the first decade of the nineteenth century, Ricardo concentrated mostly on *microeconomics*. In 1815 (and *not* prior to then), along with Malthus and Edward West he helped articulate the 'law of diminishing returns' and its consequences for land-rent determination. Society's product, call it corn, is produced by labor and by land, both working together. When you increase labor in two *equal* increments, working on fixed land, you acquire two successive increments of corn. But the second gain is presumptively *less* than the first—a technological law of nature. Thus the wage that the competitive market can pay to labor goes down when labor/land density increases; and the rent that the market will pay for land will go up. This is not a new idea—it is in Benjamin Franklin (1755) early in the eighteenth century and already in Smith (1776)—but it is an important insight that needed being made explicit.

Already Ricardo had perceived what he considers faults in Adam Smith. Posterity can be grateful that

this perception tempted Ricardo into economics, even though it can be argued convincingly that Smith is mostly the more correct thinker in the cases where Ricardo criticizes him. For example, in the capital-and-land-and-produced-capital models that both writers worked with, Smith was right to jettison the 'labor theory of value' as soon as the interest rate and/or the land rent rate becomes significantly positive.

As between the two tasks of political economy—(a) understanding how relative scarcities of factors of production constrain the alternative availabilities of different goods and their valuation relations; and (b) how society's harvest gets *distributed* among and between the different social classes of workers and owners of property—Ricardo arbitrarily declared it to be distribution that was all important. Yet nowhere in the 11 volumes of his *Works and Correspondence* (1951–73) will the reader find a *complete* and satisfactory theory of distribution that goes beyond what is in Smith, Malthus, Mill or modern mainstream writers.

Perhaps Ricardo's lack of a university education contributed to his methodological style. Frequently he asserts general 'truths' that do admit of important exceptions; repeatedly he fails to distinguish between *necessary* and *sufficient* conditions for an argument. He is a superb miniaturist but with an absence of coherence, particularly in his post-1815 textbook phase. (Ricardo's theory of comparative advantage in international trade, similar to that of the contemporary Colonel Robert Torrens, is a brilliant example of his miniaturist virtuosity.)

If it be thought that, by use of hindsight, I am a bit harsh on 1810–23 David Ricardo, let it be understood that I am even harsher on Ricardo's twentieth-century commentators: the brightest and the best of them—Piero Sraffa, George Stigler, Samuel Hollander, Mark Blaug and scores of others—while understanding that the positive interest and time-phasing of production do generically invalidate Ricardo's championing of the *labor theory of value* (as, indeed, Ricardo himself intermittently admitted), at the same time seem blind to the parallel fact that positive rent(s) on scarce land(s) must similarly invalidate the labor theory of value. (Edwin Cannan, Knut Wicksell, Jacob Viner and Lionel Robbins are honorable exceptions to this indictment of blindness about how labor/land intensities do generically invalidate any labor theory of value in the same logical way that time-intensities are known to do so.)

Around 1815, between his macro works on gold and his *Principles of Political Economy and Taxation* (1817, 1819, 1821), Ricardo did puzzle over a simple scenario involving a farm good (corn) and one or more manufactured goods (cloth, …), all produced by, say, homogeneous labor, homogeneous land and by one or more *produced* inputs (corn seed, raw materials from farming used for manufactures, durable tools). He may even have worked out, in an unpublished manuscript that has not survived, a one-sector farm-only model. In it, a determinate model of distribution could have cogently defined long-run stationary-state equilibrium for (real corn wage rate, interest rate, real corn rent rate) = $(W/P_{\text{corn}} = \bar{w},\ \bar{r},\ Rent/P_{\text{corn}} = R^*)$. How would it have gone?

Like Malthus and other predecessors, Ricardo would have specified *ex*ogenously a classical 'subsistence wage' rate, $\bar{w}$, needed to keep population just reproducing itself. Like classical successors he would have specified *ex*ogenously a minimal interest rate, $\bar{r}$, needed to keep capital(s) just reproducing themselves. Land being a classical constant, $\bar{A}$, its stationary equilibrium rent rate he would *residually* determine once variable labor-and-capital stocks, L_t and K_t, dynamically converged to their long-run, asymptotic stationary equilibrium levels: $L_t \to L^*$; $K_t \to K^*$, when K is a single scalar. Finally, short-run determination of w_t and r_t would be market-clearing rates dependent on the relative transitional abundances of L_t and K_t.

Here is how in the spirit of Prokofiev's 'Classical' Symphony, an ambitious neophyte in a 1932 Hicks–Wicksell *neo*classical workshop might synthesize these classical insights into a coherent simplistic dynamic equation system, using the notational convention $dY_t/dt = \dot{Y}_t$:

$$\text{factor supplies: } \dot{L}_t/L_t = |\bar{\alpha}|(w_t - \bar{w}),$$
$$\dot{K}_t/K_t = |\bar{\beta}|(r_t - \bar{r}),\ A_t \equiv \bar{A} = 1 \qquad (1)$$
$$\text{factor demands: } w_t = \bar{\sigma}K_t^{1/3}/L^{2/3},$$
$$r_t = \bar{\sigma}L_t^{1/3}/K_t^{2/3} \qquad (2)$$

where $\bar{\sigma}$ is an exogenous parameter denoting level of technical productivity.

When the modern crank of deduction gets turned, Ricardo's (1) and (2) do lead convergently to a determinate classical stationary state. From any *initial* $(L_0\ K_0\ \bar{A})_0$, as $t \to \infty$,

$$(w_t, r_t; L_t/\bar{A}, K_t/\bar{A}; R_t) \to (\bar{w}, \bar{r}; \tfrac{1}{3}\bar{\sigma}/[\bar{w}^2\bar{r}],$$
$$\tfrac{1}{3}\bar{\sigma}/[\bar{r}^2\,\bar{w}]; \tfrac{1}{3}\bar{\sigma}[r\ w]). \quad (3)$$

One must commend the classical tradition for so subtle an accomplishment. (Note the version worked out here differs from the Samuelson (1978) 'canonical' model, which had more faithfully adhered to the conventional labor-capital 'dose' version.)

This above supply–demand equilibrium is definitely *not* independent of the problem of how consumers wish to allocate their earned incomes among the many goods, when there are many non-corn goods in the picture. Then in it there are no 'natural prices' definable independently of how consumers choose to spend their incomes.

Definitely goods' relative price ratios do not remain close to proportionality with their respective embodied-labor contents, $P_t^i/P_t^j \neq (L_t^i/q_t^i)/(L^j/q_t^j)$.

Likewise embodied land contents $(A_t^i/q_t^i)/(A_t^j/q_t^j)$ and embodied capital contents $(K_t^i/q_t^i)/(K_t^j/q_t^j)$ similarly fail as approximations to empirical actual P^i/P^j ratios.

Since Ricardo did not coherently understand the intricacies of his own classical scenario, he never realized how unattainable was his hankering to have all relative prices determinable purely *by objective technology and cost data*. *Subjective* variabilities (a) of consumers' utilities and demand choices and (b) of intertemporal saving-and-consuming choices—which became central preoccupations of post-1870 neo-classical successors to the classical school—were thus *already* unavoidable in the classicals' own capital, land and labor models and in the competitive scenarios that Marx grappled with all his working life. This inevitability will come with pain for the small sect of Sraffian neo-Ricardians who proliferated after 1925 and especially 1960.

Ricardo was low-keyed and civil in argumentation, albeit firm in holding to his strong positions. The letters between Ricardo and Malthus, notwithstanding these scholars' deep differences, are a model of courtesy and affection. Despite Ricardo's typical unqualified dicta, he did not hesitate to change his mind and admit to previous mistakes. Marx, who abhorred classical authorities and for whom 'Parson' Malthus was a particular *bête noir*, was uncharacteristically soft toward Ricardo. Here is an illustrative incident. After Ricardo's first edition of *Principles*, but prior to its third edition, David, Humpty-Dumpty like, had declared that technical invention *must always* help *all* factors of production, including the working man. He encouraged disciples, like J. R. McCulloch, to affirm this strong view. However, on reconsideration, Ricardo came to realize that some inventions can assuredly *replace labor and reduce the demand for labor in the short and the long run*. Therefore, despite friends' warnings that this would weaken the case for *laissez faire*, Ricardo added a new chapter on Machinery to the third (1820) edition of the *Principles*, presenting exposition and numerical examples designed to show how labor's share of national income could suffer from certain possible inventions and how long-run equilibrium populations and gross-income levels could be permanently lowered. With virtually no exceptions, ancient and twentieth century authorities agreed that for once Homer had nodded. For once David must have invoked rigidity of wage rates with its resulting unemployment, thus for once denying J.-B. Say's sacred dogma that over-production is impossible in a capitalist system. Karl Marx praised the scholar for his honesty.

It is a story that casts discredit on virtually all. The Ricardo–McCulloch original sweeping position was gratuitously wrong. A new wind that raises an economy's *potential* for production and consumption most certainly can do harm to *some* competitors while helping others. (Ricardo did not, as J. S. Mill and Vilfredo Pareto were later to do, argue that gainers could always bribe losers so that *with intervention all could gain*.) Nor did Ricardo's new understanding motivate him to abandon a dogmatic defense of *laissez faire*. His critics on Machinery, both then and in our time—including Knut Wicksell, Nicholas Kaldor, Joseph Schumpeter, George Stigler, …—failed to see the absence in Ricardo's exposition of any denial of dogmatic Say's Law and its ruling out theoretical harm from under-consumption; instead his usual classical scenario of population decline dictated by the induced drop in the wage below an alleged equilibrium *subsistence* rate would, as the Machinery chapter alleged, call for an ultimate drop in population and gross product as a result of labor-saving invention(s). Just as the industrial revolution was beginning to raise wages by a genuine filter-down process, the philosophers of the Chair were getting first intimations of the complexity of how a market system metes out its distributive awards.

In that long ago age before antibiotics, David Ricardo in 1823, at the peak of his powers, died from an ear infection. Thus he never lived to see the 1836 Repeal of the Corn Laws, a vindication of his pamphleteering for free trade; and he did not live to witness Ireland's great famine of the 1840s, with its brutal confirmation of the laws of the free market.

In summary, David Ricardo, like the twentieth-century economists Friedrich Hayek and Milton Friedman, importantly pushed voters and public opinion toward libertarian *laissez faire*. Victorian England's Whig society and Manchester School owed much to him. In the modern debate about how a modern mixed economy can optimally compromise between market mechanisms and democratic rules of the road, Adam Smith and John Stuart Mill seem more subtle classical thinkers. From these two, rulers like Lenin, Stalin and Mao could have had the most to learn. But among the worldly philosophers, David Ricardo must be counted as an important shaper of the twenty-first-century mind.

See also: Asset Pricing: Emerging Markets; Economic Growth: Measurement; Economic Growth: Theory; Economics, History of; Economics: Overview; Income Distribution; Innovation and Technological Change, Economics of; Libertarianism; Malthus, Thomas Robert (1766–1834); Market Structure and Performance; Marx, Karl (1818–89); Marxian Economic Thought; Monetary Policy; Political Economy, History of; Regulation, Economic Theory of; Smith, Adam (1723–90); Stock Market Predictability

Bibliography

Franklin B 1755 *Observations Concerning the Increase of Mankind and the Peopling of Countries*

Keynes J M 1936 *The General Theory of Employment, Interest and Money*. [Reprinted in: 1977–81 *The Collected Writings of John Maynard Keynes*. Macmillan, London]
Malthus T R 1798 *An Essay on the Principle of Population as it Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet and Other Writers*. J. Johnson, London [Reprinted by Macmillan, London, 1926]
Marshal A 1890–1920 *Principles of Economics*. Macmillan, London
Marx K 1867, 1885, 1894 *Capital*, Volumes I, II, and III. Verlag von Otto Meissner, Hamburg; [Penguin Books, Harmondsworth 1976, 1978, 1981]
Niehans J 1990 *A History of Economic Theory. Classic Contributions, 1720–1980*. The Johns Hopkins University Press, Baltimore
Ricardo D 1817 *Principles of Political Economy and Taxation*. John Murray, London
Ricardo D 1951–73 *The Works and Correspondence of David Ricardo* [ed. P. Sraffa with M. H. Dobb] 9 vols. Cambridge University Press, Cambridge, UK
Samuelson P A 1978 The canonical classical model of political economy. *Journal of Economic Literature* **16**: 1415–34
Samuelson P A 1989 Ricardo was right! *Scandinavian Journal of Economics* **91**: 47–62
Schumpeter J 1954 *History of Economic Analysis*. Oxford University Press, New York
Smith A 1776 *An Inquiry into the Nature and Causes of the Wealth of Nations* [ed. E. Cannan]. The Modern Library, New York, 1937
Sraffa P 1926 The laws of returns under competitive conditions. *The Economic Journal* **36**: 535–50
Sraffa P 1960 *Production of Commodities by Means of Commodities*. Cambridge University Press, Cambridge, UK
Thornton H 1802 *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. Reprinted, Frank Cass, London; Augustus Kelley, New York, 1962

P. A. Samuelson

Rights

No concept is more central to our moral and political thought than that of rights. We talk about our rights to life, liberty, and freedom of expression; we discuss the rights of children, animals, illegal immigrants, and prisoners of war; we wonder whether we have rights to health care, a decent wage, or physician-assisted suicide. We seem to make sense of a myriad of legal and moral rights, even as we debate their boundaries. Yet few concepts so central to our thought are more deeply puzzling. The clearest evidence for this is the persistent disagreement over the content of even the most basic rights. Is a right to a free press a basic human right? Many people around the world would deny that it is. Is there a basic or human right to health care? Who has the duty to provide it, and on what grounds? Do parents have a right to discipline their children? Do workers have a human right to 'periodic holidays with pay,' as one United Nations document asserted? The problem is not simply that such assertions are controversial, but that we can feel so quickly at a loss about what to say next. Clearly, the concept of rights has been debased in recent years, as rights claims have become mere attempts to give rhetorical force to strongly held preferences, but the problem also lies in some deeper puzzles about the nature and justification of rights.

The least controversial rights are those created by laws and the shared norms of our institutions and practices, where rules allow us to coordinate and regulate our activities and to divide fairly the benefits and burdens of social cooperation. Any attempt to understand the concept of rights should begin here. If Smith gives tools to Farmer in exchange for a share of the grain Farmer will produce, then Smith has a claim to the grain after the harvest, and Farmer has a duty to turn it over. It seems natural to say that Smith has a right to the grain, and laws will respect and enforce such rights.

Governments have a right to tax their citizens to provide public goods, but laws also protect citizens by creating rights against other citizens and against the government itself. Civil society is founded on the concept of property, which is nothing more than a set of practices that allocate entitlements—in other words, rights to use and exchange things of value. Many of the functions of government create rights in this sense, by establishing entitlements and enforcing agreements.

1. *Origin of the Concept*

If it is possible to have a concept without having a word to denote it, as one view of rights maintains, then the idea of rights goes back to primitive societies and can be inferred from ancient texts. According to this view, any society with an institution of property understands the ideas of entitlements to use and claims against infringement, which are part of a theory of rights. Ancient laws also established other liberties and privileges. The Roman law of persons, for example, recognized different status for different members of society. The citizen, the resident alien, and the slave were all granted different powers and liberties, and the law determined when these people could change or lose their status. Romans thus had (and could forfeit) various personal rights.

A different view argues that rights did not exist until sometime in the late Middle Ages, when we first find a word to denote the concept (Tuck 1979). There is an important difference between the implicit, ancient understanding of practices that we regard as rights and this more explicit idea. According to this view, the ancient meaning is restricted to an objective sense of right, which refers to a state of the world. Ancient and medieval societies were dominated by ideas of the

natural differences between people. If someone was properly a citizen or a slave, or later a lord or a servant, their social rank determined their powers and privileges. Their status was meant to reflect the objective or correct natural order of the world. The concepts that come closest to describing this objective sense of right—*dikaios* in Greek and *ius* in Latin—imply an idea of a proper natural order.

The modern notion of rights, as we might call it, developed as the idea that the natural or inherited differences in status could be justified was replaced by a belief in the fundamental moral equality of all people. Historians note that this idea of human equality was not new in the late Middle Ages. It appears in the writings of the stoics and early Christians, but these ancient philosophies did not attach much importance to earthly life. As traditional social orders began to crumble in Europe, the natural differences between people came to be seen as morally arbitrary. Justice required correcting these arbitrary differences with a more egalitarian distribution of powers and opportunities. As moral thought focused on the more worldly concerns of social and political interactions among people, the subjective idea of a right emerged, as something that belongs to individuals. Thus, where ancient and medieval philosophers would say that it is right that Smith controls some thing, modern theorists would say that Smith has a right to that thing (Dagger 1989, Tuck 1979).

Rights in the subjective sense express the moral importance of the equality of persons. They justify both the claims we make against others and the duties that arise in respect of those claims. But this idea of rights is no less puzzling than that of a natural moral order. Do rights exist? What is their content and justification?

2. *An Analysis of Legal Rights*

The revolutions in the USA and France in the eighteenth century were fought in part as a defense of the natural and inalienable rights of humans. Liberal republics and democratic governments based on these rights were established, first in the USA and Europe, and then elsewhere. By that time, the modern idea of rights had fully taken hold in law, morality, and political theory.

Early in the twentieth century, Wesley Hohfeld categorized legal rights in a way that has profoundly influenced subsequent moral and legal theories (Hohfeld 1923). He argued that an assertion of the form 'Smith has a right to P' is ambiguous and could mean any of four different things (or some combination of them). It could mean, first, that Smith has a *privilege* to do P, or that he is under no obligation not to do P. Smith's right to find solace walking in a public park is a privilege, or as it is sometimes called, a mere liberty. It does not imply any corresponding duties on the part of others. The fellow on roller blades who insists on circling Smith in the park, with his boom box blasting, is also exercising a privilege, even though his activity undermines Smith's efforts to find solace. Hobbes' claim that in the state of nature 'every man has a right to every thing' has been interpreted as a privilege in Hohfeld's sense. Hobbes was asserting merely that in the state of nature a person could not be criticized as irrational for laying claim to any thing, although nobody else has a duty to provide him with any thing that he might claim (Waldron 1984).

Hohfeld's second meaning of 'Smith has a right to P' is that Smith has a *claim* to P, which imposes corresponding duties, either positive or negative, on other people to allow or enable Smith to have P. Legal theorists following Hohfeld analyze claims further by distinguishing between rights *in personam* and rights *in rem*. This distinction has implications for the kinds of duties implied by a claim. Claim rights *in personam* impose duties on particular individuals. A contract is such a right. If Smith agrees to work for Miller, then when Smith completes the work, he has a claim that requires Miller to pay the agreed wage. Claim rights *in rem* impose duties generally on all others. Property rights are an example. If Smith owns a car, then every other person has a duty not to interfere with Smith's use of the car. Each person's right not to be assaulted is also a claim right *in rem*, which means that in respect of this right, each person imposes a duty on all other people not to harm him in certain ways. In order for such rights to seem plausible, the duties must typically be negative, i.e., duties to refrain from doing something. Unlike positive duties, the millions of negative duties we impose on each other in respect of our rights *in rem* do not create any unreasonable burden. We can fulfill all of these duties simply by minding our own business and not assaulting anybody.

Hohfeld's two remaining categories of rights apply more specifically to legal contexts. 'Smith has a right to P' could mean that the law gives Smith certain *powers* to alter legal relationships, which do not entail corresponding duties. Smith exercises such powers, for example, when he sells or transfers property. Finally, 'Smith has a right to P' can mean that Smith has an *immunity* from certain kinds of legal change, i.e., he is protected from any or all other authorities altering his legal position. Constitutionally guaranteed privileges or inalienable rights are examples of immunities. Hobbes' argument, that when humans enter into civil society and give up their rights to the Leviathan they nevertheless retain the right to defend their life under all circumstances, is a combination of a privilege and an immunity.

3. *The Nature of Moral Rights*

That legal systems and other institutions give rise to rights is unproblematic, but rights also play a more fundamental role in moral and political theory, where

they are often regarded as existing independently of and prior to social institutions. The Declaration of the Rights of Man and the Citizen in France states that 'the end in view of every political association is the preservation of the natural and imprescriptable rights of man,' and the Declaration of Independence in the USA asserts that all citizens have inalienable rights to life, liberty, and the pursuit of happiness. Natural rights constitute the moral basis and justification of government and laws. They express and give meaning to the belief that all people are morally equal.

Claims to natural rights have always been controversial, however. Jeremy Bentham (1834) famously dismissed them as 'pernicious nonsense,' and wrote that 'a natural right is a son that never had a father.' The controversy is about the nature of basic moral rights and their connection to duties, as well as about the justification of basic or natural moral rights. We will address these issues in order.

The defender of basic moral rights need not assert that all moral duties are correlated to rights. Kant, for example, discusses imperfect duties, such as the duties to be charitable or kind. No individual can claim the performance of such duties as a right, and governments may choose not to enforce them. We also have legal duties, such as a duty to pay taxes, which we are morally obliged to fulfill, but the performance of which no individual can claim as a right.

Whether moral rights must be correlated to duties is more controversial. Many theorists claim that they are, but Joel Feinberg (1980) has argued that moral rights may exist without any corresponding duties. Feinberg distinguishes between what he calls justifiable rights claims and valid rights claims. The former can exist when there are good reasons for a right to exist, even if nobody is in a position to provide what the rights-holder may claim. The idea of human dignity underlies some human rights, for example, which in Feinberg's view justifies claims that certain basic human needs should be met. But in some parts of the world today, nobody is in a position to meet those needs, so they cannot be required as a matter of duty. These rights should be interpreted as developmental goals, the fulfillment of which governments and other agencies must assign the highest priority. Feinberg calls such rights 'aspirational.'

A valid right in Feinberg's sense exists when the claim is justifiable and the corresponding duty can be specified and acted upon. In such cases, the right is logically prior to the corresponding duty. Thus, if Smith has a right *in personam* against Jones, then Jones has a duty in respect of this right. Smith's claim creates the duty for Jones, and if Smith forfeits the right or chooses not to claim it, then unless some other party has the authority to make the claim on Smith's behalf, Jones' duty does not exist.

We can thus interpret moral rights as valid claims, which can be created by law or else founded on the moral belief in human equality and dignity. In order to explain rights in this way, however, we must be careful not to beg questions about the existence of rights in our understanding of claims. To claim a right is not merely to claim a truth, in the sense that to make a claim can be to assert the truth of a proposition. Claiming rights must be understood as involving a different kind of activity, more like claiming a patent for an invention or a site for a homestead in a newly settled territory. To claim something in this sense is to perform an activity different from asserting, which can create legal and moral consequences. To claim one's basic rights is to demand recognition of one's status as a moral equal, which imposes duties on others in respect of one's claims. Thus, following Feinberg (1980), we can say that a moral right is a claim, the recognition of which is called for by laws, moral principles, or an enlightened conscience.

4. *Two Views of the Justification of Rights*

This article has been describing rights by explaining their role in societies in which they are recognized. This argument might also count as a justification of rights or an argument, so to speak, for their normative existence. But to this point the argument sidesteps one of the most puzzling aspects of rights, which is about the nature of the values they promote or express. Are rights merely means for promoting other goods and preventing evils, or do they also have intrinsic value? This question has received considerable attention in recent philosophical literature, and most attempts to answer it are controversial. The following sections will consider defenses of each position.

The first approach argues that the value of rights is instrumental. Rights promote goods and prevent evils, the value of which can be explained without any reference to rights themselves. According to this account, rights provide people with certain social guarantees that are necessary for living happy and fulfilling lives. They are part of the institutional structures that give people power over their own lives and protect them from the abuses of governments or other groups.

In order for these guarantees to be effective, we may have to develop the strict prohibitions against doing certain things that rights typically require. We must refrain from treating people in certain ways, not only when the consequences would clearly be bad, but also when we believe in a particular case that the act would produce the best consequences overall. This is because deciding what to do on the basis of calculating the value of our actions in each case would produce consequences that are worse than those produced by institutions that rule out such calculations.

There is much to be said for an instrumental justification of rights. In a utilitarian theory, for example, intrinsic value is located in the aggregation of individual happiness or well-being, and rights play

an important role in promoting this value. Rights that establish property and legally enforceable contracts are effective ways to coordinate individual actions and foster ties of cooperation and trust. Rights that protect individual liberties give people the opportunity to promote their own happiness, which is generally a more effective way of achieving happiness than trusting the task to others. Rights that protect the freedoms of conscience and expression enable people to discover important truths that are necessary for reducing suffering and increasing happiness. In relatively affluent societies, a utilitarian might also defend positive rights to food, shelter, health care, a minimum income, and other basic needs, because the redistributive effects of recognizing such rights can increase social welfare. A utilitarian would thus defend most of the rights we take for granted, as well as many that are controversial.

Critics of utilitarianism may defend many of these same rights, but they will disagree about their justification. The utilitarian justification of rights depends on empirical assumptions that make it the case that rights are effective means for promoting happiness or welfare. But these assumptions may fail to hold, or the facts may be such that violating the rights of some would produce greater happiness overall. If rights are to have a more fundamental and secure role in political and moral thought, according to these critics, then they must be justified by appealing to other, more basic values.

Ronald Dworkin (1977) argues that the fundamental requirement of any legitimate political system is that it treats all people with equal respect and concern. Equality of respect means that, except in extreme circumstances, it is wrong to require some people to sacrifice their interests for the greater happiness of others. Rights play an essential role in securing equality, because they restrict the individual preferences that social and political decisions may legitimately consider, thus guaranteeing to all people a measure of independence from decisions aimed at promoting social welfare or the general good.

The kind of preferences that rights aim to restrict are primarily what Dworkin calls external preferences. These are preferences about how other people should be treated or what they should be permitted to do, because of who they are, because they have interests and desires of which others disapprove, or because their suffering is a source of happiness to others. Dworkin argues that unless a political system disallows these external preferences, it fails to show equality of respect. Under such conditions, a utilitarian concern simply to maximize the satisfaction of individual preferences lacks political legitimacy. Rights thus have a kind of lexical priority in Dworkin's system, which he expresses by saying that rights trump utility.

Another critic of utilitarianism, T. M. Scanlon (1977), argues that the basic values that a political system should promote must include not only what happens to people but also factors that affect the ability of individuals to determine what will happen to them. People care about their ability to influence outcomes, and they want to be protected from interference or control by others. The goal of rights is to foster and protect this kind of autonomy.

If Scanlon's view is to be distinguished from a sophisticated utilitarianism, however, then the value of autonomy must be defended as an objective good that does not depend on the strength of the preference for it among the people that live in any society. Scanlon argues that moral values must be determined not merely by an aggregation of preferences but by an ethically significant objective notion of their importance. To value autonomy is 'to be concerned with the conditions necessary for people to develop their own aims and interests and to make their preferences effective in shaping their own lives and contributing to the formation of social policy' (Scanlon 1977, p. 85). Rights protect people from the paternalistic interventions of others, and they ensure an acceptable distribution of control over important factors in our lives. Rights would protect religious freedom, for example, even in societies where the majority of citizens care passionately about making their religious views mandatory, while the minority who don't share these beliefs also have tepid preferences about religion. A concern for autonomy implies that people must be free to make their own decisions about religion independently of the control of others.

James Griffin (1999) argues for a different but closely related non-utilitarian justification of rights. He appeals simply to the value of human agency. Griffin argues that a fundamental moral value, more important even than happiness, is the distinctly human ability to contemplate and plan a future. We deliberate, assess, and choose in ways that enable us to envision a good life and act to realize it for our self and others. This value of personhood or agency, however, rests on certain conceptual and empirical facts. One of these facts is the condition of autonomy, which Griffin interprets as the ability to choose a course in life free from domination or control by other forces. But the value of agency also demands that we be provided with some level of education, information, and the chance to learn what others think, so that our choices will be appropriate to our deepest interests. We must have the ability to follow the course of action we have chosen, which includes an adequate supply of material resources and capabilities, as well as protection from having our efforts blocked by others. The goal of rights on Griffin's account is to enable people by protecting and providing these conditions of agency. The basic right in Griffin's system is thus a right to *pursue* happiness, which is more important than maximizing happiness itself.

These three conceptions of moral rights are non-utilitarian, because they appeal to values such as

equality, autonomy, or human agency and claim that these values are more basic than happiness or well-being. But these theorists share with utilitarians the view that the value of rights is instrumental. Rights are part of the structure of institutions that aim to promote other basic values, and they can be overridden for the sake of promoting the values that justify rights in the first place. For example, if rights are justified because they promote equality of respect, then we would be justified in violating equality in some instances in order to produce greater equality overall. On any of these accounts, we presumably should still violate the rights of one innocent person if this is necessary to protect the rights of five other innocent persons.

While these implications might seem acceptable to many people who view rights as morally important but not absolute, they are unacceptable to theorists who believe that rights have intrinsic value. The belief that the value of rights is intrinsic lies at the core of the second approach to justifying rights. According to this second approach, all instrumental accounts of the value of rights fail to explain the way in which rights determine how a person may or may not be treated. Rights aim to secure for each person a realm of independence, not only from the will of the majority but more generally from pursuit of the best overall result.

In order to have this kind of intrinsic value, rights do not have to be regarded as absolute (Nagel 1996). If there is a right not to be tortured, for example, it may nevertheless be the case that some evils are great enough that they would justify torturing one innocent person to prevent them. The threshold for such justifications would have to be set high, however. Torturing one person may be permitted to save a thousand lives, but it would not be permitted to save one life or to prevent two other innocent people from being tortured.

This intrinsic justification of rights expresses requirements for how we may treat other people in a way that is relatively independent of the value of what happens to them. It tries to explain what we are *prohibited from doing* to others in respect of their rights, which it regards as different from what we must try to *prevent from happening* to them. Thomas Nagel, who defends this view of rights, explains that, 'It is this qualified independence of the best overall results… that gives rights their distinctive character' (Nagel 1996, p. 89).

Robert Nozick (1974) assumes the truth of an intrinsic account in his influential defense of natural rights. He insists that rights should be interpreted as side constraints, which mark the bounds of value-based deliberation and set limits to what we are allowed to do in pursuit of all other goals (see also Thomson 1990). But Nozick simply accepts the traditional view of natural rights and does not explain why they should have this character. Without appealing to something such as divine sanction for natural rights, however, as did many of the traditional natural rights theorists, it seems *ad hoc* to posit rights as first principles of moral deliberation. It begs the question about why it would be wrong to torture one innocent person in order to prevent the torture of five others.

A different intrinsic account, which has been defended by Alan Gewirth (1981, 1982), tries to connect the justification of rights to the capacity for moral valuing, which human beings typically and uniquely possess. Rights express the conditions that make it possible for people to recognize and act on values, so respect for rights is necessary in order for things such as happiness or equality even to exist as moral goods. To violate a person's rights for the sake of some other value, therefore, undermines these conditions and thus makes the actions impossible to justify. While this kind of transcendental argument has drawn much interest among philosophers (see Hart 1955), many people do not find it convincing. If you kill one person in order to prevent five other people from being killed, you cannot then justify what you have done to the victim. But as long as the conditions necessary for him to recognize and act on moral values exist, it is not clear why a person could not comprehend why he might have to be sacrificed for the sake of others.

Another and currently popular intrinsic account of the value of rights appeals to our status as human beings. But rather than seeing human status as an objective moral value, this argument insists that status generates different kinds of moral requirements. Human status is a feature of each person that allows them to claim rights and demand to be recognized and treated in certain ways. Frances Kamm (1992) interprets this status as a kind of inviolability, which she identifies with the possession of rights. According to this argument, if one innocent person may be tortured in order to prevent the torture of five other innocent people, then not only have the rights of that one person been violated, but this acknowledgment of what *may* be done to someone undermines the inviolability of all people. Everyone's moral status is thus compromised by such an act (see Quinn 1993, Nagel 1996).

This argument for the intrinsic value of rights resonates with Feinberg's (1980) interpretation of rights as valid moral claims. He attempts to illustrate the significance of rights by imagining a society that is in every other way as good as possible but does not recognize moral rights. People in this imaginary society are virtuous, caring, and may even be motivated by a sense of duty, but without rights they can only receive beneficent treatment as a gift. If they fail to get what they deserve, they have no ground for complaint. What people lack in a society without rights is the capacity to claim their status as equals and to demand that these claims be recognized. Indeed, Feinberg suggests that the ability to claim rights may be necessary to having the minimal self-respect that

makes a person worthy of the love and esteem of others. A society without rights may be incapable of supporting the kind of love, compassion, and sympathy that we most esteem.

The moral importance of inviolability is thus different from the objective value of not being violated, and it is expressed in terms of moral requirements, the recognition of which each person can demand from all others. If this kind of argument could be made convincing, then it would have several important implications. First, it would explain why it is wrong to violate one person's rights, even to prevent similar violations of five other people. Second, it would explain another feature of rights that internal accounts regard as important, which is that rights imply that it is morally more important not to treat people in certain ways than it is to prevent identical harms from occurring to people. We value the reduction of harms, but rights express our idea of human dignity because they impose moral requirements that take precedence over what we value.

The justification of rights remains a matter of controversy. The difficulty is to defend a theory of rights that captures most of our deepest intuitions about them without making them seem mysterious or paradoxical in the ways we have been discussing. Most philosophers would acknowledge that much work remains to be done in defending an adequate theory of rights, just as most people would acknowledge that much more important work remains to be done to secure the protection of basic human rights for most of the world's population.

See also: Bill of Rights; Children, Rights of: Cultural Concerns; Civil Rights; Cultural Rights and Culture Defense: Cultural Concerns; Human Rights, Anthropology of; Human Rights, History of; Human Rights in Intercultural Discourse: Cultural Concerns; Human Rights: Political Aspects

Bibliography

Bentham J 1834 Anarchical fallacies. In: Bowring J (eds.) *The Collected Works of Jeremy Bentham*. Longman, London, Vol. 2

Dagger R 1989 Rights. In: Ball, T, Farr J, Hanson R (eds.) *Political Innovation and Conceptual Change*. Cambridge: Cambridge University Press, Cambridge, UK

Dworkin R 1977 *Taking Rights Seriously*. Harvard University Press, Cambridge, MA

Feinberg J 1980 The nature and value of rights. In: Feinberg J (ed.) *Rights, Justice and the Bounds of Liberty: Essays in Social Philosophy*. Princeton University Press, Princeton, NJ

Gewirth A 1981 Are there any absolute rights? *Philosophical Quarterly* **31**: 1–16

Gewirth A 1982 *Human Rights: Essays on Justification and Applications*. University of Chicago Press, Chicago

Griffin J 1999 First steps in an account of human rights. Unpublished paper presented to the Conference on Moral Theory and Moral Practice. Le Lavandou, France

Hart H L A 1955 Are there any natural rights? *The Philosophical Review* **64**: 175–91

Hohfeld W 1923 *Fundamental Legal Conceptions as Applied in Judicial Reasoning*. Yale University Press, New Haven, CT

Kamm F 1992 Non-consequentialism, the person as an end-in-itself, and the significance of status. *Philosophy and Public Affairs* **21**: 381–89

Nagel T 1996 Personal rights and public space. *Philosophy and Public Affairs* **24**: 83–107

Nozick R 1974 *Anarchy, State, and Utopia*. Basic Books, New York

Quinn W 1993 Actions, intentions and consequences: the doctrine of doing and allowing. In: Quinn W (ed.) *Morality and Action*. Cambridge University Press, Cambridge, UK

Scanlon T M 1977 Rights, goals, and fairness. *Erkenntnis* **2**: 81–94

Thomson J J 1990 *The Realm of Rights*. Harvard University Press, Cambridge, MA

Tuck R 1979 *Natural Rights Theories: Their Origin and Development*. Cambridge University Press, Cambridge, UK

Vlastos G 1962 Justice and equality. In: Brandt R (ed.) *Social Justice*. Prentice-Hall, Englewood Cliffs, NJ, pp. 31–72

Waldron J 1984 Introduction. In: Waldron J (ed.) *Theories of Rights*. Oxford University Press, New York

D. MacLean

Rights: Legal Aspects

1. Definition and Background

1.1 The Basic Logic of Rights

The language of rights generally connotes claims of entitlement to certain goods. To assert that 'I have a right …' is to articulate a normative claim that certain actions should or should not be performed by other individuals, agencies of the state, or intermediate social institutions. As such, rights are related to, but different from, preferences, interests, and needs. While rights are often invoked in the service of preferences and interests, only those rights which confer legal and/or moral authority that obligate others are enforceable by official power. Hence, whereas the assertion of a preference connotes merely that 'I want …,' to assert a right is to declare that 'I am entitled to … .' Rights thus assume priority in the normative schemes of social ordering; they take priority—or 'trump,' in Ronald Dworkin's memorable phrase—over preferences and interests. Rights similarly take precedence over needs in much the same way—but with one major difference: needs, unlike preferences, convey an objective status that can be validated by persons other than the claimants themselves. Claims to rights, also, are subject to third-party validation, but are, unlike the identification of needs, typically initiated by individual agents acting on their own behalf (Waldron 1996).

Because rights constitute relationships between the

bearers of rights and other persons or institutions, each person's rights impose correlative duties on others. These duties may be negative in that they mandate noninterference or inaction by others, as in one's claim of a right to liberty, privacy, or autonomy. Conversely, the duties imposed by rights may be positive in that others are called upon for action and assistance—as, for example, to ensure the fulfillment of contractual terms or to provide accessible facilities for voting in political elections. Rights may be primarily negative or primarily positive, but they may also involve elements of both.

Rights not only correlate with duties, they also extend by implication to all other similarly constituted relationships in society and are, thus, inclusive. This inclusiveness is the source of the frequent claim that rights are universalistic—i.e., that they can be claimed by, and impose duties on, all citizens within particular sovereign domains. Similarly, the exercise of rights typically confers responsibilities on the rights-bearer. For example, a person may have a right to drink alcohol and a right to drive a care, but these rights imply responsibilities for not engaging in both activities at once. Finally, rights are intrinsically linked to remedies. Thus, for example, not only does the right to equal treatment oblige others not to discriminate, but should discrimination occur, the right to equal treatment requires a remedy to compensate for or to correct the denial of a basic entitlement.

1.2 The Origin and Evolution of Rights

The most widely recognized origin of rights is in European social contract theory dating back to the seventeenth century, and especially to *John Locke*'s classic *Two Treatises on Government*. From this work, rights emerged as a moral imperative that is found at the very core of several of the most important political documents in the modern age—including the United States Declaration of Independence as well as the Bill of Rights, and the French Declaration of the Rights of Man and Citizen. While the concept of rights thus dates back several hundred years, and has provided a founding principle for several longstanding liberal republican orders, it is in the second half of the twentieth century that the moral ethos of rights took on widespread significance around the world. The persuasive power of rights developed, in part, in reaction to horrors of authoritarian states—Hitler's genocide, Stalin's purges, colonial repression and the like—and, in part, in recognition of deprivations suffered by ethnic and racial minorities, women, gays and lesbians, and other marginalized groups in nations throughout the world.

1.3 The Expanding Variety of Rights Claims

The evolution and diffusion of rights through the centuries and around the globe has been boiled down by analysts to three separate stages, or generations, of rights claims.

1.3.1 First generation rights. These are closely linked to the rule of law and to liberal social contract theory; they are regularly found in the constitutions of liberal democratic regimes throughout the world, as well in such transnational legal orders as the European Convention on Human Rights. First generation, or classic rights, are embedded in the logic of both property ownership and in the law itself. On one hand, such rights tend toward the more negative pole, demarcate limits on interference by the state or other social actors, and turn on conceptions of liberty, privacy, due process, and the like. On the other hand, classic, first generation rights also include more positive rights to political participation—to free speech and association, to vote, to run for office, etc.—and to fair treatment by others according to the principles of due process and equal treatment. These rights claims are the most widely recognized and the least contested in general, although specific constructions vary greatly.

1.3.2 Second generation rights. These can be traced to the social democratic traditions of Europe and have during the last century gained currency elsewhere—although not, most notably, in the United States. Second generation rights have to do primarily with entitlements to economic and social wellbeing of the sort formulated in Articles 22–27 of the Universal Declaration of Human Rights (1948)—including, for example, basic education, health care, minimum income, secure employment, workplace organization, food, housing, social security, and the like.

1.3.3 Third generation rights. These are newer in conception, different in kind, and particularly controversial. Advocates of third generation rights seek to confer entitlements to collective or communal goods such as peace, security, a healthy environment, safe natural resources, animal welfare, and group self-determination or sovereignty, and thus seek to its rights to animal. The problematic posed by third generation rights will be addressed below.

2. Contemporary Philosophical Debates About Rights

The original conceptions of so-called liberal Western thinkers in the social contract tradition grounded rights in the theocratic logic of God's divine plan and natural law. Later versions downplayed the divine hand but continued the emphasis on rights as self-evident and inalienable features sewn into the rational

order of the natural universe or human nature. By the twentieth century, the philosophical foundations for rights became thoroughly secular—whether justified by various forms of moral rationality or appeals to proven historical efficacy. It is within the tensions among these legitimating logics that many of the most important philosophical controversies about rights have been waged. We identify just a few of such persistent debates.

2.1 '*Nonsense Built on Stilts*'

This critique, that can be traced back at least to Bentham and continues today among critical 'neo-realist' philosophers, turns on the recognition of rights indeterminacy. The core contention is that if rights are indeterminate and contingent, if they always are changing, and vary with context, then they really mean nothing at all and convey no real social power. Rights, it is argued, are simply moral fictions that possess none of the objectivity, coherence, independence, and universality that philosophers, state officials, and popular culture assign to them. At best, rights are epiphenomenal reflections and rationalizations for relations grounded in other types of causal forces. Respondents to this position acknowledge that rights are social constructions, but the very fact that people believe that rights matter and act accordingly in interpersonal relations, often in the shadow of official enforcement, demonstrates the power of rights. It is further argued that all moral discourses are indeterminate and that it is, therefore, inappropriate to dismiss rights discourses as meaningless because they are unstable and contestable. Indeed, according to this argument, the indeterminacy of rights discourse, as with moral discourse more generally, is desirable in providing a flexible, dynamic, and pragmatic tradition for ordering social relations in just ways.

2.2 *Who Has Rights?*

The debate over the qualifications for rights-bearing is as old as the concept itself and has expanded among moral philosophers as specific rights claims themselves have proliferated. From the start, rights typically have been understood to be properties belonging to all individual human agents capable of claiming them. But it has been precisely these qualifying properties of rights bearers as human agents that have been deployed to limit the reach, and some might say to betray, the promise of rights. Thus, paradoxically, the claim that rights apply only to rational, disciplined, educated, human agents capable of self-government has been invoked to deny rights to children, women, ethnic and racial minorities, religious adherents, immigrants, and the like in many societies that generally recognize individual rights. Similarly, the principle that rights inherently presume individual human claimants has been central to contemporary arguments that third generation entitlements of animals and other nonhuman forms of life (e.g., trees), as well as nondivisible communal goods such as peace and a safe environment, are not suitable for rights claims. The philosophical skeptics argue that such claims lack theoretical coherence and might even undermine the social power of rights discourse itself. Such debates have generated extensive reconsideration about the very foundation of the meaning of rights as a social practice.

2.3 *Do Rights Encourage Atomism, Egoism, and Selfishness?*

The contention that rights practices encourage excessive individualism, egoism, and selfishness among citizens was perhaps most forcefully articulated by Karl Marx and is commonly asserted by moralists across the political spectrum today. Critics emphasize in particular the tendency of rights to celebrate the individual and thus to undermine citizens' sense of responsibility for their actions, duties to others, and a shared commitment to 'community.'

Theorists more favorably disposed to rights acknowledge that rights discourses presume, encourage, and protect individuals. Indeed, securing and advancing the exercise of individual citizen capacities for choice, action, self-development, and diversity against the harmful designs of other individuals, social groups, popular majorities, and the state has been the *raison d'être* of rights as a moral concept from the start. At the same time, however, most theorists understand the rights of individuals in relational rather than insular terms. Rights practices sustain particular obligations that are reciprocal among citizens as well as a general tolerance and even respect for the human dignity of others. In Kant's (1948, pp. 95–6) famous words, 'Act in such a way that you always treat humanity, whether in your own person or in the person of another, always as an end and never simply as a means.' Defenders thus argue that, far from undermining a sense of duty, rights support and enforce dutiful interaction among citizens. According to this way of thinking, rights connote a particular vision of community where individual agency is respected but balanced with conceptions of fairness, justice, and equal treatment for all. It follows that it is precisely the disregard for basic rights that encourages or permits unjust or exclusionary acts that are corrosive of community.

There is also a middle position in this debate. To begin with, it is widely recognized that some rights are more self-regarding than others. Marx's powerful indictment, after all, was that it is property rights that encourage selfishness, materialism, and competition for dominance through capital ownership. As a consequence, he argued, it was the positive communal promise of political rights which was overwhelmed, thus reducing the promise of political solidarity to an

illusion. The underlying point here is that critiques of rights that fail to distinguish among different types of rights—within the first generation as well as among the different generations of rights concepts—are unconvincing to many moral theorists.

3. Jurisprudence and Legal Theory

Contemporary discourse on rights within jurisprudence and legal theory is driven by an effort to identify a distinctively legal conception of rights. This discourse is organized around two foundational debates—pitting rights against democracy on the one hand, and natural law against legal positivism on the other. Historically, there have been profound and heated conflicts among the parties to these debates, but more recently a rapprochement has developed around the recognition that rights and law are mutually constitutive—that rights without the law are virtually unthinkable, as is law without rights. This convergence not only identifies the core common ground of legal theory; it also reveals the distance that separates the legal theory of rights from the understanding of rights that is found in moral and political philosophy. Whereas philosophical understandings of rights reach well beyond the law in search of universal moral meaning, rights in legal theory are much more constrained—principally by way of the intrinsic connection in legal thought between rights and the institutional processes in which rights under law are embedded. In particular, legal theories of rights seek to specify and to justify the responsibility of courts and judges for resolving disputes over rights, and for policing the boundaries of legitimate rights claiming.

3.1 Legal Rights vs. Democracy

The problem of the legitimation of rights claims is especially troublesome to legal officials in democracies and republics, where majority will is understood to carry considerable authority. We noted above that rights, by their very nature, are intended to protect and advance individual entitlements that restrict the actions of others, including those of popular majorities. As such, rights—both in general, and in specific instances—are often condemned as counter-democratic forces. Utilitarian theorists raise a parallel objection that begins with the assertion that rational public policy should provide the greatest good for the greatest number—thus giving considerable weight to the preferences of the majority. Rights are thus often challenged as costly obstacles to the development of democratic and efficient (see below) social policy.

Other legal theorists respond that rights do not thwart so much as give structure to utilitarian policy calculation, by assigning higher values to certain basic entitlements—especially those concerning protection for individual freedom and dignity—than to other competing considerations. Consequently, securing rights may impose costs on particular social goals, but these costs are arguably incurred in order to sustain larger, enduring social commitments. Likewise, it is argued that rights claims do not absolutely throttle democratic majorities or elected officials; rather, rights claims tend only to place specific, limited restrictions on particular policy actions that defy core social values and protection for individuals. Moreover, it is important to underline that many types of rights advance—indeed, are actually essential to—the democratic process. Such entitlements include, at the least, protections for dissenting speech, political association, free elections, due process, and basic public education.

There is no simple way out of this debate. While it may well be true that rights in general provide sustenance to democratic polities, particular rights constructions certainly do, and indeed are intended to, defy the will of elected officials and popular majorities. Hence the enduring challenge for judicial authorities is to justify by legal argument their specific rights constructions.

3.2 Rights and Justice

Historically, a divisive debate in legal theory between positivism and natural law has pitted a theory of law that recognizes little in common between morality and law against a conception of law as being inseparable from morality. The latter is virtually indistinguishable from the moral philosophy of rights discussed earlier, and represented by the work of Locke and Kant among others. The former is rooted in analytic philosophy and in a logical positivism that eschewed the search for universal truths. During the late twentieth century the gap between legal positivists and natural law theorists narrowed (but was by no means closed) by the emergence of a widespread agreement on a least-common-denominator conception of justice as the 'right to have rights.' According to this way of thinking, the guarantees associated with rights are confined to assuring predictability, reciprocity, and impartiality in the application of the law. What remains unresolved in the debate between theorists of natural law and legal positivism is whether justice, as thus defined, is the differentiating element of legal right, or only a quality that tends to be closely associated with legal rights.

While many jurists have not reached out to the broader realms of political and moral philosophy, Ronald Dworkin, (Dworkin 1977) at the end of the twentieth century the most prominent legal theorist of rights, has done so. At the core of his analysis is the distinction between policy and principle. Whereas policy is driven by utility, principle, the defining element of law, is to be observed, not because it will advance or secure an economic, political, or situation deemed desirable, but because it is a requirement of justice or fairness or some other dimension of morality.

In thus linking law to principle, and principle to justice, Dworkin identifies himself with the relatively noncontroversial common ground between positivism and (neo)natural law. On the other hand, by introducing 'some other dimension of morality' he clearly breaks new ground. For Dworkin, it is constitutions and constitutional rights that 'fuse legal and moral issues' (Dworkin 1977, p. 185). Constitutions thus provide in addition to the right to have rights, particular substantive rights as well.

For Dworkin, then, rights are inextricably linked not only to the foundational elements of positive law but, at least in their political and constitutional incarnation, to the foundational elements of the state. In this sense Dworkin can be seen as a kind of transitional figure among legal theorists, reaching out as he does to moral and political philosophy and incorporating them into his analysis of rights constructions by judges and other state officials. But unlike those of moral and political philosophers, Dworkin's theory of rights is embedded in, and inextricably linked via principle to, what is distinctively legal. Moral philosophy thus becomes an adjunct to the law and is, in effect, put at the disposal of the law, as a source of principle. In this sense, Dworkin's theory of rights returns to the starting point and, indeed, to the cornerstone of jurisprudence and legal theory, that is, it is a search for what is uniquely legal, as opposed to moral, ethical, or political.

4. *Empirical Social Science Studies of Rights Practice*

Whereas philosophic and jurisprudential inquiries into rights tend to be normative and abstract, social science research regarding rights practices is explanatory and empirical. Moreover, in the social sciences rights are treated less as moral entitlements of historical persons than as contingent resources that are mobilized both by individuals in everyday life, and by organized interests in political advocacy. Rights are not viewed as 'above politics' and power so much as constitutive of political activity and power relations. The primary objective of the social sciences thus is to assess the value of rights as resources, and to explain variation in the value of rights from time to time and from place to place. To do so, social scientists shift the locus of rights inquiry from formal systems of logic and idealized institutions to inductively organized research on rights in actual institutional settings, everyday life, and political advocacy.

4.1 *Rights and Everyday Disputing*

There has been a rapidly expanding scholarly literature on the ordinary disputing practices of citizens, especially in the United States but increasingly around the world. Many studies of 'ordinary' (non-elite) individual citizens in the US suggest that rights matter a great deal, and that these individuals are not hesitant to assert them. Indeed, some scholars demonstrate that lay people, more so than judges, clerks, police, and other legal authorities, act on the premise and the expectation that the legal system exists to protect rights. Studies from other polities around the globe have also found some evidence for this generalization. One of the most interesting findings of Ewick and Silbey (1998) is that ordinary, nonelite citizens in the US can, moreover, relate to rights in flexible, creative, and complex ways, entertaining several conflicting notions at once. Interviewees in their study could view rights instrumentally as malleable rules of a game, authoritatively as an objective source of valid law, and skeptically as a burden that must be endured, evaded, or resisted.

Other studies qualify or challenge assumptions about the salience of rights. Research on the disputes of individuals suggests, for example, that Americans and others often refrain from asserting their rights; many studies document a widespread reluctance to seek redress when rights are violated and entitlements denied. This is especially the case among lower-income and minority citizens. Moreover, some studies find evidence of considerable cultural biases and popular pressure against rights claiming, at least through formal adversarial channels, even in allegedly 'rights obsessed' cultures such as the US. Patterns of individual practice are thus complex, contradictory, counterintuitive, and often defy generalization within, much less beyond, the US context.

4.2 *The Politics of Rights*: *Mobilizing Rights in Political of Advocacy*

The value of rights as a political resource is heavily debated within the social science literature. The most stringent critics of rights strategies claim that to deploy rights is at best ineffectual, and at worst counterproductive. Some studies suggest that rights litigation had no more than negligible effect on policy outcomes. According to realist critics only political power built around electoral coalitions and influential interests can bring about changes in well-established policies and practices. Other scholars demonstrate that rights-based practices often do intrude into policy processes, but with the unfortunate consequence of making these processes rigid and inefficient.

There is, however, a substantial body of research that suggests rights strategies can make meaningful contributions to a transformative politics—albeit in complex, contingent, and unpredictable ways. This research does not so much invalidate the findings of critics as shift the focus and the standards by which rights strategies are to be judged. Instead of concentrating exclusively on the direct effect of court-articulated rights, their indirect and constitutive effects are explored and assessed. Moreover, in this research,

the criterion for judging rights strategies shifts from policy efficiency to social equity—with the salience of this latter criterion most compellingly articulated in the work of critical race and critical gender theorists.

Especially in the United States, where much of the available research has been conducted, the indirect and transformative potential of a 'politics of rights' (Scheingold 1974) has been persuasively demonstrated. Just as rights have a constitutive impact on individual disputants, so too have the connotations of entitlement associated with rights had a mobilizing impact on social movements for pay equity, disability rights, and animal rights. Insofar as a need or a grievance can be articulated as a legal or constitutional right, expectations can be altered, quiescence overcome, and a politics of rights generated. Note that this research focuses very much on the discursive power of rights to constitute social and political life. Moreover, despite their liberal connotations and associations, rights can also be deployed on behalf of substantive agendas that span the political spectrum.

Despite all of their persuasive power, however, rights alone secure very little. Nor is rights mobilization a matter of spontaneous combustion. The constitutive power of rights depends on deploying rights shrewdly, in conjunction with other political resources—including shrewd leadership working in combination with established organizations like churches (civil rights), labor unions (pay equity), and other kinds of social-action organizations. Even then a transformative politics does not necessarily follow. Political mobilization may, for example, generate countermobilization (the abortion rights vs. right-to-life movements). But as the struggle over abortion clearly demonstrates, rights do matter, and a well executed strategy of rights can at the very least destabilize the political arena—thus creating opportunities that were previously foreclosed. In sum, the increasing body of empirical research on rights advocacy in a variety of settings underscores the variable, contingent, and often-contradictory implications of a politics of rights.

See also: Bill of Rights; Civil Liberties and Human Rights; Civil Rights; Cultural Rights and Culture Defense: Cultural Concerns; Fundamental Rights and Constitutional Guarantees; Human Rights, Anthropology of; Human Rights, History of; Human Rights: Political Aspects; Rights

Bibliography

Bumiller K 1988 *The Civil Rights Society: The Social Construction of Victims*. The Johns Hopkins University Press, Baltimore, MD

Crenshaw K W 1988 Race, reform and entrenchment: Transformation and legitimation in antidiscrimination law. *Harvard Law Review* **101**: 1331–87

Dworkin R 1977 *Taking Rights Seriously*. Harvard University Press, Cambridge, MA

Ewick P, Silbey S 1998 *The Common Place of Law: Stories from Everyday Life*. University of Chicago Press, Chicago

Fuller L 1964 *The Morality of Law*. Yale University Press, New Haven, CT

Glendon M A 1987 *Rights Talk: The Impoverishment of Political Discourse*. Free Press, New York

Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK

Kant I 1948 *The Moral Law* (trans. Paton H J). Hutchinson, London

McCann M W 1994 *Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization*. University of Chicago Press, Chicago

Melden A I 1977 *Rights and Persons*. University of California Press, Berkeley, CA

Merry S E 1990 *Getting Justice, Getting Even: Legal Consciousness Among Working Class Americans*. University of Chicago Press, Chicago

Minow M 1990 *Making All the Difference: Inclusion, Exclusion, and American Law*. Cornell University Press, Ithaca, NY

Scheingold S 1974 *The Politics of Rights: Lawyers, Public Policy, and Political Change*. Yale University Press, New Haven, CT

Waldron J 1996 Rights and needs: The myth of disjunction. In: Sarat A, Hearns T (eds.) *Legal Rights: Historical and Philosophical Perspectives*. University of Michigan Press, Ann Arbor, MI

M. McCann and S. Scheingold

Right-wing Movements in the United States: Women and Gender

Right-wing movements have existed since colonial times in the USA, yet the majority of studies have focused on either what is termed the 'Old Right' of the 1950s or the 'New Right' formed during the mid-1970s and 1980s. An explosion of interest in understanding right-wing movements occurred in the aftermath of WWII when attention turned to authoritarianism, McCarthyism, and the Old Right of the 1950s. During the 1960s and 1970s, particularly with the prominence of protest movements on the left, research on right-wing movements waned, but a resurgence of interest has focused on the New Right formed during the 1970s and 1980s. While little is known about women's participation in the Old Right, recent studies have focused on understanding women's role in the New Right. This article considers the affect of women and issues of gender in both the Old and New Right.

1. Women and the Old Right

Historically, right-wing movements in the US have been based on a crucial marriage of interests between two groups. First, members of less-educated, lower economic strata, who are highly religious and drawn to the non-economic issues of right-wing movements.

This sector is intolerant of ethnic and religious minorities, with particular antagonism expressed towards Blacks and towards immigrants, previously expressed against Catholics and Jews and more recently directed against Mexicans and Asians. The second sector includes those rooted in a more highly educated, higher-income strata, less religious, tolerant of minorities, and committed above all to economic conservatism. The Old Right of the 1950s combined these two sectors, one consisting of more affluent economic conservatives, particularly reacting to the expansion of government implemented by President Franklin D. Roosevelt during the New Deal of the 1930s, and a more devoutly religious and less educated, lower middle class following, fearful of communism due to strong evangelical and fundamentalist beliefs. Anticommunism held together the Old Right, spearheaded by the campaigns of Senator Joseph McCarthy.

In terms of women's role within the Old Right, the vast majority of research makes no mention at all of female activists. It is not clear whether the lack of attention to women in the Old Right is due to the small number of female participants in the movement or to the male bias of the researchers. Because of this lack of data, there is no way of knowing how women of the Old and New Right compare in terms of number of participants, degree of involvement, or social background.

2. *Women and the Conservative Intellectual Movement*

Cleavage within the popular movements of the Right in the US is also evidenced in the conservative intellectual movement, which is comprised of three main groups. First, classical liberals or libertarians promote the free market, private property, limited government, self-reliance, and *laissez-faire*, and resist the threat of the ever-expanding state to liberty and free enterprise. Second, are traditionalists who are dismayed by the erosion of values and the emergence of a secular society and who call for a return to traditional religious and ethical absolutes. Third are militant anticommunists, ex-radicals disillusioned by the left and who became alarmed by the despotism of the Soviet Union. With the emergence of the New Right, this third strand also includes 'neo-conservatives,' primarily academics and writers who rejected the politics of the Democratic Party during the late 1960s. Converts from the left, neo-conservatives are disillusioned liberals drawn heavily from Jewish, intellectual circles of the East Coast. As with mass right-wing movements, anticommunism has acted as the glue holding together the conservative intellectual movement.

Although nearly all of the influential conservative intellectuals are male, libertarian author Ayn Rand has had a profound impact on generations of right-wing adherents. Through novels such as *The Fountainhead* and *Atlas Shrugged*, Rand introduced the philosophy of objectivism and celebrated unbridled individualism, rationalism, and *laissez-faire* capitalism. Rand inspired thousands of people, especially youth during the 1960s, to join the libertarian cause.

3. *Women and the New Right*

During the mid to late 1970s varied groups coalesced into the New Right. Much of the force behind this coalition was a counter-movement to the social movements of the 1960s, an organized protest against affirmative action, the welfare state, feminism, and gay and lesbian rights. Like the Old Right, the New Right also incorporated two distinct constituencies: social conservatism and economic or *laissez-faire* conservatives. Social conservatives are devoutly religious and are primarily concerned about moral issues and protecting the traditional family. The family represents the building block of society; it instills children with moral values and restrains the pursuit of self-interest. Implicit in this image of the family is the social conservative conception of human nature. Humans are creatures of unlimited appetites and instincts. Left on their own, the world would be a chaos of seething passions, overrun by narrow self-interest. Only the moral authority of the family and religion tames human passions, transforming self-interest into the larger good. The ideal society is one in which individuals are integrated into a moral community, bound together by religious faith and by common values. *Laissez-faire* conservatives or libertarians in contrast, aim above all to protect the economic and political liberty of the individual. Rooted in the classical liberalism associated with Adam Smith, John Locke, and John Stuart Mill, *laissez-faire* conservatives uphold the economic liberty of the free market and the political liberty based on the minimal state. The individual is viewed as an autonomous, rational, self-interested actor. Rather than viewing humans as creatures of unlimited passions, *laissez-faire* conservatives view humans as beings endowed with free will, initiative, and self-reliance. The larger good is not ensured through the maintenance of moral authority and the restriction of self-interest. Rather, the *laissez-faire* ideal poses a society in which natural harmony exists through the very pursuit of self-interest. The aim of the good society is to elevate the potential of humans by bringing their creative and productive nature to fruition.

Despite these fundamental differences, like the Old Right, these two constituencies of the New Right united in their virulent opposition to communism. The New Right is distinct from the Old Right, however, in the visibility of its female participants and leaders.

Increasingly women have taken a more active and visible role in right-wing politics. On a leadership level, such well-known figures as Phyllis Schlafly, who led the fight to defeat the Equal Rights Amendment (ERA), typifies the social conservative movement while Jeane Kirkpatrick, a professor of government, former United Nations representative, and newspaper columnist, embodies *laissez-faire* beliefs. Yet women of the New Right, like their male counterparts, are not a homogenous entity. They are drawn from different sociological pools. While the majority of activists are white, social conservatives tend to be somewhat less educated than *laissez-faire* women and are devoutly religious, primarily Catholic or Protestant fundamentalist, married with children, with many being full-time homemakers. *Laissez-faire* women tend to be younger, more secular in orientation, more likely to be single or divorced, and have smaller numbers of children, more postgraduate degrees, and typically have professional careers. In fact, in terms of social background *laissez-faire* women resemble feminist women.

3.1 Views of Gender and Feminism

Women of the New Right have contrasting beliefs regarding gender. Social conservative women believe in a strict division of gender roles as decreed by the Scriptures. Gender is envisioned as a hierarchical ordering with God and Christ at the top, followed by men, and then women. While male and female roles are each respected as essential and complementary components of God's plan, men are the spiritual leaders and decision-makers in the family. Women's role is essentially defined in terms of altruism and self-sacrifice: to support men in their positions of higher authority, and to care for the family. Because gender roles are delineated in such unambiguous terms, any blurring of these roles is viewed as a threat.

While social conservatives envision a man's world as distinct from a woman's world, no such distinctions are visible in the *laissez-faire* view of gender. There is a notable absence of commentary on gender, and particularly no mention of the need for male authority and female submission or of women's natural orientation to others. The social conservative vision of a hierarchical ordering of authority of men over women is antithetical to the *laissez-faire* conservative ethos. Faith in individual self-reliance and free will, and belief in the liberty and autonomy of every individual, extends to women as well as to men.

Clearly, there is a relationship between a woman's social location and her values and beliefs. It is not coincidental that a woman with a devout religious upbringing who is a full-time homemaker is likely to see the family at the center of the world and to adhere to a traditional ideology regarding gender. Nor is it a coincidence that a single professional women who is devoted to her career does not see the family as central and holds nontraditional attitudes about men's and women's roles. Traditional gender roles 'work' for social conservative women not simply because of ideology, but because they themselves are more likely to be in life situations which correspond to traditional roles. *Laissez-faire* women, on the other hand, are in life situations and have social resources outside of the traditional female sphere.

These varying perceptions of gender result in opposing views of gender discrimination. Because social conservatives adhere to a hierarchical ordering, they believe that positional differences between women and men do not imply inequality. Any positional differences between the sexes are attributed to 'natural' differences or to individual choice, not to gender discrimination. *Laissez-faire* women, on the other hand, are inclined to recognize and deplore the unequal treatment of women. Discrimination is denounced as interfering with the individual's ability to climb to the height of her talents. Both men and women must remain unconstrained by external interference, free to follow the dictates of their own initiative. Although *laissez-faire* women share recognition of discrimination with feminists, they react to such bias purely in voluntaristic terms. Women who are discriminated against at work should on their own initiative leave the job and enter the free market to choose another one.

These differences in views of gender extend to opposing views of feminism. While social conservatives view feminism as one of the primary forces of moral decay responsible for America's decline, *laissez-faire* women actually adhere to part of the feminist vision. Social conservative activism is a direct countermovement to the feminist movement of the 1960s and 1970s, as well as to the changes brought on by the civil rights, student, and gay and lesbian movements. Social conservatives blame feminists for attacking the status of the homemaker and for promoting issues such as the ERA, abortion, gay rights, and childcare, all of which are viewed as eroding traditional notions of gender and the family. Calling themselves the 'pro-family' movement, opposition to these very issues tops the political agenda of social conservative activists. *Laissez-faire* women, on the other hand, support gay rights, a woman's right to abortion, and daycare, seeing them as private, individual matters that should not be legislated by the state. *Laissez-faire* women depart from many feminists, however, in the means to achieve women's equality. While feminists call for federal support in addressing women's needs, *laissez-faire* women oppose any reliance on big government, including government funding of daycare or abortion. Further, the top priority of *laissez-faire* women is not addressing women's inequality; rather, issues related to the economy and defense remain the objectives of these women.

3.2 The Paradox of Right-Wing Women's Activism

Given the social conservative woman's adherence to traditional gender roles in which men are breadwinners and protectors, and women are helpmates and caretakers, some analysts point to a seeming contradiction between these beliefs and her very role as a political activist. However, social conservative women do not see a tension between their political involvement and the traditional female role. Female activism is defined within the bounds of traditional gender ideology as women altruistically working for the benefit of a larger cause, acting as moral gatekeepers to bring purity to a world filled with sin. Nor is there any apparent paradox to *laissez-faire* women's activism. Because they do not see women as bound to traditional roles, there is no seeming contradiction between their beliefs and their own role as public leaders. They do not conceive of themselves as the caretakers of society, altruistically at work for the benefit of all. Instead, they see themselves as no different from men, self-interested actors working for a political cause.

Despite the prevalent assumption that anti-feminist women are 'brainwashed' or 'lackeys' of men, the actual paradox of right-wing women is that those women who are furthest from feminists in their beliefs actually do act in their own interests as women, while *laissez-faire* women, who actually share a portion of the feminist vision, do not act in their collective gender interest. Far from suffering from false consciousness, social conservative women are well aware of their interests and act to defend their status as women. To borrow a phrase from Marx, social conservative women act as women for themselves, while *laissez-faire* conservative women remain women in themselves. Gender identity is central to the political involvement of social conservative women; recognizing their commonality with other traditional women, they seek to protect women's place as a group. *Laissez-faire* women are not motivated out of concern regarding gender; they do not act in the collective interest of women and therefore remain women in themselves. They act as members of the marketplace, not as members of their gender group, in organizing to return America to strength and freedom.

See also: Civil Rights Movement, The; Ethnic and Racial Social Movements; Feminist Movements; Fundamentalism (New Christian Right); Gay/Lesbian Movements; Historiography and Historical Thought: Southeast Asia; Identity Movements; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Labor Movements, History of; Nation-states, Nationalism, and Gender; Peace Movements, History of; Social Movements, History of: General; Social Protests, History of; Youth Movements, History of

Bibliography

Conover P J, Gray V 1983 *Feminism and the New Right: Conflict over the American Family*. Praeger, New York

Himmelstein J L 1990 *To the Right: The Transformation of American Conservatism*. University of California Press, Berkeley, CA

Klatch R E 1987 *Women of the New Right*. Temple University Press, Philadelphia, PA

Lipset S M, Raab E 1970 *The Politics of Unreason: Right Wing Extremism in America, 1790–1970*. Harper and Row, New York

Luker K 1984 *Abortion and the Politics of Motherhood*. University of California Press, Berkeley, CA

Mueller C, Dimieri T 1982 The structure of belief systems among contending ERA activists. *Social Forces* **60**: 657–75

Nash G H 1976 *The Conservative Intellectual Movement in America Since 1945*. Basic Books, New York

Staggenborg S 1998 *Gender, Family, and Social Movements*. Pine Forge Press, Thousand Oaks, CA

R. E. Klatch

Risk: Empirical Studies on Decision and Choice

Even though most people have an intuitive understanding of what is meant by *risk*, risk is a concept that defies precise definition. Ordinary citizens as well as scientific experts differentiate between choice options or courses of actions in terms of their riskiness, and such explicit or implicit judgments of perceived riskiness often affect their decisions. Some uncertainty about the outcome of a decision or action seems to be necessary to make it appear risky. This article describes three paradigms that define risk in different ways and that identify different types of variables that have been shown to affect people's perception of risk.

1. Risk Perception

The democratization of decision making and increased availability of information provide people with more choices than ever before. Economic globalization and faster social and technological change have, at the same time, introduced more uncertainty and unpredictability. Decisions under risk and uncertainty are abundant, and perceptions of risk affect those decisions. People pull their money out of financial ventures when they judge the risks to be too high or start a lawsuit when the risks of inaction outweigh the risks of litigation. The risk perceptions of individuals as citizens and consumers also affect decisions made by government agencies and corporations. Public perception that silicone breast implants put users at

risk of autoimmune disease, for example, resulted in bankruptcy for the manufacturer, despite clear scientific evidence of no silicone-related illnesses. Controversies about the licensing of technologies such as genetic engineering or the siting of facilities such as landfills or nuclear power stations tend to be fueled primarily by disagreements about posed levels of risk, rather than by disagreements about the acceptability of specific risk levels. While business or government experts have clear quantitative evaluations and definitions of the risks that products or technologies pose, based on objective data or models, members of the general public often seem to evaluate the same options very differently. Much of the early research on psychological risk dimensions (described in Sect. 2.2) was funded by the Nuclear Regulatory Commission to understand why and how public perception of the riskiness of nuclear technology could differ so dramatically from the estimates provided by nuclear engineers.

2. *Measures of Perceived Risk*

Risk perception has been studied empirically within three theoretical paradigms. Studies within the axiomatic measurement paradigm have focused on the way in which people combine objective risk information, that is, possible consequences of risky choice options such as mortality rates or financial losses and their likelihood of occurrence. Research within the psychometric paradigm has identified people's emotional reactions to risky situations that affect judgments of the riskiness of physical, environmental, and material risks in ways that go beyond their objective consequences. Studies within the sociocultural paradigm have examined the effect of group- and culture-level variables on risk perception.

Risk as an explicit or implicit variable in theories of risk preference is discussed elsewhere (see *Risk: Theories of Decision and Choice*; *Personality and Risk Taking*; Weber 1997).

2.1 *Axiomatic Measurement Studies*

The axiomatic measurement approach models perceived risk as a function of attributes of risky options that are described as probability distributions over possible outcomes. Much of this work derives from the definition of risk as variance in traditional models of risky choice (see *Risk: Theories of Decision and Choice*). Under this definition sure options carry no risk, since there is no variance around the guaranteed, single, known outcome. The wider the distribution of possible outcomes, the greater an option's riskiness becomes. Yet, studies that assessed people's judgments of the riskiness of financial gambles (see *Risk: Theories of Decision and Choice*) have shown that downside variability of outcomes affects the perception of an option's riskiness much more than upside variability, an asymmetry that is not captured by variance measures.

As a result, other axiomatic measures of perceived risk were developed that differ in the assumptions they impose on judgments of riskiness (see *Risk: Theories of Decision and Choice*). These models capture both *similarities* in people's risk judgments (with a common functional form by which probability and outcome information about risky options are combined) and *individual and group differences* (with the help of model parameters that capture differences in the relative weight given to different model components).

Weber and Bottom (1989, 1990) submitted the behavioral axioms on which these models are based to empirical tests and found support for the assumption that risk judgments are transitive (if option B is judged more risky than option A, and C more risky than B, then C will also be judged more risky than A) and monotonic (if option B is judged more risky than A, then any combination of obtaining B with probability p and X otherwise should also be more risky than the combination of obtaining A with probability p and X otherwise). Some violations of the expectation principle for risk judgments in the gain domain (Keller et al. 1986) were shown to be the result of nonnormative probability accounting, similar to that observed for preferences. People, for example, do not think that a two-stage lottery carries the same risk as a mathematically equivalent single-stage lottery, even if the final outcomes of the two lotteries and their probabilities are identical. Weber and Bottom's (1989) results supported the additive combination of gain and loss components hypothesized by the conjoint expected risk (CER) model (Luce and Weber 1986; see *Risk*: *Theories of Decision and Choice*) and ruled out, at least as descriptive models of perceived risk, other risk functions that assume that losses and gains impact risk judgments multiplicatively. The CER model has been described as the most viable model to describe single-dimensional risk appraisal (Yates and Stone 1992) and is probably the most widely applied axiomatic model of risk, but other, similar models exist (see *Risk: Theories of Decision and Choice*).

Bontempo et al. (1997) fit the CER model to judgments of the riskiness of monetary lotteries made by business students and security analysts in Hong Kong, Taiwan, the Netherlands, and the US. Cross-national differences in risk judgments followed a Chinese–Western division, with respondents from the two culturally Chinese countries having model parameters that were similar to each other but different from those of the two Western countries. The probability of a loss had a larger effect on perceived risk for Western respondents, while the magnitude of losses had a larger effect on the risk perceptions of Chinese respondents. Cross-cultural differences in risk perception were greater than differences due to occu-

pation (students vs. security analysts), suggesting that cultural upbringing and environment seem to play a larger role in shaping the perception of financial risks than professional training or expertise.

2.2 Psychometric Studies

The psychometric measurement approach treats risk perception as a multidimensional construct and uses multidimensional scaling, clustering, and factor analysis to identify its underlying psychological dimensions (Slovic et al. 1986). Research within this paradigm has found that the perceptions of the risks of hazardous technologies or activities by members of the lay public have often little to do with possible outcomes and their probabilities. Compared to technical experts, ordinary citizens overweight risk associated with infrequent, catastrophic, and involuntary events, and underweight the risk associated with frequent, familiar, and voluntary events. The psychological risk dimensions identified by the psychometric paradigm fall into two categories. The first one, *dread*, is defined by a perceived lack of control, feelings of dread, and perceived catastrophic potential. The second one, *risk of the unknown*, is defined as the extent to which the hazard is judged to be unobservable, unknown, new, or delayed in producing harmful impacts. Cross-national comparisons show that risk perceptions in a wide range of countries are affected by the *dread* and the *risk of the unknown* factor. Risk perceptions between countries differ in where respondents place particular hazards (e.g., nuclear power) within this factor space, usually in ways consistent with national differences in exposures and socioeconomic development.

2.3 Sociocultural Studies

In Douglas and Wildavsky's (1982) cultural theory, risk perception is viewed as a collective phenomenon by which members of different cultures selectively attend to different categories of danger. Each culture selects some risks for attention and chooses to ignore others. Cultural differences in risk perceptions are explained in terms of their contribution to maintaining a particular way of life. The theory identifies five distinct cultures (labeled hierarchical, individualist, egalitarian, fatalist, and hermitic, respectively) that differ in their patterns of interpersonal relationships and argues that members of these cultures therefore differ in their perceptions of risk. Hierarchically arranged groups, for example, tend to perceive industrial and technological risks as opportunities, whereas more egalitarian groups tend to perceive them as threats to their social structure. The significance of this approach to understanding risk perception is that it provides a way of accounting for the effect of group- and culture-level variables on the behavior of individuals. It suggests that culture teaches individuals where their interests lie and what variables and events pose risks to those interests and ways of life.

Implicit in this cultural theory of risk perception is the hypothesis that cultural differences in trust in institutions drive differences in perceived risk. Slovic (1997) summarizes empirical evidence for this hypothesis, for example, the fact that minority status (e.g., due to race) is associated with reduced trust in social institutions. The relationship between trust and risk perception seems to be mediated by an emotional pathway, with reduced trust resulting in stronger negative affective responses to potential hazards and increased perceptions of risk.

Weber and Hsee (1999) provide another culture-level explanation of cross-national differences in risk perception, based on differences in collectivism–individualism. Comparison of the perception of financial and other material risks by Chinese vs. American respondents supports their assumption (coined the cushion hypothesis) that social collectivism serves as mutual insurance against catastrophic losses, making the risks faced by members of the collective, in fact, smaller.

3. Relationships between Measures of Risk Perception and Applications

The axiomatic risk measures of Sect. 2.1 were developed to describe the perception of financial risks, whereas the psychometric risk measures of Sect. 2.2 were developed to describe the perception of health and safety risks. Holtgrave and Weber (1993) investigated how well either type of measure (a simplified version of the CER model and the psychometric model) would predict risk judgments for both types of risk. Respondents, who were MBA students, provided holistic risk judgments for a range of financial and health and safety risks. They also evaluated each risk on the variables used as predictors by the two types of models (probability of a loss or gain, and expected loss or gain for the CER model; dread, control, catastrophic potential, voluntariness, novelty, and equity for the psychometric model).

Regression results showed the CER model to be a better predictor of perceived risk than the psychometric model for both financial risks and health/safety risks. The predictor variables of the CER model seem to come closer to the way people evaluate risk in both domains. The psychometric model may need a dimension reflecting the probability of harm to provide a better fit, since this dimension is highly correlated with risk ratings. Another reason might be that people consider both the pros and cons of activities when judging their risks. The CER model assumes that the impact of positive outcomes can offset the impact of negative outcomes. The psychometric model, on the

other hand, focuses exclusively on the impact of the downside of activities.

The best predictor of risk perceptions for both financial and health/safety risks, however, was a hybrid model that added the dread variable of the psychometric model to the probability and outcome-based variables of the CER model. This suggests that the perception of both financial and health/safety risks has an emotional component that is not completely described by the 'objective' components of axiomatic models (see Loewenstein et al. 2000 for a summary of the role of affect in risk perception and risk taking).

Palmer and Sainfort (1993) applied the CER model in the context of genetic counseling to model the judgments made by members of a clinical population of dwarfs about the riskiness of different procreative alternatives available to them. Faulting the genetic counseling literature for equating the perceived riskiness of an adverse event (e.g., the birth of a child with a genetic disorder) with the perceived probability of the event's occurrence, Palmer showed that the perceived risk of different reproductive alternatives also reflects the severity of consequences. Therefore, measures of risk that combine both probability and outcome information (such as the CER model) are a better predictor of risk perception.

The CER model has also been used to test the prediction of Douglas and Wildavsky's (1982) socio-cultural theory that individuals with different world-views differ in predictable ways in their perceptions of risk. Palmer (1996) elicited risk judgments for financial and health/safety risks from respondents who came from either a hierarchical, individualist, or egalitarian subculture in Southern California. Hierarchists who are comfortable with determining acceptable levels of risk for technologies (Thompson et al. 1990), a process that explicitly considers and weighs gains and losses, provided risk judgments that reflected all predictor variables of the CER model (gains as well as losses, outcome levels as well as probabilities). Egalitarians, on the other hand, who are suspicious of technologies and view nature as fragile and in need of protection, suggesting that they should see risk in terms of possible harm, provided risk judgments that reflected only the loss predictor variables of the CER model. Individualists, who view risk as opportunity, given their tendency to see benefits from most activities as long as they do not interfere with market mechanisms (Thompson et al. 1990), were the group that provided the lowest risk judgments for almost all of the risky investments and activities.

4. Conclusions and Future Directions

The empirical studies of risk perception reviewed in this article show that risk perceptions are shaped by three classes of variables: the negative and positive outcomes of choice options and their likelihood; affective reactions such as dread or fear of the unknown; and social and cultural variables that influence the perception and interpretation of the consequences of risky choice options. These variables have mostly been studied in isolation. Future research should examine possible interactions. There is also need for further examination of the accuracy or appropriateness of people's perception of risk.

The fact that ordinary citizens perceive the risk of many activities and technologies in ways that fundamentally diverge from the way they are perceived by scientists moves the choice of a measure of risk onto which to base policy decisions into the political arena. Fischhoff et al. (1984) argue that such choice should not be the exclusive province of scientists, who have no special insight into what society should value.

Bibliography

Bontempo R N, Bottom W P, Weber E U 1997 Cross-cultural differences in risk perception: A model-based approach. *Risk Analysis* **17**: 479–88

Douglas M, Wildavsky A 1982 *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*. University of California Press, Berkeley, CA

Fischhoff B, Watson S R, Hope C 1984 Defining risk. *Policy Sciences* **17**: 123–39

Holtgrave D R, Weber E U 1993 Dimensions of risk perception for financial and health safety risks. *Risk Analysis* **13**: 553–8

Keller L R, Sarin R K, Weber M 1986 Empirical investigation of some properties of the perceived riskiness of gambles. *Organizational Behavior and Human Decision Processes* **38**: 114–30

Loewenstein G F, Weber E U, Hsee C K, Welch E 2001 Risk as feelings. *Psychological Bulletin* **27**: 1–20

Luce R D, Weber E U 1986 An axiomatic theory of conjoint, expected risk. *Journal of Mathematical Psychology* **30**: 188–205

Palmer C G S 1996 Risk perception: An empirical study of the relationship between world view and the risk construct. *Risk Analysis* **16**: 717–23

Palmer C G S, Sainfort F 1993 Towards a new conceptualization and operationalization of risk perception within the genetic counseling domain. *Journal of Genetic Counseling* **2**: 275–94

Slovic P 1997 Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. In: Bazerman M, Messick D, Tenbrunsel A, Wade-Benzoni K (eds.) *Psychological Perspectives to Environmental and Ethical Issues in Management*. Jossey-Bass, San Francisco, pp. 277–313

Slovic P, Fischhoff B, Lichtenstein S 1986 The psychometric study of risk perception. In: Covello V T, Menkes J, Mumpower J (eds.) *Risk Evaluation and Management*. Plenum Press, New York

Thompson M, Ellis R, Wildavsky A 1990 *Cultural Theory*. Westview Press, Boulder, CO

Weber E U 1997 The utility of measuring and modeling perceived risk. In: Marley A A J (ed.) *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*. Erlbaum, Mahwah, NJ

Weber E U, Bottom W P 1989 Axiomatic measures of perceived risk: Some tests and extensions. *Journal of Behavioral Decision Making* **2**: 113–31

Weber E U, Bottom W P 1990 An empirical evaluation of the transitivity, monotonicity, accounting, and conjoint axioms for perceived risk. *Organizational Behavior and Human Decision Processes* **45**: 253–75

Weber E U, Hsee C K 1999 Models and mosaics: Investigating cross-cultural differences in risk perception and risk preference. *Psychonomic Bulletin & Review* **6**: 611–7

Yates J F, Stone E R 1992 Risk appraisal. In: Yates J F (ed.) *Risk-taking Behavior*. Wiley, Chichester, UK

E. U. Weber

Risk Screening, Testing, and Diagnosis: Ethical Aspects

The Human Genome Organization project has increased enormously the possibilities to predict disability or disease either for an individual or for his or her offspring. Information obtained from genetic testing raises ethical and legal issues that have been debated for several decades. New issues are emerging due to rapid advances in molecular genetics.

From an ethical point of view, risk screening, genetic testing, and diagnosis belong to the most controversial and contested areas in contemporary medicine. The new genetics will have a profound impact on our self-images and on our understanding of the relations between environment and heritage. Conceptual, normative, and empirical issues are here intertwined in a complex way. To begin with, a preliminary clarification of the key concepts will be necessary.

The expression 'risk' will here be taken to refer to an adverse or negative future event that is probable but not certain. Any risk can then be analyzed in two components: the probability of occurrence of an event and the magnitude of its negativity according to certain explicitly stated norms and values (including kind, degree, duration, and timing). Risk assessment accordingly analyzes these two components, while risk management proposes strategies for handling the risk in ways that minimize the negativity and probability.

However, risks involving probabilities are not the only sort of risks that are relevant in this context. The expression 'epistemic risk' has been coined to refer to risks due to ignorance, and in particular due to ignorance of what we are ignorant about (e.g., Sahlin and Persson1994). This notion is particularly relevant to many clinical applications of molecular genetics, involving unknown factors. The ways in which these risks are communicated will then be important from an ethical point of view.

Testing should be distinguished from screening. Harper has proposed the following definition: 'Genetic testing is the analysis of a specific gene, its product of function, or other DNA and chromosome analysis, to detect or exclude an alteration likely to be associated with a genetic disorder' (Harper and Clarke 1997). Testing refers to procedures performed at the request of individuals or families, and should always be voluntary. This holds also for genetic testing of members of families known to be at high risk, such as the siblings of persons with cystic fibrosis (Wertz and Fletcher 1989).

'Screening,' however, implies application to large population groups, to entire populations or subsets of populations (e.g., pregnant women, newborns, or job applicants). Screening is not performed at the request of individuals or families but rather based on policy or public health decisions. It may be mandatory or voluntary.

Both the expressions 'genetic information' and 'genetic disorder' are vague and ambiguous concepts (Zimmern 1999). The same holds for the key concepts of health, disease, and illness. The advances in molecular and clinical genetics may profoundly change our understanding of the latter concepts. They may have an impact on health care policy, on the patient–physician relationship as well as the way in which access to care is allocated (Murray et al. 1996).

Genetic information differs from other private information in that it reveals information not only about a particular individual but also about that individual's blood relatives. Genetic test results may also provide information about population groups. Moreover, genetic information can be obtained from any cell in a person's body, not just by examining, for example, particular malfunctioning organs or tissues.

The concept of genetic information is difficult to define and delimit in a clear and uncontroversial way. Just by asking for the age of a person's parents and grandparents, genetic information in a certain sense can be obtained. The necessary and sufficient conditions for a disease or disorder to be 'genetic' are far from clear. If these expressions 'genetic information' and 'genetic disorder' are to appear in legislative texts, they have to be made precise.

There are many kinds of genetic disorders, including (1) chromosome changes, like trisomy 21 or Down's syndrome, (2) monogenetic disorders which depend on a mutation in one gene, like Huntington's chorea, and (3) polygenetic disorders like diabetes which are due to mutations in several genes in combination with environmental factors.

These distinctions are important for several reasons. What holds for monogenetic diseases concerning genetic determinism cannot be generalized to genetic disorders of other kinds. Thus, presymptomatic testing can be done for disorders in group (2) but rarely and with much less certainty, if at all, for those in group (3).

1. Public Concerns

Genetic testing may be beneficial in those cases where early detection makes a difference, for example, makes it possible to prevent or cure the disease. The validity and the reliability of the test methods will then be of crucial importance. The background of many national screening programs, like the one launched in Ontario, Canada, in 1999, to detect risk of hereditary breast, ovarian, and colon cancers is that early detection can lead to medical interventions and surveillance. This can significantly reduce the risk of developing the disease and hence be cost-effective.

The concern expressed over genetic testing and screening, however, is that it may threaten privacy and civil rights and be the basis of genetic discrimination or stigmatization of individuals as well as of groups. Moreover, it has been feared that widespread use of genetic testing will help to foster intolerance of people with genetic variations. It may be difficult, too, if not impossible, for people with certain inherited disorders to obtain a life insurance. Finally, people may feel a pressure to be tested, as a result of cost–benefit analysis. In general, of course, the negative consequences of testing people and disclosing information about their test results to blood relatives will have to be weighed against the benefits of this, if any, for those concerned. However, this weighing operation is far from simple.

Guidelines for genetic testing and screening have been proposed in the attempt to maximize the beneficial consequences and minimize the adverse ones (by e.g., UNESCO 1997, Council of Europe 1992, 1996, Nuffield Council on Bioethics 1993, Danish Council of Ethics 1993). These guidelines include discussions of genetic counseling, informed consent, disclosure to individuals and family members, confidentiality, employment, insurance, as well as public policy. The commercial availability of genetic test kits raises both operational and ethical issues, dealt with in the principles and recommendations suggested by National Institutes of Health (1997). This report discusses the validity of genetic testing, deals with genetic counseling and the informed consent process, as well as recommendations on quality assurance measures for laboratories performing genetic tests and the need for a national body with the authority to review genetic testing procedures.

2. Ethical Issues

It has often been recommended that genetic information in health care should be acquired and used 'in a manner that respects the autonomy of individuals' (Institute of Medicine 1994), but it has proved hard to explain the precise meaning of this in a noncontroversial way. The autonomy of what individuals should be respected? Sometimes respecting the autonomy of one individual may conflict with respecting the autonomy of another.

A general problem raised by all forms of genetic testing and screening is, first, how to give nondirective information, that is, information that makes autonomous decisions possible by individuals and couples, and then to obtain consent in a way that avoids exercise of pressure.

Why is nondirectiveness an important goal? In brief, the reasons are the value of autonomy and the importance for geneticists to dissociate themselves from the dark history of genetics, from abuse of genetics for eugenic purposes. But nondirectiveness in practice is not uncomplicated, and the goal is also controversial. There is now a debate on the process of genetic counseling beyond nondirectiveness (Clarke 1994, Harper and Clarke 1997).

Particular problems are raised by the testing of individuals or groups with reduced autonomy, like children, demented people, or immigrants who do not understand the language. The testing of a healthy child could arise in at least three different contexts (Chadwick 1998): predictive testing, to see if the child will develop a specific disease that runs in the family; carrier testing, to see if the child may be at risk of having an affected child; and a screening test unrelated to family history to identify children with increased susceptibility to some common disease such as diabetes or ischaemic heart disease. Children are vulnerable and their parents therefore have to be involved in the decision procedure. But how? The nature of parental consent and the degree to which children are to be involved according to their maturity need close examination. If relatives are to be informed, who should do this? The geneticist, the hospital, or those already tested? And in what way? How many relatives should be informed? How closely related to the person asking to be tested (the index person) should the relatives be for the clinical geneticist to be responsible for informing them?

Different consent models are available, each with their own pros and cons. One main alternative is called 'opt in,' which means that if you have not said yes, you have said no; another is called 'opt out,' which means that if you have not said no, you have said yes. The conditions under which any of these models may or must be used is an important ethical issue.

The challenge to genetic privacy concerns the control of personal genetic information. Privacy has two dimensions, according to the Canadian Human Rights Commission (1992): protection from the intrusion of others and protection from one's own secrets. Does this entail a right to genetic ignorance, a right not to know about one's own or one's relatives' genetic risk profile? Is there a duty for geneticists or tested persons to disclose such information to those who might benefit from it? What is the basis of such rights and duties, and what are the consequences for the various people concerned of accepting or rejecting

them? Opposite positions have been argued for by Rhodes (1998) and Takala and Hayry (2000). Several compromise positions have been put forward. Tested persons should be encouraged to share what they know about their genetic risk profile with their blood relatives, when this information would be of vital health interest to their relatives (Chadwick et al. 1997, Hermerén 2000).

The privacy and confidentiality of the information needs to be protected against third party interests, including other family members, employers, and insurers. It has been suggested that genetic information should not be given to unrelated third parties, without the explicit and informed consent of the tested persons. In many countries, access to genetic information is regulated by law. The conditions or criteria under which medical confidentiality may or must be breached are bound to be controversial (Harper and Clarke 1997).

The practical problems raised by genetic testing also have to do with how to estimate and evaluate risks, communicate them in an understandable way, and decide which risks are acceptable and which are not. This includes the risk that a person will have an increased risk for a genetic disorder and the risk that the test result is not accurate. Communicating risks is obviously an important part of the genetic counseling process.

The accuracy of the tests—their validity and reliability—and the predictive value of the test or the screening method is another important issue in this context. To estimate the predictive value, one has to be able to compute the number of false positives and false negatives as well as to know about the prevalence of the genetic disorder.

Finally, there are issues of justice and fair access. If testing and screening is to be done on an equitable basis, it should not be available on the basis of such accidents as in which geographical region a person happens to be born.

These ethical problems arise in somewhat different ways, according to the type and purpose of testing or screening. The difficulties are not the same in all cases. In particular, certain forms of screening and presymptomatic testing for late onset genetic disorders present problems from an ethical point of view. Some of the most common types of genetic testing and screening will therefore be described below along with a brief indication of the particular ethical problems they raise.

3. *Types of Testing and Screening*

3.1 *Presymptomatic Testing*

Individuals may be tested for monogenetic disorders (of which there are many though each is fairly rare) by DNA analysis before any symptoms are shown. If an early diagnosis is of value, because the disease can be prevented, presymptomatic testing is fairly uncontroversial. If the disorder cannot be prevented, the value of presymptomatic testing is more difficult to judge. This holds also in cases where there is a method of prevention but it is nevertheless impossible to guarantee that the person tested positively will not get the disease (for example, breast cancer).

In addition to the obvious difficulty of informing people in such a way that they can make autonomous decisions, presymptomatic testing raises problems about how the information obtained from the tests should be handled, and who should have access to it. Likewise, it is controversial whether blood relatives should be contacted and who should contact them in that case (the person tested, the hospital, the clinic, the geneticist, etc.). Moreover, in many countries there is great concern about how to prevent certain external third parties, like employers or insurers, from getting access to information about the test result.

3.2 *Prenatal Diagnosis*

Prenatal diagnosis can be regarded as a special case of presymptomatic testing. By prenatal diagnosis it is possible to discern whether a fetus is at risk for various disabilities or diseases.

There are several methods for prenatal diagnosis. Some are visual—either noninvasive, like ultra-sonography, or invasive, like embryoscopy or fetoscopy, which uses a camera on a needle inserted in the uterus to view the fetus. Other methods are based on analysis of fetal tissues (like amniocentesis or chorionic villus sampling). By ultrasound tomography it is possible to create fetal images on a screen.

The indications of when prenatal diagnosis should be offered include advanced maternal age, a genetic history of abnormalities in the family, repeated miscarriages, or previous infants with birth defects. Then a prenatal diagnosis may be offered to provide information to couples about what they can expect. A normal result is found in most cases. If not, prenatal diagnosis may make it possible for them to adjust and may also influence the way in which the baby is delivered.

There has been great concern about the risk of spontaneous abortion induced by the use of amniocentesis or chorionic villus sampling. This is a problem one does not have with ultrasonography. On the other hand, ultrasonography may give information that is unexpected or not wanted, for example, that some organs or limbs are malformed. With other methods, too, one may make unexpected or unwanted discoveries, for example, that the social father is not the biological father.

All this serves to underline the necessity of careful counseling before the diagnosis, so that the woman or

the couple knows what the test may show and how certain the result is. Informed consent is crucial, and the difficulties of nondirective counseling should not be underestimated.

Other ethical issues raised by this type of diagnosis are concerned with how to preserve the confidentiality of the information obtained. Since prenatal diagnostics sometimes leads to abortion, it may be controversial in the eyes of some people for the same reasons that abortion is controversial. The impact of prenatal diagnosis on our views of people with handicap has also been a matter of concern.

3.3 Preimplantation Diagnosis

Preimplantation diagnosis means that the diagnosis is carried out before (in connection with *in vitro* fertilization) the pre-embryo is implanted in the woman's uterus. If there were a history of sex-linked diseases in a family, it would then be possible to prevent a child with such a disease being born. Thus the resulting pregnancy could be normal and the woman would not have to abort the fetus at a later stage or give birth to a child with a very serious or lethal disease. Preimplantation diagnosis is becoming an established medical practice in many countries.

The first serious ethical problem that arises is to whom these tests should be offered. Usually, they are limited to severe illnesses and to persons in specific 'high risk' groups. But both these criteria are value laden, and the indications may be narrowed down or widened. There are also considerable variations not only between lay people but also between geneticists over the severity of different genetic conditions (Chadwick 1998).

Preimplantation diagnosis at the beginning of the twenty-first century is thus limited to sorting out fertilized eggs carrying genes associated with severe diseases for which there is no cure at present, like Huntington's chorea and certain X-linked diseases. Concern has been expressed that the practice may be extended to diseases that are less serious (Shenfield 1997). Obviously, there is a risk of a slippery slope in this area.

Preimplantation diagnosis also, at least theoretically, opens up possibilities for a shift of power from the woman to the geneticist, as well as for cloning and manipulation of human pre-embryos that many people find too risky or simply ethically unacceptable. In some countries, like the UK, manipulation of human pre-embryos is illegal.

3.4 Prenatal and Neonatal Screening

Newborn screening involves the analysis of blood or tissue samples taken in early infancy in order to detect inherited diseases where early intervention can minimize or eliminate the risk of later serious health problems or premature death.

In screening, a group of individuals is examined directly with DNA, RNA or biochemical analysis to find those who will later in life have an increased risk of developing a particular genetic disorder. A good screening method is reliable and valid, with a minimum of false positives (who are worried unnecessarily) and false negatives (who are lulled into false security).

The best-known example of uncontroversial neonatal screening is screening for phenylketonuria (PKU), a genetic disease. The reason is that PKU, if undetected, can lead to serious mental retardation. Moreover, it can be discovered easily and prevented through early intervention. For many other conditions, such as cystic fibrosis and galactosemia, the benefits of early detection are more controversial. Screening programs for newborns also raise the issues of freedom of choice and informed consent by parents (Hermerén 1999).

Newborn screening raises several ethical issues, among them the nature of informed parental consent and how it is obtained, confidentiality and privacy, and the interests of third parties, as well as issues of justice such as equal access to testing and treatment. Moreover, it is clear that more research needs to be carried out on the psychological responses to newborn screening, particularly when the health benefits are disputed (Harper and Clarke 1997).

People with certain genetic disorders have had difficulties in getting a life insurance. In some countries the right of insurance companies to ask for genetic information has been regulated by law. In Norway, it is forbidden. In other countries—The Netherlands and Sweden, for instance—there has been a temporary voluntary moratorium during which the insurance companies will not ask for genetic information, provided that the life insurance asked for will not exceed a certain amount. This is still a very controversial area (Sandberg 1996).

3.5 Carrier Screening

Carrier screening identifies individuals with a gene or chromosome abnormality that may cause problems either for offspring or for the person screened. The testing of blood or tissue samples can indicate the existence of particular genetic trait changes in chromosomes or changes in DNA that are associated with inherited diseases in asymptomatic individuals. Carrier screening exists for sickle cell anemia, for Tay Sachs disease, as well as for cystic fibrosis, Duchenne muscular dystrophy, hemophilia, and Huntington's chorea.

Carriers have one normal copy of a gene and one copy that varies from the normal gene. Since one gene is normal, the carriers ordinarily do not exhibit symptoms of a genetic disorder. A woman who is a

carrier of an X-linked disease such as Duchenne muscular dystrophy or fragile X syndrome, has a variant gene on one of their two X (sex-determining) chromosomes. Her risk of transmitting the variant gene to each child can be estimated accurately.

The purpose of carrier testing is to inform potential parents of their genetic risk profile so that they can make informed reproductive decisions. National screening programs are usually justified by cost–benefit analysis. The question is then which costs and whose costs are taken into account, and how the benefits are defined and evaluated. The importance of cultural differences as to what constitutes benefits, and how benefits are ranked, should not be underestimated.

The right to autonomous choices is an important issue also in this area. Whether the choices open to people found to carry a disease gene are acceptable or not, may depend on their life plans, their ethical views, the laws of their country, as well as their culture and social customs. Before a carrier screening program is decided on and implemented, these questions need to be considered. In practice there may be subtle coercion, even if a carrier screening program is formally voluntary.

Screening the subset of a population at risk rather than the entire population makes sense from a medical and economic point of view, but selecting target groups for such screening programs also creates ethical problems. For example, there have been in the past controversies over whether these screening programs in effect had a eugenic purpose, particularly when the diseases screened for are over-represented in certain ethnic groups.

3.6 Testing in the Work Place

Susceptibility screening can be used as a form of presymptomatic testing, but is also used to identify workers who may be susceptible to toxic or carcinogenic substances such as benzene in their workplace (Van Damme et al. 1997, Surralles et al. 1997, Jarvholm et al. 1999). This susceptibility may lead to occupational health problems and future severe disabilities.

Van Damme et al. (1995) have proposed a conceptual model of the complex interactions between exposure, acquired and inherited susceptibility, and risk for disease. The validity of tests for determining genotype and phenotype and their relevance for the disease must be evaluated critically to provide an objective basis for ethical discussions. The acceptability of such tests is related to a number of issues which Van Damme et al. identify and discuss.

Genetic testing of employees and job applicants is controversial. The first basic question concerning testing in the work place is simple—for whose sake is the test performed? Is the test performed in order to protect persons with increased (inherited) susceptibility to hazards at the workplace? Or is the test in the interest of the employer, so that the employer can avoid improving the working conditions and avoid hiring people who would be damaged by the work they do? Here it is essential to separate the situation of the job applicant from the situation of the person already with a job. The latter could be monitored for exposures and early effects of, for example, toxic substances on the genes.

There are methods of detecting inherited predisposition to cardiovascular diseases as well as to adverse effects from otherwise well-tolerated exposures related to the job. Clearly, with widespread use of genetic testing in the workplace there is the obvious danger of discrimination and stigmatization of individuals or groups. It may be tempting for employers to use tests to select only workers who will be able to work in the present working conditions, instead of improving the working conditions. The result could be that large groups of people would be barred from the job market.

3.7 Forensic Testing

A comparatively new use of genetic testing is to examine a possible genetic linkage between, for example, alleged fathers and their children, or between suspects and evidence discovered in criminal investigations, like hair, blood, sperm, saliva, or skin. The idea is to use genetic methods of analysis to clear the innocent and identify the guilty. However, concern has been expressed that mistakes can be made. Critics argue that it is necessary to improve the quality control of the test laboratories. Another problem is the confidentiality of the DNA profiles obtained from criminal investigations and stored in national police databanks. For how long time will they be stored? Who will have access to them and for what purpose?

4. Concluding Remarks

Genetic testing and screening should be done for 'health purposes.' This is sometimes stated as a necessary condition for genetic testing (Council of Europe 1992, 1996). Health purposes could include diagnostic and therapeutic purposes as well as medical research and health care planning. Moreover, 'health' can be defined in several ways. It is important to state more precisely what this means.

Careful counseling remains crucial. One of the present concerns is that do-it-yourself test kits for genetic disorders may be bought via the Internet or over the counter without any professional counseling to explain what the findings mean, how certain the test is, etc. There is an urgent need to find ways of regulating the economic interest of commercial biotechnological companies in order to minimize the harm of premature use and overuse of genetic testing.

See also: Bioethics: Examples from the Life Sciences; Eugenics as an International Movement; Eugenics as the Basis of Population Policy; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease-related Characteristics; Genetic Testing and Counseling: Educational and Professional Aspects; Reproductive Medicine: Ethical Aspects

Bibliography

Canadian Human Rights Commission. Privacy Commissioner 1992 *Genetic Testing and Privacy*. Privacy Commissioner of Canada, Ottawa, ON

Chadwick R 1998 *The Ethics of Genetic Screening*. Kluwer, Amsterdam, The Netherlands

Chadwick R, Levitt M, Shickle D 1997 *The Right to Know and the Right Not to Know*. Avebury, Aldershot, UK

Chadwick R, H ten Have, Husted J, Levitt M, McGleenan D, Shickle D, Wiesing U 1998 Genetic testing and ethics: European perspectives. *Journal of Medicine and Philosophy* **23**: 255–73

Clarke A (ed.) 1994 *Genetic Counselling. Practice and Principles*. Routledge, London

Clarke A, Flinter F 1996 The genetic testing of children: A clinical perspective. In: Marteau T M, Richards M P (eds.) *The Troubled Helix: Social and Psychological Implications of the New Human Genetics*. Cambridge University Press, Cambridge, pp. 164–7

Council of Europe 1992 *Draft Recommendation on Genetic Testing and Screening for Health Care Purposes*. Council of Europe, Strasbourg, France

Council of Europe 1996 *Convention on Human Rights and Biomedicine*. Council of Europe, Strasbourg, France

Danish Council of Ethics 1993 *Ethics and the Mapping of the Human Genome*. Danish Council of Ethics, Copenhagen, Denmark

Harper P S, Clarke A J 1997 *Genetics, Society and Clinical Practice*. Bios, Oxford, UK

Hepburn E 1996 Genetic testing and early diagnosis and intervention: Boon or burden? *Journal of Medical Ethics* **22**(2): 105–10

Hermerén G 1999 Neonatal screening: Ethical aspects. *Acta Padiatrica* Suppl. **432**: 99–103

Hermerén G 2000 The right to know and the right not to know. In: *Who Owns Our Genes?* Danish Council of Ethics, Copenhagen, Denmark, pp. 135–57

Institute of Medicine 1994 (Andrew L B (ed.)) *Assessing Genetic Risks: Implications for Health and Social Policy*. National Academy Press, Washington, DC

Jarvholm B, Nordstrom G, Hogstedt B, Levin J O, Wahlstrom J, Ostman C, Bergendahl C 1999 Exposure to polycyclic aromatic hydrocarbons and genotoxic effects on nonsmoking Swedish road pavement workers. *Scand. J. Work Environ. Health* **25**(2): 131–6

Murray T H 1983 Warning: screening workers for genetic risk. *Hastings Center Report* **13**(1): 5–8

Murray T H, Rothstein M A, Murray R F (eds.) 1996 *The Human Genome Project and the Future of Health Care*. Indiana University Press, Bloomington, IN

National Institutes of Health (National Human Genome Research Institute) 1997 *Promoting Safe and Effective Genetic Testing in the United States: Principles and Recommendations* [the report is available on the Internet at http://www.nhgri.nih.gov/ELSI/TFGT_final]

Nuffield Council on Bioethics 1993 *Genetic Screening: Ethical Issues*. Nuffield, London

Rhodes R 1998 Genetic links, family ties, and social bonds: Rights and responsibilities in the face of genetic knowledge. *J. Med. Philos.* **23**(1): 10–30

Sahlin N-E, Persson J 1994 Epistemic risk: the significance of knowing what one does not know. In: Brehmer B, Sahlin N-E (eds.) *Future Risks and Risk Management*. Kluwer, Dordrecht, The Netherlands, pp. 37–62

Sandberg P 1996 *Genetic Information and Life Insurance: A Proposal for an Ethical European Policy*. Norwegian University of Science and Technology, Department of Biotechnology, Trondheim, Norway

Shenfield F, Sureau C 1997 Ethics of embryo research. In: Shenfield F, Sureau C (eds.) *Ethical Dilemmas in Assisted Reproduction*. Parthenon, New York, pp. 15–21

Shenfield F, Sureau C (eds.) 1997 *Ethical Dilemmas in Assisted Reproduction*. Parthenon, New York

Surralles J, Autio K, Nylund L, Jarventus H, Norppa H, Veidebaum T, Sorsa M, Peltonen K 1997 Molecular cytogenetic analysis of buccal cells and lymphocytes from benzene-exposed workers. *Carcinogenesis* **18**(4): 817–23

Takala T, Hayry M 2000 Genetic ignorance, moral obligations and social duties. *J. Med. Philos.* **25**(1): 107–13, Discussion 114–20

UNESCO 1997 *Universal Declaration on the Human Genome and Human Rights*. UNESCO, Paris

Van Damme K, Casteleyn L, Heseltine E, Huici A, Sorsa M, van Larebeke N, Vineis P 1995 Individual susceptibility and prevention of occupational diseases: Scientific and ethical issues. *J. Occup. Environ. Med.* **37**(1): 91–9

Van Damme K, Vineis P, Sorsa M, Casteleyn L 1997 Ethical issues in genetic screening and genetic monitoring of employees. *Ann. N. Y. Acad. Sci.* **837**: 554–65

Van Damme K, Casteleyn L 1998 Ethical, social and scientific problems related to the application of genetic screening and genetic monitoring for workers in the context of a European approach to health and safety at work. *La Medicina del Lavoro*, Vol. 89, suppl. 1

Wahlstrom J 1989 Inherited mental disorders. *Acta Psychiat. Scand.* **80**(2): 111–17

Wertz D, Fletcher J 1989 *Ethics and Human Genetics: A Cross-cultural Perspective*. Springer, Berlin

Zimmern P 1999 Genetic testing: A conceptual exploration. *Journal of Medical Ethics* **25**(2): 151–6

G. Hermerén

Risk, Sociological Study of

That the world is punctuated with uncertainty, danger, hazards, disasters, and threats to what humans value has been recognized since the beginning of human life on earth. Risk is the conceptualization of that recognition—providing a means for anticipating and avoiding or adapting to untoward outcomes. By 3200 BC, ancient Babylon had the first recorded risk assessors, and in 1700 BC, the Code of Hammurabi included a classification of risk matters. Despite these

ancient beginnings, however, and despite subsequent refinements accompanying the rise of the insurance industry during the age of exploration, the centrality of risk is the child of advanced industrialization, particularly since World War II. With advanced industrialization came technologies that were larger, more sophisticated, and potentially much riskier, having the capacity to affect far greater numbers of people, as well as the things they value and the areas they occupy. At the same time, scientific and technological advances have led to remarkably increased sophistication and precision, in many cases, in the ability to detect unwanted side effects or risks.

The notion of 'risk' has evolved over the last decades of the twentieth century to become the key analytical lens for anticipating our actions' consequences for the environment and ourselves. It includes an analytic orientation and a suite of evaluation methodologies, but it is also a new consciousness—a way of looking at a world of technological and environmental uncertainty. Attention to risk is a consciousness born of vast uncertainties over the durability of nuclear peace, the resilience of the ozone protecting us, threats of global warming, the growing extinction of species and the possibility of creating new ones, and more broadly, threats of technological disasters both large and small. It is a consciousness entirely foreign to the optimistic beginning of the century just ended.

The first significant body of sociological work on risk can be traced to the 1940s, with the work on natural hazards by Gilbert White (a social geographer), gaining momentum in the 1950s with White's collaboration with a number of sociologists, and evolving into the sociological specialty of natural disaster research. A key focus of this specialty has involved the social impacts of hazard events—how did structures and individuals respond to natural disasters such as storms, floods, earthquakes, wildfires, or droughts? One of the most consistent findings from this literature was the emergence of what has become labeled a 'therapeutic community': contrary to the self-focused maximizing predictions of utility theory, citizens typically engaged in acts of altruistic aid and community building (for reviews, see Drabek 1986, Kreps 1984).

Particularly in the United States, a significant fraction of the ongoing interest in natural disasters was motivated by the concerns of the early atomic age, including the prospect that a nuclear attack could wipe out a significant fraction of a community's population; given that much the same could be said for natural phenomena, such as floods, fires, or hurricanes, these and other 'natural disasters' came to be seen as useful 'natural experiments' for studying community responses. Particularly after the accident at the Three Mile Island nuclear power plant in 1979, however, it was a different form of nuclear activity that came to provide what is in many ways the prototypical 'risk' issue.

Although the point is often forgotten today, nuclear power facilities once enjoyed widespread support, even in the communities where they were constructed. Still, even before the 1979 accident, there were early signs of rising public concern. One response of nuclear power supporters (including those who controlled some of the most powerful branches of government) was to characterize the concerns as being ill-founded or 'irrational,' doing so by demonstrating that the 'real risks' were instead quite small. The earliest efforts, however, indicated that the consequences of a plausible 'worst-case' disaster could have been quite serious, indeed. These early efforts led to little publicity, but to a change of focus—an emphasis on what soon came to be called 'probabilistic' risk assessment, or PRA, a reductionistic engineering technique that estimates the risk of a plant mishap by estimating the probability of failure of all the plant's parts and subsystems and then aggregating these probabilities. The conclusion from the first systematic application of this technique to nuclear power plants, which was stated particularly forcefully in the executive summary for the multivolume effort that is still remembered as 'the' reactor safety study (US Atomic Energy Commission 1975), was that citizens were more likely to be killed by a falling meteorite than by a malfunction at a nuclear power plant.

Four years later came the accident at Three Mile Island, near Harrisburg, the capital city of the state of Pennsylvania—an accident that by technical analysis virtually could not happen. The accident brought to public and sociological attention an incipient schism between the state, its technological experts, and citizens. European sociological interest in risk was similarly intensified by the far more serious nuclear accident at Chernobyl in April 1986. While at one level of analysis these accidents and the vast media coverage surrounding them offered dramatic signals to publics of the risks of nuclear-generated electricity, at a deeper level they revealed conflicts between expert and lay knowledge, and between 'top-down' and 'bottom-up' strategies for managing large-scale, risky technologies.

The 1979 accident, in particular, generated a number of legacies. First among them was the further institutionalization of what soon came to be known as the field of risk analysis: the Society for Risk Analysis was established in the United States (with subsequent affiliates in Europe and Asia), as was the Division of Risk and Management Science at the US National Science Foundation, along with a number of new professional journals and a substantial outburst of published literature. That outburst included a significant increase in sociological work, with much of the initial increase paralleling the three general theoretical levels in sociology—the micro, the meso, and the macro, or in simpler terms, the social psychological, the organizational/community level, and the societal.

At the micro level, much of the work was conditioned by what soon became the dominant social

science research tradition in risk studies: the psychometric tradition (Slovic 1987). A growing body of work in cognitive psychology was brought to bear in examining the extent to which lay perceptions of risk differed from actuarial data and from experts' views. Key among the findings from this work, replicated in a variety of cultures, is the idea of heuristics—cognitive rules of thumb that provide shortcuts to risk estimation. One example is the so-called 'availability heuristic,' or the tendency to judge a risk on the basis of its cognitive 'availability,' or the ease of imagining or recalling examples; such a tendency increases the perceived riskiness of a low-probability but sensational risk, such as death from botulism, relative to more probable but more prosaic risks, such as deaths from asthma.

In political circles, the early psychometric findings were sometimes taken as reinforcing the argument that the 'irrational' public should be excluded from risk decisions, but the evidence proved not to be so clear-cut. Sociological work on technological controversies, for example, had already drawn attention to the lack of unanimity among experts (see e.g., Mazur 1980), and further research showed that experts relied on similar heuristics—even in the contexts of their own expertise, and even when the experts were studied with the same protocols used to study lay perceptions. Work from a social constructivist perspective (e.g., Wynne 1992) further challenged the tendency to equate expert risk estimates with 'real' risks, with Clarke (1999) eventually going so far as to argue that many official proclamations were little more than 'fantasy documents'—efforts to deal with ultimately uncontrollable risks through ritualistic proclamations of rationality. At the same time, psychometric work pointed to the importance of a number of broader issues, such as trustworthiness and fairness, that were omitted from technical risk analyses. Hence, from the psychological literature emerged a reaffirmation of sociology's first principle: context matters.

Sociologists responded with three sustained research efforts—one that contextualized the individual actors, a second that aligned risk choices with the most powerful technological decision-makers, namely organizations, and a third that examined the consequences of technological accidents in the contexts where people live, namely communities—as well as with broader and more conceptual work at the macrosocietal level.

At the individual level, the work showed that a social actor's perception and responses to risk are shaped not just by heuristics, but also by an entire set of social, political, and institutional forces. Perhaps the most comprehensive effort to capture this wide variety of psychological and social forces is the 'social amplification' framework, a conceptualization adapted from formal communication theory (Kasperson et al. 1988), and emphasizing that, when signals are sent from a source to a receiver, those signals often flow through intermediate transmitters that 'amplify' or attenuate the message. Similarly, individuals can receive risk signals either directly, through personal experience, or indirectly, through information provided by institutions (e.g., the media) or social networks (e.g., opinion leaders). These 'amplified' interpretations can then produce behavioral responses, resulting in secondary impacts that, in turn, are often themselves amplified, producing tertiary impacts. Partly because of its interdisciplinary orientation and its capacity to integrate the cumulative psychological findings with other factors that have been shown to influence risk perception and response—social, organizational, institutional, political—the framework has generated a considerable body of generally supportive empirical research.

With mesolevel work on organizations, meanwhile, sociologists began to point out the inherent metamethodological bias of early PRA work—the Cartesian reductionism of assuming that one can understand risk by understanding its underlying fundamental elements. A counterpoint to this methodological individualism was provided by the sociological holism of organizational sociology, treating organizations, not individuals, as the key risk decision-makers in contemporary societies.

The landmark contribution to this orientation is Perrow's *Normal Accidents* (1984), which argued that certain technological systems were complicated beyond our understanding of them—and beyond our ability to make them safe. Key factors included complexity and 'tight coupling'—essentially the lack of room for error—meaning that even attempts to improve safety, such as the introduction of redundant parts and backup systems, could result in further complication of these systems, thereby increasing risk. The book argued that some of our technological systems— nuclear power plants, nuclear weapons, and perhaps others—were predictable 'system' accidents, or 'normal accidents,' waiting to happen.

A counterpoint to this argument emerged in subsequent work on 'high reliability organizations' (LaPorte and Consolini 1991). Building on a set of fieldwork-based case studies—e.g., an air traffic control system, a large electric utility system, aircraft carrier operations—this school of thought argues that complex systems can be made safe with proper component, system, and organizational design, with specialized, sometimes variable, management structures, and through the development of a corporate culture of safety. Given that a great many of society's complex technological systems do indeed operate daily without incident, but also that other empirical tests appear to support normal accident theory, the jury remains out over which of these two schools of thought offers the deeper purchase on this key problem of advanced industrialization. The issue is only likely to be resolved with further refinement of the conditions

under which 'normal' accidents or 'high reliability' operations are more likely.

Another body of mesolevel work focused on community contexts. Particularly after Erikson's study of a so-called 'man-made' or 'non-natural' disaster (1976), a growing number of studies found that the impacts of technological or non-natural disasters proved to be far more severe and long-lasting than those associated with natural disasters. In contrast to the 'therapeutic community' of natural disasters, numerous studies of technological disasters have encountered what Freudenburg and Jones (1991) call a 'corrosive community,' typified not by offers of help, but by efforts to avoid responsibility or affix blame. Similarly, while the long-term social, economic, psychological, and cultural impacts of natural disasters have generally been found to have been surprisingly small, the long-term impacts of technological disasters have been found to be far greater than would be expected on the basis of the immediate physical destruction. At Three Mile Island, for example, most observers have concluded that only small amounts of radiation escaped, but careful studies found increased reports of 'intense distress symptoms'—that is, levels of distress symptoms that are characteristic of hospitalized mental patients—with later studies finding that stress levels were actually higher some six years later than they were in the immediate postaccident period. The symptoms were not merely self-reported, but included physiological measures such as elevated catecholamine levels in blood—a form of medical measure that is generally considered 'real' even by observers who question the legitimacy of people's own reports on stress and well-being. More broadly, several studies indicate that distress levels may actually be highest among the citizens afflicted not by the worst levels of disaster, but by an *ambiguity of harm*—cases where not even the best scientific work can clearly demonstrate whether people have been seriously endangered or not (Freudenburg and Jones 1991).

At the macro- or societal level, finally, a growing body of work has responded to the call of Short (1984), in his presidential address to the American Sociological Association, to examine the importance of risks to the social fabric itself. In the early days of the twentieth century, Weber had pointed out that the 'rationality' did not mean that the citizens of modernity would know more about their surroundings and tools than the citizens of a premodern age; instead, Weber emphasized, the modern citizen would see the tools and technologies as being 'knowable,' in principle, rather than being in the realm of magic. A century later, Weber's point is, if anything, more clearly true. Collectively, we know far more about our world and tools than did our great-great-great grandparents, but individually, we actually know far less about the tools and technologies on which we depend; instead, we literally 'depend on' them to work properly. In turn, that means that we depend on whole armies of specialists, most of whom we will never meet, let alone be able to control. Even if we find that we can usually depend on the technologies, and on the people who are responsible for them, the rare exceptions can be genuinely troubling. The risks to the social fabric, in other words, include what Freudenburg (1993) has termed 'recreancy'—cases where an organization or an individual entrusted with a specialized task fails to perform in a way that fulfills the responsibility and upholds the trust.

Still more broadly, while differing in ontological and epistemological assumptions, and while differing in orientation and research method, the various perspectives share the view that risk is an identifiable feature of the world, whether real or constructed, that risks are consequential in many facets of social life, and accordingly, the topic is worthy of sociological examination as well as policy attention (Rosa 1998). In addition, the pragmatic importance of new approaches to risk management reflects the growing recognition that the modern world has generated risk problems that demand scientific understanding, but that are too complex or too ambiguous to be 'answered' by science alone. Both the scientific examination and the policy making must take place in a context that is permeated by values, and also by unknowns—both those that can be recognized in advance and those that cannot. Owing to these conditions, democracies may have entered an era of postnormal risk (Funtowicz and Ravetz 1992), an era demanding better procedures for managing growing technological risks, procedures that require scientific inputs, but that broaden the scope of peer evaluation and integrate citizen involvement. A variety of social experiments have been undertaken to refine and implement these analytic-deliberative procedures, potentially reshaping not only technology policy but also democracy itself.

Risk is the unavoidable companion of a growing technological complexity. Managing large-scale technological risks has therefore become a central challenge for all societies—perhaps even the key basis for distinguishing contemporary from past societies (Beck 1986, Giddens 1990). It is this challenge that raises the fundamental and daunting question: are contemporary societies capable of generating technological complexity, technological interdependence, and pervasive risks at a more rapid pace than our growth in knowledge of these complexities and risks—and more rapidly than we can manage them safely? Whether a form of cultural lag, an inherent paradox of technological sophistication, an epiphenomenal feature of what we currently call postmodernity, this question will remain a focal point of discourse and decision making in the coming era, both for sociology and for society more broadly.

See also: Accidents, Normal; Disasters, Sociology of; Environmental Risk and Hazards; Health Insurance:

Economic and Risk Aspects; Industrial Sociology; Law and Risk; Occupational Health; Personality and Risk Taking; Risk, Sociology and Politics of; Risk: Theories of Decision and Choice; Workplace Safety and Health

Bibliography

Beck U 1986 *Risk Society: Towards a New Modernity* (trans. Ritter M). Sage, London

Clarke L 1999 *Mission Improbable: Using Fantasy Documents to Tame Disaster*. University of Chicago Press, Chicago

Drabek T E 1986 *Human System Responses to Disaster*. Springer-Verlag, New York

Erikson K T 1976 *Everything in its Path: The Destruction of Community in the Buffalo Creek Flood*. Simon and Schuster, New York

Freudenburg W R 1993 Risk and recreancy: Weber, the division of labor, and the rationality of risk perceptions. *Social Forces* **71**(4): 909–32

Freudenburg W R, Jones T R 1991 Attitudes and stress in the presence of technological risk: A test of the Supreme Court hypothesis. *Social Forces* **69**(4): 1143–68

Funtowicz S O, Ravetz J R 1992 Three types of risk assessment and the emergence of post-normal science. In: Krimsky S, Golding D (eds.) *Social Theories of Risk*. Praeger, Westport, CT, pp. 251–73

Giddens A 1990 *The Consequences of Modernity*. Stanford University Press, Stanford, CA

Kasperson R E, Renn O, Slovic P, Brown H S, Emel J, Goble R, Kasperson J X, Ratick J 1988 The social amplification of risk: A conceptual framework. *Risk Analysis* **8**: 177–87

Kreps G A 1984 Sociological inquiry and disaster research. *Annual Review of Sociology* **10**: 309–30

LaPorte T R, Consolini P M 1991 Working in practice but not in theory: theoretical challenges of high reliability organizations. *Journal of Public Administration Research and Theory* **1**: 19–47

Mazur A 1980 *The Dynamics of Technical Controversy*. Communications Press, Washington, DC

Perrow C 1984 *Normal Accidents: Living With High-risk Technologies*. Basic Books, New York

Rosa E A 1998 Metatheoretical foundations for post-normal risk. *Journal of Risk Research* **1**: 15–44

Short Jr J F 1984 The social fabric at risk: Toward the social transformation of risk analysis. *American Sociological Review* **49**: 711–25

Slovic P 1987 Perception of risk. *Science* **236**: 280–5

US Atomic Energy Commission (AEC) 1975 *Reactor Safety Study: An Assessment of Accident Risks in US Commercial Nuclear Power Plants*. US Nuclear Regulatory Commission (WASH-1400), Washington, DC

Wynne B 1992 Risk and social learning: Reification to engagement. In: Krimsky S, Golding D (eds.) *Social Theories of Risk*. Praeger, Westport, CT, pp. 276–97

E. A. Rosa and W. R. Freudenburg

Risk, Sociology and Politics of

Risk and chance have long been held to be key elements of the human predicament; but risk awareness besets whole cultures as well as individuals. Social scientists typically have been interested in risk, from both the social-psychological stance of particular actors and in terms of cultural responses to collective exposures. At one level, sociologists have discerned a historical trend towards the taming of risk, both in the sense that risk and probability have become better understood, and that certain forms of exposure to risk have been reduced. Nineteenth- and early twentieth-century confidence in progress and control over nature suggested both that risks were diminishing—diseases could be controlled, dangers to food supply countered through scientific farm management, and so on—and that better understanding of risk and chance was possible. But this modernist confidence about the prospects for the control of risk has been moderated and to some extent undermined by subsequent developments. As an example, twentieth-century technologies were associated with new forms of risk, including risks to the whole ecosystem. It seemed that the forces of progress could lead to novel hazards as well as to enhanced security. Furthermore, the legal frameworks and intellectual tools developed for the regulation of risk made it possible to elaborate arguments about risk in ever more sophisticated ways, ironically engendering a more vivid appreciation of risk. Finally, the complexity of social development fostered distinctively social risks, related to crime and the increasing uncertainty of social life. This discussion will begin by focusing on the regulatory treatment of risk before moving on to consider the broader social experience of risk.

1. Risk Assessment: Regulators and Risks

One systematic and extensive literature on risk grew out of concerns to make policy and to legislate for risk; this literature can be categorized as concerned with 'risk assessment.' Given the ubiquity of risks, the question for regulators was customarily thought of as: 'how safe is safe enough?' Governments and official agencies recognize that, regrettably, train crashes occur, that drivers are daily involved in automobile collisions, that leaks will sometimes happen even from well-managed chemical plants, and that farm chemicals may impact the environment, agricultural workers, or even occasional consumers. Given that no complex system can be guaranteed to be perfectly safe, the leading approach was to ask about the price of additional safety. From this consequentialist world view, risk is thought of as the mathematical product of the likelihood of a hazard occurring, and the costs to which that hazard gives rise. The policy maker's art is then to minimize risk within the budget that society is willing to spend on safety.

In practical terms, this task was always much more difficult than might be supposed, and certainly more

intractable than economists routinely implied. In part, difficulties arose because of the inherent problems of trying to harmonize risk data across various fields. It may be possible to compare the costs and risks of various railway signaling systems, but (in the absence of a super-ministry of risk reduction) transport, industrial, medical, and agricultural risks cannot be brought within the same calculus. Worse still, both the likelihood of the problem and its consequences commonly defy exact specification. While there are good data on the medical and related risks of typical motor accidents on US and European freeways, the probability and the consequences of nuclear power station incidents can only be calculated in hypothetical ways. Similarly, the risks and costs of 'mad-cow disease' (bovine spongiform encephalopathy, or BSE), an unprecedented form of infection apparently spread through the food chain, could not be subjected to the risk calculus in the standard way. Further, the only way to obtain quantitative measures of many risks to humans is to extrapolate from indirect observations, such as rodent bioassays, thereby introducing additional sources of distortion into risk assessment. Without good data, all calculations will be 'rough and ready.' Nor is it even clear that there is a single 'currency' into which all sorts of harms can be converted for the purposes of cost–benefit calculations.

Despite these persistent difficulties, one line of the literature about risk has been devoted to exploring what are thought to be the public's shortcomings in relation to risk understandings. Lay people appear to make irrational and statistically unsupported assessments of the relative risks of different hazards, and to demand high levels of risk reduction without being aware of the costs. In the extreme, the public appears unwilling even to accept the cost–benefit approach, but has no systematic alternative with which to replace it. As an added complication, despite frequent public protestations of anxieties about risk, market mechanisms seem to imply that people are not necessarily very risk averse: The automobile market has tended to sell on aesthetic and performance criteria rather than on safety; and consumers continue to eat fatty foods and to avoid exercise despite well publicized links to heart disease.

The apparent discrepancies between public and expert versions of risk have given rise to a series of psychometric studies (Slovic 1992). For example, researchers have shown a tendency by people to undervalue systematically the risks to which they are routinely exposed (because of their lifestyle or occupation), and to overvalue the seriousness of risks that are novel or could be catastrophic. Moreover, there appear to be significant differences between people's assessment of the risks to which they expose themselves and those to which they believe they are subjected by others. These perceptions translate into corresponding pressures in policy. People seem more likely to demand that automobiles be engineered to lower traffic fatalities than they are to constrain themselves to drive those automobiles more cautiously.

The resulting tension between public and expert interpretations of risk has caused a problem for policy makers. If they base regulations on expert judgments—that is, keyed only to the statistical probability of harm—policies may be unpopular or even subverted, whereas basing policies on the public's apparent preferences threatens to make regulations arbitrary, unscientific, or too costly (Breyer 1993). If the citizenry really is more tolerant of self-imposed risks than of risks visited on them by others, then there is an argument for reflecting this in public policy, whatever the 'actual' exposure to risk. Similarly, if, as Slovic and colleagues report (Slovic 1992, p. 121), the public is more concerned about certain 'dread' risks than other risks which experts hold to be of equal hazard, then maybe the cost–benefit equation has to be widened to take into account people's manifest preferences. Regulators find themselves confronting a tension familiar to liberal democracies: That between people's 'revealed' preferences and the recommendations supported by expert opinion.

2. *Risk Expertise: The Reflexivity of Risk*

These problems have become amplified in particular institutional contexts in ways that social science research has sought to explicate. Risk assessments have been developed, particularly by official regulatory agencies, so as to apply to statistically representative, or in some other sense, 'average' people (Jasanoff 1990). Calculations of how long it takes to evacuate an aircraft, and thus of the requisite number of emergency exits, the width of aisles and so on, are supported by evacuation trials, but these in turn depend on notions of what is a typical passenger cohort. Similarly, arguments about exposure to pollutants have had to construct some notion of the average person as the unit for measuring the at-risk population (with, in many cases, the need to substitute the average risk animal for the average human). Yet this ideal type hardly exists in reality and may obscure threats to specific subpopulations. Women may be different from men; pregnant women are more clearly different. The young may differ from the elderly, the housebound from the active, and so on. These differences may prove to be far from 'academic' in particular contexts of exposure. For example, people who live near one of the UK's most controversial nuclear power plant sites, Sellafield, and who happen to favor a shellfish-rich diet may be exposed to a greater nuclear hazard than the average person, because of the way in which shellfish filter material from sea water. In this case, a behavioral choice exacerbates other risk factors. In other cases, however, numerous background

factors may act in synergy for the poor or socially disadvantaged, as in the case of American inner-city minority populations, who may be exposed to toxic pollutants from multiple sources (Bullard 1994). In neither the UK nor the US cases would standard risk-assessment methodologies, based on average characteristics, compute adequately the risk to specially vulnerable populations. Indeed, technical risk assessment has tended to obscure the distributive implications of risk production. It is, in this sense, a social technology that reinforces existing structures of power (Winner 1986).

Ironically, the great increase in formal examinations of risks has not so much diminished as contributed to the exacerbation of risk concerns (Beck 1986). Given the complexities of risk assessment and the high stakes involved, it is understandable that risk assessment methodologies have been subject to legal and other formal challenges (see Jasanoff 1990). These proceedings have subjected risk assessments to critical deconstruction, questioning the precise basis for the choice of methodologies. Where these methods have had consequences injurious to some but not all sections of the population, the procedure has been made to appear discriminatory. Environmental justice movements have formed to press the point that risk assessments of individual chemicals are not equivalent to a demonstration of the impact of a cocktail of pollutants. Communities experiencing extensive hazardous exposures—each element of which may be deemed not to be unduly risky on its own—argue that the overall impact of multiple sources is not well gauged by standard methodologies.

Furthermore, as Jasanoff (1990) has demonstrated, particularly in the US, the combination of adversarial cross-examination with the separation of powers and scope for judicial review of executive agencies, has meant that extensive financial and intellectual resources have been directed at deconstructing risk assessments. Given all the complexities outlined above (the hypothetical nature of many risks, the difficulty of identifying an average case, the impossibility of most human experimentation, the need for surrogate measures, and so on), there has been no prospect of finding an incontestable scientific basis for defending particular risk assessments.

Sociological research has focused also on the strategic constructions of certainty and uncertainty associated with risk assessments. Thus, Wynne (1992) has argued that risks and probabilities are made up of many kinds of not-knowing, only a fraction of which (e.g., uncertainties of extrapolation from rodent bioassays) may be acknowledged openly in formal risk assessment processes. In some cases, it is possible to establish a hierarchy of uncertainty, as between 'risks' and 'uncertainties': With risks one knows the odds, but with uncertainties, only the general parameters of the problem. Most practical questions which science-in-public has to face, however, involve an additional kind of noncertainty. This Wynne terms ignorance. Ignorance refers to aspects of a problem which are consciously or unconsciously bracketed off, and commonly not further investigated. These may be issues which lie outside the disciplinary paradigms of the sciences and are thus in a sense necessarily, rather than perniciously, excluded from day-to-day research. Nonetheless, ignorance in this sense is a different form of not-knowing than mere uncertainty; it is not adequately captured by being treated simply as extreme uncertainty.

In principle at least, more knowledge might assist in handling these kinds of not-knowing. Uncertainties might be turned into risks. New understandings might clarify specific areas of former ignorance, though, of course, there is no prospect of ignorance being overcome in general. But, in addition, Wynne argues that there is a fourth consideration, 'indeterminacy,' resulting from 'real open-endedness in the sense that outcomes depend on how intermediate actors will behave' (Wynne 1992, p. 117). In other words, the safety or reliability of systems with an organizational or human component are deeply dependent on how the systems are operated. By treating most systems as determinate, conventional risk-assessment practices typically adopt tacit, unexamined and untested sociological hypotheses about those social practices that are central to the risk-producing activities. Thus, evaluations of risks from potentially hazardous technologies do not depend only on the behavior of physical and biological systems (difficult to model though they may be) but every bit as much on the behavior of plant managers, operators, and even those exposed.

Several insights into the sociology of expertise follow from these claims. First Wynne (1992) observes that experts, when faced with the need or chance to regulate in an area of noncertainty, are tempted to treat all forms of not-knowing as statistically treatable uncertainty, even though things of which they are ignorant cannot, by definition, be quantified. Second, he argues that the public may be significantly more expert than the scientists in relation to some of the matters covered by indeterminacy. His view, as expressed with Irwin, is that 'science offers a framework which is unavoidably social as well as technical since in public domains scientific knowledge embodies implicit models or assumptions about the social world' (Irwin and Wynne 1996, pp. 2–3). In other words, insofar as expert assessments depend on assumptions about particular social, cultural, or occupational practices of lay groups, it is likely that these publics will be more expert in these matters than technical 'experts' more distant from the relevant experiential insights.

A related body of research ties this manifestation of hazard within complex sociotechnical systems to features of organizational practice and culture. Thus, Perrow's discussion of 'normal accidents' (1984) makes the case convincingly that many hazardous occur-

rences are a routine (if often unpredictable) outcome of the interaction of the manifold technical and organizational components of modern power stations and production plants. Accidents are to be expected even if no particular accident is predictable. More recently, Vaughan has applied associated insights to understanding the decision to launch the ill-fated Challenger mission (Vaughan 1996). Of course, questions still remain about the extent to which organizational arrangements give priority to concerns for production, or for safety (Clarke 1989), and many recent case studies show corporations placing the most value on production and profitability.

In sum, though risk discourse has been promoted by scientists and by scientific attitudes to the valuation of the natural world, it is clear that public experience of risk assessment has encouraged successful challenges to the official orthodoxy, as in the emergence of an environmental justice agenda, and that 'scientific' risk assessment has been undermined dialectically by the advance of legal–rational reflection. In response to such difficulties, official agencies commonly are left with no alternative but to demand 'more and better' science; yet there are few grounds for thinking that further research or rationalization will resolve the problems outlined above. Ironically, social movement campaigners find themselves confronting a similar impasse. In opposing the establishment they commonly produce counter-claims about health or environmental risks, yet find related difficulties in underpinning their risk claims. Such experiences often generate an ambivalent attitude toward scientific expertise (Yearley 1992). These phenomena serve as a good example of what Ulrich Beck (1986) has termed 'reflexive modernization.' The institutions of modernity, notably the traditions of scientific analysis and legal fact-finding, have been turned upon themselves with deconstructive consequences. For Beck, this reflexive modernization is just one facet of the 'risk society' thesis, which portrays late-millennial industrialized societies as being unusually and overwhelmingly concerned with the distribution of risks and other 'bads.' In this view, social scientific interest in risk consists not so much in the study of societal responses to particular hazards; rather it is the key to characterizing the dynamics of society as a whole.

3. *Risks and Cultures*

Beck's thesis in its widest form asserts that the most modern risks (termed by others 'late' or 'high' modern (Giddens 1991)) are reflexive in the sense that they are self-induced. Risks in the early modern period were external to the self-conscious control of social actors. Diseases would spread, bad weather would damage harvests, fires would consume urban areas as though under the influence of external, natural forces. Even if some of these risks were exacerbated by human interventions, the contemporary perception was that they were uncontrollable. In high-modernity, by contrast, risks such as the threat of disastrous nuclear reactor incidents are plainly the consequences of human activities (Beck 1986). In this view, Victorian and early-twentieth-century confidence about the progressive diminution of risk marks, in an ironic fashion, the transition from apparently external to societally-induced risk. Scientific and technological developments are caught up in this reflexive pattern because modern risks are typically the result of technological ventures (nuclear power, ozone-depleting chemicals in the earth's atmosphere); science and technology are involved in the cause, the diagnosis, and, with luck, eventual rectification of the problem.

Beck's overall analytical claim has been greeted with widespread enthusiasm by many social scientists, and the term 'risk society' has popularly been adopted. However, the details of his analysis have not met with the same welcome. For one thing, it is unclear how 'modern' all present-day risks are. The incidence of 'mad-cow disease'—eventually apparently transmitted to humans—is thought to have arisen from the low-technology business of producing cattle feed from animal protein, specifically in the context of energy-saving, low-temperature process innovations. More significantly, Beck's favored examples—such as the risk of fall-out from the Chernobyl reactor explosion—have a primitively 'democratic' quality. On the face of it, the fall-out may descend upon the poor and wealthy alike. In that sense, the 'risk society' is everyone's problem (Beck 1986). But as the environmental justice movement, particularly in the USA, has made clear, environmental 'bads' are still often distributed quite unequally along ethnic, gender, and class lines (Bullard 1994). The risks of the risk society may not be shared as evenly as Beck implies. Finally, Beck has rather little to say about the reasons for cultural differences in the salience of different kinds of risk, although it is clear, for example, that the framing of, and policy responses to, risks vary widely even between socially and economically comparable regions, such as Europe and North America (Jasanoff 1986).

The relationship of risk and culture has been explored by the anthropologist Mary Douglas and several of her colleagues. Developing Durkheim's celebrated suggestion that religious cosmologies reflect social structures, Douglas proposed that cultural views of the characteristics of nature indicate as much about the reflected character of society as about the underlying features of nature itself. Carefree, individualistic, dynamic cultures tend to view nature as resilient and able to look after itself, while cultures which are precarious or worry about protecting their boundaries tend to view nature as fragile and in need of protection. Douglas subsequently developed this approach for the examination of risk in advanced industrial societies. A society's risk anxieties, in her view, relate as much to

the cultural 'insecurities' of that society as to the actual extent of hazards (Douglas and Wildavsky 1982). Douglas' 'cultural theory' has proven difficult to test (but see Marris et al. 1998) and it raises a host of theoretical problems, such as the unit of cultural analysis and the sources of cultural change. However, her approach does serve to underline the extent to which a culture's awareness of risks may be an imputation on to nature of more general cultural anxieties.

Finally, it is clear that not all modern risks arise from human interventions in nature. Though concerns about medical and environmental risks have increased conspicuously since the mid-1970s, worries about crime, immigration, internet security, financial markets and global social dislocation have also characterized the networked and industrialized world (see also *Science and Social Movements*). Risk anxieties can thus be inscribed not only to nature and to the human body, but also to the self-understanding of modern societies.

See also: HIV Risk Interventions; Law and Risk; Personality and Risk Taking; Probability and Chance: Philosophical Aspects; Risk: Empirical Studies on Decision and Choice; Risk, Sociological Study of; Risk: Theories of Decision and Choice; Sexual Risk Behaviors

Bibliography

Beck U 1986 *Risikogesellschaft: Auf dem Weg in eine andere Moderne*. Suhrkamp, Frankfurt am Main [Trans. 1992 as *Risk Society: Towards a New Modernity*. Sage, London]

Breyer S G 1993 *Breaking the Vicious Circle: Toward Effective Risk Regulation*. Harvard University Press, Cambridge, MA

Bullard R D 1994 *Dumping in Dixie: Race, Class, and Environmental Quality*. Westview Press, Boulder, CO

Clarke L 1989 *Acceptable Risk? Making Decisions in a Toxic Environment*. University of California Press, Berkeley, CA

Douglas M, Wildavsky A 1982 *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*. University of California Press, Berkeley, CA

Giddens A 1991 *Modernity and Self-Identity*. Polity Press, Cambridge, UK

Irwin A, Wynne B 1996 Introduction. In: Irwin A, Wynne B (eds.) *Misunderstanding Science? The Public Reconstruction of Science and Technology*. Cambridge University Press, Cambridge, UK

Jasanoff S S 1986 *Risk Management and Political Culture*. Russell Sage, New York

Jasanoff S S 1990 *The Fifth Branch: Science Advisers as Policymakers*. Harvard University Press, Cambridge, MA

Marris C, Langford I, O'Riordan T 1998 A quantitative test of the cultural theory of risk perception: Comparisons with the psychometric paradigm. *Risk Analysis* **18**: 635–48

Perrow C 1984 *Normal Accidents: Living with High Risk Technologies*. Basic Books, New York

Slovic P 1992 Perception of risk: Reflections on the psychometric paradigm. In: Krimsky S, Golding D (eds.) *Social Theories of Risk*. Praeger, London

Vaughan D 1996 *The Challenger Launch Decision: Risky Technology, Culture and Deviance at NASA*. University of Chicago Press, Chicago, IL

Winner L 1986 *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press, Chicago, IL

Wynne B 1992 Uncertainty and environmental learning. *Global Environmental Change* **2**: 111–27

Yearley S 1992 Green ambivalence about science: Legal–rational authority and the scientific legitimation of a social movement. *British Journal of Sociology* **43**: 511–32

S. Yearley

Risk: Theories of Decision and Choice

Risk is a central concept in everyday life as well as in many theories of the social and behavioral sciences. People have an intuitive notion about what is meant by risk. We are able to make risk judgments and we are able to compare the riskiness of alternatives. We might agree that it is riskier to invest in stocks than in bonds, and it might seem riskier to eat mushrooms which we have collected ourselves than mushrooms bought in a supermarket. Nevertheless, some people choose to buy stocks and others choose to eat mushrooms which they collected, i.e., they are willing to take the risk.

The examples demonstrate that risk can be analyzed as a primitive and that risk can be analyzed in the context of choice. Theories which model risk as a primitive usually start from the assumption that for each individual there exists a meaningful risk ordering of alternatives which can be obtained directly by asking the individual. Theories that consider risk in the context of preference start from choice as a primitive. A definition of risk or riskiness of an alternative is then derived based on models of choice.

Throughout Sect. 1, the leading model of choice, expected utility, is considered. Based on this model, concepts of risk and riskiness are analyzed. Section 2 investigates theories that take risk as a primitive. In Sect. 3, risk value models are presented. They model choice in a way that explicitly considers risk, i.e., they combine the ideas of Sects. 1 and 2.

Throughout the article, only theories for the risk of random variables are presented. Random variables will be denoted X, Y, etc. and they will also be called lotteries or gambles. See *Risk: Empirical Studies on Decision and Choice* for what is known about the public's judgments of the riskiness of physical, environmental, and material risk in ways that go beyond modeling these risks as random variables. An overview on risk-taking behavior can be found in Yates (1992).

1. Risk in the Context of Expected Utility

The most widely used model of preference is von Neumann and Morgenstern's (1947) expected utility model:

$$X \succcurlyeq Y \text{ if and only if } E[u(X)] \geqslant E[u(Y)]$$

where $u(.)$ is the utility function of an individual and E stands for expected value. The binary relation $\succcurlyeq$ denotes 'preferred or indifferent to.' As expected utility results in an ordering of risky alternatives, it is called a theory for decision making under risk. However, utility theory offers no direct definition of risk or riskiness.

1.1 Risk Premium

One measure of risk in the context of expected utility could be the risk premium (RP), which is defined by the following equation:

$$E[u(X)] = u(E[X] - \mathrm{RP})$$

The difference $E[X] - \mathrm{RP}$ is also called the certainty equivalent of the lottery X. For reasons of simplicity, in the following only increasing utility functions are considered. If the risk premium is negative (positive), the expected utility of the lottery is larger (smaller) than the utility of the expected value of the lottery, i.e., the decision maker does (not) like the risk of the lottery, i.e., they behave in a risk seeking (risk averse) manner. A decision-maker with a risk premium of zero is labeled risk neutral. All risk premiums for nondegenerate lotteries (lotteries where no single consequence has a probability of one) are positive (negative, zero) if and only if the decision-maker's utility function is concave (convex, linear).

Can the risk premium be used to define a risk ordering on lotteries, i.e., a binary relation $\succcurlyeq_{\mathrm{R}}$, where $X \succcurlyeq_{\mathrm{R}} Y$ means that lottery X is at least as risky as lottery Y? Unfortunately, the answer is no. The risk premium depends on the initial wealth of the decision-maker. If W and W' denote two initial wealth levels, it is quite possible that $\mathrm{RP}(W+X) > \mathrm{RP}(W+Y)$ and $\mathrm{RP}(W'+X) < \mathrm{RP}(W'+Y)$. To define risk unambiguously, one must place restrictions on the utility function, or on the random variable.

1.2 Restrictions on the Utility Function

A classic restriction on the utility function is to assume that it has a quadratic form

$$u(x) = x - \beta x^2,\ \beta > 0,\ 0 \leqslant x \leqslant 1/(2\beta)$$

where x denotes possible outcomes of the gamble X. In this case, the expected utility of a gamble X is given by

$$\begin{aligned} E[u(X)] &= E[X - \beta X^2] \\ &= E[X] - \beta E[X^2] \\ &= (E[X] - \beta E[X]^2) - \beta \mathrm{Var}\,(X) \end{aligned}$$

Thus, the expected utility is a function of expected value and variance. An increase in risk (measured by variance) implies a decrease in expected utility and vice versa. The coefficient β reflects the trade-off between risk and expected value. The risk ordering of lotteries implied by variance is independent of the decision-maker's wealth level. For further cases where variance is a reasonable measure of risk, see Levy (1992). The general question to what extent expected utility can be interpreted as a risk value model is discussed in Sect. 3.

1.3 Restrictions on Random Variables

A classic restriction on the random variable is to compare only those that have identical expected values. For this case, Rothschild and Stiglitz (1970) were able to provide a definition of the risk of gambles. For $E[X] = E[Y]$, we can say that X is more risky than Y if and only if the following three statements, which are equivalent, are true

(a) Every risk-averse decision maker prefers Y to X.

(b) X is equal to gamble Y plus noise, i.e., some random variable with expected value zero.

(c) X is obtained by a mean preserving spread of Y. A mean preserving spread takes probability weight from central parts of the distribution and puts it into the tails of the distribution while holding the expected value constant.

This definition of riskiness is convincing, but it is only applicable to lotteries with equal expected value, and thus yields a partial ordering of the space of gambles. To overcome this limitation, Jia and Dyer (1996) define a risk ordering on lotteries for concave utility functions u as follows

$$X \succcurlyeq_{\mathrm{R}} Y \text{ if and only if } -E[u(X - E[X])] \geqslant -E[u(Y - E[Y])]$$

This definition of riskiness is able to order all lotteries. However, it is not clear to what extent this definition captures what people intuitively think about risk and how an ordering in terms of risk is related to an ordering in terms of preference. We will further investigate these questions in Sect. 3.

1.4 Relative Risk Aversion

All the results in this section are based on expected utility theory. However, it should be pointed out that utility functions not only measure decision-makers' risk attitude, simultaneously they measure the marginal value of the good to be evaluated. An example should illustrate this point.

Suppose one has to evaluate a lottery which gives eight oranges if on a coin flip head comes up and zero oranges if tail comes up. Assume that the certainty equivalent is three oranges, i.e., the risk premium is one orange and thus the decision-maker is considered risk averse. Now, it could well be that this positive risk premium has nothing to do with any 'intuitive' notion of risk aversion. For instance, if the decision-maker likes three oranges half as much as eight oranges, then the average value of the lottery is equal to the value of three oranges, i.e., the risk premium can equally well be explained by some strength of preference considerations. The notion of relative risk aversion, introduced by Dyer and Sarin (1982), helps to disentangle strength of preference from utility. The concept of relative risk aversion defines risk aversion relative to the strength of preference. The above example shows a decision-maker who is relatively risk neutral.

2. Risk as a Primitive

In this section we start out with the assumption that subjects are able to meaningfully order lotteries with respect to their riskiness, i.e., that a risk ordering $X \succcurlyeq_{\mathrm{R}} Y$ exists. In *Risk: Empirical Studies on Decision and Choice* it is demonstrated that individuals are able to give consistent risk judgments. In this section we investigate how these judgments can be modeled. One has to look for a function R defined on the set of gambles with the property

$$X \succcurlyeq_{\mathrm{R}} Y \text{ if and only if } R(X) \quad R(Y)$$

Every such function R will be called a risk measure. Note that we take subjects' risk judgments (risk perception) as a primitive without giving an abstract definition of risk and without relating it to preference (see Brachinger and Weber 1997 for an overview).

2.1 Naive Measures of Risk

The first class of risk measures to be discussed are so-called naive risk measures. Most of them are special cases of the following formula

$$R(X) = E[\,|X - x_0|^{\alpha} \mid X \leqslant x']$$

where x_0 is a benchmark (reference level of wealth) defining gains and losses, x' defines the range parameter that specifies what deviations are to be included in the risk measure, and α is a measure of relative impact of large and small deviations. Risk measures such as variance, semivariance, shortfall measures, probability of ruin (loss), etc., are all special cases of this risk measure. Some of these measures are quite popular, e.g., shortfall measures in the insurance and banking industry; nevertheless, they do not capture what individuals perceive as risk (see *Risk: Empirical Studies on Decision and Choice*).

The fact that risk perception is not captured by variance, the measure most popular in economics and finance, can be demonstrated by a simple example. Consider the three two-outcome gambles $X = (10, 0.5; -10, 0.5)$—this is a gamble with a 50–50 chance of receiving 10 or $-10-$, $Y = (20, 0.2; -5, 0.8)$, and $Z = (5, 0.8; -20, 0.2)$. Empirical studies find that subjects consistently judge each of Z and Y to be more risky than X (Coombs and Bowen 1971), but expected value and variance are the same for all three lotteries. Higher moments of the distribution, e.g., skewness, have an influence on risk perception too.

2.2 Axiomatic Measures of Risk: I

For a while it was quite popular to axiomatically derive measures of risk. Similar to what is done in modeling preferences, simple axioms of risk judgments were presented and measures of risk were derived mathematically. Pollatsek and Tversky (1970) were the first to use the axiomatic approach. Besides some technical axioms, they require a risk measure to be additive, i.e., the risk of the convolution of two lotteries should be equal to the sum of the risk of each lottery. This property implies, e.g., that the risk of a lottery played twice is equal to twice the risk of the lottery played once, and the risk ordering of two lotteries is not changed if they are both convoluted with the same lottery. Finally, they require that the risk of a lottery is reduced if a constant is added to all outcomes. From these conditions they derive the following risk measure

$$R(X) = \alpha \mathrm{Var}(X) - (1-\alpha) E[X], \qquad \alpha \in (0, 1]$$

Thus, according to Pollatsek and Tversky (1970), the risk ordering is generated by a linear combination of expected value and variance. This model does not fit risk judgments since, for instance, the example given at the end of Sect. 2.1 demonstrates that subjects' risk perception can not be explained just by expected value and variance.

Fishburn (1982, 1984) explored risk measurement from a purely axiomatic point of view. He considered gambles separated into gains and losses and did not specify functional forms, thus his theory remains rather general and is not considered further here.

2.3 Axiomatic Measures of Risk: II

Luce (1980) was the first to derive risk measures from functional equations characterizing the effect of resca-

ling on perceived risk. His most prominent models consider the α-transformation X_α in which each outcome of the gamble X is multiplied by $\alpha > 0$. He assumes that there exists an increasing function S (with $S(1) = 1$), such that $R(X_\alpha) = R(\alpha X) = R(X)\, S(\alpha)$ for all X and α. He also assumes that the risk of a gamble X is the expectation of a transformation T of that gamble, i.e.,

$$R(X) = E[T(X)], \text{ with } T \text{ some non-negative function with } T(0) = 0$$

These assumptions imply the following risk measure

$$R(X) = b_1\, E[\,|X|^\delta \mid X < 0]\ \text{Prob}\,(X < 0) + b_2\, E[X^\delta \mid X > 0]\ \text{Prob}\,(X > 0)$$

where δ is a positive real number, b_i are weights on the components, and Prob $(X > 0)$ denotes the probability that the random variable has outcomes greater than zero. This risk measure is a linear combination of the conditional expectation of negative and positive outcomes, where all outcomes are raised to the power δ.

Luce's model was extended by Luce and Weber (1986). Influenced by empirical work and based on a system of axioms, they proposed the Conjoint Expected Risk (CER) model that has the following form

$$\begin{aligned} R(X) = {} & a_1\ \text{Prob}\,(X = 0) + a_2\ \text{Prob}\,(X < 0) \\ & + a_3\ \text{Prob}\,(X > 0) \\ & + b_1\, E[|X|^{\delta^-} \mid X < 0]\ \text{Prob}\,(X < 0) \\ & + b_2\, E[X^{\delta^+} \mid X > 0]\ \text{Prob}\,(X > 0) \end{aligned}$$

with a_i and b_i being weights on the components, and δ^- and δ^+ being coefficients of power functions. Empirical research has demonstrated that the CER model describes risk judgments reasonably well, admittedly at the cost of a rather large number of parameters in the model.

Sarin (1987) chose another way to extend Luce's measure of risk. Instead of the α-transformation, he used a β-transformation, which just adds a factor β to all outcomes of the lottery. He then assumes that $R(X_\beta) = R(X+\beta) = R(X)\ S(\beta)$, with S a strictly decreasing function, $S(0) = 1$ and $\beta > 0$. This assumption mirrors the fact that the risk of a gamble appears to decrease when a β-transformation is applied. In conjunction with the expectation principle (see Luce's approach), Sarin derived the following risk measure

$$R(X) = k\, E[e^{cX}]$$

with $k > 0$, $c < 0$ or $k < 0$ and $c > 0$. This risk measure gives higher weight to a gamble's potential losses than to its potential gains. Both Sarin's model and the CER model are able to capture the effect that subjects strongly dislike the possibility of large losses (potential catastrophes).

All the risk measures presented in Sects. 2.2 and 2.3 have the property that adding a constant (applying the β-transformation to the gamble) changes the risk, i.e., they are not location free. This property is supported by a lot of empirical work. However, on theoretical grounds, in quite a number of applications, e.g., in capital market theory, risk should be location free. Since nearly all the models proposed in Sect. 2.1 are location free, for certain applications it might be useful to combine the ideas of Sects. 2.2 and 2.3 with the concept of a location free risk measure. Weber (1990), therefore, advocates extending Sarin's measure of risk to the following form, giving the risk measure some normative flavor

$$R(X) = k\, E[e^{c(X - E[X])}]$$

with $k > 0$, $c < 0$, which can be used in a capital market framework.

Recently, Jia et al. (1999) modified the model of Jia and Dyer (1996) presented in Sect. 1.3 to make it descriptively appealing. In Sect. 1.3, risk was defined as $R^*(X) = -E[u(X - E[X])]$ which is location free. To heal this deficiency, they propose the following measure of risk

$$R(X) = f(E[X])\ (R^*(X) - R^*(0))$$

where $f(E[X])$ is a decreasing function of the expected value and $R^*(0) = -u(0)$ a constant. Thus, an increase in expected value leads to a decrease of risk—as in Sarin's model. Their model is rather general as it only specifies that $R^*(X)$ depends on the decision-maker's utility function, but no specific form of the utility function is required.

3. *Risk Value Models*

Clearly, risk is related to preference (see also Sect. 1). To further investigate this relation, one could propose a general risk value model for evaluating gambles such that a gamble X is preferred to a gamble Y, i.e.,

$$X \succcurlyeq Y \text{ if and only if } f(V(X),\ R(X)) \geqslant f(V(Y),\ R(Y))$$

where V measures the value of a gamble, R measures the risk, and f reflects the trade-off between these two components (see Sarin and Weber 1993 for a more extensive review). Clearly, a greater value is more preferred and a lower risk is more preferred. Therefore, f is increasing in V and decreasing in R.

Risk value models have an intuitive appeal, as people like to talk about the risk and the value of alternatives. There are two ways to use risk value models:

(a) Starting from expected utility, the EU model can be reformulated as a risk value model.

(b) A measure of risk as a primitive and the expected value can be the components of a risk value model.

Both ways still have their disadvantages. If one starts with expected utility (case (a)), in most cases it is not possible to find a straightforward decomposition into risk and value. It was shown in Sect. 1 that in

order to do this one has to considerably restrict the utility function or the probability distribution. In addition, the risk measures derived from expected utility might not be descriptively correct. If one begins with a risk measure as a basis for a risk value model (case (b)), one has to first define a trade-off function *f*. It is by no means clear that the resulting risk value model will be prescriptively appealing. Future research hopefully will shed more light onto the relation between preference and risk.

See also: Bayesian Theory: History of Applications; Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision Research: Behavioral; Risk: Empirical Studies on Decision and Choice

Bibliography

Brachinger H W, Weber M 1997 Risk as a primitive: A survey of measures of perceived risk. *OR Spektrum* **19**: 235–50

Coombs C H, Bowen J N 1971 A test of VE-theories of risk and the effect of the central limit theorem. *Acta Psychologica* **35**: 15–28

Dyer J S, Sarin R K 1982 Relative risk aversion. *Management Science* **28**: 875–86

Fishburn P C 1982 Foundations of risk measurement. II. Effects of gains on risk. *Journal of Mathematical Psychology* **25**: 226–42

Fishburn P C 1984 Foundations of risk measurement. I. Risk as a probable loss. *Management Science* **30**: 396–406

Jia J, Dyer J S 1996 A standard measure of risk and risk-value models. *Management Science* **42**: 1691–1705

Jia J, Dyer J S, Butler J C 1999 Measures of perceived risk. *Management Science* **45**: 519–32

Levy H 1992 Stochastic dominance and expected utility: Survey and analysis. *Management Science* **38**: 555–93

Luce R D 1980 Several possible measures of risk. *Theory and Decision* **12**: 217–28

Luce R D, Weber E U 1986 An axiomatic theory of conjoint, expected risk. *Journal of Mathematical Psychology* **30**: 188–205

Pollatsek A, Tversky A 1970 A theory of risk. *Journal of Mathematical Psychology* **7**: 540–53

Rothschild M, Stiglitz J E 1970 Increasing risk: I. A definition. *Journal of Economic Theory* **2**: 225–43

Sarin R K 1987 Some extensions of Luce's measures of risk. *Theory and Decision* **22**: 125–41

Sarin R K, Weber M 1993 Risk-value models. *European Journal of Operational Research* **70**: 135–49

von Neumann J, Morgenstern O 1947 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

Weber M 1990 *Risikoentscheidungskalküle in der Finanzierungstheorie*. C. E. Poeschel, Stuttgart, Germany

Yates J F (ed.) 1992 *Risk-taking Behavior*. Wiley, Chichester, UK

M. Weber

Ritual

For Durkheim (1964), ritual is social activity falling within the domain of the sacred. The sacred is opposed to the profane, which does not contain ritual action. Leach (1954), however, rejects Durkheim's hard and fast distinction between sacred and profane, preferring to see these as aspects of social actions lying along a continuum. At one end are those actions that are largely technical and profane, while at the other are those which have no technical functions and are largely symbolic, aesthetic, and sacred. Most actions in between these two poles have aspects which are both sacred and profane. Ritual in Leach's sense is, then, by no means confined to ideas of the sacred. By removing ritual as exclusively to do with the sacred, Leach is able to broaden the definition of ritual. He thinks of it as a message-making kind of activity, and so as an aspect of any kind of behavior in which repetitive or customary acts signal messages. Handshakes or wearing a tie in the office are as much forms of ritual as a priest praying to ancestors at a shrine. So-called magical and secular acts are therefore both communicative and both are examples of ritual action.

Ritual for Leach is, then, a way in which humans communicate with each other different aspects of their social structure. Thus, we may be perfectly aware of social class or status differences and even be able to talk about them discursively. But we still signal or symbolize these differences in a number of short-cut ways, through distinctive accents, differences of dress and life-style, and even bodily movement and gesture. Sometimes we are not so conscious of the differences our ritual behavior communicates, and sometimes the differences exist simply through lack of knowledge of how to behave on, say, a formal occasion. In other words, the sociostructural approach argues that ritual action does not occur in a political and economic vacuum: politicoeconomic interests are always involved to some degree, even if sometimes minimally.

A second approach to the study of ritual emphasizes its logical qualities. Here, ritual communicates fundamental ideas that may be specific to a particular culture, or that may even be universal. It is an approach that owes much to structuralism in anthropology, which is predominantly associated with the work of Lévi-Strauss. It runs very much from the idea of culture as organized through binary oppositions. Thus, Needham (1961) has suggested that, when a male diviner or shaman uses his left hand to throw the cowrie shells which he or his possessory spirit will interpret in order to understand the cause and course of an illness or misfortune, he is reversing a kind of natural order. For men are normally associated with the right hand and women with the left. But since diviners fall outside the realm of normal human capacities, being able to see what other humans cannot, it makes logical sense to reverse the equation

of maleness with the right hand. There is no functional explanation for these kinds of ritual reversal for they are governed by proclivities of the human mind. That said, local political and economic circumstances can sometimes take up these proclivities and harness them to a cause. Thus, prophets leading religious movements may communicate their power and distinctiveness through rituals which reverse normal expectations. Many other kinds of ritual behavior, like some beliefs and myths, achieve their communicative effectiveness by reversing what is otherwise considered natural.

Now, while it is true that ritual may be thought of as communicating messages about status or about the way the human mind works or how the world is, it varies enormously in the degree to which it is elaborated. After all, a handshake can hardly compare with a full-scale ritual investiture of a monarch spread over days and involving thousands of people and lavish expenditure. To call both these kinds of activity 'ritual' clearly leaves out much of what is special to the latter. Similarly, some hunting and gathering peoples such as the African Mbuti forest dwellers do not perform elaborate rituals, while other African groups noted for their masquerades do. Peoples do, then, seem to differ in the relative amounts of ritual which they perform; although, in making this claim, there is a dangerous Eurocentric qualitative assumption that the significance of ritual is measured by the amount of time, materials, and people involved. Who is to say that an apparently slight activity, taking only a moment and few resources, may not in fact 'say' much more than a drawn-out ceremony of great proportions whose influence on people is in fact limited? Nevertheless, it is a useful if provisional starting point of analysis to try and understand how variations in the complexity and lavishness of ritual reflect those of social organization and cultural logic. Arising from this approach is another, which recognizes that as well as communicating messages about the current social structure or about fixed properties of the human mind, rituals also change things or enable changes to take place.

We may call this third, the transformational approach to ritual. For Van Gennep (1965), the rites of passage as marking changes of status in the human life cycle—for example, birth, puberty or adolescence, marriage, first child, first grandchild, and death—could be broken down into three phases: separation, liminality, and incorporation. In other words, the individual undergoing the ritual must first be separated from the community of which they are a member, perhaps by removal of the head hair or some other bodily alteration. They then enter the liminal phase in which they are regarded by members of the community as outside and beyond normal human society. This is the phase in which there is transition from one status to another, for example from child to adult, bachelor to husband, etc. Van Gennep saw this as a period of death, i.e., of the child, and of rebirth, i.e., of the adult. It is a theme elaborated in many later studies that showed the ritualization of death as the condition of new life or regeneration. The third and final phase is when the individual is reincorporated in the community, now no longer a child, but an adult with the rights and duties expected of this new status.

Van Gennep saw these features of rites of passage as also typifying other kinds of transition in addition to those of the life cycle, such as the seasonal or other movements of groups, fertility rites, and entering and leaving dwellings, special days or months, and cult societies. Victor Turner (1967) made the notion of liminality central to his work on ritual, extending to the idea of antistructure, by emphasizing the rejection of society on the part of people undergoing transition or in a marginal state, and also communitas, by which he meant the way in which people on ritual occasions temporarily transcended their normal, everyday social conflicts and cleavages and emphasized their common humanity. In addition to the life crisis rites of Van Gennep, Turner further distinguishes rites of affliction, which are held to reverse the effects of, for example, women's infertility, crop failure, sickness, and other misfortunes adversely affecting the lives of people living together in a single village or other dwelling. They also enable people to move from one state to another, although not as part of the human life cycle, but more as part of the unpredictability of the human and social condition in which conflicts and contradictions are not so much represented in ritual as presented, addressed, and then redressed, an approach which thus builds on and advances that of Durkheim.

A subset of this approach might conveniently be called Marxist in the sense that social inequality and exploitation of a working by a ruling class is presupposed. A Marxist approach would, for instance, claim that the official, media, and public ritual attention devoted to the British royal family evokes ordinary people's unquestioning compliance and loyalty, to the extent that they cannot see that the institution of royalty in fact underpins the system of landed privilege that defines an aristocracy and provides an imitative model for other ruling and bourgeois classes. The 1990s in the UK saw an extraordinary rise in the degree to which the media and many of these same ordinary citizens began to question the worthwhileness of royalty. Was this an example of royal rituals losing their capacity to retain subjects' loyalty, a kind of demystification of ritual? Or is this a case of the different interpretations of ritual enabling social change to occur?

Royal rituals have been a common focus of anthropological analysis, in which their transformative power has been highlighted. Gluckman (1963) was dissatisfied with studies of royal and other rituals which saw them as simply affirming the sovereign's authority and the solidarity of the society under his rule. Using the example of the Swazi *Incwala* royal

ritual in southern Africa, Gluckman claimed instead that its socially cohesive function was only possible after it had revealed the society's internal conflicts: such conflicts had to be addressed before they could be accepted as part of the normal social process. Beidelman (1966) criticizes Gluckman's over-reliance on social conflict as needing to be expunged through royal rituals, and argues for a view of them as marking out the divine or mystical properties of kingship, so providing the king with legitimate recognition to reign over his subjects.

This example moves us on to a fourth approach to ritual, which is to see them not necessarily as achieving a particular social or logical function, but as expressive in and of themselves. Beattie (1966) attacked attempts to explain rituals as practical and intellectual attempts to solve problems. Rituals draw instead on people's human predilection to engage in creative thinking, sometimes through culturally enjoined associations of ideas, and were thus closer to art and drama than to technology or science. Rituals in this sense do not 'do' things so much as express human dilemmas, symbolically and often under some kind of duress. They cannot be reduced to questions of whether they are rational or irrational, or true or false propositions, any more than aesthetics can be explained in these terms.

In many respects this explanation of ritual as expressive rather than functional, reproduces Leach's distinction between technical and aesthetic acts as occupying two extreme poles of a continuum between which they are both aspects of any activity. Nevertheless, expressed in the way that Beattie did, it freed anthropologists from having to seek practical explanations for a particular form of ritual behavior. If a ritual appeared to perform a particular function in society, this could better be regarded as an unintended consequence of an activity whose participants did not have such consequences in mind. Thus, it may well be that the British royal rituals for a long time had the effect of persuading the public to accept the stratified status quo as rightfully headed and legitimized by the monarch and the monarch's family. It may well be, moreover, that politicians and members of the ruling elite recognized, seized on, and developed this effect, so turning consequence into purpose. The expressive here becomes instrumentally effective. However, not all that which we understand as ritual behavior can be interpreted in this way. People throughout time do not normally calculate in advance what might pass as effective ritual, nor indeed separate ritual as a social category from nonritual; nor may we assume that what we define as rituals always have significant social and intellectual consequences.

Approaches to the study of rituals are not exhausted by the above, and indeed overlap in part. They represent, as much as anything, the discipline's decision to divide up its findings in this way. An alternative method is to ask how anthropologists define ritual on the basis of their own ethnographic data. A recent example is a study of the Indian Jain rite of worship by Humphrey and Laidlaw (1994). Their intriguing starting point is the fact that in some societies people may distrust ritual practices as being empty or false, and yet continue with them. Similarly, noting that for many people, religion and ritual are seen as going together, they also refer to the way in which practitioners of some religions turn on their rituals as unworthy and even frightening. The European Reformation attempted to strip off from Christian religious belief what were regarded as some unacceptable ritual practices, as Wahabi reformers in Islam are doing today. By contrast, some argue that ritual is good for religion and humanity. Mary Douglas (1973) drew some of her own ideas from trying to understand the rejection of ritual in the Catholic Church by many of its senior adherents who also, for good measure, dispensed with the Latin language as if transparent clarity was all that mattered, over and above the expressive power of faith.

The disjunction between religion and ritual led Humphrey and Laidlaw (1994) to suggest another disjunction, which is that between the intention of actors in ritual and the ritual actions themselves. Sometimes intention and act fit, and sometimes they diverge irretrievably, but sometimes they are aligned. Such lack of fit is not to be understood as the result of mistakes in intention and performance, but as evidence of the way in which ordinary actions may become endowed with the quality of ritual. For these authors, ritual is a quality that actions take on: they become ritualized. While everyday actions are intrinsically intentional, ritual acts enter a kind of phenomenological space freed of such everyday intentions, and, paradoxically, are 'intended' to do so. There is here a 'ritual commitment,' a notion close to that expressed in a classical study of ritual by Gilbert Lewis (1980). Humphrey and Laidlaw (1994), moreover, oppose the idea that ritual is necessarily communicative and expressive of messages and beliefs. One is here reminded of Sperber's (1975) view that symbols do not intrinsically mean anything. Rather it is people who interpret a symbol as such, and who thereby give it meaning. The concept of communication here has to be established. It may mean one person making an assertion and expecting a response from the hearer. Or it may simply mean making a gesture or uttering a conventional phrase ('It's hot today'), for which no response is expected and which is therefore really a form of what has been called phatic communion, a device to secure solidarity with another person. Rituals may sometimes communicate but, as Rappaport (1979) points out, so do many other kinds of phenomenon, and by privileging its expressive functions we miss what is otherwise distinctive of ritual.

The view that ritualization is a process by which actions become endowed with a ritual quality, shares something with Gerholm's (1988) notion that ritual is, so to speak, partly made up as it proceeds. That is to

say, there are at least some kinds of ritual for which precise textual rules do not exist or which have been forgotten, with the consequence that participants have little more than a sense of ritual occasion or commitment and endeavor to convey that sense in actions and words, inventing and borrowing as much as they can. This ushers in a query as to how much ritual depends on precise wording as well as actions. A Christian prayer depends heavily on personally composed as well as prescribed verbal forms, and is firmly regarded as ritual. But rituals are also organized through spatial orientation and bodily positioning. The many different kinds of Islamic prayer are also personally or formulaically verbal, but crucially require orientation to Mecca, while a particular class of prayer or devotion called *salat* is made up of a complex sequence of prostrations as well as recitations (Bowen 1989). In the event of the mosque imam taking on the recitations on behalf of the congregation, the prescribed bodily actions become the individual worshipper's ritual responsibility and must be repeated if they are performed incorrectly. Parkin (1992) has argued that it is not always the case that the verbal is important in ritual, while set spatial directions and orientations, and bodily gestures, movement, and positions are what mark off and so fundamentally characterize ritual.

A challenge is to avoid essentialist definitions of ritual which seek to set up criteria which must be met before this label can be applied to an activity. Phenomenologically, ritual is to do with the paradox of being both the agent of an action, but also being taken along by this same action. It presupposes being able to move within an action from citing its purpose to having no need, wish, or clear idea of what it is about, and yet valuing it. Nevertheless, ritual must have some kind of structure, the most elemental of which is that individuals alone or with others take up positions with regard to each other or with regard to objects, experiencing such positioning as unlike any other.

See also: Control: Social; Durkheim, Emile (1858–1917); Magic, Anthropology of; Myth in Religion; Religion, Sociology of; Ritual and Symbolism, Archaeology of; Taboo; Totemism

Bibliography

Beattie J 1966 Ritual and social change. *Man*: *The Journal of the Royal Anthropological Institute* **1**: 60–74

Beidelman T O 1966 Swazi royal ritual. *Africa* **36**: 373–405

Bowen J R 1989 *Salat* in Indonesia: The social meanings on an Islamic ritual. *Man*: *The Journal of the Royal Anthropological Institute* **24**: 600–19

Douglas M 1973 *Natural Symbols: Explorations in Cosmology*. Barrie and Jenkins, London

Durkheim E 1964 *The Elementary Forms of the Religious Life*. Allen and Unwin, London

Gerholm T 1988 On ritual: A postmodernist view. *Ethnos* **53**: 190–203

Gluckman M 1963 Rituals of rebellion in south east Africa. In: Gluckman M (ed.) *Order and Rebellion in Tribal Africa*. Cohen and West, London

Humphrey C, Laidlaw J 1994 *The Archetypal Actions of Ritual*. Oxford University Press, Oxford, UK

Leach E 1954 *Political Systems of Highland Burma: A study of Kachin Social Structure*. G. Bell, London

Lewis G 1980 *Day on Shining Red: An Essay in Understanding Ritual*. Cambridge University Press, Cambridge, UK

Needham R (ed.) 1961 The left hand of the Mugwe: An analytic note on the structure of Meru symbolism. *Africa* **31**: 28–33

Parkin D 1992 Ritual as spatial direction and bodily division. In: De Coppet D. (ed.) *Understanding Rituals*. Routledge, London

Rappaport R A 1979 *Ecology, Meaning, and Religion*. North Atlantic Books, Berkeley, CA

Sperber D 1975 *Rethinking Symbolism*. Cambridge University Press, Cambridge, UK

Turner V W 1967 *The Forest of Symbols: Aspects of Ndembu Ritual*. Cornell University Press, Ithaca, NY

Van Gennep A 1965 *The Rites of Passage*. Routledge and Kegan Paul, London

D. Parkin

Ritual and Symbolism, Archaeology of

Ritual (repetitive sequences of actions related to beliefs) and symbolism (the representation of one thing, often more abstract, by reference to another, often more concrete) have been subjects of archaeological study since the origin of the discipline. Individual artifacts, locations within archaeological sites, and even entire sites or landscapes have been used as archaeological evidence of ritual and symbolism. Archaeologists have varied in their assessment of how difficult it might be to study these subjects, with some expressing doubt that accurate information about such nonmaterial topics could be produced, and others noting that symbols are polysemous, requiring multiple interpretations. Contextual analysis and an emphasis on the active role of social agents characterize the contemporary archaeology of ritual and symbolism.

1. From Ritual Objects to Symbolic Archaeology

Archaeologists have always identified some deposits as evidence of ritual, and ascribed symbolic meanings to some artifacts. Where texts were available, as in Egyptian archaeology, they were employed to validate early interpretations. Beginning in the 1950s explicit concern was addressed to the methodological status of

archaeological inferences concerning belief and meaning. Interpretation of ideational systems, including symbolism and ritual, was viewed as more problematic than other topics, such as economics and subsistence. As early as the 1960s, some archaeologists argued that ritual and symbolism were no more difficult of access than more apparently materialist topics, questioning the analytical separation of material and ideational aspects of society. Since the 1980s, archaeologists influenced by practice theory have attempted to demonstrate that the archaeological analysis of symbolism and ritual benefits from attention to human agency. Despite changing theoretical perspectives, throughout the history of professional archaeology similar empirical materials have served as the basis for the archaeology of ritual and symbolism.

1.1 Early Archaeological Research on Ritual and Symbolism

Early fieldwork recovered material interpreted as evidence of ancient ritual and belief. Contemporary texts were sometimes available that sketched out beliefs and, as in the case of Egypt, even provided details of ritual actions.

Such early research established assumptions underlying much subsequent archaeological investigation of symbolism and ritual. First, the symbolism of material remains ideally was to be explained by reference to explicit texts, assigning material culture to the role of a container of a formalized and generally accepted singular meaning. Second, specific kinds of archaeological deposits came to occupy a central place in discussions of ritual. These included formal structures identified as the locations of rituals (temples); human burials, viewed universally as occasions for mortuary ceremonies; and 'offerings': deposits of selected objects, especially those encountered in ritual structures and burials.

1.2 The 'Ladder of Inference' and Processual Archaeology

In 1954, Christopher Hawkes proposed that in the absence of texts, archaeologists were constrained by a 'ladder of inference' which governed the probable reliability of their explanations (Hawkes 1954). According to the ladder of inference, some topics were more closely tied to material manifestations, which were more tightly circumscribed by pragmatic considerations. As a result, archaeologists were likely to recover a greater proportion of the evidence necessary for understanding these topics, and their explanations were likely to suffer less from crucial missing data or from variability in the nature of the evidence. Areas such as religion and belief were less tightly tied to materiality, and thus archaeologists were likely to lack crucial pieces of evidence to support explanation, or they might recover material that could be explained in multiple ways and have no basis to choose between explanations. In contrast, in initial programmatic statements for the processual archaeology of the early 1960s, Lewis Binford explicitly argued that all aspects of culture were open to scientific archaeological analysis if questions were properly framed. Neither symbolism nor ritual was excluded from archaeological investigation.

Important contributions to the archaeology of ritual and symbolism were developed in processual archaeology. Drawing on information theory, symbols could be treated as messages intended to communicate clear meanings (Wobst 1977). Information theory provided a means to support some of the assumptions necessary for archaeologists to treat ancient symbols as if they had a single consistent meaning. If symbols were media to convey messages, then there would have been strong social pressures, and even adaptive advantages, to having shared interpretations of their meaning. The application of communication models to the archaeological study of symbols facilitated the development of an archaeological theory of style, the definition and interpretation of sociopolitical 'status badges,' and the interpretation of differences in material culture as markers of ethnic boundaries and cultural identities.

Robert Drennan (1983), employing processual systems theory, developed an explicit set of expectations for ritual subsystems, which could be linked to specific characteristics of archaeological remains to create predictions for the material correlates of ritual. Following social anthropologist Roy Rappaport, Drennan operationally defined ritual as specific sequences of repeated action directed at intervening in relations between humans and supernatural forces or beings. Among the key features he and others used to attempt to objectively identify remains stemming from ritual were associations between otherwise rare items, and the recovery of such rare items from public spaces. The systems approach assumed a separation between an everyday sphere of activities and a separate sphere, the domain of ritual, detectable by distinctive location and unusual artifacts.

Processual approaches required external bodies of information for the formulation of models of ritual and symbolic systems. Specific ethnographic analogy, the 'direct historic approach,' drew on the study of living societies believed to be descended from the archaeological population. General analogies were made from comparative ethnographic samples such as those contained in the Human Relations Area Files. (These were abstracts of ethnographic samples selected to provide a cross-section of known variation in human activity. They provided a beginning point for statistical generalizations about human behavior, including ritual.) The power of explanation rested on

the strength of arguments linking the analogies to the test case, requiring extremely broad uniformitarian generalization. The level of generalization required made explanations of ritual and symbolism employed extremely vague, and tended to substitute the operations of superorganic entities for individual actions.

1.3 Postprocessual Directions in the Archaeology of Ritual and Symbolism

In a series of edited volumes, Ian Hodder (1982a, 1982b, 1989) and colleagues initiated a range of challenges to the processual analysis of ritual and symbolism. Critiques of the information theory models embedded in processual analyses were central to these analyses. Demonstrations of diversity *within* cultural traditions, based on historical and ethnographic documents, highlighted problems with uniformitarian and normative assumptions shared by processual and preprocessual archaeologies.

While there is variation among scholars participating in the postprocessual critique, they share a fundamental view of the nature of symbolic meaning. Rather than seeing archaeologically recovered objects as vehicles carrying pre-existing meanings, they argue that symbolic meaning was negotiated through the use of cultural forms, including the permanent ones recovered by archaeologists. This view of symbolism requires attention to contextual relations in order to understand how symbolic meaning was negotiated. Even when archaeologists operating under this model express uncertainty about their ability to understand the specific meanings symbols had for past actors, they routinely make arguments about the deployment of symbolism as part of strategies of actors and factions within past societies.

The meaning of symbols and their social effects are seen as mutually constituted. Contrasts within and between contexts are required to argue for a relationship between symbolism and any specific form of meaning under these assumptions. The constitution of symbolic meaning is seen as a constant feature of social life. Rituals provided marked settings in which the meanings of symbols were negotiated, but they were not the only or even the most significant settings in which symbolic meaning developed.

2. Archaeology of Symbolism

Relations of substitution are fundamental to the definition of symbols. Archaeologists commonly use specific objects as signs of categories, for example, the identification of Clovis stone points as markers of a Palaeolithic population in the Americas (see *Ethnic Identity/Ethnicity and Archaeology*). This basic archaeological operation rests on the assumption that the makers and users of these points would have recognized a commonality among themselves and a dissimilarity from others, albeit on a largely nondiscursive level (see Gardin and Peebles 1992). Ethnoarchaeological research has demonstrated that the degree to which the use of common material forms conveys either a conscious or unconscious identification is highly variable (Hodder 1982a).

Some symbolic relations of substitution by past populations are treated as more conscious (see *Prehistoric Art*).The embedding of symbolism in concrete things has been a focus of extensive empirical and theoretical debate (e.g., Robb 1999). For example, some artifacts have been interpreted as 'status badges' created as deliberate signs of rank. Some archaeologists seek to understand the specific properties of materials which lend themselves to being used in symbolic transactions, or to reinforce culturally accepted meanings. Others are concerned with the way that symbolic meanings are interpreted. The latter approach includes a consideration of the experiential contexts within which objects with symbolic significance were used, including rituals.

3. Archaeology of Ritual

The archaeology of ritual has moved away from an early history of identification of specific material objects, settings, and representations as evidence of the presence of a separate ritual or symbolic sphere or system, to the use of contextual associations to identify repeated sequences of action like those that are recognized as ritual in living societies. Certain classes of artifacts have been described as 'ritual' objects: examples include ancient Chinese bronze vessels, North American copper and shell objects, and ceramic figurines in many areas of the world. The identification of objects as ritual objects has been criticized as a facile means to defer our understanding of unusual objects. Recently, Ruth Whitehouse (in Wilkins 1996) has argued that 'ritual objects' can be a real category of material with unique potential for informing about ancient belief and action.

What distinguishes ritual contexts in recent work is not simply their separation from presumed household locations. Historical archaeologists specializing in the study of the African diaspora have recovered assemblages of artifacts arguably used in healing rituals on the household level. In place of defining specific objects and places as inherently associated with ritual, contemporary approaches emphasize the repetition of activity that is central to the definition of ritual. Archaeologists working on the British Neolithic proposed the concept of 'structured deposition' as a way to recognize specific archaeological deposits that were likely to be the result of ritual action (Richards and Thomas 1984).

4. The Material Base for an Archaeology of Ritual and Symbolism

Remarkably, despite theoretical differences, the specific material bases for an archaeology of ritual and symbolism have remained fairly stable and have been constantly under investigation. Offerings or caches, burials, particular spatial locales, and visual images have consistently provided the subject matter for archaeological analyses.

4.1 Caches and Offerings

Caches, hoards, or 'offering deposits' figure as primary evidence of ritual in several regional traditions in archaeology, particularly North America, Central America, the eastern Mediterranean, and Neolithic Europe (e.g. Bradley 1991). Interpretations of caches vary from purely materialistic to ideational. Some analysts stress the economic effect of the withdrawal from circulation of a large quantity of specialized goods at one time. Others propose structural-symbolic interpretations in which materials employed worked to create specific associations between ideas about the supernatural world and the locale of the offering, consecrated by the deposit. Explanations of both kinds have been offered for the same deposits, and indeed, they are not mutually exclusive, since a ritual action can have both pragmatic and symbolic effects.

4.2 Burials

Mortuary analysis, the study of sites and artifacts related to human burial, has occupied a central place in the archaeology of ritual and symbolism (e.g., Chapman et al. 1981). Studies conducted under the normative assumptions of culture history *described* burial customs, including the laying out of the body according to symbolic beliefs (such as the eastward orientation of the head), and the inclusion of artifacts as 'grave goods.' Among mortuary activities, the removal of skeletal elements and the practice of secondary burial have been seen as unambiguous evidence of ritual practice. The deliberate burial of animals has generally also been regarded as evidence of ritual activity.

Processual archaeology directed attention to the generation of covering laws or assumptions that could link the forms of burials, and particularly, diversity among them, to social differences. Burials were seen as reflecting the social identity of the deceased. They operated as mechanisms to assert claims for territory. Postprocessual archaeologists examine burial assemblages as the material media of ancient rituals, and demonstrate that mortuary remains can mask social relations as easily as they can reflect them (e.g., contributions to Hodder 1982b).

4.3 Temples and 'Sacred Landscapes'

The identification of temple sites through spatial segregation, specific construction, or a combination, characterizes much archaeology addressed at understanding ritual and symbolism. While buildings denominated temples may sometimes be unambiguously identified, usually through written documents, such as ancient Egyptian and Maya inscriptions, in other cases the identification of 'temples' may be challenged. Other architectural settings that are assumed not to have seen use in everyday life, such as the ball courts of Central America, the Caribbean, and the US Southwest, are also viewed as sites of ritual.

Spatial locales that can be viewed as nonordinary by nature are sometimes treated as effectively unproblematic locales for an archaeology of ritual. Caves and natural springs are prominent foci for some definitions of 'sacred geographies.' In contrast, the investigation of entire landscapes, such as those of megalithic Europe, emphasizes the social marking of space through ritual action (see *Settlement and Landscape Archaeology*).

4.4 Visual Images

Artistic representations of artifacts and buildings recovered archaeologically have been a staple in reconstructing ritual. They have been approached in two complementary ways. On one hand, researchers have argued that scenes represented mythologies or beliefs about the supernatural. Some scenes have been identified as models for ritual enactment. For example, the elaborately-dressed individuals buried at the coastal Peruvian Sipan site have been identified with the clearly nonhuman figures depicted in supposed mythological scenes on pots of the Moche kingdom. The implication that such acts as ritual decapitation and mountaintop sacrifice actually took place within ancient Peruvian society has been supported by the recovery of caches of skulls and of frozen bodies on the high Andes.

In other cases, representations have been interpreted as scenes recording ritual practices. For example, Classic Maya *stelae* have been identified as showing members of the ruling class engaged in ritual bloodletting and the sacrificial burning of paper and resin. The availability of written texts in societies like those of the ancient Maya and Mesopotamia has led to ambitious attempts to reconstruct specific ritual practices from an emic perspective.

4.5 Animal Remains

Analysis of faunal assemblages is a relatively recent addition to the repertoire of methods used to reconstruct ancient ritual practices (Anderson and Boyle

1996). Ritual studies begin with the assumption that the use of any materials in unusual proportions within a site may signal the location of activities different from those oriented toward the practical demands of day-to-day life. With faunal remains, an additional criterion can be added: the identification of parts of animals that cannot have been used for subsistence.

5. *Future Directions*

Contextual analysis has become a fundamental part of archaeological investigation of ritual and symbolism. Through contextual analyses, specific suites of artifacts are identified that can be interpreted as evidence of ritual action. Continued emphasis on combining data to produce multifaceted arguments for ritual, and on placing active social agents at the center of reconstructions of meaning, can be expected.

For example, archaeologists in many regions have converged on the concept of feasting as a means to interpret particular assemblages of artifacts, animal bone, and plant remains. While feasting itself need not be seen as a ritual, it is often described in those terms, and is always interpreted as of symbolic significance. Studies of ritual feasting employ remarkably similar language and concepts. Drinking, the ingestion of drugs, and dancing have all been identified as part of ritual feasts. This is leading to the identification of specific categories of material remains that provide evidence of presumed ritual activities. Attention is also being directed to the experiential effects of ancient ritual practices on the development of forms of subjectivity.

Archaeologists are otherwise less inclined to propose specific meanings for symbols, except in literate traditions where texts continue to serve to validate interpretations. Continuing debate will focus on how to create interpretations of symbols that incorporate multivocality and capture the recursion between practice and meaning.

See also: Cognitive Archaeology; Death, Anthropology of; Folk Religion; Household Archaeology; Magic, Anthropology of; Ritual; Sacrifice; Symbolism (Religious) and Icon

Bibliography

Anderson S, Boyle K (eds.) 1996 *Ritual Treatment of Human and Animal Remains*. Oxbow Books, Oxford, UK

Bradley R J 1991 *The Passage of Arms: An Archaeological Analysis of Prehistoric Hoards and Votive Deposits*. Cambridge University Press, Cambridge, UK

Chapman R, Kinnes I, Ransborg K (eds.) 1981 *The Archaeology of Death*. Cambridge University Press, New York

Drennan R D 1983 Ritual and ceremonial development at the early village level. In: Flannery K V, Marcus J (eds.) *The Cloud People*. Academic Press, New York

Gardin J-P, Peebles C S (eds.) 1992 *Representations in Archaeology*. Indiana University Press, Bloomington, IN

Garwood P (ed.) 1991 *Sacred and Profane: Proceedings of a Conference on Archaeology, Ritual and Religion*. Oxford Committee for Archaeology, Institute of Archaeology, Oxford, UK

Hawkes C 1954 Archaeological theory and method: Some suggestions from the Old World. *American Anthropologist* **56**: 155–168

Hodder I (ed.) 1982a *Symbols in Action: Ethnoarchaeological Studies of Material Culture*. Cambridge University Press, Cambridge, UK

Hodder I (ed.) 1982b *Symbolic and Structural Archaeology*. Cambridge University Press, Cambridge, UK

Hodder I (ed.) 1989 *The Meanings of Things: Material Culture and Symbolic Expression*. Unwin-Hyman, London

Richards C, Thomas J 1984 Ritual activity and structured deposition in Later Neolithic Wessex. In: Bradley R, Gardiner J (eds.) *Neolithic Studies: A Review of Some Current Research*. B. A. R. British Series, Oxford, UK

Robb J E 1998 The archaeology of symbols. *Annual Review of Anthropology* **27**: 329–346

Robb J E (ed.) 1999 *Material Symbols. Center for Archaeological Investigations*. Southern Illinois University, Carbondale, IL

Wilkins J B (ed.) 1996 *Approaches to the Study of Ritual: Italy and the Ancient Mediterranean*. Accordia Research Centre, London

Wobst M H 1977 Stylistic behavior and information exchange. In: Cleland C L (ed.) *For the Director: Research Essays in Honor of James B. Griffin*. University of Michigan Museum of Anthropology, Papers in Anthropology, No. 61, Ann Arbor, MI

R. A. Joyce

Robinson, Joan (1903–83)

Joan Robinson was born on October 31, 1903, in Camberley, Surrey, into an upper middle-class English family characterized by vigorous dissent and independence of mind. Her great-grandfather was F. D. Maurice, the Christian Socialist who lost his Chair of Theology at King's College, London, on a point of principle; her father was Major General Sir Frederick Maurice, the victim of the infamous Maurice debate in 1918 and subsequently Principal of what is now Queen Mary College in the University of London. Her mother was Helen Margaret Marsh, the daughter of Frederick Howard Marsh, Professor of Surgery and Master of Downing at Cambridge. Joan Robinson was born into a milieu of controversy. She was educated at St. Paul's Girls' School in London, where she read history, and from 1922 to 1925 at Girton College, Cambridge, where she was Gilchrist Scholar. There she read for the Economies Tripos, searching for rational argument and hoping to understand poverty and its cures. She graduated in 1925 with second-class honours ('a great disappointment'–her words) although it was not until 1948, when women

were given full membership of the university, that she was admitted to the degree. In 1926, she married Austin Robinson, the Cambridge economist, and the couple sailed for India for what was to become the first of many visits. They stayed for over two years and returned to Cambridge in 1929, making it their intellectual and domestic base for the rest of their lives. They had two daughters and five grandchildren. Joan joined the Cambridge Faculty of Economics and Politics as a Faculty Assistant Lecturer in 1934, becoming part of an outstanding group of scholars in Maynard Keynes, A. C. Pigou, Dennis Robertson, Austin Robinson, Maurice Dobb, Richard Kahn, Gerald Shove and Piero Sraffa. She became a University Lecturer in 1937, Reader in 1949, and Professor in 1965. She was elected to a Professorial Fellowship at Newnham and made Honorary Fellow of Girton in 1965, and Honorary Fellow of Newnham in 1971; she was the first woman to be made an Honorary Fellow of King's (in 1979). From 1958 to 1971 she was an FBA. She 'retired' from her Chair in 1971, remaining as active as ever in Cambridge and abroad, until her death in Cambridge on August 3, 1983.

Early in her career Joan Robinson set out her views on methodology for economics, (Robinson 1932a, Harcourt 1990a). Her procedure was to identify and make explicit her assumptions and to develop a 'box of tools' for understanding and exploring the analytical object, to construct a logically sound analysis of the processes concerned: these successive stages she saw as independent of each other. Extending these views, she suggested (1937b, pp. 99–102) that controversy could result from lack of understanding, lack of knowledge, errors of logic, from working from different assumptions but, most commonly in economics, from differences in (political) objectives: the goals of a theory could direct its analysis as well as its policy conclusions. Throughout her life she argued that economics is not neutral, it is always ideologically driven. Nevertheless she believed that it was possible to develop pure analysis independently of dispute.

Her early contributions to economic theory were mainly in the realm of value theory (microeconomics, as it is called now) and *The Economics of Imperfect Competition* (1933), her first book, generously acknowledged her debts to Marshall (via Pigou), Kahn, Austin Robinson, Shove, and Sraffa. It was a controversial and much celebrated success, revealing her great clarity of thought and exceptional analytical powers. It was her interpretation, within a Marshallian context, of several contemporary theoretical issues. In particular, it was a response to the heretical claims about decreasing costs and increasing returns being made in the *Economic Journal*, especially Sraffa (1926), Symposium (see Robertson et al. 1930), and including Robinson (1932b). Robinson replaced an analytical basis in Marshall's free competition with one in imperfect competition, where free competition became a special case. She also moved the focus from Marshall's long period to a short period. Within that context the book also presented new concerns. On the basis of the downward sloping demand curve for each firm, Robinson introduced the marginal revenue curve as an integral analytical construct. She showed how the response of relative factor shares to a change in demand depended on whether you used her partial analysis of the elasticity of substitution or Hicks', aggregate. She reconceived Marshall's 'representative' firm as an 'equilibrium' firm where each such firm faced a less than infinitely elastic demand curve for its product in a monopolistic market. What then became 'the industry' in such a market? And was the simple profit-maximising rationale too reductionist to depict the firm? How was a monopolist to be identified theoretically? Or a duopolist? How indeed was imperfect competition definable?

The Economics of Imperfect Competition provoked; each stage of the theoretical process, even her geometrical method, incited reaction. More controversial than her analysis were its implications for policy. She represented perfect competition theory as demonstrating that the economy left alone would generate the optimal outcomes: it suggested a political philosophy of *laissez faire*. Imperfect competition showed that left alone the economy could become stuck at a position of unemployed capacity and labour and indeed, at that time, this was the case. The debates on the elasticity of substitution further illustrated what the emerging critique of perfect competition theory could mean for the status quo. The issues raised in this book were to fill the journals for decades to come. She herself came to reject its analyses (postwar) as a 'shameless fudge.'

Robinson's interests were not confined to microeconomics. During this time she was a member of a group of economists who met to discuss John Maynard Keynes's *Treatise on Money* (1930) and his emerging new theory. His ideas appealed to her, they represented a break from the orthodox and they were focused on the pressing problem of the persistence of high levels of unemployment. As a member of the Circus, with Meade, Kahn, Austin Robinson, and Sraffa, Robinson contributed to the critical scrutiny of drafts of Keynes's *The General Theory* (1936). A formidable group, but they were persuaded by the dissenting view of the behaviour of investment and of saving in Keynes's new theory of output and employment. Immediately following the book's publication, she wrote to develop its themes and analysis. *Essays in the Theory of Employment* (1937a) and the popular *Introduction to the Theory of Employment* (1937b) were both published in the year following *The General Theory* (and in the same year that her second daughter was born). She sent Keynes drafts of the two books as well as the proofs, and the related correspondence reveals his respect for Robinson in her role as his interpreter. *Introduction* is still one of the best guides to *The General Theory*.

During the interwar years, Michal Kalecki had a major impact on Robinson's thinking. His pre-emptive version of the theory of effective demand, which interwove imperfect competition with a macroeconomic theory of distribution, was based on Karl Marx's schema of reproduction and his analysis of the realisation crisis. She integrated his double-sided relation between investment and profits and his short-period theory of income distribution to develop her theory of employment and output. Kalecki provided a structural base to Robinson's theoretical explorations and extensions of Keynesianism and later of growth. During the war Robinson wrote *An Essay on Marxian Economics* (1942). Her interest was partly attributable to Kalecki's and Sraffa's Marxist-based work, and partly to the structure of society portrayed by, and fundamental to, such classical and Marxist analysis; her continued dissatisfaction with orthodox theory was also reflected in her essay which compared Marxian with mainstream ideas.

Robinson's interpretation of Marx again illustrates her approach to constructing theory. Responses to her book were to her methodology, as well as to her critical remarks about both Marxian and neoclassical economics. She extracted from Marx's writing the circuits of deductive logic with their attendant assumptions and political implications. The methodology of her censorship was not so inclusive as to embrace Marx's theory of history and of the processes of change. Yet it was these elements—the centrality of historical analysis to Marx's method and of class to understanding effective demand—that attracted Robinson to Marx's writings. The pervasive implications of time and history underlay her vision of capitalism and informed all of her subsequent economic research even though she ultimately was to abandon the classical method. Her book was a challenge to her colleagues to read and learn from Marx, if selectively and from a Keynesian perspective. It remains a useful introduction to Marxist economics.

In the early postwar years Robinson became committed to developing a dynamic analysis of capitalism. Inspired by Harrod's (1939, 1948) studies of long-period growth, *The Rate of Interest and Other Essays* (1952) was a collection in preparation for this project, in particular. The generalization of *The General Theory* attempted to extend Keynes into the long period. This essay was an interpretation of Keynes in terms of the Marshallian short period, characterized by uncertainty, incompatible decisions and unrealizable expectations, by a new definition of equilibrium. The elements of Robinson's growth theory were coming together in a non-neoclassical and long-period framework.

Integral with Robinson's exploration of long-period accumulation was an increasing dissatisfaction with the conceptual and logical foundations of neoclassical theory and its method, for which she became most well known, if not best understood. She prefaced her constructive contributions to dynamic theory with the critique of the neoclassical production function, Robinson (1953–4). This article fired a debate that continued for more than two decades. The implications of these exchanges were for the legitimacy of the very concept of capital and hence of the aggregate production function, fundamental to neoclassical theory. In particular, Robinson targeted its distribution analysis and its corollary, value theory. Ultimately, she criticized its method, that 'the neo-neoclassics still refuse to understand the difference between a comparison of timeless equilibrium positions and the effects to be expected from a change taking place at a particular moment' (1978a, p. xix). That particular moment was characterised by its own inherited structure of decisions, expectations, and uncertainties. She rejected the neoclassical method of long-period equilibrium as a point towards which the actual economy is moving; the implied stationarity of that theory meant that the movement towards it, and the rest point were logical, not historical. Her own account would not be in equilibrium terms but would take into account the actual features of the economy, the existence of uncertainty, and disappointed expectations. Her understanding of capitalism was based on classes and relations of conflict, while the neoclassicals begin with technical factors and harmonious distributive shares. What she wanted to put in the place of neoclassical theory was not a rival complete theory, which she considered would be only another 'box of tricks,' but an approach that started from institutions, history, and the 'rules of the game' (Harcourt 1990b). Robinson's optimism about the power of economic debate was continually frustrated, yet she determinedly defended what seemed to her to be right and just arguments. Her conservative colleagues, she felt, treated new ideas 'like heresies.'

The Accumulation of Capital (1956) continued the 'generalization' within a classical framework. The questions she set out to explore were: what are the conditions for cumulative long-term growth of income and capital (Golden Age) and what is the outcome of this process in terms of growth of gross and net output and the distribution of income between wages and profits, given a certain evolution through time of the labor force and technology? Inspired by Harrod's 'fruitful provocation,' *The Accumulation of Capital* brought together Marx's historicism, Harrod's economic dynamics, and Wicksell's capital theory. True to her methodology, she started out by presenting her most basic model and then devoted the rest of this remarkable book to relaxing the 'most severely simple' assumptions and tracing out the implications. 'The whole argument,' she proposed, 'is set out … as an analytical construction, with the minimum of controversy' (1956, p. x). In contrast to the mathematical style endemic to American economics, Robinson wrote her entire work of pure theory in literary form. It was a difficult book to assimilate and several years

later she followed it up with a hands-on guide. *Exercises in Economic Analysis* (1960) fills in the gaps, clarifies obscurities, and extends the arguments from *The Accumulation of Capital*. Its object, Robinson stated, was 'to sort out questions' (1960, p. xx). It offers readers an opportunity to formulate for themselves the analytical examples she derives in *The Accumulation of Capital*. This book (1960) was immediately followed by *Essays in the Theory of Economic Growth* (1962) which presents a retrospective introduction for the 'too terse' earlier (1956) book. Although there was a hint of re-switching in Robinson (1953–4) as well as in *The Accumulation of Capital*, Robinson dismissed the phenomenon at the time and saw it as a distraction from the main methodological point of her critique of neoclassical capital.

Dissatisfied with her quest for a long-period theory of growth, Robinson looked more to Kalecki's theory of cycles as a depiction of capitalism. She was a pioneer in the development of what was to become post-Keynesian theory, its bases in Sraffa's interpretation of Ricardo and in Kalecki's interpretation of the theory of employment. To this she brought 'a general framework of long- and short-period analysis which will enable us to bring the insights of Marx, Keynes, and Kalecki into coherent form and apply them to the contemporary scene' (1978b pp. 1, 18). Within this framework, post-Keynesian theory offered a 'method of analysis which takes account of the difference between the future and the past' (1978b, p. 12). Here was a broad conceptual structure that brought together Robinson's major theoretical concerns. From this analysis, post-Keynesian theory is evolving in several, not always compatible, directions, establishing its authority in the body of economic theory.

Robinson was a charismatic teacher; she inspired successive generations of students to develop their own ideas from her methods of analysis and brilliant insights, to think critically and for themselves. She spent many years demonstrating that 'mainstream teaching has been inculcating defective methodology'; she wanted to provide a logically sound, historically based dynamic approach to first-year economic analysis. *An Introduction to Modern Economics* (1973) with John Eatwell was a response to the 'fashionable textbooks [of] mainstream' economics. It presented a history of thought, a classical-cum-Kaleckian/post-Keynesian analysis. It was an ambitious, presumptuous book, at once a critique of neoclassical theory, and a terse introduction to the alternative. Robinson overestimated the intuitive powers of her audience and it proved too difficult for beginners; it remains a challenge for more advanced students.

Throughout her life Robinson pursued an active interest in the Third World, in India and China in particular, and in the Soviet Union. Her initial writing on development was based on the growth models developed in *The Accumulation of Capital*. Kalecki influenced this book. Development, she argued, should proceed by appropriate land reform and industrialization in both agriculture and manufacture, mindful of the balance of payments constraint and a minimum real wage or its equivalent. Socialist development, she continued, could mobilize savings and enable the state to organize the investment of the surplus, so achieving faster if not more efficient accumulation and growth than under capitalism; socialism, furthermore, led to a more equitable distribution of income (although the standard of living she saw as a political decision). She was less enthusiastic about the Soviet-style socialism of Eastern Europe than she was about China. Her mission to eradicate poverty and to meet basic needs always guided her judgements. She leaves a body of work; some reprinted in her *Collected Economic Papers* as well as Robinson (1979), on development planning and socialism that is relevant and provocative.

Joan Robinson depicted her work as an agenda for discussion as well as formulating the questions. Yet she has left an important legacy in her substantial literary output. She wrote for colleagues, students, and for the general, interested reader. As well as her many books, there is a stream of challenging evaluative essays written through the 50 years of her career. Her writing includes a theoretically rigorous critique of neoclassical economics that as yet goes unanswered; the questions she has raised present a formidable program for further research. She leaves an exacting methodology, a theoretical structure, and significant analytical details of a new, unified system of political economy.

See also: Capitalism; Capitalism: Global; Demand for Labor; East Asian Studies: Economics; Economic Growth: Measurement; Economic Growth: Theory; Economic History; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Economics: Overview; Employment and Labor, Regulation of; Firm Behavior; Labor Supply; Market Areas; Market Structure and Performance; Marxian Economic Thought; Policy History: State and Economy; Political Economy, History of; Soviet Studies: Economics; Unemployment: Structural

Bibliography

Harcourt G C 1990a Joan Robinson's early views on method. *History of Political Economy* **22**: 411–27

Harcourt G C 1990b Robinson, Joan Violet. In: Lord Blake, Nichols C J (eds.) *The Dictionary of National Biography* 1981–1985. Oxford University Press, Oxford, UK, pp. 346–7

Harcourt G C 1996 Some reflections on Joan Robinson's changes of mind and their relationship to post-Keynesianism and the economics profession. In: Marcuzzo M C, Pasinetti L, Roncaglia A (eds.) *The Economics of Joan Robinson*. Routledge, London, pp. 317–29

Harrod R F 1939 An essay in dynamic theory. *Economic Journal* **49**: 14–33

Harrod R F 1948 *Towards a Dynamic Economics*. Macmillan, London
Keynes J M 1930 *A Treatise on Money*, 2 Vols. In: *Collected Writings*. Macmillan, London, Vols. V, VI
Keynes J M 1936 *The General Theory of Employment, Interest and Money*. In: *Collected Writings*, Macmillan, London, Vol. VII, 1973
Robertson D H, Shove G F, Sraffa P 1930 Increasing returns, and the representative firm: A symposium. *Economic Journal* **40**: 79–116
Robinson J 1932a *Economics is a Serious Subject: The Apologia of an Economist to the Mathematician, the Scientist and the Plain Man*. Heffer, Cambridge, UK
Robinson J 1932b Imperfect competition and falling supply price. *Economic Journal* **42**: 544–54
Robinson J 1933 *The Economics of Imperfect Competition*. Macmillan, London
Robinson J 1937a *Essays in the Theory of Employment*. Macmillan, London
Robinson J 1937b *Introduction to the Theory of Employment*. Macmillan, London
Robinson J 1942 *An Essay on Marxian Economics*. Macmillan, London
Robinson J 1952 *The Rate of Interest and Other Essays*. Macmillan, London
Robinson J 1952–80 *Collected Economic Papers*. Blackwell, Oxford, UK, Vols. I–VI
Robinson J 1953–4 The production function and the theory of capital. *Review of Economic Studies* **21**: 81–106
Robinson J 1956 *The Accumulation of Capital*. Macmillan, London
Robinson J 1960 *Exercises in Economic Analysis*. Macmillan, London
Robinson J 1962 *Essays in the Theory of Economic Growth*. Macmillan, London
Robinson J 1978a *Contributions to Modern Economics*. Blackwell, Oxford, UK
Robinson J 1978b Keynes and Ricardo. *Journal of Post Keynesian Economics* **1**: 12–8
Robinson J 1979 *Aspects of Development and Underdevelopment*. Cambridge University Press, Cambridge, UK
Robinson J, Eatwell J 1973 *An Introduction to Modern Economics*. McGraw-Hill, Maidenhead, UK
Sraffa P 1926 The laws of returns under competitive conditions. *Economic Journal* **36**: 535–50

G. C. Harcourt and P. M. Kerr

Robustness in Statistics

Robust statistics is concerned with the description and the correction of problems caused by a discrepancy between attributes of the data being analyzed and the implicit or explicit assumptions used in modeling the data. But, of course, robustness is a word used frequently and in diverse situations. In statistics, it also conveys various meanings, always involving the stability of the method under study. The arithmetic mean, for example, is an optimal estimator of the expectation if the data follow exactly a normal distribution. Tiny deviations of the actual distribution of the data from the normal ideal, however, destroy the optimality and the arithmetic mean can be an arbitrarily bad estimator even though the data has a very nearly normal distribution. The robustness of estimators may also be defined without reference to a model. The arithmetic mean of the three observations (14, 13.7, 26.3), for example, is equal to 18.0 and the lack of stability is expressed through the large effect of the single observation 26.3 on the estimate.

Implicit and explicit assumptions in statistics are of various types. In a generalized linear model, for example, they include the link function, which determines the scale on which the relationship between the response and the explanatory variables is of the form of a linear combination of terms. Another assumption concerns the cumulative distribution for the response, which determines certain aspects of the inherent randomness of the data, such as the form of the relationship between the standard deviation and the expectation of the response variable or the rate of growth of the extreme $(1-\alpha)$ quantile of the response variable as α tends to zero. The best-known phenomenon observed in the data concerns the so-called *outliers*. This is a term used in describing various incidents in a data base. It may refer to *gross* errors in observations due to faulty transcription, mistakes in the manipulation of a measurement instrument, errors in the preparation of a particular experiment, confusion as to which experiment should have been performed, or some other such mistake. The frequency of such gross errors is much higher than one would expect at first thought. We will also use the term *wild value* for a distinct outlier. The term 'outlier' may also refer to the effect, that a sizable proportion of the data seems to behave differently from the rest. In a linear regression, for example, it is sometimes observed that any one regression surface fits only part of the data. This might be due to an overly simple model for the surface, or to the fact that the data is a mixture of cases that cannot be separated into more homogeneous pieces based on the available variables.

Of course, the data can be used to check assumptions and it is certainly good statistical practice to do so. To check details of a model, however, requires huge amounts of data and it is misleading to suggest that diagnostic checking of model assumptions can solve the robustness problem. Good practice in statistics requires the use of methods that do not suffer the drawbacks of the arithmetic mean of loosing performance under tiny violations of the idealistic assumptions. This is especially true in routine and less thoroughly scrutinized applications of statistical methodology. Robust methods retain very good performance, while being less sensitive to assumptions.

Robustness seems similar to nonparametric or, more precisely, distribution-free statistics. This is a superficial and even wrong view. Converting data to ranks, for example, in order to liberate the statistical

analysis from distributional assumptions, does not give particularly good protection from the deleterious effects of atypical observations.

1. Indicators of Robustness

1.1 Sensitivity Functions

Let $T(y_1, \ldots, y_n)$ be an estimator of a real parameter, calculated with n observations, $(y_1, \ldots, y_n)$. The sensitivity function is a simple idea leading to a better understanding of the estimator. The question being asked is: *'How would the estimator have come out, if the data had not been exactly the ones I observed?'* To answer it, one considers the *sensitivity function*

$$\mathrm{SC}(y) = (n+1)(T(y, y_1, \ldots, y_n) - T(y_1, \ldots, y_n)) \qquad (1)$$

that is the change in the estimate induced by an additional observation at an arbitrary value y. The multiplier $(n+1)$ in this formula provides a natural scaling of the size of the change.

Examples. For the arithmetic mean $T(y_1, \ldots, y_n) = \bar{y} = (y_1 + \ldots + y_n)/n$ one finds

$$\mathrm{SC}(y) = (n+1)\left(\frac{y + y_1 + \cdots + y_n}{n+1} - \frac{y_1 + \cdots + y_n}{n}\right) = y - \bar{y}. \qquad (2)$$

Even in this easy case, the sensitivity function is of interest. It shows that the arithmetic mean is sensitive to the distance of the additional observation y from the average of the remaining data. The larger this istance, the larger the change.

For the trimmed mean, $T(y_1, \ldots, y_5) = (y_{(2)} + \ldots + y_{(4)})/3$, obtained by deleting the largest and the smallest observation, the sensitivity depends on the position of the additional observations y. If y is between $y_{(1)}$ and $y_{(5)}$, it will not be trimmed from the sample and $\mathrm{SC}(y) = 6((y_{(2)} + y_{(3)} + y_{(4)} + y)/4 - (y_{(2)} + y_{(3)} + y_{(4)})/3)$. If on the other hand, y is larger than $y_{(5)}$, then y will be trimmed and the sensitivity becomes $\mathrm{SC}(y) = 6((y_{(2)} + y_{(3)} + y_{(4)} + y_{(5)})/4 - (y_{(2)} + y_{(3)} + y_{(4)})/3)$, which is independent of y. The same reasoning applies in the case where y is smaller than $y_{(1)}$.

For the least-squares estimate of a regression coefficient, the algebra is a bit more difficult, partly because the data is of a more complex structure. Suppose we observe the couples of explanatory variables and responses, $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, and that we add an observation at $(\boldsymbol{u}, v)$. The least-squares estimate of the regression coefficients is $\boldsymbol{T}(y_1, \ldots, y_n) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, where $\boldsymbol{X}$ is the design matrix. After some algebra, one has

$$\mathrm{SC}(\boldsymbol{u}, v) = (v - \boldsymbol{u}^T\boldsymbol{T})\frac{(n+1)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{u}}{1 + \boldsymbol{u}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{u}}. \qquad (3)$$

Again, this is a revealing formula, because it makes it clear that the estimated regression coefficients are sensitive to the size of the residual, the first factor, but also to the position of $\boldsymbol{u}$ among the explanatory variables as expressed by the second factor.

In practice, it is not really necessary to do the computations illustrated in these examples. The sensitivity function can easily be computed numerically for any given estimator and underlying data set and one can thus construct the plot of $\mathrm{SC}(y)$ against y, which in most cases will serve as a good basis for judging the robustness qualities of the estimator. From such a plot one can deduce several indicators, the simplest one being the *maximal sensitivity*

$$g^* = \sup_y |\mathrm{SC}(y)| \qquad (4)$$

where $|\mathrm{SC}|$ denotes the absolute value of SC. In the above examples, the maximal sensitivity is equal to ∞ for the mean and the least-squares estimate of the regression coefficient—the first and the third example. It is finite for the trimmed mean. An estimator with infinite maximal sensitivity is necessarily *non-resistant* in the sense that a single wild observation that behaves differently from the rest can destroy the estimate and lead to rather arbitrary conclusions.

1.2 Influence Functions

The sensitivity function has a more formal counterpart in the *influence function*, which is defined as a kind of derivative of a statistical functional. In many cases it is simply the limit as the sample size n tends to infinity of the sensitivity function computed for a sample generated from a fixed model.

Examples. Let $y_1, \ldots, y_n$ be a sample from some distribution F with expectation μ_F. As n increases, the sensitivity function of the arithmetic mean converges to the influence function of the expectation, which is $\mathrm{IF}_F(y) = y - \mu_F$. The influence function of the trimmed mean obtained by trimming a fraction of α at each end is

$$\mathrm{IF}_F(y) = \begin{cases} (y - \mu_F)/(1-2\alpha), & \text{for } y \leqslant F^{-1}(1-\alpha); \\ \mathrm{sign}(y - \mu_F)F^{-1}(1-\alpha)/(1-2\alpha), & \text{otherwise} \end{cases} \qquad (5)$$

where F is any distribution symmetric around μ_F. In the regression case, one can choose an arbitrary error distribution F with finite variance and an arbitrary true regression coefficient θ. The influence function is

$$\mathrm{IF}_F(\boldsymbol{u}, v) = (v - \boldsymbol{u}^T\theta)\,(\mathrm{ave}(\boldsymbol{x}\boldsymbol{x}^T))^{-1}\boldsymbol{u} \tag{6}$$

where the matrix $\mathrm{ave}(\boldsymbol{x}\boldsymbol{x}^T)$ denotes the average design matrix.

The influence function is very similar to the sensitivity function and can thus be interpreted similarly. In particular, the *gross error sensitivity* is

$$\gamma^*_F = \sup_y |\mathrm{IF}_F(y)|. \tag{7}$$

The squared influence function is directly linked to the asymptotic variance of an estimator. Let $Y_1, \ldots, Y_n$ be independent and identically distributed with distribution F_θ for some real parameter θ. One can show that most estimators of θ satisfy the following approximate relationship:

$$T(Y_1, \ldots, Y_n) \approx \theta + \mathrm{ave}(\mathrm{IF}_{F_\theta}(Y_i)). \tag{8}$$

If this approximation is sufficiently precise, it implies that the asymptotic variance of the estimate T is

$$\mathrm{Var}_{\mathrm{asy}}(T) = E_{F_\theta}((\mathrm{IF}_{F_\theta}(Y))^2) = \int (\mathrm{IF}_{F_\theta}(t))^2 F_\theta(\mathrm{d}t). \tag{9}$$

In this case, one can also apply the central limit theorem to conclude that

$$P(T \leqslant t) \approx \Phi\left(\frac{\sqrt{n}(t-\theta)}{\sqrt{\mathrm{Var}_{\mathrm{asy}}(T)}}\right) \tag{10}$$

where Φ denotes the normal distribution function with density $\varphi(t) = \exp(-t^2/2)/\sqrt{2\pi}$. This in turn shows that

$$[T \pm 1.96\sqrt{\mathrm{Var}_{\mathrm{asy}}(T)}/\sqrt{n}] \tag{11}$$

is an approximated 95 percent confidence interval for θ.

1.3 Breakdown and High Resistance

A limitation of the influence function is its local nature. It is based on the change of the value of the estimator from one sample to another one that differs by adding a single small 'contamination.' The *breakdown point* is the simplest global measure of robustness. It is the largest percentage of wild values in a data set that an estimator can tolerate without giving wild estimates. This notion is inspired by the trimmed means, computed as the mean of a sample after deleting 100α percent of the largest and of the smallest data points. This estimator has a breakdown point of 100α percent, since any contamination by wild data below this fraction is being trimmed. A special case of the trimmed means is the median, which for an odd-sized sample ($n = 2k+1$) has a breakdown point of $k/n \approx 50$ percent. This is so, because the median is computed as the $(k+1)$st value in the sample and anything less than k wild values will not lead to a wild estimate. More generally, global indicators describe the properties of an estimator under contamination of a sizeable portion of the data.

The distributions used in modeling data sets containing outliers are *heavy-tailed*. One can, for example, model a sample containing a sizeable portion of outliers with a *contaminated distribution*. A sample generated from $(1-\varepsilon)F + \varepsilon H$ is a mixture of one part, $(1-\varepsilon)$, of a 'good' distribution F and another part, ε, of a 'bad' distribution G. The *contaminated normal distribution*, $(1-\varepsilon)\Phi(y) + \varepsilon\Phi(y/\sigma)$, describes observations that can either result from a standard normal distribution, or from a normal distribution having a larger variance $\sigma > 1$. As ε grows from 0 to 50 percent, a larger and larger fraction of the data is expected to exhibit a distinct behavior. A global robustness indicator for estimators is based on behavior under positive fractions ε, whereas local properties, such as the influence function, merely consider limits as ε tends to zero.

An estimator with a breakdown point of 50 percent is called *highly resistant*. Such estimators can be very revealing when applied to data containing a large proportion of outliers, because they lead to a clear separation of the 'good' and the 'bad' data. The median is an example of a highly resistant location estimator. The residuals from the arithmetic mean of the three observations (14.0, 13.7, 26.3) are equal to $(14.0 - 18.0 = -4.0, -4.3, 8.3)$, whereas the residuals from the median are $(14.0 - 14.0 = 0.0 - 0.3, 12.3)$. The highly resistant median produces residuals that make it immediately evident that the third observation is a wild value. The ability of estimators to reveal hidden outliers is explained by the fact that they have bounded sensitivity with regard to single observations, but also, and more importantly, that even a largish fraction of outliers does not exert an undue influence on the estimate. The median absolute deviation (MAD) is a highly resistant scale estimator. It is defined as

$$\mathrm{MAD}(y_1, \ldots, y_n) = \mathrm{median}(|y_1 - m|, \ldots, |y_n - m|) \tag{12}$$

where m denotes the median, median $(y_1, \ldots, y_n)$.

2. Robust Estimators

2.1 Least-squares, maximum likelihood, and M-estimators

Least-squares estimators,

$$T(y_1, \ldots, y_n) = \arg\min_\theta \sum_{i=1}^{n} (y_i - \mu_i(\boldsymbol{\theta}))^2 \qquad (13)$$

have an unbounded influence function and a breakdown point of 0 percent. Maximum likelihood estimators (m.l.e.) for independent data,

$$T(y_1, \ldots, y_n) = \arg\max_\theta \sum_{i=1}^{n} \log(f_i(y_i, \boldsymbol{\theta}))$$

$$= \text{solution of} \left(\sum_{i=1}^{n} \partial \log(f_i(y_i, \boldsymbol{\theta}))/\partial\boldsymbol{\theta} = \mathbf{0} \right) \qquad (14)$$

can be robust, depending on the properties of the densities f_i. Generalizing either of these two concepts, one obtains *M-estimators*, defined by

$$T(y_1, \ldots, y_n) = \arg\min_\theta \sum_{i=1}^{n} \rho_i(y_i, \boldsymbol{\theta}) \qquad (15)$$

where the choice of ρ_i is in principle quite free. The corresponding estimating equation is of the form

$$\sum_{i=1}^{n} \psi_i(y_i, \boldsymbol{T}(y_1, \ldots, y_n)) = \mathbf{0} \qquad (16)$$

where $\psi_i(y, \boldsymbol{\theta}) = \partial\rho_i(y_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$.

Example. For illustration, we consider here briefly the linear regression case, where $\mu_i(\boldsymbol{\theta}) = \boldsymbol{x}_i^T\boldsymbol{\theta}$, $f_i(y_i, \boldsymbol{\theta}) = f(y_i - \boldsymbol{x}_i^T\boldsymbol{\theta})$, $\rho_i(y_i, \boldsymbol{\theta}) = \rho(y_i - \boldsymbol{x}_i^T\boldsymbol{\theta})$, and $\psi_i(y_i, \boldsymbol{\theta}) = (\psi(y_i - \boldsymbol{x}_i^T\boldsymbol{\theta})\, \boldsymbol{x}_i)$ with $\psi(t) = \rho'(t)$. The M-estimator with $\rho(t) = t^2/2$ (rho-function) and $\psi(t) = t$ (psi-function) is nothing else but the least-squares estimator, which is the same as the m.l.e. for Gaussian errors. Any m.l.e. can be thought of as an M-estimator with $-\rho(t) = \log(f(t))$, but thinking in terms of M-estimators is an advantage and an important conceptual step since it frees us from any artificial constraints. The influence function of the M-estimator at the error distribution F and the true regression parameter $\boldsymbol{\theta}$ is

$$\mathrm{IF}_F(\boldsymbol{u}, v) = \frac{\psi(v - \boldsymbol{u}^T\boldsymbol{\theta})}{\int_{-\infty}^{\infty} \psi'(t) F(\mathrm{d}t)} (\mathrm{ave}(\boldsymbol{x}\boldsymbol{x}^T))^{-1}\boldsymbol{u}. \qquad (17)$$

The first factor in this influence function involves the residual at the contaminating point $(\boldsymbol{u}, v)$. It is bounded whenever $\psi = \rho'$ is bounded. The deficiencies in the stability of the least-squares estimator are therefore best explained by the fact that the derivative $\rho'(t) = \psi(t) = t$ is unbounded.

M-estimators are defined by a rho-function and an associated psi-function. As shown above, boundedness of the psi-function is a necessary condition to achieve robustness. Careful tuning of the shape of the psi-function permits both robustness and efficiency. It is well-known that the least-squares estimator, which has psi-function $\psi(t) = t$, is optimally efficient when the error distribution is exactly Gaussian. To preserve this Gaussian efficiency as much as possible, the linearity of the psi-function in a neighborhood of the origin $t = 0$ should remain unaltered. The *Huber estimator* is based on

$$\psi_H(t) = \min(1, \max(-1, t)) = \begin{cases} -1, & \text{if } t \leqslant -1; \\ t, & \text{if } -1 < t \leqslant 1 \\ 1, & \text{if } t > 1. \end{cases} \qquad (18)$$

The reader should draw this function to note that it does indeed preserve the linearity near the origin, while at the same time enforcing a bound. In this definition, the interval $[-1, 1]$ plays a special and undeserved role. It goes without saying that in any practical use of the estimator this interval must be adapted to the data.

Example. The Huber regression estimator is defined as the solution to the estimating equation

$$\sum_{i=1}^{n} \psi_H \left(\frac{y_i - x_i^T\boldsymbol{\theta}}{k \times S} \right) \boldsymbol{x}_i = \mathbf{0} \qquad (19)$$

where $S = S(r_1, \ldots, r_n)$ is a scale estimate based on residuals and $k > 0$ is a further tuning constant. Consider S as known and computed before solving the estimating equation (19). Dividing the argument of the psi-function by a multiple of a residual scale achieves the adaptation we were looking for. It is quite evident that the tuning constant k introduces a trade-off between robustness and Gaussian efficiency. For sufficiently large values of k the estimator is equal to the least-squares estimator because all scaled residuals $(y_i - x_i^T\boldsymbol{\theta})/(k \times S)$ will lie in the interval $[-1, 1]$ and thus in the linear part of Huber's psi-function. For small values of k, the Huber estimator loses Gaussian efficiency but gains in robustness. In the limit, as $k \to 0$, the scaled residuals are either exactly zero, smaller than -1, or larger than 1. In this case, the function ψ_H returns as values either 0, -1, or 1 and the above estimating equation turns into

$$\sum_{i=1}^{n} \mathrm{sign}(y_i - x_i^T\boldsymbol{\theta})\boldsymbol{x}_i = \mathbf{0} \qquad (20)$$

which corresponds to the so-called L_1-estimator defined by $\psi(t) = \text{sign}(t)$ and $\rho(t) = |t| =$ absolute value of t. A good compromise choice for the denominator $k \times S$ is $1.3 \times \text{MAD}$ (residuals).

2.2 Theoretical Justification of Huber's Estimator

A concise discussion of many of the topics discussed in this entry can be found in the two books by Huber (1981, 1996). Huber's location estimators T_k in the case of known scale ($S = 1$) are solutions of

$$\sum_{i=1}^{n} \psi_H \left(\frac{y_i - T_k(y_1, \ldots, y_n)}{k} \right) = \sum_{i=1}^{n} \psi_k(y_i - T_k(y_1, \ldots, y_n))/k = 0 \quad (21)$$

where

$$\psi_k(t) = \min(k, \max(-k, t)) = \begin{cases} -k, & \text{if } t \leqslant -k; \\ t, & \text{if } -k < t \leqslant k \\ k, & \text{if } t > k. \end{cases} \quad (22)$$

The corresponding rho-function is

$$\rho_k(t) = \begin{cases} -k(t + k/2), & \text{if } t \leqslant -k; \\ t^2/2, & \text{if } -k < t \leqslant k \\ k(t - k/2), & \text{if } t > k, \end{cases} \quad (23)$$

that is quadratic in the middle and linear in the tails. These M-estimators are m.l.e.s in the sense that $\exp(-\rho_k(t))$ can be standardized to become a density, $f_k(t)$. From the form of $\rho_k(t)$ it follows that $f_k(t)$ is like a Gaussian in the center and has tails that decrease exponentially, i.e.,

$$f_k(t) = \begin{cases} (1-\varepsilon_k)\exp(-t^2/2)/\sqrt{2\pi}, & \text{if } -k \leqslant t \leqslant k \\ (1-\varepsilon_k)\exp(k^2/2 - k|t|)/\sqrt{2\pi}, & \text{if } |t| > k \end{cases} \quad (24)$$

where $0 \leqslant \varepsilon_k \leqslant 1$ decreases as k increases. These distributions were discovered by Huber and have been discussed in his seminal paper (Huber 1964). The above formula shows that $f_k(t) = (1-\varepsilon_k)\varphi(t) + \varepsilon_k h_k(t)$ is a *contaminated density* with contamination fraction ε_k. Huber showed that f_k is in some mathematical sense the worst among all contaminated densities with a contaminating fraction of ε_k. As a consequence, the corresponding Huber estimator T_k is the estimator that best protects against a contamination of 100 ε_k percent of the data. The sets of possible error distributions considered in Huber's paper contain a positive fraction $\varepsilon_k > 0$ of an unspecified 'bad' density. The Huber estimator is therefore the best one in a global sense. Even if a sizeable portion of the data are not of any real use, this estimator will perform as well as possible under the circumstances. All of this, however, is only true for the simplest case of a location parameter with known scale. In more general regression models Huber's result no longer holds.

Indicators of robustness such as the breakdown point and the gross error sensitivity lead rather naturally to optimality problems. One is interested, for example, in finding an estimator that satisfies the two conditions of consistency and bounded gross error sensitivity and which, given these conditions, is most precise. This problem is of a local nature, because we do not consider sizeable fractions of outliers, but rather limits for very small fractions. The problem is again solved by the Huber estimators, which shows—not very surprisingly—that a globally best estimator is also locally best. The estimators T_k form a family of least variable estimators with gross error sensitivities that increase as k increases. In other words, they provide a bridge between the median ($k = 0$) and the mean ($k = \infty$) and between a minimal value of the gross error sensitivity and an infinite gross error sensitivity. Considering only local properties defined by the influence function has the advantage of being more easily generalized. Thus, it is possible to solve a variety of problems of this type. The book by Hampel et al. (1986) provides a detailed discussion of related results.

2.3 Redescending M-estimators

The form of the Huber estimator comes as a surprise because it never entirely rejects gross errors among the observations. In order to completely reject such outliers, the psi-function $\psi(t)$ would have to be equal to zero for sufficiently large $|t|$. The *biweight estimator* proposed by Tukey is a convenient example. It is based on

$$\psi_T(t) = \begin{cases} t(1-t^2)^2, & \text{if } -1 \leqslant t \leqslant 1 \\ 0, & \text{if } |t| > 1. \end{cases} \quad (25)$$

Such an estimator is called redescending, because the psi-function remains roughly linear near the origin, reaches a bound, and subsequently redescends towards zero. The reader is invited to draw Tukey's biweight

psi-function in order to appreciate these qualities. The biweight regression estimator solves the same estimating equation (19) as Huber's estimator, except that we replace ψ_H by ψ_T. This small change has important repercussions. Because ψ_T is not an increasing function, the estimating equation of the biweight estimator may have multiple solutions. The estimate is therefore not well defined. However, among all the solutions of the estimating equations, an appropriate one can usually be picked out by looking at the resulting residuals. Also, when multiple solutions do exist, several of them may be of interest. An appropriate choice for the denominator $k \times S$ when using the biweight is $7 \times$ MAD (residuals). The rho-function of the biweight estimator,

$$\rho_T(t) = \begin{cases} (1-(1-t^2)^3)/6, & \text{if } -1 \leqslant t \leqslant 1; \\ 1/6, & \text{if } |t| > 1 \end{cases} \tag{26}$$

is shaped like an vase with a flat rim and the solutions of the estimating equation are local minima of the criterion function

$$\text{criterion}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \rho_T \left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\theta}}{k \times S} \right). \tag{27}$$

Redescending M-estimators such as the biweight are of interest in the quest for highly resistant regression estimators. Inspection of the influence function (17) of the regression M-estimator reveals a weakness, namely the unboundedness of the second factor. In cases where one observation is far from the other observations in the sense that $\boldsymbol{u}$ is distant from the other observations taken at $\boldsymbol{x}_i$, the influence of this distant observation on the estimated regression parameter will be considerable. Redescending estimators for which the first factor in the influence function can be equal to zero correct this problem. Whenever the data give support to several regression equations, the redescending M-estimator can be of great value in pointing out the different solutions, each one as a separate local minimum of the criterion.

Rousseeuw's *least-median-of-squares* (LMS) estimator is another highly resistant regression estimator. It is defined by the criterion

$$\text{criterion}(\boldsymbol{\theta}) = \text{median}\{(y_1 - \boldsymbol{x}_1^T \boldsymbol{\theta})^2, \ldots, (y_n - \boldsymbol{x}_n^T \boldsymbol{\theta})^2\}. \tag{28}$$

The LMS is a useful, often revealing, estimate. It is, however, quite unstable, not with regard to outliers, but with regard to other features of the data.

2.4 Finite Samples

A robust data analysis must be valid and efficient beyond a narrow set of circumstances. The classical regression estimator, for example, is chosen to be the best possible for normally distributed observations

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim \text{normal}(0, \sigma^2). \tag{29}$$

Robustness demands joint optimality for a broader set of error distributions. One way in which this can be achieved, is by choosing a finite number of challenges, in the simplest case a couple such as {normal, Student's t_2}, the idea being that the estimator would have to perform well, whether the errors were generated by one or the other of the distributions. For such a set of challenges, the performance of an estimator is determined by a set of efficiencies, one for each challenge. Typically, an estimator that is highly efficient for one of the challenges will have a low efficiency for all the others. In the example of a couple of challenges including the normal distribution and the heavy-tailed Student's t_2 distribution, the least-squares estimator, for example, has optimal effiency for the Gaussian but is inefficient for the t_2. Robust estimators on the other hand must have a high efficiency for all challenges. How to compute estimators that maximize jointly the whole set of efficiencies is discussed in Morgenthaler and Tukey (1991).

2.5 Iteratively Re-weighted Least-squares

The computation of the least-squares estimator is straightforward because the estimating equation is linear. A program for computing the weighted least-squares solution is all that is needed to find the values of M-estimators. To illustrate the idea, we consider the solution of (21). Rewriting the estimating equation as

$$\sum_{i=1}^{n} \left[\frac{\psi_k(y_i - T_k(y_1, \ldots, y_n))}{(y_i - T_k(y_1, \ldots, y_n))} (y_i - T_k(y_1, \ldots, y_n)) \right] = 0 \tag{30}$$

an algorithm immediately suggests itself. We could start with a reasonable guess T_k^0 for the value of T_k, such as the median of the observations. Next, we compute weights

$$w_i^0 = \frac{\psi_k(y_i - T_k^0)}{y_i - T_k^0} \tag{31}$$

and then solve the estimating equation, $\Sigma_{i=1}^n w_i^0 (y_i - T_k(y_1, \ldots, y_n)) = 0$. From this, we obtain the next guess for T_k, i.e.,

$$T_k^1 = \sum_{i=1}^{n} w_i^0 y_i / \sum_{i=1}^{n} w_i^0. \tag{32}$$

With this new value for T_k we are back at the start of the loop, ready to recompute the weights, and then

the next value for T_k. Repeating the loop a fixed number of times or until convergence leads to practical values for M-estimates.

3. Robust Inference

The sampling distributions of robust estimators can often be approximated sufficiently well by their asymptotic normal distributions. For the regression M-estimator the following formula gives an estimated variance matrix of the estimated regression coefficients

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(kS)^2\left(\psi\left(\frac{r_i}{kS}\right)\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{w}\psi'\left(\frac{r_i}{kS}\right)\right)^2} \tag{33}$$

where $r_1, \ldots, r_n$ are the residuals. Using the approximate normality of the estimate, this equation allows one to compute robust confidence intervals for individual parameters. The robustness in the context of confidence intervals means that the coverage of the parameter is equal to the nominal coverage rate even under deviations from the model. Furthermore, the interval should not achieve this feat by being overly long.

Distribution-free tests are very resistant in one respect. Their size, that is the probability of falsely rejecting the null hypothesis, is stable. The power, however, depends, of course, on the distribution of the observations and need not be stable under changes to this distribution. Robust tests must have both, stable sizes and good power under a variety of circumstances. This point illustrates once more the fact that robustness and distribution-freeness are two quite distinct aspects of statistical inference.

See also: Distributions, Statistical: Approximations; Distributions, Statistical: Special and Continuous; Distributions, Statistical: Special and Discrete; Semiparametric Models; Statistics: The Field

Bibliography

Box G E P, Leonard T, Wu C F (eds.) 1983 *Scientific Inference, Data Analysis, and Robustness Proceedings of a Conference Conducted by the Mathematics Research Center, The University of Wisconsin–Madison, November 4–6, 1981*. Academic Press, New York

Gnanadesikan R 1977 *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York

Hampel F R, Ronchetti E M, Rousseeuw P J, Stahel W A 1986 *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York

Hoaglin D C, Mosteller F, Tukey J W (eds.) 1983 *Understanding Robust and Exploratory Data Analysis*. Wiley, New York

Hoaglin D C, Mosteller F, Tukey J W (eds.) 1985 *Exploring Data Tables, Trends, and Shapes*. Wiley, New York

Huber P J 1964 Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**: 73–101

Huber P J 1981 *Robust Statistics*. Wiley, New York

Huber P J 1996 *Robust Statistical Procedures (2nd edn.) CBMS-NSF Regional Conference Series in Applied Mathematics* **68**. SIAM, Philadelphia

Jurečová J, Sen P K 1996 *Robust Statistical Procedures: Asymptotics and Interrelations*. Wiley, New York

Kadane J B (ed.) 1984 *Robustness of Bayesian Analyses*. North-Holland/Elsevier, Amsterdam

Launer R L, Wilkinson G N (eds.) 1979 *Robustness in Statistics: Proceedings of a Workshop Sponsored by the Mathematics Division, Army Research Office, held at Army Research Office, Weiss Buildings, April 11–12, 1978*. Academic Press, New York

Morgenthaler S, Ronchetti E, Stahel W A (eds.) 1993 *New Directions in Statistical Data Analysis and Robustness*. Birkhäuser, Basel

Morgenthaler S, Tukey J W (eds.) 1991 *Configural Polysampling: A Route to Practical Robustness*. Wiley, New York

Rieder H 1994 *Robust Asymptotic Statistics*. Springer-Verlag, Berlin

Rousseeuw P J, Leroy A M 1987 *Robust Regression and Outlier Detection*. Wiley, New York

Sprent P 1993 *Applied Nonparametric Statistical Methods*, 2nd edn. Chapman and Hall, London

Stahel W, Weisberg S (eds.) 1991 *Directions in Robust Statistics and Diagnostics, Parts I and II*. Springer-Verlag, Berlin

Staudte R G, Sheather S J 1990 *Robust Estimation and Testing*. Wiley, New York

Wilcox R R 1997 *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, San Diego

S. Morgenthaler

Rogers, Carl Ransom (1902–87)

Carl R. Rogers was arguably America's most influential psychotherapist and one of its most prominent psychologists. His therapeutic approach, known most often as 'client-centered therapy,' was validated through voluminous scientific research and taught widely, profoundly affecting the fields of counseling and psychotherapy throughout the world and permeating many other helping professions. A leader in the humanistic psychology movement of the later twentieth century, Rogers' writings and personal example influenced individuals personally and professionally in many continents.

1. Education and Influences

Rogers was born on January 2, 1902 in suburban Oak Park, Illinois, where he spent his childhood with four brothers, a sister, and conservative, Protestant

parents. Spending teenage years on the family farm in Glen Ellen, Illinois, he developed a love of nature and serious working knowledge of scientific method, conducting agricultural experiments on a plot he and his brothers managed. He majored in agriculture at the University of Wisconsin, Madison, but switched to history after experiencing the call to religious work. Motivated by the 'social gospel' as much as religious conviction, he eschewed his parents' dogmatism and attended liberal Union Theological Seminary in New York City, also taking courses at adjoining Teachers College, Columbia University. There, religious doubt combined with fascination with psychology. Influenced by instructors Leta Hollingworth, Goodwin Watson, William Heard Kilpatrick, the leading proponent of John Dewey's progressive education philosophy, and others, Rogers transferred to Teachers College to pursue a doctorate in clinical psychology.

At Columbia, exposure to the testing and measurement movement of E. L Thorndike was balanced by his experiences in clinical work and fellowship at the Institute for Child Guidance, where he encountered Freudian thought, a lecture by Alfred Adler, Rorschach testing, and other psychoanalytic and psychiatric approaches. Seeking to integrate these influences, Rogers came to appreciate the importance of attending to and honoring clients' inner experience while also striving to assess objectively any changes which resulted from treatment. His doctoral dissertation, *Measuring Personality Adjustment in Children Nine to Thirteen Years of Age* (1931), combined both subjective and objective measures. Rogers' 'Personality Adjustment Inventory' was subsequently published by YMCA Press and used for over 50 years.

His next major learning experience was the 12 years working with thousands of troubled children and adults in Rochester, New York, where he gradually developed his nondirective approach to counseling and psychotherapy. During this period, he was influenced by students of Otto Rank, especially Jessie Taft whose 'relationship therapy' shifted emphasis from past content to a focus on the patient's self-insight and self-acceptance within the therapeutic relationship.

2. *Positions Held*

In Rochester, Rogers worked at the Society for Prevention of Cruelty to Children (1928–38), where he directed the Child Study Department, then became director of the Rochester Guidance Center (1939–40). Based on his practical experience and first book, *Clinical Treatment of the Problem Child* (1939), he was offered a full Professorship at Ohio State University, where he taught from 1940 to 1944. After an interim year training United Service Organization (USO) workers to counsel returned servicemen, he moved to the University of Chicago, where he developed and ran the internationally renowned Counseling Center and taught in the Psychology Department from 1945 to 1957. He held joint appointments in the Departments of Psychology and Psychiatry at the University of Wisconsin from 1957 to 1963 before leaving academic life and moving to the Western Behavioral Sciences Institute in La Jolla, California (1963–72). He and others then formed their own organization, Center for Studies of the Person, where Rogers remained until his death in 1987.

3. *Client-centered Therapy*

Rogers' approach to counseling and psychotherapy was based on a core hypothesis about human growth and personality change, which he summarized in 1950:

> Client-centered therapy operates primarily upon one central and basic hypothesis which has undergone relatively little change with the years. This hypothesis is that the client has within himself the capacity, latent if not evident, to understand those aspects of his life and of himself which are causing him pain, and the capacity and the tendency to reorganize himself and his relationship to life in the direction of self-actualization and maturity in such a way as to bring a greater degree of internal comfort. The function of the therapist is to create such a psychological atmosphere as will permit this capacity and this strength to become effective rather than latent or potential.

While other therapies might profess similar belief, Rogers' method of creating the therapeutic psychological atmosphere was radically different from other approaches commonly employed. Avoiding questions, interpretation, or other directive techniques, Rogers' initial 'nondirective method' relied exclusively on careful listening and skillful 'reflection of feelings,' leading to client insight and positive action. Although he always remained primarily nondirective in his practice, Rogers soon recognized that the counselor's *attitudes* were more important than his particular techniques. Still later he clarified that it was the therapeutic *relationship*, which the attitudes helped create, that was growth producing. Rogers called this therapeutic relationship 'client-centered' and described three key 'conditions' in the relationship that bring about positive change in clients.

First is to accept the client as he or she is, as a person of inherent worth possessing both positive and negative feelings and impulses. Rogers called this acceptance and prizing of the person 'unconditional positive regard.' Second is 'the therapist's willingness and sensitive ability to understand the client's thoughts, feelings, and struggles from the client's point of view … to adopt their frame of reference.' Third is to be 'genuine, or whole, or congruent in the relationship … It is only as [the therapist] is, in that re-

lationship, a unified person, with his experienced feeling, his awareness of his feelings, and his expression of those feelings all congruent or similar, that he is most able to facilitate therapy.'

When a counselor communicates this unconditional positive regard, empathic understanding, and congruence so that the client perceives them, the 'necessary and sufficient conditions for therapeutic personality change' are present. Rogers argued and demonstrated that clients have within themselves the ability and tendency to understand their needs and problems, to gain insight, to reorganize their personality, and to take constructive action. What clients need is not the judgment, interpretation, advice, or direction of experts, but supportive counselors and therapists to help them rediscover and trust their 'inner experiencing' (a concept borrowed from Gendlin), achieve their own insights, and set their own direction.

4. *Recording Therapy*

Rogers was able to demonstrate these propositions cogently, in part because he was the first person to record and publish complete cases of psychotherapy. This was a remarkable achievement before the invention of tape recorders, requiring a microphone in the counseling room connected to alternating phonograph machines in an adjoining room, which cut grooves in blank record disks that had to be changed every three minutes. With graduate student Bernard Covner, Rogers and his team recorded thousands of disks, involving scores of clients. These recordings were pivotal in the clinical training of psychotherapists which, in the 1940s, Rogers may have been the first to offer in an American university setting.

Beyond audio recording of therapy sessions, Rogers was among the first to make cinematic recordings of counseling and psychotherapy. The American Academy of Psychotherapists became a leading distributor of training tapes and movies, with Rogers the most frequent therapist portrayed. A still widely distributed set of training films showed Rogers, gestalt therapist Frederick Perls, and rationale–emotive therapist Albert Ellis each demonstrating his method with the same client.

5. *Scientific Research and Theory*

Recording actual therapy sessions also provided the data with which Rogers and his colleagues conducted more scientific research on one therapeutic approach than had ever been undertaken before. Rogers devised and used numerous instruments for measuring the variables of client-centered therapy and its outcomes: therapist acceptance, empathy and congruence; client expression of feelings, insight, self-concept, self-acceptance, self-ideal, level of experiencing; clients' positive actions, social adjustment, and numerous other variables. In 1959 the American Psychological Association awarded Rogers its first 'Distinguished Scientific Contribution Award' for developing an original method to objectify the description and analysis of the psychotherapeutic process, for formulating a testable theory of psychotherapy and its effects on personality and behavior, and for extensive systematic research to exhibit the value of the method and explore and test the implications of the theory.

As the award citation suggests, Rogers was interested in psychological *theory* and in the effects of therapy on *personality* as well as behavior. Building upon the Gestalt and phenomenological movements in psychology, and work of his students Victor Raimy and Arthur Combs, he developed a 'self-theory' of personality which is still included in many textbooks (Rogers 1959). The theory describes how an individual's concept of self emerges, how the process of socialization causes individuals to distrust their feelings and sense of self, how experiences which are inconsistent with the concept of self become denied and distorted causing personal distress and psychological problems, and how the therapeutic relationship can help the individual restructure the sense of self, allowing previously denied and distorted experience into awareness, leading to reduction in stress and openness to new experiencing.

6. *Impact on the Professions*

More than anyone, Rogers helped spread professional counseling and psychotherapy beyond psychiatry and psychoanalysis to other helping professions. While not the first to use the term 'client' for the recipient of therapy, he popularized it, emphasizing that 'a person seeking help was not treated as a dependent patient but as a responsible client,' and that those in psychological distress were not necessarily 'sick,' requiring treatment by medical specialists. Rather all people could be helped by the growth-producing conditions of empathic understanding, unconditional positive regard and congruence, and professionals from many fields could be trained to provide these conditions. Thus counselors, social workers, clergy, medical workers, youth and family workers, and other helping professionals could employ counseling methods and client-centered attitudes in their work. Rogers' books *Counseling and Psychotherapy* (1942) and *Client-Centered Therapy* (1951) described the principles of effective therapy and presented ample case studies from recorded sessions to illustrate his points. In 1982, surveys in the *Journal of Counseling Psychology* and *American Psychologist* still ranked Carl Rogers as the most influential author and psychotherapist.

Rogers also impacted the helping professions through leadership in many professional associations. He served on the executive committee of the American

Association of Social Workers, was Vice-President of the American Association of Orthopsychiatry (social workers), and President of the American Association of Applied Psychology, the American Psychological Association, and the American Academy of Psychotherapists, among other distinguished positions.

7. *Applications to Other Fields*

Through teaching, writing, and personal example, Rogers continually applied his theory and method to other fields—education, parenting, group leadership, and the health professions, to name a few. In each instance he demonstrated how the facilitative conditions of empathy, positive regard, and congruence could unleash growth, creativity, learning, and healing in children, students, group members, patients, and others. Applied to education, his views on 'student-centered learning' coincided with and contributed to the 'open education' movement in the USA, Great Britain, and elsewhere. His book *Freedom to Learn* (1969) went through two new editions over 25 years.

He expanded his theory and practice into the areas of group leadership and group facilitation over several decades. In the 1940s Rogers, Thomas Gordon, and colleagues had experimented with 'group-centered leadership,' whereby the leader's acceptance, understanding, genuineness, and willingness to let the group set its own directions stimulated great energy, creativity, and productivity among group members. In the late 1950s and 1960s, Gordon, Richard Farson, Rogers, and associates extended this approach to what Rogers called the 'basic encounter group'—an unstructured group experience in which 'normal' group members came to greater self-understanding, spontaneity, improved communication, and genuineness in relationships. Rogers led scores of encounter groups in professional, business, religious, medical, academic, personal growth, and organizational settings. *Look* magazine called Rogers an 'elder statesmen of encounter groups.' *Carl Rogers on Encounter Groups* (1970) was a major seller, and his and Farson's filmed encounter group, 'Journey into Self,' won an Academy Award (Oscar) for best full-length feature documentary.

Recognizing the ever-widening applicability of the client-centered, student-centered, group-centered approach, Rogers increasingly used a broader term, 'person-centered,' to describe his work. In the 1970s and 1980s he experimented with a person-centered approach to resolving intergroup and international conflict. Through workshops and filmed encounter groups involving different races, generations, and political persuasions—including Catholic and Irish protagonists from Northern Ireland—Rogers demonstrated how the same growth-promoting conditions useful in all helping relationships can enhance communication and understanding among antagonistic groups. He and colleagues led person-centered workshops for groups of 100 to 800 participants in Brazil, Mexico, South Africa, Hungary, Soviet Union, other newly emerging democracies, and other countries. They organized a gathering of international leaders in Austria, about resolving tensions in Central America, which vividly demonstrated the potential of the person-centered approach for resolving international conflict. Testimonials suggested these efforts in professional development and citizen diplomacy helped foster democratization in several countries.

8. *Humanistic Psychology*

Rogers was a leader in what Abraham Maslow described as a 'third force' in psychology, which became prominent in the latter half of the twentieth century. 'Humanistic psychology,' as it came to be known, differed from psychoanalysis and behaviorism in at least three ways. (See *Freud, Sigmund (1856–1939)*; *Psychoanalysis: Overview*; *Skinner, Burrhus Frederick (1904–90)*; *Behaviorism*; *Allport, Gordon W (1897–1967)*.) First, this psychology gave more emphasis and credence to the individual's *phenomenal field*—e.g., the client-centered therapist empathizes with the client's frame of reference rather than evaluating or diagnosing from the outside; the existential psychotherapist helps the patient find 'meaning' in his life. (See *Lewin, Kurt (1890–1947)*.) Second, this psychology focused not just on remediation of psychological problems, but on psychological *health*, *creativity*, *self-actualization*—what Rogers described as 'the fully functioning person.' The goal was more than 'adjustment,' rather helping people experience their full human potential. Third, it was a psychology interested in what distinguishes human beings from other species. Choice, will, freedom, values, feelings, goals, and other *humanistic* concerns were all central subjects of study.

As Rogers' career and that of leading behavioral psychologist B. F. Skinner were parallel in timing, productivity, and influence, their views were inevitably contrasted. Meeting on several occasions, including a six-hour debate–dialogue in 1962 (Kirschenbaum and Henderson 1989a), one of their exchanges on *Some Issues Concerning the Control of Human Behavior* (Rogers and Skinner 1956) became one of the most reprinted articles in the behavioral sciences. Rogers became a leading spokesperson for the humanistic psychology movement.

9. *Enduring Influence*

Major factors in Rogers' impact on psychology and the helping professions were his personal example, longevity, and prolific writings. He presented a vivid role model of the person-centered approach for almost

six decades, demonstrating his theories through teaching, lecturing, live demonstrations and workshops, and audio-visual recordings. By all accounts, he embodied his theories by being an exceptional listener and communicator and a decent, honorable person. He wrote 17 books and over 200 professional articles and research studies. Millions of copies of his books have been printed, in over 60 foreign language editions. His most popular book *On Becoming a Person* (1961), written for both a professional and general audience, continues to spread his ideas.

Critics of Rogers' work have argued (sometimes fairly, sometimes not) that client-centered therapy is superficial, unworkable with some populations, and unmindful of recent advances in behavioral, drug, or alternative therapies; that Rogers' views on human nature are unrealistically optimistic and underestimate human evil; that encounter groups and humanistic psychology have fostered widespread selfishness, narcissism, and moral permissiveness; and that Rogers' experiments with organizational change were naive and counter-productive. Nevertheless, in the USA, although no longer a central figure in popular psychology, Rogers' influence endures. In 1972, he was awarded the American Psychological Association's Distinguished Professional Contribution Award, becoming the first psychologist to receive the organization's highest scientific *and* professional honor. The citation read:

> His commitment to the whole person has been an example which has guided the practice of psychology in the schools, in industry and throughout the community. By devising, practicing, evaluating and teaching a method of psychotherapy and counseling which reaches to the very roots of human potentiality and individuality, he has caused all psychotherapists to reexamine their procedures in a new light. Innovator in personality research, pioneer in the encounter movement, and respected gadfly of organized psychology, he has made a lasting impression on the profession of psychology.

After Rogers' death in 1987, the greatest new interest in his work has been outside the USA, including Russia, Eastern Europe, and Latin America. As a Japanese counselor explained in the 1960s, Rogers helped 'teach me … to be democratic and not authoritative.' His life's work demonstrated how supportive, growth-producing conditions can unleash healing, responsible self-direction, and creativity in individuals and groups in all walks of life. As countries strive to resolve intergroup tensions and practice self-government and self-determination, many have recognized in Rogers' work not only useful methods for helping professionals, but a positive, person-centered, democratic philosophy consistent with their national aspirations. Rogers eventually recognized the political implications of his theories and methods and explored these in *Carl Rogers on Personal Power* (1977). At Rogers' memorial service, Richard Farson eulogized him as 'a quiet revolutionary.'

10. Personal Life

Upon college graduation, Rogers married childhood friend Helen Elliott (1901–79), an aspiring commercial artist. Their two children, David (1926–94) and Natalie (1928–), distinguished themselves professionally, David in medicine (e.g., Dean of Medicine, Johns Hopkins; President, Robert Wood Johnson Foundation) and Natalie in art therapy and 'client-centered expressive therapy.' For many years the Rogers spent the academic winter quarter in the Caribbean and Mexico, where Carl read widely and developed a number of his theories, Helen painted, and both snorkeled. They regularly traveled to visit their children and six grandchildren. Carl pursued life-long hobbies of photography, making mobiles, and gardening. When Helen became ill in her seventies, Carl cared for her until her death in 1979. Thereafter he remained involved in work, cultivated old and new friendships with both men and women, and traveled widely. He was active until his death at age 85, on February 4, 1987 in La Jolla, California.

See also: Counseling Psychology; Group Psychotherapy, Clinical Psychology of; Interpersonal Psychotherapy; Person-centered Psychotherapy; Psychology: Overview; Psychotherapy, History of: Psychiatric Aspects; Psychotherapy Process Research

Bibliography

Hart J T, Tomlinson T M 1970 *New Directions in Client-Centered Therapy*. Houghton Mifflin, Boston

Kirschenbaum H 1979 *On Becoming Carl Rogers* [a biography]. Delacorte Press, New York

Kirschenbaum H, Henderson V L (eds.) 1989a *Carl Rogers: Dialogues*. Houghton Mifflin, Boston [Conversations with Buber, Tillich, Skinner, Polanyi, Bateson, May and Niebuhr]

Kirschenbaum H, Henderson V L (eds.) 1989b *Carl Rogers Reader*. Houghton Mifflin, Boston

Lietaer G, Rombauts J, Van Balen R (eds.) 1990 *Client-centered and experiential psychotherapy in the nineties*. Leuven University Press, Louvain, Belgium

Rogers C R 1931 *Measuring Personality Adjustment in Children Nine to Thirteen Years of Age*. Ph.D. thesis

Rogers C R 1939 *Clinical Treatment of the Problem Child*. Houghton Mifflin, Boston

Rogers C R 1942 *Counseling and Psychotherapy: Newer Concepts in Practice*. Houghton Mifflin, Boston

Rogers C R 1951 *Client-Centered Therapy*. Houghton Mifflin, Boston

Rogers C R 1959 A theory of therapy, personality, and interpersonal relationships as developed in the client-centered framework. In: Koch S (ed.) *Psychology: A Study of a Science, Vol III*. McGraw-Hill, New York

Rogers C R 1961 *On Becoming a Person*. Houghton Mifflin, Boston

Rogers C R 1969 *Freedom to Learn: A View of What Education Might Become*, 1st edn. C. E. Merrill, Columbus, OH

Rogers C R 1970 *Carl Rogers on Encounter Groups*, 1st edn. Harper & Row, New York

Rogers C R 1977 *Carl Rogers on Personal Power*. Delacorte Press, New York
Rogers C R, Dymond R F 1954 *Psychotherapy and Personality Change*. University of Chicago Press, Chicago
Rogers C R, Skinner B F 1956 Some issues concerning the control of human behavior. *Science* **124**(3231): 1057–66

H. Kirschenbaum

Rokkan, Stein (1921–79)

Stein Rokkan was one of the world's leading social scientists after World War II. He played a unique role in organizing international social science cooperation in the 1960s and 1970s, and was the driving force in the development of infrastructure services for comparative research. But he was also a prolific writer, and it was the combination of all his contributions to comparative research, international cooperation, and infrastructure development that made him the comparativist of his time.

1. Stein Rokkan, the Norwegian

Rokkan was born July 4, 1921 in Vågan on the Lofoten Islands, on the northern periphery of Norway, and grew up in Narvik. Between 1938 and 1948 he attended the University of Oslo, where he first studied philology and then political philosophy, in which he received his doctorate with a dissertation on David Hume. From 1948 to 1951 Rokkan continued his studies in the USA and England: at Columbia University, where his work with Paul Lazersfeld acquainted him with the methods of modern social research, and at the London School of Economics, where he met T. H. Marshall, whose historical-sociological research profoundly influenced his own.

Rokkan's educational career already contained the seeds of a rare combination of highly diverse talents: above all his great openness towards the 'new' social sciences while maintaining a firm grounding in the 'old' cultural sciences. His extensive linguistic, philosophical, and historical studies gave him a keen eye for the significance of cultural differences in international comparison, and thus for the often misleading simplicity of purely quantitative approaches. At the same time he recognized the potential of such approaches in comparative research and became one of their most important supporters. Rokkan's thought was interdisciplinary from early on. His scholarly interests included the entire spectrum of the social sciences, and this characterized his core scientific work: a political sociology in the tradition of classic macrosociology, the historical-comparative analysis of whole societies and polities.

From the very beginning, Rokkan combined a genuine international orientation with a deep devotion to his homeland that shaped his comparative approach. From his intimate knowledge of Norway he developed concepts (e.g., the center–periphery dichotomy) and questions (e.g., on the interaction of mass movements and institutional change) that he then applied to comparative research. This made him aware of the complexity of each single 'case' and led him to assess realistically the possibilities and limits of international comparisons.

Rokkan developed his system of thought, which was rooted in his origin and education, systematically throughout his entire scholarly career until his early death on July 22, 1979 in Bergen. From 1951 to 1958 he worked at the Institute of Social Research in Oslo, where he soon became engaged in various comparative research projects. In the same period he initiated, together with Henry Valen, the Norwegian Program of Electoral Research which was oriented on the work of the Survey Research Center in Ann Arbor, Michigan, but already showed Rokkan's signature in its emphasis on context variables.

In 1958 Rokkan became research professor at the Christian Michelsen Institute in Bergen, where in 1966 he was awarded a university chair and founded a new department of comparative politics. During this time he was also engaged in efforts to increase cooperation between the Scandinavian countries, which resulted in the publication of *Scandinavian Political Studies* since 1966, and in the founding of the Nordic Political Science Association in 1975.

2. Rokkan and the Internationalization of the Social Sciences

Rokkan was active at the international level from the very beginning of his professional career. Developing international research programs, organizing international projects and conferences, and building institutions and infrastructures for comparative research became the focus of his activities from the early 1960s. Three contexts of these activities can be distinguished: UNESCO and the International Social Science Council, the International Committee on Political Sociology, and the European Consortium for Political Research.

Rokkan, then Secretary-General of the International Sociological Association, participated in the founding of the International Social Science Council (ISSC) by UNESCO in 1952. In 1961, he proposed to the ISSC a broad program in cross-national and cross-cultural research which he then directed for over 15 years. He became a member of the Executive Committee of the ISSC in 1965 and served as president of the ISSC from 1973 to 1977.

The 1961 program led to a broad spectrum of activities to develop infrastructure services for com-

parative research and to promote concerted interdisciplinary research. Whole series of conferences were organized on comparative survey research and aggregate statistical comparisons of nations in the first half of the 1960s. Some of the further work was channeled into the European Coordination Centre for Research and Documentation in Social Sciences, set up in Vienna in 1963; other activities were headed up by two ISSC standing committees for data archives and comparative research established in 1966–7. Rokkan remained the driving force behind most of the subsequent activities, which in the 1970s also included summer schools and workbooks for training on comparative analysis and the promotion of an International Federation of Data Organization.

In the late 1960s, with the support of UNESCO Rokkan launched another series of conferences on national developments, asking what conclusions could be drawn from the early European experiences of state formation and nation-building for developmental processes outside Europe. Rokkan here developed his basic ideas about a macromodel of European political development. This work was also linked to the International Committee on Political Sociology that Rokkan, together with Shmuel Eisenstadt, Morris Janowitz, and Martin Lipset, had founded in 1960 and for which he served as Secretary from 1960 to 1970. Originally affiliated only with the International Sociological Association, in which Rokkan was vice-president from 1966 to 1970, it became a joint ISA/IPSA committee in 1970 when Rokkan was elected president of the International Political Science Association. During the 1960s the committee was an exclusive group of international scholars with a common interest in studying major historical changes. It was in this second context that Rokkan developed his ideas on democratization, cleavage structures, and party systems.

The third context was the European Consortium for Political Research. Rokkan was one of its founders and its first Chair from 1970 to 1976. Here too he pursued his central goals of developing joint research programs and improving infrastructure in the broadest sense: from bibliographic tools, research documentation, and data archives to computer applications (including cartography), and the training of the next generation of social scientists.

Rokkan's decisive role in the internationalization of the social sciences and the promotion of comparative research resulted from a congruence of personality and historical situation. In the 1960s the social sciences were expanding rapidly in Western Europe, but they suffered from national fragmentation. International cooperation was needed, and succeeded first via the USA. Only later did a genuine 'Europeanization' begin. Rokkan's broad education and international orientation proved invaluable in building bridges between the USA and Europe and reconciling different European traditions. His enormous talent for combining personal networking and institution-building was a crucial asset in this start-up phase and the close linkage he saw between programmatic conceptualization, cooperative organization, and infrastructure-building gave this endeavor a cohesiveness that seems to be lacking today.

Rokkan's historical role has somewhat overshadowed his scientific contributions, apart from a few famous single hypotheses and typologies. His lasting contribution will be his attempt to develop a macro-model of the political development of Europe. In particular, his comparative studies on the structuring of European mass democracies, and his analyses of the processes of state formation and nation-building in Europe, but also his general approach to comparative research.

3. *The Structuring of Mass Politics in Western Europe*

Rokkan's work on the democratization of the European nation-states concentrated on the institutional and organizational variations in the structuring of mass politics up to World War II. Here he developed his well-known ideas of four institutional thresholds of democratization and of historical cleavage structures which in the process of democratization are transformed into party systems via the formation of alliances and oppositions.

In institutional terms, democratization for Rokkan meant the successive lowering or removal of four thresholds: the threshold to political opposition (legitimation), to political participation (incorporation), to access to parliament (representation), and to participation in government (executive power).

The lowering of the first two thresholds, through the recognition of civic rights and the extension of suffrage, ushered in the age of mass politics with mass parties and mass election campaigns. For Rokkan, an unbroken tradition of representation, a history of continuous center building, and a Protestant nationalization of territorial culture were the main factors favoring an earlier, more gradual, and stable democratization.

The lowering of the third threshold implied a transition from majoritarian to proportional representation, easing the access of minorities to representation in the legislature. In his view, the pressures for proportional representation were higher in the ethnically and/or religiously heterogeneous countries, and in general in the smaller nations with their greater international dependence.

The introduction of universal (male) suffrage in all Western European countries at the end of World War I and the transition to proportional representation in most of them tended to stabilize the party systems of the 1920s for a long time, in most cases until the 1960s (Rokkan's famous 'freezing' thesis). In Rokkan's view,

the party systems of the 1920s reflected older cleavages that had emerged in the earlier processes of state formation and nation-building and became politicized and organized in the process of democratization.

Cleavages are understood as strong and enduring conflicts that tend to polarize a political system. Rokkan distinguished between four or five basic cleavages that emerged at different critical junctures in European history. The *National Revolution* (1789 and after) produced two cleavages, with even older roots in the early state formation processes and the Reformation:

(a) the *center–periphery cleavage*: conflicts between the central nation-building culture and ethnically linguistically distinct subject populations in the peripheries;

(b) the *state–church cleavage*: conflicts between the aspirations of the centralizing, standardizing, and mobilizing nation-state and the corporate claims of the church.

Two further cleavages were produced by the *Industrial Revolution* in the nineteenth century:

(c) the *land–industry* cleavage: conflicts between the landed interests and the rising class of commercial/industrial entrepreneurs; and

(d) the *worker–employer* cleavage: conflicts between owners/employers and tenants/workers.

The class cleavage was overlaid by a fifth cleavage between proletarian internationalism and nation-accepting socialism as a result of the *International Revolution* (1917 and after).

According to Rokkan, the differences between the European party systems largely were produced by the first three cleavages, reflected in the presence or absence of regional, Christian, or agrarian parties. Whereas Christian parties emerged mainly in the religiously mixed and some of the Catholic countries, agrarian parties are found only in Protestant areas. The class cleavage, by contrast, produced working-class parties in every country, but their strength and unity differed greatly due to the socialist–communist split.

4. *State Formation and Nation-building in Europe*

In analyzing the historical development of the different cleavage structures, Rokkan took a decisive step beyond traditional comparative research: he no longer regarded the units of comparison, the nation-states, as given, but rather thematized their development itself. In these later works, Rokkan was influenced greatly by Hirschman's concepts of exit, voice, and loyalty, which he used in his analyses of the development of the nation-state in terms of boundary-building, internal structuring, and system-building.

Rokkan distinguished between four basic processes of loyalty-producing system-building: state formation and nation-building; two processes mainly of boundary-building and two processes of mainly restructuring, the establishing of political citizenship by equalizing the rights to participation and the establishing of social citizenship by redistributing resources and benefits. In the older countries these processes tended to follow each other in time, whereas in the newer countries the problems associated with them often cumulated.

In the European context, according to Rokkan, these processes were influenced by six givens: first, the heritage of the Roman Empire (law and citizenship); second, the tradition of the Germanic kingdoms (legislative/judicial assemblies); third, the crucial role of the supraterritorial Catholic Church; fourth, the growth of a central city belt; fifth, the development of feudal and manorial agrarian structures; and sixth, the emergence of literatures in vernacular languages.

The European world of states and nations emerged from the disintegration of the Western Roman Empire, after centuries of territorial fragmentation and retrenchment, followed by processes of expansion and reorganization. For Rokkan, 'the alphabet and the city decided the fate of Europe' (Rokkan 1999) in these processes. As a result of various waves of migration, a new 'ethnic-linguistic infrastructure' arose in the Middle Ages. The alphabet made it possible to turn spoken into written standards of communication, and this vernacularization provided a basis for the later nation-building.

The city belt in turn was a decisive factor in center formation and early state building. The revival of trade between the Orient, the Mediterranean, and the North Sea had produced a belt of independent cities from Italy to Flanders. The polycephality of the city-state Europe either delayed state formation (Germany, Italy) or led to confederations (The Netherlands, Switzerland). By contrast, monocephalic city structures conducive to centralized state development formed at the western and later also at the eastern end of the city belt. The rise of Atlantic capitalism, however, brought with it a European division of labor that strengthened the Western states, while the Eastern states largely became incorporated into empires.

The variations in city structure and thus the tripartite division of Europe into a Western Europe of unitary, centralized nation-states, a federalist Central Europe, and an Eastern Europe of multi-ethnic empires constitute the 'west–east axis' in Rokkan's well-known typological–topological model or conceptual map of Europe. It helps to explain the great paradox of European development: that the strongest and most durable systems emerged at the periphery of the city belt and the old Empire.

While the 'west–east axis' defines differences in economic conditions of state formation, the 'south–north axis' denotes differences in cultural conditions of nation building, depending on the geopolitical distance northward from Rome. Rokkan interpreted the Reformation as the first major step in defin-

ing territorial nations in Northern Europe, whereas in Catholic Europe the Church remained supra-territorial.

The conditions for successful center-building and territorial consolidation as well as for cultural standardization and identity-building, therefore, varied across Europe in a typical and topological way. The easier these processes of territorial and cultural boundary-building, Rokkan generalized, the more gradual and less violent were the internal restructuring of the nation-states through the establishment of citizenship rights.

Towards the end of his life Rokkan attempted to integrate his comparative studies into a comprehensive model of Europe spanning the entire history of state formation, nation-building, and mass politics in Europe and identifying the crucial variables in the long and complex process that led up to the current constellations of territories, economies, and systems of political alignment. This macro-model remained a sketch exemplifying the nature of most of Rokkan's studies, which were essentially research programs.

5. *Rokkan's Comparative Approach*

Rokkan defined the decisive challenge of cross-cultural, cross-societal, and cross-national research as follows: 'We propose to direct our attention to one single, central set of issues: the possibilities of translating "grand theory" into empirically workable "typologies of macro-settings" for variations in human behaviour, and the consequences of such typologies for decisions on the cultural and the geographical range of comparisons at the level of communities, households, and individuals' (Rokkan 1970b).

From this fundamental orientation, Rokkan developed a series of basic elements that characterize his comparative approach: the need for context-specific comparisons, the multi-dimensionality of contexts, the twin concepts of structuring and boundary-building, the dynamics of freezing and unfreezing at critical junctures, and finally the method of retrospective diachronics.

For Rokkan model-building always had to be context-specific: 'No single explanatory or intervening variable can be linked up with a dependent variable in isolation from the context, whether across systems or across stages' (Rokkan 1999). Because such macro-contexts are comparable only to a certain degree, Rokkan was skeptical of universalizing comparisons and instead preferred region- and period-specific models, particularly in his cleavage model and his conceptual map of Europe.

The macrocontexts are in principle multidimensional, since for Rokkan they always resulted from interactions between three different systems, the political, economic, and cultural systems. Due to their relative autonomy, he started in his macromodel from configurations, rather than hierarchies of factors: 'There is no economic determinism in the model, nor a geopolitical, nor a cultural: in this sense it seeks to combine the traditions of Karl Marx with those of Max Weber and Emile Durkheim.'

While all three systems have a territorial dimension, only the political system is ultimately capable of controlling territory. Territory is a crucial concept in Rokkan's comparative analysis of political systems. It is a starting point for his most basic twin concepts of structuring and boundary-building in the formation of territorial systems. Such systems essentially are structured through the formation of political, economic, and cultural centers and the incorporation of peripheries. Their consolidation presupposes a process of political, economic, and cultural boundary-building which in turn contributes to the emergence of territory-wide cleavages. Center–periphery relations and cleavage structures are the two dimensions that for Rokkan underlay the institutional and organizational structuring of political systems.

With his twin concepts, Rokkan went beyond traditional comparative research that was limited to comparing structures and focused on the formation of systems as a process of structuring and boundary-building. In this historical process critical junctures are crucial. This concept of Rokkan's refers to political, economic, and cultural 'revolutions,' historical events and processes which he used to distinguish between periods: between 'critical junctures' when basic new decisions are taken, structures are 'unfrozen' and boundaries are 'opened,' and periods in which the institutional and organizational structures remain 'frozen' and the boundaries 'closed.' This attempt at periodization underlies Rokkan's basic method of retrospective diachronic analysis which addresses the question: 'given an observed contrast in the values of variables at time t_i, what combinations of variables for earlier phases t_{i-t}, t_{i-2}, and so on, can best account for these differences?' (Rokkan 1999).

In this way, Rokkan attempted to explain systematically current structural differences between the European societies and polities as arising from their long-term development. In many cases, his efforts to do so took the form of research programs. With his contributions to developing a social science infrastructure, he wished to create a basis for carrying out these programs. In both respects, Rokkan has left us an important legacy.

See also: Comparative History; Comparative Studies: Method and Design; Data Archives: International; Human–Environment Relationship: Comparative Case Studies; International Communication: History; International Research: Programs and Databases; International Science: Organizations and Associations; Science and Technology: Internationalization

Bibliography

Lipset S, Rokkan S (eds.) 1967 *Party Systems and Voter Alignments: Cross-national Perspectives*. Free Press, New York

Rokkan S 1970a *Citizens, Elections, Parties. Approaches to the Comparative Study of the Processes of Development*. Universitetsforlaget, Oslo, Norway

Rokkan S 1970b Cross-cultural, cross-societal and cross-national research. In: *UNESCO, Main Trends of Research in the Social and Human Sciences*. Mouton, Paris, p. 1

Rokkan S 1999 *State Formation, Nation-building, and Mass Politics in Europe. The Theory of Stein Rokkan*. Oxford University Press, Oxford, UK

Rokkan S et al (eds.)1987 *Centre–Periphery Structures in Europe: An ISSC Workbook in Comparative Analysis*. Campus, Frankfurt, Germany

Rokkan S, Dogan M (eds.) 1969 *Quantitative Ecological Analysis in the Social Sciences*. MIT Press, Cambridge, MA

Rokkan S, Eisenstadt S (eds.) 1973 *Building States and Nations*, 2 Vols. Sage, Beverly Hills, CA

Rokkan S, Merritt R (eds.) 1966 *Comparing Nations. The Use of Quantitative Data in Cross-national Research*. Yale University Press, New Haven, CT

Saelen K 1981 Stein Rokkan: A bibliography. In: Torsvik P (ed.) *Mobilization, Center–Periphery Structures and Nation-building*. Universitetsforlaget, Bergen, Norway

P. Flora

Romanticism: Impact on Social Thought

Romanticism was a broad intellectual movement within modern Enlightenment that aimed at bringing the arts together with philosophy, history, and religion and attempted to aesthetisize life. The central objective was a philosophico–political reconciliation of the fundamental connectedness of life, which was seen as damaged in the present, with the individual, with love, and with beauty. Although it was a wide European movement, it was in particular German romanticism that significantly influenced art and literature as well as the philosophy, psychology, and history of modernity. The key participants in the circles of Berlin and Jena as well as of Heidelberg, as the most important groups, were August Wilhelm Schlegel (1767–1845), Dorothea Caroline Michaelis/Schlegel/Schelling (1763–1809), Friedrich Schlegel (1772–1829), Brendel Mendelssohn/Dorothea Veit/Dorothea Schlegel (1763–1839), Friedrich von Hardenberg (Novalis, 1772–1801), Friedrich Schleiermacher (1768–1834), Rahel Antonie Levin Markus/Varnhagen von Ense (1771–1833), Henriette Herz (1764–1847), Karoline Günderrode (1780–1806), Sophie Schubart/Merau /Brentano (1770–1806), Clemens Brentano (1778–1842), Bettina Brentano/von Arnim (1785–1859), Achim von Arnim (1781–1831), Friedrich Wilhelm Joseph Schelling (1775–1854), Johann Gottlieb Fichte (1762–1814), in the broader environment Friedrich Hölderlin (1770–1843); as well as Johann Joseph Görres (1776–1848), Jacob Grimm (1785–1863), and Wilhelm Karl Grimm (1786–1859).

The impact of romanticism on the emerging social sciences can be demonstrated across the following areas: (a) knowledge and hermeneutics; (b) culture and society; (c) critique of modernity; and (d) reason and aesthetics.

1. Knowledge and Hermeneutics

Whereas Enlightenment thought starts out from the assumption of the intelligibility of the world and lets its knowledge of the world predetermine its view of the arts, romanticism turns this relation around and—with reference to Kant's critique of reason—denies philosophy's primacy with regard to knowledge. Science (*Wissenschaft*) was to be 'beautiful'; analogies and metaphors were to enrich and complement scientific-scholarly practices; and the 'novel of life,' in which all arts and sciences unite, should unfold from an ever-open, progressive concept of form. 'The absolute,' 'freedom,' and 'nature' become key concepts of philosophical debate. Within such core of romanticist perspectives, however, important differences and contradictions co-exist. Friedrich Schleiermacher, for insistence, turns against Friedrich Schlegel's progressive universal poetics (*progressive Universalpoesie*) and insists on the boundaries and relative separateness of the diverse aspects of reality (religion, science, art, aesthetics). In addition, the new mythologization appears to him as an unwarranted intervention of human action into the absolute. He contradicts Schlegel's characterization of infinity with its inclination towards dissolving ethics, since under conditions of achieved finite freedom the autonomy of the other needs to be respected. Even though he rejects any ahistorical contractual thinking, he thus combines an ethics of obligation with an ethics of virtue.

Within this framework, the task of hermeneutics according to Schleiermacher (see *Hermeneutics, History of*) emerges from the following movement of thought. If thinking occurs in intersubjective individuality, then there needs to be an art of understanding, which compares that which is one's own critically with that which is of the other. Thus, such art can approach the infinite spirit that is always already at work in everything individual. Schleiermacher's hermeneutics—which can well be seen as part of an aesthetics, of a 'theory of artistic activity' (Schleiermacher 1984, p. 154)—investigates writing and speech under the following aspects: with regard to general grammar, in psychological-subjective perspective with regard to the genesis of authorship and work, by comparative procedure and, finally, by divinatory procedure. Understanding occurs in a historical process that encompasses both reader and author. Although understanding aims at agreement, this process can nevertheless not be closed (see Frank

1977). Not least Walter Benjamin (1991) has demonstrated that it is only the reader as the extended author who completes the work of art. Thus, he showed that the romantic way of 'raising reflection in the work to a higher power' contained a 'more modern' theory of art than can be found in Goethe. The impact of hermeneutics continued through Wilhelm Dilthey to Hans-Georg Gadamer (Ferraris 1998).

Beyond hermeneutics, theology, the study of religion, classic studies, and philology, psychology and psychoanalysis show signs of a romantic heritage. Early romanticism showed a particular interest in phenomena at the margins of reason, such as magnetism, occultism, somnambulism, and telepathy (Johann Wilhelm Ritter 1776–1810, founder of electrical chemistry). Late romanticism—Gotthilf Heinrich Schubert (1780–1860)—was concerned with the analysis of the human unconscious and systematic studies in the symbolism of dreams as a linguistic system (*Symbolik des Traums*, 1814), in particular, can be considered as predecessors to Freud and his cultural criticism (*Das Unbehagen in der Kultur*, 1912). Themes of the split I (the motif of the *Doppelgänger*), trauma, loss of identity, melancholy, and delusion as well as, in particular, the imagery world of sentiments in, for example, the works of Johann Ludwig Tieck, Jean Paul, Heinrich von Kleist, or E. T. A. Hoffmann were well known to Freud.

2. *Culture and Society*

There is a clear impact of romanticist thinking on Jacob Grimm and Wilhelm Grimm (1785–1863 and 1778–1859), whose historically and empirically comparative research on language opened the space for the identity-constituting role of the emerging *Germanistik* (German studies) and at the same time already separated the natural sciences from the human sciences.

Romanticist impact, though, is particularly evident in that problem area that is currently discussed under the denomination of 'identity' (Friese 2001). Johann Gottfried Herder (1744–1803; *Über den Ursprung der Sprache*, 1772; *Ideen zur Philosophie der Geschichte der Menschheit*, 1784–1791), whose work connects Rousseau to romanticism, already meant what is now called 'collective identity' when he spoke of the inalienable 'character' (*Geburtsstamina*) of a people. Following up on Herder and relating to motifs in Friedrich W. Schelling—motifs that found their continuation in the theory of archetypes as developed by C. G. Jung and Mircea Eliade, Aby Warburg but also in Ruth Benedict's work—the search was for universal and constant features of culture that found regionally diverse expressions. These concepts have in common that they postulate a universalism in terms of the laws of the human psyche on the one hand, while working with a particularism in terms of the variety of different 'identities' of peoples on the other. The assumption of 'collectively shared' representations was also one of the conceptual pillars of the 'psychology of peoples' (*Völkerpsychologie*) and of the emerging sociology. It was Wilhelm Wundt who provided the most comprehensive view of psychology and the cultural sciences. His late writings (*Grundrisse der Psychologie*, 1898, *Völkerpsychologie*, 1917) focused on those psychic phenomena that are significant for the social life of human beings and that are at the roots of community and of the emergence of creations of the collective spirit. The reference to Herder's 'spirit of the people' (*Volksgeist*), which describes the supra-individual features of the members of a people, is as evident as the one to Wilhelm von Humboldt whose comparative linguistic investigations entailed a comparison of 'people's characters.' The impact of Wundt—whose writings were known by Franz Boas, Bronislaw Malinowski, but also by Emile Durkheim—extended as far as Lucien Lévy-Bruhl.

Next to the emphasis on linguistic and psychical factors for the formation of identities, the distinction between *Gemeinschaft* (mostly translated as 'community') and *Gesellschaft* ('society'; see Schleiermacher, *Theorie des geselligen Betragens*, 1799) as well as the particular relations between society and state (Fichte, *Beiträge zur Berichtigung der Urtheile des Publikums über die französische Revolution*, 1793), that were established in romanticism and reaching from republicanism to a thinking in terms of natural rights to a skepticism towards the state, lived on in the emerging social sciences (see *Community/Society: History of the Concept*; *State and Society*). Even though they were considerably altered, these conceptual oppositions had an echo in virtually all dualisms between 'traditional community' and 'modern society' that were established by 'classical' sociology and maintained from Max Weber, Georg Simmel, Emile Durkheim onwards to the present day.

3. *A Modern Critique of Modernity*

The crisis of the modern subject was made a central literary topic in romanticism (and was continued, *inter alia*, in Robert Musil or Franz Kafka). Friendship and love (Schlegel 1988) were counterposed to modern loneliness, loss of self, and the break of the individual with the icy machinery of society (as presented in Ludwig Tieck or E. T. A. Hoffmann). At the same time, forms of sociality were theorized as well as practiced in which an unadapted, anticonventional subjectivity mediated between the romantic individualism and the freedom of the subject on the one hand and the exchange with others on the other. Such forms of sociality were considered as the ideal of education (*Bildung*) in a comprehensive humanistic sense. Romanticism thus created an intellectual space in which women in particular, who were excluded from the university as an institution, occupied a central place.

The crisis of the modern subject, the romantic-enlightened cultural criticism and its dream of the 'noble savage' were in the background of the emerging field of anthropology (Georg Friedrich Creuzer, 1771–1858) and in the 'imaginary ethnography' of the nineteenth century (Kramer 1977, Said 1979). It continues to be effective as the escapism into the exoticism of the other as well as in the current constructions of (cultural) identity and 'difference' in the era of 'globalization.' The modern critique of modernity, as prefigured by romanticism, is taken up in a variety of quite different strands of thought. Karl Marx states the human alienation under conditions of capitalism; Friedrich Nietzsche celebrates the 'great separation' against both Christendom and bourgeois morality; Max Weber condemns the 'iron cage', Martin Heidegger the being 'fallen to the they' (*Verfallenheit im Man*); Theodor W. Adorno denounces the administered world and the culture industry and Walter Benjamin laments the 'loss of aura'. The (social) philosophy of the early twentieth century reclaims 'authenticity' and 'emancipation' and all these perspectives share in the critique of the Modern world as developed in romanticism and they point to a—indeed often denied—romantic heritage (in which, in particular, the connection between aesthetic experience and renewed religion, characteristic of early romanticism, is visible; see *Eudaemonism*). The romantic critique of modernity, however, is an integral moment of the Enlightenment, and not its subjective-irrational counterpart. This latter verdict has often been cast, and peculiar intellectual alliances can be found sharing it. It stretches from Hegel and the Left-Hegelians to Heinrich Heine's rejection of romanticism to Nietzsche and Carl Schmitt's critique of 'romanticism as subjective occasionalism' as well as to George Lukács. It finally culminates in the reproach that the modern-bureaucratic mass annihilation is a result of the romantic critique of modernity—a modernity that precisely created such horror.

4. Reason and Aesthetics

For August W. Schlegel, formulating the program of romanticism in his lectures on 'arts and literature' (*Schöne Kunst und Literatur*, 1801/04), reason is founded not just on empirical perception, and the autonomous validity of poetry is thus emphasized. Romanticism, however, never abandons the concept of reason. Even when the technical-instrumental use of reason against nature is opposed, a comprehensive concept of reason is sustained. Despite all criticism of idealist systemic thinking, romanticism moves within the path of Kantian philosophy, behind which there is no falling back, and it shares in Kant's rejection of the utilitarian grounding of morality in selfishness. Schlegel knows perfectly well the distinction between concepts of the intellect, which refer to the world of appearances, and concepts of reason, which refer to ideas and states that since ideas can be judged only by a reason that has enlightened itself about the limits of the possibility of knowledge.

In this context, early romanticism reacted to a situation in which the Enlightenment attempts to deal with the question of religion were difficult to sustain. Kant's critique made the argument for belief on the basis of reason impossible. Lessing intervened into the enlightened criticism of the bible by publication of the Reimarus papers; the dispute over this inheritance culminates in F. H. Jacobi's reproach of Spinozism (The dispute over atheism, 1789–99). Fichte justified his 'Attempt at a critique of revelation' (*Versuch einer Kritik der Offenbarung*, 1792) with the argument that revelation constrains the moral freedom of the human being. For Schleiermacher, the bible is a historical expression in the ever-incomplete path of Christendom, in the infinite perfectability of the world.

At the same time, romanticism addresses the problem of Enlightenment after the unity of reason and the tangible world could be cast only as the telos of action in the sense of a regulative idea, which never becomes real. Kant's critique of reason had demonstrated that the claim that nature is the bearer of moral norms could not be justified. Nevertheless, human action in the process of history can provide experience as to whether reality is constituted by the actualization of regulative ideas. Such a concept of Enlightenment in which the relation between ideas and the world of appearances can be sensually experienced in the course of an infinite rapprochement has been expressed by Novalis and F. Schlegel by means of infinitesimal calculus, which lets the tendency towards the infinite be grasped within finitude.

A different path towards the unity of idea and world of appearances is provided by imagination and the arts. Fichte described productive imagination as 'the faculty that is suspended in the middle between determination and nondetermination, between the finite and the infinite' (*Grundlage der Wissenschaftslehre*, 1794, II, 360). The objective is here not reconciliation, but rather the insistence on remaining within a space of contradictions.

Fichte exerted decisive influence on Friedrich Schlegel's aesthetic conceptions such as the 'progressive universal poetics' and infinite irony as poetic reflection (see *Imagination: History of the Concept*). Such poetcs can 'remain suspended in the middle on the wings of poetic reflection and free of all real or ideal interest, it can ever again produce such reflection and multiply it as in an infinite row of mirrors; […] the romantic art of poetry is still in its becoming; it is even its authentic being that it can ever only become and can never be completed' (F. Schlegel 1988, II, pp. 114–15; *Athenäums–Fragment*, p. 116; see also *Lyceum-Fragment*, p. 108). The concept of infinite irony reassumes such a moment of infinite reflexivity, a reflexivity that is not limited by any fixation. Hegel—in the Phenomenology of Spirit and in the

Aesthetics—attacked such conception, just as he criticized the transfer of Fichte's philosophy into the arts.

F. Schlegel referred here to Socratic irony, as presented by Plato (see Plato, *Symposium*, 216e), which according to Quintilian (*Institution oratoria*, IX, 2,44) brings the opposite of what is said to understanding. Irony is, in a dialogical reference, a double negation. It reflects on the possibility of things being different from what they seem to be and thus opens an 'ironic space of play' (Allemann 1956, p. 10). In contrast to allegory and metaphor (see *Metaphor and its Role in Social Thought: History of the Concept*), irony demands a principled and reflective understanding of the ambiguity what is meant and what is said, of essence and appearance. Friedrich Schlegel develops irony in particular in the Lyceum and Atheneum fragments and explains it in his essay 'On unintelligibilty' (*Über die Unverständlichkeit*, 1800): 'The French Revolution, Fichte's *Wissenschaftslehre* and Goethe's *Meister* are the grandest tendencies of the age' (Fragment 216)—'I wrote this fragment with the most honest intention and almost without any irony' (237). The 'system of irony' (239)—in the plural forms of irony, from the raw to the fine, dramatic and double irony to the 'irony of irony' (239)—is turned here from a rhetorical trope into a basic element of philosophical reasoning. Irony becomes the infinite play back and forth between opposite moments, none of which can be led into univocality. Infinite irony is thus distanced from any kind of identity or mediation—as, for instance, in the sense of Hegel's dialectic. When 'idea' is characterized as 'a concept completed to the level of irony […] the change, ever again recreating itself, between two thoughts in dispute' (Fragment 121, 115), then it is not a conceptually mediated identity that is at issue, but the persistent creation of difference. 'Irony of irony,' however, contains infinite regress, namely 'when one speaks without irony about irony, as it has just been the case' (239), thus when irony can no longer be mastered, 'when one is no longer able to step outside of irony, as it appears to have been the case in this essay about unintelligibility; […] when irony, so to say, ironizes the poet […] when irony has to be made against one's will […] when irony turns wild and lets itself no longer be governed' (239–40).

Schlegel's writings on irony have been received as well as criticized by, among others, Schelling, Schleiermacher, Novalis, Karl Wilhelm Solger (1780–1819), and Søren Kierkegaard (1813–55). Even though Solger (*Vorlesungen über Ästhetik*, 1819) remains with the negativity of irony that excludes absolute mediation, irony, originally thought of as infinite, is turned by him—and by Hegel—into a dialectically conceived moment of artistic activity. Following up on both Solgers and Hegel, Kierkegaard (*Über den Begriff der Ironie mit ständiger Rücksicht auf Sokrates*, 1841) regarded the infinite irony of romanticism as a subjective moment, seeing itself capable of separation from the positivity of reality. 'For irony, everything turns into nothing' (Kierkegaard 1993, p. 263)—and therefore Kierkegaard counterposes a controlled, limited irony to the infinite romantic irony.

Although the recasting of romantic irony thus appeared to have reached its end, recent attempts at rehabilitating irony refer back to this heritage, multiply broken and rewritten as it may be. Richard Rorty (1989, p. 123) attempted 'to cleanse romanticism of the last traces of German idealism'—but 'he overlooks that German romanticism has no need for him, since the inventor of theoretical irony had already taken his distance from early German idealism himself' (Bohrer 2000, pp. 11–12). Rorty (1989, p. xv) then designs a liberal society in which irony has become universal, thus reducing irony to a plain liberal principle. The idea here is that a plural society with its variety of life-projects, language-games, and contingencies can no longer be determined or mastered from an allegedly superior viewpoint.

Beyond such political thinking, the heritage of romanticism has a particular current significance in theories of language. The determined opposition between what is meant and what is said, between the authentic and the inauthentic is, once it has been separated from its foundation in transcendental philosophy, dissolved in 'infinite suspense' and deconstructed as the vertigo of infinite recombinability in Paul de Man and as the emphasis on principled 'undecidability' in apparently fixed semantic orders of meaning in Jacques Derrida. The recent theoretical approaches, however, can hardly be considered a continuation of romanticism, but rather as the rewriting of romanticism, Enlightenment, and modernity as an inextricable philosophical complex, as 'the rewriting of some traits that had been claimed by modernity' (Lyotard 1988, p. 43).

See also: Culture: Contemporary Views; Enlightenment; Hermeneutics, Including Critical Theory; Kant, Immanuel (1724–1804); Knowledge, Sociology of; Modernity; Religion: Morality and Social Control

Bibliography

Allemann B 1956 *Ironie und Dichtung*. Neske, Pfullingen, Germany

Benjamin W 1919/1991 Der Begriff der Kunstkritik in der deutschen Romantik. In: Tiedemann R, Schweppenhäuser H (eds.) *Gesammelte Schriften*. Suhrkamp, Frankfurt am Main, Germany, Vol. I 1, pp. 7–122

Bohrer K H (ed.) 2000 *Sprachen der Ironie—Sprachen des Ernstes*. Suhrkamp, Frankfurt am Main, Germany

Ferraris M 1998 *L'Ermeneutica*. Roma/GLF editori, Laterza, Italy

Frank M 1977 *Das individuelle Allgemeine. Textstrukturierung und -interpretation nach Schleiermacher*. Suhrkamp, Frankfurt am Main, Germany
Friese H (ed.) 2001 *Identities*. Berghahn, Oxford, UK
Kierkegaard S 1841/1993 Über den Begriff der Ironie mit ständiger Rücksicht auf Sokrates. In: Hirsch E, Gerdes H (eds.) *Gesammelte Werke*. Gütersloher Verlagshaus, Gütersloh, Germany
Kramer F 1977 *Verkehrte Welten. Zur imaginären Ethnographie des 19. Jahrhunderts*. Syndikat, Frankfurt am Main, Germany
Lyotard J-F 1988 Réécrire la modernité. In: Lyotard J-F (ed.) *L'Inhumain: causeries sur le temps*. Galilée, Paris, pp. 33–44
Rorty R 1989 *Contingency, Irony and Solidarity*. Cambridge University Press, Cambridge, UK
Said E W 1979 *Orientalism* 1st Vintage Book ed., Vintage, New York
Schlegel F 1800/1988 Über die Unverständlichkeit. In: Behler E, Eichner H (eds.) *Kritische Schriften und Fragmente 1794–1828*. Ferdinand Schöningh, Paderborn/München, Germany, Vol. II, pp. 235–42
Schleiermacher F 1984 Ästhethik (1819/25). In: Lehnerer Th (eds.) *Über den Begriff der Kunst (1831/32)*. Meiner, Hamburg, Germany

H. Friese

Rousseau, Jean-Jacques (1712–78)

1. Life and Work

Rousseau was born on June 28, 1712 and raised in Geneva, the center of Calvinism. Jean Calvin, not a native of Geneva, published his magisterial *Christinae Religionis Institutio*, the central work on education, in 1536. Its final version, dated 1559, justified general priesthood by way of the doctrine of predestination. Calvin dissolves the 'two empires' of laymen and ecclesiastical office in favor of a parish without differences, one which sees itself as an elected body (Calvin 1957). Clergymen are chosen teachers whose sermon is intended to educate without the possibility of it being associated with primacy or hierarchy (Calvin 1958). Rousseau's life and work is, in many ways, a manner of dealing with Calvinism. The concept of *equality* determines *The Social Contract*, the theory of education, and Rousseau's natural theology. Calvin's doctrine of grace is translated into the predestination of nature, along with a rigorous inwardness which has the only true form of education on its side. Virtue becomes a model for society given the education of man's nature. Calvinistic rigor fascinated Rousseau. Following the banning and public burning of his books, he dispensed with the citizenship of the Republic of Geneva in 1763, but not his membership in Calvin's church, which he had renewed in 1754.

Rousseau's dramatic life (Cranston 1991a, 1991b, 1997, Incoprom SA et al. 1999) is characterized by ascent outside the elite and by being an outsider. The young Rousseau completed an apprenticeship as an engraver and in terms of his education he was largely self-taught. His constantly intermittent, restless, and wavering career ran between Geneva and Paris, later included Switzerland, and England. Rousseau himself attempted to gain control over his biography with the writing of his *Confessions* (Rousseau 1961), which was published posthumously in 1782 and have determined the image of Rousseau ever since then. In 1740, after various wanderings and low-level activities Rousseau took up a position as a private tutor in Lyon, where he wrote his first text on education (*Mémoire à Mr de Mably*: Rousseau 1969, pp. 1–32). In 1742 Rousseau went to Paris, publishing his *Dissertation sur la musique moderne* in 1743. In 1749 Rousseau wrote the articles on music for Diderot's *Encyclopédie*. His first major literary success came a year later with the *Discours sur les sciences et les art*, which was awarded a prize by the Academy of Dijon. From 1756 on, Rousseau, living now in the *Eremitage* of Montmorency, concentrated on his three major works, the *Nouvelle Héloise* (1761), a successful, widely read novel throughout Europe, the *Contrat social* (1762), Rousseau's theory of society, and *Emile ou de l'éducation* (1762), the theory of 'new education.' The latter two books were banned directly after publication and publicly burned in Paris and Geneva.

From then on Rousseau was on the run. He was granted asylum by the Prussian governor in Neuchâtel soon after his flight from Paris. At his place of refuge in Môtiers he was subjected to attacks by such differing authors as the general procurator of Geneva, Tronchin, and his arch-enemy Voltaire ('Sentiment des citoyens', Voltaire 1961, pp. 715–18). At the same time Rousseau became the idol of the generation of 'Sturm und Drang': young *literati* from all over Europe visited Rousseau in Môtiers and made him famous. In January 1766, his flight having taken him to various places, Rousseau made for England following an invitation from David Hume. Both had different preconceptions of each other and their relationship soon broke down. Rousseau returned to France and worked, with increasing paranoia, on his autobiography, which can be viewed as an attempt at self-therapy. His precarious financial situation was revived at short notice in May 1778, when Rousseau and Therèse Levasseur, whom he had married in 1768 following a long-standing affair, moved to Ermenonville into an estate owned by the Marquis de Girardin, Rousseau's final benefactor. His sudden death on July 2, 1778 ended a very unlikely literary and philosophical career which was intended to be *à rebour*—contrary to convention.

2. Education and Society

The archbishop of Paris, Christophe de Beaumont, justified the banning and burning of Rousseau's *Emile*

primarily because of its denial of original sin. Rousseau assumes that human nature is not tainted. 'There is no original sin in the heart of man' 1969, p. 322). In his statement of defense against Beaumont, printed at the beginning of 1763, Rousseau formulated a far-reaching dualism which has dominated political and educational conceptions ever since. The two aspects of this dualism are *nature* and *society*, which are treated as contradictions. The present social order contradicts human nature in every respect, society is the alienation of nature, and this explains the vices of men and the evils of life. The assumption of original sin then is superfluous: man could live without sin if nature and society corresponded and harmonized (Rousseau 1969, 966ff.). In this respect Rousseau had a post-Augustinian concept of education in mind, and this is developed in *Emile*.

Essentially, the theory has three axes: the political difference of *homme civil* and *homme naturel*, the assumption of phases of 'natural development,' and the anthropological difference of *amour de soi* and *amour propre*. Rousseau's concept of education is constructed from these axes. This concept is peculiar *not* because it stresses 'natural education,' a term that was established in educational discourses long before Rousseau (Mercier 1961), but because of its paradoxical attempt to solve the contradiction between nature and society by way of education. To this end, Rousseau, as early as in the draft of *Emile*, the *Manuscrit Favre*, draws a distinction between two contrary types of education: that of *nature* and that of *society* (Rousseau 1969, p. 58). These refer to two ways of life—the life of natural and of civil man—both described contrary to Hobbes, who, in his *Leviathan* (1968, originally published 1651) distinguished between natural and social condition. Society, for Hobbes, tames and cultivates nature, and thus repression of nature is unavoidable. Rousseau reversed this argument. For him, the social condition is that of ongoing civil war (*bellum omnium contra omnes*), while the natural condition is considered to be of presocial peaceableness. Education for 'public business,' as suggested by Hobbes (1968, 289ff.) is completely ignored.

Rousseau justified his fundamental thesis in the second *Discours*, dated 1765, which was devoted to the evolution of inequality (Rousseau 1964, pp. 109–94). Rousseau explains *inegalité* in the form of society's corruption of human nature, which was originally in a position to live harmoniously with people's own needs. *Homme sauvage* ('savage man'), as he terms it, is considered to exist on the basis of his own strengths, while *homme civilisé* ('civilized man') develops social needs which render him dependent and weaken his nature (Rousseau 1964, 135ff.). The division of labor, the development of knowledge, and connected social differentiation (Rousseau 1964, p. 143ff.) forcefully brought about a progressive inequality among men. Leave is taken of the condition of innocence and natural freedom (p. 152), and the invention of the *first* society irresistibly causes all those following (p. 178) for the prize of alienation of mankind from its own nature. In this respect *society* is the fall of men, not nature.

The theory of education in *Emile* picks up the evolution thesis and elaborates it with regard to the development of the child. This forces a reversal: the social condition does not follow the natural condition. Moreover, the social condition is led back to the natural condition, given that education can develop nature on the basis of its own potentials. Rousseau's risky experiment of thought goes as follows: if children are seen as the *homme sauvage*, education would have to apply to this condition entirely, provided that the social condition can be excluded for the duration of education. This assumption leads to *Emile*'s basic scenario: an *élève imaginaire*, described in greater detail merely as an 'orphan,' is exclusively educated by a 'governor' outside any society. The scene is anonymous countryside far away from corrupt cities and raw villages i.e., independent from the social condition as Rousseau viewed them (Rousseau 1969, pp. 264, 267, 279). The place in which the story is told is described simply *as au milieu des champs* ('in the middle of fields') (p. 277) without any details of the origin and personal history of the two protagonists. As such, the tale is not a novel of education but a treatise which is intended to describe the paradigm of true education. The name 'Emile' is presumably reminiscent of Plutarch (Shanks 1927), as it has no biographical meaning. Accordingly, the governor is not given a name and is thus not distinguishable. Both are paradigms, not persons.

'Natural' is *negative* education. Rousseau largely draws on the phases of human development from the *Histoire Naturelle* by Buffon and de Leclerc Comte (1989, pp. 1–114). Again it is not the schema of 'development,' formulated at the beginning of the *Manuscrit Favre* (Rousseau 1969, p. 60) that is original, but the linkage with the axes of educational theory. Rousseau understands the entire childhood as the 'age of nature,' which must take place outside the social condition. Thus the entire first education is negative. It does not consist at all of teaching virtue or truth, but in avoiding vices and errors in order to grant innocence of heart and integrity of mind (Rousseau 1969, p. 323). The concept of negative education is aimed against Locke (317ff.), whose *Some Thoughts Concerning Education* had been available in French in a translation by Pierre Coste since 1695 and had had considerable influence. Locke's central idea (1989, p. 142), that 'Children are to be treated as rational creatures,' aroused passionate contradiction in Rousseau. For him '*Reasoning* with children' (Locke 1989, p. 142) was nothing other than the reversal of nature's plan. Reason (*raison*) is the end of education, but the end cannot determine the beginning: 'If children could understand reason there would be no need to educate

them' (Rousseau 1969, p. 317). Childhood must have its own manner (*manière*) of seeing, thinking, and feeling, which is not that of adults (p. 319). Locke's concept of reason is that of adults outside the world of children.

It is above all this idea of the separate world of childhood, that justifies Rousseau's fame in education—along with the paradisaical metaphor of pure education outside society and the spontaneity of children led by immediate interests (*intérests présents*) (Rousseau 1969, p. 358). Rousseau seems to be the originator of modern education, which stresses the 'new' image of child and childhood. In this respect, however, the fact that *Emile* clearly has a theological center is overlooked. The treatise ends by giving preference to republic, not to nature, and the whole theory presents *two* concepts of education with differing gender forms. The center of *Emile* is the concept of 'natural religion,' justified in the *Profession de foi du Vicaire Savoyard*, a treatise within the treatise (Rousseau 1969, pp. 565–635). Rousseau takes the role of the vicar, who, against the materialism of the Parisian philosophy, announces a supreme being or an active creator who is assumed to be 'King of Earth' (Rousseau 1969, p. 382). Being in unison with creation cannot mean anything sinister: 'Where everything is good nothing can be injust' (p. 588). The good can be seen in creation, justice emanates from the good, only a just person can live happily (p. 589), but the enthusiasm for virtue (p. 596) is an inner principle (p. 598)—that of the heart and not of reason (p. 627). Consequently, the key to the profession of faith is as follows: 'We can be men without being scholars' (p. 601).

What Rousseau called 'natural religion' (Rousseau 1969, p. 607) is a theology of the heart (Osterwalder 1995, cf. Chevallier 1994), it does not require any positive dogma. Moreover, inner enlightenment needs no teachers (*pasteurs*) (Rousseau 1969, p. 609), only the inspiration of belief in seeing God's creation. Rousseau radicalizes the Calvinistic doctrine here: the metaphor of 'heart' signals independence even from the community of belief. 'Equality' is therefore not a social construction which would require comparison with others, but a sovereignty in solitude. 'Sovereignty' or inner strength presupposes identity and distance. Rousseau justified this difference with the relation between two instances of the soul called *amour de soi* and *amour propre*. The fact that two instances are mentioned violates again a convention of literature. In French virtue literature such as La Rochefoucauld's *Maximes* (1665) or Blaise Pascal's *Pensées* (1670), 'amour propre' is the formula for egoism and self-love that has to be restrained by education. Rousseau used the formula in a new way: he took over the negative meaning of *amour propre* but added a second, positive, instance—that of *amour de soi*. The positive instance suggests that man looks only for himself and is content with the satisfaction of his own true needs. The negative instance, *amour propre*, leads to comparison with others; man is never satisfied because no man will give preference to others more than to himself (Rousseau 1969, p. 493).

Loving and tender passions occur in *amour de soi*; hate and temper, i.e., from comparison with others, occur in *amour propre*. Thus, man is all the happier the fewer needs he has and the more he can avoid comparisons with others; in contrast, the factors which make man suffer ills and evils (*méchants*) are *too many* needs and *too many* opinions i.e., the dependency on others. Consequently, Rousseau develops the concept of *éducation solitaire* (Rousseau 1969, p. 341), which isolates the fictitious Emile throughout his childhood, totally controls learning (*la liberté bien réglée*: p. 321), and stipulates the didactic scenario on the theory of natural requirements and needs. Therefore Emile does not receive lessons, is not provided with any written works, and is excluded from all forms of cultural education. Only the *age de raison*—youth—requires formal lessons; all of childhood should apply exclusively to nature, which, of course, is only experienced in a didactic manner i.e., extremely artificial. The governor (and he alone) arranges the entire experience, and it is not by chance that *Robinson Crusoe* is the only literary evidence that Emile is given to read (Rousseau 1969, 455ff.).

The concept of *éducation solitaire* refers to the education of man (*faire un homme*: Rousseau 1969, p. 248). However, in the beginning of the fifth book Rousseau introduces 'Sophie ou la femme' i.e., he is forced to react to the difference in sex. The education of 'the woman' is given treatment as generalized as was previously the case with 'the child.' But now 'the child' is recognized as a *male* child and not as a sexless 'person' in general. Rousseau's *homme sauvage* is stated as male, namely 'strong,' 'sovereign,' and 'independent.' Sophie, the generalized women, is educated in a manner that she can see herself as being in complementary dependence on males. The education of the sexes is therefore not equal, it is dissimilar (Rousseau 1969, 700ff.). This is subject to the precondition of sexual difference that cannot be dissolved or harmonized (p. 697). Because men want to seduce, they are dependent in the crucial moment, but women have to agree and thus can say 'yes' or 'no.' This female strength can only be compensated by binding the woman to the house and forcing her to virtuousness publicly. The 'mutual dependency' is therefore not symmetric: men are dependent on women because of their désires (*desirs*), women are dependent on men as a result of their wishes *and* needs (*besoins*). Accordingly, 'by the law of nature itself women are at the mercy of men' (Rousseau 1969, p. 702), not vice versa.

Rousseau's treatise about true education ends with a grand tour (Rousseau 1969, pp. 826–55) which is intended to introduce Emile to basics of government theory (p. 833) i.e., the doctrines of Rousseau's own *Contrat social*. Endeavors are made regarding 'le

dernière perfection' (p. 799): it is envisaged that the well-educated *man* now is to become a respectable *citizen*. The citizen's know-how is taught with the theory of the 'social contract' (839ff.); Emile has to learn what defines the status of citizens (*citoyens*) and thus the constitution of society (p. 840). The small *republic* is given preference—one which can create an ideal relationship between population and government (p. 840). Obviously the Republic of Geneva, idealized by Rousseau during his lifetime, is the paragon for this thesis. The ideal republic is rare and thus hard to find. The governor concludes his teachings on government as follows: the obligations of the citizen presuppose a true country (*patrie*), and a country is true insofar as it can be considered a model of political virtue. Those who do not find their *patries* will not face political abysm, 'car qui n'a pas une patrie a du moins un pays' (Rousseau 1969, p. 858).

The connection of education and society fails. Rousseau had two views on society: first the ideal of the *contrat social* creating a model for social order that is mainly granted by the concept of *volonté générale*—the general will behind all singular wills (Rousseau 1964, 361ff.); and second the polemical description of general decadence within existing society. In no place does society comply with the ideal; therefore the grand tour ends without result. The couple Emile and Sophie, educated differently, are not released into society as citizens but led to marriage and family (p. 867). More is not possible: the ideal society, inspired by the *volonté générale* and thus safe from the corruption of power, is not realized anywhere. The social ideal has no social place. Decadent society, on the other hand, is morally unacceptable; real society has its place but not a legitimate form. Thus, loneliness (*solitude*) is in the end stronger than sociality (Starobinski 1971), because education will not change society but can only develop man's nature. The great project of 'humanization' though natural education (as stated by Ravier 1941) fails because of its own ambitions and contradictions. 'Nature,' in the framework of Rousseau's *Emile*, is an artificial experience, learning is bound by didactic parameters, control is total, and sentiment for the children's *separate* world arises from the fiction of a good and equal nature that does not mirror any child's individuality.

3. *Reception and Enduring Influence*

The reception of Rousseau's work demonstrates the often dramatic and always radical conflict between convinced supporters and equally convinced adversaries. Rousseau divides his readers into the twenty-first century, not least because he wrote against reality. His ideals of antiquity, especially those of Sparta and Athens, suggest an *âge d'or* (Terasse 1970, Leduc-Fayette 1974), which can be understood as an anticipation of the future, the renovation of the 'true society' that has been lost in history. It is therefore not by chance that Rousseau is the hero of the Jacobins; his cult was established in the French Revolution (Barny 1986) against all conservative theories that negate revolution in favor of the long-term development of society without the sentiment of decadence. Rousseau's *Social Contract* represents the new society that conservatives can only deny. The tension between freedom and equality attributed to, and paradoxically and provocatively described by, Rousseau characterizes one major part of political theory up to John Rawls's reformulation of the *Contrat social*. The same applies to the theory of 'natural education': Rousseau's *Emile* is a key source for Tolstoy and the reform movements of the nineteenth century, a central inspiration for Piaget (Oelkers 1996) and the development of child psychology, and a milestone for progressive education.

Of central importance are Rousseau's dualisms and thus his language. He stresses contradictions and paradoxes between nature and society, men and citizens, children and adults, and, not least, male and female. Even John Dewey, who rejected dualism, accepted that Rousseau—and only Rousseau—was the founder of the theory of 'natural development' (Dewey 1985, 211ff.) as opposed to artificial schooling, and was thus to be read in a dualistic manner. Educational theorists are irritated by Rousseau, who is regarded as a 'classical writer' of education (Rorty 1998) *because* of the unsolved paradoxes. Rousseau's concept of education makes no attempt at solving the paradoxes and dilemmas, but leaves them open, thus puzzling readers with keen contradictions which are right and wrong at the same time. What counts for Rousseau is not logic but the heart, and this seems to be part of his success. His radical conclusions seem to be obvious: politics within a new society following the education of 'new man' according to his own true nature. The famous ascription to 'return to nature,' which was part of the nineteenth century's reading of Rousseau and nowhere stated by him directly, was regarded as the emancipation from alienation, and thus a project of the left. But Rousseau was at the same time radical *and* conservative—stating that society should return to the golden age, education should leave schooling for true nature, man should be first man and then citizen. It is this rigor without reality that fascinated Rousseau's readers—disciples as well as opponents.

Rousseau's theory of education is provocatively puzzling: 'negative education' has no objective, and nature develops itself. However, the learning process is subject to extreme regulations (Rorty 1998, p. 248). The present should not be made a victim of the future (Rousseau 1969, p. 309), but every form of education is a deal with the future, and this is true for Rousseau's theory too: the ages of childhood and of youth are

clearly defined, as are the phases of education, so nature is not chance but necessity, aiming at the future. Children should be educated in a manner solely dependent on *things* (p. 311), yet the governor dominates the whole of Emile's education. Nature should lead the way (p. 259), but every effort possible must be made to avoid the first wrong step (p. 317). Children stand like savages outside the law and are completely natural (p. 320), but education must take everything in hand so as to exclude chance (p. 342) without being able to act solely on the basis of the necessity of nature. The first education should be that of the senses alone (p. 380), but that requires a rational plan of education which cannot simply be drawn from nature. In consequence Rousseau's educational scene is one of extreme regulation: Emile does not play, he does not develop anything of his own and is not allowed to listen to music, and his learning differs in every way from 'amusement' (p. 407). According to Rousseau, education should be 'realistic' (p. 418) and this is possible only when undertaken in an extremely artificial manner. Because society is excluded, educational rigor can rule. The basis for this is nature in the sense of self-love (amour de soi). Self-love is always good and always in compliance with the natural order (p. 491)—which begs the question why education does not concentrate fully on this *amour de soi*.

Rousseau wrote the counterclaim to sensualistic educational theory, which dominated the pedagogy of the eighteenth century. Children, according to Rousseau, should not be viewed as empty vessels or *tabulae rasae*, but instead as parts of nature which develop of their own accord. Children cannot simply be influenced, and education is not merely the establishment of habits and customs; moreover, it is the child's nature that limits all education. This fundamental concept is weakened by the implicit sensualism, the education of the senses (Rousseau 1969, 415ff.) that is necessary because education is inconceivable without any influence. But the provocation remains, and it defines Rousseau's standing as an educational author. Education is limited by nature, nature has nothing to do with sin, the child is innocent because of nature's original goodness, thus education can take place without any of the burdens of history and society. It is a renewal of mankind with every new child. This is the fundamental basis of Rousseau's theory, which is still read and discussed today because it provokes and stimulates educational thought without attempting to solve the puzzles. Rousseau is read today because he defines the problems, not their solutions.

See also: Democracy; Democratic Theory; Educational Philosophy: Historical Perspectives; Enlightenment; French Revolution, The; Rationalism

Bibliography

Barny R 1986 Rousseau dans la Révolution: le personnage de Jean-Jacques et les débuts du culte révolutionnaire (1787–1791). In: Mason H T (ed.) *Studies on Voltaire and the Eighteenth Century*. The Voltaire Foundation at the Taylor Institution, Oxford, UK, Vol. 24

Buffon G L, Leclerc Comte de 1989 *Histoire naturelle de l'homme et des animaux*. Jean de Bonnot, Paris

Calvin J 1957 *Institution de la Religion chrétienne. Livre troisième. Édition nouvelle publiée par La Société Calviniste de France*. Labor et Fides, Geneva, Switzerland

Calvin J 1958 *Institution de la Religion chrétienne. Livre quatrième. Édition nouvelle publiée par La Société Calviniste de France*. Labor et Fides, Geneva, Switzerland

Chevallier M 1994 *Pierre Poiret (1646–1719). Du protestantisme à la mystique*. Labor et Fides, Geneva, Switzerland

Cranston M 1983 *Jean-Jacques. The Early Life and Work of Jean-Jacques Rousseau 1712–1754*. University of Chicago Press, Chicago

Cranston M 1991 *The Noble Savage. Jean-Jacques Rousseau 1754–1762*. University of Chicago Press, Chicago

Cranston M 1997 *The Solitary Self. Jean-Jacques Rousseau in Exile and Adversity*. University of Chicago Press, Chicago

Dewey J 1985 Essays on education and politics 1915. In: Boydston J A (ed.) *The Middle Works, 1899–1924*. Southern Illinois University Press, Carbondale, IL, Vol. 8

Hobbes T 1968 *Leviathan* [ed. MacPherson C B]. Penguin, Harmondsworth, UK

Incoprom S A, L'Institut de France, Cybele Productions 1999 *Le CD-Rom Jean-Jacques Rousseau*. Incoprom SA, L'Institut de France, Cybele Productions, Geneva, Switzerland

Leduc-Fayette D 1974 *Jean-Jacques Rousseau et le mythe de l'Antiquité*. J. Vrin, Paris

Locke J 1989 *Some Thoughts Concerning Education*. [eds. Yolton J W, Yolton J S]. Clarendon Press, Oxford

Mercier R 1961 L'enfant dans la société du XVIIIe siècle (before *Emile*). Ph.D. thesis, University of Paris

Oelkers J 1996 Piaget et l'éducation nouvelle. In: Barraletta J-M, Perret-Clermont A-N (eds.) *Jean Piaget et Neuchâtel. L'apprenti et le savant*. Éditions Payot, Lausanne, Switzerland, pp. 165–76

Osterwalder F 1995 Die pädagogischen Konzepte des Jansenismus im ausgehenden 17. Jahrhundert und ihre Begründung. Theologische Ursprünge des modernen pädagogischen Paradigmas. In: *Jahrbuch für Historische Bildungsforschung*. Juventa Verlag, Weinheim, Germany, Vol. 2, pp. 59–83

Ravier A 1941 *L'éducation de l'homme nouveau. Essai Historique et Critique sur le Livre de l'Émile de J.-J. Rousseau*, Vols. 1 and 2. Éditions Spes, Issoudun, France

Rorty A O 1998 Rousseau's educational experiments. In: Rorty A O (ed.) *Philosophers on Education. Historical Perspectives*. Routledge, London, pp. 238–54

Rousseau J-J 1961 *Les Confessions. Autres textes autobiographiques*. In: Gagnebin B, Raymond M (eds.) *Oeuvres complètes*. Gallimard, Paris, Vol. 1

Rousseau J-J 1964 *Du Contrat Social. Écrits politiques*. In: Gagnebin B, Raymond M (eds.) *Oeuvres complètes*. Gallimard, Paris, Vol. 3

Rousseau J-J 1969 *Émile. Education—Morale—Botanique*. In: Gagnebin B, Raymond M (eds.) *Oeuvres complètes*. Gallimard, Paris, Vol. 4

Shanks L P 1927 A possible source for Rousseau's name *Emile*. *Modern Language Notes* **17**: 243–4

Starobinski J 1971 *Jean-Jacques Rousseau. Le transparence et l'obstacle*. Gallimard, Paris
Terrasse J 1970 *Jean-Jacques Rousseau et la quête de l'âge d'or*. Palais des Académies, Brussels, Belgium
Voltaire 1961 *Mélanges*. Gallimard, Paris

J. Oelkers

Rule of Law

The rule of law is commonly understood in contrast to arbitrary exercise of power, which is, above all, the evil that the rule of law is supposed to curb. The rule of law, though not always the specific verbal formulation, embodies ideals that have been central to political and constitutional discourse at least since Aristotle, who contrasted 'the rule of the law' with 'that of any individual.' In any of its many versions, the meaning of the rule of law is contested; all, however, share the idea that law can and should contribute in salutary, some say indispensable, ways to channeling, constraining, and informing—rather than merely serving—the exercise of power, particularly public power.

The antagonistic juxtaposition of rule of law and arbitrariness has led many jurists to view the rule of law as a purely negative ideal (Raz 1979), devoted to 'damage control' (Shklar 1998a, p. 5), particularly damage done by government. However that view is not universal. Several writers (e.g., Allan 1993, Dworkin 1985, Selznick 1999) offer a more affirmative, morally ambitious account of what is required to govern according to the rule of law. Negatively, constraint by clear legal rules is central. Positively, space needs to be made to include underlying principles of morality and purpose in legal systems. The differences between these views go deep, involving larger differences over the nature and the point of law. This account begins with the better known negative conception, but seeks to point beyond it.

1. *Elements of the Rule of Law*

1.1 *Scope*

Central to most conceptions of the rule of law is the conviction that there should be no privileged exemptions to it. Everyone comes within the scope of the law, even those who make it. This has a political and a social aspect.

1.1.1 Government under law. The political aspect of rule of law is that governments and public officials are subject to the existing law. Legislatures can of course change law, but even 'sovereign' legislatures must comply with existing law so long as it remains in force, and can only change it in conformity with existing higher-order legal rules. To make this subjection effective, there must exist legal ways of challenging and forcing the government and political officials, even the very highest, to submit to the law.

There have been many polities in which the ideal of subjecting government to law was unknown and would seem outlandish. Even where the idea exists, the extent to which it is realized within a political order will vary markedly between polities and over time. It is, after all, a complex ideal even more complex to realize, not a faithful description of a state of affairs. Where the ideal is professed or understood, differences in attainment are crucial indicators of the sway of the rule of law. To the extent that the idea or the practice is lacking, so is this crucial element of the rule of law.

1.1.2 Equality before the law. The social aspect of the rule of law is the equal subjection of all citizens to the law. This, it was traditionally thought, required that law should take the form of an 'abstract rule which does not mention particular cases or individual nominated persons, but which is issued in advance to apply to all cases and all persons in the abstract' (Neumann 1986, p. 213; see also Hayek 1960, Oakeshott 1983). This is to prevent the use of law against specific groups or individuals, by making all equally liable, in advance, if they come within its terms. A corollary is that adjudicators should then weigh, impartially and impersonally, legal issues that come before them on the basis of such laws. They should be 'blind,' as the bearer of the scales of justice is classically represented, to inequalities of status, power, wealth, and other resources.

These conditions are the source of much controversy. Contemporary law associated with the welfare and regulatory activities of modern governments is frequently neither general nor abstract but targeted, detailed, and specific. A matter of continuing controversy is whether these forms of law are compatible with the rule of law (Hayek 1973–1979, Kamenka and Tay 1975, Neumann 1986, Oakeshott 1983, Unger 1976, but see Selznick 1992, 1999); if so, how, and if not, what should be sacrificed to what. A matter of even longer dispute is whether such formal equalization counts for much if it merely leaves substantive social and economic inequalities to play themselves out with more significant effect, and even if the legal arena allows or ensures that the 'haves come out ahead' (Galanter 1974). Liberals have commonly believed that the existence of social inequalities does not cancel out the worth of the rule of law, and may even make its impartiality more significant; precisely because inequalities confer advantages in the world,

they should not do so before the law. Many on the political left disagree, arguing that differences in the world render the 'blindness' of the law false or inconsequential, or both.

1.2 Character

A second element of the rule of law has to do with the character of laws. It can be expressed with minimal and deceptive simplicity: 'the law should be such that people will be able to be guided by it' (Raz 1979, p. 213). Reflection on that simple condition can generate tomes; it has certainly generated lists (see Fuller 1969, Raz 1979, Walker 1988). In order that law guide behavior, it must exist, that is, it must take the form of general requirements, typically rules, that people can consult before they act. Even where it exists, it will not guide behavior unless its requirements can be known. Therefore it must be public, understandable (by someone—it might take a lawyer), and relatively clear and determinate in its requirements. These requirements must be possible to perform. So the law cannot be internally contradictory (more accurately, the legal system must provide ways to resolve contradictions which inevitably occur). It must be prospective, for if it is retrospective people will neither know it, nor be able to rely on what they do know for fear of it being retrospectively changed. It must be relatively stable, for if it were constantly to change people would be reluctant to trust any present incarnation. Exercises of public power made unannounced, or announced in peremptory or arbitrary decrees, retrospective in effect, in ignorance or in violation of existing law, defy the rule of law.

Of course, whatever the character of the laws, the rule of law will amount to little unless laws are taken seriously, and enforced. So the law must be authoritative, that is, taken as binding by those whom it claims to bind. To make it so, it must be treated so, at least by the legal institutions that administer it, and it must be administered in ways that conform to its publicly announced terms.

1.3 Institutions

In the most influential English account of the concept in the last several centuries, Dicey (1961) took the rule of law to be a peculiarly English virtue associated with peculiarly English institutions, unavailable across the Channel but successfully implanted across the Atlantic. From what Judith Shklar aptly describes as this 'unfortunate outburst of Anglo-Saxon parochialism,' the rule of law emerges, 'both trivialized as the particular patrimony of one and only one national order, and formalized, by the insistence that only one set of inherited procedures and court practices could sustain it' (Shklar 1998b, p. 26).

As an ideal of some degree of generality, longevity, and abstraction, however, the rule of law is tied neither to any one national experience nor to any set ofm institutions. Rule-of-law regimes have often embodied different judgments about how to implement rule-of-law ideals, and have been realized within different legal and other histories and traditions that have influenced the particular shapes of the institutions that emerge. Institutional variety and possibility are too rich, and local traditions often too dense and particular, for it to be fruitful to identify the rule of law with just those particular institutions and arrangements which seem to work at home. Moreover, even within one society, problems change, and so do the solutions most appropriate to solve them. These differences are not automatically fatal, since the rule of law is not a recipe for detailed institutional design. It is rather a value, or an interconnected cluster of values, which might inform such design, and which might be—and have been—pursued in a variety of ways (see Selznick 1999).

This is not to deny that the ideal has been better served in some nations and by some institutions than by others. Institutional learning can and does occur. Some arrangements have been shown to work well in many contexts; others less so or only in some contexts. It is, therefore, appropriate to ponder the relationship between particular, realized, institutional forms, and rule-of-law ideals. Success does not, however, depend upon faithful mimicry (for even if whole institutional complexes could be duplicated, it is another job to graft them), and it often thrives on innovation, even misinterpretation. Montesquieu understood the English to have prospered because of the institutional separation of powers he mistakenly attributed to them. Even though this was not quite so, the Americans were so impressed with the philosopher's account that they institutionalized a strict institutional separation of powers embedded in that other American innovation in aid of the rule of law, a binding, written constitution. They have been widely copied. The details vary, but there is a rather simple and generalizable institutional insight at the bottom of this: at the very least, those who judge the legality of exercises of power should not be the same as those who exercise it; and they should be shielded from interference. This is why the separation of the judiciary from other branches of government is seen as central in rule-of-law states, even where they do not follow America in other aspects of the separation of powers. It is also why the rule of law is thought to depend on measures to secure the independence of the judiciary from political and other pressures. Further resources, often institutional traditions and conventions as much as written laws, are important in order to ensure that judicial decisions are grounded within plausible interpretations of existing laws. Institutional measures to guarantee litigants a fair hearing are also crucial. Access to legal institutions should be available to those who would make use

of them. The details can go on forever, but it is better not to amass them aimlessly, as though such means, rather than the values that inspire them, were the point. Rather, it is preferable to think to them in their particular contexts, from the stand point of the idea of the rule of law, and of the values it is thought to serve, together with close, but not slavish, attention to stories of institutional success and failure.

1.4 Legal Culture

As lawyers often forget, but social scientists never should, the rule of law depends upon more than the scope, character, and institutions of the law. It also depends upon a host of—apparently 'soft' but actually crucial—cultural supports, among them socialization into the values of the rule of law, at least of the professionals who have to administer it, and assumptions more or less widespread in the society that laws do and should count. Such assumptions are commonly revealed in the degree of willingness to use laws 'on the books' to protect and advance one's interests, in demand for law as much as in supply as well as in levels of unreflective but routine law abidingness in a society. These crucial aspects of legal culture and practice are variable among and within societies and we know little about how to affect them in ways that might advance the rule of law. Lawyers give more thought to laws and institutions than to what might be necessary to make them count, but if no one is listening it does not matter too much what the law is saying. And who listens, why and how, and with what effects are empirical questions to which the law offers few answers, though the sociology of law might. If the answer is 'only lawyers,' the rule of law is not in good (see Örkény 1999).

2. The Value of the Rule of Law

Though the rule of law carries the patina of age and the danger of congratulatory cliché, what it expresses are not truisms. Neither the goal nor the means of the rule of law are self-evident goods. Many have thought that what mattered was who had power, not whether or how that power might be constrained. Indeed, if the right people had power, they should be unconstrained. Charismatic leaders, as Weber points out (1968, p. 243) are the sort who say 'it is written … but I say unto you …'; often they have been happier when 'it' was not written at all. Nor was it, for example, a Russian tradition that the ruler, whether Tsar or Party, should be constrained, either by laws or to produce laws.

For anarchists, on the other hand, constraint was not enough. Institutions of public power had to go; this was Marx's long-term view too. In the meantime, like many radicals, he considered fraudulent those claims that law significantly restrained the power of the powerful within capitalist societies. In bourgeois societies, many Marxists have alleged, the rule of law is an ideological mystification which masks the truth: that the bourgeoisie rules, typically with aid of the instruments of law (see Krygier 1994). Many subsequent 'critical' thinkers, Marxist or not, have tended to agree on this point.

In the modern world, it is the liberal tradition that values the rule of law most highly. Liberty being central to liberalism, liberals are concerned with ways of averting threats to it and with securing it. An overarching source of such security is taken to be the rule of law. What is at stake here?

2.1 Freedom from Fear

According to Thomas Hobbes, fear of others is what drives people to agree to subjection to a political sovereign, and in turn justifies that subjection. The sovereign, by being empowered to issue and enforce laws, saves us from others, and them from us. Moreover, and therefore, it saves us and them from a major reason to fear each other. When we all are equally subject to known laws, both we and they have assurances which would otherwise be unavilable.

Many thinkers who agree with Hobbes on little else agree with him that this is a necessary and salutary work for states to do. But as John Locke observed, if we have reason to fear our neighbors who are just individuals, how should we avoid terror of that 'mortal God,' the state, which Hobbes called Leviathan? The solution is to arrange an institutional order which both consolidates power, and then curbs it, so preventing its arbitrary exercise; the challenge is to achieve such a solution. One very old attempt to achieve it—suggested by liberalism though not its invention—is to require rulers to operate under law, and by means of laws of a character that conforms to the rule of law.

The stakes on this view are high. According to those belonging to the tradition Judith Shklar calls the 'liberalism of fear,' we have a simple and stark choice,

> a very basic dichotomy. The ultimate spiritual and political struggle is always between war and law. … The institutions of judicial citizen protection may create rights, but they exist in order to avoid what Montesquieu took to be the greatest of human evils, constant fear created by the threats of violence and the actual cruelties of the holders of military power in society. The Rule of Law is the one way ruling classes have of imposing controls upon each other (Shklar 1998b, p. 25).

Reflecting on the blood-stained twentieth century, the eminent Marxist historian E P Thompson voiced similar sentiments, in the process scandalizing many

other Marxists, who traditionally had little time for law. He insisted that

> the rule of law itself, the imposing of effective inhibitions upon power and the defence of the citizen from power's all-intrusive claims, seems to me to be an unqualified human good. To deny or belittle this good is, in this dangerous century when the resources and pretentions of power continue to enlarge, a desperate error (Thompson 1977, p. 266).

What Shklar, a lifelong liberal, and Thompson, a somewhat rueful Marxist or ex-Marxist, share is the insistence that the value of the rule of law lies primarily in what it shields us against. When they warn complacent beneficiaries of the rule of law to value what they have, it is by comparison with the perils that they believe to flow from its lack.

2.2 *Coordination and Cooperation*

Fear is not, however, the only reason for the rule of law, nor is government its only subject. We all are. Madison wrote that if we were angels we would not need any laws. Unfortunately, so the familiar argument goes, we are not, so we do. However, even angels and indeed all but the omniscient, particularly if there are a lot of them about, might benefit from the rule of law. We can all become confused and lose our way, not necessarily due to our or others' evil but merely to the superfluity of possibilities in an unordered world. All the more when that world is, as ours is, one of large and mobile societies full of strangers, where ties of kinship, locality and familiarity can only partially bind, reassure, or inform.

The predicament of a member of such a 'civilized' society, identified by Adam Smith in the eighteenth century, is that he 'stands at all times in need of the co-operation and assistance of great multitudes, while his whole life is scarce sufficient to gain the friendship of a few persons' (Smith 1981, Vol. 1, p. 26). The rule of law can provide crucial information and security, 'a basis for legitimate expectations' (Rawls 1971, p. 238), without which the 'co-operation and assistance of great multitudes' will necessarily be a more chancy affair. It enables fellow citizens to know a good deal about each other, though many of them are strangers, to coordinate their actions, and to feel some security and predictability in their dealings with them. For though it does not make everything predictable, it ties down much that would otherwise be up for grabs. It establishes fixed and knowable points in the landscape, on the basis of which the strangers who routinely interact in modern societies can do so with some security, autonomy, and ability to choose.

To the extent that a legal system approximates the ideal of rule of law, citizens have, or can obtain, in advance, clear understanding of their and others' legal obligations, and they can reasonably have faith that the law will constrain other citizens and officials of state in ways that they can predict. Rules of the road are an example; everyone has to obey them, and everyone suffers from confusion, or a good deal worse, if they do not apply to all, are retrospective, secret, incomprehensible, contradictory, require the impossible, or are corruptly administered. If laws do not suffer from these defects, citizens share intersubjective cues as to the rights and responsibilities of people they might know little else about. These are invaluable cues (and clues) to have.

3. *Justice*

On the account so far, and according to many writers, the rule of law might provide certain basic conditions for a society to exist free of certain fears and informed of certain things, but it has little to do with the moral quality of what goes on there. According to one prominent interpretation, there is no special connection between the rule of law and higher values such as justice. The former merely sets out principles of legal efficacy, such as those one might address to a knife-maker: 'make it sharp'; or to a poisoner: 'make it kill quickly and quietly' (see Hart 1965, p. 1286). What it is used for is a quite separate question. Moreover, proponents of this view point out that the rule of law is perfectly compatible with iniquity in the content of laws. The Republic of South Africa, prior to 1989, is a common example in this connection. On the other hand, it is common for there to be a 'general and drastic deterioration in legality' (Fuller 1969, p. 40) when despots do their worst, and there is no surprise in that. There is nothing in the logic of despotism that inclines it to respect the rule of law, and there are many costs in doing so for which there is no despotic rationale. So at the very least one might claim a strong empirical inconsistency between horrific government acts and the rule of law.

Many writers might concede this, but still agree with Joseph Raz that 'the rule of law is a purely negative value. It is merely designed to minimize the harms to freedom and dignity which the law may cause in its pursuit of its goals however laudable these may be' (Raz 1979, p. 228). However, as we have seen, the rule of law has a horizontal function—in facilitating interaction and cooperation among citizens—which is as important as its vertical one—in relation to governments. To the extent that it performs the former role, it goes well beyond the merely negative virtue that Raz allows it.

By contrast to this negative conception, another tradition has it that the rule of law is altogether more, and more morally charged, than mere formal or procedural regularity. For Ronald Dworkin, the rule of law 'is the ideal of rule by an accurate public conception of individual rights' (Dworkin 1985,

pp. 11–12). That will often involve formal regularity, but not only, and not always. For Philip Selznick, the constraints emphasized in the formal conception are precious and need to be respected for what they secure. Yet the weight given to formality in the rule of law might generate vices of its own—particularly that of blindness to particularity—that subvert the doing of justice that the rule of law might otherwise be thought to support.

Even apart from the high price of exaggerated 'legalism,' legal systems are not always or only threatened by unrestrained power. There are other sources of arbitrariness in social life, and among them can be a refusal to take account of particulars to which rules are being applied. Moreover, in well established legal systems, with strong rule-of-law institutions, traditions, and legal expectations, there might well be room to vindicate larger values carried within the principles, traditions and purposes of law, in the service of a 'higher instrumentalism' which is not limited to, but need not necessarily subvert, a legal order whose injunctions overall are clear, predictable, and obeyed.

According to this view, formal regularity should not be the limit of one's ambitions for the rule of law. Rather, it should be supplemented by a

> thicker, more positive vision [that] speaks to more than abuse of power. It responds to values that can be realized, not merely protected, within a legal process. These include respect for the dignity, integrity, and moral equality of persons and groups. Thus understood, the rule of law enlarges horizons even as it conveys a message of restraint (Selznick 1999, p. 26).

The move from negative to positive conception, on Selznick's view, is not merely a matter of normative expansion. It has a sociological basis and plausibility, for

> we cannot really separate the negative and positive aspects of the rule of law. Indeed it would be highly unsociological to try to do so, for we would then miss the moral and institutional dynamics which create demands for justice, and which induce rulers to accept accountability (1999 p. 25).

One objection which the first, agnostic, interpretation has to the second, justice-impregnated one is that it threatens to do away with the specificity and usefulness of the concept:

> [i]f the rule of law is the rule of the good law then to explain its nature is to propound a complete social philosophy. But if so the term lacks any useful function. We have no need to be converted to the rule of law just in order to discover that to believe in it is to believe that good should triumph (Raz 1979, p. 211).

Yet there is a space between a purely procedural account of the rule of law, whose only connection—if any—with virtue is external, and one that wipes out the distinction between the rule of law and the rule of good law. Legal orders typically embody and generate certain values, both in their animating principles and in the complaints they provoke when their practices flout the values that give them legitimacy. These include values such as due process or natural justice and, in particular legal traditions, much more. The rule of law might well be argued to be incomplete to the extent that such values are not honored. Attempts to vindicate such values, often implicit in legal principles and traditions if not in rules, might be considered a service to the rule of law, even if it goes beyond purely procedural fairness and notwithstanding that it might fall short of justice more globally conceived.

4. *Tensions and Challenges within the Rule-of-law Ideal*

Like most important ideals, the rule of law contains internal tensions and can lead to conflict with commitment to other social ideals, because the ideals themselves are under strain, or because different interpretations of the same ideal, or attempts to realize different ideals, have different institutional logics. Thus, whether one thinks it an assault on the rule of law, or a vindication of a particular vision of it, the law generated by an active welfare state sits ill with some of the procedural conditions mentioned earlier. The simplest ways to deal with such tensions is to hold fast to one interpretation of one ideal and reject whatever might compromise it in another (Hayek 1973–79, Walker 1988). Another is to minimize the conflict and pretend that no price is being paid. More complex, but perhaps more realistic, is to acknowledge that the rule of law is not consistent with every value that one holds dear, and that, consequently, compromises in one or other direction might be unavoidable. We ask many things of law, as we do of life, and not all of them are consistent with each other. The rule of law is a mighty source of good in large modern polities, but not the only source, nor one which is automatically better the more you have of it. If power in a polity is unconstrained, the negative conception of the rule of law speaks to precious virtues, the absence of which can be vicious. But if power is already substantially constrained by law, the rule of law might tolerate, even on occasion require, that some space be made for wisdom, judgment, particularity, and substantive justice.

All the more is this the case, if one amplifies the lawyer's and philosopher's understanding, of what law does and how it does it, with what might be gleaned from other social sciences. Sociology of law does not suggest any linear connection between the rigidity and precision of laws and their real-world

reliability (see Bardach and Kagan 1982). Social causation is too complex, and the mediation and supplementation of formal by informal institutions, norms, and social practices too pervasive, for such a connection to be plausible, or for the extrapolations of legal dogmatics to carry much explanatory or predictive weight.

Yet those who sought the protection of the rule of law against abuse of power and social disorganization were onto something precious. Our experience of despotism, and of rule-less chaos, gives us good reason to take their exhortations very seriously indeed. Our experience, however, of incompetence, legalism, and arbitrary manipulation of hard-edged rules, even by non-despots, gives us reason to be suspicious of too narrow an understanding of the rule of law, or, if that understanding appeals, not to exaggerate its proper reach.

See also: Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Governments; Hobbes, Thomas (1588–1679); Justice and Law; Law and Democracy; Rechtsstaat (Rule of Law: German Perspective)

Bibliography

Allan T R S 1993 *Law, Liberty and Justice. The Legal Foundations of British Constitutionalism*. Clarendon Press, Oxford, UK

Bardach E, Kagan R 1982 *Going by the Book*. Temple University Press, Philadelphia, PA

Dicey A V 1961 *Introduction to the Study of the Law of the Constitution*, 10th edn. Macmillan, London

Dworkin R 1985 Political judges and the rule of law. In: Dworkin R (ed.) *A Matter of Principle*, Harvard University Press, Cambridge, MA

Fuller L L 1969 *The Morality of Law*, rev. edn. Yale University Press, New Haven, CT

Galanter M 1974 Why the 'haves' come out ahead: Speculations on the limits of legal change. *Law and Society Review* **9**: 95–160

Hart H L A 1965 The morality of law (book review). *Harvard Law Review* **78**: 1281–94

Hayek F A 1960 *The Constitution of Liberty*. Routledge and Kegan Paul, London

Hayek F A 1973–79 *Law Legislation and Liberty* (3 Vols.), University of Chicago Press, Chicago

Kamenka E, Tay A E-S 1975 Beyond bourgeois individualism—the contemporary crisis in law and legal ideology. In: Kamenka E, Neale R S (eds.) *Feudalism, Capitalism and Beyond*. Edward Arnold, London

Krygier M 1994 Marxism, communism, and the rule of law. In: Krygier M (ed.) *Marxism and Communism. Posthumous Reflections on Politics, Society, and Law*, Rodopi, Amsterdam

Neumann F L 1986 *The Rule of Law*. Berg, Leamington Spa, UK

Oakeshott M 1983 The rule of law. In: Oakeshott M (ed.) *On History and Other Essays*. Blackwell, Oxford

Örkény S 1999 Rules of law: The complexity of legality in Hungary. In: Krygier M, Czarnota A (eds.) *The Rule of Law after Communism*. Ashgate/Dartmouth, Aldershot, UK

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Raz J 1979 The rule of law and its virtue. In: Raz J (ed.) *The Authority of Law*. Clarendon Press, Oxford, UK, pp. 210–29

Selznick P 1992 *The Moral Commonwealth*. University of California Press, Berkeley, CA

Selznick P 1999 Legal cultures and the rule of law. In: Krygier M, Czarnota A (eds.) *The Rule of Law after Communism*. Ashgate/Dartmouth, Aldershot, UK

Shklar J 1998a The liberalism of fear. In: Shklar J (ed.) *Political Thought and Political Thinkers*. University of Chicago Press, Chicago

Shklar J 1998b Political theory and the rule of law. In: Shklar J (ed.) *Political Thought and Political Thinkers*. University of Chicago Press, Chicago

Smith A 1981 *An Inquiry into the Nature and Causes of the Wealth of Nations*. Liberty Fund, Indianapolis, IN

Thompson E P 1977 *Whigs and Hunters. The Origins of the Black Acts*. Penguin, Harmondsworth

Unger R M 1976 *Law in Modern Society*. Free Press, New York

Walker G de Q 1988 *The Rule of Law*. Melbourne University Press, Melbourne, Australia

Weber M 1968 *Economy and Society. An Outline of Interpretive Sociology*. Bedminster, New York

M. Krygier

Rules in the Legal Process

The ancient Romans had two words for law: *lex*, the specific rules written in statute books, and *ius*, enduring principles of justice. Contemporary 'legal positivists' view *lex*, the legal rules made by legislatures, courts, and administrative agencies, as the essence of law, and those legal rules are the focus of this article. Yet the idea that law also includes a body of principles, in some sense superior to the rules written by the current political regime, persists in many modern languages, in legal theory (Dworkin 1967, Fuller 1964), in popular consciousness (Ewick and Silbey 1998), and in the recurrent political effort to entrench 'higher law' principles in written constitutions.

To distinguish rules from principles, legal scholars emphasize the specificity and binding quality of legal rules. Thus, Lawrence Friedman (1975, p. 26) defines a legal rule as 'a proposition of law couched in general terms,' consisting of a statement of facts and a statement of the legal consequences which follow from those facts, for example, 'Any person who drives a motor vehicle through a red light is guilty of a violation and shall be fined $50.' Social scientists, on the other hand, focus on the distinction between legal rules and customary rules or social norms. Thus Max Weber

defined legal rules as norms which have some likelihood of being enforced coercively by an officially-designated staff of legal officials (Rheinstein 1954, pp. 1vi, 14–15), a definition that emphasizes the political origin of legal rules and the contingencies surrounding their actual use.

1. *Types of Legal Rules*

Legal rules can be unwritten, as in preliterate societies (Hoebel 1964), or written and published, as in larger, literate political systems. Some legal rules are specific, such as those which prohibit driving over 60 miles per hour, while others, such as those that impose liability for accidents caused by driving 'unreasonably dangerously,' articulate vague standards. Highway speed limits are primary rules, designed to regulate ordinary citizens and businesses. Secondary rules, in contrast, guide the making and application of primary rules by government officials and judges (Hart 1961). Thus secondary rules—often stated in constitutions as well as in statutes—prescribe which officials have authority to make speed laws (and other laws) for a particular territory, who has authority to enforce them, who has authority to adjudicate disputes about their proper application, and who has authority to build the roads and install traffic lights. Other secondary rules govern the making of rules by legislatures, while 'procedural rules' govern the process of litigation and adjudication. In political systems that value the rule of law, one fundamental secondary rule is that legal officials must be guided in their decisions and actions by preexisting, publicly accessible primary and secondary rules.

From a sociolegal perspective, it is useful to distinguish 'high consensus' legal rules (such as statutes condemning theft or breach of contract) from rules that are politically, morally, or economically controversial. Controversial rules are generally much harder to enforce than high consensus rules, and they pose more difficult moral dilemmas for officials and judges who are expected to enforce them.

2. *The Political, Social, and Institutional Functions of Legal Rules*

The most obvious function of legal rules is to serve justice, increasing the likelihood that different decision-makers will treat like cases alike. Political scientists, however, emphasize the political control function of legal rules (Farber and Frickey 1991): political authorities entrench their policy choices in statutes and administrative rules to increase the odds that those policies will be implemented uniformly by far-flung bureaucrats, law enforcement officials, and judges. For political authorities, legal rules also serve a broader communicative function, informing citizens and businesses about their obligations and the decisions they might expect from government administrators, law enforcers, and judges. A tax code provides a familiar example. In consequence, legal rules often are said to have a constitutive function, shaping citizens' conceptions about what is fair, permissible, and proper (Sarat and Kearns 1993, Edelman and Suchman 1997, pp. 503–6, Kagan and Skolnick 1993).

By imposing controls on administrative officials, police, and judges, legal rules tend to serve a useful function for the citizenry: restraining official power, at least in the sense of making it more predictable. Published legal rules provide criteria for holding officials accountable via judicial review or public criticism. For this reason, legal rules, when combined with a credible governmental commitment to adhere to them, serve yet another function: adding legitimacy to governmental and judicial authority—which helps explain their ubiquity in the governance of nation-states. To Karl Marx and to contemporary 'critical legal theorists' (Kennedy 1997, Kairys 1982), this legitimating function has an ominous side, shrouding repressive policies (and, one might add, nonrepressive policies) and coercive penalties in the language of neutrality and restraint of power.

In contemporary democratic countries, legal rules are not infrequently remade in response to popular political movements, litigation by public interest law firms, and electoral competition. Hence they have still another function: protecting individual rights. Partly because legal rules lend a graspable concreteness to general ideals and notions of justice, ordinary citizens can brandish them as shields and swords against government officials, businesses, and fellow citizens, both in and outside of court. This is the function referred to by the ancient Greek playwright Euripides in *The Suppliant Women* when he has Theseus praise the development of written codes of law:

Once written laws exist
People of few resources and the rich
Both have the same recourse to justice. Now
A man of means, if badly spoken of,
Will have no better standing than the weak;
And if the less is in the right, he wins
Against the great.

Many sociolegal scholars would argue that Euripides praised written legal rules too fulsomely. For as rules accumulate within a legal system, acquiring complicated subsections and qualifying clauses, they make the legal process so complex that it becomes inaccessible unless one can afford a lawyer. Rather than guaranteeing equality, therefore, contemporary legal systems tend to provide an advantage to sophisticated, better-financed litigants (Galanter 1974). Many sociolegal scholars might have Thrasymachus (the legal cynic of Plato's *Dialogues*) respond to Theseus by saying something like this:

All written laws are parchment words

Whose import shifts in each man's mind.
Thus Justice is tossed this way and that from
Judge to judge, and advocates with clever minds
Oft bend the law to serve their wealthy friends.

After communist revolutions in Russia and Cuba, new regimes established 'people's courts' in which lay judges, freed from 'bourgeois' legal rules, were to 'do justice' based on the community's sense of fairness and social good. But the unhappy experience of 'socialist justice' (Markovits 1986, Hendley 1996, Krygier 1990) reminds us that written legal rules have an important political function: preserving judicial autonomy and blunting the risk of tyranny. Moreover, the 'peoples' courts' experience suggests that Theseus was probably more correct than Thrasymachus. Except in homogenous, slow-changing societies, an adjudicatory system based on written legal rules, however imperfect, is likely to approach predictability and even-handedness more closely than systems that eschew fixed rules.

Finally, legal rules enhance the efficiency of legal processes. When legal rules are clear and predictably applied, they become weighty 'bargaining chips' in negotiating settlements before or during litigation (Mnookin and Kornhauser 1979) and hence most disputes can be resolved without adjudication, while litigation is more costly when legal predictability declines (Priest 1993).

3. *The Dilemma of Rules and Discretion*

Because human activity is varied and changing, legal rule-makers suffer from relative ignorance of fact and indeterminacy of aim (Hart 1961, p. 125). Rules, therefore, are often underinclusive and often overinclusive—that is, covering activities that are irrelevant or marginal to the primary evil that the rulemakers wished to deal with. Thus, if a motor vehicle code states 'Any person who drives a motor vehicle through a red light is guilty of a violation and shall be fined $50,' it ostensibly compels punishment of a motorist who proceeds carefully through a malfunctioning traffic light which is stuck on red, or an ambulance driver rushing a patient to the hospital (Goffman 1971 p. 102). To apply the rule in such cases is disparaged as legalism (Friedman 1975, pp. 247–8, Nonet and Selznick 1978). Rule-based decision-making can lead to legalism because rules, in their quest for neutrality, often disregard individual litigants' special attributes and circumstances; this makes it easier for officials to apply law's coercive powers, but blunts their sensitivity to justice claims not embodied in the written rules (Noonan 1976).

Many governments seek to prevent the anger generated by legalism by granting legal officials some discretion to make exceptions. American criminal prosecutors, for example, can generally decline to prosecute violations of the law 'in the interests of justice.' Juries, by virtue of the secrecy of their deliberations, can decline to convict when they feel justice requires it (Kadish and Kadish 1973). Another antidote for legalism is to substitute general standards for very specific rules; for example, requiring custody decisions to be based on 'the best interests' of the particular child, taking all relevant facts and considerations into account. Yet granting scores of frontline decision-makers discretion to suspend rules or use their own judgement in implementing vague standards recreates the risk that legal rules were invented to reduce; that is, decisions that vary across different individual decision-makers or erode official policies.

Most legal institutions, therefore, employ a mix of rule and discretion (Mashaw 1983). Some, such as tax collection agencies, may be instructed to emphasize rule-based decision-making, while others, such as juvenile courts or patent offices, may be allowed to rely on experienced legal decision-makers to apply general standards sensibly and consistently. Generally, an institution's legal rules are more prescriptive when lawmakers and politically potent interest groups do not trust it to make acceptable decisions. Thus American regulatory statutes, regulations, and permits often contain more prescriptive legal rules than their Western European counterparts (Rubin 1997, Kagan and Axelrad 2000, Rose-Ackerman 1995, Vogel 1986).

4. *Can Legal Rules 'Decide Cases?'*

Beginning with the American 'legal realists' of the 1920s and 1930s (Ackerman 1974, White 1972, Twining 1973), some scholars have contested the assumption that judges can be 'bound' by pre-existing legal rules. In political science, adherents to the 'attitudinal model' hold that each judge's own political values are the best predictors of individual decisions (Rowland and Carp 1983, Spaeth and Segal 1999). More fundamentally, legal philosophers refer to the 'indeterminacy' of legal rules, arguing that no matter how long one stares at the words of a constitutional bill of rights, a statute, or a set of judicial precedents, the words alone cannot determine which of two conflicting interpretations (and legal outcomes) in a particular case is 'correct.' Moreover, in complex legal systems, it is argued, decision-makers inevitably use their own judgment in deciding which of many rules or precedents 'applies' to a case. The implication is that all legal decisions are policy decisions, rationalized only after the fact by reference to one particular rule or rule interpretation.

Other scholars, however, argue that within any particular legal institution—a national judiciary, an administrative agency—legal professionals reduce indeterminacy by assigning mutually understood meanings to the rules they work with. They learn to do so through social interaction: in law school classes and legal argument; by explaining and defending their rule

interpretations to others; by absorbing institution-specific conventions concerning the purpose of particular rules; and interaction with senior officials or higher authorities who resolve disputes about particular cases. Conventional methods of rule application become part of the socially constructed role of 'judge,' 'regulatory inspector,' or 'police officer.' Legal officials are socialized to regard adherence to their institutions' processes or rule interpretations as obligatory, as essential to their legitimacy as power wielders and their identity as legal professionals (Fuller 1958, Dworkin 1967, Kagan 1978, pp. 88–90, Feeley and Rubin, 1998, p. 338, Kennedy 1997). Indeed, in complex, pluralistic societies, legal officials often find adherence to rules a comforting practice, relieving them of the sense of personal responsibility for decisions which disappoint claimants or impose harsh penalties. Legal officials who stray from their institution's conventions learn that their decisions can be criticized or reversed; thus even judges who wish to maximize their policy preferences must act strategically, avoiding decisions that cannot be linked to plausible interpretations of existing rules (Epstein and Knight 1998).

Conventional rule interpretations do not dictate the result in every case, or totally suppress the individual decision-maker's inclination to decide each case as the decision-maker personally thinks best, but they substantially restrain those impulses. The amount of constraint depends on the degree of homogeneity and hierarchical supervision in particular legal institutions, and on whether the legal rule in question is highly specific (as in highway speed limits) or is complex, requiring the weighing of many facts (as in a standard demanding payment of 'compensation' for accidents caused by 'negligent' driving). Yet even in the US Courts of Appeals, where politically appointed Republican and Democratic judges usually divide on party lines in the 20 percent of cases that generate dissenting opinions, in the other 80 percent judges with different political preferences agree on which legal rules and precedents are controlling (Gottschall 1986). A stark illustration of the constraining effect of 'interpretive communities' and socially constructed judicial roles is provided by Cover's (1975) account of American judges in Northern states in the 1850s; they returned escaped slaves to Southern 'slave-catchers,' notwithstanding the judges' (and their communities') moral opposition to slavery, because a recently enacted federal statute seemed so clearly to require that painful action.

5. *Styles of Rule Application*

For many judges and administrative officials, decision-making is a two-step process: the decision-maker first 'looks backwards' to pre-existing rules to discern what they seem to require, but then 'looks forward' to assess the consequences of the rule-based decision in terms of the decision-maker's understanding of fairness or good public policy (Cardozo 1921, pp. 19–25, Wasserstrom 1961, Kagan 1978). Even if he feels the rule-based decision would not 'make sense,' he might nevertheless feel bound to follow the rule; that would be a legalistic decision-making style. In another style—unauthorized discretion—decision-makers ignore the rule and follow their own judgment (Kagan, 1978, p. 93). A third style, retreatism, is to temporize, push the parties to settle the dispute, or find other ways of avoiding hard decisions. Finally, in what has been called legal policymaking, the legal decision-maker reinterprets the pre-existing rule, adopting innovative readings or principled exceptions which support the preferred outcome (Feeley and Rubin 1998).

Which of these four modes of rule application predominate in a legal institution is influenced by the political system and the legal tradition in which it operates. Common law systems appear to allow judges more scope for creative rule interpretation than Western European judiciaries, which are more closely bound by statutory codes and legal treatises (Damaska 1986). Yet judicial policymaking does occur in civil law judiciaries, such as France's (Lasser 1995), and common law judges in Great Britain tend to feel more tightly bound by existing precedents than their counterparts in the United States—partly because American legal training encourages a more policy-oriented mode of legal reasoning, and partly because the fragmentation of lawmaking power in the United States makes judges less vulnerable to prompt 'overrides' by the legislature than are judges in parliamentary systems (Atiyah and Summers 1987, Damaska 1990).

Rule application styles also vary across legal institutions in the same legal system. Variations in local political culture produce different rule enforcement styles in different police departments (Wilson 1968), criminal courts (Levin 1972), and regulatory agencies (Hutter 1988). Variations in rule application also arise from the personality, role conceptions, and political commitments of individual decision-makers; from the skill and persistence with which particular parties or their legal advocates present their arguments; and from the likelihood that the decision will be reported in the news media or subjected to professional review and criticism. US Court of Appeals judges are less likely to decide cases on the basis of their political values when they serve on panels with judges from another political party (Cross and Tiller 1998). Regulatory agencies tend to enforce rules legalistically when they are subjected to public criticism for perceived laxity (Bardach and Kagan 1982) and to adopt a retreatist style when potential complainants remain silent or violators enjoy the backing of political authorities (Gunningham 1987). Legal officials are more likely to bend the rules when they confront individual subject face to face, as in 'streetcorner

bureaucracies' (Lipsky 1980) than when decision-making is based on paper dossiers or occurs in highly formal settings (Kagan 1978, p. 152). Overall, the way that legal rules are employed is shaped by the interaction of a legal institution's political environment, its clientele, its legal culture, and its mode of recruiting, training, and evaluating its personnel.

See also: Injustice: Legal Aspects; Justice and its Many Faces: Cultural Concerns; Justice and Law; Law: Overview; Legal Positivism; Natural Law; Procedure: Legal Aspects

Bibliography

Ackerman B 1974 Law and the modern mind by Jerome Frank. *Daedalus* Winter: 119–30

Atiyah P S, Summers R S 1987 *Form and Substance in Anglo-American Law: A Comparative Study of Legal Reasoning. Legal Theory, and Legal Institutions*. Clarendon Press, Oxford, UK

Bardach E, Kagan R A 1982 *Going By The Book: The Problem of Regulatory Unreasonableness. A Twentieth Century Fund Study*. Temple University Press, Philadelphia, PA

Cardozo B 1921 *The Nature of the Judicial Process*. Yale University Press, New Haven, CT

Cover R 1975 *Justice Accused: Anti-Slavery and the Judicial Process*. Yale University Press, New Haven, CT

Cross F, Tiller E 1998 Judicial partisanship and obedience to legal doctrine, whistle blowing on the Federal Courts of Appeals. *Yale Law Journal* **107**: 2155

Damaska M 1986 *The Faces of Justice and State Authority: A Comparative Approach to the Legal Process*. Yale University Press, New Haven, CT

Damaska M 1990 Reflections on American constitutionalism. *American Journal of Comparative Law* **38**: 421–43

Dworkin R 1967 The model of rules. *University of Chicago Law Review* **35**: 14

Edelman L, Suchman M 1997 The legal environments of organizations. *Annual Review of Sociology* **23**: 479–515

Epstein L, Knight J 1998 *The Choices Justices Make*. CQ Press, Washington, DC

Euripides 1938 *The Complete Greek Drama*. In: Oates W, O'Neill E (eds.)

Euripides 1995 *The Suppliant Women*. [Trans. Warren R, Scully S] Oxford University Press, New York

Ewick P, Silbey S S 1998 *The Common Place of Law: Stories from Everyday Life*. University of Chicago Press, Chicago

Farber D, Frickey P 1991 *Law and Public Choice*. University of Chicago Press, Chicago

Feeley M M, Rubin E L 1998 *Judicial Policy Making and the Modern State*. Cambridge University Press, New York

Friedman L M 1975 *The Legal System: A Social Science Perspective*. Russell Sage Foundation, New York

Fuller L 1958 Positivism and fidelity to law—a reply to Professor Hart. *Harvard Law Review* **71**: 630–72

Fuller L 1964 *The Morality of Law*. Yale University Press, New Haven, CT

Galanter M 1974 Why 'the haves' come out ahead: Speculations on the limits of legal change. *Law & Society Review* **9**: 95

Goffman E 1971 *Relations in Public*. Harper and Row, New York

Gottschall J 1986 Reagan's appointments to the US Courts of Appeals. *Judicature* June/July: 49–54

Gunningham N 1987 Negotiated noncompliance: A case study of regulatory failure. *Law and Policy* **9**: 69–81

Hoebel E A, 1964 *The Law of Primitive Man*. Harvard University Press, Cambridge, MA

Hart H L A 1961 *The Concept of Law*. Clarendon Press, Oxford, UK

Hutter B 1988 *The Reasonable Arm of the Law? The Law Enforcement Procedures of Environmental Health Officials*. Clarendon Press, Oxford, UK

Kadish M, Kadish S 1973 *Discretion to Disobey: A Study of Lawful Departures from Legal Rules*. Stanford University Press, Stanford, CA

Kagan R A 1978 *Regulatory Justice: Implementing a Wage–Price Freeze*. Russell Sage Foundation, New York

Kagan R A, Axelrad L (eds.) 2000 *Regulatory Encounters: Multinational Corporations and American Adversarial Legalism*. University of California Press, Berkeley, CA

Kagan R A, Skolnick J 1993 Banning smoking: Compliance without enforcement. In: Rabin R, Sugarman S (eds.) *Smoking Policy: Law, Politics and Culture*. Oxford University Press, Oxford, UK

Kairys D (ed.) 1982 *The Politics of Law*. Pantheon Books, New York

Kennedy D 1997 *A Critique of Adjudication*. Harvard University Press, Cambridge, MA

Krygier M 1990 Marxism and the rule of law: Reflections after the collapse of communism. *Law & Social Inquiry* **15**: 633

Lasser M de S-O-1'E 1995 Judicial (self-) portraits: Judicial discourse in the French legal system, *Yale Law Journal* **104**: 1325–409

Levin M 1972 Urban politics and judicial behavior. *Journal of Legal Studies* **1**: 193

Lipsky M 1980 *Street Level Bureaucracy*. Russell Sage Foundation, New York

Markovits I 1986 Pursuing one's rights under socialism. *Stanford Law Review* **38**: 689–761

Mashaw J 1983 *Bureaucratic Justice: Managing Social Security Disability Claims*. Yale University Press, New Haven, CT

Mnookin R, Kornhauser L 1979 Bargaining in the shadow of the law: The case of divorce. *Yale Law Journal* **88**: 950

Nonet P, Selznick P 1978 *Law and Society in Transition: Toward Responsive Law*. Harper Colophon, New York

Noonan J T Jr. 1976 *Persons and Masks of the Law*. Farrar, Straus and Giroux, New York

Priest G, 1993 Justifying the Civil Jury. In: Robert Litan (ed.) *Verdict: Assessing the Civil Jury System*. Brookings Institution, Washington, DC

Rose-Ackerman S 1995 *Controlling Environmental Policy: The Limits of Public Law in Germany and the United States*. Yale University Press, New Haven, CT

Rheinstein M (ed.) 1954 *Max Weber on Law in Economy and Society*. Simon and Schuster, New York

Rowland C K, Carp R A 1983 The relative effect of maturation, time period, and appointing president on district judges' policy choices. *Political Behavior* **5**: 109

Rubin E L 1997 Discretion and its discontents. *Chicago Kent Law Review* **72**: 1299

Sarat A, Kearns T 1993 Beyond the Great Divide. In: Sarat A, Kearns T (eds.) *Law in Everyday Life*. University of Michigan Press, Ann Arbor, MI

Spaeth H, Segal J 1999 *Majority Rule or Minority Will: Adherence to Precedent on the US Supreme Court*. Cambridge University Press, New York

Twining W 1973 *Karl Llewellyn and the Realist Movement*. Weidenfeld, London
Vogel D 1986 *National Styles of Regulation: Environmental Policy in Great Britain and the United States*. Cornell University Press, Ithaca, NY
Wasserstrom R 1961 *The Judicial Decision*. Stanford University Press, Stanford, CA
White G E 1972 From sociological jurisprudence to realism. *Virginia Law Review* **58**: 999
Wilson J Q 1968 *Varieties of Police Behavior*. Harvard University Press, Cambridge, MA

R. A. Kagan

Rumors and Urban Legends

Rumors may be defined as unverified or false brief reports (*rumors*, strictly speaking) or short stories (*urban legends*) with surprising content. They are passed on among a social milieu as true and current and expressing something of that group's fears and hopes.

1. *Mythology and Ethics of Rumor*

Ancient civilizations bequeathed us a mythological and ethical concept of rumor. In *The Aeneid* (IV, 173–197), Virgil (first century BC) personifies the rumor through the features of the goddess Fama, a monster with innumerable eyes, wings, ears, and mouths, and a swift messenger of truth and falsehood. This fabulous description gave inspiration to Shakespeare for *Henry IV* (1598), in giving *Rumor* a character of its own, its costume covered with painted tongues. Judeo-Christian tradition reproves gossip as a sin, ranking it with slander. The Talmud puts *lachon hara* ('ill tongue') on equal terms with atheism, adultery, and murder. In Matthew's Gospel (12: 33–7), it is said that people will also be judged by what comes out of their mouths, and curses those who make a baseless accusation against someone.

The philosophers of the Enlightenment apply the rationalist analysis to 'fables' (myths, legends, rumors). For Fontenelle (*De l'origine des fables* 1684), they are the result of peoples' ignorance, the strength of the imagination, the verbal transmission of narratives, and the explanatory function of the myth. For Voltaire, doubt does not only relate to supernatural legends, but also to historical narratives. In *Le Siècle de Louis XIV* (1751), he shows how a rumor could transform a minor war operation—the French army's crossing of the Rhine River in 1672—into a grandiose epic.

Rumors and legends are deemed to be the result of people's ignorance, fancy and credulity: A philosophical reprobation can be added to the moral reprobation of rumors.

2. *Variety of Psychosociological Approaches to Rumors*

The first scientific studies of rumors started in the 1940s in the USA, during the growth of communication and public opinion researches in social sciences, and being in the state of war, which drew the ruling classes' attention to the prominent part that rumors play.

Allport and Postman ([1947] 1965) formulated concepts that remain relevant today. First, they established that a given subject would more likely be the subject of rumors if perceived as important and ambiguous by the members of a social group. For example, the Japanese attack of Pearl Harbor in 1941 gave way to rumors on account of the importance of this first military attack, which made the Americans fear other assaults, and the ambiguity about the extent of casualties.

Allport and Postman attribute an affective function to rumors, when easing emotional pressure (fear, hostility) of individuals, and a cognitive function, when answering the need for information and explanation.

Lastly, through an experimental study (linear transmission of information by word of mouth), Allport and Postman pointed out three basic processes that lead to the gradual distortion of the message. *Leveling* is the rapid reduction of the volume of details. Rumors are inclined to become a short statement with the form 'subject + predicate': The subject is any person, animal, thing, institution, social or ethnic group; the predicate is the state or action attributed to the subject. *Sharpening* consists of selecting, and often magnifying, details. For example, numerical accentuation increases quantity; temporal accentuation brings rumors up to date; size accentuation exaggerates pronounced features (like a caricaturist who turns a small man into a dwarf). Lastly, *assimilation* is the alteration of details to fit them into the prejudices, stereotypes, and interests of the rumor transmitters. A rumor is the result of a psychological and collective process reflecting the lowest common denominator of ideas, values, concerns and feelings of the people making up a social group.

It remains necessary to analyze the rumor's degree of actuality and to study the mechanisms of distortion of facts or factual information. However, it is most important to understand how and why false news is spread and believed.

Shibutani ([1966], 1977) distinguishes among three different kinds of communication channels, in a descending order of social control and institutionalization: *formal channels* (media, official statements), *auxiliary channels* (clubs, circles), and *informal channels* (the word of mouth). At the normal rate of the communication system, information supply and demand are balanced: the main group of information is conveyed by formal channels, and just a few rumors

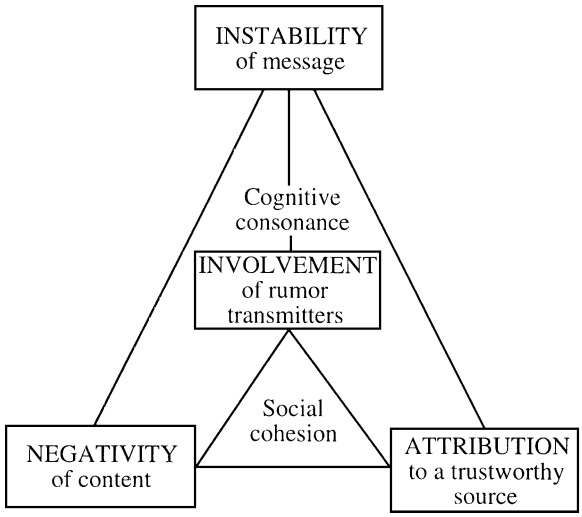

Figure 1
The rumor syndrome adapted from Rouquette (1990)

are spread by informal channels. On the contrary, in a critical rate (war, social or economic crisis, natural disaster, etc.), information demand exceeds supply, formal channels fall into discredit and informal channels abound with rumors.

Today, the distinction between official and informal channels must be moderated to the extent that rumors are also circulating by flyer, fax, e-mail, Internet, and even the mass media.

The inquiry into the 'Rumor in Orléans' by Morin (1971) is a model case study. Rather than the experimental method, he prefers the clinical approach, investigating *in vivo* the spread, then the extinction, of a rumor. Rather than take an interest in critical periods, he studies everyday life rumors, in a French provincial town. In spring 1969, it was said in Orléans that young girls vanished in dress shops owned by Jews to supply a 'white slavery' trade (i.e., prostitution). Morin tried to understand why so many people had believed these false allegations. The rumor drew its scenario from the pseudo-information of popular culture about white young girls abducted and forced into prostitution in exotic countries. Two 'catalyses' converted this scenario into 'reality.' First was a popular magazine's publishing of an unverified anecdote about the failed abduction of a young girl, found unconscious and drugged in back of a dress shop in Grenoble in France. The other was the opening of a clothes department for teenagers, named '*Aux Oubliettes*' (i.e., secret dungeon) in a medieval decorated basement of a clothing shop in Orléans. The rumor did not affect the shopkeeper, who belonged to the traditional bourgeoisie, but fixed itself on several shops owned by Jews. For Morin, the reactivation of two old myths—white slavery and the Jew—allowed the population to symbolically express the deep fears caused by social and cultural changes at the end of the 1960s. Feminine emancipation and clothing and sexual liberalization fascinate and worry mothers and daughters: the rumor explicates the idea that miniskirts lead girls to prostitution. Second, urban modernization and growing anonymity put people in fear of losing the control of their city. The rumor expresses, through metaphors, the 'bad' city hidden behind the 'good' one: the trick dressing room, the vault under the shop, the clothes tradesman who also deals in young women, and the Jew as a figure of the one who hides his difference. One of the fundamental contributions by Morin is to have shown that a rumor is carrying a hidden message, an implicit moral, and that the study of rumors must undertake a hermeneutical work to reveal their symbolic meanings.

Morin has also suggested that, in the fight against a rumor, antirumor (denial) was less effective than counter-rumor. Thus, the rumor of Orléans died out, not so much due to police and press denials, but to an emerging counterrumor claiming that the rumor-mongers were anti-Semites and fascists.

3. A Unified Pattern: The Rumor Syndrome by Rouquette

Rouquette (1975, 1990) has built up an unified pattern of rumor in four typical features (see Fig. 1). When information or a narrative presents these characteristics, it is more than likely a rumor or an urban legend.

Instability refers to the alterations of the message, mainly at the time of its formation but also when the rumor is adapted to a new cultural environment. The presence of variants for a rumor or legend is a basic characteristic of the phenomenon.

Involvement is the fact that the rumor transmitter is feeling concerned by the message. The more rumor-mongers are involved in a rumor, the more they will believe it and pass it on. A rumor is any information or story 'worth telling.' Involvement is at its maximum when a rumor leads to behavior (e.g., panic).

Negativity is a dominating feature of rumors. Searchers estimate that 90 percent of rumors are '*noir*,' reporting news of aggression, accident, failure, scarcity, scandal, etc.; and only 10 percent are '*rose*,' containing optimistic content and attaining imaginary shared hopes and desires.

Lastly, *attribution* is the alleged source of a rumor. The rumor transmitters mention a trustworthy source, who assures the narrative's truth.

Rouquette's pattern offers a strong consistency of its features through six links.

Instability and involvement. Contrary to accepted opinion, the transmission of a message is less accurate when the subjects who spread the rumors are involved —there is an unconscious interference by interest, in their remembrance of the subject—than when they are

not concerned by the rumor. In transforming the message, the rumor transmitters tend to keep their cognitive consonance, in other words, the mental consistency of information, opinions, and beliefs they have about a situation.

Instability and negativity. The distortions of the message are commonly negative.

Instability and attribution. The rumor changes and enriches its attributions as it circulates. For example, a relay becomes a source (when rumors are passed on by the press or by a company's internal memos), a denial becomes a confirmation (in a French variant of the flyer about 'LSD tattoos,' the heading 'The Narcotic Bureau Confirms' appeared just after the organization denied the allegations. Cf. Campion-Vincent and Renard 1992).

The next three links show the role that rumors play in social cohesion.

Involvement and negativity. The discrediting of 'others' is implicitly corollary to giving importance to 'ourselves.' This sets up a mental and moral convergence between the rumor transmitter and his or her audience, therefore strengthening the social cohesion.

Involvement and attribution. Calling upon a trustworthy source weaves social ties that give value to the rumor transmitter.

Negativity and attribution. The more the rumor is negative, the stronger the guarantee of veracity.

Rouquette (1989) has also expressed the idea that rumors are multiple 'answers' to a small number of 'problems': how to account for an unexplained fact, how to vindicate an attitude or a belief, how to instigate behavior (warning rumors), how to intensify the social differentiation, or how to favor social integration. These problems take shape through themes, whether they are universal or specific of a given culture at a given time. In the Western modern world, the recurring themes of rumors and urban legends relate to just a few preoccupations (Kapferer 1990, Renard 1999): health, sex, money, new techniques, strangers, urban violence, nature, the change of manners, and the supernatural.

4. *Folklore's Contribution to the Study of Rumors*

Since rumors and legends are subjects of research that usually depend on distinct sciences—sociology for the former and folklore for the latter—their profound identity has been hidden for a long time. It was not until the emergence of a new concept—'modern or urban legend' (Mullen 1972, Brunvand 1981)—that sociologists and folklorists collaborated profitably and renewed their perspective on their own research topic. Mullen has emphasized the great resemblance between rumor and legend: Both are the result of a collective interactive process, they are usually first spread by word of mouth, their message is passed on as the truth, and their social functions are similar. Rumor and legend are two aspects of the same phenomenon but one aspect sometimes outweighs the other. Studying the legendary aspect of rumors reveals their mythical deep-rootedness. Studying the 'rumoral' aspect of legends brings enlightenment to the transformations of content and the production of variants.

This aspect appears in the following tendencies: briefness of content, instability, and actualization, active and brief circulation in a certain time and place and among a given social group. The legendary aspect expresses complementary tendencies: the patterning of the content into a narrative form; the exploitation of motifs and themes, sometimes of ancient origin, emerging from the collective imagination and having strong symbolic contents; insertion into long, sometimes secular, periods of time, spreading internationally (Brunvand 1981, 1993, etc.).

A rumor may expand—become 'authentic'—and embody itself in a legendary narrative; reciprocally, a legend may be reduced to a mere statement. For instance, 'There are alligators in the sewers of New York' is the brief rumor form of the story reporting that New Yorkers brought back baby alligators from Florida, and then disposed of them when they became burdensome by flushing them down the toilet, where they ended up in the sewer (Brunvand 1981).

A similar legend may arise in both a traditional version and a modern version. For instance, the narrative motif 'Animal lives in person's body' is treated in Africa as witchcraft, whereas in the West it is explained by pseudomedical reasons.

In his pertinently headed article 'From Trolls to Turks,' Tangherlini (1995) shows many correlations between rumors and urban legends, circulating in modern Denmark, and the traditional narratives of the Danish folklore. He proves that beliefs about immigrants are revived from old motifs concerning supernatural beings (elves, trolls, witches): loathsome food customs, abduction of women or children, physical violence.

It is not accidental that interest in urban legends first arose among the folklorists, at least in English- and German-speaking countries. Rumors and contemporary popular narratives are actually modern equivalents of ancient legends, all of them expressing social or symbolic thought.

See also: Collective Beliefs: Sociological Explanation; Collective Memory, Psychology of; Panic, Sociology of; Prejudice in Society; Public Opinion: Microsociological Aspects

Bibliography

Allport G W, Postman L [1947] 1965 *The Psychology of Rumor*. Russell and Russell, New York

Brunvand J H 1981 *The Vanishing Hitchhiker: American Urban Legends and Their Meanings*. Norton, New York
Brunvand J H 1993 *The Baby Train and Other Lusty Urban Legends*. Norton, New York
Campion-Vincent V, Renard J-B 1992 *Légendes urbaines: Rumeurs d'aujourd'hui*. Payot, Paris
Kapferer J-N 1990 *Rumors: Uses, Interpretations and Images*. Transaction Publishers, New Brunswick, NJ
Morin E 1971 *Rumour in Orléans*. Blond, London
Mullen P B 1972 Modern legend and rumor theory. *Journal of the Folklore Institute* **9**: 95–109
Renard J-B 1999 *Rumeurs et légendes urbaines*. Presses Universitaires de France, Paris
Rosnow R L, Fine G A 1976 *Rumor and Gossip: The Social Psychology of Hearsay*. Elsevier, New York
Rouquette M-L 1975 *Les Rumeurs*. Presses Universitaires de France, Paris
Rouquette M-L 1989 La rumeur comme résolution d'un problème mal défini. *Cahiers internationaux de sociologie* **86**: 117–22
Rouquette M-L 1990 Le syndrome de rumeur. *Communications* **52**: 119–23
Shibutani T [1966] 1977 *Improvised News: A Sociological Study of Rumor*. Bobbs-Merrill, Indianapolis, IN
Tangherlini T R 1995 From trolls to Turks: Continuity and change in Danish legend tradition. *Scandinavian Studies* **67**(1): 32–62

J.-B. Renard

Rural Geography

While it is widely accepted today that those working in the rural context can make a significant contribution to the theory and practice of human geography more broadly, this has not always been the view. At times rural geography has been regarded as a poor relation of modern geography, (see Cloke 1980) trailing behind the 'real' intellectual advances in the discipline and generally presenting a largely traditional and often atheoretical perspective on key debates. Moreover, the very existence of a sub-discipline of geography defined according to a set of shifting, obscured, and contested spatial boundaries has been questioned at times, particularly in the context of geographers' interests in national, international, and global systems and processes. This review of rural geography will focus on how the sub-discipline has developed and, in particular, how its various phases have been influenced by broader theoretical and conceptual ideas within geography as a whole.

1. From Regionalism to Political Economy

The regional approaches that dominated geography before 1960 favored rural spaces and places. Descriptive studies of land use frequently focused on the countryside, while work on population change and settlement also tended to be approached from a rural perspective. Such dominance was put down to the importance of agriculture as a land use and to the emphasis that regional approaches placed on describing identifiable physical characteristics. The demise of regional geography in the 1960s, however, shifted attention away from the countryside; new directions in human geography saw the introduction of more 'relevant' approaches in which the sheer size and visibility of urban problems ensured that they received most attention. During the next 15 years interest in rural geography declined while the work that did continue became:

> entrenched in its agricultural roots and seemingly oblivious to the theoretical strides being achieved within urban studies (Cloke 1989 p. 65).

Following this rather fallow period, rural geography experienced something of a resurgence during the 1970s (particularly in Britain, Canada, and Europe). Textbooks specifically devoted to rural areas emerged (for example, Clout 1972) claiming that the urban emphasis of recent geographical research had resulted in an imbalance in terms of wider geographical inquiry and a profound absence of knowledge about rural places and people. This renewed focus on rural geography yielded a large amount of data on the countryside in developed countries, but rural studies at this time continued to follow a highly traditional pattern. Much of the research was descriptive, establishing rural areas as interesting places to study but lacking explanation and avoiding theoretical debate. Emphasis was placed on identifying the rural as different or unique (leading to rather sterile debates on the definition of 'rurality,' see Cloke 1977) rather than on showing how an understanding of patterns and processes operating in rural areas could inform human geography more broadly.

The willingness of rural geographers to engage with theory marks perhaps a key watershed in the development of the sub-discipline. During the mid 1980s those working within the rural context began to argue that rural research needed to be informed by theoretical approaches being developed outside the field and that a failure to locate studies within a broader theoretical and conceptual framework was seriously undermining their value. Although some reference was made through the 1960s and 1970s to debates taking place in radical rural sociology—especially the use of Marxist approaches in the study of agriculture—the theoretical concepts used were not generally taken up by rural geographers. It was thus only with the development of political economic approaches that the need to engage with theory outside rural geography started to influence its scope and direction to any significant degree. An awareness of the importance of processes of restructuring and capital accumulation to rural economies and societies

had begun to demonstrate to rural researchers the relevance of macroeconomic and political transformations to the detail of change within the countryside. So, considerably later than their urban counterparts, rural researchers turned to political economy approaches to inform their analyses and explanations. Space is insufficient here for a detailed review of the nature and contribution of political economic approaches to the study of rural geography (but see Cloke and Little 1990, Marsden et al. 1986, Phillips and Williams 1984, Urry 1984). It is possible, however, to identify a number of key themes that emerged (and were sometimes transformed) through the use of such approaches.

2. *Economic and Social Restructuring*

During the 1980s changes in the patterns and processes of capital accumulation were visible in the transformation of agricultural production and shifts in the location of industry. Understanding changes in rural areas required grounding such changes conceptually within more global patterns of restructuring and capital accumulation, essentially within a political economy framework. Particularly important in this context were arguments surrounding the uneveness of restructuring and the ways in which global processes of transformation impacted differently on different places at different times.

The adoption of political economy approaches provided research on agricultural geography in particular with a rich and highly fruitful arena for debate (see, in particular, Buttel and Newby 1980, Goodman and Redclift 1985, Marsden et al. 1986) as studies sought to detail and explain the major shifts taking place in the agricultural and associated industries. Research concentrated initially on the capitalization of agriculture and on the implications of new technologies on the methods of farming and on the ownership of the means of agricultural production. Later, however, the debate expanded to consider the broader transformation of food production including issues such as the role of 'downstream' and 'upstream' industries together with the social relations of agricultural production (especially changes to the family farm) (Goodman et al. 1987, Whatmore et al. 1987). Although not all work on agriculture adopted an explicitly political economic frameworks (indeed some consciously rejected political economy perspectives), the influence on this aspect of rural geography was profound, providing a much more informed understanding of the changes taking place in rural areas and of their links with global patterns of economic change.

The application of political economy approaches to another major research direction in rural geography, the restructuring and relocation of industrial development, was also reflected in increased output (see, for example, Bradley and Lowe 1984). More importantly, perhaps, it encouraged research into the distributional impacts of shifts in the location of industrial capital, including a consideration of the effects on different sections of the labor market and the implications for the changing social structure of rural areas. Such issues were also incorporated in the second key theme to be highlighted in the discussion of political economy and rural geography, that of class and social inequality.

3. *Inequality in Rural Society*

The renewed focus on rural geography that took place in the 1970s and early 1980s, produced a number of studies from Britain and the USA on rural deprivation and 'welfare' issues, but it was only with the application of political economic approaches that such studies began to address issues of social inequality and to develop class-based analyses of the causes and effects of rural deprivation. For the first time research included not only an examination of the nature and extent of deprivation in rural areas but also the relationship between individuals and groups according to access to resources. A 'service' based model of rural deprivation in which the welfare of rural people was seen to be a function of the availability of certain essential public and private services, was gradually replaced (or in some cases augmented) by one that related deprivation to people's personal access to resources in terms of income, property, status, and power (Moseley 1982, Newby 1979). Detailed studies of access to housing, employment, transport, health care, etc., emerged through which the inequalities existing in contemporary rural societies were highlighted (for summaries see Champion and Watkins 1991, Troughton 1995). Such studies were then used in the broader examination of access to the countryside and of the transformation of rural societies and the gentrification of rural communities (see Little 1987a, Marsden et al. 1993).

4. *State–Society Relations*

Political economy approaches highlighted the centrality of power relations to the composition and well-being of the rural population. While the class-based analysis of access to rural resources contributed to an understanding of rural residents' personal power (especially financial but also in terms of status), the wider discussion of the structures and operation of power prompted researchers to explore the relationship between the state and the rural community (see Cloke and Little 1990). Importantly, this work drew on urban-based debates on the nature of the local state and on the changing mechanisms and practices of central government. At a time when the developed economies, particular the USA and Britain, were seeing attempts by the center to curtail the responsibilities of the local state, rural research began to

investigate the implications for rural areas and to document the localized power struggles and resistance that accompanied the changing balance of power.

Political economy approaches to central–local relations provided a framework for the investigation of power in a variety of ways. Studies focused on inter agency relations, the changing role of the private sector, and the role of key actors, elites, and institutions within the policymaking process. An area of particular attention, (especially within British rural geography), was that of town and country planning, and a number of studies on the negotiation and distributional consequence of planning decisions contributed to the wider understanding of power relations and political control within the regulatory process in the countryside. One particular strand of such research sought to highlight the role and status of environmental pressure groups in an examination of the priorities of the dominant interests in the countryside. Much of this work revolved around a discussion of the changing status and power of agriculture and raised questions concerning the position of farmers and landowners in relation to other interests within both urban and rural communities.

Forms of explanation that originated beyond the rural led to questions concerning the validity of 'the rural' as a legitimate category of analysis or even subject of research activity (see Hoggart 1990). The spatial uneveness of patterns and processes of capital accumulation and the specificity of local level responses to economic and social restructuring were used in defence of a focus on the rural, but research moved away from attributing causal properties to rural areas in the explanation of issues such as population migration, inequality, and access to resources. As well as contributing in its own right to the development of rural geography, the adoption of political economy perspectives, it may be argued, helped to open up the subject to other theoretical approaches. Particularly important here was the emergence of feminist perspectives in rural geography in the late 1980s (see Darque and Gasson 1991, Little 1986,1987b, Whatmore 1991).

5. *Feminist Rural Geography*

The experience of women in the countryside were referred to fleetingly in the early development of rural geography, in, for example, the discussion of labour relations within family farm businesses (Sachs 1983, Symes and Marsden 1983) and the sociology of the rural community (see Davidoff et al. 1976). Only in the late 1980s, however, did a sustained attempt to incorporate explicitly feminist perspectives into the subject emerge. As with the adoption of political economy approaches, the introduction of feminist perspectives was a response to theoretical ideas being applied within other areas of geography and drew on a wide range of literature from urban/social geography as well as from other academic disciplines such as sociology and women's studies. It also reflected a need to go beyond the description of the constraints and opportunities experienced by women living in the countryside in explaining their position. Useful data on the detail of women's lives had been collected in research on issues such as employment, access, and service provision (for a summary see Little 1991). Rural women in developed countries, for example, had lower activity rates than their urban counterparts and that their paid work was generally characterized by poor conditions and low wages. Women's lives were also disproportionately constrained by limited transport and other public and private services. It was argued, however, that such work failed to go beyond the examination of women's roles and that it provided little explanation for the basis of women's inequality. Lacking alternative forms of explanation, such work was in danger of attributing 'rurality' with causal powers and identifying characteristics of rural economies and societies as wholly responsible for the experiences of rural women.

In keeping with work in other parts of the discipline, feminist rural geographers argued that studies of women needed to be located within a broader consideration of gender relations and their operation in the countryside. An understanding of women's experiences in all areas from their contribution to the family farm to their access to child care, had to recognize, it was asserted, the underlying power relations that exist between men and women and needed to acknowledge the role of patriarchy as a fundamental element of social control in the countryside (see Little 1987b, Whatmore 1991). Having argued that the experiences of rural women needed to be conceptualized in terms of more universal patriarchal systems and gender relations, the research also sought to demonstrate the specificity of the experiences of rural women, showing how rural communities exhibited a particularly traditional form of gender relations and were dominated, as a result, by more conventional gender roles and attitudes towards gender equality. Women's position was seen as a function of a particular view of rural communities, one in which the family, and women's reproductive role within it, were central. Research also argued, in the discussion of the daily detail of women's lives and experiences, that women's ability to take part in paid work as well as their position within the rural community were very much shaped by this dominant understanding of the 'appropriate' role of women in the countryside.

Research on rural women and gender relations has continued to develop this theme and in so doing has drawn on and contributed to the 'cultural turn' within rural geography. More recent research has explored notions of gender identity in the context of the (re)construction of rural cultures, suggesting that certain aspects of women's (and men's) gender ident-

ities are prioritized within certain dominant cultural constructions of rurality within developed economies (see Hughes 1997, Agg and Phillips 1998, Little 1997). Again, such work has explored the implications of more traditional views of gender identity within rural culture for the reality of women's lives. Furthermore, some researchers have shown how dominant constructions of the relationship between gender and rurality will be experienced differently by groups of women and men; such as gay men and lesbians (see Bell and Valentine 1995, Kramer 1995) or young women (Dahlström 1996). Increasingly, work on gender identity and on sexuality has led to a discussion of the construction of femininity and masculinity within a rural context (Brandth 1995, Woodward 1998).

6. *Rural Geography and the Cultural Turn*

Recent evidence of another 'resurgence' in rural geography (documented by Cloke 1997) has accompanied the adoption of cultural approaches in geography and the social sciences generally. Cloke maintains that three important foci of cultural studies have provided a particular impetus for the study of rural spaces and societies: landscape, the rural other, and the spatiality of nature. In addition, a willingness by rural researchers to engage with important theoretical dualisms such as society/space, nature/culture, self/other has brought rural studies back into the mainstream of geographical inquiry.

A widely quoted springboard for new cultural approaches in rural geography came in the form of Chris Philo's 1992 article on 'neglected rural geographies.' In this short piece Philo challenged the implicit and explicit preoccupation of rural geographers with the experiences of 'Mr Average,' arguing that both the content and the practice of rural geography were dominated by middle-aged, middle-class, white, males. He suggested that this emphasis be redressed in the study of 'other' elements of the rural population and that in so doing researchers take on board the variety that exists in the composition and experience of the rural population. The fact that such variety has been 'hidden' from both academic and lay discourses of rural spaces reflects the dominance of particular constructions of rurality.

Following Philo's article there has indeed been a concerted attempt by rural geographers to focus on the experiences of previously neglected, and frequently marginalized groups. Studies of, for example, young people (Jones 1999, Valentine 1997), black and ethnic minority people (Snipp 1996), New Age Travellers, (Halfacree 1996) have helped to increase our understanding of the diversity of rural lifestyles. Moreover, many studies of 'other' groups within rural society have attempted to locate their findings within a theorization of difference and in so doing to draw on critical debates in social and cultural theory. These attempts to theorize the experience of difference and diversity have demonstrated the existence of multiple ways of seeing rural spaces and drawn attention to the important power relations involved in the privileging of certain views over others. Thus, it has been argued that the experiences of groups such as the young or black people in the countryside are profoundly shaped by their (in)ability to be accommodated within dominant construction of the rural. This sense of being 'out of place' will influence their own perceptions or constructions of rural spaces and reinforce their marginalization.

An interest in the dominant, multiple, and contested constructions of rurality has led to (and drawn from) an interest in the different representations of the countryside. Studies have sought to demonstrate the various ways in which rural spaces and people have been represented in different arena, identifying the key signs and symbols that are selected to reflect rurality and exploring their cultural significance. Research has also attempted to show how these images have been used in particular situations and how they have contributed to the projection of idealized and sanitized rural spaces, masking the existence of poverty and marginalization and creating a sense of staged authenticity that differs drastically from 'reality.' Their ability to influence lived behavior at the local level in addition to some sort of wider notion of rurality has been a central concern of some studies as they investigate the circulation of symbolic meanings.

Debates surrounding the construction of rurality and the power relations bound up in the representation of rural spaces and people have raised questions about the role of human agency and the relationship between humans and non-humans in the control of the countryside. Nowhere is this more relevant than in the study of 'nature' and the direction of environmental change (see Murdoch 1997, Whatmore and Thorne 1998). While not confined to rural/environmental studies, issues concerning the nature/society dualism and the changing relationship between 'humans' and 'animals' have opened up major areas of research within rural geography from the development of food networks and the 'quality' of agricultural produce (Arce and Marsden 1995) to the role and purpose of different forms of landscape/environmental protection (Winter and Gaskell 1998).

Exploring the different conceptualizations, structures, and mechanisms of power has once again highlighted the theme of policy and governance within rural geography. As noted above, political economy approaches encouraged the examination of both formal and informal structures of power and their operation within the policy process. Recently there has been a re-examination and development of such work in relation to a broader conceptualization of governance and to the changing social, economic, and cultural context of rural policymaking (see Goodwin

1998). Research has considered, for example, the changing practices and mechanisms of governance in rural areas in the analysis of partnerships and the growth of community participation. Such work has also been informed by cultural perspectives in the examination of the values and attitudes underlying shifts in the regulatory framework.

Clearly, the cultural turn in rural geography has had a major impact on the direction and content of research within the subdiscipline. It has also had a far-reaching influence on the methods employed by rural researchers; on the way data is collected, analyzed, and used. In common with other areas of geography, there has been a significant growth in the adoption of ethnographic methods in rural research and in the use of 'grounded theory' and while such approaches may have characterized some of the early rural 'community studies,' it is only recently that they have been more widely applied to a range of economic, environmental, and policy-based issues in rural studies. Those working on rural lifestyles and marginalization have argued, in particular, that a real understanding of the lives of those living in the countryside needs to emerge from the detailed analysis of their behavior, attitudes, and beliefs and that this can only be obtained through ethnographic methods. In addition, rural researchers have started to make use of a variety of other sources of data, drawing on art, literature, diaries, and other cultural artefacts, both historical and contemporary, in their work. With a shift in the methods used by (some) rural researchers has come something of a change in attitudes towards research. The sensitivity of work on marginalization and deprivation has opened rural geographers' eyes to the potential dangers of research tourism and encouraged them to think more carefully about the way their work may be used to the advantage or disadvantage of particular groups at particular times.

7. Conclusion

This short article has highlighted some of the main 'stages' in the evolution of rural geography since the 1950s. The goal has been to identify key shifts in thought within rural geography and show how they have influenced the nature and direction of rural studies. An impression that the subject has moved smoothly from one dominant theoretical influence to another is inevitable but misleading.

Rural geography is in a buoyant period, having been theoretically invigorated by the application of perspectives from cultural and feminist studies. Rural researchers appear more open to the influence of perspectives from outside the narrow confines of the subject and the willingness to engage with and contribute to debates in other parts of academia as well as other areas of geography, has raised the profile of rural geography generally and ensured that the wider applicability of their arguments is acknowledged.

See also: Development and Urbanization; Development: Rural Development Strategies; Geography; Internal Migration (Further Types): Industrialized Countries; Internal Migration (Rural–Urban): Industrialized Countries; Peasants and Rural Societies in History (Agricultural History); Rural Planning: General; Rural Sociology; Urban Geography

Bibliography

Agg J, Phillips M 1998 Neglected gender dimensions of rural social restructuring. In: Boyle P, Halfacree K (eds.) *Migration into Rural Areas: Theories and Issues*. Wiley, Chichester, UK

Bell D, Valentine G 1995 Queer country: rural lesbian and gay lives. *Journal of Rural Studies* **11**: 113–22

Bradley T, Lowe P (eds.) 1984 *Locality and Rurality*. Geo Books, Norwich, UK

Brandth B 1995 Rural masculinity in transition: gender images in tractor advertisements. *Journal of Rural Studies* **11**: 123–33

Buttel F, Newby H (eds.) 1980 *The Rural Sociology of the Advanced Societies*: Critical Perspective. Allenheld Osmun, Montclair, USA

Champion T, Watkins C (eds.) 1991 *People in the Countryside*. Chapman, London

Cloke P J 1977 An index of rurality for England and Wales. *Regional Studies* **11**: 31–46

Cloke P J 1980 New emphases for applied rural geography. *Progress in Human Geography* **4**: 181–217

Cloke P 1989 Rural geography and political economy. In: Peet R, Thrift N (eds.) *New Models in Geography 1*. Unwin Hyman, London

Cloke P 1997 Country backwater to virtual village? Rural studies and 'the cultural turn.' *Journal of Rural Studies* **13**: 367–75

Cloke P, Little J 1990 *The Rural State: Limits to planning in rural society*. Oxford University Press, Oxford, UK

Clout H D 1972 *Rural Geography: An Introductory Survey*. Pergamon, Oxford, UK

Dahlström M 1996 Young women in the male periphery: Experiences from the Scandinavian North. *Journal of Rural Studies* **12**: 259–71

Darque M, Gasson R (eds.) 1991 Women in Agriculture. Special issue of *Journal of Rural Studies*. **7**: 1–137

Davidoff L, L'Esperance J, Newby H 1976 Landscape with figures. In: Mitchell J, Oakley A (eds.) *The Rights and Wrongs of Women*. Penguin, Harmondsworth, UK

Goodman D, Sorj B, Wilkinson J 1987 *From Farming to Biotechnology: A theory of agro-industrial development*. Blackwell, Oxford

Goodman D, Redclift M 1985 Capitalism, petty commodity production and the farm enterprise. *Sociologia Ruralis* **15**: 231–247

Goodwin M 1998 The governance of rural areas: some emerging research issues and agendas. *Journal of Rural Studies* **14**: 5–12

Halfacree K H 1996 Out of place in the country: travellers and the rural idyll. *Antipode* **28**: 42–72

Hoggart K 1990 Let's do away with rural. *Journal of Rural Studies* **6**: 245–57

Hughes A 1997 Rurality and cultures of 'womanhood.' In: Cloke P, Little J (eds.) *Contested Countryside Cultures: Otherness, Marginalisation and Rurality*. Routledge, London

Jones O 1999 Tomboy tales: the rural, nature and the gender of childhood. *Gender, Place and Culture* **6**: 117–36

Kramer J 1995 Bachelor farmers and spinsters: gay and lesbian identities and communities in rural North Dakota. In: Ball D, Valentine G (eds.) *Mapping Desire*. Routledge, London
Little J 1986 Feminist perspectives in rural geography: an introduction. *Journal of Rural Studies* **2**: 1–8
Little J 1987a Gentrification and the influence of local level planning. In: Cloke P (ed.) *Rural Planning: Policy into Action?* Harper and Row, London
Little J 1987b Gender relations in rural areas: the importance of women's domestic role. *Journal of Rural Studies* **3**: 335–42
Little J 1991 Women and the rural labor market: a policy evalution. In: Champion T, Watkins C (eds.) *People in the Countryside*: *Studies of Social Change in Rural Britain*. Paul Chapman, London
Little J 1997 Employment, marginality and women's self-identity. In: Cloke P, Little J (eds.) *Contested Countryside Cultures: Otherness, Marginalisation and Rurality*. Routledge, London
Marsden T, Arce A 1995 Constructing quality: emerging food networks in the rural transition. *Environment and Planning A* **27**: 1261–79
Marsden T, Munton R, Whatmore S, Little J 1986 Towards a political economy of capitalist agriculture: a British perspective. *International Journal of Urban and Regional Research* **10**: 498–521
Marsden T, Murdoch J, Lowe P, Munton R, Flynn A 1993 *Constructing the Countryside*. UCL Press, London
Moseley M J (ed.) 1982 *Power, Planning and People in Rural East Anglia*. Centre of East Anglian Studies, Norwich, UK
Murdoch J 1997 Inhuman/non-human/human: actor network theory and the prospects for a non-dualistic and symmetrical perspective on nature and society. *Environment and Planning D* **15**: 731–56
Newby H 1979 *Green and Pleasant Land? Social change in rural England*. Hutchinson, London
Phillips D, Williams A 1984 *Rural Britain: a social geography*. Blackwell, Oxford, UK
Philo C 1992 The child in the country–ward, C. *Journal of Rural Studies* **8**: 193–207
Sachs S E 1983 *Invisible Farmers. Women's Work in Agricultural Production*. Rhinehart Allenheld, Totowa, NJ
Snipp C M 1996 Understanding race and ethnicity in rural America. *Rural Sociology* **61**: 125–42
Symes D G, Marsden T K 1983 Complementary roles and asymmetrical lives: farmer's wives in a large farm environment. *Sociologia Ruralis* **23**: 229–41
Troughton M J 1995 Rural Canada and Canadian rural geography: an appraisal. *The Canadian Geographer* **39**: 290–305
Urry J 1984 Capitalist restructuring, recomposition and the regions. In: Bradley T, Lowe P (eds.) *Locality and Rurality*. Geo Books, Norwich, UK
Valentine G 1997 A safe place to grow up? Parenting, perceptions of children's safety and the rural idyll. *Journal of Rural Studies* **13**: 137–48
Whatmore S 1991 *Farming Women: Gender, Work and Family Enterprise*. Macmillan, Houndmills, UK
Whatmore S 1991 Life cycle or patriarchy? Gender divisions in family farming. *Journal of Rural Studies* **7**: 71–6
Whatmore S, Thorne L 1998 Wild(er)ness: reconfiguring the geographies of wildlife. *Transactions of the Institute of British Geographers* **23**: 435–54
Whatmore S, Munton R, Marsden T, Little J 1987 Interpreting a relational typology of farm businesses in Southern England. *Sociologia Ruralis* **27**(2/3): 103–22
Winter M, Gaskell P 1998 The *Agenda 2000* debate and CAP reform in Great Britain: is the environment being sidelined? *Land Use Policy* **15**: 217–231
Woodward R 1998 'It's a man's life!': soldiers, masculinity and the countryside. *Gender, Place and Culture* **5**: 277–300

J. Little

Rural Industrialization in Developing Nations and Gender

1. *Expansion of Rural Industrialization*

During the 1970s, transformative industries spread throughout the rural zones of many developing nations. These enterprises shared five features: they all involved industries engaged in the manufacture of basic consumer goods; the goods were produced in small- to medium-sized firms; working conditions and wages were marginal relative to those in the formal economy of state-regulated employer–worker relations; many, though not all, of the workers were women laboring in their own homes; and the women recruited for such labor tended to be young and single (Arias 1988, Wilson 1991).

Women have always participated in the domestic chores and contributed to the income of rural households, but heretofore their work was unpaid and involved activities that were labeled as 'female.' Rural industrialization changed all that and unleashed paid labor and contractual employment outside of the household yet within the rural setting (Bachrach Ehlers 1993, Boris and Daniels 1989). Until then, female contact with paid work and manufacturing employment only occurred through migration to the city.

The most common transformative industries now present in developing rural areas are, in decreasing order: the canning of fruits, vegetables, meats, and marine products; making clothing and fabrics; the manufacture of goods, candies, and electronics; and the transformation of popular crafts into marketable 'folk art.' Until recently, this manufacturing generally occurred in 'traditional' industries, which relied on the intensive use of poorly-paid labor as the preferred strategy for adapting to markets that are substantial, but fragmented and ever-changing, and to consumer demands that may be intense but ephemeral. Under these circumstances, access to a ready niche of low-cost labor has been crucial to sustaining the economic viability of these firms (Bradley 1989).

The expansion of industry into rural areas has been particularly intense in Asia (notably in China, Bangladesh, India, Indonesia, Malaysia, and Taiwan) and Latin America (especially Mexico, Paraguay, Puerto Rico, and the Dominican Republic), as well as in the

less-developed countries of southern Europe (Greece, Portugal, Turkey, and to a lesser extent Spain and Italy). Rural industrialization is more common in countries with high fertility and a large share of the population living outside of cities.

The association between rural industrialization and female employment across such a wide variety of countries suggests that, under the present global economic regime, certain industries have come upon a powerful political-economic strategy that takes advantage of and reinforces gender relations within rural environments. As a result, rural women have increasingly become one of the best options for securing a steady supply of labor that is cheap, docile, flexible, and unorganized, that is widely available in many developing settings (Dwyer and Bruce 1988, Mencher 1988).

The rapid expansion of rural industrialization and its effective rooting in many peasant societies has been interpreted from two principal perspectives: that of the labor market and that of the household. The notion of the *social construction of difference* is essential to understanding the issue that both perspectives seek to theorize: how to construct and maintain a workforce that will accept the low salaries, unstable working conditions, and discriminatory treatment of women that prevail in rural manufacturing?

The endless search for and creation of niches of cheap manual labor is made possible by what Maruani (1991) calls the 'social construction of difference,' which can be understood as a dynamic mechanism permitting the selective definition of jobs and wages along the lines of gender, race, legal status, and other demographic criteria in such a way as to invent, recreate, and conserve hierarchies that legitimate segregation, inequality, and discrimination. Through the social construction of difference, rural industrialists have been able to keep labor costs down by creating social categories of people who can be offered lower wages for work under precarious conditions (Maruani 1991).

2. *The Labor Market Perspective*

The creation of niches of low-cost manual labor might be explained simply as a consequence of macroeconomic changes and recent policies of structural adjustment, not only in poor sectors, but in the traditional sectors that have assured the survival of rural society despite its long state of crisis. Indeed, the dyad of rural industrialization/female work appears to be strongly associated with the deterioration and stagnation of tasks characteristic of rural life: agriculture, cattle raising, and sheep herding. This association implies an irremediable crisis of the proverbially male occupations of rural society that affirmed the role of men as principal providers for their households.

In the context of local poverty and family crisis, women have been compelled to accept whatever offers of work have come along, however undesirable. This acceptance was facilitated by rural women's low level of education and information about the norms and rules of paid labor in formal urban settings. But education and information are not the only factors. Conserving female labor as a low-cost resource results from a social process of 'deskilling' of those tasks carried out by women.

3. *Deskilling as a Social Construction*

In rural society, it is commonly accepted that female employment occurs thanks to abilities acquired in the home, such as housekeeping, a knowledge of the kitchen, and dexterity at sewing, which together open the door to cannaries and workshops that make shoes, clothes, toys, and other consumer goods. As these abilities are taken to be 'natural,' or at least learned at home, employers treat them as 'unskilled' jobs to which a low value may be assigned relative to the manufacturing tasks learned by men.

This 'cheapening' of female labor also arises from seeing the work of women in terms of 'feminine' attributes: that they are careful, attentive, fine, delicate, and sensitive. These qualities are seen as part of the innate character of woman, rather than as part of the skill set associated with paid labor and priced in the market.

Another factor in women's low wages and hours of work is their responsibility for reproductive tasks and to domestic responsibilities (England 1993). Indeed, Dwyer and Bruce (1988) have calculated that most of the part-time labor done in the world today is carried out by women. This limits the possibilities for women to approach work as a career developed over the long term. Because of domestic responsibilities, in-home paid labor, which is easier to combine with domestic tasks, has been embraced among rural women. They know that it is poorly paid, but for some it is the only combination possible for making money while not neglecting household obligations (Fernández-Kelly and García 1989).

In this way, there has been an effective adaptation of the rural market for manufacturing labor to the situation of rural women. The instability of the rhythms of production, a key entrepreneurial element in cheapening the cost of manual labor, is manipulated as a mechanism to achieve labor 'flexibility,' which in turn emanates from and is adapted to the needs of female workers. The irregularity and instability of production, the continual opening and closing of factory doors, and payment by piece rate (according to the number of units produced) all form part of the prevailing economic model, which comes to exemplify and derogate the women themselves: it is *they* who are unskilled, *they* who require labor flexibility, and *they*

who earn what they want, when they want it. In this way, rural manufacturing has been able to maintain and recreate the precarious conditions of the labor market such that they are viewed as a concession on the part of employers to women's domestic responsibilities. This social disqualification of female labor might diminish over time, but that process is slowed by another deeply rooted social fact: gender differences within the family.

4. *The Household Perspective*

It is only recently that investigators began to underscore the importance of *household survival strategies* as crucial mechanisms that permit poor households to organize production and consumption, not simply to adapt to changes associated with the life cycle, but to surmount the repeated economic crises that have buffeted Third World economies in recent years. We know that family survival strategies include not only cooperation, but also conflict, domination, and inequality, such that men are often able to impose their preferences and decisions on women (Dwyer and Bruce 1988, Ward 1993). The household is an arena in which the tasks and obligations of members, whose interests and powers may clash, are subject to continuous contestation.

Gender inequalities are manifested in the labor force behavior of family members. The broader political economy extends into intradomestic strug gles. Rural men and women conceive of, are conceived of, and for that matter offer themselves very differently to the labor market, at least in certain core matters that not only signal differences, but reflect broader inequalities and hierarchies with respect to gender: the family–individual relation and conceptualization of male and female work.

4.1 *Family–Individual Relations*

During the 1970s, the availability of paid labor in cannaries, workshops, factories, and *maquiladoras* was, in many cases, something entirely new in the life of rural communities (Ward 1993). Given economic crisis, rural families could not afford to pass up opportunities for work within their own communities. They began to redefine priorities and reorganize to harness emergent opportunities; and they did this in accord with traditional gender hierarchies. It was clearly a complex undertaking trying to incorporate these new labor opportunities while conforming to the idea that the female condition is permanent and immutable. With the aim of mitigating the social effects that women's wages and the power they potentially bring, the family positioned itself as the crucial intermediary in recruiting and organizing the new systems of work that arrived in the countryside (Hareven 1982, Rosado 1988).

In general, families looked to assure the permanence of two basic principles: to assure new income for the family, and to safeguard 'morality' in the face of the compromises entailed in taking paid labor. One form of mitigating the potentially disruptive effects of female work and extra-familial income was to send only young, single women into the workforce, i.e., those who lived in their paternal home and therefore had much less power to negotiate within the family. The selection of young unmarried women assured that the new earnings would be incorporated directly into the family circle. Under the authority of the father, the income generated by daughters was socially defined as an obligatory family *contribution*, much the same as the fulfilling of any other domestic obligation of a daughter or sister.

The contrasting family responsibilities of men and women are also evident in what happens to the earnings. As has been shown in a variety of societies, men experience much greater latitude than women in deciding what to do with their money, even when they are unmarried (Dwyer and Bruce 1988). In the end, the conservation of family norms and values was more important to male family members than the struggle to better working conditions or wages for their wives and daughters.

4.2 *Male Work as 'Obligation'*

The conceptualization of work and what one expects to achieve through it also distinguishes men from women. For men, work is viewed as an obligation through which the man must (but also can) fulfill the role of fundamental family provider. Through work, he hopes to generate income, sufficient to cover the basic necessities of those who depend on him, although this, as we know, is an ideological construction of relative recency in history, with little relation to economic reality (Rothstein 1995).

This social construction has assigned men both the obligations and the rights of a family provider. This means that men have the freedom to seek out the best pay and working conditions, rejecting jobs that cannot support their family. This makes it socially acceptable for men to reject poorly paid jobs and to migrate far from home in search of better jobs. Through mobility, men may avoid the low salaries offered in agriculture, agribusiness, manufacturing, and the *maquiladoras* that were built throughout the rural world in the 1970s (Arias and Wilson 1997). This helps explain the association in the countryside between an abundant demand for labor and at the same time the ongoing departure of men for other regions or countries in search of work.

This paradox illustrates another characteristic of male migration: that they can move away for work while still meeting their family obligations and not disturbing (indeed, reinforcing) the social framework in which they form a part, for they know that mothers,

wives, and sisters will be charged with maintaining their filial and familial obligations (Massey et al. 1987). In most societies, as Dwyer and Bruce (1988) indicate, the time that men actually devote to their children has always been meager and is easily reduced without repercussion.

The situation of the rural woman is entirely different. For her, liberty and the freedom to disengage from domestic obligations do not exist, so that her justification in leaving to look for work, or to find better working conditions, is truly limited and is supposed to be arranged through a long and costly negotiation with other women to take charge of the domestic responsibilities that she will no longer be able to meet.

4.3 Feminine Work as Complementary 'Help'

The paid work of women is governed by very different notions; it is seen as complementary help. Women's work is, and has always been seen as a form of altruistic collaboration, that is voluntary, generous, and absolute in comparison with male labor (Folbre 1988). That a woman is hardworking and that her efforts indeed 'help' and 'complement' the family income speaks well of her in moral and marital terms, but may limit her cash contribution to the family. In this gendered process, women's paid work, as 'complementary help,' is devalued. The efforts of women are socially recognized, but they are not seen as important work and are assigned an inferior status relative to men's work (Dwyer and Bruce 1988, Safilios-Rothschild 1988).

To speak of work as 'complementary' because it is female is a way of downgrading and devaluing, *a priori*, the contribution of women to household resources. Moreover, as Dwyer and Bruce (1988) and Ward (1993) have pointed out, in many societies it is socially accepted that men can retain a part of their earnings for themselves (and often the family never knows exactly *what* proportion of male earnings are retained). In contrast, wives and daughters are expected to hand over to the family *all*resources obtained outside the household (Mencher 1988). Thus, women's earnings may be a resource families count on more than men's. Thus, seeing women's work as secondary 'complementary help' is a social construction rather than an accurate portrayal of the situation.

Seeing female labor as 'help' makes it socially manageable by the family. But, the notion of women's work as 'help' limits it several ways. First, it reaffirms that the female priority remains reproduction and domestic responsibilities; if the work of the woman is only 'help,' it is by definition insufficiently important to require renegotiation of the distribution of roles and tasks within the household. Second, the notion of female paid labor as 'help' assures that it remains of lower value, and in this way, that the woman is always available to leave it to resume her function in the family if needed. Third, viewing female labor as 'help' guarantees that in the face of family economic crises, the woman can (and *must*) enter the market to work and to accept conditions and wages that men are not expected to take, but that once the immediate crisis passes, she will return to her fundamental priority of the family (Hareven 1982). Finally, the notion of female work as 'help' is a key mechanism to avoid discrediting socially or maritally the work and earnings of men. Facing transitional labor markets that have shifted the balance of work opportunities between the sexes, the notion of 'help' contributes to the maintenance of traditional obligations and rights within the family, in obvious detriment to the position of the women themselves.

Thus, the conceptualization of female work as 'help' of a 'complementary' nature can be seen, in the end, as a porous category in which the social devaluation occurring in the labor market intersects with the social hierarchies of gender within the household. This intersection is what permits employers continued access to sources of low-cost labor, on the one hand, and the perpetuation of female work as subordinate to family obligations, on the other.

5. Recent Changes

As expected, the experience of female labor outside the household has affected gender relations within communities and families. Notwithstanding social resistance from employers as well as their families, women have begun to modify their relationships within the family and with employers. In rural societies in which the paid work of women forms an integral or intermittent part of local economic life, and where it is possible to encounter two or more generations of female workers, they have been renegotiating relations with their families that are less unequal, particularly with respect to education, mobility, and control over earnings, against the backdrop of a labor market ever more oriented to labor rights (Rosado 1988).

National and international struggles have been undertaken by nongovernmental organizations (NGOs) and women themselves during the 1990s, especially in Southeast Asia. The globalization of the economy has helped women to understand their conditions better and has extended the struggle of female workers to a geographic scale never before seen. Increasingly women have contested not only wages and work conditions that function outside the law, but those that violate basic human rights, such as the right of association, the limitation of child labor, and exposure to hazardous conditions, as well as combating different types of labor market discrimination. The struggle by female workers and NGOs has now succeeded to the point where some large international companies operating at the global level have developed norms assuring human and labor rights for workers connected to them under various subcontracting arrangements. Al-

though this is only a beginning, it indicates that rural industrialization may have entered a new phase of development.

See also: Boserup, Ester (1910–99); Development, Economics of; Development: Rural Development Strategies; East Asian Studies: Gender; Economic Development and Women; Economic Globalization and Gender; Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Southeast Asia; Gender Ideology: Cross-cultural Aspects; Islam and Gender; Labor Markets, Labor Movements, and Gender in Developing Nations; Land Rights and Gender; Latin American Studies: Gender; Lone Mothers in Nations of the South; Nation-states, Nationalism, and Gender; Near Middle East/North African Studies: Gender; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Southeast Asian Studies: Gender

Bibliography

Arias P 1988 El empleo a domicilio en el medio rural: La nueva manufactura. *Estudios Sociológicos*, **6**: 405–36

Arias P, Wilson F 1997 *La aguja y el surco. Cambio regional, consumo y relaciones de género en la industria de la ropa en México*. Universidad de Guadalajara, Guadalajara, Mexico

Bachrach Ehlers T 1993 Belts, business, and Bloomingdale's: An alternative model for Guatemalan artisan development. In: Nash J (ed.) *Crafts in the World Market*. State University of New York, Albany, NY, pp. 181–96

Boris E, Daniels C R (eds.) 1989 *Homework. Historical and Contemporary Perspectives on Paid Labor at Home*. University of Illinois Press. Urbana, IL

Bradley H 1989 *Men's Work, Women's Work*. University of Minnesota Press, Minneapolis, MN

Dwyer D, Bruce J (eds.) 1988 *A Home Divided. Women and Income in the Third World*. Stanford University Press, Stanford, CA pp. 1–19

England P (ed.) 1993 *Theory on Gender. Feminism on Theory*. Aldine de Gruyter, New York, pp. 3–22

Fernández-Kelly M, García A 1989 Hispanic women and homework. Women in the informal economy of Miami and Los Angeles. In: Boris B, Daniels C R (eds.) *Homework. Historical and Contemporary Perspectives on Paid Labor at Home*. University of Illinois Press, Urbana, IL, pp. 165–79

Folbre N 1988 The black four of hearts: toward a new paradigm of household economics. In: Dwyer D, Bruce J (eds.) *A Home Divided. Women and Income in the Third World*. Stanford University Press, Stanford, CA, pp. 248–62

Hareven T K 1982 *Family Time and Industrial Time*. Cambridge University Press, Cambridge, MA

Maruani M 1991 La construcción social de las diferencias de sexo en el mercado de trabajo. *Revista de Economía y Sociología del Trabajo*. Septiembre–diciembre 1991, Madrid, pp. 129–37

Massey D S, Alarcón R, Durandt J, González H 1987 *Return to Aztlan*. University of California Press, Berkeley, CA

Mencher J P 1988 Women's work and poverty: Women's contribution to household maintenance in south India. In: Dwyer D, Bruce J (eds.) *A Home Divided. Women and Income in the Third World*. Stanford University Press, Stanford, CA, pp. 99–119

Rosado G 1988 Las mujeres de San Pablo. trabajo y vida cotidiana. In: Aranda J (ed.) *Las mujeres en el campo*. Universidad Autónoma Benito, Oaxaca de Juarez, Mexico, pp. 147–62

Rothstein F A 1995 Gender and multiple income strategies in rural Mexico. In: Bose C E, Acosta-Belén E (eds.) *Women in the Latin American Development Process*. Temple University Press, Philadelphia, PA, pp. 167–93

Safilios-Rothschild C 1988 The impact of agrarian reform on men's and women's Incomes in rural Honduras. In: Dwyer D, Bruce J (eds.) *A Home Divided. Women and Income in the Third World*. Stanford University Press, Stanford, CA, pp. 216–28

Ward K B 1993 Reconceptualizing world system theory to include women. In: England P (ed.) *Theory on Gender. Feminism on Theory*. A de Gruyter, New York, pp. 43–68

Wilson F 1991 *Sweaters. Gender, Class and Workshops-based Industry in Mexico*. Macmillan, London, UK

P. Arias

Rural Planning: General

The heading 'rural planning' as a sector of government involvement is often absent from plan documents in many countries. But policy and legislation related to agricultural, transportation and communications, education, health and human services, and social welfare, sectors generally addressed in plans, have had far reaching and profound impacts on the physical conditions of, and infrastructure availability in, rural areas. They have also determined the range and quality of services accessible to rural populations. When headings such as rural planning, rural development, integrated rural development are present in plan documents, as discrete sectors requiring attention and having related funding lines in national and state budgets, these sectors are often defined in terms of engagement with social welfare agendas. The goals of such sectoral policies have often included statements about the need to redress inequalities in access to public services and redistributive justice. A broader definition of rural planning as encompassing policy that has both explicit and implicit intention to impact rural conditions, economic, social, and physical is adopted here.

During the formative period of the United States in the nineteenth century, agriculture production was a central element in its' robust growth and development. Similarly, an agriculture-based economy and a predominantly agrarian population was the legacy of heavily populated, newly independent, countries like China and India, which faced the task of nation building in the second half of the twentieth century.

But over the years these countries have differed significantly in their approach to rural planning. The interventions of their governments in the rural sector define a spectrum of strategies and approaches, from total centralized control to indirect, market-driven change. Furthermore, their strategies, which took divergent paths at the start of nation formation, have over time sharply changed direction.

The United States (US), India, and China provide examples of countries that have developed rural planning-related policies based on quite differing ideologies about planning and the role of rural areas in national development. Their positions have ranged from the free market, laissez faire orientation typical of 'First World' countries like the US to policies of 'Second World' countries like China that adopted a centralized, cooperatives based model of rural planning of the Soviet Union. Seeking a more intermediate path between these two extremes were countries like India, a founding member of the block of so-called 'Third World' countries that sought to chart a distinct path to development. Examining the development of the rural planning concept in countries like the US and India provides a glimpse of the differing interpretations of rural planning that have been evolving around the world in countries where the government's ability to control, plan, and implement has been limited by, and has been cognizant of, the significant role of the private sector in the farming endeavor. Study of the China model provides insight on the strengths and limitations of sustained, centralized efforts to control developments in rural economy and society.

1. *The United States Model of Rural Planning*

The US is acknowledged to be one of the most open and laissez faire economies in the world. Within the US, the idea that attention needs to be turned specifically to rural planning is not a widely embraced concept. But models of economic growth that guide policy are accepted and in these the agricultural sector is factored as an integral component of the national economy. Historically, US agricultural policy has defined and shaped rural conditions and had far-reaching consequences on the development of rural areas, but it has not been recognized as an instrument for rural planning.

This American ambivalence towards acknowledging the existence of rural planning is understandable given historical antecedents. After American independence, the feudalistic land law that prevailed under British rule was replaced with a freehold land tenure system under which public lands were sold to raise cash for the new government. To facilitate sales the US Congress ordered a land survey in 1785 that established a grid system of land division as the dominant pattern for the area west of Pittsburgh. Its indelible stamp ordered the transportation and communication systems and helped give shape to rural administrative areas. Subsequent land laws and purchase allowed 'a mass distribution of land to those who tilled the soil [that] was unique in the history of the world up to that time' Cocharane (1979, p. 57). The philosophy of land ownership as vesting freehold in an individual, and the establishment of a unique American tradition of independent and self reliant farm families scattered in a regular pattern, each in isolation, on their own individual pieces of land, has perhaps contributed to a national antipathy to identifying government interventions as responsible for rural development. Nonetheless these early interventions shaped the rural context, with plans that were not always openly agreed to or publicly articulated, and embedded it with desired outcomes that might have been contested if formally acknowledged in planning documents.

The movement of the US agricultural enterprise from the subsistence cultivation of the pioneer to the high tech, mechanized, computerized, agro industry of today's corporate farm is not the product of free-market forces but of significant government actions. The role of the government in shaping US farming is noted and generally accepted in the literature pertaining to American agriculture (Rasmussen 1960, Cocharane 1979, McConnell 1953). Cocharane (1979, p. 277) delineates a broad and inclusive set of categories in which government has been active, asserting that it has been involved in almost every aspect of agricultural development. With the passage of the Agricultural Adjustment Acts of 1933 and 1938 the Department of Agriculture has supported and managed agricultural production pursuing policies affecting both the supply and distribution systems of agricultural production (Cocharane 1979, pp. 141–2).

1.1 *Modern Farming in the US*

Government activities in the transport, research, and education sectors have transformed American agriculture from subsistence to cash cropping for regional and national markets and later to capital-intensive agribusiness. Significant government support of educational and research institutions assisting in agricultural development through scientific and technological innovations has been sustained for more than a century. The Morrill Act of 1862 enabled government to fund land grant colleges in each state and the Hatch Act of 1887 allowed annual allocations for state agricultural experimentation stations devoted to basic agriculture research (Rasmussen 1960). The impact of this activity on shaping the scientific base of modern American farming and increasing productivity and efficiency was significant and to date has had resonance and wielded influence around the world.

In response to the worldwide Depression of 1929 New Deal agricultural economists created the Agricultural Adjustment Acts of 1933 and 1938, to set the price and allowable level of farm production. The central theme in these legislative efforts was to stabilize agricultural prices and income by offering subsidy payments to farmers to reduce production, price support measures, and commodity storage. In the postwar period from 1945 to 1975 the modern farm evolved in the context of rapid advances in national/international transport the emergence of the United States as a major world power, and the corresponding change in its domestic and foreign policy, and tremendous advances in scientific/technological farming in the United States including development of high-yield hybrids, fertilizers, pest controls, and extensive mechanization. The modern farm featured large-scale production for national and international markets. Cultivation for sale rather than family subsistence became the central production objective of the farm enterprise. However the ideological underpinnings of the earlier time of minimum government control in the farm enterprise persisted. There was support for government to be an agent responsible for presenting new options and alternatives which might voluntarily be subscribed to, or not, by the American farmer.

1.2 Corporate Farming in the US

A shift from modern to corporate farming after 1975 was triggered by changes in the role the United States played in world affairs and the tack it took with respect to its foreign policy following World War II. The United States emerged as a leading proponent of global free trade. This posture meant that the highly efficient farms in the US began to function as the 'world's breadbasket.' Farmers produced and exported crops to India, Russia, and other countries when these countries faced domestic shortfalls in food production. Advances in national and international transport aided this expansion of the market. 'Green revolution' technology packages of high-yielding hybrids, super fertilizers, and pest controls were developed leading to unexpectedly large crop yields. Specialization and mono-crop farms become common. Although energy intensive in their use of new products and machinery, they were viable as energy costs were kept relatively low by the government. A process of consolidation of family and modern farms into larger and larger vertically and horizontally integrated corporate entities occurred. The resulting corporate super farms have changed the rural landscape tremendously as fewer and fewer farms account for more and more of the of total farm sales and total farm profits in the United States. This concentration is apparent in the fact that in 1997 two percent of all farms in the US produced 50 percent of total farm sales.

1.3 Services-oriented Rural Planning

In the 1980s and 1990s it has been increasingly obvious in the United States that although the link between agriculture policy and rural planning has traditionally been very strong it may be less important in the emerging 'global' economy of the twenty-first century. Today the link that needs to be forged for rural planning is with economic sectors such as manufacturing and services. The emerging 'new economy' is based on the growth of the service sector and geared to the information industry that has centralized population in metropolitan areas. The impact of this centralization on populations that remain rural and dispersed is significant. For example in rural areas public services in education, health, and social welfare provision are of lower quality than in metropolitan areas. Currently US rural planning efforts have focused on providing quality services despite the delivery difficulties posed by lower population densities that mean that greater distances must be covered to reach targeted rural populations. A second strand of rural planning pertains to national settlement patterns and the belief that it is important to maintain numbers of people in rural areas. Planning efforts creating rural enterprise and empowerment zones have involved job-training programs as an economic development policy to attract entrepreneurs. The second element of this strategy aims to reduce costs for entrepreneurship in rural areas by programs that make capital available for business development. The scale of such efforts is however quite modest. Finally, the current paradigm of rural planning focuses on equalizing rural communication and infrastructure with urban. Supporting the expansion of Internet and high-speed communications links, wireless, and satellite connections all aimed at linking dispersed populations are an integral part of this agenda.

2. The 'Third World' Approach to Rural Planning

Numerous countries attained independence from European colonial rule at the end of World War II. Some, like the Peoples' Republic of China, became 'Second World' countries that adopted planning models emulating Soviet style control by a strong central government. Another block of countries attempted to forge a separate third path to development differing from United States style capitalism and Soviet style socialism. This 'Third World' forged constitutions endorsing centralized planning by government to attain modernization by facilitating rapid industrialization of both the private and public sectors. In these models rural planning tended to be a key element in national economic development. Rural planning measures aimed to increase production but also to redistribute wealth on the grounds of social justice and human welfare. The latter provided legitimacy for governmental planning which was essential

in countries recently released from colonization. Furthermore the majority of these nations' populations were rural. In approaching the development of rural areas many of these countries emulated the US model. They aimed to increase and improve agricultural production through the application of scientific research that was to be developed, disseminated, and implemented in farming practice through agricultural extension services. This was seen as the modern way for rural planning to support national economic development.

Third World countries aimed to use agricultural production to raise capital for industrialization and the development of related infrastructure, as well as for the provision of urban as well as rural based populations. However they tempered their desire to increase agriculture production with the broader objective of bringing about an increased living standard and quality of life for economically disadvantaged populations in the countryside. Their emphasis on rural planning for human development reflected their recognition that the majority of their populations were rural, asset poor, involved in subsistence production, and members of lower status social groups of social structures that were feudal in character. Over time, these countries evolved a planning model for 'integrated rural development' that recognized gender and class differences and attempted to address social inequalities in access to rural resources; a process differentiated from earlier production oriented agricultural development models. This more recent approach to rural planning attempted to combine and balance between agricultural and economic development and social welfare and redistributive justice.

3. *The Indian Model*

At the point of independence from British rule India embraced rural planning directly. Initial Indian national-planning efforts to achieve rapid socioeconomic change concentrated on industrialization at chosen metropolitan centers, on the assumption that the benefits would 'trickle down' through the economy. Planners believed that although initially benefits would be concentrated in the core cities they would spread to rural areas. However, they also recognized the need to promote complementary rural planning to reduce existing and anticipated urban-rural disparities in wealth.

Rural planning measures in India conspicuously emulated the US model of agricultural development. They included: irrigation projects; the promotion of scientifically developed cash crops through the provision of improved seeds, equipment and fertilizer (the so-called Green Revolution approach to making agriculture productive); providing credit for agricultural investments to stimulate agricultural production; and, the establishment of a system of extension services to introduce scientific farming techniques at the village level. They also regulated markets to stimulate trade in agricultural commodities; invested in roads and communications; and founded cooperatives to buy and sell agricultural products and facilitate the marketing of goods in rural areas.

3.1 *Redistribution and Equity*

The national policy of investing in industry and commercial agriculture increased the gap between those involved in the corporate formal sectors of the economy, urban and rural, and those in the informal, traditional, subsistence sectors. Rural landless and land poor, squeezed out of traditional positions in village society, migrated to the city in unprecedented numbers. Cities were unable to accommodate the massive influx of displaced, often unskilled, rural laborers and their families and urban environments deteriorated as slum settlements of the poor proliferated. Responding to this phenomenon inspired a new set of rural planning policies. Stemming migration to the city by improving rural life and diversifying the rural economy emerged as prime planning objectives. Making rural living more attractive included provision of postal service, healthcare facilities, clean drinking water, and electricity in the countryside. As this policy evolved, its objectives also expanded to include increasing equity and distributing the benefits of development to the poor more explicitly. This change inspired such planning measures as putting a ceiling on individual land holdings, reforming tenancy to enhance the rights to the land of those who cultivated it, initiating programs to support traditional artisans with better tools, establishing credit and marketing networks to allow them to compete with industry or at least sustain them until alternative jobs were created. Finally, the Indian government emphasized free, mandatory elementary school education throughout rural areas to improve skills the poor needed to compete for development benefits.

In 1976 A. T. Mosher an international agricultural development expert summarized the various phases of Indian rural development planning in *Thinking about Rural Development*. He described some common characteristics. They were:

(a) Recognition that although economic development is important it is not an essential precondition to the social transformation that is considered the central objective of a rural development strategy;

(b) Attention to increasing efficiency with the technical revolution of fertilizers and high yield varieties of cereals. The attempt was to boost both food grain production and commercial crops and monoculture production of commodities for which there was an international market;

(c) Development of programs that addressed equity and asset redistribution; and

(d) Enhancing rural citizen's participation in the evolution of locally appropriate planning programs. Elements of these phases of rural planning policy and objectives in India—moral and social transformation, economic efficiency, redistribution and social justice, and popular participation in decision making—are to be found in the rural planning approaches of many countries around the world although there is a great deal of variation both in emphasis in policy and in the capacity to implement.

4. *Current Priorities and Approaches in Rural Planning*

Rural planning oriented to improving physical infrastructure and the quality of life for rural populations was an important policy agenda in developing countries in the earlier periods of their post-independence histories. More recently, as rates of urbanization have remained high in these countries and cities have experienced hyper growth, despite efforts to stem rural migration and population increases, rural planning in many countries has become a lower priority endeavor. The worldwide liberalization of economies attention is focused on World cities of over 10 million people, and, as a result, notions of intervention in rural areas to improve rural conditions have been given far less significance.

Concerns about the long-term negative impacts of the 'Green Revolution' have emerged in many parts of the World. They are related to degradation of soil quality; depletion of groundwater tables; the long-term sustainability of modern agricultural cultivation on prime land; and, as supply of such land is exhausted, about the viability of applications of this technology package on poorer quality, more fragile lands; and rising concern about genetically modified crops entering the food system of human populations. These concerns have led to calls for a modified, mixed approach to rural development that recognizes traditional, sustainable practices. In contexts of risk prone environments, development planners are describing a need for a rural planning strategy in which 'top down' science driven Green Revolution strategies of intensifying agricultural cultivation are tempered with the understanding that can be attained by tapping the knowledge of rural cultivators who are following traditional, self sustaining forms of agricultural production (Pichón et al. 1999). This position, labeled 'sustainable agriculture', has stimulated rural planning efforts based on more traditional modes of agriculture (National Research Council 1989). The task of helping rural places has been defined as a set of two complementary concerns related to how people will earn a living (Lapping et al. 1989) and how they will do it so that natural, environmental resources will be conserved (Sargent et al. 1991).

In the 'First World' rural planning is coming full circle. From a 'universalistic' modern-science driven notion of development and progress there is a renewed orientation to the specificities of the local, rural, asset base of both physical and human capital. In much of the developing world, facing the waves of change engulfing countries in the throes of economic liberalization, rural planning has been placed much lower in priority to programs that address issues of rapid urbanization and globalization. Thus programs related to social justice, participatory decision making, sustainability, and the environment have taken second place to economic transformation of rural production.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Development: Sustainable Agriculture; Environment and Development; Environmental and Resource Management; Environmentalism: Preservation and Conservation; Farmland Preservation; Green Revolution; Greening of Technology and Ecotechnology

Bibliography

Cocharane W W 1979 *The Development of American Agriculture: A Historical Analysis*. University of Minnesota Press, Minneapolis, MN

Lapping M B, Daniels T L, Keller J W 1989 *Rural Planning and Development in the United States*. Guilford Press, New York

McConnell G 1953 *The Decline of Agrarian Democracy*. Atheneum/University of California Press, New York

Mosher A T 1976 *Thinking About Rural Development*. Agricultural Development Council, New York

National Research Council 1989 *Alternative Agriculture*. National Academy Press, Washington, DC

Pichón J F, Uquillas J E, Frechione J 1999 *Traditional and Modern Natural Resource Management in Latin America*. University of Pittsburgh Press, Pittsburgh, PA

Rasmussen W D 1960 *Readings in the History of American Agriculture*. University of Illinois Press, Urbana, IL

Sargent F O, Lusk P, Rivera J A, Varela M 1991 *Rural Environmental Planning for Sustainable Communities*. Island Press, Washington, DC

H. C. Dandekar

Rural Sociology

Rural sociology has traditionally been defined as the study of social organization and social processes that are characteristic of geographical zones where population sizes and densities are relatively low. Rural societies, however, do not exist in isolation or a social vacuum, and thus the subject matters of rural sociology increasingly reflect the larger processes of the regional differentiation and allocation of populations and economic activities, and of social relations within

a society as a whole (and, increasingly, within global economy and society).

Perhaps the single most crucial fact about this field is that, as a formal sociological specialty, rural sociology is largely US in its origins, and that rural sociology emerged within the land-grant agricultural college system and in close association with major US historical trends in agrarian politics. In this article I will begin with the history of US rural sociology, particularly in relation to major contemporary emphases within the field. In the second section I will focus on the major branches of contemporary rural sociology, and especially on the significance of the 'new rural sociology' during the last two decades of the twentieth century.

1. *The Origins of Rural Sociology*

In a sense rural sociology, much like the discipline at large, has had very strong roots in nineteenth century social thought. Many of the major issues of concern to the classical theorists were in some sense germane to rural sociology, and have remained important to the subdiscipline a century later. Among the most important of the preoccupations of nineteenth century social theory was whether village and farm life was more socially cohesive than, or morally or socially superior to, metropolitan life, and whether rural social relations would be resilient in the face of capitalist industrialization. Emile Durkheim (1858–1917) and Karl Marx (1818–83) both argued that urban-industrial capitalism, the advancing division of labor, and modern technologies and organizational practices were inexorable—and, on balance, progressive—social forces that would eventually supersede or supplant rural social forms. Ferdinand Tönnies (1855–1936) concurred that urbanization and industrial capitalism were tending to lead to the decline of intimate village communal bonds, though Tönnies was more optimistic than Durkheim and Marx that rural social relations could be sustained. For most of its history, rural sociology has reflected these nineteenth century debates over the desirability and resilience of traditional rural social organization.

It is now generally agreed that the pioneering rural sociological studies in the USA were those by W. E. B. DuBois in the late 1890s on the social consequences of the postbellum crop-lien system in Southern plantation agriculture. But while DuBois was the pioneering American rural sociologist and there were several prominent nonland grant rural sociologists (e.g., F. H. Giddings of Columbia University) in the field's early years, the character of rural sociology in the twentieth century was largely shaped by its articulation with the land grant college system. Rural sociology as a recognized subdiscipline largely had its origins in the preoccupation of land grant college leaders with the challenge of Populism. Founded and funded by industrial and other elites after the turn of the century, the Country Life movement promoted the view that the problems of rural society were not due to the shortcomings of industrial capitalism as the Populists had held. These problems were rather seen to be due to a lack of organization, poor infrastructure, and technological backwardness in rural areas, which in turn led rural society to lag behind urban USA.

In 1908 President Theodore Roosevelt recognized this reform movement by appointing a Presidential Commission on Country Life. The next six decades of US rural sociology were largely prefigured by the County Life Commission Report, which was published in 1909. The Report did acknowledge many of the social problems of rural USA (e.g., the inequities of the crop-lien system, widespread tenancy, land speculation). But the dominant thrust of the Commission Report was the case it made that an expanded effort to modernize rural USA technologically and structurally was integral to improving the condition of rural USA. The Commission recommended the harnessing of the social sciences, particularly agricultural economics and rural sociology, to support the technological modernizationist vision for rural America. The land grant colleges responded to the Country Life Commission Report by hiring rural sociologists and establishing rural sociology units or departments. Rural sociological studies in land grant colleges of agriculture were seen to be important in order to help remove social barriers to technological modernization and to stabilize rural communities.

Sorokin and Zimmerman's (1929) influential treatise *Principles of Rural–Urban Sociology* exemplified the first three decades of US rural sociology. *Principles* synthesized the European classical tradition (particularly that of Tönnies) and land grant technological progressivism with the emerging empirical tradition of research on rural people and communities. The Sorokin–Zimmerman perspective was essentially a theory of rural–urban continuum, in which the character of social relations was posited to be shaped by the size, density, and related parameters of settlement patterns.

Very soon after *Principles* was published, rural sociology would be confronted with the Great Depression and rural squalor, and most rural sociologists embraced New Deal reformism. The Great Depression and the fashioning of the Depression-era New Deal opened up vast opportunities for rural sociological scholarship on rural communities and peoples which was closely linked with the New Deal reform agenda (e.g., land tenure reform, encouragement of cooperatives, commodity programs) (see the summaries in Buttel et al. 1990, Chap. 1, Larson et al. 1992).

Reform-minded rural sociology, however, was short lived as a result of the ideological shift to the right that had largely swept away most New Deal rural programs by the mid-1940s (Larson et al. 2000). The jettisoning of the New Deal rural programs induced the land

grant colleges to refocus their efforts on the technological upgrading of the countryside. In the late 1940s and early 1950s a handful of prescient rural sociologists developed the diffusion of innovations perspective on agricultural innovation and change which provided sociological support to the land grant technological project. The diffusion of innovations perspective was a social-psychological perspective on farmer technology adoption behavior that posited that innovative and progressive technologies and value orientations tend to diffuse from metropolitan institutions to rural people and localities, and that persons with more modern orientations are most likely to adopt improved metropolitan-developed technologies and practices (see Rogers 1995 for a contemporary summary, and Fliegel 1993 for a historical perspective and sympathetic critique). Shortly thereafter the social psychology of farm practice adoption, and of related rural behaviors such as educational aspirations, provided a comfortable niche for rural sociology within the land grant college system.

Rural sociology now exists in dozens of world nations. Much of the impetus for the global spread of rural sociology was due to the missionary zeal of US rural sociologists who from the 1940s through the 1960s sowed the seeds of US-style rural sociology across an impressive range of countries across the globe. Since the 1970s, however, rural sociology has become increasingly diverse and less influential globally. US rural sociology increasingly borrows from the ideas of European and Third World scholars (e.g., Goodman and Redclift 1991, Llambi 1990, van der Ploeg 1992), and is sometimes criticized for its parochialism and quantitative emphasis.

2. *The Major Focuses of Contemporary Rural Sociology*

Rural sociology has increasingly become a less coherent subdiscipline, and more of an 'umbrella' for an assemblage of research, much of which no longer pertains primarily to rural people or rural social relations. Rural sociology now has six major branches: rural population, rural community, natural resources and environment, agriculture, sociology of international rural development, and sociology of agricultural science and technology.

From the very inception of rural sociology, sociological analysis of rural population and rural community dynamics through census and social survey data has been central to the field. In the early days of rural sociology, rural population and rural community studies were very closely articulated. In large part this was because census of population data for counties and rural places were the most accessible data sources for characterizing rural communities and farm structures. Even by the end of the twentieth century there remains a significant relationship between rural population and community studies (see, e.g., Fuguitt et al. 1989, Garkovich 1989).

While there has been a continuing basis for articulation between rural population and community studies, these areas of work have become increasingly differentiated so that they are now seen as quite distinct specialties. The use of census and survey data to document rural–urban differences, as in Sorokin and Zimmerman's (1929) *Principles*, fell into disrepute, and the rural–urban continuum perspective was slowly but surely undermined empirically (Newby 1980). Rural population research has followed the larger subdiscipline of demography in becoming more quantitative, while the rural sociological analysis of communities, though it has a strong quantitative wing, has continued to have a significant cadre of researchers who use qualitative or field methodologies (Wilkinson 1991). Studies of regional and labor market inequalities (Lyson and Falk 1993, Lobao 1990), rural poverty dynamics and racial inequalities (Duncan 1999), and rural gender inequalities (Sachs 1996) have complemented the traditional rural sociological emphasis on community studies.

While there has been a long and significant tradition of rural sociological scholarship on the relations among people, communities, and natural resources (Field and Burch 1988), the sociology of natural resources and environment in the contemporary sense did not emerge until the early 1970s. Rural sociology pioneered in development of the larger subdiscipline of environmental sociology and natural resource sociology. Rural sociologists have figured prominently in both environmental sociology (Redclift and Woodgate 1997) and the more management-oriented field of the sociology of natural resources (Field and Burch 1988).

As noted earlier, many of the most prominent US rural sociologists of the post-World War II period devoted major segments of their careers to encouraging the diffusion of rural sociology across the globe, particularly in the developing world. Part of the early impulse for elaborating an international rural sociology was the intellectual one of wanting to promote a comparative approach to understanding rural social organization. But equally important was the fact that the postwar period was an emerging era of 'developmentalism' (i.e., an era of faith in the efficacy of planned social change and development in the decolonizing world, see McMichael 2000). Rural sociologists trained in the adoption-diffusion tradition were particularly well represented among the first cohort of international rural sociologists (see Fliegel 1993, Rogers 1995). Scholars adhering to the diffusion of innovations perspective found their work very compatible with modernization theory, the dominant theory at the time in the sociology of development. This modernizationist/developmentalist tradition of rural sociology predominated for nearly three decades, but ultimately it would be undermined as a result of

critical assessments of the Green Revolution experience and of the role of rural sociology in international assistance (e.g., Koppel 1995).

Of all the major rural sociological specializations, the sociology of agricultural science and technology is, in one sense, one of its oldest, and in another sense, rural sociology's newest major speciality area. There were significant studies of technological change and labor displacement in Southern plantation agriculture during the 1930s through the 1950s. And, as noted earlier, the diffusion of agricultural innovations predominated in US rural sociology from the late 1940s through the late 1960s (see the summaries in Buttel et al. 1990 and Fliegel 1993).

Diffusionism, however, declined for three main reasons, each of which corresponded with an important new dimension in sociological research on agricultural science and technology. The first reason, as suggested earlier, was the growing criticism of the ability of the technological modernization project to generate equitable social benefits. This criticism was originally focused on the shortcomings of the Green Revolution, but was soon extended to the publicly funded agricultural-technological project in the advanced countries as well (see Newby 1980). Second, beginning in the late 1970s the Busch–Lacy group drew on world systems theory and the sociology of science to develop an influential critical perspective on the social causes and consequences of technological change in agriculture (Busch and Lacy 1983). Busch and Lacy's approach to the sociology of agricultural science demonstrated that social factors shape the content of new technologies, that new technologies tend to have social costs as well as benefits, and that rural sociologists need to take a more agnostic position toward new technologies than was generally the case among diffusionists. Third, in the 1970s and 1980s US rural sociology was increasingly influenced by the rise of Marxism in sociology at large. Neo-Marxism played a particularly prominent role in the sociology of agricultural science and technology (Kloppenburg 1988, Goodman and Redclift 1991). The neo-Marxist literature on agricultural science focused sociological attention on the class character of agricultural research institutions and agricultural technologies, and in so doing was often critical of diffusionism for having obscured how agricultural research is implicated in class and power relations.

It was noted at the outset of this article that rural sociology's history has been one of alternating between opposing views about the distinctiveness and resilience of traditional rural social structures. Rural sociology came into its own during the post-World War II era when it rejected rural romanticism and transcended New Deal reformism by embracing the technological-modernizationist assumptions. In the 1970s and early 1980s, however, the balance between these two views would be equalized to a degree. As noted, emerging controversies over agricultural technologies and their social and environmental impacts removed some of the luster from the technological-modernizationist perspective in rural sociology. In addition, the 1970s were a period of 'rural renaissance' in the USA, as the population of nonmetropolitan counties grew faster than that of metropolitan counties due to net metro-to-nonmetro migration (Fuguitt et al. 1989).

These new realities of rural USA during the 1970s, along with the growing popularity within sociology of various 'critical' theories, led to what is often referred to as the 'new rural sociology.' These new theories were actually quite diverse, drawing heavily from neo-Marxism but also from neo-Weberianism, the peasant studies literature, and critical theories drawn from hermeneutics and related traditions (e.g., Mann 1990, Friedmann 1978, Friedland et al. 1981). To the degree there was a coherence within the new rural sociology, it was its critique of modernizationism and diffusionism within rural sociology, and its insistence on spotlighting the role of class and power relations. The 'new rural sociology' was originally most prominent in the sociological literature on agricultural change (see the summary in Buttel et al. 1990), but most recently it exhibited a major shift toward research on cross-border agricultural commodity chains and on the globalization of agrifood systems more generally (Bonanno et al. 1994, McMichael 2000).

Perhaps the clearest indication of the importance of this new scholarly tradition was that even scholars who have been ambivalent about the new rural sociology have been obliged to address its premises and hypotheses (van der Ploeg 1992). The new rural sociology has also directly or indirectly influenced most other areas of rural sociology (see, e.g., Lyson and Falk 1993, Lobao 1990). But despite the fact that the new rural sociology is the most pervasive new trend during the last quarter century of rural sociology, rural sociology remains highly diversified. The new rural sociology, even in its heyday of the 1980s through the mid-1990s, did not come close to dominating the field in the way the diffusion paradigm did in the 1950s.

See also: Agricultural Change Theory; Development: Rural Development Strategies; Development: Sustainable Agriculture; Diffusion, Sociology of; Human Ecology; Innovation, Theory of; Social Change: Types

Bibliography

Bonanno A, Busch L, Friedland W, Gouveia L, Mingione E (eds.) 1994 *From Columbus to ConAgra*. University Press of Kansas, Lawrence, KS

Busch L, Lacy W B 1983 *Science, Agriculture, and the Politics of Research*. Westview Press, Boulder, CO

Buttel F H, Larson O F, Gillespie G W Jr 1990 *The Sociology of Agriculture*. Greenwood Press, Westport, CT

Duncan C M 1999 *Worlds Apart*. Yale University Press, New Haven, CT
Field D R, Burch W R 1988 *Rural Sociology and the Environment*. Greenwood Press, Westport, CT
Fliegel F C 1993 *Diffusion Research in Rural Sociology*. Greenwood Press, Westport, CT
Friedland W H, Barton A H, Thomas R J 1981 *Manufacturing Green Gold*. Cambridge University Press, New York
Friedmann H 1978 World market, state, and family farm: Social bases of household production in an era of wage labor. *Comparative Studies in Society and History* **20**: 545–86
Fuguitt G V, Brown D L, Beale C L 1989 *Rural and Small Town America*. Russell Sage Foundation, New York
Garkovich L 1989 *Population and Community in Rural America*. Praeger, New York
Goodman D, Redclift M 1991 *Refashioning Nature*. Routledge, New York
Kloppenburg J Jr 1988 *First the Seed*. Cambridge University Press, New York
Koppel B M (ed.) 1995 *Induced Innovation Theory and International Agricultural Development*. Johns Hopkins University Press, Baltimore, MD
Larson O F, Zimmerman J N, Moe E O 1992 *Sociology in Government*. Westview Press, Boulder, CO
Llambi L 1990 Transitions to and within capitalism: Agrarian transitions in Latin America. *Sociologia Ruralis* **30**: 174–96
Lobao L M 1990 *Locality and Inequality: Farm and Industry Structure and Socioeconomic Conditions*. State University of New York Press, Albany, NY
Luloff A E, Swanson L E (eds.) 1990 *American Rural Communities*. Westview Press, Boulder, CO
Lyson T A, Falk W W (eds.) 1993 *Forgotten Places*. University Press of Kansas, Lawrence, KS
Mann S A 1990 *Agrarian Capitalism in Theory and Practice*. University of North Carolina Press, Chapel Hill, NC
McMichael P 2000 *Development and Social Change*, 2nd edn. Pine Forge Press, Thousand Oaks, CA
Newby H 1980 Trend report—rural sociology. *Current Sociology* **28**: 1–141
Redclift M, Woodgate G 1997 *International Handbook of Environmental Sociology*. Edward Elgar, Cheltenham, UK
Rogers E M 1995 *Diffusion of Innovations*, 4th edn. Free Press, New York
Sachs C 1996 *Gendered Fields*. Westview Press, Boulder, CO
Sorokin P A, Zimmerman C C 1929 *Principles of Rural–Urban Sociology*. Henry Holt, New York
van der Ploeg J D 1992 The reconstitution of locality: Technology and labour in modern agriculture. In: Marsden T, Lowe P, Whatmore S (eds.) *Labour and Locality*. David Fulton, London, pp. 19–43
Wilkinson K 1991 *The Community in Rural Society*. Greenwood Press, Westport, CT

F. H. Buttel

Russian Revolution, The

The term is used in different ways. It refers to varying events and time frames. Normally it means the major upheaval of 1917 that replaced the tsarist empire with Bolshevik rule. In this case the term can refer to either the demise of the monarchy in the so-called *February Revolution* (February 23/March 8 N. S. to March 2/15) or to the *overthrow of October 25–26* (November 7–8) when the Bolsheviks seized power in the Petrograd Soviet of Workers' and Soldiers' Deputies. But 'Russian Revolution' is often extended to denote longer processes and obviously related sequences of events of the same type. In this case, the term encompasses the 1918–20 Civil War immediately following the October Revolution as well as the subsequent early Soviet period up to the *Stalin Revolution* of 1929–30 or—the widest definition—the entire transformation of Russia from the turn of the century and so-called *First Revolution* of 1905–6 all the way down to the Stalin Era. Looking at it this way, the 'Russian Revolution' covers Russia's peculiar political and socioeconomic modernization crisis which ended in the establishment of the Soviet Communist developmental dictatorship (1900–1930).

Substantive reasons and the question of research feasibility suggest narrowing this range of definitions somewhat. 'Russian Revolution' should be restricted to the convulsive transition from the Russian *ancien régime* to the Bolshevik form of government. The term properly denotes the two core events of the transfer of power: the overthrows of February and October and their consequences. However, there are valid reasons for including the Civil War because it was by nature the inevitable, belatedly-waged war of revolution. Both time frames (1917 and 1917–21) lend themselves to differentiation of processes into *politics and power struggle*, *socioeconomic* change, and *cultural* and *nationalist* developments.

1. The Political Constitution and System of Government

Political change in 1917 consisted of the succession from the monarchy to the 'emergency democracy' (Geyer), the Provisional Government (March 2–15 to October 26–November 8), to the soviet-based rule of the Bolsheviks with the 'Left Socialist Revolutionary Party' as a junior partner. Neither the direction of change nor the options in the year of revolution can be understood without a look at preceding developments. By the same token, the consolidation and transformation of the outcome cannot be explained without reference to the Civil War. The impetus for criticism of the autocracy definitely came from circles just outside the seat of power. 'Liberal' forces in the aristocracy, the industrial-commercial upper class, which had developed a bourgeois self-image, and the recently established academic intelligentsia rallied to the cause of the constitutional restraint of the absolute monarch. But already during the grave State crisis of 1905–6 widely differing views were held on the extent of the say that parliament (the Duma) should get in govern-

ment. Following the restabilization of tsarist rule only moderate demands were still voiced by the groups that were allowed to operate legally. Nonetheless the 'revolution,' a term which cannot withstand any close scrutiny, clearly engendered a new political system and apparently a new distribution of power, although opinion is divided on the latter claim. Its legacy included a parliament, a party spectrum (albeit an incomplete one) and a politically outspoken press. Moreover, regional self-government (*zemstva*) appeared alongside the central government. Provincial authorities gained more and more powers. World War I accelerated this shift. Ultimately the *zemstva* guaranteed the last remnants of public administration.

Only this background explains why a protest by workers' wives against hunger and cold swelled to a mass uprising, which toppled the monarchy. The last week of the 300-year Romanov rule began with an International Women's Day demonstration, which differed from numerous past rallies only in the number of participants. A repeat performance represented a new quality. In the next few days the movement escalated into a general strike of workers in the capital, and local military units joined the protest. The revolution, in the strict legal sense as a new edition of sovereignty, took place on February 28 when a Duma committee without authorization of the tsar appointed itself as the Provisional Committee for the Restoration of Public Order, thereby seizing power. At the same time the Workers' Soviet, which had convened the day before on the lines of the autumn 1905 council, elected an analogous body. This established the *two principal political institutions* which constituted the so-called *February Regime* ('dual power') of the following year. They represented two different social classes, requiring mediation of their interests by the Provisional Government, first formed by the Liberals (embodied by the *Constitutional Democratic* Party), after May 5–8 together with the moderate wings of the *Socialist Revolutionary* and *Menshevik Social Democratic* parties. The remarkable correspondence between social status and party orientation including institutional representation is indicative first of the marked polarization in the society and second of the absolutely crucial requisite for survival of the February order: building a bridge between the two fronts.

This became harder and harder to do from early summer on. Several factors contributed. With the government and the ruling parties it was a growing hardening of personal animosities and of positions on issues. But a more serious aggravation was the result of omissions and a bad decision. The *Kerensky* Offensive of June/July, named after the premier, was a fatal error. It ended in a fiasco, which not only caused foreign policy damage. It also discredited the Provisional Government at home. Economic depression worsened simultaneously. Unemployment spread as more and more companies failed. Strikes and demonstrations proliferated, bringing the State to the brink of anarchy. The formation of new coalitions and two 'Reconciliation Congresses' (with different compositions) were unable to check the rapid erosion of authority of the February Regime.

This benefited the *Bolshevik* Social Democratic Party, which declared an irreconcilable struggle with the 'capitalist government' when their undisputed leader, V. I. Lenin, returned to Petrograd from Swiss exile on April 3/16. They only had to channel the workers' and soldiers' misery and disenchantment over the failure of the new system. With all this grist for their mills, the Bolsheviks easily won great popularity. The spring splinter party was the strongest political force by the time the capital city councils were re-elected in September. What had happened and was about to happen was clearly evidenced by the dramatic change of leadership in the Petrograd Workers Soviet. The Bolshevik Trotzky replaced the Menshevik Chkheidze as chairman.

The 'Red October' (Daniels 1967) has produced much controversy. Soviet historiography, adhering to ideology, saw it as the investing of the first socialist state in the history of the world by the power of the masses. Others insisted on interpreting it as a military coup. There is no question that the seizure of power in the city garrisons from October 22–November 4 was executed before everyone's eyes by the Bolshevik-dominated Military Revolution Committee in defence of the Soviet. It is also undisputed that the Bolsheviks were able to obtain the support of a majority of delegates to the Second All-Russian Congress of Soviets, which convened on October 25. With this backing they managed to elect a presidium, more than half of whose members were their loyal supporters, while another third consisted of their Left Socialist Revolutionary sympathizers. To top it all off, the 'Right' Social Revolutionaries and the Mensheviks did them the favor of resigning their seats. Finally, the unchallenged occupation of strategic buildings and points in the city suggests that, (unlike in Moscow a little later) there was no resistance worth mentioning. However, it is equally obvious that armed force was used against a legitimate government that had become unpopular but had been elected by parliament. The storming of the Winter Palace and the arrest of ministers pointedly attests to the nature of the overthrow. The maneuver, accomplished by the Bolsheviks with the help of their Red Guard, was technically a military rebellion, and politically a *coup d'état*. But it was possible only because it was undergirded by the majority of the Soviet deputies and the tolerance, if not the sympathy, of the urban lower classes, and locally stationed soldiers in particular. Thus, the 'art of insurrection' preached and wholeheartedly practised by Lenin more or less boiled down to picking up the power that was lying in the street (N. Sukhanov).

The only question remaining was whether the Bolsheviks could maintain control. The wishes of the rebellious and potentially rebellious masses were

quickly satisfied. Early on October 26, the Soviet Congress passed a Peace Decree and a Land Decree. The intention of obtaining an immediate ceasefire secured the support of the soldiers for the overthrow. Distribution of large estates among village communes sanctioned transfers that had taken place much earlier and neutralized the peasantry. Overcoming the other parties and adherents of the *ancien régime* entailed a real struggle. It was inextricably tied to the independence movements of non-Russian nations and lasted until March 1921. Beginning in spring 1918 it escalated into the bloodiest war fought on Russian soil since the early seventeenth century. It actually started, still without recourse to weapons, right after the October coup. In the end the Bolsheviks maintained the upper hand so that they could begin the peaceful construction of 'their' State. However, at this juncture the fundamental decisions on its structure had already been made, and they were only superficially and temporarily amended, if at all, by the proclamation of the New Economic Policy.

The most important trend reflected Lenin's practical 'interpretation' of the slogan, 'All Power to the Soviets,' as 'All Power to the Bolsheviks,' which is what he had in mind in the first place. This usurpation was gradual. The cooperation between the Bolsheviks and the Left Socialist Revolutionaries in the October night session became a formal coalition in mid-November. Even this collaboration was based on the exclusion of not only the Liberals, but also the Mensheviks and 'Right' Socialist Revolutionaries. This confrontation manifested itself in the *alternative* between the *soviets* and the *Constituent Assembly*. When the latter was forcibly dispersed on January 6–19, 1918, it was not only the final curtain on the attempt to erect a democracy in post-tsarist Russia. It also dissolved most of the common ground between the coalition partners. The fate of the bipartisan government was sealed. All that was needed to precipitate its collapse was the separate peace treaty of Brest–Litovsk in March. Now one-party rule and the fusion of party and State began in earnest. The system which lasted until 1991 was established.

Moreover, at least three other features of the new government and political system had their roots in the post-October clashes. First, the life-and-death struggle fostered the emergence of extraordinary organs and extreme centralization of decision making. Both measures were only partially rescinded. The most important institutional legacy was the Cheka, which was soon an army of its own, operating in the hinterland and authorized to administer martial law. As a secret police, the Cheka later became a synonym for totalitarian rule because it was empowered to execute arbitrary force beyond and above the law and was only subordinated to the top Party leaders. A second fundamental decision was subjecting the economy to the control of the Party and the government. Starting in December 1917, the regime nationalized first the banks and big industry; then, during the Civil War, commerce and small business as well. Officially control was assumed by a pyramid of economic councils; *de facto*, however, since the unions had been forced to march in step, by members of the only legal party. In addition, an increasing intolerance of intra-party criticism reinforced the concentration of power and authority. The relative laxness which the Civil War encouraged in 1919 yielded at the conclusion of hostilities to a 'prohibition of factions' (March 1921), which Stalin used as a decisive weapon in the power struggle for Lenin's succession. Finally, the Red Army was created in the fight for the spoils of the October Revolution. Even though its structure and personnel underwent several reshuffles, one birthmark was never effaced—its orientation on the regime and the Party.

2. *Social Upheaval*

The revolutionary movement united political and social objectives from the start. It laid its claims in the name of the lower classes, which it always viewed as exploited. The Socialist Revolutionaries acted on behalf of all the 'enslaved,' but primarily the peasants, while the Social Democrats marched in the name of the urban 'proletariat.' The Liberals, on the other hand, considered themselves to be 'class neutral.' But actually this camp had a social stratum of its own, ranging from the reform-oriented aristocracy and the intelligentsia to the commercial-industrial bourgeoisie. In the first revolution these various tendencies underwent symbiosis. The nearly identical timing of the general strike in the capital and the peasants' storming of the agricultural estates in the autumn of 1905 shocked the State to such an extent that it agreed to concessions to the democratic demands of the Liberals.

In February 1917 a similar convergence took place, with one fundamental difference. Instead of the pincer attack of workers and peasants, an alliance of workers and garrisoned (not frontline) soldiers led the assault. This time the *peasants* remained aloof. The agrarian revolution took place later, but more violently than ever before, *after* the collapse of the tsarist state and its rule by force in summer 1917. Later the rural population divorced itself from the urban revolutionary movement and went its own separate way until the autumn of 1918. The core significance of the upheaval can be summed up with *one* result: the wildcat (i.e., spontaneous) redistribution of the property of the landed aristocracy. The peasants seized that which they had always believed to be their rightful possession. The predominant form of the operations was characteristic of the peasants' ties to tradition. Most villages acted collectively and allotted the land to commune members according to the dictates of custom to be used for a certain length of time. Ownership remained in the hands of the *obshchina*, and private

property came into *de facto* existence only because hardly any redistribution occurred thereafter. Following this, their revolution, the peasants were sated. Events and problems beyond their village left them cold. Since the New Economic Policy legalized the reform, they could perceive themselves as the winners of the revolution despite the horrors of the Civil War. However, this redistribution wreaked havoc on the economy. It abolished the large cultivated holdings which yielded the bulk of the farm produce sold elsewhere and reduced the average acreage to a level which hardly sufficed for efficient and stable production for the market. The outcome was a crisis-proneness that provided a convincing argument for a centrally planned economy.

The consequences of the February 1917 general strike of the Petrograd *workers* were more complicated. From the standpoint of the protagonists, the medium-term ramifications were negative. The rebels managed to obtain their demands only in the short run. Wages increased substantially, and working hours were reduced, even though the principal demand, a law stipulating an eight-hour workday, was not achieved. Personal and social security were markedly improved. But the galloping inflation and unemployment following the economic collapse offset these benefits from early summer on. This setback made it possible for the syndical factory committee movement to extend the say in management it had gained in Petrograd in March to full control. Its victory was hollow because the victors inherited the administration of the bankruptcy. To add insult to injury, the expectations it cherished after the October overthrow were completely dashed. If the Bolsheviks had a typical way of dealing with 'their' constituency, it was revealed in their relations with the factory committees. Before the overthrow they adopted the demands of the committees; after the Bolsheviks seized power, they also succeeded in gaining control of these organs. At the beginning of the year the 'anarchistic-decentralized' factory committees were muzzled by the unions, which were themselves on the way to becoming a 'transmission belt' (Lenin) of the Party. All that remained was the disappointment over unkept promises, which triggered spreading strikes, the rebellion of Kronstadt sailors and intraparty criticism ('Workers' Opposition'). All this dissidence was suppressed by spring 1921, in many cases with extreme brutality.

3. *Disintegration of the Empire and National Independence Movements*

Liberal political and sociological views of the Russian Revolution have overlooked too much evidence that it was also a *national* movement. Incipient emancipation campaigns were exhibited in the first 'crisis of authority' in 1905–06. Social unrest in the Baltic provinces and in Poland were particularly violent. Both there and in the Caucasus it expressed a desire for autonomy. The first authentically freely- elected State Duma had several national factions and a total of about 45 percent non-Russian deputies. Following the 'prewar' and war period, which gave these movements an enormous boost, and the collapse of the tsarist empire, the national outbursts flared up as attempts to secede by force. Favored by the 'an-archy' in the literal sense created by the power vacuum, the 'old' nations like Poland and Finland and non-Slavic Baltic and Caucasian regions declared independence from the Petrograd central government. Nuclei of the old empire such as Ukraine joined the breakaway trend. It was not without cause that the June 1917 resolution of the Ukrainian *Rada* (Council) declaring it would act as a formally independent body plunged the Provisional Government into an extremely severe crisis. The Bolsheviks supported this national emancipation only verbally and for tactical reasons. Prior to their power seizure and during the Civil War in the peripheral areas of the empire, they espoused the doctrine that nation states had to reclaim their historical rights in the 'bourgeois phase' of the revolution. However, the stronger the *socialist* self-awareness of the new regime became, the more it again embraced the necessity for a unitary state. In fact, the Red Army acted from the beginning as an arm and an instrument of Great Russian centralism which regained control of the old empire—with the exception of Finland, Poland, and the new Baltic republics—by the end of the Civil War. If one looks very closely, this resulted inevitably from the political foundations of the new State. Since only one political force was legal, it had to be centralist. Therefore the Soviet Union established in 1922 was only formally a federal system. In truth it was an anachronistic multinational empire in a postwar world organized into nation states and dominated by nationalisms.

4. *'Iconoclasm'*

Last but not least, scholars have recently remembered to highlight the extremely radical cultural upheaval the Russian Revolution entailed. In the first place, it had a salient 'ideocratic' dimension. The Leninist version of radical Marxism was an ideology with the political power to stake a claim to shaping all sectors of public life, and ultimately also of all private life of the people under its sway, from work to education, according to fundamentally new, theoretically preconceived principles. The October overthrow ushered in the first regime in modern history to make a specified interpretation of world history and destiny, including a programmatic, militant atheism, obligatory. The radical Bolshevik Revolution unleashed an unprecedented outburst of all kinds of 'anti-establishment' movements, almost entirely borrowed from Western

European modernism. They were of a sociopolitical (women's liberation), intellectual-aesthetic (abstract art, '*Proletkult*') and pedagogical nature (polytechnic education, public health). Even though this experimental exuberance was a source of aggravation to the Party and government leadership and its wings were gradually clipped after the conclusion of the Civil War, it survived in isolated pockets until definitively forced into submission by Stalin. In the aesthetic sector, the 'revolutionary culture' is universally credited with significant achievements.

See also: Communism; Communism, History of; Communist Parties; Dictatorship; Dictatorship in History; Revolution; Revolutions, Sociology of; Revolutions, Theories of; Soviet Studies: Politics; Soviet Studies: Society; Totalitarianism

Bibliography

Altrichter H 1997 *Rußland 1917. Ein Land auf der Suche nach sich selbst*. Schöningh, Paderbor, Germany

Brovkin V 1987 *The Mensheviks After October. Socialist Opposition and the Rise of Bolshevik Dictatorship*. Cornell University Press, Ithaca, NY

Brovkin V N 1994 *Behind the Front Lines of the Civil War. Political Parties and Social Movements in Russia, 1918–1922*. Princeton University Press, Princeton, NJ

Daniels R V 1967 *Red October: The Bolshevik Revolution of 1917*. Sribner, New York

Figes O 1996 *A People's Tragedy. The Russian Revolution 1891–1925*. Cape, London

Figes O 1989 *Peasant Russia, Civil War. The Volga Countryside in Revolution* (*1917–1921*). Clarendon, Oxford, UK

Galili Z 1989 *The Menshevik Leaders in the Russian Revolution: Social Realities and Political Strategies*. Princeton University Press, Princeton, NJ

Gill G J 1979 *Peasants and Governments in the Russian Revolution*. Macmillan, London

Hasegawa T 1981 *The February Revolution: Petrograd 1917*. University of Washington Press, Seattle, WA

Koenker D P, Rosenberg W G 1989 *Strikes and Revolution in Russia*, 1917. Princeton University Press, Princeton, NJ

Leggett G 1981 *The Cheka: Lenin's Political Police*. Clarendon, Oxford, UK

Pipes R 1990 *The Russian Revolution*, 1st edn. Knopf, New York

Rabinowitch A 1978 *The Bolsheviks Come to Power*, 1st edn. N W Norton, New York

Stites R 1989 *Revolutionary Dreams. Utopian Vision and Experimental Life in the Russian Revolution*. Oxford University Press, Oxford, UK

Wildman A K 1979, 1987 *The End of the Russian Imperial Army*. Princeton University Press, Princeton, NJ, 2 vols

M. Hildermeier

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

S

Sacrifice

There is no scholarly consensus regarding the meaning of the term 'sacrifice.' In its widest usage the term, which derives from the Latin *sacrificium*, is applied to virtually any form of gift-giving to a deity. A more restricted application limits its use to situations in which the gift is mutilated or destroyed. An even more restrictive usage is suggested by the Oxford English Dictionary (Simpson and Weiner 1989/1933), whose primary definition is 'the slaughter of an animal (often including the subsequent consumption of it by fire) as an offering to God or a deity.'

The term sacrifice is, therefore, best regarded as a polythetic category which includes heterogeneous modes of behavior motivated by intentions of variable kinds. These include commemoration, initiation, expiation, propitiation, establishing a covenant, or affecting a transition in social status. Regarding sacrifice not as constituting a class of phenomena having a fixed set of features in common but as a multivocal term applying to phenomena that merely have 'family resemblances' in common has the advantage of mitigating somewhat the Euro-centric bias inherent in the work of many of the most influential thinkers on the subject, among them Robertson Smith (1894/1927), Hubert and Mauss (1898), and Durkheim (1912). 'The Sacrifice' for this generation of scholars was the central feature of religion, a Judaic–Christian bias still evident as late as 1956, the year in which E. E. Evans-Pritchard published his classic, *Nuer Religion*. Tellingly, Evans-Pritchard did not identify any word in the Nuer language that unambiguously corresponded to the English 'sacrifice.' This lack of correspondence appears widespread. Benveniste (1969, p. 223), for example, could discover no common word for sacrifice in the Indo-European languages.

1. *The Terminology of Sacrifice*

In more elaborate forms than straightforward ritualized gift-giving, sacrifice typically involves a set of four terms. First, there is the sacrificer, the executor of the ritual. Second, there is the sacrifier, the beneficiary of the sacrifice. This may be the executor of the ritual, individual members of the community, or the community *in toto*. Third, there is the object sacrificed. Although this gift may be anything, ethnographic descriptions usually emphasize the bloody sacrifice of a living being, in which case blood typically is described as connoting such ideas as fertility, renewal, replenishment,vitality, life force, and empowerment.

Access by the sacrifier to these desirables is acquired when this substance is shed in ritual by the act of forcing life out of the victim. Substitution, or the use of a surrogate in cases when the offering is that of a human being, is a common occurrence, so frequently resorted to that some writers, including two who have made important contributions to the topic—de Heusch (1986) and Bloch (1992, p. 32)—consider substitution to be a key feature of sacrificial rituals. Fourth, there is the deity to whom the sacrifice is made.

2. *Approaches to the Sacrifice*

A number of distinguished thinkers have advanced what might be termed 'approaches' to the study of sacrifice, which afford insightful, though overlapping, perspectives. Among the most useful are the following five.

2.1 *Sacrifice as Gift*

For Tylor (1871/1913), following Plato, the sacrifice was a gift made to an anthropomorphic divinity, much as a gift might be given to a chief, whose acceptance made the deity indebted to humanity. This approach was popular with nineteenth century scholars, who attributed to the act of sacrificing such motives as expiating sins, obtaining benefits, nourishing the deity, and establishing good relations between human beings and spirits. This theory lays stress on the beliefs of the givers rather than on the ritual performance itself, which in this interpretation is relegated to logically subservient status. This approach is at the same time too broad and too narrow. It is overly broad in that these motives undoubtedly are widespread in many rituals of sacrifice all over the world, yet excessively narrow since sacrificial rituals are employed to do much more than simply offer gifts.

2.2 *Sacrifice as Communion*

Robertson Smith (1894/1927), in contrast to Tylor, granted priority to ritual, a ranking modern ethno-

graphic fieldwork tends to validate. Robertson Smith's main point was that early Semitic sacrifice, and by implication sacrifice among nonliterate peoples, was a feast human beings communally shared with their deity. He understood the immolated victim to be the clan totem, an animal that shared in the blood of its human kin. It was also the clan ancestor, which was the clan deity, so in consuming their animal victim members of the community were eating their deity. This act of consumption conferred spiritual power.

Later research in the field undermined much of Robertson Smith's argument, but unquestionably the sacrifice does offer a means by which human beings may conjoin with their deity and establish a relationship, and here Robertson Smith did have insight. He strongly influenced Hubert and Mauss, whose *Essai sur la nature et la fonction du sacrifice*, which appeared in 1898, remains the basic work on the topic. Like all the studies of the time, it depended entirely upon secondary and tertiary sources rather than original ethnography, and the authors further restricted their sources to Sanskrit and Hebrew texts. To them the most significant feature about the sacrificial feast was that it provided the mechanism by which communion, in the sense of communication between human beings and deity, could be brought about. It was made possible by the fact that the victim represented both sacrifier and divinity, thus breaking down the barrier between the sacred and the profane (a moral contrast that played a major role in their approach). Yet, although the sacrifical act conjoined the sacred and profane, it could also disjoin them since while there were sacrifices that brought the spirit into the human world and made the profane sacred, there were also sacrifices that removed the spirit from the human world and made what had become sacred, profane. The former Hubert and Mauss called rituals of 'sacralization;' the latter, rituals of 'desacralization.' Even though Evans-Pritchard—who was very much influenced by their ideas—found this distinction among the Nuer (1956, pp. 197–230), it is by no means as universal as they supposed.

Hubert and Mauss were intrigued by those sacrifices in which it is the deity that offers his life for the benefit of his worshippers. As they saw it, the exemplary sacrifice was that of a god, who, through unqualified self-abnegation, offers himself to humankind, in this way bestowing life on lesser beings. The most celebrated example of this is the central ritual of the Catholic religion, the Mass, where Christ's sacrifice on the Cross is re-enacted. In a sacrifice of this nature the relationship between the deity and worshippers raises the question of relative status. From Hubert and Mauss' perspective the relationship is one of grave imbalance since the deity gives more than it gets. This may appear so, but Valeri (1985, pp. 66–7) has advanced a sophisticated argument designed to prove that sacrifice actually creates a bond of 'mutual' indebtedness between the human and the divine, which makes each party dependent upon the other. This proposition may not be true universally, but support for Valeri's argument comes from the bear sacrifices carried out by the Ainu of Japan. In these, the immolation of bears brings fertility to the community and nurture to the spirits (Irimoto 1996).

Durkheim's (1912) distinctive contribution, which otherwise relied heavily upon Hubert and Mauss, and Robertson Smith, was to promote the importance of sacrificial activities in bringing about social cohesion and the values that contribute to this cohesion. In Timor, a South Pacific island, a ritual is performed in which a local king is figuratively slain by members of two ethnically different communities (Hicks 1996). The king, who comes from one of these groups, symbolizes their unity as a single social entity at the same time that he symbolizes their local god. In addition, by periodically immolating their king the two groups regenerate their notions of 'society' and 'god' as epistemological categories.

2.3 Sacrifice as Causality

For Hocart (1970, p. 217) the most compelling reason for performing collective rituals was the 'communal pursuit of fertility.' A sacrifice is offered to the gods in order that life and the fertility that engenders and sustains it are acquired by the community, a view that focuses upon sacrifice as a practical device to affect empirical transformation. Catholic rituals before the Reformation were thought of in this way; afterwards they were reinterpreted as a symbolic process (Muir 1997). For Catholics the Mass materially transformed the bread into Christ's body; for Protestants the bread merely represented Christ's sacrifice.

2.4 Sacrifice as Symbol System

More contemporary scholars, such as Valeri (1985), have emphasized the symbolic character of sacrificial behavior, analyzing its structure and isolating its motifs within the wider context of a society's ideology. In this approach, the array of symbols isolated is then apprehended as a semantic system whose meaning requires decoding. Valeri's interpretation of Hawaiian sacrifice, for example, requires that sacrificial rituals be understood as a symbolic action 'that effects transformations of the relationships of sacrifier, god, and group by representing them in public' (1985, pp. 70–1). Although vulnerable to the criticism that it relies too heavily on the subjective bias of the analyst, this approach has the advantage of enabling sacrificial practices to be understood within a wider social context, at the same time as it allows an analysis of sacrifice to open up a distinctive perspective on society.

2.5 Sacrifice as Catharsis

Another approach to sacrifice is to focus on violence. Girard (1972), most notably, has suggested that sacrifice offers a socially controlled, and therefore, socially acceptable, outlet for the aggressive urges of human beings. The typical manner by which this is accomplished is to use a sacrificial victim as a scapegoat. A more recent scholar who has stressed the importance of violence in sacrifice is Bloch (1992), but his approach differs from that of Girard in that for Bloch sacrifice requires violence because human beings need 'to create the transcendental in religion and politics,' and that this is attainable through violence done to the victim (Bloch 1992, p. 7). One difficulty with this approach is that behavior that one might arguably class as sacrificial may involve no violence at all.

3. Origins of Sacrifice

Sacrificial rituals have an extensive pedigree. One of the earliest for which a reasonably detailed record is available dates from about 3,000 years ago. The Vedic Aryans held sacrifice to be their central religious activity (Flood 1996, pp. 40–1), and it consisted of putting milk, grains of rice and barley, and domesticated animals into a sacred fire that would transport the offerings to supernatural entities. As the Agnicayana, this ritual is still carried out in Kerala, southwest India.

Evidence for sacrificial behavior extends further back than Vedic times, to the civilizations of Sumer 5,000 years ago and China 4,000 years ago, and its presence seems accounted for even further in the past. This long history, taken with the widespread practice of sacrifice all over the world, suggests that a case could be made that sacrificial behavior is one of the natural dispositions common to the species *Homo sapiens*. As Needham (1985, p. 177) has averred for ritual in general, it may be that sacrifice, 'Considered in its most characteristic features … is a kind of activity—like speech or dancing—that man as a ceremonial animal happens naturally to perform.'

See also: Commemorative Objects; Exchange in Anthropology; Honor and Shame; Religion: Morality and Social Control; Ritual; Trade and Exchange, Archaeology of

Bibliography

Benveniste E 1969 *Le Vocabulaire des institutions indo-européennes, I. Economie, parenté, société; II. Pouvoir, droit, religion*. Les Editions de Minuit, Paris

Bloch M 1992 *Prey into Hunter: The Politics of Religious Experience*. Cambridge University Press, Cambridge, UK

Durkheim E 1912/1968 *Les Formes élémentaires de la vie religieuse: le système totémique en Australie*. Presse Universitaires de France, Paris

Evans-Pritchard E E 1956 *Nuer Religion*. Clarendon Press, Oxford, UK

Flood G O 1996 *An Introduction to Hinduism*. Cambridge University Press, Cambridge, UK

Frazer J G 1922 *The Golden Bough: A Study in Magic and Religion,* Vol. 1. Macmillan, London

Girard R 1972 *La Violence et le sacré*. Bernard Grasset, Paris

de Heusch L 1986 *Le Sacrifice dans les religions africaines*. Editions Gallimard, Paris

Hicks D 1996 Making the king divine: A case study in ritual regicide from Timor. *Journal of the Royal Anthropological Institute* **2**: 611–24

Hocart A M 1970 *Kings and Councillors: An Essay in the Comparative Anatomy of Human Society*. University of Chicago Press, Chicago

Hubert H, Mauss H 1898 Essai sur la nature et la fonction du sacrifice. *Année sociologique* **2**: 29–138

Irimoto T 1996 Ainu worldview and bear hunting strategies. In: Pentikäinen J (ed.) *Shamanism and Northern Ecology* (Religion and Society 36). Mouton de Gruyter, Berlin, pp. 293–303

Muir E 1997 *Ritual in Early Modern Europe*. Cambridge University Press, Cambridge, UK

Needham R 1985 *Exemplars*. University of California Press, Berkeley, CA

Simpson J A, Weiner E S C 1933/1989 *Oxford English Dictionary*, 2nd. edn. Clarendon Press, Oxford, UK

Smith W R 1894/1927 *Lectures on the Religion of the Semites*, 3rd edn. Macmillan, New York

Tylor E B 1871/1913 *Primitive Culture*, 4th edn. Murray, London, Vol. 2

Valeri V 1985 *Kinship and Sacrifice: Ritual and Society in Ancient Hawaii* (trans. Wissing P). University of Chicago Press, Chicago

D. Hicks

Safety, Economics of

The economics of safety is concerned with the causes and prevention of accidents resulting in economic losses. The primary focus is on accidents causing workplace injuries and diseases. The topic is examined using several variants of economic theory. One set of theories applies when there is an ongoing relationship between the party who is generally the source of the accidents and the party who generally bears the costs of the accidents. Negotiations (explicit or implicit) between the parties can provide *ex ante* economic incentives to reduce accidents. Another set of theories applies when there is no ongoing relationship between the party who causes the accident and the party who suffers the consequences. *Ex post* financial penalties for the party who caused the harm can deter others from causing accidents.

1. 'Pure' Neoclassical Economics

The neoclassical theory of work injuries is examined in Thomason and Burton (1993). Workplace accidents are undesirable by-products of processes that produce

goods for consumption (Oi 1974). Several assumptions are used in the neoclassical analysis of these processes. Employers and workers have an ongoing relationship. Employers maximize profits. Workers maximize utility, not just pecuniary income. Workers at the margin are mobile and possess accurate information concerning the risks and consequences of work injuries. One implication of these assumptions is that employers pay higher wages for hazardous work than for nonhazardous work. On the assumption that accident insurance is not provided by employers or the government, the equilibrium wage for the hazardous work will include a risk premium equal to the expected costs of the injuries, which includes lost wages, medical care, and the disutility caused by the injury. If workers are risk averse, the premium will also include a payment for the uncertainty about who will be injured.

The employer has an incentive to invest in safety in order to reduce accidents and thus the risk premium. The firm will make safety investments until the marginal expenditure on safety is equal to the marginal reduction in the risk premium. Since there is a rising marginal cost to investments in safety, equilibrium will occur with a positive value for the risk premium, which means that in equilibrium there will be some work injuries.

Variants on neoclassical economics theory can be generated by changing some of the previous assumptions. For example, if all workers purchase actuarially fair insurance covering the full costs of work injuries, they will be indifferent about being injured. Furthermore, if injured, workers will have no incentive to return to work since the insurance fully compensates them for their economic losses. The change in worker behavior resulting from the availability of insurance is an example of the moral hazard problem, in which the insurance increases the quantity of the events being insured against (i.e., the occurrence of injuries and the durations of the resulting disabilities). These variants on neoclassical economics with their implications for employee behavior are discussed in Burton and Chelius (1997).

1.1 Evidence Consistent with the Neoclassical Model

There is a burgeoning literature on compensating wage differentials for the risks of workplace death and injury. Ehrenberg and Smith (1996) report that generally studies find industries having the average risk of job fatalities (about 1 per 10,000 workers per year) pay wages that are 0.5 percent to 2 percent higher than the wages for comparable workers in industries with half that level of risk. Viscusi (1993) reviewed 17 studies estimating wage premiums for the risks of job injuries, and 24 studies estimating premiums for the risks of job fatalities. The total amount of risk premiums implied by these various estimates is substantial: probably the upper bound is the Kniesner and Leeth (1995) figure of $200 billion in risk premiums in the US in 1993.

The Viscusi (1993) survey of studies of risk premiums for workplace fatalities included evidence from the United Kingdom (UK), Quebec, Japan, and Australia. More recently, Siebert and Wie (1994) found evidence for risk premiums for work injuries in the UK, and Miller et al. (1997) provided additional evidence on compensating differentials for risk of death in Australia.

1.2 Qualifications Concerning the Neoclassical Theory

The model that postulates workplace safety results from risk premiums that provide financial incentives to employers to invest in safety has been challenged. Some critics assert that risk premiums are inadequate because workers lack sufficient information about the risks of work injuries and/or have limited mobility to move to less hazardous jobs. Ehrenberg and Smith (1996) conclude that workers have enough knowledge to form accurate judgments about the relative risks of various jobs. However, Viscusi (1993) refers to a sizable literature in psychology and economics documenting that individuals tend to overestimate low probability events, such as workplace injuries, which could result in inappropriately large risk premiums. Both Ehrenberg and Smith (1996) and Viscusi (1993) argue there is sufficient worker mobility to generate risk premiums. Another challenger to the neoclassical model is Ackerman (1988), who argues that the labor market does not generate proper risk premiums because certain costs resulting from work injuries are not borne by workers but are externalized.

The empirical evidence on risk premiums has also been challenged or qualified. Ehrenberg and Smith (1996), for example, conclude that the studies of compensating wage differentials for the risks of injury or death on the job 'are generally, but not completely, supportive of the theory.' Viscusi (1993) concludes: 'The wage-risk relationship is not as robust as is, for example, the effect of education on wages.'

The most telling attack on the compensating wage differential evidence is by Dorman and Hagstrom (1998), who argue that most studies have been specified improperly. They find that, after controlling for industry-level factors, the only evidence for positive compensating wage differential pertains to unionized workers. For nonunion workers, they argue that properly specified regressions suggest that workers in dangerous jobs are likely to be paid less than equivalent workers in safer jobs.

The Dorman and Hagstrom challenge to the previously generally accepted view—that dangerous work results in risks premiums for most workers—is likely to result in new empirical studies attempting to resuscitate the notion of risk premiums. However, even if every empirical study finds a risk premium for workplace fatalities and injuries, the evidence would not conclusively validate the neoclassical economics

approach. Ehrenberg (1988) provides the necessary qualification: if there are any market imperfections (such as lack of information or mobility), the 'mere existence of *some* wage differential does *not* imply that it is a fully compensating one.' And if the risk premium is not fully compensating (or is more than fully compensating), then *inter alia*, the market does not provide the proper incentive to employers to invest in safety.

Another qualification concerning the use of risk premiums as a stimulus to safety is that institutional features in a country's labor market may aid or impede the generation of risk premiums. In the US, the relative lack of government regulation of the labor market and the relative weakness of unions probably facilitate the generation of risk premiums. In countries like Germany, where unions are relatively strong and apparently opposed to wage premiums for risks, and where labor markets appear to be less flexible than those in the US because of government regulation of the labor market, wages are less likely to reflect underlying market forces that produce risk premiums.

2. *Modified Neoclassical Economics and the Old Institutional Economics*

Some economists rely on a 'modified' version of neoclassical economics theory that recognizes limited types of government regulation of safety and health are appropriate in order to overcome some of the attributes of the labor market that do not correspond to the assumptions of pure neoclassical economics. These attributes include the lack of sufficient knowledge and mobility by employees, and the possible lack of sufficient knowledge or motivation of employers about the relationship between expenditures on safety and the reduction in risk premiums. The 'old institutional economists' (OIE) would agree with these critiques of lack of knowledge and mobility, and would also emphasize factors such as the unequal bargaining power of individual workers as another limitation of the labor market.

An example of government promulgation of information in order to overcome the lack of knowledge in the labor market is the Occupational Safety and Health Act (OSHAct) Hazard Communication standard, which requires labeling of hazardous substances and notification to workers and customers and which is estimated to save 200 lives per year (Viscusi 1996). (Other aspects of the OSHAct are discussed below.)

2.1 *Experience Rating*

An example of a labor market intervention that could improve incentives for employers to invest in safety is the use of experience rating in workers' compensation programs, which provide medical and cash benefits to workers injured at work. Firm-level experience rating is used extensively in the workers' compensation programs in the US, and is used to some degree in many other countries, including Canada, France, and Germany. Firm-level experience rating determines the workers' compensation premium for each firm above a minimum size by comparing its prior benefit payments to those of other firms in the industry.

In the pure neoclassical economics model, the introduction of workers' compensation with experience rating should make no difference in the safety incentives for employers compared to the incentives provided by the labor market without workers' compensation. Under assumptions such as perfect experience rating, risk neutrality by workers, and actuarially fair workers' compensation premiums, which are explicated by Burton and Chelius (1997), the risk premium portion of the wage paid by the employer will be reduced by an amount exactly equal to the amount of the workers' compensation premium. Also, under these assumptions, the employer has the same economic incentives to invest in safety both before and after the introduction of the workers' compensation program. Under an alternative variant of the pure neoclassical economics approach, in which the assumption of perfect experience rating is dropped, the introduction of workers' compensation will result in reduced incentives for employers to reduce accidents.

In contrast, the OIE approach argues that the introduction of workers' compensation with experience rating should improve safety because the limitations of knowledge and mobility and the unequal bargaining power for employees mean that the risk premiums generated in the labor market are inadequate to provide employers the safety incentives postulated by the pure neoclassical economics approach. Commons (1934), a leading figure in the OIE approached, claimed that unemployment is the leading cause of labor market problems, including injuries and fatalities, because slack labor markets undercut the mechanism that generates compensating wage differentials. Commons asserted that experience rating provides employers economic incentives to get the 'safety spirit' that would otherwise be lacking. The modified neoclassical economics approach also accepts the idea that experience rating should help improve safety by providing stronger incentives to employers to avoid accidents, although they place less emphasis on the role of unemployment in undercutting compensating wage differentials and more emphasis on the failure of employers to recognize the cost savings possible from improved safety without the clear signals provided by experience-rated premiums.

A number of recent studies of the workers' compensation program provide evidence that should help assess the virtues of the various economic theories. However, the evidence from US studies is inconclusive. One survey of studies of experience rating by Boden (1995) concluded that 'research on the safety impacts

has not provided a clear answer to whether workers' compensation improves workplace safety.' In contrast, a recent survey by Butler (1994) found that most recent studies provide statistically significant evidence that experience rating 'has had at least some role in improving workplace safety for large firms.' Burton and Chelius (1997) sided with Butler. The beneficial effect of experience rating on safety has been found in Canada by Bruce and Atkins (1993) and in France by Aiuppa and Trieschmann (1998).

The mildly persuasive evidence that experience rating improves safety is consistent with the positive impact on safety postulated by the OIE approach and the modified neoclassical economists, and is inconsistent with the pure neoclassical view that the use of experience rating should be irrelevant or may even lead to reduced incentives for employers to improve workplace safety.

2.2 Collective Action

While the modified neoclassical economists and the OIE largely would agree on the desirability of several prevention approaches, such as the use of experience rating in workers' compensation, the OIE endorsed another approach that many modified neoclassical economists would not support, namely the use of collective bargaining and other policies that empower workers. Collective bargaining agreements can minimize unsafe activities or at least explicitly require employers to pay a wage premium for unsafe work. Many agreements also establish safety committees that assume responsibility for certain activities, such as participation in inspections conducted by government officials. If workers are injured, unions can help them obtain workers' compensation benefits, thereby increasing the financial incentives for employers to improve workplace safety.

The beneficial effects predicted by the OIE approach for these collective efforts appear to be achieved. Several studies, including Weil (1999) concluded that OSHA enforcement activity was greater in unionized firms than in nonunionized firms. Moore and Viscusi (1990) and other researchers found that unionized workers receive larger compensating wage differentials for job risks than unorganized workers. Hirsch et al. (1997) found that 'unionized workers were substantially more likely to receive workers' compensation benefits than were similar nonunion workers.' Experience rating means that these higher benefit payments will provide greater economic incentives for these unionized employers to reduce injuries.

There are also safety committees in some jurisdictions that are utilized in nonunionized firms. Burton and Chelius (1997) reviewed several studies involving such committees in the US and Canada, and found limited empirical support for their beneficial effects. Reilly et al. (1995) provided a more positive assessment for their accomplishments in the UK. These studies provide some additional support for the beneficial role of collective action postulated by the OIE.

3. The New Institutional Economics

New institutional economics (NIE) authors generally argue that market forces encourage efficient forms of economic organization without government assistance and that opportunities for efficiency-improving public interventions are rare. This section considers only the Coase theory/transaction costs economics strain of NIE.

3.1 The Coase Theory and Transaction Costs Economics

In the absence of costs involved in carrying out market transactions, changing the legal rules about who is liable for damages resulting from an accident will not affect decisions involving expenditures of resources that increase the combined wealth of the parties. The classic example offered by Coase (1988) involves the case of straying cattle that can destroy crops growing on neighboring land. The parties will negotiate the best solution to the size of the herd, the construction of a fence, and the amount of crop loss due to the cattle whether or not the cattle-rancher is assigned liability for the crop damages.

Coase recognized that the assumption that there were no costs involved in carrying out transactions was 'very unrealistic.' According to Coase (1988):

> In order to carry out a market transaction, it is necessary to discover who it is that one wishes to deal with … and on what terms, to conduct negotiations …, to draw up the contract, to undertake the inspection needed to make sure that the terms of the contract are being observed, and so on. These operations are often … sufficiently costly … to prevent many transactions that would be carried out in a world in which the pricing system worked without cost.

The goal of transactions costs economics is to examine these costs and to determine their effect on the operation of the economy. As scholars of transaction costs have demonstrated, when transactions costs are significant, changing the legal rules about initial liability can affect the allocation of resources.

3.2 Evidence Concerning Changes in Liability Rules

Workplace safety regulation provides a good example of a change in liability rules, since in a relatively short period (1910–20), most US states replaced tort suits (the employer was only responsible for damages if negligent) with workers' compensation (the employer

is required to provide benefits under a no-fault rule) as the basic remedy for workplace injuries. Evidence from Chelius (1977) 'clearly indicated that the death rate declined after workers' compensation was instituted as the remedy for accident costs.' This result suggests that high transaction costs associated with the determination of fault in negligence suits were an obstacle to achieving the proper incentives to workplace safety, and that the institutional features of workers' compensation, including the no-fault principle and experience rating, provided a relatively more efficient approach to the prevention of work injuries.

An interesting study that also suggests that institutional features can play a major role in determining the effects of changing liability rules is Fishback (1987). He found that fatality rates in coal mining were generally higher after states replaced negligence law with workers' compensation. Fishback suggested that the difference between his general results and those of Chelius might be due to high supervision costs in coal mining in the early 1900s.

The transaction economics component of the NIE theory thus appears to provide a useful supplement to neoclassical economics, since institutional features, such as liability rules, can have a major impact on the economic incentives for accident prevention.

4. *Law and Economics*

Law and economics (L&E) theory draws on neoclassical economics and transaction cost economics, but is distinctive in the extent to which it examines legal institutions and legal problems. This section examines the tort law branch of L&E theory, while the next section examines the employment law branch.

4.1 *Theoretical Stimulus of Tort Law to Safety*

Tort law typically is used when one party harms another party, and the parties do not have an ongoing relationship. However, even though workers and employers have a continuing relationship, tort suits were generally used as a remedy for workplace accidents in the US until workers' compensation programs were established. When negligence is the legal standard used for tort suits, if the employer has not taken proper measures to prevent accidents and thus is at fault, the employer will be liable for all of the consequences of the injury. The standard for the proper prevention measure was developed by Judge Learned Hand and restated by Posner (1972) as:

> The judge (or jury) should attempt to measure three things in ascertaining negligence: the magnitude of the loss if an accident occurs; the probability of the accident's occurring; and the burden (cost) of taking precautions to prevent it. If the product of the first two terms, the expected benefits, exceeds the burden of precautions, the failure to take those precautions is negligence.

Posner argued that proper application of this standard would result in economically efficient incentives to avoid accidents. Burton and Chelius (1977) examine some qualifications to this conclusion.

4.2 *Evidence on the Tort Law Stimulus to Safety*

There are two types of empirical evidence that suggest skepticism is warranted about the stimulus to workplace safety from tort suits. First, tort suits were used as the remedy for workplace injuries in the late 1800s and early 1900s. As previously noted, Chelius (1977) found that the replacement of the negligence remedy with workers' compensation led to a general reduction in workplace fatalities. The Fishback (1987) contrary result for a specific industry—coal mining—provides a qualification to this general result.

Second, in other areas of tort law, there is a major controversy among legal scholars about whether the theoretical incentives for safety resulting from tort suits actually work. One school of thought is exemplified by Landes and Posner (1987), who state that 'although there has been little systematic study of the deterrent effect of tort law, what empirical evidence there is indicates that tort law … deters.' An opposing view of the deterrent effects of tort law is provided by Priest (1991), who finds almost no relationship between actual liability payouts and the accident rate for general aviation and states that 'this relationship between liability payouts and accidents appears typical of other areas of modern tort law as well, such as medical malpractice and product liability.'

A study of tort law by Schwartz (1994) distinguished a strong form of deterrence (as postulated by Landes and Posner) from a moderate form of deterrence, in which 'tort law provides a significant amount of deterrence, yet considerably less that the economists' formulae tend to predict.' Schwartz surveyed a variety of areas where tort law is used, including motorist liability, medical malpractice, and product liability, and concluded that that the evidence undermines the strong form of deterrence but provides adequate support for the moderate form of deterrence. As to workers' injuries, Schwartz cited the Chelius and Fishback studies and concluded 'it is unclear whether a tort system or workers' compensation provides better incentives for workplace safety.' Burton and Chelius (1997) concluded, based on both the ambiguous historical experience of the impact of workers' compensation on workplace safety and the current controversy over the deterrence effect in other areas of tort law, that 'the law and economics theory concerning tort law does not provide much assistance in

designing an optimal policy for workplace safety and health.'

Further examinations of the relative merits of workers' compensation and the tort system in dealing with workplace accidents are Dewees et al. (1996) and Thomason et al. (1998).

5. Government Mandate Theory vs. Law and Economics Theory

5.1 The Government Mandate Theory

The government mandate theory argues that the government promulgation of health and safety standards and enforcement of the standards by inspections and fines will improve workplace safety and health. This is basically a legal theory, although many of the supporting arguments involve reinterpretation or rejection of studies conducted by economists. For example, proponents of the theory object to the evidence on compensating wage differentials and on the deterrent effect of experience rating and object in principle to economists' reliance on cost-benefit analysis. Supporters of the theory, such as McGarity and Shapiro (1996), also provide a positive case why OSHA is necessary:

> OSHA's capacity to write safety and health regulations is not bounded by any individual worker's limited financial resources. Likewise, OSHA's capacity to stimulate an employer to action does not depend upon the employees' knowledge of occupation risks or bargaining power.

The government mandate theory would not be endorsed by many economists, including the OIE. Commons and Andrews (1936), for example, criticized at length the punitive approaches that used factory inspectors in the form of policemen, since this turned employers into adversaries with the law. The minimum standards supported by the OIE were those developed by a tripartite commission, involving employers, employees, and the public, rather than standards promulgated by the government. While the OIE theory is, thus, unsympathetic to the government mandate theory, the sharpest attack is derived from the L&E theory.

5.2 The L&E Theory Concerning Government Regulations

L&E scholars make a distinction between mandatory, minimum terms (standards) and those terms that are merely default provisions (or guidelines) that employers and employees can agree to override. Most employment laws, including workplace safety laws, create standards and are thus objectionable to the L&E scholars.

Willborn (1988) articulated the standard economic objections to mandatory terms. Employers will treat newly imposed standards like exogenous wage increases and in the short run will respond by laying off workers. In the long run, employers will try to respond to mandates by lowering the wage. The final wage-benefits-standards employment package will make workers worse off than they were before the imposition of standards—otherwise the employers and workers would have bargaining for the package without legal compulsion.

5.3 Evidence on the Effects of OSHA Standards

The evidence, as reviewed by Burton and Chelius (1997), suggests that the OSHAct has done little to improve workplace safety, thus lending more support to the L&E theory than to the government mandate theory. OSHA's ineffectiveness in part may be due to the lack of inspection activity, since the average establishment is only inspected once every 84 years. But the evidence also suggests that allocating additional resources to plant inspections may be imprudent. Smith (1992) concluded, after an exhaustive review of the studies of OSHA inspections, that the evidence 'suggests that inspections reduce injuries by 2 percent to 15 percent,' although the estimates often are not statistically significant (and thus cannot be distinguished confidently from zero effect).

Several studies, including Scholz and Gray (1990) and Weil (1996) provide a more favorable assessment of the OSHA inspection process. However, even Dorman (1996), who supports an aggressive public policy to reduce workplace injuries, provided a qualified interpretation of such evidence: 'even the most optimistic reading indicates that … more vigorous enforcement alone cannot close the gap between U.S. safety conditions and those in other OECD countries.'

In addition to the questionable effectiveness of OSHA inspections, some of the standards promulgated by OSHA have been criticized as excessively stringent. Viscusi (1996) examined OSHA standards using an implicit value of life of $5 million (derived from the compensating wage differential studies) as the standard for an efficient regulation. Four of the five OSHA safety regulations, but only one of the five OSHA health regulations adopted as final rules, had costs per life saved of less than $5 million. This evidence caused Burton and Chelius (1997) to provide a strong critique of the government mandate theory:

> To be sure, cost-benefit analysis of health standards issued under the OSHAct is not legal, and so those standards that fail the cost-benefit test (considering both the lives saved plus injuries and illnesses avoided) do not violated the letter and presumably the purpose of the law. But to the extent that the rationale offered by the government mandate theorists for regulation of health is that workers lack enough information to make correct decisions and therefore the government is in

a better position to make decisions about how to improve workplace health, the evidence on the variability of the cost/benefit ratios for OSHA health standards is disquieting. Rather than OSHA standards reflecting interventions in the marketplace that overcome deficiencies of the marketplace, the explanation of why the stringency of regulation varies so much among industries would appear at best to be a result of technology-based decisions that could well aggravate the alleged misallocation of resources resulting from operation of the market and at worst could reflect relative political power of the workers and employers in various industries.

6. Conclusions

This review suggests that understanding the economics of workplace safety involves a rather eclectic mix of theories. Burton and Chelius (1997) were least impressed with the arguments and evidence pertaining to the pure neoclassical economics and the government mandate theories. They concluded that among the other economic theories pertaining to safety, no single theory provides an adequate understanding of the causes and prevention of workplace accidents. Rather, a combination of the theories, though untidy, is needed.

Bibliography

Ackerman S 1988 Progressive law and economics—and the new administrative law. *Yale Law Journal* **98**: 341–68

Aiuppa T, Trieschmann J 1998 Moral hazard in the French workers' compensation system. *Journal of Risk and Insurance* **65**: 125–33

Boden L 1995 Creating economic incentives: lessons from workers' compensation systems. In: Voos P (ed.) *Proceedings of the 47th Annual Meeting*. Industrial Relations Research Association, Madison, WI

Bruce C J, Atkins F J 1993 Efficiency effects of premium-setting regimes under workers' compensation: Canada and the United States. *Journal of Labor Economics* **11**: Part 2: S38–S69

Burton J, Chelius J 1997 Workplace safety and health regulations: Rationale and results. In: Kaufman B (ed.) *Government Regulation of the Employment Relationship*. 1st edn. Industrial Relations Research Association, Madison, WI

Butler R 1994 Safety incentives in workers' compensation. In: Burton J, Schmidle T (eds.) *1995 Workers' Compensation Year Book*. LRP Publications, Horsham, PA

Chelius J R 1977 *Workplace Safety and Health*. American Enterprise Institute for Public Policy Research, Washington, DC

Coase R H 1988 *The Firm, the Market, and the Law*. University of Chicago Press, Chicago

Commons J R 1934 *Institutional Economics: Its Place in Political Economy*. Macmillan, New York

Commons J R, Andrews J B 1936 *Principles of Labor Legislation*, 4th edn. MacMillan, New York

Dewees D, Duff D, Trebilcock M 1996 *Exploring the Domain of Accident Law: Taking the Facts Seriously*. Oxford University Press, Oxford

Dorman P 1996 *Markets and Mortality: Economics, Dangerous Work, and the Value of Human Life*. Cambridge University Press, Cambridge

Dorman P, Hagstrom P 1998 Wage compensation for dangerous work revisited. *Industrial and Labor Relations Review* **52**: 116–35

Ehrenberg R G 1988 Workers' Compensation, Wages, and the Risk of Injury. In: Burton J (ed.) *New Perspectives on Workers' Compensation*. ILR Press, Ithaca, NY

Ehrenberg R G, Smith R S 1996 *Modern Labor Economics: Theory and Public Policy* 6th edn. Addison-Wesley, Reading, MA

Fishback P V 1987 Liability rules and accident prevention in the workplace: Empirical evidence from the early twentieth century. *Journal of Legal Studies* **16**: 305–28

Hirsch B T, Macpherson D A, Dumond J M 1997 Workers' compensation recipiency in union and nonunion workplaces. *Industrial and Labor Relations Review* **50**: 213–36

Kniesner T, Leeth J 1995 Abolishing OSHA. *Regulation* **4**: 46–56

Landes W M, Posner R A 1987 *The Economic Structure of Tort Law*. Harvard University Press, Cambridge, MA

McGarity T, Shapiro S 1996 OSHA's Critics and Regulatory reform. *Wake Forest Legal Review* **31**: 587–646

Miller P, Mulvey C, Norris K 1997 Compensating differentials for risk of death in Australia. *Economic Record* **73**: 363–72

Moore M J, Viscusi W K 1990 *Compensation Mechanisms for Job Risks: Wages, Workers' Compensation, and Product Liability*. Princeton University Press, Princeton, NJ

Oi W Y 1974 On the economics of industrial safety. *Law and Contemporary Problems* **38**: 669–99

Posner R 1972 A theory of negligence. *Journal of Legal Studies* **1**: 29–66

Priest G L 1991 The modern expansion of tort liability: its sources, its effects, and its reform. *Journal of Economic Perspectives* **5**: 31–50

Reilly B, Paci P, Holl P 1995 Unions, safety committees and workplace injuries. *British Journal of Industrial Relations* **33**: 275–88

Schwartz G T 1994 Reality in the economic analysis of tort law: does tort law really deter? *UCLA Law Review* **42**: 377–444

Scholz J T, Gray W B 1990 OSHA enforcement and workplace injuries: a behavioral approach to risk assessment. *Journal of Risk and Uncertainty* **3**: 283–305

Siebert S, Wie X 1994 Compensating wage differentials for workplace accidents: evidence for union and nonunion workers in the UK. *Journal of Risk and Uncertainty* **9**: 61–76

Smith R 1992 Have OSHA and workers' compensation made the workplace safer? In: Lewin D, Mitchell O, Sherer P (eds.) *Research Frontiers in Industrial Relations and Human Resources*. Industrial Relations Research Association, Madison, WI

Thomason T, Burton J F 1993 Economic effects of workers' compensation in the United States: Private insurance and the administration of compensation claims. *Journal of Labor Economics* **11**: Part 2: S1–S37

Thomason T, Hyatt D, Roberts K 1998 Disputes and dispute resolution. In: Thomason T, Burton J, Hyatt D (eds.) *New Approaches to Disability in the Workplace*. Industrial Relations Research Association, Madison, WI

Viscusi W K 1993 The value of risks to life and health. *Journal of Economic Literature* **31**: 1912–46

Viscusi W K 1996 Economic foundations of the current regulatory reform efforts. *Journal of Economic Perspectives* **10**: 119–34

Weil D 1996 If OSHA is so bad, why is compliance so good? *Rand Journal of Economics* **27**: 618–40

Weil D 1999 Are mandated health and safety committees substitutes for or supplements to labor unions? *Industrial and Labor Relations Review* **52**: 339–60
Willborn S 1988 Individual employment rights and the standard economic objections: theory and empiricism. *Nebraska Legal Review* **67**: 101–39

J. F. Burton, Jr.

Sample Surveys: Cognitive Aspects of Survey Design

Cognitive aspects of survey design refers to a research orientation that began in the early 1980s to integrate methods, techniques, and insights from the cognitive sciences into the continuing effort to render data from sample surveys of human populations more valid and reliable, and to understand systematically the threats to such reliability and validity.

1. Background and History

While a sample survey of human beings can be of good and measurable accuracy only if it is accomplished through probability sampling (*Sample Surveys: The Field* and *Sample Surveys: Methods*), probability sampling is only the first step in assuring that a survey is successful in producing results accurate enough for use in predicting elections, for gauging public opinion, for measuring the incidence of a disease or the amount of acreage planted in a crop, or for any of the other myriad of important uses that depend on survey data. Questions must be worded not only in ways that are unbiased, but that are understandable to respondents and convey the same meaning to respondents as was intended by the survey's author. If the questions refer to the past, they must be presented in ways that help respondents remember the facts accurately and report truthfully. And interviewers must be able to understand respondents' answers correctly to record and categorize them appropriately.

These and other nonsampling issues have been of concern to survey researchers for many decades (e.g., Payne 1951). Researchers were puzzled by the fact that changing the wording slightly in a question, sometimes resulted in a different distribution of answers and sometimes did not; worried that sometimes the context of other questions affected answers to a particular question, and sometimes did not; concerned that sometimes respondents were able to remember things accurately and sometimes were not. In the late 1970s and early 1980s, as survey data came to be used ever more extensively for public policy purposes, concerns for the validity of those data became especially widespread.

These concerns led to a cross-disciplinary research endeavor that would bring the insights and methods of the cognitive sciences, especially cognitive psychology, to bear on the problems raised by surveys, and that would encourage researchers in the cognitive sciences to use surveys as a means of testing, broadening, and generalizing their laboratory-based conclusions. Recent work in the field has incorporated theories and viewpoints from fields not usually classified as the cognitive sciences, notably linguistics, conversational analysis, anthropology, and ethnography.

Described below are some of the practical and theoretical achievements of the movement to date. These include methodological transfers from the cognitive sciences to survey research, the broad establishment of cognitive laboratories in governmental and nongovernmental survey research centers, theoretical formulations that account for many of the mysteries generated by survey responses, and practical applications to the development and administration of on-going surveys. Some references appear at the close describing these achievements more fully and speculating on the future course of the field.

2. Methodological Transfers

Much work in the field is rooted in a classic information processing model applied to the cognitive tasks in a survey interview. The model suggests that the tasks the respondent must accomplish, in rough order in which they must be tackled (though respondents go back and forth among them), are comprehension/interpretation, retrieval, judgment, and communication. These models have been intertwined with the earlier models of the survey interview as a social interaction (Sudman and Bradburn 1974) through the efforts of such researchers as Clark and Schober (1992) and Sudman et al. (1996) to take into account both the social communication and individual cognitive processes that play a part in the survey interview.

Guided by this model, one of the earliest achievements of the movement to study cognitive aspects of survey design was the adoption of some of the methods of the cognitive psychology laboratory for the pretesting of survey questionnaires. In particular, these methods are aimed at insuring that questions are comprehensible to respondents and that the meanings transmitted are those intended by the investigator. They also can point to problems in retrieval and communication. Such cognitive pretesting includes 'think-aloud protocols,' in which specially recruited respondents answer the proposed questionnaire aloud to a researcher, and describe, either concurrently with each question or retrospectively at the end of the questionnaire, their thought processes as they attempt to understand the question and to retrieve or construct an answer. Also used is behavioral coding, based on

procedures originally developed by Cannell and co-workers (e.g., Cannell and Oskenberg 1988), which note questions for which respondents ask for clarification or for which their responses are inadequate. Other methods include reviews of questionnaires by cognitively trained experts, the use of focus groups, and for questionnaires that are automated, measures of response latency. Detailed descriptions of these methods and their applicability to detect problems at various stages of the question answering process appear in Sudman et al. (1996), Schwarz and Sudman (1996), and Forsyth and Lessler (1991).

3. *Interviewer Behavior*

Although survey interviewing began with investigators themselves merely going out and having conversations with respondents, as the survey enterprise aspired to greater scientific respectability and as it became sufficiently large-scale so that many interviewers were needed for any project, standardized interviewing became the norm. Printed questionnaires were to be followed to the letter; interviewers were instructed to answer respondents' requests for clarification by merely re-reading the question or by saying that the meaning of any term is 'whatever it means to you.' While it is probably true that skilled interviewers have always deviated from these rigid standards in the interest of maintaining respondent cooperation and getting valid data, the robot-like interviewer delivering the same 'stimulus' to every respondent was considered the ideal. Working with videotapes of interviews for the National Health Interview Survey and for the General Social Survey and using techniques of conversational analysis, Suchman and Jordan (1990) showed that misunderstandings frequently arose because of the exclusion of normal conversational resources from the survey interview, and that these misunderstandings resulted not only in discomfort on the part of respondents, but in data that did not meet the needs of the survey designer. Suchman and Jordan recommended a more collaborative interviewing style, in which interviewer and respondent work together to elicit the respondent's story and then fill in the questionnaire. Note that this recommendation applies almost exclusively when the information sought is about the respondent's activities or experiences rather than about his/her attitudes or opinions. Both theoretical and practical advances have followed on these insights.

On a theoretical level, several researchers have used variants of the 'cooperativeness' principle advanced by Grice (1975) to help explain some of the anomalies that have puzzled survey researchers over the decades. Many are understandable as respondents bringing to the survey interview the communicative assumptions and skills that serve well in everyday life. The maxims of the cooperativeness principle require a participant in a conversation to be truthful, relevant, informative, and clear (Sudman et al. 1996, p. 63). Thus, for example, it has long been known that respondents are willing to report on their attitudes towards fictitious issues or nonexistent groups, and investigators have taken this willingness as evidence of the gullibility of respondents and perhaps of the futility of trying to measure attitudes at all. But to 'catch on' to the 'trick' nature of such a question, a respondent would have to assume that the interviewer violated all the maxims of the cooperativeness principle. Parsimony suggests, instead, that the respondent would assume that the interviewer knew what s/he was talking about, and answer accordingly. Several puzzles about context effects, discussed below, also seem understandable in light of Grice's maxims.

On a practical level, freeing interviewers to be more conversational with respondents is not costless; when interviewers explain more, interviews take longer, and are thus more costly in interviewers' wages and respondents' patience. Schober and Conrad (e.g., 1997) have embarked on a program of research that has demonstrated that conversational interviewing generates better data than does standardized interviewing when the mappings of the respondent's experiences onto the survey's concepts are complicated, but indeed at the cost of longer interviews. These investigators are now experimenting with interviewing styles that fall at several points along the continuum between fully standardized and fully conversational, seeking a point that balances the benefits in accuracy to be gained by a more conversational style with the costs in interview length.

4. *Factual Questions vs. Questions about Attitudes or Opinions*

The distinction between questions that ask respondents to report on facts, usually autobiographical, and those that ask for an attitude or an opinion has been an important one in survey research. Indeed, the explicit origins of the movement to study cognitive aspects of survey design lie in a desire to improve the accuracy of factual reports from respondents; the extension of concern to attitudinal questions came somewhat later, although such a concern was certainly prefigured by the work reported in Turner and Martin (1984). The distinction remains important for interviewer behavior, in the sense that the kinds of conversational interviewing being investigated, are designed to help respondents better comprehend the intent of the question and to recall more accurately their experiences. Interviewer interventions designed for these purposes, when the question is aimed at getting a factual report might well bias the respondent's answer when the aim of the question is to elicit an attitude or opinion. The distinction is also valid in the sense that the cognitive theories about

autobiographical memory that have been used to understand problems of the recall of factual information are less applicable to attitudes and opinions, and the cognitive techniques that have been developed to aid respondent recall for autobiographical questions are not appropriate for the recall of attitudes or opinions. But in important ways the distinction obscures similarities between the two types of questions, as will be discussed in the section on context effects, below.

4.1 Improving the Accuracy of Factual Reports

Very often survey questions ask for the frequency with which a respondent has accomplished a behavior (e.g., visited a doctor) or experienced an event (e.g., had an illness) during a specified time period (usually starting sometime in the past and ending with the date of the interview) called the reference period. Thus the respondent, having understood what events or experiences that should be included in the report, has a two-fold task: he or she must recall the events or experiences, and also determine whether those events or experiences fell inside the reference period. Theories from the cognitive sciences about how experiences are stored in autobiographical memory and retrieved therefrom have been marshaled to understand both phases of the respondent's recall task and to improve their accuracy.

In a phenomenon called 'telescoping,' respondents are known often to move events or experiences forward in time, reporting them as falling within the reference period although in fact they occurred earlier. One of the earliest contributions to the literature on cognitive aspects of survey design was the proposal of a technique dubbed 'landmarking' to control telescoping (Loftus and Marburger 1983). Here, rather than inquiring about the last six months or the last year, a respondent is supplied with (or is asked to supply) a memorable date, and then is asked about events or experiences since that date. The related concept of bounding is also useful in controlling telescoping. In a panel survey, in which respondents are interviewed repeatedly at regular intervals, informing them of the events or experiences reported at the last interview will prevent such events or experiences from being placed in the time period between the previous interview and the current one. An analogous technique useable in a single interview is the two-time-frame procedure in which the respondent is first asked the report events or experiences during a long reference period (e.g., the last 6 months) and then for a shorter one that is really of interest to the investigator (e.g., the last 2 months). This technique both seems to relieve the pressure on respondents to report socially desirable events or experiences and to stress the interviewer's interest in correct dating, encouraging respondents to strive for greater accuracy. A theoretical explanation of telescoping, taking into account the workings of memory for the storage of timing and the effects of rounding to culturally stereotypical values (e.g., 7 days, 10 days, 30 days), is provided by Hutenlocher et al. (1990).

The task of counting or estimating the number of events or experiences in the reference period can also provide a challenge to the respondent. Issues of frequency regularity and similarity influence whether the respondent is likely to actually count the instances or to employ an estimation strategy, as well as influencing whether counting or estimation is likely to be more accurate. In general, respondents are more likely to retrieve from memory and count events or experiences if their number is few and to estimate if their number is large. But regular events (e.g., going to the dentist every 6 months) and similar events (e.g., trips to the grocery store) are more likely to be estimated. And investigators have recently found that estimation (perhaps by retrieving a rate and applying that rate to the length of the reference period) is usually more accurate, than the attempt to retrieve and count regular, similar events or experiences. For a full discussion of these issues, see Sudman et al. (1996).

4.2 Context Effects

Context effects occur when the content of the questions preceding a particular question (or even following that question in a self-administered questionnaire in which the respondent is free to return to earlier questions after reading later ones) influence how some respondents answer that question. Our understanding of the term has been broadened recently to include also the effects on the distribution of responses of the response categories offered in a closed-ended question. Investigators have used the maxims of Grice's cooperativeness principle to approach a systematic understanding of these effects, which have seemed mysterious for many years. Full treatments of these issues may be found in Sudman et al. (1996) and Tourangeau (1999); of necessity only some selected findings can be presented here.

A pair of puzzling context effects that have fascinated researchers for years are assimilation effects and contrast effects. An assimilation effect occurs when the content of the preceding items moves the responses to the target item in the direction of the preceding items; a contrast effect when the movement is in the opposite direction. These effects can now be understood through the inclusion/exclusion model of the judgment process in survey responding, proposed by Schwarz and Bless (1992). The model holds that when a respondent has to make a judgment about a target stimulus, s/he must construct a cognitive representation, not only of the target stimulus, but also of a standard with which to compare the target

stimulus. To construct these representations respondents must retrieve information from memory; since not all information in memory can be retrieved, respondents retrieve only that which is most easily accessible. And information can be accessible either because it is 'chronically' accessible or because it is temporarily accessible; it is the temporary accessibility of information that is affected by the context of the question. Positive temporarily accessible information supplied by the context and included in the representation of the target will render the judgment of the target more positive: negative context information thus included will render the judgment more negative. Both these processes produce assimilation effects. Contrast effects are created when the context suggests that information be excluded from the representation of the target, or included in the representation of the standard to which the target is to be compared. For example, Schwarz and Bless (1992) were able to manipulate the frequency of positive evaluations of a political party by using a reference to a popular politician in a preceding question. They were able to create an assimilation effect, resulting in more positive evaluations, by encouraging respondents to include the politician in their representations of the party (by having them recall that the politician had been a member of the party). They were also able to create a contrast effect, resulting in less positive evaluations of the political party, by encouraging respondents to exclude the politician from their representation of the party (by having them recall that his present position took him out of party politics). The model predicts inclusion of more information when the target is a general question, thus encouraging assimilation effects; it also predicts contrast effects when the target is narrow. Thus, a preceding question about marital happiness seems to result in higher reports of general happiness for those who are happily married, but lower reports of general happiness for those who are unhappy in their marriages. This influence of the preceding question can be eliminated by having respondents see the questions about marital happiness and general happiness as part of a single unit (either by an introduction that links the two questions or by a typographical link on a printed questionnaire); then, in obedience to the conversational maxim that speakers should present new information, respondents exclude their marital happiness from their evaluation of their general happiness.

Similar recourse to conversational maxims helps explain the influence of the response alternatives presented in closed answer questions on the distribution of responses. For example, Schwarz et al. (1985) found that 37.5 percent of respondents reported watching TV for more than two and a half hours a day when the response scale ranged from 'up to $2\frac{1}{2}$ hours' to 'more than $4\frac{1}{2}$ hours,' but only 16.2 percent reported watching more than two and a half hours a day when the response scale ranged from 'up to $\frac{1}{2}$ hour' to 'more than $2\frac{1}{2}$ hours.' Respondents take the response scales to be conversationally informative about what typical TV watching habits are and calibrate their report accordingly.

5. *Some Achievements of the Movement to Study Cognitive Aspects of Surveys*

5.1 *The Impact of Theories*

Perhaps the most profound effect of the movement to study cognitive aspects of survey design is the introduction of theory into the research area. There has been space above to present only a few of the many such theoretical advances, but their utility should be clear. For example, consulting the maxims of cooperativeness in conversation gives us a framework for predicting and preventing response effects. In the same sense, the inclusion/exclusion model, much more fully developed than is presented here, makes predictions for what to expect from various manipulations of question ordering and content. These predictions offer guidance for the experimental advance of the research domain, and their testing systematically advances our knowledge, both of the phenomena and of the applicability of the theory.

5.2 *The Cognitive Laboratories and Their Impact*

Cognitive interviews in laboratory settings have become standard practice in several US government agencies (including the National Center for Health Statistics, the Bureau of Labor Statistics, and the Census Bureau), as well as in government agencies around the world, and at many academic and commercial survey organizations. Work in these laboratories supplements the usual large-scale field pretests of surveys. For illustrative purposes it is worth describing two major efforts at US government agencies.

The Current Population Survey (CPS) is a large-scale government survey, sponsored by the Bureau of Labor Statistics and carried out by the Census Bureau. From interviews with some 50,000 households monthly are derived major government statistical series, including the unemployment rate. In the early 1990s the CPS underwent its decennial revision, this time using the facilities of the government cognitive laboratories as well as traditional field testing of proposed innovations. In particular, this research found that the ordering and wording of the key questions regarding labor force participation caused some respondents to report themselves as out of the labor force when in fact they were either working part time or looking for work. It also found that the concept of layoff, which the CPS takes to mean

temporary furlough from work with a specific date set for return, was not interpreted that way by respondents. Respondents seemed to understand 'on layoff' as a polite term for 'fired.' Appropriate revision of these questions, followed by careful field testing and running the new and old versions in parallel for some months resulted in estimates of the unemployment rate in which policy makers and administrators can have greater confidence. More details can be found in Norwood and Tanur (1994).

The US Decennial Census has always asked for a racial characterization of each resident. Originally designed to distinguish whites (mostly free) from blacks (mostly slaves) for purposes of counting for apportionment (the US Constitution originally mandated that a slave be counted as 3/5 of a man), the uses to which answers to the race identification question was put evolved over time. Civil Rights legislation starting in the 1960s designated protected groups, and it thus became important to know the population of such groups. At the same time, individuals were becoming both more conscious of their racial self-identification and less willing to be pigeonholed into a small set of discrete categories, especially as more individuals were self-identifying with more than one racial group. The 1990 Census asked two questions in this general area—one on racial identity followed by one on Hispanic ethnic identification. Experimental work (Martin et al. 1990) indicated that the ordering of these questions mattered to respondents. For example, when the Hispanic ethnicity question appeared first, fewer people chose 'other' as their racial category. This and other research made a convincing case that the complexity of Americans' racial and ethnic self-identification needed a more complex set of choices on the Census form. A 4-year program of research resulted in the conclusion that allowing respondents to choose as many racial categories as they believe apply to them is a solution superior to the provision of a multiracial category (see Tucker et al. 1996). (For more details on this and other census research see *Censuses: History and Methods*.) The 2000 Decennial Census allowed respondents to choose as many racial categories as they wished; at this writing it is not yet known what proportion of the population took advantage of that opportunity, or the effects those choices will have on the analyses of the data.

6. Conclusions

The movement to study cognitive aspects of surveys has thus given us some theoretical insights and practical approaches. Ideally, we could hope for a theory of the survey designing and responding process, but we are still far from that goal, and may perhaps never be able to attain it. But the frameworks so far provided offer clear guidance of how to proceed in an effort to reach systematic understanding. Far fuller accounts of the results of the movement to study cognitive aspects of survey design and ideas of future directions can be found in Sirken et al. (1999a), Sirken et al. (1999b), Sudman et al. (1996), and Tanur (1992).

See also: Probability: Formal; Probability: Interpretations; Sample Surveys, History of; Sample Surveys: Methods; Sample Surveys: Model-based Approaches; Sample Surveys: Nonprobability Sampling; Sample Surveys: Survey Design Issues and Strategies; Sample Surveys: The Field

Bibliography

Cannel C F, Oskenberg L 1988 Observation of behavior in telephone interviews. In: Groves R M, Biemer P B, Lyberg L E, Massey J T, Nichols II W L, Waksberg J (eds.) *Telephone Survey Methodology*. Wiley, New York

Clark H H, Schober M F 1992 Asking questions and influencing answers. In: Tanur J. (ed.) *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*. Sage, New York

Forsyth B H, Lessler J 1991 Cognitive laboratory methods: A taxonomy. In: Biemer P, Groves R M, Lyberg L E, Mathiowetz N, Sudman S (eds.) *Measurement Errors in Surveys*. Wiley, New York

Grice H P 1975 Logic and conversation. In: Cole P, Morgan J L (eds.) *Syntax and Semantics, Vol 3: Speech Acts*. Academic Press, New York

Hippler H J, Schwarz N, Sudman S (eds.) 1987 *Social Information Processing and Survey Methodology*. Springer-Verlag, New York

Huttenlocher J, Hedges L V, Bradburn N M 1990 Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**: 196–213

Loftus E F, Marburger W 1983 Since the eruption of Mt. St. Helens, did anyone beat you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition* **11**: 114–20

Martin E, Demaio T J, Campanelli P C 1990 Context effects for census measures of race and Hispanic origin. *Public Opinion Quarterly* **54**: 551–66

Norwood J L, Tanur J M 1994 Measuring unemployment in the nineties. *Public Opinion Quarterly* **58**: 277–94

Payne S L 1951 *The Art of Asking Questions*. Princeton University Press, Princeton, NJ

Schober M F, Conrad F G 1997 Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly* **61**: 576–602

Schwarz N, Bless H 1992 Constructing reality and its alternatives: Assimilation and contrast effects in social judgment. In: Martin L L, Tesser A. (eds.) *The Construction of Social Judgment*. Erlbaum, Hillsdale, NJ

Schwarz N, Hippler H J, Deutsch B, Strack F 1985 Response categories: Effects on behavioral reports and comparative judgments. *Public Opinion Quarterly* **49**: 388–95

Schwarz N, Sudman S (eds.) 1996 *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. Jossey-Bass, San Francisco

Sinaiko H, Broedling L A (eds.) 1976 *Perspectives on Attitude Assessment Surveys and their Alternatives*. Pendleton, Champaign, IL

Sirken M G, Herrmann D J, Schechter S, Schwarz N, Tanur J M, Tourangeau R (eds.) 1999a *Cognition and Survey Research*. Wiley, New York

Sirken M G, Jabine T, Willis G, Martin E, Tucker C (eds.) 1999b *A New Agenda for Interdisciplinary Survey Research Methods: Proceedings of the CASM II Seminar*. US Department of Healthy and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD

Suchman L, Jordan B 1990 Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association* **85**: 232–53

Sudman S, Bradburn N M 1974 *Response Effects in Surveys: A Review and Synthesis*. Aldine, Chicago

Sudman S, Bradburn N M, Schwarz N 1996 *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco

Tanur J M (ed.) 1992 *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*. Sage, New York

Tourangeau R 1999 Context effects on answers to attitude questions. In: Sirken M G, Hermann D J, Schwarz N, Tanur J M, Tourangeau R (eds.) *Cognition and Survey Research*. Wiley, New York

Tucker C, McKay R, Kojetin B, Harrison R, de al Puente M, Stinson L, Robison E 1996 Testing methods of collecting racial and ethnic information: Results of the Current Population Survey Supplement on Race and Ethnicity. Bureau of Labor Statistics, Statistical Notes, No. 40

Turner C F, Martin E (eds.) 1984 *Surveying Subjective Phenomena*. Sage, New York

J. M. Tanur

Sample Surveys, History of

Sampling methods have their roots in attempts to measure the characteristics of a nation's population by using only a part instead of the whole population. The movement from the exclusive use of censuses to the at least occasional use of samples was slow and laborious. Two intertwined intellectual puzzles had to be solved before the move was complete. The first such puzzle was whether it is possible to derive valid information about a population by examining only a portion thereof (the 'representative' method). And the second puzzle concerns the method for choosing that portion. Issues of choice themselves fall into two categories: the structuring of the population itself in order to improve the accuracy of the sample, and whether to select the units for inclusion by purposive or random methods.

This article considers the development of methods for the sampling of human populations. We begin by describing two antecedents of modern survey methods—the development of government statistics and censuses and the general development of statistical methodology, especially for ratio estimation. Then we deal with the movement from censuses to sample surveys, examining the debate over representative methods and the rise of probability sampling, including the seminal 1934 paper of Jerzy Neyman. We describe the impact in the United States of the results obtained by Neyman on sample surveys and the contributions of statisticians working in US government statistical agencies. We end with a brief description of the subsequent establishment of random sampling and survey methodology more broadly and examine some history of the movement from the study of the design of sampling schemes to the study of other issues arising in the surveying of human populations.

1. *Antecedents of Modern Sample Survey Methods*

In order to determine facts about their subjects or citizens, and more often about their number, governments have long conducted censuses. Thus, one set of roots of sample survey methodology is intertwined with the history of methods for census-taking. The origins of the modern census are found in biblical censuses described in the Old Testament (e.g., see the discussions in Duncan (1984) and Madansky (1986) as well as in censuses carried out by the ancient Egyptians, Greeks, Japanese, Persians, and Romans (Taeuber (1978)). For most practical purposes we can skip from biblical times to the end of the eighteenth century and the initiation of census activities in the United States of America, even though there is some debate as to whether Canada, Sweden, or the United States should be credited with originating the modern census (Willcox 1930) (see *Censuses: History and Methods*).

Another antecedent of sampling lies in the problem of estimating the size of a population when it is difficult or impossible to conduct a census or complete enumeration. Attempts to solve this problem even in the absence of formal sampling methods were the inspiration of what we now know as ratio estimation. These ideas emerged as early as the seventeenth century. For example, John Graunt used the technique to estimate the population of England from the number of births and some assumptions about the birth rate and the size of families. Graunt's research, later dubbed political arithmetic, was based on administrative records (especially parish records) and personal observation, as in his seminal work, *Natural and Political Observations Made Upon the Bills of Mortality*, published in 1662. To accomplish what Graunt did by assumption, others looked to hard data from subsets of a population. Thus, Sir Frederick Morton Eden, for example, found the average number of people per house in selected districts of Great Britain in 1800. Using the total number of households in Great Britain from tax rolls (and allowing for those houses missing from the rolls), Eden estimated the population of Great Britain as nine million, a figure confirmed by the first British census of 1801 (Stephan (1948).

Even earlier, in 1765 and 1778, civil servants published population estimates for France using an

enumeration of the population in selected districts and counts of births, deaths, and marriages, for the country as a whole (see Stephan (1948). It was Pierre Simon de Laplace, however, who first formally described the method of ratio estimation during the 1780s, and he then employed it in connection with a sample survey he initiated to estimate the population of France as of September 22, 1802. He arranged for the government to take a sample of administrative units (communes), in which the total population, y, and the number of registered births in the preceding year, x, were measured. Laplace then estimated the total population of France by $\hat{Y} = Xy/x$, where X was the total number of registered births. Laplace was the first to estimate the asymptotic bias and variance of $\hat{Y}$, through the use of a superpopulation model for the ratio $p = y/x$ (see Laplace (1814)). Cochran (1978) describes this derivation and links it to results which followed 150 years later. Quetelet later applied the ratio estimation approach in Belgium in the 1820s but abandoned it following an attack by others on some of the underlying assumptions (see Stigler 1986, pp. 162–6).

2. *From Censuses to Surveys*

The move from censuses to sample surveys to measure the characteristics of a nation's population was slow and laborious. Bellhouse (1988), Kruskal and Mosteller (1980), and Seng (1951) trace some of this movement, especially as it was reflected in the discussions regarding surveys that took place at the congresses of the International Statistical Institute (ISI). The move succeeded when investigators combined random selection methods with the structuring of the population and developed a theory for relating estimates from samples, however structured, to the population itself.

As early as the 1895 ISI meeting, Kiaer (1896) argued for a 'representative method' or 'partial investigation,' in which the investigator would first choose districts, cities, etc., and then units (individuals) within those primary choices. The choosing at each level was to be done purposively, with an eye to the inclusion of all types of units. That coverage tenet, together with the large sample sizes recommended at all levels of sampling, was what was judged to make the selection representative. Thus the sample was, approximately, a 'miniature' of the population. Kiaer used systematic sampling in his 1895 survey of workers as a means of facilitating special tabulations from census schedules for a study of family data in the 1900 Norwegian census. Kiaer actually introduced the notion of random selection in his description of this work by noting that at the lowest level of the sample structure 'the sample should be selected in a haphazard and random way, so that a sample selected in this manner would turn out in the same way as would have been the case had the sample been selected through the drawing of lots … ' (Kiaer 1897, p. 39)

The idea of less than a complete enumeration was opposed widely, and Kiaer presented arguments for sampling at ISI meetings in 1897, 1901, and 1903. Lucien March, in a discussion of Kiaer's paper at the 1903 meeting, formally introduced the concepts of simple random sampling without replacement and simple cluster sampling, although not using these names (Bellhouse 1988).

Arthur Lyon Bowley did a number of empirical studies on the validity of random sampling, motivated at least in part by a 1912 paper by Franicis Ysidro Edgeworth (1912). Bowley used large sample normal approximations (but did not conceptualize what he was doing as sampling from a finite population) to actually test the notion of representativeness. Then, he carried out a 1912 sample survey on poverty in Reading, in which he drew the respondents at random (see Bowley 1913), although he appears to have equated simple random sampling with systematic sampling (Bellhouse 1988).

At about the same time, in what appears to be an independent line of discovery, Tchouproff and others were overseeing the implementation of sample survey methods in Russia, especially during the First World War and the period immediately following it. This work was partially documented after the Russian Revolution in *Vestnik Statistiki*, the official publication of the Central Statistical Administration (Zarkovic 1956, 1962). Even though there is clear evidence that Tchouproff had developed a good deal of the statistical theory for random sampling from finite populations during this period (see Seneta 1985), whether he actually implemented forms of random selection in these early surveys is unclear. He later published the formulae for the behavior of sample estimates under simple random sampling and stratified random sampling from finite populations in Tchouproff (1918a, 1918b, 1923a, 1923b).

In a seemingly independent development of these basic ideas for sampling in the context of his work on agricultural experiments, Neyman in Splawa-Neyman (1923) described them in the form of the drawing of balls without replacement from an urn. In the resulting 1925 paper, Splawa-Neyman (1925) gave the basic elements of the theory for sampling from finite populations and its relationship with sampling from infinte populations. These results clearly overlapped those of Tchouproff, and two years later following his death, his partisans took Neyman to task for his lack of citation to the Russian work as well as to earlier work of others (see Fienberg and Tanur 1995, 1996) for a discussion of the controversy).

By 1925, the record of the ISI suggests that the representative method was taken for granted, and the discussions centered around how to accomplish representativeness and how to measure the precision of sample-based estimates, with the key presentations

being made by Bowley (1926) and in the same year by the Danish statistician Adolph Jensen. Notions of clustering and stratification were put forward, and Bowley presented a theory of proportionate stratified sampling as well as the concept of a frame, but purposive sampling was still the method of choice. It was not until Gini and Galvani made a purposive choice of which returns of an Italian census to preserve, and found that districts chosen to represent the country's average on seven variables were, in that sense, unrepresentative on other variables, that purposive sampling was definitively discredited (Gini 1928, Gini and Galvani 1929).

3. *Neyman's 1934 Paper on Sampling*

In their work on the Italian census, Gini and Galvani seemed to call into question the accuracy of sampling. Neyman took up their challenge in his classic 1934 paper presented before the Royal Statistical Society, 'On the two different aspects of the representative method.' In it he compared purposive and random sampling, and concluded that it wasn't sampling that was problematic but rather Gini and Galvani's purposive selection.

Elements of synthesis were prominent in the paper as well. Neyman explicitly uncoupled clustering and purposive sampling, saying, 'In fact the circumstance that the elements of sampling are not human individuals, but groups of these individuals, does not necessarily involve a negation of the randomness of the sampling' (1952, p. 571). He calls this procedure 'random sampling by groups' and points out that, although Bowley did not consider it theoretically, he used it in practice in London, as did O. Anderson in Bulgaria. Neyman also combined stratification with clustering to form 'random stratified sampling by groups,' and he provided a method for deciding how best to allocate samples across strata (optimal allocation). The immediate effect of Neyman's paper was to establish the primacy of the method of stratified random sampling over the method of purposive selection, something that was left in doubt by the 1925 ISI presentations by Jensen and Bowley. But the paper's longer-term importance for sampling was the consequence of Neyman's wisdom in rescuing clustering from the clutches of those who were the advocates of purposive sampling and integrating it with stratification in a synthesis that laid the groundwork for modern-day multistage probability sampling.

Surprisingly, for many statisticians the memorable parts of Neyman's paper were not these innovations in sampling methodology but Neyman's introduction of general statistical theory for point and interval estimation, especially the method of confidence intervals (see *Estimation: Point and Interval* and *Frequentist Inference*).

As pathbreaking as Neyman's paper was, a number of its results had appeared in an earlier work by Tschuprow (1923a 1923b), in particular the result on optimal allocation in stratified sampling (see Fienberg and Tanur 1995, 1996 for discussions of this point). The method was derived much earlier, by the Danish mathematician Gram (1883) in a paper dealing with calculations for the cover of a forest based on a sample of trees. Gram's work has only recently been rediscovered and seems not to have been accessible to either Tchouproff or Neyman.

Neyman provided the recipe for others to follow and he continued to explain its use in convincing detail to those who were eager to make random sampling a standard diet for practical consumption (e.g., see Neyman 1952 for a description based on his 1937 lectures on the topic at the US Department of Agriculture Graduate School, as discussed below).

4. *The Development of Random Sampling in the United States*

One might have thought that the resolution of the controversy over the representative method and the articulation of the basic elements of a theory of sample surveys by Neyman would have triggered extensive application of random sampling throughout the world. Surprisingly, this was not the case. With some notable exceptions, for example, in England (see Cochran 1939 and Yates 1946) and India, the primary application occurred in the United States and this led to a spate of new and important methodological developments.

As late as 1932, however, there were few examples of probability sampling anywhere in the US federal government (Duncan and Shelton (1978) and the federal statistical agencies had difficulty responding to the demand for statistics to monitor the effects of the programs of President Franklin Roosevelt's New Deal. In 1933, the American Statistical Association (ASA) set up an advisory committee that grew into the Committee on Government Statistics and Information Services (COGSIS), sponsored jointly by ASA and the Social Science Research Council. COGSIS helped to stimulate the use of probability sampling methods in various parts of the Federal government, and it encouraged employees of statistical agencies to carry out research on sampling theory. For example, to establish a technical basis for unemployment estimates, COGSIS, and the Central Statistical Board which it helped to establish, organized an experimental Trial Census of Unemployment as a Civil Works Administration project in three cities, using probability sampling, and carried out in late 1933 and early 1934. The positive results from this study led in 1940 to the establishment of the first large-scale, ongoing sample survey on employment and unemployment using probability sampling methods. This survey later

became known as the Current Population Survey and it continues to the present day.

Another somewhat indirect outcome of the COGSIS emphasis on probability sampling took place at the Department of Agriculture Graduate School where Deming, who recognized the importance of Neyman's 1934 paper, invited Neyman to present a series of lectures in 1937 on sampling and other statistical methods (Neyman (1952). These lectures had a profound impact on the further development of sampling theory not simply in agriculture, but across the government as well as in universities.

Among those who worked on the probability-sampling-based trial Census of Unemployment at the Bureau of the Census was Hansen, who was then assigned with a few others to explore the field of sampling for other possible uses at the Bureau, and went on to work on the 1937 sample Unemployment Census. After working on the sample component of the 1940 decennial census (under the direction of Deming), Hansen worked with others to redesign the unemployment survey based on new ideas on multistage probability samples and cluster sampling (Hansen and Hurwitz 1942, 1943). They expanded and applied their approach in various Bureau surveys, often in collaboration and interaction with others, and this effort culminated in 1953 with the publication of a two-volume compendium of theory and methodology (Hansen et al. 1953a, 1953b).

Cochran's (1939) paper, written in England and independently of the US developments, is especially notable because of its use of the analysis of variance in sampling settings and the introduction of superpopulation and modeling approaches to the analysis of survey data. In the 1940s, as results from these two separate schools appeared in various statistical journals, we see some convergence of ideas and results. The theory of estimation in samples with unequal probabilities of selection also emerged around this time (see Horvitz and Thompson 1952, Hansen et al. 1985) (see *Sample Surveys: Methods*).

Statisticians have continued to develop the theoretical basis of alternative methods of probability sampling and statistical inference from sampling data over the past fifty years (see, e.g., Rao and Bellhouse 1990, Särndal et al. 1992). The issue of statistical inference for models from survey data remains controversial, at least for those trained from a traditional finite sampling perspective.

5. *Market Research and Polling*

The 1930s, and especially the period after World War II, however, saw a flowering of survey methodology in market research and polling, as well as in the social sciences more broadly. Initial stimulation came from a number of committees at the Social Science Research Council and social scientists such as Hadley Cantril, Paul Lazarsfeld, Rensis Likert, William Ogburn, and Samuel Stouffer (e.g., see Stephan 1948, Converse 1987 for a discussion). Stouffer actually spent several months in England in 1931–32, and learned about sampling and other statistical ideas from Bowley, Karl and Egon Pearson, and R. A. Fisher. He then participated in the Census of Unemployment at the Bureau of the Census.

Market research and polling trace their own prehistory to election straw votes collected by newspapers, dating back at least to the beginning of the nineteenth century. Converse (1987) points out, however, a more serious journalistic base; election polls were taken and published by such reputable magazines as the *Literary Digest* (which had gained a reputation for accuracy before the 1936 fiasco). Then, as now, election forecasting was taken as the acid test of survey validity. A reputation for accuracy in 'calling' elections was thought to spill over to a presumption of accuracy in other, less verifiable areas.

There was a parallel tradition in market research, dating back to just before the turn of the twentieth century, attempting to measure consumers' product preferences and the effectiveness of advertising. It was seen as only a short step from measuring the opinions of potential consumers about products to measuring the opinions of the general public about other objects, either material or conceptual. By the mid-1930s there were several well-established market research firms. Many of them conducted election polls in 1936 and achieved much greater accuracy than did the *Literary Digest*. It was the principals of these firms (e.g., Archibald Crossley, George Gallup, and Elmo Roper) who put polling—election, public opinion, and consumer—on the map in the immediate pre-World War II period.

The polling and market research surveys of Crossley, Gallup, Roper, and others were based on a sampling method involving 'quota controls,' but did not involve random sampling. Stephan (1948) observed the close link between their work and the method of purposive sampling that had gained currency much earlier in government and academic research circles, but which by the 1930s had been supplanted by random sampling techniques.

6. *From Sampling Theory to the Study of Nonsampling Error*

The 1940s saw a rapid spread of probability sampling methods to a broad array of government agencies. It was, however, only after the fiasco of the 1948 presidential pre-election poll predictions (Mosteller et al. 1949) that market research firms and others shifted towards probability sampling. Even today many organizations use a version of probability sampling with quotas (Sudman 1967).

Amidst the flurry of activity on the theory and practice of probability sampling during the 1940s, attention was being focused on issues of nonresponse and other forms of nonsampling error (e.g., see Deming 1944) such as difficulty in understanding questions, or remembering answers, etc.). A milestone in this effort to understand and model nonresponse errors was the development of an integrated model for sampling and nonsampling error in censuses and surveys, in connection with planning for and evaluation of the 1950 census (Hansen et al. (1951). This analysis-of-variance-like model, or variants of it, has served as the basis of much of the work on nonsampling error since (see *Linear Hypothesis*; *Nonsampling Errors*).

New developments in sampling since the mid-twentieth century have to do less with the design of samples and more to do with the structuring of the survey instrument and interview. Thus, there has been a move from face-to-face interviewing to telephone interviewing (with the attendant problems connected with random-digit dialing), and then to the use of computers to assist in interviewing. Nonsampling errors continue to be studied broadly, often under the rubric of cognitive aspects of survey design (see *Nonsampling Errors*; *Sample Surveys: Cognitive Aspects of Survey Design*).

See also: Censuses: History and Methods; Government Statistics; Sample Surveys: Methods; Sample Surveys: Survey Design Issues and Strategies; Sample Surveys: The Field; Social Survey, History of; Statistical Methods, History of: Post-1900; Statistical Systems: Censuses of Population

Bibliography

Bellhouse D 1988 A brief history of random sampling methods. In: Krishnaiah P, Rao C (eds.) *Handbook of Statistics*, Vol. 6. North Holland, Amsterdam, pp. 1–14

Bowley A 1913 Working-class households in Reading. *Journal of the Royal Statistical Society* **76**: 672–701

Bowley A 1926 Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute* **22**(1): 6–62

Cochran W 1939 The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association* **128**: 124–35

Cochran W G 1978 Laplace's ratio estimator. In: David H (ed.) *Contributions to Survey Sampling and Applied Statistics: Papers in Honor of H. O. Hartley*. Academic Press, New York, pp. 3–10

Converse J 1987 *Survey Research in the United States: Roots and Emergence 1890–1960*. University of California Press, Berkeley, CA

Deming W 1944 On errors in surveys. *American Sociological Review* **19**: 359–69

Duncan J, Shelton W 1978 *Revolution in United States Government Statistics, 1926–1976*. US Government Printing Office, Washington, DC

Duncan O 1984 *Notes on Social Measurement*. Sage, New York

Edgeworth F Y 1912 On the use of the theory of probabilities in statistics relating to society. *Journal of the Royal Statistical Society* **76**: 165–93

Fienberg S, Tanur J 1995 Reconsidering Neyman on experimentation and sampling: Controversies and fundamental contributions. *Probability and Mathematical Statistics* **15**: 47–60

Fienberg S, Tanur J 1996 Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review* **64**: 237–53

Gini C 1928 Une application de la méthode représentative aux matériaux du dernier recensement de la population italienne (1er décembre 1921). *Bulletin of the International Statistical Institute* **23**(2): 198–215

Gini C, Galvani L 1929 Di una applicazione del metodo rappresentative all'ultimo censimento italiano della popolazione (1° Decembri 1921). *Annali di Statistica* **6**(4): 1–107

Gram J 1883 About calculation of the mass of a forest cover by means of test trees (in Danish). *Tidsskrft for Skovbrug* **6**: 137–98

Hansen M, Dalenius T, Tepping B 1985 The development of sample surveys of finite populations. In: Atkinson A, Fienberg S (eds.) *A Celebration of Statistics: The ISI Centenary Volume*. Springer Verlag, New York, pp. 327–54

Hansen M, Hurwitz W 1942 Relative efficiencies of various sampling units in population inquiries. *Journal of the American Statistical Association* **37**: 89–94

Hansen M, Hurwitz W 1943 On the theory of sampling from finite populations. *Annals of Mathematical Statistics* **14**: 333–62

Hansen M, Hurwitz W, Madow W 1953a *Sample Survey Methods and Theory*, Vol. 1. Wiley, New York

Hansen M, Hurwitz W, Madow W 1953b *Sample Survey Methods and Theory*, Vol. 2. Wiley, New York

Hansen M, Hurwitz W, Marks E, Mauldin W 1951 Response errors in surveys. *Journal of the American Statistical Association* **46**: 147–90

Horvitz D, Thompson D 1952 A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**: 663–85

Jenson A 1926 Report on the representative method in statistics. *Bulletin of the International Statistical Institute* **22**: 359–80

Kiaer A 1895/1896 Observation et expériences concernant des dénombrements répresentatifs. *Bulletin of the International Statistical Institute* **9**(2): 176–83

Kiaer A 1897 The representative method of statistical surveys (in Norwegian). *Christiana Videnskabsselskabets Skrifter. II Hisoriskfilosofiske* **4** [reprinted with an English trans. by Central Bureau of Statistics of Norway, Oslo, 1976)]

Kruskal W, Mosteller F 1980 Representative sampling, iv: The history of the concept in statistics, 1895–1939. *International Statistical Review* **48**: 169–95

Laplace P 1814 *Essai philosophique sur les probabilités*. Dover, New York [trans. of 1840 6th edn. appeared in 1951 as *A Philosophical Essay on Probabilities*, trans. Truscott F W, Emory F L]

Madansky A 1986 On biblical censuses. *Journal of Official Statistics* **2**: 561–69

Mahalanobis P 1944 On large-scale sample surveys. *Philosophical Transactions of the Royal Society of London* **231**(B): 329–451

Mahalanobis P 1946 Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society* **109**: 325–78

Mosteller F, Hyman H, McCarthy P J, Marks E S, Truman D B 1949 *The Pre-election Polls of 1948*. Social Science Research Council, New York
Neyman J 1934 On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**: 558–625
Neyman J 1952 *Lectures and Conferences on Mathematical Statistics and Probability*. US Department of Agriculture, Washington, DC [The 1952 edition is an expanded and revised version of the original 1938 mimeographed edition.]
Rao J, Bellhouse D 1990 The history and development of the theoretical foundations of survey based estimation and statistical analysis. *Survey Methodology* **16**: 3–29
Särndal C, Swensson B, Wretman J 1992 *Model Assisted Survey Sampling*. Springer Verlag, New York
Seneta E 1985 A sketch of the history of survey sampling in Russia. *Journal of the Royal Statistical Society Series A* **148**: 118–25
Seng Y 1951 Historical survey of the development of sampling theories and practice. *Journal of the Royal Statistical Society, Series A* **114**: 214–31
Splawa-Neyman J 1923 On the application of probability theory to agricultural experiments. Essay on principles (in Polish). *Roczniki Nauk Rolniczych Tom* (*Annals of Agricultural Sciences*) **X**: 1–51 [Sect. 9, translated and edited by D. M. Dabrowska and T. P. Speed, appeared with discussion in *Statistical Science* (1990) **5**: 463–80]
Splawa-Neyman J 1925 Contributions of the theory of small samples drawn from a finite population. *Biometrika* **17**: 472–9 [The note on this republication reads 'These results with others were originally published in *La Revue Mensuelle de Statistique*, publ. Par l'Office Central de Statistique de la République Polonaise 1923 tom. vi, 1–29]
Stephan F 1948 History of the use of modern sampling procedures. *Journal of the American Statistical Association* **43**: 12–39
Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA
Sudman S 1967 *Reducing the Costs of Surveys*. Aldine, Chicago
Taeuber C 1978 Census. In: Kruskal W H, Tanur J M (eds.) *International Encyclopedia of Statistics*. Macmillan and the Free Press, New York, pp. 42–6
Tchouproff A A 1918a On the mathematical expectation of the moments of frequency distributions (Chap. 1 and 2). *Biometrika* **12**: 140–69
Tchouproff A A 1918b On the mathematical expectation of the moments of frequency distributions (Chap. 3 and 4). *Biometrika* **12**: 185–210
Tschuprow A A 1923a On the mathematical expectation of the moments of frequency distributions in the case of correlated observations (Chaps. i–iii). *Metron* **2**: 461–93
Tschuprow A A 1923b On the mathematical expectation of the moments of frequency distributions in the case of correlated observations (Chaps. iv–vi). *Metron* **2**: 646–80
Willcox W 1930 Census. In: Seligman E R A, Johnson A (eds.) *Encyclopedia of Social Sciences*, Vol. 3. Macmillan, New York pp. 295–300
Yates F 1946 A review of recent developments in sampling and sample surveys (with discussion). *Journal of the Royal Statistical Society* **109**: 12–42
Zarkovic S S 1956 Note on the history of sampling methods in Russia. *Journal of the Royal Statistical Society, Series A* **119**: 336–8
Zarkovich S S 1962 A supplement to 'note on the history of sampling methods in Russia.'*Journal of the Royal Statistical Society, Series A* **125**: 580–2

S. E. Fienberg and J. M. Tanur

Sample Surveys: Methods

A survey consists of a number of operations or steps of survey design. The ultimate goal of the survey is to generate estimates of population parameters, based on observations of (a sample of) the units that comprise the population. The design steps can be viewed as a chain of links. The chain is no stronger than its weakest link, for example, see *Survey Sampling: The Field*.

The steps can be labeled in the following way: (a) Research objectives are defined, i.e., the subject-matter problem is translated into a statistical problem. Researchers must define the target population they want to study and the concepts they wish to measure. Indicators (variables) of the concepts are chosen and eventually, questions are formulated. Problems in this first step usually result in relevance errors, see Hox (1997). (b) A frame of the population is developed. The frame could be a list of population units, a map, or even a set of random numbers that could be used to access a population of individuals with telephones. Coverage errors result when frame units do not completely correspond with population units. (See Groves 1989.) (c) The mode of administrating the survey is chosen. The suitable survey mode depends on budget constraints, topic, and the type of measurement that is being considered. Common modes include face-to-face interviews, telephone interviews, diaries, and administrative records. New modes related to the Internet have recently entered the scene, see Lyberg and Kasprzyk (1991) and Dillman (2000). (d) The questionnaire is developed. For each survey construct, one or more survey questions are developed. Questionnaire development is complicated since the perception of questions varies among respondents and is sensitive to a number of cognitive phenomena. Effects that are treated in the literature include: question wording, question order, order between response alternatives, context, navigational principles, and the influence of interviewers. Many large survey organizations emphasize this step and have created cognitive laboratories to improve their questionnaire work, see Sudman et al. (1996) and Forsyth and Lessler (1991). (e) The sampling design specifies the sampling unit, the method for randomly selecting the sample from the frame, and the sample size. The choice depends on the assumed population variability and the costs of sampling different units, see Särndal et al. (1992). (f)

Data are collected, i.e., observations or measurements are made of the sampled units. Typically both random and systematic errors occur at this stage, see de Leeuw and Collins (1997). (g) Data are processed. This step is needed to make estimation and tabulation possible, given that only raw data exist. The processing includes editing activities where data consistency and completeness are controlled, entry of data (data capture), and coding of variables, see Lyberg and Kasprzyk(1997).

All design steps are interdependent. A decision regarding one step has an impact on other steps. Thus, several iterations of the above steps of the design process may be necessary before the final design is determined. There follows details of the survey process.

1. Populations and Population Parameters

A population is a set of units. For example, a population might be the inhabitants of a country, the households in a city, or the companies in a sector of industry. The 'units of study' are the units used as a basis of the analysis, for example individuals, households, or companies. The 'target population' is that particular part of the population about which inference is desired. The 'study variables' are the values of the units studied, for example the age of an individual, the disposable income of a household, or the number of employees in a company. The 'population parameters' are characteristics of the population, summarizing features of the population, for example the average disposable income for households in a population. Most often interest centers on finding the population parameters for specific subsets of the population such as geographical regions or age groups. The subsets are called 'domains of study.'

2. Sampling Units, Frames, and Probability Samples

Frequently, statements about some population parameters are problematic because of a lack of time and funding necessary for surveying all units in the population. Instead we have to draw a sample, a subset of the population, and base conclusions on that sample. High quality conclusions concerning the parameter values require care when designing the sample. Assessing the quality of the conclusions made can be accomplished by using a probability sampling design, which is a design such that all units in the population is given a known probability of being included in the sample. Given that knowledge, it is possible to produce estimates of the parameters that are unbiased and of high precision. Moreover it is also possible to estimate the precision of the estimates based on the sample. This is usually in contrast to nonprobability sampling designs where, strictly speaking, the statistical properties of the estimators are unknown unless one is willing to accept distribution assumptions for the variables under study in combination with the sampling design.

Probability sampling requires a frame of sampling units. The units of study in the population can be of different types, such as individuals, households, or companies. Even if the analysis is intended to be based on these units it is often practical or even necessary to group the population units into units that are better suited for sampling. The reason for this can be that, due to imperfect frames, there is no register of the units of study and therefore the units cannot be selected directly. Instead, groups of units can be selected, for example, villages or city blocks, and interviews are conducted with the people in the villages or blocks. Here the sampling unit consists of a cluster of units of study. In many countries, the demarcation of the clusters are based on geographical units identified through areas on maps. Often the clusters are formed to be a suitable workload for an interviewer or enumerator and are called the enumeration area. The enumeration areas are often formed and updated in population censuses. In countries where the telecommunication system is well developed, interviews are often done by telephone. The sampling unit in this case is the cluster of individuals, i.e., the household members that have access to the same telephone.

The frame can take different shapes. For example, a list of N students at a university complete with names and addresses is a frame of students. In this case the sampling units and the units of study are the same. This makes it easy to select a sample of n students by numbering the students from 1 to N and generating n random numbers in a computer. Then we select the students that have the same number as the realized random numbers. Note that in doing so we deliberately introduce randomization through the random number generation. This is in contrast to methods that do not use randomization techniques but act as if data are independent identically distributed random variables regardless of how the data was obtained.

Often, the sampling units are not the same as the study units. If, for example, the units of study are individuals in a country and the target population consists of all individuals living in the country at a specific date, there is seldom a list or data file that contains the individuals. Also the cost of making such a list would be prohibitive.

Using a sequence of frames containing a sequence of sampling units of decreasing sizes can improve this situation. For example, if there exists a list of villages in the country, then this is a frame of primary sampling units. From a sample of villages, a list of all households within the villages selected can be made. This is a frame of secondary sampling units. Note that there is a need for making the list of households only for the villages selected. Using this list, some of the households may be selected for interviews. Thus, a sequence

of frames containing a sequence of sampling units has been constructed in a way that permits selection of a probability sample of units of study without enumerating all the individuals in the country. The existence of, or the possibility of, creating reliable frames is a prerequisite for making efficient sampling designs. Often the frame situation will govern the sampling procedure. Ideally the frame should cover the target population and *only* that population. It should provide the information necessary for contacting the units of study such as names, addresses, telephone numbers, or the like. If there is auxiliary information about the units, it is very practical if it is included in the frame. Auxiliary information is useful for constructing efficient sampling designs. A specific type of frame is used in 'Random Digit Dialing' (RDD). When telephone numbers are selected, a common technique is to select numbers by generating them at random in a computer. The frame is the set of numbers that could be generated in the computer. The selection probability of the cluster is proportional to the number of telephones to which the household members have access.

3. Sampling Methods and Estimators

The overall goal is to be able to make (high) quality estimates of the parameters as cheaply as possible. This is implemented using a 'sampling strategy,' i.e., a combination of sampling design (a certain combination of sampling methods) and an 'estimator,' which is as efficient as possible, i.e., gives the highest precision for a given cost. There exists a variety of sampling methods that can be applied in different situations depending on the circumstances. The most frequently employed sampling methods are:

(a) *Simple Random Sampling (SRS)*. Every unit in the population is given an equal chance of being included in the sample.

(b) *Systematic Sampling*. Every kth unit in the frame is included in the sample, starting from a randomly selected starting point. The sampling step k is chosen so that the sample has a predetermined size.

(c) *Stratified Sampling*. The units are grouped together in homogenous groups (strata) and then sampled, for example, using SRS within each group. Stratified sampling is used either for efficiency reasons or to ensure a certain sample size in some domains of study, such as geographical regions or specific age groups. It is necessary to know the values of a stratification variable in advance to be able to create the strata. Stratification is efficient if the stratification variable is related to the variables under study. If the number of units that are being selected within each stratum is proportional to the number of units in the stratum, then this is called proportional allocation. Usually this will give fairly efficient estimates. It can be improved upon using optimal allocation, which requires knowledge of the variability of the study variables or correlated variables within each stratum. In many cases the improvement is small compared to proportional stratification.

(d) *Unequal Probability Sampling*. This is a method, which can be employed either for efficiency reasons or for sheer necessity. Often the probability of selecting a cluster of units is (or is made to be) proportional to the number of units in the cluster. This is called selecting with probability proportional to size (PPS), with size in this case equal to the number of units. The measure of size can vary. The efficiency of the estimator will increase if the measure of size is related to the study variables.

(e) *Multistage Sampling*. This is a technique, which is typically used in connection with frame problems.

The above-mentioned sampling methods can be combined to produce good sampling designs. For example, the following is an example of a 'master sample design' that is used in some countries. Suppose that the object is to conduct a series of sample surveys concerning employment status, living conditions, and household expenditures. Also suppose that based on a previous census there exist enumeration areas (EAs) for the country, which are reasonably well updated so that approximations of the population sizes are available for each EA. Suppose that it is deemed appropriate to stratify the EAs according to geographical regions both because the government wants these regions as domains of study and because there is reason to believe that the consumption pattern is different within each region. Thus, the stratification of the EAs would serve both purposes, namely controlling the sample sizes for domains of study and presumably being more efficient than SRS. A number of EAs are selected within each stratum. The number of EAs to be selected is chosen to be proportional to the aggregated population figures in the regions according to the most recent census. The EAs are the primary sampling units. If the EAs vary a lot in size it could be efficient to select the EAs with PPS; otherwise SRS could be used.

The EAs so selected, constitute the master sample of primary sampling units that will be kept the same for a number of years. The actual number of years will depend on the migration rate within the regions. From this master sample a number of households are selected each time a new survey is to be conducted. When selecting the households, it is possible to make a new updating of the EAs, i.e., to make a new list of the households living in the area and then select with systematic sampling or SRS a number of households to be interviewed. The households are the secondary sampling units. Often the sampling fractions in the second stage are determined in such a way that the resulting inflation factor for the households is the same. This is called a 'self-weighting design.' In principle, a self-weighting design makes it possible to calculate the estimates and the precision of the estimates without a computer.

4. Estimators, Auxiliary Information, and Sample Weights

To each sampling design there corresponds at least one estimator, which is a function of the sample data used for making statements about the parameters. The form of the estimator depends on the sampling design used and also on whether auxiliary information is included in the function. In survey sampling, some auxiliary information is almost always available, i.e., values of concomitant variables known for all units in the population. If this information is correlated to the study variables, it can be used to increase the efficiency of the sampling strategy either by incorporating the information in the sampling method, as in stratification, or in the probability of including the unit in the sample as in PPS, or in adjusting the estimator. A very general form of an estimator is the so-called generalized regression estimator (Cassel et al. 1976), which for estimating the population total takes the form

$$t_{GR} = t_y + \hat{\beta}(t_x - T_x)$$

where x denotes the values of the auxiliary variable, T_x the known value of the population total of x, $\hat{\beta}$ is the regression coefficient estimated from the sample between x and the study variable y and

$$t_x = \sum_1^n \frac{x_i}{\alpha_i}$$

is the so-called Horvitz–Thompson estimator of X, and α_i the inclusion probability of unit i. The function t_y is the estimator of the unknown value of the parameter. The motivation for using this estimator is based on the conviction that the study variable and the auxiliary variable are linearly related. Thus, this can be seen as one example of the use of models in survey sampling. If the inclusion probabilities are the same for all units, i.e., if $\alpha_i = (n/N)$, then the Horvitz–Thompson estimator becomes the expanded sample mean, which is an unbiased estimator of the population total under the SRS scheme if the expectation is taken with respect to the design properties.

The sample weights are the inflation factors used in the estimator for inflating the sample data. They are usually a function of the inverted values of the inclusion probabilities. In the Horvitz–Thompson estimator, the sample weights are $N(\alpha_i n)^{-1}$. For the sample mean, the sample weights become N/n. As was noted earlier, to each sampling design, there exists an estimator that is natural, i.e., it is unbiased in the design sense. For example, in the case of multistage probability sampling, the sampling weights are functions of the different selection probabilities in each selection step. If the selection probabilities are carefully selected they may form a weight, which is constant for all units in the sample, and the value taken by the estimator can be calculated simply by summing the values of the units in the sample and multiplying the sum by the constant. This is called a 'selfweighted estimator' (and design). This technique was important before the breakthrough of the personal computer, because it saved a lot of manual labor but with today's powerful PCs, it has lost its merits. The generalized regression estimator can be rewritten in the form:

$$t_{GR} = \sum_1^n w_i x_i$$

where w_i are the sample weights pertinent to unit i for this estimator. The 'calibration estimator' is the form of the regression estimator where the sample weights have been adjusted so that $t_x \equiv T_x$. This adjustment causes a small systematic error, which decreases when the sample size increases.

5. Assessing the Quality of the Estimates—Precision

The quality of the statements depend among other things on the precision of the estimates. Probability sampling supports measurement of the sampling error, since a randomly selected subset of the population is a basis for the estimates. The calculation takes into consideration the randomization induced by the sampling design. Because the design might be complex, the calculation of the precision might be different from what is traditionally used in statistics. For example the variance of the sample mean $\bar{y}_s$ for estimating the population mean $\bar{y}$ under SRS with replacement is S^2/n where S^2 is the population variance, where

$$S^2 = \frac{1}{N-1}\sum_1^N (y_i - \bar{y})^2$$

and n is the sample size. This formula is similar to that used in traditional statistical analysis. However for the Horvitz–Thompson estimator, the variance is:

$$\frac{1}{2}\sum_1^N \sum_1^N \left(\frac{\alpha_i \alpha_j - \alpha_{ij}}{\alpha_{ij}} \right) \left(\frac{y_j}{\alpha_j} - \frac{y_i}{\alpha_i} \right)^2$$

where α_{ij} is the joint inclusion probability of units i and j. As can be seen, the calculation of the variance becomes complicated for complicated sampling designs, having a large number of selection stages and also unequal selection probabilities in each step. This complexity becomes even more pronounced when estimating population parameters in subgroups where the means are ratios of random variables. However some shortcut methods have been developed: ultimate cluster techniques, Taylor series, jack-knifing, and replications, see Wolter (1985). It is also evident from the formula for the variance that the variance of the

Horvitz–Thompson estimator depends on the relation between the values of the study variable y and the inclusion probability α. If the inclusion probability can be made to be approximately proportional to the values of the study variable, then the variance becomes small. That is one reason why much emphasis is put on the existence of auxiliary information in survey sampling. However, as was shown by Godambe (1955) there does not exist a uniformly minimum variance unbiased estimator in survey sampling when inference is restricted to design-based inference.

6. *Assessing the Quality of the Estimates—Total Survey Error*

The total survey error can be measured by the mean squared error (MSE), which is the sum of the variance and the squared bias of the estimate. Regular formulas for error estimation do not take all MSE components into account but rather the precision components mentioned above. Basically, a variance formula includes random variation induced by the sampling procedure and the random response variation. Components such as correlated variances induced by interviewers, coders, editors, and others have to be estimated separately and added to the sampling and response variance. The same goes for systematic errors that contribute to the bias. They have to be estimated separately and added, so that a proper estimate of MSE can be obtained.

Often the survey quality is visualized by a confidence interval, based on the estimated precision. Obviously such an interval might be too short because it has not taken the total error into account. Some of the sources of correlated variances and biases can be traced to:

(a) respondents who might have a tendency to, for instance, under-report socially undesirable behaviors;

(b) interviewers who might systematically reformulate certain questions and do so in an interviewer-specific fashion;

(c) respondents who do not participate because they cannot be contacted or they refuse;

(d) incomplete frames resulting in, e.g., undercoverage; and

(e) coders who might introduce biased measurements if they tend to erroneously code certain variable descriptions.

There are of course many other possibilities for error to occur in the estimates. There are basically two ways of assessing the quality of estimates: (a) The components of MSE can be estimated—this is a costly and time-consuming operation and (b) by a modeling approach—it might be possible to include, e.g., nonresponse errors and coverage errors in the assessment formulas. This method's success depends on the realism of the modeling of the error mechanisms involved.

One should bear in mind, however, that for each survey step there are methods designed to keep the errors small. Systematic use of such known dependable methods decreases the need for evaluation studies or heavy reliance on modeling.

See also: Sample Surveys: Nonprobability Sampling; Sample Surveys: Survey Design Issues and Strategies

Bibliography

Cassel C M, Särndal C E, Wretman J H 1976 Some results on generalized difference and generalized regression estimation for finite populations. *Biometrika* **63**: 615–20

De Leeuw E, Collins M 1997 Data collection methods and survey quality: An overview. In: Lyberg L et al. (eds.) *Survey Measurement and Process Quality*. Wiley, New York

Dillman D 2000 *Mail and Internet Surveys*. Wiley, New York

Godambe V P 1955 A unified theory of sampling from finite populations. *Journal of Royal Statistical Society, Series B* **17**: 269–78

Groves R 1989 *Survey Errors and Survey Costs*. Wiley

Forsyth B, Lessler J 1991 Cognitive laboratory methods: A taxomomy. In: Biemer P et al. (eds.) *Measurement Errors in Surveys*. Wiley, New York

Hox J 1997 From theoretical concept to survey question. In: Lyberg L et al. (eds.) *Survey Measurement and Process Quality*. Wiley, New York, pp. 47–70

Lyberg L, Kasprzyk D 1991 Data collection methods and measurement error: An overview. In: Biemer P et al. (eds.) *Measurement Errors in Surveys*. Wiley

Lyberg L, Kasprzyk D 1997 Some aspects of post-survey processing. In: Lyberg L et al. (eds.) *Survey Measurement and Process Quality*. Wiley

Särndal C E, Swensson B, Wretman J 1992 *Model Assisted Survey Sampling*. Springer-Verlag, New York

Sudman S, Bradburn N, Schwarz N 1996 *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco

Wolter K 1985 *Introduction to Variance Estimation*. Springer-Verlag

C. M. Cassel and L. Lyberg

Sample Surveys: Model-based Approaches

The theory for sample surveys has been developed using two theoretical frameworks: design-based and model-based. The design-based approach uses the probabilities with which units are selected for the sample for inference; in the model-based approach, the investigator hypothesizes a joint probability distribution for elements in the finite population, and uses that probability distribution for inference. In this article, the two approaches are described and compared, and guidelines are given for when, and how,

one should perform a model-based analysis of survey data. The use of models for survey design is discussed briefly.

1. Inference in Sample Surveys

How does one generalize from individuals in a sample to those not observed? The problem of induction from a sample was much debated by philosophers, social scientists, and mathematicians of the eighteenth and nineteenth centuries, including Immanuel Kant, Charles Peirce, John Venn, and Adolphe Quetelet. In the early years of the twentieth century, many investigators resisted the idea of using survey samples rather than censuses for 'serious statistics' because of inference issues (see *Sample Surveys, History of*). The debates involving official uses of sample surveys in these years resulted in the development of two philosophical frameworks for inference from a sample: design-based inference and model-based inference.

In design-based inference, first expounded systematically by Neyman (1934), the sample design provides the mechanism for inferences about the population. Suppose that a without-replacement probability sample (see *Sample Surveys: Methods*) of n units is to be taken from a population of N units. A random variable Z_i is associated with the ith unit in the population; $Z_i = 1$ if the unit is selected for inclusion in the sample, and $Z_i = 0$ if the unit is not selected. The joint probability distribution of $\{Z_1, \ldots, Z_N\}$ is used for inference statements such as confidence intervals (see *Estimation: Point and Interval*).

The quantity being measured on unit i, y_i, is irrelevant for inference in the design-based approach. Whether y_i is household income, years of piano lessons, or number of cockroaches in the kitchen, properties of estimators depend exclusively on properties of the random variables $\{Z_1, \ldots, Z_N\}$ that describe the probability sampling design. The Horvitz–Thompson (1952) estimator,

$$\sum_{i=1}^{N} Z_i y_i / \pi_i \tag{1}$$

where $\pi_i = P(Z_i = 1)$, is an unbiased estimator of the population total $\sum_{i=1}^{N} y_i$; the variance of the Horvitz–Thompson estimator,

$$\sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j (\pi_i \pi_j)^{-1} \mathrm{Cov}\,(Z_i, Z_j)$$

depends on the covariance structure of $\{Z_1, \ldots, Z_N\}$.

The design-based approach differs from the inferential framework used in most other areas of statistics. There, y_i is the observed value of a random variable Y_i; the joint probability distribution of the Y_i's and proposed stochastic models allow inferential statements to be made. Following questions raised by Godambe (1955) about optimal estimation and survey inference, Brewer (1963) and Royall (1970) suggested that the same model-based frameworks used in other areas of statistics also be used in finite population sampling. Thompson (1997) summarized various approaches to prediction using models.

In the model-based approach to inference, the finite population values are assumed to be generated from a stochastic model. Regression models (see *Linear Hypothesis: Regression (Basics)*) are often adopted for this purpose: if covariates $x_{i1}, x_{i2}, \ldots, x_{ip}$ are known for every unit in the population, a possible model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{2}$$

where the ε_i's are random variables with mean zero and specified covariance structure. The parameters β_j for $j = 0, \ldots, p$ may be estimated by standard techniques such as generalized least squares, and the regression equation used to predict the value of y for units not in the sample. Then the finite population total is estimated by summing the observed values of the y_i's for units in the sample and the predicted values of the response for units not in the sample.

To illustrate the two approaches, consider a hypothetical survey taken to study mental health status in a population of 10,000 women over age 65. A stratified random sample of 100 urban women and 100 rural women is drawn from a population of 6,000 urban women and 4,000 rural women over age 65. One response of interest is the score on a depression inventory. Let $\bar{y}_U$ denote the sample mean score for the urban women and let $\bar{y}_R$ denote the sample mean score for the rural women.

Under the design-based approach as exposited in Cochran (1977), $\pi_i = 1/60$ if person i is in an urban area and $\pi_i = 1/40$ if person i is in a rural area. Every urban woman in the sample represents herself and 59 other urban women who are in the population but not in the sample; every sampled rural woman represents herself plus 39 other rural women. The Horvitz–Thompson estimator for the mean depression score for the population is $N^{-1}\sum_{i=1}^{N} Z_i y_i / \pi_i$; here, the estimated population mean score under the stratified random sampling design is $0.6\bar{y}_U + 0.4\bar{y}_R$. A 95 percent confidence interval (CI) for the mean refers to the finite population: if a CI were calculated for each possible sample that could be generated using the sampling design, the selection probabilities of the samples whose CIs include the true value of the population mean depression score sum to 0.95. Inference refers to repeated sampling from the population, not to the particular sample drawn.

Any number of stochastic models might be considered in a model-based approach. Consider first a special case of the regression model in Eqn. (2), and assume that the ε_i's are independent and normally

distributed with constant variance. Model 1 has the form $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $x_i = 1$ if person i is an urban resident and $x_i = 0$ if person i is a rural resident. The stochastic model provides a link between sampled and unsampled units: women who are not in the sample are assumed to have the same mean depression score as women with the same urban/rural status who are in the sample. Under Model 1, $\hat{\beta}_0 = \bar{y}_R$ and $\hat{\beta}_1 = \bar{y}_U - \bar{y}_R$ are the least squares estimates of β_0 and β_1. Thus the predicted value of y for each urban woman is $\bar{y}_U$, and the predicted value of y for each rural woman is $\bar{y}_R$. The model-based estimate of mean depression in this population is thus

$$\frac{1}{10{,}000}[100\bar{y}_U + 100\bar{y}_R + 5{,}900\bar{y}_U + 3{,}900\bar{y}_R]$$

$$= 0.6\bar{y}_U + 0.4\bar{y}_R$$

With Model 1, the point and interval estimates of the finite population mean are the same as for the design-based approach. The 95 percent confidence interval is interpreted differently, however; 95 percent of CIs from samples with the same values of x_i that could be generated from the model are expected to contain the true value of $\beta_0 + 0.6\beta_1$.

In this example, model-based inference with Model 1 accords with the results from design-based inference. But suppose that the model adopted is Model 2: $Y_i = \mu + \varepsilon_i$. Then the predicted value of the response for all women in the finite population is $\bar{y} = (\bar{y}_U + \bar{y}_R)/2$, and the mean depression score for the finite population is also estimated by $\bar{y}$. If depression is higher among rural women than among urban women, the estimate $\bar{y}$ from Model 2 will likely overestimate the true mean depression score in the population of 10,000 women because rural women are not proportionately represented in the sample.

In the model-based approach, inference is not limited to the 10,000 persons from whom the sample is drawn but applies to any set of persons for whom the model is appropriate. Random selection of units is not required for inference as long as the model assumptions hold. If the model is comprehensive enough to include all important information, as in Model 1 above, then the sampling design is irrelevant and can be disregarded for inference.

The two types of inference differ in the conception of randomness. Design-based inference relies on the randomness involved in sample selection; it relates to the actual finite population, and additional assumptions are needed to extend the inferences to other possible populations. Model-based inference is conditional on the selected sample; the randomness is built into the model, and the model assumptions are used to make inferences about units not in the sample. Design-based inference depends on other possible samples that could have been selected from the finite population but were not, while model-based inference depends on other possible populations that could have been generated under the model but were not.

The approaches are not completely separated in practice. Rao (1997) summarized a conditional design-based approach to inference, in which inference is restricted to a subset of possible samples. Särndal et al. (1992) advocated a *model-assisted* approach, in which a population model inspires the choice of estimator, but inference is based on the sampling design. In the depression example, a model-assisted estimator could incorporate auxiliary information such as race, ethnicity, and marital status through a model of the form in Eqn. (2); however, the stratified random sampling design is used to calculate estimates and standard errors.

2. *Models in Descriptive and Analytic Uses of Surveys*

Estimating a finite population mean or total is an example of a descriptive use of a survey: the characteristics of a particular finite population are of interest. In much social science research, survey data are used for analytic purposes: investigating relationships between factors and testing sociological theories. Data from the US National Crime Victimization Survey may be used to estimate the robbery rate in 1999 (descriptive), or they may be used to investigate a hypothesized relationship between routine activities and likelihood of victimization (analytic). In the former case, the population of inference is definite and conceivably measurable through a census. In the latter, the population of inference is conceptual; the investigator may well be interested in predicting the likelihood of victimization of a future person with given demographic and routine activity variables.

Smith (1994) argued that design-based inference is the appropriate paradigm for official descriptive statistics based on probability samples. Part of his justification for this position was the work of Hansen et al. (1983), who provided an example in which small deviations from an assumed model led to large biases in inference. Brewer (1999) summarized work on design-based and model-based estimation for estimating population totals and concluded that a model-assisted generalized regression estimator (see Särndal et al. 1992), used with design-based inference, captures the best features of both approaches. Models must of course always be used for inference in nonprobability samples (see *Sample Surveys: Nonprobability Sampling*); they may also be desirable in probability samples that are too small to allow the central limit theorem to be applied for inference.

Lohr (1999 Chap. 11) distinguished between obtaining descriptive official statistics and uncovering a 'universal truth' in an analytic use of a survey. Returning to the depression example, the investigator might be interested in the relationship between de-

pression score (y) and variables such as marital status, financial resources, number of chronic health problems, and ability to care for oneself. In this case, the investigator would be interested in testing a theory that would be assumed to hold not just for the particular population of 10,000 women but for other populations as well, and should be making inferential statements about the βs in model (2). The quantity for inference in the design-based setting is b_p, the least squares estimate of β that would be obtained if the x_is and y_i were known for all 10,000 persons in the finite population. The quantity b_p would rarely be of primary interest to the investigator, though, since it is merely a summary statistic for this particular finite population. In social research, models are generally motivated by theories, and a model-based analysis allows these theories to be tested empirically.

The generalized least squares estimator of β, $\hat{\beta}_{LS}$, would be the estimator of choice under a pure model-based approach because of its optimality properties under the proposed model. This estimator is, however, sensitive to model misspecification. An alternative, which achieves a degree of robustness to the model at the expense of a possibly higher variance, is to use the design-based estimator of b_p. If the proposed stochastic model is indeed generating the finite population and if certain regularity conditions are met, an estimator that is consistent for estimating b_p will also be consistent for estimating β. Under this scenario, a design-based estimate for b_p also estimates the quantity of primary interest β and has the advantage of being less sensitive to model misspecification.

Regardless of philosophical differences on other matters of inference, it is generally agreed that two aspects of descriptive statistics require the use of models. All methods currently used to adjust for nonresponse (see *Nonsampling Errors*) employ models to relate nonrespondents to respondents, although the models are not necessarily testable. In small area estimation, sample sizes in some subpopulations of interest are too small to allow estimates of sufficient precision; models are used to relate such subpopulations to similar subpopulations and to useful covariates.

3. Models for Small Area Estimation

In small area estimation, a model is used to estimate the response in subpopulations with few or no sample observations. As an example, the US Current Population Survey (CPS) provides accurate statistics about income and poverty for the nation as a whole. It was not designed, though, to provide accurate estimates in domains such as states, counties, or school districts—the sample would have to be prohibitively large in order to provide precise estimates of poverty for every county in the USA. These domains are called small areas—the term 'small' does not refer to the size of the area or the population, but to the fact that the sample size in the domain is small or may even be zero.

Consider the states to be the small areas, and let y_k be the proportion of school-age children who are poor in state k. The direct estimate $\bar{y}_k$ of y_k is calculated using data exclusively from the CPS, and $\hat{V}(\bar{y}_k)$ is an estimate of the variance of $\bar{y}_k$. Since in some states $\hat{V}(\bar{y}_k)$ is unacceptably large, the current practice for estimating poverty at the state level (see National Research Council, 2000, p. 49) uses auxiliary information from tax returns, food stamp programs, and the decennial census to supplement the data from the CPS. A regression model for predicting y_k using auxiliary information gives predicted values

$$\hat{y}_k = \hat{\beta}_0 + \sum_{j=1}^{4} \hat{\beta}_j x_{jk}$$

where the x_{jk} s represent covariates for state k (e.g., x_{1k} is the proportion of child exemptions reported by families in poverty in state k, and x_{2k} is the proportion of people receiving food stamps in state k). The predicted value $\hat{y}_k$ from the regression equation is combined with the direct estimate $\bar{y}_k$ from the CPS according to the relative amounts of information present in each: the small area estimate for state k is

$$\tilde{y}_k = \gamma_k \bar{y}_k + (1-\gamma_k)\hat{y}_k$$

where γ_k is determined by the relative precision of $\bar{y}_k$ and $\hat{y}_k$. If the direct estimate is precise for a state, i.e., $\hat{V}(\bar{y}_k)$ is small, then γ_k is close to one and the small area estimate $\tilde{y}_k$ relies mostly on the direct estimate. Conversely, if the CPS contains little information about state k's poverty rate, then γ_k is close to zero and $\tilde{y}_k$ relies mostly on the predicted value from the regression model.

The small area model allows the estimator for area k to 'borrow strength' from other areas and incorporate auxiliary information from administrative data or other sources. Ghosh and Rao (1994) and Rao (1999) review properties of this model and other models used in small area estimation.

4. Performing a Model-based Analysis

The first step in a model-based analysis for either descriptive or analytical use is to propose and fit a model to the data. Dependence among units, such as dependence among children in the same school, can be treated using hierarchical linear models or other methods discussed in Skinner et al. (1989).

The biggest concern in a model-based analysis, as pointed out by Hansen et al. (1983), is that the model may be misspecified. Many, but not all, of the assumptions implicit in a model can be checked using the sample data. Appropriate plots of the data provide

some graphical checks of model adequacy and correctness of the assumed variance structure, as described in Lohr (1999). These assumptions can also be partially checked by performing hypothesis tests of nested models, and by fitting alternative models to the data. In the depression example, plotting the data separately for rural and urban residents would reveal inadequacy of Model 2 relative to Model 1.

Another method that can sometimes detect model inadequacy is comparison of design-based and model-based estimates of model parameters. As mentioned in Sect. 2, if the model is correct for units in the finite population, then the design-based estimates and the model-based estimates should both be consistent for the model parameters. A substantial difference in the estimates could indicate that the sample design contains information not captured in the model, and that perhaps more covariates are needed in the model.

One crucial assumption that cannot be checked using sample data is that the model describes units not n the sample. This assumption is especially important in nonprobability samples and in use of models for nonresponse adjustment or small area estimation.

5. *Models in Survey Design*

Kalton (1983) distinguished between the use of models in survey analysis and in survey design, stating that 'the use of models to guide the choice of sample design is well-established and noncontroversial.' In good survey practice, a stratified sampling design is often chosen because it is thought that there are differences among stratum means. An unequal probability design may be employed because of prior belief that large counties have more variability in total number of crime victimizations than small counties; models provide a mechanism for formalizing some of the knowledge about population structure and exploring results of alternative assumptions. Cochran (1977) illustrated the use of models for designing systematic samples. Särndal et al. (1992, Chap. 12) summarized research on optimal survey design, in which auxiliary information about the population is used to select a design that minimizes the anticipated variance of an estimator under the model and design.

Models used for design purposes do not affect the validity of estimates in design-based inference. A poor model adopted while designing a probability sample may lead to larger variance of design-based estimates, but the estimates will retain properties such as unbiasedness under repeated sampling. A good model at the design stage often leads to a design with greatly increased efficiency.

A model-based analysis can be conducted on data from any sample, probability or non-probability. Probability sampling is in theory unnecessary from a pure model-based perspective, and Thompson (1997) and Brewer (1999) concluded that certain forms of purposive non-probability sampling can be superior to probability sampling when a model-based analysis is to be conducted and the model is correct. In practice, however, there is always concern that the assumed model may miss salient features of the population, and probability sampling provides some protection against this concern. For a ratio model, with $Y_i = \beta x_i + \varepsilon_i$ and $V[\varepsilon_i] = \sigma^2 x_i$, the model-based optimal design specifies a purposive sample of the population units with the largest x values. Such a design does not allow an investigator to check whether the model is appropriate for small xs; an unequal probability sample with π_i proportional to x_i does allow such model checking, and allows inferences under either design- or model-based frameworks. As Brewer (1999) pointed out, there is widespread public perception that 'randomized sampling is fair' and that perception provides a powerful argument for using probability sampling for official statistics.

The following sources are useful for further exploration of modes of inference in sample surveys. Lohr (1999, Chap. 11) provides a more detailed heuristic discussion of the role of models in survey sampling; Thompson (1997) gives a more mathematical treatment. The articles by Smith (1994), Rao (1997), and Brewer (1999) contrast inferential philosophies, discuss appropriate use of models in analysis of survey data, and provide additional references.

See also: Sample Surveys, History of; Sample Surveys: Methods; Sample Surveys: Survey Design Issues and Strategies; Sample Surveys: The Field

Bibliography

Brewer K R W 1963 Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* **5**: 93–105

Brewer K R W 1999 Design-based or prediction-based inference? Stratified random vs. stratified balanced sampling. *International Statistical Review* **67**: 35–47

Cochran W G 1977 *Sampling Techniques*, 3rd edn. Wiley, New York

Ghosh M, Rao J N K 1994 Small area estimation: An appraisal. *Statistical Science* **9**: 55–76

Godambe V P 1955 A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society B* **17**: 269–78

Hansen M H, Madow W G, Tepping B J 1983 An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association* **78**: 776–93

Horvitz D G, Thompson D J 1952 A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**: 663–85

Kalton G 1983 Models in the practice of survey sampling. *International Statistical Review* **51**: 175–88

Lohr S L 1999 *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA

National Research Council 2000 *Small-area Income and Poverty Estimates: Priorities for 2000 and Beyond*. Panel on Estimates

of Poverty for Small Geographic Areas, Committee on National Statistics. National Academy Press, Washington, DC
Neyman J 1934 On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**: 558–606
Rao J N K 1997 Developments in sample survey theory: An appraisal. *Canadian Journal of Statistics* **25**: 1–21
Rao J N K 1999 Some recent advances in model-based small area estimation. *Survey Methology* **25**: 175–86
Royall R M 1970 On finite population sampling theory under certain linear regression models. *Biometrika* **57**: 377–87
Särndal C E, Swensson B, Wretman J 1992 *Model Assisted Survey Sampling*. Springer-Verlag, New York
Skinner C J, Holt D, Smith T M F 1989 *Analysis of Complex Surveys*. Wiley, New York
Smith T M F 1994 Sample surveys 1975–1990; an age of reconciliation? *International Statistical Review* **62**: 5–34
Thompson M E 1997 *Theory of Sample Surveys*. Chapman & Hall, London

S. L. Lohr

Sample Surveys: Nonprobability Sampling

A sample collected from a finite population is said to be a *probability sample* if each unit of the population has nonzero probability of being selected into the sample, and that probability is known. Traditional methods of probability sampling include simple and stratified random sampling, and cluster sampling. Conclusions concerning the population may be obtained by design-based, or randomization, inference. See *Sample Surveys: The Field* and *Sample Surveys: Methods*. The values of variables of interest in the population are considered as fixed quantities, unknown except for those units selected into the sample. Inference proceeds by considering the behavior of estimators of quantities of interest under the randomization distribution, based on the known selection probabilities. For example, if the N population values of variable Y are denoted $Y_1, \ldots, Y_N$ and the n sample values by $y_1, \ldots, y_n$ then $\bar{y}$, the sample mean, is a possible estimator for $\bar{Y}$, the population mean. If the sample is obtained by simple random sampling, then, with respect to this randomization distribution, $\bar{y}$ is unbiased for $\bar{Y}$ and has sampling variance $(N-n)/Nn(N-1)\Sigma_{i=1}^{N}(Y_i-\bar{Y})^2$.

Nonprobability sampling refers to any method of obtaining a sample from a population which does not satisfy the criteria for probability sampling. Nonprobability samples are usually easier and cheaper to collect than probability samples, as the data collector is allowed to exercise some choice as to which units to include in the sample. For a probability sample, this choice is made entirely by the random sampling mechanism. However, methods of design-based inference cannot be applied to assess the bias or variability of estimators based on nonprobability samples, as such methods do not allow for unknown or zero selection probabilities.

Surveys carried out by national statistical agencies invariably use probability sampling. Marsh and Scarborough (1990) also noted 'the preponderance of probability sampling in university social science.' Nonprobability sampling is much more common in market and opinion research. However, Taylor (1995 observed large national differences in the extent to which nonprobability sampling, particularly quota sampling, is viewed as an acceptable tool for market research. In Canada and the USA, probability sampling using telephone polling and random-digit dialing is the norm for public opinion surveys. In Australia and South Africa probability sampling is also prevalent, but with face-to-face interviews. On the other hand, in many European countries such as France and the UK, quota sampling is much more common.

1. *Convenience Sampling*

The easiest and cheapest way to collect sample data is to collect information on those population units which are most readily accessible. A university researcher may collect data on students. Surveys carried out through newspapers, television broadcasts or Internet sites (as described, for example, by Bradley, 1999) are necessarily restricted to those individuals who have access to the medium in question. Sometimes only a small fraction of the population is accessible, in which case the sample may consist of exactly those units which are available for observation.

Some surveys involve an element of self-selection where individuals decide whether to include themselves in the sample or not. If participation is time-consuming, or financial cost is involved, then the sample is more likely to include individuals with an interest in the subject of the survey. This may not be important. For example, an interest in participating in an experimental study of behavior might be considered to be unlikely to be associated with the outcome of the experiment. However, where the variable of interest relates to opinion on a question of interest, as is often the case in newspaper, television or Internet polls, it is likely that interest in participation is related to opinion, and it is much harder to justify using the sample data to make conclusions about a wider population.

A famous example of the failure of such a nonprobability sample to provide accurate inferences about a wider population is the Literary Digest poll of 1936. Ten million US citizens were sent postcard ballots concerning the forthcoming presidential election. Around 2 million of these were returned, a sample size which, if associated with a simple random sample, would be expected to predict the population with negligible error. However, when calibrated against the

election results, the Literary Digest poll was in error by 19 percentage points in predicting Roosevelt's share of the vote.

On the other hand, useful inferences can be made using convenience samples. Smith and Sugden (1988) considered statistical experiments, where the allocation of a particular treatment to the units under investigation is controlled, usually by randomization. In such experiments, the selection of units is not usually controlled and is often a convenience sample. For example, individuals might be volunteers. Nevertheless, inferences are often successfully extended to a wider population. Similarly, *observational studies* where neither treatment allocation nor sample selection is controlled, usually because it is impossible to do so, can be thought of as arising from convenience samples. Smith (1983) noted that Doll and Hill (1964) in their landmark study of smoking and health, used a sample entirely made up of medical practitioners. However, the validity of extending conclusions based on their data, to the general population, is now widely recognized.

Studies based on convenience samples can be an extremely effective way of conducting preliminary investigations, but it is desirable that any important conclusions drawn about a wider population are further investigated, preferably using probability samples. Where some kind of explanatory, rather than simply descriptive, inference is desired, Smith and Sugden (1988) argued that 'the ideal studies are experiments within surveys in which the scientist has control over both the selection of units and the allocation of treatments.' This approach was considered in detail by Fienberg and Tanur (1989).

2. *Quota Sampling*

When using survey data to draw an inference about a population of interest, the hope of the analyst is that sample estimators of quantities of interest are close to the corresponding population values. If a nonprobability sample has been collected, then it is instructive to observe the precision of sample estimators of known population quantities. For example, how do the sample proportions of males and females compare to known population values? If they differ substantially, then the sample is 'unrepresentative' of the population and one might have legitimate cause for concern about the reliability of estimates of unknown quantities of interest. *Purposive sampling* is a term used for methods of choosing a nonprobability sample in a way that makes it 'representative' of the population, although there is no generally agreed definition of a representative sample, and purposive sampling is often based on subjective considerations.

In *quota sampling*, the sample selection is constrained to ensure that the sample proportions of certain control variables approximately match the known population proportions. For example, if the population proportions of males and females are equal, then equal numbers of male and female units are selected into the sample. Age groups are also commonly used in designing quota samples. Sample totals for each cell of a cross-classification of two or more control variables (for example, age by sex) may also be fixed by the design. Examples are given by Moser and Kalton (1971). Quota sampling is most commonly used in market and opinion research, where control variables usually include age, sex, and socioeconomic class. Other variables such as employment status and housing tenure are also used. The known population proportions for the control variables are calculated from census data, or from surveys based on large probability samples. Variables with known population totals which are not used in setting quotas may be used for weighting in any subsequent analyses.

Where data collection involves visiting households, further constraints beyond the quotas may be applied to sample selection. For example, data collectors may be assigned a prespecified travel plan. However, where the mode of data collection involves intercepting individuals on the street for interview, then the only constraint on the data collector may be to satisfy the quotas. It is this freedom given to the data collector that provides both the biggest advantage and biggest disadvantage of quota sampling. The advantage is that with only the quota constraints to satisfy, data collection is relatively easy. Such surveys can be carried out rapidly by an individual data collector performing interviews on a busy street corner. As with any nonprobability sampling scheme, however, there is no way of assessing the bias associated with quota sampling. The sample units are necessarily selected from those which are available to the data collector, given their mode of interviewing. If availability is associated with any of the survey variables, then significant bias may occur. Advocates of quota sampling argue that the quotas control for this, but there is no way of guaranteeing that they do. Neither can design-based inference be used to assess the variability of estimates based on quota samples. Sometimes, a simple model is used to assess this variability. If one assumes that the data collectors used are drawn from a population of possible data collectors, then the 'between collector' variance combines both sampling variability and interviewer variability. Deville (1991) modeled the quota sampling process and provided some alternative measures of variability.

Studies comparing quota and probability sampling have been carried out. Moser and Stuart (1953) discovered apparent availability biases in the quota samples they investigated, with respect to the variables occupation and education. In particular, they noticed that the quota samples underestimated the proportion of population with lower levels of education. Marsh and Scarborough (1990) investigated nine possible sources of availability bias in quota samples. They

found that, amongst women, their quota sample overestimated the proportion from households with children. Both studies found that the quota samples tended to underestimate the proportion of individuals in the extreme (high and low) income groups.

Quota samples are often used for political opinion polls preceding elections. In such examples they can be externally validated against the election results and historically quota samples have often been shown to be quite accurate. Indeed Worcester (1996) argued that election forecasts using quota samples for UK elections in the 1970s were more accurate than those using probability samples. Smith (1996) presented similar evidence. However, it is also election forecasting which has led to quota sampling coming under closest scrutiny. In the US presidential election of 1948, the Crossley, Gallup, and Roper polls all underestimated Truman's share of the vote by at least five percentage points, and as a consequence, predicted the wrong election winner. Mosteller et al. (1949) in their report on the failure of the polls found one of the two main causes of error to be errors of sampling and interviewing, and concluded (p. 304) that 'it is likely that the principal weakness of the quota control method occurred at the local level at which respondents are selected by interviewers.'

The UK general election of 1992 saw a similar catastrophic failure of the pre-election opinion polls, with pre-election polls giving Labour an average lead of around 1.5 percentage points. In the election, the Conservative lead over Labour was 7 percentage points. A report by the Market Research Society Working Party (1994) into the failure of the polls identified inaccuracies in setting the quota controls as one of a number of possible sources of error. As a result the sample proportions of the key variables did not accurately reflect the proportions in the population. Lynn and Jowell (1996) attributed much of the error to the selection bias inherent in quota sampling, and argued for increased use of probability sampling methods for future election forecasts.

3. A Formal Framework

As methods of design-based inference cannot be applied to data obtained by nonprobability sampling, any kind of formal assessment of bias and variability associated with nonprobability samples requires a model-based approach (see *Sample Surveys: Model-based Approaches*). Smith (1983) considered the following framework, which can be used to assess the validity of inferences from various kinds of nonprobability samples. Let $i = 1, \ldots, N$ denote the population units, vector Y_i the values of the unknown survey variables, and vector Z_i the values of variables which are known prior to the survey. Let A be a binary variable indicating whether a unit is selected into the sample ($A_i = 1$) or not ($A_i = 0$), and let A_s be the values of A for the observed sample. Smith (1983) modeled the population values of Y and the selection process jointly through

$$f(Y, A_s \mid Z; \theta, \phi) = f(Y \mid Z; \theta) f(A_s \mid Y, Z; \phi) \quad (1)$$

where θ and ϕ are distinct model parameters for the population model and selection model respectively. Given A_s, Y can be partitioned as $(Y_s, Y_{\bar{s}})$ into observed and unobserved values.

Inferences based on the observed data model $f(Y_s \mid Z; \theta)$ and extended to the population are said to ignore the selection mechanism, and in situations where this is valid, the selection is said to be *ignorable* (Rubin, 1976); see *Statistical Data, Missing*. Selection is ignorable when

$$f(A_s \mid Y, Z; \phi) = f(A_s \mid Z; \phi) \quad (2)$$

so that the probability of making the observed selection, for given Z, is the same for all Y. A sufficient condition for this is that A and Y are conditionally independent given Z. A probability sampling scheme, perhaps using some stratification or clustering based on Z, is clearly ignorable.

Nonprobability sampling schemes based on Z (for example selecting exactly those units corresponding to a particular set of values of Z) are also ignorable. However, whether or not inferences are immediately available for values of Z not contained in the sample depends on the form of the population model $f(Y \mid Z; \theta)$ and, in particular, whether the entire θ is estimable using Y_s. If Y is independent of Z then there is no problem, but this is an assumption which cannot be verified by sample data based on a restricted sample of values of Z. If this assumption seems implausible, then post-stratification may help. Smith (1993) considered partitioning the variables comprising Y into measurement variables Y^m and stratification variables Y^q, and post-stratifying. If

$$f(Y_s^m \mid Y_s^q, Z; \xi) = f(Y_s^m \mid Y_s^q; \xi) \quad (3)$$

where ξ are parameters for the post-stratification model, then inference for any Z is available. This condition implies that, given the observed values Y_s^m of the stratification variables, Z gives no further information concerning the measurement variables. This approach provides a way of validating certain inferences based on a convenience sample, where Z is an indicator variable defining the sample.

Smith (1983) also considered ignorability for quota sampling schemes. He proposed modeling selection into a quota sample in two stages, selection into a larger sample for whom quota variables Y^q are recorded, followed by selection into the final sample, based on a unit's quota variables and the requirements to fill the quota. For the final sample, the variables of interest Y^m are recorded. Two ignorability conditions result, requiring that at neither stage does probability of selection, given Y^q and Z, depend on Y^m.

This formal framework makes clear, through expressions such as (2) and (3) when model-based inferences from nonprobability samples can and cannot be used to provide justifiable population inferences. However, it is important to realize that the assumptions required to ensure ignorability cannot be verified using the sample data alone. They remain assumptions which need to be subjectively justified before extending any inferences to a wider population. These formal concepts of ignorability confirm more heuristic notions of what is likely to comprise a good nonprobability sampling scheme. For example, opinion polls with a large element of self-selection are highly unlikely to result in an ignorable selection. On the other hand one might have much more faith in a carefully constructed quota sampling scheme, where data collectors are assigned to narrowly defined geographical areas, chosen using a probability sampling scheme, and given restrictive guidelines on choosing the units to satisfy their quota.

4. Discussion

The distinction between probability sampling and nonprobability sampling is necessarily coarse. At one extreme is a carefully constructed probability survey with no nonresponse; at the other extreme is a sample chosen entirely for the investigator's convenience. However, most surveys fall between these two extremes, and therefore strictly should be considered as nonprobability samples. Examples include quota surveys of households where the geographical areas for investigation are chosen using a probability sample, or statistical experiments where a convenience sample of units is assigned treatments using a randomization scheme. The validity of any inferences extended to a wider population depends on the extent to which the selection of units is ignorable for the inference required. This applies equally to any survey with nonresponse. The presence of nonresponders in a probability survey introduces a nonprobability element into the selection mechanism. Considerations of ignorability (of nonresponse) now need to be considered. However, surveys with probability sampling usually make a greater effort to minimize nonresponse than nonprobability surveys, where there is little incentive to do so. Furthermore, even with nonresponse, it is easier to justify ignorability of a probability sampling mechanism.

Further details concerning specific issues may be obtained from the sources referenced above. Alternative perspectives on nonprobability sampling are provided by general texts on sampling such as Hansen et al. (1953), Stephan and McCarthy (1958) and Moser and Kalton (1971).

See also: Sample Surveys, History of; Sample Surveys: Survey Design Issues and Strategies

Bibliography

Bradley N 1999 Sampling for internet surveys. An examination of respondent selection for internet research. *Journal of the Market Research Society* **41**: 387–95

Deville J-C 1991 A theory of quota surveys. *Survey Methodology* **17**: 163–81

Doll R, Hill A B 1964 Mortality in relation to smoking: ten years' observations of British doctors. *British Medical Journal* **1**: 1399–410

Fienberg S E, Tanur J M 1989 Combining cognitive and statistical approaches to survey design. *Science* **243**: 1017–22

Hansen M H, Hurwitz W N, Madow W G 1953 *Sample Survey Methods and Theory. Volume 1: Methods and Applications*. Wiley, New York

Lynn P, Jowell R 1996 How might opinion polls be improved? The case for probability sampling. *Journal of the Royal Statistical Society A* **159**: 21–8

Market Research Society Working Party 1994 *The Opinion Polls and the 1992 General Election*. Market Research Society, London

Marsh C, Scarborough E 1990 Testing nine hypotheses about quota sampling. *Journal of the Market Research Society* **32**: 485–506

Moser C A, Kalton G 1971 *Survey Methods in Social Investigation*. Heinemann, London

Moser C A, Stuart A 1953 An experimental study of quota sampling (with discussion). *Journal of the Royal Statistical Society A* **116**: 349–405

Mosteller F, Hyman H, McCarthy P J, Marks E S, Truman D B 1949 *The Pre-election Polls of 1948: Report to the Committee on Analysis of Pre-election Polls and Forecasts*. Social Science Research Council, New York

Rubin D B 1976 Inference and missing data. *Biometrika* **63**: 581–92

Smith T M F 1983 On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society A* **146**: 394–403

Smith T M F 1996 Public opinion polls: the UK general election, 1992. *Journal of the Royal Statistical Society A* **159**: 535–45

Smith T M F, Sugden R A 1988 Sampling and assignment mechanisms in experiments, surveys and observational studies. *International Statistical Review* **56**: 165–80

Stephan F F, McCarthy P J 1958 *Sampling Opinions. An Analysis of Survey Procedure*. Wiley, New York

Taylor H 1995 Horses for courses: how survey firms in different countries measure public opinion with very different methods. *Journal of the Market Research Society* **37**: 211–19

Worcester R 1996 Political polling: 95% expertise and 5% luck. *Journal of the Royal Statistical Society A* **159**: 5–20

J. J. Forster

Sample Surveys: Survey Design Issues and Strategies

A treatment of survey questions intended to be useful to those wishing to carry out or interpret actual surveys should consider several issues: the basic difference between questions asked in surveys and

questions asked in ordinary social interaction; the problems of interpreting tabulations based on single questions; the different types of survey questions that can be asked; the possibility of bias in questioning; and the insights that can be gained by combining standard survey questions with randomized experiments that vary the form, wording, and context of the questions themselves. Each of these issues is treated in this article.

1. *The Unique Nature of Survey Questioning*

A fundamental paradox of survey research is that we start from the purpose of ordinary questioning as employed in daily life, yet our results are less satisfactory for that purpose than for almost any other. In daily life a question is usually asked because one person wishes information from another. You might ask an acquaintance how many rooms there are in her house, or whether she favors the legalization of abortion in America. The assumption on both sides of the interaction is that you are interested in her answers in and of themselves. We can call such inquiries ordinary questions.

In surveys we use similar inquiries—that is, their form, their wording, and the manner of their asking are seldom sharply distinguishable from ordinary questions. At times we may devise special formats, with names like Likert-type or forced-choice, but survey questions cannot depart too much from ordinary questioning because the essential nature of the survey is communication with people who expect to hear and respond to ordinary questions. Not surprisingly, respondents believe that the interviewer or questionnaire is directly interested in the facts and opinions they give, just as would an acquaintance who asked the same questions. They may not assume a personal interest in their answers, but what they do assume is that their answers will be combined with the answers of all others to give totals that are directly interpretable.

Thus, if attitudes or opinions are inquired into, the survey is viewed as a kind of referendum and the investigator is thought to be interested in how many favor and how many oppose legalized abortion or whatever else is at issue. If facts are being asked about, the respondent expects a report telling how many people have what size homes, or whatever the inquiry is about. (By factual data we mean responses that correspond to a physical reality and could, in principle, be provided by an observer as well as by a respondent, for example, when counting rooms. By attitudinal data we mean responses that concern subjective phenomena and therefore depend on self-reports by respondents. The distinction is not airtight: for example, the designation of a respondent's 'race' can be based on self-report but also on the observations of others, and the two may differ without either being clearly 'wrong.') In this article the focus will be on attitudes, including opinions, beliefs, and values, though much of the discussion can be applied to factual data as well.

Experienced survey researchers know that a simple tally of responses to a question—what survey researchers refer to as the 'marginals'—are usually too much a function of the way the question was asked to allow for any simple interpretation. The results of questions on legalized abortion depend heavily on the conditions, definitions, and other subtleties presupposed by the question wording, and the same is true to an extent even for a question on how many rooms there are in a house. Either we must keep a question quite general in phrasing and leave the definitions, qualifications, and conditions up to each respondent—which invites unseen variations in interpretation—or we must try to make the question much more limited in focus than was usually our goal in the first place.

Faced with these difficulties in interpreting univariate results from separate questions, survey investigators can proceed in one or both of two directions. One approach is to ask a wide range of questions on an issue and hope that the results can be synthesized into a general conclusion, even though this necessarily involves a fair amount of judgment on the part of the researcher. The other direction—the one that leads to standard survey analysis—is to hold constant the question (or the index, if more than a single item is being considered) and make comparisons across time or other variables. We may not be sure of exactly what 65 percent means in terms of general support for legalized abortion, but we act on the assumption that if the question wording and survey conditions have been kept constant, we can say, within the limits of sampling error, that it represents such and such an increase or decrease from an earlier survey that asked the same question to a sample from the same population. Or if 65 percent is the figure for men and 50 percent is the figure for women, a sex difference of approximately 15 percent exists.

Moreover, research indicates that in most cases relationships are less affected by variations in the form of question than are univariate distributions—generalized as the rule of 'form-resistant correlations' (Schuman and Presser 1981). The analytic approach, together with use of multiple questions (possibly further combined on the basis of a factor analytic approach), can provide a great deal of understanding and insight into an attitude, though it militates against a single summary statement of the kind that the respondents expect to hear. (Deming's (1968, p. 601) distinction between enumerative and analytic studies is similar, but he treats the results from enumerative studies as unproblematic, using a simple factual example of counting number of children. In this article, univariate results based on attitude questions are regarded as questionable attempts to simulate actual

referenda. Thus the change in terminology is important.)

This difference between what respondents expect—the referendum point of view—and what the sophisticated survey researcher expects—the analytic point of view—is often very great. The respondent in a national survey believes that the investigator will add up all the results, item by item, and tell the nation what Americans think. But the survey investigator knows that such a presentation is usually problematic at best and can be dangerously misleading at worst. Moreover, to make matters even more awkward, political leaders often have the same point of view as respondents: they want to know how many people favor and how many oppose an issue that they see themselves as confronting. Yet it may be neither possible nor desirable for the survey to pose exactly the question the policy maker has in mind, and in any case such a question is likely to be only one of a number of possible questions that might be asked on the issue.

2. *Problems with the Referendum Point of View*

There are several reasons why answers obtained from isolated questions are usually uncertain in meaning. First, many public issues are discussed at a general level as though there is a single way of framing them and as though there are just two sides. But what is called the abortion issue, to follow our previous example, consists of a large number of different issues having to do with the reasons for abortion, the trimester involved, and so forth. Likewise, what is called 'gun control' can involve different types of guns and different kinds of controls. Except at the extremes, exactly which of these particular issues is posed and with what alternatives makes a considerable difference in the univariate results. Indeed, often what is reported as a conflict in findings between two surveys is due to their having asked about different aspects of the same general issue.

A second problem is that answers to survey questions always depend on the form in which the question is asked, because most respondents treat that form as a constraint on their answers. If two alternatives are given by the interviewer, most respondents will choose one, rather than offering a substitute of their own that they might prefer. For example, in one survey-based experiment the authors identified the problems spontaneously mentioned when a national sample of Americans was asked to name the most important problem facing the country. Then a parallel question was formulated for a comparable sample that included *none* of the four problems mentioned most often spontaneously, but instead four problems that had been mentioned by less than three percent of the population in toto, though with an invitation to respondents to substitute a different problem if they wished. Despite the invitation, the majority of respondents (60 percent) chose one of the rare problems offered explicitly, which reflected their unwillingness to go outside the frame of reference provided by the question (Schuman and Scott 1987). Evidently, the form of a question is treated by most people as setting the 'rules of the game,' and these rules are seldom challenged even when encouragement is offered.

It might seem as though the solution to the rules-of-the-game constraint is to keep questions 'open'—that is, not to provide specific alternatives. This is often a good idea, but not one that is failsafe. In a related experiment on important events and changes from the recent past, 'the development of computers' was not mentioned spontaneously nearly as often as economic problems, but when it was included in a list of past events along with economic problems, the development of computers turned out to be the most frequent response (Schuman and Scott 1987). Apparently people asked to name an important recent event or change thought that the question referred only to political events or changes, but when the legitimacy of a different kind of response was made explicit, it was heavily selected. Thus a question can be constraining even when it is entirely open and even when the investigator is unaware of how it affects the answers respondents give.

A third reason for the limitations of univariate results is the need for comparative data to make sense in interpretation. Suppose that a sample of readers of this article is asked to answer a simple yes/no question as to its value, and that 60 percent reply positively and 40 percent negatively. Leaving aside all the problems of question wording discussed thus far, such percentages can be interpreted only against the backdrop of other articles. If the average yes percentage for all articles is 40 percent, the author might feel proud of his success. If the average is 80 percent, the author might well hang his head in shame. We are all aware of the fundamental need for this type of comparison, yet it is easy to forget about the difficulty of interpreting absolute percentages when we feel the urge to speak definitively about public reactions to a unique event.

Finally, in addition to all of the above reasons, there are sometimes subtle features of wording that can affect answers. A classic example of a wording effect is the difference between 'forbidding' something and 'not allowing' the same thing (Rugg 1941). A number of survey experiments have shown that people are more willing to 'not allow' a behavior than they are to 'forbid' the same behavior, even though the practical effects of the distinction in wording are nil (Holleman 2000). Another subtle feature is context: for example, a question about abortion in the case of a married woman who does not want any more children is answered differently depending on whether or not it is preceded by a question about abortion in the case of a defective fetus (Schuman and Presser 1996 [1981]).

The problems of wording and context appear equally when an actual referendum is to be carried out

by a government: considerable effort is made by politicians on all sides of the issue to control the wording of the question to be voted on, as well as its placement on the ballot, with the battle over these decisions sometimes becoming quite fierce. This shows that there is never a single way to phrase a referendum and that even small variations in final wording or context can influence the outcome of the voting. The same is true for survey questions, but with the crucial difference that they are meant to provide information, not to determine policy in a definitive legal sense.

The analytic approach, when combined with use of multiple questions to tap different aspects of an issue, provides the most useful perspective on survey data. Rather than focusing on the responses to individual items as such, the analysis of change over time and of variations across demographic and social background variables provides the surest route to understanding both attitudinal and factual data. Almost all important scholarly work based on surveys follows this path, giving attention to individual percentages only in passing. In addition, in recent years, classic between-subjects experiments have been built into surveys, with different ways of asking a question administered to random subsamples of a larger probability sample in order to learn about the effects of question wording (Schuman and Presser 1996 [1981]). These survey-based experiments, traditionally called 'split-ballots,' combine the advantage of a probability sample survey to generalize to a much larger population with the advantage of randomized treatments to test causal hypotheses. Survey-based experiments have been used to investigate a variety of methodological uncertainties about question formulations, as we will see below, and are also employed increasingly to test hypotheses about substantive political and social issues.

3. *Types of Survey Questions*

When investigators construct a questionnaire, they face a number of decisions about the form in which their questions are to be asked, though the decisions are often not made on the basis of much reflection. The effects of such decisions were first explored in survey-based experiments conducted in the mid-twentieth century and reported in books by Cantril (1944) and Payne (1951). In 1981 (1996), Schuman and Presser provided a systematic review of variations due to question form, along with much new experimental data. Recent books by Sudman et al. (1996), Tanur (1992), Tourangeau et al. (2000), and Krosnick and Fabrigar (forthcoming) consider many of these same issues, as well as a number of additional ones, drawing especially on ideas and research from cognitive psychology (see also *Questionnaires: Cognitive Approaches*; *Sample Surveys: Cognitive Aspects of Survey Design*).

An initial important decision is whether to ask a question in *open* or *closed* form. Open questions, where respondents answer in their own words and these are then coded into categories, are more expensive in terms of both time and money than closed questions that present two or more alternatives that respondents choose from. Hence, open questions are not common in present-day surveys, typically being restricted to questions that attempt to capture rapid change and that are easy to code, as in standard inquiries about 'the most important problem facing the country today.' In this case, immediate salience is at issue and responses can usually be summarized in keyword codes such as 'unemployment,' 'terrorism,' or 'race relations.' An open-ended approach is also preferable when numerical answers are wanted, for example, how many hours of television a person watches a week. Schwarz (1996) has shown that offering a specific set of alternatives provides reference points that can shape answers, and thus it is probably better to leave such questions open, as recommended by Bradburn et al. (1979).

More generally, open and closed versions of a question often do lead to different univariate response distributions and to different multivariate relations as well (Schuman and Presser 1996, [1981]). Partly this is due to the tendency of survey investigators to write closed questions on the assumption that they themselves know how to frame the main choices, which can lead to their overlooking alternatives or wording especially meaningful to respondents. Many years ago Lazarsfeld (1944) proposed as a practical compromise the use of open questions in the early development of a questionnaire, with the results then drawn on to frame closed alternatives that would be more efficient for use in an actual survey. What has come to be called 'cognitive interviewing' takes this same notion into the laboratory by studying carefully how a small number of individuals think about the questions they are asked and about the answers they give (see several chapters in Schwarz and Sudman 1996). This may not eliminate all open/closed differences, but it helps investigators learn what is most salient and meaningful to respondents.

Even after closed questions are developed, it is often instructive to include follow-up 'why' probes of answers in order to gain insight into how respondents perceived the questions and what they see their choices as meaning. Since it is not practical to ask such follow-ups to all respondents about all questions, Schuman (1966) recommended the use of a 'random probe' technique to obtain answers from a subsample of the larger sample of questions and respondents.

When the focus is on closed questions, as it often is, a number of further decisions must be made. A frequently used format is to state a series of propositions, to each of which the respondent is asked to indicate agreement or disagreement. Although this is an efficient way to proceed, there is considerable

evidence that a substantial number of people, especially those with less education, show an 'acquiescence bias' when confronted with such statements (Krosnick and Fabrigar forthcoming). The main alternative to the agree/disagree format is to require respondents to make a choice between two or more statements. Such a balanced format encourages respondents to think about the opposing alternatives, though it also requires investigators to reduce each issue to clearly opposing positions.

Another decision faced by question writers is how to handle DK (don't know) responses. The proportion of DK answers varies not only by the type of issue—there are likely to be more to a remote foreign policy issue than to a widely discussed issue like legalization of abortion (Converse 1976–77)—but also by how much DK responses are encouraged or discouraged. At one extreme, the question may offer a DK alternative as one of the explicit choices for respondents to consider, even emphasizing the desirability of it being given if the respondent lacks adequate information on the matter. At the other extreme interviewers may be instructed to urge those who give DK responses to think further in order to provide a more substantive answer. In between, a DK response may not be mentioned by the interviewer but can be accepted when volunteered.

Which approach is chosen depends on one's beliefs about the meaning of a DK response. Those who follow Converse's (1964) emphasis on the lack of knowledge that the majority of people possess about many public issues tend to encourage respondents to consider a DK response as legitimate. Those who argue like Krosnick and Fabrigar forthcoming) that giving a DK response is mainly due to 'satisficing' prefer to press respondents to come up with a substantive choice. Still other possibilities are that DKs can involve evasion in the case of a sensitive issue (e.g., racial attitudes) and in such cases it is unclear what respondents will do if prevented from giving a DK response.

A more general methodological issue is the desirability of measuring attitude strength. Attitudes are typically defined as favorable or unfavorable evaluations of objects, but the evaluations can also be seen as varying in strength. One can strongly favor or oppose the legalization of abortion, for example, but hold an attitude toward gun registration that is weaker, or vice versa. Further, there is more than one way to measure the dimension of strength, as words like 'extremity,' 'importance,' 'certainty,' and 'strength' itself suggest, and thus far it appears that the different methods of measurement are far from perfectly correlated (Petty and Krosnick 1995). Moreover, although there is evidence that several of these strength measures are related to actual behavior, for example, donating money to one side of the dispute about the legalization of abortion, in the case of gun registration the relation has been much weaker, apparently because other social factors (i.e., the effectiveness of gun lobbying organizations) play a large role independent of attitude strength (Schuman and Presser 1996 [1981].

4. *Question Wording and Bias*

Every question in a survey must be conveyed in words and words are never wholly neutral and unproblematic. Words have tone, connotation, implication—which is why both a referendum and a survey are similar in always coming down to a specific way of describing an issue. Of course, sometimes a question seems to have been deliberately biased, as when a 'survey' mailed out by a conservative organization included the following question:

> Do you believe that smut peddlers should be protected by the courts and the Congress, so they can openly sell pornographic materials to your children?

But more typical are two versions of a question that was asked during the Vietnam War:

> If a situation like Vietnam were to develop in another part of the world, do you think the United States should or should not send troops [*to stop a communist takeover*]?

Mueller (1973) found that some 15 percent more Americans favored military action when the bracketed words were included than when they were omitted. Yet it is not entirely clear how one should regard the phrase 'to stop a communist takeover' during that period. Was it 'biasing' responses to include the phrase, or was it simply informing respondents of something they might like to have in mind as they answered? Furthermore, political leaders wishing to encourage or discourage an action can choose how to phrase policy issues, and surveys cannot ignore the force of such framing if they wish to be relevant to important political outcomes.

Another instructive example was the attempt to study attitudes during the 1982 war between Argentina and Britain over ownership of a small group of islands in the South Atlantic. It was virtually impossible to phrase a question that did not include the name of the islands, but for the Argentines they were named the Malvinas and for the British the Falkland Islands. Whichever name was used in a question could be seen as prejudicing the issue of ownership. This is an unusual example, but it shows that bias in survey questions is not a simple matter.

5. *Conclusion*

In this sense, we return to the fundamental paradox of survey research—the referendum point of view vs. the analytic point of view. We all wish at times to know

what the public as a whole feels about an important issue—whether it involves military intervention in a distant country or a domestic issue like government support for health care. Therefore, we need to remind both ourselves and the public of the limitations of univariate survey results, while at the same time taking whatever steps we can to reduce those limitations. Above all, this means avoiding the tendency to reduce a complex issue to one or two simple closed questions, because every survey question imposes a unique perspective on responses, whether we think of this as 'bias' or not.

Moreover, survey data are most meaningful when they involve comparisons, especially comparisons over time and across important social groups—provided that the questions have been kept as constant in wording and meaning as possible. From a practical standpoint, probably the most useful way to see how the kinds of problems discussed here can be addressed is to read significant substantive analyses of survey data, for example, classics like Stouffer (1955) and Campbell et al. (1960), and more recent works that grapple with variability and bias, for example, Page and Shapiro (1992) and Schuman et al. (1997).

Bibliography

Cantril H 1944 *Gauging Public Opinion*. Princeton University Press, Princeton, NJ

Bradburn N M, Sudman S, with assistence of Blair E, Locander W, Miles C, Singer E, Stocking C 1979 *Improving Interview Method and Questionnaire Design*. Jossey-Bass, San Francisco

Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. Wiley, New York

Converse J M 1987 *Survey Research in the United States: Roots & Emergence 1890–1960*. University of California Press, Berkeley, CA

Converse P 1964 The nature of belief systems in mass publics. In: Apter D E (ed.) *Ideology and Discontent*. Free Press, New York

Converse J M 1976–7 Predicting no opinion in the polls. *Public Opinion Quarterly*? **40**: 515–30

Deming W E 1968 Sample surveys: the field. In: Sills D (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 13, pp. 594–612

Holleman B H 2000 *The Forbid/Allow Asymmetry: On the Cognitive Mechanisms Underlying Wording Effects in Surveys*. Rodopi, Amsterdam

Krosnick J A, Fabrigar L A forthcoming *Designing Great Questionnaires: Insights from Social and Cognitive Psychology*. Oxford University Press, Oxford, UK

Lazarsfeld P E 1944 The controversy over detailed interviews—an offer to negotiate. *Public Opinion Quarterly* **8**: 38–60

Mueller J E 1973 *War, Presidents, and Public Opinion*. Wiley, New York

Page B I, Shapiro R Y 1992 *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. University of Chicago Press, Chicago

Payne S L 1951 *The Art of Asking Questions*. Princeton University Press, Princeton, NJ

Petty R E, Krosnick J A (eds.) 1995 *Attitude Strength: Antecedents and Consequences*. Erlbaum, Mahwah, NJ

Rugg D 1941 Experiments in wording questions: II. *Public Opinion Quarterly* **5**: 91–2

Schuman H 1966 The random probe: a technique for evaluating the validity of closed questions. *American Sociological Review* **21**: 218–22

Schuman H, Presser S 1981 *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Academic Press, New York [reprinted 1996, Sage Publications, Thousand Oaks, CA]

Schuman H, Scott J 1987 Problems in the use of survey questions to measure public opinion. *Science* **236**: 957–9

Schuman H, Steeh C, Bobo L, Krysan M 1997 *Racial Attitudes in America: Trends and Interpretations*. Harvard University Press, Cambridge, MA

Schwarz N 1996 *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation*. Erlbaum, Mahwah, NJ

Schwarz N, Sudman S (eds.) 1996 *Answering Questions*. Jossey-Bass, San Francisco

Stouffer S A 1955 *Communism, Conformity, and Civil Liberties*. Doubleday, New York

Sudman S, Bradburn N B, Schwarz N 1996 *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco

Tanur J M (ed.) 1992 *Questions About Questions: Inquiries Into the Cognitive Bases of Surveys*. Russell Sage Foundation, New York

Tourangeau R, Ripps L J, Rasinski K A 2000 *The Psychology of Survey Response*. Cambridge University Press, Cambridge, UK

H. Schuman

Sample Surveys: The Field

1. Definition of Survey Sampling

Survey sampling can be defined as the art of selecting a sample of units from a population of units, creating measurement tools for measuring the units with respect to the survey variables and drawing precise conclusions about the characteristics of the population or of the process that generated the values of the units. A more specific definition of a survey is the following (Dalenius 1985):

(a) *A survey concerns a set of objects comprising a population*. One class of population concerns a finite set of objects such as individuals, businesses, and farms. Another, concerns events during a specific time period, such as crime rates and sales. A third class concerns plain processes, such as land use or the occurrence of certain minerals in an area. More specifically one might want to define a population as, for example, all noninstitutionalized individuals 15–74 years of age living in Sweden on May 1, 2000.

(b) *This population has one or more measurable*

properties. Examples of such properties are individuals' occupations, business' revenues, and the number of elks in an area.

(c) *A desire to describe the population by one or more parameters defined in terms of these properties*. This calls for observing (a sample of) the population. Examples of parameters are the proportion of unemployed individuals in the population, the total revenue of businesses in a certain industry sector during a given time period and the average number of elks per square mile.

(d) In order to get observational access to the population, a frame is needed i.e., an operational representation, such as a list of the population objects or a map of the population. Examples of frames are business and population registers; maps where the land has been divided into areas with strictly defined boundaries; or all n-digit numbers, which can be used to link telephone numbers to individuals. Sometimes the frame has to be developed for the occasion because there are no registers available and the elements have to be listed. For general populations this is done by combining a multi-stage sampling and the listing procedure by letting the survey field staff list all elements in sampled areas only. Other alternatives would be too costly. For special populations, for example, the population of professional baseball players in the USA, one would have to combine all club rosters into one frame. In some surveys there might exist a number of frames covering the population to varying extents. For this situation a multiple frame theory has been developed (see Hartley 1974).

(e) A sample of sampling units is selected from the frame in accordance with a sampling design, which specifies a probability mechanism and a sample size. There are numerous sample designs (see *Sample Surveys: Methods*) developed for different survey situations. The situation may be such that the design chosen solves a problem (using multistage sampling when not all population elements can be listed, or when interviewer and travel costs prevent the use of simple random sampling of elements) or takes advantage of the circumstances (using systematic sampling, if the population is approximately ordered, or using stratified sampling if the population is skewed). Every sample design specifies selection probabilities and a sample size. It is imperative that selection probabilities are known, or else the design is nonmeasurable.

(f) Observations are made on the sample in accordance with a measurement design i.e., a measurement method and a prescription as to its use. This phase is called data collection. There are at least five different main modes of data collection: face-to-face interviewing, telephone interviewing, self-administered questionnaires and diaries, administrative records, and direct observation. Each of these modes can be conducted using different levels of technology. Early attempts using the computer took place in the 1970s, in telephone interviewing. The questionnaire was stored in a computer and a computer program guided the interviewer throughout the interview by automatically presenting questions on the screen and taking care of some interviewer tasks such as keeping track of skip patterns and personalizing the interview. This technology is called CATI (Computer Assisted Telephone Interviewing). Current levels of technology for the other modes include the use of portable computers for face to face interviewing, touch-tone data entry using the telephone key pad, automatic speech recognition, satellite images of land use and crop yields, 'people meters' for TV viewing behaviors, barcode scanning in diary surveys of purchases, electronic exchange of administrative records, and Internet. Summaries of these developments are provided in Lyberg and Kasprzyk (1991), DeLeeuw and Collins (1997), Couper et al. (1998) and Dillman (2000). Associated with each mode is the survey measurement instrument or questionnaire. The questionnaire is the result of a conceptualization of research objectives i.e., a set of properly worded and properly ordered questions. The design of the questionnaire is a science of its own. See for example Tanur (1992) and Sudman et al. (1996).

(g) Based on the measurements an estimation design is applied to compute estimates of the parameters when making inference from the sample to the population. Associated with each sampling design are one or more estimators that are functions of the data that have been collected to make statements about the population parameters. Sometimes estimators rely solely on sample data, but on other occasions auxiliary information is part of the function. All estimators include sample weights that are used to inflate the sample data. To calculate the error of an estimate, variance estimators are formed, which makes it possible to calculate standard errors and eventually confidence intervals. See Cochran (1977) and Särndal et al. (1992) for comprehensive reviews of the sampling theory.

2. *The Status of Survey Research*

There are many types of surveys and survey populations that fit this definition. A large number of surveys are one-time surveys aiming at measuring attitudes or other population behaviors. Some surveys are continuing, thereby allowing the estimation of change over time. An example of this is a monthly labor force survey. Typically such a survey uses a rotating design where a sampled person is interviewed a number of times. One example of this is that the person participates 4 months in a row, is rotated out of the sample for the next 4 months and then rotates back for a final 4 months. Other surveys aim at comparing different populations regarding a certain characteristic, such as the literacy level in different countries. Business

surveys often study populations where there are a small number of large businesses and many smaller ones. In the case where the survey goal is to estimate a total, it might be worthwhile to deliberately cut off the smallest businesses from the frame or select all large businesses with a probability of one and the smaller ones with other probabilities.

Surveys are conducted by many different organizations. There are national statistical offices producing official statistics, there are university-based organizations conducting surveys as part of the education and there are private organizations conducting surveys on anything ranging from official statistics to marketing. The survey industry employs more than 130,000 people only in the USA, and the world figure is of course much larger. Survey results are very important to society. Governments get continuing information on parameters like unemployment, national accounts, education, environment, consumer price indexes, etc. Other sponsors get information on e.g., political party preferences, consumer satisfaction, child day-care needs, time use, and consumer product preferences.

As pointed out by Groves (1989), the field of survey sampling has evolved through somewhat independent and uncoordinated contributions from many disciplines including statistics, sociology, psychology, communication, education and marketing research. Representatives of these disciplines have varying backgrounds and as a consequence tend to emphasize different design aspects. However, during the last couple of decades, survey research groups have come to collaborate more as manifested by, for instance, the edited volumes such as Groves et al. (1988), Biemer et al. (1991), Lyberg et al. (1997), and Couper et al. (1998). This teamwork development will most likely continue. Many of the error structures resulting from specific sources must be dealt with by multi-disciplinary teams since the errors stem from problems concerning sampling, recall, survey participation, interviewer practices, question comprehension, and conceptualization. (See *Sample Surveys: Cognitive Aspects of Survey Design*.)

The justification for sampling (rather than surveying the entire population, a total enumeration) is lower costs but also greater efficiency. Sampling is faster and less expensive compared to total enumeration. Perhaps more surprisingly, sampling often allows a more precise measurement of each sampled unit than that possible in a total enumeration. This often leads to sample surveys having quality features that are superior to those of total enumerations.

Sampling as an intuitive tool has probably been used for centuries, but the development of a theory of survey sampling did not start until the late 1800s. Main contributors to this early development, frequently referred to as 'the representative method,' were Kiaer (1897), Bowley (1913, 1926), and Tschuprow (1923). Apart from various inferential aspects they discussed issues such as stratified sampling, optimum allocation to strata, multistage sampling, and frame construction. In the 1930s and the 1940s most of the basic methods that are used today were developed. Fisher's randomization principle was applied to sample surveys and Neyman (1934, 1938) introduced the theory of confidence intervals, cluster sampling, ratio estimation, and two-phase sampling. The US Bureau of the Census was perhaps the first national statistical office to embrace and further develop the theoretical ideas suggested. For example, Morris Hansen and William Hurwitz (1943, 1949) and Hansen et al. (1953) helped place the US Labor Force Survey on a full probability-sampling basis and they also led innovative work on variance estimation and the development of a survey model decomposing the total survey mean squared error into various sampling and bias components. Other important contributions during that era include systematic sampling (Madow and Madow 1944), regression estimation (Cochran 1942), interpenetrating samples (Mahalanobis 1946) and master samples (Dalenius 1957). More recent efforts have concentrated on allocating resources to the control of various sources of error i.e., methods for total survey design, taking not only sampling but also nonsampling errors into account. A more comprehensive review of historical aspects are provided in *Sample Surveys, History of*.

3. *The Use of Models*

While early developments focused on methods for sample selection in different situations and proper estimation methods, later developments have to a large extent focused on theoretical foundations and the use of probability models for increasing the efficiency of the estimators. There has been a development from implicit modeling to explicit modeling.

The model traditionally used in the early theory is based on the view that what is observed for a unit in the population is basically a fixed value. This approach may be called the 'fixed population approach.' The stochastic nature of the estimators is a consequence of the deliberately introduced randomization among the population units. A specific feature of survey sampling is the existence of auxiliary information i.e., known values of a concomitant variable, which is in some sense related to the variable under study, so that it can be used to improve the precision of the estimators. The relationship between the variable under study and the auxiliary variables are often expressed as linear regression models, which often can be interpreted as expressing the belief (common or the sampler's own) concerning the structure of the relationship between the variables. Such modeling is used extensively in early textbooks, see Cochran 1953. A somewhat different approach is to view the values of the variables

as realizations of random variables using probability models. In combination with the randomization of the units this constitutes what is called the superpopulation approach. Model based inference is used to draw conclusions based solely on properties of the probability models ignoring the randomization of the units.

Design based inference on the other hand ignores the mechanism that generated the data and concentrates on the randomization of the units. In general, model-based inference for estimating population parameters like means of subgroups can be very precise if the model is true but may introduce biased estimates if the model is false, while design-based inference leads to unbiased, but possibly inferior estimates of the population parameters. Model assisted inference is a compromise that aims at utilizing models in such a way that, if the model is true the precision is high, but if the model is false the precision will be no worse than if no model had been used. (See *Sample Surveys: Model-based Approaches*.)

As an example, we want to study a population of families in a country. We want to analyse the structure of disposable income for the households and find out the relation between factors like age, sex, education, the number of household members, and the disposable income for a family. A possible model of the data generating process could be that the disposable income is a linear function of these background variables. There is also an element of unexplained variation between families having the same values of the background variables. Also, the income will fluctuate from year to year depending on external variation in society. All this shows that the data generating process could be represented by a probability model where the disposable income is a linear function of background variables and random errors over time and between families. The super population model would be the set of models describing how the disposable income is generated for the families. For inferential purposes, a sample of families is selected. Different types of inference can be considered. For instance we might be interested in giving a picture of the actual distribution of the disposable income in the population at the specific time when we selected the sample, or we might be interested in estimating the coefficients of the relational model either because we are genuinely interested in the model itself e.g., for prediction of a future total disposable income for the population, which would be of interest for sociologists, economists and decision makers, or for using the model as a tool for creating more efficient estimators of the fixed distribution, given for example that the distribution of sex and age is known with reasonable accuracy in the population and can be used as auxiliary information. Evidently, the results would depend on the constellation of families comprising our sample. If we use a sample design that over-represents the proportion of large households or young households with small children, compared to the population, the inference based on the sample can be misleading. Model-based inference ignores the sample selection procedure and assumes that the inference conditional on the sample is a good representation of what would have been the case if all families had been surveyed. Design-based inference ignores the data generation process and concentrates on the artificial randomisation induced by the sampling procedure. Model-assisted inference uses models as tools for creating more precise estimates. Broadly speaking, model-based inference is mostly used in the case when the relational model is of primary interest. This is the traditional way of analysing sample data as given in textbooks in statistical theory. Design-based inference on the other hand is the traditional way of treating sample data in survey sampling. It is mainly focused on giving a picture of the present state of the population. Model-assisted inference uses models as tools for selecting estimators, but relies on design properties. It too is mainly focused on picturing the present state in the population.

Modern textbooks such as Cassel et al. (1977) and Särndal et al. (1992) discuss the foundations of survey sampling and make extensive use of auxiliary information in the survey design. The different approaches mentioned above have their advocates, but most of the surveys conducted around the world still rely heavily on design-based approaches with implicit modeling. But models are needed to take nonsampling errors into account since we do not know exactly how such errors are generated. To make measurement errors part of the inference procedure, one has to make assumptions about the error structures. Such error structures concern cognitive issues, question wording and perception, interviewer effects, recall errors, untruthful answers, coding, editing and so on and so forth. Similarly, to make errors of nonobservation (frame coverage and nonresponse errors) part of the inference procedure, one needs to model the mechanisms that generate these errors. The compromise called model-assisted inference takes advantage of both design and model- based features.

Analysis of data from complex surveys denotes the situation that occurs when the survey statistician is trying to estimate the parameters of a model used for description of a random phenomenon, for example econometric or sociological models such as time series models, regression models or structural equation models. It is assumed that the data available are sample survey data that have been generated by some sampling mechanism that does not support the assumption of independent identically distributed (IID) observations on a random variable. The traditional inference developed for the estimation of parameters of the model (and not for estimating the population parameters) presupposes that IID is at hand. In some cases, traditional inference based on e.g., maximum likelihood gives misleading results. Comprehensive reviews of analysis of data from complex surveys are

provided by Skinner et al. (1989) and Lehtonen and Pahkinen (1995).

The present state of affairs is that there is a relatively well-developed sampling theory. The theory of nonsampling errors is still in its infancy, however. A typical scenario is that survey methodologists try to reduce potential errors by using for example, cognitively tested questionnaires and various means to stimulate survey participation, and these things are done to the extent available resources permit. However, not all nonsampling error sources are known and some that are known defy expression. The error reduction strategy can be complemented by sophisticated modeling of error structures. Unfortunately, a rather common implicit modeling seems to be that nonsampling errors have no serious effect on estimates. In some applications, attempts are made to estimate the total error or error components by evaluation techniques, i.e., for a subsample of the units, the survey is replicated using expensive 'gold standard' methods and the differences between the preferred measurements and the regular ones are used as estimates of the total errors. This is an expensive and time-consuming procedure that is not very suitable for long-range improvements. A more modern and realistic approach is to develop reliable and predictable (stable) survey processes that can be continuously improved (Morganstein and Marker 1997).

4. Conclusions

Obviously there are a number of future challenges in the field of survey sampling. We will provide just a few examples: (a) Many surveys are conducted in a primitive way because of limited funding and know-how. The development of more efficient designs taking nonsampling errors into account at the estimation stage is needed. There is also a need for strategies that can help allocate resources to various design stages so that total errors are minimized. Sometimes those in charge of surveys concentrate their efforts on the most visible error sources or where there is a tool available. For instance, most survey sponsors know that nonresponse might be harmful. The indicator of nonresponse error, the nonresponse rate, is both simple and visible. Therefore it might be tempting to put most resources into this error source. On the other hand, not many users are aware of the cognitive phenomena that affect the response delivery mechanism. Perhaps from a total error point of view more resources should be spent on questionnaire design. (b) Modern technology permits simultaneous use of multiple data collection modes within a survey. Multiple modes are used to accommodate respondents, to increase response rates and to allow inexpensive data collection when possible. There are, however, mode effects and there is a need for calibration techniques that can adjust the measurements or the collection instruments so that the mode effect vanishes. (c) International surveys are becoming increasingly important. Most methodological problems mentioned are inflated under such circumstances. Especially interesting is the concept of cultural bias. Cultural bias means that concepts and procedures are not uniformly understood, interpreted, and applied across geographical regions or ethnical subpopulations. To define and measure the impact of such bias is an important challenge.

See also: Databases, Core: Demography and Registers; Databases, Core: Political Science and Political Behavior; Databases, Core: Sociology; Microdatabases: Economic; Survey Research: National Centers

Bibliography

Biemer P, Groves R, Lyberg L, Mathiowetz N, Sudman S 1991 *Measurement Errors in Surveys*. Wiley, New York

Bowley A L 1913 Working-class households in reading. *Journal of the Royal Statistical Society* **76**: 672–701

Bowley A L 1926 Measurement of the precision attained in sampling. *Proceedings of the International Statistical Institute* **XII**: 6–62

Cassel C-M, Särndal C-E Wretman J 1977 *Foundations of Inference in Survey Sampling*. Wiley, New York

Cochran W G 1942 Sampling theory when the sampling-units are of unequal sizes. *Journal of the American Statistical Association* **37**: 199–212

Cochran W G 1953 *Sampling Techniques*, 1st edn. Wiley, New York

Cochran W G 1977 *Sampling Techniques*, 3rd edn. Wiley, New York

Couper M, Baker R, Bethlehem J, Clark C, Martin J, Nicholls W, O'Reilly J 1998 *Computer Assisted Survey Information Collection*. Wiley, New York

Dalenius T 1957 *Sampling in Sweden*. Almqvist and Wiksell, Stockholm, Sweden

Dalenius T 1985 *Elements of Survey Sampling*. Notes prepared for the Swedish Agency for Research Cooperation with Developing Countries (SAREC)

DeLeeuw E, Collins M 1997 Data collection methods and survey quality: An overview. In: Lyberg L, Biemer P, Collins M, DeLeeuw E, Dippo C, Scwarz N, Trewin D (eds.) *Survey Measurement and Process Quality*. Wiley, New York

Dillman D 2000 *Mail and Internet Surveys: The Tailored Design Method,* 2nd edn, Wiley, New York

Groves R 1989 *Survey Errors and Survey Costs*. Wiley, New York

Groves R, Biemer P, Lyberg L, Massey J, Waksberg J (eds.) 1988 *Telephone Survey Methodology*. Wiley, New York

Hansen M H, Hurwitz W N 1943 On the theory of sampling from finite populations. *Annals of Mathematical Statistics* **14**: 333–62

Hansen M H, Hurwitz W N 1949 On the determination of optimum probabilities in sampling. *Annals of Mathematical Statistics* **20**: 426–32

Hansen M H, Hurwitz W N, Madow W G 1953 *Sample Survey Methods and Theory (I and II)*. Wiley, New York
Hartley H O 1974 Multiple methodology and selected applications. *Sankhya, Series C* **36**: 99–118
Kiaer A N 1897 The representative method for statistical surveys (original in Norwegian). Kristiania Videnskabsselskabets Skrifter. *Historisk-filosofiske klasse* **4**: 37–56
Lehtonen R, Pahkinen E J 1995 *Practical Methods for Design and Analysis of Complex Surveys*. Wiley, New York
Lyberg L, Biemer P, Collins M, DeLeeuw E, Dippo, C, Schwarz N, Trewin D (eds.) 1997 *Survey Measurement and Process Quality*. Wiley, New York
Lyberg L, Kasprzyk D 1991 Data collection methods and measurement error: An overview. In: Biemer P, Groves R, Lyberg L, Mathiowetz N, Sudman S (eds.) *Measurement Errors in Surveys*. Wiley, New York
Madow W G, Madow L H 1944 On the theory of systematic sampling, I. *Annals of Mathematical Statistics* **15**: 1–24
Mahalanobis P C 1946 On large-scale sample surveys. *Philosophical Transactions of the Royal Society London, Series B* **231**: 329–451
Morganstein D, Marker D 1997 Continuous Quality Improvement in Statistical Agencies. In Lyberg L, Biemer P, Collins M, DeLeeuw E, Dippo C, Scwarz N, Trewin D (eds.) *Survey Measurement and Process Quality*. Wiley, New York
Neyman J 1934 On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**: 558–625
Neyman J 1938 Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* **33**: 101–16
Särndal C-E, Swensson B, Wretman J 1992 *Model Assisted Survey Sampling*. Springer-Verlag, New York
Skinner C, Holt D, Smith T M F (eds.) 1989 *Analysis of Complex Surveys*. Wiley, New York
Sudman S, Bradburn N, Schwarz N 1996 *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco, CA
Tanur J (ed.) 1992 *Questions About Questions*. Russell Sage, New York
Tschuprow A A 1923 On the mathematical expectation of the moments of frequency distributions in the case of correlated observation. *Metron* **2**: 461–93; 646–80

L. Lyberg and C. M. Cassel

Sanctions in Political Science

A sanction is an action by one actor (A) intended to affect the behavior of another actor (B) by enhancing or reducing the values available to B. Influence attempts by A using actual or threatened punishments of B are instances of negative sanctions. Influence attempts by A using actual or promised rewards to B are instances of positive sanctions. Not all influence attempts involve sanctions. Actor A may influence actor B by reason, example, or the provision of information without the use of sanctions.

1. *Concepts*

Although the definitions of positive and negative sanctions may appear simple, there are both conceptual and empirical difficulties in distinguishing between the two. Some things take the form of positive sanctions, but actually are not; e.g., giving a bonus of $100 to an employee who expected a bonus of $500, or promising not to kill a person who never expected to be killed in the first place. Likewise, some things take the form of negative sanctions, but actually are not; e.g., a threat to reduce the salary of a person who expected to be fired or the beating of a masochist.

Is withholding a reward ever a punishment? Always a punishment? Is withholding a punishment ever a reward? Always a reward? The answers depend on Actor B's perception of the situation. In order to distinguish rewards from punishments one must establish B's baseline of expectations at the moment A's influence attempt begins (Blau 1986). This baseline is defined in terms of B's expected future value position, i.e., expectations about B's future position relative to the things B values. Positive sanctions, then, are actual or promised improvements in B's value position relative to B's baseline of expectations, and negative sanctions are actual or threatened deprivations relative to the same baseline.

2. *Historical Context*

Although references to both positive and negative sanctions can be traced to the seventeenth century, the most common usage of the term 'sanctions' refers to negative sanctions. Until the mid-twentieth century, sanctions were viewed primarily as mechanisms for enforcing societal norms, including those embodied in laws. Although this normative view of sanctions continues in legal theory, ethics, political theory, and sociology (Barry 1995, Coleman 1990, Waldron 1994), a broader usage has emerged in political science during the latter part of the twentieth century. This broader usage depicts sanctions as potentially relevant to any type of influence attempt, regardless of whether it is aimed at enforcing social norms or not. This usage is typically found in discussions of influence and power (Baldwin 1971, 1989, Blau 1986, Oppenheim 1981, Lasswell and Kaplan 1950).

3. *Kinds of Sanctions*

Since sanctions are defined in terms of the values of the target of an influence attempt (actor B), they may take a variety of forms. Lasswell and Kaplan (1950) identified eight exemplary values that could serve as the basis for positive or negative sanctions: power, wealth, rectitude, physical well-being, affection, skill, and enlightenment.

Although it is sometimes suggested that positive and negative sanctions are opposites in the sense that generalizations about one type are equally applicable to the other, *mutatis mutandis*, this is often not true. Listed below are some of the hypothesized differences between positive and negative sanctions.

3.1 A's Burden of Response

When A's influence attempt is based on a promise, B's compliance obligates A to respond with a reward; whereas B's failure to comply calls for no further response from A.

3.2 Role of Costs

One important consequence of the asymmetry between positive and negative sanctions is that promises tend to cost more when they succeed, while threats tend to cost more when they fail (Boulding 1989, Parsons 1963, Schelling 1960). The difference can be summarized as follows: The bigger the threat, the higher the probability of success; the higher the probability of success, the less the probability of having to implement the threat; the less the probability of having to implement the threat, the cheaper it is to make big threats. The bigger the promise, the higher the probability of success; the higher the probability of success, the higher the probability of having to implement the promise; the higher the probability of having to implement the promise, the more expensive it is to make big promises (*ceteris paribus*).

3.3 Indicators of Success

Whereas a successful threat requires no action by A, a successful promise obligates A to implement that sanction. In a well-integrated social system, where the probability that B will comply with A's wishes is relatively high, promises will be more visible than threats. Indeed, it is precisely because threats are so successful in domestic politics that they are so difficult to detect. Since most citizens obey most laws most of the time, threats to punish lawbreakers have to be carried out with respect to only a small minority of citizens in most polities.

3.4 Efficacy

The question of the relative efficacy of positive versus negative sanctions in exercising influence has been much debated. Parson (1963) has argued that negative sanctions have more intrinsic effectiveness than positive sanctions in deterrence situations. And ever since Machiavelli and Hobbes, there have been those who argue that force is the ultimate influence technique. Identifying the conditions under which one type of sanction is more effective than another is likely to continue as a focus of research in political science as well as in other social sciences.

3.5 Legitimation

It is usually easier to legitimize demands based on positive sanctions than demands based on negative ones. An example is provided by one of the most important social institutions in the world—the institution of private property. Most societies have one set of rules specifying the conditions under which a person may be deprived of property and quite a different set of rules specifying conditions under which one person may augment another person's property.

4. Foreign Policy Sanctions

Political scientists interested in international relations and foreign policy have devoted a great deal of research to sanctions during the last two decades of the twentieth century. Most of this research has focused on economic sanctions, with comparatively little attention devoted to military, diplomatic, or other non-economic forms of sanctions. Although progress has occurred with respect to systematic gathering of empirical data (Hufbauer et al. 1990), refinement of concepts (Baldwin 1985), and development of theories, social scientific research on the role of sanctions in international relations is still in its infancy.

4.1 Methodological Problems

Five methodological problems provide useful foci for future research:

(a) *Lack of agreement on terms and concepts*. While some scholars conceive of economic sanctions in terms of mechanisms for making any type of influence attempt (Baldwin 1985), others define economic sanctions in terms of particular policy goals (Pape 1997). This lack of agreement on basic concepts not only impedes debate within the field of foreign policy studies, it is an obstacle to crossfertilization between the study of economic sanctions in foreign policy and discussions of other kinds of sanctions in other areas of social science.

(b) *Preoccupation with economic sanctions*. Research on economic sanctions has been largely insulated from research on other techniques of statecraft. This insulation sometimes gives the impression that economic sanctions are *sui generis*, rather than treating them as a particular type of statecraft. Although actual or threatened military force is a military sanction, it too is usually viewed as a unique focus of

inquiry. In order to understand when and why policy makers use economic sanctions, it is necessary to understand the alternatives among which policy makers choose. These alternatives include diplomatic, military, and/or verbal sanctions (propaganda). Future research on sanctions might usefully focus on comparative studies of alternative types of sanctions and the conditions under which each type is most likely to be cost-effective.

(c) *Neglect of positive sanctions*. Research on sanctions is heavily skewed toward negative sanctions. This is true not only of research by international relations scholars, but also of political science in general. During the last two decades, research on positive sanctions has increased (Davis 2000, Newnham 2000), but the emphasis is still disproportionately on negative sanctions.

(d) *Lack of agreement on criteria of success*. Scholars disagree as to which criteria should be used in measuring the success of economic sanctions. Some consider only the effectiveness of the influence attempt in achieving its goals (Morgan and Schwebach 1997, Pape 1997). Others argue that estimates of success should take into consideration the costs of the sanctions to the user, the costs for noncompliance inflicted on the target, the difficulty of the undertaking, and the comparative utility of alternative policy options (Baldwin 1985, 2000). This lack of agreement on how to conceive of success is not peculiar to the international relations literature. Unlike economics, political science lacks a standardized measure of value in terms of which success can be measured. There is much less agreement among political scientists as to what constitutes political success than there is among economists as to what constitutes economic success (Baldwin 1989). Developing a set of agreed upon criteria of success applicable to noneconomic, as well as economic, sanctions would be a valuable step in sanctions research.

(e) *Premature generalization*. Research on economic sanctions by international relations scholars frequently leads to premature attempts to specify policy implications. Some scholars, for example, imply that only foolish policy makers would use such policy instruments (Morgan and Schwebach 1997, Tsebelis 1990). In order to make judgements as to the wisdom of using economic sanctions in a given situation, however, one would need to know how effective such sanctions were likely to be, with respect to which goals and targets, at what cost, and in comparison with which policy alternatives (Baldwin 2000). Current research on economic sanctions rarely asks such questions.

See also: Control: Social; Deterrence; Diplomacy; Efficacy: Political; Foreign Policy Analysis; Game Theory; Law: Economics of its Public Enforcement; Legal Culture and Legal Consciousness; Legitimacy: Political; National Security Studies and War Potential of Nations; Norms; Power: Political; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Utilitarianism: Contemporary Applications

Bibliography

Baldwin D A 1971 The power of positive sanctions. *World Politics* **24**: 19–38

Baldwin D A 1985 *Economic Statecraft*. Princeton University Press, Princeton, NJ

Baldwin D A 1989 *Paradoxes of Power*. Basil Blackwell, New York

Baldwin D A 2000 The sanctions debate and the logic of choice. *International Security* **24**: 80–107

Barry B 1995 *Justice as Impartiality*. Clarendon Press, Oxford, UK

Blau P M 1986 *Exchange and Power in Social Life*. Transaction, New Brunswick, NJ

Boulding K E 1989 *Three Faces of Power*. Sage, Newbury, CA

Coleman J S 1990 *Foundations of Social Theory*. Belknap, Cambridge, MA

Davis J W 2000 *Threats and Promises: The Pursuit of International Influence*. The Johns Hopkins University Press, Baltimore, MD

Hufbauer G C, Schott J J, Elliott K A 1990 *Economic Sanctions Reconsidered*, 2nd edn. Institute for International Economics, Washington, DC

Lasswell H D, Kaplan A 1950 *Power and Society: A Framework for Political Inquiry*. Yale University Press, New Haven, CT

Morgan T C, Schwebach V L 1997 Fools suffer gladly: The use of economic sanctions in international crises. *International Studies Quarterly* **41**: 27–50

Newnham R E 2000 More flies with honey: Positive economic linkage in German Ostpolitik from Bismarck to Kohl. *International Studies Quarterly* **44**: 73–96

Oppenheim F E 1981 *Political Concepts: A Reconstruction*. University of Chicago Press, Chicago

Pape R A 1997 Why economic sanctions do not work. *International Security* **22**: 90–136

Parsons T 1963 On the concept of political power. *Proceedings of the American Philosophical Society* **107**: 232–62

Schelling T C 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA

Tsebelis G 1990 Are sanctions effective? A game-theoretic analysis. *Journal of Conflict Resolution* **34**: 3–28

Waldron J 1994 Kagan on requirements: Mill on sanctions. *Ethics* **104**: 310–24

D. A. Baldwin

Sapir, Edward (1884–1939)

1. General Career

Edward Sapir was born in Lauenburg, Germany (now Lebork, Poland) on January 26, 1884, but his family emigrated to the United States when he was five years old and eventually settled in New York City. His

brilliance earned him a full scholarship to the prestigious Horace Mann School and subsequently a Pulitzer fellowship to Columbia College, where he received his BA in 1904. He continued with graduate work in Germanic philology, but was soon drawn into Franz Boas's orbit and took up an anthropological career, for which he proved extraordinarily well fitted. In 1905 he began a series of field visits to the West that resulted in the detailed documentation of several American Indian languages and cultures, including Wishram Chinook, Takelma, Yana, Southern Paiute, and Nootka. The speed and accuracy with which Sapir collected linguistic and ethnographic data has probably never been surpassed, and his notes, even from his earliest field trips, are among the most valuable manuscripts in American Indian studies.

After receiving his doctorate from Columbia in 1909 with a dissertation on the Takelma language, Sapir taught briefly at the University of Pennsylvania, and in 1910 became Chief of the Division of Anthropology in the Geological Survey of Canada. He remained in this position until 1925, making the Nootka (Nuuchahnulth) language the focus of his research from 1910 to 1914. Between 1915 and 1920, when World War I and its aftermath brought field work to a halt, he devoted a considerable amount of time to the comparative linguistics of North American languages, establishing the Na-Dene relationship between Athabaskan, Tlingit, and Haida, and proposing expansions of the Penutian and Hokan stocks to include a large number of languages in North and Central America. During this period he also began making a name for himself as a literary and social commentator, and began publishing his experimental poetry.

Sapir married Florence Delson in 1911, and they had three children. Shortly after the birth of their third child in 1918, Florence Sapir began to manifest signs of serious illness, partly psychological in character. Although his wife's deteriorating health became a great concern to Sapir, and her death in 1924 was emotionally devastating, he remained committed to the intensive field documentation of American Indian languages, and in 1922 commenced an Athabaskan research program that eventually encompassed full-scale descriptive studies of Sarsi, Kutchin (Gwich'in), Hupa, and Navajo. This work was motivated partly by Sapir's conviction that a historical relationship could be demonstrated between the Na-Dene languages of North America and the Sino-Tibetan languages of East Asia.

After the publication of his highly influential book *Language* (1921) Sapir was regarded widely as one of the leading linguists of his generation. In 1925 he accepted Fay-Cooper Cole's offer of an academic position in the newly reorganized Department of Anthropology and Sociology at the University of Chicago. The move from Ottawa, and his second marriage to Jean McClenaghan, with whom he was to have two more children, was an emotional and intellectual watershed. He found the interdisciplinary atmosphere at Chicago stimulating and increasingly he addressed general and theoretical topics in his professional writing. He also began contributing sprightly and provocative essays on general social and cultural topics to such publications as Mencken's *American Mercury*. While much of his teaching was in linguistics and he took a leading role in establishing that discipline as an autonomous field of study, Sapir also became involved in developing a general model for social science. After he moved to Yale in 1931 as Sterling Professor of Anthropology and Linguistics he became interested particularly in a psychologically realistic paradigm for social research (see Irvine 1999) and led a seminar on personality and culture. Meanwhile linguistic work on Athabaskan, specifically Navajo, continued to absorb him, and the prospect of finding remote connections of the Na-Dene languages in Asia had led him to the study of Tibetan and Chinese.

This extraordinarily diverse agenda was abruptly suspended in the summer of 1937 when Sapir suffered a serious heart attack, from which he never fully recovered. After a year and a half of precarious health and restricted activity he died in New Haven, Connecticut on February 4, 1939, a few days after his 55th birthday.

2. *Scientific Contribution*

Sapir's scientific work can be divided into three distinct parts. First, there is his substantive work in descriptive and comparative linguistics, almost entirely devoted to North American Indian languages. Second, there is his role in establishing the paradigm for twentieth century linguistic research. Finally, there are the flashes of insight—seldom elaborated into formal hypotheses—with which Sapir from time to time illuminated the landscape of linguistic and social theory.

2.1 *Substantive Work on Languages*

The face that American Indian linguistics presents to the world at the beginning of the twenty-first century probably owes more to Sapir than to any other scholar. When Sapir took up the anthropological study of American Indian languages in 1905, the field was dominated by the classificatory concerns of John Wesley Powell's Bureau of American Ethnology, which saw as its principal task the identification of language and dialect boundaries, and the grouping of languages into families whose historical relationship was undoubted. Only a few earlier scholars like Pickering and Humboldt had been interested in the

general philological study of these languages. It was Franz Boas, Sapir's mentor at Columbia, who first proposed making comparative investigation of the grammatical structures of American Indian languages (and of non-European languages generally) a topic for sustained scientific research. Drawing his model from a German tradition of associating language and social behavior that led back through Steinthal to Humboldt, Boas impressed upon his anthropological students the necessity of understanding how linguistic 'morphology' (by which he meant grammatical structure) channeled ideas into expressive forms. Unfortunately for Boas, few of his students were equipped either by training or by intellectual inclination to carry out linguistic research of a more than superficial kind.

The only significant exception was Sapir, whom one may imagine was attracted to Boas's anthropology for precisely this reason. From his earliest fieldwork on Wishram Chinook in 1905, Sapir made grammatical analysis the centerpiece of his research, and from his first publications portrayed American Indian languages with a descriptive clarity they had seldom before enjoyed. His accomplishments are legendary to the scholars who today study these languages, and must rank in the first tier of the grammarian's art. The most renowned of his published studies include a full grammar of Takelma, a now-extinct language of southern Oregon (1922); a full grammar and dictionary of Southern Paiute (1930–31/1992); and an outline grammar of Nootka, with texts and vocabulary, prepared with the assistance of Swadesh (1939).

Sapir's grammatical descriptions are couched in a metalanguage derived from European comparative philology, tempered by Boas's insistence that the structure of every language is *sui generis*. In discussions of the evolution of grammatical theory, Sapir's grammars are sometimes portrayed as 'processual' or as early examples of 'generative' descriptions, but such labels imply a theoretical deliberation that was uncharacteristic of Sapir's work. His concern was to explicate the patterns of the language under consideration as lucidly and unambiguously as possible, not to test a general theory of linguistic structure.

The aptness with which Sapir's descriptions captured the spirit of the languages he analyzed is no better illustrated than by his model of Athabaskan grammar. Although in this case it was not embodied in a full grammatical treatment, he laid out the basic features of his descriptive system for Athabaskan in a number of shorter works, and passed it on to his students and successors through his teaching and in his files. Sixty years after his death it remains the standard descriptive model for all work in Athabaskan linguistics, regardless of the theoretical stance of the analyst.

Throughout his career Sapir maintained a deep interest in historical relationships among languages. He took pride in extending the rigor of the reconstructive method to American Indian language families, and laid the foundations for the comparative study of both Uto-Aztecan and Athabaskan. In the 1930s he returned to Indo-European linguistics and made major contributions to the Laryngeal Hypothesis (the proposal, originating with Ferdinand de Saussure, that the phonology of Proto-Indo-European included one or more laryngeal or pharyngeal consonants not attested in the extant Indo-European languages).

2.2 *The Professionalization of American Linguistics*

By 1921, Sapir was able to draw on the analytic details of the American Indian languages on which he had worked to illustrate, in his book *Language*, the wide variety of grammatical structures represented in human speech. One of the most significant impacts of this highly successful book was to provide a model for the professionalization of academic linguistics in the US during the 1920s and early 1930s.

Under Sapir's guidance, a distinctive American School of linguistics arose, focused on the empirical documentation of language, primarily in field situations. Although most of the students Sapir himself trained at Chicago and Yale worked largely if not exclusively on American Indian languages, the methods that were developed were transferable to other languages. In the mid-1930s Sapir directed a project to analyze English, and during World War II many of Sapir's former students were recruited to use linguistic methods to develop teaching materials for such strategically important languages as Thai, Burmese, Mandarin, and Russian.

A distinction is sometimes drawn between the prewar generation of American linguists—dominated by Sapir and his students and emphasizing holistic descriptions of American Indian languages—and the immediate postwar generation, whose more rigid and focused formal methods were codified by Leonard Bloomfield and others. Since many of the major figures of 'Bloomfieldian' linguistics were Sapir's students, this distinction is somewhat artificial, and a single American Structuralist tradition can be identified extending from the late 1920s through 1960 (Hymes and Fought 1981). There is little doubt that Sapir's influence on this tradition was decisive.

3. *Theoretical Insights*

From his earliest work under Boas, Sapir's independent intellectual style often carried him well beyond the bounds of the academic paradigm he was ostensibly working within. He was not, however, a programmatic thinker, and his groundbreaking work, while deeply admired by his students and colleagues, seldom resulted in significant institutional changes, at least in the short term. The cliché 'ahead of his time' is

especially apt in Sapir's case, and the influence of some of his views continues to be felt.

This is most striking in structural linguistics, where Sapir must certainly be accounted one of the most influential figures of the twentieth century. As early as 1910 Sapir was commenting on the importance of formal patterning in phonology, and in the 1920s he was among the first to enunciate the 'phonemic principle' which later figured importantly in his teaching at Chicago and Yale. Characteristically, he left it to his students and such colleagues as Leonard Bloomfield to formalize an analytic methodology, while he himself pursued the psychological implications of formal patterning.

Sapir's (1917) trenchant critique of the culture concept, particularly as defined by A. L. Kroeber, was largely ignored at the time. Sapir argued that the attribution of cultural patterning to an emergent 'superorganic' collective consciousness was an intellectual dead-end, and that research would be better directed at understanding the individual psychology of collective patterned behavior. While Kroeber's view was undoubtedly the dominant one in anthropology and general social science for much of the twentieth century, Sapir's analysis is much more consistent with recent models of human sociocultural behavior that have been developed by evolutionary psychologists. The fact that Sapir's views came from someone with extraordinary insight into the elaborate self-referential patterns of language—traditionally, the most formalized and objectified of social behaviors—can hardly be accidental, and again is consistent with recent developments in cognitive science.

In the late 1920s Sapir found an important intellectual ally in Harry Stack Sullivan, a psychiatrist whose interpersonal theory of the genesis of schizophrenia resonated with Sapir's views (Perry 1982). The two became close friends, and together with Harold Lasswell they planned and organized the William Alanson White Psychiatric Foundation, a research and teaching institution that ultimately was located in Washington, DC. In the year before he died Sapir gave serious consideration to leaving Yale and working with Sullivan in a research position at the Foundation.

Sapir's views on the history of language were linked to his view of the abstraction of patterns, and were equally controversial or misunderstood. A distinction must be drawn between Sapir's work, noted earlier, as a historical linguist within a family of languages whose relationship was secure (Athabaskan, Uto-Aztecan, Indo-European), and his explorations of much less certain relationships (Hokan, Penutian, Na-Dene) and possible interhemispheric connections (Sino-Dene). In the former, Sapir worked—with characteristic creativity and insight—with tools and models derived from a long tradition of comparative linguistics. In the latter, it was (or seemed to his contemporaries) often a matter of brilliant intuition. In fact, in this work he usually relied on an assessment of similarities in structural pattern that distinguished features susceptible to unconscious change from generation to generation (e.g., regular inflectional patterns, words) from those that are largely inaccessible to individual cognition. Although he never offered a theoretical explication, his idea of what constituted 'deep' linguistic patterns was exemplified in his classification of North and Central American Indian languages (1929). This classification, which remains influential, is still untested in its own terms.

4. *Impact and Current Importance*

A charismatic teacher, Sapir had a succession of highly motivated students both at Chicago and at Yale, the most prominent among them Morris Swadesh (who collaborated with him on Nootka research), Harry Hoijer (who codified Sapir's analysis of Athabaskan grammar), Mary R. Haas, Stanley Newman, C. F. Voegelin, George L. Trager, Zellig Harris, David G. Mandelbaum, and Benjamin L. Whorf. Through these students Sapir exercised a considerable posthumous influence on intellectual and institutional developments in both linguistics and in anthropology through the 1960s. A postwar collection of Sapir's most important general papers, edited by Mandelbaum (1949), was widely read and is still consulted. Harris' (1951) extended review of this book provides a comprehensive summary of Sapir's *oeuvre* as it was understood by his immediate circle.

Sapir is cited most frequently today for the 'Sapir–Whorf Hypothesis' of linguistic relativity, a name that inaccurately implies an intellectual collaboration between Sapir and his Yale student, Benjamin Whorf, who himself died in 1941. The more correctly designated 'Whorf theory complex' (Lee 1996) was a retrospective construct that derived largely from writings of Whorf's that were unpublished at the time of Sapir's death. Although undoubtedly stimulated by Sapir's writing and teaching, Whorf's proposal that the structure of a language to some extent determines the cognitive and behavioral habits of its speakers cannot be connected directly with Sapir's mature thought on the psychology of language and culture.

Sapir's most enduring achievement is his own descriptive linguistic work. Long after the writings of most of his contemporaries have been forgotten, Sapir's grammatical studies continue to be held in the highest esteem. In recent years his holistic analytic technique has been emulated by a number of linguists seeking an alternative to narrow formalism, particularly when working with American Indian or other indigenous languages.

Sapir's life and work was the subject of a 1984 conference (Cowan et al. 1986), from which emerged a plan to publish a standard edition of all of Sapir's work, including edited versions of unfinished manu-

scripts; by year 2000 seven volumes had appeared. A biography by Darnell (1990) is useful for the externals of Sapir's career, but her reluctance to give an intellectual account of Sapir's work, particularly in linguistics, leaves some important issues still to be addressed (Silverstein 1991).

See also: Archaeology and the History of Languages; Historical Linguistics: Overview; Language and Ethnicity; Language and Thought: The Modern Whorfian Hypothesis; Linguistic Anthropology; Linguistic Fieldwork; Linguistic Typology; Linguistics: Comparative Method; North America and Native Americans: Sociocultural Aspects; North America, Archaeology of; North America: Sociocultural Aspects; Phonology; Population Composition by Race and Ethnicity: North America; Psycholinguistics: Overview; Sapir–Whorf Hypothesis; Sociolinguistics

Bibliography

Cowan W, Foster M K, Koerner K (eds.) 1986 *New Perspectives in Language, Culture and Personality*. Benjamins, Amsterdam and Philadelphia

Darnell R 1990 *Edward Sapir: Linguist, Anthropologist, Humanist*. University of California Press, Berkeley and Los Angeles, CA

Harris Z S 1951 Review of D G Mandelbaum (ed.). Selected writings of Edward Sapir in language, culture, and personality. *Language* **27**: 288–333

Hymes D, Fought J 1981 *American Structuralism*. Mouton, The Hague, The Netherlands

Irvine J T (ed.) 1999 The psychology of culture: A course of lectures by Edward Sapir, 1927–1937. In: *The Collected Works of Edward Sapir, Vol. 3, Culture*. Mouton de Gruyter, Berlin and New York, pp. 385–686 [also published as a separate volume, *The Psychology of Culture*, Mouton de Gruyter 1993]

Lee P 1996 *The Whorf Theory Complex: A Critical Reconstruction*. Benjamins, Amsterdam and Philadelphia

Mandelbaum D M (ed.) 1949 *Selected Writings of Edward Sapir in Language, Culture, and Personality*. University of California Press, Berkeley and Los Angeles, CA

Perry H S 1982 *Psychiatrist of America: The Life of Harry Stack Sullivan*. Belknap Press, Cambridge, UK and London

Sapir E 1917 Do we need a 'superorganic'? *American Anthropologist* **19**: 441–47

Sapir E 1921 *Language*. Harcourt Brace, New York

Sapir E 1922 The Takelma language of Southwestern Oregon. In: Boas F (ed.) *Handbook of American Indian Languages, Part 2*. Bureau of American Ethnology, Washington, DC

Sapir E 1929 Central and North American Indian languages. *Encyclopaedia Britannica*, 14th edn. Vol. 5, pp. 138–41

Sapir E 1930–31/1992 *The Southern Paiute Language. The Collected Works of Edward Sapir*, Vol. 10. Mouton de Gruyter, Berlin and New York

Sapir E, Swadesh M 1939 *Nootka Texts: Tales and Ethnological Narratives with Grammatical Notes and Lexical Materials*. Linguistic Society of America, Philadelphia

Silverstein M 1991 Problems of Sapir historiography. *Historiographia Linguistica* **18**: 181–204

V. Golla

Sapir–Whorf Hypothesis

1. Nature and Scope of the Hypothesis

The Sapir–Whorf hypothesis, also known as the linguistic relativity hypothesis, refers to the proposal that the particular language one speaks influences the way one thinks about reality. Although proposals concerning linguistic relativity have long been debated, American linguists Edward Sapir (1884–1939) and Benjamin Lee Whorf (1897–1941) advanced particularly influential formulations during the second quarter of the twentieth century, and the topic has since become associated with their names. The linguistic relativity hypothesis focuses on structural differences among natural languages such as Hopi, Chinese, and English, and asks whether the classifications of reality implicit in such structures affect our thinking about reality more generally.

Analytically, linguistic relativity as an issue stands between two others: a semiotic-level concern with how speaking any natural language whatsoever might influence the general potential for human thinking (i.e., the general role of natural language in the evolution or development of human intellectual functioning), and a functional- or discourse-level concern with how using any given language code in a particular way might influence thinking (i.e., the impact of special discursive practices such as schooling and literacy on formal thought). Although analytically distinct, the three issues are intimately related in both theory and practice. For example, claims about linguistic relativity depend on understanding the general psychological mechanisms linking language to thinking, and on understanding the diverse uses of speech in discourse to accomplish acts of descriptive reference. Hence, the relation of particular linguistic structures to patterns of thinking forms only one part of the broader array of questions about the significance of language for thought.

Proposals of linguistic relativity necessarily develop two linked claims among the key terms of the hypothesis (i.e., language, thought, and reality). First, languages differ significantly in their *interpretations* of experienced reality—both what they select for representation and how they arrange it. Second, language interpretations have *influences* on thought about reality more generally—whether at the individual or cultural level. Claims for linguistic relativity thus require both articulating the contrasting interpretations of reality latent in the structures of different languages, and assessing their broader influences on, or relationships to, the cognitive interpretation of reality.

Simple demonstrations of *linguistic diversity* are sometimes mistakenly regarded as sufficient in themselves to prove linguistic relativity, but they cannot in themselves show that the language differences affect thought more generally. (Much confusion arises in this

regard because of the practice in linguistics of describing the meaningful significance of individual elements in a language as 'relative to' the grammatical system as a whole. But this latter relativity of the meaning of linguistic elements to the encompassing linguistic structure should be distinguished from broader claims for a relativity of thought more generally to the form of the speaker's language.) A variety of other arguments to the effect that distinctive perceptual or cognitive skills are required to produce and comprehend different languages likewise usually fail to establish any general effects on thought (see Niemeier and Dirven 2000).

Linguistic relativity proposals are sometimes characterized as equivalent to *linguistic determinism*, that is the view that all thought is strictly determined by language. Such characterizations of the language–thought linkage bear little resemblance to the proposals of Sapir or Whorf, who spoke in more general terms about language influencing habitual patterns of thought, especially at the conceptual level. Indeed, no serious scholar working on the linguistic relativity problem as such has subscribed to a strict determinism. (There are, of course, some who simply equate language and thought, but under this assumption of identity, the question of influence or determinism is no longer relevant.) Between the patent linguistic diversity that nearly everyone agrees exists and a claim of linguistic determinism that no one actually espouses, lies the proposal of linguistic relativity, that is the proposal that our thought may in some way be taken as relative to the language spoken.

2. *Historical Development of the Hypothesis*

Interest in the intellectual significance of the diversity of language categories has deep roots in the European tradition (Aarsleff 1988, Werlen 1989, Koerner 1992). Formulations related to contemporary ones appear during the Enlightenment period in the UK (Locke), France (Condillac, Diderot), and Germany (Hamman, Herder). They are stimulated variously by opposition to the universal grammarians, by concerns about the reliability of language-based knowledge, and by practical efforts to consolidate national identities and cope with colonial expansion. Most of this work construes the differences among languages in terms of a hierarchical scheme of adequacy with respect to reality, to reason, or to both. Later, nineteenth-century work in Germany by Humboldt and in France/Switzerland by Saussure drew heavily on this earlier tradition and set the stage for the approaches of Sapir and Whorf. Humboldt's arguments, in particular, are often regarded as anticipating the Sapir–Whorf approach. He argued for a linguistic relativity according to the formal processes used by a language (e.g., inflection, agglutination, etc.). Ultimately this remains a hierarchical relativity in which certain language types (i.e., European inflectional ones) are viewed as more adequate vehicles of thought and civilization—a view distinctly at odds with what is to follow.

Working within the US anthropological tradition of Franz Boas and stimulated by the diversity and complexity of Native American languages, Edward Sapir (1949) and Benjamin Lee Whorf (1956) reinvigorated and reoriented investigation of linguistic relativity in several ways (Lucy 1992a, Lee 1997). First, they advocated intensive first-hand scientific investigation of exotic languages; second, they focused on structures of meaning, rather than on formal grammatical process such as inflection; and third, they approached these languages within a framework of egalitarian regard. Although not always well understood, theirs is the tradition of linguistic relativity most widely known and debated today. Whorf's writings, in particular, form the canonical starting point for all subsequent discussion.

Whorf proposed a specific mechanism for how language influences thought, sought empirical evidence for language effects, and articulated the reflexive implications of linguistic relativity for scholarly thought itself. In his view, each language refers to an infinite variety of experiences with a finite array of formal categories (both lexical and grammatical) by grouping experiences together as analogically 'the same' for the purposes of speech. The categories in a language also interrelate in a coherent way, reinforcing and complementing one another, so as to constitute an overall interpretation of experience. These linguistic classifications vary considerably across languages not only in the basic distinctions they recognize but also in the assemblage of these categories into a coherent system of reference. Thus the system of categories which each language provides to its speakers is not a common, universal system, but a particular 'fashion of speaking.'

Whorf argued that these linguistic structures influence habitual thought by serving as a guide to the interpretation of experience. Speakers tend to assume that the categories and distinctions of their language are entirely natural and given by external reality, and thus can be used as a guide to it. When speakers attempt to interpret an experience in terms of a category available in their language, they unwittingly involve other language-specific meanings implicit in that particular category and in the overall configuration of categories in which it is embedded. In Whorf's view language does not blind speakers to some obvious reality, but rather it suggests associations which are not necessarily entailed by experience. Because language is such a pervasive and transparent aspect of behavior, speakers do not understand that the associations they 'see' are from language, but rather assume that they are 'in' the external situation and patently obvious to all. In the absence of another language (natural or artificial) with which to talk about experience, speakers will not be able to recognize

the conventional nature of their linguistically based understandings. Whorf argues that by influencing everyday habitual thought in this way, language can come to influence cultural institutions generally, including philosophical and scientific activity.

In his empirical research Whorf showed that the Hopi and English languages treat 'time' differently, and that this difference corresponds to distinct cultural orientations toward temporal notions. Specifically, Whorf argued that speakers of English treat cyclic experiences of various sorts (e.g., the passage of a day or a year) in the same grammatical frame used for ordinary object nouns. Thus, English speakers are led to treat these cycles as object-like in that they can be measured and counted just like tangible objects. English also treats objects as if they each have a form and a substance. Since the cyclic words get put into this object frame, English speakers are led to ask what is the substance associated with the forms *a day*, *a year*, and so forth. Whorf argues that our global, abstract notion of 'time' as a continuous, homogeneous, formless something can be seen to arise to fill in the blank in this linguistic analogy. The Hopi, by contrast, do not treat these cycles as objects but as recurrent events. Thus, although they have, as Whorf acknowledged, words for what English speakers would recognize as temporal cycles (e.g., days, years, etc.), the formal analogical structuration of these terms in their grammar does not give rise to the abstract notion of 'time' that English speakers have. (Ironically, critics of Whorf's Hopi data often miss his point about structural analogy and focus narrowly on individual lexical items.) Finally, grouping referents and concepts as formally 'the same' for the purposes of speech has led speakers to group those referents and concepts as substantively 'the same' for action generally, as evidenced by related cultural patterns of belief and behavior he describes.

3. Empirical Research on the Hypothesis

Although the Sapir–Whorf proposal has had wide impact on thinking in the humanities and social sciences, it has not been extensively investigated empirically. Indeed, some believe it is too difficult, if not impossible in principle, to investigate. Further, a good deal of the empirical work that was first developed was quite narrowly confined to attacking Whorf's analyses, documenting particular cases of language diversity, or exploring the implications in domains such as color terms that represent somewhat marginal aspects of language structure. In large part, therefore, acceptance or rejection of the proposal for many years depended more on personal and professional predilections than on solid evidence. Nonetheless, a variety of modern initiatives have stimulated renewed interest in mounting empirical assessments of the hypothesis.

Contemporary empirical efforts can be classed into three broad types, depending on which of the three key terms in the hypothesis they take as their point of departure: language, reality, or thought (Lucy 1997).

A *structure-centered approach* begins with an observed difference between languages, elaborates the interpretations of reality implicit in them, and then seeks evidence for their influence on thought. The approach remains open to unexpected interpretations of reality but often has difficulty establishing a neutral basis for comparison. The classic example of a language-centered approach is Whorf's pioneering comparison of Hopi and English described above. The most extensive contemporary effort to extend and improve the comparative fundamentals in a structure-centered approach has sought to establish a relation between variations in grammatical number marking and attentiveness to number and shape (Lucy 1992b). This research remedies some of the traditional difficulties of structure-centered approaches by framing the linguistic analysis typologically so as to enhance comparison, and by supplementing ethnographic observation with a rigorous assessment of individual thought. This then makes possible the realization of the benefits of the structure-centered approach: placing the languages at issue on an equal footing, exploring semantically significant lexical and grammatical patterns, and developing connections to interrelated semantic patterns in each language.

A *domain-centered approach* begins with a domain of experienced reality, typically characterized independently of language(s), and asks how various languages select from, encode, and organize it. Typically, speakers of different languages are asked to refer to 'the same' materials or situations so the different linguistic construals become clear. The approach facilitates controlled comparison, but often at the expense of regimenting the linguistic data rather narrowly. The classic example of this approach, developed by Roger Brown and Eric Lenneberg in the 1950s, showed that some colors are more lexically encodable than others, and that more codable colors are remembered better. This line of research was later extended by Brent Berlin, Paul Kay, and their colleagues, but to argue instead that there are cross-linguistic *universals* in the encoding of the color domain such that a small number of 'basic' color terms emerge in languages as a function of biological constraints. Although this research has been widely accepted as evidence against the validity of linguistic relativity hypothesis, it actually deals largely with constraints on linguistic diversity rather than with relativity as such. Subsequent research has challenged Berlin and Kay's universal semantic claim, and shown that different color-term systems do in fact influence color categorization and memory. (For discussions and references, see Lucy 1992a, Hardin and Maffi 1997, Roberson et al. 2000.) The most successful effort to improve the quality of the linguistic comparison in

a domain-centered approach has sought to show cognitive differences in the spatial domain between languages favoring the use of body coordinates to describe arrangements of objects (e.g., 'the man is left of the tree') and those favoring systems anchored in cardinal direction terms or topographic features (e.g., 'the man is east/uphill of the tree') (Pederson et al. 1998, Levinson in press). This research on space remedies some of the traditional difficulties of domain-centered approaches by developing a more rigorous and substantive linguistic analysis to complement the ready comparisons facilitated by this approach.

A *behavior-centered approach* begins with a marked difference in behavior which the researcher comes to believe has its roots in and provides evidence for a pattern of thought arising from language practices. The behavior at issue typically has clear practical consequences (either for theory or for native speakers), but since the research does not begin with an intent to address the linguistic relativity question, the theoretical and empirical analyses of language and reality are often weakly developed. The most famous example of a behavior-centered approach is the effort to account for differences in Chinese and English speakers' facility with counterfactual or hypothetical reasoning by reference to the marking of counterfactuals in the two languages (Bloom 1981). The interpretation of these results remains controversial (Lucy 1992a).

4. Future Prospects

Two research trends are unfolding at the present time. First, in the cognitive and psychological sciences awareness is increasing of the nature and scope of language differences. This has led to a greater number of studies focused on the possible cognitive consequences of such differences (e.g., Levinson in press, Niemeier and Dirven 2000). Second, an increasing integration is emerging among the three levels of language and thought problem (i.e., the semiotic, structural, and functional levels). On the semiotic side, for example, research on the relationship between language and thought in development is increasingly informing and informed by work on linguistic relativity (Bowerman and Levinson 2001). On the functional side, research on the relationship of cultural and discursive patterns of use is increasingly being brought into dialogue with Whorfian issues (Silverstein 1979, Friedrich 1986, Wierzbicka 1992, Hill and Mannheim 1992, Gumperz and Levinson 1996).

The continued relevance of the linguistic relativity issue seems assured by the same impulses found historically: the patent relevance of language to human sociality and intellect, the reflexive concern with the role of language in scholarly practice, and the practical encounter with linguistic diversity. To this we must add the increasing concern with the unknown implications for human thought of the impending loss of many if not most of the world's languages (Fishman 1982).

See also: Cognitive Psychology: History; Human Cognition, Evolution of; Language and Philosophy; Language and Thought: The Modern Whorfian Hypothesis; Linguistic Anthropology; Linguistics: Overview; Sapir, Edward (1884–1939); Semiotics; Wittgenstein, Ludwig (1889–1951)

Bibliography

Aarsleff H 1988 Introduction. In: Von Humboldt W *On Language: The Diversity of Human Language-structure and its Influence on the Mental Development of Mankind* (Heath P, trans.). Cambridge University Press, Cambridge, UK, pp. vii–xv

Bloom A H 1981 *The Linguistic Shaping of Thought*. Lawrence Erlbaum, Hillsdale, NJ

Bowerman M, Levinson S C 2001 *Language Acquisition and Conceptual Development*. Cambridge University Press, Cambridge, UK

Fishman J 1982 Whorfianism of the third kind: Ethnolinguistic diversity as a worldwide societal asset (The Whorfian Hypothesis: Varieties of validation, confirmation, and disconfirmation II). *Language in Society* **11**: 1–14

Friedrich P 1986 *The Language Parallax: Linguistic Relativism and Poetic Indeterminacy*. University of Texas, Austin, TX

Gumperz J J, Levinson S C (eds.) 1996 *Rethinking Linguistic Relativity*. Cambridge University Press, Cambridge, UK

Hardin C L, Maffi L (eds.) 1997 *Color Categories in Thought and Language*. Cambridge University Press, Cambridge, UK

Hill J H, Mannheim B 1992 Language and world view. *Annual Review of Anthropology* **21**: 381–406

Koerner E F K 1992 The Sapir–Whorf hypothesis: A preliminary history and a bibliographic essay. *Journal of Linguistic Anthropology* **2**: 173–8

Lee P 1997 *The Whorf Theory Complex: A Critical Reconstruction*. John Benjamins, Amsterdam, The Netherlands

Levinson S C in press *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press, Cambridge, UK

Lucy J A 1992a *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. Cambridge University Press, Cambridge, UK

Lucy J A 1992b *Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis*. Cambridge University Press, Cambridge, UK

Lucy J A 1997 Linguistic relativity. *Annual Review of Anthropology* **26**: 291–312

Niemeier S, Dirven R (eds.) 2000 *Evidence for Linguistic Relativity*. John Benjamins, Amsterdam, The Netherlands

Pederson E, Danziger E, Wilkins D, Levinson S, Kita S, Senft G 1998 Semantic typology and spatial conceptualization. *Language* **74**: 508–56

Roberson D, Davies I, Davidoff J 2000 Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology—General* **129**: 369–98

Sapir E 1949 *The Selected Writings of Edward Sapir in Language, Culture, and Personality* (Mandelbaum D G, ed.). University California Press, Berkeley, CA
Silverstein M 1979 Language structure and linguistic ideology. In: Clyne P, Hanks W, Hofbauer C (eds.) *The Elements: A Parasession on Linguistic Units and Levels*. Chicago Linguistic Society, Chicago, pp. 193–247
Werlen I 1989 *Sprache, Mensch und Welt: Geschichte und Bedeutung des Prinzips der sprachlichen Relativität*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany
Whorf B L 1956 *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (Carroll J B, ed.). MIT Press, Cambridge, MA
Wierzbicka A 1992 *Semantics, Culture, and Cognition. Universal Human Concepts in Culture-specific Configurations*. Oxford University Press, Oxford, UK

J. A. Lucy

Sauer, Carl Ortwin (1889–1975)

Carl Sauer was one of the towering intellectual figures of the twentieth century, not only in geography but also in a wider intellectual sphere. However, because of his wide-ranging thought, speculative sweep, and world perspective over a number of subjects, it is difficult to define his contribution to the social and behavioral sciences precisely and easily. But one can say that some of his underlying concerns were about scholarship, independence of thought, opposition to academic bureaucracy, a sympathy and identification with rural folk, concern for cultural diversity and environmental quality, and a distaste for the technological and scientific 'fix,' particularly the solutions offered by the emerging social sciences after 1945.

He was born in Warrenton, Missouri on December 24, 1889 of German parents, and because of his background and three years of schooling in Calw (near Stuttgart) he was influenced by German culture and literature. He completed his graduate studies at Chicago University in 1915 under the geographer Ellen Semple Churchill, the geologist Rollin D. Salisbury, and plant ecologist Henry C. Cowles, the latter two of whom made a lasting intellectual impression on him. He then taught at the University of Michigan until 1923, when he moved to geography at Berkeley where he taught for 34 years (32 as Chair) and established one of the most distinctive graduate schools of American Geography that would always be associated with 'cultural geography.' After he retired in 1955 he enjoyed 20 remarkably productive years that saw the publication of four books and a score of influential papers, all distinguished by big and speculative ideas, that were the fruits of a lifetime's reflection and unhurried reading. His reputation soared so that an 'aura of sage, philosopher-king, and even oracle surrounded him' (Hooson 1981, p. 166). He died in Berkeley on July 18, 1975.

1. Cultural Geography

Sauer rebelled against the sterile environmental determinism of contemporary geography with its emphasis on humans as response mechanisms to physical factors. If nothing else, his experience in the Economic Land Survey in the Michigan Cutovers had shown him that humans radically transformed the earth, often for the worse, and in the process created cultural landscapes. In his search for a new, humane geography, cultural anthropology seemed to offer a means of dealing with the diversity of humankind and its cultural landscapes through time. On arriving in Berkeley he found natural soul-mates in the anthropologists Alfred L. Kroeber and Robert H. Lowie. The concept of 'culture' subsequently pervaded all his teaching and writing. In *The Morphology of Landscape* he distilled an almost wholly German geographical literature, established the primacy of human agency in the formation of cultural landscapes 'fashioned out of the natural landscape by a cultural group,' and the importance of a time-based approach. In addition, he placed great importance on observation and contemplation in the field—a *Verstehen* or empathetic understanding and intuitive insight into behavior or object in order to achieve 'a quality of reasoning at a higher plane' than the tangible facts (see Sauer 1925, Williams 1983).

Sauer wrote 'Morphology' in order to 'emancipate' himself from determinist thinking and, with a few more papers on the cultural landscape of the midwest frontier, he put the epistemological game behind him and started substantive field work on early settlement and society in Lower California, Arizona and Mexico. During the 1930s he produced many works, including *Aztatlan* and *The Road to Cibola* both published in a new monograph series *Ibero-Americana* that he founded in the same year with Alfred Kroeber and the historian H. E. Bolton. (For these and many other of Sauer's publications see Leighly 1963.) These investigations drew him into the controversy about New World plant domestication and plant origins, and into collaboration with botanists, archeologists and ethnologists, whom he found congenial intellectual company.

2. Widening Horizons

All the time Sauer's horizons were getting wider, his ideas more speculative, and his ethical values more refined. Toward the end of the 1930s he wrote a slight, but ultimately influential, paper which was a sustained and biting critique of the destructive social and environmental impact that resulted from the predatory outreach of Europe, which had few counterparts at that time except, perhaps, in the writing of Karl Marx (Sauer 1938). He drew inspiration from the work of George Perkins Marsh on the human transformation of the earth and Ernst Friedrich's concept of *Raub-*

wirtschaft, or destructive exploitation (see Marsh [1865] 1965, Friedrich 1904). His experience and knowledge of Latin and Central America and their history suggested to him that the Spanish conquest had led to a devastating and permanent impoverishment of the land and of its cultures and societies. Disease, warfare and enslavement had disrupted traditional value systems. Thus, the diffusion of technologically superior societies could affect humans and their culture just as much as it could physical resources.

But two works more than any others established his world reputation and heralded a remarkable decade of multifaceted yet interrelated speculative understanding of the place of humans on earth. First was *Agricultural Origins and Dispersals* (Sauer 1952a) that flowered later into a string of publications into the human uses of the organic world, and early humans in the Americas from the Ice Age onward. Unfortunately, radiocarbon dating came too late to inform Sauer's writing, but although he may not have provided the answers, he defined the questions brilliantly.

Second, in 1956, with the collaboration of Marston Bates and Lewis Mumford, he masterminded the Princetown symposium on 'Man's Role in Changing the Face of the Earth,' the theme of which thereafter became his overriding interest. (See *Mumford, Lewis (1895–1990)*.) All his learning and concerns culminated in this volume, and in his chapter 'The Agency of Man on Earth' (Sauer 1956). The capacity of humans to alter the natural environment—the 'deformation of the pristine'—the cult of progress and waste that stemmed from mass production ('commodity fetishism'), and the alien intrusion of humans into world ecology, were included. In contemporary terms, the theme was the degradation of the environment, and it was an early and influential statement. It also had another dimension: globally, the 'imperialism of production' was as bad as the old, colonial imperialism, and might ultimately be no better than Marxist totalitarianism; mass culture was eliminating not only biological diversity but also cultural diversity, and older and less robust societies. Somehow, humans had to rise above this mindless, short-term exploitative mode. 'The high moments of history have come not when man was concerned with the comforts and displays of the flesh but when his spirit was moved to grow in grace.' Therefore, what was needed was 'an ethic and aesthetic under which man, practising the qualities of prudence and moderation, may indeed pass on to posterity a good Earth' (Sauer 1956, p. 68).

His simply articulated ideas had a resonance with many activists and intellectuals, as well as Californian avant-garde poets and literati, who extolled his work as an example of cultural and ecological sensitivity and respect, tinged with deep historical insight and scholarship, and made attractive by his simple and pithy language. He also tapped a deep spring of feeling during the 1960s and 1970s at the time of Vietnam and student unrest with their concerns at the limits of the earth and technological/political power. Yet Sauer was a complex mix. He was congenitally nonconformist but deeply conservative, and although profoundly concerned with conservation was never formally an 'environmentalist,' and indeed, he thought the movement was little more than an 'ecological binge.'

3. *Distrust of the Social Sciences*

In many ways it is ironical that Sauer has a place in this *Encyclopedia* because he had such a deep distrust and distaste for the behavioral and social sciences. Although he eschewed methodological and epistemological discussion, his writing, and particularly his personal correspondence, reveal that he was 'a philosopher in spite of himself' (Entrikin 1984). Culture history became his model of social science, and it was drawn from the natural sciences, not the social sciences. He was basically a pragmatist who was influenced by the writings of the German cultural geographers Friedrich Ratzel and Eduard Hahn, and by the methods of geology and anthropology and their emphasis on the provisional character of working hypotheses which were no more than a means to an end. His later association in Berkeley with people who worked on 'tangible things,' like plant ecologists, agricultural scientists, botanists (e.g., Ernest Babcock), experts in the evolution of population genetics (see *Wright, Sewall (1889–1988)*), and geneticists reinforced this.

He argued for a theoretical and methodological pluralism that stemmed naturally from the inherent diversity of nature and culture. The natural science idea of a dynamic balance and diversity that arose from organic evolution was embedded deeply in his thought, and he felt that the modern world had disrupted these processes so that it was out of balance. Hence his deep distrust of US capitalism and all bureaucratic systems which would destroy diversity and local community. Liberal social scientists who designed, planned, and directed community life were more likely to destroy than enhance it by imposing universalizing concepts of social organization, and by ignoring the inherent pluralism and diversity of nature and culture that stimulated the naïve curiosity about the world which was the essence of geography. Theoretical and normative social science as practiced by economists, sociologists, and political scientists, grated on him with its exaggerated confidence in the statistical and the inductive, and its 'dialectic atmosphere.' Their focus on the present precluded a better insight into the origins and evolution of any topic, and gave 'an exaggerated accent on contemporaneity.' His social science was based firmly in history and geography—this was culture history.

During the late 1930s he talked jokingly of the two patron saints of social scientists—St. Bureaucraticus, who represented rationalism and professionalism in US academic life, and St. Scholasticus, who represented social theorists who sought normative generalizations about humankind and society. Consistently, these 'Sons of Daedalus' were the target of his criticism because their well-funded procedures and programs emasculated the independence of the impressionable younger scholars (Williams 1983).

His address, 'Folkways of the Social Sciences' was an eloquent, if not audacious, plea to social scientists to 'give back the search for truth and beauty to the individual scholar to grow in grace as best he can' and to reinstate time and place into the study of the USA (Sauer 1952b).

Sauer's other *bête noire*—mass production—promoted the homogenization of society, not only in the USA but globally, as the USA firmly assumed the position of superpower during the 1950s, with what he saw as a disdain for the 'lesser breeds.'

Sauer was aware that he was 'out of step' with his colleagues and intellectuals, and spoke of himself as an 'unofficial outsider' and even as a 'peasant.' This has been typified as part of the antimodernism that characterized early twentieth century US intellectual life. But it went much deeper than that; his ideas anticipated society's fears and disenchantment with progress, science, the elimination of diversity, and the degradation of the environment. Perhaps the ultimate relevance of Sauer's work was in his groping towards a *rapprochement* between the social sciences, the humanities and the biological life sciences. By being behind he was far ahead.

See also: Agricultural Sciences and Technology; Environmental Determinism; Geography; Place in Geography

Bibliography

Entrikin J N 1984 Carl O. Sauer: Philosopher in spite of himself. *Geographical Review* **74**: 387–408

Friedrich E 1904 Wesen und geographische Verbreitung der 'Raubwirtschaft.' *Petermanns Mitteilungen* **50**: 68–79, 92–5

Hooson D 1981 Carl Ortwin Sauer. In: Blouet B W (ed.) *The Origins of Academic Geography in the United States*. Archon, Hamden, CT, pp. 165–74

Leighly J (ed.) 1963 *Land and Life: A Selection from the Writings of Carl Ortwin Sauer*. University of California Press, Berkeley, CA

Marsh G P 1965 *Man and Nature: Physical Geography as Modified by Human Action*, Lowenthal D (ed.). Harvard University Press, Cambridge, MA

Sauer C O 1925 *The Morphology of Landscape*. University of California Publications in Geography, Berkeley, CA, Vol. 2, pp. 19–53

Sauer C O 1938 Destructive exploitation in modern colonial expansion. *Comptes Rendus du Congrès International de Géographie, Amsterdam* **2** (Sect. 3c): 494–9

Sauer C O 1952a *Agricultural Origins and Dispersals*. Bowman Memorial Lectures, Series 2. American Geographical Society, New York

Sauer C O 1952b Folkways of social science. In: *The Social Sciences at Mid-Century: Papers Delivered at the Dedication of Ford Hall, April 19–21, 1951*. University of Minnesota Press, Minneapolis, MN

Sauer C O 1956 The agency of man on earth. In: Thomas W L (ed.) *Man's Role in Changing the Face of the Earth*. Chicago University Press, Chicago, IL, pp. 49–69

Williams M 1983 The apple of my eye: Carl Sauer and historical geography. *Journal of Historical Geography* **9**: 1–28

Williams M 1987 Carl Sauer and man's role in changing the face of the earth. *Geographical Review* **77**: 218–31

M. Williams

Saussure, Ferdinand de (1857–1913)

1. Saussure's Status in Twentieth-century Linguistics

Saussure is best known for the posthumous compilation of lecture notes on general linguistics taken down assiduously by students attending his courses during 1907–1911, the *Cours de linguistique générale*, edited by his former students and junior colleagues and first published in 1916 (and since 1928 translated into more than a dozen languages). During his lifetime, Saussure was most widely known for his masterly *Mémoire* of 1878 devoted to an audacious reconstruction of the Proto-Indo-European vowel system. However, it is generally agreed that his *Cours* ushered in a revolution in linguistic thinking during the 1920s and 1930s which still at the beginning of the twenty-first century is felt in many quarters, even beyond linguistics proper. He is widely regarded as 'the father of structuralism'; to many his work produced a veritable 'Copernican revolution' (Holdcroft 1991, p. 134). Indeed, essential ingredients and terms of his theory have become points of reference for any serious discussion about the nature of language, its functioning, development, and uses.

2. Formative Years and Career

Saussure was born on November 26, 1857 in Geneva, Switzerland. Although from a distinguished Geneva family which—beginning with Horace Bénédict de Saussure (1740–1799)—can boast of several generations of natural scientists, F. de Saussure was drawn early to language study, producing an '*Essai pour réduire les mots du grec, du latin et de l'allemand à un petit nombre de racines*' at age 14 or 15 (published in *Cahiers Ferdinand de Saussure* **32**: 77–101 [1978]). Following his parents' wishes, Saussure attended classes in chemistry, physics, and mathematics at the

University of Geneva during 1875–1876, before being allowed to join his slightly older classmates who had left for Leipzig the year before. So in the fall of 1876 Saussure arrived at the university where a number of important works in the field of Indo-European phonology and morphology, including Karl Verner's (1846–1896) epoch-making paper on the last remaining series of exceptions to 'Grimm's Law,' had just been published. Saussure took courses with Georg Curtius (1820–1885), the mentor of the *Junggrammatiker*, and a number of the younger professors, such as August Leskien (1840–1916), Ernst Windisch (1844–1918), Heinrich Hübschmann (1848–1908), Hermann Osthoff (1847–1909), and others in the field of Indic studies, Slavic, Baltic, Celtic, and Germanic. During 1878–1879 Saussure spent two semesters at the University of Berlin, enrolling in courses in Indic philology with Heinrich Zimmer (1851–1910) and Hermann Oldenberg (1854–1920).

After barely six semesters of formal study of comparative-historical Indo-European linguistics Saussure, then just 21, published his major lifetime work. In this 300-page *Mémoire sur le système primitif des voyelles dans les langues indo-européennes* (1879) Saussure assumed, on purely theoretical grounds, the existence of an early Proto-Indo-European sound of unknown phonetic value (designated *A) which would develop into various phonemes of the Indo-European vocalic system depending on its combination with those 'sonantal coefficients.' Saussure was thus able to explain a number of puzzling questions of Indo-European ablaut. However, the real proof of Saussure's hypotheses came only many years later, after his death, following the decipherment of Hittite and its identification as an Indo-European language. In 1927 the Polish scholar Jerzy Kuryowicz (1895–1978) pointed to Hittite cognates, i.e., related words corresponding to forms found in other Indo-European languages, that contained a laryngeal (not present in any of the other attested Indo-European languages) corresponding to Saussure's '*phonème*' *A (Szemerényi 1973). What is significant in Saussure's approach is his insistance on, and rigorous use of, the idea that the original Proto-Indo-European vowels form a coherent system of interrelated terms. Indeed, it is this emphasis on the systematic character of language which informs all of Saussure's linguistic thinking to the extent that there are not, contrary to received opinion, two Saussures, the author of the *Mémoire* and the originator of the theories laid down in the *Cours* (cf. Koerner 1998).

Having returned to Leipzig, Saussure defended his dissertation on the use of the genitive absolute in Sanskrit in February 1880, leaving for Geneva soon thereafter. Before he arrived in Paris in September 1880, he appears to have conducted fieldwork on Lithuanian, an Indo-European language of which documents reach back to the sixteenth century only, but which exhibits a rather conservative vowel system comparable with that of Ancient Greek. First-hand exposure to this language was instrumental in his explanation of the Lithuanian system of accentuation (Saussure 1896) for which he is justly famous. In 1881, Michel Bréal (1832–1915), the doyen of French linguistics, secured him a position as *Maître de Conférences* at the *École des Hautes Études*, a post he held until his departure for Geneva 10 years later. In Paris, Saussure found a number of receptive students, among them Antoine Meillet (1866–1936), Maurice Grammont (1866–1946), and Paul Passy (1859–1940), but also congenial colleagues such as Gaston Paris (1839–1903), Louis Havet (1849–1925), who had previously written the most detailed review of his *Mémoire*, and Arsène Darmesteter (1848–1888). Still, Saussure did not write any major work subsequent to his doctoral dissertation, but he wrote a series of frequently etymological papers, which illustrate his acumen in historical linguistics. It was through the posthumous publication of lectures on (in fact historical and) general linguistics that Saussure became known for his theoretical and nonhistorical views.

In 1891, the University of Geneva offered Saussure a professorship of Sanskrit and Comparative Grammar, which was made into a regular chair of Comparative Philology in 1896. It was only late in 1906 that the Faculty added the subject of General Linguistics to his teaching load. It was this decision and Saussure's three ensuing lecture series (1907, 1908–9, and 1910–11) in which he developed his thoughts about the nature of language and the manner in which it was to be studied that eventually led to the epoch-making book he did not write, the *Cours de linguistique générale*. Saussure died on February 22, 1913 at Château Vufflens, Switzerland.

3. *The Cours De Linguistique Générale*

The *Cours* appeared in 1916. By the 1920s Saussure's name began to be almost exclusively connected with this posthumous work which was based largely on extensive lecture notes carefully taken down by a number of his students. One of them was Albert Riedlinger (1883–1978), whose name appears on the title page of the *Cours* as a collaborator. It was, however, put together by Saussure's successors in Geneva, Charles Bally (1865–1947) and Albert Sechehaye (1870–1946), neither of whom had attended these lectures themselves as is frequently, but erroneously, stated in the literature. Indeed, their own focus of attention was nonhistorical linguistics, stylistics and syntax, respectively, and this had a considerable bearing on the manner in which Saussure's ideas were presented (Amacker 2000), with Historical Linguistics, the subject Saussure was most interested in, being relegated to the end of the book. (See Godel 1957, for an analysis of the editors' work; also Strozier 1988, for a close analysis of the texts.) It

was the long general introduction of the *Cours* and the part dealing with nonhistorical ('synchronic') linguistics, which made history.

3.1 Saussure's/the Cours's Legacy

The ideas advanced in the *Cours* produced something of a revolution in linguistic science; historical-comparative grammar which had dominated linguistic research since the early nineteenth century soon became a mere province of the field. At least in the manner the *Cours* had been presented by the editors, Saussure's general theory of language was seen as assigning pride of place to the nonhistorical, descriptive, and 'structural' approach. (Saussure himself did not use the last-mentioned term in a technical sense.) This emphasis on the investigation of the current state of a language or languages led to a tremendous body of work concerned with the analysis of the linguistic system or systems of language and its function(s), and a concomitant neglect of questions of language change and the field of Historical Linguistics in general, a situation still very much characteristic of the current linguistic scene. However, the field has become stronger since the mid-1980s, as sociolinguistic and typological aspects took hold in the investigation of language change.

From the 1920s onwards, notably outside of the traditional centers of Indo-European comparative linguistics, a variety of important schools of linguistic thought developed in Europe that can be traced back to proposals made in the *Cours*. These are usually identified with the respective centers from which they emanated, such as Geneva, Prague, Copenhagen, even London; more precisely these developments are to be associated with the names of Bally and Sechehaye, Roman Jakobson (1896–1982) and Nikolaj S. Trubezkoy (1890–1938), Louis Hjelmslev (1899–1965), and John Rupert Firth (1890–1960), respectively. In North America too, through the work of Leonard Bloomfield (1887–1949), Saussure's ideas became stock-in-trade among linguists, descriptivists, structuralists, and generativists (cf. Joseph 1990, for Saussure's influence on Bloomfield as well as Chomsky). In each 'school,' it is safe to say, essential ingredients of the *Cours* were interpreted differently, at times in opposition to some of Saussure's tenets as found in the book, which Saussure specialists now refer to as the 'vulgata' text, given that a number of points made in the *Cours* go back to its editors, not Saussure himself. However, it is this text that has made the impact on modern linguistics.

3.2 The Main Tenets of the Cours

At the core of Saussure's linguistic theory is the assumption that language is a system of interrelated terms, which he called '*langue*' (in contradistinction to 'parole,' the individual speech act or speaking in general). This supra-individual '*langue*' is the underlying code ensuring that people can speak and understand each other; the language-system thus has a social underpinning. At the same time, '*langue*' is an operative system embedded in the brain of everyone who has learned a given language. The analysis of this system and its functioning, Saussure maintains, is the central object of linguistics. His characterization of '*langue*' as a '*fait social*' has often led to the belief that Saussure's thinking is indebted to Émile Durkheim's (1858–1917) sociological framework. While the two were close contemporaries and shared much of the same intellectual climate, no direct influence of the latter on the former can be demonstrated; Meillet, Saussure's former student and a collaborator of Durkheim's since the late 1890s, publicly denied it when it was first proposed in 1931. For Saussure the social bond between speakers sharing the same language ('*langue*') was constitutive for the operation of this unique semiological system (see below).

The language system is a network of relationships which Saussure characterized as being of two kinds: 'syntagmatic' (i.e., items are arranged in a consecutive, linear order) and 'associative,' later on termed (by Firth and by Hjelmslev) 'paradigmatic' (i.e., pertaining to the organization of units in a deeper, not directly observable fashion dealing with grammatical and semantic relations). Since it is only in a state ('*état de langue*') that this system can be revealed, the nonhistorical, 'synchronic' approach to language must take pride of place. Only after two such language states of different periods in the development of a given language have been properly described can the effects of language change be calculated, i.e., 'diachronic,' historical linguistics be conducted. Hence the methodological, if not epistemological primacy of synchrony over diachrony.

Apart from syntagmatic vs. paradigmatic relations, several trichotomies can be found in the *Cours* which, however are usually reduced to dichotomies. Many of them have become current in twentieth-century thought, far beyond their original application, i.e., *language–langue–parole* (i.e., language in all its manifestations or 'speech'; language as the underlying system, and 'speaking,' with terms such as 'tongue' and 'discourse' or 'competence' and 'performance' being proposed to replace the *langue/parole* couple), *signe–signifié–signifiant* (sign, signified, and signifier), synchrony vs. diachrony (Saussure's 'panchrony' would be an overarching of these two perspectives).

Saussure's definition of language as 'a system of (arbitrary) signs' and his proposal of linguistics as the central part of an overall science of sign relations or 'sémiologie' have led to the development of a field of inquiry more frequently called (following Charles Sanders Peirce's [1839–1914] terminology) 'semiotics,' which more often than not deals with sign systems

other than those pertaining to language, such as literary texts, visual art, music, and architecture. (For the wider implications of Saussure's socio-semiotic ideas and a critique of the various uses of Saussure's concepts in literary theory, see Thibault 1997.)

As is common with influential works (cf., e.g., Freud's) many ingredients of Saussure's general theory of language have often been taken out of their original context and incorporated into theories outside their intended application, usually selectively and quite arbitrarily, especially in works by French writers engaged in 'structural' anthropology (e.g., Claude Lévi-Strauss) and Marxist philosophy (e.g., Louis Althusser), literary theory (e.g., Jacques Derrida), psychoanalysis (e.g., Jacques Lacan), and semiotics (e.g., Roland Barthes), and their various associates and followers. (For a judicious critique of these extra-linguistic exploitations, see Tallis 1988.) However, these uses—and abuses—demonstrate the endurance and originality of Saussure's ideas. He has achieved in linguistics a status comparable to Imanuel Kant in philosophy, in that we can, similar to Kant's place in the history of thought, distinguish between a linguistics before Saussure and a linguistics after Saussure.

Bibliography

Amacker R 2000 Le développment des idées saussurennes chez Charles Bally et Albert Sechehay. *Historigraphia Linguistica* **27**: 205–64

Bouquet S 1997 *Introduction à la lecture de Saussure*. Payot, Paris

Engler R 1976 Bibliographie saussurienne [1970–]. *Cahiers Ferdinand de Saussure* **30**: 99–138, **31**: 279–306, **33**: 79–145, **40**: 131–200, **43**: 149–275, **50**: 247–95 (1976, 1977, 1979, 1986, 1989, 1997).

Godel R 1957 *Les Sources manuscrites du Cours de linguistique générale de F. de Saussure*. Droz, Geneva, Switzerland

Harris R 1987 *Reading Saussure: A Critical Commentary on the Cours de linguistique générale*. Duckworth, London

Holdcroft D 1991 *Saussure: Signs, System, and Arbitrariness*. Cambridge University Press, Cambridge, UK

Joseph J E 1990 Ideologizing Saussure: Bloomfield's and Chomsky's readings of the Cours de linguistique générale. In: Joseph J E, Taylor T E (eds.) *Ideologies of Language*. Routledge, London and New York, pp. 51–93

Koerner E F K 1972 *Bibliographia Saussureana, 1870–1970: An Annotated, Classified Bibliography on the Background, Development and Actual Relevance of Ferdinand De Saussure's General Theory of Language*. Scarecrow Press, Metuchen, NJ

Koerner E F K 1973 *Ferdinand de Saussure: Origin and Development of his Linguistic Thought in Western Studies of Language. A Contribution to the History and Theory of Linguistics*. F. Vieweg and Sohn, Braunschweig, Germany

Koerner E F K 1988 *Saussurean Studies/Études saussuriennes. Avant-propos de R. Engler*. Slatkine, Geneva, Switzerland

Koerner E F K 1998 Noch einmal on the History of the Concept of Language as a 'système où tout se tient'. *Cahiers Ferdinand de Saussure* **51**: 21–40

Saussure F de 1879 Mémoire sur le système primitif des voyelles dans les langues indo-européennes. B. G. Teubner, Leipzig, Germany (Repr. G. Olms, Hildesheim, 1968.). In: Lehmann W P (ed.) *A Reader in Nineteenth-Century Historical Indo-European Linguistics*. Indiana University Press, Bloomington and London, 1967, pp. 218–24

Saussure F de 1896 Accentuation lituanienne. Indogermanische Forschungen. *Anzeiger* **6**: 157–66

Saussure F de 1916 *Cours de linguistique* générale ed. by Charles Bally and Albert Sechehaye, with the collaboration of Albert Riedlinger. Payot, Lausanne and Paris. (2nd ed., Paris: Payot, 1922; 3rd and last corrected ed., 1931; 4th ed., 1949; 5th ed., 1960, etc.—English transl.: (1) By Wade Baskin, *Course in General Linguistics*, Philosophical Library, London and New York, 1959 (repr., New York: McGraw-Hill, 1966; rev. ed., Collins/Fontana, London, 1974), and (2) by Roy Harris, *Course in General Linguistics*, Duckworth, London, 1983.

Saussure F de 1922 *Recueil des publications scientifiques* ed. by Charles Bally and Léopold Gautier. Payot, Lausanne; C. Winter, Heidelberg. (Repr. Slatkine, Geneva, 1970; it includes a reprint of the *Mémoire*, his 1880 dissertation (pp. 269–338), and all of Saussure's papers published during his lifetime.)

Saussure F de 1957 Cours de linguistique générale (1908–1909): Introduction. *Cahiers Ferdinand de Saussure* **15**: 6–103

Saussure F de 1967–1968, 1974 *Cours de linguistique générale. Édition critique par Rudolf Engler, 4 fasc*. Otto Harrassowitz, Wiesbaden, Germany

Saussure F de 1972 *Cours de linguistique générale*. Payot, Paris

Saussure F de 1978 (1872) Essai pour réduire les mots du grec, du latin et de l'allemand à un petit nombre de racines. *Cahiers Ferdinand de Saussure* **32**: 77–101

Strozier R M 1988 *Saussure, Derrida, and the Metaphysics of Subjectivity*. Mouton de Gruyter, Berlin and New York

Szemerényi O 1973 La théorie des laryngales de Saussure à Kuryowicz et à Benveniste: Essai de réévaluation. *Bulletin de la Société de Linguistique de Paris* **68**: 1–25

Tallis R 1988 *Not Saussure*. Macmillan, London

Thibault P J 1997 *Re-Reading Saussure: The Dynamics of Signs in Social Life*. Routledge, London

E. F. K. Koerner

Savage, Leonard J (1917–71)

L. J. Savage, always addressed by friends as Jimmie, was born in Detroit, Michigan on 20 November 1917. His father was in the real-estate business, his mother a nurse. Throughout his life he suffered from poor eyesight, a combination of nystagmus and extreme myopia, which was to affect the development of his career in many ways. Most of us accept reading as an easy, even casual, activity, whereas to him it was a serious business with the text held close to the eye. The result was that he appeared to regard the material as important and would absorb it with an intensity that often escapes those with normal vision. To have him read something you had written was, at first, an uncomfortable business, for he would politely question much of the material, but later to realize that he was being constructive and the final result benefited

enormously from his study of it. He was a superb lecturer, once remarking that he could hold himself spell-bound for an hour. When, towards the end of his life, he gave the Fisher lecture, he held the audience spell-bound for considerably more than the allotted hour.

The poor eyesight at first affected him adversely, his high-school teacher not recommending him for further education. His parents were insistent and he went initially to Wayne University and then to the University of Michigan. His attempts to do biology failed because he could not draw, and chemistry was ruled out when he dropped the beaker in the laboratory. Eventually he met a fine mathematics teacher who introduced him to a field in which the reading had to be intense, yet minimal, where his brilliance was recognized and he gained a Ph.D in the application of vectorial methods to metric geometry. After a year each at Princeton, Cornell, and Brown, 1944 found him a member of the Statistical Research Group at Columbia University, a group he described as 'one of the greatest hotbeds statistics has ever had,' a judgement that is surely sound, as almost all those who were to become the leading statisticians in the US of the 1950s were in the group or associated with it.

So he became a statistician, and although his greatest work was in the theory, as such he was able to indulge his interest in science, so that throughout his career he enjoyed, and was extremely good at, consulting with his scientific colleagues. It was not just science that interested him, for he had inherited from his father a respect for business, and had acquired from Milton Friedman, a member of the group, an appreciation of economics, that made his advice useful and generously available to all serious thinkers.

In 1946, Savage went to the University of Chicago where, from 1949, he was in the Statistics department founded by Allen Wallis, becoming chairman from 1957 until he left the university in 1960. His departure was a turning point in his life. A reason for leaving was personal difficulties in his marriage to Jane Kretschmer that they hoped could be mended by a move to Michigan with their two sons, Sam and Frank. It did not work out and they were divorced in 1964. That year he married Jean Pearce and moved to Yale. It was in New Haven on 1 November 1971 that he tragically died at the early age of 53, and statistics lost a great scholar and many of us a dear friend. The move from Chicago was partly influenced by his perception of his relations with colleagues there. He felt, as we will see below, that he had produced a sound axiomatization of statistics showing that many standard statistical procedures were unsatisfactory, so that his colleagues should either explain where the axioms were wrong, or else abandon the faulty procedures. They did neither. They were not alone in this; most statisticians acted as the Chicago faculty, refusing to contest the axioms, yet refusing to use the results (many still do so today) and Savage found it difficult to get a suitable post until Yale obliged. The monograph by Savage with others (1962), although now rather dated, indicates the sort of antagonism that existed, though, in that book, all the protagonists are on their best behavior. Savage (1981) contains most of his more important papers, tributes from others and a technical summary of his work. Many of his papers, contained therein, deal, not with technicalities, but with his growing appreciation of statistical ideas as they developed throughout his life.

Savage is justifiably famous for one tremendous contribution to statistics and the scientific method, contained in his 1954 book *The Foundations of Statistics*. Or, to be more correct, the first seven chapters of that book. Anyone contemplating study of this glorious achievement should first read his preface to the second edition in 1972, for this is one of the most honest pieces of scientific writing that I know. To appreciate the importance of the book, it is necessary to cast one's mind back to the state of statistics around 1950. Fisher made enormous advances from 1920 onwards. Neyman and Pearson had introduced new ideas and Wald was developing decision analysis. But the statistical procedures available, though undoubtedly valuable and much appreciated by scientists, especially biologists, were a disparate collection of ideas lacking coherence. Savage had the idea that statistics should be like other branches of mathematics and be based on a set of concepts, termed axioms, from which the procedures would follow as theorems. He wanted to provide a firm basis for Fisher's results. Another way of understanding is to recognize that Fisher had suggested procedures whose properties had been studied by him and others, whereas Savage turned the method around and asked what properties we wanted and discovered what procedures provided them. His expectation was that they would prove to be just what statisticians were using. To his great surprise, they were not. Interestingly, his last paper, Savage et al. (1976), is a brilliant, sympathetic review of Fisher's work.

Savage had worked with von Neumann in Princeton and had much appreciated his work on games with Morgenstern (1944), that introduced axioms leading to the concept of a utility function for the outcomes of a game. However, their treatment had used probability without justification, so Savage conceived the idea of developing an axiomatic system for decision-making under uncertainty which would not use probability in the axioms but derive it in theorems, so uniting the ideas of probability and utility. Also influential was the development by Wald (1950) of an attempt to unite statistical ideas under the concept of decision-making, using von Neumann's idea of minimax, but lacking any axiomatic foundation. Savage's 1954 book presents axioms for a single decision-maker, now often referred to as 'you,' faced with selecting a course of action in the face of uncertainty. He then proves three

theorems: first, that your various uncertainties present must combine according to the rules of probability; second, that your perceptions of the merits of the possible outcomes that result from your actions must be described by a real-valued utility function (which is itself based on probability); third, that your optimum decision is that which maximizes your expected utility (MEU), the expectation being evaluated according to the probabilities developed in the first part of this trilogy. One of the axioms he used encapsulated a simple notion that has since been seen to be of some importance. If you prefer A to B when C is true and, at the same time, prefer A to B when C is false; then you prefer A to B when you do not know whether or not C is true. It is called the 'sure-thing' principle because the preference of A over B is sure, in American parlance, at least as far as C is concerned.

The introduction of MEU was not new, it having been used by Daniel Bernoulli in the eighteenth century, but the development extended its use considerably, in particular by showing that it was the only sensible method for a single decision-maker to use. Utility had recently been explained, as we have seen, by von Neumann. The original and dramatically important result was the first, saying that probability was the only sensible description for uncertainty, so that the rules of probability were not arbitrary but dictated by the sensible and modest requirements expressed in the axioms. Others, from Laplace onwards, had used probability without justification: Savage showed that it was inevitable. So it was the extensive and unique use of probability that was Savage's main contribution in the book, and there are two ways in which his analysis was influential.

First was the fact that the probability was personal, expressing the uncertainty of the decision-maker, 'you.' It was not *the* probability, but rather *your* probability. An immediate reaction is that this contradicts the very nature of the scientific method which is supposed to yield objective results. This apparent contradiction was resolved in a later paper with Edwards and Lindeman (Savage et al. 1963), which showed in the principle of stable estimation, that under general conditions, people with different uncertainties would reach agreement on receipt of enough data. This result agrees with the observation that scientists typically disagree with one another in the early stages of an investigation but are eventually brought together as the evidence accumulates. Many now think that the reduction of uncertainty, through data, incorporated in the laws of probability, expresses the nature of induction and is part of the scientific method.

A second, influential factor was the result that probability was the *only* sensible description of uncertainty, for scientists, encouraged by statisticians, had been using other methods, for example in the use of tail-area significance tests of a null hypothesis, H, where the relevant quantity is the probability in the tail of a distribution, given that H is true. Savage's argument was that the proper evaluation is the probability of H, given the data, not of an aspect of the data, given H. Similar contradictions arise with confidence intervals for a parameter which involve probability statements about the interval, not about the parameter. These are instances where his conclusions clashed with standard practice, leading to disagreements like that at Chicago mentioned above.

There was a third way in which Savage's ideas contradicted the current practice, though this was not fully realized until Birnbaum's (1962) paper. The axiomatic development showed that if you had a statistical model with probability $p(x/\theta)$ for data x given parameter θ, then the only way the data could affect your uncertainty of θ, given x, was through the likelihood function $p(x/\theta)$ expressing how that probability varied with θ for the fixed, observed x. This is the likelihood principle. Significance tests clearly violate the principle since they use a tail-area which involves integration of $p(x/\theta)$ over x-values, that is, over data that you do not have. Indeed, many statistical procedures, even today, violate the principle. A related example of conflict with current practice is found in optional stopping. It had long been known that, under commonly occurring circumstances, it is possible to continue sampling until a null hypothesis is rejected at any preassigned significance level, so that you can opt to stop at your convenience. Savage showed that it was the significance test that was at fault, not the stopping procedure.

It is an astonishing fact, that he explains in the preface to the second edition, that he and others who worked with him, including myself, in the mid-1950s, did not appreciate these ideas and did not comprehend that, rather than his work justifying statistical practice, it altered it. The latter part of his book in which the justification is attempted, largely using the tool of minimax that von Neumann had employed with great effect in the theory of games, is today of little interest, in contrast to the brilliance of the early chapters. Savage was a true scholar, a man who studied the work of others and accepted any valid criticism of his work, to incorporate their ideas with due acknowledgment. And so he understood the ideas of Ramsey (1926) who had, in a less rigorous presentation, developed similar ideas, ideas which no one had understood until Savage. But more importantly, he appreciated the work of the Italian, de Finetti, who, from an entirely different standpoint, and one which is perhaps more forceful than Savage's, developed the concept and uniqueness of personal probability. (For English readers, the best references are de Finetti (1972, 1974, 1975), though his ideas originate in the 1930s.) Savage acquired skill in Italian and worked with him. One of the problems de Finetti had studied was your assessment of probability. You are uncertain whether it will rain tomorrow in your city; so according to their thesis, you have a probability for rain.

How are you to assess its value? De Finetti suggested a scoring rule: if you choose p as your value, you will receive a penalty score $(1-p)^2$ if it rains and p^2 if not. Scores for different events will be added. To minimize your total score the values that you give must be probabilities. The basic paper was written by Savage (1972) with due acknowledgment to the originator. With Hewitt (1955), Savage also extended significantly another result of de Finetti on exchangeability.

Savage conducted a major study with Dubins that resulted in their book (Savage and Dubins 1965). The problem here is that you are in an unfavorable gaming situation at a casino and your object is to maximize your chance of winning a fixed sum of money; how should you play? They showed that boldness pays; bet all you have or the maximum amount the casino will allow. This led to a thorough study of betting strategies and many interesting and useful results were obtained.

For reasons that are not entirely clear, the school of statistics that rests on Savage's results is called Bayesian, after the eighteenth century cleric, Thomas Bayes partly because much use is made of Bayes's theorem. Today it is a flourishing school of statistics with a large and increasing number of adherents, that co-exists alongside the earlier school, often called frequentist because probability therein refers to frequency, whereas to a Bayesian it expresses your uncertainty of an event. Bayesians believe that their approach to the handling of data, and the making of decisions that use the data, is important for social and behavioral sciences for the following reasons.

If these subjects are to be truly scientific, their arguments must depend heavily on data which, properly interpreted, can be used to defend theories in the interplay of theory and practice that is the scientific method. In the physical sciences, the data ordinarily take the form of results of planned experiments in the field or laboratory, experiments which can, and should be, repeated by others, in accord with stable estimation, their repetition connecting naturally with frequency. In the social sciences, such planning and repetition is rarely possible, so that reliance has to be placed on observational data, and the behavioral sciences are scarcely better off. It is now widely appreciated that observational data have to be analyzed with considerable care, so that any conclusions are soundly based and not confounded with other factors that the lack of planning has been unable to control. Statistics therefore becomes a more important ingredient in the social, than the physical, sciences. Furthermore, social decisions must often be made in the face of uncertainty, whereas more time and effort by an experimental scientist can substantially reduce, if not eliminate, the uncertainty. The Bayesian approach recognizes and incorporates all of these factors.

There is another sense in which the Bayesian approach may be valuable and that lies in its ability to express uncertainty about any aspect of a study, not confining probability to the repetitious, frequency aspects. Often the conclusions in the social and behavioral sciences are tentative and the personalistic perception of probability, as a measure of your belief in a hypothesis, can provide an expression of uncertainty that can be conveyed to others. We still have a long way to go in employing these ideas, when even a weather-forecaster is reluctant to say there is a probability of 0.8 that it will rain tomorrow, or, if he does, sometimes does not appreciate the meaning of what is being said.

Savage had been described as 'the Euclid of statistics' in that he did for statistics what Euclid did for geometry, laying down a set of axioms from which, as theorems, the only sensible statistical procedures emerge. As with Euclid, other geometries will emerge, so other statistical systems may arise, but, for the moment, Savage's Bayesian approach is the only systematic one available and appears to work. He was, in the sense of Kuhn, a true revolutionary, who overturned one paradigm, replacing it by another without, at first, realizing what he had done.

See also: Bayesian Statistics; Bayesian Theory: History of Applications; Decision Theory: Bayesian; Fisher, Ronald A (1890–1962); Hotelling, Harold (1895–1973); Probability: Interpretations; Risk: Theories of Decision and Choice; Statistical Methods, History of: Pre-1900; Statistics, History of; Tversky, Amos (1937–96)

Bibliography

Birnbaum A 1962 On the foundations of statistical inference. *Journal of the American Statistical Association* **57**: 269–306

de Finetti B 1972 *Probability, Induction and Statistics: The Art of Guessing*. Wiley, London

de Finetti B 1974/5 *Theory of Probability*, 2 Vols. Wiley, London

Dubins L E, Savage L J 1965 *How to Gamble if You Must: Inequalities for Stochastic Processes*. McGraw-Hill, New York

Ramsey F P 1926 Truth and probability. In: Braithwaite R B (ed.) *The Foundations of Mathematics and Other Logical Essays*. Routledge and Kegan Paul, London

Savage L J 1954 *The Foundations of Statistics*. Wiley, New York

Savage L J 1971 Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66**: 783–801

Savage L J 1972 *The Foundations of Statistics*, 2nd edn. Dover, New York

Savage L J et al. 1976 On re-reading R A Fisher. *Annals of Statistics* **4**: 441–500

Savage L J 1981 *The Writings of Leonard Jimmie Savage—A Memorial Selection*. American Statistical Association and Institute of Mathematics, Washington

Savage L J, Edwards W, Lindman H 1963 Bayesian statistical inference for psychological research. *Psychological Reviews* **70**: 193–242

Savage L J, Hewitt E 1955 Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society* **80**: 470–501

Savage L J et al. 1962 *The Foundations of Statistical Inference: A Discussion*. Methuen, London
von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ
Wald A 1950 *Statistical Decision Functions*. Wiley, New York

D. V. Lindley

Savings Behavior: Demographic Influences

In any period all economic output is either consumed or saved. By consuming, individuals satisfy their current material needs. By saving, individuals accumulate wealth that serves a multitude of purposes. When invested or loaned, wealth yields a stream of income to the holder. It is a means of providing for future material needs, for example, during retirement. Wealth provides partial protection against many of life's uncertainties, including the loss of a job or unexpected medical needs. Wealth can be passed on to future generations out of feelings of altruism, or used as a carrot to encourage desired behavior among those who hope for a bequest. Wealth provides status to the holder.

Because of the varied purposes served by saving, many behavioral models have been proposed and a variety of influencing factors have been identified. Demographic factors, including age structure, fertility, and mortality, have been found to influence saving in many studies. Additional factors that have been identified include the level of per capita income, economic growth rates, interest rates, characteristics of the financial system, fiscal policy, uncertainty, and public pension programs.

1. *Saving Trends and their Importance*

In the 1960s, low saving rates in the developing world were a serious impediment to economic development. The industrialized countries had much higher rates of saving, but international capital flows from the industrialized to the developing world were insufficient to finance needed investment in infrastructure and industrial enterprise. Since that time, the developing countries have followed divergent paths. Saving rates in sub-Saharan Africa have remained low. After increasing somewhat in the 1970s, saving rates in Latin America declined during the late 1970s and early 1980s. In contrast, South Asian and especially East Asian saving rates have increased substantially and currently are well above saving rates found in the industrialized countries. The emergence of high saving rates is widely believed to bear major responsibility for Asia's rapid economic growth up until the mid-1990s.

The industrialized countries face their own saving issue. Since the mid-1970s, they have experienced a gradual, but steady, decline in saving rates. US saving rates have reached especially low levels. Current US household saving rates are near zero. The low rates of saving raise two concerns: first, that economic growth will be unsustainable and, second, that current generations of workers will face substantially reduced standards of living when they retire.

2. *General Saving Concepts*

The national saving rate is the outcome of decisions by three sets of actors—governments, firms, and households—but analyses of saving rates are overwhelmingly based on household behavioral models. The reliance on household models is justified on two grounds. First, decisions by households may fully incorporate the decisions made by firms and governments. Firms are owned by households. When firms accumulate wealth, the wealth of households increases as well. Thus, from the perspective of the household, saving by firms and saving by households are close substitutes. Consequently, household behavior determines the private saving rate, the combined saving of firms and households. Government saving also affects the wealth of the household sector by affecting current and future taxes. By issuing debt governments can, in principle, increase consumption and reduce national saving at the expense of future generations. However, households may choose to compensate future generations (their children) by increasing their saving and planned bequests. If they do so, national saving is determined entirely by the household sector and is independent of government saving (Barro 1974).

Second, firms and governments may act as agents for households. Their saving on behalf of households may be influenced by the same factors that influence household behavior. The clearest example of this type of behavior is when firms or governments accumulate pension funds on behalf of their workers or citizens.

Households are motivated to save for a number of reasons, but current research emphasizes three motives: insurance, bequests, and lifecycle motives. Households may accumulate wealth to insure themselves against uncertain events, e.g., the loss of a crop or a job, the death of a spouse, or an unanticipated medical expense. Households may save in order to accumulate an estate intended for their descendents. Household saving may be motivated by lifecycle concerns, the divergence between the household's earnings profile and its preferred consumption profile.

The importance of these and other motives in determining national saving rates is a matter of vigorous debate among economists. Proponents of the bequest motive have estimated that as much as 80 percent of US national saving is accounted for by the bequest motive (Kotlikoff 1988). Proponents of the

lifecycle motive have estimated that an equally large portion of US national saving is accounted for by the lifecycle motive (Modigliani 1988). Clearly, the way in which changing demographic conditions impinge on national saving rates will depend on the importance of these different motives.

3. The Lifecycle Model

Both the larger debate on the determinants of saving and explorations of the impact on saving of demographic factors have been framed by the lifecycle model (Modigliani and Brumberg 1954). The key idea that motivates the model is the observation that for extended portions of our lives we are incapable of providing for our own material needs. Thus, economic resources must be reallocated from economically productive individuals concentrated at the working ages to dependents concentrated at young or old ages. Several mechanisms exist for achieving this reallocation. In traditional societies, the family plays a dominant role. Typically, the young, the old, and those of working age live in extended families supported by productive family members. In modern societies, solving the lifecycle problem is a shared responsibility. Although the young continue to rely primarily on the family, the elderly rely on the family, on transfers from workers effected by the government, and on personnel wealth they have accumulated during their working years (Lee 2000).

The lifecycle saving model is most clearly relevant to higher income settings where capital markets have developed, family support systems have eroded, and workers anticipate extended periods of retirement. Under these conditions, the lifecycle model implies that households consisting of young, working age adults will save, while household consisting of old, retired adults will dis-save. The national saving rate is determined, in part, by the relative size of these demographic groups. A young age structure yields high saving rates; an old age structure yields low saving rates. Likewise, a rise in the rate of population growth leads to higher saving rates because of the shift to a younger age structure.

In the standard formulation of the lifecycle model, changes in the growth rate of per capita income operate in exactly the same way as changes in the population growth rate. Given higher long-run rates of economic growth, young adults have greater lifetime earnings than older adults. They have a correspondingly greater impact on the aggregate saving rate because of their control of a larger share of economic resources. Consequently, an increase in either the population growth rate or the per capita income growth rate leads to higher saving.

The rate of growth effect is one of the most important and widely tested implications of the lifecycle saving model. With a great deal of consistency, empirical studies have found that an increase in the rate of per capita income growth leads to an increase in the national saving rate. Empirical research does not, however, support the existence of a positive population rate of growth effect (Deaton 1989).

The standard lifecycle model provides no basis for reconciling the divergent rate of growth effects. The impact of child dependency on household saving provides one possible explanation of why population growth and economic growth need not have the same effect on aggregate saving. Coale and Hoover (1958) were the first to point out the potentially important impact of child dependency. They hypothesized that 'A family with the same total income but with a larger number of children would surely tend to consume more and save less, other things being equal' (p. 25). This raised the possibility that saving follows an inverted-U shaped curve over the demographic transition. In low-income countries, with rapid population growth rates and young age structure, slower population growth would lead to higher saving. But in higher income countries that were further along in their demographic transitions, slower population growth would lead to a population concentrated at older, low saving ages as hypothesized in the lifecycle model. These contrasting demographic effects have been modeled in the empirical literature using the youth and old-age dependency ratios.

The variable lifecycle model incorporates the role of child dependency by allowing demographic factors to influence both the age structure of the population and the age profiles of consumption, earning, and, hence, saving. In this version of the lifecycle model the size of the rate of growth effect varies depending among other things on the number or cost of children. As with the dependency ratio model, saving follows an inverted-U shaped path over the demographic transition. However, the impact of demographic factors varies with the rate of economic growth. The model implies that in countries with rapid economic growth, e.g., East Asia, demographic factors will have a large impact on saving, but in countries with slow economic growth, e.g., Africa, demographic factors would have a more modest effect (Mason 1987).

Empirical studies of population and saving come in two forms. Most frequently, researchers have based their analyses on aggregate time series data now available for many countries. Recent estimates support the existence of large demographic effects on saving. Analysis by Kelley and Schmidt (1996) supports the variable lifecycle model. Higgins and Williamson (1997) find that demographic factors influence saving independently of the rate of economic growth. An alternative approach relies on microeconomic data to construct an age-saving profile. The impact of age structure is then assessed assuming that the age profile does not change. In the few applications undertaken, changes in age structure had an impact on saving that is more modest than found in analyses of

aggregate saving data. (See Deaton and Paxson (1997) for an example.) Until these approaches are reconciled, a firm consensus about the magnitude of demographic effects is unlikely to emerge.

4. Unresolved Issues

There are a number of important issues that have not been resolved and require additional work. First, the saving literature does not yet adequately incorporate the impact of changing institutional arrangements. Studies of saving in the industrialized countries and recent work on developing countries considers the impact of state-sponsored pension programs on saving, but the impact of family support systems has received far too little emphasis. The erosion of the extended family in developing countries is surely one of the factors that has contributed to the rise of saving rates observed in many developing countries.

Second, the role of mortality has received inadequate attention. The importance of lifecycle saving depends on the expected duration of retirement. In high mortality societies, few reach old age and many who do continue to work. Only late in the mortality transition, when there are substantial gains in the years lived late in life, does an important pension motive emerge.

Third, the saving models currently in use are static models and do not capture important dynamics. Recent simulation work that combines realistic demographics with lifecycle saving behavior shows that during the demographic transition countries may experience saving rates that substantially exceed equilibrium values for sustained periods of time (Lee 2000).

Bibliography

Barro R J 1974 Are government bonds net wealth? *Journal of Political Economy* **6** (December): 1095–117

Coale C A, Hoover E M 1958 *Population Growth and Economic Development in Low-income Countries: A Case Study of India's Prospects*. Princeton University Press, Princeton, NJ

Deaton A1989 Saving in developing countries: Theory and review. *Proceedings of the World Bank Annual Conference on Development Economics*, Supplement to the World Bank Economic Review and the World Bank Research Observer, pp. 61–96

Deaton A, Paxson C 1997 The effects of economic and population growth on national saving and inequality. *Demography* **34**(1): 97–114

Higgins M, Williamson J G 1997 Age structure dynamics in Asia and dependence on foreign capital. *Population and Development Review* **23**(2): 261–94

Kelley A C, Schmidt R M 1996 Saving, dependency and development. *Journal of Population Economics* **9**(4): 365–86

Kotlikoff L J 1988 Intergenerational transfers and savings. *Journal of Economic Perspectives* **2**(2): 41–58

Lee R D 2000 Intergenerational transfers and the economic life cycle: A cross-cultural perspective. In: Mason A, Tapinos G (eds.) *Sharing the Wealth: Demographic Change and Economic Transfers Between Generations*. Oxford University Press, Oxford, UK

Lee R D, Mason A, Miller T 2000 Life cycle saving and the demographic transition: The case of Taiwan. In: Chu C Y, Lee R D (eds.) *Population and Economic Change in East Asia, Population and Development Review*. **26**: 194–219

Mason A 1987 National saving rates and population growth: A new model and new evidence. In: Johnson D G, Lee R D (eds.) *Population Growth and Economic Development: Issues and Evidence*. University of Wisconsin Press, Madison, WI, pp. 5230–60

Modigliani F 1988 The role of intergenerational transfers and life cycle saving in the accumulation of wealth. *Journal of Economic Perspectives* **2**(2): 15–40

Modigliani F, Brumberg R 1954 Utility analysis and the consumption function: An interpretation of cross-section data. In: Kurihara K (ed.) *Post-Keynesian Economics*. Rutgers University Press, New Brunswick, NJ

A. Mason

Scale in Geography

Scale is about size, either relative or absolute, and involves a fundamental set of issues in geography. Scale primarily concerns space in geography, and this article will focus on spatial scale. However, the domains of temporal and thematic scale are also important to geographers. Temporal scale deals with the size of time units, thematic scale with the grouping of entities or attributes such as people or weather variables. Whether spatial, temporal, or thematic, scale in fact has several meanings in geography.

1. Three Meanings of Scale

The concept of scale can be confusing, insofar as it has multiple referents. Cartographic scale refers to the depicted size of a feature on a map relative to its actual size in the world. Analysis scale refers to the size of the unit at which some problem is analyzed, such as at the county or state level. Phenomenon scale refers to the size at which human or physical earth structures or processes exist, regardless of how they are studied or represented. Although the three referents of scale frequently are treated independently, they are in fact interrelated in important ways that are relevant to all geographers, and the focus of research for some. For example, choices concerning the scale at which a map

should be made depend in part on the scale at which measurements of earth features are made and the scale at which a phenomenon of interest actually exists.

1.1 Cartographic Scale

Maps are smaller than the part of the earth's surface they depict. Cartographic scale expresses this relationship, traditionally in one of three ways. A verbal scale statement expresses the amount of distance on the map that represents a particular distance on the earth's surface in words, e.g., 'one inch equals a mile.' The representative fraction (RF) expresses scale as a numerical ratio of map distance to earth distance, e.g., '1:63,360.' The RF has the advantage of being a unitless measure. Finally, a graphic scale bar uses a line of particular length drawn on the map and annotated to show how much earth distance it represents. A graphic scale bar has the advantage that it changes size appropriately when the map is enlarged or reduced. Alternatively, all three expressions of scale may refer to areal measurements rather than linear measurements, e.g., a 1-inch square may represent 1 square mile on the earth.

Given a map of fixed size, as the size of the represented earth surface gets larger, the RF gets smaller (i.e., the denominator of the RF becomes a larger number). Hence, a 'large-scale map' shows a relatively small area of the earth, such as a county or city, and a 'small-scale map' shows a relatively large area, such as a continent or a hemisphere of the earth. This cartographic scale terminology is frequently felt to be counterintuitive when applied to analysis or phenomenon scale, where small-scale and large-scale usually refer to small and large entities, respectively.

An important complexity about cartographic scale is that flat maps invariably distort spatial relations on the earth's surface: distance, direction, shape, and/or area. How they distort these relations is part of the topic of map projections. In many projections, especially small-scale maps that show large parts of the earth, this distortion is extreme so that linear or areal scale on one part of the map is very different than on other parts. Even so-called equal area projections maintain equivalent areal scale only for particular global features, and not for all features at all places on the map. Variable scale is sometimes shown on a map by the use of a special symbol or multiple symbols at different locations.

1.2 Analysis Scale

Analysis scale includes the size of the units in which phenomena are measured and the size of the units into which measurements are aggregated for data analysis and mapping. It is essentially the scale of understanding of geographic phenomena. Terms such as 'resolution' or 'granularity' are often used as synonyms for the scale of analysis, particularly when geographers work with digital representations of the earth's surface in a computer by means of a regular grid of small cells in a satellite image (rasters) or on a computer screen (pixels). Analysis scale here refers to the area of earth surface represented by a single cell.

It has long been recognized that in order to observe and study a phenomenon most accurately, the scale of analysis must match the actual scale of the phenomenon. This is true for all three domains of scale—spatial, temporal, and thematic. Identifying the correct scale of phenomena is, thus, a central problem for geographers. Particularly when talking about thematic scale, using data at one scale to make inferences about phenomena at other scales is known as the cross-level fallacy (the narrower case of using aggregated data to make inferences about disaggregated data is well-known as the ecological fallacy).

Geographers often analyze phenomena at what might be called 'available scale,' the units that are present in available data. Many problems of analysis scale arise from this practice, but it is unavoidable given the difficulty and expense involved in collecting many types of data over large parts of the earth's surface. Geographers have little choice in some cases but to analyze phenomena with secondary data, data collected by others not specifically for the purposes of a particular analysis. For example, census bureaus in many countries provide a wealth of data on many social, demographic, and economic characteristics of their populace. Frequently, the phenomenon of interest does not operate according to the boundaries of existing administrative or political units in the data, which after all were not created to serve the needs of geographic analysis. The resolution of image scanners on remote-sensing satellites provides another important example. Landsat imagery is derived from thematic mapper sensors, producing earth measurements at a resolution of about 30 by 30 meters. However, many phenomena occur at finer resolutions than these data can provide.

Most useful is theory about the scale of a phenomenon's existence. Frequently lacking this, but realizing that the available scale may not be suitable, geographers use empirical 'trial-and-error' approaches to try to identify the appropriate scale at which a phenomenon should be analyzed. Given spatial units of a particular size, one can readily aggregate or combine them into larger units; it is not possible without additional information or theory to disaggregate them into smaller units. Even given observations measured at very small units, however, there is still the problem of deciding in what way the units should be aggregated. This is known as the modifiable areal unit problem (MAUP, or MTUP in the case of temporal scale). Various techniques have been developed to study the implications of MAUP (Openshaw 1983).

1.3 Phenomenon Scale

Phenomenon scale refers to the size at which geographic structures exist and over which geographic processes operate in the world. It is the 'true' scale of geographic phenomena. Determining the scale of phenomena is clearly a major research goal in geography. It is a common geographic dictum that scale matters. Numerous concepts in geography reflect the idea that phenomena are scale-dependent or are defined in part by their scale. Vegetation stands are smaller than vegetation regions, and linguistic dialects are distributed over smaller areas than languages. The possibility that some geographic phenomena are scale independent is important, however. Patterns seen at one scale may often be observed at other scales; whether this is a matter of analogy or of the same processes operating at multiple scales is theoretically important. The mathematics of fractals has been applied in geography as a way of understanding and formalizing phenomena such as coastlines that are self-similar at different scales (Lam and Quattrochi 1992).

The belief has often been expressed that the discipline of geography, as the study of the earth as the home of humanity, can be defined partially by its focus on phenomena at certain scales, such as cities or continents, and not other scales. The range of scales of interest to geographers are often summarized by the use of terminological continua such as 'local-global' or 'micro-, meso-, macroscale.' The view that geographers must restrict their focus to particular ranges of scales is not shared universally, however, and advances have and will continue to occur when geographers stretch the boundaries of their subject matter. Nonetheless, few would argue that subatomic or interplanetary scales are properly of concern for geography.

It is widely recognized that various scales of geographic phenomena interact, or that phenomena at one scale emerge from smaller or larger scale phenomena. This is captured by the notion of a 'hierarchy of scales,' in which smaller phenomena are nested within larger phenomena. Local economies are nested within regional economies, rivers are nested within larger hydrologic systems. Conceptualizing and modeling such scale hierarchies can be quite difficult, and the traditional practice within geography of focusing on a single scale largely continues.

2. Generalization

The world can never be studied, modeled, or represented in all of its full detail and complexity. Scale is important in part because of its consequences for the degree to which geographic information is generalized. Generalization refers to the amount of detail included in information; it is essentially an issue of simplification, but also includes aspects of selection and enhancement of features of particular interest. As one studies or represents smaller pieces of the earth, one tends strongly to deal with more detailed or more fine-grained aspects of geographic features. For example, large-scale maps almost always show features on the earth's surface in greater detail than do small-scale maps; rivers appear to meander more when they are shown on large-scale maps, for instance. Studied most extensively by cartographers, generalization is in fact relevant to all three meanings of scale, and to all three domains of spatial, temporal, and thematic scale.

3. Conclusion

Issues of scale have always been central to geographic theory and research. Advances in the understanding of scale and the ability to investigate scale-related problems will continue, particularly with the increasingly common representation of geographic phenomena through the medium of digital geographic information (Goodchild and Proctor 1997). Cartographic scale is becoming 'visualization' scale. How is scale, spatial and temporal, communicated in dynamic, multidimensional, and multimodal representations, including visualization in virtual environments? Progress continues on the problem of automated generalization, programming intelligent machines to make generalization changes in geographic data as scale changes. The ability to perform multiscale and hierarchical analysis will be developed further. More profound than these advances, however, the widespread emergence of the 'digital world' will foster new conceptions of scale in geography.

Bibliography

Buttenfield B P, McMaster R B (eds.) 1991 *Map Generalization: Making Rules for Knowledge Representation*. Wiley, New York

Goodchild M F, Proctor J 1997 Scale in a digital geographic world. *Geographical and Environmental Modeling* **1**: 5–23

Hudson J C 1992 Scale in space and time. In: Abler R F, Marcus M G, Olson J M (eds.) *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography*. Rutgers University Press, New Brunswick, NJ

Lam N S-N, Quattrochi D A 1992 On the issues of scale, resolution, and fractal analysis in the mapping sciences. *The Professional Geographer* **44**: 88–98

MacEachren A M 1995 *How Maps Work: Representation, Visualization, and Design*. Guilford Press, New York

Meyer W B, Gregory D, Turner B L, McDowell P F 1992 The local-global continuum. In: Abler R F, Marcus M G, Olson J M (eds.) *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography*. Rutgers University Press, New Brunswick, NJ

Muehrcke P C, Muehrcke J O 1992 *Map Use: Reading, Analysis, Interpretation*, 3rd edn. JP Publications, Madison, WI
Openshaw S 1983 *The Modifiable Areal Unit Problem*. Geo Books, Norfolk, UK

D. R. Montello

Scaling and Classification in Social Measurement

Social measurements translate observed characteristics of individuals, events, relationships, organizations, societies, etc. into symbolic classifications that enable reasoning of a verbal, logical, or mathematical nature. Qualitative research and censuses together define one realm of measurement, concerned with assignment of entities to classification categories embedded within taxonomies and typologies. Another topic in measurement involves scaling discrete items of information such as answers to questions so as to produce quantitative measurements for mathematical analyses. A third issue is the linkage between social measurements and social theories.

1. *Classifications*

Classification assimilates perceived phenomena into symbolically labeled categories. Anthropological studies of folk classification systems (D'Andrade 1995) have advanced understanding of scientific classification systems, though scientific usages involve criteria that folk systems may not meet entirely.

Two areas of social science employ classification systems centrally. Qualitative analyses such as ethnographies, histories, case studies, etc. offer classifications—sometimes newly invented—for translating experiences in unfamiliar cultures or minds into familiar terms. Censuses of individuals, of occurrences, or of aggregate social units apply classifications—usually traditional—in order to count entities and their variations. Both types of work depend on theoretical constructions that link classification categories.

1.1 *Taxonomies*

Every classification category is located within a taxonomy. Some more general categorization, Y, determines which entities are in the domain for the focal categorization, X; so an X always must be a kind of Y. 'X is a kind of Y' is the linguistic frame for specifying taxonomies. Concepts constituting a taxonomy form a logic tree, with subordinate elements implying superordinate items.

Taxonomic enclosure of a classification category is a social construction that may have both theoretical and practical consequences. For example, if only violent crimes are subject to classification as homicides, then 'homicide' is a kind of 'violent crime,' and deaths caused by executive directives to release deadly pollution could not be homicides.

1.2 *Typologies*

A typology differentiates entities at a particular level of a taxonomy in terms of one or more of their properties. The differentiating property (sometimes called a feature or attribute) essentially acts as a modifier of entities at that taxonomic level. For example, in the USA kinship system siblings are distinguished in terms of whether they are male or female; in Japan by comparison, siblings are schematized in terms of whether they are older as well as whether they are male or female.

A scientific typology differentiates entities into types that are exclusive and exhaustive: every entity at the relevant taxonomic level is of one defined type only, and every entity is of some defined type. A division into two types is a dichotomy, into three types a trichotomy, and into more than three types a polytomy.

Polytomous typologies are often constructed by crossing multiple properties, forming a table in which each cell is a theoretical type. (The crossed properties might be referred to as variables, dimensions, or factors in the typology.) For example, members of a multiplex society have been characterized according to whether they do or do not accept the society's goals on the one hand, and whether they do or do not accept the society's means of achieving goals on the other hand; then crossing acceptance of goals and means produces a fourfold table defining conformists and three types of deviants.

Etic–emic analysis involves defining a typology with properties of scientific interest (the etic system) and then discovering ethnographically which types and combinations of types are recognized in folk meanings (the emic system). Latent structure analysis statistically processes observed properties of a sample of entities in order to confirm the existence of hypothesized types and to define the types operationally.

1.3 *Aggregate Entities*

Aggregate social entities such as organizations, communities, and cultures may be studied as unique cases, where measurements identify and order internal characteristics of the entity rather than relate one aggregate entity to another.

A seeming enigma in social measurement is how aggregate social entities can be described satisfactorily on the basis of the reports of relatively few informants, even though statistical theory calls for substantial samples of respondents to survey populations. The

key is that informants all report on the same thing—a single culture, community, or organization—whereas respondents in a social survey typically report on diverse things—their own personal characteristics, beliefs, or experiences. Thus, reports from informants serve as multiple indicators of a single state, and the number needed depends on how many observations are needed to define a point reliably, rather than how many respondents are needed to describe a population's diversity reliably. As few as seven expert informants can yield reliable descriptions of aggregate social entities, though more are needed as informants' expertise declines (Romney et al. 1986). Informant expertise correlates with greater intelligence and experience (D'Andrade 1995) and with having a high level of social integration (Thomas and Heise 1995).

2. Relations

Case grammar in linguistics defines events and relationships in terms of an actor, action, object, and perhaps instrumentation, setting, products, and other factors as well. Mapping sentences (Shye et al. 1994) apply the case grammar idea with relatively small lists of entities in order to classify relational phenomena within social aggregates. For example, interpersonal relations in a group can be specified by pairing group members with actions such as loves, admires, annoys, befriends, and angers. Mapping sentences defining the relations between individuals or among social organizations constitute the measurement model for social network research.

3. Scaling

Quantitative measurements differentiate entities at a given taxonomic level—serving like typological classifications, but obtaining greater logical and mathematical power by ordering the classification categories.

An influential conceptualization (Stevens 1951) posited four levels of quantification in terms of how numbers relate to classification categories. *Nominal* numeration involves assigning numbers arbitrarily simply to give categories unique names, such as 'batch 243.' An *ordinal* scale's categories are ordered monotonically in terms of greater-than and less-than, and numbering corresponds to the rank of each category. Numerical ranking of an individual's preferences for different foods is an example of ordinal measurement. Differences between categories can be compared in an *interval* scale, and numbers applied to categories reflect degrees of differences. Calendar dates are an example of interval measurements—we know from their birth years that William Shakespeare was closer in time to Geoffrey Chaucer than Albert Einstein was to Isaac Newton. In a *ratio* scale categories have magnitudes that are whole or fractional multiples of one another, and numbers assigned to the categories represent these magnitudes. Population sizes are an example of ratio measurements—knowing the populations of both nations, we can say that Japan is at least 35 times bigger than Jamaica.

A key methodological concern in psychometrics (Hopkins 1998) has been: how do you measure entities on an interval scale given merely nominal or ordinal information?

3.1 Scaling Dichotomous Items

Nominal data are often dichotomous yes–no answers to questions, a judge's presence–absence judgments about the features of entities, an expert's claims about truth–falsity of propositions, etc. Answers of *yes*, *present*, *true*, etc. typically are coded as 'one' and *no*, *absent*, *false*, etc. as 'zero.' The goal, then, is to translate zero-one answers for each case in a sample of entities into a number representing the case's position on an interval scale of measurement.

The first step requires identifying how items relate to the interval scale of measurement in terms of a graph of the items' characteristic curves. The horizontal axis of such a graph is the interval scale of measurement, confined to the practical range of variation of entities actually being observed. The vertical axis indicates probability that a specific dichotomous item has the value one for an entity with a given position on the interval scale of measurement. An item's characteristic curve traces the changing probability of the item having the value one as an entity moves from having a minimal value on the interval scale to having the maximal value on the interval scale.

Item characteristic curves have essentially three different shapes, corresponding to three different formulations about how items combine into a scale.

Spanning items have characteristic curves that essentially are straight lines stretching across the range of entity variation. A spanning item's line may start as a low probability value and rise to a high probability value, or fall from a high value to a low value. A rising line means that the item is unlikely to have a value of one with entities having a low score on the interval scale; the item is likely to have a score of one for entities having a high score on the interval scale; and the probability of the item being valued at one increases regularly for entities between the low and high positions on the scale.

Knowing an entity's value on any one spanning item does not permit assessing the entity's position along the interval scale. However, knowing the entity's values on multiple spanning items does allow an estimate of positioning to be made. Suppose heuristically that we are working with a large number of equivalent spanning items, each having an item characteristic curve that starts at probability 0.00 at the minimal point of the interval scale, and rises in a

straight line to probability 1.00 at the maximal point of the interval scale. The probability of an item being valued at one can be estimated from the observed proportion of all these items that are valued at one—which is simply the mean item score when items are scored zero-one. Then we can use the characteristic curve for the items to find the point on the interval scale where the entity must be positioned in order to have the estimated item probability. This is the basic scheme involved in construction of *composite scales*, where an averaged or summated score on multiple items is used to estimate an entity's interval-scale value on a dimension of interest (Lord and Novick 1968).

The more items that are averaged, the better the estimate of an entity's position on the interval scale. The upper bound on number of items is pragmatic, determined by how much precision is needed and how much it costs to collect data with more items. The quality of the estimate also depends on how ideal the items are in terms of having straight-line characteristic curves terminating at the extremes of probability. Irrelevant items with a flat characteristic curve would not yield an estimate of scale position no matter how many of them are averaged, because a flat curve means that the probability of the item having a value of one is uncorrelated with the entity's position on the interval scale. Inferences are possible with scales that include relevant but imperfect items, but more items are required to achieve a given level of precision, and greater weight needs to be given to the more perfect items.

Declivitous items have characteristic curves that rise sharply at a particular point on the horizontal axis. Idealized, the probability of the item having a value of one increases from 0.00 to the left of the inflection point to 1.00 to the right of the inflection point; or alternatively the probability declines from 1.00 to 0.00 in passing the inflection point. Realistically, the characteristic curve of a declivitous item is S-shaped with a steep rise in the middle and graduated approaches to 0.00 at the bottom and to 1.00 at the top.

The value of a single declivitous item tells little about an entity's position along the interval scale. However, an inference about an entity's scale position can be made from a set of declivitous items with different inflection points, or difficulties, that form a *cumulative scale*. Suppose heuristically that each item increases stepwise at its inflection point. Then for an entity midway along the horizontal axis, items at the left end of the scale will all have the value of one, items at the right end of the scale will all have the value zero, and the entity's value on the interval scale is between the items with a score of one and the items with a score of zero.

If the items' inflection points are evenly distributed along the interval scale, then the sum of items' zero-one scores for an entity constitutes an estimate of where the entity is positioned along the interval scale. That is, few of the items have a value of one if the entity is on the lower end of the interval scale, and many of the items are valued at one if the entity is at the upper end of the interval scale. This is the basic scheme involved in Guttman scalogram analysis (e.g., see Shye 1978, Part 5). On the other hand, we might use empirical data to estimate the position of each item's inflection point on the interval scale, while simultaneously estimating entity scores that take account of the item difficulties. This is the basic scheme involved in scaling with Rasch models (e.g., Andrich 1988).

Entities' positions on the interval scale can be pinned down as closely as desired through the use of more declivitous items with inflection points spaced closer and closer together. However, adding items to achieve more measurement precision at the low end of the interval scale does not help at the middle or the high end of the interval scale. Thus, obtaining high precision over the entire range of variation requires a large number of items, and it could be costly to obtain so much data. Alternatively, one can seek items whose characteristic curves rise gradually over a range of the interval scale such that sequential items on the scale have overlapping characteristic curves, whereby an entity's position along the interval scale is indicated by several items.

Regional items have characteristic curves that rise and fall within a limited range of the interval scale. That is, moving an entity up the interval scale increases the probability of a particular item having a value of one for a while, and then decreases the probability after the entity passes the characteristic curve's maximum value. For example, in a scale measuring prejudice toward a particular ethnic group, the probability of agreeing with the item 'they require equal but separate facilities' increases as a person moves away from an apartheid position, and then decreases as the person moves further up the scale toward a nondiscriminatory position. A regional item's characteristic curve is approximately bell-shaped if its maximum is at the middle of the interval scale, but characteristic curves at the ends of the scale are subject to floor and ceiling clipping, making them look like declivitous items.

If an entity has a value of one on a regional item, then the entity's position along the interval scale is known approximately, since the entity must be positioned in the part of the scale where that item has a nonzero probability of being valued at one. However, a value of zero on the same item can result from a variety of positions along the interval scale and reveals little about the entity's position. Thus, multiple regional items have to be used to assess positions along the whole range of the scale. The items have to be sequenced relatively closely along the scale with overlapping characteristic curves so that no entity will end up in the noninformative state of having a zero value on all items.

One could ask judges to rate the scale position of each item, and average across judges to get an item score; then, later, respondents can be scored with the average of the items they accept. This simplistic approach to regional items was employed in some early attempts to measure social attitudes. Another approach is statistical unfolding of respondents' choices of items on either side of their own positions on the interval scale in order to estimate scale values for items and respondents simultaneously (Coombs 1964).

Item analysis is a routine aspect of scale construction with spanning items, declivitous items, or regional items. One typically starts with a notion of what one wants to measure, assembles items that should relate to the dimension, and tests the items in order to select the items that work best. Since a criterion measurement that can be used for assessing item quality typically is lacking, the items as a group are assumed to measure what they are supposed to measure, and scores based on this assumption are used to evaluate individual items.

Items in a scale are presumed to measure a single dimension rather than multiple dimensions. Examining the dimensionality assumption brings in additional technology, such as component analysis or factor analysis in the case of spanning items, multidimensional scalogram analysis in the case of declivitous items, and nonmetric multidimensional scaling in the case of regional items. These statistical methods help in refining the conception of the focal dimension and in selecting the best items for measuring that dimension.

3.2 Ordered Assessments

Ordinal data—the starting point for a three-volume mathematical treatise on measurement theory (Krantz et al. 1971, Suppes et al. 1989, Luce et al. 1990)—may arise from individuals' preferences, gradings of agreement with statements, reckonings of similarity between stimuli, etc. Conjoint analysis (Luce and Tukey 1964, Michell 1990) offers a general mathematical model for analyzing such data. According to the conjoint theory of measurement, positions on any viable quantitative dimension are predictable from positions on two other quantitative dimensions, and this assumption leads to tests of a dimension's usefulness given just information of an ordinal nature. For example, societies might be ranked in terms of their socioeconomic development and also arrayed in terms of the extents of their patrifocal technologies (like herding) and matrifocal technologies (like horticulture), each of which contributes additively to socioeconomic development. Conjoint analyses could be conducted to test the meaningfulness of these dimensions, preliminarily to developing interval scales of socioeconomic development and patrifocal and matrifocal technologies. Specific scaling methodologies, like Rasch scaling and nonmetric multidimensional scaling, can be interpreted within the conjoint analysis framework.

Magnitude estimations involve comparisons to an anchor, for example: 'Here is a reference sound.... How loud is this next sound relative to the reference sound.' Trained judges using such a procedure can assess intensities of sensations and of a variety of social opinions on ratio scales (Stevens 1951, Lodge 1981). Comparing magnitude estimation in social surveys to the more common procedure of obtaining ratings on category scales with a fixed number of options, Lodge (1981) found that magnitude estimations are more costly but more accurate, especially in registering extreme positions.

Rating scales with bipolar adjective anchors like good–bad are often used to assess affective meanings of perceptions, individuals, events, etc. Such scales traditionally provided seven or nine answer positions between the opposing poles with the middle position defined as neutral. Computerized presentations of such scales with hundreds of rating positions along a graphic line yield greater precision by incorporating some aspects of magnitude estimation. Cross-cultural and cross-linguistic research in dozens of societies has demonstrated that bipolar rating scales align with three dimensions—evaluation, potency, and activity (Osgood et al. 1975). An implication is that research employing bipolar rating scales should include scales measuring the standard three dimensions in order to identify contributions of these dimensions to rating variance on other bipolar scales.

4. Measurements and Theory

Theories and measurements are bound inextricably. In the first place, taxonomies and typologies—which are theoretical constructions, even when rooted in folk classification systems—are entailed in defining which entities are to be measured, so are part of all measurements.

Second, scientists routinely assume that any particular measurement is wrong to some degree—even a measurement based on a scale, and that combining multiple measurements improves measurement precision. The underlying premise is that a true value exists for that which is being measured, as opposed to observed values, and that theories apply to true values, not ephemeral observed values. In essence, theories set expectations about what should be observed, in contrast to what is observed, and deviation from theoretical expectations is interpreted as measurement error (Kyburg 1992). A notion of measurement error as deviation from theoretical expectations is widely applicable, even in qualitative research (McCullagh and Behan 1984, Heise 1989).

Third, the conjoint theory of measurement underlying many current measurement technologies requires theoretical specification of relations between variables

before the variables' viability can be tested or meaningful scales constructed. This approach is in creative tension with traditional deductive science, wherein variable measurements are gathered in order to determine if relations among variables exist and theories are correct. In fact, a frequent theme in social science during the late twentieth century was that measurement technology had to improve in order to foster the growth of more powerful theories. It is clear now that the dependence between measurements and theories is more bidirectional than was supposed before the development of conjoint theory.

See also: Classification: Conceptions in the Social Sciences; Dimensionality of Tests: Methodology; Factor Analysis and Latent Structure: IRT and Rasch Models; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Andrich D 1988 *Rasch Models for Measurement*. Sage, Newbury Park, CA

Coombs C H 1964 *A Theory of Data*. John Wiley & Sons, New York

D'Andrade R 1995 *The Development of Cognitive Anthropology*. Cambridge University Press, New York

Heise D R 1989 Modeling event structures. *Journal of Mathematical Sociology* **14**: 139–69

Hopkins K D 1998 *Educational and Psychological Measurement and Evaluation*, 8th edn. Allyn and Bacon, Boston

Krantz D H, Luce R D, Suppes P, Tversky A 1971 *Foundations of Measurement. Volume 1: Additive and Polynomial Representations*. Academic Press, New York

Kyburg Jr. H E 1992 Measuring errors of measurement. In: Savage C W, Ehrlich P (eds.) *Philosophical and Foundational Issues in Measurement*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 75–91

Lodge M 1981 *Magnitude Scaling: Quantitative Measurement of Opinions*. Sage, Beverly Hills, CA

Lord F N, Novick M R 1968 *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA

Luce R D, Krantz D H, Suppes P, Tversky A 1990 *Foundations of Measurement. Volume 3: Representation, Axiomatization, and Invariance*. Academic Press, New York

Luce R D, Tukey J W 1964 Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology* **1**: 1–27

McCullagh B C, Behan C 1984 *Justifying Historical Descriptions*. Cambridge University Press, New York

Michell J 1990 *An Introduction to the Logic of Psychological Measurement*. Lawrence Erlbaum Associates, Hillsdale, NJ

Osgood C, May W H, Miron M S 1975 *Cross-cultural Universals of Affective Meaning*. University of Illinois Press, Urbana, IL

Romney A K, Weller S C, Batchelder W H 1986 Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist* **88**: 313–38

Shye S (ed.) 1978 *Theory Construction and Data Analysis in the Behavioral Sciences*, 1st edn. Jossey-Bass, San Francisco

Shye S, Elizur D, Hoffman M 1994 *Introduction to Facet Theory: Content Design and Intrinsic Data Analysis in Behavioral Research*. Sage, Thousand Oaks, CA

Stevens S S 1951 Mathematics, measurement, and psychophysics. In: Stevens S S (ed.) *Handbook of Experimental Psychology*. John Wiley and Sons, New York, pp. 1–49

Suppes P, Krantz D H, Luce R D, Tversky A 1989 *Foundations of Measurement. Volume 2: Geometrical, Threshold, and Probabilistic Representations*. Academic Press, New York

Thomas L, Heise D R 1995 Mining error variance and hitting pay-dirt: Discovering systematic variation in social sentiments. *The Sociological Quarterly* **36**: 425–39

D. R. Heise

Scaling: Correspondence Analysis

In the early 1960s a dedicated group of French social scientists, led by the extraordinary scientist and philosopher Jean-Paul Benzécri, developed methods for structuring and interpreting large sets of complex data. This group's method of choice was correspondence analysis, a method for transforming a rectangular table of data, usually counts, into a visual map which displays rows and columns of the table with respect to continuous underlying dimensions. This article introduces this approach to scaling, gives an illustration and indicates its wide applicability. Attention is limited here to the descriptive and exploratory uses of correspondence analysis methodology. More formal statistical tools have recently been developed and are described in *Multivariate Analysis: Discrete Variables (Correspondence Models)*.

1. Historical Background

Benzécri's contribution to data analysis in general and to correspondence analysis in particular was not so much in the mathematical theory underlying the methodology as in the strong attention paid to the graphical interpretation of the results and in the broad applicability of the methods to problems in many contexts. His initial interest was in analyzing large sparse matrices of word counts in linguistics, but he soon realized the power of the method in fields as diverse as biology, archeology, physics, and music. The fact that his approach paid so much attention to the visualization of data, to be interpreted with a degree of ingenuity and insight into the substantive problem, fitted perfectly the *esprit géometrique* of the French and their tradition of visual abstraction and creativity.

Originally working in Rennes in western France, this group consolidated in Paris in the 1970s to become an influential and controversial movement in post-1968 France. In 1973 they published the two fundamental volumes of, *L'Analyse des Données* (Data Analysis), the first on *La Classification*, that is,

unsupervised classification or cluster analysis, and the second on, *L'Analyse des Correspondances*, or correspondence analysis (Benzécri 1973), as well as from 1977 the journal *Les Cahiers de l'Analyse des Données*, all of which reflect the depth and diversity of Benzécri's work. For a more complete historical account of the origins of correspondence analysis, see Nishisato (1980), Greenacre (1984), and Gifi (1990).

2. *Correspondence Analysis*

Correspondence analysis (CA) is a variant of principal components analysis (PCA) applicable to categorical data rather than interval-level measurement data (see *Factor Analysis and Latent Structure: Overview*). For example, Table 1 is a contingency table obtained from the 1995 International Social Survey Program (ISSP) survey on national identity, tabulating responses from 23 countries on the question: 'How much do you agree/disagree with the statement: *Generally* (respondent's country) *is a country better than most other countries*?' (For example, Austrians are asked to evaluate the statement: Generally Austria is a country better than most other countries.)

The object of CA is to obtain a graphical display in the form of a spatial map of the rows (countries) and columns (question responses), where the dimensions of the map as well as the specific positions of the row and column points can be interpreted.

The theory of CA can be summarized by the following steps:

(a) Let $\mathbf{N}$ be the $I \times J$ table with grand total n and let $\mathbf{P} = (1/n)\mathbf{N}$ be the correspondence matrix, with grand total equal to 1. CA actually analyzes the correspondence matrix, which is free of the sample size. If $\mathbf{N}$ is a contingency table, then $\mathbf{P}$ is an observed bivariate discrete distribution.

(b) Let $\mathbf{r}$ and $\mathbf{c}$ be the vectors of row and column sums of $\mathbf{P}$ respectively and $\mathbf{D}_r$ and $\mathbf{D}_c$ diagonal matrices with $\mathbf{r}$ and $\mathbf{c}$ on the diagonal.

(c) Compute the singular value decomposition of the centred and standardized matrix with general element $(p_{ij} - r_i c_j)/\sqrt{r_i c_j}$:

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\top)\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\top \tag{1}$$

where the singular values are in descending order: $\alpha_1 \geqslant \alpha_2 \geqslant \ldots$ and $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$.

(d) Compute the standard coordinates $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U} \quad \mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V} \tag{2}$$

and principal coordinates $\mathbf{F}$ and $\mathbf{G}$:

$$\mathbf{F} = \mathbf{X}\mathbf{D}_\alpha \quad \mathbf{G} = \mathbf{Y}\mathbf{D}_\alpha \tag{3}$$

Notice the following:

The results of CA are in the form of a map of points representing the rows and columns with respect to a

Table 1
Responses in 23 countries to question on national pride. Source: ISSP 1995

Country	Agree strongly	Agree	Can't decide	Disagree	Disagree strongly	Missing	TOTAL RESPONSE
Austria	272	387	184	79	33	52	1007
Bulgaria	208	338	163	112	129	155	1105
Canada	538	620	223	99	33	30	1543
Czech Republic	73	156	372	289	158	63	1111
Former E. Germany	59	138	138	142	71	64	612
Former W. Germany	121	321	369	232	139	100	1282
Great Britain	151	408	282	139	25	53	1058
Hungary	66	175	268	258	143	90	1000
Ireland	167	529	149	120	14	15	994
Italy	70	324	298	281	98	23	1094
Japan	641	398	139	36	24	18	1256
Latvia	92	190	215	225	153	169	1044
Netherlands	158	758	545	417	129	82	2089
Norway	258	723	334	114	30	68	1527
New Zealand	283	499	170	41	8	42	1043
Poland	153	378	396	365	80	226	1598
Philippines	152	556	260	214	10	8	1200
Russia	272	297	352	307	134	223	1585
Slovakia	100	199	384	338	267	100	1388
Slovenia	61	204	258	366	64	83	1036
Spain	71	343	320	372	42	73	1221
Sweden	140	426	375	167	80	108	1296
USA	525	554	168	61	21	38	1367

selected pair of principal axes, corresponding to pairs of columns of the coordinate matrices—usually the first two columns for the first two principal axes. The choice between principal and standard coordinates is described below.

The total variance, called inertia, is equal to the sum of squares of the matrix decomposed in (1):

$$\sum_i \sum_j (p_{ij} - r_i c_j)^2 / (r_i c_j) \tag{4}$$

which is the Pearson chi-squared statistic calculated on the original table divided by n (see *Multivariate Analysis: Discrete Variables (Overview)*).

The squared singular values $\alpha_1^2, \alpha_2^2, \ldots$, called the principal inertias, decompose the inertia into parts attributable to the respective principal axes, just as in PCA the total variance is decomposed along principal axes.

The most popular type of map, called the symmetric map, uses the first two columns of **F** for the row coordinates and the first two columns of **G** for the column coordinates, that is both in principal coordinates as given by (3).

An alternative scaling, which has a more coherent geometric interpretation, but less aesthetic appearance, is the asymmetric map, for example, rows in principal coordinates **F** and columns in standard coordinates **Y** in (2) (or vice versa). The choice between a row-principal or column-principal asymmetric map is governed by whether the original table is considered as a set of rows or a set of columns, respectively, when expressed in percentage form.

The positions of the rows and the columns in a map are projections of points, called profiles, from their true positions in high-dimensional space onto a best-fitting lower-dimensional space. A row or column profile is the corresponding row or column of the table divided by its respective total—in the case of a contingency table the profile is a conditional frequency distribution. Each profile is weighted by a mass equal to the value of the corresponding row or column margin, r_i or c_j. The space of the profiles is structured by a weighted Euclidean distance function called the chi-squared distance and the optimal map is obtained by fitting a lower-dimensional space which fits the profiles by weighted least-squares.

Equivalent forms of (4) which show the use of profile, mass, and chi-squared distance are:

$$\sum_i r_i \sum_j \left(\frac{p_{ij}}{r_i} - c_j\right)^2 / c_j = \sum_j c_j \sum_i \left(\frac{p_{ij}}{c_j} - r_i\right)^2 / r_i \tag{5}$$

Thus the inertia is a weighted average squared distance between the profile vectors (e.g., $\frac{p_{ij}}{r_i}, j = 1, \ldots$ for a row profile, weighted by the mass r_i) and their respective average (e.g., c_j, $j = 1, \ldots$, the average row profile), where the distance is of a weighted Euclidean form (e.g., with inverse weighting of the j-th term by c_j).

An equivalent definition of CA is as a pair of classical scaling problems, one for the rows and one for the columns. For example, a square symmetric matrix of chi-squared distances can be calculated between the row profiles, with each point weighted by its respective row mass. Applying classical scaling (also known as principal coordinate analysis, see *Scaling: Multidimensional*) to this distance matrix, and taking the row masses into account, leads to the row principal coordinates in CA.

The singular value decomposition (SVD) may be written in terms of the standard coordinates in the following equivalent form, for the (i, j)-th element:

$$p_{ij} - r_i c_j = r_i c_j \left(1 + \sum_k \alpha_k x_{ik} y_{jk}\right) \tag{6}$$

which shows that CA can be considered as a bilinear model (see *Multivariate Analysis: Discrete Variables (Correspondence Models)*). For any particular solution, for example in two dimensions where the first two terms of this decomposition are retained, the residual elements have been minimized by weighted least-squares.

3. *Application*

The symmetric map of Table 1, with rows and columns in principal coordinates, is given in Fig. 1. Looking at the positions of the question responses first with respect to the first (horizontal) principal axis, they are seen to lie in their substantive order from 'strongly disagree' on the left to 'strongly agree' on the right, with 'missing' on the disagreement side. The scale values of these categories constitute an optimal scale, by which is meant a standardized interval scale for the categorical variable of agreement—disagreement, including the missing value category, which optimally discriminates between the 23 countries, that is which gives maximum between-country variance.

In the two-dimensional map the response category points form a curve known as the 'horseshoe' or 'arch' which is fairly common for data on an ordinal scale. The second dimension then separates out polarized groups towards the top, or inside the arch, where both extremes of the response scale lie, as well as the missing response in this case.

Turning attention to the countries now, they will line up from left to right in an ordination of agreement induced by the scale values. The five Eastern Bloc countries lie on the left, unfavorable, extreme of the map, with Japan, Canada, USA, and New Zealand at the other, favorable side. The countries generally follow the curved shape as well, but a country such as Bulgaria which lies in a unique position inside the curve is characterized by a relatively high polarization of responses as well as high missing values. Bulgaria's position in the middle of the first axis contrasts with the position of Great Britain, for example, which is

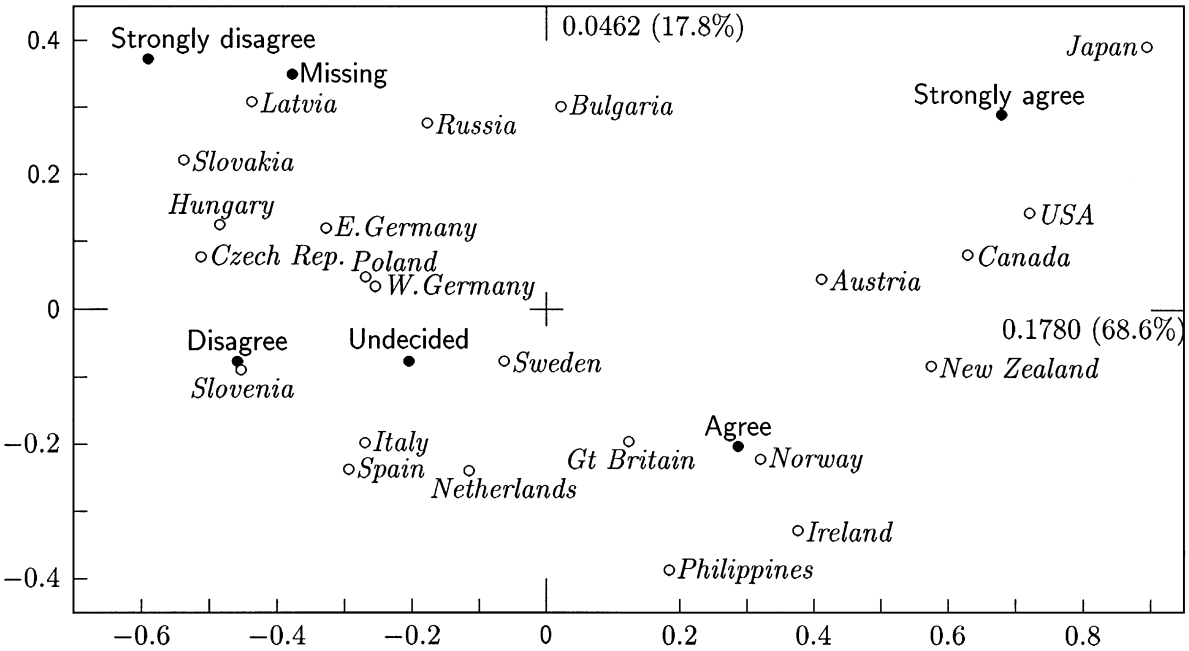

Figure 1
Symmetric correspondence analysis map of Table 1

also in the middle but due to responses more in the intermediate categories of the scale rather than a mixture of extreme responses.

The principal inertias are indicated in Fig. 1 at the positive end of each axis and the quality of the display is measured by adding together the percentages of inertia, 68.6 percent + 17.8 percent = 86.4 percent. This means that there is a 'residual' of 13.6 percent not depicted in the map, which can only be interpreted by investigating the next dimensions from third onwards. This 13.6 percent is the percentage of inertia minimized in the weighted least-squares solution of CA in two dimensions.

4. *Contributions to Inertia*

Apart from assessing the quality of the map by the percentages of inertia, other more detailed diagnostics in CA are the so-called contributions to inertia, based on the two decompositions of the total inertia, first by rows and second by columns:

$$\sum_i \sum_j (p_{ij} - r_i c_j)^2/(r_i c_j) = \sum_k \alpha_k^2 = \sum_k \sum_i r_i f_{ik}^2 = \sum_k \sum_j c_j g_{jk}^2$$

Every row component $r_i f_{ik}^2$, (respectively, column component $c_j g_{jk}^2$) can be expressed relative to the principal inertia α_k^2 of the corresponding dimension k, where α_k^2 is the sum of these components for all the rows (respectively, columns). These relative values provide a diagnostic for deciding which rows (respectively, columns) are important in the determination of the k-th principal axis.

In a similar way, for a fixed row each row component $r_i f_{ik}^2$ (respectively, column component $c_j g_{jk}^2$ for a fixed column) can be expressed relative to the total $\sum_k r_i f_{ik}^2$ (respectively, $\sum_k c_j g_{jk}^2$) across all principal axes. These relative values provide a diagnostic for deciding which axes are important in explaining each row or column. These values are analogous to the squared factor loadings in factor analysis, that is, squared correlations between the row or column and the corresponding principal axis or factor (see *Factor Analysis and Latent Structure: Overview*).

5. *Extensions*

Although the primary application of CA is to a two-way contingency table, the method is regularly applied to analyze multiway tables, tables of preferences, ratings, as well as measurement data on ratio- or interval-level scales. For multiway tables there are two approaches. The first approach is to convert the table to a flat two-way table which is appropriate to the problem at hand. Thus, if a third variable is introduced into the example above, say 'sex of respondent,' then an appropriate way to flatten the three-way table

would be to interactively code 'country' and 'sex' as a new row variable, with $23 \times 2 = 46$ categories, cross-tabulated against the question responses. For each country there would now be a male and a female point and one could compare sexes and countries in this richer map. This process of interactive coding of the variables can continue as long as the data do not become too fragmented into interactive categories of very low frequency.

Another approach to multiway data, called multiple correspondence analysis (MCA), applies when there are several categorical variables skirting the same issue, often called 'items.' MCA is usually defined as the CA algorithm applied to an indicator matrix $\mathbf{Z}$ with the rows being the respondents or other sampling units, and the columns being dummy variables for each of the categories of all the variables. The data are zeros and ones, with the ones indicating the chosen categories for each respondent. The resultant map shows each category as a point and, in principle, the position of each respondent as well. Alternatively, one can set up what is called the Burt matrix), $\mathbf{B} = \mathbf{Z}^{\top}\mathbf{Z}$, the square symmetric table of all two-way crosstabulations of the variables, including the crosstabulations of each variable with itself (named after the psychologist Sir Cyril Burt). The Burt matrix is reminiscent of a covariance matrix and the CA of the Burt matrix can be likened to a PCA of a covariance matrix. The analysis of the indicator matrix $\mathbf{Z}$ and the Burt matrix $\mathbf{B}$ give equivalent standard coordinates of the category points, but slightly different scalings in the principal coordinates since the principal inertias of $\mathbf{B}$ are the squares of those of $\mathbf{Z}$.

A variant of MCA called joint correspondence analysis (JCA) avoids the fitting of the tables on the diagonal of the Burt matrix, which is analogous to least-squares factor analysis.

As far as other types of data are concerned, namely rankings, ratings, paired comparisons, ratio-scale, and interval-scale measurements, the key idea is to recode the data in a form which justifies the basic constructs of CA, namely profile, mass, and chi-squared distance. For example, in the analysis of rankings, or preferences, applying the CA algorithm to the original rankings of a set of objects by a sample of subjects is difficult to justify, because there is no reason why weight should be accorded to an object in proportion to its average ranking. A practice called doubling resolves the issue by adding either an 'anti-object' for each ranked object or an 'anti-subject' for each responding subject, in both cases with rankings in the reverse order. This addition of apparently redundant data leads to CA effectively performing different variants of principal components analysis on the original rankings.

A recent finding by Carroll et al. (1997) is that CA can be applied to a square symmetric matrix of squared distances, transformed by subtracting each squared distance from a constant which is substantially larger than the largest squared distance in the table. This yields a solution which approximates the classical scaling solution of the distance matrix.

All these extensions of CA conform closely to Benzécri's original conception of CA as a universal technique for exploring many different types of data through operations such as doubling or other judicious transformations of the data.

The latest developments on the subject, including discussions of sampling properties of CA solutions and a comprehensive reference list, may be found in the volumes edited by Greenacre and Blasius (1994) and Blasius and Greenacre (1998).

See also: Factor Analysis and Latent Structure: Overview; Multivariate Analysis: Discrete Variables (Correspondence Models); Multivariate Analysis: Discrete Variables (Overview); Scaling: Multidimensional

Bibliography

Benzécri J-P 1973 *L'Analyse des Données* Vol I: *La Classification*, Vol. II: *L'Analyse des Correspondances*. Dunod, Paris

Blasius J, Greenacre M J 1998 *Visualization of Categorical Data*. Academic Press, San Diego, CA

Carroll J D, Kumbasar E, Romney A K 1997 An equivalence relation between correspondence analysis and classical metric multidimensional scaling for the recovery of Euclidean distances. *British Journal of Mathematical and Statistical Psychology* **50**: 81–92

Gifi A 1990 *Nonlinear Multivariate Analysis*. Wiley, Chichester, UK

Greenacre M J 1984 *Theory and Applications of Correspondence Analysis*. Academic Press, London

Greenacre M J 1993 *Correspondence Analysis in Practice*. Academic Press, London

Greenacre M J, Blasius J 1994 *Correspondence Analysis in the Social Sciences*. Academic Press, London

International Social Survey Program (ISSP) 1995 *Survey on National Identity*. Data set ZA 2880, Zentralarchiv für Empirische Sozialforschung. University of Cologne

Lebart L, Morineau A, Warwick K 1984 *Multivariate Descriptive Statistical Analysis*. Wiley, Chichester, UK

Nishisato S 1980 *Analysis of Categorical Data: Dual Scaling and its Applications*. University of Toronto Press, Toronto, Canada

M. Greenacre

Scaling: Multidimensional

The term 'Multidimensional Scaling' or MDS is used in two essentially different ways in statistics (de Leeuw and Heiser 1980a). MDS in the wide sense refers to any technique that produces a multi-dimensional geo-

metric representation of data, where quantitative or qualitative relationships in the data are made to correspond with geometric relationships in the representation. MDS in the narrow sense starts with information about some form of dissimilarity between the elements of a set of objects, and it constructs its geometric representation from this information. Thus the data are 'dissimilarities,' which are distance-like quantities (or similarities, which are inversely related to distances). This entry only concentrates on narrow-sense MDS, because otherwise the definition of the technique is so diluted as to include almost all of multivariate analysis.

MDS is a descriptive technique, in which the notion of statistical inference is almost completely absent. There have been some attempts to introduce statistical models and corresponding estimating and testing methods, but they have been largely unsuccessful. I introduce some quick notation. Dissimilarities are written as δ_{ij}, and distances are $d_{ij}(X)$. Here i and j are the objects of interest. The $n \times p$ matrix X is the configuration, with coordinates of the objects in $\mathscr{R}^p$. Often, data weights w_{ij} are also available, reflecting the importance or precision of dissimilarity δ_{ij}.

1. Sources of Distance Data

Dissimilarity information about a set of objects can arise in many different ways. This article reviews some of the more important ones, organized by scientific discipline.

1.1 Geodesy

The most obvious application, perhaps, is in sciences in which distance is measured directly, although generally with error. This happens, for instance, in triangulation in geodesy, in which measurements are made which are approximately equal to distances, either Euclidean or spherical, depending on the scale of the experiment.

In other examples, measured distances are less directly related to physical distances. For example, one could measure airplane, road, or train travel distances between different cities. Physical distance is usually not the only factor determining these types of dissimilarities.

1.2 Geography/Economics

In economic geography, or spatial economics, there are many examples of input–output tables, where the table indicates some type of interaction between a number of regions or countries. For instance, the data may have n countries, where entry f_{ij} indicates the number of tourists traveling, or the amount of grain exported, from i to j. It is not difficult to think of many other examples of these square (but generally asymmetric) tables. Again, physical distance may be a contributing factor to these dissimilarities, but certainly not the only one.

1.3 Genetics/Systematics

A very early application of a scaling technique was Fisher (1922). He used crossing-over frequencies from a number of loci to construct a (one-dimensional) map of part of the chromosome. Another early application of MDS ideas is in Boyden (1931), where reactions to sera are used to give similarities between common mammals, and these similarities are then mapped into three-dimensional space.

In much of systematic zoology, distances between species or individuals are actually computed from a matrix of measurements on a number of variables describing the individuals. There are many measures of similarity or distance which have been used, not all of them having the usual metric properties. The derived dissimilarity or similarity matrix is analyzed by MDS, or by cluster analysis, because systematic zoologists show an obvious preference for tree representations over continuous representations in $\mathscr{R}^p$.

1.4 Psychology/Phonetics

MDS, as a set of data analysis techniques, clearly originates in psychology. There is a review of the early history, which starts with Carl Stumpf around 1880, in de Leeuw and Heiser (1980a). Developments in psychophysics concentrated on specifying the shape of the function relating dissimilarities and distances, until Shepard (1962) made the radical proposal to let the data determine this shape, requiring this function only to be increasing.

In psychophysics, one of the basic forms in which data are gathered is the 'confusion matrix.' Such a matrix records how many times row-stimulus i was identified as column-stimulus j. A classical example is the Morse code signals studied by Rothkopf (1957). Confusion matrices are not unlike the input–output matrices of economics.

In psychology (and marketing) researchers also collect direct similarity judgments in various forms to map cognitive domains. Ekman's color similarity data is one of the prime examples (Ekman 1963), but many measures of similarity (rankings, ratings, ratio estimates) have been used.

1.5 Psychology/Political Science/Choice Theory

Another source of distance information is 'preference data.' If a number of individuals indicate their prefer-

Table 1
Ten psychology journals

	Journal	Label
A	*American Journal of Psychology*	AJP
B	*Journal of Abnormal and Social Psychology*	JASP
C	*Journal of Applied Psychology*	JAP
D	*Journal of Comparative and Physiological Psychology*	JCPP
E	*Journal of Consulting Psychology*	JCP
F	*Journal of Educational Psychology*	JEP
G	*Journal of Experimental Psychology*	JExP
H	*Psychological Bulletin*	PB
I	*Psychological Review*	PR
J	*Psychometrika*	Pka

ences for a number of objects, then many choice models use geometrical representations in which an individual prefers the object she is closer to. This leads to ordinal information about the distances between the individuals and the objects, e.g., between the politicians and the issues they vote for, or between the customers and the products they buy.

1.6 Biochemistry

Fairly recently, MDS has been applied in the conformation of molecular structures from nuclear resonance data. The pioneering work is Crippen (1977), and a more recent monograph is Crippen and Havel (1988). Recently, this work has become more important because MDS techniques are used to determine protein structure. Numerical analysts and mathematical programmers have been involved, and as a consequence there have been many new and exciting developments in MDS.

2. An Example

Section 1 shows that it will be difficult to find an example that illustrates all aspects of MDS. We select one that can be used in quite a few of the techniques discussed. It is taken from Coombs (1964, p. 464). The data are cross-references between ten psychological journals. The journals are given in Table 1. The actual data are in Table 2. the basic idea, of course, is that journals with many cross-references are similar.

3. Types of MDS

There are two different forms of MDS, depending on how much information is available about the distances. In some of the applications reviewed in Sect. 1 the dissimilarities are known numbers, equal to distances, except perhaps for measurement error. In other cases only the rank order of the dissimilarities is known, or only a subset of them is known.

3.1 Metric Scaling

In metric scaling the dissimilarities between all objects are known numbers, and they are approximated by distances. Thus objects are mapped into a metric space, distances are computed, and compared with the dissimilarities. Then objects are moved in such a way that the fit becomes better, until some loss function is minimized.

In geodesy and molecular genetics this is a reasonable procedure because dissimilarities correspond rather directly with distances. In analyzing input–output tables, however, or confusion matrices, such tables are often clearly asymmetric and not likely to be

Table 2
References in row-journal to column-journal

	A	B	C	D	E	F	G	H	I	J
A	122	4	1	23	4	2	135	17	39	1
B	23	303	9	11	49	4	55	50	48	7
C	0	28	84	2	11	6	15	23	8	13
D	36	10	4	304	0	0	98	21	65	4
E	6	93	11	1	186	6	7	30	10	14
F	6	12	11	1	7	34	24	16	7	14
G	65	15	3	33	3	3	337	40	59	14
H	47	108	16	81	130	14	193	52	31	12
I	22	40	2	29	8	1	97	39	107	13
J	2	0	2	0	0	1	6	14	5	59

directly translatable into distances. Such cases often require a model to correct for asymmetry and scale. The most common class of models (for counts in a square table) is $\mathbf{E}(f_{ij}) = \alpha_i \beta_j \exp\{-\phi(d_{ij}(X))\}$, where ϕ is some monotone transformation through the origin. For ϕ equal to the identity this is known as the choice model for recognition experiments in mathematical psychology (Luce 1963), and as a variation of the quasi-symmetry model in statistics (Haberman 1974). The negative exponential of the distance function was also used by Shepard (1957) in his early theory of recognition experiments.

As noted in Sect. 1.3, in systematic zoology and ecology, the basic data matrix is often a matrix in which n objects are measured on p variables. The first step in the analysis is to convert this into an $n \times n$ matrix of similarities or dissimilarities. Which measure of (dis)similarity is chosen depends on the types of variables in the problem. If they are numerical, Euclidean distances or Mahanalobis distances can be used, but if they are binary other dissimilarity measures come to mind (Gower and Legendre 1986). In any case, the result is a matrix which can be used as input in a metric MDS procedure.

3.2 Nonmetric Scaling

In various situations, in particular in psychology, only the rank order of the dissimilarities is known. This is either because only ordinal information is collected (for instance by using paired or triadic comparisons) or because, while the assumption is natural that the function relating dissimilarities and distances is monotonic, the choice of a specific functional form is not.

There are other cases in which there is incomplete information. For example, observations may only be available on a subset of the distances, either by design or by certain natural restrictions on what is observable. Such cases lead to a distance completion problem, where the configuration is constructed from a subset of the distances, and at the same time the other (missing) distances are estimated. Such distance completion problems (assuming that the observed distances are measured without error) are currently solved with mathematical programming methods (Alfakih et al. 1998).

3.3 Three-way Scaling

In 'three-way scaling' information is available on dissimilarities between n objects on m occasions, or for m subjects. Two easy ways of dealing with the occasions is to perform either a separate MDS for each subject or to perform a single MDS for the average occasion. Three-way MDS constitutes a strategy between these two extremes.

This technique requires computation of m MDS solutions, but they are required to be related to each other. For instance, one can impose the restriction that the configurations are the same, but the transformation relating dissimilarities and distances are different. Or one could require that the projections on the dimensions are linearly related to each other in the sense that $d_{ij}(X_k) = d_{ij}(XW_k)$, where W_k is a diagonal matrix characterizing occasion k. A very readable introduction to three-way scaling is Arabie et al. (1987).

3.4 Unfolding

In 'multidimensional unfolding,' information is only available about off-diagonal dissimilarities, either metric or nonmetric. This means dealing with two different sets of objects, for instance individuals and stimuli or members of congress and political issues, and dissimilarities between members of the first set and members of the second set, and not on the within-set dissimilarities. This typically happens with preference and choice data, in which how individuals like candies, or candidates like issues is known, but not how the individuals like other individuals, and so on.

In many cases, the information in unfolding is also only ordinal. Moreover, it is 'conditional,' which means that while it is known that a politician prefers one issue over another, it is not known if a politician's preference for an issue is stronger than another politician's preference for another issue. Thus the ordinal information is only within rows of the off-diagonal matrix. This makes unfolding data, especially nonmetric unfolding data, extremely sparse.

3.5 Restricted MDS

In many cases it makes sense to impose restrictions on the representation of the objects in MDS. The design of a study may be such that the objects are naturally on a rectangular grid, for instance, or on a circle or ellipse. Often, incorporating such prior information leads to a more readily interpretable and more stable MDS solution.

As noted in Sect. 3.3, some of the more common applications of restricted MDS are to three-way scaling.

4. Existence Theorem

The basic existence theorem in Euclidean MDS, in matrix form, is due to Schoenberg (1935). A more modern version was presented in the book by Torgerson (1958).

I give a simple version here. Suppose E is a nonnegative, hollow, symmetric matrix or order n, and suppose $J_n = I_n - \frac{1}{n} e_n e_n'$ is the 'centering' operator. Here I_n is the identity, and e_n is a vector with all elements equal to one. Then E is a matrix of squared

Euclidean distances between n points in $\mathscr{R}^p$ if and only if $-\frac{1}{2} J_n E J_n$ is positive semi-definite of rank less than or equal to p.

This theorem has been extended to the classical non-Euclidean geometries, for instance by Blumenthal (1953). It can also be used to show that any non-negative, hollow, symmetric E can be embedded 'nonmetrically' in $n-2$ dimensions.

5. Loss Functions

5.1 Least Squares on the Distances

The most straightforward loss function to measure fit between dissimilarities and distances is STRESS, defined by

$$\mathrm{STRESS}(X) \stackrel{\Delta}{=} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(\delta_{ij} - d_{ij}(X))^2. \qquad (1)$$

Obviously this formulation applies to metric scaling only. In the case of nonmetric scaling, the major breakthrough in a proper mathematical formulation of the problem was Kruskal (1964). For this case, STRESS is defined as,

$$\mathrm{STRESS}(X, \hat{D}) \stackrel{\Delta}{=} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(\hat{d}_{ij} - d_{ij}(X))^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(d_{ij}(X) - \bar{d}_{ij}(X))^2} \qquad (2)$$

and this function is minimized over both X and $\hat{D}$, where $\hat{D}$ satisfies the constraints imposed by the data. In nonmetric MDS the $\hat{D}$ are called disparities, and are required to be monotonic with the dissimilarities. Finding the optimal $\hat{D}$ is an 'isotonic regression problem.' In the case of distance completion problems (with or without measurement error), the $\hat{d}_{ij}$ must be equal to the observed distances if these are observed, and they are otherwise free.

One particular property of the STRESS loss function is that it is not differentiable for configurations in which two points coincide (and a distance is zero). It is shown by de Leeuw (1984) that at a local minimum of STRESS, pairs of points with positive dissimilarities cannot coincide.

5.2 Least Squares on the Squared Distances

A second loss function, which has been used a great deal, is SSTRESS, defined by

$$\mathrm{SSTRESS}(X) \stackrel{\Delta}{=} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(\delta_{ij}^2 - d_{ij}^2(X))^2. \qquad (3)$$

Clearly, this loss function is a (fourth-order) multivariate polynomial in the coordinates. There are no problems with smoothness, but often a large number of local optima results.

Of course a nonmetric version of the SSTRESS problem can be confronted, using the same type of approach used for STRESS.

5.3 Least Squares on the Inner Products

The existence theorem discussed above suggests a third way to measure loss. Now the function is known as STRAIN, and it is defined, in matrix notation, as

$$\mathrm{STRAIN}(X) \stackrel{\Delta}{=} \mathbf{tr}\{J(\Delta^{(2)} - D^{(2)}(X))J(\Delta^{(2)} - D^{(2)}(X))\} \qquad (4)$$

where $D^{(2)}(X)$ and $\Delta^{(2)}$ are the matrices of squared distances and dissimilarities, and where J is the centering operator. Since $JD^{(2)}(X)J = -2XX'$ this means that $-\frac{1}{2} J\Delta^{(2)}J$ is approximated by a positive semi-definite matrix of rank r, which is a standard eigenvalue–eigenvector computation.

Again, nonmetric versions of minimizing STRAIN are straightforward to formulate (although less straightforward to implement).

Table 3
Transformed journal reference data

0.00	2.93	4.77	1.89	3.33	2.78	0.77	1.02	1.35	3.79
2.93	0.00	2.28	3.32	1.25	2.61	2.39	0.53	1.41	4.24
4.77	2.28	0.00	3.87	2.39	1.83	3.13	1.22	3.03	2.50
1.89	3.32	3.87	0.00	5.62	4.77	1.72	1.11	1.41	4.50
3.33	1.25	2.39	5.62	0.00	2.44	3.89	0.45	2.71	3.67
2.78	2.61	1.83	4.77	2.44	0.00	2.46	1.01	2.90	2.27
0.77	2.39	3.13	1.72	3.89	2.46	0.00	0.41	0.92	2.68
1.02	0.53	1.22	1.11	0.45	1.01	0.41	0.00	0.76	1.42
1.35	1.41	3.03	1.41	2.71	2.90	0.92	0.76	0.00	2.23
3.79	4.24	2.50	4.50	3.67	2.27	2.68	1.42	2.23	0.00

6. Algorithms

6.1 Stress

The original algorithms (Kruskal 1964) for minimizing STRESS use gradient methods with elaborate step-size procedure. In de Leeuw (1977) the 'majorization method' was introduced. It leads to a globally convergent algorithm with a linear convergence rate, which is not bothered by the nonexistence of derivatives at places where points coincide. The majorization method can be seen as a gradient method with a constant step-size, which uses convex analysis methods to prove convergence.

More recently, faster linearly or superlinearly convergent methods have been tried successfully (Glunt et al. 1993, Kearsley et al. 1998).

One of the key advantages of the majorization method is that it extends easily to restricted MDS problems (de Leeuw and Heiser 1980b). Each subproblem in the sequence is a least squares projection problem on the set of configurations satisfying the constraints, which is usually easy to solve.

6.2 SSTRESS

Algorithms for minimizing SSTRESS were developed initially by Takane et al. (1977). They applied cyclic coordinate descent, i.e., one coordinate was changed at the time, and cycles through the coordinates were alternated with isotonic regressions in the nonmetric case. More efficient alternating least squares algorithms were developed later by de Leeuw, Takane, and Browne (cf. Browne (1987)), and superlinear and quadratic methods were proposed by Glunt and Liu (1991) and Kearsley et al. (1998).

6.3 STRAIN

Minimizing STRAIN was, and is, the preferred algorithm in metric MDS. It is also used as the starting

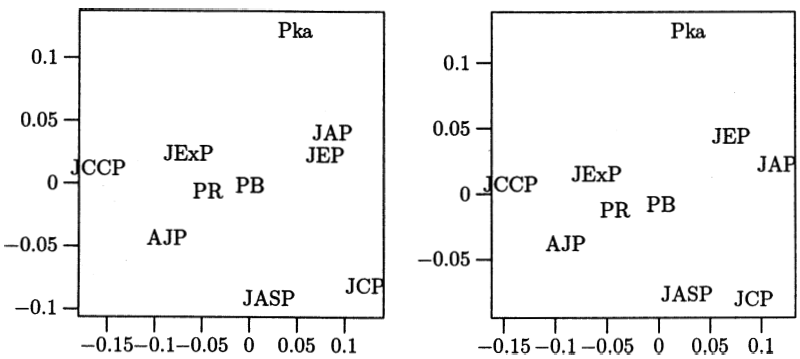

Figure 1
Metric analysis (STRAIN left, STRESS right)

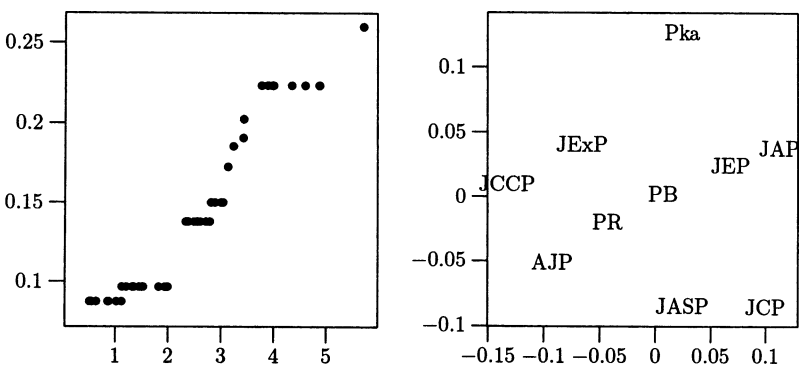

Figure 2
Nonmetric analysis (transformation left, solution right)

point in iterative nonmetric algorithms. Recently, more general algorithms for minimizing STRAIN in nonmetric and distance completion scaling have been proposed by Trosset (1998a, 1998b).

7. Analysis of the Example

7.1 Initial Transformation

In the journal reference example, suppose $\mathbf{E}(f_{ij}) = \alpha_i\beta_j \exp\{-\phi(d_{ij}(X))\}$. In principle this model can be tested by contingency table techniques. Instead the model is used to transform the frequencies to estimated distances, yielding

$$-\log\sqrt{\frac{\tilde{f}_{ij}\tilde{f}_{ji}}{\tilde{f}_{ii}\tilde{f}_{jj}}} \approx \phi(d_{ij}(X))$$

where $\tilde{f}_{ij} = f_{ij} + \frac{1}{2}$. This transformed matrix is given in Table 3.

7.2 Metric Analysis

In the first analysis, suppose the numbers in Table 3 are approximate distances, i.e., suppose that ϕ is the identity. Then STRAIN is minimized, using metric MDS by calculating the dominant eigenvalues and corresponding eigenvectors of the doubly-centered squared distance matrix. This results in the following two-dimensional configurations. The second analysis iteratively minimizes metric STRESS, using the majorization algorithm. The solutions are given in Fig. 1. Both figures show the same grouping of journals, with Pka as an outlier, the journals central to the discipline, such as AJP, JExP, PB, and PR, in the middle, and more specialized journals generally in the periphery. For comparison purposes the STRESS of the first solution is 0.0687, that of the second solution is 0.0539. Finding the second solution takes about 30 iterations.

7.3 Nonmetric STRESS Analysis

Next, nonmetric STRESS is minimized on the same data (using only their rank order). The solution is in Fig. 2. The left panel displays the transformation relating the data in Table 3 to the optimally transformed data, a monotone step function. Again, basically the same configuration of journals, with the same groupings emerges. The nonmetric solution has a (normalized) STRESS of 0.0195, and again finding it takes about 30 iterations of the majorization method. The optimal transformation does not seem to deviate systematically from linearity.

8. Further Reading

Until recently, the classical MDS reference was the little book by Kruskal and Wish (1978). It is clearly written, but very elementary. A more elaborate practical introduction is by Coxon (1982), which has a useful companion volume (Davies and Coxon 1982) with many of the classical MDS papers. Some additional early intermediate-level books, written from the psychometric point of view, are Davison (1983) and Young (1987).

More recently, more modern and advanced books have appeared. The most complete treatment is no doubt Borg and Groenen (1997), while Cox (1994) is another good introduction especially aimed at statisticians.

Bibliography

Alfakih A Y, Khandani A, Wolkowicz H 1998 Solving Euclidean distance matrix completion problems via semidefinite programming. *Computational Optimization and Applications* **12**: 13–30

Arabie P, Carroll J D, DeSarbo W S 1987 *Three-Way Scaling and Clustering*. Sage, Newbury Park

Blumenthal L M 1953 *Distance Geometry*. Oxford University Press, Oxford, UK

Borg I, Groenen P 1997 *Modern Multidimensional Scaling*. Springer, New York

Boyden A 1931 Precipitin tests as a basis for a comparitive phylogeny. *Proceedings of the Society for Experimental Biology and Medicine* **29**: 955–7

Browne M 1987 The Young–Householder algorithm and the least squares multidimensional scaling of squared distances. *Journal of Classification* **4**: 175–90

Coombs C H 1964 *A Theory of Data*. Wiley, New York

Cox T F 1994 *Multidimensional Scaling*. Chapman and Hall, New York

Coxon A P M 1982 *The User's Guide to Multidimensional Scaling: With Special Reference to the MDS(X) Library of Computer Programs*. Heinemann, Exeter, NH

Crippen G M 1977 A novel approach to calculation of conformation: Distance geometry. *Journal of Computational Physics* **24**: 96–107

Crippen G M, Havel T F 1988 *Distance Geometry and Molecular Conformation*. Wiley, New York

Davies P M, Coxon A P M 1982 *Key Texts in Multidimensional Scaling*. Heinemann, Exeter, NH

Davison M L 1983 *Multidimensional Scaling*. Wiley, New York

de Leeuw J 1977 Applications of convex analysis to multidimensional scaling. In: Barra J R, Brodeau F, Romier G, van Cutsem B (eds.) *Recent Developments in Statistics: Proceedings of the European Meeting of Statisticians, Grenoble, 6–11 September, 1976*. North Holland, Amsterdam, The Netherlands, pp. 133–45

de Leeuw J 1984 Differentiability of Kruskal's stress at a local minimum. *Psychometrika* **49**: 111–3

de Leeuw J, Heiser W 1980a Theory of multidimensional scaling. In: Krishnaiah P (ed.) *Handbook of Statistics*. North Holland, Amsterdam, The Netherlands, Vol. II

de Leeuw J, Heiser W J 1980b Multidimensional scaling with restrictions on the configuration. In: Krishnaiah P (ed.)

Multivariate Analysis. North Holland, Amsterdam, The Netherlands, Vol. v, pp. 501–22
Ekman G 1963 Direct method for multidimensional ratio scaling. *Psychometrika* **23**: 33–41
Fisher R A 1922 The systematic location of genes by means of cross-over ratios. *American Naturalist* **56**: 406–11
Glunt W, Hayden T L, Liu W M 1991 The embedding problem for predistance matrices. *Bulletin of Mathematical Biology* **53**: 769–96
Glunt W, Hayden T, Rayden M 1993 Molecular conformations from distance matrices. *Journal of Computational Chemistry* **14**: 114–20
Gower J C, Legendre P 1986 Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* **3**: 5–48
Haberman S J 1974 *The Analysis of Frequency Data*. University of Chicago Press, Chicago
Kearsley A J, Tapia R A, Trosset M W 1998 The solution of the metric STRESS and SSTRESS problems in multidimensional scaling using Newton's method. *Computational Statistics* **13**: 369–96
Kruskal J B 1964 Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**: 1–27
Kruskal J B, Wish M 1978 *Multidimensional Scaling*. Sage, Beverly Hills, CA
Luce R 1963 Detection and recognition. In: Luce R D, Bush R R, Galanter E (eds.) *Handbook of Mathematical Psychology*. Wiley, New York, Vol. 1
Rothkopf E Z 1957 A measure of stimulusa similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology* **53**: 94–101
Schoenberg I J 1935 Remarks on Maurice Fréchet's article Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert. *Annals of Mathematics* 724–32
Shepard R N 1957 Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika* **22**: 325–45
Shepard R N 1962 The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika* **27**: 125–40, 219–46
Takane Y, Young F W, de Leeuw J 1977 Nonmetric individual differences in multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika* **42**: 7–67
Torgerson W S 1958 *Theory and Methods of Scaling*. Wiley, New York
Trosset M W 1998a Applications of multidimensional scaling to molecular conformation. *Computing Science and Statistics* **29**: 148–52
Trosset M W 1998b A new formulation of the nonmetric strain problem in multidimensional scaling. *Journal of Classification* **15**: 15–35
Young F W 1987 *Multidimensional Scaling: History, Theory, and Applications*. Earlbaum, Hillsdale, NJ

J. de Leeuw

Scandal: Political

The word 'scandal' is used primarily to describe a sequence of actions and events which involve certain kinds of transgressions and which, when they become known to others, are regarded as sufficiently serious to elicit a response of disapproval or condemnation. A scandal is necessarily a public event in the sense that, while the actions which lie at the heart of the scandal may have been carried out secretly or covertly, a scandal can arise only if these actions become known to others, or are strongly believed by others to have occurred. This is one respect in which scandal differs from related phenomena such as corruption and bribery; a scandal can be based on the disclosure of corruption or bribery, but corruption and bribery can exist (and often do exist) without being known about by others, and hence without becoming a scandal.

1. *The Concept of Scandal*

The concept of scandal is very old and the meaning has changed over time. In terms of its etymological origins, the word probably derives from the Indogermanic root *skand-*, meaning to spring or leap. Early Greek derivatives, such as the word *skandalon*, were used in a figurative way to signify a trap, an obstacle or a 'cause of moral stumbling.' The idea of a trap or an obstacle was an integral feature of the theological vision of the Old Testament. In the Septuagint (the Greek version of the Old Testament), the word *skandalon* was used to describe an obstacle, a stumbling block placed along the path of the believer, which could explain how a people linked to God might nevertheless begin to doubt Him and lose their way. The notion of a trap or obstacle became part of Judaism and early Christian thought, although it was gradually prised apart from the idea of a test of faith.

With the development of the Latin word *scandalum* and its diffusion into Romance languages, the religious connotation was gradually attenuated and supplemented by other senses. The word 'scandal' first appeared in English in the sixteenth century; similar words appeared in other Romance languages at roughly the same time. The early uses of 'scandal' in the sixteenth and seventeenth centuries were, broadly speaking, of two main types. First, 'scandal' was used in a religious context to refer to the conduct of a person which brought discredit to religion, or to something which hindered religious faith or belief. Second, 'scandal' and its cognates were also used in more secular contexts to describe actions or utterances which were regarded as scurrilous or abusive, which damaged an individual's reputation, which were grossly discreditable, and/or which offended moral sentiments or the sense of decency.

It is these later, more secular senses which underlie the most common modern uses of the word 'scandal.' Although the word continues to have some use as a specialized religious term, 'scandal' is used mainly to refer to a broader form of moral transgression which is no longer linked specifically to religious codes. More precisely, 'scandal' could be defined as actions or events which have the following characteristics: their

occurrence involves the transgression of certain values, norms or moral codes; their occurrence involves a degree of secrecy or concealment, but they are known or strongly believed to exist by individuals other than those directly involved; some individuals disapprove of the actions or events and may be offended by the transgression; some express their disapproval by publicly denouncing the actions or events; and the disclosure and condemnation of the actions or events may damage the reputation of the individuals responsible for them.

While scandal necessarily involves some form of transgression, there is a great deal of cultural and historical variability in the kinds of values, norms, and moral codes which are relevant here. What counts as scandalous activity in one context, e.g., extramarital affairs among members of the political elite, may be regarded as acceptable (even normal) elsewhere. A particular scandal may also involve different types of transgression. A scandal may initially be based on the transgression of a moral code (e.g., concerning sexual relations), but as the scandal develops, the focus of attention may shift to a series of 'second-order transgressions' which stem from actions aimed at concealing the original offence. The attempt to cover up a transgression—a process that may involve deception, obstruction, false denials, and straightforward lies—may become more important than the original transgression itself, giving rise to an intensifying cycle of claim and counterclaim that dwarfs the original offence.

Scandals can occur in different settings and milieux, from local communities to the arenas of national and even international politics. When scandals occur in settings which are more extended than local communities, they generally involve mediated forms of communication such as newspapers, magazines, and, increasingly, television. The media play a crucial role in making public the actions or events which lie at the heart of scandal, either by reporting allegations or information obtained by others (such as the police or the courts) or by carrying out their own investigations. The media also become a principal forum in which disapproval of the actions or events is expressed. Mediated scandals are not simply scandals which are reported by the media; rather, they are 'mediated events' which are partly constituted by the activities of media organizations.

2. *The Nature of Political Scandal*

Scandals are common in many spheres of social life; not all scandals are political scandals. So what are the distinctive features of political scandals? One seemingly straightforward way of answering this question is to say that a political scandal is any scandal that involves a political leader or figure. But this is not a particularly helpful or illuminating answer. For an individual is a political leader or figure by virtue of a broader set of social relations and institutions which endow him or her with power. So if we wish to understand the nature of political scandal, we cannot focus on the individual alone.

Another way of answering this question would be to focus not on the status of the individuals involved, but on the nature of the transgression. This is the approach taken by Markovits and Silverstein, two political scientists who have written perceptively about political scandal. According to Markovits and Silverstein (1988), the defining feature of political scandal is that it involves a 'violation of due process.' By 'due process' they mean the legally binding rules and procedures which govern the exercise of political power. Political scandals are scandals in which these rules and procedures are violated by those who exercise political power and who seek to increase their power at the expense of due process. Since due process is fully institutionalized only in the liberal democratic state, it follows, argue Markovits and Silverstein, that political scandals can occur only in liberal democracies.

One strength of Markovits and Silverstein's account is that it analyzes political scandal in relation to some of the most important institutional features of modern states. But the main shortcoming of this account is that it provides a rather narrow view of political scandal. It treats one dynamic—the pursuit of power at the expense of process—as the defining feature of political scandal. Hence any scandal that does not involve this particular dynamic is *ipso facto* nonpolitical. This means that a whole range of scandals, such as those based on sexual transgressions, would be ruled out as nonpolitical, even though they may involve senior political figures and may have far-reaching political consequences.

Markovits and Silverstein's claim that political scandal can occur only in liberal democracies should also be viewed with some caution. It is undoubtedly the case that liberal democracies are particularly prone to political scandal, but this is due to a number of specific factors (such as the highly competitive nature of liberal democratic politics and the relative autonomy of the press) and it does not imply that political scandal is unique to this type of political organization. Political scandals can occur (and have occurred) in other types of political system, from the absolutist and constitutional monarchies of early modern Europe to the various forms of authoritarian regime which have existed in the twentieth century. But political scandals in these other types of political system are more likely to remain localized scandals and are less likely to spread beyond the relatively closed worlds of the political elite.

An alternative way of conceptualizing political scandals is to regard them as scandals involving individuals or actions which are situated within a political field (Thompson 2000). It is the political field that constitutes the scandal as political; it provides the

context for the scandal and shapes its pattern of development. A field is a structured space of social positions whose properties are defined primarily by the relations between these positions and the resources attached to them. The political field can be defined as the field of action and interaction which bears on the acquisition and exercise of political power. Political scandals are scandals which occur within the political field and which have an impact on relations within this field. They may involve the violation of rules and procedures governing the exercise of political power, but they do not have to involve this; other kinds of transgression can also constitute political scandals.

We can distinguish between three main types of political scandal, depending on the kinds of norms or codes which are transgressed. Sex scandals involve the transgression of norms or codes governing the conduct of sexual relations. In some contexts, sexual transgressions carry a significant social stigma and their disclosure may elicit varying degrees of disapproval by others. Financial scandals involve the infringement of rules governing the acquisition and allocation of economic resources; these include scandals involving bribery, kickbacks and other forms of corruption as well as scandals stemming from irregularities in the raising and deployment of campaign funds. Power scandals are based on the disclosure of activities which infringe the rules governing the acquisition or exercise of political power. They involve the unveiling of hidden forms of power, and actual or alleged abuses of power, which had hitherto been concealed beneath the public settings in which power is displayed and the publicly recognized procedures through which it is exercised.

3. *The Rise of Political Scandal*

The origins of political scandal as a mediated event can be traced back to the pamphlet culture of the seventeenth and eighteenth centuries. During the period of the English Civil War, for instance, there was a proliferation of anti-Royalist pamphlets and newsbooks which were condemned as heretical, blasphemous, scurrilous, and 'scandalous' in character. Similarly, in France, a distinctive genre of subversive political literature had emerged by the early eighteenth century, comprising the *libelles* and the *chroniques scandaleuses*, which purported to recount the private lives of kings and courtiers and presented them in an unflattering light. However, in the late eighteenth and nineteenth centuries, the use of 'scandal' in relation to mediated forms of communication began to change, as the term was gradually prised apart from its close association with blasphemy and sedition and increasingly applied to a range of phenomena which displayed the characteristics we now associate with scandal.

By the late nineteenth century, mediated scandal had become a relatively common feature of the political landscape in countries such as Britain and the USA. In Britain there were a number of major scandals, many involving sexual transgressions of various kinds, which destroyed (or threatened to destroy) the careers of key political figures such as Sir Charles Dilke (a rising star of the Liberal Party whose career was irrevocably damaged by the events surrounding a divorce action in which he was named as co-respondent) and Charles Parnell (the charismatic leader of the Irish parliamentary party whose career was destroyed by revelations concerning his affair with Mrs. Katharine O'Shea). There were numerous political scandals in nineteenth-century America too, some involving actual or alleged sexual transgressions (such as the scandal surrounding Grover Cleveland who, it was said, had fathered an illegitimate child) and many involving corruption at municipal, state and federal levels of government. America in the Gilded Age witnessed a flourishing of financial scandals in the political field, and the period of Grant's Presidency (1869–77) is regarded by many as one of the most corrupt in American history.

While the nineteenth century was the birthplace of political scandal as a mediated event, the twentieth century was to become its true home. Once this distinctive type of event had been invented, it would become a recognizable genre that some would seek actively to produce while others would strive, with varying degrees of success, to avoid. The character and frequency of political scandals varied considerably from one national context to another, and depended on a range of specific social and political circumstances. In Britain and the USA, there were significant political scandals throughout the early decades of the twentieth century, but political scandals have become particularly prevalent in the period since the early 1960s. In Britain, the Profumo scandal of 1963 was a watershed. This was a classic sex scandal involving a senior government minister (John Profumo, Secretary of State for War) and an attractive young woman (Christine Keeler), but it also involved issues of national security and a series of second-order transgressions which proved fatal for Profumo's career. In the USA, the decisive political scandal of the twentieth century was undoubtedly Watergate, a power scandal in which Nixon was eventually forced to resign as President in the face of his imminent impeachment.

Many countries have developed their own distinctive political cultures of scandal which have been shaped by, among other things, their respective traditions of scandal, the activities of journalists, media organizations, and other agents in the political field, the deployment of new technologies of communication, and the changing political climate of the time. Political scandal has become a potent weapon in struggles between rival candidates and parties in the political field. As fundamental disagreements over matters of principle have become less pronounced, questions of character and trust have become incr-

easingly central to political debate and scandal has assumed increasing significance as a 'credibility test.' In this context, the occurrence of scandal tends to have a cumulative effect: scandal breeds scandal, because each scandal exposes character failings and further sharpens the focus on the credibility and trustworthiness of political leaders.

This is the context in which President Clinton found that his political career was nearly destroyed by scandal on more than one occasion. Like many presidential hopefuls in the past, Clinton campaigned on the promise to clean up politics after the sleaze of the Reagan administration. But he soon found that members of his own administration—and, indeed, that he and his wife—were being investigated on grounds of possible financial wrongdoing. He also found that allegations and revelations concerning his private life would become highly public issues, threatening to derail his campaign is 1992 (with the Gennifer Flowers affair) and culminating in his impeachment and trial by the Senate following the disclosure of his affair with Monica Lewinsky. What led to Clinton's impeachment was not the disclosure of the affair as such, but rather a series of second-order transgressions committed in relation to a sexual harassment case instituted by Paula Jones, in the context of which Clinton gave testimony under oath denying that he had had sexual relations with Monica Lewinsky, thereby laying himself open to the charge of perjury, among other things. Clinton's trial in the Senate resulted in his acquittal, but his reputation was undoubtedly damaged by the scandal which overshadowed the second term of his presidency.

See also: Elites: Sociological Aspects; Mass Media: Introduction and Schools of Thought; Political Culture; Political Discourse; Political Trials; Public Opinion: Political Aspects

Bibliography

Allen L et al. 1990 *Political Scandals and Causes Celèbres Since 1945: An International Reference Compendium* n.d. Longman, Harlow, Essex, UK, Chicago, IL, USA: Published in the USA and Canada by St. James Press

Garment S 1992 *Scandal: The Culture of Mistrust in American Politics*. Doubleday, New York

King A 1986 Sex, money, and power. In: Hodder-Williams R, Ceaser J (eds.) *Politics in Britain and the United States: Comparative Perspectives*. Duke University Press, Durham, NC pp. 173–222

Markovits A S, Silverstein M (eds.) 1988 *The Politics of Scandal: Power and Process in Liberal Democracies*. Holmes and Meier, New York

Allen L et al. 1990 *Political Scandals and Causes Celèbres Since 1945: An International Reference Compendium* n.d. Longman, Harlow, Essex, UK, Chicago, IL, USA: Published in the USA and Canada by St. James Press

Schudson M 1992 *Watergate in American Memory: How We Remember, Forget, and Reconstruct the Past*. Basic Books, New York

Thompson J B 2000 *Political Scandal: Power and Visibility in the Media Age*. Polity, Cambridge, UK

J. B. Thompson

Schemas, Frames, and Scripts in Cognitive Psychology

The terms 'schema,' 'frame,' and 'script' are all used to refer to a generic mental representation of a concept, event, or activity. For example, people have a generic representation of a visit to the dentist office that includes a waiting room, an examining room, dental equipment, and a standard sequence of events in one's experience at the dentist. Differences in how each of the three terms, schema, frame, and script, are used reflect the various disciplines that have contributed to an understanding of how such generic representations are used to facilitate ongoing information processing. This article will provide a brief family history of the intellectual ancestry of modern views of generic knowledge structures as well as an explanation of how such knowledge influences mental processes.

1. Early Development of the 'Schema' Concept

The term 'schema' was already in prominent use in the writings of psychologists and neurologists by the early part of the twentieth century (e.g., Bartlett 1932, Head 1920, see also *Piaget's Theory of Child Development*). Though the idea that our behavior is guided by schemata (the traditional plural form of schema) was fairly widespread, the use of the term 'schema' was seldom well defined. Nevertheless, the various uses of the term were reminiscent of its use by the eighteenth century philosopher Immanuel Kant. In his influential treatise *Critique of Pure Reason*, first published in 1781, Kant wrestled with the question of whether all knowledge is dependent on sensory experience or whether there exists *a priori* knowledge, which presumably would consist of elementary concepts, such as space and time, necessary for making sense of the world. Kant believed that we depend both on our knowledge gained from sensory experience and *a priori* knowledge, and that we use a generic form of knowledge, schemata, to mediate between sensory-based and *a priori* knowledge.

Unlike today, when the idea of a schema is most often invoked for large-scale structured knowledge, Kant used the term in reference to more basic concepts. In *Critique of Pure Reason* he wrote: 'In truth, it is not images of objects, but schemata, which lie at the foundation of our pure sensuous conceptions. No

image could ever be adequate to our conception of triangles in general…The schema of a triangle can exist nowhere else than in thought and it indicates a rule of the synthesis of the imagination in regard to pure figures in space.' The key idea captured by Kant was that people need mental representations that are typical of a class of objects or events so that we can respond to the core similarities across different stimuli of the same class (see *Natural Concepts*, *Psychology of*; *Mental Representations*, *Psychology of*).

The basic notion that our experience of the world is filtered through generic representations was also at the heart of Sir Frederic Bartlett's use of the term 'schema.' Bartlett's studies of memory, described in his classic book published in 1932, presaged the rise of schema-theoretic views of memory and text comprehension in the 1970s and 1980s. In his most influential series of studies, Bartlett asked people to read and retell folktales from unfamiliar cultures. Because they lacked the background knowledge that rendered the folktales coherent from the perspective of the culture of origin, the participants in Bartlett's studies tended to conventionalize the stories by adding connectives and explanatory concepts that fit the reader's own world view. The changes were often quite radical, particularly when the story was recalled after a substantial delay. For example, in some cases the reader gave the story a rather conventional moral consistent with more familiar folktales.

Bartlett interpreted these data as indicating that: 'Remembering is not the re-excitation of innumerable fixed, lifeless, and fragmentary traces' (p. 213). Instead, he argued, memory for a particular stimulus is reconstructed based on both the stimulus itself and on a relevant schema. By schema, Bartlett meant a 'whole active mass of organised past reactions or experience.' Thus, if a person's general knowledge of folktales includes the idea that such tales end with an explicit moral, then in reconstructing an unfamiliar folktale the person is likely to elaborate the original stimulus by adding a moral that appears relevant (see *Reconstructive Memory*, *Psychology of*).

Although Bartlett made a promising start on understanding memory for complex naturalistic stimuli, his work was largely ignored for the next 40 years, particularly in the United States where researchers focused on much more simplified learning paradigms from a behavioral perspective. Then, interest in schemata returned as researchers in the newly developed fields of cognitive psychology and artificial intelligence attempted to study and to simulate discourse comprehension and memory (see *Knowledge Activation in Text Comprehension and Problem Solving*, *Psychology of*).

2. *The Rediscovery of Schemata*

Interest in schemata as explanatory constructs returned to psychology as a number of memory researchers demonstrated that background knowledge can strongly influence the recall and recognition of text. For example, consider the following passage: 'The procedure is actually quite simple. First, you arrange items into different groups. Of course, one pile may be sufficient, depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step; otherwise, you are pretty well set.' This excerpt, from a passage used by Bransford and Johnson (1972), is very difficult to understand and recall unless the reader has an organizing schema for the information. So, readers who are supplied with the appropriate title, *Washing Clothes*, are able to recall the passage much more effectively than readers who are not given the title.

Although the early 1970s witnessed a great number of clear demonstrations that previous knowledge shapes comprehension and memory, progress on specifying exactly how background knowledge becomes integrated with new information continued to be hampered by the vagueness of the schema concept. Fortunately, at this same time, researchers in artificial intelligence (AI) identified the issue of background knowledge as central to simulations of text processing. Owing to the notorious inability of computers to deal with vague notions, the AI researchers had to be explicit in their accounts of how background knowledge is employed in comprehension. A remarkable cross-fertilization occurred between psychology and AI that encouraged the development and testing of hypotheses about the representation and use of schematic knowledge.

2.1 *Schema Theory in AI Research: Frames and Scripts*

Although many researchers in the nascent state of AI research in 1970s were unconcerned with potential connections to psychology, several pioneers in AI research noted the mutual relevance of the two fields early on. In particular, Marvin Minsky and Roger Schank developed formalizations of schema-theoretic approaches that were both affected by, and greatly affected, research in cognitive psychology.

One of the core concepts from AI that was quickly adopted by cognitive psychologists was the notion of a frame. Minsky (1975) proposed that our experiences of familiar events and situations are represented in a generic way as frames that contain slots. Each slot corresponds to some expected element of the situation. One of the ways that Minsky illustrated what he meant by a frame was in reference to a job application form. A frame, like an application, has certain general domains of information being sought and leaves blanks for specific information to be filled in. As another example, someone who walks into the office of a university professor could activate a frame for a generic professor's office that would help orient the person to the situation and allow for easy identification

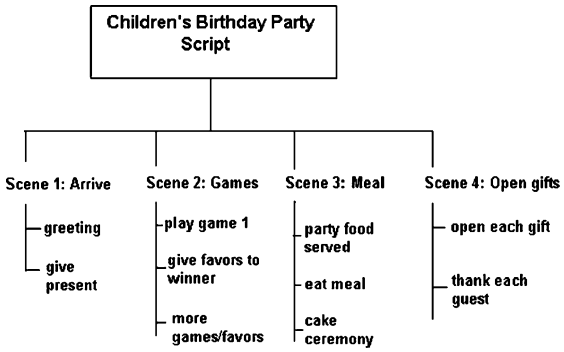

Figure 1
A child's birthday party script. The script consists of a set of distinct scenes. The slots under each scene can be filled with default information, such as 'hot dogs' for party food. Note that the cake ceremony could be broken down further into scenes such as candle lighting, blowing out candles, etc. Thus, one script or schema can be embedded inside another

of its elements. So, a person would expect to see a desk and chair, a bookshelf with lots of books, and, these days at least, a computer. An important part of Minsky's frame idea is that the slots of a frame are filled with default values if no specific information from the context is provided (or noticed). Thus, even if the person didn't actually see a computer in the professor's office, he or she might later believe one was there because *computer* would be a default value that filled an *equipment* slot in the *office* frame (see Brewer and Treyens 1981 for relevant empirical evidence).

The idea that generic knowledge structures contain slots with default values was also a central part of Roger Schank's work on scripts. Scripts are generic representations of common social situations such as going to the doctor, having dinner at a restaurant, or attending a party (Fig. 1). Teamed with a social psychologist, Robert Abelson, Schank developed computer simulations of understanding stories involving such social situations.

The motivation for the development of the script concept was Schank's observation that people must be able to make a large number of inferences quite automatically in order to comprehend discourse. This can be true even when the discourse consists of a relatively simple pair of sentences:

Q: So, would you like to get some dinner?

A: Oh, I just had some pizza a bit ago.

Even though the question is not explicitly answered, a reader of these two sentences readily understands that the dinner invitation was declined. Note that the interpretation of this exchange is more elaborate if the reader knows that it occurred between a young man and a young woman who met at some social event and had a lengthy conversation. In this context, a script for asking someone on a date may be invoked and the answer is interpreted as a clear 'brush off.'

Thus, a common feature of the AI work on frames and scripts in simulations of natural language understanding was that inferences play a key role. Furthermore, these researchers provided formal descriptions of generic knowledge structures that could be used to generate the inferences necessary to comprehension. Both AI researchers and cognitive psychologists of the time recognized that such knowledge structures captured the ideas implicit in Bartlett's more vague notion of a schema. The stage was set to determine whether frames and scripts were useful descriptors for the knowledge that people actually use when processing information (see *Inferences in Discourse, Psychology of*; *Memory for Text*).

2.2 Empirical Support for Schema Theory

If people's comprehension and memory processes are based on retrieving generic knowledge structures and, as necessary, filling slots of those knowledge structures with default values, then several predictions follow. First, people with similar experiences of common situations and events will have similar schemata represented in memory. Second, as people retrieve such schemata during processing, they will make inferences that go beyond directly stated information. Third, in many cases, particularly when processing is guided by a script-like structure, people will have expectations about not only the content of a situation but also the sequence of events.

Cognitive psychologists obtained empirical support for each of these predictions. For example, Bower et al. (1979) conducted a series of experiments on people's generation of, and use of, scripts for common events like going to a nice restaurant, attending a lecture, or visiting the doctor. They found high agreement on the actions that take place in each situation and on the ordering. Moreover, there was good agreement on the major clusters of actions, or scenes, within each script. Consistent with the predictions of schema theory, when people read stories based on a script, their recall of the story tended to include actions from the script that were not actually mentioned in the story. Bower et al. (1979) also used some passages in which an action from a script was explicitly mentioned but out of its usual order in the script. In their recall of such a passage, people tended to recall the action of interest in its canonical position rather than in its actual position in the passage.

Cognitive psychologists interested in discourse processing further extended the idea of a schema capturing a sequence of events by proposing that during story comprehension people make use of a story grammar. A story grammar is a schema that specifies the abstract organization of the parts of a story. So, at the highest

level a story consists of a setting and one or more episodes. An episode consists of an event that presents a protagonist with an obstacle to be overcome. The protagonist then acts on a goal formulated to overcome the obstacle. That action then elicits reactions that may or may not accomplish the goal. Using such a scheme, each idea in a story can be mapped onto a hierarchical structure that represents the relationship among the ideas (e.g., Mandler 1984). We can use story grammars to predict people's sentence reading times and recall because the higher an idea is in the hierarchy, the longer the reading time and the better the memory for that idea (see *Narrative Comprehension*, *Psychology of*).

Thus, the body of research that cognitive psychologists generated in the 1970s and early 1980s supported the notion that data structures like frames and scripts held promise for capturing interesting aspects of human information processing. Nevertheless, it soon became clear that modifications of the whole notion of a schema were necessary in order to account for people's flexibility in comprehension and memory processing.

3. *Schema Theory Revised: The New Synthesis*

Certainly, information in a text or the context of an actual event could serve as a cue to trigger a relevant script. But what about situations that are more novel and no existing script is triggered? At best, human information processing can only sometimes be guided by large-scale schematic structures. Traditionally, schema theorists have tended to view schema application as distinct from the processes involved in building up schemata in the first place. So, people might use a relevant schema whenever possible, but when no relevant schema is triggered another set of processes are initiated that are responsible for constructing new schemata.

More recently, however, an alternative view of schemata has been developed, again with contributions both from AI and cognitive psychology. This newer approach places schematic processing at one end of a continuum in which processing varies from data driven to conceptually driven. Data-driven (also known as bottom-up) processing means that the analysis of a stimulus is not appreciably influenced by prior knowledge. In the case of text comprehension, this would mean that the mental representation of a text would stick closely to information that was directly stated. Conceptually driven (also known as top-down) processing means that the analysis of a stimulus is guided by pre-existing knowledge, as, for example, described by traditional schema theory. In the case of text comprehension, this would mean that the reader makes many elaborative inferences that go beyond the information explicitly stated in the text. In many cases, processing will be an intermediate point on the continuum so that there is a modest conceptually driven influence by pre-existing knowledge structures.

This more recent view of schemata has been called schema assembly theory. The central idea of this view is that a schema, in the sense of a framework guiding our interpretation of a situation or event, is always constructed on the spot based on the current context. Situations vary in how much background knowledge is readily available, so sometimes a great deal of related information is accessed that can guide processing and at other times very little of such background knowledge is available. In these cases we process new information in a data-driven fashion.

This new view of schemata was synthesized from work in AI and cognitive psychology as researchers realized the need to incorporate a more dynamic view of the memory system into models of knowledge-based processing. For example, AI programs based on scripts performed poorly when confronted with the kinds of script variations and exceptions seen in the 'real world.' Accordingly, Schank (1982) began to develop systems in which appropriate script-like structures were built to fit the specific context rather than just retrieved as a precomplied structure. Therefore, different possible restaurant subscenes could be pulled together in an *ad hoc* fashion based on both background knowledge and current input. In short, Schank's simulations became less conceptually driven in their operation. Likewise, cognitive psychologists, confronted with empirical data that was problematic for traditional schema theories, have moved to a more dynamic view. There were two main sets of empirical findings that motivated cognitive psychologists to make these changes.

First, it became clear that people's processing of discourse was not as conceptually driven as classic schema theories proposed. In particular, making elaborative inferences that go beyond information directly stated in the text is more restricted than originally believed. People may be quite elaborative when reconstructing a memory that is filled with gaps, but when tested for inference making concurrent with sentence processing, people's inference generation is modest. For example, according to the classic view of schemata, a passage about someone digging a hole to bury a box of jewels should elicit an inference about the instrument used in the digging. The default value for such an instrument would probably be a shovel. Yet, when people are tested for whether the concept of shovel is primed in memory by such a passage (for example, by examining whether the passage speeds up naming time for the word 'shovel'), the data suggest that people do not spontaneously infer instruments for actions. Some kinds of inferences are made, such as connections between goals and actions or filling in some general terms with more concrete instances, but inference making is clearly more restricted than researchers believed in the early days after the re-

discovery of schemata (e.g., Graesser and Bower 1990, Graesser et al. 1997).

Second, it is not always the case that what we remember from some episode or from a text is based largely on pre-existing knowledge. As Walter Kintsch has shown, people appear to form not one, but three memories as they read a text: a short-lived representation of the exact wording, a more durable representation that captures a paraphrase of the text, and a situation model that represents what was described by the text, rather than the text itself (e.g., Kintsch 1988). It is this third type of representation that is strongly influenced by available schemata. However, whether or not a situation model is formed during comprehension depends on not only the availability of relevant chunks of knowledge, but also the goals of the reader (*Memory for Meaning and Surface Memory*; *Situation Model*: *Psychological*).

Thus, background knowledge can affect processing of new information in the way described by schema theories, but our processing represents a more careful balance of conceptually-driven and data-driven processes than claimed by traditional schema theories. Moreover, the relative contribution of background knowledge to understanding varies with the context, especially the type of text and the goal of the reader.

To accommodate more flexibility in the use of the knowledge base, one important trend in research on schemata has been to incorporate schemata into connectionist models. In connectionist models the representation of a concept is distributed over a network of units. The core idea is that concepts, which are really repeatable patterns of activity among a set of units, are associated with varying strengths. Some concepts are so often associated with other particular concepts that activating one automatically entails activating another. However, in most cases the associations are somewhat weaker, so that activating one concept may make another more accessible if needed. Such a system can function in a more top-down manner or a more bottom-up manner depending on the circumstances. To rephrase this idea in terms of traditional schema theoretic concepts, some slots are available but they need not be filled with particular default values if the context does not require it. Thus, the particular configuration of the schema that is used depends to a large extent on the processing situation (see *Cognition*, *Distributed*).

In closing, it is important to note that the success that researchers in cognitive psychology and AI have had in delineating the content of schemata and in determining when such background knowledge influences ongoing processing has had an important influence on many other ideas in psychology. The schema concept has become a critical one in clinical psychology, for example in cognitive theories of depression, and in social psychology, for example in understanding the influence of stereotypes on person perception (e.g., Hilton and von Hipple 1996). It remains to be seen whether the concept of a schema becomes as dynamic in these areas of study as it has in cognitive psychology.

See also: Connectionist Models of Concept Learning; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Knowledge (Explicit and Implicit): Philosophical Aspects; Knowledge Representation; Knowledge Spaces; Mental Representations, Psychology of; Metaphor and its Role in Social Thought: History of the Concept; Reference and Representation: Philosophical Aspects; Schemas, Social Psychology of

Bibliography

Bartlett F C 1932 *Remembering*. Cambridge University Press, Cambridge, UK

Bower G H, Black J B, Turner T J 1979 Scripts in memory for text. *Cognitive Psychology* **11**: 177–220

Bransford J D, Johnson M K 1972 Contextual prerequisites for understanding: A constructive versus interpretive approach. *Journal of Verbal Learning and Verbal Behavior* **11**: 717–26

Brewer W F, Treyens J C 1981 Role of schemata in memory for places. *Cognitive Psychology* **13**: 207–30

Graesser A C, Bower G H (eds.) 1990 *Inferences and Text Comprehension: The Psychology of Learning and Motivation*. Academic Press, San Diego, CA, Vol. 25

Graesser A C, Millis K K, Zwaan R A 1997 Discourse comprehension. *Annual Review of Psychology* **48**: 163–89

Head H 1920 *Studies in Neurology*. Oxford University Press, London

Hilton J L, von Hipple W 1996 Stereotypes. *Annual Review of Psychology* **47**: 237–71

Kant I 1958 *Critique of Pure Reason*, trans. by N. K. Smith. Modern Library, New York. Originally published in 1781

Kintsch W 1988 The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* **95**: 163–82

Mandler J M 1984 *Stories, Scripts, and Scenes: Aspects of Schema Theory*. Erlbaum Associates, Hillsdale, NJ

Minsky M 1975 A framework for representing knowledge. In: Winston P H (ed.) *The Psychology of Computer Vision*. McGraw-Hill, New York

Schank R C 1982 *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, New York

Schank R C, Abelson R P 1977 *Scripts, Plans, Goals, and Understanding*. Erlbaum Associates, Hillsdale, NJ

Whitney P, Budd D, Bramucci R S, Crane R S 1995 On babies, bath water, and schemata: A reconsideration of top-down processes in comprehension. *Discourse Process* **20**: 135–66

P. Whitney

Schemas, Social Psychology of

Schemas (or schemata) are generic knowledge structures that summarize past experiences and provide a framework for the acquisition, interpretation, and

retrieval of new information. For example, one might have a clown schema based on past encounters with, and previously learned knowledge about, clowns. When one encounters a new clown, one's clown schema may cause one to notice its painted face, expect polka-dotted pantaloons, interpret its actions as goofy, and recall the presence of a unicycle. Usually, such operations are functional, leading to more rapid, accurate, and detailed information processing. However, such operations can lead to biases or errors when they produce inaccurate information (e.g., when one mistakenly recalls a clown as having a bulbous red nose).

1. History

Many of the fundamental precepts underlying the schema concept were anticipated by the Gestalt movement in European psychology. Bartlett (1932) incorporated these ideas into his concept of schema. He posited that past memories are stored as larger, organized structures rather than as individual elements, and that newly perceived information is accommodated into these 'masses' of old information. He suggested that this accommodation process is accomplished primarily through unconscious inference and faulty memory. Consequently, his theory focused more on the dysfunctional, reconstructive, and inaccurate nature of schemas. Because of his emphasis on subjective experience and the unconscious, his views and research methods departed from the behaviorist leanings of American psychology, and were largely ignored until the cognitive revolution in the 1960s and 1970s.

In the mid-1970s, schema theory was refined by cognitive psychologists (see Brewer and Nakamura 1984 and Hastie 1981). While incorporating Bartlett's basic premises, these modern approaches also acknowledged the functional benefits of schemas in terms of increased cognitive efficiency and accuracy. Furthermore, these modern approaches extended the concept of schema by incorporating it within contemporary cognitive theory. For example, schemas are now construed as (a) selectively activated, just like any other concept in memory, (b) nested within each other, with each defined partly through reference to related schemas, (c) providing 'default values' for embedded informational features, and (d) having active, testable memory functions. These ideas stimulated considerable work on schematic processing within cognitive psychology.

2. Schematic Processing

For a schema to affect processing it must first be activated. Because of the unitary and organized nature of schemas, this is generally assumed to be an all-or-nothing process, in that the entire schema (as opposed to bits and pieces) either comes to mind completely or not at all. For example, one could not activate a chair schema without becoming aware of both the typical appearance of chairs and their normal seating function.

Research has identified a number of variables that influence when particular schemas are activated by new information, including fit, context and priming. In terms of fit, for example, a chair with typical features like four legs, armrests, and a back is more likely to activate the chair schema than is a three-legged stool. Context can also influence activation, so that an ice chest is more likely to activate a chair schema at a campsite than on a store shelf. Finally, schemas can be primed in memory through frequent or recent experiences that bring them to mind. For example, reading about chairs in this paragraph could increase the likelihood that you would activate a chair schema in response to a tree stump.

Once activated, schemas appear to have complex effects on attention. Overall, research suggests that schemas direct attention toward schema-relevant, rather than schema-irrelevant, information. However, it makes some difference whether the schema-relevant information is schema congruent or schema incongruent. Schema-incongruent information captures the most attention, presumably because it violates expectations. Schema-congruent information is attended to while an appropriate schema is in the process of being activated or instantiated, but ignored once this activation is accomplished. For example, early in the process a seat might be noticed because it helps to identify an object as a chair, but once a chair schema is activated, the seat would be ignored as obvious.

Schemas have also been shown to have powerful effects on the interpretation (or encoding) of new information, providing a framework that shapes expectancies and judgments. Research suggests that this influence is especially pronounced when the new information is ambiguous, vague or incomplete. For example, the ambiguous word 'chair' would be interpreted differently if you had a department head schema activated than if you had a furniture schema activated.

Finally, research indicates that schemas affect the retrieval of information from memory. Memory effects are often viewed as reflecting the organizational properties of schemas. Because schemas are large, organized clusters of linked information, they provide multiple routes of access for retrieving individual items of information. Such retrieval routes have been shown to facilitate memory in a number of ways. For example, experts have particularly complex and well-organized schemas and consequently exhibit better recall for domain-relevant information. Thus, as chess expertise increases, so does memory for board arrangements from actual games (though not for random arrangements).

Enhanced recall for schema-related information can reflect either retrieval or guessing mechanisms. For

example, if a newspaper article activates a prisoner execution schema, the link between such executions and electric chairs could help one to recall that an electric chair was mentioned. On the other hand, the schema may instead facilitate an educated guess. As noted previously, schemas have been found to provide defaults for filling in missing information. Thus an electric chair might be 'remembered' even if it were not mentioned in the story. Such guesses produce biases and errors when the inferred information is incorrect.

3. Criticisms of the Schema Concept

A number of criticisms were leveled at the schema concept in the 1980s (e.g., Alba and Hasher 1983, Fiske and Linville 1980). First, schemas were criticized as definitionally vague, with little consensus regarding concept boundaries. In other words, it is unclear what differentiates schemas from other kinds of knowledge, attitudes or attributions. Second, research on schemas has been attacked as operationally vague, especially because of unvalidated manipulations. For example, researchers sometimes assume that schemas are activated by cues or expectancies in the absence of manipulation checks evidencing such activation. Third, work on schemas has been denounced as nonfalsifiable, in that virtually any result can be interpreted as reflecting schema use. Fourth, it has been argued that schema theory has difficulty accounting for the complexity and accuracy of memory. For instance, there is evidence that people encode and remember more than simply the gist of events, and that schematic distortions are the exception rather than the rule. Finally, some critics have complained that schema theory and research amounts to little more than the reframing of previously known phenomena in terms of a new, schematic vocabulary. For example, biases in person perception documented as early as the 1940s have since been reframed as schema effects.

On the other hand, schema theory survived such criticisms because its strengths clearly outweigh its weaknesses. The breadth of the 'vague' schema concept allows it to encompass a wide variety of phenomena, ranging from autobiographical memory to visual perception. In many such contexts, the schema concept has been embedded in elaborate theory, allowing for more precise definition, manipulation, and measurement. With such refinements, schema theory often makes specific, falsifiable predictions that lead to clearer, if not totally unambiguous, interpretations. Additionally, modern schema research has increasingly focused on the ways that schemas facilitate, rather than interfere with, memory, accounting better for the frequent accuracy of memory. Overall, the schema concept has proven to be tremendously heuristic, leading to novel research that has documented important new phenomena.

4. Social Psychological Research

Social psychologists have applied the schema concept to complex interpersonal phenomena often using cognitive methods and theory. Given this approach, research on schemas has become an integral part of the subfield known as social cognition. Such research has not only elaborated the kinds of information processing effects described in the cognitive literature, but has also led to the identification of a wide variety of new phenomena.

4.1 Schematic Processing of Social Information

Social psychologists have examined the effects of trait schemas on the processing of schema-related information (see Taylor and Crocker 1981). This research found patterns of recall for congruent, incongruent, and irrelevant information paralleling the patterns described earlier for cognitive stimuli. Specifically, people best recall trait-incongruent information, followed by trait-congruent information, with trait-irrelevant information last. For example, if people have an impression (schema) of Jamie as honest, they would best recall his dishonest behaviors (e.g., embezzling money from the honor society), because this information is surprising and needs to be reconciled with Jamie's other behaviors. They would next best recall Jamie's honest behaviors (e.g., returning a lost wallet), because these are expected and, thus, readily fit into the existing schema. Jamie's honesty-irrelevant behaviors (e.g., consulting on statistical issues) would be worst recalled because they do not relate in any way to the honesty schema. Perhaps one of the most significant contributions of this social psychological work has been to move beyond the differential attention explanation for these findings. Such effects are now understood to also reflect behavior-to-behavior or behavior-to-schema interconnections that are created as people think about how this information can be reconciled. Incongruent information actually becomes most interconnected as people mull it over while trying to make sense of it, and irrelevant information remains least interconnected, because it warrants little reflection.

4.2 Self-schemas

Social psychologists have also used the concept of schema to understand the nature of the self, which had been debated by psychologists for a century (see Linville and Carlston 1994). Most early empirical work focused on the content of the self-concept, though such work was hindered by vague conceptions of self and by the idiosyncratic nature of self-knowledge. The conception of self as a schema provided a more coherent definition, and shifted the empirical emphasis from the content of the self to the functions of self-schemas. The consequence was a proliferation

of social psychological research in this area. Considerable empirical effort went into assessing whether the self is a unique kind of schema, distinct from other forms of social knowledge. Although there are exceptions, most studies have failed to find convincing evidence that self-schemas are unique or distinct from other forms of knowledge. Consequently, the seemingly 'unique' effects of self-schemas may simply reflect the greater complexity and personal relevance of self-knowledge.

Researchers have also explored whether people have just one, unified self-schema, or whether people have multiple self-schemas. Most contemporary theories depict people as having multiple self-schemas, with the one that is currently activated (in use) being referred to as the working or phenomenal self. Several approaches to the idea of multiple selves have been highly influential. One approach (self-complexity theory) suggests that people have different self-schemas representing their varied social roles and relationships such as student, parent, athlete, lover, and so on. When many such self-schemas implicate different attributes, then failures in one realm may be offset by successes in another, leading to emotional buffering and enhanced mental health. On the other hand, when one has few self-schemas, and these overlap considerably, then failure in any realm can have a devastating affect on one's self-esteem.

Another approach (self-discrepancy theory) identifies three kinds of self-schema: the actual self that we each believe describes us as we are, the ideal self that we aspire to become, and the ought self that others expect us to live up to. According to this theory, increased (rather than decreased) overlap among self-schemas leads to enhanced mental health. For example, discrepancies between the actual and ideal selves leads to dysphoria, and discrepancies between the actual and ought selves leads to anxiety. Theorists now construe multiple self-schemas in terms of the different kinds of information that is activated or accessible at any given time. A number of factors affect which self-schema is currently activated. The contexts or situations we find ourselves in can cue particular self-schemas. For instance, a work setting generally will cue our professional selves whereas a home setting will generally cue our familial selves.

The self-aspects that are most salient to an individual are also influenced by individual differences. For example, members of Mensa may chronically categorize themselves and others in terms of intelligence. Such 'schematics' tend to view the chronic trait as important and to rate themselves highly on that dimension compared with 'aschematics.'

A related issue involves whether self-schemas are stable or change over time. In general, schemas are viewed as fairly stable representations, and this is even more pronounced for self-schemas. However in recent years, research has focused on factors underlying the occasional malleability of the self-concept. Researchers have shown that major life events, such as losing a job or experiencing the death of a loved one, can affect the self-schema substantially. Additionally, our self-schemas can change when we obtain new self-knowledge (e.g. learning our IQ). Relatedly, our self-schemas can be influenced by how others think about and treat us. For example, research suggests that if teachers treat students as incompetent, the students tend to incorporate this view into their self-schema and behave accordingly. Finally, our self-schemas change over time, reflecting normal age-related changes in maturity, social and cognitive development, and complexity.

4.3 Biases in Person Perception

As noted earlier, schemas serve a number of cognitive functions, such as providing a framework for interpreting information and furnishing default values for unobserved or unrecalled details. When these inferences are incorrect, they produce biases and errors. For example, impressions of others can be influenced by context schemas, as when a person is viewed as more athletic when encountered at a gym than when encountered at a funeral. Similarly, impressions of a new acquaintance can be influenced by others present in the environment, even when there is no logical reason for this. For instance, relatives of physically disabled people may be viewed as having some of the same limitations simply because the disability schema has been activated. Impressions can also be influenced by *déja vu* effects. If you meet someone new who resembles someone you already know, activation of the resembled-person's schema may cause you to assume that the new person has similar traits and characteristics. Role schemas can also bias impressions. If you know two nuns, you may confuse them because they activate similar role schemas. These are just a couple examples of many ways in which schemas can alter impressions.

4.4 Types of Social Schema

As the preceding review suggests, social psychologists have identified many different types of social schema. These include previously-mentioned role schemas (e.g., occupations), relationship schemas (e.g., parent), and trait schemas (e.g., honesty). Additionally, impressions of individuals have been construed as person schemas. Event schemas represent common 'scripts' as ordered sequences of actions that comprise a social event, such as attending a wedding or dining at a restaurant. Such event schemas facilitate memory for social events, as evidenced by research showing that people have difficulty remembering events that are presented to them in an illogical order. There are also nonverbal schemas composed of series of physical acts

(sometimes called procedural schemas), such as a schema for riding a bike. Similarly, simple judgment rules or heuristics (e.g., smooth talkers are more believable) are sometimes viewed as nonverbal, procedural schemas.

Finally, a wide variety of stereotype schemas have been identified in the social literature. Research suggests that race, age, and gender schemas, in particular, are used automatically to categorize others, presumably because these cues are visually salient and our culture defines them as important. Some research has suggested that these automatic schema effects can be overridden when people have sufficient awareness, motivation and cognitive resources to do so (e.g., Devine 1989). However, this issue remains controversial.

5. Future Directions

The term *schema* is no longer in vogue, although the essential features of the concept of schemas have been incorporated within broader theories of knowledge structure and mental representation. Work in this area promises to become increasingly sophisticated in several respects. Sophisticated new cognitive theories (e.g., connectionism) provide a better understanding of how schemas form and evolve as new information is acquired. Theorists are better defining and differentiating different types of cognitive structures and mechanisms. Research is providing better evidence for the kinds of representations assumed to underlie different phenomena. Such developments are refining our understanding of the schema concept and how the mind works in general. Ultimately, this work will prove most interesting and impactful as it demonstrates that schematic representations have behavioral as well as cognitive consequences.

See also: Piaget's Theory of Child Development; Schemas, Frames, and Scripts in Cognitive Psychology; Social Psychology, Theories of

Bibliography

Alba J W, Hasher L 1983 Is memory schematic? *Psychological Bulletin* **93**: 203–31

Bartlett F C 1932 *Remembering*. Cambridge University Press, Cambridge, UK

Bartlett F C 1967 *Remembering. A Study in Experimental and Social Psychology*. Cambridge University Press, London

Brewer W F, Nakamura G V 1984 The nature and functions of schemas. In: Wyer Jr. R S, Srull T K (eds.) *Handbook of Social Cognition Vol. 1*. Erlbaum, Hillsdale, NJ

Devine P G 1989 Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* **56**: 5–18

Fiske S T, Linville P W 1980 What does the schema concept buy us? *Personality and Social Psychology Bulletin* **6**: 543–57

Hastie R 1981 Schematic principles in human memory. In: Higgins E T, Herman C P, Zanna M P (eds.) *Social Cognition: The Ontario Symposium*. Erlbaum, Hillsdale, NJ, Vol. 1

Higgins E T 1987 Self discrepancy: A theory relating self and affect. *Psychological Review* **94**: 319–40

Linville P W 1985 Self complexity and affective extremity: Don't put all of your eggs in one cognitive basket. *Social Cognition* **3**: 94–120

Linville P W, Carlston D E 1994 Social cognition of the self. In: Devine P G, Hamilton D L, Ostrom T M (eds.) *Social Cognition: Impact on Social Psychology*. Academic Press, San Diego

Smith E R 1998 Mental representation and memory. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. Oxford University Press, New York

Taylor S E, Crocker J 1981 Schematic bases of social information processing. In: Higgins E T, Herman C P, Zanna M P (eds.) *Social Cognition: The Ontario Symposium Vol. 1*. Erlbaum, Hillsdale, NJ

D. E. Carlston and L. Mae

Schizophrenia

Schizophrenia is a psychosis—a severe mental disorder in which the person's emotions, thinking, judgment, and grasp of reality are so disturbed that his or her functioning is seriously impaired. The symptoms of schizophrenia are often divided into 'positive' and 'negative.' Positive symptoms are abnormal experiences and perceptions like delusions, hallucinations, illogical and disorganized thinking, and inappropriate behavior. Negative symptoms are the absence of normal thoughts, emotions, and behavior; such as blunted emotions, loss of drive, poverty of thought, and social withdrawal.

Despite common features, different forms of schizophrenia are quite dissimilar. One person, for example, may be paranoid, constantly bothered by voices warning him or her about plots or threats, but able to show good judgment and high functioning in many areas of life. Another may be bizarre in manner and appearance, preoccupied with delusions of bodily disorder, passive and withdrawn. So marked are the differences that many researchers believe that the illness will eventually prove to be a set of different conditions which lead to somewhat similar consequences.

1. Diagnostic Difficulties

It is difficult to define schizophrenia precisely. The two most common functional psychoses are schizophrenia and bipolar disorder (also known as manic-depressive illness). The distinction between the two is not easy to make and psychiatrists in different parts of the world at different times have drawn the boundaries in different ways. Bipolar disorder is an episodic disorder in which psychotic symptoms are associated with severe alterations in mood—at times elated, agitated episodes of mania, at other times depression, with

physical and mental slowing, despair, guilt, and low self-esteem. The course of schizophrenia, by way of contrast, though fluctuating, tends to be more continuous, and the person's display of emotion is likely to be incongruous or lacking in spontaneity. Markedly illogical thinking is common in schizophrenia. Auditory hallucinations may occur in either bipolar disorder or schizophrenia, but in schizophrenia they are more likely to be commenting on the person's actions or to be conversing one with another. Delusions, also, can occur in both conditions; in schizophrenia they may give the individual the sense that he or she is being controlled by outside forces or that his or her thoughts are being broadcast or interfered with.

2. *Schizophrenia is Universal*

Schizophrenia is a universal condition and an ancient one. Typical cases are evident in the medical writings of ancient Greece and Rome, and the condition occurs today in every human society. While the content of delusions and hallucinations varies from culture to culture, the form of the illness is similar everywhere. Two World Health Organization studies, applying a standardized diagnostic approach, have identified characteristic cases of schizophrenia in developed and developing world countries from around the globe (World Health Organization 1979, Jablensky et al. 1992). One of these studies (Jablensky et al. 1992) demonstrated that the rate of occurrence of new cases (the incidence) is similar in every country studied from India to Ireland—around one per 10,000 adults each year. However, since both death and recovery rates for people with psychosis are higher in the Third World, the point prevalence of schizophrenia (the number of cases existing at any point in time) is lower in the Third World—around 3 per 1,000 of the adult population compared to 6 per 1,000 in the developed world (Warner and de Girolamo 1995). The risk of developing the illness at some time in one's life (the lifetime prevalence) is a little higher—around one percent of the developed world population. See *Mental Illness, Epidemiology of*.

3. *Conceptual Framework*

The bio-psychosocial model of illness clarifies how different factors shape schizophrenia. The model posits that the predisposition to developing an illness, its onset, and its course are each influenced by biological, psychological, and sociocultural factors. A variety of factors can affect the different phases of schizophrenia, many being environmental. Some, such as genetics, gender, and synaptic pruning are innate. Biological, psychological and social factors are involved to some extent in most phases of schizophrenia. In general, however, the research suggests that the factors responsible for the predisposition to developing an illness are more likely to be biological, that psychological factors are often important in triggering the onset, and that the course and outcome are particularly likely to be influenced by sociocultural factors.

4. *The Course and Outcome of Schizophrenia*

Wide variation occurs in the course of schizophrenia. In some cases the onset of illness is gradual, extending over months or years; in others it begins suddenly. Some have episodes of illness lasting weeks or months with full remission of symptoms between each episode; others have a fluctuating course in which symptoms are continuous; others again have little variation over the course of years. Swiss psychiatrist, Luc Ciompi, studied the onset, course, and outcome of illness in people with schizophrenia, following them into old age (Ciompi 1980). He found that the onset of the illness was either acute (less than six months from first symptoms to full-blown psychosis) or insidious in roughly equal numbers of cases; the course was episodic or continuous in approximately equal numbers of patients; and the outcome was moderate to severe disability in half the cases and mild disability or full recovery in the other half.

Despite popular and professional belief, schizophrenia does not have a progressive, downhill course with universally poor outcome. In fact, schizophrenia usually becomes less severe as the sufferer grows older. A review of outcome studies conducted in Europe and North America throughout the twentieth century reveals that, over the course of months or years, 20 to 25 percent of people with schizophrenia recover completely from the illness—all their psychotic symptoms disappear and they return to their previous level of functioning. Another 20 percent continue to have some symptoms, but are able to lead satisfying and productive lives (Warner 1994). In the developing countries, recovery rates are even better. The two World Health Organization studies mentioned above (World Health Organization 1979, Jablensky et al. 1992) have shown that good outcome occurs in about twice as many patients diagnosed with schizophrenia in the developing world as in the developed world. The reason for the better outcome in the Third World is not completely understood, but it may be that many people with mental illness in developing world villages are better accepted, less stigmatized, and more likely to find work in a subsistence agricultural economy (Warner 1994).

5. *Factors Affecting the Course of Schizophrenia*

Age of onset: The later in life the illness begins, the milder it proves to be. Onset of schizophrenia before the age of 14 is rare, but when it does begin this early it is associated with a severe course of illness. Onset after the age of 40 is also rare, and is associated with a milder course.

Gender: Women usually develop their first symptoms of schizophrenia later than men and the course of their illness tends to be less severe. The reason for these differences is not clear, but may be related to the protective effect of female hormones on brain development and function.

Stressful life events: Stress can trigger episodes of schizophrenia. People with schizophrenia are more likely to report a stressful life event preceding an episode of illness than during a period of remission. Similarly, stressful events are more likely to occur prior to an episode of schizophrenia than in the same time period for people drawn from the general population (Rabkin 1982). Life stress is also more common before the first episode of schizophrenia and so, although stress does not cause the illness, it may well influence the timing of onset.

The research indicates that the life events occurring before episodes of schizophrenia are milder and less objectively troublesome than those before episodes of other disorders such as depression (Beck and Worthen 1972), suggesting that people with schizophrenia are exquisitely sensitive to stress.

Domestic stress: The robust results of the 'expressed emotion' (EE) research, conducted in several countries in the developed and developing worlds, reveal that people with schizophrenia living with relatives who are critical or over-involved (referred to in the research as high EE) have a higher relapse rate than those living with relatives who are less critical or intrusive (low EE) (Leff and Vaughn 1985, Parker and Hadzi-Pavlovic 1990). A meta-analysis of 26 EE studies of schizophrenia conducted in 11 countries indicates that the relapse rate over a two-year follow-up period was more than twice as high, at 66 percent, for patients in families which included a high EE relative than in low EE households (29 percent) (Kavanagh 1992). Other studies have shown that relatives who are less critical and over-involved exert a positive therapeutic effect on the person with schizophrenia, their presence leading to a reduction in the patient's level of arousal (Sturgeon et al. 1981). There is no indication that the more critical and over-involved relatives are abnormal by everyday Western standards; low EE family members may be unusually low-key and permissive (Angermeyer 1983).

Several studies have shown that family psychoeducational interventions can lead to a change in the level of criticism and over-involvement among relatives of people with schizophrenia and so reduce the relapse rate (Falloon et al. 1982, Berkowitz et al. 1981). Effective family interventions provide three basic ingredients: (a) education about the illness; (b) help in developing problem-solving mechanisms; and (c) practical and emotional support (McFarlane 1983, Falloon et al. 1982, Leff and Vaughn 1985).

Substance use: Drug and alcohol abuse is more common among people with serious mental illness. Unemployment, social isolation, and alienation may contribute to these high rates. Several studies have shown that people with serious mental illness who abuse substances have a worse course of illness (Drake and Wallach 1989) but other researchers have found psychopathology to be no worse or, sometimes, lower among people with mental illness who use substances (Zisook et al. 1992, Buckley et al. 1994). One reason for this discrepancy may lie in the common finding that substance users are also more likely to be noncompliant with treatment (Drake and Wallach 1989); the poor course of illness, when it is observed, may be a result of this noncompliance rather than a direct consequence of substance use.

6. *Causes of Schizophrenia*

There is no single organic defect or infectious agent which causes schizophrenia, but a variety of factors increase the risk of developing the illness, including genetic factors and obstetric complications.

Genetic factors. Relatives of people with schizophrenia have a greater risk of developing the illness, the risk being progressively higher among those who are more genetically similar to the person with schizophrenia (see Fig. 1). For a second-degree relative the lifetime risk is about two percent (twice the risk for someone in the general population); for a first-degree relative is about ten percent, and for an identical twin (genetically identical to the person with schizophrenia) the risk is close to 50 percent (Gottesman 1991).

Studies of people adopted in infancy reveal that the increased risk of schizophrenia among the relatives of people with the illness is due to inheritance rather than environment. The children of people with schizophrenia have the same increased prevalence of the illness whether they are raised by their biological parent with schizophrenia or by adoptive parents (Gottesman 1991, Warner and de Girolamo 1995).

There is evidence implicating several genes in causing schizophrenia (Wang et al. 1995, Freedman et al. 1997), and it is likely that more than one is responsible, either through an interactive effect or by producing different variants of the disorder. See *Mental Illness, Genetics of.*

Obstetric complications. A review and meta-analysis of studies conducted prior to mid-1994 on the effect of obstetric complications in schizophrenia, reveals that complications before and around the time of birth appear to double the risk of developing the illness (Geddes and Lawrie 1995). In a more recent meta-analysis of a different sample of studies, the same researchers found the risk to be increased by a factor of 1.4.

Since these analyses were published, more recent studies have shown variable results. Data gathered at the time of birth from very large cohorts of children born in Finland and Sweden in the 1960s and 1970s indicate that various obstetric complications double

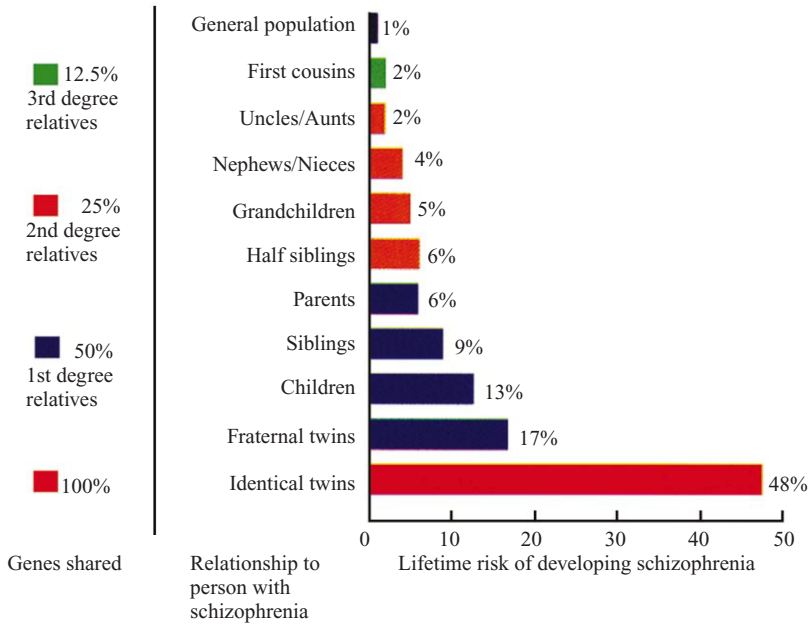

Figure 1
The average risk of developing schizophrenia for relatives of a person with the illness, compiled from family and twin studies conducted in Europe between 1970 and 1987. Reprinted by permission of the author from Gottesman (1991, p. 96) © 1991 Irving I. Gottesman

or triple the risk of developing schizophrenia (Hultman et al. 1999, Dalman et al. 1999, Jones et al. 1998). An American study shows that the risk of schizophrenia is more than four times greater in those who experience oxygen deprivation before or at the time of birth, and that such complications increase the risk of schizophrenia much more than other psychoses like bipolar disorder (Zornberg et al. 2000). Recent Scottish research, on the other hand, found the effect of obstetric complications to be less than in prior studies (Kendell et al. 2000). Other recent research suggests that only early-onset (before age 30) cases of schizophrenia are associated with obstetric complications (Byrne et al. 2000, Cannon et al. 2000).

Obstetric complications are a statistically important risk because they are so common. In the general population, they occur in up to 40 percent of births (the precise rate of occurrence depending on how they are defined) (McNeil 1988, Geddes and Lawrie 1995). The authors of the meta-analyses cited above estimate that complications of pregnancy and delivery may increase the prevalence of schizophrenia by 20 percent (Geddes and Lawrie 1995).

The obstetric complications most closely associated with the increased risk of developing schizophrenia are those that induce fetal oxygen deprivation, particularly prolonged labor (McNeil 1988), and placental complications (Jones et al. 1998, Hultman et al. 1999, Dalman et al. 1999). Early delivery is also more common for those who go on to develop schizophrenia, and infants who suffer perinatal brain damage are at a much-increased risk of subsequent schizophrenia (Jones et al. 1998). Trauma at the time of labor and delivery, and especially prolonged labor, is associated with an increase in structural brain abnormalities—cerebral atrophy and small hippocampi—which occur frequently in schizophrenia (McNeil et al. 2000).

Viruses. The risk of intrauterine brain damage is increased if a pregnant woman contracts a viral illness. We know that more people with schizophrenia are born in the late winter or spring than at other times of year, and that this birth bulge sometimes increases after epidemics of viral illnesses like influenza, measles, and chickenpox. Maternal viral infections, however, probably account for only a small part of the increased risk for schizophrenia (Warner and de Girolamo 1995).

7. *Myths about the Causes of Schizophrenia*

Parenting: Contrary to the beliefs of professionals prior to the 1970s and to the impression still promoted by the popular media, there is no evidence, even after decades of research, that family or parenting problems cause schizophrenia. As early as 1948, psychoanalysts proposed that mothers fostered schizophrenia in their

offspring through cold and distant parenting. Other theorists blamed parental schisms, and confusing patterns of communication within the family (Lidz et al. 1965, Laing and Esterton 1970). The double-bind theory, put forward by anthropologist Gregory Bateson, argued that schizophrenia is promoted by contradictory parental messages from which the child is unable to escape (Bateson et al. 1956). While enjoying broad public recognition, such theories have seldom been adequately tested, and none of the research satisfactorily resolves the question of whether differences found in the families of people with schizophrenia are the *cause* or the *effect* of psychological abnormalities in the disturbed family member (Hirsch and Leff 1975). Millions of family members of people with schizophrenia have suffered needless shame, guilt, and stigma because of this widespread misconception.

Drug abuse: Drug abuse does not cause schizophrenia, though it is possible, but by no means certain, that it can trigger the onset of the illness. Hallucinogenic drugs like LSD can induce short episodes of psychosis and heavy use of marijuana and stimulant drugs like cocaine and amphetamines may precipitate brief, toxic psychoses with features similar to schizophrenia (Bowers 1987), but there is no evidence that these drugs cause a long-lasting illness like schizophrenia. In the 1950s and 1960s, LSD was used as an experimental drug in psychiatry in Britain and America, but the proportion of these volunteers and patients who developed a long-lasting psychosis was scarcely greater than in the general population (Cohen 1960, Malleson 1971). A Swedish study found that army conscripts who used marijuana heavily were six times more likely to develop schizophrenia later in life (Andreasson et al. 1987), but this may well have been because those people who were destined to develop schizophrenia were more likely to use marijuana as a way to cope with the premorbid symptoms of the illness. Schizophrenia is preceded by a long period of prodromal symptoms, and a German study has demonstrated that the onset of drug and alcohol abuse in people with schizophrenia usually *follows* the very first negative symptom of schizophrenia (such as social withdrawal) but precedes the first positive symptom (such as hallucinations). The authors conclude that substance use is an avenue to the relief of the earliest symptoms of the illness, but is not a cause (Hambrecht and Hafner 1995).

8. *The Brain in Schizophrenia*

Physical changes in the brain have been identified in some people with schizophrenia. The analysis of brain tissue after death has revealed a number of structural abnormalities, and new brain-imaging techniques have revealed changes in both the structure and function of the brain during life. Techniques such as magnetic resonance imaging (MRI) reveal changes in the size of different parts of the brain, especially the temporal lobes. The fluid-filled spaces (the ventricles) in the interior of the temporal lobes are often enlarged and the temporal lobe tissue diminished. The greater the observed changes the greater the severity of the person's thought disorder and auditory hallucinations (Suddath et al. 1990).

Some imaging techniques, such as positron emission tomography (PET), measure the actual functioning of the brain and provide a similar picture of abnormality. PET scanning reveals hyperactivity in the temporal lobes, particularly in the hippocampus, a part of the temporal lobe concerned with episodic memory (Tamminga et al. 1992). Another type of functional imaging, electrophysiological brain recording using EEG tracings, shows that most people with schizophrenia seem to be excessively responsive to repeated environmental stimuli and more limited in their ability to blot out irrelevant information (Freedman et al. 1997). In line with this finding, those parts of the brain that are supposed to screen out irrelevant stimuli, such as the frontal lobe, show decreased activity on PET scan (Tamminga et al. 1992).

Tying in with this sensory screening difficulty, postmortem brain tissue examination has revealed problems in a certain type of brain cell—the inhibitory interneuron. These neurons damp down the action of the principal nerve cells, preventing them from responding to too many inputs. Thus, they prevent the brain from being overwhelmed by too much sensory information from the environment. The chemical messengers or neurotransmitters (primarily gamma-amino butyric acid or GABA) released by these interneurons are diminished in the brains of people with schizophrenia (Benes et al. 1991, Akbarian et al. 1993) suggesting that there is less inhibition of brain overload.

Abnormality in the functioning of these interneurons appears to produce changes in the brain cells that release the neurotransmitter dopamine. The role of dopamine has long been of interest to schizophrenia researchers, because drugs like amphetamines that increase dopamine's effects can cause psychoses that resemble schizophrenia, and drugs that block or decrease dopamine's effect are useful for the treatment of psychoses (Meltzer and Stahl 1976). Dopamine increases the sensitivity of brain cells to stimuli. Ordinarily, this heightened awareness is useful in increasing a person's awareness during times of stress or danger, but, for a person with schizophrenia, the addition of the effect of dopamine to an already hyperactive brain state may tip the person into psychosis.

These findings suggest that in schizophrenia there is a deficit in the regulation of brain activity by interneurons, so that the brain over-responds to environmental signals and lacks the ability to screen out unwanted stimuli. This problem is made worse by a decrease in the size of the temporal lobes, which ordinarily process

sensory inputs, making it more difficult for the person to respond appropriately to new stimuli. (See *MRI (Magnetic Resonance Imaging) in Psychiatry*.)

9. *Why Does Schizophrenia Begin after Puberty?*

Schizophrenia researchers have long been puzzled about why the illness normally begins in adolescence when important risk factors, such as genetic loading and neonatal brain damage, are present from birth or sooner. Recent research attempts to address the question.

Normal brain development leads to the loss of 30–40 percent of the connections (synapses) between brain cells during the developmental period from early life to adolescence (Huttenlocher 1979). Brain cells themselves do not diminish in number during this period, only their connectivity. It appears that we may need a high degree of connectivity between brain cells in infancy to enhance our ability to learn language rapidly (toddlers learn as many as twelve new words a day). The loss of neurons during later childhood and adolescence, however, improves our 'working memory' and our efficiency to process complex linguistic information (McGlashan and Hoffman 2000).

For people with schizophrenia, this normally useful process of synaptic pruning is excessive, leaving fewer synapses in the frontal lobes and medial temporal cortex (Feinberg 1983). In consequence, there are deficits in the interaction between these two areas of the brain in schizophrenia, which reduce the adequacy of working memory (Weinberger et al. 1992). One intriguing computer modeling exercise suggests that decreasing synaptic connections and eroding working memory in this way not only leads to abnormalities in the ability to recognize meaning when stimuli are ambiguous but also to the development of auditory hallucinations (Hoffman and McGlashan 1997).

It is possible, therefore, that this natural and adaptive process of synaptic elimination in childhood, if carried too far, could lead to the development of schizophrenia (Feinberg 1983). If true, this would help explain why schizophrenia persists among humans despite its obvious functional disadvantages and its association with reduced fertility. The genes for synaptic pruning may help us refine our capacity to comprehend speech and other complex stimuli, but, when complicated by environmental assaults resulting in brain injury, the result could be symptoms of psychosis. As yet, this formulation is speculative.

10. *Effective Interventions in Schizophrenia*

There is more agreement now about what is important in the treatment of schizophrenia than ever before. In establishing the World Psychiatric Association global project designed to combat the stigma and discrimination resulting from schizophrenia (Warner 2000), prominent psychiatrists from around the world recently agreed on the following principles:

People with schizophrenia can be treated effectively in a variety of settings. These days the use of hospitals is mainly reserved for those in an acute relapse. Outside of the hospital, a range of alternative treatment settings have been devised which provide supervision and support and are less alienating and coercive than the hospital (Warner 1995).

Family involvement can improve the effectiveness of treatment. A solid body of research has demonstrated that relapse in schizophrenia is much less frequent when families are provided with support and education about schizophrenia.

Medications are an important part of treatment but they are only part of the answer. They can reduce or eliminate positive symptoms but they have a negligible effect on negative symptoms. Fortunately, modern, novel antipsychotic medications, introduced in the past few years, can provide benefits while causing less severe side effects than the standard antipsychotic drugs which were introduced in the mid-1950s.

Treatment should include social rehabilitation. People with schizophrenia usually need help to improve their functioning in the community. This can include training in basic living skills; assistance with a host of day-to-day tasks; and job training, job placement, and work support. The psychosocial clubhouse is one effective model for providing many of these forms of assistance (Mosher and Burti 1989). The assertive community treatment model has proven effective in preventing relapse and hospital admission (Stein and Test 1980).

Work helps people recover from schizophrenia. Productive activity is basic to a person's sense of identity and worth. The availability of work in a subsistence economy may be one of the main reasons that outcome from schizophrenia is so much better in Third World villages (Warner 1994). Given training and support, most people with schizophrenia can work, as has been demonstrated by several north Italian employment programs (Warner 1994). However, due to problems such as work disincentives in disability pension schemes, high general unemployment and inadequate vocational rehabilitation services, the employment of people with schizophrenia in Britain and the United States has routinely been as low as 15 percent in recent years (Warner 2000).

People with schizophrenia can get worse if treated punitively or confined unnecessarily. Extended hospital stays are rarely necessary if good community treatment is available. Jails or prisons are not appropriate places of care. Yet, around the world, large numbers of people with schizophrenia are housed in prison cells, usually charged with minor crimes, largely because of the lack of adequate community treatment.

People with schizophrenia and their family members should help plan and even deliver treatment.

Consumers of mental health services can be successfully employed in treatment programs, and when they help train treatment staff, professional attitudes and patient outcome both improve (Sherman and Porter 1991, Warner 2000).

People's responses towards someone with schizophrenia influence the person's course of illness and quality of life. Negative attitudes can push people with schizophrenia and their families into hiding the illness and drive them away from help. If people with schizophrenia are shunned and feared they cannot be genuine members of their own community. They become isolated and victims of discrimination in employment, accommodation, and education (Warner 2000). The recent US Surgeon General's report on mental illness cited stigma as one of the most important obstacles to effective treatment (US Department of Health and Human Services 1999).

See also: Depression, Clinical Psychology of; Developmental Psychopathology: Child Psychology Aspects; Differential Diagnosis in Psychiatry; Mental and Behavioral Disorders, Diagnosis and Classification of; Mental Illness, Epidemiology of; Mental Illness, Genetics of; Psychiatric Assessment: Negative Symptoms; Schizophrenia and Bipolar Disorder: Genetic Aspects; Schizophrenia: Neuroscience Perspective; Schizophrenia, Treatment of

Bibliography

Akbarian S, Vinuela A, Kim J J, Potkin S G, Bunney W E J, Jones E G 1993 Distorted distribution of nicotinamide-adenine dinucleotide phosphate-diaphorase neurons in temporal lobe of schizophrenics implies anomalous cortical development. *Archives of General Psychiatry* **50**: 178–87

Andreasson S, Allebeck P, Engstrom A, Rydberg U 1987 Cannabis and schizophrenia: A longitudinal study of Swedish conscripts. *Lancet* **1987**(2): 1483–86

Angermeyer M C 1983 'Normal deviance': Changing norms under abnormal circumstances. Presented at the Seventh World Congress of Psychiatry, Vienna

Bateson G, Jackson D, Haley J 1956 Towards a theory of schizophrenia. *Behavioral Science* **1**: 251–64

Beck J, Worthen K 1972 Precipitating stress, crisis theory, and hospitalization in schizophrenia and depression. *Archives of General Psychiatry* **26**: 123–9

Benes F M, McSparran I, Bird E D, San Giovani J P, Vincent S L 1991 Deficits in small interneurons in prefrontal and cingulate cortices of schizophrenic and schizoaffective patients. *Archives of General Psychiatry* **48**: 996–1001

Berkowitz R, Kuipers L, Eberlein-Fries R, Leff J D 1981 Lowering expressed emotion in relatives of schizophrenics. *New Directions in Mental Health Services* **12**: 27–48

Bowers M B 1987 The role of drugs in the production of schizophreniform psychoses and related disorders. In: Meltzer H Y (ed.) *Psychopharmacology: The Third Generation of Progress*. Raven Press, New York

Buckley P, Thompson P, Way L, Meltzer H Y 1994 Substance abuse among patients with treatment-resistant schizophrenia: Characteristics and implications for clozapine therapy. *American Journal of Psychiatry* **151**: 385–9

Byrne M, Browne R, Mulryan N, Scully A, Morris M, Kinsella A, McNeil T, Walsh D, O'Callaghan E 2000 Labour and delivery complications and schizophrenia: Case-control study using contemporaneous labour ward records. *British Journal of Psychiatry* **176**: 531–6

Cannon T D, Rosso I M, Hollister J M, Bearden C E, Sanchez L E, Hadley T 2000 A prospective cohort study of genetic and perinatal influences in the etiology of schizophrenia. *Schizophrenia Bulletin* **26**: 351–66

Ciompi L 1980 Catamnestic long-term study on the course of life and aging of schizophrenics. *Schizophrenia Bulletin* **6**: 606–18

Cohen S 1960 Lysergic acid diethylamide: Side effects and complications. *Journal of Nervous and Mental Disease* **130**: 30–40

Dalman C, Allebeck P, Cullberg J, Grunewald C, Koster M 1999 Obstetric complications and the risk of schizophrenia. *Archives of General Psychiatry* **56**: 234–40

Drake R E, Wallach M A 1989 Substance abuse among the chronically mentally ill. *Hospital and Community Psychiatry* **40**: 1041–6

Falloon I R H, Boyd J L, McGill C W, Razani J, Moss H B, Gilderman A M 1982 Family management in the prevention of exacerbations of schizophrenia: A controlled study. *New England Journal of Medicine* **306**: 1437–40

Feinberg I 1983 Schizophrenia: Caused by a fault in programmed synaptic elimination during adolescence? *Journal of Psychiatric Research* **17**: 319–34

Freedman R, Coon H, Myles-Worsley M, Orr-Urtreger A, Olincy A, Davis A, Polymeropoulos M, Holik J, Hopkins J, Hoff M, Rosenthal J, Waldo M C, Reimherr F, Wender P, Yaw J, Young D A, Breese C R, Adams C, Patterson D, Adler L E, Kruglyak L, Leonard S, Byerley W 1997 Linkage of a neurophysiological deficit in schizophrenia to a chromosome 15 locus. *Proceedings of the National Academy of Sciences of the USA* **94**: 587–92

Geddes J R, Lawrie S M 1995 Obstetric complications and schizophrenia. *British Journal of Psychiatry* **167**: 786–93

Gottesman I 1991 *Schizophrenia Genesis: The Origins of Madness*. Freeman, New York

Hambrecht M, Hafner H 1995 Substance abuse or schizophrenia: Which comes first? Presented at the World Psychiatric Association Section of Epidemiology and Community Psychiatry Symposium, New York

Hirsch S, Leff J 1975 *Abnormality in Parents of Schizophrenics*. Oxford University Press, London

Hoffman R E, McGlashan T H 1997 Synaptic elimination, neurodevelopment, and the mechanism of hallucinated 'voices' in schizophrenia. *American Journal of Psychiatry* **154**: 1683–9

Hultman C M, Sparen P, Takei N, Murray R M, Cnattingius S 1999 Prenatal and perinatal risk factors for schizophrenia, affective psychosis, and reactive psychosis of early onset: Case control study. *British Medical Journal* **318**: 421–6

Huttenlocher P R 1979 Synaptic density in the human frontal cortex–developmental changes and effects of aging. *Brain Research* **163**: 195–205

Jablensky A, Sartorius N, Ernberg G, Anker M, Korten A, Cooper J E, Day R, Bertelsen A 1992 Schizophrenia: Manifestations, incidence and course in different cultures: A World Health Organization ten-country study. *Psychological Medicine Monograph Supplement* **20**

Jones P B, Rantakallio P, Hartikainen A-L, Isohanni M, Sipila P 1998 Schizophrenia as a long-term outcome of pregnancy,

delivery, and perinatal complications: A 28-year follow-up of the 1996 north Finland general population birth cohort. *American Journal of Psychiatry* **155**: 355–64
Kavanagh D J 1992 Recent developments in expressed emotion and schizophrenia. *British Journal of Psychiatry* **160**: 601–20
Kendell R E, McInneny K, Juszczak E, Bain M 2000 Obstetric complications and schizophrenia: Two case-control studies based on structured obstetric records. *British Journal of Psychiatry* **176**: 516–22
Laing R D, Esterton A 1970 *Sanity, Madness and the Family: Families of Schizophrenics*. Penguin Books, Baltimore
Leff J, Vaughn C 1985 *Expressed Emotion in Families*. Guilford Press, New York
Lidz T, Fleck S, Cornelison A 1965 *Schizophrenia and the Family*. International Universities Press, New York
MacGlashan T H, Hoffman R E 2000 Schizophrenia as a disorder of developmentally reduced synaptic connectivity. *Archives of General Psychiatry* **57**: 637–48
Malleson N 1971 Acute adverse reactions to LSD in clinical and experimental use in the United Kingdom. *British Journal of Psychiatry* **118**: 229–30
McFarlane W R (ed.) 1983 *Family Therapy in Schizophrenia*. Guilford Press, New York
McNeil T F 1988 Obstetric factors and perinatal injuries. In: Tsuang M T, Simpson J C (eds.) *Handbook of Schizophrenia: Nosology, Epidemiology and Genetics*. Elsevier Science, New York
Meltzer H Y, Stahl S M 1976 The dopamine hypothesis of schizophrenia: A review. *Schizophrenia Bulletin* **2**: 19–76
Mosher L, Burti L 1989 *Community Mental Health: Principles and Practice*. Norton, New York
Parker G, Hadzi-Pavlovic D 1990 Expressed emotion as a predictor of schizophrenic relapse: An analysis of aggregated data. *Psychological Medicine* **20**: 961–5
Rabkin J G 1982 Stress and psychiatric disorders. In: Goldberger L, Breznitz S (eds.) *Handbook of Stress: Theoretical and Clinical Aspects*. Free Press, New York, pp. 566–84
Sherman P S, Porter M A 1991 Mental health consumers as case management aides. *Hospital and Community Psychiatry* **42**: 494–8
Stein L I, Test M A 1980 Alternative to mental hospital treatment: I. Conceptual model, treatment program, and clinical evaluation. *Archives of General Psychiatry* **37**: 392–7
Sturgeon D, Kuipers L, Berkowitz R, Turpin G, Leff J 1981 Psychophysiological responses of schizophrenic patients to high and low expressed emotion relatives. *British Journal of Psychiatry* **138**: 40–5
Suddath R L, Christison G W, Torrey E F, Casanova M F, Weinberger D R 1990 Anatomical abnormalities in the brains of monozygotic twins discordant for schizophrenia. *New England Journal of Medicine* **322**: 789–94
Tamminga C A, Thaker G K, Buchanan R, Kirkpatrick B, Alphs L D, Chase T N, Carpenter W T 1992 Limbic system abnormalities identified in schizophrenia using positron emission tomography with fluorodeoxyglucose and neocortical alterations with deficit syndrome. *Archives of General Psychiatry* **49**: 522–30
US Department of Health and Human Services 1999 *Mental Health: A Report of the Surgeon General*. US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health, Rockville, MD
Wang S, Sun C, Walczak C A, Ziegle J S, Kipps B R, Goldin L R, Diehl S R 1995 Evidence for a susceptibility locus for schizophrenia on chromosome 6pter- p22. *Nature Genetics* **10**: 41–6
Warner R 1994 *Recovery from Schizophrenia: Psychiatry and Political Economy*. Routledge, New York
Warner R (ed.) 1995 *Alternatives to the Hospital for Acute Psychiatric Care*. American Psychiatric Press, Washington, DC
Warner R 2000 *The Environment of Schizophrenia: Innovations in Practice, Policy and Communications*. Routledge, London
Warner R, de Girolamo G 1995 *Epidemiology of Mental Problems and Psychosocial Problems: Schizophrenia*. World Health Organization, Geneva, Switzerland
Weinberger D R, Berman K F, Suddath R, Torrey E F 1992 Evidence of a dysfunction of a prefrontal-limbic network in schizophrenia: A magnetic resonance imaging and regional cerebral blood flow study of discordant monozygotic twins. *American Journal of Psychiatry* **149**: 890–7
World Health Organization 1979 *Schizophrenia: An International Follow-up Study*. Wiley, Chichester, UK
Zisook S, Heaton R, Moranville J, Kuck J, Jernigan T, Braff D 1992 Past substance abuse and clinical course of schizophrenia. *American Journal of Psychiatry* **149**: 552–3
Zornberg G L, Buka S L, Tsuang M T 2000 Hypoxia-ischemia-related fetal/neonatal complications and risk of schizophrenia and other nonaffective psychoses: A 19-year longitudinal study. *American Journal of Psychiatry* **157**: 196–202

R. Warner

Schizophrenia and Bipolar Disorder: Genetic Aspects

Schizophrenia and bipolar mood disorder (the latter sometimes called 'manic-depressive illness,') are among the most serious of all psychiatric disorders, indeed of all medical disorders. Both of these psychiatric illnesses tend to have a rather early age of onset, with most patients first becoming ill in their teens or twenties, and the symptoms are often chronic, particularly in the case of schizophrenia. Moreover, these illnesses are often severely disabling, and are associated with increased rates of educational problems, unemployment, marital difficulties, alcohol or substance abuse, and suicide. Schizophrenia and bipolar disorder each affect approximately one percent of the population in the USA and Western Europe. If one takes into account not only the numbers of people affected by schizophrenia or bipolar disorder but also the fact that many patients are seriously disabled for much of their adult lives, then the cost of these two disorders, in both economic and human terms, rivals that of diseases such as heart disease and stroke, which affect more people but tend not to strike until much later in life.

This aticle first considers diagnostic issues, including the most widely used diagnostic criteria for schizophrenia, bipolar disorder, and related disorders. Evi-

dence for genetic factors in schizophrenia and bipolar disorder is reviewed next, including some complicating factors, such as the likely presence of etiologic heterogeneity and the interaction of genes with environmental stressors. Issues of genetic counseling are then considered. The chapter concludes with a discussion of evidence that genes for schizophrenia and bipolar disorder may have a 'silver lining,' in terms of increased creative potential.

1. Diagnostic Issues

1.1 Schizophrenia and Related Disorders

The criteria currently most widely used in diagnosing schizophrenia, bipolar disorder, and related disorders are those described in the most recent (4th) edition of the *Diagnostic and Statistical Manual* (DSM-IV) of the American Psychiatric Association (1994). Briefly summarized, the diagnostic criteria for schizophrenia listed in DSM-IV require two or more of the following five characteristic symptoms: hallucinations; delusions; disorganized speech; grossly disorganized behavior; and negative symptoms, such as flattened affect. In addition, the patient must have shown significant deterioration of functioning in areas such as occupational or interpersonal relations, and have had continuous signs of illness for at least six months, including at least one month with the characteristic symptoms.

DSM-IV also recognizes several related disorders that manifest milder forms of certain symptoms that are often seen in schizophrenia. *Schizotypal* personality disorder, for example, is used for persons who have several characteristic features (e.g., magical thinking, recurrent illusions, and peculiar behavior or speech). If a person does not have such schizotypal eccentricities but *has* shown several signs of marked disinterest in interpersonal relationships (e.g., extreme aloofness, absence of any close friends), then a diagnosis of *schizoid* personality disorder is given. In *paranoid* personality disorder, there is a broad pattern of extreme suspiciousness, as shown by several signs (e.g., irrational fears that others wish to harm one). Family and adoption studies suggest that these personality disorders are part of a 'spectrum' of disorders that are genetically related to schizophrenia proper.

1.2 Bipolar Disorder and Related Disorders

In DSM-IV, the diagnosis of 'bipolar I' disorder requires a history of at least one episode of mania, which is defined as a period in which a person's mood is 'abnormally and persistently' expansive, elevated, or irritable. In addition, diagnosis of bipolar disorder requires that the mood change either involve psychotic features which last at least a week, or lead to hospitalization. The disturbance in mood must be severe enough to disrupt seriously social or occupational functioning, and/or to require hospitalization. The manic episode must also involve at least three (four, if only an *irritable* mood is present) of seven symptoms: (a) greatly inflated self-esteem, (b) increased activity or restlessness, (c) unusual talkativeness, (d) racing thoughts or flight of ideas, (e) decreased need for sleep, (f) distractibility, and (g) extremely risky actions (such as buying sprees or reckless driving) whose potentially dangerous consequences are not appreciated. The criteria for a hypomanic episode are essentially the same as those for a manic one, except that the symptoms are neither psychotic nor so severe that they severely impair social functioning or require hospitalization. The term bipolar is potentially confusing, because it does not require a history of major depression, even though most persons with bipolar disorder will also have experienced episodes of major depression. (By contrast, in major depressive disorder, a person has experienced an episode of major depression, but *not* an episode of mania.)

As in the case of schizophrenia, there appears to be a 'spectrum' of affective disorders that are genetically related to bipolar disorder but that have symptoms which are milder than those found in frank bipolar disorder. Thus DSM-IV also includes (a) bipolar II disorder (in which a patient has experienced a *hypomanic*, rather than a manic episode, as well as an episode of major depression) and (b) cyclothymic disorder, which involves a history of multiple hypomanic episodes as well as multiple periods of depressive symptoms (but not major depression). Bipolar disorder is also often accompanied by other concurrent, or 'co-morbid,' disorders, such as anxiety disorders, or alcohol and substance abuse. Moreover, even when symptoms of mania and depression are in remission, patients with a history of bipolar disorder often still meet criteria for personality disorders, particularly those with narcissistic or histrionic features.

1.3 Differential Diagnosis

A diagnosis of schizophrenia requires that diagnoses of mood disorders and schizoaffective disorders be excluded. Diagnoses of schizophrenia, mood disorders, and schizoaffective disorders all require the exclusion of neurologic syndromes caused by factors such as substance abuse, medications, or general medical condition. Until rather recently there has been a tendency, particularly in the USA, for mania with psychotic features to be misdiagnosed as schizophrenia. Accurate differential diagnosis is crucial, because misdiagnosis (and resultant inappropriate treatment) may result, on the one hand, in patients being needlessly exposed to harmful side effects of medication or, on the other hand, being deprived of

appropriate medications that may alleviate unnecessary suffering and save patients' jobs, marriages, even their lives. Mania, for example, often responds well to medications (particularly lithium and certain anticonvulsants such as carbamazepine and valproate) that are less effective in treating schizophrenia. It is important, moreover, to diagnose and treat schizophrenia and bipolar disorder as early as possible in the course of the illness because there is increasing evidence that these illnesses (and their underlying brain pathologies) tend to worsen if left untreated (e.g., Wyatt 1991).

2. *Evidence for Genetic Factors in Schizophrenia*

There are several complementary lines of evidence for an important role of genetic factors in the etiology of schizophrenia. First, there are converging lines of evidence from family, twin, and adoption studies that the risk for schizophrenia is greatly increased among schizophrenics' biological relatives. Second, investigators have, in recent years, increasingly marshalled techniques from molecular genetics to look for more direct evidence of genetic factors in schizophrenia.

2.1 *Family, Twin, and Adoption Studies*

A person's risk of developing schizophrenia increases, on average, with his or her increasing degree of genetic relatedness to a schizophrenic patient (e.g., Matthysse and Kidd 1976, Holzman and Matthysse 1990, Torrey et al. 1994). The risk of developing schizophrenia over a person's lifetime is about 0.8 percent for people in the general population, though there is a several-fold variation in the prevalence of schizophrenia across different populations that have been studied around the world (e.g., see Torrey 1987). By contrast, a person's lifetime risk is 5–10 percent if the person has a first-degree relative with schizophrenia, and is much higher—nearly 50 percent in some studies—if a person is the monozygotic (genetically identical) twin of a schizophrenic patient. While these risk figures are consistent with genetic transmission, they are not conclusive, because the degree of genetic resemblance among relatives tends to parallel their level of exposure to similar environments.

Twin studies have rather consistently reported concordance rates for schizophrenia in monozygotic (MZ) twins that are several times higher than those for dizygotic (DZ) twins (e.g., Gottesman et al. 1987, Torrey et al. 1994). It is also noteworthy that the schizophrenia concordance rate for MZ twins reared apart is quite close to that for MZ twins reared together. On the other hand, the number of such MZ twins reared apart is rather small. Moreover, it is unclear whether these twins who were reared in different settings may still have had significant contact with each other after they were separated, so that the separation of shared genetic and environmental factors may not have been complete.

The most conclusive available evidence for genetic factors in schizophrenia, therefore, comes from adoption studies. For example, Heston (1966) studied 47 adult adoptees who had been born in the USA to a schizophrenic mother, but separated shortly after birth. Five of these 'index' adoptees were found to have subsequently developed a diagnosis of schizophrenia, vs. none of 50 matched control adoptees who had been born to demographically matched, but psychiatrically healthy, mothers.

More systematic adoption studies of schizophrenia have been carried out in Scandinavia. In Denmark, for example, Kety et al. (1994) were able to identify all individuals in the entire country who had been adopted away from their biological parents at an early age and subsequently were hospitalized with a diagnosis of schizophrenia. For each of these 74 schizophrenic adoptees, a 'control' adoptee was identified who was closely matched for age, gender, and the socioeconomic status of the adoptive home. The control adoptees' biological parents had not been hospitalized for mental illness. Psychiatric diagnoses of over 1100 of the adoptees' respective biological and adoptive relatives were made after careful review of psychiatric interviews and records. Significantly higher rates of schizophrenia were found in the biological (but *not* in the adoptive) relatives of schizophrenic adoptees than in the biological relatives of control adoptees (5.0 percent vs. 0.4 percent). The prevalence of schizotypal personality disorder was also significantly elevated among the schizophrenic adoptees' biological relatives (Kendler et al. 1994). Moreover, rates of schizophrenia and related 'spectrum' disorders were significantly elevated even in the schizophrenic adoptees' biological *paternal* half-siblings, who had not even shared the same womb as the schizophrenic adoptees.

2.2 *Association and Linkage Studies*

In recent decades, advances in molecular genetics have enabled researchers to look more directly for genetic factors in schizophrenia. One strategy is to investigate 'candidate genes' for which there is a theoretical reason to suspect a role in schizophrenia. Thus several groups of investigators have looked for genes that influence susceptibility to certain infectious agents, since exposure to these agents, particularly during pre- or perinatal development, appears to increase risk for schizophrenia. For example, in a number of epidemiological studies, increased risk of exposure to influenza during the middle trimester of gestation has been found to be associated with increased risk of schizophrenic outcome. Individuals' genotypes can powerfully affect their immunologic response to infections (and their mothers' ability to combat infections while they are still *in utero*). Several studies have found that certain alleles of genes that play an important role

in immune function, such as those in the HLA complex, are more prevalent in schizophrenia. McGuffin (1989) reviewed nine such studies and concluded that there was a highly significant association between the HLA A9 allele and risk for the paranoid subtype of schizophrenia. Murray et al. (1993) suggested that prenatal exposure to influenza may increase risk of schizophrenia in the offspring because, in genetically susceptible mothers, the flu virus stimulates production of maternal antibodies that cross the placenta and disrupt fetal brain development.

Even if one does not have a good 'candidate' gene for a disorder such as schizophrenia, however, it is still possible to apply the more indirect strategy of genetic 'linkage.' This strategy makes use of the fact that genes that are located very near one another on the same chromosome (and are therefore said to be closely 'linked') will tend to be inherited together. As the result of recent advances in molecular genetics, there are now thousands of identified genes that can be used to mark various regions of different chromosomes. One can then study large numbers of 'multiplex' families in which there are at least two family members with schizophrenia, in order to examine whether there is a significant tendency for schizophrenia and alleles for particular marker genes to be transmitted together within the same family. In principle, genetic linkage studies provide an elegant approach that makes it possible to identify disease genes whose role in a disease such as schizophrenia is a complete surprise to investigators (and thus would never have been chosen as 'candidate' genes for association studies).

There are, however, some practical difficulties with linkage studies. One difficulty is that, because hundreds of different marker genes can be examined, there is a rather high probability of obtaining a spurious, or 'false positive,' linkage finding by chance. In order to screen out such false positive findings, it is important to determine whether interesting linkage results can be confirmed in new, independent, studies involving large numbers of multiplex families. In fact, linkage findings implicating genes on a particular region of chromosome 6 as risk factors for schizophrenia have recently been reported by several different research groups (e.g., see review by Gershon et al. 1998).

2.3 Etiologic Heterogeneity and Genotype–Environment Interactions

The search for genetic factors is complicated by several factors. It is likely, for example, that disorders such as schizophrenia are etiologically heterogeneous. That is, it is probable that the syndrome of schizophrenia can be produced by a number of different combinations of genes and/or environmental factors. Thus, while each of a number of different genes may well increase risk for schizophrenia, it is likely that no single gene is *necessary* for the production of most cases. One approach to the problem of heterogeneity is to identify characteristics that distinguish specific, genetically more homogeneous, subtypes of schizophrenia or bipolar disorder. Maziade et al. (1994), for example, found evidence for linkage at a locus on chromosome 11 with schizophrenia in one, but not the others, of several large pedigrees that they examined. The schizophrenics in the extended family that *did* show linkage were distinguished from the families that did not by having a particularly severe and unremitting form of schizophrenia. If this finding can be confirmed in other pedigrees, it would provide a valuable example of the subtyping strategy.

A related problem is that, while a particular gene or genes may significantly increase one's risk for developing schizophrenia or bipolar disorder, it will usually not be *sufficient* to produce the disorder; that is, most individuals who carry a susceptibility gene will not themselves become ill. (That is, there is 'incomplete penetrance' of the gene's effects.) One strategy for dealing with this latter problem is to identify more sensitive, subclinical, phenotypes that indicate the presence of the gene even in people who do not develop the illness. For example, studies of schizophrenics' families suggest that most schizophrenics carry a gene which leads to schizophrenia only 5–10 percent of the time, but causes abnormal smooth pursuit *eye movements* over 70 percent of the time. These eye movement dysfunctions should thus provide a better target for genetic linkage studies than schizophrenia itself (Holzman and Matthysse 1990).

A further complication is the likelihood of genotype–environment *interactions*. For example, there is evidence from dozens of studies that pre- and perinatal complications are significant risk factors for schizophrenia (e.g., Torrey et al. 1987). Such complications are, of course, hardly unique to schizophrenia; this suggests that pre- or perinatal insults to the developing brain may interact with genetic liability factors to produce schizophrenia. Kinney et al. (1999), for example, found that schizophrenics were much more likely than were either control subjects or the schizophrenics' own non-schizophrenic siblings to have both a major perinated complication *and* eye tracking dysfunction. Moreover, these non-schizophrenic siblings tended to have *either* a history of perinatal complications *or* eye tracking dysfunction, but not *both* in the same sibling. This pattern of findings was consistent with a two-factor model in which perinatal brain injury and specific susceptibility genes often interact to produce schizophrenia.

3. Evidence for Genetic Factors in Bipolar Disorder

As in the case of schizophrenia, several converging lines of evidence strongly implicate genetic factors in the etiology of bipolar disorder.

3.1 Family, Twin, and Adoption Studies

There is a strong tendency for bipolar disorder to run in families, and the risk of bipolar disorder in a first-degree relative of a manic-depressive is about 8 percent (vs. about 1 percent in the general population). Further evidence for a high heritability of bipolar disorder is provided by twin studies, particularly three twin studies conducted in Scandinavia in recent decades. The average concordance rate for bipolar disorder in these latter studies was 55 percent in MZ vs. only 5 percent in DZ twin pairs (see review by Vehmanen et al. 1995).

Complementary evidence for genetic factors in bipolar disorder is provided by *adoption* studies. For example, Mendlewicz and Ranier (1977) identified 29 adoptees with bipolar disorder, along with demographically matched controls who were either (a) psychiatrically normal or (b) had had polio during childhood. When the biological and adoptive parents of these adoptees were interviewed, significantly more cases of bipolar disorder, major depression, and schizoaffective disorder were found in the biological parents of bipolar adoptees than in the biological parents of either of the two control groups. The respective groups of adoptive parents, by contrast, did not differ significantly in the prevalence of these disorders.

3.2 Association and Linkage Studies

Although there have been dozens of reports of *linkage* between bipolar disorder and genetic loci on various chromosomes, few of these reports have subsequently been confirmed. Among these few are linkages to particular regions of chromosomes X, 18 and 21; each of these linkages has been confirmed by several independent groups. While this suggests that genes in these regions significantly influence susceptibility for bipolar disorder, the effects of these genes on susceptibility may be modest in size (for reviews, see Gershon et al. 1998 and Pekkarinen 1998).

A key difficulty in identifying genes in bipolar disorder is the likely presence of *etiologic heterogeneity*. One approach to overcoming this challenge is to search for markers of genetic subtypes of bipolar disorder. For example, MacKinnon et al. (1998), after identifying a strong tendency for a *subtype* of bipolar disorder with concomitant panic disorder to run in families, found strong evidence for linkage of the panic-disorder subtype to marker genes on a region of chromosome 18. There was no evidence of such linkage for bipolar disorder *without* panic disorder.

Other research has found that manic-depressive patients who respond well to lithium treatment represent a subtype that is etiologically more homogeneous, and has a stronger familial tendency, than non-lithium responders. Both linkage and association studies suggest that susceptibility to this lithium-responsive subtype is increased by certain alleles of the gene for phospholipidase C, an enzyme important in the phosphoinositol cycle that is thought to be a therapeutic target of lithium (Alda 1999).

4. Genetic Counseling

A recent study (Trippitelli et al. 1998) found that most patients with bipolar disorder and their unaffected spouses would be interested in receiving counseling about their own genetic risk, and that of their children. The majority of patients and spouses, for example, would take advantage of a test for susceptibility genes.

Even when precise gene(s) involved in a particular case of schizophrenia or bipolar disorder are unknown, genetic counselors can provide a patient's relatives with an estimated risk of developing the disorder, based on average figures from many different family studies. These risk figures, it should be noted, refer to a person's *lifetime* risk—an important distinction (e.g., for *schizophrenia*, one's risk has been cut roughly in half by age 30, and by age 50 it is extremely small). It is crucial that counseling be based on accurate differential diagnosis. Problems often encountered in counseling, such as counselees being confused by genetic information, or feeling fearful and embarrassed, may be heightened in families with bipolar disorder and schizophrenia, because these psychiatric disorders often carry a social stigma, and because many parents have (unfairly) been blamed for their child's illness. For these reasons, Kessler (1980) emphasized the importance of (a) careful follow-up, to make sure that counseling was understood, and (b) the inclusion of professionals with good psychotherapeutic skills as part of the counseling team.

5. Creativity and Liability for Schizophrenia and Bipolar Disorder

There is increasing evidence, from converging lines of research, for the idea that genetic liability for schizophrenia and bipolar disorder is associated with unusual creative potential. This idea, which has long been the subject for theoretical speculation, has received empirical support from several complementary types of studies. For example, a number of studies involving *non*-clinical samples have reported that more creative subjects tend to score higher on personality test variables that are associated with liability for schizophrenia or bipolar disorder.

5.1 Creativity in Schizophrenics' Biological Relatives

Of even greater interest are studies that have found unusual creativity in samples of the healthier bio-

logical relatives of diagnosed schizophrenics. In an Icelandic sample, for example, Karlsson (1970) found that the biological relatives of schizophrenics were significantly more likely than people in the general population to be recognized in *Who's Who* for their work in creative professions. In the adoption study noted earlier, Heston (1966) serendipitously discovered that, among the psychologically healthier 'index' adoptees (i.e., who had a biological mother with schizophrenia), there was a subgroup of psychologically healthy individuals who had more creative jobs and hobbies than the control adoptees.

Using a similar research design, Kinney et al. (2000) studied the adopted-away offspring of schizophrenic and control biological parents. The creativity of these adoptees' actual vocational and avocational activities was rated by investigators who were blind to the adoptees' personal and family histories of psychopathology. Real-life creativity was rated as significantly higher, on average, for subjects who, while not schizophrenic, did have signs of magical thinking, recurrent illusions or odd speech.

5.2 Creativity in Patients with Bipolar Disorder and Their Relatives

There is also evidence for increased creativity among subjects with major *mood* disorders and their biological relatives. Studies of eminent writers, artists, and composers carried out in the USA, the UK, and France, all found significantly higher rates of major mood disorders among these creators than among the general population (e.g., see Jamison 1990). Richards et al. (1988) extended this link by showing that measures of 'everyday,' or non-eminent, creativity were significantly higher, on average, in manic-depressive and cyclothymic subjects and their normal relatives than in control subjects who did not have a personal or family history of mood disorders.

Moreover, both creative artists and patients with mood disorders report that their creativity is significantly enhanced during periods of moderately elevated mood (e.g., Richards and Kinney 1990). These complementary findings suggest that the association between increased creative potential and genetic liability for bipolar disorder may extend not only to the millions of people with bipolar disorder, but also to tens of millions of others who, while not ill themselves, may carry genes for the disorder.

5.3 Implications of Link between Creativity and Genes for Schizophrenia and Bipolar Disorder

It is important to determine what maintains the high prevalence of genes for bipolar disorder in the population, despite the high rates of illness and death that are associated with this disorder. One interesting possibility is that genes which increase liability for bipolar disorder may also be associated with personally and socially beneficial effects, such as increased drive and creativity. The research findings suggesting an association between genes for bipolar disorder and increased creativity are also potentially of great significance in terms of how patients and their families view greater liability for bipolar disorder, as well as for combatting the social stigma that is still often attached to the disorder.

Parallel considerations apply in the case of *schizophrenia*, but perhaps with even *greater* force, because schizophrenia tends to be an even more chronic and disabling disease, have even lower fertility, and carry an even greater social stigma. As rapid advances in molecular biology and discovery of genetic markers make it possible to detect major genes for schizophrenia (and to identify individuals who carry these genes), it will become increasingly important to know whether such genes are associated with *positive*, as well as negative, behavioral phenotypes or outcomes—and to understand what genetic and/or environmental modifiers affect how these genes are expressed.

See also: Behavioral Genetics: Psychological Perspectives; Bipolar Disorder (Including Hypomania and Mania); Depression; Genetic Screening for Disease-related Characteristics; Genetics and Development; Genetics of Complex Traits Through the Life Cycle; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Mental Illness, Etiology of; Mental Illness, Genetics of; Personality Disorders; Schizophrenia, Treatment of

Bibliography

Alda A 1999 Pharmacogenetics of lithium response in bipolar disorder. *J. Psychiatry Neurosci.* **24**(2): 154–8

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV), 4th edn. American Psychiatric Association, Washington, DC

Gershon E S, Badner J A, Goldin L R, Sanders A R, Cravchik A, Detera-Wadleigh S D 1998 Closing in on genes for manic-depressive illness and schizophrenia. *Neuropsychopharmacology* **18**: 233–42

Gottesman I I, McGuffin P, Farmer A E 1987 Clinical genetics as clues to the 'real' genetics of schizophrenia. *Schizophrenia Bulletin* **13**: 23–48

Heston L L 1966 Psychiatric disorders in foster home reared children of schizophrenic mothers. *British Journal of Psychiatry* **112**: 819–25

Holzman P S, Matthysse S 1990 The genetics of schizophrenia: A review. *Psychological Science* **1**(5): 279–86

Jamison K R 1990 Manic-depressive illness and accomplishment: Creativity, leadership, and social class. In: Goodwin F, Jamison K R (eds.) *Manic-depressive Illness*. Oxford University Press, Oxford, UK

Karlsson J L 1970 Genetic association of giftedness and creativity with schizophrenia. *Hereditas* **66**: 177–81

Kendler K S, Gruenberg A M, Kinney D K 1994 Independent diagnoses of adoptees and relatives, using DSM-III criteria, in

the provincial and national samples of the Danish adoption study of schizophrenia. *Archives of General Psychiatry* **51**: 456–68
Kessler S 1980 The genetics of schizophrenia: a review. *Schizophrenia Bulletin* **6**: 404–16
Kety S S, Wender P H, Jacobsen B, Ingraham L J, Jansson L, Faber B, Kinney D K 1994 Mental illness in the biological and adoptive relatives of schizophrenic adoptees. Replication of the Copenhagen study in the rest of Denmark. *Archives of General Psychiatry* **51**: 442–55
Kinney D K, Richards R L, Lowing P A, LeBlanc D, Zimbalist M A, Harlan P 2000 Creativity in offspring of schizophrenics and controls: An adoption study. *Journal of Creativity Research* **13**(1): 17–25
Matthysse S, Kidd K K 1976 Estimating the genetic contribution to schizophrenia. *American Journal of Psychiatry* **133**: 185–91
Maziade M, Martinez M, Cliche D, Fournier J P, Garneau Y, Merette C 1994 Linkage on the 11q21–22 region in a severe form of schizophrenia. American Psychiatric Association Annual Meeting, New Research Programs and Abstracts, p. 97
McGuffin P 1989 Genetic markers: an overview and future perspectives. In: Smeraldi E, Belloni L (eds.) *A Genetic Perspective for Schizophrenic and Related Disorders*. Edi-Ermes, Milan
Mendlewicz J, Ranier J D 1977 Adoption study supporting genetic transmission of manic-depressive illness. *Journal of the American Medical Association* **222**: 1624–7
Murray R M, Takei N, Sham P, O'Callaghan E, Wright P 1993 Prenatal influenza, genetic susceptibility and schizophrenia. *Schizophrenia Research* **9**(2,3): 137
Pekkarinen P 1998 Genetics of bipolar disorder. *Psychiatria Fennica* **29**: 89–109
Richards R L, Kinney D K 1990 Mood swings and everyday creativity. *Creativity Research Journal* **3**: 202–17
Torrey E F 1987 Prevalence studies in schizophrenia. *British Journal of Psychiatry* **150**: 598–608
Torrey E F, Bowler A E, Taylor E H, Gottesman, I I 1994 *Schizophrenia and Manic-depressive Disorder*. Basic Books, New York
Trippitelli C L, Jamison K R, Folstein M R, Bartko J J, DePaulo J P 1998 Pilot study on patients' and spouses' attitudes toward potential genetic testing for bipolar disorder. *American Journal of Psychiatry* **155**(7): 899–904
Vehmanen L, Katrio J, Lonnqvist J 1995 Twin studies on concordance for bipolar disorder. *Clinical Psychiatry and Psychopathology* **26**: 107–16
Wyatt R J 1991 Neuroleptics and the natural course of schizophrenia. *Schizophrenia Bulletin* **17**: 325–51
Mackinnon D F, Xu J, McMahon F J, Simpson S G, Stine O C, McInnis M G, De Paulo J R 1998 Bipolar disorder and panic disorder in families: an analysis of chromosome 18 data. *American Journal of Psychiatry* **155**(6): 829–31
Kinney D K, Yurgelun-Todd D A, Tramer S J, Holzman P S 1998 Inverse relationship of perinatal complications with eyetracking dysfunction in relatives of patients with schizophrenia: evidence for a two-factor model. *American Journal of Psychiatry* **155**(7): 976–8
Richards R, Kinney D K, Lunde I, Benet M, Merzel A P C 1988 Creativity in manic-depressives, cyclothymes, their normal relatives and control subjects. *Journal of Abnormal Psychology*. **97**(3): 1–8

D. K. Kinney

Schizophrenia: Neuroscience Perspective

Schizophrenia is a complex mental illness characterized by acute phases of delusions, hallucinations, and thought disorder, and chronically by apathy, flat affect, and social withdrawal. Schizophrenia affects 1 percent of the world's population, independent of country or culture, and constitutes a severe public health issue (WHO 1975). Schizophrenic patients normally start to display symptoms in their late teens to early twenties, but the time of onset and the course of the illness are very variable (American Psychiatry Association 1987). Approximately a third of patients experience one acute episode after which they make a more or less full recovery. Another third are affected by the illness throughout their lives, but their symptoms are to some extent alleviated by anti-psychotic drugs. The remaining third are so chronically ill that they show little or no improvement, even with medication (Johnstone 1991).

1. *Intellectual Function*

There is a striking deterioration in intellectual and cognitive function in schizophrenia (e.g., Johnstone 1991). Patients have difficulty in initiating and completing everyday tasks, being distracted easily and tending to give up when confronted by any obstacles. These deficits are similar to the problems in initiation and planning associated with frontal-lobe lesions (Shallice and Burgess 1991). This has led many researchers to suggest that the core deficit of schizophrenia is a failure to activate frontal cortex appropriately during cognitive tasks involving planning and decision making ('task-related hypofrontality'), a notion supported in part by many functional imaging studies (see below). Schizophrenic patients also show deficits in attention and memory tasks that engage prefrontal, hippocampal, and medial temporal systems, even in drug-free populations (Saykin et al. 1994). In the rest of this entry we shall consider the direct evidence for brain abnormalities in schizophrenia.

2. *Post-mortem Neuropathology*

Studies of post-mortem brains (Harrison 1999) have observed a decrease in overall brain size and have linked schizophrenia to structural abnormalities in the prefrontal cortex and temporal lobe, especially the hippocampus and amygdala. Several studies have found that schizophrenic brains tend to have enlarged lateral ventricles compared to nonschizophrenic brains. Histological studies have shown evidence for

abnormal synaptic appearance in the cingulate and hippocampal pyramidal cells in schizophrenia. The most reproducible positive anatomical finding in post-mortem hippocampal formation has been the reduced size of neuronal cell bodies in schizophrenia. However, these changes have not been found in all studies.

3. *In Vivo Studies Of Brain Structure*

3.1 *Computerized Tomography (CT) Studies*

The main finding from CT scan studies is that the lateral ventricles are enlarged in schizophrenic patients compared to normal controls (Van Horn and McManus 1992). Although the finding of increased ventricular volume in schizophrenic patients is widespread and replicated, the difference between schizophrenic patients and normal controls is small. Overlap with the normal population is appreciable. Correlation between CT findings and symptoms has been investigated (Lewis 1990). Enlarged ventricles have been associated with chronicity of illness, poor treatment response, and neuropsychological impairment in many, but not all of these studies.

3.2 *Magnetic Resonance Imaging (MRI) Studies*

MRI studies (see Harrison 1999) have tended to confirm the finding of enlarged ventricles in schizophrenic patients, but also permit a more detailed analysis of brain structure. Temporal lobe reductions have been reported, and are especially prominent in the hippocampus, parahippocampal gyrus, and the amygdala, but have not been observed in all studies. Robust relationships between temporal lobe reductions and clinical features have not yet been found. Some studies have found frontal lobe reductions in schizophrenic patients.

Reversal or reduction of normal structural cerebral asymmetries may be related to the pathogenesis of schizophrenia (Crow 1995). Various unusual symmetries have been observed consistent with the hypothesis that failure to develop normal asymmetry is an important component of the pathology underlying some forms of schizophrenia.

There are some MRI data that provide support for a hypothesis of disconnection between brain areas in schizophrenia. These results support the existence of a relative 'fronto-temporal dissociation' in schizophrenia. Evidence for such dissociation has also been obtained in functional imaging studies (see below).

All these studies used a 'regions of interest' approach in which measurements were restricted to prespecified brain regions. More recently, techniques have been developed in which differences can be detected automatically throughout the brain. Using such techniques, Andreasen et al. (1994) observed decreased thalamus size in schizophrenic patients consistent with observations in post-mortem brains.

4. *Functional Imaging*

4.1 *Resting Studies Using Positron Emission Tomography (PET)*

More sensitive measures of brain integrity can be obtained by measuring cerebral blood flow in a patient at rest. However, these results are difficult to interpret because the pattern of blood flow may be altered by the current mental state of the patient and by medication. Early studies of this sort observed a relative reduction of blood flow in the frontal lobes of patients with schizophrenia (Ingvar and Franzen 1974). This pattern of activity became known as hypofrontality. However, subsequent studies have not always replicated this observation.

Several studies have looked at clinical correlates associated with hypofrontality, but the results are inconsistent. Among features showing a positive relationship with hypofrontality are chronicity, negative symptoms, and neuropsychological task impairment. Relations have also been observed between the pattern of blood flow and the symptomatology of patients at the time of scanning. For example, patients manifesting 'psychomotor poverty' showed reduced blood flow in dorsolateral prefrontal cortex (Liddle et al. 1992).

4.2 *The Dopamine Hypothesis*

One of the most robust findings in schizophrenia research has been the observation that drugs which block dopamine receptors are effective in reducing the severity of symptoms such as hallucinations and delusions (Seeman 1986). This led to the dopamine (DA) hypothesis of schizophrenia, which posits that schizophrenia is caused by an overactivity of dopamine receptors (Van Rossum 1966). The best way to investigate the dopamine hypothesis is the *in vivo* visualization of radioactive ligand binding to quantify dopamine receptor densities in drug-naive patients using PET.

These studies suggest that compared to healthy controls, patients with schizophrenia show a significant but mild increase in, and a larger variability of, D2 receptor density (Laruelle 1998). There is also evidence that D1 receptor density is reduced in the prefrontal cortex of schizophrenic patients (Okubo et al. 1997).

5. *Cognitive Activation Studies*

Functional neuroimaging experiments generally evaluate brain activity associated with performance of cognitive or sensori-motor tasks. Cognitive activation studies have provided further evidence for decreased frontal activity ('hypofrontality') in schizophrenia. In addition, there is increasing evidence that schizophrenics show abnormal integration between the

frontal cortex and other brain regions, including the temporal lobes, the parietal lobes, and hippocampus, during cognitive tasks.

5.1 Task-based Studies of Executive Function

Schizophrenia is characterized largely by impairments in planning and execution, and therefore tasks that involve this kind of planning and modification of behavior have been exploited in the scanner. Several studies have found that schizophrenic patients show reduced activity in the dorsolateral prefrontal cortex (DLPFC) while performing the Wisconsin card sorting task, a popular test of planning. In addition to decreased DLPFC activity, schizophrenic patients also showed abnormal responses in the temporal lobes and parahippocampal gyrus (Ragland et al. 1998). The results suggest that schizophrenia may involve a breakdown in the integration between the frontal and temporal cortex, which is necessary for executive and planning demands in healthy individuals. This interpretation moves away from the simple notion that dysfunction in isolated brain regions, explains the cognitive deficits in schizophrenia, and towards the idea that neural abnormality in schizophrenia reflects a disruption of integration between brain areas.

5.2 Willed Action

Willed actions are self-generated in the sense that the subject makes a deliberate and free choice to perform one action rather than another. Willed actions are a fundamental component of executive tasks. In normal subjects, willed actions are associated with increased blood flow in the DLPFC. Schizophrenic patients, especially those with negative signs, have difficulty with tasks involving free choices, and show an associated lack of activity in DLPFC. Activity in this region normalizes as the symptoms decrease (Spence et al. 1998). This suggests that hypofrontality depends on current symptoms. Studies of willed action also suggest that the underactivity in DLPFC observed in some schizophrenic patients is accompanied by overactivity in posterior brain regions. There is evidence of a lack of the normal reciprocal interaction between the frontal and the superior temporal cortex in schizophrenia, which supports the notion of impaired functional integration (McGuire and Frith 1996).

5.3 Memory Tasks

Memory impairments are an especially enduring feature of schizophrenia. Functional neuroimaging studies of memory have demonstrated hypofrontality, abnormal interaction between temporal and frontal cortex, and a dysfunctional cortico-cerebellar circuit in schizophrenic patients compared to control subjects. The hippocampus is a brain structure that is well known to be involved in memory. Evidence for impaired hippocampal function in schizophrenia was found in a well-controlled functional imaging study (Heckers et al. 1998). In this study, schizophrenic patients failed to recruit the hippocampus during successful retrieval, unlike normal control subjects. The schizophrenic patients also showed a more widespread activation of prefrontal areas and parietal cortex during recollection than did controls. The authors propose that this overactivation represents an 'effort to compensate for the failed recruitment of the hippocampus.' This result supports the idea that neural abnormality in schizophrenia reflects a disruption of integration between brain areas. Fletcher et al. (1999) also found evidence for abnormal integration between brain areas in schizophrenia during the performance of a memory task. They demonstrated an abnormality in the way in which left prefrontal cortex influenced activity in left superior temporal cortex, and suggested that this abnormality was due to a failure of the anterior cingulate cortex to modulate the prefronto-temporal relationship.

6. Imaging Symptoms

Functional neuroimaging is also useful for evaluating neural activity in patients experiencing specific psychotic symptoms, such as hallucinations and passivity.

6.1 Hallucinations

Hallucinations, perceptions in the absence of external stimuli, are prominent among symptoms of schizophrenia. Functional neuroimaging studies of auditory hallucinations suggest that they involve neural systems dedicated to auditory speech processing as well as a distributed network of other cortical and subcortical areas (Dierks et al. 1999). There is also evidence that the activity associated with auditory hallucinations resembles that seen when normal subjects are using inner speech (McGuire et al. 1996).

6.2 Passivity

Passivity symptoms, or delusions of control, in which patients claim that their actions and speech are being controlled by an external agent, are common in schizophrenia. Schizophrenic patients with passivity showed hyperactivation of inferior parietal lobe (BA 40), the cerebellum, and the cingulate cortex relative to schizophrenic patients without passivity and to normal controls. When patients no longer experienced passivity symptoms, a reversal of the hyperactivation of parietal lobe and cingulate was seen (Spence et al. 1997). Hyperactivity in parietal cortex may reflect the 'unexpected' nature of the

movement experienced by patients. The movement feels as though it is being caused by an external force (Frith et al. 2000).

7. Conclusions

There is considerable evidence for structural abnormalities in the brains of patients with schizophrenia, but the abnormalities identified so far are not specific to this disorder, are very variable and cannot easily be related to the symptoms. The neurotransmitter dopamine is clearly important in schizophrenia, but its precise role remains unclear. Studies of brain function are still at an early stage, but suggest that schizophrenia may be characterized by disorders of connectivity between cortical and subcortical regions. Some of the symptoms of schizophrenia can be understood in terms of these disconnections. Current advances in imaging techniques aimed at measuring connectivity in the brain are likely to have a major impact on our understanding of schizophrenia.

See also: Mental Illness, Epidemiology of; Mental Illness, Etiology of; Mental Illness, Genetics of; Psychiatric Assessment: Negative Symptoms; Schizophrenia; Schizophrenia and Bipolar Disorder: Genetic Aspects; Schizophrenia, Treatment of

Bibliography

American Psychiatry Association 1987 *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III-R), 3rd edn. American Psychiatry Association, Washington, DC

Andreasen N C, Arndt S, Swayze V II, Cizadlo T, Flaum M, O'Leary D S, Ehrhardt J C, Yuh W T 1994 Thalamic abnormalities in schizophrenia visualised through magnetic resonance image averaging. *Science* **266**: 294–8

Crow T J 1995 Aetiology of schizophrenia: An evolutionary theory. *International Clinical Psychopharmacology* **10** (Suppl. 3): 49–56

Dierks T, Linden D E J, Jandi M, Formisano E, Goebel R, Lanfermann H, Singer W 1999 Activation of Heschl's gyrus during auditory hallucinations. *Neuron* **22**: 615–21

Fletcher P C, McKenna P J, Friston K J, Frith C D, Dolan R J 1999 Abnormal cingulate modulation of fronto-temporal connectivity in schizophrenia. *NeuroImage* **9**: 337–42

Frith C D, Blakemore S-J, Wolpert D M 2000 Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews* **31**: 357–63

Harrison P J 1999 The neuropathology of schizophrenia. *Brain* **122**: 593–624

Heckers S, Rauch S L, Goff D, Savage C R, Schacter D R, Fischmaer A J, Alpert N A 1998 Impaired recruitment of the hippocampus during conscious recollection in schizophrenia. *Nature Neuroscience* **4**: 318–23

Ingvar D H, Franzen G 1974 Distribution of cerebral activity in chronic schizophrenia. *Lancet* **2**: 1484–86

Johnstone E C 1991 Defining characteristics of schizophrenia. *British Journal of Psychiatry Supplement* **13**: 5–6

Laruelle M 1998 Imaging dopamine transmission in schizophrenia. A review and meta-analysis. *Quarterly Journal of Nuclear Medicine* **42**: 211–21

Lewis S W 1990 Computerised tomography in schizophrenia 15 years on. *British Journal of Psychiatry Supplement*. **9**: 16–24

Liddle P F, Friston K J, Frith C D, Frackowiak R S 1992 Cerebral blood flow and mental processes in schizophrenia. *Journal of the Royal Society of Medicine* **85**: 224–7

McGuire P K, Frith C D 1996 Disordered functional connectivity in schizophrenia. *Psychological Medicine* **26**: 663–7

McGuire P K, Silbersweig D A, Wright I, Murray R M, Frackowiak R S, Frith C D 1996 The neural correlates of inner speech and auditory verbal imagery in schizophrenia: Relationship to auditory verbal hallucinations. *British Journal of Psychiatry* **169**: 148–59

Okubo Y, Suhara T, Suzuki K, Kobayashi K, Inoue O, Terasaki O, Someya Y, Sassa T, Sudo Y, Matsushima E, Iyo M, Tateno Y, Toru M 1997 Decreased prefrontal dopamine D1 receptors in schizophrenia revealed by PET. *Nature* **385**: 634–6

Ragland J D, Gur R C, Glahn D C, Censits D M, Smith R J, Lazarev M G, Alavi A, Gur R E 1998 Frontotemporal cerebral blood flow change during executive and declarative memory tasks in schizophrenia: A positron emission tomography study. *Neuropsychology* **12**: 399–413

Saykin A J, Shtasel D L, Gur R E, Kester D B, Mozley L H, Stafiniak P, Gur R C 1994 Neuropsychological deficits in neuroleptic naive patients with first-episode schizophrenia. *Archives of General Psychiatry* **51**: 124–31

Seeman P 1986 Dopamine/neuroleptic receptors in schizophrenia. In: Burrows G D, Norman T R, Rubenstein G (eds.) *Handbook on Studies of Schizophrenia, Part 2*. Elsevier, Amsterdam

Shallice T, Burgess P W 1991 Deficits in strategy application following frontal lobe damage in man. *Brain* **114**: 727–41

Spence S A, Brooks D J, Hirsch S R, Liddle P F, Meehan J, Grasby P M 1997 A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena (delusions of alien control). *Brain* **120**: 1997–2011

Spence S A, Hirsch S R, Brooks D J, Grasby P M 1998 Prefrontal cortex activity in people with schizophrenia and control subjects. Evidence from positron emission tomography for remission of 'hypofrontality' with recovery from acute schizophrenia. *British Journal of Psychiatry* **172**: 316–23

Van Horn J D, McManus I C 1992 Ventricular enlargement in schizophrenia. A meta-analysis of studies of the ventricle:brain ratio. *British Journal of Psychiatry* **160**: 687–97

Van Rossum J M 1966 The significance of dopamine receptor blockade for the mechanism of action of neuroleptic drugs. *Archives Internationales de Pharmacodynamie et de Therapie* **160**: 492–94

World Health Organization 1975 *Schizophrenia: A Multinational Study*. WHO, Geneva, Switzerland

C. D. Frith and S-J Blakemore

Schizophrenia, Treatment of

Schizophrenia is a brain disorder of unknown origin which severely deteriorates numerous complex functions of the central nervous system (CNS) including thought, emotion, perception, cognition, and behavior. Schizophrenia is one of the most debilitating psychiatric disorders. Due to the high lifetime prevalence (about 1.5 percent), the typical onset in early adulthood, and the strong tendency towards a chronic

course, the disorder requires a very high degree of health care provisions. About one-quarter of hospital beds are occupied by schizophrenic patients, and the total costs of treatment are enormous (e.g., US$50 billion per year in the United States). Although there is no cure for schizophrenia, the combined administration of pharmacological and psychosocial interventions considerably improves outcome, enhances quality of life in the patients affected, and enables social integration to a large extent.

1. *Causes and Pathophysiology of Schizophrenia*

The causes of schizophrenia are unknown so far. Therefore, the term 'schizophrenia' refers to an empirically defined syndrome characterized by a combination of certain symptoms which occur in a particular temporal pattern. The idea that schizophrenia is a distinct brain disorder is rooted in Emil Kraepelin's concept of dementia praecox (Kraepelin 1893). This concept emphasized one particular aspect of the disorder: the onset of persistent cognitive disturbances early in life. The term 'schizophrenia' was coined by Eugen Bleuler (1916), who wanted to emphasize the loss of coherence between thought, emotion, and behavior which represents another important feature of the disorder. Bleuler actually spoke about 'schizophrenias,' implying a group of diseases rather than one distinct disease entity. The present diagnostic systems compiled by the World Health Organization (*ICD-10* 1992) and the American Psychiatric Association (*DSM-IV* 1994) distinguish various subtypes of schizophrenia (see Table 1) which are classified according to particular symptom combinations, and according to certain aspects of the course and prognosis of the disease. However, these subtypes are defined in a phenomenological manner which neither implies distinct causes for any of the subtypes nor any particular treatment.

Although we do not know individual causes of schizophrenia, we know that the basis for the disorder is an interaction between genetic susceptibility factors and environmental components. According to numerous genetic studies the estimates of hereditability converge to about 80 percent (Owen and Cardno 1999). However, although linkage studies have yielded evidence for a number of susceptibility loci on various chromosomes, neither a single gene nor a combination of genes have been discovered so far which are definitively involved. Similarly, a number of environmental factors (e.g., birth and pregnancy complications, viral infections) are likely to play a role, but details of this role remain to be established.

Although the symptoms of schizophrenia clearly demonstrate a severe disturbance of brain functions, only subtle alterations of brain morphology have been detected. The most stable finding is a slight enlargement of brain ventricles. Recent sophisticated studies combining morphological and functional brain imaging have been suggested to indicate that subtle structural or functional lesions particularly in prefrontal and limbic neural circuits impair the integrity of these circuits (Andreasen et al. 1998). Although it is still a matter of debate whether these lesions are due to a neurodevelopmental or a neurodegenerative process (Lieberman 1999), it is clear that disturbed functioning of neuronal circuits goes along with altered neurotransmission. Dopamine, serotonin, and glutamate are the neurotransmitters most frequently implicated in the pathophysiology of schizophrenia. However, it is still not clear whether disturbed neurotransmission is the cause or the consequence of the major disease process.

Due to this scantiness of etiological and pathophysiological knowledge, the treatment of patients suffering from schizophrenia is based exclusively on empirical clinical knowledge. As outlined in detail below, the treatment approach is multimodal and based on the symptom patterns present in any individual patient (see *Mental Illness, Etiology of*).

2. *Symptoms of Schizophrenia—the Targets for Treatment*

Schizophrenia affects almost all areas of complex brain functions, and thought, perception, emotion, cognition, and behavior in particular. There have been numerous approaches to systematize the plethora of symptoms according to a variety of theoretical concepts. One widely accepted concept is the distinction between positive and negative symptoms (Andreasen

Table 1
Subtypes of schizophrenia according to the major classification systems

ICD-10	[*DSM-IV*]
Paranoid schizophrenia	Schizophrenia, paranoid type
Catatonic schizophrenia	Schizophrenia, catatonic type
Undifferentiated schizophrenia	Schizophrenia, undifferentiated type
Residual schizophrenia	Schizophrenia, residual type
Hebephrenic schizophrenia	—
—	Schizophrenia, disorganized type
Simple schizophrenia	—

Table 2
Positive and negative symptoms of schizophrenia. Areas of symptoms are listed and selected examples appear in parentheses

Positive symptoms
Hallucinations (*auditory*, *somato-tactile*, *visual*, *olfactory*)
Delusions (*persecutory*, *religious*, *thought broadcasting*, *thought insertion*)
Bizarre behavior (*unusual clothing or appearance*, *repetitive*, *stereotyped behavior*)
Positive formal thought disorders (*incoherence*, *illogicality*, *tangentiality*)
Negative symptoms
Affective flattening (*unchanging facial expression*, *paucity of expressive gestures*)
Alogia (*poverty of speech*, *increased response latency*)
Avolition-apathy (*physical anergia*, *problems with personal hygiene*)
Anhedonia-asociality (*reduction in social activities*, *sexual interest*, *closeness*)
Attention (*social inattentiveness*, *inattentiveness during testing*)

et al. 1995, Table 2) (see *Psychiatric Assessment: Negative Symptoms*). This approach groups symptoms according to whether they represent a loss of or a deficiency in a normal brain function (negative symptoms) or the appearance of abnormal phenomena (positive symptoms). From the perspective of treatment the positive–negative dichotomy is of importance because these two domains react somewhat differently to treatment. In general, positive symptoms respond fairly well to treatment with antipsychotic drugs, whereas negative symptoms are rather difficult to influence. The latter is particularly problematic, because negative symptoms are usually the limiting factor for personal and social rehabilitation. In addition to the typical symptoms, patients suffering from schizophrenia often present numerous other psychiatric problems. Among these, anxiety, sleep disturbances, obsessions and compulsions, depression and drug or substance abuse are particularly frequent. Although these additional symptoms and problems may in many cases be the consequence of the typical symptoms rather than signs of independent additional psychiatric disorders, they complicate the course of the disease and often necessitate specific treatment approaches.

The course of schizophrenia varies considerably between patients. Typically, the disease begins during early adulthood with the appearance of rather unspecific negative symptoms resulting in social withdrawal and impaired social and scholastic adjustment. Months to years later positive symptoms such as delusions and hallucinations appear either gradually or abruptly. Across time, positive symptoms tend to show an episodic course, whereas negative symptoms either remain quite stable or even progress. The associated psychiatric problems mentioned above do not show any systematic temporal relationship to the course of schizophrenia itself. Since the syndrome of schizophrenia has been defined, data on the final outcome have varied considerably, depending mainly on the diagnostic concepts applied. There is no doubt that full remission occurs, but probably in less than 20 percent of patients. In the majority of patients schizophrenia takes a chronic course (Schultz and Andreasen 1999). The major benefit from modern treatment strategies is probably not a substantial increase in the rate of full remissions, but a significant reduction in the number of extremely ill patients, and in the severity and number of episodes characterized by prominent positive symptoms.

3. *The Principles of Treatment*

Treatment of schizophrenia is usually multimodal and comprises approaches from two major areas, which are drug treatment and psychosocial interventions. In general, appropriate drug treatment is a prerequisite for the ability of the patients to comply with and actively take part in psychosocial treatments. The more effective drug treatment is, the more specialized and sophisticated psychosocial interventions can be successfully applied. Vice versa, appropriate psychosocial treatment considerably improves the compliance with drug treatment, because it enhances insight into the disease process, which initially is poor in many patients suffering from schizophrenia.

Antipsychotic drugs (also called neuroleptics) are the most important and most effective therapeutic weapon. The major targets of these drugs are positive symptoms, although newer substances might also reduce negative symptoms to some extent (see below). Treatment with antipsychotic drugs is usually a long-term treatment, whereas drugs to control accessory symptoms (anxiolytics, hypnotics, and antidepressants) are prescribed intermittently when needed. The use of electroconvulsive therapy was widespread before antipsychotic drugs were available, but today is very limited, although this treatment is effective in certain conditions (Fink and Sackeim 1996).

Among the psychosocial interventions, supportive and psychoeducative approaches are feasible for the majority of patients, whereas structured social skill training, family therapy, or complex programs in-

cluding cognitive-behavioral therapies require a considerable degree of insight and compliance. Although recent psychosocial treatment approaches incorporate also psychodynamic aspects, classical psychodynamic psychotherapy, in general, is not effective and sometimes even counterproductive.

Prior to the discovery of antipsychotic drugs, most schizophrenic patients had been hospitalized for decades or even for their entire life. Today, very differentiated treatment facilities are available, which to a large extent are community based. These include, besides classical hospitals and outpatient departments, short-term crisis intervention facilities, day-time clinics, and supported living facilities. This network of facilities considerably increases the chance for social integration and reduces the time spent overall in hospitals. However, recent aggressive approaches to shorten the duration of hospital stays dramatically might lead to a high frequency of rehospitalization, poor long-term outcome, and an overall increase in the costs of treatment.

An important limiting factor for all treatment approaches is the patients' compliance. In particular in severely ill patients compliance is often poor. In schizophrenia noncompliance is particularly due to a lack of insight into the fact of being ill, reduced ability to actively take part in the treatment process due to negative symptoms, or active avoidance of treatment due to positive symptoms such as delusions or imperative voices. To enable stable treatment compliance in patients suffering from schizophrenia, a confidential and empathetic attitude of the people involved is an important prerequisite.

4. *Particular Aspects of Treatment*

4.1 *Pharmacotherapy*

The introduction of chlorpromazine into clinical practice in 1952 is the hallmark of the pharmacological treatment of schizophrenia (Delay and Deniker 1952). Chemically, chlorpromazine is a phenothiazine and other substances of this group also possess antipsychotic properties. In the 1950s and 1960s, numerous antipsychotic drugs with different structures (in particular butyrophenones [e.g., haloperidol] and thioxanthenes [e.g., flupentixol]) were developed. Despite this diversity in chemical structure, these first-generation drugs do not differ qualitatively with respect to their profile of beneficial and undesired effects. They target mainly the positive symptoms of schizophrenia mentioned in Table 2 without preferentially affecting one or the other of these symptoms. Their effectiveness does not depend on the subtype of schizophrenia (see Table 1).

The common mode of action of the first-generation antipsychotics is believed to be blockade of the D2 subtype of dopamine receptors (Pickar 1995). Blockade of these receptors in the mesolimbic dopamine system is thought to cause a reduction in positive symptoms, and blockade in the nigrostriatal dopamine system is believed to be responsible for the typical side effects of these drugs. These are severe disturbances of motor behavior caused by a drug-induced dysfunction of the dopaminergic extrapyramidal system, which plays a pivotal role in the control of movements. Table 3 summarizes these extrapyramidal syndromes. Those which occur acutely are often very disturbing, but respond fairly well to treatment (e.g., acute dystonia to anticholinergic drugs), or at least cease upon drug discontinuation or dose reduction. Tardive extrapyramidal syndromes, however, are frequent (5–10 percent on long-term treatment with first-generation antipsychotics), often resistant to treatment approaches, and they usually persist if the antipsychotic drug is stopped. Another severe side effect of unknown cause, which might be related to the extrapyramidal system, is the neuroleptic malignant syndrome (Pelonero et al. 1998; see also Sect. 4.3). Extrapyramidal side effects are common to all first-generation drugs, whereas other, less specific side effects (e.g., sedation, weight gain, postural hypotension) occur with some but not all substances (see *Psychopharmacotherapy: Side Effects*).

Due to the central role attributed to D2 dopamine receptor blockade, it has been thought for a long time that antipsychotic effectiveness and extrapyramidal syndromes are inevitably linked to each other. However, in the late 1960s it turned out that the dibenzodiazepine clozapine is a potent antipsychotic drug that almost never induces extrapyramidal symptoms and is effective in a proportion of otherwise treatment-resistant patients (Angst et al. 1971). Moreover, it was

Table 3
Extrapyramidal syndromes induced by first-generation antipsychotic drugs.

Acute (within days)	[Late or tardive (within months to years)]
Drug-induced Parkinsonism (*bradykinesia, increased muscular tone, tremor*)	Tardive dyskinesia or dystonia (*irregular choreiform, or dystonic movements in any voluntary muscle group*)
Dystonia (*sudden-onset, sustained muscular contractions preferably in the cephalic musculature*)	Tardive tics (*repetitive, short-lasting, stereotyped movements or other kind of behavior* [*e.g., vocalization*])
Akathisia (*internal restlessness and urge to move, often coupled with pacing behavior*)	Tardive myoclonus (*brief asynchronous movements which are not stereotyped*)

shown that clozapine not only improves positive but to some extent also negative symptoms. Therefore, the introduction of clozapine into clinical practice was the first qualitative advance after the discovery of chlorpromazine, and it is justified to call this substance a second-generation antipsychotic drug. Unfortunately it turned out quickly that clozapine has a very particular and important drawback: it induces life-threatening agranulocytosis in about one percent of patients treated (Baldessarini and Frankenburg 1991). Therefore, the use of this drug is only possible when white blood cell counts are monitored in short intervals.

It is still not known why clozapine is a potent antipsychotic drug, but lacks extrapyramidal side effects. Among the possible reasons are that clozapine affects mainly the mesolimbic dopamine system, that it blocks D1 receptors to the same extent as D2 receptors, that it has a high affinity to D4 receptors, and that it also has high affinity to serotonin receptors. This theoretical framework was the basis for the development of various second-generation antipsychotics (see Table 4). Most of them have high affinities for serotonin receptors, although only one substance (olanzapine) is similar to clozapine in the equivalent binding to D1 and D2 receptors. So far there is no indication that any of these newer drugs carries a substantial risk of agranulocytosis, the most dangerous side effect of clozapine. Moreover, these substances have shown to be effective when compared to first-generation antipsychotic drugs and indeed to induce fewer extrapyramidal side effects. For two of them (olanzapine and risperidone) there are data suggesting that they also might ameliorate negative symptoms. However, it still remains to be seen how the effectiveness of these newer drugs compares to clozapine.

Treatment with antipsychotic drugs in schizophrenic patients is usually a long-term treatment continuing for many years. Intermittently, patients may need additional other psychotropic medications. Most frequently anxiolytics, hypnotics, and antidepressants are administered. The recognition and treatment of intervening depressive episodes is particularly important because depressive and negative symptoms are sometimes difficult to differentiate, but need different treatment. When administering antidepressant drugs to schizophrenic patients, one should be aware that during treatment with antidepressants positive symptoms sometimes exacerbate.

4.2 Psychosocial Treatment

It is now generally accepted that psychosocial treatment of patients with schizophrenia is not an alternative to pharmacological treatment, but rather an important complementary approach (Bustillo et al. 2000). Psychosocial measures yield little benefit for positive symptoms; in contrast, intensive psychodynamic psychotherapy might induce exacerbations. The most important and successful interventions are psychoeducative teaching and training in social skills and problem solving. Although it is still debated whether these treatments are directly effective against negative symptoms, there is no doubt that they considerably improve coping with these symptoms.

The major goal of educational approaches is to increase the knowledge about and the insight into the disease process and, as a consequence, to improve treatment compliance. Educational programs are focused on knowledge about the disease process and it's course, the mode of action and side effects of medication, early signs of exacerbation such as restlessness and insomnia, and detailed information about the various community based or hospital facilities available. These programs are particularly effective when important family members are educated as well. Moreover, approaches to improve the communication styles within the patient's network of social relationships are often useful, although the concept that particular forms of communication ('high expressed emotions') play a causative role in schizophrenia has been seriously challenged.

Schizophrenic patients' social skills are often poor. This is one major reason for the often dramatically reduced amount of social contacts. Training of social skills in patients suffering from schizophrenia is performed in a group setting. Such groups are focused on the repetitive training of behavior in various social situations of importance for everyday life. In addition, some rather basic teaching of the theoretical background of successful social interaction might be included.

Impaired problem solving in schizophrenic patients is due to a combination of disturbances in divided attention, planning, and motivation. Problem solving training includes the analysis of problems relevant to daily life and, particularly, dividing a problem into subproblems that can be successively solved. Similar to the training of social skills, the steps of the solution are practiced repeatedly.

In recent years, more complex psychotherapeutic approaches have been developed, combining cognitive, behavioral, or psychodynamic approaches, which, however, are suitable for the minority of patients only.

In particular for chronic patients who in earlier times often were hospitalized for decades, providing supervised residential living arrangements has proven to be of considerable benefit because in this setting the degree of self-sustainment can be maximized in parallel with the permanent provision of professional help.

4.3 Emergency Treatment

Schizophrenia is not only a debilitating but also a life-threatening disorder. If severe positive symptoms are present, such as verbal imperative hallucinations,

Table 4
Relative receptor affinities of selected newer antipsychotic compounds compared to the first-generation drug haloperidol.

	[Dopamine]		[Serotonin]		[Noradrenaline]		Histamine	Acetylcholine
	D_1	D_2	$5HT_1$	$5HT_{2\alpha}$	α_1	A_2	H_1	M_1
Haloperidol[a]	+ + +	+ + + +	..	+	+ +	..	..	..
Clozapine	+ +	+ +	+	+ + +	+ + +	+ + +	+ + + +	+ + + + +
Risperidone	+ +	+ + + +	+ +	+ + + + +	+ + +	+ + +	+ +	..
Olanzapine	+ + +	+ + +	..	+ + +	+ + +	..	+ + + +	+ + + + +
Seroquel	+	+ +	..	+	+ + + +	+	+ + + +	+ + +
Sertindole	+ + +	+ + + + +	+ +	+ + + +	+ + +	..	+	+
Ziprasidone	+	+ + + +	*	+ + + +	+ +	..	+	..

Adapted from Pickar, 1995
a Prototype first generation antipsychotic drug

delusions of persecution, or severe formal thought disturbances, the patient's life is threatened by a misinterpretation of his or her environment (e.g., of dangerous objects or situations) and by suicidal intentions. Suicidality can also emerge when negative symptoms are prominent. Acute suicidality and an acute confusional psychotic state both necessitate immediate treatment, including the administration of anxiolytics, antipsychotics, or both, preferably in a hospital setting. Although every effort should be undertaken to get the informed consent of the patient, it is often necessary to initiate treatment against his or her explicit will. This has to be done in accordance with the legal regulations in the respective country.

Less frequently occurring life-threatening situations in patients suffering from schizophrenia are febrile catatonia and the neuroleptic malignant syndrome, a rare side effect of antipsychotic drugs (Pelonero et al. 1998). Both conditions show similarities, including fever, disturbances of motor behavior, and reduced responsiveness to external stimuli. The differential diagnosis is difficult, but very important because in febrile catatonia treatment with antipsychotic drugs is necessary, whereas such treatment has to be immediately stopped when a neuroleptic malignant syndrome is suspected. In both of these conditions admission to an intensive care unit is necessary.

5. *Future Perspectives on the Treatment of Schizophrenia*

Despite the rapid development of the knowledge about human genetics, it is unlikely that we will see a major advance in our understanding of the genetic causes of schizophrenia in the near future. The reason is that the available epidemiological data render a major contribution of single genes quite unlikely and suggest that multiple genes contribute to the susceptibility to schizophrenia. However, it is likely that genetic investigations will improve the management of schizophrenia by enabling prediction of the treatment response to antipsychotic drugs. Very recently, for example, it has been shown that a combination of six polymorphisms in neurotransmitter-receptor-related genes enable a significant prediction of the treatment response to clozapine (Arranz et al. 2000). Although these preliminary studies must be replicated and extended to other drugs, it is probable that this approach will significantly enhance the success of drug treatment in schizophrenia. It is also likely that the near future will bring new drugs which are the result of a 'fine-tuning' of the chemical structure and the binding profile to dopamine and serotonin receptors of those substances which are presently available. Another approach which is currently being pursued is to target glutamate receptors, although first clinical trials are only in part encouraging (Farber et al. 1999). Based on a theory implicating the immune system in the pathophysiology of schizophrenia, studies are being started that use immunomodulation as a treatment. This is an interesting and promising approach, which also might help to understand the mode of action of presently available antipsychotic drugs because some of them influence the production of cytokines which are important immune mediators (Pollmächer et al. in press). Another approach based on pathophysiological theories is to target phospholipids. These are an important cell-membrane component and there is some evidence suggesting that either their uptake is deficient or that they are excessively broken down in schizophrenic patients (Horrobin 1999).

See also: Nosology in Psychiatry

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, Washington, DC

Andreasen N C, Arndt S, Alliger R, Miller D, Flaum M 1995 Symptoms of schizophrenia. Methods, meanings, and mechanisms. *Archives of General Psychiatry* **52**: 341–51

Andreasen N C, Paradiso S, O'Leary D S 1998 "Cognitive dysmetria" as an integrative theory of schizophrenia: a dysfunction in cortical-subcortical-cerebellar circuitry? *Schizophrenia Bulletin* **24**: 203–18
Angst J, Bente D, Berner P, Heimann H, Helmchen H, Hippius H 1971 Das klinische Wirkungsbild von Clozapin. *Pharmakopsychiatrie* **4**: 201–11
Arranz M J, Munro J, Birkett J, Bolonna A, Mancama D, Sodhi M, Lesch K P, Meyer J F, Sham P, Collier D A, Murray R M, Kerwin R W 2000 Pharmacogenetic prediction of clozapine response. *Lancet* **355**: 1615–16
Baldessarini R J, Frankenburg F R 1991 Drug therapy – Clozapine. A novel antipsychotic agent. *New England Journal of Medicine* **324**: 746–54
Bustillo J, Keith S J, Lauriello J 2000 Schizophrenia: psychosocial treatment. In: Sadock B J, Sadock V A (eds.) *Kaplan & Sadock's Comprehensive Textbook of Psychiatry*. 7th edn. Lippincott Williams and Wilkins, Baltimore, MD, pp. 1210–7
Delay J, Deniker P 1952 38 cas de psychoses traitées par la cure prolongée et continue de 4560 R.P. Annales médico-psychologique **110**: 364–741
Farber N B, Newcomer J W, Olney J W 1999 Glycine agonists: what can they teach us about schizophrenia? *Archives of General Psychiatry* **56**: 13–17
Fink M, Sackeim H A 1996 Convulsive therapy in schizophrenia? *Schizophrenia Bulletin* **22**: 27–39
Horrobin D F 1999 Lipid metabolism, human evolution and schizophrenia. *Prostaglandins Leukotrienes and Essential Fatty Acids* **60**: 431–7
Kraepelin E 1910 *Psychiatrie. Ein Lehrbuch fär Studierende und Ärzte. 8 Auflage*. Barth, Leipzig
Lieberman J A 1999 Is schizophrenia a neurodegenerative disorder? A clinical and neurobiological perspective. *Biological Psychiatry* **46**: 729–39
Owen M J, Cardno A G 1999 Psychiatric genetics: progress, problems, and potential. *Lancet* **354** (Suppl. 1): SI11–14
Pelonero A L, Levenson J L, Pandurangi A K 1998 Neuroleptic malignant syndrome: a review. *Psychiatric Services* **49**: 1163–72
Pickar D 1995 Prospects for pharmacotherapy of schizophrenia. *Lancet* **345**: 557–62
Pollmächer T, Haack M, Schuld A, Kraus T, Hinze-Selch D in press Effects of antipsychotic drugs on cytokine networks. *Journal of Psychiatric Research*
Schultz S K, Andreasen N C 1999 Schizophrenia. *Lancet* **353**: 1425–30
World Health Organization 1996 *International Statistical Classification of Diseases*. World Health Organization, Geneva, Switzerland

T. Pollmächer

School Achievement: Cognitive and Motivational Determinants

Although the analysis of school achievement (SA)—its structure, determinants, correlates, and consequences—has always been a major issue of educational psychology, interest in achievement as a central outcome of schooling has considerably increased during the last decade (see also *School Outcomes: Cognitive Function, Achievements, Social Skills, and Values*). In particular, large-scale projects like TIMSS (Third International Mathematics and Science Study) (Beaton et al. 1996) had a strong and enduring impact on educational policy as well as on educational research and have also raised the public interest in scholastic achievement.

1. School Achievement and its Determinants: An Overview

SA can be characterized as cognitive learning outcomes which are products of instruction or aimed at by instruction within a school context. Cognitive outcomes mainly comprise procedural and declarative knowledge but also problem-solving skills and strategies. The following facets, features, and dimensions of SA, most of which are self-explaining, can be distinguished:

(a) episodic vs. cumulative;
(b) general/global vs. domain-specific;
(c) referring to different parts of a given subject (in a foreign language e.g., spelling, reading, writing, communicating, or from the competency perspective, grammatical, lexical, phonological, and orthographical);
(d) test-based vs. teacher-rated;
(e) performance-related vs. competence-related;
(f) actual versus potential (what one could achieve, given optimal support);
(g) different levels of aggregation (school, classroom, individual level);
(h) curriculum-based vs. cross-curricular or extracurricular.

Figure 1 shows a theoretical model that represents central determinants of SA. Cognitive and motivational determinants are embedded in a complex system of individual, parental, and school-related determinants and depend on the given social, classroom, and cultural context. According to this model, cognitive and motivational aptitudes have a direct impact on learning and SA, whereas the impact of the other determinants in the model on SA is only indirect.

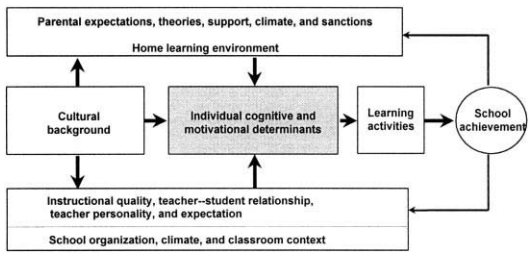

Figure 1
Interplay of individual cognitive and motivational and other determinants of SA

2. Cognitive Determinants

2.1 Intelligence

Intelligence (see *Intelligence, Prior Knowledge, and Learning*) is certainly one of the most important determinants of SA. Most definitions of intelligence refer to abstract thinking, ability to learn, and problem solving, and emphasize the ability to adapt to novel situations and tasks. There are different theoretical perspectives, for example Piagetian and neo-Piagetian approaches, information-processing approaches, component models, contextual approaches, and newer integrative conceptions such as Sternberg's triarchic theory of intelligence, or Gardner's model of multiple intelligences. Among them, the traditional psychometric approach is the basis for the study of individual differences. In this approach quantitative test scores are analyzed by statistical methods such as factor analysis to identify dimensions or factors underlying test performance (Sternberg and Kaufman 1998). Several factor theories have been proposed ranging from a general-factor model to various multiple-factor models. One of the most significant contributions has been a hierarchical model based on second-order factors that distinguishes between fluid and crystallized abilities. Fluid intelligence refers to basic knowledge-free information-processing capacity such as detecting relations within figural material, whereas crystallized intelligence reflects influences of acculturation such as verbal knowledge and learned strategies. Fluid and crystallized abilities can both be seen as representing general intelligence as some versions of a complete hierarchical model would suggest (see Gustafsson and Undheim 1996).

Substantial correlations (the size of the correlations differ from $r = 0.50$ to $r = 0.60$) between general intelligence and SA have been reported in many studies. That is, more than 25 percent of variance in post-test scores are accounted for by intelligence, depending on student age, achievement criteria, or time interval between measurement of intelligence and SA. General intelligence, especially fluid reasoning abilities, can be seen as a measure of flexible adaptation of strategies in novel situations and complex tasks which put heavy demands and processing loads on problem solvers.

The close correlation between intelligence and SA has formed the basis for the popular paradigm of underachievement and overachievement: Students whose academic achievement performance is lower than predicted on the basis of their level of intelligence are characterized as underachievers; in the reverse case as overachievers. This concept, however, has been criticized as it focuses on intelligence as the sole predictor of SA although SA is obviously determined by many other individual as well as instructional variables.

Intelligence is related to quality of instruction. Low instructional quality forces students to fill in gaps for themselves, detect relations, infer key concepts, and develop their own strategies. On the whole, more intelligent students are able to recognize solution-relevant rules and to solve problems more quickly and more efficiently. Additionally, it is just that ability that helps them to acquire a rich knowledge base which is 'more intelligently' organized and more flexibly utilizable and so has an important impact on following learning processes. Accordingly, intelligence is often more or less equated with learning ability (see Gustafsson and Undheim 1996).

Beyond general intelligence, specific cognitive abilities do not seem to have much differential predictive power, although there is occasional evidence for a separate contribution of specific abilities (Gustafsson and Undheim 1996). In Gardner's theory of multiple intelligences there are seven abilities (logical-mathematical, linguistic, spatial, musical, bodily-kinesthetic, interpersonal, and intrapersonal intelligence). Among domain-specific abilities, reading ability that is related to linguistic intelligence in Gardner's model, has some special significance. Specific abilities such as phonological awareness have turned out as early predictors of reading achievement.

2.2 Learning Styles and Learning Strategies

Styles refer to characteristic modes of thinking that develop in combination with personality factors (see *Cognitive Styles and Learning Styles*). While general cognitive or information-processing styles are important constructs within differential psychology, learning styles are considered as determinants of SA. Learning styles are general approaches toward learning; they represent combinations of broad motivational orientations and preferences for strategic processing in learning. Prominent learning style conceptions mainly differ with respect to intrinsic and extrinsic goal orientations and surface processing vs. deep processing. Intrinsic vs. extrinsic motivation deals with whether task fulfillment is a goal of its own or a means for superordinate goals. In deep processing, learners try to reach a comprehension of tasks as complete as possible to incorporate their representations into their own knowledge structures and give them personal significance. In surface processing, learners confine processing to the point where they can reproduce the material in an examination or achievement situation as well as possible (Snow et al. 1996).

Compared to general learning styles, learning strategies are more specific: They represent goal-oriented endeavors to influence one's own learning behavior. While former research mostly dealt with simple study skills that were seen as observable behavioral techniques, now cognitive and metacognitive strategies are included. Learning strategies therefore comprise cognitive strategies (rehearsal, organization, elaboration)

and metacognitive strategies (planning, monitoring, regulation) as well as resource-oriented strategies (creating a favorable learning environment, controlling attention, and sustaining concentration) (Snow and Swanson 1992). Frequently, learning strategies are assessed by means of questionnaires.

In spite of the great importance learning strategies have for understanding knowledge acquisition and learning from instruction, only few empirical data concerning the relation between learning strategies and SA are available. Most studies done in the context of schools and universities have shown positive but small correlations between learning strategies and SA. One reason may be that assessment by questionnaires is too general. On the whole, the exact conditions under which learning strategies are predictive to achievement have still to be worked out.

Besides motivational and metacognitive factors, epistemological beliefs are important determinants for learning strategies and achievement. They refer to beliefs about the nature of knowledge and learning such as: knowledge acquisition depends on innate abilities, knowledge is simple and unambiguous, and learning is quick. Epistemological beliefs can influence quality of achievement and persistence on difficult tasks (Schommer 1994).

2.3 Prior Knowledge

Although there have been early attempts to identify knowledge prerequisites for learning, the role of prior knowledge (see *Intelligence, Prior Knowledge, and Learning*) as a determinant of SA has long been neglected. Since the early 1980s, research on expertise has convincingly demonstrated that superior performance of experts is mainly caused by the greater quantity and quality of their knowledge bases (see *Expertise, Acquisition of*). Prior knowledge does not only comprise domain-specific content knowledge, that is, declarative, procedural, and strategic knowledge, but also metacognitive knowledge, and it refers to explicit knowledge as well as to tacit knowledge (Dochy 1992). Recent research has shown that task-specific and domain-specific prior knowledge often has a higher predictive power for SA than intelligence: 30 to 60 percent of variance in post-test scores is accounted for by prior knowledge (Dochy 1992, Weinert and Helmke 1998). Further results concern the joint effects of intelligence and prior knowledge on SA. First, there is a considerable degree of overlap in the predictive value of intelligence and prior knowledge for SA. Second, lack of domain-specific knowledge cannot be compensated for by intelligence (Helmke and Weinert 1999).

But learning is not only hindered by a lack of prior knowledge, but also by misconceptions many learners have acquired in their interactions with everyday problems. For example, students often have naive convictions about physical phenomena that contradict fundamental principles like conservation of motion. These misconceptions, which are deeply rooted in people's naive views of the world, often are in conflict with new knowledge to be acquired by instruction. To prevent and overcome these conflicts and to initiate processes of conceptual change is an important challenge for classroom instruction.

3. Motivational Determinants

3.1 Self-concept of Ability

The self-concept of ability (largely equivalent with subjective competence, achievement-related self-confidence, expectation of success, self-efficacy; see *Self-concepts: Educational Aspects*; *Self-efficacy: Educational Aspects*) represents the expectancy component within the framework of an expectancy *x* value approach, according to which subjective competence (expectancy aspect) and subjective importance (value aspect) are central components of motivation. Substantial correlations between self-concept of ability and SA have been found. Correlations are the higher, the more domain-specific the self-concept of ability is conceptualized, the higher it is, and the older the pupils are. Self-concept of ability is negatively related to test anxiety (see *Test Anxiety and Academic Achievement*) and influences scholastic performance by means of various mechanisms: Students with a high self-concept of ability (a) initiate learning activities more easily and quickly and tend less to procrastination behavior, (b) are more apt to continue learning and achievement activities in difficult situations (e.g., when a task is unexpectedly difficult), (c) show more persistence, (d) are better protected against interfering cognitions such as self-doubt and other worries (Helmke and Weinert 1997).

3.2 Attitude towards Learning, Motivation to Learn, and Learning Interest

There are several strongly interrelated concepts that are associated with the subjective value—of the domain or subject under consideration, of the respective activities, or on a more generalized level, of teachers and school. The value can refer to the affective aspect of an attitude, to subjective utility, relevance, or salience. In particular, attitude toward learning means the affective (negative or positive) aspect of the orientation towards learning; interest is a central element of self-determined action and a component of intrinsic motivation, and motivation to learn comprises subjective expectations as well as incentive values (besides anticipated consequences like pride, sorrow, shame, or reactions of significant others, the

incentive of learning action itself, i.e., interest in activity). Correlations between SA and those constructs are found to be positive but not very strong (the most powerful is interest with correlations in the range of $r = 0.40$). These modest relations indicate that the causal path from interest, etc. to SA is far and complex and that various mediation processes and context variables must be taken into account (Helmke and Weinert 1997).

3.3 Volitional Determinants

Motivation is a necessary but often not sufficient condition for the initiation of learning and for SA. To understand why some people—in spite of sufficient motivation—fail to transform their learning intentions into correspondent learning behavior, volitional concepts have proven to be helpful. Recent research on volition has focused on forms of action control, especially the ability to protect learning intentions against competitive tendencies, using concepts like 'action vs. state orientation' (Kuhl 1992). The few empirical studies that have correlated volitional factors with SA demonstrate a nonuniform picture with predominantly low correlations. This might be due to the fact that these variables are primarily significant for self-regulated learning (see *Self-regulated Learning*) and less important in the typical school setting, where learning activities and goals are (at least for younger pupils) strongly prestructured and controlled by the teacher (Corno and Snow 1986). Apart from that there certainly cannot be expected any simple, direct, linear correlations between volitional characteristics and SA, but complex interactions and manifold possibilities of mutual compensation, e.g., of inefficient learning strategies by increased effort.

4. Further Perspectives

In Fig. 1 only single determinants and 'main' effects of cognitive and motivational determinants on SA were considered. Actually, complex interactions and context specificity as well as the dynamic interplay between variables have to be taken into account. The following points appear important:

(a) *Interactions among various individual determinants*. For example, maximum performance necessarily requires high degrees of both intelligence and of effort, whereas in the zone of normal/average achievement, lack of intelligence can be compensated (as long as it does not drop below a critical threshold value) by increased effort (and vice versa).

(b) *Aptitude x treatment interactions and classroom context*. Whether test anxiety exerts a strong negative or a low negative impact on SA depends on aspects of the classroom context and on the characteristics of instruction. For example, high test-anxious pupils appear to profit from a high degree of structuring and suffer from a too open, unstructured learning atmosphere, whereas the reverse is true for self-confident pupils with a solid base of prior knowledge (Corno and Snow 1986). Research has shown a large variation between classrooms, concerning the relation between anxiety and achievement as well as between intelligence and achievement (Helmke and Weinert 1999). A similar process of functional compensation has been demonstrated for self-concept (Weinert and Helmke 1987).

(c) *Culture specificity*. Whereas the patterns and mechanisms of basic cognitive processes that are crucial for learning and achievement are probably universal, many relations shown in Fig. 1 depend on cultural background. For example, the 'Chinese learner' does not only show (in the average) a higher level of effort, but the functional role of cognitive processes such as rehearsal is different from the equivalent processes of Western students (Watkins and Biggs 1996).

(d) *Dynamic interplay*. There is a dynamic interplay between SA and its individual motivational determinants: SA is affected by motivation, e.g., self-concept, and affects motivation itself. From this perspective, SA and its determinants change their states as independent and dependent variables. The degree to which the mutual impact (reciprocity) of academic self-concept and SA is balanced has been the issue of controversy (skill development vs. self-enhancement approach).

See also: Academic Achievement: Cultural and Social Influences; Academic Achievement Motivation, Development of; Cognitive Development: Learning and Instruction; Intelligence, Prior Knowledge, and Learning; Learning to Learn; Motivation, Learning, and Instruction; Test Anxiety and Academic Achievement

Bibliography

Beaton A E, Mullis I V S, Martin M O, Gonzales D L, Smith T A 1996 *Mathematics Achievement in the Middle School Years*. TIMSS International Study Center, Boston

Corno L, Snow R E 1986 Adapting teaching to individual differences among learners. In: Wittrock M C (ed.) *Handbook of Research on Teaching*, 3rd edn. Macmillan, New York, pp. 255–96

Dochy F J R C 1992 *Assessment of Prior Knowledge as a Determinant for Future Learning: The Use of Prior Knowledge State Tests and Knowledge Profiles*. Kingsley, London

Gustafsson J-E, Undheim J O 1996 Individual differences in cognitive functions. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Simon & Schuster Macmillan, New York, pp. 186–242

Helmke A, Weinert F E 1997 Bedingungsfaktoren schulischer Leistungen. In: Weinert F E (ed.) *Enzyklopädie der Psychologie. Pädagogische Psychologie. Psychologie des Unter-*

richts und der Schule. Hogrefe, Göttingen, Germany, pp. 71–176
Helmke A, Weinert F E 1999 Schooling and the development of achievement differences. In: Weinert F E, Schneider W (eds.) *Individual Development from 3 to 12*. Cambridge University Press, Cambridge, UK, pp. 176–92
Kuhl J 1992 A theory of self-regulation: Action versus state orientation, self-discrimination, and some applications. *Applied Psychology: An International Review* **41**: 97–129
Schommer M 1994 An emerging conceptualization of epistemological beliefs and their role in learning. In: Garner R, Alexander P A (eds.) *Beliefs about Text and Instruction with Text*. Erlbaum, Hillsdale, NJ, pp. 25–40
Snow R E, Corno L, Jackson D 1996 Individual differences in affective and conative functions. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Simon & Schuster Macmillan, New York, pp. 243–310
Snow R E, Swanson J 1992 Instructional psychology: Aptitude, adaptation, and assessment. *Annual Review of Psychology* **43**: 583–626
Sternberg R J, Kaufman J C 1998 Human abilities. *Annual Review of Psychology* **49**: 479–502
Watkins D A, Biggs J B 1996 *The Chinese Learner: Cultural, Psychological, and Contextual Influences*. CERC and ACER, Hong Kong
Weinert F E, Helmke A 1987 Compensatory effects of student self-concept and instructional quality on academic achievement. In: Halisch F, Kuhl J (eds.) *Motivation, Intention, and Volition*. Springer, Berlin, pp. 233–47
Weinert F E, Helmke A 1998 The neglected role of individual differences in theoretical models of cognitive development. *Learning and Instruction* **8**(4): 309–23

A. Helmke and F.-W. Schrader

School Administration as a Field of Inquiry

Inquiry on the administration of education reflects a variety of intellectual and social influences. Those influences and the main trends in the field's scholarship are examined.

1. *The Ecology of an Applied Field of Inquiry*

The ecology of an applied field of inquiry reflects its indigenous theory and research and that of related fields, demands from the worlds of policy, practice, and professional preparation, and the spirit of the times which usually mirrors societal and global trends. School administration (often educational administration) is a field of many specializations in such areas as politics, organizational studies, fiscal affairs, law, and philosophical issues or, in less-disciplined oriented areas, such as school effectiveness, leadership and supervision, human resource management and labor relations, and equity issues.

Explanations about subject matter are developed in areas of this sort. As in other fields, theoretical plausibility is judged using various logical and evidentiary criteria. However, applied fields also attend to relevance for practice and implications for assorted conceptions of organizational improvement.

2. *A Brief History*

Educational administration became a recognizable academic field in the early twentieth century in North America and later, in other parts of the world (Campbell et al. 1987, Culbertson 1988). Initially, the field was oriented to practical matters and broad issues reflecting pedagogical and societal values. An example was democratic administration, especially popular before and after World War II. In the 1950s, the field began to adopt social science theories and methods. In the mid-1970s, the subjectivistic, neo-Marxist (largely as critical theory), and identity politics perspectives, already popular in the social sciences and humanities, reached the field, with postmodernism a later addition. These perspectives typically were critical of science, which generated extensive debate.

The social science emphasis led to specialization along disciplinary lines that resulted in considerable research, much of it reported in the *Encyclopedias of Educational Research* issued in roughly 10-year cycles, with 1992 the most recent, in two *International Encyclopedias of Education* issued in 1975 and 1994, and in the first *Handbook of Research on Educational Administration* (Boyan 1988). The second *Handbook* (Murphy and Louis 1999) was more fragmented and factious, often mixing research with philosophical commentary.

3. *Current Trends*

Current trends in inquiry into school administration include conflicting views of knowledge and ethics, efforts to bridge the gaps between theory on the one hand and policy and practice on the other, more attention to making sense of complexity, and an emerging literature on comparative international aspects of school administration.

3.1 *Conflicting Views*

The controversies about knowledge and appropriate methods of seeking it, and about ethics, are ultimately philosophical. The contending positions are similar across the social sciences and humanities, although each field's literature reflects its peculiarities.

The main strands of contending thought are grounded in different general philosophies. Subjectivism, a version of idealism, stresses mind, reason,

and intuition. Science is criticized for devaluing humanistic concerns in favor of objectivity. In ethics, there is usually a hierarchy of values with the highest being absolute. In educational administration, Greenfield was this movement's leading early figure, while Hodgkinson remains its best known scholar, with special contributions to ethical theory (Willower and Forsyth 1999; see also relevant entries in the 1992 *Encyclopedia of Educational Research* and the 1994 *International Encyclopedia of Education*).

Critical theory, with its historical roots in the Frankfurt School's search for a more contemporary Marxism, is devoted to a reformist agenda (revolution is out of fashion) and an ethic of emancipation. Its ideology now goes beyond class to stress race and gender, although advocates of identity politics may not endorse critical theory. Suspicious of science, described as serving the purposes of the ruling class, critical theory's literature attends mainly to unjust social arrangements. There are many adherents across subfields of education, with William Foster noteworthy in educational administration. Generally, they contend that schools serve the powerful, giving short shrift to the disadvantaged. The antidote is political action, including radicalized educators.

Postmodernists and poststructuralists reject metanarratives or broad theories of every kind, including scientific and ethical ones, believing such theories totalize and ignore difference. This leads to an emphasis on text or words; all that is left after efforts to understand and improve the world are disallowed. Texts are deconstructed, or examined for their assumptions, including the ways they oppress and 'trivialize' otherness, by what is said and omitted. This view has only recently received much attention in educational administration: the August 1998 issue of *Educational Administration Quarterly* was on postmodernism and the September 1998 *Journal of School Leadership* featured a debate on that perspective.

In applied fields such as administration, postmodernism may be selectively combined with critical theory. This was discussed by Alvesson and Deetz (1996) and illustrated in the June 1998 issue of *Administrative Science Quarterly* on critical perspectives, where the writings of Foucault on power and control were often cited. It remains to add that Derrida, perhaps the leading scholar in the postmodern-poststructuralist camp, recently (1994) appeared to argue for a relaxing of strictures against metanarratives to allow for radicalized critique, which is what deconstruction has been all along. This suggests recognition that postmodern relativism and nihilism have made it irrelevant to the world of human activity and practice.

Adherents of the views sketched often associate science with positivism, a long dead perspective that flourished in the Vienna Circle (1924–1936). Using stringent standards of verification, positivists saw metaphysics as pointless, and values as mere preferences. Although it held on longer in fields such as psychology, positivism lost favor and was folded into the more trenchant analytical philosophy. Nevertheless, positivism has been disinterred to serve as a target in philosophical disputes.

Naturalistic and pragmatist philosophies have been the main sources of support for proponents of scientific inquiry in educational administration. Science is seen as a human activity that seeks explanations of how things work, that can be subjected to public assessment. It is an open, growing, enterprise that is fallible but self-corrective, as better methods and theories displace older ones. Based on logic and evidence, its results have been highly successful from the standpoints of the development of plausible theories and of benefiting humankind. Claims about uncovering an ultimate reality or final truths are not made; they are inconsistent with a self-rectifying conception of inquiry. In ethics, absolute principles and pregiven ideologies are rejected in favor of an inquiry-based process that examines moral choices in concrete situations. Competing alternatives are appraised in terms of likely consequences using relevant concepts and theories, and by clarifying applicable principles. Such principles are derived from cumulative moral experience and are guides, not absolutes (see Willower and Forsyth 1999).

These contrasting philosophical approaches were sketched to suggest the substance of contemporary debate, because of their implications for scholarship in school administration. Qualitative research, for instance, is a mainstay of subjectivists and those critical theorists who do 'critical ethnography.' Postmodernists are less explicit about research, but several articles in the *Educational Administration Quarterly* issue on postmodernism reported field research. Qualitative studies were done in educational administration long before these views gained popularity, but have become more widespread, and the literature now includes work aimed at demonstrating injustices, rather than building theory. Changing philosophical emphases can legitimate new subject matter and methods, as shown by comparing the second *Handbook of Research on Educational Administration* (Murphy and Louis 1999) with the first (Boyan 1988).

However, philosophical influences are filtered by variations of interest and practice. Work may borrow selectively, sometimes from conflicting philosophies, or tailor philosophical ideas to special purposes, or ignore them entirely. For instance, in ethics, single concepts such as equity, caring, or community often are stressed, sometimes along with visions of the ideal school. More philosophical explorations of moral principles and their applications are scarcer. In epistemology, many writers in antiscience camps argue against positivistic 'hegemony,' while ignoring formidable views of inquiry. Hence, many disputes are about politics rather than theories of knowledge. In educational administration, examples of broad philo-

sophical efforts are Evers and Lakomski's joint work on epistemology and Hodgkinson's and Willower's respective approaches to ethics and to inquiry (See Boyan 1988 and Willower and Forsyth 1999).

The substitution of politics for philosophical substance can be found in most social sciences and humanities, reflecting the influences of critical theory, postmodernism, and identity politics. Various commentators see these influences as declining because of their one-sidedness and practical irrelevance. However, they remain visible, claiming a share of the literatures of assorted specializations, including educational administration.

3.2 Theory and Practice

In school administration, current issues command much attention; theoretical discussions have a more specialized appeal. Devolution, school site based management, and restructured decision processes are examples of a type of contemporary reform that is the object of wide attention. Such reform results from recent trends in government and politics, but also is the current incarnation of a long line of participatory-type schemes, especially seen in Western countries. They range across democratic administration, human relations, organizational development, open climates and participatory styles of leadership, organizational health, and staff empowerment. This sort of reform is oriented to practice, but has theoretical grounding because openness and participation are often held to be paths to organizational improvements, such as increased staff motivation and commitment. However, most school improvement efforts make demands on time, erode teacher autonomy, and disturb orderly routines, adding to staff overload. Counter forces include pressures to display legitimating progress and educator hopes for positive student outcomes. Such contradictory pressures are characteristic of attempts to improve school practices, but how they play out depends on contexts and contingencies. Another contemporary reform is privatization. It stresses consumer choice, efficiency, and economy. Often linked to efforts to change governmental policies on education, it is squarely in the political arena.

While reforms of policy and organization aim at changing practice, students of cognition and learning have sought to understand how practitioners solve problems and make decisions. Influenced by Simon and the 'Carnegie School,' this work emphasizes domain-general problem-solving processes and domain-specific knowledge. Research comparing expert and novice decision makers shows the importance of both. In educational administration, the investigations of Leithwood and his associates (Leithwood and Steinbach 1995) have been noteworthy (for other sources, see Willower and Forsyth 1999).

Professional preparation programs for school administrators use many approaches to transfer learning to practice. They include case studies, simulations (sometimes computer based), films, approximations of virtual reality, internships, and other school site based activities. Program improvement has benefited from the cooperative efforts of professorial and administrator associations (Willower and Forsyth 1999). Improving the practice of school administrators and educational reform is, like inquiry, ever unfinished. Current successes do not guarantee future ones, despite progress made in developing explanative theories, and continuing efforts to improve schools and the preparation of administrators. A barrier to the implementation of theory-based reforms are practice is the multiplicity and complexity of the influences that impinge on the schools, considered next.

3.3 Making Sense of Complexity

School administration appears to be more complex than many forms of management. Schools tend to be vulnerable to community and societal forces, are usually subject to substantial regulation (no surprise since they serve society's children), and have difficulty demonstrating effectiveness. The latter occurs because of the complexity of showing the school's part in student learning vs. that of family, variations in student motivation and ability, and other factors. Further, schools are commonly expected to foster subject matter achievement and acceptable social skills and conduct. Beyond that, school administrators oversee personnel who ordinarily define themselves as professionals and value latitude in their work. Thus, school organizations are characterized by external pressures and internal forces that heighten complexity. In addition, social changes have placed new demands on schools as societal arrangements that provided regulatory settings for children and youth, have begun to erode. As this occurred, schools increasingly have had to cope with an array of social ills, for instance, substance abuse, vandalism, and violence, against a backdrop that too often includes dysfunctional families and subcultures with norms that devalue learning, along with the miseducative effects on young people of ubiquitous mass media. As a result, many schools face expanded responsibilities, adding new complexities to the administration of schools.

Such complexity can be daunting. One response in education, as in business, has been a susceptibility to fads, with total quality management as a recent example. These fads may include some reasonable ideas, commonly packaged as part of a larger program. Although often treated as panaceas, they are typically short-lived. Such fads can be explained as attempts to give perceived legitimacy to organizational efforts to confront what in reality are intractable problems.

While practitioners adapted to changing social conditions and increasing complexity, those who study school administration sought to gain a better understanding of such phenomena. More sophisticated

quantitative analyses facilitated by computer technology provided ways of examining relationships among many variables in numerous combinations, clusters, and sequences. Although empirical investigations rarely employed such analyses, their availability provides possible handles on complexity. More popular recently have been field studies, done as participant observation, ethnography, or case studies. While qualitative methods are favored by those wishing to advance a particular view using illustrations from selected 'narratives,' more traditional field work has long sought to plumb complexity by attending to the details of daily activity. Disputes about qualitative vs. quantitative research are not new, mirroring the richness vs. rigor debates in sociology in the 1930s (see Waller 1934). In educational administration, the dominant view appears to favor the use of a variety of methods recognizing that each has strengths and weaknesses. In any event, empirical studies in quantitative and qualitative modes have sought to make complexity more understandable, albeit in different ways.

The increased recognition of complexity has rekindled interest in chaos theory in educational administration and the social sciences. Chaos theory, it is well to recall, is not a denial of the possibility of orderliness. Rather it is a search for odd, intricate, or many faceted patterns. Reviewing chaos scholarship, Griffiths et al. (1991) concluded enthusiasm for the theory had not resulted in meaningful social research. They suggested that chaos theory has limited applications in educational administration because of the theory's dependency on precise measures. Conceptual efforts to accommodate fluidity and fluctuation are not new. To cite others, they range from Hegelian dialectical analysis with its clash of opposites, followed by a new synthesis, to threshold effects noted only when a variable reaches a certain level, with little or no gradation, to equifinality or functional equivalents where similar results stem from different sources. Ethical theory has also been concerned with complexity. Perspectives that recognize how moral choices are embedded in elaborate contexts, and attempt to include potential consequences of action as part of choice, are illustrative. Related are cognitive-type studies of problem solving and decision making and work on types of intelligence and their relationship to problem solving.

Clearly, human experience is fluid and complex, and human behavior is often irrational and contradictory. Yet, in inquiry, the irrational is apprehended through rational processes, and experience is known through the use of concepts and explanations. Obviously, concepts and theories simplify, but so do intuition, lore, pregiven ideologies, aphorisms, and other ways of confronting experience. Nothing comprehends everything. The trick, in science and administration, is to include the relevant elements and get to the heart of the problem. The record on understanding complexity in school administration, as in the social sciences, is one of incremental gain. The methods of scientific inquiry can make complexity more comprehensible and more manageable, but there are no panaceas.

3.4 Internationalization

Educational administration is becoming internationalized. New preparation programs and journals are cropping up around the world. Established outlets such as the *Journal of Educational Administration* and *Educational Administration Quarterly* regularly feature comparative-international pieces, and research done cooperatively by investigators in different countries is increasingly appearing, as for example, in the July 1998 issue of the *Journal of School Leadership*. In addition to the *International Encyclopedias of Education*, *World Yearbooks of Education*, and others, new works are being published, for example the 1996 *International Handbook of Educational Leadership and Administration*. Scholarly associations such as the University Council for Educational Administration (in North America) and the Commonwealth Council for Educational Administration and Management continue to develop cooperative activities.

Internationalization presents opportunities for inquiry that take advantage of different societal-cultural settings to examine the impact of contingencies on behavior and outcomes. Such research can show how the relationships of the variables under study are mediated by setting variables. This, in turn, can lead to new theoretical explanations, not to mention better international understanding.

See also: Administration in Organizations; Educational Research and School Reform; School (Alternative Models): Ideas and Institutions; School as a Social System; School Effectiveness Research; School Management

Bibliography

Alvesson M, Deetz S 1996 Critical theory and postmodernism approaches to organizational studies. In: Clegg S R, Hardy C, Nord W R (eds.) *Handbook of Organizational Studies*. Sage, London

Boyan N J (ed.) 1988 *Handbook of Research on Educational Administration*. Longman, New York

Campbell R F, Fleming T, Newell T J, Bennion J W 1987 *A History of Thought and Practice in Educational Administration*. Teachers College Press, New York

Culbertson J A 1988 A century's quest for a knowledge base. In: Boyan N J (ed.) *Handbook of Research on Educational Administration*. Longman, New York

Derrida J 1994 *Specters of Marx: The State of the Debt, the Work of Mourning, and the New International*. Routlege, New York

Griffiths D E, Hart A W, Blair B G 1991 Still another approach to administration: Chaos theory. *Educational Administration Quarterly* **27**: 430–51
Leithwood K, Steinbach R 1995 *Expert Problem Solving*. State University of New York Press, Albany, NY
Murphy J, Louis K S (eds.) 1999 *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco
Waller W 1934 Insight and scientific method. *American Journal of Sociology* **40**: 285–97
Willower D J, Forsyth P B 1999 A brief history of scholarship on educational administration. In: Murphy J, Louis K S (eds.) *Handbook of Research on Educational Administration*, 2nd edn. Jossey-Bass, San Francisco, pp. 1–24

D. J. Willower

School (Alternative Models): Ideas and Institutions

Ever since its beginnings at the turn of the eighteenth to nineteenth century, the modern educational system has been accompanied by alternative schools, the numbers and broad effects of which have varied over the course of the years. This coexistence is by no means based on independence, but is the manifestation of complex reciprocal relations which exist despite fundamental differences in the ideologies of the two approaches. Phases of intensive development in the domain of traditional schools have been mirrored by similar phases of intense activity in the alternative school movement; alternative schools fulfill important functions within the general spectrum of educational possibilities.

1. *Background and Conception*

Because proponents of alternative schooling are critical of contemporary manifestations of the traditional school system, the focus of their critique has shifted over time. In general, however, their criticism is based on objections to the structures and functions of the modern school. They do not restrict their demands to a return to earlier historical epochs and structures in which schools still played a minor role. Their objectives are for duress (compulsory education and instruction) to be replaced by freedom; teacher-directed schooling and instruction by self-directed action on the part of the student; the curriculum by the individual needs of the student; competition and rivalry by community; compartmentalization by wholeness and integral methods in instruction; the separation of life and school by an interweaving of the two spheres (e.g., Holzman 1997). The fundamental criticism of traditional schooling is based on the conviction that this form of education is burdened by a basic contradiction: Although adults—and particularly adults living in modern democracies—are granted dignity, majority, and self-determination, children and adolescents at school are denied these basic rights, albeit on the pretext that they first have to learn how to exercise them. Because it is not possible to recreate the historical conditions which preceded institutionalized education and its inherent ambivalences, however, the critics of traditional schooling have to resort to so-called 'counterschool' initiatives (German: *Gegenschulen*). Although some of these are embedded in a variety of alternative organizations (e.g., flat-sharing communes, communal child care, and even dietetic measures), in general it is not society that is deschooled, but school itself. The extent to which this radical rearrangement leads to the complete disintegration of institutions is dependent on its (partly unintended) consequences.

2. *Structural and Organizational Characteristics*

Pedagogical aims are reflected in the specific organizational characteristics of any given school (e.g., the general structure of the school, the relationships between teachers, parents, and students, the organization of classes, the methods implemented). In turn, these characteristics constitute the decisive prerequisites allowing many alternative institutions to exceed the norm with regard to content and in both the pedagogical and the social fields. Indeed, the typical approach of the alternative schools can be reduced to a simple formula: in contrast to the universal institutions of the traditional system, which are characterized by high levels of internal differentiation and complicated organizational structures, alternative institutions are distinguished by social clarity and immediacy owing to the fact that they are, from the outset, geared to the particularities of their respective clients. Accordingly, the student body in an alternative school on average numbers less than 200 (cf. Mintz et al. 1994). Thus, irrespective of content-related differences, the alternative school represents a special type of school or, more specifically, a type of school offering a special kind of education. While this offer may formally be an open one, it in fact only reaches the few who know how to take it up.

Due to these invisible processes of selection, alternative schools can be expected to be more or less immune to many of the problems arising in schools which really are open to all (as is made clear by the reports and evaluations of alternative schools; cf. as a very vivid example Duke and Perry 1978). In accordance with the very definition of the alternative school, however, these assets cannot readily be transferred to the mass school system as a whole. Instead of having to simultaneously meet the needs of students from different backgrounds and with various levels of qualification, alternative schools are able to respond

to individual differences and developments emanating from a more homogeneous context of attitudes and lifestyles. The additional potential for motivation and identification (in both students and teachers) which is inherent in a free choice of school or in the decision to embark upon a rather unusual experiment constitutes another factor which cannot easily be transferred to the traditional school. As a rule, 'counterschools' are spared the arduous coordination of pedagogical subfunctions on the basis of a division of labor: because of the small size of these institutions, instruction and education, counseling and care all rest in the hands of a few. Instead of having to strictly adhere to general and at least partly formulated rules in order to ensure the smooth running of the school and its classes, confidence in the closer personal ties and connections of the school community makes it possible to act more effortlessly, easily, and flexibly. Finally, the danger of merely running out of inspiration within the confines of internal school processes is reduced, as the high levels of input from the local environment mean that the system's contours remain rather diffuse. The close links between the home and the school, which are not limited to financial support, but also involve the parents' practical cooperation, can be seen as another important asset.

3. *Developments Since the 1960s*

The concepts behind alternative schooling and its implementation reach far back in time, and in some respects they refer to the founders of modern educational theory (e.g., Rousseau or Herbart). Progressive education in the United States and so-called *Reformpädagogik* in Germany, both of which emerged after the end of the nineteenth century, may certainly be considered to be phases of activity to which a number of existing institutions can be traced back (e.g., Semel and Sadovnik 1999). However, the main wave of alternative school foundation resulted from political and social conditions including the failure or petering out of large-scale attempts at educational reform (e.g., Carnie et al. 1996); this holds especially for the Anglo-American countries since the 1960s. Most of the alternative institutions founded during this phase were set up in the United States, and the majority are still to be found there (cf. Mintz et al. 1994). Although less than 5 percent of students at elementary and secondary school actually attend alternative schools, the institutions now number several thousands in the United States, most of them elementary schools. The institutions are highly divergent in terms of their targets, methods, integration of parents and students, links with the traditional system, locus of control, and circles of addressees. It would thus seem justified to focus this overview of alternative schooling since the 1960s on the situation in the United States, all the more so because both the fundamental principle of alternative schools ('as many institutions as there are needs and groups') and current trends in development (adopting the tasks of the traditional system) are to be observed there.

One important form of alternative education is home schooling (cf. Evans 1973), where parents instruct their children at home with the permission of the state. If—as is quite often the case—these children meet up with other home students in (elementary) schools on single days of the week, there is a sliding transition to school cooperatives aiming at unconstrained, playful, spontaneous learning in open instruction. Other institutions specifically target problematic cases, so-called at-risk students, including not only chronic truants and low-performing students, but also talented nonconformers (in the broadest sense). A well-balanced program of drill and freedom is intended to render school attractive again, thus helping these students to enhance their performance.

However, this development has also led to a reverse trend, a sort of back-to-basics shift to formal 'three Rs' schools, which specifically emphasize the basic skills, discipline, and respect for the authority of parents and teachers (cf., e.g., Deal and Nolan 1978). Similar trends are to be observed at the secondary level. The large number of schools-within-schools—smaller, less complex units running different experimental programs, and affiliated to larger state schools—is particularly notable. The spectrum also includes schools attempting to set themselves up as 'just community schools' based on the principles of Kohlberg's developmental theory (cf. Kohlberg 1985). The so-called 'schools without walls' represent another form of alternative education. They not only offer a 'normal' curriculum (differing from that of the traditional school in terms of content and time), but give students the opportunity to learn in the adult world. This is intended to provide students with extra-scholastic experience and to enable them to prove themselves in the world of work. Although this approach initially seemed very promising, experience has shown that it is in fact very hard to put into practice.

4. *Recent Trends and Perspectives*

Over the past few decades, there have been signs of a kind of division of function and labor between the traditional and the alternative schools in the United States. This has been expressed in an increase in the state-controlled administration of alternative institutions. The fact that such large numbers of US alternative schools were founded at the beginning of the 1980s is probably related to this development. The radical free schools (e.g., Dennison 1969), which signaled the upheavals of the 1960s and certainly provided considerable impetus for subsequent developments, play a lesser role in today's broader stream

of initiatives. The basic conceptions and expectations of alternative schooling have also been affected, resulting in the extremely short lifespan of some of these institutions. There is still a great deal of fluctuation, making it hard to make valid evaluations of these institutions (cf. the data in Mintz et al. 1994, p. 10ff.). Even the early approaches and experiments demonstrated the great weaknesses and dangers of such institutions—the cost of their structure-related advantages (cf. also Swidler 1979). Critics have reduced this to the polemic formula that precisely those principles and premises which allowed the pure 'counterschool' initiatives of the early period to emerge later threatened their ability to survive. Wherever the concept of freedom as a counterpole to the compulsory character of the traditional institutions went hand in hand with systematic shortcomings in the specification of learning targets, there was the risk that later conflicts be all the more severe due to the differing hopes and expectations of those involved (cf. Barth 1972, especially p. 108ff). When every school-related action was dependent on the consensus of every single member of the school community, the school's fate was put in the hands of changing minorities with the power of veto, all the more so when the small size of these private institutions kindled fears of sudden financial ruin. Allowing almost all of those concerned to take an active role in the decision-making process did not in fact ensure the equality of all participants—instead of increasing the capacity to solve problems, there was a growing risk of disappointment, and the necessary basic consensus was threatened by frequent conflicts and failures in minor matters. The firm and wide-ranging commitment of parents, extending well into issues of classroom practice, often resulted not so much in the enhancement as in the disturbance of school activities. Teachers were consequently made to feel insecure and were put on the defensive instead being provided with support and control. Attempts to simultaneously cope with the factual demands, organizational affairs, and personal standards inherent in these institutions—preferably without referring to preset guidelines—led to the systematic overexertion of those concerned, and the end of many 'counterschools' was marked by the burnout of the teaching staff (Dennison described this phenomenon as early as 1969). The unconditional openness with respect to the other members of the school community and the intimacy of the small group responsible for the school meant that factual disputes often escalated into personal trials. Moreover, school affairs were often overshadowed by secondary motives such as parental self-realization or the search for security.

By integrating and promoting alternative schools, the state aims to increase the diversification of the educational spectrum, to enhance the overall efficiency of schools (which has been subject to criticism for decades; cf., e.g., Goodlad 1984), and to better meet the specific needs of certain problem groups in both school and society. The US programs of community schools, magnet schools and, more recently, charter schools provide evidence for this trend, and are looked upon with growing interest by other countries as possible models for the resolution of system-related problems. However, more detailed analyses of these models point to significant limitations in their problem-solving capacity, limitations which are ultimately rooted in the basic concepts of market and choice (cf. Lauder et al. 1999). The community schools program has already been thwarted by the US authorities' desegregation measures: although these schools were directed at the advancement of members of the underprivileged minorities, most of whom are colored, they undermined the desired objective of racial integration. Magnet schools were supposed to be a free choice of school with a special educational program which acted like a magnet, attracting students from beyond the racial boundaries marked out in residential areas and school districts. They were intended to promote racial integration, which can scarcely be achieved through the compulsory measure of busing. A number of objections have been voiced against this school form. For example, it is argued that the enlargement of catchment areas weakens the link between the school and the parental home. Furthermore, although magnet schools may well establish a balanced relation between whites and blacks in their own student bodies, there are too few magnet schools to ensure social compensation on a large scale.

Moreover, there is an imperceptible shift in their circle of addressees: instead of reaching the underprivileged, out-of-school colored population with their specific educational provision, magnet schools in fact seem to appeal to white parents and their children who would otherwise switch to private schools or the more privileged suburbs (cf. Smrekar and Goldring 1999).

The charter schools, on the other hand, which in a way transfer the principles of the private school (i.e., the autonomy of schools and a free choice of school) to the state school system, are faced with comparable problems on a different level inasmuch as unforeseen side effects have been revealed in the social process. Despite the fact that the numbers of charter schools have increased considerably over the past few years, they are still far from ever being able to constitute the majority of state schools, belonging by definition to the optional domain of educational provision. The half-hearted legislative reservations perceptible in most of the federal states against according too many privileges to these schools threaten to render the entire charter program futile (cf. Hassel 1999).

Bibliography

Barth R S 1972 *Open Education and the American School.* Agathon Press, New York

Carnie F, Tasker M, Large M (eds.) 1996 *Freeing Education. Steps Towards Real Choice and Diversity in Schools*. Hawthorn Press, Stroud, UK
Deal T E, Nolan R R (eds.) 1978 *Alternative Schools. Ideologies, Realities, Guidelines*. Nelson-Hall, Chicago, IL
Dennison G 1969 *The Lives of Children. The Story of the First Street School*. Random House, New York
Duke D L, Perry C 1978 Can alternative schools succeed where Benjamin Spock, Spiro Agnew, and B. F. Skinner have failed? *Adolescence* **13**: 375–92
Evans T 1973 *The School in the Home*. Harper & Row, New York
Goodlad J I A 1984 *Place Called School: Prospects for the Future*. McGraw Hill, New York
Hassel B C 1999 *The Charter School Challenge—Avoiding the Pitfalls, Fulfilling the Promise*. Brookings Institution Press, Washington, DC
Holzman L 1997 *Schools for Growth. Radical Alternatives to Current Educational Models*. Lawrence Erlbaum Associates, Mahwah, NJ
Kohlberg L 1985 The just community approach to moral education in theory and practice. In: Berkowitz M W, Oser F (eds.) *Moral Education: Theory and Application*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 27–87
Lauder H, Hughes D, Watson S, Waslander S, Thrupp M, Strathdee R, Simiyu I, Dupuis A, McGlinn J, Hamlin J 1999 *Trading in Futures. Why Markets in Education Don't Work*. Open University Press, Buckingham, UK
Mintz J, Solomon R, Solomon S, Muscat A (eds.) 1994 *The Handbook of Alternative Education. A Solomon Press Book*. Macmillan, New York
Ramseger J 1975 *Gegenschulen. Radikale Reformschulen in der Praxis* [*Radical Reform Schools in Practice*]. Klinkhardt, Bad Heilbrunn, Germany
Semel S F, Sadovnik A R (eds.) 1999 *'Schools of Tomorrow,' Schools of Today. What Happened to Progressive Education History of School and Schooling*. Peter Lang Verlag, New York, Vol. 8
Smrekar C, Goldring E 1999 *School Choice in Urban America. Magnet Schools and the Pursuit of Equity*. Teachers' College, Columbia University, New York
Swidler A 1979 *Organization Without Authority. Dilemmas of Social Control in Free Schools*. Harvard University Press, Cambridge, MA

A. Leschinsky

School as a Social System

A 'social system' is a set of related elements that work together to attain a goal. Social scientists have frequently employed the concept of a social system to study the school. This has been done in three ways. First, the class has been portrayed as a social system; its elements include the teacher, the students, and formal and informal groups within the class. Second, the school itself has been viewed as a social system; its components are the administration, faculty, counselors, students, academic departments, the curriculum, the extra curriculum, and social networks and other subunits within the school. Third, the school has been seen as one of a number of subunits in the larger social system of society (Gordon 1957, Loomis and Dyer 1976, Parsons 1961, Smelser 1988).

1. *Viewing the School Class as a Social System*

The most familiar application of the social system model to schools is found in Parsons' (1959) classic essay on the school class as a social system. Parsons focuses on the class, rather than the larger unit of the school, because he sees the class as the primary place where students learn and mature. According to Parsons, the two most important functions of the class are 'socialization' and 'allocation.' These functions define and motivate the academic and social processes that occur in the classroom and the interactions that occur among its various components.

1.1 *Socialization*

Socialization is the process through which students internalize the kinds of commitments that they need to play a useful role in adult society. Although the family and community also socialize students, the length of time students spend in school makes the school class a major influence in the socialization process. Time spent in school during the students' formative years provides ongoing opportunities for systematic, intensive socialization. During this time, students are expected to accept societal values and use them to guide their behavior.

Parsons claims that several structural characteristics of the class facilitate the socialization of students. First, students assigned to the same class are fairly homogeneous with respect to age and social development. They are also somewhat homogeneous with respect to social class, since class composition is constrained by the socioeconomic characteristics of the neighborhood in which the school is located. The developmental and socioeconomic homogeneity of the students assists the teacher in transmitting values to the students who become, in turn, models of appropriate behavior for each other.

Second, students in the same class are under the tutelage of one or a small number of adult teachers. The age difference between the teacher and students supports the teacher's authority. The singular role of the teacher as the representative of adult society lends legitimacy to the teacher's values.

A third structural feature of the school class is the curriculum. Pupils are exposed to the same curriculum and assigned the same tasks. A shared curriculum and similar activities allows the teacher to reinforce cultural values and attitudes in multiple ways over the school year.

A fourth structural feature of the class is its reward structure. Teachers generally establish a set of rewards

and punishments governing student behavior in a classroom. Good behavior, defined by the teacher's values, is rewarded, while disruptive actions are sanctioned. Students learn the rules of conduct that apply to social and work situations, and are motivated to obey these rules by the rewards or punishments their behaviors incur.

1.2 Allocation

Allocation is the process of sorting individuals and assigning them to groups based on their ability or skills. In a school, allocation refers to the assigning of students to instructional units according to their abilities or achievement. The purpose of this sorting process is to prepare students to attain an occupational position in society commensurate with their capabilities (Gamoran et al. 2000, Hallinan 1994, Lynch 2000, Oakes 1994).

Parsons defines achievement as excellence, relative to one's peers, in meeting the expectations of the teacher. Achievement has two components: cognitive and moral. Teachers attempt to motivate students to achieve academically and to learn the skills needed to perform in adult society. They also teach students a moral code and a set of behaviors consistent with the values of adult society. Students are evaluated on each of these two components. Society may assign greater weight to one component than the other or may treat them as equally important. As with socialization, structural differentiation occurs through a process of rewarding students for excellence and punishing them for failure to meet teacher expectations.

In the elementary school classroom, the criteria for excellence are not clearly differentiated across cognitive and moral dimensions. Teachers aim to develop both good citizenship and an achievement motivation in children. They also begin the process of classifying youth on the basis of their ability to achieve. While assigning students to classes by ability typically does not occur in elementary school, many elementary school teachers instruct their students in small, ability-based groups within the classroom. Their purpose is to facilitate instruction and to prepare students for later channeling into ability-based classes. This initial sorting sensitizes students to the differential achievement that characterizes a class and to the way rewards are allocated for performance.

At the secondary level, the cognitive and moral components of achievement are separable, and greater emphasis is placed on cognitive achievement. Parsons claims that students who achieve academic excellence in high school are better suited for technical roles in society, while those who excel in the moral or social sphere are more fit for socially oriented roles. The actual sorting of students in high school typically occurs by assigning students to classes for instruction based on their ability level. Since ability-grouped classes vary by student and teacher characteristics and by the curriculum, they represent different social systems within the school.

The conceptual attraction of the social system model of the classroom has stimulated a body of research over the past few decades. The model depicts how the teacher, formal and informal groups, and individual students interact in the classroom and perform or cooperate with the functions of socialization and allocation. Examples of research based on this model include studies of teacher expectations, the quantity and quality of teacher–student interactions, gender and race effects on task-related and social interactions, peer influences and student friendship groups.

2. Viewing the School as a Social System

Researchers are more likely to focus on the school, rather than the classroom, when utilizing the social system model to analyze education. Like the class, the school must perform the functions of socialization and allocation in order to play its assigned role in society. Viewing the school as a social system directs attention to how the parts of a school interact to carry out these functions.

2.1 Socialization

For students to be socialized successfully in school, they must cooperate in their education. They are less likely to resist if they accept the authority of their teachers and believe that the school's policies and practices are fair. When students agree with school policies and practices, they are unlikely to resist socialization. With student cooperation, a school is expected to attain its goal of graduating students who have internalized societal values and norms beyond those held by their families.

2.2 Allocation

To attain its goal of allocation, a school must locate students on an achievement hierarchy, defining each person's capability and aspirations. A critical factor in the school's ability to allocate students is that both the family and school attach high value to achievement (Parsons 1959). When adults view student achievement as a high priority, students are more likely to internalize the motivation to achieve and to cooperate with adults as they orient students toward specific adult roles. When schools succeed in this effort, they match society's human resources with its adult occupational structure.

The primary mechanism for the allocation process is the organizational differentiation of students for

instruction by ability. Middle and high schools in the United States typically assign students to Academic, Vocational, and General tracks. Students assigned to the Academic track take college preparatory courses, those assigned to the Vocational track take skills and job-related courses, and students assigned to the General track are offered both low-level academic courses and skills courses. Track assignment is a major determinant of whether a student advances to college or enters the job market after graduation.

Theoretical and empirical studies demonstrate the way various parts of a school interact to socialize students and prepare them for the labor market. Research on the school as a bureaucracy, on the formal or informal organization of the school, and on social networks within the school, illustrates this approach. These and other studies demonstrate the heuristic value of conceptualizing the school as a social system and show how this model has integrated a wide variety of studies about schooling.

3. *Viewing the School as a Subsystem in Society*

Many social scientists have used the social system model to analyze the role of various institutions in society. Education is seen as one of society's primary institutions, along with religion, the economy, and the judicial system. The aim of a social system approach to the study of schools in society is to ascertain how schooling enables society to achieve its goals.

3.1 Structural Functionalist Perspective on Schools in Society

A theoretical perspective that dominated early twentieth century societal analysis was structural functionalism (Durkheim 1956, Parsons 1951). A major premise of structural functionalism is that a society must perform a set of functions in order to survive. According to Parsons (1956), these functions are: obtaining and utilizing resources from the system's external environment; setting goals for the system and generating the motivation and effort to attain these goals; regulating the units of the system to insure maintenance and to avoid destructive conflicts; and storing and distributing cultural symbols, ideas, and values. Schools perform these functions by socializing students to societal values, by providing a common culture and language, by enabling and encouraging competition, and by preparing students for the labor market.

Structural functionalism purports that a social system must exist in a state of equilibrium or, if disrupted, must make adjustments to restore balance. If one social institution in society undergoes major change, interrelated social institutions are expected to accommodate this change and to bring society back to a stable state. For example, if the economy of a society were to change, students would need to be socialized and allocated in different ways to support the new economic structure. The social system would then be restored to balance, though it would differ structurally from its original state.

3.1.1 Limitations of the structural functionalist model. The structural functionalist perspective has been challenged on many grounds. The most common criticism questions its assumption of systemic equilibrium. Critics claim that structural functionalism ignores the processes of social change internal to a social system. They argue that the relationship among the parts of a social system tends to change over time and that the social system that emerges as a result of these changes may differ fundamentally from the original system.

Another criticism of structural functionalism regards its claim that a social system continues to operate as intended if all the parts of the system perform their functions. This assumption fails to take into account the possibility of external shocks to the system. An external shock could alter the pattern of interactions among the system's parts in such a way as to disrupt its stability, leading to a basic restructuring of the system or, possibly, to its disintegration.

Finally, the structural functionalist perspective has been criticized on ideological or political grounds. Critics claim that a structural functionalist perspective portrays the school as an institution that supports and perpetuates the existing social order and its stratification system. They argue that while schools reward students for ability and achievement, they also maintain the influence of ascribed characteristics on adult success. Structural functionalists fail to analyze the extent to which schools preserve a class-based society.

Not only does structural functionalism ignore the way schools perpetuate the status quo; it also fails to explain how ascriptive characteristics mediate the effects of achievement after graduation. By linking occupational success to organizational characteristics of schools and academic achievement, structural functionalism fails to explain the poor fit that often exists between a student's skills and abilities and the student's future place in the labor market. Critics of structural functionalism argue that a job seeker can often negotiate with a prospective employer, and that this process allows an individual's ascribed characteristics and social status to influence job placement.

In short, structural functionalism is generally viewed as a static theory, which only partially describes the interactions in a school system, or between schools and the rest of society. Even when social change is incorporated into the model, the change is not seen as leading to a radical transformation of the school. As a result, the theory fails to depict the more dynamic and controversial dimensions of schooling. Nevertheless, structural functionalism has been a useful theoretical

perspective to explain how schools and classrooms function as social systems under certain conditions in society.

3.2 Conflict Perspective on the School as a Social System

Conflict theory (Bowles and Gintis 1976, Collins 1971) is an alternative theoretical model for the analysis of the school as a social system. Conflict theory posits competition as the major force driving societal development. While structural functionalism views technological needs and economic growth as the major influences on society, conflict theory argues that competition for wealth and power is the primary state of society and social change is its inevitable result.

According to conflict theory, ascribed characteristics are the basis of elite status. Collins (1971) argues that the continuing effort of the elite to exert control over lower status groups creates an ongoing struggle for power and prestige. Conflict theory specifies conditions under which subordinates are likely to resist the domination of superiors through noncompliance and resistance. Under conditions of economic hardship, political turmoil, or cultural conflict, nonelites are more likely to resist domination and to challenge the relationship that exists between education and occupation. Their discontent typically precipitates social change.

Conflict theory not only explains the relationship between education and occupation; it also yields insights into the power struggles that occur between schools and other social groups in society. For example, during the student movement in the USA in the 1960s and 1970s, students resisted authority and the status quo. Tensions and disruptions continued until students were granted a greater voice in university affairs and in political life. Current controversies about school prayer, sex education, curriculum content, racial integration, and school vouchers are led by competing interest groups. These controversies are likely to lead to compromises that involve some redistribution of authority and power.

The power struggle that occurs in society may also be observed in the school and in the classroom. In schools, students create their own value system that may be inconsistent with the school's academic goals. The tests and grades administered by schools as a sorting mechanism may be viewed as a way to maintain the status quo, to enforce discipline and order, or to co-opt the most intelligent of the lower classes. In the classroom, students may challenge teacher authority and negotiate teacher power through resistance.

By directly addressing the relationship between conflict and social change, 'conflict theory' supplements structural functionalism in explaining the behavior of schools in society and in predicting social change. Both theories point to important aspects of the dynamics of social systems. As stressed by structural functionalists, schools do socialize and allocate students, mostly by meritocratic criteria. But conflict theorists are correct in maintaining that nonmeritocratic factors also influence the allocation process. Structural functionalists are accurate in stating that students typically cooperate with teachers in the learning process, but conflict theorists recognize that some students resist authority and rebel. Further, conflict occurs in communities, schools, and classrooms, but it is not always class-based. Relying on the insights of both theories increases our understanding of schools in society.

4. Coleman's Analysis of the School as a Social System

In his last major work on social theory, Coleman (1990) argued explicitly that the aim of sociology is to explain the behavior of social systems. He pointed out that social systems might be studied in two ways. In the first approach, the system is the unit of analysis and either the behavior of a sample of social systems are studied or the behavior of a single social system is observed over time. Coleman's (1961) study of the adolescent subculture follows this strategy. The second approach involves examining internal processes in a social system, including the relationships that exist among parts of the system. Research on the association between track level in high school and a student's growth in academic achievement is an example of linking a sub-unit of a school to an individual. Studies of the transition from school to work illustrate the relationship between two subunits of society, education and the labor market.

Social system analysis involves explaining three transitions: macro to micro, micro to micro, and micro to macro level transitions (Alexander et al. 1987, Collins 1981, Knorr-Cetina and Cicourel 1981). The macro to micro transition involves the effects of the social system itself on subunits (typically individuals) in the system. An analysis of the effects of ability group level on students' educational aspirations is an example. The micro to micro transition links characteristics of subunits in the social system to the behavior of those subunits. An example is an analysis of the effects of gender on achievement. Finally, the micro to macro transition pertains to how the behavior of subunits in a social system influences the social system as a whole. Research on the effects of student achievement on the way the school organizes students for instruction is illustrative.

Coleman argued that the micro to macro transition is the most difficult to analyze, because it requires specifying the interdependency among subunits or individuals in a social system. His theory of purposive action provides a way to model this transition. In general, Coleman's insistence that the study of social

systems includes an explicit focus on the transitions that exist between macro level and micro level processes in the social system promises to draw greater attention to the social system approach to the study of schooling in society.

5. Conclusions

Conceptualizing the school as a social system is a useful approach to the study of schools. The social system model has led to new theoretical insights about how education, as an institution, affects other societal institutions. It has also generated a significant body of empirical research that demonstrates the interdependence of sub-units in a school and of schools within larger organizational units and their effects on social outcomes. These studies have yielded a better understanding of the role schools play and the contribution they make to contemporary social life.

See also: Coleman, James Samuel (1926–95); Educational Institutions and Society; Socialization in Adolescence; Socialization in Infancy and Childhood; Socialization, Sociology of; System: Social

Bibliography

Alexander J, Giesen B, Munch R, Smelser N 1987 *The Micro-macro Link*. University California Press, Berkeley, CA

Bowles S, Gintis H 1976 *Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life*. Basic Books, New York

Coleman J S 1961 *The Adolescent Society*. Free Press, New York

Coleman J S 1990 *The Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Collins R 1971 Functional and conflict theories of educational stratification. *American Sociological Review* **36**: 1002–19

Collins R 1981 On the micro-foundations of macro-sociology. *American Journal of Sociology* **86**: 984–1015

Durkheim E 1956 *Education and Sociology*. Free Press, New York

Gamoran A, Secada W, Marrett C 2000 The organizational context of teaching and learning: Changing theoretical perspectives. In: Hallinan M (ed.) *Handbook of the Sociology of Education*. Kluwer/Academic/Plenum, New York, Chap. 2, pp. 37–64

Gordon C 1957 *The Social System of the High School*. Free Press, Glencoe, IL

Hallinan M 1994 Tracking: From theory to practice. *Sociology of Education* **67**(2): 79–84, 89–91

Knorr-Cetina K, Cicourel A (eds.) 1981 *Advances in Social Theory and Methodology*. Routledge and Kegan Paul, London

Loomis C P, Dyer E 1976 Educational social systems. In: Lommis C P (ed.) *Social Systems: The Study of Sociology*. Schenkman, Cambridge, MA, Chap. 7

Lynch K 2000 Research and theory on equality and education. In: Hallinan M (ed.) *Handbook of the Sociology of Education*. Kluwer/Academic/Plenum, New York, Chap. 4, pp. 85–106

Oakes J 1994 More than misapplied technology: A normative and political response to Hallinan on tracking. *Sociology of Education* **67**(2): 84–9, 91

Parsons T 1951 *The Social System*. Free Press, New York

Parsons T 1956 *Economy and Society: A Study in the Integration of Economic and Social Theory*. Free Press, Glencoe, IL

Parsons T 1959 The school class as a social system: Some of its functions in American society. *Harvard Educational Review* **29**: 297–318

Parsons T 1961 An outline of the social system. In: Parsons T, Shiks E, Naegele K, Pitts R (eds.) *Theories of Society*. Free Press, New York, Vol. 1

Smelser N 1998 Social structure. In: Smelser N (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA, pp. 103–29

M. Hallinan

School Effectiveness Research

1. School Effectiveness and School Effectiveness Research

In the most general sense 'school effectiveness' refers to the level of goal attainment of a school. Although average achievement scores in core subjects, established at the end of a fixed program are the most probable 'school effects,' alternative criteria like the responsiveness of the school to the community and the satisfaction of the teachers may also be considered.

Assessment of school effects occurs in various types of applied contexts like the evaluation of school improvement programs or comparing schools for accountability purposes, by governments, municipalities, or individual schools.

School effectiveness research attempts to deal with the causal aspects inherent in the effectiveness concept by means of scientific methods. Not only are assessment of school effects considered, but particularly the attribution of differences in school effects to malleable conditions. Usually, school effects are assessed in a comparative way, e.g., by comparing average achievement scores between schools. In order to determine the 'net' effect of malleable conditions, like the use of different teaching methods or a particular form of school management, achievement measures have to be adjusted for intake differences between schools. For this purpose student background characteristics like socioeconomic status, general scholastic aptitudes, or initial achievement in a subject are used as control variables. This type of statistical adjustment in research studies has an applied parallel in the strive for 'fair comparisons' between schools, known under the label of 'value-added.'

2. Strands of Educational Effectiveness Research

School effectiveness research has developed as a gradual integration of several research traditions. The roots of current 'state-of-the-art' school effectiveness

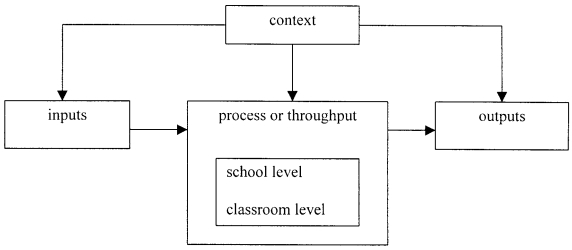

Figure 1
A basic systems model of school functioning

research are sketched by briefly referring to each of these research traditions.

The elementary design of school effectiveness research is the association of hypothetical effectiveness-enhancing conditions of schooling and output measures, mostly student achievement. A basic model from systems theory, where the school is seen as a black box, within which processes or 'throughput' take place to transform inputs into outputs. The inclusion of an environmental or context dimension completes this model (see Fig. 1). The major task of school effectiveness research is to reveal the impact of relevant input characteristics on output and to 'break open' the black box in order to show which process or throughput factors 'work,' next to the impact of contextual conditions. Within the school it is helpful to distinguish a school and a classroom level and, accordingly, school organizational and instructional processes.

Research tradition in educational effectiveness varies according to the emphasis that is put on the various antecedent conditions of educational outputs. These traditions also have a disciplinary basis. The common denominator of the five areas of effectiveness research that will be distinguished is that in each case the elementary design of associating outputs or outcomes of schooling with antecedent conditions (inputs, processes, or contextual) applies.

The following research areas or research traditions can be distinguished:

(a) Research on equality of opportunities in education and the significance of the school in this.

(b) Economic studies on education production functions.

(c) The evaluation of compensatory programs.

(d) Studies of unusually effective schools.

(e) Studies on the effectiveness of teachers, classes, and instructional procedures.

For a further discussion of each of these research traditions the reader is referred to Scheerens (1999). A schematic characterization of research orientation and disciplinary background is given in Table 1.

3. *Integrated School Effectiveness Research*

In recent school effectiveness studies these various approaches to educational effectiveness have become integrated. Integration was manifested in the conceptual modeling and the choice of variables. At the technical level multilevel analysis has contributed significantly to this development. In contributions to the conceptual modeling of school effectiveness, schools became depicted as a set of 'nested layers' (Purkey and Smith 1983), where the central assumption was that higher organizational levels facilitated effectiveness enhancing conditions at lower levels (Scheerens and Creemers 1989). In this way, a synthesis between production functions, instructional

Table 1
General characteristics of types of school effectiveness research

	Independent variable type	Dependent variable type	Discipline	Main study type
(a) (un)equal opportunities	socioeconomic status and IQ of pupils, material school characteristics	attainment	sociology	survey
(b) production functions	material school characteristics	achievement level	economics	survey
(c) evaluation compensatory programs	specific curricula	achievement level	interdisciplinary pedagogy	quasi-experiment
(d) effective schools	'process' characteristics of schools	achivement level	interdisciplinar pedagogy	case-study
(e) effective instruction	characteristics of teachers, instruction, class organization	achievement level	educational psychology	experiment, observation

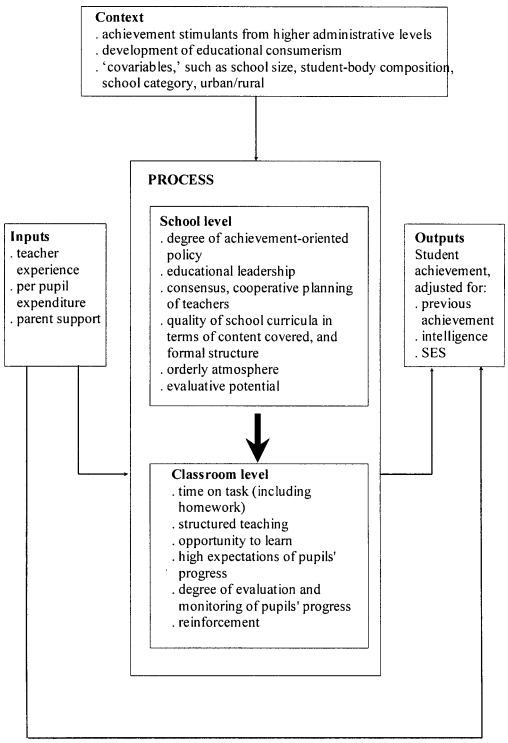

Figure 2
An integrated model of school effectiveness (from Scheerens 1990)

effectiveness, and school effectiveness became possible by including the key variables from each tradition, each at the appropriate 'layer' or level of school functioning (the school environment, the level of school organization and management, the classroom level, and the level of the individual student). Conceptual models that were developed according to this integrative perspective are those by Scheerens (1990), Creemers (1994), and Stringfield and Slavin (1992). The Scheerens model is shown in Fig. 2.

Exemplary cases of integrative, multilevel school effectiveness studies are those by Brandsma (1993), Sammons et al. (1995), and Grisay (1996).

In Table 2 (cited from Scheerens and Bosker 1997) the results of three meta-analyses and a re-analysis of an international data set have been summarized and compared to results of more 'qualitative' review of the research evidence. The qualitative review was based on studies by Purkey and Smith (1983), Levine and Lezotte (1990), Scheerens (1992), and Sammons et al. (1995). The results concerning resource input variables are based on the re-analysis of Hanushek's (1979) summary of results of production function studies that was carried out by Hedges et al. (1994). As stated before this re-analysis was criticized, particularly the unexpectedly large effect of per pupil expenditure.

The results on 'aspects of structured teaching' are taken from meta-analyses conducted by Fraser et al. (1987). The international analysis was based on the IEA Reading Literacy Study and carried out by Bosker (Scheerens and Bosker 1997, Chap. 7). The meta-analysis on school organizational factors, as well as the instructional conditions 'opportunity to learn,' 'time on task,' 'homework,' and 'monitoring at classroom level,' were carried out by Witziers and Bosker and published in Scheerens and Bosker (1997, Chap. 6). The number of studies that were used for these meta-analyses varied per variable, ranging from 14 to 38 studies in primary and lower secondary schools.

The results in this summary of reviews and meta-analyses indicate that resource-input factors on average have a negligible effect, school factors have a small effect, while instructional variables have an average to large effect. The conclusion concerning resource-input factors should probably be modified and somewhat 'nuanced,' given the results of more recent studies referred to in the above, e.g., the results of recent studies concerning class-size reduction.

There is an interesting difference between the relatively small effect size for the school level variables reported in the meta-analysis and the degree of certainty and consensus on the relevance of these factors in the more qualitative research reviews.

It should be noted that the three blocks of variables depend on types of studies using different research methods. Education production function studies depend on statistics and administrative data from schools or higher administrative units, such as districts or states. School effectiveness studies focusing at school level factors are generally carried out as field studies and surveys, whereas studies on instructional effectiveness are generally based on experimental designs.

4. *Foundational and Fundamental School Effectiveness Studies*

Foundational school effectiveness studies refer to basic questions about the scope of the concept of school effectiveness. Can a school be called effective on the basis of achievement results measured only at the end of a period of schooling, or should such a school be expected to have high performance at all grade levels? Can school effectiveness be assessed by examining results in just one or two school subjects, or should all subject matter areas of the curriculum be taken into account? Also, shouldn't one restrict the qualification of a school being effective to consistently high performance over a longer period of time, rather than a 'one shot' assessment at just one point in time?

Fortunately all of these questions are amenable to empirical research. These type of studies that are associated with the consistency of school effects over grade-levels, teachers, subject-matter areas, and time

Table 2
Reviews of the evidence from qualitative reviews, an international study, and research syntheses; coefficients are correlations with student achievement (the plus signs in the table refer to a positive assessment of the variable in question in the reviews).

	Qualitative reviews	International analyses	Research syntheses
Resource input variables:			
Pupil–teacher ratio		−0.03	0.02
Teacher training		0.00	−0.03
Teacher experience			0.04
Teachers' salaries			−0.07
Expenditure per pupil			0.20
School organizational factors:			
Productive climate culture	+		
Achievement pressure for basic subjects	+	0.02	0.14
Educational leadership	+	0.04	0.05
Monitoring/evaluation	+	0.00	0.15
Cooperation/consensus	+	−0.02	0.03
Parental involvement	+	0.08	0.13
Staff development	+		
High expectations	+	0.20	
Orderly climate	+	0.04	0.11
Instructional conditions:			
Opportunity to learn	+	0.15	0.09
Time on task/homework	+	0.00/−0.01 (n.s.)	0.19/0.06
Monitoring at classroom level	+	−0.01 (n.s.)	0.11 (n.s.)
Aspects of structured teaching:			
—cooperative learning			0.27
—feedback			0.48
—reinforcement			0.58
Differentiation/adaptive instruction			0.22

have been referred to as 'foundational studies' (Scheerens 1993) because they are aimed at resolving issues that bear upon the scope and 'integrity' of the concept of school effectiveness.

A recent review of such foundational studies is given in Scheerens and Bosker (1997, Chap. 3). Their results concerning primary schools are presented in Table 3. Consistency is expressed in terms of the correlation between two different rank orderings of schools. Results are based on arithmetic and language achievement.

Table 3
Consistency of school effectiveness (primary level)

Type of consistency	Average correlation
across time (stability) (1 or 2 years)	$r = 0.70$ (0.34–0.87)
across grades	$r = 0.50$ (0.20–0.69)
across subjects	$r = 0.70$ (0.59–0.83)

Source: Scheerens and Bosker (1997, Chap. 3)

The results summarized in Table 3 indicate that there is a reasonable consistency across cohorts and subjects, while the consistency across grades is only average. Results measured at the secondary level likewise show reasonably high stability coefficients (consistency across cohorts), and somewhat lower coefficients for stability across subjects (e.g., in a French study (Grisay 1996), coefficients based on value-added results were 0.42 for French language and 0.27 for mathematics). The average consistency between subjects at the secondary level was somewhat lower than in the case of primary schools (r about 0.50). This phenomenon can be explained by the fact that, at the secondary level, different teachers usually teach different subjects, so that inconsistency is partly due to variation between teachers.

The few studies in which factor analysis was used to examine the size of a stable school factor relative to year specific and subject specific effects have shown results varying from a school factor explaining 70 percent of the subject and cohort specific (gross) school effects, to 25 percent.

The picture that emerges from these studies on the stability and consistency of school effects is far from

being indifferentially favorable with respect to the unidimensionality of the school effects concept.

Consistency is fair, when effects at the end of a period of schooling are examined over a relatively short time interval. When grade-level and subject matter area is brought into the picture, consistency coefficients tend to be lower, particularly when different teachers teach different grades or subjects. School effects are generally seen as teacher effects, especially at the secondary school level

The message from these 'foundational studies' is that one should be careful not to overgeneralize the results of school effectiveness studies when only results in one or two subject matter areas at one point in time are measured. Another implication is that hypothetical antecedent conditions of effects are not only to be sought at the school organizational level, but also at the level of teaching and the teaching and learning process.

Fundamental school effectiveness studies are theory- and model-driven studies. Bosker and Scheerens (1994) presented alternative causal specifications of the conceptual multilevel models referred to in an earlier section. These models attempt to grasp the nature of the relationships between e.g., schools and classroom conditions. For example, whether such relationships are additive, interactive, reciprocal, or form a causal chain.

Other studies that have attempted to formalize these types of relationships are by Hofman (1995) and Creemers and Reezigt (1996). In general, it appeared to be difficult to establish the better 'fit' of one of the alternative model specifications. More complex models, based on the axioms of microeconomic theory, have been tested by de Vos (1998), making use of simulation techniques.

So far, these studies are too few to draw general conclusions about the substantive outcomes; continuation of this line of study is quite interesting, however, also from a methodological point of view.

From a substantive point of view educational effectiveness studies have indicated the relatively small effects of these conditions in developed countries, where provisions are at a uniformly relatively high level. At the same time the estimates of the impact of innate abilities and socioeconomic background characteristics—also when evaluated as contextual effects—seem to grow, as studies become more methodologically refined. Given the generally larger variation in both conditions and outcomes of schooling in developing countries—and the sometimes appallingly low levels of both—there is both societal and scientific relevance in studying school effectiveness in these settings.

5. *The Future of School Effectiveness Research*

From this article it could be concluded that school effectiveness research could be defined in a broad and a narrow sense. In the broadest sense one could refer to all types of studies which relate school and classroom conditions to achievement outcomes, preferably after taking into account the effects of relevant student background variables. In a narrower sense, state-of-the-art integrative school effectiveness studies and foundational and fundamental studies could be seen as the core.

Following the broader definition the future of school effectiveness studies is guaranteed, particularly in the sense of 'applied' studies, like cohort studies, large-scale effect studies carried out for accountability purposes, monitoring studies, and assessment studies.

State-of-the-art, fundamental and foundational school effectiveness studies are a much more vulnerable species. One of the major difficulties is the organizational complexity and costs of the 'state-of-the-art types' of study. Given the shortage of these kind of studies the more fundamental and foundational types of studies are likely to be dependent on the quality of data-sets that have been acquired for 'applied purposes.'

Therefore the best guarantee for continued fundamental school effectiveness research lies in the enhanced research technical quality of applied studies. One example consists of quasi-experimental evaluation of carefully designed school improvement projects. Another important development is the use of IRT (Item Response Theory) modeling and 'absolute' standards in assessment programs. If school effects can be defined in terms of distance or closeness in average achievement to a national or even international standard, some of the interpretation weaknesses of comparative standards belong to the past.

See also: Educational Assessment: Major Developments; Program Evaluation; School Improvement; School Management; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values; Teacher Behavior and Student Outcomes

Bibliography

Bosker R J, Scheerens J 1994 Alternative models of school effectiveness put to the test. In: Bosker R J, Creemers B P M, Scheerens J (eds.) *Conceptual and Methodological Advances in Educational Effectiveness Research*, special issue of the *International Journal of Educational Research* **21**: 159–80

Brandsma H P 1993 *Basisschoolkenmerken en de Kwaliteit van het Onderwijs* [*Characteristics of primary schools and the quality of education*]. RION, Groningen, The Netherlands

Creemers B P M 1994 *The Effective Classroom*. Cassell, London

Creemers B P M, Reezigt G J 1996 School level conditions affecting the effectiveness of instruction. *School Effectiveness and School Improvement* **7**: 197–228

Fraser B L, Walberg H J, Welch W W, Hattie J A 1987 Syntheses of educational productivity research. Special Issue of the *International Journal of Educational Research* **11**

Grisay A 1996 *Evolution des acquis cognitifs et socio-affectifs des eleves au cours des annees de college* [*Evolution of cognitive and socio-affective outcomes during the years of secondary education*]. Université de Liège, Liège, Belgium
Hanushek E A 1979 Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources* **14**: 351–88
Hedges L V, Laine R D, Greenwald R 1994 Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher* **23**: 5–14
Hofman W H A 1995 Cross-level relationships within effective schools. *School Effectiveness and School Improvement* **6**: 146–74
Levine D U, Lezotte L W 1990 *Unusually Effective Schools: A Review and Analysis of Research and Practice*. National Center for Effective Schools Research and Development, Madison, WI
Purkey S C, Smith M S 1983 Effective schools: A review. *The Elementary School Journal* **83**: 427–52
Sammons P, Hillman J, Mortimore P 1995 *Key Characteristics of Effective Schools: A Review of School Effectiveness Research*. OFSTED, London
Scheerens J 1990 School effectiveness and the development of process indicators of school functioning. *School Effectiveness and School Improvement* **1**: 61–80
Scheerens J 1992 *Effective Schooling. Research, Theory and Practice*. Cassell, London
Scheerens J 1993 Basic school effectiveness research: Items for a research agenda. *School Effectiveness and School Improvement* **4**: 17–36
Scheerens J 1999 *School Effectiveness in Developed and Developing Countries: A Review of the Research Evidence*. World Bank paper, Washington, DC
Scheerens J, Bosker R J 1997 *The Foundations of Educational Effectiveness*. Elsevier Science Ltd., Oxford, UK
Scheerens J, Creemers B P M (eds.) 1989 Developments in school effectiveness research. Special themed issue of the *International Journal of Educational Research* **13**
Stringfield S C, Slavin R E 1992 A hierarchical longitudinal model for elementary school effects. In: Creemers B P M, Reezigt G J (eds.) *Evaluation of Effectiveness*. ICO-Publication 2. GION, Groningen, The Netherlands
de Vos H 1998 *Educational Effects: A Simulation-Based Analysis*. University of Twente, Enschede, The Netherlands

J. Scheerens

School Improvement

1. School Improvement

Different terms are used to describe change processes in schools and educational systems as a whole. They include: 'educational change,' 'innovation,' 'reform,' and development.' These terms are quite often used in an interchangeable manner, and their meaning in the literature is frequently ambiguous. Many authors do not define them, but instead tend to trust everyday usage to convey the meaning of the words. Other authors define the concept by including in the definition important conditions and crucial characteristics of the processes.

In the definition of these different concepts generally, and to distinguish school improvement from other concepts, at least three points should be taken into consideration.

(a) The level in educational system where the change takes place: the school, the district, or the state. Ultimately, every change process has to reach the classroom and the level of the individual student. In order to make a distinction, educational change at national level is often called 'systemic reform.' The term 'school improvement' makes clear that the change takes place at school level.

(b) In all definitions, distinctions should be made between, on the one hand, the process of educational change, the strategies used, and the characteristics of the change process, and, on the other, the outcomes of the change process. In everyday use, the term can refer to one of these aspects or even to a combination of them: for example, to strategy and the product of change.

Fullan clarifies what he calls 'the confusion between the terms *change* and *progress*' (Fullan 1991, p. 4). Not every change is progress: the actual outcome of the change processes in terms of progress is not as important as the intended outcomes in the desired direction.

(c) When it comes to the outcomes of change, there is often no well-defined concept of what the change is aimed at: the context of education, the inputs and processes, or, ultimately, the outcomes in terms of student achievement.

The International School Improvement Project (ISIP) defines school improvement as a 'a systematic sustained effort aimed at change in learning conditions and other related internal conditions in one or more schools, with the ultimate aim of accomplishing educational goals more effectively' (Van Velzen et al. 1985). This definition specifies improvement as an innovation or planned change with specific means (change in learning and other internal conditions of the school) and specific goals (ultimately to accomplish educational goals more effectively).

2. Practice and Theory in School Improvement

School improvement as a process has been demarcated into three global phases: the initiation, the implementation, and the continuation processes. Stakeholders of school improvement are more narrowly circumscribed and consist of students, teachers, principals, and parents at school level, and of local educational authorities, consultants, and the local community at local level. In exceptional cases, where it is a regional or national strategy to improve all schools, the stakeholders at these levels are also relevant.

School improvement is a very widespread phenomenon and a wide variety of improvement efforts can be found. For example, there is a lot of improvement going on which does not aim at enhancing student outcomes at all. These types of improvement focus, for instance, on the career development of teachers, on restructuring the organization of the school, on the way decisions are made, or on the relationships between schools and their clients. Sometimes restructuring takes place at classroom level. Changes in the writing practices of elementary schoolteachers, for example, have been described in great detail. Nevertheless, the actual impact of the changes on students and on student achievement is usually left out.

Some schools practice improvement on their own and try to find their own solutions to their problems. Other schools have chosen to implement as accurately as possible improvement programs that have been developed elsewhere. This is called fidelity implementation. Often they do not consider alternative educational options. For example, they may opt for a specific program primarily because another school was satisfied with it (Stringfield 1995). Some schools are involved in improvement only because their government expects them to be (Hopkins et al. 1994). Depending on the extent to which educational policies are translated into clear outlines and contents, fidelity implementation is a more or less appropriate concept. When the educational policy is rather prescriptive with respect to curricular content and outcome levels, as it is in the UK (Hopkins et al. 1994), it is theoretically possible to check whether or not schools are implementing this policy. When the educational policy is rather open-ended, as is the Dutch policy on educational priorities (Van der Werf 1995) or the Dutch inclusion policy, it is virtually impossible to apply the notion of fidelity implementation.

Only a small part of school improvement is based on research (Stringfield 1995). Innovations are hardly ever tested before they are implemented in educational practice, and an adequate *evaluation* of their impact is rare. The same applies to the occurrence of experimental or quasi-experimental designs in improvement projects. Some school improvers have preferred forms of action research instead of research-based experiments. Sometimes projects based on changing only a couple of factors report great successes. It may be that changes in just a few important areas can alter a whole system, but the question is whether these changes—often cases of educational mass hysteria—will stand the test of time. In education, the margins for change are generally small, because of the impact of non-educational factors such as the individual characteristics of students.

Evaluations are often not carried out satisfactorily or are performed after a timespan that is too short. School improvement projects have shown that it takes time to design, develop, implement, and evaluate changes in schools—more time than is often available. Because of inadequate evaluations, questions about the causation of effects, extremely important for the school-effectiveness knowledge base, cannot be answered. School improvement, therefore, should consistently pursue assessment of its results, pay attention to its failures and try to learn from these, and limit its goals to prevent confusion of cause and effect (Hopkins et al. 1999).

In school improvement, the orientation towards educational practice and policy-making is emphasized. Although, in educational practice schools and classrooms can be found that succeed much better than others, theories cannot be based on exemplary practice only.

Insofar as theoretical notions have been developed, they have not yet been empirically and systematically tested. The typologies of school cultures (Hargreaves 1995), for example, have not been studied in educational practice, and their effects on the success of school improvement are as yet unknown. Moreover, their relationships to the first criterion for school improvement (enhancing student outcomes) are not always very clear. The same holds for the factors that are supposed to be important in different stages of educational change, outlined by Fullan (1991), and for his ideas about the essential elements in educational change at classroom level (beliefs, curriculum materials, and teaching approaches). Even though these ideas are derived from school improvement practice, their importance and their potential effects are not accounted for in detail, and have not yet been studied in research.

A contribution to theory development is the generic framework for school improvement provided by Hopkins (1996). In this framework, three major components are depicted: educational givens, a strategic dimension, and a capacity-building dimension. Educational givens cannot be changed easily. Givens can be external to the school (such as an external impetus for change) and internal (such as the school's background, organization, and values). The strategic dimension refers to the competency of a school to set its priorities for further development, to define student learning goals and teacher development, and to choose a strategy to achieve these goals successfully. The capacity-building dimension refers to the need to focus on conditions for classroom practice and for school development during the various stages of improvement. Finally, the school culture has a central place in the framework. Changes in the school culture will support teaching–learning processes which will, in turn, improve student outcomes (Hopkins 1996).

Despite the obvious gaps in theory development and testing in the field of school improvement, there are already some elements of a knowledge base. By trying to improve schools, knowledge on the implementation of classroom and school effectiveness factors in educational practice has become available. This has provided the possibility of studying, to

different degrees, the influence of factors and variables on educational outcomes (Stoll and Fink 1994).

3. *The Link Between School Improvement and School Effectiveness*

School effectiveness and school improvement conjoin because of their mutual interest, although their actual relationship may be very complicated (Reynolds et al. 1993, Creemers and Reezigt 1997). The major aim in the field of school effectiveness was always to link theory and research on the one hand, and practice and policy-making on the other. School improvement can be fostered by a knowledge base covering what works in education that can be applied in educational practice. The combination of theory, research, and development is not new in education. Almost all movements start out to make knowledge useful for educational practice and policy-making, or state their goal in terms of supplementing policy practice with a knowledge base supplied by theory and research from a cyclical point of view. The next step is to use practical knowledge for further advances in theory and research. In this way, research and improvement can have a relationship as a surplus benefit for both.

School effectiveness has led to major shifts in educational policy in many countries by emphasizing the accountability of schools and the responsibility of educators to provide all children with opportunities for high achievement, thereby enhancing the need for school improvement (Mortimore 1991). School effectiveness has pointed to the need for school improvement, in particular by focusing on alterable school factors. School improvement projects were necessary to find out how schools could become more effective. These projects were often supposed to implement effective school factors in educational practice (Scheerens 1992) and, in doing so, could yield useful feedback for school effectiveness. School improvement might point to inaccurate conceptions of effectiveness, such as the notion of linearity or one-dimensionality (Hargreaves 1995). In addition, school improvement might give more insight into the strategies for changing schools successfully in the direction of effectiveness.

The relatively short history of school effectiveness and improvement shows some successes of this linkage. Research results are being used in educational practice, sometimes with good results. School improvement findings are sometimes used as input for new research. Renihan and Renihan state that 'the effective schools research has paid off, if for no other reason than that it has been the catalyst for school improvement efforts' (Renihan and Renihan 1989, p. 365).

Most authors, however, are more skeptical (Reynolds et al. 1993). Fullan states that school effectiveness 'has mostly focused on narrow educational goals, and the research itself tells us almost nothing about how an effective school got that way and if it stayed effective' (Fullan 1991, p. 22). Stoll and Fink (1992) think that school effectiveness should have done more to make clear how schools can become effective. According to Mortimore (1991), a lot of improvement efforts have failed because research results were not translated adequately into guidelines for educational practice. Changes were sometimes forced upon a school, and when the results were disappointing the principals and teachers were blamed. Teddlie and Roberts (1993) suggest that effectiveness and improvement representatives do not cooperate automatically, but tend to see each other as competitors. Links between school effectiveness and school improvement were stronger in some countries than in others (Reynolds 1996). In the early years of school effectiveness, links were strong in the USA and never quite disappeared there. Many districts have implemented effective schools programs in recent years, but research in the field has decreased at the same time, and, because of this, school improvement is sometimes considered 'a remarkable example of (…) over-use of a limited research base' (Stringfield 1995, p. 70).

Reynolds et al. (2000) came to the conclusion based on the analyses of Dutch, British, and North American initiatives that the following principles are fundamental to a successful merger of school effectiveness and school improvement:

(a) a focus on teaching, learning, and the classroom level;
(b) use of data for decision making;
(c) a focus on pupil outcomes;
(d) addressing schools' internal conditions;
(e) enhanced consistency (through implementation of 'reliable' programs); and
(f) pulling levers to affect all levels, both within and beyond the school.

Recently, several projects (Stoll et al. 1996, Hill 1998) have started to integrate school effectiveness and school improvement. They form successful examples of the concept of sustained interactivity. These projects all share a clear definition of the problem that should be overcome, in terms of student outcomes and classroom strategies, to enhance these outcomes within the context of the school. Often, the outcomes are clearly specified for one school subject or elements of a school subject, such as comprehensive reading. The content of the projects is a balanced mix of the effectiveness knowledge base and the concepts from school improvement. The projects have detailed designs, both for the implementation of school improvement and for evaluation in terms of empirical research. By means of a research component integrated into the projects right from the start, it is possible to test effectiveness hypotheses, and to evaluate improvement outcomes at the same time. The use of control groups is essential in this respect, and various projects now incorporate control groups or choose to compare

their results to norm groups on the basis of nationwide tests. Also, many projects are longitudinal in their designs. Although most integrated projects have started recently, some of them have been running for almost a decade now, and they have been disseminated to various educational contexts. Therefore, long-term effects and context-specific effects can easily be tracked by means of follow-up measurement.

An additional feature of projects which last for several subsequent years is the possibility of testing the effectiveness of school improvement strategies and changing strategies whenever necessary. The Halton's Effective Schools Project illustrates this feature clearly (Stoll and Fink 1996, Stoll et al. 1996). The project started in Canada in 1986 with the intention of implementing British effectiveness knowledge in Halton district schools. It soon became clear that the effectiveness knowledge base in itself would not automatically lead to changes in educational practice. Over the years, the project paid a lot of attention to questions on the processes of change in schools. It focuses on the planning process in schools, the teaching and learning processes in classrooms, and staff development. Successful changes turned out to be enhanced by a collaborative school culture, a shared vision of what the school will stand for in the future, and a climate in which decisions are made in a transparent way. Based on their Halton experiences, Stoll and Fink (1996) have developed a conceptual model which links school effectiveness and school improvement through the school development planning process. The model blends the school effectiveness knowledge base with knowledge about change processes in schools. The school development planning process is at the center of the model. The process is considered to be multilayered. Two outer layers comprise invitational leadership, and continuing conditions and cultural norms. The inner cycle layer is formed by the ongoing planning cycle of assessment, planning, implementation, and evaluation of educational processes. The two central core layers refer to a strong focus on the teaching–learning processes and the curriculum, and the students in the school. The school development planning process is influenced by the context of the school and foundations such as research findings, and, in turn, influences intermediate outcomes at teacher and school levels, as well as student outcomes. Finally, the planning process is influenced by several partners (external agencies, educational networks).

In the UK, 66 percent of improvement programs now pursue goals which fit into the school effectiveness tradition of student outcomes (Reynolds et al. 1996). However, it is still not clear whether real improvements will always occur. The IQEA project (Improving the Quality of Education for All, Hopkins et al. 1994) started in 1991 as a staff development project. Gradually, a focus on classroom improvement and its effects on student achievement took over. The project does not stop with the implementation of priorities for development, but also pays explicit attention to conditions that will sustain the changes. These are staff development, involvement, leadership, coordination, inquiry and reflection, and collaborative planning (Stoll et al. 1996). In addition, when the classroom became the center of attention, the project specified the classroom-level conditions that are necessary for effective teaching and student achievement. These are authentic relationships, rules and boundaries, teacher's repertoire, reflection on teaching, resources and preparation, and pedagogic partnerships. Other promising projects in the UK are the Lewisham School Improvement Project and the Hammersmith and Fulham LEA Project. Both projects actively try to enhance student achievement by means of school effectiveness knowledge, and both projects are cooperating with a research institute (Reynolds et al. 1996).

Recent theories about school improvement stress the self-regulation of schools. Self-regulation assumes target setting and embodies the behavioral concept of *mechanisms, feedback, and reinforcement*. This self-regulatory approach to school improvement can be combined with the analogous self-regulatory feedback loops in educational effectiveness, which are again further elaborated in the upward spiraling school development planning process (Stoll and Fink 1996, Hopkins 1995).

The analysis of school improvement efforts over a period of time resulted in a distinction of several types of schools (effective vs. ineffective and improving vs. declining). Stoll and Fink (1998) and Hopkins et al. (1999) give descriptions of those schools and the features of the change processes going on. The declining effective school (the 'cruising school') has received particular attention because it points at the difficulties of keeping educational quality at the same level (Fink 1999).

See also: Educational Evaluation: Overview; Educational Innovation, Management of; Educational Institutions and Society; Educational Policy: Comparative Perspective; Educational Policy, Mechanisms for the Development of; Educational Research and School Reform; School Administration as a Field of Inquiry; School Effectiveness Research; School Management

Bibliography

Creemers B P M, Reezigt G J 1997 School effectiveness and school improvement: Sustaining links. *School Effectiveness and School Improvement* **8**: 396–429

Fink D 1999 The attrition of change: A study of change and continuity. *School Effectiveness and School Improvement* **10**: 269–95

Fullan M G 1991 *The New Meaning of Educational Change*. Teachers College Press, New York

Hargreaves D H 1995 School culture, school effectiveness and school improvement. *School Effectiveness and School Improvement* **6**: 23–46
Hill P W 1998 Shaking the foundations: Research driven school reform. *School Effectiveness and School Improvement* **9**: 419–36
Hopkins D 1995 Towards effective school improvement. *School Effectiveness and School Improvement* **6**: 265–74
Hopkins D 1996 Towards a theory for school improvement. In: Gray J, Reynolds D, Fitz-Gibbon C, Jesson D (eds.) *Merging Traditions: The Future of Research on School Effectiveness and School Improvement*. Cassell, London, pp. 30–51
Hopkins D, Ainscow M, West M. 1994 *School Improvement in an Era of Change*. Teachers College Press, New York
Hopkins D, Reynolds D, Gray J 1999 Moving on and moving up: Confronting the complexities of school improvement in the improving schools project. *Educational Research and Evaluation* **5**: 22–40
Mortimore P 1991 School effectiveness research: Which way at the crossroads? *School Effectiveness and School Improvement* **2**: 213–29
Renihan F I, Renihan P J 1989 School improvement: Second generation issues and strategies. In: Creemers B, Peters T, Reynolds D (eds.) *School Effectiveness and School Improvement*. Swets and Zeitlinger, Amsterdam/Lisse, pp. 365–77
Reynolds D 1996 County reports from Australia, the United Kingdom, and the United States of America. Introduction and overview. *School Effectiveness and School Improvement* **7**: 111–13
Reynolds D, Hopkins D, Stoll L 1993 Linking school effectiveness knowledge and school improvement practice: Towards a synergy. *School Effectiveness and School Improvement* **4**: 37–58
Reynolds D, Sammons P, Stoll L, Barber M, Hillman J 1996 School effectiveness and school improvement in the United Kingdom. *School Effectiveness and School Improvement* **7**: 133–58
Reynolds D, Teddlie C, Hopkins D, Stringfield S 2000 School effectiveness and school improvement. In: Teddlie C, Reynolds D (eds.) *The International Handbook of School Effectiveness Research*. Falmer Press, London, pp. 206–31
Scheerens J 1992 *Effective Schooling: Research, Theory and Practice*. Cassell, London
Stoll L, Fink D 1992 Effecting school change: The Halton approach. *School Effectiveness and School Improvement* **3**: 19–42
Stoll L, Fink D 1994 School effectiveness and school improvement: Voices from the field. *School Effectiveness and School Improvement* **5**: 149–78
Stoll L, Fink D 1996 *Changing our Schools: Linking School Effectiveness and School Improvement*. Open University Press, Buckingham, UK
Stoll L, Fink D 1998 The cruising school: The unidentified ineffective school. In: Stoll L, Myers K (eds.) *No Quick Fixes: Perspectives on Schools in Difficulty*. Falmer Press, London, pp. 189–206
Stoll L, Reynolds D, Creemers B, Hopkins D 1996 Merging school effectiveness and school improvement: Practical examples. In: Reynolds D, Creemers B, Bollen R, Hopkins D, Stoll L, Lagerweij N (eds.) *Making Good Schools, Linking School Effectiveness and School Improvement*. Routledge, London, pp. 113–47
Stringfield S 1995 Attempting to enhance students' learning through innovative programs: The case for schools evolving into High Reliability Organizations. *School Effectiveness and School Improvement* **6**: 67–96
Teddlie C, Roberts S P 1993 *More Clearly Defining the Field: A Survey of Subtopics in School Effectiveness Research*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA
Van der Werf M P C 1995 *The Educational Priority Policy in The Netherlands: Content, Implementation and Outcomes*. SVO, Den Haag, The Netherlands
Van Velzen W G, Miles M B, Ekholm M, Hameyer U, Robin D 1985 *Making School Improvement Work: A Conceptual Guide to Practice*. ACCO, Leuven, Belgium

B. P. M. Creemers

School Learning for Transfer

Most of what is taught in school is expected to affect learning and performance that transcend the mastery of those subjects: one learns addition to prepare the grounds for the study of multiplication, science to understand the surrounding physical and natural world, and history to gain a sense of identity and a deeper understanding of current events. Such expectations concern the transfer of what is learned in school in one domain to other realms of learning and activity, a spillover of learning that functions like a ripple effect. Thus, the study of topic A is supposed to facilitate the study of B, make it easier, faster, or better understood relative to when B is learned without the learning of A preceding it. Two approaches to the study of transfer have dominated the field since the beginning of the twentieth century: transfer as a function of similar elements between A and B, and as a function of the understanding of general principles that are transferable to a variety of situations. The history of research on transfer shows how these two major principles underlie the development of the field, even when other basic assumptions about learning and thinking have recently been challenged.

1. *Two Basic Approaches*

The expectation for transfer has as long a history as institutional learning itself. Plato argued that the study of abstract reasoning assists the solution of daily problems. Similarly, the debates about Talmudic and Biblical texts in ancient times were argued to 'sharpen minds.' The scientific study of transfer has a long history as well; it dates back to the beginning of the twentieth century with the studies by Thorndike and Woodworth (1901) in which the idea that Latin and other taxing school subjects developed one's general 'faculties of mind' was challenged. The expected transfer was not found; mastery of Latin, Greek, or geometry did not facilitate either performance or efficiency in learning other subjects. Thorndike's

findings led to his formulation of the theory of 'identical elements': transfer from A to B is likely to occur the more elements are common to both. It followed that transfer can take place only in the relatively rare case when there is clear and apparent similarity between the constituent elements of the learned topics. Thus, one would have to learn a host of independent issues and procedures as no transfer in the absence of identical elements should be expected.

Thorndike's theory was countered by a more cognitively-oriented and Gestalt-flavored alternative that emphasized the meaningful conceptual understanding of underlying principles which then could be thoughtfully applied to new situations. The more general the learned principle, the farther could it be transferred to new instances. Judd (1908) had his subjects learn to shoot at targets submerged in water after having learned principles of light refraction in water. Having learned these principles, subjects were much better at hitting the targets than those who have only practiced the activity.

As we shall see below, the two approaches—transfer as a matter of common elements and transfer as a matter of higher order cognitive processes of understanding and abstraction—although having undergone profound modifications and developments over the years, continue to this day to dominate the field.

2. *Paucity of Findings*

Research on transfer as well as practitioners' expectations for transfer are characterized by an ongoing pendulum-like oscillation between the two approaches. Such oscillation is not so much the result of evidence that supports both approaches as it is a function of the paucity of findings clearly supporting either one of them. Indeed, one of the hallmarks of the field is the great discrepancy between the expectation for transfer from school learning to the acquisition of new subjects or to the handling of daily events and the findings from controlled transfer studies. The latter often do not yield the transfer findings one would have expected. A typical case is the study by Pea and Kurland (1984) who have found that children learning LOGO programming showed no advantage in their ability to plan over children who did not master LOGO.

Transfer, when at all found in carefully carried out studies, appears to be highly specific, as when a particular procedure or principle is applied in a new situation which bears great and apparent (often analogical) similarity to the learning material, and when the process is accompanied by such facilitation as coaching, specific guidance and cueing (see Bransford and Schwartz (1999, Cox (1997, for detailed reviews). Such findings would seem to support Thorndike's theory and pessimism about transfer, as summarized by Detterman (1993): 'In short, from studies that claim to show transfer and that don't show transfer, there is no evidence to contradict Thorndike's general conclusions: Transfer is rare, and its likelihood of occurrence is directly related to the similarity between the two situations' (p.15).

3. *Transfer on the Rebound*

Whereas Thorndike's pessimism seemed to have won the day for a while, the research tradition of Judd (1908), with its emphasis on comprehension and general (nonspecific) transfer gradually showed its strength. As research of this tradition tended to show, when relevant principles or strategies are mindfully attended to, or better yet—either abstracted by learners and/or metacognitively monitored—transfer to a variety of instances can be obtained even in the absence of apparent common elements. At the same time, other research on the transfer of skills, ostensibly still in a tradition more similar to that originated by Thorndike (Thorndike and Woodworth 1901), has also yielded positive results. Thus, for example, Singley and Anderson (1989) have shown that when training to near automaticity of discrete skill components is carried out, transfer from one task (e.g., a line editor) to another (text editor) can be virtually perfect. Also more field-based research found support for impressive transfer from the study of university-taught disciplines: Nisbett et al. (1993) have shown that the study of psychology, and also to a lesser extent medicine (but not law or chemistry), improve students' abilities to apply statistical and methodological rules of inference to both scientific and daily problems.

4. *The High and the Low Roads to Transfer*

The renewed success of the two lines of research, often addressing higher order cognitions and information processing exercized by multiple activities, suggests that transfer may not be a unitary process as the two approaches differ in important ways. Observing these differences, Salomon and Perkins (1989) have developed a theory to account for the possibility that transfer takes either one of two routes (or a combination thereof), described as the *high road* and the *low road* of transfer. The low road reflects to an extent the Thorndikian line, and more recently that of Anderson's skill-acquisition and transfer theory. It is taken when skills, behaviors, or action tendencies are repeatedly practiced in a variety of situations until they are mastered to near-automaticity and are quite effortlessly applied to situations whose resemblance to the learning situations is apparent and easily perceived. Learning to drive and learning to read are two cases in point as is transfer from one text editor to another as studied by Singely and Anderson (1989). Other candidates for low road transfer are attitudes,

cognitive styles, dispositions, and belief systems the application of which to new situations is rarely a mindful process.

In contrast, the high road to transfer is characterized by the process of mindful abstraction of knowledge elements that afford logical abstraction: principles, rules, concepts, procedures, and the like. It is this mindfully abstracted, decontextualized idea ('ethnic oppression may lead to revolt') that becomes a candidate for transfer from one particular instance (the Austro-Hungarian Empire) to another (Russia and Cechenia). Simon (1980) commented in this respect that 'To secure substantial transfer of skill acquired in the environment of one task, learners need to be made explicitly aware of these skills, *abstracted from their specific task content*' (p. 82, emphasis added).

There is ample evidence to support the role of the high road in obtaining transfer. One of the studies by Gick and Holyoak (1983) illustrates this point. Participants who were given two stories and were asked to write a summary of how the two stories resembled each other (that is, their common moral), showed a 91 percent transfer to the solution of another story, relative to 30 percent transfer of a non-summary group. Bassok (1990), showed that mastering algebraic abstractions (plus examples) allowed students to view physics problems as particular instances to which the more abstract algebraic operations could be applied. Physics, on the other hand, is too particular and thus students do not expect and do not recognize any possible relationship between it and algebraic operations. Research also directs attention to the role played by self-regulation and metacognitions in the process of mindful abstraction and transfer via the high road. Following the studies and review of Campione et al. (1982), Salomon et al. (1989) have shown that students interacting with a semi-intelligent computerized Reading Partner that provides metacognitive-like guidance, tend to internalize that guidance and transfer it to new reading as well as to writing situations.

The high road/low road theory sheds light on the many failures of obtaining transfer in controlled studies. Close examination of such studies suggests that in many cases neither the low road of repeated practice nor the high road of mindful abstraction was taken. Not enough time is allocated for practice for the former, and not enough attention is given for mindful abstraction for the latter. As a consequence, neither near automatic transfer on the basis of easily recognized common elements can be attained, nor farther transfer on the basis of metacognitively guided mindful abstraction.

5. *New Approaches*

Research on transfer, until recently, did not challenge the basic paradigm and conception of transfer as a process involving change of one's performance on a new task as a result of his or her prior performance on a preceding and different task. Such a conception of transfer has recently become challenged on the basis of new theories of learning, the role of activity, and the place of cognitions therein. These challenges can be arranged along a dimension ranging from the least to the more radical, relative to traditional notions of transfer.

Bransford and Schwartz (1999), coming from a mainstream tradition of cognitive science applied to instructional issues, argue that the current ways of demonstrating transfer are more appropriate for studying only full-blown expertise. According to them, 'There are no opportunities for…[students] to demonstrate their abilities to learn to solve new problems by seeking help from other resources such as text or colleagues or by trying things out, receiving feedback, and getting opportunities to revise' (p. 68). They recommend replacing the typical transfer task which they call 'sequestered problem solving' (SPS) by a conception of transfer as 'preparation for future learning' (PFL). Thus, one would measure transfer not by having students apply previously acquired knowledge or skill to new situations, demonstrating either *knowledge how* or *knowledge that* something, but rather by demonstrating *knowledge with*—thinking, perceiving, and judging with whatever knowledge and tools are available even if that knowledge is not consciously recalled. In other words, transfer would be expected to become demonstrated by way of 'people's abilities to learn new information and relate their learning to previous experiences' (Bransford and Schwartz (1999, p. 69).

An illustration of the above can be seen in a study in which college and fifth grade students did not differ in their utilization of previous knowledge and educational experiences for the creation of a statewide plan to recover the bald eagle. This is a typical SPS way of measuring transfer, suggesting in this case that previous educational experiences did not transfer to the handling of the new problem. However, when the students were asked to generate questions that would need to be studied in preparation for the creation of a recovery plan (applying a PFL approach), striking differences were found between the two age groups in favor of the college students. Focusing on different issues, the two groups showed how, knowingly or not, previous learning has facilitated the generation of the list of questions to be asked. The use of tools and information resources would also show how previous learning of skills and knowledge becomes actively involved in the process of solving a new problem.

While Bransford and Schwartz (1999) continue to adhere to the traditional conception of learning as knowledge acquisition and to cognitions as general and transferable tools of the mind, others have developed an alternative, socio-cultural approach according to which neither the mind and its content

(e.g., arithmetic algorithms) should be treated as a neutral toolbox ready for application wherever suitable, nor should knowledge-in-use and social context of activity be taken as two independent entities. Learning is a matter of participation in a community of learners, thus knowledge is not a noun denoting *possession* but rather a verb denoting the active process of *knowing-as-construction within a social context of activity*. One underlying assumption is that learner, activity, and content become united through culturally constructed tools. A second assumption is that learning is highly situated in a particular activity context of participation, and its outcomes are 'becoming better attuned to constraints and affordances of activity systems so that the learner's contribution to the interaction is more successful' (Greeno 1997, p. 12). Seen in this light, the traditional *cognitivist* concern with the transferability of acquired knowledge across tasks becomes, from the newer *situative* perspective, an issue concerned with the task and the 'consistency or inconsistency of patterns of participatory processes across situations' (Greeno (1997, p. 12).

An important practical implication of this approach is that one would need to take into consideration the kinds of participatory activities to which school-based learning might be expected to transfer, rather than expect decontextualized school materials to transfer all on its own to other tasks, situations, and activities. Students are likely to restructure new situations to fit their previous practices and thus make the old available for transfer to the new. Illustrating this approach to learning and cognition, Saxe (1991), showed that Brazilian children's mathematical experience as street vendors affected their math learning when later on they became enrolled in school, and vice versa. '[I]n recurring activities like practices of candy selling and schooling—contexts in which knowledge is well learned and in which similar problems emerge on a repeated basis—transfer is a protracted process. In their repeated efforts to solve practice-linked problems, individuals attempt to structure problem contexts in coherent and more adequate ways' (p. 179).

The new approaches to transfer clearly deviate from the traditional views. They suggest new ways of measuring transfer, and more radically—challenge traditional assumptions by treating knowledge, skill, task, activity, and context as one participatory activity. By so doing, they emphasize the uniqueness of each activity setting. An important aspect of this approach is its ability to incorporate within the idea of participatory activity motivational and dispositional elements also. These elements have so far been neglected by much research on transfer, although they may be crucial for students' choice to treat each situation as different, or as allowing students to exercise familiar patterns of participation. Still, it appears that despite these important novelties, the principles of transfer as a function of similarity between situations (or activities) and of depth of understanding general principles (e.g., of useful participation in a community of learners) have not yet been replaced but only redefined.

See also: Competencies and Key Competencies: Educational Perspective; Learning Theories and Educational Paradigms; Learning to Learn; Metacognitive Development: Educational Implications; School Effectiveness Research; Situated Learning: Out of School and in the Classroom; Thorndike, Edward Lee (1874–1949); Transfer of Learning, Cognitive Psychology of

Bibliography

Bassok M 1990 Transfer to domain-specific problem solving procedures. *Journal of Experimental Psychology: Learning Memory and Cognition* **6**: 522–33

Bransford J, Schwartz D 1999 Rethinking transfer: A simple proposal with multiple implications. *Review of Educational Research* **24**: 61–100

Campione J C, Brown A L, Ferrara R A 1982 Mental retardation and intelligence: Contributions from research with retarded children. *Intelligence* **2**: 279–304

Cox B D 1997 A rediscovery of the active learner in adaptive contexts: A developmental-historical analysis of transfer of training. *Educational Psychologist* **32**: 41–55

Detterman D R 1993 The case for prosecution: Transfer as an epiphenomenon. In: Detterman D K, Sternberg R J (eds.) *Transfer on Trial: Intelligence, Cognition and Instruction*. Ablex, Norwood, NJ

Gick M L, Holyoak K J 1983 Schema induction and analogical transfer. *Cognitive Psychology* **15**: 1–38

Greeno J G 1997 Response: On claims that answer the wrong questions. *Educational Researcher* **26**: 5–17

Judd C H 1908 The relation of special training to general intelligence. *Educational Review* **36**: 28–42

Nisbett R E, Fong G T, Lehman D R, Cheng P W 1993 Teaching reasoning. In: Nisbett R E (ed.) *Rules for Reasoning*. Erlbaum, Hillsdale, NJ

Pea R D, Kurland D M 1984 On the cognitive effects of learning computer programming. *New Ideas in Psychology* **2**: 37–68

Salomon G, Globerson T, Guterman E 1989 The computer as a zone of proximal development: Internalizing reading-related metacognitions from a Reading Partner. *Journal of Educational Psychology* **81**: 620–7

Salomon G, Perkins D N 1989 Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychology* **24**: 113–42

Saxe G B 1991 *Culture and Cognitive Development: Studies in Mathematical Understanding*. Erlbaum, Hillsdale, NJ

Simon H A 1980 Problem solving and education. In: Tuma D T, Reif R (eds.) *Problem Solving and Education: Issues in Teaching and Research*. Erlbaum, Hillsdale, NJ

Singley M K, Anderson J R 1989 *The Transfer of Cognitive Skill*. Harvard University Press, Cambridge, MA

Thorndike E L, Woodworth R S 1901 The influence of improvement in one mental function upon the efficacy of other functions. *Psychological Review* **8**: 247–61

G. Salomon

Copyright © 2001 Elsevier Science Ltd. All rights reserved. 13579

 ISBN: 0-08-043076-7

School Management

School management is in considerable disarray, if not turmoil, and is likely to remain so in the first decade of the twenty-first century. This is partly a reflection of sweeping transformations in society, mirrored in developments in education generally and in schools in particular. It also reflects failure to effect a powerful link between school management as a field of study and school management as a field of practice, so that each informs and influences the other. More fundamental is the concern that, in each instance, management is not connected to learning as well as it should be.

1. *Definition of School Management*

Disarray is evident in the matter of definition. In some nations, notably the USA, administration is used to describe a range of functions performed by the most senior people in an organization, with management a more limited, sometimes routine set of tasks, often performed by those in supporting roles.

A conceptual difficulty relates to the distinction between leadership and management. The work of Kotter (1990) is helpful in resolving the issue. Leadership is concerned with change and entails the three broad processes of establishing direction, aligning people, and motivating and inspiring. Management is concerned with the achievement of outcomes, and the three broad processes of planning and budgeting, organizing and staffing, and controlling and problem-solving.

These difficulties are reflected in the role ambiguity of those who hold senior positions in schools, notably principals or head teachers. There is a generally held view that their work entails both leadership and management. They should certainly be leaders. In some instances managers will support them. There is concern when they act solely as managers. Kotter (1990) offers a contingent view that is as helpful for schools as it is for organizations in general. He contends that the emphasis and balance in leadership and management are contingent on two variables: the amount of change that is required and the complexity of the operation. Complex schools in a turbulent environment require high levels of leadership and management. Schools that are relatively simple but are faced with major change require leadership more than management. Schools that are complex in stable circumstances require management more than leadership. Simple schools in stable settings may require little of each.

2. *School Management as a Field of Practice*

There are three important developments in school management that are evident in almost every nation (Caldwell and Spinks 1998). Schools and systems of schools vary in the extent to which each has unfolded. The first is the shift of authority and responsibility to schools in systems of public education. Centrally determined frameworks of curriculum, standards, and accountabilities are still in force but schools have a considerable degree of freedom in how they will meet expectations. This development is known as school-based management or local management or self-management, and is frequently accompanied by the creation of school councils, school boards, or other structures for school-based decision-making. Comprehensive reform along these lines is most evident in Australia, the UK, Canada, New Zealand and the USA, but most other nations have begun the process or are planning it.

Reasons for decentralization vary but are generally consistent with the changing view of the role of government that gathered momentum in the final decade of the twentieth century. As applied to education, this view holds that governments should be mainly concerned with setting direction, providing resources, and holding schools to account for outcomes. They should also provide schools with the authority and responsibility to determine the particular ways they will meet the needs of their students, within a centrally determined framework.

The second development arises from higher expectations for schools as governments realise that the knowledge and skill of their people are now the chief resource if nations are to succeed in a global economy. There is recognition of the high social cost of failure and the high economic cost of providing a world-class system of education. As a result, there is unprecedented concern for outcomes, and governments have worked individually and in concert to introduce systems of testing at various stages of primary and secondary schooling to monitor outcomes and provide a basis for target setting in bringing about improvement. In several countries, notably the UK and some parts of the USA, rankings of schools are published in newspapers. Comparisons of national performance are now possible through projects such as the Third International Mathematics and Science Study (Martin and Kelly 1996).

The third development is change in the nature of schooling itself. Driven to a large extent by advances in information and communications technology, much of the learning that in the past could only occur in the classroom can now occur at anytime and anywhere there is access to a networked computer. Students and their teachers have access to a vast amount of information. Interactive multi-media have enriched and individualized learning on a remarkable scale. This development is not uniform across all schools or across all nations, and disparities are a matter of concern.

There is growing anxiety on the part of government that structures for the delivery of public education that

have survived for a century or more may no longer be adequate for the task. Charter schools made their appearance in Canada and the USA in the 1990s, and the number increased very rapidly, even though they are still a small fraction of the total number of schools. Charter schools receive funding from the public purse but are otherwise independent, operating outside the constraints of a school district. A schools-for-profit movement gathered momentum around the same time, with public funding enhanced by private investment through the offering of shares on the stock exchange, as instanced by the pioneering Edison Schools. In the UK, the Blair Government privatized the support services of several local education authorities and established a framework for public–private partnerships in the management of publicly owned schools. The number of private schools is increasing in most countries, reflecting growing affluence and loss of faith in the public sector. A contentious policy, gaining ground in a number of countries, is the provision of public funds (often described as 'vouchers') allowing parents to access private schools when public schools do not meet expectations.

Taken together, these developments have created conditions that call for high levels of leadership and management at the school level. Schools are more complex and are experiencing greater change than ever before. A particular challenge is to prepare, select, place, appraise, and reward those who serve, and there is growing concern in some countries about the quality and quantity of those who seek appointment. Some governments have created or are planning new institutions to address this concern, illustrated by the establishment in England in 2000 of the National College for School Leadership. There is evidence of a loss of faith in the university as an institution to shape and inform approaches to school management, and to prepare those who will serve as leaders and managers in the future.

3. *School Management as a Field of Study*

School management has emerged relatively recently as a sub-field of study within the broader field of educational management (administration). There were few formal programs and little research on the phenomenon until the 1950s. Prior to that, approaches to the management of schools paralleled those in industry. Indeed, it was generally recognized that management in education was largely a mirror image of industry, with the creation of large centralized and bureaucratized systems of public education. The knowledge base was drawn from the industrial sector.

Departments of educational administration made their appearance in universities in the US in the mid-twentieth century and soon proliferated. They were established in the UK soon after, with many other nations following suit, drawing on the US and, to a lesser extent, the UK for their intellectual underpinnings. The first quarter-century, from the early-1950s to about the mid-1970s was characterized by efforts to build theory in educational management, modeling to a large degree the discipline approach of the behavioral sciences. In the USA, *Educational Administration Quarterly* was an attempt to provide for education what *Administrative Science Quarterly* has done for administration (management). In the UK, a seminal work edited by two pioneers in the field had the title *Educational Administration and the Social Sciences* (Baron and Taylor 1969).

Separate sub-disciplines were created in fields such as economics, finance, governance, human relations, industrial relations, law, planning, policy, and politics. Large numbers of students enrolled, and departments of educational administration became a major source of revenue for schools of education.

The theory movement in educational management was the subject of powerful attack in the mid-1970s, notably by Canadian scholar T. B. Greenfield (1975). However, it was not until the late 1980s that a systematic response was mounted by the University Council for Educational Administration in the 1987 report of the National Commission on Excellence in Educational Administration (NCEE 1987). The critique was largely directed at efforts to build a scientific theory of educational management; the separation of management from learning and teaching; and fundamental considerations of context, culture, ethics, meaning, and values. Blueprints for reform were drawn up but the field was still in ferment at the turn of the twenty-first century (see Murphy and Forsyth (1999) for a review of efforts to reform the field). Evers and Lakomski have developed a new conceptual framework ('naturalistic coherentism') in an effort to bring coherence to the fields of educational administration and management. Lakomski (2001) argues that leadership ought to be replaced by better conceptions of organizational learning.

On a more pragmatic level, the critique of educational management as a field of study lies in its failure to impact practice on a large scale, and to assist in a timely manner the implementation of reform and the resolution of concerns. A promising approach to unifying the field is to place learning at the heart of the effort. This means that policy interest in the improvement of student outcomes should be the driver of this 'quest for a center,' as US scholar Joseph Murphy (1999) has described it. Murphy classified the different stages of development in the 'profession' of educational administration, as summarized in Table 1.

Much momentum was gained in the 1990s with the establishment of the International Congress for School Effectiveness and Improvement (ICSEI), which made its mission the bringing together of policy-makers, practitioners, and researchers. The periodical *School*

Table 1
Rethinking the center for the profession of educational administration (Murphy 1999)

Time frame	Center of gravity	Foundation	Engine
1820–1900 Ideology	Philosophy	Religion	Values
1901–1945 Prescription	Management	Administrative functions	Practical knowledge
1946–1985 Behavioral science	Social sciences	Academic disciplines	Theoretical knowledge
1986– Dialectic	School improvement	Education	Applied knowledge

Effectiveness and School Improvement quickly established itself as a leading international journal.

The integration of policy, practice, and research to achieve school improvement is becoming increasingly evident, illustrated in approaches to literacy in the early elementary (primary years). The examples of Australia and the UK are interesting in this regard. Levels of literacy were considered by governments in both nations to be unacceptably low. Implementation of approaches in the Early Literacy Research Project led to dramatic improvement in the state of Victoria, Australia.

Concepts such as 'whole school design' emerged from this integration. Hill and Crévola (1999) formulated a 'general design for improving learning outcomes' that proved helpful in parts of Australia and the United States. It has nine elements: beliefs and understanding; leadership and coordination; standards and targets; monitoring and assessment; classroom teaching strategies; professional learning teams; school and classroom organization; intervention and special assistance; and home, school, and community partnerships. The implications are clear as far as educational management is concerned: there must be high levels of knowledge about learning and teaching and how to create designs and formulate strategies that bring about improvement. This contributes to a particularly demanding role where decentralization has occurred, as is increasingly the case in most nations.

4. *The Future of School Management*

The picture that emerged at the dawn of the twenty-first century is one in which school management does not stand in isolation. Leadership and management, as conceptualized by Kotter (1990) are both important, and their focus must be on learning, integrated in the notion of creating a coherent and comprehensive whole-school design. Those who serve in positions of leadership are operating in a milieu of high community expectations in which knowledge and skill are the chief determinants of a nation's success in a global economy. Rapidly increasing costs of achieving expectations mean a capacity to work in new arrangements, including innovative public–private partnerships. High expectations mean a commitment to core values such as access, equity, and choice. All of this unfolds in an environment in which the nature of schooling is undergoing fundamental change.

There are major implications for universities and other providers of programs for the preparation and professional development of school leaders. An interesting development in Victoria (Australia) and the UK is that governments are turning to the private sector to specify requirements and offer a major part of training programs. Governments in both places commissioned the Hay Group to assist in this regard. The target population in the UK is serving principals, numbering about 25,000. The program has been very highly rated by both principals and employers. The most recent contribution of the Hay Group has been the specification for Victoria of 13 competencies and capabilities for school leaders. The focus is school improvement and there are four components: (a) driving school improvement (passion for teaching and learning, taking initiative, achieving focus); (b) delivering through people (leading the school community, holding people accountable, supporting others, maximizing school capability); (c) building commitment (contextual know-how, management of self, influencing others); and (d) creating an educational vision (analytic thinking, big picture thinking, gathering information).

These developments suggest that training will be a shared responsibility in the future, with a limited role for universities unless they can form effective partnerships with the profession itself and work in association with private providers. Universities will have a crucial role to play, especially in research, but again that is likely to be in a range of strategic alliances if it is to be valued. The role of the university in philosophy and critique of school management will be more highly valued to the extent that the disjunction is minimized (between study and practice and between management and learning).

A fundamental issue that awaits resolution is the extent to which those who work in school management are not weighed down by the high expectations. It was noted at the outset that applications for appointment are declining in quality and quantity. There is even a

concern in well-funded schools in the private sector. Part of the solution is to infuse leadership and management throughout the school and generally build the capacity of all to work in the new environment. The concept of 'knowledge management' is likely to be become more important, so that accounting for the resources of a school will extend beyond financial and social capital to include intellectual capital in the form of the knowledge and skill of staff. This may go only part of the way, given that schools are in many respects still operating with a design for an earlier era. Innovation in design must extend to the creation of a work place that is engaging for all that work within it. What should be retained and what should be abandoned in current designs is the challenge for the first decade of the twenty-first century. Drucker's concept of 'organized abandonment' (Drucker 1999) is as important for schools as it is for other organizations.

See also: Educational Innovation, Management of; Management: General

Bibliography

Baron G, Taylor W (eds.) 1969 *Educational Administration and the Social Sciences*. Athlone, London

Caldwell B J, Spinks J M 1998 *Beyond the Self-managing School*. Falmer, London

Drucker P F 1999 *Management Challenges for the 21st Century*. Butterworth Heinemann, Oxford

Greenfield T B 1975 Theory about organizations: A new perspective and implications for schools. In: Hughes M G (ed.) *Administering Education: International Challenge*. Athlone, London, pp. 71–99

Hill P W, Crévola C A 1999 The role of standards in educational reform for the 21st century. In: Marsh D D (ed.) *Preparing our Schools for the 21st Century*. ASCD Yearbook 1999. Association for Supervision and Curriculum Development, Alexandria, VA

Kotter J P 1990 *A Force for Change: How Leadership Differs from Management*. Free Press, New York

Lakomski G 2001 *Management Without Leadership: A Naturalistic Approach Towards Improving Organizational Practice*. Pergamon/Elsevier, Oxford, UK

Martin M O, Kelly D L (eds.) 1996 *Third International Mathematics and Science Study. Technical Report. Vol. 1: Design and Development*. Boston College, Chestnut Hill, MA

Murphy J 1999 The quest for a center: Notes on the state of the profession of educational leadership. Invited paper at the Annual Meeting of the American Educational Research Association, Montreal, April

Murphy J, Forsyth P (eds.) 1999 *Educational Administration: A Decade of Reform*. Corwin Press, Newberry Park, CA

National Commission on Excellence in Educational Administration 1987 *Leaders for America's Schools*. University Council for Educational Administration, Tempe, AZ

B. J. Caldwell

School Outcomes: Cognitive Function, Achievements, Social Skills, and Values

Achievement in US schools is as high as it was in the mid-1970s, despite the fact that increasingly more poor and minority students are remaining in schools longer than ever. Unlike in the early 1960s, where researchers debated whether there was any impact of schooling on students' learning, it is now accepted that schools (and especially teachers) impact student achievement.

Classroom learning and school achievement broadly defined continue to grow in important ways (see, for example, Biddle et al. 1997). However, much of the recent scholarship on classroom processes has been conceptual in nature, and the associated empirical research (if any) typically involves only a few classrooms. There is considerable advocacy for certain types of classroom environments—classroom environments that are notably different from traditional instruction—but solid empirical support to show the effects of new instructional models is largely absent.

Here we discuss the need for new school-based research that addresses curriculum and instructional issues in order to advance theoretical understanding of student learning in school settings. We argue that it is no longer possible to discuss 'successful schooling' on the basis of subject-matter outcomes alone. Effective schools research and policy discussions must also include the measurement of non-subject-matter outcomes of schooling.

1. Status of Achievement in Schools

The state of achievement in US schools is widely debated and incorporates at least three major camps. The first contends that current school performance is lower than that of US youth from the mid-1970s and that of international contemporaries. Indeed, several publications supported by the federal government have corroborated the assertion that US schools are failing, including *Nation at Risk* (National Commission for Excellence in Education 1983) and *Prisoners of Time* (National Education Commission on Time and Learning 1994).

A second camp argues that students' performance today is as good as it was in the mid-1970s and that international comparisons are uninterpretable (Berliner and Biddle 1995). These and other researchers note that the stability of achievement has been maintained at a time when US schools are increasingly diverse and serving more minority students and those from poor homes than ever before. This camp also: (a) notes that reports of average school achievement are often highly misleading, because students' performance varies markedly between schools; and (b) has concerns about the performance of minority students.

Importantly, there is documentable evidence not only of the stability of student achievement over time, but also of real progress. For example, in less than a decade, the number of students, including minority students, taking advanced placement tests in high schools for college credit doubled such that one-fifth of US students in 1996 entered college with credits earned from testing (Good and Braden 2000).

A third camp agrees with the second position, but asserts that traditional achievement standards are no longer applicable to a changing society—schools must address different, higher subject-matter standards. Although we accept the premise that some curriculum changes are probably necessary to accommodate the knowledge needs of a changing society, we have two major problems with advocates for new, higher standards. First, there has been no specification of what these new skills are and, typically, if examples are even provided, the rhetoric of higher standards essentially translates into moving content normally taught later (college calculus) to appear sooner (in high school). This position is specious, because it lacks sound theoretical and empirical support. Second, extant data suggest that achievement of students in states with 'higher standards' is lower than that of students in states with purported lower standards (Camilli and Firestone 1999)!

We align our arguments with advocates of the second camp, and we note that there is extensive evidence to support this position—student achievement scores are stable in the face of growing student diversity (Berliner and Biddle 1995, Good and Braden 2000). However, despite strong empirical data to show that student performance has not declined, we believe there is much room for improvement. For example, schools in urban settings need to improve student subject-matter performance, and all schools need to further enhance student non-subject-matter achievement.

The academic accomplishments of students in today's schools are difficult to understand for various reasons. Here, we review several reasons why the achievement of US students is problematic.

2. *Which Test is Appropriate?*

For one example of the difficulty of using extant assessment instruments to compare students within the USA or across cultures, one only need note the discrepancy between the measures. For example, as Bracey and Resnick (1998) report, in 1996 the National Assessment of Educational Progress (NAEP) mathematics performance indicated that only 18 percent of US fourth graders were proficient, only 2 percent were advanced, and 32 percent were below basic. In contrast, US fourth graders performed well above the average scores of the 26 countries that participated in the international TIMSS study at the fourth-grade level. Which test is the best descriptor? Are US students above average in fourth-grade mathematics, or not? This issue of 'which test' is embedded in US testing, as well. For example, in New York state there has been a raging debate about the appropriateness of a state-mandated test (i.e., in relation to other measures, such as advanced placement tests).

3. *Problems with Curriculum, Theory, and Test Alignment*

Other factors also limit efforts to relate the implemented curriculum with student achievement. For example, in mathematics, a current 'theoretical' recommendation is to avoid instruction that is a mile wide and an inch deep (i.e., in terms of the breadth and depth of curriculum concepts covered). Those who advise teachers about this problem apparently do not worry about the opposite problem—creating a curriculum that is four inches wide and a mile deep! Aligning standardized achievement tests with curriculum implementation is vital for meaningful interpretation of student scores. Appropriately, achievement tests should assess the philosophical orientation of the curriculum. Then, other tests and research can be used to assess the utility of a curriculum for various purposes (i.e., do various curricula have different effects on long-term memory of the curriculum taught or on students' ability to transfer ideas learned in a curriculum to solve new problems?).

Finally, curriculum writers often do not understand the learning theory that undergirds particular instructional practices or goals. Some curriculum theorists do not appear to understand that behavioral models are powerful when dealing with factual and conceptual knowledge (and constructivist models are exceedingly weak in this area). Also, they often fail to understand that information-processing models (cognitive science) are more powerful than behavioral or constructivist principles at addressing wide horizontal transfer of academic concepts and authentic problem solving. In addition, in applying constructivist principles, many curriculum experts do not even recognize the differences among constructivist theories.

Despite the difficulties associated with assessment, instruments, and the curricula, and instructional alignment with achievement outcomes, there are data to suggest that teachers and schools can make a difference in student achievement.

4. *Effective Teachers*

The question that was problematic is the mid-1970s—'Do teachers make a difference in student learning?'—can now be answered with a definite '*yes*' (Weinert and Helmke 1995). There are clear data to

illustrate that some teachers notably outperform other teachers in helping similar students to learn the type of material historically measured on standardized achievement tests. It is not only possible to identify effective teachers on this useful but limited measure of student learning (e.g., see Good et al. 1983), but also research has illustrated that the practices of 'successful' teachers could be taught to other teachers.

As one instance of this research base, the Missouri Mathematics Program was successful in improving students' performance in mathematics. However, a series of research studies in this program found that teachers who were involved in the training program implemented the intended program *unevenly*. Also, not surprisingly, higher levels of program implementation were associated with higher levels of student mathematical performance. Importantly, some of the 'resistance' to implementing the Missouri Mathematics Program was due to teachers' beliefs about mathematics and how it should be taught (Good et al. 1983). Teachers' beliefs are a key, but often overlooked, point in school reform. And, as Randi and Corno (1997) found, failure to take teachers' ideas into account lowers the impact of many reform interventions.

Many aspects of teaching and learning are still problematic. Although there are numerous plausible case examples of teachers who impact students' ability to think and to apply knowledge, there is still debate about how teachers obtain these 'higher order' outcomes and whether other teachers can be educated in ways that allow them to achieve comparable effects on their students. Unfortunately, research on teachers' roles in helping students to develop thinking and problem-solving skills has in recent times been limited to case-study research. Further, research that helps teachers to expand—and assess—their capacity for impacting students' thinking abilities has been inert.

5. *Effective Schools*

There is some evidence that schools (serving similar populations of students) have more effect on student achievement (as measured by conventional standardized achievement tests) than do other schools (e.g., Good and Weinstein 1986). However, unlike the correlational and experimental research on teacher effects, the early research base describing 'effective schools' was sparse and questionable. Teddlie and Stringfield's (1993) longitudinal study in Louisiana has provided strong evidence that schools make a difference. They found that schools serving similar populations of students had different 'climates' and that these school differences were associated with differential student achievement. In general, the study confirmed many of the arguments that school effects researchers had made previously. Interestingly, in their eight-year study, about one-half of the schools retained their relative effectiveness during the time period (stability rates were similar for schools that had initially been defined as effective or ineffective). More work on this issue is warranted.

There still have not been successful and replicated studies to show that factors associated with 'effective' schools can be implemented in other schools in ways that enhance student achievement. Although there are notable examples that schools can be transformed, there is comparatively little research to see if this knowledge can be used to improve other schools.

6. *Comprehensive School Programs*

In recent years, intervention using components of previous teacher and school effects research and emerging ideas (e.g., cooperative student groups) have been combined in 'comprehensive school programs' designed to transform all aspects of schooling at the same time (including governance, structure, instruction, home schooling, communication, curriculum, and evaluation!). Although these broad interventions offer potential, it should be noted that the theoretical assumptions and congruence of these programs have not been assessed. Some programs appear to have a serious misalignment between program components. For example, as McCaslin and Good (1992) noted, some schools emphasize a behavioral approach to classroom management ('do as you are told') and a constructivist approach to the curriculum ('think and argue your conception').

Recent school intervention efforts have focused upon the impact of broad programs on student achievement. Typically, research includes no observational data, and thus there is no evidence of what parts of the program are implemented and/or whether the program parts that are presented represent a theoretically integrated program or a 'patchwork quilt' of separate and conflicting program parts. Further, the more complex and comprehensive the intervention model, the more likely teachers will alter program parts (often permanently). These 'teachable' moments go unnoticed by reformers. Perhaps unsurprisingly, there is growing consensus that such programs have not had consistent, demonstrable effects on students' achievement (Slavin 1999).

7. *Little Support for New Research*

In the past decade, policy leaders have tended to agree with advocates of camp one about the effectiveness of schools. Those in camp one (and who believe that schools are failing radically) have been able to convince policymakers to invest widely in charter and voucher plans, even in the absence of any convincing

pilot data. To these advocates, both the problem and solution are evident. Those in the second camp (stable student achievement over time) have been 'represented' by researchers attempting to implement comprehensive school reform. Unfortunately, as noted above, such research attempts fail to show the effects of program implementation. This group of researchers does not question the knowledge base sufficiently and hence fails to develop research designs that could improve its conceptions of 'good practice.'

Those who fall into the third camp (new societal requirements mandating new processes and outcomes of schooling) are prepared to reject extant theory and research on the basis that it is outdated and irrelevant. Indeed, it is not only that monolithic state-mandated standardized tests have gotten in the way of scientific research, but also that there is a growing willingness of professional groups to advocate desirable teaching processes on theoretical grounds alone. Prestigious groups, such as the National Council of Teachers of Mathematics (NCTM), have encouraged curriculum development without advocacy for research on instruction and have encouraged teachers to teach less and to use more student group work. Although it is perhaps the case that at an aggregated level, teachers as a group should teach less, the real issue is about when teachers should teach and when students should explore alone or with other students in small or large groups.

8. *Controllable and Uncontrollable Influences on Student Learning*

It has become better understood in the last 25 years of the twentieth century that schools can influence students' learning independent of home background. However, it is also better understood that the effects of schools are mediated by other factors. Bracey and Resnick (1998) argue that many factors impacting students are within the control of the school district, including: aligning textbooks with standards; aligning the broader curriculum with standards; employing high-quality teachers and providing them with appropriate professional development; and having modern textbooks and laboratory equipment. Although these factors are controllable in theory, it should be noted that some school districts have considerably fewer resources to work with than do other districts. It seems unconscionable that all schools do not receive these controllable resources, which are known to impact achievement.

Bracey and Resnick (1998) have argued that there are various out-of-school factors that can influence student achievement, including: teenage pregnancy rates; percentage of female-headed households; poverty rate; parental and maternal educational levels; the percentage of students for whom English is not their native language; number of violent incidents per year; and number of annual policy visits/disciplinary actions that occur at particular schools. These opportunity-to-learn variables are seen as conditions that impact student achievement but are largely beyond the direct influence of the school system. Accordingly, various groups have noted that achievement problems of students from low-income families need to be addressed by other social agencies. Clearly, poor achievement in schools is often outside the control of individual schools, but not outside the influence of the broader society. If a society chooses to hold individual schools 'accountable' for achievement, it is incumbent on that society to spread its wealth in ways that address factors that individual schools cannot address.

9. *Improving Performance: Aligning Theory and Research*

Massive investments of public funds in student testing have occurred over the last two decades of the twentieth century and, as noted, some policymakers have called for even more testing and for raising extant standards. The research for higher standards has led to the development of tests in some states that are designed, absurdly, so that most students fail them. Oddly, despite this massive investment in testing, no information has been obtained to describe adequately what students learn in schools and how conditions can be changed to further enhance student learning. We believe that school reform efforts need to bring theory and empirical research together. It is time to stop advocacy from prestigious groups that do not support their reform calls in the form of theory and data.

There is a growing sense of the importance of instructional balance in order to scaffold the learning needs of students in complex classroom settings (Good and Brophy 2000, Grouws and Cebulla 2000). However, these assertions demand more research examination in various contexts. In contrast, this advice is currently not echoed by professional groups. These groups continue to overemphasize (in our opinion) problem solving and application without concomitant attention to the development of conceptual knowledge. The ability to solve problems also requires the ability to recognize and find problems as well as the computational and conceptual skills necessary for addressing problems. Simply put, there can be too little or too much attention to concepts, facts, or problem solving.

10. *Non-subject-matter Outcomes of Schooling*

Americans have always argued that their students should not simply be 'academic nerds,' but that students should also be participants in broader society (have jobs, participate in drama or sports, etc.).

Increasingly, parents and citizens have expressed interest in schools being more responsive to the non-subject-matter needs of youth. Often, this argument is expressed only in terms of improving achievement—if children feel secure, they will achieve better. But increasingly, the assertion is made that some outcomes of schooling are important whether or not they impact subject-matter learning.

Despite the considerable advocacy of policy leaders for higher standards, many citizens are pushing hard for students to learn responsibility. For example, when asked what first comes to their minds when Americans think about today's teenagers, 67 percent described them as 'rude, irresponsible, and wild,' and 41 percent reported that they did not think teens had enough to do and that they were not being taught the value of hard work (schools are not doing their 'job') (Farkas et al. 1997). This idea is echoed in the 1999 Phi Delta Kappan/Gallup poll, where 43 percent of parents reported that the main emphasis of schools should be on students' academic skills, and 47 percent of parents said they believe that the main emphasis of schools should be helping students to learn to take responsibility. And earlier in a 1996 Phi Delta Kappan/Gallup poll, parents indicated they would prefer their children to receive bad grades and be active in school activities rather than making an excellent grade and not participating in extracurricular activities (60 percent to 20 percent). It is clear that parents and citizens, reacting to a perception of adolescents' moral decline, believe that schools should play a greater role in teaching non-subject-matter content, such as responsibility and civic behavior.

Although citizens want schools to provide students with more non-subject-matter knowledge, there are problems with the systematic addition of this 'content' to the curriculum. One problematic issue is the difficulty in finding consensus on what a moral curriculum would look like. Subsequently, what assessment criteria should be used to measure the curriculum's effectiveness? Is the curriculum successful if more students vote? Or is it effective if students are able to negotiate conflicts calmly? As we have argued, achievement tests do not always represent adequately the curricula they are supposed to measure. Therefore, it is reasonable to think that appropriate accountability and assessment tools would be difficult to implement for non-subject-matter content—at least without adequate empirical research.

Policymakers, however, are not only supporting past emphases of the impact of schools on achievement, but are pressuring for higher and higher standards. In some instances, the standards have been pushed so high (and artificially high) that in Virginia and in Arizona over 90 percent of students failed exams that presumably represented efforts by state government to improve standards. At some level, increasing standards may be a laudable goal; however, developing standards that only 10 percent of our students pass seems to be a ridiculous and self-defeating effort. Given the exaggerated, negative views of youth, and testing policies designed to guarantee youth's failure, it seems important to explore motivation. Are such practices designed to suggest that youth, especially minority youth, do not deserve resources in their schools?

11. *Educators on Noncognitive Factors*

Rothstein (2000) has noted that, in part, achievement measures are reified in making decisions about the effectiveness of schooling, because there are many of these measures, but few ways to measure noncognitive outcomes. He noted that there are 10 core outcomes that motivate citizens to invest in schools so that students can contribute to society and lead productive adult lives. The dimensions he identifies are literacy, competencies in mathematics, science, and technology, problem solving, foreign languages, history knowledge, responsible democratic citizenship, interest in creative arts, sound wellness practices, teamwork and social ethics, and equity (narrowing the gap between minority and white students in the other nine areas).

Wanlass (2000) has drawn attention to the numerous problems in contemporary society, including the effects of a global economy, increased ethnic diversity, advanced technology, and increasing risk for various problems such as intolerance (hate crimes), overpopulation, and a dramatic depletion of natural resources. In this context, she argued a need to help youth develop the affective dispositions required for the modern world (leadership, tolerance, philanthropic concern). She suggested several strategies, capitalizing on the unique strengths and talents of more students and the need to better develop the unique abilities of nonacademic areas as well as academic concerns (performing arts, social interaction, leadership, civic-mindedness, social advocacy, etc.).

Goodenow (1992) has argued that students who feel more connected to their schools are more likely to be academically and socially successful. Students who felt more personally valued and invested in were more likely to place higher value and have higher expectations for classroom success than were students who did not feel valued. This finding was replicated in a large-scale study (Blum and Rinehart 1996), which surveyed more than 90,000 students nationwide and found that those who felt more connected to their schools were less likely to engage in risky behavior, including: smoking, doing drugs, engaging in sexual activity, and repeated truancy, and they were less likely to drop out of school. In this study, school connectedness was defined as students who: (a) felt they were treated fairly, (b) reported they got along with other teachers and students, and (c) felt close to people at school. Connectedness was also explored in

terms of how schools were perceived by their teachers and administrators. A school that encouraged more connectedness exhibited less student prejudice and had higher daily attendance rates, lower drop out rates, and more teachers employed with masters degrees. Although these data are correlational in nature, it suggests that some factors thought to be uncontrollable, may in fact be controllable.

Researchers have also investigated the role of non-subject-matter variables as they influence students' motivation to achieve. For example, Nichols (1997) examined the nature of the relationship between students' perceptions of fairness and their reported levels of motivation to achieve. Of the four dimensions of fairness identified through factor analytic techniques (personal, rules, punishment, teacher), a *personal* sense of fairness more strongly mediated motivational variables than did other dimensions. Similar to Goodenow (1992), this data set indicated that students who felt more personally valued were more likely to be motivated to achieve in school settings. However, even if students' personal sense of fairness was not linked directly to an achievement motivation, it would seem that this is a desirable outcome of schooling. Students should feel respected and safe, whether or not these variables influence achievement.

12. Conclusion

Clearly, more public debate is called for, if more focused judgments are to be reached concerning those aspects of schools that are defined as most critical. However, given the growing evidence that citizens are concerned about non-subject-matter outcomes of schooling, it seems incumbent on policymakers to 'proxy' these concerns by collecting data on some of the many outcomes that could be measured. There is a need for research and arguments focused on non-subject-matter variables, as well. For example, it seems that high subject-matter specialization and excessively high levels of student obedience (as opposed to student initiatives) are not a prescription for forging a professional community well prepared for engaging in creative work and problem solving.

A recent poll of teenagers found that the three greatest pressures they experience are to get good grades (44 percent), to get into college (32 percent), and to fit in socially. In contrast, fewer teens report feeling pressure to use drugs or alcohol (19 percent) or to be sexually active (13 percent). What modern youth are concerned about varies markedly from the view presented by the media. Unfortunately, policymakers at present seem more intent on blaming youth and holding them accountable than on understanding them. The collection of reliable data on various non-subject-matter outcomes of schooling might help to assure that youth are developing the proactive skills necessary for life in a democracy (e.g., civility, a willingness to examine ideas critically, and ambition).

The historical concern for evidence about subject-matter growth should continue—particularly if these measures are collected in ways that provide a basis for curriculum improvement. However, it is important to recognize that US public schools are about much more than subject-matter acquisition. More research on students and how they mediate social, emotional, and academic contexts might enable educators to design better school programs that recognize students both as academic learners and as social beings (McCaslin and Good 1992).

See also: Curriculum as a Field of Educational Study; Educational Assessment: Major Developments; Educational Evaluation: Overview; Educational Research for Educational Practice; School Effectiveness Research; Teacher Behavior and Student Outcomes

Bibliography

Berliner D, Biddle B 1995 *The Manufactured Crisis: Myth, Fraud and the Attack on America's Public Schools*. Addison-Wesley, New York

Biddle B, Good T, Goodson I 1997 *Teachers: The International Handbook of Teachers and Teaching*. Kluwer, Dordrecht, The Netherlands, Vols. 1 and 2

Blum R W, Rinehart P M 1996 *Reducing the Risk: Connections That Make a Difference in the Lives of Youth*. University of Minnesota, Division of General Pediatrics and Adolescent Health, Minneapolis, MN

Bracey G W, Resnick M A 1998 *Raising the Bar: A School Board Primer on Student Achievement*. National School Boards Association, Alexandria, VA

Camilli G, Firestone W 1999 Values and state ratings: An examination of the state-by-state indicators in quality counts. *Educational Measurement: Issues and Practice* **35**: 17–25

Farkas S, Johnson J, Duffett A, Bers A 1997 *Kids These Days: What Americans Really Think About The Next Generation*. Public Agenda, New York

Good T L, Braden J 2000 *The Great School Debate: Choice, Vouchers, and Charters*. Erlbaum, Mahwah, NJ

Good T L, Brophy J 2000 *Looking in Classrooms*, 8th edn. Longman, New York

Good T L, Weinstein R 1986 Schools make a difference: Evidence, criticisms, and new directions. *American Psychologist* **41**: 1090–7

Good T L, Grouws D A, Ebmeier H 1983 *Active Mathematics Teaching*. Longman, New York

Goodenow C 1992 School motivation, engagement, and sense of belonging among urban adolescent students. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA

Grouws D, Cebulla K 2000 Elementary and middle school mathematics at the crossroads. In: Good T (ed.) *American Education: Yesterday, Today, and Tomorrow*. Ninety-Ninth Yearbook of the National Society for the Study of Education. University of Chicago Press, Chicago, Chapter 5, pp. 209–55

McCaslin M, Good T 1992 Compliant cognition: The misalliance of management and instructional goals and current school reform. *Educational Researcher* **21**: 4–17
National Commission for Excellence in Education 1983 *A Nation At Risk: The Imperatives for Education Reform*. US Department of Education, National Commission for Excellence in Education, Washington, DC
National Education Commission on Time and Learning 1994 *Prisoners of Time*. US Government Printing Office, Washington, DC
Nichols S 1997 *Students in the Classroom: Engagement and Perceptions of Fairness*. Master's thesis, University of Arizona
Randi J, Corno L 1997 Teachers as innovators. In: Biddle B, Good T, Goodson I (eds.) *International Handbook of Teachers and Teaching*. Kluwer, Dordrecht, The Netherlands, Vol. 2, Chap. 12, pp. 1163–221
Rothstein R 2000 Toward a composite index of school performance. *Elementary School Journal* **100**(5): 409–42
Slavin R 1999 How Title I can become the engine of reform in America's schools. In: Orfield G, Debray E (eds.) *Hard Work for Good Schools: Facts Not Fads In Title I Reform*. The Civil Rights Project, Harvard University, Cambridge, MA, Chap. 7, pp. 86–101
Teddlie C, Stringfield S 1993 *Schools Make a Difference: Lessons Learned From a 10-Year Study of School Effects*. Teachers College Press, New York
Wanlass Y 2000 Broadening the concept of learning and school competence. *Elementary School Journal* **100**(5): 513–28
Weinert L, Helmke A 1995 Interclassroom differences in instructional quality and interindividual differences in cognitive development. *Educational Psychologist* **30**: 15–20

T. L. Good and S. L. Nichols

Schooling: Impact on Cognitive and Motivational Development

Up until the mid-1970s, neither the general public nor the social sciences could agree on whether schooling has a significant impact on cognitive development, how large this impact is or may be, and which variables are responsible for any potential effects it may have. For example, even in 1975, Good et al. asked skeptically, 'Do schools or teachers make a difference? No definite answer exists because little research has been directed on the question in a comprehensive way' (Good et al. 1975, p. 3). Things have changed since then, because the findings from numerous empirical studies have banished any serious doubts concerning whether schooling impacts on both the cognitive and motivational development of students.

This article will report the arguments questioning the importance of school for the mental development of children and adolescents and assess their scientific validity; present the empirical evidence on the strong impact of school on cognitive development, while simultaneously showing the limits to its generalizability; and sketch the findings on the role of school in motivational development. The final section presents some conclusions on the relations between cognitive and motivational development under the influence of school.

1. *Scientific Doubts Regarding the Impact of Schooling on Cognitive Development*

Compared with the impact of social origins and/or stable individual differences in intellectual abilities on cognitive development, that of schooling was long considered marginal. Nonetheless, the Coleman Report (Coleman et al. 1966) was still a great shock to many politicians, educators, and teachers when it concluded, 'that school brings little influence to bear on a child's achievement that is independent of his background and general social context' (p. 325). Just as radically, Jencks et al. (1972) confirmed 'that the character of schools' output depends largely on a single input, namely, the characteristics of the entering children. Everything else—the schools' 'budget, its policies, the characteristics of the teachers—is either secondary or completely irrelevant' (p. 256). According to their findings, elementary school contributed 3 percent or less to the variance in cognitive development, and secondary school 1 percent or less.

Even though both Coleman et al.'s (1966) and Jencks et al.'s (1972) general statements were modified and differentiated slightly in comparisons between private and public schools, between middle-class and lower-class children, and between the effects of different school variables, the final conclusion remains unchanged: 'Nevertheless, the overall effect of elementary school quality on test scores appears rather modest' (Jencks et al. 1972, p. 91). Viewed from the perspective of the beginning of the twenty-first century, such statements and conclusions are, without exception, underestimations of the impact of schooling on cognitive development in childhood and adolescence.

There are various reasons for these underestimations. First, two aspects of cognitive development that should be kept strictly separate are frequently confounded: the growth and acquisition of competencies, knowledge, and skills on the one side compared with the change in interindividual differences in cognitive abilities on the other. We now know that schools are necessary and very influential institutions for the acquisition of that declarative and procedural knowledge that cannot be learned in the child's environment outside school, but that the quantity and quality of schooling has only a limited power to modify individual differences in abilities.

Second, Coleman et al. (1966), Jencks et al. (1972), and many other studies, compared very similar schools in the USA or other industrialized countries. Naturally, such schools reveal more commonalities than

differences due to the similarities in teacher training, curricular goals, educational traditions, the budget of the educational institutions, the state-regulated quantity of teaching, and standardized achievement tests. The methodological problems in such research did not become evident until the publication of findings from cross-cultural studies. These showed that when schools in the Third World—or even achievement measured in ethnic groups with no or very low school education—are taken into account alongside schools in industrialized countries, massive effects of the quantity and quality of instruction can be ascertained (Anderson et al. 1977).

The third reason for underestimating the impact of schools was that scientific interest in educational sociology focused on the role of social class and the family in mental development, whereas developmental psychology was dominated by Jean Piaget's constructivist model. Whereas theories of cognitive development deal with general mechanisms of thought, action, and experience, models of school learning in the strict sense are concerned with the acquisition of specific skills and small bits of knowledge. Piaget and his co-workers emphasized 'that this form of learning is subordinate to the laws of development and development does not consist in the successive accumulation of bits of learning since development follows structuration laws that are both logical and biological' (Inhelder and Sinclair 1969, p. 21).

2. *Empirical Confirmation of the Impact of Schooling on Cognitive Development*

In many social scientists, the skeptical attitude toward the role of schools in recent decades has been overcome to a large extent (although not completely) by the results of numerous empirical studies. This applies not only to the acquisition of specific cognitive competencies but also to the promotion of intelligence.

After reviewing the relevant literature, Rutter (1983, p. 20) came to the conclusion that 'the crucial component of effective teaching includes a clear focus on academic goals, an appropriate degree of structure, an emphasis on active instruction, a task-focused approach, and high achievement expectations.' Although recent years have seen increasingly strong criticism of Rutter's description of the successful classroom, it is broadly confirmed empirically that active teachers, direct instruction, and effective classroom management on the one side accompanied by active, constructive, and purposeful learners on the other side are necessary conditions for the effective acquisition of competencies, knowledge, and skills. Naturally, a major role can also be assigned to intrinsically motivated, self-organized, and cooperative learning (Weinert and Helmke 1995).

Nonetheless, the impact of schools on cognitive development ranges far beyond teaching declarative and procedural knowledge and also includes the promotion of intellectual abilities. For example, Ceci (1991) summarized his state-of-the-art review as follows:

> Of course, schooling is not the complete story in the formation and maintenance of IQ-scores and IQ-related cognitive processes … Children differ in IQ and cognitive processes prior to entering school, and within a given classroom there are sizeable individual differences despite an equivalence of schooling. … Thus, the conclusion seems fairly clear: Even though many factors are responsible for individual and group differences in the intellectual development of children, schooling emerges as an extremely important source of variance, notwithstanding historical and contemporary claims for the contrary. (p. 719)

Looking at the available data from industrialized countries, the Third World, and, above all, cross-cultural studies on the impact of schooling on cognitive development in general and the effectiveness of specific features of schools in particular, substantial effects can be ascertained on intellectual abilities; on metacognitive strategies; on the acquisition of verbal, mathematical, scientific, logical, and technological competencies; and on various forms of domain-specific and cross-domain knowledge. These findings provide unequivocal support for Geary's (1995) hypothesis that schools are necessary cultural conditions for the acquisition of those abilities and skills that cannot be learned in the child's immediate environment: this means almost all the cognitive competencies necessary for a good life and a successful career in our modern scientifically and technologically shaped world.

Up to which level cognitive development is enhanced and which competencies are acquired depends strongly on the quality and quantity of schooling. This is confirmed by findings from a number of large-scale international studies that not only compared school achievement but also assessed important features and variables in the individual national school systems. A current example is the Third International Mathematics and Science Study (TIMSS) carried out by the International Association for the Evaluation of Educational Achievement (IEA; see Baumert et al. 1997).

In association with numerous national and regional studies on educational productivity, school effectiveness, or school improvement, it has been possible to identify a variety of combinations and configurations of variables in the school system, the individual school, the classroom, the teacher, instruction, and the school context that impact on various dimensions and aspects of the cognitive development of students. At the same time, these findings have also led to scientific suggestions for improving school systems.

Alongside mean effect sizes, the long-term effects of exceptional schools and teachers are of particular interest. For example, Pederson et al. (1978) reported

the strong long-term effects of one excellent teacher in a single-case study. Compared with the students of other teachers, the children she had taught in their first two years at school were significantly more successful in their later academic careers as well as in their later jobs. More detailed analyses of this effect revealed that this teacher attained above-average effects on school performance in the first two school grades as well as on her students' attitudes toward work and their initiative, but no great increase in intellectual abilities. Even more important than the direct teacher effects were the indirect effects of a good, successful, and positively experienced start to school on all the children's further cognitive and motivational development.

Although it has to be said that this is only one single-case study, its findings could be replicated as a trend, though not so emphatically, in several representative longitudinal studies (Weinert and Helmke 1997). However, great methodological care is necessary when gathering, analyzing, and interpreting longitudinal data. Only multilevel analyses (that view teacher or instruction variables as being moderated by classroom contexts, and classroom effects as being influenced by features of the school, the school system, and the sociocultural environment) permit valid estimations of the effects of certain clusters of variables on students' cognitive development and growth in knowledge.

Such studies have shown that individual variables make only a very limited contribution to explaining the variance in student outcomes. There are many reasons for this: one of the most important is the complex 'multideterminedness' of students' academic achievement. Haertel et al. (1983) characterize the ensuing theoretical and methodological problems as follows:

> … classroom learning is a multiplicative diminishing-returns function of four essential factors—student ability and motivation, and quality and quantity of instruction … Each of these essential factors appear to be necessary but insufficient by itself to classroom learning; that is, all four of these factors appear to be required at least at minimum levels for the classroom learning to take place. It also appears that the essential factors may substitute, compensate, or trade-off for one another in diminishing rates of return. (p. 75)

3. *The Limited Impact of School on the Modification of Individual Differences in Cognitive Abilities and Competencies*

Of the four essential conditions for classroom learning identified by Haertel et al. (1983), only two (the quantity and quality of instruction) are under the control of the school and the teacher. The other two (ability and motivation) are features of students. Already when entering school, children exhibit large interindividual differences in both their cognitive competencies and their motivational tendencies.

Through genetic factors and through covariations between genotype and environmental influences, interindividual differences in intellectual abilities already become moderately stable early in childhood. Round about the seventh year of life, astonishingly high stability coefficients of 0.5 to 0.7 are already found, despite the relatively low reliability of many tests at this age level. Assuming that more intelligent students learn, on average, more quickly, more easily, and better than less intelligent students under the same conditions of instruction, it can be anticipated that interindividual differences in mental abilities and cognitive development will not diminish under the influence of school but become increasingly more stable. This is also supported by the results of many empirical studies: For example, during the course of elementary schooling, stability coefficients in verbal intelligence tests increase to about 0.8, nonverbal intelligence to 0.7, mathematical competencies to 0.7, reading and writing to 0.7, and the development of positive attitudes toward learning, self-confidence, and test anxiety to 0.6 in each case (Weinert and Helmke 1997). This recognizable trend in elementary school continues in secondary school, with stabilities in IQ scores already attaining values between 0.8 and 0.9 from the age of 12 years onward, thus permitting almost perfect long-term prediction.

These and other empirical findings can be used to draw some theoretical conclusions on classroom learning that may be modified, but not falsified, by variations in school contexts:

(a) Cognitive abilities and competencies develop dramatically in childhood and adolescence—not least under the influence of the school. At the same time, interindividual differences already stabilize at a relatively early age and then remain more or less invariant.

(b) This trend applies not only for intellectual abilities but also for the acquisition of cognitive competencies and domain-specific knowledge. This is particularly apparent when the knowledge acquired is cognitively demanding. In this case, it is necessary to assume a relationship between the level of intelligence and the quality of the knowledge ('intelligent knowledge').

(c) Naturally, the stabilities in individual differences are in no way perfect, so that changes can be observed in students' relative positions in the ranking of abilities and academic achievements. However, very strong changes are often an exception due to idiosyncrasies rather than the rule.

(d) All attempts to level out these individual differences in cognitive abilities and competencies that stabilize during the course of schooling and to make achievements and achievement dispositions equal for all students have generally been unsuccessful. The same applies to the concept of mastery learning when

practiced over a longer period of time and on demanding tasks.

(e) Of course, very significant changes in individual differences result when students are taught for different lengths of time and/or with different achievement aspirations in various subjects or the same subject. Whereas, for example, some students broaden their basic knowledge of the world only slightly through the physics they learn at school, others acquire a high degree of expertise in the subject. Despite the doubtful validity of the claim put forward in the novice–expert research paradigm that intellectual abilities and talents are irrelevant for the acquisition of expertise, the extent of deliberate practice seems to be decisive for the achievement of excellence.

(f) Regardless of the level of individual abilities and regardless of the stability of interindividual differences in ability, it is still necessary for all students to acquire all their cognitive competencies through active learning. Many studies have shown that the quantity and quality of instruction are useful and, in part, even necessary for this.

4. *The Impact of Schooling on Motivational Development*

The available statistical meta-analyses confirm the belief shared by most laypersons that student motivation is an essential personal factor for successful learning at school. However, motivational tendencies and preferences are not just conditions of learning but are also influenced by learning and teaching in the classroom. Their systematic promotion is simultaneously an important goal of school education.

To analyze motivational conditions and consequences of school learning, it is necessary to distinguish strictly between dispositional motives and current motivation. Whereas motives (e.g., achievement motive, affiliation motive, attribution style, test anxiety, self-confidence) are understood as personal traits that are relatively stable and change only slowly, motivation concerns current approach or avoidance tendencies that not only depend on dispositional motives but are also influenced strongly by situational conditions (stimuli, incentives, rewards, threats, etc.). Naturally, the long-term modification of motivational development through the school predominantly involves the students' dispositional motives, attitudes, and value orientations.

School enrollment between the fifth and seventh year of life seems to be a sensitive period of development for this. The cognitive competencies acquired during preschool years enable first graders to compare themselves with others, to make causal interpretations over the differences and changes in achievement they register, and to form anticipations regarding future achievements. At the same time, the school class offers a variety of opportunities for comparisons on the levels of behavior, achievement, evaluation, and ability. Teachers play an important role in this.

Almost all preschool-age children are optimists: They look forward to attending school, most anticipate that they will do well, they have only a low fear of failure, and they believe that they will be able to cope with all the demands of the school. This more-than- optimistic attitude changes during the first two years at school. In general, children become realists, although most of them continue to hold positive expectations regarding their own success. This basic attitude stabilizes during the further course of schooling—naturally, with large interindividual differences due, in part, to personal academic successes and failures, in part to a stable anticipation bias (Weinert and Helmke 1997).

This general developmental trend is modified more or less strongly by specific features of the school, the classroom, the teacher, and the methods of instruction (for overview see Pintrich and Schunk 1996). A few examples of this will be given below. These come either from studies in which the motivational effects of teaching and learning were observed in the classroom, or from studies assessing the effects of planned interventions.

Personal causation (the feeling of being an origin and not a pawn) is an important motivational condition for successful classroom learning. If students are to experience personal causation, they must have opportunities to (a) set themselves demanding but realistic goals, (b) recognize their own strengths and weaknesses, (c) be self-confident about the efficacy of their own actions, (d) evaluate whether they have attained the goals they have set themselves, and (e) assume responsibility not only for their own behavior but also, for example, that of their classroom peers.

Psychological interventions designed to enhance the 'origin atmosphere' in the classroom and increase the 'origin experiences' of the students have led not only to the intended motivational changes but also to broad improvements in average academic achievement. Similarly positive outcomes are reported from studies designed to improve the goal orientation of learning, to strengthen achievement motivation, and to modify the individual patterns of attributions used in intuitive explanations of success and failure. For example, efforts have been made in both classrooms and individual students to modify the typical attribution style for learned helplessness (attributing failure to stable deficits in ability and success to luck or low task difficulty) so that given abilities and varying effort as a function of task difficulty become the dominant personal attribution patterns.

Positive modifications to student motivational development depend decisively on the teacher's attitudes, instructional strategies, and feedback behavior. As a result, many intervention programs for enhancing motivation focus on both students and teachers

simultaneously. For example, a large study of 'self-efficacious schools' is being carried out at the present time in Germany. It is testing the assumption that enhanced self-efficacy beliefs may affect not only teacher behavior (by reducing burnout syndromes and enhancing a proactive attitude) but also student learning (by increasing motivation and achievement).

For several decades, theories of learning motivation have been dominated by the assumption that basic needs are satisfied through achievement, social support, rewards, self-evaluation, and so on. Frequently the motivational mechanisms were those underlying expectancy × value models, cost-benefit calculations, or instrumental behavior models. In comparison, intrinsic motives for learning have played a much smaller role. Intrinsic motivation addresses needs and goals that are satisfied by the process of learning itself and by the possibility of experiencing and enjoying one's own expression of competence and increasing knowledge in a preferred content domain. In this theoretical framework, working and learning are not a means to an end, but are more or less an end in themselves; learning activities and learning goals are not clearly distinguished (e.g., exploratory drive, experience of flow, effectancy motivation, etc.). Underlying these concepts is the assumption '… that intrinsic motivation is based in the innate, organismic needs for competence and self-determination. It energizes a high variety of behaviors and psychological processes for which the primary rewards are the experience of effectance and autonomy' (Deci and Ryan 1985, p. 32).

The reason for a careful consideration of intrinsic motivation is that in educational philosophy intrinsic motivated learning and behavior are regarded as the most important goals of schooling. The worry is that strengthening extrinsic motivation by making external incentives and rewards available is likely to be detrimental to achieving this goal. The ideas underlying such fears come from the results of some studies investigating the 'overjustification effect,' which show that the use of rewards can undermine intrinsic motivation. However, the results of several recent studies indicate that this fear is unjustified. Furthermore, the relation between dispositional motives, actual motivation, and learning activities is not simple and linear but is influenced strongly by interconnections with cognitive processes, metacognitive strategies, and volitional skills.

5. *Conclusions*

Schools are cultural and educational institutions for promoting cognitive development and imparting those competencies that the child cannot acquire within the immediate environment. Throughout the world, schools fulfil this evolutionary, psychological, and cultural function more or less well. The quantity and quality of instruction are a major source of the variance in the genesis of interindividual differences in cognitive competencies. Moreover, the development and promotion of competent, self-regulated learning is an equally important goal of schools in modern societies. The development of intrinsic and extrinsic motives becomes particularly important here in association with metacognitive and volitional competencies. This is why recent decades have seen not only studies examining which relations can be found between the classroom atmosphere, the teacher's behavior, and the subjective experiences of the students on the one side and their motivational and cognitive development on the other side, but also numerous programs designed to promote motivation systematically that have been tried out with, in part, notable success.

The relation between cognitive and motivational development and their interaction are of particular interest in both theory and practice. This is just as true for the influence of academic success on the development of a positive self-concept as for the importance of the self-concept for individual progress in learning (Weinert and Helmke 1997).

See also: School Effectiveness Research; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values

Bibliography

Anderson R C, Spiro R J, Montague W E (eds.) 1977 *Schooling and the Acquisition of Knowledge*. Erlbaum, Hillsdale, NJ

Baumert J, Lehmann R et al. 1997 *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: Deskriptive Befunde*. Leske and Budrich, Opladen, Germany

Ceci S J 1991 How much does schooling influence general intelligence and its cognitive components? A reassessment of evidence. *Developmental Psychology* **27**: 703–22

Coleman J S, Campbell E R, Hobson C J, Mc Partland J, Mood A, Weingold F D, York R L 1966 *Equality of Educational Opportunity*. Government Printing Office, Washington, DC

Deci E, Ryan R M 1985 *Intrinsic Motivation and Self-determination in Human Behavior*. Plenum Press, New York

Geary D C 1995 Reflections on evolution and culture in children's cognition. *American Psychologist* **50**: 24–36

Good T L, Biddle B J, Brophy J E 1975 *Teachers Make a Difference*. Holt, Rinehart, & Winston, New York

Haertel G D, Walberg H J, Weinstein T 1983 Psychological models of educational performance. A theoretical synthesis of constructs. *Review of Educational Research* **53**: 75–91

Inhelder B, Sinclair H 1969 Learning cognitive structures. In: Mussen P H, Langer J, Covington M (eds.) *Trends and Issues in Developmental Psychology*. Holt, Rinehart, & Winston, New York, pp. 2–21

Jencks C, Smith M, Acland H, Bane M J, Cohen D, Gintis H, Heyns B, Michelson S 1972 *Inequality*. Basic Books, New York

Pederson E, Faucher T A, Eaton W W 1978 A new perspective on the effects of first grade teachers on children's subsequent adult status. *Harvard Educational Review* **48**: 1–31

Pintrich P R, Schunk D 1996 *Motivation in Education: Theory, Research, and Applications*. Prentice Hall, Englewood Cliffs, NJ
Rutter M 1983 School effects on pupil progress: Research findings and policy implications. *Child Development* **54**: 1–29
Weinert F E, Helmke A 1995 Learning from wise mother nature or big brother instructor: The wrong choice as seen from an educational perspective. *Educational Psychologist* **30**: 135–42
Weinert F E, Helmke A (eds.) 1997 *Entwicklung im Grundschulalter*. Beltz, Weinheim, Germany

F. E. Weinert

Schools, Micropolitics of

'Micropolitics' is a specific perspective in organization theory. It focuses on 'those activities taken within organizations to acquire, develop, and use power and other resources to obtain one's preferred outcomes in a situation in which there is uncertainty or dissent' (Pfeffer 1981, p. 7) because it considers them as most important for the constitution and the workings of organizations.

This article will explore the uses and usefulness of this concept in educational research. It starts by outlining the micropolitical view on educational organizations. Then, it gives some examples of representative micropolitical research in order to illustrate typical topics and research strategies. Finally, it discusses criticisms of micropolitics and indicates areas for development.

1. *The Concept of Micropolitics*

1.1 *Main Elements*

Traditional organizational theories of different origins seem to converge in describing organizations as goal oriented and rationally planned, characterized by stable objective structures, a high degree of integration, and a common purpose of its members. Conflicts between the members of an organization are considered costly and irrational 'pathologies' to be eradicated as soon as possible.

However, conflicts are endemic in organizational life (Ball 1987). Different stakeholders pursue their own interests which are not necessarily identical with the formulated goals. Coalitions are formed, meetings are boycotted, and formal structures are ignored. As such phenomena frequently occur in organizations, it is not helpful to exclude them from organizational theorizing: models which can cope with this 'dark side of organizational life' (Hoyle 1982, p. 87) are needed.

The micropolitical perspective was originally developed in the realm of profit organizations (e.g., Bacharach and Lawler 1980, Pfeffer 1981). In a 1982 seminal paper, Eric Hoyle (1982) put forward arguments for exploring the concept also for education. In 1987 Stephen Ball (1987) published the book *The Micro-Politics of the School* in which he used empirical material to push the development of the concept. In its wake the term 'micropolitics' appeared more frequently in various research papers (see Blase 1991 for a representative collection). What are the main elements of a micropolitical view of schools as presented in this early research?

(a) The micropolitical perspective is based on a specific view of organizations. They are seen as characterized by diverse goals, by interaction and relationships rather than structures, by diffuse borders and unclear areas of influence rather than clear-cut conditions of superordination and delegation, by continuous, unsystematic, reactive change rather than by longer phases of stable performance and limited projects of development. The main focus of micropolitical approaches is not the organization as it aspires to be in mission statements or organizational charts, but the organization-in-action and, in particular, the 'space between structures' (Hoyle 1982, p. 88) which produces enough ambiguity to allow political activities to flourish.

(b) A micropolitical approach is based on a specific image of actors. The members of organizations are seen as pursuing their own interests in their daily work. The actors may do this as individuals, they may coalesce in loosely associated interest sets, or use subdivisions of the organization (e.g., departments, professionals versus administration) as power bases. In order to protect or enhance their organizational room for maneuver, they aim to retain or obtain control of resources, such as the following (see Kelchtermans and Vandenberghe 1996, p. 7, Ball 1987, p. 16):

(i) material resources, such as time, funds, teaching materials, infrastructure, time tabling, etc.

(ii) organizational resources such as procedures, roles, and positions which enable and legitimate specific actions and decisions. They define 'prohibited territories' and relationships of subordination or autonomy. Thereby, they have their bearing on the actors' chances to obtain other resources (e.g., through career, participation, etc.).

(iii) normative or ideological resources, such as values, philosophical commitments, and educational preferences. A particularly important ideological resource is the 'definition of the organization' which indicates what is legitimate and important in the organizational arena.

(iv) informational resources: the members of an organization are interested in obtaining organizationally relevant information and to have their expertise, knowledge, and experience acknowledged.

(v) social resources: affiliation to, support from, and reputation with influential groups within and outside

the school are assets which may be mobilized for one's organizational standing.

(vi) personal resources: also personal characteristics—such as being respected as a person, being grounded in an unchallenged identity as a teacher, etc.—are resources in organizational interactions.

(c) A micropolitical perspective pays special attention to interaction processes in organizations. These are interpreted as strategic and conflictual struggle about the shape of the organization which, in consequence, defines the members' room for maneuver. This has the following implications:

(i) Organizational interaction is power-impregnated. It draws on the resources mentioned above for influencing the workings of an organization. However, these resources have to be mobilized in interaction in order to become 'power.'

(ii) Power is in need of relationship. In order to satisfy his or her interests an actor needs some interaction from other actors. Power builds on dependency, which makes the relationship reciprocal, but, as a rule, asymmetric.

(iii) Power is employed from both ends of the organizational hierarchy. There are advantages of position, however, 'micropolitical skills may be well distributed throughout the organization and position is not always an accurate indicator of influence' (Ball 1994, p. 3824).

(iv) Actors use a multitude of strategies and tactics to interactively mobilize their resources and to prevent others from doing so, such as, for example, setting the agenda of meetings, boycotting and escalating, avoiding visibility, confronting openly or complying, carefully orchestrating scenes behind closed doors or in public arenas, etc.

(v) Organizational interaction is seen as conflictual and competitive. 'Confronted by competition for scarce resources and with ideologies, interests and personalities at variance, bargaining becomes crucial' (Gronn 1986, p. 45). The relationship between control and conflict (or domination and resistance) is conceived 'as the fundamental and contradictory base of organizational life' (Ball 1994, p. 3822).

(vi) Organizational change is nonteleological, pervasive, and value-laden since it advances the position of certain groups at the expense of the relative status of others (see Ball 1987, p. 32). Schools are not organizations easily changed; a 'quick-fix orientation' in innovation will not be appropriate to their cultural complexities (see Corbett et al. 1987, p. 57).

2. *Research in the Micropolitics of Education*

2.1 *Methods and Strategies*

If part of the micropolitical activities which are interesting for the researcher take place in privacy, if these processes are often overlain by routine activities (see Ball 1994, p. 3822), and if experienced actors regularly use strategies such as covering up actions, deceiving about real intentions etc., then researching the micropolitics of schools will have to cope with some methodological problems.

To come to grips with their object of study micropolitical studies focus on critical incidents, persons and phases that disrupt routine, for example, the appointment of a new principal, the introduction of new tasks or programs, instances of structural reorganization (e.g., a school merger), etc. They pay special attention to the processes of (re-)distribution of resources, rewards, and benefits (see Ball 1994, p. 3823) or analyze instances of organizational conflict to find out how new organizational order has been established. These strategies provide a quicker access to indicative processes, but are certainly in danger of inducing an overestimation of the proportion of conflict in organizations. Micropolitical studies try to enhance their credibility through triangulation of methods, collection of multiple perspectives, long-term engagement in the field, critical discourse in a team of researchers, and feedback from participants and external researchers.

Although there have been some attempts to tackle organizational micropolitics with quantitative means (see Blickle 1995), most studies are based on qualitative methods. There have been interview and observation studies, however, the typical strategy is a case study approach mainly based on interviews, field notes from participant or nonparticipant observation, transcripts of meetings, analyses of documents, and cross-case comparison of different sites.

A good example of such a case study is Sparkes' (1990) three-year study of a newly appointed departmental head's attempts at innovation: the head seeks to orientate Physical Education teaching towards a child-centered approach emphasizing mixed-ability grouping and individual self-paced activities which run counter to the then prevalent 'sporting ideology.' In the beginning the new head believes in rational change and democratic participation. The head tries to make other department members understand his own particular teaching approach before introducing changes to the curriculum. However, not all staff can be convinced through dialogue.

In his management position the departmental head can control the agenda of department meetings, time, and procedure of arriving at decisions, etc. He secures the agreement of crucial staff members before the meetings and networks with other influential persons outside his department. His proposals for innovation are nevertheless met with resistance by some staff. However, this resistance cannot prevent the structural changes being introduced. However, these structural changes cannot guarantee teachers' conformity to the head's educational goals within the confines of individual classrooms. Thus, he goes on to consolidate

his domination of the department by careful selection of new staff and exclusion of the unwilling.

Although being successful in that his proposals for curriculum change were implemented within one year, the departmental head ended up in a situation that was characterized by conflict and uncertainty and, thus, in many respects, counteracted his own aspirations of democratic leadership. Sparkes suggests that his findings have some bearing beyond the specific case. He warns that the 'ideology of school-centered innovation … fails to acknowledge the presence and importance of conflict and struggle both within, and between, departments … the commonly held view of teacher participation in the innovative process as 'good' in itself is highly questionable. The findings presented indicate that qualitative differences exist in the forms of participation available to those involved and that, to a large extent, these are intimately linked to the differential access that individuals have to a range of power resources' (Sparkes 1990, p. 177).

2.2 Research Topics

Typical topics of micropolitical studies are the following:

Processes of change: when organizational 'routine games' are disrupted e.g., through educational, curricular, or organizational innovations, through the recruitment of new principals or teachers, through the introduction of quality assurance systems, then intensified processes of micropolitical relevance are to be expected. Some studies investigate how externally mandated reform is dealt with in schools: when Ball and Bowe (1991, p. 23) studied the impact of the 1988 British Educational Reform Act, they found that 'the implementation of externally initiated changes is mediated by the established culture and history of the institution, and that such changes, ''their acceptance and their implementation become sites as well as the stake of internal dispute.'' The reform puts some possible alternative definitions about the meaning and the workings of schools on the agenda which have to be 'processed' in the local environments through disputes over topics such as, for example, 'serving a community versus marketing a product' or 'priority to managerial or professional views of schooling.'

Leadership and the relationships in the organization: traditionally, educational ethnography had a strong classroom orientation. Micropolitics may be understood as a move to apply ethnographic methods to study the interaction of the members of the school outside the classroom. Although parents', pupils' and ancillary staff's voices are included in some studies, most research focuses on the interaction of teachers. Since steering and regulation of schoolwork are at the heart of the micropolitical approach, it is no wonder that issues of leadership have been frequently investigated (see Ball 1987, p. 80). Holders of senior positions must acquire micropolitical skills for their career advancement and they 'have much to lose if organizational control is wrested from them' (Ball 1994, p. 3824) which makes them perfect foci for study. While many of these studies concentrated on sketching vivid images of a 'politics of subordination,' Blase and Anderson (1995) developed a framework for a 'micropolitics of empowerment' through facilitative and democratic leadership.

Socialization and professional development of teachers: the politically impregnated organizational climate in schools is the context in which individual teachers develop their professional identity (Kelchtermans and Vandenberghe 1996). During the first years of their career teachers concentrate on acquiring knowledge and competencies with respect to academic and control aspects of classroom instruction. Later on they build up a 'diplomatic political perspective,' because they feel under 'constant scrutiny' by pupils, parents, fellow teachers, and administrators (see Blase 1991, p. 189, Kelchtermans and Vandenberghe 1996). A pervasive feeling of 'vulnerability' induces many teachers to develop a range of protective strategies such as low risk grading, documentation of assessment and instruction, avoidance of risky topics and extracurricular activities, etc. (see Blase 1991, p. 193). Thus, it is no wonder that some authors claim that the acquisition of political competence is fundamental to teachers' career satisfaction (see Ball 1994, p. 3824). Since micropolitical perceptions and competencies are almost absent in beginning teachers' interpretive frameworks, attention to the induction phase is important (Kelchtermans and Vandenberghe 1996, p. 12).

Understanding schools as organizations: since micropolitics is an approach to organization theory all these studies aim to contribute to our understanding of schools as organizations. There are many studies which prove the intensity of power-impregnated activity in schools and analyze typical causes and process forms. It has been argued that political control strategies are a characteristic part of school life 'precisely because the professional norms of teacher autonomy limit the use of the effectiveness of more overt forms of control' (Ball 1994, p. 3824).

In Iannaconne's (1991) view, the unique contribution a micropolitical approach can offer lies in exploiting the metaphor of 'society' for schools: a typical school is organized like a 'caste society' since a majority of its members are 'subject to its laws without the right to share in making them' (Iannaconne 1991, p. 469). From the characteristics of a 'caste society' some hypotheses about schools may be derived, e.g.:

(a) Such societies are characterized by great tensions between the castes and a fragile balance of power.

(b) They tend to conceal the internal differences of a caste from others. Consequently, critical self-examination of castes and open debate over policy issues is difficult.

(c) They tend to avoid fundamental conflicts and to displace them into petty value conflicts.

'In schools, as in small rural towns, the etiquette of gossip characterizes teacher talk. Faculty meetings wrangle over trivial matters, avoiding philosophic and ethical issues like the plague' (Iannaconne 1991, p. 469).

3. *Criticism and Areas for Development*

The first wave of research and conceptual work in micropolitics at the beginning of the 1990s has triggered some criticism which will be discussed in order to identify possible limitations, but also areas for development.

(a) Many authors of the micropolitical approach revel in examples of illegitimate strategies: secrecy, deception, and crime add much to the thrill of micropolitical studies. Some critics argue that talking about the pervasiveness of 'illegitimate power' produces itself a 'micropoliticization' of practice and will undermine confidence in organizations. This claim implies that a second meaning of 'micropolitics' is introduced: it does not (only) refer to an approach in organization theory (which analyses both legitimate and illegitimate interactions and their effects for the structuration of an organization), but refers to those specific activities and attitudes in organizations which are based on power politics with illegitimate means. This conceptual move, however, produces problems: how should it be possible to draw the line between 'good' and 'bad organizational politics' since it were precisely situations characterized by diverse goals and contested expertise which made a concept like 'micropolitics' necessary?

(b) Although the proponents of micropolitics are right to criticize a presupposition of organizational consensus, it is unsatisfactory to interpret any consensus as a form of domination (see Ball 1987, p. 278, Ball 1994, p. 3822). Many early studies did not include cooperative or consensual interactions (see Blase 1991, p. 9). It will be necessary to develop conceptual devices to account for relationships of co-agency which neither dissolve any cooperation into subtle forms of domination nor gloss over the traces of power which are at the core of collaboration and support (see Altrichter and Salzgeber 2000, p. 104).

(c) In focusing on strategic aspects of action micropolitical approaches are again in danger of overemphasizing the 'rationality' of organizational action. To understand the workings of organizations it may be necessary to consider actions which are mainly influenced by emotions and affect, which have not been carefully planned according to some interests, as well as unintended consequences of otherwise motivated actions.

(d) Although being right to put the spotlight on the potential instability of organizational processes, early micropolitical studies had some difficulty in conceptualizing the relative stability and durability of organizations, which we can observe every day in many organizations. For this purpose, it will be necessary to bring agency and structure in a balanced conceptual relationship: micropolitics will not deal with 'relationships rather than structures' (Ball 1994, p. 3822) but precisely with the interplay between interaction and structure.

(e) The understanding of politics displayed in many early micropolitical studies was quite unsatisfactory, not to say unpolitical. Politics was often seen as the opposite of truth and reason; it was associated with treason, conspiracy and accumulation of influence as if fair and democratic ways of negotiation in organizations were inconceivable. More recent studies, for example, Blase and Anderson (1995), seem to overcome this bias. It was also criticized that in some micropolitical studies the external relationships of the organization were dissolved into the interactions between individual actors. However it must be said that some studies took care to account for the societal embedding of organizational interactions (see Ball and Bowe 1991). In the meantime there is general agreement that this must be a feature of sound micropolitical studies (see Blase 1991, p. 237).

Problems (c) and (e) may be overcome by reference to Giddens' (1992) theory of structuration, to Crozier and Friedberg's (1993) concept of organizational games, and to Ortmann et al.'s (1990) studies of power in organizational innovation. On this conceptual basis, organizational structure may be seen as an outcome of and resource for social actions which enables and restricts actions, but never fully determines them. Organizations are webs woven from concrete interactions of (self-)interested actors. They acquire some spatial-temporal extension through the actors' use and reproduction of 'games.' The 'game' is the point of conceptual intersection in which the analytically divided elements of agency and structure, of individual and organization meet. The game works as an 'indirect integration mechanism which relates the diverging and/or contradictory action of relatively autonomous actors' (Crozier and Friedberg 1993, p. 4). In order to take action and to pursue their own interests members of organizations must at least partly use structural resources of the organization i.e., they are forced to play within the organizational rules, and thereby, they reproduce these rules. The functioning of an organization may be explained as 'result of a series of games articulated among themselves, whose formal and informal rules indirectly integrate the organization members' contradictory power strategies' (Ortmann et al. 1990, p. 56). The interlocked character of the games and the high amount of routine elements in social practices are major reasons for the relative stability of organizations (see Altrichter and Salzgeber 2000, p. 106).

See also: Educational Innovation, Management of; Educational Leadership; Group Processes in Organizations; Leadership in Organizations, Psychology of; Organization: Overview; Organizational Decision Making; School as a Social System; School Improvement; School Management

Bibliography

Altrichter H, Salzgeber S 2000 Some elements of a micropolitical theory of school development. In: Altrichter H, Elliott J (eds.) *Images of Educational Change*. Open University Press, Buckingham, UK, pp. 99–110

Bacharach S, Lawler E 1980 *Power and Politics in Organizations*. Jossey-Bass, San Francisco, CA

Ball S J 1987 *The Micro-Politics of the School*. Routledge, London

Ball S J 1994 Micropolitics of schools. In: Husen T, Postlethwaite T N (eds.) *The International Encyclopedia of Education*, 2nd edn. Pergamon, Oxford, UK, pp. 3821–6

Ball S J, Bowe R 1991 Micropolitics of radical change. Budgets, management, and control in British schools. In: Blase J (ed.) *The Politics of Life in Schools*. Sage, Newbury Park, CA, pp. 19–45

Blase J (ed.) 1991 *The Politics of Life in Schools*. Sage, Newbury Park, CA

Blase J, Anderson G L 1995 *The Micropolitics of Educational Leadership*. Cassell, London

Blickle G 1995 Wie beeinflussen Personen erfolgreich Vorgesetzte, KollegInnen und Untergebene? [How do persons successfully influence superiors, colleagues, and subordinates?] *Diagnostica* **41**: 245–60

Corbett H D, Firestone W A, Rossman G B 1987 Resistance to planned change and the sacred in school cultures. *Education Administration Quarterly* **23**: 36–59

Crozier M, Friedberg E 1993 *Die Zwaenge kollektiven Handelns*. Hain, Frankfurt/M., Germany [1980 *Actors and Systems*. University of Chicago Press, Chicago, IL]

Giddens A 1992 *Die Konstitution der Gesellschaft*. Campus, Frankfurt/M., Germany [1984 *The Constitution of Society*. Polity Press, Cambridge, UK]

Gronn P 1986 Politics, power and the management of schools. In: Hoyle E, McMahon A (eds.) *World Yearbook of Education*. Kogan Page, London

Hoyle E 1982 Micropolitics of educational organizations. *Educational Management and Administration* **10**: 87–98

Iannaconne L 1991 Micropolitics of education–what and why. *Education and Urban Society* **23**: 465–71

Kelchtermans G, Vandenberghe R 1996 Becoming political: a dimension in teachers' professional development. *Paper presented at the AREA-conference*. New York (ERIC-document: ED 395-921)

Ortmann G, Windeler A, Becker A, Schulz H-J 1990 *Computer und Macht in Organisationen [Computers and power in organizations]*. Westdeutscher Verlag, Opladen, Germany

Pfeffer J 1981 *Power in Organizations*. Pitman, Boston, MA

Sparkes A C 1990 Power, domination and resistance in the process of teacher-initiated innovation. *Research Papers in Education* **5**: 153–78

H. Altrichter

Schumpeter, Joseph A (1883–1950)

Joseph A. Schumpeter is one of the best known and most flamboyant economists of the twentieth century. His main accomplishments are to have introduced the entrepreneur into economic theory through *Theorie der wirtschaftlichen Entwicklung* (1912) and to have produced a magnificent history of economic thought, *History of Economic Analysis* (posthumously published in 1954). Schumpeter's interests were wide ranging, and his writings also include contributions to sociology and political science. His analysis of democracy, which can be found in his most popular work, *Capitalism, Socialism, and Democracy* (1942), is generally regarded as an important addition to the theory of democracy.

1. Early Life and Career

Joseph Alois Schumpeter was born on February 8, 1883 in the small town of Triesch in the Austro-Hungarian Empire (today Trest in Slovakia). His father was a cloth manufacturer and belonged to a family which had been prominent in the town for many generations; and his mother, who was the daughter of a medical doctor, came from a nearby village. Schumpeter's father died when he was four years old, an event that was to have great consequences for the future course of his life. Some time later his mother moved to Graz and then to Vienna, where she got married to a retired officer from a well known, aristocratic family, Sigismund von Kéler. Schumpeter was consequently born into a small-town bourgeois family but raised in an aristocratic one; and his personality as well as his work were to contain a curious mixture of values from both of these worlds.

After having graduated in 1901 from the preparatory school for the children of the elite in the Empire, Theresianum, Schumpeter enrolled at the University of Vienna. His goal from the very start was to become an economist. Around the turn of the century the University of Vienna had one of the best educations in economics in the world, thanks to Carl Menger and his two disciples Eugen von Böhm-Bawerk and Friedrich von Wieser. Schumpeter had the latter two as his teachers, and he also attended lectures in mathematics on his own. When Schumpeter received his doctorate in 1906, he was 23 years old and the youngest person to have earned this degree in the Empire.

After having traveled around for a few years and added to his life experience as well as to his knowledge of economics, Schumpeter returned in 1908 to Vienna to present his *Habilitationsschrift*. Its title was *Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie (*1908; *The Nature and Essence of Theoretical Economics)*, and it can be characterized as a work in economic methodology from an analytical perspective. In 1909 Schumpeter was appointed as-

sistant professor at the University of Czernowitz and thereby became the youngest professor in German-speaking academia. It was also at Czernowitz that Schumpeter produced his second work, the famous *Theorie der wirtschaftlichen Entwicklung* (1912; 2nd edn. from 1926 trans. as *The Theory of Economic Development*). At the University of Graz, to which he moved as a full professor in 1911, Schumpeter produced a third major study, *Epochen der Dogmen- und Methodengeschichte* (1914; trans. in 1954 as *Economic Doctrine and Method*). Over a period of less than ten years Schumpeter had produced three important works, and one understands why he referred to the third decade in a scholar's life as 'the decade of sacred fertility.'

1.1 The Works from the Early Years

Das Wesen is today best known for being the place where the term 'methodological individualism' was coined. In its time, however, this work played an important role in introducing German students to analytical economics, which Gustav von Schmoller and other members of the Historical School had tried to ban from the universities in Germany. *Das Wesen* is well written and well argued, but Schumpeter would later regard it as a youthful error and did not allow a second edition to appear. The reason for Schumpeter's verdict is not known, but it is probably connected to his vigorous argument in the book that economics must sever all links with the other social sciences.

While the emphasis in *Das Wesen* had been exclusively on statics, in his second book, *Theorie der wirtschaflichen Entwicklung*, it was on dynamics. Schumpeter very much admired Walras' theory of general equilibrium, which he had tried to explicate in *Das Wesen*, but he was also disturbed by Walras' failure to handle economic change. What was needed in economics, Schumpeter argued, was a theory according to which economic change grew out of the economic system itself and not, as in Walras, a theoretical scheme in which economic change was simply explained as a reaction to a disturbance from outside the economic system.

Schumpeter starts the argument in *Theorie* by presenting an economy where nothing new ever happens and which therefore can be analyzed with the help of static theory, that is all goods are promptly sold, no new goods are produced or wanted, and profits are zero. Schumpeter then contrasts this situation of a 'circular flow' with a situation in which the activities of the entrepreneur sets the whole economic system in motion, and which can only be explained with the help of a dynamic theory. The model that Schumpeter now presented in *Theorie* would later be repeated and embellished upon in many of his writings: the entrepreneur makes an innovation, which earns him a huge profit, this profit attracts a swarm of other, less innovative entrepreneurs and as a result of all these activities a wave of economic change begins to work its way through the economic system, a business cycle, in brief, has been set in motion. The theory of business cycles that one can find in *Theorie* is fairly rudimentary, and Schumpeter would later spend much time in improving upon it.

Much of Schumpeter's analysis in *Theorie* is today forgotten and few economists are today interested in the way that Schumpeter explains interest, capital and profit by relating these to the activities of the entrepreneur. Some parts of *Theorie*, however, are still very much alive and often referred to in studies of entrepreneurship. In one of these Schumpeter discusses the nature of an innovation, and explains how it differs from an invention; while an invention is something radically new, an innovation consists of an attempt to turn something—for example, an invention—into a money-making enterprise. In another famous passage Schumpeter discusses the five main forms of innovations:

> (1) The introduction of a new good … or of a new quality of a good. (2) The introduction of a new method of production … (3) The opening of a new market … (4) The conquest of a new source of supply of raw materials or half-manufactured goods … (5) The carrying out of the new organization of any industry, like the creation of a monopoly position … or the breaking up of a monopoly position.

Schumpeter's description of what drives an entrepreneur is also often referred to, as is his general definition of entrepreneurship: 'the carrying out of new combinations.'

Schumpeter's third work from these early years is a brief history of economics, commissioned by Max Weber (1914–20) for a handbook in political economy (*Grundriss der Sozialökonomik*). Its main thesis is that economics came into being when analytical precision, of the type that one can find in philosophy, came into contact with an interest in practical affairs, of the type that is common among businessmen; and to Schumpeter this meeting took place in the works of the Physiocrats. *Epochen der Dogmen- und Methodengeschichte* was later overshadowed by the much longer *History of Economic Analysis*, but while the latter work tends to be used exclusively as a reference work, Schumpeter's study from 1914 can be read straight through and is also more easy to appreciate.

1.2 Difficult Mid-Years

By the mid-1910s Schumpeter had established himself as one of the most brilliant economists in his generation, and like other well-known economists in the Austro-Hungarian Empire he hoped for a prominent political position, such as finance minister or economic adviser to the Emperor. During World War I Schumpeter wrote a series of memoranda, which were

circulated among the elite of the Empire, and through which he hoped to establish himself as a candidate for political office. The various measures that Schumpeter suggested in these memoranda are of little interest today, for example, that the Empire must not enter into a customs union with its powerful neighbor, Germany. What is of considerably more interest, however, is that they provide an insight into Schumpeter's political views when he was in his early 30s. Schumpeter's political profile at this time can be described in the following way: he was a royalist, deeply conservative, and an admirer of tory democracy of the British kind.

Schumpeter's attempt during World War I to gain a political position failed, but his fortune changed after the Empire had fallen apart and the state of Austria came into being. In early 1919 Schumpeter was appointed finance minister in a joint Social Democratic and Catholic Conservative government in Austria, led by Karl Renner. Already by the fall of 1919, however, Schumpeter was forced to resign, mainly because the Social Democrats felt that he had betrayed them and could not be trusted. The Social Democrats in particular thought that Schumpeter had gone behind their backs and stopped their attempts to nationalize parts of Austria's industry. Schumpeter always denied this, and in retrospect it is difficult to establish the truth. What is clear, however, is that the Austrian Social Democrats advocated a democratic and peaceful form of socialism, while Schumpeter detested all forms of socialism.

Loath to go back to his position at the University of Graz, Schumpeter spent the early part of the 1920s working for a small bank in Vienna, the Biedermann Bank. Schumpeter did not take part in the everyday activities of the bank but mainly worked as an independent investor. These activities ended badly, however, and by the mid-1920s Schumpeter was forced to leave the Biedermann Bank and had by this time accumulated a considerable personal debt. For many years to come, Schumpeter would give speeches and write articles in order to pay back his debts. Adding to these economic misfortunes was the death in 1926 of his beloved wife Anna ('Annie') Reisinger, a 20 years younger working class girl to whom Schumpeter had only been married for a year. From this time on, Schumpeter's friends would later say, there was a streak of pessimism and resignation in his personality.

In 1925 Schumpeter was appointed professor of public finance at the University of Bonn, and it is clear that he was both pleased and relieved to be back in academia. While his ambition, no doubt, extended to politics and business, he always felt an academic at heart. The great creativity that Schumpeter had shown in the 1910s was, however, not to return in the 1920s, and Annie's death was no doubt an important reason for this. Schumpeter's major project during the 1920s was a book in the theory of money, but this work was never completed to his satisfaction Schumpeter 1970). Schumpeter did, however, write several interesting articles during these years, some of which are still worth reading. One of these is *'The Instability of Capitalism'* (1928), in which Schumpeter suggests that capitalism is undermining itself and will eventually turn into socialism. Another worthwhile article is *'Gustav vs. Schmoller und die Probleme von heute'* (1926), in which Schumpeter argues that economics has to be a broad science, encompassing not only economic theory but also economic history, economic sociology and statistics. Schumpeter's term for this type of general economics is *'Sozialökonomik.'*

To illustrate what economists can accomplish with the help of sociology, Schumpeter wrote an article on social classes (1927), which is still often referred to. Schumpeter starts out by drawing a sharp line between the way that the concept of class is used in economics and in sociology. In economics, Schumpeter says, class is basically used as a category (e.g., wage-earner/non-wage-earner), while in sociology class is seen as a piece of living reality. An attempt was also made by Schumpeter to link his theory of the entrepreneur to the idea of class.

In mentioning Schumpeter's articles on social classes, something must also be said about two of his other famous studies in sociology: *'The Fiscal Crisis of the Tax State'* (1918) and *'The Sociology of Imperialisms'* (1918–19). The former study can be described as a pioneer effort in the sociology of taxation, which focuses on the situation in Austria just after World War I but which also attempts to formulate some general propositions about the relationship between taxation and society. Schumpeter, for example, suggests that if the citizens demand more and more subventions from the state, but are unwilling to pay for these through ever higher taxes, the state will collapse. In the article on imperialism, Schumpeter argues that imperialism grows out of a social situation which is *not* to be found in capitalism; and whatever imperialist forms that still exist represent an atavism and a leftover from feudal society. Schumpeter's famous definition of imperialism is as follows: 'imperialism is the objectless disposition on the part of the state to unlimited forcible expansion.'

2. *The Years in the United States*

By 1930 Schumpeter was still depressed about his life and quite unhappy about his career; he had found no one to replace his wife on an emotional level and he felt that he deserved a better professional position than Bonn. When Werner Sombart's chair at the University of Berlin became vacant, Schumpeter had high hopes that he would get it. When this did not happen, however, he instead accepted an offer from Harvard University; and in 1932 he moved to Cambridge, Massachusetts for good. Schumpeter was slowly to regain his emotional composure, and also his capacity

to work, in the United States. In 1937 he got married to economist Elizabeth Boody, who regarded it as her duty in life to take care of Schumpeter and manage his worldly affairs. With the support of his wife, Schumpeter produced three major works: *Business Cycles* (1939), *Capitalism, Socialism and Democracy* (1942), and *History of Economic Analysis* (published posthumously in 1954).

2.1 The Works from the American Period

Of these three works, the one for which Schumpeter had the highest hopes was definitely *Business Cycles* (1939), which appeared exactly 25 years after his last book in Europe. This work is more than 1100 pages long and it traces business cycles in Great Britain, the United States and Germany during the years 1787–1938. Schumpeter had hoped that *Business Cycles* would get the kind of reception that Keynes' *General Theory* got, and he was deeply disappointed by the lack of interest that his colleagues and other economists showed. The impression one gets from histories of economic thought, however, is that *Business Cycles* is rarely read today, mainly because few people are persuaded by Schumpeter's theory that business cycles are set off by innovations and that their motion can be best described as a giant Kondratieff cycle, modified by Juglar and Kitchin cycles; Schumpeter's handling of statistics is also considered poor. To a large extent this critique is true, although it should also be noted that a number of themes and issues in this and other of Schumpeter's works are increasingly being referred to in evolutionary economics.

After many years of excruciatingly hard work on *Business Cycles*, Schumpeter decided that he needed an easier task to work on than the recasting of economic theory that was on his agenda since the time in Bonn. In the late 1930s Schumpeter therefore began working on what he saw as a minor study of socialism. This minor study, however, quickly grew in size, and when it finally appeared in 1942—under the title *Capitalism, Socialism, and Democracy*—it was nearly 400 pages long. Of all of Schumpeter's works (1942), this book was to become the most popular one, especially in the United States where it is still considered a standard work in political science. *Capitalism, Socialism, and Democracy* has also been translated into more languages than any other work by Schumpeter, and it is still in print.

Capitalism, Socialism, and Democracy contains an interesting history of socialist parties (part V), a fine discussion of the nature of socialism (part III), and an excellent discussion of the work of Karl Marx (part I). What has made this work into a social science classic, however, are its two remaining parts with their discussion of capitalism (part II) and the nature of democracy (part IV). Capitalism, Schumpeter says, can be characterized as a process of continuous economic change, through which new enterprises and industries are continuously being created and old ones continuously being destroyed. Schumpeter's famous term for this process is 'creative destruction.' Perfect competition of the type that economists talk about, he also says, is a myth; and monopolies, contrary to common belief and economists' dogma, are often good for the economy. Without monopolies, for example, very expensive forms of research and investments would seldom be undertaken.

Schumpeter's theory of monopolies has led to a large number of studies of innovation. By 'the Schumpeterian hypothesis' two propositions are meant: that large firms are more innovative than small firms, and that innovation tends to be more frequent in monopolistic industries than in competitive ones. Each of these hypotheses can in their turn be broken down into further hypotheses. Many decades later the results of all of these studies are still inconclusive, and there is no consensus if monopolistic forms of enterprise further or obstruct innovation (for an overview of research on the Schumpeterian hypothesis, see Kamien and Schwartz 1982).

Schumpeter's overall theory in *Capitalism, Socialism, and Democracy*—the idea that capitalism is on its way to disappear and be replaced by socialism—has, on the other hand, led to little concrete research. One reason for this is probably that many readers of Schumpeter's book have felt that his argument is very general in nature and at times also quite idiosyncratic. Schumpeter argues, for example, that capitalism tends to breed intellectuals, and that these are hostile to capitalism; that the bourgeois class is becoming unable to appreciate property; and that the general will of the bourgeoisie is rapidly eroding. 'Can capitalism survive?' Schumpeter asks rhetorically, and then gives the following answer, 'No. I do not think it can.' Schumpeter's conclusion about the inevitable fall of capitalism has often been sharply criticized (e.g., Heertje 1981). His more general attempt to tackle the question of how social and economic institutions are needed for an economic system to work properly has, however, remained unexplored. There also exists a certain tendency to cite catchy phrases from Schumpeter's work without paying attention to their organic place in his arguments.

What has always been much appreciated in *Capitalism, Socialism, and Democracy* is Schumpeter's theory of democracy. Schumpeter starts out by criticizing what he calls 'the classical doctrine of democracy,' which he characterizes in the following way: 'the democratic method is that institutional arrangement for arriving at political decisions which realizes the common good by making the people itself decide issues through the election of individuals who are to assemble in order to carry out its will.' What is wrong with this theory, he argues, is that there is no way of establishing what the common good is. There is also the additional fact that politicians do not have interests

of their own, according to this theory; their function is exclusively to translate the desires of the people into reality.

To the classical type of democracy Schumpeter counterposes his own theory of democracy, which he defines in the following manner: 'the democratic method is that institutional arrangement for arriving at political decisions in which individuals acquire the power to decide by means of a competitive struggle for the people's vote.' Some critics have argued that Schumpeter's theory is elitist in nature, since democracy is reduced to voting for a political leader at regular intervals. This may well be true, and Schumpeter's theory of democracy no doubt coexists in his work with contempt for the masses. Regardless of this critique, however, Schumpeter's theory is generally regarded as being considerably more realistic than the classical doctrine of democracy.

World War II was a very trying time for Schumpeter since he felt emotionally tied to German speaking Europe, even though he by now was a US citizen. Schumpeter did not support the Nazis but like many conservatives felt that Communism constituted a more dangerous enemy than Hitler. Schumpeter's intense hatred of Roosevelt, whom he suspected of being a socialist, also put him on a collision course with the academic community in Cambridge during World War II. In his diary Schumpeter expressed his ambivalence towards Hitler, Jews, Roosevelt and many other things; and in his everyday life he more and more withdrew into solitude and scholarship. The War, in brief, was a very difficult and demoralizing time for Schumpeter.

From the early 1940s Schumpeter started to work feverishly on a giant history of economic thought that was to occupy him till his death some ten years later. He did not succeed in completing *History of Economic Analysis*, which was instead put together by Elizabeth Boody Schumpeter after several years of work. *History of Economic Analysis* is many times longer than Schumpeter's history of economics from 1914, and it is also considerably more sophisticated in its approach. Schumpeter's ambition with this work, it should first of all be noted, was not only to write the history of economic theory, but also to say something about economic sociology, economic history, and economic statistics, the four 'fundamental fields' of economics or *Sozialökonomik*. It should also be mentioned that *History of Economic Analysis* begins with a very interesting methodological discussion of the nature of economics. This is then followed by a brilliant, though occasionally willful discussion of economics from classical Greece to the years after World War I. The section on twentieth century economics was unfortunately never completed, but it is clear that Schumpeter was very discontent with the failure of his contemporaries to create a truly dynamic economic theory.

Schumpeter died in his sleep on January 8, 1950 and the cause was put down as cerebral hemorrhage. Some of his friends thought that the real reason was overwork, while his diary shows that he had been tired of life for a long time. Schumpeter's private notes also show that behind the image of brilliant lecturer and iconoclastic author, which Schumpeter liked to project, there was a person who felt deeply unhappy with the way that his life had evolved and with his failure to produce a theory that would revolutionize economic thought. 'With a fraction of my ideas,' he wrote in dismay a few years before his death, 'a new economics could have been founded.'

See also: Business History; Capitalism; Democracy; Democracy, History of; Democracy: Normative Theory; Democratic Theory; Determinism: Social and Economic; Development, Economics of; Development: Social; Economic Growth: Theory; Economic History; Economic Sociology; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Economics: Overview; Economics, Philosophy of; Entrepreneurship; Entrepreneurship, Psychology of; Imperialism, History of; Imperialism: Political Aspects; Innovation and Technological Change, Economics of; Innovation, Theory of; Marxian Economic Thought; Policy History: State and Economy; Political Economy, History of; Social Democracy; Socialism; Statistical Systems: Economic

Bibliography

Allen R L 1991 *Opening Doors: The Life and Work of Joseph Schumpeter*. Transaction Publishers, New Brunswick, NJ

Augello M 1990 *Joseph Alois Schumpeter: A Reference Guide*. Springer-Verlag, New York

Harries S E (ed.) 1951 *Schumpeter: Social Scientist*. Books for Libraries Press, Freeport, New York

Heertje A (ed) 1981 *Schumpeter's Vision: Capitalism, Socialism and Democracy After 40 Years*. Praeger, New York

Kamien M I, Schwartz N L 1982 *Market Structure and Innovation*. Cambridge University Press, Cambridge, UK

Schumpeter J A 1908 *Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie*. Duncker and Humblot, Leipzig

Schumpeter J A 1912 *Theorie der wirtschaftlichen Entwicklung*. Duncker & Humblot, Leipzig (actually appeared in 1911, 2nd edn. 1926). [trans. into English in 1934 as *The Theory of Economic Development*, Harvard University Press, Cambridge, MA]

Schumpeter J A 1914 *Epochen der Dogmen- und Methodengeschichte*. In: Bücher et al. (eds.) *Grundriss der Sozialökonomik. I. Abteilung, Wirtschaft und Wirtschaftswissenschaft*. J.C.B. Mohr, Tübingen, Germany, pp. 19–124

Schumpeter J A 1926 Gustav vs. Schmoller und die probleme von heute. *Schmollers Jahrbuch* **50**: 337–88

Schumpeter J A 1939 *Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process*. McGraw-Hill, New York

Schumpeter J A 1942 *Capitalism, Socialism, and Democracy*. Harper and Brothers, New York

Schumpeter J A 1954 *History of Economic Analysis*. Oxford University Press, New York
Schumpeter J A 1954 *Economic Doctrine and Method*, Oxford University Press, New York
Schumpeter J A 1970 *Das Wesen des Geldes*. Vandenhoeck and Ruprecht, Göttingen, Germany
Schumpeter J A 1989 The instability of capitalism. In: Schumpeter J A (ed.) *Essays*. Transaction Publishers, New Brunswick, NJ
Schumpeter J A 1991 The crisis of the tax state. In: Schumpeter J A (ed.) *The Economics and Sociology of Capitalism*. Princeton University Press, Princeton, NJ
Schumpeter J A 1991 The sociology of imperialisms. In: Schumpeter J A (ed.) *The Economics and Sociology of Capitalism*. Princeton University Press, Princeton, NJ
Schumpeter J A 1991 Social classes in an ethically homogenous milieu. In: Schumpeter J A (ed.) *The Economics and Sociology of Capitalism*. Princeton University Press, Princeton, NJ
Schumpeter J A 1991 *The Economics and Sociology of Capitalism*. Princeton University Press, Princeton, NJ
Schumpter J A 2000 *Briefe/Letters*. J.C.B. Mohr, Tübingen, Germany
Stolper W F 1994 *Joseph Alois Schumpeter: The Private Life of a Public Man*. Princeton University Press, Princeton, NJ
Swedberg R 1991 *Schumpeter: A Biography*. Princeton University Press, Princeton, NJ
Weber M (ed.) 1914–20 *Grundriss der Sozialökonomik*. J.C.B. Mohr, Tübingen, Germany

R. Swedberg

Schütz, Alfred (1899–1959)

Alfred Schütz was born in Vienna on April 13, 1899. Here he studied law and social sciences from 1918 to 1921. After receiving his Doctorate of Law he continued his studies in the social sciences until 1923. His teachers included Hans Kelsen, an eminent lawyer, Ludwig von Mises, and Friederich von Wieser, prominent representatives of the Austrian school of economics. Equally important as his university studies was his participation in the informal academic life of Vienna, e.g., his participation in the private seminars of von Mises, where Schütz also cultivated his philosophical interests. In addition to his scholarly activities, Schütz held a full-time job as a bank attorney, a dual life which lasted until 1956. After the annexation of Austria to the Third Reich, Schütz and his family escaped to New York. On his arrival, Schütz got in touch with Talcott Parsons whose *Structure of Social Action* he intended to review. But despite their intense correspondence (Schütz and Parsons 1978), the contact between Schütz and Parsons broke down. Instead, Schütz became affiliated with the Graduate Faculty at the New School for Social Research in New York where he started teaching as Lecturer in 1943, advancing to a full Professor of Sociology in 1952 and also of Philosophy in 1956. He died on May 20, 1959 in New York.

1. Major Aims of Schütz's Theory

Alfred Schütz's social theory focuses on the concept of the life world, which is understood as a social-cultural world as it is perceived and generated by the humans living in it. Schütz identifies everyday interaction and communication as being the main processes in which the constitution of the life world's social reality takes place. In this sense, his approach pursues two major goals: the development of a theory of action which reveals the constitution of social reality with its meaning structure given in the commonly shared typical patterns of knowledge, and a description of the life world with its multiple strata of reality emerging from these constituting processes. These leitmotifs merge in a theory of life world constitution—a project which occupied Schütz during the final years of his life.

2. Main Intellectual Sources of the Schützian Approach

Schütz's conception originated in the intellectual discourse in the social sciences and philosophy of the early decades of this century. Schütz developed his position from within a triangular interdisciplinary field marked by Max Weber's historising interpretative sociology (*verstehende Soziologie*), the Austrian economic approach represented by Ludwig von Mises, and the philosophical theories of the stream of consciousness formulated by H. Bergson and E. Husserl. From the very onset, he adopted Max Weber's view of social reality as a meaningful social-cultural world, as well as his concept of social action as a meaning oriented behavior, and shared Weber's search for an interpretative method in sociology. He also shared the methodological individualism advocated by Weber and by the Austrian economists who denoted individual action as the starting point of social research. At the same time, he accepted the need for a generalizing theory of action, as stressed by his teacher Ludwig von Mises, and criticized Weber for neglecting to develop the basic framework of such a theory and especially for not inquiring into the constitution of meaning attached to social action. But Schütz's criticism also addressed the 'general theory of action' based on an '*a priori*' postulate of *rational choice* put forth by Ludwig von Mises. Schütz rejected this conception since it imposed an abstract framework on actors' orientation of action and ignored their actual stock of knowledge.

In order to analyze how the meaning attached to action is revealed, Schütz referred to the philosophical concepts of Henri Bergons and Edmund Husserl. Both offer insights into the stream of lived experience and into the acts of consciousness in which our world is meaningfully constituted. Schütz adopted the Bergsonian idea of the stream of consciousness in his first scholarly attempts in 1924–28 (Schütz 1982), only to later recognize the difficulties in Bergson's intuitivism

and to turn to Husserl's phenomenology which ultimately gave the name to his approach. Husserl was concerned with the acts of consciousness which establish the meaningful taken-for-grantedness of the world in humans' natural attitude. His term 'life world,' which Schütz later assumes, is aimed at the reality constituted in those acts. Based primarily on this philosophical approach, Schütz wrote his first monograph, *Der sinnhafte Aufbau der sozialen Welt*, in 1932, in which he developed the basic concept of his theory (Schütz 1932). Starting with the criticism on Max Weber mentioned above, he devoted himself to the process in which humans create the social world as a reality that is meaningful and understandable to them.

3. *Theory of the Life World with its Everyday Core*

In order to grasp the constitution of the social world, Schütz (1932) had to transcend the realm of consciousness and perception analyzed by Husserl. He included both acts of consciousness as well as human action and interaction into the constitutive process under scrutiny. Departing from the transcendental philosophical approach, he developed his own *mundane* phenomenology which analyzes the constitution of the meaningful reality within social relationships in the everyday world.

He proceeded in three steps dealing with three problems constitutive for any theory of action which is searching for the construction of social reality.

(a) How does a meaningful orientation of human action emerge?

(b) How can we understand the Other?

(c) How is socially-shared, intersubjective valid knowledge generated?

Following the principle of methodological individualism, Schütz started with the analysis of single individual action. Relying on Bergson's concept of the inner stream of lived experience and on Husserl's investigations on intentionality and temporality of consciousness, Schütz first explored the constitution of meaning in subjects as a temporal process. The meaning attached to a lived experience emerges from successive acts of selective reflection aimed towards it and framing it into a context of other experiences. The schemes of experience on which the primary meaning of action, i.e., its project is based, arise from this basic temporal dynamics and plasticity of consciousness. The temporality of the meaning attached to an action manifests itself in two kinds of motivation attached to its different temporal dimensions: '*in-order-to-motives*' which direct action toward the future, and '*because-motives*' representing the roots of the action in the past. However, the meaning of the project of action changes during the time when the projected action is going on. The final meaning of an action can be found in the experience perceived by the subject when looking at the changes that have emerged in the meaning structure of the action when the action is completed. The meaning attached to an action and consequently the schemes of experience are thus not only affected by the acts of consciousness but also by the action itself.

In his second step, Schütz proceeded to show how the schemes of experience are shaped by interaction and communication. He perceived communication as a process in which two subjective streams of consciousness are coordinated within a social interaction (*Wirkensbeziehung*). Thus, communication signifies an interaction in which the meaning of ego's action consists in the intention to evoke a reaction of the alter. Actions in this sense have the function of signs which are mutually indicated and interpreted. The ultimate meaning of my action is revealed in the reaction of the Other and vice versa, therefore communication generates a *chain of motives* where my *in-order-to-motives* become *because-motives* of the Other and provide a common stock of shared patterns of interpretation which allows for mutual understanding even if each of the actors are always referring to their own schemes of experience. In this concept of understanding, based on interaction, Schütz offers his own solution to the problem of intersubjectivity posed by Husserl.

The constitution and appropriation of shared knowledge primarily takes place in long-lasting face-to-face interaction (we-relations) where the mutual expectations are learned, verified, and sedimented to typical patterns that can be applied to more remote and anonymous strata of the social world. As a result, the meaning structure of the social world is characterized by *typifications* of actions, situations, and persons generated in interaction and communication.

In these three steps, Schütz laid down the main features of his theory of the life world. In his later work, Schütz (1962, 1964, 1966) determined this communicatively created social reality as the world of everyday life, in which typical patterns are taken for granted, and which represents the intersubjective common core of the reality we live in. Later on (Schütz 1962, 1970), he disclosed further structural characteristics of this everyday core of the life world. Since its typical structure greatly depends on action, it is also the pragmatic orientation selecting the areas where typification processes take place. Both typicality and this kind of selection, which Schütz designated as *pragmatic relevance*, represent two generative principles of order in the everyday world. They determine its formal structure and at the same time, when realized in action, they shape this structure into a concrete social–cultural world, which is thus characterized by social distribution and differentiation of knowledge. A third moment structuring the everyday world can be found in the rootedness of its constitution in individual action. Here actors and their bodies represent the

central point of the *everyday world and its temporal, spatial, and social dimensions* that are arranged into spheres of past and future, of within-reach and distant and of intimacy and anonymity, respectively, to the ctors' own position (Schütz and Luckmann 1973).

Everyday reality is nevertheless not identical with the life world on the whole. By suspending their pragmatic interests, actors are able to modify their everyday experiences and perceive them as objects of a game, fantasy, art, science, or, if the conscious attention is completely absent, as a dream. There are areas—even in the realm of everyday action directed by the principal of pragmatic relevance—which are beyond the sphere of actors everyday practice and therefore transcend it. All these modifications represent different provinces of meaning transcending the everyday world and constituting *multiple realities* (Schütz 1962) which make up the life world. Nevertheless, among the different provinces of the life world, the everyday core denotes a *paramount reality* since it is only here that actors, driven by the *fundamental anxiety* when facing the finality of their lives, have to master their living conditions and are engaged in communicative processes producing common knowledge which makes mutual understanding possible.

Schütz (1962) viewed communication as a substantial constitutive mechanism of social reality and stressed the role of *language* in this process. He considered language—including its syntax and semantics—as an objectivation of sedimented, social provided stock of knowledge that preserves relevances and typification inherent to cultures and social groups, and thus as crucial for the constitution of the life world (Schütz and Luckmann 1989). Schütz saw language as delineating an essential case of social objectified systems of appresentation which bridges the transcendence between different areas and meaning provinces of the life world. Another important integrative mechanism in the life world are symbolic structures, which are often based on language, and intermediate between the everyday world and non-everyday realities such as religion, arts, or politics (Schütz 1962).

Schütz did not restrict his study of the structure of the life world solely to theoretical research. He also applied his theory as a general interpretative scheme to several social fields and problems. He explored courses of intercultural communication and the social distribution of knowledge, the social conditions of equality in modern societies as well as the fields of music and literature (Schütz 1964).

4. *Consequences of the Theory of Life World for the Methodology of the Social Sciences*

Schütz considered the construction of ideal types to be the general methodological device in the social sciences since sociology, economy, political science, etc., explain social phenomena by modeling ideal actors and courses of action which they endorse with special features as explanatory variables. However, he did not see ideal types in the Weberian sense, i.e., as a scientific model which does not need to correspond to reality in all respects. For Schütz, the ideal-typical method is legitimized by his findings on the typicality of everyday world which is the object of social research. Social reality can be approached by type-construction on the scientific level because its immanent structure is typified itself. Thus, the Schützian theory of the life world and its structures represents a methodological tool to bridge the gap between sociological theoretical reasoning and its object. The methodological rule which Schütz derived from his theoretical approach consists correspondingly in the *postulate of adequacy* (Schütz 1962, 1964) between everyday and scientific typifications. This postulate holds that ideal types featured by the social sciences have to be constructed in correspondence to the structure of the everyday typifications, so that everyday actors can take the model for granted when they act under the conditions stated in the ideal type. In this sense, social scientists as well as everyday actors have to follow the same frame of reference given by the structure of the life world but their cognitive attitude differ in several respects: scientists neither share the pragmatic interest of everyday actors nor do they share their everyday rationality restricted by their beliefs in the grantedness of typical knowledge. Opposing T. Parson's functionalism and C. G. Hempel's methodologic positivism, Schütz stressed that scientists must not impose their own theoretical concepts and rationality on their object of study—the life world. First, they must discover and then respect its immanent meaning structure.

5. *The Significance of the Schützian Approach for the Social Sciences*

Since the 1960s, Schütz's theory has drawn sociologists' attention to inquiries into everyday interaction, communication as well as to the insight that social reality has to be considered as a construction produced within these processes. Sociologists started to examine the practices of everyday action, communication and interpretation from which social reality emerges. H. Garfinkel's (1967) *Ethnomethology*, which was aimed at finding the formal properties of everyday practices, led to a series of case studies covering a wide range of everyday life in society and its institutions. Continuing this line of research, examinations of everyday communication were provided by E. A. Shegloff and H. Sacks (1995), whose *Conversational Analysis* became a widespread method in qualitative sociological research. A. Cicourel's (1964) *Cognitive Sociology* revealed the constructed character of data in social institutions as well as in science and thus initiated a

series of studies in the sociology of organization and in the sociology sciences. *Milieus* as formations of everyday interaction are the subject of the *Social Phenomenology* R. Grathoff (1986).

P. L. Berger and Th. F. Luckmann's (1966) idea of *the Social Construction of Reality* (as being a general process in which cultural worlds emerge) triggered new impulses in both the sociology of knowledge and the sociology of culture reconceived now in Schützian terms. In this context, Schütz's impact can also be seen in the sociology of language and the sociology of religion. Contemporary Marxian theory saw the way in which Schütz focused on the everyday practice as a possible mediation between social structure and individual consciousness. The term *phenomenological sociology* was coined in the 1970s (G. Psathas 1973) to address the spectrum of approaches oriented and inspired by Schütz's theory. Under this label, Schütz's approach became one of the general paradigms in the interpretative social science and theory of action. The diffusion and empirical application of the Schützian approach enforced the search for qualitative research methods which would reveal data pertaining to the social construction of social reality in everyday life. Aside from ethnomethodology and conversational analysis mentioned above, this quest especially led to a refinement in the techniques of narrative interviews and in biographical research.

Once established in the 1970s, the Schützian paradigm influenced the mainstream of sociological theory which became sensitive not only to the social construction of life world but also to the phenomenological background of the Schützian theory. The 'life world,' in the sense of a basic social reality provided by humans in their 'natural' intercourse, became one of the central terms in social theory (J. Habermas 1981). The everyday construction of social reality was recognized as a crucial mechanism in which society emerges (P. Bourdieu 1972, A. Giddens 1976, Z. Baumann 1991). The phenomenological conception of meaning constitution reformulated as an *autopoiesis* (self-creation) of social and psychic systems influenced the development of the contemporary sociological system theory (N. Luhmann 1996).

Beyond of the scope of sociology, other humanities also gained innovative impulses from the Schützian approach. In philosophy, Schütz's theory led to a critical assessment of the Husserlian view of intersubjectivity and to the conceptions of a *worldly phenomenology* (L. Embree 1988) or of a *philosophy of modern anonymity* (M. Natanson 1986). In the literature, Schütz's constructionism inspired the *esthetics of reception* which pointed out the beholder's participation in co-creating the autonomous reality of literary works (H. R. Jauss 1982). Schütz's concept of the structure of life world also affected theorizing in social geography (B. Werlen 1993), educational theory (K. Meyer-Drawe 1984), and political science (E. Voegelin 1966).

See also: Constructivism/Constructionism: Methodology; Culture, Sociology of; Ethnomethodology: General; Everyday Life, Anthropology of; Husserl, Edmund (1859–1938); Interactionism: Symbolic; Interpretive Methods: Micromethods; Knowledge, Sociology of; Methodological Individualism: Philosophical Aspects; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Verstehen und Erklären, Philosophy of; Weber, Max (1864–1920)

Bibliography

Bauman Z 1991 *Modernity and Ambivalence*. Polity Press, Oxford, UK

Berger P L, Luckmann Th 1966 *The Social Construction of Reality*. Doubleday, Garden City, New York

Bourdieu P 1972 *Esquisse d'une Théorie de la Pratique, précédé de trois études d'ethnologie kabyle*. Droz S.A., Geneva, Switzerland

Cefai D 1998 *Phénoménologie et les Sciences Sociales: la Naissance d'une Philosophie Anthropologique*. Libraire Droz, Geneva, Switzerland

Cicourel A V 1964 *Method and Measurement in Sociology*. The Free Press of Glencoe, New York

Embree L 1988 *Wordly Phenomenology: The Continuing Influence of Alfred Schutz on North American Human Sciences*. University Press of America, Washington DC

Garfinkel G 1967 *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ

Giddens A 1976 *New Rules of Sociological Method*. Hutchinson & Co., London, UK

Grathoff R 1986 *Milieu und Lebenswelt*. Suhrkamp, Frankfurt/M, Germany

Habermas J 1981 *Theorie des kommunikativen Handelns*. Suhrkamp, Frankfurt/M, Germany

Jauss H P 1982 *Ästhetische Erfahrung und literarische Hermeneutik*. Suhrkamp, Frankfurt/M, Germany

Luhmann N 1996 *Die neuzeitlichen Wissenschaften und die Phänomenologie*. Picus, Vienna, Austria

Meyer-Drawe K 1984 *Leiblichkeit und Sozialität*. Fink, Munich, Germany

Natanson M 1986 *Anonymity: A Study in The Philosophy of Alfred Schutz*. Indiana University Press, Bloomington, IN

Psathas, G (ed.) 1973 *Phenomenological Sociology. Issues and Applications*. John Wiley & Sons, New York

Sacks H 1995 *Lectures on Conversation*. Blackwell, Oxford, UK

Schütz A 1932 *Der sinnhafte Aufbau der sozialen Welt*. Springer, Vienna

Schütz A 1962 *Collected Papers I*. Nijhoff, The Hague, The Netherlands

Schütz A 1964 *Collected Papers II*. Nijhoff, The Hague, The Netherlands

Schütz A 1966 *Collected Papers III*. Nijhoff, The Hague, The Netherlands

Schütz A 1970 *Reflections on the Problem of Relevance*. Yale University Press, New Haven, CT

Schütz A 1982 *Life Forms and Meaning Structure*. Routledge and Kegan Paul, London

Schütz A 1995 *Collected Papers IV*. Nijhoff, The Hague, The Netherlands

Schütz A, Luckmann T 1973 *Structures of Life World I*. Northwestern University Press, Evanston, IL

Schütz A, Luckmann T 1989 *Structures of Life World II*. Northwestern University Press, Evanston, IL
Schütz A, Parsons T 1978 *Theory of Social Action: The Correspondence of Alfred Schütz and Talcott Parsons*. Indiana University Press, Bloomington, IN
Srubar I 1988 *Kosmion die Genese der pragmatischen Lebensweltheorie von Alfred Schütz und ihr anthropologischer Hintergrund*. Suhrkamp, Frankfurt am Main, Germany
Wagner H R 1983 *Alfred Schütz. An Intellectual Biography*. University of Chicago Press, Chicago
Werlen B 1993 *Society, Action and Space. An Alternative Human Geography*. Routledge, London
Voegelin E 1966 *Anamnesis. Zur Theorie der Geschichte und Politik*. Piper, Munich

I. Srubar

Science and Development

During the past half-century, conventional understandings of science, development, and their relationship have changed radically. Formerly, science was thought to refer to a clear and specific variety of Western knowledge with uniformly positive effects on society. Formerly, development was viewed as a unidirectional process of social change along Western lines. Formerly, science was viewed as a powerful contributor to the developmental process. Each of these ideas has been subjected to insightful criticism. This article will examine science and development, concluding that the relationship between the two is problematic, partly because of the complexity of the concepts themselves. Three major theories of development are considered, together with the main types of research institutions in developing areas.

1. Science

Much of what is termed science in developing areas is far from what would be considered 'pure science' in the developed world. The 'root concept' of science involves *research*, the systematic attempt to acquire new knowledge. In its modern form, this involves experimentation or systematic observation by highly trained specialists in research careers, typically university professors with state-of-the-art laboratory equipment. These scientists seek to contribute to a cumulative body of factual and theoretical knowledge, testing hypotheses by means of experiments and reporting their results to colleagues through publication in peer-reviewed journals.

Yet when a new variety of seed is tested by a national research institute and distributed to farmers in Africa, this is described as the result of 'science.' When the curator of a botanical exhibit has a college degree, he or she may be described as 'the scientist.' When a newspaper column discusses malaria or AIDS, 'scientific treatments' are recommended. Seeds, educated people, and advice are not science in the abstract and lofty sense of the pursuit of knowledge for its own sake or systematically verified facts about the world. But they are science from the standpoint of those who matter—local people who spend scarce resources on their children's education, development experts who determine how and where to spend funds, politicians who decide whether to open a new university, corporate personnel who open a new factory in a developing region.

Perhaps the most important shift in recent thinking about science is a broadening of the scholarly view to include the ideas of science found among ordinary people. These are often *more* extended in developing areas, because of the association of science with 'modern' things and ideas. 'Science' in its extended sense includes technological artifacts, trained expertise, and knowledge of the way the world works. The importance of this point will be clear in the conclusion.

Given the fuzziness of the boundaries that separate science from other institutions, and the dependence of modern research on sophisticated technical equipment, the term 'technoscience' is often used to denote the entire complex of processes, products, and knowledge that flows from modern research activities. Even if we recognize that the term 'science' has extended meanings, it is useful to draw a distinction between (a) the institutions that produce knowledge and artifacts and (b) the knowledge that is produced. That is, on the one hand, there are organizations, people, and activities that are devoted to the acquisition of knowledge and things that can be produced with knowledge. These constitute the modern organization of *research*. On the other hand, there are *claims* involving knowledge and artifacts—often significantly transformed as they leave the confines of the research laboratory. What makes claims and practices 'scientific' is their association with scientific institutions.

Modern research capacity is concentrated in industrialized countries. Indeed, with respect to the global distribution of scientific and technical personnel, scientific organizations, publications, citations to scientific work, patents, equipment, and resources, scientific institutions display extremely high degrees of inequality. The most common indicator of scientific output is publications. In 1995, Western Europe, North America, Japan, and the newly industrialized countries accounted for about 85 percent of the world total. Leaving aside countries and allies of the former Soviet Union, developing areas contributed less than 9 percent of the world total. Much the same applies to technological output measured in patents and expenditures on research and development (UNESCO 1998).

Yet if we shift our focus from the question of inequality to the question of diffusion, an entirely different picture arises. To what extent have the *idea*

and practice of research spread throughout the world? The main issues here involve who conducts research, on what subjects, and what happens to the results. Each of these topics is the subject of analysis and controversy.

Scientific research in developing countries began during the pre-independence era with the establishment of universities and research institutes. Research was conducted on crops and commodities for export as well as conditions (e.g., disease) that affected the profits sought by external agents from their control over the land, labor, and property of colonized peoples. Methodologies and organizational models for research were brought by European colonists to Asia, Africa, and Latin America. During the era of independence and throughout the 1970s, number of types of entities engaged in the generation of knowledge multiplied. The main scientific organizations now fall into five main types, or sectors: academic departments, state research institutes, international agencies, private firms and, to a lesser degree, nongovernmental organizations.

2. *Development*

The concept of development involves several dimensions of transformation, including the creation of wealth (that is, rapid and sustained economic growth) and its distribution in a fashion that benefits a broad spectrum of people rather than a small elite (that is, a reduction in social inequality). Cultural transformation (recognition of and attendant value placed on local traditions and heritage) has also been viewed as an important aspect of the process since the early 1980s. There is general agreement that development in the second half of the twentieth century is not a mere recapitulation of the process of industrialization that characterized Europe and North America in the eighteenth and nineteenth centuries.

Three theoretical perspectives, with many variations, have dominated development studies: modernization, dependency, and institutional. One way of distinguishing these theories is by their position on the ways in which relationships external to a country affect the process of change. Since scientific institutions and knowledge claims are of external origin, each of these perspectives views science and technology as important in the development process, with very different assessments of the costs and benefits.

2.1 *Modernization*

The oldest approach, sometimes called modernization theory, focused on the shift from a traditional, rural, agricultural society to a modern, urban, industrial form. Transformations internal to a country (such as formal education, a market economy, and democratic political structures) are emphasized, while external relationships are de-emphasized. However, science was the exception to this, available to benefit developing nations through *technology transfer* from Western sources. This idea relied on two assumptions. One was the 'hardness' of technological artifacts—their alleged independence from people and culture, their seeming ability to produce certain effects 'no matter what.' The second was the 'linear model' of technology development in which the (a) discoveries of basic science lead to the (b) practical knowledge of applied science and finally to (c) technological applications such as new products. In retrospect, both of these assumptions were simplistic in any context, but in the developing world they were especially problematic.

The assumption of 'hardness' has been replaced by the generalization that the uses, effects, and even the meanings of technological artifacts are affected by the context of use. First, effective technologies, from automobiles to indoor plumbing, do not typically stand alone, but are embedded in *systems* that provide infrastructure (roads, sewage treatment) which is often lacking. Second, the provision of artifacts such as buildings and computers is much easier than their maintenance, which requires both resources and knowledge. Third, introduction of new technology involves a multiplicity of consequences—positive and negative, short term and long term, economic and ecological. Many of these consequences are unpredictable, even in those rare cases where such foresight is attempted.

The case of the Green Revolution is illustrative. In the 1960s, widespread food shortages, population growth, and predicted famine in India prompted major international foundations to invest research and technology transfer efforts towards the goals of increasing agricultural productivity and the modernization of technology. What resulted were new kinds of maize, wheat, and rice. These modern varieties promised higher yields and rapid maturity, but not without other inputs and conditions. They were, rather, part of a 'package' that required fertilizers as well as crop protection inputs such as pesticides, herbicides, and fungicides—sometimes even irrigation and mechanization. Moreover, seed for these varieties had to be purchased anew each year.

The consequences of the Green Revolution are still debated, and there is little doubt that many of them were positive. Famine in India was averted through increased yields, but the benefits of the technology required capital investments that were only possible for wealthier farmers. Not only did the adoption of new technology increase dependence on the suppliers of inputs, but it was claimed to increase inequality by hurting the small farmer—one intended beneficiary of the technology. The actual complexity of the outcomes is revealed by one of the most sophisticated assessments—modern seed varieties do reach small farmers, increase employment, and decrease food prices, but the benefits are less than expected because the poor are

increasingly landless workers or near landless farm laborers (Lipton and Longhurst 1989).

What is important for the question of the relationship between science and development is that the products and practices of the Green Revolution were research-based technology. This technology was often developed in *international* research institutes funded by multilateral agencies such as the World Bank and bilateral donors such as the US Agency for International Development. Since the combined resources of these donors dwarf those of many poor countries, their developmental and research priorities constitute a broad global influence on the nature of science for development. The largest and most visible of these organizations form a global research network, the Consultative Group for International Agricultural Research (CGIAR) which grew from 4 to 13 centers during the 1970s as support by donors quadrupled. The influence of this network of donors and international agencies was clearly evident in the early 1990s when environmental concerns led to an emphasis on 'sustainability' issues. This led to a change in CGIAR priorities, as the older emphasis on agricultural productivity shifted to the relatively more complex issue of natural resource management.

2.2 Dependency

Modernization theory emphasized internal factors while making an exception of science. Dependency theory and its close relative, world system theory, emphasized the role of *external relationships* in the developmental process. Relationships with developed countries and particularly with multinational corporations were viewed as barriers. Economic growth was controlled by forces outside the national economy. Dependency theory focused on individual nations, their role as suppliers of raw materials, cheap labor, and markets for expensive manufactured goods from industrialized countries. The unequal exchange relationship between developed and developing countries was viewed as contributing to poor economic growth. World system theory took a larger perspective, examining the wider network of relationships between the industrialized 'core' countries, impoverished 'peripheral' countries, and a group of 'semiperipheral' countries in order to show how some are disadvantaged by their position in the global system. Because of their overspecialization in a small number of commodities for export, the unchecked economic influence of external organizations, and political power wielded by local agents of capital, countries on the periphery of the global capitalist system continue to be characterized by high levels of economic inequality, low levels of democracy, and stunted economic growth.

What is important about the dependency account is that science is not viewed in benign terms, but rather as one of a group of institutional processes that contribute to underdevelopment. As indicated above, research is highly concentrated in industrialized countries. Dependency theory adds to this the notion that most research is also conducted for their benefit, with problems and technological applications selected to advance the interests of the core. The literature on technology transfer is also viewed in a different light. The development of new technology for profit is associated with the introduction and diffusion of manufactured products that are often unsuited to local needs and conditions, serving to draw scarce resources away from more important developmental projects. The condition of dependency renders technological choice moot.

This concern with choice, associated with the argument that technology from abroad is often *imposed* on developing countries rather than *selected* by them, has resurfaced in many forms. In the 1970s it was behind the movement known as 'intermediate' technology, based on the work of E. F. Schumacher, which promoted the use of small-scale, labor-intensive technologies that were produced locally rather than of complex, imported, manufactured goods. These 'appropriate' technologies might be imported from abroad, but would be older, simpler, less mechanized, and designed with local needs in mind. What these viewpoints had in common was a critical approach to the adoption of technology from abroad.

By the late 1980s and 1990s even more radical positions began to surface, viewing Western science as a mechanism of domination. These arguments were more closely related to ecological and feminist thought than to the Marxist orientation of dependency theory. Writers such as Vandana Shiva proposed that Western science was reductionist and patriarchal in orientation, leading to 'epistemic violence' through the separation of subject and object in the process of observation and experimentation (1991). 'Indigenous knowledge' and 'non-Western science' were proposed as holistic and sustainable alternatives to scientific institutions and knowledge claims. Such views had an organizational base in nongovernmental organizations (NGOs), which received an increasing share of development aid during this period, owing to donor distrust of repressive and authoritarian governments in developing areas. NGOs have been active supporters of local communities in health, community development, and women's employment, even engaging in research in alternative agriculture (Farrington and Bebbington 1993).

2.3 Institutional Theory

Institutional theory seeks to explain why nations are committed to scientific institutions as well as what forms these take. The central theme is that organizational structures developed in industrialized countries are viewed by policy makers, donors, and other states as signals of progress towards modern institutional

development and hence worthy of financial support. Regardless of the positive or negative consequences of their activities, the introduction and maintenance of certain forms in tertiary education and government serves to communicate this commitment. Institutional theory provides an account of the growth and structure of the academic and state research sectors, as successful organizations in industrialized nations operate as models far from their original contexts.

Academic departments consist of researchers grouped by subject, each of whom is relatively free to select research projects. They bear the closest resemblance to the root concept of science introduced at the beginning of this article. But research requires time and resources. In areas such as sub-Saharan Africa, laboratories and fieldwork are poorly funded, if at all, since many institutions can barely afford to pay salaries. Professors teach, consult, and often maintain other jobs. Research is conducted as a secondary activity and professional contacts with other scientists in Europe and the US are few.

Equally important to the scientific establishment are state research institutes. These organizations are agencies of the state, they are charged with performing research with relevance to development, with health and agriculture the two most important content areas. They are linked to ministries, councils, and international agencies as well as systems (such as Extension Services in agriculture) that deliver technology to users—again based on a model from the developed world.

3. *Relationships Between Science and Development*

The popularity of dependency arguments and the resurgence of interest in indigenous forms of knowledge implies continued competition for older views of the uniformly positive effects of science. Institutional theory provides an alternative account of the spread of science and its organizational forms. But two features of current scholarship may prove more significant in the long run.

First, extreme diversity exists among developing areas in terms of their economic, social, and cultural patterns. It makes decreasing sense to speak of 'development' as an area of study. Latin American nations, for example, are generally far better positioned than the nations of sub-Saharan Africa. There is even wide variation within countries, as the case of India makes clear. While much of India qualifies as a developing area, it is among the world's top producers of scientific work, has a technically skilled, English-speaking labor force second only to the US, and is a leading exporter of computer software for corporations.

Second, 'science' is viewed as having many dimensions, many effects, and fuzzy institutional boundaries, but it is always a feature of the modern, industrial, interconnected world. Science cannot be the cause of modernization because, in its diverse institutional articulations and its evolving fit with society, science exemplifies the meaning of modernization itself.

See also: Biomedical Sciences and Technology: History and Sociology; Development: Social-anthropological Aspects; Development Theory in Geography; Innovation and Technological Change, Economics of; Research and Development in Organizations; Science and Technology, Anthropology of; Science and Technology: Internationalization; Science and Technology, Social Study of: Computers and Information Technology; Technology, Anthropology of

Bibliography

Baber Z 1996 *The Science of Empire: Scientific Knowledge, Civilization, and Colonial Rule in India*. State University of New York Press, Albany, NY

Farrington J, Bebbington A 1993 *Reluctant Partners: Non-governmental Organisations, the State, and Sustainable Agricultural Development*. Routledge, London

Gaillard J 1991 *Scientists in the Third World*. University of Kentucky Press, Lexington, KY

Kloppenburg J 1988 *First the Seed: The Political Economy of Plant Biotechnology*. Cambridge University Press, Cambridge, UK

Lipton M, Longhurst R 1989 *New Seeds and Poor People*. Johns Hopkins University Press, Baltimore, MD

Pearse A 1980 *Seeds of Plenty, Seeds of Want: Social and Economic Implications of the Green Revolution*. Clarendon Press, Oxford, UK

Schumacher E F 1973 *Small is Beautiful: Economics as if People Mattered*. Harper and Row, New York

Shahidullah S 1991 *Capacity Building in Science and Technology in the Third World*. Westview Press, Boulder, CO

Shiva V 1991 *The Violence of the Green Revolution*. Zed Books, London

Shrum W, Shenhav Y 1995 Science and technology in less developed countries. In: Jasanoff S, Markle G, Peterson J, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA

Stewart F 1977 *Technology and Underdevelopment*. Westview Press, Boulder, CO

UNESCO 1998 *World Science Report 1998*. Elsevier, Paris

Yearley S 1988 *Science, Technology, and Social Change*. Unwin Hyman, London

W. Shrum

Science and Industry

The premodern industrial craft economy provided the initial intersection of industry with science through scientific instrument making. The development of scientific inquiry through craft-based production, and its effects, can be seen in Galileo and the telescope, chang-

ing the world-picture and the place of humans within it. From as early as Leeuwenhoek's microscope to as late as the making of the Hubble telescope in the 1980s, lens and mirror construction was an uncertain art, not fully amenable to scientific analysis and control.

The invention of moveable type transformed the transfer of knowledge, through storage and retrieval devices, and facilitated the development of modern scientific as well as popular literature. However, it was the development of the steam engine by a scientifically informed inventor, James Watt, which led to a systematization of practice that could be analyzed scientifically, becoming the basis for the next advance. This process gave rise to a new term, technology, to denote an innovation process as well as its results.

Technology is the feedback link between science and industry. Invented in the eighteenth century, this new concept represented, on the one hand, the systematization of craft, through the creation of engineering disciplines, originally to quantify and sytematize the construction of military fortifications (Calvert 1967). On the other hand, technology was also derived from the extension of science. This occurred through the creation of applied sciences such as solid state physics and the invention of semiconductor devices such as the transistor. This article surveys the relationship between science and industry since the 1700s.

1. *Early Science and Industrial Development*

Science originated in the seventeenth century as organized investigation of the natural world according to relatively secure methodological principles. In this era, practical and theoretical concerns each provided between 40–60 percent of the impetus to research, with some overlap (Merton 1938). Well before science was institutionalized in universities and research institutes, individual scientists, loosely connected through scientific societies, provided an occasional basis for industrial development. For example, in the absence of reliable navigational techniques, commerce was impeded by the need for ships to stay close to shorelines. In response to a prize offered by the British government, astronomers used their observational techniques and knowledge base to develop star charts useful to navigators. Their involvement in solving a commercial problem was secured without incurring explicit costs. These were considered to be assumable by the astronomers themselves, with navigational research considered as an offshoot of their government or academic responsibilities that could be carried out at marginal cost.

Clockmakers approached the problem from the opposite stance by adapting a mechanical device to the solution of the navigational problem. Of lower status and with lesser financial resources and institutional backing than the astronomers, a clockmaker who arrived at a mechanical solution to the problem had great difficulty in being taken seriously by the judges of the competition. Moreover, as an independent craftsman he was dependent upon receiving intermediate financial dispensations from government to improve his device. Nevertheless, science and craft intersected to overcome blockages to the flow of trade by providing reliable navigational methods (Sobel 1996).

Science became more directly involved in industrial production in seventeenth century Germany when professors of pharmacy invented medical preparations in the course of their experimentation. With support from some German princely states, and in collaboration with entrepreneurs, firms were formed to commercialize these pharmaceutical discoveries. Thus, the academic spin-off process, with governmental and industrial links, was adumbrated at this time (Gustin 1975).

2. *Incremental Innovation: Learning by Doing*

In an era when most industry was craft based, incremental improvements arose primarily from workers' experience with the process of production. For example, in the course of firing bricks a worker might notice that a brick had attained exceptional strength under high heat and then attempt, through trial and error, to duplicate what the observation of a chance event had brought to light. Eventually, the conditions that produced the original improved brick might be approximately reproduced and a sufficiently high rate of production achieved, with the 'experimenter' knowing more or less how but not really why the useful result had been achieved (Landes 1969).

Scientific advance was also built upon what is now called 'learning by doing' (Lundvall and Borras 1997). By interviewing various practitioners, researchers began to collate and systematize their local knowledge into broader syntheses. Thus, advances in understanding of stratigraphy derived from miners' practical experience in eighteenth-century Italy (Vaccari 2000).

Much, if not most, innovation still takes place through craft-based experience. Indeed, scientific principles and methods such as those developed in operations research have recently been applied to systematize incremental innovation. Incremental innovation itself has been scientized through the application of statistical techniques, pioneered in Japan during the 1930s by the disciples of Edward Deming, the US researcher, who only later gained consulting opportunities and renown in his own country.

3. *The Industrialization of Science*

The connection between academe, science, and industry strengthened with the invention of the chemistry laboratory as a joint teaching and research format

by Justus Liebig at the University of Giessen in the mid-nineteenthth century (Brock 1997). Having achieved reliable analytical methods, Liebig could assign an unsolved problem to a student and expect a solution to arise in due course, with a minimum of supervision from the master or his assistants. The teaching laboratory model then spread to other experimental fields. As an organizational innovation combined with replicable methods of investigation, the laboratory allowed training and original investigation to be coordinated and expanded, creating a larger market for research equipment.

Alexander Von Humboldt theorized the unity of teaching and research as an academic model as well as a practical tenet (Oleson and Voss 1979). The incorporation of science into the university along with methods to revive classical knowledge led to the development of research as an academic practice. The research university model was transferred from Germany to the US in the mid-nineteenth century and eventually became the base for technology transfer and firm formation (Jencks and Riesman 1968).

Liebig, himself, attempted to take the development of technology from academic research a step further by starting businesses based upon his scientific discoveries. His mix of successes and failures foreshadowed the contemporary science-based firm. Nevertheless, this was an unusual combination of roles of researcher and entrepreneur in one person at the time (Jones 1993). More typically, the professor's students made technology arising from scientific research into companies. Thus, the emblem of the Zeiss optical firm incorporates a portrait of the original founders, including professor and student.

4. *The Foundation of Firms Based on Scientific Research*

With the invention of the laboratory, instrument making was internalized within science. As science and the need for research equipment grew, instrument production began to be externalized and made into an industry. Scientific instrument making is an early source of firm formation linked to academic research. For example, scientist-initiated instrument firms grew up around MIT and Harvard in the late nineteenth century, along with consulting firms such as A. D. Little providing a formal overlay of relationships between university and industry, beyond personal ties between teacher and former student.

Until quite recently, scientific instrument making was a specialized niche industry, having more to do with academia than industry. A shift toward dual use industrial and scientific technologies began to occur with the development of electronic instrumentation. Oscilloscopes served as research tools, for example, to record nerve impulses in physiology, but were also utilized to provide quality assurance in industrial production. The formation of the Hewlett Packard and Varian Corporations, based upon innovations in electronics in the physics department at Stanford University in the 1930s, exemplify this transitional phase in the relationship between scientific instrumentation and industry (Lenoir 1997).

Industrialized science is exemplified today by mass-produced scientific equipment such as the sequencing machines crucial to the human genome project. What is new is the breakdown of the distinction between scientific instrumentation and the means of industrial production. This has occurred through the emergence of technologies such as computers that have broad application to science, business, art, and other techniques (Ellul 1964). Indeed, the computer is a machine with such protean implications that it became the basis of a science itself and contributed to the creation of a new class of sciences of the artificial.

5. *The Rise of Science-based Industry*

Although it had its precursors in China, with gunpowder, paper making, and the organizational technology of bureaucracy, the science–industry interface has been rationalized in the West and North and transferred to the East and South in the modern era. Karl Marx, the initiator of the theory of science-based industry, even though he lost confidence and retreated to a labor theory of value, was far seeing. In the late nineteenth century, he had one example on which to base his thesis, the British chemist Perkin's research on aniline-based dyes that was translated into an industry in Germany.

According to a close observer, 'Science and technology may have converged circa 1900' (Wise 1983). The rise of the chemical and electrical industries in the late nineteenth century fulfilled some of the promise of science for industrial development. Nevertheless, even in these industries the translation of science into useful products often had an intervening phase based on 'cut and try methods that only later became rationalized as in the unit method of scaling up chemical production (Servos 1980).

Thomas Alva Edison, the inventor of the electric light, was also the inventor of the systematic production of inventions. His 'idea factory,' staffed by formally trained scientists and craftspersons, provided a support structure for the application of Edison's basket of techniques to a series of technical problems whose solution was the basis for the creation of new industries and firms (Israel 1998).

One of these companies, the General Electric Corporation (GE), took the connection between science and industry one step further in the US by hiring an academic scientist to organize a research laboratory for the firm. GE's hiring of Willis Whitney brought the consulting function of academics to industrial problems within the firm and also created a new source of

product development, the corporate R&D laboratory (Wise 1983).

A reverse relationship between science and engineering exists when corporations have 'first mover' rather than 'follower' business strategies. Siemens, for example, developed new business on the basis of advanced research in semiconductors during the early postwar period while its competitor, AEG, waited to see if a sufficient market developed before entering a field. Indeed, Siemen's confidence in science was so great that it did not employ sufficient engineers to refine and lower the cost of its production (Serchinger 2000).

The function of the corporate R&D lab was at least threefold: (a) maintain contact with the academic world and other external sources of information useful to the firm; (b) assist in solving problems in production processes and product development originating within the firm; and (c) originate new products within and even beyond the existing areas of activity of the firm.

A few corporate labs, typically in 'public' industries such as telecommunications took on a fourth function as quasi-universities, contributing to the advance of science itself (Nelson 1962). Smaller firms typically focused on the first two functions, close to production, while larger firms more often spanned the entire range of activities (Reich 1987).

6. *University–Industry–Government Relations*

Models for close public–private cooperation that were invented before World War II have recently been revived and expanded (Owens 1990). During the early postwar period, an institutional division of labor arose in R&D, with industry supporting applied research; government funding basic research in universities and research institutes; and universities providing knowledge and trained persons to industry (Reingold 1987). This 'virtuous circle' still accurately describes much of the university–industry–government relationship (Kevles 1977)

Socialist countries developed a variant of this format, attempting to create closer connections to research and production through central planning mechanisms. However, by locating the R&D function in so-called branch institutes, socialist practice actually created a distance between research and production that was even greater than in capitalist firms that located their R&D units geographically apart from production while retaining organizational ties.

Recently, socialist and, to some extent, large corporate formats for R&D have been decomposed and recombined in new formats. In the US, where these developments have taken their most advanced form to date, several innovations can be identified including:

(a) Universities extending their functions from research and training into technology development and firm formation through establishment of new organizational mechanisms such as centers, technology transfer offices, and incubator facilities.

(b) The rise of start-up, high-tech firms, whether from universities, large corporate laboratories, or previous failed spin-offs, providing a new dynamic element in the science–industry relationship, both as specialized research organizations and as production units rooted in scientific advance.

(c) A new role for government in encouraging collaboration among various academic and industrial units, large and small, in industrial innovation going beyond the provision of traditional R&D funding, including laws changing 'the rules of the game' and programs to encourage collaboration among firms and between universities and firms.

An industrial penumbra appears around scientific institutions such as universities and research laboratories, creating feedback loops, as well as conflicts of interest and commitment between people involved in the two spheres. Over time, as conflicts are resolved, new hybrid forms of science-based industry and research, as well as new roles such as the industrial and entrepreneurial scientist, are institutionalized (Etzkowitz 2001).

7. *Conclusion*

Economic development is increasingly based on science and technology at the local, regional, national, and multinational levels. The science–industry connection, formerly an ancillary and subsidiary aspect of both science and industry, now moves to the center of the stage of economic development strategy as the university and other knowledge-creating organizations become a source of new industry. Political entities at all of these levels develop policies and programs to enhance the science–industry interface, and especially to encourage high-tech firm formation (Etzkowitz et al. 2000).

The 'endless frontier' model of 'knowledge flows,' transferred to industry through publications and graduates was insufficient to induce industrial innovation in most cases (Brown 1999). Closer connections were required. A series of organizational innovations to enhance technology transfer took place along a continuum: providing training in entrepreneurship, translation of research findings into intellectual property rights, and encouraging the early stages of firm formation.

The potential for participation in knowledge-based economic development revivifies the discredited nineteenth century notion of progress. During the colonial era, technology transfer typically took place in a format that helped maintain political control, and the higher forms of tacit knowledge were kept secret by employing expatriate engineers. Nevertheless, as in India, a steel industry, created by local entrepreneurs, provided an economic base for a political indepen-

dence movement. Moreover, a seemingly over-expanded higher educational system, originally put in place to train lower level bureaucrats and technicians, has become an engine of growth for India's software industry.

The synthesis of science and industry is universalized, as both industrial and natural heritages, such as machine-tool industry in German and biodiversity in Brazil, are integrated into new global configurations, often through local firms in strategic alliance with multinationals. Given the decreasing scale of scientific equipment and the increasing availability of higher education, countries and regions in virtually every part of the world can take advantage of opportunities to develop niches. A trend toward worldwide technological innovation has transformed the previous designations of third and second worlds into the 'emerging world.'

New candidates for science-based economic growth bridge the expected gaps between 'long waves' of economic growth. Heretofore, the technology-based long waves identified by Freeman and his co-workers were made possible by a few key technologies; that is, information technology and biotechnology (Freeman and Soete 1997). The potential for growth from technologies such as solar photovoltaics, multimedia, and the Internet, and new materials arising from nanotechnology are so numerous that there is little or no technical reason, only political, for business cycles with declines as well as rises.

See also: History of Science; History of Technology; Industrial Society/Post-industrial Society: History of the Concept; Industrial Sociology; Innovation: Organizational; National Innovation Systems; Reproductive Rights in Developing Nations; Science, Technology, and the Military; Scientific Revolution: History and Sociology

Bibliography

Brock W 1997 *Justus von Liebig: The Chemical Gatekeeper*. Cambridge University Press, Cambridge, UK

Brown C G 1999 Patent policy to fine tune commercialization of government sponsored university research. *Science and Public Policy* December: 403–14

Calvert M 1967 *The Mechanical Engineer in America, 1830–1910: Professional Cultures in Conflict*. Johns Hopkins University Press, Baltimore

Carnavale A 1991 *America and the New Economy*. Jossey Bass, San Francisco

Clow A 1960 The industrial background to John Dalton. In: Cardwell D S L (ed.) *John Dalton and the Progress of Science*. Manchester University Press, Manchester, UK

Davis L 1984 *The Corporate Alchemist: Profit Makers and Problem Makers in the Chemical Industry*. William Morrow, New York

Dyson F 1999 *The Sun, the Genome and the Internet*. Oxford University Press, New York

Ellul J 1964 *The Technological Society*. Knopf, New York

Etzkowitz H 2001 *The Second Academic Revolution: MIT and the Rise of Entrepreneurial Science*. Gordon and Breach, London

Etzkowitz H, Gulbrandsen M, Levitt J 2000 *Public Venture Capital*. Harcourt, New York

Freeman C, Soete L 1997 *The Economics of Industrial Innovation*. MIT Press, Cambridge, MA

Gustin B 1975 The emergence of the German chemical profession. Ph.D. dissertation, University of Chicago

Israel P 1998 *Edison: A Life of Invention*. Wiley, New York

Jencks C, Riesman D 1968 *The Academic Revolution*. Doubleday, New York

Johnson J 1990 *The Kaiser's Chemists: Science and Modernization in Imperial Germany*. University of North Carolina Press, Chapel Hill, NC

Jones P 1993 Justus Von Liebig, Eben Horsford and the development of the baking powder industry. *Ambix* **40**: 65–74

Kevles D 1977 The NSF and the debate over postwar research policy, 1942–45. *ISIS* **68**

Landes D 1969 *The Unbound Prometheus*. Cambridge University Press, Cambridge, UK

Lenoir T 1997 *Instituting Science: The Cultural Production of Scientific Disciplines*. Stanford University Press, Stanford, CA

Lundvall B-A, Borras S 1997 *The Globalising Learning Economy*. The European Commission, Brussels

Merton R K 1938 *Science, Technology and Society in Seventeenth Century England*. St. Catherines Press, Bruges

Nelson R 1962 The link between science and invention: The case of the transistor. In: *The Rate and Direction of Inventive Activity*. Princeton University Press, Princeton, NJ, pp. 549–83

Oleson A, Voss J 1979 *The Organization of Knowledge in Modern America*. Johns Hopkins University Press, Baltimore

Owens L 1990 MIT and the federal 'Angel' academic R&D and federal–private cooperation before World War II. *ISIS* **81**: 181–213

Reich L 1987 Edison, Coolidge and Langmuir: Evolving approaches to American industrial research. *Journal of Economic History* **47**: 341–51

Reingold N 1987 Vannever Bush's new deal for research; or the triumph of the old order. *Historical Studies in the Physical and Biological Sciences* **17**: 299–344

Serchinger R 2000 Wirstschaftswunder in Prezfeld, Upper Franconia; Interactions between science, technology and corporate strategies in Siemens semiconductor rectifier research & development, 1945–1956. *History and Technology* **335**: 82

Servos J 1980 The industrial relations of science: Chemical engineering at MIT, 1900–1939. *ISIS* **71**: 531–49

Sobel D 1996 *Longitude*. Penguin, Baltimore

Vaccari E 2000 Mining and knowledge of the Earth in eighteenth century Italy. *Annals of Science* **57**(2): 163–80

Wise G 1983 Ionists in industry: Physical chemistry at General Electric, 1900–1915. *ISIS* **74**: 7–21

H. Etzkowitz

Science and Law

The relationship between law and science has occupied scientists, philosophers, policymakers, and social analysts since the early modern period. Nature, according

to science's early modern practitioners, was governed by law-like regularities—God's laws—and the work of science lay in discerning and revealing these laws of nature. In time, however, human law came to be seen as a fallible institution in need of rationalization, to be made more like a science in order to avoid arbitrariness. For early twentieth-century legal reformers, science provided a model of regularity for law to aspire to. At the same time, law and science were viewed by their practitioners as independent institutions, each with its own organization, practices, objectives, and ethos. Similarities and differences between the two fields were noted by many observers, with frequent commentary on conflicts between the law's desire for justice and science's commitment to the truth.

By the end of the twentieth century, a new preoccupation with the law's instrumental uses of science emerged, which led to divergent schools of thought about the epistemological status of scientific evidence and the ways in which legal and policy institutions should interact with scientific experts. The growth of the state's regulatory powers increased governments' dependence on scientific methods and information, and disputes developed about the extent to which science could reliably answer the questions put to it by legislators, regulators, and litigants. An influx of technically complex disputes caused judicial systems to reassess their handling of scientific and technical evidence. Analysts of these processes grappled with questions about the law's capacity to render justice under conditions of scientific and social uncertainty, as well as the continued relevance of the lay jury and the role of legal proceedings in the production of scientific knowledge.

This article reviews the resulting scholarship on the nature and consequences of the law–science relationship under four major headings. One concerns the role of these two institutions in shaping the authority structures of modernity, particularly in legitimating the exercise of power in democratic societies. A second relates to the law's impact on the objectives, status, and professional practices of scientific disciplines relevant to the resolution of social problems. A third focuses on responses by courts and their critics to the challenges posed by experts in the legal system, including shifts in the rules governing expert testimony. Fourth and finally, an emerging line of research looks at the law as a site for generating scientific and technical knowledge and inquires into the implications of collaboration between law and science for individual and social justice.

1. *Law and Science in the Transition to Modernity*

Since the origins of modern scientific thought, the term 'law' has been used to describe both the regularities discernible in nature and the rules laid down by religious or secular authorities to guide human conduct. Both kinds of laws, it was once popularly assumed, embodied principles that should work always and everywhere in the same fashion, even though Robert Boyle, an early modern founder of experimental science, cautioned against too literal an assimilation of the behavior of matter under natural law to the behavior of human agents under civil law (Shapin 1994, pp. 330–1). As certain facts about human nature and behavior (such as the approximate equality of humans, their vulnerability, and the limits of their resources and altruism) came to be accepted as given, modern legal theorists adopted more synthetic views regarding the relationship between science and law. According to the renowned British scholar of jurisprudence H. L. A. Hart, for example, these natural facts underwrite a 'natural law' which human societies require in order to maintain their salient moral characteristics (Hart 1961). The American legal theorist Lon Fuller took a more sociological tack in his account of the 'morality of law.' Echoing Robert Merton's well-known essay on the *norms of science*, Fuller argued that law and science resemble each other because each institution possesses an internal morality resulting from its distinctive arrangements, practices, and fiduciary demands (Fuller 1964). Later legal scholarship compared the 'cultures' of law and science and likewise called attention to their normative and procedural particularities (Goldberg 1994, Schuck 1993), but these institutional characteristics were now seen as a source of conflict, or culture clash.

As elite institutions, science and law historically have drawn upon each other's practices to build their authority. Thus, early modern experimentalists in the natural sciences enlisted the support of witnesses, both real and 'virtual' (Shapin and Schaffer 1985), as if they were proving a case in a court of law. The Enlightenment of the eighteenth century and the rise of liberal democracies brought science, now prized for its practical utility (Price 1985), into a more openly instrumental partnership with the law. In the USA, Thomas Jefferson drafted the first Patent Act of the fledgling American republic in 1793, thereby implementing the constitutional grant of power to Congress 'to promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries.' Where Jefferson saw the state as sponsoring science, some modern scholars have seen science's open, self-regulating, nonhierarchical modes of governance as offering a model or prototype for democracy (Polanyi 1962). Others, however, regard the relationship of scientific and political authority as one of mutual dependency. Democratically constituted governments, they argue, need science and technology in order to answer constant public criticism and make positive demonstrations of their worth to citizens. States, therefore, have harnessed science and technology to instrumental projects in both war and peace.

In these legitimation exercises, the public attests to the state's technical performance, much as scientists bore witness to each other's observations in early modern experimental laboratories (Ezrahi 1990). The law's procedural forms in this way have underwritten the credibility of both modern science and the state.

The law's own dependence on science has grown by distinctive pathways in the USA, as could be expected in a society where public decisions are exceptionally transparent and significant policy change commonly proceeds through litigation. By the beginning of the twentieth century, ideological and political shifts fostered the idea that social policies enacted by the state could and should be responsive to findings in the sciences. An icon of the Progressive era, US Supreme Court Justice Louis D. Brandeis, spurred this development through his historic 'Brandeis brief' in *Muller v. Oregon* (1908). Written to defend the state's 10-hour working-day restriction on women's employment, the 100-page brief is widely regarded as the first attempt to bring social science insights into the courts. Brandeis, then a public interest lawyer in Boston, argued successfully that women were sufficiently different from men to justify special legislative protection, thereby distinguishing the Oregon case from the Court's earlier, infamous decision in *Lochner v. New York* (1905), which struck down a similar state law curtailing working hours for (male) bakers.

Social science evidence received a further boost when, in overturning the 'separate but equal' doctrine that was the basis for racial segregation in public schools, the Supreme Court in *Brown v. Board of Education* (1954) relied on psychological studies of segregation's 'detrimental effect upon the colored children.' In later years, students with disabilities would benefit from similar sorts of evidence in claiming rights to special education (Minow 1990). Employment discrimination, racial segregation, antitrust, capital punishment, and environmental liability were among the many litigated subjects that brought social science testimony into the US courts in the latter decades of the twentieth century. These cases kept alive the view that objective scientific information could be invoked in legal proceedings to combat various forms of oppression and social injustice.

If science was allied with the search for justice in US courts, its role in policymaking was primarily as the voice of reason. Government's increasing involvement in welfare policy in the New Deal, intensified by the rise of social regulation in the 1970s, created new demands for rational decisions concerning a host of costly and potentially explosive problems, such as race and the environment. As can be observed in the context of environmental regulation (Jasanoff 1990), federal agencies proved increasingly less able to take refuge in generalized claims of expertise. Much work went into separating science from politics and creating new institutions to deliver putatively independent advice to government, although episodes like the 1995 dissolution of the Office of Technology Assessment, an advisory body to the US Congress, called attention to the fragility of such efforts. New types of scientific expertise also evolved to support governmental decisions with substantial distributive impacts, as discussed in the next section.

The relationship between law and science in modern European states was in some ways more covert than in the USA, but nonetheless pervasive. The crises of industrialization in the nineteenth century and the ensuing rise of the welfare state ushered in a demand for credible public action that states could not satisfy without rapid increases in their capacity to order society through newly accredited social science disciplines (Wagner et al. 1991). In turn, the classifications offered by the human sciences in particular supplied the basis for new forms of self-understanding, normalization, and social control. Whether in medicine, psychology, criminology or studies of intelligence, science offered the means for dividing people into categories of normal and pathological, mainstream and deviant. The scientific study of human characteristics thus enabled members of society to see and discipline each other without the need for constant, formal regulatory oversight by the state (Foucault 1990). Science, on this account, not only provided supports for legal authority but actually replaced some of the ordering functions of positive law.

European institutions at first resisted the pressures that fed attempts to separate science from politics in US legal contexts. Expert advisory bodies were routinely constituted to include political as well as technical expertise, although judgments about the need for lay or nonprofessional expertise varied from country to country. European courts, for their part, dealt with more focused and on the whole less politically salient disputes than in the USA, so that there were fewer reasons to contest expert judgments. By the 1990s, however, a number of developments tried public confidence in European governments' uses of science, chief among them the growing political influence of the European Union's technical bureaucracies and the fallout from the UK government's botched response to the bovine spongiform encephalopathy (BSE) crisis, which cast doubt on the integrity and impartiality of experts. European lawmakers, regulators, and courts began reconsidering the adequacy of their institutional arrangements for science-based decision-making, as the issue of science and governance rose to new prominence on political agendas.

2. *Disciplines, Professions, and the Law*

The interactivity of law and science has spurred new disciplinary formations and helped to change professional norms and practices, especially in the human sciences, which serve both as allies in regulation and

law enforcement and as uneasy subjects of the law's skeptical gaze. This dual relationship with the law has prompted novel lines of scientific research and theorizing, accompanied by periodic legal encroachment upon science's professional autonomy. Developments in the fields of risk analysis and psychology clearly illustrate these patterns.

The scientific study of risk is an outgrowth of state efforts to manage the hazards of industrial production and an extension of modern governments' growing reliance on formal decision-making tools to rationalize difficult political decisions. Uncertainty about technology's risks and benefits rose through the twentieth century along with the size and complexity of systems of production. From village-based economies, in which injured parties presumed they knew the agents as well as the causes of harms inflicted on them, the global economy pulled consumers into webs of relationships that diffused both the knowledge of risks and the responsibility for regulation and compensation. Epitomizing the plight of industrialization's innocent victims, millions of asbestos workers in mines, shipyards, and other industries were unknowingly exposed to severe health hazards, and hundreds of thousands died of consequent illnesses. Resulting lawsuits financially crippled asbestos manufacturers without adequately recompensing injured workers. A 1984 accident at a US subsidiary's chemical plant in Bhopal, India, again produced hundreds of thousands of victims, with massive scientific and legal uncertainties clouding the settlement of claims.

As the pervasiveness of industrial hazards became apparent, legislatures sought to ensure that regulation would not wait for proof in the form of 'body counts': risk, not harm, became the basis for action under new laws safeguarding health, safety, and the environment, and state agencies struggled to develop credible processes for assessing and mitigating risks before anyone was injured. The sciences, for their part, responded with attempts to mobilize existing knowledge and methods to provide more useful tools for analyzing risks and crafting preventive strategies.

From relatively modest beginnings, rooted largely in the engineering and physical sciences (as in assessing the risks of meltdown in nuclear power plants), risk assessment in support of regulatory standards grew into an important, and increasingly contested, branch of the social sciences. Risk studies acquired the indicia of professionalization, with dedicated journals, societies, curricula, and centers at major universities. One focus of research and development in this field, particularly in the USA, was the construction of reliable models for measuring risk, from the use of animal species as surrogates for humans to sophisticated methods of representing uncertainty in risk estates (National Research Council 1994). Risk assessment presented not only methodological difficulties for scientists and engineers, but also institutional challenges for governments needing to persuade industry and the public that their decisions were scientifically sound and politically balanced (Jasanoff 1990). Lawsuits questioning the risk assessment practices of federal agencies were not uncommon and led to several significant judicial decisions, including Supreme Court rulings on the validity of the occupational standard for benzene in 1980 and the air quality standard for ozone and particulates in 2001.

Other researchers meanwhile approached risk from the standpoint of social psychology, asking why lay observers frequently diverged from experts in their judgments concerning the relative significance of different types of risk. This question engendered a sizeable body of experimental and survey work on risk perception (Slovic 2000), as well as theoretical critiques of such work, showing problems in both the framing and characterization of lay attitudes (Irwin and Wynne 1994). While some social scientists occupied themselves with identifying and measuring 'real' risks, or perceptions of them, others treated *risk* itself as a new organizing principle in social life, noting that vulnerability to risk did not map neatly onto earlier divisions of society according to race, class or economic status (Beck 1992). By contrast, the US environmental justice movement generated data suggesting that social inequality still mattered in the distribution of industrial risks, which fell disproportionately on poor and minority populations (Bullard 1990). This strategic, though controversial, research led in 1994 to a presidential executive order demanding that US federal agencies should consider the equity implications of their actions; equity analysis emerged in this way as an offshoot of existing methodologies for assessing risk and offered formal justification for legal claims of environmental *in*justice.

Unlike risk analysis, which is a product of twentieth-century concerns, law and the mental health disciplines have existed in a relationship of close reciprocity at least since 1843, when the British House of Lords laid down the so-called M'Naghten rule for determining whether a criminal defendant is entitled to relief on grounds of insanity. The rule holds that a person may be released from criminal responsibility for actions taken while he or she was unable to distinguish between right and wrong as a result of mental disease or defect.

More recently, the human sciences' ability to sort people's behavior and capacities into seemingly objective, 'natural' categories has served the needs of both democratic and totalitarian states in a growing variety of legal settings. The need to determine people's mental states in civil as well as criminal proceedings was an important driver of the legal system's enrolment of professionals from these fields. In US law, for example, psychiatric evidence was used to determine whether a criminal accused could stand trial, whether conditions for sentencing leniency or parole were met, and whether a convicted criminal posed a sufficient threat of long-term dangerousness to

deserve the death penalty. While these intersections provided new professional opportunities for psychiatry and psychology, spawning a thriving pedagogic specialty of law and mental health (Gutheil and Appelbaum 2000), entanglements with the law also disclosed frequent divisions among experts and undermined disciplinary authority in other ways.

The most visible challenge to professional autonomy occurred in the 1974 US case of *Tarasoff v. Regents of the University of California* (redecided in 1976), in which the California Supreme Court ruled that psychiatrists were required to warn (later changed to a duty to protect) third parties threatened by their patients. Although the decision did not open the floodgates to liability as many therapists had feared, *Tarasoff* was widely seen as influencing therapeutic practice, often in counterproductive ways. In another notable decision affecting the mental health field, the US Supreme Court in *Barefoot v. Estelle* (1983) essentially ignored an *amicus* brief by the American Psychiatric Association stating that psychiatric predictions of dangerousness are wrong in as many as two out of three cases. The Court expressed confidence in juries' capacity to evaluate expert evidence even if it is weak or deficient.

By the end of the twentieth century, the alliance between law and the human sciences—joined, eventually, by the nascent science of genomics—helped to define a host of novel disorders and syndromes that were invoked to extenuate or penalize various types of conduct. Post-traumatic stress disorder (PTSD), for example, along with its specific manifestations such as rape trauma syndrome, entered the *DSM-IV*, the official diagnostic handbook for psychologists. So codified, PTSD became not only an identifiable and presumably treatable medical condition, but also a basis for asserting new legal claims (as in lawsuits for stress-related psychological injury in the workplace) or defending oneself against them (as in murder trials of sexually abused defendants). Possibly the most notorious example of synergy between law and psychology occurred in a rash of so-called recovered memory and child abuse cases, which became especially prevalent in the US in the 1980s (Hacking 1995). Initially gaining credibility from psychiatric experts, who testified to the reliability of recovered childhood memories, these cases caused scandal as witnesses recanted and it became apparent that some experts had more likely instilled than elicited the shocking 'memories.' The profession eventually organized to deny the scientific validity of the claims and repudiate the expertise of the claimants (Loftus and Ketcham 1994).

Research on eyewitness identification, another branch of research dealing with memory and recall, also received its chief impetus from the legal process. Prompted by insupportably high error rates on the part of eyewitnesses, psychologists tested people's ability to recognize persons they had encountered in stressful situations and discovered that many factors, such as race, inhibit accurate identification under these circumstances (Wells 1988). These findings began to have an impact on police procedure by the late 1990s as awareness dawned that relatively simple changes, such as sequential rather than simultaneous presentation of suspects in line-ups, could substantially improve the reliability of eyewitness testimony.

3. *Scientific Evidence and Expert Witnesses*

Whereas civil law systems relegated the production of scientific and technical evidence largely to national forensics labs and court-appointed experts, common law proceedings traditionally left it to the parties themselves to produce evidentiary support for their claims. The adversary process, with its right of cross-examination, was assumed to be equal to the task of testing the evidence, no matter how arcane or technical, thereby permitting the fact-finder—the judge or jury—to ascertain the truth. Although commentators sometimes deplored the practice of treating experts as hired guns, the parties' basic right to present their evidence was not in doubt, and courts rarely used their legally recognized power to appoint independent experts who might offer a more disinterested view (Jasanoff 1995).

By the turn of the century, several factors combined to challenge this hands-off attitude. The sheer volume of cases demanding some degree of technical analysis was one contributing cause, especially in the US, where low entry barriers to litigation and the inadequacy of social safety nets routinely brought into the courts controversies that were resolved by other means in most common law jurisdictions. Highly publicized instances of expertise gone awry, as in the recovered memory cases, shook people's confidence in the power of cross-examination to keep out charlatans presenting themselves as scientists. The economic consequences of successful products liability lawsuits strained the system from yet another direction, especially in a growing number of mass tort actions that threatened industries with bankruptcy. It was tempting to blame many of these developments on the perceived weaknesses of the adversary system: the passivity and low scientific acumen of judges, the technical illiteracy of juries, the gamesmanship of lawyers, the bias or incompetence of experts selected by parties. These deficits led to calls for reforming the process by which expert testimony was admitted into the courtroom.

The initial round of critique marked a turn away from the liberal ideals of the 1970s and proved to be politically powerful, although it was neither methodologically rigorous nor of lasting scholarly value. An early broadside accused courts of deciding cases on the basis of 'junk science,' a rhetorically useful concept that took hold with opponents of the burgeoning tort

system but proved difficult to characterize systematically (Huber 1991). The barely concealed program of this and related writings was to remove as much as possible of the testing of scientific evidence from the adversary system, particularly from the jury's purview, and to decide these issues either in judicially managed pretrial proceedings or with the aid of scientists appointed by the courts. Undergirding this literature was an almost unshakable faith in science's self-corrective ability and a technocratic conviction that truth would eventually win out if only the legal system left scientific fact-finding to scientists (Foster and Huber 1997). These attacks also tacitly presumed that mainstream expert opinion could be identified on any issue relevant to the resolution of legal disputes. These positions would later be thrown into doubt, but not until the spate of polemical work had left its mark on American scientific and social thought.

Studies by practicing scientists, physicians and some legal academics reiterated and amplified the theme of the law's misuse of science. Critics cited as evidence a number of tort cases in which multimillion dollar jury verdicts favorable to plaintiffs conflicted with the opinions of respected scientists who denied any causal connection between the alleged harmful agent, such as a drug or workplace toxin, and the harm suffered. Particularly noteworthy was the litigation involving Bendectin, a medication prescribed to pregnant women for morning sickness and later suspected of causing birth defects in their children. Juries frequently awarded damages despite assertions by epidemiologists that there was no statistically significant evidence linking Bendectin to the claimed injuries. Research on the role of experts in these cases showed suggestive differences in behavior (such as higher rates of repeat witnessing) between plaintiffs' and defendants' experts (Sanders 1998), leading some to question whether judges were adequately screening offers of scientific testimony. Another episode that drew considerable critical commentary was litigation involving silicone gel breast implants. Close to a half-million women surgically implanted with these devices sued the manufacturer, Dow Corning, claiming injuries that ranged from minor discomfort to permanent immune system damage. The company's initial settlement offer broke down under the onslaught of lawsuits, but epidemiological studies, undertaken only after the commencement of legal action, indicated no causal connection between the implants and immune system disorders. Publication of these results in the prominent *New England Journal of Medicine* led its executive editor to join the chorus of accusation against the legal system's apparent misuse of science (Angell 1996).

Rising discontent about the quality of scientific evidence led the US Supreme Court to address the issue for the first time in 1993. In *Daubert v. Merrell Dow Pharmaceuticals*, the Court set aside the so-called *Frye* rule that had governed the admissibility of expert testimony for the preceding 70 years. Instead of simply demanding that scientific evidence should be 'generally accepted' within the relevant peer community, the Court urged judges to screen science in pretrial proceedings, in accordance with standards that scientists themselves would use. The Court offered four sample criteria: (a) was the evidence based on a testable theory or technique, and had it been tested; (b) had it been peer reviewed; (c) did it have a known error rate; and (d) was the underlying science generally accepted? Two more major evidence decisions of the 1990s cemented the high court's message that trial judges should play a vastly more proactive gate-keeping role when confronted by scientific and technical evidence. As judicial involvement on this front increased, new complaints appeared on the horizon: that judges were using the *Daubert* criteria as an inappropriately inflexible checklist rather than as the guidelines they were meant to be; that mechanical application of *Daubert* and its progeny was trumping deserving claims; that judges were usurping the jury's constitutional role; and that misinterpretation of *Daubert* was introducing unscientific biases and impermissibly raising the burden of proof in civil cases.

In most civil law jurisdictions, by contrast, the inquisitorial approach to testing evidence, coupled with the state's near monopoly in generating forensic science, precluded much controversy about the legitimacy of experts or the quality of their testimony. Tremors could be detected, however, arising from such episodes as the discovery of substandard or fabricated evidence in British trials of Irish terrorists and child abuse suspects and the French government's cover-up of the contamination of the blood supply with the HIV-AIDS virus. Ironically, as US courts moved to consolidate their gate-keeping function in the 1990s, constricting the entry routes for experts, a countermove could be discerned in some European legal systems to open the courts, and policy processes more broadly, to a wider diversity of expert opinion (Van Kampen 1998).

4. *Sites of Knowledge-making*

Investigation of the law-science relationship took a considerably more sophisticated turn in the 1990s as it attracted the attention of a new generation of researchers in science and technology studies. From this disciplinary vantage point, the use–misuse framing that dominated popular writing appeared superficial in the extreme when set against the observed richness of interactions between the two institutions. Rather, as this body of work began to document, it was the hybridization of law and science that held greatest interest and significance for science and society. At this active frontier of social problem solving, one could observe in fine-grained detail how important normative and epistemological commitments either reinforced or contradicted each other in contemporary

societies. For example, even within the common law's adversarial environment, the ritualistic character of legal proceedings successfully held the line against certain forms of radical skepticism, leaving untouched traditional hierarchies of expert authority (Wynne 1982), a belief in the ultimate accessibility of the truth (Smith and Wynne 1989), and an abiding faith in technological progress (Jasanoff 1995).

At the same time, the legal process increasingly came to be recognized as a distinctive site for the production of otherwise unavailable knowledge and expertise. Even in cases where litigation arguably produced anomalous results—such as Bendectin and breast implants—resort to the courts demonstrably added to society's aggregate store of information. Other cases were unambiguously progressive. A particularly striking example was the importation of molecular biological techniques into the arenas of criminal identification and paternity testing in the form of 'DNA typing.' Not only did the legal system's needs provide the initial context for the technique's development, but subsequent contestation over the reliability of DNA-based identification had impacts on the field of population genetics and spurred the formation of new testing methods and agencies, as well as the standardization of commonly used tests (Lynch and Jasanoff 1998). As the technique became black-boxed, new market niches for it emerged within the legal process, most spectacularly as a means of exonerating mistakenly convicted persons. The success of DNA as a forensic tool thus helped to destabilize the legal system's longstanding commitment to eyewitness evidence, and even its reliance on the seemingly unassailable authority of fingerprints (Cole 2001), thereby facilitating critiques of a possibly antiquated privileging of visual memory (Wells 1988). But the massive uptake of forensic DNA analysis by the criminal justice system also opened up new areas of social concern, such as privacy and database protection, which called for continued vigilance and ingenuity on the part of the legal system.

The rise of DNA typing, like that of risk analysis and post-traumatic stress disorder, captured the dynamics of an epoch in which the establishment of any form of social order became virtually unthinkable without the mutual engagement of law and science. In such a time, commentary deploring the law's alleged misuse of science appeared increasingly reductionist and out of touch with everyday reality. The emergence of science and law as a special topic within science and technology studies responded to the inadequacies of earlier analyses and showed greater intellectual promise. A major contribution of this work was to abandon once for all the pared down 'clashing cultures' model of the law–science relationship and to put in its place a more nuanced picture of the positive interplay between normative and cognitive authority. By exploring in detail how the law not only 'uses' science, but also questions it and often compensates for its deficiencies, the new scholarship on science and law invited reflection on the intricate balancing of truth and justice in technologically advanced societies.

See also: Biotechnology; Disciplines, History of, in the Social Sciences; Expert Testimony; Expert Witness and the Legal System: Psychological Aspects; Intellectual Property, Concepts of; Intellectual Property: Legal Aspects; Law: History of its Relation to the Social Sciences; Legal Process and Social Science: United States; Norms in Science; Power in Society; Professions, Sociology of; Science and Technology Studies: Experts and Expertise; Science and the State; Science, Economics of; Truth and Credibility: Science and the Social Study of Science

Bibliography

Angell M 1996 *Science on Trial: The Clash of Medical Evidence and the Law in the Breast Implant Case*. Norton, New York

Beck U 1992 *Risk Society: Towards a New Modernity*. Sage, London

Bullard R D 1990 *Dumping in Dixie: Race, Class, and Environmental Quality*. Westview, Boulder, CO

Cole S 2001 *Manufacturing Identity: A History of Criminal Identification Techniques from Photography Through Fingerprinting*. Harvard University Press, Cambridge, MA

Ezrahi Y 1990 *The Descent of Icarus: Science and the Transformation of Contemporary Democracy*. Harvard University Press, Cambridge, MA

Foster K R, Huber P W 1997 *Judging Science: Scientific Knowledge and the Federal Courts*. MIT Press, Cambridge, MA

Foucault M 1990 *The History of Sexuality*. Vintage, New York

Fuller L 1964 *The Morality of Law*. Yale University Press, New Haven, CT

Goldberg S 1994 *Culture Clash*. New York University Press, New York

Gutheil T G, Appelbaum P S 2000 *Clinical Handbook of Psychiatry and the Law*, 3rd edn. Lippincott Williams & Wilkins, Philadelphia, PA

Hacking I 1995 *Rewriting the Soul*. Princeton University Press, Princeton, NJ

Hart H L A 1961 *The Concept of Law*. Oxford University Press, Oxford, UK

Huber P 1991 *Galileo's Revenge: Junk Science in the Courtroom*. Basic Books, New York

Irwin A, Wynne B (eds.) 1994 *Misunderstanding Science*. Cambridge University Press, Cambridge, UK

Jasanoff S 1990 *The Fifth Branch: Science Advisers as Policymakers*. Harvard University Press, Cambridge, MA

Jasanoff S 1995 *Science at the Bar: Law, Science and Technology in America*. Harvard University Press, Cambridge, MA

Loftus E F, Ketcham K 1994 *The Myth of Repressed Memory: False Memories and Allegations of Sexual Abuse*. St. Martin's Press, New York

Lynch M, Jasanoff S (eds.) 1998 Contested identities: Science, law and forensic practice. *Social Studies of Science* **28**: 5–6

Minow M 1990 *Making all the Difference: Inclusion, Exclusion, and American Law*. Cornell University Press, Ithaca, NY
National Research Council (NRC) 1994 *Science and Judgment in Risk Assessment*. National Academy Press, Washington, DC
Polanyi M 1962 The republic of science. *Minerva* **1**: 54–73
Price D K 1985 *America's Unwritten Constitution: Science, Religion, and Political Responsibility*. Harvard University Press, Cambridge, MA
Sanders J 1998 *Bendectin on Trial: A Study of Mass Tort Litigation*. University of Michigan Press, Ann Arbor, MI
Schuck P 1993 Multi-culturalism redux: Science, law, and politics. *Yale Law and Policy Review* **11**: 1–46
Shapin S 1994 *A Social History of Truth*. University of Chicago Press, Chicago
Shapin S, Schaffer S 1985 *Leviathan and the Air-pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, Princeton, NJ
Slovic P 2000 *The Perception of Risk*. Earthscan, London
Smith R, Wynne B (eds.) 1989 *Expert Evidence: Interpreting Science in the Law*. Routledge, London
Van Kampen P T C 1998 *Expert Evidence Compared: Rules and Practices in the Dutch and American Criminal Justice System*. Intersentia, Antwerp
Wagner P, Wittrock B, Whitley R (eds.) 1991 *Discourses on Society: The Shaping of the Social Science Disciplines*. Kluwer, Dordrecht, The Netherlands
Wells G L 1988 *Eyewitness Identification: A System Handbook*. Carswell, Toronto
Wynne B 1982 *Rationality and Ritual: The Windscale Inquiry and Nuclear Decisions in Britain*. British Society for the History of Science, Chalfont St. Giles, UK

S. Jasanoff

Science and Religion

Understanding the present relationship between science and religion requires a recognition of broad historical trends. The subject has been most commonly treated in depth by historians of science, though scholars in both pure science and theology have also been active. Several posts in the UK, including a professorship at Oxford, and at least three major journals are devoted to the subject of science and religion (*Zygon*, *Perspectives on Science and Christian Faith*, and *Science and Christian Belief*), and numerous organizations have been formed to promote its study both in the US and UK.

Scholarship has become increasingly sophisticated, and frequent attacks on earlier 'simplistic' views have had the effect of excluding from the debate a great many ordinary people. This is quite unnecessary, though it is true that both 'science' and 'religion' are not eternally unchanging entities and do vary over time and space. In what follows, science is simply the study and systematic observation of the natural world, while religion may be seen generally as humanity's quest after God, a quest taking a variety of different forms. The subject is, of course, an emotive issue for some, from vociferous antagonists of religion (like the geneticist Richard Dawkins) to equally uncompromising supporters (such as the American 'creationists').

In the following account the main emphasis will be on Christianity, not for partisan reasons but because, historically, that religion more than any other has related closely to the emerging sciences.

1. *The Social Relations of Science and Christianity*

This topic has become the subject of a number of well-known statements or theses, five of which will now be considered. Some are about specific periods in history (as the seventeenth century) while others relate to allegedly more timeless issues. Some bear well-known names of their chief proponents.

1.1 *The Merton Thesis*

In 1938, the American sociologist Robert Merton argued that the emergence of science in the seventeenth century was promoted by the ascendancy of Protestant religion. Since then this contentious position has been revisited again and again, and it is hard to resist the conclusion that, in modified form, the thesis has more than an element of truth. Certainly, there is a correlation between visibility in science and religious allegiance. In the French *Académie des Sciences* from 1666 to 1885, the ratio of Protestants to Catholics was 80:18, despite the preponderance of Catholics in the country. Others have shown a similarly high proportion of Puritans (not merely Protestants) in the membership of the early Royal Society in England, and Merton argued that Puritan attitudes did much to encourage its growth.

There have been objections to such 'counting of heads,' not least because of the difficulty of defining a Puritan. If such a person is considered theologically rather than politically some of the difficulties disappear. On this basis a Puritan is one who holds strongly to the teaching of the Bible as opposed to the church or tradition, not necessarily a supporter of the Parliamentary cause in revolutionary England. Yet it may still be argued that such correlations do not prove that Puritan theology encouraged science, for may not both have emerged from a common cause? Could not both have been an expression of new movements of social and economic change and of a libertarian philosophy? This may well be true, but a general correlation in the 1600s between promotion of science and a generally Protestant loyalty to the Bible seems inescapable.

1.2 The Hooykaas Thesis

This goes further than the Merton thesis might suggest, and argues that the origins of modern science do not merely correlate neatly with Protestant beliefs but are directly derived from them. It has been implied by many authors since the 1970s but is particularly associated with the Dutch historian of science Reijer Hooykaas (1906–94) in his epochal *Religion and the Rise of Modern Science*. To make such an assertion is to stand traditional views on their head. It was once customary to see science as a truly Greek inheritance, at last freed from religious shackles at the Renaissance. Hooykaas invites us to view it rather as a product of Biblical theology, freshly released at the Reformation, and subverting much of Greek philosophy which had for 1500 years inhibited any real rise of experimental science.

Serious evidence exists in support of this thesis. There are explicit declarations by many well-known scientific figures from Francis Bacon to Isaac Newton and beyond that their science was inspired theologically. Then there is a remarkable congruity between Biblical and scientific credos. At least five points of intersection can be identified. (a) Most profound of all, perhaps, was what Hooykaas called the 'demythologization of nature,' so that nature was no longer to be seen as divine or even animate, but more like a machine than anything else. It is not God but a creation by him, and therefore a proper object of study and experiment. Such was the theology of the Old and New Testaments, and it was eloquently proclaimed by statements from 'the father of chemistry' Robert Boyle and many others. (b) Moreover, the idea that nature worked by laws is at the very foundation of science, and is also Biblical. Amongst those who wrote of laws impressed by God on nature were Descartes, Boyle, and Newton. Men looked for laws when they recognized the law-giver. (c) But science can only discover these laws by experimentation. Manipulation of nature had been regarded by many Greeks as socially unacceptable (except for slaves) or even impious. Among the most urgent advocates of an experimental method, based on widespread Biblical approval of manual techniques for testing, was Francis Bacon who urged 'men to sell their books, and to build furnaces.' (d) Then again a further religious impetus to science came from the Biblical exhortations to see the heavens and earth as manifesting the glory of their Creator. The astronomer Kepler, it is said, asserted that in his researches he 'was thinking God's thoughts after him.' This religious motivation became a strong emphasis in Calvinistic theology. (e) Finally, the Biblical mandate for humanity to exert 'dominion' over nature opened up possibilities of scientific work 'for the glory of God and the relief of man's estate' (Bacon). As Hooykaas wrote:

> The Biblical conception of nature liberated man from the naturalistic bonds of Greek religiosity and philosophy and gave a religious sanction to the development of technology [Hooykaas, *Religion and the Rise of Modern Science*, 1973, p. 67].

One cannot dismiss the burgeoning literature linking science to Protestantism in the seventeenth century as mere rhetoric. But it can be argued that Hooykaas underestimates the numbers of Roman Catholics who were eminent in science, though even here caution is needed. The oft-quoted counter example of the Catholic Copernicus must not obscure the astronomer's indebtedness to both his Lutheran assistant Rheticus, and the liberal legacies of Erasmus within his own church. Attitudes towards nature varied widely within Catholicism, and there is always the danger of generalising too widely. But in its essence the Hooykaas thesis appears to be substantially correct.

1.3 The Lynn White Thesis

If Christian theology has been one of the formative influences on modern science it does not follow that this has always been in the best interests of the world and its inhabitants. That opinion has been forcibly expressed in a further thesis, first proposed by the American historian Lynn White in 1966/7. He argued that much of the damage to our environment springs from a misuse of science and technology for which 'Christianity bears a huge burden of guilt.' More specifically he locates the problem in the 'realisation of the Christian dogma of man's transcendence of, and rightful mastery over, nature' (White 1967). He urges a return not to the primitive Christianity of the New Testament but to the animistic world of St Francis of Assisi which saw the earth and its inhabitants as brothers rather than instruments. In White's view the Biblical call to 'dominion' must have been understood in terms of exploitation.

However, careful historical examination of the evidence suggests that few, if any, pioneers in science or theology held this particular view. John Calvin explicitly repudiated it, as did the noted eighteenth century writer William Derham. They, and many others, urged an interpretation of dominion as 'responsible stewardship.' In more recent times ideologically driven human conquest of nature has been on the Marxist rather than the Christian agendas. It is now widely acknowledged that, as a historical generalization, the Lynn White thesis stands largely discredited.

1.4 The Conflict Thesis

This thesis is much older than the others we have examined, and much better known. The thesis states that science and religion have been in perpetual conflict and that, eventually, science will vanquish

religion. It was most notoriously promulgated by the Victorian naturalist T. H. Huxley, but has been advocated by many others in the late nineteenth century and is probably best not attributed to any one individual. Two books arguing the point are *History of the Conflict between Religion and Science* by J. W. Draper (first published in 1875), and *A History of the Warfare of Science with Theology in Christendom* by A. D. White (1895). They achieved enormous circulation and are still in print.

The essence of the argument is that where scientific conclusions have been challenged by the church there has usually been an eventual retreat by the ecclesiastical authorities. Two classic cases were those of Galileo in the seventeenth century and of Darwin nearly 250 years later. It was in the aftermath of the latter *imbroglio* that the conflict thesis was formulated. Many more examples were unearthed by Draper and White and an impressive case assembled (especially by the latter).

Manifestly, it is hard to reconcile this thesis with the three previously considered. If science owed so much to religion how could they possibly be in conflict? The thesis flies in the face of massive evidence that points to a powerful alliance between science and Christianity since at least the seventeenth century. A noteworthy feature of the conflict thesis is that it was first proposed at a time when history of science hardly existed and historical scholarship had certainly not explored the nuances perceived by Merton, Hooykaas, and their colleagues. Detailed examination of the relatively minor cases urged by Draper and White reveals that many were badly documented, some plainly apocryphal and others greatly exaggerated. It seems that the authors sometimes saw what they were looking for and built their generalizations on a slender foundation. To be sure, Galileo and Darwin were assailed by organized religion but they seem to have been relatively isolated exceptions to the general rule that Christianity and science coexisted in harmony. Partial explanations for the persecution of Galileo lay in power struggles within the church, while Darwin's problems arose in part from a deeply divided society in industrial Britain.

Today the books of Draper and White are rated not as serious works of historical scholarship but as highly polemical tracts reflecting the tensions existing in the social and cultural environment in which the authors lived. Draper wrote as a man deeply disenchanted with the Roman Catholic church, not least for its recent declaration of papal infallibility. White, on the other hand, was President of one of the first explicitly nonsectarian colleges in the US (Cornell) and had suffered badly at the hands of the religious establishment. Each had his own axe to grind in taunting the organized church. The result was not history but myth.

Yet a conflict thesis is commonly held in the world today, so one needs to asks how such a tendentious myth could have become so entrenched in Western culture. To understand the reason it is necessary to recall the plight of English science in the last 40 years of Victoria's reign. Underfunded by government, inadequately supported by secondary education, unpopular with the general public for a variety of reasons and a poor relation in those citadels of the establishment, the universities of Oxford and Cambridge, English science was crying out for recognition and support. It was falling seriously behind in the race with Continental science (especially German) and had no chance of occupying a leading place in English culture. In search of a weapon with which to acquire some kind of hegemony Thomas Henry Huxley and some close allies determined on a course of action. They would seek by all means within their power to undermine the privileged position of the Anglican church and put science in its place. A detailed strategy was worked out, an important element being the propagation of a conflict thesis in which the church was always portrayed as the vanquished party in its perennial fights with science. The books of Draper and White were at their service, and so was evolution which, as Huxley said, was a good stick with which to beat the church. With all the discipline of a military campaign the attack was launched and the conflict thesis enshrined as a self-evident truth in English culture. It was a demonstrably wrong but it carried the day, at least with the general public (see Russell 1989).

1.5 The Complementarity Thesis

In the 1960s and 1970s the scientist/philosopher Donald MacKay made much of the paradigm of complementarity. Derived from modern physics, where, for instance, wave and particle descriptions of light are complementary, the notion was applied to explanations from science and theology. It was the antithesis of a reductionist view where objects or events were nothing but phenomena of one kind (usually materialist). This view did not preclude traffic from one area to another, and did not deny their mutual influence. Indeed MacKay was committed strongly to the Hooykaas thesis. But it did mean that science and theology could co-exist as complementary enterprises.

We may call this the complementarity thesis. Its merits include the ability to account for several phenomena of modern and recent science. One of these is the large number of scientists who hold to Christian beliefs. In Huxley's day they included many of the most eminent physical scientists as Faraday, Joule, Stokes, and Lord Kelvin. Statistical surveys undertaken since that day have confirmed the same general trend, which is quite the opposite of common expectation. In the early years of the twenty-first century the numbers appear to be increasing. The only reasonable explanation is that such people regard their

scientific and religious activities as complementary and nonthreatening. Membership data and corporate activities of organisations like the American Scientific Affiliation and Christians in Science appear to bear this out.

In modern times an interesting tendency strengthens still further the case for a complementary approach. That is the rebirth of natural theology, in which the natural universe is seen to testify to the power and goodness of God. Such arguments go back to Biblical times, but were advocated with fresh vigour in the late eighteenth and early nineteenth centuries, reaching their peak with *Natural Theology* by William Paley, first appearing in 1802. The argument for a Designer from the apparent design of the natural world received hard blows from the philosophical critique of Hume and the evolutionary theories of Darwin. The proponents of the conflict thesis added their opposition. Yet over 100 years later the disclosures of modern science have caused some cynics to have second thoughts, and have encouraged scientist/theologians like John Polkinghorne to revive a form of natural theology that does justice to a natural world infinitely more complex than that perceived by Paley or even Darwin.

2. *Beyond Traditional Christianity*

It remains to note that orthodox Christianity has no monopoly in engagement with science, though historically it happens to have been the most prominent faith to do so. Much literature on science and religion has focused, therefore, on the Christian faith. A modern variant of Christianity is sufficiently different to justify the term heterodox; that is the so-called process theology that denies omniscience to God and regards natural processes as contributing to his self-fulfillment. It has been extensively used in discussions about indeterminacy in nature, about human free will and about the phenomenon of evil. Though favoured by some philosophers and theologians it cuts little ice with the majority of scientists.

The other great monotheistic religions of Judaism and, to a smaller extent, Islam have encouraged their followers in the study of nature though, operating in very different cultures and with different presuppositions, their effects have been rather unlike those of Christianity. The eastern mystical religions have arguably been less important for the growth of science, though some parallels between atomic physics and Buddhism have been emphasized by Franz Capra. The early promise of Chinese science was not fulfilled because the concept of scientific law was hardly present, reflecting, according to the historian Joseph Needham, the absence of any monotheistic law-giver in ancient Chinese culture.

Finally, the growth of environmental awareness has led some to espouse an almost pantheistic view of the universe as an alternative to the view of it as a machine which can be used or abused at whim. They have been encouraged by the *Gaia* hypothesis of James Lovelock which emphasizes the complexity of earth and its biosphere and its capacity for self-regulation. This had led several (including Lovelock) to posit the earth as in some sense 'alive.' From this it has been but a step to invoke the earth (*Gaia*) as a mother and even a goddess. If taken to extremes this view may lead to a greater sensitivity to the needs for sustainable development and other desirable attitudes to the earth. But it may also lead to a reversion to a prescientific animistic or even divine universe in which the benefits of a thriving science are rejected as minimal. If, however, the values of Christian monotheism are integrated with these views it may be possible for science to prosper, but always in a spirit of responsible stewardship. At present, we seem a long way from either of these positions.

See also: Creationism, Evolutionism, and Antievolutionism; Religion and Politics: United States; Religion and Science; Religion: Evolution and Development; Religion, History of; Religion, Psychology of; Religion, Sociology of

Bibliography

Barbour I 1990 *Religion in an Age of Science*. Harper, San Francisco

Brooke J H 1991 *Science and Religion: Some Historical Perspectives*. Cambridge University Press, Cambridge, UK

Finocchiaro M A 1989 *The Galileo Affair: A Documentary History*. University of California Press, Berkeley, CA

Harrison P 1998 *The Bible, Protestantism and the Rise of Natural Science*. Cambridge University Press, Cambridge, UK

Hooykaas R 1972 *Religion and the Rise of Modern Science*. Scottish Academic Press, Edinburgh, UK

Jeeves M A, Berry R J 1998 *Science, Life and Christian Belief*. Apollos, Leicester, UK

Kaiser C 1991 *Creation and the History of Science*. Marshall Pickering, London

Lindberg D C, Numbers R (eds.) 1998 *God and Nature: Historical Essays on the Encounter Between Christianity and Science*. University of California Press, Berkeley, CA

Livingstone D N, Hart D G, Noll M A (eds.) *Evangelicals and Science in Historical Perspective*. Oxford University Press, New York

Moore J M 1979 *The Post-Darwinian Controversies: A Study of the Protestant Struggle to Come to Terms with Darwin in Great Britain and America 1870–1900*. Cambridge University Press, Cambridge, UK

Nebelsick H P 1992 *The Renaissance, the Reformation and the Rise of Science*. T. & T. Clark, Edinburgh, UK

Polkinghorne J 1990 *One World: The Interaction of Science and Theology*. SPCK, London

Polkinghorne J 1991 *Reason and Reality: The Relationship Between Science and Theology*. SPCK, London

Russell C A 1985 *Cross-currents: Interactions Between Science and Faith*. Inter-Varsity Press, Leicester, UK (corrected and reprinted, Christian Impact, London, 1996)

Russell C A 1989 The conflict metaphor and its social origins. *Science and Christian Belief* **1**: 3–26
Russell C A 1994 *The Earth, Humanity and God.* University College Press, London
White L 1967 *Science* **155**: 1203

C. A. Russell

Science and Social Movements

The social movements that are discussed in this article refer to major political processes of popular mobilization that have significant societal significance, and include organized forms of protest as well as broader and more diffuse manifestations of public opinion and debate. In relation to science, social movements conceived in this way can be seen to have served as seedbeds for new modes of practicing science and organizing knowledge more generally, and also as sites for critically challenging and reconstituting the established forms of scientific activity.

1. *Introduction*

Since they fall between different academic specializations and disciplinary concerns, the relations between science and social movements have tended to be a neglected subject in the social sciences. For students of social movements in sociology and political science, the relations to science have been generally of marginal interest and received little systematic attention (Dellaporta and Diani 1999), while in science and technology studies, social movements usually have been regarded more as contextual background than as topics of investigation in their own right. Considering their potential significance as 'missing links,' both in the social production of knowledge and in broader processes of political and social change, however, the relations between science and social movements deserve closer scrutiny. For one thing, new ideas about nature and society have often first emerged outside the world of formal scientific activity, within broader social and political movements. Social movements have also often provided audiences, or new publics, for the spreading and popularization of scientific findings and results. Movements of religious dissent were among the most significant disseminators of the new experimental philosophy in the mid-seventeenth century, for example, while, in the nineteenth and twentieth centuries, new approaches to medicine, social relations, gender roles, and the environment found receptive audiences within social and political movements. Most visibly, social movements and their spokespersons have offered critical perspectives on both the form and content of science, as well as on the uses to which science has been put. In recent decades, a range of so-called new social movements have criticized the dominant biases and perspectives of many scientific and technological fields. This article first presents these different forms of interaction between science and social movements in general, historical terms. It then discusses contemporary relations between science and social movements, and, in particular, the environmental movement.

2. *Historical Perspectives*

A fundamental insight of the sociology of knowledge, as it developed in the 1920s and 1930s, was to suggest that modern science emerged in the seventeenth century out of a more all-encompassing struggle for political freedom and religious reform (Scheler 1980/1924; Merton 1970/1938). What eventually came to be characterized as modern science represented an institutionalized form of knowledge production that, at the outset, had been inspired by more sweeping social and political transformations. The historical project of modernity did not begin as a new scientific method, or a new mechanical worldview, or, for that matter, as a new kind of state support for experimental philosophy in the form of scientific academies. As for the Reformation, it arose as a more deep-seated challenge to traditional ways of thought in social and religious life. It was a protest against the Church (which is why the members of the movement were called protestants), and it was an encompassing social and cultural movement that articulated and practiced alternative, or oppositional forms of religion, politics, and learning as part of its political

Table 1
A schematic representation of the scientific revolution from movement to institution

From movement…	*To institution*
Social reform	Reform of philosophy
Millenarian vision	Experimental program
Connection to radical politics	Rejection of politics
Decentralized structure	Central academy
Democratic/open to all	Elitist/professional
Spiritual (absolute) knowledge	Instrumental (probabalistic) knowledge
Technical-economic improvement	Scientific development
Informal communication pamphlets, correspondence	Formal communication journals, papers

struggle (Mandrou 1978). The teachings of Paracelsus, Giordano Bruno, and Tomaso Campanella, to name some of the better known examples, combined a questioning of established religious and political authority with an interest in scientific observation, mathematics, mechanics, and technical improvements (Merchant 1980). But as the broader movements were replaced by more formalized institutions in the course of the seventeenth century, the political and social experiments came to be transformed into scientific experiments; the political and religious reformation, tinged with mysticism and filled with mistrust of authority, was redefined and reconstituted, at least in part, as a scientific revolution (see Table 1).

In the later seventeenth and early eighteenth centuries, the scientific 'aristocracy' that had emerged in London and Paris at the Royal Society and the *Academie des Sciences* was challenged by dissenting groups and representatives of the emerging middle classes, some of whom fled from Europe to the colonies in North America, and some of whom established scientific societies, often in provincial areas in opposition to the established science of the capital cities. Many of them shared with the academicians and their royal patrons a belief in what Max Weber termed the protestant ethic, that is, an interest in the value of hard work and the virtue of making money, and most had an interest in what Francis Bacon had termed 'useful knowledge.' The social movements of the enlightenment objected, however, to the limited ways in which the Royal Society and the Parisian Academy had organized the scientific spirit and institutionalized the new methods and theories of the experimental philosophy. The various attempts to democratize scientific education in the wake of the French Revolution and to apply the mechanical philosophy to social processes—that is, to view society itself as a topic for scientific research and analysis—indicate how critique and opposition helped bring about new forms of scientific practice. Adam Smith's new science of political economy developed in the Scottish hinterland, and many of the first industrial applications of experimentation and mechanical philosophy took place in the provinces, rather than in the capital cities, where the scientific academies were located (Russell 1983).

Many of the political and cultural trends of the nineteenth century—from romanticism and utopianism to socialism and populism—also began as critical movements that at a later stage, and in different empirical manifestations, came to participate in reconstituting the scientific enterprise. Romanticism, for example, first emerged as a challenge to the promethean ambitions of science, in the guise, for example, of Mary Shelley's mad Doctor Frankenstein, while socialism, in the utopian form promulgated by Robert Owen, was a reaction, among other things, to the problematic ways in which the technological applications of scientific activity were being spread into modern societies. Romantic writers and artists, like William Blake, Johann Goethe, and later Henry David Thoreau, were critical of the reductionism and cultural blindness of science and of the 'dark satanic mills' in which the mechanical worldview was being applied. They sought to mobilize the senses and the resources of myth and history in order to envision and create other ways of knowing (Roszak 1973). Some of their protest was transformed into constructive artistic creation, and later in the century, the romantic revolt of the senses inspired both the first waves of environmentalism in the form of conservation societies and the emergence of a new kind of holistic scientific discipline, to which was given the name ecology: household knowledge.

In the late nineteenth century, the labor movement also sought a deep-going, fundamental political, and cultural transformation of society, but it would see its critical message translated, in the twentieth century, into packages of reforms and a more welfare-oriented capitalism, on the one hand, and into the 'scientific socialism' of Leninand Stalin on the other (Gouldner 1980). Once again, however, science benefited from this institutionalization of the challenge of social movements; the knowledge interests of the labor movement, for example, entered into the new social science disciplines of economics and sociology. A political challenge was once again transformed into programs of scientific research and state policy; but while new forms of scientific expertise were developed, little remained of the broader democratization of knowledge production that the labor movement, in its more radical days, had articulated.

In the early twentieth century, the challenge to established science was mobilized most actively in the colonies of European imperialism, as well as in the defeated imperialist powers of Germany and Italy; the critique was primarily of scientific civilization writ large, and of what Mahatma Gandhi in India called its 'propagation of immorality.' In the name of modernism, science stood for the future and legitimated, in the colonies as well as in Europe, a wholesale destruction of the past and of the 'traditional' knowledges—the other ways of knowing—that had been developed in other civilizations (Tambiah 1990). These social movements had an impact on the development of political ideology on both the right and left, but also inspired new sciences of ethnography and anthropology, as well as the sociology of knowledge. Even more important perhaps were the various attempts to combine the artisanal knowledges of the past and of other peoples with the modern science and technology of the present in new forms of architecture, design, and industrial production. Many of the regional development programs of the 1930s and 1940s, in Europe and North America, can trace their roots back to the cultural critique of modernism that was inspired by the Indian independence movement and by such 'movement intellectuals' as William Morris and

Patrick Geddes in Britain. Both Roosevelt's New Deal and the Swedish model welfare state can be said to have mobilized civilizationally critical perspectives in their projects of social reconstruction.

3. *Science and Contemporary Social Movements*

In recent decades, social movements have also served to challenge and reorient the scientific enterprise. Out of the anti-imperialist and student movements of the 1960s and the environmentalist, feminist, and identity movements of the 1970s and 1980s have emerged a range of alternative ideas about science, in form, content, and meaning, that have given rise to new scientific theories, academic fields, and technological programs (Schiebinger 1993, Harding 1998). Out of critique have grown the seeds of new, and often more participatory, ways of sciencing—from technology and environmental impact assessment to women's studies, queer theory, and postcolonial discourses. What were in the 1960s and 1970s protest movements of radical opposition largely have been emptied of their political content, but they have given rise to new branches of, and approaches to, science and technology. While the more radical, or oppositional, voices have lost much of their influence, the more pragmatic and scientific voices have been given a range of new opportunities. This is not to say that there is no longer a radical environmental opposition or a radical women's liberation movement, but radicals and reformists increasingly have drifted apart from one another, and in most countries now work in different organizations, with little sense of a common oppositional movement identity. There has been a fragmentation of what was a coherent movement into a number of disparate bits and pieces (Melucci 1996).

The new social movements rose to prominence in the downturn of a period of institutional expansion and economic growth. They emerged in opposition to the dominant social order and to its hegemonic scientific technological regime, which had been largely established during and immediately after World War II (Touraine 1988). The war led to a fundamental transformation in the world of science and technology, and to the emergence of a new relation, or contract, between science and politics. Unlike previous phases of industrialization, in which science and engineering had lived parallel but separate identities, World War II ushered in the era of technoscience. The war effort had been based on an unprecedented mobilization of scientists to create new weapons, from radar to the atomic bomb, and to gather and conduct intelligence operations. In the process, 'little' science was transformed into big science, or industrialized science.

Especially important for the social movements that were to develop in the 1960s and beyond was the fact that scientific research was placed at the center of postwar economic development. Many of the economically significant new products—nylon and other synthetic textiles, plastics, home chemicals and appliances, television—were based directly on scientific research, and the new techniques of production were also of a different type: it was the era of chemical fertilizers and insecticides, of artificial petrochemical-based process industries, and food additives. The forms of big science also differed from the ways in which science had been organized in the past. The big science laboratories—both in the public and private sectors—were like industrial factories, and scientific-technical innovation came to be seen as an important concern for business managers and industrial organizers. The use of science in society had become systematized, and, as the consequences of the new order became more visible, new forms of mistrust and criticism developed (Jamison and Eyerman 1994).

One wing of the public reacted to the destruction of the natural environment, what an early postwar writer, Fairfield Osborn, termed 'man's war with nature.' The exploitation of natural resources was increasing and, in the 1940s and 1950s, it began to be recognized that the new science-based products were more dangerous for the natural environment than those that had come before. But it would only be with Rachel Carson, and her book *Silent Spring* in 1962, that an environmental movement began to find its voice and its characteristic style of expression. It was by critically evaluating specific instances of scientific technology, particular cases of what Osborn had called the 'flattery of science' that the environmentalist critique would reach a broader public. Carson singled out the chemical insecticides for detailed scrutiny and assessment, but her point was more general. Carson's achievement was to direct the methods of science against science itself, but also to point to another way of doing things, the biological or ecological way—what she called in her book the road not taken.

Another source of inspiration for the new social movements came from philosophers and social historians who questioned the more general impact of technoscience on the human spirit. It was one-dimensional thinking which critical theorists like Herbert Marcuse reacted against, the dominance of an instrumental rationality over all other forms of knowing. For Lewis Mumford, another major source of inspiration, it was the homogenization of the landscape that was most serious, the destruction of the organic rhythms and flows of life that had followed in the wake of postwar economic growth, as well as the dominance of what he termed the 'megamachine,' the use of technology for authoritarian purposes.

In the 1970s, a range of new social movements, building on these and other sources of inspiration, came to articulate an oppositional, or alternative approach to science and technology (Dickson 1974). The so-called new social movements represented an integrated set of knowledge interests, which combined the critique of Rachel Carson with the liberation

dialectics of Marcuse and the direct democracy of the student movement. The new movements involved a fundamental critique of modern science's exploitative attitude to nature, as well as an alternative organizational ideal—a democratic, or participatory ideal—for the development of knowledge. There were also distinct forms of collective learning in the new social movements of environmentalism and feminism, as well as grass-roots engineering activities that went under the name of appropriate or alternative technology.

4. Social Movements as Cognitive Praxis

On the basis of these historical and contemporary relations between science and social movements, many social movements can be characterized as producers of science (Eyerman and Jamison 1991). The critical ideas and new public arenas that are mobilized by social movements often provide the setting for innovative forms of cognitive praxis,combining alternative worldviews, or cosmological assumptions with alternative organizational and practical–technical criteria. In the case of environmentalism, the cosmology was, to a large extent, the translation of a scientific paradigm into a socioeconomic paradigm; in the 1970s, the holistic concepts of systems ecology were transformed into political programs of social ecology, and an ecological worldview was to govern social and political interactions. Technology was to be developed under the general perspective that 'small is beautiful' (in the influential phrase of E. F. Schumacher), and according to the assumption that large-scale, environmentally destructive projects should be opposed and stopped. At the same time, new contexts for education and experimentation and the diffusion of research were created in the form of movement workshops and, in the Netherlands, for example, in the form of science shops, which allowed activist groups to gain access to the scientific expertise at the universities (Irwin 1995).

In the 1980s, this cognitive praxis was decomposed largely into a disparate cluster of organizations and individuals, through processes of professionalization and fragmentation. The knowledge interests of the environmental movement were transformed into various kinds of professional expertise, which made it possible to incorporate parts of the movement into the established political culture, and to shift at least some of the members of the movement from outsider to insider status. Some of the alternative technical projects proved commercially viable—biological agriculture, wind energy plants, waste recycling—and gave rise to a more institutionalized form of environmental politics, science, and technology (Hajer 1995).

Similar processes can be identified in relation to other social movements of recent decades (Rose 1994). The political struggles for civil rights, women's and sexual liberation, and ethnic and national identity have inspired new approaches to knowledge that have since been institutionalized and transformed into established scientific fields, such as women's studies, gay and lesbian studies, African-American studies, as well as in new areas of medicine and technology.

At the same time, science itself has been reconstituted, partly as a result of the critical perspectives and cognitive challenges posed by social movements. Many of the postmodern theories of the cultural and human sciences have been inspired by the experiences of social and political movements. Out of the alternative public spaces that have been created by social and political movements has emerged a new kind of scientific pluralism, in terms of organization, worldview assumptions, and technical application. In the transformations of movements into institutions, a significant channel of cognitive and cultural change can thus be identified. It may be hoped, in conclusion, that these interactions between science and social movements will receive more substantial academic attention in the future.

See also: History of Science; Innovation, Theory of; Kuhn, Thomas S (1922–96); Marx, Karl (1818–89); Social Movements, History of: General; Social Movements: Psychological Perspectives; Social Movements, Sociology of

Bibliography

Dellaporta D, Diani M 1999 *Social Movements. An Introduction*. Routledge, London/Blackwell, Oxford, UK

Dickson D 1974 *Alternative Technology and the Politics of Technical Change*. Fontana, London

Eyerman R, Jamison A 1991 *Social Movements. A Cognitive Approach*. Penn State University Press, University Park, PA

Gouldner A 1980 *The Two Marxisms. Contradictions and Anomalies in the Development of Theory*. Seabury University Press, New York

Hajer M 1995 *The Politics of Environmental Discourse. Ecological Modernization and the Policy Process*. Oxford University Press, Oxford, UK

Harding S 1998 *Is Science Multicultural? Postcolonialisms, Feminisms, and Epistemologies*. Indiana University Press, Bloomington, IN

Irwin A 1995 *Citizen Science. A Study of People, Expertise and Sustainable Development*. Routledge, London

Jamison A, Eyerman R 1994 *Seeds of the Sixties*. University of California Press, Berkeley, CA

Mandrou R 1978 *From Humanism to Science 1480–1700*. Penguin Books, Harmondsworth, UK

Melucci A 1996 *Challenging Codes: Collective Action in the Information Age*. Cambridge University Press, Cambridge, UK

Merchant C 1980 *The Death of Nature. Women, Ecology, and the Scientific Revolution*. Harper and Row, San Francisco

Merton R 1970 (1938) *Science, Technology, and Society in Seventeenth-century England*. Harper and Row, New York

Rose H 1994 *Love, Power and Knowledge. Toward a Feminist Transformation of the Sciences*. Polity Press, Cambridge, UK

Roszak T 1973 *Where the Wasteland Ends: Politics and Transcendence in Post-industrial Society*. Anchor, New York

Russell C 1983 *Science and Social Change 1700–1900*. Macmillan, London
Scheler M 1980 (1924) *Problems of a Sociology of Knowledge*. Routledge, London
Schiebinger L 1993 *Nature's Body: Gender in the Making of Modern Science*. Beacon, Boston
Tambiah S 1990 *Magic, Science, Religion and the Scope of Rationality*. Cambridge University Press, Cambridge, UK
Touraine A 1988 *The Return of the Actor. Social Theory in Postindustrial Society*. University of Minnesota Press, Minneapolis, MN

A. Jamison

Science and Technology, Anthropology of

The anthropology of science and technology is an expanding arena of inquiry and intervention that critically examines the cultural boundary that sets science and technology off from the lives and experiences of people. Most research in this arena draws on ethnographic fieldwork to make visible the significance and force of this boundary, as well as meanings and experiences that cut across it or otherwise get hidden. Key categories of projects include juxtaposing shared systems of cultural meaning, following flows of metaphors back and forth across the boundary, and retheorizing or relocating the boundary itself in ways that reconnect science and technology to people. While advancing understanding of how people position science and technology in both specialized and everyday worlds, the anthropology of science and technology also calls attention to cultural possibilities whose realization might help expand participation in decision-making about science and technology.

1. *Helping STS Fulfill its Dual Objectives*

The unique contribution of this field to science and technology studies (STS) is that it helps revive and refocus questions regarding relations among science, technology, and people. STS research and researchers are held together by their diverse, yet collective, efforts to trouble and transform the dominant, but simplistic, model or image of science and technology in society. According to the dominant model, researchers live in specialized technical communities whose deliberations are essentially opaque and presumably free of cultural content. Knowledge, in the singular, is created by these bright, well-trained people located inside the academy and then diffuses outside into the public arena through mechanisms of education, popularization, policy, and the benefits of new technologies. Its social significance is evaluated exclusively in the public arena, where knowledge is used, abused, or ignored. The outward travel of knowledge preserves the autonomy of creation and establishes a sharp boundary between science and technology on the one side and people on the other, including the actual lives of scientists and technologists.

STS has sought to engage this model in two ways. First, it brings together researchers who analyze the conceptual and the social dimensions of science and technology simultaneously and in historical perspective. Second, by offering new ways of thinking, STS promises to afford society new pathways for confronting and resolving problems that involve science and technology. STS thus offers the dual trajectories of theory and intervention, of proposing new frameworks of interpretation, and participating critically in societal problem solving. These activities complement one another. Good theory defines pathways that make a difference, and successful acts of critical participation depend upon novel theoretical insights.

During the 1980s and early 1990s, a major focus within STS was on a philosophical debate between 'objectivism' and '(social) constructivism.' Constructivism provided a major theoretical advance over 1970s research on the 'public understanding of science' and the 'impacts' of science and technology, which tended to take for granted the internal contents of science and technology (see *Technology Assessment*). One important body of constructivist work was often labeled 'anthropology of science' because it relied on the direct observation of scientists in laboratories to link scientific practices to knowledge development. For more information on this formative strand in the anthropology of science (see *Laboratory Studies: Historical Perspectives*; *Actor Network Theory*). By questioning how science and technology gain internal contents, the philosophical debate between objectivism and constructivism concentrated attention on the science/technology side of the cultural boundary with people. Also, a focus on the emergence and stabilization of either new scientific knowledge or new technological artifacts tended to maintain the more general cultural separation between science and technology (see *Technology, Social Construction of*).

The newer strands in the anthropology of science and technology explicitly foreground the interventionist potential of STS by moving back and forth across the boundary between science/technology and people, investigating its force and making visible meanings and experiences that get hidden when it is taken for granted. In this work, the commitment to ethnographic practices forces attention to alternative pathways for critical participation as an integral part of theoretical innovation (see also *Ethnography*).

2. *Helping Anthropology Rethink Culture*

As recently as 1988, the American Anthropological Association (AAA) rejected proposed panels on the anthropology of science and technology on the grounds such work did not fit under the AAA umbrella. Things had changed by 1992 as a series of

panels on 'cyborg anthropology' and on the work of Donna Haraway attracted standing-room-only audiences in a ballroom setting.

Key to this shift was a growing recognition that anthropological debates over the status of cultural analysis and cultural critique were similar to developments and deliberations in STS. Anthropologists grew hopeful that analyzing and critiquing the dominant model of science and technology in cultural terms might further people's understanding of how sciences and technologies, including anthropology, live in society. Emerging questions included: what are the implications of accounting for science, technology, and people in cultural terms? Might the finding of wider cultural meanings in science and technology improve searches for alternative configurations? To what extent has anthropology itself depended on the dominant model of science and technology in society? How can one participate critically in science and technology through ethnographic work? As it emerged from symbolic anthropology during the 1970s, the concept of culture drew from the predominant model of language as an underlying grammar, structure, or system of symbols and meanings (see *Culture: Contemporary Views*). Beneath surface differences in speech and action among competent participants in a given community lies a more fundamental sharedness, a bounded 'culture.' This concept of cultures as sets of shared assumptions or presuppositions depends on a contrast with the concept of 'nature.' Where nature provides people with base needs and desires, culture provides content and meaning.

Until the 1980s, it appeared difficult to apply the concept of culture to science and technology. While theorizing culture as a bounded system helped in accounting for differences across cultures, this approach offered no means of accounting for differences within cultures. Yet research questions in STS typically focused on the latter. Also, much research in anthropology itself, for example, kinship theory, depended on the nature/culture distinction for its legitimacy. Treating anthropological work itself as a cultural enterprise could have threatened to undermine the discipline.

As Franklin (1995) put it, by the 1990s 'several trajectories coalesce[d] to produce … momentum' for an emerging anthropology of science and technology. Feminist anthropology had made visible the role of biological assumptions in gender and kinship studies (see *Kinship in Anthropology*). Poststructuralism made it possible to think about power–knowledge relationships and to make the conceptual limitations of the 'human' a focal point for social theory (see *Knowledge, Sociology of*; *Postmodernism in Sociology*). Postmodernist critiques of science and technology called attention to the processes of their production, making 'progress' a contingent effect (see *Cultural Studies of Science*). In revealing how anthropologists 'write culture' by turning people into 'others,' postmodernism in anthropology introduced 'cultural critique' as an anthropological practice (see *Cultural Critique: Anthropological*). Cross-cultural comparisons of knowledge systems began to reposition Western science as ethnoscience (see *Knowledge, Anthropology of*; *Indigenous Knowledge: Science and Technology Studies*; *Postcoloniality*). The rise of interdisciplinary cultural studies juxtaposed 'popular' with 'high' cultural forms, revealing power relations between the two and retheorizing culture as a site of active work (see *Cultural Studies of Science*).

Feminist critiques of science demonstrated its saturation with gender metaphors, and feminist critiques of reproductive technologies provided compelling accounts of women's experiences that could not be counted unproblematically as the benefits of innovation (see *Feminist Epistemology*; *Gender and Technology*). Cross-disciplinary interests in emerging transnational forms demanded simultaneous attention to technology, knowledge, and capital (see *Globalization, Anthropology of*; *Postcoloniality*; *Capitalism: Global*). Growing demands for accountability across the academy in general fueled interest in how ethnographic research can make a difference in the arena under study (see *Ethnography*; *Advocacy in Anthropology*).

During the early 1990s, cultural anthropologists studying science and technology actively resisted the label 'anthropology of science and technology' on the grounds it masked these trajectories, confining diverse work to a bounded subdiscipline. The label works only to mark a collection of intellectual activities within which competition exists not to achieve domination but to make a difference and where border crossing is accepted to enhance the life chances of what the dominant model submerges. Ongoing projects fall into roughly three categories, with individual researchers and studies often contributing to more than one.

3. *Juxtaposing Cultural Systems of Meaning*

The publication of Sharon Traweek's *Beamtimes and Lifetimes* in 1988 marked the emergence of projects that identify and juxtapose cultural systems of meaning in science and technology. Rather than focusing solely on theory change in science, Traweek provides 'an account of how high energy physicists see their own world; how they have forged a research community for themselves, how they turn novices into physicists, and how their community works to produce knowledge' (1988, p. 1). Studying the culture of high-energy physicists extends and transforms the anthropological project of cross-cultural comparison by studying people who live in more than one culture at the same time, in this case those of the international physics community, Japan, and the United States.

Demonstrating sharedness among high-energy physicists and then setting physicists off as a community

achieves two key contributions. It repositions the dominant model of science and technology, which physicists both embrace and embody, into one among many possible cultural perspectives without suggesting that all perspectives on physical knowledge are equal. The ethnographic approach also locates the researcher within the power relations that constitute the field of study and makes visible ways in which scientists function as people who live social lives.

The juxtaposition of shared systems of meaning has proven fruitful in analyzing public controversies over science and technology. Articulating subordinate perspectives and locating these alongside dominant ones demonstrates that factual claims gain meaning in terms of more general frameworks of interpretation. Such work also intervenes in power relations in ways that highlight mediation and collaboration as possible pathways for resolution (see also *Scientific Controversies*).

Representative contributions in this category juxtapose Brazilian spiritists, parapsychologists, and proponents of bacterial theories of cancer with 'orthodox' perspectives (David Hess); antinuclear weapons groups with nuclear weapons scientists (Hugh Gusterson); creation science with evolutionary science (Christopher Toumey); antinuclear power with pronuclear power groups (Gary Downey); artificial intelligence researchers with expert system users (Diana Forsythe); advocates of midwifery with obstetrics (Robbie Davis-Floyd); Marfan scientists with technicians and activists (Deborah Heath); a unified Europe with separate nation–states seeking space travel (Stacia Zabusky); nuclear power plant operators with plant designers (Constance Perin); and competing workgroups in industry (Frank Dubinskas). Recent work questions the culture/people relationship by understanding cultures as, for example, shifting discourses (Gusterson) or recursive, cross-cutting perspectives in a social arena (Hess). Future work will likely review the notion of sharedness, no longer asserting it as a condition of anthropological analysis but exploring what gets accomplished when it can be demonstrated empirically.

4. *Following Flows of Metaphors Across the Boundary*

The publication of Emily Martin's *The Woman in the Body* in 1987 marked the emergence of projects that follow flows of metaphors from general cultural life into science and technology and back again into people's lives and experiences. Writing a cultural history of the present, Martin demonstrates how medical conceptions of menstruation and menopause draw upon metaphors of production in portraying these as breakdown, the failure to produce. In addition to contrasting descriptions from medical textbooks with an ethnographic account of the actual experiences of women, Martin also experiments with re-imagining menstruation and menopause as positive processes of production.

Following flows of metaphors brings a new approach to the study of public understanding and of what the dominant model of science and technology characterizes as impacts. It calls direct attention to the existence and life of the cultural boundary that separates science and technology from people. How do people import scientific facts and technological artifacts into their lives and worlds? How does the dominant model of science and technology live alongside and inform other cultural meanings?

Like the strategy of juxtaposition, following metaphors focuses attention on how people involve themselves with science and technology, making visible those meanings and experiences that do not have a place within the dominant model alongside those that do. What sorts of effects do scientific facts have in people's bodies and lives? How do experiences with technologies go beyond the simply positive, negative, or neutral? How do emerging sciences and technologies contribute to the fashioning of selves? Such work makes visible patterns and forms of difference that may not correlate with demographic categories of race, gender, and class.

Finally, following metaphors introduces the possibility of speculating on alternative cultural possibilities, including narrative experiments with counterfactual scenarios. While sharing with readers an anthropological method of critical thinking, such speculation also forces more explicit attention to how anthropological accounts participate critically within their fields of study. How, for example, might choices of theory and method combine with aspects of researchers' identities to shape pathways of intervention?

Work along these lines follows how sonography and amniocentesis accelerate the introduction of medical expertise into pregnancies (Rayna Rapp); the cultural origins of Western science and its shifts over time (David Hess); how people calculate in diverse settings (Jean Lave); the cultural meanings of medicine in everyday life (Margaret Lock); the narrative construction of scientists' autobiographies (Michael Fischer); creative adjustments when medical expertise fails to account for pregnancy loss (Linda Layne); the flows of meanings that constituted the reproductive sciences as sciences (Adele Clarke); what living with gender assumptions means for women scientists (Margaret Eisenhart); how Bhopal lives as fact and image in different enunciatory communities (Kim Fortun); and how patents travel around the world (Marianne De Laet).

Future work is likely to elaborate questions of scale and identity. How do metaphors gain effects at different scales and how do meanings that live at different scales combine in people's lives and selves? How do people respond to meanings that summon

and challenge them, positioning themselves in searches for identities that work (see also *Identity in Anthropology*).

5. *Retheorizing the Boundary between Science/Technology and People*

The publication of Donna Haraway's *Simians, Cyborgs, and Women* in 1991 marked the emergence of projects that retheorize the boundary itself. Moving through the project of following flows of metaphors to reimagine categories that 'implode' on one another, Haraway calls for 'pleasure in the confusion of boundaries and … responsibility in their construction' (1991, p. 150). Claiming the cyborg (see *Cyborg*) as a feminist icon and key marker of what she calls the 'New World Order,' Haraway forges attention to the contemporary dissolution of boundaries between human and animal, human and machine, and physical and nonphysical. Challenging the 'god-trick' of universalism, she poses 'situated knowledges' (see *Situated Knowledge: Feminist and Science and Technology Studies Perspectives*) as a means of holding simultaneously to a radical historical contingency and a no-nonsense commitment to faithful accounts of a 'real' world.

This project calls attention to burgeoning collections of activities involving science and technology that live across the purported separation between them or across their boundary with people. It includes the exploration of emerging fields that do not fit conventional disciplinary categories, such as biotechnology, biomedicine, and bioengineering, and documents the decline and loss of a distinction between basic and applied science. Following novel activities in research and production motivates the invention of new labels for the anthropological object of study, including, for example, 'technoscience' and 'technoculture.'

The project of retheorizing the boundary between science/technology and people highlights the presence of the nature/culture distinction in anthropology of science and technology as well as in other areas of STS inquiry and intervention. Pressing questions include: through what sorts of processes do analytic findings and interpretations become naturalized as facts in everyday, or popular, modes of theorizing? In what ways might analytic accounts, including claims about culture, depend upon facts from popular theorizing? How might new modes of theorizing about analysis contribute to rethinking the nature/culture distinction itself?

Finally, by calling attention to the difficulty of living within existing categories while attempting to theorize and embody new ones, retheorizing the boundary sharpens the question of intervention. Locating experiences that live across categories invites researchers to examine how emergent categories impact and inflect old ones. Through what sorts of pathways might reformulations actually intervene and achieve change in specific cases? How might it be possible to assess the extent to which reformulations and refigurations prove, in fact, to be helpful, and to whom?

Contributions to this diverse project explore how ideas of the natural help constitute cultural ways of knowing (Marilyn Strathern); the situatedness of practices in machine design and machine use (Lucy Suchman); opportunities for a 'cyborg anthropology' that studies people without starting with the 'human' (Gary Downey, Joseph Dumit, Sarah Williams); how reproductive technologies blur the facts of life (Sarah Franklin); emerging refigurations to constitute and categorize artificial life (Stefan Helmreich); the activities of biotechnology scientists who move outside the academy to gain academic freedom (Paul Rabinow); the importance of 'good hands' and 'mindful bodies' in laboratory work (Deborah Heath); how practices of tissue engineering materialize new life forms (Linda Hogle); experiences with PET scanning that escape the nature/culture distinction (Joseph Dumit); the contemporary production of 'cyborg babies' (Robbie Davis-Floyd and Joseph Dumit); experiences of computer engineers that belie the separation of human and machine (Gary Downey); possibilities for reinvigorating general anthropology by locating technology and humans in the unified frame of cyborg (David Hakken); and how African fractal geometry escapes classification in either cultural or natural terms alone (Ron Eglash) (See also *Actor Network Theory* for an approach that defines both humans and nonhumans as 'actants'.)

Future work is likely to forge novel alliances among theoretical perspectives previously separated by the nature/culture distinction. Actively locating academic theorizing in the midst of popular theorizing will likely force modalities of intervention into focus. Finally, to the extent that the anthropology of science and technology succeeds in challenging and replacing the simplistic dominant model by blurring and refiguring the boundary between science/technology and people, wholly new projects of inquiry and intervention will have to emerge to take account of the changing context. See bibliography for additional reviews.

See also: Actor Network Theory; Common Sense, Anthropology of; Cultural Studies of Science; Culture: Contemporary Views; Ethnography; Gender and Technology; Globalization, Anthropology of; History of Science; History of Technology; Identity in Anthropology; Indigenous Knowledge: Science and Technology Studies; Interpretation in Anthropology; Knowledge, Anthropology of; Science, New Forms of; Science, Sociology of; Scientific Culture; Scientific Disciplines, History of; Symbolism in Anthropology; Technology, Anthropology of

Bibliography

Davis-Floyd R, Dumit J (eds.) 1988 *Cyborg Babies: From Techno-Sex to Techno-Tots*. Routledge, New York

Downey G L, Dumit J (eds.) 1998 *Cyborgs & Citadels: Anthropological Interventions in Emerging Sciences and Technologies*. SAR Press, Santa Fe, NM

Franklin S 1995 Science as culture, cultures of science. *Annual Review of Anthropology* **24**: 163–84

Haraway D 1991 *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York

Heath D, Rabinow P (eds.) 1993 Bio-politics: The anthropology of the new genetics and immunology. *Culture, Medicine, Psychiatry* **17**(special issue)

Hess D J 1995 *Science and Technology in a Multicultural World: The Cultural Politics of Facts and Artifacts*. Columbia University Press, New York

Hess D J, Layne L L (eds.) 1992 The anthropology of science and technology. *Knowledge and Society* **9**

Layne L L (ed.) Anthropological approaches in science and technology studies. *Science, Technology, & Human Values* **23**(1): Winter (special issue)

Martin E 1987 *The Woman in the Body*. Beacon Press, Boston

Nader L (ed.) 1996 *Naked Science: Anthropological Inquiry into Boundaries, Power, and Knowledge*. Routledge, New York

Traweek S 1988 *Beamtimes and Lifetimes: The World of High Energy Physicists*. Harvard University Press, Cambridge, MA

G. L. Downey

Science and Technology: Internationalization

Science is predominantly done in national scientific communities, but scientific information regularly crosses boundaries. Various conditions influence this flow, which means that it is not symmetric in all directions. Internationalization is the uneven process wherein cross-border linkages of communication and collaboration in science and technology among countries multiply and expand. Linkages involve individual scientists and their institutions, but also, increasingly, governments through treaties and conventions that include strong science and technology components. The growing density of interconnections between large industrial corporations regarding research in the precompetitive phase as well as technological alliances is also relevant.

1. Some Distinctions

1.1 Scientific Internationalism

Scientific internationalism emerged with the development of national academies of science (Crawford et al. 1993). Later, correspondence and interchange between individual scientists channeled through international scientific unions and congresses appearing in the nineteenth century (Elzinga and Landström 1996). The flare-up of nationalism and World War I caused a temporary rupture in this internationalism, both in mode of organization and spirit. After the war, transnational scientific relations were gradually repaired, even with Germany, but at the cost of promoting a strict neutralist ideology, nurturing an image of science disembodied from and standing above society.

1.2 Scientific Internationalization and Economic Globalization

The multiplier effect of interconnections between transnational corporations (TNCs) and financial institutions that shape trade related agreements and intellectual property regimes influences, but must not be confused with, scientific internationalization. It is a 'globalization' process in the economic sphere, driven by for-profit motives, facilitated by a combination of neoliberal politics, privatization and information-technological developments. Internationalization of science, by contrast, is ultimately predicated on trust and solidarity among intellectual peers, even though it is interpenetrating increasingly with economic globalization.

A further distinction is between quantitative and qualitative aspects of internationalization. Quantitative studies deal with numbers and patterns of cross-border linkages, and multinational interconnectivity or collaboration and cooperation. They provide indications of changing trends, but for deeper insights need complementation by studies of qualitative changes, e.g., emergence of new institutional arrangements and incentive systems that facilitate international exchange, reorientation of local research agendas, or harmonization of approaches to policy and priorities, like foresight methodologies, at national and regional levels.

Internationalization of industrial research and development (R&D) is now recognized as an important research topic. Statistical surveys and empirical case studies confirm increases in the numbers of multi-country patents held by individual firms, as well as a proliferation of technological alliances within and between a triad of trading blocks (NAFTA; North American Free Trade Agreement, EU; European Union, and Developing Asian Economies plus Japan). In this respect it is actually more appropriate to speak of a 'triadization' than economic globalization.

Pertinent literature on internationalization of technology also refers to other types of interaction between TNCs, and it deals with various R&D management strategies in this context. But in this respect also, extant overviews are limited largely to providing taxonomies and typologies of discernible patterns (*Research Policy* 1999); (for further details, see *National Innovation Systems* and *Technological Innovation*).

Qualitative aspects of internationalization also include epistemic change: i.e., in the intellectual contents of scientific fields, namely, dominant perspectives, methodologies and theories. Thematic integration appears along research fronts, as well as sectoral lines where problems transcend national boundaries (acid rain, global climate change, AIDS), or are too costly for a single nation to handle alone (e.g., CERN). In these instances, and more generally, when international research programs foster coordination of national contributions they also force standardization of data formats, preferred instrumentation and experimental practices.

In what follows, transformations of interconnectivity consonant with internationalism is probed in cartographic, institutional and epistemic dimensions, with political aspects also thrown in relief.

2. *Tracing the Span and Patterns of International Networks*

At the time of writing, the volume of science, national outputs per monies allocated per gross national product (GNP), and the distribution of research efforts and patenting across the globe are mapped regularly. This is done with the help of publication counts (of papers) and reviews of who cites whom (Price 1986) to trace citation patterns and coauthorship linkages (see *Scientometrics*). Visibility of scientists from different countries or in regions is compared. Evaluations of research performance use science and technology indicators as proxies for effectivity, quality and international standing of research groups and their institutions.

The span of global networks increased during the last decades of the twentieth century. Coauthorship patterns reveal growing contacts of scientists across national borders during the 1970s and 1980s (Hicks and Katz 1996), with a certain slackening in the 1990s (UNESCO 1998). Leading scientific producers such as the USA engage less in international co-authorships than do smaller countries—the larger the national or regional scientific community, the greater is 'self-reliance.' Generally, internationalism is only played up in large countries when scientific leadership is in decline, or other countries possess desired specialty knowledge. An interesting anomaly is that India and China also figure lower on polls of cross-border co-authorships than might be expected. Latin America has remained constant, while Africa, especially its sub-Saharan part, stands out as most disadvantaged.

Concentration of resources, prestige, authority and recognition thus display regional variations, with densities following continental contours dominated by an Anglophone region. The predominance of English as the main language of scientific communication, as evidenced in the Science Citation Index (SCI), is also increasing. International coauthorship furthermore reveals subclusters (e.g., the Scandinavian countries) influenced by geographical vicinity, historical traditions, and linguistic as well as cultural affinities. The number of Third World countries now participating in world science has increased, but the vast majority still belong to the scientific periphery (Schott 1998). The share of liberal democracies in world science 1986 was nearly five times their share of world population, while the poorer countries' share in scientific production counted for only one-tenth of their share of the world population. With a few remarkable exceptions (India, Brazil, China) this global gap became even more exaggerated during the 1990s (Schott 1993).

Increased connectivity, scope and participation in scientific communication across national borders, in other words, cannot be equated with decreasing hierarchization. Rather, 'globalization' of institutional models and participation in science is accompanied by a deglobalization in dispersion of science. This contrasts sharply with the notion of science as a public good. Scientific knowledge remains highly concentrated where it is first created, namely among OECD countries. The scientific centers in advanced industrial nations, by virtue of prestige and scientific achievement (often measured by relative density of Nobel laureates in specific fields) also exercise influence over work done in peripheral countries. A Eurocentric skew persists in power, resources, problem selection, and overriding perspectives on the content and function of science and technology.

The end of the Cold War opened a new era of pluricentrism, primarily around three large regional blocks: North America, the European Union, plus Japan and developing economies in Asia. These cleavages are reflected in transnational citation impact and coauthorship clusters, and they coincide with current density patterns in international trade and technology alliances (EC 1997, pp. 10–11, 93).

3. *Driving Factors*

Public spending on R&D is more intense in the USA than in the EU, while in terms of patenting, Japan has caught up with the USA, and the EU rates third (EC 1997a, pp. 53, 93). Economic globalization is a second driving force to reckon with, also uneven. With R&D investments concentrated in a few high-technology industries in the world, these lead strategic reconfigurations in S&T landscapes. Rapid advances in information and communication technologies interlock with economic development, providing new vehicles for rapid interaction (e.g., the Internet), and spurring further alliances and integration of knowledge in pre-competitive phases. Simultaneously, science is being drawn more deeply into the economic globalization process by policy responses, as countries

and whole regions (e.g., the EC through its Framework Programs) facilitate technological development for highly competitive world markets. The EC's mobility schemes for graduate students and post docs must be seen in this light, and so also—by extension—events like the turn of the millennium announcement of a cooperative bilateral university-level agreement between two of the world's most prestigious institutions, MIT (Massachussetts Institute of Technology, USA) and Cambridge University (UK), and the EC's a new integrative policy for 'the European Research Area.' Such events herald a new phase in the intermeshing of the two processes, scientific internationalism and economic globalization.

An ideological driving factor is the traditional cosmopolitan ethos inherent to academe. Associated with modernity, it incorporates the idea of progress, with a history parallel to the emergence of the nation state, democracy and secularization. This means it is in fact culturally bound, as are its CUDOS components enunciated by Robert Merton as the social glue of modern scientific institutions: intellectual Communism, Universalism, Disinterestedness, and Organized Skepticism (see *Norms in Science*). Recent deconstruction of such norms, standards and models of science by scholars in the newer sociology of science highlights the role of particularism, drawing attention to social mechanisms of negotiation between various actors, and how these shape or 'stabilize' scientific consensus that can never be final. This captures the cultural diversity of scientific practices, but not the basic ideological import of universalism as an ideal in scientific internationalism.

A recent factor of internationalization resides in the global nature of environmental problems and the demand for concerted political action on a global scale to address them. Increasing numbers of treaties and conventions with strong science and technology components have been drawn up in this vein. On a converging track one finds pressures of escalating costs of large-scale facilities and calls to share budgetary burdens in newer as well as more traditional fields. The end of the Cold War was also a triggering factor. In its wake appear programs to aid Eastern and Central European scientists in their transition to capitalism and market-driven incentives, and new forms of international science and technology cooperation.

Finally, there are the local pressures in smaller nations that push scientists in settings of relative isolation in all parts of the world to integrate their work more closely with research fronts. Here the benefits of internationalization and the added intellectual stimulation it entails sometimes has the makings of a self-fulfilling prophecy. Pressures to increase local visibility and gain international recognition by exposure to wider peer control of scientific quality get entrenched in procedures and new funding opportunities at national research councils and universities, affecting the behavior of individuals and groups of researchers, and success in getting grants.

4. *Explanatory Models*

Traditional literature on the history of international scientific organizations usually distinguishes two types of fora, scientific non and intergovernmental organizations (scientific NGOs and scientific IGOs). These are taken to represent two different institutional mechanisms for fostering internationalization. In the innovation literature, on the other hand, the focus has mostly been on firms and their role in international diffusion of technologies.

In general, there are two broad strands of theorizing about international organizations. One is rooted in the assumption that networks and more stable forms of organization arise in response to considerations of efficiency and rational goal-oriented behavior, in the course of which less viable alternative forms of interaction are circumvented. An economistic variant of this approach is implicit in an evolutionary theory of technological advance and diffusion. Here 'market' is taken as the selection mechanism that filters out a set of particular technologies from many potentially possible ones, and these get diffused across the globe (Nelson and Winter 1982).

Explicitly or implicitly, the market is regarded as being determined by the tastes, preferences and purchasing power of potential technology users, who are treated as exogenous. Internationalization, in turn, becomes largely a unidirectional product of technological regimes or trajectories which cross and are taken up in socially constituted selective environments. Depending on similarities in the selection mechanisms, or 'learning' between national innovation systems, the transfer of ideas and technologies may be harmonized successively at a global level.

Theoretical assumptions in the historiography of scientific organizations, or alternatively in regime-theory within the study of international relations, run somewhat parallel to this. Instead of the prominence attributed to an economic imperative, in these literatures ideological or political imperatives, and even communities of experts, may be foregrounded. One therefore gets the picture of internationalization primarily as the product of ideologically driven self-direction by the scientific community (in the case of scientific NGOs) or of governments' guiding hands, plus rules and experts (in the case of scientific IGOs).

In all these accounts, the organizations in question are regarded more or less as mechanisms through which other agencies act. They are not depicted as purposive actors with an autonomy, power or culture of their own, even if regime-theory has been criticized for giving too much prominence to experts while

obfuscating the role of the nation states that invest them with authority.

In contrast to this, a sociological strand of theorizing takes its point of departure in Weber's view of bureaucracies and focuses squarely on issues of legitimacy and power. Its advocates seek to explain a much broader range of impacts organizations can have, among others by virtue of their role in constructing actors, interests and social purpose. Emphasized are complexity, multiplicity and flexibility, with actors incorporated as endogenous to the processes of change. In this perspective it becomes interesting to consider how fora for international interaction, once they are created, take on a life of their own, exercising power autonomously in ways unintended and unanticipated by scientific associations or governments at the outset. The same can be said about conventions or regimes introduced to regulate and standardize intercourse between firms internationally in the fields of trade or intellectual property rights; norms are taken to have significant repercussions on the character of the interface between science and industry, and on the use of expertise in other realms.

The constructivist approach associated with sociological institutionalism thus explains the emergence of the relatively autonomous powers of new international fora, in terms of the rational-legal authority they embody, emphasizing the new interests for the parties involved, and the concomitant learning process in which certain organizational models become diffused across the globe. New bureaucracies as they develop are seen to provide a generic cultural form that shapes the various forums in specific ways in their respective domains (firms, scientific NGOs, and scientific IGOs). New actors are seen to be created, responsibilities specified, and authority delineated, defining and binding the roles of both old and new actors, giving them meaning and normative values. In this model, culture, imagery, and rhetoric are held to be forceful ingredients in the life of international organizations, especially in the way these play out their roles in constructing social worlds with a global reach (Finnemore 1993, Barnett and Finnemore 1999).

With the foregoing in mind, the next section highlights some empirical facets, with particular regard to the two most obvious institutional mechanisms (and associated with them, a few typically prominent actors and programs) pertinent to internationalization of—mostly academic, but also governmentally directed—science (as distinct from internationalization of R&D in industrial enterprise—for this see *National Innovation Systems*).

5. *Two Institutional Mechanisms: Scientific NGOs and IGOs*

The numbers and influence of scientific NGOs and IGOs have grown tremendously since the 1980s. Now they not only find themselves interacting, but also pulled in different directions by lobbies of both transnational corporations (TNCs) driven by the profit motive, and nongovernmental civic society organizations (social movement NGOs). The latter are frequently fired by an ethic of equality and justice. Truth, politics, money, and human equality or justice, then, are the four 'logics' that meet in international forums, when strategies to tackle global problems are negotiated, e.g., Rio 1992; in the process new international groups of experts join the scene. This is an important field for future studies (Rayner and Malone 1998).

Analytically, it is useful to draw a distinction between autoletic and heteroletic organizations, especially between ones meant to serve science as an end in itself, and ones that are created and sustained by governmental action (Elzinga and Landström 1996). This is parallel to the distinction between policy for science, and science for policy, serving to mark institutional separation of science from politics, an aspect central to arguments regarding the integrity and objectivity of expert knowledge under strain (see *Scientific Controversies*).

5.1 *Scientific NGOs*

In general, nongovernmental mechanisms operate directly between research communities of different countries, without the intervening medium of governments. They are autoletic, the premise being that communities of scientists are best be left to themselves to organize their transnational contacts for common goals. A unique example is the International Council of Scientific Unions (ICSU), the umbrella organization that in 1931 sprang from an older cosmopolitan ideal (Greenaway 1991). It coordinated the Second International Polar Year (1932–3), the forerunner of a series of programs of global, often multidisciplinary studies that began in 1952 with the plan for the International Geophysical Year (1957). Present-day successors are the International Geosphere-Biosphere Program (IGBP) and the World Climate Research Program (WCRP). These cut across disciplines pertinent to research into global environmental problems. Each has several subprograms dealing with specific themes and aspects of global climate change.

Other major programs cover biodiversity (DIVERSITAS) and the International Human Dimensions of Global Environmental Change program (IHDP), cosponsored with the International Social Science Council (ISSC), ICSU's smaller sister organization for the social sciences, created in 1952. A milestone event is the World Conference on Science in Budapest (ICSU 1999), sponsored jointly with the United Nations Educational, Scientific and Cultural Organization (UNESCO), one of the world's most wide-ranging intergovernmental organizations. Since

its foundation in 1945 UNESCO has worked constantly to link the peripheries to the centers of science.

5.2 Scientific IGOs

Scientific IGOs typify the second (heteroletic) type of mechanism for internationalization. They promote scientific interchange via governmental channels. The point of departure is not scientific knowledge production as such, but the use of science for a particular purpose that form the basis for concerted action. In such contexts science is made a vehicle for the promotion of cultural, economic, political, and other goals at regional and global levels. UNESCO has already been mentioned. The World Meteorological Organization (WMO) and the World Health Organization (WHO) are two of many others within the UN family; in the domain of technology too there are many IGOs, some of which set and regulate standards. The World Bank is another significant actor. A recent addition is the Intergovernmental Panel on Climate Change (IPCC), which follows up on the science produced under the auspices of IGBP and WCRP, harmonizing research-based statements in advice to governments. This has important repercussions for scientific and technological pursuits.

Some IGOs serve stakeholder interests in specific regions of the world, OECD being an example. Since 1963 it has been a pacesetter in developing R&D statistics, and science and technology policy doctrines, contributing to some harmonization between countries. A recent innovation was the creation of the Forum for Megascience (1992), renamed the Global Science Forum. It is responsible for periodical reviews of very large-scale projects so costly (in the order of billions of US dollars per year) and complex that they require multinational collaboration. Thence science and international diplomacy must meet to produce unique management cultures (Watkins 1997). Megaprojects can be concentrated at one site (e.g., CERN) or distributive (e.g., the Human Genome Project). Deep ocean drilling, climate change, thermonuclear fusion experimentation, as well as Antarctic research are further examples. Antarctica is a continent shaped by science as key vehicle in an international regime outside the UN (Elzinga, in Crawford et al. 1993).

6. Intermesh and Thematic Integration

Proliferation and interpenetration of scientific NGOs and IGOs over past decades has gone hand-in-hand with a corresponding growth in numbers of civic NGOs and corporate lobbies. These also interact with scientific bodies, adding to the hybrid criss-cross of connections in which scientific and political agendas converge and blend. Leading scientists respond to the strain by constantly reemphasizing the need to protect the objectivity and integrity of scientific knowledge claims (Greenaway 1991). Conflicts have emerged around attempts to commercialize and privatize national databases, with both ICSU and IGOs, like the WMO, coming out strongly in favor of a policy for open access to information that is invaluable for world research on global problems. Similar tensions have evolved over intellectual property in biotechnology, where scientists are more apt to be of two minds (see *Intellectual Property*, *Concepts of*).

In the wake of globalization, in Third World countries in particular, both scientists and politicians have reacted against the design of Trade Related Intellectual Property Rights (TRIPS) as negotiated within the World Trade Organization (WTO). Around these and other issues, new tensions arise to pull scientists in several directions along the four different 'logics' (delineated at the beginning of Sect. 4).

Thematic integration of research agendas is a second qualitative dimension needing further study. How problems are framed, data, as well as concept formation involves interpretation and epistemological imperatives (see *Situated Knowledge: Feminist and Science and Technology Studies Perspectives*). This is apparent in research on global climate change, where large international scientific programs (IGBP and WCRP) work hand-in-hand with IPCC to orchestrate problem sets, preferred methodologies and modeling criteria. Core sets of concepts spun in the world's leading scientific countries with epistemic communities of climatologists around General Circulation Models (GMCs) have a bearing on the type of data field workers should look for, and the format in which the data are cast; special funding programs exist that enroll scientists from the scientific peripheries.

Creating global consensus around a core of scientifically accepted knowledge and streamlining homogeneous accounts to spur concerted political action have epistemological implications beyond science. As activities, they contribute to the formation of common world views. Concepts such as 'global warming potential' (of greenhouse gases) help to build bridges between science and political decision-making. The very notion of 'global climate,' as well as 'Earth system science' are other examples where conceptual work facilitate cognitive integration, both over disciplinary boundaries in science and in interfaces with citizens at large (Jasanoff and Wynne, in Rayner and Malone 1998). They change our world picture in a reimagining of humankind in its encounters with nature (e.g., ideas like anthropogenic 'fingerprints'). Representations of local climates, the atmosphere, and circulations systems in the oceans are linked up with representations of human ecology (e.g., land use), which in turn are linked to conceptions of risk and truth (see *Risk, Sociology and Politics of*). Internationalization of science resonates with gradual reorientation of perspectives and scientific practices in unitary fashion. To what extent alternatives to expertise as avenues of

knowledge production are foreclosed remains a contentious issue.

See also: History of Science; History of Technology; International Organization; International Science: Organizations and Associations; Science and Technology, Social Study of: Computers and Information Technology; Scientific Academies, History of; Universities and Science and Technology: Europe; Universities and Science and Technology: United States

Bibliography

Barnett M, Finnemore M 1999 The politics, power, and pathologies of international organizations. *International Organization* **53**: 699–732

Crawford C, Shinn T, Sörlin S (eds.) 1993 *Denationalizing Science*. Kluwer, Dordrecht, The Netherlands

European Commission 1997 *Second European Report on S&T Indicators*. Office for Official Publications of the European Communities, Luxembourg.

Elzinga A, Landström C (eds.) 1996 *Internationalism and Science*. Taylor Graham, London

Finnemore M 1993 International organizations as teachers of norms—the United Nations Educational and Cultural Organization and science policy. *International Organization* **47**: 565–97

Greenaway F 1991 *Science International: A History of the International Council of Scientific Unions*. Cambridge University Press, Cambridge, UK

Hicks D, Katz J S 1996 Where is science going? *Science, Technology and Human Values* **21**: 379–406

ICSU 1999 *Science International* [Sept: Special Issue] ICSU Secretariat, Paris

Nelson R, Winter S 1982 *An Evolutionary Theory of Economic Change*. Harvard University Press, Cambridge, MA

Price D de S 1986 *Little Science, Big Science and Beyond*. Columbia University Press, New York

Rayner S, Malone E (eds.) 1998 *Human Choice and Climate Change*. Batelle Press, Columbus, OH, vol. 1

Research Policy 1999 Special Issue **28**, 107–36 (The internationalization of Industrial R&D)

Schott S 1993 World science: Globalization of institutions and participation. *Science, Technology and Human Values* **18**: 196–208

Schott T 1998 Between center and periphery in the scientific world system: Accumulation of rewards, dominance and self-reliance in the center. *Journal of World-systems Research* **4**: 112–44

UNESCO 1998 *World Science Report*. Elsevier, Paris

Watkins J D 1997 Science and technology in foreign affairs. *Science* **277**: 650–1

A. Elzinga

Science and Technology, Social Study of: Computers and Information Technology

Roots and origins being elusive and often illusory, I will date this article from about 1980, when some key collaborations between sociotechnical systems (STS) scholars and maverick computer scientists began. This passes lightly over some early important work critiquing automation, such as that of J. D. Bernal, the early days of the STS movement in Europe, particularly in England, Germany, and Scandinavia, and the neo Marxian analyses of labor process, many of which came to inform later STS work described below.

The nexus of work that currently links STS with computer and information science is a very complex one, with roots in all the areas described below, and held together by a strong, invisible college. This group shares a common concern with how computers shape, and are shaped by, human action, at varying levels of scale. The links with STS include concerns about computer design within the social construction of technology; computers as an agent of social or organizational change; ethics and/or computing (Introna and Nissenbaum 2000); critical studies of computers and computer/information science; applied, activist, and policy research on issues such as the 'digital divide,' or the unequal distribution of computing and information technology across socioeconomic strata and regions of the world.

Some contributions from this part of STS that have been used by scholars in many other parts of the field are Suchman's 'situated action' perspective; Star's 'boundary objects' (Star and Griesemer 1989); Forsythe's methodological questions about 'studying up' and the politics of the anthropology of computing (1993); Henderson's work on engineering drawings as 'conscription devices' (1999); Edwards' work on computing and the Cold War, and its model of 'closed and green' worlds (1996); Berg's critique of rationalization in medicine via computing (1997); Heath and Luff's study of 'centers of control' via the technologies and interactions of the London Underground work force (2000); Woolgar's program building and analytic work—much of it concerning the World Wide Web—from the Virtual Society? Program based at Brunel University (Grint and Woolgar 1997); Bowker's concept of 'infrastructural inversion,' taken from information management at Schlumberger Corporation (1994); and Yates' historical examination of information control techniques in American business (1989, 1996, 1999); Hanseth et al.'s work on standards and ethics in information technology (1996); and Bowker and Star's work on large-scale classification schemes (1999).

1. Automation and the Impact of Computerization

Many early studies of computers and society, or computers and organizations, concerned computing

as an automation of human work. This included two major analytical concerns: the replacement of humans by machines in the workplace, and the 'deskilling' of existing functional jobs (Braverman 1998). Later, the analytic picture became much richer, as researchers realized that automation and deskilling were just two dimensions of the picture. An early breakthrough article by two social-computer scientists defined this territory as 'the web of computing' (Kling and Scacchi 1982), a term which became a touchstone for the next generation. It refers to the co-construction of human work and machine work, in the context of complex organizations and their politics. Much of this research took place in business schools; some in departments of sociology or industrial psychology. In Scandinavia and other regions with strong labor unions, partnerships between the unions and researchers were important, at first focused on job replacement and deskilling, and later, on design created through partnerships with users, social scientists, and computer designers (discussed below).

The social impact of computing remains a strong research strain worldwide, despite the anthropological fact that finding 'untouched tribes' is increasingly difficult. Thus the problematics of interest to STS have shifted from *initial impact* on a group or institution, to understanding *ongoing dynamics* such as usage, local tailoring of systems, shifts in design approaches, understanding the role of infrastructure in helping to form impacts, and the ecology of computers, paper, telephones, fax machines, face-to-face communication, and so forth commonly found in most offices, and increasingly in homes and schools.

2. *The Emulation of Human Cognition and Action (Artificial Intelligence and Robotics)*

The early program of artificial intelligence (AI) work, beginning in an organized way during World War II, was to create a machine that would emulate human thinking (classic AI) and action (robotics). Some of the emulation work came in the form of intelligent tutoring systems and expert systems (Wenger 1987) meant to replace or supplement human decision making. The decision-making side of the research continues to be strong in management schools and in government agencies; it is often critiqued by STS researchers, although some do use these tools (Gilbert and Heath 1985, Byrd et al. 1992). The discussion about whether emulation of human thinking would be possible dates far back; it came to the attention of many STS researchers with the work of philosopher Hubert Dreyfus (1979, 1992), who argued for the irreducibility of human thought and therefore its impossibility in computers. Later STS researchers (including Forsythe, Star, and Suchman) formed direct collaborations with AI researchers. These partnerships took both critical and system-building forms; in both cases, the social science researchers acted as informants about the nature of the 'real world' as opposed to the emulated world of AI. AI changed during the late 1980s and early 1990s—some computer scientists spoke of the 'AI winter,' referring to funding problems unknown in the 1970–88 period. As personal computing and e-mail spread, and later, the Web (1994→), and the early promises of AI seemed not to be paying off, AI began to lose its prestige within computer science. Many AI researchers changed to problems in information science, software engineering, or cognitive science. A branch of AI, distributed artificial intelligence, continued to interact with the STS community (Huhns 1987, Huhns and Gasser 1989). Their interests were in modeling and supporting spatially and temporally distributed work and decision practices, often in applied settings. This reflected and bridged to STS concerns with community problem-solving, communication and translation issues, and the division of labor in large scientific projects.

3. *The Enterprise of Computing, Its Military Roots and the Role of Activism*

In 1983 US President Ronald Reagan introduced his (in)famous 'Star Wars' (officially known as the Strategic Defense Initiative, or SDI) military defense program, a massively expensive attempt to create a distributed laser and missile-based system in space that would protect the USA from foreign attack. The proposal put computer scientists, and especially software engineers, at the center of the design effort.

An immediate outcry was raised by the computing community as well as by alarmed STS scholars close to it. There were concerns that a system of that magnitude was untestable; tools were lacking even to emulate the testing. There were concerns about its viability on other grounds. There were also concerns about its ecological impact on the field of computer science (akin to those raised by organismal biologists about the Human Genome Project)—that the lion's share of research funds would be funneled into this project, orphaning others.

A new group, Computer Scientists for Social Responsibility (CPSR) (www.cpsr.org), was formed in Silicon Valley as a focus for these concerns. The group quickly found common ground in many areas, including activist, ethical, policy and intellectual questions. It has flourished and now sponsors or co-sponsors three conferences a year in which STS scholars often participate: Computers, Freedom and Privacy (policy); Directions in Advanced Computing (DIAC) (intellectual and design directions, and their political and ethical bases and outcomes); and the Participatory Design Conference (PDC) (brings together users, community organizers, computer and social scientists to work on issues of co-design and

appropriate technology) (Kyng and Mathiassen 1997). It, like its counterparts in the UK (Computers and Social Responsibility) and elsewhere, became an important meeting ground between STS and computer/information science. This was especially true of those critical of military sponsorship of computing agendas (at one point in the US some 98 percent of all funding of computer science came from the military, especially ARPA (formerly DARPA)—the Advanced Research Projects of the Army, arguably the critical actor in the creation of the internet (Abbate 1999).

Beginning in the 1980s, a grassroots movement sometimes called 'community computing' arose, and linked with CPSR and other similar organizations. It attempted to increase access to computing for poor people and those disenfranchised from computing for a variety of reasons (Bishop 2000). This often involved the establishment of 'freenets,' or publicly available free computing. Terminals were placed in venues such as public libraries, churches, and after-school clubs; some were distributed directly to people in need. This general problematic has come to be called the 'digital divide,' and forms an active nexus of research across many areas of interest to STS.

An important part of the attempt to enfranchise everyone arose in the world of feminist activism and scholarship, as early computer hackers were nearly all male, and there were sexist and structural barriers to women as both users and professionals. Another meeting place between STS and computer/information science was formed by numerous conferences on women in computer science, gender and computing, and feminist analyses of the problem choice and ethics of computer science. An excellent series of conferences, whose proceedings are published in book form every four years, is sponsored by IFIPS (the International Federation of Information Processing Societies, Working Group 9.1, entitled Women, Work and Computing). See, for example, Grundy et al. (1997). There is a resurgence of interest in this topic with the advent of the Web (see, e.g., Wakeford 1999).

4. *Design*

Beginning in the early 1980s, but reaching full strength in the late 1980s, a number of STS and STS-linked scholars began studying information technology design. This took two forms: studying the work of doing design, and doing design as part of a multidisciplinary team. Henderson's work (cited above) on the visual design practices of engineers is a good example of the former; other important works in the area include Kunda (1992); and a special issue on Design of the journal *Computer Supported Collaborative Work* (Volume 5:4, 1996). STS scholars participating directly in design did so in a number of ways. Anthropologists (such as Nardi 1993, Nardi and O'Day 1999) and Orr (1996) looked at issues such as the culture of programming, its practices, and the role of technicians (Barley and Orr 1997). Some, such as Nardi, conducted usability tests for prototypes of systems. Sociologists examined work practices, organizational processes, and informed the computer scientists about the ways in which these issues could shape or block design (Star and Ruhleder 1996).

5. *From 'The Unnamable' to Social Informatics: Shaping an Invisible College Linked to STS*

An invisible college of social scientists, including STS researchers, who do–study–critique computer and information science has steadily grown since the early 1980s. A brief overview of some of the 'sister travelers' now growing in strength in STS itself includes the following.

5.1 *Organization Theory and Analysis*

People who study organizations have for some time been concerned with the use and impact of computing within them. During the 1990s, STS has become an increasingly important theoretical resource for this field, and has resulted in a number of organization science conferences inviting keynote addresses by STS scholars; reading STS work and attempting to apply it in organizational design and studies. The work of Latour and actor network theory has been especially important (Orlikowski et al. 1995). The policy area of STS has also become increasingly important where issues of privacy, employee rights, intellectual property, and risk are related to computer and information systems (regular updated reports and references can be found on Phil Agre's 'Red Rock Eater' electronic news service: http://dlis.gseis.ucla.edu/people/pagre/rre.html).

5.2 *Computer-supported Cooperative Work (CSCW) and the Participatory Design Movement (PD)*

Cognitive and experimental psychologists have been involved with the design of computer and information systems since at least the 1950s. They formed a field that came to be known as Human–Computer Interaction (HCI). This field focused originally on very small-scale actions, such as measuring keyboard strokes, attention and interface design, ergonomics of workstations, and usability at the level of the individual user. In the late 1980s, part of this field began to stretch out to more organizational and cultural issues (Grudin 1990, Bannon 1990). The impact of the

personal computer and the consequent decentralization of computing, and the rapid spread of networked computing to non-computer specialists were important factors. This wing of HCI began to draw together computer scientists, sociologists, anthropologists, and systems analysts of like mind. In 1986 there was a search to put a name to this. Two of the major competitors were 'office information systems' (to replace the old 'office automation'), and 'computer-supported cooperative work (CSCW),' which won out. CSCW is an interdisciplinary and international group that studies a range of issues including the nature of cooperation, critiques of computer science, building systems to support cooperative work (both local and high distributed). STS scholars have been part of this field since the beginning; Star was a founding co-editor of the journal. Her work and that of Suchman's have been widely used in CSCW.

Closely linked with CSCW is PD, the practice of designing computer systems with user communities and social informaticians. Annual conferences are held in tandem with CSCW. They have picked up and actively use STS approaches. The roots of PD have been reviewed recently by Asarco (2000); they include corporate initiatives, community development, social movements, and the Scandinavian School, discussed in the next section.

5.3 *The 'Scandinavian School'*

In the 1950s the (powerful) trade unions in Scandinavia helped to pass the 'codetermination legislation.' This law stated that unions must be involved in technological design—originally motivated by concerns about deskilling and job loss through automation. A form of sociotechnical systems analysis evolved into a set of techniques for studying work places and processes. As computers arrived in the workplace, this came to include progressive computer scientists and social scientists, many of whom now participate in STS publishing and conferences, as well as CSCW and PD (Greenbaum and Kyng 1991, Bjerknes et al. 1987, Bødker 1991, Neumann and Star 1996).

5.4 *Computers and Education*

The use of computers for science education began in the 1960s. The study of this, based in schools of education and science departments, has had both critical components and basic research. In recent years, the addition of distance education via computers, internet classes, and the ubiquity of computers in college classes has become a critical component of this community. One branch is called Computer-Supported Cooperative Learning (CSCL), and has a lively relationship with both STS and CSCW (Koschmann 1996). STS concepts are beginning to be used across all areas of science education, and recent STS conferences reflect the links strengthening between science education and STS.

5.5 *Former Library Schools as an Emergent Nexus for STS Work*

Since the 1980s, schools of library science have experienced massive closures, due to a complex of corporatization, declining public sphere funding for public libraries, and a general move to close professional schools whose faculty and students are mostly women (social work and physical therapy schools have met similar fates). In the USA, a few of the surviving library schools have reinvented themselves as new 'Schools of Information,' most dropping the word 'library' from the title. Sites include the University of Michigan, University of Illinois at Urbana-Champaign, University of North Carolina, and Indiana University. Their faculty now includes social scientists, computer and information scientists, and library scientists. STS work is central to many of the programs, and several faculty members from the STS world who work with information technology or information itself have found positions there. The structural shape of the programs differs in countries with different forms of funding, and different configurations of the public sphere. Kling's 'Social Informatics' page is an excellent resource (http://www.slis.indiana.edu/SI/), as is Myers' 'Qualitative Research in Information Science' (http://www.auckland.ac.nz/msis/isworld/) (see also Kling 2000). However, STS work is still very influential, for example, at the Royal School of Librarianship in Copenhagen.

6. *The Web: Cultural Studies, Economic Impact, Social Practices, and Ethics*

The public availability of the World Wide Web from 1994 and the commercialization of the Internet–Web, combined with falling prices for computers with good graphics and cheap memory, has changed the face of computing dramatically. Many STS scholars have been involved in different facets of studying the Web. Cultural studies of chat rooms, home pages, MOOs and MUDs, and social inequities in distribution and use have exploded (see e.g., Turkle 1995). Some of this is adapted by STS scholars as studying the use and development of technology; some as a lens through which technology mediates culture. E-commerce, including scientific publishing, has come to overlap with STS work (e.g., Knorr-Cetina and Preda 1998). Ethical areas, such as privacy, electronic stalking and harassment, hate speech sites, and identification of

minors on the Web (which are protected by human subjects regulations, but impossible to identify in many e-venues) are among these issues (Friedman and Nissenbaum 1996).

7. *Challenges and Future Directions*

One of the difficult things for STS scholars who work in this area is the process of tacking back and forth between sympathetic communities housed in information and computer science, STS itself, and one's home discipline. This appears at many levels: job openings, publications, choosing conferences to attend, and the growing call from within industry and computer science for ethnographers, social science evaluators, and co-designers. Sometimes this latter call takes the form of guest talks to technical groups, keynote addresses at technical meetings, consulting, or being asked for free advice about complex social problems within an organization. As with STS itself, there are a growing number of PhD programs within social informatics, broadly speaking, and the field is both spreading and converging. This may, in the long run, ease the juggling problem. At present, most researchers in the area pursue a double career, publishing and participating in both computer/IS and STS communities, with some bridge-building back to the home disciplines.

Another challenge lies in the area of methods. Many social informatics/STS researchers inherited methodological practices from their home disciplines, and some of these make an uneasy fit with current technological directions. For example, there is now a great deal of survey research and of 'ethnography' being done on the Web, including intense study of chat rooms, e-discussion lists, and other forms of online behavior. On the survey side, sampling and validity problems loom large, although they are old problems with venerable methods literatures addressing their solution. The question of sampling only from those with e-mail hookups is similar to the old question of sampling only those who have telephones. Every sample has limits. However, the validity issue, such as the practices of filling out e-mail forms, how the survey appears in the context of many other e-mail messages, and the impact of the genre itself on the content is only now beginning to be explored. Similarly, for ethnographers, it is clear that e-mail messages are not the ready-made fieldnotes that many early studies (i.e., from the early 1990s) delighted in. Forms of triangulation between online and offline research are now coming to the fore in sophisticated methods discussions; another contribution lies in the socially little-known interaction between the built information environment and the phenomenology of users (Wakeford and Lyman 1999).

In conclusion, growing overlaps between complex computer/information science groups, STS, and the other social worlds listed above means new opportunities for STS. STS students are finding employment in corporate research and development (R&D) settings, new information schools, government policy employers concerned about information technology, and in the non-profit sector concerned with issues such as the digital divide.

A note on access: much of this sort of material exists in the proceedings of conferences and is indexed by corporate or organizational authors. It is, in libraries, usually held in special Computer Science libraries on campus, and is, by social science standards, badly indexed. Seek help from the reference librarian to find this material. Working groups or special interest groups (SIGs in computer science terminology) can be powerful loci of change, with thousands of members, both professional and academic. Some are of direct interest to STS scholars, such as SIGSOC, the Special Interest Group on Computers and Society run by the ACM (Association for Computer Machinery, the dominant professional organization for computer scientists in the USA) or ACM- SIGCHI (SIG on Computers and Human Interaction), which has grown so large it is now functionally its own professional organization. Although all professions have jargon and acronyms, computer and information science is highly dense with them. Scholars seeking to build bridges to the computer/information science community can consult many online 'acronym servers' such as Free On-Line Dictionary of Computing at http://foldoc.doc.ic.ac.uk/foldoc/index.html; acronym finder at http://www.acronymfinder.com/.

8. *Selected Journal Resources Publishing Some STS Work*

Accounting, Management and Information Technology (which, in 2000, was due to change its name to *Information and Organization*)
Organization Science
Computer Supported Cooperative Work (CSCW): *The Journal of Collaborative Computing*
Information Systems Research
The Scandinavian Journal of Information Science
CSCL
Learning Sciences
Journal of the American Society of Information Sciences
Human–Computer Interaction
The Information Society
Information Technology and People
CMC Magazine
Electronic Journal on Virtual Culture: http://www.monash.edu.au/journals/ejvc/

See also: Artificial Intelligence in Cognitive Science; Communication: Electronic Networks and Pub-

lications; Communication: Philosophical Aspects; Computers and Society; Digital Computer: Impact on the Social Sciences; Human–Computer Interaction; Human–Computer Interface; Information and Knowledge: Organizational; Information Society; Information Technology; Information Theory; Mass Communication: Technology; Science and Technology: Internationalization; Telecommunications and Information Policy

Bibliography

Abbate J 1999 *Inventing the Internet*. MIT Press, Cambridge, MA

Asarco P M 2000 Transforming society by transforming technology: The science and politics of participatory design. *Accounting, Management and Information Technologies* **10**: 257–90

Bannon L 1990 A pilgrim's progress: From cognitive science to cooperative design. *AI and Society* **4**(2): 59–75

Barley S, Orr J (eds.) 1997 *Between Craft and Science: Technical Work in US Settings*. IRL Press, Ithaca, NY

Berg M 1997 *Rationalizing Medical Work: Decision-support Techniques and Medical Practices*. MIT Press, Cambridge, MA

Bishop A P 2000 Technology literacy in low-income communities. (Re/mediating adolescent literacies). *Journal of Adolescent and Adult Literacy* **43**: 473–76

Bjerknes G, Ehn P, Kyng M (eds.) 1987 *Computers and Democracy: A Scandinavian Challenge*. Avebury, Aldershot, UK

Bødker S 1991 *Through the Interface: A Human Activity Approach to User Interface Design*. Erlbaum, Hillsdale, NJ

Bowker G 1994 Information mythology and infrastructure. In: Bud-Frierman L (ed.) *Information Acumen: The Understanding and Use of Knowledge in Modern Business*. Routledge, London, pp. 231–47

Bowker G, Star S L 1999 *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge, MA

Bowker G, Star S L, Turner W, Gasser L (eds.) 1997 *Social Science, Information Systems and Cooperative Work: Beyond the Great Divide*. Erlbaum, Mahwah, NJ

Braverman H 1998 Labor and monopoly capital: The degradation of work in the twentieth century. 25th Anniversary edn. *Monthly Review Press*, NY

Byrd T A, Cossick K L, Zmud R W 1992 A synthesis of research on requirements analysis and knowledge acquisition techniques. *MIS Quarterly* **16**: 117–39

Dreyfus H L 1979 *What Computers Can't Do: The Limits of Artificial Intelligence*, Revd. edn. Harper & Row, New York

Dreyfus H L 1992 *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press, Cambridge, MA

Edwards P N 1996 *The Closed World: Computers and the Politics of Discourse in Cold War America*. MIT Press, Cambridge, MA

Forsythe D E 1992 Blaming the user in medical informatics: The cultural nature of scientific practice. Knowledge and society. *The Anthropology of Science and Technology* **9**: 95–111

Forsythe D E 1993 The construction of work in artificial intelligence. *Science, Technology, and Human Values* **18**: 460–80

Friedman B, Nissenbaum H 1996 Bias in computer systems. *ACM Transactions on Information Systems* **14**: 330–48

Gilbert N, Heath C (eds.) 1985 *Social Action and Artificial Intelligence*. Gower, Aldershot, UK

Greenbaum J, Kyng M 1991 *Design at Work: Cooperative Design of Computer Systems*. Erlbaum, Hillsdale, NJ

Grint K, Woolgar S 1997 *The Machine at Work: Technology, Work, and Organization*. Polity Press, Cambridge, UK

Grudin J 1990 The computer reaches out: The historical continuity of interface design. In: Chew J, Whiteside J (eds.) *Proceedings of the CHI '90 Conference on Human Factors in Computing Systems*, April 1–5; ACM Press, Seattle, WA, pp. 261–68

Grundy A F et al. (eds.) 1997 Spinning a web from past to future. *Proceedings of the 6th International IFIP Conference*. Bonn, Germany, May 24–27, Springer, Berlin

Heath C, Luff P 2000 *Technology in Action*. Cambridge University Press, New York

Hanseth O, Monteiro E, Hatling M 1996 Developing information infrastructure: The tension between standardization and flexibility. *Science, Technology, and Human Values* **21**: 407–27

Henderson K 1999 *On Line and on Paper: Visual Representations, Visual Culture, and Computer Graphics in Design Engineering*. MIT Press, Cambridge, MA

Huhns M (ed.) 1987 *Distributed Artificial Intelligence*. Morgan Kaufmann, Los Altos, CA

Introna L, Nissenbaum H 2000 The politics of search engines. *IEEE Spectrum* **37**: 26–28

Kling R 2000 Learning about information technologies and social change: The contribution of social informatics. *The Information Society* **16**: 21–232

Kling R, Scacchi W 1982 The web of computing: Computing technology as social organization. *Advances in Computers* **21**: 3–78

Knorr-Cetina K D, Preda A 1998 The epistemization of economic transactions. In: Sales A, Adikhari K (eds.) *Knowledge, Economy and Society*. Sage, New York

Koschmann T (ed.) 1996 *CSCL, Theory and Practice of an Emerging Paradigm*. Erlbaum, Mahwah, NJ

Kunda G 1992 *Engineering Culture: Control and Commitment in a High-tech Corporation*. Temple University Press, Philadelphia, PA

Kyng M, Mathiassen L (eds.) 1997 *Computers and Design in Context*. MIT Press, Cambridge, MA

Luff P, Hindmarsh J, Heath C (eds.) 2000 *Workplace Studies: Recovering Work Practice and Informing System Design*. Cambridge University Press, New York

Nardi B A 1993 *A Small Matter of Programming: Perspectives on End User Computing*. MIT Press, Cambridge, MA

Nardi B A, O'Day V 1999 *Information Ecologies: Using Technology with Heart*. MIT Press, Cambridge, MA

Neumann L, Star S L 1996 Making infrastructure: The dream of a common language. In: Blomberg J, Kensing F, Dykstra-Erickson E (eds.) *Proceedings of PDC '96 (Participatory Design Conference)*. Computer Professionals for Social Responsibility, Palo Alto, CA, pp. 231–40

Orlikowski W, Walsham G, Jones M, DeGross J (eds.) 1995 *Information Technology and Changes in Organizational Work. Proceedings of IFIP WG8.2 Conference, Cambridge, UK*. Chapman and Hall, London

Orr J E 1996 *Talking About Machines: An Ethnography of a Modern Job*. ILR Press, New York

Star S L 1988 The structure of ill-structured solutions: Heterogeneous problem-solving, boundary objects and distri-

buted artificial intelligence. In: Huhns M, Gasser L (eds.) *Distributed Artificial Intelligence* 2. Morgan Kauffmann, Menlo Park, CA, pp. 37–54
Star S L, Griesemer J 1989 Institutional ecology, 'translations,' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939, *Social Studies of Science*, **19**: 387–420. (Reprinted in Mario Biagioli (ed.) *The Science Studies Reader*. Routledge, London, pp. 505–24)
Star S L, Ruhleder K 1996 Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* **7**: 111–34
Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster, New York
Wakeford N, Lyman P 1999 Going into the (virtual) field. *American Behavioral Scientist* **43**:
Wakeford N 1999 Gender and the landscapes of computing at an internet café. In: Crang P, Dey J (eds.) *Virtual Geographies*. Routledge, London
Wenger E 1987 *Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge*. Morgan Kaufmann, Los Altos, CA
Yates J 1989 *Control Through Communication: The Rise of System American Management*. Johns Hopkins University Press, Baltimore, MD
Yates J 1996 Exploring the black box: Technology, economics, and history. *Technology and Culture* **37**: 61–620
Yates J 1999 Accounting for growth: Information systems and the creation of the large corporation. *Journal of Economic History* **59**: 540–2

S. L. Star

Science and Technology Studies: Ethnomethodology

Ethnomethodology is a sociological approach to the study of practical actions which has influenced the development of constructionist, discourse analytic, and related approaches in science and technology studies (S&TS). Early ethnomethodological studies of ordinary activities and social scientific research practices developed an orientation to local practices, situated knowledge, and concrete discourse which later became prominent in science and technology studies. In addition to being a precursor to S&TS, ethnomethodology continues to offer a distinctive approach to practical actions in science and mathematics which rivals more familiar versions of social constructionism.

1. *Ethnomethodological Research Policies*

In the 1960s, Harold Garfinkel coined the term ethnomethodology as a name for a unique sociological approach to practical actions and practical reasoning (Garfinkel 1974, Heritage 1984, p. 45, Lynch 1993, pp. 3–10). Garfinkel was influenced by existential phenomenology (especially Schutz 1962) and sociological theories of action (especially Parsons 1937). As usually defined, ethnomethodology is the investigation of 'folk methods' for producing the innumerable practical and communicative actions which constitute a society's form of life.

1.1 *Ethnomethodological Indifference*

Unlike methodologists in the philosophy of science who accord special status to scientific methods, ethnomethodologists examine methods for composing and coordinating ordinary as well as scientific activities. Garfinkel deemed all methods to be worthy of detailed study. This research policy is known as ethnomethodological indifference (Garfinkel and Sacks 1970, pp. 345–6, Lynch 1993, pp. 141–2). According to this policy, any method is worthy of study, regardless of its professional status, relative importance, adequacy, credibility, value, and necessity. This does not mean that ethnomethodologists treat all methods as equally 'good'; instead, it means that they do not believe that preconceptions about the importance and validity of particular methods should determine a choice of research topic.

Ethnomethodological indifference is similar in some respects to the more familiar policies of symmetry and impartiality in the Strong Programme in the sociology of scientific knowledge (Bloor 1976, pp. 4–5). However, there is a significant difference between the two, which has to do with a less obvious implication of the term 'ethnomethodology.' In addition to being a name for an academic field that studies 'folk methods,' the term refers to 'methodologies'—systematic inquiries and knowledge about methods—which are internal to the practices studied. Many social scientists (including many in S&TS) assume that the persons they study perform their activities unreflexively or even unconsciously, and that the intervention of an outside analyst is necessary for explicating, explaining, and criticizing the tacit epistemologies that underlie such activities. In contrast, ethnomethodological indifference extends to the privileges ascribed to social scientific analysis and criticism:

> Persons doing ethnomethodological studies can 'care' no more or less about professional sociological reasoning than they can 'care' about the practices of legal reasoning, conversational reasoning, divinational reasoning, psychiatric reasoning, and the rest. (Garfinkel and Sacks 1970, p. 142)

Perhaps more obviously than other domains of practice, scientific research involves extensive methodological inquiry and debate. Science is not alone in this respect. In modern (and also many ancient) societies, large bodies of literature articulate, discuss, and debate the practical, epistemic, and ethical character of a broad array of activities, including legal procedures,

food preparation, dining, sexuality, child rearing, gardening, and fly fishing. Any effort to explain such practices sociologically must first come to terms with the fact that explanations of many kinds (including social explanations) are embedded reflexively in the history, production, and teaching of the activities themselves.

1.2 Ethnomethodological Reflexivity

Garfinkel (1967, p. 1) spoke of the 'reflexive' or 'incarnate' character of accounting practices and accounts. By this he meant that social activities include endogenous practices for displaying, observing, recording, and certifying their regular, normative, and 'rational' properties. In other words, social agents do not just act in orderly ways, they examine, record, and reflexively monitor their practices. This particular sense of 'reflexivity' goes beyond the humanistic idea that individuals 'reflect' on their own situations when they act, because it emphasizes collective, and often highly organized, accounting practices. The fact that the persons and organized groups that social scientists study already observe, describe, and interpret their own methodic practices can provoke challenges to the authority of social science methods when the latter conflict with native accounts. Whether or not a social scientist agrees with, for example, official accounts of scientific method given by the subjects of an ethnographic or historical investigation, it is necessary to pay attention to the more pervasive way in which methodic understandings, and understandings of method, play a constitutive role in the practices under study. For ethnomethodologists, the research question is not just 'How do a society's (or scientific field's) members formulate *accounts* of their social world?,' but 'How do members perform actions so as to make them account-*able*; that is, observable and reportable in a public domain of practice?' Accordingly, accounts are not simply interpretations made after actions take place; instead, actions display their accountability for others who are in a position to witness, interpret, and record them.

1.3 Topic and Resource

The idea that explanations, explanatory concepts, and, more generally, natural language resources are common to professional sociology and the activities sociologists study is a source of long-standing consternation and confusion in the social sciences (Winch 1990, [1958]). As Garfinkel and Sacks (1970, p. 337) noted:

> The fact that natural language serves persons doing sociology, laymen or professionals, as circumstances, as topics, and as resources of their inquiries furnishes to the technology of their inquiries and to their practical sociological reasoning *its* circumstances, *its* topics, and *its* resources.

An injunction developed by ethnomethodologists (Zimmerman and Pollner 1970), and adopted by proponents of discourse analysis (Gilbert and Mulkay 1984) is that social analysts should not confuse topic and resource. That is, they should not confuse the common sense explanations given by participants in the social activities studied with the analytic resources for studying those same activities. A variant of this injunction in social studies of science warns the analyst not to adopt, for polemical purposes, the very vocabulary of scientific authority that social studies of science make problematic (Woolgar 1981, Ashmore 1989). While this may be good advice in particular cases, when taken as a general policy the injunction not to confuse topic and resource may seem to encourage a retreat from the very possibility of describing or explaining, let alone criticizing, the practices in question. If every possible analytic resource is already found in the social world as a problematic social phenomenon, then what would a sociologist have to say that is not already incorporated into the practices and disputes studied? While acknowledging that this is a problem for any sociological explanation of science, Latour (1986) recommends a 'semiotic turn' which would examine the terms of the (scientific) tribe from the vantage point provided by an abstract theoretical vocabulary. But, unless one grants special epistemic status to semiotics, this 'turn' simply reiterates the problem.

2. Ethnomethodology and the Problem of Description

Ethnomethodologists do not agree upon a single solution to the problem of reflexivity, and contradictory ways of handling it are evident in the field, but some ethnomethodologists believe that sociological descriptions do not require special theoretical or analytical auspices in order to overcome reflexivity. As Sharrock and Anderson (1991, p. 51) argue, under most circumstances in which descriptions are made and accepted as adequate, there is no need to overcome epistemological scepticism. Ethnomethodologists treat reflexivity as ubiquitous, unavoidable, and thus 'irremediable.' Consequently, reflexivity is not a 'methodological horror' (Woolgar 1988, p. 32) that makes description or explanation impossible or essentially problematic, because an ethnomethodologist is in no worse (or better) shape than anyone else who aims to write intelligible, cogent, and insightful descriptions of particular states of affairs. Even so, there still remains the question of the scientific, or other, grounds of ethnomethodological descriptions.

2.1 The Possibility of 'Scientific' Descriptions of Human Behavior

A possible basis for ethnomethodological descriptions is presented in an early argument written by Sacks (1992, pp. 802–5), the founder of a field that later came to be called conversation analysis. This was a brief, but intriguing, argument about the possibility of developing a science that produces stable, reproducible, naturalistic accounts of human behavior. Sacks observed that natural scientists already produce accounts of human behavior when they write natural language descriptions of methods for reproducing observations and experiments. Social scientists who attempt to describe 'methods' in a broad range of professional and non-professional activities are in no worse position than natural scientists who attempt to describe their own particular methods. In other words, Sacks conceived of the replication of experiments as a particular (but not necessarily special) case of the reproduction of social structures. While social scientists do not aim to produce 'how to' manuals, they can write accounts of social actions which are 'true' in a praxiological sense: adequate to the production and reproduction of the relevant practices.

2.2 Replication and the Reproduction of Social Structure

It is well established in the sociology of science that replication is problematic. Instead of being a methodological foundation for establishing natural facts, particular instances of experimental replication often beg further questions about the detailed conditions and competencies they involve (Collins 1985). To say that replication is problematic does not imply that it is impossible, but it does raise the question of how scientists manage to secure assent to the adequacy of their experiments. The phenomenon of just how scientists conduct experiments, and how they establish particular results in the relevant communities, has become a major topic for socio-historians and ethnographers of science (Gooding et al. 1989). Consequently, in the more general case of the reproduction of social structure, it seems reasonable to conclude that Sacks identifies a topic for ethnomethodological research, but not a grounding for a possible sociological research program (Lynch and Bogen 1994).

Ethnomethodological studies of the reproduction of order in science, other professions, and daily life, address the topic of the production of instructed actions. Instructed actions include an open-ended variety of actions performed in accordance with rules, plans, recipes, methods, programs, guidelines, maps, models, sets of instructions, and other formal structures. In addition to examining efforts to reproduce initial observations, ethnomethodologists have studied a series of topics: deriving mathematical proofs, following instructions for photocopying, and reproducing standard laboratory protocols (Garfinkel 1986, 1991, Garfinkel et al. 1981, Livingston 1986, Suchman 1987).

2.3 Ethnomethodology and Social Constructionism

Ethnomethodology has an ambivalent relation to social constructionism. Early laboratory studies (Latour and Woolgar 1979, Knorr 1981) integrated selected themes from ethnomethodology into constructionist arguments, but many ethnomethodologists prefer to speak of the 'production' rather than the 'construction' of social orders (Lynch 1993, Button and Sharrock 1993). While it may be so that no fact is ever free of 'construction' in the broadest sense of the word, local uses of the term in empirical science (though not in mathematics and certain branches of theory) refer to research artifacts which are distinguished from a field of natural objects (Lynch 1985). Ethnomethodologists prefer not to speak indiscriminately of the construction or manufacture of knowledge, in order to preserve an 'indifferent' orientation to the way distinctions between constructed and unconstructed realities are employed in practical action and situated argument.

2.4 Normative and Ethical Considerations

Ethnomethodology has been criticized for treating normative aspects of the activities studied as 'mere phenomena' (Habermas 1984, p. 106). For example, when studying a contentious court case (Goodwin 1994) or jury deliberation (Maynard and Manzo 1993), an ethnomethodologist does not assess the discursive practices by reference to ideal standards of validity, rationality, or justice. This does not suppose that normative considerations are 'mere' phenomena. Instead, it supposes that peoples' methods (the phenomena studied by ethnomethodologists) are just that: situated actions that, for better or worse, incorporate normative judgments and ethical claims.

Instead of recasting such judgments and claims in terms of one or another transcendent framework, ethnomethodologists attempt to explicate the way normative judgments are produced, addressed, and fought over in specific circumstances. When ethnomethodologists investigate highly charged uses of, and appeals to, normative judgment, they enable readers to examine specific configurations of action and reasoning that do not neatly fall under recipe versions of norms and values. Consequently, rather than invoking or developing a normative theory, ethnomethodologists invite their readers to consider intricate situations of practice and contestation that no single general framework can possibly forecast or resolve.

Practical and ethical dilemmas are no less salient for ethnomethodologists than for cultural anthropologists,

lawyers, and other participant-investigators. The policy of ethnomethodological indifference does not relieve an investigator of ethical choices and responsibilities. Like other investigators, ethnomethodologists may in some circumstances find it advisable to respect norms of privacy and propriety, while in others they may feel compelled to expose wrongdoing. However, as a body of doctrines, research policies, and exemplary studies, ethnomethodology does not supply a set of rules or ethical guidelines for making such difficult choices. This does not mean that ethnomethodologists proceed without ethics, but that their ethical judgments, like many of their other judgments, have a basis in communal life that is not encapsulated by any academic school, theory, or method.

See also: Ethnology; Ethnomethodology: General; Parsons, Talcott (1902–79); Reflexivity in Anthropology; Reflexivity: Method and Evidence

Bibliography

Ashmore M 1989 *The Reflexive Thesis: Wrighting the Sociology of Scientific Knowledge*. University of Chicago Press, Chicago

Bloor D 1976 *Knowledge and Social Imagery*. Routledge and Kegan Paul, London

Button G (ed.) 1991 *Ethnomethodology and the Human Sciences*. Cambridge University Press, Cambridge, UK

Button G, Sharrock W 1993 A disagreement over agreement and consensus in constructionist sociology. *Journal of Theory Social Behaviour* **23**: 1–25

Collins H M 1985 *Changing Order: Replication and Induction in Scientific Practice*. Sage, London

Garfinkel H 1967 *Studies in Ethnomethodology*. Prentice Hall, Englewood Cliffs, NJ

Garfinkel H 1974 On the origins of the term 'ethnomethodology.' In: Turner R (ed.) *Ethnomethodology*. Penguin, Harmondsworth, UK

Garfinkel H (ed.) 1986 *Ethnomethodological Studies of Work*. Routledge and Kegan Paul, London

Garfinkel H 1991 Respecification: Evidence for locally produced, naturally accountable phenomena of order, logic, reason, meaning, method, etc. in and as of the essential haecceity of immortal ordinary society (I)—an announcement of studies. In: Button G (ed.) *Ethnomethodology and the Human Sciences*. Cambridge University Press, Cambridge, UK

Garfinkel H, Lynch M, Livingston E 1981 The work of a discovering science construed with materials from the optically discovered pulsar. *Philosophy of the Social Sciences* **11**: 131–58

Garfinkel H, Sacks H 1970 On formal structures of practical actions. In: McKinney J C, Tiryakian E A (eds.) *Theoretical Sociology: Perspectives and Development*. Appleton-Century-Crofts, New York

Gilbert G N, Mulkay M 1984 *Opening Pandora's Box: An Analysis of Scientists' Discourse*. Cambridge University Press, Cambridge, UK

Gooding D, Pinch T, Schaffer S (eds.) 1989 *The Uses of Experiment*. Cambridge University Press, Cambridge, UK

Goodwin C 1994 Professional vision. *American Anthropologist* **96**: 606–33

Habermas J 1984 *The Theory of Communicative Action. Volume I: Reason and the Rationalization of Society*. Beacon Press, Boston

Heritage J 1984 *Garfinkel and Ethnomethodology*. Polity Press, Oxford, UK

Knorr K 1981 *The Manufacture of Knowledge*. Pergamon, Oxford, UK

Latour B 1986 Will the last person to leave the social studies of science please turn on the tape recorder. *Social Studies of Science* **16**: 541–8

Latour B, Woolgar S 1979 *Laboratory Life: The Social Construction of Facts*. Sage, London

Livingston E 1986 *The Ethnomethodological Foundations of Mathematics*. Routledge and Kegan Paul, London

Lynch M 1985 *Art and Artifact in Laboratory Science*. Routledge and Kegan Paul, London

Lynch M 1993 *Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science*. Cambridge University Press, New York

Lynch M, Bogen D 1994 Harvey Sacks's primitive natural science. *Theory, Culture and Society* **11**: 65–104

Maynard D, Manzo J 1993 On the sociology of justice: Theoretical notes from an actual jury deliberation. *Sociological Theory* **11**: 171–93

Parsons T 1937 *The Structure of Social Action*. Free Press, New York, Vols. I & II

Sacks H 1992 *Lectures on Conversation*. Blackwell, Oxford, UK, Vol. I

Schutz A 1962 *Collected Papers*. Martinus Nijhoff, The Hague, Vol. I

Sharrock W, Anderson B 1991 Epistemology: Professional scepticism. In: Button G (ed.) *Ethnomethodology and the Human Sciences*. Cambridge University Press, Cambridge, UK

Suchman L 1987 *Plans and Situated Actions*. Cambridge University Press, Cambridge, UK

Winch P 1990[1958] *The Idea of a Social Science and its Relation to Philosophy*, 2nd edn. Humanities Press, Atlantic Highlands, NJ

Woolgar S 1981 Interests and explanations in the social study of science. *Social Studies of Science* **11**: 365–94

Woolgar S 1998 *Science: The Very Idea*. Tavistock, London

Zimmerman D H, Pollner M 1970 The everyday world as a phenomenon. In: Douglas J D (ed.) *Understanding Everyday Life: Toward the Reconstruction of Sociological Knowledge*. Aldine, Chicago

M. Lynch

Science and Technology Studies: Experts and Expertise

1. Introduction

Science is a social activity to construct, justify, and critique cognitive claims based on widely accepted methodologies and theories within relevant communi-

ties. The term science is used here to include claims of true statements about any phenomenon, be it natural, anthropological, social, or even metaphysical. Methods or procedures are clearly distinct among the various domains of science, but the main assertion here is that the overall goal of claiming truthful statements remains the essence of scientific inquiry throughout all disciplines (similar attempts in Campbell 1921, pp. 27–30). Using scientific expertise, however, is not identical with generating scientific statements (Lindblom and Cohen 1979, p. 7ff.).

In a policy arena, scientific experts are expected to use their skills and knowledge as a means of producing arguments and insights for identifying, selecting, and evaluating different courses of collective action. Since such advice includes the prediction of likely consequences of political actions in the future, experts are also in demand to give advice on how to cope with uncertain events and how to make a prudent selection among policy options, even if the policy-maker faces uncertain outcomes and heterogeneous preferences (Cadiou 2001, p. 27). Many policy-makers expect scientific experts to help construct strategies that promise to prevent or mitigate the negative and promote the positive impacts of collective actions. In addition, scientific expertise is demanded as an important input to design and facilitate communication among the different stakeholders in debates about technology and risk.

Based on these expectations, scientific expertise can assist policy-makers to meet five major functions (similar in Renn 1995):

(a) providing factual insights that help policy-makers to identify and frame problems and to understand the situation (*enlightenment function*);

(b) providing instrumental knowledge that allows policy-makers to assess and evaluate the likely consequences of each policy option (*pragmatic or instrumental function*);

(c) providing arguments, associations, and contextual knowledge that helps policy-makers to reflect on their situation and to improve and sharpen their judgment (*reflexive function*);

(d) providing procedural knowledge that helps policy-makers to design and implement procedures for conflict resolution and rational decision making (*catalytic function*); and

(e) providing guidelines or designing policy options that assist decision-makers in their effort to communicate with the various target audiences (*communicative function*).

These five functions touch on crucial aspects of policy-makers' needs. First, insights offered by experts help policy-makers to understand the issues and constraints of different policy options when designing and articulating policies. Policy-makers need background information to develop standards, to ground economic or environmental policies on factual knowledge, and to provide information about the success or failure of policies. Second, scientific methods and their applications are needed to construct instrumental knowledge in the format of 'if–then' statements and empirically tested theories; this knowledge leads to the articulation of means-ends oriented policies and problem solving activities. Third, scientific reasoning and understanding help policy-makers to reflect on their activities and to acknowledge social, cultural, institutional, and psychological constraints as well as opportunities that are not easily grasped by common sense or instrumental reasoning. However, scientific statements may also restrict policy-makers as they are directed towards adopting a single perspective in analyzing and framing a problem. Fourth, policy-makers may use scientists to design procedures of policy formulation and decision making in accordance with normative rules of reasoning and fairness. These procedures should not interfere with the preferences of those who are involved in the decision-making process, but provide tools for making these preferences the guiding principle of policy selection. To meet this function, scientists need to play a role similar to a chemical catalyst by speeding up (or if necessary slowing down) a process of building consensus among those who are entitled to participate in the policy-making process (Fishkin 1991). Lastly, scientific experts can help to design appropriate communication programs for the purpose of legitimizing public policies as well as preparing target audiences for their specific function or role in the task of risk management.

This article focuses predominantly on the influence of scientific and technical expertise, in particular the results of technology assessments, on public policy-making. The second section deals with the risks and challenges of technical experts providing input to policy design and implementation. The third section addresses the influence of systematic knowledge for policy-making. The fourth section provides some theoretical background for the role of expertise in deliberative processes. The fifth section focuses on cultural differences in the use of expertise for policy-making. The last section summarizes the main points of this article.

2. *Using Scientific Expertise for Policy-making: Risks and Challenges*

The interaction between experts and policy-makers is a major issue in technology management today and is likely to become even more important in the future. This is due in the first instance to the increased interactions between human interventions and natural responses and, secondarily, to the increased complexity of the necessary knowledge for coping with economic, social, and environmental problems. Population growth and migration, global trade, inter-

national market structures, transboundary pollution, and many other conditions of modern life have increased the sensitivity to external disturbances and diminished the capability of social and natural systems to tolerate even small interventions. Although contested by some (Simon 1992), most analysts agree that ecological systems have become more vulnerable as the impact of human intervention has reached and exceeded thresholds of self-repair (Vitousek et al. 1986).

Given this critical situation, what are the potential contributions of expertise to the policy process? In principle, experts can provide knowledge that can help to meet the five functions mentioned above and to anticipate potential risks before they materialize. But they can do this only to the degree that the state of the art in the respective field of knowledge can provide reliable information pertaining to the policy options. Many policy-makers share assumptions about expertise that turn out to be wishful thinking or illusions (Funtowicz and Ravetz 1990, Jasanoff 1990, 1991, Rip 1992, Beck 1992). Most prominent among these are:

(a) *illusion of certainty:* making policy-makers more confident about knowing the future than is justified;

(b) *illusion of transferability:* making policy-makers overconfident that certainty in one aspect of the problem applies to all other aspects as well;

(c) *illusion of 'absolute' truth:* making policy-makers overconfident with respect to the truthfulness of evidence;

(d) *illusion of ubiquitous applicability:* making policy-makers overconfident in generalizing results from one context to another.

These illusions are often reinforced by the experts themselves. Many experts feel honored to be asked by powerful agents of society for advice. Acting under the expectation of providing unbiased, comprehensive, and unambiguous advice, they often fall prey to the temptation to oversell their expertise and provide recommendations far beyond their realm of knowledge. This overconfidence in one's own expertise gains further momentum if policy-maker and advisor share similar values or political orientations. As a result policy-makers and consultants are prone to cultivate these illusions and act upon them.

In addition to these four types of illusions, experts and policy-makers tend to overemphasize the role of systematic knowledge in making decisions. As much as political instinct and common sense are poor guides for decision making without scientific expertise, the belief that scientific knowledge is sufficient to select the correct option is just as short sighted. Most policy questions involve both systematic as well as anecdotal and idiosyncratic knowledge (Wynne 1989). Systematic knowledge often provides little insight into designing policies for concrete issues. For example, planning highways, supporting special industries, promoting health care for a community and many other issues demand local knowledge on the social context and the specific history of the issue within this context (Wynne 1992, Jasanoff 1991). Knowledge based on local perspectives can be provided only by those actors who share common experiences of the issue in question. The role of systematic versus particularistic knowledge is discussed in more detajobil in the next section.

3. *The Relevance of Systematic Expertise for Policy-making*

There is little debate in the literature that the inclusion of expertise is essential as a major resource for designing and legitimizing technological policies (Jasanoff 1990). A major debate has evolved, however, on the status of scientific and technical expertise for representing all or most of the knowledge that is relevant to these policies. This debate includes two related controversies: the first deals with the problem of objectivity and realism; the second one with the role of subjective and experiential knowledge that non-experts have accumulated over time. This is not the place to review these two controversies in detail (see Bradbury 1989, Shrader-Frechette 1991). Depending on which side one stands on in this debate, scientific evidence is either regarded as one input to fact-finding among others or as the central or even only legitimate input for providing and resolving knowledge claims. There is agreement, however, among all camps in this debate that systematic knowledge is instrumental for understanding phenomena and resolving problems. Most analysts also agree that systematic knowledge should be generated and evaluated according to the established rules or conventions of the respective discipline (Jaeger 1998, p. 145). Methodological rigor aiming to accomplish a high degree of validity, reliability and relevance remains the most important yardstick for judging the quality of scientific insights. Constructivist scholars in science and technology studies do not question the importance of methodological rules in securing credible knowledge but are skeptical whether the results of scientific inquiries represent objective or unambiguous descriptions of reality (Latour and Woolgar 1979, Knorr-Cetina 1981). Rather, they see scientific results as products of specific processes or routines that an elite group of knowledge producers has framed as 'objective' and 'real.' The 'reality' of these products is determined by the availability of research routines and instruments, prior knowledge and judgments, and social interests (see also Beck 1992, although he regards himself as a moderate realist).

For the analysis of scientific input to policy-making, the divide between the constructivists and the realists matters only in the degree to which scientific input is used as a genuine knowledge base or as a final arbiter for reconciling knowledge conflicts. A knowledge

discourse deals with different, sometimes competing claims that obtain validity only through compatibility checks with acknowledged procedures of data collection and interpretation, proof of theoretical compatibility and conclusiveness, and the provision of intersubjective opportunities for reproduction (Shrader-Frechette 1991, pp. 46ff.). Obviously many research results do not reach the maturity of proven facts, but even intermediary products of knowledge, ranging from plain hypotheses, via plausible deductions to empirically proven relationships, strive for further perfection (cf. the pedigree scheme of Functowicz and Ravetz 1990). On the other hand, even the most ardent proponent of a realist perspective will admit that only intermediary types of knowledge are often available when it comes to assess and evaluate risks.

What does this mean for the status and function of scientific expertise in policy contexts? First, scientific input has become a major element of technological decision-making in all technologically developed countries. The degree to which the results of scientific inquiry are taken as ultimate evidence to judge the appropriateness and validity of competing knowledge claims is contested in the literature and also contested among policy-makers and different social groups. Frequently, the status of scientific evidence becomes one of the discussion points during social or political deliberation depending on the context and the maturity of scientific knowledge in the technological policy arena under question. For example, if the issue is the effect of a specific toxic substance on human health, subjective experience may serve as a heuristic tool for further inquiry and may call attention to deficits in existing knowledge, although toxicological and epidemiological investigations are unlikely to be replaced with intuitions from the general public. If the issue, by contrast, is siting of an incinerator, local knowledge about sensitive ecosystems or traffic flows may be more relevant than systematic knowledge about these impacts in general (a good example for the relevance of such particularistic knowledge can be found in Wynne 1989). Second, the resolution of competing claims of scientific knowledge usually are governed by the established rules within the relevant disciplines. These rules may not be perfect and even contested within the community. Yet they are regarded as superior to any other alternative (Shrader-Frechette 1991, pp. 190ff.). Third, many technological decision options require systematic knowledge that is either not available or still in its infancy or in an intermediary status. Analytic procedures are then demanded by policy-makers as a means to assess the relative validity of each of the intermediary knowledge claims, to display their underlying assumptions and problems, and to demarcate the limits of 'reasonable' claims, that is, to identify the range of those claims that are still compatible with the state of the art in this knowledge domain (Shrader-Frechette 1991). Fourth, knowledge claims can be systematic and scientific as well as idiosyncratic and anecdotal. Both forms of knowledge have a legitimate place in technological decision-making. How they are used depends on the context and the type of knowledge required for the issue in question (Wynne 1992). All four points show the importance of knowledge for technology policy and decision making, but also make clear that choosing the right management options requires more than looking at the scientific evidence alone.

4. *Scientific Evidence in Deliberative Processes*

Given the delicate balance between anecdotal, experiential, and systematic knowledge claims, policy-making depends on deliberative processes in which competing claims are generated or ignored, sorted, selected, and highlighted. The first objective is to define the relevance of different knowledge claims for making legitimate and defendable choices. The second objective is to cope with issues of uncertainty and to assign trade-offs between those who will benefit and those who will suffer from different policy options. The third objective is to take into account the wider concerns of the affected groups and the public at large. These three elets of deliberation are related to coping with the problems of complexity, uncertainty, and ambiguity.

How can deliberative processes deal with the problems of complexity, uncertainty, and ambiguity? To respond to this question, it is necessary to introduce the different theoretical concepts underlying deliberative processes. The potential of deliberation has been discussed primarily in three schools of thought:

(a) The utility-based theory of rational action (basics in Fisher and Ury 1981, Raiffa 1994; review of pros and cons in Friedman 1995): in this concept, deliberation is framed as a process of finding one or more option(s) that optimize the payoffs to each participating stakeholder. The objective is to convert positions into statements of underlying interests. If all participants articulate their interest, it is either possible to find a new win-win option that is in the interest of all or at least does not violate anybody's interest (Pareto optimal solution) or to find a compensation that the winner pays to the losers to the effect that both sides are at least indifferent between the situation without the preferred policy option and no compensation and the implementation of this option plus compensation (Kaldor–Hicks solution). In this context systematic knowledge is required to inform all participants of the likely consequences of each decision option. The evaluation of desirability of each option is a matter of individual preferences and in this perspective outside of the deliberation process.

(b) Theory of communicative action (Habermas 1987, Webler 1995): this concept focuses on the communicative process of generating preferences,

values, and normative standards. Normative standards are those prescriptions that do not apply only to the participants of the discourse but also to society as a whole or at least a large segment of the external population. Normative standards in technological arenas include, for example, exposure limits or performance standards for technologies. They apply to all potential emitters or users regardless whether they were represented at the discourse table or not. The objective here is to find consensus among moral agents (not just utility maximizers) about shared meaning of actions based on the knowledge about consequences and an agreement on basic human values and moral standards. Systematic knowledge in this context helps the participants to provide insights into the potential effects of collective decision options and help them to reorganize their preferences according to mutually desirable outcomes.

(c) Theory of social systems (Luhmann 1986, Eder 1992): the (neo)functional school of sociology pursues a different approach to deliberation. It is based on the assumption that each stakeholder group has a separate reservoir of knowledge claims, values, and interpretative frames. Each group-specific reservoir is incompatible with the reservoir of the other groups. This implies that deliberative actions do not resolve anything. They represent autistic self-expressions of stakeholders. In its cynical and deconstructivist version, deliberation serves as an empty but important ritual to give all actors the illusion of taking part in the decision process. In its constructive version deliberation leads to the enlightenment of decision-makers and participants. Far from resolving or even reconciling conflicts, deliberation in this viewpoint has the potential to decrease the pressure of conflict, provide a platform for making and challenging claims, and assist policy-makers in making them cognizant of different interpretative frames (Luhmann 1993). Deliberations help to reframe the decision context, to make policy-makers aware of public demands, and enhance legitimacy of collective decisions through reliance on formal procedures (Skillington 1997). In this understanding of deliberation, reaching a consensual conclusion is neither necessary nor desirable. The process of talking to each other, exchanging arguments, and widening one's horizon is all what deliberation is able to accomplish. It is an experience of mutual learning without a substantive message. Systematic knowledge in this context is never free of context and prescriptive assumptions. Hence, each group will make knowledge claims according to its interests and strategic goals. Integration of knowledge is based on rhetoric, persuasion skills, and power rather than established rules of 'discovering the truth.'

These three understandings of deliberation are not mutually exclusive although many proponents of each school argue otherwise. Based on the previous arguments, it is quite obvious that the rational actor approach provides a theoretical framework to understand how actors in deliberative processes deal with complexity and partially uncertainty. The communicative action approach provides a theoretical structure for understanding and organizing discourses on ambiguities and moral positions. In particular, this concept can highlight those elements of deliberation that help participants to deal competently with moral and normative issues beyond personal interests. The system–analytic school introduces some skepticism towards the claim of the other schools with respect to the outcomes of deliberation. Instead it emphasizes the importance of procedures, routines, and learning experiences for creating links or networks between the major systems of society. Deliberation is the lubricant that helps each of the collective social actors to move mostly independently in society without bumping into the domains of the other actors.

Deliberative processes aimed at integrating experts, stakeholders, policy-makers, and the public at large can be organized in many different forms. Practical experiences have been made with advisory committees, citizen panels, public forums, consensus conferences, formal hearings, and others (see Rowe and Frewer 2000). A hybrid model of citizen participation (Renn et al. 1993) has been applied to studies on energy policies and waste disposal issues in West Germany, for waste-disposal facilities in Switzerland, and to sludge-disposal strategies in the United States (Renn 1999).

5. *Cultural Styles in Using Scientific Expertise*

The way that knowledge and expertise are included in policy processes depends on many factors. Comparative research on the influence of systematic knowledge in policy processes emphasizes the importance of cultural context and historic developments (Solingen 1993). In addition, state structures and institutional arrangements significantly influence the type of inclusion of expertise into the decision-making processes. There has been a major shift in modern states to an organized and institutionalized exchange of science organizations with policy-making bodies (Mukerji 1989, Jasanoff 1990). Although science has become a universal enterprise, the specific meaning of what science can offer to policy-makers differs among cultures and nations (Solingen 1993). The situation is even more diverse when one investigates the use of science in different countries for advising policy-makers. Scientific and political organizations partially determine what aspects of life are framed as questions of knowledge and what of "subjective" values. In addition, national culture, political traditions, and social norms influence the mechanisms and institutions for integrating expertise in the policy arenas (Wynne 1992).

In one line of work, policy scholars have developed a classification of governmental styles that highlight

four different approaches to integrating expert knowledge into public decisions (Brickman et al. 1985, Jasanoff 1986, O'Riordan and Wynne 1987, Renn 1995). These styles have been labeled inconsistently in the literature, but they refer to common procedures in different nations. The 'adversarial' approach is characterized by an open forum in which different actors compete for social and political influence in the respective policy arena. The actors in such an arena need and use scientific evidence to support their position. Policy-makers pay specific attention to formal proofs of evidence because policy decisions can be challenged on the basis of insufficient use or neglect of scientific knowledge. Scientific advisory boards play an important role as they help policy-makers to evaluate competing claims of evidence and to justify the final policy selection (Jasanoff 1990).

A sharp contrast to the adversarial approach is provided by the fiduciary style (Renn 1995). The decision-making process is confined to a group of patrons who are obliged to make the 'common good' the guiding principle of their action. Public scrutiny or involvement is alien to this approach. The public can provide input to and arguments for the patrons but is not allowed to be part of the negotiation or policy formulation process. Scientists outside the policy-making circles are used as consultants at the discretion of the patrons and are selected according to prestige or personal affiliations. Their role is to provide enlightenment and information. Patrons' staff generate instrumental knowledge. This system relies on producing faith in the competence and the fairness of the patrons involved in the decision-making process.

Two additional styles are similar in their structure but not identical. The consensual approach is based on a closed circle of influential actors who negotiate behind closed doors. Representatives of important social organizations or groups and scientists work together to reach a predefined goal. Controversy is not visible and conflicts are often reconciled before formal negotiations take place. The goal of the negotiation is to combine the best available evidence with the various social interests that the different actors represent. The corporatist style is similar to the consensual approach, but is far more formalized. Well-known experts are invited to join a group of carefully selected policy-makers representing major forces in society (such as employers, unions, churches, professional associations, and environmentalists). Invited experts are asked to offer their professional judgment, but they often do not need to present formal evidence for their claims. This approach is based on trust in the expertise of scientists.

These four styles are helpful in characterizing and analyzing different national approaches to policy-making. The American system is oriented toward the adversarial style and the Japanese system toward the consensual. The policy style of northern Europe comes closest to the corporatist approach, whereas most southern European countries display a fiduciary approach. All these systems, however, are in transition. Interestingly the United States has tried to incorporate more consensual policies into its adversarial system, while Japan is faced with increasing demands for more public involvement in the policy process. These movements towards hybrid systems have contributed to the genesis of a new regulatory style, which may be called 'mediative.' There has been a trend in all technologically developed societies to experiment with opening expert deliberations to more varied forms of stakeholder or public participation. In the United States, it has taken the form of negotiated or mediated rule making, in Europe it has evolved as an opening of corporatist clubs to new groups such as the environmental movement. It is too early to say whether this new style will lead to more convergence among the countries or to a new set of cultural differentiations (Renn 1995).

6. *Conclusions*

The economic and political structures of modern societies underwent rapid transitions in the late twentieth century. This transition was accompanied by globalization of information, trade, and cultural lifestyles, an increased pluralism of positions, values, and claims, the erosion of trust and confidence in governing bodies, an increased public pressure for participation, and growing polarization between fundamentalist groups and agents of progressive change. The resulting conflicts put pressure on political systems to integrate different outlooks and visions of the future and to provide justifications of governmental decisions on the basis of both facts and values. In this situation, policy-making institutions discovered an urgent need for policy advice, as well as new modes of integrating expertise with values and preferences. Research on advisory processes indicates that the following points will need to be addressed.

(a) Using scientific expertise in the policy arena is one element in the quest of modern societies to replace or amend the collective learning process of trial and error by more humane methods of anticipation, in which the possibility of errors is reduced. This process is socially desired, though it cannot reduce the uncertainties of change to zero. Anticipation both necessitates and places new demands on expertise in the service of policy-making.

(b) Scientific expertise can serve five functions: enlightenment; pragmatic or instrumental; reflexive; catalytic; and communicative. All five are in demand by policy-makers, but expertise may also distort their perspective on the issue or prescribe a specific framing of the problem. That is why many policy analysts demand that scientific input be controlled by democratic institutions and be open to public scrutiny.

(c) Scientific expertise is also used for legitimizing

decisions and justifying policies that may face resistance or opposition. Expertise can, therefore, conflict with public preferences or interests. In addition, policy-makers and experts pursue different goals and priorities. Expertise should be regarded as one crucial element of policy-making among others. Scientific advice is often mandated by law, but its potential contributions may vary from one policy arena to another. In particular, scientific expertise cannot replace public input in the form of locally relevant knowledge, historical insights, and social values.

(d) The influence of expertise depends on the cultural meaning of expertise in different social and political arenas. If systematic expertise is regarded as the outcome of a socially constructed knowledge system, its authority can be trumped by processes that are seen to be more democratic. If, however, systematic expertise is seen as an approximation of reality or truth, it gains a privileged status among different sources of knowledge and inputs, even if it is the product of a democratically imperfect process. The degree to which these different understandings of expertise are accepted or acknowledged within a political arena will affect the practical influence and power of experts in collective decision-making.

(e) Scientific expertise is absorbed and utilized by the various policy systems in different styles. One can distinguish four styles: adversarial, fiduciary, consensual, and corporatist. A new mediative style seems to be evolving from the transitions toward more open procedures in decision-making. This style seems to be specifically adjusted to postmodern societies. Scientific expertise in this style pursues a 'system and problem oriented' approach to policy-making, in which science, politics, and economics are linked by strategic networks.

(f) Organizing and structuring discourses on the selection of policy options is essential for the democratic, fair, and competent management of public affairs. The mere desire to initiate a two-way-communication process and the willingness to listen to public concerns are not sufficient. Deliberative processes are based on a structure that assures the integration of technical expertise, regulatory requirements, and public values. Co-operative discourse is one model among others that has been designed to meet that challenge.

No one questions the need to initiate a common discourse or dialogue among experts, policy-makers, stakeholders and representatives of affected publics. This is particularly necessary if highly controversial subjects are at stake. The main challenge of deliberative processes will continue to be how to integrate scientific expertise, rational decision-making, and public values in a coherent form.

See also: Expert Systems in Cognitive Science; Expert Testimony; Expert Witness and the Legal System: Psychological Aspects; Expertise, Acquisition of; Medical Expertise, Cognitive Psychology of; Policy History: Origins; Policy Knowledge: Universities; Policy Networks

Bibliography

Beck U 1992 *Risk Society: Toward a New Modernity* (trans. Ritter M A). Sage, London

Bradbury J A 1989 The policy implications of differing concepts of risk. *Science, Technology, and Human Values* **14**(4): 380–99

Brickman R S, Jasonoff S, Ilgen T 1985 *Controlling Chemicals: The Politics of Regulation in Europe and the United States*. Cornell University Press, Ithaca, NY

Cadiou J-M 2001 The changing relationship between science, technology and governance. *The IPTS Report* **52**: 27–9

Campbell N 1951 (original 1921) *What is Science?* Dover, New York

Eder K 1992 Politics and culture: On the sociocultural analysis of political participation. In: Honneth A, McCarthy T, Offe C, Wellmer A (eds.) *Cultural–Political Interventions in the Unfinished Project of Enlightenment*. MIT Press: Cambridge, MA, pp. 95–120

Fisher R, Ury W 1981 *Getting to Yes: Negotiating Agreement without Giving In*. Penguin Books, New York

Fishkin J 1991 *Democracy and Deliberation: New Directions of Democratic Reform*. Yale University Press, New Haven, CT

Friedman J (ed.) 1995 *The Rational Choice Controversy*. Yale University Press, New Haven, CT

Funtowicz S O, Ravetz J R 1990 *Uncertainty and Quality in Science for Policy*. Kluwer, Dordrecht and Boston

Habermas J 1987 *Theory of Communicative Action. Vol. II: Reason and the Rationalization of Society*. Beacon Press, Boston

Jaeger C 1998 Current thinking on using scientific findings in environmental policy making. *Environmental Modeling and Assessment* **3**: 143–53

Jasanoff S 1986 *Risk Management and Political Culture*. Russell Sage Foundation, New York

Jasanoff S 1990 *The Fifth Branch: Science Advisers as Policy-makers*. Harvard University Press, Cambridge, MA

Knorr-Cetina K D 1981 *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Pergamon Press, Oxford, UK

Latour B, Woolgar S 1979 *Laboratory Life: The Social Construction of Scientific Facts*. Sage, Beverley Hills and London

Lindblom C E, Cohen D K 1979 *Usable Knowledge: Social Science and Social Problem Solving*. Yale University Press, New Haven, CT

Luhmann N 1986 The autopoiesis of social systems. In: Geyer R F, van der Zouven J (eds.) *Sociocybernetic Paradoxes: Observation, Control and Evolution of Self-steering Systems*. Sage, London, pp. 172–92

Luhmann N 1993 *Risk: A Sociological Theory*. Aldine de Gruyter, New York

Mukerji C 1989 *A Fragile Power: Scientists and the State*. Princeton University Press, Princeton, NJ

O'Riordan T, Wynne B 1987 Regulating environmental risks: A comparative perspective. In: Kleindorfer P R, Kunreuther H C (eds.) *Insuring and Managing Hazardous Risks: From Seveso to Bhopal and Beyond*. Springer, Berlin, pp. 389–410

Raiffa H 1994 *The Art and Science of Negotiation*, 12th edn. Cambridge University Press, Cambridge, UK

Renn O 1995 Style of using scientific expertise: A comparative framework. *Science and Public Policy* **22**: 147–56
Renn O 1999 A model for an analytic deliberative process in risk management. *Environmental Science and Technology* **33**(18): 3049–55
Renn O, Webler T, Rakel H, Dienel P C, Johnson B B 1993 Public participation in decision making: A three-step-procedure. *Policy Sciences* **26**: 189–214
Rip A 1992 The development of restrictedness in the sciences. In: Elias N, Martins H, Whitely R (eds.) *Scientific Establishments and Hierarchies*. Kluwer. Dordrecht and Boston, pp. 219–38
Rowe G, Frewer L J 2000 Public participation methods: A framework for evaluation. *Science, Technology & Human Values* **225**(1): 3–29
Shrader-Frechette K 1991 *Risk and Rationality*. *Philosophical Foundations for Populist Reforms*. University of California Press, Berkeley, CA
Simon J L 1992 There is no environmental, population, or resource crisis. In: Tyler-Miller G (ed.) *Living in the Environment*. Wadsworth, Belmont, pp. 29–30
Skillington T 1997 Politics and the struggle to define: A discourse analysis of the framing strategies of competing actors in a 'new' participatory forum. *British Journal of Sociology* **48**(3): 493–513
Solingen E 1993 Between models and the State: Scientists in comparative perspective. *Comparative Politics* **26**(1): 19–27
Vitousek P M, Ehrlich A H, Matson P H 1986 Human appropriation of the products of photosynthesis. *Bio Science* **34**: 368–73
Webler T 1995 'Right' discourse in citizen participation. An evaluative yardstick. In: Renn O, Webler T, Wiedemann P (eds.) *Fairness and Competence in Citizen Participation*. Kluwer, Dordrecht and Boston, pp. 35–86
Wynne B 1989 Sheepfarming after Chernobyl. *Environment* **31**: 11–15, 33–9
Wynne B 1992 Uncertainty and environmental learning. Reconceiving science and policy in the preventive paradigm. *Global Environmental Change* **2**: 111–27

O. Renn

Science and the Media

The phrase 'science and the media' as a single concept encompasses two major social institutions. Science is both a system of reliable knowledge about the natural world and a complex social system for developing and maintaining that knowledge. Similarly, the media comprise a variegated system for collecting and presenting information as well as being an institution with significant economic, political, and social impact. Though long perceived as separate realms, recent scholarship has highlighted their mutual dependence. Increasing episodes of tension and conflict in the late twentieth century led scholars, scientists, media producers, and social critics to explore the interactions of science and the media. Working scientists and media producers sought guidance on how to use each others' resources more effectively, while analysts asked how such issues as political and economic interests, rhetorical conventions, and audience responses shed light on the roles of science and media in shaping each other. One effect of recent scholarship has been to elide the difference between 'media' (all forms of media) and 'the media' (particularly the major institutions of newspapers, magazines, television, and radio), an elision that will be evident throughout this article.

1. *Science and the Media: A Brief History*

The history of science and the media is the history of a growing mesh: increasing use of media by science, increasing attention to scientific ideas by media institutions, and increasing tensions caused by the rising interaction.

Through the middle of the twentieth century, science and media were largely seen as separate realms. Science had emerged in the nineteenth century from natural philosophy as a systematic approach to understanding the natural world. With the advent of research-based industries such as chemical dyes and electronic communication, science began to attract the attention of capitalists and government leaders who understood the value of controlling scientific knowledge. By the end of World War II, the emergence of 'big science' had made science a powerful institutional force as well as a system of accredited knowledge with increasingly far-reaching applications. The media had also evolved in the nineteenth century from small, local organizations into nationally and internationally-circulated publications that served as tools of merchant capitalism and political control, full of advertising, business and political news, and the ideology of the industrial revolution. By the early twentieth century, new electronic media such as movies and radio provided the means for newly-developed forms of 'publicity' and propaganda to help produce modern mass society, as well as to increase public access to new forms of cultural entertainment.

Since the late seventeenth century, natural philosophers had relied on print media, particularly professional journals such as the *Philosophical Transactions* and books such as Isaac Newton's *Principia*, as tools for disseminating the results of their investigations. These publications were not only records of experiments or philosophical investigations, but also served as active carriers of the rhetorical structures that enabled natural philosophers to convince people outside their immediate vicinity of the truth of their claims (Shapin and Schaffer 1985).

Early in the nineteenth century, new forms of science media emerged, particularly a genre known variously as 'popularization,' *vulgarisation* [French], or *divulgacion* [Spanish]. These books and magazines provided entrée into the knowledge of the natural world for non-scientist readers. They ranged from textbooks for young women through to new mass-circulation magazines filled with instruction on the latest achievements

of a rapidly industrializing society. The so-called 'great men' of late nineteenth century England (most notably Thomas Huxley and John Tyndall) became evangelists for science, lecturing widely to enlist public support for the enterprise of rational understanding of the natural world; they converted their lectures into articles and books to spread their reach even further. At the same time, scientists' use of media for disseminating knowledge among the scientific community grew dramatically, with the growth of new journals and the founding of publishing houses committed to scientific information.

The growth of both science and media in the twentieth century forced specialization into the fields, leading to a distinction between what some scientists have called 'communication in science' and 'communication about science.' For scientists in their daily work, broad synthetic journals such as *Science* and *Nature* (published respectively in the USA and the UK) were joined by topical journals in chemistry, physics, biology, and their many sub-disciplines. For non-scientists, magazines and other media (such as radio and new 'wire services' for newspapers) increasingly focused on particular audiences, such as teachers, inventors, 'high culture' readers, general media users, and so on. Although scientific ideas appeared in some entertainment media, science in most non-technical media became a component of news coverage for different audiences. In the 1920s and 1930s, a small group of professional journalists began to write almost exclusively about science. (These developments were clearest in the highly developed countries of Europe and North America. The patterns in the rest of the world have been little studied and so cannot be described.)

By the second half of the twentieth century, the interaction of science and media had become a complex web, in which new scientific research appeared in abstracts, journals, press reports, and online, while other forms of science–not necessarily derivative of the research reports–appeared regularly in media outlets such as newspapers, televisions, websites, radio programs, puppet shows, traveling circuses, and museums. Moreover, as science became central to modern culture, the presence of science in media such as poetry, sculpture, fine art, entertainment films, and so on, became more evident. In the journalistic media, many reports circulated internationally, through cooperative agreements between public broadcasting systems and worldwide news services. The rise of the World Wide Web in the 1990s provided further opportunities for information about science to circulate more fully between the developed and developing world.

However, from about the mid-century onward, no matter what media were involved, many working scientists perceived presentations in non-technical media to be incorrect, oversimplified, or sensationalized. Thus, tensions developed between the institutions of science and those of the media. Those tensions led eventually to the development of an analytic field of science and media, populated largely by media sociologists, communication content scholars, and sociologists of science.

2. *Understanding Science and Media*

Analysis of science and media has led to a deeper understanding of the essential role of media in creating scientific knowledge, the ways that media presentations shape understandings of nature, and the role of public debate in constituting scientific issues.

The earliest attempts to understand the relationship of science and media came from two directions: the overwhelming growth of scientific information, leading to attempts by scientists and scientific institutions to learn how to manage scientific publications, and emerging tensions between scientists and journalists, as specialized reporting developed in the 1930s. Research and publications before and after World War II led by the 1970s to two common understandings.

First, sociologists of science had come to understand the fundamental role of communication in the production of reliable knowledge about the natural world. Scientists do not work in isolation, but must constantly present their ideas to colleagues for acknowledgement, testing, modification, and approbation. The process of communication, occurring in various media, is the 'essence of science' (Garvey 1979), leading to science becoming 'public knowledge' (Ziman 1968). At the same time, the apparent differences between the goals of science (methodical, tentative, but nonetheless relatively certain statements about the natural world) and the needs of public media (rapid, attention-getting narratives about issues directly related to readers and viewers, regardless of certainty) had also been explicated (Krieghbaum 1967), leading to practical attempts to 'bridge the gap' between science and the media but also to recurring fears that the gap was unbridgeable.

Beginning in the 1970s, new developments in the sociology of scientific knowledge opened up new ways of conceiving of the relationship between science and media (Barnes 1974, Latour and Woolgar 1979). In particular, attention to the rhetorical goals of scientists in their use of media suggested that a distinction between communication in science and communication about science could not be maintained. Instead, some researchers began to discuss the 'expository' nature of science, showing how scientists tailored their communication to meet the needs of specific media and specific communication contexts (Bazerman 1988, Shinn and Whitley 1985). These researchers highlighted the interaction of the idealized intellectual goals of science (for the production of reliable knowledge about the natural world) with the social and institutional goals of scientists and their employers or

patrons (for priority, status in the community, ownership of ideas and thus patents, and so on).

In this conception of science and media, scientific knowledge itself might differ in different presentations—in some cases, appearing as a 'narrative of science' (focusing on the methodological and theoretical structures leading to particular knowledge of the natural world), in other cases as a 'narrative of nature' (focusing on the relationships between organisms or entities, with emphasis on the 'story' linking different aspects of those organisms and entities) (Myers 1990). The very boundary between 'science' and 'non-science' (or 'mere' popularization), between communication in science and communication about science, was shown to be highly mutable, itself an object of rhetorical construction used for political purposes by participants in scientific controversies to establish control over elements of a debate (Hilgartner 1990).

A second new line of inquiry highlighted the institutional interdependence of science and the media. Pointing to the growing post-World War II need for scientific institutions to generate public and political support, and to the media's need to draw on scientific developments for constant infusions of drama and 'newness,' the new research identified 'selling science' as a major element of the political economy linking science and the media (Nelkin 1987).

Finally, work on the images of science that appeared over time helped show the ways that social concerns interacted with scientific developments to create cultural representations of science (Weart 1988, LaFollette 1990, Nelkin and Lindee 1995).

At the same time that these developments were occurring in the understanding of science communication, new approaches were developing in understanding the relationship between science and the public. Spurred by concerns in the scientific community about lack of 'scientific literacy,' the studies ultimately questioned the idea that the public has a 'deficit' of knowledge that needs to be 'improved' (Irwin and Wynne 1996). Instead, the new approach focused on the possibility of engaging the public in discussions about scientific issues, recognizing the contingent nature of understanding of scientific issues and thus the alternative meanings that might emerge in democratic discussions instead of authoritarian pronouncements (Sclove 1995). Much of the new work looked at issues of uncertainty and risk, highlighting the social negotiations required to make personal and policy decisions in a context of incomplete information (Friedman et al. 1999). The new work clearly implied that science and the media cannot be perceived as separate institutions with separate goals. Instead, better understanding comes from analyzing science and media across audiences, looking at the political and economic interests of media producers, at the rhetorical meanings of media conventions, and at the audience responses to media content (especially in terms of meaning-making).

3. *Approaches to Studying Science and Media*

3.1 *Political and Economic Interests of Media Producers*

The most detailed and well-developed understandings of science and the media point to the political and economic interests that motivate the producers of media representations of science. In many countries, the media depend on government subsidies or at least on political tolerance. Moreover, media are expensive to produce. While the presence of the media in democratic societies is often seen as fundamental to political liberty, the economic need to appeal to broad audiences (even for government-controlled media) often leads to editorial decisions that emphasize broad coverage and sensationalism over depth and sobriety. Science media are no different, and even the most prestigious scientific journals such as *Science* and *Nature* have been criticized for highlighting the most 'newsworthy' research, for hyping scientific reports beyond their value, for engaging in 'selling science.' Newspaper and television coverage of science is routinely criticized for its sensationalism, particularly in the coverage of controversial issues such as food contamination (e.g., the BSE or 'mad cow' epidemic in Britain in the 1980s and 1990s), genetic modification of food, and global climate change. Journalists (and their employers) routinely defend their practices on the basis of 'what readers/viewers want,' which is often judged by sales success. Entertainment media are criticized even more heavily for creating movie plots that involve impossible science or focus on highly unlikely scenarios; producers disclaim any responsibility for accuracy, claiming only to be providing (profitable) escapism. These recurring disputes between the scientific community and media producers highlight the irreconcilable tension between the goals of science and the goals of the media.

3.2 *Rhetorical Meanings of Media Conventions*

An emerging area of understanding is the way that media conventions shape the rhetorical meaning of science communication. Scientific papers reporting original experimental research, for example, have developed a standardized format (called 'IMRAD,' for introduction, methods, results, analysis, discussion). The IMRAD format, conventionally written in an impersonal tone, with passive grammatical constructions, hides the presence of the researcher in the research, highlighting the objective character of scientific knowledge that is claimed to exist independently of its observer. In journalistic reports of scientific controversies, the conventions of objective reporting require 'balance'—the presentation of both sides of a controversy (Bazerman 1988). Though 99.9 percent of all researchers may hold that a new claim

(such as the eventually discredited 1989 announcement of a new form of 'cold nuclear fusion') is scientifically untenable, the journalistic norm of balance may lead to approximately equal attention to all claims. Similarly, journalistic goals of story telling often lead to emphasis on the human, narrative dimensions of an issue over the theoretical or mathematical components of the science. New research is exploring the meaning of visual representations of science, at every level from the technical outputs of genome analyzers to mass media manipulations of false-color images produced by orbiting astronomical telescopes (Lynch and Woolgar 1990).

3.3 Audience Responses to Media Content

One of the most difficult areas to investigate is the audience response to particular media presentations. Virtually no work, for example, has been done on how scientists respond to specialized science media, despite anecdotal stories suggesting that changes in information presentation (such as World Wide Web access to scientific databases) may have dramatic impacts on how scientists conceive of the natural world and frame their hypotheses and theories. Some work has been done at broader public levels, such as responses to media presentations of risk information. But, although the psychology of risk perceptions, for example, has been extensively investigated, the factors affecting individual responses to media coverage of specific risky incidents are so varied as to defy measurement. Though the complexity leads some researchers to reject the possibility of quantitative analysis, others insist that social scientists can learn to isolate appropriate variables. Certainly some areas of audience response could be investigated more carefully; audience reaction to images of science in entertainment films, for example, has never been systematically tested, despite frequent statements that 'the image of scientists in the movies is bad.'

The shift away from separate analysis of 'science' and 'media' to an integrated understanding of 'science and media' has led in recent years to the possibility of cross-cutting analyses, such as those of producers, rhetorical structures, and audience reception. These analyses will ultimately yield a much richer understanding of the interaction of science and media. Less clear is how such improved understanding will contribute to the practical concerns of those in the scientific community and elsewhere, who worry about possible linkages among science literacy, the image of science, and public support for science.

See also: Educational Media; Genetics and the Media; Media and Child Development; Media and History: Cultural Concerns; Media Ethics; Media Events; Media, Uses of; Public Broadcasting; Public Relations in Media; Research Publication: Ethical Aspects

Bibliography

Barnes B 1974 *Scientific Knowledge and Sociological Theory*. Routledge and Kegan Paul, London

Bazerman C 1988 *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. University of Wisconsin Press, Madison, WI

Friedman S M, Dunwoody S, Rogers C L (eds.) 1999 *Communicating Uncertainty: Media Coverage of New and Controversial Science*. Erlbaum Associates, Mahway, NJ

Garvey W D 1979 *Communication: The Essence of Science–Facilitating Information Exchange Among Librarians, Scientists, Engineers and Students*. Pergamon Press, New York

Hilgartner S 1990 The dominant view of popularization: conceptual problems, political uses. *Social Studies of Science* **20**(3): 519–39

Irwin A, Wynne B (eds.) 1996 *Misunderstanding Science? The Public Reconstruction of Science and Technology*. Cambridge University Press, Cambridge, UK

Krieghbaum H 1967 *Science and the Mass Media*. New York University Press, New York

LaFollette M 1990 *Making Science Our Own: Public Images of Science, 1910–1955*. University of Chicago Press, Chicago

Latour B, Woolgar S 1979 *Laboratory Life*. Sage, Beverly Hills, CA

Lynch M, Woolgar S 1990 *Representation in Scientific Practice*. MIT Press, Cambridge, MA

Myers G 1990 *Writing Biology: Texts in the Social Construction of Scientific Knowledge*. University of Wisconsin Press, Madison, WI

Nelkin D 1987 *Selling Science: How the Press Covers Science and Technology*. Freeman, New York

Nelkin D, Lindee M S 1995 *The DNA Mystique: The Gene as a Cultural Icon*. Freeman, New York

Sclove R 1995 *Democracy and Technology*. Guilford, New York

Shapin S, Schaffer S 1985 *Leviathan and the Air-pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, Princeton, NJ

Shinn T, Whitley R (eds.) 1985 *Expository Science: Forms and Functions of Popularisation* D Reidel, Dordrecht, The Netherlands, Vol. 9

Weart S 1988 *Nuclear Fear: A History of Images*. Harvard University Press, Cambridge, MA

Ziman J M 1968 *Public Knowledge*: *An Essay Concerning the Social Dimension of Science*. Cambridge University Press, Cambridge, UK

B. V. Lewenstein

Science and the State

1. Introduction

Since the sixteenth century, the 'scientific revolution' has generated a host of new, intellectual, rhetorical, and institutional strategies for undermining traditional political authorities and constructing alternative ones. Science became a vital resource in the modern attempt to discredit traditional political hierarchies and spiritual transcendental sources of authority in

the context of public affairs. By providing new rationales for order compatible with the novel modern commitments to the values of individualism, voluntarism, and egalitarianism, science and technology paradoxically also provided new grounds for novel modern forms of hierarchy and authority committed to the use of knowledge in the reconstruction of society. The widely shared notion that science goes along with progress appeared to suggest that knowledge, when applied to public affairs, can in fact depoliticize public discourse and action. This faith produced wide mandates for large-scale social and political engineering and monumental state-sponsored technological projects in both authoritarian and democratic states.

While such ideals and projects were from the very beginning criticized by observers such as Edmund Burke (1729–97), throughout the twentieth century the record of the relations between science and politics turned out to be sufficiently mixed to discard earlier hopes. The controversial role of scientists in the production and military deployment of means of mass destruction, the role of scientists in the Nazi experiments on human beings and the endorsement of racism, and their involvement in the industrial pollution of the environment, diminished the public trust in science fostered by such developments as medical breakthroughs and spectacular space flights.

Since the closing decades of the twentieth century, an increasing number of social scientists and historians have been pointing out that new uncertainties in the sphere of politics have been converging with newly recognized uncertainties of science and risks of technology to form novel postmodern configurations of state, science, and society (Eisenstadt 1999, Beck 1992). It has become increasingly recognized that, in ethnically and religiously heterogeneous and multicultural societies, science and technology can no longer be attached to uncontroversial comprehensive values that privilege claims of neutrality, objectivity, and rationality; nor can the earlier belief that knowledge can depoliticize public discourse and action be sustained any longer.

2. *The Authorities of Science and the State*

The belief that science can provide a secular version of the synoptic divine view of human affairs was politically most significant. Until its erosion towards the end of the twentieth century, the modernist configuration of science and the state was formed over the course of at least four centuries. The secularization of political power in the modern state has put the dilemma of arbitrary rule at the center of modern political theory and practice. Whereas legal and constitutional constraints appeared to be the most appropriate remedy to this problem, the rise of modern science and technology encouraged faith in the possibility of restraint and moderation through public enlightenment and by means of scientific and technical advice. Not surprisingly, the redemptive drive behind the integration of science and politics had its origins in premodern religious visions and values (Manuel 1974). In both its authoritarian and democratic versions, this modernist program rested on the belief that the modern state can remedy the defects of the present sociopolitical order and realize an ideal order. Science contributed to this modernist outlook the faith in the power of secular knowledge to control nature and reconstruct society. In his dedication to Lorenzo di Medici of Florence in *The Prince* (1513), Niccolo Machiavelli implies that the holder of the God's-eye view of the entire state is neither the King, who looks down at the people from the top of the hierarchy, nor the people, who see the King at the top from their lower state. It is a person like himself, a man of knowledge. The King, to be sure, has knowledge of the people, and the people have knowledge of the King but, because of his intellectual perspective, 'a man of humble and obscure condition' like Machiavelli can claim to see both the King and his subjects and understand the nature of their interactions.

Political theory and political science thus claimed to have inherited the God's-eye view of the whole polity and evolved a secular vision of the state as an object of knowledge. In turn, the nation-state has often found it useful to adopt the frame, if not always the content, of the inclusive scientific outlook in order to legitimate its interventions in the name of the general common good and its expressions in public goals such as security, health, and economic welfare. If, from the inclusive perspective of the state, all citizens, social groups, or institutions appeared as parts which the supreme power had the capacity to fit together or harmonize, science in its various forms often provided the technical means to rationalize, and the rhetorical strategies to depoliticize, such applications of power.

The potential of science not only as a source of knowledge but also as a politically useable authority was already recognized by Thomas Hobbes. This founder of modern political theory was in contact with Francis Bacon, admired William Harvey's findings about the circulation of the blood, and criticized Robert Boyle's position on the vacuum (Shapin and Schaffer 1985). In his influential works on the state, Hobbes tried to buttress his claims by appealing to the authority of scientific, especially mathematical, certainties, hoping to enlist the special force of a language, which claims to generate proofs as distinct from mere opinions (Skinner 1996). Although he was suspicious of the experimental science of Boyle, Hobbes followed Machiavelli in describing politics in the language of causes rather than motives. Thomas Sprat, the historian of the Royal Society (1667), claimed that when compared to the contentious languages of religions, the temperate discourse of science can advance con-

sensus across diverse social groups. In both the experimental and mathematical scientific traditions, however, knowledge, whether based on inferences or observations (or on their combination), was expected to end disputes.

The association between science and consensus was regarded as deeply significant in a society which had begun to view authority and order as constructed from the bottom up rather than coming from above (Dumont 1986). Reinforced by the growing image of science as a cooperative, nonhierarchical, and international enterprise in which free individuals come to uncoerced agreements on the nature of the universe, scientific modes of reasoning and thinking appeared to suggest a model for evolving discipline and order among equals (Polanyi 1958, 1962).

Deeply compatible with the notion that legitimate power and authority are constituted by social contracts, the rise of modern scientific knowledge appeared to accompany and reinforce the rise of the modern individual and his or her ability to challenge hierarchical forms of knowledge and politics. From Hobbes and Descarte, through Kant, J. S. Mill, and late-modern democratic thinkers such as Popper, Dewey, and Habermas, the commitment to the centrality of individual agency becomes inseparable from a growing faith in the possibility of expanding public enlightenment backed up by the demonstrated success of scientific modes of reasoning and acting.

3. *Power and the Social Sciences*

In succeeding centuries, one of the most persistent ambitions of Western society has been to make conclusions touching the 'things' of society, law, and politics appear as compelling and impervious to the fluctuations of mere opinion as conclusions touching the operation of heaven and earth. The faith that natural scientists are bound by the 'plain' language of numbers to speak with an authority which cannot be corrupted by fragile human judgment was gradually extended to the fields of engineering and the social sciences (Porter 1986). The notion that experts who are disciplined by respect for objective facts are a symbol of integrity and can therefore serve as guardians of public virtues against the villains of politics and business was widely interpreted to include, beyond natural scientists, other categories of experts who speak the language of numbers. Social science disciplines like economics, social statistics, sociology, political science, and psychology adopted modes of observing, inferring, and arguing which appeared to deploy in the sphere of social experience notions of objective facts and discernible laws similar to those that the natural sciences had developed in relation to physical nature. The social sciences discovered that the language of quantification is not only a powerful tool in the production of knowledge of society but also a valuable political and bureaucratic resource for depersonalizing, and thus legitimizing, the exercise of power (Porter 1995). In the context of social and political life the separation, facilitated by expert languages, between facts and values, causal chains and motives, appeared profoundly consequential. Until effectively challenged, mostly since the 1960s, such scientific and technical orientations to human affairs appeared to produce the belief that economic, social, political, and even moral 'facts' can be used to distinguish subjective or partisan from professional and apolitical arguments or actions. The apparent authority of science in the social and political context motivated the principal founders of modern ideologies like socialism, fascism, and liberalism to enlist science to their views of history, society, and the future.

While theoretical and mathematified scientific knowledge remained esoteric, as religious knowledge was for centuries in the premodern state, the ethos of general enlightenment and the conception of scientific knowledge as essentially public made scientific claims appear agreeable to democratic publics even when these claims remained, in fact, elusive and removed from their understanding. In fields such as physics, chemistry, and medicine, machines, instruments, and drugs could often validate in the eyes of the lay public claims of knowledge which theories alone could not substantiate. The belief that science depoliticizes the grounds of state policies and actions, and subordinates them to objective and, at least in the eyes of some, transparent, professional standards, has been highly consequential for the uses of science in the modern state.

Political leaders and public servants discovered that the appearance of transparency and objectivity can make even centralized political power seem publicly accountable (Price 1965, Ezrahi 1990). Moreover, the realization that the authority of science could in various contexts be detached from substantive scientific knowledge and deployed without the latter's constraint often enhanced the political usefulness of both natural and social scientists independently of the value actually accorded to their professional judgments. Still, the uses of expert scientific authority in the legitimization of state politics and programs has often empowered scientists to actually exercise influence on the substance of government actions or to effectively criticize them from institutional bases outside the government (Jasanoff 1990). The actual uses of scientific expertise could often be just the consequence of a political demand for expert authority.

4. *Science in Democratic and Authoritarian States*

In democratic societies, the integration of scientific authority, and sometimes also of scientific knowledge and technology, into the operations of the modern

state encouraged the development of a scientifically informed public criticism of the government. This process was greatly augmented by the rise of the modern mass media and the ability of scientists to call public attention to governmental failures traceable to nonuse or misuse of scientific expertise. The role of scientists in empowering public criticism of the uses of nuclear weapons and nuclear energy (Balough 1991), in grounding public criticism of the operation of such agencies as the National Aeronautics and Space Administration (NASA) and the Food and Drugs Administration (FDA), and in the criticism of public and private agencies for polluting the environment, illustrates the point. In authoritarian states like the Soviet Union, such cooperation between scientists, the mass media, and the public in criticizing government policies and actually holding the government accountable was, of course, usually repressed. In the authoritarian modernist version of science and the state, science was used mostly to justify, and partly to direct, comprehensive planning and control. Here, the force of scientific knowledge and authority rarely constrained the rationalization of centralization or warranted public skepticism towards the government or criticism of its actions (Scott 1998).

Even in democratic states, however, science and technology were massively used to promote the goals of reconstruction, coordination, mass production, and uniformity. The redemptive role assumed by the political leadership in the name of equality, public welfare, and uniformity was no less effective in rationalizing state interventions than the totalistic utopianism enlisted to justify the direction and manipulation of society in authoritarian regimes. In all the variants of the modern state, bodies of expert knowledge such as statistics, demography, geography, macroeconomics, city planning, public medicine, mental health, applied physics, psychology, geology, and others were used to rationalize the creation of government organizations and semipublic regulatory bodies which had the mandate to shift at least part of the discretion held formerly by political, legal, and bureaucratic agents to experts (Price 1965, Wagner et al. 1991).

In many democratic states, the receptiveness to experts within the state bureaucracy was manifest in the introduction of merit-based, selective, personnel recruitment procedures and a widening use of examinations. The cases of the former USSR and Communist China suggest, by contrast, the ease with which experts and their expertise could be controlled by political loyalists. Even where it was facilitated, however, the welding of bureaucrats and professionals was bound to be conflict-ridden. The hierarchical authority structure of bureaucratic organizations has inevitably come to both challenge and be challenged by the more horizontal structure of professional peer controls (Wilson 1989, Larson 1977). Nevertheless, the professionalization of fields of public policy and regulation boosted the deployment of more strictly instrumental frames of public actions in large areas. The synoptic gazes of power and science could thus converge in supporting such projects as holistic social engineering, city planning, and industrial scientific agriculture (Scott 1998).

5. *Science, War, and Politics*

The most dramatic and consequential collaboration between science and the modern state took place during wars. Such collaboration was usually facilitated by the atmosphere of general mobilization, defining the war effort as the top goal of the nation. In a state of emergency, even democratic states have suspended, or at least limited, the competitive political process. Under such conditions, democracies have temporarily, albeit voluntarily, experienced the usual conditions of the authoritarian state: centralized command and control and a nationally coordinated effort. In the aftermath of such wars, politicians and experts often tried to persist in perpetuating the optimal conditions of their collaboration, regarding the resumption of usual political processes as disruptive of orderly rational procedures. In authoritarian nationalist or socialist states, open competitive politics was repressed as a matter of routine along with the freedom of scientific research, and the liberty of scientists to diffuse their findings and interpret them to students and the public at large (Mosse 1966). In such countries, the political elite usually faced the dilemma of how to exploit the resources of science for the war effort and for advancing its domestic goals without becoming vulnerable to the revolutionary potential of science as the expression of free reason, criticism, and what Robert K. Merton called 'organized skepticism' (1973). In the democratic state, the special affinities between the participatory ethos of citizens, free to judge and evaluate their government, and the values of scientific knowledge and criticism did not usually allow the coherences and the clarities of the war period to survive for long, and the relations between science and the state had to readjust to conditions of open political contests and the tensions between scientific advice to, or scientific criticism of, the state. Limitations imposed by open political contests were mitigated, however, in political cultures which encouraged citizens to evaluate and judge the government with reference to the adequacy of its performance in promoting public goals. The links between instrumental success and political legitimation preserved in such contexts the value of the interplay between expert advice to the government and to its public critics in the dynamic making of public policy and the construction of political authority. Political motives for appealing to the authority of science could of course be compatible with the readiness to ignore knowledge and

adequate performance, although it did not need to contradict the willingness to actually use scientific knowledge to improve performance. Discrepancies between the uses of scientific authority and scientific knowledge became widespread, however, because of the growing gaps between the timeframes of science and politics in democratic societies. In fields such as public health, economic development, security, and general welfare, instrumentally effective policies and programs must usually be measured over a period of years, and sometimes even decades. But the politicians whose status has come to be mediated by the modern mass media need to demonstrate their achievements within the timeframe of months or even days or weeks. In contemporary politics, one month is a long time, and the life expectancy of issues on the public agenda is usually much shorter. Such a state of affairs means that political actors often lack the political incentives to invest in the costly resources that are required for instrumental success whose payoffs are likely to appear only during the incumbency of their political successors.

6. *Science and the Transformation of the Modern Democratic State*

These conditions have encouraged leaders in democratic states increasingly to replace policy decisions and elaborate programs with grand political gestures (Ezrahi 1990). Where instant rewards can be obtained by well-advertised commitment to a particular goal or policy, the political motive to invest in substantive moves to change reality and improve governmental performance over the longer run may easily decline.

Besides the tendency of democratic politics to privilege the present over the future in the allocation of public resources, the position of science in substantiating long-term instrumental approach to public policy was further destabilized by the fragmentation of the normative mandates of state policies. Roughly since the 1960s, the publics of most Western democracies have become increasingly aware of the inherent constraints on the determination and ordering of values for the purpose of guiding public choices. Besides the influence of ethnic, religious, linguistic, or cultural differences on this process, such developments as the feminist movement reflect even deeper pressures to reorder basic values. Feminist spokespersons demanded, for example, that financial and scientific resources directed to control diseases be redistributed to redress gender discrimination against the treatment of specifically female diseases. In part, the feminist critique has been but one aspect of wider processes of individuation, which, towards the later part of the twentieth century, have increased the diversity of identities, tastes, lifestyles, and patterns of association in modern societies. This new pluralism implied the necessity of continually renegotiating the uses of science and technology in the social context and locally adapting it to the diverse balances of values and interests in different communities. In the course of the twentieth century, leading scientific voices like J. B. S. Haldane, J. B. Bernal, and Jacques Monod tended to treat almost any resistance to the application of advanced scientific knowledge in human affairs as an 'abuse of science' resulting from irrationalism, ignorance, or prejudice. These and many other scientists failed to anticipate the changes in social and political values which complicated the relations of science and politics during the twentieth century and the constraints imposed by the inherently competitive value environment of science and policy making in all modern states.

The constant, often unpredictable, changes in the value orientations of democratic publics and other related developments have clearly undermined confidence in the earlier separation between facts and values, science and politics, technological and political structures of action. Such a central element of policy making as risk assessment, for example, which for a long time was regarded as a matter best left to scientists, was gradually understood as actually a hybrid process combining science and policy judgments (Jasanoff 1990, Beck 1992). In contemporary society, in which the respective authorities of science and the state were demystified, knowledge has come to be regarded as too complex to directly check power, and power as too diffused to direct or repress the production and diffusion of knowledge.

Historically, one of the most important latent functions of science in the sociocultural construction of the democratic political universe was to warrant the faith in an objective world of publicly certifiable facts which can function as standards for the resolution or assessment of conflicts of opinions. To the extent that scientists and technologists could be regarded as having the authority to state the valid relations between causes and effects, they were believed to be vital to the attribution of responsibility for the desirable or undesirable consequences of public actions. Policies were regarded within this model of science and politics as hypotheses, which are subject to tests of experience before a witnessing public (Ezrahi 1990). Elements of this conception have corresponded to the modern political experience. Although the public gaze has always been mediated, and to some extent manipulated, by hegemonic elites, publics could usually see when a war effort failed to achieve the stated objectives, when a major technology like a nuclear reactor failed and endangered millions of citizens, or when economic policy succeeded in producing affluence and stability. The expansion of this conception of government responsibility and accountability to regions outside the West even made it possible to redescribe fatal hunger in wide areas in India, for instance, as the consequence of policy rather

than natural disaster (Sen 1981). Still, in large areas, the cultural and normative presuppositions of the neutral public realm as a shared frame for the nonpartisan or nonideological description and evaluation of reality have not survived the proliferation of the modern mass media and the spread of political and cultural pluralism. In contemporary democracies, the pervasive mediating role of largely commercialized electronic communication systems has diminished the power of the state to influence public perceptions of politics and accentuated the weight of emotional and esthetic factors relative to knowledge or information in the representation and construction of political reality. Influenced by diverse specialized and individualized electronic media, the proliferation of partly insulated micro-sociocultural universes within the larger society has generated a corresponding proliferation of incommensurable notions of reality, causality, and factuality.

7. *Transformations of the Institutional Structure and Intellectual Orientations of Science*

Changes in the political and institutional environment of science in the modern state have been accompanied by changes in the internal institutional and intellectual life of science itself. In the modern democratic state, the status of science as a distinct source of certified knowledge and authority was related to its institutional autonomy vis-à-vis the structure of the state and its independence from private economic institutions. Freedom of research and academic freedom were regarded as necessary conditions for the production and diffusion of scientific knowledge and, therefore, also for its powers to ground apolitical advice and criticism in the context of public affairs. Such institutional autonomy was, of course, never complete. The very conditions of independence and freedom had to be negotiated in each society and reflected at least a tacit balance between the needs of science and the state as perceived by the government. But the international universalistic ethos of science tended to ignore or underplay such local variations (Solingen 1994).

The institutional arrangements that secured a degree of autonomy and freedom to scientists seemed to warrant the distinction between basic and applied research, between pure research directed to the advancement of knowledge and research and development directed to advance specific industrial, medical, or other practical goals. An important function of this distinction was to balance the adaptation of science to the needs of the modern state and the preservation of the internal intellectual traditions and practices of scientific research and academic institutions. In practice, the relations between science and the state were more symbiotic than the public ethos of science would have suggested. Scientific institutions, in order to function, almost always required public and political support and often also financial assistance. On the other hand, the state could not remain indifferent to the potential uses of scientific knowledge and authority to both secure adequate responses to problems and facilitate the legitimation of its actions by reference to apolitical expert authority. Thus, while the separation between truth and power and the respect for the boundaries between pure and applied sciences, as well as between politics and administration, were considered for a long time as the appropriate way to think about the relation between science and politics (Price 1965), the relations between science and the state, as well as between scientists and politicians, turned out to be much more interactive and conflict-ridden. Since scientists and politicians respectively controlled assets useful for each other's work, they were bound to be more active in trying to pressure each other to cooperate and engage in mutually useful exchanges (Gibbons et al. 1994).

Not all these exchanges were useful to either science or the state in the long run. The mobilization of scientists to the war efforts of the modern nation-state, while it boosted the status of scientists domestically in the short run, had deleterious effects on the international network of scientific cooperation, as well as on the independence of science in the long run. Especially in cases of internally controversial wars, like the American involvement in Vietnam, the mobilization of science inevitably split the scientific community and politicized the status of science. Nevertheless, the unprecedented outlays of public money made available to science in the name of national defense allowed many research institutions to acquire expansive advanced facilities and instruments that permitted the boosting of pure science in addition to militarily related research and development. But the gains in size and in potential scientific advance were obtained at the cost of eroding the glorious insulation of scientific research and exposing its delicate internal navigational mechanisms to the impact of external political values and institutions (Leslie 1993). Yielding to the pressures of the modern nation-state to make science more relevant and useful to more immediate social goals inevitably reduced the influence of internal scientific considerations and the priorities of scientific research (Guston 2000). Following such developments, the university, the laboratory, and the scientific community at large could often appear less elitist and more patriotic, but also more detached from their earlier affinities to humanistic culture and liberal values.

Like the partnership between science and the state, the partnership between science and the market often facilitated by their collaboration during the war effort and postwar privatization of projects and services has undermined the autonomy of science and the ethos of basic research in many late modern states. But, whereas the interpenetration between science and the

state subjected the internal intellectual values of science to the pressures of public goals and pressing political needs, the links between science and the private sector subordinated scientific norms to the private values of profit making. The conversion of scientific expertise into capital opened the way for substantial private support for the intellectual pursuits of science. But the linkages between science, industry, and capital only reinforced the decline of science and its institutions as a bastion of enlightenment culture, progress, and universal intellectual values. Thus, while scientific and technical knowledge and skills were increasingly integrated into a wider spectrum of industrial productions and commercial services, science as a whole became more amorphous, less recognizable or representable as a set of distinct institutions, and a community with a shared ethos (Gibbons et al. 1994).

An instructive illustration of these processes is the pressure exerted by both the state and private business firms on scientists to compromise the cherished professional norm of publicity. In the context of science, the transparency of methodology and the publicity of research results have for a long time been reinforced by both ethical commitment and practical needs. The methodologies and findings of any research effort are the building blocks, the raw materials, with which scientists in discrete places produce more knowledge. Transparency and publicity are necessary for the orderly flow of the research process and the operation of science as a cooperative enterprise (Merton 1973). In addition, publicity has been a necessary element of the special status of science as public knowledge. From the very beginning, pioneering scientists such as Bacon, Boyle, and Lavoisier distinguished scientific claims from the claims of magic, alchemy, cabala, and other esoteric practices by the commitment to transparency. Transparency was also what rendered claims of scientific knowledge appear publicly acceptable by democratic citizens (Tocqueville 1945). Turning scientific research into a state secret in order to gain an advantage in war or into an economic (commercial) secret in order to gain an advantage in the market was not congenial to sustaining the cooperative system of science or the early luster of science as an embodiment of noble knowledge and virtues which transcended national loyalties and sectoral interests. In the aftermath of such developments, the virtues of objectivity, distinterestedness, universality, and rationality which were associated with earlier configurations of science (Polanyi 1958, 1962, Merton 1973) appeared unsustainable. Moreover, the fact that the resources of science and technology could be enlisted not only by liberal and democratic but also by fascist, communist, and other authoritarian regimes accentuated the image of science as an instrument insufficiently constrained by internal norms and flexible enough to serve contradictory, unprogressive, and extremely controversial causes.

8. *The Postmodern Condition and the Reconfiguration of Science and the State*

These developments had a profoundly paradoxical impact on the status of science in the late-modern state. Especially since the closing decades of the twentieth century, while such projects as the decoding of the human genome indicate that scientific knowledge has advanced beyond even the most optimistic predictions, the authority of science in society and in the context of public affairs has suffered a sharp decline. In order to consider this state of affairs as reversible, one needs to believe that the insular autonomy of science can be restored and that the political and economic environments of science can become more congenial for sustaining this condition. Such expectations seem unwarranted. On the other hand, an apparent decline in the distinct social and political value of scientific authority may not necessarily undermine the impact of scientific knowledge on the ways states and governments act. What appears against the past as the decline of scientific authority in the larger social and political context may even be redescribed as a reconfiguration of the authority of expertise in the postmodern state. The earlier distinctions between science, politics, law, ethics, economy, and the like may have become irrelevant in complex contemporary societies, in which the use of scientific knowledge requires finer, more intricate, and continual adjustments and readjustments between science and other expressions of truth or power.

In the absence of hegemonic ideological or national frames, the macropolitical sphere of the state divides into a multitude of—often just temporary and amorphous—subgroups, each organized around a particular order of values and interests. In each of these particular universes scientific knowledge and expertise can occupy a privileged authority as a means to advance shared goals. When these groups clash or compete in the wider political arena, their experts tend to take sides and function as advocates. The state and its regulatory institutions enlist their own experts to back up their particular perspectives. Within this new, more open-ended and dynamic system, science is more explicitly and reflexively integrated into social economic and political values. In the context of the new pluralist polity, the authority and knowledge of science may therefore be less distinctly visible and less relevant to the symbolic, ideological, or cultural aspects of politics. At the same time, it is no less present in the organization, management, and the normative constitution of social order.

See also: Academy and Society in the United States: Cultural Concerns; Development and the State; Higher Education; History of Science: Constructivist Perspectives; Science and Industry; Science, Economics of; Science Funding: Asia; Science Funding: Europe; Science Funding: United States; Science,

Social Organization of; Science, Sociology of; Science, Technology, and the Military; Scientific Academies, History of; Scientific Academies in Asia; State and Society; Universities, in the History of the Social Sciences

Bibliography

Balough B 1991 *Chain Reaction: Expert Debate and Public Participation in American Commercial Nuclear Power 1945–1975*. Cambridge University Press, New York

Beck U 1992 *Risk Society: Towards a New Modernity*. Sage Publications, London

Dumont L 1986 *Essays on Individualism, Modern Ideology in Anthropological Perspective*. University of Chicago Press, Chicago

Eisenstadt S N 1999 *Paradoxes of Democracy: Fragility, Continuity and Change*. The Woodrow Wilson Center Press, Washington, DC

Ezrahi Y 1990 *The Descent of Icarus: Science and the Transformation of Contemporary Democracy*. Harvard University Press, Cambridge, MA

Ezrahi Y 1996 Modes of reasoning and the politics of authority in the modern state. In: Olson D R, Torrance N (eds.) *Modes of Thought: Explorations in Culture and Cognition*. Cambridge University Press, Cambridge, UK

Gibbons M, Nowotny H, Limoges C, Schwartzman S, Scott P, Trow M (eds.) 1994 *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. Sage, London

Guston D H 2000 *Between Politics and Science: Assuring the Integrity of Research*. Cambridge University Press, New York

Jasanoff S 1990 *The Fifth Branch: Science Advisers as Policymakers*. Harvard University Press, Cambridge, MA

Larson M S 1977 *The Rise of Professionalism*. University of California Press, Berkeley, CA

Leslie S W 1993 *The Cold War and American Science*. Columbia University Press, Cambridge, MA

Manuel F E 1974 *The Religion of Isaac Newton*. Clarendon Press, Oxford, UK

Merton R K 1973 The normative structure of science. In: Storer N W (ed.) *The Sociology of Science*. University of Chicago Press, Chicago

Mosse G L 1966 *Nazi Culture*. Schoken Bookes, New York

Polanyi M 1958 *Personal Knowledge: Towards a Post Critical Philosophy*. University of Chicago Press, Chicago

Polanyi M 1962 *The Republic of Science: Its Political and Economic Theory*. Minerva Press, pp. 54–73

Porter T M 1986 *The Rise of Statistical Thinking 1820–1900*. Princeton University Press, Princeton, NJ

Porter T M 1995 *Trust In Numbers, The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ

Price D K 1965 *The Scientific Estate*. Belknap, Cambridge, MA

Scott J C 1998 *Seeing Like a State*. Yale University Press, New Haven, CT

Sen A K 1981 *Poverty and Famines: An Essay on Entitlements and Deprivation*. Clarendon Press, Oxford, UK

Shapin S, Schaffer S 1985 *Leviathan and the Air Pump: Hobbes, Boyle and the Experimental Life*. Princeton University Press, Princeton, NJ

Skinner Q 1996 *Reason and Rhetoric in the Philosophy of Hobbes*. Cambridge University Press, Cambridge, UK

Solingen E (ed.) 1994 *Scientists and The State: Domestic Structures and the International Context*. University of Michigan Press, Ann Arbor, MI

Sprat Th [1667] 1958 *History of the Royal Society*. Washington University Press, St Louis, MO

Tocqueville A de 1945 Democracy in America. In: Bradley Ph (ed.) Vintage Books, New York

Wagner P, Weiss C H, Wittrock B, Wollman H 1991 *Social Sciences and Modern States*. Cambridge University Press, Cambridge, UK

Wilson J Q 1989 *Bureaucracy: What Government Agencies Do and Why They Do It*. Basic Books, New York

Y. Ezrahi

Science, Economics of

Determining the principles governing the allocation of resources to science as well as the management and consequences of the use of these resources are the central issues of the economics of science. Studies in this field began with the assumption that science was a distinct category of public spending that required rationalization. They have moved towards the view that science is a social system with distinct rules and norms. As the systems view has developed, the focus of the economics of science has moved from the effects of science on the economy to the influence of incentives and opportunities on scientists and research organizations.

There is a productive tension between viewing science as a social instrument and as a social purpose. In the first view, science is a social investment in the production and dissemination of knowledge that is expected to generate economic returns as this knowledge is commercially developed and exploited. This approach has the apparent advantage that the standard tools of economic analysis might be directly employed in choosing how to allocate resources to science and manage their use. In the second approach, science is assumed to be a social institution whose norms and practices are distinct from, and only partially reconcilable with, the institutions of markets. While this second approach greatly complicates the analysis of resource allocation and management, it may better represent the actual social organization of science and the behavior of scientists, and it may therefore ultimately produce more effective rules for resource allocation and better principles for management. Both approaches are examined in this article, although it is the first that accounts for the majority of the economics of science literature (Stephan 1996).

1. *The Economic Analysis of Science as a Social Instrument*

In arguing for a continuing high rate of public funding of science following World War II, US Presidential

Science Advisor Vannevar Bush (1945) crafted the view that science is linked intrinsically to technological and economic progress as well as being essential to national defense. The aim of 'directing' science to social purposes was already well recognized, and had been most clearly articulated before the war by John D. Bernal (1939). What distinguished Bush's argument was the claim that science had to be curiosity-driven and that, in companies, such research would be displaced by the commercial priorities of more applied research. The view that science is the wellspring of economic growth became well established within the following generation, giving rise to statements like 'Basic research provides most of the original discoveries from which all other progress flows' (United Kingdom Council for Scientific Policy 1967).

The concept of science as a source of knowledge that would be progressively developed and eventually commercialized became known as the 'linear model.' In the linear model, technology is science reduced to practical application. The 'linear model' is an oversimplified representation that ignores the evidence that technological change is often built upon experience and ingenuity divorced from scientific theory or method, the role of technological developments in motivating scientific explanation, and the technological sources of instruments for scientific investigation (Rosenberg 1982). Nonetheless, it provides a pragmatic scheme for distinguishing the role of science in commercial society.

If science is instrumental in technological progress and ultimately economic growth and prosperity, it follows that the economic theory of resource allocation should be applicable to science. Nelson (1959) and Arrow (1962) demonstrated why market forces could not be expected to generate the appropriate amount of such investment from a social perspective. Both Arrow and Nelson noted that in making investments in scientific knowledge, private investors would be unable to capture all of the returns to their investment because they could not charge others for the use of new scientific discoveries, particularly when those discoveries involved fundamental understanding of the natural world. Investment in scientific knowledge therefore had the characteristics of a 'public good', like publicly accessible roads. This approach established a basis for justifying science as a public investment. It did not, however, provide a means for determining what the level of that investment should be.

Investments in public goods are undertaken, in principle, subject to the criterion that benefits exceed costs by an amount that is attractive relative to other investments of public funds. To employ this criterion, a method for determining the prospective returns or benefits from scientific knowledge is required. The uncertainty of scientific outcomes is not, in principle, a fundamental barrier to employing this method. In practice, it is often true that the returns from investments in public good projects are uncertain, and prospective returns often involve attributing to new projects the returns from historical projects. Griliches (1958) pioneered a methodology for retrospectively assessing the economic returns on research investment, estimating that social returns of 700 percent had been realized in the period 1933–55 from the $2 million of public and private investments on the development of hybrid corn from 1910–55. Other studies of agricultural innovation as well as a limited number of studies of industrial innovation replicated Griliches' findings of a high social rate of return (see Steinmueller 1994 for references). Mansfield (1991) provides a fruitful approach for continuing to advance this approach. Mansfield asked R&D executives to estimate the proportion of their company's products and processes commercialized in 1975–85 that could not have been developed, or would have been substantially delayed, without academic research carried out in the preceding 15 years. He also asked them to estimate the 1985 sales of the new products and cost savings from the new processes. Extrapolating the results from this survey to the total investment in academic research and the total returns from new products and processes, Mansfield concluded that this investment had produced the (substantial) social rate of return of 28 percent.

The preceding discussion could lead one to conclude that the development of a comprehensive methodology for assessing the rate of return based on scientific research was only a matter of greater expenditure on economic research. This conclusion would be unwarranted. Efforts to trace the returns from specific government research efforts (other than in medicine and agriculture) have been less successful. The effort by the US Department of Defense Project Hindsight to compute the returns from defense research expenditures not only failed to reveal a positive rate of return, but also rejected the view that 'any simple or linear relationship exists between cost of research and value received' (Office of the Director of Defense Research and Engineering 1969). Similar problems were experienced when the US National Science Foundation sought to trace the basic research contributions underlying several major industrial innovations (National Science Foundation 1969). In sum, retrospective studies based on the very specific circumstances of 'science enabled' innovation or upon much broader claims that science as a whole contributes a resource for commercial innovation seem to be sustainable. When these conditions do not apply, as in the cases of specific research programs with uncertain application or efforts to direct basic research to industrial needs, the applicability of retrospective assessment, and therefore its value for resource allocation policy, is less clear.

More fundamentally, imputing a return to investments in scientific research requires assumptions about the 'counter-factual' course of developments that

would have transpired in the absence of specific and identified contributions of science. In examples like hybrid corn or the poliomyelitis vaccine, a reasonable assumption about the 'counter-factual' state of the world is a continuation of historical experience. Such assumptions are less reasonable in cases where scientific contributions enable a particular line of development but compete with alternative possibilities or where scientific research results are 'enabling' but are accompanied by substantial development expenditures (David et al. 1992, Mowery and Rosenberg 1989, Pavitt 1993).

For science to be analyzed as a social instrument, scientific activities must be interpreted as the production of information and knowledge. As the results of this production are taken up and used, they are combined with other types of knowledge in complex ways for which the 'linear model' is only a crude approximation. The result is arguably, and in some cases measurably, an improvement in economic output and productivity. The robustness and reliability of efforts to assess the returns to science fall short of standards that are employed in allocating public investment resources. Nonetheless, virtually every systematic study of the contribution of science to economy has found appreciable returns to this social investment. The goals of improving standards for resource allocation and management may be better served, however, by analyzing science as a social institution.

2. *Science as a Social Institution*

The economic analysis of science as a social system begins by identifying the incentives and constraints that govern the individual choices of scientists and this may reflect persistent historical features of science or contemporaneous policies. Incentives may include tangible rewards such as monetary awards, intangible, but observable, rewards such as status, and less observable rewards such as personal satisfaction. Similarly, constraints should be interpreted broadly, including not only financial limitations but also constraints stemming from institutional rules, norms, and standards of practice. The following simplified account suggests one of several ways of assembling these elements into a useful analytical framework.

Becoming a scientist requires substantial discipline and persistence in educational preparation as well as skills and talents that are very difficult to assess. Scientific training may be seen as a filter for selecting from prospective scientists those who have the ability and drive to engage in a scientific career. In addition, the original work produced during research training demonstrates the capacity of the researcher and provides a means for employers to assess the talents of the researcher (David 1994). Analyzing science education as an employment filter is a complement to more traditional studies of the scientific labor market such as those reviewed by Stephan (1996). The employment filter approach may also waste human resources by making schooling success the only indicator of potential for scientific contribution. If, for example, the social environment of the school discourages the participation or devalues the achievement of women or individuals from particular ethnic groups, the filter system will not perform as a meritocracy.

The distinctive features of science as a social system emerge when considering the incentives and constraints facing employed scientists. Although there is a real prospect of monetary reward for outstanding scientific work (Zuckerman 1992), many of the incentives governing scientific careers are related to the accumulation of professional reputation (Merton 1973). While Merton represented science as 'universalist' (open to claims from any quarter), the ability to make meaningful claims requires participation in scientific research networks, participation that is constrained by all of the social processes that exclude individuals from such social networks or fail to recognize their contribution. The incentive structure of seeking the rewards from professional recognition, and the social organization arising from it, is central to the 'new economics of science' (Dasgupta and David 1994).

The new economics of science builds upon sociological analyses (Cole and Cole 1973, Merton 1973, Price 1963) of the mechanisms of cumulative reinforcement and social reward within science. From an economic perspective, the incentive structure governing science is the result of the interactions between the requirement of public disclosure and the quest for recognition of scientific 'priority', the first discovery of a scientific result. Priority assures the alignment of individual incentives with the social goal of maximizing the scientific knowledge base (Dasgupta and David 1987). Without the link between public disclosure and the reward of priority, it seems likely that scientists would have an incentive to withhold key information necessary for the further application of their discoveries (David et al. 1999).

As Stephan (1996) observes, the specific contribution of the new economics of science is in linking this incentive and reward system to resource allocation issues. Priority not only brings a specific reward of scientific prestige and status but also increases the likelihood of greater research support. Cumulative advantage therefore not only carries the consequence of attracting attention, it also enables the recruitment of able associates and students and provides the means to support their research. These effects are described by both sociologists of science and economists as the Matthew effect after Matthew 25:29, 'For to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away.' As in the original parable, it

may be argued that this allocation is appropriate since it concentrates resources in hands of those who have demonstrated the capacity to produce results.

The race to achieve priority and hence to collect the rewards offered by priority may, however, lead to inappropriate social outcomes because priority is a 'winner take all' contest. Too many resources may be applied to specific races to achieve priority and too few resources may be devoted to disseminating and adapting scientific research results (Dasgupta and David 1987, David and Foray 1995), a result that mirrors earlier literature on patent and technology discovery races (Kamien and Schwartz 1975). Moreover, the mechanisms of cumulative advantage resulting from achieving priority may reduce diversity in the conduct of scientific research. This system has the peculiarity that the researchers who have the greatest resources and freedom to depart from existing research approaches are the same ones who are responsible for creating the status quo.

The principal challenges to the view that science is a distinct social system are the growing number of scientific publications by scientists employed in private industry (Katz and Hicks 1996) and the argument that scientific knowledge is tightly bound to social networks (Callon 1994). Private investments in scientific research would appear to question the continuing validity of the 'public good' argument. For example, Callon (1994) contends that scientific results are, and have always been, strongly 'embedded' within networks of researchers and that 'public disclosure' is therefore relatively useless as a means of transfer for scientific knowledge. Gibbons et al. (1994) argue that research techniques of modern science have become so well distributed that public scientific institutions are no longer central to scientific activity.

While the arguments of both Callon and Gibbons et al. suggest that private scientific research is a direct substitute for publicly funded research, other motives for funding and publication such as gaining access to scientific networks suggest that public and private research are complementary (David et al. 1999). The growing reliance of industry on science provides a justification for investing in science to improve the 'absorption' of scientific results (Cohen and Levinthal 1989, Rosenberg 1990). Employed scientists need to be connected with other scientific research colleagues who identify 'membership' in the scientific community with publication, and labor force mobility for employed scientists requires scientific publication. Thus, it is premature to conclude that the growing performance of scientific research in industry or publication of scientific results by industrial authors heralds the end of the need for public support of science.

The growing significance of private funding of scientific research does, however, indicate the need to improve the socioeconomic analysis of the incentive and governance structures of science. Empirical work on the strategic and tactical behavior of individual scientists, research groups, and organizations is urgently needed to trace the implications of the changing environment in which the social institutions of science are evolving. Ultimately, these studies should be able to meet the goal of developing better rules for allocating and managing the resources devoted to science.

See also: Innovation, Theory of; Research and Development in Organizations; Research Funding: Ethical Aspects; Science and the State; Science Funding: Asia; Science Funding: Europe; Science Funding: United States; Science, Technology, and the Military

Bibliography

Arrow K J 1962 Economic Welfare and the Allocation of Resources for Invention. *The Rate and Direction of Inventive Activity*. National Bureau of Economic Research, Princeton University Press, Princeton, NJ

Bernal J D 1939 *The Social Function of Science*. MIT Press, Cambridge MA. MIT Press Paperback edition 1967

Bush V 1945 *Science: The Endless Frontier: A Report to the President on a Program for Postwar Scientific Research*. United States Office of Scientific Research and Development, Washington DC. National Science Foundation reprint 1960

Callon M 1994 Is Science a Public Good? Fifth Mullins Lecture, Virginia Polytechnic Institute, 23 March 1993. *Science, Technology and Human Values* **19**: 395–424

Cohen W M, Levinthal D A 1989 Innovation and learning: The two faces of R&D. *Economic Journal* **99**(397): 569–96

Cole J R, Cole S 1973 *Social Stratification in Science*. University of Chicago Press, Chicago, IL

Dasgupta P, David P A 1987 Information disclosure and the economics of science and technology. In: Feiwel G R (ed.) *Arrow and the Ascent of Modern Economic Theory*. New York University Press, New York pp. 519–40

Dasgupta P, David P A 1994 Toward a new economics of science. *Research Policy* **23**(5): 487–521

David P A 1994 Positive feedbacks and research productivity in science: reopening another black box. In: Granstrand O (ed.) *Economics of Technology*. North Holland, Amsterdam and London, pp. 65–89

David P A, Foray D 1995 Accessing and expanding the science and technology knowledge base. *Science and Technology Industry Review* **16**: 13–68

David P A, Foray D, Steinmueller W E 1999 The research network and the new economics of science: From metaphors to organizational behaviours. In: Gambardella A, Malerba F (eds.) *The Organization of Economic Innovation in Europe*. Cambridge University Press, Cambridge, UK pp. 303–42

David P A, Mowery D, Steinmueller W E 1992 Analysing the economic payoffs from basic research. *Economics of Innovation and New Technology* **2**: 73–90

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M 1994 *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. Sage, London

Griliches Z 1958 Research costs and social returns: hybrid corn and related innovations. *Journal of Political Economy* (October): 419–31

Kamien M I, Schwartz N L 1975 Market structure and innovation: a survey. *Journal of Economic Literature* **13**(1): 1–37

Katz J S, Hicks D M 1996 A systemic view of British science. *Scientometrics* **35**(1): 133–54
Mansfield E 1991 Academic research and industrial innovation. *Research Policy* **20**(1): 1–12
Merton R 1973 *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press, Chicago, IL
Mowery D C, Rosenberg N 1989 *Technology and the Pursuit of Economic Growth*. Cambridge University Press, Cambridge, UK
National Science Foundation 1969 *Technology in Retrospect and Critical Events in Science* (*TRACES*). National Science Foundation, Washington, DC
Nelson R R 1959 The simple economics of basic scientific research. *Journal of Political Economy* **67**(June): 297–306
Office of the Director of Defense Research and Engineering 1969 Project Hindsight: Final Report. Washington, DC
Pavitt K 1993 What do firms learn from basic research. In: Foray D, Freeman C (eds.) *Technology and the Wealth of Nations: The Dynamics of Constructed Advantage*. Pinter Publishers, London
Price D J, de Solla 1963 *Little Science*, *Big Science*. Columbia University Press, New York
Rosenberg N 1982 *Inside the Black Box: Technology and Economics*. Cambridge University Press, Cambridge, UK
Rosenberg N 1990 Why do firms do basic research (with their own money). *Research Policy* **19**(2): 165–74
Steinmueller W E 1994 Basic science and industrial innovation. In: Dodgson M, Rothwell R (eds.) *Handbook of Industrial Innovation*. Edward Elgar, London pp. 54–66
Stephan P E 1996 The economics of science. *Journal of Economic Literature* **XXXIV**(September): 1199–235
United Kingdom Council for Scientific Policy (Great Britain) 1967 Second Report on Science Policy. HMSO, London
Zuckerman H 1992 The proliferation of prizes: Nobel complements and Nobel surrogates in the reward system of science. *Theoretical Medicine* **13**: 217–31

W. E. Steinmueller

Science Education

The launch of Sputnik in 1957 almost single-handedly conferred on science education the status of an Olympic sport. A series of international comparison studies along with the increasing importance of science and technology in the global economy have nurtured this image (e.g., Schmidt et al. 1997). Stakeholders, including policy makers, natural scientists, textbook publishers, test designers, classroom teachers, business leaders, and pedagogical researchers, have responded with competing strategies for improving science education. In some countries, most notably the United States, powerful groups have mandated untested policies and unachievable educational goals—often driven by a single aspect of the problem such as textbooks, assessments, curriculum frameworks, peer learning, or technological innovation.

Experience with failed policy initiatives and ambiguous research findings has highlighted the systemic, intricate, complex, interconnected nature of science education. Innovations succeed under some circumstances, but not others. Students regularly fail to learn what they are taught, prefer ideas developed from personal experiences, and make inferences based on incomplete information. Curriculum designers often neglect these aspects of learners and expect students to absorb all the information in a textbook or learn to criticize research on ecology by studying logic puzzles (Linn and Hsi 2000, Pfundt and Duit 1991).

International studies of science learning in over 50 countries raise questions about the complex relationships between achievement and curriculum, or between interest in science and science learning, or between teachers' science knowledge and student success that many have previously taken for granted. This systemic character of science education demands a more nuanced and contextual understanding of the development of scientific understanding and the design of science instruction, as well as new research methods, referred to as design studies, to investigate the impact of innovations.

1. Developing Scientific Understanding

1.1 Knowledge Integration

Most researchers would agree that science learners engage in a process of knowledge integration, making sense of diverse information by looking for patterns and building on their own ideas (Bransford et al. 1999, Bruer 1993). Knowledge integration involves linking and connecting information, seeking and evaluating new ideas, as well as revising and reorganizing scientific ideas to make them more comprehensive and cohesive.

Designing curriculum materials to support and guide the process of knowledge integration has proven difficult. Most textbooks and even hands-on activities reflect a view of learners as absorbing information rather than attempting to integrate new ideas with their existing knowledge. Recent research provides guidance to curriculum designers by describing the interpretive, cultural, and deliberate dimensions of knowledge integration. Learners interpret new material in light of their own ideas and experiences, frequently relying on personal perspectives rather than instructed ideas. For example, science learners often believe that objects in motion come to rest, based on their extensive observations of the natural world. Learning happens in a cultural context where group norms, expectations, and supports shape learner activity. For example, when confronted with the Newtonian view that objects in motion remain in motion, many learners conclude that objects come to rest on the playground but remain in motion in science class, invoking separate norms for these distinct contexts. Individuals make deliberate decisions about their science learning, develop commitments about reusing what they learn, pay attention to some science debates

but not others, and select or avoid science as a career. For example, some students desire a cohesive account of topics like motion and seek to explain new phenomena such as the role of friction in nanotechnology, while others report with pride that they have 'forgotten everything taught in science class.' The interpretive, cultural, and deliberate dimensions of science learning apply equally to pre-college students, preservice and in-service science teachers, research partnerships, and lifelong science learners. These perspectives help clarify the nature of knowledge integration and suggest directions for design of science instruction.

1.2 Interpretive Nature of Science Learning

Learners develop scientific expertise by interpreting the facts, processes, and inquiry skills they encounter in terms of their own experiences and ideas. Experts in science develop richly connected ideas, patterns, and representations over the years and regularly test their views by interpreting complex situations, looking for anomalies, and incorporating new findings. For example, expert physicists use free body diagrams to represent mechanics problems while novices rely on the formulas they learn in class.

Piaget (1971) drew attention to the ideas that students bring to science class such as the notion that the earth is round like a pancake, or that heavier objects displace more volume. Piaget offered assimilation and accommodation as mechanisms to account for the process of knowledge integration. Vygotsky (1962) distinguished spontaneous ideas developed from personal experience such as the view that heat and temperature are the same from the instructed distinction between heat and temperature. Recent research calls for a more nuanced view of knowledge integration by showing that few learners develop a coherent perspective on scientific phenomena: most students develop 'knowledge in pieces' and retain incohesive ideas in their repertoire (diSessa 2000). Even experts may give incohesive views of phenomena when asked to explain at varied levels of granularity. For example, scientists may have difficulty designing a picnic container to keep food safe even when they have expert knowledge of molecular kinetic theory (Linn and Hsi 2000).

Science textbooks, lectures, films, and laboratory experiments often reinforce students' incoherent views of science; they offer disconnected, inaccessible ideas, and avoid complex personally-relevant problems. For example, many texts expect students to gain understanding of friction by analyzing driving on icy roads. But non-drivers and those living in warm climates find this example unfamiliar and inaccessible. Similarly, research on heat and temperature reveals that difficulties in understanding the particulate nature of matter stand in the way of interpreting the molecular-kinetic model. The engineering-based heat flow model offers a more descriptive and accessible account of many important aspects of heat and temperature such as wilderness survival or home insulation or thermal equilibrium.

Designing instruction to stimulate the interpretive process means carefully selecting new, compelling ideas to add to the views held by students and supporting students as they organize, prioritize, and compare these various ideas. This process of weighing alternative accounts of scientific phenomena can clash with student ideas about the nature of science and of science learning. Many students view science knowledge as established and science learning as memorization. To promote knowledge integration, students need to interpret dynamic examples of science in the making and they need to develop norms for their own scientific reasoning.

Students need a nuanced view of scientific investigation that contrasts methodologies and issues in each discipline. General reasoning skills and critical thinking are not sufficient. Instead, students need to distinguish the epistemological underpinnings of methodologies for exploring the fossil record, for example, from those required for study of genetic engineering. They need to recognize pertinent questions for research on earthquake-resistant housing, DNA replication, and molecular modeling. Effective knowledge integration must include an understanding of the ethical and moral dilemmas involved in diverse scientific areas, as well as the nature of scientific advances.

Research on how students make sense of science suggests some mechanisms to promote the interpretive process of knowledge integration. Clement (1991) calls for designing bridging analogies to help students make effective connections. For example, to help students understand the forces between a book and a table, Clement recommends comparing the experience of placing the book on a spring, a sponge, a mattress, and possibly, releasing it on air. Linn and Hsi (2000) call on designers to create pivotal cases that enable learners to make subtle connections and reconsider their views. For example, a pivotal scientific visualization of the relative rates of heat flow in different materials helped students interpret their personal observations that, at room temperature, metals feel colder than wood. To help students sort out alternative experiences, observations, school ideas, and intuitions, research shows the benefit of encouraging students to organize their knowledge into larger patterns, motivating students to critique alternative ideas, and establishing a taste for cohesive ideas.

1.3 Cultural Context of Science Learning

All learning occurs in a cultural context where communities respond to competing perspectives, confer status on research methods, and establish group norms and expectations. Students benefit from the cultural

context of science, for example, when they find explanations provided by peers more comprehensible than those in textbooks or when peers debate alternative views. Expert scientists have well established mechanisms for peer review, norms for publications, and standards for inquiry practices. Group norms are often institutionalized in grant guidelines, promotion policies, and journal publication standards. The cultural context of the classroom has its own characteristics that may or may not support and encourage the development of cohesive, sustained inquiry about science. Students may enact cultural norms that exclude those from groups underrepresented in science from the discourse (Wellesley College Center for Research on Women 1992, Keller 1983) or limit opportunities for students to learn from each other. Textbooks and standardized tests may privilege recall of information over sustained investigation. Images of science in curriculum materials and professional development programs may neglect societal or policy issues.

Contemporary scientific controversies, such as the current international debate about genetically modified foods, rarely become topics for science classes. As a result, students may develop flawed images of science in the making and lack ability to make a critical evaluation of news accounts of personally-relevant issues. For example, to make decisions about the cultivation and consumption of genetically-modified foods, students would ideally compare the risks from traditional agricultural practices such as hybridization to the risks from genetic modification. They would also weigh issues of economics, world hunger, and individual health. In addition, students studying this controversy would learn to distinguish comments from scientists supported by the agricultural industry, environmental protection groups, and government grants. Examining knowledge integration from a cultural perspective helps to clarify the universality of controversy in science and the advantages of creating classroom learning communities that illustrate the role of values and beliefs in scientific work (Brown 1992, Bransford and Brown 1999).

1.4 Deliberative Perspective on Science Learning

Students make deliberate decisions about science learning, their own progress, and their careers. Lifelong science learners deliberately reflect on their views, consider new accounts of scientific problems, and continuously improve their scientific understanding; they seek a robust and cohesive account of scientific phenomena. Designing instruction that develops student responsibility for science learning creates a paradox. In schools, science curriculum frameworks, standards, texts, and even recipe-driven hands-on experiences leave little opportunity for independence. Yet, the curriculum cannot possibly provide all the necessary information about science; instead, students need supervised practice in analyzing their own progress in order to guide their own learning.

Many instructional frameworks offer mechanisms leading to self-guided, intentional learning (Linn and Hsi 2000, White and Frederiksen 1998). Vygotsky (1962) drew attention to creating a 'zone of proximal development' by designing accessible challenges so that students, supported by instruction and peers, could continue to engage in knowledge integration. Vygotsky argued that, when students encounter new ideas within their zone of proximal development and have appropriate supports, they can compare and analyze spontaneous and instructed ideas, achieve more cohesive understandings, and expand their zone of proximal development. With proper support students can even invent and refine representations of their experimental findings (diSessa 2000). Others have shown that engaging students in guided reflection, analysis of their own progress, and critical review of their own or others, arguments establishes a more deliberative stance towards science learning (Linn and Hsi 2000, White and Frederikson 1998).

2. Designing Science Instruction

Design of science instruction occurs at the level of state and national standards, curriculum frameworks for science courses, materials such as textbooks or software, and activities carried out by both students and teachers. In many countries, a tension has emerged between standards that mandate fleeting coverage of a list of important scientific topics and concerns of classroom teachers that students lack opportunity to develop a disposition towards knowledge integration and lifelong science learning. As science knowledge explodes, citizens face increasingly complex science-related decisions, and individuals need to regularly update their workplace skills. To make decisions about personal health, environmental stewardship, or career advancement, students need a firm foundation in scientific understanding, as well as experience interpreting complex problems. To lead satisfying lives, students need to develop lifelong learning skills that enable them to revisit and refine their ideas and to guide their own science learning in new topic areas.

Recent research demonstrates the need to design and test materials to be sure they are promoting knowledge integration and setting learners on a path towards lifelong learning. Frameworks for design of science instruction for knowledge integration call for materials and activities that feature accessible ideas, make thinking visible, help students learn from others, and encourage self-monitoring (Bransford et al. 2000, Linn and Hsi 2000, White and Frederiksen 1998).

2.1 Designing Accessible Ideas

To promote knowledge integration, students need a designed curriculum that includes pivotal cases and

bridging analogies to help them learn. Rather than asking experts to identify the most sophisticated ideas, designers need to select the most accessible and generative ideas to add to the mix of student views. College physics courses generally start with Newton rather than Einstein; in pre-college courses one might start with everyday examples from the playground rather than the more elegant but less understandable frictionless problems.

To make the process of lifelong knowledge integration accessible, students need some experience with sustained, complex inquiry. Carrying out projects such as developing a recycling plan for a school or researching possible remedies for the worldwide threat of malaria, engage students in the process of scientific inquiry and can establish lifelong learning skills. Often, however, science courses neglect projects or provide less knowledge integration intensive experiences such as a general introduction of critical thinking or hands-on recipes for solving unambiguous problems. By using computer learning environments to help guide students as the carry out complex projects curriculum designers can foster a robust understanding of inquiry (Driver et al. 1996).

Projects take instructional time, require guidance for individual students, bring the complexities of science to life, and depend on well-designed questions. Students often confuse variables such as food and appetite, rely on flawed arguments from advertisements or other sources, and flounder because they lack criteria for critiquing their own progress. For example, when students critique projects, they may comment on neatness and spelling, rather than looking for flaws in an argument. Many teachers avoid projects because they have not developed the pedagogical skills necessary to mentor students, deal with the uncertainties of contemporary science dilemmas, or design researchable questions. Research shows that computer learning environments can make complex projects more successful by scaffolding inquiry, providing help and hints, and freeing teachers to interact with students about complex science issues (Feurzeig and Roberts 1999, Linn and Hsi 2000, White and Frederiksen 1998).

2.2 Making Thinking Visible

Students, teachers, and technological tools can make thinking visible to model the process of knowledge integration, illustrate complex ideas, and motivate critical analysis of complex situations. Learning environments, such as WorldWatcher (http://www.worldwatcher.nwu.edu/), Scientists in Action (http://peabody.Vanderbilt.edu), and the Web-Based Integrated Science Environment (WISE—http://wise.berkeley.edu) guide students in complex inquiry and make thinking visible with scientific visualizations. In these projects, students successfully debate the causes of frog deformities, evaluate the water quality in local streams, and design houses for desert climates. Research demonstrates benefits of asking students to make their thinking visible in predictions, reflections, assessments of their progress, and collaborative debate.

Technological learning environments also make student ideas visible with embedded performance assessments that capture ability to critique arguments, make predictions, and reach conclusions. Such assessments help teachers and researchers identify how best to improve innovations and provide an alternative to high stakes assessments (Heubert and Hauser 1998, Bransford et al. 1999).

2.3 Helping Students Learn from Others

When students collaboratively investigate science problems, they can prompt each other to reflect, provide explanations in the language of their peers, negotiate norms for critiques of arguments, and specialize in specific aspects of the problem. Several programs, including Kids as Global Scientists (http://www.onesky.umich.edu/) and Project Globe (http://www.globe.gov/), orchestrate contributions from students around the world to track and compare extreme weather.

Technological learning environments support global collaborations as well as collaborative debate and equitable discussion. In collaborative debate (see Science Controversies On-Line: Partnerships in Education, SCOPE, http://scope.educ.washington.edu/), students research contemporary controversies on the Internet with guidance from a learning environment, prepare their arguments often using visual argument representations, and participate in a classroom debate where every student composes a question for each presenter. Teachers have a rich sample of student work to use for assessment. Online scientific discussions engage many more students than do class discussions and also elicit more thoughtful contributions (Linn and Hsi 2000).

2.4 Promoting Autonomy and Lifelong Learning

New pedagogical practices often implemented in computer learning environments can nudge students towards deliberate, self-guided learning. Projects with personally-relevant themes capitalize on the intentions of students and motivate students to revisit ideas after science class is over. For example, students who studied deformed frogs brought news articles to their teacher years after the unit was completed. Projects can offer students a rich context, such as the rescue of an endangered animal or the analysis of the earthquake safety of their school, that raises investment in science learning.

Students need to monitor their own learning to deal with new sources of science information, such as the Internet, where persuasive messages regularly appear along with public service announcements. Helping students jointly form partnerships where multiple forms of expertise are represented, develop common norms and criteria for evaluating arguments, and deliberately review their progress prepares for situations likely to occur later in life.

3. Research Methods

Researchers have responded to the intricate complexities in science education with new research methods informed by practices in other design sciences, including medicine and engineering. In the design sciences, researchers create innovations like science curricula, drugs, or machines and study these innovations in complex settings. In education, these innovations, such as technology enhanced science projects, build on increased understanding of science learning.

Research in a design science is typically directed by a multi-disciplinary partnership. In education, partners bring expertise in a broad range of aspects of learning and instruction, including technology, pedagogy, the science disciplines, professional development, classroom activity structures, and educational policy; collaborators often have to overcome perceptions of status differences among the fields represented. By working in partnership, individuals with diverse forms of expertise can jointly contribute to each others' professional development.

Practices for design studies come from design experiments and Japanese lesson study (Brown 1992, diSessa 2000, Lewis 1995). Design studies typically start when the partnership creates an innovation, such as a new learning environment, curriculum, or assessment and have as their goal the continuous improvement of the innovation. The inspiration for innovative designs can come from laboratory investigations, prior successes, spontaneous ideas, or new technologies. Many technology-enhanced innovations have incorporated elements that have been successful in laboratory studies, like one-on-one tutoring or collaborative learning. Other innovations start with scientific technologies such as real time data collection (Bransford et al. 1999).

In design studies, partners co-design innovations and evaluations following the same philosophy so that assessments are sensitive to the goals of instruction. Partners often complain that standardized, multiple choice tests fail to tap progress in knowledge integration and complex scientific understanding. Results from assessment allow the partnership to engage in principled refinement of science instruction and can also inform future designers. Often, innovations become more flexible and adaptive as they are refined, making it easier for new teachers to tailor instruction to their students and curricular goals. Design studies may also be conducted by small groups of teachers engaging in continuous improvement of their instruction, inspired by the Japanese lesson study model. When teachers collaborate to observe each other, provide feedback, and jointly improve their practice, they develop group norms for design reviews.

Methodologies for testing innovations in complex settings and interpreting results include an eclectic mix of approaches from a broad range of fields, including classroom observations, video case studies, embedded assessments of student learning, student performance on design reviews, classroom tests, and standardized assessments, as well as longitudinal studies of students and teachers. When partnerships design rubrics for interpreting student work they develop a common perspective.

Design study research has just begun to address the complexities of science education. For example, current investigations reveal wide variation among teachers implementing science projects. In classrooms, some teachers spend up to five minutes with each group of students, while others spend less than one minute per group. In addition, some teachers, after speaking to a few small groups, recognize a common issue and communicate it to their whole class. This practice, while demanding for teachers, has substantial benefits. Developing sensitivity to student dilemmas when complex projects are underway requires the same process of knowledge integration described above and may only emerge in the second and subsequent uses of the project. The design study approach to science instruction succeeds when teachers and schools commit to multiple refinements of the same instructional activities, rather than reassigning teachers and selecting new curriculum materials annually.

4. Emerging Research Topics and Next Steps

Emerging areas for science education research include professional development and policy studies. The interpretive, cultural, and deliberate character of learning applies equally to these fields. Teachers may value lifelong science learning but have little experience connecting the science in the curriculum to their own spontaneous ideas and few insights into how to support this process in students.

Teachers taking a knowledge integration approach to instruction face complex questions such as whether to introduce genetics by emphasizing genotypes, phenotypes, the human genome, or a treatment perspective. They may discover that some students come to class believing that two unrelated people who look alike could be twins. Effective professional development should help teachers with these sorts of specific questions rather than providing general

grounding in the science discipline or glitzy experiments that confuse rather than inform students. Inspired by the Japanese lesson study approach, more and more research groups are convening and studying collaborative groups of science teachers who face similar instructional decisions. Like science students, science teachers need opportunities and encouragement to build on their spontaneous ideas about learning, instruction, and the nature of science to become proficient in guiding their own professional development.

Science policy-makers may also be viewed through this lens of interpretive, cultural, and deliberate science learning. Policy-makers frequently hold ideas about science learning that might be at odds with those held by teachers. The status differences between policy-makers, natural scientists, and classroom teachers can interfere with open and effective communication. Building a cohesive perspective on pedagogy, science teaching, and science learning has proven very difficult in almost every country. The popularity of high-stakes assessments that might be insensitive to innovation underscores the dilemma. Recent news reports that high-stakes assessments are motivating absenteeism, cheating, and unpromising teaching practices increases the problem (Heubert and Hauser 1998).

Science educators face many complex, pressing problems including the connection between science and technology literacy; the dual goals of excellence and equitable access to science careers; the tradeoffs between a focus on public understanding of science and career preparation; the role and interpretation of high stakes tests; and the challenges of balancing the number of topics in the curriculum with the advantages of science project work. If we form a global partnership for science education and jointly develop a cohesive research program on lifelong learning, we have an unprecedented opportunity to collaboratively design and continuously improve teaching, instruction, and learning.

See also: Discovery Learning, Cognitive Psychology of; Gender and School Learning: Mathematics and Science; Scientific Concepts: Development in Children; Scientific Reasoning and Discovery, Cognitive Psychology of; Teaching and Learning in the Classroom; Teaching for Thinking

Bibliography

Bransford J, Brown A L, Cocking R R, National Research Council (US) 1999 *How People Learn: Brain, Mind, Experience, and School*. National Academy Press, Washington DC

Brown A 1992 Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of Learning Sciences* **2**(2): 141–78

Bruer J T 1993 *Schools for Thought: A Science of Learning in the Classroom*. MIT Press, Cambridge, MA

Clement J 1991 *Non-formal Reasoning in Science: The Use of Analogies, Extreme Cases, and Physical Intuition*. Lawrence Erlbaum Associates, Mahwah, NJ

diSessa A 2000 *Changing Minds: Computers, Learning, and Literacy*. MIT Press, Cambridge, MA

Driver R, Leach J, Millar R, Scott P 1996 *Young People's Images of Science*. Open University Press, Buckingham, UK

Feurzeig W, Roberts N (eds.) 1999 *Modeling and Simulation in Science and Mathematics Education*. Springer, New York

Heubert J, Hauser R (eds.) 1998 *High Stakes: Testing for Tracking, Promotion, and Graduation*. National Academic Press, Washington, DC

Keller E F 1983 *Gender and Science*. W H Freeman, San Francisco, CA

Lewis C 1995 *Educating Hearts and Minds: Reflections on Japanese Preschool and Elementary Education*. Cambridge University Press, New York

Linn M C, Hsi S 2000 *Computers, Teachers, Peers: Science Learning Partners*. Lawrence Erlbaum Associates, Mahwah, NJ

Piaget J 1971 *Structuralism*. Harper & Row, New York

Pfundt H, Duit R 1991 *Students' Alternative Frameworks*, 3rd edn. Institute for Science Education at the University of Kiel/Institut für die Pädagogik der Naturwissenschaften. Kiel, Germany

Schmidt W H, Raizen S A, Britton E D, Bianchi L J, Wolfe R G 1997 *Many Visions, Many Aims: A Cross-national Investigation of Curricular Intentions in School Science*. Kluwer, Dordrecht, Germany

Vygotsky L S 1962 *Thought and Language*. MIT Press, Cambridge, MA

Wellesley College Center for Research on Women 1992 *How Schools Short Change Girls*. American Association of University Women Education Foundation, Washington, DC

White B Y, Frederiksen J R 1998 Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction* **16**(1): 3–118

M. C. Linn

Science Funding: Asia

Science funding refers to national expenditure, from both public and private sources, for the institutionalization and promotion of a variety of scientific activities conventionally termed research and development (R&D). These may take the form of basic, applied, and development research undertaken or sponsored across a range of science and technology (S&T) institutions in national S&T innovation systems. In the postwar era, the concept of science funding assumed considerable importance in the national policies of Asian countries. The government, both as a patron and as a powerful mediator, played a significant part in shaping the structure and direction of science funding.

In the latter half of the twentieth century, the belief in science as a symbol of 'progress' was transformed into an established policy doctrine in the Asian region.

Creating wealth from knowledge and achieving social, political, and military objectives came to be closely associated with deliberately fostering S&T activities through funding scientific research. In the mid 1990s, the Asian region accounted for 26.9 percent of the world's gross expenditure on research and development (GERD), which was US$470 billions in 1994. While Japan and newly industrializing countries (NICs) accounted for 18.6 percent, South East Asia, China, India, and other South Asian countries accounted for 0.9 percent, 4.9 percent, 2.2 percent, and 0.3 percent respectively. Within the Asian region, whereas Japan and NICs accounted for about 69 percent, other countries accounted for 31 percent of total science funding.

Three main interrelated approaches underlie the concept of science funding in Asian countries. The first approach underlines the importance and essential nature of public or government funding of R&D as a 'public good' that improves the general interest and welfare of the society or nation as a whole. The public good approach emphasizes the funding and generation of knowledge that is non-competitive, open for public access, and non-exclusive. As private firms and the market are unable to capture all of the benefits of scientific research, they often tend to under-invest in R&D, which is also a key component of the process of innovation. It is in this context that public funding becomes important by bearing the social costs of generating scientific knowledge or information. Further, numerous studies have shown that public funding of scientific research as a public good yields several social and economic benefits (see Arrow 1962, Pavitt 1991). In contrast to institutionalized and 'codified' forms of knowledge, scientific capacity in the form of 'tacit knowledge' is seen to be embodied in research personnel trained over a period of time. Thus, publicly supported training in higher educational settings and networks of professional associations and academies forms an important component of science funding.

A closely related second approach to science funding is state sponsorship of military and defense related strategic R&D. Even though there is considerable spinoff from military R&D to the non-military sectors of economy, the main rationale for the importance given to such research is national security. Most countries in the world, including those in the Asian region, spend more money on military and defense-related scientific research activities than in the civilian R&D domain.

The third approach essentially recognizes the importance of private sources of funding for scientific research. Although private patrons have always played an important part in supporting S&T-related research activities, the postwar decades, particularly the 1980s and 1990s, witnessed a remarkable shift from public to private sources of funding of science. The significance of private science funding has grown since the 1980s, as the 'linear model,' which was based on the perceived primacy of basic research, began to lose its credibility. Insights from studies on the economics of innovation came to substantiate the view that it is 'downstream' development spending in the spectrum of R&D 'that plays a crucial role in determining who gets to capture the potential rents generated by scientific research' (Rosenberg 1991, p. 345). Private industrial firms, rather than government funding, dominate this segment of R&D in Asia as in other regions. Second the importance of private industry as a source of science funding attracted considerable attention from the mid 1980s, as the success of Japan and other East Asian NICs came to be analyzed from the perspective of how these countries transferred the burden of science funding from public to private sources.

Table 1
GERD by source of funding from the late 1970s (A) to the mid-1990s (B)

	Govt		Private industry		Other nat. sources	
	A	B	A	B	A	B
Japan	30	22	70	67	—	10
NICs	43	33	52	63	3	4
SE Asia	86	58	14	27	—	15
India	89	82	12	16	—	1
China	100	82	—	15	—	3
OSA	85	78	6	8	7	14

Source: World Science Report, Unesco, Paris, 1998; CAST ASIA II, 22–30 March 1982.
OSA: Other South Asian Countries include Bangladesh, Nepal, Myanmar, Pakistan, Sri Lanka, and Mongolia.
NICS include South Korea, Taiwan, Singapore, Hong Kong.
South East Asia includes Thailand, Philippines, Malaysia, Indonesia, and Vietnam.

As the expenditure on military R&D is largely met by governments, the pattern of GERD in the Asian region can be explored in terms of public and private sources. As Table 1 shows, three broad subregions are discernible in the Asian region insofar as the source of science funding is concerned. While private industry accounts for approximately two-thirds, government accounts for one-third of the total R&D expenditure in the case of Japan and NICs. There has been a significant transformation in the increase of private funding of science in South Korea. Private sources accounted for about 2.3 percent of GDP, that is, 80 percent of total R&D funds, which is one of the highest levels in the world. In Japan, Singapore, and Taiwan, over 65 percent of total R&D spending is financed by private industry. Much of this transformation is the result of state mediation and institutional mechanisms which offer appropriate incentives for the private sector to invest in R&D.

In the second subregion of South East Asia, there has been a perceptible change in private industry's share of total R&D, which increased from 14 percent to 27 percent, whereas the government's share witness-

Table 2
GERD and military expenditure as a percentage of GDP

	GERD 1994	Military expenditure 1985	1995
Japan	3.02	1.00	1.00
NICs	1.60	5.45	4.05
SE Asia	0.50	2.95	2.20
India	0.80	3.50	2.40
China	0.50	4.90	2.30
Other Asian countries	0.38	4.60	3.50

Source: World Science Report, Unesco, Paris, 1998; World Development Reports 1998/1999; and 1999/2000.

ed a decline from 86 percent to 58 percent between the late 1970s and the mid-1990s. In Southern Asia, which comprises China, India, and other South Asian countries, the government continues to be the main patron for science funding. The private sector accounts for less than 16.5 percent. In China and India, while the proportion of private funding of R&D increased to 15 percent and 4.4 percent respectively during the late 1970s and the mid-1990s, the private sector is likely to play a more dominant role in the 2000s. This is because these countries are witnessing considerable foreign direct investment and economic restructuring which are fostering liberalization and privatization. Except in the case of the Philippines and Sri Lanka, where about a quarter of total R&D funding is met by foreign sources, foreign funding hardly plays a significant role in the Asian region as whole.

Although there is no established correlation between GERD as a percentage of GDP and economic growth rates, this indicator has assumed great significance in S&T policy discourse in the 1990s. Whereas the industrially advanced countries spend between 2 percent to 3 percent of GDP on R&D activities, middle-income countries such as the NICs and some others in the South East Asian region spend over 1.5 percent. The poorer developing countries spend less than 1 percent and the countries falling at the bottom of economic and human development rank index spend even less than 0.5 percent. This general trend is also largely borne out in Asia, as shown in Table 2. However, the relation between different types or forms of R&D and economic growth has also come into sharper focus.

Even though the of share private industry in overall R&D effort is emerging as one of the determinants of economic dynamism, some studies draw attention to various national technological activities (in relation to the national science base) which have a direct bearing on national economic performance measured in terms of productivity and export growth (see Pavitt 1998, p. 800). For instance, as shown in Table 3, the most economically dynamic NICs such as Taiwan and South Korea began with relatively weak scientific strength, as did India in the early 1980s. Nevertheless, they outperformed India by the early 1990s in terms of the change in their share of world publications, as well as in the share of registered patents in USA. Taiwan and South Korea not only spent a much higher proportion on R&D as a percentage of GDP, but the privately dominated R&D structure was such that it gave high priority to patenting. In actual terms, while Japan and South Korea filed 335,061 and 68,446 patents in their respective countries, India filed only 1545 patents in 1995–6. Further, whereas high technology exports as a proportion of total manufacturing exports in Japan, South Korea, Hong Kong, and Singapore registered 39 percent, 39 percent, 27 percent, and 71 percent respectively, India registered hardly 10 percent in 1996, which is much lower than the figure for China (21 percent) for the same year.

As indicated in Table 2, there is a good deal of military burden in most of the Asian countries. Even though there was drastic reduction in military expenditure as a percentage of GDP between 1985 and 1995 in the region, all the Asian countries and subregions still spend three to four times more on military activities than on civilian R&D. While Pakistan maintained over 6 percent, Sri Lanka increased its military expenditure as a percentage of GDP from 2.9 percent to 4.6 percent between 1985 and 1995. Japan is the only country in the Asian region that limited its military expenditure to 1 percent of GDP while spending over three times this figure on civilian R&D (3.1 percent) during the same period.

Table 3
Trends in scientific and technological performance in selected Asian countries

County	Change in share of world publications, 1993/1982	Change in share of US patents, 1993/1983	Publications per million population, 1980–84
Taiwan	5.97	12.81	23.3
South Korea	5.45	29.79	8.00
Singapore	3.53	03.20	71.6
Hong Kong	2.37	02.42	45.90
India	0.83	02.45	18.10

Source: As given in Pavitt (1998, p. 800).

In terms of the sectoral focus of R&D funding in the Asian region, a contrasting picture emerged. Japan and NICs directed their science funding to high technologies, capital, exports of engineering goods, advanced materials, modern biology, and basic research in information and communication technologies. In the 1990s, Japan, South Korea, and Singapore registered greater proportions of private sector funding than even Germany and the USA. Japan's R&D expenditures have increased eightfold over the 20 year period from 1971 to 1993, which is the highest rate of increase among the industrially developed nations.

At the other extreme are the Southern Asian countries, including China, India, Pakistan, Bangladesh, Nepal, Myanmar, and Sri Lanka, where R&D funding related to agriculture and to the manufacturing sector assumed equal importance, as over 60 percent of their populations was dependent on agriculture. Another revealing feature about countries such as China, India, and Pakistan was the importance given to military and strategic R&D, which consumed 45 to 55 percent of the total R&D budget.

In most of the South East Asian countries, the contribution of agriculture to the GDP witnessed a considerable decline in contrast to Southern Asian countries. In terms of sectoral contribution to the GDP, none of the South East Asian countries accounted for more than 26 percent (the figure for Vietnam) for agriculture in 1998. The manufacturing and service sectors witnessed unprecedented growth rates between 1980 and 1998. Though these countries spend no more than 0.5 percent of GDP on R&D, the thrust of science funding is directed to manufacturing, industrial, and service-related activities. Agricultural research consumes only 5–10 percent of the total R&D funding, much less than in the Southern Asian countries.

See also: Infrastructure: Social/Behavioral Research (Japan and Korea); Research and Development in Organizations; Research Ethics: Research; Research Funding: Ethical Aspects; Science, Economics of; Science Funding: United States; Science, Social Organization of; Scientific Academies in Asia; Universities and Science and Technology: Europe; Universities and Science and Technology: United States

Bibliography

Arrow K 1962 Economic welfare and the allocation of resources for invention. In: Nelson R (ed.) *The Rate and Direction of Inventive Activities*. Princeton University Press, Princeton, NJ, pp. 609–25

Pavitt K 1991 What makes basic research economically useful. *Research Policy* **20**, pp. 109–19

Pavitt K 1998 The social shaping of the national science base. *Research Policy* **27**, pp. 793–805

Rosenberg N 1991 Critical issues in science policy research. *Science and Public Policy* **18**, pp. 335–46

UNESCO 1998 *World Science Report*. Paris

World Development Report 1998/1999

World Development Report 1999/2000

V. V. Krishna

Science Funding: Europe

European governments invest considerable sums of money in science. This article examines the reasons why they do this, covering briefly the historical context of European science funding and highlighting current issues of concern. The focus is on government funding of science, rather than funding by industry or charities, since government has historically been the largest funder of 'science' as opposed to 'technology.' As an approximate starting point, 'science' refers to research that is undertaken to extend and deepen knowledge rather than to produce specific technological results, although the usefulness of this distinction will be questioned below. By 'science policy' what is meant is the set of objectives, institutions, and mechanisms for allocating funds to scientific research and for using the results of science for general social and political objectives (Salomon 1977). 'Europe' here only refers to Western Europe within the European Union, excluding the Eastern European countries.

1. Background

Government funding of science in Europe started in a form that would be recognizable today only after World War II, although relations between science and the state can be traced back at least as far as the Scientific Revolution in the seventeenth century (Elzinga and Jamison 1995). The history of science funding in Europe can be summarized broadly as a movement from a period of relative autonomy for scientists in the postwar period, through stages of increasing pressures for accountability and relevance, resulting in the situation today, where the majority of scientists are encouraged to direct their research towards areas that will have some socially or industrially relevant outcome.

However, this account is too simplistic. Many current concerns about science funding are based on the idea that 'pure' or 'basic' science (autonomous research concerned with questions internal to the discipline) is being sacrificed in place of 'applied' research (directed research concerned with a practical outcome), incorrectly assuming that there is an unproblematic distinction between the two (see Stokes 1997). Looking back, it can be seen that even in the late 1950s there were expectations that science should provide practical outcomes in terms of economic and social benefits, and the work that scientists were doing

at this time was not completely 'pure,' because much of it was driven by Cold War objectives. This is a demonstration of the broader point that in science policy, the categories used to describe different types of research are problematic, and one must be careful when using the traditional terminology. With these caveats in place, it is possible to trace the major influences on European science funding.

In the 1950s and 1960s, much of the technologically oriented funding of research was driven by military objectives and attempts to develop nuclear energy. In terms of science funding, this was a period of institutional development and expansion in science policy (Salomon 1977). The autonomy that scientists enjoyed at this time was based on the assumption that good science would spontaneously generate benefits. Polanyi (1962) laid out the classic argument to support this position, describing a self-governing 'Republic of Science.' He argued that because of the essential unpredictability of scientific research, government attempts to direct science would be counterproductive because they would suppress the benefits that might otherwise arise from undirected research. This influential piece can be seen as a response to Bernal's work (1939), which was partly influenced by the Soviet system and which argued that science should be centrally planned for the social good.

Another important concept of the time was the 'linear model' propounded by US science adviser Vannevar Bush (1945). In this model for justifying the funding of science, a one-way conceptual line was drawn leading from basic research to applied research to technological innovation, implying that the funding of basic research would ultimately result in benefits that would be useful to society.

But pressures on science from the rest of society were increasing. In the 1970s, there was a growing awareness of environmental problems (often themselves the results of scientific and technological developments), and European countries experienced the oil crises, with accompanying fiscal restrictions. There were increasing pressures on scientists to be accountable for the money they were spending on research. Also at this time the social sciences, especially economics, provided new methods for understanding the role of scientific research in industrial innovation and economic growth (see Freeman 1974).

In the 1980s Europe realized it had to respond to the technological and economic challenges of Japan and the US, and because of the ending of the Cold War, military incentives for funding science were no longer so pressing. Technology, industrial innovation, and competitiveness were now the main reasons for governments to fund science. Academic studies of innovation also began to question the linear model of the relationship between science and technology, described above, arguing that the process was actually more complicated (e.g., Mowery and Rosenberg 1989). This led to pressures on the previous 'contract' between government and scientists (Guston and Keniston 1994), which had been based on the assumptions of the linear model. Rather than presuming that science would provide unspecified benefits at some unspecified future time, there were greater and more specific expectations of scientists in return for public funding. This was accompanied by reductions in the growth of science budgets, producing a 'steady state' climate for scientific research, where funding was not keeping up with the rapid pace at which research was growing (see Ziman 1994).

Science policy work at this time produced tools and data for measuring and assessing science. Various techniques were developed, such as technology assessment, research evaluation, technology management, indicator-based analysis, and foresight (Irvine and Martin 1984).

In the 1990s, there was greater recognition of the importance of scientific research for innovation, with the development of new hi-tech industries that relied on fundamental scientific developments (such as biotechnology), in conjunction with other advanced technologies. There were also growing pressures for research to be relevant to social needs. Gibbons et al. (1994) argued that the 1990s have witnessed an increasing emphasis on problem-oriented, multidisciplinary research, with knowledge production having spread out to many diverse locations, and that distinctions between basic and applied science, and between science and technology, are becoming much more difficult to make.

2. *The Influence of the European Union*

Moving from a general historical context to look more specifically at the European level shows that research funding from the European Union (EU), in the form that it currently takes, did not start until 1984 with the first 'Framework Programme.' This funded pre-competitive research (i.e., research that is still some way from market commercialization) following an agenda influenced by industrial needs (Sharp 1997). From the 1960s onward, the Organization for Economic Cooperation and Development (OECD) had been a more influential multinational organization than the EU in terms of national science policies (Salomon 1977). In particular, the OECD enabled countries to compare their research activities with those of other countries, and encouraged greater uniformity across nations.

Currently EU research funding only comprises a few percent of the total research funding of all the member states (European Commission 1994), although it has been more important in the 'less favored' regions of Europe (Peterson and Sharp 1998). Consequently, in terms of science funding, the national sources are more important than the EU. However, EU programs do have an influence on the funding priorities of national governments. In theory, the EU

does not fund research that is better funded by nation states (according to the 'principle of subsidiarity,' see Sharp 1997), so it does not fund much basic research, but is primarily involved in funding research that is directed towards social or industrial needs.

The most important impact of the EU has been in stimulating international collaboration and helping to form new networks, encouraging the spread of skills. One of the requirements of EU funded projects is that they must involve researchers from at least two countries (Sharp 1997). This could be seen as part of a wider political project that is helping to bind Europe together. It is possible that many of these collaborations might have happened without European encouragement because of a steady rise in all international collaborations (Narin et al. 1991). However, it is likely that through its collaborative programs and their influence, the EU will play an increasingly important role in the future of research funding in the member countries (Senker 1999).

3. *Individual Countries in Europe*

Since it is the individual countries in Europe that are responsible for the majority of science funding, the organization of their research systems deserves attention.

All the countries have shown the general trends outlined above, but the historical and cultural differences among the European nations lead to considerable diversity in science funding arrangements. It is possible to compare the different countries by looking at the reasons why they fund science and the ways in which research is organized.

European nations, like those elsewhere, have traditionally funded science to encourage economic development, although most countries also attach importance to advancing knowledge for its own sake. Some countries such as Sweden and Germany have emphasized the advancement of knowledge, and other countries, such as Ireland, have put more emphasis on economic development (Senker 1999). Since the 1980s, the economically important role of science has been emphasized in every country. This has often been reflected at an organizational level with the integration of ministerial responsibilities for science funding with those for technology and higher education.

We can compare individual countries in terms of differences in the motivations behind funding research. Governments in France and Italy have traditionally promoted 'prestige' research, and have funded large technology projects, such as nuclear energy. These reasons for funding research, even though they are less significant in the present climate, have had long-lasting effects on the organization of the national research systems. The UK is notable in that the importance of science for economic competitiveness is emphasized more than in other European countries, and industrial concerns have played a larger role (Rip 1996).

Organizational differences between countries can tell us something about the way research funding is conceptualized and can reflect national attitudes toward the autonomy and accountability of researchers. In the different European countries, the locus of scientific research varies. In some countries, the universities are most important (e.g., Scandinavia, Netherlands, UK), and funds are competed for from research councils (institutions that mediate between scientists and the state, see Rip 1996). In this type of system there will usually be some additional university funding that provides the infrastructure, and some of the salaries. The level of this funding varies between countries, which results in differences in scientists' dependence on securing research council funds and has implications for researcher autonomy. In other countries, a great deal of scientific research is carried out in institutions that are separate from the universities (e.g., France and Italy).

The situation is not static, and scientific research in the university sector has been growing in importance across the whole of Europe (Senker 1999). For example, in France, the science funding system has traditionally been centralized with most research carried out in the laboratories of the Centre National de la Recherche Scientifique (CNRS). Now the situation is changing and universities are becoming more involved in the running of CNRS labs, because universities are perceived to be more flexible and responsive to user needs (Senker 1999). Germany is an interesting case because there is a diversity of institutions involved in the funding of science. There is a division of responsibility between the federal state and the Länder, which are responsible for the universities. There are also several other types of research institute, including the Max Planck institutes, which do basic research, and the more technologically-oriented Fraunhofer institutes. Resulting institutional distinctions between different types of research may lead to rigidities in the system (Rip 1996). In all countries in Europe, there is an attempt to increase coordination between different parts of the national research system (Senker 1999).

4. *Current Trends*

As has been emphasized throughout, European governments have demanded increasing relevance of scientific results and accountability from scientists in return for funding research. Although the situation is complex, it is clear that these pressures, and especially the rhetoric surrounding them, increased significantly during the 1990s. This has led to worries about the place for serendipitous research in a 'utilitarian–instrumental' climate (Nowotny 1997, p. 87).

These pressures on science to be useful are not the only notable feature of the current funding situation. The views of the public are also becoming more important in decisions concerning the funding of science.

The risks and detrimental effects of science are of particular concern to the public, possibly because the legitimacy of the authority of politicians and scientists is being gradually eroded (Irwin and Wynne 1996). Throughout Europe, there has been a growth in public distrust in technological developments, which has led to pressures for wider participation in the scientific process. This is related to the current (and somewhat desperate) emphasis on the 'public understanding of science,' which is no longer simply about educating the public in scientific matters, but has moved towards increasing participation in the scientific process (see Gregory and Miller 1998). Concerns about the environmental effects of scientific developments can be traced back to the 1960s, but recent incidents in the 1980s and 1990s have led to a more radical diminution of public faith in scientific experts (with issues such as climate change, Chernobyl, and genetically modified foods).

The public distrust of science may also be due to the fact that scientists, by linking their work more closely either to industrial needs or to priorities set by government, are losing their previously autonomous and potentially critical vantage point in relation to both industry and government.

Certain European countries, especially the Netherlands and Scandinavia, which have a tradition of public participation, are involving the public more in debates and priority setting on scientific and technological issues. This has been described as a 'post-modern' research system (Rip 1996). As the distinction between science and technology becomes less clear, in this type of research system there is also a blurring of boundaries between science and society.

5. *Implications*

An implication of these changes in science funding is that the growing importance of accountability and of the role of the public in scientific decisions may have epistemological effects on the science itself, since scientific standards will be determined not only by the scientific community but by a wider body of actors often with divergent interests (Funtowicz and Ravetz 1993).

If norms are linked to institutions, and if institutions are changing because of the greater involvement of external actors in science, and of science in other arenas, then the norms may be changing too (Elzinga 1997). This is an issue that was touched on in the 1970s and 1980s by the proponents of the 'finalization thesis' who argued that, as scientific disciplines become more mature, they become more amenable to external steering (Böhme et al. 1983). The importance of external influences leads to worries about threats to traditional values of what constitutes 'good' science (Elzinga 1997, see also Guston and Keniston 1994 for US parallels). There may be an emergence of new standards of evaluation of scientific research.

European science funding has changed considerably since it was institutionalized, partly because of its success in generating new technologies and partly because of its failures and their social consequences. It is becoming more difficult to categorize science, technology, and society as separate entities (Jasanoff et al. 1995), or to think of pure scientists as different from those doing applied research. Wider society has become inextricably linked with the progress of science and the demands placed on scientists and science-funding mechanisms are starting to reflect this restructuring. This tendency is likely to continue into the future.

See also: Infrastructure: Social/Behavioral Research (Western Europe); Kuhn, Thomas S (1922–96); Research and Development in Organizations; Research Funding: Ethical Aspects; Science, Economics of; Science Funding: Asia; Science Funding: United States; Science, Social Organization of; Universities and Science and Technology: Europe

Bibliography

Bernal J D 1939 *The Social Function of Science*. Routledge, London

Böhme G, Van Den Daele W, Hohlfeld R, Krohn W, Schäfer W 1983 *Finalization in Science: The Social Orientation of Scientific Progress*. Reidel, Dordrecht, The Netherlands

Bush V 1945 *Science: The Endless Frontier*. USGPO, Washington, DC

Elzinga A 1997 The science–society contract in historical transformation. *Social Science Information* **36**: 411–45

Elzinga A, Jamison A 1995 Changing policy agendas. In: Jasanoff S, Markle G E, Petersen J, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA

European Commission 1994 *The European Report on Science and Technology Indicators*. European Commission, Luxembourg

Freeman C 1974 *The Economics of Industrial Innovation*. Penguin, Harmondsworth, UK

Funtowicz S O, Ravetz J R 1993 Science for the post-normal age. *Futures* **25**: 739–56

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M 1994 *The New Production of Knowledge*. Sage, London

Gregory J, Miller S 1998 *Science in Public: Communication, Culture and Credibility*. Plenum, New York

Guston D, Keniston K 1994 *The Fragile Contract: University Science and the Federal Government*. MIT Press, London

Irvine J, Martin B 1984 *Foresight in Science: Picking the Winners*. Pinter, London

Irwin A, Wynne B (eds.) 1996 *Misunderstanding Science? The Public Reconstruction of Science and Technology*. Cambridge University Press, Cambridge, UK

Jasanoff S, Markle G E, Petersen J, Pinch T (eds.) 1995 *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA
Mowery D, Rosenberg N 1989 *Technology and the Pursuit of Economic Growth*. Cambridge University Press, Cambridge, UK
Narin F, Stevens K, Whitlow E 1991 Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics* **21**: 313–23
Nowotny H 1997 New societal demands. In: Barre R, Gibbons M, Maddox J, Martin B, Papon P (eds.) *Science in Tomorrow's Europe*. Economica International, Paris
Peterson J, Sharp M 1998 *Technology Policy in the European Union*. Macmillan, Basingstoke, UK
Polanyi M 1962 The republic of science: its political and economic theory. *Minerva* **1**: 54–73
Rip A 1996 The post-modern research system. *Science and Public Policy* **23**: 343–52
Salomon J J 1977 Science policy studies and the development of science policy. In: Spiegel-Rösing I, Solla Price D (eds.) *Science, Technology and Society: A Cross-disciplinary Perspective*. Sage, London
Senker J 1999 *European Comparison of Public Research Systems*. Report prepared for the European Commission. SPRU, Sussex, UK
Sharp M 1997 Towards a federal system of science in Europe. In: Barre R, Gibbons M, Maddox J, Martin B, Papon P (eds.) *Science in Tomorrow's Europe*. Economica International, Paris
Stokes D E 1997 *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution Press, Washington, DC
Ziman J 1994 *Prometheus Bound*. Cambridge University Press, Cambridge, UK

J. Calvert and B. R. Martin

Science Funding: United States

The pursuit of knowledge has historically been shaped by patronage relationships. The earliest scientists were astronomers and mathematicians supported at ancient courts in the Middle East and China. In Renaissance Europe, observers of and experimenters with nature had patrons in the aristocracy or royal families. In nineteenth century America, industrial philanthropists took up their cause.

Since World War II, research in the United States has received its support largely from government, with help from industry and private foundations. This article focuses on the postwar history of government funding for research in the United States. That history has been characterized by a creative tension between autonomy and accountability, embodied in successive waves of invention of new institutional arrangements. Should science set its own agenda, or should it answer to the public? Could the two goals be reconciled? These issues have echoed through five decades of research policy.

1. Autonomy for Prosperity

In the eighteenth and nineteenth centuries, the US federal government funded research only if it was applied directly to practical goals. The earliest federal efforts were in surveying and geological exploration, activities that contributed both to nation-building and to the search for mineral wealth. A second major wave of federal effort in the latter half of the nineteenth century took up agricultural research in partnership with the States, through land grant colleges and agricultural experiment stations. Scientists were also brought in to solve the immediate problems of wartime. World War I was known as 'the chemists' war,' because of the use of chemical warfare agents. World War II became 'the physicists' war,' with the invention of the atomic bomb. And in the 1930s, a small federal laboratory for health research was established, which later became the National Institutes of Health, the nation's research arm for biomedical sciences (Dupree 1964, Kevles 1978, Strickland 1989).

By the time of World War II, these efforts had grown into a set of research programs in mission agencies, focused on food, health, and defense. These activities were carried out largely in government laboratories, and were focused on immediate, practical goals. In later decades, two other agencies joined this group. The National Aeronautics and Space Administration (NASA) was formed in response to the launching of a Russian satellite, the Sputnik, in 1957 (Logsdon 2000). And the Department of Energy, when it was established in the 1970s, incorporated major research elements, including the high-energy physics laboratories. This array of government research efforts continues to anchor the mission-oriented portion of a pluralistic system of funding for US research.

The other dimension of that system is fundamental research. It is anchored in universities, and interwoven with graduate education and the preparation of new generations of researchers. This second dimension has its origins in moments of national crisis, when government research efforts were expanded temporarily through the cooptation of university researchers. After such an interlude in World War I, influential US scientists tried to convince the federal government to keep the research relationship with universities going, but to no avail. It was only after the success of the crucial atomic bomb project that the achievements of science carried enough weight to make credible proposals for permanent, across-the-board government support for research (Kevles 1978).

The spokesperson for this plan was Vannevar Bush, a university scientist who had been active in the government projects around the war. He called for the formation of a National Research Foundation, to provide basic support for research, without the specific targets that had been imposed during the war. He argued that unfettered research, carried out largely in

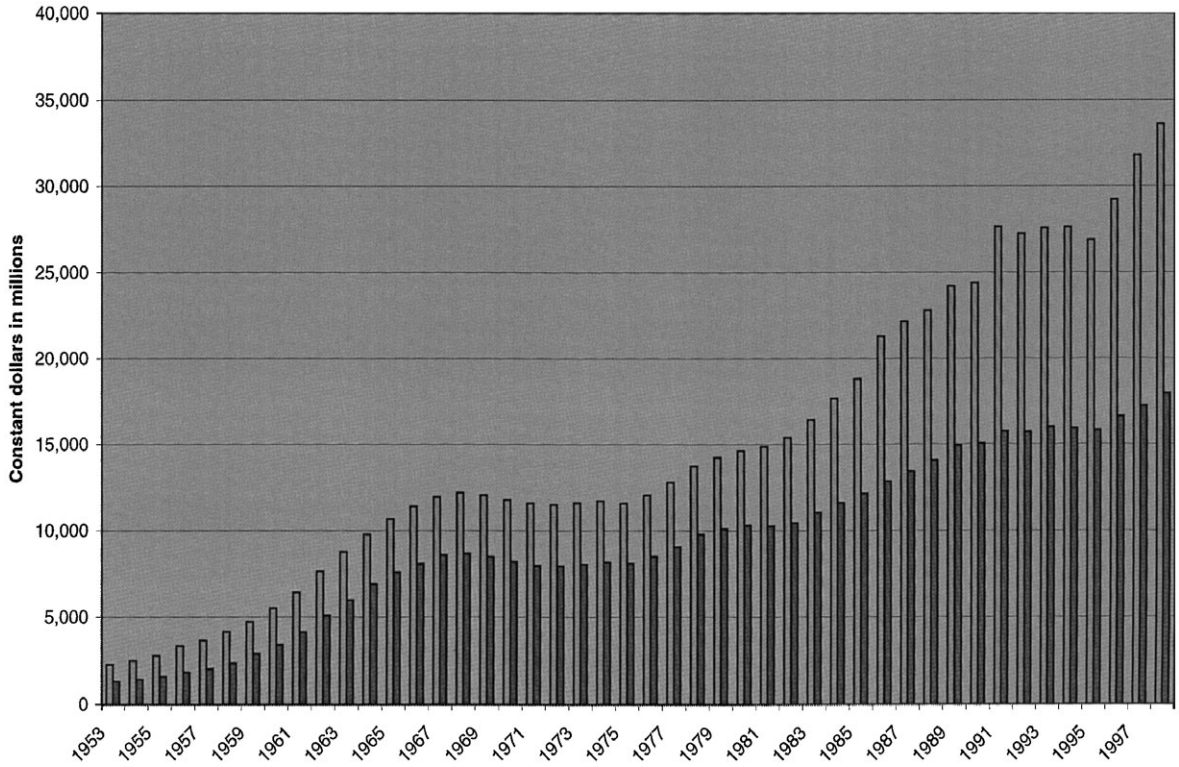

Figure 1
US basic research spending, 1953–1998, total and Federal

universities, would build a base of human and knowledge resources that would help solve problems across the range of challenges in health, defense, and the economy (Bush 1990 [1945]).

But to make these contributions, researchers needed freedom—both freedom in the laboratory to choose and pursue research problems, and organizational freedom at the level of government agencies, to set the larger research agenda in directions that were free from political control. Bush's model of the relationship between science and society has been called the 'autonomy for prosperity' model (Cozzens 2001). Most of the experimentation over the next few decades with ways to maintain autonomy while being useful took the form of add-ons to this model.

Bush's model built a protective shell of organizational autonomy around the agencies that provided funds for basic research. The National Science Foundation (NSF) and the National Institutes of Health (NIH), both of which grew very fast in the postwar period, developed strategies that insulated direct decisions around research from the political context. One funding mechanism that embodied the essence of the autonomy-for-prosperity model in its early days was the science development program. Under these programs in the 1950s and 1960s, federal block grants allowed universities to build their educational and research capacities, with very few strings attached (Drew 1985). The second strategy was the project grant/peer review system of funding. Under this mechanism, the government hands over choice of research topics as well as the judgment of what is to count as quality into the hands of researchers. 'Peer review' for project selection became both the major form of quality control in the federal funding system and an important symbol of scientific freedom (Chubin and Hackett 1990).

Ironically, the autonomy-protecting institutional shell that Bush and his successors designed undermined the practical effectiveness that had won credibility for his plan to start with. As government funding for research grew rapidly in the 1950s, beyond the scale of Vannevar Bush's wildest dreams, the institutional shell unintentionally created a protected space where new researchers could do their work without any grounding in the practical problems of the world. As soon as the practical grounding was lost, a gap was created between science and society that needed to be bridged if autonomy was really going to be turned into prosperity.

2. Knowledge Transfer and Knowledge Mandating

Among the first set of funding methods developed to bridge this gap were knowledge transfer mechanisms. Knowledge transfer mechanisms do not threaten either laboratory or organizational autonomy, because they leave the protective shell in place while focusing on diffusing or disseminating research-based knowledge. For example, in the 1960s and 1970s, many federal programs addressed the 'information explosion' with 'science information services,' under the rubric of 'information policy.' The government encouraged first journals, then abstracting and indexing services, to provide access to the exploding journal literature, and later added extension services. The emphasis was on providing infrastructure for communication, not shaping its content; science was to speak to the public, not with it. These mechanisms protect research autonomy by segmenting knowledge processes over time.

As the size of the research enterprise continued to grow, however, the pendulum swung back and the accountability issue inevitably arose. Elected representatives of the public asked 'What are we getting for all this spending?' In the absence of specific answers, a backlash eventually appeared. Policymakers began to demand that research solve societal problems directly, rather than through the diffuse claim of 'autonomy for prosperity.' In 1969, for example, Congress passed 'the Mansfield amendment,' limiting support from the Department of Defense to goal-directed research. The Nixon administration phased out science development programs. And in the 1970s, several 'knowledge mandating' programs appeared. At NSF, for example, the Program of Research Applied to National Needs (RANN) grew out of a Presidential concern about 'too much research being done for the sake of research.' (Mogee 1973). At NIH, the 'War on Cancer' emerged during the same period (Strickland 1989, Rettig 1977).

In their more programmed forms, such programs threatened both individual and organizational autonomy. But even in softer forms, in which money was only redirected from one priority to another, they still threatened organizational autonomy, since external forces were dictating research directions. There is an important lesson in the history of these programs: over time, organizational autonomy won out over societal knowledge mandating. RANN was abolished, and over the years, both NSF and NIH have found ways to look programmatic without doing central planning.

3. Knowledge Sharing

By the mid-1970s, a third approach began to be developed in the United States to bridge the gap between the science produced inside the protective institutional shell and the problems of the world surrounding it. This new form threatened neither the individual autonomy of researchers nor the organizational autonomy of funding organizations. This new form can be called 'knowledge-sharing.'

This approach first took the form of an emphasis on partnerships, and within that emphasis, an early focus on partnerships with industry. New centers involved industry partners in setting strategic plans. The interchange of people and information became the rule. The critical element in these was a two-way dialog, which replaced the one-way science-to-society diffusion of information of knowledge transfer and the one-way society-to-science mechanism of knowledge mandating. In the two-way dialog, scientists became strategic thinkers, able to formulate their own problems creatively, but steeped again, as in the 1950s, in the problems articulated by some external set of partners. Another new element was the explicit link to education, at both graduate and undergraduate levels. The Engineering Research Centers of the NFS, for example, were intended to produce a 'new breed' of engineers, better prepared than previous generations to participate in R&D in industry because they understood the needs and culture of industry.

This new partnership model raised questions: Was knowledge mandating being changed from a public function to a private one under this scheme? A change in the law that allowed universities to hold intellectual property rights in results produced under government grants further heightened these concerns. Public knowledge seemed to be on a path to privatization. At the same time, several well-publicized instances of accusations of fraud in science seemed to undermine the trusting relationships among citizens, government, and science (Guston 1999).

In the late 1980s, however, a second crucial step in the development of the partnership model took place. This was a step back toward the public. Centers were urged to form partnerships, not only with industry, but also with State and local governments, citizen groups, and schools. The benefits that industry gained from earlier collaborations were now available to other parts of society, in a new pattern that has been called 'partner pluralism' (Cozzens 2001). Strategic plans began to be shaped by two-way dialog across many sectors, and the education of a new breed of researcher able to bridge the cultures of university, school, and public service sector began.

Another manifestation of knowledge-sharing and an attempt to restore public trust arrived with a round of attention among researchers to public awareness of science and science education. The 1990s heard a new note in the discussions of the public among science leaders. Research leaders stressed that scientists need to learn more about the public, as well as the public learning more about science. Even the venerable National Academy of Sciences recommended that the institutions of science open their doors to the public,

urging, for example, that NIH take more advice from the public in priority-setting. These were signs of the beginning of a two-way dialog.

4. The Knowledge Society

As US research entered the twenty-first century, many believed that knowledge is the key resource in the new economy. For a society to be innovative, however, its creative capacity must be widely shared. This goal will not be achieved unless its scientists become strategic thinkers steeped in society's problems and issues, and government funding agencies remain public through partner pluralism. Accountability for research in the knowledge society is achieved through the engagement of many societal actors in the research enterprise. By placing many actors on an equal footing and encouraging two-way dialog, research policy in the twenty-first century will help stimulate shared creativity, and open a new path to prosperity.

See also: Academy and Society in the United States: Cultural Concerns; Research Funding: Ethical Aspects; Science Funding: Asia; Science Funding: Europe; Science, Social Organization of

Bibliography

Bush V 1990 [1945] *Science—the Endless Frontier*. National Science Foundation, Washington, DC, reprinted from the original

Chubin D E, Hackett E J 1990 *Peerless Science: Peer Review and US Science Policy*. State University of New York Press, Albany, New York

Cozzens S E 2001 Autonomy and accountability for 21st century science. In: de la Mothe J (ed.) *Science, Technology, and Governance*. Pinter, London

Drew D E 1985 *Strengthening Academic Science*. Praeger, New York

DuPree A H 1964 *Science in the Federal Government: A History of Policies and Activities to 1940*. Harper & Row, New York

England J M 1983 *A Patron for Pure Science: The National Science Foundation's Formative Years*. National Science Foundation, Washington, DC

Guston D H 2000 *Between Politics and Science: Assuring the Productivity and Integrity of Research*. Cambridge University Press, New York

Kevles D J 1978 *The Physicists: The History of a Scientific Community in Modern America*. Knopf, New York

Logsdon J M 1970 *The Decision to Go to the Moon: Project Apollo and the National Interest*. MIT Press, Cambridge, MA

Mogee M E 1973 *Public Policy and Organizational Change: The Creation of the RANN Program in the National Science Foundation*. MSc. Thesis, George Washington University, Washington, DC

Morin A J 1993 *Science Policy and Politics*. Prentice-Hall, Englewood Cliffs, NJ

Mukerji C 1989 *A Fragile Power: Scientists and the State*. Princeton University Press, Princeton, NJ

Rettig R A 1977 *Cancer Crusade: The Story of the National Cancer Act of 1971*. Princeton University Press, Princeton, NJ

Smith B L R 1989 *American Science Policy since World War II*. Brookings Institution, Washington, DC

Stokes D E 1997 *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution, Washington, DC

Strickland S P 1988 *The Story of the NIH Grants Program*. University Press of America, Lanham, MD

S. E. Cozzens

Science, New Forms of

In recent times, there has been a growing sense that 'science' is not merely expanding in its grasp of the natural world, but that the activity itself is changing significantly. For this we use the term 'new forms of science,' understanding both the practice and the reflection on it. Since science is now a central cultural symbol, changes in the image of science rank in importance with those in the practice itself. Our situation now is one of the differentiation of a practice that hitherto seemed unified, and of conflict replacing consensus on many issues, some of which had been settled centuries ago. The prospect for the immediate future is for a social activity of science that no longer has a unified core for its self-consciousness and its public image.

The main baseline for this 'novelty' is roughly the century preceding World War II, when the 'academic' mode of scientific activity became dominant. This was characterized by the displacement of invention and diffusion by disciplinary research, and of amateurs (with private means or patronage) by research professionals (perhaps with a partial sinecure in advanced teaching). In the latter part of this period, the mathematical–experimental sciences took the highest prestige away from the descriptive field sciences. The social sciences were increasingly pressured to conform to the dominant image. And throughout, the consciousness of the activity, reflected in scholarly history and philosophy as well as in popularization, was triumphalist and complacent. There could be no way to doubt that the idealistic scientists, exploring nature for its own sake, inevitably make discoveries that eventually enrich all humanity both culturally and materially.

Of course, any 'period' in a social activity is an intellectual construct. Closer analyses will always reveal change and diversity within it; and it contains both relics of the deeper past and germs of the future. But occasionally there occurs a transforming event, so that with all the continuities in practice there is a discontinuity in consciousness. In the case of recent science, this was the Manhattan Project, in which the first atomic bombs were designed and constructed in a

gigantic scientific engineering enterprise. It was of particular symbolic significance that the project was conceived and managed, not by humble 'applied scientists' or engineers, but by the elite theoretical physicists. It was science itself that was implicated in the moral ambiguities and political machinations that ensued in the new age of nuclear terror.

After the Bomb, it was no longer possible to build scholarly analyses on assumptions of the essential innocence of science and the beneficence of scientists. Prewar studies in the sociology of science, such as those of Merton (with his 'four norms'), which assumed a purity in the endeavor and its adherents, could later find echoes only in those of the amateur philosopher Polanyi (1951), himself embattled against J. D. Bernal and the Marxist 'planners' of science (1939). But it took about a decade for a critical consciousness to start to take shape. Those scientists who shared the guilt and fear about atomic weapons still saw the cure in educating society, as through the *Bulletin of the Atomic Scientists*, rather than wondering whether science itself had taken a wrong turning somewhere. Leo Szilard (1961) imagined an institute run by wise scientists that would solve the world's problems; and Jacob Bronowski (1961) faced the problem of the responsibility of science for the Bomb, only to deny it absolutely. The vision of Vannevar Bush (1945), of science as a new and endless frontier for America, continued the prewar perspective into postwar funding.

The first significant novelty in the study of science was accomplished in the 1950s by Derek J. de Solla Price (1963). Although his quantitative method seemed at times to be a caricature of science itself, he was submitting scientific production to an objective scrutiny that had hitherto seemed irrelevant or verging on irreverent. And he produced the first 'limits to growth' result; with a doubling time (steady over three centuries!) of 15 years, compared to some 70 in the economy, sooner or later society would refuse to foot the bill. Since the prevailing attitude among scientists was that society's sole contribution to science should be a blank check, Price's bleak forecast, however vague in its timing, was not well received. He went on to give the first characterization of the new age of science; but he only called it 'big,' as opposed to the 'little' of bygone ages.

A more serious disenchantment motivated Thomas S. Kuhn. He had imbibed the naive progressive image of science, whereby truths were piled on truths, and scientists erred only through bad methods or bad attitudes. Then it all cracked, and he wrote his seminal work on *Scientific Revolutions* (1962), all the more powerful because of its confusions and unconscious ironies. Whether he actually approved of routinized puzzle-solving 'normal science' was not clear for many years; to idealists like Popper, Kuhn's flat vision of science represented a threat to science and to civilization (1970).

Through the turbulent 1960s, more systematically radical social critiques began to emerge. Of these, the most cautious was that of John Ziman's *Public Knowledge* (1968), and the most ambiguous that of Jerome Ravetz (combining *Gemeinschaft* craftsmen's science, Nuclear Disarmament, and the Counterculture, but ending his book with a prayer by Francis Bacon) (1996, [1971]). An attempt at a coherent Marxist perspective was mounted by Steven and Hilary Rose (1976). This was enlivened by their internal tension between Old and New Left, but with the complete discrediting of Soviet socialism, it was soon relegated to purely historical significance. A serious attempt at an objective social study of science was made by N. D. Ellis (*The Scientific Worker* 1969) but there was then no audience for such an approach. Another premature social analysis of science came to light around then. This was the work of Ludwik Fleck (1979 [1935]) on the 'genesis and development of a scientific fact'; at the time of its publication and for long after, the very title smacked of heresy.

At the same time, the new dominant social practice of science, either 'industrialized' (Ravetz) or 'incorporated' (the Roses) took shape. The prophetic words of President Eisenhower on the corrupting influence of state sponsorship were little heeded; and an American science that was both 'pure' and 'big' produced its first muckraker, in Dan Greenberg (1967), a Washington correspondent for *Science* magazine. There was a short but sharp protest at the militarization of science in the Vietnam War (Allen 1970), out of which came various attempts at organizing for 'social responsibility in science.'

On the ideological front, there was a collapse of attempts to preserve science as the embodiment of the Good and the True. Somewhat too late for the 1960s but revolutionary nonetheless, Paul Feyerabend published *Against Method*, in which even the rearguard defenses of Popper (1963) and Lakatos (1970) were demolished. Then came the tongue-in-cheek naive anthropology of *Laboratory Life* (Latour and Woolgar 1977), considering scientists as a special tribe producing 'inscriptions' which were intended to gain respect and to rebut hostile criticism. A radically skeptical theory of scientific knowledge was elaborated in sociological studies called 'constructivism.'

As prewar science receded in time and relevance, scholars could settle down to the analysis of this new form of social practice of science in its own terms. Early in the postwar period there had been a conception that scientists in the public arena should be 'on tap but not on top,' a sort of 'drip' theory of their function. But more sophisticated analyses began to emerge. Among these was 'mandated science' (Salter 1988), in which the pitfalls of well-intentioned involvement were chronicled. The special uses and adaptations of science in the regulatory process and in the advisory function were analyzed (Jasanoff 1990). In such studies, it became clear that neither 'science'

nor 'scientist' can be left as elements outside the process; indeed, the old Latin motto 'who guards the guardians?' reminds us of the essentially recursive nature of the regulatory process. For (as Jasanoff showed) those who regulate or advise act by certain criteria and are also selected for their position; and who sets the criteria, and who chooses the agents, can determine the whole process before the scientist-advisors even get to work.

Meanwhile the existing trends in the social practice of science were intensified. Physics lost the preeminence it had gained through winning the war with the Atomic Bomb. Civil nuclear power (which was really engineering anyway) failed to live up to its early promise, producing not only disasters (near and actual) in practice but also intractable problems in waste management. The core of basic physics research, high-energy studies, was caught in a cul-de-sac of ever more expensive machines hunting ever more esoteric particles. It may also have suffered from its overwhelming reliance, not covert but not widely advertised either, on the US military for its funding. Then physics was caught up in the debate over 'Star Wars,' which was controversial not merely on its possible political consequences, but even on the question of whether it could ever work as claimed. In this case, it appeared that the technical criteria of quality were totally outweighed by the political and fiscal.

The focus of leading-edge research shifted to biology, initiated by the epochal discovery of the structure of the genetic material DNA, the 'double helix.' There was a steady growth in the depth of knowledge and in the power of tools in molecular biology. This was punctuated in the 1970s by an episode of discovery of, and reaction to, possible hazards of research. In an unprecedented move, the 'recombinant DNA' research community declared a moratorium on certain experiments, until the hazards had been identified and regulated. Critics of the research interpreted this as an admission that the hazards were real rather than merely hypothetical. There was a brief period (1976–7) of confrontation in the USA, complicated by some traditional town–gown antagonisms and by relics of Vietnam activism. But it soon died down, having served partly to stimulate interest in the research but also planting the seeds of future disputes.

Another novelty in this period was the emergence of ethics as a systematic concern of science. Qualms about the uses of science in warfare go back a long way (Descartes 1638 has a significant statement), as well as disputes over priority and honesty. But with the Bomb, the possibility of science being turned to evil ends became very real; and with the loss of innocence in this respect, other sorts of complaints could get a hearing. Medical research is the most vulnerable to criticism, since the subjects are humans (or other sentient beings) and there must be a balance between their personal interests and those of the research (and hence of the broader community). But some cases went too far, as that of the African-American men who (in the 1930s) were not given treatment for syphilis, so that their degeneration could be recorded for science. Also, there were revelations of some of the more bizarre episodes of military science at the height of the Cold War, such as subjecting unwitting human subjects either to nuclear radiation or to psychotropic drugs. Since some universities shared responsibility with the military for such outrages, the whole of science could not but be tarnished. Out of all this came a problem of the sort that could not have been imagined in the days of little science: activism, to the point of terror tactics, against scientists and labs accused of cruelty to animals.

A further loss of innocence was incurred in the realization of the Marxist vision of science as 'the second derivative of production,' albeit under conditions that Marxists did not anticipate. Now 'curiosity-driven,' or even 'investigator-initiated' research, forms a dwindling minority within the whole enterprise. The days when the (British) Medical Research Council could simply 'back promising chaps' are long since gone. Increasingly, research is mission oriented, with the missions being defined by the large institutions, public and private, possessing the means and legitimacy to do so. Again, it is in biology, or rather bioengineering, where the change is most remarked. With genetic engineering, science is increasingly caught up in the globalization, or rather commodification, of everything that can be manipulated, be it life-forms or even human personality. When firms are accused of bio-piracy (appropriating plants or genetic materials from local people, and then patenting them), this is not exactly new, since it was also practiced by the illustrious Kew Gardens in Victorian times. But such practices are no longer acceptable in the world community (a useful survey of these issues is to be found in Proctor 1991, Chap. 17).

The social organization of science has changed in order to accommodate to these new tasks and circumstances. The rather vague term 'mode 2' has been coined to describe this novel situation (Gibbons et al. 1994). In the terms of this analysis of a condition which is still developing, scientists are reduced to proletarians, neither having control over the intellectual property in the products of their work (since it is done largely on contracts with various restrictions on publication), nor even the definable skills of disciplinary science (since most projects fall between traditional boundaries). The discussion in the book provides a useful term for this sort of science: 'fungible.' For the research workers become a convertible resource, being shipped from project to project, to be reassigned or discarded as convenient. There are all sorts of tensions within such a new dispensation, perhaps most notably the contradictory position of the educational institutions, whose particular sort of excellence may not be compatible with

the demands of the new labor market and research regime.

With the fragmentation and globalization of politics, other new sorts of scientific practice are emerging, not continuous with the evolving mainstream institutions but either independent of, or in opposition to, them. The leading environmental organizations have their own scientific staff who debate regularly with those promoting controversial projects. Increasingly, 'the public' (or representatives of the more aware and critical sections of it) is drawn into deliberative processes for the evaluation and planning of developments in technology and medicine. In some respects a 'stakeholder society' in science is emerging, engaged in debate on issues that had hitherto been left for the experts and politicians to decide. Movements for 'community research' are emerging in the USA (with a focus in the Loka Foundation) and elsewhere.

This development can be seen as an appropriate response to an emerging new major task for science. In the process of succeeding so brilliantly in both discovery and invention, science has thrown up problems that can be loosely defined as risk. The tasks here have significant differences from those of the traditional scientific or technical problems. For one, problems of risk and safety are inherently complex, involving the natural world, technical systems, people, institutions, values, uncertainties, and ignorance. Isolated, reductionist studies of the 'normal science' sort may be an essential component, but cannot determine a policy. Further, 'safety' can never be conclusively established, not merely because it is impossible to prove impossibility, but more because the evidence of causation is either only suggestive (as from toxicological experiments with acute doses on non-human species) or indirect (as from epidemiological studies).

In the debates on managing risk, methodology becomes politicized, as it is realized that the assignment of burden of proof (reflected in the design of statistical studies and of experiments) can be critical in determining the result that goes forward into the policy process. Conflict of interest becomes difficult to avoid, since most opportunities for achieving expertise are gained through work with the interests promoting, rather than with those criticizing, new developments. Also, the debates take place in public forums, such as the media, activist campaigning, or the courts, each of which produces its characteristic distortions on the process.

One response to this increased public involvement has been a variety of recommendations, from respected and even official sources, for greater openness and transparency in the science policy process. In some respects, this is only prudence. In the UK, the controversy over 'mad cow disease' (or the BSE catastrophe) could fester for a decade only because of official control of knowledge and ignorance. In its aftermath, the development of genetically engineered food crops was threatened by 'consumer power' in Europe and beyond. Expressed negatively, we can be said to be living in a 'risk society' (Beck 1992), where new forms of political activity are fostered by the consequences of an inadequately controlled science–technology system (Sclove 1995, Raffensperger 1999). The assumption of beneficence of that system is now wearing thin; thus the mainstream British journal *New Scientist* made this comment on the mapping of the human genome: 'For all our nascent genetic knowledge, if we end up with our genes owned by rich corporations and [with] a genetic underclass, will we really have advanced that much?' (2000)

There is no doubt that a significantly modified practice and image of science will be necessary for these new tasks of managing risks in the current social and political context. If nothing else, traditional science assumed that values were irrelevant to its work (this was, paradoxically, one of its great claims to value), and that uncertainties could be managed by technical routines (such as statistics). This was the methodological basis for the restriction of legitimacy to experts and researchers whose 'normal science' training had rigorously excluded such considerations. With uncertainty (extending even to salient ignorance) and value loading involved in any study of risk or safety, we are firmly in a period where a 'post-normal' conception of science is appropriate (Funtowicz and Ravetz 1993).

The core of this new conception is 'quality,' since 'truth' is a luxury in contexts where, typically, facts are uncertain, values in dispute, stakes high, and decisions urgent. Under these circumstances, there is a need for an 'extended peer community' for the assurance of the quality of the scientific inputs, supplementing the existing communities of subject specialists and clients. And this new body of peers will have its own 'extended facts.' The materials for this new sort of science involve not only the operations of the natural world, but also the behavior of technical systems and the social systems of control. Therefore, information of new sorts becomes relevant, including local knowledge and commitments, investigative journalism, and published confidential reports. Although 'post-normal science' has its political aspects, this proposed extension of legitimacy of participation in a scientific process is not based on political objectives. It is a general conception of the appropriate methodological response to this new predicament of science, where the tasks of managing risks and ensuring safety cannot be left to the community of accredited scientific experts alone.

This account of new forms of science would be incomplete without mention of a tendency which was thought to have been laid to rest long ago, as superstition unworthy of a civilized society. Coming in through East Asian practices of medicine and South Asian practices of consciousness, enriched cosmologies, involving 'vibrations' and 'energies' have become popular, even chic. The reduction of reality to its

mathematical dimensions, which was the metaphysical core of the Scientific Revolution, is now being eroded in the practice of unconventional medicine. Although this has not yet touched the core of either research or the policy sciences, it is a presence, alien to our inherited scientific culture, which cannot be ignored.

We are coming to the end of a period lasting several centuries, when in spite of all the major developments and internal divisions and conflicts, it made sense to speak of 'science' as an activity on which there was little disagreement on fundamentals. The very successes of that science have led to new challenges, from which significantly novel forms of science are emerging, characterized by equally novel forms of engagement with society.

See also: Cultural Studies of Science; Foucault, Michel (1926–84); History of Science; History of Science: Constructivist Perspectives; Innovation, Theory of; Kuhn, Thomas S (1922–96); Polanyi, Karl (1886–1964); Popper, Karl Raimund (1902–94); Reflexivity, in Science and Technology Studies; Science, Sociology of; Scientific Knowledge, Sociology of; Social Science, the Idea of

Bibliography

Allen J (ed.) 1970 *March 4: Scientists, Students, and Society*. MIT Press, Cambridge, MA

Beck U 1992 *Risk Society: Towards a New Modernity*. Sage, London

Bernal J D 1939 *The Social Function of Science*. Routledge, London

Bronowski J 1961 *Science and Human Values*. Hutchinson, London

Bush V 1945 *Science: The Endless Frontier*. Government Printing Office, Washington, DC

Descartes R 1638 *Discours de la Méthode* (6th part)

Editorial 2000 *New Scientist*, July 1

Ellis N D 1969 The scientific worker. PhD thesis, University of Leeds

Feyerabend P 1975 *Against Method*. New Left Books, London

Fleck L 1979 *Genesis and Development of a Scientific Fact*. Translation from German edn., 1935. University of Chicago Press, Chicago

Funtowicz S, Ravetz J R 1993 Science for the post-normal age. *Futures* **25**: 739–55

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M 1994 *The New Production of Knowledge*. Sage, London

Greenberg D 1967 *The Politics of Pure Science*. New American Library, New York

Jasanoff S 1990 *The Fifth Branch: Science Advisers as Policymakers*. Harvard University Press, Cambridge, MA

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Lakatos I 1970 History of science and its rational reconstructions. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK

Latour B, Woolgar S 1977 *Laboratory Life: The Social Construction of Scientific Facts*. Sage, Beverley Hills, CA

Merton R 1957 *Social Theory and Social Structure*. The Free Press, Glencoe, IL

Polanyi M 1951 *The Logic of Liberty*. Routledge, London

Popper K R 1963 *Conjectures and Refutations*. Routledge, London

Popper K R 1970 Normal science and its dangers. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK

Price D J 1963 *Little Science, Big Science*. Cambridge University Press, Cambridge, UK

Proctor R N 1991 *Value-free Science?* Harvard University Press, Cambridge, MA

Raffensperger C 1999 Editorial: Scientists making a difference. *The Networker: The Newsletter of the Science and Environmental Health Network* **4**

Ravetz J R 1996 [1971] *Scientific Knowledge and its Social Problems*, new edn. Transaction Publishers, New Brunswick, NJ

Rose S, Rose H 1976 *The Political Economy of Science*. Macmillan, London

Salter L 1988 *Mandated Science: Science and Scientists in the Making of Standards*. Kluwer Academic Publishers, Dordrecht, Holland

Sclove R E 1995 *Democracy and Technology*. The Guilford Press, New York

Szilard L 1961 *The Voice of the Dolphins and Other Stories*. Simon & Schuster, New York

Ziman J 1968 *Public Knowledge*. Cambridge University Press, Cambridge, UK

J. R. Ravetz

Science, Social Organization of

The sociology of science has been divided since about 1980 between those contending that science gains sociological significance because of its organizational location and forms, and those arguing that it should be understood for its knowledge-building practices. The two groups have tended to treat social organization in completely different ways, and have consciously developed their ideas in opposition to the others. However, both have used some notion of the social organization of science to explain the constitution of facts about nature and the development of ways of reworking nature for strategic advantage.

The tensions between the schools have been productive of new approaches to organizations and the operation of power. A scholar like Jasanoff, interested in political process and expertise, has used constructivist theories of scientific knowledge to show how central science has become to the legal system and regulatory structures (Jasanoff 1990). In contrast, a researcher like Bruno Latour, interested in the social struggles involved in making science 'real,' has demonstrated how the laboratory has become productive of powers that shape contemporary life (Latour 1993). This work, and much more like it, has begun to suggest

the fundamental ways that contemporary political systems and other major institutions depend on science and technology for their forms, operations, and legitimacy.

It would have been hard for sociologists to entirely avoid the correlation between the growth of modern States and the so-called scientific revolution, and the questions it raises about power and control of the natural world. Courts in the fifteenth and sixteenth centuries used scientists to help design military technology, do political astrology, and make the earth into a showplace of power (Grafton 1999, Masters 1998). Even members of the Royal Society in the seventeenth century (often depicted as a politically independent organization for science) dedicated much effort to addressing public problems with their research (Webster 1976). Practical as well as conceptual arts like cartography gained importance in Italy, France, and England as a tool for trade and territorial control. Double entry bookkeeping provided a way of legitimating both commerce and governmental actions through the management of 'facts' (Poovey 1998). Botany and the plant trade, particularly in Spain, The Netherlands, and France, enriched the food supply and increased the stock of medicinal herbs. Italian engineers tried to tame rivers, the Dutch and French built canals (Masters 1998), the English deployed medical police to ensure public health (Carroll 1998), and the French, English, and Germans worked on forestry (Scott 1998). Military engineering throughout Europe was revolutionized with a combination of classical architectural principles, and new uses of cannon fire and other weapons (Mukerji 1997). As Patrick Carroll argues, technoscience was not a product of the twentieth century, but already part of the culture of science (the engine science) of the seventeenth century (Carroll 1998). The point of science and engineering in this period was human efficacy, the demonstration of human capacities to know the world and transform it for effect. One manifestation of this was the cultivation of personal genius and individual curiosity, yielding a dispassionate science, but another was the constitution of political territories that were engineered for economic development and political legibility, and managed (in part) to maintain the strength and health of the population (Mukerji 1997, Scott 1998).

1. *Functionalist Foundations of the Sociology of Science*

The early sociologists of science, such as Merton and Ben-David, took for granted both historical and contemporary links between science and state power. They were steeped in the sociological literature on organizations that defined technology, at least, as central to the organization of major institutions—from the military to industry. Merton in his dissertation recognized the historical interest of courts and States in scientists and engineers, but traced the development of a disinterested culture of science, emanating from the places where thought was at least nominally insulated from political pollution: universities, scientific societies, and professional journals (Merton 1973, Crane 1972). Looking more directly at politics and science, Ben-David (1971), still interested in how social organization could promote excellence in science, considered differences in the organization of national science systems and their effects on thought. He assumed a kind of Mannheimian sociology of knowledge to argue that systems of research impact the progress of science. If location in the social system shaped what people could know, then the social organization of science was necessarily consequential for progress in science and engineering (Ben-David 1971).

These approaches to the organization of science were grounded in their historical moment—the Cold War period. Science and engineering were essential to the power struggle of East and West. It was commonly held that WWII had been won through the successful effort to dominate science and technology. Policymakers in the US and Europe wanted to gain permanent political advantages for their countries by constructing a system of research in science and engineering that would continue to dominate both thought and uses of natural resources. Western ideology touted the freedom of thought allowed in the non-Communist world as the route to the future. According to the Mertonians, history seemed to support this posture. England with its Royal Society independent of the government was the one which produced Newton—not France and Italy with their systems of direct state patronage (Crane 1972, Merton 1973). The result was a clear victory for the open society, and reason to be confident about American science, which was being institutionalized inside universities rather than (in most cases) national laboratories.

2. *The Sociology of Scientific Knowledge*

For all that the sociology of scientific knowledge (SSK) first presented itself as a radical subfield at odds with the functionalist tradition, it continued Merton's impulse to associate science less with politics than philosophy (Barnes et al. 1996). Taking the laboratory (rather than the Department of Defense), as the center of calculation for science (and science studies) was an effective way to imagine that power was not at stake in science—even in the US. Moreover, SSK presented good reasons to look more closely at scientific knowledge. In the Mertonian tradition, sociologists had discussed the relationship of organizational structures to scientific progress—as though sociologists could

and would know when science was flourishing. Sociologists of scientific knowledge were not prepared to make such judgments about scientific efficacy, nor passively willing to accept the word of scientists as informants on this. For SSK researchers, what was at stake was the philosophical problem of knowledge—how you know when an assertion about nature is a fact (Barnes et al. 1996, Collins 1985). This was precisely what Merton (1973) had declared outside the purview of sociology, but SSK proponents now transformed into a research question. How did insiders to the world of science make these kinds of determinations for themselves?

The implications for studying the organization of science were profound. Laboratories were the sites where facts were made, so they were the organizational focus of inquiry (Knorr-Cetina 1981, Latour and Woolgar 1979, Lynch 1985). Making a scientific fact was, by these accounts, a collective accomplishment, requiring the coordination of the cognitive practices both within and across laboratories. Laboratories had their own social structures—some temporary and some more stable. In most cases laboratory members distinguished between principal investigators and laboratory technicians, who had different roles in the constitution of knowledge. The significance of research results depended on the social credibility of the researchers. Science was work for gentlemen whose word could be trusted—not tradesmen or women even if they did the work (Shapin and Schaffer 1985). There were alliances across laboratories forged with ideas and techniques that were shared by researchers and were dedicated to common problems or ways of solving them (Pickering 1984). To make scientific truths required more than just an experiment that would confirm or disconfirm an hypothesis. The results had to be witnessed and circulated within scientific communities to make the 'facts' known (Shapin and Schaffer 1985). Scientific paradigms needed proponents to defend and promote them. Creating a scientific fact was much like a military campaign; it required a high degree of coordination of both people and things. It was a matter of gaining the moral stature in the scientific community to have a scientist's ideas taken as truthful. The tools accomplishing these ends were both social and cognitive (Shapin and Schaffer 1985).

Members of these two schools (Mertonian and SSK) may have envisioned the task of articulating a sociology of science differently, but they shared a basic interest in socially patterned ways of thinking about, mobilizing, and describing nature (or the nature of things). Now that the dust has settled on their struggle for dominance in sociology the epistemological break assumed to exist between them has come to seem less profound. Power circulates through laboratories, and scientific experts circulate through the halls of power (Haraway 1989, Jasanoff 1990, 1994, Mukerji 1989). Ways of organizing research affect both the constitution of knowledge and ways of life (Rabinow 1996). The world we know is defined and engineered through patterns of cognition that include manipulation of nature—in the laboratory and beyond.

3. *The Politics of Knowledge-making and Knowledge Claims*

Contemporary research in the sociology of science has shed new light on organizations that sociologists thought they understood before, but never examined for their cognitive processes and relations to nature. Jasanoff's work on the regulatory system in the US, for example, does not simply argue that concern about pollution and scientific regulation of other aspects of life has stimulated new research and yielded new information for policymakers—although that is true. She has shown how the legitimacy of the State has come to depend (to a surprising extent) on its claims to provide at least a minimal level of well-being for the population. Safe air, a healthy food supply, and good drinking water have become taken-for-granted aspects of political legitimacy that depend not only on manipulations of nature but also on the development of new strategies of reassurance. Regulators have not been able simply to ask scientists to use existing expertise to assess and ameliorate problems. They have had to cultivate sciences pertinent to the problems, and face the controversies about the results of new lines of research. The result is a new set of cognitive tools for science honed for policy purposes, and new pressures on political actors to understand and work with at least some of these measurements (Jasanoff 1990, 1994).

Similarly, political legitimacy in rich countries also rests on the government's ability to confront and address medical problems. As Steven Epstein pointed out in his study of AIDS research, the population now generally expects doctors, scientists, and policymakers to solve problems and keep the population healthy. Any fundamental disruption of this faith in expertise leads to anger and (in the US case he studied) political activism. Government officials (like public health workers) in these instances need not only to advocate and support research but also to make the government's health system seem responsive to public needs. This means that the dispassionate pursuit of scientific truths cannot dictate the practice of research, or the dispersement of drugs. Research protocols cannot be entirely determined by experts, but require debate with the activists as well as professional politicians for whom the problem is an issue (Epstein 1996).

Science is therefore used to manage public health (the body politic), and for creating a healthful environment for the citizenry. It is also used to design and manage infrastructures for the society as a whole. Computers are used to design roads, and manage toll systems.

They are employed by hospitals to define illnesses, codify medical practices, and determine courses of treatment for individuals. The military has not only used scientists and engineers (in Cold War style) to develop new weapons and create new modes for their delivery, but has mobilized these groups to develop national communication infrastructures—from road systems to airports to the Internet (Edwards 1996). This engineering is oddly an outgrowth of territorial politics—in this period when States are supposed (by theories from post-modernist to Marxist) to be dying due to globalization (Castells 1996). However, responsibility for health and well- being are circumscribed by national boundaries, and legal responsibility for them is kept within territorial boundaries and remains largely a problem for States.

4. *Post-colonial Studies of Science and Technology*

The territorial dimensions of legitimacy and public health are particularly apparent in the burgeoning post-colonial literature on science, technology, and medicine, showing the export to poor countries of research involving dangerous materials or medical risks (Rafael 1995). The export of risk was perhaps most dramatically illustrated by the US use of an atoll in the Pacific Ocean for bomb tests, but there have been numerous less dramatic instances of this. Manufacturers using dangerous materials have been encouraged to set up factories in Third World countries where economic growth has been considered a higher priority than environmental safety (Jasanoff 1994). Medical researchers in need of subjects have frequently turned to colonial populations for them (Rafael 1995). These practices make it obvious that dangerous research and manufacture are not necessarily abandoned because of the risks. They are simply done in those less powerful places within a state's sphere of influence where legitimacy is not at stake. In these instances, political regulation at home has not necessarily created a moral revolution in science, but a physical dissociation of practices and responsibility. Outside the social world of Western gentlemen, there has often been little concern on the part of researchers about their moral stature. Governments have not systematically worried about the normative consequences of technological development. Instead, poor people have been treated (like prisoners at home) as a disposable population of test subjects—just what AIDS patients refused to be (Epstein 1996, Jasanoff 1994, Rafael 1995).

This pattern is both paralleled by, and connected to, the export of high technology into post-colonial areas of the world. The growth of manufacturing in Third World countries, using computerized production systems, and the growth of computing itself in both India and Africa for example, testify to another pattern of export of science and technology (Jasanoff 1994, Jules-Rosette 1990). These practices are usually presented as ways that corporations avoid local union rules, and deploy an educated workforce at low cost (Castells 1996). However, they are also ways large corporations have found to limit in-migration of labor from post-colonial regions, and to avoid political responsibility for the health of laboring people from poorer parts of the world. Less a case of exporting risk, it is a way of exporting responsibility for it (Jasanoff 1994).

5. *Commercial Stakes in Scientific Thought*

It would be easy to think that while politics is driving some areas of research, there remains a core of pure science like the one desired by Merton. However, commercial as well as political forces work against this end (Martin 1991). Even the publishing system that was supposed to buffer science from the workings of power has turned out to be corruptible. It is not simply that scientific texts (the immutable mobiles of Latour and Woolgar (1979) have found political uses; texts have not been stabilized by being put in print. As Adrian Johns (1998) has shown, simply because printing technology could fix ideas and stabilize authorship has not meant that the publishing industry would use it this way. In seventeenth-century England—the heyday of the scientific revolution—publications were often pirated, changed for commercial purposes, or reattributed. Scientific authorship never did unambiguously extend empiricism and accurate beliefs about nature, or give scientific researchers appropriate recognition for their work. Publishing in science was just another part of the book trade, and was managed for profit. To this day, commercial pressures as well as peer review shapes the public record in science. The purpose of the science section in newspapers is to sell copies, not promote scientific truth. However, scientists still frequently use the popular press to promote their research and advance their careers in science (Epstein 1996).

The practices of science and engineering themselves are not so clearly detached either. Commercial interests in biotechnology and computing have powerful effects on the organization of research, and relations between the university and industry (Rabinow 1996). Even though the Cold War is over, scientists still often cannot publish what they learn from DOD research (which includes work in computing). They rely on external funding, and so must study what is of interest to the government. They are under careful supervision of administrators when they export engineering practices from the laboratory into the world (Mukerji 1989). Dreams of democracy served by science and technology seem hard to sustain.

6. Beyond Power/Knowledge

The broad range of contemporary studies of the politics of science and engineering do not just manifest a revival of interest in the political dimensions/connections of science and engineering in the post-Foucaultian world of power/knowledge, but they manifest new understandings of how power operates through multiple organizational forms. Scientists have (both historically and in the present) aided in the articulation of a system of power based on the principles familiar to the Frankfurt School—the domination of nature for the domination of people. Researchers in science studies who focus on the political mobilization of research for organizational advantage are now making clearer how strategic scientific management of the natural world works and does not work. It seems that governments and industry historically have not so much attained the ideas about nature they needed or paid for, but that scientists, in pursuing knowledge, have also produced means of dominating nature that have been used (and to some extent contained) by those institutions (Mukerji 1989). Modern states, in funding the cultivation of cognitive systems for learning about and managing natural resources, nuclear power, chemical pollutants, and viruses, have generated new patterns of domination but have also opened themselves up to new questions of legitimacy and, in some cases (Epstein 1996), to a redistribution of expertise. The system of scientific and engineering research is not just productive of ideas, but also transportation systems, research animals, laboratories themselves, and new technologies (like the Internet) (Edwards 1996, Kohler 1994, Rabinow 1996). The result is not just a brains trust of scientists, but an entire sociotechnical environment built for strategic effect (Cronon 1991). The cognitive systems of science and engineering are not just ways of coordinating thought through language to reach the truth, but ways of making the world again to reflect and carry human intelligence (and stupidity) about nature.

See also: Academy and Society in the United States: Cultural Concerns; Disciplines, History of, in the Social Sciences; Human Sciences: History and Sociology; Kantian Ethics and Politics; Normative Aspects of Social and Behavioral Science; Paradigms in the Social Sciences; Research and Development in Organizations; Scientific Academies in Asia; Scientific Knowledge, Sociology of; Truth and Credibility: Science and the Social Study of Science; Universities, in the History of the Social Sciences

Bibliography

Barnes B, Bloor D, Henry J 1996 *Scientific Knowledge*. Chicago University Press, Chicago

Ben-David J 1971 *The Scientist's Role in Society*. Prentice-Hall, Englewood Cliffs, NJ

Bowker G, Star S L 1999 *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge, MA

Carroll P 1998 *Ireland: Material Construction of the Techno-scientific State*. Dissertation, University of California, San Diego, CA

Castells M 1996 *The Rise of the Network Society*. Blackwell, London

Collins H 1985 *Changing Order*. Sage, Beverly Hills, CA

Crane D 1972 *Invisible Colleges*. University of Chicago Press, Chicago

Cronon W 1991 *Nature's Metropolis*. Norton, New York

Edwards P N 1996 *The Closed World*. MIT Press, Cambridge, MA

Epstein S 1996 *Impure Science*. University of California Press, Berkeley, CA

Grafton A 1999 *Cardano's Cosmos*. Harvard University Press, Cambridge, MA

Habermas J 1975 *Legitimation Crisis*. Beacon Press, Boston

Haraway D 1989 *Primate Visions*. Routledge, New York

Jasanoff S 1990 *The Fifth Branch*. Harvard University Press, Cambridge, MA

Jasanoff S 1994 *Learning from Disaster*. University of Pennsylvania Press, Philadelphia, PA

Johns A 1998 *The Nature of the Book*. University of Chicago Press, Chicago

Jules-Rosette B 1990 *Terminal Signs: Computers and Social Change in Africa*. Mouton de Gruyter, New York

Knorr-Cetina K 1981 *The Manufacture of Knowledge*. Pergamon Press, New York

Kohler R 1994 *Lords of the Fly*. University of Chicago Press, Chicago

Latour B 1993 *We Have Never Been Modern*. Harvard University Press, Cambridge, MA

Latour B, Woolgar S 1979 *Laboratory Life*. Sage Publications, Beverly Hills, CA

Lynch M 1985 *Art and Artifact in Laboratory Science*. Routledge and Kegan Paul

Martin B 1991 *Scientific Knowledge in Controversy*. State University of New York Press, Albany, NY

Masters R 1998 *Fortune is a River*. Free Press, New York

Merton R K 1973 *The Sociology of Science*. University of Chicago Press, Chicago

Mukerji C 1989 *A Fragile Power*. Princeton University Press, Princeton, NJ

Mukerji C 1997 *Territorial Ambitions and the Gardens of Versailles*. Cambridge University Press, New York

Pickering A 1984 *Constructing Quarks*. University of Chicago Press, Chicago

Poovey M 1998 *A History of the Modern Fact*. University of Chicago Press, Chicago

Proctor R 1991 *Value-free Science?* Harvard University Press, Cambridge, MA

Rabinow P 1996 *The Making of PCR*. University of Chicago Press, Chicago

Rafael V 1995 *Discrepant Histories*. Temple University Press, Philadelphia, PA

Scott J C 1998 *Seeing Like a State*. Yale University Press, New Haven, CT

Shapin S, Schaffer S 1985 *Leviathan and the Air-pump*. Princeton University Press, Princeton, NJ

Webster C 1976 *The Great Instauration*. Holmes and Meier, New York

C. Mukerji

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Science, Sociology of

To make science the object of sociological analysis directs attention to the production and consumption of scientific knowledge in diverse cultural contexts, institutional structures, local organizations, and immediate settings. The sociology of science divides into three broad lines of inquiry, each distinguished by a particular mix of theories and methods. The earliest systematic studies (mostly from the 1950s to the early 1970s) focus on the structural contexts of scientists' behavior: what rules govern the pursuit of scientific knowledge, how are scientists judged and rewarded, how is scientific research broken up into dense networks of specialists? In the 1980s, sociologists shift their attention to the practices through which scientific knowledge is constructed—at the laboratory bench or in the rhetoric of professional papers. Starting in the 1990s, science is put in more encompassing societal contexts, as sociologists examine scientists as purveyors of cognitive authority, and explore their linkages to power, politics, and the economy.

1. Precursors

It is remarkable how much the literature in sociology of science is bunched into the last third of the twentieth century. Perhaps only after the deployment of nuclear weapons, or only after genetic engineering raised eugenic nightmares, could sociologists begin to think about science as a social problem rather than as a consistent solution; or maybe earlier generations of sociologists were guided by epistemological assumptions that rendered true scientific knowledge immune from social causes—thus putting it outside the orbit of sociological explanation.

1.1 Classical Anticipations

'Science' is nowhere indexed in Max Weber's encyclopedic *Economy and Society*, a measure of his unwillingness or inability to see it as a consequential factor in human behavior or social change. Weber's interest in science was largely methodological and political. Could the causal models employed so effectively in the natural sciences be used as well to study social action? Does the objectivity and neutrality of the social scientist preclude involvement in political activity?

Emile Durkheim also sought to institutionalize sociology by making its methods appear scientifically precise, but at the same time considered scientific knowledge as an object of sociological study. Durkheim suggested that basic categories of thought and logic (such as time and space) are social in origin, in that they correspond to fundamental social categories (such as the division of a society by family or gender). However, as human societies grew in size and as their institutions became functionally differentiated, a distinctively scientific pursuit of knowledge was gradually insulated from such social causes. The observable facts of modern science, Durkheim concluded, were in accord with the reality of the physical world—a position that forestalled examinations of how observable facts are also shaped by the culture and communities in which they arise.

Karl Marx's materialism would seem to commit him to the idea that all beliefs arise amid historically specific conditions of production, as they are shaped by the goals and interests of a ruling class. The rise of science in seventeenth-century Europe is intimately bound with the rise of industrial capitalism and, for Marx, can be explained in terms of the utilities of science-based technologies for improving productivity and enlarging surplus value. But although the rate of scientific growth may be explained by its congruence with the interests of the bourgeoisie, Marx seems to suggest that the content of scientific claims inside professionalized research networks is nonideological—that is, an objective account of natural reality.

1.2 Science in the Sociology of Knowledge

Even more surprising is the failure of systematic sociological studies of science to emerge from a blossoming sociology of knowledge in the 1920s and 1930s. Neither Max Scheler nor Karl Mannheim, authors of foundational treatises on the social determinants of knowledge, inspired sustained inquiry into the social determinants of science—probably because both distinguished scientific knowledge from other kinds in a way that truncated what sociology could say about it. Scheler isolated the content of scientific knowledge—and the criteria for ascertaining validity—by describing these as absolute and timeless essences, not shaped by social interests. The effects of social structure (specifically, the power of ruling elites) is limited to selections of problems and beliefs from that self-contained and essential realm of ideas.

Mannheim sustained the neo-Kantian distinction between formal knowledge of the exact sciences and socio-historical knowledge of culture. Phenomena of the natural world are invariant, Mannheim suggests, and so therefore are criteria for deciding truth (i.e., impartial observations based on accurate measurements). In contrast, cultural phenomena become meaningful only as they are constructed through interest-laden judgments of significance, which are neither impartial nor invariant, and thus they are amenable to sociological explanation.

Robert K. Merton's 1938 classic *Science, Technology and Society in Seventeenth-century England* (see Merton 1973) tackles a fundamental problem:

why did modern science emerge with a flourish in seventeenth-century England? His answer has become known as the 'Merton Thesis:' an ethos of Puritanism that provided both the motivating force and legitimating authority for the pursuit of scientific inquiry. Certain religious values—e.g., God is glorified by an appreciation of his handiwork in Nature, or Blessed Reason separates human from beast—created a cultural context fertile for the rise of science. Merton also explains shifts in the foci of research attention among the early modern 'natural philosophers' by connecting empirical inquiry to the search for technological solutions to practical problems in mining, navigation, and ballistics.

2. *Social Organization of the Scientific Community*

When concerted sociological studies of science began in the late 1950s and 1960s, research centered on the institutions or social structures of science—with relatively less attention given to the routine practices involved in making knowledge or to the wider settings in which science was conducted. This work was largely inspired by theories of structural-functional analysis, which ask how the community of scientists is organized in order to satisfy modern society's need for certified, reliable knowledge. One distinctive feature of this first phase is a reliance on quantitative methods of analysis. With statistical data drawn from surveys of scientists and from the *Science Citation Index* (and other bibliometric sources), sociologists developed causal models to explain individual variations in research productivity and used topographical techniques such as multidimensional scaling to map the dense networks of scientists working at a research front.

2.1 *Norms of Science*

The shift from analyzing science in society to analyzing its internal social organization was effected in Merton's 1942 paper on the normative structure of science (in Merton 1973). Written under the shadow of Nazism, Merton argues that the success of scientists in extending certified knowledge depends, at once, on a salutary political context (namely democracy, which allows science a measure of autonomy from political intrusions and whose values are said to be congruent with those of science—quite unlike fascism) and on an internal institutionalized ethos of values held to be binding upon the behavior of scientists. This ethos comprised the famous norms of science: scientists should evaluate claims impersonally (universalism), share all findings (communism), never sacrifice truth for personal gain (disinterestedness) and always question authority (organized skepticism). Behavior consonant with these moral expectations is functional for the growth of reliable knowledge, and for that reason young scientists learn through precept how they are expected to behave, conformity is rewarded, and transgressions are met with outrage.

Subsequent work ignored Merton's conjectures about science and democracy, as sociologists instead pursued implications of the four social norms. Studies of behavioral departures from these norms—ethnocentrism, secrecy, fraud, plagiarism, dogmatism—precipitated debates over whether such deviance is best explained by idiosyncratic characteristics of a few bad apples or changing structural circumstances (such as commercialization of research) that might trigger increases in such behavior. Sociologists continue to debate the possibility that Merton's norms are better explained as useful ideological justifications of scientists' autonomy and cognitive authority. Other research suggests that the norms guiding scientific conduct vary historically, vary among disciplines, vary among organizational contexts (university research vs. military or corporate research), and vary even in their situational interpretation, negotiation and deployment—raising questions about whether the norms identified by Merton are functionally necessary for enlarging scientific knowledge.

2.2 *Stratification and Scientific Careers*

The norm of universalism in particular has elicited much empirical research, perhaps because it raises questions of generic sociological interest: how is scientific performance judged, and how are inequalities in the allocation of rewards and resources best described and explained? With effective quantitative measures of individual productivity (number of publications or citations to one's work), resources (grant dollars), and rewards (prizes, like the Nobel), sociologists have examined with considerable precision the determinants of individual career success or failure. Competition among scientists is intense, and the extent of inequality high: the distribution of resources and rewards in science is highly skewed. A small proportion of scientists publish most research papers (and those papers collect most citations), compete successfully for research grants and prestigious teaching posts, achieve international visibility and recognition, and win cherished prizes.

Debate centers on whether these observed inequalities in the reward system of science are compatible with the norm of universalism—which demands that contributions to knowledge be judged on their scientific merit, with resources and opportunities meted out in accordance with those judgments. The apparent elitism of science may result from an 'accumulation of advantage': work by relatively more eminent or well-positioned scientists is simply noticed

more and thus tends to receive disproportional credit—which (over time) enlarges the gap between the few very successful scientists and everybody else. Such a process may still be universalistic because it is functional for the institutional goal of science: giving greater attention to research of those with accomplished track-records may be an efficient triage of the overwhelming number of new candidate theories or findings. Others suggest that particularism contributes to the stratification of scientists—old boy networks that protect turf and career reputations by rewarding sycophants. The underrepresentation of women in the higher echelons of science has called attention to sometimes subtle sexism that occurs early in the scientific career (restricted access to well-connected mentors, or essential research equipment, or opportunities to collaborate and assignment to trivial problems or mind-numbing tasks).

2.3 *Institutionalization of the Scientific Role*

A separate line of sociological inquiry (exemplified in work by Joseph Ben-David 1991 and Edward Shils) seeks an explanation for how science first became a remunerable occupation—and later, a profession. How did the role of the scientist emerge from historically antecedent patterns of amateurs who explored nature part-time and generally at their own expense? The arrival of the 'scientist' as an occupational self-identification with distinctive obligations and prerogatives is inseparable from the institutionalization of the modern university (itself dependent upon government patronage). Universities provided the organizational form in which the practice of science could become a full-time career—by fusing research with teaching, by allowing (ironically) the historic prestige of universities as centers of theology and scholasticism to valorize the new science, and by providing a bureaucratic means of paying wages and advancing careers.

The scientific role has also been institutionalized in corporate and government labs. The difficulties of transporting a 'pure science' ideal of university-based research into these very different organizational settings have been the object of considerable sociological attention. Scientists in industry or government face a variety of competing demands: their research is often directed to projects linked to potential profits or policy issues rather than steered by the agenda of their discipline or specialty; the need to maintain trade secrets or national security hampers the ability of scientists in these settings to publicize their work and receive recognition for it. And, as Jerome Ravetz (1971) suggests, the intrusion of 'bureaucratic rationality' into corporate and state science compromises the craft character of scientific work: largely implicit understandings and skills shared by the community of scientists and vital for the sustained accumulation of scientific wisdom have little place in accountabilities driven by the bottom-line or policy-relevance.

2.4 *Disciplines and Specialties*

Sociologists use a variety of empirical indicators to measure the social and cognitive connections among scientists: self-reports of those with whom a scientist exchanges ideas or preprints, subject-classifications of publications in topical indexes or abstract journals, lineages of mentor–student relationshships or collaborations, patterns of who cites whom or is cited with whom ('co-citation'). The networks formed by such linkages show occasional dense clusters of small numbers of scientists whose informal communications are frequent, who typically cite each other's very recent papers, and whose research focusses on some new theory, innovative method, or breakthrough problem. Emergence of these clusters—for example, the birth of radio astronomy in England after WWII, as described by David Edge and Michael Mulkay (1976)—is a signal that science has changed, both cognitively and socially: new beliefs and practices are ensconced in new centers for training or research with different leaders and rafts of graduate students. Over time, these specialties evolve in a patterned way: the number of scientists in the network becomes much larger and connections among them more diffuse, the field gets institutionalized with the creation of its own journals and professional associations, shattering innovations become less common as scientists work more on filling in details or adding precision to the now-aging research framework. As one specialty matures, another dense cluster of scientists emerges elsewhere, as the research front or cutting edge moves on.

3. *Sociology of Scientific Knowledge*

A sea-change in sociological studies of science began in the 1970s with a growing awareness that studies of the institutional and organizational contexts shaping scientists' behavior could not illuminate sufficiently the processes that make science science: experimental tinkering, sifting of evidence, negotiation of claims, replacement of old beliefs about nature with new ones, achievement of consensus over the truth. All of these processes—observation, getting instruments and research materials (e.g., mice) to work, logic, criteria for justifying a finding as worthy of assent, choices among theories, putting arguments into words or pictures, persuading other scientists that you are correct—are uncompromisingly social, cultural, and historical phenomena, and so sociologists set about to explain and interpret the content of scientific knowledge by studying the routine practices of scientific work.

This research is guided by constructivist theories (and, less often, ethnomethodology), which center attention on the practically accomplished character of social life. Rather than allow *a priori* nature or given social structures to explain behavior or belief, constructivist sociologists examine how actors incessantly make and remake the structural conditions in which they work. Such research relies methodologically on historical case studies of scientific debate, up-close ethnographic observations of scientific practices, and on interpretative analysis of scientific texts.

3.1 Sociology of Discovery

Diverse studies of scientific discovery illustrate the range of sociological perspectives brought to bear on these consequential events. An early line of inquiry focusses on the social and cognitive contexts that cause the timing and placing of discoveries: why did these scientists achieve a breakthrough then and there? Historical evidence points to a pattern of simultaneous, multiple and independent discoveries—that is, it is rare for a discovery to be made by a scientist (or a local team) who are the only ones in the world doing research on that specific question. Because honor and recognition are greatest for solutions to the perceivedly 'hottest' problems in a discipline, the best scientists are encouraged by the reward system of science to tackle similar lines of research. But these same social structures can also forestall discovery or engender resistance to novel claims. Cognitive commitments to a long-established way of seeing the natural world (reinforced by reputations and resources dependent upon those traditional perspectives) can make it difficult for scientists to see the worthiness of a new paradigm. Resistance to new ideas seems to be greatest among older scientists, and in cases where the proposed discovery comes from scientists with little visibility or stature within the specialty or discipline that would be transformed.

More recent sociological research considers the very idea of 'discovery' as a practical accomplishment of scientists. Studies inspired by ethnomethodology offer detailed descriptions of scientific work 'first-time-through,' taking note of how scientists at the lab bench decide whether a particular observation (among the myriad observations) constitutes a discovery. Other sociologists locate the 'moment' of discovery in downstream interpretative work, as scientists narrate first-time-through research work with labels such as 'breakthrough.' Such discovery accounts are often sites of dissensus, as scientists dispute the timing or implications of an alleged discovery amid ongoing judgments of its significance for subsequent research initiatives or allocations of resources. These themes—interests, changing beliefs, ordinary scientific work, post-hoc accountings, dissent, persuasion—have become hallmarks of the sociology of scientific knowledge.

3.2 Interests and Knowledge-change

In the mid-1970s to the 1980s, sociology of science took root at the Science Studies Unit in Edinburgh, as philosopher David Bloor developed the 'strong programme,' while Barry Barnes, Steven Shapin, Donald MacKenzie, and Andrew Pickering developed its sociological analog—the 'interest model.' The goal is to provide causal explanations for changes in knowledge—say, the shift from one scientific understanding of nature to a different one. Scientists themselves might account for such changes in terms of greater evidence, coherence, robustness, promise, parsimony, predictive power, or utility of the new framework. Sociologists, in turn, account for those judgments in terms of social interests of scientists that are either extended or compromised by a decision to shift to the new perspective. What becomes knowledge is thus contingent upon the criteria used by a particular community of inquirers to judge competing understandings of nature, and also upon the goals and interests that shape their interpretation and deployment of those criteria. Several caveats are noted: interests are not connected to social positions (class, for example, or nationality, discipline, specialty) in a rigidly deterministic way; social interests may change along with changes in knowledge; choices among candidate knowledge-claims are not merely strategic—that is, calculations of material or symbolic gains are bounded by considerable uncertainty and by a shared culture of inquiry that provides standards for logical or evidential adequacy and for the proper use of an apparatus or concept.

Drawing on historical case studies of theoretical disputes in science—nineteenth-century debates over phrenology and statistical theory, twentieth-century debates among high-energy physicists over quarks—two different kinds of interests are causally connected to knowledge-change. Political or ideological commitments can shape scientists' judgments about candidate knowledge claims: the development of statistical theories of correlation and regression by Francis Galton, Karl Pearson and R. A. Fisher depended vitally on the utility of such measures for eugenic objectives. Different social interests arise from the accumulated expertise in working with certain instruments or procedures, which incline scientists to prefer theories or models that allow them to capitalize on those skills.

3.3 Laboratory Practices and Scientific Discourse

As sociologists moved ever closer to the actual processes of 'doing science,' their research divided into

two lines of inquiry: some went directly to the laboratory bench seeking ethnographic observations of scientists' practices *in situ*; others examined scientists' discourse in talk and texts—that is, their accounting practices. These studies together point to an inescapable conclusion: there is nothing not-social about science. From the step-by-step procedures of an experiment to writing up discovered facts for journal publication, what scientists do is describable and explicable only as social action—meaningful choices contingent on technical, cognitive, cultural, and material circumstances that are immediate, transient, and largely of the scientists' own making.

Laboratory ethnographies by Karin Knorr-Cetina (1999), Michael Lynch (1993), and Bruno Latour and Steve Woolgar (1986) reveal a science whose order is not to be found in transcendent timeless rules of 'scientific method' or 'good lab procedures,' but in the circumstantial, pragmatic, revisable, and iterative choices and projects that constitute scientific work. These naturalistic studies emphasize the local character of scientific practice, the idea that knowledge-making is a process situated in particular places with only these pieces of equipment or research materials or colleagues immediately available. Never sure about how things will turn out in the end, scientists incessantly revise the tasks at hand as they try to get machines to perform properly, control wild nature, interpret results, placate doubting collaborators, and rationalize failures. Even methodical procedures widely assumed to be responsible for the objective, definitive, and impersonal character of scientific claims—experimental replication, for instance—are found to be shot-through with negotiated, often implicit, and potentially endless judgments about the competence of other experimentalists and the fidelity of their replication-attempts to the original (as Harry Collins (1992) has suggested).

Ethnographic studies of how scientists construct knowledge in laboratories compelled sociologists then to figure out how the outcomes of those mundane contextual practices (hard facts, established theories) could paradoxically appear so unconstructed—as if they were given in nature all along, and now just found (not made). Attention turned to the succession of 'inscriptions' through which observations become knowledge—from machine-output to lab notebook to draft manuscript to published report. Scientists' sequenced accounts of their fact-making rhetorically erase the messy indeterminacy and opportunism that sociologists have observed at the lab bench, and substitute a story of logic, method, and inevitability in which nature is externalized as waiting to be discovered. Such studies of scientific discourse have opened up an enduring debate among constructivist sociologists of science: those seeking causal explanations for scientists' beliefs treat interests as definitively describable by the analyst, while others (Gilbert and Mulkay 1984, Woolgar 1988) suggest that sociologists must respect the diversity of participants' discursive accounts of their interests, actions, or beliefs—and thus treat actors' talk and text not as mediating the phenomena of study but as constituting them.

3.4 *Actor-networks and Social Worlds*

After sociologists worked to show how science is a thoroughly social thing, Bruno Latour (1988) and Michel Callon (1986) then retrieve and reinsert the material: science is not only about facts, theories, interests, rhetoric, and power but also about nature and machines. Scientists accomplish facts and theories by building 'heterogenous networks' consisting of experimental devices, research materials, images and descriptive statistics, abstract concepts and theories, the findings of other scientists, persuasive texts—and, importantly, none of these are reducible to any one of them, nor to social interests. Things, machines, humans, and interests are, in the practices of scientists, unendingly interdefined in and through these networks. They take on meanings via their linkages to other 'actants' (a semiotic term for anything that has 'force' or consequence, regardless of substance or form). In reporting their results, scientists buttress claims by connecting them to as many different actants as they can, in hopes of defending the putative fact or theory against the assault of real or potential dissenters. From this perspective, length makes strength, that is, the more allies enrolled and aligned into a network—especially if that network is then stabilized or 'black boxed'—the less likely it is that dissenters will succeed in disentangling the actants and thereby weaken or kill the claim. Importantly for this sociology of science, the human and the social are decentered, in an ontology that also ascribes agency to objects of nature or experimental apparatuses.

Actor-network theory moved the sociological study of science back outside the laboratory and professional journal—or, rather, reframed the very idea of inside and outside. Scientists and their allies 'change the world' in the course of making secure their claims about nature, and in the same manner. In Latourian vernacular, not only are other scientists, bits of nature or empirical data enlisted and regimented, but also political bodies, protest movements, the media, laws and *hoi polloi*. When Louis Pasteur transformed French society by linking together microbes, anthrax, microscopes, laboratories, sick livestock, angry farmers, nervous Parisian milk-drinkers, public health officials, lawmakers, and journalists into what becomes a 'momentous discovery,' the boundary between science and the rest of society is impossible to locate. Scientists are able to work autonomously at their benches precisely because so many others outside the lab are also 'doing science,' providing the life

support (money, epistemic acquiescence) on which science depends.

The boundaries of science also emerge as theoretically interesting in related studies derived from the brand of symbolic interactionism developed by Everett Hughes, Herbert Blumer, Anselm Strauss, and Howard Becker (and extended into research on science by Adele Clarke 1990, Joan Fujimura, and Susan Leigh Star). On this score, science is work—and, instructively, not unlike work of any other kind. Scientists (like plumbers) pursue doable problems, where 'doability' involves the articulation of tasks across various levels of work organization: the experiment (disciplining research subjects), the laboratory (dividing labor among lab technicians, grad students, and postdocs), and 'social worlds' (the wider discipline, funding agencies, or maybe animal-rights activists). Scientific problems become increasingly doable if 'boundary objects' allow for cooperative intersections of those working on discrete projects in different social worlds. For example, success in building California's Museum of Vertebrate Zoology in the early twentieth century depended upon the standardization of collection and preparation practices (here, the specimens themselves become boundary objects) that enabled biologists to align their work with trappers, farmers, and amateur naturalists in different social worlds. As in actor-network theory, sociologists working in the 'social worlds' tradition make no assumption about where science leaves off and the rest of society begins—those boundaries get settled only provisionally, and remain open to challenge from those inside and out.

4. *Science as Cultural Authority*

It is less easy to discern exactly what the sociology of science is just now, and where it is headed. Much research centers on the position of science, scientists, and scientific knowledge in the wider society and culture. Science is often examined as a cognitive or epistemic authority; scientists are said to have the legitimate power to define facts and assess claims to truth. This authority is not treated as an inevitable result of the character or virtue of those who become scientists, the institutional organization of science (norms, for example) or of the 'scientific method.' It is, rather, an accomplished resource pursued strategically by a profession committed not only to extending knowledge but also to the preservation and expansion of its power, patronage, prestige, and autonomy.

No single theoretical orientation or methodological program now prevails. Constructivism remains appealing as a means to render contingent and negotiable (rather than 'essential') those features of scientific practice said to justify its epistemic authority. But as the agenda in the sociology of science shifts from epistemological issues (how is knowledge made?) to political issues (whose knowledge counts, and for what purposes?), constructivism has yielded to a variety of critical theories (Marxism, feminism and postmodernism) that connect science to structures of domination, hierarchy, and hegemony. A popular research site among sociologists of science is the set of occasions where scientists find their authority challenged by those whose claims to knowledge lack institutional legitimacy.

4.1 *Credibility and Expertise*

Steven Shapin (1994) (among others) has identified credibility as a constitutive problem for the sociology of science. Whose knowledge-claims are accepted as believable, trustworthy, true or reliably useful—and on what grounds? Plainly, contingent judgments of the validity of claims depend upon judgments of the credibility of the claimants—which has focussed sociological attention on how people use (as they define) qualities such as objectivity, expertise, competence, personal familiarity, propriety, and sincerity to decide which candidate universe becomes provisionally 'real.' A long-established line of sociological research examines those public controversies that hinge, in part, on 'technical' issues. Case studies of disputes over environmental and health risks find a profound ambivalence: the desire for public policy to be decided by appropriate legislative or judicial bodies in a way that is both understandable and accountable to the populace runs up against the need for expert skills and knowledge monopolized by scientific, medical or engineering professionals. Especially when interested publics are mobilized, such disputes often become 'credibility struggles' (as Steven Epstein (1998) calls them). In his study of AIDS politics, Epstein traces out a shift from activists' denunciation of scientists doing research on the disease to their gaining a 'seat at the table' by learning enough about clinical trials and drug development to participate alongside scientists in policy decisions. In this controversy, as in many others, the cultural boundaries of science are redrawn to assign (or, alternatively, to deny) epistemic authority to scientists, would-be scientists, citizens, legislators, jurists, and journalists.

4.2 *Critique of Science*

Recent sociological studies have themselves blurred the boundaries between social science and politics by examining the diverse costs and benefits of science. Whose agenda does science serve—its own? global capital? political and military elites? colonialism? patriarchy? the Earth's? Studies of molecular biology and biotechnology show how the topics chosen for

scientific research—and the pace at which they are pursued—are driven by corporate ambitions for patents, profits, and market-share. Related studies of the Green Revolution in agricultural research connect science to imperialist efforts to replace indigenous practices in less developed countries with 'advanced technologies' more consonant with the demands of global food markets. Feminist researchers are equally interested in the kinds of knowledge that science brings into being—and, even more, the potential knowledges not sought or valorized. In the nineteenth century, when social and natural science offered logic and evidence to legitimate patriarchal structures, other styles of inquiry and learning practiced among women (Parisian salons, home economics, midwifery, and cookery) are denounced as unscientific and, thus, suspect. Other feminists challenge the hegemony of scientific method, as a way of knowing incapable of seeing its own inevitable situatedness and partiality; some suggest that women's position in a gender-stratified society offers distinctive epistemic resources that enable fuller and richer understandings of nature and culture.

These critical studies share an interest in exposing another side of science: its historical complicity with projects judged to be inimical to goals of equality, human rights, participatory democracy, community and sustainable ecologies. They seek to fashion a restructured 'science' (or some successor knowledge-maker) that would be more inclusive in its practitioners, more diverse in its methods, and less tightly coupled to power.

See also: Actor Network Theory; Cultural Studies of Science; Laboratory Studies: Historical Perspectives; Norms in Science; Science and Technology Studies: Ethnomethodology; Science, Social Organization of; Scientific Controversies; Scientific Culture; Scientific Knowledge, Sociology of; Strong Program, in Sociology of Scientific Knowledge; Truth and Credibility: Science and the Social Study of Science

Bibliography

Barber B, Hirsch W (eds.) 1962 *The Sociology of Science*. Free Press, Glencoe, IL

Barnes B, Bloor D, Henry J 1996 *Scientific Knowledge: A Sociological Analysis*. University of Chicago Press, Chicago

Barnes B, Edge D (eds.) 1982 *Science in Context*. MIT Press, Cambridge, MA

Ben-David J 1991 *Scientific Growth*. University of California Press, Berkeley, CA

Callon M 1986 Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Bliuex Bay. In: Law J (ed.) *Power, Action and Belief*. Routledge & Kegan Paul, London

Clarke A E 1990 A social worlds research adventure. In: Cozzens S E, Gieryn T F (eds.) *Theories of Science in Society*. Indiana University Press, Bloomington, IN

Clarke A E, Fujimura J H (eds.) 1992 *The Right Tools for the Job: At Work in Twentieth-century Life Sciences*. Princeton University Press, Princeton, NJ

Collins H M 1992 *Changing Order: Replication and Induction in Scientific Practice*. University of Chicago Press, Chicago

Edge D O, Mulkay M J 1976 *Astronomy Transformed*. Wiley, New York

Epstein S 1998 *Impure Science: AIDS, Activism, and the Politics of Knowledge*. University of California Press, Berkeley, CA

Gieryn T F 1999 *Cultural Boundaries of Science: Credibility on the Line*. University of Chicago Press, Chicago

Gilbert G N, Mulkay M 1984 *Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse*. Cambridge University Press, Cambridge, UK

Hagstrom W O 1965 *The Scientific Community*. Basic Books, New York

Harding S 1986 *The Science Question in Feminism*. Cornell University Press, Ithaca, NY

Jasanoff S, Markle G E, Petersen J C, Pinch T (eds.) 1995 *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA

Kloppenburg J R Jr. 1988 *First the Seed: The Political Economy of Plant Biotechnology 1492–2000*. Cambridge University Press, Cambridge, UK

Knorr-Cetina K 1999 *Epistemic Cultures*. Harvard University Press, Cambridge, MA

Knorr-Cetina K, Mulkay M J (eds.) 1983 *Science Observed*. Sage, Beverly Hills, CA

Latour B 1988 *Science in Action*. Harvard University Press, Cambridge, MA

Latour B, Woolgar S 1986 *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, NJ

Long J S, Fox M F 1995 Scientific careers: Universalism and particularism. *Annual Review of Sociology* **21**: 45–71

Lynch M 1993 *Scientific Practice and Ordinary Action*. Cambridge University Press, Cambridge, UK

Merton R K 1973 *The Sociology of Science*. University of Chicago Press, Chicago

Mulkay M J 1979 *Science and the Sociology of Knowledge*. George Allen & Unwin, London

Mulkay M J 1980 Sociology of science in the West. *Current Sociology* **28**: 1–184

Nelkin D (ed.) 1984 *Controversy: The Politics of Technical Decisions*. Sage, Beverly Hills, CA

Ravetz J R 1971 *Scientific Knowledge and its Social Problems*. Oxford University Press, Oxford, UK

Shapin S 1994 *A Social History of Truth*. University of Chicago Press, Chicago

Shapin S 1995 Here and everywhere: Sociology of scientific knowledge. *Annual Review of Sociology* **21**: 289–321

Woolgar S 1988 *Science: The Very Idea*. Tavistock, London

Zuckerman H 1988 The sociology of science. In: Smelser N (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA

T. F. Gieryn

Science, Technology, and the Military

Technology has changed warfare since time immemorial. The invention of gunpowder, artillery, and rifles revolutionized warfare. Individual scientists, for cen-

turies, have advised the military on specific problems: Archimedes reportedly helped the tyrant of Syracuse in devising new weaponry against the Romans in 212 BC; Leonardo da Vinci supplied us with a variety of drawings of new armaments; and, since the emergence of 'modern' science in the sixteenth and seventeenth centuries, many prominent scientists, including Tartaglia, Galileo, Newton, Descartes, Bernouilli, and Euler, have devoted some of their time and intellect to helping solve military problems.

This article first argues that World War II and the subsequent Cold War produced a dramatic change in the way scientists became involved in the weapons innovation process. Next, it shows that concerns about the resulting 'arms race' brought about a new type of studies—defense technology assessment studies—that dealt with the impact of new weapons systems on national and international security. Many of the newly developed weapons were perceived to have a negative impact, which raised the question of whether and how the weapons innovation process could be influenced. The article discusses a variety of analytical approaches aimed at understanding the dynamics of the weapons innovation process. It argues that a sociotechnical network approach is the most promising one to provide valuable insights for influencing this innovation process. This approach also provides a suitable framework for investigating the relationship between civil and military technological innovation, a subject of growing interest that is discussed in the final section.

1. *Weapons Innovation Becomes Organized*

World War II produced a dramatic change in the way scientists became involved in military matters. Science and scientists in great numbers were mobilized for weapons innovation in a highly organized and concentrated effort. In the USA these scientists contributed, mainly under the auspices of the newly established Office of Scientific Research and Development, to the development of a variety of new technologies, including the atomic bomb, radar, the proximity fuse, and also penicillin. The decisive contribution of scientists to these war efforts implied a fundamental shift in the role of science and technology in future military affairs.

Immediately after the war, US science policy pioneer Vannevar Bush (1945), drawing on the war experience, advised the President:

> [t]here must be more—and more adequate—military research in peacetime. It is essential that the civilian scientists continue in peacetime some portion of those contributions to national security which they have made so effectively during the war.

His advice stood in sharp contrast to Thomas Alva Edison's suggestion, many years before, during World War I, to the Navy, that it should bring into the war effort at least *one* physicist in case it became necessary to 'calculate something' (Gilpin 1962, p. 10).

For the first time in history military research and development (R&D) became a large-scale institutionalized process even in peacetime, indeed on a scale not seen before; it was legitimized as well as fueled by the climate of the Cold War. In the decades following the war, weapons were replaced in a rapid process of 'planned obsolescence.' The R&D was carried out in national laboratories, the defense industry, laboratories of the military services, and at universities to varying degrees in different countries. A United Nations study of 1981 estimated that annually some $100 billion, that is, some 20–25 percent of all R&D expenditures, were devoted to military R&D.

The resulting 'qualitative' arms race in nuclear, conventional, and biological and chemical weapons between the NATO and Warsaw Pact countries during the Cold War raised the question of whether national and international security actually decreased, rather than increased, as a result of 'destabilizing' weapons innovations. A related question was whether military technological developments could be steered or directed so as not to undermine international arms control agreements. After the end of the Cold War in 1990, observers wondered why, particularly in the Western industrialized countries, military R&D efforts continued nearly unabated, while the original threat had disappeared.

The systematic and organized involvement of science and technology in developing new armaments raised more questions of both societal and social-scientific interest. For instance, to what extent have military R&D and defense relations influenced academic research (mainly a concern in the USA) and the course—or even the content—of scientific and technological developments more generally? What is the relationship between (developments in) civil and military technology: are these separate developments or do they, on the contrary, profit from each other through processes of diffusion and spin-off?

Addressing these questions has become an interdisciplinary task, drawing on and integrating insights from many fields. This challenge has been taken up, though still on a limited scale, within the framework of the science and technology (S&T) studies that have emerged since the 1970s.

This article focuses on the origin and nature of 'defense technology assessment studies' and the related research on possibilities for influencing the weapons innovation process.

2. *Defense Technology Assessment Studies*

The declared purpose of weapons innovation is enhancing national security (often broadly interpreted as

including intervention and power projection). However, during the Cold War many argued that the ever-continuing weapons innovation process was actually counter-productive. The issue was not only that the huge amounts spent on armament might be a waste of resources, but also that the huge military R&D effort and the resulting new weapons caused a rapid decrease rather than an increase of national security (e.g., York 1971, p. 228).

Since the 1960s, many studies, often carried out by scientists who had become concerned about the escalating arms race, have dealt with the impact of new weapons systems and new military technologies on national and international security. These studies pointed out, for instance, that the anti-ballistic missile (ABM) systems, consisting of many land-based anti-missile missiles, as proposed by the USA in the 1960s, would actually stimulate the Soviet Union to deploy even more nuclear missiles. Also, a similar ABM system by the Soviet Union would trigger the USA to deploy multi-warhead missiles (MIRVs—Multiple Independently Targetable Re-entry Vehicles). These missiles, carrying up to twelve nuclear warheads, could thus saturate the capabilities of the Soviet ABM interception missiles. Actually, the development and deployment of MIRVed missiles by the USA even preceded a possible Soviet ABM system. As the then US Defense Secretary, Robert McNamara, wrote (quoted in Allison and Morris 1975, p. 118):

> Because the Soviet Union *might* [emphasis in original] deploy extensive ABM defenses, we are making some very important changes in our strategic missile forces. Instead of a single large warhead our missiles are now being designed to carry several small warheads … . Deployment by the Soviets of a ballistic missile defense of their cities will not improve their situation. We have *already* [emphasis added] taken the necessary steps to guarantee that our strategic offensive forces will be able to overcome such a defense.

This weapons innovation 'dynamics' was aptly encapsulated by Jerome Wiesner, former science adviser to President John F. Kennedy, as 'we are in an arms race with ourselves—and we are winning.' In the 1990s, when the USA continued its efforts to develop anti-satellite (ASAT) technology capable of destroying an adversary's satellites, Wiesner's words might rightly have been paraphrased as 'we are in an arms race with ourselves—and we are losing.' For the irony here is that it is the US military system that, more than any other country's defense system, is dependent on satellites (for communication, reconnaissance, eavesdropping, and so on), which would be highly vulnerable to a hostile ASAT system. The most likely route, however, for hostile countries to obtain the advanced ASAT technology would not be through their own R&D, but through the *proliferation*, that is, diffusion of US ASAT technology, once it had been developed. Again, the USA would very likely decrease rather than increase its own security by developing ASAT technology.

Many of the early and later 'impact assessments' of weapons innovations assessed their potential of circumventing and undermining existing international agreements that aimed to halt the arms race, like the Anti-Ballistic Missile (ABM) Treaty (1972) and the accompanying Strategic Arms Limitation Agreements (SALT, 1972), the SALT II Treaty (1979), and a Comprehensive Test Ban Treaty (CTBT, 1996). In addition, assessments were made, both by independent scientists and governmental agencies, of the potential of civil technologies to spill over into military applications: for in such cases, countries could, under the guise of developing civil technologies make all preparations needed for developing nuclear, chemical, or biological weapons.

Through these 'dual-use' technologies, a proliferation of weapons could occur, or at least the threshold lowered for obtaining those weapons, without formally violating the Non Proliferation Treaty (1971), the Comprehensive Test Ban Treaty (1996), the Biological Weapons Convention (1972), or the Chemical Weapons Convention (1993). The *Arms Control* readings (York 1973) from the Magazine *Scientific American* provide an instructive sample of early defense technology assessments.

In the 1980s, both the USA and NATO emphasized the importance of a stronger conventional defense, which was then believed to be feasible because of new 'emerging technologies,' including sensor and guidance technologies, C^3I (Command, Control, Communications and Intelligence) technologies (like real time data processing), electronic warfare, and a variety of new missiles and munitions. The emphasis on high technology in the conventional weapons area, in its turn, triggered a great variety of studies (pro and con) by academic and other defense analysts, not only in the USA, but also in Europe. These included analyses of the technical feasibility of the proposed systems, the affordability of acquiring sufficient numbers of ever more costly weapons, and the associated consequences for national defense and international security.

The results of these defense technology assessments were fed into the public discussion, with the aim of containing the arms race and reaching international arms control agreements, or preventing their erosion.

The impact of these assessments on the weapons innovation process and its accompanying military R&D was often limited. Many, therefore, considered the weapons innovation process to be 'out of control,' providing another topic of S&T studies on military technology.

3. *Influencing the Weapons Innovation Process*

What does it mean to 'bring weapons innovation under political control'? The concept seems obvious at

one level but it is actually not trivial and needs elaboration.

In national politics there is often no consensus on the kinds of armament that are desirable or necessary. Those who say that politics is not in control may actually mean that developments are not in accordance with their political preferences, whereas those who are quite content with current developments may be inclined to say that politics is in control. Neither position is analytically satisfactory. But neither is it satisfactory to say that politics is in control simply because actual weapons innovations are the outcome of the political process, which includes lobbying of defense contractors, interservice rivalry, bureaucratic politics, arguments over ideology and strategic concepts, and so forth, as Greenwood (1990) has suggested. Rather than speaking of control, one should ask whether it would be possible to influence the innovation process in a systematic way, or to steer it according to some guiding principle (Smit 1989). This implies that the basic issue of 'control,' even for those who are content with current developments, concerns whether it would be possible to change their course if this was desired.

A plethora of studies have appeared on what has been called the technological arms race (see e.g. Gleditsch and Njolstad 1990). Many of them deal with what President Eisenhower in his much-cited farewell address called the military-industrial complex, later extended to include the bureaucracy as well. These studies belong to what has been called the bureaucratic-politics school or domestic structure model (Buzan 1987, Chap. 7), in contrast to the action-reaction models (Buzan 1987, Chap. 6), which focus on interstate interactions as an explanation of the dynamics of the arms race. A third approach, the technological imperative model (Buzan 1987, Chap. 8) sees technological change as an independent factor in the arms race, causing an unavoidable advance in military technology, if only for its links with civil technological progress—though a link whose importance is under debate (see also the last section of this article). By contrast, Ellis (1987), in his social history of the machine gun, has shown the intricate interweaving of weapons innovation with social, military, cultural, and political factors. To some extent, these studies might be considered complementary, focusing on different elements in a complex pattern of weapons development and procurement. For instance, the 'reaction' behavior in the interstate model might be translated into the 'legitimation' process of domestically driven weapons developments.

Many of these studies are of a descriptive nature. Some (Allison and Morris 1975, Brooks 1975, Kaldor 1983) are more analytical and try to identify important determinants in the arms race. These factors are predominantly *internal* to each nation and partly linked up with the lengthy process—10 to 15 years—of the development of new weapons systems. Other studies (Rosen 1991, Demchak 1991) relate military innovation, including military technology, to institutional and organizational military factors and to the role of civilians. There are quite a number of case studies on the development of specific weapons systems, and empirical studies on the structure of defense industries (Ball and Leitenberg 1983, Kolodziej 1987, Gansler 1980, 1987) or arms procurement processes (Cowen 1986, Long and Reppy 1980).

However, hardly any of these studies focus on the question of how the (direction) of the weapons innovation process and the course of military R&D might be influenced. One task for future S&T studies, therefore, would be to combine insights from this great variety of studies for a better understanding of these innovation processes. Some steps have already been taken. The lengthy road of developing new weapons systems implies that it will be hard to halt or even redirect a system at the end when much investment has been made. Influencing weapons innovations, therefore, implies a continuous process of assessment, evaluation, and (re-)directing, starting at the early stages of the R&D process (see also *Technology Assessment*). Just striving for 'technological superiority,' one of the traditional guiding principles in weapons development, will lead to what is seemingly an autonomous process. Seemingly, because technology development is never truly autonomous. The appearance of autonomy results from the fact that many actors (i.e., organizations) are involved in developing technology, and no single actor on its own is able to steer the entire development. Rather, all actors involved are connected within a network—we may call it a sociotechnical network—working together and realizing collectively a certain direction in the weapons innovation process.

Network approaches, appearing in several areas of science and technology studies since the mid 1980s, in which the positions, views, interests, and cultures of the actors involved are analyzed, as well as their mutual links and the institutional settings in which the actors operate, open up an interesting road for dealing with the question of influencing military technological developments (Elzen et al. 1996). Such approaches emphasize the interdependencies between the actors and focus on the nature of their mutual interactions (see also *Actor Network Theory*).

From a network approach it is evident why one single organization by itself cannot determine technological developments. At the same time, network approaches have the power to analyze the way these developments may be *influenced* by actors in a network. Networks both enable and constrain the possibilities of influencing technology developments. Analyzing them may provide clues as to how to influence them.

Weapons innovation and its associated military R&D differ in one respect from nearly all other technologies, in that there is virtually only one

customer of the end product—that is, the state. (Some civil industries, like nuclear power and telecommunications in the past, also show considerable similarities in market structure—monopolies or oligopolies coupled with one, or at most a few dominant purchasers. They are also highly regulated, and markedly different from the competitive consumer goods sectors (see Gummett 1990). Moreover, only a specific set of actors comprises the sociotechnical networks of military technological developments, including the defense industry, the military, the defense ministry, and the government. The defense ministry, as the sole buyer on the monopsonistic armament market, has a crucial position. In addition, the defense ministry is heavily involved in the whole R&D process by providing, or refunding to industry, much of the necessary funds. Yet the defense ministry, in its turn, is dependent on the other actors, like the defense industry and military laboratories, which provide the technological options from which the defense ministry may choose. S&T studies of this interlocked behavior offer a promising approach for making progress on the issue of regulatory regimes. Not steering from a central position, but adopting instead an approach of 'decentralized regulation' in which most actors participate then seems the viable option for influencing military technological developments. In this connection various 'guiding principles' could play a role in a regulatory regime for military technological innovation (see also Enserink et al. 1992). Such guiding principles could include the 'proportionality principle,' 'humanitarian principles,' and 'limiting weapons effects to the duration of conflict' (contrary, for instance, to the use of current anti-personnel mines).

Additional phenomena that should in any event be taken into account in such network approaches are the increasing international cooperation and amalgamation of the defense industry, the possibly increasing integration of civil and military technology, and the constraining role of international arms control agreements. The intricate relation between civil and military technology will briefly be discussed in the final section.

4. *Integration of Civil and Military Technology*

Sociotechnical network approaches seem particularly useful for studying the relation between civil and military technology. This issue has assumed increasing interest because of the desire to integrate civil and military technological developments. Technologies that have both civil and military applications are called dual-use technologies. The desire for integration originates (a) from the need for lower priced defense products because of reductions in procurement budgets, and (b) from the new situation that in a number of technological sectors innovation in the civil sector has outstripped that in the military sector. Examples of such sectors are computers and information and communication technology, where such integration already emerges. Such integration, of course, could be at odds with a policy of preventing weapons proliferation as discussed before.

Research on the transformations needed to apply civil technologies in the military sector and vice versa have only just begun. The extent to which civil and military technologies diverge depends not only on the diverging needs and requirements, but also on the different institutional, organizational, and cultural contexts in which these developments occur.

Several case studies illustrate how intricately interwoven the characteristics of technology may be with the social context in which it is being developed or in which it functions. MacKenzie (1990) conducted a very detailed study of the development of missile guidance technologies in relation to their social context, and the technological choices made for improving accuracy, not only for missiles but also for aircraft navigation. He showed how different emphases in requirements for missile accuracy and for civil (and military) air navigation resulted in alternative forms of technological change: the former focusing on accuracy, the latter on reliability, producibility, and economy.

A number of historical studies have addressed the question of the *relation* between civil and military technology. Some of them have shown that in a number of cases in the past, the military successfully guided technological developments in specific directions that also penetrated the civil sector. Smith (1985), for instance, showed that the manufacturing methods based on 'uniformity' and 'standardization' that emerged in the USA in the nineteenth century, were more or less imposed (though not without difficulties and setbacks) by the Army Ordnance Department's wish for interchangeable parts. Noble (1985) investigated, from a more normative perspective, three technical changes in which the military have played a crucial role—namely, interchangeable parts manufacture, containerization, and numerical control—arguing that different or additional developments would have been preferable from a different value system. Studies going back to the nineteenth or early twentieth centuries, though interesting from a historical perspective, may not always be relevant for modern R&D and current technological innovation processes. Systematic studies of the interrelation between *current* military and civil technological developments are only of recent date (see Gummett and Reppy 1988). They point out that it may be useful to distinguish not only between different levels of technology, such as generic technologies and materials, components, and systems (Walker et al. 1988), but also between products and manufacturing or processing technologies (Alic et al. 1992). In certain technological areas the distinguishing features between civil and military technology may be found at the level

of system integration, rather than at the component level. In conclusion one may say that military technology, for many reasons, is a fascinating field for future science and technology studies: its links with a broad scope of societal institutions, its vital role in international security issues, the need for influencing its development, and its increasing integration with civil technology, particularly in the sector of information and communication technologies, which may revolutionize military affairs.

See also: Innovation: Organizational; Innovation, Theory of; Military and Disaster Psychiatry; Military Geography; Military History; Military Psychology: United States; Military Sociology; Research and Development in Organizations; Science and Technology, Social Study of: Computers and Information Technology

Bibliography

Alic J A, Branscomb L M, Brooks H, Carter A B, Epstein G L 1992 *Beyond Spinoff: Military and Commercial Technologies in a Changing World*. Harvard Business School Press, Boston

Allison G T, Morris F A 1975 Armaments and arms control: Exploring the determinants of military weapons. *Daedalus* **104**(3): 99–129

Ball N, Leitenberg M 1983 *The Structure of the Defense Industry*. Croom Helm, London and Canberra

Brooks H 1975 The military innovation system and the qualitative arms race. *Daedalus* **104**(3): 75–97

Buzan B 1987 *An Introduction to Strategic Studies: Military Technology and International Relations*. Macmillan, Basingstoke, UK

Cowen R 1986 *Defense Procurement in the Federal Republic of Germany: Politics and Organization*. Westview Press, Boulder, CO

Demchak C C 1991 *Military Organizations, Complex Machines. Modernization in the US Armed Services*. Cornell University Press, Ithaca, NY

Ellis J 1987 *The Social History of the Machine Gun*. The Cresset Library, London. (Reprint of 1975 ed.).

Elzen B, Enserink B, Smit W A 1996 Socio-technical networks: How a technology studies approach may help to solve problems related to technical change. *Social Studies of Science* **26**(1): 95–141

Enserink B, Smit W A, Elzen B 1992 Directing a cacophony—weapon innovation and international security. In: Smit W A, Grin J, Voronkov L (eds.) *Military Technological Innovation and Stability in a Changing World. Politically Assessing and Influencing Weapon Innovation and Military Research and Development*. VU University Press, Amsterdam, The Netherlands, pp. 95–123

Gansler J S 1980 *The Defense Industry*. MIT Press, Cambridge, MA

Gansler J S 1989 *Affording Defense*. MIT Press, Cambridge, MA

Gilpin R 1962 *American Scientists and Nuclear Weapons Policy*. Princeton University Press, Princeton, NJ

Gleditsch N P, Njolstad O (eds.) 1990 *Arms Races. Technological and Political Dynamics*. Sage, London

Greenwood T 1990 Why military technology is difficult to constrain. *Science Technology and Human Values* **15**(4): 412–29

Gummett P H 1990 Issues for STS raised by defense science and technology policy. *Social Studies of Science* **20**: 541–58

Gummett P H, Reppy J (eds.) 1988 *The Relations Between Defense and Civil Technologies*. Kluwer, Dordrecht, The Netherlands

Hacker B C 1994 Military institutions, weapons, and social change: Towards a new history of military technology. *Technology and Culture* **35**(4): 768–834

Kaldor M 1983 *The Baroque Arsenal*. Sphere, London

Kolodziej E A 1987 *Making and Marketing Arms: The French Experience and its Implications for the International System*. Princeton University Press, Princeton, New Jersey

Long F A, Reppy J (eds.) 1980 *The Genesis of New Weapons*. Pergamon, New York

MacKenzie D 1990 *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. MIT Press, Cambridge, MA

Mendelsohn E, Smith M R, Weingart P (eds.) 1988 *Science, Technology and the Military*. Sociology of the Sciences Yearbook, Vol. XII/1. Kluwer, Boston

Molas-Gallart J 1997 Which way to go? Defense technology and the diversity of 'dual-use' technology transfer. *Research Policy* **26**: 367–85

Noble D F 1985 Command performance: A perspective on the social and economic consequences of military enterprise. In: Smith M R (ed.) *Military Enterprise and Technological Change—Perspectives on the American Experience*. MIT Press, Cambridge, MA, pp. 329–46

Rosen S P 1991 *Winning the Next War. Innovation and the Modern Military*. Cornell University Press, Ithaca

Sapolsky H M 1977 Science, technology and military policy. In: Spiegel-Rösing I M, de Solla Price D (eds.) *Science, Technology and Society. A Cross-disciplinary Perspective*. Sage, London, pp. 443–72

Smith M R 1985 Army ordnance and the "American system" of manufacturing, 1815–1861. In: Smith M R (ed.) *Military Enterprise and Technological Change—Perspectives on the American Experience*. MIT Press, Cambridge, MA, pp. 39–86

Smit W A 1989 Defense technology assessment and the control of emerging technologies. In: Borg M ter, Smit W A (eds.) *Non-provocative Defence as a Principle of Arms Reduction*. Free University Press, Amsterdam, The Netherlands, pp. 61–76

Smit W A 1995 Science, technology, and the military: relations in transition. In: Jasanoff S, Markle G E, Peterson J C, Pirich T (eds.) *Handbook of Science and Technology Studies*. Sage, London, pp. 598–626

Smit W A, Grin J, Voronkov L (eds.) 1992 *Military Technological Innovation and Stability in a Changing World. Politically Assessing and Influencing Weapon Innovation and Military Research and Development*. VU University Press, Amsterdam

The Annals of the American Academy of Political and Social Science 1989 Special Issue: Universities and the Military. **502** (March)

Walker W, Graham M, Harbor B 1988 From components to integrated systems: Technological diversity and integration between military and civilian sectors. In: Gummett Ph, Reppy J (eds.) *The Relations Between Defence and Civil Technologies*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 17–37

York H F 1971 *Race to Oblivion: A Participants View of the Arms Race*. Simon and Schuster, New York
York H F (comp.) 1973 *Arms Control*. Readings from Scientific American. Freeman, San Francisco

W. A. Smit

Scientific Academies, History of

Scientific academies are associations of scientific practitioners like the *Academie Royale des Sciences* of Paris, the Royal Society of London, the Berlin Academy, the Russian Academy of Science, or the US National Academy of Science. Since their inception in seventeenth-century Europe, the distinctive feature of scientific academies has been the regulation of their membership and activities according to corporate protocols, statutes, and bylaws (which, nevertheless, may be developed in accordance with the policies of the state authorities that sometimes sponsor them). Typical activities of scientific academies include: the publication of journals, monographs, and the collected work of famous scientists, the award of prizes and medals (the Nobel Prize being the most famous example), the organization of scientific meetings and, more rarely, of research projects (sometimes in collaboration with other academies). They may also take an advising role in matters of science policy, and may be called upon to advise governments on scientific and technological projects.

The sociological and historical significance of scientific academies is tied to their crucial role in the development of science as a peer-based system, in constituting the 'scientist' as a distinct socio-professional role, in establishing networks of communication and collaboration, and, more generally, in fostering the so-called internationalism of science (Crawford 1992). Historically, scientific academies have functioned as institutions demarcating the elite of scientific practitioners from other socioprofessional groups (which may have included other scientific practitioners pursuing different methodological programs or more practical subjects). By doing so, they have created the conditions of possibility for corporate authority in science (though, historically, different academies have taken different stances with respect to their role as judges of scientific and technical claims). They have also constituted themselves as crucial 'lobbying' entities in negotiations between corporate science and the state.

But while academies are seen as a cornerstone in the development of science as a peer-based and peer-regulated system, the very notion of peer has undergone important changes since the seventeenth century and remains contested even today. There is a circular relationship between the definition of peer and academy. Academies are constituted by peers, but they also constitute what peer means. A peer is not only someone who has specific technical competencies in a given scientific field, but also a person who, through a process of training and professional socialization, has developed (and has been given the opportunity to develop) a specific corporate sociability and set of values. The ingredients of such a sociability are rooted in the historical, disciplinary, and socio-political context in which a given academy developed, but there are some features that have cut across these various contexts. One of them is gender. A historically conspicuous feature of scientific academies is their distinctly male corporate culture which, until the beginning of the twentieth century, led to the exclusion of women from their membership—an exclusion that was less marked in other less corporate sites of science (Schiebinger 1989).

Another theme that runs through the literature on academies is their relative 'independence.' Some have seen academies as providing an institutional boundary between science and society, thereby shielding the allegedly ever-threatened 'purity' of science (Ben-David 1971). Others, instead, have taken the empirically more defensible view that academies are institutions constituted within specific socio-political ecologies—ecologies to whose development they contribute from within. The former view sees science as in need of being or becoming as independent as possible from the social nomos, while the latter treats it as a set of practices that have articulated increasingly finer and more pervasive relationships with the socio-political order precisely by developing more specialized and professionalized institutional forms. In one case, institutionalization means separation; in the other it means integration within an increasingly articulated interplay of social institutions.

These considerations bear directly on the problem of defining 'scientific academy.' It is easy to compile a long and inclusive list of past or present academies, but it would be much harder, perhaps impossible, to categorize them as a long-standing institutional type like, say, the university. Depending on historical and national contexts, 'scientific academy' refers to different types of associations with different structures, functions, methodological orientations, membership requirements, notions of intellectual property and authorship, funding arrangements, and affiliations with private patrons or state governments and bureaucracies. Since their inception in seventeenth-century Europe, scientific academies have dealt with the production, reward, and dissemination of scientific knowledge, the training of practitioners, the evaluation of patent applications, and the advising of state and political authorities on scientific and techical matters. Some academies have been established by practitioners themselves and tend to be independently funded. Others are state-sponsored and tend to function as a body of scientific experts for the state that

supports them. Certain academies have a distinctly international membership, while others draw it from a specific country (though they may also extend limited forms of membership to a few foreign practitioners). Some select their members from one specific discipline, others gather practitioners from different fields. Some promote and fund the production of new knowledge, others concerns themselves with its dissemination. Most modern academies do not offer stipends to researchers (and may even require their members to contribute annual fees), while some earlier academies (Paris, Berlin, and St Petersburg) did provide their members with salaries and research funds (McClellan 1985).

Because scientific academies have assumed most of the functions now typical of other institutions, it is possible to find partial analogies between academies and these other components of the social system of science. For instance, it may be difficult in some cases to draw a sharp demarcation between scientific academies and other institutions such as discipline-based professional associations, national laboratories and research institutes, foundations (like Rockefeller in the US), national agencies for the development of science (CNRS in France, NSF or NIH in the US, CNR in Italy, etc.), or even university science departments.

Similar problems emerge when we try to distinguish scientific academies from other less scientific-sounding associations which, nevertheless, promote or promoted the production of knowledge about nature. In this category we find early modern philosophical salons (usually established and run by women (Lougee 1976)), religious orders that supported collective work in mathematics, astronomy, and natural history (like the Society of Jesus), provincial academies of natural philosophy, belles lettres, and history (common in Europe and America throughout the nineteenth century), nonprofessional associations for the popularization of science (which may be seen as the modern heirs to earlier provincial academies), or contemporary associations of amateur scientists. The same taxonomical difficulties emerge when one looks for some essential demarcation between academies and collection-based institutions like the Smithsonian Institution in Washington and other museums of science and natural history, and botanical gardens. In the premodern period, such institutions tended to be associated with academies and, as shown by the American Museum of Natural History in New York, may still perform some of the research functions of their historical ancestors. The same may be said about astronomical observatories. Many of them (Greenwich, Paris, Berlin, St Petersburg, etc.) were developed within scientific academies or soon became annexed to them, but later evolved into independent institutions or became linked to other state agencies (like the *Bureau de Longitude* in nineteenth-century France).

In sum, one either runs the risk of treating 'academy' as a synonym for 'scientific institution' (in the same way some apply the term 'laboratory' to anything from large-scale international organizations like CERN and private industrial entities like Bell Labs, to early modern alchemical workshops or to Uraniborg—Tycho Brahe's sixteenth-century aristocratic observatory-in-a-castle (Christianson 2000)) or, in a reverse move, of identifying academies solely with institutions bearing that name (but likely to display remarkably different structures and functions), while leaving out a slew of other associations whose activities may in fact overlap with those pursued, at one point or another, by 'proper' academies.

A possible solution to this definitional puzzle is to take a genealogical (rather than taxonomical) approach. In any given historical and national context there is a division of labor and functions between scientific academies and the surrounding institutions of science and of the state bureaucracy. It is this institutional ecology that we need to look at to understand which role academy fulfills in any given context.

If we take peer-based management to be the distinctive trait of academies, then a reasonable starting point of their genealogy is the development of legal statutes and bylaws; that is, of instruments for the establishment of their corporate form of life. This makes the Royal Society of London (chartered in 1662) the first scientific academy. This criterion excludes a number of well-known seventeenth-century academies (such as the Lincei (1603–1631), the Cimento (1657–1667), and even the early *Academie Royale des Sciences* (which began to gather around 1666 but obtained its first statute only in 1699).

Academies that did not have legal statutes or charters gathered at the will of a prince who sponsored them, and operated within the rules and system of exchange typical of the patronage system. They can be seen as the extension of the model of informal court gatherings and aristocratic literary salons to the domain of natural philosophy. They were prince-dependent not only in the sense that they were funded by princes, but their work, publications, membership, and sociability were structured by their direct, quasi-personal relationship to the patrons who, in several cases, participated in them. In these contexts, academicians were more princely subjects than collaborating scientific peers (Biagioli 1996).

The next step—the development of charters, statutes, and bylaws—marked the transition from the patronage framework to institutions which, while still connected to royal power, had a more formal and bureaucratic relationship to the state. This institutional development followed two different models, one exemplified by the Royal Society of London, the other by the *Academie Royale des Sciences*. These two models matched quite closely the structure and discourse of royal power in those two countries: absolutism in France and constitutional monarchy in England.

In fact, despite its name, the Royal Society was a private institution (some may say a large gentlemen's club) that, while chartered by the king, received negligible funding and virtually no supervision from the crown (Hunter 1989). Its members were nominated and approved by the Society itself. Similar protocols were followed for the selection of the president and other officials. There were no paid positions except that of the secretary and the curator of experiments, and the operating budget came mostly from annual membership fees which were typically hard to collect. The training of young practitioners was not one of the Society's stated goals.

The *Academie des Sciences*, instead, was much more of a state agency. The French crown funded it, provided stipends, and closely controlled its projects, membership, and publications. Its members were grouped according to disciplinary taxonomies, and organized hierarchically according to seniority (with students at the bottom) and, sometimes, social status. The king selected crown bureaucrats to run its operation (Hahn 1971).

These two models and their various permutations were adopted during the remarkable development (almost an explosion) of scientific academies in the eighteenth century. Driven by a mix of national pride, emulation, and the state's growing appetite for technical expertise, virtually all European states, from England to Russia, developed such academies (McClellan 1985). Academies proliferated not only across but within nations, moving from the capitals to the periphery (Roche 1978). Reflecting the logic of political absolutism, state academies tended to follow the French paradigm, while provincial and private academies, having little or no connections to central state power, looked to the Royal Society as their primary model. The center–periphery distinction maps quite well on the disciplinary scope of early scientific academies: provincial institutions often included literature and history sections, while centrally located academies tended to specialize in natural philosophy.

Until the French Revolution, academies constituted the primary site of scientific activity, epitomized the very notion of scientific institution, and provided science with a kind of social and professional legitimation it had struggled to achieve in previous periods. Some military schools (especially in France) shared the academies' concern with research. But with the exception of Bologna, Pavia, and Gottingen, early modern universities usually limited themselves to the teaching of (not the research in) natural philosophy.

If the eighteenth century was the golden age of scientific academies, the nineteenth century saw the beginning of their decline, or at least their slow but steady reframing from research institutions into prestige-bearing ones. New kinds of institutions emerged and older ones were reformed as the social system of science developed in scale and specialization. The remarkable complexity of this scenario is such that only a few general trends can be discussed here, and only in reference to their impact on the role of academies.

First, starting with Germany, the restructuring of the university led to an increased representation of the sciences in its curriculum and, more importantly, to a slow but steady trend toward research, not only teaching, in the natural sciences. In the second half of the nineteenth century, this trend was followed by other countries including the US, where new universities (first Johns Hopkins and then the so-called land-grant state universities) were developed with an eye on European research models—models that were slowly adopted by Ivy League universities as well. This trend eroded much of the academies' pedagogical and research role.

Second, with the slow but progressive rise of science-based industry, the private sector (often in collaboration with universities) provided more niches for research and for the employment for fast-growing cadres of scientists. Widespread industrial development eroded another aspect of the academies' traditional function: that of judge of technological innovation. The emergence of the patent system, first in England but, after 1800, in most other European countries and the US, placed the reward and protection of technological innovation in the domain of the law and of the market and disconnected it from its authoritative assessment by academies (as had been the case in early modern France and other countries).

Third, the establishment or growing role of state or military technical agencies and schools (like, say, the Corps of Engineers and Geological Survey in the US, the *Ecole Politechnique* and the *Ecole des Ponts et Chaussees* in France) greatly reduced the role of scientific academies as providers of technical and scientific expertise to the state (Shinn 1980).

Fourth, with the branching of science in increasingly numerous disciplines, a slew of professional associations emerged and, with them, increasingly specialized journals. Not only did these discipline-based associations erode the academies' quasi-monopoly on scientific publishing and translations, but they also took up much of their function as international nodes of communication, collaboration, and standardization of methodologies, terminology, and units of measurement. The 'internationalism' of science which largely had been connected not only to networks of communication among academies but also to the export and import of academicians across national boundaries (most conspicuously in the eighteenth century) was replaced by more specific disciplinary and professional networks and by the migration of students (and later post doctorates) between universities and, later, institutes and laboratories. Similarly, the academies' role in the diffusion of scientific knowledge was eroded by these trends, as well as by the development of a wider nonprofessional audience for science and of

the popular publications and magazines that catered to them.

All these trends continued and were further accelerated in the twentieth century, when the additional creation of national agencies for the promotion and funding of science and the growing role of private foundations in the patronage of science took on yet another role that had been traditionally played by academies. Even large-scale, state-sponsored technical and military projects (which, historically, had seen the direct or indirect participation of scientific academies) have become commonly structured around collaborations between the state, the university, and industry. International scientific collaborations are also primarily managed by universities and national science agencies and, in the case of science-based industry, through joint ventures, licensing, patent sharing, or mergers and acquisitions. (These trends, however, are typical of Western contexts. In the former Soviet Union, the Soviet Academy of Science played a dominant role in research, training, and rewards at a time when its western counterparts had largely shed those functions (Vucinich 1984).)

In sum, it is possible to identify at least three phases in the genealogy and role of academies within the broader ecology of the social system of science: (a) patronage-based princely academies without statutes and corporate protocols up to the end of the seventeenth century; (b) peer-based academies with statutes, publications, and, in many cases, formalized relations with the state since the late seventeenth century, and mostly during the eighteenth century; and (c) dramatic increase of the scale and complexity of the social system of science associated with the multiplication of more specialized scientific institutions taking up most of the traditional roles of scientific academies (since the early nineteenth century).

The net result of this historical trajectory is that academies have moved from being informal gatherings structured by the curiosity and philosophical interests of princes and gentlemen, to comprehensive and multifunctional corporate entities crucial to the production and legitimation of science and its social system, to prestigious institutions that are marginally involved in research but still maintain an important advising role to the state and other constituencies, and bestow professional recognition on leading scientists through membership and awards. In many ways, the changing role of prizes awarded by academies encapsulates this trajectory: In the early modern period, prizes were seen as means to promote new research (not unlike today's grants) but now they function as rewards of important work that scientists have done in the past.

The importance of the largely symbolic role of modern scientific academies, however, should not be underestimated. In an age when 'lobbying' has become the dominant mode of corporate participation in political decisions (especially in the US), academies remain the authoritative agents of corporate science, no matter how accurately they may be able to represent the interests and concerns of such a large and diversified community.

See also: Centers for Advanced Study: International/Interdisciplinary; History of Science; History of Science: Constructivist Perspectives; Scientific Disciplines, History of; Universities, in the History of the Social Sciences

Bibliography

Ben-David J 1971 *The Scientist's Role in Society: A Comparative Study*. Prentice-Hall, Englewood Cliffs, NJ

Biagioli M 1996 Etiquette, interdependence, and sociability in seventeenth-century science. *Critical Inquiry* **22**: 193–238

Boehm L, Raimondi E 1981 *Universita' Accademie e Societa' Scientifiche in Italia e in Germania dal Cinquecento al Settecento*. Il Mulino, Bologna, Italy

Cahan D 1984 The institutional revolution in German physics, 1865–1914. *Historical Studies in the Physical and Biological Sciences* **15**: 1–65

Cavazza M 1990 *Settecento inquieto: Alle origini dell'Istituto delle Scienze di Bologna*. Mulino, Bologna, Italy

Christianson J R 2000 *On Tycho's Island*. Cambridge University Press, Cambridge, UK

Crawford E 1992 *Nationalism and Internationalism in Science, 1880–1930: Four Studies of the Nobel Population*. Cambridge University Press, Cambridge, UK

Crosland M 1992 *Science Under Control: The French Academy of Sciences, 1795–1914*. Cambridge University Press, Cambridge, UK

Daniels G 1976 The process of professionalization of American science: The emergent period, 1820–1860. In: Rheingold N (ed.) *Science in America Since 1820*. Science History Publications, Canton, MA, pp. 63–78

Fox R, Weisz G 1980 *The Organization of Science and Technology in France, 1808–1914*. Cambridge University Press, Cambridge, UK

Frangsmyr T (ed.) 1989 *Science in Sweden: The Royal Swedish Academy of Science, 1739–1989*. Science History Publications, Canton, MA

Frangsmyr T (ed.) 1990 *Solomon's House Revisited*. Science History Publications, Canton, MA

Graham L 1967 *The Soviet Academy of Sciences and the Communist Party, 1927–1932*. Princeton University Press, Princeton, NJ

Hahn R 1971 *The Anatomy of a Scientific Institution*. University of California Press, Berkeley, CA

Heilbron J L 1983 *Physics at the Royal Society During Newton's Presidency*. William Clark Memorial Library, Los Angeles

Hunter M 1989 *Establishing the New Science*. Boydell, Woodbridge, MA

Jasanoff S 1990 *The Fifth Branch: Science Advisers as Policymakers*. Harvard University Press, Cambridge, MA

Kevles D 1978 *The Physicists: The History of a Scientific Community in Modern America*. Knopf, New York

Knowles Middleton W E 1971 *The Experimenters: A Study of the Academia del Cimento*. Johns Hopkins University Press, Baltimore, MD
Kohlstedt S 1976 *The Formation of the American Scientific Community*. University of Illinois Press, Urbana, IL
Lougee C 1976 *Le Paradis des Femmes: Women, Salons, and Social Stratification in Seventeenth-century France*. Princeton University Press, Princeton, NJ
MacLeod R, Collins P (eds.) 1981 *The Parliament of Science: The British Association for the Advancement of Science, 1831–1981*. Kluwer, Northwood, UK
McClellan J E 1985 *Science Reorganized*. Columbia University Press, New York
Moran B 1992 *Patronage and Institutions*. Boydell, Woodbridge
Morrell J, Thackray A 1981 *Gentlemen of Science: Early Years of the British Association for the Advancement of Science*. Clarendon Press, New York
Oleson A, Brown S C (eds.) *The Pursuit of Knowledge in the Early American Republic: American Scientific and Learned Societies from Colonial Times to the Civil War*. Johns Hopkins University Press, Baltimore, MD
Pyeson L, Sheets-Pyeson S 1999 *Servants of Nature*. Norton, New York
Roche D 1978 *Le Siecle de Lumieres en Province*. Mouton, Paris
Schiebinger L 1989 *The Mind Has No Sex?* Harvard University Press, Cambridge, MA
Shinn T 1980 *Savoir scientifique et pouvoir social: L'Ecole Polytechnique, 1794–1914*. Presses de la Fondation Nationale des Sciences Politiques, Paris
Vucinich A 1984 *Empire of Knowledge: The Academy of Sciences of the USSR, 1917–1970*. University of California Press, Berkeley, CA

M. Biagioli

Scientific Academies in Asia

The criterion for selection of the institutions described in this article is that academic research is their main function. Hence, academies that focus mainly on teaching, and policy-oriented research institutes, are not included. Taiwan's Academia Sinica is examined in terms of organization and research accomplishment, as well as its role in society. Similar research academies in mainland China and in Japan are also briefly introduced.

1. The Academia Sinica, Taiwan

1.1 Organization

The Academia Sinica was founded in mainland China 1928 in with two major missions: to undertake research in science and the humanities and to instigate, coordinate and encourage academic research. In 1949, after the Chinese civil war, the Academia was moved to Taiwan and re-established at the present site in 1954. The Academia has been headed by seven presidents since its founding, and the current president, Dr. Lee Yuan-Tseh, a Nobel laureate in chemistry, took office in 1994.

The upper-level organization of the Academia Sinica comprises three parts: the convocation, the council, and the central advisory committee. The convocation is a biennial meeting attended by preeminent Chinese scholars from all over the world who are elected to be academicians of the Academia Sinica. The year 2000 saw the 26th convocation and most of the 193 academicians—an honorary lifetime title—gathered to elect new academicians and council members. They also proposed research policies for the Academia Sinica. The council consists of 18 *ex officio* members, including all directors of research institutes, and 35 elected members. Specific functions of the council include review of research projects, evaluation of proposals related to institutional changes, promotion of academic cooperation within and outside Taiwan and, perhaps most importantly, to present a shortlist of three candidates for the presidency of the Academia Sinica to the President of the Republic of China, who make the final decision. The central advisory committee, established in 1991, includes chairpersons of the advisory committees of individual institutes and three to nine specialists nominated by the president of the Academia Sinica. Their tasks are to recruit scholars of various disciplines as well as to suggest long-term and interdisciplinary research plans to the president. The committee is also responsible for evaluating large-scale cross-institutional research projects, applications for postdoctoral research posts, and the annual awards for junior researchers' publications.

But the core of the Academia Sinica is made up of the 25 institutes/preparatory offices classified into three divisions: Mathematics and Physical Sciences, Life Sciences, and Humanities and Social Sciences (see Table 1). In 2000, 815 research staff (including 13 distinguished research fellows, 324 research fellows, 197 associate research fellows, 138 assistant research fellows, and 143 research assistants) conducted active research either individually or in groups within, as well as across, institutions. In addition to the research staff, postdoctoral researchers, contracted research assistants, and administrative officers add up to approximately 3,000 people working in the Academia Sinica.

1.2 Research Focus and Achievements

There are six fundamental principles or basic goals guiding academic research in Academia Sinica (Lee 1999) Examples of research accomplishments, landmark research projects and significant publications are discussed below.

1.2.1 A balance between scientific and humanitarian research. Critics have condemned the dominance of

Table 1
Number of research staff at Academia Sinica, Taiwan: 1994–2000

Researchnstitutes	1994	1995	1996	1997	1998	1999	2000
Division of Mathematics & Physical Sciences							
Institute of Mathematics	30	33	32	34	32	31	31
Institute of Physics	41	41	39	39	42	43	42
Institute of Chemistry	29	28	27	27	28	30	27
Institute of Earth Sciences	37	36	35	34	37	37	34
Institute of Information Sciences	28	29	30	30	32	34	36
Institute of Statistical Sciences	29	31	31	32	30	32	32
Institute of Atomic and Molecular Sciences	21	22	22	22	22	24	24
Institute of Astronomy and Astrophysics[a]	1	2	4	7	10	10	13
Institute of Applied Science & Engineering[a]							2
Subtotal	216	222	220	225	233	241	241
No. increase		6	−2	5	8	8	0
Rate of increase (percent)		2.78	−0.90	2.27	3.56	3.43	0.00
Division of Life Sciences							
Institute of Botany	41	41	41	39	38	39	40
Institute of Zoology	28	30	30	29	30	29	27
Institute of Biological Chemistry	30	31	33	34	35	34	36
Institute of Molecular Biology	80	84	82	79	82	85	87
Institute of Biomedical Sciences	51	48	49	54	54	55	54
Institute of BioAgricultural Sciences[a]						8	11
Subtotal	230	234	235	235	239	250	255
No. increase		4	1	0	4	11	5
Rate of increase (percent)		1.74	0.43	0.00	1.70	4.60	2.00
Division of Humanities & Social Sciences							
Institute of History and Philology	65	66	66	69	60	58	56
Institute of Ethnology	39	40	25	27	28	27	27
Institute of Modern History	53	51	51	49	48	49	46
Institute of Economics	48	48	49	49	49	46	46
Institute of American and European Studies	35	34	37	34	34	33	32
Sun Yat-Sen Institute for Social Sciences and Philosophy	48	48	44	47	45	47	46
Institute of Sociology			17	18	19	21	18
Institute of Chinese Literature and Philosophy[a]	14	16	17	19	17	22	21
Institute of Taiwan History[a]	1	6	10	13	14	15	16
Institute of Linguistics[a]					12	11	11
subtotal	303	309	316	325	326	329	319
No. increase		6	7	9	1	3	−10
Rate of increase (percent)		1.98	2.27	2.85	0.31	0.92	−3.04
Total	749	765	771	785	798	820	815
No. increase	16	6	14	13	22	−5	66
Rate of increase (percent)	2.14	0.78	1.82	1.66	2.76	−0.61	

[a] Preparatory Office

scientific over humanitarian research in Taiwan. This imbalance came about because of the general need to modernize the country's economy in the postwar years. During this process, most resources were allocated to technology-oriented research. As a result, natural science and the life sciences enjoyed a major share of manpower as well as of budget. In later years, especially after the lifting of martial law on Taiwan in the 1980s, the environment for social sciences research improved greatly. The humanities and social sciences (at least in the Academia Sinica) have received relatively fair treatment if adequacy of regular budget is used as an indicator. This is perhaps a rare phenomenon worldwide.

The Academia Sinica claims to have achieved balanced development among its three divisions:

Mathematics and Physical Sciences, Life Sciences, and Humanities and Social Sciences. This has been achieved by setting different criteria for the various divisions with regard to both their research direction and evaluation, and by building up a lively academic community that allows researchers from various divisions to work on the same topic.

A recently completed study on Taiwanese aboriginals is one example of such a project. Under the coordination of a researcher from the Institute of Ethnology, staff from Biomedical Sciences, along with scholars from universities and hospitals, jointly investigated the migratory history, the hereditary traits from blood samples, and the ethnic differences of aboriginals in order to establish possible genetic polymorphic markers. This type of joint endeavor by different disciplines continues and is given priority in the funding process. An ongoing study of cardiovascular-related disease, the second leading cause of death in Taiwan, is another example. The study intends to collect multifaceted community data to help combat this disease. Funded by the Academia Sinica, researchers in epidemiology, social demography, economics and statistics have formed an interdisciplinary team to tackle the issue from various angles.

1.2.2 A balance between indigenous and international research. In the late 1990s the question of the relative priority of indigenization versus internationalization of academic research was much debated among social scientists in Taiwan. Supporters of indigenization emphasized the particularistic or the unique aspect of social science studies in Taiwan and the importance of avoiding the influence of dominant Western models. To others, however, internalization of social sciences is regarded as an inevitable global trend fitting into the theme of 'knowledge without national boundaries.'

Regarding debates on the nature of the research, the Academia Sinica takes a balanced stand. On the one hand, it encourages active participation in the valued conventional research areas. On the other hand, focusing on Taiwan's particular social issues and disseminating relevant research findings is considered to be important for the intellectual community and for mankind in general. Hence, the Academia Sinica has funded large-scale research projects that have both of the above purposes in goal.

A recent group project on the long-term development of capitalism in Taiwan has the potential to extend the Taiwanese experience to other Asian economies. This study encompasses the history of agricultural and industrial developments in Taiwan, trade and navigational expansion, macroeconomic performance, and the role the Taiwan Development Company played in the consolidation of capitalism in the territory. From the colonial era to the period of Japanese rule and the postwar era, Taiwan has gone through significant social transitions in its capitalist development. Although each stage may be characterized by different sets of institutions, one common factor emerges from the historical process: the expansion of exports (from tea, sugar, rice, processed foods, to light manufactured goods and producer goods). The exploration of the origin and the evolution of capitalist development in Taiwan will not only benefit the local academies, but will enhance the comparative study of economic development in other countries.

1.2.3 A balance between basic and applied research. Academic research is the main function of the Academia Sinica. However, increasing demands are being made on it to provide research results for applied use. The Academia Sinica recognizes the necessity to respond to important social phenomena and the need for technological development in its research endeavors. Heavy emphasis has been placed on the implementation of findings from basic research. An equal amount of effort has also been allocated to promoting applied research that may shed light on the consequent academic research. Two illustrations from Information Science and Life Science will highlight this recent focus.

A study on natural language understanding is directed towards the construction of a computer program with a knowledge system that is capable of understanding human perception of various recognition systems. The project has successfully developed a concept recognition mechanism called the 'Information Map.' This map arranges human knowledge in a hierarchical fashion with a cross-referencing capability. Using the information map, concept understanding can be reduced to intelligent semantic search in the knowledge system. This project has already produced a semantic 'Yellow Page' search for a telephone company and an automatic question and answer agent on the internet. A popular product of its Chinese phonetic input system is a typing software package widely used by the public. Another well-known research project with excellent applied values is the method for detecting differentially expressed genes. The research group has developed a DNA microarray with colorimetric detection system to simultaneously monitor the expression of thousands of genes in a microarray format on nylon membrane. Testing on filter membranes and quantifying the expression levels of the target genes in cells under different physiological or diseased states will reduce each screening process to a couple of days.

It is clear that the applied research is basically restricted to non-social sciences. In social science divisions, market-oriented or consumer-involved

studies remain quite rare. Researchers are mostly committed to basic research funded by the institute or by the National Science Council. Although policy study is ostensibly given a high priority, the fact that institutes to be established in the near future will still specialize in the basic disciplines reflects a general emphasis on this area. At present, few policy-oriented research projects are undertaken by individuals.

1.2.4 Coordination and promotion of academic research in Taiwan. Under the new organizational law, which allows more flexibility for research institutes to establish ad hoc research centers, the Academia Sinica will be given the responsibility of proposing a national academic research agenda, coordinating academic research in Taiwan, and training intellectual manpower. These tasks reflect the central role of the Academia Sinica in Taiwanese academies. In order to meet these requirements, the Academia has attempted to collaborate with various universities by exchanging teaching and research staffs. *Ad hoc* research committees and specific research programs including scholars from different institutes have also been established.

The committees on Sinological Research, on Social Problems, on Taiwan Studies, and on Mainland China Studies—all established since 1997—exemplify such an effort. These committees are interdisciplinary in nature, and comprise scholars from within as well as outside the Academia Sinica. The committee may focus on a few selected research issues and organize related seminars/conferences. The committee is also allowed to form various taskforces to plan future collaborative research topics. Besides the *ad hoc* research committees, the promotion of large-scale cross-institutional research projects has become important to the Academia Sinica. A so-called 'thematic project' will share a common research framework and will include several individual research topics proposed by researchers from different institutes and universities. The study of the organization-centered society represents one of these group efforts. In the investigation of modern social structure, nine subprojects were proposed, all funded exclusively by the Academia Sinica. Their findings on the importance of impersonal trust in Taiwan's economic development—instead of the traditional interpersonal trust—as key to organizational success, has given rise to the future research perspective that a modern society such as Taiwan is organization-centered. Similar thematic projects aiming to promote collaborative work among various academic institutes have been encouraged by the granting of funds.

However, it should be pointed out that playing a central role does not equate to having the central planning function. Taiwanese social scientists, in comparison with life scientists or Japanese colleagues, have a strong inclination to pursue individual researches. Various researchers joining together in a large communal laboratory or generations of scholars working in the same topic are not common at all. The thematic project in the Academia Sinica, or the joint project promoted by the National Science Council, is more of an initiative to encourage collaborative teamwork rather than a reaction to the present research demand. Whether individual projects maintain their importance or are replaced by group efforts will not change the expectations placed on the Academia Sinica—to respect individual research freedom and to facilitate research needs in Taiwan.

1.2.5 Encouragement of international collaboration. Active participation in international research has always been a priority in Taiwan. Researchers are encouraged to present their findings at international meetings and to build collaborative relationships with foreign research groups. Renowned scholars abroad are also frequently invited to visit and to work with research staff at the Academia Sinica. In line with this trend, a proposal has been made to establish an international graduate school at Academia Sinica (Chang 2000). The aim of this program is to attract highly qualified young people to do their Ph.D. degree under the guidance of top researchers in Taiwan. The competitive program is intended to provide the opportunity for independent inquiry as well as dynamic peer interaction; it is assumed that the supportive research environment will facilitate the training of future intellectual leaders and creative scholars. Although this program is yet to be finalized, the Academia Sinica has clearly revealed its interest in taking a concrete step towards globalization by investment in brilliant young minds.

1.2.6 Feedback into society. It is the firm belief among academicians that any type of feedback to society from members of the Academia Sinica must be based on solid academic research. Several feedback patterns have been adopted. With regard to emergent social issues, researchers with related knowledge and research findings are encouraged to air their suggestions by organizing conferences or public hearings. The problem of the over-plantation of betel nuts on hills and mountains and their harmful effect on the environmental protection is an example. Short-term research projects—such as the analysis of juvenile delinquency initiated by the committee on social problems—are another possible strategy.

Furthermore, the Academia Sinica opens its campus annually and individual research institutes sponsor introductory lectures to interested students and visitors. Numerous data sets from social science or biological researches, as well as valuable original historical files, are gradually released for public use.

1.3 The Role of the Academia Sinica

Whether the Academia Sinica should play a role beyond pure academic research and beyond research-based feedback to various social problems has always hotly debated. Some of the more articulate members have been quite vocal in insisting that it should have a stronger role in economic and social life. When it comes to how far the institute should involve itself in politics, however, the issue is more delicate. The subject is closely linked with the role of contemporary intellectuals in Taiwan (Chang 2000), where it is considered permissible for intellectuals to vocalize their moral conscience with regard to significant political issues. Nevertheless, it is precisely because most intellectuals are respected for their professional scholarship, not necessarily the correctness of their political views, that the appropriateness of actual political participation is seriously questioned.

For most ordinary members of the Academia Sinica, the pressure to produce excellent research is the foremost stimulus. Academic ambition is factored into the evaluation of promotion, the review criteria for assessing institutes' research accomplishment, the process of planning new research development, and in the regular report to the Legislative Yuan. Nevertheless, considering that research staff are government employees, the public perhaps has a right to voice concerns about the public utility of the Academia Sinica, whatever its academic credentials and quality of research.

When challenged about its value to the Taiwanese taxpayer, the Academica Sinica usually reminds the public of its past research accomplishments and its dynamic future role, concrete evidence of which can be found in several newly established research institutes.

When it comes to past achievements, the most senior institute in the Humanities and Social Science division—the Institute of History and Philology—is often cited. It was established in the same year as the Academia Sinica (1928). Early collective projects such as the An-Yang excavation, the study of Chinese dialects, and the reconstruction of ancient histories gained international fame for the institute. The institute is also engaged in the systematic compilation and organization of valuable Chinese historical documents, which contributes enormously to the field of Sinology and further enhances its academic status.

Far from resting on the academic achievements of the past, the Academia Sinica is constantly trying to stay on the cutting edge of research, as can be seen by the recently established institutes. In the division of Mathematics and Physical Sciences, the Institute of Astronomy and Astrophysics (1993) and the Institute of Applied Science and Engineering Research (1999) are the two latest research institutes. The separation between pure and applied science, especially in the life sciences, is obviously not applicable any more. Within the division of Humanities, the Institute of Taiwan History (1993) and the Institute of Linguistics (1997) were formed after drastic social changes had taken place in Taiwan. The Taiwan History institute is in the forefront of indigenous research in Taiwan and it has become the focal coordinating agency for Taiwan studies.

In short, Taiwan's Academia Sinica is a government-sponsored research institute. With funding available from the regular budget, academic research has been its main prescribed task. The highly trained research staff represents the research elite in Taiwan and has full liberty in deciding on individual projects. In recent years, the Academia Sinica has made a conscious effort to promote major interdisciplinary research programs both in fulfillment of its leadership role in the Taiwanese academies and as a response to changing societal expectations. A review of recent developments within the Academia Sinica reveals its intention to gain a greater global profile on the basis of its academic performance and generous research resources.

2. Research Academies in Mainland China

2.1 The Chinese Academy of Sciences

The Chinese Academy of Sciences was founded in 1949, the same year in which the Academia Sinica moved to Taiwan. With basic research as its main task, this academy has perhaps the largest organization of any institution of its type in the world. Besides the headquarters in Beijing, 22 branches made up of no fewer than 121 research institutes are scattered all over the country. Among five academic divisions (Mathematics and Physics, Chemistry, Biological Science, Earth Science, Technological Science), more than 40,000 scientists and technical professionals work for the Academy. Among them, nearly 10,000 are regarded as basic research staff, and 230 members of the Academy (out of the current 584 members who are elected as the most preeminent scientists in mainland China) also actively engage in research at the Academy. Members of the Academy enjoy the highest academic prestige. They play a planning and consulting role in China's scientific and technological development as well as providing reports and suggestions regarding important research issues.

A review of the general research orientation of the Chinese Academy of Sciences reveals at least two important characteristics which may differentiate it from the Academia Sinica in Taiwan. First, basic research with highly applied values was a major priority of the Academy from the beginning. High-tech research and development centers are growing rapidly and the support staff composed of well-trained professionals has become a major facilitating force. Second, collaboration with industrial sectors and with foreign institutes has contributed to the Academy's

research resources. The cooperative relationship has been substantial and extensive in that more than 3,000 enterprises have joined the industry–education–research development program. The international exchange program involves 7,000 personnel annually. This program has benefited both the research staff and the postgraduates of the Academy and fulfilled an important training function: more than 14,000 staff and graduate students have received advanced training abroad since 1978 and over 8,000 have completed their studies and returned to the Academy.

2.2 The Chinese Academy of Social Sciences

The Chinese Academy of Social Sciences was formally established in 1977 from the former Philosophy and Social Sciences division of the Chinese Academy of Sciences. The central headquarters in Beijing is made up of 14 research institutes employing 2,200 staff. Among these centralized institutes are Economics, Archeology, History, Law, and Ethnology. As is the case with the Academy of Sciences, there are many branch institutes throughout China so that the staff complement totals 4,200 in 31 research institutes.

According to much of the publicity material on the Academy of Social Sciences, the needs of the country appear to be of the utmost importance in the selection of research projects. The material and spiritual development and the democratization of the nation are constantly cited as basic motives to conduct relevant studies. This bias probably owes more to the fact that funding is from central government than to any policy implications. For example, the national philosophy and social sciences committee has organized several selected research topics every five years, such as the study of changes among rural families under the economic reform coordinated by Beijing University during the 7th national Five-year Plan. A substantial proportion of research undertaken by the Academy of Social Sciences consists of special commissions of this sort.

Because the availability of funding is the key to the commencement of any research, there is a substantial reliance on foreign funding. Funding from foreign foundations is usually generous enough to greatly enhance the possibility of conducting extensive studies across different regions of the nation. But collaboration with foreign institutes, especially in social science surveys, tends to be limited to data collection. Highly quality academic written manuscript from the collaborative project in Academy of Social Sciences is relatively inadequate and waits to be promoted in the future.

As in the Academy of Sciences, there is a long-standing international academic exchange program in the Academy of Social Sciences. More than 4,100 research staff and graduate students have participated in this program since 1978 and positive outcomes are revealed in new research projects as well as publications. The 82 academic journals published by the Academy cover various disciplines such as sociology, law, history, literature, world economics, etc.

3. Research Academies in Japan

There are basically two lines of research institutions in Japan: one under the Ministry of Education and the other under the Science and Technology agencies. University-affiliated research institutes, as well as independent national research institutes with graduate schools, come under the jurisdiction of Ministry of Education. As of 1997, among 587 Japanese universities, 62 had affiliated research institutes of which 20 were open to all researchers in Japan. There are also 14 independent research institutes in Japan, unaffiliated with any university, carrying out major academic research projects. These so-called inter-university research institutes are set up because of a specific demand to undertake academic research that requires resources and manpower beyond the university boundary. The National Laboratory for High Energy Physics was the first of this kind to be established (1971). The famous National Museum of Ethnology (1974) and the National Institute of Multimedia Education (1997), which aim at scientific research, data collection, and curricula development, have a substantial research complement of their own and are staffed by visiting scholars from abroad as well as local ones. The actual contribution of the inter-university research institutes lies in the nature of basic research. Large-scale facilities and data resources as well as human resources seconded from universities throughout Japan are considered important mechanisms in enhancing the progress of scientific research in Japan.

Other national research institutes, mostly concerned with natural sciences but also including social sciences (such as the noted National Institute of Population and Social Security Research and Economic Research Institute), fall into the domain of Science and Technology agencies. The population and social security research institute was founded in 1996 by combining two government research organizations: the Institute of Population Problems and the Social Development Research Institute. It is now affiliated with the Ministry of Health and Welfare. A research staff of 45 is located in seven research departments. Although policy-oriented research comprises a major proportion of the institute's work and is self-defined by the staff, academic research is still encouraged through both institutional and individual efforts. Surveys concerning population and social security are carried out to produce primary and secondary data for policy formulation. At the same time, these data also give rise to future academic studies on social and economic issues.

There is also a national advisory board on the scientific development of Japan. The Science Council

of Japan, attached to the Prime Minister's office, was established in 1949. Unlike the Japan Academy or academicians and academy members in Taiwan and mainland China, the 210 distinguished scientists from all fields who sit on the board are not given honorary lifetime titles but serve three-year terms of office. The council enjoys great academic prestige and represents Japanese scientists domestically as well as internationally. The council has the right to advise the government to initiate important scientific research programs, and the government may seek professional recommendations from the council as well. The council is also actively engaged in bilateral scientific exchanges and other forms of international participation. The council is also expected to coordinate academic research in Japan and facilitate the implementation of important decisions concerning academic development in Japan.

With a few exceptions, such as the Nihon University Population Research Institute, most academies in Japan are national. But the restrictions stemming from their organizational structure (they come under the jurisdiction of various government ministries) may translate into less research freedom or a higher demand for policy-oriented studies. In addition, those academies or research institutes affiliated with universities are usually also expected to carry out teaching functions at the individual researcher's level.

4. Conclusion

This article has briefly outlined the national character of research organizations in Taiwan, mainland China, and Japan. The importance of the government's role in academic development in this region can be clearly observed. The private sector, in contrast, plays only a minor role, if any, in academic research. However, several differences may be distinguished among the three territories. Taiwan's Academia Sinica is perhaps foremost in terms of both research autonomy and social services research. Benefiting from the cultural tradition and the expectations of the society in which it functions, researchers there also appear to enjoy more resources in funding and in social prestige. The motivation for research can be stated in purely academic terms and no policy orientation is required in order to receive adequate funding. In addition, Taiwanese scholars have shown a stronger preference for individual research projects. Mainland China, in comparison, launched its modern social science sector in the late 1970s, a time when a few basic disciplines such as sociology were still under suspicion. That contextual factor has certainly introduced an added element to the Chinese academy—the importance of correct political attitudes. As a consequence, research aims are required to be framed within the mainstream ideology and group efforts are more likely to be observed. Japan has a different tradition regarding academies. Although research autonomy is encouraged and political factors are not necessarily emphasized, research still tends to be applied in nature. This is largely because of structural factors, in that most academies are affiliated with various government ministries or with research institutes that are concerned with policy formulation and teaching as well as pure research. Also, Japanese social scientists are more inclined to participate in group projects headed by a leader in the field. Whether this is consistent with the national character remains to be explored.

With the relatively positive outlook for economic growth in the near future, the academies in Taiwan, mainland China, and Japan may experience substantial concomitant development and may thus reach new horizons in certain research fields. Nevertheless, academic collaboration within the region itself is still comparatively rare and may be a focus in the future agenda. Hopefully, a unique Asian perspective may be developed from persistent and extensive social science studies in these areas.

See also: Centers for Advanced Study: International/Interdisciplinary; History of Science; Science Funding: Asia; Scientific Academies, History of; Universities, in the History of the Social Sciences

Bibliography

Academia Sinica 1998 *Academia Sinica: A Brief Introduction*. Academia Sinica, Taipei, Taiwan

Brief Description of The Chinese Academy of Social Sciences. The Chinese Academy of Social Sciences, Beijing

Brief Description of the Science of Council of Japan. Science Council of Japan, Tokyo

Chang S. I 2000 Rationale for an International Graduate School at Academia Sinica. *Report presented at the Preliminary Discussion of the Mentoring of Graduate Students to Become Future Intellectual Leaders*. Academia Sinica, Taipei, Taiwan

Introduction to the National Institute of Population and Social Security Research in Japan. 2000 National Institute of Population and Social Security Research, Tokyo

Lee Y-T 1999 The Academic Development of Academia Sinica: Its Present Situation and its Future Vision. *The Newsletter of the Legislative Yuan*, The Legislative Yuan, Taipei, Taiwan

The Glorious Fifty Years of The Chinese Academy of Sciences: 1949—1999. 1999 The Chinese Academy of Sciences, Science Press

The Weekly Newsletter of Academia Sinica 2000 The Personnel Office of Academia Sinica

C.-C. Yi

Scientific Concepts: Development in Children

This article examines the development of children's scientific understanding. It is organized into four

sections: initial understanding; development of physical concepts; development of biological concepts; and learning processes.

1. Initial Understanding of Scientific Concepts

A large amount of recently obtained evidence indicates that infants' conceptual understanding is considerably more sophisticated than previously assumed. Traditionally, researchers relied on children's verbal explanations and/or their actions as measures of their conceptual understanding. These methods often underestimated infants' and very young children's understanding, owing to their inarticulateness and poor motor coordination. However, a new method, the violation-of-expectation paradigm, has made it possible to assess infants' physical knowledge by examining how long they look at 'possible' and 'impossible' events. A typical experiment involves habituating the child to a series of physically possible events and then presenting either a different possible event or an impossible event. The assumption is that children who understand the impossibility of the one event will look longer at it, because they are surprised to see the violation of the principle.

Studies using this violation of expectation paradigm have revealed impressive initial understanding of physical concepts. Even infants possess certain core concepts and understand several principles that govern the mechanical movement of objects (Spelke and Newport 1998). For example, 4-month-olds have a notion that one solid object cannot move through the space occupied by another solid object. In the studies that demonstrated this phenomenon, infants were first habituated to a display in which a ball was dropped behind a screen, which was then removed to reveal the ball on the floor. They then saw two events. In the consistent condition, the ball was again dropped, and when the screen was removed, infants saw the ball resting on a platform above the stage floor. In the inconsistent event, the ball was resting on the stage floor under the platform. Infants looked longer at the inconsistent event than at the consistent one, as if they were surprised to see the violation of the object–solidity principle. Other studies with a similar approach have also revealed infants' understanding of gravity and other physical regularities. Infants appear to understand that an unsupported object should move downward and that objects do not ordinarily move without any external force being applied.

Infants' knowledge of physical regularities is gradually refined over the first year, as demonstrated by their understanding of collisions. Most 3-month-olds appear surprised to see a stationary object move when not hit by another object. Six-month-olds can appreciate how the features of objects affect a collision. They appeared to be surprised when an object moves further following a collision with a small moving object than it does after colliding with a larger moving object. Later during the first year, infants responded differently to events in which the object that was struck moved at an angle perpendicular to the motion of the object that struck it than when it moved in a more standard path.

Other researchers, however, raise concerns about the use of such paradigms to draw inferences about infants' conceptual knowledge (e.g., Haith and Benson 1998). They argue that differential looking only indicates that infants discriminate between two events and that perceptual rather than conceptual features might drive this discrimination. Thus, infants' visual preference or looking time might have been incorrectly interpreted as evidence of an appreciation of physical principles.

Early conceptual understanding is not limited to understanding of physics principles. By age 3 years, children can distinguish living from nonliving things. They recognize that self-produced movement is unique to animals. In one study that made this point, preschoolers were shown brief videotapes in which animals or inanimate artifacts moved across the screen. Then the children were asked to make judgments about internal causes (does something inside this object make it move) and external causes (did a person make this move). Children typically attributed the cause of the animate object's motion to internal features ('it moves itself;' 'something inside makes it move'). In contrast, they were more likely to attribute the motion of an artifact to an external agent (Gelman 1990).

Preschoolers are also sensitive to differences in the 'types of stuff' inside animate and inanimate objects. They draw inferences about identity and capacity to function based on internal parts, associating, for example, internal organs, bones, and blood with animals and 'hard stuff' or 'nothing' with inanimate objects. After hearing a story about a skunk that was surgically altered so that it looked like a raccoon, young children reported that the animal was still a skunk, despite its altered appearance. However, children did not reason in the same way when they heard similar stories about artifacts; a key that was melted down and stamped into pennies was no longer a key (Keil 1989).

2. Developing Understanding of Physical Concepts

Rudimentary understanding of basic concepts does not imply full-blown appreciation of physical principles. Even older children's concepts and theories often involve substantial misconceptions. Understanding of physical concepts undergoes substantial change with age and experience. One good example involves physical causality and mechanical movement. When Event A precedes Event B, many 3- and 4-year-olds fail to choose A consistently as the cause, whereas

5- and 6-year-olds are considerably more likely to choose A. Children also hold intuitive theories of motion that are inconsistent with fundamental mechanical principles. For example, when asked to predict how a ball would travel after rolling through a spiral tube, only one-fourth of 9-year-olds and less than half of 11-year-olds correctly predicted the ball's trajectory.

Misconceptions also occur with other concepts. For example, children's conceptions of matter, weight, volume, and density undergo substantial change with age. Most 3-year-olds have undifferentiated notions of the roles of density, weight, and volume in producing buoyancy of objects placed in liquids. Most 4- and 5-year-olds have some conception of density, although their judgments are also affected by other features of the objects (i.e., weight and volume). Eight- and 9-year-olds, in contrast, rely consistently on density in judging whether objects will sink or float.

On some physical tasks, children's increasing understanding can be characterized as a series of increasingly adequate rules. One such task is Siegler's (1976) balance scale task. On each side of the scale's fulcrum were four pegs on which metal weights could be placed. In each trial, children were shown a configuration of weights on pegs and were asked to predict whether the scale would balance or whether one side would go down after release of a lever that held the scale motionless. Most children based their predictions on one of four rules. The large majority of 5-year-olds relied solely on weight (Rule I). This involved predicting that the scale would balance if both sides had the same amount of weight and that the side with more weight would go down if the two sides had different amounts of weight. Nine-year-olds often used Rule II. This involved predicting that the side with more weight would go down when one side had more weight, but predicting that the side with its weight further from the fulcrum would go down when weights on the two sides were equal. Some 9-year-olds and most 13–17-year-olds used Rule III. They considered both weight and distance on all problems, and predicted correctly when weights, distances, or both were equal for the two sides. However, when one side had more weight and the other side's weights were further from the fulcrum, children muddled through, not relying consistently on any identifiable approach. Rule IV allowed children to solve all balance scale problems. It involved choosing the side with greater torque ($W_L D_L$ vs. $W_R D_R$) when one side had more weight (W) and the other had its weight further from the fulcrum (D). Few children or adults used Rule IV. Similar sequences of rules have been shown to characterize the development of a variety of tasks, including projection of shadow, probability, water displacement, conservation of liquid and solid quantity, and time, speed, and distance.

A related way of conceptualizing the development of scientific understanding is as a succession of increasingly adequate mental models. One good example of such a succession involves understanding of the Earth as an astronomical object (Vosniadou and Brewer 1992). Some children, particularly young ones, conceive of the Earth as a flat, solid, rectangular shape. A slightly more sophisticated mental model is to think of the Earth as a disk. Three yet more advanced incorrect approaches are the 'dual Earth model,' which includes a flat, disk-like Earth where people live and a spherical Earth up in the sky; the 'hollow sphere model,' in which people live on flat ground inside a hollow sphere; and the 'flat sphere model,' in which, people live on flat ground on top of a hollow sphere. All three of these models allow children to reconcile their perception that the Earth looks flat with their teachers' and textbooks' insistence that the Earth is round. The proportion of children who possess the correct 'spherical model' of the Earth increases from 15 percent to 40 percent to 60 percent from first to third to fifth grade.

3. *Developing Understanding of Biological Concepts*

Young children have a concept of living things, but it does not perfectly match the concept of older children and adults. Until about age 7 years, most children do not view plants as living things. In addition, 3-year-olds fairly often attribute life to powerful, complex, or moving inanimate objects, such as robots and the Sun. A similar mix of understandings and misunderstandings is evident in preschoolers' views regarding internal parts of living things. They know that animals have bones and hearts, but have little idea of their functions.

Children's understanding of other uniquely biological concepts, such as growth, inheritance, and illness, also undergoes substantial change with age and experience. Preschool children have some appreciation of biological growth. They expect animals to grow, appreciate that growth can only occur in living things, and understand that growth is directional (small to big). However, preschoolers also believe that living things may or may not grow, and have difficulty accepting that even small things, such as a worm or butterfly, grow. Not until age 5 or 6 years do children realize the inevitability of growth; one cannot keep a baby pet small and cute just because one wants to do so.

Children's understanding of inheritance, another uniquely biological process, also develops with age and experience. Preschoolers understand that like begets like: Dogs have baby dogs, rabbits have baby rabbits, and offspring generally share biological properties with parents (Wellman and Gelman 1998). They also believe that animals of the same family share physical features even when they are raised in different environments. For example, preschoolers believe that

a rabbit raised by monkeys would still prefer carrots to bananas. However, other studies suggest that not until 7 years of age do children understand birth as part of a process mediating the acquisition of physical traits and nurturance as mediating the acquisition of beliefs. Only older children clearly distinguish between properties likely to be affected by heredity and properties likely to be affected by environment. For example, not until school age do children expect a boy to resemble his biological father in appearance but to resemble his adoptive father in beliefs. Thus, younger children seem to have different intuitions about the mechanisms of inheritance than do older children and adults.

Even preschoolers show some understanding of illness, yet another biological process. For example, preschoolers have a notion that an entity can induce illness or be contaminated and that contamination may occur through the workings of invisible, physical particles. Four- and 5-year-olds have some understanding of contagion; they believe that a child is more likely to get sick from exposure to another person who caught the symptom by playing with a sick friend than from another person who developed the symptom through other means. Although preschoolers may have a general idea that germs can cause symptoms, they do not differentiate the effects of symptoms caused by germs from those caused by poison, for example. The mature concept of illness, which is characterized as uniting various components such as its acquisition, symptoms, treatment, and transmission, is not mastered until much later (Solomon and Cassimatis 1999).

As with physical concepts, children's understanding of biological concepts involves substantial developmental change. For some concepts, changes involve enrichment, as children learn more and more details and phenomena relevant to the concepts. For others, developmental changes involve radical conceptual reorganization. There are some interesting parallels between the redefining and restructuring involved in the history of scientific understanding and the changes that occur within an individual lifetime (Carey 1985).

4. *Learning Processes*

Acquisition of scientific understanding involves the discovery of new rules and concepts through direct experience, as well as through instruction. Children's misconceptions can be overcome through experience that contradicts them. Only recently, however, have researchers directly examined the learning processes involved in the acquisition of scientific concepts. One approach that has proved particularly useful for learning about changing understanding of scientific concepts is the microgenetic method. This approach involves observing changing performance on a trial-by-trial basis, usually as children gain experience that promotes rapid change. Thus, the approach yields the type of high-density data needed to understand change processes (Siegler and Crowley 1991).

One example of the usefulness of the approach is provided by Siegler and Chen's (1998) study of preschoolers' learning about balance scales. The children were presented problems in which the two sides of the scale had the same amount of weight, but one side's weight was further from the fulcrum. The goal was to see if children acquired Rule II, which correctly solves such problems, as well as problems on which weight on the two sides varies but distance does not. Children's rules were assessed in a pretest and posttest in which children were asked to predict which side of the scale would go down or whether it would remain balanced. In the feedback phase between the pretest and post-test, children were repeatedly asked to predict which side of the balance, if either, would go down if a lever that held the arm motionless was released; then the lever was released and the child observed the scale's movement; and then the child was asked to explain the outcome they had observed. The trial-by-trial analysis of changes in children's predictions and explanations allowed the examination of the learning processes involved in rule acquisition. Four learning processes were identified. The first component of learning involves *noticing* potential explanatory variables (e.g., the role of distance) which previously had been ignored. The second step involves *formulating* a rule that incorporated distance as well as weight. To be classified as formulating a rule, children needed to explain the scale's action in one trial by stating that a given side went down because its disks were further from the fulcrum, and then in the next trial to predict that the side with its disks further from the fulcrum would go down. The third component involves *generalizing* the rule to novel problems by using it in most trials after it was formulated. Finally, the last component involves *maintaining* the new rule under less facilitative circumstances, by using the new rule in the posttest, where no feedback was given.

The componential analysis proved useful for understanding learning in general and also developmental differences in learning. The key variable for learning of both older and younger children, and the largest source of developmental differences in learning, involved the first component, noticing the potential explanatory role of distance from the fulcrum. Most 5-year-olds noticed the potential role of distance during learning, whereas most 4-year-olds did not. Children of both ages who noticed the role of distance showed high degrees of learning; those of both ages who did not, did not. The same componential analysis of children's learning of scientific concepts has proved useful in examining children's learning about water displacement and seems applicable to many other concepts also.

Although children often discover new rules, modify their mental models, or acquire new concepts through

direct experience both in physical and biological domains, direct observation of the natural world is often inadequate for learning new concepts. Indeed, daily experience sometimes hinders children's understanding. For example, children's misconceptions about the shape and motion of the Earth might result from the fact that the world looks flat. Well planned instruction is essential for helping children to overcome these misconceptions and to gain more advanced understanding.

The effects of instruction on children's scientific understanding can be illustrated by a study of the acquisition of the variable control principle (Chen and Klahr 1999). The variable control principle involves manipulating only one variable at a time so that an unconfounded experiment can be conducted and valid inferences can be made about the results. Most early elementary school children do not discover the variable control concept on their own. This observation led Chen and Klahr (1999) to test whether second, third and fourth graders could learn the concept through carefully planned instruction. Children were asked to design experiments to test the possible effects of different variables (whether the diameter of a spring affects how far it stretches, whether the shape of an object affects the speed with which it sinks in water, etc.). An unconfounded design would contrast springs that differed only in diameter but not in length, for example. Providing direct instruction proved to be effective in acquiring the control of variables concept. Both older and younger children who received instruction in designing tests in a specific task were able to understand the rationale and to apply the principle to other tasks. However, the older children were more able to extend the principle to novel contexts. When receiving training in designing tests involving the diameter of a spring, for example, second graders were able to apply the concept only to testing other variables involving springs (e.g., wire size). Third graders used the principle in designing experiments involving other mechanical tasks, such the speed with which objects sank in water. Only fourth graders, however, were able to apply the principle to remote contexts, such as experiments on the causes of plant growth.

In summary, recent research has revealed that infants, toddlers, and preschoolers have considerably greater scientific knowledge than previously recognized. However, their knowledge is far from complete, though. Developmental changes in children's scientific understanding involve both enrichment and structural reorganization. Older children possess more accurate and coherent rules and mental models. These understandings arise, at least in part, from their richer experience, their more advanced abilities to encode and interpret that experience, and their superior ability to separate their theories from the data (Kuhn et al. 1995). Older children also generalize the lessons from instruction more effectively than do younger children.

Although there is general agreement that early understanding of scientific concepts is surprisingly strong, there are disagreements about exactly what knowledge or concepts to attribute to infants, toddlers, and preschoolers. Even less is known about how children progress from their initial understanding of scientific concepts to more advanced understanding. Addressing the issue of how change occurs remains a major challenge, as well as a fruitful direction for future research.

See also: Cognitive Development in Childhood and Adolescence; Cognitive Development in Infancy: Neural Mechanisms; Concept Learning and Representation: Models; Early Concept Learning in Children; Infant Development: Physical and Social Cognition; Piaget's Theory of Child Development; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Carey S 1985 *Conceptual Change in Childhood*. MIT Press, Cambridge, MA

Chen Z, Klahr D 1999 All other things being equal: acquisition and transfer of the control of variables strategy. *Child Development* **70**: 1098–120

Gelman R 1990 First principles organize attention to and learning about relevant data; number and the animate–inanimate distinction as examples. *Cognitive Science* **14**: 79–106

Haith M, Benson J 1998 Infant cognition. In: Damon W, Kuhn D, Siegler R S (eds.) *Handbook of Child Psychology*. *Vol. 2. Cognition*, *Perception and Language*, 5th edn. J. Wiley, New York, pp. 199–254

Keil F C 1989 *Concepts*, *Kinds*, *and Cognitive Development*. MIT Press, Cambridge, MA

Kuhn D, Garcia-Mila M, Zohar A, Andersen C 1995 Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development* **60**: 1–127

Siegler R S 1976 Three aspects of cognitive development. *Cognitive Psychology* **8**: 481–520

Siegler R S, Chen Z 1998 Developmental differences in rule learning: a microgenetic analysis. *Cognitive Psychology* **36**: 273–310

Siegler R S, Crowley K 1991 The microgenetic method: a direct means for studying cognitive development. *American Psychologist* **46**: 606–20

Solomon G E A, Cassimatis N L 1999 On facts and conceptual systems: young children's integration of their understandings of germs and contagion. *Developmental Psychology* **35**: 113–26

Spelke E S, Newport E L 1998 Nativism, empiricism, and the development of knowledge. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology*. *Vol. 1. Theoretical Models of Human Development*, 5th edn. J. Wiley, New York, pp. 275–340

Vosniadou S, Brewer W F 1992 Mental models of the earth: a study of conceptual change in childhood. *Cognitive Psychology* **24**: 535–85

Wellman H M, Gelman S A 1998 Knowledge acquisition in foundational domains. In: Damon W, Kuhn D, Siegler R S (eds.) *Handbook of Child Psychology. Vol. 2. Cognition, Perception and Language*, 5th edn. J. Wiley, New York, pp. 575–630

Z. Chen and R. Siegler

Scientific Controversies

Science in general can be an object of controversy such as in disputes between science and religion. Particular scientific findings can also generate controversies either within or outside science. The importance of scientific controversy has been recognized by scholarship within science and technology studies (S&TS) since the 1970s. Indeed the study of controversies has become an important methodological tool to gain insight into key processes that are not normally visible within the sciences.

What makes something a scientific controversy? It is important to distinguish longstanding disputes, such as that between science and religion, or the merits of sociobiological explanation as applied to humans, or whether the fundamental constituents of matter are particles or waves, from more localized disputes such as over the existence of a new particle or a new disease transmitting entity. The latter sorts of controversy are more like a 'hot spot' that erupts for a while on the surface of science than a deeply entrenched long-running battle. Also controversies are not be confused with the bigger sea changes which science sometimes undergoes during scientific revolutions. Although defining a scientific revolution is itself contested, the all-pervasive nature of the changes in physics brought about by quantum mechanics and relativity seems different from, for example, the controversy over the detection of large fluxes of gravitational radiation or over the warped zipper model of DNA. Similarly, long-running debates on the relative impacts of nature and nurture on human behavior have a different character from more episodic controversies, such as the possibility of interspecies transfer of prion disease. Of course, intractable disputes and revolutions share some of the features associated with controversies, but it is the bounded nature of controversies which has led to their becoming an object of study in their own right, especially within the tradition of S&TS associated with the sociology of scientific knowledge (SSK).

One metaphor for understanding why controversies have taken on such methodological importance is that of 'punching' a system. On occasions scientists gain insight into natural systems by punching, or destabilizing, them. For example, one may learn more about the laws of momentum by bouncing one billiard ball off another than by watching a stationary billiard ball. Similarly Rutherford famously used scattering experiments in which gold foil was bombarded with alpha particles to uncover the structure of the atom and in particular the presence of the nucleus. The methodological assumption underpinning the study of controversies is similar. By studying a scientific controversy, one learns something about the underlying dynamics of science and technology and their relations with wider society. For instance, during a controversy the normally hidden social dimensions of science may become more explicit. Sites of contestation are places to facilitate the investigation of the metaphors, assumptions, and political struggles embedded within science and technology.

We can note four different influential approaches towards the study of scientific controversies. The school of sociological research associated with Robert Merton (1957) first recognized the importance of controversies within science. Of particular interest to Merton was the existence of priority disputes. Many well-known controversies center on who is the first scientist to make a particular scientific discovery. A second approach toward the study of scientific controversy developed in the 1960s as concerned citizens increasingly protested what they took to be the negative effects of science and technology. Here the source of controversy is the perceived negative impact of science and technology on particular groups and it is the study of these political responses that forms the core of the analysis. The new SSK which emerged in the 1970s and which largely displaced the Mertonian School provides a third approach towards the study of controversies. Here the focus is on controversies at the research frontiers of science where typically some experimental or theoretical claim is disputed within an expert community. Modern S&TS owe a heavy debt to SSK but are less likely to make distinctions between the content of science and its impact. Within this fourth approach, controversies are seen as integral to many features of scientific and technological practice and dissemination. Their study forms a key area of the discipline today.

1. *Merton and Priority Disputes*

Merton's interest in priority disputes stemmed from his claim that science has a particular normative structure or 'institutional ethos' with an accompanying set of rewards and sanctions. Because so much of the reward structure of science is built upon the recognition of new discoveries, scientists are particularly concerned to establish the priority of their findings. Such priority disputes are legion, such as the famous fight between Newton and Leibnitz over who first discovered the calculus.

It was Thomas Kuhn (1962) who first raised a fundamental problem for the analysis of priority

disputes. A priority dispute is predicated upon a model of science, later known as the 'point model' of scientific discovery, which can establish unambiguously who discovered what and when. Asking the question of who discovered oxygen, Kuhn showed that the crucial issue is what counts as oxygen. If it is the dephlogisticated air first analyzed by Priestly then the discovery goes to him, but if it is oxygen as understood within the modern meaning of atomic weights then the discovery must be granted to Lavoisier's later identification. The 'point model' requires discovery to be instantaneous, and for discoveries to be recognized and dated. A rival 'attributional model' of discovery, first developed by Augustin Brannigan (1981), draws attention to the social processes by which scientific discoveries are recognized and 'attributed.' This approach seems to make better sense of the fact that what counts as a discovery can vary over time. In short, it questions the Eureka moment of the point model.

For example, Woolgar (1976), in his analysis of the pulsar's discovery, shows that the date of the discovery varies depending on what stage in the process is taken to be the defining point of the discovery. If the discovery is the first appearance of 'scruff' on Jocelyn Bell's chart recording of signals from the radio telescope, then it will be dated earlier than when it was realized that the unambiguous source of this 'scruff' was a star. This case was particularly controversial because it was alleged by the dissonant Cambridge radio astronomer Fred Hoyle that the Nobel Prize winners for this discovery should have included Jocelyn Bell, who was then a graduate student. Priority disputes can touch in this way on the social fabric of science, such as its gender relationships and hierarchical structure.

Despite the challenge posed by the attributional model, it is the point model of discovery that is embedded in the reward system of science. As a result, priority disputes still abound. In modern technoscience, discovery can mean not only recognition, but also considerable financial reward, as for example with patents, licensing arrangements, or stock in a biotech company. In such circumstances, priority disputes have added salience. One has only to think of the unseemly battle between Robert Gallo and the Pasteur Institute over priority in the discovery that HIV is the cause of AIDS. In this case, there was not only scientific priority at stake, but also the licensing of the lucrative blood test for identifying AIDS. The controversy could only be settled by intervention at the highest political level. The Presidents of the USA and France, Ronald Reagan and Jacques Chirac, agreed to share the proceeds from the discovery. Again what was at stake scientifically was not simply who was first past the post; the protagonists initially claimed to have isolated different retroviruses and disagreed over the effectiveness of the various blood tests. This case was marked by additional controversy because of allegations of scientific misconduct raised against Gallo that led to Congressional and National Institute of Health (NIH) investigations.

2. *Controversy over the Impact of Science and Technology*

That a priority dispute could require the intervention of national political leaders is an indication of just how important science and technology can become for the wider polity. In response to the AIDS crisis, activist groups have campaigned and pressured scientists and government officials to do more scientifically. They have also intervened in matters of research design, such as the best way to run clinically controlled trials. Such activist engagement dates back to the political protests that science and technology generated in the 1960s in the context of Vietnam-era issues such as war and environmentalism. There has been increasing recognition that science and technology are neither neutral nor necessarily beneficial and that many developments stemming from modern science and technology, such as nuclear power, petrochemical industries, and genetic engineering, raise profound and controversial issues for a concerned citizenry.

Dorothy Nelkin, a pioneer in analyzing these types of disputes identified four types of political, economic, and ethical controversies that engage the public in the US (Nelkin 1995). One set revolves around the social, moral, and religious impact of science. Issues such as the teaching of evolution in US schools, animal rights, and the use of fetal tissue fall into this first category. A second type of controversy concerns a clash between the commercial and economic values surrounding science and technology and that of the environmental movement. Ozone depletion, toxic waste dumps, and greenhouse gases are pertinent examples. A third set has been provoked by health hazards arising from the transformation of food and agricultural practices by the use of modern science and technology. Genetically modified foods, the carcinogenic risks posed by food additives, and the use of bovine growth hormones in the dairy industry all belong in this category. A fourth group centers on conflicts between individual rights and group rights: a conflict that has been heightened by new developments in science and technology. For example, the mass fluoridation of water to improve dental health denies individuals the right to choose for themselves whether they want fluoride in their water supply.

Research on these sorts of controversies has focused mainly on the interest politics of the groups involved. How and why do they get involved in political action over science and technology; what underlying political values do such groups exhibit; and how do they effectively intervene to protest some perceived deleterious development stemming from science, technology, or medicine? The positions taken by the

participants are consistent with their interests, although these interests may not enable the outcome or closure of a debate to be predicted. For instance, the demise of nuclear power had as much to do with economics as with political protest. Since scientists themselves often play an active part in these disputes, a full analysis will touch upon how scientists deploy their science for political aims. But, by and large, this research tradition has avoided using the entry of scientists into these disputes to examine the core processes by which scientific knowledge is developed and certified. In short, the attention was focused upon seeing how scientists became political rather than upon how politics might itself shape scientific knowledge. Political controversies were treated as analytically separable from epistemic controversies and as resolved by distinct processes of closure (Engelhardt and Caplan 1987). Typically, epistemic controversies were thought to be closed by application of epistemic and methodological standards, while political controversies were closed through the intervention of 'non-scientific factors,' such as economic and political interests.

3. *Scientific Controversy and the Sociology of Scientific Knowledge*

With the emergence of the SSK in the late 1970s, it was no longer possible to avoid examining how scientific knowledge was shaped and how this shaping contributed to the dynamics of controversies. A key tenet of this new sociology of science, as formulated by David Bloor (1991) in his 'Strong Programme,' was that of symmetry. This principle called upon sociologists to use the same explanatory resources to explain both successful and unsuccessful knowledge claims. It raised to methodological status the necessity of examining the processes by which science distinguishes the wheat of truth from the chaff of error. SSK soon turned its attention towards examining scientific controversies because it is during such controversies that this symmetry principle can be applied to good effect. With each side alleging that it has 'truth' on its side, and disparaging the theoretical and experimental efforts of the other, a symmetrical analysis can explain both sides of the controversy using the same sorts of sociological resources. This differs from the earlier interest approach to controversies in that it applies this symmetrical sociological analysis to the very scientific claims made by the participants.

Bloor and his colleagues of the Edinburgh school pursued their program mainly through theoretical analysis supported by historical case studies. H. M. Collins and the 'Bath School' by contrast, developed an empirical method for studying the SSK in contemporaneous cases: a method based primarily upon the study of scientific controversies. One early application of the method was to the study of parapsychology (Collins and Pinch 1982). Collins and Pinch suggested that controversies such as that provoked by parapsychology were resolved by boundary crossing between two different forums of scientific debate: the constitutive and the contingent. Generalizing from several case studies of controversies, Collins (1981) argued that during controversies scientific findings exhibited 'interpretative flexibility,' with the facts at stake being debated and interpreted in radically different ways by the parties in the controversy.

This interpretative flexibility did not last forever: by following a controversy over time, researchers could delineate the process of 'closure' by which controversy vanished and consensus emerged. Collins defined the group of scientists involved in a controversy as the 'core set.' Only a very limited set of scientists actively partook in controversies; the rest of the scientific community depended upon the core set for their expert judgment as to what to believe. This was particularly well illustrated by Martin Rudwick (1985) in his study of the great Devonian Controversy in the history of geology.

As researchers followed controversies from their inception to the point of closure, it became necessary to address matters of scientific method as they were faced in practice by the participants. Factors that had usually been seen as issues of method or epistemology thus became open to sociological investigation: for example, the replication of experiments, the role of crucial experiments, proofs, calibration, statistics, and theory. In addition, other factors such as reputation, rhetoric, and funding were shown to play a role in the dynamics of controversies.

An important finding of this research was what Collins (1992) called the 'experimenter's regress.' Controversies clearly were messy things and were very rarely resolved by experiments alone. Collins argued that in more routine science experiments were definitive because there was an agreed-upon outcome which scientists could use as a way of judging which scientists were the competent practitioners. If one could get one's experiment to work, one had the requisite skills and competence; if one failed, one lacked the skills and competence. The trouble was that when there was a dispute at the research frontiers there was no agreed upon outcome by which to judge the competent practitioners. Experiments had to be built to investigate a claimed new phenomenon, but failure to find the new phenomenon might mean either there was no new phenomenon to be found or that the experimenter failing to find it was incompetent. This regress was only broken as a practical matter by the operation of a combination of factors such as rhetoric, funding, and prior theoretical dispositions.

Often the losing side in a scientific controversy continues to fight for its position long after the majority consensus has turned against it. Those who continue will meet increasing disapprobation from

their colleagues and may be forced to leave science altogether. 'Life after death' goes on at the margins and often finally passes away only when the protagonists themselves die or retire (Simon 1999). The uncertain side of science is clearest during moments of controversy. Most scientists never experience controversies directly, and often it is only after exposure to a controversy that scientists become aware of the social side of science, start reading in science studies, and even employ ideas drawn from science studies to understand what has happened to them.

This work on scientific controversy has been exemplified by a number of case studies of modern science such as memory transfer, solar neutrinos, gravity waves, high-energy physics, and famously cold fusion. Historians have also taken up the approach used by sociologists and the sociological methods have been extended to a number of historical case studies. Such studies pose particular methodological challenges because often the losing viewpoint has vanished from history. Shapin and Schaffer's (1985) study of the dispute between Robert Boyle and Thomas Hobbes over Boyle's air pump experiments was a landmark in research on scientific controversy, because it showed in a compelling way how the wider political climate, in this case that of Restoration Britain, could shape the outcome of a controversy. It showed how that climate could help institutionalize a new way of fact-making experiments in the Royal Society at the same time. In addition, it drew attention to the literary and technological dimensions of building factual assent in science. By documenting the witnesses to particular experimental performances, a culture of 'virtual witnessing' was born.

The SSK approach to scientific controversy has also been influential in the study of technology. The social construction of technology (SCOT) framework uses concepts imported from the study of scientific controversy such as 'interpretative flexibility' and 'closure.' A variety of competing meanings are found in technological artifacts and scholars study how 'closure mechanisms' such as advertising produce a stable meaning of a technology (Pinch and Bijker 1987, Bijker 1995).

Another influential approach to the study of controversies in sciences and technology has been that developed by Bruno Latour and Michel Callon. Again, the initial impetus came from studies of scientists. Callon's (1986) article on a controversy over a new method of harvesting scallops is one of the first articulations of what later became known as Actor Network Theory (ANT). Callon argues that the outcome of a controversy cannot be explained by reference to the social realm alone, but the analyst must also take account of the actions of non-human actors, such as scallops, which play a part in shaping the outcome. Subsequently Latour's work on how 'trials of strength' are settled in science and technology has become especially influential within the new SSK. Such struggles, according to Latour (1987), involve aligning material and cognitive resources with social ones into so-called 'immutable mobiles' or black boxes, objects which remained fixed when transported along scientific networks and which contain embedded within them sets of social, cognitive, and material relationships.

Latour and Woolgar (1979) in their now classic study of a molecular biology laboratory showed that literary inscriptions play a special role in science. They indicated how controversies could be analyzed in terms of whether certain modalities are added to or subtracted from scientific statements making them more or less fact-like. The role of discourse in scientific controversies has been examined in great depth in a study of the oxidative phosphorylation controversy by Gilbert and Mulkay (1984). They showed how particular repertoires of discourse, such as the 'empiricist repertoire' and the 'contingent repertoire,' are used selectively by scientists in order to bolster their own claims or undermine those of their opponents. Subsequently there has been much work on how a variety of rhetorical and textual resources operate during controversies (e.g., Myers 1990). Sometimes the resolution of controversy is only possible by drawing boundaries around the relevant experts who can play a role in the controversy. Sometimes particular scientific objects cross such boundaries and form a nexus around which a controversy can be resolved. Such 'boundary work' (Gieryn 1983) and 'boundary objects' (Star and Griesemer 1989) form an important analytical resource for understanding how controversies end.

In addition to analyzing scientific controversies, SSK has itself become a site of controversy. Most notably, lively controversies have occurred over the viability of interest explanations, over the extent to which the sociology of science should itself be reflexive about its methods and practices, and over the role of non-human actors. The 'science war' involving debates between natural scientists and people in science studies over the methods and assumptions of science studies and cultural studies of science is another area of controversy that is ripe for sociological investigation.

4. *Scientific Controversy in Science and Technology Studies Today*

In contemporary S&TS, the sites of contestation chosen for analysis have become more heterogeneous. One strength of the new discipline of S&TS is the wide terrain of activities involving science and technology that it examines. For example, similar methods can be used to examine controversies involving science and technology in the courtroom, the media, quasi-governmental policy organizations, and citizens' action groups. Indeed, many of the most contentious political issues facing governments and citizens today involve

science and technology: issues such as genetically modified foods, gene therapy, and *in vitro* fertilization.

The study of controversies in modern techno-science—with its porous boundaries between science, technology, politics, the media, and the citizenry—also calls for the analyst to broaden the array of analytical tools employed. Although the fundamental insights produced by SSK remain influential, such insights are supplemented by an increased understanding of how macro-political structures such as the state and the legal system enable and constrain the outcome of scientific controversies. Examples of this sort of work include:

Jasanoff's (1990) investigations of how technical controversies are dealt with by US agencies such as the Environmental Protection Agency (EPA) and the Food and Drug Administration (FDA),

Lynch and Jasanoff's (1998) work on science in legal settings such as the use of DNA evidence in courtrooms, and

Epstein's (1996) work on the AIDS controversy.

In the latter case Epstein deals not only with the dispute about the science of AIDS causation, but also turns to social movements research to understand how AIDS activists outside of science got sufficient influence actually to affect the design of clinical trials by which new AIDS drugs are tested.

Particularly interesting methodological issues have been raised by the study of controversies that overtly impinge upon politics. When studying controversies within science SSK researchers were largely able to adopt the neutral stance embodied in the symmetry principle of the Strong Programme (see *Epigenetic Inheritance*). However, some scholars have argued that, when dealing with cases where analysis could have a direct impact upon society, it is much harder to maintain neutrality. Researchers studying these sorts of disputes, such as whether Vitamin C is an effective cancer cure, find they can become 'captured' by the people they are studying. This complicates the possibility of producing the sort of neutral analysis sought after in SSK. A number of solutions have been proposed to this dilemma (see Ashmore and Richards 1996).

Several authors have attempted to produce typologies of scientific controversies (Engelhardt and Caplan 1987, Dascal 1998). Unfortunately, such typologies are not as useful as they could be because they are confounded by their underlying epistemological assumptions. For example, within an SSK approach it makes little sense to work with a category of, say, 'sound argument,' for closing a controversy because in SSK 'sound argument' is seen as part of the controversy. What counts as a 'sound argument' can be contested by both sides. Martin and Richards in their (1995) review adopt a fourfold typology. This review is particularly useful because it distinguishes between the different types of epistemological assumptions underlying the different analytical frameworks employed: namely, positivist, group politics, SSK, and social structural.

Traditional history and the philosophy of science according to one account (Dascal 1998) are becoming more cognizant of the phenomenon of scientific controversy. But the call to examine scientific controversy by historians and philosophers makes strange reading to scholars immersed in S&TS. It is as if the historians and philosophers of controversy simply have ignored or failed to read the relevant literatures. Thus neither Nelkin's (1992) influential volume, *Controversies* nor a special edition of *Social Studies of Science* edited by H. M. Collins (1981), *Knowledge and Controversy*, which sets out the SSK approach towards scientific controversy, are referenced. That the best way to study scientific controversy is still controversial within the academy could scarcely be more obvious.

See also: Epigenetic Inheritance

Bibliography

Ashmore M, Richards E 1996 The politics of SSK: Neutrality, commitment and beyond. Special issue of *Social Studies of Science* **26**: 219–445

Bijker W 1995 *Of Bicycles, Bakelites, and Bulbs: Towards a Theory of Sociotechnical Change*. MIT Press, Cambridge, MA

Bloor D 1991 *Knowledge and Social Imagery*, 2nd edn. University of Chicago Press, Chicago

Brannigan A 1981 *The Social Basis of Scientific Discoveries*. Cambridge University Press, Cambridge, UK

Callon M 1986 Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieux Bay. In: Law J (ed.) *Power, Action and Belief: A New Sociology of Knowledge?* Sociological Review Monograph, Routledge, London, pp. 196–229

Collins H M 1981 Knowledge and controversy: Studies in modern natural science. Special issue of *Social Studies of Science* **11**: 1

Collins H M 1992 *Changing Order*, 2nd edn. University of Chicago Press, Chicago

Collins H M, Pinch T J 1982 The construction of the paranormal, nothing unscientific is happening. In: Wallis R (ed.) *On the Margins of Science: The Social Construction of Rejected Knowledge*. Sociological Review Monograph, University of Keele, Keele, UK

Dascal M 1998 The study of controversies and the theory and history of science. *Science in Context* **11**: 147–55

Engelhardt T H, Caplan A L 1987 *Scientific Controversies: Case Studies in the Resolution and Closure of Disputes in Science and Technology*. Cambridge University Press, Cambridge, UK

Epstein S 1996 *Impure Science: AIDS, Activism and the Politics of Knowledge*. University of California Press, Berkeley, CA

Gieryn T 1983 Boundary work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review* **48**: 781–95

Gilbert N G, Mulkay M K 1984 *Opening Pandora's Box*. Cambridge University Press, Cambridge, UK

Jasanoff S 1990 *The Fifth Branch: Science Advisors as Policy Makers*. Harvard University Press, Cambridge, MA

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
Latour B 1987 *Science in Action*. Harvard University Press, Cambridge, MA
Latour B, Woolgar S W 1979 *Laboratory Life*. Sage, London
Lynch M, Jasanoff S (eds.) 1998 Contested identities: Science, law and forensic practice. Special issue of *Social Studies of Science* **28**: 5–6
Martin B, Richards E 1995 Scientific knowledge, controversy, and public decision making. In: Jasanoff S, Markle G E, Petersen J C, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA
Merton R K 1957 Priorities in scientific discoveries: A chapter in the sociology of science. *American Sociological Review* **22**: 635–59
Myers G 1990 *Writing Biology: Texts and the Social Construction of Scientific Knowledge*. University of Wisconsin Press, Madison, WI
Nelkin D (ed.) 1992 *Controversies: Politics of Technical Decisions*, 3rd edn. Sage, Newbury Park, CA
Nelkin D 1995 Science controversies: The dynamics of public disputes in the United States. In: Jasanoff S, Markle G E, Petersen J C, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA
Pinch T J, Bijker W E 1987 The social construction of facts and artifacts. Or how the sociology of science and the sociology of technology might benefit each other. In: Bijker W E, Hughes T S, Pinch T J (eds.) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, MA
Rudwick M 1985 *The Great Devonian Controversy: The Shaping of Scientific Knowledge Among Gentlemanly Specialists*. University of Chicago Press, Chicago
Shapin S, Schaffer S 1985 *Leviathan and the Air-pump: Hobbes, Boyle and the Experimental Life*. Princeton University Press, Princeton, NJ
Simon B 1999 Undead science: Making sense of cold fusion after the (arti)fact. *Social Studies of Science* **29**: 61–87
Star S L, Griesemer J 1989 Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939. *Social Studies of Science* **19**: 387–420
Woolgar S 1976 Writing an intellectual history of scientific development: The use of discovery accounts. *Social Studies of Science* **6**: 395–422

T. Pinch

Scientific Culture

1. Modern Scientific Culture

All human societies develop knowledge about their natural and social worlds. Even hunter-gatherers possess remarkable local knowledge about plants, animals, and climate. The great world civilizations had a more complex division of labor that allowed priests, doctors, smiths, farmers, and other specialists to develop more elaborate and less local knowledge systems about astronomy, psychology, medicine, metallurgy, agriculture, and other fields. Researchers tend to reserve the term 'science' for the elaborated, written knowledge systems of the great world civilizations—such as ancient Greek or Chinese science—and they tend to describe other, more local knowledge systems as ethnomedicine, ethnoastronomy, ethnobotany, and so on.

Many attempts have been made to describe the distinctive features of the type of science that emerged in Western Europe around 1500–1700 (Cohen 1997). The term 'modern science' is somewhat clearer than 'Western science,' not only because Western science was built up in part from non-Western sources, but also because Western science rapidly became globalized. However, two clarifications are in order. First, the term 'scientist' did not emerge until the nineteenth century; in earlier centuries the term 'natural philosopher' was more common. Second, the term 'modern' is used here to refer to a society characterized by a family of institutions that includes modern science, constitutional democracy, capitalism, religious pluralism, social mobility, and a universalistic legal system. Although some of those institutions can be found in some of the world's societies prior to 1500, their development as a system occurred in the West and accelerated after 1500. As an institution, science soon found a niche in modern societies by providing research and/or ideological support for the emerging capitalist industries, the state, and the churches (e.g., ballistics, mining, navigation, taxation, public health, critiques of magic).

Modern societies provided more than financial resources that supported the growth of modern science as an institution; the culture of modernity provided intellectual resources that contributed to the emergence of modern scientific inquiry. Three of the central values of modern scientific culture are empiricism, formalism, and mechanism. Although each of the distinctive features of modern scientific culture can be found in other scientific cultures (and may not be found across all modern scientific disciplines), as a family of features they have some value in characterizing modern scientific culture, and in showing its points of confluence with the cultures of modernity more generally.

Empiricism refers to the value placed on observations as a means for resolving disputes about natural and social knowledge. In some sciences, empiricism was developed as experimentalism; in other words, observations could be brought into the laboratory and subjected to experiments that were, in principle, reproducible by competent members of the scientific community. However, other fields remained non-experimental, such as the fieldwork-based sciences. Because observations are always subject to interpretation, their use as a resource for dispute resolution was in turn rooted in broader societal cultures. Scientists had to trust each other not to lie (Shapin 1994), and they required societies and journals in which they could debate and share results. Those

requirements were met in the European societies that fostered the emergence of a 'relatively autonomous intellectual class' (Ben-David 1991, p. 304) and a public sphere of open debate (Habermas 1989).

The relationship between empiricism and the broader society went beyond institutional requirements; other sectors of society were also characterized by an empirical cultural style. For example, constitutional democracies and markets were based on the 'empiricism' of elections and consumer purchases. Likewise, some Protestant churches replaced church dogma with the empiricism of direct interpretation of Bibles, and they emphasized knowledge of God through study of his works (Hooykaas 1990, p. 191, Merton 1973, Chap. 11). In this sense, the empiricism of science emerged as part of a more general way of resolving conflicting opinions about the world through recourse to data gathering and evidence.

Formalism refers to the value placed on increasingly higher levels of generalization. As generalization progresses in science, concepts and laws tend to become increasingly abstract and/or explicit. In some fields, the generalizations took the form of mathematical laws that encompassed a wide range of more specific observations. In physics, for example, it became possible to apply the same set of formal laws to the mechanics of both terrestrial and celestial objects, and space and time were analyzed in terms of geometry (Koyré 1965). In other fields, the generalizations took the form of increasingly formal and abstract taxonomies and systems of classification, as in early modern biology (Foucault 1970).

Again, broader institutions and values contributed to the emergence of this specific form of inquiry. When researchers found support for societies that published journals where an archive of research could be located, they found the institutional resources that allowed them to conjugate their work with that of others. Likewise, as Western colonial powers expanded, they sent out scientific expeditions that incorporated local knowledges and new observations into existing research fields. As 'Western' science became increasingly cosmopolitan, it also became more abstract. Again, however, the relationship with the broader society went beyond institutional influences to shared values. For example, scientific formalism developed in parallel with modernizing, Western political systems and social contract theories that emphasized the abstraction of general laws from particular interests, or of a general good from individual wills.

A third common feature is the value of mechanism as a form of explanation. Over time, modern science tended to rule out occult astrological forces, vitalistic life forces, and so on. When the concept of force was retained, as in Newton's gravitational force, it was subjected to the restraints of a formal law. Again, general social and cultural resources supported this development. The attack on occult forces was consistent with reformation campaigns against popular magic (Jacob 1988), and the disenchantment of the world that mechanistic models depended on was both supported by Christianity and intensified by some forms of Protestantism. The development and spread of clocks and constitutions provided an early metaphor of mechanism, and as new technologies and social charters were developed, new metaphors of mechanism continued to emerge (Kammen 1987, pp. 17–8, Shapin 1996, pp. 30–7). Sciences that violate the cultural value of mechanism, such as parapsychology (the study of claimed paranormal phenomena), are rejected. Likewise, the incorporation of local sciences has tended to occur after filtering out occult or vital forces. For example, in response to demand from patients and cost efficiency concerns, acupuncture and Chinese medicine are being incorporated into cosmopolitan biomedicine. However, even as the practices are being incorporated, the vitalistic concept of 'chi' and Chinese humoral concepts are being translated into mechanistic concepts consistent with modern biology and physics.

Just as values such as empiricism, formalism, and mechanism have been used to describe the intellectual culture of modern science, so a related family of concepts has been used to describe its institutional culture. Most influential was sociologist Robert Merton's (1973, Chap. 13) list of four central norms. Although subsequent research suggested that norms were frequently violated and better conceptualized as an occupational ideology (Mulkay 1976), Merton's analysis did have the advantage of pointing to the fundamental preconditions for the existence of modern scientific culture. Perhaps the basic underlying institutional value is autonomy. In other words, there is a value placed on leaving alone a certified community of qualified peers to review and adjudicate the credibility of various claims of evidence and consistency, rather than have the function ceded to the fiat of kings, dictators, church leaders, business executives, and others who do not understand the research field. The value of autonomy creates an interesting tension between science and another modern institution, constitutional democracy. Although scientific communities have suffered greatly under nondemocratic conditions, their demand for some degree of autonomy based on expertise also entails a defense of a type of elitism within a modern, democratic social order that values egalitarianism. The tension is reduced by valuing egalitarianism within the field of science, that is, by reserving it for those persons who have obtained the credentials to practice as a scientist.

2. *Variations in Scientific Culture(s)*

Going beyond the focus on modern science as a whole, much research on scientific culture has been devoted to its variations over time and across disciplines. Historians occasionally use the concept of periods as a way

of ordering cultural history, for example in the divisions of music history from baroque to classical to romantic. Although periodization is very approximate and relatively subjective, it nonetheless helps to point to some of the commonalities across scientific disciplines within the same time period, and some of the disjunctures in the history of science over time (Foucault 1970). Changes in conceptual frameworks and research programs across disciplines within a time period usually are part of broader cultural changes. For example, in the nineteenth century political values were frequently framed by grand narratives of progress (shared with the 'white man's burden' of colonialism among other ideologies), and likewise new scientific fields such as cosmology, thermodynamics, and evolutionary biology evidenced a concern with temporal issues. In the globalized information society of the twenty-first century, concerns with complex systems have become more salient, and scientists in many disciplines are developing research programs based on ideas of information processing and systemic complexity.

Science is to some extent universal in the sense that, for example, physicists throughout the world agree upon the basic laws and empirical findings of physics. However, there are also significant cultural variations in science across geographic regions. The variations are more salient in the humanities and social sciences, where distinctive theoretical frameworks often have a regional flavor. For example, one may speak of French psychoanalysis or Latin American dependency theory. The variations are also obvious for the institutional organization of scientific communities, laboratories, and university systems (Hess 1995).

Furthermore, variations in the institutional organization are sometimes associated with differing research traditions in the content of a field. For example, the salary and power structure of German universities in the first third of the twentieth century favored a more theoretical approach to genetics and a concern with evolution and development. In contrast, the relatively collegial governance structures of the American universities, together with their location in agricultural schools, contributed to the development of American genetics in a more specialized, empirical, and applied direction (Harwood 1993). Sometimes variations in institutional structures and conditions are also associated with differences in laboratory technologies, which in turn restrict the selection of research problems. For example, in Japan, funding conditions for physics have in part shaped the emergence of detector designs that differ substantially from those in the USA (Traweek 1988).

In addition to the temporal and national cultures of science, a third area of variation in scientific cultures involves disciplinary cultures. For example, both high-energy physics and molecular biology are considered 'big science,' but their disciplinary cultures are substantially different (Knorr-Cetina 1999). Regarding their intellectual cultures, in high-energy physics data are widely recognized as heavily interpretable, whereas in molecular biology the contact with the empirical world is much more hands-on and direct. Regarding the institutional cultures of the two fields, laboratories in high-energy physics are very large-scale, with the individual's work subordinated to that of the cooperative team, even to the point of near anonymity in publication. In contrast, in molecular biology the laboratories are smaller, even for the genome project, and competition among individual researchers remains salient, with frequent disputes over priority in a publication byline.

3. *Multiculturalism and Science*

A third approach to the topic of scientific culture involves the ongoing modernization of scientific cultures. Although modern science is increasingly international and cosmopolitan, its biases rooted in particular social addresses have constantly been revealed. Science has been and remains largely restricted to white men of the middle and upper classes in the developed Western countries (Harding 1998, Rossiter 1995, Schiebinger 1989). As scientific knowledge and institutions have spread across the globe and into historically excluded groups, they have sometimes undergone changes in response to perspectives that new groups bring to science. For example, in the case of primatology, humans and monkeys have coexisted for centuries in India. Consequently, it is perhaps not surprising that Indian primatologists developed new research programs based on monkey–human interactions that challenged the romanticism of Western primatologists' focus on natural habitats (Haraway 1989). More generally, as scientific disciplines become globalized, different national cultural traditions intersect with the globalized disciplinary cultures to reveal unseen biases and new possibilities for research.

As women and under-represented ethnic groups have achieved a place in scientific fields, they have also tended to reveal hidden biases and develop new possibilities. Although the reform movements of multiculturalism in science can be restricted to the institutional culture of science (such as reducing racism and sexism in hiring practices), they sometimes also extend to the intellectual culture or content of science. Theories and methods—particularly in the biological and social sciences—that seem transparently cosmopolitan and truthful to white, male colleagues often appear less so to the new groups. Whether it is the man-the-hunter theory of human evolutionary origins (Haraway 1989), the 'master' molecule theory of nucleus to cytoplasm relations (Keller 1985), or biological measures of racial inferiority (Harding 1993), scientific disciplines find themselves continually challenged to prove their universalism and to modify

concepts and theories in response to anomalies and new research.

In some cases the members of excluded groups do more than replace old research methods and programs with new ones; they also create new research fields based on their identity concerns. A prime example is the work of the African American scientist George Washington Carver. Although best known in American popular culture for finding many new uses for the peanut, Carver's research was embedded in a larger research program that was focused on developing agricultural alternatives to King Cotton for poor, rural, African American farmers (Hess 1995).

4. Conclusions

It is important not to think of the embeddedness of scientific cultures in broader cultural practices as a problem of contamination. The broader cultures of modern science provide a source of metaphors and institutional practices that both inspire new research and limit its possibilities. For example, if evolutionary theory could not be thought before the progressivist culture of the nineteenth century, it cannot help but to be rethought today. Not only have new research findings challenged old models, but the broader cultural currents of complex systems and limits to growth have also inspired new models and empirical research (DePew and Weber 1995). In turn, today's concepts and theories will be rethought tomorrow. The broader societal cultures are not weeds to be picked from the flower bed of scientific culture(s) but the soil that both nurtures and limits its growth, even as the soil itself is transformed by the growth that it supports.

See also: Academy and Society in the United States: Cultural Concerns; Cultural Psychology; Cultural Studies of Science; Culture in Development; Encyclopedias, Handbooks, and Dictionaries; Ethics Committees in Science: European Perspectives; History of Science; History of Science: Constructivist Perspectives; Scientific Academies in Asia

Bibliography

Ben-David J 1991 *Scientific Growth*. University of California Press, Berkeley, CA

Cohen F 1997 *The Scientific Revolution*. University of Chicago Press, Chicago

DePew D, Weber B 1995 *Darwinism Evolving*. MIT Press, Cambridge, MA

Foucault M 1970 *The Order of Things*. Vintage, New York

Habermas J 1989 *The Structural Transformation of the Public Sphere*. MIT Press, Cambridge, MA

Haraway D 1989 *Primate Visions*. Routledge, New York

Harding S (ed.) 1993 *The Racial Economy of Science*. Routledge, New York

Harding S (ed.) 1998 *Is Science Multicultural?* Indiana University Press, Bloomington, IN

Harwood J 1993 *Styles of Scientific Thought*. University of Chicago Press, Chicago

Hess D 1995 *Science and Technology in a Multicultural World*. Columbia University Press, New York

Hooykaas R 1990 Science and reformation. In: Cohen, I (ed.) *Puritanism and the Rise of Science*. Rutgers University Press, New Brunswick, NJ

Jacob M 1988 *The Cultural Meaning of the Scientific Revolution*. Knopf, New York

Kammen M 1987 *A Machine that Would Go of Itself*. Knopf, New York

Keller E 1985 *Reflections on Gender and Science*. Yale University Press, New Haven, CT

Knorr-Cetina K 1999 *Epistemic Cultures*. Harvard University Press, Cambridge, MA

Koyré A 1965 *Newtonian Studies*. Chapman and Hall, London

Merton R 1973 *Sociology of Science*. University of Chicago Press, Chicago

Mulkay M 1976 Norms and ideology in science. *Social Science Information* **15**: 637–56

Rossiter M 1995 *Women Scientists in America*. Johns Hopkins University Press, Baltimore, MD

Schiebinger L 1989 *The Mind has no Sex?* Harvard University Press, Cambridge, MA

Shapin S 1994 *A Social History of Truth*. University of Chicago Press, Chicago

Shapin S 1996 *The Scientific Revolution*. University of Chicago Press, Chicago

Traweek S 1988 *Beamtimes and Lifetimes*. Harvard University Press, Cambridge, MA

D. Hess

Scientific Disciplines, History of

The scientific discipline as the primary unit of internal differentiation of science is an invention of nineteenth century society. There exists a long semantic prehistory of *disciplina* as a term for the ordering of knowledge for the purposes of instruction in schools and universities. But only the nineteenth century established real disciplinary communication systems. Since then the discipline has functioned as a unit of structure formation in the social system of science, in systems of higher education, as a subject domain for teaching and learning in schools, and finally as the designation of occupational and professional roles. Although the processes of differentiation in science are going on all the time, the scientific discipline as a basic unit of structure formation is stabilized by these plural roles in different functional contexts in modern society.

1. Unit Divisions of Knowledge

Disciplina is derived from the Latin *discere* (learning), and it has often been used since late Antiquity and the early Middle Ages as one side of the distinction *disciplina* vs. *doctrina* (Marrou 1934). Both terms

meant ways of ordering knowledge for purposes of teaching and learning. Often they were used synonymously. In other usages *doctrina* is more intellectual and *disciplina* more pedagogical, more focused on methods of inculcating knowledge. A somewhat later development among the church fathers adds to *disciplina* implications such as admonition, correction, and even punishment for mistakes. This concurs with recent interpretations of discipline, especially in the wake of Michel Foucault, making use of the ambiguity of *discipline* as a term always pointing to knowledge and disciplinary power at the same time (cf. Hoskin in Messer-Davidow et al. 1993). Finally, there is the role differentiation of teaching and learning and the distinction *doctrina/disciplina* is obviously correlated with it (Swoboda 1979).

One can still find the same understandings of *doctrina* and *disciplina* in the literature of the eighteenth century. But what changed since the Renaissance is that these two terms no longer refer to very small particles of knowledge. They point instead to entire systems of knowledge (Ong 1958). This goes along with the ever more extensive use by early modern Europe of classifications of knowledge and encyclopedic compilations of knowledge in which disciplines function as unit divisions of knowledge. The background to this is the growth of knowledge related to developments such as the invention of printing, the intensified contacts with other world regions, economic growth and its correlates such as mining and building activities. But in these early modern developments there still dominates the *archival function of disciplines*. The discipline is a place where one deposits knowledge after having found it out, but it is not an active system for the production of knowledge.

2. *Disciplines as Communication Systems*

A first premise for the rise of disciplines as production and communication systems in science is the specialization of scientists and the role differentiation attendant on it (Stichweh 1984, 1992). Specialization is first of all an intellectual orientation. It depends on a decision to concentrate on a relatively small field of scientific activity, and, as is the case for any such decision, one needs a social context supporting it, that is, other persons taking the same decision. Such decisions are rare around 1750 when encyclopedic orientations dominated among professional and amateur scientists alike, but they gained in prominence in the last decades of the eighteenth century. Second, specialization as role differentiation points to the educational system, which is almost the only place in which such specialized roles can be institutionalized as occupational roles. From this results a close coupling of the emerging disciplinary structures in science and the role structures of institutions of higher education. This coupling is realized for the first time in the reformed German universities of the first half of the nineteenth century and afterwards quickly spreads from there to other countries. Third, role differentiation in institutions of higher education depends on conditions of organizational growth and organizational pluralization. There has to be a sufficient number of organizations which must be big enough for having differentiated roles and these organizations must be interrelated in an ongoing continuity of interactions.

The emergence of *communities of specialists* is a further relevant circumstance. In this respect the rise of disciplines is synonymous with the emergence of *scientific communities* theorized about since Thomas Kuhn (Kuhn 1970). Scientific communities rest on the intensification of interaction, shared expertise, a certain commonality of values, and the orientation of community members towards problem constellations constitutive of the respective discipline. Modern science is not based on the achievements of extraordinary individuals but on the epistemic force of disciplinary communities.

Scientific communities are *communication systems*. In this respect the emergence of the scientific discipline is equivalent to the invention of new communication forms specific of disciplinary communities. First of all one may think here of new forms of *scientific publications*. In the eighteenth century a wide spectrum of publication forms existed; they were not, however, specialized in any way. There were instructional handbooks at the university level, journals of a general scientific nature for a regional public interested in utility, and academy journals aiming at an international public, each covering a wide subject area but with rather limited communicative effects. It was only after 1780 that in France, in Germany, and finally, in England, nationwide journals with a specific orientation on such subjects as chemistry, physics, mineralogy, and philology appeared. In contrast to isolated precursors in previous decades, these journals were able to exist for longer periods exactly because they brought together a *community of authors*. These authors accepted the specialization chosen by the journal; but at the same time they continually modified this specialization by the cumulative effect of their published articles. Thus the status of the scientific publication changed. It now represented the only communicative form by which, at the macrolevel of the system of science—defined originally by national but later by supranational networks—communication complexes specialized along disciplinary lines could be bound together and persist in the long run (Stichweh 1984, Chap. 6, Bazerman 1988).

At the same time the scientific publication became a formal principle interfering in every scientific production process. Increasingly restrictive conditions were defined regarding what type of communication was acceptable for publication. These conditions

included the requirement of identifying the problem tackled in the article, the sequential presentation of the argument, a description of the methods used, presentation of empirical evidence, restrictions on the complexity of the argument accepted within an individual publication, linkage with earlier communications by other scientists—using citations and other techniques—and the admissibility of presenting speculative thoughts. In a kind of feedback loop, publications, as the ultimate form of scientific communication, exercised pressure on the scientific production process (i.e., on research) and were thereby able to integrate disciplines as social systems.

This reorganization of the *scientific production process* adheres to one new imperative: the *search for novelties*. The history of early modern Europe was already characterized by a slow shift in the accompanying semantics associated with scientific truth, from an imperative to preserve the truth to an interest in the novelty of an invention. The success achieved in organizing traditional knowledge, as well as tendencies towards empirical methods and increased use of scientific instruments, worked toward this end. In this dimension, a further discontinuity can be observed in the genesis of the term *research* in the years after 1790. In early modern times the transition from the preservation to the enlargement of knowledge could only be perceived as a continual process. In contrast, *research* from about 1800 refers to a fundamental, and at any time realizable, questioning of the entire body of knowledge until then considered as true. Competent scientific communication then had to be based on research in this sense. What was communicated might be a small particle of knowledge, as long as it was a new particle of knowledge. Scientific disciplines then became research disciplines based on the incessant production of novelties.

The link between scientific disciplines and organizations of higher education is mediated by two more organizational structures. The first of these are *disciplinary careers*. Specialized scientists as members of disciplinary communities do not need only specialized occupational roles. Additionally there may be a need for careers in terms of these specialized roles. This again is a condition which sharply distinguishes eighteenth from nineteenth century universities. Around 1750 you still find, even in German universities, hierarchical career patterns which implied that there was a hierarchical succession of chairs inside of faculties and a hierarchical sequence of faculties by which a university career was defined as a progression of steps through these hierarchized chairs. One could, for example, rise from a chair in the philosophical faculty to an (intellectually unrelated) chair in the medical faculty. The reorganization of universities since early nineteenth century completely discontinued this pattern. Instead of a succession of chairs in one and the same university, a scientific career meant a progression through positions inside a discipline, which normally demands a career migration between universities. This presupposes intensified interactions and competitive relations among universities which compete for qualified personnel and quickly take up new specializations introduced elsewhere. In Germany such regularized career paths through the national university system were especially to be observed from around 1850.

This pattern is again closely related to *disciplinary curricula*, meaning that one follows one's disciplinary agenda not only in one's research practice and personal career, but furthermore that there exist institutional structures favoring teaching along lines close to current disciplinary core developments. The unity of teaching and research is one famous formula for this, but this formula does not yet prescribe disciplinary curricular structures which would demand that there should be a complete organization of academic studies close to the current intellectual problem situation and systematics of a scientific discipline. Only if this is the case does there arise a *professionalization of a scientific discipline*, which means that a systematic organization of academic studies prepares for a non-academic occupational role which is close to the knowledge system of the discipline. Besides professionalization there is then the effect that the discipline educates its own future research practitioners in terms of the methods and theories constitutive of the discipline. A discipline doing this is not only closed on the level of the disciplinary communication processes, it is also closed on the level of socialization practices and the attendant recruitment of future practitioners (on the operational closure of modern science see Luhmann 1990, Stichweh 1990).

3. *The Modern System of Scientific Disciplines*

It is not sufficient to analyze disciplines as individual knowledge producing systems. One has to take into account that the invention of the scientific discipline brings about first a limited number, then many scientific disciplines which interact with one another. Therefore it makes sense to speak of a modern system of scientific disciplines (Parsons 1977, p. 300ff., Stichweh 1984) which is one of the truly innovative social structures of the modern world.

First of all, the modern system of scientific disciplines defines an *internal environment* (*milieu interne* in the sense of Claude Bernard) for any scientific activity whatsoever. Whatever goes on in fields such as physics, sociology, or neurophysiology, there exists an internal environment of other scientific disciplines which compete with that discipline, somehow comment on it, and offer ideas, methods, and concepts. There is *normal science* in a Kuhnian sense, always involved with problems to which solutions seem to be at hand in the disciplinary tradition itself; but normal science is

always commented upon by a parallel level of interdisciplinary science which arises from the conflicts, provocations and stimulations generated by other disciplines and their intellectual careers.

In this first approximation it is already to be seen that the modern system of scientific disciplines is a very dynamic system in which the dynamism results from the intensification of the interactions between ever more disciplines. Dynamism implies, among other things, ever changing disciplinary boundaries. It is exactly the close coupling of a *cognitively defined discipline* and a *disciplinary community* which motivates this community to try an expansionary strategy in which the discipline attacks and takes over parts of the domain of other disciplines (Westman 1980, pp. 105–6). This was wholly different in the disciplinary order of early modern Europe, in which a classificatory generation of disciplinary boundaries meant that the attribution of problem domains to disciplines was invariable. If one decided to do some work in another domain, one had to accept that a change over to another discipline would be necessary to do this.

Closely coupled to this internally generated and self-reinforcing dynamics of the modern system of scientific disciplines is the openness of this system to new disciplines. Here again arises a sharp difference to early modern circumstances. In early modern Europe there existed a closed and finite catalogue of scientific disciplines (Hoskin 1993, p. 274) which was related to a hierarchical order of these disciplines (for example philosophy was a higher form of knowledge than history, and philosophy was in its turn subordinated to faculty studies such as law and theology). In modern society no such limit to the number of disciplines can be valid. New disciplines incessantly arise, some old ones even disappear or become inactive as communication systems. There is no center and no hierarchy to this system of the sciences. Nothing allows us to say that philosophy is more important than natural history or physics more scientific than geology. Of course, there are asymmetries in influence processes between disciplines, but no permanent or stable hierarchy can be derived from this.

The modern system of scientific disciplines is a global system. This makes a relevant difference from the situation of the early nineteenth century, in which the rise of the scientific discipline seemed to go along with a strengthening of national communities of science (Crawford et al. 1993, Stichweh 1996). This nationalization effect, which may have had to do with a meaningful restriction of communicative space in newly constituted communities, has since proved to be only a temporary phenomenon, and the ongoing dynamics of (sub-) disciplinary differentiation in science seems to be the main reason why national communication contexts are no longer sufficient infrastructures for a rapidly growing number of disciplines and subdisciplines.

4. *The Future of the Scientific Discipline*

The preponderance of subdisciplinary differentiation in the late twentieth century is the reason most often cited for the presumed demise of scientific discipline postulated by a number of authors. But one may object to this hypothesis on the ground that a change from disciplinary to subdisciplinary differentiation processes does not at all affect the drivers of internal differentiation in modern science: the relevance of an internal environment as decisive stimulus for scientific variations, the openness of the system to disciplinary innovations, the nonhierarchical structure of the system. Even if one points to an increasing importance of interdisciplinary ventures (and to problem-driven interdisciplinary research) which one should expect as a consequence of the argument on the internal environment of science, this does not change the fact that disciplines and subdisciplines function as the form of consolidating interdisciplinary innovations. And, finally, there are the interrelations with the external environments of science (economic, political, etc.), which in twentieth and twenty-first century society are plural environments based on the principle of functional differentiation. Systems in the external environment of science are dependent on sufficiently stable addresses in science if they want to articulate their needs for inputs from science. This is true for the educational environment of science which has to organize school and higher education curricula in disciplinary or interdisciplinary terms, for role structures as occupational structures in the economic environment of science, and for many other demands for scientific expertise and research knowledge which always must be able to specify the subsystem in science from which the respective expertise may be legitimately expected. These interrelations based on structures of internal differentiation in science which have to be identifiable for outside observers are one of the core components of modern society which, since the second half of the twentieth century, is often described as *knowledge society*.

See also: Disciplines, History of, in the Social Sciences; History and the Social Sciences; History of Science: Constructivist Perspectives; Human Sciences: History and Sociology; Knowledge Societies; Scientific Academies, History of; Scientific Culture; Scientific Revolution: History and Sociology; Teaching Social Sciences: Cultural Concerns; Universities and Science and Technology: Europe; Universities and Science and Technology: United States; Universities, in the History of the Social Sciences

Bibliography

Bazerman C 1988 *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. University of Wisconsin Press, Madison, WI

Crawford E, Shinn T, Sörlin S 1993 *Denationalizing Science. The Contexts of International Scientific Practice*. Kluwer, Dordrecht, The Netherlands
Hoskin K W 1993 Education and the genesis of disciplinarity: The unexpected reversal. In: Messer-Davidow E, Shumway D R, Sylvan D J (eds.) *Knowledges: Historical and Critical Studies in Disciplinarity*. University Press of Virginia, Charlottesville, VA, pp. 271–305
Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago
Luhmann N 1990 *Die Wissenschaft der Gesellschaft*. Suhrkamp, Frankfurt am Main, Germany
Marrou H I 1934 'Doctrina' et 'Disciplina' dans la langue des pères de l'église. *Archivus Latinitatis Medii Aevi* **9**: 5–25
Messer-Davidow E, Shumway D R, Sylvan D J (eds.) 1993 *Knowledges: Historical and Critical Studies in Disciplinarity*. University Press of Virginia, Charlottesville, VA
Ong W J 1958 *Ramus, Method, and the Decay of Dialogue: From the Art of Discourse to the Art of Reason*. Harvard University Press, Cambridge, MA
Parsons T 1977 *Social Systems and the Evolution of Action Theory*. Free Press, New York
Stichweh R 1984 *Zur Entstehung des Modernen Systems Wissenschaftlicher Disziplinen—Physik in Deutschland 1740–1890*. Suhrkamp, Frankfurt am Main, Germany
Stichweh R 1990 Self-organization and autopoiesis in the development of modern science. In: Krohn W, Küppers G, Nowotny H (eds.) *Selforganization—Portrait of a Scientific Revolution. Sociology of the Sciences*, Vol. XIV. Kluwer Academic Publishers, Boston, pp. 195–207
Stichweh R 1992 The sociology of scientific disciplines: On the genesis and stability of the disciplinary structure of modern science. *Science in Context* **5**: 3–15
Stichweh R 1996 Science in the system of world society. *Social Science Information* **35**: 327–40
Swoboda W W 1979 Disciplines and interdisciplinarity: A historical perspective. In: Kockelmans J J (ed.) *Interdisciplinarity and Higher Education*. University Park, London, pp. 49–92
Westman R S 1980 The astronomers's role in the sixteenth century: A preliminary study. *History of Science* **18**: 105–47

R. Stichweh

Scientific Discovery, Computational Models of

Scientific discovery is the process by which novel, empirically valid, general, and rational knowledge about phenomena is created. It is, arguably, the pinnacle of human creative endeavors. Many academic and popular accounts of great discoveries surround the process with mystery, ascribing them to a combination of serendipity and the special talents of geniuses. Work in Artificial Intelligence on computational models of scientific reasoning since the 1970s shows that such accounts of the process of science are largely mythical. Computational models of scientific discovery are computer programs that make discoveries in particular scientific domains. Many of these systems model discoveries from the history of science or simulate the behavior of participants solving scientific problems in the psychology laboratory. Other systems attempt to make genuinely novel discoveries in particular scientific domains. Some have produced new findings of sufficient worth that the discoveries have been published in mainstream scientific journals. The success of these models provides some insights into the nature of human cognitive processes in scientific discovery and addresses some interesting issues about the nature of scientific discovery itself (see *Scientific Reasoning and Discovery, Cognitive Psychology of*).

1. Computational Models of Scientific Discovery

Most computational models of discovery can be conceptualized as performing a recursive search of a space of possible states, or expressions, defined by the representation of the problem. Procedures are used to search the space of legal states by manipulating the expressions and using tests of when the goal or subgoals have been met. To manage the search, which is typically subject to potential combinatorial explosion, heuristics are used to guide the selection of appropriate operators. This is essentially an application of the theory of human problem solving as heuristic search within a symbol processing system (Newell and Simon 1972).

For example, consider BACON (Langley et al. 1987) an early discovery program which finds algebraic formulas as parsimonious descriptions of quantitative data. States in the problem search space of BACON include simple algebraic formulas; such as P/D or P^2/D, where, for instance, P is the period of revolution of planets around the sun and D is their distance from the sun. Tests in BACON attempt to find how closely potential expressions match the given quantitative data. Given quantitative data for the planets of the solar system, one step in BACON's discovery path finds that neither P^2/D nor P/D are constant and that the first expression is monotonically increasing with respect to the second. Given this relation between the expressions BACON applies its INCREASING operator to give the product of the terms, i.e., P^3/D^2. This time the test of whether the expression is constant, within a given margin of error, is true. $P^3/D^2 =$ constant is one of Kepler's planetary motion laws. For more complex cases with larger numbers of variables, BACON uses discovery heuristics based on notions of symmetry and the conservation of higher order terms to pare down the search space. The heuristics use the underlying regularities within the domain to obviate the need to explore parts of the search space that are structurally similar to previously explored states.

Following such an approach, computational models have been developed to perform tasks spanning a full

spectrum of theoretical activities including the formation of taxonomies, discovering qualitative and quantitative laws, creation of structural models and the development of process models (Langley et al 1987, Shrager and Langley 1990, Cheng 1992). The range of scientific domains covered is also impressive, ranging from physics and astronomy, to chemistry and metallurgy, to biology, medicine, and genetics. Some systems have produced findings that are sufficiently novel to be worthy of publication in major journals of the relevant discipline (Valdes-Perez 1995).

2. Scope of the Models

Computational models of scientific discovery have almost exclusively addressed theory formation tasks. However, the modeling of experiments has not been completely neglected as models have been built that design experiments, to a limited extent, by using procedures to specify what properties should be manipulated and measured in an experiment and the range of magnitudes over which the properties should be varied (Kulkarni and Simon 1988, Cheng 1992). For these systems, the actual experimental results are either provided by the user of the system or generated by a simulated experiment in the software. Discovery systems have also been directly connected to a robot which manipulates a simple experimental setup so that data collected from the instruments can be fed to the system directly, so eliminating any human intervention (Huang and Zytkow 1997). Nevertheless, few systems have simulated or supported substantial experimental activities, such as observing or creating new phenomena, designing experiments, inventing new experimental apparatus, developing new experimental paradigms, establishing the reliability of experiments, or turning raw data into evidence. This perhaps reflects a fundamental difference between the theoretical and experimental sides of science. While both clearly involve abstract conceptual entities, experimentation is also grounded in the construction and manipulation of physical apparatus, which involves a mixture of sophisticated perceptual abilities and motor skills. Developing models of discovery that include such capabilities would necessarily require other areas within AI beyond problem solving, such as image processing and robotics.

The majority of discovery systems model a single theory formation task. The predominance of such systems might be taken as the basis for a general criticism of computational scientific discovery. The models are typically poor imitations of the diversity of activities in which human scientists are engaged and, perhaps, it is from this variety that scientific creativity arises. Researchers in this area counter such arguments by claiming that the success of such single task systems is a manifestation of the underlying nature of the process of discovery, that it is composed of subprocesses or tasks that are relatively autonomous. More complex activity can be modeled by assembling systems that perform one task into a larger system, with the inputs to a particular component subsystem being the outputs of other systems. The handful of models that do perform multiple tasks demonstrate the plausibility of this claim (e.g., Kulkarni and Simon 1988, Cheng 1992). The organization of knowledge structures and procedures in those systems exploits the hierarchical decomposition of the overall process into tasks and subtasks.

This in turn raises the general question about the number and variety of different tasks that constitute scientific discovery and the nature of their interactions. What distinct search spaces are involved and how is information shared among them? Computational models of scientific discovery provide some insight into this issue. At a general level, many models can be characterized in terms of two spaces, one for potential hypotheses and the other a space of instances or sets of data (Simon and Lea 1974). Scientific discovery is then viewed as the search of each space mutually constrained by the search of the other. Inferring a hypothesis dictates the form of the data needed to test the hypothesis, while the data itself will determine whether the hypothesis is correct and suggest in what ways it should be amended (Kulkarni and Simon 1988). This is an image of scientific discovery that places equal importance on theory and experiment, portraying the overall process as a dynamic interaction between both components. This approach is applicable both to disciplines in which individual scientists do the theorizing and experimenting and to disciplines in which these activities are distributed among different individuals or research groups. The search of the theoretical and experimental spaces can be further decomposed into additional subspaces; for example, Cheng (1992) suggests three subspaces for hypotheses, models, and cases under the theory component, and spaces of experimental classes, setups, and tests under the experimental component.

3. Developing Computational Models of Discovery

One major advantage of building computational models over other approaches to the study of scientific discovery is the precision that is imposed by writing a running computer program. Ambiguities and inconsistencies in the concepts used to describe discovery processes become apparent when attempting to encode them in a programming language. Another advantage of modeling is the ability to investigate alternative methods or hypothetical situations. Different versions of a system may be constructed embodying, say, competing representations to investigate the difficulty

of making the discovery with the alternatives. The same system can be run with different sets of data, for example, to explore whether a discovery could have been made had there been less data, or had different data been available.

Many stages are involved in the development of the models, including: formulation of the problem, engineering appropriate problem representations, selecting and organizing data, design and redesign of the algorithm, actual invocation of the algorithm, and filtering and interpretation of the findings of the system (Langley 1998). Considering the nature and relative importance of these activities in the development of systems provides further insight into the nature of scientific discovery. In particular, the design of the representation appears to be especially critical to the success of the systems. This implies that generally in scientific discovery finding an effective representation may be fundamental to the making of discoveries. This issue has been directly addressed by computational models that contrast the efficacy of different representations for modeling the same historical episode (Cheng 1996). Consistent with work in cognitive science, diagrammatic representations may in some cases be preferable to informationally equivalent propositional representations. Although computational models argue against any special abilities of great scientists beyond the scope of conventional theories of problem solving, the models suggest that the ability of some scientists to modify or create new representations may be an explanation, at least in part, of why they were the ones to succeed.

4. Conclusions and Future Directions

Given the extent of the development work necessary on a discovery system, it seems appropriate to attribute discoveries as much to the developer as to the system itself, although without the system many of the novel discoveries would not have been possible. This does not imply that machine discovery is impossible, but that care must be taken in delimiting the capabilities of discovery systems. Further, the ability of the KEKEDA system (Kulkarni and Simon 1988) to change its goals to investigate any surprising phenomenon it discovers suggests that systems can be developed that would filter and interpret the output of existing systems, by constraining the search of the space defined by the outputs of those systems using metrics based on notions of novelty. Developing such a system, or other systems that find problems or that select appropriate representations, will require the system to possess a substantially more extensive knowledge of the target domain. Such knowledge based systems are costly and time consuming to build, so it appears that the future of discovery systems will be more as collaborative support systems for domain scientists rather than fully autonomous systems (Valdes-Perez 1995). Such systems will exploit the respective strengths of domain experts and the computational power of the models to compensate for each others' limitations.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence in Cognitive Science; Artificial Intelligence: Search; Deductive Reasoning Systems; Discovery Learning, Cognitive Psychology of; Intelligence: History of the Concept; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning; Scientific Reasoning and Discovery, Cognitive Psychology of

Bibliography

Cheng P C-H 1992 Approaches, models and issues in computational scientific discovery. In: Keane M T, Gilhooly K (eds.) *Advances in the Psychology of Thinking*. Harvester-Wheatsheaf, Hemel Hempstead, UK, pp. 203–236

Cheng P C-H 1996 Scientific discovery with law encoding diagrams. *Creativity Research Journal* **9**(2&3): 145–162

Huang K-M, Zytkow J 1997 Discovering empirical equations from robot-collected data. In: Ras Z, Skowron A (eds.) *Foundations of Intelligent Systems*. Springer, Berlin

Kulkarni D, Simon H A 1988 The processes of scientific discovery: The strategy of experimentation. *Cognitive Science*, **12**: 139–75

Langley P 1998 The computer-aided discovery of scientific knowledge. In: *Proceedings of the First International Conference on Discovery Science*. Springer, Berlin

Langley P, Simon H A, Bradshaw G L, Zytkow J M 1987 *Scientific Discovery: Computation Explorations of the Creative Processes*. MIT Press, Cambridge, MA

Newell A, Simon H A 1972 *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ

Shrager J, Langley P (eds.) 1990 *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufmann, San Mateo, CA

Simon H A, Lea G 1974 Problem solving and rule induction: a unified view. In: Gregg L W (ed.) *Knowledge and Cognition*. Lawrence Erlbaum, Potomac, MD, pp. 105–127

Valdes-Perez R E 1995 Some recent human/computer discoveries in science and what accounts for them. *AI Magazine* **16**(3): 37–44

P. C.-H. Cheng

Scientific Evidence: Legal Aspects

Expertise, scientific and otherwise, has been part of the legal landscape for centuries (Hand 1901). Over the last decades of the twentieth century the role of scientific evidence in the law has expanded rapidly, in both regulatory settings and in litigation. Statutes and

treaties routinely require agencies to provide scientific justifications for regulatory decisions within and among nations. National and international policy disputes are increasingly fought out within the risk assessment rhetoric of science (Craynor 1993).

Courts are another large consumer of science. In both criminal and civil cases, parties believe scientific testimony will make their case stronger. As science's role has grown, so has interest in the relationship between law and science. This essay reviews the present state of knowledge concerning the science–law relationship and mentions several areas ripe for investigation.

1. *What Law Wants from Science*

In both the administrative and the courtroom context, law is most frequently interested in acquiring scientific answers to practical questions of causation (e.g., does exposure to airborne asbestos cause lung cancer) and measurement (e.g., what is the alcohol level in a driver's blood). Both of these questions entail questions as to whether specific techniques (e.g., a breathalyzer) are capable of producing reliable measurements. Theoretical questions *per se* are frequently of secondary interest, useful insofar as they help courts and regulators to chose among conflicting measurements, extrapolations, or causal assertions.

Courts and agencies are both interested in questions of general and specific causation. Agencies are interested in the effects of nitrous oxide on the environment and the specific level of emissions from a particular plant. Courts may be interested in whether a given level of airborne asbestos causes lung cancer. And they may need to determine if a particular plaintiff's lung cancer was caused by asbestos exposure. It is often much more difficult to achieve a scientifically rigorous answer to this latter type of question.

An important difference between courts and agencies is the level of proof necessary to reach a conclusion. In administrative settings agencies may prevail if they can show that a substance poses a danger or a risk. In a courtroom however, in order to prevail, the plaintiff will be required to show both general and specific causation. Scientific evidence that is sufficient in the regulatory context may be considered insufficient in the court context. For example, animal studies showing a relationship between saccharin consumption and cancer may be sufficient to cause an agency to impose regulations on human exposure to the sweetener, but would not be sufficient to permit a groups of plaintiffs to prevail on a claim that they were actually injured by consuming saccharin in diet soft drinks, or even on the general causation claim that saccharin at human dose levels causes cancer in any humans.

2. *Legal Control of Scientific Evidence*

One of the more interesting aspects of the use of science in courts is the legal effort to control the terms of the law–science interaction. Judicial control of scientific experts is shaped by whether the court is operating in an inquisitorial or an adversarial system (van Kampen 1998). In inquisitorial systems, e.g., Belgium, Germany, France, Japan, the judge plays a large role in the production of evidence. Experts are almost always court appointed and are asked to submit written reports. Parties may be given the opportunity to object to a particular expert, question the expert about the opinion rendered, or hire their own expert to rebut the court-appointed expert; but it is very difficult to attack the admissibility of an expert's report successfully (Langbein 1985, Jones 1994). In adversarial systems the parties usually have far greater control over the production of evidence, including the selection and preparation of expert witnesses. Court-appointed experts are rare (Cecil and Willging 1994).

A similar pattern can be observed in the regulatory arena. Jasanoff (1991) compares the relatively closed, consensual, and non-litigious British regulatory approach to the open, adversarial and adjudicatory style found in the United States. Clearly, legal organization and legal culture shape the way in which legal systems incorporate science. In turn, legal organization is related to larger cross-cultural differences. For example, open, adversarial approaches are more pronounced in societies that are more individualistic and lower on measures of power distance (Hofstede 1980).

Judicial control of scientific expert testimony has become a high profile issue in the United States. Critics complain that the combination of adversarial processes and the use of juries in both criminal and civil trials encourages the parties to introduce bad science into the trial process. They argue, with some empirical support, that the use of experts chosen by the parties produces individuals who by their own account stray relatively further from Merton's four norms of science—universalism, communism, disinterestedness, and organized skepticism (Merton 1973). They submit that jurors are too easily swayed by anyone who is called a scientist regardless of the quality of the science supporting the expert's position. (See *Expert Witness and the Legal System*: *Psychological Aspects*.) Many now call upon judges to play a greater role in selecting and supervising experts, a role closer to that found in inquisitorial systems (Angell 1996).

The preference for an inquisitorial style makes several assumptions about the nature of science and the proper relationship between law and science. Similar assumptions underlie the admissibility rules employed by courts. The history of admissibility decisions in American courts provides a useful way to explore these assumptions.

Precisely because the knowledge possessed by the scientist is beyond the ken of the court, the judge (or

the jury) often is not in a good position to determine if what the expert is offering is helpful. In a 1923 decision a federal court offered one solution to this problem; acquiesce to the judgment of the community of elites who are the guardians of each specialized area of knowledge. Under the so-called *Frye* rule experts may testify only if the subject of their testimony had reached 'general acceptance' in the relevant scientific community. In 1993 in *Daubert* vs. *Merrell Dow Pharmaceuticals*, *Inc*., and subsequent opinions, the United States Supreme Court moved away from the *Frye* rule. In its place federal courts have developed a nonexclusive multifactor test designed to assess the validity of the science underlying the expert's testimony. Factors the courts have mentioned include: whether the theory or technique underlying the expert's testimony is falsifiable and had been tested; the error rate of any techniques employed; whether the expert's research was developed independent of litigation; whether the testifying expert exercised the same level of intellectual rigor that is routinely practiced by experts in the relevant field; whether the subject matter of the testimony had been subjected to peer review and publication in refereed journals; and, in a partial retention of the *Frye* test, whether the theory or technique had achieved general acceptance in the relevant scientific community.

The *Daubert*-inspired test offers an alternative to acquiescence, a do-it-yourself effort on the part of the courts. A number of the factors are consistent with an adversarial legal system that institutionalizes mistrust of all claims to superior authority. The adversary system itself is consistent with the values of a low-power distance culture that affords relatively less legitimacy to elite authoritative opinion and, therefore, is skeptical that scientists are entitled to a privileged language of discourse from which others are excluded. It is also consistent with a view of science strongly influenced by other forces in society. These are perspectives on the scientific enterprise that are associated with those who adopt a social constructionist view of science (Shapin 1996, Pickering 1992).

3. *Legal Understanding of the Scientific Enterprise*

Daubert, however, offers anything but a social constructionist test if by this we mean that scientific conclusions are solely the result of social processes within the scientific community. At its core, the opinion requires of judges that they become sufficiently knowledgeable about scientific methods so that they can fairly assess the validity of evidence offered at trial. This requirement that scientific testimony must pass methodological muster reflects a positivist approach that is slanted toward a Baconian view of science. The opinion cites with favor a Popperian view of how to distinguish the scientific enterprise from other forms of knowledge (Popper 1968). In this regard, the opinion is not unique.

In their use of scientific evidence, both courts and administrative agencies seem to distinguish the process of science from its products. They accept the constructionist insight that the process of doing science is a social enterprise and is subject to the buffeting, often distorting winds of social, political, economic, and legal influences. At the same time, courts, agencies, and legislatures cling to a realist belief that the products of science may state a truth about the world, or at least something so similar to truth as it is commonly understood at a given point in history that the particular discipline of law does not need to concern itself with the difference. The legal system's view of science adopts a strong version of what Cole (1992, p.x) calls a realist–constructivist position, i.e., science is socially constructed both in the laboratory and in the wider community, but the construction is constrained by input from the empirical world. It rejects what he calls a relativist–constructionist position that claims nature has little or no influence on the cognitive content of science.

The focus on methods is a search for some assurance that the expert has given the empirical world a reasonable opportunity to influence and constrain the expert's conclusions. Ultimately, the law's epistemology with respect to science holds that there are a set of (social) practices often given the shorthand name 'the scientific method' which increase the likelihood that someone will make positive contributions to knowledge; a set of practices that scientists themselves frequently point to as the sources of past scientific success (Goldman 1999).

There is a large dose of pragmatism in all of this, of course, and the *Daubert* rule itself has been cited as an example of 'the common law's genius for muddling through on the basis of experience rather than logic' (Jasanoff 1995, p. 63). Not surprisingly, some have criticized the courts for failing to adopt a philosophically coherent admissibility rule (Schwartz 1997). The court's admissibility rulings do seem to have proceeded in happy obliviousness to the 'science wars' that arguably began with Fleck (1979), flourished with Kuhn (1962) and raged for much of the last half of the twentieth century between the defenders of a more traditional, positivist view of science and those critics who emphasize its historical, political, social, and rhetorical aspects (Leplin 1997, Latour 1999). The same could be said of administrative use of science.

The rejection of relativist views of science does not mean that all legal actors hold identical views. It would be valuable to map the beliefs of legal actors on central disputes in the science wars, and how these beliefs impact their use of science. For example, if, as seems likely, plaintiff personal injury lawyers in the United States hold a more relativist view of science, how, if at all, does this affect their selection and

preparation of experts and, in turn, how are these experts received in courts?

4. *Law–Science Interdependence*

Law's approach to science should not be understood in terms of science alone but rather in terms of the law–science interaction. There are several dimensions to this relationship. First, although modern legal systems may recognize that scientists are influenced by the social world around them, permitting radical deconstruction that would undermine science's claim to special status is difficult to imagine (Fuchs and Ward 1994). The modern state is increasingly dependent upon science as a source of legitimacy. By turning to science for solutions to complex environmental and safety issues, legislatures are able to avoid making difficult political choices while giving the appearance of placing decision-making in the hands of apparently neutral experts who are held in very high esteem relative to other elites (Lawler 1996). The advantages of this approach are so great that agencies frequently engage in a 'science charade' in which they wrap their decision in the language of science even when there is very little research supporting a regulation (Wagner 1995).

Second, and related to the first observation, the law–science interaction is one of the important ways in which the state helps to produce and maintain science's dominant position in modern Western society. In both its administrative and judicial actions, the law defines what is and what is not scientific knowledge, and thereby assists science in important boundary maintenance work of excluding many from the community of 'scientists.' For example, in the case of *United States* vs. *Starzecpyzel*, the court permitted the state's handwriting experts to testify on the question of whether the defendants had forged documents, but only if they did not refer to themselves as 'scientists.' Moreover, legal decisions contain an implicit epistemology that reinstitutionalizes one view of the nature of scientific knowledge. When courts and other institutions of the state reject a relativist view of science that argues the empirical world has little, if any, influence on what is accepted as true by the scientific community, they help to marginalize anyone who adopts this position. There can be little doubt that the view is given little or no attention in law. A cursory search for the names of individuals most associated with these debates in American federal cases finds over 100 references to Popper alone, but no more than one or two passing references to the more prominent critics of traditional understandings of the scientific enterprise.

Third, the exact contours of the law–science interaction are shaped by a society's legal structure and its legal culture. For example, the American legal system's realist–constructivist understanding of science fits neatly with its own normative commitment to both 'truth' and 'justice' as legitimate dispute resolution goals. Ideally, cases should be correctly decided, should arrive at the truth. A realist view of science is consistent with the idea that a court may arrive at the correct outcome. But the truth is contested, and the courts should also give litigants the sense that they were listened to; that they received procedural justice (Tyler 1990). A constructivist understanding of the scientific enterprise legitimates the right of each party to find an expert who will present their view of the truth. We may hypothesize that all legal systems tend toward an epistemology of science and the scientific enterprise that fit comfortably within their dominant methods of law-making and dispute settlement. We might expect, therefore, that inquisitorial systems would be more skeptical of a constructivist view of the scientific enterprise than is the case in adversarial legal systems.

5. *Effects of Growing Interdependence*

As this essay attests, science's impact on legal processes grows apace. In many areas, an attorney unarmed with scientific expertise operates at a significant disadvantage. A growing number of treatises attempt to inform lawyers and judges of the state of scientific knowledge in various areas (Faigman et al. 1997). At the level of the individual case, there is some evidence that scientific norms are altering the way scientifically complex lawsuits are tried. Restrictive admissibility rules are one of several ways that courts in the United States may restrict traditional adversary processes when confronted with cases involving complex scientific questions.

What of law's effects on science? If the supply of science on some issue is affected by demand, legal controversy should draw scientists to research topics they might otherwise have eschewed. There is evidence that this does occur. The subject matter of scientific research is shaped by legal controversy in a wide number of areas, from medical devices (Sobol 1991) to herbicides (Schuck 1986). Some have argued that law's interest has an additional effect of causing the production of 'worse' science (Huber 1991), but it might also be argued that in areas such as DNA testing, law's interest has produced better science and better technology than might otherwise have existed. On the other hand, the threat of legal controversy may impede research into some areas, such as the development of new contraceptive devices.

Even if we believe all science is produced through a process of social construction, it is also the case that most scientists believe that their work is constrained by the empirical world (Segerstråle 1993). Moreover, they often attempt to surround themselves with a social structure that keeps society at a distance. Scientific groups are increasingly attempting to propa-

gate ethical standards for those who offer expert testimony in courts, reflecting the widespread belief that scientists who typically appear in legal arenas differ from their colleagues on these dimensions. Finally, legal use of science affects the closure of scientific debates (Sanders 1998). On the one hand, law may perpetuate controversy by supplying resources to both sides. On the other hand, it may help to bring a dispute to closure by authoritatively declaring one side to be correct.

See also: Expert Systems in Cognitive Science; Expert Testimony; Expert Witness and the Legal System: Psychological Aspects; Parties: Litigants and Claimants; Science and Law; Science and Technology Studies: Experts and Expertise; Statistics as Legal Evidence

Bibliography

Angell M 1996 *Science on Trial: The Clash of Medical Evidence and the Law in the Breast Implant Case*. Norton, New York

Cecil J S, Willging T E 1994 Court appointed experts. In *Federal Judicial Center Reference Manual on Scientific Evidence*. West, St. Paul, MN

Cole S 1992 *Making Science: Between Nature and Society*. Harvard University Press, Cambridge, MA

Craynor C F 1993 *Regulating Toxic Substances: A Philosophy of Science and the Law*. Oxford University Press, New York

Daubert vs. *Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993)

Faigman D L, Kaye D, Saks M, Sanders J 1997 *Modern Scientific Evidence: The Law and Science of Expert Testimony*. West Group, St. Paul, MN

Fleck L 1979 *Genesis and Development of a Scientific Fact*. University of Chicago Press, Chicago; IL

Fuchs S, Ward S 1994 What is deconstruction, and where and when does it take place? Making facts in science, building cases in law. *American Sociological Review* **59**: 481–500

Goldman A I 1999 *Knowledge in a Social World*. Clarendon Press, Oxford UK

Hand L, Learned 1901 Historical and practical considerations regarding expert testimony. *Harvard Law Review* **15**: 40

Hofstede G H 1980 *Culture's Consequences: International Differences in Work-related Values*. Sage, Beverly Hills, CA

Huber P W 1991 *Galileo's Revenge: Junk Science in the Courtroom*. Basic Books, New York

Jasanoff S 1991 Acceptable evidence in a pluralistic society. In: Mayo G D, Hollander R D (eds.) *Acceptable Evidence: Science and Values in Risk Management*. Oxford University Press, Oxford; UK

Jasanoff S 1995 *Science at the Bar: Law, Science, and Technology in America*. Harvard University Press, Cambridge, MA

Jones C A G 1994 *Expert Witnesses: Science, Medicine, and the Practice of Law*. Clarendon Press, Oxford

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL

Langbein J H 1985 The German advantage in civil procedure. *University of Chicago Law Review* **52**: 823–66

Latour B 1999 *Pandora's Hope: Essays on the Reality of Science Studies*. Harvard University Press, Cambridge, MA

Lawler A 1996 Support for science stays strong. *Science* **272**: 1256

Leplin J 1997 *A Novel Defense of Scientific Realism*. Oxford University Press, New York

Merton R K 1973 *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press, Chicago, IL

Pickering A (ed.) 1992 *Science as Practice and Culture*. University of Chicago Press, Chicago, IL

Popper K R 1968 *The Logic of Scientific Discovery*. Hutchinson, London

Sanders J 1998 *Bendectin On Trial: A Study of Mass Tort Litigation*. University of Michigan Press, Ann Arbor, MI

Schuck P H 1986 *Agent Orange On Trial: Mass Toxic Disaster in the Courts*. Belknap Press of Harvard University Press, Cambridge, MA

Schwartz A 1997 A 'dogma of empiricism' revisited: Daubert vs. Merrell Dow Pharmaceuticals, Inc. and the need to resurrect the philosophical insight of Frye vs. United States. *Harvard Journal of Law and Technology* **10**: 149

Segerstråle U 1993 Bringing the scientist back in the need for an alternative sociology of scientific knowledge. In: Brante T, Fuller S, Lynch W (eds.) *Controversial Science: From Content to Contention*. State University of New York Press, Albany, NY

Shapin S 1996 *The Scientific Revolution*. University of Chicago Press, Chicago, IL

Sobol R B 1991 *Bending The Law: The Story of the Dalkon Shield Bankruptcy*. University of Chicago Press, Chicago, IL

Tyler T R 1990 *Why People Obey the Law*. Yale University Press, New Haven, CT

United States v. *Starzecpyzel*, 880 F. Supp. 1027 (S.D.N.Y. 1995)

van Kampen P T C 1998 *Expert Evidence Compared: Rules and Practices in the Dutch and American Criminal Justice System*. Intersentia Rechtswetenschappen, Antwerp, Belgium

Wagner W E 1995 The science charade in toxic risk regulation. *Columbia Law Review* **95**: 1613–723

J. Sanders

Scientific Instrumentation, History and Sociology of

In the course of the last seventy years, historical and sociological analysis of instrumentation has changed considerably. World War II serves as an important watershed. Before the war, 'instrumentation' referred mainly to scientific instruments. They figured in experimentation, whose purpose was to demonstrate the truths of theory by making theoretical claims visible. After 1945 the ways in which instrumentation has been perceived and the functions attributed to it have multiplied and expanded, taking into account not only devices in the science laboratory, but also apparatus used in industry, government, health care, the military, and beyond. Instrumentation is now identified in many areas of science and technology

studies as central to research, engineering, industrial production, and to the processes of innovation. It is perceived as a mechanism that conditions the content of knowledge and affects the organization of work and even broader societal interactions.

In the writings of early twentieth-century students of science, instrumentation was seldom discussed and never highlighted. Historian-philosophers of science like Gaston Bachelard (1933, 1951) saw science chiefly in terms of the development of new theory. In this idealist historiographical tradition, experimentation received little attention, and instrumentation was only treated as a prop for experiments whose function was to document the discoveries embodied in scientific theories. Instruments did not invite study. Questions of instrument design, construction, and use, and their limitations, went unattended.

This is not to suggest, however, that there was no interest in scientific devices at the time. A few scholars were fascinated by them, and they strove to preserve apparatus and to unearth new devices. They were interested in the technical intricacies and novelty of instruments, for example those of Galileo Galilei. Devices were often viewed as antiquities, and due to this focus the distinction between the work of curators of science museums and instrument scholarship was sometimes a fine one. Here again, instrumentation was not treated as an active component in the knowledge production process, nor was it regarded as problematic in terms of its invention, use, or impact on the organization of research and science and technology communities.

Stimulating new perspectives in scientific instrumentation arose in the 1970s and 1980s in connection with historical and sociological investigations of post-1945 big science. For a long time most of the research that portrayed instrumentation as a central component of science and technology focused on devices in the physical sciences. The classic study by John Heilbron and Robert Seidel (1989) of the Berkeley cyclotron in the 1930s is emblematic of the new place of instrumentation in contemporary historiography. Issues of design, finance, engineering, and construction lay at the center of the cyclotron study. The cyclotron was portrayed as an instrument whose technical and social facets involved uncertainties. It was not a 'pure' instrument that reflected science's drive to probe the physical world. While the cyclotron in part served this objective, the instrument also reflected the economic and institutional environment of the San Francisco region, the hope for better healthcare, financial concessions wrung from the government, and involvement by wealthy research foundations and industry. This history demonstrates that scientific instrumentation may be guided by the scientific community, but that it is sometimes spawned by circumstances and forces outside the pale of science.

The idea that instruments are not neutral devices that serve science but elements that give structure to the scientific community first took root in studies of radio astronomy. This provocative concept was quickly extended to the sphere of high-energy physics at large (Krige 1996), oceanography (Mukerji 1992), and space science (Krige 2000). David Edge and Michael Mulkay (1976) first demonstrated that a scientific discipline, radio astronomy, which emerged in the 1950s and 1960s, was directly linked to or even defined by the design, construction, and diffusion of an altogether novel device: the radio telescope (itself an outgrowth of microwave technical research). The radio telescope discovered astronomical bodies and events. It contributed importantly to the birth of a new speciality, with its own university departments, journals, and national and international congresses. A new scientific instrument transformed knowledge, and it also affected the very institution of science.

In a more speculative, even iconoclastic representation of scientific instruments, they are depicted as the key to research career success, and yet more assertively, as decisive motors in determining what is true in science. In some areas of science, the equipment crucial to carrying out telling experiments is extremely scarce due to the expense of acquisition and operation. By virtue of possessing a monopoly in an area of instrumentation, a scientist or laboratory can exercise control over the production of the best experimental data. Studies in this vein have been done for the fields of biology (Latour and Woolgar 1979) and physics (Pickering 1984, Pinch 1986).

In this perspective Bruno Latour (1987) has suggested that scientific instruments yield not merely professional and institutional advantage, but more important, what is true and false, valid and invalid in science. A researcher's dealings with instruments empower him or her to be heard and to be 'right' during scientific controversies. By dint of possessing a strategic apparatus, a laboratory is well positioned to establish what is and is not a sound claim. Latour insists that arguments and findings based on 'weaker' instruments, on mathematics, and on rational evaluation are a poor match against a truly powerful scientific instrument. Analyses of this sort are diametrically opposite to those of pre-World War II idealist historiography.

A balanced, subtle, intellectual, and social contribution of instrumentation in the work of scientific research is found in the writings of Peter Galison. In *How Experiments End* (1987), this author argues that in twentieth-century microscopic experimental physics, instrument-generated signals are often crucial to settling rival claims, and he suggests that the work of separating noise from signals constitutes a key component in this process. In *Image and Logic*, Galison (1997) states that, contrary to what is often argued, scientific findings are not the outcome of interactions between theory and experimentation. He factors in a third element, namely instrumentation. Science thus derives from a triangular exchange between theory,

experimentation, and instrumentation. Galison speaks of a 'trading zone': a language and realm where these three currents merge, and where intelligibility is established. The brand of philosophical realism espoused by Ian Hacking (1983, 1989) likewise accords a central position to instrumentation. He insists that physical entities exist to the extent that instruments generate unarguably measurable effects. The classical example given by Hacking is the production of positrons whose presence induces palpable technical effects.

Today, other science and technology studies identify instrumentation as an element which influences and sometimes structures the organization of work. Instruments are depicted as rendering obsolete some activities (functions) and some groups, as stimulating fresh functions, and as helping create the backdrop for organizational transformations. In the case of early big science, high-energy physics, highly specialized physicists were replaced in the task of particle detection and tracking by armies of low skilled women observers because of the emergence of alternative large-scale photographic technologies and protocols. In parallel, new technical roles, framed by new work arrangements, arose for engineering and technician cadres who were assigned to design or maintain novel devices or to assure effective interfaces between instrument packages. In a different sphere, the introduction of the first electronic calculators and computers near the end of World War II transformed occupations and work organization as they spelled the end of the human calculators (mostly women) who during the early years of the war contributed to the military's high-technology programs. Due to the advent of the electronic computer in the late 1940s and 1950s, small horizontally organized work groups supplanted the former vastly bigger and vertically structured work system.

Outside science too, for example in the military and industry, instrumentation is also currently viewed as having a structuring impact on the organization of labor. Control engineering, connected as it is with chains of cybernetic-based instrument systems, is depicted as having profoundly modified the organization of certain activities in military combat operations, fire control, and guidance. Other studies highlight instrumentation as a force behind changes in the composition and organization of many industrial tasks and occupations, through automation and robotics (Noble 1984, Zuboff 1988). These alter the size of a work force and its requisite skills, and the internal and external chains of hierarchy and command.

Throughout the 1980s and 1990s, students of science and technology tracked and analyzed the development, diffusion, and implantation of devices in spheres increasingly distant from science and the laboratory. The concept 'instrumentation' took on a broader and different meaning from the initial historical and sociological concept of 'scientific instrumentation.'

Studies of medical instrumentation led the way in this important transformation. Stuart Blume's (1992) investigations of the emergence and diffusion of cat-scan devices, NMR imaging, and sonography illuminated the links between academic research, industrial research, the non-linear processes of industrial development of medical instrumentation, and the far-flung applications of such apparatus. Blume's and similar analyses had the effect of introducing significant complexities into the earlier fairly clear-cut notion of instrumentation. They blur the former understanding of a 'scientific instrument' by showing the multiple sites of its origins and development. They further show that an instrument may possess multiple applications, some in science and others in engineering, industry, metrology, and the military. One thing stands out with clarity; a scientific instrument is frequently coupled to industry in powerful ways: design, construction, diffusion, and maintenance. This link is historically situated in the late nineteenth and the twentieth centuries, and it appears to grow constantly in strength. This situation is pregnant with material and epistemological implications for science and beyond, as experiment design, laboratory work practices, reasoning processes, the organization of industry, and daily life are now all so interlocked with instrumentation.

With only a few exceptions, however, little historical and sociological work has thus far concentrated on firms specifically engaged in the design, construction, and diffusion of instrumentation. It was not until the late nineteenth and early twentieth centuries that the military general staffs and politicians of certain countries (Germany, France, Great Britain, and somewhat later the US, the USSR, and Japan) began to perceive that their nation's fate was tightly bound up with the quantity and quality of the instrumentation that endogenous industry could conceive and manufacture. In part because of this new consciousness, as well as the growth of scientific research, the number of companies involved in instrument innovation and production increased impressively. Mari Williams (1994) indicates that instrument companies must be thought of as key components in national systems of education, industry, government policy, and the organization of science. During much of the nineteenth century, France enjoyed the lead. England challenged the French instrument industry at the turn of the century, but by this time leadership clearly belonged to recently united Germany. Success in the scientific instrument industry appears to have been associated with tight organic bonds between specific firms and specific research laboratories, which was the case in England for companies close to the Cavendish. Germany's immense success was the product of an organic association between the military, government, industrial manufacturing, and instrument making firms (Joerges and Shinn 2001). In the twentieth century, such proximity (not only geographic but perhaps more

particularly in terms of working collaborations and markets) similarly proved effective in the United States.

Does there exist an instrument-making or instrument-maker culture? This question too has received little attention, and any answer would have to be historically limited. Nevertheless, one often-cited study does address this issue, albeit indirectly. For a sample of post-World War II US instrument specialists, the sociologist Daniel Shimshoni (1970) looked at instrument specialists' job mobility. He discovered that when compared to other similarly trained personnel, instrument specialists changed jobs more frequently than other categories of employees. However, the reasons behind this high mobility received scanty attention. One interpretation (not raised by Shimshoni) is that instrument specialists change employers in order to carry their instruments into fresh environments. Alternatively, once instrument specialists have performed their assigned tasks, perhaps employers encourage their departure from the firm, and instrument men are consequently driven to seek work elsewhere.

One theme that has gained considerable attention during the 1980s and 1990s is the relationship between innovation and instrumentation. In an influential study, the sociologist of industrial organization and innovation Eric Von Hippel (1988) has explored the sites in which instrument innovations arise, and has examined the connections between those sites and the processes of industrial innovation. He indicates that a sizable majority of industrially relevant instrument novelties comes from inside industry, and definitely not from academia or from firms specialized in instrumentation. The instruments are most frequently the immediate and direct consequence of locally experienced technical difficulties in the realms of product design, manufacture, or quality control. Instrumentation often percolates laterally through a company, and is thus usually home used as well as home grown. Many devices do not move beyond the firm in which they originate. Hence, while in some instances dealings with academia and research may be part of industry practice, instrumentation is normally only loosely tied to science. When the connection between instrumentation development and science is strong, it is, moreover, often the case that industry-spawned instrumentation percolates down to the science laboratory rather than academia-based devices penetrating industry.

Some sociologists of innovation suggest that instrumentation in industry is linked to research and development, and through it to in-house technology and to company performance. According to some studies, during the 1970s and 1980s instrumentation correlated positively with industrial performance and company survival for US plants in a range of industrial sectors (Hage et al. 1993). Firms that exhibited small concern for advanced technology tended to stumble or close when compared with companies that actively sought new technologies. According to the Von Hippel hypothesis, an important fraction of innovative technology would take the form of in-house instrument-related innovation.

The connection between academia, instrument innovation, and economic performance has been approached from a different perspective in a more narrowly researched and fascinatingly speculative study carried out by the economic historian Nathan Rosenberg (1994). Rosenberg suggests that many key instrument innovations having a great impact on industry spring from fundamental research conducted in universities. To illustrate this claim, he points to the university-based research on magnetic spin and resonance conducted by Felix Bloch at Stanford University in the late 1940s and 1950s. This basic physical research (theoretical and experimental) gave rise to nuclear magnetic resonance (NMR) and to NMR instrumentation; and in turn, NMR apparatus and related devices have spread outward in American industry, giving rise to new products and affecting industrial production operations. The best-known use of NMR instrumentation is in the area of medical imaging. Rosenberg suggests that the spillover effect of academic instrument research has been underestimated, and that the influence of academic research through instrumentation is characterized by a considerable multiplier effect.

In the final analytical orientation indicated here, instrumentation is represented as a transverse epistemological, intellectual, technical, and social force that promotes a convergence of practices and knowledge among scientific and technological specialities. Instrumentation acts as a force that partly transcends the distinctions and differentiations tied to specific divisions of labor associated with particular fields of practice and learning.

This transcending and transverse function is reserved to a particular category of instrumentation: 'research-technology' (Shinn 1993, 1997, 2000, Joerges and Shinn 2001). Research-technology is built around 'generic' devices that are open-ended general instruments. They result from basic instrumentation research and instrument theory. Examples include automatic control and registering devices, the ultracentrifuge, Fourier transform spectroscopy, the laser, and the microprocessor. Practitioners transfer their products into academia, industry, state technical services, metrology, and the military. By adapting their generic products to local uses, they participate in the development of narrow niche apparatus. The research-technologist operates in an 'interstitial' arena between established disciplines, professions, employers, and institutions. It is this in-between position that allows the research-technologist to design generalist non-specific generic devices, and then to circulate freely in and out of niches in order to diffuse them. Through this multi-audience diffusion, a form of

practice-based universality arises, as a generic instrument provides a lingua franca and guarantees stable outcomes to multiple groups involved in a sweep of technical projects.

Throughout the 1980s and 1990s studies devoted to instrumentation, or studies in which instrumentation plays a leading role, grew appreciably. The impressive number of instrument-related articles appearing in many of the major journals of the social studies of science and technology testify to the fact that the theme is now strong. Two general tendencies characterize today's enquiries into instrumentation: first, a great diversity in the number of analytic currents and research schools which perceive instrumentation as basic to past and contemporary cognitive and organizational activities; and second, considerable variety in the spheres of activity that are putatively affected by instrumentation and in the mechanisms that allegedly underpin instrument influence.

See also: Experiment, in Science and Technology Studies; History of Technology; Innovation, Theory of; Technological Innovation; Technology, Anthropology of

Bibliography

Bachelard G 1933 *Les intuitions atomistiques*. Boivin, Paris

Bachelard G 1951 *L'activité rationaliste de la physique contemporaine*. Presses Universtaires de France, Paris

Blume S 1992 *Insight and Industry: On the Dynamics of Technological Change in Medicine*. MIT Press, Cambridge, MA

Edge D O, Mulkay M J 1976 *Astronomy Transformed: The Emergence of Radio Astronomy in Britain*. Wiley, New York

Galison P 1987 *How Experiments End*. University of Chicago Press, Chicago

Galison P 1997 *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, Chicago

Hacking I 1983 *Representing and Intervening*. Cambridge University Press, Cambridge, UK

Hacking I 1989 The divided circle: a history of instruments for astronomy, navigation and surveying. *Studies in the History and Philosophy of Science* **20**: 265–370

Hage J, Collins P, Hull F, Teachman J 1993 The impact of knowledge on the survival of American manufacturing plants. *Social Forces*. **72**: 223–46

Heilbron J L, Seidel R W 1989 *Lawrence and his Laboratory: A History of the Lawrence Berkeley Laboratory*. University of California Press, Berkeley, CA

Joerges B, Shinn T 2001 *Instrumentation Between Science, State and Industry*. Kluwer, Dordrecht, The Netherlands

Krige J 1996 The ppbar project. In: Krige J (ed.) *History of CERN*. North-Holland, Amsterdam, Vol. 3

Krige J 2000 Crossing the interface from R&D to operational use: The case of the European meteorological satellite. *Technology and Culture* **41**(1): 27–50

Latour B 1987 *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge, MA

Latour B, Woolgar S 1979 *Laboratory Life: The Social Construction of Scientific Facts*. Sage, Beverley Hills, CA

Mukerji C 1992 Scientific techniques and learning: Laboratory 'signatures' and the practice of oceanography. In: Bud R, Cozzens S (eds.) *Invisible Connections: Instruments, Institutions and Science*. SPIE Optical Engineering Press, Bellingham, WA

Noble D F 1984 *Forces of Production: A Social History of Industrial Automation*, 1st edn. Knopf, New York

Pickering A 1984 *Constructing Quarks: A Sociological History of Particle Physics*. University of Chicago Press, Chicago

Pinch T J 1986 *Confronting Nature: The Sociology of Solar-neutrino Detection*. Reidel, Dordrecht, The Netherlands

Rosenberg N 1994 *Exploring the Black Box: Technology, Economics, and History*. Cambridge University Press, Cambridge, UK

Shimshoni D 1970 The mobile scientist in the American instrument industry. *Minerva* **8**(1): 58–89

Shinn T 1993 The Bellevue grand electroaimant, 1900–1940: Birth of a research-technology community. *Historical Studies in the Physical Sciences* **24**: 1, 157–87

Shinn T 1997 Crossing boundaries: The emergence of research technology communities. In: Leydesdorf L, Etzkowitz H (eds.) *Universities and the Global Knowledge Economy: A Triple Helix of University–Industry–Government Relations*. Cassel Academic Press, London

Shinn T 2000 Formes de division du travail scientifique et convergence intellectuelle (Forms of division of scientific labor and intellectual convergence). *Revue Française de Sociologie* **41**: 447–73

Von Hippel E 1988 *The Sources of Innovation*. Oxford University Press, Oxford, UK

Williams M E 1994 *The Precision Makers: A History of the Instrument Industry in Britain and France 1870–1939*. Routledge, London

Zuboff S 1988 *In the Age of the Smart Machine: The Future of Work and Power*. Basic Books, New York

T. Shinn

Scientific Knowledge, Sociology of

1. Origins

The central concern of the sociology of knowledge is the idea that what people take to be certain is an accident of the society in which they are born and brought up. It is obvious that religious or political truths are largely affected by their social settings. If the truths are moral then such relativism is more troubling. But if the truths are scientific, then the idea of social determination is widely regarded as subversive, or self-defeating—since the science which purports to show that science is socially situated is itself socially situated. The sociology of knowledge, as formulated by Mannheim (1936), avoided these dangerous and dizzying consequences by putting scientific and mathematical knowledge in a special category; other kinds of knowledge had roots in society but the knowledge of the natural sciences was governed by nature or logic. Scientific method, then, when properly applied, would insulate scientists from social influences and their knowledge should be more true and more

universal than other kinds of knowledge. From the early 1970s, however, groups of sociologists, philosophers, historians, and other social scientists began programs of analysis and research which treated scientific knowledge as comparable with other kinds of knowledge. Their work broke down the barrier between ordinary knowledge and scientific knowledge.

The post-1970s Sociology of Scientific Knowledge (SSK), which included major contributions from historians of science, shows us how to understand the impact of social influences on all knowledge. We can use the field's own approach to query its intellectual origins. A suitable model is a well-known paper by a historian, Paul Forman (1971), who argued that the rise of quantum theory owed much to the 'Weltanschauung' of Weimar Germany. In the same way it could be argued that the 1960s saw the growth of postwar prosperity in Europe which allowed the development of new markets with young people becoming powerful consumers of goods and producers of culture—much of it consisting of a rebellion against traditional forms of expression; it saw the development of the birth control pill, making possible a (temporary) breakdown of sexual mores; and it saw experiments with perception-affecting drugs, while protests against the Vietnam war threatened old lines of institutional authority. The new order had its counterpart in academe. Antipsychiatry questioned the barriers between the sane and the insane, while in Europe at least, romanticized versions of Marxism dominated debate. SSK's questioning of the traditional authority of science can plausibly be seen as a product of this social ferment.

The emblematic book of the period as far as science was concerned was Thomas Kuhn's (1962) *The Structure of Scientific Revolutions*. Kuhn's book was influential in helping to reinforce the intellectual atmosphere in which radical ideas about science could flourish. Kuhn's notion of the incommensurability of paradigms, and paradigm revolution, provided a way of thinking about scientific knowledge as a product as much of culture as of nature. The book was not widely discussed until late in the 1960s, when the new social movements were at their height, and this adds force to the argument.

Philosophy of science in the early 1970s was also affected by this turn. The so-called 'Popper–Kuhn' debate (Lakatos and Musgrave 1970) pitted Karl Popper's (e.g., 1959) notion that scientific theories could be falsified with near certainty, even if they could not be proved, against Thomas Kuhn's (1961, 1962) claim that what was taken to be true or false varied in response to sudden revolutions in scientific world view. The 'Duhem–Quine thesis' (Losee 1980) showed that scientific theories were supported by networks of observations of ideas such that no one observation could threaten any other part of the network directly. Imre Lakatos's ([1963]/1976) brilliant analysis of the history of Euler's theorem showed that falsification of a theory was only one choice when faced with apparently recalcitrant observations. Lakatos's work was particularly attractive to the social analyst of science because it dealt with the detailed history of a real case rather than arising from general principles. Lakatos was showing how mathematicians argued (or could have argued in principle). This was the world of philosophical activity into which the sociology of scientific knowledge was born.

Like any other group of knowers, sociologists of scientific knowledge prefer models of the genesis of their own ideas which reflect their self-image as 'rational creatures' or, in Mannheim's phrase, 'free floating intellectuals.' In these models Kuhn plays a much smaller role. There are two intellectual stories. One turns on Ludwik Fleck's ([1935]/1979) *Genesis and Development of a Scientific Fact*, which Kuhn cites in the preface of his 1962 book, and which already contained many of the key ideas found therein. Furthermore, Fleck's book anticipated the developments in sociology of scientific knowledge which came in the 1970s in that, unlike *The Structure of Scientific Revolutions*, it included a case study of a contemporaneous scientific episode, Fleck's own research on the Wasserman reaction for the diagnosis of syphilis. Indeed, Fleck's book remains unique in sociology of science in that his social analysis was of a pioneering piece of scientific research which he himself was pioneering. Thus it is the most perfect piece of participant observation yet done in the social study of scientific knowledge.

Fleck's book, however, was not widely known until many years after Kuhn's was published, and not until long after the sociology of scientific knowledge had established its own way of doing things. Fleck, then, was the first to set out some of the crucial ideas for SSK, but he had little influence on the new movement except through his influence on Kuhn, while Kuhn helped open up the social and intellectual space for what came after rather than providing intellectual foundations or methodological principles; in other words, Kuhn provided the intellectual space but not, to use his own term, the paradigm.

The origin of SSK's intellectual and methodological *paradigm* has to be understood by looking at more direct influences. In the 1970s a group of philosophers, concerned with defending science against what they perceived as its critics, put forward the argument that sociological analysis could be applied only to knowledge that was in error, whereas true knowledge remained insulated from social forces. The intellectual source for the sociologists who inspired these initial attacks and fought against them was yet another philosopher, Ludwig Wittgenstein, whose work was central to the anthropological debate about whether there were universal standards of rationality (Wilson 1970). Readings of Ludwig Wittgenstein's later philosophy (1953, 1956, Winch 1958) were the crucial theoretical resource for the work in sociology of

scientific knowledge that had its origins in the mid-1970s. The 'Edinburgh School' and its 'strong program' argued on philosophical grounds that 'symmetrical analysis'—analysis that treated true knowledge and false knowledge equally—was possible. Historical studies conducted in this framework showed how social influence affected scientific conclusions, in science in general, in early studies of the brain, in the development of statistics, and in high-energy physics (Bloor 1973, 1983, Barnes 1974, Shapin 1979, MacKenzie 1981, Pickering 1984). Independently, a group at the University of Bath applied Wittgensteinian ideas to contemporary episodes of science, particularly controversies in physics and the fringe sciences, and developed the methodology of interviewing at networks of competing scientific centers. Its program became known as the 'Empirical Programme of Relativism' (EPOR; Collins 1975, 1985, Pinch 1977, 1986, Travis 1980, 1981). These approaches were to become quite widely adopted in the early years of SSK.

As the 1970s turned into the 1980s other groups, more strongly influenced by anthropological practice, entered and influenced the field. The Edinburgh–Bath approach was influenced by anthropology in that much of the discussion of the meaning of Wittgensteinian philosophy for the social sciences turned on questions to do with the knowledge of isolated 'tribes' and its relationship to Western knowledge. The newer contributors, however, took their methodological approach as well as their philosophy from anthropology. Their work founded the tradition of 'laboratory studies.' In the Strong Program the source of material tended to be historical archives; in EPOR the method was interviews with members of the 'core-set' of scientific laboratories contributing to a scientific controversy; in laboratory studies the method was a long sojourn in, or deep study of, a single laboratory (Latour and Woolgar 1979, Knorr Cetina 1981). The setting for the laboratory studies tended to be the life sciences rather than the hard sciences or fringe sciences.

Another important input was the interpretive tradition in sociology associated with phenomenology and ethnomethodology. Indeed the term 'social construction' is a widely used description of the approach: 'social construction of scientific knowledge' was taken from the title of a well-known book by Peter Berger and Thomas Luckman (1967)—though Berger's *Invitation to Sociology* (1963) probably had more direct impact on the field. Ethnomethodology strongly influenced all those who practiced SSK and especially the detailed analyses of scientific practice such as those by Lynch (1985 [but written much earlier than its publication date]), Woolgar (1976), and especially the move of some authors into what became known as 'discourse analysis' (Knorr Cetina and Mulkay 1983).

As the 1980s turned into the 1990s the concentration on language that began with the move to discourse analysis helped SSK to come to seem a part of the much larger movement known as 'postmodernism.' The influence of French philosophers of literature and culture such as Derrida became marked. Since then 'cultural studies of science,' which share with SSK what could broadly be called a 'social constructivist' approach to science, have attracted large numbers of followers in the humanities as well as the social sciences. The respective methods and, more especially, the attitudes to methodology, make it possible to separate cultural studies from SSK. Crucially, SSK practitioners still take natural science as the touchstone where matters of method are concerned, though the model of science is very far from the narrow statistically-based notion of science that informed the 'scientific sociology' of the 1950s and early 1960s and continues to dominate much mainstream sociological practice in the USA. SSK stresses the more general values of careful empirical observation and repeatability. Cultural studies, on the other hand, takes philosophical literary criticism, or semiotics, as the model to be followed.

2. *Concepts*

Among sociology's traditional topics was the study of the social factors which affected scientists' choice of research topic. The sociology of scientific knowledge, however, took it that the *outcomes* of research projects were also affected by their social setting. SSK showed that theoretical and experimental procedures did not determine, or fully determine, the conclusions of scientific research. The philosophical ideas described above provided an important starting point, but sociology developed its own concepts set in the working world of the scientist rather than the abstract world of the philosopher. One strand of research showed how difficult it was to transfer the skills of physics experimentation between settings, that it was even harder to know that such skills had been transferred, and that it was harder still to know when an experiment had been satisfactorily completed. This meant we did not know whether a negative experiment should be taken to contradict a positive one. This argument became known as 'the experimenter's regress' (Collins 1975, 1985), and revealed the scope for 'interpretive flexibility' regarding the outcomes of passages of experimentation and theorization. It also means that even while SSK stresses the importance of empirical research and repeatability, it is aware that, by themselves, such procedures cannot turn disputed knowledge into certainty in either the natural or social sciences.

Even if experimental science produced widely accepted data, as Pinch (1981) showed, their 'evidential significance' could vary and the same finding could count as revolutionary or insignificant. Sociological studies also described the process between the dis-

ordered activities of the laboratory and the orderly 'findings' (Latour and Woolgar 1979, Knorr Cetina 1981) or 'discoveries' (Brannigan 1981) reported—or constructed—in the published literature. Latour and Woolgar demonstrated the way that 'modalities'—phrases that qualified a finding or referred to the particular time and place of its generation—were successively stripped from publications as the finding became established. What they called 'inversion' was the way the stripping of the modalities, and like processes, comprised the establishment of the finding—a reversal of the accepted direction of the causal arrow. It was also shown that, with some exceptions, 'distance lends enchantment,' that is, certainty about scientific outcomes tends to increase as the intepreter's distance from the laboratory increases (Collins 1985, MacKenzie 1998). This explained why commentators' accounts often seemed far more confident than the reports of the scientists themselves, and explained much about the way scientific knowledge was diffused beyond the specialists.

Mechanisms were described by which potentially open-ended disputes were brought to a close around a particular interpretation, with EPOR concentrating on fringe sciences, the 'French School' concentrating on the interplay of actors within networks (Latour 1987), and the 'Edinburgh School' concentrating on large-scale political influences on the content of ideas (Shapin 1979, MacKenzie 1981, Shapin and Schaffer 1987). More recently, work on closure has broadened to include studies in the social construction of technology, the public understanding of science and law and science.

3. *Significance*

As the recent so-called 'science wars' have revealed, outsiders consider that the importance of SSK and related analyses of science lies in the impact they have on the perception of the meaning of science in the wider world. SSK is widely seen as a radical relativist attack on science. But both the philosophical and political radicalism of the social analysis of science varies from program to program and study to study.

For example, 'epistemological relativism' implies that one social group's way of justifying its knowledge is as good as another's and that there is no external vantage point from which to judge between them; all that can be known can be known only from the point of view of one social group or another. Ontological relativism is the view that, in social groups such as those described above, reality itself is different. A combination of epistemological and/or ontological relativism can be referred to as 'philosophical relativism.' This attitude is nicely captured by McHugh in the following quotation: 'We must accept that there are no adequate grounds for establishing criteria of truth except the grounds that are employed to grant or concede it—truth is conceivable only as a socially organized upshot of contingent courses of linguistic, conceptual and social behaviour' (1971, p. 329). Philosophical relativism is, then, a philosophically radical viewpoint.

A still more philosophically radical position is the 'actor (or actant) network theory' (ANT) most closely associated with Michel Callon and Bruno Latour. The approach was first signaled by Latour and Woolgar when, in 1986, they changed the subtitle of *Laboratory Life* (Latour and Woolgar 1979) from *The Social Construction of Scientific Facts* to *The Construction of Scientific Facts* so as to signify that the 'social' in their view no longer deserved special mention in the shaping of scientific knowledge. Subsequently, as Callon and Latour developed their ideas, scientific facts came to be seen as emerging from the interplay of 'actants' in the 'text,' of life—suggesting that ANT should most properly be included under cultural studies of science rather than SSK. Within Callon and Latour's 'text,' terms such as 'social' and 'natural' are themselves outcomes of the interplay of actants rather than original causes. Therefore, to draw attention to the social as a special factor is to begin the discussion at too low a level of generality. Under this approach 'constructivism' has ceased to be *social* and among the actants nonhumans are given as much reality-forming agency as humans (Callon 1986). The philosophical radicalness of ANT is clear in its refusal to accept even the notion of human and nonhuman as primary.

Methodological relativism, by contrast, says nothing direct about reality or the justification of knowledge. Methodological relativism is an attitude of mind recommended by some of those who practice SSK; it says that the sociologist or historian of science should act as though the beliefs about reality of any competing groups being investigated are not caused by the reality itself. The intention is to limit analysis to the kind of causes of scientific beliefs that are located in the domain of the social. Methodological relativism is meant to push exploration of the social causes of belief to the limit without having it cut off by the argument that a belief was widely accepted because it was rational. Methodological relativism is, then, not at all radical as a philosophy, it is a (social) scientific method.

Relativism may be called time-bounded if it claims only that scientific procedures are often inconclusive in the short term whatever the deeper significance of the consensuses that are attained in the long term. This kind of relativism, which has no philosophical potency, is all that is needed to support much current social analysis of contemporary science, especially controversies that impact on the public domain (e.g., Collins and Pinch 1993, Richards 1991, Epstein 1996).

The relationship between philosophical radicalism and political radicalism is, however, perverse. Time-bounded relativism has consequences for contemporary scientific controversies such as engage

environmentalists and the like: it has this power because it shows that, contrary to the way these things are treated in textbooks and the like, the 'short-term' period when scientific disputes are being resolved is often many decades long. The empirical and analytic tools of SSK explain why this is likely to be so and it means that we should not expect speedy science-based resolutions to our current technological dilemmas. From this it can be argued that the decision-making power once monopolized by scientific experts should be more widely shared (Wynne 1987). Methodological relativism, which is philosophically quietist, tends to be disturbing to the scientific community as it blatantly ignores scientific consensus however well established. To go the other extreme, ANT, the most philosophically radical position of all, by removing humans from the central position they hold in *social* constructivism, recreates a relationship between scientists and their nonhuman objects of study which is similar to that which held before the revolutions of the 1960s and 1970s (Collins and Yearley 1992, Callon and Latour 1992, Fuller 2000). What about philosophical relativism? Scientists and philosophers seem to believe that philosophical relativism can be used to justify nonscientific beliefs such as astrology or creationism, though its proponents deny it fervently, and at least one critic (Fuller 2000) has argued that it is politically quietist. SSK, of course, argues that as knowledge moves from its seat of creation to a wider audience, it tends to be stripped down and simplified, so it is not surprising that the subtleties of the various relativist positions have escaped the more outspoken critics of the social analysis of scientific knowledge.

Bibliography

Barnes B S 1974 *Scientific Knowledge and Sociological Theory*. Routledge and Kegan Paul, London

Barnes B S, Bloor D, Henry J 1996 *Scientific Knowledge: A Sociological Analysis*. Athlone Press, London

Berger P L 1963 *Invitation to Sociology*. Anchor Books, Garden City, NY

Berger P L, Luckman T 1967 *The Social Construction of Reality*. Allen Lane, London

Bloor D 1973 Wittgenstein and Mannheim on the sociology of mathematics. *Studies in the History and Philosophy of Science* **4**: 173–91

Bloor D 1983 *Wittgenstein: A Social Theory of Knowledge*. Macmillan, London

Brannigan G 1981 *The Social Basis of Scientific Discoveries*. Cambridge University Press, New York

Callon M 1986 Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In: Law J (ed.) *Power, Action and Belief: A New Sociology of Knowledge?* Routledge & Kegan Paul, London, pp. 196–233

Callon M, Latour B 1992 Don't throw the baby out with the bath school! In: Pickering A (ed.) *Science as Practice and Culture*. University of Chicago Press, Chicago, pp. 343–68

Collins H M 1975 The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. *Sociology* **9**(2): 205–24

Collins H M 1985 *Changing Order: Replication and Induction in Scientific Practice*. Sage, Beverly Hills & London [2nd edn., University of Chicago Press, 1992]

Collins H M, Pinch T J 1993 *The Golem: What Everyone Should Know About Science*. Cambridge University Press, Cambridge & New York [subsequent editions, 1994, 1998]

Collins H M, Yearley S 1992 Epistemological chicken. In: Pickering A (ed.) *Science as Practice and Culture*. University of Chicago Press, Chicago, pp. 301–26

Epstein S 1996 *Impure Science: AIDS, Activism and the Politics of Knowledge*. University of California Press, Berkeley, Los Angeles & London

Feyerabend P K 1975 *Against Method*. New Left Books, London

Fleck L 1979 *Genesis and Development of a Scientific Fact*. University of Chicago Press, Chicago [first published in German in 1935]

Forman P 1971 Weimar culture, causality and quantum theory, 1918–1927: Adaptation by German physicists and mathematicians to a hostile intellectual environment. In: McCormmach R (ed.) *Historical Studies in the Physical Sciences*, No 3. University of Pennsylvania Press, Philadelphia, PA, pp. 1–115

Fuller S 2000 *Thomas Kuhn: A Philosophical History for our Times*. University of Chicago Press, Chicago

Knorr Cetina K D 1981 *The Manufacture of Knowledge*. Pergamon, Oxford, UK

Knorr Cetina K D, Mulkay M (eds.) 1983 *Science Observed: Perspectives on the Social Study of Science*. Sage, London & Beverley Hills

Kuhn T S 1961 The function of measurement in modern physical science. *ISIS* **52**: 162–76

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Lakatos I 1976 *Proofs and Refutations*. Cambridge University Press, Cambridge [originally published in *British Journal for the Philosophy of Science* 1963 **XIV**: 1–25, 120–39, 221–45, 296–342]

Lakatos I, Musgrave A (eds.) 1970 *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK

Latour B 1987 *Science in Action*. Open University Press, Milton Keynes, UK

Latour B, Woolgar S 1979 *Laboratory Life: The Social Construction of Scientific Facts*. Sage, London & Beverly Hills [2nd edn., 1986]

Losee J 1980 *A Historical Introduction to the Philosophy of Science*. Oxford University Press, Oxford, UK

Lynch M 1985 *Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory*. Routledge and Kegan Paul, London

McHugh P 1971 On the failure of positivism. In: Douglas J D (ed.) *Understanding Everyday Life*. Routledge and Kegan Paul, London, pp. 320–35

MacKenzie D 1981 *Statistics in Britain 1865–1930*. Edinburgh University Press, Edinburgh, UK

MacKenzie D 1998 The certainty trough. In: Williams R, Faulkner W, Fleck J (eds.) *Exploring Expertise: Issues and Perspectives*. Macmillan, Basingstoke, UK

Mannheim K 1936 *Ideology and Utopia: An Introduction to the Sociology of Knowledge*. University of Chicago Press, Chicago

Pickering A 1984 *Constructing Quarks: A Sociological History of Particle Physics*. Edinburgh University Press, Edinburgh, UK

Pinch T J 1977 What does a proof do if it does not prove? In: Mendelsohn E, Weingart P, Whitley R (eds.) *The Social Production of Scientific Knowledge*. Reidel, Dordrecht, The Netherlands
Pinch T J 1981 The sun-set: The presentation of certainty in scientific life. *Social Studies of Science* **1**(11): 131–58
Pinch T J 1986 *Confronting Nature: The Sociology of Solar-neutrino Detection*. Reidel, Dordrecht, The Netherlands
Popper K R 1959 *The Logic of Scientific Discovery*. Harper & Row, New York
Richards E 1991 *Vitamin C and Cancer: Medicine or Politics*. Macmillan, Basingstoke, UK
Shapin S 1979 The politics of observation: Cerebral anatomy and social interests in the Edinburgh Phrenology Disputes. In: Wallis R (ed.) *On the Margins of Science: The Social Construction of Rejected Knowledge, Sociological Review Monograph, 27*. Keele University Press, Keele, UK, pp. 139–78
Shapin S, Schaffer S 1985 *Leviathan and the Air Pump: Hobbes, Boyle and the Experimenal Life*. Princeton University Press, Princeton, NJ
Travis G D L 1980 On the importance of being earnest. In: Knorr K, Krohn R, Whitley R (eds.) *The Social Process of Scientific Investigation: Sociology of the Sciences Yearbook, 4*. Reidel, Dordrecht, The Netherlands, pp. 165–93
Travis G D L 1981 Replicating replication? Aspects of the social construction of learning in planarian worms. *Social Studies of Science* **11**: 11–32
Wilson B (ed.) 1970 *Rationality*. Blackwell, Oxford, UK
Winch P G 1958 *The Idea of a Social Science*. Routledge and Kegan Paul, London
Woolgar S 1976 Writing an intellectual history of scientific developments: The use of discovery accounts. *Social Studies of Science* **6**: 395–42
Wittgenstein L 1953 *Philosophical Investigations*. Blackwell, Oxford, UK
Wittgenstein L 1956 *Remarks on the Foundations of Mathematics*. Blackwell, Oxford, UK
Wynne B 1987 *Risk Management and Hazardous Wastes: Implementation and the Dialectics of Credibility*. Springer, Berlin

H. M. Collins

Scientific Reasoning and Discovery, Cognitive Psychology of

The cognitive psychology of scientific reasoning and discovery refers to the study of the cognitive processes that scientists use in all aspects of science. Researchers have used interviews and historical records, cognitive experiments on components of scientific thinking, computational models based on particular scientific discoveries, and investigations of scientists as they reason live, or '*in vivo*,' in an effort to uncover the thinking and reasoning strategies that are important in science. In this article, six different approaches to scientific reasoning are discussed. One important point to note is that scientific thinking builds upon many different cognitive components such as induction, deduction, analogy, problem solving, priming, and categorization, that are objects of study in their own rights. Research specifically concerned with scientific thinking tends to use content domains that are from an established domain of science (such as physics, biology, or chemistry), or looks at how different cognitive processes such as concepts and deduction are used together in areas like experiment design. Useful books on the nature of scientific thinking are Tweeney et al. (1982), Giere (1992), and Klahr et al. (2000).

1. *Interviews and the Historical Record*

Two frequently used and related approaches that have been used to investigate scientific thinking have been interviews with scientists and the analysis of historical records and documents such as notebooks. One of the earliest accounts of scientific thinking and reasoning was the interview of Albert Einstein conducted by the Gestalt Psychologist Max Wertheimer (1959). Wertheimer argued that a key strategy used by Einstein was to search for invariants. Wertheimer saw the velocity of light as a key invariant around which Einstein built his theory. Wertheimer incorporated his analysis into a Gestalt theory of thought. More recently, researchers have conducted interviews in the context of principles from cognitive science. For example, Paul Thagard (1999) has conducted many interviews with the scientists who proposed that ulcers are caused by bacteria. Thagard has pointed to the important roles of serendipity, observation, and analogy in this discovery. A related line of inquiry is the use of historical documents. Using the scientists' lab books, biographical, and autobiographical materials, researchers attempt to piece together the reasoning strategies that the scientists used in making a discovery. For example, Nersessian (1992) has conducted extensive analyses of the physicist Faraday's notebooks and has argued that the key to understanding his discoveries is in terms of his use of mental models. By mapping out the types of mental models that Faraday used and showing how these types of models shaped the discoveries that Faraday made, Nersessian offered a detailed account of the mental processes that led to a particular discovery. Another cognitive approach using the historical record is to take a real scientific discovery, such as Monod and Jacob's Nobel Prize-winning discovery of a mechanism of genetic control and give people the same problem, using a simulated scientific laboratory, and determine whether people use the same discovery strategies that the scientists used to make the discovery, such as focusing on unexpected findings (Dunbar 1993).

2. *Scientific Reasoning as Problem Solving and Concept Formation*

Two common approaches to scientific thinking have been to see it as a way of discovering new concepts or

as a form of problem solving. Beginning with Bruner et al.'s (1956) classic experiments in which college students were asked to induce the rule that determines whether an item is, or is not, a member of a category, these researchers attempted to discover the types of inductive reasoning strategies used to acquire new concepts. Bruner et al. argued that much of science consists of inducing new concepts from data and that the memory loads that different strategies require will make certain types of inductive reasoning strategies more common than others. More recently, Holland et al. (1986) provided an account of the different inductive learning procedures that could be used to acquire new concepts in science.

Herbert Simon (1977) argued that concept formation is a form of problem solving, thus the two approaches can be seen as complimentary (see Klahr et al. 2000). Simon argued that scientific thinking consists of a search in a problem space with the two main spaces being a hypothesis space and an experiment space. The hypothesis and experiment spaces refer to all possible hypotheses, experiments, and operators that can be used to get from one part of the space to the next part of the space, such as grouping common elements (the grouping operator) in sets of results to form a new hypothesis. Simon has taken specific scientific discoveries and mapped out the types of heuristics (or strategies), such as heuristics for designing experiments that a scientist used in searching the experiment space. Using the notion of searching in a problem space, other researchers have analyzed the types of search heuristics that are used in all aspects of scientific thinking and have conducted experiments on the problem-solving heuristics that people use in designing experiments, formulating hypotheses, and interpreting results (Dunbar 1993, Klahr et al. 2000). These approaches specify the types of knowledge that an individual must possess and the heuristics that are used to formulate hypotheses, design experiments, and interpret data.

3. *Errors in Scientific Thinking*

One of the most frequently investigated aspects of scientific thinking and reasoning has been the finding that both scientists and participants in psychology experiments attempt to confirm their hypothesis when they conduct an experiment, sometimes called 'confirmation bias' (see Tweeney et al. 1982). Following from the writings of the philosopher Karl Popper, many psychologists have assumed that attempting to confirm a hypothesis is a faulty reasoning strategy. Numerous studies have revealed that, when given an hypothesis to test, people will design experiments that will confirm their hypothesis and not conduct experiments that could falsify their own hypothesis. This is a pervasive phenomenon that is difficult to eradicate; even when given instructions to falsify hypotheses, people find it difficult to do. Thus, many researchers have concluded that both people in psychology experiments as well as scientists at large make this faulty reasoning error. However, Klayman and Ha (1987) argued that conducting experiments that confirm a hypothesis is not necessarily a scientific reasoning error. They argued that if the prior probability of confirming one's hypothesis is low, then even if the scientist is attempting to confirm the hypothesis, it can still be disconfirmed. One other interpretation of the phenomenon of confirmation bias is that early in developing a theory or a hypothesis people will attempt to confirm the hypothesis; however once the hypothesis is fleshed out and confirmed, people will attempt to conduct disconfirming experiments (see Tweeney et al. 1980).

4. *Science 'In vivo': How Scientists Think in Naturalistic Contexts*

One important issue in scientific reasoning and discovery is that most accounts have tended to use indirect evidence such as lab notebooks, biographies, and interviews with scientists to determine the thinking and reasoning strategies that scientists use. Another approach is to conduct experiments on isolated aspects of scientific thinking. See Dunbar (1995) for an analysis of these standard approaches that he has termed '*in vitro*.' Both approaches, while very informative, do not look at scientists directly. Thus, a complimentary approach has been to investigate real scientists' thinking and reasoning strategies while they are conducting real research. Using this '*in vivo*' approach, Dunbar (1999) has identified the specific ways that scientists use analogies, deal with unexpected findings, and use collaborative reasoning strategies in their research. He found that scientists use analogies to similar entities (or 'local analogies') when fixing experimental problems, analogies to entities from the same class of items (or 'regional analogies') when formulating new hypotheses, and analogies to very dissimilar domains ('long-distance analogies') when explaining scientific issues to others. Furthermore, Dunbar found that over half the findings that scientists obtain are unexpected, and that the scientists have specific strategies for dealing with these unexpected findings: First, scientists provide methodological explanations using local analogies that suggest ways of changing their experiments to obtain the desired result. If the changes to experiments do not provide the desired results, then the scientists switch from blaming the method to formulating hypotheses; this involves the use of 'regional analogies,' as well as collaborative reasoning in which groups of scientists build models and theories together. Dunbar has further brought back these '*in vivo*' findings into the cognitive laboratory to conduct controlled experi-

ments, which, taken together, have been used to build new accounts of the ways that analogy, collaborative reasoning, and causal reasoning are used in scientific thinking (Dunbar 1999).

5. The Development of Scientific Thinking Skills

Beginning with the work of Piaget, many researchers have noted that children are similar to scientists. This 'child-as-scientist' metaphor has two main strands. First, children's acquisition of new concepts and theories is said to be similar to the large conceptual changes that occur in scientific fields. Researchers investigating this view have pointed to parallels between changes in children's concepts, such as their concepts of heat and temperature, and changes in the concepts of heat and temperature in the history of physics (see Chi 1992, Carey 1992). The second strand of the child-as-scientist metaphor is that children reason in identical ways to scientists, ranging from deduction and induction to experimental design. Some researchers have argued that there is little difference between a scientist and a three-year-old; while scientists clearly have considerably more knowledge of specific domains than children, their underlying competencies are viewed as the same. Other researchers have taken this 'child-as-scientist' view even further and argued that infants are basically scientists. Yet other researchers have argued that there are fundamental differences between children and scientists and that scientific thinking skills follow a developmental progression (see Klahr et al. 2000 for an overview of this debate).

6. Cognitively Driven Computational Discovery: Twenty-first Century Scientific Discovery

One development that has occurred in many sciences is the placing of vast amounts of information in computer databases. In the year 2000 the entire human genome was mapped and now exists in databases. Similar developments have occurred in physics, where the entire universe has been mapped and put on a database. In the case of the human genome, data consists of long sequences of nucleotides, each of which is represented by a letter such as A for Adenine. Databases consist of strings of letters such as ATGTC with each letter representing a particular nucleotide. These strings or sequences can extend for hundreds of millions of nucleotides without interruption. Buried in these sequences are genes, and most of the genes are of unknown function. One goal of researchers is to search the database, find genes, and determine the function of the genes and how the genes interact with each other. This new wave of scientific investigation incorporates many of the principles of analogical reasoning, inductive reasoning and problem-solving strategies discussed here, as well as many of the algorithms (neural nets, Markov models, and production systems) discovered by cognitive psychologists in the 1980s and 1990s to discover the functions of genetic sequences and the properties of certain types of matter in the universe. One interesting aspect of this development has been that rather than computer programs being expected to make an entire discovery from beginning to end, they are a tool that can be used by scientists to help make a discovery. The cognitive psychology of scientific reasoning has moved from being a description of the scientific mind to an active participant in scientific practice.

See also: Discovery Learning, Cognitive Psychology of; History of Science; Informal Reasoning, Psychology of; Piaget's Theory of Child Development; Problem Solving and Reasoning: Case-based; Problem Solving and Reasoning, Psychology of; Problem Solving: Deduction, Induction, and Analogical Reasoning; Reasoning with Mental Models; Scientific Concepts: Development in Children

Bibliography

Bruner J S, Goodnow J J, Austin G A 1956 *A Study of Thinking*. Wiley, New York

Carey S 1992 The origin and evolution of everyday concepts. In: Giere R N (ed.) *Minnesota Studies in the Philosophy of Science. Vol. XV: Cognitive Models of Science*. University of Minnesota Press, Minneapolis, MN, pp. 89–128

Chi M 1992 Conceptual change within and across ontological categories. Examples from learning and discovery in science. In: Giere R N (ed.) *Minnesota Studies in the Philosophy of Science. Vol. XV: Cognitive Models of Science*. University of Minnesota Press, Minneapolis, MN, pp. 129–86

Dunbar K 1993 Concept discovery in a scientific domain. *Cognitive Science* **17**: 397–434

Dunbar K 1995 How scientists really reason: Scientific reasoning in real-world laboratories. In: Sternberg R J, Davidson J E (eds.) *Mechanisms of Insight*. MIT Press, Cambridge, MA, pp. 365–95

Dunbar K 1999 The scientist in vivo: How scientists think and reason in the laboratory. In: Magnani L, Nersessian N, Thagard P (eds.) *Model-based Reasoning in Scientific Discovery*. Kluwer Academic/Plenum Publishers, New York, pp. 89–98

Giere R N (ed.) 1992 *Minnesota Studies in the Philosophy of Science. Vol. XV: Cognitive Models of Science*. University of Minnesota Press, Minneapolis, MN

Holland J H, Holyoak K J, Nisbett R E, Thagard P R 1986 *Induction: Processes of Inference, Learning, and Discovery*. MIT Press, Cambridge, MA

Klahr D, Dunbar K, Fay A, Penner D, Schunn C 2000 *Exploring Science: The Cognition and Development of Discovery Processes*. MIT Press, Cambridge, MA

Klayman J, Ha Y W 1987 Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review* **94**: 211–28

Nersessian N J 1992 How do scientists think? Capturing the dynamics of conceptual change in science. In: Giere R N (ed.) *Minnesota Studies in the Philosophy of Science. Vol. XV:*

Cognitive Models of Science. University of Minnesota Press, Minneapolis, MN
Simon H A 1977 *Models of Discovery: and Other Topics in the Methods of Science*. D Reidel Publishing, Dordrecht, The Netherlands
Thagard P 1999 *How Scientists Explain Disease*. Princeton University Press, Princeton, NJ
Tweney R D, Doherty M E, Mynatt C R (eds.) 1981 *On Scientific Thinking*. Columbia University Press, New York
Wertheimer M 1959 *Productive Thinking*. Harper and Row, New York

K. Dunbar

Scientific Revolution: History and Sociology of

1. The Term 'Scientific Revolution'

The term 'Scientific Revolution' was introduced to denote what has been considered since the nineteenth century one of the most important discontinuities in the history of European science (H F Cohen 1994, I B Cohen 1985, Lindberg and Westman 1990). It covered roughly the period between Copernicus and Newton, and led from Aristotelian natural philosophy (see *Aristotle (384–322 BC)*) deriving its dogmatic authority from the Church to the establishment of the classical sciences and their institutions, i.e., the period between 1500 and 1700 (see *Scientific Disciplines, History of Scientific Academies, History of*).

This period can be characterized as one in which a new social group, the 'engineer-scientists' (consisting of engineers, scientists, inventors, artists, and explorers), emerged and became institutionalized. This group confronted traditional natural philosophy with the challenges of practice and experience, but also engaged in self-contained explanations of natural phenomena, expecting that science is a means to master nature, as Francis Bacon put it. Among the lasting achievements of the Scientific Revolution were the establishment of heliocentric astronomy, classical mechanics, as well as numerous contributions to optics, chemistry, physiology and other areas of modern science (for an overview see Butterfield 1965, Dijksterhuis 1986, Hall 1954). Many of these achievements became, in fact, the basis for technological breakthroughs. However, these breakthroughs occurred, as a rule, much later than anticipated by the protagonists of the Scientific Revolution. Further, the intellectual breakthroughs responsible for the revolution's lasting impact on the development of science were mainly attained towards its end and only a generation or more after their proclamation by the early pioneers (Damerow et al. 1992).

The Scientific Revolution has become a paradigmatic reference for all approaches to the history and philosophy of science (see *History of Science*). At least since the Enlightenment it has been conceived as the triumph of the scientific method over the irrationalism of religious beliefs (see *Science and Religion*). Opposition to this view, together with a growing professionalization of historical studies, has opened up the space for other accounts, including attempts—such as that by Pierre Duhem—to deny that the Scientific Revolution actually represented a radical break, claiming instead that it merely constituted an episode within a continuous accumulation of scientific knowledge since the Middle Ages (Duhem 1996). Furthermore, the traditional emphasis on the role of outstanding protagonists of the Scientific Revolution, such as Bacon, Galileo, and Descartes, and their individual 'discoveries' (see e.g., Koyré 1978) has, in recent scholarship, increasingly receded in favor of an analysis of contexts tracing scientific achievements back to cultural, social, and economic conditions (Osler 2000, Porter and Teich 1992, Shapin 1996). The Scientific Revolution, on the other hand, itself provided a model for historical explanations when Thomas Kuhn (see *Kuhn, Thomas S (1922–96)*) radically challenged traditional historiography with his notion of 'scientific revolution' conceived of as a general structure of knowledge development (Kuhn 1962).

2. Historical and Sociological Context

From a sociological perspective, the Scientific Revolution appears as part of a wider social process in which technical knowledge assumed a new role in the organization of European societies (see *Renaissance*; *Enlightenment*). This process took off in the late Middle Ages and was primarily rooted in the larger cities, which saw an ever more diversified and developed artisanal culture and a growing accumulation of merchandise capital. The cities of early modern Europe thus offered favorable conditions for the rapid growth of technical knowledge and the reflection of this growth in political, philosophical and religious thinking. Large-scale ventures involving technical expertise, such as projects of military architecture, water regulation, military adventures, or seafaring expeditions (see *History of Technology*), involved types of resources, social mobility, and an outlook on the world available only in urban centers such as Florence, Venice, Paris and London, which in fact became, long after they had attained an outstanding economic role, also the nuclei of the Scientific Revolution. Historians such as Edgar Zilsel (Zilsel 2000) have considered the early modern engineering projects as a decisive condition for the practical orientation and empirical knowledge base that distinguish the science of this period from its medieval antecedents.

Beginning in the fifteenth century, ambitious practical ventures (such as the building of the cupola of Florence Cathedral, the search for a sea route to India, or the development of a new military technology) increasingly relied on expert knowledge comprising both logistic and technological competencies exceeding those of traditional practitioners and artisans. Such competencies could only be gained on the basis of broader reflection on the relevant practical and theoretical knowledge resources available. This reflection became the specialty of the new group of engineer-scientists such as Filippo Brunelleschi, Christopher Columbus, Leonardo da Vinci, Niccolo Tartaglia, and Galileo Galilei. While the states of the time (if not actually at war) competed with each other in the pursuit of practical ventures, for example, building projects and seafaring expeditions, the knowledge acquired in such ventures was nevertheless constantly spread among the engineer-scientists employed by them. In the social fabric of the early modern period, engineer-scientists occupied a place similar to that earlier conquered by Renaissance humanists, administrators and artists (see *Art History*). These practically oriented intellectuals were, as a rule, highly mobile, offering their services to whatever patronage was available. At the same time, they constituted, as a social group, a collective memory, accumulating and transmitting the new knowledge long before appropriate institutions of learning emerged and in spite of the frequent political and military turnovers of the time.

The characteristic features of the engineer-scientists and their work become understandable against the background of their uncertain social status and their dependence on the patronage of courts and city governments (see e.g., Biagioli 1993) with rapidly changing power structures. Examples are their incessant engagement with projects for potential future patrons; their usually unrealistic promises regarding the practical benefits of their theories, inventions or projects; the secretiveness with which they treated their discoveries; their frequent involvement in priority struggles; as well as the striving to ennoble their practical knowledge by the claim of creating 'new sciences.'

The social and political ambitions of the engineer-scientists were reflected in their pursuit of a literary culture of technical knowledge, largely emulating the humanist culture of the courts, including their reference to ancient Greek and Roman canons. Their contribution to the literary culture was in turn welcomed by those interested in challenging the transcendent, religious legitimization of the feudal order as an argument for the possibility of an immanent explanation of both the natural and the social world. This affinity, together with the fact that an all-encompassing explanation of the world on the basis of Aristotelian philosophy had been adopted as the official doctrine of the Catholic Church, brought the engineer-scientists almost unavoidably into conflict with its power structures (Feldhay 1995).

Other developments contributed to turning the growth of technological and scientific knowledge into a force driving profound changes in the established structures of European society. The invention of printing offered a revolutionary new means of dissemination that challenged the exclusiveness of a literary culture based on manuscripts. The new dissemination technique contributed to overcoming the traditional separation between various branches of practical knowledge, confined by a transmission process relying exclusively on participation and oral communication as well as restricted by guild regulations. It also bridged the social gulf between such practical knowledge and the theoretical knowledge transmitted via scholarly texts at universities, monasteries, and courts. As a result, knowledge resources of different provenance were integrated and became widely available. Together with the accumulation of new knowledge in the context of the great practical ventures of the time, this process formed part of a veritable explosion of knowledge, both in the sense of its expansion and of its spreading across traditional social barriers.

In reaction to this knowledge explosion and its growing significance for the functioning of the advanced European societies, new institutions of learning such as the Accademia del Cimento (1657), the Royal Society (1660), and the Académie Royale des Sciences (1666) emerged in the middle of the seventeenth century (see *Knowledge Societies*; *Scientific Academies, History of*). Traditional institutions, such as those of the Church, on the other hand, had to accommodate to the new situation, as may be illustrated by the prominent involvement of Jesuits in the scientific culture of the time (Wallace 1991). Parallel to this process of institutionalization, science had, towards the end of the period here under consideration, gradually emancipated itself from the expectation of immediate practical benefits and could increasingly be pursued for its own sake. In summary, as a result of the Scientific Revolution, not only the production and transmission of technological knowledge but also its representation by scientific theories became an essential factor in the social, economic, and cultural development of European societies. This is the context in which scientists such as Huygens, Leibniz, Hook, Newton, and Wallis, traditionally identified with the completion of the Scientific Revolution by the creation of classical terrestrial and celestial mechanics, achieved their celebrated results.

3. *Structures and Achievements*

The cognitive, social, and material structures and achievements of the Scientific Revolution have been extensively studied and discussed amid much contro-

versy. Recent scholarship in the context of a cultural history of science has emphasized the possibility of an integrative treatment of these various dimensions. From the point of view of the traditional history of ideas, the Scientific Revolution appears primarily as the renewal of a knowledge development going back to antiquity, as a renaissance of Greek science. In fact, the ancient tradition of mathematical science, and in particular the 'Elements' of Euclid, provided the protagonists of the Scientific Revolution with the canonical model for a mathematical theory, a model which they systematically applied to new areas such as ballistics (Tartaglia) and even ethics (Spinoza). But the ancient tradition also had in stock designs for a theory of nature not based on Aristotelian views which they could hence exploit in their struggle against scholasticism, in particular Platonism and atomism. The ancient tradition finally offered a substantial corpus of knowledge in such domains as geometry, mathematical astronomy, and mechanics, serving as a point of departure for new scientific endeavors. The perception of their work as a renewal of antique science was typical of the self-image of the protagonists of the Scientific Revolution, who honored each other by such titles as that of a 'new Archimedes.' In short, the characterization of the Scientific Revolution as a renaissance of antique science by historians of ideas such as Koyré is in agreement with the claims of contemporary scientists to have accomplished a radical break with Aristotelian scholasticism (Koyré 1978).

This characterization, however, is in apparent conflict with the results of studies inaugurated by scholars in the tradition of Duhem, pointing to a conceptual continuity between early modern science and its medieval predecessors (and even with contemporary scholasticism), in spite of the anti-Aristotelian polemics (Duhem 1996). Such studies have provided evidence that the intellectual means available to the engineer-scientists engaged in creating new sciences were essentially still rooted in traditional conceptual frameworks. For instance, Galileo Galilei and his contemporaries allowed their investigations of motion and mechanics to be shaped by such Aristotelian notions as the distinction between natural and violent motion and the assumption that violent motion is caused by a moving force (Damerow et al. 1992). Studies of medieval natural philosophy have revealed, on the other hand, that the advanced explanations of phenomena such as projectile motion, by which early modern scientists distinguished themselves from Aristotelian natural philosophy, make use of concepts of causation such as that of 'impetus' (or 'impressed force') and techniques for conceptualizing change such as Oresme's doctrine of the 'latitude of forms' that had already been developed by late antique or medieval commentators on Aristotle (Clagett 1968).

Such contrasting accounts of early modern science appear less incompatible if other dimensions of the development of scientific knowledge are taken into account. For instance, epistemological considerations have suggested that one should differentiate between the claim of Renaissance engineer-scientists to have created a new science or a new scientific method, on the one hand, and the knowledge base they shared with their contemporaries, on the other hand. Using the terminology introduced by Elkana (1988) we can say that their ambitious claim resulted from their 'image of knowledge,' which was determined by their fragile social status. This image is to be distinguished from the shared 'body of knowledge' comprising the antique and medieval heritage which determined what challenges they were able to master. The role of this shared knowledge for the Scientific Revolution has been analyzed not only from the point of view of the history of ideas but also from the viewpoint of its actual function in the social and material contexts of this revolution. It has thus turned out that the material culture of the Scientific Revolution decisively shaped the way in which the knowledge of antique and medieval science was taken up or newly interpreted.

It has become evident, for instance, that the knowledge resources available to the engineer-scientists, largely structured by traditional conceptual frameworks, were challenged by their application to the new objects of a rapidly expanding range of experience, acquired in the context of the practical ventures in which the engineer-scientists were engaged. While investigations of 'challenging objects' such as the pendulum, the trajectories of artillery, purified metals, the dissected human body, the terrestrial globe, and the planetary system often remained less successful than their protagonists hoped and claimed, they nevertheless triggered elaborations and modifications of these traditional frameworks creating the foundations for the accomplishments of the age of classical science. Columbus searched for the sea route to India, but discovered America. Kepler tried in the Pythagorean tradition to unravel the harmonies of the world but from the harmonies he found only three laws of planetary motion that became the starting point of Newtonian cosmology. Galileo explained ballistics and pendulum motion on the basis of the impetus concept, but in fact he contributed to a new mechanics, which expelled this concept once and for all from science. Harvey defended Aristotle's claim of the primacy of the heart among the organs, but became the founder of an anti-Aristotelian theory, a mechanistic medicine. All these achievements resulted from coping with challenging new objects while relying on essentially traditional intellectual means.

Classical science, often considered an accomplishment of the Scientific Revolution, was actually established for the most part only in the course of the eighteenth century. Even classical mechanics, the pilot and model science of the Scientific Revolution, assumed the formulation familiar from today's physics textbooks only in the aftermath of Newton's pion-

eering work. The characteristics of classical science, relatively stable theoretical frameworks and generally accepted standards for the production of knowledge serving as the canonical reference for a scientific community, were, in any case, not yet shared by the scientific endeavors of the Scientific Revolution.

The conceptual framework of classical science comprising basic concepts such as inertia, a new methodical canon including the experimental method and mathematical techniques such as the differential calculus, new images of knowledge such as the mechanization of the world view, and the new institutions of classical science such as the academies were nevertheless consequences of the Scientific Revolution, as a belated result of reflection on its accumulated experience.

See also: Evolution: Diffusion of Innovations; History of Science; History of Technology; Innovation, Theory of; Kuhn, Thomas S (1922–96); Physical Sciences: History and Sociology; Technological Innovation

Bibliography

Biagioli M 1993 *Galileo, Courtier: The Practice of Science in the Culture of Absolutism*. University of Chicago Press, Chicago

Butterfield H 1965 *The Origins of Modern Science 1300–1800*. The Free Press, New York

Clagett M H (ed.) 1968 *Nicole Oresme and the Medieval Geometry of Qualities and Motions*. University of Wisconsin Press, Madison, WI

Cohen H F 1994 *The Scientific Revolution: A Historiographical Inquiry*. University of Chicago Press, Chicago

Cohen I B 1985 *Revolution in Science*. Belknap Press, Cambridge, MA

Damerow P, Freudenthal G, MacLaughlin P, Renn J 1992 *Exploring the Limits of Preclassical Mechanics*. Springer, New York

Dijksterhuis E J 1986 *The Mechanization of the World Picture: Pythagoras to Newton*. Princeton University Press, Princeton, NJ

Duhem P 1996 *Essays in the History and Philosophy of Science*. Hackett, Indianapolis, IN

Elkana Y 1988 Experiment as a second order concept. *Science in Context* **2**: 177–96

Feldhay R 1995 *Galileo and the Church: Political Inquisition or Critical Dialogue?* Cambridge University Press, Cambridge, UK

Hall A R 1954 *The Scientific Revolution 1500–1800: The Formation of the Modern Scientific Attitude*. Longmans, Green, London

Koyré A 1978 *Galileo Studies*. Humanities Press, Atlantic Highlands, NJ

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Lindberg D C, Westman R S (eds.) 1990 *Reappraisals of the Scientific Revolution*. Cambridge University Press, Cambridge, UK

Osler M J (ed.) 2000 *Rethinking the Scientific Revolution*. Cambridge University Press, Cambridge, UK

Porter R, Teich M (eds.) 1992 *The Scientific Revolution in National Context*. Cambridge University Press, Cambridge, UK

Shapin S 1996 *The Scientific Revolution*. University of Chicago Press, Chicago

Wallace W A 1991 *Galileo, the Jesuits, and the Medieval Aristotle*. Variorum, Aldershot, UK

Zilsel E 2000 *The Social Origins of Modern Science*. Kluwer, Dordrecht, The Netherlands

P. Damerow and J. Renn

Scientometrics

1. Introduction

Scientometrics can be defined as the study of the quantitative aspects of scientific communication, R&D practices, and science and technology (S&T) policies. The objective is to develop indicators of the intellectual and social organization or the sciences using network relations between scientific authors and texts. The specialty has developed in relation to the increased capacities of computer storage and information retrieval of scientific communications (e.g., citation analysis). Archival records of scientific communications contain institutional address information, substantive messages (e.g., title words), and relational information from which one is able to reconstruct patterns and identify the latent characteristics of both authors and document sets. Using scientometric techniques, one is thus able to relate institutional characteristics at the level of research groups with developments at the level of scientific disciplines and specialties.

Citations, for example, can be used for retrieving documents on the basis of author names, or vice versa. The scientometric representation is formal: it remains in need of an interpretation. The focus on uncertainty contained in the distribution relates scientometrics additionally to the (neo-evolutionary) study of complex and adaptive systems. Simulation models are increasingly used for the study of the role of science-based technologies in innovation processes. However, the specialty remains data driven because of its mission to provide indicators to S&T policy processes and R&D management.

2. A Metric of Science?

In 1978, the journal *Scientometrics* was launched as a new medium to stimulate the quantitative study of scientific communication. Derek de Solla Price, one of the founding fathers of the specialty, proclaimed in his introduction the development of scientometrics as the

emergence of a 'relatively hard' social science. This claim has generated discussion from the very beginning of the specialty. In that same year (1978), leading sociologists of science published an edited volume entitled *Toward a Metric of Science: The Advent of Science Indicators*, dedicated to 'Paul F. Lazarsfeld (1901–76), Master of Quantitative and Qualitative Social Research' (Elkana et al. 1978).

The systematic comparison of science indicators across fields of science was made possible by the creation of the *Science Citation Index* by Eugene Garfield of the Institute of Scientific Information (Garfield 1979). A preliminary version of this index became available in 1962. The creation of the database has stimulated the development of new perspectives on studies in various traditions. For example, the growth of scientific disciplines and specialties can be discussed in quantitative terms using this database (e.g., Price 1963), 'invisible colleges' can be explained in terms of network structures (Crane 1969), and theories of citations can perhaps be developed and tested (cf. Braun 1998).

The development of the specialty went hand in hand with the need for a means of legitimating science policies (Wouters 1999). Narin (1976) elaborated an instrumentarium for the systematic development of the biennial series of *Science Indicators* which the National Science Foundation of the USA began providing in 1972. (In 1987, the name of this series was changed to *Science and Engineering Indicators*.) With the further development of S&T policies in other nation-states and with the gradual emergence of such policies at the European level, scientometrics became a booming business during the 1980s. By comparing radio-astronomy facilities at the international level, Martin and Irvine (1983) showed the feasibility of comparing research groups in terms of quantitative performance indicators.

In a special issue of *Social Studies of Science* about performance indicators, Collins (1985) raised the question of the unit of analysis: what is being assessed in terms of what? The authors, the papers, or the cognitions shaped in terms of sociocognitive interactions among authors and discourses? In relation to French traditions of linguistic analysis, Callon et al. (1983) proposed using words and their co-occurrences instead of citations and co-citations (Small 1973) as units of analysis. Citations can be considered as subtextual codifications, while words indicate the variation at the level of texts. Words may change both in meaning and in terms of their observable frequency distributions.

3. *Methodologies*

The availability of well-organized relational databases on an annual basis challenges the scientometrician to develop comparative statistics. How is the structure in each of these years to be depicted, and how may changes in structure be distinguished from and related to changes in the observable variation? Can the difference over time be identified as 'growth of science,' or is it mainly a difference between two measurement errors? How significant are differences when tested against simulation results?

In principle, the idea of a dynamic mapping of science requires an independent operationalization of structural (that is, latent) dimensions of the maps and observable variation which is to be pencilled into these maps. Science, however, develops not only in terms of the variation, but also by changing its structural dimensions. Because of the prevailing reflexivity within the science system, previous structures can be felt as constraints and at the same time be used as resources. The construction of structure may historically be stabilized, but reflexive actors are able to deconstruct and to assess the previous constructions with hindsight.

The methodological apparatus for the mapping of science in terms of multivariate statistics (multidimensional scaling, cluster analysis, etc.) has been a product of the 1980s (e.g., Van Raan 1988). The 1990s have provided the evolutionary turn: what does history mean in relation to (envisaged?) future options? How can the system itself be informed reflexively with respect to its self-organizing capacities (Leydesdorff 1995)? The relation to technometrics and the measurement of 'systems of innovation' has become central to a shifting research agenda. The development of the sciences has increasingly been contextualized in relation to science-based technologies and innovation systems (Gibbons et al. 1994).

3.1 *Comparative Statistics of Science*

Various methods for the mapping of science can be based on relational indicators such as citations, words, co-occurrences of each of these categories, etc. Clustering, however, requires the choice of a similarity criterion and a clustering algorithm. The distinction between positional (or factor analytical) and relational (or graph analytical) approaches is also relevant: the network is expected to contain an architecture in which actors have a position. A nearby position does not necessarily imply a relation. The methodological reflection may thus help to clarify the theoretical analysis (Burt 1982).

The mapping requires a specific perspective for the projection. Each perspective assumes a position or a window. If the multidimensional space is highly structured, the different positions may provide nearly incommensurable projections. In an article on the development of the relation between scientometrics and other subfields of S&T studies, Leydesdorff and Van den Besselaar (1997) showed that the mappings depicting science studies from the perspective of the journals *Scientometrics, Social Studies of Science*, and *Research Policy*, respectively, are increasingly different

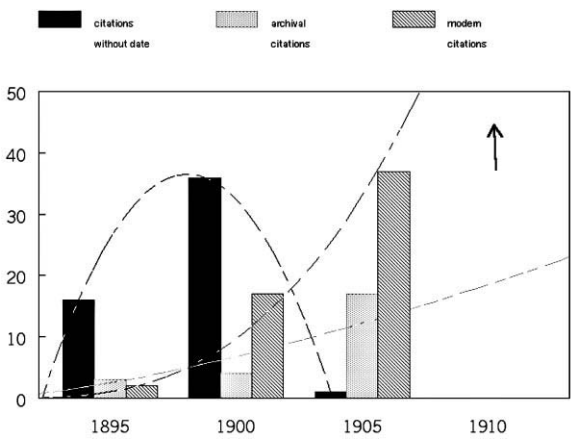

Figure 1
The emergence of the modern citation in the February issues of *Journal of the American Chemical Society* (1890–1910) (after Leydesdorff and Wouters 1999, p. 175).

during the 1990s. Citation relations among these core journals tend to decrease. The authors characterize *Social Studies of Science* as the 'codifier' of the field (along the historical axis), *Scientometrics* as the 'formalizer,' and *Research Policy* as the 'utilizer.' Only a few scholars in 'Science, Technology, and Innovation Studies' have developed competences for communicating across these subdisciplinary boundaries.

3.2 Time-series Methodologies

During the 1980s a debate raged in the community concerning the scientometric indication of a 'decline of British science.' Eventually, a special issue of *Scientometrics* was devoted to the debate in 1991 (Martin 1991). Some agreement could be reached that the inclusion and exclusion of data types *and* the framework for the comparison can be crucial for the dynamic evaluation. Should one compare with reference to a previous stage (for example, in terms of an *ex ante* fixed journal set), or should one with hindsight reconstruct the relevant data? For example, not only has the number of biotechnology journals changed, but also our understanding of 'biotechnology' is changing continuously. The sociological understanding of scientific knowledge production and control seems to have eroded Price's dream of developing scientometrics as a relatively 'hard social science.'

Using time-series analysis, one is always able to increase the fit of the curve by allowing for higher-order polynomials. Here again, the theoretical appreciation has to guide the choices of the parameters. For example, if one wishes to measure growth, it may be useful to include a second- or third-order polynomial in addition to the linear fit. Figure 1, for example, depicts the emergence of the modern citation itself as a historical phenomenon, but using theoretically informed helplines. Scientific citation emerged around the turn of the twentieth century as a means to search in both the textual and the social dimensions of science. Thus, the citation can be considered as an indicator of the complexity of sociocognitive interaction in science after its institutionalization in the nineteenth century.

3.3 Neo-evolutionary Methodologies

The networks of texts and the networks of authors operate upon each other in a selective mode. The distributions are expected to contain information, and this information may be increasingly codified by recurrent selections. As the systems 'lock-in' (in terms of their mutual information), the closure of the communication into a paradigm is one among various possibilities. The uncertainties which prevail in these networks can interactively generate codifications, which can be expected to perform a 'life-cycle.' Both participants and observers are able to hypothesize these structures reflexively, and new information can be expected to induce an update. Thus, the codification drives the knowledge production process. Codification also provides instruments for local control of otherwise global developments.

The relational operation is recursive. For example, citations refer to other texts and/or to citations in other texts. The networks resulting from this operation are expected to have an architecture (which can be mapped at each moment in time). Operations are expected to be reproduced if they are able to further the production of new knowledge and the latter's retention into cognitive structures. What is functional, however, is decided at a next moment in time, that is, at the level of a reflexive hypernetwork that overlays the historically generated ones. The distributions indicate the patterns of the expected operations. Thus, the process of new knowledge claims is propelled and made more precise and selective in tradeoffs of references in social and cognitive dimensions.

Scientometric studies can be helpful in revealing the patterns of intellectual and social organization which may have remained (partially) latent to the knowledgable actors involved. Simulation studies using scientometric mappings as input enable us to indicate the difference that the moves of the players can make. The complexity of the scientists' worlds is reflected in the scientometric reconstructions. The recognition of these objectified reconstructions recursively assumes and potentially refines the cognition within the discourses at both levels.

Over time, the cognitive reconstruction becomes thoroughly selective: citations may be 'obliterated by incorporation' into the body of knowledge, and social factors may play a role in further selections, e.g., in

terms of reputations. In this co-evolution between communications and authors, distributions of citations function, among other things, as contested boundaries between specialties. Since the indicators are distributed, the boundaries remain to be validated. Functions are expected to change when the research front moves further. By using references, authors position their knowledge claims within one specialty area or another. Some selections are chosen for stabilization, for example, when codification into citation classics occurs. Some stabilizations are selected for globalization at a next-order level, for example, when the knowledge component is integrated into a technology.

4. Conclusion

The focus on evolutionary dynamics relates scientometrics increasingly with the further development of evolutionary economics (Leydesdorff and Van den Besselaar 1994). How can systems of innovation be delineated? How can the complex dynamics of such systems be understood? How is the (potentially random) variation guided by previously codified expectations? How can explorative variation be increased in otherwise 'locked-in' trajectories of technological regimes or paradigms?

From this perspective, the indication of newness may become more important than the indication of codification. The Internet, of course, offers a research tool for what has also now been called 'sitations' (Rousseau 1997). 'Webometrics' may develop as a further extension of scientometrics relating this field with other subspecialities of science and technology studies, such as the public understanding of science or the appropriation of technology and innovation using patent statistics.

See also: Communication: Electronic Networks and Publications; History of Science; Libraries; Science and Technology, Social Study of: Computers and Information Technology; Science and Technology Studies: Experts and Expertise

Bibliography

Braun T (ed.) 1998 Topical discussion issue on theories of citation. *Scientometrics* **43**: 3–148

Burt R S 1982 *Toward a Structural Theory of Action*. Academic Press, New York

Callon M, Courtial J-P, Turner W A, Bauin S 1983 From translations to problematic networks: an introduction to co-word analysis. *Social Science Information* **22**: 191–235

Collins H M 1985 The possibilities of science policy. *Social Studies of Science* **15**: 554–8

Crane D 1969 Social structure in a group of scientists. *American Sociological Review* **36**: 335–352

Elkana Y, Lederberg J, Merton R K, Thackray A, Zuckerman H 1978 *Toward a Metric of Science: The Advent of Science Indicators*. Wiley, New York

Garfield E 1979 *Citation Indexing*. Wiley, New York

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M 1994 *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. Sage, London

Leydesdorff L, Van den Besselaar P (eds.) 1994 *Evolutionary Economics and Chaos Theory: New Directions in Technology Studies*. Pinter, London

Leydesdorff L, Van den Besselaar P 1997 Scientometrics and communication theory: Towards theoretically informed indicators. *Scientometrics* **38**: 155–74

Leydesdorff L, Wouters P 1999 Between texts and contexts: Advances in theories of citation? *Scientometrics* **44**: 169–182

Leydesdorff L A 1995 *The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications*. DSWO Press, Leiden University, Leiden, Netherlands

Martin B R 1991 The bibliometric assessment of UK scientific performance. A reply to Braun, Glänzel and Schubert. *Scientometrics* **20**: 333–57

Martin B R, Irvine J 1983 Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy* **12**: 61–90

Narin F 1976 *Evaluative Bibliometrics*. Computer Horizons Inc., Cherry Hill, NJ

Price D, de Solla 1963 *Little Science, Big Science*. Columbia University Press, New York

Raan A F J Van (ed.) 1988 *Handbook of Quantitative Studies of Science and Technology*. Elsevier, North-Holland, Amsterdam

Rousseau R 1997 Sitations: An exploratory study. *Cybermetrics* **1**: 1 http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Small H 1973 Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* **24**: 265–9

Wouters P 1999 The citation culture. PhD Thesis, University of Amsterdam

L. Leydesdorff

Screening and Selection

1. Introduction

In order to assist in selecting individuals possessing either desirable traits such as an aptitude for higher education or skills needed for a job or undesirable ones such as having an infection, illness or a propensity to lie, screening tests are used as a first step. A more rigorous selection device, e.g., diagnostic test or detailed interview is then used for the final classification.

For some purposes, such as screening blood donations for a rare infection, the units classified as positive are not donated but further tests on donors may not be given as most will not be infected. Similarly, for estimating the prevalence of a trait in a population, the screening data may suffice provided an appropriate estimator that incorporates the error rates of the test is used (Hilden 1979, Gastwirth 1987, Rahme and Joseph 1998). This article describes the measures of accuracy used to evaluate and compare screening tests and issues arising in the interpretation of the results. The importance of the prevalence of the trait on the population screened and the relative costs of the two types of misclassification are discussed. Methods for estimating the accuracy rates of screening tests are briefly described and the need to incorporate them in estimates of prevalence is illustrated.

2. *Basic Concepts*

The purpose of a screening test is to determine whether a person or object is a member of a particular class, C or its complement, $C-$. The test result indicating that the person is in C will be denoted by S and a result indicating non-membership by $S-$. The *accuracy* of a test is described by two probabilities:

$$\eta = \mathrm{P}[S \mid C]$$

being the probability that someone in C is correctly classified, or the sensitivity of the test; and

$$\theta = \mathrm{P}[S- \mid C-]$$

being the probability that someone not in C is correctly classified, or the specificity of the test. Given the prevalence, $\pi = \mathrm{P}(C)$, of the trait in the population screened, from Bayes' theorem it follows that the predictive value of a positive test (PVP) is

$$\mathrm{P}[C \mid S] = \pi\eta/[\pi\eta + (1-\pi)(1-\theta)].$$

Similarly, the predictive value of a negative test is

$$\mathrm{P}[C- \mid S-] = (1-\pi)\theta/[(1-\pi)\theta + \pi(1-\eta)].$$

In the first two sections, we will assume that the accuracy rates and the prevalence are known. When they are estimated from data, appropriate sampling errors for them and the PVP are given in Gastwirth (1987).

For illustration, consider an early test for HIV, having a sensitivity of 0.98 and a specificity of 0.93, applied in two populations. The first has a very low prevalence, 1.0×10^{-3}, of the infection while the second has a prevalence of 0.25. From Eqn. (1), the PVP in the first population equals 0.0138, i.e., only about one-and-one-half percent of individuals classified as infected would actually be so. Nearly 99 percent would be false positives. Notice that if the fraction of positives, expected to be 0.0797, in the screened data were used to estimate prevalence, a severe overestimate would result. Adjusting for the error rates yields an accurate estimate. In the higher prevalence group, the PVP is 0.8235, indicating that the test could be useful in identifying individuals. Currently-used tests have accuracy rates greater than 0.99, but even these still have a PVP less than 0.5 when applied to a low prevalence population. A comprehensive discussion is given in Brookmeyer and Gail (1994).

The role of the prevalence, also called the base rate, of the trait in the screened population and how well people understand its effect has been the subject of substantial research, recently reviewed by Koehler (1996). An interesting consequence is that when steps are taken to reduce the prevalence of the trait prior to screening, the PVP declines and the fraction of false-positives increases. Thus, when high-risk donors are encouraged to defer or when background checks eliminate a sizeable fraction of unsuitable applicants prior to their being subjected to polygraph testing the fraction of classified positives who are truly positive is small.

In many applications screening tests yield data that are ordinal or essentially continuous, e.g., scores on psychological tests or the concentration of HIV antibodies. Any value, t, can be used as a cut-off point to delineate between individuals with the trait and 'normal.' Each t generates a corresponding sensitivity and specificity for the test and the user must then incorporate the relative costs of the two different errors and the likely prevalence of the trait in the population being screened to select the cut off. The receiver operating characteristic (ROC) curve displays the trade-off between the sensitivity and specificity defined by various choices of t and also yields a method for comparing two or more screening tests.

To define the ROC curve, assume that the distribution of the measured variable (test score or physical quantity) is $F(x)$ for the 'normal' members of the population but is $G(x)$ for those with the trait. The corresponding density functions are $f(x)$ and $g(x)$ respectively and $g(x)$ will be shifted (to the right (left) if large (small) scores indicate the trait) of $f(x)$ for a good screening test. Often one fixes the probability ($\alpha = 1 - y\theta$ or one minus the specificity) of classifying a person without the characteristic as having it at a small value. Then t is determined from the equation $F(t) = 1 - \theta$. The sensitivity, η, of the test is $1 - G(t)$. The ROC curve plots η against $1 - \theta$. A perfect test would have $\eta = 1$ so the closer the ROC is to the upper left corner in Fig. 1, the better the screening test. In Fig. 1 we assume f is a normal density with mean 0 and variance 1 while g_1 is normal with mean 1 and the same variance. For comparison we also graphed the ROC for a second test, which has a density g_2 with mean 2 and variance 1. Notice that the ROC curve for the

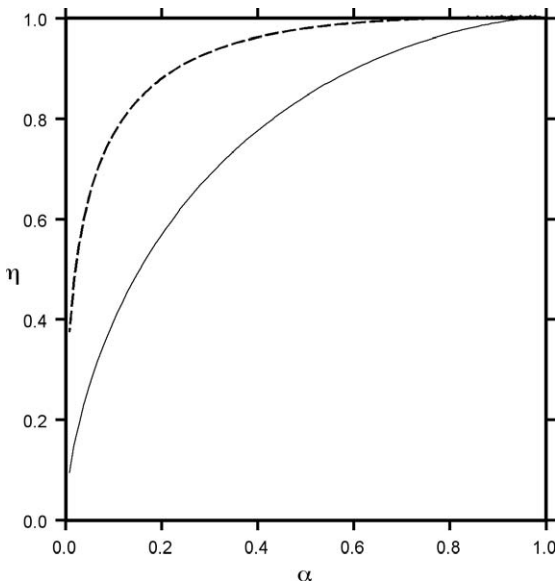

Figure 1
The ROC curves for screening tests 1 and 2. The solid line is the curve when the diseased group has mean 1 and the dashed curve is for the second group with mean 2

second test is closer to the left corner (0,1) than that of the first test.

A summary measure (Campbell 1994), which is useful in comparing two screening tests is the area, A, under the ROC. The closer A is to its maximum value of 1.0, the better the test is. In Fig. 1, the areas under the ROC curves for the two tests are 0.761 and 0.922, respectively. Thus, the areas reflect the fact that the ROC curve for the second test is closer to what a 'perfect' test would be. This area equals the probability a randomly chosen individual will have a higher score on the screening test than a normal one. This probability is the expected value of the Mann–Whitney form of the Wilcoxon test for comparing two distributions and methods for estimating it are in standard non-parametric statistics texts.

Non-parametric methods for estimating the entire ROC curve are given by Wieand et al. (1989) and Hilgers (1991) obtained distribution-free confidence bounds for it. Campbell (1994) uses the confidence bounds on F and G to construct a joint confidence interval for the sensitivity and one minus the specificity in addition to proposing alternative confidence bounds for the ROC itself. Hseih and Turnbull (1996) determine the value of t that maximizes the sum of sensitivity and specificity. Their approach can be extended to maximizing weighted average of the two accuracy rates, suggested by Gail and Green (1976). Wieand et al. (1989) also developed related statistics focusing on the portion of the ROC lying above a region, α_1 and α_2 so the analysis can be confined to values of specificity that are practically useful. Greenhouse and Mantel (1950) determine the sample sizes needed to test whether both the specificity and sensitivity of a test exceed pre-specified values.

The area A under the ROC can also be estimated using parametric distributions for the densities f and g. References to this literature and an alternative approach using smoothed histograms to estimate the densities is developed in Zou et al. (1997). They also consider estimating the partial area over the important region determined by two appropriate small values of α.

The tests used to select employees need to be reliable and valid. Reliability means that replicate values are consistent while validity means that the test measures what it should, e.g., successful academic performance. Validity is often assessed by the correlation between the test score (X) and subsequent performance (Y). Often X and Y can be regarded as jointly normal random variables, especially as monotone transformations of the raw scores can be used in place of them. If a passing score on the screening or pre-employment test is defined as $X \geqslant t$ and successful performance is defined as $Y \geqslant d$, then the sensitivity of the test is $\mathrm{P}[X > t \mid Y > d]$, the specificity is $\mathrm{P}[X < t \mid Y < d]$ and the prevalence of the trait is $\mathrm{P}[Y > d]$ in the population of potential applicants. Hence, the aptitude and related tests can be viewed from the general screening test paradigm.

When the test and performance scores are scaled to have a standard bivariate normal distribution, both the sensitivity and specificity increase with the correlation, ρ. For example, suppose one desired to obtain employees in the upper half of the performance distribution and used a cut-off score, t, of one-standard deviation above the mean on the test (X). When $\rho = 0.3$, the sensitivity is 0.217 while the specificity is 0.899. If $\rho = 0.5$, the sensitivity is 0.255 and the specificity is 0.937. The use of a high cut-off score eliminates the less able applicants but also disqualifies a majority of applicants who are in the upper half of the performance distribution. Reducing the cut-off score to one-half a standard deviation above the average raises the sensitivities to 0.394 and 0.454 for the two values of ρ but lowers the corresponding specificities to 0.777 and 0.837. This trade-off is a general phenomenon as seen in the ROC curves.

3. *The Importance of the Context in Interpreting the Results of Screening Tests*

In medical and psychological applications, an individual who tests positive for a disease or condition on a screening test will be given a more accurate confirmatory test or intensive interview. The cost of a 'false positive' screening result on a medical exam is

often considered very small relative to a 'false negative,' which could lead to the failure of suitable treatment to be given in a timely fashion. A false positive result presumably would be identified in a subsequent more detailed exam. Similarly, when government agencies give employees in safety or security sensitive jobs a polygraph test, the loss of potentially productive employee due to a false positive was deemed much less than the risk of hiring an employee who would might be a risk to the public or a security risk.

One can formalize the issue by including the costs of various errors and the prevalence, π, of the trait in the population being screened in determining the cut-off value of the screening test. Then the expected cost, which weights the probability of each type of error by its cost is given by

$$w\alpha(1-\pi)+(1-w)\pi G(t).$$

Here the relative costs of a false positive (negative) are w and $1-w$, respectively and as before $t = F^{-1}(1-\alpha)$. The choice of cut-off value, t_o, minimizing the expected cost satisfies:

$$\frac{g(t)}{f(t)} = \frac{w(1-\pi)}{(1-w)\pi} \tag{1}$$

Whenever the ratio, $g{:}f$, of the density functions is a monotone function the previous equation yields an optimum cut-off point, t_o, which depends on the costs and prevalence of the trait. Note that for any value of π, the greater the cost of a false positive, the larger will be the optimum value, t_o. This reflects the fact that the specificity needs to be high in order to keep the false positive rate low.

Although the relative costs of the two types of error are not always easy to obtain and the prevalence may only be approximately known, Eqn. (1) may aid in choosing the critical value. In practice, one should also assess the effect slight changes in the costs and assumed prevalence have on the choice of the cut-off value.

The choice of t that satisfies condition (1) may not be optimal if one desires to estimate the prevalence of the trait in a population rather than classifying individuals. Yanagawa and Tokudome (1990) determine t when the objective is to minimize the relative absolute error of the estimator of prevalence on the basis of the screening test results.

The HIV/AIDS epidemic raised questions about the standard assumptions about the relative costs of the two types of error. A 'false positive' classification would not only mean that a well individual would worry until the results of the confirmatory test were completed, they also might have social and economic consequences if friends or their employer learned of the result.

Similar problems arise in screening blood donors and in studies concerning the association of genetic markers and serious diseases. Recall that the vast majority of donors or volunteers for genetic studies are doing a public service and are being screened to protect others or advance knowledge. If a donation tests positive, clearly it should not be used for transfusion. Should a screened-positive donor be informed of their status? Because the prevalence of infected donors is very small, the PVP is quite low so that most of the donors screened positive are 'false.' Thus, blood banks typically do not inform them and rely on approaches to encourage donors from high-risk groups to exclude themselves from the donor pool (Nusbacher et al. 1986). Similarly, in a study (Hartge et al. 1998) of the prevalence of mutations in two genes that have been linked to cancer the study participants were not notified of their results.

The screening test paradigm is useful in evaluating tests used to select employees. The utility of a test depends on the costs associated with administering the test and the costs associated with the two types of error. Traditionally, employers focused on the costs of a false positive, hiring an employee who does not perform well, such as termination costs, and the possible loss of customers. The costs of a false negative are more difficult to estimate.

The civil-rights law, which was designed to open job opportunities to minorities, emphasized the importance of using appropriate tests, i.e., tests that selected better workers. Employers need to check whether the tests or job requirements (e.g., possession of a high school diploma) have a disparate impact upon a legally protected group. When they exclude a significantly greater fraction of minority members than majority ones, the employer needs to validate it, i.e., show it is predictive of on the job performance. Arvey (1979) and Paetzold and Willborn (1994) discuss these issues.

4. *Estimating the Accuracy of the Screening Tests*

So far, we have assumed that we can estimate the accuracy of the screening tests on samples from two populations where the true status of the individuals is known with certainty. In practice, this is often not the case and can lead to biased estimates of the sensitivity and specificity of a screening test, as some of the individuals believed to be normal have the trait, and vice versa.

If one has samples from only one population to which to apply both the screening and confirmatory test, then one cannot estimate the accuracy rates. The data would be organized into a 2×2 table, with four cells, only three of which are independent. There are five parameters, however, the two accuracy rates of the two tests plus the prevalence of the trait in the

population. In some situations, the prevalence of the trait may vary amongst sub-populations. If one can find two such sub-populations and if the accuracy rates of both tests are the same in both of those sub-populations, then one has two 2×2 tables with six independent cells, with which to estimate six parameters. Then estimation can then be carried out (Hui and Walter 1980).

This approach assumes that the two tests are conditionally independent given the true status of the individual. When this assumption is not satisfied, Vacek (1985) showed that the estimates of sensitivity and specificity of the tests are biased. This topic is an active area of research, recently reviewed by Hui and Zhou (1998). A variety of latent-class models have been developed that relax the assumption of conditional independence (see Faraone and Tsuang 1994 and Yang and Becker 1997 and the literature they cite).

5. *Applications and Future Concerns*

Historically screening tests were used to identify individuals with a disease or trait, e.g., as a first stage in diagnosing medical or psychological conditions or select students or employees. They are being increasingly used, often in conjunction with a second, confirmatory test, in prevalence surveys for public health planning. The techniques developed are often applicable, with suitable modifications, to social science surveys.

Some examples of prevalence surveys illustrate their utility. Katz et al. (1995) compared two instruments for determining the presence of psychiatric disorders in part to assess the needs for psychiatric care in the community and the available services. They found that increasing the original cut-off score yielded higher specificity without a substantial loss of sensitivity. Similar studies were carried out in Holland by Hodiamont et al. (1987) who found a lower prevalence (7.5 percent) than the 16 percent estimate in York. The two studies, however, used different classification systems illustrating that one needs to carefully examine the methodology underlying various surveys before making international comparisons. Gupta et al. (1997) used several criteria based on the results of an EKG and an individual's medical history to estimate the prevalence of heart disease in India.

Often one has prior knowledge of the prevalence of a trait in a population, especially if one is screening similar populations on a regular basis, as would be employers, medical plans, or blood centers. Bayesian methods incorporate this background information and can yield more accurate estimates (see Geisser 1993, and Johnson, Gastwirth and Pearson 2001). A cost-effective approach is to use an inexpensive screen at a first stage and retest the positives with a more definitive test. Bayesian methodology for such studies was developed by Erkanli et al. (1997).

The problem of misclassification arises often in questionnaire surveys. Laurikka et al. (1995) estimated the sensitivity and specificity of self-reporting of varicose veins. While both measures were greater than 0.90, the specificity was lower (0.83) for individuals with a family history than those with negative histories. Sorenson (1998) observed that often self-reports are accepted as true and found that the potential misclassification could lead to noticeable (10 percent) errors in estimated mortality rates. The distortion misclassification errors can have on estimates of low prevalence traits because of the high fraction of false positive classifications, was illustrated in the earlier discussion of screening tests for HIV/AIDS. Hemenway (1997) applies these concepts to demonstrate that surveys typically overestimate rare events; in particular the self-defense uses of guns. Thus it is essential to incorporate the accuracy rates into the prevalence estimate (Hilden 1979, Gastwirth 1987, Rahme and Joseph 1998).

Sinclair and Gastwirth (1994) utilized the Hui-Walter paradigm to assess the accuracy of both the original and re-interview (by supervisors) classifications in labor force surveys. In its evaluation, the Census Bureau assumes that the re-interview data are correct; however, those authors found that both interviews had similar accuracy rates. In situations where one can obtain three or more classifications, all the parameters are identifiable (Walter and Irwig 1988) Gastwirth and Sinclair (1998) utilized this feature of the screening test approach to suggest an alternative design for judge–jury agreement studies that had another expert, e.g., law professor or retired judge, assess the evidence.

6. *Conclusion*

Many selection or classification problems can be viewed from the screening test paradigm. The context of the application determines the relative costs of a misclassification or erroneous identification. In criminal trials, society has decided that the cost of an erroneous conviction far outweighs the cost of an erroneous acquittal. While, in testing job applicants, the cost of not hiring a competent worker is not as serious. The two types of error vary with the threshold or cut-off value and the accuracy rates corresponding to these choices is summarized by the ROC curve.

There is a burgeoning literature in this area as researchers are incorporating relevant covariates, e.g., prior health status or educational background into the classification procedures. Recent issues of *Biometrics and Multivariate Behavioral Research, Psychometrika* and *Applied Psychological Measurement* as well as the medical journals cited in the article contain a variety of

articles presenting new techniques and applications of them to the problems discussed.

See also: Selection Bias, Statistics of

Bibliography

Arvey R D 1979 *Fairness in Selecting Employees*. Addison-Wesley, Reading, MA

Brookmeyer R, Gail M H 1994 *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, New York

Campbell G 1994 Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* **13**: 499–508

Erkanli A, Soyer R, Stangl D 1997 Bayesian inference in two-phase prevalence studies. *Statistics in Medicine* **16**: 1121–33

Faraone S V, Tsuang M T 1994 Measuring diagnostic accuracy in the absence of a gold standard. *American Journal of Psychiatry* **151**: 650–7

Gail M H, Green S B 1976 A generalization of the one-sided two-sample Kolmogorov–Smirnov statistic for evaluating diagnostic tests. *Biometrics* **32**: 561–70

Gastwirth J L 1987 The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data (with discussion). *Statistical Science* **2**: 213–38

Gastwirth J L, Sinclair M D 1998 Diagnostic test methodology in the design and analysis of judge–jury agreement studies. *Jurimetrics Journal* **39**: 59–78

Geisser S 1993 *Predictive Inference*. Chapman and Hall, London

Greenhouse S W, Mantel N 1950 The evaluation of diagnostic tests. *Biometrics* **16**: 399–412

Gupta R, Prakash H, Gupta V P, Gupta K D 1997 Prevalence and determinants of coronary heart disease in a rural population of India. *Journal of Clinical Epidemiology* **50**: 203–9

Hartge P, Struewing J P, Wacholder S, Brody L C, Tucker M A 1999 The prevalence of common *BRCA1* and *BRCA2* mutations among Ashkenazi jews. *American Journal of Human Genetics* **64**: 963–70

Hemenway D 1997 The myth of millions of annual self-defense gun uses: A case study of survey overestimates of rare events. *Chance* **10**: 6–10

Hilden J 1979 A further comment on 'Estimating prevalence from the results of a screening test.' *American Journal of Epidemiology* **109**: 721–2

Hilgers R A 1991 Distribution-free confidence bounds for ROC curves. *Methods and Information Medicine* **30**: 96–101

Hodiamont P, Peer N, Syben N 1987 Epidemiological aspects of psychiatric disorder in a Dutch health area. *Psychological Medicine* **17**: 495–505

Hseih F S, Turnbull B W 1996 Nonparametric methods for evaluating diagnostic tests. *Statistics Sinica* **6**: 47–62

Hui S L, Walter S D 1980 Estimating the error rates of diagnostic tests. *Biometrics* **36**: 167–71

Hui S L, Zhou X H 1998 Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* **7**: 354–70

Johnson W O, Gastwirth J L, Pearson L M 2001 Screening without a gold standard: The Hui–Walter paradigm revisited. *American Journal of Epidemiology* **153**: 921–4

Katz R, Stephen J, Shaw B F, Matthew A, Newman F, Rosenbluth M 1995 The East York health needs study—Prevalence of DSM-III-R psychiatric disorder in a sample of Canadian women. *British Journal of Psychiatry* **166**: 100–6

Koehler J J 1996 The base rate fallacy reconsidered: descriptive, normative and methodological challenges (with discussion). *Behavioral and Brain Sciences* **19**: 1–53

Laurikka J, Laara E, Sisto T, Tarkka M, Auvinen O, Hakama M 1995 Misclassification in a questionnaire survey of varicose veins. *Journal of Clinical Epidemiology* **48**: 1175–8

Nusbacher J, Chiavetta J, Naiman R, Buchner B, Scalia V, Horst R 1986 Evaluation of a confidential method of excluding blood donors exposed to human immunodeficiency virus. *Transfusion* **26**: 539–41

Paetzold R, Willborn S 1994 *Statistical Proof of Discrimination*. Shepard's/McGraw Hill, Colorado Springs, CO

Rahme E, Joseph L 1998 Estimating the prevalence of a rare disease: adjusted maximum likelihood. *The Statistician* **47**: 149–58

Sinclair M D, Gastwirth J L 1994 On procedures for evaluating the effectiveness of reinterview survey methods: Application to labor force data. *Journal of American Statistical Assn* **91**: 961–9

Sorenson S B 1998 Identifying Hispanics in existing databases. *Evaluation Review* **22**: 520–34

Vacek P M 1985 The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* **41**: 959–68

Walter S D, Irwig L M 1988 Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *Journal of Clinical Epidemiology* **41**: 923–37

Wieand S, Gail M H, James B R, James K 1989 A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**: 585–92

Yanagawa T, Tokudome S 1990 Use of screening tests to assess cancer risk and to estimate the risk of adult T-cell leukemia/lymphoma. *Environmental Health Perspectives* **87**: 77–82

Yang I, Becker M P 1997 Latent variable modeling of diagnostic accuracy. *Biometrics* **52**: 948–58

Zou K H, Hall W J, Shapiro D E 1997 Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* **16**: 214–56

J. L. Gastwirth

Search, Economics of

The economics of search study the implications of market frictions for economic behavior and market performance. 'Frictions' in this context include anything that interferes with the smooth and instantaneous exchange of goods and services. The most commonly-studied problems arise from imperfect information about the location of buyers and sellers, their prices, and the quality of the goods and services that they trade. The key implication of these frictions is that individuals are prepared to spend time and other resources on exchange; they *search* before buying or selling. The labor market has attracted most

theoretical and empirical interest in this area of research, because of the heterogeneities that characterize it and the existence of good data on flows of workers and jobs between activity and inactivity, which can be used to test its propositions.

1. Historical Background

The first formal model of individual behavior, due to Stigler (1961), was in the context of a goods market: choosing the optimal number of sellers to search before buying at the lowest price. Stigler's rule, known as a fixed-sample rule, was abandoned in favor of sequential stopping rules: choosing the optimal reservation price and buying at the first store encountered which sells at or below the reservation price (see McCall 1970 for an early influential paper).

The first big momentum to research in the economics of search came with the publication of Phelps et al. (1970), which showed that search theory could be used to analyze the natural rate of unemployment and the inflation-unemployment trade off, the central research questions of macroeconomics at that time. Although interest in the inflation-unemployment trade-off has since waned, interest in search theory as a tool to analyze transitions in the labor market and equilibrium unemployment has increased. The momentum in this direction came in the 1980s, when contributions by Diamond (1982), Mortensen (1982) and Pissarides (1985) showed that search theory could be used to construct equilibrium models of the labor market with more accurate predictions than the traditional neoclassical model (see Pissarides 2000, Mortensen and Pissarides 1999a, 1999b for reviews). The appearance of comprehensive data on job and worker flows, which can be studied with the tools of search theory, also contributed to interest in this direction (see Leonard 1987, Dunne et al. 1989, Davis et al. 1996, Blanchard and Diamond 1990). This article reviews major developments since the mid 1980s, with explicit reference to labor markets.

2. Job Search

An individual has one unit of labor to sell to firms, which create jobs. The valuation of labor takes place under the assumptions that agents have infinite horizons and discount future income flows at the constant rate r, they know the future path of prices and wages and the stochastic processes that govern the arrival of trading partners, and they maximize the present discounted value of expected incomes. Let U_t be the expected present discounted value of a unit of labor before trade at time t (the 'value' of an unemployed worker) and W_t the expected value of an employed worker. During a short time interval δt the unemployed worker receives income $b\delta t$, and a job offer arrives with probability $a\delta t$. The frictions studied in the economics of search are summarized in the arrival process. In the absence of frictions, $a \to \infty$. With frictions and search, $a > 0$; with no search, $a = 0$. A large part of the literature is devoted to specifying the arrival process, an issue addressed in Sect. 3.1. The choice of search intensity has also been studied, by making a an increasing function of search effort, but this issue is not addressed here (see Pissarides 2000, Chap. 5).

If a job offer arrives, the individual has the option of taking it, for an expected return $W_{t+\delta t}$; or not taking it and keeping instead return $U_{t+\delta t}$. If no offer arrives, the individual's return is $U_{t+\delta t}$. Therefore, with discount rate r, U_t satisfies the Bellman equation

$$U_t = b\delta t + a\delta t \frac{\max(W_{t+\delta t}, U_{t+\delta t})}{1+r\delta t} + (1-a\delta t)\frac{U_{t+\delta t}}{1+r\delta t} \quad (1)$$

Rearrangement of terms yields

$$rU_t = b + a(\max(W_{t+\delta t}, U_{t+\delta t}) - U_{t+\delta t}) + \frac{U_{t+\delta t} - U_t}{\delta t} \quad (2)$$

Taking the limit of (2) as $\delta t \to \infty$, and omitting subscripts for convenience, yields

$$rU = b + a(\max(W, U) - U) + \dot{U} \quad (3)$$

where $\dot{U}$ denotes the rate of change of U.

Equation (3) is a fundamental equation in the economics of search. It can be given the interpretation of an arbitrage equation for the valuation of an asset in a perfect capital market with risk-free interest rate r. This asset yields coupon payment b and at some rate a, it gives its holder the option of a discrete change in its valuation, from U to W. Optimality requires that the option is taken (and the existing valuation given up) if $W \geqslant U$. The last term, $\dot{U}$, shows capital gains or losses due to changes in the market valuation of the asset. In most of the economics of labor-market search, however, research concentrates on 'steady states,' namely, on situations where the discount rate, transition rates, and income flows are all constant. With infinite horizons there are then stationary solutions to the valuation equations, obtained from (3) with $\dot{U} = 0$.

One simple way of solving (3) is to assume that employment is an 'absorbing state,' so when a job that offers wage w is accepted, it is kept for life. Then, $W = w/r$, and if the individual is sampling from a known wage offer distribution $F(w)$, the stationary version of (3) satisfies

$$rU = b + a(\int \max(w/r, U)\, \mathrm{d}F(w) - U) \quad (4)$$

The option to accept a job offer is taken if $w/r \geqslant U$, giving the *reservation wage* eqn.

$$\xi = rU \tag{5}$$

The reservation wage is defined as the minimum acceptable wage, and it is obtained as the solution to (4) and (5).

Partial models of search and empirical research in the duration of unemployment have explored generalized forms of Eqn. (4) to derive the properties of transitions of individuals from unemployment to employment (see Devine and Kiefer 1991). For a known $F(w)$ with upper support A, Eqn. (4) specializes to

$$\xi = b + \frac{a}{r}\int_{\xi}^{A}(w-\xi)\,\mathrm{d}F(w) \tag{6}$$

Various forms of (6) have been estimated by the empirical literature or used in the construction of partial models of the labor market. The transition from unemployment to employment (the unemployment 'hazard rate') is $a(1-F(\xi))$ and it depends both on the arrival of offers and on the individual's reservation wage. a, r, and the parameters of the wage offer distribution can be made to depend on the individual's characteristics. The empirical literature has generally found that unemployment compensation acts as a disincentive on individual transitions through the influence of b on the reservation wage, but the effect is not strong. The offer arrival rate increases the hazard rate, despite the fact that the reservation wage increases in a. A number of personal characteristics influence reservation wages and transitions, including age, education, and race.

3. Two-sided Matching for Given Wages

Recent work in the economics of search has focused mainly on the equilibrium implications of frictions and search decisions. An equilibrium model needs to specify the decisions of firms and solve for the offer arrival rate a. In addition, a mechanism is needed to ensure that search is an ongoing process. The latter is achieved by introducing a probability $\lambda\delta t$ that a negative shock will hit a job during a short time interval δt. When the negative shock arrives the job is closed down ('destroyed'), and the worker has to search again to find another job. For the moment, λ is assumed to be a positive constant (see Sect. 5).

3.1. The Aggregate Matching Function

To derive the equilibrium offer arrival rate, suppose that at time t there are u unemployed workers and v vacant jobs. In a short time interval δt each unemployed worker moves to employment with probability $a\delta t$, so in a large market the total flow of workers from unemployment to employment, and the total flow of jobs from vacant state to production, are both $au\delta t$. A key assumption in the equilibrium literature is that the total flows satisfy an *aggregate matching function*. The aggregate matching function is a black box that gives the outcome of the search process in terms of the inputs into search. If the u unemployed workers are the only job seekers and they search with fixed intensity of one unit each, and firms also search with fixed intensity of one unit for each job vacancy, the matching function gives

$$m = m(u, v) \tag{7}$$

with m standing for the flow of matches, au. The function is usually assumed to be continuous and differentiable, with positive first partial derivatives and negative second derivatives, and to satisfy constant returns to scale (see Petrongolo and Pissarides 2000 for a review). A commonly-used matching function in the theoretical literature, derived from the assumption of uncoordinated random search, is the exponential

$$m = v(1-\mathrm{e}^{-ku/v}), \quad k > 0 \tag{8}$$

The empirical literature, however, estimates a log-linear (constant elasticity) form, which parallels the Cobb–Douglas production function specification, with the elasticity on unemployment estimated in the range 0.5–0.7.

The fact that job matching is pairwise implies that the transition rates of jobs and workers are related Poisson processes. Given $au = m$, the rate at which workers find jobs is $a = m/u$. If q is the rate of arrival of workers to vacant jobs, then total job flows $vq = au$, and so $q = m/v$. The equilibrium literature generally ignores individual differences and treats the average rates m/u and m/v as the rates at which jobs and workers, respectively, arrive to each searching worker and vacant job. By the properties of the matching function,

$$q = m\left(\frac{u}{v}, 1\right) \tag{9}$$

$$\equiv m(\theta^{-1}, 1) \equiv q(\theta) \tag{10}$$

with $q'(\theta) < 0$ and elasticity $-\eta(\theta) \in (-1, 0)$. Here, θ is a measure of the *tightness* of the market, of the ratio of the inputs of firms into search to the inputs of workers. Similarly, the transition rate of workers is

$$a = m\left(1, \frac{v}{u}\right) \equiv \theta q(\theta) \tag{11}$$

By the elasticity properties of $q(\theta)$, $\partial a/\partial\theta > 0$. In the steady state, the inverse of the transition rates, $1/q(\theta)$ and $1/\theta q(\theta)$, are the expected durations of a vacancy and unemployment respectively.

The influence of tightness on the transition rates is independent of the level of wage rates. If there are more vacant jobs for each unemployed worker, the arrival rate of workers to the typical vacancy is lower and the arrival rate of job offers to the typical unemployed worker is higher, irrespective of the level of wages. When a worker and a firm meet and are considering whether or not to stay together, they are not likely to take into account the implications of their action for market tightness and the transition rates of other unmatched agents. For this reason, the influence of tightness on the transition rates is known as a *search externality*. Several papers in the economics of search have explored the efficiency properties of equilibrium given the existence of the externality (see Sect. 4.3.2 and Diamond 1982, Mortensen 1982, Pissarides 1984, Hosios 1990).

3.2 Job Creation

A job is an asset owned by the firm and is valued in a perfect capital market characterized by the same risk-free interest rate r. Suppose that in order to recruit a worker, a firm has to bear set-up cost K to open a job vacancy and in addition has to pay a given flow cost c for the duration of the vacancy. The flow cost can be interpreted as an advertising and recruitment cost, or as the cost of having an unfilled position in what might be a complex business environment (which is not modeled).

Let V be the value of a vacant position and J the value of a filled one. Reasoning as in the case of the value of a job seeker, U, the Bellman equation satisfied by V is

$$rV = -c + q(\theta)(J - V) \tag{12}$$

The vacant job costs c and the firm is offered the option to take a worker at rate $q(\theta)$.

Since the firm has to pay set-up cost K to create a job, it will have an incentive to open a job if (and only if) $V \geqslant K$. The key assumption made about job creation is that the gains from job creation are always exhausted, so jobs are created up to the point where

$$V = K \tag{13}$$

Substitution of V from (12) into (13) yields

$$J = K + \frac{rK + c}{q(\theta)} \tag{14}$$

Competition requires that the expected present discounted value of profit when the worker arrives, the value of a filled job J, should be just sufficient to cover the initial cost K and the accumulated costs for the duration of the vacancy, interest on the initial outlay rK and ongoing costs c for the expected duration of the vacancy, $1/q(\theta)$.

Let productivity be a constant p in all jobs and the wage rate a constant w. With break-up rate λ, the value of a job satisfies the Bellman equation

$$rJ = p - w - \lambda J \tag{15}$$

The flow of profit to the firm is $p - w$ until a negative shock arrives that reduces its value to 0. Replacing J in (14) by its expression in (15) yields the *job creation condition*

$$\frac{p - w}{r + \lambda} - K - \frac{rK + c}{q(\theta)} = 0 \tag{16}$$

Equation (16) determines θ for each w and parallels the conventional labor demand curve. A higher wage rate makes it more expensive for firms to open jobs, leading to lower market tightness. Frictions slow down the arrival of suitable workers to vacant jobs and so the firm incurs some additional recruitment costs. If the arrival rate is infinitely fast, as in Walrasian economics, $q(\theta)$ is infinite and the last term in (16) disappears. The assumptions underlying (13) ensure that at the margin, the recruitment cost is just covered.

4. Wage Setting

With search frictions, there is no supply of labor that can be equated with demand to give wages. The conventional supply of labor is constant here: there is a fixed number of workers in the market, which is usually normalized to unity, and each supplies a single unit of labor.

In search equilibrium there are local monopoly rents. Firms and workers who are together can start producing immediately. If they break up, they can start producing only after they find another partner through an expensive process of search. The value of the search costs that they save by staying together correspond to a pure economic rent; they could be taken away from them and they would still stay together. Wages need to share those rents.

Two approaches dominate in the literature. The first and more commonly-used approach employs the solution to a Nash bargaining problem (see Diamond 1982, Pissarides 2000). The Nash solution allocates the rents according to each side's 'threat points.' The threat points in this case are the returns from search, and the Nash solution gives to each side the same surplus over and above their expected returns from search. Generalizing this solution concept, wages are determined such that the net gain accruing to the

worker from the match, $W-U$, is a fixed proportion β of the total surplus, the sum of the worker's and the firm's surplus, $J-V$. This sharing rule can be obtained as the maximization of the product

$$(W-U)^{\beta}(J-V)^{1-\beta}, \quad \beta \in (0, 1) \tag{17}$$

The coefficient β can be given the interpretation of bargaining strength, although strictly speaking, bargaining strength in the conventional Nash solution is given by the threat points U and V (see Binmore et al. 1986 for an interpretation of β in terms of rates of time preference).

The second approach to wage determination postulates that the firm 'posts' a wage rate for the job, which the worker either takes or leaves. The posted wage may be above the worker's reservation wage because of some 'efficiency wage' type arguments, for example, in order to reduce labor turnover, encourage more effort or attract more job applicants.

4.1 Bargaining

The value of an employed worker, W, satisfies the Bellman eqn.

$$rW = w - \lambda(W-U) \tag{18}$$

The worker earns wage w and gives up the employment gain $W-U$ when the negative shock arrives. No worker has an incentive to quit into unemployment or search for another job whilst employed, for as long as $w \geqslant b$, which is assumed.

With (15) and (17) in place, the Nash bargaining solution to (18) gives the sharing rule

$$W-U = \beta(W-U+J-V). \tag{19}$$

Making use of the value equations and sharing rule yields

$$w = rU + \beta(p-(r+\lambda)K - rU) \tag{20}$$

$$= (1-\beta)b + \beta[p-(r+\lambda)K + (rK+c)\theta] \tag{21}$$

There is a premium on the reservation wage which depends on the worker's bargaining strength and the net surplus produced. Wages depend positively on unemployment income and the productivity of the job, the first because of the effect that unemployment income has on the cost of unemployment and the second because of the monopoly rents and bargaining. Wages also depend on market tightness. In more tight markets they are higher, because the expected duration of unemployment in the event of disagreement is less. Empirical evidence supporting this wage equation has been found by a number of authors (e.g., Blanchflower and Oswald 1994), although it should be noted that similar wage equations can also be derived from other theoretical frameworks.

4.2 Equilibrium

Equation (21) replaces the conventional labor supply curve and closes the system. When combined with the job creation condition (16) it gives unique solutions for wages and market tightness. A variety of intuitive properties are satisfied by this equilibrium. For example, higher labor productivity implies higher wages and tightness; higher unemployment income implies higher wages but lower tightness. It remains to obtain the employment rate in equilibrium.

The labor force size is fixed, so by appropriate normalizations, if at some time t unemployment is u_t, employment is $1-u_t$. In a short time interval δt, $a_t u_t \delta t$ workers are matched and $\lambda(1-u_t)\delta t$ workers lose their jobs. Given that $a_t = \theta_t q(\theta_t)$, the evolution of unemployment is given by

$$u_{t+\delta t} = u_t + \lambda(1-u_t)\delta t - \theta_t q(\theta_t) u_t \delta t \tag{22}$$

Dividing through by δt and taking the limit as $\delta t \to 0$ yields

$$\dot{u} = \lambda(1-u) - \theta q(\theta) u \tag{23}$$

Because the solution for θ is independent of u, this is a stable differential equation for unemployment with a unique equilibrium

$$u = \frac{\lambda}{\lambda + \theta q(\theta)} \tag{24}$$

Equation (24) is often referred to as the Beveridge curve, after William Beveridge who first described such a 'frictional' equilibrium. Plotted in space with vacancies on the vertical axis and unemployment on the horizontal is a convex-to-the origin curve. In the early literature the impact of frictions on the labor market was measured by the distance of this curve from the origin (see also Pissarides 2000, Blanchard and Diamond 1989).

4.3 Wage Posting

Wage posting is an alternative to wage bargaining, with different implications for search equilibrium. The firm posts a wage for the job and the worker who searches it either takes it or leaves it. Three different models of wage posting are examined.

4.3.1 Single offers, no prior information. Workers search sequentially one firm at a time, they discover the firm's offer after they have made contact and they have to accept it or reject it (with no recall) before they can sample another firm. The worker who contacts a firm that posts wage w_i has two options. Accept the wage offer of the firm and enjoy expected return W_i, obtained from (18) for $w = w_i$, or reject it and continue search for return U. The firm that maximizes profit chooses w_i subject to $W_i \geqslant U$. Since the firm will have no incentive to offer the worker anything over and above the minimum required to make workers accept its offer, in this model wages are driven to the worker's reservation wage, $w_i = b$ (Diamond 1971).

In terms of the bargaining solution, the 'Diamond' (see (20), (21)) equilibrium requires $\beta = 0$. The model can then be solved as before, by replacing β by 0 in the job creation condition and wage equation. Tightness, and consequently unemployment, absorb all shocks other than those operating through unemployment income, which also change wages. (There is a paradox in this model, in that if wages are equal to unemployment income no worker will have an incentive to search.)

4.3.2 Competitive search. The next model increases the amount of information that workers have before they contact the firm (Moen 1997). Workers can see the wage posted by each firm but because of frictions they cannot be certain that they will get the job if they apply. If the probability of a job offer across firms is the same, all workers will apply for the highest-wage job. Queues will build up and this will reduce the probability of an offer. In equilibrium, the wage offer and queue characterizing each job have to balance each other out, so that all firms get applicants.

The length of the queue is derived from the matching process. The firm that posts wage w_i makes an offer on average after $1/q(\theta_i)$ periods and a job applicant gets an offer on average after $1/\theta_i q(\theta_i)$ periods. Implicit in this formulation is the assumption that more than one firm offers the same wage, and firms compete for the applicants at this wage. Workers apply to only one job at a time.

Suppose now there is a firm, or group of firms, such that when workers join their queue they derive expected income stream $\bar{U}$, the highest in the market. The constraint facing a firm when choosing its wage offer is that the worker who applies to it derives at least as much expected utility as $\bar{U}$. The expected profit of the firm that posts wage w_i solves

$$rV_i = -c + q(\theta_i)\,(J_i - V_i) \tag{25}$$

$$rJ_i = p - w_i - \lambda J_i \tag{26}$$

The worker's expected returns from applying to the firm posting this wage satisfy the system of equations

$$rU_i = b + \theta_i q(\theta_i)(W_i - U_i) \tag{27}$$

$$rW_i = w_i - \lambda(W_i - \bar{U}) \tag{28}$$

The firm chooses w_i to maximize V_i subject to $U_i \geqslant \bar{U}$. The first-order maximization conditions imply that all firms offer the same wage, which satisfies

$$W - U = \frac{\eta}{1-\eta}(J - V) \tag{29}$$

Comparison of (29) with (19) shows that the solution is indeed similar to the Nash solution but with the share of labor given by the (negative of the) elasticity of $q(\theta)$, (which is equal to the unemployment elasticity of the underlying matching function). The rest of the model can be solved as in the Nash case.

There is a special significance to the share of labor obtained in this formulation. In the case where wages are determined according to the Nash rules, the firm and worker choose the wage after they meet, so it is unlikely that they will internalize the search externalities; they do not take into account the effect of their choices on the transition rates of unmatched agents. It can be shown that with constant returns to scale there is a unique internalizing rule which requires $\beta = \eta$, the solution of the wage posting model considered here (Hosios 1990, Pissarides 1984, 2000). For this reason, this particular wage posting model is often called the *competitive search* equilibrium. The key assumption that gives efficiency is the relaxation of the informational restrictions on workers, which lets them know both the firm's wage offer and the length of the queue associated with it.

4.3.3 Wage differentials. The third version of the wage posting model relaxes the assumption that workers have the choice of at most one wage offer at a time but does not allow workers knowledge of the wage offer before they apply (Burdett and Judd 1983, Burdett and Mortensen 1998, Montgomery 1991). The easiest way to introduce this is to allow workers the possibility of search on the job, i.e., to let them continue looking for a better job after they accepted one. Suppose for simplicity that job offers arrive to searching workers at the same rate a, irrespective of whether they are employed or unemployed. Suppose also that job search is costless (except for the time cost) and job changing is costless. Then, if a worker is earning w now and another offer paying w' comes along, the worker accepts the new offer if (and only if) $w' > w$. The worker's reservation wage is the current wage.

Unemployment pays b, so no firm can pay below b and attract workers. Consider a firm that pays just above b. Anyone applying for a job from the state of unemployment will accept its offer, but no one else will. Employed job seekers will have no incentive to quit their jobs to work at a wage close to b. In addition, this firm's workers will be quitting to join other firms, which may be paying above b. So a firm paying a low wage will have high turnover and will be waiting long before it can fill its vacancies.

A firm paying a high wage will be attracting workers both from unemployment and from other firms paying less than itself. Moreover, it will not be losing workers to other firms. So high-wage firms will have fewer vacant positions.

Now suppose the two firms have access to the same technology. The low-wage firm enjoys a lot of profit from each position, but has a lot of vacant positions. The high-wage firm enjoys less profit from each position, but has them filled. It is possible to show under general conditions that high wage and low wage firms will co-exist in equilibrium.

Burdett and Mortensen (1998) show this by assuming that there is a distribution of wage offers for homogeneous labor, $F(w)$, and demonstrating that (a) no two firms will offer the same wage, (b) firms can choose any wage between the minimum b and a maximum $\bar{w}$, and enjoy the same profit in the steady state. The maximum is given by

$$\bar{w} = p - \left(\frac{\lambda}{a+\lambda}\right)^2 (p-b) \tag{30}$$

where as before p is the productivity of each worker and λ the job destruction rate.

The interesting result about wage posting in this model is that once the assumption that the worker can only consider one offer at a time is relaxed, a distribution of wage offers for homogeneous labor arises. The distribution satisfies some appealing properties. As $a \to 0$, the upper support tends to $\bar{w} = b$, the Diamond solution. At the other extreme, as $a \to \infty$, it can be shown that all wage offers converge to p, the competitive solution. Intuitively, $a = 0$ maximizes the frictions suffered by workers and $a = \infty$ eliminates them.

5. Job Destruction

A body of empirical literature shows that there is a lot of 'job churning,' with many jobs closing down and new ones opening to take their place. It is also found that both the job creation and job destruction rates, especially the latter, vary a lot over the cycle (see Leonard 1987, Dunne et al. 1989, Davis et al. 1996). This is an issue addressed by search theorists. Returning to the model with a Nash wage rule, the job creation flow is the matching rate $m(u, v)$. The job creation *rate* is defined as the job creation flow divided by employment, $m(u, v)/(1-u)$, which is

$$\frac{m(u, v)}{1-u} = \theta q(\theta)\frac{u}{1-u} \tag{31}$$

Since both θ and u are endogenous variables of the model that respond to shocks, the model predicts a variable job creation rate that can be tested against the data (see Mortensen and Pissarides 1994, Cole and Rogerson 1996).

But the job destruction flow is $\lambda(1-u)$, so the job destruction rate is a constant λ, contrary to observation. Two alternative ways of making it variable are considered.

5.1 Idiosyncratic Shocks

In the discussion so far, the productivity of a job is p until a negative shock arrives that reduces it to zero. Generalizing this idea, suppose that although initially the productivity is p, when a shock arrives it changes it to some other value px. The component p is common to all jobs but x is specific to each job. It has distribution $G(x)$ in the range [0, 1]. New jobs start with specific productivity $x = 1$; over time shocks arrive at rate λ that transform this productivity to a value between 0 and 1, according to the distribution G. The firm has the choice of either continuing to produce at the new productivity, or closing the job down. The idea is that initially firms have a choice over their product type and technique and choose the combination that yields maximum productivity. Over time techniques are not reversible, so if the payoffs from a given choice change, the firm has the choice of either continuing in the new environment or destroying the job.

As in the case where workers are faced with take-it-or-leave-it choices from a given wage distribution, Mortensen and Pissarides (1994) show that when the firm has the choice of either taking a productivity or leaving it, its decision is governed by a reservation productivity, denoted R. The reservation productivity depends on all the parameters of the model. Profit maximization under the Nash solution to the wage bargain implies that both workers and firms agree about the optimal choice of R, i.e., which jobs should be closed and which should continue in operation. With knowledge of R, the flow of job closures generalizes to $\lambda G(R)(1-u)$, the fraction of jobs that get shocks below the reservation productivity. The job destruction rate then becomes $\lambda G(R)$, which responds to the parameters of the economy through the responses of R to shocks.

An interesting feature of the job creation and job destruction rates, which conforms to observation, is

that because the job creation rate depends on unemployment, which is a slow-moving variable, whereas the job destruction rate does not depend on it, the job destruction rate is more volatile than the job creation rate. For example, a rise in general productivity p, associated with a positive cyclical shock, reduces job destruction immediately by reducing the reservation productivity, but increases job creation at first and then reduces it, as tightness rises at first but unemployment falls in response.

5.2 Technological Progress and Obsolescence

Another way of modeling job destruction borrows ideas from Schumpeter's theory of growth through creative destruction (see Aghion and Howitt 1994, Caballero and Hammour 1994). Jobs are again created at the best technology but once created, their technology cannot be updated. During technological progress, the owners of the jobs have the option of continuing in production with the initial technology or closing the job down and opening another, more advanced one. As new jobs are technologically more advanced, the wage offers that workers can get from outside improve over time. There comes a time when the worker's outside options have risen sufficiently to render the job obsolete.

Formally, the model can be set up as before, with the technology of the job fixed at the frontier technology at creation time and wages growing over time because of growth in the returns from search. The value of a job created at time 0 and becoming obsolete at T is

$$J_0 = \int_0^T \mathrm{e}^{-(r+\lambda)t}\,[p(0)-w(t)]\,\mathrm{d}t \tag{32}$$

As before, r is the discount rate and λ is the arrival rate of negative shocks that may lead to earlier job destruction. $p(0)$ is the initial best technology and $w(t)$ the growing wage rate. The job is destroyed when $w(t)$ reaches $p(0)$; i.e., the job life that maximizes J_0 is defined by $w(T^*) = p(0)$.

A useful restriction to have when there is growth is to assume that both unemployment income and the cost of recruitment grow at the exogenous rate of growth of the economy. As an example, consider the wage eqn. (21) with the restriction $K = 0$ and $b(t) = b\mathrm{e}^{gt}$, $c(t) = c\mathrm{e}^{gt}$, where g is the rate of growth of the economy. T^* then satisfies

$$\mathrm{e}^{gT^*} = \frac{(1-\beta)p}{(1-\beta)b+\beta c\theta} \tag{33}$$

Jobs are destroyed more frequently when growth is faster, when unemployment income is higher and when the tightness of the market is higher.

Job destruction in this model has two components, the jobs destroyed because of the arrival of shocks, $\lambda(1-u)$, and those destroyed because of obsolescence. The latter group was created T^* periods earlier and survived to age T^*, $\theta q(\theta)u\mathrm{e}^{-\lambda T^*}$ (note that in the steady state the job creation flow is a constant $\theta q(\theta)u$). Therefore, the job destruction rate now is $\lambda + \theta q(\theta)u\mathrm{e}^{-\lambda T^*}/(1-u)$, which varies in response to changes in T^* but also in response to changes in the job creation rate $\theta q(\theta_u)/(1-u)$.

See also: Behavioral Economics; Consumer Economics; Consumption, Sociology of; Economics: Overview; Information, Economics of; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Supply; Market Research; Market Structure and Performance; Stigler, George Joseph (1911–91); Transaction Costs and Property Rights; Wage Differentials and Structure; Work, Sociology of

Bibliography

Aghion P, Howitt P 1994 Growth and unemployment. *Review of Economic Studies* **61**: 477–94

Binmore K G, Rubinstein A, Wolinsky A 1986 The Nash bargaining solution in economic modelling. *Rand Journal of Economics* **17**: 176–88

Blanchard O J, Diamond P A 1989 The Beveridge curve. *Brookings Papers on Economic Activity* **1**: 1–60

Blanchard O J, Diamond P A 1990 The cyclical behavior of the gross flows of US workers. *Brookings Papers on Economic Activity* **2**: 85–155

Blanchflower D G, Oswald A J 1994 *The Wage Curve*. MIT Press, Cambridge, MA

Burdett K, Judd K 1983 Equilibrium price distributions. *Econometrica* **51**: 955–70

Burdett K, Mortensen D T 1998 Wage differentials, employer size, and unemployment. *International Economic Review* **39**: 257–73

Caballero R J, Hammour M L 1994 The cleansing effect of recessions. *American Economic Review* **84**: 1350–68

Cole H L, Rogerson R 1999 Can the Mortensen–Pissarides matching model match the business cycle facts? *International Economic Review* **40**: 933–59

Davis S J, Haltiwanger J C, Schuh S 1996 *Job Creation and Destruction*. MIT Press, Cambridge, MA

Devine T J, Kiefer N M 1991 *Empirical Labor Economics: The Search Approach*. Oxford University Press, Oxford

Diamond P A 1971 A model of price adjustment. *Journal of Economic Theory* **3**: 156–68

Diamond P A 1982 Wage determination and efficiency in search equilibrium. *Review of Economic Studies* **49**: 217–27

Dunne T, Roberts M J, Samuelson L 1989 Plant turnover and gross employment flows in the manufacturing sector. *Journal of Labor Economics* **7**: 48–71

Hosios A J 1990 On the efficiency of matching and related models of search and unemployment. *Review of Economic Studies* **57**: 279–98

Leonard J S 1987 In the wrong place at the wrong time: The extent of frictional and structural unemployment. In: Lang K, Leonard J S (eds.) *Unemployment and the Structure of Labor Markets*. Basil Blackwell, New York

McCall J J 1970 Economics of information and job search. *Quarterly Journal of Economics* **84**: 113–26
Moen E R 1997 Competitive search equilibrium. *Journal of Political Economy* **105**: 385–411
Montgomery J 1991 Equilibrium wage dispersion and inter-industry wage differentials. *Quarterly Journal of Economics* **106**: 163–79
Mortensen D T 1982 The matching process as a non-cooperative/bargaining game. In: McCall J J (ed.) *The Economics of Information and Uncertainty*. University of Chicago Press, Chicago, IL
Mortensen D T, Pissarides C A 1994 Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* **61**: 397–415
Mortensen D T, Pissarides C A 1999a Job reallocation, employment fluctuations, and unemployment. In: Woodford M, Taylor J (eds.) *Handbook of Macroeconomics*. North-Holland, Amsterdam
Mortensen D T, Pissarides C A 1999b New developments in models of search in the labor market. In: Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*. North-Holland, Amsterdam
Petrongolo B, Pissarides C A 2000 Looking into the black box: A survey of the matching function. *Journal of Economic Literature* **39**: 390–431
Phelps E S et al. 1970 *Microeconomic Foundations of Employment and Inflation Theory*. Norton, New York
Pissarides C A 1984 Efficient job rejection. *Economic Journal* **94**: 97–108
Pissarides C A 1985 Short-run equilibrium dynamics of unemployment, vacancies, and real wages. *American Economic Review* **75**: 676–90
Pissarides C A 2000 *Equilibrium Unemployment Theory*, 2nd edn. MIT Press, Cambridge, MA
Stigler G J 1961 The economics of information. *Journal of Political Economy* **69**: 213–25

C. A. Pissarides

Second Language Acquisition

Humans are all born with the capacity to learn and to use a language; but they are not born with a language. It is not part of human genetical endowment that 'horse' means 'equine quadruped,' that the past tense is marked by '-ed,' or that the negation follows the finite verb; this knowledge must be derived from the input with which the learner is confronted. The ways which lead this innate language faculty to the knowledge of a particular linguistic system vary considerably, depending on factors such as age, nature of input and whether this task is undertaken for the first time ('first language acquisition,' FLA) or not ('second language acquisition,' SLA). SLA is not a homogeneous phenomenon, for at least two reasons. First, it need not wait until the learner has completed FLA; hence, there is a continuous transition from bilingual FLA, in which a child is exposed more or less simultaneously to two systems from birth, to the adult's struggles with a new kind of linguistic input. Second, there is a wide range of ways in which the human language faculty gains access to a second linguistic system, ranging from metalinguistic description, as in traditional Latin classes, to language learning by everyday communication, as in the case of a foreign worker. In the history of mankind, explicit teaching of a language is a relatively late phenomenon, and untutored learning was, and probably still is, the most common case; but due to its practical importance, SLA in the classroom still dominates research.

Linguists and laymen alike tend to consider children's way to their mother tongue to be the most important type of language acquisition. This view seems most natural; but it leads easily to a distorted picture of how the human language faculty functions, and what its typical manifestations are. FLA is a very complex mixture of cognitive, social and linguistic developments, and it is not easy to isolate its purely linguistic components. The acquisition of the English tense and aspect system, for example, not only requires the learning of a particular mapping of forms and meanings, but also the development of the concept of time itself. Moreover, most people learn more than one language, albeit to different degrees of perfection. Therefore, the *normal* manifestation of the human language faculty is a 'learner variety,' i.e., a linguistic system which comes more or less close to the linguistic habits of a particular social group. In a child's case, the final learner variety is usually a 'perfect replication' of these habits; children who grow up in multilingual communities often achieve two or even three such perfect replications. Adults who set out to learn another language hardly ever reach a stage where they speak like those from whom they learn; their 'learner varieties' normally fossilize at an earlier stage. This does not mean that their final learner variety is less of a language, or less efficient; there is no reason to assume that a linguistic system which says 'He swam yesterday' is a superior manifestation of the human language faculty than a system which says 'He swimmed yesterday' or even 'He swim yesterday.' It is just the way the English do it, and deviations from their norms are stigmatized. If the study of language acquisition, and of SLA in particular, should inform us about the nature of the human language faculty, then it must not focus on issues of perfect replication and why it fails sometimes, but try to clarify how the human language faculty deals under varying conditions with particular forms of linguistic input to which it has access. The first step to this end is to isolate the crucial factors which play a role in this process, and to look at ways they can vary. The second step is to investigate what happens under varying constellations. The final step is to draw generalizations from these findings and to turn them into a theory not just of language acquisition, but the nature of human language itself (Klein 1986).

The picture which research on SLA offers at the time of writing is much less systematic. As with so many other disciplines, it has its origin in practical concerns; researchers were looking for scientific ways to improve foreign language teaching, and this seems impossible without a deeper understanding of the principles of SLA. Therefore, most empirical work in this field is still in the classroom. A second source of inspiration was research on FLA, which started much earlier and therefore set the theoretical and methodological stage. More recently, work in theoretical linguistics has increasingly influenced research on SLA. These and other influences, for example from cognitive and social psychology, resulted in a very scattered picture of theories, methods and findings. Rather than reviewing this research, the following discussion will concentrate on three key issues (useful surveys are found in Ellis 1994, Ritchie and Bhatia 1996, Mitchell and Myles 1998, Braidi 1999).

1. SLA and Foreign Language Instruction

The pedagogical background of SLA research has led naturally to a particular view on SLA, for which two assumptions are constitutive:

(a) There is a well-defined target of the acquisition process—the language to be learned. This *target language* is a clearly fixed entity, a structurally and functionally balanced system, mastered by those who have learned it in childhood, and more or less correctly described in grammars and dictionaries.

(b) SLA learners miss this target at varying degrees and in varying respects—they make errors in production as well as in comprehension, because they lack the appropriate knowledge or skills.

This is the *target deviation perspective*. It is the teacher's task to erase, or at least to minimize, the deviations; it is the researcher's task to investigate which 'errors' occur when and for which reasons. As a consequence, learners' performance in production or comprehension is not studied very much in its own right, as a manifestation of learning capacity, but in relation to a set norm; not in terms of what learners do, but in terms of what they fail to do. The learners' utterances at some time during the process of acquisition are considered to be more or less successful attempts to reproduce the structural properties of target language utterances. Learners try to do what the mature speaker does, but do it less well. Three reasons make the target deviation perspective so natural and attractive, in fact, almost self-evident. First, it is the natural perspective of the language teacher: language teaching is a normative process, and the teacher is responsible for moving students as closely to some norm as possible. Second, it is also the natural perspective of all of those who had to learn a second language in the classroom—and that means, also, of practically every language researcher. Third, the target deviation perspective provides the researcher with a simple and clear design for empirical work. There is a yardstick against which the learners' production and comprehension can be measured: the target language, or actually what grammar books and dictionaries say about it. What is measured is the differences between what learners do and what the set norm demands. Therefore, the dominant method in SLA research was, and is, error analysis: Learners' errors are marked and then either counted and statistically analyzed, or they are interpreted individually (Corder 1981, Ellis 1994, pp. 561–664).

There are two problems with this perspective. First, it does not tell us what learners *do* but what they are *not able* to do. Second, its results reflect not just the principles according to which the human language faculty functions, but the efficiency of a particular teaching method. Therefore, this approach may be of eminent importance to the language teacher, but it is of limited value if we want to understand the nature of human language.

2. FLA and SLA

Experience shows that FLA normally leads to 'perfect command' of the target language, whereas SLA hardly ever does. Why this difference? Is perfect attainment of a second language possible at all? Does the learning process only stop at an earlier point, or does it follow different principles?

The last question has found two opposite answers. The *identity hypothesis*, advocated by many researchers in the early 1970s, claims that the underlying processes are essentially the same across all types of acquisition. Under this view, the fact that the learner already knows a language plays no role: there is no transfer from the 'source language' (Odlin 1989). Evidence came mainly from the order in which certain grammatical phenomena, such as inflectional morphemes or the position of negation, are acquired. It turned out, however, that these similarities are quite isolated; there are hardly any supporters of the identity hypothesis anymore. Under the opposite view, it is mainly structural differences between source and target language that cause problems for learners. This *contrastive hypothesis* has given rise to a number of contrastive grammars for pedagogical purposes. But while there are many clear cases in which learners' first language interferes in the learning process, structural contrasts can at best account for some properties of the acquisitional process. In acquisition outside the classroom, for example, all learners regularly develop a particular type of 'learner variety' which is essentially independent of source and target language (see Klein and Perdue 1997). The net result of thirty years of research is simply that there are similarities as well as dissimilarities.

The varying success in final attainment could be due to (a) age differences, or (b) to the fact that there is already a language which blocks the acquisition of a

second language. The second possibility is ruled out by the fact that school age children normally have no problem in learning a second language to perfection; hence, the varying success must be an age effect. Apparently, the capacity to learn a language does not disappear, but it deteriorates with age. Since this capacity is stored in the brain, it seems plausible to assume that changes in the brain are responsible for the age effect. The clearest statement of this view is Lenneberg's theory of a biologically fixed 'critical period,' during which the brain is receptive for language; it ranges approximately from birth to puberty. After this period, linguistic knowledge can only be learned in a different form, roughly like the knowledge of historical or geographical facts (Lenneberg 1967). This theory has the seductive charm of simple solutions, and hence has been welcomed with great enthusiasm. But as far as is known, all potentially relevant changes in the brain occur in the first four years of life, rather than around puberty. Moreover, all available evidence shows that the capacity to learn a new language deteriorates only gradually; there is no clear boundary at puberty or at any other time. Finally, it could be shown that 'perfect attainment' is perhaps rare but definitely possible after puberty (see Birdsong 1999). It appears, therefore, that there is no clear biological threshold to language acquisition; the age effect is due to a much wider array of factors (Singleton 1989).

3. *SLA and Theoretical Linguistics*

The apparent ease and speed with which children, despite deviant and insufficient input, become perfect speakers of their mother tongue has led Noam Chomsky and other generative grammarians to assume that a great deal of the necessary linguistic knowledge is innate. Since every newborn can learn any language, this innate knowledge must be universal, and it is this 'universal grammar' (UG) which is the proper object of linguistic theory. Since languages also differ in some respects (otherwise, SLA would be superfluous), the competence of mature speakers is supposed to include a 'peripheral part,' which includes all idiosyncratic properties and must be learned by input analysis, and a 'core.' The core consists of a number of universal principles—the UG. Initially, these principles include a number of 'open parameters,' i.e., variable parts which must be fixed by input analysis. Chomsky made this point only for FLA, and only in the mid 1980s was the question raised whether UG is still 'accessible' in SLA.

A number of empirical studies tested the potential 'resetting' of various parameters. Spanish, for example, allows the omission of a subject pronoun, a property which is structurally linked to other features such as a relatively rich inflectional word order and relatively free word order; these and other properties form the 'pro-drop parameter.' English children have set this parameter the opposite way when acquiring their language. Are adult English learners of Spanish able to 'reset' it, or do they have to learn all of these properties by input analysis? Results are highly controversial (see e.g., Eubank 1991, Epstein et al. 1997). Although inspired by theoretical linguistics, most empirical research in this framework keeps the traditional 'target deviation perspective'; with only a few exceptions, it deals with acquisition in the classroom, hence reflecting the effects of teaching methods. Moreover, there is no agreement on the definition of the parameters itself; in fact, more recent versions of generative grammar have essentially abandoned this notion. Finally, it is an open issue as to which parts of linguistic knowledge form the core and which parts belong to the periphery, and hence must be learned from the input. These language-specific parts clearly include the entire lexicon, the inventory of phonemes, inflectional morphology, all syntactic properties in which languages can differ—in short, almost everything. It seems more promising, therefore, to look at how learners construct their learner varieties by input analysis.

4. *Learner Varieties*

The alternative to the target deviation perspective is to understand the learners' performance at any given time as an immediate manifestation of their capacity to speak and to understand: form and function of these utterances are governed by principles, and these principles are those characteristic of the human language faculty. Early attempts in this direction are reflected in notions such as 'interlanguage,' 'approximate systems' and so on. Since the 1980s, most empirical work on SLA outside the classroom has taken this 'learner variety perspective' (von Stutterheim 1986, Perdue 1993, Dietrich et al. 1995). In its most elaborate form, it can be characterized by three key assumptions (Klein and Perdue 1997).

(a) During the acquisitional process, learners pass through a series of learner varieties. Both the internal organization of each variety at a given time, as well as the transition from one variety to the next, are essentially systematic in nature.

(b) There is a small set of principles which are present in *all* learner varieties. The actual structure of an utterance in a learner variety is determined by a particular interaction of these principles. The kind of interaction may vary, depending on various factors, as the learner's source language. With ongoing input analysis, the interaction changes. Picking up some component of noun morphology from the input, for example, may cause the learner to modify the weight of other factors to mark the grammatical status of a noun phrase. Therefore, learning a new feature is not adding a new piece of puzzle which the learner has to

put together. Rather, it entails a sometimes minimal, sometimes substantial reorganization of the whole variety, where the balance of the various factors approaches the balance characteristic of the target language successively.

(c) Learner varieties are not imperfect imitations of a 'real language' (the target language), but systems in their own right. They are characterized by a particular lexical repertoire and by a particular interaction of structural principles. Fully developed languages, such as Spanish, Chinese or Russian, are only special cases of learner varieties. They represent a relatively stable state of language acquisition—that state where learners stop learning because there is no difference between their variety and the variety of their social environment, from which they get input.

Thus, the process of language acquisition is not to be characterized in terms of errors and deviations, but in terms of the twofold systematicity which it exhibits: the inherent systematicity of a learner variety at a given time, and the way in which such a learner variety evolves into another one. If we want to understand the acquisitional process, we must try to uncover this two fold systematicity, rather than look at how and why a learner misses the target.

See also: First Language Acquisition: Cross-linguistic; Foreign Language Teaching and Learning; Language Acquisition; Language Development, Neural Basis of

Bibliography

Birdsong D (ed.) 1999 *Second Language Acquisition and the Critical Period Hypothesis*. Erlbaum, Mahwah, NJ

Braidi S M 1999 *The Acquisition of Second Language Syntax*. Arnold, London

Corder P 1981 *Error Analysis and Interlanguage*. Oxford University Press, Oxford, UK

Dietrich R, Klein W, Noyau C 1995 *Temporality in a Second Language*. Benjamins, Amsterdam

Ellis R 1994 *The Study of Second Language Acquisition*. Oxford University Press, Oxford, UK

Epstein S, Flynn S, Martohardjono G 1997 Second language acquisition. Theoretical and experimental issues in contemporary research. *Behavioural and Brain Sciences* **19**: 677–758

Eubank L (ed.) 1991 *Point–Counterpoint Universal Grammar in the Second Language*. Benjamins, Amsterdam

Klein W 1986 *Second Language Acquisition*. Cambridge University Press, Cambridge, UK

Klein W, Perdue C 1997 The basic variety, or couldn't natural languages be much simpler? *Second Language Research* **13**: 301–47

Lenneberg E 1967 *Biological Foundations of Language*. Wiley, New York

Mitchell R, Myles F 1998 *Second Language Learning Theories*. Arnold, London

Odlin T 1989 *Language Transfer*. Cambridge University Press, Cambridge, UK

Perdue C (ed.) 1993 *Adult Language Acquisition Crosslinguistic Perspectives*. Cambridge University Press, Cambridge, UK, 2 Vols

Ritchie W C, Bhatia T (eds.) 1996 *Handbook of Second Language Acquisition*. Academic Press, New York

Singleton D 1989 *Language Acquisition: The Age Factor*. Multilingual Matters, Clevedon, UK

Stutterheim C von 1986 *Temporalität in der Zweitsprache* (Temporality in second language). De Gruyter, Berlin

W. Klein

Second World War, The

1. The Second World War: A Narrative

World War II was an event of massive significance. For at least fifty years after its end in 1945 it continued to condition societies and ideas throughout the world. Much of the politics of the second half of the twentieth century can be read as occurring in an 'after-war' context. The war exacted a death toll of at least 60 million, and probably tens of millions more than that (figures for China and the rest of Asia are mere guesses and the USSR's sacrifice has risen from seven to 20 to 29 or more million as time has passed, circumstances varied, and the requirements of history altered). A majority of the casualties were civilians, a drastic change from World War I when some 90 percent of deaths were still occasioned at the fronts. Moreover, the invention of the atom bomb during the war and its deployment by the USA at Hiroshima and Nagasaki (August 6 and 9, 1945) suggested that, in any future nuclear conflict, civilians would compose 90 percent or more of the victims. When this apparent knowledge was added to the revelations of Nazi German barbarism on the eastern front and the Nazis' massacre of European Jewry, either in pit killings or when deliberately transported to such death camps as Auschwitz-Birkenau, Treblinka, Sobibor, Chelmno, Belzec, and Majdanek, another casualty of the war seemed to be optimism itself. Certainly at 'Auschwitz' and perhaps at Hiroshima, 'civilization,' the modernity of the Enlightenment, the belief in the perfectibility of humankind, had led not to hope and life but instead to degradation and death.

This more or less open fearfulness, with its automatic resultant linking of a pessimism of the intellect to any optimism of the will among postwar social reformers, may be the grandest generalization that can be made about the meaning of World War II. Big history, however, should not lose sight of microhistory. Actually World War II was fought on many fronts, at different times, for different reasons, and with different effects. In this sense, there was a multiplicity of World War II's.

In September 1939, a war broke out between Nazi Germany and authoritarian Poland. The liberal democratic leadership of Britain and France intervened

saying that they would defend Poland, although in practice they made do with 'phoney war' until the Nazi forces took the military initiative, first in Denmark and Norway, and then in the Low Countries and France in April–May 1940. In June 1940, Fascist Italy entered the war as had been envisaged in the 'Pact of Steel' signed with its Nazi ally in May 1939. War now spread to the Italian empire in North and East Africa and, from October 1940, to the Balkans. Italian forces botched what they had hoped would be a successful Blitzkrieg against Greece and the effect over the next year was to bring most of the other Balkan states into the conflict. Often these fragile states dissolved into multiple civil wars, the most complicated and the one of most lasting significance began in Yugoslavia from March–April 1941.

On June 22, 1941, the Nazi Germans invaded the Soviet Union, commencing what was, in some eyes, the 'real' World War II, and certainly the one that was inspired by the most direct ideological impulse and which unleashed the most horrendous brutality. In the course of the campaign in the east it is estimated that the Germans sacked 1,710 towns and 70,000 villages. During the epic siege of Leningrad from 1941 to 1944, a million or so of the city's inhabitants starved to death. In their invasion, the Germans were joined by an assortment of anticommunist allies and friends, including military forces from authoritarian Romania and Fascist Italy. Many Lithuanians, Latvians, and Estonians, and quite a few anti-Soviet elements within the USSR (Ukrainian nationalists, people from the Caucasus, and others) acted an auxiliaries of Nazi power. The Nazis were even embarrassed by a 'Russian' army under General A. A. Vlasov, willing to fight on their side against Stalin and his system. Volunteers also came from pro-fascist circles in France, and from Spain and Portugal, states ruled by clerical and reactionary dictators who hated communists but were not fully reconciled to the radical thrust of much Nazi-fascist rhetoric and some Nazi-fascist policy.

On December 7, 1941, the war widened again when the Japanese airforce attacked the American Pacific fleet at anchor at Pearl Harbor in Hawaii. In the following weeks, the Japanese army and navy thrust south and east, dislodging the British from Singapore by January 15, 1942. They went on to seize the Philippines and Dutch East Indies (later Indonesia) and were in striking distance of Australia before being checked at the Battle of the Coral Sea in June 1942. They simultaneously continued the terrible campaign that they had been waging in China since 1937 (or rather since 1931, when they had attacked Manchukuo or Manchuria). In their special wars, the militarist Japanese leadership tried to throw off what they called the imperialist yoke of US capitalism and the older 'white' metropolitan empires. The purity of their anti-imperial motives was damaged, however, by their own commitment to empire and by their merciless killing of Asians. Moreover, as in Europe, their invasions often touched off varieties of civil war provoked by the highly complex stratification of society in the region. At the forefront of such campaigns were often local nationalists who imagined communities subservient neither to European powers nor the Japanese.

On December 11, 1941, Germany and Italy, loyal to the terms of the anti-Comintern pact, had also declared war on the USA, somewhat ironically so, given that Japan, checked by military defeats in an unofficial war against the USSR at Khalkin Gol and Nomonhan in 1938–9, had now engaged with other enemies. The Italians would find out the implications in North Africa, the Germans after the Allied invasion of France on 'D-Day' (June 6, 1944), as well as in Italy from September 8, 1943 (the Fascist dictator, Benito Mussolini, overthrown on July 25, was thereafter restored as a sort of German puppet in northern Italy; allied forces moved slowly up the peninsula from the south, liberating Rome on June 4, 1944 but only reaching Milan at the end of the war in late April 1945). Of the participants in the war, the USA, once fully mobilized, possessed the biggest and most productive economy, and was therefore of crucial importance in the eventual defeat of the anti-Comintern powers. The campaign the Americans fought with the most passion, and with an evident racism of their own, was the war against Japan. In another sense, the USA had a relatively soft war, not disputed on its own territory and not requiring the sort of physical or spiritual sacrifice obligatory from most other combatants. The USA's special World War II was not really a visceral one.

If the war was fought in many different ways, it is equally true that the variety of conflicts did not all end at the same time and in the same way. The Nazi armies surrendered on May 8, 1945, the Japanese on August 15. But matters were more complicated than that. France's special World War II would commemorate 'victory' from the date of the liberation of Paris on August 25, 1944 (and General Charles De Gaulle would firmly proclaim that Paris and France had liberated themselves). In most of Nazi-fascist occupied Europe and in parts of Asia, partisan movements had never altogether accepted defeat. Communists were invariably prominent in such resistance, even if quite a few still envisaged themselves as fighting as much for the Soviet revolution as for the liberty of their own nation state. Every successive 'liberation,' in Europe most frequently coming after the military victory of the Red Army, and in the Pacific that of the USA, had its own special character. Yugoslavia and China were two especially complicated places where the resistance was very strong but where it was contested, not only by the Nazi-fascists and the Japanese but also by local anti-communist and nationalist or particularist forces. The effects and memory of their wars were by definition to be very different from such societies as the USA, Australia, and the UK which did not endure foreign

occupation, though the last, with its severe experience of bombing, was itself different from the other two.

In sum, World War II was not just an enormously influential event but also an extraordinarily complicated one. Its complexity has in turn stimulated many passionate and long-lasting debates about its historical meaning.

2. *The Causes of War*

For many years it was customary to argue that World War II as compared to World War I had a simple cause. This was 'Hitler's war.' Mainstream analysis continues to ascribe to the Nazi dictator great responsibility for the invasion of Poland and for the spreading of the war thereafter, and especially for the launching of Operation Barbarossa against the USSR. Nonetheless, from the 1960s, the course of historiography, especially as exemplified in the rise of social history, did not favor a 'Great Man' view of the past and tended to urge that even dictators had limits to their free will. As early as 1964, English radical historian A. J. P. Taylor (1964) argued in a book entitled *The Origins of the Second World War* that the several crises which led up to September 1939 and what he sardonically called the 'War for Danzig' needed to be understood in a variety of contexts, including the peace settlements at the end of World War I, the course of German political and social history, the institutionalization of the Russian revolution, with its victorious but feared and paranoid communist and then Stalinist regime, and the lights and shadows of democratic liberalism in Western Europe.

Taylor wrote with a flaunted stylistic brilliance and practiced a brittle historical cleverness. He was destined to be misunderstood and often gloried in the misunderstanding. His book thus produced an enormous controversy, the first of many to be sparked by attempts to define the meaning of World War II. At the same time, Taylor's idiosyncrasies ensured that his work could easily enough be dismissed by those mainstream historians who liked to feel the weight of their commitment to Rankean principles and to make their liking evident.

Nonetheless the issues raised by Taylor did not go away. In West Germany, the so-called 'Fischer controversy,' sparked by the Hamburg liberal historian Fritz Fischer's *Griff nach der Weltmacht*, a massively documented study of German aims during World War I and also published in 1961, raged through the decade. Although Fischer, then and thereafter, wrote almost exclusively about World War I and about imperial Germany, he was read as commenting on World War II and, indeed, on Germany's divided fate in its aftermath. Two issues were prominent. Was imperial Germany an aggressive power in a way that bore comparison with the Nazi regime during the 1930s? Was the motive of the German leadership as much domestic as foreign—did they seek foreign adventure and even world war in order to divert the pressure building for social democracy? Was the appalling conflict from 1939 to 1945 caused by the 'German problem,' which may have begun as early as 1848 and may have continued after 1945?

Fischer's work was more directly influential than Taylor's and the issue of the relationship between *Innenpolitik* and *Aussenpolitik* fitted neatly into the preoccupations of new social historians who, by the 1970s, were scornfully dismissing the work of 'old fashioned diplomatic historians.' For all that, specialist works on the causation of World War II continued to privilege the power of Hitler and acknowledge the ideological thrust of Nazism. English independent Marxist, Tim Mason, may have tried to pursue a Fischerian line in asking how much the diplomatic crises of 1938–9 were prompted by the contradictions of Nazi economic and social policy, but his essays remained at the periphery of most analysis. Rather such firmly Rankean historians as Gerhard Weinberg and Donald Watt assembled evidence which, in their eyes, only confirmed that the war was caused by Hitler. The newest and most authoritative English-language biographer of the Führer, Ian Kershaw, despite his background in social history, does not disagree. Hitler may have been erratic as an executive. Nazi totalitarian state and society, contrary to its propaganda about militant efficiency and a people cheerfully bound into a *Volksgemeinschaft*, may in practice have often been ramshackle. But the dictator, Kershaw argued, did possess power. Indeed, so all-embracing was his will that Germans strove to 'work towards' their Führer, to accept his ideas and implement his policies before he had fully formulated them. For a generation in the wake of the Fischer controversy, scholarship on Nazism had separated into 'intentionalists' (advocates of Great Man history) and 'functionalists' (those who preferred to emphasize the role of structures and contexts and who were especially alert to the 'institutional darwinism' of the Nazi regime). Now Kershaw, often a historian of the golden mean, seemed to have found a way to resolve and end that conflict.

The only major variants on this reaffirmation that Hitler had provoked the war came from certain conservative viewpoints. Some anticommunist historians focused on the Ribbentrop–Molotov pact (August 23, 1939), placing responsibility for the war on 'Stalin' and the Russian revolution in what seemed to others a highly tendentious effort to blame the victim. More common was the view expressed most succinctly by Zionist historian Lucy Dawidowicz that the whole conflict was in essence a 'war against the Jews.' In this interpretation, German nationalism, German anticommunism, German racism towards Slavs, Nazi repression of the socialist and communist left, none in any sense equated with Hitlerian anti-

Semitism. Hitler (or, in the variant recently made notorious by Daniel Goldhagen, the Germans) wanted to kill Jews; that was the purpose of the Nazi regime; that was the aim of its wars.

In other ex-combatant societies, further examples of local focus are evident. In England, fans of appeasement still existed, the most prominent being John Charmley. For him, the problem with the war lay simply in Britain's engagement in it. As far as the British empire was concerned, Nazism did not need to be fought and the USSR should never have been an ally. Worst of all was the fact that the USA had dominated the post war world. Charmley has never quite said so, but implication of his work is that the 'real' World War II for Britain, and the one it most dramatically lost, was implicitly fought against its American 'cousins.'

The Asian-Pacific conflict has similarly been subject to historical revision. In the 1960s Gabriel Kolko and other American 'new leftists' applied a Marxian model to their nation's foreign policy, being very critical of the gap between its idealistic and liberal theory and its realist and capitalist practice. They were in their turn duly subjected to withering fire from more patriotic and traditional historians. Nonetheless a consensus grew that, at least in regard to the onset of the American–Japanese war, US policy makers carried some responsibility. By the 1980s, liberal historian John Dower was even urging that the two rivals had been 'enemies of a kind,' neither of them guiltless of racism and brutality.

In Japan, by contrast, an officially sponsored silence long hung over everything to do with the war. The Ministry of Education was particularly anxious that schoolchildren not be exposed to worrying facts about such terrible events as the massacre in Nanking in 1937, the practice of germ warfare, the exploitation of 'comfort women,' and the many other examples of Japanese murder, rape, and pillage in Asia and the Pacific. Nonetheless, a stubborn undercurrent of opinion exemplified in the work of historian Ienaga Saburo continued to contest the Ministry line and, by the 1990s, the Japanese leadership had gone further than ever before in admitting some of the misdeeds of its militarist predecessors. Fully critical history may still not be especially appreciated in Tokyo. But, by the end of the 1990s, Japan was not the only society to behave that way in that world in which the ideology of economic rationalism had achieved unparalleled hegemony backed by what American democratic historian Peter Novick has called 'bumper sticker' lessons from the past.

3. *The Course of the War*

In the preceding paragraphs it has not always been possible to keep fully separate discussions about the coming of World War II from what happened after hostilities commenced. In any case, military history in the pure sense, just like diplomatic history, after 1945 soon lost ground professionally. To most eyes, the military history of the war can swiftly enough be told. Of the anti-Comintern states, Germany and Japan but not Italy, won rapid initial victories, exulting in their respective Blitzkriegs. But their triumphs were always brittle. Germany and its allies may have been good at getting into wars, but their ideologies made it difficult for them thereafter to contemplate any policy except the complete liquidation of the enemy. Hitler and the Japanese imperial and military leadership thus made no attempts to offer compromise from a position of strength, and Mussolini's occasional flirtation with the idea always involved Nazi sacrifice in their wars, especially that in the east, rather than Italian loss.

Nor did the anti-Comintern states make the most of the huge territories, which they had conquered, and the immense material resources that they therefore controlled. Nazi Germany is something of a case study in this regard. In the West, where the war was always gentler, the Nazis found plenty of direct and indirect collaborators. They were thus, for example, able to harness a very considerable proportion of the French economy to their cause. They also started to construct a new economic order that was not utterly unlike some of the developments, which would occur in Western Europe after Nazi-fascism had been defeated. With extraordinary contradiction for a state built on an utter commitment to racial purity, Nazi Germany, already before 1939, needed immigrants to staff its economy. Once the war began, this requirement became still more pressing. One partial solution was to import workers by agreement with its ally Italy and by arrangement with the friendly Vichy regime in France. Not all such French 'guest-workers' came unwillingly and not all had especially bad wars. In Germany they joined other, more reluctant, immigrants from the east, who were often little more than slave laborers. Poles and Soviet prisoners of war constituted the majority of these; at first they were frequently worked to death. However, as the Nazi armies turned back and the war settled into one of attrition and retreat, as symbolized by the great defeat at Stalingrad (November 1942–January 1943), the Germans began to treat even laborers from the east in a way that allowed some minimum chance of their survival and which also permitted some tolerable productivity from their labor.

The exception was, of course, the Jews, who, from September–October 1941, became the objects of the 'Final Solution,' a devotion to murder confirmed by officials who attended the Wannsee conference in January 1942. In terms of fighting a war, the adoption of the policy of extermination was, of course, counterproductive in many senses, among which was the economic. In staffing and fueling the trains, which transported the Jews to the death camps, in the low

productivity of these and the other camps, the Nazis wasted resources needed at the front. Saul Friedlander will write elsewhere in this volume about the debates about the meaning of the Holocaust. It is worth noting in this segment, however, that each combatant society has argued about its particular experience of being visited by the Nazi war machine and therefore of being exposed to collaboration with it. Two important examples occurred in France and the USSR.

The spring of 1940 brought disaster to the Third French Republic. Now was the time of what Marc Bloch, one of the founders of the great structuralist historical school of the *Annales*, a patriotic Frenchman and a Jew, called the 'strange defeat.' Military historians have demonstrated that France was not especially inferior in armament to the invading Germans. Rather, France lost for reasons to do with morale and the domestic divisions of French society. As a result, by June 1940 the French state and empire had collapsed. During the next four years, the inheritance of the Third Republic was disputed between the Vichy regime headed by Marshal Pétain within the rump of French metropolitan territories, the 'Free French' under General Charles De Gaulle, resident in London, and there, however reluctantly and churlishly, dependent on Allied goodwill and finance, and a partisan movement which gradually became more active in the occupied zones. This last was typically divided between communists and other forces, some of which disliked communists as much as they hated the invaders. The years of Axis occupation were thus also the time of the 'Franco-French civil war,' with killings and purge trials extending well beyond liberation.

The meaning of war, occupation, and liberation in France has been much disputed after 1945. It took a film maker, Marcel Ophuls in *Le Chagrin et la pitié* (1971), and an American historian, Robert Paxton, to break a generation of silence about the troubling implications of this period of national history, although Socialist President François Mitterrand (1981–95), with his own equivocal experience of Vichy, was scarcely an unalloyed advocate of openness during his term in office. Historian Henry Rousso has brilliantly examined the 'Vichy syndrome' and done much to expose some of the obfuscations favored by many different leadership groups in post-1945 French politics. Perhaps his work does not go far enough, however. With its fall, and then with decolonization after 1945, France had lost its political empire, promising to become just another European state. However, this loss was curiously compensated by the rise and affirmation of French culture. In almost every area of the humanities, such French intellectuals as De Beauvoir, Braudel, Barthes, Foucault, Lévi-Strauss, Lyotard, Baudrillard, and Nora, charted the way to postmodernity. They did so sometimes invoking the pro-Nazi philosopher Martin Heidegger and almost always without much reckoning of the collapse of the French nation state in 1940. The intellectuals of France were much given to forgetting their nation's fall, the better to affirm their own rights to cultural imperium.

If France provides a case study of the transmutation of defeat into victory, the USSR offers the reverse example of a victor whose people would eventually learn that 'actually' they had lost. The nature of the Soviet war effort is still in need of research. Sheila Fitzpatrick, a historian of Soviet social history, has depicted a population by 1939 brutalized and depressed by the tyranny, incompetence, and contradictions of Stalinism. Her work does not really explain, however, how that same populace fought so stubbornly 'for the motherland, for Stalin.' No doubt the absurd murderousness of Nazi policies gave them little alternative. No doubt the aid which eventually flowed through from the USA was of great significance. But something does remain unexplained about how 'Stalin's Russia' won its 'Great Patriotic War.'

4. *The Consequences of the War*

By now very clear, however, are the consequences of the war for the USSR. In the short term the war made the Soviet state the second superpower, the global rival to the USA, and gave Stalin, until his death in 1953, an almost deified status. However, as the postwar decades passed, it became clear that the USSR and its expanded sphere of influence in Eastern Europe were not recovering from the war with the speed being dramatically exemplified in Western Europe and Japan. Indeed, the history of the USSR, at least until Gorbachev's accession to the party secretaryship in 1985 and, arguably, until the fall of the Berlin Wall and the collapse of communism (1989–91), should best be read as that of a generation who had fought and won its wars (including the terrible domestic campaigns for collectivization as well as the purges of the 1930s). Leaders and people were unwilling or unable to move beyond that visceral experience. After 1945, the USSR became the archetypal place 'where old soldiers did not die nor even fade away.' Brezhnev, Chernenko, and their many imitators further down the power structure were frozen into a past that had an ever-diminishing connection with a world facing a new technological and economic revolution.

Other ex-combatant societies were more open to change than the USSR but a certain sort of remembering and a certain sort of forgetting can readily enough be located in them, too. Memory proved most threatening in Yugoslavia. There the victorious partisans under Josip Broz Tito seemed for a time to have won a worthwhile victory. The special history of their campaigns against the Nazi-fascist occupiers of their country allowed them claims to independence from the USSR, which they duly exercised after 1948. At the same time the barbarity, during the war, of the collaborationist Croat fascist regime under Ante

Pavelić, whose savagery embarrassed even the Germans, seemed to suggest that the region was indeed best administered by a unitary state. The fall of communism, however, also brought down a communist Yugoslavia in which such Serb leaders as Slobodan Milosevic, unable to believe in official ideals, increasingly recalled the nationalism of wartime Cetniks rather than the internationalist Marxism once espoused by the partisans. This memory of war and murder justified new wars and new murders, even if with the somewhat ironical result that, by the end of the 1990s, the (Serb) winners of the last war had become losers and the (Slovene, Croat, Bosnian Moslem, and Kosovar) losers had become winners.

In Italy, the country with the largest communist party in the West and a polity which had a 'border idiosyncrasy,' bearing some comparison with Yugoslavia's role in the Eastern Bloc, the inheritance of war and fascism similarly possessed peculiar features. Postwar Italy began by renouncing fascism, empire, and war, in 1946 abandoning the monarchy that had tolerated the imposition of Mussolini's dictatorship and, in 1947, adopting a constitution which made considerable claim that the new Republic would be based on labor. In practice, however, Italy took its place in the Cold War West. Its purging of ex-Fascists was soon abandoned and both the power elites and the legislative base of the new regime exhibited much continuity with their Fascist predecessors. Nonetheless, from the 1960s, an ideology of antifascism was accorded more prominence in a liberalizing society. From 1978 to 1985, Sandro Pertini, an independent socialist who had spent many years in a Fascist jail and been personally involved in the decision to execute Mussolini, became Italy's president. Widely popular, he seemed an embodiment of the national rejection of the Fascist past.

Once again, however, the process of memory was taking a turn and a different useable past was beginning to emerge. Left terrorists in the 1970s had called themselves the new Resistance and declared that they were fighting a Fascist-style state—the governing Christian Democrats were thought to be merely a mask behind which lurked the Fascist beast. The murder of Aldo Moro in 1978 drove Italians decisively away from this sort of rhetoric and, in the 1980s and 1990s, Italians sought instead a 'pacification' with the past in which ex-Fascists had as much right to be heard as ex-partisans. Among historians, 'anti-anti-Fascists,' led by Renzo De Felice, the biographer of Mussolini, provided evidence and moral justification for this cause. Media magnate and conservative politician Silvio Berlusconi joined those who agreed that Italy's World War II had lost its ethical charge.

Among the Western European ex-combatants states perhaps the UK was the place where the official myth of the war survived with least challenge. A vast range of British society and behavior was influenced by Britain's war. The Welfare State, as codified by the postwar Labor government under Clement Attlee, was explained and justified as a reward for the effort of the British people in the 'people's war.' Wartime conservative Prime Minister, Winston Churchill, despite his many evident limitations, remained a national icon. British comedies from the Goons in the 1950s to *Dad's Army* in the 1970s and 1980s to *Goodnight Sweetheart* in the 1990s were obsessively set in the war. From the alleged acuteness of their secret service activities to the alleged idealism of their rejection of Nazism, the British have constantly sought to preserve the lion's share of the positives of World War II for themselves. The suspicion of a common European currency and the many other examples of continuing British insularity in turn reflect the British cherishing of the fact that they fought alone against the Nazi-fascists from 1940 to 1941, and express their associated annoyed perplexity that somehow their wartime sacrifice entailed a slower route to postwar prosperity compared with that of their continental neighbors.

Memory after memory, history after history, World War IIs, in their appalling plenitude, still eddy around. As the millennium ended, another historian wrote a major book about the meaning of an aspect of the war, and about the construction of that meaning. Peter Novick's *The Holocaust in American Life* (1999) daringly wondered whether the privileging of the Nazi killing of the Jews in contemporary Jewish and even gentile American discourse is altogether a positive. Being a historical victim at one time in the past, he argued cogently, can obscure as well as explain. His caution is timely. It is probably good that World War IIs are with us still; it will be better if the interpretation of so many drastic events can still occasion democratic debate, courteous, passionate, and humble debate, and if we can therefore avoid possessing a final solution to its many problems.

See also: Cold War, The; Contemporary History; First World War, The; Genocide: Historical Aspects; Holocaust, The; International Relations, History of; Military History; War: Anthropological Aspects; War: Causes and Patterns; War Crimes Tribunals; War, Sociology of; Warfare in History

Bibliography

Bartov O 1986 *The Eastern Front 1941–5: German Troops and the Barbarisation of Warfare*. St. Martin's Press, New York

Bosworth R J B 1993 *Explaining Auschwitz and Hiroshima: History Writing and the Second World War 1945–1990*. Routledge, London

Bosworth R J B 1998 *The Italian Dictatorship: Problems and Perspectives in the Interpretation of Mussolini and Fascism*. Arnold, London

Calder A 1969 *The People's War: Britain 1939–45*, Pantheon, London

Charmley J 1993 *Churchill: The End of Glory—A Political Biography*. Hodder and Stoughton, London

Dawidowicz L 1986 *The War Against the Jews 1933–45*, rev. edn. Penguin, Harmondsworth, UK
Dower J W 1986 *War Without Mercy: Race and Power in the Pacific War*. Pantheon, New York
Fischer F 1967 *Germany's Aims in the First World War*. Chatto and Windus, London
Fischer F 1986 *From Kaiserreich to Third Reich: Elements of Continuity in German History, 1871–1945*. Allen and Unwin, London
Fitzpatrick S 1999 *Everyday Stalinism: Ordinary Life in Extraordinary Times: Soviet Russia in the 1930s*. Oxford University Press, New York
Gorodetsky G 1999 *Grand Delusion: Stalin and the German Invasion of Russia*. Yale University Press, New Haven, CT
Ienaga S 1979 *Japan's Last War: World War II and the Japanese 1931–1945*. Australia University Press, Canberra, Australia
Hogan M J (ed.) 1996 *Hiroshima in History and Memory*. Cambridge University Press, Cambridge, UK
Kershaw I 1999 *Hitler 1889–1936: Hubris*. W. W. Norton, New York
Kershaw I 2000 *Hitler 1936–1945: Nemesis*. Allen Lane, London
Kolko G 1968 *The Politics of War: The World and United States Foreign Policy 1943–5*. Random House, New York
Milward A 1977 *War, Economy and Society 1939–1945*. University of California Press, Berkeley, CA
Novick P 1999 *The Holocaust in American Life*. Houghton Mifflin, Boston
Parker R A C 1990 *Struggle for Survival: The History of the Second World War*. Oxford University Press, Oxford, UK
Rousso H 1991 *The Vichy Syndrome: History and Memory in France Since 1944*. Harvard University Press, Cambridge, MA
Taylor A J P 1964 *The Origins of the Second World War*, rev. edn. Penguin, Harmondsworth, UK
Thorne C 1986 *The Far Eastern War: States and Societies 1941–5*. Unwin, London
Tumarkin N 1994 *The Living and the Dead: The Rise and Fall of the Cult of World War II in Russia*. Basic Books, New York
Watt D C 1989 *How War Came: The Immediate Origins of the Second World War 1938–1939*. Pantheon Books, New York
Weinberg G L 1994 *A World at Arms: A Global History of World War II*. Cambridge University Press, Cambridge, UK

R. J. B. Bosworth

Secondary Analysis: Methodology

Secondary analysis refers to a set of research practices that involve utilizing data collected by someone else or data that has been collected for another purpose (e.g., administrative records). It is used, to varying degrees, across a wide range of disciplines and throughout the world. Given this breadth, it is not surprising that it has taken on numerous forms. It is also conducted for several distinct reasons. Although not a research methodology, *per se*, several features distinguish it from other research activities. In turn, these features create opportunities and limitations for the secondary analyst. Central issues concern: (a) data availability, access, and documentation; (b) maintaining confidentiality and privacy pledges made by primary researchers; and (c) proprietary rights and data ownership.

1. Secondary Analysis as a 'Methodology'

The use of secondary data in behavioral and social sciences is ubiquitous, appearing in a number of traditional (i.e., quantitative) and untraditional (i.e., qualitative) forms. Despite its rather extended roots in the social and behavioral sciences, it is not widely celebrated as a method. As partial evidence of its relative obscurity, consider the fact that between 1978 and the present, a search of *PsychInfo* uncovered only 36 articles and books using the keyword 'secondary analysis.' Furthermore, the phrase 'secondary analysis' appeared in fewer than half of the article or book titles. Books on the topic are also relatively scarce; fewer than a dozen were uncovered through the same search (e.g., Boruch et al. 1981, Elder et al. 1993, Hyman 1972, Stewart 1984).

Ironically, the apparent obscurity of secondary analysis as a methodology is due to its pervasiveness. That is, the use of secondary data sources is so commonplace in many fields (e.g., economics, education, and sociology) that there is little need in calling attention to it as a separable methodology. This makes sense because unlike ethnography, survey research, or quasi-experimentation which each have distinctive methodological procedures and practices, secondary analysis does not involve a new or different set of tactics. Even from a statistical point of view, there is little to distinguish it from primary analysis, and most of the measurement, design, and statistical issues facing the secondary analyst are largely the same as those faced by a primary analyst. The obvious exception is that the secondary analyst is constrained by scope and nature of the design (e.g., the sample, sample sizes, attrition, missing data, measures, and research design) inasmuch as these have been specified by someone else (McCall and Appelbaum 1991). As such, secondary analysis boils down to a data resource, not a methodology, *per se*. However, the use of secondary data—especially when micro data records are used—does involve a unique set of logistical, ethical and practical considerations.

2. Varieties of Secondary Analysis

Because primary data can assume a number of forms (e.g., data based on cross-sectional surveys, administrative records, panel surveys, observations), secondary analysis has taken on a variety of forms. Two

main categories can be identified: (a) some data are developed for the explicit purpose of serving as a national resource; and (b) other forms of data are byproducts of the actions of individuals and organizations. The latter are included in the definition of secondary analysis because they play a large role in contemporary research. Given the increased use of the Internet, more administrative records, facts, and statistics (public and private) also will be readily available for use in research.

2.1 Traditional Forms of Secondary Analysis

The most common forms of secondary data include population censuses, continuous or regular surveys, national cohort studies, multisource data sets, administrative records, and one-time surveys or research studies. In the US, interest in secondary analysis was prompted by the appearance of public opinion polls and surveys in the 1940s and 1950s, and continued in the 1960s, when the federal government began, in earnest, gathering data through large-scale surveys based on representative sampling (Hyman 1972).

Of particular interest to social science researchers are the large scale, nationally representative, longitudinal (panel) surveys. In the USA, even cursory literature searches reveal hundreds of publications in economics and sociology that use data drawn from, for example, the Panel Study of Income Dynamics (PSID), the National Longitudinal Survey of Youth (NLSY), and the General Social Survey (GSS). In the United Kingdom, the General Household Survey (GHS) has been conducted annually since the beginning of the 1970s. The reuse of large-scale survey data is also evident across many nations simultaneously, as when Ravillion and Chen (1997) examined the correspondence between the rate of growth in gross domestic product and changes in poverty rates. The willingness of domestic and foreign governments to invest in data collection prompted Cherlin (1991) to speculate that the term secondary analysis will become obsolete. Many large-scale surveys are being sponsored by governments with no primary data analyses in mind.

Whereas large-scale surveys like the PSID are deliberately fielded to answer a multitude of research questions about populations and subgroups, substantial data gathering also is undertaken by governments as byproducts of their operation, to monitor their own processes and outcomes, or to assess environmental conditions or events (e.g., weather, rainfall). These are generally referred to as administrative or governmental records. The use of governmental records in research has a long tradition; over 100 years ago, Emile Durkheim used government statistics to examine some of the causes of suicide (Cherlin 1991).

Individual research studies also are fodder for secondary analysis. These reanalyzes are undertaken to check the accuracy of statistical conclusions in primary studies; for testing new statistical procedures or substantive hypotheses; resolving conflicts among researchers; for testing supplemental hypotheses suggested in primary analyses. The latter are usually conducted by the primary analyst, and this category represents the prevalent use of the term 'secondary analysis.' In recent years, secondary analysts reanalyzed multiple studies that use the same instrumentation, a tactic more akin to the spirit of meta-analysis.

A particularly important role for secondary analysis of individual studies is testing or demonstrating the superiority of new statistical methods. A thoughtful example is provided by Muthen and Curran (1997). They showed that latent growth curve models produced larger intervention effects than had been previously reported by exerting greater control over extraneous sources of error. More generally, in the past 25 years, efforts to address causal questions with extant data have produced substantial advances in statistical modeling (Duncan 1991). And, reanalysis of data from program evaluation studies has been undertaken to assure policy-makers that the results of primary studies are technically sound (Boruch et al. 1981).

2.2 Less Traditional Forms of Secondary Analysis

Although not a traditional form of secondary analysis, Webb et al. (1965) demonstrated that a substantial amount of 'data' are produced by individuals and organizations (public and private) as byproducts of their daily transactions. In their now classic text *Unobtrusive Measures: Nonreactive Research in the Social Sciences*, they identified a host of unconventional ways in which researchers can reuse existing data. In making their case, they showed how physical traces (erosion and accretion); the content of running records (e.g., mass media, judicial records), private records (sales records, credit card purchases), and contrived observations (e.g., hidden cameras) can be used as sources of data in research. The literature is filled with studies based on creative uses of these artifacts. Examples include assessing: the type, amount, and 'quality' of household garbage to assess the dietary and recycling habits of Arizona residents; the amount of trash left in the streets by revelers at Marti Gras to estimate the size of the daily crowd; and the differential wear and tear seen in separate sections to ascertain the popularity of sections of the *International Encyclopedia of the Social Sciences*. Finally, whereas secondary analysis has been historically viewed as a quantitative endeavor, reanalysis of qualitative data is now regarded as a legitimate part of the enterprise.

It would appear that Webb et al. (1965) had a substantial influence on subsequent generations of primary researchers. The most underacknowledged uses of secondary data are within primary evaluation

studies, where a mix of new and existing data are increasingly being used to evaluate the effectiveness of social interventions. The use of mixed method evaluation designs represents a core feature of contemporary evaluation theory and practice (Lipsey and Cordray 2000).

2.3 Advantages and Disadvantages of Secondary Analysis

Secondary analysis offers several advantages over primary data collection in behavioral and social sciences, but it also has its shortcomings. On the positive side of the ledger, re-use of data is efficient. These efficiencies include: (a) replication (or not) of findings across investigators; and (b) the discovery of biases in conventional statistical methods. Both of these benefit science by winnowing false hypotheses more quickly and offering new evidence (estimates) in their place. Testing new hypotheses, beyond those envisioned by the data developers, add to the efficiency of knowledge acquisition. These benefits have served as partial justification for the cost of conducting large scale, nationally representative surveys (using cross-sectional and panel designs) that query individuals on a wide array of topics. Some topics involving special populations (e.g., twins) or long time frames cannot be investigated without reliance upon archives and data sharing among investigators.

Alternatively, secondary data impose limits on what can be studied, where, and over what period of time. McCall and Appelbaum (1991) suggest using a Feasibility Matrix of Sample × Measure × Assessment Age as a tool for prescreening the potential utility of secondary data sets. In addition, technical problems (e.g., selection biases, sample attrition, and missing data) can be so severe as to limit the value of the primary data set. As such, assessing data quality probably needs to be included in the McCall–Appelbaum matrix. The existence of data can short-circuit one facet of the scientific process, tempting the analyst to 'mine' the data rather than initiate analyzes based on a theory or hypothesis. Although interesting results may emerge, such practices can lead to unreliable findings or findings limited to a single operationalization. The extent to which these problems in secondary analysis have influenced knowledge development is unknown. But, a mismatch between theory and data can be avoided with proper consideration of the relevance and quality of each data source.

3. What is Unique About Secondary Analysis?

Using data generated by someone else does raise several issues that make secondary analysis somewhat unique. Obviously, secondary analysis is possible only when it is available, easy to access, and in a form that is usable. Availability and access can be facilitated or impeded by a number of factors. In particular, because data do not 'speak for themselves,' they must be well documented. Data from government-sponsored, large surveys, panels and so on are often routinely archived and well documented. This is not uniform across all types of secondary data. Data sharing among individual investigators can become contentious, especially in light of questions about proprietary rights and data ownership. Researchers are obliged to honor original pledges of confidentiality and privacy. Balancing the desire to share data with these ethical requirements may require configuring the data in a way that may limit how it is disclosed and the methodological options available to the secondary analyst.

3.1 Availability, Access and Documentation

Establishment of data archives, advances in computer technology, and the emergence of the World Wide Web (WWW) have greatly facilitated access to data and solved some of the early problems that plagued the first generation of secondary analysts (e.g., poor documentation, noncommon data formats and language). Since the 1960s, the Interuniversity Consortium for Political and Social Research (ICPSR) has functioned as a major repository and dissemination service in the US. The National Archives have served a similar function for some governmental data. Increasingly, these roles have been devolved to individual governmental agencies that have developed skills in storing and disseminating their own data. In addition, the Internet has the potential for revolutionizing the use and transfer of data.

Secondary analysis also has been institutionalized within journal and professional codes of conduct. To facilitate access to data appearing in scientific journals, many journals have adopted policies whereby contributing authors are expected to make the data from their studies available for a designated period (usually three to five years) of time. Similarly, professional associations (e.g., the American Psychological Association) have incorporated data sharing into their codes of ethical behavior. Whereas archives and agencies require that data be properly documented, journals and professional associations are generally silent on this aspect of the data sharing process. Because data documentation is a process that records analytic decisions as they unfold over the course of the study, primary study authors need to be aware of these nontrivial editorial and ethical demands (see Boruch et al. 1981).

3.2 Ethical Considerations and Disclosure

The need to protect the confidentiality and privacy of research participants is sometimes at odds with the desire to make data available to others. Resolving these competing values requires careful consideration

at the time that primary research is conducted, documented, stored, and disseminated. Concealing the identity of participants can often be accomplished by removing personal identifiers from the data file. To the extent that identification is still possible through deductive disclosure (combining information to produce a unique profile of an individual), alternative procedures are needed. If identifiers are needed to link records (as in longitudinal research), additional layers of protection are needed. Fortunately, Boruch and Cecil (1979) provide a comprehensive treatment of the available statistical procedures (e.g., inoculating raw data with known amounts of error, randomized response techniques, collapsing categories) and institutional procedures (e.g., third parties who would serve as 'data brokers') that can be used to relax these problems. Data sharing among individual researchers requires explicit attention to ethical, statistical, organizational, and logistical (e.g., video editing and image modification) issues throughout the research process.

3.3 Proprietary Rights and Data Ownership

Changing policies, practices, and technology will facilitate data sharing and, as a consequence, increase use of data collected by someone else. Alternatively, secondary analysis of data for the purposes of addressing disputes among analysts can be quite difficult to negotiate. When conflict arises, it often revolves around who 'owns' the data and if, when, how and how much of it should be disclosed to others. Establishing proprietary rights to research and evaluation data is not a simple matter. Data ownership depends on how the research was financed, policies of the sponsoring and host organizations, conditions specified in laws (e.g., Freedom of Information Act, Privacy Act), data sharing polices of the journal in which the research is published, and ethical guidelines of professional associations to which authors belong. As researchers embark on primary or secondary analyses, it is necessary to understand these avenues and constraints.

4. A Summary Note

Access to quantitative and qualitative data from governments, businesses, and individual researchers has greatly facilitated the practice of secondary analysis. Changes in information technology—notably greater use of the World Wide Web—will undoubtedly enhance these practices even further at primary and secondary levels of research and evaluation.

See also: Archives and Historical Databases; Data Archives: International; Databases, Core: Demography and Registers; Intellectual Property: Legal Aspects; Intellectual Property Rights: Ethical Aspects; International Research: Programs and Databases; Meta-analysis: Overview; Privacy of Individuals in Social Research: Confidentiality; Unobtrusive Measures

Bibliography

Boruch R F, Cecil J S 1979 *Assuring the Confidentiality of Social Research Data*. University of Pennsylvania Press, Philadelphia, PA

Boruch R F, Wortman P M, Cordray D S and Associates (eds.) 1981 *Reanalyzing Program Evaluations: Policies and Practices for Secondary Analysis of Social and Educational Programs*, 1st edn. Jossey-Bass, San Francisco, CA

Cherlin A 1991 On analyzing other people's data. *Developmental Psychology* **27**(6): 946–8

Duncan G J 1991 Made in heaven: Secondary data analysis and interdisciplinary collaborators. *Developmental Psychology* **27**(6): 949–51

Elder Jr. G H, Pavalko E K, Clipp E C 1993 *Working with Archival Data: Studying Lives*. Sage Publications, Newbury Park, CA

Hyman H H 1972 *Secondary Analysis of Sample Surveys: Principles, Procedures and Potentialities*. Wiley, London

Lipsey M W, Cordray D S 2000 Evaluation methods for social intervention. *Annual Review of Psychology* **51**: 345–75

McCall R B, Appelbaum M I 1991 Some issues of conducting secondary analysis. *Developmental Psychology* **27**(6): 911–17

Muthen B O, Curran P J 1997 General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Method* **2**: 371–402

Ravillion M, Chen S H 1997 What can new survey data tell us about recent changes in distribution and poverty? *World Bank Economic Review* **11**: 357–82

Stewart D W 1984 *Secondary Research: Information Sources and Methods*. Sage, Beverly Hills, CA

Webb E J, Campbell D T, Schwartz R D, Sechrest L 1965 *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Rand McNally, Chicago, IL

D. S. Cordray

Secrecy, Anthropology of

A simple definition of secrecy as the deliberate concealment of information has served cross-cultural comparison (e.g., Tefft 1980, Bok 1982). Under this definition, people in cultures everywhere have secrets. Since a secret may turn out to be 'empty'—no hidden information in fact exists—some anthropologists have suggested that definition be refocused on the practice or the 'doing' of secrets rather than on a secret's

informational content. Anthropologists have been primarily interested in what could be called public, or institutionalized, secrecy: those persistent practices of hiding information within kinship, religious, political, or economic groups. Areas of research have included concealments of ritual knowledge and the social functions of secret societies. Uses of more personal, or private, secrets have also attracted attention insofar as these relate to cultural constructions of personhood and to the dynamics of interpersonal relationships. The study of secrets has posed obvious methodological and ethical problems for ethnographers.

1. *The Paradox of Secrecy*

The revelation of secrets is as important as the keeping of them. Ethnographers have called this the 'paradox' of secrecy (Bellman 1984). For a secret to persist within a social order, it must eventually be told to someone who may in turn keep the secret until it passes along again. Furthermore, the social consequence of secrets rests on the fact that people know of their existence even though content remains hidden.

Public awareness of the existence of secrets can afford prestige and power to those who know what others do not. Many accounts of secrecy revisit Georg Simmel's ground-breaking analysis of secrecy and secret societies (1950 [1908]), including his observation that secrets are a form of 'inner property.' Simmel suggested that secrets are a type of 'adornment'—a jewel or bauble that seduces public attention. Secrets with social consequence thus must be at least partly transparent and also periodically transacted and revealed.

2. *Public Secrecy*

Simmel's commodification opened the door to a political economy of secrecy. If secrets are property, then they are produced, exchanged, and consumed. Secrets are public because they have exchange value within a knowledge marketplace, and because such exchange has political consequence. Systems of political hierarchy may be grounded in an unequal distribution of secrets. Those in the know may dominate those who do not know so long as the ignorant grant that hidden knowledge has value.

2.1 *Secret Societies*

Anthropologists have pursued Simmel's original concern with secret societies, documenting a variety of secret groups, associations, lodges, and clubs in cultures around the world the members of which are pledged to maintain secrets, ritual and otherwise. The early comparative category 'secret society' was catholic enough to encompass Melanesian men's houses and grade societies, Australian totemic cults, Native American medicine lodges, West African male and female associations, and more (Webster 1908). Secrecy itself—the one attribute that these diverse organizations had in common—thus enticed anthropological attention (as Simmel predicted).

Liberal theories of democratic process and of the capitalist marketplace are suspicious of secrecy as they are of cabals and monopolies. Democracy requires an informed citizenry, and the market is supposed to depend on the free flow of information. From this perspective, secrecy generally goes against the public good. Structural-functionalist analysis, however, argued that secret societies commonly serve important social functions, including the education of children, preservation of political authority, punishment of lawbreakers, stimulation of the economy, and the like (Little 1949). Such 'social' secret societies address important community needs, unlike 'anti-social' secret societies whose goals are criminal or revolutionary (Wedgewood 1930).

2.2 *Secrecy and Power*

In the latter years of the twentieth century, anthropological attention returned to issues of power and inequality. Neutral definitions of culture as more-or-less shared knowledge gave way to new concerns with the contradictions, gaps, variation, and disparity in that knowledge. Secret knowledge and secret societies might have social functions, but they also maintain ruling political and economic regimes. The distribution of public secrets typically parallels that of other rights and powers. Adults hide information from children, often until the latter have been ritually initiated into adulthood. Men refuse to share ritual knowledge with women. Family and lineage groups own various sorts of genealogical, medical, or historical knowledge, sharing these only with kin.

Anthropologists have used a political economy of secrets to account for various forms of inequality. For example, the power of Kpelle elders over youth in Liberia was grounded in their management of the secrets of the Poro society (Murphy 1980). Beyond West Africa, anthropologists have argued that older, male keepers of secrets thereby acquired authority over women and the other uninitiated in various societies of Melanesia (Barth 1975), Native America (particularly Southwestern Pueblo cultures), and Australia (Keen 1994). Male appropriation of religious ritual, of access to the supernatural, of technology, of medicine, of sacred texts, of history, or of other valued information cemented men's authority over the ignorant.

Furthermore, women, children, and others in subordinate political position may be forced to reveal what they know, or otherwise denied rights to secrecy. Rights to have secrets are as consequential as the right to reveal information. Conversely, secrets can be a device to resist power. The dominated conceal what they know from authority. Women share knowledge kept from men. Slaves commune in secret languages. Children construct hidden worlds that evade adult awareness. In this reading, secrecy is a weapon of the weak that functions to resist and deflect supervision from above. Secrecy can also preserve a person's idiosyncratic beliefs and practices from public deprecation as in the case, for example, of middle-class English witches (Luhrmann 1989).

Alongside preservation of regimes of power, secrecy also contributes to perceptions of individual identity. Self-understanding may transform after a person has acquired secret information. Boys, for example, come to redefine themselves as men after progressing through an initiation ritual during which they learn adult secrets (Herdt 1990).

2.3 Rights to Reveal Secrets

Many analysts of secrecy systems have concluded that many secrets are not actually secret. Women know men's tricks: those spirit voices overheard during ritual are really bamboo flutes or bullroarers. Children pretend not to know that their parents are not really Santa Claus. Kin from one lineage are familiar with the supposedly secret names and genealogies of their neighbors. In oral societies, the leakiness of secret knowledge, in fact, helps maintain its viability within a community. If a secret holder dies, others are around to reveal lost knowledge if need be.

Systems of secrecy often rely upon inequalities in rights to reveal knowledge rather than on an effective concealment of information—rights that the Kpelle summarize in the term 'you cannot talk it' (Bellman 1984). Folk copyrights of this sort regulate who may speak about what (Lindstrom 1990). Even though someone may know the content of concealed information, that person may not repeat this knowledge in public without the copyright to do so. Family groups in Vanuatu, for example, own songs, myths, genealogies, and medical recipes that others may not publicly reveal without that family's permission. In private contexts, however, public secrets are surreptitiously discussed.

Knowledge is regulated not just by restricted transmission—by secrecy—but also by copyrights that limit who speaks publicly. Folk systems of copyright transform secrets into 'open' secrets. Those supposed not to know must pretend not to know. And those supposed to know pretend to think that only they, in fact, do know. Anthropological analyzes of the social and psychological dynamics of open secrecy prefigured work on other similar institutions, including the homosexual 'closet' (Sedgwick 1990, see also Taussig 1999 on public secrecy).

3. Personal Secrecy

Secrecy becomes possible given particular assumptions about personhood. The person must comprise some sort of inner self where secrets are stored along with an intention and capacity to conceal information. One can imagine a culture where notions of the person lacking such an inner self might deny the possibility of secrecy. No such culture exists although Western conceptions of childhood have often presumed that psychologically immature children lack competence with secrecy—that aptitudes to intentionally conceal information emerge as part of the child developmental process (Bok 1982).

Western historians, too, have suggested that there have been different regimes of secrecy in the past, related to shifts in constructions of personhood. Simmel (1950) connected the evolution of secrecy to that of individualization and urbane modernity (his argument recalls similar evolutionary accounts of personal privacy.) In pre-urban societies, lack of individualism and everyday intimacies of contact made secrecy difficult. Simmel supposed that secrets increased with developing opportunities for personal reserve and discretion.

More recent historians, stimulated by the work of Michel Foucault (1978) have argued instead that modernity shrinks opportunities for personal secrecy—that bureaucratic power structures increasingly have come to rely upon the monitoring of individuals, either by themselves or by institutional devices that extract information. According to Foucault, the origins of such extraction trace back to the Christian practice of confession. People are obliged to reveal their secrets to ecclesiastical, juridical, and other authorities. Revelation to authority confirms ones subjugation within a social order. The modern individual possesses inner capacities to conceal information but also contradictory urges and duties to reveal those secrets.

Stimulated by the work of Erving Goffman (1959) and others, students of interpersonal relationships have remarked the significance of secrets, masks, and 'face' in an everyday micropolitics of self-presentation. Such studies have noted more egalitarian uses of revelation. People strengthen their relationships, creating communities of trust, by revealing secret information. This may be a secret about themselves, or a secret about another that they pass along, often in return for the promise 'not to tell.' Such secrets are a kind of gossip (Bok 1982), the exchange of which remarks and maintains sentiments of friendship. Personal secrets are a social currency that people invest in

their relationships. Whereas public secrets maintain political value insofar as their transmission is restricted, the value of personal secrets flows from their easy everyday revelation.

4. The Study of Secrecy

Anthropology's interest in cross-cultural description as a whole can be taken to be the desire to learn and reveal other people's secrets (Taussig 1999). Methodologically, ethnographers face obvious problems of access to concealed information, but the study of secrecy raises even larger ethical puzzles. Anthropological codes of ethics generally require ethnographers to ensure that research does not harm the safety, dignity, or privacy of the people they study. Some researchers have described the structure and function of secrecy systems while avoiding details of secret knowledge content. Others have promised not to make their publications available to those members of a community (women, often) who should not have access to secret information. A few have refrained from publishing at all and restrict access to their fieldnotes. Ethical issues are thorniest where the secrets that anthropologists probe help maintain unjust social orders.

See also: Censorship and Secrecy: Legal Perspectives; Emotions, Sociology of; Guilt; Interpersonal Trust across the Lifespan; Knowledge, Anthropology of; Ritual; Trust: Philosophical Aspects; Trust, Sociology of

Bibliography

Barth F 1975 *Ritual and Knowledge Among the Baktaman of New Guinea*. Yale University Press, New Haven, CT

Bellman B L 1984 *The Language of Secrecy: Symbols & Metaphors in Poro Ritual*. Rutgers University Press, New Brunswick, NJ

Bok S 1982 *Secrets: On the Ethics of Concealment and Revelation*, 1st edn. Pantheon Books, New York

Foucault M 1978 *The History of Sexuality Volume 1: An Introduction*. Pantheon Books, New York

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY

Herdt G 1990 Secret societies and secret collectives. *Oceania* **60**: 361–81

Keen I 1994 *Knowledge and Secrecy in an Aboriginal Religion*. Oxford University Press, Oxford, UK

Lindstrom L 1990 *Knowledge and Power in a South Pacific Society*. Smithsonian Institution Press, Washington, DC

Little K L 1949 The role of the secret society in cultural specialization. *American Anthropologist* **51**: 199–212

Luhrmann T M 1989 The magic of secrecy. *Ethos* **17**: 131–65

Murphy W P 1980 Secret knowledge as property and power in Kpelle society: Elders versus youth. *Africa* **50**: 193–207

Sedgwick E K 1990 *Epistemology of the Closet*. University of California Press, Berkeley, CA

Simmel G 1950 [1908] The secret and the secret society. In: Wolff K (ed.) *The Sociology of Georg Simmel*. Free Press, Glencoe, IL

Taussig M 1999 *Defacement: Public Secrecy and the Labor of the Negative*. Stanford University Press, Stanford, CA

Tefft S K 1980 *Secrecy: A Cross-cultural Perspective*. Human Sciences Press, New York

Webster H 1908 *Primitive Secret Societies: A Study in Early Politics and Religion*, 2nd edn. rev. Macmillan, New York

Wedgewood C H 1930 The nature and function of secret societies. *Oceania* **1**: 129–45

L. Lindstrom

Secular Religions

The term 'secular religions' can be used to describe certain apparently secular enterprises that appear to share common features with enterprises usually thought of as 'religious.' 'Secular religion' is one of several terms—including 'civil religion,' 'invisible religion,' para-religion, and 'quasi-religion'—which draw attention to religious and religious-like beliefs and activities which do not fit easily into the Western folk conception of religion as a distinct institutional structure focused on a transcendent being or beings.

1. Examples of 'Secular Religions'

At first blush, the very notion of 'secular religion' would appear to be an oxymoron. In both popular and sociological usage, the term 'religion' is typically associated with the realm of the 'sacred' and contrasted with that of the 'secular.' However, a number of scholars have pointed out striking similarities between ostensibly religious and ostensibly secular undertakings. Secular ideologies it is argued may, like religions, unite followers into a community of shared beliefs. They may provide adherents with a sense of meaning and ultimate purpose. They may inspire in believers a sense of transcendence usually associated with religion. Attempts to find parallels between the sacred and the secular have been especially common in studies of the political realm and in studies of therapeutic organizations and activities.

1.1 Political 'Religions'

Numerous scholars have highlighted the religious aspects of political movements and ideologies. Communism, for example, has often been regarded as a secular religion. Zuo (1991) has recently described the veneration of Chairman Mao during the Chinese Cultural Revolution as a political religion replete with

sacred beings (Mao himself), sacred texts (the Little Red Book), and ritual (political denunciations). The term 'political religion' has also been employed to describe attempts made in developing societies to rally support for the concept of the nation. Crippin (1988) has gone so far as to argue the nationalism is the religion *par excellence* in modern society and that it is displacing more traditional forms of religion.

O'Toole (1977) has used the term 'sect' to describe certain political groups operating in Canada, including the Socialist Labor Party, followers of De Leon who wait for a Communist millennium which they regard as imminent. Such social movements as environmentalism, the animal rights movement, and the health food movement have been described as quasi-religions insofar as they provide adherents with a coherent worldview and sense of purpose at the same time that they command a great deal of loyalty.

1.2 Therapeutic 'Religions'

Many scholars have drawn attention to ritual aspects of western medical practice. Others have pointed out that psychiatrists have much in common with shamans and other religious healers. A number of researchers have pointed out the similarities that exist between self-help groups and religious organizations (Galanter 1989, Rudy and Greil 1988). One family within the self-help movement that is perhaps more obviously 'religious' in character than most of the others includes Alcoholics Anonymous (AA) and other 12-step or co-dependency groups. A few among many of the religious characteristics of AA are a conception of the sacred, ceremonies and rituals, creedal statements, experiences of transcendence, and the presence of an AA philosophy of life.

1.3 Other Examples

Other examples of attempts to point out analogies between apparently secular enterprises and religious ones abound in the social scientific literature. Within the realm of sport, attention has been paid to ritual elements and experiences of transcendence to be found in cricket, baseball, football, and the Olympic games. In the sphere of business, some writers have drawn attention to the sectarian characteristics of certain types of business, such as home-party sales organizations and direct sales organizations. It has now become commonplace among students of corporate culture to regard such ordinary activities as meetings, presentations and retirement dinners as rituals. Among the voluntary organizations that have been interpreted as quasi-religions are the Boy Scouts, groups of hobbyists and collectors, and 'fan clubs.' Jindra (1994) has described the phenomenon of Star Trek 'fandom' as religious in that it provides participants with a common worldview and inspires high levels of commitment. Several writers have attempted to characterize atheism itself as a religious enterprise, characterizing Ethical Culture and other humanist groups as analogous to religions. Still others have viewed faith in science as the dominant religious perspective of the contemporary era.

2. 'Secular Religions' and the Problem of Defining Religion

It may be argued that an interest in secular religion is a natural outgrowth of the functionalist perspective in classical sociology, which tended to view religion as serving a necessary function for the maintenance of society by uniting members of a society into a common moral universe. If, as many early sociologists thought, supernaturalistic conceptions of the universe were destined to recede in the face of industrialization and the increasing influence of science, the question arose as to what phenomena might serve as the 'social cement' of future societies. Comte's call for a 'religion of humanity' qualifies as an early sociological mention of the notion of a 'secular religion.' Durkheim's expectation that a 'cult of the individual' might play a similar role in the maintenance of society as that traditionally played by religions represents another early effort to provide intellectual justification for the idea that apparently secular ideologies may have religious characteristics.

As appealing and intuitive as the idea that secular enterprises may share important features with religions might be, the notion of a 'secular religion' must confront theoretical problems concerning the appropriate sociological definition of 'religion' and of 'the sacred.' Functional definitions of religion emphasize that the essential element in religion is the provision of an encompassing system of meaning or the ability to connect people to the ultimate conditions of their existence. Substantive definitions of religion argue that what distinguishes religion from other types of human activity is its reference to the sacred, the supernatural, or the superempirical.

The advantage of functional definitions is a breadth and inclusiveness that allows one to look at beliefs and practices not commonly referred to as religious but which may nonetheless resemble religious phenomena in important ways. One major disadvantage is that they may have the effect of so broadening the concept of religion that it becomes meaningless. While substantive definitions avoid this problem, they may result in allowing traditional Western conceptions of the nature of religion to determine what is to be included in the 'religious' category. Viewed from the perspective of the debate over the definition of religion, the theoretical problem with the concept of 'secular

religion' is that it seems to rely simultaneously on both a broad functional approach to religion and on a narrower substantive approach. The idea that environmentalism or nationalism could be properly called a religion relies on a functional definition, while the idea that such religions are 'secular' rather than 'sacred' necessarily depends on a substantive definition.

One might argue that both functional and substantive definitions of religion share the weakness that they privilege social scientists' conceptions of religion over folk conceptions of religion, that is to say the ways that people use the term 'religion' in everyday life. Some social scientists would therefore argue that the search for a scientifically valid definition of religion is futile and that the best that can be done is to define religion ethnographically, treating it as a 'category of discourse' with meanings that are changeable over time. The position that scholars take with regard to these definitional issues will necessarily influence the way in which they approach the study of 'secular religions.' This is perhaps one reason why there is at present no consensus within the social sciences with regard to the most appropriate method for studying phenomena residing on the border between the sacred and the secular.

3. *Approaches to the Study of 'Secular Religions'*

Within contemporary sociology and anthropology, there exist several different research traditions that focus on the boundary between the religious and the non-religious.

3.1 *The Notion of Civil Religion*

While Rousseau coined the term 'civil religion,' its development as a social scientific concept is attributed to Robert Bellah (1967). Working within the functionalist tradition, which sees religion as integrative for society, Bellah argued for the existence of a US civil religion, an 'institutionalized collection of sacred beliefs about the American nation,' which binds US citizens together in spite of denominational pluralism. A key tenet of US civil religion is the conception of the USA as a nation with a divinely ordained mission. Although the civil religion concept was developed in the US context, it has been applied to the analysis of many states including Canada, Israel, and Malaysia (see *Christianity: Evangelical, Revivalist, and Pentecostal*).

3.2 *The Implicit Religion Tradition*

Although the term 'implicit religion' was coined and popularized by Edward Bailey (1983), the concept owes much to the work of Thomas Luckmann (1967). Working with a broad functional definition of religion as the transcending of biological nature and the formation of a self, Luckmann argues that religion is a human universal. Luckmann maintains that traditional religions have become irrelevant in the modern world but that religion, rather than disappearing, has become personalized, privatized, and 'invisible' (see *Religiosity: Modern*). Following Luckmann's lead, Bailey's implicit religion approach looks for the experience of the sacred within events of everyday life ordinarily dismissed as profane. Thus, in a study of interaction in an English public house, Bailey interprets the ethos of 'being a man,' mastering alcohol and respecting the selves of others as implicitly religious.

3.3 *The Study of 'Religious' Forms*

There also exist a relatively large number of studies of general social 'forms' which are deemed to be relevant in both sacred and secular contexts. General discussions of 'secular ritual,' for example, fall into this category. Goffman's (1967) work on the functional significance of such rituals of everyday life as 'saving face' and showing 'deference' is particularly well known. Collins (1981) treats 'interaction ritual chains' as a key element in his attempt to lay a micro-sociological foundation for macro-sociology. Once conversations, social greetings, and athletic contests are allowed to count as ritual, ritual becomes virtually coterminous with social life. For this reason, some scholars reject 'ritual' as a meaningless category, while others argue that the ubiquity of ritual simply means that ritualizing is a fundamental human activity.

Many sociological studies of the commitment process either explicitly or implicitly present the commitment process as operating in much the same way in both sacred and secular contexts (Kanter 1972). A number of writers have developed models of the identity change process which highlight the similarity between religious conversion and other forms of identity change (Galanter 1989, Greil and Rudy 1983). Studies that employ the term 'sect' in the analysis of apparently secular organizations have already been discussed.

3.4 *The 'Quasi-religion' Approach*

The quasi-religion approach popularized by Greil (1993) and his colleagues relies on an ethnographic approach to the term 'religion,' regarding it, not as an objective category susceptible to social scientific definition, but as a claim negotiated in the course of social interaction. Greil distinguishes between para-religions and quasi-religions. Para-religions, which are ostensibly nonreligious entities that nonetheless deal with matters of ultimate concern, resemble the enterprises referred to in this article as 'secular religions.'

The term 'quasi-religion,' on the other hand, refers to groups, like AA or certain occult groups, which deal with matters of transcendence or ultimate concern, but which do not see themselves or present themselves unambiguously as religious. The concern here is not to determine whether a particular group is or is not a quasi-religion but to examine the process by which particular groups attempt to claim or repudiate the religious label and by which other groups and individuals respond to these claims.

4. The Theoretical Significance of Secular Religions

Pursuing the analogy between ostensibly secular enterprises and 'religion' as it is usually conceived raises important questions concerning the proper definition of religion, the process of secularization and the nature of religion in contemporary societies. The study of secular religions puts into bold relief the problematic nature of social scientific attempts to define religion. How one defines religion has important consequences for how one thinks about secular religions and for whether or not one regards the concept as useful. Attention to religious border phenomena presses one to consider the value of an ethnographic approach to definition that conceptualizes religion as a category of discourse whose precise meaning and implications are continually being negotiated in the course of social interaction. The study of secular religions also directs attention to difficulties in specifying and evaluating the secularization thesis, which claims that the influence of religion is declining in contemporary societies (see *Secularization*). It should be clear that one's approach to the question of secularization (including whether one even conceives of secularization as a theoretical possibility) depends on one's definition of religion and on one's treatment of religious border phenomena.

Several of the approaches to secular religion discussed here imply that traditional Western conceptions of religion as an institutionalized set of beliefs and practices focusing on a transcendent deity is beginning to lose its hold over many people. Likewise, many people seem to be tending to give expression to their experiences of transcendence in institutional contexts that have not typically been thought of as religious. If this is, in fact, the case, the line between religion and nonreligion may be expected to become even more blurred and the study of religious border phenomena even more central to the social scientific enterprise.

See also: Communism; Laicization, History of; Nationalism, Sociology of; New Religious Movements; Religion, Sociology of; Religiosity, Sociology of

Bibliography

Bailey E 1983 The implicit religion of contemporary society: An orientation and plea for its study. *Religion* **13**: 69–83

Bellah R N 1967 Civil religion in America. *Daedalus* **96**: 1–21

Collins R 1981 On the micro-foundations of macro-sociology. *American Journal of Sociology*. **86**: 984–1014

Crippin T 1988 Old and new gods in the modern world: Toward a theory of religious transformation. *Social Forces* **67**: 316–36

Galanter M 1989 *Cults: Faith, Healing, and Coercion*. Oxford University Press, New York

Goffman E 1967 *Interaction Ritual: Essays on Face-to-Face Behavior* 1st edn. Anchor Books, Garden City, NY

Greil A L 1993 Explorations along the sacred frontier: Notes on para-religions, quasi-religions, and other boundary phenomena. In: Bromley D, Hadden J K (eds.) *Handbook of Cults and Sects in America: Assessing Two Decades of Research and Theory Development*, (Volume 2 of *Religion and the Social Order*). JAI Press, Greenwich, CT pp. 153–72

Greil A L, Rudy D R 1983 Conversion to the world view of Alcoholics Anonymous: A refinement of conversion theory. *Qualitative Sociology* **6**: 5–28

Jindra M 1994 Star trek fandom as a religious phenomenon. *Sociology of Religion* **55**: 27–51

Kanter R M 1972 *Commitment and Community: Communes and Utopias in Sociological Perspective*. Harvard University Press, Cambridge, MA

Luckmann T 1967 *The Invisible Religion: The Problem of Religion in Modern Society*. Macmillan, New York

O'Toole R 1977 *The Precipitous Path*: *Studies in Political Sects*. Peter Martin, Toronto, Canada

Rudy D R, Greil A L 1988 Is Alcoholics Anonymous a religious organization? *Sociological Analysis* **50**: 41–51

Zuo J P 1991 Political religion: The case of the cultural revolution in China. *Sociological Analysis* **52**: 99–110

A. L. Greil

Secularization

In its precise historical sense, 'secularization' refers to the transfer of persons, things, meanings, etc., from ecclesiastical or religious to civil or lay use. In its broadest sense, often postulated as a universal developmental process, secularization refers to the progressive decline of religious beliefs, practices, and institutions.

1. The Term 'Secularization'

Etymologically, the term secularization derives from the medieval Latin word *saeculum*, with its dual temporal-spatial connotation of secular *age* and secular *world*. Such a semantic connotation points to the fact that social reality in medieval Christendom was structured through a system of classification which

divided 'this world' into two heterogeneous realms or spheres, 'the religious' and 'the secular.' This was a particular, and historically rather unique, variant of the kind of universal dualist system of classification of social reality into sacred and profane realms, postulated by Émile Durkheim.

In fact, Western European Christendom was structured through a double dualist system of classification. There was, on the one hand, the dualism between 'this world' (the City of Man) and 'the other world' (the City of God). There was, on the other hand, the dualism within 'this world' between a 'religious' and a 'secular' sphere. Both dualisms were mediated, moreover, by the sacramental nature of the church, situated in the middle, simultaneously belonging to both worlds and, therefore, able to mediate sacramentally between the two.

The differentiation between the cloistered *regular* clergy and the *secular* clergy living in the world was one of the many manifestations of this dualism. The term secularization was first used in canon law to refer to the process whereby a religious monk left the cloister to return to the world and thus become a secular priest. In reference to an actual historical process, however, the term secularization was first used to signify the lay expropriation of monasteries, landholdings, and the mortmain wealth of the church after the Protestant Reformation. Thereafter, secularization has come to designate any transfer from religious or ecclesiastical to civil or lay use.

1.1 Secularization as a Historical Process

Secularization refers to the historical process whereby the dualist system within 'this world' and the sacramental structures of mediation between this world and the other world progressively break down until the medieval system of classification disappears. Max Weber's expressive image of the breaking of the monastery walls remains perhaps the best graphic expression of this radical spatial restructuration initiated by the Protestant Reformation. This process, which Weber conceptualized as a general reorientation of religion from an other-worldly to an inner-worldly direction, is literally a process of secularization. Religious 'callings' are redirected to the *saeculum*. Salvation and religious perfection are no longer to be found in withdrawal from the world but in the midst of worldly secular activities. In fact, the symbolic wall separating the religious and the secular realms breaks down. The separation between 'this world' and 'the other world,' for the time being at least, remains. But, from now on, there will be only one single 'this world,' the secular one, within which religion will have to find its own place. To study what new systems of classification and differentiation emerge within this one secular world and what new place religion will have within it is precisely the analytical task of the theory of secularization.

Obviously, such a concept of secularization refers to a particular historical process of transformation of Western European Christian societies and might not be directly applicable to other non-Christian societies with very different modes of structuration of the sacred and profane realms. It could hardly be applicable, for instance, to such 'religions' as Confucianism or Taoism, insofar as they are not characterized by high tension with 'the world' and have no ecclesiastical organization. In a sense those religions which have always been 'worldly' and 'lay' do not need to undergo a process of secularization.

In itself such a spatial-structural concept of secularization describes only changes in the location of Christian religion from medieval to modern societies. It tells very little, however, about the extent and character of the religious beliefs, practices, and experiences of individuals and groups living in such societies. Yet the theory of secularization adopted by the modern social sciences incorporated the beliefs in progress and the critiques of religion of the Enlightenment and of Positivism and assumed that the historical process of secularization entailed the progressive decline of religion in the modern world. Thus, the theory of secularization became embedded in a philosophy of history that saw history as the progressive evolution of humanity from superstition to reason, from belief to unbelief, from religion to science.

2. The Secularization Paradigm

Secularization might have been the only theory within the social sciences that was able to attain the status of a paradigm. In one form or another, with the possible exception of Alexis de Tocqueville, Vilfredo Pareto, and William James, the thesis of secularization was shared by all the founders. Paradoxically, the consensus was such that for over a century the theory of secularization remained not only uncontested but also untested. Even Durkheim's and Weber's work, while serving as the foundation for later theories of secularization, offer scant empirical analysis of modern processes of secularization, particularly of the way in which those processes affect the place, nature and role of religion in the modern world. Even after freeing themselves from some of the rationalist and positivist prejudices about religion, they still share the major intellectual assumptions of the age about the future of religion.

For Durkheim, the old gods were growing old or already dead and the dysfunctional historical religions would not be able to compete with the new functional gods and secular moralities which modern societies were bound to generate. For Weber, the process of

intellectual rationalization, carried to its culmination by modern science, had ended in the complete disenchantment of the world, while the functional differentiation of the secular spheres had displaced the old integrative monotheism, replacing it with the modern polytheism of values and their unceasing and irreconcilable struggle. The old churches remain only as a refuge for those 'who cannot bear the fate of the times like a man' and are willing to make the inevitable 'intellectual sacrifice.'

Only in the 1960s one finds the first attempts to develop more systematic and empirically grounded formulations of the theory of secularization in the works of Acquaviva (1961), Berger (1967), Luckmann (1963), and Wilson (1966). It was then, at the very moment when theologians were celebrating the death of God and the secular city, that the first flaws in the theory became noticeable and the first systematic critiques were raised by Martin (1969) and Greeley (1972) in what constituted the first secularization debate. For the first time it became possible to separate the theory of secularization from its ideological origins in the Enlightenment critique of religion and to distinguish the theory of secularization, as a theory of the modern autonomous differentiation of the secular and the religious spheres, from the thesis that the end result of the process of modern differentiation would be the progressive erosion, decline and eventual disappearance of religion. Greeley (1972) already pointed out that the secularization of society, which he conceded, by no means implied the end of church religiosity, the emergence of 'secular man,' or the social irrelevance of religion in modern secular societies. Yet after three decades the secularization debate remains unabated. Defenders of the theory tend to point to the secularization of society and to the decline of church religiosity in Europe as substantiating evidence, while critics tend to emphasize the persistent religiosity in the United States and widespread signs of religious revival as damaging counterevidence that justify discarding the whole theory as a 'myth.'

3. *The Three Subtheses of the Theory of Secularization*

Although it is often viewed as a single unified theory, the paradigm of secularization is actually made up of three different and disparate propositions: secularization as differentiation of the secular spheres from religious institutions and norms, secularization as general decline of religious beliefs and practices, and secularization as privatization or marginalization of religion to a privatized sphere. Strictly speaking, the core and central thesis of the theory of secularization is the conceptualization of the historical process of societal modernization as a process of functional differentiation and emancipation of the secular spheres—primarily the state, the economy, and science—from religion and the concomitant specialized and functional differentiation of religion within its own newly found religious sphere. The other subtheses, the decline and privatization of religion, were added as allegedly necessary structural consequences of the process of secularization. Maintaining this analytical distinction should allow the examination and testing of the validity of each of the three propositions independently of each other and thus refocus the often fruitless secularization debate into comparative historical analysis that could account for obviously different patterns of secularization.

3.1 *The Differentiation and Secularization of Society*

The medieval dichotomous classification of reality into religious and secular realms was to a large extent dictated by the church. In this sense, the official perspective from which medieval societies saw themselves was a religious one. Everything within the *saeculum* remained an undifferentiated whole as long as it was viewed from the outside, from the perspective of the religious. The fall of the religious walls put an end to this dichotomous way of thinking and opened up an entire new space for processes of internal differentiation of the various secular spheres. Now, for the first time, the various secular spheres—politics, economics, law, science, art, etc.—could come fully into their own, become differentiated from each other, and follow what Weber called their 'internal and lawful autonomy.' The religious sphere, in turn, became a less central and spatially diminished sphere within the new secular system, but also a more internally differentiated one, specializing in 'its own religious' function and losing many other 'non-religious' (clerical, educational, social welfare) functions (Luhmann 1977). The loss in functions entailed as well a significant loss in hegemony and power.

It is unnecessary to enter into the controversial search for first causes setting the modern process of differentiation into motion. It suffices to stress the role which four related and parallel developments played in undermining the medieval religious system of classification: the Protestant Reformation; the formation of modern states; the growth of modern capitalism; and the early modern scientific revolution. Each of the four developments contributed its own dynamic to modern processes of secularization. The four of them together were certainly more than sufficient to carry the process through.

The Protestant Reformation by undermining the universalist claims of the Roman Catholic church helped to destroy the old organic system of Western Christendom and to liberate the secular spheres from religious control. Protestantism also served to legit-

imate the rise of bourgeois man and of the new entrepreneurial classes, the rise of the modern sovereign state against the universal Christian monarchy, and the triumph of the new science against Catholic scholasticism. Moreover, Protestantism may also be viewed as a form of internal secularization, as the vehicle through which Christian religious contents were to assume institutionalized secular forms in modern societies, thereby erasing the religious/secular divide.

If the universalist claims of the church as a salvation organization were undermined by the religious pluralism introduced by the Reformation, its monopolist compulsory character was undermined by the rise of a modern secular state which progressively was able to concentrate and monopolize the means of violence and coercion within its territory. In the early absolutist phase the alliance of throne and altar became even more accentuated. The churches attempted to reproduce the model of Christendom at the national level, but all the territorial national churches, Anglican, Lutheran, Catholic, and Orthodox, fell under the caesaro–papist control of the absolutist state. As the political costs of enforcing conformity became too high, the principle *cuius regio eius religio* turned into the principle of religious tolerance and state neutrality towards privatized religion, the liberal state's preferred form of religion. Eventually, new secular *raison d'état* principles led to the constitutional separation of church and state, even though some countries such as England and the Scandinavian Lutheran countries may have maintained formal establishment.

Capitalism, that revolutionizing force in history which according to Marx 'melts all that is solid into air and profanes all that is holy,' had already sprouted within the womb of the old Christian society in the medieval towns. No other sphere of the *saeculum* would prove more secular and more unsusceptible to religious and moral regulation than the capitalist market. Nowhere is the transvaluation of values which takes place from Medieval to Puritan Christianity as radical and as evident as in the change of attitude towards 'charity'—that most Christian of virtues—and towards poverty. Following Weber (1958), one could distinguish three phases and meanings of capitalist secularization: in the Puritan phase, 'asceticism was carried out of monastic cells into everyday life' and secular economic activities acquired the meaning and compulsion of a religious calling; in the utilitarian phase, as the religious roots dried out, the irrational compulsion turned into 'sober economic virtue' and 'utilitarian worldlines'; finally, once capitalism 'rests on mechanical foundations,' it no longer needs religious or moral support and begins to penetrate and colonize the religious sphere itself, subjecting it to the logic of commodification (Berger 1967).

The conflict between the church and the new science, symbolized by the trial of Galileo, was not as much about the substantive truth or falsity of the new Copernican theory as it was about the validity of the claims of the new science to have discovered a new autonomous method of obtaining and verifying truth. The conflict was above all about science's claims to differentiated autonomy from religion. The Newtonian Enlightenment established a new synthesis between faith and reason, which in Anglo-Saxon countries was to last until the Darwinian crisis of the second half of the nineteenth-century. Across the Channel, however, the Enlightenment became patently radicalized and militantly anti-religious. Science was transformed into a scientific and scientistic worldview which claimed to have replaced religion the way a new scientific paradigm replaces an outmoded one.

As each of these carriers—Protestantism, the modern state, modern capitalism, and modern science—developed different dynamics in different places and at different times, the patterns and the outcomes of the historical process of secularization varied accordingly. Yet it is striking how few comparative historical studies of secularization there are which would take these four, or other, variables into account. David Martin's *A General Theory of Secularization* is perhaps the single prominent exception. Only when it comes to capitalism has it been nearly universally recognized that economic development affects the 'rates of secularization,' that is, the extent and relative distribution of religious beliefs and practices. This positive insight, however, turns into a blinder, when it is made into the sole variable accounting for different rates of secularization. As a result, those cases in which no positive correlation is found, as expected, between rates of secularization and rates of industrialization, urbanization, proletarianization, education, in short, with indicators of socio-economic development, are classified as 'exceptions' to the 'rule.'

3.2 *The Decline of Religion Thesis*

The assumption, often stated but mostly unstated, that religion in the modern world was declining and would likely continue to decline has been until recently a dominant postulate of the theory of secularization. It was based primarily on evidence from European societies showing that the closer people were involved in industrial production, the less religious they became, or at least, the less they took part in institutional church religiosity. The theory assumed that the European trends were universal and that non-European societies would evince similar rates of religious decline with increasing industrialization. It is this part of the theory which has proven patently wrong.

One should keep in mind the inherent difficulties in comparative studies of religion. Globally, the evidence is insufficient and very uneven. There is often no consensus as to what counts as religion and even when there is agreement on the object of study, there is likely to be disagreement on which of the dimensions of religiosity (membership affiliation, official vs. popular

religion, beliefs, ritual and nonritual practices, experiences, doctrinal knowledge, and their behavioral and ethical effects) one should measure and how various dimensions should be ranked and compared.

Nevertheless, one can prudently state that since World War II, despite rapid increases in industrialization, urbanization, and education, most religious traditions in most parts of the world have either experienced some growth or maintained their vitality (Whaling 1987). The main exceptions were: the rapid decline of primal religions mostly in exchange for more 'advanced' ones (mainly Islam and Christianity), the sudden and dramatic decline of religion in communist countries, a process which is now being reversed after the fall of communism, and the continuous decline of religion throughout much of Western Europe (and, one could add, some of its colonial outposts such as Argentina, Uruguay, New Zealand, and Quebec).

What remains, therefore, as significant and overwhelming evidence is the progressive and apparently still continuing decline of religion in Western Europe. It is this evidence which has always served as the empirical basis for most theories of secularization. Were it not for the fact that religion shows no uniform sign of decline in Japan or the United States, two equally modern societies, one could still perhaps maintain the 'modernizing' developmentalist assumption that it is only a matter of time before the more 'backward' societies catch-up with the more 'modern' ones. But such an assumption is no longer tenable. Leaving aside the evidence from Japan, a case which should be crucial, however, for any attempt to develop a *general* theory of secularization, there is still the need to explain the obviously contrasting religious trends in Western Europe and the United States.

Until very recently, most of the comparative observations as well as attempts at explanation came from the European side. European visitors have always been struck by the vitality of American religion. The United States appeared simultaneously as the land of 'perfect disestablishment' and as 'the land of religiosity *par excellence*.' Yet, Europeans rarely felt compelled to put into question the thesis of the general decline of religion in view of the American counterevidence. Religious decline was so much taken for granted that what required an explanation was the American 'deviation' from the European 'norm.' The standard explanations have been either the expedient appeal to 'American exceptionalism' or the casuistic strategy to rule out the American evidence as irrelevant, because American religion was supposed to have become so 'secular' that it should no longer count as religion (Luckmann 1963).

From a global–historical perspective, however, it is the dramatic decline of religion in Europe that truly demands an explanation. A plausible answer would require a search for independent variables, for those independent carriers of secularization present in Europe but absent in the United States. Looking at the four historical carriers mentioned above, neither Protestantism nor capitalism would appear as plausible candidates. The state and scientific culture, however, could serve as plausible independent variables, since church–state relations and the scientific worldviews carried by the Enlightenment were significantly different in Europe and America.

What the United States never had was an absolutist state and its ecclesiastical counterpart, a caesaro–papist state church. It was the caesaro–papist embrace of throne and altar under absolutism that perhaps more than anything else determined the decline of church religion in Europe. Consistently throughout Europe, nonestablished churches and sects in most countries have been able to withstand the secularizing trends better than the established church. It was the very attempt to preserve and prolong Christendom in every state and thus to resist modern functional differentiation that nearly destroyed the churches in Europe.

The Enlightenment and its critique of religion became themselves independent carriers of processes of secularization wherever the established churches became obstacles to the modern process of functional differentiation or resisted the emancipation of the cognitive–scientific, political–practical, or aesthetic–expressive secular spheres from religious and ecclesiastical tutelage. In such cases, the Enlightenment critique of religion was usually adopted by social movements and political parties, becoming in the process a self-fulfilling prophecy. By contrast, wherever religion itself accepted, perhaps even furthered, the functional differentiation of the secular spheres from the religious sphere, the radical Enlightenment and its critique of religion became superfluous. Simply put, the more religions resist the process of modern differentiation, that is, secularization in the first sense, the more they will tend in the long run to suffer religious decline, that is, secularization in the second sense.

3.3 The Privatization of Religion Thesis

As a corollary of the thesis of differentiation, religious disestablishment entails the privatization of religion. Religion becomes indeed 'a private affair.' Insofar as freedom of conscience, 'the first freedom' as well as the precondition of all modern freedoms, is related intrinsically to 'the right to privacy'—to the modern institutionalization of a private sphere free from governmental intrusion as well as free from ecclesiastical control—and inasmuch as the right to privacy serves as the very foundation of modern liberalism and of modern individualism, then indeed the privatization of religion is essential to modern societies.

There is, however, another more radical version of the thesis of privatization which often appears as a corollary of the decline of religion thesis. In modern secular societies, whatever residual religion remains

becomes so subjective and privatized that it turns 'invisible,' that is, marginal and irrelevant from a societal point of view. Not only are traditional religious institutions becoming increasingly irrelevant but, Luckmann (1963) adds, modern religion itself is no longer to be found inside the churches. The modern quest for salvation and personal meaning has withdrawn to the private sphere of the self. Significant for the structure of modern secular societies is the fact that this quest for subjective meaning is a strictly personal affair. The primary 'public' institutions (state, economy) no longer need or are interested in maintaining a sacred cosmos or a public religious worldview.

The functionalist thesis of privatization turns problematic when instead of being a falsifiable empirical theory of dominant historical trends, it becomes a prescriptive normative theory of how religious institutions ought to behave in the modern world. Unlike secular differentiation, which remains a structural trend that serves to define the very structure of modernity, the privatization of religion is a historical option, a 'preferred option' to be sure, but an option nonetheless. Privatization is preferred internally as evinced by general pietistic trends, by processes of religious individuation, and by the reflexive nature of modern religion. Privatization is constrained externally by structural trends of differentiation which force religion into a circumscribed and differentiated religious sphere. Above all, privatization is mandated ideologically by liberal categories of thought which permeate modern political and constitutional theories.

The theory of secularization should be free from such a liberal ideological bias and admit that there may be legitimate forms of 'public' religion in the modern world, which are not necessarily anti-modern fundamentalist reactions and which do not need to endanger either modern individual freedoms or modern differentiated structures. Many of the recent critiques and revisions of the paradigm of secularization derive from the fact that in the 1980s religions throughout the world thrust themselves unexpectedly into the public arena of moral and political contestation, demonstrating that religions in the modern secular world continue and will likely continue to have a public dimension.

See also: Religion: Evolution and Development; Religion, Phenomenology of; Religion, Sociology of; Secular Religions

Bibliography

Acquaviva S S 1961 *L'ecclissi del sacro nella civiltà industrialle*. Edizioni di Communità, Milan [1979 *The Decline of the Sacred in Industrial Society*. Blackwell, Oxford]

Berger P 1967 *The Sacred Canopy*. Doubleday, Garden City, NY

Casanova J 1994 *Public Religions in the Modern World*. University of Chicago Press, Chicago

Dobbelaere K 1981 *Secularization: A Multidimensional Concept*. Sage, Beverly Hills, CA

Greeley A 1972 *Unsecular Man: The Persistence of Religion*. Schocken, New York

Greeley A 1989 *Religious Change in America*. Harvard University Press, Cambridge, MA

Hadden J K, Shupe A (eds.) 1989 *Secularization and Fundamentalism Reconsidered*. Paragon House, New York

Luckmann T 1963 *Das Problem der Religion in der modernen Gesellschaft*. Rombach, Freiburg [1967 *Invisible Religion*. Macmillan, New York]

Luhmann N 1977 *Funktion der Religion*. Suhrkamp Verlag, Frankfurt [1984 *Religious Dogmatics and the Evolution of Societies*. Mellen Press, New York]

Martin D 1969 *The Religious and the Secular*. Shocken, New York

Martin D 1978 *A General Theory of Secularization*. Blackwell, Oxford, UK

Stark R, Bainbridge W S 1985 *The Future of Religion*. University of California Press, Berkeley, CA

Weber M 1958 The protestant ethic and the spirit of capitalism. Scribner's Sons, New York

Whaling F (ed.) 1987 *Religion in Today's World: The Religious Situation of the World from 1945 to the Present Day*. T & T Clark, Edinburgh, UK

Wilson B 1966 *Religion in Secular Society*. Watts, London

Wilson B 1985 Secularization: The inherited model. In: Hammond P (ed.) *The Sacred in a Secular Age*. University of California Press, Berkeley, CA

J. Casanova

Segregation Indices

Measuring the residential segregation of racial and ethnic populations became very important to social scientists during the civil rights movement in the United States (Duncan and Duncan 1955; Taueber and Taueber 1965). Until the mid-1970s, when important critiques of the dissimilarity index stimulated efforts to develop alternatives, research almost exclusively relied upon this measure, and focused on the residential segregation of blacks from whites in the United States. Despite the many alternatives proposed, only the isolation index has really joined the dissimilarity index as a measure frequently used in empirical research on residential segregation. However, applications have extended substantially, to other racial and ethnic populations, and also to groups within and across races, defined by income, poverty, nativity, or other characteristics. Virtually all segregation indices are constructed by aggregating the observed distributions of the 'minority' and referent populations across a metropolitan area or city divided into small geographic areas—usually census tracts or census blocks in the United States. These geographic units are sufficiently small to be taken as rough approximations of neighborhoods whose residents would acknowledge and share some sense of common spatial identity, and hence some sense of community.

Massey and Denton's (1988) review and analysis of

20 indices that either had been or could be proposed for measuring distinct aspects of residential segregation represents the major watershed in the field. Massey and Denton suggested that the 20 indices measured five distinct dimensions of residential segregation:

(a) Evenness: indices that measure how the observed areal distributions of the minority and majority group populations in a metropolitan area deviate from distributions representing various definitions and criteria for equality or randomness.

(b) Exposure: indices based on the probability that the next person a minority group member encounters in their neighborhood or geographic area will also be a member of that group, or will instead be a member of the referent group.

(c) Concentration: indices that measure the degree to which a minority population is concentrated in a relatively small number of compact tracts.

(d) Centralization: indices that measure the degree to which a minority population resides close to the center or 'downtown' areas of a city.

(e) Clustering: indices that measure the degree to which the minority population disproportionately resides in neighborhoods or geographic units that adjoin one another, and hence form clusters or enclaves.

1. *Measures of Evenness*

The index of dissimilarity remains by far the most frequently used in empirical research on residual segregation. It is built on the premise that a metropolitan area or city would be integrated completely if all of its census tracts (or other areal units) had exactly the proportion of minority group and majority group residents as the metropolitan area or city as a whole. The index is calculated as the average deviation of the tracts or areal units from this proportionate representation, where the deviation is tract i is $|p_i - P|$. Thus,

$$D = \left(\sum_{1}^{n} [t_i |p_i - P|]\right) \Big/ [2TP(1-P)]$$

The resulting dissimilarity score can be interpreted as representing the percentage of the metropolitan area's minority residents who would have to move from tracts or areal units where the minority group is over-represented to areal units where they are under-represented to achieve proportionate representation or complete integration.

The primary criticism of the dissimilarity index is that it is insensitive to moves (or 'transfers') of minority group members among tracts where they are over-represented or under-represented. Several indices developed in literatures measuring income inequalities overcome this violation of the 'transfers principle.' The Gini coefficient averages the differences in the proportions of the minority group in every pair of tracts or areal units in the metropolitan area, where $|p_i - p_j|$ represents the difference in these proportions for tracts i and j

$$G = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} [t_i t_j |p_i - p_j|]\right) \Big/ [2T^2 P(1-P)]$$

The Atkinson measures allow researchers to develop indices that weight tracts where the minority group is under-represented ($p_i < P$) and over-represented ($p_i > P$) equally ($b = 0.5$), or to weight under-representation ($0 < b < 0.5$) or over-representation ($0.5 > b > 1.0$) more heavily

$$A = [P/(1-P)] \left| (1/PT) \left(\sum_{1}^{n} (1-p_i)^{(1-b)} p_i^b t_i \right) \right|^{1/(1-b)}$$

The theoretical or conceptual value of the Atkinson indices lies in applications where one might be more interested in reducing under-representation (e.g., integrating 'lily-white' suburbs) or in reducing over-representation (e.g., desegregating all black or Hispanic ghettos or barrios).

Entropy or information indices are based on the concept that entropy is maximized and information minimized when distributions are in equilibrium. In segregation applications, this would occur when the proportions of the minority and majority group populations were equal (50 percent for two groups) in a metropolitan area. The entropy index measures the percentage of the metropolitan area's entropy that is represented by the average deviation of entropy in each tract or areal unit from that in the metropolitan area as a whole. Thus, where entropy in each tract i and in the metropolitan area or city are represented respectively, by

$$E_i = p_i \log(1/p_i) + (1-p_i) \log(1/(1-p_i))$$

and

$$E = P \log(1/P) + (1-P) \log(1/(1-P))$$

then

$$H = \left(\sum_{1}^{n} [t_i (E - E_i)]\right) \Big/ ET$$

2. *Measures of Exposure*

Scores on the entropy index depend on P, the proportion that the minority group represents of the metropolitan area or city population. This violates a

criterion that proposes the segregation indices should have 'compositional invariance,' that is, independent of the proportions of minority and majority group members in different cities or metropolitan areas. However, in many critical applications such as school desegregation, the very low percentages of minority or majority group students in a district constitute major barriers to increasing racial integration or balance. The conceptual appropriateness and theoretical value of indices that are influenced by the relative proportions of minority and majority group members is thus undeniable, and they are among the few measures besides the dissimilarity index that are used extensively in empirical research.

The interaction and isolation measures are designed specifically to measure segregation defined by the degree to which the relative sizes of the minority and majority populations, as well as their distributions across neighborhoods or areal units, affect the chances of their availability to interact with one another in those neighborhoods. The interaction index is the average, weighted by the proportion of the metropolitan area's minority group living in tract i, of the majority group's proportion in each tract or areal unit i. It can be interpreted as representing the probability that the next person that a minority resident of a neighborhood or tract encounters will belong to the majority population.

$$ {}_xP_y^* = \sum_1^n [(x_i/X)(y_i/t_i)] $$

Conversely, the isolation index measures the likelihood that minority residents in a tract encounter only each other, and is calculated as the minority-weighted average of the minority group's proportion in each other or areal unit i

$$ {}_xP_x^* = \sum_1^n [(x_i/X)(x_i/t_i)] $$

Because the interaction and isolation indices depend on the composition of the population, ${}_xP_y^*$ and ${}_xP_x^*$ will be equal only if the majority and minority groups represent equal proportions of the population. The correlation ratio, Eta^2, adjusts for the effects of proportion minority of the population, and indeed can be classified as an evenness measure.

$$ V = Eta^2 = ({}_xP_x^* - P)/(1-P) $$

3. *Measures of Concentration*

Concentration indices seek to measure the extent to which the minority population occupies less of the metropolitan area's or city's geographic area than the majority or total population, and hence also lives at higher densities. If the minority population was uniformly distributed throughout the city or metropolitan area, the proportion of the total minority population in tract or areal unit i would be equal to that tract's proportion of the total area in the metropolitan area (a_i/A). Delta measures concentration by aggregating the deviations from this expectation

$$ \Delta = 0.5\left(\sum_1^n |(x_i/X) - (a_i/A)|\right) $$

Like the index of dissimilarity, the score on delta can be interpreted as representing the percentage of the minority group population that would have to move from areas of higher to lower than average density for the group to live at uniform densities across the metropolitan area.

Massey and Denton (1988) proposed new measures of concentration. The absolute concentration index compares the observed distribution of the minority group in the metropolitan area against the theoretical minimum and maximum areas they could occupy. If tracts are ordered and the total population is cumulated from smallest to largest in area, the minimum area that could accommodate the minority group population is reached at tract n_1 where the cumulated population reaches or exceeds the total minority population (X) of the metropolitan area. Analogously, the maximum area that the minority population could occupy is obtained by cumulating the total population from the tracts largest in area to smallest, and finding the point n_2 where the cumulated population equals or exceeds the minority total.

$$ ACO = 1 - \left\{\left[\sum_1^n (x_i a_i/X) - \sum_1^n (t_i a_i/T_1)\right] \Big/ \left[\sum_{n_2}^n (t_i a_i/T_2) - \sum_1^{n_1} (t_i a_i/T_2)\right]\right\} $$

The relative concentration index compares the relative concentration observed for the minority and majority populations with the maximum ratio possible, which occurs if the minority population were concentrated in the smallest possible area, and the majority population spread across the largest.

$$ RCO = \left\{\left[\sum_1^n (x_i a_i/X)\right] \Big/ \left[\sum_1^n (y_i a_i/Y)\right] - 1\right\} \Big/ \left\{\left[\sum_1^n (t_i a_i/T_1)\right] \Big/ \left[\sum_{n_2}^n (t_i a_i/T_2)\right] - 1\right\} $$

The segregation indices usual range from 0.0 to 1.0. In contrast, the relative concentration index ranges from -1.0 to 1.0, with the negative values indicating that the minority population is less concentrated than the majority.

4. Measures of Centralization

Because older and cheaper housing is usually closer to the center of the city, and because housing discrimination prevented blacks and other minorities in the United States from suburbanizing for decades after whites did, minority populations are often concentrated in neighborhoods closer to the center of the city. Both centralization measures order the metropolitan area's tracts or areal units by increasing distance from the central business district. The absolute centralization index aggregates the deviations of the cumulative proportion of the minority (X_i) reached at each tract i from the proportion of the land area reached by that tract

$$ACE = \sum_{1}^{n} (X_{i-1}A_i) - \sum_{1}^{n} (X_i A_{i-1})$$

The relative centralization index, aggregates the deviations of the cumulative proportions of the minority (X_i) and majority (Y_i) populations reached at each tract i from what would be expected were they comparably distributed

$$RCE = \sum_{1}^{n} (X_{i-1}Y_i) - \sum_{1}^{n} (X_i A_{i-1})$$

The ACE ranges from 0.0 to 1.0, but the RCE from -1.0 to 1.0, with negative values indicating that the minority population is less centralized than the majority.

5. Measures of Clustering

The final dimension of residential segregation measures the degree to which tracts with disproportionately large minority populations adjoin one another, creating, for example, large ghettos or barrios, or are scattered throughout the metropolitan area. This is a form of the geographer's contiguity and checkerboard problems, and drawing on this literature, Massey and Denton (1988) proposed as a measure of absolute clustering.

$$ACL = \left\{ \sum_{1}^{n} \left[(x_i/X) \sum_{1}^{n} (c_{ij}x_j) \right] - \left[(X/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} (c_{ij}x_j) \right] \right\} \Big/ \left\{ \sum_{1}^{n} \left[(x_i/X) \sum_{1}^{n} (c_{ij}t_j) \right] - \left[(X/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} (c_{ij}) \right] \right\}$$

where $c_{ij} = \exp(-d_{ij})$, that is, the negative exponential of the distance d_{ij} between the centroids of tracts i and j (and where $d_{ij} = (0.6a_i)^{0.5}$). The negative exponential is used to estimate the otherwise massive contiguity matrix whose elements equal 1.0 when tracts i and j are contiguous, and 0.0 when they are not. The formula takes the average number of minority residents in nearby tracts as a proportion of the total population in those tracts. The absolute clustering index has a minimum of 0.0 and can approach, but never reach, 1.0.

The spatial proximity index compares the average proximity of minority group residents to one another (P_{xx}), and the majority group residents with one another (P_{yy}) to the average proximity among the residents in the total population (P_{tt}), weighted by the fraction that each group represents in the population. Thus

$$SP = (XP_{xx} + YP_{yy})/TP_{tt}$$

where

$$P_{xx} = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}x_j x_j / X^2$$

and P_{yy} and P_{tt} are calculated by analogues of P_{xx}. Comparing the relative intragroup proximities of the minority and majority populations produces a measure of relative clustering.

$$RCL = (P_{xx}/P_{yy}) - 1$$

Clustering can also be measured by extending the concepts underlying the interaction and isolation measures to estimate how these probabilities should decay with distance. The distance–decay interaction and isolation indices can be constructed by aggregating, over each tract i, the probability that the next person one meets anywhere in the city is respectively a majority or minority resident from tract j, and that this probability exponentially decays with increasing distance between i and j. Thus

$$DP_{xy} = \sum_{1}^{n} \left[(x_i/X) \sum_{1}^{n} (k_{ij}y_j/t_{ij}) \right]$$

and

$$DP_{xx} \sum_{1}^{n} \left[(x_i/X) \sum_{1}^{n} (k_{ij}x_j/t_j) \right]$$

where

$$k_{ij} = [t_{ij} \exp(-d_{ij})] \Big/ \left[\sum_{1}^{n} (t_j \exp(-d_{ij}) \right]$$

6. Conclusion

Few empirical researchers have used more than one or two of these 20 indices in studying residential segregation, and the full spectrum remains of interest primarily to those devoted to constructing measures of the phenomenon in its varied dimensions. Massey and Denton (1988) used factor analysis to extract one index from each dimension and to define a concept of

hypersegregation for blacks in the United States based on their high segregation scores on each of the five dimensions. This has extended the use of measures beyond the dissimilarity and interaction/isolation indices. However, the literature is still far from developing an over-arching framework that might suggest how several measures might be used systematically to identify how urban areas vary in the five-dimensional space suggested or with what consequences. One might imagine, for example, identifying and comparing urban areas where high unevenness results from low integration in the suburbs rather than large ghettos; those with concentrated and perhaps centralized ghettos or barrios vs. others with comparably high unevenness or isolation but less clustered distributions of the minority population. The full set of indices has much potential value that remains to be developed.

See also: Ethnic Conflicts; Locational Conflict (NIMBY); Population Composition by Race and Ethnicity: North America; Race and Urban Planning; Racial Relations; Residential Concentration/Segregation, Demographic Effects of; Residential Segregation: Sociological Aspects; Urban Geography; Urban Sociology

Bibliography

Duncan O D, Duncan B 1955 A methodological analysis of segregation indices. *American Sociological Review* **20**: 210–17

Massey D S, Denton N A 1988 The dimensions of residential segregation. *Social Forces* **67**: 281–315

Taueber K E, Taueber A F 1965 *Negroes in Cities: Residential Segregation and Neighborhood Change*. Aldine, Chicago

R. Harrison

Selection Bias, Statistics of

In almost all areas of scientific inquiry researchers often want to infer, on the basis of sample data, the characteristics of a relationship that holds in the relevant population. Frequently this is the relationship between a dependent variable and one or more explanatory variables. In some cases, however, although one may have complete information on the explanatory variables, information on the dependent variable is lacking for some observations or 'units.' Furthermore, whether or not this information is present may not be conditionally independent (given the model) of the value taken by the dependent variable itself. This is a case of selection bias.

1. Selection Bias

The following illustrates the basic problem. Suppose one hypothesizes the following relationship, in the population, between a dependent variable, Y^*, and a set of $j = 1,\ldots, J$ explanatory variables, X

$$y_i^* = x_i'\beta + \varepsilon_i \tag{1}$$

Here the unit-specific values of Y^* are denoted y_i^* (where i denotes the ith observation) and each unit's values of the X variables are contained in the vector x_i. ε is a random error and the parameter vector, β, is to be estimated. Specific assumptions about the distribution of ε will determine the choice of estimation technique.

If Y^* is considered as a latent variable, a variety of different statistical models can be generated from Eqn. (1) depending on the manner in which Y^* is actually observed (though still, at this point, assuming that Y^* is observed, in some fashion, for all units). To do this, write a second equation that defines the relationship between the observed variable Y, and the latent variable, Y^*. For example, where Y^* is completely observed

$$y_i = y_i^* \tag{2a}$$

but one might also have

$$\begin{aligned} y_i &= 1 \quad \text{if } y_i^* > 0 \\ y_i &= 0 \quad \text{otherwise} \end{aligned} \tag{2b}$$

In the case of Eqn. (2(b)) Y is the binary observed realization of the underlying latent variable, Y^*. More complex relationships between Y^* and Y are also possible.

To capture the idea of selection bias, suppose that there is another latent variable Z^* such that whether or not we observe a value for Y depends on the value of Z^* which is given by

$$z_i^* = w_i' \alpha + v_i \tag{3}$$

Here w is a vector of explanatory variables with coefficients α, and v is a random error term. The observed variable Z is defined as

$$\begin{aligned} z_i &= 1 \quad \text{if } z_i^* i > 0 \\ z_i &= 0 \quad \text{otherwise} \end{aligned} \tag{4}$$

Finally, define the observation equation for Y as depending on Z^* as follows

$$\begin{aligned} y_i &= y_i \quad \text{if } z_i = 1 \\ y_i &\text{ unobserved if } z_i = 0 \end{aligned} \tag{5}$$

Equations (1), (2), and (5), together with Eqns. (3) and (4), link the latent variable Y^* to its observed counterpart, Y, when observation of the latter depends on the value of another variable, Z^*. Selection bias occurs when there is a nonzero covariance between the error terms ε and v. More complex selection bias models can be generated by having, for instance, more than one Z^* variable. The censored regression (or Tobit) model is a simpler case in which whether Y is observed or not depends on whether it exceeds (or, in other cases, falls below) a given threshold value.

2. The Problem

Selection bias is a problem because if one tries to estimate β using normal statistical methods the resulting estimates will be poor and potentially misleading. For example, if Eqn. (2(a)) specifies the relationship between Y and Y^*, discarding cases where Y is unobserved and running an OLS regression on the remainder will give estimates that are biased and inconsistent.

Where might this kind of selection bias arise? Suppose one draws a random sample from a population and carries out a survey. Some respondents may refuse to answer some, though not all, questions. If the response to such a question played the role of dependent variable in a regression analysis there would be a selection bias problem if the probability of having responded to the item was not independent of the response value, given the specification of the regression model. This may not always be so: in the simplest case nonresponse might be random. Alternatively, response might be independent of the response value controlling for a set of measured covariates. In this case there is selection on observables (Heckman and Robb 1985). But often neither of these cases holds and a nonzero residual correlation between the selection Eqn. (3) and the outcome Eqn. (1) means that there is selection on unobservables.

In studies of the criminal justice system in which the population is all those charged with a crime, sentence severity is observed only for those found guilty. In studies of the effectiveness of university education the outcome (say, examination results) is only observed for those who had completed that period of education. In studies of women's earnings in paid work the earnings of women who are not in the labor force cannot be observed. To these examples could be added very many more. Selection bias is a pervasive problem. As a consequence a variety of methods has been proposed to deal with it.

The problem of selection bias arises because of the nonzero correlation of the errors in Eqns. (1) and (3), and this arises commonly in the following context. Suppose there is a nonexperimental comparison of two groups, one exposed to some policy treatment, the other not. In this case the outcome of interest, Y, could be observed for members of both groups. Let Z now be the indicator of group membership (so that, for instance, $Z = 0$ means membership of the comparison group and $Z = 1$ means membership of the treatment group). In the case where Eqn. (2(a)) holds write the equation for Y as

$$y_i = x_i' \beta + \gamma z_i + \varepsilon_i \tag{6}$$

and interest would focus on the estimate of γ as the gross effect on the outcome measure of being in the treatment, rather than the comparison, group. The problem of selection bias still arises to the extent that ε and v have a nonzero correlation. The difference between this and the type of selection bias discussed initially (namely that now Y is observed for both groups) is more apparent than real as an alternative formulation shows. For each unit Y is observed as the response to membership of either the treatment or comparison group, but not both. Let Y_1 be the outcome given treatment and Y_0 given no treatment. Then the ith individual unit's gain from treatment is simply $\Delta_i = Y_{1i} - Y_{0i}$. But Δ cannot be measured since one cannot simultaneously observe a unit's values of Y_1 and Y_0. Instead

$$Y_i = ZY_{1i} + (1 - Z)Y_{0i} \tag{7}$$

is observed. Eqn. (7) thus specifies the incomplete observation of two latent variables, Y_0 and Y_1. This set-up can easily be extended to cases with more than two groups. As might be expected, this approach is commonly encountered in the evaluation literature but one of its earliest formulations was by Roy (1951) as the problem of self-selection.

3. The Solutions

Broadly speaking there are two ways to address problems of selection bias: *ex ante* by research design, and *ex post* by statistical adjustment. The modern literature on correcting for selection bias is itself biased towards the latter, beginning with the work of Heckman in the 1970s (Heckman 1976, 1979) whose so-called 'two step' method is widely used and hard-wired into many econometrics programs. The argument that underlies this technique is the following. Assume the set-up as defined by Eqns. (1), (2(a)), (3), (4), and (5) and a nonzero covariance between ε and v. Then we can write the 'outcome equation' (i.e., the regression equation for Y, given its observability) as

$$E(y_i | z_i = 1, x_i) = x_i' \beta + E(\varepsilon_i | z_i = 1) \tag{8}$$

Because there are only observations of Y when $z = 1$ there is an extra term for the conditional expectation

of ε. Because of the nonzero covariance between ε and v and if, as is usually assumed, $E(\varepsilon) = 0$, then this conditional expectation cannot be zero. One can write

$$E(y_i|z_i = 1, x_i) = x'_i\,\beta + E(\varepsilon_i|v_i > w_i'\alpha) \qquad (9)$$

Using a standard result for the value of a truncated bivariate normal distribution, the second term on the right-hand side of this equation is given by

$$E(\varepsilon_i|v_i > w'_i\,\alpha) = \rho\sigma_\varepsilon\,\sigma_v\frac{\phi(w_i'\,\alpha)}{\Phi(w'_i\,\alpha)} \qquad (10)$$

Here the σs are the standard deviations of the respective error terms from Eqns. (1) and (3); ρ is the correlation between them; ϕ and Φ are, respectively, the density and distribution functions of the standard normal and their ratio, as it appears in Eqn. (10), is termed the 'inverse Mill's ratio.' Heckman's two-step method requires first running a probit regression with Z as the dependent variable, estimated using all the observations in the data. Then, for those observations for which Y is observed, the probit coefficient estimates are used to compute the value of the inverse Mill's ratio. This is then inserted as an extra variable in the OLS regression in which Y is the dependent variable to give

$$E(y_i|z_i = 1, x_i) = x_i'\,\beta + \theta\hat{\lambda}_i \qquad (11)$$

where λ is the inverse Mill's ratio and the 'hat' indicates that it is estimated. Its coefficient, θ, is then itself an estimate of $\rho\sigma_\varepsilon\sigma_v$. This is the covariance between ε and v (σ_v is set to unity in the probit). This approach is extended readily to the case in which an outcome is observed for both groups ($Z = 0$ and $Z = 1$).

What are the assumptions of this approach and what are the properties of the resulting estimator when we apply this method to data from a sample of the population? First, it is assumed that all the other requirements for an OLS regression are met. Second, in the set-up outlined above, the joint distribution of ε and v should be bivariate normal (though, in general, weaker assumptions suffice, namely that v be normally distributed; and the expectation of ε, conditional on v, should be linear in v: (see Olsen 1980)).

Given that the assumptions hold, the two-step estimator yields consistent estimates of the population β in Eqn. (1). The standard errors are incorrect (due to heteroscedasticity and the use of an estimated value of λ) but these are corrected readily. An alternative to the two-step approach is to estimate both the selection and outcome equations simultaneously using maximum likelihood (ML). The resulting estimates are asymptotically unbiased and more efficient than those from the two-step method (see Nelson 1984, who compares OLS, the two-step method and ML).

This two-step method is, in any case, applicable only when the relationship between Y^* and Y is as given by Eqn. (2(a)). When this relationship is given by, for example, Eqn. (2(b)), the two-stage method is inconsistent. In these cases ML is the most feasible option. If the outcome variable is itself binary, the joint log-likelihood of the selection and outcome equations has the same general as for a bivariate probit but one in which there are three, rather than four, possible outcomes. They are $z = 1$ and $y = 1$; $z = 1$ and $y = 0$; and $z = 0$ (in which case y is not observed).

Although the two-step method is probably the most widely used approach to correcting for selection bias it has been subjected to much criticism. Among the main objections are the following:

(a) *sensitivity to distributional assumptions*. Practical implementation of the method renders it particularly sensitive in this respect. If the assumptions are not met the estimator has no desirable properties (i.e., it is not even consistent).

(b) *identification and robustness*. It is common to find that the estimate of the inverse Mill's ratio is highly correlated either with the explanatory variables or with the intercept in Eqn. (11). The extent to which such problems will arise depends mainly on three things. They are:

(i) the specification of the selection equation;

(ii) the sample correlation between the two sets of explanatory variables, X and W (call this q);

(iii) the degree of sample selection in the sample (p, the proportion of cases for which $z = 1$).

For example, if X and W are identical, the two-equation system is identified only because of the nonlinearity of the probit equation. But for some ranges of p the probit function is almost linear, with the result that the estimated λ will be close to a linear function of the explanatory variables in the probit; and thus it will be highly correlated with the X variables in the outcome equation. In general, the correlation between the inverse Mill's ratio estimate and the other explanatory variables in this equation will be greater the greater is q and the closer is p to zero or one (this issue is discussed in detail in Breen 1996). If the selection equation does not discriminate well between the selected and unselected observations the estimated inverse Mill's ratio will be approximately a constant, and there will therefore be a high correlation between it and the intercept of the outcome equation.

Both these objections reflect genuine difficulties in applying the model. In principle the solution to the identification and robustness problems is simple: ensure that the probit equation discriminates well and do not rely on the nonlinearity of the probit for identification. In practice it may be rather difficult to achieve these things. On the other hand, the issue of distributional assumptions may be even less tractable, obliging recourse to semiparameteric or other approaches (Lee 1994; Cosslett 1991).

An alternative is to try to assess the likely degree of bias that sample selection might induce in a given case. Rubin (1977) presents a Bayesian approach in which the investigator computes the likely degree of bias, conditional on a prior belief about the relationship between the parameters of the distribution of Y in the selected and nonselected samples. The mean and the variance of the latter sample might, for example, be expressed as a function of their values in the selected sample, conditioning on the values of observed covariates. It is then straightforward to express the extent of selection bias for plausible values of the parameters of this function and to place a corresponding Bayesian probability interval around any estimates. Rosenbaum (1995, 1996) suggests and illustrates other sensitivity analyses for selection bias in nonexperimental evaluations.

4. *Program Evaluation and Selection Bias*

For several years now the literature on selection bias has been dominated by discussion of how to evaluate programs (such as job training programs) when randomized assignment to treatment and control group is not possible. In a very widely cited paper, Lalonde (1986) compared the measured effectiveness of labor market programs using a randomized assignment design with their effectiveness computed using the Heckman method and showed that they led to quite different results (but see also Heckman and Hotz 1989). While some have seen this as a fatal criticism of the method, one consequence has been the development of greater awareness of the need to ensure the suitability of different selection bias correction approaches for specific cases. Another has been a greater concern to identify other possible sources of bias in nonrandomized program evaluations.

In the statistical literature, matching methods commonly are advocated in program evaluations in the absence of randomization. They involve the direct comparison of treated and untreated units that share common, or very similar, values on a set of relevant covariates, X. The rationale for this is the assumption that, conditional on X, and assuming only selection on observables (all of which are in X), the observed outcome for nonparticipants has the same distribution as the unobserved outcome that participants would have had had they not participated (see Heckman et al. 1997). This then allows one to estimate Δ. In passing, note that matching is also used commonly to impute missing values in item nonresponse in surveys (Little and Rubin 1987). A central issue is how to carry out such matching. If there are K covariates, then the problem is one of matching pairs, or sets, of participants and nonparticipants in this K-dimensional space. But by a result due to Rosenbaum and Rubin (1983) (see also Rosenbaum 1995), matching using a scalar quantity called the propensity score is equally effective. The propensity score is simply the probability of being in the treatment, rather than the comparison group, given the observed covariates. Matching on the propensity score controls bias due to all observed covariates and, even if there is selection on unobservables, the method nevertheless produces treatment and comparison groups with the same distribution of X variables. This is so under the assumption that the true propensity score is known, but Rosenbaum and Rubin (1984) show that an estimated propensity score (typically using a logit model) appears to perform at least as well.

The use of matching draws attention to the problem of the possibly different distributions of covariates (or, equally, of propensity scores) in the treatment and comparison groups. There are two important aspects: first, the support of the propensity score may be different, so some ranges of propensity score values may be present in one group and not the other. More simply, some participants may have no comparable non-participants. Second, the distributions of the set of common values of propensity scores (i.e., which appear in both groups) may be different (Heckman et al. 1997, 1998). Hitherto, in practice, both of these sources of bias in evaluating the program's effects commonly would have been confounded with selection bias proper. The use of propensity scores can remove the former but, if there is also selection on unobservables (i.e., selection bias proper), matching cannot be expected to solve the problem. Indeed, it is possible that these different forms of bias may have opposite signs, so that correcting for some but not others may make the overall bias greater. Nevertheless, the ability to draw such finer distinctions among biases is a valuable development, not least because it allows one to focus on methods for controlling selection bias, free from the contaminating effects of other biases.

5. *Conclusion*

This article has only skimmed the surface of selection bias statistics. In particular, it has looked only at models for cross-sectional data: other approaches are possible given other sorts of data. For example, given longitudinal data, with observations of Y for each unit at two (or more) points in time, one prior to the program to be evaluated, the other after it, then it may be possible to specify and test a model which assumes that unobserved causes of selection bias are unit-specific and time-invariant and can therefore be removed by differencing (Heckman and Robb 1985, Heckman et al. 1997, 1998). Considerations of this sort draw attention to the *ex ante* approach to dealing with selection bias that was referred to earlier. Here the emphasis falls on designing research so as to remove, or at least reduce, selection bias. Random assignment is one possibility, though not always feasible in practice. Nevertheless, quasi-experimental approaches, interrupted time-series, and longitudinal

designs with repeated measures on the outcome variable are among a range of possibilities that could be considered in designing ones research to minimize selection bias problems (see Rosenbaum 1999 and associated comments). Even so, there will still be many instances in which analysts use data over whose collection they have had no control and where the use of *ex post* adjustments will be unavoidable.

To conclude: much work on the selection bias problem has been undertaken since the 1980s yet there is no widespread agreement on which statistical methods are most suitable for use in correcting for the problem. Some broad guidelines for practitioners can, however, be discerned:

(a) the need to ensure that methods to correct for selection bias are appropriate and that their requirements (such as distributional properties) are met;

(b) the need to be aware of other, possibly confounding, sources of bias;

(c) the usefulness of analyses of the sensitivity of conclusions to various possible magnitudes of selection bias; and of the sensitivity of the selection-bias corrected results to the assumptions of whatever method is employed; and

(d) the desirability of designing research so that selection bias problems are, as far as possible, eliminated without the need for complex and sometimes fragile *ex post* adjustment.

See also: Longitudinal Research: Panel Retention; Mortality Differentials: Selection and Causation; Nonequivalent Group Designs; Screening and Selection

Bibliography

Breen R 1996 *Regression Models: Censored, Sample-selected or Truncated Data*. Sage, Thousand Oaks, CA

Cosslett S 1991 Semiparametric estimation of a regression model with sample selectivity. In: Barnett W A, Powell J, Tauchen G (eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge, UK

Heckman J J 1976 The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5**: 475–92

Heckman J J 1979 Sample selection bias as a specification error. *Econometrica* **47**: 153–61

Heckman J J, Hotz V J 1989 Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association* **84**(408): 862–74

Heckman J J, Ichimura H, Smith J, Todd P E 1998 Characterizing selection bias using experimental data. *Econometrica* **66**: 1017–98

Heckman J J, Ichimura H, Todd P E 1997 Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* **64**: 605–54

Heckman J J, Robb R 1985 Alternative methods for evaluating the impact of interventions. In: Heckman J J, Singer B (eds.) *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, New York

Lalonde R J 1986 Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* **76**: 604–20

Lee L F 1994 Semi-parametric two-stage estimation of sample selection models subject to Tobit-type selection rules. *Journal of Econometrics* **61**: 305–44

Little R D, Rubin D B 1987 *Statistical Analysis with Missing Data*. Wiley, New York

Nelson F D 1984 Efficiency of the two-step estimator for models with endogenous sample selection. *Journal of Econometrics* **24**: 181–96

Olsen R J 1980 A least squares correction for selectivity bias. *Econometrica* **48**: 1815–20

Rosenbaum P R 1995 *Observational Studies*. Springer-Verlag, New York

Rosenbaum P R 1996 Observational studies and nonrandomized experiments. In: Ghosh S, Rao C R (eds.) *Handbook of Statistics*. Elsevier, Amsterdam, Vol. 13

Rosenbaum P R 1999 Choice as an alternative to control in observational studies. *Statistical Science* **14**: 259–304

Rosenbaum P R, Rubin D B 1983 The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55

Rosenbaum P R, Rubin D B 1984 Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**: 516–24

Roy A D 1951 Some thoughts on the distribution of earnings. *Oxford Economic Papers* **3**: 135–46

Rubin D B 1977 Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* **72**: 538–43

R. Breen

Self-concepts: Educational Aspects

The capacity to reflect one's own capabilities and actions is uniquely human. Early in life, young children begin to form beliefs about themselves which may serve as reference mechanisms for perceiving the world and oneself, and for regulating emotion, motivation, and action. Research on self-related beliefs can be traced back to the seminal writings of William James (1893) who distinguished the self as 'I' ('existential self') and as 'me' ('categorical self'), the latter implying cognitions about the self as an object of thinking (as represented in a sentence like 'I think about me'). Research was continued throughout the decades after that and gained a central status in personality and social psychology as well as educational research after the cognitive paradigm shift in

the 1950s and 1960s. At that time, behavioristic approaches were gradually replaced by cognitive perspectives, thus making it possible to acknowledge the importance of cognitive processes in human agency. Since then, the number of studies on self-concepts has increased drastically. This also applies to studies on the educational relevance of self-concepts (Hansford and Hattie 1982, Helmke 1992).

1. Definition of the Term 'Self-concept'

The term 'self-concept' is used in two interrelated ways. Talking about *the* self-concept of a person has to be differentiated from referring to a number of *different* self-concepts of this person. Usage of the term in self-concept research implies that 'the' self-concept may be defined as the total set of a person's cognitive representations of him- or herself which are stored in memory. In the plural sense, different self-concepts of a person are subsets of such representations relating to different self-attributes (like abilities, physical appearance, social status, etc.). In both variants, the term implies that self-concepts are self-representations which are more or less enduring over time, in contrast to situational self-perceptions. The definition is open as to whether these representations refer to factual reality of personal attributes, or to fictitious (e.g., possible or ideal) attributes.

There are two definitional problems which remain conceptually unresolved and impede scientific progress. First, it is unclear whether self-related emotions should be included or not. One may argue that emotions are different from cognition and should be regarded as separate constructs. From such a perspective, (cognitive) self-concepts and (emotional) self-related feelings might be subsumed under umbrella constructs like *self-esteem*, comprising both cognitive and affective facets, but should not be mixed up conceptually. On the other hand, prominent measures of the construct take more integrative perspectives by combining cognitive and affective self-evaluative items. One example are H. Marsh's widely used Self-Description Questionnaires (SDQ: Marsh 1988) measuring academic self-concepts both by cognitive items (e.g., 'I am good at mathematics') and by affective items (e.g., 'I enjoy doing work in mathematics).'

Second, there is no common agreement on the conceptual status of self-related cognitions linking the self to own actions and the environment. An example are self-efficacy beliefs pertaining to own capabilities to be successful in solving specific types of tasks, thus cognitively linking the self (own capabilities) to own actions (task performance) and to environmental demands (a domain of tasks; cf. Pajares 1996; see *Self-efficacy: Educational Aspects*). It may be argued that such self-representations should be regarded as part of a person's self-concept as well. However, terms like self-concept, on the one hand, and self-efficacy beliefs, control beliefs etc., on the other, have been used as if they were relating to conceptually distinct phenomena, and have been addressed by different traditions of research. These research traditions tend to mutually ignore each other in spite of overlapping constructs and parallel findings, implying that more 'cross-talk' among researchers should take place in order to reduce conceptual confusion and the proliferation of redundant constructs which still prevails at the beginning of the twenty-first century (Pajares 1996).

2. Facets and Structures of Self-concepts

Self-related cognitions can refer to different attributes of the self, and can imply descriptive or evaluative perspectives pertaining to the factual or nonfactual (e.g., ideal) reality of these attributes. Research has begun to analyze the structures of these different representational facets.

2.1 Representations of Attributes: Hierarchies of Self-concepts

Self-concepts pertaining to different attributes may vary along two dimensions: the domain of attributes which is addressed, and their generality. It may be theorized that self-concepts are hierarchically organized along these two dimensions in similar ways as semantic memory networks can be assumed to be, implying that more general self-concepts are located at top levels of the self-concept hierarchy, and more specific self-concepts at lower levels. In educational research, the hierarchical model put forward by Shavelson et al. (1976) stimulated studies on facets of self-concepts differing in generality. This model implied that a 'general self-concept' is located at the top of the hierarchy; an 'academic self-concept' pertaining to one's academic capabilities as well as social, emotional, and physical nonacademic self-concepts at the second level; self-concepts relating to different school subjects and to social, emotional, and physical subdomains at the third level; and more specific self-concepts implying evaluations of situation-specific behavior at lower levels.

Studies showed that correlations between academic self-concepts pertaining to different subjects tend to be near zero, in contrast to performance which typically is correlated positively across academic domains. An explanation has been provided by Marsh's internal–external frame of reference model (I/E model) positing that the formation of academic self-concepts may be based on internal standards of comparison implying within-subject comparison of abilities across domains, as well as external standards implying between-

subjects social comparison with other students (Marsh 1986). Applying internal standards may lead to negative correlations between self-concepts pertaining to different domains (e.g., perceiving high ability in math may lead to lowered estimates of verbal abilities, and vice versa). External standards would imply positive correlations since achievement in domains like mathematics and languages is positively correlated across students. The model assumes that students use both types of standards, implying that opposing effects may cancel out. Accordingly, the Shavelson et al. (1976) model has been reformulated by assuming that students hold math-related and verbal-related academic self-concepts, but no general academic self-concept (Marsh and Shavelson 1985).

2.2 *Different Perspectives on Attributes*

Self-representations can imply descriptive as well as evaluative accounts of attributes (e.g., 'I am tall' vs. 'I am an attractive woman'). In academic self-concepts, this distinction may often be blurred because descriptions of academic competence may use comparison information implying some evaluation as well (e.g., 'I am better at math than most of my classmates'). Any subjective evaluation of academic competence may use a number of different standards of evaluation, e.g. (a) social comparison standards and (b) intraindividual comparison across domains as addressed by Marsh's I/E model (see above), as well as (c) intraindividual comparison of current and past competence (implying an evaluation of one's academic development), (d) mastery-oriented comparison relating one's competence to content-based criteria of minimal or optimal performance in an academic domain, and (e) cooperative standards linking individual performance to group performance.

Beyond existing attributes ('real self'), self-representations can pertain to attributes of the self which do not factually exist, but might exist ('possible selves'), are wanted by oneself ('ideal self'), wanted by others ('ought self'), etc. Concepts of nonreal selves may be as important for affect and motivated self-development as concepts of the real self. For example, perceived discrepancies between ideal and real self may be assumed to trigger depressive emotion, whereas discrepancies between ought and real self may give rise to anxiety (Higgins 1987).

3. *Self-concepts and Academic Achievement*

The relation between self-concept and academic achievement is one of the most often analyzed problems in both self-concept and educational research. In the first stage of research on this problem, connections between the two constructs were analyzed by correlational methods based on cross-sectional field studies. Results implied that self-concepts and achievement may be positively linked. However, the magnitude of the correlation proved to depend on the self-concept and achievement constructs used. For example, in the meta-analysis provided by Hansford and Hattie (1982), the average correlation between self-concept and achievement/performance measures was $r = 0.22$ for general self-esteem, and $r = 0.42$ for self-concept of ability. When both self-concepts and achievement are measured in more domain-specific ways, correlations tend to be even higher (e.g., correlations for self-concepts and academic achievement in specific school subjects). This pattern of relations implies that the link gets closer when self-concept and criterion measure are matched, and when both are conceptualized in domain-specific ways.

In the second stage, researchers began to analyze the causal mechanisms producing these relations. From a theoretical perspective, the 'skill development' model maintained that academic self-concepts are the result of prior academic achievement, whereas the 'self-enhancement' approach implied that self-concepts influence students' achievement (cf. Helmke and van Aken 1995). In a number of longitudinal studies, both hypotheses were tested competitively by using cross-lagged panel analysis and structural equations modeling. Results showed that both hypotheses may be valid, thus implying that self-concepts and achievement may be linked by reciprocal causation: Prior academic achievement exerted effects over time on subsequent academic self-concepts, and self-concepts influenced later achievement, although effect sizes for the latter causal path were less strong and consistent (cf. Helmke and van Aken 1995, Marsh and Yeung 1997). This evidence suggests that academic self-concepts are partly based on information about own academic achievement, and may contribute to later academic learning and achievement.

Not much is known about the exact nature of these mechanisms. Investigators have just begun to explore possible mediators and moderators of self-concept/achievement relations. Judging from theory and available evidence, the following may be assumed.

3.1 *Effects of Achievement on Self-concepts*

Academic feedback of achievement (e.g., by grades) may underly the formation of self-representations of academic capabilities. The match between feedback and self-representations may depend on the cumulativeness and consistency of feedback, its salience, the reference norms used, and the degreee of congruency with competing information from parents, peers, or one's past. This implies that the impact of achievement on self-concepts may differ between schools, teachers, and classrooms using different standards and practices

of feedback (cf. Helmke 1992). Finally, beyond exerting effects of academic self-concepts, achievement feedback can be assumed to influence students' general sense of self-worth (Pekrun 1990), thus affecting students' overall psychological health and personality development as well.

3.2 Effects of Self-concepts on Achievement

Self-concepts may influence the formation of self-efficacy and success expectations when being confronted with academic tasks. These expectations may underly academic emotions and motivation, which may in turn influence effort, strategies of learning, and on-task behavior directly affecting learning and performance (Pekrun 1993). Research has shown that self-concepts implying moderate overestimation of own capabilities may be optimal for motivation and learning gains, whereas an underestimation may be rather detrimental. However, self-evaluations precisely reflecting reality may also be suboptimal for learning and achievement, indicating that there may be a conflict between educational goals of teaching students to be self-realistic vs. highly motivated (Helmke 1992).

4. Development of Self-concepts

4.1 Basic Needs Underlying Self-concept Development

Two classes of basic human needs seem to govern the development of self-representations (cf. Epstein 1973). One are general needs for maximizing pleasure and minimizing pain, from which needs for self-enhancement follow, implying motivation to establish and maintain a positive view of the self. The second category comprises needs to perceive reality and foresee the future in such ways that adaptation and survival are possible, thus implying needs for self-perceptions which are consistent with reality, and consistent with each other (needs for consistency). Self-enhancement and consistency may converge when feedback about the self is positive. However, they may be in conflict when feedback is negative: The need for self-enhancement would suggest not accepting negative feedback, whereas needs for reality-oriented consistency would imply endorsing it.

4.2 Developmental Courses across the Life Span

The development of self-concepts is characterized by the interplay of mechanisms driven by these conflicting basic needs. In general, in the absence of strong information about reality, humans tend to overestimate their capabilities in self-enhancing ways. At the same time, they search for self-related information, and tend to endorse this information even if it is negative on condition that such negative information is salient, consistent, and cumulative. For example, it has repeatedly been found that many children entering school drastically overestimate their competences to master academic demands, but adjust their self-evaluations downward when diverging cumulative feedback is given (Helmke 1992). The process of realistically interpreting self-related feedback may be impeded in young children by their lack of sufficiently sophisticated metacognitive competences. Nevertheless, the interplay of beginners' optimism and experts (relative) realism may characterize not only the early elementary years, but later phases of entering new institutions as well (university, marriage, a new job etc.), thus implying a dynamic interplay of self-related optimism and realism across much of the life span.

4.3 Impact of Educational Environments

The influence of social environments was addressed early in the twentieth century by symbolic interactionism postulating that interactions with significant others may shape the development of self-concepts (cf. the concept of the 'looking glass self,' Cooley 1902). Generally, a number of environmental variables may be influential. Self-concepts can develop according to direct attributions of traits and personal worth by other persons on condition that such attributions are consistent with other sources of information and interpreted accordingly. A second source of self-related information are indirect, implicit attributions which are conveyed by others' emotional and instrumental behavior towards the developing person. Of specific importance are acceptance and support by others implying attributions of personal worth, thereby influencing the development of a person's general self-esteem (Pekrun 1990).

Beyond attributions, social environments may define situational conditions for the development of knowledge, skills, motivation, and behavior, which may in turn contribute to self-perceptions of own competences. For example, instruction may build up knowledge and skills influencing knowledge-specific academic self-concepts; support of autonomy may foster the acquisition of self-regulatory abilities and, thereby, the development of related self-concepts of abilities; and consistent behavioral rules may enhance the cognitive predictability of students' environments which may also positively affect their overall sense of competence.

As outlined above, concerning academic self-concepts, feedback of achievement implying information about abilities and effort may be of specific importance, and the effects of such feedback may depend on

the reference norms used. Feedback given according to competitive, interindividually referenced norms implies that the self-concepts of high achievers may benefit, whereas low achievers may have difficulties protecting their self-esteem. In contrast, intraindividual, mastery-oriented and cooperative norms may be better suited to foster low achievers' self-concepts as well. Finally, classroom conditions may also be influential. Specifically, grouping of students may influence students' relative, within-classroom achievement positions, thus affecting their academic self-concepts when competitive standards of grading are used. For example, being in a low-ability class would help an average student to maintain positive academic self-concepts, whereas being in a class of highly gifted students would enforce downward adjustment of self-evaluation ('big-fish-little-pond effect,' Marsh 1987).

5. *Summary of Implications for Educational Practice*

Fostering students' self-concepts may be beneficial for their achievement and personality development. Furthermore, since positive self-esteem may be regarded as a key element of psychological health and well-being, nurturing self-esteem may be regarded as an educational goal which is valuable in itself. From the available research summarized above, it follows that education may contribute substantially to self-concept development. Concerning parents as well as academic institutions like the school, acceptance and support may be of primary importance for the development of general self-esteem. In addition, giving students specific feedback on achievement, traits, and abilities may help in shaping optimistic self-concepts and expectancies which nevertheless are grounded in reality. In designing feedback, it may be helpful to use mastery-oriented, individual, and cooperative reference norms instead of relying on social comparison standards and competitive grading. Furthermore, educational environments may help students' self-concept development by providing high-quality instruction, consistent normative and behavioral structures implying predictability, as well as sufficient autonomy support fostering students' sense of controllability and competence.

See also: Motivation, Learning, and Instruction; School Achievement: Cognitive and Motivational Determinants; Schooling: Impact on Cognitive and Motivational Development; Self-development in Childhood; Self-efficacy; Self-efficacy: Educational Aspects; Self-regulated Learning

Bibliography

Bracken B A (ed.) 1996 *Handbook of Self-concept*. Wiley, New York

Cooley C H 1902 *Human Nature and the Social Order*. Scribner, New York

Covington M V 1992 *Making the Grade*. Cambridge University Press, New York

Covington M V, Beery R 1976 *Self-worth and School Learning*. Holt, Rinehart and Winston, New York

Epstein S 1973 The self-concept revisited or a theory of a theory. *American Psychologist* **28**: 404–16

Hansford B C, Hattie J A 1982 The relationship between self and achievement/performance measures. *Review of Educational Research* **52**: 123–42

Hattie J 1992 *The Self-concept*. Erlbaum, Hillsdale, NJ

Helmke A 1992 *Selbstvertrauen und schulische Leistungen*. Hogrefe, Göttingen, Germany

Helmke A, van Aken M A G 1995 The causal ordering of academic achievement and self-concept of ability during elementary school: A longitudinal study. *Journal of Educational Psychology* **87**: 624–37

Higgins E T 1987 Self-discrepancy: A theory relating self and affect. *Psychological Review* **94**: 319–40

James W 1893 *Psychology*. Fawcett, New York

Kulik J A, Kulik C L 1992 Meta-analytic findings on grouping processes. *Gifted Child Quarterly* **36**: 73–7

Marsh H W 1986 Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal* **23**: 129–49

Marsh H W 1987 The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology* **79**: 280–95

Marsh H W 1988 *Self Description Questionnaire I (SDQ): Manual and Research Monograph*. Psychological Corporation, San Antonio, TX

Marsh H W, Shavelson R J 1985 Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist* **20**: 107–25

Marsh H W, Yeung A S 1997 Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology* **89**: 41–54

Pajares F 1996 Self-efficacy beliefs in academic settings. *Review of Educational Research* **66**: 543–78

Pekrun R 1990 Social support, achievement evaluations, and self-concepts in adolescence. In: Oppenheimer L (ed.) *The Self-concept*. Springer, Berlin, pp. 107–19

Pekrun R 1993 Facets of adolescents' academic motivation: A longitudinal expectancy-value approach. In: Maehr M, Pintrich P (eds.) *Advances in Motivation and Achievement*. JAI Press, Greenwich, CT, Vol. 8, pp. 139–89

Shavelson R J, Hubner J J, Stanton G C 1976 Self-concept: Validation of construct interpretations. *Review of Educational Research* **46**: 407–41

Skaalvik E M, Rankin R J 1995 A test of the internal/external frame of reference model at different levels of math and verbal self-perception. *American Educational Research Journal* **32**: 161–84

R. Pekrun

Self-conscious Emotions, Psychology of

Shame, guilt, embarrassment, and pride are members of a family of 'self-conscious emotions' that are evoked by self-reflection and self-evaluation. This self-evaluation may be implicit or explicit, consciously ex-

perienced or transpiring beyond our awareness. But either way, the self is the object of self-conscious emotions.

In contrast to 'basic' emotions (e.g., anger, fear, joy) which are present very early in life, self-conscious emotions have been described as 'secondary,' 'derived,' or 'complex' emotions because they emerge later and hinge on several cognitive achievements—recognition of the self separate from others and a set of standards against which the self is evaluated. For example, Lewis et al. (1989) showed that the capacity to experience embarrassment coincides with the emergence of self-recognition. Very young children first show behavioral signs of embarrassment during the same developmental phase (15–24 months) in which they show a rudimentary sense of self. Moreover, within this range, children who display self-recognition (in a 'rouge' test) are the same children who display signs of embarrassment in an unrelated task.

1. Shame and Guilt

The terms 'shame' and 'guilt' are often used interchangeably. When people do distinguish the two, they typically suggest that shame arises from public exposure and disapproval of a failure or transgression, whereas guilt is more 'private,' arising from one's own conscience. Recent empirical research has failed to support this public/private distinction. For example, Tangney et al. (1996) found that people's real-life shame and guilt experiences are each most likely to occur in the presence of others. More important, neither the presence of others nor others' awareness of the respondents' behavior distinguished between shame and guilt. Overall, there are surprisingly few differences in the types of events that elicit shame and guilt. Shame is somewhat more likely in response to violations of social norms, but most types of events (e.g., lying, cheating, stealing, failing to help another) result in guilt for some people and shame for others.

So what is the difference between shame and guilt? Empirical research supports Helen Block Lewis's (1971) notion that shame involves a negative evaluation of the global self, whereas guilt involves a negative evaluation of a specific behavior. This differential emphasis on self ('*I* did that horrible thing') vs. behavior ('I *did* that horrible *thing*') leads to different affective experiences.

Shame is an acutely painful emotion typically accompanied by a sense of shrinking, 'being small,' and by a sense of worthlessness and powerlessness. Shamed people also feel exposed. Although shame doesn't necessarily involve an actual observing audience, there is often the imagery of how one's defective self would appear to others. Not surprisingly, shame often leads to a desire to escape or hide—to sink into the floor and disappear.

In contrast, guilt is typically less painful and devastating because the primary concern is with a specific behavior, not the entire self. Guilt doesn't affect one's core identity. Instead, there is tension, remorse, and regret over the 'bad thing done' and a nagging preoccupation with the transgression. Rather than motivating avoidance, guilt typically motivates reparative action.

1.1 Implications of Shame and Guilt for Interpersonal Adjustment

Research has consistently shown that, on balance, guilt is the more adaptive emotion, benefiting relationships in a variety of ways (Baumeister et al. 1994, Tangney 1995). Three sets of findings illustrate the adaptive, 'relationship-enhancing functions' of guilt, in contrast to the hidden costs of shame.

First, shame typically leads to attempts to deny, hide, or escape; guilt typically leads to reparative action—confessing, apologizing, undoing. Thus, guilt orients people in a more constructive, proactive, future-oriented direction, whereas shame orients people toward separation, distance, and defense.

Second, a special link between guilt and empathy has been observed at the levels of both emotion states and dispositions. Studies of children, college students, and adults (Tangney 1995) show that guilt-prone individuals are generally empathic individuals. In contrast, shame-proneness is associated with an impaired capacity for other-oriented empathy and a propensity for 'self-oriented' personal distress responses. Similar findings are evident when considering feelings of shame and guilt 'in the moment.' Individual differences aside, when people describe personal guilt experiences, they convey greater empathy for others, compared to descriptions of shame experiences. It appears that by focusing on a bad behavior (as opposed to a bad self), people experiencing guilt are relatively free of the egocentric, self-involved focus of shame. Instead, their focus on a specific behavior is likely to highlight the consequences for distressed others.

Third, there is a special link between shame and anger, again observed at both the dispositional and state levels. At all ages, shame-prone individuals are also prone to feelings of anger and hostility (Tangney 1995). Moreover, once angered, they tend to manage their anger in an aggressive, unconstructive fashion. In contrast, guilt is generally associated with constructive means of handling anger. Similar findings have been observed at the situational level, too. In a study of couples' real-life episodes of anger, shamed partners were significantly more angry, more aggressive, and less likely to elicit conciliatory behavior from the offending partner (Tangney 1995). What accounts for this link between shame and anger? When feeling

shame, people initially direct hostility inward ('*I'm* such a bad person'). But this hostility may be redirected outward in a defensive attempt to protect the self by shifting the blame elsewhere ('Oh what a horrible person I am, and damn it, how could you make me feel that way!').

In sum, studies employing diverse samples, measures and methods converge. All things equal, it's better if your friend, partner, child, or boss feels guilt than shame. Shame motivates behaviors that interfere with interpersonal relationships. Guilt helps keep people constructively engaged in the relationship at hand.

1.2 Implications of Shame and Guilt for Psychological Adjustment

Although guilt appears to be the more 'moral' or adaptive emotion when considering social adjustment, is there a trade-off vis-à-vis individual psychological adjustment? Does the tendency to experience guilt or shame leave one vulnerable to psychological problems?

Researchers consistently report a relationship between proneness to shame and a whole host of psychological symptoms, including depression, anxiety, eating disorder symptoms, subclinical sociopathy, and low self-esteem (Harder et al. 1992, Tangney et al. 1995). This relationship is robust across measurement methods and diverse populations.

There is more controversy regarding the relationship of guilt to psychopathology. The traditional view is that guilt plays a significant role in psychological symptoms. Clinical theory and case studies make frequent reference to a maladaptive guilt characterized by chronic self-blame and obsessive rumination over one's transgressions. On the other hand, recent theory and research has emphasized the adaptive functions of guilt, particularly for interpersonal behavior. Tangney et al. (1995) argued that once one makes the critical distinction between shame and guilt, there's no compelling reason to expect guilt over specific behaviors to be associated with poor psychological adjustment.

The empirical research is similarly mixed. Studies employing adjective checklist-type (and other globally worded) measures find both shame-proneness and guilt-proneness correlated with psychological symptoms. On the other hand, measures sensitive to the self vs. behavior distinction show no relationship between proneness to 'shame-free' guilt and psychopathology.

1.3 Development of Guilt Distinct from Shame

The experience of self-conscious emotions requires the development of standards and a recognized self. In addition, a third ability is required to experience guilt (about specific behaviors) independent of shame (about the self)—the ability to make a clear distinction between self and behavior. Developmental research indicates that children do not begin making meaningful distinctions between attributions to ability (enduring characteristics) vs. attributions to effort (more unstable, volitional factors) until about age eight—the same age at which researchers find interpretable differences in children's descriptions of shame and guilt experiences (Ferguson et al. 1991).

2. Embarrassment

Miller (1996) defines embarrassment as 'an aversive state of mortification, abashment, and chagrin that follows public social predicaments' (p. 322). Indeed, embarrassment appears to be the most 'social' of the self-conscious emotions, occurring almost without exception in the company of others.

2.1 Causes of Embarrassment

In Miller's (1996) catalog of embarrassing events described by several hundred adolescents and adults, 'normative public deficiencies' (tripping in front of a large class, forgetting someone's name, unintended bodily induced noises) were at the top of the list. But there were many other types of embarrassment situations—awkward social interactions, conspicuousness in the absence of any deficiency, 'team transgressions' (embarrassed by a member of one's group), and 'empathic' embarrassment.

The diversity of situations that lead to embarrassment has posed a challenge to efforts at constructing a comprehensive 'account' of embarrassment. Some theorists believe that the crux of embarrassment is negative evaluation by others (Edelmann 1981, Miller 1996). This social evaluation account runs into difficulty with embarrassment events that involve no apparent deficiency (e.g., being the center of attention during a 'Happy Birthday' chorus). Other theorists subscribe to the 'dramaturgic' account, positing that embarrassment occurs when implicit social roles and scripts are disrupted. A flubbed performance, an unanticipated belch, and being the focus of 'Happy Birthday' each represent a deviation from accustomed social scripts.

Lewis (1992) distinguished between two types of embarrassment—embarrassment due to exposure and embarrassment due to negative self-evaluation. According to Lewis (1992), embarrassment due to exposure emerges early in life, once children develop a rudimentary sense of self. When children develop standards, rules and goals (SRGs), a second type of embarrassment emerges—'embarrassment as mild shame' associated with failure in relation to SRGs.

2.2 Functions of Embarrassment

Although there is debate about the fundamental causes of embarrassment, there is general agreement about its

adaptive significance. Gilbert (1997) suggests that embarrassment serves an important social function by signaling appeasement to others. When untoward behavior threatens a person's standing in an important social group, visible signs of embarrassment function as a nonverbal acknowledgment of shared social standards, thus diffusing negative evaluations and the likelihood of retaliation. Evidence from studies of both humans and nonhuman primates supports this remedial function of embarrassment (Keltner and Buswell 1997).

2.3 Embarrassment and Shame

Is there a difference between shame and embarrassment? Some theorists essentially equate the two emotions. A more dominant view is that shame and embarrassment differ in intensity of affect and/or severity of transgression. Still others propose that shame is tied to perceived deficiencies of one's core self, whereas embarrassment results from deficiencies in one's presented self.

Recent research suggests that shame and embarrassment are indeed quite different emotions—more distinct, even, than shame and guilt. For example, comparing adults' personal shame, guilt, and embarrassment experiences, Tangney et al. (1996) found that shame was a more intense, painful emotion that involved a greater sense of moral transgression. But controlling for intensity and morality, shame and embarrassment still differed markedly along many affective, cognitive, and motivational dimensions. When shamed, people felt greater responsibility, regret, and self-directed anger. Embarrassment was marked by more humor, blushing, and a greater sense of exposure.

2.4 Individual Differences in Embarrassability

As with shame and guilt, people vary in their propensity to experience embarrassment. These individual differences are evident within the first years of life, and are relatively stable across time. Research has shown that embarrassability is associated with neuroticism, high levels of negative affect, self-consciousness, and a fear of negative evaluation from others. Miller (1996) has shown that this fear of negative evaluation is not due to poor social skills, but rather a heightened concern for social rules and standards.

3. Pride

Of the self-conscious emotions, pride has received the least attention. Most research comes from developmental psychology, particularly in the achievement domain.

3.1 Developmental Issues

There appear to be substantial developmental shifts in the types of situations that induce pride, the nature of the pride experience itself, and the ways in which pride is expressed. For example, Stipek et al. (1992) observed developmental changes in the criteria children use for evaluating success and failure—and in the types of situation that lead to pride. Children under 33 months respond positively to task-intrinsic criteria (e.g., completing a tower of blocks), but they do not seem to grasp the concept of competition (e.g., winning or losing a race to complete a tower of blocks). It is only after 33 months that children show enhanced pride in response to a competitive win.

There are also developmental shifts in the importance of praise from others. Stipek et al. (1992) reported that all children 13–39 months smiled and exhibited pleasure with their successes. But there were age differences in social referencing. As children neared two years of age, they began to seek eye contact with parents upon completing a task, often actively soliciting parental recognition, which in turn enhanced children's pleasure with achievements. Stipek et al. (1992) suggests that the importance of external praise may be curvilinear across the lifespan. Very young children take pleasure in simply having immediate effect on their environment; as they develop self-consciousness (at about two), others' reactions shape their emotional response to success and failure. Still later, as standards become increasingly internalized, pride becomes again more autonomous, less contingent on others' praise and approval.

3.2 Two Types of Pride?

Both Tangney (1990) and Lewis (1992) have suggested that there are two types of pride. Paralleling the self vs. behavior distinction of guilt and shame, Tangney (1990) distinguished between pride in self ('alpha' pride) and pride in behavior ('beta' pride). Similarly, Lewis (1992) distinguished between pride (arising from attributing one's success to a specific action) and hubris (pridefulness arising from attributions of success to the global self). Lewis (1992) views hubris as largely maladaptive, noting that hubristic individuals are inclined to distort and invent situations to enhance the self, which can lead to interpersonal problems.

4. Future Research

Future research will no doubt focus on biological and social factors that shape individual differences in self-conscious emotions. In addition, we need to know more about the conditions under which guilt, shame, pride, and embarrassment are most likely to be adaptive vs. maladaptive. Finally, more cross-cultural research is needed. Kitayama et al. (1995) make the

compelling argument that, owing to cultural differences in the construction of the self, self-conscious emotions may be especially culturally sensitive.

See also: Culture and Emotion; Emotion and Expression; Emotion, Neural Basis of; Emotions, Evolution of; Emotions, Psychological Structure of; Shame and the Social Bond

Bibliography

Baumeister R F, Stillwell A M, Heatherton T F 1994 Guilt: An interpersonal approach. *Psychological Bulletin* **115**: 243–67

Edelmann R J 1981 Embarrassment: The state of research. *Current Psychological Reviews* **1**: 125–38

Ferguson T J, Stegge H, Damhuis I 1991 Children's understanding of guilt and shame. *Child Development* **62**: 827–39

Gilbert P 1997 The evolution of social attractiveness and its role in shame, humiliation, guilt, and therapy. *British Journal of Medical Psychology* **70**: 113–47

Harder D W, Cutler L, Rockart L 1992 Assessment of shame and guilt and their relationship to psychopathology. *Journal of Personality Assessment* **59**: 584–604

Keltner D, Buswell B N 1997 Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin* **122**: 250–70

Kitayama S, Markus H R, Matsumoto H 1995 Culture, self, and emotion: A cultural perspective on 'self-conscious' emotion. In: Tangney J P, Fischer K W (eds.) *Self-conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford Press, New York, pp. 439–64

Lewis H B 1971 *Shame and Guilt in Neurosis*. International Universities Press, New York

Lewis M 1992 *Shame: The Exposed Self*. Free Press, New York

Lewis M, Sullivan M W, Stanger C, Weiss M 1989 Self-development and self-conscious emotions. *Child Development* **60**: 146–56

Mascolo M F, Fischer K W 1995 Developmental transformation in appraisals for pride, shame, and guilt. In: Tangney J P, Fischer K W (eds.) *Self-conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford, New York, pp. 64–113

Miller R S 1996 *Embarrassment: Poise and Peril in Everyday Life*. Guilford Press, New York

Stipek D J, Recchia S McClintic S 1992 Self-evaluation in young children. Monographs of the Society for Research in Child Development 57 (1, Serial No. 226) R5–R83

Tangney J P 1990 Assessing individual differences in proneness to shame and guilt: Development of the self-conscious affect and attribution inventory. *Journal of Personality and Social Psychology* **59**: 102–11

Tangney J P 1995 Shame and guilt in interpersonal relationships. In: Tangney J P, Fischer K W (eds.) *Self-conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford Press, New York, pp. 114–39

Tangney J P, Burggraf S A, Wagner P E 1995 Shame-proneness, guilt-proneness, and psychological symptoms. In: Tangney J P, Fischer K W (eds.) *Self-conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford, New York, pp. 343–67

Tangney J P, Miller R S, Flicker L, Barlow D H 1996 Are shame, guilt and embarrassment distinct emotions? *Journal of Personality and Social Psychology* **70**: 1256–69

Tangney J P, Wagner P, Gramzow R 1992 Proneness to shame, proneness to guilt, and psychopathology. *Journal of Abnormal Psychology* **101**: 469–78

J. P. Tangney

Self-development in Childhood

In addition to domain-specific self-concepts, the ability to evaluate one's overall worth as a person emerges in middle childhood. The level of such global self-esteem varies tremendously across children and is determined by how adequate they feel in domains of importance as well as the extent to which significant others (e.g., parents and peers) approve of them as a person. Efforts to promote positive self-esteem are critical, given that low self-esteem is associated with many psychological liabilities including depressed affect, lack of energy, and hopelessness about the future.

1. Introduction

Beginning in the second year of life toddlers begin to talk about themselves. With development, they come to understand that they possess various characteristics, some of which may be positive ('I'm smart') and some of which may be negative ('I'm unpopular'). Of particular interest is how the very nature of such self-evaluations changes with development as well as among individual children and adolescents across two basic evaluative categories, (a) domain-specific self-concepts, i.e., how one judges one's attributes in particular arenas, e.g., scholastic competence, social (for a complete treatment of self-development in childhood and adolescence, see Harter 1999).

Developmental shifts in the nature of self-evaluations are driven by changes in the child's cognitive capabilities (see *Self-knowledge*: *Philosophical Aspects* and *Self-evaluative Process*, *Psychology of*). Cognitive-developmental theory and findings (see Piaget 1960, 1963, Fischer 1980) alert us to the fact that the young child is limited to very specific, concrete representations of self and others, for example, 'I know my ABCs' (see Harter 1999). In middle to later childhood, the ability to form higher-order concepts about one's attributes and abilities (e.g., 'I'm smart') emerges. There are further cognitive advances at adolescence, allowing the teenager to form abstract concepts about

the self that transcend concrete behavioral manifestations and higher-order generalizations (e.g., 'I'm intelligent').

2. Developmental Differences in Domain-specific Self-concepts

Domain-specific evaluative judgments are observed at every developmental level. However, the precise nature of these judgments varies with age (see Table 1). In Table 1, five common domains in which children and adolescents make evaluate judgments about the self are identified: Scholastic competence, Physical competence, Social competence, Behavioral conduct, and Physical appearance. The types of statements vary, however, across three age periods, early childhood, later childhood, and adolescence, in keeping with the cognitive abilities of each age period.

2.1 Early Childhood

Young children provide very concrete accounts of their capabilities, evaluating specific behaviors. Thus, they communicate how they know their ABCs, how they can run very fast, how they are nice to a particular friend, how they don't hit their sister, and how they possess a specific physical feature such as pretty blond hair. Of particular interest in such accounts is the fact that the young child typically provides a litany of virtues, touting his or her positive skills and attributes. One cognitive limitation of this age period is that the young child cannot distinguish the wish to be competent from reality. As a result, they typically overestimate their abilities because they do not yet have the skills to evaluate themselves realistically. Another cognitive characteristic that contributes to potential distortions is the pervasiveness of all-or-none thinking. That is, evaluations are either all positive or all negative. With regard to self-evaluations, they are typically all positive. (Exceptions to this positivity bias can be observed in children who are chronically abused, since severe maltreatment is often accompanied by parental messages that make the child feel inadequate, incompetent, and unlovable. Such children will also engage in all-or-none thinking but conclude that they are all bad.)

2.2 Middle to Later Childhood

As the child grows older, the ability to make higher-order generalizations in evaluating his or her abilities and attributes emerges. Thus, rather than cite prowess at a particular activity, the child may observe that he or she is good at sports, in general. This inference can further be justified in that the child can describe his or her talent at several sports (e.g., good at soccer, basketball, baseball). Thus, the higher-order generalization represents a cognitive construction in which an over-arching judgment (good at sports) is defined in terms of specific examples which warrant this conclusion. Similar processes allow the older child to conclude that he or she is smart (e.g., does well in math, science, and history). The structure of a higher-order generalization about being well behaved could include such components as obeying parents, not getting in trouble, and trying to do what is right. A generalization concerning the ability to make friends may subsume accounts of having friends at school, making friends easily at camp, and developing friendships readily upon moving to a new neighborhood. The perception that one is good-looking may be based on one's positive evaluation of one's face, hair, and body.

During middle childhood, all-or-none thinking diminishes and the aura of positivity fades. Thus, children do not typically think that they are all virtuous in every domain. The more common pattern is for them to feel more adequate in some domains than others. For example, one child may feel that he or she is good at schoolwork and is well behaved, whereas he or she is not that good at sports, does not think that he or she is good-looking, and reports that it is hard to make friends. Another child may report the opposite pattern.

There are numerous combinations of positive and negative evaluations across these domains that chil-

Table 1
Developmental changes in the nature of self-evaluative statements across different domains

Domains	Early childhood specific behaviors	Later childhood generalizations[a]	Adolescence abstractions[a]
Scholastic competence	I know my A,B,C's	I'm smart in school	I'm intelligent
Athletic competence	I can run very fast	I'm good at sports	I'm athletically talented
Social competence	I'm nice to my friend, Jason	It's easy for me to make friends	I'm popular
Behavioral conduct	I don't hit my sister	I'm well behaved	I think of myself as a moral person
Physical appearance	I have pretty blond hair	I'm good looking	I'm physically attractive

[a]Examples in the table represent positive self-evaluations. However, during later childhood and adolescence, negative judgments are also observed

dren can and do report. Moreover, they may report both positive and negative judgments within a given domain, for example, they are smart in some school subjects (math and science) but 'dumb' in others (english and social studies). Such evaluations may also be accompanied by self-affects that also emerge in later childhood, for example, feeling proud of one's accomplishments but ashamed of one's perceived failures (see also *Self-conscious Emotions*, *Psychology of*). This ability to consider both positive and negative characteristics is a major cognitive–developmental acquisition. Thus, beginning in middle to later childhood, these distinctions result in a profile of self-evaluations across domains.

Contributing to this advance is the ability to engage in social comparison. Beginning in middle childhood one can use comparisons with others as a barometer of the skills and attributes of the self. In contrast, the young child cannot simultaneously compare his or her attributes to the characteristics of another in order to detect similarities or differences that have implications for the self. Although the ability to utilize social comparison information for the purpose of self-evaluation represents a cognitive-developmental advance, it also ushers in new, potential liabilities. With the emergence of the ability to rank-order the performance of other children, all but the most capable children will necessarily fall short of excellence. Thus, the very ability and penchant to compare the self with others makes one's self-concept vulnerable, particularly if one does not measure up in domains that are highly valued. The more general effects of social comparison can be observed in findings revealing that domain-specific self-concepts become more negative during middle and later childhood, compared to early childhood.

2.3 *Adolescence*

For the adolescent, there are further cognitive-developmental advances that alter the nature of domain-specific self-evaluations. As noted earlier, adolescence brings with it the ability to create more abstract judgments about one's attributes and abilities. Thus, one no longer merely considers oneself to be good at sports but to be athletically talented. One is no longer merely smart but views the self more generally as intelligent, where successful academic performance, general problem-solving ability, and creativity might all be subsumed under the abstraction of intelligence. Abstractions may be similarly constructed in the other domains. For example, in the domain of behavioral conduct, there will be a shift from the perception that one is well behaved to a sense that one is a moral or principled person. In the domains of social competence and appearance, abstractions may take the form of perceptions that one is popular and physically attractive.

These illustrative examples all represent positive self-evaluations. However, during adolescence (as well as in later childhood), judgments about one's attributes will also involve negative self-evaluations. Thus, certain individuals may judge the self to be unattractive, unpopular, unprincipled, etc. Of particular interest is the fact that when abstractions emerge, the adolescent typically does not have total control over these new acquisitions, just as when one is acquiring a new athletic skill (e.g., swinging a bat, maneuvering skis), one lacks a certain level of control. In the cognitive realm, such lack of control often leads to overgeneralizations that can shift dramatically across situations or time. For example, the adolescent may conclude at one point in time that he or she is exceedingly popular but then, in the face of a minor social rebuff, may conclude that he or she is extremely unpopular. Gradually, adolescents gain control over these self-relevant abstractions such that they become capable of more balanced and accurate self-representations (see Harter 1999).

3. *Global Self-esteem*

The ability to evaluate one's worth as a person also undergoes developmental change. The young child simply is incapable, cognitively, of developing the verbal concept of his/her value as a person. This ability emerges at the approximate age of eight. However, young children exude a sense of value or worth in their behavior. The primary behavioral manifestations involve displays of confidence, independence, mastery attempts, and exploration (see Harter 1999). Thus, behaviors that communicate to others that children are sure of themselves are manifestations of high self-esteem in early childhood.

At about the third grade, children begin to develop the concept that they like, or don't like, the kind of person they are (Harter 1999, Rosenberg 1979). Thus, they can respond to general items asking them to rate the extent to which they are pleased with themselves, like who they are, and think they are fine, as a person. Here, the shift reflects the emergence of an ability to construct a higher-order generalization about the self. This type of concept can be built upon perceptions that one has a number of specific qualities; for example, that one is competent, well behaved, attractive, etc. (namely, the type of domain-specific self-evaluations identified in Table 1). It can also be built upon the observation that significant others, for example, parents, peers, teachers, think highly of the self. This process is greatly influenced by advances in the child's ability to take the perspective of significant others (Selman 1980). During adolescence, one's evaluation of one's global worth as a person may be further elaborated, drawing upon more domains and sources of approval, and will also become more abstract. Thus, adolescents can directly acknowledge

that they have high or low self-esteem, as a general abstraction about the self (see also *Self-esteem in Adulthood*).

4. Individual Differences in Domain-specific Self-concepts as well as Global Self-esteem

Although there are predictable cognitively based developmental changes in the nature of how most children and adolescents describe and evaluate themselves, there are striking individual differences in how positively or negatively the self is evaluated. Moreover, one observes different profiles of children's perceptions of their competence or adequacy across the various self-concept domains, in that children evaluate themselves differently across domains. Consider the profiles of four different children. One child, Child A, may feel very good about her scholastic performance, although this is in sharp contrast to her opinion of her athletic ability, where she evaluates herself quite poorly. Socially she feels reasonably well accepted by her peers. In addition, she considers herself to be well behaved. Her feelings about her appearance, however, are relatively negative. Child A also reports very high self-esteem. Another child, Child B, has a very different configuration of scores. This is a boy who feels very incompetent when it comes to schoolwork. However, he considers himself to be very competent, athletically, and feels well received by peers. He judges his behavioral conduct to be less commendable. In contrast, he thinks he is relatively good-looking. Like Child A, he also reports high self-esteem.

Other profiles are exemplified by Child C and Child D, neither of whom feel good about themselves scholastically or athletically. They evaluate themselves much more positively in the domains of social acceptance, conduct, and physical appearance. In fact, their profiles are quite similar to each other across the five specific domains. However, judgments of their self-esteem are extremely different. Child C has very high self-esteem whereas Child D has very low self-esteem. This raises a puzzling question: how can two children look so similar with regard to their domain-specific self-concepts but evaluate their global self-esteem so differently? We turn to this issue next, in examining the causes of global self-esteem.

5. The Causes of Children's Level of Self-esteem

Our understanding of the antecedents of global self-esteem have been greatly aided by the formulations of two historical scholars of the self, William James (1892) and Charles Horton Cooley (1902). Each suggested rather different pathways to self-esteem, defined as an overall evaluation of one's worth as a person (see reviews by Harter 1999, Rosenberg 1979). James focused on how the individual assessed his or her competence in domains where one had aspirations to succeed. Cooley focused on the salience of the opinions that others held about the self, opinions which one incorporated into one's global sense of self (see *Self: History of the Concept*).

5.1 Competence–Adequacy in Domains of Importance

For James, global self-esteem derived from the evaluations of one's sense of competence or adequacy in the various domains of one's life relative to how important it was to be successful in these domains. Thus, if one feels one is successful in domains deemed important, high self-esteem will result. Conversely, if one falls short of one's goal in domains where one has aspirations to be successful, one will experience low self-esteem. One does not, therefore, have to be a superstar in every domain to have high self-esteem. Rather, one only needs to feel adequate or competent in those areas judged to be important. Thus, a child may evaluate himself or herself as unathletic; however, if athletic prowess is not an aspiration, then self-esteem will not be negatively affected. That is, the high self-esteem individual can discount the importance of areas in which one does not feel successful.

This analysis can be applied to the profiles of Child C and Child D. In fact, we have directly examined this explanation in research studies by asking children to rate how important it is for them to be successful (Harter 1999). The findings reveal that high self-esteem individuals feel competent in domains they rate as important. Low self-esteem individuals report that areas in which they are unsuccessful are still very important to them. Thus, Child C represents an example of an individual who feels that social acceptance, conduct, and appearance, domains in which she evaluates herself positively, are very important but that the two domains where she is less successful, scholastic competence and athletic competence are not that important. In contrast, Child D rates all domains as important, including the two domains where he is not successful; scholastic competence and athletic competence. Thus, the discrepancy between high importance coupled with perceptions of inadequacy contribute to low self-esteem.

5.2 Incorporation of the Opinions of Significant Others

Another important factor influencing self-esteem can be derived from the writings of Cooley (1902) who metaphorically made reference to the 'looking-glass self' (see Oosterwegel and Oppenheimer 1993). According to this formulation, significant others (e.g., parents and peers) were social mirrors into which one gazed in order to determine what they thought of the

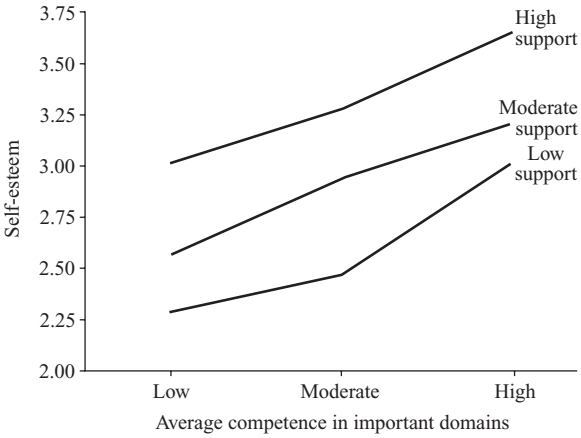

Figure 1
Findings on how competence in domains of importance and social support combine to predict global self-esteem

self. Thus, in evaluating the self, one would adopt what one felt were the judgments of these others whose opinions were considered important. Thus, the approval, support, or positive regard from significant others became a critical source of one's own sense of worth as a person. For example, children who receive approval from parents and peers will report much higher self-esteem than children who experience disapproval from parents and peers.

Findings reveal that both of these factors, competence in domains of importance and the perceived support of significant others, *combine* to influence a child's or adolescent's self-esteem. Thus, as can be observed in Fig. 1, those who feel competent in domains of importance and who also report high support, rate themselves as having the highest self-esteem. Those who feel inadequate in domains deemed important and who also report low levels of support, rate themselves as having the lowest self-esteem. Other combinations fall in between (data from Harter 1993).

6. Conclusions

Two types of self-representations that can be observed in children and adolescents were distinguished, evaluative judgments of competence or adequacy in specific domains and the global evaluation of one's worth as a person, namely overall self-esteem. Each of these undergoes developmental change based on age-related cognitive advances. In addition, older children and adolescents vary tremendously with regard to whether self-evaluations are positive or negative. Within a given individual, there will be a profile of self-evaluations, some of which are more positive and some more negative. More positive self-concepts in domains considered important, as well as approval from significant others, will lead to high self-esteem. Conversely, negative self-concepts in domains considered important, coupled with lack of approval from significant others, will result in low self-esteem. Self-esteem is particularly important since it is associated with very important outcomes or consequences. Perhaps the most well-documented consequence of low self-esteem is depression. Children and adolescents (as well as adults) with the constellation of causes leading to low self-esteem will invariably report that they feel depressed, emotionally, and are hopeless about their futures; the most seriously depressed consider suicide. Thus, it is critical that we intervene for those experiencing low self-esteem. Our model of the causes of self-esteem suggests strategies that may be fruitful, for example, improving skills, helping individuals discount the importance of domains in which it is unlikely that they can improve, and providing support in the form of approval for who they are as people. Future research, however, is necessary to determine the different pathways to low and high self-esteem. For example, for one child, the sense of inadequacy in particular domains may be the pathway to low self-esteem. For another child, lack of support from parents or peers may represent the primary cause. Future efforts should be directed to the identification of these different pathways since they have critical implications for intervention efforts to enhance feelings of worth for those children with low self-esteem (see also *Self-concepts*: *Educational Aspects*). Positive self-esteem is clearly a psychological commodity, a resource that is important for us to foster in our children and adolescents if we want them to lead productive and happy lives.

See also: Identity in Childhood and Adolescence; Personality Development in Childhood; Self-concepts: Educational Aspects; Self: History of the Concept; Self-knowledge: Philosophical Aspects; Self-regulation in Childhood

Bibliography

Cooley C H 1902 *Human Nature and the Social Order*. Charles Scribner's Sons, New York

Damon W, Hart D 1988 *Self-understanding in Childhood and Adolescence*. Cambridge University Press, New York

Fischer K W 1980 A theory of cognitive development: the control and construction of hierarchies of skills. *Psychological Review* **87**: 477–531

Harter S 1993 Causes and consequences of low self-esteem in children and adolescents. In: Baumeister R F (ed.) *Self-esteem: The Puzzle of Low Self-regard*. Plenum, New York

Harter S 1999 *The Construction of the Self: A Developmental Perspective*. Guilford Press, New York

James W 1892 *Psychology: The Briefer Course*. Henry Holt, New York

Oosterwegel A, Oppenheimer L 1993 *The Self-system: Developmental Changes Between and Within Self-concepts*. Erlbaum, Hillsdale, NJ

Piaget J 1960 *The Psychology of Intelligence*. Littlefield, Adams, Patterson, NJ
Piaget J 1963 *The Origins of Intelligence in Children*. Norton, New York
Rosenberg M 1979 *Conceiving the Self*. Basic Books, New York
Selman R L 1980 *The Growth of Interpersonal Understanding*. Academic Press, New York

S. Harter

Self-efficacy

Self-efficacy refers to the individual's capacity to produce important effects. People who are aware of being able to make a difference feel good and therefore take initiatives; people who perceive themselves as helpless are unhappy and are not motivated for actions. This article treats the main concepts related to self-efficacy, their theoretical and historical contexts, their functions and practical uses, as well as developmental and educational/therapeutic aspects.

1. Concepts

Everything that happens is caused to happen (Aristotle). Making changes means being a cause or providing a cause that produces a change. As the true causes are difficult to identify, the terms *conditions* and *contingencies* are often used instead. An effect is contingent upon a condition or a set of conditions if it always occurs if the condition or the set of conditions is met. Such conditions are sufficient but not necessary for producing the effect.

Here we are interested in human actions as necessary conditions of change (see *Motivation and Actions, Psychology of*). Actions, too, depend on conditions. Considering person-related conditions of effective actions, we can differentiate aspects like knowledge, initiative, perseverance, intelligence, experience, physical force, help from others, and more. Thus, instead of saying that an actor is able to produce a certain effect, we can say more elaborately that an actor is endowed with certain means or conditions that enable him or her to attain certain goals (Fig. 1).

We say that individuals (or groups) are in *control* of a specific goal if they are able to produce the corresponding changes (horizontal line of Fig. 1). More elaborately, they are in control of a specific goal if they are aware of the necessary *contingencies* and if they are *competent* enough to make these contingencies work (both diagonal lines in Fig. 1). Control is complete if these contingencies are necessary and sufficient; control is partial if the contingencies are necessary but not sufficient. Instead of control, Bandura (1977) introduced the word *efficacy*, more specifically *self-efficacy*. We use control and self-efficacy interchangeably.

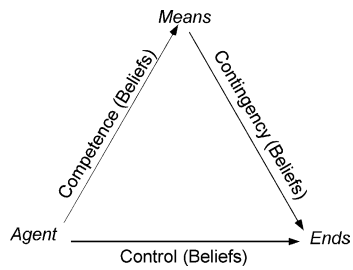

Figure 1
Means–ends relations and agency as components of control (adapted from Skinner et al. 1988 and Flammer 1990)

Controlling means putting control into action. This is not equivalent to having control or being in control: People are in control of certain states of affairs, if they can put control into action, even if they do not. For example, somebody has control over buying a new bicycle, even if he or she does not buy it.

People can have control over certain events without knowing it; they will then probably miss possible chances to activate such control. On the other hand, people may believe that they have some control, but in fact do not. That may make them feel good as long as there is no need to put this control into work. Obviously, it is important that people do not only have control, but that they also know that they have control. Not being in control of an important situation is equivalent to being helpless in this respect (Seligman 1975). It has been proved that the psychological effects of helplessness (HL) are different depending on whether the helpless person believes themself to be helpless for ever (chronic HL), whether being helpless is unique (personal vs. universal HL), and whether helplessness is related to a specific domain or to most domains of life (specific vs. global HL). In the worst case, helpless people are (a) deeply sad about not having control, (b) demotivated to take initiatives or to invest effort and perseverance, (c) cognitively blind for any alternative or better view of the state of the world, and (d) devaluate themselves.

Obviously, at least in subjectively important domains, we prefer self-efficacy to helplessness: self-efficacy beliefs provide us with security and pride. When we lack self-efficacy in important domains we either strive for self-efficacy (by fighting, learning, or training) or search for compensation. A common type of compensation consists of seeking help or delegating personal control (= *indirect control* or *proxy control*), e.g., to pay a gardener caring for one's garden or to put a doctor in charge with one's health or to pray to God for a favor in a seemingly hopeless situation. Another way of compensating lacking (primary) control is to use *secondary control* (Rothbaum et al. 1982). While control (i.e., primary control) consists of making the

world fit with one's goals and aspirations, secondary control accommodates personal aspirations or personal interpretations of the actual state in order to make them fit with the world (see *Coping across the Lifespan*).

2. *History of the Concept of Self-efficacy*

In the 1950s, Rotter (1954) suggested the concept *locus of control*, meaning the place where control of desired reinforcement for behavior is exerted. Internal control means control within the person, external control means control outside the person, possibly in powerful others, in objective external conditions, or in chance or luck. Rotter and his associates have developed valid measuring instruments that have been used in thousands (!) of studies demonstrating that internal locus of control is positively correlated with almost all desirable attributes of humans.

Fritz Heider (1944) who studied the subjective attributions for observed actions had already suggested the concepts of internality and externality. The true origin, i.e., the 'cause' of an observed action, is either attributed to the person (internal, personal liability) or to person-independent conditions (external, no personal liability). Heider's work has triggered a large research tradition on causal attributions. Results of this research were of use for successful differentiations of subjective interpretations concerning experiences of helplessness (see above; see also *Attributional Processes: Psychological*). Consequently, attribution theory remains an important element of self-efficacy theory.

Modern self-efficacy theory goes beyond Rotter's theory insofar as it is more differentiated (e.g., contingency vs. competence, primary vs. secondary), distinctively referred to specific domains of actions (e.g., health, school), and elaborated to also include aspects other than personality (e.g., motivation, development).

3. *Self-efficacy as an Important Element of a Happy and Successful Person*

Individuals with high self-efficacy beliefs also report strong feelings of well-being and high self-esteem in general (Bandura 1997, Flammer 1990). They are willing to take initiative in related domains, to apply effort if needed, and persevere in efforts as long as they believe in their efficacy. Potentially stressful situations produce less subjective stress in highly self-efficient individuals. However, while self-efficacy acts as a buffer against stress, it can also—indirectly—produce stress insofar as it can induce overly ambitious individuals to assume more responsibilities than they are able to cope with in sheer quantity.

Moreover, self-efficacy has been reported to exert a positive influence on recovery from surgery or illness and on healthy lifestyles. It is not surprising that high self-efficacy beliefs enhance school success; likewise, school failure inhibits relative self-efficacy beliefs, again partly depending on the individual's attributional patterns. Interestingly, it has been demonstrated repeatedly and in several cultures that in most domains healthy and happy individuals tend to slightly overestimate themselves. Realistic estimation of self-efficacy is rather typical for persons vulnerable to depressed mood, and clear underestimation increases the chance for a clinical (reactive) depression. On the other hand, major overestimation might result in painful and harmful clashes with reality.

4. *The Development of Self-efficacy Beliefs*

Evidently, the newborn baby does not have self-efficacy beliefs in our sense. The basic structure of the self-efficacy beliefs develops within the first three or four years. According to Flammer's (1995) analysis, the infant's development towards the basic understanding of self-efficacy proceeds through a developmental sequence consisting of the acquisition of (a) the basic event schema (i.e., that classes of events happen), (b) the elementary causal schema (conditions, i.e., actions, events), (c) the understanding of personally producing effects, (d) the understanding of success and failure in aiming at nontrivial goals (visible as pride and as shame, respectively), and (e) the discovery of being not only the origin of one certain change but also capable of producing such changes.

Obviously, this development proceeds in domains that are accessible by the infant so far. Later on, this development will have to be extended to further domains. As to the domain of school success, within the second half of the first decade of life, the child learns more and more differentiations of means towards the same ends. Thus, he or she gradually abandons a *global concept* of simply being or not being able and singles out—probably in this sequence—the factor *effort* (more effort is needed to solve tasks—a typical lesson to be learned early in school), the factors individual ability and task difficulty (higher difficulty requiring more ability), and finally the understanding of the compensatory relation between effort and ability (it is possible to reach the same goals by being less capable but more hard-working).

In adolescence and early adulthood more lessons have to be learned. More and more domains become accessible to personal control due to increased cognitive, physical, or economic strength and social power. This is exciting, indeed. However, individuals have permanently to select from the choices which are offered to them (Flammer 1996). Trying to control everything results in overburdening. One thing is to deselect control domains because they compete with higher priority control domains; another thing is to be forced to renounce control because no accessible

contingencies seem to exist. As long as there are enough attractive alternatives available, it is not painful, but it can severely hurt handicapped individuals and old people when they lose control of important domains. Old people are well advised both not to resign too early and to search for compensations. Such compensations consists of artifacts of all kinds (from memory aids to hearing aids), but they also include the above mentioned compensations like indirect control (social resources) and secondary control. Indeed, it seems that the extent and the importance of secondary control increases with the lifetime (Heckhausen and Schulz 1995). Under certain conditions, Baltes and Silverberg (1994) have even suggested that people in old people's homes adjust better if in certain domains they give up personal control at all. Alluding to the concept of learned helplessness, they called such behavior *learned dependency*. Learned dependency helps to avoid certain social conflicts; the only remaining personal control may be the control of giving in.

5. *Educational and Therapeutic Aspects*

According to the development of the basic structure of self-efficacy, contingent behavior by the caregivers is crucial already within the first weeks of life. Caregivers' behavior should be predictable, i.e., contingent at least upon the baby's actual behavior, and as far as ever possible upon the baby's perceptions, feelings, and intentions. This requires an enormous potential of sensitivity towards the child. Fortunately, researchers have demonstrated that parental empathy is partly a natural gift with the majority of attentive parents. Studies have shown that contingent behavior fosters children's happiness, but also their willingness to learn and their curiosity. If caregivers are judged as not reacting contingently enough, we also have to consider that some babies show quite unorganized behavior and make the caregiver's task very difficult ('difficult babies'). In such cases it is difficult to decide whether the noncontingency has originated from the caregiver or from the baby.

The subsequent steps in the development of self-efficacy require that caregivers provide freedom for experimentation, let the child try by himself or herself, and comment on the successes and failures in a way that the child can establish and maintain confidence in his or her efficacy (Schneewind 1995). Nevertheless, caregivers should try to prevent the child from dangerous and frequent hopeless experiences. Psychotherapy with individuals who have severely undermined self-efficacy beliefs is difficult. Teaching and trying to convince them that they are really capable even when they believe not to be does not help much. Helping them to recall from memory prior success experiences instead of being impressed by failures only is more efficient. Even more efficient are new and successful experiences. Helpless individuals not only interpret failures to their disadvantage, they also play down their contribution to eventual success. Parallel to these findings, memory research has demonstrated that depressed people's memories of own actions are biased towards recalling more of their failures than their own success.

This leaves us with an important contrast: while children and healthy adults tend to overestimate their self-efficacy, individuals who have lost confidence in themselves immunize such devastating beliefs by not trying anymore, by self-damaging attributions, and by recalling their biography in a way that is consistent with their beliefs. Given the pervasive influence of positive beliefs in self-efficacy, it is important to help individuals with establishing and maintaining self-efficacy beliefs at a high level, and to guide failure-expecting persons to positive experiences.

6. *Conclusion*

Within the last decades, theory and research have established self-efficacy beliefs as important elements in the understanding of human action and human well-being in a very large sense. However, little is known so far about differences in self-efficacy beliefs in different life domains and among different cultures. Further research should include more systematic comparisons between cultures, between life domains, and—if possible—between historical times. In addition, it is suggested that in the future investigators consider more seriously the fact that all changes are due to a multitude of necessary conditions. More specifically, there is a need for researchers to consider the efficacy and efficacy-beliefs of interacting people, that is, to examine concepts such as shared control or 'common efficacy.'

See also: Control Behavior: Psychological Perspectives; Learned Helplessness; Motivation and Actions, Psychology of; Self-efficacy and Health; Self-efficacy: Educational Aspects; Self-regulation in Adulthood; Self-regulation in Childhood

Bibliography

Baltes M M, Silverberg S B 1994 The dynamics between dependency and autonomy. In: Featherman D L, Lerner R M, Parlmutter M (eds.) *Life-span Development and Behavior*. Erlbaum, New York

Bandura A 1977 Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* **84**: 191–215

Bandura A 1997 *The Exercise of Control*. Freeman, New York

Flammer A 1990 *Erfahrung der eigenen Wirksamkeit* [*Experiencing One's Own Efficacy*]. Huber, Bern, Switzerland

Flammer A 1995 Developmental analysis of control beliefs. In: Bandura A (ed.) *Self-efficacy in Changing Societies*. Cambridge University Press, New York

Flammer A 1996 *Entwicklungstheorien* [*Theories of Development*]. Huber, Bern, Switzerland

Heckhausen J, Schultz R 1995 The life-span theory of control. *Psychological Review* **102**: 284–304
Heider F 1944 Social perception and phenomenal causality. *Psychological Review* **51**: 358–74
Rothbaum F, Weisz J R, Snyder S S 1982 Changing the world and changing the self: a two-process model of perceived control. *Journal of Personality and Social Psychology* **42**: 5–37
Rotter J B 1954 *Social Learning and Clinical Psychology*. Prentice-Hall, Englewood Cliffs, NJ
Schneewind K A 1995 Impact of family processes on control beliefs. In: Bandura A (ed.) *Self-efficacy in Changing Societies*. Cambridge University Press, New York
Seligman M E P 1975 *Helplessness. On Depression, Development and Death*. Freeman, San Francisco
Skinner E A, Chapman M, Baltes P B 1988 Control, means–ends, and agency beliefs: A new conceptualization and its measurement during childhood. *Journal of Personality and Social Psychology* **54**: 117–33

A. Flammer

Self-efficacy and Health

The quality of human health is heavily influenced by lifestyle habits. By exercising control over several health habits people can live longer healthier and slow the process of aging (see *Control Beliefs: Health Perspectives*). Exercise, reduce dietary fat, refrain from smoking, keep blood pressure down, and develop effective ways of coping with stressors. If the huge health benefits of these few lifestyle habits were put into a pill it would be declared a spectacular breakthrough in the field of medicine. The recent years have witnessed a major change in the conception of human health and illness from a disease model to a health model. It is just as meaningful to speak of levels of vitality as of degrees of impairment. The health model, therefore, focuses on health promotion as well as disease prevention. Perceived self-efficacy plays a key role in the self-management of habits that enhance health and those that impair it.

1. *Perceived Self-efficacy*

Perceived self-efficacy is concerned with people's beliefs in their capabilities to exercise control over their own functioning and over environmental events. Such beliefs influence what courses of action people choose to pursue, the goals they set for themselves and their commitment to them, how much effort they put forth in given endeavors, how long they persevere in the face of obstacles and failure experiences, their resilience to adversity, whether their thought patterns are self-hindering or self-aiding, how much stress and depression they experience in coping with taxing environmental demands, and the level of accomplishments they realize (Bandura 1997, Schwarzer 1992).

In social cognitive theory, perceived self-efficacy operates in concert with other determinants in regulating lifestyle habits. They include the positive and negative outcomes people expect their actions to produce. These outcome expectations may take the form of aversive and pleasurable physical effects, approving and disapproving social reactions, or self-evaluative consequences expressed as self-satisfaction and self-censure. Personal goals, rooted in a value system, provide further self-incentives and guides for health habits. The perceived sociostructural facilitators and impediments operate as another set of determinants of health habits.

Self-efficacy is a key determinant in the causal structure because it affects health behavior both directly, and by its influence on these other determinants. The stronger the perceived efficacy, the higher the goal challenges people set for themselves, the more they expect their efforts to produce desired outcomes, and the more they view obstacles and impediments to personal change as surmountable.

There are two major ways which a sense of personal efficacy affects human health. At the more basic level, such beliefs activate biological systems that mediate health and disease. The second level is concerned with the exercise of direct control over habits that affect health and the rate of biological aging.

2. *Impact of Efficacy Beliefs on Biological Systems*

Stress is an important contributor to many physical dysfunctions (O'Leary 1990). Perceived controllability appears to be the key organizing principle in explaining the biological effects of stress. Exposure to stressors with the ability to exercise some control over them has no adverse physical effects. But exposure to the same stressors without the ability to control them impairs immune function (Herbert and Cohen 1993b, Maier et al. 1985). Epidemiological and correlational studies indicate that lack of behavioral or perceived control over stressors increases susceptibility to bacterial and viral infections, contributes to the development of physical disorders and accelerates the rate of progression of disease (Schneiderman et al. 1992).

In social cognitive theory, stress reactions arise from perceived inefficacy to exercise control over aversive threats and taxing environmental demands (Bandura 1986). If people believe they can deal effectively with potential stressors, they are not perturbed by them. But, if they believe they cannot control aversive events, they distress themselves and impair their level of functioning. Perceived inefficacy to manage stressors activates autonomic, catecholamine and opioid systems that modulate the immune system in ways that

can increase susceptibility to illness (Bandura 1997, O'Leary 1990).

The immunosuppressive effects of stressors is not the whole story, however. People are repeatedly bombarded with taxing demands and stressors in their daily lives. If stressors only impaired immune function people would be highly vulnerable to infective agents that would leave them chronically bedridden with illnesses or quickly do them in. Most human stress is activated while competencies are being developed and expanded. Stress aroused while gaining a sense of mastery over aversive events strengthens components of the immune system (Wiedenfeld et al. 1990). The more rapid the growth of perceived coping efficacy, the greater the boost of the immune system. Immunoenhancement during development of coping capabilities vital to effective adaptation has evolutionary survival value.

The field of health functioning has been heavily preoccupied with the physiologically debilitating effects of stressors. Self-efficacy theory also acknowledges the physiologically strengthening effects of mastery over stressors. As Dienstbier (1989) has shown, a growing number of studies is providing empirical support for physiological toughening by successful coping.

Depression is another affective pathway through which perceived coping efficacy can affect health functioning. It has been shown to reduce immune function, and to heighten susceptibility to disease (Herbert and Cohen 1993a). The more severe the depression, the greater the reduction in immunity. Perceived self-efficacy to exercise control over things one values highly produces bouts of depression (Bandura 1997).

Social support reduces vulnerability to stress, depression, and physical illness. But social support is not a self-forming entity waiting around to buffer harried people against stressors. People have to go out and find, create, and maintain supportive relationships for themselves. This requires a robust sense of social efficacy. Perceived social inefficacy contributes to depression both directly, and by curtailing development of social supports (Holahan and Holahan 1987). Social support, in turn, enhances perceived self-efficacy. Mediational analyses show that social support alleviates depression and fosters health-promoting behavior only to the extent that it boosts personal efficacy.

3. *Self-efficacy in Promoting Healthful Lifestyles*

Lifestyle habits can enhance or impair health (see *Health Behaviors*). This enables people to exert some behavioral control over their vitality and quality of health. Efficacy beliefs affect every phase of personal change: whether people even consider changing their health habits; whether they enlist the motivation and perseverance needed to succeed should they choose to do so; and how well they maintain the habit changes they have achieved (Bandura 1997).

3.1 *Initiation of Change*

People's beliefs that they can motivate themselves and regulate their own behavior play a crucial role in whether they even consider changing detrimental health habits. They see little point in even trying if they believe they do not have what it takes to succeed. If they make an attempt, they give up easily in the absence of quick results or setbacks. Among those who change detrimental health habits on their own, the successful ones have stronger perceived self-efficacy at the outset than nonchangers and subsequent relapsers.

Efforts to get people to adopt healthful practices rely heavily on persuasive communications in health education campaigns. Health communications foster adoption of healthful practices mainly by raising beliefs in personal efficacy, rather than by transmitting information on how habits affect health, by arousing fear of disease, or by increasing perception of one's personal vulnerability or risk (Meyerowitz and Chaiken 1987). To help people reduce health-impairing habits requires a change in emphasis, from trying to scare people into health, to enable them with the skills and self-beliefs needed to exercise control over their health habits. In community-wide health campaigns, people's pre-existing beliefs that they can exercise control over their health habits, and the efficacy beliefs enhanced by the campaign, both contribute to health-promoting habits (Maibach et al. 1991).

3.2 *Adoption of Change*

Effective self-regulation of health behavior is not achieved through an act of will. It requires development of self-regulatory skills. To build people's sense of efficacy, they must develop skills on how to influence their own motivation and behavior. In such programs, they learn how to monitor their health behavior and the social and cognitive conditions under which they engage in it; set attainable subgoals to motivate and guide their efforts; draw from an array of coping strategies rather than rely on a single technique; enlist self-motivating incentives and social supports to sustain the effort needed to succeed; and apply multiple self-influence consistently and persistently (Bandura 1997, Perri 1985). Once equipped with skills and belief in their self-regulatory capabilities, people are better able to adopt behaviors that promote health, and to eliminate those that impair it. A large body of evidence

reveals that the self-efficacy belief system operates as a common mechanism through which psychosocial treatments affect different types of health outcomes (Bandura 1997, Holden 1991).

3.3 Maintenance of Change

It is one thing to get people to adopt beneficial health habits. It is another thing to get them to adhere to them. Maintenance of habit change relies heavily on self-regulatory capabilities and the functional value of the behavior. Development of self-regulatory capabilities requires instilling a resilient sense of efficacy as well as imparting skills. Experiences in exercising control over troublesome situations serve as efficacy builders. Efficacy affirmation trials are an important aspect of self-management because, if people are not fully convinced of their personal efficacy, they rapidly abandon the skills they have been taught when they fail to get quick results or suffer reverses. Like any other activity, self-management of health habits includes improvement, setbacks, plateaus, and recoveries. Studies of refractory detrimental habits show that a low sense of efficacy increases vulnerability to relapse (Bandura 1997, Marlatt et al. 1995). To strengthen resilience, people need to develop coping strategies not only to manage common precipitants of breakdown, but to reinstate control after setbacks. This involves training in how to manage failure (see *Health: Self-regulation*).

4. Self-management Health Systems

Healthcare expenditures are soaring at a rapid rate. With people living longer and the need for healthcare services rising with age, societies are confronted with major challenges on how to keep people healthy throughout their lifespan, otherwise they will be swamped with burgeoning health costs. Health systems generally focus heavily on the supply side with the aim of reducing, rationing, and curtailing access to health to contain health costs. The social cognitive approach works on the demand side by helping people to stay healthy through good self-management of health habits. This requires intensifying health promotion efforts and restructuring health delivery systems to make them more productive.

Efficacy-based models have been devised combining knowledge of self-regulation of health habits with computer-assisted implementation that provides effective health-promoting services in ways that are individualized, intensive and highly convenient (DeBusk et al. 1994). In this type of self-management system, people monitor their health habits. They set short-term goals for themselves and receive periodic feedback of progress towards their goals along with guides on how to manage troublesome situations. Efficacy ratings identify areas in which self-regulatory skills must be developed and strengthened if beneficial changes are to be achieved and maintained. The productivity of the system is vastly expanded by combining self-regulatory principles with the power of computer-assisted implementation. A single implementer, assisted with a computerized coordinating and mailing system, provides intensive individualized training in self-management for large numbers of people simultaneously. The self-management system reduces health risk factors, improves health status, and enhances the quality of life in cost-effective ways (Bandura 1997).

The self-management system is well received by participants because it is individually tailored to their needs, provides continuing personalized guidance and informative feedback that enables them to exercise considerable control over their own change; is a home-based program that does not require any special facilities, equipment, or attendance at group meetings that usually have high dropout rates; serves large numbers of people simultaneously under the guidance of a single implementer; is not constrained by time and place; and provides valuable health-promotion services at low cost. By combining the high individualization of the clinical approach with the large-scale applicability of the public health approach, the self-management system includes the features that ensure high social utility. Linking the interactive aspects of the self-management model to the Internet can vastly expand its availability for preventive and promotive guidance.

Chronic disease has become the dominant form of illness and the major cause of disability. The treatment of chronic disease must focus on self-management of physical conditions over time rather than on cure. This requires, among other things, pain amelioration, enhancement and maintenance of functioning with growing physical disability and development of self-regulative compensatory skills. Holman and Lorig (1992) have devised a prototypical model for the self-management of different types of chronic diseases. Patients are taught cognitive and behavioral pain control techniques; proximal goal setting combined with self incentives as motivators to increase levels of activity; problem solving and self-diagnostic skills; and the ability to locate community resources and to manage medication programs. How healthcare systems deal with clients can alter their efficacy in ways that support or undermine their restorative efforts. Clients are, therefore, taught how to take greater initiative for their healthcare in dealings with health personnel. These skills are developed through modeling of self-management skills, guided mastery practice, and enabling feedback.

The self-management program retards the biological progression of the disease, raises perceived self-regulatory efficacy, reduces pain and distress, fosters better cognitive symptom management, lessens the

impairment of role functions, improves the quality of life, and decreases the use of medical services. Both the perceived self-efficacy at the outset, and the efficacy beliefs instilled by the self-management program predict later health status and functioning (Holman and Lorig 1992).

5. *Childhood Health Promotion Models*

Many of the lifelong habits that jeopardize health are formed during childhood and adolescence (see *Childhood Health*). For example, unless youngsters take up the smoking habit as teenagers they rarely become smokers in adulthood. It is easier to prevent detrimental health habits than to try to change them after they have become deeply entrenched as part of a lifestyle. The biopsychosocial model provides a valuable public health tool for societal efforts to promote the health of its youth.

Health habits are rooted in familial practices. But, schools have a vital role to play in promoting the health of a nation. This is the only place where all children can be easily reached so it provides a natural setting for promoting healthful habits and building self-management skills.

Effective health promotion models include several major components. The first component is informational. It informs people of the health risks and benefits of different lifestyle habits. The second component develops the social and self-regulative skills for translating informed concerns into effective preventive action. As noted earlier, this includes self-monitoring of health practices, goal setting, and enlistment of self-incentives for personal change. The third component builds a resilient sense of self-regulatory efficacy to support the exercise of control in the face of difficulties that inevitably arise. Personal change occurs within a network of social influences. Depending on their nature, social influences can aid, retard, or undermine efforts at personal change. The final component, therefore, enlists and creates social supports for desired changes in health habits (see *Social Support and Health*).

Educational efforts to promote the health of youth usually produce weak results. They are heavy on didactics but meager on personal enablement. They provide factual information about health, but do little to equip children with the skills and self-beliefs that enable them to manage the emotional, and social pressures to adopt detrimental health habits. Managing health habits involves managing emotional states and diverse social pressures for unhealthy behavior not just targeting a specific health behavior for change.

Health promotion programs that encompass the essential elements of the self-regulatory model prevent or reduce injurious health habits. Health knowledge can be conveyed readily, but changes in values, attitudes, and health habits require greater effort. The more behavioral mastery experiences provided in the form of role enactments, the greater the beneficial changes (Bruvold 1993). The more intensive the program and the better the implementation, the stronger the impact (Connell et al. 1985). Comprehensive approaches that integrate guided mastery health programs with family and community efforts are more successful in promoting health and preventing adoption of detrimental health habits, than are programs in which the schools try to do it alone (Perry et al. 1992).

6. *Efficacy Beliefs in Prognostic Judgment and Health Outcomes*

Much of the work in the health field is concerned with diagnosing maladies, forecasting the likely course of different physical disorders and prescribing appropriate remedies. Medical prognostic judgments involve probabilistic inferences from knowledge of varying quality and inclusiveness about the multiple factors governing the course of a given disorder. One important issue regarding medical prognosis concerns the scope of determinants included in a prognostic model. Because psychosocial factors account for some of the variability in the course of health functioning, inclusion of self-efficacy determinants in prognostic models enhances their predictive power (Bandura 2000).

Recovery from medical conditions is partly governed by social factors. Recovery from a heart attack provides one example. About half the patients who experience heart attacks have uncomplicated ones. Their heart heals rapidly, and they are physically capable of resuming an active life. But the psychological and physical recovery is slow for those patients who believe they have an impaired heart. The recovery task is to convince patients that they have a sufficiently robust heart to resume productive lives. Spouses' judgments of patients' physical and cardiac capabilities can aid or retard the recovery process (see *Social Support and Recovery from Disease and Medical Procedures*). Programs that raise and strengthen spouses' and patients' beliefs in their cardiac capabilities enhance recovery of cardiovascular capacity (Taylor et al. 1985). The couple's joint belief in the patients' cardiac efficacy is the best predictor of improvement in cardiac functioning. Those who believe that their partners have a robust heart are more likely to encourage them to resume an active life than those who believe their partner's heart is impaired and vulnerable to further damage. Pursuit of an active life strengthens the cardiovascular system.

Prognostic judgments are not simply inert forecasts of a natural course of a disease. Prognostic expectations can affect patients' beliefs in their physical efficacy. Therefore, diagnosticians not only foretell,

but may partly influence the course of recovery from disease. Prognostic expectations are conveyed to patients by attitude, word, and the type and level of care provided them. People are more likely to be treated in enabling ways under positive than under negative expectations. Differential care that promotes in patients different levels of personal efficacy and skill in managing health-related behavior can exert stronger impact on the trajectories of health functioning than simply conveying prognostic information. Prognostic judgments have a self-confirming potential. Expectations can alter patients' sense of efficacy and behavior in ways that confirm the original expectations. The self-efficacy mechanism operates as one important mediator of self-confirming effects.

7. *Socially Oriented Approaches to Health*

The quality of health of a nation is a social matter not just a personal one. It requires changing the practices of social systems that impair health rather than just changing the habits of individuals. Vast sums of money are spent annually in advertising and marketing products and promoting lifestyles detrimental to health. With regard to injurious environmental conditions, some industrial and agricultural practices inject carcinogens and harmful pollutants into the air we breathe, the food we eat, and the water we drink, all of which take a heavy toll on health. Vigorous economic and political battles are fought over environmental health and where to set the limits of acceptable risk.

We do not lack sound policy prescriptions in the field of health. What is lacking is the collective efficacy to realize them. People's beliefs in their collective efficacy to accomplish social change by perseverant group action play a key role in the policy and public health approach to health promotion and disease prevention (Bandura 1997, Wallack et al. 1993). Such social efforts take a variety of forms. They raise public awareness of health hazards, educate and influence policymakers, devise effective strategies for improving health conditions, and mobilize public support to enact policy initiatives.

While concerted efforts are made to change socio-structural practices, people need to improve their current life circumstances over which they command some control. Psychosocial models that work best in improving health and preventing disease promote community self-help through collective enablement (McAlister et al. 1991). Given that health is heavily influenced by behavioral, environmental, and economic factors, health promotion requires greater emphasis on the development and enlistment of collective efficacy for socially oriented initiatives.

See also: Control Beliefs: Health Perspectives; Health Behavior: Psychosocial Theories; Health Behaviors; Health Education and Health Promotion; Health: Self-regulation; Self-efficacy; Self-efficacy: Educational Aspects

Bibliography

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ

Bandura A 1997 *Self-efficacy: The Exercise of Control*. W H Freeman, New York

Bandura A 2000 Psychological aspects of prognostic judgments. In: Evans R W, Baskin D S, Yatsu F M (eds.) *Prognosis of Neurological Disorders*, 2nd edn. Oxford University Press, New York, pp. 11–27

Bruvold W H 1993 A meta-analysis of adolescent smoking prevention programs. *American Journal of Public Health* **83**: 872–80

Connell D B, Turner R R, Mason E F 1985 Summary of findings of the school health education evaluation: Health promotion effectiveness implementation and costs. *Journal of School Health* **55**: 316–21

DeBusk R F et al. 1994 A case-management system for coronary risk factor modification after acute myocardial infarction. *Annals Of Internal Medicine* **120**: 721–9

Dienstbier R A 1989 Arousal and physiological toughness: Implications for mental and physical health. *Psychological Review* **96**: 84–100

Herbert T B, Cohen S 1993a Depression and immunity: A meta-analytic review. *Psychological Bulletin* **113**: 472–86

Herbert T B, Cohen S 1993b Stress and immunity in humans: a meta-analytic review. *Psychosomatic Medicine* **55**: 364–79

Holahan C K, Holahan C J 1987 Self-efficacy, social support, and depression in aging: a longitudinal analysis. *Journal of Gerontology* **42**: 65–8

Holden G 1991 The relationship of self–efficacy appraisals to subsequent health related outcomes: A meta–analysis. *Social Work in Health Care* **16**: 53–93

Holman H, Lorig K 1992 Perceived self-efficacy in self-management of chronic disease. In: Schwarzer R (ed.) *Self-Efficacy: Thought Control of Action*. Hemisphere Pub. Corp, Washington, DC, pp. 305–23

Maibach E, Flora J, Nass C 1991 Changes in self-efficacy and health behavior in response to a minimal contact community health campaign. *Health Communication* **3**: 1–15

Maier S F, Laudenslager M L, Ryan S M 1985 Stressor controllability, immune function, and endogenous opiates. In: Brush F R, Overmier J B (eds.) *Affect, Conditioning, and Cognition: Essays on the Determinants of Behavior*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 183–201

Marlatt G A, Baer J S, Quigley L A 1995 Self-efficacy and addictive behavior. In: Bandura A (ed.) *Self-efficacy in Changing Societies*. Cambridge University Press, New York, pp. 289–315

McAlister A L, Puska P, Orlandi M, Bye L L, Zbylot P 1991 Behaviour modification: Principles and illustrations. In: Holland W W, Detels R, Knox G (eds.) *Oxford Textbook of Public Health: Applications in Public Health*, 2nd edn. Oxford University Press, Oxford, UK, pp. 3–16

Meyerowitz B E, Chaiken S 1987 The effect of message framing on breast self-examination attitudes intentions and behavior. *Journal of Personality and Social Psychology* **52**: 500–10

O'Leary A 1990 Stress, emotion, and human immune function. *Psychological Bulletin* **108**: 363–82
Perri M G 1985 Self-change strategies for the control of smoking, obesity, and problem drinking. In: Shiffman S, Wills T A (eds.) *Coping and Substance Use*. Academic Press, Orland pp. 295–317
Perry C L, Kelder S H, Murray D M, Klepp K 1992 Community-wide smoking prevention: Long-term outcomes of the Minnesota heart health program and the class of 1989 study. *American Journal of Public Health* **82**: 1210–6
Schneiderman N, McCabe P M, Baum A (eds.) 1992 *Stress and Disease Processes: Perspectives in Behavioral Medicine*. Erlbaum, Hillsdale, NJ
Schwarzer R (ed.) 1992 *Self-efficacy: Thought Control of Action*. Hemisphere Pub. Corp, Washington, DC
Taylor C B, Bandura A, Ewart C K, Miller N H, Debusk R F 1985 Exercise testing to enhance wives' confidence in their husbands' cardiac capabilities soon after clinically uncomplicated acute myocardial infarction. *American journal of Cardiology* **55**: 635–8
Wallack L, Dorfman L, Jernigan D, Themba M 1993 *Media Advocacy and Public Health: Power for Prevention*. Sage, Newbury Park, CA
Wiedenfeld S A, Bandura A, Levine S, O'Leary A, Brown S, Raska K 1990 Impact of perceived self-efficacy in coping with stressors on components of the immune system. *Journal of Personality and Social Psychology* **59**: 1082–94

A. Bandura

Self-efficacy: Educational Aspects

Current theoretical accounts of learning and instruction postulate that students are active seekers and processors of information (Pintrich et al. 1986). Research indicates that students' cognitions influence the instigation, direction, strength, and persistence of achievement behaviors (Schunk 1995).

This article reviews the role of one type of personal cognition: self-efficacy, or one's perceived capabilities for learning or performing behaviors at designated levels (Bandura 1997). The role of self-efficacy in educational contexts is discussed to include the cues that students use to appraise their self-efficacy. A model of the operation of self-efficacy is explained, along with some key findings from educational research. The entry concludes by describing the role of teacher efficacy and suggesting future research directions.

1. *Self-efficacy Theory*

Self-efficacy can affect choice of activities, effort, persistence, and achievement (Bandura 1997, Schunk 1991). Compared with students who doubt their learning capabilities those with high self-efficacy for accomplishing a task participate more readily, work harder, persist longer when they encounter difficulties, and demonstrate higher achievement.

Learners acquire information to appraise self-efficacy from their performance accomplishments, vicarious (observational) experiences, forms of persuasion, and physiological reactions. Students' own performances offer reliable guides for assessing efficacy. Successes raise self-efficacy and failures lower it, but once a strong sense of self-efficacy is developed a failure may not have much impact (Bandura 1986).

Learners also acquire self-efficacy information from knowledge of others through classroom social comparisons. Similar others offer the best basis for comparison. Students who observe similar peers perform a task are apt to believe that they, too, are capable of accomplishing it. Information acquired vicariously typically has a weaker effect on self-efficacy than performance-based information because the former can be negated easily by subsequent failures.

Students often receive persuasive information from teachers and parents that they are capable of performing a task (e.g., 'You can do this'). Positive feedback enhances self-efficacy, but this increase will be temporary if subsequent efforts turn out poorly. Students also acquire efficacy information from physiological reactions (e.g., heart rate, sweating). Symptoms signaling anxiety might be interpreted to mean that one lacks skills.

Information acquired from these sources does not automatically influence self-efficacy; rather, it is cognitive appraised (Bandura 1986). In appraising efficacy, learners weigh and combine perceptions of their ability, the difficulty of the task, the amount of effort expended, the amount of external assistance received, the number and pattern of successes and failures, similarity to models, and credibility of persuaders (Schunk 1991).

Self-efficacy is not the only influence in educational settings. Achievement behavior also depends on knowledge and skills, outcome expectations, and the perceived value of outcomes (Schunk 1991). High self-efficacy does not produce competent performances when requisite knowledge and skills are lacking. Outcome expectations, or beliefs concerning the probable outcomes of actions, are important because students strive for positive outcomes. Perceived value of outcomes refers to how much learners desire certain outcomes relative to others. Learners are motivated to act in ways that they believe will result in outcomes they value.

Self-efficacy is dynamic and changes as learning occurs. The hypothesized process whereby self-efficacy operates during learning is as follows (Schunk 1996). Students enter learning situations with varying degrees of self-efficacy for learning. They also have goals in mind, such as learning the material, working quickly, pleasing the teacher, and making a high grade. As they engage in the task, they receive cues about how well they are performing, and they use these cues to assess

their learning progress and their self-efficacy for continued learning. Perceived progress sustains motivation and leads to continued learning. Perceptions of little progress do not necessarily diminish self-efficacy if learners believe they know how to perform better, such as by working harder, seeking help, or switching to a more effective strategy (Schunk 1996).

2. *Factors Affecting Self-efficacy*

There are many instructional, social, and environmental factors that operate during learning. Several of these factors have been investigated to determine how they influence learners' self-efficacy. For example, research has explored the roles of goal setting, social modeling, rewards, attributional feedback, social comparisons, progress monitoring, opportunities for self-evaluation of progress, progress feedback, and strategy instruction (Schunk 1995).

As originally conceptualized by Bandura, self-efficacy is a domain-specific construct. Self-efficacy research in education has tended to follow this guidance and assess students' self-efficacy within domains at the level of individual tasks. In mathematics, for example, students may be shown sample multiplication problems and for each sample judge their confidence for solving similar problems correctly. Efficacy scales typically are numerical and range from low to high confidence. After completing the efficacy assessment students are presented with actual problems to solve. These achievement test problems corresponding closely to those on the self-efficacy test, although they are not identical. Such specificity allows researchers to relate self-efficacy to achievement to determine correspondence and prediction (Pajares 1996). Other measures often collected by self-efficacy researchers include persistence, motivation, and self-regulation strategies. Following a pretest, students receive instruction complemented by one or more of the preceding educational variables. After the instruction is completed students receive a post-test; in some studies follow-up maintenance testing is done.

A general finding from much educational self-efficacy research is that educational variables influence self-efficacy to the extent that they convey to learners information about their progress in learning (Schunk 1995). For example, much research shows that specific proximal goals raise self-efficacy, motivation, and achievement better than do general goals (Schunk 1995). Short-term specific goals provide a clear standard against which to compare learning progress. As learners determine that they are making progress, this enhances their self-efficacy for continued learning. In contrast, assessing progress against a general goal (e.g., 'Do your best') is difficult; thus, learners receive less clear information about progress, and self-efficacy is not strengthened as well.

3. *Predictive Utility of Self-efficacy*

Self-efficacy research has examined the relation of self-efficacy to such educational outcomes as motivation, persistence, and achievement (Pajares 1996). Significant and positive correlations have been obtained across many studies between self-efficacy assessed prior to instruction and subsequent motivation during instruction. Initial judgments of self-efficacy have been found also to correlate positively and significantly with post-test measures of self-efficacy and achievement collected following instruction.

Multiple regression has been used to determine the percentage of variability in skillful performance accounted for by self-efficacy. Schunk and Swartz (1993) found that post-test self-efficacy was the strongest predictor of children's paragraph writing skills. Shell et al. (1989) found that although self-efficacy and outcome expectations predicted reading and writing achievement, self-efficacy was the strongest predictor.

Several studies have tested causal models. Schunk (1981) employed path analysis to reproduce the correlation matrix comprising long-division instructional method, self-efficacy, persistence, and achievement. The best model showed a direct effect of method on achievement and an indirect effect through persistence and self-efficacy, an indirect effect of method on persistence through self-efficacy, and a direct effect of self-efficacy on persistence and achievement. Schunk and Gunn (1986) found that the largest direct influence on achievement was due to use of effective learning strategies; achievement also was heavily influenced by self-efficacy.

4. *Teacher Self-efficacy*

Self-efficacy is applicable to teachers as well as students. Ashton and Webb (1986) postulated that self-efficacy should influence teachers' activities, efforts, and persistence. Teachers with low self-efficacy may avoid planning activities that they believe exceed their capabilities, may not persist with students having difficulties, may expend little effort to find materials, and may not reteach content in ways students might better understand. Teachers with higher self-efficacy might develop challenging activities, help students succeed, and persevere with students who have trouble learning. These motivational effects enhance student learning and substantiate teachers' self-efficacy by suggesting that they can help students learn.

Correlational data show that self-efficacy is related to teaching behavior. Ashton and Webb (1986) found that teachers with higher self-efficacy were likely to have a positive classroom environment (e.g., less student anxiety and teacher criticism), support students' ideas, and meet the needs of all students. High teacher self-efficacy was positively associated with use of praise, individual attention to students, checking on

students' progress in learning, and their mathematical and language achievement. Tschannen-Moran et al. (1998) discuss teacher self-efficacy in greater depth.

5. Future Research Directions

The proliferation of self-efficacy research in education has enlightened understanding of the construct but also has resulted in a multitude of measures. In developing efficacy assessments it is imperative that researchers attempt to be faithful to Bandura's (1986) conceptualization of self-efficacy as a domain-specific measure. Research will benefit from researchers publishing their instruments along with validation data to include reliability and validity.

As self-efficacy research continues in settings where learning occurs, it will be necessary to collect longitudinal data showing how self-efficacy changes over time as a consequence of learning. This focus will require broadening of self-efficacy assessments from reliance on numerical scales to qualitative data. Researchers also should relate measures of teacher self-efficacy to those of student self-efficacy to test the idea that these variables reciprocally influence one another.

Finally, research is needed on the role of self-efficacy during self-regulation. Self-regulation refers to self-generated thoughts and actions that are systematically oriented toward attainment of one's learning goals (Zimmerman 1990). Self-efficacy has the potential to influence many aspects of self-regulation, yet to date only a few areas have been explored in research. This focus will become more critical as self-regulation assumes an increasingly important role in education.

Bibliography

Ashton P T, Webb R B 1986 *Making a Difference: Teachers' Sense of Efficacy and Student Achievement*. Longman, New York

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ

Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman, New York

Pajares F 1996 Self-efficacy beliefs in academic settings. *Review of Educational Research* **66**: 543–78

Pintrich P R, Cross D R, Kozma R B, McKeachie W J 1986 Instructional psychology. *Annual Review of Psychology* **37**: 611–51

Schunk D H 1981 Modeling and attributional effects on children's achievement: A self-efficacy analysis. *Journal of Educational Psychology* **73**: 93–105

Schunk D H 1991 Self-efficacy and academic motivation. *Educational Psychologist* **26**: 207–31

Schunk D H 1995 Self-efficacy and education and instruction. In: Maddux J E (ed.) *Self-efficacy, Adaptation, and Adjustment: Theory, Research, and Application*. Plenum, New York, pp. 281–303

Schunk D H 1996 Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal* **33**: 359–82

Schunk D H, Gunn T P 1986 Self-efficacy and skill development: Influence of task strategies and attributions. *Journal of Educational Research* **79**: 238–44

Schunk D H, Swartz C W 1993 Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary Educational Psychology* **18**: 337–54

Shell D F, Murphy C C, Bruning R H 1989 Self-efficacy and outcome expectancy mechanisms in reading and writing achievement. *Journal of Educational Psychology* **81**: 91–100

Tschannen-Moran M, Hoy A W, Hoy W K 1998 Teacher efficacy: Its meaning and measure. *Review of Educational Research* **68**: 202–48

Zimmerman B J 1990 Self-regulating academic learning and achievement: The emergence of a social cognitive perspective. *Educational Psychology Review* **2**: 173–201

D. H. Schunk

Self-esteem in Adulthood

Self-esteem refers to a global judgment of the worth or value of the self as a whole (similar to self-regard, self-respect, and self-acceptance), or to evaluations of specific aspects of the self (e.g., appearance self-esteem or academic self-esteem). The focus of this article is on global self-esteem, which has distinct theoretical importance and consequences (Baumeister 1998, Rosenberg et al. 1995). Although thousands of studies of self-esteem have been published (Mruk 1995), many central questions about the nature, functioning, and importance of global self-esteem remain unresolved (Baumeister 1998).

1. The Importance of Self-esteem

Global self-esteem is a central aspect of the subjective quality of life. It is related strongly to positive affect and life satisfaction (Diener 1984), less anxiety (Solomon et al. 1991), and fewer depressive symptoms (Crandall 1973). High and low levels of self-esteem appear to be associated with different motivational orientations. High self-esteem people focus on self-enhancement and 'being all that they can be,' whereas low self-esteem people focus on self-protection and avoiding failure and humiliation (Baumeister et al. 1989). Low self-esteem has been proposed as a cause of many social problems, such as teenage pregnancy, aggression, eating disorders, and poor school achievement. However, evidence that low self-esteem is a cause, rather than a symptom, of these problems is scarce (Baumeister 1998, Dawes 1994, Mecca et al. 1989), and some researchers have suggested that high

self-esteem may actually be the cause of social problems such as aggression (Baumeister 1998). Because of unresolved issues regarding the nature and functioning of self-esteem and how to measure it, firm conclusions about whether high self-esteem is or is not socially useful seem premature.

2. Issues in the Self-esteem Literature

2.1 Trait or State?

Psychologists assume typically that self-esteem is a psychological trait (i.e., that it is stable over time and across situations). In support of this view, self-esteem tends to be highly stable over long periods of time (see, e.g., Rosenberg 1979). Self-esteem is also a state, however, changing in response to events and experiences in the course of life (Heatherton and Polivy 1991). James (1890) suggested that self-esteem has qualities of both a state and a trait, rising and falling in response to achievements and setbacks relevant to one's aspirations. On the other hand, he recognized that people tend to have average levels of self-esteem that are not linked directly to their objective circumstances. Although research supports James' intuition, for some people self-esteem is relatively stable and trait-like across time, whereas for others it is more state-like, fluctuating daily (Kernis and Waschull 1995).

2.2 Affect or Cognitive Judgment?

Researchers disagree about whether self-esteem is fundamentally a feeling or a judgment about the self. Global self-esteem and mood are correlated strongly, leading some to conclude that affect is a component of self-esteem (e.g., Brown 1993, Pelham and Swann 1989). Others argue that self-esteem is a cognitive judgment, based on standards of worth and accessible information about how well an individual is meeting those standards (see, e.g., Moretti and Higgins 1990). Current mood may be one source of information on which judgments of self-worth are based (Schwarz and Strack 1999). It seems likely that the relationship between self-esteem and mood is complex: both mood and self-esteem may be affected independently by life events; mood may be a source of information on which judgments of self-esteem are based; and mood may also be a consequence of having high or low self-esteem.

2.3 Where Does Self-esteem Come From?

Why are some people high and others low in self-esteem? James (1890) suggested that global self-esteem is determined by successes divided by pretensions, or how well a person is doing in areas that are important. Although it might seem logical that high self-esteem results from success in life (e.g., being smart, attractive, wealthy, and popular), these objective outcomes are related only weakly to self-esteem. For example, socioeconomic status (Twenge and Campbell 1999), physical attractiveness as rated by observers (Diener et al. 1995, Feingold 1992), obesity (Miller and Downey 1999), school achievement (Rosenberg et al. 1995), and popularity (Wylie 1979) are related only weakly to global self-esteem. A stronger relationship is observed between global self-esteem and how well people *believe* they are doing in important domains, but the direction of this relationship is unclear. Believing one is doing well in important domains might cause high self-esteem, or people may think they are doing well because they have high self-esteem. Experimental studies have demonstrated that specific self-evaluations are sensitive to manipulated success or failure, but evidence that global self-esteem responds to such feedback is very scarce (Blascovich and Tomaka 1991).

Mead (1934) and Cooley (1902) proposed that self-esteem develops in social relationships. Cooley (1902) argued that subjectively-interpreted feedback from others is a main source of information about the self. The self-concept arises from imagining how others perceive and evaluate the self. These 'reflected appraisals' affect self-perceptions and self-evaluations, resulting in what Cooley (1902) described as the 'looking glass self.' Mead (1934) argued that the looking glass self is a product of, and essential to, social interaction. To interact smoothly and effectively with others, people need to anticipate how others will react to them, and so they need to learn to see themselves through the eyes of others, either the specific people with whom they interact, or a generalized view of how most people see the self, or a 'generalized other.' Research indicates that self-esteem is related only weakly to others' evaluations of the self, but is related strongly to beliefs about others' evaluations (Shrauger and Schoeneman 1979). Evidence regarding the causal direction of this effect is scarce.

A third view, which can encompass these others, is that self-esteem is a judgment of self-worth constructed on the basis of information and standards for the self that are available and accessible at the moment. People may differ in the self-standards that are chronically accessible to them (see, e.g., Higgins 1987). For example, some people may judge their self-worth chronically according to whether they are competent in important domains, whereas others judge their self-worth chronically according to whether others approve of them (Crocker and Wolfe 2001). In this view, self-esteem will be stable over time if the standards used to evaluate the self and information about how well an individual is doing relative to those standards are stable. When circumstances make alternative standards salient, or alter beliefs about how well the

individual is doing relative to those standards, then self-esteem changes (Crocker 1999, Quinn and Crocker 1999).

2.4 Defensive or Genuine?

Genuine self-esteem is usually assumed to be synonymous with self-worth, self-respect, and self-acceptance, despite awareness of one's flaws and shortcomings. Yet, many studies have demonstrated that people who are high in self-esteem are more likely to make excuses for failure, derogate others when threatened, and have unrealistically positive views of themselves (see Baumeister 1998, Blaine and Crocker 1993, Taylor and Brown 1988 for reviews), behaviors which appear to be quite defensive. These findings have fueled the suspicion that many people who are outwardly high in self-esteem inwardly harbor serious doubts about their self-worth, and have defensive, rather than genuinely high, self-esteem. Although the distinction between genuine and defensively high self-esteem has a long history in psychology, researchers have had little success at distinguishing these two types of high self-esteem empirically.

One view is that implicit, or nonconscious, evaluations of the self are dissociated from conscious self-evaluations (see, e.g., Greenwald and Banaji 1995). According to this view, genuine high self-esteem results from having high explicit (conscious) and implicit (nonconscious) self-esteem, whereas defensively high self-esteem results from having high explicit self-esteem and low implicit self-esteem. A measure of implicit self-esteem based on the Implicit Associations Test (Greenwald et al. 1998) shows the predicted dissociation between implicit and explicit measures, but to date research has not demonstrated that defensive behaviors such as blaming others for failure are associated uniquely with the combination of high explicit and low implicit self-esteem. Another view is that people who have stable high self-esteem are relatively nondefensive in the face of failure, whereas people with unstable high self-esteem are both defensive and hostile when they fail (see Kernis and Waschull 1995 for a review). According to Kernis and his colleagues, people with unstable high self-esteem have a high level of ego-involvement in everyday events, consequently their self-esteem is at stake even when relatively minor negative events occur. Considerable evidence has accumulated supporting the view that people with unstable high self-esteem are defensive, whereas people with stable high self-esteem are not. Crocker and Wolfe (2001) argue that instability of self-esteem results when outcomes in a person's life are relevant to their contingencies, or conditions, of self-worth. Consequently, Crocker and Wolfe argue that people are defensive when they receive negative or threatening information in domains in which their self-esteem is contingent. To date, however, they have not provided empirical support for their view.

In general, the issue of defensive vs. genuine self-esteem has focused attention on dimensions of self-esteem that go beyond whether it is high or low. This broader perspective on multiple dimensions of self-esteem might help resolve several issues in the self-esteem literature (Crocker and Wolfe 2001). For example, the role of self-esteem in social problems such as substance abuse and eating disorders may be linked to instability or contingencies of self-esteem as well as, or instead of, level of self-esteem.

2.5 A Cultural Universal?

Psychologists have long assumed that there is a universal need to have high self-esteem. Consistent with this view, most people have high self-esteem, and will go to great lengths to achieve, maintain, and protect this. Yet, almost all of this research has been conducted in a North American cultural context, leaving open the possibility that the need for high self-esteem is a culturally specific phenomenon (see Heine et al. 1999 for a review). Levels of self-esteem in Asians are related to how much time they have spent in the USA or Canada (Heine et al. 1999). Furthermore, Asians and Asian-Americans, on average, do not show the same self-enhancing tendencies so characteristic of North Americans. Heine et al. argue that there are fundamental cultural differences in the nature and importance of self-esteem. Taking Japan as an example, they argue that, whereas in the USA and Canada people are self-enhancing and motivated to achieve, maintain, and protect high self-esteem, in Japan people are self-critical and motivated to improve the self. This self-critical orientation in Japan, they argue, is adaptive in a culture that values self-criticism and self-improvement, and considers them to be evidence of commitment to the group. Consequently, in contrast to North America, self-criticism may lead to positive self-feelings resulting from living up to cultural standards that value self-criticism. The notion that the need for self-esteem and the motivation to self-enhance are not universal is a crucial development in self-esteem research.

3. Measurement of Self-esteem

In 1974, Wylie reviewed research on self-esteem and criticized researchers for developing idiosyncratic measures of self-esteem, rather than well-established, psychometrically valid and reliable instruments. Although measurement of self-esteem has improved somewhat since Wylie's critique (see Blascovich and Tomaka 1991 for a review), there remains a tendency

for researchers to develop idiosyncratic measures for particular studies. Issues in the measurement of self-esteem tend to reflect issues in its conceptualization. Measures of trait self-esteem assess global judgments of self-worth, self-respect, or self-regard that encourage respondents to consider how they usually or generally evaluate themselves (see, e.g., Rosenberg 1965), or assess evaluations of the self in several domains and create a composite score, on the assumption that such a composite is an indicator of global self-esteem (see, e.g., Coopersmith 1967). Measures of state self-esteem, on the other hand, focus on how the person feels at a specific moment in time. Some state measures assess momentary evaluations of the self in one or more domains such as appearance or performance (see, e.g., Heatherton and Polivy 1991). Others assess current mood or self-related affect with self-ratings on items such as feeling proud, important, and valuable (see, e.g., Leary et al. 1995). Both types of state self-esteem measures appear to be responsive to positive and negative events. Researchers interested in state self-esteem tend not to measure momentary or current self-worth, self-regard, or self-respect (i.e., global state self-esteem). Consequently, studies of state self-esteem and studies of trait self-esteem tend to measure different constructs, making results across these types of studies difficult to compare.

4. Future Directions

Several important issues remain to be addressed by research. First, additional progress is needed in measurement and conceptualization of self-esteem, and in identifying and validating different types of self-esteem (e.g., defensive vs. genuine, implicit vs. explicit, stable vs. unstable, and contingent vs. noncontingent). Possible cultural and subcultural differences in the nature and functioning of self-esteem are a very important area needing further exploration. Only after progress has been made in these areas will researchers be able to provide definitive answers to questions about the social importance of self-esteem.

See also: Intrinsic Motivation, Psychology of; Self-efficacy; Self-evaluative Process, Psychology of; Self: History of the Concept; Self-regulation in Adulthood; Well-being (Subjective), Psychology of

Bibliography

Baumeister R F 1998 The self. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, pp. 680–740

Baumeister F, Tice D M, Hutton D G 1989 Self-presentational motivations and personality differences in self-esteem. *Journal of Personality* **57**: 547–79

Blaine B, Crocker J 1993 Self-esteem and self-serving biases in reactions to positive and negative events: An integrative review. In: Baumeister R F (ed.) *Self-esteem: The Puzzle of Low Self-regard*. Erlbaum, Hillsdale, NJ, pp. 55–85

Blascovich J, Tomaka J 1991 Measures of self-esteem. In: Robinson J P, Shaver P R, Wrightsman L S (eds.) *Measures of Personality and Social Psychological Attitudes*. Academic Press, San Diego, CA, pp. 115–60

Brown J D 1993 Motivational conflict and the self: The double-bind of low self-esteem. In: Baumeister R F (ed.) *Self-esteem: The Puzzle of Low Self-regard*. Plenum, New York, pp. 117–30

Cooley C H 1902 *Human Nature and the Social Order*. Schocken, New York

Coopersmith S 1967 *The Antecedents of Self-esteem*. W. H. Freeman, San Francisco, CA

Crandall R 1973 The measurement of self-esteem and related constructs. In: Robinson J, Shaver P R (eds.) *Measures of Social Psychological Attitudes*. Institute for Social Research, Ann Arbor, MI

Crocker J 1999 Social stigma and self-esteem: Situational construction of self-worth. *Journal of Experimental Social Psychology* **35**: 89–107

Crocker J, Wolfe C T 2001 Contingencies of self-worth. *Psychological Review* **108**: 593–623

Dawes R M 1994 *House of Cards: Psychology and Psychotherapy Built on Myth*. Free Press, New York

Diener E 1984 Subjective well-being. *Psychological Bulletin* **95**: 542–75

Diener E, Wolsic B, Fujita F 1995 Physical attractiveness and subjective well-being. *Journal of Personality and Social Psychology* **69**: 120–29

Feingold A 1992 Good-looking people are not what we think. *Psychological Bulletin* **111**: 304–41

Greenwald A G, Banaji M R 1995 Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* **102**: 4–27

Greenwald A G, McGhee D E, Schwarz J L K 1998 Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* **74**: 1464–80

Heatherton T F, Polivy J 1991 Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social Psychology* **60**: 895–10

Heine S J, Lehman D R, Markus H R, Kitayama S 1999 Is there a universal need for positive self-regard? *Psychological Review* **106**: 766–94

Higgins E T 1987 Self-discrepancy: A theory relating self and affect. *Psychological Review* **94**: 319–40

James W 1890 *The Principles of Psychology*. Harvard University Press, Cambridge, MA, Vol. 1

Kernis M H, Waschull S B 1995 The interactive roles of stability and level of self-esteem: Research and theory. In: Zanna M P (ed.) *Advances in Experimental Social Psychology*. Academic Press, San Diego, CA, Vol. 27, pp. 93–141

Leary M R, Tambor E S, Terdal S K, Downs D L 1995 Self-esteem as an inter-personal monitor: The sociometer hypothesis. *Journal of Personality and Social Psychology* **68**: 518–30

Mead G H 1934 *Mind, Self, and Society*. University of Chicago Press, Chicago, IL

Mecca A M, Smelser N J, Vasconcellos J (eds.) 1989 *The Social Importance of Self-esteem*. University of California Press, Berkeley, CA

Miller C T, Downey K T 1999 A meta-analysis of heavyweight and self-esteem. *Personality and Social Psychology Review* **3**: 68–84
Moretti M M, Higgins E T 1990 Relating self-discrepancy to self-esteem: The contribution of discrepancy beyond actual self-ratings. *Journal of Experimental Social Psychology* **26**: 108–23
Mruk C 1995 *Self-esteem: Research, Theory, and Practice*. Springer, New York
Pelham B W, Swann W B Jr 1989 From self-conceptions to self-worth: On the sources and structure of global self-esteem. *Journal of Personality and Social Psychology* **57**: 672–80
Quinn D M, Crocker J 1999 When ideology hurts: Effects of feeling fat and the protestant ethic on the psychological well-being of women. *Journal of Personality and Social Psychology* **77**: 402–14
Rosenberg M 1965 *Society and the Adolescent Self-image*. Princeton University Press, Princeton, NJ
Rosenberg M 1979 *Conceiving the Self*. Basic Books, New York
Rosenberg M, Schooler C, Schoenbach C, Rosenberg F 1995 Global self-esteem and specific self-esteem: Different concepts, different outcomes. *American Sociological Review* **60**: 141–56
Schwarz N, Strack F 1999 Reports of subjective well-being: Judgmental processes and their methodological implications. In: Kahneman D, Diener D, Schwarz N (eds.) *Well-being: Foundations of Hedonic Psychology*. Russell Sage, New York, pp. 61–84
Shrauger J S, Schoeneman T J 1979 Symbolic interactionist view of self-concept: Through the looking-glass darkly. *Psychological Bulletin* **86**: 549–73
Solomon S, Greenberg J, Pyszczynski T 1991 A terror management theory of social behavior: The psychological functions of self-esteem and cultural worldviews. In: Zanna M P (ed.) *Advances in Experimental Social Psychology*. Academic Press, San Diego, CA, Vol. 24, pp. 91–159
Taylor S E, Brown J D 1988 Illusion and well-being: A social-psychological perspective on mental health. *Psychological Bulletin* **103**: 193–210
Twenge J M Campbell W K 1999 *Does Self-esteem Relate to Being Rich, Successful, and Well-educated? A Meta-analytic Review*. Manuscript, Case-Western Reserve University
Wylie R C 1974 *The Self-concept* (rev. edn.). University of Nebraska Press, Lincoln, NE
Wylie R C 1979 *The Self-concept: Theory and Research on Selected Topics*, 2nd edn. University of Nebraska Press, Lincoln, NE, Vol. 2

J. Crocker

Self-evaluative Process, Psychology of

This article explores some of the processes associated with evaluative response to one's own self. Evaluation refers to a response registering the idea/feeling that some aspect of one's world is good or bad, likeable or dislikable, valuable or worthless. It is one of our most fundamental psychological responses. Indeed, research on the psychology of meaning has revealed that evaluation is the single most important aspect of meaning. Evaluative responses are fast; there is evidence that we have an evaluation of some things even before we are completely able to recognize them. Evaluative responses are sometimes automatic—we do not set out to make them and we often cannot turn them off. Moreover, evaluation often colors our interpretation of situations. For example, ambiguous actions of people we like are interpreted more benevolently than the same actions of people we do not like. An evaluative response attached to the self is often termed self-esteem and we will use the terms self-esteem and self-evaluation interchangeably.

1. *Individual Differences in Self-evaluation*

There are literally thousands of studies reported since the 1950s measuring self-esteem and comparing persons who are high with persons who are low on this dimension (see *Self-esteem in Adulthood*). Most frequently, self-esteem is assessed by self-report. One of the most popular measures (Rosenberg 1965) consists of ten items like 'I am a person of worth' followed by a series of graded response options, e.g., strongly agree, agree, disagree, strongly disagree. Such measures have proven to be reliable and valid. However, they are subject to the same general criticisms of any self-report measure: Scores can be distorted by the tendency to agree with an item regardless of its content and the tendency to try to create a favorable impression. Moreover, there may be aspects of one's self-evaluation that are not easily accessible to conscious awareness. To address some of these concerns 'implicit' measures of self-esteem are currently being explored (Greenwald and Banaji 1995). Most of these measures work by priming the self, i.e., making the self salient, and then measuring the impact of self-salience on other evaluative responses. For example, when the self is primed, the more positive the self-evaluation the faster one should be in making other positive evaluative judgments. As of this writing, implicit measures of self-evaluation show great promise but it is still unclear what impact such measures will have on our ultimate understanding of self-evaluation.

Individual differences in self-evaluation have been associated with a variety of psychological traits. For example, compared to persons with low self-esteem, persons with high self-esteem tend to achieve more in school, be less depressed, better adjusted, less socially anxious and more satisfied with life, etc. Going through this research almost leads one to draw the conclusion that all the good things in life are positively associated with self-esteem. Indeed, the intuition that high self-esteem is good is so compelling that the state of California even put together a task force to promote self-esteem. There are, of course, many arguments for the positive impact of self-esteem. However, most of

the studies rely on correlational methods. Correlational methodology makes it difficult to know if self-esteem is a cause or an effect of these other variables. For example, it may be that high self-esteem leads to school achievement but it may also be that school achievement improves self-esteem. It may also be that the correlation between achievement and self-esteem is not causal at all; each may be caused by the same third variable, e.g., general health.

Recent research is beginning to correct the simple view of self-esteem as always 'good.' For example, it may be persons who are high in self-esteem rather than persons who are low in self-esteem that are most likely to be aggressive (Baumeister et al. 1996). Why? Persons high in self-esteem have more to lose when confronted by failure or a personal affront. Related to this suggestion is the observation that self-esteem may be stable in some persons but unstable in others. Stability of self-esteem is consequential (Kernis and Waschull 1995). Persons whose self-esteem is high on the average but whose self-evaluation fluctuates over time score higher on a hostility measure than persons who are high in self-esteem but whose self-evaluation is stable. Perhaps it is persons who aspire to feel positive about themselves but are unsure of themselves that tend to fluctuate in their self-evaluation and to respond aggressively to threats to self-esteem.

2. *Self Motives*

The idea that persons strive to maintain a positive self-evaluation is obvious. It is not difficult to notice that people respond positively to success and compliments and negatively to failure and insults. They tend to seek out persons who respect their accomplishments and situations in which they can do well. In spite of the obviousness and ubiquity of a self-enhancement motive, at least two other motives have captured some research attention. One is the motive for self-knowledge, i.e., a self-assessment motive, and the other is a consistency or self-verification motive.

Feeling good about ourselves can take us only so far. It is also important to have accurate knowledge about the self. Leon Festinger (1954) suggested that we have a drive to evaluate our abilities and opinions. Indeed, we seem to be fascinated by information about ourselves. We want to know what others think of us. We go to psychotherapists to learn more about ourselves. We are even curious about what the stars have to say about our lives (note the popularity of horoscopes). Systematic research has varied the diagnosticity of experimental tasks, i.e., the extent to which we believe the task is truly revealing of our abilities. Under some conditions, for example, when certainty is low or when we are in a particularly good mood, we prefer diagnostic feedback to flattering feedback. Thus, there is evidence for the self-assessment motive.

Another motive that has received some research attention is the tendency to verify one's current view of the self (Swann 1990). According to this point of view, people are motivated to confirm their self-view. They will seek out persons and situations that provide belief consistent feedback. If a person has a positive view of self, he or she will seek out others who also evaluate them positively and situations in which they can succeed. Note that this is exactly the same expectation that could be derived from a self-enhancement point of view. However, self-verification and self-enhancement predictions diverge when a person has a negative view of self. The self-verification hypothesis predicts that persons with a negative self-view will seek out others who also perceive them negatively and situations that will lead to poor performance. There is some evidence for the self-verification prediction. On the other hand, neither the self-assessment motive nor the self-verification motive appears to be as robust as the self-enhancement motive (Sedikides 1993).

3. *Self-enhancement*

As noted above, self-enhancement processes are frequent, easy to observe, and robust. William James and a host of contemporary workers have focused on particular mechanisms by which self-evaluation is affected. For example, James suggests that threats to self-esteem are stronger if they involve abilities on which the self has 'pretensions' or aspires to do well. Others have noted that feelings of success and failure that affect self-evaluation often come from comparison with other persons. Still other investigators have shown the impact of self-enhancement on cognition. For example, the kinds of causal attributions people make for their own successes and failures are often self-serving. Persons tend to locate the causes of success internally (due to my trying, ability) and the causes of failure externally (due to bad luck, task difficulty). The self-enhancement mechanisms proposed in the psychological literature have been so numerous and so diverse that the collection of them has sometimes been dubbed the 'self zoo.' However, three general classes of mechanism encompass many of the proposed self-enhancement/protection mechanisms: Social comparison, inconsistency reduction, and value expression.

3.1 *Social Comparison*

One large class of self-enhancement mechanisms concerns social comparisons. The Self-Evaluation Maintenance (SEM) model (Tesser 1988), for example, proposes that when another person does better than we do at some activity, our own self-evaluation is affected. The greater our '*closeness*' to the other person

(through similarity, contiguity, personal relationship, etc.) the greater the effect on our self-evaluation. Being outperformed by another can lower self-evaluation by inviting unflattering self-comparison, or it can raise self-evaluation, a kind of 'basking in reflected glory'. (Examples of basking are seen in statements like, 'That's my friend Bob, the best widget maker in the county.') The *relevance*of the performance domain determines the relative importance of these opposing processes. If the performance domain is important to one's self-definition, i.e., high relevance, then the comparison process will be dominant. One's self-evaluation will be threatened by a close other's better performance. If the performance domain is unimportant to one's self-definition, i.e., low relevance, then the reflection process will be dominant. One's self-evaluation will be augmented by a close other's better performance. Thus, combinations of relative performance, closeness and relevance, are the antecedents of self-esteem threat or enhancement.

The assumption that people are motivated to protect or enhance self-evaluation, combined with the sketch of how another's performance affects self-evaluation, provides the information needed to predict self-evaluation maintenance behavior. An example: Suppose Nancy learns that she made a B+ on the test. The only other person from Nancy's dormitory in this chemistry class, Kaela, made an A+. This should be threatening to Nancy: Kaela outperformed her; Kaela is psychologically close (same dormitory); and chemistry is high in relevance to Nancy who is studying to be a doctor. What can Nancy do to reduce this threat and maintain a positive self-evaluation? She can change the performance differential by working harder herself or by preventing Kaela from doing well, e.g., hide the assignments, put the wrong catalyst in Kaela's beaker. She can reduce her psychological connection to Kaela, e.g. change dorms and avoid the same classes. Alternatively, she can convince herself that this performance domain is not self-relevant, e.g., chemistry is not highly relevant to the kind of medicine in which she is most interested. Laboratory experiments have produced evidence for each of these modes of dealing with social comparison threat to self-evaluation.

3.2 Cognitive Consistency

The number of variations within this approach to self-evaluation regulation is also substantial. An example of this approach is cognitive dissonance theory (Festinger 1957). According to dissonance theory, self-esteem is threatened by inconsistency. Holding beliefs that are logically or 'psychologically' inconsistent, i.e., dissonant, with one another is uncomfortable. For example, suppose a student agrees to a request to write an essay in favor of a tuition increase at her school. Her knowledge that she is opposed to a tuition increase is dissonant with her knowledge that she agreed to write an essay in favor of a tuition increase. One way to reduce this threatening dissonance is for the student to change her attitude to be more in favor of a tuition increase.

Note that social comparison mechanisms and consistency reduction mechanisms are both self-enhancement strategies, yet they seem to have little in common. Threat from dissonance rarely has anything to do with the performance of another, i.e., social comparison. Similarly, inconsistency is generally irrelevant to an SEM threat, whereas other's performance is crucial. Attitude change is the usual mode of dissonance threat reduction; on the other hand, changes in closeness, performance, or relevance are the SEM modes.

3.3 Value Expression

The notion that expressing one's most cherished values can affect self-esteem also has a productive history in social psychology. Simply expressing who we are, affirming our important values seems to have a positive effect on self-evaluation. According to self-affirmation theory (e.g., Steele 1988), self-evaluation has at its root a concern with a sense of global self-integrity. Self-integrity refers to holding self conceptions and images that one is 'adaptively and morally adequate, that is, as competent, good, coherent, unitary, stable, and capable of free choice, capable of controlling important outcomes, and so on' (Steele 1988, p. 262). If the locus of a threat to self-esteem is self-integrity then the behavior to reduce that threat is self-affirmation or a declaration of the significance of an important self-value. Again, note that as a self-enhancement strategy, affirming a cherished value is qualitatively different from the SEM behaviors of changing closeness, relevance or performance or the dissonance behavior of attitude change.

3.4 Putting It All Together

We have briefly described three classes of self-enhancement mechanisms: Social comparison, cognitive consistency and value expression. Each of these mechanisms is presumed to regulate self-evaluation, yet they are strikingly different from one another. These differences raise the question of whether self-evaluation is a unitary system or whether there are three (or more) independent self-evaluation systems. The goal of the self-enhancement motive is to maintain positive self-esteem. If there is a unitary self-evaluation system, the various self-evaluation mechanisms should substitute for one another. For example, if behaving inconsistently reduces self-evaluation then a positive social comparison experience, being part of the same

system, should be able to restore self-evaluation. On the other hand, if there are separate self-evaluation systems then a positive social comparison experience will not be able to restore a reduction in self-evaluation originating with inconsistent behavior. One would have to reduce the inconsistency to restore one's self-evaluation. Recent research favors the former interpretation. At least under certain circumstances, the three self-enhancement mechanisms are mutually substitutable for one another in maintaining self-evaluation. In short, self-evaluation appears to be a unitary system with multiple processes for regulating itself (Tesser et al. 1996).

4. *The Origins of Self-evaluation*

Psychologists have only recently begun to think about the origins of the self-enhancement motive. One line of research suggests that the self-enhancement motive grows out of an instinct for self-preservation coupled with knowledge of our own mortality. Although we as individuals may not live on, we understand that the culture of which we are a part does live on. 'Immortality' comes from our connection with our culture. Self-evaluation is a psychological indicator of the extent to which we are connected and acceptable to our culture and, hence, an index of our own 'immortality' (Pysczynski et al. 1997). Another line of work builds on the observation that evolution has predisposed us to be social, gregarious animals who are highly dependent on group living. We wish to maintain a positive self-esteem because self-esteem is a kind of 'sociometer' that indicates the extent to which we are regarded positively or negatively by others (Leary et al. 1995). Also taking an evolutionary perspective, the SEM model builds on the sociometer idea in two ways. It suggests that groups differ in the power they have to affect self-esteem. Compared to psychologically distant groups, psychologically close groups are typically more consequential to our well being and they have greater impact on our self-esteem. The SEM model also suggests that division of labor is fundamental to groups to maximize efficiency and to avoid conflict. Consequently, self-evaluation is more sensitive to feedback regarding the self's own niche in the group. See Beach and Tesser (in press) for discussion.

5. *Summary*

Self-evaluation has a productive history in psychology. Individual differences in self-esteem tend to be correlated with a number of positive attributes such as school achievement, general happiness, and lack of depression. There is a strong tendency for people to maintain a positive self-esteem but there is also evidence of motives for self-accuracy and for self-verification. At least three processes affect self-evaluation: social comparison, cognitive consistency, and value expression. Although these processes are qualitatively different from one another, they are substitutable for one another in maintaining self-esteem. Self-enhancement is thought to have evolutionary roots in the individual's connections to groups.

See also: Cognitive Dissonance; Self-esteem in Adulthood; Self-monitoring, Psychology of; Self-regulation in Adulthood; Social Comparison, Psychology of

Bibliography

Baumeister R F, Smart L, Boden J M 1996 Relation of threatened egotism to violence and aggression: The dark side of high self-esteem. *Psychological Review* **103**: 5–33

Beach S R H, Tesser A in press Self-evaluation maintenance and evolution: Some speculative notes. In: Suls J, Wheeler L (eds.) *Handbook of Social Comparison*. Lawrence Erlbaum Associates, Mahwah, NJ

Festinger L 1954 A theory of social comparison processes. *Human Relations* **7**: 117–40

Festinger L 1957 *A Theory of Cognitive Dissonance*. Row, Peterson, Evanston, IL

Greenwald A G, Banaji M R 1995 Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review* **102**: 4–27

James W 1905 *The Principles of Psychology*. Holt, New York, Vol. 1

Kernis M H, Waschull S B 1995 The interactive roles of stability and level of self-esteem: Research and theory. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, San Diego, CA, Vol. 27, pp. 93–141

Leary M R, Tambor E S, Terdal S K, Downs D L 1995 Self-esteem as an interpersonal monitor: The sociometer hypothesis. *Journal of Personality and Social Psychology* **68**: 518–30

Pysczynski T, Greenberg J, Solomon S 1997 Why do we need what we need? A terror management perspective on the roots of human social motivation. *Psychological Inquiry* **8**: 1–20

Rosenberg M 1965 *Society and the Adolescent Self-image*. Princeton University Press, Princeton, NJ

Sedikides C 1993 Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology* **65**(2): 317–38

Steele C M 1988 The psychology of self-affirmation: Sustaining the integrity of the self. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, San Diego, CA, Vol. 21, pp. 261–302

Swann W B 1990 To be adored or to be known? The interplay of self-enhancement and self-verification. In: Sorrentino R M, Higgins E T (eds.) *Handbook of Motivation & Cognition*. Guilford Press, New York, Vol. 2, pp. 408–48

Tesser A 1988 Toward a self-evaluation maintenance model of social behavior. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, San Diego, CA, Vol. 21, pp. 181–227

Tesser A, Martin L, Cornell D 1996 On the substitutability of self-protective mechanisms. In: Gollwitzer P M, Bargh J A (eds.) *The Psychology of Action: Linking Motivation and Cognition to Behavior*. Guilford Press, New York, pp. 48–67

A. Tesser

Self-fulfilling Prophecies

A self-fulfilling prophecy occurs when an originally false social belief leads to its own fulfillment. The self-fulfilling prophecy was first described by Merton (1948), who applied it to test anxiety, bank failures, and discrimination. This article reviews some of the controversies surrounding early self-fulfilling prophecy research, and traces how those controversies have led to modern research on relations between social beliefs and social reality.

Self-fulfilling prophecies did not receive much attention until Rosenthal and Jacobson's (1968) Pygmalion study. Teachers were led to believe that randomly selected students would show dramatic increases in IQ over the school year. Results seemed to show that, especially in the earlier grade levels, those students gained more in IQ than other students. Thus, the teachers' initially false belief that some students would show unusual IQ gains became true.

1. *Controversy, Replication, and Meta-analysis*

Rosenthal and Jacobson's study (1968) was highly controversial. Although it seemed to explain the low achievement of disadvantaged students, it was criticized on methodological and statistical grounds. This controversy inspired attempts at replication. Only about one third of these early attempts succeeded (Rosenthal and Rubin 1978). Critics concluded that the phenomenon was unreliable. Proponents concluded that this demonstrated the existence of self-fulfilling prophecies because, if chance differences were occurring, replications would only succeed 5 percent of the time.

This controversy inspired Rosenthal's work on meta-analysis—statistical techniques for summarizing the results of multiple studies. Rosenthal and Rubin's (1978) meta-analysis of the first 345 studies of interpersonal expectancy effects conclusively demonstrated that self-fulfilling prophecies are a real and reliable phenomenon. That meta-analysis also showed that they were neither pervasive (nearly two-thirds of the studies failed to find the effect) nor powerful (effect sizes, in terms of correlation or regression coefficients, averaged 0.2–0.3).

This, however, did not end the controversy. Although few modern researchers dispute the existence of self-fulfilling prophecies in general, several do dispute the claims that teacher expectations influence student *intelligence* (Snow 1995). For example, Snow (1995) concluded that: (a) the expectancy effect disappears if one removes five students with implausible IQ increases (100 points within one year) from the Rosenthal and Jacobson (1968) study, and (b) the literature fails to demonstrate an effect of teacher expectations on IQ.

2. *Self-fulfilling Stereotypes*

Some researchers saw in self-fulfilling prophecies an explanation for social inequalities. Thus, in the 1970s, research began addressing the self-fulfilling effects of stereotypes. The main ideas were that; if most stereotypes were inaccurate and if self-fulfilling prophecies were common and powerful (as was believed) then negative stereotypes regarding intelligence, achievement, motivation, etc., may produce self-fulfilling prophecies that lead individuals from devalued groups to objectively confirm those stereotypes.

The early experimental research seemed to support this perspective:

(a) White interviewers' racial stereotypes could undermine the performance of Black interviewees.

(b) Males acted more warmly towards, and evoked warmer behavior from, female interaction partners erroneously believed to be more physically attractive.

(c) When interacting with a sexist male who was either physically attractive or who was interviewing them for a job, women altered their behavior to appear more consistent with traditional sex stereotypes.

(d) Teachers used social class as a major basis for expectations and treated students from middle class backgrounds more favorably than students from lower class backgrounds (see reviews by Jussim et al. 1996 and Snyder 1984).

3. *Widespread Acceptance and More Questions*

The role of self-fulfilling prophecies, however, in creating social problems remained unclear because many of the early self-fulfilling prophecy experiments suffered important limitations. In most, if the expectancy manipulation was successful, perceivers developed erroneous expectations. However, under naturalistic conditions perceivers may develop accurate expectations, which, by definition, do not create self-fulfilling prophecies (because self-fulfilling prophecies begin with an initially false belief). Three categories of research have addressed the limitations of the early experiments in different ways.

3.1 *Naturalistic Studies of Teacher Expectations*

Longitudinal, quantitative investigations of naturally occurring teacher expectancies addressed the accuracy problem directly. All assessed relations between

teacher expectations and students' past and future achievement. If teacher expectations predicted future achievement beyond effects accounted for by students' past achievement, results were interpreted as providing evidence consistent with self-fulfilling prophecies. These studies also provided data capable of addressing two related questions:

(a) How large are naturally occurring teacher expectation effects?

(b) Do teacher expectations predict students' achievement more because they create self-fulfilling prophecies or more because they are accurate?

The results were consistent:

(a) In terms of standardized regression coefficients, the self-fulfilling effects of teacher expectations were about 0.1–0.2.

(b) Teacher expectations were strongly based on students' past achievement.

(c) Teachers' expectations predicted students' achievement more because they were accurate than because they led to self-fulfilling prophecies.

(d) Even teachers' perceptions of differences between students from different demographic groups (i.e., stereotypes) were mostly accurate (Jussim et al. 1996).

3.2 Naturalistic Studies of Close Relationships

Recent research has begun investigating the occurrence of self-fulfilling prophecies in close relationships. Children come to view their math abilities in a manner consistent with their mothers' sex stereotypes (Jacobs and Eccles 1992). New college roommates change each others' self-perceptions of academic and athletic ability (McNulty and Swann 1994). People who feel anxious that their romantic partners will reject them often evoke rejection from those partners (Downey et al. 1998). Furthermore, the more positive the illusions one holds regarding one's romantic partner the longer that relationship is likely to continue and the more positively one's romantic partner will come to view him or herself (Murray et al. 1996). As with the teacher expectation studies, however, self-fulfilling prophecy effect sizes average about 0.2.

3.3 Nonconscious Priming of Stereotypes

Chen and Bargh (1997) avoided inducing false expectations by nonconsciously priming a stereotype, and observing its effect on social interaction. First, they subliminally presented to perceivers either African-American or White faces. Perceivers and targets (all of whom were White) were then placed into different rooms where they communicated through microphones and headphones. These interactions were recorded and rated for hostility. Perceivers primed with an African-American face were rated as more hostile, and targets interacting with more hostile perceivers reciprocated with greater hostility themselves. The expectancy effect size was 0.23.

4. Moderators

Failures to replicate and generally small effect sizes prompted some researchers to begin searching for moderators—factors that inhibit or facilitate self-fulfilling prophecies (Jussim et al. 1996). This research has identified some conditions under which powerful self-fulfilling prophecies occurred and many conditions under which self-fulfilling prophecies did not occur. Identified moderators include characteristics of perceivers, targets, and the situation.

4.1 Perceiver Moderators

(a) Perceivers motivated to be accurate or sociable are not likely to produce self-fulfilling prophecies.

(b) Perceivers motivated to confirm a particular belief about a target, or to arrive at a stable impression of a target are more likely to produce self-fulfilling prophecies.

(c) Perceivers with a rigid cognitive style or who are certain of their beliefs about targets are more likely to produce self-fulfilling prophecies.

4.2 Target Moderators

(a) Unclear self-perceptions lead people to become more vulnerable to self-fulfilling prophecies.

(b) When perceivers have something targets want (such as a job), targets often confirm those beliefs in order to create a favorable impression.

(c) When targets desire to facilitate smooth social interactions, they are also more likely to confirm perceivers' expectations.

(d) When targets believe that perceivers hold a negative belief about them, they often act to disconfirm that belief. Similarly, when their main goal is to defend a threatened identity, or express their personal attributes, they are also likely to disconfirm perceivers' expectations.

(e) Self-fulfilling prophecies are stronger among students from at least some stigmatized social groups (African-American students, students from lower social class backgrounds, and students with a history of low achievement).

4.3 Situational Moderators

(a) Self-fulfilling prophecies are most common when people enter new situations, such as kindergarten or military service.

(b) Experimental studies conducted in educational contexts were much more likely to obtain self-fulfilling prophecies if the expectancy manipulation occurred early in the school year (presumably, because teachers were more open to the information at that time).

5. Accumulation

Small self-fulfilling prophecy effects, if they accumulate over time, might lead to large differences between targets. If, for example, stereotype-based expectations lead to small differences in the intellectual achievement of students from middle class or poor backgrounds each year, those differences may accumulate over time to lead to large social class differences in achievement. This argument lies at the heart of claims emphasizing the power of self-fulfilling prophecies to contribute to social problems.

Self-fulfilling prophecies in the classroom, however, do not accumulate. All studies examining this issue have failed to find accumulation and, instead, have generally found that teacher expectation effects dissipate over time (Smith et al. 1999). Whether self-fulfilling prophecies accumulate outside of the classroom is currently unknown.

6. Future Directions

6.1 Naturalistic Studies Outside of Classrooms

Nearly all of the early naturalistic research focused on teacher expectations. Thus, the recent emergence of research on self-fulfilling prophecies in close relationships has been sorely needed, and will likely continue. The expectations that parents, employers, therapists, coaches, etc. develop regarding their children, employees, clients, athletes, etc. are all rich areas for future research.

6.2 Stereotype Threat

Stereotype threat refers to concern that one's actions may fulfill a negative cultural stereotype of one's group (Steele 1997). Such concerns may, paradoxically, lead to the fulfillment of those stereotypes. For example, African-American students who believe they are taking a test of intelligence (triggering potential concern about confirming negative cultural stereotypes regarding African-American intelligence) perform worse than White students; however, when led to believe that the same test is one of 'problem-solving,' the differences evaporate. Similar patterns have occurred for women taking standardized math tests (Steele 1997), and for middle class and poor students on intelligence tests (Croizet and Claire 1998). Stereotype threat is a relatively new concept in the social sciences, and has been thus far used primarily to explain demographic differences in standardized test performance. In addition, it helps identify how cultural stereotypes (beliefs about the widespread beliefs regarding groups) may be self-fulfilling, even in the absence of a specific perceiver with an inaccurate stereotype. As such, it promises to remain an important topic in the social sciences for some time.

7. Conclusion

Self-fulfilling prophecies are pervasive in the sense that they occur in many different contexts. They are not pervasive in the sense that self-fulfilling prophecy effect sizes are typically small, and many studies have failed to find them.

Because of the alleged power of expectancy effects to create social problems, teachers have sometimes been accused of perpetrating injustices based on race, class, sex, and other demographic categories. This accusation is unjustified. Teacher expectations predict student achievement primarily because those expectations are accurate. Furthermore, even when inaccurate, teacher expectations do not usually influence students very much; and even when they do influence students, such influence is likely to dissipate over time.

Sometimes, however, both inside and outside the classroom, self-fulfilling prophecies can be powerful. In the classroom, the effects among some groups (low achievers, African-Americans, students from lower social class backgrounds) have been quite powerful. Although self-fulfilling prophecies in the classroom do not accumulate, they can be very long lasting—detectable as many as six years after the original teacher-student relationship (Smith et al. 1999).

Outside the classroom, recent research has demonstrated the potentially important role of self-fulfilling prophecies in close relationships, and in the maintenance of socio-cultural stereotypes. Thus, self-fulfilling prophecies occur in a wide variety of contexts and are a major phenomenon linking social perception to social behavior.

See also: Decision Making, Psychology of; Personality and Conceptions of the Self; Self-concepts: Educational Aspects; Stereotypes, Social Psychology of; Teacher Behavior and Student Outcomes

Bibliography

Chen M, Bargh J A 1997 Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology* **33**: 541–60

Croizet J, Claire T 1998 Extending the concept of stereotype and threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin* **24**: 588–94

Downey G, Freitas A L, Michaelis B, Khouri H 1998 The self-fulfilling prophecy in close relationships: Rejection sensitivity and rejection by romantic partners. *Journal of Personality and Social Psychology* **75**: 545–60

Jacobs J E, Eccles J S 1992 The impact of mothers' gender-role stereotypic beliefs on mothers' and children's ability perceptions. *Journal of Personality and Social Psychology* **63**: 932–944

Jussim L, Eccles J, Madon S 1996 Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology* **28**: 281–388

Merton R K 1948 The self-fulfilling prophecy. *Antioch Review* **8**: 193–210

McNulty S E, Swann W B 1994 Identity negotiation in roommate relationships: The self as architect and consequence of social reality. *Journal of Personality and Social Psychology* **67**: 1012–23

Murray S L, Holmes J G, Griffin D W 1996 The self-fulfilling nature of positive illusions in romantic relationships: Love is not blind, but prescient. *Journal of Personality and Social Psychology* **71**: 1155–80

Rosenthal R, Jacobson L 1968 *Pygmalion in the classroom: Teacher expectations and pupils' intellectual development*. Holt, Rinehart, and Winston, New York

Rosenthal R, Rubin D B 1978 Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences* **1**: 377–415

Smith A E, Jussim L, Eccles J 1999 Do self-fulfilling prophecies accumulate, dissipate, or remain stable over time? *Journal of Personality and Social Psychology* **77**: 548–65

Snow R E 1995 Pygmalion and intelligence? *Current Directions in Psychological Science* **4**: 169–71

Snyder M 1984 When belief creates reality. *Advances in Experimental Social Psychology* **18**: 247–305

Steele C M 1997 A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist* **52**: 613–29

L. Jussim

Self: History of the Concept

It is often asserted that the concept of the self emerges only in early modern times in connection with the concern for subjectivity that is taken to be characteristic of modernity. While it is true that the term 'self' as a noun, describing that which in a person is really and intrinsically this person, comes into use only from the seventeenth century onwards, what can be called the question of selfhood was not unfamiliar to earlier thinkers. This is, in its core, the question whether there is—and if so, what is—some unity, or at least continuity and coherence, of a human being over the life-course beyond their bodily constitution, beyond the unity of the body.

In the history of human thought, this question has been answered in a great variety of different ways. Broadly, empirical traditions tended to doubt the existence of such unity. They could observe changes, even radical transformations, in the human mind and had thus no ground to postulate any *a priori* unity. Transcendental traditions, in contrast, tended to argue that there had to be a unity of apperception, or of consciousness. Otherwise, not even the question of the self could be asked.

In the intellectual space between these two positions, so to say, the issue could be addressed in ways that are more specific to the social and psychological sciences. Then, the faculty of memory, for instance, could be seen as enabling a sense of continuous self to emerge. Human subjects have the ability to narrate their lives. Or, selves could be seen as formed in the interaction with others, that is, in the mirroring of an image of oneself through the responses to one's own words and actions by others. In moral philosophy, the existence of a continuous self was seen as a precondition for holding human beings accountable for their past deeds. Selfhood was thus linked to moral and political responsibility.

1. *Selfhood in the Social Sciences*

In these forms, the question of the self was already posed at the historical moment when the social sciences, by and large in the contemporary understanding of the term, arose towards the end of the eighteenth century. It is noteworthy, then, that the emerging social sciences rather reduced the range of ways of exploring selfhood across the nineteenth century. In some of their fields, like liberal political philosophy and political economy/economics, they postulated a rational self, able to make choices and to be responsible for her, or rather: his deeds. In other fields, in particular in the sociological way of reasoning, the orientations and behaviors of human beings were seen as determined by their social position. This opposition has become known as the one between an under- and an oversocialized conception of the human being (Wrong 1961). More cautiously, the empirical and behavioral strands of social research restricted themselves to observing human behavior and refrained from making any assumptions about selfhood at all. Regularities emerged here only through the aggregation of observations.

All these approaches had in common, though, was that they aimed at regularizing and stabilizing human orientations and behavior. Whether human beings were under- or oversocialized or just happened to behave according to patterns, the broader question of selfhood, namely whether and how a sense of continuity and coherence in a human being forms, was rather neglected or answered *a priori* by theoretical postulate.

In this light, it seems appropriate to say that the social sciences have developed a serious interest in

questions of human selfhood only late, systematically only in the early twentieth century. Furthermore, they have largely ignored ways of representing the human self that were proposed in other areas, in philosophy for instance, but maybe even more importantly in literature. Thus, issues such as individuality and subjectivity, the possible idiosyncrasy of a life-course and life-project, have long remained outside the focus of the social sciences, which have tended to look at the question from the perspective of a fully developed, stable, personal identity, rather than making the continuity and coherence of the self an issue open to investigation.

This focus can hardly be understood otherwise than through the perceived need to expect a political stability of the world on the basis of a presupposed coherence of human orientations and actions (see *Freedom/Liberty: Impact on the Social Sciences*).

The ground for an interest in such broader issues was prepared outside of what is conventionally recognized as social science, namely by Friedrich Nietzsche, and later by Sigmund Freud. Nietzsche radically rejected the problems of moral and political philosophy and thus liberated the self from the impositions of the rules of the collective life. Freud located the drives toward a fuller realization of one's self in the human psyche and connected the history of civilization with the repression of such drives. Against such a 'Nietzschean-Freudian' background (Rorty 1989), Georg Simmel and George Herbert Mead could observe the ways in which identities are formed in social interaction and conceptualize variations of self-formation in different social contexts. From then on, a sociology and social psychology of selfhood and identity has developed which no longer relies on presuppositions about some essence of human nature and is able to connect its findings to both child psychology and phenomenology. It emphasizes the socially constructed nature of selfhood, but remains capable, at least in principle, to analyze the specific social contexts of self-formation thus working towards a comparative-historical sociology of selfhood. This broadened debate on selfhood has created a semantic space in which aspects of the self are emphasized in various ways (Friese 2001). This space can be described by connecting the concept of self to notions of modernity, of meaning, and of difference.

2. *Selfhood and Modernity*

The idea of human beings as autonomous subjects is often taken to be characteristic of modernity as an era, or at least for the self-understanding of modernity as an ethos (see *Modernity: History of the Concept*). Such a view entails a conception of the self as rather continuous and coherent, since only on such a basis is the choice of a path of action and a way of life as well as the acceptance of the responsibility for one's deeds conceivable. The concept of selfhood is here conditioned by the need to maintain a notion of human autonomy and agentiality as a basic tenet of what modernity is about, namely the possibility to shape the world by conscious human action. Such modernism, however, is not necessarily tied to the atomist and rationalist individualism of some versions of economic and political thought, most notably neoclassical economics and rational choice theory.

In the light of the twentieth century developments in sociology and (social) psychology, as mentioned above, the view that connects selfhood to modernity mostly—in all its more sophisticated forms—starts out from an assumption of constitutive sociality of the human being. Not absolute autonomy, but rather the conviction that human beings have to construct their self-identities can then be seen as characteristically modern (Hollis 1985). Unlike the concept of the rational, autonomous self, this concept is open towards important qualifications in terms of the corporeality, situatedness, and possible nonteleological character of human action (Joas 1996). Without having to presuppose the self-sustained individual of modernity, the aim is to demonstrate how autonomous selves develop through social interactions over certain phases of the life-course (Joas 1998, Straub 2001).

The commitment to autonomy and agentiality, characteristic of modernity, becomes visible rather in the fact that formation of self (or, of self-identity) is here understood as the forming and determining of the durably significant orientations in a life. It is related to the formation of a consciousness of one's own existence and thus biographically predominantly to the period of adolescence. Crises of identity occur accordingly during growing up; more precisely one should speak of life crises during the formation of one's identity. Self-identity once constituted is seen as basically stable further on. No necessary connection is presupposed between self-formation and individuality; theoretically, human beings may well form highly similar identities in great numbers. Since the very concept of identity is connected to continuity and coherence of self, however, stability is turned into a conceptual assumption.

Objections against such a conceptualization go in two different directions. On the one hand, this discourse, which often has its roots in (social) psychology, stands in a basic tension to any culturalist concept of identity, which emphasizes meaning. On the other hand, doubts about the presupposition of continuity and coherence of selfhood employ notions of difference and alterity that are not reducible to the idea that selves are formed by relating to others, by intersubjectivity.

3. *Selfhood and Meaning*

Some critics of the close connection between self-formation and modernity argue that every form of selfhood is dependent on the cultural resources that

are at hand to the particular human being when giving shape to their important orientations in life. Human beings give meaning to their lives by interpreting their situations with the help of moral-cultural languages that precede their own existence and surround them. Cultural determinism is the strong version of such theorizing, mostly out of use nowadays, but many current social theories adopt a weaker version of this reasoning which indeed sustains the notion of the continuity and coherence of the self but sees this self as strongly embedded in cultural contexts.

Such a view of the self has its modern source in romanticism (see *Romanticism: Impact on Social Thought*). Philosophically, it emerged as a response to the rationalist leanings of the Enlightenment, and politically, against the conceptions of abstract freedom in individualist liberalism. Unlike cultural determinism, however, which has a strongly oversocialized conception of the human being and thus hardly any concept of self at all, romanticism emphasizes agency and creativity in the process of self-formation and self-realization. Human beings are seen in their singularity, but the relation to others is an essential and inescapable part of their understanding of their own selves.

Charles Taylor's inquiry into *The Sources of the Self* is a most recent and forceful restatement of this conception. Taylor starts out from the familiar argument that the advent of modernity indicates that common frameworks for moral evaluation can no longer be presumed to exist. The key subsequent contention is then that the ability and inclination to question any existing such framework of meaning does not lead into a sustainable position that would hold that no such frameworks are needed at all, a view he calls the 'naturalist supposition' (Taylor 1989, p. 30). If such frameworks of meaning are what gives human beings identity, allows them to orient themselves in social and moral space, then they are not 'things we invent' and may as well not invent. They need to be seen as 'answers to questions which inescapably pre-exist for us,' or, in other words, 'they are contestable answers to inescapable questions' (Taylor 1989, pp. 38, 40). Taylor develops here the contours of a concept of inescapability as part of a moral-social philosophy of selfhood under conditions of modernity. Meaning-centered conceptions of selfhood have recently been underlined as a basis of communitarian positions in moral and political philosophy, such as Sandel's concept (1982) of the 'encumbered self.'

4. *Selfhood and Otherness*

Arguably, these two concepts of selfhood remain within the frame of a debate in which the under- and the oversocialized views of the human being occupy the extreme points. The introduction of interaction and intersubjectivity in self-formation has created intermediate theoretical positions, and, possibly more importantly, they have allowed different accentuations of selfhood without making positions mutually incompatible. Nevertheless, the modernity-oriented view still emphasizes a self that acts upon the world, whereas the meaning-oriented view underlines the fact that the self is provided the sense of their actions by the world of which they are a part. Since the mid-1970s, in contrast, theoretical and empirical developments have tended to break up the existing two-dimensional mode of conceptualization.

On the one hand, the notion of a 'decentering of the subject' has been proposed mainly from poststructuralist discussions. On the other hand, the observation of both multiple and changing basic orientations in human beings has led to the proliferation of the—infelicitously coined—term 'postmodern identity' as a new form of selfhood. In both cases, a strong concept of self has been abandoned, in the one case on the basis of philosophical reflection, in the other, grounded on empirical observation. Both the ideas of a 'decentring' and of a 'postmodern self' question some major implications of the more standard sociological view of selfhood, namely the existence of the human self as a unit and its persistence as the 'same' self over time.

In the former perspective, the philosophical maxim of thinking identity and difference as one double concept, rather than as one concept opposed to another, is translated into sociological thinking as the need to think self and other as a relation, rather than as a subject and a context or object. Such notions rather underline the nonidentitarian character of being by pointing to the issue of 'the other in me' (Emmanuel Lévinas) as a question of co-constitution rather than of interaction. While emphasized in recent poststructuralist thought, similar ideas can be found, for instance, in Arendt's (1978, pp. 183–7) insistence on the 'two-in-one,' on the relation to oneself as another, as the very precondition for thought. In a broad sense, they go back to the ancient view of the friend as an 'other self.' And Theunissen (1977/1965) had already conceptualized self-other relations on the basis of an understanding of 'the Other' as referring to all 'concepts through which philosophy interprets the presence and present of the fellow/co-human being (*Mitmensch*) or of the transcendental original form of this human being.'

In one particular formulation, Cavell's reflections provide an example for a thinking about selfhood that does not presuppose an idea of identity, coherence, or consistency (Cavell 1989, 1990). He cautions against 'any fixed, metaphysical interpretation of the idea of a self' and against the idea of 'a noumenal self as one's "true self" and of this entity as having desires and requiring expression.' In contrast, Cavell suggests that the 'idea of the self must be such that it can contain, let us say, an intuition of partial compliance with its idea

of itself, hence of distance from itself' or, in other words, he advocates the idea of 'the unattained but attainable self' (Cavell, 1990, pp. 31, 34). Cavell proposes here a relation of the unattained and the attainable as constitutive for the self, that is, he makes the very question of attainability a central feature of a theory of selfhood.

5. *Selfhood and Sociohistorical Transformations*

The development of this threefold semantic space has considerably enriched the conceptualization and analysis of selfhood in the social sciences during the twentieth century. If one considers the three fields of inquiry as explorations of the various dimensions of selfhood, rather than as mutually exclusive perspectives, the question of the constitution of the human self emerges as a *problématique* that concerns all human beings and for which there may be a variety of processes which can be determined generally only in their forms but not in their results. As a consequence, however, it becomes much more difficult to conclude from the prevailing character of selfhood on social structure and political order. This has often been seen as a 'weakness' of symbolic interactionism, for instance, which Mead's view of the self is said to have inaugurated, as a social theory that allegedly cannot address issues of societal constitution. But by the same move Mead allows for and recognizes a plurality of selves that has returned to the center of discussion today—after the renewal of a regressive synthesis of identity and society in Talcott Parsons's work (to which—what is often overlooked—Erik Erikson contributed as well). The question of the relation between selfhood and society and politics thus needs to be rephrased as a question of comparative-historical sociology rather than merely of social theory.

Some indications as to how this relation should be understood can be found when reading the historical sociology of twentieth century societies in this light. The social upheaval during the second half of the nineteenth century with industrialization, urbanization and the phrasing of 'the social question' is often seen as a first 'modern' uprooting of established forms of selfhood, as a first massive process of 'disembedding' (Giddens 1990). The development towards so-called mass societies during the first half of the twentieth century lets the question of the relation between individuation and growth of the self emerge. Totalitarianism has been analyzed in terms of an imbalance between imposed individuation and delayed self-formation, the result having been the tendency towards 'escape from freedom' and into stable collective identities (Fromm 1941, Arendt 1958). The first three decades after the Second World War are then considered as a form of re-embedding of selves into the institutional frames of democratic welfare societies. Most recently, the indications of dissolution and dismantling of the rather comprehensive set of social institutions of the interventionist welfare state are one of the reasons to focus sociological debate again on questions of selfhood and identity. During this period, as some contributions argue, no longer continuity and coherence but transience, instability, and inclination to change are said to be marks of the important life orientations of contemporary human beings (see Shotter and Gergen 1989, Lash and Friedman 1992, Kellner 1995). However, many of these analyses are challenged on grounds of the limited representativity of the empirical sample or on conceptual grounds.

Given the theoretical insights described above as the emergence of the threefold semantic space of selfhood, any attempt to offer a full-scale reformulation of the issue of the varieties of selfhood in different sociohistorical configurations would be adventurous. Forms of self-constitution and substantive orientations of human selves are just likely to be highly variable across large populations with varieties of experiences and sociocultural positions. However, the notion of recurring crises of modernity and the identification of historically distinct processes of disembedding and re-embedding could be the basis for a socially more specific analysis of the formation and stability of social selves (Wagner 1994).

The social existence of the 'modern' idea that human beings construct their selves is what many societies have in common throughout the 1800s and 1900s. As such, it does not give any guidance in defining different configurations. Therefore, three qualifying criteria have been introduced. First, the existence of the idea of construction of selfhood still leaves open the question whether all human beings living in a given social context share it and are affected by it. The social permeation of the idea may be limited. Second, human beings in the process of constructing their selves may consider this as a matter of choice, as a truly modernist perspective would have it. In many circumstances, however, though a knowledge and a sense of the fact of social construction prevails, it may appear to human beings as almost natural, as in some sense pre-given or ascribed, which social self they are going to have. Third, the stability of any self one has chosen may vary. Such a construction of selfhood may be considered a once-in-a-lifetime occurrence, but may also be regarded as less committing and, for instance, open to reconsideration and change at a later age. These criteria widen the scope of constructability of selfhood. All conditions of construction have existed for some individuals or groups at any time during the past two centuries in the West. But the widening of the scope of construction may mark the transitions from one to another social configuration of modernity. These transitions entail social processes of disembedding and provoke transformations of social selves, in the course of which not only other selves are acquired but the possibility of construction is also more widely perceived.

See also: Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Identity and Identification: Philosophical Aspects; Identity in Anthropology; Identity: Social; Interactionism and Personality; Modernity; Person and Self: Philosophical Aspects; Personality and Conceptions of the Self; Self: Philosophical Aspects

Bibliography

Arendt H 1958 *The Origins of Totalitarianism*, 2nd edn. World Publishing Company, Cleveland, OH, and Meridian Books, New York

Arendt H 1978 *The Life of the Mind, Vol. 1: Thinking*. Harcourt, Brace, Jovanovich, New York

Cavell S 1989 *The New yet Unapproachable America. Lectures After Emerson After Wittgenstein*. Living Batch Press, Albuquerque, NM

Cavell S 1990 *Conditions Handsome and Unhandsome: The Constitution of Emersonian Perfectionism*. The University of Chicago Press, Chicago

Friese H 2001 Identity: Desire, name and difference. In: Friese H (ed.) *Identities*. Berghahn, Oxford, UK

Fromm E 1941 *Escape from Freedom*. Holt, Rinehart, and Winston, New York

Giddens A 1990 *The Consequences of Modernity*. Polity, Cambridge

Hollis M 1985 Of masks and men. In: Carrithers M, Collins S, Lukes S (eds.) *The Category of the Person: Anthropology, Philosophy, History*. Cambridge University Press, Cambridge, UK and New York

Joas H 1996 *The Creativity of Action* [trans. by Gaines J, Keast P]. Polity, Cambridge

Joas H 1998 The autonomy of the self. The meadian heritage and its postmodern challenge. *European Journal of Social Theory* **1**(1): 7–18

Kellner D 1995 *Media Culture, Cultural Studies, Identity and Politics Between the Modern and the Postmodern*. Routledge, London and New York

Lash S, Friedman J (eds.) 1992 *Modernity and Identity*. Blackwell, Oxford, UK

Rorty R 1989 *Contingency, Irony, and Solidarity*. Cambridge University Press, Cambridge, UK

Sandel M J 1982 *Liberalism and the Limits of Justice*. Cambridge University Press, Cambridge, UK and New York

Shotter J, Gergen K J (eds.) 1989 *Texts of Identity*. Sage, London

Straub J 2001 Personal and collective identity. A conceptual analysis. In: Friese H (ed.) *Identities*. Berghahn, Oxford, UK

Taylor C 1989 *Sources of the Self. The Making of Modern Identity*. Harvard University Press, Cambridge, MA

Theunissen M 1977/1965 *Der Andere*. de Gruyter, Berlin

Wagner P 1994 *A Sociology of Modernity: Liberty and Discipline*. Routledge, London and New York

Wrong D H 1961 The oversocialized conception of man in modern sociology. *American Sociological Review* **26**: 183–93

P. Wagner

Self-knowledge: Philosophical Aspects

'Self-knowledge'—here understood as the knowledge a creature has of its own mental states, processes, and dispositions—has come to be, along with consciousness, an increasingly compelling topic for philosophers and scientists investigating the structure and development of various cognitive capacities. Unlike consciousness, however, the capacity for self-knowledge is generally held to be the exclusive privilege of human beings—or, at least, of creatures with the kind of conceptual capacity that comes with having a language of a certain rich complexity. Prima facie, self-knowledge involves more than simply having sensations (or sensory experiences). Many say it involves even more than having moods, emotions, and 'first-order' intentional states, paradigmatically beliefs and desires, but also hopes, fears, wishes, occurrent thoughts, and so on. Some philosophers disagree, claiming that first-order intentional states (individuated by 'propositional content') are themselves properly restricted to language-using creatures and, by that token, to self-knowing creatures. Nevertheless, it is generally agreed that self-knowledge involves, minimally, the capacity for a conceptually rich awareness or understanding of one's own mental states and processes, and hence the capacity, as it is often put, to form 'second-order' beliefs about these, a capacity that can be extended to, or perhaps only comes with, forming beliefs about the mental states and processes of others.

More generally, then, self-knowledge involves the capacity for adopting what Daniel Dennett calls the *intentional stance* (Dennett 1987)—understanding oneself (and others) as having experiences and ways of representing the world that shape agential behavior. But this capacity should not be understood in purely spectatorial terms. The more sophisticated a creature's intentional understanding, the more sophisticated it will be at acting in a world partially constituted by agential activity. Thus, the capacity for intentional understanding continually plays into, and potentially modifies, an agent's own intentionally directed activity, a point I will return to below.

1. Empirical Findings: Testing the Limits of Intentional Understanding

A number of experiments have been designed to test for the presence of such a capacity in nonhuman animals (primarily hominids) and very young children, with interestingly controversial results. Some investigators claim that the evidence for intentional understanding (including a fairly sophisticated awareness of self) is persuasively strong in the case of certain primates (chimpanzees and bonobos, though for a deflationary interpretation of this evidence based on

further empirical trials, see Povinelli and Eddy (1996)). Experiments with young children (most famously the 'false-belief task') seem clearly to indicate that intentional understanding comes in degrees, passing through fairly discernible stages. How precisely to characterize these stages and what constitutes the mechanism of change is hotly debated (Astington et al. 1988), yet there is an emerging pattern in the empirical data that is noteworthy for standard philosophical discussions of self-knowledge. First, under certain, not very unusual conditions, young children will make mistakes in attributing basic mental states to themselves (desires and beliefs) that are inconceivable from the adult point of view because of the direct first-personal awareness adults seem to have of these states. Second, the errors children make in self-attribution mirror in kind and frequency the errors they make in attributing mental states to others; there seems to be no dramatic asymmetry between self-knowledge and knowledge of other minds, at least for children (Flavell et al. 1995, Gopnik 1993). Moreover, though children become increasingly more sophisticated in their intentional self- and other-attributions, this symmetry and systematicness in kind and frequency of error seems to persist in notable ways into adulthood, with subjects 'confabulating' explanations for their actions and feelings that are purportedly based on direct introspective awareness of their own minds (Nisbett and Wilson 1977). To psychologists investigating these phenomena, this pattern of results suggests that, for both children and adults, judgments about mental states and processes are affected (perhaps unconsciously) by background beliefs about how such states are caused and how they in turn cause behavior, and these beliefs can be systematically wrong, at least in detail. There is nothing special about the capacity for self-knowledge that avoids the error-causing beliefs embedded in our 'folk-psychology.'

2. *The Philosophical Project: Explaining the Special Character of Self-knowledge*

Although judgments about our own mental states and processes are clearly not infallible, there are distinctive features of self-knowledge that require explanation. These features, and the interrelations among them, are variously presented in the philosophical literature, but they are generally held to consist in the following (Wright et al. 1998, pp. 1–2).

(a) *Immediacy*—the knowledge we have of our own sensations, emotions, and intentional states is normally not based on behavioral or contextual evidence; it is groundless or immediate, seemingly involving some kind of 'special access' to our own mind.

(b) *Saliency (or, transparency)*—knowledge of our own minds, especially with regard to sensations and occurrent thoughts, is unavoidable; if we have a particular thought or sensation, we typically know that we have it (such mental states are 'self-intimating').

(c) *(First-person) authority*—judgments about our own minds (expressed in sincere first-person claims) are typically treated by others as correct simply by virtue of our having made them. Normally, such claims require no justification and are treated as true, so long as there is no overwhelmingly defeating counter-evidence to suggest that we are wrong about ourselves.

The traditional account against which most discussions of self-knowledge react is drawn from René Descartes (Descartes 1911). Viewing the mind as a non-physical substance directly knowable only from the first-person point of view, Descartes had a ready explanation of first-person authority: we clearly and distinctly *perceive* the contents of our own minds, which are by their nature transparent to us. Alternatively, since we have no special perceptual access to the minds of others, the only way we can make judgments about them is by observing their external bodily movements and inferring from their similarities to ours that they are likewise caused by mental states and processes. Such weakly inferred third-person judgments are perforce dramatically less secure than first-person judgments, which are themselves infallible. On this view, the commonly acknowledged asymmetry between first- and third-person judgments is exaggerated in both directions.

This Cartesian account has been discredited in almost everyone's estimation—and not just empirically, but conceptually. Even the picturesque metaphor of the 'mind and its contents' was held by Gilbert Ryle to encourage such bad habits of thought that nothing worthwhile could be said in its defense. The 'metaphysical iron curtain' it introduced between knowledge of self and knowledge of others was, in Ryle's view, sufficient to show the bankruptcy of the Cartesian picture by making impossible our everyday practices of correctly applying, and being able to correct our applications of, 'mental conduct' concepts either to others or to ourselves. Since this would undermine the possibility of a coherent language containing such concepts, it undermines the possibility of our conceptualizing ourselves intentionally at all. Thus, so far from providing an explanation of our capacity for self-knowledge, Descartes' theory discredits it. We find out about ourselves in much the same way that we find out about other people, Ryle is famous for saying: 'A residual difference in the supplies of the requisite data makes some differences in degree between what I can know about myself and what I can know about you, but these differences are not all in favor of self-knowledge' (Ryle 1949, p. 155). Although Ryle exhaustively illustrated his view that the ways we find out about ourselves are as various as the kinds of things we find out, thus debarring any simple philosophical analysis of self- and other-knowledge, he is

generally thought to have promoted the doctrine of 'logical behaviorism' (not Ryle's term), according to which all talk about so-called 'mental' states and processes (our own or anyone else's) is really just talk about actual and potential displays of overt behavior. In fact, though clearly suspicious of philosophical defenses of first-person authority based on a principle epistemic privilege, Ryle is more profitably read as following Wittgenstein in trying to reveal how the 'logical (or 'grammatical) structure' of a pervasive way of thinking about the 'mind and its contents' inevitably generates self-defeating explanations of various phenomena, including our capacity for self-knowledge. Any substantial account of such phenomena must therefore exemplify a different kind of logical structure altogether (see discussions of Wittgenstein's view in McDowell 1991, Wright 1991).

3. *Current Proposals: 'Causal–Perceptual' vs. 'Constitutive' Explanations of Self-knowledge*

This proposed opposition between different 'logical types' of explanation is given more substance in contemporary work with philosophers pursuing two quite different methodological approaches to understanding the purported immediacy, saliency, and authority of self-knowledge. The first 'causal–perceptual' approach focuses on the mechanics of self-knowledge, including the nature of the states we are purportedly making judgments about, the nature of the states that constitute those judgments, and the nature of the connection between them. The second so-called 'constitutive' approach focuses on the larger implications of our capacity for self-knowledge, including what role it plays in specifying conditions for the possibility of our powers and practices of agency. Adopting either one of these approaches does not rule out addressing the questions featured by the alternative approach, although the sorts of answers given are materially shaped by which questions occupy center field.

3.1 *Causal–Perceptual Accounts of Self-knowledge*

According to some philosophers, the basic defects of the Cartesian view stem not from the idea that first-person knowledge is a form of inner perception, but from its dualistic metaphysics and an overly restricted conception of what perception involves. On Descartes's 'act-object' model of perception, mental states are objects before the mind's 'inner eye' which we perceive (as we do external objects) by the act of discerning their intrinsic (nonrelational) properties. Without mounting any radical 'grammatical' critique, problems with this view can be readily identified, especially once the mind is understood to be part of the material world (see Shoemaker 1996, Wright et al. 1998).

(a) There is no organ of inner perception analogous to the eye (but we do have a capacity for sensing the position of our bodies, by proprioception, and this does not require any specialized organ; perhaps introspection is like proprioception).

(b) Sensations could possibly count as 'objects' of inner perception, but intentional states (beliefs and desires) are less plausible candidates. First, phenomenologically, we do not seem to be aware of our intentional states as discernibly perceptible in the way that sensations are; there is no perceptual experience that precedes, or comes with, our judgments about such states—no 'qualia.' Second, intentional states are individuated by their propositional content, and their propositional content is plausibly determined by (causal) relational features these states bear to the subject's physical or social environment, rather than intrinsic features of the states themselves. If the 'externalist' view of content is right, then authoritative self-knowledge of these states cannot be based on inwardly perceiving their intrinsic properties (see Ludlow and Martin 1998).

(c) In contrast with external perception, there are no mediating perceptual experiences between the object perceived and judgments based on those perceivings. If mental states and processes are objects of inner perception, they must be 'self-intimating' objects, where the object of perception is itself a perceptual experience of that object. Taken literally, this idea is hard to make coherent, though some have found it invitingly appropriate in the case of sensations as a way of marking what is so unusually special and mysterious about them.

Many of these problems (and more) are satisfactorily addressed by adopting a 'tracking' model of inner perception (Armstrong 1968). We do have 'special access' to our own minds in so far as the second-order beliefs we form about our (ontologically distinct) first-order states and processes are *reliably caused*, perhaps *subcognitively*, by the very first-order states and processes such beliefs are about. There need be no distinctive phenomenology that characterizes such access, and no need therefore to give different accounts of our self-knowledge of intentional states versus sensations, imaginings, and other phenomenologically differentiated phenomena. Of course, if all such states are functionally defined—that is, dispositionally, in terms of their causal relations to other mental states, to perceptual inputs and to behavioral outputs—there is nothing about them that makes them in principle accessible only to the person whose states they are. But this is a strength of the approach. Such states are accessible in a *special way* to the subject (by normally and reliably causing the requisite second-order beliefs), thereby accounting for the immediacy, saliency, and authority of our first-person judgments.

Yet because causal mechanisms can break down in particular cases, first-person error may be expected (as in self-deception). Such errors may also be detected: since first-order states are constitutively related to a subject's linguistic and nonlinguistic behavior, significant evidence of noncohesiveness between the subject's sayings and doings may be used to override her self-ascriptions. Moreover, there is no inconsistency between this account of the special character of self-knowledge and psychologists' findings of systematic error in self- and other attribution. It may be just as they suggest: since the concepts in terms of which we express our second-order beliefs (about self and other) are functionally defined in terms of folk-psychological theory, systematic error may result from the theory's deficiencies.

3.2 Constitutive Accounts of Self-knowledge

The causal–perceptual approach to self-knowledge has many advantages, including, it would seem, the ontological and conceptual *independence* of first-order states from the second-order beliefs that track them in self-knowers. For this means there is no conceptual bar on attributing beliefs, desires, and so forth to creatures that lack the capacity for self-knowledge but still behave rationally enough to be well predicted from the intentional stance. Why then do some philosophers insist that having such first-order intentional states is constitutively linked to having authoritative self-knowledge—and, therefore, presumably, to a subject's forming second order beliefs about her first order states for which 'truth is the default condition' (Wright 1991)?

To begin with a negative answer: if causal–perceptualists are right in saying that the subject's status as an authoritative self-knower is merely contingently dependent on a reliable causal mechanism linking first- and second-order states, then we could not account for the special role self-knowledge plays in constituting that subject as a rational and responsible agent (Bilgrami 1998, Burge 1996, Shoemaker 1996). How so? Consider, for instance, the causal–perceptualist account of error in self-attribution: The subject forms a false second-order belief due to a breakdown in the causal mechanism normally linking that belief to the appropriate first-order state. Her lapse of rationality consists in her saying one thing (because of the second-order belief) and doing another (because of her first-order states). But why count this a lapse of *rationality* (as surely we do), rather than simply a self-misperception, analogous (except for the causal pathway) to misperceptions of others? What imperative is there for a subject's self-attributions to line up with her other sayings and doings in order for her to be a coherent subject, responsible and responsive to the norms embedded in our folk-psychological conception of rational agency?

Different authors take various tacks in answering this question, but there is a discernible common thread. An agent cannot act linguistically or nonlinguistically in ways that respond self-consciously to these norms, unless she knows in a directive sense what she is doing—i.e., unless she recognizes herself as acting in accord with certain intentional states that she knows to be hers, and knows in a way that acknowledges her own agency in producing and maintaining them.

Moreover, an agent is self-consciously responsive to norms that govern her own intentional states (such as *believing that p*), not just by recognizing what counts as norm-compliant behavior, but actually by *regulating* herself (her judgments, thoughts, and actions) in accord with such norms. On this picture, if a person's sincere claims about herself do not cohere with the (other) things she does and says, we may say she fails to know herself. But, just as importantly, we say that she acts unknowingly. That is, there is some sense (perhaps pathological) in which she acts, but without authoring her own acts as a free and responsible agent. Hence, she does not *authorize* such acts; they are out of her reflective control. To fail in self-knowledge on this account is not to be wrong in the passively spectatorial sense that one can be wrong about others; it is to fail more actively in exercising one's powers of agency.

Constitutive explanations of self-knowledge display a very different logical structure from causal–perceptual accounts. Consequently, they suggest a different way of conceptualizing the intentional capacities of self-knowers, a way that has become needlessly complicated through retaining the traditional architecture of first-order states and second-order beliefs about them that are ontologically, if not conceptually, distinct and that supposedly underlie and get expressed in first-person claims (McGeer 1996). This distinction makes sense on a causal–perceptualist account of self-knowledge; but it only confuses and obscures on the constitutive account, leading, for instance, to inappropriate descriptions of what distinguishes us from other animals. The constitutive explanation of self-knowledge focuses on what it means to be an intentional creature of a certain sort, an intentional creature that is capable of actively regulating her own intentional states, revising and maintaining them in self-conscious recognition of various norms. No doubt there are automatic mechanisms also involved in regulating intentional states in an environmentally useful way, mechanisms we may well share with nonlinguistic creatures and which explain why adopting the intentional stance towards them works as well as it does (Shoemaker 1996, essay 11). But in our own case, we increase the range and power of the intentional stance by understanding what rational, responsible agency requires of us and molding ourselves to suit. The apparatus of first- and second-order states may play some useful role in further elucidating this capacity for intentional self-regulation, but more likely

it encourages bad habits of thought—a lingering vestige of the dying, but never quite dead, Cartesian tradition in epistemology and philosophy of mind.

See also: Culture and the Self (Implications for Psychological Theory): Cultural Concerns; Person and Self: Philosophical Aspects; Self-development in Childhood; Self: History of the Concept; Self: Philosophical Aspects

Bibliography

Armstrong D M 1968 *A Materialist Theory of the Mind*. Routledge and Kegan Paul, London

Astington J W, Harris P L, Olson D R (eds.) 1988 *Developing Theories of Mind*. Cambridge University Press, Cambridge, UK

Bilgrami A 1988 Self-knowledge and resentment. In: Wright C, Smith B C, Macdonald C (eds.) *Knowing our Own Minds*. Clarendon Press, Oxford, UK

Burge T 1996 Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society* **96**: 91–111

Dennett D C 1987 *The Intentional Stance*. MIT Press, Cambridge, MA

Descartes R 1911 Meditations on first philosophy. In: Haldane E S, Ross G R T (eds.) *The Philosophical Works of Descartes*. Cambridge University Press, Cambridge, UK, Vol. 1, pp. 131–200

Flavell J H, Green F L, Flavell E R 1995 *Young Children's Knowledge About Thinking*. University of Chicago Press, Chicago

Gopnik A 1993 How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* **16**: 1–14

Ludlow P, Martin N (eds.) 1998 *Externalism and Self-knowledge*. Cambridge University Press, Cambridge, UK

McDowell J 1991 Intentionality and interiority in Wittgenstein. In: Puhl K (ed.) *Meaning Scepticism*. W. de Gruyter, Berlin, pp. 168–9

McGeer V 1996 Is 'Self-knowledge' an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy* **93**: 483–515

Nisbett R E, Wilson T D 1977 Telling more than we can know: Verbal reports on mental processes. *Psychological Review* **84**(3): 231–59

Povinelli D J, Eddy T J 1996 What young chimpanzees know about seeing. *Monographs of the Society for Research in Child Development* **61**(3): 1–190

Ryle G 1949 *The Concept of Mind*. Hutchinson's University Library, London

Shoemaker S 1996 *The First Person Perspective and Other Essays*. Cambridge University Press, Cambridge, UK

Wright C 1991 Wittgenstein's later philosophy of mind: Sensation, privacy and intention. In: Puhl K (ed.) *Meaning Scepticism*. W. de Gruyter, Berlin, pp. 126–47

Wright C, Smith B C, Macdonald C (eds.) 1998 *Knowing Our Own Minds*. Clarendon Press, Oxford, UK

V. McGeer

Self-monitoring, Psychology of

According to the theory of self-monitoring, people differ in the extent to which they monitor (i.e., observe and control) their expressive behavior and self-presentation (Snyder 1974, 1987). Individuals high in self-monitoring are thought to regulate their expressive self-presentation for the sake of public appearances, and thus be highly responsive to social and interpersonal cues to situationally appropriate performances. Individuals low in self-monitoring are thought to lack either the ability or the motivation to regulate their expressive self-presentations for such purposes. Their expressive behaviors are thought instead to reflect their own inner states and dispositions, including their attitudes, emotions, self-conceptions, and traits of personality.

Research on self-monitoring typically has employed multi-item, self-report measures to identify people high and low in self-monitoring. The two most frequently employed measuring instruments are the 25 true–false items of the original Self-monitoring Scale (Snyder 1974) and an 18-item refinement of this measure (Gangestad and Snyder 1985; see also Lennox and Wolfe 1984). Empirical investigations of testable hypotheses spawned by self-monitoring theory have accumulated into a fairly sizable published literature (for a review of the literature, see Gangestad and Snyder 2000).

1. Major Themes of Self-monitoring Theory and Research

Soon after its inception, and partially in response to critical theoretical issues of the times, self-monitoring was offered as a partial resolution of the 'traits vs. situations' and 'attitudes and behaviors' controversies in personality and social psychology. The propositions of self-monitoring theory suggested that the behavior of low self-monitors ought to be predicted readily from measures of their attitudes, traits, and dispositions whereas that of high self-monitors ought to be best predicted from knowledge of features of the situations in which they operate. Self-monitoring promised a 'moderator variable' resolution to debates concerning the relative roles of person and situation in determining behavior. These issues set the agenda for the first generation of research and theorizing on self-monitoring, designed primarily to document the relatively 'situational' orientation of high self-monitors and the comparatively 'dispositional' orientation of low self-monitors (for a review, see Snyder 1987).

In a second generation of research and theorizing, investigations moved beyond issues of dispositional and situational determination of behavior to examinations of the linkages between self-monitoring and

interpersonal orientations. Perhaps the most prominent of these programs concerns the links between expressive control and interpersonal orientations, as revealed in friendships, romantic relationships, and sexual involvements (e.g., Snyder et al. 1985). Other such programs of research concern advertising, persuasion, and consumer behavior (e.g., Snyder and DeBono 1985), personnel selection (e.g., Snyder et al. 1988), organizational behavior (Caldwell and O'Reilly 1982, Kilduff 1992), socialization and developmental processes (e.g., Eisenberg et al. 1991, Graziano and Waschull 1995), cross-cultural studies (e.g., Gudykunst 1985).

Central themes in these programs of research have been that high self-monitors live in worlds of public appearances created by strategic use of impression management, and that low self-monitors live in worlds revolving around the private realities of their personal identities and the coherent expression of these identities across diverse life domains. Consistent with these themes, research on interpersonal orientations has revealed that high, relative to low, self-monitors choose as activity partners friends who will facilitate the construction of their own situationally-appropriate images and appearances (e.g., Snyder et al. 1983). Perhaps because of their concern with images and appearances, high self-monitors have romantic relationships characterized by less intimacy than those of low self-monitors. Also consistent with these themes, explorations of consumer attitudes and behavior have revealed that high self-monitors value consumer products for their strategic value in cultivating social images and public appearances, reacting positively to advertising appeals that associate products with status and prestige; by contrast, low self-monitors judge consumer products in terms of the quality of the products stripped of their image-creating and status-enhancing veneer, choosing products that they can trust to perform their intended functions well (e.g., DeBono and Packer 1991). These same orientations manifest themselves in the workplace as well, with high self-monitors preferring positions that call for the exercise of their self-presentational skills; thus, for example, high self-monitors perform particularly well in occupations that call for flexibility and adaptiveness in dealings with diverse constituencies (e.g., Caldwell and O'Reilly 1982) whereas low self-monitors appear to function best in dealing with relatively homogeneous work groups.

It should be recognized that, although these programs of research, for the most part, have not grounded their hypotheses or interpretations in self-monitoring's traditionally fertile ground—issues concerning the dispositional vs. situational control of behavior—they do nevertheless reflect the spirit of the self-monitoring construct. That is, their guiding themes represent clear expressions of self-monitoring theory's defining concerns with the worlds of public appearances and social images, and the processes by which appearances and images are constructed and sustained. However, it should also be recognized that these lines of research go beyond showing that individual differences, in concern for cultivating public appearances, affect self-presentational behaviors. These programs of research have demonstrated that these concerns, and their manifestations in expressive control, permeate the very fabric of individuals' lives, affecting their friendship worlds, their romantic lives, their interactions with the consumer marketplace, and their work worlds.

2. *The Nature of Self-monitoring*

Despite their generativity, the self-monitoring construct and its measure have been the subject of considerable controversy over how self-monitoring ought to be interpreted and measured. The roots of this controversy are factor analyses that clearly reveal that the items of the Self-monitoring Scale are multifactorial, with the emergence of three factors being the most familiar product of these factor analyses (Briggs et al. 1980). These factor analyses, and attempts to interpret them, have stimulated a critically important question: Is self-monitoring truly a unitary phenomenon?

Although there is widespread agreement about the multifactorial nature of the items of the Self-monitoring Scale, there exist diverging viewpoints on the interpretation of this state of affairs. One interpretation is that some criterion variables represented in the literature might relate to one factor, other criterion variables to a second independent factor, and yet others to still a third factor—an interpretation which holds that self-monitoring is not a unitary phenomenon (e.g., Briggs and Cheek 1986).

Without disputing the multifactorial nature of the self-monitoring items, it is nevertheless possible to construe self-monitoring as a unitary psychological construct. Taxonomic analyses have revealed that the self-monitoring subscales all tap, to varying degrees, a common latent variable that may reflect two discrete or quasidiscrete self-monitoring classes (Gangestad and Snyder 1985). In addition, the Self-monitoring Scale itself taps a large common factor accounting for variance in its items and correlating, to varying degrees, with its subscales; this general factor approximates the Self-monitoring Scale's first unrotated factor (Snyder and Gangestad 1986). Thus, the Self-monitoring Scale may 'work' to predict diverse phenomena of individual and social functioning because it taps this general factor; this interpretation is congruent with self-monitoring as a unitary, conceptually meaningfully psychological construct.

Although much of the debate about the nature of the self-monitoring construct has focused on con-

trasting interpretations of the internal structure of the Self-monitoring Scale, it is possible to consult another source of evidence with which to address the major issues of the self-monitoring controversy—the literature on the Self-monitoring Scale's relations with criterion variables external to the scale itself (i.e., behavioral, behavioroid, and performance measures of phenomena relevant to self-monitoring theorizing). Based on a quantitative review of the literature on the Self-monitoring Scale's relations with behavioral and behavioroid external criterion variables, it appears that, with some important exceptions, a wide range of external criteria tap a dimension directly measured by the Self-monitoring Scale (Gangestad and Snyder 2000).

Based on this quantitative appraisal of the self-monitoring literature, it is possible to offer some specifications of what self-monitoring is and what it is not, specifications that may guide the next generations of theory and research on self-monitoring. That is, it is possible to identify 'exclusionary messages' about features of self-monitoring theory that should not receive the attention heretofore accorded them (e.g., delimiting the scope of self-monitoring as a moderator variable such that claims about peer-self agreement ought no longer be made, although claims about behavioral variability may yet be made), and to identify 'inclusionary messages' about features that should define the evolving agenda for theory and research on self-monitoring (e.g., focusing on the links between self-monitoring and strategic motivational agendas associated with engaging in, or eschewing, impression management tactics that involve the construction of social appearances and the cultivation of images).

3. *Conclusions*

To some extent, the productivity and generativity of the self-monitoring construct may derive from the fact that it appears to capture one of the fundamental dichotomies of psychology—whether behavior is a product of forces that operate from outside of the individual (exemplified by the 'situational' orientation of the high self-monitor) or whether it is governed by influences that guide from within the individual (typified by the 'dispositional' orientation of the low self-monitor). In theory and research, self-monitoring has served as a focal point for issues in assessment, in the role of scale construction in theory building, and in examining fundamental questions about personality and social behavior, particularly those concerning how individuals incorporate inputs from their own stable and enduring dispositions and inputs from the situational contexts in which they operate into agendas for action that guide their functioning as individuals and as social beings.

See also: Impression Management, Psychology of; Personality and Conceptions of the Self; Self-conscious Emotions, Psychology of; Self-evaluative Process, Psychology of; Self-knowledge: Philosophical Aspects

Bibliography

Briggs S R, Cheek J M 1986 The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality* **54**: 106–48

Briggs S R, Cheek J M, Buss A H 1980 An analysis of the self-monitoring scale. *Journal of Personality and Social Psychology* **38**: 679–86

Caldwell D F, O'Reilly C A 1982 Boundary spanning and individual performance: The impact of self-monitoring. *Journal of Applied Psychology* **67**: 124–27

DeBono K G, Packer M 1991 The effects of advertising appeal on perceptions of product quality. *Personality and Social Psychology Bulletin* **17**: 194–200

Eisenberg N, Fabes R, Schaller M, Carlo G, Miller P 1991 The relation of parental characteristics and practices to children's vicarious emotional responding. *Child Development* **62**: 1393–1408

Gangestad S, Snyder M 1985 'To carve nature at its joints': On the existence of discrete classes in personality. *Psychological Review* **92**: 317–49

Gangestad S, Snyder M 2000 Self-monitoring: appraisal and reappraisal. *Psychological Bulletin* **126**: 530–55

Graziano W G, Waschull S B 1995 Social development and self-monitoring. *Review of Personality and Social Psychology* **15**: 233–60

Gudykunst W B 1985 The influence of cultural similarity, type of relationship, and self-monitoring on uncertainty reduction processes. *Communication Monographs* **52**: 203–17

Kilduff M 1992 The friendship network as a decision-making resource: dispositional moderators of social influences on organization choice. *Journal of Personality and Social Psychology* **62**: 168–80

Lennox R, Wolfe R 1984 Revision of the self-monitoring scale. *Journal of Personality and Social Psychology* **46**: 1349–64

Snyder M 1974 Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology* **30**: 526–37

Snyder M 1987 *Public appearances, Public realities: The Psychology of Self-monitoring*. W. H. Freeman, New York

Snyder M, Berscheid E, Glick P 1985 Focusing on the exterior and the interior: Two investigations of the initiation of personal relationships. *Journal of Personality and Social Psychology* **48**: 1427–39

Snyder M, Berscheid E, Matwychuk A 1988 Orientations toward personnel selection: Differential reliance on appearance and personality. *Journal of Personality and Social Psychology* **54**: 972–9

Snyder M, DeBono K G 1985 Appeals to image and claims about quality: Understanding the psychology of advertising. *Journal of Personality and Social Psychology* **49**: 586–97

Snyder M, Gangestad S 1986 On the nature of self-monitoring: Matters of assessment, matters of validity. *Journal of Personality and Social Psychology* **51**: 125–39

Snyder M, Gangestad S, Simpson J A 1983 Choosing friends as activity partners: The role of self-monitoring. *Journal of Personality and Social Psychology* **45**: 1061–72

M. Snyder

Self-organizing Dynamical Systems

Interactions, nonlinearity, emergence and context, though omnipresent in the social and behavioral sciences, have proved remarkably resistant to understanding. New light may be shed on these problems by virtue of the introduction and development of theoretical concepts and methods of self-organization and the mathematical tools of (nonlinear) dynamical systems (Haken 1996, Kelso 1995). Self-organized dynamics promises both a language for, and a strategy toward, understanding human behavior on multiple levels of description. A key problem on any given level of description is to identify the essential dynamical variables characterizing the formation and change of behavioral patterns so that the pattern dynamics, the rules governing behavior, may be found and their predictions pursued. Dynamics offers not only a concise mathematical description of different kinds and classes of behavior. Dynamical models provide new measures and predict new phenomena often not observed before in studies of individual human and social behavior. And they may help explain effects previously attributed to more static, nondynamical mechanisms. Therein lies the promise of dynamics for the social and behavioral sciences at large. Dynamics offers a new way to think about, perhaps even solve, old problems.

Here, following a brief historical introduction, the main ideas of dynamical systems are described in a nontechnical fashion. The connection between notions of self-organization and nonlinear dynamical systems is then addressed as a foundation for understanding behavioral pattern formation, flexibility, and change. A flavor of the approach, which stresses an intimate relationship between theory and experiment, is addressed in the context of a few examples from the behavioral and social sciences. Finally, some extensions of the approach are discussed briefly, including a way to include meaningful information into the self-organizing dynamics.

1. *A Short History of Dynamical Systems*

For the eighteenth century Scottish philosopher David Hume, all the sciences bore a relation to the human mind. In his *Treatise of Human Nature* (1738), Hume first divided the mind into its contents: ideas and impressions. Then he added dynamics, noting the impossibility of simple ideas forming more complex ones without some bond of union between them. Hume's three dynamical laws for the association of ideas—resemblance, contiguity, and cause and effect—were thought to be responsible for controlling all mental operations. A kind of attraction, Hume thought, existed in the mental world: the motion of the mind was conceived as analogous to the motion of a body, as described earlier by Newton. Mental 'stuff' was governed (somehow) by dynamics.

All this has a strangely contemporary ring to it. Dynamics is not only the language of the natural sciences, but has permeated the social, behavioral, and neurosciences as well (Arbib et al. 1998, Kelso 1995, Vallacher and Novak 1997). Cognitive science, whose theory and paradigms are historically embedded in the language of the digital computer, may be the most recent of the human sciences to fall under the spell of dynamics (Port and van Gelder 1995, Thelen and Smith 1994).

Which factors led to the primacy of dynamics as the language of science, whether natural or human? First, was the insight of the French mathematician Henri Poincaré to put dynamics into geometric form. Poincaré introduced an array of powerful methods for visualizing how things behave, somewhat independent of the things themselves.

Second, following Poincaré's lead, were developments in understanding nonlinear dynamical systems from a mathematical point of view. These are sets of differential equations that can exhibit multiple solutions for the same parameters and are capable of highly irregular, even unpredictable behavior.

Third, was the ability to visualize how these equations behave by integrating them numerically on a digital computer. Parameters can be varied and parameter spaces explored without doing a 'real' experiment or collecting data of any kind. In fact, computer simulation and visualization is often the only possible method of exploring complex behavior.

Finally, were quite specific examples of cooperative (or 'collective') behavior in physics (e.g., fluids, lasers), chemistry (e.g., chemical reactions), biology (e.g., morphogenesis), and later, in the behavioral sciences. The behavioral experiments (Kelso 1984) and consequent theoretical modeling (Haken et al. 1985) showed for the first time that dynamic patterns of human behavior, how behavior forms, stabilizes, and changes, may be described quite precisely in the mathematical language of nonlinear dynamical systems. Not only were new effects predicted and observed, it was possible to derive these emergent patterns and the pattern dynamics by nonlinearly coupling the component parts. Since these early empirical and theoretical developments, ideas of self-organization and dynamics have infiltrated the social, behavioral, economic, and cognitive sciences, not to speak of literature and the arts.

2. Dynamical Systems: Some Terminology

What exactly is a dynamical system? It is not possible here to do justice to the mathematics of dynamical systems and its numerous technical applications, for example, in time series analysis (e.g., Strogatz 1994). For present purposes, only the basics are considered. Fundamentally, a dynamical system pertains to anything—such as the behavior of a human being or a social group—that evolves over time. Mathematically, if $X_1, X_2, \ldots, X_n$ are the variables characterizing the system's behavior, a dynamical system is a system of equations stipulating the temporal evolution of $\vec{X}$, the vector of all permissible X-values. Therein lies the rub: in the social and behavioral sciences one has to find these Xs and identify their dynamics on a chosen level of description.

If $\vec{X}$ is a continuous function of time, then the dynamics of $\vec{X}$ is typically described by a set of first order ordinary differential equations (ODEs). The order of the equation refers to the highest derivative that appears in the equation. Thus, $\dot{X} = F(X)$ is a first order equation where the dot above the X denotes the first derivative with respect to time and $F(X)$ gives the vector field. ODEs are of greatest interest when F is nonlinear because, depending on the particular form of F, a wide variety of behaviors can arise.

Another important class of dynamical system is difference equations or maps. Maps are often useful when the behavior of a system appears to be organized around discrete temporal events. Maps and ODEs provide complementary ways to model and study nonlinear dynamical behavior, but we shall not pursue maps further here (see, however, van Geert 1995 as an example of the approach).

A vectorfield is defined at each and every point of the state or phase space of the system, defined by the values of the variables $X_1, X_2, \ldots, X_n$. As these states evolve in time they can be represented graphically as a trajectory. The state space is filled with trajectories. At any point in a given trajectory, a velocity vector can be determined by differentiation. These velocity vectors prescribe a velocity vectorfield, which is assumed to represent the flow or behavior of the dynamical system (Abraham and Shaw 1982). This flow often has characteristic tendencies particularly in dissipative systems, named thus because their state space volume decreases or dissipates over time toward attractors. In other words, trajectories converge over time (technically, as time goes to infinity) to subsets of the phase space (a given attractor's basin of attraction). One may even think of an attractor as a representation of the goal of the behaving system (Saltzman and Kelso 1987).

The term transient defines the segment of a trajectory starting from some initial condition in the basin of attraction until it settles into an attractor. Attractors can be fixed points, in which all initial conditions converge to some stable rest state. Attractors can be periodic, exhibiting preferred rhythms or orbits on which the system settles regardless of where it starts. Or, there can be so-called strange attractors; strange because they exhibit deterministic chaos, a type of irregular behavior resembling random noise, yet often containing pockets of quite ordered behavior.

The presence of chaos in physical systems is ubiquitous, and there is some evidence suggesting that it may also play an important role in certain biological systems, including the human brain and its disorders (Kelso and Fuchs 1995). Chaos means sensitive dependence on initial conditions: nudging a chaotic system by just a hair can change its future behavior dramatically. In the vernacular, human behavior certainly seems 'chaotic' sometimes. However, although an area of active interest in many research fields, it remains unclear whether chaos, and its close relative, fractals (Mandelbrot 1982) provide any essential insights in the social and behavioral sciences.

What may be rather more important to appreciate is that parameters, so-called control parameters, can move a system through a vast array of different kinds of dynamic behavior of which chaos is only one. Thus, when a parameter changes smoothly, the attractor in general also changes smoothly. In other words, sizeable changes in the input have little or no effect on the resulting output. However, when the control parameter passes through a critical point or threshold in an intrinsically nonlinear system an abrupt change in the attractor can occur. This sensitive dependence on parameters is called a bifurcation in mathematics, or a nonequilibrium phase transition in physical theories of pattern formation (Haken 1977).

3. Concepts of Self-organization

Dynamical theory, by itself, does not give us any essential insights into how patterns of behavior may be created in complex systems that contain multiple (often different) elements interacting nonlinearly with each other and their environment. This is where the concept of self-organization is helpful (Haken 1977, Nicolis and Prigogine 1977).

Self-organization refers to the spontaneous formation of pattern and pattern change in complex systems whose elements adapt to the very patterns of behavior they create. Think of birds flocking, fish schooling, bees swarming. There is no 'self' inside the system ordering the elements, dictating to them what to do and when to do it. Rather, the system, which includes the environment, literally organizes itself. Inevitably, when the elements form a coupled system with the world, coordinated patterns of behavior arise. Emergent pattern reflects the collective behavior of the elements. Collective behavior as a result of self-organization reduces the very large number of degrees of freedom into a much smaller set of relevant

dynamical variables called, appropriately enough, collective variables. The enormous compression of degrees of freedom near critical points can arise because events occur on different timescales: the faster individual elements in the system may become 'enslaved' to the slower, emergent pattern or collective variables, and lose their status as relevant behavioral quantities (Haken 1977). Alternatively, one may conceive of a hierarchy of timescales for various processes underlying human behavior. On a given level of the hierarchy are pattern variables subject to constraints (e.g., of the task) that act as boundary conditions on the pattern dynamics. At the next level down are component processes and events that typically operate on faster timescales. Notice in this scheme (Kelso 1995), the key is to chose a level of description and understand the relation between adjacent levels, not reduce to some 'fundamental' lower level.

So what causes behavioral patterns to form? And what causes pattern change? It is here that the connection between processes of self-organization and dynamical systems theory becomes quite explicit. In complex systems that are open to exchanges of information with their environment, naturally occurring environmental conditions or intrinsic, endogenous factors may take the form of control parameters in a nonlinear dynamical system. When the control parameter crosses a critical value, instability occurs, leading to the formation of new (or different) patterns.

Dynamic instability is the generic mechanism underlying spontaneous self-organized pattern formation and change in all systems coupled to their internal or external environment. The reason is that near instability the individual elements, in order to adjust to current conditions (control parameter values), must order themselves in a new or different way. Fluctuations are always present, constantly testing whether a given behavioral pattern is stable. Fluctuations are not just noise; rather, they allow the system to discover new, more adaptive behavioral patterns.

The patterns that emerge and change near instabilities have a rich and highly nonlinear pattern, or coordination dynamics. Herein lies the basis of the hypothesis that human beings and the basic forms of behavior they produce, may be understood as self-organizing dynamical systems (Kelso 1995). Humans—complex systems consisting of multiple, interacting elements—produce behavioral patterns that are captured by collective variables. The resulting dynamical rules are essentially nonlinear and thus capable of producing a rich repertoire of behaviors. By this hypothesis, all human behavior—even human cognitive development (Magnusson 1995, Thelen and Smith 1994) and learning (Zanone and Kelso 1992) which occur on very different time scales—arises because complex material systems, through the process of self-organization, create a dynamic, pattern forming system that is capable of both behavioral simplicity (such as fixed point behavior) and enormous, even creative, behavioral complexity (such as chaotic behavior). No new principles, it seems, need be introduced (but see Sect. 5).

4. *Finding Dynamical Laws in Behavioral and Social Systems*

In the social and behavioral sciences the key collective variables characterizing the system's state space are seldom known *a priori* and have to be identified. Science always needs special entry points, places where irrelevant details may be pruned away while retaining the essential aspects of the behavior one is trying to understand. Unique to the perspective of self-organized dynamical systems is its focus on qualitative change, places where behavior bifurcates. The reason is that qualitative change affords a clear distinction between one pattern and another, thereby enabling one to identify the key pattern variables that define behavioral states.

Likewise, any parameter that induces qualitative behavioral change qualifies as a control parameter. The payoff from knowing collective pattern variables and control parameters is high: they enable one to obtain the dynamical rules of behavior on a chosen level of description. By adopting the same strategy the next level down, the individual component dynamics may be studied and identified. It is the interaction between these that creates the patterns at the level above, thereby building a bridge across levels of description.

5. *The Laws that Bind Us*

It has long been recognized that social relationships are an emergent product of the process of interaction. Commonly studied relationships are the mother–infant interaction, the marriage bond and the patient-therapist relation, all of which are extremely complicated and difficult to understand. A more direct approach may be to study simpler kinds of social coordination, with the aim of determining whether such emergent behavior is self-organized, and if so what its dynamics are.

As an illustrative example, consider two people coordinating their movements. In this particular task each individual is instructed to oscillate a limb (the lower leg in this case) in the same or opposite direction to that of the other (Schmidt et al. 1990). Obviously the two people must watch each other in order to do the task. Then, either by an instruction from the experimenter or by following a metronome whose rate systematically increases, the social dyad also must speed up their movements.

When moving their legs up and down in the same direction, the two members of the dyad can remain

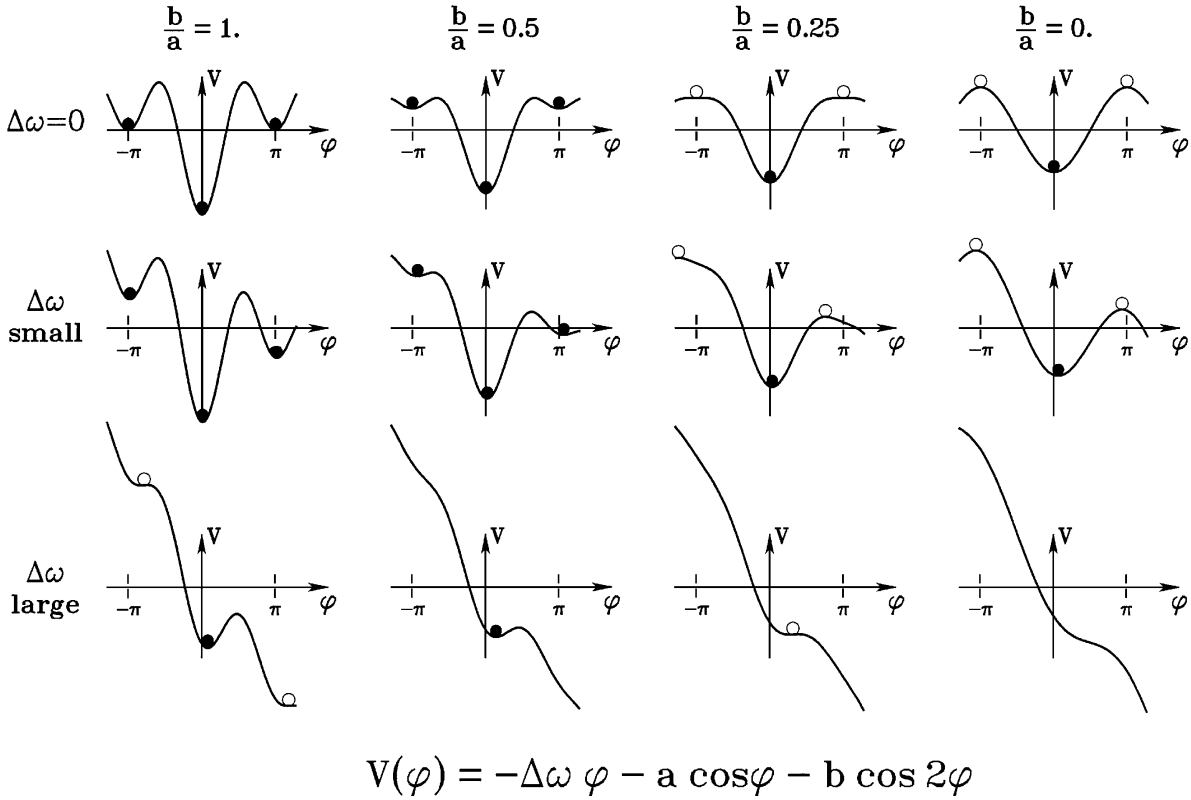

Figure 1
The potential, V(ϕ) of Eqn. (2) (with $\Delta\omega = 0$) and Eqn. (4) (with $\Delta\omega \neq 0$). Black balls symbolize stable coordinated behaviors and white balls correspond to unstable behavioral states (see text for details)

synchronized throughout a broad range of speeds. However, when moving their legs in the opposite direction (one person's leg extending at the knee while the other's is flexing), such is not the case. Instead, at certain critical speeds participants spontaneously change their behavior so that the legs now move in the same direction.

How might these social coordination phenomena be explained? The pattern variable that changes qualitatively at the transition is the relative phase, ϕ. When the two people move in the same direction, they are in-phase with each other, $\phi = 0$. When they move in different directions, their behavior is antiphase ($\phi = \pm\pi$ radians or ± 180 degrees).

The phase relation is a good candidate for a collective variable because it clearly captures the spatiotemporal ordering between the two interacting components. Moreover, ϕ changes far more slowly than the variables that might describe the individual components (position, velocity, acceleration, electromyographic activity of contracting muscles, etc.) Importantly, ϕ changes abruptly at the transition.

The simplest dynamics that captures all the observed facts is

$$\dot{\phi} = -a\sin\phi - 2b\sin 2\phi \tag{1}$$

where ϕ is the relative phase between the movements of the two individuals, $\dot{\phi}$ is the derivative of ϕ with respect to time, and the ratio b/a is a control parameter corresponding to the movement rate in the experiment. An equivalent formulation of Eqn. (1) is

$$\dot{\phi} = -\partial V(\phi)/\partial\phi$$
$$\text{with } V(\phi) = -a\cos\phi - b\cos 2\phi \tag{2}$$

In the literature, this is called the HKB model of coordinated behavior, after Haken et al. (1985) who formulated it as an explanation of the coordination of limb movements within a single person (Kelso 1984). Figure 1 (top) allows one to develop an intuitive understanding of the behavior of Eqn. (2), as well as to connect to the key concepts of stability and instability in self-organizing dynamical systems introduced

earlier. The dynamics can be visualized as a particle moving in a potential function, $V(\phi)$. The minima of the potential are points of vanishing force, giving rise to stable solutions of the HKB dynamics. As long as the speed parameter (b/a) is slow, meaning the cycle period is long, Eqn. (2) has two stable fixed point attractors, collective states at $\phi = 0$ and $\phi = \pm\pi$ rad. Thus, two coordinated behavioral patterns coexist for exactly the same parameter values, the essentially nonlinear feature of bistability. Such bi- and in general multistability is common throughout the behavioral and social sciences. Ambiguous figures (old woman or young girl? Faces or vase?) are well-known examples (e.g., Kruse and Stadler 1995).

As the ratio b/a is decreased, meaning that the cycle period gets shorter as the legs speed up, the formerly stable fixed point at $\phi = \pm\pi$ rad. becomes unstable, and turns into a *repellor*. Any small perturbation will now kick the system into the basin of attraction of the stable fixed point (behavioral pattern) at $\phi = 0$. Notice also that once there, the system's behavior will stay in the in-phase attractor, even if the direction of the control parameter is reversed. This is called hysteresis, a basic form of memory in nonlinear dynamical systems.

What about the individual components making up the social dyad? Research has established that these take the form of self-sustaining oscillators, archetypal of all time-dependent behavior whether regular or not. The particular functional form of the oscillator need not occupy the reader here. More important is the nature of the nonlinear coupling that produces the emergent coordination. The simplest, perhaps fundamental coupling that guarantees all the observed emergent properties of coordinated behavioral patterns—multistability, flexible switching among behavioral states, and primitive memory, is

$$K_{12} = (\dot{X}_1 - \dot{X}_2)\{\alpha + \beta(X_1 - X_2)^2\} \tag{3}$$

where X_1 and X_2 are the individual components and α and β are coupling parameters. Notice that the 'next level up,' the level of behavioral patterns and the dynamical rule that governs them (Eqns.(1) and (2)), can be derived from the level below, the individual components and their nonlinear interaction. One may call this constructive reductionism: by focusing on adjacent levels, under the task constraint of interpersonal coordination, the individual parts can be put together to create the behavior of the whole.

The basic self-organized dynamics, Eqns. (2) and (3) have been extended in numerous ways, only a few of which are mentioned here.

(a) *Critical slowing down* and *enhancement of fluctuations*. Introducing stochastic forces into Eqns. (1) and (2) allows key predictions to be tested and quantitatively evaluated. Critical slowing is easy to understand from Fig. 1 (top). As the minima at $\phi = \pm\pi$ become shallower and shallower, the time it takes to adjust to a small perturbation takes longer and longer. Thus, the local relaxation time is predicted to increase as the instability is approached because the restoring force (given as the gradient in the potential) becomes smaller. Likewise, the variability of ϕ is predicted to increase due to the flattening of the potential near the transition point. Both predictions have been confirmed in a wide variety of experimental systems, including recordings of the human brain.

(b) *Symmetry breaking*. Notice that Eqns. (1) and (2) are symmetric: the dynamical system is 2π periodic and is identical under left-right reflection (ϕ is the same as $-\phi$). This assumes that the individual components are identical, which is seldom, if ever the case. Nature thrives on broken symmetry. To accommodate this fact, a term $\Delta\omega$ is incorporated into the dynamics

$$\dot{\phi} = \Delta\omega - a\sin\phi - 2b\sin 2\phi,$$

and

$$V(\phi) = -\Delta\omega\phi - a\cos\phi - b\cos 2\phi \tag{4}$$

for the equation of motion and the potential respectively. Small values of $\Delta\omega$ shift the attractive fixed points (Fig. 1 middle) in an adaptive manner. For larger values of $\Delta\omega$ the attractors disappear entirely (Fig. 1 bottom) causing the relative phase to drift: no coordination between the components occurs. Note, however, that the dynamics still retain some curvature (Fig. 1 bottom right): even though there are no attractors there is still *attraction* to where the attractors used to be. The reason is that the difference ($\Delta\omega$) between the individual components is sufficiently large that they do their own thing, while still retaining a tendency to cooperate. This is how global integration, in which the component parts are locked together, is reconciled with the tendency of the parts to function locally, as individual autonomous units (Kelso 1995).

(c) *Information: a new principle?* Unlike the behavior of inanimate things, the self-organizing dynamics of human behavior is fundamentally informational, though not in the standard sense of data communicated across a channel (Shannon and Weaver 1949). Rather, collective variables are context-dependent and intrinsically meaningful. Context-dependence does not imply lack of reproducibility. Nor does it mean that every new context requires a new collective variable or order parameter. As noted above, within- and between-person coordinated behaviors are described by exactly the same self-organizing dynamics. One of the consequences of identifying the latter is that in order to modify or change the system's behavior, any new information (say a task to be learned, an intention to change one's behavior) must be expressed in terms of parameters acting on system-relevant collective variables. The benefit of identifying the latter is that one knows what to modify. Likewise, the collective variable dynamics—prior to the introduction of new information—influences how that information is used. Thus, information is not lying out there as mere data: information is meaningful

to the extent that it modifies, and is modified by, the collective variable dynamics.

(d) *Generalization*. The basic dynamics (Eqns. (1–4)) can readily be elaborated as a model of emergent coordinated behavior among many anatomically different components. Self-organized behavioral patterns such as singing in a group or making a 'wave' during a football game are common, yet unstudied examples. Recently, Neda et al. (2000) have examined a simpler group activity: applause in theater and opera audiences in Romania and Hungary. After an exceptional performance, initially thunderous incoherent clapping gives way to slower, synchronized clapping. Measurements indicate that the clapping period suddenly doubles at the onset of the synchronized phase, and slowly decreases as synchronization is lost. This pattern is a cultural phenomenon in many parts of Europe: a collective request for an encore. Increasing frequency (decreasing period) is a measure of the urgency of the message, and culminates in the transition back to noise when the performers reappear. These results are readily explained by a model of a group of globally coupled nonlinear oscillators (Kuramoto 1984)

$$\frac{\mathrm{d}\phi_k}{\mathrm{d}t} = \omega_k + \frac{K}{N}\sum_{j=1}^{N} \sin(\phi_j - \phi_k) \tag{5}$$

in which a critical coupling parameter, K_c determines the different modes of clapping behavior. K is a function of the dispersion (D) of clapping frequencies

$$K_c = \sqrt{\frac{2}{\pi^3}}D \tag{6}$$

During fast clapping, synchronization is not possible due to the large dispersion of clapping frequencies. Slower, synchronized clapping at double the period arises when small dispersion appears. Period doubling rhythmic applause tends not to occur in big open-air concerts where the informational coupling among the audience is small. K can also be societally imposed. In Eastern European communities during communist times, synchronization was seldom destroyed because enthusiasm was often low for the 'great leader's' speech. For people in the West, the cultural information content of different clapping patterns may be quite different. Regardless, the mathematical descriptions for coordinated behavior—of social dyads and the psychology of the crowd—are remarkably similar.

6. Conclusion and Future Directions

The theoretical concepts and methods of self-organizing dynamics are likely to play an ever greater role in the social, behavioral, and cognitive sciences, especially as the interactions among disciplines continues to grow. Up to now, the use of nonlinear dynamics is still quite restricted, and largely metaphorical. One reason is that the tools are difficult to learn, and require a degree of mathematical sophistication. Their implementation in real systems is nontrivial, requiring a different approach to experimentation and observation. Another reason is that the dynamical perspective is often cast in opposition to more conventional theoretical approaches, instead of as an aid to understanding. The former tends to emphasize decentralization, collective decision-making and cooperative behavior among many interacting elements. The latter tends to focus on individual psychological processes such as intention, perception, attention, memory, and so forth. Yet there is increasing evidence that intending, perceiving, attending, deciding, emoting, and remembering have a dynamics as well. The language of dynamics serves to bridge individual and group processes. In each case, dynamics must be filled with content, with key variables and parameters obtained for the system under study. Every system is different, but what we learn about one may aid in understanding another. What may be most important are the principles and insights gained when human brains and human behavior are seen in the light of self-organizing dynamics.

See also: Chaos Theory; Computational Psycholinguistics; Dynamic Decision Making; Emergent Properties; Evolution: Self-organization Theory; Hume, David (1711–76); Neural Systems and Behavior: Dynamical Systems Approaches; Stochastic Dynamic Models (Choice, Response, and Time)

Bibliography

Abraham R H, Shaw C D 1982 *Dynamics: The Geometry of Behavior*. Ariel Press, Santa Cruz, CA

Arbib M A, Erdi P, Szentagothai J 1998 *Neural Organization: Structure, Function and Dynamics*. MIT Press, Cambridge, MA

Haken H 1977 *Synergetics, an Introduction: Non-equilibrium Phase Transitions and Self-organization in Physics, Chemistry and Biology*. Springer, Berlin

Haken H 1996 *Principles of Brain Functioning*. Springer, Berlin

Haken H, Kelso J A S, Bunz H 1985 A theoretical model of phase transitions in human hand movements. *Biological Cybernetics* **51**: 347–56

Kelso J A S 1984 Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology: Regulatory, Integrative and Comparative* **15**: R1000–4

Kelso J A S 1995 *Dynamic Patterns: The Self-organization of Brain and Behavior*. MIT Press, Cambridge, MA

Kelso J A S, Fuchs A 1995 Self-organizing dynamics of the human brain: Critical instabilities and Sil'nikov chaos. *Chaos* **5**(1): 64–9

Kuramoto Y 1984 *Chemical Oscillations, Waves and Turbulence*. Springer-Verlag, Berlin

Magnusson D 1995 Individual development: A holistic, integrated model. In: Moen P, Elder G H Jr, Lüscher K (eds.)

Examining Lives in Context. American Psychological Association, Washington, DC
Mandelbrot B 1982 *The Fractal Geometry of Nature*. Freeman, New York
Neda Z, Ravasz E, Brechet Y, Vicsek T, Barabasi A L 2000 The sound of many hands clapping. *Nature* **403**: 849–50
Nicolis G, Prigogine I 1977 *Self-organization in Nonequilibrium Systems*. Wiley, New York
Port R F, van Gelder T 1995 *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, MA
Saltzman E L, Kelso J A S 1987 Skilled actions: A task dynamic approach. *Psychological Review* **94**: 84–106
Schmidt R C, Carello C, Turvey M T 1990 Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance* **16**: 227–47
Shannon C E, Weaver W 1949 *The Mathematical Theory of Communication*. University of Illinois Press, Chicago
Strogatz S H 1994 *Nonlinear Dynamics and Chaos*. Addison-Wesley, Reading, MA
Thelen E, Smith L B 1994 *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA
Vallacher R R, Nowak A 1997 The emergence of dynamical social psychology. *Psychological Inquiry* **8**: 73–99
Van Geert P 1995 Learning and development: A semantic and mathematical analysis. *Human Development* **38**: 123–45
Zanone P G, Kelso J A S 1992 The evolution of behavioral attractors with learning: Nonequilibrium phase transitions. *Journal of Experimental Psychology: Human Perception and Performance* **18/2**: 403–21

J. A. S. Kelso

Self: Philosophical Aspects

There are two related ways in which philosophical reflection on the self may usefully be brought to bear in social science. One concerns the traditional problem about personal identity. The other concerns the distinctive kinds of social relations into which only selves can enter. Sustained reflection on the second issue reveals striking analogies between some of the social relations into which only selves can enter and certain first personal relations that only a self can bear to itself. These analogies entail two possibilities that bear on the first issue concerning personal identity, namely, the possibility that there could be group selves who are composed of many human beings and the possibility that there could be multiple selves within a single human being.

Section 1 outlines the classical problem of personal identity that was originally posed by Locke and continues to generate philosophical debate. Section 2 discusses some of the distinctive social relations into which only selves can enter. Section 3 draws several analogies between such social relations and first personal relations. Section 4 shows how these analogies point to the possibilities of group and multiple selves. It closes with some cautionary remarks about how overlooking these two possibilities can lead to confusions about methodological individualism.

1. The Philosophical Problem about Personal Identity

To have a self is to be self-conscious in roughly the sense that Locke took to be the defining mark of the person. In his words, a person is 'a thinking, intelligent being, that has reason and reflection, and can think itself as itself in different times and places' (Locke 1979). In general, all of these capabilities—for thought, reason, reflection, and reflexive self-reference—tend to be lumped together by philosophers under the common heading of self-consciousness. Except in some generous accounts of animal intelligence, it is generally assumed that the only self-conscious beings are human beings.

Locke famously disagreed. This was not because he believed that there are other species of animal besides the human species that actually qualify as persons (though he did discuss at some length a certain parrot that was reported to have remarkable conversational abilities). It was rather because he thought that the condition of personal identity should not be understood in biological terms at all; it should be understood, rather, in phenomenological terms. A person's identity extends, as he put it, exactly as far as its consciousness extends. He saw no reason why the life of a person so construed, as a persisting center of consciousness, should necessarily coincide with the biological lifespan of a given human animal. In defense of this distinction between personal and animal identity he offered the following thought experiment: first, imagine that the consciousnesses of a prince and a cobbler are switched each into the other's body and, then, consider who would be who thereafter. He thought it intuitively obvious that the identity of the prince is tied to his consciousness in such a way that he would remain the same person—would remain himself—even after having gained a new, cobbling body (similar remarks apply to the cobbler).

Many objections have been raised against Locke's argument and conclusion. It still remains a hotly disputed matter among philosophers whether he was right to distinguish the identity of the person from the identity of the human animal (Parfit 1984, Perry 1975, Rorty 1979). This philosophical dispute has not drawn much attention from social scientists who have generally assumed against Locke that the life of each individual self or person coincides with the life of a particular human being. The anti-Lockean assumption is so pervasive in the social sciences that it serves as common ground even among those who take

opposing positions on the issue concerning whether all social facts are reducible to facts about individuals or whether there are irreducibly social facts (see *Individualism versus Collectivism: Philosophical Aspects*; *Methodological Individualism: Philosophical Aspects*)). Both sides take for granted that the only individuals in question are human beings.

Prima facie, it seems reasonable that social scientists should disregard Locke's distinction between the self or person and the human being. For something that neither Locke nor neo-Lockeans have ever done is offer an *actual* instance of his distinction. All they have argued for is the *possibility* that a person's consciousness could be transferred from one human body to another and, in every case, they have resorted to thought experiments in their attempts to show this. There is a significant difference between Locke's original thought experiment about the prince and the subsequent neo-Lockean experiments. Whereas his simply stipulates that a transfer of consciousness has occurred, the latter typically describe some particular process by which the transfer of consciousness is supposed to be accomplished—as in, for example, brain transplantation or brain reprogramming. However, although this might make the neo-Lockean thought experiments seem more plausible, there is no realistic expectation that what they ask us to imagine will ever actually come to pass (Wilkes 1988). Because this is so, they have no particular bearing on the empirical concerns of social science.

Nevertheless, it is important to be clear about whether the anti-Lockean assumption that informs social science is correct. For the way in which we conceive the boundaries that mark one individual self off from another will inevitably determine how we conceive the domain of social relations that is the common subject of all the social sciences. So we need to be absolutely clear about whether it is right to assume, against Locke, that the identity of the individual is at bottom a biological fact.

As the next two sections will show, there are reasons to side with Locke. But these reasons differ from Locke's own in several respects. Unlike his, they are not derived from thought experiments. Furthermore, they support a somewhat different version of his distinction between the identity of the self or person and the human being. Finally, they invite the expectation that this version of his distinction can be realized in fact and not just in imagination—which makes it relevant to social science after all.

2. *The Sociality of Selves: Rational Relations*

Starting roughly with Hegel, various philosophers have mounted arguments to the effect that self-consciousness is a social achievement that simply cannot be possessed by beings who stand outside of all social relations (Wittgenstein 1953, Davidson 1984). The common conclusion of these arguments is overwhelmingly plausible. But even if the conclusion were shown to be false, it would remain true that many self-conscious beings are social and, furthermore, their social relations reflect their self-consciousness in interesting ways. In order to gain a full appreciation of this fact, we must shift our attention away from the phenomenological dimension of self-consciousness that Locke emphasized in his analysis of personal identity, and shift our attention to the other dimension of self-consciousness that he emphasized in his definition of the person as a 'being with reason and reflection.'

One need not cite Locke as an authority in order to see that reflective rationality is just as crucial to selfhood as its phenomenological dimension. For it obviously will not do to define the self just in phenomenological terms, as a center of consciousness, and this is so even if Locke was right to analyze the identity of the self or person in such terms. After all, some sort of center of consciousness is presupposed by all sentience, even in the case of animals that clearly do not have selves, such as wallabies or cockatoos. It also will not do to suppose that selfhood is in place whenever the sort of consciousness that goes together with sentience is accompanied by the mere capacity to refer reflexively to oneself. For, assuming that wallabies and cockatoos can represent things (which some philosophers deny (Davidson 1984)), it is very likely that some of their representations will function in ways that are characteristic of self-representation. Consider, for example, mental maps in which such animals represent their spatial relations to the various objects they perceive; such maps require a way to mark the distinction between their own bodies and other bodies, a distinction which is not inaptly characterized as the distinction between self and other (Evans 1982). Self-consciousness in the sense that goes together with selfhood involves a much more sophisticated sort of self-representation that incorporates a conception of oneself as a thinking thing and, along with it, the concept of thought itself. Armed with this concept, a self-conscious being cannot only represent itself; it can self-ascribe particular thoughts and, in doing so, it can bring to bear on them the critical perspective of reflective rationality.

Whenever selves in this full sense enter into social relations, they enter into relations of reciprocal recognition. Each is self-conscious; each conceives itself as an object of the other's recognition; each conceives the other as so conceiving itself (i.e., as an object of the other's recognition); each conceives both as having a common nature and, in so doing, each conceives both as so conceiving both (i.e., as having a common nature). Since this reciprocally recognized common nature is social as well as rational, it affords the possibility of the social exercise of rational capacities. Thus, corresponding to the capacity to ascribe thoughts to oneself, there is the capacity to ascribe

thoughts to others; corresponding to the capacity to critically evaluate one's own thoughts there is the capacity to critically evaluate the thoughts of others; and together with all this comes another extraordinary social capacity, the capacity for specifically rational influence in which selves aim to move one another by appealing to the normative force of reasons. Such rational influence is attempted in virtually every kind of social engagement that takes place among selves, including: engaging in ordinary conversation and argument; offering bribes; lodging threats; bargaining; cooperating; collaborating. In all of these kinds of engagement reasons are offered up in one guise or another in the hope that their normative force will move someone else.

3. *Analogies Between First Personal and Social Relations*

It is because selves have a rational nature that analogies arise between their first personal relations to their own selves and their social relations to one another. Here are three examples.

3.1 *Critical Evaluation*

It is a platitude that the principles of rationality are both impersonal and universal. They lay down normative requirements that *anyone* must meet in order to be rational, and they constitute the standards by which *everyone* evaluates their own and everyone else's rational performances. It follows that when we evaluate our own thoughts by these standards, we are really considering them as if they were anyone's—which is to say, as if they were not our own. We sometimes describe this aspect of critical self-evaluation with the phrase 'critical distance.' There is a specific action by which we accomplish this distancing: we temporarily suspend our commitment to act on our thoughts until after we have determined whether keeping the commitment would be in accord with the normative requirements of rationality. If we determine that it would not be in accord with those requirements, then the temporary suspension of our commitment becomes permanent, and our critical distance on our thoughts gives way to literally disowning them. But before getting all the way to the point of disowning our thoughts we must first have gotten to a point where our relations to our thoughts are rather like our relations to the thoughts of others. For, whether we decide to disown the thoughts under critical scrutiny or to reclaim them, the normative standards on the basis of which we do so are necessarily conceived by us as impersonal and universal. In consequence, there is a straightforward sense in which we are doing the same thing when we apply the standards to our own thoughts and when we apply them to the thoughts of others. There is this difference, though: whereas self-evaluation involves distancing ourselves from our own thoughts, evaluating the thoughts of others involves a movement in the opposite direction. We must project ourselves into their points of view so that we can appeal to their own sense of how the principles of rationality apply in their specific cases. This sort of criticism that involves projection is sometimes called 'internal criticism,' in contrast with the sort of 'external criticism' that proceeds without taking into account whether its target will either grasp it or be moved by it (Williams). This difference between internal criticism of others and self-criticism—the difference between projection and distancing—does nothing to undermine the analogies between them. On the contrary, for by these opposite movements two distinct selves can bring themselves to the very same place. That is, when one self projects itself into another's point of view for the purposes of internal criticism, it can reach the very same place that the other gets to by distancing itself from its own thoughts for the purposes of self-criticism. Moreover, there is a sense in which what the two selves do from that place is not merely analogous, but the very same thing. Each of the two selves brings to bear the same critical perspective of rationality on the same particular self's thoughts. The only difference that remains is that in the one case it is a social act directed at another, whereas in the other case it is a self-directed act.

3.2 *Joint Activities*

Just as internal criticism of another replicates what individual selves do when they critically evaluate themselves, so also, joint activities carried out by many selves may replicate the activities of an individual self. For the purposes of such activities, distinct selves need to look for areas of overlap between (or among) their points of view, such as common ends, common beliefs about their common ends, and common beliefs about how they might act together so as to realize those ends. These areas of commonly recognized overlap can then serve as the basis from which the distinct selves deliberate and act together and, as such, the areas of overlap will serve as a common point of view which is shared by those selves. Whenever distinct selves do deliberate together from such a common point of view, they will be guided by the very same principles that ordinarily govern individual deliberation, such as consistency, closure (the principle that one should accept the implications of one's attitudes), and transitivity of preferences. All such principles of rationality taken together define what it is for an individual to be fully or optimally rational. For this reason, they generally apply separately to each individual but not to groups of individuals. (Thus, different parties to a

dispute might be rational even though they disagree, but they cannot be rational if they believe contradictions.) The one exception is joint activities in which such groups of individuals deliberate together from a common point of view. To take a limited example, you and I might deliberate individually about a philosophical problem or we might do so together. If we do it individually, each of us will aim to be consistent in our reasoning because that is what the principles of rationality require of us individually. But these principles do not require us to do this together; which is to say, you and I may disagree about many things without compromising our ability to pursue our individual deliberations from our separate points of view in a completely rational manner. What would be compromised, however, is our ability to work on a philosophical problem together. We cannot do that without resolving at least some of our disagreements, namely, those that bear on the problem. This means that if we truly are committed to working on the problem together, then we should recognize a sort of conditional requirement of rationality, a requirement to achieve as much consistency between us as would be necessary to carry out our joint project. Although we needn't achieve overall consistency together in order to do this—that is, we needn't agree on every issue in every domain—nevertheless, we must resolve our disagreements about the relevant philosophical matters. When we resolve these disagreements we will be doing in degree the very same thing that individuals do when they respond to the rational requirement of consistency. The same holds for the other principles of rationality such as closure and transitivity of preferences. For, in general, whenever distinct individuals pursue ends together, they temporarily cease deliberating and acting separately from their individual points of view in order to act and deliberate together from their common point of view. They will also strive to achieve by social means within a group—albeit in diminished degree—the very sorts of rational relations and outcomes that are characteristic of the individual self (Korsgaard 1989).

3.3 Rationality over Time

Although there is some sense in which individual points of view are extended in time, the deliberations that proceed from such points of view always take place in time from the perspective of the present. The question immediately arises, why should an individual take its future attitudes into account when it deliberates from the perspective of the present? The question becomes poignant as soon as we concede that an individual's attitudes may change significantly over time. Thus, consider Parfit's Russian nobleman who believes as a young man that he should give away all of his money to the poor, but who also believes that he will become much more conservative when he is older and will then regret having given the money away. What should he do? Parfit answers as follows: if he recognizes any normative pressure toward an altruistic stance from which he is bound to take the desires of others into account, then he must extend that stance to his own future desires as well—regardless of the fact that he now disapproves of them (Parfit 1984). Other philosophers have also insisted on parity between the normative demands of prudence and altruism. Among moral philosophers, the parity is often taken in the positive way that Parfit does, as showing that we have reason to be both prudent and altruistic (Sidgewick 1981, Nagel 1970). But some philosophers have taken it in a negative way, as showing that we lack reason to be either. According to them, it follows from the fact that deliberation always proceeds from the perspective of the present that rationality does not require us to take our own future attitudes into account any more than it requires us to take the attitudes of others into account. If they are right, then the rational relations that hold within an individual over time are comparable to social relations within a group (Elster 1979, Levi 1985). However, anyone who is struck by this last suggestion—that an individual over time is, rationally speaking, analogous to a group—would do well to bear in mind what the first two analogies brought out, which is that the social exercise of rational capacities within a group, can, in any case, replicate individual rationality.

4. Group Selves and Multiple Selves

It might seem implausible that the normative analogies between first personal and social relations could ever jeopardize the metaphysical distinction between them, which is grounded in whatever facts constitute the identity and distinctness of selves. According to Locke these facts are phenomenological; according to his opponents, the facts are biological. In both cases, the metaphysical facts on offer would clearly be adequate to ground the distinction between first personal and social relations, and this is so no matter how deep and pervasive the analogies between these two sorts of relations might be at the normative level.

Yet there is a way of thinking about the self that calls this into question. It is the very way of thinking that predominated in Locke's initial definition of the person as a being with 'reason and reflection.' When we emphasize this rational dimension of the self, we will find it natural to use a normative criterion in order to individuate selves. According to this normative criterion, we have one self wherever we find something that is subject to the normative requirements that define individual rationality. There is some disagreement among philosophers about exactly what these requirements are. But, even so, we can say this much

with confidence: it is a sufficient condition for being a rational individual who is subject to the requirements of rationality that the individual have some conception of what those requirements are and is in some degree responsive to them—in the sense of at least trying to meet them when it deliberates and accepting criticism when it fails to meet them (Bilgrami 1998). Of course, such criticism might issue from without as well as within, since a rational individual in this sense would certainly have the conceptual resources with which to comprehend efforts by others to wield rational influence over it. This ensures that there is a readily available public criterion by which such rational individuals can be identified for social purposes: there is one such individual wherever there is an interlocutor whom we can rationally engage. It must be admitted that, due to the analogies between first personal and social relations, the boundaries between such individuals will not always be clear. Yet this does not mean that the normative conception of the self is inadequate or unworkable. It only means that the conception has unfamiliar and, indeed, revisionist consequences. For, if we press the analogies to their logical limit, it follows that there can be group selves comprising many human beings and, also, multiple selves within a single human being (Rovane 1998).

Take the group case first. We have seen that, for the purposes of joint activities, distinct selves must find areas of overlap between their individual points of view which can then serve as a common point of view from which they deliberate and act together. We have also seen that such joint deliberations replicate individual rationality in degree. Now, suppose that this were carried to the following extreme: a group of human beings resolved to engage only in joint activities; accordingly, they resolved to rank all of their preferences about such joint activities together and to pool all of their information; and, finally, they resolved always to deliberate jointly in the light of their jointly ranked preferences and pooled information. The overlap between their points of view would then be so complete that they no longer had separate points of view at all. They would still have separate bodies and separate centers of consciousness. But, despite their biological and phenomenological separation, they would nevertheless all share a single rational point of view from which they deliberated and acted *as one*. Consequently, they could also be rationally engaged in the ways that only individual selves can be engaged. There is good reason to regard such a group as a *bona fide* individual in its own right (Korsgaard 1989, Rovane 1998, Scruton 1989).

Now take the multiple case. Recall that it might be rational for an individual to disregard its future attitudes when it deliberates about what to do in the present, in just the way that it might be rational (though possibly immoral) to disregard the attitudes of others. If the normative demands of rationality are thus confined to one's own present attitudes, then one's future self will be no more bound by one's present deliberations than one is presently bound to take one's future attitudes into account. In other words, one's future self will be free to disregard any long-term intentions or commitments that one is thinking of undertaking now. The point is not that there are no such things as long-term intentions or commitments. The point is that when one makes them one does not thereby bind one's future self in any causal or metaphysical sense. It is always up to one's future self to comply or not as it wishes, and one could not alter this aspect of one's situation except by undermining one's future agency altogether. Although this may seem alarming, it does not follow that individual selves do not endure over time. All that follows is that the unity of an individual self over time is constituted by the same sorts of facts that constitute the unity of a group self. Just as the latter requires shared commitments to joint endeavors so too does the former, the only difference being that the latter involves sharing across bodies while the former involves sharing across time. This serves to bring out that the metaphysical facts in terms of which Locke and his opponents analyze the identity of the self over time are, in a sense, normatively impotent. The fact that there is a persisting center of consciousness, or a persisting animal body, does not suffice to bind the self together rationally over time. And, arguably (though this is more counterintuitive), such facts do not even suffice to bind the self together rationally in the present either. That is why we are able to understand the phenomenon of dissociative identity disorder, formerly labeled multiple personality disorder. Human beings who suffer from this disorder exhibit more than point of view, each of which can be rationally engaged on its own. Since none of these points of view encompasses all of the human host's attitudes, there is no possibility of rationally engaging the whole human being. At any given time one can engage only one part of it or another. Each of these parts functions more or less as a separate self, as is shown precisely by the susceptibility of each to rational engagement on its own. Although these multiple selves cannot be rationally engaged simultaneously, there is good reason to suppose that they nevertheless exist simultaneously. While one such self is talking, its companions are typically observing and thinking on their own (this generally comes out in subsequent conversation with them). Furthermore, these multiple selves who coexist simultaneously within the same body are sometimes 'co-conscious' as well—that is, they sometimes have direct phenomenological access to one another's thoughts. But their consciousness of one another's thoughts does not put them into the normative first person relation that selves bear to their own thoughts; for they do not take one another's thoughts as a basis for their deliberations and actions. Thus neither their shared body nor their shared consciousness suffices to bind multiple personalities together in the sense that is

required by the normative conception of the self. On that conception, they qualify as individual selves in their own rights (Braude 1995, Rovane 1998, Wilkes 1988).

Between these two extreme cases of group and multiple selves, the normative conception provides for a wide spectrum of cases, exhibiting differing degrees of rational unity within and among human beings. The conception says that whenever there is enough rational unity—though it is hard to say how much is enough—there is a rational being with its own point of view who can be rationally engaged as such and who, therefore, qualifies as an individual self in its own right.

5. *Implications for Methodological Individualism*

It is not the case that all of the facts about the Columbia Philosophy Department can be reduced to facts about its members atomistically described. Nor is it the case that all of the facts about that department can be explained by appealing to facts about its members atomistically described. Why? Because sometimes the members of the department stop reasoning from their own separate human-size points of view and begin to reason together from the point of view of the department's distinctive projects, needs, opportunities and so on. When the members of the department do this, they give over a portion of their human lives to the life of the department in a quite literal sense, so that the department itself can deliberate and act from a point of view that cannot be equated with any of their individual points of view. Outside the context of this article, this claim about the Columbia Philosophy Department might naturally be taken to contradict methodological individualism. This is because it is generally assumed that the only individuals there are or can be human individuals. That is what leads many to interpret the facts about the Columbia Philosophy Department—insofar as they cannot be properly described or explained by appealing to facts about its individual human members atomistically described—as irreducibly social phenomena. However, this interpretation fails to take account of the way in which the department instantiates the rational properties that are characteristic of the individual self (vs. a social whole).

See also: Person and Self: Philosophical Aspects; Personality and Conceptions of the Self; Self: History of the Concept; Social Identity, Psychology of

Bibliography

Bilgrami A 1998 Self-knowledge and resentment. In: Wright C et al. (eds.) *Knowing Our Own Minds*. Oxford University Press, Oxford, UK

Braude S E 1995 *First Person Plural: Multiple Personality Disorder and the Philosophy of Mind*. Rowman Littlefield, Lanham, MD

Davidson D 1984 *Inquires into Truth and Interpretation*. Oxford University Press, New York

Elster J 1979 *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge University Press, New York

Evans G 1982 *Varieties of Reference*. Oxford University Press, New York

Gilbert M 1989 *On Social Facts*. Routledge, London

Korsgaard C 1989 Personal identity and the unity of agency: A Kantian response to parfit. *Philosophy and Public Affairs* **28**(2)

Levi I 1985 *Hard Choices*. Cambridge University Press, New York

Locke J 1979. In: Niddich P (ed.) *An Essay Concerning Human Understanding*. Oxford University Press, Oxford, UK

Nagel T 1970 *The Possibility of Altruism*. Clarendon, Oxford, UK

Parfit D 1984 *Reasons and Persons*. Clarendon, Oxford, UK

Perry J (ed.) 1975 *Personal Identity*. University of California Press, Berkeley, CA

Rorty A O (ed.) 1979 *The Identities of Perons*. University of California Press, Berkeley, CA

Rovane C 1998 *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton University Press, Princeton, NJ

Scruton R 1989 Corporate persons. *Proceedings of the Aristotelian Society* **63**

Sidgewick H 1981 *The Methods of Ethics*. Hackett, Indianapolis, IN

Wilkes K V 1988 *Real People: Personal Identity Without Thought Experiments*. Oxford University Press, New York

Wittgenstein L 1953 *Philosophical Investigations*. Macmillan, New York

C. Rovane

Self-regulated Learning

Self-regulated learning refers to how students become masters of their own learning processes. Neither a mental ability nor a performance skill, self-regulation is instead the self-directive process through which learners transform their mental abilities into task-related skills in diverse areas of functioning, such as academia, sport, music, and health. This article will define self-regulated learning and describe the intellectual context in which the construct emerged, changes in researchers' emphasis over time as well as current emphases, methodological issues related to the construct, and directions for future research.

1. *Defining Self-regulated Learning*

Self-regulated learning involves metacognitive, motivational, and behavioral processes that are personally initiated to acquire knowledge and skill, such as goal setting, planning, learning strategies, self-reinforcement, self-recording, and self-instruction. A self-regulated learning perspective shifts the focus of educational analyses from students' learning abilities and instructional environments as fixed entities to students' self-initiated processes for improving their

methods and environments for learning. This approach views learning as an activity that students do for themselves in a proactive way, rather than as a covert event that happens to them reactively as a result of teaching experiences. Self-regulated learning theory and research is not limited to asocial forms of education, such as discovery learning, self-education through reading, studying, programmed instruction, or computer assisted instruction, but can include social forms of learning such as modeling, guidance, and feedback from peers, coaches, and teachers. The key issue defining learning as self-regulated is not whether it is socially isolated but rather whether the learner displays personal initiative, perseverence, and adaptive skill in pursuing it. Most contemporary self-regulation theorists have avoided dualistic distinctions between internal and external control of learning and have envisioned self-regulation in broader, more interactive terms. Students can self-regulate their learning not only through covert cognitive means but also through overt behavioral means, such as selecting, modifying, or constructing advantageous personal environments or seeking social support. A learner's sense of self is not limited to individualized forms of learning but includes self-coordinated collective forms of learning in which personal outcomes are achieved through the actions of others, such as family members, team-mates, or friends, or through use of physical environment resources, such as tools. Thus, covert self-regulatory processes are viewed as reciprocally interdependent with behavioral, social, and environmental self-regulatory processes.

Self-regulation is defined as a variable process rather than as a personal attribute that is either present or absent. Even the most helpless learners attempt to control their functioning, but the quality and consistency (i.e., quantity) of their processes are low. Novice learners rely typically on naive forms of self-regulatory processes, such as setting nonspecific distal goals, using nonstrategic methods, inaccurate forms of self-monitoring, attributions to uncontrollable sources of causation, and defensive self-reactions. By contrast, expert learners display powerful forms of self-regulatory processes, especially during the initial phase of learning. Student efforts to self-regulate their learning has been analyzed in terms of three cyclical learning phases. Forethought phase processes anticipate efforts to learn and include self-motivational beliefs, such as self-efficacy, outcome expectations, intrinsic interest, as well as task analysis skills, such as planning, goal setting, and strategy choice. Performance phase processes seek to optimize learning efforts and include use of time management, imagery, self-verbalization, and self-observation processes. Self-reflection phase processes follow efforts to learn and provide understanding of the personal implication of outcomes. They include self-judgment processes, such as self-evaluation and attributions, and self-reactive processes, such as self-satisfaction and adaptive/defensive inferences. Because novice learners fail to use effective forethought processes proactively, such as proximal goal setting and powerful learning strategies, they must rely on reactive processes occurring after learning attempts that often have been unsuccessful. Such unfortunate experiences will trigger negative self-evaluations, self-dissatisfaction, and defensive self-reflections—all of which undermine self-motivation necessary to continue cyclical efforts to learn. By understanding self-regulation in this cyclical interactive way, qualitative as well as quantitative differences in process can be identified for intervention.

2. *Intellectual Context for Self-regulated Learning Research*

Interest in students' self-regulated learning as a formal topic emerged during the 1970s and early 1980s out of general efforts to study human self-control. Promising investigations of children's use of self-regulatory processes like goal setting, self-reinforcement, self-recording, and self-instruction, in such areas of personal control as eating and task completion prompted educational researchers and reformers to consider their use by students during academic learning. Interest in self-regulation of learning was also stimulated by awareness of the limitations of prior efforts to improve achievement that stressed the role of mental ability, social environmental background of students, or qualitative standards of schools. Each of these reform movements viewed students as playing primarily a reactive rather than a proactive role in their own development. In contrast to prior reformers who focused on how educators should adapt instruction to students based on their mental ability, sociocultural background, or achievement of educational standards, self-regulation theorists focused on how students could proactively initiate or substantially supplement experiences designed to educate themselves.

Interest in self-regulation of learning emerged from many theoretical sources during the 1970s and 1980s. For example, operant researchers adapted the principles and technology of B. F. Skinner for personal use, especially the use of environmental control, self-recording, and self-reinforcement. Their preference for the use of single-subject research paradigms and time series data was especially useful for individuals seeking greater self-regulation of learning. During this same time period, phenomenological researchers shifted from monolithic, global views of how self-perceptions influenced learning to hierarchical, domain-specific views and began developing new self-concept tests to assess functioning in specific academic domains. The research of Hattie Marsh, Shavelson, and others was especially influential in gaining new currency for the role of academic domain self-concepts in learning. During this era, social learning researchers, such as Bandura, Schunk, and Zimmerman,

shifted their emphasis from modeling to self-regulation and renamed the approach as social cognitive. They identified a new task-specific motive for learning—self-efficacy belief, and they linked it empirically to other social cognitive processes, such as goal setting, self-observation, self-judgment, and self-reaction. During this era, researchers such as Corno, Gollwitzer, Heckhausen, Kuhls, and others, resurrected volitional notions of self-regulation to explain human efforts to pursue courses of learning in the face of competing events. In their view, self-regulatory control of action can be undermined by ruminating, extrinsic focusing, and vacillating—which interfere with the formation and implementation of an intention.

Also during the 1970s and 1980s, suppressed writings of Vygotsky were published in English that explained more fully how children's inner speech emerges from social interactions and serves as a source of self-control. Cognitive behaviorists, such as Meichenbaum, developed models of internalization training on the basis of Vygotsky's description of how overt speech becomes self-directive. During this same era, cognitive constructivists shifted their interest from cognitive stages to metacognition and the use of learning strategies to explain self-regulated efforts to learn. The research and theory of Flavell played a major role in effecting this transition by describing self-regulation in terms of metacognitive knowledge, self-monitoring, and control of learning.

Research on self-regulation was also influenced by the emergence of goal theories during the 1970s and 1980s. Locke and Lathan showed that setting specific, proximal, challenging but attainable goals greatly influenced the effectiveness of learners' efforts to learn. Theorists such as Ames, Dweck, Maehr, Midgley, and Nicholls identified individual differences in goal orientations that affected students' efforts to learn on their own. These researchers found that learning or mastery goal orientations facilitated persistence and effort during self-directed efforts to learn, whereas performance or ego goal orientations curtailed motivation and achievement. During this same period, another perspective emerged that focused on the role of intrinsic interest in learning. Deci, Harackiewicz, Lepper, Ryan, Vallerand, and others demonstrated that perceptions of personal control, competence, or interest in a task were predictive of intrinsic efforts to learn on one's own. This focus on intrinsic motivation was accompanied by a resurgence of research on the role of various forms of interest in self-directed learning by a host of scholars in Europe as well as elsewhere, such as Eccles, Hidi, Krapp, Renninger, Schiefele, Wigfield, and others. During these same decades, research on self-regulation of learning was influenced by cybernetic conceptions of how information is processed. Researchers such as Carver, Scheier, and Winne demonstrated the important role of executive processes, such as goal setting and self-monitoring, and feedback control loops in self-directed efforts to learn.

3. *Changes in Emphasis Over Time*

Before the 1980s, researchers focused on the impact of separate self-regulatory processes, such as goal setting, self-efficacy, self-instruction, volition, strategy learning, and self-management with little consideration for their broader implications regarding student learning of academic subject matter. Interest in the latter topic began to coalesce in the mid-1980s with the publication of journal articles describing various types of self-regulated learning processes, good learning strategy users, self-efficacious learners, and metacognitive engagement, among other topics (Jan Simons and Beukhof 1987, Zimmerman 1986). By 1991, a variety of theories and some initial research on self-regulated learning and academic achievement was published in special journal articles and edited textbooks (Maehr and Pintrich 1991, Zimmerman and Schunk 1989). These accounts of academic learning, which described motivational and self-reactive as well as the metacognitive aspects of self-regulation, spurred considerable research. By the mid-1990s, a number of edited texts had been published chronicling the results of this first wave of descriptive research and experimental studies of self-regulated learning (Pintrich 1995, Schunk and Zimmerman 1994). The success of these empirical studies stimulated interest in systematic interventions to students' self-regulated learning and the results of these implementations emerged in journal articles and textbooks by the end of the 1990s (Schunk and Zimmerman 1998).

4. *Emphases in Current Theory and Research*

There is much current interest in understanding the self-motivational aspects of self-regulation as well as the metacognitive aspects (Heckhausen and Dweck 1998). Self-regulated learners are distinguished by their personal initiative and associated motivational characteristics, such as higher self-efficacy beliefs, learning goal orientations, favorable self-attributions, and intrinsic motivation, as well as by their strategic and self-monitoring competence. The issue of self-motivation is of both practical as well as theoretical importance. On the practical side, researchers often confront apathy or helplessness when they seek to improve students' use of self-regulatory processes, and they need viable methods for overcoming this lack of motivation. On the theoretical side, researchers need to understand how motivational beliefs interact with learning processes in a way that enhances students' initiative and perseverance. A number of models have included motivational and learning features as interactive components, such as Pintrich's self-schema model, Boerkaert's three-layered model, Kuhl's action/state control model, and Bandura, Schunk, and Zimmerman's cyclical phase model. These models are designed to transcend conceptual barriers between

learning and motivational processes and to understand their reciprocal interaction. For example, causal attributions are not only expected to affect students' persistence and emotional reactions but also adaptations in their methods of learning. Each of these models seeks to explain how learning can become self-motivating and can sustain effort over obstacles and time.

A second issue of current interest is the acquisition of self-regulatory skills in deficient populations of learners. Study skill training programs have been instituted with diverse populations and age groups of students. These intervention programs have involved a variety of formats, such as separate strategy or study skill training classes, strategy training classes linked explicitly to subject matter classes, such as in mathematics or writing, or personal trainer services in classrooms and counseling centers. This research is uncovering a fascinating body of evidence regarding the metacognitive, motivational, and behavioral limitations of at-risk students. For example, naive learners often overestimate their knowledge and skill, which can lead to understudying, nonstrategic approaches, procrastination, and faulty attributions. Successful academic functioning requires accurate self-perceptions before appropriate goals are set and strategies are chosen. Many of these interventions are predicated on models that envision self-regulatory processes as cyclically interdependent, and a goal of these approaches is to explain how self-fulfilling cycles of learning can be initiated and sustained by students.

5. *Methodological Issues*

Initial efforts to measure self-regulated learning processes relied on inventories in which students are asked to rate their use of specific learning strategies, various types of academic beliefs and attitudes, typical methods of study, as well as their efforts to plan and manage their study time. Another method is the use of structured interviews that involved open-ended questions about problematic learning contexts, such as writing a theme, completing math assignments, and motivating oneself to study under difficult circumstances. The latter form of assessment requires students to create their own answers, and experimenters to train coders to recognize and classify various qualitative forms of self-regulatory strategies. Although both of these approaches have reported substantial correlations with measures of academic success, they are limited by their retrospective or prospective nature—their focus on the usual or preferred methods of learning. As such, they depend on recall or anticipatory knowledge rather than on actual functioning under taxing circumstances. To avoid these limitations, there have been efforts to study self-regulation on-line at various key points during learning episodes using speak-aloud, experimenter questioning, and various performance and outcome measures. Experimental studies employing the latter measures offer the best opportunity to test the causal linkage among various self-regulatory processes, but, even this approach has potential shortcomings, such as inadvertent cuing or interference with self-regulatory processes.

Two common research design issues have emerged in academic interventions with at-risk populations of students: the lack of a suitable control group, and the reactive effects of self-regulatory assessments. Educators have asked self-regulation researchers to provide assistance to students who are often in jeopardy of expulsion, and it is unethical to withhold treatment from some of these students for experimental purposes. One solution is to use intensive within-subject, time-series designs in which treatments are introduced in sequential phases. However, as students begin to collect and graph data on themselves, they become more self-observant and self-reflective—which can produce unstable baselines. This, in turn, can confound causal inferences about the effectiveness of other self-regulatory components of the intervention.

6. *Future Directions for Research and Theory*

6.1 *Role of Technology*

The computer has been recommended as an ideal instrument to study and enhance students' self-regulation. Program menus can be faded when they are no longer needed and performance processes and outcomes can be logged in either a hidden or overt fashion. Computers provide the ultimate feedback to the experimenter or the learner because results can be analyzed and graphed in countless ways to uncover underlying deficiencies. Winne is developing a specially designed computer learning environment designed to study and facilitate self-regulation by the student.

6.2 *Developmental Research*

Relatively little attention has been devoted to forms of self-regulation that can be performed by young children. Very young children have difficulty observing and judging their own functioning, are not particularly strategic in their approach to learning, and tend to reason intuitively from beliefs rather than evidence. Their self-judgments of competence do not match their teachers' judgments until approximately the fifth grade. However, there is reason to expect that simple forms of self-regulatory processes begin to emerge during the early years in elementary school. For example, there is evidence that children can make self-comparisons with their earlier performance at the outset of elementary school. These issues are con-

nected to the key underlying issue of how self-regulation of learning develops in naturalistic as well as designed instructional contexts.

6.3 Out-of-school Influences

There is increasing research showing that non-academic factors such as peer groups, families, and part-time employment strongly affect school achievement. Schunk and Zimmerman have discussed the forms of social influence that these social groups can have on students' self-regulatory development. Brody and colleagues have found that parental monitoring of their children's activities and standard setting regarding their children's performance were very predictive of the children's academic as well as behavior self-regulation, which in turn was predictive of their cognitive and social development. Martinez-Pons has recently reported that parental modeling and support for their children's self-regulation was predictive of the youngsters' success in school. More attention is needed about the psychosocial origins of self-regulatory competence in academic learning.

6.4 Role of Teachers

The recent focus on academic intervention research has uncovered evidence that teachers often conduct classrooms where it is difficult for students to self-regulate effectively. For example, teachers who fail to set specific instructional goals, are ambiguous or inconsistent about their criteria for judging classroom performance, give ambiguous feedback about schoolwork, make it difficult for students to take charge of their learning. Of course, students who enter such classes with well-honed self-regulatory skills possess personal resources that poorly self-regulated learners do not possess. Such self-regulatory experts can turn to extra-classroom sources of information, can deduce subtle unspecified criteria for success, and can rely on self-efficacious beliefs derived from earlier successful learning experiences.

Regarding the development of self-regulated learners, few teachers ask students to make systematic self-judgments about their schoolwork, and as a result, students are not prompted or encouraged to use self-regulatory subprocesses such as self-observation, self-judgment, and self-reactions. Students who lack awareness of their functioning have little reason to try to alter their personal methods of learning. Finally, teachers who run classrooms where students' have little personal choice over their goals, methods, and outcomes of learning can undermine students' perceptions of control and assumption of responsibility for their classroom outcomes.

See also: Learning to Learn; Motivation, Learning, and Instruction; Self-efficacy: Educational Aspects

Bibliography

Heckhausen J, Dweck C S (eds.) 1998 *Motivation and Self-regulation Across the Lifespan*. Cambridge University Press, Cambridge, UK

Jan Simons P R, Beukhof R (eds.) 1987 *Regulation of Learning*. SVO, Den Haag, The Netherlands

Maehr M L, Pintrich P R (eds.) 1991 *Advances in Motivation and Achievement: Goals and Self-regulatory Processes*. JAI Press, Greenwich, CT, Vol. 7

Pintrich P (ed.) 1995 *New Directions in College Teaching and Learning: Understanding Self-regulated Learning*. Jossey-Bass, San Francisco, No. 63, Fall

Schunk D H, Zimmerman B J (eds.) 1994 *Self-regulation of Learning and Performance: Issues and Educational Applications*. Erlbaum, Hillsdale, NJ

Schunk D H, Zimmerman B J (eds.) 1998 *Self-regulated Learning: From Teaching to Self-reflective Practice*. Guilford Press, New York

Zimmerman B J (ed.) 1986 Self-regulated learning. *Contemporary Educational Psychology* **11**, Special Issue

Zimmerman B J, Schunk D H (eds.) 1989 *Self-regulated Learning and Academic Achievement: Theory, Research, and Practice*. Springer, New York

B. J. Zimmerman

Self-regulation in Adulthood

Self-regulation is one of the principal functions of the human self, and it consists of processes by which the self manages its own states and actions so as to pursue goals, conform to ideals and other standards, and maintain or achieve desired inner states. Many experts use self-regulation interchangeably with the everyday term self-control, although some invoke subtle distinctions such as restricting self-control to refer to resisting temptation and stifling impulses.

1. Scope of Self-regulation

Most knowledge about self-regulation can be grouped into four broad and one narrower category. The most familiar is undoubtedly impulse control, which refers to regulating one's behavior so as not to carry out motivated acts that could have harmful or undesirable consequences. Dieting, responsible sexual behavior, recovery from addiction, and control of aggression all fall in the category of impulse control.

A second category is affect regulation, or the control of emotions. The most common form of this is the attempt to bring oneself out of a bad mood or bring an end to emotional distress. In principle, however, affect regulation can refer to any attempt to alter any mood or emotion, including all attempts to induce, end, or prolong either positive or negative emotions. Affect regulation is widely regarded as the most difficult and problematic of the major spheres, because most people

cannot change their moods and emotions by direct control or act of will, and so people must rely on indirect strategies, which are often ineffective.

The third category is the control of thought. This includes efforts to stifle unwanted ideas or to concentrate on a particular line of thought. It can also encompass efforts to direct reasoning and inference processes, such as in trying to make a convincing case for a predetermined conclusion or to think an issue carefully through so as to reach an accurate judgment.

The fourth major category is performance regulation. Effective performance often requires self-regulation, which may include making oneself put forth extra effort or persevere (especially in the face of failure), avoid choking under pressure, and make optimal tradeoffs between speed and accuracy.

A narrower category of self-regulation involves superordinate regulation, which is sometimes called self-management. These processes cut across others and involve managing one's life so as to afford promising opportunities. Choosing challenges or tasks (such as college courses) that are appropriately suited to one's talents, avoiding situations that bring hard-to-resist temptations, and conserving one's resources during stressful periods fall in this category.

2. *Importance of Self-regulation*

The pragmatic importance of self-regulation can scarcely be understated. Indeed, most of the personal and social problems afflicting citizens of modern, highly developed countries involve some degree of failure at self-control. These problems include drug and alcohol abuse, addiction, venereal disease, unwanted pregnancy, violence, gambling, school failure and dropping out, eating disorders, drunk driving, poor physical fitness, failure to save money, excessive spending and debt, child abuse, and behavioral binge patterns. Experts regularly notify people that they could live longer, healthier lives if they would only quit smoking, eat right, and exercise regularly, but people consistently fail to regulate themselves sufficiently in those three areas. Longitudinal research has confirmed the enduring benefits of self-regulation. Mischel and his colleagues (e.g., Mischel 1996) found that children who were better able to exercise self-control (in the form of resisting temptation and delaying gratification) at age 4 years were more successful socially and academically over a decade later. Thus, these regulatory competencies appear to be stable and to yield positive benefits in multiple spheres over long periods of time. Self-regulation enables people to resist urges for immediate gratification and adaptively pursue their enlightened self-interest in the form of long-term goals.

Self-regulation also has considerable theoretical importance. As one of the self's most important and adaptive functions, self-regulation is central to the effective functioning of the entire personality. The processes by which the self controls itself offer important insights into how the self is structured and it operates. Higgins (1996) has analyzed the 'sovereignty of self-regulation,' by which he means that self-regulation is the master or supreme function of the self.

3. *Feedback Loops*

Psychologists have borrowed important insights from cybernetic theory to explain self-regulatory processes (e.g., Powers 1973). The influential work by Carver and Scheier (1981, 1982, 1998) analyzed self-awareness and self-regulation processes in terms of feedback loops that control behavior. In any system (including mechanical ones such as thermostats for heating/cooling systems), control processes depend on monitoring current status, comparing it with goals or standards, and initiating changes where necessary.

The basic form of a feedback loop is summarized in the acronym TOTE, which stands for test, operate, test, exit. The test phase involves assessing the current status and comparing it with the goals or ideals. If the comparison yields a discrepancy, some operation is initiated that is designed to remedy the deficit and bring the status into line with what is desired. Repeated tests are performed at intervals during the operation so as to gauge the progress toward the goals. When a test reveals that the desired state has been reached, the final (exit) step is enacted, and the regulatory process is ended.

Feedback loops do not necessarily exist in isolation, of course. Carver and Scheier (1981, 1982) explained that people may have hierarchies of such loops. At a particular moment, a person's actions may form part of the pursuit of long-term goals (such as having a successful career), medium-term goals (such as doing well in courses that will furnish qualifications for that career), and short-term goals (such as persisting to finish a particular assignment in such a course). Once a given feedback loop is exited, indicating the successful completion of a short-term act of self-regulation (e.g., finishing the assignment), the person may revert back to aiming at the long-term goals.

Emotion plays a central role in the operation of the feedback loop (Carver and Scheier 1998). Naturally, reaching a goal typically brings positive emotions, but such pleasant feelings may also arise simply from making suitable progress toward the goal. Thus, emotion may arise from the rate of change of discrepancy between the current state and the goal or standard. Meanwhile, negative emotions may arise not only when things get worse but also simply when one stands still and thereby fails to make progress. Carver and Scheier (1998) phrase this as a 'cruise control' theory of affect and self-regulation, invoking the analogy to an automobile's cruise control mechanism. Like that mechanism, emotion kicks in to regulate the process whenever the speed deviates from the prescribed rate of progress toward the goal.

Although this analysis has emphasized feedback loops that seek to reduce discrepancies between real and ideal circumstances, there are also 'negative feedback loops' in which the goal is to increase the discrepancy between oneself and some standard. For example, people may seek to increase the discrepancy between themselves and the average person.

4. *Self-regulation Failure*

Two main types of self-regulatory failure have been identified (for a review, see Baumeister et al. 1994). The more common is underregulation, in which the person fails to exert the requisite self-control. For example, the person may give in to temptation or give up prematurely at a task. The other category is misregulation, in which person exerts control over self but does so in some counterproductive manner so that the desired result is thwarted.

Several common themes and factors have been identified in connection with underregulation. One is a failure to monitor the self, which prevents the feedback loop from functioning. Thus, successful dieters often count their calories and keep track of everything they eat. When they cease to monitor how much they eat, such as when watching television or when distracted by emotional distress, they may eat much more without realizing it.

Alcohol contributes to almost every known pattern of self-control failure, including impulsive violence, overeating, smoking, gambling and spending, and emotional excess. This pervasive effect probably occurs because alcohol reduces people's attention to self and thereby impairs their ability to monitor their own behavior (see Hull 1981). Emotional distress likewise contributes to underregulation, and although there are multiple pathways by which it produces this effect, one of them is undoubtedly impairment of the self-monitoring feedback loop.

A perennial question is whether underregulation occurs because the self is weak or because the impulse is too strong. The latter invokes the concept of 'irresistible impulses,' which is popular with defense lawyers and addicts hoping to imply that no person could have resisted the problematic urge. After reviewing the evidence, Baumeister et al. (1994) concluded that most instances of underregulation involve some degree of acquiescence by the individual, contrary to the image of a valiant person being overwhelmed by an irresistible impulse. For example, during a drinking binge, the person will actively participate in procuring and consuming alcohol, which contradicts claims of having lost control of behavior and being passively victimized by the addiction. Yet clearly people would prefer to resist these failures and so at some level do experience them as against their will. Self-deception and inner conflict undoubtedly contribute to muddying the theoretical waters. At present, then, the relative contributions of powerful impulses and weak or acquiescent self-control have not been fully understood, except to indicate that both seem to play some role.

Misregulation typically involves some erroneous understanding of how self and world interact. In one familiar example, sad or depressed people may consume alcohol in the expectation that it will make them feel better, when in fact alcohol is a depressant that often makes them feel even worse. The strategy therefore backfires. By the same token, when under pressure, people may increase their attention to their own performance processes, under the assumption that this extra care will improve performance, but such attention can disrupt the smooth execution of skills, causing what is commonly described as 'choking under pressure.'

5. *Strength and Depletion*

The successful operation of the feedback loop depends on the 'operate' stage, in which the person actually changes the self so as to bring about the desired result. Recent work suggests that these operations involve a limited resource akin to energy or strength. The traditional concept of 'willpower' thus is being revived in psychological theory.

Willpower appears to be a limited resource that is depleted when people exert self-control. In laboratory studies, if people perform one act of self-regulation, they seem to be impaired afterwards. For example, after resisting the temptation to eat forbidden chocolates, people are less able to make themselves keep trying on a difficult, frustrating task (Baumeister et al. 1998). Whenever there are multiple demands on self-regulation, performance gradually deteriorates and afterwards the self appears to suffer from depletion of resources (Muraven and Baumeister 2000).

Although the nature of this resource is not known, several important conclusions about it are available. First, it appears that the same strength or energy is used for many different acts of self-regulation, as opposed to each sphere of self-control using a different facility. Trying to quit smoking or keep a diet may therefore impair a person's ability to persist at work tasks or to maintain control over emotions. Second, the resource is sufficiently limited that even small exertions cause some degree of depletion. These effects do not necessarily imply pending exhaustion; rather, people conserve their resources when partly depleted. Third, the resource is also used for making decisions and choices, taking responsibility, and exercising initiative. Fourth, there is some evidence that willpower can be increased by regular exercise of self-control. Fifth, rest and sleep seem to be effective at replenishing the resource.

The effects of ego depletion provide yet another insight into self-control failure. This helps explain the familiar observation that self-control tends to deteriorate when people are working under stress or

pressure, because they are expending their resources to coping with the stress and therefore have less available for other regulatory tasks.

6. Ironic Processes

Another contribution to understanding self-regulatory failure invokes the notion of different mental processes that can work at cross-purposes. Wegner (1994) proposed that effective self-regulation involves both a monitoring process, which remains vigilant for threat or danger (including, for example, the temptation to do something forbidden), and an operating process, which compels the self to act in the desirable fashion. When the monitoring process detects a threat, the operating process helps to prevent disaster, such as when the dieter refuses the offer of dessert. Unfortunately, the monitoring process is generally more automatic than the operating process, so the monitor may continue to notice temptations or other threats even when the operating process is not working. Upon completion of a diet, for example, the dieter may stop saying no to all tempting foods, but the monitoring process continues to draw attention to every delicious morsel that becomes available, so that the person feels constantly tempted to eat. Likewise, when people are tired or depleted, the monitoring system may continue to seek out troublesome stimuli, with disastrous results.

7. Conclusion

Effective self-regulation depends on three ingredients. First, one must have clear and unconflicting standards, such as clear goals or ideals. Second, one must monitor the self consistently, such as by keeping track of behavior. Third, one must have the wherewithal to produce the necessary changes in oneself, and that may include willpower (for direct change) or knowledge of effective strategies (for indirect efforts, such as for changing emotions). Problems in any area can impair self-regulation.

Several areas for future research on self-regulation are promising. The nature of the psychological resource that is expended must be clarified, and brain processes may shed valuable light on how self-regulation works. Interpersonal aspects of self-regulation need to be explicated. The replenishment of depleted resources will also help elucidate the nature of self-control in addition to being of interest in their own right. Individual differences in self-regulation and developmental processes for acquiring self-control need considerably more study. Applications to clinical, industrial, and relationship phenomena need to be examined.

Effective self-regulation includes the ability to adapt oneself to diverse circumstances. It is a vital aspect of human success in life.

Bibliography

Baumeister R F, Bratslavsky E, Murave M, Tice D 1998 Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* **74**: 1252–65

Baumeister R F, Heatherton T F, Tice D 1994 *Losing Control: How and Why People Fail at Self-regulation*. Academic Press, San Diego, CA

Carver C S, Scheier M F 1981 *Attention and Self-regulation*. Springer, New York

Carver C S, Scheier M F 1982 Control theory: A useful conceptual framework for personality—social, clinical and health psychology. *Psychological Bulletin* **92**: 111–35

Carver C S, Scheier M F 1998 *On the Self-regulation of Behavior*. Cambridge University Press, New York

Higgins E 1996 The 'self digest': Self-knowledge serving self-regulatory functions. *Journal of Personality and Social Psychology* **71**: 1062–83

Hull J G 1981 A self-awareness model of the causes and effects of alcohol consumption. *Journal of Abnormal Psychology* **90**: 586–600

Mischel W 1996 From good intentions to willpower. In: Gollwitzer P M, Bargh J A (eds.) *The Psychology of Action*. Guilford, New York

Muraven M, Baumeister R F 2000 Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin* **126**: 247–59

Powers W T 1973 *Behavior: the Control of Perception*. Aldine Publishing Company, Chicago

Wegner D M 1994 Ironic processes of mental control. *Psychological Review* **101**: 34–52

R. F. Baumeister

Self-regulation in Childhood

Self-regulation has several meanings in the psychology literature, some of which have conceptual links with each other. Three definitions of self-regulation are provided below, but only the first is the theme of this article.

1. Definitions

Self-regulation refers to the development of children's ability to follow everyday customs and valued norms embraced and prescribed by their parents and others. Self-regulation is a vital constituent of the socialization process. A broad construct, self-regulation encompasses a diverse set of behaviors such as compliance, the ability to delay actions when appropriate, and modulation of emotions, motor, and vocal activities as suitable to norm-based contexts. The *self* in self-regulation is an essential feature, and involves a sentient self—one that recognizes and understands the reasons for standards and evaluates one's own actions in relation to others' feelings and needs. Thus the hallmark of self-regulation is the ability to act in accordance with various family and social values in

the absence of external monitors, across a variety of situations, but neither slavishly nor mindlessly.

A second meaning of self-regulation refers to various physiological or psychobiological processes that function adaptively to situational demands, and often do not involve normative standards. Examples include the regulation of intensity of arousal subsequent to emotion producing events, or the control of centrally regulated physiological and perceptual systems to novel events. Centrally regulated systems include electrodermal responses, vagal tone and heart rate, and attention control. Variation in these responses is often studied for links to temperament styles, control of arousal states, and emotion control and coping (e.g., Rothbart et al. 1995).

A third view of self-regulation comes from recent perspectives in motivation. Here, with an emphasis on an individual's goals and the processes that shape the person's actions, a major role is given to the self's role and to evaluation of gains and losses (e.g., Heckhausen and Dweck 1998). Autonomy, control, self-integrity, and efficacy are essential in order for the individual to be effective in the pursuit of desired goals. Within the motivational framework, substantial attention is directed to understanding how adults facilitate or inhibit the child's autonomy.

As noted earlier, this entry is concerned with the development of self-regulation to family and sociocultural standards. Related articles include *Coping across the Lifespan*; *Control Behavior: Psychological Perspectives*; and *Self-regulation in Adulthood.*

2. *Historical Contexts*

The quest to understand the origins and development of self-regulation was a late twentieth century conceptual endeavor grounded in a clinical issue: why do some children as young as two and three years of age reveal high levels of dysregulated behaviors? Examples include resistance to parental requests, difficulty with family routines, high levels of activity, and unfocused attention often associated with an inability 'to wait.'

At the time, earlier findings (e.g., impulse control, resistance to discipline) had implicated factors such as ineffective parenting, as well as distortions in the child's own cognitive or language skills. With rare exceptions, few attempts had been made to understand antecedents, age-related correlates, and consequences of early self-regulated or dys-regulated behaviors. It seemed that a more comprehensive and integrative perspective could lead to a broadened developmental model, and provide insights about multiple contributors to early-appearing problematic behaviors. This was the impetus for a new view of *self-regulation* (Kopp 1982). Family and social standards were increasingly emphasized because of their relevance for self-regulation (Kopp 1987, Gralinski and Kopp 1993).

Self-regulation is not a unique psychological construct, rather it has ideational and behavioral links to courses of action variously labeled as self-management, self-control, and will. These emphases are longstanding and can be found in biblical passages, early philosophical writings, in manuals for parents during the seventeenth century and on to the present, and in the essays of William James, Freud, and John Dewey. More recent thinking can be found in several domains of psychology: developmental, personality, social learning, psychopathology. The enduring historical importance underscores the crucial fact that *individual* adherence to behavioral norms provides essential support for the structure and function of family and sociocultural groups.

Societal norms are grounded in historical time and locale, and are reflected by caregivers' socialization practices. Changes in practices have implications for the kinds of behaviors expected of children (including those associated with self-regulation). The dramatic changes in the sociocultural scene in the USA since the mid-1900s provide an example. At the end of the twentieth century, parenting was primarily authoritarian and children were expected to conform to standards without dissent. However, a mid-century transformation in child rearing practices occurred. The change was linked to two decades of economic hardship, a devastating war, quests for family life among returning veterans, and a pediatrician (Benjamin Spock) whose counsel to mothers about trusting their own judgement would resonate with new parents. In time, as a more relaxed style of parenting emerged, there was increasing tolerance for children to assert themselves about norms. Parents began to reveal a willingness to negotiate with their children about a variety of things including how norms could be met. At the beginning of the twenty-first century, a reciprocal style of parenting (often termed authoritative; Baumrind 1967) is far more common than the more inflexible authoritarian approach.

For social scientists, this changing socialization orientation led to new or recast research topics. Studies focused on the characteristics and consequences of authoritative parenting, children's autonomy needs, and understanding how the self-determination needs of the individual balance with the value based requirements of social–cultural groups. The challenges to define and explain the balancing process continue.

3. *Socialization and Self-regulation*

3.1 *Principles*

Despite changing emphases in socialization, three principles remain a constant; these have implications for the development of self-regulation. First, most sociocultural groups expect primary authority figures such as parents to begin the active process of socialization. These individuals actively expose older infants

and toddlers to family norms and practices. Later, these socializing agents will be joined by others who typically include teachers, age mates, friends, and neighbors. New socializing agents may reinforce previous norms, expand the interpretation of norms, or introduce new ones. Children are expected to adapt their behavior accordingly.

Second, parents tend to be mindful—perhaps implicitly—of a child's developmental capabilities, the family's functional needs, and sociocultural standards when they teach norms. Protection of the child from harm seems to be the most salient *initial* child-rearing value. Then the content of parents' socialization efforts progresses to other family concerns, and from there to the broader context of neighborhood, community, and societal norms. Across sociocultural groups, there are assumptions that toddlers and preschoolers will learn a few specific family and neighborhood norms, whereas school-aged children will adopt family and cultural customs and moral standards, formal laws, and economic values. Adolescents are expected to prepare themselves for social, emotional, and economic independence using as infrastructure the norms of the family and culture.

The third principle is that socialization and self-regulation are adaptive, bidirectional processes. Children and adolescents are not passive recipients in the socialization process. With the growth of their own cognitive skills and recognition of their own self needs, children show an increasing desire to have a say in defining the everyday customs that they encounter, and the ways customs are followed. Thus each succeeding generation in modern, democratic sociocultural groups modifies the content or enactment of customs and norms in some way. An implication for children is that effective self-regulation represents thoughtful, rather than indiscriminate, adherence to family practices and social norms.

3.2 Studying Self-regulation in the Context of Socialization

Self-regulation is often used as a conceptual frame for research rather than a design element with the three themes of compliance, delay, and behavioral modulation. Rather, most research has been focused on the study of compliance during the preschool period (typically, from 3 to 5 years of age). Although this emphasis is warranted—adherence to norms is considered to be a crucial developmental task for this age (Sroufe and Rutter 1984)—important knowledge gaps exist. There is fragmented information about effective integration of compliance, delay, and behavioral modulation, and developmental trends among younger and older children.

Recent descriptive studies also reveal considerable complexity in young children's responses to parental requests. This finding has prompted efforts to refine operational definitions, expand the research contexts for the study of self-regulation, and utilize data-collection procedures that involve multiple measurement techniques and informants. These recent efforts have largely focused on children without major behavioral problems.

Lastly, considerable research has highlighted parental and child correlates of self-regulation components, albeit sometimes within a limited age period and one or two contexts. The most complete database is available for the preschool years.

3.3 Research Data: Parents, Children, and Social Norms

The timing of active socialization originates in child behaviors such as the onset of walking. When very young children locomote on their own, they discover opportunities to explore, sometimes with the potential for physical or psychological harm. Thus parents start to identify acceptable and unacceptable behaviors for the child, and attempt to obtain some measure of compliance using a variety of attention getting techniques (Schaffer 1996). These beginnings mark the dynamic and occasionally formidable interplay between parents' socialization efforts and young children's trajectory toward self-regulation.

Although parents use their child's behavior as a socializing cue, they also rely on family needs (e.g., the composition of a family, living arrangements), as well as societal norms. A toddler in a large family with limited living space is likely to be exposed to somewhat different restrictions to an only child living in spacious quarters. However, three overarching themes unite parents' initial socialization efforts: young children must be protected from harming themselves or others (LeVine 1974, Gralinski and Kopp 1993); they must not tamper with family members' possessions; and they must learn to respect others' feelings (Dunn 1988). As children become older and their social environments increase, parental cautions about normative standards extend beyond the immediate family to peers, neighbors, and teachers, and focus on conventions and moral values, among other topics.

In addition to structuring the content of socialization norms, the how of parental socialization is crucial. There is unequivocal consensus about the effect of child-rearing styles at least with Euro-American middle-class families. Sensitive and knowledgeable parenting is correlated with children's compliance to norms, across age periods (Kochanska and Aksan 1995).

3.4 A Self-regulation Developmental Trajectory

Learning to deal with parents' norm-related requests is an all-powerful challenge. Children must bring their cognitive, linguistic, social, and emotional resources to situations that demand self-regulation. The task is most difficult for young children: they have cognitive

Table 1
Self-regulation: developmental landmarks

Landmarks	Components of self-regulation		
	Compliance	Delay/waiting (object rewards, others' attention)	Modulation of emotions, vocalizations, and motor activities
Emergent: responses tend to be situation specific and unpredictable	Early 2nd year	Early 2nd year	Early PS[a]
Functional: a few predictable norm responses; some cognizance of self role	End 2nd year; early PS	Mid-PS	Mid-PS
Integrative: response coherence across two self-regulation components; nascent causal reasoning	Late PS	Late PS	Late PS
Internalized: understands, accepts others' values and self role for social norms	*Late PS*[b]; *S-A*[c]	*S-A*	*S-A*
Future oriented: planful behavior for some norm contexts	*Mid-late S-A*	*Mid-late S-A*	*Mid-late S-A*

a Preschool period. b Italics indicate presumed progression. c School age years.

and language limitations; they are inclined to explore anything that looks interesting; they long for autonomy and self-assertion. It is not surprising that self-regulation takes years to mature into an effective process.

Table 1 depicts the three components of self-regulation along with ages associated with developing behavioral landmarks. The ages represent approximations; the landmarks highlight behaviors that typify increasing maturity to parental norm-based requests.

Using compliance as an exemplar, *Emergent* signifies that responses to a parent prohibition (e.g., child does not stand on a kitchen chair) may occur sporadically. *Functional* portrays fairly regular compliance to a rule (e.g., child does not stand on chairs, whether in the living room or kitchen); children may show signs of remorse when they do not comply. *Integrative* refers to predictability (i.e., coherence) in children's behavior *across different kinds* of norm-based situations (e.g., a child does not yell in a market, respects a sibling's possessions, and waits to be served dinner).

Extant data reveal that compliance to family *do's* and *dont's* tends to be easier for young children than norm-based situations that require waiting or modulation of ongoing behaviors (also shown in Table 1). This disparity may be related to additional regulatory demands imposed on children when timing or refined behavioral nuances are critical for an effective response. In these instances, the use of strategic behaviors (e.g., self-distraction, conscious suppression efforts) may be crucial for effective self-regulation.

Despite inevitable setbacks, children become more responsive to specific family and other social rules. The growth of self-regulation is facilitated by greater understanding of people and events, widening sensitivity to everyday happenings in the family and neighborhood, and increasing ability to talk about self needs in relation to others' needs. The importance of language in self-regulation is exemplified by the transition from outright refusals common among toddlers to attempts to talk about and negotiate task demands among older preschoolers (Klimes-Dougan and Kopp 1999). Still, children of this age may have difficulty with anger control, waiting for parental attentiveness, and sharing possessions.

An extensive database on topics related to self-regulation in the preschool years suggests that correlates include gender (many girls learn behavioral controls sooner than boys), temperament characteristics such as controlled attention and inhibition of motor acts, verbal skills, effective self-distraction, and competent use of strategies (Eisenberg and Fabes 1992, Metcalfe and Mischel 1999).

Overall, continuing efforts of parents to socialize their children gradually dovetail with children's growth of their own behavioral cognitive, social, and motivational repertoire. Few details exist about how parents and children collaborate with their own resources so that children increasingly take on responsibility for self-regulation. What can be said with certainty is that self-regulation gradually emerges when children understand the reasons for values and standards, possess a cognizant self, are able to suppress an immediate goal that runs counter to family norms, and voluntarily assume responsibility for their own actions across a variety of situations.

Self-generated and self-monitored adherence to norms *is* self-regulation. This autonomous, sometimes conscious, effort to frame activities with normative values stands in contrast to early forms of compliance that do not rely on knowledge or self-awareness. Effective self-regulation is by definition an adaptive,

developmental process in that children have to discover how to meet their own self needs while in general following societal standards across many settings. The task is a changing one because each age period is associated with new socialization demands imposed by parents, teachers, other individuals, and the larger social context.

4. Self-regulation: Its Value and Directions

The construct of self-regulation has heightened awareness of dys-regulation and its long-term implications. Findings from related research point to long-standing problems when young children have difficulty with compliance, delay, and impulse control. To date, however, the developmental antecedents of these problems have been elusive. A renewed focus on the toddler years should be useful: two important developments occur in the second and third years that have relevance for the growth of self-regulation. These are emergent selfhood and the control of attention and consciousness.

It is well known that effective norm-based behaviors require reconciliation of conflicts between self-goals and social norm demands. Among mature individuals, these conflicts are often met with reflection about courses of action, informal appraisals of costs and benefits, and plans for making reparations should they be necessary. This complex cognitive strategy is not available to young children. Understanding how very young children begin to sort through competing self and social goals may provide insights into effective paths to self-regulation.

With respect to conscious learning about social norms, this almost certainly occurs, but when and how are not well understood. However, consciousness demands psychic energy, so it is in children's best interests to sort out those family and social norms that require relatively habitual responses from those that demand additional attentional or memory efforts from themselves. How children learn to differentiate and classify standards may yield understanding about strategies useful for modulating behaviors in novel norm-based contexts, and why some children falter in these situations.

See also: Parenting: Attitudes and Beliefs; Self-development in Childhood; Self-regulation in Adulthood; Socialization and Education: Theoretical Perspectives; Socialization in Infancy and Childhood; Socialization, Sociology of

Bibliography

Baumrind D 1967 Child care practices anteceding three patterns of preschool behavior. *Genetic Psychology Monographs* **75**: 43–88

Dunn J 1988 *The Beginning of Social Understanding*. Blackwell, Oxford, UK

Eisenberg N, Fabes R A 1992 Emotion, regulation, and the development of social competence. In: Clark M S (ed.) *Review of Personality and Social Psychology: Vol. 14. Emotion and Social Behavior*. Sage, Newbury Park, CA, pp. 119–50

Gralinski J H, Kopp C B 1993 Everyday rules for behavior: Mothers' requests to young children. *Developmental Psychology* **29**: 573–84

Heckhausen J M, Dweck C S 1998 *Motivation and Self-regulation Across the Life Span*. Cambridge University Press, New York

Klimes-Dougan B, Kopp C B 1999 Children's conflict tactics with mothers: A longitudinal investigation of the toddler and preschool years. *Merrill-Palmer Quarterly* **45**: 226–41

Kochanska G, Aksan N 1995 Mother child mutually positive affect, the quality of child compliance to requests and prohibitions, and maternal control as correlates of early internalization. *Child Development* **66**: 236–54

Kopp C B 1982 Antecedents of self-regulation: a developmental perspective. *Developmental Psychology* **18**: 199–214

Kopp C B 1987 The growth of self-regulation: caregivers and children. In: Eisenberg N (ed.) *Contemporary Topics in Developmental Psychology*. Wiley, New York, pp. 34–55

LeVine R A 1974 Parental goals: A cross-cultural view. In: Leichter H J (ed.) *The Family as Educator*. Teachers College Press, New York, pp. 56–70

Metcalfe J, Mischel W 1999 A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychological Review* **106**: 3–19

Rothbart M K, Posner M I, Hershey K L 1995 Temperament, attention, and developmental psychopathology. In: Cicchetti D, Cohen D J (eds.) *Manual of Developmental Psychopathology*. Wiley, New York, Vol. 1, pp. 315–40

Schaffer H R 1996 *Social Development*. Blackwell, Oxford, UK

Sroufe L A, Rutter M 1984 The domain of developmental psychopathology. *Child Development* **55**: 17–29

C. B. Kopp

Semantic Knowledge: Neural Basis of

Semantic knowledge is a type of long-term memory, commonly referred to as semantic memory, consisting of concepts, facts, ideas, and beliefs (e.g., Tulving 1983). Semantic memory is thus distinct from *episodic* or *autobiographical* memories, which are unique to an individual and tied to a specific time and place. For example, answering the question 'What does the word *breakfast* mean?' requires semantic memory. In contrast, answering the question 'What did you have for breakfast yesterday?' requires episodic memory, to retrieve information about events in our personal past, as well as semantic memory, to understand the question. Semantic memory therefore includes the information stored in our brains that represents the meaning of words and objects.

Understanding the nature of meaning, however, has proven to be a fairly intractable problem: especially

regarding the meaning of words. One important reason why this is so is that words have multiple meanings. The specific meaning of a word is determined by context, and comprehension is possible because we have contextual representations (see Miller 1999 for a discussion of lexical semantics and context). Cognitive neuroscientists have begun to get traction on the problem of meaning in the brain by limiting inquiry to concrete objects as represented by pictures, and by their names.

Consider, for example, the most common meaning of two concrete objects; a camel and a wrench. Camel is defined as 'either of two species of large, domesticated ruminants (genus camelus) with a humped back, long neck, and large cushioned feet,' the object wrench is defined as 'any number of tools used for holding and turning nuts, bolts, pipes, etc.' (Webster's New World Dictionary 1988). Two things are noteworthy about these definitions. First, they are largely about features; camels are large and have a humped back; wrenches hold and turn things. Second, different types of features are emphasized for different types of object. The definition of the camel consists of information about its visual appearance, whereas the definition of the wrench emphasizes how it is used. Differences in the types of feature that define different objects have played, and continue to play, a central role in models of how semantic knowledge is organized in the human brain.

Another point about these brief definitions is that they include only part, and perhaps only a small part, of the information we may possess about these objects. For example, we may know that camels are found primarily in Asia and Africa, that they are known as the 'ships of the desert,' and that the word camel can also refer to a color and a brand of cigarettes. Similarly, we also know that camels are larger than a bread box and weigh less than a 747 jumbo jet. Although this information is also part of semantic memory, little is known about the brain bases of these associative and inferential processes. Neuroscientists are, however, beginning to gain insights into the functional neuroanatomy associated with identifying objects and retrieving information about specific object features and attributes.

1. *Semantic Representations*

A central question for investigators interested in the functional neuroanatomy of semantic memory has been to determine how information about object concepts is represented in the brain. A particularly influential idea guiding much of this work is that the features and attributes that define an object are stored in the perceptual and motor systems active when we first learned about that object. For example, information about the visual form of an object, its typical color, its unique pattern of motion, would be stored in or near regions of the visual system that mediate perception of form, color, and motion. Similarly, knowledge about the sequences of motor movements associated with the use of an object would be stored in or near the motor systems active when that object was used. This idea has a long history in behavioral neurology. Indeed, many neurologists at the beginning of the twentieth century assumed that the concept of an object was composed of the information about that object learned through direct sensory experience and stored in or near sensory and motor cortices (e.g., Lissauer 1988, 1890).

2. *Semantic Deficits Result from Damage to the Left Temporal Lobe*

The modern era of study on the organization of semantic knowledge in the human brain began with Elizabeth Warrington's seminal paper 'The selective impairment of semantic memory' (Warrington 1975). Warrington reported three patients with progressive dementing disorders that provided neurological evidence for a semantic memory system. There were three main components to the disorder. First, it was selective. The disorder could not be accounted for by general intellectual impairment, sensory or perceptual problems, or an expressive language disorder. Second, the disorder was global, in the sense that it was neither material- nor modality-specific. Object knowledge was impaired regardless of whether objects were represented by pictures or their written or spoken names. Third, the disorder was graded. Knowledge of specific object attributes (e.g., does a camel have two, four, or six legs?) was more impaired than knowledge of superordinate category information (i.e., is a camel a mammal, bird, or insect?).

Following Warrington's report, similar patterns of semantic memory dysfunction have been reported in patients with brain damage resulting from a wide range of etiologies. These have included patients with progressive dementias such as Alzheimer's disease and semantic dementia, herpes encephalitis, and closed head injury (for review, see Patterson and Hodges 1995). Consistent with the properties established by Warrington, these patients typically had marked difficulty producing object names under a variety of circumstances, including naming pictures of objects, naming from the written descriptions of objects, and generating lists of objects that belong to a specific category (e.g., animals, fruits and vegetables, furniture, etc.). In addition, the deficits were associated primarily with damage to the left temporal lobe, suggesting that information about object concepts may be stored, at least in part, in this region of the brain.

3. Brain Damage Can Lead to Category-specific Semantic Deficits

Other patients have been described with relatively selective deficits in recognizing, naming, and retrieving information about different object categories. The categories that have attracted the most attention are animals and tools. This is because of a large and growing number of reports of patients with greater difficulty naming and retrieving information about animals (and often other living things) than about tools (and often other man-made objects). Reports of the opposite pattern of dissociation (greater for tools than animals) are less frequent. However, enough carefully-studied cases have been reported to provide convincing evidence that these categories can be doubly dissociated as a result of brain damage. The impairment in these patients is not limited to visual recognition. Deficits occur when knowledge is probed visually and verbally, and therefore assumed to reflect damage to the semantic system or systems (for review, see Forde and Humphreys 1999).

While it is now generally accepted that these disorders are genuine, their explanation, on both the cognitive and neural levels, remains controversial. Two general types of explanation have been proposed. The most common explanation focuses on the disruption of stored information about object features. Specifically, it has been proposed that knowledge about animals and tools can be disrupted selectively because these categories are dependent on information about different types of features stored in different regions of the brain. As exemplified by the definitions provided previously, animals are defined primarily by what they look like, and functional attributes play a much smaller role in their definition. In contrast, functional information, specifically how an object is used, is critical for defining tools. As a result, damage to areas where object form information is stored leads to deficits for categories that are overly-dependent on visual form information, whereas damage to regions where object use information is stored leads to deficits for categories overly-dependent on functional information. The finding that patients with category-specific deficits for animals also have difficulties with other visual-form-based categories, such as precious stones, provides additional support for this view (e.g., WarPrington and Shallice 1984).

This general framework for explaining category-specific disorders was first proposed by Warrington and colleagues in the mid- to late 1980s (e.g., Warrington and Shallice 1984). Influential extensions and reformulation of this general idea have been provided by a number of investigators, including Farah and McClelland (1991), Damasio (1990), Caramazza et al. (1990), and Humphreys and Riddoch (1987).

The second type of explanation focuses on broader semantic distinctions (e.g., animate v. inanimate objects), rather than on features and attributes, as the key to understanding category-specific deficits. A variant of this argument has recently been proposed by Caramazza and Shelton (1998) to counter a number of difficulties with feature-based formulations. Specifically, these investigators note that a central prediction of at least some feature-based models is that patients with a category-specific deficit for animals should have more difficulty in answering questions that probe knowledge about visual information about animals (does an elephant have a long tail?), than about function information (is an elephant found in the jungle?). As Caramazza and Shelton (1998) show, at least some patients with an animal-specific knowledge disorder (and, according to their argument, all genuine cases) have equivalent difficulty with both visual and functional questions about animals. As a result of these and other findings, Caramazza and Shelton argue that category-specific disorders cannot be explained by feature-based models. Instead, they propose that such disorders reflect evolutionary adaptations for animate objects, foods, and perhaps by default, tools, and other manufactured objects (the 'domain-specific hypothesis'; Caramazza and Shelton 1998).

4. Functional Brain Imaging Reveals that Semantic Knowledge about Objects is Distributed in Different Regions of the Brain

Functional brain imaging allows investigators to identify the neural systems active when normal individuals perform different types of task. These studies have confirmed findings with brain damaged subjects, and have begun to extend knowledge of the neural basis of semantic memory. Although functional brain imaging is a relatively new tool, the current evidence suggests that information about different object features is stored in different regions of the cerebral cortex. One body of evidence in support of this claim comes from experiments using word generations tasks. Subjects are presented with the name of an object, or a picture of an object, and required to generate a word denoting a specific feature or attribute associated with that object. In one study, positron emission tomography (PET) was used to investigate differences in patterns of brain activity when subjects generated the name of an action typically associated with an object (e.g., saying the word 'pull' in response to a static, achromatic line-drawing of a child's wagon), relative to generating an associated color word e.g., saying 'red' in response to the wagon) (Martin et al. 1995). Relative to generating a color word, action-word generation activated a posterior region of the left temporal lobe, called the middle temporal gyrus. This location was of particular interest because it was just anterior to sites known to be active during motion

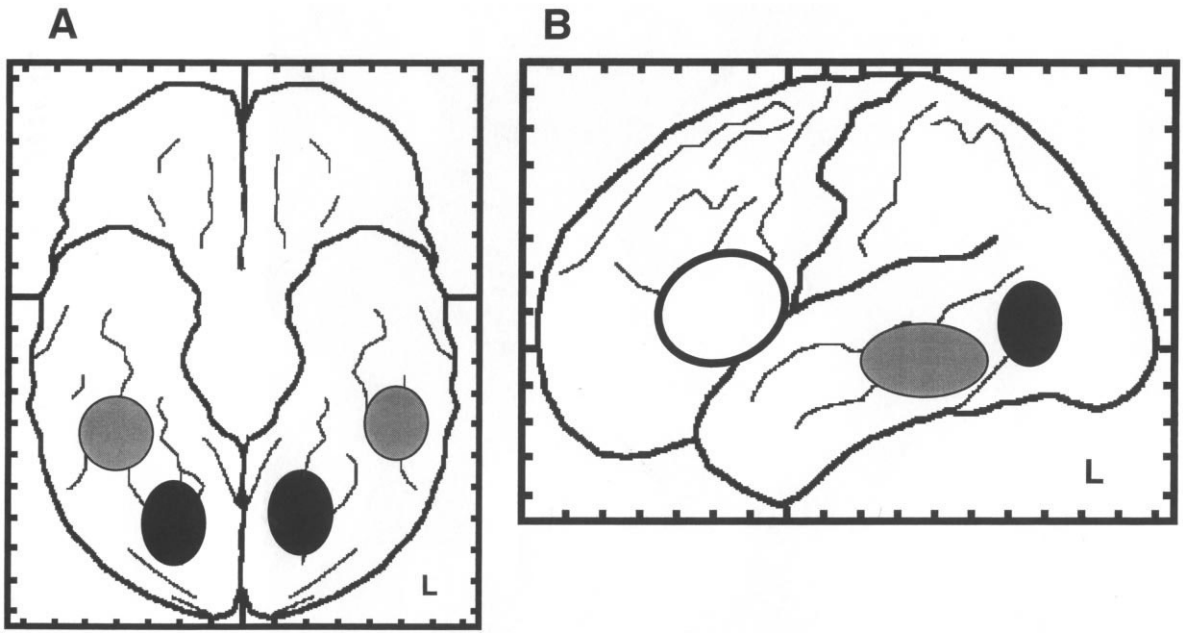

Figure 1
Brain regions active during retrieval of color and action knowledge. **A** Ventral view of the human brain showing approximate locations of regions in the occipital cortex active during color perception (black ovals), and color word generation in response to object pictures and object names (gray ovals). This region is also active when subjects generate color imagery in response to verbal commands, and in subjects with auditory word–color synethesia (for review, see Martin 2001). **B** Lateral view of the left hemisphere showing location of the region active during motion perception (area MT; black oval), and region of the middle temporal gyrus active when subjects generated action words to object pictures, and object names (gray oval). Location of this region is based on over 20 studies encompassing nine native languages (for review, see Martin 2001). The white oval indicates the approximate location of the region of inferior prefrontal cortex implicated in selecting, retrieving, and maintaining lexical and phonological information (for review, see Martin and Chao 2001)

perception (area MT). In contrast, relative to generating action words, color-word generation activated a region on the ventral, or underside of the temporal lobe, called the fusiform gyrus. This location was of particular interest because it was just anterior to sites known to be active during color perception. Finally, the same pattern of results was found when subjects were presented with the written names of objects, rather than an object picture (Fig. 1) (for review see Martin 2001).

These findings, and findings from other studies using a wide range of semantic processing tasks, provide support for two important ideas about the neural representation of semantic knowledge. First, there is a single semantic system in the brain, rather than separate semantic systems for different modalities of input (visual, auditory) or types of material (pictures of objects, words) (e.g., Vandeberghe et al. 1996). This system includes multiple brain regions, especially in the temporal and frontal lobes of the left hemisphere (for review see Price et al. 1999, Martin 2001). Second, information about object features and attributes are not stored in a single location, but rather are stored as a distributed network of discrete cortical areas. Moreover, the locations of the sites appear to follow a specific plan that parallels the organization of sensory systems, and, as will be reviewed below, motor systems, as well.

5. *Ventral Occipitotemporal Cortex and the Representation of Object Form*

Another body of evidence that object concepts may be represented by distributed feature networks comes from studies contrasting patterns of neural activity associated with naming, and performing other types of tasks with objects from different categories. A feature common to all concrete objects is that they have physical shape or form. Evidence is accumulating that suggests that many object categories elicit distinct

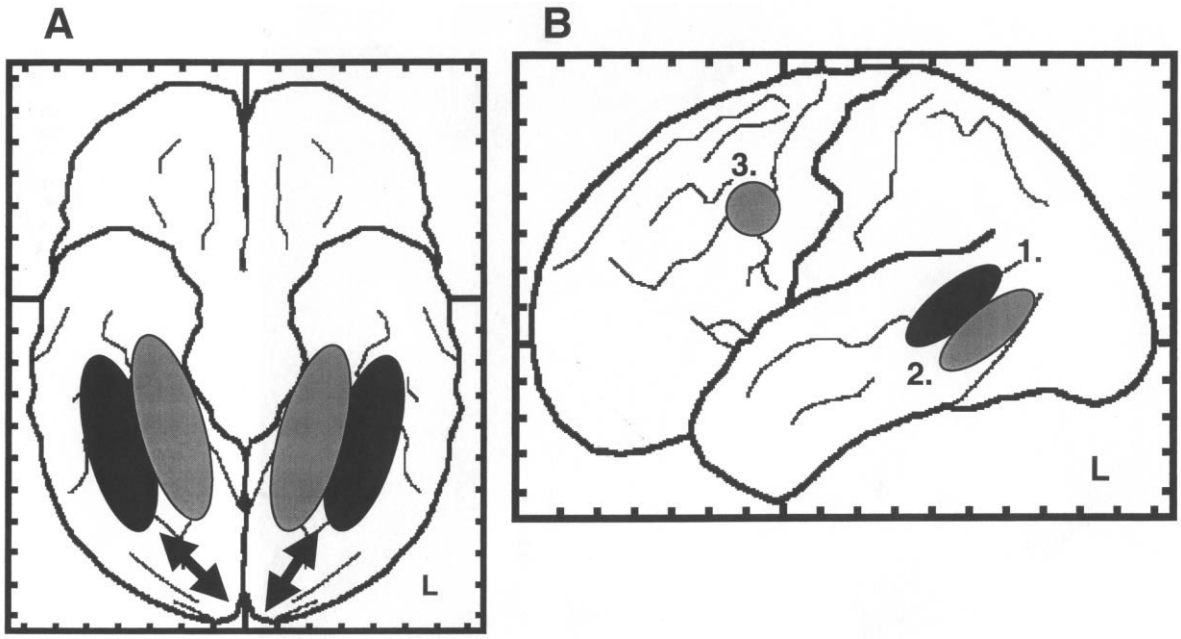

Figure 2
Brain regions showing object category-related patterns of activation. Approximate location of regions in the ventral temporal cortex **A**, and lateral cortex **B**, active during viewing, naming, and retrieving information about animals (black ovals) and tools (gray ovals). Blecause of their anatomic proximity to visual object form (and form-related features like color), motion, and motor processing areas, it has been suggested that information about how an object appears, moves, and is used, is stored in these regions (see text for details). These regions interact with lower-level visual processing areas in occipital cortex (double arrows); especially when discriminating among items of similar appearance (e.g., four-legged animals, faces). 1. Superior temporal sulcus (STS). 2. Middle temporal gyrus. 3. Premotor cortex

patterns of neural activity in regions involved in object form processing (ventral occipital and temporal cortex). Moreover, the locations of these category-related activations appear to be consistent across individual subjects and processing tasks (e.g., naming object pictures, matching pictures, reading object names). This seems to be especially so for objects defined primarily by visual form-related features such as animals, faces, and landmarks.

Early reports using PET found that naming (Martin et al. 1996) and matching (Perani et al. 1995) pictures of animals resulted in greater activation of the left occipital cortex than performing these same tasks with pictures of tools. Because the occipital cortex is involved primarily in the early stages of visual processing, it was suggested that this activity reflected top-down activation from more anterior sites in the occipitotemporal object processing stream (Martin et al. 1996). This may occur whenever detailed information about visual features or form is needed to identify an object. Specifically, naming objects that differ from other members of the same category by relatively subtle differences in visual form (four-legged animals) may require access to stored information about visual detail. Retrieving this information, in turn, may require participation of early visual processing areasn the occipital cortex. A subsequent report showing that unilateral occipital lesions could result in greater difficulty in naming and retrieving information about animals than tools provided converging evidence for this view (Tranel et al. 1997).

However, most patients with semantic deficits for animals have had lesions confined to the temporal lobes (for review, see Gainotti et al. 1995). Functional brain imaging of normal individuals has now provided evidence for category-related activations in the ventral region of the temporal lobes. This has been accomplished by using functional magnetic resonance imaging (fMRI), which provides better spatial resolution than was possible using PET.

A number of investigators have found that distinct regions of the ventral temporal cortex show differential responses to different object categories. In one study, viewing, naming, and matching pictures of animals, as well as answering written questions about animals, were found to activate the lateral region of the fusiform

gyrus, relative to performing these tasks with pictures and names of tools. In contrast, the medial fusiform was more active for tools than animals. A similar, but not identical, pattern of activation was found for viewing faces (lateral fusiform) relative to viewing houses (medial fusiform) (Chao et al. 1999). Other investigators have also reported face-related activity in the lateral region of the fusiform gyrus, and house-related activity in more medial regions of the ventral temporal lobe, including the fusiform, lingual, and parahippocampal gyrus (Fig. 2) (for review, see Kanwisher et al. 2001, Martin 2001).

These findings suggest that different object categories elicit activity in different regions of ventral temporal cortex, as defined by the location of their peak activation. Moreover, the typological arrangement of these peaks was consistent across subjects and tasks. Importantly, however, the activity associated with each object category was not limited to a specific region of the ventral occipitotemporal cortex, but rather was distributed over much of the region (Chao et al. 1999).

Additional evidence for the distributed nature of object representations in the ventral temporal cortex comes from single cell recordings from intracranial depth electrodes implanted in epileptic patients (Kreiman et al. 2000a). Recordings from the regions of the medial temporal cortex (entorhinal cortex, hippocampus, and amygdala), which receive major inputs from the ventral temporal regions described above, identified neurons that showed highly selective responses to different object categories, including animals, faces, and houses. Moreover, the responses of the neurons were category-specific rather than stimulus-specific. That is, animal-responsive cells responded to all pictures of animals, rather than to one picture or a select few.

Studies reporting similar patterns of neural activity when subjects view and imagine objects provide further support that object information is stored in these regions of cortex. For example, regions active during face perception are also active when subjects imagine famous individuals (O'Craven and Kanwisher 2000). Similar findings have been reported for viewing and imagining known landmarks (O'Craven and Kanwisher 2000), houses, and even chairs (Ishai et al. 2000). In addition, the majority of category-selective neurons recorded from human temporal cortex also responded selectively when the patients were asked to imagine these objects (Kreiman et al. 2000b).

Taken together, the data suggest that the ventral occipitotemporal cortex may be best viewed, not as a mosaic of discrete category-specific areas, but rather as a lumpy feature-space, representing stored information about features of object form shared by members of a category. How this feature space is organized, and why its topological arrangement is so consistent from one subject to another, are critical questions for future investigations.

6. *Lateral Temporal Cortex and the Representation of Object Motion*

Information about how objects move through space, and patterns of motor movements associated with their use, are other features that could aid object identification. This would be especially true for categories of manufactured objects such as tools that have a more variable mapping between their name and their visual form than a category such as four-legged animals. Thus access to these additional features may be required to identify them as unique entities. Here again, evidence is accumulating that naming and identifying objects with motion-related attributes activate areas close to regions that mediate perception of object motion (posterior region of the lateral temporal lobe) with different patterns of activity associated with biological and manufactured objects.

A number of laboratories using a variety of paradigms with pictures and words have reported that tools elicit greater activity in the posterior, left middle temporal gyrus than animals and other object categories (for review, see Martin 2001). Moreover, the active region was just anterior to area MT, and overlapped with the region active in the verb generation studies discussed above. Damage to this region has been reported selectively to impair tool recognition and naming (Tranel et al. 1997). In contrast, naming animals and viewing faces elicits greater activity in the superior temporal sulcus (STS) (Fig. 2). This region is of particular interest because of its association with the perception of biological motion in monkeys as well as humans (for review, see Allison et al. 2000). As suggested for the ventral temporal cortex, neurons in the lateral temporal cortex may also be tuned to features that objects within a category share. The nature of these features is unknown; however, based on its anatomical proximity to visual motion processing areas, this region may be tuned to features of motion associated with different objects. In support of this conjecture, increased activity in the posterior lateral temporal cortex has been found when subjects viewed static pictures of objects that imply motion (Kourtzi and Kanwisher 2000, Senior et al. 2000), and when subjects focused attention on the direction of eye gaze (Hoffman and Haxby 2000). Investigation of the differences in the properties of motion associated with biological and manufactured objects may provide clues to the organization of this region.

7. *Ventral Premotor Cortex and the Representation of Use-associated Motor Movements*

If activations associated with different object categories reflect stored information about object properties, then one would expect tools to elicit activity in

motor-related regions. Several laboratories have reported this association. Specifically, greater activation of the left ventral premotor cortex has been found for naming tools relative to naming animals, viewing pictures of tools relative to viewing pictures of animals, faces, and houses, and generating action words to tools (Fig. 2). Mental imagery (e.g., imagining manipulating objects with the right hand) has also resulted in ventral premotor activation (Fig. 2) (for review, see Martin 2001).

Electrophysiological studies have identified cells in monkey ventral premotor cortex that responded not only when objects were grasped, but also when the animals viewed objects they had had experience of manipulating (Jeannerod et al. 1995). The ventral premotor activation noted in the human neuroimaging studies may reflect a similar process. These findings are consistent with reports of patients with greater difficulty in naming tools than animals following damage to the left lateral frontal cortex (for review, see Gainotti et al, 1995) and suggest that the left ventral premotor cortex may be necessary for naming and retrieving information about tools.

8. *Conclusions and Future Directions*

Evidence from functional brain imaging studies provides considerable support for feature-based models of semantic representation. However, some of the findings could be interpreted as evidence for the 'domain-specific' hypothesis as well. For example, the clustering of activations associated with animals and faces, on the one hand, and tools and houses, on the other, may be viewed as consistent with this interpretation. Other evidence suggests, however, that all nonbiological object representations do not cluster together. For example, it has been reported that activity associated with a category of objects of no evolutionary significance (chairs) was located lateral to the face-responsive region (in the inferior temporal gyrus), rather than medially where tools and houses elicit their strongest activity (Ishai et al. 1999). Both functional brain imaging and patient studies suggest that object knowledge is represented in distributed cortical networks. There do not seem to be single regions that map on to whole object categories. Nevertheless, there may be a broader organization of these networks that reflects evolutionarily adapted, domain-specific knowledge systems for biological and nonbiological kinds of object. This possibility remains to be explored.

Although progress is being made in understanding the neural substrate of some aspects of meaning, many important questions and issues remain. For example, semantic representations are prelexical. Yet, to be of service, they must be linked intimately to the lexicon. Little is known about how the lexicon is organized in the brain, and how lexical and semantic networks interact (Damasio et al. 1996).

Another important issue concerns the neural basis of retrieval from semantic memory. Semantic knowledge, like all stored information, is not useful unless it can be retrieved efficiently. Studies of patients with focal lesions have shown that the left lateral prefrontal cortex is involved critically in word retrieval, even in the absence of a frank aphasia (e.g., Baldo and Shimamura 1998). Functional brain imaging studies have confirmed this association (Fig. 2). Moreover, recent evidence suggests that different regions of the left inferior prefrontal cortex may mediate selection among competing alternatives in semantic memory, whereas other regions may be involved in retrieving, manipulating, and maintaining semantic information in working memory (for review, see Martin and Chao 2001). Much additional work will be needed to describe the role that different regions of prefrontal cortex play in semantic processing.

Another critical question will be to determine how semantic object representations are modified by experience. Some of the findings discussed here suggest that the typological arrangement among object categories in the cortex is relatively fixed. Other evidence suggests a much more flexible organization, in which the development of expertise with an object category involves a particular portion of the fusiform gyrus (Gauthier et al. 1999). Longitudinal studies tracking changes in the brain associated with learning about completely novel objects should help to clarify this issue.

There is also little known about where object information unrelated to sensory or motor properties is stored (e.g., that camels live in Asia and Africa). Similarly, little is known about the representation of abstract concepts (e.g., honor, liberty, and justice), metaphors, and the like.

Finally, questions concerning the neural systems involved in object category formation have been almost totally neglected. E. E. Smith and colleagues have shown that exemplar-based and rule-based categorization activate different neural structures (Smith et al. 1999). This finding suggests that categorization processes may be a fruitful area for future investigation. The advent of techniques to converge fMRI data with technologies such as magnetoencephalography (MEG) that provide temporal information on the order of milliseconds, should provide a wealth of new information on the neural basis of semantic knowledge.

See also: Comprehension, Cognitive Psychology of; Dementia, Semantic; Evolution and Language: Overview; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Meaning and Rule-following: Philosophical Aspects; Memory for Meaning and Surface Memory; Semantic Similarity, Cognitive Psychology of; Semantics; Sentence Comprehension, Psychology of; Word, Linguistics of; Word Meaning: Psychological Aspects

Bibliography

Allison T, Puce A, McCarthy G 2000 Social perception and visual cues: role of the STS. *Trends in Cognitive Science* **4**: 267–78

Baldo J V, Shimamura A P 1998 Letter and category fluency in patients with frontal lobe lesions. *Neuropsychology* **12**: 259–67

Caramazza A, Shelton J R 1998 Domain-specific knowledge systems in the brain the animate-inanimate distinction. *Journal of Cognitive Neuroscience* **10**: 1–34

Caramazza A, Hillis A E, Rapp B C, Romani C 1990 The multiple semantics hypothesis: multiple confusions? *Cognitive Neuropsychology* **7**: 161–89

Chao L L, Haxby J V, Martin A 1999 Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience* **2**: 913–19

Damasio A R 1990 Category-related recognition deficits as a clue to the neural substrates of knowledge. *Trends in Neurosciences* **13**: 95–8

Damasio H, Grabowski T J, Tranel D, Hichwa R D, Damasio A R 1996 A neural basis for lexical retrieval. *Nature* **380**: 499–505

Farah M J, McClelland J L 1991 A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General* **120**: 339–57

Forde E M E, Humphreys G W 1999 Category-specific recognition impairments: a review of important case studies and influential theories. *Aphasiology* **13**: 169–93

Gainotti G, Silveri M C, Daniele A, Giustolisi L 1995 Neuroanatomical correlates of category-specific semantic disorders: a critical survey. *Memory* **3**: 247–64

Gauthier I, Tarr M J, Anderson A W, Skudlarski P, Gore J C 1999 Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience* **2**: 568–73

Hoffman E A, Haxby J V 2000 Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience* **3**: 80–4

Humphreys G W, Riddoch M J 1987 On telling your fruit from your vegetables: A consideration of category-specific deficits after brain damage *Trends in Neurosciences* **10**: 145–48

Ishai A, Ungerleider L G, Haxby J V 2000 Distributed neural systems for the generation of visual images. *Neuron* **28**: 979–90

Ishai A, Ungerleider L G, Martin A, Schouten J L, Haxby J V 1999 Distributed representation of objects in the ventral visual pathway. *Proceedings of the National Academy of Sciences, USA* **96**: 9379–84

Jeannerod M, Arbib M A, Rizzolatti G, Sakata H 1995 Grasping objects: The cortical mechanisms of visuomotor transformation. *Trends in Neurosciences* **18**: 314–20

Kanwisher N, Downing P, Epstein R, Kourtzi Z 2001 Functional neuroimaging of visual recognition. In: Cabeza R, Kingstone A (eds.) *Handbook of Functional Neuroimaging of Cognition*. MIT Press, Cambridge, MA

Kourtzi, Z, Kanwisher N 2000 Activation in human MT/MST by static images with implied motion. *Journal of Cognitive Neuroscience* **12**: 48–55

Kreiman G, Koch C, Fried I 2000a Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience* **3**: 946–53

Kreiman G, Koch C, Fried I 2000b Imagery neurons in the human brain. *Nature* **408**: 357–61

Lissauer H. 1988 [1890] A case of visual agnosia with a contribution to theory. *Cognitive Neuropsychology* **5**: 157–92

Martin A 2001 Functional neuroimaging of semantic memory. In: Cabeza R, Kingstone A (eds.) *Handbook of Functional Neuroimaging of Cognition*. MIT Press, Cambridge, MA

Martin A, Chao L L 2001 Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology*. **11**: 194–201

Martin A, Haxby J V, Lalonde F M, Wiggs C L, Ungerleider L G 1995 Discrete cortical regions associated with knowledge of color and knowledge of action. *Science* **270**: 102–5

Martin A, Wiggs C L, Ungerleider L G, Haxby J V 1996 Neural correlates of category-specific knowledge. *Nature* **379**: 649–52

Miller G A 1999 On knowing a word. *Annual Review of Psychology* **50**: 1–19

O'Craven K M, Kanwisher N 2000 Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience* **6**: 1013–23

Patterson K, Hodges J R 1995 Disorders of semantic memory. In: Baddeley A D, Wilson B A, Watts F N (eds.) *Handbook of Memory Disorders*. Wiley, New York

Perani D, Cappa S F, Bettinardi V, Bressi S, Gorno-Tempini M, Matarrese M, Fazio F 1995 Different neural systems for the recognition of animals and man-made tools. *Neuroreport* **6**: 1637–41

Price C J, Indefrey P, van Turennout M 1999 The neural architecture underlying the processing of written and spoken word forms. In: Brown C M, Hagoort P (eds.) *Neurocognition of Language Processing*. Oxford University Press, New York

Senior C, Barnes J, Giampietro V, Simmons A, Bullmore E T, Brammer M, David A S 2000 The functional neuro-anatomy of implicit motion perception of 'representational momentum'. *Current Biology* **10**: 16–22

Smith E E, Patalano A L, Jonides J 1999 Alternative strategies of categorization. *Cognition* **65**: 167–96

Tranel D, Damasio H, Damasio A R 1997 A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia* **35**: 1319–27

Tulving E 1983 *Elements of Episodic Memory*. Oxford University Press, New York

Vandenberghe R, Price C, Wise R, Josephs O, Frackowiak R S 1996 Functional anatomy of a common semantic system for words and pictures. *Nature* **383**: 254–6

Warrington E K 1975 The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology* **27**: 635–57

Warrington E K, Shallice T 1984 Category specific semantic impairments. *Brain* **107**: 829–54

Webster's New World Dictionary 1988 3rd college edn. Simon and Schuster, New York

A. Martin

Semantic Processing: Statistical Approaches

The development of methods for representing meaning is a critical aspect of cognitive modeling and of applications that must extract meaning from text input. This ability to derive meaning is the key to any

approach that needs to use or evaluate knowledge. Nevertheless, determining how meaning is represented and how information can be converted to this representation is a difficult task. For example, any theory of meaning must describe how the meaning of each individual concept is specified and how relationships among concepts may be measured. Thus, appropriate representations of meaning must first be developed. Second, computational techniques must be available to permit the derivation and modeling of meaning using these representations. Finally, some form of information about the concepts must be available in order to permit a computational technique to derive the meaning of the concepts. This information about the concepts can either be more structured human-based input, for example dictionaries or links among related concepts, or less structured natural language.

With the advent of more powerful computing and the availability of on-line texts and machine-readable dictionaries, novel techniques have been developed that can automatically derive semantic representations. These techniques capture effects of regularities inherent in language to learn about semantic relationships among words. Other techniques have relied on hand-coding of semantic information, which is then placed in an electronic database so that users may still apply statistical analyses on the words in the database. All of these techniques can be incorporated into methods for modeling a wide range of psychological phenomena such as language acquisition, discourse processing, and memory. In addition, the techniques can be used in applied settings in which a computer can derive semantic knowledge representations from text. These settings include information retrieval, natural language processing, and discourse analysis.

1. *The Representation of Meaning*

Definitions of semantics generally encompass the concepts of knowledge of the world, of meanings of words, and of relationships among the words. Thus, techniques that perform analyses of semantics must have representations that can account for these different elements. Because these computational techniques rely on natural language, typically in the form of electronic text, they must further be able to convert information contained in the text to the appropriate representation. Two representational systems that are widely used in implementations of computational techniques are feature systems and semantic networks. While other methods of semantic representation (e.g., schemas) also account for semantic information in psychological models, they are not as easily specified computationally.

1.1 *Feature Systems*

The primary assumption behind feature systems is that if a fixed number of basic semantic features can be discovered, then semantic concepts can be defined by their combination, or composition of these features. Thus, different concepts will have different basic semantic features or levels of these features. Concepts vary in their relatedness to each other, based on the degree to which they have the same semantic features (e.g., Smith and Medin 1981). One approach to define empirically the meaning of words using a feature system was the use of the semantic differential (Osgood et al. 1957). By having people rate how words fell on a Likert scale of different bipolar adjectives (e.g., fair–unfair, active–passive), a word could be represented as the combination of ratings. Through collecting large numbers of these ratings, Osgood could define a multidimensional semantic space in which the meaning of a word is represented as a point in that space based on the ratings on each scale, and words could be compared among each through measuring distances in the space.

Feature representations have been widely used in psychological models of memory and categorization. They provide a simple representation for encoding information for input into computational models, most particularly connectionist models (e.g., Rumelhart and McClelland 1986). While the representation permits assigning concepts based on their features, it does not explicitly show relations among the concepts. In addition, in many models using feature representations, the dimensions are hand-created; therefore, it is not clear whether the features are based on real-world constructions, or are derived psychological constructions developed on the fly, based on the appropriate context. Some of the statistical techniques described below avoid the problem of using humanly defined dimensions by automatically extracting relevant dimensions.

1.2 *Semantic or Associative Networks*

A semantic network approach views the meaning of concepts as being determined by their relations to other concepts. Concepts are represented as nodes with labeled links (e.g., IS-A or Part-of) as relationships among the nodes. Thus, knowledge is a combination of information about concepts and how those concepts relate to each other. Based on the idea that activation can spread from one node to another, semantic networks have been quite influential in the development of models of memory. Semantic networks and spreading activation have been widely used for modeling sentence verification times and priming, and have been incorporated into many localist connectionist models.

Semantic networks permit economies of storage because concepts can inherit properties shared by other concepts (e.g., Collins and Quillian 1969). This basis has made semantic networks a popular approach for the development of computer-based lexicons, most particularly within the field of artificial intelligence.

Nevertheless, many of the assumptions for relatedness among concepts must still be based on handcrafted networks in which the creator of the network uses their knowledge to develop the concepts and links.

2. *Statistical and Human Approaches to Deriving Meaning*

No matter what type of representation is used for meaning, some form of information must be stored within the representation for it to be useful. There are two primary techniques to derive representations and fill in the lexical information. The first is to handcraft the information through human judgments about definitions, categorization, organization, or associations among concepts. The second is to use automated methods to extract information from existing on-line texts. The former approach relies on human expertise, and it is assumed that humans' introspective abilities and domain knowledge will provide an accurate representation of lexical relations. The latter approach assumes that the techniques can create a useful and accurate representation of meaning from processing natural language input.

2.1 *Human Extraction of Meaning for Knowledge Bases*

Long before the existence of computing, humans developed techniques for recording the meaning of words. These lexical compilations include dictionaries, thesauri, and ontologies. While not strictly statistical techniques for deriving meaning, human-based methods deserve mention for several reasons. Many of these lexical compilations are now available on-line. For example, there are a number of machine readable dictionaries on-line (e.g., Wilks et al. 1996), as well as projects to develop large ontologies of domain knowledge. Because of this availability, statistical techniques can be applied to these on-line lexicons to extract novel information automatically from a lexicon, for example to discover new semantic relations. In addition, the information from existing lexical entries can be used in automatic techniques for categorizing new lexical items.

One notable approach has been the development of WordNet (see Fellbaum 1998). WordNet is a handbuilt on-line lexical reference database which represents both the forms and meanings of words. Lexical concepts are organized in synonym sets (synsets) which represent semantic and lexical relations among concepts including synonymy, antonymy, hyponymy, meronymy, and morphological relations. Automated techniques (described in Fellbaum 1998) have been applied to WordNet for a number of applications including discovering new relations among words, performing automated word-sense identification, and applying it to information retrieval through automated query expansion.

Along with machine-readable dictionaries and ontologies, additional human-based approaches to deriving word similarity have collected large numbers of word associations to derive word-association norms (e.g., Deese 1965) and ratings of words on different dimensions (e.g., Osgood et al. 1957). The statistics from these collections are incorporated into computational cognitive models. In human-generated representations, however, there is much overhead involved in collecting the set of information before the lexical representation can be used. Because of this, it is not easily adapted to new languages or new domains. Further, handcrafted derivations of relationships among words do not provide a basis for a representational theory.

2.2 *Automatic Techniques for Deriving Semantics*

'You shall know a word by the company it keeps' (Firth 1957). The context in which any individual word is used can provide information about both the word's syntactic role and the semantic contributions of the word to the context. Thus, with appropriate techniques for measuring the use of words in language, we should be able to infer the meaning of those words. While much research has focused on automatically extracting syntactic regularities from language, there has been a recent increase in research on approaches to extracting semantic information. This research takes a structural linguistic approach in its assumption that the structure of meaning (or of language in general) can be derived or approximated through distributional measures and statistical analyses.

In order automatically to derive a semantic representation from analyzing language, several assumptions must be fulfilled. First it is assumed that information about the cooccurrence of words within contexts will provide appropriate information about semantic relationships. For example, the fact that 'house' and 'roof' often occur in the same context or in similar contexts informs that there must be a relationship between them. Second, assumptions must be made about what constitutes the context in which words appear. Some techniques use a moving window which moves across the text analyzing five to 10 words as a context; others use sentences, paragraphs, or complete documents as the complete context in which words appear. Finally, corpora of hundreds of thousands to millions of words of running text are required so that there are enough occurrences of information about how the different words appear within their contexts.

A mathematical overview of many statistical approaches to natural language may be found in Manning and Schütze (1999), and a review of statistical techniques applied to corpora may be found in

Boguraev and Pustejovsky (1996). Below, we focus on a few techniques that have been applied to psychological modeling, have shown psychological plausibility, and/or have provided applications that may be used more directly within cognitive science. The models described all use feature representations of words, in which words are represented as vectors of features. In addition, the models automatically derive the feature dimensions rather than having them predefined by the researcher.

2.2.1 The HAL model. The HAL (Hyperspace Analog to Memory) model uses lexical cooccurrence to develop a high-dimensional semantic representation of words (Burgess and Lund 1999). Using a large corpus (320 million words) of naturally occurring text, they derive vector representations of words based on a 10-word moving window. Vectors for words that are used in related contexts (within the same window) have high similarity. This vector representation can characterize a variety of semantic and grammatical features of words, and has been used to investigate a wide range of cognitive phenomena. For example, HAL has been used for modeling results from priming and categorization studies, resolving semantic ambiguity, modeling cerebral asymmetries in semantic representations, and modeling semantic differences among word classes, such as concrete and abstract nouns.

2.2.2 Latent Semantic Analysis. Like HAL, Latent Semantic Analysis (LSA) derives a high-dimensional vector representation based on analyses of large corpora (Landauer and Dumais 1997). However, LSA uses a fixed window of context (e.g., the paragraph level) to perform an analysis of cooccurrence across the corpus. A factor analytic technique (singular value decomposition) is then applied to the cooccurrence matrix in order to derive a reduced set of dimensions (typically 300 to 500). This dimension reduction causes words that are used in similar contexts, even if they are not used in the same context to have similar vectors. For example, although 'house' and 'home' both tend to occur with 'roof,' they seldom occur together in language. Nevertheless, they would have similar vector representations in LSA. In LSA, vectors for individual words can be summed to provide measures of the meaning of larger units of text. Thus, the meaning of a paragraph would be the sum of the vectors of the words in that paragraph. This basis permits the comparison of larger units of text, such as comparing the meaning of sentences, paragraphs, or whole documents to each other.

LSA has been applied to a number of different corpora, ranging from large samples of language that children and adults would have encountered to specific corpora on particular domains, such as individual course topics. For the more general corpora, it derives a generalized semantic representation of knowledge similar to that general knowledge acquired by people in life. For the domain-specific corpora, it generates a representation more similar to that of people knowledgeable within that domain area.

LSA has been used both as a theoretical model and as a tool for the characterization of semantic relatedness of units of language (see Landauer et al. 1998 for a review). As a theoretical model, LSA has been used to model the speed of acquisition of new words by children, its scores overlap those of humans on standard vocabulary and subject matter tests, it mimics human word sorting and category judgments, it simulates word-word and passage-word lexical priming data, and it accurately estimates textual coherence and the learnability of texts by individual students. The vector representation in LSA can be applied within other theoretical models. For example, propositional representations based on LSA-derived vectors have been integrated into the Construction-Integration model, a symbolic connectionist model of language (see Kintsch 1998). As an application, LSA has been used to measure the quality and quantity of knowledge contained in essays, for matching user queries to documents in information retrieval, and for performing automatic discourse segmentation.

2.2.3 Connectionist approaches. Connectionist modeling uses a network of interacting processing units operating on feature vectors to model cognitive phenomena. It has been widely used to model aspects of language processing. Although in some connectionist models words or concepts are represented as vectors in which the features have been predefined (e.g., McClelland and Kawamoto 1986), recent models have automatically derived the representation. Elman (1990) implemented a simple recurrent network that used a moving window analyzing a set of sentences from a small lexicon and artificial grammar. Based on a cluster analysis of the activation values of the hidden units, the model could predict syntactic and semantic distinctions in the language, and was able to discover lexical classes based on word order. One current limitation, however, is that it is not clear how well the approach can scale up to much larger corpora. Nevertheless, like LSA, due to the constraint satisfaction in connectionist models, the pattern of activation represented in the hidden units goes beyond direct cooccurrence, and captures more of the contextual usage of words.

3. Conclusions

Statistical techniques for extracting meaning from online texts and for extending the use of machine-readable dictionaries have become viable approaches

for creating semantic-based models and applications. The techniques go beyond modeling just cooccurrence of words. For example, the singular value decomposition in LSA or the use of hidden units in connectionist models permits derivation of semantic similarities that are not found in local cooccurrence but that are seen in human knowledge representations. The techniques incorporate both the idea of feature vectors, from feature-based models, and the idea that words can be defined by their relationships to other words found in semantic networks.

3.1 Advantages and Disadvantages of Statistical Semantic Approaches

Compared to many models of semantic memory, statistical semantic approaches are quite parsimonious, using very few assumptions and parameters to derive an effective representation of meaning. They further avoid problems of human-based meaning extraction since the techniques can process realistic environmental input (natural language) directly into a representation. The techniques are fast, requiring only hours or days to develop new lexicons, and can work in any language. They typically need large amounts of natural language to derive representation; thus, large corpora must be obtained. Nevertheless, because they are applied to large corpora, the lexicons that are developed provide realistic representations of tens to hundreds of thousands of words in a language.

3.2 Theoretical and Applied Uses of Statistical Semantics

Within cognitive modeling, statistical semantic techniques can be applied to almost any model that must incorporate meaning. They can therefore be used in modeling in such areas as semantic and associative priming, lexical ambiguity resolution, anaphoric resolution, acquisition of language, categorization, lexical effects on word recognition, and higher-level discourse processing. The techniques can provide useful additions to a wide range of applications that must encode or model meaning in language; for example, information retrieval, automated message understanding, machine translation, and discourse analysis.

3.3 The Future of Statistical Semantic Models in Psychology

While it is important that the techniques provide effective representations for applications, it is also important that the techniques have psychological plausibility. The study of human language processing can help inform the development of more effective methods of deriving and representing semantics. In turn, the development of the techniques can help improve cognitive models. For this to happen, strong ties must be formed between linguists, computational experts, and psychologists. In addition, the techniques and lexicons derived from them are not widely available to all researchers. For the techniques to succeed, better distribution, either through Web interfaces or through software, will allow them to be more easily incorporated into a wider range of cognitive models.

See also: Connectionist Models of Language Processing; Lexical Access, Cognitive Psychology of; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Lexical Semantics; Lexicon; Memory for Meaning and Surface Memory; Semantic Knowledge: Neural Basis of; Semantic Similarity, Cognitive Psychology of; Semantics; Word Meaning: Psychological Aspects

Bibliography

Boguraev B, Pustejovsky J 1996 *Corpus Processing for Lexical Acquisition*. MIT Press, Cambridge, MA

Burgess C, Lund K 1999 The dynamics of meaning in memory. In: Dietrich E, Markman A B (eds.) *Cognitive Dynamics: Conceptual and Representational Change in Humans and Machines*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 117–56

Collins A M, Quillian M R 1969 Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* **8**: 240–7

Deese J 1965 *The Structure of Associations in Language and Thought*. Johns Hopkins University Press, Baltimore, MD

Elman J L 1990 Finding structure in time. *Cognitive Science* **14**: 179–211

Fellbaum C 1998 *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA

Firth J R 1968 A synopsis of linguistic theory 1930–1955. In: Palmer F (ed.) *Selected Papers of J.R. Firth*. Longman, New York, pp. 32–52

Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, New York

Landauer T K, Dumais S T 1997 A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* **104**: 211–40

Landauer T K, Foltz P W, Laham D 1998 An introduction to Latent Semantic Analysis. *Discourse Processes* **25**: 259–84

Manning C D, Schütze H 1999 *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA

McClelland J L, Kawamoto A H 1986 Mechanisms of sentence processing: Assigning roles to constituents. In: Rumelhart D E, McClelland J L (eds.) PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol. 2*. MIT Press, Cambridge, MA, pp. 272–325

Osgood C E, Suci G J, Tannenbaum P H 1957 *The Measurement of Meaning*. University of Illinois Press, Urbana, IL

Rumelhart D E, McClelland J L 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA, Vol. 1

Smith E E, Medin D L 1981 *Categories and Concepts*. Harvard University Press, Cambridge, MA

Wilks Y, Slator B, Guthrie L 1996 *Electric Words: Dictionaries, Computers and Meanings*. MIT Press, Cambridge, MA

P. W. Foltz

Semantic Similarity, Cognitive Psychology of

Semantic similarity refers to similarity based on meaning as opposed to form. The term is used widely throughout cognitive psychology and related areas such as psycholinguistics, memory, reasoning, and neuropsychology. For example, semantically similar words are confusable with each other, and can prime each other, with the consequence that verbal memory performance is heavily dependent on similarity of meaning (see *False Memories, Psychology of*; *Priming, Cognitive Psychology of*). In the context of reasoning, people draw inductive inferences on the basis of semantic similarity, for example, inferring that properties of cows are more likely to be true of horses than to be true of hedgehogs (Heit 2001)

But what is semantic similarity? In experimental work, semantic similarity is often estimated directly from participants' explicit judgments. However, there are several advantages to representing semantic similarity within a standardized framework or model. These advantages include the greater consistency and other known mathematical properties that can be imposed by a formal model, and the possible benefits of data reduction and the extraction of meaningful dimensions or features of the stimuli being analyzed. The history of research concerning semantic similarity can be captured in terms of the succession of models that have been applied to this topic. The models themselves can be divided into generic models of similarity and a number of more specialized models aimed at particular aspects of semantic similarity.

1. General Models of Similarity

Although the following accounts are meant to address similarity in general terms, they can be readily applied to semantic similarity. The two classical, generic approaches to modeling similarity are spatial models of similarity and Tversky's (1977) contrast model. More recently, structured approaches have addressed limitations of both of these classical accounts.

1.1 Spatial Representations

Spatial models seek to represent similarity in terms of distance in a psychological space (Shepard 1980). An item's position is determined through its coordinate values along the relevant dimensions; nearby points thus represent similar items, whereas distant items are psychologically very different. For example, a spatial representation of mammal categories might represent different animals on dimensions of size and ferocity, with lions and tigers being close on both of these dimensions (Fig. 1). The relevant space is derived using the technique of multidimensional scaling (MDS), a statistical procedure for dimensionality reduction. MDS works from experimental participants' judgments, typically in matrices of proximity data such as pairwise confusions or similarity ratings between items. Spatial models have been used widely for visualization purposes and as the heart of detailed cognitive models, e.g., of categorization and recognition memory (Nosofsky 1991). Less widely used are related statistical procedures such as hierarchical clustering (Shepard 1980) although they have also been used in the context of semantic similarity.

1.2 The Contrast Model

The traditional alternative to spatial models is Tversky's (1977) contrast model. This account was developed to address perceived limitations of spatial models which bring these into conflict with behavioral data (but also see Nosofsky 1991). Chief among these are violations of the so-called metric axioms that underlie any spatial scheme, such as asymmetries in human similarity judgments that are not well captured by spatial distance which should be symmetrical. For example, people judged the similarity of North Korea to China to be greater than the similarity of China to North Korea. In the contrast model (Eqn. (1)), the similarity between items a and b is a positive function of the features common to both items and a negative function of the distinctive features of a and also the distinctive features of b. Each of these three feature sets is governed by a weighting parameter which allows the model to capture asymmetries according to the nature of a particular task. According to the focusing hypothesis, greater attention is given to distinctive features of the first item in a comparison than of the second item, hence $\alpha > \beta$. When China is more familiar than North Korea, having more known distinctive features, then the similarity from China to North Korea should be lower than the similarity of North Korea to China.

$$S(a, b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \quad (1)$$

Although applications of the contrast model to the modeling of specific cognitive tasks are fewer than those of spatial models, an application to semantic similarity can be found in Ortony (1979), who applied the model to distinctions between literal and metaphorical similarity.

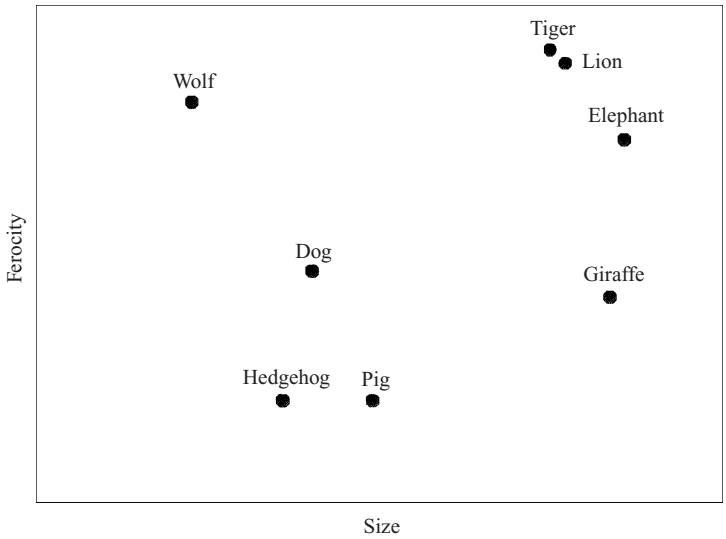

Figure 1
Multidimensional scaling representation of mammals

1.3 Structured Representations

Despite the efforts aimed at establishing the superiority of either the contrast model or spatial models, it has been argued that both approaches share a fundamental limitation in that they define similarity based on oversimplified kinds of representations: points in space or feature sets. Arguably most theories of the representation of natural objects, visual textures, sentences, etc., assume that these cannot be represented in line with these restrictions (see *Feature Representations in Cognitive Psychology*). Instead, they seem to require structured representations: complex representations of objects, their parts and properties, and the interrelationships between them. Descriptions such as SIBLING-OF (Linus, Lucy) and SIBLING-OF (Lucy, Linus) cannot be translated in an obvious way to either lists of features or points in space that represent the similarity between these two items as well as their differences from BROTHER-OF (Linus, Lucy). (See also Fodor and Pylyshyn (1988) for a critique of attempts in connectionist networks to represent relational structure in a featural framework.)

The perceived need for accounts of similarity to work with structured representations has given rise to the structural alignment account (see Markman (2001) for an overview) which has its roots in research on analogical reasoning (see *Mental Models*, *Psychology of*). Structural alignment operates over structured representations such as frames (see *Schemas*, *Frames*, *and Scripts in Cognitive Psychology*) consisting of slots and fillers. The comparison process requires that at least some of the predicates, that is relations such as ABOVE (x, y), are identical across the comparison. These identical predicates are placed in correspondence. The alignment process then seeks to build maximal structurally consistent matches between the two representations.

Structural alignment has been implemented in a variety of computational models (e.g., Falkenhainer et al. 1990, Goldstone and Medin 1994) that have been used to capture behavioral data, especially similarity judgments and analogical reasoning. Experimental results have supported an important prediction of the structural alignment account, that there will be a greater impact of alignable differences compared with nonanalignable differences. Nonalignable differences between two representations are elements of one object that have no correspondence in the other. In contrast, alignable differences refer to representational elements that have corresponding roles in the two representations but fail to match exactly. For example, imagine two tables, one with a flower on top and the other with a bowl on top in addition to a chair beneath it. Comparing the two scenes, flower vs. bowl would be an alignable difference, whereas the chair would be a nonalignable difference.

2. Specialized Models of Semantic Similarity

2.1 Semantic Differentials

Leaving behind generic models of similarity, semantic similarity has also been captured through a variety of specially built models and approaches. The first of these is Osgood et al.'s (1957) semantic differential.

This approach used psychometric techniques to compute the psychological distance between concepts. Participants were required to rate concepts on 10–20 bipolar scales for a set of semantically relevant dimensions, such as positive–negative, feminine–masculine. The ratings of words on these dimensions would essentially form their coordinates in semantic space. The approach is thus related to spatial models of similarity with the difference primarily in the way that the semantic space is derived. Whereas models based on multidimensional scaling require pairwise proximity data for all concepts of interest, the semantic differential approach requires that all concepts of interest are rated on all relevant dimensions. Therefore, the semantic differential method suffers from the difficulty that the relevant dimensions must be stipulated in advance. Still, semantic differentials have been widely used, largely because the data are straightforward to collect, analyze, and interpret.

2.2 Semantic Networks and Featural Models

Osgood et al.'s work motivated the semantic feature model of Smith et al. (1974) which addressed how people verify statements such as 'a robin is a bird' and 'a chicken is a bird,' considering the defining and characteristic features of these concepts. Verifying the latter statement would be slower owing different characteristic features for chicken vs. bird. The semantic feature model was developed in contrast to Collins and Quillian's (1969) semantic network model. In this model semantic meaning is captured by nodes that correspond to individual concepts. These nodes are connected by a variety of links representing the nature of the relationship between the nodes (such as an IS-A link between robin and bird). Semantic similarity (or distance) is captured in terms of the number of links that must be traversed to reach one concept from another. Although the semantic network model was a groundbreaking attempt at capturing structured relations, it still suffered from some problems. For example, the model correctly predicts that verifying 'a robin is an animal' is slower than verifying 'a robin is a bird' owing to more links being traversed for the first statement. However, this model does not capture typicality effects such as the difference between robin and chicken.

2.3 High-dimensional Context Spaces

Osgood et al.'s semantic differential can also be seen as a predecessor of contemporary corpus-derived measures of semantics and semantic similarity. For example, Burgess and Lund's (1997) hyperspace analogue to language model (HAL) learns a high-dimensional context from large-scale linguistic corpora that encompass many millions of words of speech or text. The model tracks lexical co-occurrences throughout the corpus and from these derives a high-dimensional representational space. The meaning of a word is conceived of as a vector. Each element of this vector corresponds to another word in the model, with the value of an element representing the number of times that the two words co-occurred within the discourse samples that constitute the corpus. For example, the vector for dog will contain an element reflecting the number of times that the word 'bone' was found within a given range of words in the corpus. These vectors can be viewed as the coordinates of points (individual words) in a high-dimensional semantic space. Semantic similarity is then a matter of distance between points in this space.

Several other such usage-based models have been proposed to date; similar in spirit to HAL, for example, is Landauer and Dumais's (1997) latent semantic analysis (see also *Semantic Processing*: *Statistical Approaches*). The basic approach might be seen to be taking to its logical consequence Wittgensteir's famous adage of 'meaning as use.' Its prime advantage over related approaches such as spatial models of similarity and the semantic differential lies in the ability to derive semantics and thus measures of semantic similarity for arbitrarily large numbers of words without the need for any especially collected behavioral data. The ability of these models to capture a wide variety of phenomena, such as results in semantic priming and effects of semantic context on syntactic processing, has been impressive.

2.4 Connectionist Approaches

The final approach to semantic similarity to be discussed shares with these context-based models a statistical orientation, but connectionist modeling has been popular particularly in neuropsychological work on language and language processing. In connectionist models, the semantics of words are represented as patterns of activations, or banks of units representing individual semantic features. Semantic similarity is then simply the amount of overlap between different patterns, hence these models are related to the spatial accounts of similarity. However, the typically nonlinear activation functions used in these models allow virtually arbitrary re-representations of such basic similarities. The representation schemes utilized in these models tend to be handcrafted rather than derived empirically as in other schemes such as multidimensional scaling and high-dimensional context spaces. However, it is often only very general properties of these semantic representations and the similarities between them that are crucial to a model's behavior, such as whether these representations are 'dense' (i.e., involve the activation of many semantic features) or 'sparse,' so that the actual semantic features chosen are not crucial. For example, this distinction between dense and sparse representation has been used to capture patterns of semantic errors

associated with acquired reading disorders (Plaut and Shallice 1993) and also patterns of category specific deficits following localized brain damage (Farah and McClelland 1991).

3. Conclusion

There is a wide variety of models of semantic similarity available. Underlying this array of approaches is a fundamental tension that is as yet unresolved. Many of the models reviewed can be classed as loosely spatial in nature. These models have all been applied extensively to behavioral data, yet there are fundamental limitations to spatial approaches with respect to their representational capacity, made clear by the successes of the contrast model and the structural alignment approach. The behavioral evidence that relational structure is important to similarity seems particularly compelling. This leaves this area of research with two contrasting and seemingly incompatible strands of models, each of which successfully relates to experimental data. It is possible only to speculate on how this contradiction might ultimately be resolved. One possibility is that spatial models and structured representations capture different aspects or types of semantic similarity: one automatic, precompiled, effortless, and in some ways shallow, and one the result of more in-depth, or line, or metacognitive processing. Such a distinction between types of processing of similarity would be compatible with the success of models based on, for example, context spaces in capturing phenomena such as semantic priming, and the success of models such as those based on structural alignment in capturing phenomena such as analogy, or the understanding of novel concept combinations. Whether such a distinction will take shape or whether there will one day be a single, all-encompassing account remains for future research.

See also: Categorization and Similarity Models; Categorization and Similarity Models: Neuroscience Applications; Dementia, Semantic; Lexical Semantics; Semantic Knowledge: Neural Basis of; Semantic Processing: Statistical Approaches; Semantics

Bibliography

Burgess C, Lund K 1997 Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes* **12**: 177–210

Collins A M, Quillian M R 1969 Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* **8**: 240–7

Falkenhainer B, Forbus K D, Gentner D 1990 The structure-mapping engine: Algorithm and examples. *Artial Intelligence* **41**: 1–63

Farah M J, McClelland J L 1991 A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General* **120**: 339–57

Fodor J, Pylyshyn Z 1988 Connectionism and cognitive architecture: A critical analysis. *Cognition* **28**: 3–71

Goldstone R L, Medin D L 1994 The time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**: 29–50

Heit E 2001 Properties of inductive reasoning. *Psychonomic Bulletin and Review* **7**: 569–92

Landauer T K, Dumais S T 1997 A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**: 211–40

Markman A B 2001 Structural alignment, similarity, and the internal structure of category representations. In: Hahn U, Ramscar M (eds.) *Similarity and Categorization*. Oxford University Press, Oxford, UK

Nosofsky R M 1991 Stimulus bias, asymmetric similarity and classification. *Cognitive Psychology* **23**: 94–140

Ortony A 1979 Beyond literal similarity. *Psychological Review* **86**: 161–80

Osgood C E, Suci G J, Tannenbaum P H 1957 *The Measurement of Meaning*. University of Illinois Press, Urbana, IL

Plaut D C, Shallice T 1993 Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology* **10**: 377–500

Shepard R N 1980 Multidimensional scaling, tree-fitting, and clustering. *Science* **210**: 390–7

Smith E E, Shoben E J, Rips L J 1974 Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review* **81**: 214–41

Tversky A 1977 Features of similarity. *Psychological Review* **84**: 327–52

U. Hahn and E. Heit

Semantics

Semantics is the study of meaning communicated through language, and is usually taken to be one of the three main branches of linguistics, along with phonology, the study of sound systems, and grammar, which includes the study of word structure (morphology) and of sentence structure (syntax). This entry surveys some of the main topics of current semantics research.

1. Introduction

Traditionally, the main focus of linguistic semantics has been on word meaning, or lexical semantics. Since classical times writers have commented on the fact, noticed surely by most reflecting individuals, that the meaning of words changes over time. Such observa-

tions are the seeds of etymology, the study of the history of words. Over longer stretches of time, such changes become very obvious, especially in literate societies. Words seem to shift around: some narrow in meaning such as English 'queen,' which earlier meant 'woman, wife' but now means 'wife of a king.' Others become more general, while still others shift to take on new meaning or disappear altogether. Words are borrowed from language to language. The study of such processes is now part of historical semantics (Fisiak 1985). Another motivation for the study of word meaning comes from dictionary writers as they try to establish meaning correspondences between words in different languages, or, in monolingual dictionaries, seek to provide definitions for all the words of a language in terms of a simple core vocabulary. In lexicology similarities and differences in word meaning are a central concern.

The principled study of the meaning of phrases and sentences has only become established in linguistics relatively recently. Thus it is still common for descriptive grammars of individual languages to contain no separate section on semantics other than providing a lexicon. Nonetheless it has always been clear that one can identify semantic relations between sentences. Speakers of English know from the semantics of negation that nominal negation has different effects than sentence negation, so that 'No-one complained' may aptly be used to answer 'Who complained?,' while 'Someone did not complain' may not. Two sentences may seem to say essentially the same thing, even be paraphrases of each other, yet one may be more suited to one context than another–like the pair 'Bandits looted the train' and 'The train was looted by bandits.' A single sentence may be internally inconsistent, such as 'Today is now tomorrow,' or seem to be repetitive or redundant in meaning, such as 'A capital city is a capital city.' Another feature of sentence meaning is the regularity with which listeners draw inferences from sentences, and often take these to be part of the meaning of what was said. Some inferential links are very strong, such as entailment. Thus we say that 'Bob drank all of the beer' entails 'Bob drank some of the beer' (assuming the same individual Bob, beer, etc.), because it is hard to think of a situation where acceptance of the second sentence would not follow automatically from acceptance of the first. Other inferential links are weaker and more contextually dependent: from the utterance 'Bob drank some of the beer' it might be reasonable to infer 'Bob didn't drink all of the beer,' but it is possible to think of situations where this inference would not hold. We might say that a speaker of the first sentence is implying the second, in a certain context. Speakers of all languages regularly predict and use such inferential behavior to convey their meaning, such that often more meaning seems to be communicated than is explicitly stated. All these aspects of sentence meaning are under study in various semantic frameworks.

Semanticists share with philosophers an interest in key issues in the use of language, notably in reference. We use this term to describe the way in which speakers can pick out, or name, entities in the world by using words as symbols. Many scholars, especially formal semanticists, accept Frege's distinction between reference (in German, *Bedeutung*) and sense (*Sinn*); see Frege (1980). Reference is the act of identifying an entity (the referent) while sense is the means of doing so. Two different linguistic expressions such as 'the number after nine' and 'the number before eleven' differ in sense but they both share the same referent, 'ten.' For semanticists it is particularly interesting to study the various mechanisms that a language offers to speakers for this act of referring. These include names such as 'Dublin,' nouns such as 'cat,' which can be used to refer to a single individual, 'your cat,' or a whole class, 'Cats are carnivorous,' quantified nominals such as 'many cats,' 'some cats,' 'a few cats,' etc. Linguists as well as philosophers have to account for language's ability to allow us to refer to nonexistent and hypothetical referents such as 'World War Three,' 'the still undiscovered cure for cancer,' 'the end of the world.'

Semanticists also share interests with psychologists, for if sense is the meaning of an expression, it seems natural to many semanticists to equate it with a conceptual representation. Cognitive semanticists, in particular, (for example Lakoff 1987, Talmy 2000), but also some generative linguists (Jackendoff 1996), seek to explore the relationship between semantic structure and conceptual structure. One axis of the debate is whether words, for example, are simply labels for concepts, or whether there is a need for an independent semantic interface that isolates just grammatically relevant elements of conceptual structure. As Jackendoff (1996) points out, many languages make grammatical distinctions corresponding to the conceptual distinctions of gender and number, but few involve distinctions of colour or between different animal species. If certain aspects of concepts are more relevant to grammatical rules, as is also claimed by Pinker (1989), this may be justification for a semantic interface.

2. *Approaches to Meaning*

Even in these brief remarks we have had to touch on the crucial relationship between meaning and context. Language of course typically occurs in acts of communication, and linguists have to cope with the fact that utterances of the same words may communicate different meanings to different individuals in different contexts. One response to this problem is to hypothesize that linguistic units such as words, phrases, and sentences have an element of inherent meaning that does not vary across contexts. This is sometimes called inherent, or simply, sentence meaning. Language

users, for example speakers and listeners, then enrich this sentence meaning with contextual information to create the particular meaning the speaker means to convey at the specific time, which can then be called speaker meaning. One common way of reflecting this view is to divide the study of meaning into semantics, which becomes the study of sentence meaning, and pragmatics, which is then the study of speaker meaning, or how speakers use language in concrete situations. This is an attempt to deal with the tension between the relative predictability of language between fellow speakers and the great variability of individual interpretations in interactive contexts. One consequence of this approach is the view that the words that a speaker utters underdetermine their intended meaning.

Semantics as a branch of linguistics is marked by the theoretical fragmentation of the field as a whole. The distinction between formal and functional approaches, for example, is as marked in semantics as elsewhere. This is a large subject to broach here but see Givón (1995) and Newmeyer (1998) for characteristic and somewhat antagonistic views. One important difference is the attitude to the autonomy of levels of analysis. Are semantics and syntax best treated as autonomous areas of study, each with its own characteristic entities and processes? A related question at a more general level is whether linguistic processes can be described independently of general psychological processes or the study of social interaction. Scholars in different theoretical frameworks will give contradictory answers to these questions of micro- and macro-autonomy. Autonomy at both levels is characteristic of semantics within generative grammar; see, for example, Chomsky (1995). Functionalists such as Halliday (1996) and Harder (1996) would on the other hand argue against microautonomy, suggesting that grammatical relations and structure cannot be understood without reference to semantic function. They also seek motivation for linguistic structure in the dynamics of communicative interaction. A slightly different external mapping is characteristic of cognitive semantics, for example Lakoff (1987) and Langacker (1987), where semantic structures are correlated to conceptual structures.

Another dividing issue in semantics is the value of formal representations. Scholars are divided on whether our knowledge of semantics is sufficiently mature to support attempts at mathematical or other symbolic modeling; indeed, on whether such modeling serves any use in this area. Partee (1996), for example, defends the view of formal semanticists that the application of symbolic logic to natural languages, following in particular the work of Montague (1974), represents a great advance in semantic description. Jackendoff (1990), on the other hand, acknowledges the value of formalism in semantic theory and description but argues that formal logic is too narrow adequately to describe meaning in language. Other scholars, such as Wierzbicka (1992), view the search for formalism as premature and distracting.

There has been an explosive increase in the research in formal semantics since Montague's (1974) proposal that the analysis of formal languages could serve as the basis for the description of natural languages. Montague's original theory comprised a syntax for the natural language, say English, a syntax for the logical language into which English should be translated (intensional logic), rules for the translation, and rules for the semantic interpretation of the intensional logic. This and subsequent formal approaches are typically referential (or denotational) in that their emphasis is on the connection of language with a set of possible worlds, including the real, external world and the hypothetical worlds set up by speakers. Crucial to this correspondence is the notion of truth, defined at the sentence level. A sentence is true if it correctly describes a situation in some world. In this view, the meaning of a sentence is characterized by describing the conditions which must hold for it to be true. The central task for such approaches is to extend the formal language to cope with the semantic features of natural language while maintaining the rigor and precision of the methodology. See the papers in Lappin (1996) for typical research in this paradigm.

Research in cognitive semantics presents an alternative strategy. Cognitive semanticists reject what they see as the mathematical, antimentalist approach of formal semantics. In their view meaning is described by relating linguistic expressions to mental entities, conventionalized conceptual structures. These semanticists have proposed a number of conceptual structures and processes, many deriving from perception and bodily experience and, in particular, conceptual models of space. Proposals for underlying conceptual structures include image schemas (Johnson 1987), mental spaces (Fauconnier 1994), and conceptual spaces (Gärdenfors 1999). Another focus of interest is the processes for extending concepts, and here special attention is given to metaphor. Lakoff (1987) and Johnson (1987) have argued against the classical view of metaphor and metonymy as something outside normal language, added as a kind of stylistic ornament. For these writers metaphor is an essential element in our categorization of the world and our thinking processes. Cognitive semanticists have also investigated the conceptual processes which reveal the importance of the speaker's perspective and construal of a scene, including viewpoint shifting, figure-ground shifting, and profiling (Langacker 1987).

3. *Topics in Sentence Semantics*

Many of the semantic systems of language, for example tense (see *Tense, Aspect, and Mood, Linguistics of*), aspect, mood, and negation, are marked

grammatically on individual words such as verbs. However, they operate over the whole sentence. This 'localization' is the reason that descriptive grammars usually distribute semantic description over their analyses of grammatical forms. Such semantic systems offer the speaker a range of meaning distinctions through which to communicate a message. Theoretical semanticists attempt to characterize each system *qua* system, as in for example Verkuyl's (1993) work on aspect and Hornstein's (1990) work on tense. Typological linguists try to characterize the variation in such systems across the world's languages, as in the studies of tense and aspect by Comrie (1976, 1985), Binnick (1991), and Bybee et al. (1994). We can sketch some basic features of some of these systems.

3.1 Situation Type and Aspect

Situation type and aspect are terms for a language's resources that allow a speaker to describe the temporal 'shape' of events. The term situation type is used to describe the system encoded in the words of a language, while aspect is used for the grammatical systems which perform a similar role. To take one example, languages typically allow speakers to describe a situation either as static, as in 'The bananas are ripe,' or as dynamic, as in 'The bananas are ripening.' Here the state is the result of the process but the same situation can be viewed as more static or dynamic, as in 'The baby is asleep' and 'The baby is sleeping.' As these examples show, this distinction is lexically marked: in English, for example, adjectives are typically used for states, and verbs for dynamic situations. There are, however, a group of stative verbs, such as 'know,' 'understand,' 'love,' 'hate,' which describe static situation types. There are a number of semantic distinctions typically found amongst dynamic verbs, for example the telic/atelic (bounded/unbounded) distinction and the punctual/durative distinction. Telic verbs describe processes which are seen as having a natural completion, which atelic verbs do not. A telic example is 'Matthew was growing up,' and an atelic example is 'Matthew was drinking.' If these procsses are interrupted at any point, we can automatically say 'Matthew drank,' but not 'Matthew grew up.' However, atelic verbs can form telic phrases and sentences by combining with other grammatical elements, so that 'Matthew was drinking a pint of beer' is telic. Durative verbs, as the term suggests, describe processes that last for a period of time, while punctual describes those that seem so instantaneous that they have no detectable internal structure, as in the comparison between 'The man slept' and 'The light flashed.' As has often been observed, if an English punctual verb is used with a durative adverbial, the result is an iterative meaning, as in 'The light flashed all night,' where we understand the event to be repeated over the time mentioned.

Situation type typically interacts with aspect. Aspect is the grammatical system that allows the speaker choices in how to portray the internal temporal nature of a situation. An event, for example, may be viewed as closed and completed, as in 'Joan wrote a book,' or as an ongoing process, perhaps unfinished, as in 'Joan was writing a book.'

The latter verb form is described as being in the progressive aspect in English, but similar distinctions are very common in the languages of the world. In many languages we find described a distinction between perfective and imperfective aspects, used to describe complete versus incomplete events; see Bybee et al. (1994) for a survey. As mentioned above, aspect is intimately associated both with situation type and tense. In Classical Arabic the perfective is strongly associated with past tense (Comrie 1976, Binnick 1991). In English, for example, stative verbs are typically not used with progressive aspect, so that one may say 'I know some French' but not 'I am knowing some French.' Staying with the progressive, when it is used in the present tense in English (and in many other languages) it carries a meaning of proximate future or confident prediction as in 'We're driving to Los Angeles' or 'I'm leaving you.' The combination of the three semantic categories of tense, situation type, and aspect produces a complex system that allows speakers to make subtle distinctions in relating an event or describing a situation.

3.2 Modality

Modality is a semantic system that allows speakers to express varying attitudes to a proposition. Semanticists have traditionally identified two types of modality. One is termed epistemic modality, which encodes a speaker's commitment to, or belief in, a proposition, from the certainty of 'The ozone layer is shrinking' to the weaker commitments of 'The ozone layer may/might/could be shrinking.' The second is deontic modality, where the speaker signals a judgment toward social factors of obligation, responsibility, and permission, as in the various interpretations of 'You must/can/ may/ ought to borrow this book.' These examples show that similar markers, here auxiliary verbs, can be used for both types. When modality distinctions are marked by particular verbal forms, these are traditionally called moods. Thus many languages, including Classical Greek and Somali, have a verb form labeled the optative mood for expressing wishes and desires. Other markers of modality in English include verbs of propositional attitude, as in 'I know/believe/think/doubt/ that the ozone layer is shrinking,' and modal adjectives, as in 'It is certain/probable/likely/possible that the ozone layer is shrinking.'

A related semantic system is evidentiality, where a speaker communicates the basis or source for present-

ing a proposition. In English and many other languages this may be done by adding expressions like 'allegedly,' 'so I've heard,' 'they say,' etc., but certain languages mark such differences morphologically, as in Makah, a Nootkan language spoken in Washington State (Jacobsen 1986, p. 10):
wiki·caxaw: 'It's bad weather' (seen or experienced directly);
wiki·caxakpi·d: 'It looks like bad weather' (inference from physical evidence);
wiki·caxakqad/I: 'It sounds like bad weather' (on the evidence of hearing); and
wiki·caxakwa·d: 'I'm told there's bad weather' (quoting someone else).

3.3 Semantic Roles

This term describes the speaker's semantic repertoire for relating participants in a described event. One influential proposal in the semantics literature is that each language contains a set of semantic roles, the choice of which is partly determined by the lexical semantics of the verb selected by the speaker. A characteristic list of such roles is:

agent: the initiator of some action, capable of acting with volition;

patient: the entity undergoing the effect of some action, often undergoing some change in state;

theme: the entity which is moved by an action, or whose location is described;

experiencer: the entity which is aware of the action or state described by the predicate but which is not in control of the action or state;

beneficiary: the entity for whose benefit the action was performed;

instrument: the means by which an action is performed or something comes about;

location: the place in which something is situated or takes place;

goal: the entity towards which something moves;

recipient: the entity which receives something; and

source: the entity from which something moves.

In an example like 'Harry immobilized the tank with a broomstick,' the entity Harry is described as the agent, the tank as the patient, and the broomstick as the instrument. These roles have also variously been called deep semantic cases, thematic relations, participant roles, and thematic roles.

One concern is to explain the matching between semantic roles and grammatical relations. In many languages, as in the last example, there is a tendency for the subject of the sentence to correspond to the agent and for the direct object to correspond to a patient or theme; an instrument often occurs as a prepositional phrase. Certain verbs allow variations from this basic mapping, for example the we find with English verbs such as 'break': 'The boy broke the window with a stone' (subject = agent); 'The stone broke the window' (subject = instrument); 'The window broke' (subject = patient). Clearly verbs can be arranged into classes depending on the variations of mappings they allow, and not all English verbs pattern like 'break.' We can say 'The admiral watched the battle with a telescope,' but 'The telescope watched the battle' and 'The battle watched' sound decidedly odd.

From this literature emerges the claim that certain mappings are more natural or universal. One proposal is that, for example, there is an implicational hierarchy governing the mapping to subject, typically such as: agent > recipient/benefactive > theme/patient > instrument > location. In such a hierarchy each left element is more preferred than its right neighbor, so that moving rightward along the string gives us fewer expected subjects. The hierarchy also makes certain typological claims: if a language allows a certain semantic role to be subject, it will allow all those to its left. Thus if we find that a language allows the role instrument to be subject, we predict that it allows the roles to the left, but we do not know if it allows location subjects.

One further application of semantic roles is in lexical semantics, where the notion allows verbs to be classified by their semantic argument structure. Verbs are assigned semantic role templates or grids by which they may be sorted into natural classes. Thus, English has a class of transfer, or giving verbs, which in one type includes the verbs 'give,' 'lend,' 'supply,' 'pay,' 'donate,' 'contribute.' These verbs encode a view of the transfer from the perspective of the agent and may be assigned the pattern <*agent*, theme, recipient>, as in 'The committee donated aid to the famine victims.' A second subclass of these transfer verbs encodes the process from the perspective of the recipient. These verbs include 'receive,' 'accept,' 'borrow,' 'buy,' 'purchase,' 'rent,' 'hire,' and have the pattern <*recipient*, theme, source>, as in 'The victims received aid from the committee.'

3.4 Entailment, Presupposition, and Implication

These terms relate to types of information a hearer gains from an utterance but which are not stated directly by the speaker. These phenomena have received a lot of attention because they seem to straddle the putative divide between semantics and pragmatics described above, and because they reveal the dynamic and interactive nature of understanding the meaning of utterances. Entailment describes a relationship between sentences such that on the basis of one sentence, a hearer will accept a second, unstated sentence purely on the basis of the meaning of the first. Thus sentence A entails sentence B, if it is not possible to accept A but reject B. In this view a sentence such as 'I bought a dog today' entails 'I bought an animal today'; or 'President Kennedy was assassinated yesterday' entails 'President Kennedy is now dead.' Clearly these sentential relations depend on lexical relations: a speaker who understands the meaning of

the English word 'dog' knows that a dog is an animal; similarly the verb 'assassinate' necessarily involves the death of the unfortunate object argument. Entailment then is seen as a purely automatic process, involving no reasoning or deduction, but following from the hearer's linguistic knowledge. Entailment is amenable to characterization by truth conditions. A sentence is said to entail another if the truth of the first guarantees the truth of the second, and the falsity of the second guarantees the falsity of the first.

Presupposition, on the other hand, is a more complicated notion. In basic terms, the idea is simple enough: that a speaker communicates certain assumptions aside from the main message. A range of linguistic elements communicates these assumptions. Some, such as names, and definiteness markers such as the articles 'the' and 'my,' presuppose the existence of entities. Thus 'James Brown is in town' presupposes the existence of a person so called. Other elements have more specific presuppositions. A verb such as 'stop' presupposes a preexisting situation. So a sentence 'Christopher has stopped smoking' presupposes 'Christopher smoked.' If treated as a truth-conditional relation, presupposition is distinguished from entailment by the fact that it survives under negation: 'Christopher has not stopped smoking' still presupposes 'Christopher smoked,' but the sentence 'I didn't buy a dog today' does not entail 'I bought an animal today.'

There are a number of other differences between entailment and presupposition that cast doubts on the ability of a purely semantic, truth-conditional account of the latter. Presuppositions are notoriously context sensitive, for example. They may be cancelled without causing an anomaly: a hearer can reply 'Christopher hasn't stopped smoking, because he never smoked' to cancel the presupposition by what is sometimes called metalinguistic negation. This dependency on context has led some writers to propose that presupposition is a pragmatic notion, definable in terms of the set of background assumptions that the speaker assumes is shared in the conversation. See Beaver (1997) for discussion.

A third type of inference is Grice's conversational implicature (1975, 1978). This is an extremely context-sensitive type of inference which allow participants in a conversation to maintain coherence. So, given the invented exchange below,

A: Did you give Mary the book?
B: I haven't seen her yet.

It is reasonable for A to infer the answer 'no' to her question. Grice proposed that such inferences are routinely relied on by both speakers and hearers, and that this reliance is based on certain assumptions that hearers make about a speaker's conduct. Grice classified these into several different types, giving rise to different types of inference, or, from the speaker's point of view, what he termed implicatures. The four main maxims are called Quality, Quantity, Relevance, and Manner (Grice 1975, 1978). They amount to a claim that a listener will assume, unless there is evidence to the contrary, that a speaker will have calculated their utterance along a number of parameters: they will tell the truth, try to estimate what their audience knows, and package their material accordingly, have some idea of the current topic, and give some thought to their audience being able to understand them. In our example above, it is A's assumption that B's reply is intended to be relevant that allows the inference 'no.'

Implicature has three characteristics: first, that it is implied rather than said; second, that its existence is a result of the context i.e., the specific interaction. There is no guarantee that in other contexts 'I haven't seen her' will be used to communicate 'no.' Third, implicature is cancelable without causing a contradiction. Thus the implicature 'no' in our example can be cancelled if B adds the clause 'but I mailed it to her last week.'

These three notions—entailment, presupposition, and implicature—can all be seen as types of inference. They are all produced in conversation, and are taken by participants to be part of the meaning of what a speaker has said. They differ in a number of features and crucially in context sensitivity. The attempt to provide a unified analysis of them all is a challenge to semantic and pragmatic theories. See Sperber and Wilson (1995) for an attempt at such a unified approach.

4. *Future Developments*

Although semantics remains theoretically a very diverse field it is possible to detect some shared trends which seem likely to develop further. One is a move away from a static view of sentences in isolation, detached from the speaker/writer's act of communication, toward dynamic, discourse-based approaches. This has always been characteristic of functional approaches to meaning but has also been noticeable in formal approaches as they move away from their more philosophical origins. Among examples of this we might mention discourse representation theory (Kamp and Reyle 1993) and dynamic semantics (Groenendijk et al. 1996).

Another development which seems likely to continue is a closer integration with other disciplines in cognitive science. In particular, computational techniques seem certain to make further impact on a range of semantic inquiry, from lexicography to the modeling of questions and other forms of dialogue. A subfield of computational semantics has emerged and will continue to develop; see Rosner and Johnson (1992) for example.

See also: Etymology; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Lexical

Semantics; Lexicology and Lexicography; Lexicon; Semantic Knowledge: Neural Basis of; Semantic Processing: Statistical Approaches; Semantic Similarity, Cognitive Psychology of; Word Meaning: Psychological Aspects

Bibliography

Beaver D 1997 Presuppositions. In: Van Bentham J, Ter Meulen A (eds.). *Handbook of Logic and Language*. Elsevier, Amsterdam, pp. 939–1008

Binnick R I 1991 *Time and the Verb: A Guide to Tense and Aspect*. Oxford University Press, Oxford, UK

Bybee J, Perkins R, Pagliuca W 1994 *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press, Chicago

Chomsky N 1995 *The Minimalist Program*. MIT Press, Cambridge, MA

Comrie B 1976 *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge University Press, Cambridge, UK

Comrie B 1985 *Tense*. Cambridge University Press, Cambridge, UK

Fauconnier G 1994 *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge University Press, Cambridge, UK

Fisiak J (ed.) 1985 *Historical Semantics—Historical Word Formation*. Mouton de Gruyter, Berlin

Frege G 1980 *Translations from the Philosophical Writings of Gottlob Frege* [ed. Geach P, Black M]. Blackwell, Oxford, UK

Gärdenfors P 1999 Some tenets of cognitive semantics. In: Allwood J, Gärdenfors P (eds.). *Cognitive Semantics: Meaning and Cognition*. John Benjamins, Amsterdam, pp. 12–36

Givón T 1995 *Functionalism and Grammar*. John Benjamins, Amsterdam

Grice H P 1975 Logic and conversation. In: Cole P, Morgan J (eds.). *Syntax and Semantics, Speech Acts*. Academic Press, New York, Vol. 3 pp. 43–58

Grice H P 1978 Further notes on logic and conversation. In: Cole P (ed.). *Syntax and Semantics 9: Pragmatics*. Academic Press, New York, pp. 113–28

Groenendijk J, Stokhof M, Veltman F 1996 Coreference and modality. In: Lappin S (ed.). *The Handbook of Contemporary Semantic Theory*. Blackwell, Oxford, UK, pp. 179–214

Halliday M A K 1994 *An Introduction to Functional Grammar*, 2nd edn. Edward Arnold, London

Harder P 1996 *Functional Semantics: A Theory of Meaning, Structure and Tense in English*. Mouton de Gruyter, Berlin

Hornstein N 1990 *As Time Goes By: Tense and Universal Grammer*. MIT Press, Cambridge, MA

Jackendoff R 1990 *Semantic Structures*. MIT Press, Cambridge, MA

Jackendoff R 1996 Semantics and cognition. In: Lappin S (ed.). *The Handbook of Contemporary Semantic Theory*. Blackwell, Oxford, UK, pp. 539–60

Jacobsen W H Jr 1986 The heterogeneity of evidentials in Makah. In: Chafe W, Nichols J (eds.). *Evidentiality: The Linguistic Coding of Epistemology*. Ablex, Norwood, NJ, pp. 3–28

Johnson M 1987 *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press, Chicago

Kamp H, Reyle U 1993 *From Discourse to Logic*. Kluwer, Dordrecht, The Netherlands

Lakoff G 1987 *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, Chicago

Langacker R W 1987 *Foundations of Cognitive Grammar*. Stanford University Press, Stanford, CA

Lappin S (ed.) 1996 *The Handbook of Contemporary Semantic Theory*. Blackwell, Oxford, UK

Lehmann W P 1992 *Historical Linguistics*, 3rd edn. Routledge, London

Montague R 1974 *Formal Philosophy: Selected Papers of Richard Montague* [ed. Thomason R H]. Yale University Press, New Haven, CT

Newmeyer F J 1998 *Language Form and Language Function*. MIT Press, Cambridge, MA

Partee B H 1996 The development of formal semantics in linguistic theory. In: Lappin S (ed.). *The Handbook of Contemporary Semantic Theory*. Blackwell, Oxford, UK, pp. 11–38

Pinker S 1989 *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, MA

Rosner M, Johnson R (eds.) 1992 *Computational Linguistics and Formal Semantics*. Cambridge University Press, Cambridge, UK

Sperber D, Wilson D 1995 *Relevance: Communication and Cognition*, 2nd edn. Blackwell, Oxford, UK

Talmy L 2000 *Toward a Cognitive Semantics*. MIT Press, Cambridge, MA

Verkuyl H J 1993 *A Theory of Aspectuality: The Interaction Between Temporal and Atemporal Structure*. Cambridge University Press, Cambridge, UK

Wierzbicka A 1992 *Semantics, Culture, and Cognition. Universal Concepts in Culture-specific Configurations*. Oxford University Press, Oxford, UK

J. I. Saeed

Semiotics

Semiotics is an interdisciplinary field that studies 'the life of signs within society' (Saussure 1959, p. 16). While 'signs' most commonly refers to the elements of verbal language and other vehicles of communication, it also denotes any means of representing or knowing about an aspect of reality. As a result, semiotics has developed as a close cousin of such traditional disciplines as philosophy and psychology. In the social sciences and humanities, semiotics has become an influential approach to research on culture and communication particularly since the 1960s. This article describes the classical origins of semiotics, exemplifies its application to contemporary culture, and outlines its implications for the theory of science.

1. Origins: Logic and Linguistics

Two different senses of 'signs' can be traced in the works of Aristotle (Clarke 1990, p. 15) (see *Aristotle (384–322 BC)*). First, the mental impressions that

people have, are signs which represent certain objects in the world to them. Second, spoken and written expressions are signs with which people are able to represent and communicate a particular understanding of these objects to others. The first sense, of mental impressions, points to the classical understanding of signs, not as words or images, but as naturally occurring evidence that something is the case. For example, fever is a sign of illness, and clouds are a sign that it may rain, to the extent that these signs are interpreted by humans. The second sense of signs as means of communication points towards the distinction that came to inform most modern science between conventional signs, especially verbal language, and sense data, or natural signs. Modern natural scientists could be said to study natural signs with reference to specialized conventional signs.

Given the ambition of semiotics to address the fundamental conditions of human knowledge, it has been an important undercurrent in the history of science and ideas. The first explicit statement regarding natural signs and language as two varieties of one general category of signs came not from Aristotle, but from St. Augustine (c. AD 400). This position was elaborated throughout the medieval period with reference, in part, to the understanding of nature as 'God's Book' and an analogy to the Bible. However, it was not until the seventeenth century that the term semiotic emerged in John Locke's *An Essay concerning Human Understanding* (1690). Here, Locke proposed a science of signs in general, only to restrict his own focus to 'the most usual' signs, namely, verbal language and logic (Clarke 1990, p. 40).

1.1 Peircean Logic and Semiotics

Charles Sanders Peirce (1839–1914) was the first thinker to recover this undercurrent in an attempt to develop a general semiotic. He understood his theory of signs as a form of logic, which informed a comprehensive system for understanding the nature of being and of knowledge. The key to the system is Peirce's definition of the sign as having three aspects:

> A sign, or *representamen*, is something which stands to somebody for something in some respect or capacity. It addresses somebody, that is, creates in the mind of that person an equivalent sign, or perhaps a more developed sign. That sign which it creates I call the *interpretant* of the first sign. That sign stands for something, its *object*.

An important implication of this definition is that signs always serve to mediate between objects in the world, including social facts, and concepts in the mind. Contrary to various skepticist positions, from Antiquity to postmodernism, Peirce's point was not that signs are what we know, but how we come to know what we can justify saying that we know, in science

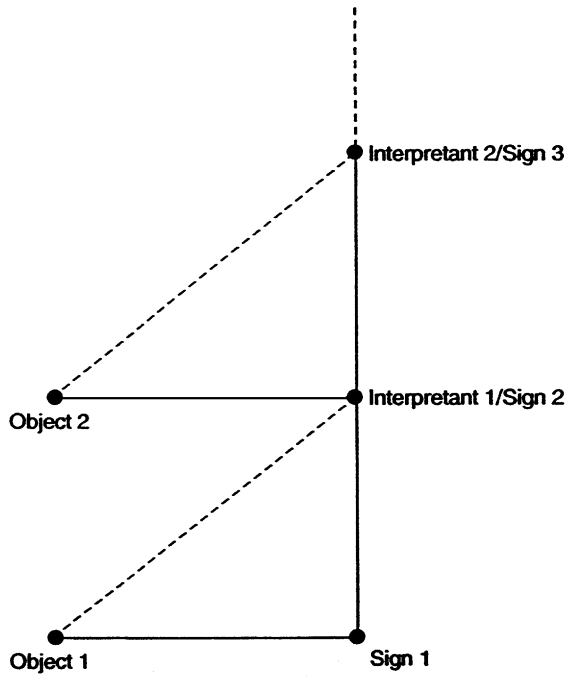

Figure 1
The process of semiosis

and in everyday life. Peircean semiotics married a classical, Aristotelian notion of realism with the modern, Kantian position (see *Kant, Immanuel (1724–1804)*) that humans necessarily construct their understanding of reality in the form of particular cognitive categories.

A further implication is that human understanding is not a singular internalization of reality, but a continuous process of interpretation, what is called semiosis. This is witnessed, for example, in the process of scientific discovery, but also in the ongoing coordination of social life. Figure 1 illustrates the process of semiosis, noting how any given interpretation (interpretant) itself serves as a sign in the next stage of an unending process which differentiates the understanding of objects in reality. Although Peirce's outlook was that of a logician and a natural scientist, semiosis can be taken to refer to the processes of communication by which cultures are maintained and societies reproduced and, to a degree, reformed.

One of the most influential elements of Peirce's semiotics outside logic and philosophy has been his categorization of different types of signs, particularly icon, index, and symbol. Icons relate to their objects through resemblance (e.g., a realistic painting); indices have a causal relation (e.g., fever as a symptom of illness); and symbols have an arbitrary relation to their object (e.g., words). Different disciplines and fields have relied on these types as analytical instruments in

order to describe the ways in which humans perceive and act upon reality.

1.2 Saussurean Linguistics and Semiology

Compared with Peirce, the other main figure in the development of semiotics, Ferdinand de Saussure (1857–1913), placed a specific focus on verbal language (see *Saussure, Ferdinand de (1857–1913)*), what Peirce had referred to as symbols. Probably the main achievement of Saussure was to outline the framework for modern linguistics (see *Linguistics: Overview*). In contrast to the emphasis that earlier philology had placed on the diachronic perspective of how languages change over time, Saussure proposed to study language as a system in a synchronic perspective. Language as an abstract system (*langue*) could be distinguished, at least for analytical purposes, from the actual uses of language (*parole*). The language system has two dimensions. Along the syntagmatic dimension, letters, words, phrases, etc., are the units that combine to make up meaningful wholes, and each of these units has been chosen as one of several possibilities along a paradigmatic dimension, for example, one verb in preference to another. This combinatory system helps to account for the remarkable flexibility of language as a medium of social interaction.

To be precise, Saussure referred to the broader science of signs, as cited in the introduction to this article, not as semiotics, but as semiology. Like Locke, and later Peirce, Saussure relied on classical Greek to coin a technical term, and the two variants are explained, in part, by the fact that Peirce and Saussure had no knowledge of each other's work. Moreover, in Saussure's own work, the program for a semiology remained undeveloped, appearing almost as an aside from his main enterprise of linguistics. It was during the consolidation of semiotics as an interdisciplinary field from the 1960s that this became the agreed term, overriding Peirce's 'semiotic' and Saussure's 'semiology', as symbolized by the formation in 1969 of the International Association for Semiotic Studies. It was also during this period that Saussure's systemic approach was redeveloped to apply to culture and society.

One important legacy of Saussure has been his account of the arbitrariness of the linguistic sign. This sign type is said to have two sides, a signified (concept) and a signifier (the acoustic image associated with it), whose relation is conventional or arbitrary. While this argument is sometimes construed in skepticist and relativist terms, as if sign users were paradoxically free to choose their own meanings, and hence almost destined to remain divorced from any consensual reality, the point is rather that the linguistic system as a whole, including the interrelations between signs and their components, is arbitrary, but fixed by social convention. When applied to studies of cultural forms other than language, or when extended to the description of social structures as signs, the principle of arbitrariness has been a source of analytical difficulties.

2. Applications: Media, Communication and Culture

It was the application of semiotics to questions of meaning beyond logic and linguistics which served to consolidate semiotics as a recognizable, if still heterogeneous field from the 1960s. Drawing on other traditions of aesthetic and social research, this emerging field began to make concrete what the study of 'the life of signs within society' might mean. The objects of analysis ranged from artworks to mass media to the life forms of premodern societies, but studies were united by a common interest in culture in the broad sense of worldviews that orient social action. The work of Claude Lévi-Strauss (1963) on structural anthropology was characteristic and highly influential of such research on shared, underlying systems of interpretation which might explain the viewpoints or actions of individuals in a given context. It was Saussure's emphasis on signs as a system, then, which provided a model for structuralist theories (see *Structuralism*) about social power (e.g., Louis Althusser) and about the unconscious as a 'language' (e.g., Jacques Lacan) (see *Psychoanalysis: Overview*).

Of the French scholars who were especially instrumental in this consolidation, Roland Barthes, along with A. J. Greimas, stands out as an innovative and systematic theorist. His model of two levels of signification, reproduced in Fig. 2, has been one of the most widely copied attempts to link concrete sign vehicles, such as texts and images, with the 'myths' or ideologies which they articulate. Building on Louis Hjelmslev's formal linguistics, Barthes (1973) suggested that the combined signifier and signified (expressive form and conceptual content) of one sign (e.g., a picture of a black man in a French uniform saluting the flag) may become the expressive form of a

Figure 2
Two Levels of Signification (reproduced by permission of Roland Barthes [Random House Group Ltd (Jonathan Cape)] from *Mythologies*, p. 115, first published in French by de Seuil in 1957)

further, ideological content (e.g., that French imperialism is not a discriminatory or oppressive system). Barthes' political agenda, shared by much semiotic scholarship since then, was that this semiotic mechanism serves to naturalize particular worldviews, while oppressing others, and should be deconstructed (see *Critical Theory: Contemporary*).

In retrospect, one can distinguish two ways of appropriating semiotics in social and cultural research. On the one hand, semiotics may be treated as a methodology for examining signs, whose social or philosophical implications are interpreted with reference to another set of theoretical concepts. On the other hand, semiotics may supply also the theoretical framework, so that societies, psyches, and images are understood not just in procedural, but also in actual conceptual terms as signs. The latter position is compatible with a variant of semiotics, which assumes that also biological and even cosmological processes are best understood as semioses. This ambitious and arguably imperialistic extension of logical and linguistic concepts into other fields has encountered criticism, for example, in cases where Saussure's principle of arbitrariness has been taken to apply to visual communication or to the political and economic organization of society.

A semioticization of, for instance, audiovisual media (for example, see *Visual Images in the Media*) can be said to neglect their appeal to radically different perceptual registers, which are, in certain respects, natural rather than conventional (e.g., Messaris 1994) (for example, see *Cognitive Science: Overview*). Similarly, a *de facto* replacement of social science by semiotics may exaggerate the extent to which signs, rather than material and institutional conditions, determine the course of society, perhaps in response to semioticians' political ambition of making a social difference by deconstructing signs. In recent decades, these concerns, along with critiques of Saussurean formalist systems, have led to attempts to formulate a social semiotics, integrating semiotic methodology with other social and communication theory (e.g., Hodge and Kress 1988, Jensen 1995). On the relations between semiotics and other fields, see Nöth 1990.

A final distinction within semiotic studies of society and culture arises from the question of what is a 'medium'. Media research has drawn substantially on semiotics to account for the distinctive sign types, codes, narratives and modes of address of different media such as newspapers, television or the Internet (for example, see *Mass Media: Introduction and Schools of Thought*). A particular challenge has been the film medium, which was described by Christian Metz (1974) not as a language system, but as a 'language' drawing on several semiotic codes. Beyond this approach to media as commonly understood, semioticians, taking their inspiration from Lévi-Strauss' anthropology, have described other objects and artifacts as vehicles of meaning. One rich example is Barthes' study of the structure of fashion (see *Fashion, Sociology of*), as depicted in magazines (Barthes 1985). The double meaning of a 'medium' is another indication both of the potential theoretical richness of semiotics and of the pitfalls of confusing different levels of analysis.

3. *Theory of Signs and Theory of Science*

It is this double edge which most likely explains the continuing influence of semiotic ideas since Antiquity, mostly, however, as an undercurrent of science. The understanding articulated in Aristotle's writings of signs as evidence for an interpreter of something else which is at least temporarily absent, or as an interface which is not identical with the thing that is in evidence to the interpreter, may be taken as a cultural leap that has made possible scientific reflexivity as well as social organization beyond the here and now. Further, the sign concept has promoted the crucial distinction beginning with Aristotle between necessary and probable relations, as in logical inferences. Returning to the examples above, fever is a sure sign that the person is ill, since illness is a necessary condition of fever. But clouds are only a probable sign of rain. Empirical social research is mostly built on the premise that probable signs can lead to warranted inferences.

The link between a theory of signs and a theory of science has been explored primarily from a Peircean perspective. For one thing, Peirce himself made important contributions to logic and to the theory of science generally, notably by adding to deduction and induction the form of inference called abduction, which he detected both in everyday reasoning and at the core of scientific innovation. For another thing, the wider philosophical tradition of pragmatism (for example, see *Pragmatism: Philosophical Aspects*), which Peirce inaugurated, has emphasized the interrelations between knowledge, education, and action, and it has recently enjoyed a renaissance at the juncture between theory of science and social theory (e.g., Bernstein 1991, Joas 1993). By contrast, the Saussurean tradition has often bracketed issues concerning the relationship between the word and the world (Jakobson 1981, p. 19), even if critical research in this vein has recruited theories from other schools of thought in order to move semiology beyond the internal analysis of signs.

For future research, semiotics is likely to remain a source of inspiration in various scientific domains, as it has been for almost 2500 years. In light of the record from the past century of intensified interest in theories of signs, however, the field is less likely to develop into a coherent discipline. One of the main opportunities for semiotics may be to develop a meta-framework for understanding how different disciplines and fields conceive of their 'data' and 'concepts'. The semiotic heritage offers both systematic analytical procedures

and a means of reflexivity regarding the role of signs in society as well as in the social sciences.

See also: Communication: Philosophical Aspects

Bibliography

Barthes R 1973 *Mythologies*. Paladin, London

Barthes R 1985 *The Fashion System*. Cape, London

Bernstein R J 1991 *The New Constellation*. Polity Press, Cambridge, UK

Bouissac P 1998 (ed.) *Encyclopedia of Semiotics*. Oxford University Press, New York

Clarke Jr. D S 1990 *Sources of Semiotic*. Southern Illinois University Press, Carbondale, IL

Greimas A J, Courtés J 1982 *Semiotics and Language: An Analytical Dictionary*. Indiana University Press, Bloomington, IN

Hodge R and Kress G 1988 *Social Semiotics*. Polity Press, Cambridge, UK

Jakobson R 1981 Linguistics and poetics. In: *Selected Writings*. Mouton, The Hague, The Netherlands, Vol. 3

Joas H 1993 *Pragmatism and Social Theory*. University of Chicago Press, Chicago, IL

Jensen K B 1995 *The Social Semiotics of Mass Communication*. Sage, London

Lévi-Strauss C 1963 *Structural Anthropology*. Basic Books, New York

Messaris P 1994 *Visual 'Literacy': Image, Mind, and Reality*. Westview Press, Boulder, CO

Metz C 1974 *Language and Cinema*. Mouton, The Hague, The Netherlands

Nöth W 1990 *Handbook of Semiotics*. Indiana University Press, Bloomington, IN

Peirce C S 1982 *Writings of Charles S. Peirce: A Chronological Edition*. Indiana University Press, Bloomington, IN

Peirce C S 1992–98 *The Essential Peirce*. Indiana University Press, Bloomington, IN, Vols 1–2

Posner R, Robering K, Sebeok T A (eds.) 1997–98 *Semiotik: Ein Handbuch zu den zeichentheoretischen Grundlagen von Natur und Kultur/Semiotics: A Handbook on the Sign-Theoretic Foundations of Nature and Culture*. Walter de Gruyter, Berlin, Vols. 1–2

de Saussure F 1959 *Course in General Linguistics*. Peter Owen, London

Sebeok T A 1994 *Encyclopedic Dictionary of Semiotics*, 2nd edn. Mouton de Gruyter, Berlin, Vols 1–3

K. B. Jensen

Semiparametric Models

Much empirical research in the social sciences is concerned with estimating conditional mean functions. For example, labor economists are interested in estimating the mean wages of employed individuals, conditional on characteristics such as years of work experience and education. The most frequently used estimation methods assume that the conditional mean function is known up to a set of constant parameters that can be estimated from data, possibly by ordinary least squares. Models in which the only unknown quantities are a finite set of constant parameters are called 'parametric.' The use of a parametric model greatly simplifies estimation, statistical inference, and interpretation of the estimation results but is rarely justified by theoretical or other *a priori* considerations. Estimation and inference based on convenient but incorrect assumptions about the form of the conditional mean function can be highly misleading. Semiparametric statistical methods reduce the strength of the assumptions required for estimation and inference, thereby reducing the opportunities for obtaining misleading results. These methods are applicable to a wide variety of estimation problems in economics and other fields.

1. Introduction

A conditional mean function gives the mean of a dependent variable Y conditional on a vector of explanatory variables X. Denote the mean of Y conditional on $X = x$ by $\boldsymbol{E}(Y|x)$. For example, suppose that Y is a worker's weekly wage (or, more often in applied econometrics, the logarithm of the wage) and X includes such variables as years of work experience and education, race, and sex. Then $\boldsymbol{E}(Y|x)$ is the mean wage (or logarithm of the wage) when experience and the other explanatory variables have the values specified by x. As an illustration, the solid line in Fig. 1 shows an estimate of the mean of the logarithm of weekly wages, $\log W$, conditional on years of work experience, EXP, for white males with 12 years of education who work full time and live in urban areas of the North Central USA. The estimate was obtained by applying a nonparametric method (explained in Sect. 2) to data from the 1993 Current Population Survey (CPS). The estimated conditional mean of $\log W$ increases steadily up to approximately 30 years of experience and is flat thereafter.

In most applications, $\boldsymbol{E}(Y|x)$ is unknown and must be estimated from data on the variables of interest. In the case of estimating a wage function, the data consist of observations of individuals' wages, years of experience, and other characteristics. The most widely used method for estimating $\boldsymbol{E}(Y|x)$ is not the nonparametric method mentioned previously, but rather a method that assumes that $\boldsymbol{E}(Y|x)$ is known up to finitely many constant parameters. This gives a 'parametric model' for $\boldsymbol{E}(Y|x)$. Often, $\boldsymbol{E}(Y|x)$ is assumed to be a linear function of x, in which case the parameters can be estimated by ordinary least squares (OLS), among other ways. A linear function has the form $\boldsymbol{E}(Y|x) = \beta' x$, where β is a vector of coefficients. For

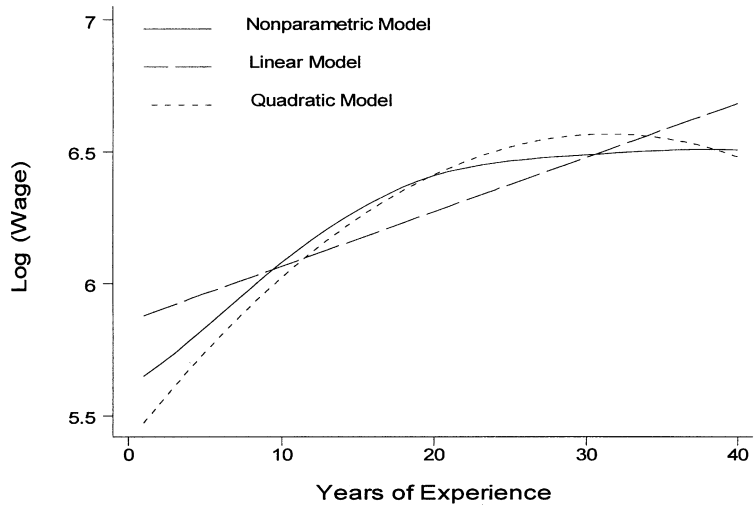

Figure 1
Estimates of $\boldsymbol{E}(\log W \mid EXP)$

example, if x consists of an intercept and the two variables x_1 and x_2, then β has three components, and $\beta' x = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. OLS estimators are described in many textbooks. See, for example, Goldberger (1998).

The OLS estimator of $\boldsymbol{E}(Y \mid x)$ can be highly misleading if $\boldsymbol{E}(Y \mid x)$ is not linear in the components of x, that is if there is no β such that $\boldsymbol{E}(Y \mid x) = \beta' x$. This problem is illustrated by the dashed and dotted lines in Fig. 1, which show two parametric estimates of the mean of the logarithm of weekly wages conditional on years of work experience. The dashed line is the OLS estimate that is obtained by assuming that $\boldsymbol{E}(\log W \mid EXP)$ is the linear function $\boldsymbol{E}(\log W \mid EXP) = \beta_0 + \beta_1 EXP$. The dotted line is the OLS estimate that is obtained by assuming that $\boldsymbol{E}(\log W \mid EXP)$ is quadratic: $\boldsymbol{E}(\log W \mid EXP) = \beta_0 + \beta_1 EXP + \beta_2 EXP^2$. The nonparametric estimate (solid line) places no restrictions on the shape of $\boldsymbol{E}(\log W \mid EXP)$. The linear and quadratic models give misleading estimates of $\boldsymbol{E}(\log W \mid EXP)$. The linear model indicates that $\boldsymbol{E}(\log W \mid EXP)$ steadily increases as experience increases. The quadratic model indicates that $\boldsymbol{E}(\log W \mid EXP)$ decreases after 32 years of experience. In contrast, the nonparametric estimate of $\boldsymbol{E}(\log W \mid EXP)$ becomes nearly flat at approximately 30 years of experience. Because the nonparametric estimate does not restrict the conditional mean function to be linear or quadratic, it is more likely to represent the true conditional mean function.

The opportunities for specification error increase if Y is binary. For example, consider a model of the choice of travel mode for the trip to work. Suppose that the available modes are automobile and transit. Let $Y = 1$ if an individual chooses automobile and $Y = 0$ if the individual chooses transit. Let X be a vector of explanatory variables such as the travel times and costs by automobile and transit. Then $\boldsymbol{E}(Y \mid x)$ is the probability that $Y = 1$ (the probability that the individual chooses automobile) conditional on $X = x$. This probability will be denoted $\boldsymbol{P}(Y = 1 \mid x)$. In applications of binary response models, it is often assumed that $\boldsymbol{P}(Y \mid x) = G(\beta' \mathrm{x})$, where β is a vector of constant coefficients and G is a known probability distribution function. Often, G is assumed to be the cumulative standard normal distribution function, which yields a 'binary probit' model, or the cumulative logistic distribution function, which yields a 'binary logit' model (see *Multivariate Analysis: Discrete Variables (Logistic Regression)*). The coefficients β can then be estimated by the method of maximum likelihood (Amemiya 1985). However, there are now two potential sources of specification error. First, the dependence of Y on x may not be through the linear *index* $\beta' x$. Second, even if the index $\beta' x$ is correct, the 'response function' G may not be the normal or logistic distribution function. See Horowitz (1993, 1998) for examples of specification errors in binary response models and their consequences.

Many investigators attempt to minimize the risk of specification error by carrying out a 'specification search' in which several different models are estimated and conclusions are based on the one that appears to fit the data best. Specification searches may be unavoidable in some applications, but they have many undesirable properties and their use should be minimized. There is no guarantee that a specification search will include the correct model or a good approximation to it. If the search includes the correct model, there is no guarantee that it will be selected by the investigator's model selection criteria. Moreover, the

search process invalidates the statistical theory on which inference is based.

The rest of this entry describes methods that deal with the problem of specification error by relaxing the assumptions about functional form that are made by parametric models. The possibility of specification error can be essentially eliminated through the use of nonparametric estimation methods. These are described in Sect. 2. They assume that $\boldsymbol{E}(Y|x)$ is a smooth function but make no other assumptions about its shape or functional form. However, nonparametric methods have important disadvantages that seriously limit their usefulness in applications. Semiparametric methods, which are described in Sect. 3 offer a compromise. They make assumptions about functional form that are stronger than those of a nonparametric model but less restrictive than the assumptions of a parametric model, thereby reducing (though not eliminating) the possibility of specification error. In addition semiparametric methods avoid the most serious practical disadvantages of nonparametric methods.

2. Nonparametric Models

In nonparametric estimation $\boldsymbol{E}(Y|x)$ is assumed to satisfy smoothness conditions such as differentiability, but no assumptions are made about its shape or the form of its dependence on x. Härdle (1990) and Fan and Gijbels (1996) provide detailed discussions of nonparametric estimation methods. One easily understood and frequently used method is called 'kernel estimation.' This method was used to produce the solid line in Fig. 1. To describe the kernel method simply, assume that X is a continuously distributed, scalar random variable. Let $\{Y_i, X_i : i = 1, \ldots, n\}$ be a random sample of n observations of (Y, X). Let K be a probability density function that is bounded, continuous, and symmetrical about zero. For example, K may be the standard normal density function. Let $\{h_n\}$ to be a sequence of positive numbers that converges to 0 as $n \to \infty$. For each $n = 1, 2, \ldots$ and $i = 1, \ldots, n$ define the function $w_{ni}(\cdot)$ by

$$w_{ni}(x) = \frac{K[(x - X_i)/h_n]}{\sum_{i=1}^{n} K[(x - X_i)/h_n]} \tag{1}$$

Then the kernel nonparametric estimator of $\boldsymbol{E}(Y|x)$ is

$$H_n(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i \tag{2}$$

$H_n(x)$ is a weighted average of the observed values of Y. Observations Y_i for which X_i is close to x get higher weight than do observations for which X_i is far from x. It can be shown that if $h_n \to 0$ and $nh_n/(\log n) \to \infty$ as $n \to \infty$, then $H_n(x) \to \boldsymbol{E}(Y|x)$ with probability 1. Thus, if n is large, $H_n(x)$ is likely to be very close to $\boldsymbol{E}(Y|x)$. Härdle (1990) provides a detailed discussion of the statistical properties of kernel nonparametric estimators.

Nonparametric estimation minimizes the risk of specification error, but the price of this flexibility can be high. One important reason for this is that the precision of a nonparametric estimator decreases rapidly as the number of continuously distributed components of X increases. This phenomenon is called the 'curse of dimensionality.' As a result of it, impracticably large samples are usually needed to obtain acceptable estimation precision if X is multidimensional, as it often is in social science applications. For example, a labor economist may want to estimate mean log wages conditional on years of work experience, years of education, and one or more indicators of skill levels, thus making the dimension of X at least 3. See *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)* for further discussion of the curse of dimensionality.

Another problem is that nonparametric estimates can be difficult to display, communicate, and interpret when X is multidimensional. Nonparametric estimates do not have simple analytic forms, so displaying and interpreting them can be difficult. If X is one- or two-dimensional, then the estimate of $\boldsymbol{E}(Y|x)$ can be displayed graphically as in Fig. 1, but only reduced-dimension projections can be displayed when X has three or more components. Many such displays and much skill in interpreting them can be needed to fully convey and comprehend the shape of the estimate of $\boldsymbol{E}(Y|x)$.

Another problem with nonparametric estimation is that it does not permit extrapolation. That is, it does not provide predictions of $\boldsymbol{E}(Y|x)$ at points x that are outside of the support (or range) of the random variable X. This is a serious drawback in policy analysis and forecasting, where it is often important to predict what might happen under conditions that do not exist in the available data. Finally, in nonparametric estimation, it can be difficult to impose restrictions suggested by economic or other theory. Matzkin (1994) discusses this issue.

Semiparametric methods permit greater estimation precision than do nonparametric methods when X is multidimensional. In addition, semiparametric estimates are easier to display and interpret than nonparametric ones and provide limited capabilities for extrapolation and imposing restrictions derived from economic or other theory models.

3. Semiparametric Models

The term 'semiparametric' refers to models in which there is an unknown function in addition to an unknown finite dimensional parameter. For example, the binary response model $\boldsymbol{P}(Y = 1|x) = G(\beta' x)$ is

semiparametric if the function G and the vector of coefficients β are both treated as unknown quantities. This section describes two semiparametric models of conditional mean functions that are important in applications. The section also describes a related class of models that have no unknown finite-dimensional parameters but, like semiparametric models, mitigate the disadvantages of fully nonparametric models.

In addition to the estimation of conditional mean functions, semiparametric methods can be used to estimate conditional quantile and hazard functions, binary response models in which there is heteroskedasticity of unknown form, transformation models, and censored and truncated mean- and median- regression models, among others. Space does not permit discussion of these models here. Horowitz (1998) and Powell (1994) provide more comprehensive treatments in which these models are discussed.

3.1 Single Index Models

In a semiparametric single index model, the conditional mean function has the form

$$\boldsymbol{E}(Y \mid x) = G(\beta' x) \tag{3}$$

where β is an unknown constant vector and G is an unknown function. The quantity $\beta' x$ is called an 'index.' The inferential problem is to estimate G and β from observations of (Y, X). G in (3.1) is analogous to a link function in a generalized linear model, except in Eqn. (3) G is unknown and must be estimated.

Model (3) contains many widely used parametric models as special cases. For example, if G is the identity function, then Eqn. (3) is a linear model. If G is the cumulative normal or logistic distribution function, then Eqn. (3) is a binary probit or logit model. When G is unknown, Eqn. (3) provides a specification that is more flexible than a parametric model but retains many of the desirable features of parametric models, as will now be explained.

One important property of single index models is that they avoid the curse of dimensionality. This is because the index $\beta' x$ aggregates the dimensions of x, thereby achieving 'dimension reduction.' Consequently, the difference between the estimator of G and the true function can be made to converge to zero at the same rate that would be achieved if $\beta' x$ were observable. Moreover, β can be estimated with the same rate of convergence that is achieved in a parametric model. Thus, in terms of the rates of convergence of estimators, a single index model is as accurate as a parametric model for estimating β and as accurate as a one-dimensional nonparametric model for estimating G. This dimension reduction feature of single index models gives them a considerable advantage over nonparametric methods in applications where X is multidimensional and the single index structure is plausible.

A single-index model permits limited extrapolation. Specifically, it yields predictions of $\boldsymbol{E}(Y \mid x)$ at values of x that are not in the support of X but are in the support of $\beta' X$. Of course, there is a price that must be paid for the ability to extrapolate. A single index model makes assumptions that are stronger than those of a nonparametric model. These assumptions are testable on the support of X but not outside of it. Thus, extrapolation (unavoidably) relies on untestable assumptions about the behavior of $\boldsymbol{E}(Y \mid x)$ beyond the support of X.

Before β and G can be estimated, restrictions must be imposed that insure their identification. That is, β and G must be uniquely determined by the population distribution of (Y, X). Identification of single index models has been investigated by Ichimura (1993) and, for the special case of binary response models, Manski (1988). It is clear that β is not identified if G is a constant function or there is an exact linear relation among the components of X (perfect multicollinearity). In addition, (3.1) is observationally equivalent to the model $\boldsymbol{E}(Y \mid X) = G^*(\gamma + \delta\beta' x)$, where γ and $\delta \neq 0$ are arbitrary and G^* is defined by the relation $G^*(\gamma + \delta v) = G(v)$ for all v in the support of $\beta' X$. Therefore, β and G are not identified unless restrictions are imposed that uniquely specify γ and δ. The restriction on γ is called 'location normalization' and can be imposed by requiring X to contain no constant (intercept) component. The restriction on δ is called 'scale normalization.' Scale normalization can be achieved by setting the β coefficient of one component of X equal to one. A further identification requirement is that X must include at least one continuously distributed component whose β coefficient is nonzero. Horowitz (1998) gives an example that illustrates the need for this requirement. Other more technical identification requirements are discussed by Ichimura (1993) and Manski (1988).

The main estimation challenge in single index models is estimating β. Given an estimator b_n of β, G can be estimated by carrying out the nonparametric regression of Y on $b_n X$ (e.g., by using the kernel method described in Sect. 2). Several estimators of β are available. Ichimura (1993) describes a nonlinear least squares estimator. Klein and Spady (1993) describe a semiparametric maximum likelihood estimator for the case in which Y is binary. These estimators are difficult to compute because they require solving complicated nonlinear optimization problems. Powell et al. (1989) describe a 'density-weighted average derivative estimator' (DWADE) that is noniterative and easily computed. The DWADE applies when all components of X are continuous random variables. It is based on the relation

$$\beta \propto \boldsymbol{E}[p(X)\partial G(\beta' X)/\partial X] = -2\boldsymbol{E}[Y\partial p(X)/\partial X] \tag{4}$$

where p is the probability density function of X and the second equality follows from integrating the first by

parts. Thus, β can be estimated up to scale by estimating the expression on the right-hand side of the second equality. Powell et al. (1989) show that this can be done by replacing p with a nonparametric estimator and replacing the population expectation $\boldsymbol{E}$ with a sample average. Horowitz and Härdle (1996) extend this method to models in which some components of X are discrete. They also give an empirical example that illustrates the usefulness of single index models. Ichimura and Lee (1991) investigate a multiple index generalization of Eqn. (3).

3.2 Partially Linear Models

In a partially linear model, X is partitioned into two nonoverlapping subvectors, X_1 and X_2. The model has the form

$$\boldsymbol{E}(Y \mid x_1, x_2) = \beta' x_1 + G(x_2) \tag{5}$$

where β is an unknown constant vector and G is an unknown function. This model is distinct from the class of single index models. A single index model is not partially linear unless G is a linear function. Conversely, a partially linear model is a single index model only in this case. Stock (1989, 1991) and Engle et al. (1986) illustrate the use of Eqn. (5) in applications. Identification of β requires the 'exclusion restriction' that none of the components of X_1 are perfectly predictable by components of X_2. When β is identified, it can be estimated with an $n^{-1/2}$ rate of convergence regardless of the dimensions of X_1 and X_2. Thus, the curse of dimensionality is avoided in estimating β.

An estimator of β can be obtained by observing that Eqn. (5) implies

$$Y - \boldsymbol{E}(Y \mid x_2) = \beta'[X_1 - \boldsymbol{E}(X_1 \mid x_2)] + U \tag{6}$$

where U is an unobserved random variable satisfying $\boldsymbol{E}(U \mid x_1, x_2) = 0$. Robinson (1988) shows that under regularity conditions β can be estimated by applying OLS to Eqn. (6) after replacing $\boldsymbol{E}(Y \mid x_2)$ and $\boldsymbol{E}(X_1 \mid x_2)$ with nonparametric estimators. The estimator of β, b_n, converges at rate $n^{-1/2}$ and is asymptotically normally distributed. G can be estimated by carrying out the nonparametric regression of $Y - b_n' X_1$ on X_2. Unlike b_n, the estimator of G suffers from the curse of dimensionality; its rate of convergence decreases as the dimension of X_2 increases.

3.3 Nonparametric Additive Models

Let X have d continuously distributed components that are denoted $X_1, \ldots X_d$. In a nonparametric additive model of the conditional mean function,

$$\boldsymbol{E}(Y \mid x) = \mu + f_1(x_1) + \cdots + f_d(x_d) \tag{7}$$

where μ is a constant and $f_1, \ldots, f_d$ are unknown functions that satisfy a location normalization condition such as

$$\int f_k(v) w_k(v)\, \mathrm{d}v = 0, \quad k = 1, \ldots, d \tag{8}$$

where w_k is a non-negative weight function. An additive model is distinct from a single index model unless $\boldsymbol{E}(Y \mid x)$ is a linear function of x. Additive and partially linear models are distinct unless $\boldsymbol{E}(Y \mid x)$ is partially linear and G in Eqn. (5) is additive.

An estimator of $f_k (k = 1, \ldots, d)$ can be obtained by observing that Eqn. (7) and (8) imply

$$f_k(x_k) = \int \boldsymbol{E}(Y \mid x) w_{-k}(x_{-k})\, \mathrm{d}x_{-k} \tag{9}$$

where x_{-k} is the vector consisting of all components of x except the kth and w_{-k} is a weight function that satisfies $\int w_{-k}(x_{-k})\, \mathrm{d}x_{-k} = 1$. The estimator of f_k is obtained by replacing $\boldsymbol{E}(Y \mid x)$ on the right-hand side of Eqn. (9) with nonparametric estimators. Linton and Nielsen (1995) and Linton (1997) present the details of the procedure and extensions of it. Under suitable conditions, the estimator of f_k converges to the true f_k at rate $n^{-2/5}$ regardless of the dimension of X. Thus, the additive model provides dimension reduction. It also permits extrapolation of $\boldsymbol{E}(Y \mid x)$ within the rectangle formed by the supports of the individual components of X. Hastie and Tibshirani (1990) and *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)* discuss an alternative estimation procedure called 'backfitting.' This procedure is widely used, but its asymptotic properties are not yet well understood.

Linton and Härdle (1996) describe a generalized additive model whose form is

$$\boldsymbol{E}(Y \mid x) = G[\mu + f_1(x_1) + \cdots + f_K(x_d)] \tag{10}$$

where $f_1, \ldots, f_d$ are unknown functions and G is a known, strictly increasing (or decreasing) function. Horowitz (2001) describes a version of Eqn. (10) in which G is unknown. Both forms of Eqn. (10) achieve dimension reduction. When G is unknown, Eqn. (10) nests additive and single index models and, under certain conditions, partially linear models.

The use of the nonparametric additive specification (7) can be illustrated by estimating the model

$$\begin{aligned}\boldsymbol{E}(\log W \mid EXP, EDUC) \\ = \mu + f_{EXP}(EXP) + f_{EDUC}(EDUC)\end{aligned}$$

where W and EXP are defined as in Sect. 1, and $EDUC$ denotes years of education. The data are taken from the 1993 CPS and are for white males with 14 or fewer

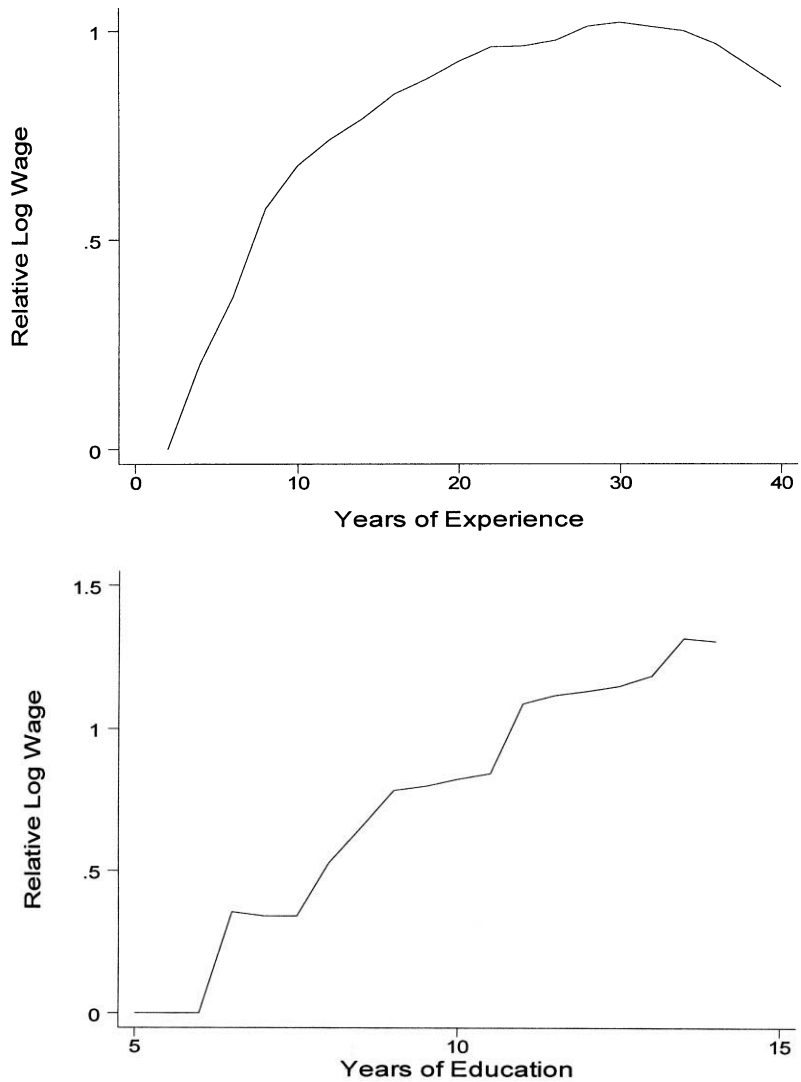

Figure 2
(a) Estimate of f_{EXP} in additive nonparametric model of $\boldsymbol{E}(\log W \mid EXP, EDUC)$. (b) Estimate of f_{EDUC} in additive nonparametric model of $\boldsymbol{E}(\log W \mid EXP, EDUC)$

years of education who work full time and live in urban areas of the North Central US. The results are shown in Fig. 2. The unknown functions f_{EXP} and f_{EDUC} are estimated by the method of Linton and Nielsen (1995) and are normalized so that $f_{EXP}(2) = f_{EDUC}(5) = 0$. The estimates of f_{EXP} (Fig. 2a) and f_{EDUC} (Fig. 2b) are nonlinear and differently shaped. Functions f_{EXP} and f_{EDUC} with different shapes cannot be produced by a single index model, and a lengthy specification search might be needed to find a parametric model that produces the shapes shown in Fig. 2. Some of the fluctuations of the estimates of f_{EXP} and f_{EDUC} may be artifacts of random sampling error rather than features of $\boldsymbol{E}(\log W \mid EXP, EDUC)$. However, a more elaborate analysis that takes account of the effects of random sampling error rejects the hypothesis that either function is linear.

4. *Conclusions*

This article has described several semiparametric methods for estimating conditional mean functions. These methods relax the restrictive assumptions made by linear and other parametric models, thereby reducing (though not eliminating) the likelihood of

seriously misleading inference. The value of semiparametric methods in empirical research has been demonstrated. Their use is likely to increase as their availability in commercial statistical software packages increases.

Bibliography

Amemiya T 1985 *Advanced Econometrics*. Harvard University Press, Cambridge, MA

Engle R F, Granger C W J, Rice J, Weiss A 1986 Semiparametric estimates of the relationship between weather and electricity sales. *Journal of the American Statistical Association* **81**: 310–20

Fan J, Gijbels I 1996 *Local Polynomial Modelling and its Applications*. Chapman & Hall, London

Goldberger A S 1998 *Introductory Econometrics*. Harvard University Press, Cambridge, MA

Härdle W 1990 *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, UK

Hastie T J, Tibshirani R J 1990 *Generalized Additive Models*. Chapman & Hall, London

Horowitz J L 1993 Semiparametric and nonparametric estimation of quantal response models. In: Maddala G S, Rao C R, Vinod H D (eds.) *Handbook of Statistics*. Elsevier, Amsterdam, Vol. II, pp. 45–72

Horowitz J L 1998 *Semiparametric Methods in Econometrics*. Springer-Verlag, New York

Horowitz J L 2001 Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica* **69**: 599–631

Horowitz J L, Härdle W 1996 Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association* **91**: 1632–40

Ichimura H 1993 Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**: 71–120

Ichimura H, Lee L-F 1991 Semiparametric least squares estimation of multiple index models: single equation estimation. In: Barnett W A, Powell J, Tauchen G (eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge, UK, pp. 3–49

Klein R W, Spady R H 1993 An efficient semiparametric estimator for binary response models. *Econometrica* **61**: 387–421

Linton O B 1997 Efficient estimation of additive nonparametric regression models. *Biometrika* **84**: 469–73

Linton O B, Härdle W 1996 Estimating additive regression models with known links. *Biometrika* **83**: 529–40

Linton O B, Nielsen J P 1995 A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**: 93–100

Manski C F 1988 Identification of binary response models. *Journal of the American Statistical Association* **83**: 729–38

Matzkin R L 1994 Restrictions of economic theory in nonparametric methods. In: Engle R F, McFadden D L (eds.) *Handbook of Econometrics*. North-Holland, Amsterdam, Vol. 4, pp. 2523–58

Powell J L 1994 Estimation of semiparametric models. In: Engle R F, McFadden D L (eds.) *Handbook of Econometrics*. North-Holland, Amsterdam, Vol. 4, pp. 2444–521

Powell J L, Stock J H, Stoker T M 1989 Semiparametric estimation of index coefficients. *Econometrica* **51**: 1403–30

Robinson, P M 1988 Root-*N*-consistent semiparametric regression. *Econometrica* **56**: 931–54

Stock J H 1989 Nonparametric policy analysis. *Journal of the American Statistical Association* **84**: 567–75

Stock J H 1991 Nonparametric policy analysis: An application to estimating hazardous waste cleanup benefits. In: Barnett W A, Powell J, Tauchen G (eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge, UK, pp. 77–98

J. L. Horowitz

Senescence: Genetic Theories

Senescence is the progressive deterioration of vitality that accompanies increasing age. Like other features of organismal life histories, patterns of senescence vary between individuals within populations, between populations of the same species, and between species, suggesting that they are modifiable by genetic factors and subject to evolutionary change. In this article, the various evolutionary forces that might direct genetic modifications of senescence are considered, and a theoretical framework for understanding the evolution of life histories is presented. The secondary problem of the maintenance of genetic variation for life history traits is also reviewed.

1. Medawar's Principle

The modern evolutionary theory of senescence begins with Medawar who argued that '…the efficacy of natural selection deteriorates with increasing age' (1952, p. 23). A simple hypothetical example similar to a case considered by Hamilton (1966) illustrates the principle. Consider two genetic variants in humans, both having age-specific effects as follows: each variant confers complete immunity against a lethal disease, but only for one particular year of life. The first variant gives immunity to 12 year-olds, while the second variant confers immunity at the age of 60 years.

What are the relative selective advantages of the genetic variants? If, for simplicity, the effects of parental care are ignored and it is also assumed that menopause always comes before 60 years, then it is immediately obvious that the second variant is selectively neutral, having no effect at all on the ability of carriers to transmit genes to the next generation, whereas the first variant has a significant selective advantage. This example illustrates the general principle that natural selection is most effective in the

young. To obtain a more exact and quantitative understanding of the relation between organismal age and the force of selection, it is necessary to develop a description of selection in age-structured populations.

2. *Age-structured Populations*

Some organisms, such as annual plants, complete their life cycles in discrete fashion, exhibiting no overlap of parental and offspring generations. However, most higher organisms, including humans, have overlapping generations. Under the latter circumstances, the description of population composition and growth requires two kinds of information: age-specific survival, and age-specific fertility. Survivorship, denoted $l(x)$, is defined as the probability of survival from birth or hatching until age x. A survivorship curve is a graph of $l(x)$ versus x, where x ranges from zero to the greatest age attained in the population. The survivorship is initially 100 percent at birth and then declines to zero at the maximum observed age; it cannot increase with increasing age. If a cohort of 1,000 age-synchronized individuals are followed throughout their lives, then 500 of them will be alive at the age when $l(x)$ is 0.50, 100 will be alive when $l(x)$ is 0.10, and so on. Age-specific fertility, represented as $m(x)$, is defined as the average number of progeny produced by a female of age x.

One of the fundamentals of demography is that, under a wide range of conditions, a population having fixed $l(x)$ and $m(x)$ schedules will eventually attain a stable age-structure. That is, after a period of time the proportions of the population in each age-class will reach unchanging values. If the survivorship or fertility schedules are altered, then a different age-structure will evolve.

Prior to attaining the stable age distribution, population growth is likely to be erratic, but once the stable age distribution is reached, then, under the assumption of unlimited resources, the population will grow smoothly. In particular, the population will exhibit exponential growth described by the following equation:

$$N(t) = N(0)\,\mathrm{e}^{rt} \tag{1}$$

where $N(t)$ is population size as a function of time t, $N(0)$ is initial population size, e is the natural exponential, and r is the Malthusian parameter, also known as the intrinsic rate of increase of the population. The parameter r combines the effects of age-specific survival and fertility and translates them into a population growth rate. The value of r is the implicit solution to the following equation, known as the Euler–Lotka equation:

$$\int \mathrm{e}^{-rx}\, l(x)\, m(x)\, \mathrm{d}x = 1 \tag{2}$$

The significance of the Malthusian parameter is that it reflects fitness, in the Darwinian sense, in an age-structured population. If there are several genotypes in a population, and if those genotypes differ with respect to age-specific survival or fertility patterns, then each genotype will have a particular r value. Those r's specify the rate at which a population consisting of only that genotype would grow, once the stable age distribution has been attained. The r's also specify the relative fitnesses of the genotypes in a genotypically mixed population. The genotype with the highest r has the highest fitness and will be favored by natural selection under conditions that allow population growth. Much of the evolutionary theory relating to senescence and life histories uses the Malthusian parameter r as a surrogate for Darwinian fitness, essentially asking what changes in the $l(x)$ and $m(x)$ schedules would maximize the intrinsic rate of increase.

There is one other quantity that arises as a measure of fitness in populations with overlapping generations. Fisher (1930) defined 'reproductive value,' which is the expected number of progeny that will be produced by an individual of age x over the rest of its lifetime, given that it has survived to age x. Reproductive value is not the same as fitness, because it does not take into account the chances of surviving to age x.

3. *Hamilton's Perturbation Analysis*

Hamilton (1966) asked the following question: What sorts of small genetic changes in the $l(x)$ or $m(x)$ schedules will be favored by natural selection? To answer this question he employed the Malthusian parameter r as a measure of fitness, assuming that the modifications of $l(x)$ and $m(x)$ that lead to the highest value of r will be the ones to evolve. He also approximated the continuous functions described above with their discrete-time counterparts. The discrete-time rate of population increase is:

$$\lambda = \mathrm{e}^{r} \tag{3}$$

The discrete-time version of the Euler–Lotka equation is

$$\sum l_x m_x \lambda^{-x} = 1 \tag{4}$$

Age-specific survival is expressed in descrete time as:

$$l_x = p1\; p2\; p3 \ldots p_x \tag{5}$$

where p_x is the probability of surviving the duration of the xth age class given that one has survived to the beginning of age class x.

Now consider the evolutionary fate of a mutation which causes a small change in the ability to survive at some particular age a. The new mutation will be favored by natural selection if it causes an increase in r, or, what is equivalent in the discrete-time case, an increase in $ln\ \lambda$. The effect of the perturbation is studied by examining the partial derivative of λ with

respect to p_a. Hamilton obtained a closed form of this derivative and was able to conclude the following:

(a) The force of selection, as indicated by the partial derivative is highest at the youngest pre-reproductive ages, begins to decline when reproduction commences and drops to zero when reproduction ceases.

(b) If a mutation causes a gain in survival at a particular age a_1 and an equal loss in survival at age a_2, then such a mutation will increase in the population only if $a_1 < a_2$.

(c) If a mutation causes a gain in fertility at a particular age a_1 and an equal loss of fertility at age a_2, then such a mutation will increase in the population only if $a_1 < a_2$.

(d) If a mutation causes a loss in survival at a particular age and an increase in fertility at that same age, then the limits of the loss in survival that can be tolerated are set by the inverse of the reproductive value. That is, if the reproductive value at age x is large, then only a small reduction of survival can be exchanged for a gain in fertility, but if the remaining reproductive value is small, then a large reduction in survival can evolve.

(For further explication of these results, see Roughgarden 1996, p. 363.)

Hamilton's general conclusion is that 'for organisms that reproduce repeatedly senescence is to be expected as an inevitable consequence of the working of natural selection' (1966, p. 26). This is a view that is clearly consistent with Medawar (1952). For a technical discussion of the validity of the assumption that the Malthusian parameter is equivalent to fitness in age-structured populations, see Charlesworth (1980), whose models extend the early results of Haldane (1927) and Norton (1928).

4. *Pleiotropy*

Pleiotropy means that a single gene affects two or more characters. In the context of life history evolution, pleiotropy means that a single gene affects the fitness of the organism at two or more ages. It is convenient to categorize the combinations of age-specific pleiotropic effects as shown in Table 1. If a new mutation improves fitness in both young and old animals, then it is likely to be favored by natural selection, and will increase in the population. Conversely, a gene that decreases fitness in both young and old organisms will be eliminated by natural selection. The more interesting cases in Table 1 are those in which the fitness effects on young and old organisms are negatively correlated, a condition referred to as 'negative pleiotropy' or 'antagonistic pleiotropy.' Medawar's principle suggests that mutations that improve early fitness at the expense of late fitness will be favored by natural selection, while those with the converse effects will be eliminated.

Table 1
Pleiotropic effects of mutations affecting life history

Effect on fitness in the young	Effect on fitness in the old	Evolutionary fate
+	+	Increase
−	−	Decrease
−	−	Decrease
−	+	Decrease
+	−	Increase

Table 2
Fitness parameters for a one-locus model of antagonistic pleiotropy

Genotype:	A1A1	A1A2	A2A2
Fitness in young:	High	Medium	Low
Fitness in old:	Low	Medium	High

The possibility that genes might increase fitness at one age and also decrease it at another was mentioned by early theorists, but the first strong advocate of this mechanism of the evolution of senescence was Williams (1957), who noted that natural selection will tend to maximize vigor in the young at the expense of vigor later in life. An example of negatively pleiotropic gene action of the sort that Williams proposed is shown in Table 2. Williams argued that, in the course of selecting for the allele A1 which is beneficial at young ages, the deleterious effects of allele A2 on the old are brought along; in this scenario, senescence evolves as an incidental consequence of adaptation at earlier ages. The exact mathematical conditions for the increase of antagonistic, pleiotropic mutations have been derived (Charlesworth 1980), verifying that such mutations can indeed increase in populations.

While the theoretical basis for antagonistic pleiotropy is sound and widely accepted, it is unclear whether there exists the special sort of genetic variation that this mechanism requires. While it is easy to imagine physiological situations in which there could occur trade-offs between the fitness of the young and the old, there are few, if any, actual cases of such variation identified (Finch 1990, p. 37), even though a half century has passed since Williams' proposal. Negative correlations between life history characters are sometimes construed as evidence for pleiotropy, but this interpretation overlooks the fact that phenotypic correlations arise from factors other than pleiotropy, including the correlation of environmental factors and the correlation of alleles at genetically linked loci (linkage disequilibrium). What is required for the antagonistic pleiotropy model is not just evidence of trade-offs in life history traits, which is abundant, but a demonstration that there exist trade-offs in life history characters that are mediated by alternative alleles at specific polymorphic loci. Until such genes are characterized and shown to play a role

in life history evolution, the antagonistic pleiotropy model will remain an interesting theoretical construct, but one of unknown, and possibly negligible, biological significance.

5. *Mutation Accumulation*

Germ-line mutations, which are changes in the DNA sequence in sperm and egg cells, occur at low but nonzero rates, largely as a result of proof-reading errors in the enzymes that replicate DNA. This slow, steady input of genetic variants has the potential to corrupt the gene pool, since almost all of the novel variants that have some effect are deleterious. However, natural selection works against the corrupting effect by removing carriers of deleterious mutations. A balance is reached between the steady input of deleterious genes through mutation and their removal by natural selection. One of the characteristics of the equilibrium balance state is that, for any particular gene, the deleterious alleles are present at low frequencies, usually much less than 1 percent. The low rate of occurrence of each of many hereditary human diseases is thought to reflect the mutation-selection balance operating at many genes, each of which is capable of being mutated to produce a deleterious condition (Hartl and Clark 1997).

The classical mutation-selection balance model is appropriate for mutations that have deleterious effects early in life, but what happens when the disability is expressed only late in life? Medawar (1952) suggested that natural selection will be unable to counteract the feeble pressure of repetitive mutation if the mutant genes make their effects known at advanced ages, either post-reproductively or at ages not attained by most of the members of the species. This follows naturally from his proposal that the force of selection declines with increasing age. Under such conditions, the deleterious mutations would gradually accumulate, unchecked by natural selection. In this view, senescence is a process driven entirely by mutation. This mechanism for the evolution of senescence is distinct from, but not mutually exclusive of, antagonistic pleiotropy. While the pleiotropy process suggests that senescence is the incidental consequence of adaptation, the mutation accumulation model invokes deterioration without adaptation (Partridge and Barton 1993).

Charlesworth (1980) has analyzed a deterministic model of an age-structured population with recurrent mutation. He derived an approximation for the frequency of heterozygous carriers of deleterious alleles and found that the equilibrium frequency is inversely proportional to the selection intensity. The significance of this result is that when there is only very weak selection pressure, as at advanced ages, then mutant alleles can attain high frequencies under the influence of recurrent mutation. This result verifies the earlier conjectures of Medawar and Williams. In contrast to the situation with antagonistic pleiotropy, there is experimental evidence for the kinds of genetic variation that the model requires, namely spontaneous mutations with age-specific effects on vital rates (Mueller 1987, Pletcher et al. 1998, 1999).

6. *Postponement of Deleterious Effects*

Medawar (1952) also considered the case of Huntington's chorea, a grave and ultimately fatal nervous disorder that usually manifests itself in middle-aged patients. He suggested that there could be selection in favor of genetic modifiers which have as their main effect the postponement of the effects of the Huntington's gene or other genes causing hereditary disorders. This suggestion, which very much resembles an earlier proposal of R. A. Fisher concerning the evolution of dominance, is unlikely to be correct. While it makes sense that there would be some selection in favor of delaying the mutant effect, Charlesworth (1980, p. 219) has shown that the selection pressure exerted on such a hypothetical modifier gene would be exceedingly small, on the order of the mutation rate. This is because the modifier has an effect on fitness only when it co-occurs with the Huntington's or other disease gene, which is at mutation-selection balance and present in only a small fraction of the population. Under such conditions the evolutionary fate of the modifier is likely to be determined by genetic drift or other stochastic factors rather than the minuscule selective pressure.

7. *The Variation Problem*

While the primary concern of theorists has been to explain the degeneration of vitality associated with old age, there is a secondary problem that can also be addressed with these models. Genetic variation is the raw material upon which natural selection operates to produce adaptations and new species. The mechanisms by which variation is maintained in populations are therefore of considerable interest to evolutionary geneticists. To what extent do genetic models of senescence tend to maintain variation in life histories within populations? Several authors have addressed the question and come to two different answers depending upon the theoretical construct employed. Curtsinger et al. (1994) analyzed deterministic one- and two-locus models of antagonistic pleiotropy and asked under what conditions polymorphisms would be maintained. The conditions for stable polymorphism were found to be rather restrictive, especially with weak selection. The conditions were also found to be very sensitive to dominance parameters; in particular, reversal of dominance properties with respect to the two traits is often required for polymorphism, but seems improbable on biochemical grounds.

Tuljapurkar (1997) gives an overview of modeling strategies and describes some of his own models in which mortality is assumed to depend on both organismal age and random variables in the environment. In these models, the relative fitnesses are measured by a stochastic growth rate, which reflects average vital rates and environmental variability. Results from several related models show that phenotypic combinations that differ in age-specific fertility can be equally fit in a range of stochastic environments. The paper concludes that polymorphisms for length of reproductive life can be readily maintained by selection in temporally varying environments.

8. *Future Directions*

Two important challenges to the genetical theories of senescence arise from recent experimental work. The first challenge concerns mortality rates at advanced ages. Observations of survival in experimental organisms are usually presented in terms of age-specific survivorship, as defined in Sect. 2, but if sample sizes are sufficiently large then the survival data can also be analyzed in terms of hazard functions, which define the instantaneous risk of death as a function of age. Unlike survivorship, the hazard function can be nonmonotonic. Many experimental studies of moderate sample size have documented that the hazard increases approximately exponentially with age, a dynamic generally referred to as the Gompertz law (Finch 1990). Recent experiments have been done on an unusually large scale, making it possible to estimate hazards at very advanced ages. For Drosophila, nematode worms, and Med-flies, hazard functions increase exponentially in the early part of the life history, as expected, but at the most advanced ages the hazard functions decelerate, bending over and producing unexpected 'mortality plateaus' (see Vaupel et al. 1998, for a review of the experimental evidence and data on human populations).

The existence of mortality plateaus at advanced, post-reproductive ages poses a challenge for mutation accumulation models, which predict, under a wide range of assumption, a 'wall' of high mortality at the age when reproduction is completed. A preliminary attempt to accommodate mortality plateaus into antagonistic pleiotropy models has failed (Pletcher and Curtsinger 1998, Wachter 1999). One possible solution is that the mortality plateaus are caused by population heterogeneity of both genetic and nongenetic origin (Pletcher and Curtsinger 2000, Service 2000). It has so far proved very difficult to measure the relevant heterogeneity and determine whether it is of sufficient magnitude to produce the plateaus. A second possibility that can explain some features of the plateaus is a model of positive pleiotropy, which causes late-life mortality rates to avoid inflation because of the positively correlated effects of alleles selected for early survival (Pletcher and Curtsinger 1998). Models of positive pleiotropy merit further investigation.

The second experimental challenge to current theory concerns genetic variance. The mutation accumulation model predicts that genetic variance for mortality should increase at advanced ages. Recent experiments document instead a decline of genetic variance at advanced ages in experimental populations of Drosophila (Promislow et al. 1996). Hughes and Charlesworth (1994) initially reported that genetic variance for mortality increases with age in their Drosophila populations, but a re-analysis of the data show close concordance with the Promislow result (Shaw et al. 1999). At present no one knows why genetic variance declines at advanced ages; it could be related to the mortality plateaus described above.

Three other lines of research appear to hold promise. Group selection is a process by which collections of organisms succeed or fail as a collective, either becoming extinct or spawning new groups in competition with other groups. Groups selection arguments are sometimes invoked to explain altruistic behaviors, but evolutionary biologists typically disdain group selection, because the process tends to be much weaker than selection between individuals, and also seems to work only in very small populations (Williams 1966). However, group selection might play a role in shaping post-reproductive mortality rates, when individual-level selection is essentially inoperative. This type of model could be particularly relevant to human evolution, and is essentially unexplored. A related type of model that needs further development involves the evolution of vital rates in combination with kin selection, taking into account the effects of post-reproductive survival, parental care, and fitness effects mediated through relatives in the kin group (Roach 1992). Finally, as noted by Tuljapurkar (1997), the theoretical methods are limited to small perturbations and local analyses, under which conditions the population is always close to demographic equilibrium. There does not exist at present a theory that can accommodate large mutational changes in vital rates in combination with non-equilibrium demographics.

See also: Aging and Health in Old Age; Aging, Theories of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Differential Aging; Lifespan Development, Theory of; Old Age and Centenarians; Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

Bibliography

Charlesworth B 1980 *Evolution in Age-structured Populations*. Cambridge University Press, Cambridge, UK

Curtsinger J W, Service P, Prout T 1994 Antagonistic pleiotropy, reversal of dominance, and genetic polymorphism. *American Naturalist* **144**: 210–28
Finch C E 1990 *Longevity, Senescence, and the Genome*. University of Chicago Press, Chicago
Fisher R A 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK
Haldane J B S 1927 A mathematical theory of natural and artificial selection, Part IV. *Mathematical Proceedings of the Cambridge Philosophical Society* **32**: 607–15
Hamilton W D 1966 The moulding of senescence by natural selection. *Journal of Theoretical Biology* **12**: 12–45
Hartl D L, Clark A G 1997 *Principles of Population Genetics*, 3rd edn. Sinauer Associates Inc., Sunderland, MD
Hughes K A, Charlesworth B 1994 A genetic analysis of senescence in Drosophila. *Nature* **367**: 64–6
Medawar P B 1952 *An Unsolved Problem of Biology*. H K Lewis, London
Mueller L D 1987 Evolution of accelerated senescence in laboratory populations of Drosophila. *Proceedings of the National Academy of Sciences of the USA* **84**: 1974–77
Norton H T J 1928 Natural selection and Mendelian variation. *Proceedings of the London Mathematical Society* **28**: 1–45
Partridge L, Barton N H 1993 Optimality, mutation, and the evolution of ageing. *Nature* **362**: 305–11
Pletcher S D, Curtsinger J W 1998 Mortality plateaus and the evolution of senescence: Why are old-age mortality rates so low? *Evolution* **52**: 454–64
Pletcher S D, Curtsinger J W 2000 The influence of environmentally induced heterogeneity on age-specific genetic variance for mortality rates. *Genetical Research Cambridge* **75**: 321–9
Pletcher S D, Houle D, Curtsinger J W 1998 Age-specific properties of spontaneous mutations affecting mortality in *Drosophila melanogaster*. *Genetics* **148**: 287–303
Pletcher S D, Houle D, Curtsinger J W 1999 The evolution of age-specific mortality rates in *Drosophila melanogaster*: Genetic divergence among unselected lines. *Genetics* **153**: 813–23
Promislow D E L, Tatar M, Khazaeli A A, Curtsinger J W 1996 Age-specific patterns of genetic variance in *Drosophila melanogaster*. I. Mortality. *Genetics* **143**: 839–48
Roach D A 1992 Parental care and the allocation of resources across generations. *Evolutionary Ecology* **6**: 187–97
Roughgarden J 1996 *Theory of Population Genetics and Evolutionary Ecology. An Introduction*. Prentice Hall, Upper Saddle River, NJ
Service P M 2000 Heterogeneity in individual mortality risk and its importance for evolutionary studies of senescence. *American Naturalist*. **156**: 1–13
Shaw F, Promislow D E L, Tatar M, Hughes L, Geyer C 1999 Towards reconciling inferences concerning genetic variation in senescence in *Drosophila melanogaster*. *Genetics* **152**: 553–66
Tuljapurkar S 1997 The evolution of senescence. In: Wachter K W, Finch C E (eds.) *Between Zeus and the Salmon*. National Academy Press, Washington, DC, pp. 65–77
Vaupel J W, Carey J R, Christensen K, Johnson T E, Yashin A I, Holm N V, Iachine I A, Kanisto V, Khazaeli A A, Liedo P, Longo V D, Zeng Y, Manton K G, Curtsinger J W 1998 Biodemographic trajectories of longevity. *Science* **280**: 855–60
Wachter K W 1999 Evolutionary demographic models for mortality plateaus. *Proceedings of the National Academy of Sciences of the USA* **96**: 10544–7
Williams G C 1957 Pleiotropy, natural selection, and the evolution of senescence. *Evolution* **11**: 398–411
Williams G C 1966 *Adaptation and Natural Selection, A Critique of Some Current Evolutionary Thought*. Princeton University Press, Princeton, NJ

J. W. Curtsinger

Sensation and Perception: Direct Scaling

One aim of psychophysics is to measure the subjective intensity of sensation. We easily hear that a tone of given intensity has a particular loudness, and that a tone of higher intensity sounds louder still. The measurement problem lies in the question 'How much louder?' Direct scaling is a particular way of answering that question: the observer is asked to assign numbers corresponding to the subjective magnitudes of given stimulus intensities, thus providing the capability of saying that, for example, one tone sounds twice or ten times as loud as another.

1. *History*

1.1 *Fechner's Solution*

G. Fechner first proposed a solution, albeit an indirect one, in 1860 (Boring 1950). He accepted Weber's law—that the difference threshold (the physical size of a difference in intensity needed to tell one signal from another) grows in proportion to signal intensity, determining a constant of proportionality, the Weber fraction, unique to each sensory continuum. Fechner assumed that all difference thresholds are subjectively constant and therefore can serve as the unit of measurement for scales of sensation. He concluded that subjective magnitude grows by constant differences as stimulus magnitude grows by constant ratios, a logarithmic relation now known as Fechner's law.

1.2 *Category Rating and Magnitude Scaling*

In Fechner's laboratory, there also evolved a direct approach to measuring sensation magnitude. This was the method of absolute judgment (later called category rating), in which the subject assigns stimuli to categories according to some subjective aspect. Originally designed to study esthetic judgment, it was adapted by 1920 to studying sensation magnitude; early findings using this method supported Fechner's law.

In the 1950s, S. S. Stevens refined another and related direct scaling procedure in which the subject estimates the apparent magnitude of a signal by assigning to it a numerical value—hence, magnitude estimation. Data from this method were better fitted by a power law.

2. Methods

All the direct scaling methods require an observer to judge a series of stimuli varying along (at least) one dimension. For instance, the experimenter might present a pure tone at several sound pressure levels covering a 1000:1 range and request judgments of loudness, or present a light disk at several luminanoes covering a 10:1 range and request judgments of brightness.

2.1 Category Rating

The observer is instructed to assign each stimulus to a category, which may be designated descriptively (very soft, soft, … very loud) or numerically (1, 2, … , 7); the assignments should create subjectively equal category intervals. The allowable response values, including the extreme values, are specified by the experimenter. Usually the number of categories ranges from 5 to 11, although smaller and larger values are sometimes used. Typically, the stimulus range does not exceed 10:1, although it may do so.

2.2 Magnitude Estimation and Production

In magnitude estimation, the observer is instructed to assign to each stimulus a (positive) number so that the number reflects its apparent intensity. There are no limits on allowable response values. Originally, one stimulus value was designated as standard and a numerical value assigned to it as modulus. Later, however, the use of designated standard and assigned modulus was abandoned and the observer instructed to choose any numbers appropriate to represent apparent magnitudes. Typically the stimulus range is at least 100:1 and is often greater.

In magnitude production, stimulus and judgmental continua are interchanged; the experimenter presents a set of numbers, and the observer produces values on some intensive continuum to effect a subjective match for each presented number.

2.3 Cross-modal Matching

The use of numbers, which has provoked many objections, can be avoided by instructing the observer to match intensities on one physical continuum directly to intensities on another. For example, the observer may produce luminances to match sound pressure levels so that brightness of a light disk is equivalent to loudness of a tone, or the observer may produce handgrip forces to match odor intensities so that perceived effort is equivalent to strength of smell.

3. Achievements of Direct Scaling

3.1 A Quantitative Phenomenology

These methods made possible a quantitative phenomenology of the various sensory systems. For example, in hearing, the importance of salient variables, such as frequency and intensity, to the ability of the observer to detect a sound or to tell the difference between two sounds, had long been known. Furthermore, by having the observer match tones of varying frequency and intensity to a fixed reference tone, equal loudness contours were established. What remained unknown were the loudness relations among these contours. Magnitude estimation, by providing a numerical reference scale, permits the observation that one tone sounds twice, or ten times, or half as loud as another. Thus, magnitude estimation has been used to map the effects of variables known to influence perception for loudness, brightness, tastes and smells, and tactile sensations, as well as internal states such as perceived effort, fatigue, and satiety (Stevens 1975). It has also been used to study continua without a physical measure (e.g., seriousness of crime, severity of punishment, and their relationship) (Sellin and Wolfgang 1964). Category rating has been employed to study the cognitive integration of perceptual dimensions (Anderson 1981).

3.2 Intermodal Comparisons

Direct scaling also provides the capability to compare and contrast phenomena that occur in several modalities, such as temporal and spatial summation, sensory adaptation and recovery, and spatial inhibition (brightness contrast and auditory masking), at levels above absolute threshold (Marks 1974). For example, a persisting olfactory stimulus of constant intensity produces a smell intensity that diminishes in strength over time. That diminution can be assessed by asking an observer to assign numbers to the apparent magnitude of the smell at different elapsed times, e.g., after 5 s, 10 s, 30 s, 60 s, … ; in this way, the course of sensory adaptation can be traced for this stimulus and for others of differing initial intensities. When the procedure is repeated in a different modality, the parameters of the adaptation curves can be compared.

3.3 Stevens' Power Law

The third major achievement of direct scaling is the discovery that, to at least a first approximation, equal stimulus ratios produce equal judgmental ratios. This nearly universal relation is called Steven's power law, after S. S. Stevens who established it: subjective

magnitude is proportional to intensity raised to some power. Further, that exponent takes distinctive values for each stimulus continuum, ranging from 0.3 for luminance to 2.0 or more for electric shock.

Later experiments showed that the magnitude exponents predicted cross-modal matching exponents: if continua A and B had magnitude exponents a and b, and if a cross-modal match of B to A was obtained, the resulting exponent was a/b. Indeed, the cross-modal exponents for a variety of continua are connected in a transitive network (Daning 1983), a finding with important theoretical implications.

4. Problems of Direct Scaling

4.1 Use of Numbers as Responses

Critics questioned treating the numbers that subjects assigned to stimulus magnitudes as if they were measurements. However, the practice was validated by the discovery that the results obtained with number as the matching continuum are in agreement with the results obtained with cross-modal matching (Stevens, 1975). Furthermore, in cross-modal matching, some other stimulus continuum can substitute for number, since the choice of a reference continuum is, for most purposes, arbitrary. Whether the numbers used by an observer constitute a direct measure of sensation is an unresolved, and perhaps unresolvable, question (Teghtsoonian 1974).

4.2 The Psychophysical Regression Effect

Since magnitude estimation (assigning numbers to physical intensities) and magnitude production (assigning physical intensities to match numbers) are inverses of each other, they should produce the same exponent. However, this is not the case. The size of the difference in exponents may be quite large, with the result that a precise value of the exponent for a given stimulus continuum cannot be specified; no satisfactory combination rule has been agreed upon. The size, and indeed the direction, of the difference depends on the range of physical intensities (or numbers) presented; for practical purposes, the smallest effect is exhibited when the range is large. Some portion, at least, of this regression effect depends on sequential effects, the influence exerted by previous stimuli and judgments on subsequent judgments.

4.3 Category vs. Ratio Scales

Although both category rating and magnitude estimation purport to give scales of sensation magnitude, they do not, in general, agree with each other: both scales can be characterized as power functions, but the exponents are not the same. Much discussion has centred on whether subjects can make two kinds of judgments or whether the experimenter transforms a single kind of magnitude judgment by treating it as ratio or interval. However, an experimental determination of the sources of variance in the type of scale produced shows that instructions to judge either ratios or intervals account for almost none of the variance; much more important are such methodological variables as whether judgmental end-points are assigned or free and whether the range of stimuli is small or large. With free end-points and a large range, category judgment and magnitude estimation produce the same results; the obtained power functions have stable exponents characteristic of magnitude estimation (Montgomery 1975).

4.4 Local vs. Global Psychophysics

A long-standing conundrum in psychophysics has been the relation between the local (thresholds, both absolute and difference) and the global (measurements of subjective magnitude at levels above absolute threshold) (Luce and Krurnhansl (1988). Fechner believed he had combined the two by using difference thresholds as subjective units to yield a logarithmic scale of sensation. Stevens (1961) proposed, while honoring Fechner, to repeal his law and to substitute a power scale of sensation; he believed that one could not derive measures of sensation magnitude from threshold determinations. An early attempt to reconcile these two positions was R. Teghtsoonian's argument (1973) that both difference thresholds and power law exponents are indices of dynamic range. He proposed that there is a common scale of sensory magnitude for all perceptual continua and that the observer's dynamic range for each continuum maps completely onto that common scale. For example, the least sound intensity experienced has the same subjective magnitude as the least luminance, and the greatest sound intensity to which the auditory system responds has the same subjective magnitude as the greatest luminance. Thus the mapping of widely divergent dynamic ranges for the several perceptual continua onto a single subjective magnitude range determines their power law exponents, and the mapping of widely divergent difference thresholds onto a single subjective difference threshold determines their Weber fractions.

See also: Fechnerian Psychophysics; Memory Psychophysics; Psychophysical Theory and Laws, History of; Psychophysics; Scaling: Correspondence Analysis; Stevens, Stanley Smith (1906–73).

Bibliography

Anderson N H 1981 *Foundations of Information Integration Theory*. Academic Press, New York

Boring E G 1950 *A History of Experimental Psychology*, 2nd edn. Appleton-Century-Crofts, New York
Daning R 1983 Intraindividual consistencies in cross-modal matching across several continua. *Perception and Psychophysics* **33**: 516–22
Luce R D, Krumhansl C L 1988 Measurement, scaling, and psychophysics. In: Atkinson R C, Hernstein R J, Lindzey G, Luce R D (eds.) *Stevens' Handbook of Experimental Psychology, 2nd edn. Vol 1: Perception and Motivation*. Wiley, New York
Marks L E 1974 *Sensory Processes: The New Psychophysics*. Academic Press, New York and London
Montgomery H 1975 Direct estimation: Effect of methodological factors on scale type. *Scandinavian Journal Psychology* **16**: 19–29
Sellin J T, Wolfgang M E 1964 *The Measurement of Delinquency*. Wiley, New York
Stevens S S 1961 To honor Fechner and repeal his law. *Science* **133**: 80–6
Stevens S S 1975 *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. Wiley, New York
Teghtsoonian R 1971 On the exponents in Stevens's law and the constant in Ekman's law. *Psychology Review* **78**: 71–80
Teghtsoonian R 1974 On facts and theories in psychophysics: Does Ekman's law exist? In: Moskowitz H R, Scharf B, Stevens J C (eds.) *Sensation and Measurement: Papers in Honor of S. S. Stevens*. Reidel, Dordrecht, The Netherlands

M. Teghtsoonian

Sensation Seeking: Behavioral Expressions and Biosocial Bases

Sensation seeking is a personality trait defined as the tendency to seek varied, novel, complex, and intense sensations and experiences and the willingness to take risks for the sake of such experience (Zuckerman 1979, 1994). The standard measure of the trait is the Sensation Seeking Scale (SSS). In the most widely used version (form V) it consists of four subscales: (a) Thrill and Adventure Seeking (through risky and unusual sports or other activities); (b) Experience Seeking (through the mind and senses, travel, and an unconventional lifestyle); (c) Disinhibition (through social and sexual stimulation, lively parties, and social drinking); (d) Boredom Susceptibility (aversion to lack of change and variety in experience and people). Form V uses a total score based on the sum of the four subscales. More recently a single scale called 'Impulsive Sensation Seeking' has been developed which is a combination of items measuring the tendency to act impulsively without planning ahead, and adaptations of items from the SSS which assess the general need for excitement without mention of specific interest or activities (Zuckerman 1993).

Similar constructs and measures have been developed by other researchers including: change seeking, stimulus variation seeking, excitement seeking, arousal seeking, novelty seeking and venturesomeness. Most of these scales correlate very highly with the SSS and usually have the same kind of behavioral correlates.

Novelty seeking, a construct and scale devised by Cloninger (1987), not only correlates highly with Impulsive Sensation Seeking, but is based on a similar kind of biobehavioral model. Individuals high on novelty seeking are described as impulsive, exploratory, fickle, and excitable. They are easily attracted to new interests and activities, but are easily distracted and bored. Those who are low on this trait are described as reflective, rigid, loyal, stoic, frugal, orderly, and persistent. They are reluctant to initiate new activities and are preocccupied with details, and think very carefully and long before making decisions.

1. *Theoretical and Research Origins of the Construct*

The first version of the SSS was an attempt to provide an operational measure of the construct 'optimal levels of stimulation and arousal' (the level at which one feels and functions best). The author was conducting experiments on sensory deprivation (SD) in which participants volunteered to spend some length of time (from 1 to 24 hours in the author's experiments and up to two weeks in those of other investigators) in a dark soundproof room. It was theorized that persons with high optimal levels of stimulation would be most deprived by the absence of stimulation. Experiments show that the high sensation seeker became more restless over the course of an eight-hour experiment but did not experience more anxiety than the low sensation seekers (defined by the first version of the SSS). The construct changed as research using the SSS was extended to the broader domain of life experience.

2. *Behavioral Expressions*

Contrary to what we expected, the volunteers for the SD experiments were more commonly high sensations seekers than lows. This was because they had heard that people had weird experiences in SD, like hallucinations. High sensation seekers volunteered for any kind of unusual experiments, like hypnosis or drug studies. They did not volunteer for more ordinary experiments. The concept of sensation seeking centered more around the need for novel stimulation or inner experiences rather than stimulation *per se*. Surveys of drug use and preferences showed that high sensation seekers tended to be polydrug users of both stimulant and depressant drugs whereas lows did not use any drugs (Segal et al. 1980). It was the novel experience of drugs, not their effect on arousal, that

attracted high sensation seekers. Of course hallucinatory drugs had a particular attraction for those who scored high on this trait. They were not deterred by the legal, social, or physical risks entailed in drug use. Sensation seeking in preadolescents predicts their later alcohol and drug use and abuse.

Similarly sensation seekers proved to be attracted to risky sports that provided intense or novel sensations and experiences, like mountain climbing, parachuting, scuba diving, and car racing. They were not found in ordinary field sports or among compulsive exercisers. Outside of sports the high sensation seekers were found to drive their cars faster and more recklessly than lower sensation seekers.

High sensation seekers are attracted to exciting vocations such as firefighting, emergency room work, flying, air-traffic control, and dangerous military assignments. When they are stuck in monotonous desk jobs they report more job dissatisfaction than low sensation seekers.

In interpersonal relationships high sensation seekers tend to value the fun and games aspects rather than intimacy and commitment (Richardson et al. 1988). They tend to have more premarital sexual experience with more partners and engage in 'risky sex.' There is a high degree of assortative mating based on sensation seeking, i.e., highs tend to marry highs and lows tend to marry lows. Couples coming for marital therapy tend to have discrepant scores on the trait. Divorced persons are higher on the trait than monogamously marrieds.

Not all expressions of sensation seeking are risky. High sensation seekers like designs and works of art that are complex and emotionally evocative (expressionist). They like explicit sex and horror films and intense rock music. When watching television they tend to frequently switch channels ('channel surfing'). High sensation seekers enjoy sexual and nonsense types of humor (Ruch 1988). Low sensation seekers prefer realistic pastoral art pictures, media forms like situation comedies, and quiet background music.

3. *Psychopathology*

Sensation seeking is an essentially normal trait and most of those who are very high or low on the trait are free from psychopathology. However, persons with certain kinds of disorders involving a lack of impulse control tend to be high sensation seekers. Sensation seeking is a primary motive for those with antisocial personality disorder; their antisocial behavior involves great risks where the only motive is sometimes the increase of excitement. Other disorders with a high number of sensation seekers include those with conduct disorders, borderline personality disorders, alcohol and drug abusers, and bipolar (manic-depressive) disorders. It has been discovered that there are genetic and biological trait links between many of these disorders and sensation seeking. For instance, as yet unaffected children of bipolars tend to be high sensation seekers as well as showing similar differences on an enzyme to be discussed.

4. *Genetics*

Studies of twins raised in intact families show a relatively high degree of heritability (60 percent), compared to other personality traits that are typically in the range 30–50 percent. Analyses show no effects for the shared family environment; the environment that is important is that outside of the family which affects each twin differently. A study of twins separated shortly after birth and adopted into different families confirms these results (Hur and Bouchard 1997).

A specific gene has been found to be related to novelty seeking (Ebstein et al. 1996). The gene produces one class of receptors for the neurotransmitter dopamine. One of the two major forms of the gene is found more often in high sensation seekers. This form of the gene has also been found in high percentages of heroin abusers, pathological gamblers, and children with attention deficit hyperactivity disorder. The neurotransmitter dopamine as well as the other two monoamines in the brain, norepinephrine and serotonin, are theorized to underlie the three behavioral mechanisms involved in sensation seeking: strong approach, and weak arousal and inhibition.

5. *Biochemistry*

Males who are high in sensation seeking trait have high levels of the hormone testosterone compared to average levels in lower sensation seekers (Daitzman and Zuckerman 1980). This finding is consistent with the difference between men and women in the trait and the finding that sensation seeking peaks in the late teens and declines with age in both sexes.

The enzyme monoamine oxidase (MAO) type B is lower in high sensation seekers than in those who score low on the trait. This is also consistent with age and gender differences since women are higher than men on MAO at all ages and MAO rises in the brain and blood platelets with age. Type B MAO is a regulator of the monoamines, particularly dopamine, and low levels imply a lack of regulation perhaps related to the impulsivity characteristic of many high sensation seekers. Low levels of MAO are also found in disorders characterized by poor behavioral control: attention deficit hyperactivity disorder, antisocial and borderline personality disorders, alcoholism, drug abuse, pathological gambling disorder, and mania (bipolar disorder). MAO is part of the genetic predisposition for these disorders as shown by the finding that the enzyme is low in as yet nonaffected children of alcoholics and those with bipolar disorder. Evidence

of behavioral differences in newborn infants related to MAO levels also show the early effects on temperament. MAO differences are also related to behavioral traits in monkeys analogous to those of high and low sensation seeking humans.

6. Psychophysiology

Differences in the psychophysiological responses of the brain and autonomic nervous system as a function of stimulus intensity and novelty have been found and generally replicated (Zuckerman 1990). The heart rate response reflecting orienting to moderately intense and novel stimuli is stronger in high sensation seekers than in lows, perhaps reflecting their interest in novel stimuli (experience seeking) and disinterest in repeated stimuli (boredom suceptibility).

The cortical evoked potential (EP) reflects the magnitude of the brain cortex response to stimuli. Augmenting–reducing is a measure of the relationship between amplitude of the EP as a function of the intensity of stimuli. A high positive slope (augmenting) is characteristic of high sensation seekers (primarily those of the disinhibition type) and very low slopes, sometimes reflecting a reduction of response at the highest stimulus intensities (reducing), is found primarily in low sensation seekers. These EP augmenting–reducing differences have been related to differences in behavioral control in individual cats and strains of rats analogous to sensation seeking behavior in humans (Siegel and Driscoll 1996).

7. Evolution

Comparative studies of humans and other species using the same biological markers suggest that the trait of impulsive sensation seeking has evolved in the mammalian line (Zuckerman 1984). Exploration and foraging is risky but adaptive in species that must move frequently to avoid exhaustion of the resources in an area. The balance between sensation seeking and fear determines exploration of the novel environment. Our own hominid species that came out of Africa and settled the entire earth in about 100,000 years had to have at least a moderate degree of sensation seeking. The hunting of large animals by men, warfare, and the seeking of mates outside of the group all involved risks which may have been overcome by the sensation-seeking pleasure in such activity. Individual differences in the trait are seen in human infants before the major effects of socialization are seen. This suggests that impulsive sensation seeking is a basic trait related to individual differences in the approach, arousal, and inhibition behavioral mechanisms in humans and other species.

8. Future Directions

One gene (DRD4) has been associated with sensation seeking but it only accounts for 10 percent of the genetic variance. The search for other major genes will continue and the understanding of what these genes do in the nervous system will fill in the large biological gap between genes and sensation-seeking behavior. Longitudinal studies that begin with genetic markers like the DRD4, biological markers like MAO, and behavioral markers like reactions to novelty, will be used to find out how specific environmental factors interact with dispositions to determine the expressions of sensation seeking, for instance why one sensation seeker becomes a criminal and another a firefighter who does skydiving on weekends..

See also: Gender Differences in Personality and Social Behavior; Genetic Studies of Personality; Personality and Crime; Personality and Risk Taking; Temperament and Human Development; Temperament: Familial Analysis and Genetic Aspects

Bibliography

Cloninger C R 1987 A systematic model for clinical description and classification of personality variants. *Archives of General Psychiatry* **44**: 573–88

Cloninger C R, Sigvardsson S, Bohman M 1988 Childhood personality predicts alcohol abuse in young adults. *Alcoholism: Clinical and Experimental Research* **12**: 494–505

Daitzman R J, Zuckerman M 1980 Disinhibitory sensation seeking, personality, and gonadal hormones. *Personality and Individual Differences* **1**: 103–10

Ebstein R P, Novick O, Umansky R, Priel B, Oser Y, Blaine D, Bennett E R, Nemanov L, Katz M, Belmaker R H 1996 Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait, novelty seeking. *Nature Genetics* **12**: 78–80

Hur Y, Bouchard T J Jr 1997 The genetic correlation between impulsivity and sensation seeking traits. *Behavior Genetics* **27**: 455–63

Richardson D R, Medvin N, Hammock G 1988 Love styles, relationship experience, and sensation seeking: A test of validity. *Personality and Individual Differences* **9**: 645–51

Ruch W 1988 Sensation seeking and the enjoyment of structure and content of humor: Stability of findings across four samples. *Personality and Individual Differences* **9**: 861–71

Segal B S, Huba, G J, Singer J L 1980 *Drugs, Daydreaming and Personality: Study of College Youth*. Erlbaum, Hillsdale, NJ

Siegel J, Driscoll P 1996 Recent developments in an animal model of visusal evoked potential augmenting/reducing and sensation seeking behavior. *Neuropsychobiology* **34**: 130–5

Teichman M, Barnea Z, Rahav G 1989 Sensation seeking, state and trait anxiety, and depressive mood in adolescent substance users. *International Journal of the Addictions* **24**: 87–9

Zuckerman M 1979 *Sensation Seeking: Beyond the Optimal Level of Arousal*. Erlbaum, Hillsdale, NJ

Zuckerman M 1984 Sensation seeking: A comparative approach to a human trait. *Behavioral and Brain Sciences* **7**: 413–71

Zuckerman M 1990 The psychophysiology of sensation seeking. *Journal of Personality* **58**: 313–45
Zuckerman M 1993 Sensation seeking and impulsivity: A Marriage of traits made in biology? In: McCown W G, Johnson J L, Shure M B (eds.) *The Impulsive Client: Theory, Research, and Treatment*. American Psychological Association, Washington, DC, pp. 71–91
Zuckerman M 1994 *Behavioral Expressions and Biosocial Bases of Sensation Seeking*. Cambridge University Press, New York

M. Zuckerman

Sensitive Periods in Development, Neural Basis of

In all animals including man, the organization of brain and behavior is not fully determined by genetic information. Information from the environment, mediated by sensory organs, plays an important role in shaping the central nervous system (CNS) or a given behavior to its adult appearance. Many studies have shown that the environmental influence on development is not only different between species, but also varies along the time of development. In many cases, it is only a very short period of time where external stimulation affects the development. The same stimulus, given earlier or later in life, may have no effect or no visible effect. This phenomenon, a time limited influence of external stimulation on the wiring of the CNS or on the performance of a given behavior, has been called, with only subtle differences in meaning, 'sensitive' or 'critical periods,' or 'sensitive phases.' I shall use 'sensitive period' here just because it is the most frequent term.

1. *Examples of Developmental Phenomena with Sensitive Periods*

1.1 *Imprinting and Song Learning*

Probably the best known example of early external influences of the environment on the organization of behavior is the so-called 'imprinting' process (Lorenz 1935), by which a young bird restricts its social preference to a particular animal or object. In the course of 'filial imprinting,' for example, a young chick or duck learns about the object that it has followed when leaving the nest (Hess 1973). Young zebra finches, in the course of 'sexual imprinting' (Immelmann 1972), learn the features of an object that subsequently releases courtship behavior in fully-grown birds. In addition to these two phenomena, many other paradigms of imprinting have been described, as for example homing in salmons, habitat imprinting, acoustic imprinting, or celestial orientation in birds.

All imprinting phenomena are characterized by at least two criteria (Immelmann and Suomi 1982): First, learning about the object which the bird is following later on, or to which courtship behavior is directed, is restricted to a sensitive period early in development. In the case of filial imprinting, this phase is quite short (several hours) and it starts directly after hatching. In sexual imprinting, which has been investigated mainly in birds, which are hatching underdeveloped and with closed eyes, the sensitive period starts at the day of eye opening, and may last for several days. The second feature of all imprinting paradigms examined so far is that the information storage is rather stable. The preference for an object to follow or to court, which has been established in the course of the sensitive period, cannot be altered later on. Whenever the bird has a chance to choose between the imprinted object and another one, it will choose the familiar imprinted object.

Song learning is, at the first glance, a little bit more complicated than imprinting. It has been shown to comprise two parts (Konishi 1965, Marler 1970). Early in life a young male bird (only males, in most avian species, sing) learns about the song he himself is singing later when adult, and he is learning mainly from his father. At the time of learning, the young male is not able to sing by himself. This time span is called the *acquisition* period, and it is thought that the male is storing some template of the song he has heard during this phase of learning. When the bird grows older, he starts singing by himself, and it has been shown that he tries to match his own song with the previously acquired template. During this 'crystallization period,' the young male selects its final song from a bigger set of other songs that he was singing at the beginning. Thereafter, this selected song remains stable and shows only minor variation. The other songs are, in most cases, not uttered any longer.

Song learning thus shows the same characteristics as imprinting. It occurs during a sensitive period, and after the crystallization period, the song that is selected cannot easily be altered. Recent research indicated that a second event like crystallization could also be demanded in imprinting. At least formally, one can separate an acquisition period (which is the 'classical' sensitive period) and a second event that may be called stabilization also in imprinting (Bischof 1994).

As already mentioned, development is an interplay between genetic instruction and acquired information. This is also the case in imprinting and song learning (Bischof 1997). In imprinting, not only the behavior for which the object is learned is genetically determined, there are also some genetic constraints which at least lead to advantages for certain objects to be learned easier than others. In filial imprinting, it has been shown for example that there is a natural preference for red over other colors, and this prefer-

ence can be enhanced or diminished by selective breeding. In song learning, only a small variety of songs can be learned by most species, and there is some indication that certain features of song are innate and not alterable by early learning.

1.2 Plasticity of the Visual Cortex in Cats

The best-known example of sensitive periods in neural development is the plasticity of neurons of the visual cortex in cats (Hubel and Wiesel 1970). In the adult cat, most neurons in area 17 are driven by visual stimulation of the left as well as the right eye, and are thus defined as binocular. If one eye of a kitten is briefly sutured closed in its early postnatal life, the access of the eyes to cortical neurons is dramatically altered. There is an obvious lack of binocular neurons in the visual cortex of such kittens, and most of the neurons are driven exclusively or are at least strongly dominated by, the non-deprived eye. These changes in 'ocular dominance distribution' are only observed if monocular deprivation occurs during postnatal development; the same deprivation in an adult cat does not cause any change. Thus, there is a sensitive period during which the alteration of the visual input affects the wiring of neurons within the visual cortex. The wiring, which has been established after the end of this sensitive period, remains stable for the rest of life. Most of the results obtained in cats were later on confirmed and extended by research on other animals, including monkey, ferret, rat, and mouse (Berardi et al. 2000).

As in imprinting and song learning, there is evidence that in addition to early learning, genetic instruction plays an important role for the organization of the visual cortex. The basic pattern of wiring is already there at birth, and this basic pattern is either stabilized during the sensitive period if the sensory information is adequate, or can be altered if the information coming from the eyes deviates from normal. For example, if one eye is closed as in the experiments above, or if the eyes are not aligned appropriately as it is the case in squinting.

2. Similarity of Events with Sensitive Periods on the Behavioral and on the Neuronal Level

The short overviews in Sect. 1 already show that in addition to the fact that the environment affects the organization of brain and behavior only during a sensitive period, there are three features that are similar in all examples. First, there is some genetic preorganization that is maintained or altered by sensory information during the sensitive period. Second, the shape of the sensitive period over time is very similar in all cases: the effect of environmental information increases quickly to a maximum, and becomes smaller much slower, going asymptotically to zero. Third, if the sensitive period is over, the brain structure or the behavior is not easily altered again. However, so far our comparison was only on phenomena, the effects may be due to very different mechanisms. It is therefore necessary to go into more detail and to discuss the mechanisms by which storage of information occurs in the different paradigms, and how the start, the duration and the end of sensitive periods, respectively, is determined. Sects. 2.1–2.3 will show that there is indeed a lot of similarity on the level of mechanisms.

2.1 Control of Sensitivity

The control of the time span over which external information is able to affect brain and behavior, was initially thought to be due to a genetically determined window which was opened for some time during development allowing external information to access the CNS. However, this turned out to be too simple an idea because the environmental sensitivity could be shifted in time due to experimental conditions (Bateson 1979). Dark rearing, for example, delays the sensitive period during which ocular dominance can be shifted by monocular deprivation, and the sensitive period for imprinting lasts longer if the young bird is isolated and thus not able to see the appropriate stimulus which leads to imprinting. In imprinting, one can also show that the sensitive period is prolonged if the stimulus is not optimal. For example, exposing a zebra finch male on another species like a Bengalese finch leads to a prolongation of the sensitive period.

The ideas, which were raised to explain these phenomena, were as follows (Bischof 1997): the natural onset of the sensitive periods coincides with the functioning of the sensory systems involved. Thus, the sensitive period for sexual imprinting in zebra finches starts with about 10 days when the eyes are fully open, while in the precocial chicks, filial imprinting starts directly after birth because these birds are born with open eyes. The sensitive period for monocular deprivation should also start at eye opening. This is roughly correct, but the effect is quite low at the beginning, and some recent results indicate that eye opening may trigger some intermediate events that then lead to enhanced sensitivity of the affected area of the brain.

Why does the sensitive period have a time course with a quite sharp increase to a maximum of sensitivity, but a slow, asymptotic decrease? One idea is that the sensitive period is some self-terminating process. If we suppose that an object is described by a limited amount of information bits, or, on the storage side, there is a limited store for the information which has to be acquired, it is easy to imagine that the probability for storage of a given bit of information is

high at the beginning and goes asymptotically to zero, dependent on the amount of information already stored (Bischof 1985).

2.2 Sites and Modes of Storage

For all the examples of phase specific developmental phenomena in Sect. 1, the locations within the brain are known where the plastic events can be observed. Therefore, it is possible to compare the changes of wiring between different examples. It was the visual cortex plasticity where it was detected first that the anatomical basis for the development of neuronal specificity was a segregation of previously overlapping neuronal elements within area 17 (LeVay et al. 1978). Thus, the specification of neurons was caused by a reduction of preexisting neuronal elements. This principle was also found in imprinting and song learning. In both paradigms, the spine density of neurons within the areas that are involved in the learning process was substantially reduced in the course of the sensitive period (Bischof 1997). This indicates that pruning of preexisting elements is an essential part of the physiological mechanisms underlying phase dependent developmental plasticity and learning. The reduction of spine density is stable thereafter; it can not be enhanced by any treatment when it has occurred once. This is also an indication that the reduction of spines may be the anatomical basis of imprinting like learning and developmental plasticity.

2.3 Ultrastructural Events

It has long been speculated that the machinery causing learning induced changes during imprinting, song learning, and cortical plasticity may be different from that causing changes in adult learning, because there is a decrease in spine density instead of an increase, and the changes are stable and cannot be reversed. However, concerning the basic machinery, no significant differences were found. Developmental learning can obviously also be explained by 'Hebbian' synapses which strengthen their connections if pre- and postsynaptic neurons fire together, and disconnect if the activity is asynchronous. To cause changes in the postsynaptic neuron, NMDA receptors are involved as in other learning paradigms, causing Long Term Potentiation (LTP) or Long Term Depression (LTD) as well as the cascades of second messengers which finally lead to the activation of the genome which then causes long term changes in synaptic efficiency.

The difference to adult learning obviously lies in the fact that plasticity is limited to a certain time span. Many ideas have been developed which systems may gate plasticity (Katz 1999). One of the earliest ideas was that myelination delimits plasticity. Unspecified projections, adrenergic, serotoninergic or cholinergic have been shown to play a role where it was investigated. Recent links from experiments with knockout animals point towards neurotrophic agents which may only be available to a limited amount; when the resource is exhausted, plasticity is no longer possible (Berardi and Maffei 1999). Another very interesting finding is that inhibition plays a role (Fagiolini and Hensch 2000); neurons may have to reach a genetically determined balance of inhibition and excitation to become plastic. Whether these latter ideas can also be applied to the other early learning paradigms has to be examined.

3. Is Generalization to Humans Possible?

One has to be very careful if one generalizes examples from one species to another, and this is even more true for generalization from animals to humans. However, there are some hints that at least part of the results described in Sect. 2 can be applied to humans (Braun 1996). On the neuronal level, it is globally accepted that amblyopia, a visual deficit, is based on the mechanisms described for the development of the visual cortex. It has been shown that if one corrects in humans the misalignment of the eyes which causes amblyopia during early development, the connection between eyes and cortical neurons, and this is no longer possible in adults (Hohmann and Creutzfeldt 1975).

On the behavioral level, it has been shown that language learning has so much in common with song learning (Doupe and Kuhl 1999) that it is intriguing to speculate that the neuronal machinery may be similar in both paradigms. However, one has to be aware that the similarity is as yet only on the phenomenological level.

Since the early days of imprinting research, it has also been discussed whether aggressiveness, social competence, and similar things are imprinted (Leiderman 1981), and whether this can also be applied to humans. The frightening idea was that in this case parents could easily make big mistakes if they did not confront their children with the appropriate surrounding. However, evidence is sparse even in animals that social competence is severely influenced by early experience. If there is an influence, ways are available, even in the case of sexual imprinting in the zebra finch, to at least overcome temporarily the effects of imprinting. However, that imprinting effects are in most cases only covered but not eliminated, may be reason enough to pay some attention to the conditions under which children grow up.

See also: Birdsong and Vocal Learning during Development; Brain Development, Ontogenetic Neurobiology of; Neural Plasticity in Visual Cortex;

Prenatal and Infant Development: Overview; Visual Development: Infant

Bibliography

Bateson P 1979 How do sensitive periods arise and what are they for? *Animal Behaviour* **27**: 470–86

Berardi N, Maffei L 1999 From visual experience to visual function: Roles of neurotrophins. *Journal of Neurobiology* **41**: 119–26

Berardi N, Pizzorusso T, Maffei L 2000 Critical periods during sensory development. *Current Opinion Neurobiology* **10**: 138–45

Bischof H-J 1985 Environmental influences on early development: A comparison of imprinting and cortical plasticity. In: Bateson P P G, Klopfer P H (eds.) *Perspectives in Ethnology, Vol. 6. Mechanisms*. Plenum Press, New York, pp. 169–217

Bischof H-J 1994 Sexual imprinting as a two-stage process. In: Hogan J A, Bolhuis J J (eds.) *Causal Mechanisms of Behavioural Development*. Cambridge University Press, Cambridge, UK, pp. 82–7

Bischof H-J 1997 Song learning, filial imprinting, and sexual imprinting: Three variations of a common theme? *Biomedical Research–Tokyo* **18**(Suppl. 1): 133–46

Braun K 1996 Synaptic reorganization in early childhood experience and learning processes: Relevance for the development of mental diseases. *Zeitschrift für Klinische Psychologie Psychiatrie und Psychotherapie* **44**: 253–66

Doupe A J, Kuhl P K 1999 Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience* **22**: 567–631

Fagiolini M, Hensch T K 2000 Inhibitory threshold for critical-period activation in primary visual cortex. *Nature* **404**: 183–6

Hess E H 1973 *Imprinting: Early Experience and the Developmental Psychobiology of Attachment*. Van Nostrand Reinhold, New York

Hohmann A, Creutzfeldt O D 1975 Squint and the development of binocularity in humans. *Nature* **254**: 613–4

Hubel D H, Wiesel T N 1970 The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *Journal Physiology* (*London*) **206**: 419–36

Immelmann K 1972 The influence of early experience upon the development of social behaviour in estrildine finches. *Proceedings of the 15th International Ornithological Congress. The Hague* 316–38

Immelmann K, Suomi S J 1982 Sensitive phases in development. In: Immelmann K, Barlow G W, Petrinovich L, Main M (eds.) *Behavioural Development*. Cambridge University Press, Cambridge, UK, pp. 395–431

Katz L C 1999 What's critical for the critical period in visual cortex? *Cell* **99**: 673–6

Konishi M 1965 The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Zeitschrift für Tierpsychologie* **22**: 770–83

Leiderman P H 1981 Human mother-infant social bonding: Is there a sensitive phase? In: Immelmann K (ed.) *Behavioral Development. The Bielefeld Interdisciplinary Project*. Cambridge University Press, Cambridge, UK

Le Vay S, Stryker M P, Shatz C J 1978 Ocular dominance columns and their development in layer IV of the cat's visual cortex: A quantitative study. *Journal of Comparative Neurology* **179**: 223–44

Lorenz K 1935 Der Kumpan in der Umwelt des Vogels. *Journal für Ornithologie*. **83**: 137–213, 289–413

Marler P 1970 A comparative approach to vocal learning: Song development in white-crowned sparrows. *Journal of Comparative and Physiological Psychology*. **71**(Suppl.): 1–25

H.-J. Bischof

Sentence Comprehension, Psychology of

In the process of mapping from form (speech or printed text) to meaning, listeners and readers have the task of combining individual word meanings into sentence meanings. This article examines the cognitive challenges posed by this task, describes some experimental techniques that psycholinguists have used to understand the task, summarizes some basic empirical phenomena of sentence comprehension, and surveys the range of cognitive theories that have been developed to explain how people comprehend sentences.

1. *The Tasks of Sentence Comprehension*

Listeners, and skilled readers, recognize individual words in a manner which appears effortless but which actually hides a wealth of complex cognitive processes (e.g., see *Lexical Access, Cognitive Psychology of*; *Word Recognition, Cognitive Psychology of*). They can go further and identify the message being conveyed by a discourse, they can create a mental model of the scenario being described, and they can engage in further social interaction with the person who produced the words.

The gap between word and message is bridged by the process of sentence comprehension. A sentence conveys meaning. The meaning of a sentence is composed out of the meanings of its words, guided by the grammatical relations that hold between the words of the sentence. The psychology of sentence comprehension is concerned with the cognitive processes that permit a reader or listener to determine how the word meanings are to be combined in a way that satisfies the intention of the writer or speaker.

The reader/listener's task of mapping from print or sound to meaning is not trivial. Small differences in the input can make great changes in meaning. Consider the contrast between 'Tom gave his dog biscuits' and 'Tom gave him dog biscuits.' The difference between 'his' and 'him' signals a difference in the 'case' of the pronoun, an aspect of its morphology that carries information about the relation it has to a verb or some other 'case-assigning' word in a sentence. The form 'him' signals the accusative case, which means that the pronoun has some immediate relation to the verb, in this case the indirect object or 'recipient' of the action denoted by the verb. The form 'his,' on the other hand, signals the possessive or 'genitive' case, which means

that the pronoun is immediately related to the following noun 'dog.' This whole noun phrase in turn takes the recipient role. Morphologically signaled case actually plays a minor role in English, although it plays a very important role in many other languages. Languages like English signal the structural relations between their words primarily by word order (compare 'The dog bit the man' and 'The man bit the dog'), and English, like all languages, signals structural relations by the choice of particular lexical items (compare 'The man helped the child to first base' and 'The man helped the child on first base').

Listeners and readers must be sensitive to subtle pieces of information about the structure of sentences. Their task is complicated by the fact that they must also be sensitive to structural relations that span arbitrarily long distances. Consider sentences like 'The boy likes the girl,' 'The boy from the small town likes the girl,' 'The boy from the small town where I grew up likes the girl,' and so forth. There is no limit to the amount of material that can be included in the sentence to modify 'the boy.' Nonetheless, the form of the final verb, 'likes,' must agree in number with the initial subject, 'the boy.' Or consider a sentence like 'Which girl did the boy from the small town … like?' The initial question phrase, 'which girl,' must be understood to be the direct object of the final verb 'like,' even though the two phrases can be separated by an arbitrarily long distance. Readers and listeners (and writers and speakers) are clearly sensitive to such 'long distance dependencies' (cf. Clifton and Frazier 1989, Fodor 1978).

A third problem faced by listeners and readers is the ubiquitous ambiguity (temporary or permanent) of language. Ambiguity begins in the speech stream, which often can be segmented at different points into different words (see, e.g., *Speech Perception*, see also Cutler et al. 1997). A given word form can correspond to different lexical concepts (e.g., the money vs. river meaning of 'bank'). And a string of words can exhibit different structural relations, which may or may not be resolved by the end of a sentence (consider, e.g., the temporary ambiguity in 'I understood the article on word recognition was written by a real expert'). Readers and listeners are generally not aware of such ambiguities (although puns force awareness), but that only means that their cognitive processes adequately resolve the ambiguities in the course of sentence comprehension.

1.1 The Role of Psycholinguistics in the Development of Cognitive Psychology

The modern study of sentence comprehension began at the start of the cognitive revolution, in the late 1950s and early 1960s. Prior to this time, complex behaviors were claimed to be the result of (possibly complex) stimulus–response contingencies. Sentences were presumed to be produced and understood by chaining strings of words together, under the control of environmental and internal stimuli.

Psycholinguistic phenomena of sentence comprehension (and language acquisition; see e.g., *Language Acquisition*) demonstrated the inadequacy of stimulus–response chaining mechanisms. The sheer fact that we are always understanding and producing completely novel sentences was hard enough to explain in a stimulus–response framework, but even sharper arguments against behaviorism were constructed. Consider just one example. Long-distance dependencies, discussed above, could not be analyzed in terms of stimulus–response chains. If stimuli are chains of words, then the effective stimulus that links a 'which' question phrase with the point in the sentence where it has to be interpreted (often referred to as a 'gap') would have to be different for each added word that intervenes between the 'which' word ('stimulus') and its gap ('response'). In the limit, this means that an arbitrarily large (potentially infinite) number of stimuli would have been associated with a response, which is not possible within stimulus–response theory.

1.2 The Role of Linguistic Knowledge in Sentence Comprehension

Arguments like the one just given were among the strongest reasons to develop a psychology of cognitive processes (see Miller 1962). Cognitive psychologists took arguments like this to support the claim that the mind had to have processes that operated on structures more abstract than stimulus–response chains. Many psycholinguists took the new generative grammars being developed by Chomsky (1957, see also Chomsky 1986) to provide descriptions of the structures that cognitive processes could operate on. One important structure is phrase structure, the division of a sentence into its hierarchically arranged phrases and the labeling of these phrases. For example, 'I understood the article on word recognition was written by an expert' is divided into the subject 'I' and the verb phrase 'understood … expert'; this verb phrase is divided into the verb 'understood' and its embedded sentence complement 'the article … was written by an expert'; this sentence is similarly divided into subject and verb phrase, and so forth. Another structure (identified only after several years of development of linguistic theory) is the long-distance dependency between 'moved items' or 'fillers' and their 'traces' or 'gaps' (as in the relation between a 'which' phrase and its gap, discussed earlier). Still other structures involve case relations (e.g. the distinction between accusative and genitive case discussed earlier) and thematic role relations (e.g., the distinction between theme or affected object and recipient in 'Tom gave his dog biscuits').

Early psycholinguists devoted much attention to the 'psychological reality' of grammatical structures, arguing whether they are really involved in sentence

comprehension (Fodor et al. 1974). As psycholinguistic theory developed, it became apparent that the real debate involves not whether these structures are real but how they are identified in the course of sentence comprehension and how, in detail, the mind creates and represents them (Frazier 1995). It now seems impossible to explain how we understand sentences without theorizing about how people assign structure to sentences. However, as we will see later, there is very lively debate about just how we do identify sentence structure.

2. *Ways of Studying Sentence Comprehension*

Sentence comprehension seems almost effortless and automatic. How can one observe the fine-grain details of such a smoothly flowing process? Early psycholinguists focused on what was remembered once a sentence was comprehended. They learned some interesting things. Grammatical structure can predict which sentences will be confused in recall and recognition and it can predict what words from a sentence will be good recall probes. While these findings supported the 'psychological reality' of grammatical structures, other findings indicated that the gist of a sentence was typically more salient in memory than its specific form or even its grammatical structure. (See Tanenhaus and Trueswell 1995 for a more detailed review.)

2.1 *Online Measures*

Showing that the end state of comprehension is gist tells us little about how meaning is arrived at. The early psycholinguists recognized this, and developed 'online' tasks to probe the early stages of sentence comprehension. Some used a technique of sentence verification, measuring the time taken to decide whether a sentence was true or false (e.g., when spoken of a picture). Others (e.g., Just et al. 1982, see also Haberlandt 1994) measured the time readers took to read each word or phrase of a sentence when sentence presentation was under their control (pressing a button brings on each new word), thereby getting a more precise look at the difficulty readers experience word by word and allowing the development of theories about how readers construct interpretations of sentences. The development of techniques for measuring eye movements during reading (Rayner et al. 1989; see also *Eye Movements in Reading*), coupled with the discovery that the eye generally seems to be directed toward whatever input the reader is processing, allowed even more sensitive and less disruptive ways of measuring the course of reading.

In addition to devising techniques for measuring which parts of sentences readers found easy or hard, psycholinguists have developed ways to assess just what a reader or listener might be thinking about at any given time. One can interrupt a spoken sentence with a probe (e.g., a word to name or to recognize from earlier in the sentence) and use probe reaction times to make inferences about how highly activated the probe word is. For example, probes for fillers such as 'girl' in 'Which girl did the boy from the small town like' are sometimes observed to be faster after the word that assigns it a thematic role ('like') than at other points in the sentence, as if the filler was reactivated at the gap. (Note that this technique is methodologically tricky and can mislead a researcher if it is misused; McKoon et al.1994.) One can also measure ERPs (event-related brain potentials, electrical traces of underlying brain activities measured at the scalp) at critical points in sentences, for example, when a sentence becomes implausible or ungrammatical or simply difficult to process (Kutas and van Petten 1994). There appear to be distinct signatures of different types of comprehension disruption. For instance, implausibility or violation of expectations commonly leads to an 'N400,' a negative-going potential shift peaking about 400 ms. after the introduction of the implausibility. This allows a researcher to test theories about just when and how disruption will appear in hearing or reading a sentence.

3. *Phenomena of Sentence Comprehension*

3.1 *Clausal Units*

Early theories of processing, guided by linguistic analyses that posited 'transformational' rules whose domains could be as large as a full clause, proposed that full sentence analysis and interpretation took place only at a clause boundary. These theories were supported by evidence that memory for verbatim information declines across a clause boundary, that clauses resists interruption by extraneous sounds, and that readers pause at ends of clauses and sentences to engage in 'wrap-up' processes (Fodor et al. 1974, Just and Carpenter 1980).

3.2 *Immediacy*

However, a great deal of evidence now exists indicating that readers do substantial interpretation on a word-by-word basis, without waiting for a clause boundary. The initial evidence for this claim came from Marslen-Wilson's (1973) work on 'fast shadowing' (repeating what one hears, with no more than a quarter-second delay). He showed that if the shadowed message contained a mispronunciation of a word's end (e.g., 'cigareppe'), the shadower would fairly often correct it spontaneously. Since this happened more often in

normal than in scrambled sentences, the listener must be using grammatical structure or meaning within a quarter-second or so to constrain recognition of words.

Evidence from measuring eye movements in reading led to a similar conclusion (see *Eye Movements in Reading*). For instance, if a person reads 'While Mary was mending the sock fell off her lap,' the reader's eyes are likely to fixate a long time on the word 'fell' or to regress from that word to an earlier point in the sentence. This happens because the verb 'fell' is grammatically inconsistent with the normally preferred structural analysis of '…mending the sock'; 'the sock' must be the subject of 'fell,' not the object of 'mending.' Similarly, a person who reads 'When the car stopped the moon came up' is likely to fixate a longer than normal time on 'moon.' Here, 'moon' is simply implausible as the direct object of 'stop,' not ungrammatical. This pattern of results indicates that readers (and presumably listeners) create and evaluate grammatical and meaning relations word by word, without waiting for a clause boundary.

It is tempting to hypothesize that readers and listeners always perform a full semantic analysis of sentences word by word, with never a delay. This 'immediacy hypothesis' may be too strong. Some evidence indicates that, while true lexical ambiguities (e.g., the difference between the bank of a river and a bank for money) may be resolved word by word, sense ambiguities (e.g., the difference between a newspaper as something to read and a newspaper as an institution) may not be (Frazier and Rayner 1990). Similarly, determining the antecedent of a pronoun may be a task that readers do not always complete at the first opportunity. Nonetheless, one secure conclusion is that readers and listeners generally understand a great deal of what a sentence conveys with little or no lag.

3.3 Garden-pathing

Another much-studied phenomenon is 'garden-pathing.' When readers and listeners construct word-by-word interpretations of sentences, they sometimes make mistakes. Bever (1970) initiated the study of garden-pathing with his infamous sentence 'The horse raced past the barn fell' (compare 'The car driven into the garage caught fire' if you doubt that the former sentence is actually possible in English). Readers are 'led down the garden path' by their preference to take 'horse' as the agent of 'race,' and is disrupted when the final verb 'fell' forces a revision so that 'raced past the barn' is taken as a relative clause that modifies 'horse,' and 'horse' is taken to be the theme of 'race' and the subject of 'fell.'

The mistakes readers and listeners make can be extremely diagnostic of the decision rules they follow. A simple empirical generalization (which may or may not describe the cognitive process a reader or listener is engaging in; see below) is that when a choice between two analyses has to be made, the reader/listener initially favors (a) the choice that is simpler in terms of grammatical structure, or (b) in the absence of complexity differences, the choice that permits new material to be related to the most recently processed old material. In the 'horse raced' sentence, the main verb analysis is simpler than the relative clause analysis. In 'While Mary was mending the sock fell off her lap,' a reader/listener prefers to relate 'the sock' to the recently processed verb 'mending' rather than to material that has not yet been received. In both cases, these preferences are disconfirmed by material that follows, resulting in a garden-path.

A great deal of experimental work using techniques of self-paced reading and eye movement measurement has indicated that reading is slowed (and regressive eye movements are encouraged) at points where garden-paths occur. Similarly, experimental research using ERPs has indicated the existence of distinct patterns of neural response to being garden-pathed. Research using these, and other, techniques has gone further and demonstrated the existence of subtle garden-paths that may not be apparent to conscious introspection (e.g., the temporary tendency to take 'the article…' as the direct object of 'understand' in 'I understood the article on word recognition was written by a real expert'). Normally, it appears that the rules readers and listeners use to decide on the grammatical structures of sentences function so smoothly that their operation cannot be observed. But just as the analysis of visual illusions allows insight into the normal processes of visual perception by identifying when they go astray, identifying when a rule for analyzing sentence structure gives the wrong answer can go far in determining what rule is actually being followed.

3.4 Lexical and Frequency Effects

Once the existence and basic nature of garden-paths were discovered, researchers realized that garden-paths did not always occur. The sentence presented earlier, 'The car driven into the garage caught fire' does not seem to lead to a garden-path. This may be due in part to the fact that the verb 'driven' is unambiguously a participle, not a main verb, and the normally preferred simpler analysis is grammatically blocked. However, other cases indicate that more subtle effects exist. Sentences with verbs that are obligatorily transitive are relatively easy (e.g., '*The dictato*r captured in the coup was hated' is easier than 'The dictator fought in the coup was hated'). Sentences with verbs like 'cook,' whose subject can take on the thematic role of theme/affected object, are particularly easy ('The soup cooked in the pot tasted good').

Sentences in which the subject is implausible as agent of the first verb but plausible as its theme are easier than sentences in which subject is a plausible agent (e.g., 'The evidence examined by the lawyer was unreliable' is easier than 'The defendant examined by the lawyer was unreliable,' although this difference may depend on a reader being able to see the start of the disambiguating 'by' phrase while reading the first verb). Consider a different grammatical structure, prepositional phrase attachment. It is easier to interpret a prepositional phrase as an argument of an action verb than as a modifier of a noun (e.g., 'The doctor examined the patient with a stethoscope' is easier than 'The doctor examined the patient with a broken wrist'), which is consistent with a grammatical analysis in which the former sentence is structurally simpler than the latter. However, this difference reverses when the verb is a verb of perception or a 'psychological' verb and the noun that may be modified is indefinite rather than definite (e.g., 'The salesman glanced at a customer with suspicion' is harder than 'The salesman glanced at a customer with ripped jeans') (Spivey-Knowlton and Sedivy 1995; see also Frazier and Clifton 1996, Mitchell 1994, Tanenhaus and Trueswell 1995, for further discussion and citation of experimental articles).

These results indicate that grammatical structure is not the only factor that affects sentence comprehension. Subtle details of lexical structure do as well. Further, the sheer frequency with which different structures occur and the frequency with which particular words occur in different structures seems to affect sentence comprehension: For whatever reason, more frequent constructions are easier to comprehend. For example, if a verb is used relatively frequently as a participle compared to its use as a simple past tense, the difficulty of the 'horse raced' type garden-path is reduced. Similarly, the normal preference to take a noun phrase following a verb as its direct object (which leads to difficulty in sentences like 'I understood the article on word recognition was written by an expert') is reduced when the verb is more often used with sentence complements than with direct objects. MacDonald et al. (1994) review these findings, spell out one theoretical interpretation which will be considered shortly, and discuss the need to develop efficient ways of counting the frequency of structures and to decide on the proper level of detail for counting structures.

3.5 *Effects of Context*

Sentences are normally not interpreted in isolation, which raises the question of how context can affect the processes by which they are understood. One line of research has emphasized the referential requirements of grammatical structures. One use of a modifier, such as a relative clause, is to select one referent from a set of possible referents. For instance, if there are two books on a table, you can ask for the 'book that has the red cover.' Some researchers (e.g., Altmann and Steedman 1988) have suggested that the difficulty of the relative clause in a 'horse raced' type garden-path arises simply because the sentence is presented out of context and there is no set of referents for the relative clause to select from. This suggestion entails that the garden-path would disappear in a context in which two horses are introduced, one of which was raced past a barn. Most experimental research fails to give strong support to this claim (cf. Mitchell 1994, Tanenhaus and Trueswell 1995). However, the claim may be correct for weaker garden-paths, for example the prepositional phrase attachment discussed above ('The doctor examined the patient with a broken wrist'), at least when the verb does not require a prepositional phrase argument. And a related claim may even be correct for relative clause modification when temporal relations, not simply referential context, affect the plausibility of main verb vs. relative clause interpretations (cf. Tanenhaus and Trueswell 1995).

3.6 *Effects of Prosody*

Most research on sentence comprehension has focused on reading. However, some experimental techniques permit the use of spoken sentences, which raises the possibility of examining whether the prosody of a sentence (its rhythm and melody) can affect how it is comprehended. The presence of a prosodic boundary (marked in English by a change in pitch at the end of the prosodic phrase, elongation of its final word, and possibly the presence of a pause) can affect the interpretation of a sentence. It can even eliminate the normal preferences that result in garden-paths. For example, Kjelgaard and Speer (1999) showed that the proper pattern of prosodic boundaries can eliminate the normal difficulty of sentences like 'Whenever Madonna sings a song is a hit.' As these authors emphasize, though, there is more to prosody than just putting in pauses. The entire prosodic pattern of an utterance (including the presence and location of 'pitch accents'; Schafer et al. 1996) can affect how it is interpreted.

4. *Psychological Models of Sentence Comprehension*

Psycholinguists have learned a great deal about the phenomena of sentence comprehension. This does not mean that they agree about the cognitive processes of readers and listeners that produce these phenomena.

Early theories of how people understand sentences claimed that the rules that make up a linguistic theory of a language are known (implicitly) to language users,

who use the rules directly in comprehending language (see Fodor et al. 1974). This did not work with the rules provided by early transformational grammars. Their rules operated on domains as large as a clause, and it soon became clear that people did not wait for the end of a clause to understand a sentence.

These early direct incorporation theories were replaced by what can be called 'detective-style' theories. Readers and listeners were presumed to search for clues of any kind to the relations among words in a sentence. Psycholinguistic theorizing became little more than listing cues that people used.

In the 1970s, as grammars changed to use more restrictive types of rules (especially emphasizing phrase-structure rules), a new breed of grammar-based theories developed. These theories added descriptions of decision strategies to the claim that people use the rules provided by the grammar, focusing on how they choose among various alternative rules in analyzing a sentence (Frazier 1987; cf. Frazier and Clifton 1996, Mitchell 1994, for further discussion of these theories). They claimed, for instance, that readers and listeners incorporate each new word into a sentence structure in the simplest and quickest possible way. This led to compelling accounts of garden-path phenomena, and the theories under discussion are often referred to as 'garden-path' theories.

In the 1970s and 1980s, linguistic theory moved away from positing broad, generally applicable rules to making claims about information contained in individual lexical items (constrained by some extremely broad general principles). This encouraged the development of 'lexicalist' theories of sentence comprehension (see MacDonald et al.1994, for the most completely developed statement of such a theory). These theories emphasize the specific contribution of individual words, rather than the effects of conformity to generally applicable word configurations. An individual word can provide a variety of kinds of information, including the possible and preferred thematic roles for its arguments, the frequency with which it is used in various constructions, and the entities it can refer to, as well as information about the phrase structure configurations it can appear in. Lexicalist theories propose that all these kinds of information are available and are used in deciding among alternative analyses.

It is clear that contemporary theories of sentence comprehension differ in the range of information that they claim guides sentence analysis. Garden-path theories are 'modular' (cf. Fodor 1983) in that they claim that only certain necessarily relevant types of information affect initial decisions about sentence structure. Lexicalist theories are generally non-modular in allowing many different kinds of information to affect these decisions. The theories also differ in what can be called depth-first vs. breadth-first processing (see Clifton 2000, for discussion). Garden-path theories are typically depth-first (although logically they need not be); they typically claim that a single analysis is built initially and then evaluated. Lexicalist theories are typically breadth-first; they claim that several different analyses are activated in parallel (generally being projected from the head of a phrase), and that all available types of information are used to choose among them. Experimental data have thus far not been able to settle the debate between theorists favoring depth-first models and theorists favoring breadth-first ones.

Garden-path theories explain garden-paths elegantly, and provide a more adequate account of how sentence structures are created than lexicalist theories do. On the other hand, they are forced to ascribe the effects of lexical structure, frequency, and context described above to a 'reanalysis' stage, which follows discovery that the initially preferred analysis is incorrect. Lexicalist theories, on the other hand, provide a natural account of effects of lexical structure, frequency, and context, especially when implemented as connectionist constraint-satisfaction models. But they have not yet been developed in a way that adequately explains where sentence structures come from (it is unappealing to say that the large parts of the grammar are stored with each lexical entry) or that explains why structurally complex sentences are overwhelmingly the ones that lead to garden-paths.

5. *Beyond Parsing*

This article has focused on how readers and listeners use their knowledge of language to compose sentence meanings from words. While much of the power of language comes from readers' and listeners' abilities to 'follow the rules' and arrive at grammatically licensed interpretations of sentences, it is also clear that people have the ability to go beyond such literal interpretation. The context in which an utterance is heard or a sentence is read can influence its interpretation far more profoundly than the context effects discussed above. The meaning of words can be altered depending on the context in which they occur (consider 'bank' in the context of a trip to the river to fish vs. a purchase of a house, or 'red' in the context of a fire truck vs. a red-haired woman). The reference of a definite noun phrase can depend the discourse context. A listener can determine whether a speaker who says 'Can you open this door?' is asking about the adequacy of a job of carpentry vs. requesting some assistance. Speakers can rely on listeners to get their drift when they use figurative language such as metaphors and possibly irony ('My copy editor is a butcher'). And very generally, what a listener or reader takes a speaker or writer to mean may depend on their shared goals and mutual knowledge (cf. Clark 1996).

Nonetheless, all these varied uses of language depend on the listener/reader's ability to put words

together into sentence meanings, which makes the study of the topics considered in this article an important branch of cognitive psychology.

See also: Eye Movements in Reading; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Syntactic Aspects of Language, Neural Basis of; Syntax; Syntax–Semantics Interface; Text Comprehension: Models in Psychology

Bibliography

Altmann G, Steedman M 1988 Interaction with context during human sentence processing. *Cognition* **30**: 191–238

Bever T G 1970 The cognitive basis for linguistic structures. In: Hayes J R (ed.) *Cognition and the Development of Language*. Wiley, New York, pp. 279–352

Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands

Chomsky N 1986 *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York

Clark H H 1996 *Using Language*. Cambridge University Press, Cambridge, UK

Clifton Jr C 2000 Evaluating models of sentence processing. In: Crocker M, Pickering M, Clifton Jr C (eds.) *Architecture and Mechanisms of Language Processing*. Cambridge University Press, Cambridge, UK, pp. 31–55

Clifton Jr C, Frazier L 1989 Comprehending sentences with long-distance dependencies. In: Carlson G N, Tanenhaus M K (eds.) *Linguistic Structure in Language Processing*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 273–318

Cutler A, Dahan D, Van Donselaar W 1997 Prosody in the comprehension of spoken language: A literature review. *Language and Speech* **40**: 141–201

Fodor J A 1983 *Modularity of Mind*. MIT Press, Cambridge, MA

Fodor J A, Bever T G, Garrett M F 1974 *The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar*. McGraw-Hill, New York

Fodor J D 1978 Parsing strategies and constraints on transformations. *Linguistic Inquiry* **9**: 427–74

Frazier L 1987 Sentence processing: A tutorial review. *Attention and Performance* **12**: 559–86

Frazier L 1995 Representational issues in psycholinguistics. In: Miller J L, Eimas P D (eds.) *Handbook of Perception and Cognition, Vol. 11: Speech, Language, and Communication*. Academic Press, San Diego

Frazier L, Clifton Jr C 1996 *Construal*. MIT Press, Cambridge, MA

Frazier L, Rayner K 1990 Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language* **29**: 181–200

Haberlandt K F 1994 Methods in reading research. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA, pp. 1–31

Just M A, Carpenter P A 1980 A theory of reading: From eye fixations to comprehension. *Psychological Review* **85**: 329–354

Just M A, Carpenter P A, Woolley J D 1982 Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General* **111**: 228–38

Kjelgaard M M, Speer S R 1999 Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language* **40**: 153–94

Kutas M, Van Petten C K 1994 Psycholinguistics electrified: Event-related brain potential investigations. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA, pp. 83–143

MacDonald M C, Pearlmutter N J, Seidenberg M S 1994 Lexical nature of syntactic ambiguity resolution. *Psychological Review* **101**: 676–703

Marslen-Wilson W D 1973 Linguistic structure and speech shadowing at very short latencies. *Nature* **244**: 522–3

McKoon G, Ratcliff R, Ward G 1994 Testing theories of language processing: An empirical investigation of the on-line lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**: 1219–28

Miller G A 1962 Some psychological studies of grammar. *American Psychologist* **17**: 748–62

Mitchell D C 1994 Sentence parsing. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, San Diego, CA, pp. 375–410

Rayner K, Sereno S, Morris R, Schmauder R, Clifton Jr C 1989 Eye movements and on-line language comprehension processes. *Language and Cognitive Processes* **4**: SI 21–50

Schafer A, Carter J, Clifton Jr C, Frazier L 1996 Focus in relative clause construal. *Language and Cognitive Processes* **11**: 135–63

Spivey-Knowlton M, Sedivy J C 1995 Resolving attachment ambiguities with multiple constraints. *Cognition* **55**: 227–67

Tanenhaus M K, Trueswell J C 1995 Sentence comprehension. In: Miller J, Eimas P (eds.) *Handbook of Perception and Cognition: Speech, Language, and Communication*. Academic Press, San Diego, CA, Vol. 11 pp. 217–62

C. Clifton, Jr.

Sequential Decision Making

Sequential decision making describes a situation where the decision maker (DM) makes successive observations of a process before a final decision is made, in contrast to dynamic decision making (see *Dynamic Decision Making*) which is more concerned with controlling a process over time.

Formally a sequential decision problem is defined, such that the DM can take observations $X_1, X_2, \ldots$ one at a time. After each observation X_n the DM can decide to terminate the process and make a final decision from a set of decisions D, or continue the process and take the next observation X_{n+1}. If the observations $X_1, X_2, \ldots$ form a random sample, the procedure is called sequential sampling.

In most sequential decision problems there is an implicit or explicit cost associated with each observation. The procedure to decide when to stop taking observations and when to continue is called the *stopping rule*. The objective in sequential decision making is to find a stopping rule that optimizes the

decision in terms of minimizing losses or maximizing gains including observation costs. The optimal stopping rule is also called the optimal strategy or the optimal policy.

A wide variety of sequential decision problems have been discussed in the statistics literature, including search problems, inventory problems, gambling problems, and secretary-type problems, including sampling with and without recall. Several methods have been proposed to solve the optimization problem under specified conditions, including dynamic programming, Markov chains, and Bayesian analysis.

In the psychological literature, sequential decision problems are better known as *optional stopping problems*. One line of research using sequential decision making is concerned with seeking information in situations such as buying houses, searching for a job candidate, price searching, or target search. The DM continues taking observations until a decision criterion for acceptance is reached. Another line of research applies sequential decision making to account for information processing in binary choice tasks (see *Diffusion and Random Walk Processes*; *Stochastic Dynamic Models (Choice, Response, and Time)*), and hypothesis testing such as in signal detection tasks (see *Signal Detection Theory: Multidimensional*). The DM continues taking observations until either of two decision criteria is reached. Depending on the particular research area, observations are also called offers, options, items, applicants, information, and the like. Observation costs include explicitly not only possibly money, but also time, effort, aggravation, discomfort, and so on.

Contrary to the objective of statisticians or economists, psychologists are less interested in determining the optimal stopping rule, and more interested in discussing the variables that affect human decision behavior in sequential decision tasks. Optimal decision strategies are considered as normative models, and their predictions are compared to actual choice behavior.

1. Sequential Decision Making with One Decision Criterion

In sequential decision making with one decision criterion the DM takes costly observations X_n, $n = 1,\ldots$ of a random process one at a time. After observing $X_n = x_n$ the DM has to decide whether to continue sampling observations or to stop. In the former case, the observation X_{n+1} is taken at a cost of c_{n+1}; in the latter case the DM receives a net payoff that consists of the payoff minus the observation costs. The DM's objective is to find a stopping rule that maximizes the expected net payoff.

The optimal stopping rule depends on the specific assumptions made about the situation: (a) the distribution of X is known, not known or partly known, (b) X_i are distributed identically for all i, or have the similar distribution but with different parameters, or have different distributions, (c) the number of possible observations, n, is bounded or unbounded, (d) the sampling procedure, e.g., it is possible to take the highest value observed so far when stopping (sampling with recall) or only to take the last value when stopping (sampling without recall), and (e) the cost function, c_n, is fixed for each observation or is a function of n. Many of these problems have been studied theoretically by mathematicians and experimentally by psychologists. Pioneering experimental work was done in a series of papers by Rapoport and colleagues (1966, 1969, 1970, 1972).

1.1 Unknown Sample Distribution: Secretary-type Problems

Kahan et al. (1967) investigated decision behavior in a sequential search task where the DM had to find the largest of a set of $n = 200$ numbers, observed one at a time from a deck of cards. The observations were taken in random order without replacement. The DM could only declare the current observation as the largest number (sampling without recall), and could compare the number with the previous presented numbers. No explicit cost for each observation was taken, i.e., $c = 0$. The sample distribution was unknown to the DM. A reward was paid only when the card with the highest number was selected, and nothing otherwise. This describes a decision situation that is known as *the secretary problem* (a job candidate search problem; for various other names, see, e.g., Freeman 1983) which, in its simplest form, makes explicit the following assumptions (Ferguson 1989): (a) only one position is available, (b) the number n of applicants is known, (c) applicants are interviewed sequentially in random order, each order being equally likely, and (d) all applicants can be ranked without ties—the decision to reject or accept an applicant must be based only on the relative ranks of the applicants interviewed so far, (e) an applicant once rejected cannot later be recalled, and (f) the payoff is 1 when choosing the best of the n applicants, 0 otherwise.

The optimal strategy for this kind of problem is to reject the first $s-1$, $s \geqslant 1$, items (cards, applicants, draws) and then choose the first item that is best in the relative ranking so far. With

$$a_s \equiv \frac{1}{s} + \frac{1}{s+1} + \cdots + \frac{1}{n-1} \tag{1}$$

the optimal strategy is to stop if $a_s < 1$ and to continue if $a_s > 1$, which can easily be determined for small n. For large n, the probability of choosing the best item is

approximated by $1/e$ and the optimal s by n/e. ($e = 2.71\ldots$). (For derivations, see, e.g., DeGroot 1970, Freeman 1983, Gilbert and Mosteller 1966).

Kahan et al. (1967) reported that about 40 percent of their subjects did not follow the optimal strategy but stopped too late and rejected a card that should have been accepted. The failure of the strategy for describing behavior was assigned to its inadequacy for the described task. Although at the beginning of the experiment the participants did not know anything about the distribution (requirement), they could learn about the distribution by taking observations (partly information). To guarantee ignorance of the distribution, Gilbert and Mosteller (1966) recommended supplying only the rank of the observation made so far and not the actual value. Seale and Rapoport (1997) conducted an experiment following this advice. They found that participants (with $n = 40$ and $n = 80$) stopped earlier than prescribed by the optimal stopping rule. They proposed simple decision rules or heuristics to describe the actual choice behavior. Using a cutoff rule, the DM rejects the first $s-1$ applicants and then chooses the next top-ranked applicant, i.e., the candidate. The DM simply counts the number of applicants and then stops on the first candidate after observing $s-1$ applicants. Under a candidate count rule, the DM counts the number of candidates and chooses the jth candidate. A successive non-candidate rule requires the DM to choose the first candidate after observing at least k consecutive noncandidates following the last candidate.

The secretary problem has been extended and generalized in many different directions within the mathematical statistics field. Each of the above assumptions has been relaxed in one way or another. (Ferguson 1989, Freeman 1983). However, the label of secretary problem tends to be used only when the distribution is unknown and the decision to stop or to continue depends only on the relative ranking of the observations taken so far and not on their actual values.

1.2 Known Sample Distribution

Rapoport and Tversky (1966, 1970) investigated choice behavior when the mean and the variance of the distribution was known to the DM. The cost for each observation was fixed but the amount varied across experimental conditions, and the number of possible observations n was unbounded (1966) or bounded and known (1970). Behavior for sampling with and without recall was compared. When sampling is without recall only the value of the last observation, $X_n = x_n$, can be received, and the payoff is this value minus the total sampling cost, i.e, $x_n - cn$. The optimal strategy is to find a stopping rule that maximizes the expected payoff $E(X_N - cN)$. When sampling is with recall, the highest value observed so far can be selected and the payoff is $\max(x_1,\ldots, x_n) - cn$ and the optimal strategy is to find a stopping rule that maximizes $E(\max(X_1,\ldots, X_N) - cN)$. In the following, v with subscripts and v^* denote the expected gain from an (optimal) procedure.

1.2.1 Number of Observations Unbounded. If n is unbounded, i.e., if the number of observations that can be taken is unlimited, and X_1, $X_2\ldots$ are sampled from a known distribution function $F(x)$, the optimal strategy is the same for both sampling with and without recall. In particular, the optimal strategy is to continue to take observations whenever the observed value $x_j < v^*$, and to stop taking observations as soon as an observed value $x_j \geqslant v^*$, where v^* is the unique solution of

$$\int_{v^*}^{\infty} (x - v^*)\, dF(x) = c \quad -\infty < v^* < \infty. \tag{2}$$

When the observations are taken from a standard normal distribution with density functions $\phi(x)$ and distribution function $\Phi(\mathrm{x})$, we have that

$$v^* = \frac{\phi(v^*) - c}{1 - \Phi(v^*)}. \tag{3}$$

Although sampling with and without recall have the same solution, they seem to be different from a psychological point of view. Rapoport and Tversky (1966) found that the group sampling without recall took significantly fewer observations than the participants sampling with recall. The mean number of observations for both groups decreased with increasing cost c, and the difference with respect to the number of observations taken was diminished. However, the participants in both groups took fewer observations than prescribed by the optimal strategy. This nonoptimal behavior of the participants was attributed to a lack of thorough knowledge of the distributions.

1.2.2 Number of Observations Bounded. If n, $n \geqslant 2$, is bounded, i.e., if not more than n observations can be taken, the optimal stopping rules for sampling with and without recall are different. For sampling without recall, an optimal procedure is to continue taking observations whenever $x_j < v_{n-j} - c$ and to stop as soon as $x_j \geqslant v_{n-j} - c$, where $j = 1, 2\ldots n$ indicates the number of observations which remain available and

$$v_{j+1} = (v_j - c) - \int_{v_j - c}^{\infty} (x - (v_j - c))\, dF(x). \tag{4}$$

With $v_1 = E(X) - c$, the sequence can be computed successively. Again, assuming a standard normal distribution $v_{j+1} = \phi(v_j - c) + \Phi(v_j - c)$.

For sampling with recall, the optimal strategy is to continue the process whenever a value $x_j < v^*$ and to stop taking observations as soon as an observed value $x_j \geqslant v^*$, where v^* is as in Eqn. (2), which is the same solution as for n unbounded. (For derivations of the strategies, see DeGroot 1970, Sakaguchi 1961.)

Rapoport and Tversky (1970) investigated choice behavior within this scenario. Sampling was done both with and without recall. The number of observations that could be taken as well as observation cost varied across experimental groups. One third of the participants did not follow the optimal strategy. Under both sampling procedures and all cost conditions, they took on average fewer observations than predicted by the corresponding optimal stopping rules. There were no systematic differences due to cost, as observed in their previous study. They concluded that 'the optimal model provides a reasonable good account of the behavior of the subjects' (p. 119).

1.3 Different Sample Distributions for Each Observation

Most research concerned with sequential decision making assumes that the observations are sampled from the same distribution, i.e., X_i are distributed identically for all i. For many decision situations, however, the observations may be sampled from the same distribution family with different parameters, or from different distributions. Especially in economic areas, such as price search, it is reasonable to assume that the distributions from which observations are taken change over time. The sequence of those samples has been called *nonstationary series*. Of particular interest are two special nonstationary series: *ascending* and *descending* series. For ascending series, the observations are drawn from distributions, usually from normal distributions, with increasing mean as i increases; for descending series the mean of the distribution decreases as i increases, i indicating the sample index. For both cases, experiments have been conducted to investigate choice behavior in a changing environment. Shapira and Venezia (1981) compared choice behavior for ascending, descending and constant (identically distributed X_i) series. In one experiment (numbers from a deck of cards), the distributions were known to the DM; no explicit observation costs were imposed; sampling occurred without recall; and the number of observations that could be taken was limited to $n = 7$. The variance of the distributions varied across experimental groups. An optimal procedure was assumed to continue taking observations whenever $x_j < v_{n-j}$, and to stop as soon as $x_j \geqslant v_{n-j}$, where $j = 1, 2 \ldots n$ indicates the number of observations which remain available. $k = 1, \ldots, n$ indicates the specific distribution for the jth observation. Thus

$$v_{j+1} = v_j - \int_{v_j}^{\infty} (x - v_j)\, dF_k(x). \tag{5}$$

With $v_1 = E(X_1)$ the sequence can be computed successively. Assuming a standard normal distribution $v_{j+1} = \phi_k(v_j) + \Phi_k(v_j)$.

Across all conditions, 58 percent of the participants behaved in an optimal way. The proportion of optimal stopping did not depend on the type of series but on the size of the variance. Nonoptimal stopping (24 percent stopped too early; 18 percent too late) depended on the series and on the size of the variance. In particular, participants stopped too early on ascending and too late on descending series. A similar result was observed by Brickman (1972). In this study, departing from the optimal stopping rule was attributed to an inadequacy of the stopping rule taken for the particular experimental conditions (assuming complete knowledge of the distributions). In a secretary problem design (see Sect. 1.1), Corbin et al. (1975) were less concerned with optimal choice behavior than with the processes by which the participants made their selections, and with factors that influenced those processes. The emphasis of the investigation was on decision making heuristics rather than on the adequacy of optimal models. With the same optimal stopping rule for all experimental conditions, they found that stopping behavior depended on contextual variables such as the ascending or descending trend of the inspected numbers of the stack.

2. Search Problems—Multiple Information Sources

In a sequential decision making task with multiple information sources, the DM has the option to take information sequentially from different sources. Each information source may provide valid information with a particular probability and at different cost. The task is not only to decide to stop or to continue the process but also, if continuing, which source of information to consult.

Early experimental studies were done by Kanarick et al. (1969), Rapoport (1969), Rapoport et al. (1972). A typical task is to find an object (e.g., a black ball) which is hidden in one of several possible locations (e.g., in one of several bins containing white balls). The optimal search strategy depends on further task specifications, such as whether the object can move from one location to another, how many objects are to be found, and whether the search process may stop before the object has been found. Rapoport (1969)

investigated the case when a single object that could not move was to be found in one of r, $r \geqslant 2$, possible locations. The DM was not allowed to stop the process before the target was found. All of the following were known to the DM: the a priori probability p_i, $p_i > 0$ that the object is in location i, $i = 1, 2, \ldots, r$, with $\sum_{i=1}^{r} p_i = 1$; a miss probability α_i, $0 < \alpha_i < 1$, that even if the object is in location i it will not be found in a particular search of that location ($1-\alpha_i$ is referred to the respective detection probability); and a cost, c_i, for a single observation at location i. The objective of the DM is to find a search strategy that minimizes the expected cost. For $i = 1, \ldots, r$ and $j = 1, 2, \ldots$ let Π_{ij} denote the probability that the object is found at location i during the jth search and the search is terminated. Then

$$\Pi_{ij} = p_i \alpha_i^{j-1} (1-\alpha_i), \quad i = 1, \ldots, r, \; j = 1, 2, \ldots \tag{6}$$

If all values of Π_{ij}/c_i for all values of i and j are arranged in order of decreasing magnitude, the optimal strategy is to search according to this ordering (for derivations, see DeGroot 1970). Ties may be ordered arbitrarily among themselves. The optimal strategy is determined by the detection probabilities and observation costs, and optimal search behavior implies a balance between maximizing the detection probability and minimizing the observation cost. Rapoport (1969) found that participants did not behave optimally. They were more concerned with maximizing the probability of detecting the target than with minimizing observation cost. Increasing the difference of observation cost c_i among the $i = 1, 2, 3, 4$ locations showed that the deviation from the optimal strategy even increased. Rapoport et al. (1972) varied the search problems by allowing the DM to terminate the search at any time; adding a terminal reward, R, for finding the target; and a terminal penalty, B, for not finding the target. Most participants showed a bias toward maximizing detection probability vs. minimizing search cost per observation, similar to the previous study.

3. *Sequential Decision Making with Two or More Possible Decisions*

A random sample $X_1, X_2, \ldots$ is generated by an unknown state of nature, Θ. The DM can take observations one at a time. After observing $X_n = x_n$ the DM makes inferences about Θ based on the values of $X_i, \ldots, X_n$ and can decide whether to continue sampling observations or to stop the process. In the former case, observation X_{n+1} is taken; in the latter, the DM makes a final decision $d \in D$. The consequences to the DM depend on the decision d and the value θ.

The statistical theory for this situation was developed by Wald during the 1940s. It has been used to test hypotheses and estimate parameters. In psychological research, sequential decision making of this kind is usually limited to two decisions $D = \{d_1, d_2\}$, and applied to binary choice tasks (see *Diffusion and Random Walk Processes*; *Stochastic Dynamic Models (Choice, Response, and Time)*; *Bayesian Theory: History of Applications*).

The standard theory of sequential analysis by Wald (1947) does not include considerations of observation costs $C(n)$, losses for terminal decisions $L(\theta, d)$, and *a priori* (subjective) probabilities π of the alternative states of nature. *Deferred decision theory* generalizes the original theory by including these variables explicitly. The objective of the DM is to find a stopping rule that minimizes expected loss (called risk) and expected observation cost. The form of that optimal stopping rule depends mainly on the assumptions about the number of observations that can be taken (bounded or unbounded), and on the assumption of cost per observation (fixed or not) (see DeGroot 1970). Birdsall and Roberts (1965), Edwards (1965), and Rapoport and Burkheimer (1971) introduced the idea of deferred decision theory as normative models of choice behavior to the psychological community. Experiments investigating human behavior in deferred decision tasks have been carried out by Pitz and colleagues (e.g., Pitz et al. 1969), and by Busemeyer and Rapoport (1988). Rapoport and Wallsten (1972) summarize experimental findings.

For illustration, assume the decision problem in its simplest form. Suppose two possible states of nature θ_1 or θ_2, and two possible decisions d_1 and d_2. Cost c per observation is fixed and the number of observations is unbounded. The DM does not know which of the states of nature, θ_1 or θ_2 is generating the observation, but there are a priori probabilities π that it is θ_1 and $(1-\pi)$ that it is θ_2. Let w_i denote the loss for a terminal decision incurred by the DM in deciding that θ_i is not the correct state of nature when it actually is ($i = 1, 2$). No losses are assumed when the DM makes a correct decision. Let π_n denote the posterior probability that θ_1 is the correct state of nature generating the observations after n observations have been made. The total posterior expected loss is $r_n = \min\{w_1\pi_n, w_2(1-\pi_n)\} + nc$. The DM's objective is to minimize the expected loss. An optimal stopping rule is specified in terms of decision boundaries, α and β. If the posterior probability is greater than or equal to α, then decision d_1 is made; if the posterior probability is smaller than or equal to β, then d_2 is selected; otherwise sampling continues.

See also: Decision Making (Naturalistic), Psychology of; Decision Making, Psychology of; Decision Research: Behavioral; Dynamic Decision Making; Multi-attribute Decision Making in Urban Studies

Bibliography

Birdsall T G, Roberts R A 1965 Theory of signal detectability: Deferred decision theory. *The Journal of Acoustical Society of America* **37**: 1064–74

Brickman P 1972 Optional stopping on ascending and descending series. *Organizational Behavior and Human Performance* **7**: 53–62

Busemeyer J R, Rapoport A 1988 Psychological models of deferred decision making. *Journal of Mathematical Psychology* **32**(2): 91–133

Corbin R M, Olson C L, Abbondanza M 1975 Context effects in optional stopping decisions. *Organizational Behavior and Human Performance* **14**: 207–16

De Groot M H 1970 *Optimal Statistical Decisions*. McGraw-Hill, New York

Edwards W 1965 Optimal strategies for seeking information: Models for statistics, choice response times, and human information processing. *Journal of Mathematical Psychology* **2**: 312–29

Ferguson T S 1989 Who solved the secretary problem? *Statistical Science* **4**(3): 282–96

Freeman P R 1983 The secretary problem and its extensions: A review. *International Statistical Review* **51**: 189–206

Gilbert J P, Mosteller F 1966 Recognizing the maximum of a sequence. *Journal of the American Statistical Association* **61**: 35–73

Kahan J P, Rapoport A, Jones L E 1967 Decision making in a sequential search task. *Perception & Psychophysics* **2**(8): 374–6

Kanarick A F, Huntington J M, Peterson R C 1969 Multisource information acquisition with optional stopping. *Human Factors* **11**: 379–85

Pitz G F, Reinhold H, Geller E S 1969 Strategies of information seeking in deferred decision making. *Organizational Behavior and Human Performance* **4**: 1–19

Rapoport A 1969 Effects of observation cost on sequential search behavior. *Perception & Psychophysics* **6**(4): 234–40

Rapoport A, Tversky A 1966 Cost and accessibility of offers as determinants of optional stopping. *Psychonomic Science* **4**: 45–6

Rapoport A, Tversky A 1970 Choice behavior in an optional stopping task. *Organizational Behavior and Human Performance* **5**: 105–20

Rapoport A, Burkheimer G J 1971 Models of deferred decision making. *Journal of Mathematical Psychology* **8**: 508–38

Rapoport A, Lissitz R W, McAllister H A 1972 Search behavior with and without optional stopping. *Organizational Behavior and Human Performance* **7**: 1–17

Rapoport A, Wallsten T S 1972 Individual decision behavior. *Annual Review of Psychology* **23**: 131–76

Sakaguchi M 1961 Dynamic programming of some sequential sampling design. *Journal of Mathematical Analysis and Applications* **2**: 446–66

Seale D A, Rapoport A 1997 Sequential decision making with relative ranks: An experimental investigation of the 'secretary problem.' *Organizational Behavior and Human Decision Processes* **69**(3): 221–36

Shapira Z, Venezia I 1981 Optional stopping on nonstationary series. *Organizational Behavior and Human Performance* **27**: 32–49

Wald A 1947 *Sequential Analysis*. Wiley, New York

A. Diederich

Sequential Statistical Methods

Statistics plays two fundamental roles in empirical research. One is in determining the data collection process: the experimental design. The other is in analyzing the data once it has been collected. For the purposes of this article, two types of experimental designs are distinguished: *sequential* and *nonsequential*. In a sequential design the data that accrue in an experiment can affect the future course of the experiment. For example, an observation made on one experimental unit treated in a particular way may determine the treatment used for the next experimental unit. The term 'adaptive' is commonly used as an alternative to sequential. In a nonsequential design the investigator can carry out the entire experiment without knowing any of the interim results.

The distinction between sequential and nonsequential is murky. An investigator's ability to carry out an experiment exactly as planned is uncertain, as information that becomes available from within and outside the experiment may lead the investigator to amend the design. In addition, a nonsequential experiment may give results that encourage the investigator to run a second experiment, one that might even simply be a continuation of the first. Considered separately, both experiments are nonsequential, but the larger experiment that consists of the two separate experiments is sequential.

In a typical type of nonsequential design, 20 patients suffering from depression are administered a drug and their improvements are assessed. An example sequential variation is the following. Patients' improvements are recorded 'in sequence' during the experiment. The experiment stops should it happen that at least nine of the first 10 patients, or no more than one of the first 10 patients improve(s). On the other hand, if between two and eight of the first 10 patients improve then sampling continues to a second set of 10 patients, making the total sample size equal to 20 in that case. Another type of sequential variation is when the dose of the drug is increased for the second 10 patients should it happen that fewer than four of the first 10 improve.

Much more complicated sequential designs are possible. For example, the first patient may be assigned a dose in the middle of a range of possible doses. If the patient improves then the next patient is assigned the next lower dose, and if the first patient does not improve then the next patient is assigned the next higher dose. This process continues, always dropping the dosage if the immediately preceding patient improved, and increasing the dosage if the immediately preceding patient did not improve. This is called an 'up-and-down' design.

Procedures in which batches of experimental units (such as groups of 10 patients each) are analyzed before proceeding to the next stage of the experiment

are called 'group-sequential.' Designs such as the up-and-down design, in which the course of the experiment can change after each experimental unit responds are called 'fully sequential.' So a fully sequential design is a group-sequential design in which the group size is one. Designs in which the decision of when to stop the experiment depends on the accumulating results are called 'sequential stopping.' Using rules to determine which treatments to assign to the next experimental unit or batch of units is called 'sequential allocation.'

Designs of most scientific experiments are sequential, although perhaps not formally so. Investigators usually want to conserve time and resources. In particular, they do not want to continue an experiment if they have already learned what they set out to learn, and this is so whether their conclusion is positive or negative, or if finding a conclusive answer would be prohibitively expensive. (Where the investigator discovers that the standard deviation of the observations is much larger than originally thought is an example of an experimental design that would be prohibitively expensive to continue because the required sample size would be large.)

Sequential designs are difficult or impossible to use in some investigations. For example, results might take a long time to obtain, and waiting for them would mean delaying other aspects of the experiment. Suppose one is interested in whether grade-schoolers diagnosed with attention deficit hyperactivity disorder (ADHD) should be prescribed Ritalin. The outcome of interest is whether children on Ritalin will be addicted to drugs as adults. Consider assigning groups of 10 children to Ritalin and 10 to a placebo, and waiting to observe their outcomes before deciding whether to assign an additional group of 10 patients to each treatment. The delay in observation means that it would probably take hundreds of years to get an answer to the overall question. The long-term nature of the endpoint means that any reasonable experiment addressing this question would necessarily be nonsequential, with large numbers of children assigned to the two groups before any information at all would become available about the endpoint.

1. *Analyzing Data from Sequential Experiments—Frequentist Case*

Consider an experiment of a particular type, say one to assess extrasensory perception (ESP) ability. A subject claiming to have ESP is asked to choose between two colors. The *null hypothesis* of no ability is that the subject is only guessing, in which case the correct color has a probability of 1/2. Suppose the subject gets 13 correct out of 17 tries. How should these results be analyzed and reported? The answer depends on one's statistical philosophy. Frequentists and Bayesians take different approaches. Frequentist analyses depend on whether the experiment's design is sequential, and if it is sequential the conclusions will differ depending on the actual design used.

In the nonsequential case the subject is given exactly 17 tries. The frequentist P-value is the probability of results as extreme as or more extreme than those observed. The results are said to be 'statistically significant' if the P-value is less than 5 percent. A convention is to include both 13 or more successes and 13 or more failures. (This 'two-sided' case allows for the possibility that the subject has ESP but has inverted the 'extrasensory' signals.) Assuming the null hypothesis and that the tries are independent, the probabilities of the number of successes is binomial. Binomial probabilities can be approximated using the normal distribution. The z-score for 13 out of 17 is about 2, and so the probability of 13 or more successes or 13 or more failures is about 0.05 (the exact binomial probability is 0.049), and so the results are statistically significant at the 5 percent level and the null hypothesis is rejected.

Now suppose the experiment is sequential. The frequentist significance level is now different, and it depends on the actual design used. Suppose the design is to sample until the subject gets at least four successes and at least four failures—same data, different design. Again, more extreme means more than 13 successes (and exactly four failures) or more than 13 failures (and exactly four successes). The total probability of these extreme values is 0.021—less than 0.049—and so the results are now more highly significant than if the experiment's design had been nonsequential.

Consider another sequential design, one of a type of group-sequential designs commonly used in clinical trials. The experimental plan is to stop at 17 tries if 13 or more are successes or 13 or more are failures, and hence the experiment is stopped on target. But if after 17 tries the number of successes is between five and 12 then the experiment continues to a total of 44 tries. If at that time, 29 or more are successes or 29 or more are failures then the null hypothesis is rejected. To set the context, suppose the experiment is nonsequential, with sample size fixed at 44 and no possibility of stopping at 17; then the exact significance level is again 0.049. When using a sequential design, one must consider all possible ways of rejecting the null hypothesis in calculating a significance level. In the group-sequential design there are more ways to reject than in the nonsequential design with the sample size fixed at 17 (or fixed at 44). The overall probability of rejecting is greater than 0.049 but is somewhat less than $0.049 + 0.049$ because some sample paths that reject the null hypothesis at sample size 17 also reject it at sample size 44. The total probability of rejecting the null hypothesis for this design is actually 0.080. Therefore, even though the results beyond the first 17 observations are never observed, the fact that they might have been observed makes 13 successes of 17 no

Table 1
Summary of experimental designs and related significance levels

Stopping rule	Significance level
After 17 observations (nonsequential)	0.049
After at least 4 successes and 4 failures	0.021
After 17 or 44 observations, depending on interim results	0.08
Stop when 'you think you know the answer'	Undefined

longer statistically significant (since 0.08 is greater than 0.05). The three designs above are summarized in Table 1. The table includes a fourth design in which the significance level cannot be found.

To preserve a 0.05 significance level in group-sequential or fully sequential designs, investigators must adopt more stringent requirements for stopping and rejecting the null hypothesis; that is, they must include fewer observations in the region where the null hypothesis is rejected. For example, the investigator in the above study might drop 13 successes or failures in 17 tries and 29 successes or failures in 44 tries from the rejection region. The investigator would stop and claim significance only if there are at least 14 successes or at least 14 failures in the first 17 tries, and claim significance after 44 tries only if there are at least 30 successes or at least 30 failures. The *nominal significance levels* (those appropriate had the experiment been nonsequential) at $n = 17$ and $n = 44$ are 0.013 and 0.027, and the overall (or adjusted) significance level of rejecting the null hypothesis is 0.032. (No symmetric rejection regions containing more observations allow the significance level to be greater than this but still smaller than 0.05.) With this design, 13 successes out of 17 is not statistically significant (as indicated above) because this data point is not in the rejection region. The above discussion is in the context of significance testing. But the same issues apply in all types of frequentist inferences, including confidence intervals.

The implications of the need to modify rejection regions depending on the design of an experiment are profound. In view of the penalties that an investigator pays in significance level that are due to repeated analyses of accumulating data, investigators strive to minimize the number of such analyses. They shy away from using sequential designs and so may miss opportunities to stop or otherwise modify the experiment depending on accumulating results.

What happens if investigators fail to reveal that other analyses did occur, or that the experiment might have continued had other results been observed? Any frequentist conclusion that fails to take the other analyses into account is meaningless. Strictly speaking, this is a breach of scientific ethics when carrying out frequentist analyses. But it is difficult to find fault with investigators who do not understand the subtleties of frequentist reasoning and who fail to make necessary adjustments to their inferences.

For more information about the frequentist approach to sequential experimentation, see Whitehead (1992).

2. *Analyzing Data from Sequential Experiments—Bayesian Case*

When taking a Bayesian approach (see *Bayesian Statistics*) (or a likelihood approach), conclusions are based only on the observed experimental results and do not depend on the experiment's design. So the murky distinction that exists between sequential and nonsequential designs is irrelevant in a Bayesian approach. In the example considered above, 13 successes out of 17 tries will give rise to the same inference in each of the designs considered. Bayesian conclusions depend only on the data actually observed and not otherwise on the experimental design (Berger and Wolpert 1984, Berry 1987).

The Bayesian paradigm is inherently sequential. Bayes's theorem prescribes the way learning takes place under uncertainty. It specifies how an observation modifies one's state of knowledge (Berry 1996). Moreover, each observation that is planned has a probability distribution. After 13 successes in 17 tries, the probability of success on the next try can be found. This requires a distribution, called a 'prior distribution,' for the probability of success on the first of the 17 tries. Suppose the prior distribution is uniform from zero to one. (This is symmetric about the null hypothesis of 1/2, but it is unlikely to be anyone's actual prior distribution in the case of ESP because it gives essentially all the probability to *some* ESP ability.) The predictive probability of a success on the 18th try is then $(13+1)/(17+2) = 0.737$, called 'Laplace's rule of succession' (Berry 1996, p. 204). Whether to take this 18th observation can be evaluated by weighing the additional knowledge gained (having 14 successes out of 18, with probability 0.737, or 13 successes out of 18, with probability 0.263) with the costs associated with the observation.

Predictive probabilities are fundamental in a Bayesian approach to sequential experimentation. They indicate how likely it is that the various possibilities for future data will happen, given the data currently available. Suppose that after 13 successes of 17 tries one is entertaining taking an additional 27 observations. One may be interested in getting at least 30 successes out of the total of 44 observations—which means at least 17 of the additional 27 observations are successes. The predictive probability of this is about 50 percent. Or one may be interested in getting successes

in at least 1/2 (22) of the 44 tries. The corresponding predictive probability is 99.5 percent.

The ability to use Bayes's theorem for updating one's state of knowledge and the use of predictive probabilities makes the Bayesian approach appealing to researchers in the sequential design of experiments. As a consequence, many researchers who prefer a frequentist perspective use the Bayesian approach in the context of sequential experimentation. If they are interested in finding the frequentist operating characteristics (such as significance level and power), these can be calculated by simulation.

The next section (Sect. 4) considers a special type of sequential experiment. The goals of the section are to describe some of the calculational issues that arise in solving sequential problems and to convey some of the interesting aspects of sequential problems. It takes a Bayesian perspective.

3. *Sequential Allocation of Experiments: Bandit Problems*

In many types of experiments, including many clinical trials, experimental units are randomized in a balanced fashion to the candidate treatments. The advantage of a balanced design is that it gives maximal information about the differences between treatments. In some types of experiment, including some clinical trials, it may be important to obtain good results on the units that are part of the experiment. Treatments—or *arms*—are assigned based on accumulating results; that is, assignment is sequential. The goal is to maximize the overall effectiveness—of those units in the experiment, but also, perhaps, including those units not actually in the experiment but whose treatment might benefit from information gained in the experiment.

Specifying a design is difficult. The first matter to be considered is the arm selected for the initial unit. Suppose that the first observation is X_1. The second component of the design is the arm selected next, given X_1 and also given the first arm selected. The third component depends on X_1 and the second observation X_2, and on the corresponding arms selected. And so on. A design is *optimal* if it maximizes the expected number of successes. An arm is optimal if it is the first selection of an optimal design.

Consider temporarily an experiment with n units and two available arms. Outcomes are dichotomous: arm 1 has success probability p_1 and arm 2 has success probability p_2. The goal is to maximize the expected number of successes among the n units. Arm 1 is standard and has known success proportion p_1. Arm 2 has unknown efficacy. Uncertainty about p_2 is given in terms of a prior probability distribution. To be specific, suppose that this is uniform on the interval from 0 to 1.

Table 2
Possible designs and associated expected number of successes

Design	Expected number of successes
{1; 1, 1}	$\mathbf{2p_1}$
{1; 1, 2}	$p_1+p_1^2+(1-p_1)/2$
{1; 2, 1}	$p_1+p_1/2+(1-p_1)p_1$
{1; 2, 2}	$p_1+1/2$
{2; 1, 1}	$1/2+p_1$
{2; 1, 2}	$1/2+(1/2)p_1+(1/2)(1/3)$
{2; 2, 1}	$\mathbf{1/2+(1/2)(2/3)+(1/2)p_1}$
{2; 2, 2}	$\mathbf{2(1/2)=1}$

If $n=1$ then the design requires only an initial selection, arm 1 or arm 2. Choosing arm 1 has expected number of successes p_1. Choosing arm 2 has conditional expected number of successes p_2, and an unconditional expected number of successes, the prior probability of success, which is 1/2. Therefore arm 1 is optimal if $p_1 \geqslant 1/2$ and arm 2 is optimal if $p_1 \leqslant 1/2$. (Both arms—and any randomization between them—are optimal when $p_1 = 1/2$.)

The problem is more complicated for $n \geqslant 2$. Consider $n=2$. There are two initial choices and two choices depending on the result of the first observation. There are eight possible designs. One can write a design as $\{a; a_S, a_F\}$, where a is the initial selection, a_S is the next selection should the first observation be a success, and a_F is the next selection should the first observation be a failure. To find the expected number of successes for a particular design, one needs to know such quantities as the probability of a success on arm 2 after a success on arm 2 (which is 2/3) and the probability of a success on arm 2 after a failure on arm 2 (which is 1/3). The possible designs and their associated expected numbers of successes are given in Table 2.

It is easy to check that only three of these expected numbers of successes (shown in bold) are candidates for the maximum. If $p_1 \geqslant 5/9$ then {1; 1, 1} is optimal; if $1/3 \leqslant p_1 \leqslant 5/9$ then {2; 2, 1} is optimal; and if $p_1 \leqslant 1/3$ then {2; 2, 2} is optimal. For example, if $p_1 = 1/2$ then it is optimal to use the unknown arm 2 initially. If the outcome is a success, then a decision is made to 'stay with a winner' and use arm 2 again. If a failure occurs, then the decision is made to switch to the known arm 1.

Enumeration of designs is tedious for large n. Most designs can be dropped from consideration based on theoretical results (Berry and Fristedt 1985). For example, there is a breakeven value of p_1, say p_1^*, such that arm 1 is optimal for $p_1 \geqslant p_1^*$. Also, one need consider only those designs that continue to use arm 1 once it has been selected. But many designs remain. Backward induction can be used to find an optimal design (Berry and Fristedt 1985).

Table 3
The optimal expected proportion of successes for selected values of n and for fixed $p_1 = 1/2$

n	1	2	5	10	20	50	100	200	500	1,000	10,000
Proportion of successes	0.500	0.542	0.570	0.582	0.596	0.607	0.613	0.617	0.621	0.622	0.624

Table 4
The breakeven values of p_1* for selected values of n

n	1	2	5	10	20	50	100	200	500	1000	10,000
p_1*	0.500	0.556	0.636	0.698	0.758	0.826	0.869	0.902	0.935	0.954	0.985

Table 3 gives the optimal expected proportion of successes for selected values of n and for fixed $p_1 = 1/2$. Asymptotically, for large n, the maximal expected proportion of successes is 5/8, which is the expected value of the maximum of p_1 and p_2. Both arms offer the same chance of success on the current unit, but only arm 2 gives information that can help in choosing between the arms for treating later units.

Table 4 gives the breakeven values of p_1^* for selected values of n. This table shows that information is more important for larger n. For example, if $p_1 = 0.75$ then arm 1 would be optimal for $n = 10$, but it would be advisable to test arm 2 when $n = 100$; this is so even though arm 1 has probability of 0.75 of being better than arm 2.

When there are several arms with unknown characteristics, the problem is still more complicated. Optimal designs may well indicate selection of an arm that was used previously and set aside in favor of another arm because of inadequate performance. For the methods and theory for solving such problems, see Berry (1972), and Berry and Fristedt (1985). The optimal designs are generally difficult to describe. Berry (1978) provides easy to use sequential designs that are not optimal but that perform reasonably well.

Suppose the n units in the experiment are a subset of the N units on which arms 1 and 2 can be applied. Berry and Eick (1995) consider the case of two arms with dichotomous response and show how to incorporate all N units into the design problem. They find the optimal Bayes design when p_1 and p_2 have independent uniform prior distributions. They compare this with various other sequential designs and with a particular nonsequential design: balanced randomization to arms 1 and 2. The Bayes design performs best on average, of course, but it is robust in the sense that it outperforms the other designs for essentially all pairs of p_1 and p_2.

4. Further Reading

The pioneers in sequential statistical methods were Wald (1947) and Barnard (1944). They put forth the sequential probability ratio test (SPRT), which is of fundamental importance in sequential stopping problems. The study of the SPRT dominated the theory and methodology of sequential experimentation for decades.

For further reading about Bayesian vs. frequentist issues in sequential design, see Berger (1986), Berger and Berry (1988), Berger and Wolpert (1984), and Berry (1987, 1993). For further reading about the frequentist perspective, see Chow et al. (1971) and Whitehead (1992). For further reading about Bayesian design issues, see Berry (1993), Berry and Stangl (1996), Chernoff and Ray (1965), Cornfield (1966), and Lindley and Barnett (1965). For further reading about bandit problems, see Berry (1972, 1978), Berry and Eick (1995), Berry and Fristedt (1985), Bradt et al. (1956), Friedman et al. (1964), Murphy (1965), Rapoport (1967), Rothschild (1974), Viscusi (1979), and Whittle (1982/3). There is a journal called *Sequential Analysis* that is dedicated to the subject of this article.

See also: Clinical Treatment Outcome Research: Control and Comparison Groups; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Psychological Treatments: Randomized Controlled Clinical Trials; Quasi-Experimental Designs

Bibliography

Barnard G A 1944 *Statistical Methods and Quality Control*, Report No. QC/R/7. British Ministry of Supply, London

Berger J O 1986 *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Springer, New York

Berger J O, Berry D A 1988 Statistical analysis and the illusion of objectivity. *American Scientist* **76**: 159–65

Berger J O, Wolpert R L 1984 *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, CA

Berry D A 1972 A Bernoulli two-armed bandit. *Annals of Mathematical Statistics* **43**: 871–97

Berry D A 1978 Modified two-armed bandit strategies for certain clinical trials. *Journal of the American Statistical Association* **73**: 339–45

Berry D A 1987 Interim analysis in clinical trials: The role of the likelihood principle. *American Statistician* **41**: 117–22
Berry D A 1993 A case for Bayesianism in clinical trials (with discussion). *Statistics in Medicine* **12**: 1377–404
Berry D A 1996 *Statistics: A Bayesian Perspective*. Duxbury Press, Belmont, CA
Berry D A, Eick S G 1995 Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine* **14**: 231–46
Berry D A, Fristedt B 1985 *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London
Berry D A, Stangl D K 1996 Bayesian methods in health-related research. In: Berry D A, Stangl D K (eds.) *Bayesian Biostatistics*. Marcel Dekker, New York, pp. 1–66
Bradt R N, Johnson S M, Karlin S 1956 On sequential designs for maximizing the sum of *n* observations. *Annals of Mathematical Statistics* **27**: 1060–70
Chernoff H, Ray S N 1965 A Bayes sequential sampling inspection plan. *Annals of Mathematical Statistics* **36**: 1387–407
Chow Y S, Robbins H, Siegmund D 1971 *Great Expectations*. Houghton Mifflin, Boston
Cornfield J 1966 Sequential trials, sequential analysis and the likelihood principle. *American Statistician* **20**: 18–23
Friedman M P, Padilla G, Gelfand H 1964 The learning of choices between bets. *Journal of Mathematical Psychology* **1**: 375–85
Lindley D V, Barnett B N 1965 Sequential sampling: Two decision problems with linear losses for binomial and normal random variables. *Biometrika* **52**: 507–32
Murphy R E Jr. 1965 *Adaptive Processes in Economic Systems*. Academic Press, New York
Rapoport A 1967 Dynamic programming models for multistage decision making tasks. *Journal of Mathematical Psychology* **4**: 48–71
Rothschild M 1974 A two-armed bandit theory of market pricing. *Journal of Economic Theory* **9**: 185–202
Viscusi W K 1979 *Employment Hazards: An Investigation of Market Performance*. Harvard University Press, Cambridge, MA
Wald A 1947 *Sequential Analysis*. Wiley, New York
Whitehead J 1992 *The Design and Analysis of Sequential Clinical Trials*. Horwood, Chichester, UK
Whittle P 1982/3 *Optimization Over Time*. Wiley, New York, Vols. 1 and 2

D. A. Berry

Serial Verb Constructions

1. Definition and Importance

The term 'serial verb construction' (SVC) is usually applied to a range of apparently similar syntactic constructions in different languages, in which several verbs occur together within one clause or unit, without evidence of either subordination or coordination of the verbs. For example, sentences like (1), from the West African language Twi, are typical of what most scholars would regard as an SVC:

(1) Kofi de pono no baae
Kofi take-PAST table the come-PAST
'Kofi brought the table' or more literally,
'Kofi took the table and came with it'

SVCs have attracted attention from linguists for three main reasons. First, syntacticians have found them interesting because the possibility of having more than one main (i.e., nondependent, nonauxiliary) verb within a clause or clause-like unit challenges traditional assumptions that a clause contains exactly one predicate (see Foley and Olson 1985, pp. 17, 57). Second, because of their prevalence both in languages of West Africa and in certain Creole languages, they have become central to debates among Creole language researchers, being seen either as evidence for the importance of substrate (i.e., African mother-tongue) influence on the Creole, or as deriving from universal properties of human language which are manifested in the creolization process (see *Pidgin and Creole Languages*. Third, serial verbs show a tendency to be reanalyzed as other grammatical categories (complementizers and prepositions), with implications for the processes of historical change (see *Historical Linguistics*: *Overview*).

While different researchers use different criteria for defining serial verbs, the following set would probably be accepted by most and is based on one described by McWhorter (1997, p. 22) as a 'relatively uncontroversial distillation of the conclusions of various scholars':

(2) (a) an SVC contains only one overt subject;
(b) it contains no overt markers of coordination or subordination;
(c) it expresses concomitant actions (either simultaneous or consecutive) or a single action;
(d) it falls under one intonational contour;
(e) it has tense–modality–aspect marking on none of the verbs, or on one only (usually the first or last in the series) or on all of them; and
(f) it contains a second verb which is not obligatorily subcategorized for by the first verb.

Even the above six conditions leave it unclear where exactly to draw the line around SVCs, particularly in languages which have little morphology. In some cases of what may appear to be SVCs, we may be dealing simply with unmarked coordination of verb phrases expressing simultaneous actions.

2. Serial Verb Types

Serial constructions can be classified into different types on the basis of the functions of the verbs in the series relative to one another.

The sentences in (3) provide examples of different functional types of SVC which have been discussed in the literature.

(3) (a) *Directional complements*: V_2 is *go*, *come*, or another common intransitive motion verb which functions to indicate the directionality of the action denoted by V_1.
Olú gbé àga wá
Olu take chair come
'Olu brought a chair' (Yoruba, West African)
(b) *Other motion verb complements*: similar to type (a), but V_2 may be transitive motion verb with its own object expressing the goal of the action denoted by V_1
Kofi tow agyan no wɔɔ Amma
Kofi throw-PAST arrow the pierce-PAST Amma
'Kofi shot Amma with the arrow'
(c) *Instrumental constructions*: V_1 is *take* or a semantically similar common verb, while V_2 denotes an action performed with the aid of the object of V_1.
Kofi teki a nefi koti a brede
Kofi take the knife cut the bread
'Kofi cut the bread with a knife' (Sranan, Caribbean Creole)
(d) *Dative constructions*: V_2 is usually *give*, with an object which denotes the (semantic) indirect object of V_1.
Ogyaw ne sika mãã me
he-leave-PAST his money give-PAST me
"He left me his money" (Twi, West African)
(e) *Comparative constructions*: a verb meaning *pass* or *surpass* is used with an object to express comparison. In this case 'V_1' is often in fact an adjective.
Amba tranga pasa Kofi
Amba strong pass Kofi
'Amba is stronger than Kofi' (Sranan, Caribbean Creole)
(f) *Resultative constructions*: V_2 denotes the result or consequence of an action denoted by V_1.
Kofi naki Amba kiri
Kofi hit Amba kill
'Kofi struck Amba dead' (or more literally,
'Kofi hit Amba and killed her') (Sranan, Caribbean Creole)
(g) *Idiomatic constructions (lexical idioms)*: These are cases where the meaning of the verbs together is not derivable from the meanings of the verbs separately. They are not found in all serializing languages, but some of the West African group are particularly rich in them.
Anyi-Baule (West African):
bɔ 'hit' nĩã 'look' bɔ…nĩã 'say, tell'
Yoruba (West African):
là 'cut open' yé… 'understand' là…yé'explain'

3. Distribution in the World's Languages

Serial verb constructions have been identified in many languages but it is clear that they occur in areal clusters and are not spread evenly among the languages of the world. So far serializing languages have been identified in the following areas: West Africa (Kwa and related language families), the Caribbean (Creole languages which have a historical relationship with Kwa languages), Central America (e.g., Misumalpan), Papua New Guinea (Papuan languages and Tok Pisin—Melanesian Pidgin English), South-east Asia (e.g., Chinese, Vietnamese, Thai). There are isolated reports of SVC-like constructions from elsewhere. How credible these are depends on which criteria are adopted to define SVCs.

4. Grammatical Analyses

Numerous researchers have put forward proposals for grammatical analyses of SVCs. The analyses may be classified into two types, with some degree of overlap. *Semantic* analyses typically seek to account for SVCs in terms of how they break down verbal concepts into more basic semantic components (take–carry–come for *bring*, for example). *Syntactic* analyses differ mainly according to whether they treat SVCs as a phrase structure phenomenon (usually under a version of X-bar theory), as a form of complementation, subordination, or secondary predication (i.e., involving more than one clause-like unit), or as a lexical phenomenon (e.g., as single but disjoint lexical items (see *Syntax*)).

The following phrase structure (4) proposed for SVCs in the literature is typical of those advocated by a number of researchers:

(4) $VP \rightarrow V\ XP\ VP$ where X is N or P

This produces a right-branching tree with a theoretically infinite number of verbs in a series. Such a structure would allow lexical relations and relations like 'subject-of' and 'object-of' to hold between verbs in the series, and captures the intuition that verbs (or verb phrases) function as complements to verbs earlier in the sequence.

Some researchers have treated SVCs as involving complementation or subordination, with clauses or clause-like units embedded within the VP and dependent on a higher verb. For example, Larson (1991, pp. 198–205) has argued in favor of analyzing serial constructions as a case of 'secondary predication.' A structure similar to that for *Carol rubbed her finger raw* in English can also, he says, account for serial combinations of the *take… come* and *hit… kill* types (a) and (f) above).

Foley and Olson (1985) propose a distinction between 'nuclear' and 'core' serialization. In nuclear serialization the serial verbs all occur in the nucleus of the clause, in other words the verbs form a single unit which must share arguments and modifiers. The core layer of the clause consists of the nuclear layer plus the

core arguments of the verb (Foley and Olson 1985, p. 34). In 'core' serialization two cores, each with own nucleus and arguments, are joined together. Languages may exhibit both kinds of serialization or just one. Examples (a) and (b) below (Foley and Olson 1985, p. 38) illustrate this difference; (a) is an example of core, and (b) of nuclear, serialization:

(5) (a) fu fi fase isoe
he sit letter write
'He sat down and wrote a letter'
(Barai, Papuan)

(b) Fu fase fi isoe
He letter sit write
'He sat writing a letter'

This evidence suggests that 'serialization' may not be a unitary phenomenon, and although a single account which explains the two or more different construction types (such as Foley and Olson's) would be welcome, it may be that separate analyses may be needed as more different types of SVC come to light.

5. *The Creolists' Debates*

SVCs are among the most salient grammatical features of many Creoles of the Caribbean area which distinguish them grammatically from their European lexifier languages. As such they have attracted attention from creolists, who, finding similar structures in West African languages which historically are connected with the Creoles in question, have claimed SVCs are evidence of substrate influence on the grammar of the Creoles concerned.

An alternative view has been offered by Bickerton (1981) and Byrne (1987), who regard the similarities between SVCs in the Creole and West African languages as coincidental. Instead, they argue, SVCs result in Creoles from universal principles of grammar, and are a consequence of the rudimentary verbal structure (V but not VP) which exist, they say, in a Creole at the earliest stage of development. They point out the existence of SVCs or serial-like structures in other pidgins and Creoles such as Tok Pisin (New Guinea Pidgin) and Hawaiian Creole English, which have no historical connections with West Africa.

McWhorter (1997) argues against Bickerton and Byrne, claiming that the Caribbean Creoles not only share SVCs with West African languages, but that their SVCs are structurally similar to each other in ways that SVCs from other areal language groupings are not. Using the nuclear/core distinction proposed by Foley and Olson (see above), he argues that Papuan languages serialize at the nuclear level, while Kwa languages, Caribbean Creoles, Chinese, and other Southeast Asian languages serialize at the core level. Considering also the range of pidgins and Creoles which lack SVCs altogether, he concludes that 'SVCs have appeared around the world in precisely the creoles which had serializing substrata' (1997, p. 39).

SVCs have thus become a major site of contention between 'substratists' and 'universalists' in Creole studies, with both sides remaining committed to their own positions (see *Pidgin and Creole Languages*).

6. *SVCs and Language Change*

Particular verbs which participate in serial constructions in some cases appear to have been reanalyzed as members of another syntactic category, e.g., from verb to preposition (*give*→*for*) or complementizer (*say*→*that*). In such cases, the position of the verb usually makes it amenable to such a reanalysis (e.g., if it typically occurs immediately before its objects, it may be reanyalzed as a preposition if the semantics encourage this interpretation). Reanalyzed serial verbs lose, to varying degrees, their verbal properties and take on, again to varying degrees, the morphological characterisics of their new category. Lord (1973) describes cases of reanalysis of verbs as prepositions, comitative markers, and a subordinating conjunction in Yoruba, Gã, Ewe, and Fon. Verbs with the meaning '*say*' are susceptible to reinterpretation as complementizers (introducing sentential clauses), while those with the meaning '*finish*' have a tendency to be interpreted as completive aspect markers (see *Grammaticalization*).

7. *Related Constructions*

Lack of a satisfactory definition of 'serial verb construction' makes it difficult to decide what may count as a related kind of structure. Chinese has a range of verbal structures which bear resemblances to SVCs. One of these, the class of *co-verbs*, is widely believed to be the result of the reanalysis of serial verbs as prepositions. Others, so-called *resultatives*, resemble secondary predications.

The Bantu languages, though related to the serializing languages of West Africa, do not seem to have SVCs. However, some have verbal chains where the second and subsequent verb have different marking from the first. In many Bantu languages, the verb meaning *say* is homophonous with a complementizer. These phenomena remain to be explained by a general theory.

See also: Syntactic Aspects of Language, Neural Basis of; Syntax; Valency and Argument Structure in Syntax

Bibliography

Bickerton D 1981 *Roots of Language*. Karoma, Ann Arbor, MI

Byrne F 1987 *Grammatical Relations in a Radical Creole: Verb Complementation in Saramaccan*. Benjamins, Amsterdam

Foley W A, Olson M 1985 Clausehood and verb serialization. In: Nichols J A, Woodbury A C (eds.) *Grammar Inside and Outside the Clause: Some Approaches to Theory from the Field.* Cambridge University Press, Cambridge, UK, pp. 17–60
Larson R K 1991 Some issues in verb serialization. In: Lefebvre C (ed.) *Serial Verbs: Grammatical Comparative and Cognitive Approaches.* Benjamins, Amsterdam, pp. 184–210
Lefebvre C J (ed.) 1991 *Serial Verbs: Grammatical, Comparative and Cognitive Approaches.* Benjamins, Amsterdam
Lord C 1973 Serial verbs in transition. *Studies in African Lingusitics* **4**(3): 269–96
McWhorter J H 1997 *Towards a New Model of Creole Genesis.* Peter Lang, New York
Sebba M 1987 *The Syntax of Serial Verbs: An Investigation into Serialization in Sranan and other Languages.* Benjamins, Amsterdam

M. Sebba

Service Economy, Geography of

Producer services are types of services demanded primarily by businesses and governments, used as inputs in the process of production. We may split the demand for all services broadly into two categories: (a) demands originating from household consumers for services that they use, and (b) demands for services originating from other sources. An example of services demanded by household consumers is retail grocery services, while examples of producer services are management consulting services, advertizing services, and computer systems engineering.

1. *Development of the Term 'Producer Services'*

Producer Services have emerged as one of the most rapidly growing industries in advanced economies, while the larger service economy has also exhibited aggregate growth in most countries. In the 1930s, Fisher (1939) observed the general tendency for shifts in the composition of employment as welfare rose, with an expanding service sector. Empirical evidence of this transformation was provided by Clark (1957), and the Clark–Fisher model was a popular description of development sequences observed in nations that were early participants in the Industrial Revolution. However, this model was criticized as a necessary pattern of economic development by scholars who observed regions and nations that did not experience the pattern of structural transformation encompassed in the Clark–Fisher model. Moreover, critics of the Clark–Fisher model also argued that the size of the service economy relative to that of goods and primary production required efforts to classify the service economy into categories more meaningful than Fisher's residual category 'services.'

The classification of service industries based upon their nature and their source of demand was pioneered in the 1960s and 1970s. The classic vision of services developed by Adam Smith ('they perish the very instant of their performance') was challenged as scholars recognized that services such as legal briefs or management services can have enduring value, similar to investments in physical capital. The term producer services became applied to services with 'intermediate' as opposed to 'final' markets.

While this distinction between intermediate or producer services and final or consumer services is appealing, it is not without difficulties. Some industries do not fit neatly into one category or another. For example, households as well as businesses and governments demand legal services, and hotels—a function generally regarded to a consumer service—serve both business travelers and households. Moreover, many of the functions performed by producer service businesses are also performed internally by other businesses. An example is the presence of in-house accounting functions in most businesses, while at the same time there are firms specializing in performing accounting services for their clients. Notwithstanding these difficulties of classification, there is now widespread acceptance of the producer services as a distinctive category of service industry. However, variations in the classification of industries among nations, as well as differences in the inclusion of specific industries by particular scholars leads to varying sectoral definitions of producer services. Business, legal, and engineering and management services are included in most studies, while the financial services are often considered to be a part of the producer services. It is less common to consider services to transportation, wholesaling, and membership organizations as a component of the producer services.

Table 1 documents for the US economy growth in producer services employment between 1985 and 1995 by broad industrial groups. The growth of producer services over this time period was double the national rate of job growth, and this growth rate was almost identical in metropolitan and rural areas. While financial services grew relatively slowly, employment growth in business and professional services was very strong. Key sectors within this group include temporary help agencies, advertizing, copy and duplicating, equipment rental, computer services, detective and protective, architectural and engineering, accounting, research and development, and management consulting and public relations services.

2. *Early Research on the Geography of Producer Services*

Geographer's and regional economist's research on producer services only started in the 1970s. Pioneering

Table 1
US employment change in producer Services, 1985–95

Sector	Job growth (thousands)	Percentage growth
Finance, insurance, and real estate	994	16.6
Business and professional services	4,249	79.2
Legal services	275	40.0
Membership organizations	599	38.6
Total, producer services	6,117	45.0
Total, all industries	19,039	23.4

research includes the investigations of office location in the United Kingdom by Daniels (1975), and the studies of corporate control, headquarters, and research functions related to corporate headquarters undertaken in the United States by Stanback (1979). Daniels (1985) provided a comprehensive summary of this research, distinguishing between theoretical approaches designed to explain the geography of consumer and producer services. The emphasis in this early research focused on large metropolitan areas, recognizing the disproportionate concentration of producer services employment in the largest nodal centers. This concentration was viewed as a byproduct of the search for agglomeration economies, both within freestanding producer services and in the research and development organizations associated with corporate headquarters. This early research also identified the growth of international business service organizations that were concentrated in large metropolitan areas to serve both transnational offices of home-country clients and to generate an international client base.

Early research on producer services also documented tendencies for decentralization from central business districts into suburban locations, as well as the relatively rapid development of producer services in smaller urban regions. In the United States, Noyelle and Stanback (1983) published the first comprehensive geographical portrait of employment patterns within the producer services, differentiating their structure through a classification of employment, and relating this classification to growth trends between 1959 and 1976. While documenting relatively rapid growth of producer services in smaller urban areas with relatively low employment concentrations, this early research did not point towards employment decentralization, although uncertainties related to the impact of the development of information technologies were recognized as a factor that would temper future geographical trends.

The geography of producer services was examined thoroughly in the United Kingdom through the use of secondary data in the mid-1980s (Marshall et al. 1988). This research documented the growing importance of the producer services as a source of employment, in an economy that was shedding manufacturing jobs. It encouraged a richer appreciation of the role of producer services by policy-makers and scholars and called for primary research to better understand development forces within producer service enterprises.

While many scholars in this early period of research on producer services emphasized market linkages with manufacturers or corporate headquarter establishments, other research found industrial markets to be tied broadly to all sectors of the economy, a fact also borne out in input-output accounts. Research also documented geographic markets of producer service establishments, finding that a substantial percentage of revenues were derived from nonlocal clients, making the producer services a component of the regional economic base (Beyers and Alvine 1985).

3. *Emphasis in Current Research on the Geography of Producer Services*

In the 1990s there has been an explosion of geographic research on producer services. Approaches to research in recent years can be grouped into studies based on secondary data describing regional patterns or changes in patterns of producer service activity, and studies utilizing primary data to develop empirical insights into important theoretical issues. For treatment of related topics see *Location Theory*; *Economic Geography*; *Industrial Geography*; *Finance, Geography of*; *Retail Trade*; *Urban Geography*; and *Cities: Capital, Global, and World.*

3.1 *Regional Distribution of Producer Services*

The uneven distribution of producer service employment continues to be documented in various national studies, including recent work for the United States, Canada, Germany, and the Nordic countries. While these regions have exhibited varying geographic trends in the development of producer services, they share in common the relatively rapid growth of these industries. They also share in common the fact that relatively few regions have a concentration of or a share of employment at or above the national average.

In Canada, the trend has been towards greater concentration, a pattern which Coffey and Shearmur

(1996) describe as uneven spatial development. Utilizing data for the 1971 to 1991 time period, they find a broad tendency for producer service employment to have become more concentrated over time.

The trend in the United States, Germany, and the Nordic countries differs from that found in Canada. In the Nordic countries producer service employment is strongly concentrated in the capitals, but growth in peripheral areas of Norway and Finland is generally well above average, while in Denmark and Sweden many peripheral areas exhibit slow growth rates (Illeris and Sjøholt 1995). In Germany, the growth pattern exhibits no correlation between city size and the growth of business service employment (Illeris 1996). The United States also exhibits the uneven distribution of employment in the producer services. In 1995 only 34 of the 172 urban-focused economic areas defined by the US Bureau of Economic Analysis had an employment concentration in producer services at or above the national average, and these were predominately the largest metropolitan areas in the country. However, between 1985 and 1995 the concentration of employment in these regions diminished somewhat, while regions with the lowest concentrations in 1985 tended to show increases in their level of producer services employment. Producer service growth rates in metropolitan and rural territory have been almost identical over this same time period in the United States. There is also considerable evidence of deconcentration of employment within metropolitan areas from central business districts into 'Edge Cities' and suburban locations in Canada and the United States (Coffey et al. 1996a, Coffey et al. 1996b, Harrington and Campbell 1997). Survey research in Montreal indicates that this deconcentration is related more strongly to new producer service businesses starting up in suburban locations than to the relocation of existing establishments from central city locations (Coffey et al. 1996b).

3.2 Producer Services and Regional Development

Numerous studies have now been undertaken documenting the geographic markets of producer service establishments; for summaries see Illeris (1996) and Harrington et al. (1991). Illeris' summary of these studies indicates typical nonlocal market shares at 35 percent, but if nonlocal sales are calculated by weighting the value of sales the nonlocal share rises to 56 percent. In both urban and rural settings establishments are divided into those with strong interregional or international markets, and those with primarily localized markets. Over time establishments tend to become more export oriented, with American firms tending to have expanded interregional business, while European firms more often enter international markets (O'Farrell et al. 1998). Thus, producer services contribute to the economic base of communities and their contribution is rising over time due to their relatively rapid growth rate.

Growth in producer service employment in the United States has occurred largely through the expansion in the number of business establishments, with little change in the average size of establishments. Between 1983 and 1993 the number of producer service establishments increased from 0.95 million to 1.34 million, while average employment per establishment increased from 11 to 12 persons. The majority of this employment growth occurred in single unit establishments, nonpayroll proprietorships, and partnerships. People are starting these new producer service establishments because they want to be their own boss, they have identified a market opportunity, their personal circumstances lead them to wish to start a business, and they have identified ways to increase their personal income. Most people starting new companies were engaged in the same occupation in the same industry before starting their firm, and few report that they started companies because they were put out of work by their former employer in a move designed to downsize in-house producer service departments (Beyers and Lindahl 1996a).

The use of producer services also has regional development impacts. Work in New York State (MacPherson 1997) has documented that manufacturers who make strong use of producer services as inputs are more innovative than firms who do not use these services, which in turn has helped stimulate their growth rate. Similar positive impacts on the development of clients of producer service businesses has been documented in the United Kingdom (Wood 1996) and more broadly in Europe (Illeris 1996).

3.3 Demand Factors

Considerable debate has raged over the reasons for the rapid growth of the producer services, and one common perception has been that this growth has occurred because of downsizing and outsourcing by producer service clients to save money on the acquisition of these services. However, evidence has accumulated that discounts the significance of this perspective. Demand for producer services is increasing for a number of reasons beyond the cost of acquiring these services, based on research with both the users of these services and their suppliers (Beyers and Lindahl 1996a, Coffey and Drolet 1996). Key factors driving the demand for producer services include: the lack of expertise internal to the client to produce the service, a mismatch between the size of the client and the volume of their need for the service, the need for third-party opinions or expert testimony, increases in government regulations requiring use of particular services, and the need for assistance in managing the complexity of firms or to stay abreast of new technologies.

Evidence also indicates that producer service establishments often do business with clients who have in-house departments producing similar services. However, it is rare that there is direct competition with these in-house departments. Instead, relationships are generally complementary. Current evidence indicates that there have been some changes in the degree of in-house vs. market acquisition of producer services, but those selling these services do not perceive the balance of these in-house or market purchases changing dramatically (Beyers and Lindahl 1996a).

3.4 Supply and Competitive Advantage Considerations

The supply of producer services is undertaken in a market environment that ranges from being highly competitive to one in which there is very little competition. In order to position themselves in this marketplace, producer service businesses exhibit competitive strategies, and as with the demand side, recent evidence points towards the prevalence of competitive strategies based on differentiation as opposed to cost (Lindahl and Beyers 1999, Hitchens et al. 1996, Coffey 1996). The typical firm tries to develop a marketplace niche, positioning itself to be different from competitors though forces such as an established reputation for supplying their specialized service, the quality of their service, their personal attention to client needs, their specialized expertise, the speed with which they can perform the service, their creativity, and their ability to adapt quickly to client needs. The ability to deliver the service at a lower cost than the client could produce it is also a factor considered important by some producer service establishments.

3.5 Flexible Production Systems

The flexible production paradigm has been extended from its origin in the manufacturing environment in recent years into the producer services (Coffey and Bailly 1991). There are a variety of aspects to the issue of flexibility in the production process, including the nature of the service being produced, the way in which firms organize themselves to produce their services, and their relationships with other firms in the production and delivery process. The labor force within producer services has become somewhat more complex, with the strong growth of the temporary help industry that dispatches a mixture of part-time and full-time temporary workers, as well as some increase in the use of contract or part-time employees within producer service establishments. However, over 90 percent of employment remains full-time within the producer services (Beyers and Lindahl 1999).

The production of producer services is most frequently undertaken in a manner that requires the labor force to be organized in a customized manner for each job, although a minority of firm's approach work in a routinized manner. Almost three-fourths of firms rely on outside expertise to produce their service, and half of producer service firms engage in collaboration with other firms to produce their services to extend their range of expertise or geographic markets. The services supplied by producer service establishments are also changing frequently; half of the establishments interviewed in an US study had changed their services over the previous five years. They do so for multiple reasons, including changing client expectations, shifts in geographic or sectoral markets, changes in government regulations, and changes in information technologies with their related changes on the skills of employees. These changes most frequently produce more diversified service offerings, but an alternative pathway is to become more specialized or to change the form or nature of services currently being produced (Beyers and Lindahl 1999).

3.6 Information Technologies

Work in the producer services typically involves face to face meetings with clients at some point in the production process. This work may or may not require written or graphical documents. In some cases the client must travel to the producer service firm's office, and in other cases the producer service staff travels to client. These movements can be localized or international in scale. However, in addition to these personal meetings, there is extensive and growing use of a variety of information technologies in the production and delivery of producer services work. Computer networks, facsimile machines, telephone conversations, courier services, and the postal system all play a role, including a growing use of the Internet, e-mail, and computer file transfers between producers and clients. Some routine functions have become the subject of e-commerce, but much of the work undertaken in the producer services is nonroutine, requiring creative interactions between clients and suppliers.

3.7 Location Factors

The diversity of market orientations of producer service establishments leads to divergent responses with regard to the choice of business locations (Illeris 1996). Most small establishments are located convenient to the residential location of the founder. However, founders often search for a residence that suits them, a factor driven by quality of life considerations for many rural producer service establishments (Beyers and Lindahl 1996b). Businesses with spatially dispersed markets are drawn to locations convenient to travel networks, including airports and the Interstate Highway System in the US. Businesses

with highly localized markets position themselves conveniently to these markets. Often ownership of a building, or a prestigious site that may affect trust or confidence of clients is an influencing factor. Establishments that are part of large multi-establishment firms select locations useful from a firm-network-of-locations perspective, which may be either an intra- or interurban pattern of offices.

4. Methodological Considerations

Much of the research reported upon here has been conducted in the United States, Canada, or the United Kingdom. While there has been a rapid increase in the volume of research reported regarding the themes touched upon in the preceding paragraphs, the base of field research is still slender even within the nations just mentioned. More case studies are needed in urban and rural settings, focused on marketplace dynamics, production processes, and the impact of technological development on the evolution of the producer services. The current explosion of e-commerce and its use by producer service firms is a case in point. Recent research has tended to be conducted either by surveying producer service enterprises or their clients, but only rarely are both surveyed to obtain answers that can be cross-tabulated. Individual researchers have developed their protocol for survey research, yielding results that tend to be noncomparable to other research. Means need to be developed to bring theoretical and empirical approaches into greater consistency and comparability.

There are variations in the organization of production systems related to the producer services among countries. Some countries internalize these functions to a relatively high degree within other categories of industrial activity, and there needs to be greater thought given to ways in which research on these functions within these organizations can be measured.

5. Future Directions of Theory and Research

Research on the producer services will continue at both an aggregate scale, as well as at the level of the firm or establishment. Given the recent history of job growth in this sector of advanced economies, there will certainly be studies documenting the ongoing evolution of the geographical distribution of producer services. This research needs to proceed at a variety of spatial scales, ranging from intrametropolitan, to intermetropolitan or interregionally within nations, and across nations. International knowledge of development trends is currently sketchy, especially in developing countries where accounts may be less disaggregate than in developed countries with regard to service industries.

Research at the level of the firm and establishment must continue to better understand the start-up and evolution of firms. While there is a growing body of evidence regarding motivations and histories of firm founders, there is currently little knowledge of the movement of employees into and among producer service establishments. There is also meager knowledge of the distribution of producer service occupations and work in establishments not classified as producer services. Case studies are needed of collaboration, subcontracting, client-seller interaction, processes of price formation, the interplay between the evolution of information technologies and service-product concepts, and types of behavior that are related to superior performance as measured by indicators such as sales growth rate, sales per employee, or profit. There is also a pressing need for international comparative research, not just in the regions that have been relatively well researched (generally the US, Canada, and Europe), but also in other parts of the planet. There is also a pressing need for the development of more robust models and theory in relation to the geography of producer services.

See also: Geodemographics; Market Areas; Services Marketing

Bibliography

Beyers W B, Alvine M J 1985 Export services in postindustrial society. *Papers of the Regional Science Association* **57**: 33–45

Beyers W B, Lindahl D P 1996a Explaining the demand for producer services: Is cost-driven externalization the major factor? *Papers in Regional Science* **75**: 351–74

Beyers W B, Lindahl D P 1996b Lone eagles and high fliers in rural producer services. *Rural Development Perspectives* **12**: 2–10

Beyers W B, Lindahl D P 1999 Workplace flexibilities in the producer services. *The Service Industries Journal* **19**: 35–60

Clark C 1957 *The Conditions of Economic Progress*. Macmillan, London

Coffey W J 1996 Forward and backward linkages of producer service establishments: Evidence from the Montreal metropolitan area. *Urban Geography* **17**: 604–32

Coffey W J, Bailly A S 1991 Producer services and flexible production: An exploratory analysis. *Growth and Change* **22**: 95–117

Coffey W J, Drolet R 1996 Make or buy? Internalization and externalization of producer service inputs in the Montreal metropolitan area. *Canadian Journal of Regional Science* **29**: 25–48

Coffey W J, Drolet R, Polèse M 1996a The intrametropolitan location of high order services: Patterns, factors and mobility in Montreal. *Papers in Regional Science* **75**: 293–323

Coffey W J, Polèse M, Drolet R 1996b Examining the thesis of central business district decline: Evidence from the Montreal metropolitan area. *Environment and Planning A* **28**: 1795–1814

Coffey W J, Shearmur R G 1996 *Employment Growth and Change in the Canadian Urban System, 1971–94*. Canadian Policy Research Networks, Ottawa, Canada
Daniels P W 1975 *Office Location*. Bell, London
Daniels P W 1985 *Service Industries, A Geographical Appraisal*. Methuen, London
Fisher A 1939 Production, primary, secondary, and tertiary. *Economic Record* **15**: 24–38
Harrington J W, Campbell H S Jr 1997 The Suburbanization of Producer Service Employment. *Growth and Change* **28**: 335–59
Harrington J W, MacPherson A D, Lombard J R 1991 Interregional trade in producer services: Review and synthesis. *Growth and Change* **22**: 75–94
Hitchens D M W N, O'Farrell P N, Conway C D 1996 The competitiveness of business services in the Republic of Northern Ireland, Wales, and the south East of England. *Environment and Planning A* **28**: 1299–1313
Illeris S 1996 *The Service Economy. A Geographical Approach*. Wiley, Chichester, UK
Illeris S, Sjøholt P 1995 The Nordic countries: High quality services in a low density environment. *Progress Planning* **43**: 205–221
Lindahl D P, Beyers W B 1999 The creation of competitive advantage by producer service establishments. *Economic Geography* **75**: 1–20
MacPherson A 1997 The role of producer service outsourcing in the innovation performance of New York State manufacturing firms. *Annals of the Association of American Geographers* **87**: 52–71
Marshall J, Wood P A, Daniels P W, McKinnon A, Bachtler J, Damesick P, Thrift N, Gillespie A, Green A, Leyshon A 1988 *Services and Uneven Development*. Oxford University Press, Oxford, UK
Noyelle T J, Stanback T M 1983 *The Economic Transformation of American Cities*. Rowman and Allanheld, Totowa, NJ
O'Farrell P N, Wood P A, Zheng J 1998 Regional influences on foreign market development by business service companies: Elements of a strategic context explanation. *Regional Studies* **32**: 31–48
Stanback T M 1979 *Understanding the Service Economy: Employment, Productivity, Location*. Allanheld and Osmun, Totowa, NJ
Wood P 1996 Business services, the management of change and regional development in the UK: A corporate client perspective. *Transactions of the Institute of British Geographers* **21**: 649–65

W. B. Beyers

Services Marketing

Marketing, as a philosophy and as a function, has already reached the maturity stage. In the 1950s, the basic emphasis was on consumer goods and the mass marketing approach; in the 1960s, it was the marketing potentiality of durable goods; and in the 1970s; the emphasis was on industrial goods.

Only in the 1980s did service organizations start to take a professional interest in marketing approaches and tools. More recently also, the public services, and the nonprofit services in general, have begun to participate in marketing. Despite its recent arrival, services marketing is undoubtedly the most innovative and, thanks to new technology, seems as though it may change the old approach completely. The trend is moving from mass marketing to one-to-one marketing, a typical feature of services promotion.

1. *Peculiarities of Services Marketing*

If services marketing is becoming so important, it is necessary to understand its peculiarities. Everything is related to the *intangibility of services*: they are 'events' more than 'goods,' and as a consequence it is impossible to store them. They are 'happenings,' where it is difficult to forecast what will be produced and what the consumer will get. The relationship becomes increasingly important, especially when the consumer becomes a 'part-time producer'—a 'pro-sumer' (producer + consumer).

Interaction with people is very important in many services, such as restaurants, air transportation, tourism, banking, and so on. The interaction is not only between producer and consumer, but also among the users, as happens in schools among students.

Because of these elements there is another very important aspect: in services marketing it is difficult to standardize performance, and quality can be very different from one case to another. Quality control ahead of the event is impossible, because nobody knows in advance the service that will actually be provided to the customer. Many external factors can affect the level of quality, e.g., the climate, or the unpredictable behavior of customers. Consequently, it is not always possible to fulfill promises at the point of delivery, even if the organization does its best to deliver the required level of quality.

Another peculiarity is that services are not 'visible' and so cannot represent a 'status symbol,' unless the service can be matched with tangible objects, such as credit cards, university ties, football scarves, etc. But the most important aspect is that services cannot be stored, and therefore production and use are simultaneous. This creates problems in terms of production capacity, because it is difficult to match fluctuations in demand over seasons, weeks, or days.

It is possible to synchronize production and use by means of different tools. The most used approach is pricing, because varying service prices may be applied at different periods of time (e.g., telephone charges, or airline tickets). There are other tools, such as 'user education,' implemented by the service producer. In this case, the producer tries to 'educate' the consumer to ask for a particular, better-quality, service during the so-called 'valley periods' when sales are slow.

Other solutions involve the employment of part-time workers during 'peak periods,' or the implemen-

tation of maintenance activity in the 'valley periods'; making reservations is a typical approach in the case of theatres or in the health care service. In all cases it is necessary to overcome fluctuating demand with a 'synchromarketing activity.'

2. Critical Success Factors in Services Marketing

Because of the peculiarities of the services, there are many critical success factors in services marketing: (a) service industrialization, (b) strategic image management, (c) customer satisfaction surveys, (d) operations, personnel, and marketing interaction, (e) internal marketing, (f) user education, and participation, (g) managerial control of costs, and investments, (h) relationship with the public authorities, (i) quality control, and (j) synchromarketing.

Of these, further comment can be devoted to internal marketing, whose goal is 'to create an internal environment which supports customer consciousness and sales mindedness among the personnel' (Gronroos 1990). 'Internal marketing means applying the philosophy and practices of marketing to people who serve the external customers, so that the best possible people can be employed and retained, and they will do the best possible work' (Berry and Parasuraman 1991).

Another point to emphasize is quality control, because, as has been said, the quality of the service is perceived to be more than objective. Many studies say that the focal point in services is the perceived quality, as a result of a comparison between the expected quality and the image of the service really provided. As a consequence, it is very important to survey customer opinion on a regular basis, and to check whether, via this opinion, service is becoming better or worse. A further point is related to the importance of a company's image as a factor affecting, positively or negatively, customers' judgment. This is why it is necessary to manage a company's image with a strongly strategic approach, outlining what the company is, what people think it is, and how it wants people to think of it. Image improvement is not only a matter of communication; it also involves personnel behavior, services provided, and material facilities.

3. Future Perspectives

At the start of the twenty-first century, services marketing is very important in the developed countries, because of the relevance of services in terms of GNP, employment, etc., and we can predict that it is going to increase further in importance, thanks to new technologies that facilitate the relationship between producer and customer. There will be an increasing amount of one-to-one marketing, which is particularly suitable for services. Another important development will concern the internationalization of services (e.g., McDonald's, Club Med, Manchester United, Hertz, etc.), where the critical success factors will be related to the improvement of 'high tech' in the back office and 'high touch' in the front line: Industrialization and personalization at the same time, to achieve high customer satisfaction in every country. Sales of products now often include the offer of some services, as, for example, in the automotive industry, where this is represented by the after sales service, a very important aspect affecting the buyer decision process. Offers such as this are the key issues that, in most cases, actually make the difference, given the fact that the basic product is very frequently similar to other products, and the difference is provided by the service included in the product.

See also: Advertising Agencies; Advertising and Advertisements; Advertising: General; Advertising, Psychology of; Computers and Society; Internet: Psychological Perspectives; Market Research; Marketing Strategies; Markets: Artistic and Cultural; Service Economy, Geography of

Bibliography

Berry L L, Parasuraman A 1991 *Marketing Services—Competing Through Quality*. The Free Press
Cherubini S 1981 *Il Marketing dei Servizi*. Franco Angeli Editore
Cowell D 1984 *The Marketing of Services*. Heinemann
Donnelly J H, George W R (eds.) 1981 *Marketing of Services*, AMA's Proceedings Series.
Eiglier P, Langeard E 1987 *Servuction—Le Marketing des Services*. McGraw-Hill, New York
Eiglier P, Langeard E, Lovelock C H, Bateson J E G, Young R F 1977 *Marketing Consumer Services: New Insights*. Marketing Science Institute
Fisk R P 2000 *Interactive Services Marketing*. Hougthon Mifflin
Gronroos C 1990 *Service Management and Marketing. Managing the Moments of Truth in Service Competition*. Lexington Books
Heskett J L, Sasser W E, Hart C W L 1986 *Service Breakthroughs. Changing the Rules of the Game*. The Free Press, New York
Lovelock C H 1984 *Services Marketing*. Prentice-Hall, Englewood Cliffs, NJ
Normann R 1984 *Service Management. Strategy and Leadership in Service Businesses*. Wiley, New York
Payne A 1993 *The Essence of Services Marketing*. Prentice-Hall, Englewood Cliffs, NJ
Payne A, McDonald M H B 1997 *Marketing Planning for Services*. Butterworth-Heinemann, Oxford, UK
Zeithamal V A 1996 *Services Marketing*. McGraw-Hill, New York

S. Cherubini

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

13936

 ISBN: 0-08-043076-7

Set

Settlement and Landscape Archaeology

For contemporary archaeology, settlement and landscape approaches represent an increasingly important focus that is vital for a core mission of the discipline to describe, understand, and explain long-term cultural and behavioral change. Despite this significance, few syntheses of this topic have been undertaken (cf. Parsons 1972, Ammerman 1981, Fish and Kowalewski 1990, Billman and Feinman 1999). Yet settlement and landscape approaches provide the only large-scale perspective for the majority of premodern societies. These studies are reliant on archaeological surface surveys, which discover and record the distribution of material traces of past human presence/habitation across a landscape (see *Survey and Excavation (Field Methods) in Archaeology*). The examination and analysis of these physical remains found on the ground surface (e.g., potsherds, stone artifacts, house foundations, or earthworks) provide the empirical foundation for the interpretation of ancient settlement patterns and landscapes.

1. Historical Background

Although the roots of settlement pattern and landscape approaches extend back to the end of the nineteenth century, archaeological survey has only come into its own in the post World War II era. Spurred by the analytical emphases of Steward (1938), Willey's Virú Valley archaeological survey (1953) provided a key impetus for settlement pattern research in the Americas. In contrast, the landscape approach, which has a more focal emphasis on the relationship between sites and their physical environments, has its roots in the UK. Nevertheless, contemporary archaeological studies indicate a high degree of intellectual cross-fertilization between these different surface approaches.

1.1 Early Foundations for Archaeological Survey in the Americas and England

The American settlement pattern tradition stems back to scholars, such as Morgan (1881), who queried how the remnants of Native American residential architecture reflected the social organization of the native peoples who occupied them. Yet the questions posed by Morgan led to relatively few immediate changes in how archaeology was practiced, and for several decades few scholars endeavored to address the specific questions regarding the relationship between settlement and social behavior that Morgan posed. When surface reconnaissance was undertaken by archaeologists, it tended to be a largely unsystematic exercise carried out to find sites worthy of excavation.

In the UK, the landscape approach, pioneered by Fox (1922), was more narrowly focused on the definition of distributional relationships between different categories of settlements and environmental features (e.g., soils, vegetation, topography). Often these early studies relied on and summarized surveys and excavations that were carried out by numerous investigators using a variety of field procedures rather than more uniform or systematic coverage implemented by a single research team. At the same time, the European landscape tradition generally has had a closer link to romantic thought as opposed to the more positivistic roots of the North American settlement pattern tradition (e.g., Sherratt 1996).

1.2 The Development of Settlement Archaeology

By the 1930s and 1940s, US archaeologists working in several global regions recognized that changing patterns of social organization could not be reconstructed and interpreted through empirical records that relied exclusively on the excavation of a single site or community within a specific region. For example, in the lower Mississippi Valley, Phillips et al. (1951) located and mapped archaeological sites across a large area to analyze shifting patterns of ceramic styles and settlements over broad spatial domains and temporal contexts.

Yet the most influential and problem-focused investigation of that era was that of Willey in the Virú Valley. Willey's project was the first to formally elucidate the scope and potential analytical utility of settlement patterns for understanding long-term change in human economic and social relationships. His vision moved beyond the basic correlation of environmental features and settlements as well as beyond the mere definition of archtypical settlement types for a given region. In addition to its theoretical contributions, the Virú program also was innovative methodologically, employing (for the first time in the Western Hemisphere) vertical air photographs in the location and mapping of ancient settlements. Al-

though Willey did not carry out his survey entirely on foot, he did achieve reasonably systematic areal coverage for a defined geographic domain for which he could examine changes in the frequency of site types, as well as diachronic shifts in settlement patterns.

Conceptually and methodologically, these early settlement pattern projects of the 1930s and 1940s established the intellectual underpinnings for a number of multigenerational regional archaeological survey programs that were initiated in at least four global regions during the 1950s and 1960s. In many ways, these later survey programs were integral to the theoretical and methodological re-evaluations that occurred in archaeological thought and practice under the guise of 'the New Archaeology' or processualism. The latter theoretical framework stemmed in part from an expressed emphasis on understanding long-term processes of behavioral change and cultural transition at the population (and so regional) scale. This perspective, which replaced a more normative emphasis on archtypical sites or cultural patterns, was made possible to a significant degree by the novel diachronic and broad scalar vantages pieced together for specific areas through systematic regional settlement pattern fieldwork and analysis.

1.3 Large-scale Regional Survey Programs

During the 1950s through the 1970s, major regional settlement pattern programs were initiated in the heartlands of three areas where early civilizations emerged (Greater Mesopotamia, highland Mexico, and the Aegean), as well as in one area known for its rich and diverse archaeological heritage (the Southwest USA). The achievements of the Virú project also stimulated continued Andean settlement pattern surveys, although a concerted push for regional research did not take root there until somewhat later (e.g., Parsons et al. 1997, Billman and Feinman 1999).

Beginning in 1957, Robert M. Adams (e.g., 1965, 1981) and his associates methodically traversed the deserts and plains of the Near East by jeep, mapping earthen tells and other visible sites. Based on the coverage of hundreds of square kilometers, these pioneering studies of regional settlement history served to unravel some of the processes associated with the early emergence of social, political, and economic complexity in Greater Mesopotamia. Shortly thereafter, in highland Mexico, large-scale, systematic surveys were initiated in the area's two largest mountain valleys (the Basin of Mexico and the Valley of Oaxaca). These two projects implemented field-by-field, pedestrian coverage of some of the largest contiguous survey regions in the world, elucidating the diachronic settlement patterns for regions in which some of the earliest and most extensive cities in the ancient Americas were situated (e.g., Sanders et al. 1979, Blanton et al. 1993). After decades, about half of the Basin of Mexico and almost the entire Valley of Oaxaca were traversed by foot.

In the Aegean, regional surveys (McDonald and Rapp 1972, Renfrew 1972) were designed to place important sites with long excavation histories in broader spatial contexts. Once again, these investigations brought new regional vantages to areas that already had witnessed decades of excavation and textual analyses. Over the same period, settlement pattern studies were carried out in diverse ecological settings across the US Southwest, primarily to examine the differential distributions of archaeological sites in relation to their natural environments, and to determine changes in the numbers and sizes of settlements across the landscape over time. In each of the areas investigated, the wider the study domain covered, the more diverse and complex were the patterns found. Growth in one part of a larger study area was often timed with the decrease in the size and number of sites in another. And settlement trends for given regions generally were reflected in episodes of both growth and decline.

Each of these major survey regions (including much of the Andes) is an arid to semiarid environment. Without question, broad-scale surface surveys have been most effectively implemented in regions that lack dense ground cover, and therefore the resultant field findings have been most robust. In turn, these findings have fomented long research traditions carried out by trained crews, thereby contributing to the intellectual rewards of these efforts. As Ammerman (1981, p. 74) has recognized, 'major factors in the success of the projects would appear to be the sheer volume of work done and the experience that workers have gradually built up over the years.'

1.4 Settlement Pattern Research at Smaller Scales of Analysis

Although settlement pattern approaches were most broadly applied at the regional scale, other studies followed similar conceptual principles in the examination of occupational surfaces, structures, and communities. At the scale of individual living surfaces or house floors, such distributional analyses have provided key indications as to which activities (such as cooking, food preparation, and toolmaking) were undertaken in different sectors (activity areas) of specific structures (e.g., Flannery and Winter 1976) or surfaces (e.g., Flannery 1986, pp. 321–423). In many respects, the current emphasis on household archaeology (e.g., Wilk and Rathje 1982) is an extension of settlement pattern studies (see *Household Archaeology*). Both household and settlement pattern approaches have fostered a growing interest in the non-elite sector of complex societies, and so have spurred the effort to understand societies as more than just undifferentiated, normative wholes.

At the intermediate scale of single sites or communities, settlement pattern approaches have compared the distribution of architectural and artifactual evidence across individual sites. Such investigations have clearly demonstrated significant intrasettlement variation in the functional use of space (e.g., Hill 1970), as well as distinctions in socioeconomic status and occupational history (e.g., Blanton 1978). From a comparative perspective, detailed settlement pattern maps and plans of specific sites have provided key insights into the similarities and differences between contemporaneous cities and communities in specific regions, as well as the elucidation of important patterns of cross-cultural diversity.

2. *Contemporary Research Strategies and Ongoing Debates*

The expansion of settlement pattern and landscape approaches over the last decades has promoted the increasing acceptance of less normative perspectives on cultural change and diversity across the discipline of archaeology. In many global domains, archaeological surveys have provided a new regional-scale (and in a few cases, macroregional-scale) vantage on past social systems. Settlement pattern studies also have yielded a preliminary means for estimating the parameters of diachronic demographic change and distribution at the scale of populations, something almost impossible to obtain from excavations alone. Nevertheless, important discussions continue over the environmental constraints on implementation, the relative strengths and weaknesses of different survey methodologies, issues of chronological control, procedures for population estimation, and the appropriate means for the interpretation of settlement pattern data.

2.1 *Environmental Constraints*

Although systematic settlement pattern and landscape studies have been undertaken in diverse environmental settings including heavily vegetated locales such as the Guatemalan Petén, the eastern woodlands of North America, and temperate Europe, the most sustained and broadly implemented regional survey programs to date have been enacted in arid environments. In large part, this preference pertains to the relative ease of finding the artifactual and architectural residues of ancient sites on the surface of landscapes that lack thick vegetal covers. Nevertheless, archaeologists have devised a variety of means, such as the interpolation of satellite images, the detailed analysis of aerial photographs, and subsurface testing programs, that can be employed to locate and map past settlements in locales where they are difficult to find through pedestrian coverage alone. In each study area, regional surveys also have to modify their specific field methodologies (the intensity of the planned coverage) and the sizes of the areas that they endeavor to examine to the nature of the terrain and the density of artifactual debris (generally nonperishable ancient refuse) associated with the sites in the specified region. For example, sedentary pottery-using peoples generally created more garbage than did mobile foragers; the latter usually employed more perishable containers (e.g., baskets, cloth bags). Consequently, other things being equal, the sites of foragers are generally less accessible through settlement and landscape approaches than are the ancient settlements that were inhabited for longer durations (especially when ceramics were used).

2.2 *Survey Methodologies and Sampling*

Practically since the inception of settlement pattern research, archaeologists have employed a range of different field survey methods. A critical distinction has been drawn between full-coverage and sample surveys. The former approaches rely on the complete and systematic coverage of the study region by members of a survey team. In order to ensure the full coverage of large survey blocks, team members often space themselves 25–50 m apart, depending on the specific ground cover, the terrain, and the density of archaeological materials. As a consequence, isolated artifact finds can occasionally be missed. But the researchers generally can discern a reasonably complete picture of settlement pattern change across a given region. Alternatively, sample surveys by definition are restricted to the investigation of only a part of (a sample of) the study region. Frequently such studies (because they only cover sections of larger regions) allow for the closer spacing of crew members.

Archaeologists have employed a range of different sampling designs. Samples chosen for investigation may be selected randomly or stratified by a range of diverse factors, including environmental variables. Nevertheless, regardless of the specific sampling designs employed, such sample surveys face the problem of extrapolating the results from their surveyed samples to larger target domains that are the ultimate focus of study. Ultimately, such sample surveys have been shown to be more successful at estimating the total number of sites in a given study region than at defining the spacing between sites or at discovering rare types of settlement. The appropriateness of sample design can only be decided by the kinds of information that the investigator aims to recover. There is no single correct way to conduct archaeological survey, but certain methodological procedures have proven more productive in specific contexts and given particular research aims.

2.3 *Chronological Constraints and Considerations*

One of the principal strengths of settlement pattern research is that it provides a broad-scale perspective

on the changing distribution of human occupation across landscapes. Yet the precision of such temporal sequences depends on the quality of chronological control (see *Chronology, Stratigraphy, and Dating Methods in Archaeology*). The dating of sites during surveys must depend on the recovery and temporal placement of chronologically diagnostic artifacts from the surface of such occupations. Artifacts found on the surface usually are already removed from their depositional contexts. Finer chronometric dating methods generally are of little direct utility for settlement pattern research, since such methods are premised on the recovery of materials in their depositional contexts. Of course, chronometric techniques can be used in more indirect fashion to refine the relative chronological sequences that are derived from the temporal ordering of diagnostic artifacts (typically pottery).

In many regions, the chronological sequences can only be refined to periods of several hundred years in length. As a result, sites of shorter occupational durations that may be judged to be contemporaneous in fact could have been inhabited sequentially. In the same vein, the size of certain occupations may be overestimated as episodes of habitation are conflated. Although every effort should be made to minimize such analytical errors, these problems in themselves do not negate the general importance of the long-term regional perspective on occupational histories that in many areas of the world can be derived from archaeological survey alone. Although the broad-brush perspective from surveys may never provide the precision or detailed views that are possible from excavation, they yield an encompassing representation at the population scale that excavations cannot achieve. Adequate holistic perspectives on past societies rely on the multiscalar vantages that are provided through the integration of wide-ranging archaeological surveys with targeted excavations.

2.4 Population Estimation

One of the key changes in archaeological thought and conceptualization over the past half-century has been the shift from essentialist/normative thinking about ancient societies to a more populational perspective. But the issue of how to define past populations, their constituent parts, and the changing modes of interaction between those parts remains challenging at best. Clearly, multiscalar perspectives on past social systems are necessary to collect the basic data required to estimate areal shifts in population size and distribution. Yet considerable debate has been engendered over the means employed by archaeologists to extrapolate from the density and dispersal of surface artifacts pertaining to a specific phase to the estimated sizes of past communities or populations.

Generally, archaeologists have relied on some combination of the empirically derived size of a past settlement, along with a comparative determination of surface artifact densities at that settlement, to generate demographic estimates for a given community. When the estimates are completed for each settlement across an entire survey region, extrapolations become possible for larger study domains. By necessity, the specific equations to estimate past populations vary from one region to another because community densities are far from uniform over time or space. Yet due to chronological limitations, as well as the processes of deposition, disturbance, and destruction, our techniques for measuring ancient populations remain coarse-grained. Although much refinement is still needed to translate survey data into quantitative estimates of population with a degree of precision and accuracy, systematic regional surveys still can provide the basic patterns of long-term demographic change over time and space that cannot be ascertained in any other way.

2.5 The Interpretation of Regional Data

Beyond the broad-brush assessment of demographic trends and site distribution in relation to environmental considerations, archaeologists have interpreted and analyzed regional sets of data in a variety of ways. Landscape approaches, which began with a focused perspective on humans and their surrounding environment, have continued in that vein, often at smaller scales. Such studies often examine in detail the placement of sites in a specific setting with an eye toward landscape conservation and the meanings behind site placement (Sherratt 1996). At the same time, some landscape studies have emphasized the identification of ancient agrarian features and their construction and use.

In contrast, the different settlement pattern investigations have employed a range of analytical and interpretive strategies. In general, these have applied more quantitative procedures and asked more comparatively informed questions. Over the last 40 years (e.g., Johnson 1977), a suite of locational models derived from outside the discipline has served as guides against which different sets of archaeological data could be measured and compared. Yet debates have arisen over the underlying assumptions of such models and whether they are appropriate for understanding the preindustrial past. For that reason, even when comparatively close fits were achieved between heuristically derived expectations and empirical findings, questions regarding equifinality (similar outcomes due to different processes) emerged. More recently, theory-building efforts have endeavored to rework and expand these locational models to specifically archaeological contexts with a modicum of success. Continued work in this vein, along with the integration of some of the conceptual strengths from both landscape and settlement pattern approaches are requisite to under-

standing the complex web of relations that govern human-to-human and human-to-environment interactions across diverse regions over long expanses of time.

3. Looking Forward: The Critical Role of Settlement Studies

The key feature and attribute of archaeology is its long temporal panorama on human social formations. Understanding these formations and how they changed, diversified, and varied requires a regional/populational perspective (as well as other vantages at other scales). Over the last century, the methodological and interpretive toolkits necessary to obtain this broad-scale view have emerged, diverged, and thrived. The emergence of archaeological survey (and settlement pattern and landscape approaches) has been central to the disciplinary growth of archaeology and its increasing ability to address and to contribute to questions of long-term societal change. At the same time, the advent of settlement pattern studies has had a critical role in moving the discipline as a whole from normative to populational frameworks.

Yet settlement pattern work has only recently entered the popular notion of this discipline, long wrongly equated with and defined by excavation alone. Likewise, many archaeologists find it difficult to come to grips with a regional perspective that has its strength in (broad) representation at the expense of specific reconstructed detail. Finally, the potential for theoretical contributions and insights from settlement pattern and landscape approaches (and the wealth of data collected by such studies) has only scratched the surface. In many respects, the growth of regional survey and analysis represents one of the most important conceptual developments of twentieth-century archaeology. Yet at the same time, there are still so many mountains (literally and figuratively) to climb.

See also: Chronology, Stratigraphy, and Dating Methods in Archaeology; Household Archaeology; Survey and Excavation (Field Methods) in Archaeology

Bibliography

Adams R M 1965 *Land Behind Baghdad: A History of Settlement on the Diyala Plains*. University of Chicago Press, Chicago

Adams R M 1981 *Heartland of Cities: Surveys of Ancient Settlement and Land Use on the Central Floodplain of the Euphrates*. University of Chicago Press, Chicago

Ammerman A J 1981 Surveys and archaeological research. *Annual Review of Anthropology* **10**: 63–88

Billman B R, Feinman G M (eds.) 1999 *Settlement Pattern Studies in the Americas: Fifty Years Since Virú*. Smithsonian Institution Press, Washington, DC

Blanton R E 1978 *Monte Albán: Settlement Patterns at the Ancient Zapotec Capital*. Academic Press, New York

Blanton R E, Kowalewski S A, Feinman G M, Finsten L M 1993 *Ancient Mesoamerica: A Comparison of Change in Three Regions*, 2nd edn. Cambridge University Press, Cambridge, UK

Fish S K, Kowalewski S A (eds.) 1990 *The Archaeology of Regions: A Case for Full-coverage Survey*. Smithsonian Institution Press, Washington, DC

Flannery K V (ed.) 1986 *Guilá Naquitz: Archaic Foraging and Early Agriculture in Oaxaca, Mexico*. Academic Press, Orlando, FL

Flannery K V, Winter M C 1976 Analyzing household activities. In: Flannery K V (ed.) *The Early Mesoamerican Village*. Academic Press, New York, pp. 34–47

Fox C 1923 *The Archaeology of the Cambridge Region*. Cambridge University Press, Cambridge, UK

Hill J N 1970 *Broken K Pueblo: Prehistoric Social Organization in the American Southwest*. University of Arizona Press, Tucson, AZ

Johnson G A 1977 Aspects of regional analysis in archaeology. *Annual Review of Anthropology* **6**: 479–508

McDonald W A, Rapp G R Jr (eds.) 1972 *The Minnesota Messenia Expedition. Reconstructing a Bronze Age Environment*. University of Minnesota Press, Minneapolis, MN

Morgan L H 1881 *Houses and House Life of the American Aborigines*. US Department of Interior, Washington, DC

Parsons J R 1972 Archaeological settlement patterns. *Annual Review of Anthropology* **1**: 127–50

Parsons J R, Hastings C M, Matos R 1997 Rebuilding the state in highland Peru: Herder-cultivator interaction during the Late Intermediate period in Tarama-Chinchaycocha region. *Latin American Antiquity* **8**: 317–41

Phillips P, Ford J A, Griffin J B 1951 *Archaeological Survey in the Lower Mississippi Alluvial Valley, 1941–47*. Peabody Museum of Archaeology and Ethnology, Cambridge, MA

Renfrew C 1972 *The Emergence of Civilisation. The Cyclades and the Aegean in the Third Millennium BC*. Methuen, London

Sanders W T, Parsons J R, Santley R S 1979 *The Basin of Mexico: Ecological Processes in the Evolution of a Civilization*. Academic Press, New York

Sherratt A 1996 'Settlement patterns' or 'landscape studies'? Reconciling reason and romance. *Archaeological Dialogues* **3**: 140–59

Steward J H 1938 *Basin Plateau Aboriginal Sociopolitical Groups*. Bureau of American Ethnology, Washington, DC

Wilk R R, Rathje W L (eds). 1982 *Archaeology of the Household: Building a Prehistory of Domestic Life. American Behavioral Scientist* **25**(6)

Willey G R 1953 *Prehistoric Settlement Patterns in the Virú Valley, Peru*. Bureau of American Ethnology, Washington, DC

G. M. Feinman

Sex Differences in Pay

Differences in pay between men and women remain pervasive in late twentieth-century labor markets, with women's average earnings consistently below men's

even when differences in hours of work are taken into account. This 'gender pay gap' may result from unequal pay within jobs, but is also related to the different types of jobs occupied by men and women. Considerable debate has arisen over the extent to which it is evidence of discrimination in labor markets or simply a result of the individual attributes and choices of men and women. The size of the pay gap may also be influenced by the institutional framework of pay bargaining and regulation in different countries. The article reviews a range of factors thought to contribute to the gender pay gap, and the strategies designed to eradicate it. It draws primarily on evidence from affluent industrialized nations.

1. *The Gender Pay Gap: Trends and Comparisons*

While the gap between men's and women's average earnings has narrowed in most countries in the late twentieth century, significant inequality remains. Moreover, substantial cross-national differences are apparent. Table 1 illustrates some of these variations in selected countries. To enhance comparison, the data are drawn as far as possible from one source (International Labour Office [ILO] *Yearbook of Labour Statistics*). Where available, the figures are for hourly earnings, as data for longer periods produce deflated estimates of women's relative earnings due to the tendency for women to work fewer hours than men. Also, while the most widely available and reliable data are for manufacturing, women's employment in industrialized nations tends to be concentrated more in service industries. Thus earnings in the ILO's more inclusive category of 'non-agricultural activity' are also reported where available, although these are less satisfactory for cross-national comparison due to variations in the range of industries included. Even within countries, series breaks in data collection may limit the accuracy of comparisons over time. Some caution is therefore needed in interpreting cross-national differences and trends. Nevertheless, a number of broad observations can be made on the basis of the data presented in Table 1.

Looking first at trends within countries, an increase in women's average earnings relative to men's since 1970 is evident in all of the countries listed except Japan. In many cases this has occurred mainly in the 1970s, although some countries (for example, Belgium, Canada, and the USA) show marked improvement since 1980. In several cases, however, closing of the gender pay gap has slowed in the 1990s.

The table also provides some indication of cross-national variation in the size of the gender pay gap. Figures for Canada and the USA are from different sources and not strictly comparable with data from the other countries. Among the others listed, the picture of cross-national variance appears broadly similar for manufacturing and 'non-agricultural' data. On the basis of the hourly manufacturing data, which are the most reliable for comparative purposes, Sweden stands out having the narrowest gender pay gap in the mid-1990s. The large gap in Japan reflects the use of monthly rather than hourly earnings, and thus the impact of different working hours between men and

Table 1
Women's earnings as a percentage of men's[i], selected countries and years[ii]

	Manufacturing				Non-agricultural industries[iii]			
	1970	1980	1990	1995	1970	1980	1990	1995
Australia	64	79	83	85	65	86	88	90
Belgium	68	70	75	79	67	69	75	79
Canada[iv]	—	—	—	—	60	64	68	73
Denmark	74	86	85	85[ix]	72	85	83	—
France	77[vii]	77	79	79[x]	—	79	81	81[x]
Japan[v]	45	44	41	44[ix]	51	54	50	—
Netherlands	72	75	75	75	74	78	78	76
New Zealand	66[viii]	71	74	77	72[viii]	77	81	81
Sweden	80	90	89	90	—	—	—	—
UK	58	69	68	71	60	70	76	79
USA[vi]	—	—	—	—	—	65	78	81

Sources: International Labour Office (ILO) *Yearbook of Labour Statistics*. ILO, Geneva, various issues; Statistics Canada, *Earnings of Men and Women*, Cat No 13-217 Annual, Statistics Canada, Ottawa, various issues; United States Department of Labor, Women's Bureau *Women's Earnings as a Percent of Men's*, http://www.dol.gov/dol/wb/public/wb_pubs/7996.htm (accessed October 10, 1999).
i. Percentages are based on hourly earnings except for Canada and Japan. ii. Countries have been selected on the basis of data availability and to provide a span across continents. iii. ILO 'non-agricultural' activity groups. Data for Canada and the USA are from different sources and are based on all industries. iv. For Canada, percentages are based on yearly earnings for full-year, full-time workers. The use of yearly wages deflates the figures as women work fewer hours than men on average, even when the comparison is limited to full-time workers. v. For Japan, percentages are based on monthly earnings. vi. For the USA, percentages are based on hourly earnings, but these are for workers paid an hourly wage and are not directly comparable with hourly data for other countries. vii. 1972. viii. 1974. ix. 1992. x. 1993 — Data unavailable.

women. However, OECD approximations of hourly earnings in manufacturing in Japan still show a very large pay gap (OECD 1988, p. 212), and ILO data indicate that on this issue Japan looks more like other East Asian countries (for example, South Korea and Singapore) than the other countries listed in the table.

Overall, the data raise several questions. Why is there a gender pay gap? What accounts for its variation over time and across nations? What can, or should, be done to eradicate it? These questions are addressed in the following sections which examine possible explanations for the gender pay gap, the effect of institutional factors on its variation across countries, and the strategies most likely to assist in its elimination.

2. *Explaining the Gender Pay Gap*

One of the most direct ways in which sex differences in pay have arisen historically is through the establishment of different rates of pay for men and women in wage-setting processes. The assumption that men, as 'breadwinners,' should be entitled to higher pay than women underpinned pay determination in many countries until well into the twentieth century. In some countries this was institutionalized in the form of a 'family wage' for men. While early challenges to these assumptions were made in several countries, equal pay rates for men and women were usually only achieved where men's jobs were threatened by cheaper female labor (see, for example, Ryan and Conlon 1989, pp. 99–100). In most countries it was not until the 1960s and 1970s, following the growth of second wave feminism and increasing involvement of women in paid employment, that the inequity of different rates of pay for men and women was responded to directly through policies of equal pay for equal work.

However, much of the pay difference between men and women results not from different pay rates in the same jobs, but from the location of men and women in different jobs. This sex segregation of the labor market has proved remarkably resistant to change, and in Walby's (1988) view, accounts for the persistence of the gender pay gap since World War II in spite of the increasing human capital (i.e., education and labor market experience) of women over that time period (see *Sex Segregation at Work*).

Segregation is significant for pay inequality to the extent that female-dominated sectors of the labor market deliver lower pay. Female-dominated occupations, for example, may be low-skilled or involve skills that have been undervalued. They may also be less likely to provide discretionary payments such as overtime or bonuses. Empirical studies demonstrate that the proportion of women in an occupational group is negatively associated with wage levels, with typically around one-third of the gender pay gap shown to be due to occupational segregation by sex (Treiman and Hartmann 1981). Results for this type of analysis are highly dependent on the level of disaggregation of occupations applied. Finer disaggregation uncovers greater inequality, as vertical segregation exists within broad occupational groups, with men more likely to be in the higher status, higher paid, jobs.

Concentration of women in a small number of occupational groups is also significant for pay outcomes. Grimshaw and Rubery (1997), for example, demonstrate that in seven OECD countries around 60 percent of women are concentrated in just 10 occupational groups out of a total of between 50 and 80, with little change in this situation since the mid-1980s. Moreover, their analysis shows a wage penalty associated with this concentration of employment. Women within these occupations, on average, earn less than the all-occupation average.

Another form of labor market division potentially affecting pay differences is the distribution of men and women between firms. Blau (1977) has shown that in the USA women and men are differently distributed among firms, with women more likely to be employed in comparatively low-paying firms. These sorts of divisions may contribute to sex differences in earnings within occupations.

A further type of division that has become of increasing significance is that between full-time permanent and other, less regular, types of employment. Part-time work is highly female dominated in most countries, and on average tends to be less well remunerated per hour than full-time work. Waldfogel (1997, p. 215), for example, shows that part-time work carries a wage penalty for the women in her sample. Thus, although part-time positions may assist women to retain careers by facilitating the combination of work and family responsibilities, the concentration of women in this type of work may contribute to the gender pay gap.

In sum, sex differences in pay can arise from different pay rates in particular jobs or the distribution of female employment into lower paying jobs. In both cases, pay differences may be due to forms of discrimination or to non-discriminatory influences. Non-discriminatory influences include the individual attributes and choices of men and women, and although these may reflect broader social inequities, they can be distinguished from overt forms of discrimination within the labor market. Both areas are considered below.

2.1 *Individual Attributes and Choices*

A wide body of literature has addressed the extent to which the gender pay gap can be explained by individual attributes and choices. Human capital theory suggests that productivity related differences between men and women, such as education, skill, and labor market experience, may account for sex differ-

ences in earnings. Typically, analyses decompose earnings differences between men and women into a component that can be 'explained' by productivity factors, and another that represents the 'unexplained' elements of the gender pay gap that are ascribed to labor market discrimination (Oaxaca 1973). Whether human capital differences themselves might be evidence of social inequality (for example, reflecting differential access to training or restricted choices about labor force attachment) has not been an explicit part of this analytical approach.

Numerous studies have been conducted for different countries, and although there are difficulties of measurement and interpretation that complicate such analyses, human capital variables typically account for less than half, and often considerably smaller proportions, of the gender pay gap (Treiman and Hartmann 1981, p. 42). The studies which explain most have been those including detailed estimates of employment experience or labor force attachment (for example, Corcoran and Duncan 1979). Where women's more intermittent labor force attachment has been captured effectively, as in longitudinal surveys or work histories, the negative effect on wages has been clearly demonstrated (Waldfogel 1997, pp. 210–11).

Alongside intermittent labor force attachment, several studies show a wage penalty for women associated with the presence of children. In a regression model utilizing data from 1968–88, Waldfogel (1997, p. 212) identifies a penalty in hourly wages associated with having a child, and a larger penalty for two or more children, even after controlling for actual employment experience and factors such as education. The effects of family and domestic labor responsibilities are thus likely to be cumulative, lowering women's earnings through reducing employment experience and the capacity to retain career paths. While some may interpret such findings as evidence that a proportion of the gender pay gap is non-discriminatory and simply due to individual choices, others may observe that women's disproportionate responsibility for family care affects the range of choices available (see *Motherhood: Economic Aspects*).

2.2 Discrimination in Employment

There are several types of labor market discrimination with implications for sex differences in pay. For example, employers may discriminate against women when recruiting and promoting staff, preferring to hire men for higher status positions, and investing more in training and career support for men. This may be 'statistical discrimination,' reflecting assumptions that women will be more likely than men to leave work, or have lower levels of commitment to it, once they have family responsibilities. However, it is not clear that women leave jobs more frequently than men (see England 1992, p. 33), hence the rationality of this type of discrimination is questionable. Less direct forms of discrimination may be the result of customary practice, with long-standing procedures and organizational cultures effectively hindering the advancement of women (see Cockburn 1991).

Discrimination may also be evident in the way pay rates are established. While the overt use of different rates of pay for men and women discussed earlier is no longer widespread, female-dominated areas of employment may be relatively underpaid. England's (1992) analysis provides some evidence for this by showing that the sex composition of an occupation explains between 5 percent and 11 percent of the gender pay gap even when factors such as different types of demands for skill and effort, and industrial and organizational characteristics are controlled for. Undervaluation could be the result of women's comparatively low bargaining power, and could also reflect gender-biased estimations of the value of skills in female-dominated occupations. England (1992), for example, shows that 'nurturing' skills carry a wage penalty, thus suggesting that this type of work is devalued because of its association with typical 'women's work.' Such findings indicate that policies of 'comparable worth' or 'equal pay for work of equal value'—that is, comparisons of dissimilar jobs to produce estimations of job value free of gender-bias—have an important role to play.

3. The Role of Institutions

While the factors considered thus far contribute to an understanding of the gender pay gap in a general sense, they are frequently less useful in explaining differences between countries. For example, differences between countries in the level of occupational segregation appear to be unrelated to performance on sex differences in pay. In Japan, the combination of low levels of occupational segregation and a large gender pay gap may be explained partly by women's relative lack of access to seniority track positions in large firms (Anker 1997, p. 335)—that is, by other types of segregation. However, a different type of explanation is necessary to understand why some countries that are highly sex-segregated by occupation (such as Sweden and Australia) have comparatively narrow gender pay gaps. This anomaly suggests the importance of institutional factors in influencing the gender pay gap.

Blau and Kahn (1992), for example, point out that the effect of occupational concentration on sex differences in pay will be influenced by the overall wage distribution in any country—where wages are relatively compressed, the effect of segregation on earnings will be minimized. Wage distribution is in turn likely to be a product of the institutional framework for wage bargaining, with more centralized and regulated systems conducive to lower levels of wage dispersion,

and, therefore, a narrower gender pay gap. Cross-national statistical evidence supports these links, showing association between centralized wage fixation and high levels of pay equity (Whitehouse 1992, Gunderson 1994). This relationship is likely to result not only from wage compression, but also from an enhanced capacity to implement equal pay measures in more centralized systems. Rubery (1994) notes that the degree of centralization has implications for several matters of relevance to pay equity outcomes, including the maintenance of minimum standards and the scope for equal value comparisons. Decentralized pay systems tend to provide more limited scope for comparisons to support equal pay for work of equal value cases, and results may be limited to specific enterprises, or to individuals. Overall, trends in wage bargaining arrangements may be more influential on pay equity outcomes than specific gender equity measures (Rubery 1994, Rubery et al. 1997, Whitehouse 1992). Wage bargaining systems are, however, country-specific, and embedded in national employment systems (Rubery et al. 1997). Translation of institutional structures across countries is therefore unlikely to be a viable proposition.

4. *Policy options*

As the gender pay gap cannot be attributed to one type of cause, a number of strategies will be necessary to attempt its elimination. While addressing pre-market impediments to women's advancement (such as sex-role stereotypes and their impact on educational and employment choices) is part of this agenda, the main strategies are those designed to remove barriers within the labor market. International conventions such as the ILO's Equal Remuneration Convention (No. 100) and the United Nations' Convention on Elimination of all forms of Discrimination Against Women have provided some impetus for action, and most countries have now implemented some form of equal pay legislation or prohibition against discrimination in employment. However, the efficacy of such measures is highly variable.

The most direct strategies are legislative provisions requiring the payment of equal pay for equal work, and equal pay for work of equal value. Equal pay for equal work provisions have been most effective where they have included a requirement to remove differential pay rates for men and women in industrial agreements. The rapid narrowing of the gender pay gap in Australia and the UK in the 1970s, for example, demonstrates the effectiveness of measures that removed direct discrimination in collective agreements (Gregory et al. 1989, Zabalza and Tzannatos 1985). Improvements in Sweden in the same decade also reflect the advantage of widespread coverage by collective agreements, and in that case predate the introduction of equal pay legislation. However, while equal pay requirements may be quite effective initially where they equalize minimum rates between men and women across a comprehensive set of collective agreements, the level of segregation in the labor market means that most provisions for equal pay for equal work apply to only a small proportion of women, as few men and women are in fact doing the same work.

Provisions for equal pay for work of equal value (comparable worth) open up a wider field for contestation, but this strategy has proved difficult to implement. Historical bias in the way 'female' jobs are valued has not been easy to eradicate as even quite detailed job evaluation methods retain aspects of gender bias and may in fact perpetuate existing hierarchies (Steinberg 1992). Moreover, cases have proved complex and time consuming. Comparable worth does permit reconsideration of the valuation of work, however, and is particularly important given the apparent resistance of patterns of occupational segregation to change. It will be most effective where the scope for comparison is wide, and results apply collectively to types of jobs rather than to individuals.

Apart from strategies that deal directly with pay, provisions to prohibit discrimination in employment have been adopted in most countries. Anti-discrimination legislation and equal employment opportunity or affirmative action measures aim to prevent sex discrimination in hiring and placement, and in some cases seek to correct for past sex discrimination by requiring attention to the unequal distribution of men and women within organizations. While it is difficult to estimate the impact of such measures on the gender pay gap, they have no doubt restricted the scope for overt discrimination and contributed to a gradual change in customs and attitudes.

Given the impact of family responsibilities on women's earnings noted earlier, erosion of the gender pay gap will also require the use of strategies to assist in the combination of work and family responsibilities. Ultimately this also requires a more even division of domestic labor between men and women to assist women to retain career progression and employment experience. Paid parental and family leave, and measures to deliver flexibility with job security while combining employment and caring responsibilities will be part of this agenda, although experience thus far suggests that encouragement to fathers to share these types of provisions will also be necessary.

Finally, it must be emphasized that the gender pay gap is dependent on a wide range of policy and institutional factors, most of which are not designed with gender equity goals in mind. In particular, wage bargaining arrangements and employment policies may affect the size of the gender pay gap by affecting outcomes such as wage dispersion. Trends away from centralized and regulated forms of pay bargaining are therefore of some concern as they may increase dispersion, which in turn may erode gains made

through equal pay or comparable worth strategies. In short, the pursuit of pay equity cannot be limited to a single agenda, but requires multiple policy measures, and—like most endeavors attempting significant social change—will require a long period of time to achieve.

Bibliography

Anker R 1997 Theories of occupational segregation by sex: An overview. *International Labour Review* **136**: 315–39

Blau F 1977 *Equal Pay in the Office*. DC Heath and Company, Lexington, MA

Blau F, Kahn L 1992 The gender earnings gap: Learning from international comparisons. *American Economic Review, Papers and Proceedings* **82**: 533–8

Cockburn C 1991 *In the Way of Women: Men's Resistance to Sex Equality in Organizations*. ILR Press, Ithaca, NY

Corcoran M, Duncan G 1979 Work history, labor force attachments, and earnings differences between the races and sexes. *Journal of Human Resources* **14**: 3–20

England P 1992 *Comparable Worth: Theories and Evidence*. Aldine de Gruyter, New York

Gregory R G, Anstie R, Daly A, Ho V 1989 Women's pay in Australia, Great Britain and the United States: the role of laws, regulations, and human capital. In: Michael R T, Hartmann H I, O'Farrell B (eds.) *Pay Equity: Empirical Inquiries*, National Academy Press, Washington, DC, pp. 222–42

Grimshaw D, Rubery J 1997 *The Concentration of Women's Employment and Relative Occupational Pay: a Statistical Framework for Comparative Analysis*. Labour Market and Social Policy, Occasional Paper No 26, Organisation for Economic Cooperation and Development, Paris

Gunderson M 1994 *Comparable Worth and Gender Discrimination: an International Perspective*. International Labour Office, Geneva, Switzerland

Oaxaca R 1973 Male–female wage differentials in urban labor markets. *International Economic Review* **14**: 693–709

OECD 1988 *Employment Outlook*. Organization for economic cooperation and development, Paris

Rubery J 1994 Decentralisation and individualisation: The implications for equal pay. *Economies et Sociétés* **18**: 79–97

Rubery J, Bettio F, Fagan C, Maier F, Quack S, Villa P 1997 Payment structures and gender pay differentials: Some societal effects. *The International Journal of Human Resource Management* **8**: 131–49

Ryan E, Conlan A 1989 *Gentle Invaders: Australian Women at Work*. Penguin, Ringwood, Victoria

Steinberg R J 1992 Gendered instructions. Cultural lag and gender bias in the Hay system of job evaluation. *Work and Occupations* **19**: 387–423

Treiman D J, Hartmann H I (eds.) 1981 *Women, Work and Wages: Equal Pay for Jobs of Equal Value*. National Academy Press, Washington, DC

Walby S 1988 Introduction. In: Walby S (ed.) *Gender Segregation at Work*. Open University Press, Milton Keynes, UK

Waldfogel J 1997 The effect of children on women's wages. *American Sociological Review* **62**: 209–17

Whitehouse G 1992 Legislation and labour market gender inequality: An analysis of OECD countries. *Work, Employment & Society* **6**: 65–86.

Zabalza A, Tzannatos Z 1985 *Women and Equal Pay: The Effect of Legislation on Female Employment and Wages in Britain*. Cambridge University Press, Cambridge, UK

G. Whitehouse

Sex Hormones and their Brain Receptors

Nerve cells react not only to electrical and chemical signals from other neurons, but also to a variety of steroid factors arising from outside the brain. Steroids, particularly gonadal steroids, have profound effects on brain development, sexual differentiation, central nervous control of puberty, the stress response, and many functions in the mature brain including cognition and memory (Brinton 1998, McEwen 1997).

The brain contains receptors for all five classes of steroid hormones: estrogens, progestins, androgens, glucocorticoids, and mineralocorticoids. Steroid hormone receptors are not uniformly distributed, but occupy specific loci within the brain (McEwen 1999, Shughrue et al. 1997). As expected the hypothalamus has a rich abundance of steroid receptors, particularly receptors for the sex hormones estrogen, progesterone, and testosterone. Remarkably, the limbic system is also a site for steroid action and contains estrogen, androgen, and glucocorticoid receptors. The cerebral cortex is also a site of steroid action and expresses receptors for estrogens and glucocorticoids. Some of these receptors arise from separate genes while others arise from alternative splicing of a single gene. For example the progesterone receptor exists in multiple isoforms derived from the alternative splicing of one gene (Whitfield et al. 1999) whereas estrogen receptors have multiple receptor subtypes (ERα and ERβ) that originate from separate genes, as well as multiple isoforms (Warner et al. 1999, Keightley 1998).

1. Structure of Steroid Receptors

All genomic steroid hormone receptors are composed of at least three domains: the N-terminal (or A/B) region, the DNA binding domain containing two zinc-fingers (region C), and the ligand-binding domain to which the hormone binds (region E). Select steroid receptors, such as the estrogen receptors, have an additional C-terminal (or F) region, which serves to modify receptor function (Tsai and O'Malley 1994, Warner et al. 1999) (Fig. 1A).

The amino-terminal A/B region of steroid receptors is highly variable and contains a transactivation domain, which interacts with components of the core transcriptional complex. Region C of these receptors contains a core sequence of 66 amino acids that is the most highly conserved region of the nuclear hormone receptor family. This region folds into two structures that bind zinc in a way characteristic of type II zinc

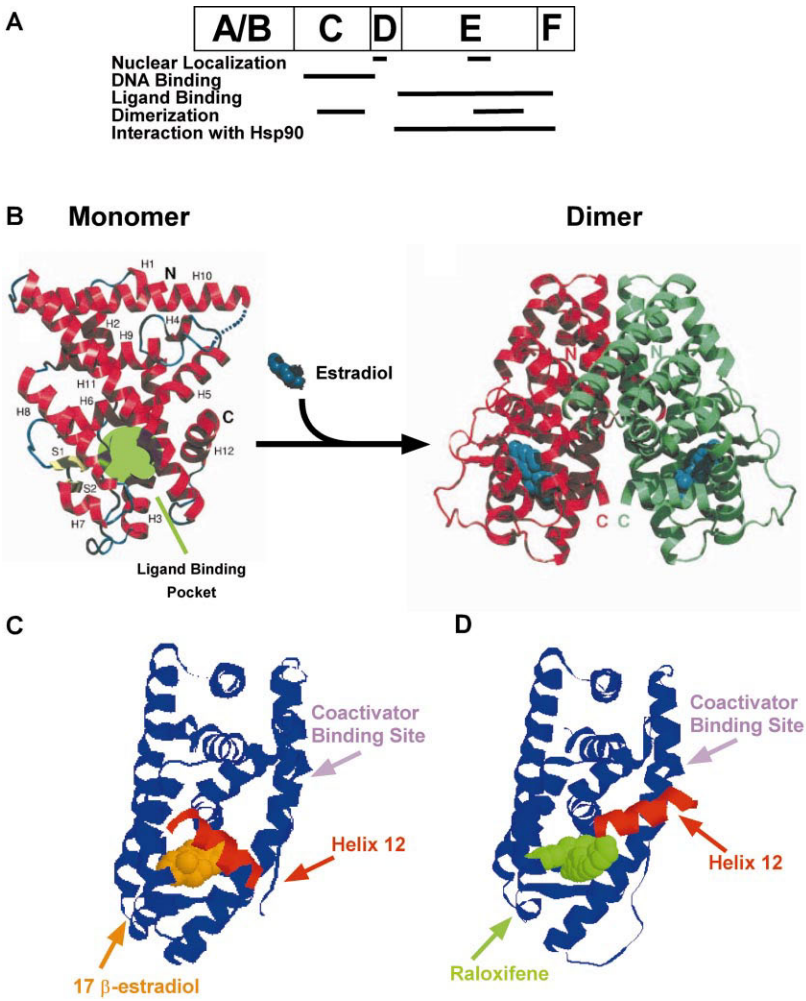

Figure 1
Structural aspects of Steroid Receptors. A. The steroid receptors can be divided into 6 functional domains A, B, C, D, E, and F. The function of each domain is indicated by the solid lines. B. Ribbon representation of the ligand binding domain of estrogen receptor α indicating the ligand binding pocket. Upon ligand binding the receptor dimerizes and then can act as a transcription factor (adapted from Brzozowski et al. 1997). C. Ribbon representation of ERα demonstrating the conformational alterations induced by 17β-estradiol. Upon binding estradiol, helix 12 folds to cover the ligand binding pocket leaving the coactivator binding site free. D. Ribbon representation of ERα demonstrating the conformational alterations induced by raloxifene. Upon binding raloxifene, helix 12 is displaced obscuring the coactivator binding site (adapted from Brzozowski et al. 1997)

fingers. These two type II zinc clusters are believed to be involved in specific DNA binding, thus giving the C regions its designation as the DNA binding domain. In addition, the C region plays a role in receptor dimerization (Fig. 1A and B). The D region contains a hinge portion that may be conformationally altered upon ligand binding. The E/F-region is functionally complex since it contains regions important for ligand binding and receptor dimerization, nuclear localization, and interactions with transcriptional co-activators and corepressors (Horwitz et al. 1996). It has been observed that a truncated receptor lacking regions A/B and C can bind hormone with an affinity similar to that of the complete receptor suggesting that region E is an independently folded structural domain. Amino acids within region E that are conserved amongst all members of the nuclear receptor family are thought to be responsible for the conformation of a hydrophobic pocket necessary for steroid binding whereas the non-conserved amino acids may be

important for ligand specificity. The conserved amino acid sequences within the ligand binding domain of the nuclear receptors make up a common structural motif, which is composed of 11–12 individual α helices with helix 12 being absent in some members of the superfamily of steroid receptors. Although domain F is not required for transcriptional response to hormone, it appears to be important for modulating receptor function in response to different ligands.

Of prime importance in estrogen receptor function is helix 12 (Fig. 1C), found in the most carboxyl region of the ligand-binding domain, which is thought to be responsible for transcriptional activation function (AF-2) activity (Warner et al. 1999). This region realigns over the ligand-binding pocket when associated with agonists, but takes on a different conformation with antagonists (Fig. 1). Such conformational changes are thought to affect interactions with coactivators, since helix 12 is required for interaction with several of the coactivator proteins and mutations that are known to abolish AF-2 transcriptional activity also abolish the interaction of the nuclear receptors with several of their associated proteins (McKenna et al. 1999, Warner et al. 1999). The ability of different ligands to determine the conformational state of the receptor has profound consequences on the ability of the steroid receptor to drive gene transcription.

2. *Classical Mechanism of Sex Steroid Action*

Classically, sex steroids exert their effects by binding to intracellular receptors and modifying gene expression (Tsai and O'Malley 1994). Steroid receptors are ligand-activated transcription factors that modulate specific gene expression. In the presence of hormone, two receptor monomers dimerize and bind to short DNA sequences located in the vicinity of hormone-regulated genes (Fig. 2). These specific DNA sequences, or hormone response elements (HRE), contain palindromic or directly repeating half-sites (Tsai and O'Malley 1994). Since an HRE exerts its action irrespective of its orientation and when positioned at a variable distance upstream or downstream from a variety of promoters (Tsai and O'Malley 1994), it is an enhancer. Steroid receptors represent inducible enhancer factors that contain regions important for hormone binding, HRE recognition, and activation of transcription.

The mechanism of transactivation by nuclear receptors has recently achieved further complexity by the discovery of an increasing number of coregulators. Coregulator proteins can modulate the efficacy and direction of steroid-induced gene transcription. Coregulators include coactivators ERAP160 and 140, and RIP160, 140, and 80 which were biochemically identified by their ability to specifically interact with the hormone binding domain of the receptor in a ligand-dependent manner (McKenna et al. 1999) (Fig. 2). For example, the interaction between estrogen receptors and coregulators was promoted by estradiol whereas antiestrogens did not promote this interaction. Further studies led to the identification of many other co-activators including glucocorticoid receptor interacting protein 1 (GRIP 1), steroid receptor coactivator 1 (SRC 1), and transcriptional intermediary factor 2 (TIF 2). When cotransfected with nuclear receptors, including estrogen receptor, these coactivators are capable of augmenting ligand-dependent transactivation. In addition, the phospho-CREB-binding protein (CBP) and the related p300 have been demonstrated to be estrogen receptor-associated proteins and involved in ligand-dependent transactivation. It has also been shown that co-activator RIP140 interacts with the ER in the presence of estrogen, and this interaction enhances transcriptional activity between 4- to 100-fold, depending on promoter context.

3. *Rapid Signaling Effects of Sex Steroids*

Although steroids induce many of their effects in the brain through activation of genomic receptors, non-transcriptional actions of estrogens, progestins, glucocorticoids, and aldosterone have been observed in a variety of tissues, including the brain (Brinton 1993, 1994, Watson et al. 1995, Pappas et al. 1995). These nontranscriptional actions are characterized by short-term effects that range from milliseconds to several minutes. Additionally, these effects still occur in the presence of actinomycin D or cyclohexamide, known transcriptional blockers. The first suggestion of the non-transcriptional actions of steroids came when it was observed that progesterone induced rapid anesthetic and sedative actions when injected subcutaneously (Brinton 1994). Then in the 1960s and 1970s numerous studies suggested that estrogen modulated the electrical activity of a variety of nerve cells (Foy et al. 1999). Subsequent studies utilizing various techniques and preparations have shown that these effects occur too rapidly to be genomic in nature (Foy and Teyler 1983).

Morphological studies of neuronal development conducted *in vitro* have shown that estrogenic steroids exert a growth-promoting, neurotrophic effect on hippocampal and cortical neurons via a mechanism that requires activation of NMDA receptors (Brinton et al. 1997). *In vivo* studies have revealed a proliferation of dendritic spines following 17β-estradiol treatment that can be prevented by blockade of NMDA receptor-channels, though not by AMPA or muscarinic receptor antagonists (Woolley 1999). Other reports have provided evidence that chronic 17β-estradiol treatment increases the number of NMDA receptor binding sites, and NMDA receptor-mediated responses (Woolley 1999).

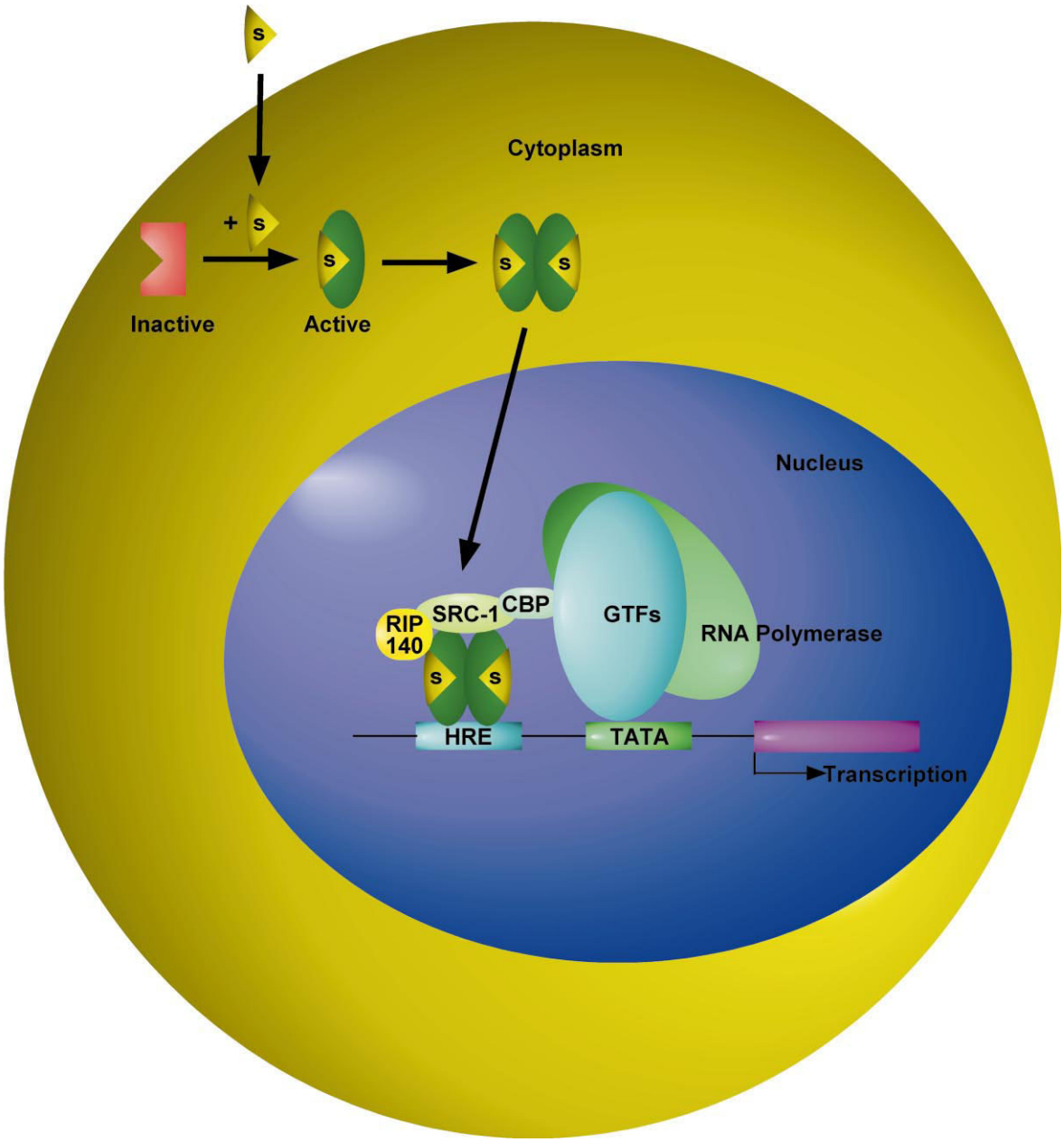

Figure 2
Classical mechanism of steroid action. Upon binding of steroid (S) to an inactive steroid receptor, the receptor is activated and two receptor-ligand monomers dimerize and bind to the hormone response element (HRE). Coactivators such as RIP140, CBP, and SRC-1 bind to and link the hormone receptor with the general transcription factors (GTF) and RNA polymerase of the transcription machinery to alter transcription

Recent studies suggest a direct link between the estrogen receptor and the mitogen-activated protein kinase (MAPK) signaling cascade (Singh et al. 2000). MAPKs are a family of serine-threonine kinases that become phosphorylated and activated in response to a variety of cell growth signals. In neuronal cells, estrogen resulted in neuroprotection that was associated with a rapid activation of the MAPK

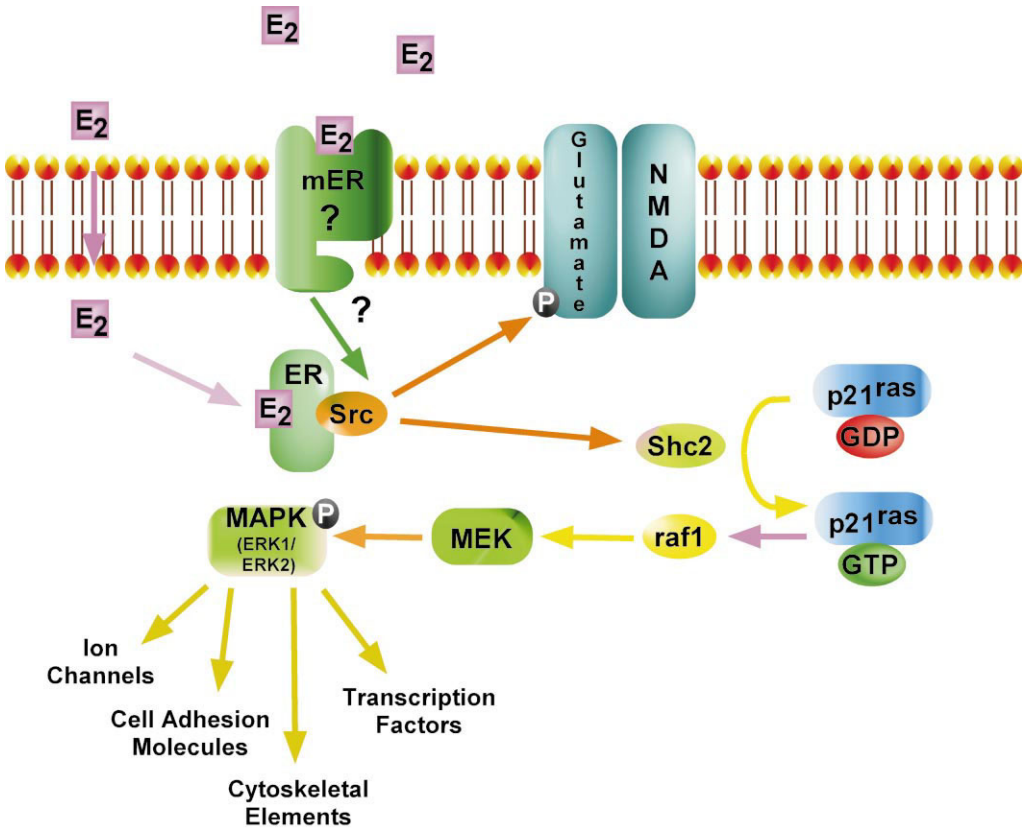

Figure 3
Rapid signaling effects of sex steroids. Estradiol, acting through a possible membrane bound or cytoplasmic receptor, causes the transient activation of c-src-tyrosine kinase. Activated c-src can potentiate the glutamate response through the NMDA receptor. Additionally, c-src can cause the phosphorylation of p21 (ras)-guanine nucleotide activating protein leading to the activation of MAP Kinase, which has many downstream effects on cell survival and growth

signaling pathway (Singh et al. 2000). These neuroprotective effects, which occurred within 5 minutes of estrogen exposure are thought to occur through the transient activation of c-src-tyrosine kinases and tyrosine phosphorylation of p21(ras)-guanine nucleotide activating protein (Fig. 3). Additionally, the potentiation of the NMDA receptor mediated neuronal response by estrogen is thought to be mediated by c-src-tyrosine kinases (Bi et al. 2000). It is not yet clear whether these effects require the classical estrogen receptor (ERα/β) or if there is an as of yet undiscovered membrane receptor that mediates these rapid effects of estrogen (Razandi et al. 1999). Thus, while steroid receptors are ligand-induced transcriptional enhancers, they are also activators of intracellular signaling pathways that can profoundly influence neuronal function and survival.

Challenges remain in our understanding of steroid action and sites of action in brain. Remarkably, our knowledge of genomic sites of steroid action has greatly expanded in the past decade while the membrane sites of steroid action remain a challenge to fully characterize. Our understanding of steroid effects in brain has led to the expansion of the role of steroid receptors beyond sexual differentiation and reproductive neuroendrocrine function to regulators of nearly every aspect of brain function including cognition. The full range of mechanisms by which steroids influence such a wide array of brain functions remains to be discovered.

See also: Gender Differences in Personality and Social Behavior; Gender-related Development

Bibliography

Bi R, Broutman G, Foy M R, Thompson R F, Baudry M 2000 The tyrosine kinase and mitogen-activated protein kinase pathways mediate multiple effects of estrogen in hippocampus.

Proceedings of the National Academy of Science, USA **97**: 3602–7
Brinton R D 1993 17 β-estradiol induction of filopodial growth in cultured hippocampal neurons within minutes of exposure. *Molecular Cell Neuroscience* **4**: 36–46
Brinton R D 1994 The neurosteroid 3 alpha-hydroxy-5 alpha-pregnan-20-one induces cytoarchitectural regression in cultured fetal hippocampal neurons. *Journal of Neuroscience* **14**: 2763–74
Brinton R D 1998 Estrogens and Alzheimer's disease. In: Marwah J, Teitelbaum H (eds.) *Advances in Neurodegenerative Disorders, Vol. 2: Alzheimer's and Aging*. Prominent Press, Scottsdale, AZ, pp. 99
Brinton R D, Proffitt P, Tran J, Luu R 1997 Equilin, a principal component of the estrogen replacement therapy premarin, increases the growth of cortical neurons via an NMDA receptor-dependent mechanism. *Experimental Neurology* **147**: 211–20
Foy M R, Teyler T J 1983 17-alpha-Estradiol and 17-beta-estradiol in hippocampus. *Brain Research Bulletin* **10**: 735–9
Foy M R, Xu J, Xie X, Brinton R D, Thompson R F, Berger T W 1999 17 β-estradiol enhances NmDA receptor mediated EPSPs and long-term potentiation in hippocamfal CA1 cells. *Journal of Neurophysiology* **81**: 925–8
Horwitz K B, Jackson T A, Bain D L, Richer J K, Takimoto G S, Tung L 1996 Nuclear receptor coactivators and co-repressors. *Molecular Endocrinology* **10**: 1167–77
Keightley M C 1998 Steroid receptor isoforms: Exception or rule? *Molecular & Cellular Endocrinology* **137**: 1–5
McEwen B S 1997 Hormones as regulators of brain development: Life-long effects related to health and disease. *Acta Paediatrica* **422** (Suppl.): 41–4
McEwen B S 1999 Clinical review 108: The molecular and neuroanatomical basis for estrogen effects in the central nervous system. *Journal of Clinical Endocrinology & Metabolism* **84**: 1790–7
McKenna N J, Lanz R B, O'Malley B W 1999 Nuclear receptor coregulators: Cellular and molecular biology. *Endocrine Review* **20**: 321–44
Pappas T C, Gametchu B, Watson C S 1995 Membrane estrogen receptors identified by multiple antibody labeling and impeded-ligand binding. *FASEB* **9**: 404–10
Razandi M, Pedram A, Greene G L, Levin E R 1999 Cell membrane and nuclear estrogen receptors (ERs) originate from a single transcript: Studies of ERalpha and ERbeta expressed in Chinese hamster ovary cells. *Molecular Endocrinology* **13**: 307–19
Shughrue P J, Lane M V, Merchenthaler I 1997 Comparative distribution of estrogen receptor-alpha and -beta mRNA in the rat central nervous system. *Journal of Comprehensive Neurology* **388**: 507–25
Singh M, Setalo G J, Guan X, Frail D E, Toran-Allerand C D 2000 Estrogen-induced activation of the mitogen-activated protein kinase cascade in the cerebral cortex of estrogen receptor-alpha knock-out mice. *Journal of Neuroscience* **20**: 1694–1700
Tsai M J, O'Malley B W 1994 Molecular mechanisms of action of steroid/thyroid receptor superfamily members. *Annual Review of Biochemistry* **63**: 451–86
Warner M, Nilsson S, Gustafsson J A 1999 The estrogen receptor family [Review] [50 refs]. *Current Opinions in Obstetrics & Gynecology* **11**: 249–54
Watson C S, Pappas T C, Gametchu B 1995 The other estrogen receptor in the plasma membrane: Implications for the actions of environmental estrogens. *Environmental Health Perspectives* **103** (Suppl.): 41–50
Whitfield G K, Jurutka P W, Haussler C A, Haussler M R 1999 Steroid hormone receptors: Evolution, ligands, and molecular basis of biologic function. *Journal of Cell Biochemistry* **32– 3** (Suppl.): 110–22
Woolley C S 1999 Electrophysiological and cellular effects of estrogen on neuronal function. *Critical Review of Neurobiology* **13**: 1–20

R. D. Brinton and J. T. Nilsen

Sex Offenders, Clinical Psychology of

The clinical psychology of sex offenders involves assessment, treatment, and prevention. Clinical assessment involves the careful description of the problem and an estimation about the risk for recidivism, or re-offense. Clinical treatment involves interventions to reduce the risk of recidivism. Clinical prevention involves interventions before a person becomes a sex offender to reduce the risk of recidivism. There is much more research on assessment and treatment of sex offenders than there is on prevention of sex offending.

1. *Who are Sex Offenders?*

A sex offender is anyone who has forced another person to engage in sexual contact against their will. Sex offending may or may not involve physical force. For example, a person can use psychological force to get another person to have sex, as in the case of a power differential between two persons (e.g., employer–employee). However, coercive sex that involves physical force is more likely to be regarded as sex offending than coercive sex that does not. Another issue is the ability of the victim to give consent. Minors and developmentally disabled persons are usually considered to lack the ability to give consent. A person's ability to give consent may also become impaired, such as in the case of substance abuse. Thus, a person engaging in sex with a person who is unable to consent or whose ability to give consent is impaired might be considered a sex offender.

The term sex offender usually is associated with a person who has been apprehended in a legal context for their coercive sexual behavior. Nevertheless, not all persons who engage in sexually coercive behaviors are caught. Thus, two persons could engage in the same sexually coercive behavior, but the one who is caught would be considered a sex offender and the one who is not caught would not. Moreover, legal statutes

vary from jurisdiction to jurisdiction. For example, sexual penetration may be required to be considered a sex offender in some jurisdictions, while other jurisdictions may have broader definitions that include nonpenetrative forms of sexual contact. Legal jargon may obscure the meaning and impact of sexually coercive behavior, as well. For example, terms such as 'indecent liberties' and 'indecent assault' may both refer to rape, as defined as forced sexual intercourse. However, indecent liberties or indecent assault charges carry very different legal meanings and punishments than does rape. Similarly, rape committed by a stranger is more likely to be prosecuted than rape committed by someone known to the victim. Nevertheless, the behavior in both instances is rape. Because of these inconsistencies and vagaries in legal definitions of sex offending, the current article will adopt a broad approach to sexual aggression, focusing on coercive sexual behavior whether or not the perpetrator has been apprehended.

There is often disagreement between perpetrators and victims on whether sexual aggression actually occurred, or the seriousness of it. A comprehensive discussion of the veracity of perpetrator and victim reports is beyond the scope of this article. However, current definitions of sexual aggression give more weight to victims' perceptions of the occurrence of sexual aggression, given perpetrators' tendencies to defend and minimize aggressive behavior and the inherent power differential between victims and offenders. It is usually disadvantageous to bring the negative attention to oneself that accompanies accusing someone of sexually victimizing them. In courtroom settings, victims' personal and sexual histories are examined in cases of sexual abuse in a manner that is unparalleled for other types of offenses. Nevertheless, child custody disputes appear to be one context in which there may be some risk, albeit small, for false accusations of sexual abuse on the part of a parent seeking to disqualify the parenting fitness of the other parent.

The Diagnostic and Statistical Manual (*DSM*) of the American Psychiatric Association includes several sexual disorders, or paraphilias, that sex offenders may have. These include fetishism (sexual arousal associated with nonliving objects), transvestic fetishism (sexual arousal associated with the act of cross-dressing), voyeurism (observing unsuspecting nude individuals in the process of disrobing or engaging in sexual activity), exhibitionism (public genital exposure), frotteurism (touching a nonconsenting person's genitalia or breasts, or rubbing one's genitals against a nonconsenting person's thighs or buttocks), pedophilia (sexual attraction or contact involving children), sexual masochism (sexual arousal associated with suffering), and sexual sadism (sexual arousal associated with inflicting suffering). Although the fetish disorders do not involve contact with an actual person, these disorders can result in arrest when a fetishist steals the fetish items (e.g., undergarments). Rape often does not involve sexual arousal directly associated with suffering and is more commonly classified in *DSM* as a component of antisocial personality disorder than as a sexual disorder. Such a classification appears justified, in that many rapists engage in both sexual and nonsexual forms of aggression and other rule-violating behaviors. The focus of this article will be on rape and child molestation because there exists more research on these topics than on any of the other types of sexual offenses. The emphasis in the literature on these two sex offenses reflects the fact that these may create more harm to victims than the other disorders.

2. *History*

The clinical study of sex offenders had its beginnings in research on sexuality. Some of the earliest scholarly writings on sexual deviance were detailed case studies by psychiatrist Krafft-Ebing (1965/1886). Krafft-Ebing postulated that all sexual deviations were the result of masturbation. The case study method focused exclusively on highly disturbed individuals without matched control cases, which precluded objective considerations of etiology (Rosen and Beck 1988).

Kinsey and colleagues (1948) conducted large-scale normative surveys of sexual behavior, resulting in major works on male and female sexuality. Because adult–child sexual contact was relatively common in his samples, Kinsey underplayed the negative effects of such behavior (Rosen and Beck 1988). Work on sexual deviance continued at the Kinsey Institute after Kinsey's death (Gebhard et al. 1965).

A major advance in the assessment of sexual arousal was the development of the laboratory measures of penile response by Freund (1963). Freund's measure, known as the penile plethysmograph, involved an inflatable tube constructed from a condom, by which penile volume change in response to erotic stimuli (e.g., nude photographs) was measured by air displacement (Rosen and Beck 1988). Less intrusive penile measures to assess circumference changes were later developed (Bancroft et al. 1966; Barlow et al. 1970). Among behaviorists, penile response to deviant stimuli (e.g., children, rape) became virtually a gold standard of measurement for sexual deviance (e.g., Abel et al. 1977).

The emphasis of behaviorists on the role of sexual arousal in sex offending came under criticism from feminist theories. Rape was conceptualized as a 'pseudosexual' act of anger and violence rather than as a sexual disorder. This approach was popularized by Groth (1979). More recent conceptualizations have incorporated sexual, affective, cognitive, and developmental motivational components of sex offending (Hall et al. 1991).

3. *Risk Factors for Sex Offending*

One major risk factor for being a sex offender is being male. Less than 1 percent of females perpetrate any form of sexual aggression, whereas the percentage of men who are rapists ranges from 7 to 10 percent, and the percentage of men who admit to sexual contact with children is 3 to 4 percent. Evolutionary psychologists suggest that mating with multiple partners provides males with a reproductive advantage. Thus, sexually aggressive behavior may be a by-product of evolutionary history. Nevertheless, most men are not sexually aggressive. Thus, being a male may be a risk factor for being sexually aggressive, but certain environmental conditions may be necessary for someone to become sexually aggressive. For example, aggressive behavior, including sexually aggressive behavior, is accepted and socialized among males more than females in most societies.

Another risk factor for males becoming sexually aggressive, particularly against children, is personal sexual victimization. Boys who have been sexually victimized are more likely to engage in sexualized behaviors (e.g., sexual touching) immediately following sexual victimization than boys who have not. Moreover, in recent studies, over half of adult sex offenders have reported being sexually abused during childhood. However, the sexual abuse of males is not invariably associated with becoming sexually abusive, as the majority of males who are sexually abused do not become abusers.

The single best predictor of sex offending is past sex offending. Persons who have sexually offended multiple times in the past are more likely to sexually offend in the future than those who have limited histories of sex offending. Recidivism rates for child molesters and rapists are similar at 25 to 32 percent. Persons who have committed multiple sex offenses have broken the barriers against offending and have lowered the threshold for re-offending by developing a pattern of behavior. Thus, sex offenders having multiple offenses are the group that poses the greatest risk to community safety. Moreover, the predictive utility of past sex offending suggests that interventions with first-time offenders and interventions to prevent sex offending from starting are critical to avoid the establishment of an ingrained pattern of behavior.

4. *Motivational Factors for Sexual Offending*

What causes men to become sex offenders? One common explanation has been the sexual preference hypothesis that sex offenders are more sexually aroused by coercive than by consenting sexual activity. Learning theories posit that sexual arousal to coercive sexual activity is conditioned (see *Classical Conditioning and Clinical Psychology*). For example, a pedophile may have had sexually arousing experiences during childhood with peers and may never have outgrown these experiences. The fusion of violent and sexual images in the media may influence some men to associate sexual arousal with aggressive activity. The sexual preference hypothesis appears to be most appropriate for some sex offenders against children, particularly those who offend against boys. For many men who molest children, their sexual arousal to children exceeds their sexual arousal to adults, as assessed by genital measures. However, some child molesters, particularly incestuous offenders, may be more sexually aroused by adults than by children. The sexual preference hypothesis is less applicable to men who rape adults. Most rapists' sexual arousal is greatest in response to adult consenting sexual activity and less in response to sexual activity that involves physical force. However, a minority of rapists do exhibit a sexual preference for rape over consenting adult sexual activity. These rapists would be considered sexual sadists.

Another common explanation of sex offending has been anger and power. Some feminist scholars have gone as far as to say that sex offending is a 'pseudo-sexual' act and that it is not about sex. However, recent feminist scholarship recognizes the sexual and aggressive aspects of sex offending. If sex offending is a violent, rather than a sexual, act, why does the perpetrator not simply assault the victim rather than sexually assaulting the victim? Anger and power are most relevant in explaining the rape of women by males. Adversarial relationships with women may cause some men to attempt to enforce their sense of superiority over women via rape. Some child molesters may also have anger and power motives for their offending. However, depression is a more common precursor to child molesting than anger. One source of depression may be perceived or actual social incompetence in peer relationships. Sexual contact with children may represent a maladaptive method of coping with this depression.

Excuses to justify sexually aggressive behavior may be the motivation for some forms of sex offending. Such excuses deny or minimize the impact of sexual aggression and are known as cognitive distortions. Cognitive distortions are common in acquaintance rape situations and in incest. An acquaintance rapist may contend that rape cannot occur in the context of a relationship or that the existence of a relationship justifies any type of sexual contact that the rapist desires. Common cognitive distortions among incest offenders are that sexual contact with their child is a form of affection, or sexual education, or that it merely amounts to horseplay. Thus, the motivation for a person employing cognitive distortions is that the sexual contact is 'normal' and is not aggressive.

It is likely that these different types of motivation for sex offending may interact for many sex offenders. Sexual arousal to coercive sexual behavior, emotional problems, and cognitive distortions may coexist for

some sex offenders, and a singular explanation of the basis of the problem may be inadequate. Nevertheless, complex explanations of sex offending are less likely to result in treatment interventions than are explanations that identify the major motivational issues.

5. *Treatment Interventions for Sex Offenders*

The most common forms of treatment for sex offenders have been behavioral, cognitive-behavioral, and psychohormonal interventions. Behavioral methods have typically involved interventions to reduce sexual arousal to inappropriate stimuli (e.g., children, rape) (see *Behavior Therapy: Psychiatric Aspects*). Aversive conditioning, which pairs sexual arousal to the inappropriate stimulus with an aversive stimulus (e.g., foul odor, thoughts about punishment for sex offending) is perhaps the most widely used behavioral treatment with sex offenders. Cognitive-behavioral methods focus on the influence of cognitive factors, including individual beliefs, standards, and values, on sex-offending behavior (see *Cognitive Therapy*). Common cognitive-behavioral treatments often involve cognitive restructuring, empathy enhancement, social skills training, and self-control techniques. Relapse prevention, which has been adapted from the treatment of addictive behaviors, has gained relatively wide acceptance in the treatment of sex offenders. This cognitive-behavior method involves self-control via anticipating and coping with situations following treatment that create high risk for relapse (e.g., social contact with potential victims) (see *Relapse Prevention Training*). Psychohormonal treatments involve the use of antiandrogen drugs to reduce the production of testosterone and other androgens. Such androgen reduction suppresses sexual arousal. The antiandrogen that has been most commonly used is medroxyprogesterone (Depo Provera).

Evidence across recent studies of rapists, child molesters, and exhibitionists suggests that cognitive-behavioral and psychohormonal interventions may be more effective than no treatment or behavioral treatments. Recidivism rates among sex offenders, as measured by arrests for sex offenses, are 25–32 percent. Sex offender recidivism rates in studies of behavioral treatments have been approximately 40 percent. Both cognitive-behavioral and psychohormonal treatments have yielded recidivism rates of 13 percent. Why is there such a high rate of recidivism in studies of behavioral treatments? It could be contended that behavioral treatments are too narrowly focused on sexual arousal and ignore other motivational issues (e.g., emotional, cognitive). However, psychohormonal treatments also primarily focus on sexual arousal reduction and result in relatively low recidivism rates. Thus, it is possible that any positive effects of behavioral treatments may 'wear off' over time. It is also possible that the sex offenders in the behavioral treatment studies were at higher risk for recidivism, although the recidivism rates for all forms of treatment are based on both inpatient and outpatient samples. Thus, there is support for cognitive-behavioral and psychohormonal treatments being the most effective treatments with sex offenders. Nevertheless, the evidence of this treatment effectiveness is somewhat limited.

A major limitation of psychohormonal treatments is compliance. These treatments usually involve intramuscular injections and are effective only as long as they are complied with. Sex offender refusal and dropout rates with psychohormonal treatments range from 50 to 66 percent. Moreover, a study directly comparing the relative effectiveness of cognitive-behavioral vs. antiandrogen treatments within the same population of sex offenders is needed. Such a study could clarify whether the effective suppression of sexual arousal achieved by antiandrogen treatments is sufficient to reduce recidivism or whether the more comprehensive aspects of cognitive-behavioral treatments offer necessary adjuncts. Also unknown is whether a combined cognitive behavioral plus antiandrogen approach would be superior to either approach alone.

An encouraging development is evidence of effectiveness of cognitive-behavioral methods in reducing recidivism among adolescent sex offenders. Interventions with adolescents may prevent the development of a history of sex offending that is associated with reoffending. However, there are extremely few outcome studies of adolescent sex offender treatment.

Most of the available treatment research on sex offenders has been conducted with European American populations. It is unknown whether these treatments are equally effective with other groups. For example, cognitive-behavioral methods are individually based. However, individually-based interventions may not be as effective in cultures in which there is a strong group orientation. Thus, individual change might be offset by group norms. For example, in some patriarchal groups, misogynous attitudes may be normative and sexual aggression permissible. Conversely, there may be protective cultural factors that could be mobilized to prevent sex offending. Thus, the context in which interventions are used needs to be examined in future research.

6. *Prevention*

There have been virtually no prevention studies that have examined perpetration of sex offending as an outcome measure. Yet the costs to society of sex offending, in terms of damage to and rehabilitation of victims and of incarceration and rehabilitation of offenders, implore us to seek ways to prevent the problem before it occurs. The recidivism rates of the most effective treatments are at 13 percent, which is

significantly better than other forms of treatment or no treatment. Yet, it could be contended that a 13 percent recidivism rate is unacceptably high. Perhaps the effective components of cognitive-behavioral interventions could be adapted for proactive use in prevention programs.

Most sexual abuse prevention programs have focused on potential victims for interventions. However, perpetrators, not victims, are responsible for sex offenses. A completely effective sex offense prevention program would eliminate the need for victim programs. Prevention programs for potential victims are critically important in terms of empowerment. However, there has been a disproportionate amount of attention in sexual abuse prevention to victims. Relatively simple modifications of existing victim prevention programs could potentially go a long way toward preventing perpetration of sex offenses. For example, most children's sexual abuse prevention programs present the concept of 'bad touch,' which usually is instigated by someone else. Missing from most of these programs, however, is the idea that not only should someone else not 'bad touch' you, but you also should not 'bad touch' someone else. Such an intervention could reach many potential perpetrators who would not otherwise receive this information. The impact of efforts to prevent sex offense perpetration is unknown. Thus, there is a great need for the development and evaluation of interventions to prevent sex offending.

See also: Childhood Sexual Abuse and Risk for Adult Psychopathology; Rape and Sexual Coercion; Regulation: Sexual Behavior; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Perversions (Paraphilias); Treatment of the Repetitive Criminal Sex Offender: United States

Bibliography

Abel G G, Barlow D H, Blanchard E B, Guild D 1977 The components of rapists' sexual arousal. *Archives of General Psychiatry* **34**: 895–903

Bancroft J H, Gwynne Jones H E, Pullan B P 1966 A simple transducer for measuring penile erection with comments on its use in the treatment of sexual disorder. *Behaviour Research and Therapy* **4**: 239–41

Barbaree H E, Marshall W L, Hudson S M 1993 *The Juvenile Sex Offender*. Guilford Press, New York

Barlow D, Becker R, Leitenberg H, Agras W W 1970 A mechanical strain gauge for recording penile circumference change. *Journal of Applied Behavior Analysis* **3**: 73–6

Buss D M, Malamuth N M 1996 *Sex, Power, Conflict: Evolutionary and Feminist Perspectives*. Oxford University Press, New York

Freund K 1963 A laboratory method for diagnosing predominance of homo- or heteroerotic interest in the male. *Behaviour Research and Therapy* **1**: 85–93

Gebhard P H, Gagnon J H, Pomeroy W B, Christenson C V 1965 *Sex Offenders: An Analysis of Types*. Harper & Row, New York

Groth A N 1979 *Men Who Rape: The Psychology of the Offender*. Plenum, New York

Hall G C N 1995 Sexual offender recidivism revisited: A meta-analysis of recent treatment studies. *Journal of Consulting and Clinical Psychology* **63**: 802–9

Hall G C N 1996 *Theory-based Assessment, Treatment, and Prevention of Sexual Aggression*. Oxford University Press, New York

Hall G C N, Andersen B L, Aarestad S, Barongan C 2000 Sexual dysfunction and deviation. In: Hersen H, Bellack A S (eds.) *Psychopathology in Adulthood*, 2nd edn. Allyn and Bacon, Boston

Hall G C N, Barongan C 1997 Prevention of sexual aggression: Sociocultural risk and protective factors. *American Psychologist* **52**: 5–14

Hall G C N, Hirschman R, Beutler L E (eds.) 1991 Special section: Theories of sexual aggression. *Journal of Consulting and Clinical Psychology* **59**: 619–81

Heilbrun K, Nezu C M, Keeney M, Chung S, Wasserman A L 1998 Sexual offending: Linking assessment, intervention, and decision making. *Psychology, Public Policy, and Law* **4**: 138–74

Kinsey A C, Pomeroy W B, Martin C E 1948 *Sexual Behavior in the Human Male*. W B Saunders, Philadelphia

Krafft-Ebing R von 1965 Psychopathia Sexualis. Putnam, New York (Original work published in 1886)

Marshall W L, Fernandez Y M, Hudson S M, Ward T 1998 *Sourcebook of Treatment Programs for Sexual Offenders*. Plenum Press, New York

Prentky R A, Knight R A, Lee A F S 1997 Risk factors associated with recidivism among extrafamilial child molesters. *Journal of Consulting and Clinical Psychology* **65**: 141–9

Rosen R C, Beck J C 1988 *Patterns of Sexual Arousal: Psychophysiological Processes and Clinical Applications*. Guilford, New York

G. C. N. Hall

Sex Preferences in Western Societies

Since 1950, a great number of authors working in the field of anthropology, demography, sociology, and psychology in North America and Europe have tried to determine parental sex preferences. The authors have used various approaches and samples, but in general results are relatively consistent. In this entry, the four methods most frequently used in this field will be reviewed: (a) first-child preference, (b) only-child preference, (c) sex preference for the next child, and (d) the parity-progression ratio technique.

1. First-child Preference

Many authors have tried to determine the sex preference of adult individuals with regard to their firstborn child with a question such as 'For your first child would you prefer a girl or a boy?' More than 30 studies

conducted in North America, for the most part with samples of college students or respondents recruited for fertility studies, have indicated that most women and men prefer a boy rather than a girl for their firstborn child. These findings give the strong impression that the preference for a boy as a firstborn has been universal among people in Western societies since 1950.

Since these results are based on a hypothetical situation that might not happen in the respondent's life, Steinbacher and Gilroy (1985) have argued that an assessment of sex preference when the women are pregnant would be more valid. A review of the literature in English and French found 16 empirical investigations in which it was possible to identify the maternal sex preference of *pregnant* women. In eight of these studies, the information concerning the expectant father's preference was also available, either from the fathers themselves (in three studies) or from their pregnant wives (in five studies).

These studies clearly indicate that first-time pregnant women more often prefer a *girl* than a boy, especially after 1981, when a preference for a girl is shown in six out of seven studies. The data concerning expectant fathers are different; in fact, in seven out of the eight studies, men preferred a boy rather than a girl for a first child.

Concerning the variables associated with the sex preference for a first child, two studies have presented data on pregnant women (Uddenberg et al. 1971, Steinbacher and Gilroy 1985). First, Uddenberg et al. (1971) found that women who grew up with only female siblings more often preferred a son first in a small sample of 81 first-time pregnant women in Sweden. (Some other studies showed the opposite: namely, that the more sisters a woman has, the greater her preference for a girl.) In addition, women who desire a girl are psychologically more autonomous than women who desire a boy. Interestingly, Uddenberg et al. (1971) also found no significant difference between the age or the social classes of the women who preferred a girl or a boy, or who expressed no specific preference. Second, the study by Steinbacher and Gilroy (1985), which dealt with 140 first-time pregnant women in the USA, reported that older women more often chose the no-preference category, and that those who agreed strongly with the women's movement preferred a girl rather than a boy. Otherwise, they did not find any significant relationship in variables such as race, income, marital status, or religion.

Interestingly, in this type of literature, the no-preference percentage has varied from 25 percent to 59 percent since 1981. An important point is to determine whether in fact many of the women who claim to have no preference use this answer to hide a preference, as has been suggested by Pharis and Manosevitz (1980). Marleau et al. (1996) consequently checked the validity of this traditional sex preference question by comparing the answers to that question with the answers to a 'feminine/masculine' scale which assessed how pregnant women imagined the sex of their future baby. It was shown that women having expressed no sex preference on a direct question had in fact no explicit image of their baby as male or female on the 'feminine/masculine' scale. This experiment seems to confirm the validity of the classical direct question.

2. *Only-child Preference*

In reviewing the literature relative to only-child preference, Marleau and Maheu (1998) identified many studies in which it was possible to identify women's and/or men's sex preference(s). Two subgroups of studies are found. The first one consists of 11 studies in which the subjects were forced to select the sex of a child on the basis of the hypothetical situation that they would have only one child in their whole lives. In the second subgroup, five studies were designed to elicit the number of children the subjects desired in their lives. Some subjects declared that they wanted only one child, and it was possible from these answers to determine the sex of this child for further analysis. The results of these two subgroups of studies were collapsed together for the final analysis. It should be noted that all of these studies were made in the USA between 1951 and 1991, and that in the majority of studies the samples consisted of college or university students.

The mains results indicate that women, in general, prefer a boy to a girl for an only child. However, in three of the five most recently published studies, women more often prefer a *girl* rather than a boy for an only child. Results for men show that in nearly all studies, at least 70 percent prefer a boy.

Some of these authors have tried to determine whether any variables are related to the sex preference. The most frequently found connection has been to education; in the two most recent decades, the data indicate that women who have reached university level more often prefer a girl than a boy for an only child. Men, whether they have an university education or not, prefer a boy more often. Pooler (1991) has identified two other variables: the variable 'wife retaining her own name' and the variable 'religion.' Female students who agree with the idea of retaining their own name after marriage more often prefer a girl. In addition, Jewish students prefer a girl whereas Catholics and Protestants prefer a boy.

Some authors have hypothesized that the women's preference for a female only child could be attributed to the fact that the perceived traditional female role disadvantage appears now to be significantly diminishing in Western societies. For example, Hammer and McFerran (1988) showed, significantly, that all subgroups of females (except unmarried noncollege females) would prefer to be reborn as a female.

3. The Preference for Sex of the Next Child

Another method consists in asking individuals their sex preference for a next child, based on the existing family composition. This type of question is found habitually in fertility surveys. In general, the data indicate that women with only one child more often prefer a child of the opposite sex. Moreover, a high percentage of women prefer a child of the opposite sex when they already have two or three children of the same sex. (Men rarely participate in these fertility surveys.)

For example, Marleau and Saucier (1993) showed that almost half of the women who already had a child hoped that their second child would be of the opposite sex in data from the Canadian Fertility Survey of 1984. Nearly 80 percent of women who already had two boys desired a girl for their next child, whereas nearly 50 percent of those with two girls preferred a boy.

Some authors have worked with other measures, especially with the mean number of desired children. Here the results are mixed. Some studies have found that women with a boy and a girl intend to have the same mean number of children as those who have two children of the same sex, but other studies have found that the mean number of children desired is higher for the latter group.

Other authors have worked with a measure such as the use of contraception by women who have already had children. A study done by Krishnan (1993) with the Canadian Fertility Survey of 1984 on women aged between 18 and 49 showed a son preference: the women who already had two sons were more likely to use contraception than those who had two girls.

4. Parity Progression Ratio

A method used by many authors is the parity-progression ratio. This technique consists in observing real behavior rather than being satisfied with verbal statements of attitude as in the methods mentioned above.

The parity progression ratio is the proportion of couples at a given parity who have at least one additional child. If certain sex compositions of existing children are associated with a lower than average progression ratio, the inference is made that the predominant sex in those compositions is preferred. More than 30 studies were identified in the literature. In general, the data indicate that couples with one child continue to bear children regardless of the sex of the first child. Parents with two children of the same sex are more likely to go on than those with one child of each sex.

For example, it was shown (Marleau and Saucier 1996) in a large sample from the Canadian General Social Survey of 1990 that 58 percent of the couples with two children of the same sex were more likely to have another child as compared with 53 percent of couples with two children of both sexes. On the other hand, those who stopped childbearing the more often were those with a boy first and a boy second. No such differences in behavior occurred in couples with three children.

This method is interesting because it can be computed from large databases collected for other purposes. But, it remains a weakness of this method that it gives good results only if sex preferences are relatively homogeneous in the population studied.

5. Conclusion

When we compare the findings reviewed above, we note that a global tendency is revealed by the first three, namely the increasing preference among women for a girl over a boy since 1980. More research will be needed to verify whether this trend will continue and to understand the reasons for this recent shift. A further trend is revealed by the first method, namely the increasing proportion of women who have chosen the no-preference option since 1980. This trend is visible for both pregnant and nonpregnant women.

See also: Family Size Preferences; Family Systems and the Preferred Sex of Children; Family Theory and the Realities of Childbearing Behavior; Family Theory: Economics of Childbearing; Fertility Control: Overview; Fertility: Proximate Determinants

Bibliography

Hammer M, McFerran J 1988 Preference for sex of child: A research update. *Individual Psychology* **44**: 481–91

Krishnan V 1993 Gender of children and contraceptive use. *Journal of Biosocial Science* **25**: 213–21

Marleau J D, Maheu M 1998 Un garçon ou une fille? Le choix des hommes et des femmes à l'égard d'un seul enfant. *Population* **5**: 1033–42

Marleau J D, Saucier J-F 1993 Préférence des femmes canadiennes et québécoises non enceintes quant au sexe du premier enfant. *Cahiers québécois de démographie* **22**: 363–72

Marleau J D, Saucier J-F 1996 Influence du sexe des premiers enfants sur le comportement reproducteur: une étude canadienne. *Population* **2**: 460–3

Marleau J D, Saucier J-F, Bernazzani O, Borgeat F, David H 1996 Mental representations of pregnant nulliparous women having no sex preference. *Psychological Reports* **79**: 464–6

Pharis M E, Manosevitz M 1980 Parental models of infancy: A note on gender preferences for firstborns. *Psychological Reports* **47**: 763–8

Pooler W S 1991 Sex of child preferences among college students. *Sex Roles* **9/10**: 569–76

Steinbacher R, Gilroy F G 1985 Preference for sex of child among primiparous women. *The Journal of Psychology* **124**: 283–8

Uddenberg N, Almgren P E, Nilsson A 1971 Preference for sex of the child among pregnant women. *Journal of Biosocial Science* **3**: 267–80

J. D. Marleau and J.-F. Saucier

Sex-role Development and Education

Today's debate on sex or gender is dominated by two antagonistic paradigms: sociobiology and constructivism. On evolutionary terms man's characteristics are determined by natural selection operating at the level of genes. Man has survived millions of years of competitive struggle for existence. This by definition implies egotism as a core feature of his genes, i.e., a readiness to maximize own reproductive chances at the cost of others. Egotistic genes may produce 'altruistic' behavior in case the net reproductive success of own genes (embodied in close relatives) can be increased (kinship altruism) or future repayments in emergencies can be secured (reciprocal altruism). Gene egotism has implications for sex differences. Across all species females' investment in reproduction is greater: they produce fewer and larger, i.e., more costly gametes; among mammals they invest time in carrying and nursing the young and among humans in taking care of them during an extended phase of dependency. Both sexes—being but vehicles of egotistic genes—strive to maximize reproductive success. Different strategies, however, will be efficient. In view of their high pre-investment, taking care of the young pays for females and given that females will take care, it pays better for males to spread their genes as widely as possible. For females sexual reserve is the better strategy—it allows them to test a male's fidelity and increase his investments; and it pays to select high status males, since their greater resources increase chances of survival for the few young a female will be able to bear. In contrast, males profit from quick seductions and from selecting females according to beauty and youthfulness—criteria that indicate high reproductive capacity. Further assumptions are needed: given that genes happen to reside half of the time of their existence in male, the other half in female bodies, they have to be assumed to operate differently in different environments. Also, the adaptive value of present characteristics can be justified only by reference to life conditions of our ancestors—of which we know little. Constructivism draws a contrary picture. Sex differences—so its core claim—are but social constructions. In fact, the classificatory system itself is a modern Western design stipulating the following features (see Tyrell 1986). (a) Reference to physical aspects, i.e., the presence/absence of a penis (some cultures focus more on social activities like childcare or warfare). (b) Binary, i.e., a strictly exclusive categorization (some cultures allow for or even ascribe a positive status to hermaphrodites). (c) Inclusive, i.e., all individuals are classified even those with unclear genetic make-up (using an operation to improve outward appearance if necessary). (d) Irreversible (except via operation) (some cultures allow for social sex role changes). (e) Ascriptive from birth on (some cultures define children as neutrals and ascribe sex role membership only in initiation rituals). The very assumption of large sex differences is a modern idea that from the constructivist perspective is mistaken: it is an essentialist reification of what in fact is but a cooperative interactive achievement. Humans—so the basic tenet—don't 'have' a sex and they 'are' not males or females, rather they 'act' and 'see' each other as such. Studies on transsexuality analyze the ways in which individuals perform and recognize gender. The two paradigms are based on opposite assumptions. In sociobiology man is but a vehicle for powerful genes, in constructivism he is an omnipotent creator. In sociobiology he is a lone wolf entering social relations only for reproductive concerns, in constructivism he is a social being whose very (even sexual) identity is dependent on interactive co-construction. Nevertheless, both paradigms concur in their ahistorical approach. In sociobiology genes that survived under the living conditions of our ancestors determine human dispositions forever—across all periods and societies. In constructivism man is created ever anew in each interaction situation. Both approaches simply ignore the power of history—consolidated in collective traditions and reproduced in biographical learning processes. These will come to the fore in the following analysis of present sex role understanding. Three aspects will be discussed: its empirical description, historical emergence, and ontogenetic development.

1. *Sex Differences—a Descriptive Perspective*

Sex differences can be analyzed on various levels: on the level of the individual, the culture, the social structure.

1.1 *Psychological Level*

Recent meta-analyses of US data show that with greater educational equality sex differences in cognitive performance have largely disappeared over the past few decades except for a slight overrepresentation of males at the very top of mathematical and the very bottom of verbal abilities and an average higher male performance in spatial ability tasks. Male spatial superiority is even greater among

Mexicans but has not been found among Eskimos. These latter findings suggest a connection between the development of spatial understanding and cultural differences in degree of supervision and control exerted upon young girls.

It has been claimed that morality is gendered: women are maintained to be more flexible and care-oriented, men to be more rigidly oriented to abstract principles and more autonomous. These differences are seen to arise from differences in the structure of self shaped by the early experience of female mothering. Girls can maintain the primary identification with the first caretaker (relational self), while boys—in order to become different—have to distance themselves (autonomous self) (Gilligan and Wiggins 1988).

Neither flexibility nor care, however, are specific to women. Flexibility is a correlate of a modern moral understanding. Kant had still ascribed exceptionless validity to negative duties given that God—not man—was held responsible for any harm resulting from compliance. On innerwordly terms, however, impartially minimizing harm is given priority over strict obedience to rules. If exceptions are at all deemed justifiable they will more likely be conceded by those who are aware of possible costs incurred by anyone affected. Individuals who are personally involved will be more knowledgeable of such costs. This may explain why women—in agreement with Gilligan—were found to judge more flexibly with respect to abortion, yet at the same time more rigidly with respect to the issue of draft resistance than men (Döbert and Nunner-Winkler 1985). Besides, those in power can insist more rigidly on their convictions, thus flexibility might be the virtue of subordinates. Care is not part of a relational self-structure produced in early childhood—rather it is part of the female role obligation. Indeed, preschool girls showed no more empathic concerns than boys (Nunner-Winkler 1994), but a majority of (especially older) German subjects more often justified strictly condemning working mothers by referring to their dereliction of duty and to their egotistic strivings for self-fulfillment than to the harm their children might suffer.

1.2 Cultural Level

Two aspects—although empirically concurring—need to be distinguished: gender stereotypes and gender-role obligations.

Gender stereotypes are collectively shared assumptions about the different 'nature' of men and women. Across cultures men are assumed to be aggressive, independent, and assertive, women to be emotional and sensitive, emphatic, and compliant. Contradictory evidence does not detract from such persuasions—immunity to empirical refutation is the very core of stereotypes and there are mechanisms to uphold them. Expectations guide the way observations are perceived, encoded, and interpreted, e.g., noncompliance is perceived as a sign of strength if shown by a man, of dogmatism if shown by a woman. Behavior that conforms to expectations is encoded on abstract terms, discrepant behavior with concrete situational details. This eases making use of the 'except clause' when interpreting deviant cases (e.g., for a woman she is extraordinarily assertive). Thus, suitably framed and interpreted even conflicting observations can stabilize stereotypes.

With gender-role obligations, women are assigned to the private sphere, men to the public realm. Thus, taking care of children and household chores is seen to be primarily women's task, breadwinning men's. These roles are defined by contrasting features. Family roles are ascribed, diffuse, particularistic, affective and collectivity-oriented; occupational roles are achieved, specific, universalistic, affectively neutral, and self-oriented (Parsons 1964). Identifying and living in agreement with one's gender role will influence ways of reacting, feeling, and judging. Thus, gender differences might be understood as a correlate not primarily of genetic dispositions or of a self-structure shaped in infancy, but rather of the culturally institutionalized division of labor between the sexes.

1.3 Sociostructural Level

Gender stereotypes and role obligations influence career choice and commitment to the occupational sphere. In consequence, there is a high gender segregation of the workforce. The proportion of women is over 90 percent in some fields (e.g., secretary, receptionist, kindergarten-teacher) and less than 5 percent in others (e.g., mechanic, airplane pilot). Jobs that are considered women's work tend to offer fewer opportunities for advancement, less prestige, and lower pay than jobs occupied primarily by men. Worldwide the gender gap in average wage is 30–40 percent and it shows little sign of closing. Top positions in economy, politics, and sciences are almost exclusively filled by men, and part-time working is almost exclusively a female phenomenon. Both men and women tend to hold negative attitudes towards females in authority. Women entering male occupations are critically scrutinized, males entering female occupations (e.g., nursing) in contrast easily win acceptance and promotion.

2. The Historical Emergence of Gender Differences

There are two (partly independent) dimensions implied in the debate on gender roles: the hierarchical one of equality vs. inequality of rights and the horizontal one of difference vs. sameness in personality

make-up. We begin with equality. According to medieval understanding individuals find themselves in different social positions by the will of God and it is God who commanded that women obey men. Enlightenment declared all men to be equal—irrespective of gender (or race). Thus, a new justification was needed if the subordination of women (or black slaves) was to be maintained. This instigated a search for 'natural' differences between women and men (between blacks and whites) that soon succeeded in specifying differences in brain size or the shape and position of sexual organs (in IQ)—much to the detriment of women (or blacks). Legal discrimination has largely discontinued. Women (and blacks) are granted full rights to vote or to participate in the educational system.

The assumption of gender differences, however, is still prevalent. It arose in consequence of the industrialization process. In agricultural economies women had their own sphere of control (house, garden, cattle, commercialization of surplus products) and their contribution was essential for subsistence: 'These women in no way resemble the 19th century image of women as chaste, coy, demure … Peers describe them as wild, daring, rebellious, unruly' (Bock and Duden 1977). Industrialization led to a separation of productive and reproductive work, i.e., to the contrast between familial and occupational roles described above. With rapid urbanization and increasing anonymity around the turn of the twentieth century, antimodernist discontent grew. Increasingly, female 'complementary virtues,' e.g., empathy, emotionality, sensitivity came to be seen as bulwark against the cold rationality of the structure of capitalist economy and bureaucratic administration. This sentiment was (and partly still is) shared even by feminists deeply committed to secure legal, political, and social equality for women.

3. *Ontogenetic Development*

Each generation of newborns is an invasion of barbarians—nonetheless, within a decade or two most turn into useful members of their specific society. How is this effected? Education is too narrow a term in that it refers primarily to methods purposefully applied in order to produce desired results. Children, however, are influenced not primarily by planned educational actions, but rather by the entirety of their life conditions. They are not merely passive objects to social instruction, rather—in mostly implicit learning processes—they actively reconstruct the basic rule systems underlying their experiences. This way they acquire knowledge systems, value orientations, action dispositions. In this (self-)socialization process different learning mechanisms are at work: classical and instrumental conditioning produce response tendencies; through bestowal and withdrawal of love a conformity disposition is shaped; through parental authority and firmness the internalization of values is furthered; children imitate behavior of models they deem interesting and they implicitly recognize regularities and rule structures. With increasing cognitive and ego development reflexive self-distancing from and consciously taking a stance towards one's previous learning history becomes possible.

How are sex roles acquired? Increasingly parents advocate identical educational goals for boys and girls. Nevertheless, unwittingly, especially fathers tend to treat them differently—handling male infants more roughly and disapproving of sissy behavior (Golombok and Fivush 1994). Also, from early on, children prefer same-sex playmates (Maccoby 1990). Such early experiences may leave some traces. More direct sex role learning, however, seems to depend on sociocognitive prerequisites. A change has been documented to occur in children's understanding of concepts. They shift from focusing on externally observable surface features to basic definitional criteria (in the case of nominal terms) or to the assumption of stable and essential inner characteristics that all members of a given category share (in the case of natural kind terms). Sex is treated like a natural kind term, i.e., sex is understood to remain constant despite outward changes, to denote some 'essential' even if unobservable commoness, and to allow for generalizing new information across all members of the same category (Gelman et al. 1986). This constitutes a universal formal frame of reference (that makes stereotypical thinking so irresistible). It needs to be filled with content. Children learn what is typical and appropriate for men and women in their culture by beginning to selectively observe and imitate exemplary same-sex models (Slaby and Frey 1975). Largely, this learning process is intrinsically motivated (Kohlberg 1966), i.e., by the desire to be a 'real boy/girl' and become a 'real man/woman.' It proceeds by (mostly implicitly) reconstructing those gendered behavioral, expressive, and feeling rules that are institutionalized in the given culture. In our culture there are many cues from which children will read what constitutes sex-appropriate demeanor. The sexual division of labor is seen already in the family: women are more likely to sacrifice their career for the family (and men their family-life for their career) and even full-time-employed mothers spend considerably more time on childcare and housework than fathers. In the school, teachers tend to give more attention to boys and praise them for the quality of their work (while praising girls for neatness). Curricula may segregate the sexes, offering home economics to girls, contact sport or mechanic training to boys. The social structure of the school impresses the idea of male authority over women with most principals in elementary school being male although most teachers are female. In fact, it has been found that first-graders attending schools

with a female principal display less stereotypical views on gender roles than children in schools with a male principal. These early lessons on sex differences are reinforced in public: in politics and business top positions are mostly filled by men and books, films, TV, and advertisements depict men as dominant, powerful, and strong, women as beautiful, charming, and yielding. Thus, consciously treating boys and girls alike in family and kindergarten will be of little avail in counterbalancing the impressive overall picture of structural sexual asymmetry and alleged personality differences between the sexes in personality characteristics and behavioral dispositions.

4. *The Future of Sex Roles*

With modernization, ascriptive categories lose importance. Social systems increasingly come to be differentiated in subsystems each fulfilling a specific function and operating according to its own code (Luhmann 1998). Gender is the code of the family; it cannot substitute for the codes of other subsystems, e.g., in science it is the truth of a statement, on the market it is the purchasing power, in court it is the lawfulness of the sentence that counts; the gender of the author, the customer, the judge are (or should be) irrelevant. True, there still exists inequality between the genders. Nevertheless, all over the world it has been drastically reduced over the past few decades. In all countries women have profited more by the educational expansion of the 1960s and 1970s; while at the beginning of the twentieth century in less than 1 percent of 133 countries analyzed voting rights were conceded to women, today all of them with male franchise (over 90 percent) have extended suffrage to women (Ramirez et al. 1997). Many countries have inserted a clause concerning equal social participation rights for women in their constitution and set up special institutions; also some have introduced affirmative actions. Increasingly, women come into prestigious positions which will improve chances for succeeding women who now meet with role models and old girls' networks. Nevertheless, merely increasing the proportion of women on top will not suffice. It may only increase the tendency of a split up of female biographies with women opting for either family or career (while men can have both). A real change requires a more equal distribution of productive and reproductive work between the genders. In this respect Sweden has been quite successful by institutionalizing an egalitarian welfare regime (e.g., providing publicly financed daycare centers across the whole country; granting a generous leave of absence to both parents at child birth, and offering a specific paternity leave to fathers which 83 percent make use of). In Sweden we find a high birthrate (1.8) along with a high rate of female employment (75.5 percent). This stands in contrast to the situation in Germany where both rates are low (1.3; 59.7 percent). The German welfare system is described as paternalistic (e.g., women are offered extended publicly financed maternity leaves yet there are hardly any daycare centers for infants or all-day schools and the disapproval of working mothers is especially high (Garhammer 1997). Such differences between countries may indicate that social policy does have an important part to play in reducing or reproducing gender inequalities.

See also: Education and Gender: Historical Perspectives; Education (Primary and Secondary Schools) and Gender; Gender and School Learning: Mathematics and Science; Gender Differences in Personality and Social Behavior

Bibliography

Bock G, Duden B 1977 *Arbeit aus Liebe—Liebe als Arbeit. Zur Entstehung der Hausarbeit im Kapitalismus*. Frauen und Wissenschaft. Beiträge zur Berliner Sommeruniversität für Frauen Juli 1976. Berlin, 118ff

Döbert R, Nunner-Winkler G 1985 Value change and morality. In: Lind G, Hartmann H A, Wakenhut R (eds.) *Moral Development and the Social Environment*. Precedent Publishing, Inc., Chicago, pp. 125–53

Garhammer M 1997 Familiale und gesellschaftliche Arbeitsteilung—ein europäischer Vergleich. *Zeitschrift für Familienforschung* **9**: 28–70

Gelman S A, Collman P, Maccoby E E 1986 Inferring properties from categories versus inferring categories from properties: The case of gender. *Child Development* **57**: 396–404

Gilligan C, Wiggins G 1988 The origins of morality in early childhood relationships. In: Gilligan C, Ward J V, Taylor J M (eds.) *Mapping the Moral Domain: A Contribution of Women's Thinking to Psychological Theory and Education*. Harvard University Press, Cambridge, MA, pp. 110–38

Golombok S, Fivush R 1994 *Gender Development*. Cambridge University Press, New York

Kohlberg L 1966 A cognitive-developmental analysis of children's sex-role concepts and attitudes. In: Maccoby E E (ed.) *The Development of Sex Differences*. Leland Stanford Junior University, Stanford, CA, pp. 82–173

Luhmann N 1998 *Die Gesellschaft der Gesellschaft*. Suhrkamp, Frankfurt a.M

Maccoby E E 1990 Gender and relationships. A developmental account. *American Psychologist* **4**: 513–20

Nunner-Winkler G 1994 Der Mythos von den zwei Moralen. *Deutsche zeitschrift für Philosophie* **42**: 237–54

Parsons T 1964 *The Social System*. The Free Press of Glencoe, London

Ramirez F, Soysal Y, Shanahan S 1997 The changing logic of political citizenship: Cross-national acquisition of women's suffrage rights, 1890–1990. *American Sociological Review* **62**: 735–45

Slaby R G, Frey K S 1975 Development of gender constancy and selective attention to same-sex models. *Child Development* **46**: 849–56

Tyrell H 1986 Geschlechtliche Differenzierung und Geschlechterklassifikation. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* **38**: 450–89

G. Nunner-Winkler

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Sex Segregation at Work

All societies organize work through a sexual division of labor in which women and men typically perform different tasks. These tasks may also depend as well on people's age, race, ethnicity, nativity, and so forth. Although this sexual division of labor divides market and non market work by sex, sex segregation usually refers to the sexual division of labor among paid workers. Thus, sex segregation is the tendency for the sexes to do different kinds of paid work in different settings.

Customarily, 'segregation' denotes physical separation, as in school segregation by race. However, the term 'sex segregation' refers to physical, functional, or nominal differentiation in the work that women and men do. 'Men's work' does not simply differ from 'women's work,' however; predominantly male jobs tend to be better and more highly valued. This occurs both in vertical segregation which assigns men higher-level jobs than women in the same occupation, and in horizontal segregation in which the sexes pursue different occupations.

1. *Sex Segregation and Sex Inequality*

Segregation—whatever its form—is a key engine of inequality. The domains reserved for members of dominant and subordinate groups are not just different, they are unequal, with the more desirable domains reserved for members of dominant groups. Because economic, social, and psychological rewards are distributed through people's jobs, segregation both facilitates and legitimates unequal treatment. Sex segregation reinforces sex stereotypes, thereby justifying the sexual division of labor it entails, and it reduces equal-status cross-sex contacts that could precipitate challenges to differentiation. In sum, workplace segregation benefits men and harms women (Cotter et al. 1997).

1.1 *Consequences of Sex Segregation*

Job, establishment, and occupational segregation account for most of the differences in the rewards that men and women garner from employment. The most important of these differences is earnings. Estimates of the importance of occupational segregation for the pay gap between the sexes vary widely, depending on the analytic approach. An estimated one-third of the earnings gap results from occupational segregation in the USA and abroad (Anker 1998). Establishment-level data for the USA, Norway, and Sweden by Petersen and his colleagues attribute about three-quarters of the gap to occupational segregation (Petersen and Morgan 1995). This approach may still underestimate the effect of occupational segregation on the pay gap: the amount of sex segregation in metropolitan labor markets affects how much segregation affects women's and men's earnings with all women losing pay and all men gaining in highly segregated markets, and no pay gap if occupational segregation were very low (Cotter et al. 1997, p. 725).

Men dominate the more lucrative jobs partly because employers reserve such jobs for men and partly because female work is devalued (England 1992). The under-valuation of female activities not only shrinks women's pay, it also reduces their economic independence, familial power, and sense of entitlement.

Sex segregation also creates disparities in the sexes' authority, mobility opportunities, working conditions, and chance to acquire skills. Segregation disproportionately relegates women to dead-end jobs or jobs on short career ladder, thus reducing their aspirations and opportunity for mobility ladders (Baron et al. 1986). This vertical segregation creates a 'glass ceiling' that concentrates women in the lower level positions and reduces their authority.

2. *Theories of Segregation*

The most common theories of sex segregation emphasize the preferences of workers and employers. Hypothetically, gender-role socialization or a sexual division of domestic labor that encourages men to pick jobs that maximize earnings and women to select jobs that facilitate child-rearing lead to sex-differentiated preferences (England 1992). Men's vested interests also hypothetically prompt them to exclude women from 'men's jobs.' The universality of the assumed sex-specific preferences limits the value of preference explanations: because preferences theoretically vary between, not within the sexes, the sexes should be completely segregated. Empirical evidence casts doubt on preference theories: youthful occupational preferences only loosely related to the sex composition of adults' occupation (see Jacobs 1989, 1999). Moreover, the sexes similarly value high pay, autonomy, prestige, and advancement opportunities, which should minimize between-sex differences.

Explanations focusing on employers' preferences stemming from sex biases and attempts to minimize employment costs through statistical discrimination are limited by their emphasis on hard-to-measure motives. Although it stands to reason that employers influence job-level segregation because they assign workers to jobs, it is difficult to learn why. However, by examining the effects of training and turnover costs and skill on women's exclusion and sex segregation, Bielby and Baron (1986) showed that employers discriminated statistically against women by imputing to individuals stereotypical characters of their sex, but contrary to neoclassical economic theory, this practice was neither efficient nor rational.

In sum, standard theoretical approaches to segregation predict universally high segregation rather than explaining variation in its level. Although these theories have stimulated considerable research, it is difficult to directly test them, so the importance of women's and employers' 'choices' remains a matter of debate. More fruitful explanations (summarized later) try to explain variation in segregation across aggregates or over time.

3. Measuring Segregation

Segregation is a property of aggregates, not individuals. Its measurement depends on how it is conceptualized. The most common conceptualization—the sexes' different representation across work categories—is measured by the index of dissimilarity, D, whose formula is $\Sigma\, f_i - m_i/2$ (f_i and m_i represent the percentages of female and male workers in each work entity). D is the percentage of either female or male workers who would have to change to a work category in which they are underrepresented for the sexes' distributions across categories to be identical. The size of D depends both on the extent of segregation and the relative sizes of the work entities (e.g., occupation) so it is unsuitable for comparing segregation across populations with work entities of different relative sizes. A size-standardized variant of D which holds entity size constant permits comparing levels of segregation across populations.

Segregation is also conceptualized as the concentration of the sexes in work entities dominated by one sex. This conception is relevant to the thesis that women are disproportionately crowded into relatively few occupations. A simple measure of concentration is the proportion of each sex in the n occupations that employ the most men or women. (For an index of concentration, see Jacobs (1999) who warns that concentration measures assume that female and male occupations are aggregated to the same degree.)

A third conception of segregation is each sex's probability of having a coworker in their job or occupation of the same or the other sex, designated as P* (Jacobs 1999). The value of P* depends on both the extent of segregation and the sex composition of the population.

Occupations can be described in terms of their sex composition, and are sometimes characterized as 'integrated' or 'segregated' or 'female- or male dominated' based on whether one sex is disproportionally represented. For example, in the USA, billing clerk is a segregated occupation because 89 percent of its incumbents are female, whereas women are just 47 percent of the labor force. The individual-level experience of segregation is the sex composition of her or his job or occupation, so a billing clerk holds a predominantly female occupation.

In 1992, Charles proposed using log linear methods to study sex segregation. Because log linear methods adjust for both the varying sizes of work entities and the sexes' respective shares of the labor force they are useful for studying why the extent of segregation varies cross-nationally and over time. They also allow researchers to estimate the impact of independent variables on a sex's representation in specific occupations or occupational categories, net of the occupational structure (Charles 1992, p. 488, 1998).

All measures of segregation are sensitive to the level of aggregation of the work units of interest. If segregation exists within units—which it usually does—the more aggregated the units, the less segregation will be observed. Although fully disaggregated data, such as establishment-level data, are preferable, they are hard to come by. Cross-national data are particularly likely to be aggregated because researchers need to make work units comparable across countries and because the data some countries collect are broadly aggregated.

4. The Extent of Segregation

Most research on segregation focuses on the sexes' concentration in different occupations. (An occupation comprises jobs that involve similar activities within and across establishments.) In a 'typical' country in the early 1990s about 55 percent of workers held either 'male' or 'female' jobs (Anker 1998, p. 5). Although countries vary considerably in the extent to which the sexes are segregated occupationally, in the 1990s most developed nations had segregation indices between 50 and 70 (Anker 1998). In 1997, the segregation index for detailed occupations in the USA was 53.9 (Jacobs 1999).

The sexes are most segregated in the Middle East and Africa. In the early 1990s, segregation indices across detailed occupations were at least 70 for Tunisia, Kuwait, and Jordan, largely because of Muslim proscriptions against contact between the sexes. Most of the EU societies and 'transition economies' (former Soviet satellites) had indices in the mid-50s. Segregation was lowest in Asian/Pacific nations, with indices ranging from 36 (China) to 50 (Japan). China's low segregation level results from its large, sex-integrated agricultural sector.

Among the Western developed countries, Italian and US workers were least segregated, and Scandinavians most segregated. In these countries, occupational segregation is positively related to women's labor force participation; indeed the same factors facilitate women's labor force participation and encourage sex segregation—paid parental leave and part-time work which lead to statistical discrimination against women and the shift of customarily female domestic tasks into the market economy (Charles 1992). Also contributing to cross-national differences in segregation is the relative sizes of manufacturing and service sectors, since men tend to dominate the

former and women the latter (Charles 1998). Variation in national commitment to gender ideology also explains cross-national differences in segregation (Charles 1992).

4.1 Segregation Across Jobs

Much higher levels of sex segregation exist in establishment-level studies of job segregation (a job is a position in an establishment whose incumbents perform particular tasks) because people in the same occupation work in different establishments or different jobs (Bielby and Baron 1984; Tomaskovic-Devey 1993). For example, male lawyers are more likely than female lawyers to work for law firms, and within law firms men dominate litigation. Organizations' characteristics and personnel practices affect how segregated they are. Small and large establishments are the most segregated. Small establishments are segregated because they employ workers of only one sex; whereas large bureaucracies have segregative personnel practices, such as sex segregated job ladders (Bielby and Baron 1984).

5. Trends in Segregation

In the last quarter of the twentieth century segregation fell in most of the world, although women continued to dominate clerical and service occupations, and most customarily male production jobs remained inaccessible to women (Charles 1998). Segregation increased during the 1970s and 1980s in Asian/Pacific countries; remained constant in Southern and Central European countries; and declined in the USA, Western Europe, the Middle East and North Africa, and other developing countries (Anker 1998, p. 328). In most countries declines reflected increased occupational integration as well as shifts in the occupational structure in which heavily segregated occupations shrank.

In the USA, increased integration resulted primarily from women's movement into customarily male occupations. The higher pay in customarily male jobs draw women. Large numbers of men enter customarily female occupations only when they become much more attractive, and such changes are rare. The size-standardized index was stable between 1990 and 1997, indicating that after 1990 real occupational integration halted, although fewer persons worked in heavily segregated occupations (Jacobs 1999).

5.1 Explaining the Decline in Segregation

According to the theoretical framework summarized above, the more similar the sexes become in their preferences and skills, the less segregated they will be. According to cross-national research, the narrowing of the education gap between the sexes has contributed to integration. In the USA, for example, as fewer women majored in education and more majored in business, educational segregation declined and occupations integrated. The shrinking experience gap between the sexes may also reduce segregation if it relieves employers' misgivings over hiring women for jobs that involve firm-specific skills.

However, employers' hiring and job-assignment practices are the proximate determinants of the level of segregation. Many establishments employ women in customarily male jobs only when there are not enough qualified men. Thus, organizational and occupational growth both foster integration if they cause a shortfall of male labor. Shortages have also brought about integration when occupations deteriorated in their pay, autonomy, or working conditions (e.g., pharmacist, typesetter; Reskin and Roos 1990) and thus became less attractive to men.

Employers who face economic penalties for segregating and whose job assignments are monitored are more likely than are others to integrate jobs (Baron et al. 1991). Thus, anti-discrimination and affirmative action regulations have fostered integration by making it illegal for employers to exclude workers from jobs on the basis of their sex. These rules have made a difference largely through 'class-action' lawsuits against employers. Anti-discrimination laws are ineffective in reducing segregation when regulations are not enforced and when class-action lawsuits are not permitted.

Finally, according to development theory, modernization replaces ascription with achievement in filling social positions. However, modernization is associated with the growth of economic sectors that most societies label as 'women's work.' Thus, the net effect of modernization, as measured by GNP, has been segregative because it expands female dominated clerical and sales occupations (Charles 1992).

5.2 Nominal Integration

What appears to be genuine integration may be a stage in a process of re-segregation in which women are replacing men as the predominant sex. In the USA, for example, insurance adjusting and typesetting re-segregated over a 20-year period when these occupations changed from male to female dominated. For individual women, the corresponding process is the 'revolving door' through which women in male occupations pass back into customarily female occupations (Jacobs 1989).

5.3 Explaining Stability

Because the amount of sex segregation has changed so little, scholars have devoted as much attention to explaining stability as to explaining change. Unchanging levels of segregation may result from stability in

the causal variables or from processes whose opposing effects cancel out each other.

Gender ideology strongly favors stability. Many occupations are sex-labeled, representing a cultural consensus that they are appropriate for one sex but not the other (e.g., childcare worker, automobile mechanic). Although these labels are not deterministic (indeed they sometimes vary across societies), they predispose employers to prefer sex-typical workers and workers to prefer sex-typical jobs. Although state policies have altered gender ideology, it is usually a conservative force.

Work organizations also resist change in customary business practices. Organizations' structures and practices reflect the cultural environment present when they were founded, and inertia preserves these structures and practices. Such practices include sex-segregated promotion ladders, credentials that are required for jobs, and the use of workers' informal networks to fill jobs. For example, recruitment through networks perpetuates the sex composition of jobs because people's acquaintances tend to be of their same sex. Statistical discrimination or stereotype-based job assignments perpetuate the sex composition of jobs because they prevent employers from discovering that their stereotypes are unfounded. Thus, barring external pressures to alter job assignments, job segregation is unlikely to change.

Of course, stable levels of segregation may be misleading if they are the result of counteracting forces that have opposing effects on the level of sex segregation. For example, in the 1980s Japan's occupations became slightly more integrated, but this was offset by the increased number of workers employed in occupations that were either male or female dominated. The net effect of the growth of female dominated clerical and service sectors in most economically advanced countries will cancel out some shift toward greater integration within occupations.

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on its Effectiveness; Affirmative Action Programs (India): Cultural Concerns; Affirmative Action Programs (United States): Cultural Concerns; Affirmative Action, Sociology of; Discrimination; Education (Higher) and Gender; Equality of Opportunity; Feminist Legal Theory; Gender and School Learning: Mathematics and Science; Gender and the Law; Modernization, Sociological Theories of; Sex Differences in Pay

Bibliography

Anker R 1998 *Gender and Jobs: Sex Segregation of Occupations in the World*. International Labour Office, Geneva, Switzerland

Baron J N, Davis-Blake A, Bielby W T 1986 The structure of opportunity: How promotion ladders vary within and among organizations. *Administrative Science Quarterly* **31**: 248–73

Baron J N, Mittman B S, Newman A E 1991 Targets of opportunity: Organizational and environmental determinants of gender integration within the California Civil Service 1979–1985. *American Journal of Sociology* **96**: 1362–1401

Bielby W T, Baron J N 1984 A woman's place is with other women. In: Reskin B F (ed.) *Sex Segregation in the Workplace: Trends, Explanations, Remedies*. National Academy Press, Washington, DC

Bielby W T, Baron J N 1986 Men and women at work: Sex segregation and statistical discrimination. *American Journal of Sociology* **91**: 759–99

Charles M 1992 Cross-national variation in occupational sex segregation. *American Sociological Review* **57**: 483–502

Charles M 1998 Structure, culture, and sex segregation in Europe. *Research in Social Stratification and Mobility* **16**: 89–116

Cotter D A, DeFiore J A, Hermsen J M, Kowalewski B, Vanneman R 1997 All women benefit: The macro-level effects of occupational integration on earnings inequality. *American Sociological Review* **62**: 714–34

England P 1992 *Comparable worth. Theories and Evidence*. Aldine de Gruyter, New York

Gross E 1968 Plus ça change … ? The sexual structure of occupations over time. *Social Problems* **16**: 198–208

Hakim C 1992 Explaining trends in occupational segregation: The measurement, causes, and consequences of the sexual division of labor. *European Sociological Review* **8**: 127–52

Jacobs J A 1989 *Revolving Doors*. Stanford University Press, Stanford, CA

Jacobs J A 1999 The sex segregation of occupations: Prospects for the 21st century. Forthcoming In: Powell G A (ed.) *Handbook of Gender in Organizations*. Sage, Newbury Park, CA

Petersen T, Morgan L A 1995 Separate and unequal: Occupation-establishment sex segregation and the gender wage gap. *American Journal of Sociology* **101**: 329–65

Reskin B F 1993 Sex segregation in the workplace. *Annual Review of Sociology* **19**: 241–70

Reskin B F, Roos P A 1990 *Job Queues, Gender Queues: Explaining Women's Inroads Into Customarily Male Occupations*. Temple University Press, Philadelphia, PA

Tomaskovic-Devey D 1993 *Gender and Racial Inequality at Work*. Cornell University, Ithaca, NY

B. F. Reskin

Sex Therapy, Clinical Psychology of

Sex therapy is an approach to the treatment of sexual problems. Results of psychophysiological studies of sexual responses in sexually functional subjects, in the 1960s, allowed gynecologist William Masters and psychologist Virginia Johnson to develop a therapeutic format for the treatment of sexual inadequacy. Apart from organic causes, sexual problems may be attributed to interpersonal difficulties and to problems with emotional functioning. Masters and Johnson were convinced that 'adequate' stimulation would result in sexual response in an engaged partner. Adequate stimulation feels good, is rewarding, and facilitates focus of attention on those feelings. Orig-

inally, adequate stimulation in therapies was restricted to behaviors comprising the classical heterosexual foreplay, coitus, and orgasm sequence. This criterion for adequate stimulation has been abandoned because of the rich diversity in sexual practices that exist in the real world.

Masters and Johnson hypothesized that fear of failure, and as a consequence becoming a spectator of one's own feelings, is the most important cause for sexual inadequacy. They devised a very ingenious procedure to assist people in becoming engaged in feelings of sexual excitement. Three important steps were specified. In the first step people learn to accept that mutual touching feels good. In this step demands for sexual performance, and anxieties related to such demands, are precluded by an instruction not to touch genitals or secondary sex characteristics (e.g., breasts). When the first step leads to acceptance of positive bodily feelings, the second step demands including genitals and breasts in mutual touching and caressing. Masters and Johnson have suggested some variations to accommodate specific sexual problems (e.g., premature ejaculation, erectile problems). The second step has to result in the physiological signs of sexual excitement—vaginal lubrication in women and erection of the penis in men. In the classical sequence the first and second step prepared for the third step which consisted of coital positions, again with variations, and stimulation through coitus to orgasm. Although coitus may be relevant from a reproductive point of view, it is not always and certainly not the only way to experience the ultimate pleasures of sex.

1. *Desire, Excitement, and Orgasm*

Masters and Johnson, on the basis of their psychophysiological studies, proposed a model of sexual response consisting of three phases: an excitement phase, an orgasm phase, and a resolution phase. The excitement phase and orgasm phase may be recognized as the second and third step of the classic treatment format. When other therapists began to apply Masters and Johnson's format it became clear that many patients do not easily engage in interactions prescribed by the sex therapy format. Apparently, an initial motivational step was missing. People who hesitate or avoid intimate interactions may lack desire, which often means that they do not expect the interaction to be rewarding. In 1979 Helen Kaplan proposed adding a desire phase, preceding the three phases specified by Masters and Johnson. Since Kaplan's proposal it has become clear that the prevalence of lack of desire is considerable. With hindsight most people accept that lack of desire must be the most important sexual problem. It is a problem for the individual who does not arouse desire in his or her partner. In most instances, not feeling desire is in itself unproblematic, but lack of sexual desire may become a problem in the relationship.

Sex therapy was a fresh and new treatment. Sexual problems were openly discussed, there was no time-consuming delving into past conflicts, and there were suggestions for a direct reversal of symptoms of sexual failure. Masters and Johnson preferred working with couples because the interaction within the couple often contributes in important ways to the sexual difficulties. Other therapists have offered treatment to individuals and to groups of individuals or couples.

An alternative to Masters and Johnson's therapy format is the mimicking of normal sexual development through the use of masturbation. This has been an important step for many women, especially those who missed this aspect of discovery of their own sexuality. In this approach people learn to induce sexual excitement through masturbation to eventually apply this 'skill' in interaction with a partner.

Some approaches developed from behavior therapy and rational emotive therapy focused on performance anxiety and fear of failure. Others used interventions from couple and group therapy. It is fair to say that nowadays almost all approaches to sexual difficulties incorporate elements from the Masters and Johnson sex therapy format.

2. *Sexual Dysfunctions in Men*

2.1 *Diagnostic Procedures*

The aim of the initial clinical interview is to gather detailed information concerning current sexual functioning, onset of the sexual complaint, and the context in which the difficulty occurred. This information gathering may be aided by the use of a structured interview and paper-and-pencil measures regarding sexual history and functioning. An individual and conjoint partner interview, if possible, can provide additional relationship information and can corroborate data provided by the patient. The initial clinical interview should help the clinician in formulating the problem. It is important to seek the patient's agreement with the therapist's formulation of the problem. When such a formulation is agreed upon, the problem may guide further diagnostic procedures.

Many men with erectile dysfunction may be wary of psychological causes of their problem. Psychological causes seem to imply that the man himself is responsible for his problem. This may add to the threat to his male identity that he is already experiencing by not being able to function sexually. Considering the way a man may experience his problem, it can be expected that it will not be easy to explain to him the contribution of psychological factors. A clinician knowledgeable in biopsychosocial aspects of sexual functioning should be able to discuss the problem openly with the patient. Dysfunctional performance is meaningful performance in the sense that misinformation, emotional states, and obsessive concerns about performance provide information about the

patient's 'theory' of sexual functioning. When contrasting this information with what is known about variations in adequate sexual functioning, it is often clear that one cannot but predict that the patient must fail. For the clinician a problem arises when, even with adequate stimulation and adequate processing of stimulus information according to the clinician's judgment, no response results, either at a physiological or a psychological level.

At this point, a number of assessment methods aimed at identifying different components or mechanisms of sexual functioning may be considered. In principle, two main strategies may be followed: In the first, although a psychological factor interfering with response cannot be inferred from the report of the patient, one can still suspect some psychological factor at work. Possibly the patient is not aware of this factor, thus he cannot report on it. Eliminating this psychological influence may result in adequate response. The second strategy applies when even with adequate (psychological) stimulation and processing, responding is prevented by physiological dysfunction. Physiological assessment may then aid in arriving at a diagnostic conclusion. The biopsychosocial approach predicts that it is inadequate to choose one of these strategies exclusively. The fact that sexual functioning is always psychophysiological functioning means that there may always be an unforeseen psychological or biological factor.

2.2 Psychological Treatments of Sexual Dysfunctions in Men

2.2.1 Approaches to treatment. The most important transformation of the treatment of sexual dysfunctions occurred after the publication of Masters and Johnson's (1970) *Human Sexual Inadequacy*. First of all, they brought sex into the treatment of sexual problems. Before the publication of their seminal book, sexual problems were conceived as consequences of (nonsexual) psychological conflicts, immaturity, and relational conflicts. In most therapies for sexual problems sex was not a topic in the therapeutic transactions. There were always things 'underlying,' 'behind,' and 'besides' the sexual symptoms that deserved discussion. Masters and Johnson proposed to attempt directly to reverse the sexual dysfunction by a kind of graded practice and focus on sexual feelings. If sexual arousal depends directly on sexual stimulation, that very stimulation should be the topic of discussion. Here the second important transformation occurred: A sexual dysfunction was no longer something pertaining to an individual; rather, it was regarded as a dysfunction of the couple. It was assumed they did not communicate in a way that allowed sexual arousal to occur when they intend to 'produce' it. Masters and Johnson thus initially considered the couple as the 'problem' unit. Treatment goals were associated with the couple concept: The treatment goal was orgasm through coital stimulation. This connection between treatment format and goals was lost once Masters and Johnson's concept was used in common therapeutic practice. People came in for treatment as individuals. Male orgasm through coitus adequately fulfills reproductive goals, but it is not very satisfactory for many women because they do not easily achieve orgasm through coitus. What has remained over the years, since 1970, is a direct focus on dysfunctional sex and a focus on sexual sensations and feelings as a vehicle for reversal of the dysfunction.

What Masters and Johnson tried to achieve in their treatment model is a shift in their patients' focus of attention. Let us look at one of Masters and Johnson's interventions to elucidate this point. People with sexual dysfunctions tend to wait and look for the occurrence of feelings, instead of feeling what occurs—hence, the spectator role. Their attention is directed towards something that is not there or does not exist, which is frustrating. In simplest form, Masters and Johnson propose to redirect attention by using the following steps: first of all they manipulate expectations by instructing the patient about what is allowed to occur and what is not. It is explained to the patient that nonsexual feelings are to be accepted as a way to accept sexual feelings later on, and therefore sexual areas are excluded in the initial homework tasks. From a psychological point of view this manipulation is ingenious; it directs attention away from sex—when you feel a caress on your arm it may be pleasant but (now) it is not sexual—however, at the same time, it defines sexual feelings as feelings in 'sexual areas.'

To attain a direct approach of sexual function numerous variants of couple, communication, and group therapy have been used. Rational emotive therapy has been used to change expectation and emotions (see *Behavior Psychotherapy: Rational and Emotive*). To remedy biographical memories connected with sexual dysfunction, psychoanalytic approaches have been used as well as cognitive behavior therapy approaches (see *Behavior Therapy: Psychological Perspectives*) . There are specific interventions for some dysfunctions; for example, premature ejaculation has been treated with attempts at heightening the threshold for ejaculatory release (stop–start or squeeze techniques).

Recently, as a spin-off of research into cardiac vascular smooth muscle pharmacology, drugs have become available which act by relaxing smooth muscles in the spongiose and cavernous structures in the penis. This relaxation is necessary to allow blood flow into the penis, thus causing an erection. Some of these drugs (e.g., sildenafil—Viagra®) support the natural neurophysiological reaction to sexual stimuli. Others act locally in the penis without sexual stimu-

lation. To slow down speed of ejaculation in men with premature ejaculation SSRIs (serotonin selective reuptake inhibitors) seem to be helpful.

Smooth muscle relaxants like sildenafil are helpful in older men when less naturally occuring transmitters are available. Men with vascular or neurodegenerative diseases (e.g., diabetes, multiple sclerosis) may also benefit from the use of smooth muscle relaxants. Although these drugs are very effective, they do not help every man with erectile problems.

Pharmacological treatment for erectile disorder may be an important step in restoring sexual function. Most couples will need information and advice to understand what they may expect from this treatment. For many it will not bring the final resolution of their relationship problems. In addition to drug treatment they will need some form of sex therapy or psychotherapy.

2.2.2 Validated treatments for male sexual dysfunctions. It has been difficult to get an overview of treatments for sexual dysfunctions because any proposal about how to approach dysfunctions was valid. This has changed through the introduction of criteria for validated or evidence-based practice by the American Psychological Association (APA). From the timely review by O'Donohue et al. (1999) of psychotherapies for male sexual dysfunctions it appears that the state-of-the-art is far from satisfactory. Following the criteria of APA's Task Force, they found no controlled outcome studies for male orgasmic disorder, sexual aversion disorder, hypoactive sexual desire disorder, and dyspareunia in men. For premature ejaculation and for erectile disorder there is evidence for the usefulness of psychological treatment. But effects are limited and often unstable over time.

Although the evidence-based practice movement should be firmly supported, unqualified support would be disastrous for the practice of the treatment of sexual problems. The care for patients with sexual problems must be continued even without proof according to the rules of 'good clinical practice.' The sensible clinician will learn to be very careful about any claims concerning either diagnostic procedures or treatments.

3. Sexual Dysfunctions in Women

3.1 Diagnostic Methods

Similar to the procedures in men, initial interviews should help the clinician in formulating the problem and in deciding whether sex therapy is indicated. Since sexual problems can be a consequence of sexual trauma it is necessary to ask if the woman ever experienced sexual abuse. An important issue is the agreement between therapist and patient about the formulation of the problem and the nature of the treatment. To reach a decision to accept treatment, the patient needs to be properly informed about what the diagnosis and the treatment involve. Dependent on the nature of the complaint, the initial interviews may be followed by medical assessments. In contrast to the assessment of men, the psychobiological assessment of women's sexual problems is not well developed.

3.2 Psychological Treatments of Sexual Dysfunctions in Women

3.2.1 Approaches to treatment. Similar to men, the treatment of sexual dysfunction in women contains many elements from Masters and Johnson sex therapy. As noted before, an important addition, especially in women, is the use of masturbation to discover their own sexuality.

Low sexual desire is generally treated with sensate focus exercises to minimize performance pressure, and communication training. In the treatment of sexual aversion the focus is on decreasing anxiety, the common core of sexual aversions. Behavioral techniques, like exposure, are most commonly used.

Treatment of sexual arousal disorder generally consists of sensate focus exercises and masturbation training, with the emphasis on becoming more self-focused and assertive in asking for adequate stimulation.

For the treatment of primary (lifelong) anorgasmia there exists a well-described treatment protocol. Basic elements of this program are education, self-exploration and body awareness, and directed masturbation. Because of the broad range of problems behind the diagnosis of secondary (not lifelong) and situational anorgasmia, there is no major treatment strategy for this sexual disorder. Dependent on the problem, education, disinhibition strategies, and assertiveness training are used. It is important to identify unrealistic goals for treatment like achieving orgasm during intercourse without clitoral stimulation.

For dyspareunia (genital pain often associated with intercourse), there are multiple possible somatic and psychological causes. A common picture is vulvar vestibulitis, pain at small inflamed spots at the lower side of the vaginal opening. However, often there is no clear organic cause for the pain. Treatment should be tuned to the specific causes diagnosed and can vary from patient to patient. Behavioral interventions typically include prohibition of intercourse, finger exploration of the vagina, first by the woman, then by her partner. Sensate-focus exercises may be used to increase sexual arousal and sexual satisfaction. Pelvic floor muscle exercises and relaxation training can be recommended in case of vaginismus or a high level of muscle tension in the pelvic floor.

Treatment of vaginismus commonly involves exposure to vaginal penetration by using dilators of increasing size or the women's fingers. Pelvic floor muscle exercises may be used to provide training in discrimination of vaginal muscle contraction and relaxation, and to teach voluntary control over muscle spasm.

Pharmacological treatment of sexual disorders of women is just beginning. Smooth muscle relaxants have been used in women to ameliorate sexual arousal and as a consequence hypoactive sexual desire. It appears that drugs like sildenafil produce smooth muscle relaxation and increased genital blood flow, but they have no effect on the subjective experience of sexual response. In women with subphysiological levels of testosterone—mainly in the postmenopause—testosterone patches appear to have an effect on mood, energy, and libido.

3.2.2 Validated treatments for women's sexual dysfunctions. Reviews of treatments for sexual dysfunctions in women following the criteria for validated or evidence-based practice have been published (O'Donohue et al. 1997, Heiman and Meston 1997, Baucom et al. 1998). Heiman and Meston conclude that treatments for primary anorgasmia fulfil the criteria of 'well-established,' and secondary anorgasmia studies fall into the 'probably efficacious' group. They conclude with some reservations that vaginismus appears to be successfully treated if repeated practice with vaginal dilators is included in the treatment. Their reservations are due to a lack of controlled or treatment comparison studies of vaginismus. All authors conclude that adequate data on the treatment of sexual desire disorder, sexual arousal disorder, and dyspareunia is lacking.

Although the evidence-based practice movement deserves support, care for patients with sexual problems must be continued even without proof according to the rules of 'good clinical practice.'

4. Future Directions

Sex therapy bloomed in the 1970s and 1980s, but reviews of evidence-based treatments suggest that developments stagnated and very few new studies have been undertaken. The recent shift to biological approaches will continue, at least for a while. Viagra and testosterone patches will shortly be followed by more centrally acting drugs (e.g., dopamine agonists). The search for drugs has provoked a wide range of studies into the biological basis of sexual function. This work inspires behavioral and cognitive neuroscience studies, which may provide a framework and new tools to better understand sexual emotions and sexual motivation.

See also: Psychological Treatments, Empirically Supported; Sexuality and Gender

Bibliography

Baucom D H, Shoham V, Mueser K T, Daiuto A D, Stickle T R 1998 Empirically supported couple and family interventions for marital distress and adult mental health problems. *Journal of Consulting and Clinical Psychology* **66**: 53–88

Heiman J R, Meston C M 1997 Evaluating sexual dysfunction in women. *Clinical Obstetrics and Gynecology* **40**: 616–29

Janssen E, Everaerd W 1993 Determinants of male sexual arousal. *Annual Review of Sex Research* **4**: 211–45

Kaplan H S 1995 *The Sexual Desire Disorders: Dysfunctional Regulation of Sexual Motivation*. Brunner/Mazel, New York

Kolodny R C, Masters W H, Johnson V E 1979 *Textbook of Sexual Medicine*. 1st edn., Little, Brown, Boston

Laan E, Everaerd W 1995 Determinants of female sexual arousal: Psychophysiological theory and data. *Annual Review of Sex Research* **6**: 32–76

Laumann E O, Paik A, Rosen R C 1999 Sexual dysfunction in the United States: Prevalence and predictors. *Journal of the American Medical Association* **281**: 537–44

Masters W H, Johnson V E 1970 *Human Sexual Inadequacy*. 1st edn., Little, Brown, Boston

O'Donohue W T, Dopke C A, Swingen D N 1997 Psychotherapy for female sexual dysfunction: A review. *Clinical Psychology Review* **17**: 537–66

O'Donohue W T, Geer J H (eds.) 1993 *Handbook of Sexual Dysfunctions: Assessment and Treatment*. Allyn and Bacon, Boston

O'Donohue W T, Swingen D N, Dopke C A, Regev L V 1999 Psychotherapy for male sexual dysfunction: A review. *Clinical Psychology Review* **19**: 591–630

W. Everaerd

Sexual Attitudes and Behavior

The focus in this article is on changes in premarital, homosexual, and extramarital sexual attitudes and behaviors as revealed in nationally representative surveys during the second half of the twentieth century. Attention will be mainly on shared sexual attitudes and behaviors in various societies. A shared attitude is a cultural orientation that pressures us to have positive or negative feelings toward some behavior, and to think about that behavior in a particular way.

1. Sexual Attitudes: Premarital Sexuality

Findings from various countries will be examined and compared but because of the large number of national surveys conducted in the USA, we will start there. The first national representative sample of adults in the USA that used a series of scientifically designed questions to measure premarital sexual attitudes was completed in 1963 (Reiss 1967). The National Opinion Research Center (NORC) at the University of Chicago was contracted to do the survey. Reiss composed 24 questions about premarital sexual relationships that

formed two unidimensional scales and several subscales (Reiss 1967, Chap. 2). The highest acceptance was to the question asking about premarital coitus for a man when engaged. Even on this question only 20 percent (30 percent of males and 10 percent of females) said they agreed that such premarital coitus was acceptable (Reiss 1967, p. 31). The date of this survey (early 1963) was strategic because it was at the start of the rapid increase in premarital sexuality that came to be known as the sexual revolution, and which soon swept the USA and much of the western world.

Two years later, in 1965, NORC fielded a national survey that contained four questions on premarital sexual attitudes. One of the questions also asked about the acceptance of premarital coitus for males when engaged. Scott (1998) reports an average acceptance rate of 28 percent, consisting of 37 percent of males and 21 percent of females in this survey. Acceptance was designated by a person checking either of the last two categories ('wrong only sometimes' and 'not wrong at all'), and disapproval was indicated by a person checking either of the first two categories ('always wrong' and 'almost always wrong'). The 20 percent acceptance rate in the 1963 national survey and the 28 percent rate in this 1965 survey lends credence to a rough estimate that at the start of the sexual revolution in premarital sex (1963–5), about one quarter of adults in the USA accepted premarital coitus while three quarters disapproved of it.

In 1970 Albert Klassen and his colleagues at the Kinsey Institute conducted the next nationally representative survey (1989). Klassen also used the NORC to conduct his survey. He used four questions and asked about the acceptability of premarital intercourse for males and females, when in love and when not in love. Reiss reported that the responses to questions about males specifying love and those specifying engagement, were less than 2 percent apart, and so the Klassen question about premarital sex for males in love can be compared to the two previous surveys. Klassen reports that 52 percent (60 percent of males and 45 percent of females) chose the two acceptant categories of 'wrong only sometimes' or 'not wrong at all' (Klassen et al. 1989, p. 389). In just seven years acceptance of coitus rose from 20 percent in 1963, 28 percent in 1965, to 52 percent in 1970. The questions used in all three national surveys seem comparable and, most importantly, the size of the difference from 1963 to 1970 is so large, that it is hard not to conclude that in those seven years, something that can be called a revolution began to evidence itself in American attitudes toward premarital coitus.

Reiss considered the increased autonomy of females and young people as the key factor in the increase of premarital sexual attitudes and developed the *Autonomy Theory* explanation of the sexual revolution around this concept (Reiss 1967, Chap. 10, Reiss and Miller 1979). In the l960s a higher percentage of females were employed than ever before and this meant more autonomy for them, and more autonomy for their children from parental controls and indoctrination. Females more than males were impacted by this increased autonomy and the three studies show much greater proportionate changes in female attitudes than in males. Adult males' acceptance doubled from 30 percent in 1963 to 60 percent in 1970, whereas adult females' acceptance increased 3.5-fold from 10 percent to 45 percent during the same period. The autonomy theory predicted that since female autonomy was increasing the most, female premarital permissiveness would increase the most. This is precisely what happened.

Starting in 1972 the NORC introduced the General Social Survey (GSS) to gather national data annually or biennially on adults in the USA concerning premarital sexuality and a wide range of other nonsexual attitudes and behaviors. These data afford a basis here to examine the change in premarital sexual attitudes from 1972 to 1998. Unfortunately, the GSS researchers did not ask a question modeled after that in the l963, 1965, and 1970 surveys, which all specified gender and the presence of love or engagement. The GSS question basically taps a respondent's global response to the acceptability of premarital coitus. It asked: 'If a man and a woman have sex relations before marriage, do you think it is always wrong, almost always wrong, wrong only sometimes or not wrong at all' (Davis and Smith 1999, p. 235). The 'always wrong' response to the GSS question is the only response that clearly excludes acceptance of coitus even for an engaged or in love male and so it will be considered as the response indicating rejection of such behavior.

The 1972 GSS survey reported that 63 percent accepted premarital coitus under some condition and 37 percent checked the 'always wrong' category. That was a significant growth from the 52 percent acceptance reported by Klassen for 1970. By 1975 the GSS surveys reported that acceptance had risen to 69 percent. The rise was small after 1975 and acceptance was 74 percent in 1998 (Davis and Smith 1999, p. 235). Looking at all the national surveys in the USA from 1963 to 1998 the evidence is that the period of most rapid change in acceptance of premarital coitus was from 1963 to 1975, with an overall increase from 20 percent to 69 percent. That is the period that we can most accurately label as a premarital sexual revolution. In all the surveys discussed the data showed that females, much more than males, increased their acceptance of premarital coitus and this led to more gender equality in attitudes towards premarital coitus. The gender comparisons in 1963 were 10 percent female acceptance to 30 percent male acceptance. By 1975 the comparisons were 65 percent female acceptance to 74 percent male acceptance.

Studies in a number of European societies indicate that even after the increased acceptance of premarital coitus in the USA, many Western countries were still more acceptant than the USA. For example, using

data from the International Social Survey Program (ISSP) of 1994, comparing the USA to five other societies, Scott reports that only Ireland was less acceptant of premarital coitus than the USA. Germany and Sweden were much more acceptant, and even Britain and Poland were significantly more acceptant (Scott 1998, p. 833). Unlike in the USA, Scott reports increases in acceptance of premarital coitus continuing in Britain during the 1980s and 1990s.

There are other national surveys that can be studied such as the 1971 and 1992 national surveys in Finland that show similar trends to what was found in the USA (Kontula and Haavio-Mannila 1995, Chap. 12). Frequent mention of comparable changes in premarital sexual attitudes can be found in the *International Encyclopedia of Sexuality* in its accounts of 31 societies (Francoeur 1997). There are also a number of other European countries with national surveys taken in the 1980s and 1990s but lacking earlier national surveys for comparison. Nevertheless, what evidence we have on these other societies seems to support a significant increase in the acceptance of premarital coitus similar to what was happening in the USA, although not necessarily in the exact same years.

2. *Sexual Attitudes: Homosexuality*

In 1973, a question on homosexual behavior was first asked in the GSS national survey in the USA. No distinction was made between male and female homosexuality. Nineteen percent accepted homosexual behavior as 'wrong only sometimes' or 'not wrong at all.' That did not vary a great deal until the 1993 GSS survey where acceptance jumped to over 29 percent. It then rose to 34 percent in 1996 and to 36 percent in 1998 (Davis and Smith 1999). One can only speculate as to why in the early 1990s this change accelerated in the USA. Perhaps the changes in the 1980s toward greater civil rights for homosexuals encouraged the increase in acceptance of homosexual behavior itself. However changes in the intensity of feelings cannot be indicated by the simple percent distribution on the GSS question on homosexuality. Clearly, using just one question to measure a sexual attitude, while useful, does not afford us sufficient information.

There are national data on homosexuality from other countries. Inglehart, using his two World Value Surveys, compared 20 societies in 1981 and 1990 on the question of homosexuality. He reported that in all but three of the 20 (Ireland, Japan, South Africa) there was an increase in acceptance of homosexuality between 1981 and 1990 (Inglehart 1997, p. 279). In addition, the ISSP (1994) reported that Poland and Ireland were less acceptant of homosexuality than the USA, whereas Britain, West Germany, East Germany, and Sweden were more acceptant (Scott 1998, p. 833).

When we compare changes in males and females in the USA using the GSS data we find that females changed more than males in accepting homosexuality. Homosexual attitudes have traditionally been one of the very few sexuality areas where females equal or exceed males in the acceptance of a sexual behavior. This type of male/female difference was commonly found also in Western European countries studied in the World Values Surveys. But this higher female level of acceptance of homosexuality was not typically found in Eastern European or Asian countries (Inglehart 1997).

3. *Sexual Attitudes: Extramarital Sexuality*

Extramarital sexual attitudes present a very different trend from either premarital or homosexual attitudes. The GSS data for the USA show that the acceptance of extramarital sexuality fell significantly between 1973 and 1988. Acceptance of extramarital sexuality (answering 'wrong only sometimes' and 'not wrong at all') in 1973 was 16 percent but by 1988 this had dropped to only 8 percent. Although male acceptance stayed higher than that of females, both genders showed close to a 50 percent decrease in acceptance over that period. The fear of HIV/AIDS may have played a role. Negative experiences with extramarital sexuality during the era of rapidly increasing divorce rates in the 1970s may also have contributed to this conservative trend.

This more conservative shift in extramarital attitudes was seen in only a minority of the 20 countries on which Inglehart presents data for the period 1981–90 (Inglehart 1997, p. 367). France, Northern Ireland, Sweden, Argentina, and South Africa showed the sharpest decreases in their acceptance of extramarital sexuality. Meanwhile, Mexico, Italy, Finland, and Hungary evidenced the strongest changes toward greater acceptance of extramarital sex. This finding presents a most interesting puzzle as to why some countries changed to be more restrictive while others became more acceptant of extramarital sexuality.

Adding more to this puzzle is the finding by Inglehart that from 1981 to 1990, 16 (of 19) countries increased their belief that a child needs two parents to be happy (1997, p. 287). Thus, there was clearly more agreement in most of these countries on the importance of the two-parent family than on extramarital sexuality. It must be that in some countries, extramarital sexuality was not seen as a challenge to the stability of the two-parent family. There is a need here to also study the various types of extramarital coital relationships to distinguish the impact on stable relationships of having a casual vs. a love affair and/or a consensual vs. a nonconsensual affair (Reiss 1986, Chap. 3). These sexual complexities are but one of the many sexual conundrums waiting to be deciphered by more detailed research that can clarify and elaborate the survey research data presented here.

4. Relation of Sexual Behaviors and Sexual Attitudes

The relation of sexual behavior to the attitude changes we have noted can be explored in several national studies in the USA. The 1982 National Survey of Family Growth (NSFG) interviewed females 15 to 44 years old. These 15–44 year old females were asked about their first coitus, and from this we can obtain retrospective reports for those in this sample who were teenagers in the early 1960s. Hopkins reports that the average percent nonvirginal for 15–19 year old females in 1962 was 13 percent and this rose to 30 percent by 1971 (Hopkins 1997). Zelnik and Kantner (1980) undertook three nationally representative samples of teenage females in the USA during the 1970s. They report that the percentage of 15–19 year old females who had experienced coitus rose from 30 percent in 1971 to 43 percent in l976, and finally to 50 percent in 1979 (Zelnik and Kantner 1980). The 50 percent rate dropped to 45 percent in 1982 and rose back to about 50 percent in 1988 and has stayed close to that level since then (Singh and Darroch 1999, Reiss 1997, Chap. 3). There were many other changes in teenage sexual relationships, such as improved contraception, that cannot be discussed here. Finally, it should be noted that premarital coital behavior, just like premarital attitudes, changed much more for females than for males (Laumann et al. 1994).

When we compare these behavioral changes with the attitudinal changes discussed above, it is apparent that the large increase in the acceptance of premarital coitus starting in the 1960s was very much in line with the contemporaneous increases in teenage coital behavior. Using the GSS surveys allows one to compare attitudes with behavior in all three areas of premarital, homosexual, and extramarital for specific years. These data show a relatively close relationship between attitudes and behaviors in all three areas. For example, among those who said premarital sex is 'always wrong' 32 percent had premarital coitus in the last year, while among those who said premarital coitus was 'not wrong at all,' 86 percent had premarital coitus in the last year (Smith 1994, p. 89). The comparable figures for homosexuality are 1 percent vs. 15 percent, and for extramarital sex 2 percent vs. 18 percent. These are very large differences and they support the interactive relationship of attitudes and behavior in our sexual lives which others have commented upon (Reiss 1967 Chap. 7, Klassen et al. 1989, p. 253).

Many of the same countries that were noted for increases in sexual attitudes since the 1960s also have national data supporting changes in sexual behavior, particularly in premarital sexuality (Inglehart 1997). A large national survey in England presents data supporting a very close association of attitudes and behaviors in premarital, homosexual, and extramarital sexuality (Johnson et al. 1994, p. 245). Also, in the area of homosexuality, Laumann reports similarities in his 1992 American data concerning homosexuality with data from other national surveys (Laumann et al. 1994). Finally, Francouer's encyclopedic work written by experts from 31 societies also supports a close connection between attitude changes and behavior changes in many of the countries studied (Francoeur 1997). The puzzle concerning how, and in what temporal sequences attitudes and behavior influence each other is one that requires careful research and theoretical attention.

5. Conclusions

The representative national surveys examined leads to several important conclusions about sexual attitudes and behaviors in the post Kinsey era. It seems clear that there has been a sexual revolution in the area of premarital sexuality in the USA and in a large number of other Western countries. This is evidenced in both attitudes and behaviors and more strongly on the part of females than males. There has also been a more moderate increase in the acceptance of homosexuality. Finally, it was found that the acceptance of extramarital sexuality has actually decreased in the USA and elsewhere while increasing in a few other countries.

There is a dearth of theories regarding why such sexual changes have occurred. The autonomy theory argues that the key variable in premarital sexual attitude change is the rise in autonomy (Reiss 1967, Reiss and Miller 1979). Such a change is also part of increased social acceptance of gender equality and of premarital sexuality, particularly in a gender equal relationship. The extensive examination by Hopkins of national data in the USA to test the autonomy theory's ability to explain premarital sexuality trends from 1960–90, strongly supports the increases in female gender equality as a key determinant of changes in autonomy, which in turn produced changes in premarital sexual attitudes and behaviors (Hopkins 1997, Chap. 6).

6. Future Directions

Reiss has delineated the nature of a new sexual ethic that he finds is increasingly popular in many countries in the Western world. He calls this new ethic, *HER Sexual Pluralism*, meaning that the moral yardstick in a sexual relationship is now the degree of *Honesty*, *Equality*, *and Responsibility* present. The older norms that judged people by whether they had performed a specific sexual behavior, have increasingly been replaced by focusing on the HER relationship parameters (Reiss 1997, Chaps. 1 and 10). This new ethic fits with the increased acceptance of premarital and homosexual sexuality, for the HER ethic does not use

marriage as the Rubicon of good and bad sexuality. The drop in the acceptance of extramarital sexuality in many countries may reflect the difficulty in carrying out two HER relationships simultaneously. HER sexual pluralism is well integrated with the more gender equal type of evolving Western society and is predicted to become the dominant sexual ethic of the twenty-first century.

One other theoretical explanation of sexual trends comes from Inglehart (1997). He postulates that as capitalist societies become more affluent and increasing numbers of their citizens feel secure, there has occurred a rise in 'non materialist' values. These new values stress well being and quality of life, over accumulation of more economic wealth. Inglehart argues that as part of this major emphasis on quality of life, we are witnessing a liberation and pluralization of sexual values, particularly in the premarital and homosexual areas. Inglehart's thesis is quite compatible with the autonomy theory because both theories are placing the birth of the changes in sexuality as a key part of the development of a new type of society; a society that is more autonomous and more concerned with the quality of life, than with economic survival. The fact that young people in these societies evidence these sexuality trends more than older people lends further support to the future growth of these changes. There appears to be a change in our basic social institutions in much of the world, and with that, a change in sexual relationships is occurring in line with the emerging HER sexual pluralism ethic.

Many aspects of sexual attitudes and behaviors could not be discussed in this article. Even the surveys in the three areas examined illustrate the need for a more detailed examination of the many nuances of sexuality in each area. In addition, sexual science requires better coverage of peoples outside the Western World (Barry and Schlegel 1980, Reiss 1986). We can make progress by combining qualitative and quantitative methods in our work and linking those many disciplines that study sexuality. One way that significant progress can be encouraged is by the establishment of a multidisciplinary Ph.D. degree in sexual science, and in the Spring of 1999 the Kinsey Institute started work to produce just such a degree program (Reiss 1999). This program will be immensely helpful in expanding sexual science's ability to explain our sexual lives. Theory is explanation, and science without theory is just bookkeeping. In all science we need to know why something is the way we find it, not just describe what is found. With the growth of our scientific explanations we will be better able to contain the myriad of sexual problems that plague sexual relationships worldwide.

See also: Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Prostitution; Rape and Sexual Coercion; Rationality and Feminist Thought; Regulation: Sexual Behavior; Reproductive Rights in Affluent Nations; Sexual Behavior and Maternal Functions, Neurobiology of; Sexual Behavior: Sociological Perspective; Sexual Orientation: Historical and Social Construction; Sexuality and Gender

Bibliography

Barry H, Schlegel A (eds.) 1980 *Cross-Cultural Samples and Codes*. University of Pittsburgh Press, Pittsburgh, PA

Davis J A, Smith T W 1999 *General Social Surveys, 1972–1998*. University of Connecticut, Storrs, CT

Francoeur R T 1997 *The International Encyclopedia of Sexuality*. Continuum Publishing, New York, 3 Vols.

Hopkins K W 1997 An explanation for the trends in American teenagers' premarital coital behavior and attitudes between 1960–1990. Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN

Inglehart R 1997 *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ

Johnson A M, Wadsworth J, Wellings K, Field J, Bradshaw S 1994 *Sexual Attitudes and Lifestyles*. Blackwell Scientific Publications, Oxford, UK

Klassen A D, Williams C J, Levitt E E 1989 *Sex and Morality in the US*. Wesleyan University Press, Middletown, CT

Kontula O, Haavio-Mannila E 1995 *Sexual Pleasures: Enhancement of Sex Life in Finland, 1971–1992*. Dartmouth Publishing, Aldershot, UK

Laumann E O, Gagnon J H, Michael R T, Michales S 1994 *The Social Organization of Sexuality*. University of Chicago Press, Chicago

Reiss I L 1967 *The Social Context of Premarital Sexual Permissiveness*. Holt, Rinehart & Winston, New York

Reiss I L, Miller B C 1979 Heterosexual permissiveness: a theoretical analysis. In: Burr W, Hill R, Nye I, Reiss I (eds.) *Contemporary Theories About the Family*. Free Press of MacMillan, New York, Vol. 1

Reiss I L 1986 *Journey Into Sexuality: An Exploratory Voyage*. Prentice-Hall, Englewood Cliffs, NJ

Reiss I L 1997 *Solving America's Sexual Crises*. Prometheus Books, Amherst, NY

Reiss I L 1999 Evaluating sexual science: Problems and prospects. *Annual Review of Sex Research* **10**: 236–71

Scott J 1998 Changing attitudes to sexual morality: A cross-national comparison. *Sociology* **32**: 815–45

Singh S, Darroch J E 1999 Trends in sexual activity among adolescent American women: 1982–1995. *Family Planning Perspectives* **31**: 212–19

Smith T W 1994 Attitudes toward sexual permissiveness: Trends, correlates, and behavioral connections. In: Rossi A S (ed.) *Sexuality Across the Life Course*. University of Chicago Press, Chicago

Zelnik M, Kantner J 1980 Sexual activity, contraceptive use and pregnancy among Metropolitan-area teenagers: 1971–1979. *Family Planning Perspectives* **12**: 230–7

I. L. Reiss

Copyright © 2001 Elsevier Science Ltd. 13973
All rights reserved.

 ISBN: 0-08-043076-7

Sexual Behavior and Maternal Functions, Neurobiology of

During the second half of the twentieth century, interdisciplinary research efforts have generated considerable information about the physiological mechanisms that control mammalian reproductive behaviors. The approach has used field and laboratory studies with animals to elucidate general principles that are beginning to be assimilated by the social sciences as they toil to understand the sexual and parental behaviors of our own species. What follows is a review of basic research on sexual behavior, sexual differentiation, and maternal functions in mammals including humans.

1. *Sexual Behavior*

Mammalian sexual behavior is facilitated in males and females by the hormonal secretions of the testes and ovaries, respectively. In females the display of sexual behavior shows cycles closely associated with fluctuations in the hormonal output of the ovaries. In many species, including rodents commonly used in laboratory experiments, the cyclic display of behavior includes both changes in the ability of the female to copulate as well as her motivation or desire to engage in sexual behavior. In these species intercourse is often physically impossible except during a brief period of optimal hormonal stimulation of central and peripheral tissues. The central effects of ovarian hormones (i.e., estrogen (E) and progesterone (P)) in the facilitation of female sexual behavior are mediated by receptors found in several brain regions including the ventromedial nucleus of the hypothalamus (VMH), the preoptic area, and the midbrain central gray. The effects of E and P on neurons of the VMH appear to be sufficient to facilitate the display of the postural adjustment necessary for copulation in female rats (i.e., the lordosis reflex), and the display of lordosis is prevented by lesions of the VMH (Pfaff et al. 1994). Female rats with VMH lesions can show lordosis if other neural systems that normally inhibit the display of lordosis are also surgically removed (Yamanouchi et al. 1985). The rewarding or reinforcing aspects of female copulation are likely to be mediated by dopaminergic systems that include the nucleus accumbens. The rewarding aspects of female sexuality appear to be activated only when the female can control when and how often she has contact with the male during mating, i.e., when the female can 'pace' the copulatory encounter (Erskine 1989).

In sharp contrast with nonprimate species, female monkeys are capable of engaging in sexual behavior at all phases of the ovulatory cycle and after ovariectomy. This fact which of course also applies to women, has been often used to question the importance of ovarian hormones in the modulation of female sexuality in primates. Other work, however, has shown that in spite of having the ability to copulate throughout the menstrual cycle, female monkeys show fluctuations in sexual motivation that are predictable from the pattern of ovarian E production across the cycle. For the behavioral effects of E to be evident, female monkeys must be tested in situations in which they are given the opportunity to choose between avoiding contact with the male or alternatively to approach the male and solicit his attention. In a series of elegant experiments conducted under naturalistic conditions Wallen and associates (Wallen 1990, 2000) have shown that in rhesus females, the willingness to leave an all-female group in order to approach a sexually active male peaks at the time of maximal E production at the end of the follicular phase. Further, under the same conditions, females never approach the male if their ovarian functions are suppressed by pharmacological manipulations. As argued by Wallen, in rhesus females and perhaps also in women, ovarian hormones do not determine the ability to engage in sexual behavior but have salient effects on sexual desire. Almost nothing is known about where E acts in the brain to facilitate sexual motivation in female primates.

In male mammals, castration results in a reduction in sexual behavior and replacement therapy with testosterone (T) or its metabolites restores the behavior to pre castration levels. In nonprimate species both the motivational and performance aspects of male sexuality are affected by lack of T. In men and possibly other primates, lack of T seems to affect sexual motivation more than sexual ability (Wallen 2000). Thus, men with undetectable circulating levels of T can reach full erections when shown visual erotic materials, but report little sexual interest in the absence of T replacement. In men, erectile dysfunction is often due to non endocrine causes such as vascular pathologies or damage to peripheral nerves. In the brain T seems to act primarily in the medial preoptic area (MPOA) to facilitate male sexual behavior, but it is evident that other brain regions as well as the spinal cord and other peripheral sites need hormonal stimulation for the optimal display of sexual behavior in males. Lesions of the MPOA have immediate and profound disruptive effects on male sexual behavior and these effects are remarkably consistent across species. Lesions of the MPOA however do not equally affect all components of male sexual behavior and some of the effects of the lesions are paradoxical. For example, in male rats lesions of the MPOA that virtually abolish male sexual behavior do not seem to affect willingness to work on an operant task when the reward is access to a receptive female. Similarly, male mice continue to show courtship behavior directed to females after receiving large lesions of the MPOA. It has been suggested that lesions of the MPOA selectively affect consummatory aspects of male behavior leaving appetitive or motivational components intact. Not all the data fit this proposed dichotomy. For

example rhesus monkeys that show copulatory deficits after MPOA damage do not lose all the consummatory aspects of male behavior; after the lesions the animals are capable of achieving erections and they frequently masturbate to ejaculation (Nelson 2000).

The view that male sexual behavior results primarily from T action in the MPOA and female sexual behavior from E and P action in the VMH is not without detractors. In a recent study of male sexual behavior in the rat, it was found that androgen antagonists implanted directly into the MPOA did block male sexual behavior as would be expected. However, the most effective site for androgen antagonist blockade was in the VMH! (McGinnis et al. 1996). There is also considerable overlap between the neural systems that are active during female sexual behavior and those that are active in the male during copulation. These studies may indicate that the appetitive aspects of sexual behavior, seeking out a partner, courtship, etc. are under the control of similar neural systems in both sexes (Wersinger et al. 1993).

2. *Sexual Differentiation*

In addition to the activational effects gonadal hormones have on adult sexual behavior, there is an extensive literature base to show that gonadal hormones also affect the development of brain systems that regulate male and female sexual behavior. These long-lasting effects are often referred to as 'organizational effects' to distinguish them from the concurrent facilitative actions that gonadal hormones have in the adult. Organizational hormone actions are thought to occur primarily during the period of sex differentiation of the nervous system. In species with short gestation, such as rats and hamsters, this developmental period occurs around the time of birth, whereas in those species with longer gestation times, such as primates, sex differentiation of the nervous system takes place during fetal development (Nelson 2000).

Normal males are exposed to androgens throughout this early developmental period as a result of the testes becoming active during fetal development. When genetic female rats were treated with T throughout this time they showed significant masculinization of behavior and genital morphology. As adults, these experimental females displayed most of the elements of male sexual behavior including the ejaculatory reflex. When males were deprived of androgens during this time they were less likely to show the consummatory responses associated with masculine copulation. Further, in the absence of androgens during early development, male rodents develop into adults who show most of the elements of female sexual behavior if treated with ovarian hormones. For many laboratory rodents the behavioral masculinizing effects of T treatment result from metabolites of T, namely estradiol and reduced androgens such as dihydrotestosterone. There is good evidence that for some species it is the estrogenic metabolite that is essential for behavioral masculinization.

Ovarian hormones do not appear to play a role in the development of the neural systems that underlie feminine sexual behavior in mammals. When female laboratory rodents, such as the golden hamster, are treated with estrogen during early development they actually show reduced levels of female sexual behavior as adults. However, their levels of male-like behavior, such as mounting receptive females, are increased, findings that are consistent with the concept that estrogen is a masculinizing metabolite of T in males.

The findings on the effects of gonadal hormones during early development of the female are not easily interpreted, because an independent definition of just what is feminine and what is masculine is not available. On the one hand it is clear that if a female rodent is exposed to high levels of androgen throughout early development she will develop a male-like anatomy and will, as an adult, show all the elements of male sexual behavior. On the other hand, female rodents are often exposed to low levels of androgens normally during gestation as a result of developing next to a male in the uterus. These females are often more dominant than other females and less attractive to males (VomSaal 1989). But they still copulate as females and reproduce, suggesting that normal variations in female phenotype may result from normal variation in androgen exposure during development. This variability in female behavior and attractiveness may be an important part of any normal female population. The problem lies in defining the limits of normal female variation resulting from androgen (or estrogen) exposure, as opposed to what androgen effects might be interpreted as masculinization (Fang and Clemens 1999). The criteria for making this distinction have not yet been defined.

Numerous sex differences have been reported for the mammalian nervous system and most of these probably result from the differential exposure to gonadal hormones that occurs during sex differentiation. It also presumed that the sex differences in behavior that we have noted result from these differences in the nervous system of males and females, but few models are available to show a strong correlation between sex differences in the CNS and sex differences in behavior (but see Ulibarri and Yahr 1996). Gonadal hormones can influence the development of the nervous system in a number of ways such as altering anatomical connectivity, neurochemical specificity, or cell survival. For example, while both male and female rats have the same number of nerve cells in the dorsomedial nucleus of the lumbosacral cord prior to sex differentiation, in the absence of androgen many of these cells die in the female, a normal process referred to as 'apoptosis.' This differential cell-death rate leaves the adult male rat with a larger dorsomedial nucleus than the female. In some brain regions it is suspected that hormones may actually promote cell death result-

ing in nuclei that are larger in the female than in the male. There are also numerous examples of sex differences in the peripheral nervous system as well.

For many years sex differences in human behavior were regarded as reflections of differences in how male and female children are reared. However, the accumulation of volumes of work showing that sex differences in non humans are strongly influenced by differential hormone exposure has forced scientists from many fields to re-evaluate the nurture hypothesis. Most would probably now agree to several general statements. 1. The brains of male and female humans are structurally very different. 2. These differences may reflect, at least in part, the different endocrine histories that characterize the development of men and women. 3. Some of the behavioral differences between males and females may result from these differences in their nervous systems. Disagreement occurs when we try to specify how much of one trait or another is due to biological factors and how much to experience, and it must be recognized at the outset that a clean separation of nature from nurture is not possible.

Most of the evidence for biological factors operating to produce sex differences in human behavior comes from clinical or field psychology studies. A number of syndromes involve variation in androgen or estrogen levels during early development: congenital adrenal hyperplasia (CAH) is a syndrome in which the adrenal gland produces more androgen than normal. Turners Syndrome is characterized by regression of the ovaries and reduced levels of androgen and E from an early age. Hypogonadism is a condition where boys are exposed to lower than normal levels of androgen. There are also populations of girls and boys who were exposed to the synthetic estrogen, diethylstilbestrol (DES) during fetal development as well as populations whose mothers were treated with androgenic-like hormones during pregnancy. A number of studies point to a change in girls play behavior as a result of exposure to androgens during fetal development. CAH girls or girls exposed to exogenous androgenic compounds are often found to show an increased preference for playing with toys preferred by boys and less likely to play with toys preferred by untreated girls. These androgen-exposed girls also are more likely to be regarded as 'tomboys' than their untreated sibs or controls and engage in more rough and tumble play than controls (Hines 1993). Some have argued that variation in androgen or E levels during early development may affect sexual orientation, but a general consensus on this point has not been reached at this time.

3. *Maternal Functions*

For rodents, especially in the case of the laboratory rat, the endocrine and neural mechanisms responsible for the onset and maintenance of maternal behavior are relatively well understood (Numan 1994). This understanding has stemmed to a large extent from the careful description of behaviors shown by maternal rats; these behaviors are easy to identify and to quantify in the laboratory. Such a rich and objective behavioral description is often lacking for other species including our own. In rats, all components of maternal care of the young (except milk production) can be induced in the absence of the hormonal changes that normally accompany pregnancy, parturition, and lactation. Thus, virgin female rats become maternal if they are repeatedly exposed to pups of the right age, a process referred to as sensitization. Hormones, nevertheless, play a critical role in the facilitation of maternal behavior and are necessary for the coordination of maternal care and the arrival of the litter. Using different experimental paradigms, several laboratories have identified E as the principal hormone in the facilitation of maternal behavior. When administered systemically or when delivered directly into the MPOA, E triggers the display of maternal behavior under many experimental conditions. Prolactin from the anterior pituitary or a prolactin-like factor from the placenta also play a role, albeit secondary, in the facilitation of maternal functions; administration of prolactin or analogs of this hormone enhances the activational effects of E on maternal behavior. Also, mice lacking functional prolactin receptors show poor maternal care. Other peptides and steroids have been implicated in the support of maternal functions (Nelson 2000). For example, central infusions of oxytocin facilitate maternal behavior under some conditions and the effects of P on food intake and fuel partitioning are crucial to meet the energetic challenges of pregnancy and lactation (Wade and Schneider 1992).

The integrity of the MPOA is necessary for the display of normal maternal behavior. Damage of the MPOA using conventional lesions or chemical lesions that spare fibers of passage interferes with normal maternal behavior under several endocrine conditions. Similar behavioral deficits are seen after knife cuts that interrupt the lateral connections of the MPOA. Cuts that interrupt other connections of the MPOA do not reproduce the behavioral deficits seen after complete lesions (Numan 1994). Normally, male rats do not participate in the care of the young, but exposing males to pups for several days can induce components of maternal behavior. Compared to females, adult males require more days of sensitization with pups and tend to show less robust maternal behavior. This sex difference is not evident before puberty (Stern 1987) suggesting that sexual differentiation of behavior may not be completed until after sexual maturation. When males are induced to care for the pups, the behavior is sensitive to the disruptive effects of lesions of the MPOA. Since MPOA damage affects both male sexual behavior and the display of maternal care, it is possible that there is partial overlap between the neural circuits

that support these two behavioral functions. Alternatively, MPOA lesions may affect more fundamental aspects of the behavioral repertoire of the animals, and such a deficit then becomes evident in different ways under different social conditions. The effects of MPOA lesions in females are also complex and not fully understood. In addition to affecting maternal functions MPOA damage results in a facilitation of the lordosis reflex concurrently with a reduction in the females' willingness to approach a sexually active male in testing situations that permit female pacing. Fine-grained analyses of the functional anatomy of the MPOA are needed to further elucidate the precise role of this area in mammalian reproductive functions.

Studies of monogamous mammalian species where males and females remain together after mating offer the opportunity to study paternal care as well as social bonding between parents and between parents and offspring. One promising model is that of the prairie vole in which both the female and the male care for the young (DeVries and Villalba 1999). Studies of these monogamous rodents suggest that bonding of the female to the male and to her young may be enhanced by oxytocin, a peptide hormone synthesized in the hypothalamus and secreted by the posterior pituitary. In the male, paternal care appears to result, not from oxytocin, but from another hypothalamic hormone, vasopressin. In addition to these posterior pituitary hormones investigators have found evidence for a role of the endogenous opiates in strengthening the bond between mother and offspring (Keverne et al. 1999).

See also: Queer Theory; Sex Hormones and their Brain Receptors; Sexual Behavior: Sociological Perspective; Sexual Orientation: Biological Influences

Bibliography

Carter C S, Lederhendler I I, Kirkpatrick B (eds.) 1999 *The Integrative Neurobiology of Affiliation*. MIT Press, Cambridge, MA

DeVries G J, Villalba C 1999 Brain sexual dimorphism and sex differences in parental and other social behaviors. In: Carter C S, Lederhendler I I, Kirkpatrick B (eds.) *The Integrative Neurobiology of Affiliation*. MIT Press, Cambridge, MA, pp. 155–68

Erskine M 1989 Solicitation behavior in the estrous female rat: A review. *Hormones and Behaviour* **23**: 473–502

Fang J, Clemens L G 1999 Contextual determinants of female–female mounting in laboratory rats. *Animal Behaviour* **57**: 545–55

Haug M R, Whalen R E, Aron C, Olsen K L (eds.) 1993 *The Development of Sex Differences and Similarities in Behavior*. Kluwer, Boston, MA

Hines M 1993 Hormonal and neural correlates of sex-typed behavioral development in human beings. In: Haug M, Whalen R E, Aron C, Olsen K L (eds.) *The Development of Sex Differences and Similarities in Behavior*. Kluwer, Boston, MA, pp. 131–50

Keverne E G, Nevison C M, Martel F L 1999 Early learning and the social bond. In: Carter C S, Lederhendler I I, Kirkpatric B (eds.) *The Integretive Neurobiology of Affiliation*, MIT Press, Cambridge, MA, pp. 263–74

McGinnis M Y, Williams G W, Lumia A R 1996 Inhibition of male sex behavior by androgen receptor blockade in preoptic area or hypothalamus, but not amygdala or septum. *Physiology & Behavior* **60**: 783–89

Nelson R J 2000 *An Introduction to Behavioral Endocrinology*, 2nd edn. Sinauer Associates, Sunderland, MA

Numan M 1994 Maternal behavior. In: Knobil E, Neil J D (eds.) *The Physiology of Reproduction*, 2nd edn. Raven Press, New York, Vol. 2. pp. 221–302

Pfaff D W, Schwartz-Giblin S, McCarthy M M, Kow L-M 1994 Cellular and molecular mechanisms of female reproductive behaviors. In: Knobil E, Neil J D (eds.) *The Physiology of Reproduction*, 2nd edn. Raven Press, New York, Vol. 2. pp. 107–220

Pfaus J G 1996 Homologies of animals and human sexual behaviors. *Hormones and Behavior* **30**: 187–200

Stern J M 1987 Pubertal decline in maternal responsiveness in Long-Evans rats: Maturational influences. *Physiology & Behavior* **41**: 93–99

Ulibarri C, Yahr P 1996 Effects of androgen and estrogens on sexual differentiation of sex behavior, scent marking, and the sexually dimorphic area of the gerbil hypothalamus. *Hormones and Behaviour* **30**: 107–30

VomSaal F S 1989 Sexual differentiation in litter-bearing mammals: Influence of sex of adjacent fetuses in utero. *Journal of Animal Science* **67**: 1824–40

Wade G N, Schneider J E 1992 Metabolic fuels and reproduction in female mammals. *Neuroscience and Biobehavioral Reviews* **16**: 235–72

Wallen K 1990 Desire and and ability: Hormones and the regulation of female sexual behavior. *Neuroscience and Biobehavioral Reviews* **14**: 233–41

Wallen K 2000 Risky business: Social context and hormonal modulation of primate sexual desire. In: Wallen K, Schneider J E (eds.) *Reproduction in Context*, MIT Press, Cambridge, MA, pp. 289–323

Wallen K, Schneider J E (eds.) 2000 *Reproduction in Context*. MIT Press, Cambridge, MA

Wersinger S R, Baum M J, Erskine M S 1993 Mating-induced FOS-like immunoreactivity in the rat forebrain: A sex comparison and a dimorphic effect of pelvic nerve transection. *Journal of Neuroendocrinology* **5**: 557–68

Yamanouchi K, Matsumoto A, Arai Y 1985 Neural and hormonal control of lordosis behavior in the rat. *Zoological Sciences* **2**: 617–27

A. A. Nunez and L. Clemens

Sexual Behavior: Sociological Perspective

This article considers social science research in the field of human sexual behavior since the start of the nineteenth century. Sexual behavior is understood here in a broad sense to include not just sexual acts but

also the associated verbal interactions and emotions (most notably love), as well as sexual desires, fantasies, and dysfunctions.

1. Overview

The main disciplines considered here are the sociology and anthropology of sexuality, the psychology and psychopathology of sexual behavior, and sexology. These disciplines began to take form in the nineteenth century, and were influenced by other intellectual currents and areas of knowledge which, though not discussed in this article, need to be indicated:

(a) Eugenics, in the form given it by Francis Galton from the 1860s. Eugenicist preoccupations were shared by most of the leading sexologists of the late nineteenth and early twentieth centuries (notably Havelock Ellis and Auguste Forel).

(b) The history of sexuality. This developed in the nineteenth century, initially as the history of erotic art and practices, and of prostitution; and later as the history of sexuality in the ancient world and in other cultural contexts.

(c) The ethology of sexuality. This emerged at the end of the nineteenth century, and has grown in influence since the 1960s, related partly to the development of sociobiology.

Three phases can be identified in the development of social science research on sexuality: the nineteenth century (when the emphasis was on the study of prostitution and the psychopathology of sexual behavior); the period 1900–45 (that of the great sexological syntheses, the pioneering anthropological monographs and the first sex surveys); and the period beginning in 1946 (marked in particular by an expansion of quantitative research on the general population).

2. 1830–99: From the 'Pathological' to the 'Normal'

It would be inaccurate to portray the nineteenth century in the industrialized countries as uniformly puritanical. The period was, of course, characterized by a widespread double standard in sexual morality (much less restrictive for young men than for young women), repression of masturbation, hypocrisy over the expression of love and sexual desires, censorship of literature and erotic art, and so on. But the nineteenth century also saw the development of feminism, the struggle for the civil rights of homosexuals, and the introduction of contraceptive methods in many countries. It was in the nineteenth century also that sexual behavior emerged as a major subject of scientific study.

A characteristic of much work on sexuality in this period is how a concentration on the 'pathological' and 'deviant' was used to cast new light on the 'normal'; that is, on the behavior most widespread in the population. For example, the quantitative study of prostitution preceded that of sexuality in marriage; the 'perversions' (referred to today as 'paraphilias') were examined before heterosexual intercourse between married couples; and the first scientific description of the orgasm (in 1855 by the French physician, Félix Roubaud) actually appeared in a study on impotence.

The first quantitative research on sexual behavior was conducted in the 1830s, much of it using the questionnaire technique being developed at this time. The first major empirical study in this field, based on quantification and combining sociological and psychological perspectives, was the investigation of prostitution in Paris conducted by the physician Alexandre Parent-Duchâtelet (1836).

The second main current of research in the nineteenth century was concerned with the psychopathology of sexuality. The years 1886 and 1887 were a decisive period. In 1886 was published the first edition of Richard von Krafft-Ebing's *Psychopathia sexualis* which presented a systematic classification of 'sexual perversions.' In 1887, the psychologist Alfred Binet (Binet 2001) published an article with the title 'Le fétichisme dans l'amour' ('Erotic fetishism'). Binet's text was the origin of an intellectual fashion for labeling the various 'sexual perversions' as 'isms' (the psychiatrist Charles Lasègue had coined the term 'exhibitionists' in 1877 but not the word 'exhibitionism'). By 'fetishism,' Binet referred to the fact of being particularly—or indeed exclusively—sexually excited by one part of the body or aspect of character, or by objects invested with a sexual significance (for example, underwear or shoes). He argued that the fetishism of any given individual was usually formed in childhood or adolescence, through a psychological process of association, during or after an experience that stirred the first strong sexual feelings. For Binet, many 'sexual perversions' as well as homosexuality should be considered as different forms of 'erotic fetishism.' Finally, he asserted that 'pathological' fetishism (such as obsessional fetishism for certain objects) was merely an 'exaggerated' form of the fetishism characteristic of 'normal' love (that of the majority of people). For a time this notion provided the unifying perspective for the psychopathology of sexuality, beginning with that elaborated by Krafft-Ebing in successive editions of his *Psychopathia sexualis*.

The years that followed saw the generalization of the terms 'sadism' and 'masochism' (popularized by Krafft-Ebing), 'narcissism' (invented by Havelock Ellis and Paul Näcke in 1898–99), 'transvestism' (introduced by Magnus Hirschfeld around 1910). The psycho-analysis of Sigmund Freud integrated these various expressions and, most importantly, Binet's

ideas about the lasting influence of childhood sexual impressions. This period also saw the development—encouraged by Alfred Binet and Pierre Janet—of the analysis of the sexual content in ordinary daydreams. The first questionnaire-based surveys of what in the twentieth century came to be referred to as sexual fantasies were carried out in the United States in the 1890s, in the context of research on adolescence led by G. S. Hall.

3. *1900–45: Large-scale Sexological Syntheses, Pioneering Anthropological Monographs and the First Sex Surveys*

The large volume of research conducted between 1900 and the end of World War II can be divided into three main currents (for a general view of the most significant contributions from this period, see E. Westermarck (1936).

The first current is that of sexological research. It was in this period that sexology acquired an institutional status. The first sexological societies were set up in Germany in the years after 1910, and in 1914 Albert Eulenburg and Iwan Bloch founded the period's most important journal of sexology (the *Zeitschrift für Sexualwissenchaft*). The first Institute for Sexual Science was opened by Magnus Hirschfeld in Berlin in 1919. In the 1920s the first international conferences of sex research were held. This period also saw publication of large-scale works of synthesis, in particular those by Auguste Forel (Swiss), Albert Moll, Hermann Rohleder, Magnus Hirschfeld (German), followed later by René Guyon (French) and Gregorio Marañon (Spanish). In the 1920s and 1930s, sexology became more self-consciously 'political.' A stated aim was to advance the 'sexual liberation' of young people and women, a cause advocated in influential books by B. Lindsey and W. Evans, Bertrand Russell, and Wilhelm Reich. Also published in these years were a number of extremely successful works popularizing sexological questions (the best known being those of the Englishwoman, Marie Stopes, and of the Dutch gynecologist, Th. H. Van de Velde). The aim of these manuals was to promote an enjoyment of marital sex, and the emphasis was accordingly on sexual harmony, orgasm, sexual dysfunctions, and no longer—as had been the case at the end of the nineteenth century—on 'sexual perversions.' The most representative and influential expression of the various tendencies in sexology at this time were the seven volumes of *Studies in the Psychology of Sex* (1900–28) by the Englishman, Havelock Ellis.

The second current of research is in sexual anthropology. Broadly speaking these were either comparative studies of marriage and sexual life (E. Crawley, W. I. Thomas, W. G. Sumner, A. Van Gennep, R. Briffault, K. Wikman and, foremost, E. Westermarck) or anthropological monographs, in particular those by B. Malinowski, M. Mead and G. Gorer. The most important of these monographs is that of Bronislaw Malinowski (1929) on the natives of the Trobriand Islands in New Guinea: topics examined include prenuptial sexuality, marriage and divorce, procreation, orgiastic festivals, erotic attraction, sexual practices, orgasm, the magic of love and beauty, erotic dreams and fantasies, as well as the morals of sex (decency and decorum, sexual aberrations, sexual taboos).

The third and final current from this period is quantitative studies of sex behavior. Before 1914 this research was conducted mainly in Russia, Germany and Scandinavia. Between the wars, it developed primarily in the United States (R. Pearl, G. V. Hamilton, K. B. Davis, R. L. Dickinson, and L. Beam). These works prepared the way for and in many respects prefigured the research conducted by Kinsey and his co-workers from 1938.

4. *Post-1946: Empirical Research in the Age of Sexual Liberalization*

In many countries, the second half of the twentieth century was a period of sexual liberalization. The improved status of women was reflected in a greater recognition of their rights in sexual matters (with implications for partner choice, use of contraception and abortion, as well as sexual pleasure). One consequence of this change was to encourage research into contraception: the contraceptive pill became available from 1960, and sterilization for contraceptive purposes was the most widely used means of birth control in the world by the end of the 1970s. Research was also encouraged into the physiology of the orgasm, and particularly the female orgasm, in the 1950s and 1960s (E. Gräfenberg, A. M. Kegel, A. C. Kinsey, W. H. Masters and V. E. Johnson, etc.). Another important factor of change was the arrival at adolescence in the 1960s of the postwar baby boom generation; economic affluence was the context for their demands for greater sexual freedom. This aspiration was reflected in a fall in age of first intercourse, especially for young women, and, related to this, a decline in the norm of female virginity at first marriage (or formation of first stable union). This liberalization reached its peak in the developed countries at the end of the 1970s and was brought to an abrupt halt by the AIDS epidemic, awareness of which began to develop, first in the United States, from 1981.

In the course of the last fifty years of the twentieth century, the social sciences have made a major contribution to the understanding of human sexuality. Special mention must be made of the contributions from historical demography, history (R. Van Gulik, K. J. Dover, M. Foucault, P. Brown, and others),

ethnology (V. Elwin, G. P. Murdock, C. S. Ford and F. A. Beach), the psychology of sexuality (see Eysenck and Wilson 1979), but also research originating in gay and lesbian studies conducted from a perspective of 'social constructionism' (sexuality is not a biological given but is socially constructed), as well as research on sexual identity, transsexualism, pornography and fantasies (for a general overview of the research mentioned above see: Ariès and Béjin (eds.) 1982, Allgeier and Allgeier 1988, McLaren 1999). Many of the new insights into sexual behavior acquired in this period have come from quantitative-based empirical research.

The most influential of this research in the 1940s and 1950s was that directed in the United States by Alfred Kinsey. Between 1938 and 1954, Kinsey and his co-researchers interviewed more than 16,000 volunteers. While the personal information they collected was probably reliable, the sample constructed by Kinsey's team was not representative of the US adolescent and adult population. Kinsey et al. (1948, 1953) distinguished the following 'sources of sexual outlet': masturbation, nocturnal emissions (or sex dreams), premarital heterosexual petting, premarital coitus (or intercourse), marital coitus, extramarital coitus, intercourse with prostitutes, homosexual responses and contacts, animal contacts. They established that the sexual history of each individual represents a unique combination of these sources of outlet and showed that between individuals there could be wide variation in 'total sexual outlet' (the sum of the orgasms derived from the various sources of sexual outlet). They also identified a number of sociological patterns. For example, compared with less educated people, better educated men and women had first heterosexual intercourse later, but had greater acceptance and experience of masturbation, heterosexual petting, foreplay and orogenital sexual practices. Also, according to these researchers, people with a premarital petting experience were more likely to have a stable marriage. In these two volumes there were curious omissions, the most striking being the almost total neglect of the emotions, notably love. And the interpretations given by the authors were sometimes debatable, such as the presentation of premature ejaculation as almost 'normal' because it happened to be widespread in the United States, or in considering women's erotic imagination to be much less developed than men's on the grounds that it seemed less responsive to sexually explicit images.

It was in large part because of these two volumes that sexology between the 1950s and the end of the 1970s often resembled little more than what has been described as 'orgasmology' (see Ariès and Béjin (eds.) 1982, pp. 183 et seq.). However, they do represent an important stage in the development of the sociology of sexuality.

A large number of quantitative surveys on sexuality were conducted in the 1960s and 1970s, influenced in part by Kinsey's research, but which gave much greater attention to the sexual attitudes, personality, family background, feelings and even fantasies of the people being interviewed. It has also to be noted that this research was increasingly based on representative samples.

These surveys were conducted on adult populations (Sweden, England, France, Finland, United States) but also on adolescents (England, United States, Denmark), young students and workers (West Germany) and homosexuals (United States, West Germany, in particular). They were conducted in a climate of increased politicization of sexuality that recalled the 1920s and often had utopian aspirations. The adoption of alternative sexual lifestyles (exemplified by the 'communes') was advocated by some; others celebrated the revolutionary potential of the (chiefly clitoral) orgasm.

But this climate changed rapidly in the 1980s with the emergence of AIDS. This epidemic, for which no vaccine existed, demonstrated the need for up-to-date empirical data on sexual behavior as a basis for encouraging prevention behaviors by the population (more careful partner selection, and the use of HIV testing and condoms, etc.). In response, large-scale surveys of the general population, often using probability samples, were carried out in the 1990s, in Europe and the United States (see Wellings et al. 1994, Laumann et al. 1994, Kontula and Haavio-Mannila 1995, Béjin 2001) and in the developing world (Cleland and Ferry (eds.) 1995).

The sex surveys of the 1990s cannot be summarized here, but a number of their shared characteristics can be identified. They are based either on interviews or questionnaires (either face-to-face, or self-administered, or by telephone), and have been facilitated by the unquestionably greater willingness in recent decades to talk about sex, as reflected in better participation and response rates and fewer abandons. The theories advanced to interpret the data collected are often, though not always, those which Laumann et al. (1994, pp. 5–24) refer to as 'scripting theory,' 'choice theory' and 'social network theory.' The first postulates that, because of their exposure to an acculturation process, individuals usually follow 'sexual scripts' which prescribe with whom, when, where, how, and why they should have sex. The second places the emphasis on the costs (in time, money, emotional and physical energy, personal reputation, etc.) of sexual behavior. The third seeks to understand why some types of sexual relations occur between people with similar social characteristics whereas others (more unconventional) involve socially more contrasted individuals.

One noteworthy finding from these surveys is that, compared with the developed countries, those of Sub-Saharan Africa are characterized by earlier occurrence of first heterosexual intercourse and a higher level of multiple partnership, but also by a substantially higher

percentage of people who had not had intercourse during the previous month. In other words, in the countries with a 'young' population structure, heterosexual activity tends to begin sooner but occurs less frequently and extends over a shorter period.

5. *Conclusion: Future Directions*

To simplify, it can be said that in the nineteenth century the social sciences (including sexology) focused primarily on the forms of sexuality considered to be 'deviant' or 'perverse.' In addition, they gave priority to an exploration of behavior before beginning to study the psychological aspects (personality, sexual desires and fantasies).

The same process occurred in the twentieth century, but on a broader scale, taking as subject the general population and thus an 'average' sexuality. Initially, between the 1920s and the end of the 1960s, the emphasis was on sexual practices (in particular, 'sexual technique,' and the orgasm). In a second period, however, especially since the start of the 1970s, attention has focused increasingly on sexual desire and fantasies. The fact, for example, that Viagra (the erection-enhancing drug first marketed in 1998) only works for men who feel attracted to their partners, illustrates the need for an understanding of the interior aspects of sexuality, in particular of the psychological blockages, desires and fantasies. This suggests that one trend in the future will be a growth of research on sexual fantasies and the complexities of sexual orientation and identity, and into the effects of pornography, cybersex and sexual addictions.

A second probable trend in future research is the development of the comparative study of 'national' sexualities as revealed in the sex surveys of the 1990s. The data on sexuality that has been assembled needs to be subjected to analysis and interpretation. A comparative analysis of national sex surveys offers an excellent means of identifying the influence of culture on sexual attitudes and behavior as well as on desires and fantasies. It should be possible, for example, to compare the respective influence of cultures that are predominantly hedonistic or ascetic in orientation. Other factors to be assessed include religious beliefs and practices, population aging, democratization, and the relationship between the sexes. This material is potentially the basis for new syntheses in the sociology of sexuality, comparable in scope to the great sociosexological syntheses produced in the first thirty years or so of the twentieth century.

A third trend in future research will be a continuing exploration of the themes developed since the 1970s whose common point is their focus on sexual behavior that is more or less coercive in character: paedophilia, sexual tourism, sexual violence, sexual harassment, and sexual mutilations, notably those inflicted on women.

A fourth line of research for the future follows from the growing numbers of elderly and old people in the population, who increasingly expect to remain sexually active. Their sexuality will surely be the subject of further and more detailed research.

The fifth and final trend for the future concerns a partial renewal of social science research on sexuality from the recent developments in the ethology of sexuality. More generally, indeed, interdisciplinary perspectives can be expected to inform and enrich all areas of social science research in the field of sexual behavior.

See also: Eugenics, History of; Family and Kinship, History of; Heterosexism and Homophobia; Sex Therapy, Clinical Psychology of; Sexual Attitudes and Behavior; Sociobiology: Overview

Bibliography

Allgeier A R, Allgeier E R 1988 *Sexual Interactions*. Heath, Lexington, MA

Ariès P, Béjin A (eds.) 1982 *Sexualités occidentales*. Seuil, Paris [1985 *Western Sexuality*. Basil Blackwell, Oxford, UK]

Béjin A 2001 *Les fantasmes et la vie sexuelle des Français*. Payot, Paris

Binet A 2001 *Le fétichisme dans l'amour* [1st edn. 1887]. Payot, Paris

Cleland J, Ferry B (eds.) 1995 *Sexual Behaviour and AIDS in the Developing World*. Taylor and Francis, London

Ellis H 1900–28 *Studies in the Psychology of Sex* (7 Vols.). Davis, Philadelphia, PA

Eysenck H J, Wilson G 1979 *The Psychology of Sex*. Dent, London

Kinsey A C, Pomeroy W B, Martin C E 1948 *Sexual Behavior in the Human Male*. Saunders, Philadelphia, PA

Kinsey A C, Pomeroy W B, Martin C E, Gebhard P H 1953 *Sexual Behavior in the Human Female*. Saunders, Philadelphia, PA

Kontula O, Haavio-Mannila E 1995 *Sexual Pleasures. Enhancement of Sex Life in Finland, 1971–1992*. Dartmouth, Aldershot, UK

Laumann E O, Gagnon J H, Michael R T, Michaels S 1994 *The Social Organization of Sexuality. Sexual Practices in the United States*. University of Chicago Press, Chicago, IL

McLaren A 1999 *Twentieth-century Sexuality: A History*. Basil Blackwell, Oxford, UK

Malinowski B 1929 *The Sexual Life of Savages in North-Western Melanesia*. Routledge, London

Parent-Duchâtelet A 1836 *De la prostitution dans la ville de Paris*. Baillière, Paris

Wellings K, Field J, Johnson A M, Wadsworth J 1994 *Sexual Behaviour in Britain*. Penguin Books, Harmondsworth, UK

Westermarck E A 1936 *The Future of Marriage in Western Civilization*. Macmillan, London

A. Béjin

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Sexual Harassment: Legal Perspectives

After a remarkably swift development in law and popular consciousness, sexual harassment remains the subject of controversy and debate. The concept of sexual harassment is largely an American invention. Like many other concepts that have mobilized social action in the United States, this one emerged in the context of law reform. In the 1970s, US feminists succeeded in establishing sexual harassment as a form of sex discrimination prohibited by Title VII of the Civil Rights Act of 1964 (the major federal statute prohibiting discrimination in employment). Since then, the concept has taken hold in broader arenas such as organizational practice, social science research, media coverage, and everyday thought. Although Title VII remains the primary legal weapon against sexual harassment in the US, traditional anti-discrimination law squares uneasily with the overtly sexual definition of harassment that emanated from 1970s feminist activism and ideas. Just at the time when this narrow sexual definition of harassment has come under criticism in the US, however, that same definition has begun to spread to other nations—inviting inquiry around the globe. Today, sexual harassment's conceptual boundaries are being rethought in the law, in the wider culture, and in feminist thought.

1. *Definitions and Origins*

In the United States, harassment is predominantly defined in terms of unwanted sexual advances. In the eyes of the public and the law, the quintessential case of harassment involves a powerful male supervisor who makes sexual advances toward a female subordinate. Harassment is an abuse of sexuality; it connotes men using their workplace power to satisfy their sexual needs.

This sexual model of harassment was forged in early Title VII law. Some of the earliest cases were brought by women who had been fired for refusing their bosses' sexual advances. Lower courts at first rejected these claims, reasoning that the women had been fired because of their refusal to have affairs with their supervisors and not 'because of [their] sex' within the meaning of the law. The appellate courts reversed; they held employers responsible for the bosses' conducts as a form of sex discrimination now called quid pro quo harassment.

The results were a step forward: It was crucial for the courts to acknowledge that sexual advances can be used as a tool of sex discrimination. But the reasoning spelled trouble, because the courts' logic equated the two. The courts said the harassment was based on sex under Title VII because the advances were driven by a sexual attraction that the male supervisor felt for a woman but would not have felt for a man. By locating the sex bias in the sexual attraction presumed to underline the supervisor's advances, these decisions singled out (hetero) sexual desire as the *sine qua non* of harassment. Had the supervisor demoted the plaintiff or denigrated her intelligence, the court would have had far more difficulty concluding that the conduct was a form of sexism prescribed by law.

Even at the time, there were broader frameworks for understanding women's experiences as sex harassment. Carroll Brodsky's book, *The Harassed Worker* (1976), for example, articulated a comprehensive non-sexual definition of harassment as 'treatment that persistently provokes, pressures, frightens, intimidates, or otherwise discomforts another person.' Rather than a form of sexual exploitation, Brodsky saw harassment as 'a mechanism for achieving exclusion and protection of privilege in situations where there are no formal mechanisms available.' (Brodsky 1976, p. 4). In his usage, 'sexual harassment' referred not simply to sexual advances, but to all uses of sexuality as a way of tormenting those who felt 'discomfort about discussing sex or relating sexually.' (Brodsky 1976, p. 28). A few Title VII decisions had already recognized sexual taunting and ridicule as a mechanism for male supervisors and co-workers to drive women away from higher-paying jobs and fields reserved for men. Indeed, the very concept of harassment as a form of discrimination in the terms and conditions of employment was first invented in race discrimination cases, where judges had discovered that employers could achieve racial segregation not only through formal employment decisions such as hiring and firing, but also through informal, everyday interactions that create an atmosphere of racial inferiority in which it is more difficult for people of color to work.

Analogizing to the race cases, the courts might have likened bosses' demands for sexual favors from women to other discriminatory supervisory demands—such as requiring black women to perform heavy cleaning that is not part of their job description in order to keep them in their place, or requiring women to wear gendered forms of dress or to perform stereotypically feminine duties not considered part of the job when men do it. Or judges might have located the sexism in a male boss's exercise of the paternalist prerogative to punish as an employee someone, who dares to step out of her place, as a woman by refusing the boss sexual favors; sociological analysis reveals that male bosses have penalized female employees for other non-sexual infractions that represent gender insubordination rather than job incompetence (Crull 1987).

But the courts relied instead on a sexualized framework put forward by some feminist lawyers and activists. Toward the mid-1970s, US cultural-radical feminists moved toward a simplistic view of heterosexuality as the lynchpin of women's oppression (Willis 1992, p. 144). Given this ideological commitment, it is not surprising that these early feminists conceived of women's workplace harassment as a

form of unwanted sexual advances analogous to rape. Lin Farley's book, *Sexual Shakedown*, defined harassment as 'staring at, commenting upon, or touching a woman's body; requests for acquiescence in sexual behavior; repeated nonreciprocated propositions for dates; demands for sexual intercourse; and rape.' (Farley 1978, p. 15). A few years later, Catharine MacKinnon argued that harassment is discriminatory precisely because it is sexual in nature—and because heterosexual sexual relations are the primary mechanism through which male dominance and female subordination are maintained (MacKinnon 1979).

In the US, less than a decade later, the consolidation of this narrow view of women's workplace harassment was largely complete. The 1980 Equal Employment Opportunity Commission (EEOC) guidelines defined sex harassment as 'unwelcome sexual advances, requests for sexual favors, and other verbal or physical conduct of a sexual nature'—a definition courts have read to require overtly sexual conduct for purposes of proving both *quid pro quo* harassment (which involves conditioning employment opportunities on submission to sexual advances) and hostile work environment harassment (which involves creating an intimidating or hostile work environment based on sex). Indeed, in hostile environment cases, the lower courts have tended to exonerate even serious sexist misconduct if it does not resemble a sexual advance (Schultz 1998).

Media coverage has infused this view of harassment into popular culture. The 1991 Anita Hill-Clarence Thomas controversy helped solidify the view of harassment as sexual predation. Hill, at the time a novice lawyer in her mid-twenties, claimed that Thomas, her then-supervisor at the Department of Education and later Chair of the EEOC, had pressured her to go out with him and regaled her with lewd accounts of pornographic films and his own sexual prowess (Mayer and Abramson 1994). Soon afterward, the news media broke the story of Tailhook, in which drunken Navy pilots sexually assaulted scores of women at a raucous convention (Lancaster 1991). Later in the 1990s, public attention turned to the harassment lawsuit of a former Arkansas state employee, Paula Jones, who alleged that President Bill Clinton had made crude sexual advances towards her while he was the Governor of Arkansas.

Organizations seeking to avoid legal liability for sexual harassment have also defined it in terms of overtly sexual conduct. Many employers have adopted sexual harassment policies, but Schultz's research reveals that these policies are rarely if ever integrated into broader anti-discrimination programs. Indeed, some employers' efforts to avert sexual harassment may undermine their own efforts to integrate women fully into the workplace, as firms adopt segregationist strategies designed to limit sexual contact between men and women (such as prohibiting men and women from traveling together). Such policies reinforce perceptions of women as sexual objects and deprive them of the equal training and opportunity the law was meant to guarantee.

2. *Current Challenges*

In recent years, the prevailing understanding of sexual harassment has come under challenge in the US. Civil libertarians have voiced concern that imposing vicarious liability on employers for their employees' sexual harassment gives employers a powerful incentive to curb workers' freedom of speech and sexual expression in the workplace. Critics say harassment law incorporates vague standards—including the requirement that the hostile work environment harassment be 'sufficiently severe and pervasive to alter the conditions of the victim's employment and create an abusive working environment'—(*Meritor Savings Bank vs. Vinson*, p. 67)—that permit employers to adopt broad policies that chill sexual speech (Strossen 1995, Volokh 1992). Many employers may not limit their policies to 'unwelcome' conduct, because they will want to avoid costly, contentious inquiries into whether particular sexual interactions were unwelcome. Because few employers have a stake in protecting their employees' freedom of expression (and only government employers have any First Amendment obligation to do so), many firms may simply adopt broad, across-the-board proscriptions on sexual activity and talk on the part of employees (Rosen 1998, Hager 1998). Although there has been no systematic research in this area, some alarming incidents have been reported. (Schuldt 1997), (Grimsley 1996, p. A1).

Such concerns have resonated with a new generation of feminist legal scholars, who have begun to worry about the extent to which equating workplace sexual interaction with sex discrimination replicates neo-Victorian stereotypes of women's sexual sensibilities (Abrams 1998, Franke 1997, Strossen 1995). While civil libertarians have urged repealing or restricting the range of employer liability under Title VII—often, proposing instead to hold individual harassers responsible for their own sexual misconduct under common law (Hager 1998, Rosen 1998), younger feminist scholars have focused on reforming sex harassment law to bring it in line with traditional anti-discrimination goals. Kathryn Abrams defines harassment as sex discrimination not because (hetero) sexual relations inherently subordinate women, but because she believes male workers use sexual advances to preserve masculine control and norms in the workplace; she would limit liability to cases in which the harasser manifestly disregards the victim's objections or ambivalence toward his advances (Abrams 1998). Katherine Franke argues that harassment is a 'technology of sexism' through which men police the boundaries of gender; she would focus on whether men have used sexuality to press women or other men

into conventional 'feminine' or 'masculine' roles (Franke 1997). Both Abrams and Franke attempt to break with the old equation of sexuality and sexism. Yet neither makes a decisive break, for each retains the idea that (hetero) sexual objectification is the key producer of gender (for Abrams, of gender subordination in the workplace, for Franke, of gender performance throughout social life).

Bolder analyses seek to jettison altogether sexual harassment's conceptual underpinnings in sexuality. Janet Halley's queer theory-based critique emphasizes both the cultural/psychic dangers of outlawing the expression of sexuality and the heavier-handed repression such an approach places on sexual minorities. To the extent that harassment law focuses on whether sexual conduct is offensive to a reasonable person, judges and juries will rely on their 'common sense' to evaluate the advances—and other actions—of gays, lesbians, bisexuals, and other sexual dissidents as inherently more offensive than those of heterosexuals (Halley 2000). As Kenji Yoshino has observed, the courts have conditioned liability in same-sex sexual harassment cases on the harasser's sexual orientation. Conduct that courts consider an unwanted sexual advance when the harasser is homosexual is deemed innocuous horseplay when the harasser is heterosexual—an approach that sets up a two-tiered system of justice that has nothing to do with the victim's injury (Yoshino 2000). (See *Sexual Orientation and the Law*.)

For related reasons, Schultz has argued that sex harassment law should abandon its emphasis on sexual misconduct and focus on gender-based exclusion from work's privileges. She argues for reconceptualizing harassment as a means of preserving the masculine composition and character of highly-valued types of work and work competence (Schultz 1998). In Schultz's view, the prevailing sexual understanding of harassment is too narrow, because it neglects more common, non-sexual forms of gender-based mistreatment and discrimination that keep women in their place and prevent them from occupying the same heights of pay, prestige, and authority as men. Indeed, Schultz contends, the centrality of occupational identity to mainstream manhood leads some men to harass other men they regard as unsuitably masculine for the job. At the same time, sexual model risks repression of workers' sexual talk or interaction—even where they do not threaten gender equality on the job. Schultz's call to move away from a sexuality-based harassment jurisprudence builds on the earlier work of Regina Austin, who recognized in 1988 that most workplace harassment was not 'sexual' in nature and proposed a tort of worker abuse to protect employees from class-based mistreatment at the hands of their bosses (Austin 1988). Although Austin was concerned with structures of class and race, recently a more individual dignitary approach has been revived by Anita Bernstein and by Rosa Ehrenreich, who propose protecting employees' rights to equal respect through anti-discrimination and tort law, respectively (Bernstein, 1997, Ehrenreich 1999). Recognizing harassment as just another form of discriminatory treatment restores Title VII's protections to those who allege discriminatory abuse based on characteristics other than gender or even race, such as religion and disability (Goldsmith 1999).

3. *Future Directions*

In the United States, the sexual model of harassment always rested uneasily alongside traditional employment discrimination law—which is concerned with work, not sexuality. Although the future is far from certain, the US Supreme Court's recent sexual harassment decisions seem poised to restore harassment law to its traditional focus.

The Court's 1998 decisions in *Ellerth vs. Burlington Industries* and *Faragher vs. Boca Raton* held that a company's vicarious liability for a supervisor's harassment turns on whether the harassment involves a 'tangible employment action'—such as hiring, firing, or promotion—not on the content of the misconduct or its characterizations as quid pro quo or hostile environment harassment. In the absence of such a tangible action, companies can avoid liability by proving that they adequately corrected harassment a victim reported (or reasonably should have reported) through acceptable in-house channels. By creating a loophole for companies who investigate harassment through their own procedures, the Court sought to counter any current incentives for managers to ban sexual interactions across the board. By adhering to vicarious liability where harassment involves the same tangible employment decisions as more traditional forms of discriminatory treatment, the Court acknowledged that harassment is simply a form of employment discrimination subject to the usual legal rules (White 1999).

The Court's decision in *Oncale vs. Sundowner Offshore Services* further reconciles sexual harassment with the traditional discrimination approach. In *Oncale*, the Court held that male-on-male harassment is actionable under Title VII, taking pains to emphasize that harassment is not to be equated with conduct that is sexual in content or design. 'We have never held that workplace harassment … is … discrimination because of sex merely because the words used have sexual content or connotations,' said the Court (*Oncale vs. Sundowner Offshore Services*, *Inc*., p. 80). By the same token, 'harassing conduct need not be motivated by sexual desire to support an inference of discrimination on the basis of sex.' (*Oncale vs. Sundowner Offshore Services*, *Inc*., p. 80). 'The critical issue,' stressed the Court, 'is whether members of one sex are exposed to disadvantageous terms or conditions of employment to which members of the other

sex are not exposed, (*Oncale vs. Sundowner Offshore Services, Inc.*, p. 80).

The social sciences are also beginning to look beyond the sexual model, providing evidence of the need to conceptualize sex harassment in broader terms. In the early years, most workers attempted to document the prevalence of harassment, albeit with limited results due to their implicit reliance on a conceptually narrow (yet often vague) notion of harassment specified by lists of sexual behaviors derived from the EEOC guidelines (Welsh 1999). Over time, the research has become more multi-faceted, as a burgeoning scholarship has focused greater attention on defining harassment, theorizing its causes, documenting its consequences, and developing predictors. Yet much research remains wedded to the sexual view. Prominent theories of harassment posit that men's tendency to stereotype women as sexual objects is 'primed' by the presence of sexual materials or behavior to induce a 'sex-role spillover' effect that leads men to sexualize women inappropriately in the workplace (Fiske 1993, Gutek 1992). Predictive models search for characteristics that predispose men to harassment, such as a propensity to sexualize women over whom they have some supervisory authority (Pryor et al. 1995). Further examples abound (see Welsh 1999, Borgida and Fiske 1995). Nonetheless, recent research has opened up a broader horizon. A few studies have included measures of gender-based harassment that is not necessarily sexual in content or design: The results suggest that such harassment is more widespread than overtly sexual forms (Frank et al. 1998, Fitzgerald et al. 1988). Social psychologists have begun to look beyond sexual objectification to explore how other gender-based stereotypes and motives can combine with occupational identities and institutional contexts to produce a variety of forms of workplace harassment and discrimination that are not all motivated by sexual attraction (Fiske and Glick 1995). Even some economists have moved away from the conventional view of harassment as sexual coercion that is unrelated to the internal dynamics of labor markets (Posner 1989) to develop innovative theories to explain how incumbent workers can obtain the power to exclude and disadvantage aspiring entrants (Lindbeck and Snower 1988). These developments align social psychological and economic theories more closely with those of sociologists, who have long emphasized harassment's connection to structural features of organizations such as gross numerical imbalance, standardless selection processes, and the absence of managerial accountability (Reskin 2000, Schultz 1991).

In tension with such efforts to transcend a narrow definition of sex harassment in the United States are developments in some other parts of the globe, where the sexual model championed by early US cultural-radical feminists seems to be gaining headway. According to some commentators, the American understanding of sex harassment has been disseminated abroad so successfully that it now forms the foundation for international debates on sex harassment (Cahill 2000). Following closely on the heels of legal developments in the USA, for example, the European Union took steps to condemn and outlaw workplace sex harassment as a violation of women's dignity, and it defined this concept in terms strikingly similar to the definition promulgated by the US EEOC. Encouraged by EU initiatives, feminists in Europe have drawn on particular features of the US model to promote versions of sex harassment law that resonate with their own traditions. In Austria, Cahill shows, feminists capitalized on desires to signal Austria's compliance with the economic First World's laws to press for a harassment law that adopts both the sexualized substantive definition and the privatized enforcement mechanisms of the US approach. The importation of these features allows critics to reject not only the law but also the very existence of harassment as a non-indigenous, imperialist export (Cahill 2000). In France, as Saguy shows, French feminists won a law that criminalizes the use of 'orders, threats, constraint or serious pressure in the goal of obtaining sexual favors, by someone abusing the authority conferred by his position' (Saguy 2000, p. 19 and note 34). This approach incorporates the American cultural-radical feminist view of harassment as a form of sexual abuse yet simultaneously signals distance from what the French perceive to be US sexual prudishness by highlighting that harassment is an abuse of hierarchical authority—an idea that conforms to conventional French views of hierarchical power as inimical to equality. Neither French feminists nor lawmakers connect harassment to a larger system of workplace gender inequality that relegates women to inferior jobs; the gender segregation of work is accepted as a 'neutral' background condition rather than challenged as the structural context of inequality in which sex harassment flourishes and which it fosters. Friedman's work suggests that even in Germany, where there is a tradition of using law to promote workers' empowerment, transplanting the US model of harassment-as-sexual overtures serve to exacerbate what German feminists have decried as a conservative judicial tendency to define sex harassment as a violation of female sexual honor which requires moral purity (Friedman 2000).

It is perhaps ironic that just as German (and other) feminists are trying to highlight the specificity of sexual harassment, a new generation of US feminists are striving to incorporate sexual harassment into broader frameworks for understanding the dynamics of ingroup/outgroup exclusion among different groups of workers—a project that might well be assisted by expanding the German concept of mobbing as a pervasive pattern of workplace harassment that targets an individual for exclusion and abuse at the hands of coworkers and supervisors' (Friedman 2000,

p. 6). What is needed is a structural analysis of how a variety of different forms of harassment—including sexual advances, gender-based and other forms of taunting, physical threats, verbal denigration, work sabotage, heightened surveillance, and social isolation—can be used by socially dominant groups to code and to claim scarce social resources (such as good jobs and a sense of entitlement to them) to the exclusion of others. In this project, Americans have as much to learn from other countries' experiences as vice versa.

It is cause for optimism that so many scholars, activists and policymakers are struggling toward such a broader—and deeper—understanding of harassment. Inspired by new research, new feminisms, and newer social movements (such as queer theory) that are calling into question the old top-down, male-female sexual model, the law of sex harassment awaits being overhauled to fit the world of the twenty-first century.

See also: Autonomy at Work; Gender and the Law; Gender Differences in Personality and Social Behavior; Heterosexism and Homophobia; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Male Dominance; Organizations: Authority and Power; Psychological Climate in the Work Setting; Rape and Sexual Coercion; Regulation: Sexual Behavior; Sex Differences in Pay; Sex Segregation at Work; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Harassment: Social and Psychological Issues; Sexuality and Gender; Workplace Environmental Psychology

Bibliography

Abrams K 1998 The new jurisprudence of sexual harassment. *Cornell Law Review* **83**: 1169–1230

Arvey Richard D, Cavanaugh M A 1995 Using surveys to assess the prevalence of sexual harassment: Some methodological problems. *Journal of Social Issues* **51**(1): 39–50

Austin R 1988 Employer abuse, worker resistance, and the tort of intentional infliction of emotional distress. *Stanford Law Review* **41**: 1–59

Bernstein A 1997 Treating sexual harassment with respect. *Harvard Law Review* **111**: 45–527

Borgida E, Fiske S T (eds.) 1995 Special issue: gender stereotyping, sexual harassment, and the law. *Journal of Social Issues* **51**(1): 1–207

Brodsky C 1976 *The Harassed Worker*. Lexington Books, Lexington, MA

Burlington Industries vs. Ellerth. 1998. 524 US 742

Cahill M 2000 The legal problem of sexual harassment and its international diffusion: A case study of Austrian sexual harassment law. (unpublished manuscript)

Crull P 1987 Searching for the causes of sexual harassment: An examination of two prototypes. In: Bose C, Feldberg R, Sokoloff N (eds.) *Hidden Aspects of Women's Work*. Praeger Press, New York, pp. 225–44

Ehrenreich R 1999 Dignity and discrimination: Toward a pluralistic understanding of workplace harassment. *Georgetown Law Journal* **88**: 1–64

Faragher vs. City of Boca Raton. 1998. 524 US 775

Farley Lin 1978 *Sexual Shakedown: The Sexual Harassment of Women on the Job*. McGraw-Hill, New York

Fiske T 1993 Controlling other people: The impact of power on stereotyping. *American Psychologist* **48**: 621–8

Fiske S T, Glick P 1995 Ambivalence and stereotypes cause sexual harassment: A theory with implications for organizational change. *Journal of Social Issues* **51**: 97–115

Fitzgerald L F, et al. 1988 The incidence and dimensions of sexual harassment in academia and the workplace. *Journal of Vocational Behavior* **32**: 152–75

Frank E et al. 1998 Prevalence and correlates of harassment among US women physicians. *Archives Internal Medicine 1998* **158**: 352–8

Franke K 1998 What's wrong with sexual harassment? *Stanford Law Review* **49**: 691–772

Friedman G 2000 Dignity at work: workplace harassment in Germany and the United States. (unpublished manuscript).

Goldsmith E 1999 God's House or the Law's. *Yale Law Journal* **108**: 1433–40

Grimsley K D 1996 In combating sexual harassment, companies sometimes overreact. *Washington Post* (Dec. 23): A1

Gutek B 1992 Understanding sexual harassment at work. *Notre Dame Journal of Law, Ethics & Public Policy* **6**: 335–58

Hager M 1998 Harassment as a tort: Why title VII hostile environment liability should be curtailed. *Connecticut Law Review* **30**: 375–439

Halley J 2000 Sexuality harassment. (unpublished manuscript).

Kanter R M 1997 *Men and Women of the Corporation*. Basic Books, New York

Lancaster J 1991 Navy 'gauntlet' probed: Sex harassment alleged at fliers' convention. *Washington Post* (Oct. 30): A1

Lindbeck A, Snower D 1988 *The Insider-Outsider Theory of Employment and Unemployment*. MIT Press, Cambridge, MA

MacKinnon C 1979 *The Sexual Harassment of Working Women*. Yale University Press, New Haven, CT

Mayer J, Abramson J 1994 *Strange Justice: the Selling of Clarence Thomas*. Houghton Mifflin, Boston

Meritor Federal Savings Bank vs. Vinson. 1986. 477 US 57.

Oncale vs. Sundowner Offshore Services, Inc. 1998. 523 US 75.

Posner R 1989 An economic analysis of sex discrimination laws. *University of Chicago Law Review* **56**: 1311–35

Pryor J B et al. 1995 A social psychological model for predicting sexual harassment. *Journal of Social Issues* **51**: 69–84

Reskin B 2000 The proximate causes of employment discrimination. *Contemporary Sociology* **29**: 319–28

Rosen J 1998 In defense of gender blindness. *The New Republic* **218**: 25–35

Saguy A C 2000 Sexual harassment in France and the United States: Activists and public figures defend their definitions. In: Lamont M, Thevenot L (eds.) *Rethinking Comparative Cultural Sociology: Polities and Repertoires of Evaluation in France and the United States*. Cambridge University Press, Cambridge, UK, pp. 56–93

Schuldt Gretchen 1997 Ex-Miller exec. copied page with anatomical word: man suing over firing says he showed copy of female co-worker. *Milwaukee Sentinel* (June 26), 1

Schultz V 1991 Telling stories about women and work: Judicial interpretations of sex segregation in Title VII Cases raising the lack of interest argument. *Harvard Law Review* **103**: 1749–843

Schultz V 1998 Reconceptualizing sexual harassment. *Yale Law Journal* **107**: 1683–1804

Strossen N 1995 *Defending Pornography: Free Speech, Sex, and the Fight for Women's Rights*. Anchor Books, New York
Volokh E 1992 Freedom of speech and workplace harassment. *UCLA Law Review* **39**: 1791–872
Welsh S 1999 Gender and sexual harassment. *Annual Review of Sociology* **25**: 169–90
White R H 1999 There's nothing special about sex: The supreme court mainstreams sexual harassment. *William & Mary Bill of Rights Journal* **7**: 725–53
Willis E 1992 Radical feminism and feminist radicalism. In: Wills E (ed.) *No More Nice Girls: Countercultural Essays*, Wesleyan University Press, London, 117–50
Yoshino K 2000 The epistemic contract of bisexual erasure. *Stanford Law Review* **52**: 353–461

V. Schultz and E. Goldsmith

Sexual Harassment: Social and Psychological Issues

Sexual harassment is generally defined as unwelcome sexual advances, requests for sexual favors, or other verbal or physical conduct of a sexual nature that is either a condition of work or is severe and pervasive enough to interfere with work performance or to create a hostile, intimidating work environment. It may consist of words, gestures, touching, or the presence of sexual material in the work environment. It typically involves a pattern of behavior over a period of time, rather than a single event. We might think about the former as 'an episode of sexual harassment.' In perhaps 90 percent of the episodes, women are the recipients and men are the initiators, but both sexes can harass and both can be harassed by the same sex or the other sex. If sexual harassment meets certain criteria (e.g., unwelcome, severe, and pervasive), it is illegal in many countries, but not all behavior commonly considered sexual harassment violates the law.

1. A Short History of Research on Sexual Harassment

It is no doubt safe to assume that sexual harassment has been around for a long time, but it has been labeled, studied, and legislated for only about 20 years. In 1978, journalist, Lin Farley wrote *Sexual Shakedown* to bring attention to the phenomenon. In 1979, legal scholar, Catharine MacKinnon wrote an influential book that would provide a legal framework for dealing with sexual harassment in the US. MacKinnon argued that sexual harassment was a form of sex discrimination (i.e., denies women equal opportunity in the workplace) and therefore Title VII of the 1964 Civil Rights Act, which forbids discrimination on the basis of sex (among other social categories), should apply. A year after her book was published, the US Equal Employment Opportunity Commission established guidelines on sexual harassment. Early empirical studies of sexual harassment in the workplace and academia started appearing in print about the same time. By 1982, at least one journal (*Journal of Social Issues*) had produced a whole issue devoted to scholarship on the topic. Today sexual harassment is studied by scholars in many countries who work in many fields, including law, psychology (clinical, forensic, organizational, social), sociology, management and human resources, history, anthropology, communication, and the humanities. Within the social sciences, sexual harassment is studied through quantitative techniques that focus on the measurement of constructs and determination of base rate statistics and qualitative case-study techniques focusing on specific occupations such as wait staff and female coal miners.

2. Measurement of Sexual Harassment

In the early 1980s, researchers frequently used the term 'social-sexual behaviors' to distinguish a set of behaviors that might constitute sexual harassment from a legally defined measure of sexual harassment. These sets typically included behaviors unlikely to be considered sexual harassment either under the law or by a majority of the population. By including a broad range of behaviors, researchers could learn whether people's views of specific behaviors differ over time (or across samples). It would also allow researchers to see if legal and illegal social-sexual behaviors have common antecedents and consequences.

More recently, however, sexual harassment is what researchers say they are measuring. This has caused some confusion because many people seem to interpret statistics on sexual harassment to indicate the percentage of the workforce that would have a strong legal claim of sexual harassment. This is not true. Researchers have not attempted to capture the legal definition of harassment because: (a) the legal definition changes as the law develops, so the legal definition is a moving target; (b) laws vary from country to country; (c) targets may experience negative consequences of sexual harassment without having the harassment rise to meet a legal definition; (d) there is no reason to believe that we can learn about sexual harassment only by measuring it to conform to its legal definition. Some scholars now make explicit the point that a lay definition of sexual harassment does not necessarily imply that a law has been broken.

A global item, say, 'Have you ever been sexually harassed,' is rarely used to measure sexual harassment because some researchers contend that it results in an under-reporting of the phenomenon. Workers seem reluctant to acknowledge that they have been sexually

harassed. In addition, asking respondents if they have been sexually harassed places a great cognitive load on them, as they would have to determine first what constitutes sexual harassment and then determine if they had experienced any behavior that met those criteria. In many studies, a single question asking the respondent if she has been sexually harassed is used as an indicator of acknowledging or labeling sexual harassment rather than an indicator of sexual harassment, *per se*.

Most studies measure sexual harassment by asking respondents if they have experienced any of a list of behaviors that might be considered sexual harassment. These measures are generally (but not always) designed to be suitable for both sexes. In some cases, respondents are asked whether they have experienced a list of behaviors that might be considered sexual harassment, and later in the survey asked which of those behaviors they consider sexual harassment, allowing the researcher to determine which of a broad range of behaviors respondents have experienced that they consider sexual harassment.

Multi-item measures of sexual harassment are also based on a list of behaviors that people may have experienced. The best known of these measures is the Sexual Experiences Questionnaire (SEQ) developed by Louise Fitzgerald and her colleagues. The SEQ has experienced many changes; the number of questions asked, the wording of questions, and the wording of responses have all been modified, and its refinement is ongoing. The number of subscales that emerge from it has also changed over time so it is important to keep in mind that all the studies using the SEQ are not necessarily using the same set of questions, scored the same way, and at this point it cannot be used to assess changes over time or differences across studies.

3. *The Prevalence of Sexual Harassment*

A number of studies have relied on random sample surveys including studies of government workers, the military, specific geographical areas, and specific work organizations. Some of these have obtained quite high response rates, some studies have been conducted in Spanish as well as in English, and a few longitudinal analyses have been conducted.

In measuring prevalence, there is some debate about the timeframe that should be considered. Researchers typically inquire about experiences within the past year, the prior two years, or throughout the person's entire work life, depending on the purpose of the study. While critics have cautioned that retrospective measures will introduce inaccuracy or bias, researchers active in the field are less concerned. It is difficult to know if one is currently being sexually harassed and qualitative studies provide convincing examples of events initially not labeled sexual harassment that came to be so labeled at a later date.

3.1 *The Prevalence of Sexual Harassment of Women*

The random-sample studies together suggest that from about 35 percent to 50 percent of women have been sexually harassed at some point in their working lives, where sexual harassment refers to behavior that most people consider sexual harassment. Estimates are higher among certain groups such as women who work in male-dominated occupations. The most commonly reported social-sexual behaviors are the less severe ones, involving sexist or sexual comments, undue attention or body language. Sexual coercion is, fortunately, much rarer, involving 1–3 percent of many samples of women. In contrast, a study of allegations in cases tried in court showed a much higher incidence of severe behaviors, with 22 percent involving physical assault, 58 percent nonviolent physical contact, and 18 percent violent physical contact.

The incidence of sexual harassment of women appears to be rather stable. Three US Merit Systems Protections Board studies spanning 14 years show that 42–44 percent of women in the federal workforce have experienced one or more of a list of potentially sexually-harassing behaviors within the previous 24 months. In addition, the number of charges filed with the US Equal Employment Opportunity Commission may be leveling off in the range of 15,000–16,000 per year (in a labor force of about 120 million people).

3.2 *The Prevalence of Sexual Harassment of Men*

Men have been included in studies of sexual harassment from the very beginning. In her study of a random sample of working men and women in Los Angeles County, Barbara Gutek found that some time during their working lives, from 9 percent to 35 percent of men (depending on definition of harassment) had experienced some behavior initiated by one or more women that they considered sexual harassment. The US Merit Systems Protection Board studies found that from 14 percent to 19 percent of men in the Federal workforce experienced at least one episode of a sexually harassing experience (initiated by either men or women) within the previous two years. These surveys revealed that about one-fifth of the harassed men were harassed by another man. Data from the 1988 US Department of Defense Survey of Sex Roles in the Active Duty Military revealed that about one-third of the men (but about 1 percent of the women) who experienced at least one of nine types of uninvited, unwanted sexual attention during the previous 12 months, reported that the initiator was the same sex. The harassment of men by other men tends to be of two types: lewd comments that were considered offensive and attempts to enforce male gender role behaviors. Harassment by women is somewhat different, consisting of negative remarks and/or unwanted sexual attention.

The available research on sexual harassment of men, admittedly much less than the research on women, suggests that many of the behaviors women might find offensive are not considered offensive to men when the initiators are women and/or they report few negative consequences. In addition, a disproportionate percentage of men's most distressing experiences of sexual harassment come from other men. Presumably these are especially distressing because the recipient's masculinity and/or sexual orientation are being called into question.

3.3 The Prevalence of Sexual Harassment Among Other Groups

Relatively few studies have either focused on or found consistent differences in the experience of sexual harassment beyond the consistent differential rate of sexual harassment of women vs. men. It may be the case that younger and unmarried women are somewhat more likely targets of sexual harassment than older and married women. Lesbian women may be more likely to be sexually harassed than heterosexual women, or they may simply be more likely to label their experiences sexual harassment. Although several authors have suggested that in the US women of color (Asian, African-American, Hispanic, and American Indian) are more likely to be targets of sexual harassment than Caucasian women, the evidence is far from clear. Several random sample surveys found no clear link between ethnicity and the experience of sexual harassment but qualitative studies suggest that women of color experience sexual harassment frequently. Two kinds of arguments have been advanced for reasons why minority women might experience relatively more sexual harassment. A direct argument relies on stereotyping of minorities. Although the stereotypes of African-American women differ from stereotypes of Chicanas or Asian-American women, in each case the stereotype might place these women at greater risk. An indirect argument relies on concepts of power and marginality. As women of color are less powerful and more marginal by virtue of their ethnicity than white women, they may be more prone to sexual harassment.

4. Explanations for Sexual Harassment

Why sexual harassment exists has been of interest to social scientists since the phenomenon acquired a label. The various explanations can be subsumed into four categories: natural/biological perspectives, organizational perspectives, sociocultural explanations, and individual differences perspectives. These explanations tend to be broad in scope, not easily testable in a laboratory.

4.1 Natural/Biological Explanations

There are two natural/biological perspectives: a hormonal model and an adaptive/evolutionary explanation. While intriguing, neither is supported by available data.

4.2 Organizational Explanations

There are two organizational perspectives: sex-role spillover and organizational power. Sex-role spillover, defined as the carry over into the workplace of gender-based expectations that are irrelevant or inappropriate to work, occurs because gender role is more salient than work role and because under many circumstances, men and women fall back on sex role stereotypes to define how to behave and how to treat those of the other sex. Sex-role spillover tends to occur most often when the gender ratio is heavily skewed in either direction, i.e., when the job is held predominantly either by men or by women. In the first situation (predominantly male), nontraditionally employed women are treated differently than their more numerous male co-workers, are aware of that different treatment, report relatively frequent social-sexual behavior at work, and tend to see sexual harassment as a problem. In the second situation (predominantly female), female workers hold jobs that take on aspects of the female sex-role and where one of those aspects is sex object (e.g., cocktail waitress, some receptionists), women may become targets of unwanted sexual attention, but may attribute the way they are treated to their job, not their gender. Several studies find some support for this perspective.

The earliest writings on sexual harassment were about men abusing the power that comes from their positions in organizations to coerce or intimidate subordinate women. Some subsequent statements on the power perspective are gender neutral, suggesting that although men tend to harass women, in principle if women occupied more positions of power, they might harass men in equal measure. This interpretation of sexual harassment as an abuse of organizational power is contraindicated by research showing that about half or more of harassment comes from peers. In addition, both customers and subordinates are also sources of harassment. Sexual harassment by subordinates has been documented primarily in academic settings where, for example, studies find up to half of female faculty at universities had experienced one or more sexually harassing behaviors by male students.

While power cannot explain all sexual harassment, various kinds of power—formal organizational power and informal power stemming from ability to influence—remain potent explanations for at least some sexual harassment. For example, the fact that sexual harassment by customers is fairly common can be

explained, at least in part, by the emphasis employers place on customer satisfaction and the notion that the 'customer is always right.'

Some researchers focus less on broad theoretical perspectives and more on the types of organizational factors that need to be included in models of sexual harassment. Thus far, contact with the other sex and a unprofessional and/or sexualized work environment have been identified as correlates of sexual harassment.

4.3 Sociocultural Explanations

There are at least two ways of thinking of the broader sociocultural context. One is that behavior at work is merely an extension of male dominance that thrives in the larger society. Overall, there is general agreement in the literature about the characteristics of the sex stratification system and the socialization patterns that maintain it. The exaggeration of these roles can lead to sexual harassment. For example, men can sexually harass women when they are overly exuberant in pursuing sexual self-interest at work, or they feel entitled to treat women as sex-objects, or when they feel superior to women and express their superiority by berating and belittling the female sex.

The second way of thinking of the broader sociocultural context is to study the sociocultural system itself and examine how and why status is assigned. According to this view, sexual harassment is an organizing principle of our system of heterosexuality, rather than the consequence of systematic deviance.

4.4 Individual Difference Explanations

Although the data suggest that most sexual harassers are men, most men (and women) are not sexual harassers. This makes the study of personality characteristics particularly relevant. The search for individual-level characteristics of perpetrators does not negate any of the other explanations, but helps to determine, for example, which men in a male-dominated society or which men in powerful positions in organizations harass women when most men do not.

John Pryor developed a measure of the Likelihood to Sexually Harass (LSH) in men, consisting of 10 vignettes that place the respondent in a position to grant someone a job benefit in exchange for sexual favors. Currently the most widely known individual difference measure used in the study of sexual harassment, the LSH, has been validated in a number of studies. For example, undergraduate men who score relatively high on the LSH demonstrate more sexual behavior in a lab experiment and hold more negative attitudes toward women relative to those who report a lower likelihood to sexually harass.

Sexual harassment may be an attempt to immediately gratify the desire for discrimination, intimidation, or sexual pleasure and therefore those people with low self-control may also be more likely to sexually harass. Some research findings support this gender-neutral explanation.

5. Factors Affecting Judgments of Sexual Harassment

The most widely published area of research on sexual harassment measures people's perceptions about sexual harassment. One set of studies attempts to understand which specific behaviors (e.g., repeated requests for a date, sexual touching, stares or glances, a sexually oriented joke) respondents consider to be sexual harassment. The other set of studies attempts to understand factors that affect the way respondents perceive behavior that might be considered sexual harassment. Typically, respondents are asked to read a vignette in which factors are manipulated and then they are asked to make judgments about the behavior in the vignette. The factors that are manipulated include characteristics of the behavior (e.g., touching vs. comments), characteristics of the situation (e.g., the relationship between the initiator and recipient), and characteristics of the initiator and recipient (e.g., sex, age, attractiveness, occupation). In addition, characteristics of the rater (e.g., sex, age) are typically measured.

5.1 The Effects of Rater Sex on Judgments of Sexual Harassment

Sex is the most frequently studied feature in studies about perceptions of sexual harassment. In all, hundreds of studies have been done. These studies have been reviewed using traditional methods and meta-analyses. The conclusions of these reviews are two: (a) women consistently rate vignettes and specific behaviors more sexually harassing than men and (b) the average difference is small, raising questions about the practical significance of these results. Practical significance is important because in the United States it would appear that these studies have had an indirect influence on the Ninth Circuit's 1991 decision, *Ellison vs. Brady*. In that case, the court adopted a new legal standard in hostile environment cases of sexual harassment, the reasonable woman standard, which replaces the traditional reasonable person standard in that Circuit. Juries are asked to evaluate the events from the perspective of a reasonable woman: taking into account all the facts, would a reasonable woman consider the plaintiff to be sexually harassed? The predominant claim is the new standard would force judges and juries to look at the case from the perspective of the complainant, who is typically a woman. This would presumably make it less difficult for a plaintiff to make a convincing claim of hostile work environment harassment. While some recent research suggests that a reasonable woman standard as currently implemented may not result in different

judgments than the traditional reasonable person standard, the magnitude of the gender gap in perceptions about sexual harassment may not justify a change in standards, regardless of the effect of standard itself.

5.2 *The Effects of Other Rater Characteristics on Judgments of Sexual Harassment*

Although gender is far and away the most widely studied factor in the study of sexual harassment perceptions, other factors have been studied. For example, when the initiator is higher status than the recipient, judges generally respond more positively toward the recipient, more negatively toward the initiator, and perceive more harassment than when the initiator is not a supervisor. In addition, several studies that compare students with workers find that students have a broader, more lenient view of social-sexual behavior relative to samples of workers who are typically somewhat older and have more work experience.

6. *Some Remaining Issues*

While much is known, we lack a complete picture of sexual harassment. Must concern about sexual harassment eliminate any kind of dating flirtation at work? How responsible should employers be for the behavior of their employees? How should targets of harassing behavior respond in order to eliminate harassment without damaging their own career possibilities?

See also: Gender and Place; Male Dominance; Sexual Attitudes and Behavior; Sexual Harassment: Legal Perspectives; Workplace Safety and Health

Bibliography

Bowes-Sperry L, Tata J 1999 A multiperspective framework of sexual harassment. In: Powell G N (ed.) *Handbook of Gender and Work*. Sage Publications, Inc., Thousand Oaks, CA pp. 263–80

Brewer M B, Berk R A (eds.) 1982 Beyond Nine to Five: Sexual harassment on the job. *Journal of Social Issues* **38**(4): 1–4

Estrich S 1991 Sex at work. *Stanford Law Review* **43**: 813–61

Borgida E, Fiske S T 1995 (eds.) Gender stereotyping, sexual harassment, and the law. *Journal of Social Issues* **51**(1)

Franke K M 1997 What's wrong with sexual harassment? *Stanford Law Review* **49**: 691–772

Gutek B A 1985 *Sex and the Workplace*. Jossey-Bass Publishers, San Francisco

Gutek B A, Done R 2001 Sexual harassment. In: Unger R K (ed.) *Handbook of the Psychology of Women and Gender*. Wiley, New York

Harris V. 1993 Forklift Systems, 114 S. Ct. 367

MacKinnon C A 1979 *Sexual Harassment of Working Women: A Case of Sex Discrimination*. Yale University Press, New Haven, CT

O'Donohue W (ed.) 1997 *Sexual Harassment: Theory, Research, and Treatment*. Allyn and Bacon, Boston

Pryor J B, McKinney K(Issue eds.) 1995 Special Issue: Research advances in sexual harassment. *Basic and Applied Social Psychology* **17**(4)

Stockdale M S 1996 *Sexual Harassment in the Workplace: Perspectives, Frontiers, and Response Strategies*. Sage Publications, Inc., Newbury Park, CA

Welsh S 1999 Gender and sexual harassment. *Annual Review of Sociology* **25**: 169–90

Williams C L, Giuffe P A, Dellinger K 1999 Sexuality in the workplace: Organizational control, sexual harassment, and the pursuit of pleasure. *Annual Review of Sociology* **25**: 73–93

B. A. Gutek

Sexual Orientation and the Law

Sexual orientation refers to any classification based on sexual desire or conduct, such as heterosexuality, same-sex sexuality, or bisexuality. It is a classification that includes all sexual orientations, just as race is a classification that includes all races (White, Black, Asian, etc.) However, just as discussions of race tend to focus on marginalized racial groups (such as people of color), discussions of sexual orientation commonly refer to minority sexual orientations, gay, lesbian, or bisexual. As a result, this article will reflect that focus on gay people. It will, however, also note the growing literature on heterosexuality and transgender people.

1. *Homosexuality and Heterosexuality as Historically Contingent*

The study of the law relating to sexually marginalized people, also known as 'queer legal theory,' evolved in the 1980s and 1990s out of feminist legal theory, postmodern theory, and critical race theory. The most influential work informing queer legal theory is Michel Foucault's *History of Sexuality, Vol. 1* (Foucault 1978), which posited that sexual orientation is socially constructed rather than naturally or divinely ordained. This point has been further developed by Judith Butler's performativity theory of sexual orientation and gender, which posits that identities are performed rather than biologically or divinely ordained (Butler 1990).

1.1 *From Conduct to Status and Back to Conduct*

The social construction of sexual orientation is revealed by the different ways in which society and law have viewed same-sex sexuality. Different terms used

to describe same-sex sexuality reflect these changes. Prior to the late nineteenth century, same-sex sexuality was viewed as sinful conduct, redeemable through repentance. The terms 'homosexual' and 'heterosexual' did not appear in English until 1892, along with the idea that sexual orientation was a status. These terms are products of German and English medical research on same-sex sexuality, which saw same-sex sexuality as a status, gave it a name, contrasted it with the previously unmarked status of heterosexuality, and assigned it social meaning as a constellation of characteristics that a person is (rather than particular sexual acts that one does). The terms 'invert' or 'homosexual' refer to this medicalized understanding of same-sex sexuality as a sickness rather than sinful conduct. Oscar Wilde's 1895 trials and imprisonment for gross indecency are credited with introducing into popular culture the status of the male homosexual as effete, artistic, self-centered, and sybaritic.

A third stage, beginning in the 1940s and 1950s, relied on the scientific research of Alfred Kinsey and Evelyn Hooker to contend that same-sex sexuality was a normal variation from opposite-sex sexuality, rather than a disease (Hooker 1957, Kinsey et al. 1948). The nascent homophile movement adopted the term 'gay' to distance itself from the medical judgment associated with 'homosexual.' A fourth stage, emerging through queer theory in the 1990s, rejected the very notion of sexual orientation status (evil or neutral), pointing out the indeterminate and potentially subordinating qualities associated with status-based notions of identity. This stage dubbed minority sexual orientations 'queer' (reclaiming the epithet), treating this anti-status as an epistemological category. To be queer is to believe that racial, sexual, and sexual orientation subordination is wrong. Queer theory arguably uncouples status from conduct, so that one person can be both queer and actively heterosexual.

1.2 Fluidity Between Sexual Orientation and Gender Identity

Sexual orientation and gender issues overlap, but are also distinct in important ways. Both are associated with gender nonconformity. They differ, however, in their fundamental premises. Conventional approaches to sexual orientation presuppose two sexes, and categorize a woman as lesbian if she is with another woman (homosexual literally meaning same-sex), and heterosexual if she is with a man (heterosexual literally meaning different-sex). Transgender theory, in contrast, posits that there are more than two sexes, and that people exist along a continuum of masculinities and femininities. This reasoning renders gay theory's primary assumption absurd; to speak of a person being gay or heterosexual ceases to make sense when there are many more than the two options of being a woman-with-a-woman or woman-with-a-man.

'Transgender' generally refers to gender norm transgression. Some transgender people, known as transsexuals, undergo medical treatment to change from one sex to another. Other categories of transgender people include transvestites (people who cross-dress but do not undergo medical gender reassignment), and transgendered people (those who do not conform to gender norms for the sex they were assigned at birth, but neither do they undergo medical treatment to reassign their sex).

Some cultures do not draw bright lines to distinguish same-sex sexuality from gender identity issues. In Bolivia, for example, men who engage in same-sex sexuality do so within a cultural understanding that one of the participants, in some sense, is female. Men who engage in same-sex sexuality call themselves *gente de ambiente*, or people of the atmosphere (West and Green 1997). Three subcategories of *gente de ambiente* highlight the overlap between sexual orientation and transgender identity in Bolivian culture. *Travestis*, or transvestites, view themselves as women trapped in men's bodies. *Camuflados*, the camouflaged, see themselves as men, but take the receptive role in intercourse, and often pass as heterosexual in public spaces. The third category, *hombres*, or just men, take the penetrating role in intercourse, and are distinguished from heterosexual men in that they respond to sexual overtures from other men. In Bolivia, and elsewhere, these sexual orientation and gender categories are fluid. A man might sleep with men exclusively, only occasionally, or for money. He might desire to sleep with men, but refrain for fear of social or legal condemnation. The very difficulty in deciding the characteristics of identity categories forms the thesis of much postmodern scholarship that identity generally, and sexual identity in particular, is indeterminate because it varies historically, culturally, and within a particular individual's lifetime.

2. Changing Focus of Sexual Orientation Research

Sexual orientation law is a synergy of gay rights advocacy and theoretical academic writings. Advocates bring cases and support legislation countering anti-gay discrimination. Scholars both craft theories that inform these efforts and evaluate the cases and statutes that become law. Two important areas involve anti-sodomy laws and the ban on same-sex marriage.

2.1 Anti-sodomy Laws

Because the criminalization of same-sex sexuality is the most obvious expression of state condemnation, sexual orientation law began by challenging anti-sodomy laws. In the US case *Bowers* vs. *Hardwick*, 478 US 186 (1986), the US Supreme Court upheld Geor-

gia's statute criminalizing sodomy against a constitutional privacy challenge. The Court issued extraordinarily homophobic declarations, such as the concurring opinion's citation of Blackstone (1859) to describe same-sex sodomy as 'an offense of "deeper malignity" than rape, a heinous act "the very mention of which is a disgrace to human nature."' The gratuitous nastiness inspired both numerous critiques of the decision (Goldstein 1988, Thomas 1992) and challenges to other laws disadvantaging gay people.

One strand of this critique challenges the validity of anti-sodomy statutes, arguing that sodomy is an historically contingent category. For example, early Roman–Dutch law (imported to Colonial South Africa) interpreted sodomy as including a wide range of nonconforming sexual acts, such as anal penetration, bestiality, masturbation, oral penetration, penetration with an inanimate object, interfemoral intercourse, and heterosexual intercourse between a Jew and a Christian (West and Green 1997, p. 6). This insight undermines the logic of sodomy law defenders, who, along with the US Supreme Court in *Bowers* vs. *Hardwick*, rest their defense on 'millennia of moral teaching' (Goldstein 1988). If what counts as reprehensible conduct differs markedly among people, places, and times, theorists reason, then one cannot invoke a uniform condemnation of the conduct.

While many governments have decriminalized same-sex sexuality, the criminal ban remains strong in some places. In 2000 in Malaysia, for example, former deputy Prime Minister Anwar Ibrahim was convicted of sodomy and sentenced to nine years in prison after a 14-month trial. At the opposite extreme, the 1996 South African Constitution explicitly forbade sexual orientation discrimination. The ban on gays in the military, like anti-sodomy laws, excludes gay people from full citizenship. While the ban remains in the US, other countries, such as Israel, allow gay people to serve in the military.

2.2 The Ban on Same-sex Marriage

Although anti-sodomy laws are rarely enforced, their existence impedes other anti-discrimination claims, such as gay people's rights to marry or to serve in the military. Since the 1980s, many countries (i.e., South Africa) and US states (i.e., Georgia) have decriminalized same-sex sexuality. Legal advocacy and scholarship has moved on to contest discrimination in other areas. In *Romer* vs. *Evans*, 517, US 620 (1996), the US Supreme Court invalidated a state constitutional amendment that forbade any state entity from protecting gay, lesbian, or bisexual people from discrimination. Advocates and legal scholars, however, have paid far more attention to marriage litigation. In 2001, the Netherlands became the first country to allow same-sex couples to marry. Other countries such as Denmark, France, Greenland, Hungary, Iceland, Norway and Sweden each recognize some type of partnership, which accords same-sex couples (and sometimes opposite-sex unmarried couples) various benefits accorded to married couples.

Parts of Canada and Spain also recognize same-sex partnerships, and domestic partners enjoy some employment benefits in Argentina, Canada, Israel, New Zealand, and South Africa. Other countries, such as Australia, Austria, Belgium, Brazil, the Czech Republic, Germany, Portugal, Spain, and the UK recognize narrower rights, such as succession rights in private housing and some inheritance rights. Australia, New Zealand, Belgium, Denmark, Finland, Germany, Iceland, The Netherlands, Norway, South Africa, Sweden, and the UK recognize same-sex relationships for immigration purposes. Smaller governmental units, such as cities, counties, and provinces around the world, ban discrimination on the basis of sexual orientation.

In the USA, only one state, Vermont, accords same-sex relationships recognition akin to marriage; the federal government and a majority of the states explicitly refuse to recognize same-sex marriage. The Vermont Supreme Court in *Baker* vs. *Vermont*, 744 A.2d 864 [Vt. 1999] found that banning same-sex marriage (or its equivalent under another name) violated the Common Benefits Clause of the Vermont constitution. That clause provides that 'government is, or ought to be, instituted for the common benefit, protection, and security of the people, nation, or community, and not the particular emolument or advantage of any single person, family, or set of persons, who are only a part of that community.' Following the court's instructions, the Vermont legislature created a structure parallel to marriage for same-sex couples, calling it civil union.

3. Theoretical Approaches to Sexual Orientation Law

The emphasis in current theory and research on sexual orientation law is to deconstruct and reconstruct legal regulations that subordinate gay and transgendered people. This rubric includes various ideological and methodological approaches.

3.1 Ideology

Ideological diversity includes both classical liberal and critical approaches. Liberal approaches suggest that gay people are sufficiently similar to heterosexuals to justify equal legal treatment. This approach suggests that altering legal rules to lift the ban on gay participation in marriage or the military will not fundamentally change these institutions. A critical approach, in contrast, contests essentialized notions of identity, seeing it as socially constructed rather than

reflecting essential commonalties among people who engage in same-sex sexuality. This critical approach often seeks a more comprehensive restructuring of legal regulations than simply adding gay people to existing institutions. Instead, critical approaches propose incorporating sexual orientation analysis into other antisubordination discourses, such as feminism, critical race studies, and class analysis. Critical theorists might further propose alternatives to marriage, such as domestic partnership for all people, contending that only new institutions can alleviate the sex/gender hierarchies inherent in marriage.

3.2 Methodology

Postmodern, empirical, and legal economic approaches have contributed to the literature on sexual orientation law. One empirical approach compiles data about rates of arrest and prosecution for sexual offenses, and points out that consensual same-sex sexual activity is unfairly singled out for criminal penalty (Eskridge 1999). The two major premises of this approach, that like parties should be treated alike, and further that the state should generally refrain from interfering in private consensual activities, together lead to the conclusion that the state statutes criminalizing sodomy should be invalidated.

One postmodern approach closely examines the language and reasoning of a judicial opinion to decode the cultural context of the decision, revealing, for example, the Supreme Court's strategic and biased collapse of status and conduct in *Bowers* vs. *Hardwick*. This approach focuses on how the majority decision in *Hardwick* collapsed the distinction between status and conduct, and ignored the many cross-sex couples who commit sodomy by engaging in acts such as fellatio, cunnilingus, and anal intercourse. The majority in *Hardwick* framed the question as 'whether the Federal Constitution confers a fundamental right upon homosexuals to engage in sodomy' focusing the legal inquiry on both a bad act (sodomy) and a bad status (homosexuality). Under this reasoning:

> [I]f sodomy is bad … then homosexuals and heterosexuals who do it are bad. If homosexuals are bad, we are bad whether we've engaged in sodomy or not. To hold both of these positions with consistency, you have to be willing to say that many, many heterosexuals are bad. *This* the majority Justices never acknowledged … They wanted the badness of each to contaminate the other—while heterosexual personhood remained out of the picture, protected from the taint with which it was logically involved. (Halley 1999, pp. 7–8)

By leaving the categories of sexual orientation status and conduct ambiguous, legal doctrine governing consensual sexuality 'remained always ready to focus on "act" or "status" according to the expediencies of the situation' (Halley 1999, p. 8)

An emerging strand of legal scholarship uses legal economic premises to examine heterosexuality as well as same-sex sexuality. US Court of Appeals Judge Richard Posner's book *Sex and Reason* (1992 posited a bioeconomic theory of sexuality that combined sociobiology with legal economics. Posner's often controversial analysis (describing, for example, the economic efficiency in some contexts of female infanticide and baby selling, which he renames 'parental-right selling') provoked a firestorm of response. Yet recent scholarship has built on Posner's economic approach to posit a bargaining theory of sexuality, suggesting that legal regulation should equalize the bargaining positions between men and women (Hirshman and Larson 1998).

4. Future Directions of Theory and Research

The youth of sexual orientation legal research makes it unpredictable. Four likely future trends emerge. First, future research may well include increased use of empirical methods such as compilations data from court records in criminal or family law cases. Second, queer legal theory's emphasis on post-identity analysis and legal doctrine's focus on identity as a foundational category require that future scholarship address the tensions between deconstructing identity and reconstructing a legal regime that does not discriminate on the basis of sexuality or gender performance. Third, scholars are likely to try to resolve the analytical tension between a binary construction of sex that underlies gay theory and a fluid construction of sex that underlies transgender approaches. Fourth and finally, future scholarship and advocacy may focus on the legal regulation of heterosexuality and bisexuality, further developing the literature on how legal regulation misconstrues identity as fixed or natural. On an ideological level, theoretical research is likely to continue to develop in both liberal (assimilationist) and critical (utopian) strands.

See also: Civil Rights; Family Law; Feminist Legal Theory; Gay/Lesbian Movements; Gender and the Law; Privacy: Legal Aspects; Queer Theory; Regulation: Sexual Behavior; Sex Segregation at Work; Sexual Attitudes and Behavior; Sexual Orientation: Historical and Social Construction; Social Class and Gender

Bibliography

Blackstone W 1859 *Commentories*. Harpe Brothers, New York, Vol. IV

Butler J 1990 *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York

Eskridge W 1999 *Gaylaw: Challenging the Apartheid of the Closet*. Harvard University Press, Cambridge, MA

Foucault M 1978 *The History of Sexuality, Vol. I: An Introduction*. Pantheon, New York
Goldstein A 1988 History, homosexuality and political values: Searching for the hidden determinants of Bowers v. Hardwick. *Yale Law Journal*. **97**: 1073–103
Halley J 1999 *Don't: A Reader's Guide to the Military's Anti-Gay Policy*. Duke University, Durham, NC
Hirshman L, Larson J 1998 *Hard Bargains: The Politics of Sex*. Oxford University Press, New York
Hooker E 1957 The adjustment of the male overt homosexual. *Journal of Projective Techniques* **21**: 18–31
International Lesbian and Gay Association, http://www.ilga.org/
Kinsey A, Pomeroy W, Martin C 1948 *Sexual Behavior in the Human Male*. W. B. Saunders, Philadelphia, PA
Posner R 1992 *Sex and Reason*. Harvard University Press, Cambridge, MA
Robson R 1992 *Lesbian (Out)law: Survival Under the Rule of Law*. Firebrand Books, Ithaca, NY
Rubinstein W 1997 *Sexual Orientation and the Law* (2nd edn.). West, St. Paul, MN
Symposium: InterSEXionality: Interdisciplinary perspectives on queering legal theory. 1998. *Denver University Law Review* **75**: 1129–464.
Thomas K 1992 Beyond the privacy principle. *Columbia Law Review* **92**: 1431–516
Valdes F 1995 Queers, sissies, dykes and tomboys: Deconstructing the conflation of 'sex,' 'gender,' and 'sexual orientation' in Euro-American law and society. *California Law Review* **83**: 1–377
West D, Green R 1997 *Sociolegal Control of Homosexuality: A Multinational Comparison*. Plenum Press, New York

M. M. Ertman

Sexual Orientation: Biological Influences

How humans develop sexual orientations is a question that people have pondered for at least a century. In the later part of the twentieth century, scientists began to articulate sophisticated biological theories of how sexual orientations develop. This article critically surveys contemporary biological theories of the development of sexual orientations. In particular, it examines three recent studies of how sexual orientations develop and their theoretical underpinnings. The focus is on these studies, because not only are they positively cited by almost every scientist trying to develop a biological theory of sexual orientation, but also because they are typical in their assumptions and methodology.

1. *What Is a Sexual Orientation?*

A person's sexual orientation concerns his or her sexual desires and fantasies towards others in virtue of their sex (gender). However, a person's sexual orientation is only one part of a person's sexual interest generally. People have a wide range of sexual tastes. Some are attracted to people of certain ages, people of certain body types, of certain races, of certain hair colors, of certain personality types, of certain professions, as well as to people of a certain sex and a certain sexual orientation. Further, people are not only sexually interested in certain sorts of people, some also have quite specific interests in certain sorts of sexual acts, certain venues for sex, or a certain frequency of having sex. We recognize that people can be sorted into all sorts of groups in virtue of their sexual interests, but most contemporary scientific studies focus only on the sex of the people a person is sexually attracted to as an essential feature about him or her. Doing so may be culturally salient but it is not scientifically justified.

2. *What Makes a Theory a Biological One?*

To say that sexual orientation is biologically based is an ambiguous claim; there are various senses in which it is trivially true that sexual orientation is biological. Everything psychological is biologically based. Humans can have sexual orientations while inanimate objects and one-celled organisms cannot because of our biological/psychological make-up. The same sort of claim is true with respect to having a favorite type of music: humans but not single-celled organisms can have a favorite type of music. Even though a preference for classical music seems a paradigmatic example of a learned trait, such a preference is also biological in that a certain cognitive complexity is required in order to have such a preference. Sexual orientation is at least biologically based in the sense that musical preferences are. The central claim of biological research on sexual orientation makes a much bolder claim. It says a person's sexual orientation is inborn or determined at a very early age and, as a result, a person's sexual orientation is 'wired into' his or her brain. To understand the significance of such a claim, it is useful to contrast three models of the role genes and other biological factors might play in sexual orientation (Byne 1996, Stein 1999, pp. 123–7).

According to the *permissive model*, genes or other biological factors influence neuroanatomical structures on which experience inscribes sexual orientation, but biological factors do not directly or indirectly determine sexual orientation. Something like the permissive model correctly describes the development of musical preferences. Various genetic and biological factors make it possible for our experiences with various kinds of music to shape our musical preferences.

Contrast the permissive model with the *direct model*, according to which genes, hormones, or other biological factors directly influence the brain structures that underlie sexual orientation. According to the

direct model, the neurological structures responsible for the direction of a person's sexual attraction toward men or women develop at a very early age as a result of a person's genes or other biological factors. One version of the direct model sees genes in the q28 region of the X chromosome as coding for a set of proteins that causes the INAH-3 region of the hypothalamus to develop so as to determine a person's sexual orientation (LeVay and Hamer 1994).

The direct and the permissive models can be contrasted with the *indirect model*, according to which genes code for (and/or other biological factors influence) temperamental or personality factors that shape how a person interacts with his or her environment and experiences of it, which, in turn, affects the development of his or her sexual orientation. On this view, the same gene (or set of genes) might predispose to homosexuality in some environments, to heterosexuality in others, and have no effect on sexual orientation in others. An example of such a theory is Daryl Bem's theory of sexual orientation according to which biological factors code for childhood personality types and temperaments (for example, aggressiveness, willingness to engage in physical contact, and so on) (Bem 1996, Peplau et al. 1998). In societies like ours where there are significantly different gender roles typically associated with men and women, these different personality and temperament types get molded into gender roles that, in turn, play a crucial role in the development of sexual orientation. The three biological theories of the development of sexual orientation discussed below accept the direct model. However, what evidence there is for these theories is equally consistent with the indirect model.

3. *Is Sexual Orientation Wired into the Brain?*

In 1991, Simon LeVay published a study of the size of INAH-3, a particular cell group in the hypothalamus (LeVay 1991). Starting from the assumption that there are neurological differences between men and women, LeVay decided to look for *sexual-orientation* differences in some of the areas of the hypothalamus that seem to exhibit sex differentiation. He reasoned as follows: given that most people who are primarily attracted to women are men and most people who are primarily attracted to men are women, in order to discover where sexual orientation is reflected in the brain, we should look in parts of the brain that are structured differently for men and women. This picture is based on seeing gay men as having female-typical characteristics and seeing lesbians as having male-typical characteristics. Seeing gay men and lesbians as gender inverts has a certain cultural salience but its scientific merit has been subject to serious criticism (Stein 1999, pp. 202–5, Byne 1994).

To examine the hypothalamus, LeVay had to study portions of human brain tissue that are accessible only after the person has died. Further, LeVay needed to know the sexual orientations of the people associated with the brain tissue he was studying. LeVay's study was made possible as a result of the AIDS epidemic, which had the result of making available brains from people whose self-reported sexual histories are to some extent part of their medical records. LeVay examined 41 brains: 19 of them from men LeVay presumed to be gay because they had died of complications due to AIDS and their medical records suggested that they had been exposed to HIV (the virus that causes AIDS) through sexual activity with other men; six of them from men of undetermined sexual orientation who also died of AIDS and who LeVay presumed were heterosexual; 10 of them from men of undetermined sexual orientation who died of causes other than AIDS and who were also presumed to be heterosexual; and six of them from women all of whom were presumed to be heterosexual, one whom died of AIDS and five who died from other causes. LeVay found that, on average, the INAH-3 of the presumed gay men were significantly smaller than those of the presumed heterosexual men and about the same size as those of the women. From this, he inferred that gay men's INAH-3 are in a sense 'feminized.' Although LeVay rather cautiously concluded that his 'results do not allow one to decide if the size of INAH-3 in an individual is the cause or consequence of that individual's sexual orientation or if the size of INAH-3 and sexual orientation co-vary under the influence of some third unidentified variable,' he also said that his study illustrates that 'sexual orientation in humans is amenable to study at the biological level' (LeVay 1991, p. 1036). In media interviews after the publication of his study he made even stronger claims; he said, for example, that the study 'opens the door to find the answer' to the question of 'what makes people gay or straight' (Gelman 1992).

4. *Is Sexual Orientation Inherited?*

Various studies suggest that sexual orientation runs in families (e.g., Pillard and Weinrich 1986). Such studies show that a same-sex sibling of a homosexual is more likely to be a homosexual than a same-sex sibling of a heterosexual is to be a homosexual; more simply, for example, the brother of a gay man is more likely to be gay than the brother of a straight man. These studies do not establish that sexual orientation is genetic because most siblings, in addition to sharing a significant percentage of their genes, share many environmental variables, that is, they are raised in the same house, are fed the same meals, attend the same schools and have many of the same adult role models. For these reasons, disentangling inherited and environmental influences requires more sophisticated studies.

Heritability studies done by Michael Bailey and Richard Pillard assessed sexual orientation in identical

twins, fraternal twins, nontwin biological siblings, and similarly-aged unrelated adopted siblings (Bailey and Pillard 1991, Bailey et al. 1993). If sexual orientation is genetic, then, first, all identical twins should have the same sexual orientation and, second, the rate of homosexuality among the adopted siblings should be equal to the rate of homosexuality in the general population. If, on the other hand, identical twins are as likely to have the same sexual orientation as adopted siblings, this suggests that genetic factors make very little contribution to sexual orientation.

In both twin studies, subjects were recruited through ads placed in gay publications that asked for homosexual or bisexual volunteers with twin or adoptive siblings. Volunteers were encouraged to reply to the ad 'regardless of the sexual orientation' of their siblings. In both of these studies, the percentage of identical twins who are both homosexual is substantially higher than the percentage with respect to fraternal twins. For example, 48 percent of the identical twins of lesbians were also lesbians, 16 percent of the fraternal twin sisters were lesbians, 14 percent of the nontwin biological sisters were lesbians, as were 6 percent of adoptive sisters. These results show that sexual orientation is at least partly not the result of genetic factors. However, the higher concordance rate is consistent with a genetic effect because identical twins share all of their genes while fraternal twins, on average, share only half of their genes. Also consistent with a genetic effect is the result that the concordance rates for both types of twins are higher than the concordance rates for adopted siblings.

5. *Is Sexual Orientation Genetic?*

Building on heritability studies, Dean Hamer and his collaborators obtained DNA samples from gay brothers in families in which homosexuality seemed surprisingly common. These samples were analyzed using *linkage analysis*, a technique for narrowing the location of a gene for some trait, to see if there was any particular portion of the X chromosome that was identical in the pairs of brothers at an unexpectedly high frequency (Hamer et al. 1993). Hamer found that a higher than expected percentage of the pairs of gay brothers had the same genetic sequences in a particular portion of the q28 region of the X chromosome (82 percent rather than the expected 50 percent). In other words, he found that gay brothers are much more likely to share the same genetic sequence in this particular region than they are to share the same genetic sequence in any other region of the X chromosome. Hamer's study suggests that the q28 region is the particular place where sexual orientation differences are inscribed. This study does *not*, contrary to popular belief, claim to identify any *particular* genetic sequence associated with homosexuality. At best, it has found that many of the pairs of homosexual brothers had the same genetic sequences in this portion of the X chromosome. When Hamer is at his most precise, he says that his study shows that 'at least one subtype of male sexual orientation is genetically influenced' (Hamer et al. 1993, p. 321). In other contexts, Hamer is less careful. For example, in his book, *Science of Desire*, he talks of 'gay genes,' going so far as to use this term in the book's subtitle (Hamer and Copeland 1994).

6. *Problems with These (and Other) Studies*

6.1 *Lack of Confirmation*

Independent confirmation is the earmark of the scientific method. LeVay's results have not been independently confirmed (Byne 1996) and Hamer's study has been recently disconfirmed (Rice et al. 1999). Although various research teams have confirmed the twin study results, a recent and more sophisticated study by Bailey undermines the methodology of early twin studies (Bailey et al. in press). Bailey systematically recruited subjects from a registry of identical and fraternal twins in Australia. In women, the percentage of identical twins of bisexuals and homosexuals who are also either bisexual or homosexual was between 24 and 30 percent (depending on how the boundaries of these groups are drawn), while the percentage of same-sex fraternal twins of bisexual and homosexual women was between 10 and 30 percent. Not only is the difference between identical and fraternal twins significantly smaller than in previous studies but the percentages for identical twins with the same sexual orientations are dramatically lower. Although these results can be read as consistent with the direct model, the evidence is much weaker than earlier heritability studies suggested.

6.2 *Problems with the Subject Pool*

This Australian study shows that the results of earlier twin and family studies (Bailey and Pillard 1991, Bailey et al. 1993, Hamer et al. 1993), which recruited subjects through HIV clinics, lesbian and/or gay organizations, newspapers, and other nonsystematic methods, must have been inflated by sampling bias. In particular, it suggests that gay men and lesbians with identical twins of the same sexual orientation are more likely to participate in such studies. If an experiment makes use of a subject pool that is in some way biased, then this gives rise to doubts about the conclusions based on it. LeVay's subject pool is also biased. In particular, the homosexual population in his study is made up exclusively of men who died from complications due to AIDS and who told hospital staff that they had engaged in same-sex sexual activities.

6.3 Methods of Classification

A related problem with such studies concerns the determination of subjects' sexual orientations. Conclusions based on studies that inaccurately assign sexual orientations to subjects are weak. LeVay assumed, on the basis of no particular evidence, that all the women in his study were heterosexual and that all the men in his studies whose hospital records did not indicate same-sex sexual activity were heterosexual. Other studies that rely on a family member to report a person's sexual orientation may be similarly problematic. Further, by assuming that there are two sexual orientations—gay and straight—and that people can be easily and reliably classified as one or the other on the basis of their behavior or their self-report, scientific research accepts, without strong justification, that our cultural presumptions about human sexual desires are scientifically valid (Stein 1999, pp. 201–13).

6.4 Undefended Assumptions

Generally, most biological research on sexual orientation accepts without argument a quite particular picture of sexual orientation. For example, many studies in the emerging research program unquestioningly accept the inversion assumption, according to which lesbians and gay men are seen as gender inverts and many studies assume the direct relevance of animal models of sexual behavior to human sexual orientation (Stein 1999, pp. 164–79). More crucially, such studies typically accept that a person's sexual orientation is a deep scientific property about her and that sexual orientation is a 'window into a person's soul.' This view of the centrality of sexual orientation to human nature is neither culturally universal nor scientifically established (Stein 1990, Stein 1999, pp. 93–116).

7. Conclusion

Although there is some evidence that is consistent with biological factors playing an indirect role in the development of sexual orientations, there is no convincing evidence that biological factors are a direct cause of sexual orientation. How human sexual desires develop is an interesting research question, but we are probably rather far from answering it.

See also: Culture as Explanation: Cultural Concerns; Feminist Theory: Radical Lesbian; Sexual Preference: Genetic Aspects

Bibliography

Bailey J M, Pillard R C 1991 A genetic study of male sexual orientation. *Archives of General Psychiatry* **48**: 1089–96

Bailey J M, Pillard R C, Neale M C, Agyei Y 1993 Heritable factors influence sexual orientation in women. *Archives of General Psychiatry* **50**: 217–23

Bailey J M, Dunne M P, Martin N G in press The distribution, correlates, and determinants of sexual orientation in an Australian twin sample

Bem D J 1996 Exotic becomes erotic: A developmental theory of sexual orientation. *Psychological Review* **103**: 320–35

Byne W 1994 The biological evidence challenged. *Scientific American* **270**(May): 50–5

Byne W 1996 Biology and sexual orientation: Implications of endocrinological and neuroanatomical research. In: Cabaj R, Stein T (eds.) *Textbook of Homosexuality and Mental Health*. American Psychiatric, Press, Washington, DC

Gelman D 1992 Born or bred? *Newsweek*, February **24**: 46–53

Hamer D, Copeland P 1994 *The Science of Desire: The Search for the Gay Gene and the Biology of Behavior*. Simon and Schuster, New York

Hamer D H, Hu S, Magnuson V L, Hu N, Pattatucci A 1993 A linkage between DNA markers on the X chromosome and male sexual orientation. *Science* **261**: 321–7

LeVay S 1991 A difference in hypothalamic structure between heterosexual and homosexual men. *Science* **253**: 1034–7

LeVay S, Hamer D H 1994 Evidence for a biological influence in male homosexuality. *Scientific American* **270**(May): 44–9

Peplau L A, Garnets L D, Spalding L R, Conley T D, Veniegas R C 1998 A critique of Bem's 'exotic becomes erotic' theory of sexual orientation. *Psychological Review* **105**: 387–94

Pillard R C, Weinrich J 1986 Evidence of familial nature of male homosexuality. *Archives of General Psychiatry* **43**: 808–12

Rice G, Anderson C, Risch N, Ebers G 1999 Male homosexuality: Absence of linkage to microsatellite markers at xq28. *Science* **284**: 665–7

Stein E D (ed.) 1990 *Forms of Desire: Sexual Orientation and the Social Constructionist Controversy*. Garland, New York

Stein E D 1999 *The Mismeasure of Desire: The Science, Theory, and Ethics of Sexual Orientation*. Oxford University Press, New York

E. Stein

Sexual Orientation: Historical and Social Construction

Sexual orientation is a much used but fundamentally ambiguous concept. It came into use in discussions of sexuality in the 1970s largely as a synonym for homosexual desire and object choice, less frequently for heterosexual patterns. 'Sexual orientation' suggests an essential sexual nature. The task of historical or social constructionist approaches is to suggest that this belief is what itself needs investigation. Constructionist approaches seek to do two broad things: to understand the emergence of sexual categorizations (such as 'the homosexual' or 'the heterosexual' in western cultures since the nineteenth century) within their specific historical and cultural contexts; and to interpret the sexual meanings, both subjective and social, which allow people to identify with, or reject,

these categorizations. It is, thus, largely preoccupied, not with what causes individual desires or orientations, but with how specific definitions develop within their historic contexts, and the effects these definitions have on individual self-identifications and collective meanings.

1. *The Rise of Constructionist Approaches to Sexuality*

The classical starting point for social constructionist approaches is widely seen as an essay on *The Homosexual Role* by the British sociologist, Mary McIntosh (1968). Its influence can be traced in a range of historical studies from the mid-1970s (Weeks 1977, Greenberg 1988), and it has been anthologized frequently (e.g., Stein 1992). What is important about the work is that it asks what was at the time a new question: not, as had been traditional in the sexological tradition from the late nineteenth century, what are the causes of homosexuality, but rather, why are we so concerned with seeing homosexuality as a condition that has causes? And in tackling that new question, McIntosh proposed an approach that opened up a new research agenda through seeing homosexuals 'as a social category, rather than a medical or psychiatric one.' Only in this way, she suggests, can the right questions be asked, and new answers proposed. Using Kinsey (Kinsey et al. 1948), McIntosh makes a critical distinction between homosexual behavior and 'the homosexual role.' Homosexual behavior is widespread; but distinctive roles have developed only in some cultures, and do not necessarily encompass all forms of homosexual activity. The creation of a specialized, despised, and punished role or category of homosexual, such as that which developed in Britain from the early eighteenth century, was designed to keep the bulk of society pure in rather the same way that the similar treatment of some kinds of criminal keeps the rest of society law-abiding. McIntosh drew on a variety of intellectual sources, from structural functionalism to dramaturgical approaches, but clearly central to her argument was a form of labeling theory. The creation of the homosexual role was a form of social control designed to minoritize the experience, and protect and sustain social order and traditional sexual patterns.

If McIntosh put on the agenda the process of social categorization, another related but distinctive approach that shaped social constructionism came from the work of Gagnon and Simon, summarized in their book *Sexual Conduct: The Social Sources of Human Sexuality* (1974). Drawing again on the work of Kinsey and symbolic interactionist traditions, they argued that, contrary to the teachings of sexology, sexuality, far from being the essence of 'the natural,' was subject to sociocultural shaping to an extraordinary degree. The importance culture attributes to sexuality may, therefore, they speculated, not be a result of its intrinsic significance: society may have had a need to invent its importance and power at some point in history. Sexual activities of all kinds, they suggested, were not the results of an inherent drive but of complex psychosocial processes of development, and it is only because they are embedded in social scripts that the physical acts themselves become important. These insights suggested the possibility of exploring the complex processes by which individuals acquired subjective meanings in interaction with significant others, and the effects of 'sexual stigma' on these developmental processes (Plummer 1975).

By the mid-1970s, it is possible to detect the clear emergence of a distinctive sociological account, with two related concerns. One focused on the social categorization of sexuality, asking questions about what historical factors shaped sexual differences which appeared as natural but were in fact cultural. The other was concerned primarily with methods of understanding the shaping of subjective meanings through sexual scripting, which allowed a better understanding of the balance between individual and collective sexual meanings.

A third theoretical element now came into play: that represented by the work of Foucault (1976/1979). Foucault's essay is often seen, misleadingly, as the starting point of constructionist approaches, but there can be no doubt of the subsequent impact of what was planned as a brief prolegomena to a multivolumed study. Like Gagnon and Simon, Foucault appeared to be arguing that 'sexuality' was a 'historical invention.' Like McIntosh, and others who had been influenced by her, he saw the emergence of the concept of a distinctive homosexual personage as a historical process, with the late nineteenth century as the key moment. The process of medicalization, in particular, was seen as a vital explanatory factor. Like McIntosh, he suggested that psychologists and psychiatrists have not been objective scientists of desire, as the sexological tradition proclaimed, but on the contrary 'diagnostic agents in the process of social labeling' (McIntosh 1968). But at the same time, his suggestion that people do not react passively to social categorization—'where there is power, there is resistance'—left open the question of how individuals react to social definitions, how, in a word that now became central to the debate, *identities* are formed in the intersection of social and subjective meanings.

2. *Sexual Behavior, Sexual Categories, and Sexual Identities*

The most crucial distinction for social constructionism is between sexual behavior, categories, and identities. Kinsey et al (1948) had shown that there was no

necessary connection at all between what people did sexually and how they identified themselves. If, in a much disputed figure, 37 per cent of the male population had had some sort of sexual contact with other men to the point of orgasm, yet a much smaller percentage claimed to be exclusively homosexual, identity had to be explained by something other than sexual proclivity or practice. Yet at the same time, by the 1970s, many self-proclaimed homosexuals were 'coming out,' in the wake of the new lesbian and gay movement. Many saw in the historicization of the homosexual category a way of explaining the stigma that homosexuality carried. What was made in history could be changed in history. Others, however, believed clearly that homosexuality was intrinsic to their sense of self and social identity, essential to their nature. This was at the heart of the so-called social constructionist–essentialist controversy in the 1970s and 1980s (Stein 1992). For many, a critique of essentialism could also be conceived of as an attack on the very idea of a homosexual identity, a fundamental challenge to the hard won gains of the lesbian and gay movement, and the claim to recognition of homosexuals as a legitimate minority group. This was the source of the appeal of subsequent theories of a 'gay gene' or 'gay brain,' which suggested that sexual orientation was wired into the human individual.

It is important to make several clear points in response to these debates, where social scientific debates became a marker of social movement differences. First, the distinction between behavior, categories, and identities need not necessarily require the ignoring of questions of causation, it merely suspends them as irrelevant to the question of the social organization of sexuality. Foucault himself stated that: 'On this question I have absolutely nothing to say' (cited in Halperin 1995). The really important issue is not whether there is a biological or psychological propensity that distinguishes those who are sexually attracted to people of the same gender from those who are not. More fundamental are the meanings these propensities acquire, however, or why ever they occur, the social categorizations that attempt to demarcate the boundaries of meanings, and their effect on collective attitudes and individual sense of self. Social categorizations have effects in the real world, whether or not they are direct reflections of inherent qualities and drives.

The second point to be made is that the value of the argument about the relevance of theories of a 'homosexual role' does not depend ultimately on the validity of the variants of role theory (cf. Whitam and Mathy 1986; Stein 1992). The use of the word 'role' was seen by McIntosh (1968) as a form of shorthand, referring not only to a cultural conception or a set of ideas but also to a complex of institutional arrangements which depended on and reinforced these ideas. Its real importance as a concept is that it defined an issue that required exploration. Terms such as constructionism and roles are in the end no more than heuristic devices to identify and understand a problem in studying sexuality in general and homosexuality in particular. It is transparently obvious that the forms of behavior, identity, institutional arrangements, regulation, beliefs, ideologies, even the various definitions of the 'sexual,' vary enormously through time and across cultures and subcultures. A major objective of historical and social constructionist studies of the erotic has been to problematize the taken for granted, to denaturalize sexuality in order to understand its human dimensions and the coils of power in which it is entwined, how it is shaped in and by historical forces and events. The historicization of the idea of the homosexual condition is an excellent pioneering example of this.

The third point that requires underlining is that, regardless of evidence for the contingency of sexual identities, this should not imply that personal sexual identities, once acquired, can readily be sloughed off. The fact that categories and social identities are shaped in history does not in any way undermine the fact that they are fully lived as real. The complex relationship between societal categorization and the formation of subjectivities and sexual identities has in fact been the key focus of writing about homosexuality since the mid-1970s. On the one hand, there is a need to understand the classifying and categorizing processes which have shaped our concepts of homosexuality—the law, medicine, religion, patterns of stigmatization, formal and informal patterns of social regulation. On the other, it is necessary to understand the level of individual and collective reception of, and battle with, these classifications and categorizations. The best historical work has attempted to hold these two levels together, avoiding both sociological determinism (you are what society dictates) or extreme voluntarism (you can be anything you want): neither is true (see discussion in Vance 1989).

Some of the most interesting work has attempted to explore the subcultures, networks, urban spaces, or even rural idylls that provided the space, the conditions of possibility, for the emergence of distinctive homosexual identities. McIntosh's suggestion that the late seventeenth century saw the emergence of a subcultural context for a distinctive homosexual role in England has been enormously influential. Her rediscovery of the London mollies' clubs has been the starting point of numerous historical excavations (e.g., Trumbach 1977; Bray 1982). There is now plentiful work which attempts to show that subcultures and identities existed before the late seventeenth century, for example, in the early Christian world (Boswell 1980), or in other parts of Europe (see essays in Herdt 1994), just as there have been scholars who have argued that we cannot really talk about homosexual identities until the late nineteenth, or even mid-twentieth centuries (see essays in Plummer 1981). There is a real historical debate. As a result, it would

now seem remarkable to discuss sexual identities (and their complex relationship to social categorizations) without a sense of their historical and social context. Sexual identities are made in history, not in nature.

3. Heterosexuality and Homosexuality

Identities are not, however, created in isolation from other social phenomena. In particular, a history of homosexual identities cannot possibly be a history of a single homogeneous entity, because the very notion of homosexuality is dependent, at the very least, on the existence of a concept of heterosexuality, which in turn presupposes a binary notion of gender. Only if there is a sharply demarcated difference between men and women does it become meaningful to distinguish same sex from other sex relationships. Social constructionism, Vance (1989) noted, had paid little attention to the construction of heterosexuality. But without a wider sense that 'the heterosexual' was also a social construction, attempts to explain the invention of 'the homosexual' made little sense. One of the early attractions of the first volume of Foucault's *The History of Sexuality* was precisely that it both offered an account of the birth of the modern homosexual, and put that into a broader historical framework: by postulating the invention of sexuality as a category in western thought, and in delineating the shifting relationships between men and women, adults and children, the normal and the perverse, as constituent elements in this process. Foucault himself was criticized for putting insufficient emphasis on the gendered nature of this process, but this was more than compensated for by the developing feminist critique of 'the heterosexual institution,' with its own complex history (MacKinnon 1987; Richardson 1996).

Central to these debates was the perception that sexuality in general is not a domain of easy pluralism, where homosexuality and heterosexuality sit easily side by side. It is structured in dominance, with heterosexuality privileged, and that privilege is essentially male oriented. Homosexuality is constructed as a subordinate formation within the 'heterosexual continuum,' with male and female homosexuality having a different relationship to the dominant forms. In turn, once this is recognized, it becomes both possible and necessary to explore the socially constructed patterns of femininity and masculinity (Connell 1995).

Although the constructionist debates began within the disciplines of sociology and history, later developments, taking forward both theoretical and political (especially feminist) interventions, owed a great deal to postructuralist and deconstructionist literary studies, and to the emergence of 'queer studies.' Whereas history and sociology characteristically have attempted to produce order and pattern out of the chaos of events, the main feature of these approaches is to show the binary conflicts reflected in literary texts. The texts are read as sites of gender and sexual contestation and, therefore, of power and resistance (Sedgwick 1990). Sedgwick's work, and that of the American philosopher, Butler (1990) were in part attempts to move away from the essentialist/constructionist binaries by emphasizing the 'performative' nature of sex and gender. This in turn opened up what might be called the 'queer moment' that radically challenged the relevance of fixed sexual categorizations. For queer theorists, the perverse is the worm at the center of the normal, giving rise to sexual and cultural dissidence and a transgressive ethic, which constantly works to unsettle binarism and to suggest alternatives.

4. Comparative Perspectives

Much of the debate about the homosexual/heterosexual binary divide was based on the perceived Western experience, and was located in some sense of a historical development. Yet from the beginning, comparisons with non-Western sexual patterns were central to constructionist perspectives. Foucault (1976/1979) compared the Western 'science of sex' with the non-Western 'erotic arts.' It was the very fact of different patterns of 'institutionalised homosexuality,' that formed the starting point of McIntosh's essay, where she had identified two key regulating elements: between adults and nonadults (as in the intergenerational sex which denoted the passage from childhood to adulthood in some tribal and premodern societies), and between the genders (as in the case of the native American 'berdache'). For some writers, these patterns are the keys to understanding homosexuality in premodern times. Historians have traced the evolution over time of patterns of homosexual life which have shifted from intergenerational ordering, through categorization around gender and class, to recognizably modern forms of egalitarian relationships (for a historical perspective see Trumbach 1998). So it is not surprising that constructionist approaches have led to an efflorescence of studies of sexuality in general, and homosexuality in particular, in other cultures, tribal, Islamic, southern (Herdt 1994). This comparative framework increasingly has been deployed within contemporary western societies to highlight the difficulty of subsuming behavior within a confining definition of condition or fixed orientation.

5. Beyond Constructionism

Historical and social constructionism has advanced and changed rapidly since the 1970s. The 'category' that early scholars were anxious to deconstruct has become 'categories' which proliferate in contemporary societies. 'Roles,' neat slots into which people could be expected to fit as a response to the bidding of the agents of social control, have become 'performances'

(Butler 1990) or 'necessary fictions' (Weeks 1995), whose contingencies demand exploration. 'Identities,' which once seemed categoric, are now seen as fluid, relational, hybrid: people are not quite today what they were yesterday, or will be tomorrow. Identities have come to be seen as built around personal 'narratives,' stories people tell each other in the various interpretive communities to which they belong (Plummer 1995). Individual identities, it is increasingly recognized, are negotiated in the ever-changing relationship between self and other, within rapidly changing social structures and meanings. Sexual orientation may, or may not, be a product of genetics, psychosocial structuring, or environmental pressures. That issue, which has tortured sexology for over a century, may or may not be resolved at some stage in the future. For the constructionist, however, other questions are central: not what causes the variety of sexual desires, 'preferences' or orientations that have existed in various societies at different times, but how societies shape meanings around sexual diversity, and the effects these have on individual lives.

See also: Feminist Theory: Radical Lesbian; Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender Differences in Personality and Social Behavior; Gender Ideology: Cross-cultural Aspects; Heterosexism and Homophobia; Masculinities and Femininities; Queer Theory; Sex-role Development and Education; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Preference: Genetic Aspects; Sexuality and Gender; Sexuality and Geography

Bibliography

Altman D, Vance C, Vicinus M, Weeks J et al. 1989 *Homosexuality, Which Homosexuality*? GMP Publishers, London

Boswell J 1980 *Christianity, Social Tolerance, and Homosexuality: Gay People in Western Europe from the Beginning of the Christian Era to the Fourteenth Century*. University of Chicago Press, Chicago

Bray A 1982 *Homosexuality in Renaissance England*. GMP Publishers, London

Butler J 1990 *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York

Connell R W 1995 *Masculinities*. University of California Press, Berkeley, CA

Foucault M 1976/1979 *Histoire de la sexualite, 1, La Volunte de Savoir*. Gallimard, Paris [trans. 1979 as *The History of Sexuality, Vol. 1, An Introduction*. Allen Lane, London]

Gagnon J H, Simon W 1973 *Sexual Conduct: The Social Sources of Human Sexuality*. Aldine, Chicago

Greenberg D E 1988 *The Construction of Homosexuality*. University of Chicago Press, Chicago

Halperin D 1995 *Saint Foucault: Towards a Gay Hagiography*. Oxford University Press, New York

Herdt G (ed.) 1994 *Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History*. Zone Books, New York

Kinsey A C, Pomeroy W B, Martin C E 1948 *Sexual Behavior in the Human Male*. Saunders, Philadelphia

McIntosh M 1968 The homosexual role. *Social Problems* **16**(2): 182–92

MacKinnon C A 1987 *Feminism Unmodified: Discourses on Life and Law*. Harvard University Press, Cambridge, MA

Plummer K 1975 *Sexual Stigma: An Interactionist Account*. Routledge and Kegan Paul, London

Plummer K 1995 *Telling Sexual Stories: Power, Change and Social Worlds*. Routledge, London

Richardson D (ed.) 1996 *Theorising Heterosexuality: Telling it Straight*. Open University Press, Buckingham, UK

Sedgwick E K 1990 *Epistemology of the Closet*. University of California Press, Berkeley, CA

Stein E (ed.) 1992 *Forms of Desire: Sexual Orientation and the Social Constructionist Controversy*. Routledge, New York

Trumbach R 1977 London's sodomites: Homosexual behavior and Western culture in the 18th century. *Journal of Social History* **11**(1): 1–33

Trumbach R 1998 *Sex and the Gender Revolution, Heterosexuality and the Third Gender in Enlightenment London*. Chicago University Press, Chicago, Vol. 1

Vance C S 1989 Social Construction Theory: Problems in the History of Sexuality. In: Altman D, Vance C, Vicinus M, Weeks J et al. (eds.) *Homosexuality, Which Homosexuality?* GMP Publishers, London, pp. 13–34

Weeks J 1977 *Coming Out: Homosexual Politics in Britain from the Nineteenth Century to the Present*. Quartet Books, London

Weeks J 1995 *Invented Moralities: Sexual Values in an Age of Uncertainty*. Columbia University Press, New York

Whitam F L, Mathy R M 1986 *Male Homosexuality in Four Societies*. Praeger, New York

J. Weeks

Sexual Perversions (Paraphilias)

1. Definition

Modern usage tends to favor the word 'paraphilia' rather than 'deviance' or 'perversion.' In this article these latter terms will be used only in their historical context. The word 'paraphilia' means a love of (*philia*) the beyond or irregular (*para*), and is used instead of those words which today have pejorative implications that are not always relevant. The term itself is used to describe people, usually men, with intense sexual urges that are directed towards nonhuman objects, or the suffering or humiliation of oneself or one's partner, or more unacceptably, towards others who are incapable of giving informed consent, such as children, animals, or unwilling adults. People who are paraphiliacs often exhibit three or four different aspects, and clinical psychiatric conditions (personality disorders or depression) may sometimes be present.

Paraphilias include: exhibitionism (the exposure of ones genitals to a stranger, sometimes culminating in masturbation); voyeurism or peeping (the observance of strangers undressing or having sexual intercourse,

without their being aware of the voyeur, who usually masturbates); fetishism (the use of inanimate objects for arousal, usually articles of women's underwear, although if these are used to cross-dress, transvestic fetishism is the diagnosis); frotteurism (the rubbing of the genitals against the buttocks, or the fondling of an unsuspecting woman, usually in a crowded situation, so that detection of the perpetrator is unlikely). Pedophilia refers to men sexually attracted to children, some to girls, some to boys and some to either sex. Some pedophiles are attracted sexually only to children, the exclusive types, but some are attracted to adults as well, the nonexclusive type.

Sexual sadism and sexual masochism (S&M) involve the infliction or reception of pain respectively to necessitate arousal and orgasm. Sadistic fantasies that involve obtaining complete control of the victim are particularly dangerous and may result in death. Some masochistic behaviors that involve self-asphyxiation as part of a masturbatory ritual can result in accidental death. Other paraphilias include the use for sexual purposes of corpses (necrophilia), animals (zoophilia), feces (coprophilia), enemas (klismaphilia) and urine (urophilia).

It is of particular interest that just as masturbation is nowadays excluded from being considered a deviant act, so homosexuality, which figured so largely in this area in earlier times, was removed as a paraphilia in the 1974 edition of the *Diagnostic and Statistical Manual of Mental Disorders*.

2. Greek Mythology

The ancient Greeks often ascribed to their gods extreme forms of sexual behavior. Whether this reflected fears or wishful thinking or actual practices is obviously very difficult to evaluate. Zeus's myriad affairs were often conducted in the form of animals or even inanimate objects, such as a cloud or a shower of gold, and his objects of desire were male or female, as indeed were those of the other gods and heroes. Artemis and Athena both eschewed sex altogether and remained adamantly virginal. The mighty Hercules cross-dressed to be humiliated by Queen Omphale, while Theseus and Achilles both donned women's clothes without apparent later loss of esteem (Licht 1969, Bullough 1976).

3. History

3.1 The Hunter-gatherer Past

In most species sexuality and reproduction are of necessity tied together and overlap, but in modern humans it is the relatively recent separation of sexuality from reproduction that represents a crucial evolutionary phase. Despite huge sociological changes we are genetically still at the stage of hunter-gatherers, who had a shorter life, no menopause and long periods between childbirth due to prolonged breast feeding. Perhaps because of better nutrition, and possibly exposure to electric light (which affects the hypothalamus), girls in the developed countries in the 1990s have their first period at around 13 years of age, and ovulation starts soon afterwards, with pregnancy being a distinct possibility. Usually most women have conceived and had their children by 30–40 years of age, which, assuming many will reach 80 years of age, leaves some 40–50 years of life to enjoy non-reproductive sex. Sex therefore takes on new meanings and fulfils other roles, perhaps as a recreational activity, which allows time for sexual variant behavior (Short 1976).

3.2 Sexuality

Numerous so-called Venus figurines and cave paintings have been found from between 25,000 and 10,000 years ago that depict almost all of the sexual acts with which we are familiar today. What the meanings and significance of these depictions were to the people who created them we can only speculate. Early studies showed that paraphilias were present universally in every culture and throughout every historical period. Often tied up with religious rites, sacred prostitution, and/or phallic cults, public shows of self-immolation, sadism, fetishism, and homosexuality flourished. Iwan Bloch believed that every sensory organ could function as an erotogenous zone, and so form the basis for a so-called perversion. Freud pointed out that the ancients laid stress on the instinct itself, whereas we today are preoccupied with its object. Thus it was considered wrong to be passive, that is, to be penetrated anally by one's slave, but acceptable to penetrate him or her. Today the object, that is, whom a man penetrates, is important.

3.3 Classifications

The enormous variations found in human sexual behavior have always been part of the heritage of humankind, but it was not until the eighteenth century in the West that such activities were labeled and put into categories, so that attempts at a scientific classification could be made. These categories varied from culture to culture and showed a great deal of fluidity, as each culture constructed the behavior in different ways. The essential behaviors remained constant, however, and it was rather how each society considered them, whether they approved or disapproved, that created the particular sexual climate. Classifications thus offer an insight into the current thinking of each particular time.

The Marquis de Sade (1740–1814), while imprisoned in the Bastille, wrote a full description of most perversions and suggested a classification about

100 years before his successors, Krafft-Ebing, Havelock Ellis, and Freud. He wrote many short stories on sexual themes including incest and homosexuality. He also foresaw the women's rights issue, which he treated in a compassionate and modern way, yet this did not preclude his use of women as victims in his writings. His *120 Days of Sodom* is perhaps his main work on perversion, although all his works contain little homilies about sex and morals. As de Sade lived in an age that was both turbulent and brutal, it says something for the humanity of the man that he implored us not to laugh at or sneer at those with deviant impulses, but rather to look upon them as one might a cripple, with pity and understanding. De Sade offers the following classification for sexuality (Gorer 1962):

(a) People with weak or repressed sexual desires—the nondescript majority, he saw as cannon fodder for the next two groups;

(b) Natural perverts (born so);

(c) Libertines—people who imitate group (b), but wilfully rather than innately.

De Sade's admiration was for group (b). The *120 Days of Sodom*, written while Sade was in the Bastille, contains examples of all these, and the range extends from foot fetishism, various types of voyeurism, obsessional rituals and bestiality, to frank sadism and murder. Sade described many sexual acts involved with body fluids, sperm, blood, urine, saliva, and feces.

Jeremy Bentham (1748–1832), the utilitarian philosopher who believed that nature has placed humankind under the governance of two sovereign masters, *pain* and *pleasure*, published his essay on pederasty in 1785 in which he argued cogently for the decriminalization of sodomy (then punishable by hanging), and other sexual acts. He challenged the notion current at the time that such behaviors weaken men, threaten marriage, and diminish the population. He went on to discuss these punishments as an irrational antipathy to pleasure, and he highlighted the dangers that prohibitions incur, i.e., possible blackmail and false accusations (Crompton 1978).

In the nineteenth century, Darwin in his theory of evolution was able to explain the presence of gill slits and the tail present in the early human embryo as well as the appendix, as evidence of our evolutionary past. Ernst Haeckel, Darwin's forceful adherent, developed the concept that ontogeny recapitulated phylogeny (see *Evolution, Natural and Social: Philosophical Aspects*). It was argued from this that if elements from our past physical evolution were so preserved, why should this not apply to our psychological history? Thus sexual perversions were deemed to reflect a return to an earlier developmental stage in phylogeny that had somehow become fixated, so that perverts, like other races and women, were at a lower point on the evolutionary ladder. Following on from Darwin's observations of hermaphroditism in nature, Karl A. Ulrichs (whose pseudonym was Numa Numantius) believed male homosexuals to be female souls in men's bodies.

Succeeding generations of doctors and sexologists argued these points and many offered alternative classifications of the perversions to reflect their views. These are mainly of historical value now, but perhaps we could look at Krafft-Ebing's classification as an example. He believed that life was a struggle between instinct and civilization, and that mental illness created no new instincts, but only altered those that already existed. He believed that hereditary taint and excessive masturbation were among the causes of sexual perversion. His classification, which is self-explanatory, was as follows:

Group 1: deviation of the sex impulse. Too little sexual feeling (impotence, frigidity); too much (satyriasis and nymphomania), and sexual feeling appearing at the wrong time (in childhood or old age) or being directed wrongly (at children, the elderly or animals).

Group 2: Sexual release through other forms of activity (inflicting or receiving pain or sexual bondage).

Group 3: Inverted sexuality (homosexuality, bisexuality, and transvestism).

Krafft Ebing's *Psychopathia Sexualis* ran into 12 editions in his lifetime, with many revisions, and although he did write some of the more explicit details in Latin, the *British Medical Journal* in 1893 lamented that the entire book had not been written in Latin, 'in the decent obscurity of a dead language.' Like Sade before him, Krafft Ebing spoke of 'perverts' as 'stepchildren of nature,' and asked for a more medical and sympathetic approach rather that mere legal strictures.

His belief that most perversions were mainly congenital was challenged by Alfred Binet who argued that with fetishism (a term he coined), for example, the fetish may take many different forms, for example, an obsession with the color of the eyes, bodily odors, various types of dress, or different inanimate objects. Heredity could not dictate this choice, rather a chance phenomena that occurred in childhood was more likely as a cause. Furthermore, on this argument, the fetishist could just as well become a sadist, masochist, or homosexual if early childhood events had been so conducive. As many individuals were exposed to such childhood events and did not develop a perversion, however, Binet had to conclude that there could well be a congenital morbid state in those who did so. Albert Moll argued further that as all biological organs and functions were subject to variations and anomalies, why should the sexual behaviors be any different?

Moll further believed that the sense of smell was always an important factor in mammalian sexuality, and the advent of clothing in human culture diminished this. Wilhelm Fliess, an ENT surgeon and colleague of Freud, drew attention to the erectile

tissues of the nose and its similarity to that of the penis and clitoris. The upright posture adopted by humans had distanced them from the ground, away from feces and smell, so diminishing the importance of olfaction in human arousal, and forming what Freud described as an abandoned erotogenous zone (Sulloway 1992).

Today these mechanisms have been somewhat elaborated. A region in the nasal septum (Jacobson's organ), is linked by nerves to the hypothalamus, a region of the brain that controls sex-hormone secretion. Small volatile chemicals known as pheromones, found in sweat around the genital and axillary regions, are known to stimulate this nasal pathway. Subtle differences in the pheromone balance depend on the individual's genetic make-up, and it is possible that these differences give a biological basis for avoidance of attraction to one's near relatives who exhibit similar profiles. The importance of smell in animal sexuality is well known, and it is of interest that to many fetishists it is soiled rather than clean female underwear that arouses them (Goodman 1998).

Freud believed that each individual from infancy to adulthood repeated the moral development of the race, from sexual promiscuity, to perversion, and then on to heterosexual monogamy. There was a correct developmental path, during which the infant, from the stage of polymorphous perversity, went through various phases of development, and negotiated the Oedipal situation and castration complex. Freud believed that perversions arose because of arrested development leading to fixations on this pathway due to sexual trauma. Those who did not deviate when so exposed in childhood, Freud believed to have been protected by constitutional factors, namely an innate sense of propriety that had been acquired through moral evolution. He classified as deviants those who have different sexual aims (from normal vaginal intercourse), for example, sadists, fetishishists, and those who have a different love object (from the normal heterosexual partner), such as pedophiles or homosexuals (Weeks 1985) (see *Psychoanalysis: Current Status*).

4. *Modern Times*

At the beginning of the twenty-first century, technology-led cultures have had a profound effect on sexual behavior, which has expanded to fill the different niches in the manner of Darwin's 'tangled bank.' We may see even greater fluidity of such behaviors in the future. Just two areas will be briefly considered.

4.1 *Fetishism and Fashion*

Today the fashion industry makes much use of fetishism in its designs. Magazines cater for a whole range of people, both gay and straight, with fetishistic and S&M interests. Relatively new phenomena have appeared in recent times, such as 'vogueing' where the devotee dresses up as a facsimile body of an admired personality either of the same sex (homovestism), or of the opposite sex (the well-known transvestism). Common examples are Elvis Presley and Madonna lookalikes. Voguers compete with each other for realness. Fans who obtain objects, such as articles of clothing or signed photographs from their idols, or who in the USA search the dustbins of stars for trinkets (so-called 'trashcanners') and later use these in their sexual fantasies, resemble classical fetishists in their behavior (Gammon and Makinen 1994, Goodman 1998).

4.2 *Technology and Sexuality*

Every technological invention has altered society in some way and this applies equally to the field of sexual behavior. Newspapers allow wide advertisement of sexual services, the telephone made possible obscene telephone calls, while commercial sex lines have utilized this phenomenon so that men (it is mainly men who use them) can dial and indulge in sexual talk for the price of the call. The automobile is often used as a place of sexual assignation, the home video for viewing obscene material, and the camcorder for making erotic home movies.

The term 'cybersex' is used for people who wish to indulge in erotic fantasies through the Internet. The Internet provides a means of getting in touch with other like-minded individuals worldwide, so that fantasies, which may previously have existed only in the mind of one individual, can be exchanged, enhanced, and embellished. Arrangements have been made for such people to meet in order to carry out acts, which have included pedophilia, rape, and even murder. The perpetrator(s) may take on a different age or gender persona ('gender bending'), to entrap the unwary, especially children. As a result of this, groups have been set up to combat these problems, for example, the Cyberangels in the UK, who monitor the net and inform the police where necessary.

Psychologically some individuals in tedious relationships may find that occasional looks at erotica enlivens their sex lives, but others have become obsessed and addicted to on-line sex ('hot chatting'), and indeed have needed therapy to help them cope (Durkin and Bryant 1995). Virtual reality, which offers the possibility of taking part in a virtual sexual scenario, will further increase the scope of sexual variant behavior.

5. *Recent Scientific Advances in the Understanding of the Paraphilias*

5.1 *The Brain*

The human brain reached its present size and proportions about 50,000–100,000 years ago. The brain

consists of regions which formed at different epochs in vertebrate evolution. It is the hypothalamus that is largely concerned with the hormonal control of reproduction. Certain nuclei found here in homosexual men have been claimed to be more female than male-like, and similar findings have been made in male to female transsexuals (Swaab et al. 1997). Sexual fantasy depends on the cerebral cortex, which is of such complexity that an almost infinite variety of mental responses are possible which ensures a unique plasticity to the range of sexual behaviors.

Furthermore, following head injury or the use of certain drugs, paraphilic behavior may become evident in individuals who did not show such tendencies before. Medical conditions such as temporal lobe epilepsy have been associated with fetishism in some patients. New noninvasive techniques for studying the brain, such as functional magnetic resonance imaging and positron emission tomography are helping to elucidate its functions.

5.2 Genes and the Developing Fetus

The contributions of genes and development to adult sexual behavior have been a topic of intense debate for many years. Recently, new discoveries have added fuel to this argument, and new concepts have been considered.

In humans the Y-chromosome slows down the growth rate in the male fetus, which is consequently born relatively more immature compared with the female and is therefore more vulnerable. Thus males have a higher perinatal mortality and a higher incidence of accidental death, and later in life they show an increased vulnerability to cardiac disease and certain forms of cancer. Mental handicap, which includes autism and epilepsy, is more common in men than women and perhaps the preponderance of the paraphilias in males is also a reflection of this vulnerability (Ounsted and Taylor 1972).

Do genes play some part in determining human sexual behaviors and orientation? Certainly in *Drosophila*, the fruit fly, where male and female behavior were once considered separate and exclusive, recent experimental manipulations of various genes have produced bisexual and homosexual behaviors in males. Even courtship chains, both heterosexual and homosexual (like something out of the Marquis de Sade), have been seen, behaviors that never occur in the wild (Yamamoto et al. 1996). In humans, some paraphiliac behavior does seem to run in families (Gorman 1964, Cryan et al. 1992), and suggestions of a gene occurring on the X-chromosome that predisposes to male homosexuality have been made (Hamer and Copeland 1994).

Developmental factors are also of seeming importance in future behaviors and it is the presence of the two hormones, testosterone, and estrogen, that are thought to influence the developing fetus. This has been shown in mice and other species. When a male fetus is placed in between two female fetuses leakage of female hormones may feminize the male. Testosterone can masculinize the female if a female is between two males (Vom Saal and Bronson 1980).

Stress applied to pregnant rats just one week before delivery resulted in homosexual and bisexual behaviors in males (Ward and Weisz 1980) and Dorner et al. (1983) thought that the stress of war in pregnant women resulted in a higher incidence of male homosexuality in the German population.

Sex hormones, particularly testosterone, are known to interact with the immune system. It has been suggested that the mother's immune system triggered by a male fetus can affect male psychosexual development, especially if she has had many pregnancies. This would further inhibit her immune system and may explain the preponderance of elder brothers in the families of homosexuals and certain pedophiles (Blanchard and Bogaert 1996) (see *Homosexuality and Psychiatry*).

6. The Future

Freud and the sexologists considered the libido in terms of the combustion engine and electricity, as well as using biological ideas that were extant at the time. Today as knowledge in all fields converges, concepts from one area are fertilized as never before by ideas from others. Cybernetics, the study of control and communication in artificial neural networks, has been applied to biological systems such as the brain. Chaos theory, a concept derived from nonlinear dynamics and used initially to predict the weather, has been applied to numerous other areas of research which include fetal development and sexual behavior, as well as various branches of psychology (Goodman 1997).

Waddington (1975) portrayed fetal development as a mountainous terrain, in which the fetus is the ball that rolls down the valleys, which depict the possible developmental pathways. These pathways represent the culmination of millions of years of evolution and are relatively resistant to change. Both genetic factors and early environmental stresses may divert the fetus onto another pathway however, although the system does have a degree of stability. It may well be that certain paraphiliacs and individuals with homosexual or transsexual identities have been diverted from the more common (but not necessarily more normal) path of heterosexual identity. If sexual orientation and behavior are linked to cognition, as seems to be true, then variations could have evolutionary possibilities by throwing up individuals who have the ability to think differently from their peers, offering no little advantage in the struggle for survival. The epistemological solipsism of the developing brain needs sex for its development in the world and not just for pro-

creation (Freeman 1995). Sexual behavior and its variants therefore may merely be a reflection of this process. We should contemplate it with a sense of awe.

See also: Rape and Sexual Coercion

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*, 4th edn. American Psychiatric Association, Washington DC

Blanchart R, Bogaert A F 1996 Homosexuality in men and number of older brothers. *American Journal of Psychiatry* **153**(1): 27–31

Bullough V L 1976 *Sexual Variance in Society and History*. Wiley, New York

Crompton L 1978 Offences against oneself, by Jeremy Bentham, Part 1 (ca. 1785). *Journal of Homosexuality* **3**(4): 389–405

Cryan E M J, Butcher G J, Webb M G T 1992 Obsessive-compulsive disorder and paraphilia in a monozygotic twin pair. *British Journal of Psychiatry* **161**: 694–8

Dorner G, Schenk B, Schmiedel B, Ahrens L 1983 Stressful events in prenatal life of bi- and homosexual men. *Experimental Clinical Endocrinology* **81**: 83–7

Durkin K F, Byant C D 1995 "Log on to sex": Some notes on the carnal computer and erotic cyberspace as an emerging research frontier. *Deviant Behavior: An Interdisciplinary Journal* **16**: 179–200

Freeman W J 1995 *Societies of Brains. A Study in the Neuroscience of Love and Hate*. Lawrence Erlbaum, Hillsdale, NJ

Gammon L, Makinen M 1994 *Female Fetishism: A New Look*. Lawrence & Wishart, London

Goodman R E 1997 Understanding human sexuality: Specifically homosexuality and the paraphilias in terms of chaos theory and fetal development. *Medical Hypotheses* **48**(3): 237–43

Goodman R E 1998 The paraphilias: An evolutionary and developmental perspective. In: Freeman H, Pullen I, Stein G, Wilkinson G (eds.) *Seminars in Psychosexual Disorders*. Gaskell, London, pp. 142–55

Gorer G 1963 *The Life and Ideas of the Marquis de Sade*. Peter Owen, London

Gorman G F 1964 Fetishism occurring in identical twins. *British Journal of Psychiatry* **110**: 255–6

Hamer D, Copeland P 1994 *The Science of Desire: The Search for the Gay Gene and the Biology of Behavior*. Simon and Schuster, New York

Licht H 1969 *Sexual Life in Ancient Greece*. Panther Books, London

Ounsted C, Taylor D C 1972 The Y-chromosome message, a point of view. In: Ounsted C, Taylor D C (eds.) *Gender Differences: Their Ontogeny and Significance*. Churchill Livingstone, Edinburgh, UK

Short R V 1976 The evolution of human reproduction. *Proceedings of the Royal Society of London, Biology* **195**: 3–24

Swaab D F, Zhou J N, Fodor M, Hofman M A 1997 Sexual differentiation of the human hypothalamus: Differences according to sex, sexual orientation and transsexuality. In: Ellis L, Ebertz L (eds.) *Sexual Orientation. Toward Biological Understanding*. Praeger, Westport, CT, pp. 129–50

Sulloway F J 1992 *Freud. Biologist of the Mind*. Harvard University Press, Cambridge, MA

Vom Saal F S, Bronson F H 1980 Sexual characteristics of adult female mice are correlated with their blood testosterone levels during prenatal development. *Science* **208**: 597–9

Waddington C H 1975 *The Evolution of an Evolutionist*. Edinburgh University Press, Edinburgh, UK

Ward I L, Weisz J 1980 Maternal stress alters plasma testosterone in foetal males. *Science* **207**: 328–9

Weeks J 1985 *Sexuality and its Discontents*. Routledge and Kegan Paul, London

Yamamoto D, Ito H, Fujitani K 1996 Genetic dissection of sexual orientation: Behavioral, cellular, and molecular approaches in *Drosophila melanogaster*. *Neuroscience Research* **26**: 95–107

R. E. Goodman

Sexual Preference: Genetic Aspects

In many animals, females preferentially mate with males that are adorned with extravagant traits like bright feathers or complicated courtship behavior. As a result such sexual preference has led to the evolution of many elaborate male signals as for example nightingale song or peacock feathers. Many male extravagant traits are thus clearly caused by female sexual preferences, but why did female preference evolve in the first place. Several different hypotheses have been proposed to explain the evolution of female choice. Here, these hypotheses and the available empirical evidence that might help to distinguish between them are discussed.

1. Sexual Selection and the Sex Roles

Males and females in most animal species differ to a large extent in their morphology and their behavior. Female birds are, for example, often drab and coy compared to their colorful and sexually active mates. In many species, males compete with other males for access to females and often use specialized horns, teeth or other weapons, whereas their mates often do not fight with other females. The basic reason behind these sex differences is gamete size that often differs by several orders of magnitude between the sexes. Females produce relatively large gametes, the eggs, and accordingly can only produce relatively few of them. Males, on the other hand, produce large numbers of tiny sperm cells. Due to this difference in the number of gametes available, female reproductive success usually is limited by the number of eggs produced, and multiple matings by females have at most only a small effect on female reproductive success. Male reproductive success, however, is mainly limited by the number of eggs fertilized and thus by the number of mates obtained and not so much by sperm

number (Clutton-Brock and Vincent 1991). Males therefore can benefit from attractiveness and from competition for mating partners whereas females probably benefit more by choosing the best ones among the many available mates. In accordance with this expectation, females often seem to carefully examine the available males and reject mating attempts by nonpreferred males (Andersson 1994). In peacocks, for example, hens prefer males with large colorful feathers that are distributed symmetrically, and in some other birds, females prefer males with large song repertoires. The few species where males heavily invest in offspring, e.g., some katydids with large parental investment, as well as seahorses and wading birds with exclusive paternal care, confirm the rule. In these species females usually compete for mates and males choose among females because males are limited by the number of offspring they can care for and females are limited by the number of males they can obtain to care for their offspring. These studies on sex role reversed species show that the sex difference in the benefit of additional matings after the first one is the driving force behind sexual selection. With standard sex roles, male traits that impair survival but render males especially attractive to females can evolve since male traits that are preferred by females will cause an increased mating frequency. Due to the effect of mating frequency on male reproductive success, these traits can lead to an increased lifetime mating success even when male survival decreases. The splendid feathers of male peacocks for example probably decrease male survival because this shiny ornament will also attract predators and make escape less efficient. Less adorned males may live longer. However, when they leave, on average, fewer offspring than attractive males, genes for being less adorned will go extinct. Compared to the evolution of male traits under the influence of female choice, the evolution of female preference is less easy to explain. In the next section the possible routes for the evolution of female preference are described.

2. *Theoretical Models for the Evolution of Preferences*

Several hypothesis have been presented to explain the evolution of female preference that causes females to choose their mating partners among the available males (see Kirkpatrick and Ryan 1991, Maynard Smith 1991, for reviews). The simplest possibility would be that males differ in the amount of resources provided or in their fertilization rate. According to this hypothesis, choosy females increase the number of offspring produced and they thus benefit directly from their choice. In some birds, for example, those males are preferred as mates that provide superior territories and outstanding paternal care. Another possible explanation for the evolution of female preferences is that choice has an influence on the genetic quality of the offspring. According to this hypothesis females benefit indirectly when offspring of preferred males have superior viability or increased mating success compared to the offspring of nonpreferred males. And finally, female preference might have evolved in another context than sexual selection, that is, female birds might prefer males with blue feathers because it is adaptive to react positively to blue when blue berries constitute a valuable resource so that female sensory physiology became tuned to this color. This model has accordingly been termed sensory exploitation hypothesis, meaning that males exploit female sensory physiology to attain attractiveness (Ryan and Keddy-Hector 1992).

For all of these models, genetic variation in preference is essential since evolution of all traits rests on the existence of genetic variance. Despite the importance of such data, not very many species have been examined for genetic variance in female preference. In most of the examined species, ranging from insects to mice, within population variation in female preference seems generally to be influenced by additive genetic variance (Bakker and Pomiankowski 1995). Such genetic variance in preference means that this trait can easily evolve whenever there is a selective advantage to a specific preference. In the following paragraph the different hypotheses for the evolution of female preferences will be discussed in more detail and the genetic aspects of these models will be concentrated on.

When females benefit directly from their choice by increased lifetime fecundity, it is easy to see that any new choice will evolve as long as the extra benefit gained is larger than the extra cost paid for female choice. For example, females might prefer healthy mates because this might reduce the risk of infection during mating. Choosing males with intact and shiny feathers might thus reduce the risk of ectoparasite transfer during copulation. For models that rest on indirect benefits of female preference, females only gain from choice regarding the genetic quality of their offspring. For these indirect benefit models of female preference, the male trait also needs to have genetic variance. One of these hypotheses, Fisher's arbitrary sexual selection model, predicts that female choice should evolve in response to the evolution of male traits. If females prefer males with a specific trait size, a linkage disequilibrium (a nonrandom association of genes) will build up, since the genes for the preference and the genes for the male trait will co-occur together more frequently than by chance. The existence of choosy females will, in addition, cause increased reproductive success of these preferred males. Since these males also carry a disproportional share of the preference allele, due to the linkage disequilibrium that is caused by the preference, the preference evolves in response to the evolution of the male trait (Kirkpatrick and Ryan 1991). If sexual selection is strong and if preference and male trait are heritable, both the

male trait and the female preference are predicted to evolve by a positive feedback that eventually might lead to runaway selection. The distinctive and sufficient genetic condition for Fisher's arbitrary sexual selection model thus is a genetic correlation between female preference and male traits. According to the good genes hypothesis, another hypothesis that suggests indirect benefits of female preferences, choosy females benefit not only by producing preferred sons but also by producing offspring with superior viability. For such a process to work, male attractiveness needs to indicate male viability. Speaking in genetic terminology, there has to be a genetic correlation between male signaling traits and viability. Since a genetic correlation between female preference and male signaling traits will also build up as a result under the good genes model, such a correlation thus cannot be taken as evidence for the arbitrary sexual selection model but it means that one of the indirect benefit models helps to maintain the preference. To provide evidence that the sensory exploitation model helps to explain the evolution of female preference, one has to show that female preference evolved before the male trait and thus independent of mate choice. The usual way is to show that female preference is more ancestral in phylogeny than the male trait. One of the best studied examples of this hypothesis is the tungara frog, *Physalaemus pustulosus*, where females have a preference for a male acoustic signal that is not produced by the males, but males of a closely related species do produce it. The suggested explanation for this pattern is that female choice evolved first in the ancestor of both species and the male trait evolved later in only one of those species. However, the loss of the attractive male signal during phylogeny cannot be excluded as an alternative explanation.

3. *Testing Genetic Aspects of Sexual Selection Models*

The good genes hypothesis predicts that females benefit from choosing attractive males because these males will produce offspring with superior viability. The critical problem for this hypothesis is that attractive males need to have higher viability. Why should such a correlation between attractiveness and viability exist? The most prominent explanation is the following: Attractive signals are costly and only males with superior viability are able to afford these costs so that only those males can attain high attractiveness. This view of condition-dependent signaling is supported by empirical data showing that in various insects, spiders, and frogs the production of attractive courtship signals consumes more energy than the production of less attractive signals. Also, males that suffer from disease or are starving are usually not as attractive as healthy competitors. Such a process of condition-dependent signaling leads to a genetic correlation between male signaling traits and male viability when there is genetic variation for both traits. If females prefer males with signaling traits that are correlated with viability, female choice can evolve since it will become associated with high viability and a rare female choice allele can thus be expected to increase in frequency when the costs of female choice are low. If females choose among males on the basis of a male trait, a genetic correlation will build up between male trait and female preference at the same time. Since this prediction is identical to the critical and sufficient condition for the arbitrary sexual selection model, experimental separation of these two non-exclusive hypotheses is difficult.

The most frequent method to examine whether the good genes hypothesis contributes to maintain female preference is to compare the viability of the offspring of preferred males with the viability of average or nonpreferred males. In some studies using this method, female choice seems to have an astonishingly large effect on offspring viability, in other studies, no significant effect was observed (see Møller and Alatalo 1999 for a review).

Theoretical models show that these indirect sexual selection benefits are unlikely to increase offspring fitness by more than a few percent (Kirkpatrick 1996). Since direct benefits are suggested to incur larger benefits, evolution of female preferences is predicted to be dominated by direct benefits if these can be obtained by female choice. However, even a small indirect benefit to sexual preferences may be large enough for quick evolution when choice is not very costly. With an exhaustive literature review Møller and Alatalo (1999) showed that a significant positive correlation coefficient between male viability and attractiveness exists on average and they therefore proposed that females can gain a few percent regarding the viability of their offspring when they carefully choose among the potential mates.

In general, the evolution and maintenance of sexual preference seems to be due to various factors. Empirical evidence exists in favor of each of the most prominent hypotheses—direct benefits, sensory exploitation, Fisher's arbitrary sexual selection and good gene models. These hypotheses are not mutually exclusive and some of them might work synergistically. In the future, more studies examining the quantitative genetics of sexually selected traits are necessary to evaluate the importance of the different hypotheses (Bakker 1999).

4. *The Maintenance of Genetic Variation in Male Traits*

4.1 *Theoretical Expectation*

Both indirect sexual selection models depend on the existence of genetic variance in male traits and will in

turn also cause strong selection on these traits. For theoretical reasons it has often been argued that genetic variance of traits under strong selection will usually be smaller than the genetic variance of traits that are less strongly selected (Fisher 1930). This difference occurs because under strong selection, all but the most beneficial alleles will disappear quickly from a population. In line with these arguments, life history traits that are closely connected to fitness have a lower additive genetic variance than morphological traits that are believed to be under weaker selection. If strong selection would deplete genetic variance in male attractiveness or viability, the benefit of female preference would decrease and if there are costs to female choice, females would accordingly benefit from refraining from choice and saving these costs. However, females seem to choose strongly among males in most animal species (Andersson 1994). Furthermore, female choice and the effects of sexual selection on male traits are most obvious in lekking species, where males court females at courtship territories and do not provide any resources for the females. In these cases, direct benefits are unlikely and it is not easy to see why sensory exploitation should be more frequent in these species. The indirect benefit models are thus the only ones that are likely to explain why females strongly select among the available males in lekking species and elsewhere when males do not provide resources. However, when females exert sexual selection on male viability or attractiveness, the genetic variance in male quality should decrease and the benefit of the preference will also decrease in response. The important question thus is whether male genetic variance is large enough to counterbalance the costs of being choosy.

4.2 Empirical Evidence for Genetic Variance among Males

There is ample evidence for genetic variance in male signaling traits and in attractiveness: crosses between populations that differ in the male signaling trait show that the difference is inherited; heritability estimates from father-son comparisons show that genetic variation exists even within populations and artificial selection experiments have proven that sexually selected male traits can evolve quickly. Despite strong sexual selection on male signaling traits, sexually selected traits generally seem to have heritabilities that are as large as the values for traits that are assumed to be only under weak selection (Pomiankowski and Møller 1995). The deviation from the theoretical expectation is even more impressive when one compares the additive genetic variance (another measure for genetic variance that does not depend on the extent of environmental influence on the trait under consideration) between sexually selected traits and traits that are assumed to be not under sexual selection. Sexually selected traits have significantly larger additive genetic variance than other traits, showing that the existing genetic variation is sufficient for both indirect sexual selection models.

4.3 Possible Reasons for the Maintenance of Genetic Variance

The extent of genetic variation present in natural populations thus seems to contradict the theoretical expectations. It is, therefore, important to understand how extensive genetic variance can be maintained in the face of strong sexual selection. Among the hypotheses put forward to explain these data that seem to contradict the theory, the following three causes for the maintenance of genetic variance shall be discussed in some detail because they have received considerable credit: capture of genetic variance in condition, host-parasite coevolution, and meiotic drive.

Male signals are only likely to honestly indicate male viability, when attractive signals are costly because otherwise males with low viability will also be able to produce these attractive signals. When signals are costly, only males in good condition will be able to produce attractive signals (Zahavi 1975). Due to this process, male signal quality will indicate a male's condition. Since male condition is assumed to possess large genetic variance because many traits will influence condition, male signals will capture genetic variance in condition and can in this way maintain the genetic variance in male viability necessary for the good genes model (Rowe and Houle 1996).

Another hypothesis suggests that host-parasite coevolution is important in maintaining genetic variance in male signaling traits (Hamilton and Zuk 1982, Westneat and Birkhead 1998). Let us assume there are male traits that indicate that a male is free of parasites or resistant to parasites. If such resistance is heritable, females would clearly benefit from choosing those males and female preference could accordingly evolve. With increasing resistance, selection on the parasite would lead to new types of parasites that in turn will cause the selection of new resistance. This arms race can in principle lead to a persistent and large advantage to female preference for resistant males because genetic variance in males is maintained since superior male genotypes change with time. There is some indication that such a process also works in humans, where scents from potential partners with dissimilar MHC-genes (an immunologically important group of genes) are preferred (Wedekind and Furi 1997).

Preference of females for males resistant to the action of meiotic drive has also recently been proposed as a scenario that allows persistent benefits of female choice. In stalk-eyed flies, females prefer males with longer eyestalks, and lines selected for longer eyestalks show increased resistance to meiotic drive (Wilkinson et al. 1998). It was, therefore, suggested that females

benefit from choosing males with large eyestalks because resistance to meiotic drive is more frequent in these males. Theoretical simulations, however, revealed that the predicted process cannot occur, but they showed that the avoidance of males possessing meiotic drive does allow persistent benefits for female preference (Reinhold et al. 1999).

5. *Genetics of Sexually Selected Traits: X-chromosomal Bias*

Based on reciprocal crosses between two *Drosophila* species, Ewing (1969) long ago proposed that a disproportional part of the genes that influence traits important for mate recognition reside on the X-chromosome. Recently, two reviews revealed that X-chromosomal genes actually have a disproportionate effect on sex and reproduction related traits in humans and on sexually selected traits in animals. Using large molecular databases, Saifi and Chandra (1999) compared the linkage of mutations influencing traits related to sex and reproduction with all other traits. Their analysis shows that genes influencing traits related to sex and reproduction are several times more likely to be linked to the X-chromosome than other traits. With a different method using literature data on reciprocal crosses, Reinhold (1998) found that the influence of X-chromosomal genes is much stronger for traits that are likely to be under sexual selection than for other genes. On average, about one third of the difference between the two parental lines used for the reciprocal crosses was caused by X-chromosomal genes when sexually selected traits were considered. For those traits that were classified to be not under sexual selection, this value was much smaller—on average two percent of the difference was due to X-linked genes—and was not significantly different from zero.

Such a bias towards the X-chromosome can be expected for traits that are influenced by sexually antagonistic selection (Rice 1984) and for sex-limited traits that are under fluctuating selection (Reinhold 1999). Antagonistic selection occurs if the optimal phenotype differs for males and females and if a genetic correlation between the sexes prevents the phenotypes to reach their evolutionary stable optimum. Under sexually antagonistic selection, sex-linked traits can be expected to evolve faster than other traits because sex-linked traits almost always differ in their expression in the two sexes. Rare recessive X-chromosomal genes, for example, will always be expressed when they occur in the heterogametic sex (the sex that has two different sex chromosomes; in humans the males, they posses an X as well as a Y-chromosome) and will be expressed almost not at all in the homogametic sex. This difference in expression then provides the raw material selection can work on, so that preferential X-linkage can be expected for traits under sexually antagonistic selection. Under fluctuating selection the fate of an allele is influenced by its geometric mean fitness a fitness measure that is equivalent to the case of effective interest rates when the interest rate on a financial investment fluctuates in time. This fitness measure is influenced by the extent of expression of a trait. Between autosomal and X-chromosomal genes there is such a difference in the extent of expression: (a) X-chromosomal genes coding for sex limited male traits in heterogametic males are only expressed to one third because the other two thirds of all X-chromosomes are present in females that do not express the genes under consideration; (b) Autosomal sex-limited genes, i.e. genes that are expressed in only one sex and do not reside on the sex chromosomes but lie on any of the other chromosomes, are, in contrast, expressed to one half provided they are not totally recessive. Due to this difference in expression, autosomal genes coding for the same phenotype as X-chromosomal genes have a disadvantage compared to X-linked genes. As a consequence, X-chromosomal genes coding for sex limited traits are expected to evolve more easily than autosomal genes (Reinhold 1999). The observed X-bias for sexually selected traits can accordingly be explained by the effect of fluctuating selection on these sex-limited traits.

See also: Genetics and Mate Choice; Sex Hormones and their Brain Receptors; Sexual Attitudes and Behavior; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexuality and Gender; Y-chromosomes and Evolution

Bibliography

Andersson M 1994 *Sexual Selection*. Princeton University Press, Princeton, NJ

Bakker T C M 1999 The study of intersexual selection using quantitative genetics. *Behaviour* **136**: 1237–65

Bakker T C M, Pomiankowksi A 1995 The genetic basis of female mate preferences. *Journal of Evolutionary Biology* **8**: 129–71

Clutton-Brock T H, Vincent A C J 1991 Sexual selection and the potential reproductive rates of males and females. *Nature* **351**: 58–60

Ewing A W 1969 The genetic basis of sound production in *Drosophila pseudobscura* and *D. persimilis*. *Animal Behaviour* **17**: 555–60

Fisher R A 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK

Hamilton W D, Zuk M 1982 Heritable true fitness and bright birds: A role for parasites. *Science* **218**: 384–7

Kirkpatrick M 1996 Good genes and direct selection in the evolution of mating preferences. *Evolution* **50**: 2125–40

Kirkpatrick M, Ryan M J 1991 The evolution of mating preferences and the paradox of the lek. *Nature* **350**: 33–8

Møller A P, Alatalo R V 1999 Good-genes effects in sexual selection. *Proceedings of the Royal Society of London Series B* **266**: 85–91

Pomiankowski A, Møller A P 1995 A resolution of the lek paradox. *Proceedings of the Royal Society of London Series B* **260**: 21–9
Reinhold K 1998 Sex linkage among genes controlling sexually selected traits. *Behavioural Ecology and Sociobiology* **44**: 1–7
Reinhold K 1999 Evolutionary genetics of sex-limited traits under fluctuating selection. *Journal of Evolutionary Biology* **12**: 897–902
Reinhold K, Engqvist L, Misof B, Kurtz J 1999 Meiotic drive and evolution of female choice. *Proceedings of the Royal Society of London Series B* **266**: 1341–5
Rice W R 1984 Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38**: 735–42
Rowe L, Houle D 1996 The lek paradox and the capture of genetic variance by condition dependent traits. *Proceedings of the Royal Society of London Series B* **263**: 1415–21
Ryan M J, Keddy-Hector A 1992 Directional patterns of female mate choice and the role of sensory biases. *The American Naturalist* **139**: S4–S35
Saifi G M, Chandra H S 1999 An apparent excess of sex- and reproduction-related genes on the human X chromosome. *Proceedings of the Royal Society of London Series B* **266**: 203–9
Smith J M 1991 Theories of sexual selection. *Trends in Ecology and Evolution* **6**: 146–51
Wedekind C, Furi S 1997 Body odour preferences in men and women: Do they aim for specific MHC combinations or simply heterozygosity? *Proceedings of the Royal Society of London Series B* **264**: 1471–9
Westneat D F, Birkhead T R 1998 Alternative hypotheses linking the immune system and mate choice for good genes. *Proceedings of the Royal Society of London B* **265**: 1065–73
Wilkinson G S, Presgraves D C, Crymes L 1998 Male eye span in stalk eyed flies indicate genetic quality by meiotic drive suppression. *Nature* **391**: 276–79
Zahavi A 1975 Mate selection—a selection for a handicap. *Journal of Theoretical Biology* **53**: 205–14

K. Reinhold

Sexual Risk Behaviors

By the mid-1980s, the menace of AIDS had made indiscriminate sexual behavior a serious risk to health. Under certain conditions sexual behavior can threaten health and psychosocial well-being. Unwanted pregnancies and sexually transmitted diseases (STDs) such as gonorrhea and syphilis have long been potential negative consequences of sexual contact. Yet today HIV infection receives most attention. Because of this, the article will focus on the risk of HIV infection through sexual contact, while HIV infection through needle sharing, maternal-child transmission, and transfusion of blood and blood products will not be addressed. This article is divided into five sections. Section 1 defines sexual risk behavior and delimits the scope. Section 2 discusses the research methodology in this field. Section 3 presents epidemiological data on sexual behavior patterns and HIV infection status among relevant population groups. Section 4 examines the determinants of risk behavior. Section 5 concludes with a discussion of strategies for reducing risk exposure.

1. *Definitions*

Sexuality is an innate aspect of humanity and is oriented toward sensory pleasure and instinctual satisfaction. It goes beyond pure reproduction and constitutes an important part of sensual interaction. Sexual behavior embodies the tension between biological determination, societal and cultural norms, and an individual's personal life choices. Sexuality and sexual behavior are subject to constant change. Sexual mores and practices differ widely both within and between cultures, and also from epoch to epoch. One defining feature of sexuality is, therefore, the diversity of sexual conduct and the attitudes surrounding it.

Sexual risk behavior can be described as sexual behavior or actions which jeopardize the individual's physical or social health. High-risk sexual practices, such as unprotected intercourse with infected individuals, constitute unsafe sexual conduct, and therefore deserve to be classified as a health risk behavior or behavioral pathogen (Matarazzo et al. 1984) on a par with smoking, lack of exercise, unhealthy diet, and excessive consumption of alcohol. Among the STDs, which include gonorrhea, syphilis, genital herpes, condylomata, and hepatitis B among others, HIV infection and AIDS are by far the most dangerous. It has linked love and sexuality with disease and death. If an HIV-infected person fails to inform a partner of his or her serostatus (positive HIV antibody test), unprotected sexual intercourse resulting in the partner's infection has also legal consequences.

In general, unprotected sexual intercourse (i.e., without condoms) with an HIV-infected partner is of risk. HIV may be acquired asymptomatically and transmitted unknowingly. There is a 'hierarchy' in the level of risk involved in sexual practices. High-risk sexual practices include unprotected anal intercourse (especially for the receptive partner) and unprotected vaginal intercourse (again, more for the receptive partner) with infected individuals, as well as any other practices resulting in the entry or exchange of sexual fluids or blood among the partners. Oral sex, on the other hand, carries only a low risk (see *Sexually Transmitted Diseases: Psychosocial Aspects*). Petting is not considered sexual risk behavior. The risk of sexual behavior is a function of the partner's infection status, the sexual practices employed, and the protective measures used. Sex with a large number of sexual partners is also risky because of the higher probability of coming into contact with an infected partner. The risk can be minimized or ruled out entirely by appropriate protective measures: low-risk sexual practices, use of condoms, and avoidance of sexual contact with HIV positive partners.

Sexual risk behavior also applies to the area of family planning. Unwanted pregnancies may occur if the sexual partners do not use safe methods of contraception. If the partners do not wish to have children, sexual contact risks an undesired outcome. Unwanted pregnancy can have consequences ranging from changes in life plans, in the partnership, and in decisions to abort with the ensuing emotional stress and medical risks. Deviant sexual behavior such as exhibitionism or pedophilia cannot be defined as risk behavior.

The analysis of sexual risk behavior falls in the domains of psychology, sociology, medicine, and the public health sciences (Bancroft 1997, McConaghy 1993). Research is available in the areas of health psychology, health sociology and social science AIDS research (von Campenhourdt et al. 1997; see the journal *AIDS Care*; see *HIV Risk Interventions*). The terms used in studies more frequently than 'sexual risk behavior' are 'HIV risk behavior,' 'HIV protective behavior,' 'AIDS-preventive behavior,' 'HIV-related preventive behavior,' 'safer sexual behavior,' 'safe sex,' or 'safer sex' (DiClemente and Peterson 1994, Oskamp and Thompson 1996). These descriptions ought to be given preference over the term 'sexual risk behavior' in order to prevent pathologizing certain sexual practices; it is not the sexual practice as such which presents the risk, but the partner's infection status. Although this term will be used further here, this issue ought to be kept in mind.

2. *Research on Sexual Risk Behavior*

Assessment of sexual risk behavior is complicated by problems in research methodology. Fundamentally, research in sexology faces the same methodological problems as other social sciences. Studies of sexual behavior require special survey instruments and must take into account factors influencing data acquisition and results (see Bancroft 1997). The most important methods of data collection include questionnaires, personal interviews, telephone surveys, and self-monitoring. None of these methods is clearly superior to the others. The quality of retrospective data depends on the selected period, the memory capacity of study participants, the frequency of the behavior, and whether the behavior is typical. In general, in research on sexuality no adequate external criteria are available to validate reporting. Psychophysiological data exist only to a limited extent, and field experiments and participatory observations are usually not applicable. Since the different survey methods and instruments involve differential advantages and disadvantages, a combination of methods would be appropriate. However, this is often not feasible due to practical constraints including limited resources and constrained access to target groups.

When conducting scientific surveys on sexual behavior, it is important to inform and instruct study participants about the purpose of the investigation. Questions must be phrased in a neutral and unprejudiced manner. For sensitive issues, such as taboo topics and unusual sexual practices, the phrasing of questions and their position in the questionnaire are critical. When surveying populations with group-specific language codes (e.g., minorities that have a linguistic subculture), it is necessary to discuss sexual terms with the interviewees beforehand. The willingness to participate in the study and to answer survey questions can be influenced by the fear of reprisals if the answers become public. Depending on the objectives of the study and the study sample, relative anonymity (via telephone interview) can play an important role in the willingness to participate and the openness of responses.

Independent of the selected survey instrument, exaggerations, understatements, and socially desirable answers are to be expected, particularly to questions about sexual risk behavior. Study participants report their own behavior based on role expectations and their self-image. Thus, it is important to assess and control for appropriate motivational and dispositional variables, such as social desirability and the tendency for self-disclosure and attitudes toward sexuality. The person of the researcher, in particular his or her sex, age, and sexual orientation, has an important influence on response behavior. For example, for a heterosexual study sample it may be advisable that male participants be questioned by a male interviewer and female participants by a female interviewer. A heterosexual interviewer must be prepared not to be accepted by homosexual participants. This applies in particular when questions address highly intimate sexual experiences and sexual risk behavior.

Generalizability is one of the most serious problems when reporting data from scientific surveys on sexuality. Access to the study sample and the willingness to participate are more closely linked to relevant personal characteristics of the participants (sexual experiences, attitude toward sexuality) compared to other research. Participants tend to show, for example, greater self-disclosure, a broader repertoire of sexual activities, less guilt, and less anxiety than nonparticipants. Since randomized or quota selection are rarely possible, an exact description of the sampling frame is essential to permit an estimate of the representativity of the sample and the generalizability of the results. Generally, study findings are often based on reports by individuals who are, to a substantial degree, self-selected. This must be taken into consideration when evaluating the findings of studies on sexual behavior.

3. *Epidemiology*

Data on the epidemiology of sexual risk behavior can be acquired from different sources. Numbers can be obtained through studies on sexual risk behavior, focusing on the behavior itself. As mentioned above,

methodological problems often plague such studies. Even if rates of condom use and numbers of sexual partners are assessed correctly, they may not provide the specific information needed to present the epidemiology of sexual risk behavior. The finding that about 50 percent of heterosexual men do not regularly use condoms is only relevant if they do not use condoms in sexual encounters with potentially infected partners. This information, however, is rarely available from existing studies. Numbers can also be extrapolated from incidence and prevalence rates on sexually transmitted disease or unwanted pregnancies.

These data focus on consequences rather than the risk behavior itself, and extrapolation to the prevalence of risk behaviors is complicated by the fact that it is sometimes difficult to distinguish how an infection was transmitted (sexual contact, transfusions, or needle sharing). In addition, reporting may be incomplete or inaccurate for the following reasons: (a) not all countries require compulsory registration of HIV infections, (b) cases are often recorded anonymously which may result in multiple registrations of the same case, and (c) there may be a high number of unreported cases that do not find their way into the statistics.

Despite these methodological problems, epidemiological data are necessary and useful for behavioral research. The available epidemiological data on sexual risk behavior from European and US American sources are presented below for five different subgroups: adolescents, ethnic groups, homosexuals, heterosexuals, and prostitutes. These subgroups are not exhaustive, but they elucidate the differences and specificity of epidemiological data on sexual risk behavior (see *Sexually Transmitted Diseases: Psychosocial Aspects*).

3.1 Adolescents

Adolescents comprise a special group concerning sexual risk behavior. Their initiation into sexuality may shape their sexual behavior for many years to come. They have as yet no rigid sexual behavior patterns and are subject to influences from their peers, parents, the media, and other sources. Compared to the 1970s, teenagers are sexually active at a younger age. Many teenagers (about 25 percent of males and 40 percent of females) have had sexual intercourse by the time they reach 15 years of age, and the mean age of first sexual intercourse is between 16 and 17 years. Because adolescents are sexually active earlier in their lives, they also engage in sexual risk behavior at an earlier age. The majority have several short-term sexual relationships, and by the end of their teens about half report having had more than four partners. The majority of adolescents report being exclusively heterosexual, but an increasing number of teenage males report being bisexual or homosexual.

Data on infection rates show that sexually transmitted diseases are widespread among adolescents. The human papilloma virus (HPV) is likely to be the most common STD among adolescents, with a prevalence of 28–46 percent among women under the age of 25 in the US (Centers for Disease Control and Prevention 2000). The prevalence of HIV infection is increasing slightly among adolescents, but accurate data are difficult to obtain. Due to its long incubation period, those with AIDS in their twenties probably contracted the virus as adolescents, and such cases are on the rise. Teenage pregnancy is still an issue, although numbers generally have decreased. The numbers are still high, especially in the US, where 11 percent of all females between ages 15 and 19 become pregnant (Adler and Rosengard 1996). Data on behavior indicate that about 75 percent of adolescents use contraception. Adolescents in steady relationships predominantly use the pill, while adolescents with casual sexual encounters mainly use condoms. Most adolescents are aware of the risk of pregnancy when sexually active, and they use condoms for the purpose of contraception rather than for protection against STDs. Most of them have sufficient knowledge about HIV infection and AIDS, although erroneous assumptions still prevail. Personal vulnerability to AIDS is perceived to be fairly low, and awareness of other sexually transmitted diseases is practically nonexistent. Alcohol and drug use, common among adolescents, further influence sexual behavior among adolescents including reduced frequency of condom use.

3.2 Ethnic Groups

There are distinct differences in HIV infection among different ethnic groups. Data on infection rates are available mainly from studies conducted in the US. These show that African–Americans face the highest risk of contracting HIV. Despite making up about 12 percent of the US population, the prevalence among African–Americans is 57 percent of HIV diagnoses and 45 percent of AIDS cases. Almost two-thirds of all reported AIDS cases in women are among African–Americans. The incidence rate of reported AIDS cases is eight times the rate of whites (Centers for Disease Control and Prevention 1999). The Hispanic population has the next highest prevalence rates. Hispanics accounted for about 20 percent of the total number of new AIDS cases in 1998, while their population representation was about 13 percent. Their rate is about four times that of whites. It is likely that differences in sexual behavior underlie these statistics, but most studies of behavior have focused on one specific group rather than comparing them with each other.

Possible reasons for differences in infection rates could be that (a) members of ethnic groups often have a lower socioeconomic status and have less access to health care, (b) they are less educated and score lower on HIV risk behavior knowledge, (c) they com-

municate less with partners about sexual topics, and men often have a dominant role in relationships, such that women have difficulty discussing promiscuity and condom use, and (d) men often are less open about their sexual orientation and their HIV status compared to white men.

3.3 Homosexuals

Homosexual men (men who have sex with men) are a high-risk group for contracting HIV infection. Estimates suggest that 5 percent to 8 percent of all homosexual men are HIV positive. Unprotected sexual contact among homosexual men accounts for about 50 percent of all HIV infections (Robert Koch-Institut 1999). Compared to other groups, homosexual men are more likely to engage in sexual risk behavior, including receptive or insertive anal sex and higher numbers of sexual partners. However, prevention campaigns seem to have influenced the sexual behavior of this group. An estimated three-quarters of homosexual men now report using condoms, especially with anonymous partners. There also seems to be a decline in the overall number of sexual partners. However, some homosexual men continue to expose themselves to considerable risk of HIV infection. Although more homosexual men are using condoms, they often do not use them consistently in all potentially risky sexual encounters. Also, especially those who still engage in unprotected sex are usually sexually very active and/or have multiple partners. Studies on sexual behavior often overlook this fact when reporting rates of condom use. Homosexual women are at practically no risk of HIV infection through sexual behavior.

3.4 Heterosexuals

Heterosexuals in monogamous relationships are at low risk for contracting HIV. However, often both partners have had previous sexual contacts, and there is no guarantee that all of those past partners were not infected. Most cases of HIV transmission in this setting are male to female and result from the man's past high-risk sexual contacts, including homosexual relations and encounters with prostitutes. Sex tourism to countries with high rates of HIV adds particular risk. Differences in sexual behavior between men and women, combined with the relatively higher efficiency of HIV transmission from male to female, have the consequence that in developed countries only 5 percent of all HIV infections in men are due to unprotected heterosexual intercourse, whereas for women the figure is 33 percent.

Rates of condom use among heterosexual couples vary from study to study, but about 40–60 percent of sexually active individuals report not using condoms. No conclusive data are available in which settings condoms are used. Up to 23 percent of heterosexual men and 35 percent of heterosexual women report having had two or more sexual partners in the past five years, and up to 6 percent of men and 3 percent of women admit to extramarital sex in the past 12 months. Thus, condom use in these settings is of greatest relevance (Johnson et al. 1994, Laumann et al. 1994). Most individuals in monogamous relationships use condoms as a method of contraception rather than as a means of protection against infection.

3.5 Prostitutes

There seems to be a distinction between 'drug-related' and 'professional' prostitutes. HIV infections are fairly uncommon among professional prostitutes in developed countries; studies report rates of 0.5 percent to 4 percent. Rates of infection among drug-related prostitutes are much higher, with about 30 percent being HIV infected. Behavioral patterns differ between the two groups. Drug-related prostitutes may be forced by either financial need or dependence on a drug supplier to conform to the wishes of their clients, often including sexual intercourse without condoms. This is less the case for professional prostitutes who have more of a choice in their clients and who can insist on using condoms. Rates for unprotected sex among drug-related prostitutes range from 41 percent to 74 percent. In contrast, rates for professional prostitutes range from 20 percent to 50 percent. Rates may be even higher for male prostitutes (Kleiber and Velten 1994).

4. Determinants of Risk Behavior

Models that attempt to describe and explain HIV risk and protective behavior include the following factors: cultural factors (e.g., ethnic group, social norms), social and environmental factors (e.g., membership in subgroups, knowing HIV-infected individuals), demographic factors (socioeconomic status, marital status), biographic factors (e.g., sexual orientation, attitude toward health), and psychosocial factors (e.g., level of knowledge about STD risk, self-efficacy, attitude toward condoms). Sociological models emphasize the way risk behavior is influenced by social class, educational level and the overall social context.

Social disadvantage often goes hand in hand with lack of opportunities for health maintenance and medical care, and with a higher level of risk-taking behavior. Sexual behavior is imbedded in the mode of living of the individual and is closely tied in with the social environment. Some sociological theories focus on the aspect of communication in intimate relationships and economic factors (e.g., financial dependence). Especially among subgroups such as substance abusers, prostitutes and the socioecono-

mically disadvantaged, social and economic conditions can be the central factor that controls behavior.

Psychological models focus on determinants of risk behavior that entail processes taking place within the individual. Although a large number of studies have been published since the mid-1980s, and despite a long tradition of research into risk-taking behavior even before the era of AIDS (Yates 1992), there is no comprehensive and sufficiently well-founded theory to explain sexual risk behavior (Bengel 1996). Most significant in the field of sexual risk behavior are the social-psychological models (e.g., Theory of Reasoned Action and Planned Behavior, Theory of Protection Motivation, and the AIDS Risk Reduction Model; see *Health Behavior: Psychosocial Theories*). For the purpose of this article the AIDS Risk Reduction Model is presented because it entails the most direct links to preventive strategies.

It distinguishes between (a) demographic and personality variables, (b) labeling stage variables, (c) commitment stages variables, and (d) enactment stage variables (Catania et al. 1990). Demographic factors such as gender, age, and education, as well as personality factors such as impulsivity or the readiness to take risks contribute little in explaining sexual risk behavior (Sheeran et al. 1999). Also of limited predictive value are labeling stage variables such as knowledge of AIDS, sexual experience, and threat appraisal or risk perception. The assumption is that each individual makes a personal assessment of the risk of infection or disease (perceived vulnerability and perceived severity). A heterosexual, monogamous male, for instance, may perceive the menace of AIDS as a severe health issue, but may feel personally invulnerable or nonsusceptible. However, some individuals who conduct manifest sexual risk behaviors may underestimate their personal risk as compared to others and thereby perform an 'optimistic bias' (see *Health Risk Appraisal and Optimistic Bias*). Many studies have shown significant but tenuous correlations between threat appraisal variables and risk behavior.

Commitment stage variables influence behavior more substantially and include: (a) social influence: perception of social pressure from significant others to use or not use a condom and of a sexual partner's attitude toward condoms; (b) beliefs about condoms: attitudes toward condoms, intentions, and perceived barriers to condom use; (c) self-efficacy: confidence in the ability to protect oneself against HIV; and (d) pregnancy prevention: condom use for contraceptive purposes. Social pressure, self-efficacy, attitudes toward condoms, as well as previous use of condoms correlate closely and significantly with HIV protective behavior.

The extent to which protective or risky behavior is displayed also depends on situative and interactive factors, the enacting stage variables. Particularly among casual sexual encounters, lack of immediate condom availability can be the decisive determinant of risk behavior. The nature of the relationship and, in particular, the level of communication about safe sex play a central role in risk behavior. Can the partners communicate about HIV and protective behavior, or are they afraid of offending the partner and jeopardizing what could otherwise be a valued romance?

The significance of the influencing factors above varies, depending on the target behavior (e.g., condom use, sexual practices) and on target group (e.g., homosexuals, adolescents, prostitutes; see, e.g., Flowers et al. 1997). All available theoretical approaches and models assume that HIV-protective behavior is governed by a rational decision process. Emotional and motivational factors, as well as planned behavior and action control, largely have been disregarded in these models and have also been insufficiently researched. After experiencing a risk situation, individuals change both their risk perception and their appraisal of the options available for risk management. Especially when uncertainty or fear about HIV infection is high, cognitive coping (e.g., 'I know my partner's friends, so he is not HIV infected') and behavioral coping (e.g., seeking HIV antibody testing) are deployed.

5. *Strategies for Behavioral Change*

Reducing the rates of infections with HIV and other sexually transmitted diseases and preventing unwanted pregnancies constitute major tasks for health science and policy. Although sexual risk behavior in most important target groups is difficult to assess, and explanatory models lack empirical validation and are incomplete, preventive programs must be developed and implemented (Kelly 1995). A societal agreement on target groups and on methods used in such prevention programs is essential. The need to prevent the spread of AIDS has triggered lively and controversial discussions in many countries: should the emphasis be on information and personal responsibility, or should regulatory measures be employed to stem the tide of the disease? Controversy has raged among scientists about the ability to influence sexual behavior and the right of the state to intervene in such intimate affairs. There is agreement, however, that sexual risk behavior cannot be regarded as an isolated mode of personal conduct, but must be seen in the context of an individual's lifestyle and social environment.

Prevention programs promote the use of condoms as the basic method of protection. They promulgate a message of personal responsibility to prevent risk: 'protect yourself and others.' AIDS prevention programs in Western European and North American countries have pursued two objectives: (a) to convey basic information on modes of transmission and methods of protection and (b) to motivate the popu-

lation to assess individual risk and undertake behavioral change if needed. These recommendations start by urging 'safe sex,' that is, use of condoms and avoidance of sexual practices in which body fluids are exchanged, in any situation where infection is a potential risk. Individuals are also advised to reduce the number of sexual partners, to avoid anonymous sexual partners and to reduce the use of substances that may result in loss of control.

Prevention programs use mass communicative, personal, and structural measures. Mass communication involves dissemination of information via media such as radio, TV, newspaper, and posters, as well as distribution of brochures and informational leaflets. These media may be intended for all audiences or may be aimed at a specific group. The information is conveyed in simple, concrete, everyday language, describing the modes of transmission and the options available for protection. Personal measures include telephone hotlines, events, and seminars for special target groups, street-working, individual counseling, and information for sex tourists. These person-to-person prevention programs aim at fostering recognition of the problem, improving the level of knowledge, and changing attitudes, intentions, and behaviors of members of particular target groups. Structural measures include the provision of sterile syringes for intravenous drug users, easy access to condoms, and the improvement of the social situation of prostitutes. Preventive measures must be tailored to the lifestyle, environment, and language of each target group given its specific risk behavior pattern and risk of HIV infection. Programs that rely entirely on the generation of fear enhance risk perception but offer no coping alternatives. Specific messages alone, for instance an appeal for condom use, are also often inadequate to bring about (permanent) behavior changes. Communication among sexual partners should be encouraged as one of the crucial target parameters of prevention.

Evaluation of prevention programs has suggested that they have been successful at conveying the most important information about AIDS. In certain sections of the population there are nevertheless uncertainties, irrational assumptions, and false convictions about risk of infection through such activities as kissing or work contacts. The acceptance level of this information varies widely depending on the target group and is lowest among intravenous drug users. As expected, outside of those groups at especially high risk, the greatest fear of infection prevails among persons below the age of 35 and among singles. Only moderate behavioral change has been found among intravenous drug users. Rapid and significant changes in behavior have been found among homosexual males, especially in cities (fewer sexual partners, fewer high-risk sexual practices, increased use of condoms). Yet only a minority practices safe sex all the time, and some of the behavioral changes are relatively unstable. Some experts fear that the messages of preventive campaigns are wearing off and that some individuals are now less concerned about becoming infected than in the past. This may be due to impact of the new antiviral drugs, which have changed the perception of HIV from that of a death sentence to that of a chronic, manageable disease.

See also: Adolescent Health and Health Behaviors; Health Behavior: Psychosocial Theories; Health Behaviors; Health Risk Appraisal and Optimistic Bias; HIV Risk Interventions; Regulation: Sexual Behavior; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexually Transmitted Diseases: Psychosocial Aspects

Bibliography

Adler N E, Rosengard C 1996 Adolescent contraceptive behavior: Raging hormones or rational decision making? In: Oskamp S, Thompson S C (eds.) *Understanding and Preventing HIV Risk Behavior. Safer Sex and Drug Use*. Sage, Thousand Oaks, CA

Bancroft J (ed.) 1997 *Researching Sexual Behavior. Methodological Issues*. Indiana Press, Bloomington, IN

Bengel J (ed.) 1996 *Risikoverhalten und Schutz vor AIDS: Wahrnehmung und Abwehr des HIV-Risikos—Situationen, Partnerinteraktionen, Schutzverhalten* (Risk behavior and protection against AIDS: Perception of and defense against the risk of HIV—situations, partner interactions, protective behavior). Edition Sigma, Berlin

von Campenhourdt L, Cohen M, Guizzardi G, Hausser D (eds.) 1997 *Sexual Interactions and HIV-Risk*. Taylor & Francis, London

Catania J A, Kegeles S M, Coates T J 1990 Towards an understanding of risk behavior: An AIDS risk reduction model. *Health Education Quarterly* **17**: 53–72

Centers for Disease Control and Prevention 1999 *HIV/AIDS among African–Americans* (On-line). Available: http://www.cdc.gov/hiv/pubs/facts/afam.pdf

Centers for Disease Control and Prevention 2000 *Tracking the hidden epidemics. Trends in the STD epidemics in the United States* (On-line). Available: http://www.cdc.gov/nchstp/dstd/Stats_Trends/Stats_and_Trends.htm

DiClemente R J, Peterson J L (eds.) 1994 *Preventing AIDS*. Plenum, New York

Flowers P, Sheeran P, Beail N, Smith J A 1997 The role of psychosocial factors in HIV risk-reduction among gay and bisexual men: A quantitative review. *Psychology and Health* **12**: 197–230

Johnson A M, Wadsworth J, Wellings K, Field J 1994 *Sexual Attitudes and Lifestyles*. Blackwell, Oxford, UK

Kelly J A 1995 *Changing HIV Risk Behavior: Practical Strategies*. Guilford Press, New York

Kleiber D, Velten D 1994 *Prostitutionskunden. Eine Untersuchung über soziale und psychologische Charakteristika von Besuchern weiblicher Prostituierter in Zeiten von AIDS*. (Clients of prostitutes. An investigation into social and psychological characteristics of customers of female prostitutes in the era of AIDS). Nomos Verlagsgesellschaft, Baden-Baden

Laumann E O, Gagnon J H, Michael R T, Michaels S 1994 *The Social Organization of Sexuality. Sexual Practices in the United States*. University of Chicago Press, Chicago
Matarazzo J D, Weiss S M, Herd J A, Miller N E, Weiss S M (eds.) 1984 *Behavioral Health: A Handbook of Health Enhancement and Disease Prevention*. Wiley, New York
McConaghy N (ed.) 1993 *Sexual Behavior. Problems and Management*. Plenum, New York
Oskamp S, Thompson S C (eds.) 1996 *Understanding and Preventing HIV Risk Behavior. Safer Sex and Drug Use*. Sage, Thousand Oaks, CA
Robert Koch-Institut 1999 *AIDS/HIV Halbjahresbericht I/99. Bericht des AIDS-Zentrums im Robert Koch-Institut über aktuelle epidemiologische Daten* (AIDS/HIV semiannual report I/99. Report by the AIDS Center at the Robert Koch-Institute on Current Epidemiological Data). Berlin
Sheeran P, Abraham C, Orbell S 1999 Psychosocial correlates of heterosexual condom use: A meta-analysis. *Psychology Bulletin* **125**: 90–132
Yates J F 1992 *Risk-taking Behavior*. Wiley, Chichester, UK

J. Bengel

Sexuality and Gender

The meaning of the terms sexuality and gender, and the ways that writers have theorized the relationship between the two, have changed considerably over the last 40 years. The term sexuality has various connotations. It can refer to forms of behavior, it may include ideas about pleasure and desire, and it is also a term that is used to refer to a person's sense of sexual being, a central aspect of one's identity, as well as certain kinds of relationships. The concept of gender has also been understood in relation to varying criterion. Prior to the 1960s, it was a term that referred primarily to what is coded in language as masculine or feminine. The meaning of the term gender has subsequently been extended to refer to personality traits and behaviors that are specifically associated either with women or men; to any social construction having to do with the male/female distinction, including those which demarcate female bodies from male bodies; to gender being thought of as the existence of materially existing social groups 'men' and 'women' that are the product of unequal relationships. In this latter sense, gender as a socially meaningful category is dependant on a hierarchy already existing in any given society, where one class of people (men) have systematic and institutionalized power and privilege over another class of people (women) (Delphy 1993). The term *patriarchy* or, more recently, the phrase 'patriarchal gender regimes,' is used as a way of conceptualizing the oppression of women which results.

More recently, the notion of gender as social practice has emerged and is associated with the work of Judith Butler (1990), who argues that gender is performatively enacted through a continual citation and reiteration of social norms. Butler offers a similar analysis of sexuality, claiming that far from being fixed and naturally occurring, (hetero) sexuality is 'unstable,' dependant on ongoing, continuous, and repeated performances by individuals 'doing heterosexuality,' which produce the illusion of stability. There is no 'real' or 'natural' sexuality to be copied or imitated: heterosexuality is itself continually in the process of being reproduced.

Such ideas are part of the establishment of a new canon of work on sexuality and gender that has emerged since the 1960s. This newer approach differs radically from the older tradition put forward by biologists, medical researchers, and sexologists, which developed through the late nineteenth century and was profoundly influential during the first half of the twentieth century. The traditional approach to understanding sexuality and gender has been primarily concerned with establishing 'natural' or 'biological' explanations for human behavior. Such analyses are generally referred to as *essentialist*. More recent approaches, although not necessarily denying the role of biological factors, have emphasized the importance of social and cultural factors; what is now commonly known as the *social constructionist* approach.

The sociological study of sexuality emerged in the 1960s and 1970s and was informed by a number of theoretical approaches that were significant at the time; notably symbolic interactionism, labeling theory and deviancy theory, and feminism. The work of writers such as John Gagnon and William Simon (1967, 1973) in the US, and Mary McIntosh (1968) and Kenneth Plummer (1975) in the UK, was particularly important in establishing a different focus for thinking about sexuality. A primary concern of such works was to highlight how sexuality is social rather than natural behavior and, as a consequence, a legitimate subject for sociological enquiry.

Gagnon and Simon developed the notion of *sexual scripts* which, they argued, we make use of to help define who, what, where, when, how, and—most anti-essentialist of all—why we have sex. A script refers to a set of symbolic constructs which invest actors and actions, contexts and situations, with 'sexual' meaning—or not as the case may be. People behave in certain ways according to the meanings that are imputed to things; meanings which are specific to particular historical and cultural contexts; meanings that are derived from scripts learnt through socialization and which are modified through ongoing social interactions with others. Most radical of all, Gagnon and Simon claimed that not only is sexual conduct socially learnt behavior, but the reason for wanting to engage in sexual activity, what in esssentialist terms is referred to as sexual 'drive' or 'instinct,' is in fact a socially learnt goal. Unlike Freud, who claimed the opposite to be true, Gagnon and Simon suggested that social motives underlie sexual actions. They saw gender as central to this and detailed how in con-

temporary Western societies sexual scripts are different for girls and boys, women and men. Here gender is seen as a central organizing principle in the interactional process of constructing sexual scripts. In this sense gender can be seen as constitutive of sexuality, at the same time as sexuality can be seen as expressive of gender. Thus, for example, Gagnon and Simon argue that men frequently express and gratify their desire to appear 'masculine' through specific forms of sexual conduct. For example, for young men in most Western cultures first sexual intercourse is a key moment in becoming a 'real man,' whereas this is not the same for young women. It is first menstruation rather than first heterosex that marks being constituted as 'women.'

Another important contribution to the contemporary study of sexuality, which has posed similar challenges to essentialist theories, is the *discourse analysis* approach. One example is the work of Michel Foucault (1979) and his followers, who claim that sexuality is a modern 'invention' and that by taking 'sexuality' as their object of study, various discourses, in particular medicine and psychiatry, have produced an artificial unity of the body and its pleasures: a compilation of bodily sensations, pleasures, feelings, experiences, and actions which we call 'the sexual.' Foucault understands sex not as some essential aspect of personality governed by natural laws that scientists may discover, but as an *idea* specific to certain cultures and historical periods. Foucault draws attention to the fact that the history of sexuality is a history of changing forms of regulation and control over sexuality. What 'sexuality' is defined as, its importance for society, and to us as individuals may vary from one historical period to the next. Furthermore, Foucault argues, as do interactionists, that sexuality is regulated not only through prohibition, but is produced through definition and categorization, in particular through the creation of sexual categories such as, for example, 'heterosexual' and 'homosexual.' Foucault argues that, while both heterosexual and homosexual behavior has existed in all societies, there was no concept of a person whose sexual identity is 'homosexual' until relatively recently. Although there is some disagreement among writers as to precisely when the idea of the homosexual person emerged, it has its origins in the seventeenth to nineteenth centuries, with the category lesbian emerging somewhat later than that of male homosexuality. Such analyses have also highlighted how medical and psychiatric knowledge during the late nineteenth and early twentieth centuries was a key factor in the use of the term 'homosexual' to designate a certain type of person rather than a form of sexual conduct.

A major criticism of Foucault's work is that insufficient attention is given to examining the relationship between sexuality and gender. Feminist writers in particular have pointed out how in Foucault's account of sexuality there is little analysis of how women and men often have different discourses of sexuality. Sexuality is employed as a unitary concept and, such critic's claim, that sexuality is male. Despite such criticisms, many feminists have utilized Foucauldian perspectives.

A further challenge to essentialist ideas about sexuality and gender is associated with psychoanalysis, in particular the reinterpretation of Freud by Jacques Lacan. For Lacan and his followers sexuality is not a natural energy that is repressed; it is language rather than biology that is central to the construction of 'desire.' Lacanian psychoanalysis has had a significant influence on the development of feminist theories of sexuality and gender, although some writers have been critical of Lacan's work (Butler 1990).

At the same time as social scientists and historians were beginning to challenge the assumption that sexual desires and practices were rooted in 'nature,' more and more people were beginning to question dominant ideas about gender roles and sexuality. The late 1960s/early 1970s saw the emergence of both women's and gay and lesbian liberation movements in the US and Europe. An important contribution to analyses of sexuality and gender at that time was the distinction feminists, along with some sociologists and psychologists, sought to make between the terms sex and gender. Sex referred to the biological distinction between females and males, whereas gender was developed and used as a contrasting term to sex. Gender refers to the social meanings and value attached to being female or male in any given society, expressed in terms of the concepts femininity and masculinity, as distinct from that which is thought to be biologically given (sex). Feminists have used the sex/gender distinction to argue that although there may exist certain biological differences between females and males, societies superimpose different norms of personality and behavior that produce 'women' and 'men' as social categories. It is this reasoning that led Simone de Beauvoir (1964) to famously remark 'One is not Born a Woman.'

More recently, a new understanding of gender has emerged. Rather than viewing sex and gender as distinct entities, sex being the foundation upon which gender is superimposed, gender has increasingly been used to refer to any social construction to do with the female/male binary, including male and female bodies. The body, it is argued, is not free from social interpretation, but is itself a socially constructed phenomenon. It is through understandings of gender that we interpret and establish meanings for bodily differences that are termed sexual difference (Nicholson 1994). Sex, in this model, is subsumed under gender. Without gender we could not read bodies as differently sexed; gender provides the categories of meaning for us to interpret how the body appears to us as 'sexed.'

Feminists have critiqued essentialist understandings of both sexuality and gender and have played an

important role in establishing a body of research and theory that supports the social constructionist view. However, feminist theories of sexuality are not only concerned with detailing the ways in which our sexual desires and practices are socially shaped; they are also concerned to ask how sexuality relates to gender and, more specifically, what the relationship is between sexuality and gender inequality? It is this question which perhaps more than any other provoked discussion and controversy between feminists during the 1970s and 1980s.

Most feminists would agree that historically women have had less control in sexual encounters than their male partners and are still subjected to a double standard of sexual conduct that favors men. It is, for example, seen as 'natural' for boys to want to have sex and with different partners, whereas exactly the same behavior that would be seen as understandable and extolled in a boy is censured in a girl. Sexually active women are subject to criticism and are in danger of being regarded as a 'slut' or a 'slag.' Where feminists tend to differ is over the importance of sexuality in understanding gendered power differences. For many radical feminists sexuality is understood to be one of the key mechanisms through which men have regulated women's lives. Sexuality, as it is currently constructed, is not merely a reflection of the power that men have over women in other spheres, but is also productive of those unequal power relationships. Sexuality both reflects and serves to maintain gender divisions. From this perspective the concern is not so much how sexual desires and practices are affected by gender inequalities, but, more generally, how constructions of sexuality constrain women in many aspects of their daily lives from restricting their access to public space to shaping health, education, work, and leisure opportunities (Richardson 1997). Fears of sexual violence, for instance, may result in many women being afraid to go out in public on their own, especially at night. It is also becoming clearer how sexuality affects women's position in the labor market in numerous ways; from being judged by their looks as right or wrong for the job, to sexual harassment in the workplace as a common reason given by women for leaving paid employment.

Other feminists have been reluctant to attribute this significance to sexuality in determining gender relations. They prefer to regard the social control of women through sexuality as the outcome of gendered power inequalities, rather than its purpose. There is then a fundamental theoretical disagreement within feminist theories of sexuality, over the extent to which sexuality can be seen as a site of male power and privilege, as distinct from something that gendered power inequalities act upon and influence.

In the 1990s a new perspective on sexuality and sexual politics emerged fueled by the impact of HIV and AIDS on gay communities and the anti-homosexual feelings and responses that HIV/AIDS revitalized, especially among the 'moral right.' One response by scholars was queer theory; a diverse body of work that aims to question the assumption in past theory and research that heterosexuality is 'natural' and normal. Queer theory is often identified, especially in its early stages of development, with writers associated with literary criticism and cultural studies, and generally denotes a postmodernist approach to understanding categories of gender and sexuality. The work of Eve Sedgwick (1990), Judith Butler (1990), and Teresa de Lauretis (1991), for instance, might be taken as key to the development of queer theory.

A principal characteristic of queer theory is that it problematizes sexual and gender categories in seeking the deconstruction of binary divides underpinning and reinforcing them such as, for instance, woman/man; feminine/masculine; heterosexual/homosexual; essentialist/constructionist. While queer theory aims to develop existing notions of gender and sexuality, there are broader implications of such interventions. As Sedgwick (1990) and others have argued, the main point is the critique of existing theory for its heterosexist bias rather than simply the production of theory about those whose sexualities are marginalized such as, for example, lesbians and gay men.

Sexuality is the primary focus for analysis within queer theory and, while acknowledging the importance of gender, the suggestion is that sexuality and gender can be separated analytically. In particular, queer theorists are centrally concerned with the homo/heterosexual binary and the ways in which this operates as a fundamental organizing principle in modern societies. The emphasis is on the centrality of homosexuality to heterosexual culture and the ways in which the hetero/homo binary serves to define heterosexuality at the center, with homosexuality positioned as the marginalized 'other.' Feminist perspectives, on the other hand, have tended to privilege gender in their analyses—the woman/man binary—and, as I have already outlined above, are principally concerned with sexuality insofar as it is seen as constitutive, as well as determined by, gendered power relations.

Of particular significance for the development of our understanding of the relationship between queer and feminism is a rethinking of the distinction between sexuality and gender.

The relationship between sexuality and gender has, then, been theorized in different ways by different writers. These can be grouped into five broad categories. First, some theories place greater emphasis on gender insofar as concepts of sexuality are understood to be largely founded upon notions of gender (Gagnon and Simon 1973, Jackson 1996). For example, it is impossible to talk meaningfully about heterosexuality or homosexuality without first having a notion of one's sexual desires and relationships as being directed to a particular category of gendered persons. Others propose a different relationship, where sexu-

ality is understood to be constitutive of gender. The radical feminist Catherine MacKinnon (1982), for example, suggests that it is through the experience of 'sexuality,' as it is currently constructed, that women learn about gender, learn what 'being a woman' means. 'Specifically, "woman" is defined by what male desire requires for arousal and satisfaction and is socially tautologous with "female sexuality" and the "female sex"' (MacKinnon 1982). A third way of understanding the relationship between sexuality and gender, which moves away from causal relationships where one is seen as more or less determining of the other, is one which allows for the possibility that the two categories can be considered as analytically separate, if related, domains. Gayle Rubin (1984) in her account of what she terms a 'sex/gender system,' and others who have been influenced by her work, such as Eve Sedgwick (1990), make this distinction between sexuality and gender, which means that it is possible for sexuality to be theorized apart from the framework of gender difference. This is a model favored by many queer theorists. Alternatively, we may reject all of these approaches in favor of developing a fourth model which relies on the notion that sexuality and gender are inherently co-dependent and may not usefully be distinguished one from the other (Wilton 1996).

The stage in our contemporary understandings of sexuality and gender are such that there can be no simple, causal model that will suffice to explain the interconnections between them. However, rather than wanting to privilege one over the other, or seeking to analytically distinguish sexuality and gender or, alternatively, to collapse the two, we might propose a fifth approach which investigates 'their complex interimplication' (Butler 1997). It is this articulation of new ways of thinking about sexuality and gender in a dynamic, historically, and socially specific relationship that is one of the main tasks facing both feminist and queer theory (Richardson 2000).

See also: Female Genital Mutilation; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender and Reproductive Health; Gender Differences in Personality and Social Behavior; Gender Ideology: Cross-cultural Aspects; Heterosexism and Homophobia; Lesbians: Historical Perspectives; Male Dominance; Masculinities and Femininities; Prostitution; Queer Theory; Rape and Sexual Coercion; Rationality and Feminist Thought; Regulation: Sexual Behavior; Sex Offenders, Clinical Psychology of; Sex-role Development and Education; Sex Therapy, Clinical Psychology of; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexual Risk Behaviors; Sexuality and Geography; Teen Sexuality; Transsexuality, Transvestism, and Transgender

Bibliography

Butler J 1990 *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York

Butler J 1997 Critically queer. In: Phelan S (ed.) *Playing with Fire: Queer Politics, Queer Theories*. Routledge, London

de Beauvoir S 1964 *The Second Sex*. Bantam Books, New York

de Lauretis T 1991 Queer theory: lesbian and gay sexualities: an introduction. *Differences: Journal of Feminist Cultural Studies* **3**(2): iii–xviii

Delphy C 1993 Rethinking sex and gender. *Women's Studies International Forum* **16**(1): 1–9

Foucault M 1979 *The History of Sexuality*. Allen Lane, London, Vol. 1

Gagnon J H, Simon W (eds.) 1967 *Sexual Deviance*. Harper & Row, London

Gagnon J H, Simon W H 1973 *Sexual Conduct: The Social Sources of Human Sexuality*. Harper & Row, London

Jackson S 1996 Heterosexuality and feminist theory. In: Richardson D (ed.) *Theorising Heterosexuality*. Open University Press, Buckingham, UK

MacKinnon C A 1982 Feminism, Marxism, method and the state: an agenda for theory. *Signs: Journal of Women in Culture and Society* **7**(3): 515–44

McIntosh M 1968 The homosexual role. *Social Problems* **16**(2): 182–92

Nicholson L 1994 Interpreting gender. *Signs: Journal of Women in Culture and Society* **20**(1): 79–105

Plummer K 1975 *Sexual Stigma: An Interactionist Account*. Routledge and Kegan Paul, London

Rich A 1980 Compulsory heterosexuality and lesbian existence. *Signs: Journal of Women in Culture and Society* **5**(4): 631–60

Richardson D 1997 Sexuality and feminism. In: Robinson V, Richardson D (eds.) *Introducing Women's Studies: Feminist Theory and Practice*, 2nd edn. Macmillan, Basingstoke, UK

Richardson D 2000 *Re: Thinking Sexuality*. Sage, London

Rubin G 1984 Thinking sex: notes for a radical theory of the politics of sexuality. In: Vance C S (ed.) *Pleasure and Danger: Exploring Female Sexuality*. Routledge, London

Sedgwick E K 1990 *Epistemology of the Closet*. University of California Press, Berkeley, CA

Wilton T 1996 Which one's the man? The heterosexualisation of lesbian sex. In: Richardson D (ed.) *Theorising Heterosexuality*. Open University Press, Buckingham, UK

D. Richardson

Sexuality and Geography

Geographical questions concerning sexuality are ideally investigated by combining empirical work with theories of how and why particular sexualities garner spatial force and cultural meaning. In practice, most

theoretical treatments of sexuality in geography have dealt with normative forms of heterosexuality in industrial contexts. Queries about heterosexuality have centered partially on theorizing why and how procreation is given symbolic, spatial, and practical centrality in different cultural and historical contexts. Other theoretical queries have involved analyzing how heterosexual metaphors and expectations inform the structure and meaning of language, discourse, and place. For instance, what is the historical and spatial significance of saying master bedroom, cockpit, capitalist penetration, motherland, virgin territory, mother-earth, or father-sky? In contrast, empirical work has centered on documenting where nonheterosexualities take place, how states and societies regulate nonheterosexual places, bodies, and practices, and the sociospatial difficulties and resistances faced by those who desire nonheterosexual places and identities. Some empirical work also addresses the marginalized places of sex workers in heterosexual sex-work trades.

The disparity between theoretical vs. empirical work reflects how knowledge and power intersect spatially. Most persons today are compelled to assume, and are hence familiar with, the bodily and spatial arrangements of heterosexual gestures and public displays of affection, procreational practices (marriages, buying a 'family' home, having children) and forms of entertainment. Consequently, little explanatory or descriptive analytical space is needed. In contrast, the bodily and spatial arrangements of those assuming non-normative sexualities are commonly hidden from view and relatively unknown. They therefore require documentation before theorization can proceed.

Intersections between geography and sexuality cannot be understood singularly by listing research accomplished to date. Rather, discussion about how geographers have formally and informally navigated sexuality debates is needed to contextualize how external politics have informed internal concerns and research agendas.

1. The Sociology of Sexuality Research in Geography

Few geographers have, until recently, explored how sexuality shapes and is shaped by the social and spatial organization (what geographers call the sociospatiality) of everyday life. Fewer still have explored how geography as a discipline is shaped by heterosexual norms and expectations, or what is called heteronormativity. The ways that modern heterosexuality informs everyday life are not seen or studied because it is taken as natural. Hence, most people unconsciously use it to structure the sociospatiality and meaning of their lives, living out and within what Butler (1990) calls a 'heterosexual matrix.' Many cultural settings in which heterosexuality predominates place great practical and ideological value on procreation. Here, couplings of persons with genitalia qualified as oppositely sexed are presumed to be biologically natural, feeding into culturally constructed ideals of gender identities and relations (WGSG 1997). In modern industrial contexts, boys and girls are expected to grow up to be fathers and mothers who settle into a nuclear household to procreate. Given modern heterosexuality's centrality in determining what is natural, it commonly grounds notions of ontology (being), epistemology (ways of knowing), and truth.

The socially constructed 'naturalness' of heterosexuality in modern contexts is produced and reproduced at a variety of geographical scales. Many nation-states, for example, define and regulate heterosexuality as the moral basis of state law and moral order (e.g., The Family), forbidding sexual alliances not ostensibly geared towards procreation (Nast 1998).

Such laws and related political, cultural, and epistemological strictures have historically limited the kinds of spatial questions geographers have thought, imagined, or asked. Today, precisely when the existence and sovereignty of nation-states is being challenged by transnational actors so, too, is the relevance of the nuclear family being questioned, suggesting partial societal and political symbiosis between family and state types. Nevertheless, for the time being, heterosexual strictures predominate across sociospatial domains, helping to account for geography's generalized disciplinary anxiety towards: (a) those who conduct research on, or question, how heterosexuality is sociospatially made the norm, (b) geographers who hold queer identities, and (c) those engaged in queer identity research. The word queer, here, refers to persons practicing non-normative kinds of sexuality. The word is not meant to efface sexuality differences, but to stress the oppositional contexts through which those who are non-normative are made marginal (see Elder et al. in press). By the late 1980s geographers in significant numbers had brought analytical attention to the violence involved in sustaining what Rich (1980) has named, 'compulsory heterosexuality' (e.g., Foord and Gregson 1986, Bell 1991, Valentine 1993a, 1993b, WGSG 1997, Nast 1998). Part of this violence has to do with the fact that most geographic research fails to question how heterosexuality is spatially channeled to shape the discipline and everyday life.

Heterosexuality's incorporation into the micropractices and spaces of the discipline means that heteronormativity dominates the ways geographers investigate space. Witness the sustained heterosexist practices and spaces of geography departments and conferences (e.g., Chouinard and Grant 1995, Valentine 1998), heterosexist geographical framings or epistemologies of space, nature, and landscape, and the discipline's heterosexist empirical interests, divisions, and focii (e.g., Rose 1993, Binnie 1997a). Even

feminist geographical analyses rarely show how constructions of gender relate to the heterosexual practices to which they are attached Foord and Gregson 1986, Townsend in WGSG 1997, p. 54). In these ways, heterosexuality underpins geographical notions of self, place, sex, and gender (Knopp 1994). Consequently, geographers reproduce hegemonic versions of hetero sex and oedipal (or, nuclear) family life, helping to make these appear legitimate, natural, morally right, and innocent.

2. *Theoretical Research on Heterosexuality and Geography*

Perhaps the first geographers to *theorize* heterosexuality's impact were Foord and Gregson (1986). Using a realist theoretical framework, they argued that gender divisions of male and female derive from procreation and, hence, heterosexuality. Later, Rose (1993) demonstrated how the language, interests, methods, theories, and practices of certain subfields in geography (particularly time geography and humanistic geography) bely a masculinity that, respectively, ignores women's lives or renders them nostaglic objects of desire, not unlike how landscapes and nature have been romanticized and feminized in the discipline. Yet the masculinity Rose describes in heterosexual; it is heteromasculinity. Heteromasculinity can additionally be argued to reside in epistemologies and empirical focii of geographic research generally (Knopp 1994), especially physical geography and some Marxian analyses of space.

Nast (1998) theorizes that modern heteromasculinity is not a singular entity set in structural opposition to femininity asin traditional binary theories of gendered binaries. Rather, the heteromasculine is structured along two imagined, symbolized, and enaced avenues: (a) the filial (son-like) and (b) the paternal. The filial is constructed as hyperembodied and hence celebrates-as-it-constructs men's superior physical strength and courage. The paternal is different in that it celebrates men's superior intellectual strength, objectivity, and cunning and is written about as though disembodied. These *two* masculinities are antagonistically inter-related in an imaginary and symbolic framework supportive of the prototypical industrial, oedipal family.

In the context of globalization and nationalisms, Gibson-Graham (1999) and Nast (1998), respectively, theorize how heterosexualities and geographies discursively and practically make one another. Gibson-Graham analyzes the striking rhetorical parallels between some Marxian theoretical constructions of capital and some social constructions of rape that cast women as victims. Thus, many Marxian theories script capitalism as violently irrepressible, a force that destroys noncapitalist social relations with which it comes into contact. Such theorization gathers force from heterosexualized language and imagery, such that capitalism and its effects are rendered *naturally* and forever singular, hard, rapacious, and spatially penetrative. Analogously, raped women have commonly been scripted as victims of men's *naturally* and singularly superior strength and virility. Rape, once initiated, is judged as unstoppable or inevitable. Women are therefore encouraged to submit (like noncapitalist groups) to heteromasculinity's logic-violence. For Gibson-Graham, what needs to be resisted is not the 'reality' that women and markets are naturally rapable or that men and capitalism are monolithically superior and rapacious. Rather, imaginaries and discursive formations need to be deployed that allow for difference and agency and that support, facilitate, and valorize resistance and change.

Nast (1998) drawing upon Foucaultian notions of discourse as *useful*, similarly shows how heteronormativity is dispersed across sociospatial circuits of everyday life, including discourses and practices of the body, nation, transnation, and world. Focusing on modern nationalist language and imagery, she shows that many representations of nation-states depend upon images of a pure *maternal* and pure nuclear family. In these instances, heteromasculine control over women's bodies and procreation is commonly linked to eugenics or racialized notions of national purity. Obversely, some nation-states, particularly fascist ones, are represented using phallic language and imagery that glorify the oedipal (nuclear) family *father*. Cast as superior protector of family and nation, he and his sons defend against that deemed weak and polluting, including other 'races' and the maternal. Nuclear familial imagery and practices have also been used in the context of industrial interests providing, for example, practical and aesthetic means for systematically alienating and reproducing labor. Here, the nuclear family is a core setting for producing and socializing industrial workers, with the family and family home fetishized as natural protected places of privacy and peace.

3. *Empirical Research on Sexuality and Geography in Sociospatial Context*

Most sexuality research in geography consists of empirical studies of sexuality. In many cases the research is undertheorized, partially reflecting the nascent state of sexuality research wherein specificities of sexualities are mapped out before they are theorized across contexts. Many of those documenting the spatial contours of sexualized bodies and practices have focused on: (a) marginalized heterosexuals, such as sex workers, and (b) queer bodies and places. The former is heteronormative to the degree that the researchers do not specify that it is heterosexuality

with which they work; they also do not theorize heterosexuality's specific sociospatial force in the contexts they study. Consequently their work reinforces popular senses that when one speaks of sex, it is naturally and singularly heterosex. In contrast, working against the grain of heterosexual norms and [expectations] because they are researchers who study the spatiality of queer bodies and places call attention to the specific sexuality(ies) through and against which they work.

Perhaps because in some postindustrial contexts the nuclear family and procreational economies went into decline in the 1970s (alongside the Fordist family), social movements involving queer sexualities obtained greater popular recognition and political and material might, especially where related to gay men. It was at this time that a first wave of queer geographical work broke surface, dealing primarily with white gay male investors in urban contexts and drawing largely upon Marxian theories of rent. Many of these works were unpublished and bore descriptive titles (but see Weightman 1980; for unpublished works, see Elder et al. in press).

In tandem with increased scholarly production was an increase in the scholarly *visibility* of queer geographers (see Elder et al. in press). Evidence of increased networking among sexuality researchers is *Mapping Desire*, the first edited collection on sexuality and geography (Bell and Valentine 1995). The collection, like other works produced towards the late 1980s and 1990s, reflects broad geographical interests, ranging from heterosexual prostitution in Spain, to state surveillance and disciplinary actions towards Black men and women in apartheid South Africa, to the 'unsafeness' of the traditional nuclear family home for many lesbians. The works consists mostly of case studies detailing how marginalized identities are lived out, negotiated, and contested. Ironically, given the collection's title, no spatial theories of desire are presented.

Institutional recognition of queer geographers or sexuality research was not sought in English-speaking contexts until the 1990s and only then, in the United States and Britain. In the 1980s, members of the Association of American Geographers (AAG) formed the Lesbigay Caucus, subsequently holding organizational meetings at annual AAG meetings and co-sponsoring sessions. In 1992 a small group of British geographers created The Sexuality and Space Network, a research and social network affilated loosely with the Institute of British Geographers. That same year, members of the Lesbigay Caucus decided to work for creation of a Sexuality and Space Specialty Group within the AAG which, unlike the caucus, would sponsor sessions and research committed to exploring inter-relations between sexuality and space. The Specialty Group was granted official status in 1996, becoming the first group of its kind in the discipline.

Empirical work on sexuality and geography is gendered, especially in the context of queer research. Most scholars writing about lesbian spaces are women documenting geographies of women's fear and discrimination and how these are resisted at different scales and in different contexts. In contrast, most researchers addressing gay male spaces are men. In contrast to lesbian-related research, that related to gay men speaks mostly to proactive market measures successfully taken in securing urban-based private property and terrorities, particularly queer cultural districts produced through large investments in real estate and small business. In either case, western contexts prevail.

Thus, Patricia Meono-Picado (Jones et al. 1997) documents how Latina lesbians in New York used spatial tactics to oppose the homophobic programming of a Spanish-language radio station, Valentine (1993b) and McDowell (in Bell and Valentine 1995) chart how some lesbians tenuously negotiate their identities in the workplace, Rothenberg (in Bell and Valentine 1995) discusses the informal networks lesbians use to create community and safety, and Johnston (1998) speaks of the importance of gyms as safe public spaces for alternative female embodiments. Valentine also discusses discriminations faced by lesbians everyday (1993a) and the considerable diversity within and among lesbian communities. (in Jones et al. 1997). Finally, Chouinard and Grant (1995) argue against marginalization of disabled and lesbian women in the academy, comparing their exclusions as disabled and lesbian women, respectively.

In contrast, the urban market emphasis present in much gay male research is partly evident in titles such as, '*The Broadway Corridor: Gay Businesses as Agents of Revitalization in Long Beach, California*' (Ketteringham 1983 cited in Elder et al. in press). The title secondarily speaks to a phenomenon present in much research, particularly about gay male spaces, namely, the exclusionary use of the word '*gay*'. While a study may pertain to gay men only, the word 'gay,' like heteropatriachal designations of 'man,' is commonly deployed as though it speaks for all queer communities. It thereby obliterates the specificities and identities of those not gay-male-identified (Chouinard and Grant (1995). Perhaps the most prominent and prolific scholar pioneering research into urban markets and queer life is Larry Knopp. Though his work centers on gay male gentrification, he has consistently attempted to theorize gendered spatial differences in queer opportunity structures and life (e.g., Knopp 1990, Lauria and Knopp 1985).

Gay men's desires and differential abilities to procure spatial security at scales and densities greater than that obtained by lesbians (for example, Boystown in Chicago, the Castro district in San Francisco, the West End in Vancouver, Mykonos and Korfu in Greece, and cruising areas for western gay men in Bangkok) have been a source of scholarly debate.

Castells (1983) argued early on, for example, that urban-based gay men form distinct neighborhoods because, as men, they are inherently more territorial. Adler and Brennan (1992) disagree, contending that lesbians, as women, are relatively disadvantaged economically and more prone to violence and attack. Consequently, they are less able to secure territory, legitimacy, visibility (see also Knopp 1990).

More recent analyses have taken a comparative or cross-cultural look at queer life, entertaining both greater analytical specificity and diversity along lines of gender, 'race,' disability, and/or national or rural–urban location (e.g., Brown 2000, Bell and Valentine 1995, Chouinard and Grant 1995). There is also nascent research on state regulation of sexuality, national identity, and citizenship (e.g., Binnie 1997b, Nast 1998) and on intersections of sexuality and the academy (JGHE 1999).

Despite the increased presence of queer research and researchers in the discipline, sexuality research, especially that which deconstructs heterosexuality or which makes queer sexuality visible, is still considered radical (Chouinard and Grant 1995, Valentine 1998, JGHE 1999). Negative reactions point to anxieties over research that expose heterosexuality's artifice, research that implicitly argues against the privileges and privileging power relations upon which heterosexual ontologies of truth and gender are grounded. The depth of fear is made poignantly clear in a recent article by Valentine (1998) about the ongoing homophobic harassment levied at her by someone in the discipline who collapses her sexuality (she is a lesbian) with her sexuality research on lesbians, explicitly naming both a moral abomination.

4. *Future Research*

A number of areas for future research suggest themselves. First, theorization is needed in addressing sexuality–geography links. What kinds of spatial and symbolic work do normative and queer sexualities practically accomplish across historical and cultural place? More work is needed that explores how normative heterosexualities vary across time and place, in keeping with diverse sociocultural expressions, uses and values of kinship and family structures. Moreover, the nuclear family is apparently in decline unevenly across place and time, in keeping with shifts in places of industrialization, begging the question as to whether some sexualities and affective familial patterns are better suited to certain political economies than to others (see also Knopp 1994). If this is the case, what kinds of sexual identities and desires work well in postindustrial places, and how are industrial restructuring processes affecting oedipal family life and the commoditization of normative and oppositionally sexed identities? What sorts of alternative sociospatial alliances are needed to allow diverse sexualities to coexist? Are gay white men between the ages of 25 and 40 ideally situated economically and culturally to take advantage of the decline of the nuclear family in postindustrial contexts, which do not depend as much upon, or value, procreation?

Second, sexuality–geography research needs to be more theoretically and empirically attuned to differences produced through normative and oppositional constructions of race, gender, class, disability, age, religious beliefs, and nation. Little gender research has been done in the context of queer communities, for example. Are sociospatial interactions between gay male and lesbian communities devoid of gendered tensions? Does patriarchy disappear in spaces of gay male communities (Chouinard and Grant 1995, Knopp 1990)? Are lesbians or gay men situated outside the signifying strictures of heteronormative codes of femininity and masculinity? And how are constructions of race expressed sexually and geographically? Much work needs to be done on how persons of color have been put into infantilized, subordinate positions and places associated with the white-led Family. African-American men after emancipation, for example, were consistently constructed as rapist-sons sexually desirous of the white mother. What kind of work do these and other similarly sexed constructions accomplish in colonial and neocolonial contexts (Nast 2000)? Similarly, why does sex with young boys and men of color, particularly in Southeast Asia, have such wide market appeal among privileged gay white men, resulting in considerable sexualized tourism investment? And what kind of cultural, political, and economic work is achieved through sado-masochistic constructions threaded across sexed and raced communities? What precisely is being constructed and why? Finally, how do sociospatial debates about, and constructions of sexuality, continue to shape the theories, practices, and empirical concerns of geography?

See also: Gender and Environment; Gender and Place; Queer Theory; Sexuality and Gender

Bibliography

Adler S, Brennan J 1992 Gender and space: Lesbians and gay men in the city. *International Journal of Urban and Rural Research* **16**: 24–34

Bell D J 1991 Insignificant others: Lesbian and gay geographies. *Area* **23**: 323–9

Bell D, Valentine G (eds.) 1995 *Mapping Desire*. Routledge, New York

Binnie J 1997a Coming out of geography: Towards a queer epistemology? *Environment and Planning D: Society and Space* **15**: 223–37

Binnie J 1997b Invisible Europeans: Sexual citizenship in the New Europe. *Environment and Planning A* **29**: 237–48

Brown M P 2000 *Closet Geographies*. Routledge, New York

Butler J P 1990 *Gender Trouble* Routledge, New York

Castells M 1983 *The City and the Grassroots*. University of California Press, Berkeley, CA
Chouinard V, Grant A 1995 On being not even anywhere near "the project": Ways of putting ourselves in the picture. *Antipode* **27**: 137–66
Elder G, Knopp L, Nast H in press Geography and sexuality. In: Gaile G, Willmott C (eds.) *Geography in America at the Dawn of the 21st Century*. Oxford University Press, Oxford, UK
Foord J, Gregson N 1986 Patriarchy: towards a reconceptualisation. *Antipode* **18**: 186–211
Gibson-Graham J K 1998 Queerying globalization. In: Nast H, Pile S (eds.) *Places Through the Body*. Routledge, New York
Johnston L 1998 Reading the sexed bodies and spaces of gyms. In: Nast H J, Pile S (eds.) *Places Through the Body*. Routledge, New York
Jones J P III, Nast H J, Roberts S H (eds.) 1997 *Thresholds in Feminist Geography*. Rowman and Littlefield, Lanham, MD
JGHE Symposium: Teaching Sexualities in Geography 1999 *The Journal of Geography in Higher Education* **23**: 77–123
Knopp L 1990 Some theoretical implications of gay involvement in an urban land market. *Political Geography Quarterly* **9**: 337–52
Knopp L 1994 Social justice, *sexuality*, and the city. *Urban Geography* **15**: 644–60
Lauria M, Knopp L 1985 Towards an analysis of the role of gay communities in the urban renaissance. *Urban Geography* **6**: 152–69
Nast H 1998 Unsexy geographies. *Gender, Place and Culture* **5**: 191–206
Nast H J 2000 Mapping the 'unconscious': Racism and the oedipal family. *Annals of the Association of American Geographers* **90**(2): 215–55
Rich A 1980 Compulsory heterosexuality and lesbian existence. *Signs* **5**: 631–60
Rose G 1993 *Feminism and Geography*. University of Minnesota Press, Minneapolis, MN
Valentine G 1993a (Hetero)sexing space: Lesbian perceptions and experiences of everyday spaces. *Environment and Planning D: Society and Space* **11**: 395–413
Valentine G 1993b Negotiating and managing multiple sexual identities: Lesbian time-space strategies. *Transactions of the Institute of British Geographer* **18**: 237–48
Valentine G 1998 "Sticks and Stones May Break my Bones": A personal geography of harassment. *Antipode* **30**: 305–32
Weightman B 1980 Gay bars as private places. *Landscape* **24**: 9–17
Women and Geography Study Group (WGSG) 1997 *Feminist Geographies: Explorations in Diversity and Difference*. Longman, Harlow, UK

H. J. Nast

Sexually Transmitted Diseases: Psychosocial Aspects

Sexually transmitted diseases (STDs) including HIV place an enormous burden on the public's health. In 1995, the World Health Organization (WHO) estimated that, worldwide, there were 333 million new cases of chlamydia, gonorrhea, syphilis, and trichomoniasis in 15- to 49-year-olds. In 1999, WHO estimated that there were 5.6 million new HIV infections worldwide, and that 33.6 million people were living with AIDS. Although it is not who you are, but what you do that determines whether you will expose yourself or others to STDs including HIV, STDs disproportionately affect various demographic groupings. For example, STDs are more prevalent in urban settings, in unmarried individuals, and in young adults. They are also more prevalent among the socioeconomically disadvantaged and various ethnic groups. For example, in 1996 in the United States, the incidence of reported gonorrhea per 100,000 was 826 among black non-Hispanics, 106 among Native Americans, 69 among Hispanics, 26 among white non-Hispanics, and 18.6 among Asian/Pacific Islanders (Aral and Holmes 1999).

Demographic differences in STD rates are most likely explained by differences in sexual behaviors or disease prevalence. Thus, demographic variables are often referred to as risk markers or risk indicators. In contrast, sexual and healthcare behaviors that directly influence the probability of acquiring or transmitting STDs represent true risk factors. From a psychosocial perspective, it is these behavioral risk factors (and not risk markers) that are critical for an understanding of disease transmission (see *Sexual Risk Behaviors*).

1. Transmission Dynamics

To understand how behavior contributes to the spread of an STD, consider May and Anderson's (1987) model of transmission dynamics: $Ro = \beta cD$, where, Ro = Reproductive Rate of Infection. When Ro is greater than 1, the epidemic is growing; when Ro is less than 1, the epidemic is dying out; and when $Ro = 1$, the epidemic is in a state of equilibrium, β = measure of infectivity or tranmissability, c = measure of interaction rates between susceptibles and infected individuals, and D = measure of duration of infectiousness

Each of the parameters on the right hand side of the equation can be influenced by behavior. For example, the transmission rate (β) can be lowered by increasing consistent and correct condom use or by delaying the onset of sexual activity. Transmissibility can also be reduced by vaccines, but people must utilize these vaccines and before that, it is necessary for people to have participated in vaccine trials. Decreasing the rate of new partner acquisition will reduce the sexual interaction rate c and, at least for bacterial STDs, duration of infectiousness D can be reduced through the detection of asymptomatic or through early treatment of symptomatic STDs. Thus, increasing care seeking behavior and/or increasing the likelihood that one will participate in screening programs can affect the reproductive rate. Given that STDs are

important cofactors in the transmission of HIV, early detection and treatment of STDs will also influence HIV transmissibility (β). Finally, D can also be influenced by patient compliance with medical treatment, as well as compliance with partner notification.

Clearly, there are a number of different behaviors that, if changed or reinforced, could have an impact on the reproductive rate R_o of HIV and other STDs. A critical question is whether it is necessary to consider each of these behaviors as a unique entity, or whether there are some more general principals that can guide our understanding of any behavior. Fortunately, even though every behavior is unique, there are only a limited number of variables that need to be considered in attempting to predict, understand, or change any given behavior.

2. *Psychosocial Determinants of Intention and Behavior*

Figure 1 provides an integration of several different leading theories of behavioral prediction and behavior change (cf. Fishbein et al. 1992). Before describing this model, however, it is worth noting that theoretical models such as the one presented in Fig. 1 have often been criticized as 'Western' or 'US' models that don't apply to other cultures or countries. When properly applied, however, these models are culturally specific, and they require one to understand the behavior from the perspective of the population being considered. Each of the variables in the model can be found in almost any culture or population. In fact, the theoretical variables contained in the model have been assessed in over 50 countries in both the developed and the developing world. Moreover, the relative importance of each of the variables in the model is expected to vary as a function of both the behavior and the population under consideration (see *Health Behavior: Psychosocial Theories*).

2.1 *Determinants of Behavior*

Looking at Fig. 1, it can be seen that any given behavior is most likely to occur if one has a strong intention to perform the behavior, if one has the necessary skills and abilities required to perform the behavior, and if there are no environmental constraints preventing behavioral performance. Indeed, if one has made a strong commitment (or formed a strong intention), has the necessary skills and abilities, and if there are no environmental constraints to prevent behavioral performance, the probability is close to one that the behavior will be performed (Fishbein et al. 1992).

Clearly, very different types of interventions will be necessary if one has formed an intention but is unable to act upon it, than if one has little or no intention to perform the behavior in question. Thus, in some populations or cultures, the behavior may not be performed because people have not yet formed intentions to perform the behavior, while in others, the problem may be a lack of skills and/or the presence of

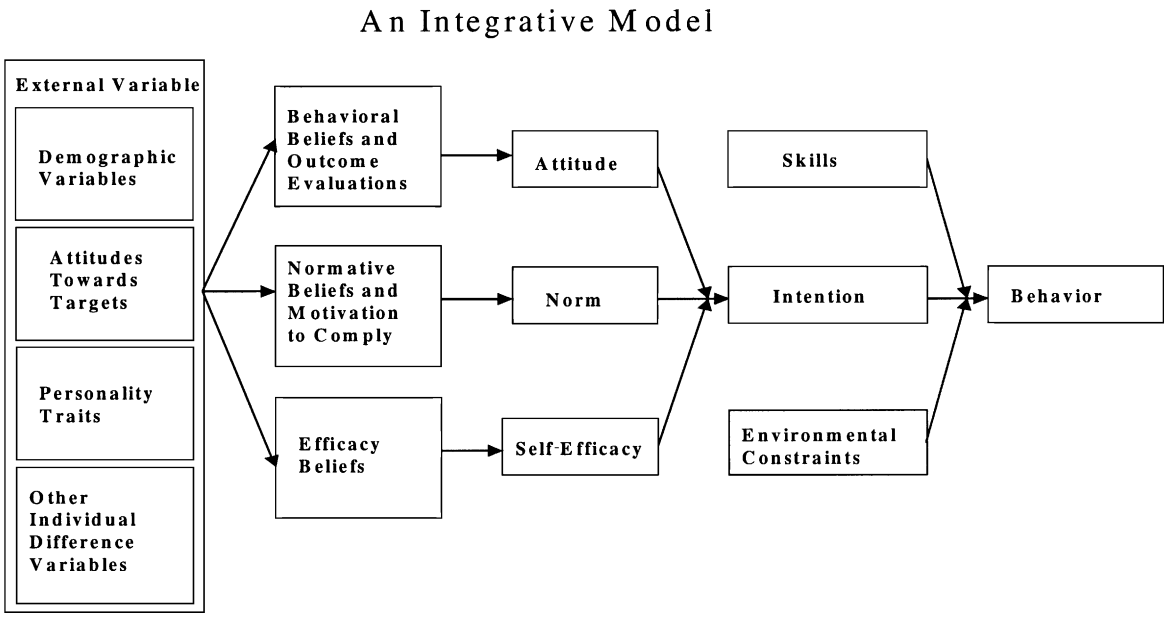

Figure 1
An integrative model

environmental constraints. In still other cultures, more than one of these factors may be relevant. For example, among female commercial sex workers (CSWs) in Seattle, Washington, only 30 percent intend to use condoms for vaginal sex with their main partners, and of those, only 40 percent have acted on their intentions (von Haeften et al. 2000). Clearly if people have formed the desired intention but are not acting on it, a successful intervention will be directed either at skills building or will involve social engineering to remove (or to help people overcome) environmental constraints.

2.2 Determinants of Intentions

On the other hand, if strong intentions to perform the behavior in question have not been formed, the model suggests that there are three primary determinants of intention: the attitude toward performing the behavior (i.e., the person's overall feelings of favorableness or unfavorableness toward performing the behavior), perceived norms concerning performance of the behavior (including both perceptions of what others think one should do as well as perceptions of what others are doing), and one's self-efficacy with respect to performing the behavior (i.e., one's belief that one can perform the behavior even under a number of difficult circumstances) (see *Self-efficacy and Health*). As indicated above, the relative importance of these three psychosocial variables as determinants of intention will depend upon both the behavior and the population being considered. Thus, for example, one behavior may be determined primarily by attitudinal considerations while another may be influenced primarily by feelings of self-efficacy. Similarly, a behavior that is driven attitudinally in one population or culture may be driven normatively in another. Thus, before developing interventions to change intentions, it is important first to determine the degree to which that intention is under attitudinal, normative, or self-efficacy control in the population in question. Once again, it should be clear that very different interventions are needed for attitudinally controlled behaviors than for behaviors that are under normative influence or are related strongly to feelings of self-efficacy. Clearly, one size does not fit all, and interventions that are successful in one culture or population may be a complete failure in another.

2.3 Determinants of Attitudes, Norms, and Self-efficacy

The model in Fig. 1 also recognizes that attitudes, perceived norms, and self-efficacy are all, themselves, functions of underlying beliefs—about the outcomes of performing the behavior in question, about the normative proscriptions and/or behaviors of specific referents, and about specific barriers to behavioral performance. Thus, for example, the more one believes that performing the behavior in question will lead to 'good' outcomes and prevent 'bad' outcomes, the more favorable one's attitude toward performing the behavior. Similarly, the more one believes that specific others think one should (or should not) perform the behavior in question, and the more one is motivated to comply with those specific others, the more social pressure one will feel (or the stronger the norm) with respect to performing (or not performing) the behavior. Finally, the more one perceives that one can (i.e., has the necessary skills and abilities to) perform the behavior, even in the face of specific barriers or obstacles, the stronger will be one's self-efficacy with respect to performing the behavior.

It is at this level that the substantive uniqueness of each behavior comes into play. For example, the barriers to using and/or the outcomes (or consequences) of using a condom for vaginal sex with one's spouse or main partner may be very different from those associated with using a condom for vaginal sex with a commercial sex worker or an occasional partner. Yet it is these specific beliefs that must be addressed in an intervention if one wishes to change intentions and behavior. Although an investigator can sit in their office and develop measures of attitudes, perceived norms, and self-efficacy, they cannot tell you what a given population (or a given person) believes about performing a given behavior. Thus, one must go to members of that population to identify salient behavioral, normative, and efficacy beliefs. One must understand the behavior from the perspective of the population one is considering.

2.4 The Role of 'External' Variables

Finally, Fig. 1 also shows the role played by more traditional demographic, personality, attitudinal, and other individual difference variables (such as perceived risk). According to the model, these types of variables primarily play an indirect role in influencing behavior. For example, while men and women may hold different beliefs about performing some behaviors, they may hold very similar beliefs with respect to others. Similarly rich and poor, old and young, those from developing and developed countries, those who do and do not perceive they are at risk for a given illness, those with favorable and unfavorable attitudes toward family planning, etc., may hold different attitudinal, normative or self-efficacy beliefs with respect to one behavior but may hold similar beliefs with respect to another. Thus, there is no necessary relation between these 'external' or 'background' variables and any given behavior. Nevertheless, external variables such as cultural and personality differences should be reflected in the underlying belief structure.

3. The Role of Theory in HIV/STD Prevention

Models, like the one represented by Fig. 1, have served as the theoretical underpinning for a number of successful behavioral interventions in the HIV/STD arena (cf. Kalichman et al. 1996). For example, the US Centers for Disease Control and Prevention (CDC) have supported two large multisite studies based on the model in Fig. 1. The first, the AIDS Community Demonstration Projects (Fishbein et al 1999), attempted to reach members of populations at risk for HIV/STD that were unlikely to come into contact with the health department. The second, Project RESPECT, was a multisite randomized controlled trial designed to evaluate the effectiveness of HIV/STD counseling and testing (Kamb et al. 1998). Project RESPECT asked whether prevention counseling or enhanced prevention counseling were more effective in increasing condom use and reducing incident STDs than standard education.

Although based on the same theoretical model, these two interventions were logistically very different. In one, the intervention was delivered 'in the street' by volunteer networks recruited from the community. In the other, the intervention was delivered one-on-one by trained counselors in an STD clinic. Thus, one involved community participation and mobilization while the other involved working within established public health settings. In addition, one was evaluated using the community as the unit of analysis while the other looked for behavior change at the individual level.

Despite these logistic differences, both interventions produced highly significant behavioral change. In addition, in the clinic setting (where it was feasible to obtain biologic outcome measures), the intervention also produced a significant reduction in incident STDs. The success of these two interventions is due largely to their reliance on established behavioral principles. More important, it appears that theory-based approaches that are tailored to specific populations and behaviors can be effective in changing STD/HIV related behaviors in different cultures and communities (cf. NIH 1997).

4. The Relation between Behavioral and Biological Outcome Measures

Unfortunately, behavior change (and in particular, self-reported behavior change), is often viewed as insufficient evidence for disease prevention, and several investigators have questioned the validity of self-reports of behavior and their utility as outcome measures in HIV/STD Prevention research. For example, Zenilman et al. (1995) investigated the relationship between STD clinic patients self-reported condom use and STD incidence. Patients coming to an STD clinic in Baltimore, Maryland, who agreed to participate in the study were examined for STDs upon entry and approximately 3 months later. At the time of the follow-up exam, participants were asked to report the number of times they had sex in the past month and the number of times they had used a condom while having sex. Those who reported 100 percent condom use were compared with those who reported sometime use or no use of condoms with respect to incident (or new) STDs.

Zenilman et al. (1995) found no significant relationship between self-reported condom use and incident STDs. Based on this finding, they questioned the veracity of the self-reports and suggested that intervention studies using self-reported condom use as the primary outcome measure were at best suspect and at worst, invalid. Such a view fails to recognize the complex relationship between behavioral and biological measures.

4.1 Transmission Dynamic Revisited

Consider again the May and Anderson (1987) model of transmission dynamics: $Ro = \beta cD$. It is important to recognize that the impact on the reproductive rate, of a change in any one parameter, will depend upon the values of the other two parameters. Thus, for example, if one attempted to lower the reproductive rate of HIV by reducing transmission efficiency (either by reducing STDs or by increasing condom use), the impact of such a reduction would depend upon both the prevalence of the disease in the population and the sexual mixing patterns in that population. Clearly, if there is no disease in the population, decreases (or increases) in transmission efficiency can have very little to do with the spread of the disease. Similarly, a reduction in STD rates or an increase in condom use in those who are at low risk of exposure to partners with HIV will have little or no impact on the epidemic. In contrast, a reduction in STDs or an increase in condom use in those members of the population who are most likely to transmit and/or acquire HIV (that is, in the so-called core group), can, depending upon prevalence of the disease in the population, have a big impact on the epidemic (cf. Pequegnat et al. 2000).

To complicate matters further, it must also be recognized that changes in one parameter may directly or indirectly influence one of the other parameters. For example, at least some people have argued that an intervention program that increased condom use successfully, could also lead to an increase in number of partners (perhaps because now one felt somewhat safer). If this were in fact the case, an increase in condom use would not lead necessarily to a decrease in the reproductive rate. In other words, the impact of a given increase (or decrease) in condom use on STD/HIV incidence will differ, depending upon the values of the other parameters in the model. In addition, it's important to recognize that condom use behaviors are very different with 'safe' than with 'risky' partners. For example, condoms are used much more frequently

with 'occasional' or 'new' partners than with 'main' or 'regular' partners (see e.g., Fishbein et al 1999).

Thus, one should not expect to find a simple linear relation (i.e., a correlation) between decreases in transmission efficiency and reductions in HIV seroconversions. Moreover, it should be recognized that many other factors may influence transmission efficiency. For example, the degree of infectivity of the donor; characteristics of the host; and the type and frequency of sexual practices all influence transmission efficiency; and variations in these factors will also influence the nature of a relationship between increased condom use and the incidence of STDs (including HIV).

In addition, although correct and consistent condom use can prevent HIV, gonorrhea, syphilis, and probably chlamydia, condoms are less effective in interrupting transmission of herpes and genital warts. Although one is always better off using a condom than not using a condom, the impact of condom use is expected to vary by disease. Moreover, for many STDs, transmission from men to women is much more efficient than from women to men. For example, with one unprotected coital episode with a person with gonorrhea, there is about a 50 to 90 percent chance of transmission from male to female, but only about a 20 percent chance of transmission from female to male.

It should also be noted that one can acquire an STD even if one always uses a condom. Consistent condom use is not necessarily correct condom use, and incorrect condom use and condom use errors occur with surprisingly high frequencies (Fishbein and Pequegnat 2000, Warner et al. 1998). In addition, at least some 'new' or incident infections may be 'old' STDs that initially went undetected or that did not respond to treatment. Despite these complexities, it is important to understand when, and under what circumstances, behavior change will be related to a reduction in STD incidence. Unfortunately, it is unlikely that this will occur until behaviors are assessed more precisely and new (or incident) STDs can be more accurately identified. From a behavioral science or psychosocial perspective, the two most pressing problems and the greatest challenges are to assess correct, as well as consistent, condom use, and to identify those at high or low risk for transmitting or acquiring STDs (see *HIV Risk Interventions*).

4.2 Assessing Correct and Consistent Condom Use

Condom use is most often assessed by asking respondents how many times they have engaged in sex and then asking them to indicate how many of these times they used a condom. One will get very different answers to these questions depending upon the time frame used (lifetime, past year, past 3 months, past month, past week, last time), and the extent to which one distinguishes between the type of sex (vaginal, anal, or oral) and type of partner (steady, occasional, paying client). Irrespective of the time frame, type of partner or type of sex, these two numbers (i.e., number of sex acts and number of times condoms were used) can be used in at least two very different ways. In most of the literature (particularly the social psychological literature), the most common outcome measure is the percent of times the respondent reports condom use (i.e., for each subject, one divides the number of times condoms were used by the number of sex acts).

Perhaps a more appropriate measure would be to subtract the number of time condoms were used from the number of sex acts. This would yield a measure of the number of unprotected sex acts.

Clearly, if one is truly interested in preventing disease or pregnancy, it is the number of unprotected sex acts and not the percent of times condoms are used that should be the critical variable. Obviously, there is a difference in one's risk of acquiring an STD if one has sex 1,000 times and uses a condom 900 times than if one has sex 10 times and uses a condom 9 times. Both of these people will have used a condom 90 percent of the time, but the former will have engaged in 100 unprotected sex acts while the latter will have engaged in only 1.

By considering the number of unprotected sex acts rather than the percent of times condoms are used, it becomes clear how someone who reports always using a condom can get an STD or become pregnant. To put it simply, consistent condom use is not necessarily correct condom use, and incorrect condom use almost always equates to unprotected sex.

4.3 How Often Does Incorrect Condom Use Occur?

There is a great deal of incorrect condom use. For example, Warner et al. (1998) asked 47 sexually active male college students (between 18 and 29 years of age) to report the number of times they had vaginal intercourse and the number of times they used condoms in the last month. In addition the subjects were asked to quantify the number of times they experienced several problems (e.g., breakage, slippage) while using condoms. Altogether, the 47 men used a total of 270 condoms in the month preceding the study. Seventeen percent of the men reported that intercourse was started without a condom, but they then stopped to put one on; 12.8 percent reported breaking a condom during intercourse or withdrawal; 8.5 percent started intercourse with a condom, but then removed it and continued intercourse; and 6.4 percent reported that the condom fell off during intercourse or withdrawal. Note that all of these people could have honestly reported always using a condom, and all could have transmitted and/or acquired an STD!

Similar findings were obtained from both men and women attending STD clinics in the US (Fishbein and Pequegnat 2000). For example, fully 34 percent of

women and 36 percent of men reported condom breakage during the past 12 months, with 11 percent of women and 15 percent of men reporting condom breakage in the past 3 months. Similarly, 31 percent of the women and 28 percent of the men reported that a condom fell off in the past 12 months while 8 percent of both men and women report slippage in the past 3 months. Perhaps not surprisingly women are significantly more likely than men to report condom leakage (17 percent vs. 9 percent). But again, the remarkable findings are the high proportion of both men and women reporting different types of condom mistakes. For example, 31 percent of the men and 36 percent of the women report starting sex without a condom and then putting one on, while 26 percent of the men and 23 percent of the women report starting sex with a condom and then taking it off and continuing intercourse. This probably reflects the difference between using a condom for family planning purposes and using one for the prevention of STDs including HIV.

The practice of having some sex without a condom probably reflects incorrect beliefs about how one prevents pregnancy. Irrespective of the reason for these behaviors, the fact remains that all of these 'errors' could have led to the transmission and/or acquisition of an STD despite the fact that condoms had, in fact, been used.

In general, and within the constraints of one's ability to accurately recall past events, people do appear to be honest in reporting their sexual behaviors including their condom use. Behavioral scientists must obtain better measures, not of condom use per se, but of correct and consistent condom use, or perhaps even more important, of the number of unprotected sex acts in which a person engages.

4.4 Sex With 'Safe' and 'Risky' Partners

Whether one uses a condom correctly or not is essentially irrelevant as long as one is having sex with a safe (i.e., uninfected) partner. However, to prevent the acquisition (or transmission) of disease, correct and consistent condom use is essential when one has sex with a risky (i.e., infected) partner. Although it is not possible to always know whether another person is safe or risky, it seems reasonable to assume that those who have sex with either a new partner, an occasional partner, or a main partner who is having sex outside of the relationship are at higher risk than those who have sex only with a main partner who is believed to be monogamous. Consistent with this, STD clinic patients with potential high risk partners (i.e., new, occasional, or nonmonogamous main partners) were significantly more likely to acquire an STD than those with potential low risk partners (18.5 percent vs. 10.4 percent; Fishbein and Jarvis 2000).

One can also assess people's perceptions that their partners put them at risk for acquiring HIV. For example, the STD clinic patients were also asked to indicate, on seven-place likely (7)/unlikely (1) scales, whether having unprotected sex with their main and/or occasional partners would increase their chances of acquiring HIV. Not surprisingly those who felt it was likely that unprotected sex would increase their chances of acquiring HIV (i.e., those with scores of 5, 6, or 7) were, in fact, significantly more likely to acquire a new STD (19.4 percent) than those who perceived that their partner(s) did not put them at risk (i.e., those with scores of 1, 2, 3, or 4—STD incidence = 12.6 percent).

Somewhat surprising, clients' perceptions of the risk status of their partners were not highly correlated with whether they were actually with a potentially safe or dangerous partner ($r = .11$, $p < .001$). Nevertheless, both risk scores independently relate to STD acquisition. More specifically, those who were 'hi' on both actual and perceived risk were almost four times as likely to acquire a new STD than were those who are 'lo' on both actual and perceived risk. Those with mixed patterns (i.e., hi/lo or lo/hi) are intermediate in STD acquisition (Fishbein and Jarvis 2000). So, it does appear possible to not only find behavioral measures that distinguish between those who are more or less likely to acquire a new STD, but equally important, it appears possible to identify measures that distinguish between those who are, or are not, having sex with partners who are potentially placing them at high risk for acquiring an STD.

4.5 STD Incidence as a Function of Risk and Correct and Consistent Condom Use

If this combined measure of actual and perceived risk is 'accurate' in identifying whether or not a person is having sex with a risky (i.e., an infected) or a safe (i.e., a noninfected) partner, condom use should make no difference with low-risk partners, but correct and consistent condom use should make a major difference with high-risk partners. More specifically, correct and consistent condom use should significantly reduce STD incidence among those having sex with potential high-risk partners. Consistent with this, while condom use was unrelated to STD incidence among those with low-risk partners, correct and consistent condom use did significantly reduce STD incidence among those at high risk. Not only are these findings important for understanding the relationship between condom use and STD incidence but they provide evidence for both the validity and utility of self-reported behaviors. In addition, these data provide clear evidence that if behavioral interventions significantly increase correct and consistent condom use among people who perceive they are at risk and/or who have a new, an occasional, or a nonmonogamous main partner, there will be a significant reduction in STD incidence. On

the other hand, if, among this high-risk group, we only increase consistency of use without increasing correct use and/or if we only increase condom use among those at low risk, then we will see little or no reduction in STD (or HIV) incidence.

5. *Future Directions*

Clearly it is time to stop using STD incidence as a 'gold standard' to validate behavioral self-reports and to start paying more attention to understanding the relationships between behavioral and biological outcome measures. It is also important to continue to develop theory-based, culturally sensitive interventions to change a number of STD/HIV related behaviors. While much of the focus to date has been on increasing consistent condom use, interventions are needed to increase correct condom use and to increase the likelihood that people will come in for screening and early treatment, and, for those already infected, to adhere to their medical regimens. Indeed, it does appear that we now know how to change behavior, and that under appropriate conditions, behavior change will lead to a reduction in STD incidence. While many investigators have called for 'new' theories of behavior change, 'new' theories are probably unnecessary. What is needed, however, is for investigators and interventionists to better understand and correctly utilize existing, empirically supported theories in developing and evaluating behavior change interventions.

See also: Health Behavior: Psychosocial Theories; HIV Risk Interventions; Self-efficacy and Health; Sexual Risk Behaviors; Vulnerability and Perceived Susceptibility, Psychology of

Bibliography

Aral S O, Holmes K K 1999 Social and behavioral determinants of the epidemiology of STDs: Industrialized and developing countries. In: Holmes K K, Sparling P F, Mardh P-A, Lemon S M, Stamm W E, Piot P, Wasserheit J N (eds.) *Sexually Transmitted Diseases*. McGraw-Hill, New York

Fishbein M, Higgins D L, Rietmeijer C 1999 Community-level HIV intervention in five cities: Final outcome data from the CDC AIDS community demonstration projects. *American Journal of Public Health* **89**(3): 336–45

Fishbein M, Bandura A, Triandis H C, Kanfer F H, Becker M H, Middlestadt S E 1992 *Factors Influencing Behavior and Behavior Change: Final Report—Theorist's Workshop*. National Institute of Mental Health, Rockville, MD

Fishbein M, Jarvis B 2000 Failure to find a behavioral surrogate for STD incidence: What does it really mean? *Sexually Transmitted Diseases* **27**: 452–5

Fishbein M, Pequegnat W 2000 Using behavioral and biological outcome measures for evaluating AIDS prevention interventions: A commentary. *Sexually Transmitted Diseases* **27**(2): 101–10

Kalichman S C, Carey M P, Johnson B T 1996 Prevention of sexually transmitted HIV infection: A meta-analytic review of the behavioral outcome literature. *Annals of Behavioral Medicine* **18**(1): 6–15

Kamb M L, Fishbein M, Douglas J M, Rhodes F, Rogers J, Bolan G, Zenilman J, Hoxworth T, Mallotte C K, Iatesta M, Kent C, Lentz A, Graziano S, Byers R H, Peterman T A, the Project RESPECT Study Group 1998 HIV/STD prevention counseling for high-risk behaviors: Results from a multicenter, randomized controlled trial. *Journal of the American Medical Association* **280**(13): 1161–7

May R M, Anderson R M 1987 Transmission dynamics of HIV infection. *Nature* **326**: 137–42

NIH Consensus Development Panel 1997 Statement from consensus development conference on interventions to prevent HIV risk behaviors. *NIH Consensus Statement* **15**(2): 1–41

Pequegnat W, Fishbein M, Celantano D, Ehrhardt A, Garnett G, Holtgrave D, Jaccard J, Schachter J, Zenilman J 2000 NIMH/APPC workgroup on behavioral and biological outcomes in HIV/STD prevention studies: A position statement. *Sexually Transmitted Diseases* **27**(3): 127–32

von Haeften I, Fishbein M, Kaspryzk D, Montano D 2000 Acting on one's intentions: Variations in condom use intentions and behaviors as a function of type of partner, gender, ethnicity and risk. *Psychology, Health & Medicine* **5**(2): 163–71

Warner L, Clay-Warner J, Boles J, Williamson J 1998 Assessing condom use practices: Implications for evaluating method and user effectiveness. *Sexually Transmitted Disases* **25**(6): 273–7

Zenilman J M, Weisman C S, Rompalo A M, Ellish N, Upchurch D M, Hook E W, Clinton D 1995 Condom use to prevent incident STDs: The validity of self-reported condom use. *Sexually Transmitted Diseases* **22**(1): 15–21

M. Fishbein

Shamanism

Shamanism is a tradition of part-time religious specialists who establish and maintain personalistic relations with specific spirit beings through the use of controlled and culturally scripted altered states of consciousness (ASC). Shamans employ powers derived from spirits to heal sickness, to guide the dead to their final destinations, to influence animals and forces of nature in a way that benefits their communities, to initiate assaults on enemies, and to protect their own communities from external aggression.

Shamans exercise mastery over ASC and use them as a means to the culturally approved end of mediating between human, animal, and supernatural worlds. Shamans draw upon background knowledge, conveyed through myth and ritual, which renders intelligible the potentially chaotic experience of ASC. The criterion of control helps to distinguish shamanism from the use of ASC in other traditions.

Shamanism has long been a subject of inquiry and controversy in diverse academic disciplines, with many hundreds of accounts of shamanic practices published

by the early 1900s. It has also been a topic of spiritual interest to the wider public in Europe and North America for the past several decades. Depending on the perspective taken, shamanism is either the most archaic and universal form of human spirituality or a culturally distinct religious complex with historical roots in Siberia and a path of diffusion into North and South America. The latter view is adopted in the following.

1. *Siberian Origin and Common Features*

The term 'shaman' is drawn from seventeenth century Russian sources reporting on an eastern Siberian people, the Evenk (Tungus). In the classic work on the subject by the historian of religion Mircea Eliade, 'shamanism' is used to refer to a complex of beliefs and practices diffused from northern Asia to societies in central and eastern Asia, as well as through all of North and South America. Eliade's comparative work specifies a core constellation of ideas common to shamanic traditions. The classic shamanic initiation involves the novice's selection—frequently unwanted and resisted—by a spirit(s), a traumatic and dangerous series of ordeals, followed by a death and rebirth that sometimes involves violent dismembering and subsequent reconstitution of the fledgling shaman's body. The archetypal shamanic cosmology is vertically tiered, with earth occupying the middle level and a cosmic tree or world mountain serving as a connecting path for shamans to travel to other cosmic planes (up or down) in pursuit of their 'helping spirits.'

Shamans use a variety of techniques to enter the ASC in which they communicate with their spirit helpers: sensory deprivation (e.g., fasting, meditation), repetitive drumming and/or dancing, and ingestion of substances with psychoactive properties (e.g., plant hallucinogens). The last of these means is particularly important in the shamanism of Central and South America, where an impressive array of plants and animal-derived chemicals have been used for their hallucinogenic properties (e.g., peyote, datura, virola, and poison from the *Bufo marinus* toad).

Among the most commonly reported characteristics of the shaman's ASC are transformations into animals, and flights to distant places. Shamanic animals include the jaguar (lowland South America), the wolf (North American Chukchee), and bears, reindeer and fish (Lapp). Shamanic flights are undertaken to recapture the wandering souls of the sick, to intercede on behalf of hunters with spirits that control animal species, and to guide the dead to their final destination.

The fact that shamans are often called upon to heal the sick should not lead to the conclusion that they are always concerned with the well being of their fellows. While the Hippocratic Oath of Western medicine prohibits physicians from doing harm, shamans frequently engage in actions meant to sicken, if not kill, their adversaries. This is an important corrective to romanticized images of the shamanic vocation.

2. *Popular and Professional 'Shamanism'*

Beginning in the 1960s there developed a convergence in the professional and popular interests in shamanism. Inspired by the experimentation with hallucinogenic drugs on the part of American and European youth, attention was drawn to the physiology and psychology of altered states of consciousness. In academic circles, this led to speculations about the neurochemical foundations of hallucinatory experiences and their potential therapeutic benefits. Investigations were undertaken into the imagery of shamanic states of consciousness to determine how much could be attributed to universal, physiologically related sensory experiences. One often cited collection of essays, published in the journal *Ethos* (Prince 1982), sought to link the shaman's ASC to the production of the naturally occurring opiates known as endorphins. Scholars argued that this might account for the healing effect of shamanic therapies.

In the popular cultures of North America and Europe, stress came to be placed on shamanism as an avenue to self-exploration, a means by which persons from any culture could advance their quest for spiritual understandings. The legacy of this development is to be found in the forms of 'neoshamanism' among New Age religions, which draw eclectically from more traditional shamanic practices. Native North and South American societies have been special sources of inspiration for practitioners of neoshamanism, but not always with the willing support of the indigenous people.

At the same time, and in response to the evolution of research in symbolic anthropology, anthropologists have contributed increasingly sophisticated ethnographic analyzes of the metaphysical underpinnings of shamanism in specific societies. These anthropologists concentrate their attention on the complexities of shamanic cosmologies and the relationships between shamanic symbolism and crucial features of the natural world (e.g., important animal and plant species, meteorological phenomena, and geographic landmarks).

An especially detailed analysis of a shamanic tradition comes from the Warao of the Orinoco Delta in Venezuela. The German-born and American-trained anthropologist Johannes Wilbert documents the mythic cosmology that supports the work of three distinct kinds of shamans, each of which are responsible for the care and feeding of specific gods. The highest ranking shaman, the Wishiratu ('Master of Pains' or Priest Shaman), responds when the Kanobos (Supreme Beings) send sickness to a Warao village. The Bahanarotu is also a curer, but has a special

responsibility for a fertility cult centered on an astonishingly complex supernatural location, the House of Tobacco Smoke, where four insect couples and a blind snake play a game of chance under the watchful eye of a swallow-tailed kite. The Hoarotu (Dark Shaman) has the unpleasant but essential task of feeding human flesh to the Scarlet Macaw, the God of the West. The preferred food of the other gods is tobacco smoke, which is also the vehicle by which Warao shamans enter ASC—by hyperventilating on long cigars made of a strong black leaf tobacco.

3. Research Directions

3.1 Dead Ends and Detours

There have been a number of dead ends and detours in shamanism studies. An unfortunate amount of energy was spent in fruitless discussions of the mental health of shamans—whether or not their abnormal behavior justified the application of psychiatric diagnoses (e.g., schizophrenia). Most scholars now recognize that the culturally patterned nature of shamans' behavior and the positive value placed on their social role make the application of mental illness labels inappropriate.

Another distraction came when cultural evolutionists hypothesized that shamanism was an intermediary stage in the developmental sequence of religious forms, in-between magic and institutionalized religion. Attempts to treat shamanism exclusively as an archaic expression of human religiosity flounder on the perseverance and even revitalization of shamanic traditions in cultures around the world. For example, the remarkable resiliency of shamanic practices is evident in the continuity between contemporary *curanderismo* (curing) on Peru's north coast and shamanic iconography dating to the pre-Hispanic Chavin culture (approx. 900–200 BC).

3.2 Positive Directions

A still limited amount of research has been directed toward the important question of how effective are shamanic treatments. This work has been troubled by serious theoretical issues (e.g., what constitutes 'efficacy') and thorny methodological questions (e.g., is the double blind, randomized trial possible under the unusual circumstances of shamanic rituals?). The previously cited endorphin theory has not been confirmed ethnographically. Another approach has adapted psychiatric questionnaires to before and after treatment interviews with shamanic patients.

A fruitful line of investigation has focused on the relationship between shamanism and culturally constructed notions of gender. Shamans sometimes 'bend' gender roles (e.g., transvestitism) and the sicknesses they treat can be entangled in gender-based conflicts. Some scholars suggest that where women dominate in the shaman role there is a less adversarial model of supernatural mediation than with male shamans. The best examples of gender-focused research come from East Asia (especially Korea), where the majority of shamans are women, and from the Chilean Mapuche, among whom female and cross-dressing male shamans have predominated since at least the sixteenth century.

A final body of research worth noting focuses on the capacity of shamanic traditions to survive and thrive even under the most disruptive social and political conditions. Agents of culture change, whether they are Soviet-era ideologues in Siberia or Christian missionaries in the Amazon, have been repeatedly frustrated in their attempts to obliterate shamanic practices. In recent decades, shamans have even become central figures in the politics of ethnicity and in antidevelopment protests. Scholars have shown that shamans accomplish this survival feat not by replicating ancient traditions, but by continuously reinventing them in the light of new realities and competing symbolic structures. While this strategy may occasion debates about what constitutes an 'authentic' shaman, it is nevertheless the key to shamanism's continuing success.

4. A Plea for Terminological Precision

The most serious threat to the academic study of shamanism lies in the broad application of the term to any religiously inspired trance form. To so dilute the concept as to make it applicable to Kung trance dancers, New Age spiritualists, and Mexican Huichol peyote pilgrims is to render it meaningless. The value of a scientific term is that it groups together phenomena that are alike in significant regards, while distinguishing those that are different. The promiscuous use of the term shaman will ultimately leave it as generic as 'spiritual healer,' and just as devoid of analytic value.

See also: Alternative and Complementary Healing Practices; Healing; Religion and Health; Spirit Possession, Anthropology of

Bibliography

Atkinson J M 1992 Shamanisms today. *Annual Review of Anthropology* **21**: 307–30

Bacigalupo A M *Shamans of the Cinnamon Tree, Priestesses of the Mood: Gender and healing among the Chilean Mapuche*. Manuscript in preparation

Brown M F 1989 Dark side of the shaman. *Natural History* **11** 8–10

Eliade M 1972 *Shamanism: Anchaic Techniques of Ecstasy*. Princeton University Press, Princeton, NJ

Furst P T (ed.) 1972 *Flesh of the Gods: The Ritual Use of Hallucinogens*. Praeger, New York

Hultkrantz A 1992 *Shamanic Healing and Ritual Drama*. Crossroad, New York

Joralemon D 1990 The selling of the shaman and the problem of informant legitimacy. *Journal of Anthropological Research* **46**: 105–18

Joralemon D, Sharon D 1993 *Sorcery and Shamanism: Curanderos and Clients in Northern Peru*. University of Utah, Salt Lake City, UT
Kalweit H 1988 *Dreamtime & Inner Space: The World of the Shaman*, 1st edn. Shambhala, Boston
Kendall L 1985 *Shamans, Housewives, and Other Restless Spirits: Women in Korean Ritual Life*. University of Hawaii Press, Honolulu, HI
Kleinman A, Sung L H 1979 Why do indigenous practitioners successfully heal? *Social Science & Medicine* **13(1B)**: 7–26
Peters L G, Price-Williams D 1980 Towards an experiential analysis of shamanism. *American Ethnologist* **7**: 398–418
Prince R 1982 Shamans and endorphins: Hypotheses for a synthesis. *Ethos* **10**(4): 409–23
Reichel-Dolmatoff G 1975 *The Shaman and the Jaguar*. Temple University Press, Philadelphia, PA
Shirokogoroff S M 1935 *The Psychomental Complex of the Tungus*. Kegan, Paul, Tranch and Trubner, London
Silverman J 1967 Shamanism and acute schizophrenia. *American Anthropologist* **69**: 21–31
Sullivan L E 1988 *Icanchu's Drum: An Orientation to Meaning in South American Religions*. Macmillan, New York
Wilbert J 1993 *Mystic Endowment: Religious Ethnography of the Warao Indians*. Harvard University Press, Cambridge, MA

D. Joralemon

Shame and the Social Bond

Many theorists have at least implied that emotions are a powerful force in social process. Although Weber didn't refer to emotions directly, his emphasis on values implies it, since values are emotionally charged beliefs. Especially in his later works, Durkheim proposed that collective sentiments created social solidarity through moral community. G. H. Mead proposed emotion as an important ingredient in his social psychology. For Parsons it is a component of social action in his AGIL scheme (Parsons and Shils 1955).

Marx implicated emotions in class tensions in the solidarity of rebelling classes. Durkheim proposed that '… what holds a society together—the 'glue' of solidarity—and [Marx implied that] what mobilizes conflict—the energy of mobilized groups—are emotions' (Collins 1990).

But even the theorists who dealt with emotions explicitly, Durkheim, Mead, and Parsons, did not develop concepts of emotion, investigate their occurrence, nor collect emotion data. Their discussions of emotion, therefore, have not borne fruit. The researchers whose work is reviewed here took the step of investigating a specific emotion.

1. *Seven Pioneers in the Study of Social Shame*

Five of the six sociologists reviewed acted independently of each other. In the case of Elias and Sennett, their discovery of shame seems forced upon them by their data. Neither Simmel nor Cooley define what they mean by shame. Goffman only partially defined embarrassment. The exception is Helen Lynd, who was self-conscious about shame as a concept. Lynd's book on shame was contemporaneous with Goffman's first writings on embarrassment and realized their main point: face-work meant avoiding embarrassment and shame.

Helen Lewis's empirical work on shame (1971) was strongly influenced by Lynd's book. She also was sophisticated in formulating a concept of shame, and in using systematic methods to study it. Sennett's work involved slight outside influence. He approvingly cited the Lynd book on shame in *The Hidden Injuries of Class* (1972), and his (1980) has a chapter on shame.

1.1 *Simmel: Shame and Fashion*

Shame plays a significant part in only one of Simmel's essays, on fashion (1904). People want variation and change, he argued, but they also anticipate shame if they stray from the behavior and appearance of others. Fashion is the solution to this problem, since one can change along with others, avoiding being isolated, and therefore shame (p. 553). Simmel's idea about fashion implies conformity in thought and behavior among one group in a society, the fashionable ones, and distance from another, those who do not follow fashion, relating shame to social bonds.

There is a quality to Simmel's treatment of shame that is somewhat difficult to describe, but needs description, since it characterizes most of the other sociological treatments reviewed here. Simmel's use of shame is casual and unselfconscious. His analysis of the shame component in fashion occurs in a single long paragraph. Shame is not mentioned before or after. He doesn't conceptualize shame or define it, seeming to assume that the reader will know the meaning of the term. Similar problems are prominent in Cooley, Elias, Sennett, and Goffman. Lynd and Lewis are exceptions, since they both attempted to define shame and locate it with respect to other emotions.

1.2 *Cooley: Shame and the Looking Glass Self*

Cooley (1922), like Simmel, was direct in naming shame. For Cooley, shame and pride both arose from self-monitoring, the process that was at the center of his social psychology. His concept of 'the looking glass self,' which implies the social nature of the self, refers directly and exclusively to pride and shame. But he made no attempt to define either emotion. Instead he used the vernacular words as if they were self-explanatory.

To give just one example of the ensuing confusion: in English and other European languages, the word pride used without qualification usually has an inflection of arrogance or hubris (pride goeth before the fall). In current usage, in order to refer to the kind

of pride implied in Cooley's analysis, the opposite of shame, one must add a qualifier like justified or genuine. Using undefined emotion words is confusing.

However, Cooley's analysis of self-monitoring suggests that pride and shame are the basic social emotions. His formulation of the social basis of shame in self-monitoring can be used to amend Mead's social psychology. Perhaps the combined Mead–Cooley formulation can solve the inside–outside problem that plagues psychoanalytic and other psychological approaches to shame, as I suggest below.

1.3 Elias: Shame in the Civilizing Process

Elias undertook an ambitious historical analysis of what he calls the 'civilizing process' (1994). He traced changes in the development of personality and social norms from the fifteenth century to the present. Like Weber, he gave prominence to the development of rationality. Unlike Weber, however, he gave equal prominence to emotional change, particularly to changes in the threshold of shame: 'No less characteristic of a civilizing process than "rationalization" is the peculiar molding of the drive economy that we call "shame" and "repugnance" or "embarrassment".'

Using excerpts from advice manuals, Elias outlined a theory of modernity. By examining advice concerning etiquette, especially table manners, body functions, sexuality, and anger, he suggests that a key aspect of modernity involved a veritable explosion of shame.

Elias showed that there was much less shame about manners and emotions in the early part of the period he studied than there was in the nineteenth century. In the eighteenth century, a change began occurring in advice on manners. What was said openly and directly earlier begins only to be hinted at, or left unsaid entirely. Moreover, justifications are offered less. One is mannerly because it is the right thing to do. Any decent person will be courteous; the intimation is that bad manners are not only wrong but also unspeakable, the beginning of repression.

The change that Elias documents is gradual but relentless; by a continuing succession of small decrements, etiquette books fall silent about the reliance of manners, style, and identity on respect, honor, and pride, and avoidance of shame and embarrassment. By the end of the eighteenth century, the social basis of decorum and decency had become virtually unspeakable. Unlike Freud or anyone else, Elias documents, step by step, the sequence of events that led to the repression of emotions in modern civilization.

By the nineteenth century, Elias proposed, manners are inculcated no longer by way of adult to adult verbal discourse, in which justifications are offered. Socialization shifts from slow and conscious changes by adults over centuries to swift and silent indoctrination of children in their earliest years. No justification is offered to most children; courtesy has become absolute. Moreover, any really decent person would not have to be told. In modern societies, socialization automatically inculcates and represses shame.

1.4 Richard Sennett: Is Shame the Hidden Injury of Class?

Although *The Hidden Injuries of Class* (1972) carries a powerful message, it is not easy to summarize. The narrative concerns quotes from interviews and the authors' brief interpretations. They do not devise a conceptual scheme and a systematic method. For this reason, readers are required to devise their own conceptual scheme, as I do here.

The book is based on participant observation in communities, schools, clubs, and bars, and 150 interviews with white working class males, mostly of Italian or Jewish background, in Boston for one year beginning in July of 1969 (p. 40–1).

The hidden injuries that Sennett and Cobb discovered might be paraphrased: their working class men felt that first, because of their class position, they were not accorded the respect that they should have received from others, particularly from their teachers, bosses and even from their own children. That is, these men have many complaints about their status. Secondly, these men also felt that their class position was at least partly their own fault. Sennett and Cobb imply that social class is responsible for both injuries. They believe that their working men did not get the respect they deserved because of their social class, and that the second injury, lack of self-respect, is also the fault of class, rather than the men's own fault, as most of them thought.

Sennett and Cobb argue that in US society, respect is largely based on individual achievement, the extent that one's accomplishments provide a unique identity that stands out from the mass of others. The role of public schools in the development of abilities forms a central part of Sennett and Cobb's argument. Their informants lacked self-respect, the authors thought, because the schooling of working class boys did not develop their individual talents in a way that would allow them to stand out from the mass as adults. In the language of emotions, they carry a burden of feelings of rejection and inadequacy, which is to say chronic low self-esteem (shame).

From their observations of schools, Sennett and Cobb argue that teachers single out for attention and praise only a small percentage of the students, usually those who are talented or closest to middle-class. This praise and attention allows the singled-out students to develop their potential for achievement. The large majority of the boys, however, are ignored and, in subtle ways, rejected.

There are a few working class boys who achieve their potential through academic or athletic talent. But the large mass does not. For them, rather than

opening up the world, public schools close it off. Education, rather than becoming a source of growth, provides only shame and rejection. For the majority of students, surviving school means running a gauntlet of shame. These students learn by the second or third grade that is better to be silent in class rather than risk humiliation of a wrong answer. Even students with the right answers must deal with having the wrong accent, clothing, or physical appearance. For most students, schooling is a vale of shame.

1.5 Helen Lynd: Shame and Identity

During her lifetime, Helen Lynd was a well-known sociologist. With her husband, Robert, she published the first US community studies, *Middletown* and *Middletown in Transition*. But Lynd was also profoundly interested in developing an interdisciplinary approach to social science. In her study *On Shame and the Search for Identity* (1961), she dealt with both the social and psychological sides of shame. She also clearly named the emotion of shame and its cognates, and located her study within previous scholarship, especially psychoanalytic studies. But Lynd also modified and extended the study of shame by developing a concept, and by integrating its social and psychological components.

In the first two chapters, Lynd introduced the concept of shame, using examples from literature to clarify each point. In the next section, she critiques mainstream approaches in psychology and the social sciences. She then introduces ideas from lesser known approaches, showing how they might resolve some of the difficulties. Finally, she has an extended discussion of the concept of identity, suggesting that it might serve to unify the study of persons by integrating the concepts of self, ego, and social role under the larger idea of identity.

Lynd's approach to shame is much more analytical and self-conscious than the other sociologists reviewed here. They treated shame as a vernacular word. For them, shame sprung out of their data, unavoidable. But Lynd encounters shame deliberately, as part of her exploration of identity.

Lynd explains that shame and its cognates get left out because they are deeply hidden, but at the same time pervasive. She makes this point in many ways, particularly in the way she carefully distinguishes shame from guilt.

One idea that Lynd develops is profoundly important for a social theory of shame and the bond, that sharing one's shame with another can strengthen the relationship: 'The very fact that shame is an isolating experience also means that … sharing and communicating it … can bring about particular closeness with other persons' (Lynd 1961, p. 66). In another place, Lynd went on to connect the process of risking the communication of shame with the kind of role-taking that Cooley and Mead had described: 'communicating shame can be an experience of … entering into the mind and feelings of another person' (p. 249).

Lynd's idea about the effects of communicating and not communicating shame was pivotal for Lewis's (1971) concepts of acknowledged and unacknowledged shame, and their relationship to the state of the social bond, as outlined below.

1.6 Goffman: Embarrassment and Shame in Everyday Life

Although shame goes largely unnamed in Goffman's early work, embarrassment and avoidance of embarrassment is the central thread. Goffman's *Everyperson* is always desperately worried about his image in the eyes of the other, trying to present herself with her best foot forward to avoid shame. This work elaborates, and indeed, fleshes out, Cooley's abstract idea of the way in which the looking glass self leads directly to pride or shame.

Interaction Ritual (1967) made two specific contributions to shame studies. In his study of face work, Goffman states what may be seen as a model of 'face' as the avoidance of embarrassment, and losing face as suffering embarrassment. This is an advance, because it offers readily observable markers for empirical studies of face. The importance of this idea is recognized, all too briefly, at the beginning of Brown and Levinson's (1987) study of politeness behavior.

Goffman's second contribution to the study of shame was made in a concise essay on the role of embarrassment in social interaction (1967). Unlike any of the other shame pioneers in sociology, he begins the essay with an attempt at definition. His definition is a definite advance, but it also foretells a limitation of the whole essay, since it is behavioral and physiological, ignoring inner experience. Framing his analysis in what he thought of as purely sociological mode, Goffman omitted feelings and thoughts. His solution to the inside–outside problem was to ignore most of inner experience, just as Freud ignored most of outside events.

However, Goffman affirms Cooley's point on the centrality of the emotions of shame and pride in normal, everyday social relationships, 'One assumes that embarrassment is a normal part of normal social life, the individual becoming uneasy not because he is personally maladjusted but rather because he is not … embarrassment is not an irrational impulse breaking through social prescribed behavior, but part of this orderly behavior itself' (1967, pp. 109, 111).

Even Goffman's partial definition of the state of embarrassment represents an advance. One of the most serious limitations of current contributions to the sociology of emotions is the lack of definitions of the emotions under discussion. Much like Cooley, Elias and Sennett, Kemper (1978) offers no definitions of emotions, assuming that they go without saying. Hochshild (1983) attempts to conceptualize various

emotions in an appendix, but doesn't go as far as to give concrete definitions of emotional states. Only in Retzinger (1991, 1995) can conceptual and operational definitions of the emotions of shame and anger be found.

2. Lewis's Discovery of Unacknowledged Shame

Helen Lewis's book on shame (1971) involved an analysis of verbatim transcripts of hundreds of psychotherapy sessions. She encountered shame because she used a systematic method for identifying emotions, the Gottschalk–Gleser method (Gottschalk et al. 1969, Gottschalk 1995), which involves use of long lists of keywords that are correlated with specific emotions.

Lewis found that anger, fear, grief, and anxiety cues showed up from time to time in some of the transcripts. She was surprised by the massive frequency of shame cues. Her most relevant findings:

(a) Prevalence: Lewis found a high frequency of shame markers in all the sessions, far outranking markers of all other emotions combined.

(b) Lack of awareness: Lewis noted that patient or therapist almost never referred to shame or its near cognates. Even the word embarrassment was seldom used. In analyzing the context in which shame markers occurred, Lewis identified a specific context: situations in which the patient seemed to feel distant from, rejected, criticized, or exposed by the therapist.

However, the patient's showed two different, seemingly opposite responses in the shame context. In one, the patient seemed to be suffering psychological pain, but failed to identify it as shame. Lewis called this form overt, undifferentiated shame. In a second kind of response, the patient seemed not to be in pain, revealing an emotional response only by rapid, obsessional speech on topics that seemed somewhat removed from the dialogue. Lewis called this second response bypassed shame.

3. Shame, Anger, and Conflict

In her transcripts, Lewis found many episodes of shame that extended over long periods of time. Since emotions are commonly understood to be brief signals (a few seconds) that alert us for action, the existence of long-lasting emotions is something of a puzzle. Lewis's solution to this puzzle may be of great interest in the social sciences, since it provides an emotional basis for longstanding hostility, withdrawal, or alienation.

She argued that her subjects often seemed to have emotional reactions to their emotions, and that this loop may extend indefinitely. She called these reactions 'feeling traps.' The trap that arose most frequently in her data involved shame and anger. A patient interprets an expression by the therapist as hostile, rejecting or critical, and responds with shame or embarrassment. However, the patient instantaneously masks the shame with anger, then is ashamed of being angry. Apparently, each emotion in the sequence is brief, but the loop can go on forever. This proposal suggests a new source of protracted conflict and alienation, one hinted at in Simmel's treatment of conflict.

Although Lewis didn't discuss other kinds of spirals, there is one that may be as important as the shame–anger loop. If one is ashamed of being ashamed, it is possible to enter into a shame–shame loop that leads to silence and withdrawal. Elias's work on modesty implies this kind of loop.

4. Shame and the Social Bond

Finally, Lewis interpreted her findings in explicitly social terms. She proposed that shame arises when there is a threat to the social bond, as was the case in all of the shame episodes she discovered in the transcripts. Every person, she argued, fears social disconnection from others.

Lewis's solution to the outside–inside problem parallels and advances the Darwin–Mead–Cooley definition of the social context of shame. She proposed that shame is a bodily and/or mental response to the threat of disconnection from the other. Shame, she argued, can occur in response to threats to the bond from the other, but in can also occur in response to actions in the 'inner theatre,' in the interior monologue in which we see ourselves from the point of view of others. Her reasoning fits Cooley's formulation of shame dynamics, and also Mead's (1934) more general framework: the self is a social construction, a process constructed from both external and internal social interaction, in role-playing and role-taking.

5. Shame as the Social Emotion

Drawing upon the work of these pioneers, it is possible to take further steps toward defining shame. By shame, I mean a large family of emotions that includes many cognates and variants, most notably embarrassment, humiliation and related feelings such as shyness, that involve reactions to rejection or feelings of failure or inadequacy. What unites all these cognates is that they involve the feeling of a threat to the social bond. That is, I use a sociological definition of shame, rather than the more common psychological one (perception of a discrepancy between ideal and actual self). If one postulates that shame is generated by a threat to the bond, no matter how slight, then a wide range of cognates and variants follow: not only embarrassment, shyness, and modesty, but also feelings of rejection or failure, and heightened self-consciousness of any kind. Note that this definition usually subsumes the psychological one, since most ideals are social, rather than individual.

If, as proposed here, shame is a result of threat to the bond, shame would be the most social of the basic emotions. Fear is a signal of danger to the body, anger a signal of frustration, and so on. The sources of fear

and anger, unlike shame, are not uniquely social. Grief also has a social origin, since it signals the loss of a bond. But bond loss is not a frequent event. Shame on the other hand, following Goffman, since it involves even a slight threat to the bond, is pervasive in virtually all social interaction. As Goffman's work suggests, all human beings are extremely sensitive to the exact amount of deference they are accorded. Even slight discrepancies generate shame or embarrassment. As Darwin (1872) noted, the discrepancy can even be in the positive direction; too much deference can generate the embarrassment of heightened self-consciousness.

Especially important for social control is a positive variant, a sense of shame. That is, shame figures in most social interaction because members may only occasionally feel shame, but they are constantly anticipating it, as Goffman implied. Goffman's treatment points to the slightness of threats to the bond that lead to anticipation of shame. For that reason, my use of the term shame is much broader than its vernacular use. In common parlance, shame is an intensely negative, crisis emotion closely connected with disgrace. But this is much too narrow if we expect shame to be generated by even the slightest threat to the bond.

6. *Conclusion*

The classic sociologists believed that emotions are crucially involved in the structure and change of whole societies. The authors reviewed here suggest that shame is the premier social emotion. Lynd's work, particularly, suggests how acknowledgement of shame can strengthen bonds, and by implication, lack of acknowledgment can create alienation. Lewis's work further suggests how shame–anger loops can create perpetual hostility and alienation. Acknowledged shame could be the glue that holds relationships and societies together, and unacknowledged shame the force that divides them.

See also: Civilizational Analysis, History of; Emotion: History of the Concept; Emotions, Evolution of; Emotions, Sociology of; Identity: Social; Moral Sentiments in Society; Norms; Values, Sociology of

Bibliography

Brown P, Levinson S C 1987 *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge

Collins R 1990 Stratification, emotional energy, and the transient emotions. In: Kemper T D (ed.) *Research Agendas in the Sociology of Emotions*. State University of New York Press, Albany, NY, pp. 27–57

Cooley C H 1922 *Human Nature and the Social Order*. C. Scribner's Sons, New York

Darwin C 1872 *The Expression of Emotion in Men and Animals*. John Murray, London

Elias N 1994 *The Civilizing Process*. Blackwell, Oxford, UK

Goffman E 1967 *Interaction Ritual*. Anchor, New York

Gottschalk L A 1995 *Content Analysis of Verbal Behavior*. Lawrence Erlbaum Associates, Hillsdale, NJ

Gottschalk L, Winget C, Gleser G 1969 *Manual of Instruction for Using the Gottschalk–Gleser Content Analysis Scales*. University of California Press, Berkeley, CA

Hochshild A R 1983 *The Managed Heart*. University of California Press, Berkeley, CA

Kemper T O 1978 *A Social Interactional Theory of Emotions*. Wiley, New York

Lewis H B 1971 *Shame and Guilt in Neurosis*. International Universities Press, New York

Lynd H M 1961 *On Shame and the Search for Identity*. Science Editions, New York

Mead G H 1934 *Mind, Self, and Society*. University of Chicago Press, Chicago

Parsons T, Shils E 1951 *Toward a General Theory of Action*. Harvard University Press, Cambridge

Retzinger S M 1991 *Violent Emotions*. Sage, Newbury Park, CA

Retzinger S M 1995 Identifying shame and anger in discourse. *American Behavioral Scientist* **38**: 1104–13

Sennett R 1980 *Authority*. Alfred Knopf, New York

Sennett R, Cobb J 1972 *The Hidden Injuries of Class*, 1st edn. Knopf, New York

Simmel G 1904 Fashion. *International Quarterly* **X**: 130–55 [Reprinted in the *American Journal of Sociology* **62**: 541–59]

T. J. Scheff

Shared Belief

1. *Introducing the Concept of Shared Belief*

The importance of the notion of shared belief has been emphasized by philosophers, economists, sociologists, and psychologists at least since the 1960s (the earlier contributors include Schelling 1960, Scheff 1967, Lewis 1969, and Schiffer 1972). This article deals with the strongest kind of shared belief, to be called 'mutual belief.' While a belief can be shared in the trivial sense of two or more people having the same belief, mutual belief requires the kind of strong sharing which requires at least the participants' awareness (belief) of their similar belief. Mutual belief is one central kind of collective attitude, examples of others being collective intentions, wants, hopes, and fears. Understandably, collective attitudes are central explanatory notions in the social sciences, as one of the tasks of these sciences is to study collective phenomena, including various forms of collective thinking and acting (see below for illustrations).

There is some special, additional interest in the notion of shared belief as mutual belief. First, mutual beliefs serve to characterize social or intersubjective existence in a sense that does not rely on the participants' making agreements or contracts. Thus, many social relations, properties, and events arguably involve mutual beliefs (cf. Lewis 1969, Ruben 1985, Lagerspetz 1995, Tuomela 1995). As a simple example, think of the practice of two persons, *A* and *B*, shaking

hands. It presupposes that *A* believes that *B* and *A* are shaking hands, and that *A* also believes that *B* believes similarly; and *B* must believe analogously. Similarly, communication has been argued by many philosophers, especially Grice, to involve mutual belief (cf. Schiffer 1972, Grice 1989). Second, characterizations of many other collective attitudes in fact depend on the notion of mutual belief (cf. Balzer and Tuomela 1997).

In social psychology and sociology theoreticians often speak about *consensus* instead of mutual belief. The notion of consensus—with the core meaning 'mutual belief'—has been regarded as relevant to such topics as public opinion, values, mass action, norms, roles, communication, socialization, and group cohesion. It can also be mentioned that fads, fashions, crazes, religious movements, and many other related phenomena have been analyzed partly in terms of shared beliefs, consensus, shared consensus, mutual belief, or some similar notions. As pointed out by the sociologist Scheff (1967), such analyses have often gone wrong because they have treated consensus merely as shared first-order belief. Thus, as Scheff argues, consensus as mere first-order agreement does not properly account for 'pluralistic ignorance' (where people agree but do not realize it) and 'false consensus' (where people mistakenly think that they agree). Scheff proposes an analysis in terms of levels of agreement corresponding to a hierarchy of so-called loop beliefs (e.g., in the case of two persons *A* and *B*, *A* believes that *B* believes that *A* believes that something *p*). As is easily seen, pluralistic ignorance and false consensus are second-level phenomena. The third level will have to be brought in when speaking about people's awareness of these phenomena. Other well-known social psychological notions requiring more than shared belief are Mead's concept of 'taking the role of the generalized other,' Dewey's 'interpenetration of perspectives,' and Laing's metaperspectives (see, for example, Scheff 1967).

There are two different conceptual–logical approaches to understanding the notion of mutual (or, to use an equivalent term, common) belief: (a) the *iterative* account, and (b) the *reflexive* or *fixed-point* account. According to the iterative account, mutual belief is assumed to mean iteratable beliefs or dispositions to believe (cf. Lewis 1969, Chap. 2, and, for the weaker account in terms of dispositions to come to believe, Tuomela 1995, Chap. 1). In the two-person case, mutual belief amounts to this according to the iterative account: *A* and *B* believe that *p*, *A* believes that *B* believes that *p* (and similarly for *B*), *A* believes that *B* believes that *A* believes that *p* (and similarly for *B*); and the iteration can continue as far as the situation demands. In the case of loop beliefs there is accordingly mutual awareness only in a somewhat rudimentary sense. As will be seen, in many cases one needs only two iterations for functionally adequate mutual belief: *A* and *B* believe that *p* and they also believe that they believe that *p*. However, there are other cases in which it may be needed to go higher up in the hierarchy.

The fixed-point notion of mutual belief can be stated as follows: *A* and *B* mutually believe that *p* if and only if they believe that *p* and also believe that it is mutually believed by them that *p*. No iteration of beliefs is at least explicitly involved here. Correspondingly, a clear distinction can be made between the iterative or the level-account and the fixed-point account (to be briefly commented on in Sect. 4).

One can speak of the individual (or personal) mode and the group mode of having an attitude such as belief or intention. This article deals with the general notion of mutual belief, be it in the individual mode or in the group mode. Some remarks on the present distinction are anyhow appropriate here. The group-mode sense, expressible by 'We, as a group, believe that *p*,' requires that the group in question is collectively committed to upholding its mutual belief or at least to keeping the members informed about whether it is or can be upheld. This contrasts with mutual belief in an aggregative individual mode involving only personal commitments to the belief in question. When believing in the group-mode sense, group members are accordingly committed to a certain shared view of a topic and to group-mode thoughts such as 'We, as a group, believe that *p*.' Group-mode beliefs are central in the analysis of the kinds of beliefs that structured social groups such as organizations and states have. According to the 'positional' account defended by Tuomela (1995), the group members authorized for belief or view formation collectively accept the views which will qualify as the beliefs the group has. These views are group-mode views accepted for the group and are strictly speaking acceptances of something as the group's views rather than beliefs in the strict sense.

2. *Mutual Beliefs More Precisely Characterized*

This section will be concerned with what mutual beliefs involve over and above plain shared beliefs, and the Sect. 3 takes up the problem of how many layers of hierarchical beliefs are conceptually or psychologically needed. (The discussion below draws on the treatment in Tuomela 1984 and 1995.)

Consider now the iterative account starting with the case of two persons, *A* and *B*. Recall from Sect 1 that according to the standard iterative account *A* and *B* mutually believe that *p* if and only if *A* believes that *p*, *B* believes that *p*, *A* believes that *B* believes that *p* (and similarly for *B*), *A* believes that *B* believes that *A* believes that *p* (and similarly for *B*), and so on, in principle ad infinitum. In the general case, the account defines that it is mutually believed in a group, say *G*, that *p* if and only if *a*) everyone in *G* believes that *p*, *b*) everyone believes that everyone believes that *p*, *c*) everyone believes that everyone believes that everyone

believes that p, and so on *ad infinitum*. The word 'everyone' can of course be qualified if needed and made dependent on some special characteristics; for example, it can be restricted to concern only every fully fledged, adequately informed, and suitably rational member of the group.

A major problem with the iterative account is that it seems in some cases it is psychologically realistic. It loads the people's minds with iterated beliefs which people, after all, do not experientially have and possibly, because of lack of rationality and memory, cannot have. Thus, the account must be improved to make it better correspond to psychological reality. One way to go is to operate partly in terms of lack of disbelief. A person's lack of disbelief that p is defined as that it is not the case the person believes the negation of p. One can try to define mutual belief schematically by saying that it is mutually believed in G that p if and only if everyone believes that p, iterated n times, and from level $n+1$ on everyone lacks the disbelief that p. With $n = 2$, viz. two levels actually present, this definition says that mutual belief amounts to everyone's believing that p and everyone's believing that everyone believes that p and that everyone lacks the disbelief that p from the second level on.

The basic reason for using this amended iterative approach is that no more levels of positive belief than people actually have should be required. Provided that the value of n can be established, the present account satisfies this. It must be assumed here that the agents in question are to some degree intelligent and rational as well as free from emotional and other disturbances so that they, for instance, do not lack a higher-order belief when they really ought to have one in order to function optimally.

While the iterative account amended with the use of the notion of lack of disbelief seems viable for some purposes, an alternative analysis is worth mentioning. This analysis amends the original iterative account in that it requires iterative beliefs up to some level n and from that level on requires only the disposition to acquire higher-order beliefs in appropriate conditions. These appropriate conditions, serving to determine the value of n, include both background conditions and more specific conditions needed for acquiring the belief in question. First, the agents in G must be assumed to be adequately informed and share both general cultural and group-specific information; especially, they must have the same standards of reasoning so that they are able to 'work out the same conclusions' (see Lewis 1969, Chap. 2, and Heal 1978). Second, they must be sufficiently intelligent and rational, and they must be free from cognitive and emotional disturbances to a sufficient degree (so that there are no psychological obstacles for adding the $n+1$st order belief when a genuine mutual belief is involved and for not adding it, or indeed for adding its negation, when a genuine mutual belief is not involved). A typical releasing condition for the disposition to acquire a higher-order belief would be simply that the agents are asked about higher-order beliefs (or are presented with other analogous problems concerning them).

There is not much to choose between the two amended accounts, and the matter will not be discussed further here. With the conceptual machinery at hand, one can deal with, for instance, the mentioned phenomena of pluralistic ignorance and false consensus. An account of pluralistic ignorance, concerning something p, must obviously include the idea that everyone believes that p. Second, it must say that the agents do not have the belief that they agree. Compatibly with this, it can still be required that that they do not disbelieve that they believe that p. As to false consensus, it must be required that not everyone believes that p and that, nevertheless, it is believed by everyone that everyone believes that p.

3. *Mutual Beliefs and the Level Problem*

Let us next discuss the level problem. Given the iterative account, the problem is how many iterative levels of belief are conceptually, epistemically, or psychologically needed for success in various cases. The criteria here are functionality and success. For instance, how many levels of belief does successful joint action require? The level question is difficult to answer in general terms. However, it can be conjectured that in certain cases at least two levels (viz. $n = 2$), relative to the base level, are needed for mutual belief, but no more. In some other cases this is not enough. Generally speaking, there are very seldom epistemic or psychological reasons for going beyond the fourth level. For many agents level $n = 4$ probably is psychologically impossible or at least very difficult to handle in one's reasoning.

The base level is the level zero at which the sentence or proposition p is located, and p will concern or be about an agent's or agents' actions, dispositions to act, intentions, beliefs (or something related); or, alternatively, p can be (or concern) a belief content. In the analysis below it is accepted that the beliefs—at least those going beyond the level $n = 2$—need not be proper occurrent or standing beliefs or even subconscious ones but only dispositions to form the belief in question. (See Audi 1994, for the relevant notion of a disposition to form a belief.)

The claim that $n = 2$ is not only necessary but also sufficient in many typical cases is allowed to involve social loop beliefs where the content of mutual belief could be, for instance, that A believes that B believes that A will act in a certain way, or believes that such and such is the case. What the base level will be is relative to the topic at hand. That at least two levels are needed can be regarded as a conceptual truth (related to the notion 'social existence'); that in many typical cases no more is needed or that in some other

cases four layers are needed can be regarded as general psychological theses concerning the actual contents of people's minds in those situations.

The present view allows that a need may arise to add layers indefinitely. There follows a recursive argument to show that such a situation is possible in principle. Given an analysis of mutual belief in terms of iteration of beliefs such as ours, then at each level under consideration one may start asking questions about people's realizing (or believing) such and such, where 'such and such' refers to the level at hand. This game obviously can go on indefinitely as far as the involved agents' reasoning capacities (or whatever relevant capacities) permit. This is an inductive argument (in the mathematical sense—the first step is obvious) for the open and infinite character of the hierarchy of nested beliefs.

To discuss the level problem in more concrete terms, let us consider mutual belief in the case of two agents, A and B. First we let our base-level sentences be related to the agents' performances of their parts as follows:

(a) A will do his part of X (in symbols, $p(A)$);

(b) B will do his part of X ($p(B)$).

Next, let us assume:

(i) A believes that (a) and (b);

(ii) B believes that (a) and (b).

This is a social or mutual recognition of the agents' part-performances in the context of joint action.

The sentence p in (a) and (b) may alternatively concern other things than action. Thus it may represent a belief content, for example $p =$ The earth if flat, and we will consider this case below.

Considering joint action, it can be argued that $n = 2$ is both necessary and sufficient for successful performance. In this example, A will obviously have to believe that B will do her part. This gives her some social motivation to perform her own part. Concentrating on our reference point individual A and genuine joint action X (in which the agents' performances of their parts are interdependent), A must also believe (or be disposed to believe, at any rate, viz. come to form the belief if asked about the matter) that B believes that A will do her part. For otherwise A could not believe with good reason that B will do her part, for B would—if she does not so believe—lack the social motivation springing from her belief that A will do her part. However, if she did lack that motivation, A could not on this kind of ground defensibly form the belief that B will perform her part. This argument shows that it must be required that the agents have a loop belief or at least a disposition to have a loop belief:

(iii) A believes that (i) and (ii);

(iv) B believes that (i) and (ii).

For instance, (iii) gives the loop 'A believes that B believes that $p(A)$,' and this means that in the present kind of case $n = 2$ is necessary and normally also sufficient (relative to the chosen base level).

An analogous claim can be defended in the case of belief. Here the defense can be given in terms of the recognition by the agents that they share a belief. Suppose that each member of the Flat Earth Society not only believes that the Earth is flat, but—because of being organized into a belief group—the members all share this information. This second-order belief indicates the first level at which social and mutual belief can be spoken of in an interesting sense, for at this level people recognize that the others in the group believe similarly, and this establishes a doxastic social connection; at level one that recognition is, however, missing. This is a typical case of a social property, and again $n = 2$ is both necessary and normally sufficient for acting in matters related to the shape of the Earth *qua* a member of the society.

Suppose next the members of the Flat Earth Society now begin to reflect upon this second-order belief of theirs and come to form the belief that indeed it does exist. (Perhaps a sociologist has just discovered that they share the second-order belief and told them about it.) They might wonder how strong this second-order awareness is, and so on. All this entails that on some occasions third-order beliefs may be needed, although the standard case clearly would be second-order beliefs accompanied by a disposition to go higher up, if the need arises. Analogously, in the case of joint action in some very special cases, for example, double loop beliefs may be needed for rational action.

The reader should be reminded that not all social notions depend on mutual belief. For instance, latent social influence and power do not require it, nor does a unilateral love relation between two persons.

4. *Mutual Belief as Shared We-belief*

What has been called 'we-attitudes' are central for an account of social life (see Tuomela, 1995, Chap. 1, Tuomela and Balzer 1999). A we-attitude is a person's attitude (say belief, want, fear, etc.) concerning something p (a proposition, or sentence) such that (1) this person has the attitude in question and believes that (2) everyone in the group has the attitude and also believes that there is a mutual belief to the effect that (2) in the group. Of these, clause (1) is required because there cannot of course be a truly shared attitude without all the members participating in that attitude. Clause (2) gives a social reason for adopting and having the attitude A, and (3) strengthens the reason by making it intersubjective. (A shared we-attitude can be either in the group mode or in the individual mode; cf. Sect. 1.)

A shared we-belief is a we-belief (be it in the individual or in the group mode) which (ideally) all the group members have. To take an example, a shared we-belief that the Earth is flat entails that the group members believe that the Earth is flat, that the group members believe that the Earth is flat, and also that it

is a mutual belief in the group that the group members believe that the Earth is flat.

Shared we-beliefs can be related to the reflexive or fixed-point account of mutual belief. According to the simplest fixed-point account, mutual belief is defined as follows: it is a mutual belief in a group that p if and only if everyone in the group believes that p and also that it is mutually believed in the group that p. It can thus be seen that the account of mutual belief given by the fixed-point theory is equivalent to the definiens in the definition of a shared we-belief. In the fixed-point approach the syntactical infinity involved in the iterative approach is cut short by a finite fixed-point formula, that is an impredicative construct in which the joint notion to be 'defined' already occurs in the definiens. Under certain rationality assumptions about the notion of belief it can be proved that the iterative approach, which continues iterations ad infinitum, gives the fixed-point property as a theorem (see Halpern and Moses 1992 and, for a more general account, Balzer and Tuomela 1997).

The fixed-point account is in some contexts psychologically more realistic, as people are not required to keep iterative hierarchies in their minds. Note, however, that it depends on context whether the iterative approach or the fixed-point approach is more appropriate. Thus, in the case of successful joint action, at least loop beliefs must be required.

See also: Action, Collective; Collective Behavior, Sociology of; Collective Beliefs: Sociological Explanation; Collective Identity and Expressive Forms; Collective Memory, Anthropology of; Collective Memory, Psychology of; Genes and Culture, Coevolution of; Human Cognition, Evolution of; Memes and Cultural Viruses; Religion: Culture Contact; Religion: Evolution and Development; Religion, Sociology of

Bibliography

Audi R 1994 Dispositional beliefs and dispositions to believe. *Nous* **28**: 419–34

Balzer W, Tuomela R 1997 A fixed point approach to collective attitudes. In: Holmström-Hintikka G, Tuomela R (eds.) *Contemporary Action Theory II*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 115–42

Grice P 1989 *Studies in the Ways of Words*. Harvard University Press, Cambridge, MA

Halpern J, Moses Y 1992 A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence* **54**: 319–79

Heal J 1978 Common knowledge. *Philosophical Quarterly* **28**: 116–31

Lagerspetz E 1995 *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Lewis D 1969 *Convention, A Philosophical Study*. Harvard University Press, Cambridge, MA

Ruben D-H 1985 *The Metaphysics of the Social World*. Routledge, London

Scheff R 1967 Toward a sociological model of consensus. *American Sociological Review* **32**: 32–46

Schelling R 1960 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA

Schiffer S 1972 *Meaning*. Oxford University Press, Oxford, UK

Tuomela R 1984 *A Theory of Social Action*. Reidel, Dordrecht, The Netherlands

Tuomela R 1995 *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford Series in Philosophy, Stanford University Press, Stanford, CA

Tuomela R, Balzer W 1999 Collective acceptance and collective social notions. *Synthese* **117**: 175–205

R. Tuomela

Sherrington, Sir Charles Scott (1857–1952)

> Like this old Earth that lolls through sun and shade,
> Our part is less to make than to be made.
> (Sherrington 1940, p. 259)

To characterize the broad span of Sherrington's engagement he should be described as 'Pathologist, Physiologist, Philosopher and Poet.' It is somewhat arbitrary to pick out only two of them because it is the combination of all these elements that made him both great and unique. However, hardly any other physiologist since 1800 made such a fundamentally important contribution to the knowledge and to the development of concepts in modern neurophysiology. Moreover, when advanced in years, he was the first to attempt a contribution to the solution of the age-old problem of brain/body and mind on a well-founded neuroscientific basis.

1. The Life

Born in London on November 27, 1857, Sherrington was brought up in Ipswich and Norwich. From his earliest years he felt a strong attraction to poetry, a love which never left him throughout his life. Some of his poems were published in various magazines and later were partly collected in a small volume (Sherrington 1925).

In 1876 Sherrington first enrolled at St. Thomas' Hospital Medical School in London and in 1881 matriculated at Cambridge University, where he received his B.A. in 1884 and his M.D. in 1885. During his undergraduate years he started his research in neuroanatomy and neuropathology, publishing his first papers on these studies in 1884 together with John Newport Langley (Langley and Sherrington 1884). This was the beginning of a very productive publishing activity with almost 320 scientific publications (for Sherrington's complete bibliography see Fulton 1952). During his initial postgraduate period at Cambridge

Sherrington's scientific interests concentrated on neurohistology and pathology. This period was interrupted by expeditions to epidemics of Asian Cholera in Spain and Italy. The autopsied material taken during these expeditions he analyzed together with J. Graham Brown and Charles S. Roy in England (1885) and with Rudolf Virchow and Robert Koch in Germany (1886). In 1890 Sherrington was appointed lecturer in physiology at St. Thomas' Hospital in London where he mainly performed clinical work. A year later he received the position of the Professor-Superintendent at the Brown Institute at the University of London, where he became remarkably productive with a variety of publications which reflected his pathological, but also his increasing neurophysiological interest. In 1895 Sherrington received his first appointment to a full professorship of physiology at the University College in Liverpool, where he investigated intensely the organization of the spinal cord.

At age 56 (in 1913) Sherrington entered probably the busiest period of his scientific career when he was appointed to the chair of physiology at Oxford University, from which he did not retire until 1935, when he was 78. It was during his Oxford time that he, together with Edgar Adrian, received the Nobel Prize for Medicine in 1932 (the greatest tribute among numerous honours and rewards, like, for example the Presidency of the Royal Society 1920–25, some 21 international honorary doctorates, and many medals and honorary memberships of British and international societies and academies, among these the corresponding membership of *l'Institut de France*). Three of his students at Oxford were later Nobel Prize winners: Howard Walter Florey, John Carew Eccles and Ragnar Granit.

In the early 1930s Sherrington performed his last physiological experiments and he felt free to turn to his other great love, the philosophy of the nervous system. Having retired to his old hometown of Ipswich in 1935, his scientific interest was increasingly enriched by philosophical considerations that became evident in the Gifford Lectures, held between 1937 and 1938. Obviously, it was also Sherrington's interest in Natural Philosophy that caused him to work on Goethe (Sherrington 1942) and Jean Fernel, a French physician and philosopher of the sixteenth century (Sherrington 1946).

Though his body was crippled by arthritis, his intellectual capacities remained undimmed until his death in Eastbourne on March 4 1952.

2. *The Physiologist*

Regarding the nervous system as a functional unity resulting from its well-organised parts, Sherrington investigated a variety of motor and sensory systems: the topographical organisation of the motor cortex, the binocular integration in vision, as well as the different receptor systems. However, his main efforts concentrated on the spinal cord and its reflex function in motor control. He explicitly admitted that spinal reflex behaviour is not 'the most important and far-reaching of all types of nerve behavior' but represents the advantage that 'it can be studied free from complication with the psyche' (Sherrington 1906/1947, Foreword, p. xiii).

In the spinal cord, isolated from the brain, he carefully analyzed almost all spinal reflex-types from the 'simple' two-neurone reflex arc of the stretch reflex (first described by him), via the complex partly nociceptive flexion and crossed extension reflex and co-ordinated rhythmic scratching and stepping reflexes, up to goal-directed reflex movements of the limbs towards irritating stimuli. Right from the beginning he demonstrated that spinal reflexes are not following a stereotyped all-or-nothing behavior, but that the adaptability of reflexes, based on graduated amplitude, irradiation, and mutual interference between different reflex types, is of great importance for motor control. Sherrington was the first to suggest that inhibition had at least the same functional importance for motor coordination—and other nervous functions—as excitation, not only at the spinal level but also at the brain level and for descending motor commands. This view also opened up new insights into the background of motor disorders observed in various neurological syndromes.

Though some of the reflexes and other motor and sensory phenomena described by Sherrington had been noted earlier, his experimental observations and their analyses were altogether more exact and careful than those of former investigations. Thus, he disproved e.g., several of Pflügers's reflex laws and the assumption of Descartes (1677) and John and Charles Bell (1826) that reciprocal inhibition is localized peripherally in the muscle. However, the higher experimental sophistication alone would not for sure have received that particularly high enduring recognition. Indeed, Sherrington's most important contribution to neuroscience resulted from his unique capability to condense his immense knowledge on sensory and motor functions and on the structure of the nervous system (partly derived from an intense contact with the famous Spanish neurohistologist Santiago Ramón y Cajal) in an almost visionary synopsis of the integrative action of the nervous system, which laid the foundation for the great progress of neuroscience in the twentieth century and which has lost none of its basic validity. This synopsis made his book *The Integrative Action of the Nervous System* (1906) a milestone in neurophysiology, which has been compared 'in importance with Harvey's' 300 years older '*De motu cordis* in marking a turning point in the history of physiological thought' (Fulton 1952, p. 174). For Sherrington, spinal reflexes did not have a 'purpose' but a 'meaning.' He replaced the formerly presumed 'soul' of the spinal cord (e.g., Legallois 1812, Pflüger

1853) as the driving force for reflexes and even voluntary movements by an astonishing exactly considered function of spinal interneuronal activity. He replaced the Cartesian reflex behaviorist view by the still-relevant view that the performance of coordinated movements as a basis for animal and 'conscious' human behavior requires an integration of 'intrinsically' generated brain functions with 'extrinsically' induced reflex functions. In this context he suggested that the pure 'apsychical' reflex behavior without a contribution of 'mind' loses in importance with phylogenetic ascendance, playing the smallest role in man: 'The spinal man is more crippled than is the spinal frog' (Sherrington 1906/1947, Foreword, p. xiv).

As a true scientist Sherrington never felt that he had arrived at any incontrovertible truth and in a typically unpretentious understatement he concluded: 'I have dealt very imperfectly with a small bit of a large problem…. The smallness of the fragment is disappointing' (1931 p. 27); this a statement from a man about whom Frederick George Donnan wrote in 1921 to the Nobel Committee: 'In the field of physiology and medicine, Professor Sherrington's works remind one of that of Kepler, Copernicus and Newton in the field of mechanics and gravitation' (in H. Schück et al., 1962, p. 310).

3. *The Philosopher*

Like for example, Descartes before and Eccles after him, Sherrington realised that mere experimental investigation of the nervous system was not sufficient to solve the fundamental problem of the body/brain–mind relation, particularly the difficulty of explaining how it actually works. In the Rede Lecture on *The Brain and its Mechanism* (Sherrington 1933) in which he considered the complexity of the brain functions in motor control and behavior and in which he tentatively approached this problem, Sherrington took a fully dualistic view. He even negated any scientific right to 'conjoin mental experience with the physiological…. The two … seem to remain disparate and disconnected' (1933, pp. 22–3).

But later in the Gifford Lectures, which he held in Edinburgh in 1937–1938 (published as *Man on his Nature* in Sherrington 1940 and as an extensively revised new edition in 1951) he took up the great enigma of the brain–mind relation in the light of his new neuroscientific knowledge. Rejecting the assumption of a mysterious mind of a 'heavenly' origin he stated, 'Ours is an earthly mind which fits our earthly body' (Sherrington 1940, p. 164). 'Mind' for Sherrington was coupled to motor acts, 'mind' only being recognizable via motor acts. He suggested that 'mind' is graduable, that a recognizable 'mind' is developing in the developing brain in humans and to varying degrees in animals. Regarding the relation between the two concepts of 'energy' (assuming 'matter' also as 'energy' in the physical sense) and 'mind', he concluded that 'mind' as a function of the living brain is mortal, while the 'energy' the matter—is immortal, just changing its state when the brain is dying. In his discussions he entirely avoided the complicating problem of an 'immortal soul', which he, as being himself oriented towards Natural Religion, left to the Revealed Religions: 'When on the other hand the mind-concept is so applied as to insert into the human individual an immortal soul, again a trespass is committed' (Sherrington 1940, p. 355).

When Sherrington with something like resignation stated, 'But the problem of how that [body-mind] liaison is effected remains unsolved; it remains where Aristotle left it more than 2000 years ago' (Sherrington 1906/1947, Foreword, p. xxiii) he probably underestimated the contribution of his step towards the direction of a solution by defining 'mind' as a function of the brain requiring a complex integrative action of the nervous system; a view which was taken up and further developed by one of his students, Sir John C. Eccles (1970, Eccles and Gibson 1979).

4. *And Now?*

As stated in Sect. 3 the experimentally well-based theories and concepts, which Sherrington had creatively developed with great visionary imagination, formed a well-recognised secure foundation for future developments in the neurosciences.

The half-century after Sherrington's death has brought an immense increase in neuroscientific knowledge. First, particularly, the technique of microrecordings from nerve cells and the analysis of ionic membrane mechanisms confirmed and extended the knowledge and hypotheses based on Sherrington's activities on brain and spinal functions in sensory perception, motor control, and behavior. Then new techniques allowed for analytic investigation of more and more detailed neuronal structures, functions, and interactions down to the molecular level. New gentechnical approaches enabled the 'production' of innumerable variants of mice with defined genetic defects at the molecular level.

However, the bewildering mass of results originating from these experiments opened up a host of new questions. What is required now is a new 'Sherringtonian' approach which comprises a complex critical analysis of today's entire neuroscientific knowledge (from molecular mechanisms up to the level to complex behaviour) in an integrative synopsis and which develops new concepts to explain the sophisticated brain functions and their relation to mind as expressed by behavior. Despite the exponential increase in detailed neuroscientific knowledge, we have not really come closer to a solution of that problem since Sherrington.

See also: Behavioral Neuroscience; Cognitive Neuroscience; Consciousness, Neural Basis of; Cross-modal

(Multi-sensory) Integration; Functional Brain Imaging; Motor Control; Motor Cortex; Motor Skills, Psychology of; Neural Plasticity of Spinal Reflexes; Neural Representations of Direction (Head Direction Cells); Psychophysiology; Theory of Mind; Topographic Maps in the Brain; Vestibulo-ocular Reflex, Adaptation of the

Bibliography

Bell J, Bell C 1926 *Anatomy and Physiology of the Human Body*. Edinburgh

Descartes R 1677 *De homine* ed. by de la Forge. Amsterdam

Eccles J C 1970 *Facing Reality*. Springer-Verlag, New York

Eccles J C, Gibson W C 1979 *Sherrington—His Life and Thought*. Springer International, Berlin

Fulton J F 1952 Sir Charles Scott Sherrington, O.M. *Journal of Neurophysiology* **15**: 1667–190

Granit R 1966 *Charles Scott Sherrington—An Appraisal*. Nelson, London

Langley J N, Sherrington C S 1884 Secondary degeneration of nerve tracts following removal of the cortex of the cerebrum in the dog. *Journal of Physiology* **5**: 49–65

Legallois M 1812 *Expériences sur la principe de la vie*. D'Hautes, Paris

Pflüger E 1853 *Die sensorischen Funktionen des Rückenmarks der Wirbeltiere*. Springer, Berlin

Schück H, Sohlman R, Österling A, Liljestrand G, Westgren A, Siegbahn M, Schou A, Ståle N K 1962 *Nobel, the Man and His Prizes*. The Nobel Foundation, Elsevier, Amsterdam

Sherrington C S 1906/1947 *The Integrative Action of the Nervous System*, 2nd edn, Yale University Press, New Haven, CT, with a new Foreword, 1947

Sherrington C S 1925 *The Assaying of Brabantius and Other Verse*. Oxford University Press, London

Sherrington C S 1931 Quantitative management of contraction in lowest level co-ordination. Hughlings Jackson Lecture. *Brain* **54**: 1–28

Sherrington C S 1933 *The Brain and its Mechanism*. Cambridge University Press, Cambridge, UK

Sherrington C S 1940/1951 *Man on his Nature*, 2nd rev. edn. 1951. Cambridge University Press, Cambridge, UK

Sherrington C S 1942 *Goethe on Nature and on Science*. Cambridge University Press, Cambridge, UK

Sherrington C S 1946 *The Endeavour of Jean Fernel*. Cambridge University Press, Cambridge, UK

E. D. Schomburg

Short-term Memory, Cognitive Psychology of

The concept of short-term memory has been of theoretical significance to cognitive psychology since the late 1950s. Some investigators have even argued that 'all the work of memory is in the short-term system' (Shiffrin 1999, p. 21). However, others have claimed that short-term memory is an archaic concept and there is no need to distinguish the processes involved in short-term memory from other memory processes (Crowder 1993).

1. *Definitions and Terminology*

'Short-term memory' refers to memory over a short time interval, usually 30 s or less. Another term for the same concept is 'immediate memory.' Both these terms have been distinguished from the related terms 'short-term store' and 'primary memory,' each of which refers to a hypothetical temporary memory system. However, the term 'short-term memory' has also been used by many authors to refer to the temporary memory system. Thus, here the term 'short-term memory' is used in both senses.

It is important to distinguish 'short-term memory' from the related concepts 'working memory' (see *Working Memory, Psychology of*) and 'sensory memory'. Some authors have used the terms 'short-term memory' and 'working memory' as synonymous, and indeed the term 'working memory' has been gradually replacing the term 'short-term memory' in the literature (and some authors now refer to 'short-term working memory'; see Estes 1999). However, 'working memory' was originally adopted to convey the idea that active processing as well as passive storage is involved in temporary memory. 'Sensory memory' refers to memory that is even shorter in duration than short-term memory. Further, sensory memory reflects the original sensation or perception of a stimulus and is specific to the modality in which the stimulus was presented, whereas information in short-term memory has been coded so that it is in a format different from that originally perceived.

2. *Historical Development and Empirical Observations*

An initial description of short-term memory was given by James (1890), who used the term 'primary memory' and described it as that which is held momentarily in consciousness. The intense study of short-term memory began almost 70 years later with the development of the distractor paradigm (Brown 1958, Peterson and Peterson 1959).

2.1 *The Distractor Paradigm*

In this paradigm, a short list of items (usually five or fewer) is presented to subjects for study. If the subjects recall the list immediately, perfect performance results because the list falls within their memory span. However, before the subjects recall the items, they engage in an interpolated activity, the distractor task,

which usually involves counting or responding to irrelevant material. The purpose of the distractor task is to prevent the subjects from rehearsing the list items. The length of the distractor task varies, and its duration is called the 'retention interval.' By comparing performance after various retention intervals, it was found that the rate of forgetting information from short-term memory is very rapid, so that after less than 30 s little information remains about the list of items.

Another central finding from the distractor paradigm involves the serial position curve, which reveals accuracy for each item in the list as a function of its position. The curve is typically bowed and symmetrical with higher accuracy for the initial and final positions than for the middle positions. The distractor task typically requires serial recall (i.e., recall in the order in which the items were presented; see *Representation of Serial Order, Cognitive Psychology of*). Two types of errors occur with this procedure. Transposition errors are order errors in which subjects recall an item in the wrong position. Nontransposition errors are substitution errors in which subjects replace an item with one not included in the list. For example, if subjects receive the list BKFH and recall VKBH, they make a nontransposition error in the first position and a transposition error in the third. The bowed serial position curve reflects the pattern of transposition errors. The pattern of nontransposition errors shows instead a relatively flat function with errors increasing slightly across the list positions.

Errors in the distractor paradigm can also be classified by examining the relation between the correct item and the item that replaces it. In the example, the subjects replace B with V. Because those two letters have similar-sounding names, the error is called an 'acoustic confusion error', and such errors occur often in the distractor paradigm. Even visually presented items are typically coded in short-term memory in an acoustic or speech representation. This fact was first illustrated in an experiment in which subjects were shown a list of six consonants for immediate serial recall. The errors were classified by a confusion matrix, in which the columns indicate the letter presented and the rows indicate the letter actually recalled. A high correlation was found between the confusion matrix resulting from this memory task when subjects recalled a list of visually presented letters, and the confusion matrix resulting from a listening task when subjects heard one letter at a time and simply named it with no memory requirement (Conrad 1964).

2.2 *The Free Recall Task*

Another paradigm commonly used in the early investigation of short-term memory was the free recall task, in which subjects are given a relatively long list of items and then recall them in any order they choose. When the list is recalled immediately, the subjects show greater memory for the most recent items. This 'recency effect' was attributed to the fact that the final items in the list, but not the earlier ones, are still in short-term memory at the termination of the list presentation. Support for this conclusion came from the observation that if presentation of the list is followed by a distractor task, then there is no recency effect, although as in the case of immediate recall there is a 'primacy effect,' or advantage for the initial items in the list. The explanation offered for the elimination of the recency effect with the distractor task is that the final list items are no longer in short-term memory after the distractor activity. Further support for this explanation came from finding that other variables like list length and presentation rate had differential effects on the recency and earlier sections of the serial position curve. Specifically, subjects are less likely to recall an item when it occurs in a longer list for all except the most recent list positions. Likewise, subjects are less likely to recall an item in a list presented at a faster rate for all but the recency part of the serial position curve.

3. *Theoretical Accounts*

The most widely accepted account of short-term memory was presented in the 1960s in what was subsequently termed the 'modal model' because of its popularity. The core assumption of that model is the distinction between short-term memory, which is transient, and long-term memory, which is relatively permanent. The fullest description of the modal model was provided by Atkinson and Shiffrin (1968), who also distinguished between short-term memory and sensory memory.

3.1 *The Buffer Model*

Atkinson and Shiffrin's 'buffer model,' as it has been called, is characterized along two dimensions (Atkinson and Shiffrin 1968). Following a computer analogy, the first dimension involves the structural features of the system, analogous to computer hardware, and the second dimension involves the 'control processes,' or operations under the control of the subjects, analogous to computer software. The structural features of the buffer model include the sensory registers (with different registers for each sense), short-term store, and long-term store. The control process emphasized is rote rehearsal, which takes place in part of short-term store called the 'buffer.' The rehearsal buffer has a small capacity with a fixed number of slots (about four).

To account for performance in the free recall task, it is assumed that each item enters the buffer and when the buffer is full the newest item displaces a randomly selected older item. While an item is in the buffer,

information about it is transferred to long-term store, with the amount of information transferred a linear function of the time spent in the buffer. Although information about an item thereby gets transferred to long-term store, the item remains in the buffer until it is displaced by an incoming item. At test, subjects first respond with any items in the buffer and then try to retrieve other items from long-term store, with the number of retrieval attempts fixed (see *Memory Retrieval*). These assumptions allow the model to account for the various empirical observations found with the free recall task. The model accounts for the recency effect and its elimination with a distractor task because the final items are still in the buffer immediately after list presentation but get displaced from the buffer by the interpolated material of the distractor task. The model accounts for the primacy effect by assuming that the buffer starts out empty so the initial items reside in the buffer longer than subsequent items because they are not subject to displacement until the buffer is full. The effect of list length is due to the fixed number of attempts to retrieve information from long-term store. The longer the list, the smaller is the likelihood of finding a particular item. The effect of presentation rate is due to the linear function for transferring information from short-term to long-term store. More information is transferred when the rate is slower so that retrieving an item from long-term store is more likely at a slower rate.

There have been numerous refinements and expansions of the buffer model since it was first proposed. These updated versions have been termed 'SAM' (search of associative memory) and 'REM' (retrieving effectively from memory). However, these refinements have been focused on the search processes in long-term memory, and the short-term memory component of the model remains largely intact (Shiffrin 1999).

3.2 The Perturbation Model

Whereas the buffer model explains the results of the free recall task, a popular model proposed by Estes (1972) explains the results of the distractor paradigm. According to this 'perturbation model', the representation in memory of each item in a list is associated with a representation of the experimental context. This contextual representation is known as a 'control element.' It is assumed that if the control element is activated, then the representations of the items associated with it are activated in turn, allowing for their recall. Forgetting is attributed in part to the fact that the context shifts with time so that subjects may be unable to activate the appropriate control element after a delay. Reactivation of the item representations by the control element does not only occur at the time of test. In addition, there is a reverberatory loop providing a periodic recurrent reactivation of each item's representation, with the difference in reactivation times for the various items reflecting the difference in their initial presentation times. This reverberatory activity provides the basis for the short-term memory of the order of the items in a list. Because of random error in the reactivation process, timing perturbations result, and these perturbations may be large enough to span the interval separating item representations, thereby leading to interchanges in the order of item reactivations and, hence, transposition errors during recall. Because such interchanges can occur in either the forward or backward direction for middle items in the list but in only one direction for the first and last items, the bowed serial position function for transposition errors is predicted. The symmetry in the serial position function is predicted with the assumption that timing perturbations do not start until all the list items have been presented. Because interchanges can also occur between the list items and the interpolated distractor items, the gradually increasing proportion of nontransposition errors is also predicted.

The original version of the perturbation model included only the perturbation process responsible for short-term memory, but subsequent research documented the need to include a long-term memory process in addition to the short-term perturbation process (Healy and Cunningham 1999). Other extensions allowed the perturbation model to account for a wide range of findings in the distractor paradigm and also to provide insights into other memory processes such as those responsible for memory distortions in eyewitness situations (Estes 1999; see also *Eyewitness Memory: Psychological Aspects*).

3.3 Theoretical Controversies and Complications

Although models of memory typically include the distinction between short- and long-term memory, some investigators have pointed to problems with some of the evidence establishing the need to postulate a distinct short-term memory. For example, Craik and Lockhart (1972) argued that short-term memory cannot be distinguished by the exclusive use of speech coding, as suggested by Conrad (1964). Craik and Lockhart proposed an alternative framework called 'levels of processing,' according to which information is encoded at different levels and the level of processing determines the subsequent rate of forgetting (see *Memory: Levels of Processing*).

More recent arguments against the need for a separate short-term memory were made by Crowder (1993), who pointed out that the rapid forgetting across retention intervals in the distractor paradigm is not found on the first trial of an experiment. He also pointed out that a recency effect like that in immediate free recall is found in a number of tasks relying exclusively on long-term memory. Healy and McNamara (1996) dismissed some of these arguments with two general considerations. First, information

can be rapidly encoded in long-term memory, so that, for example, recall can derive from long-term memory rather than short-term memory on the first trial of an experiment before interference from previous trials has degraded the long-term memory representation. Second, although recency effects occur in many memory paradigms, the specific properties and causes of the serial position functions differ across paradigms.

Memory models also typically make the distinction between sensory and short-term memory. However, an important exception is Nairne's 'feature model' (Nairne 1990). Instead of distinguishing between these two types of memory stores or processes, the feature model distinguishes between two types of memory trace features—modality independent (which involve speech coding) and modality dependent (which are perceptual but not sensory). According to this model, during recall subjects compare the features of the memory trace to the features of various item candidates. Forgetting occurs as the result of feature overwriting, by which a new feature can overwrite an old feature but only if the two features are of the same type (modality dependent or modality independent).

Although short-term memory is distinguished from both sensory and long-term memory by most contemporary models of memory, including a recent 'composite model' proposed by Estes (1999) to describe the current state of theorizing, many current models have broken down short-term memory into different subcomponent processes. The most popular of these models is Baddeley's 'working memory model' (Baddeley 1992), which includes an attentional control system called the 'central executive' (see *Attention: Models*) and two 'slave systems,' the phonological loop (responsible for speech coding) and the visuospatial sketchpad (responsible for the coding of both spatial and visual information). Recent neuropsychological evidence (Martin and Romani 1994) has led to a further breakdown into semantic and syntactic components in addition to the components of Baddeley's system.

Despite these controversies and complications, it seems clear that the concept of 'short-term memory' will continue in the future to play an important role in our theoretical understanding of cognitive processes.

See also: Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Long-term Potentiation (Hippocampus); Memory Models: Quantitative; Short-term Memory: Psychological and Neural Aspects; Visual Memory, Psychology of; Working Memory, Neural Basis of; Working Memory, Psychology of

Bibliography

Atkinson R C, Shiffrin R M 1968 Human memory: A proposed system and its control processes. In: Spence K W, Spence J T (eds.) *The Psychology of Learning and Motivation: Advances in Research and Theory 2*. Academic Press, San Diego, pp. 89–195

Baddeley A-A. 1992 Is working memory working? The fifteenth Bartlett lecture. *Quarterly Journal of Experimental Psychology* **44**: 1–31

Brown J 1958 Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology* **10**: 12–21

Conrad R 1964 Acoustic confusions in immediate memory. *British Journal of Psychology* **55**: 75–84

Craik F I M, Lockhart R S 1972 Levels of processing: A framework for memory research. *Journal of Verb Learning and Verb Behavior* **11**: 671–84

Crowder R G 1993 Short-term memory: Where do we stand? . *Memory and Cognition* **21**: 142–45

Estes W K 1972 An associative basis for coding and organization in memory. In: Melton A W, Martin E (eds.) *Coding Processes in Human Memory*. Winston, Washington, DC, pp. 161–90

Estes W K 1999 Models of human memory: A 30-year retrospective. In: Izawa C (ed.) *On Human Memory: Evolution, Progress, and Reflections on the 30th Anniversary of the Atkinson–Shiffrin Model*. Erlbaum, Mahwah, NJ, pp. 59–87

Healy A F, Cunningham T C 1999 Recall of order information: Evidence requiring a dual-storage memory model. In: Izawa C (ed.) *On Human Memory: Evolution, Progress, and Reflections on the 30th Anniversary of the Atkinson–Shiffrin Model*. Erlbaum, Mahwah, NJ, pp. 151–64

Healy A F, McNamara D S 1996 Verbal learning and memory: Does the modal model still work? *Annual Review of Psychology* **47**: 143–72

James W 1890 *The Principles of Psychology*. Holt, New York

Martin R C, Romani C 1994 Verbal working memory and sentence comprehension: A multiple-components view. *Neuropsychology* **8**: 506–23

Nairne J S 1990 A feature model of immediate memory. *Memory and Cognition* **18**: 251–69

Peterson L R, Peterson M J 1959 Short-term retention of individual verbal items. *Journal of Experimental Psychology* **58**: 193–8

Shiffrin R M 1999 30 years of memory. In: Izawa C (ed.) *On Human Memory: Evolution, Progress, and Reflections on the 30th Anniversary of the Atkinson–Shiffrin Model*. Erlbaum, Mahwah, NJ, pp. 17–33

A. F. Healy

Short-term Memory: Psychological and Neural Aspects

1. Short-term Memory

'Short-term memory' refers to a number of systems with limited capacity (in the verbal domain, roughly the 'magical' number 7 ± 2 items: Miller 1956) concerned with the temporary retention (in the range of seconds) of a variety of materials. Knowledge of the functional and anatomical organization of short-term memory in humans, and its role in cognition as at the turn of the twenty-first century, is herewith presented, drawing data from three main sources of evidence: (a)

behavioral studies in normal individuals and in brain-injured patients with selective neuropsychological short-term memory deficits; (b) correlations between the anatomical localization of the cerebral lesion and the short-term memory disorder of brain-damaged patients; (c) correlations between the activation of specific cerebral areas and the execution of tasks assessing short-term retention in normal subjects. The two more extensively investigated aspects of short-term memory are considered: verbal and visual/spatial. 'Short-term memory' is closely related to the concept of 'working memory' (see *Working Memory, Psychology of Working Memory, Neural Basis of*). The present article focuses on the 'storage' and 'rehearsal' components of the system, rather than on the cognitive operations and executive functions currently associated with 'working memory.' However, the section devoted to the uses of short-term memory illustrates some of the working aspects of short-term retention.

2. *Historical Origin of the Construct 'Short-term Memory'*

Suggestions of a distinction between two types of memory, one concerned with temporary retention, and the other having the function of a storehouse for materials which have been laid out of sight, date back at least to John Locke's '*Essay Concerning Human Understanding*' (1700). William James (1895) revived the distinction, suggesting the existence of a limited-capacity 'primary memory,' embracing the present and the immediate past, and subserving consciousness (see *Consciousness, Neural Basis of*). Psychological research in the nineteenth century and in the first half of the twentieth century was, however, mainly concerned with the diverse factors affecting learning and retention, in the context of a basically unitary view of human memory. It was only in the 1950s that short-term memory became the object of systematic behavioral studies in normal subjects (Baddeley 1976). In the late 1960s the division of human memory into a short- and a long-term system became a current view (Atkinson and Shiffrin 1968).

3. *Functional Architecture of Short-term Memory*

3.1 *Evidence from Normal Subjects*

Three main behavioral phenomena suggest the existence of a discrete limited-capacity system, concerned with short-term retention (Baddeley 1976, Baddeley 1986, Glanzer 1972). (a) Accuracy in recalling a short list of stimuli (e.g., trigrams of consonants or of words) decreases dramatically in a few seconds if the subjects' repetition of the memory material (rehearsal) is prevented by a distracting activity such as counting backwards by threes. (b) In the immediate free recall of a sequence of events, such as words, the final five to six stimuli on the list are recalled better that the preceding ones. This 'recency effect' vanishes after a few seconds of distracting activity, and is minimally affected by factors such as age, rate of presentation of the stimuli, and word frequency. (c) In the immediate serial recall of verbal material (memory span), the subject's performance is affected by factors such as phonological similarity and word length, with the effects of semantic factors being comparatively minor. Each of these phenomena subsequently proved to be considerably more complex than initially thought. They were also interpreted as compatible with a unitary, single-system, view of human memory. This account proved untenable, however, mainly on the basis of neuropsychological evidence.

These empirical observations illustrate the main characteristics of 'short-term memory': a retention system with limited capacity, where the memory trace, in the time range of seconds, shows a decay, which may be prevented through rehearsal. Material stored in short-term memory has a specific representational format, which, in the case of the extensively-investigated verbal domain, involves phonological codes, separately from lexical-semantic representations stored in long-term memory. The latter contribute, however, to immediate retention, e.g., in verbal span tasks.

The functional architecture of phonological short-term memory has been investigated in detail using effects which break down storage and rehearsal subcomponents. The effect of phonological similarity, whereby the immediate serial recall of auditory and visual verbal material is poorer for sequences of phonologically similar stimuli than for dissimilar ones, reflects the coding which takes place in the phonological short-term store. The effect of word length, whereby the immediate serial recall of auditory and visual verbal material is poorer for sequences of long words than for short ones, is held to reflect the activity of the process of rehearsal, abolished by 'articulatory suppression,' i.e., a continuous uttering of an irrelevant speech sound. Suppression, while disrupting rehearsal, also reduces immediate memory performance in span tasks. The interaction between phonological similarity, input modality and articulatory suppression, with suppression abolishing phonological similarity only when the stimuli are presented visually, suggests that rehearsal participates in the process of conveying visual verbal material to the phonological short-term store. Finally, some phonological judgments (such as rhyme and stress assignment) are held to involve the articulatory components of rehearsal, because the performance of normal subjects is selectively impaired by suppression (Burani et al. 1991).

In the nonverbal domain a similar distinction is drawn between visual and spatial short- and long-term memory systems, with the relevant representational

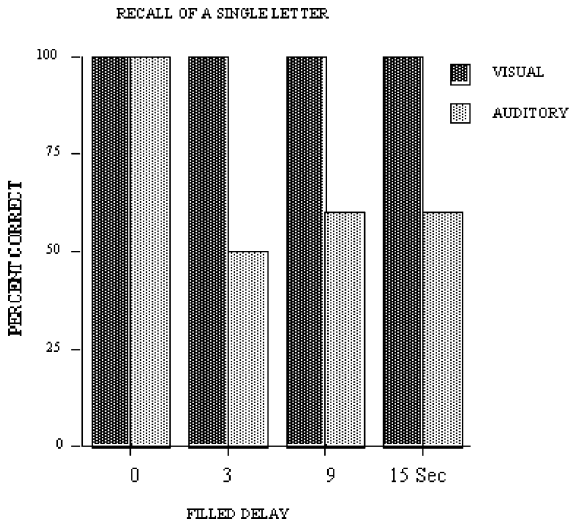

Figure 1
Short-term forgetting of a single letter by patient PV after 3 to 15 seconds of delay filled by arithmetic distracting activity. With an auditory input, dramatic forgetting occurred within a few seconds. In the visual modality, the patient's performance was at ceiling. Immediate recall was fully accurate in both input modalities, ruling out perceptual and response-related deficits, differently from the short-term memory disorder (Source: data from Basso A, Spinnler H, Vallar G, Zanobio M E 1982 Left hemisphere damage and selective impairment of auditory-verbal short-term memory. *Neuropsychologia* **20**: 263–274)

format being in terms of the shape or spatial location of the stimulus (Baddeley 1986, Della Sala and Logie 1993). Also in the visuo-spatial domain, similarity, recency and interference effects have been observed. Visuo-spatial short-term memory is likely to comprise storage and rehearsal (pictorial and spatial) components. In the case of spatial locations, rehearsal may be conceived in terms of planned movements (e.g., ocular, manual reaching, locomotion) towards a target coded in a spatial reference frame, (e.g., egocentric).

3.2 Evidence from Brain-injured Patients

Studies in patients with localized brain damage provide unequivocal evidence that supports the independence of short- and long-term memory systems, conjuring up a double dissociation of deficits. In patients with 'global amnesia,' which may be characterized as a selective impairment of the declarative or explicit component of long-term memory (see *Declarative Memory, Neural Basis of*; *Episodic and Autobiographical Memory: Psychological and Neural Aspects*) verbal and visuo-spatial short-term memory are unimpaired, with immediate serial span, the recency effect in immediate free recall, short-term forgetting being within normal limits. This functional dissociation suggests a serial organization of the two systems, with temporary retention being a necessary condition for long-term storage (Atkinson and Shiffrin 1968).

Since the late 1960s selective impairments of short-term memory have been reported (Vallar and Papagno 1995, Vallar and Shallice 1990). The more extensively investigated area concerns auditory-verbal (phonological) short-term memory. Patients show a disproportionate reduction of the auditory-verbal span to an average of less than three items (digits, letters, or words); the recency effect in immediate free recall of auditory-verbal lists of words is abolished; short-term forgetting is abnormally rapid. The disorder is modality-specific, with the level of performance being better when the material is presented visually (Fig. 1). This input-related dissociation has two main implications: (a) discrete phonological and visual short-term memory components exist; (b) in the input–output processing chain, the phonological short-term store should be envisaged as an input system, rather than an output buffer store. This division argues against a monolithic view of the system as a single store, which is amodal, i.e., not specific for the different sensory modalities (see Atkinson and Shiffrin 1968). Additional support to this input locus of the system comes from the observation that in some of these patients speech production is entirely preserved.

Patients with defective phonological short-term memory may show unimpaired long-term verbal and visuo-spatial learning (Fig. 2). This observation further corroborates the complete independence of short- and long-term memory systems, but is incompatible with a serial organization, in which defective short-term memory entails a long-term memory impairment (see Atkinson and Shiffrin 1968), suggesting a parallel architecture instead. After early perceptual analysis, information may enter short- or long-term memory, either of which may be selectively disrupted by brain damage. The learning abilities of these patients are, however, dramatically impaired when the phonological material to be learned does not possess pre-existing lexical-semantic representations in long-term memory. This is the case of pronounceable meaningless sequences of letters, such as non-words, or words of a language unknown to the subject (Fig. 2).This phonological learning deficit indicates that, within a specific representational domain, temporary retention in short-term memory is necessary for stable long-term learning and retention, as predicted by serial models.

Other brain-injured patients show deficits of short-term retention in the visual and spatial domain, even though these patterns of impairment have been explored less extensively than the phonological disorder. Immediate retention of short sequences of spatial

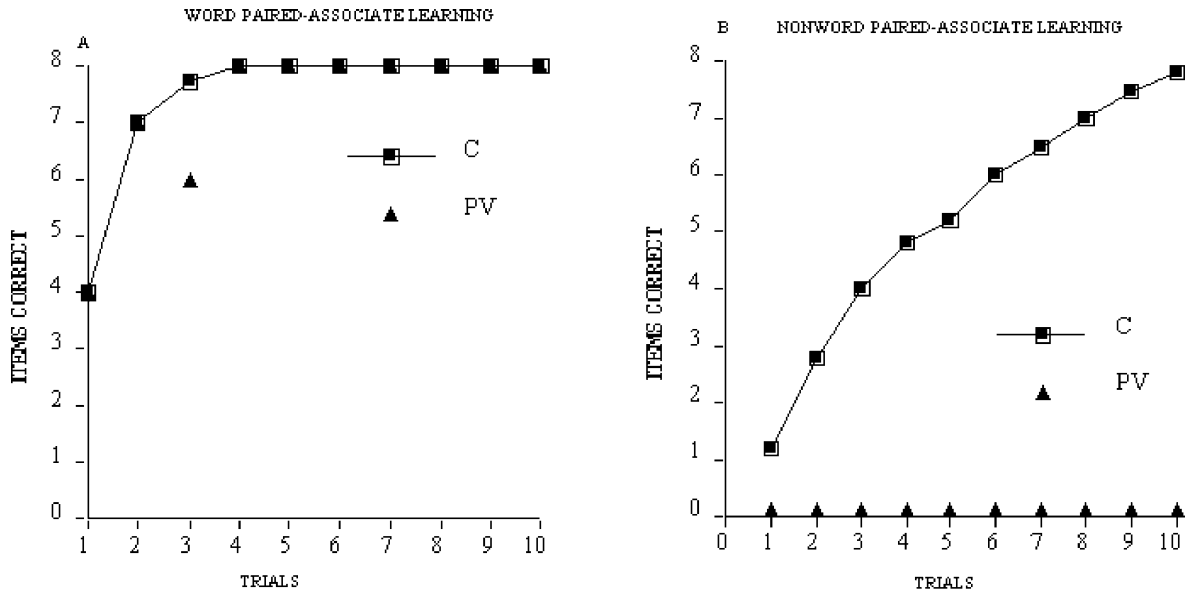

Figure 2
Word (A) and nonword (B) paired-associate learning by patient PV and matched control subjects (C), with auditory presentation. The patient's performance was within the normal range with native language (Italian) words, dramatically defective with unfamiliar nonwords (Russian words transliterated into Italian) (Source: Baddeley A D, Papagno C, Vallar G 1988 When long-term learning depends on short-term storage. *Journal of Memory and Language* **27**: 586–595)

locations, as assessed by a block tapping task (a spatial analogue of digit span), may be selectively defective, with verbal memory being unaffected, in both its short- and long-term components. In other patients, the deficit may involve the visual (shape) component of short-term memory, while impairments in the short-term visual recognition of unfamiliar faces, objects, voices, colours have been also described. The disorder of patients with defective visual imagery (as assessed, for instance, by tasks requiring colour or size comparisons) may be interpreted in terms of an impaired visual short-term memory store (see *Visual Imagery, Neural Basis of*). A deficit of visuo-spatial short-term memory may also disrupt long-term learning of unfamiliar non-verbal material, as assessed by recognition memory for unfamiliar faces and objects. This extends to the visuo-spatial domain the conclusion that long-term acquisition requires short-term storage.

The process of rehearsal has been traditionally described as an activity which, through rote repetition, refreshes the short-term memory trace, preventing its decay. The precise characteristics of rehearsal have been elucidated in more detail in recent years, particularly in the domain of phonological memory. Rehearsal may be conceived in terms of the recoding of the memory trace from input (auditory-verbal) to output-related (articulatory) representations and vice-versa. More specifically, rehearsal of verbal material has long been regarded as 'articulatory' in nature, involving output-related verbal processes—such as the motor programming of speech production in a phonological assembly system or output buffer—the actual articulation of the material to be rehearsed (subvocal rehearsal), or both.

Anarthric subjects, who are unable to utter any articulated speech sound due to congenital disorders or brain damage acquired in adult age, may, nonetheless, show a preserved immediate memory, including verbal rehearsal. This suggests that the process is 'central' and does not require the activity of the peripheral musculature. Brain-damaged patients with a selective impairment of rehearsal (Vallar et al. 1997) show a defective immediate verbal span, as do patients with damage to the phonological short-term store. Both components of phonological memory contribute to immediate retention, though the phonological store provides a major retention capacity. The ability to perform phonological judgments is disrupted, however, by damage to the rehearsal process, but not to the phonological store. By contrast, a defective rehearsal leaves the recency effect in immediate free recall of auditory lists of words largely unimpaired. This recency effect is largely reduced or absent in patients with damage to the phonological short-term store.

4. The Uses of Short-term Memory

Is there a use for a system providing the temporary retention of a limited amount of stimuli, besides infrequent situations such as the following? A friend tells us an unfamiliar eight-digit number, which we have to dial on a telephone placed on the other side of a large street and we have no paper and pencil to write it down. The answer is positive. Short-term retention contributes to the stable acquisition of new information in long-term memory. More specifically, phonological short-term memory plays an important role in learning new vocabulary and participates in the processes of speech comprehension and production (see *Speech Perception*; *Speech Production, Psychology of*).

4.1 Long-term Learning

The observation that patients with defective auditory-verbal span are also impaired in learning unfamiliar pronounceable letter sequences gives rise to the possibility that phonological memory may contribute to a relevant aspect of language development, the acquisition of vocabulary (Fig. 2). Similarly, subjects with a developmental deficit of phonological memory are impaired in vocabulary acquisition and in non-word learning. An opposite pattern is provided by subjects with a congenital cognitive impairment which selectively spares phonological short-term memory. Acquisition of vocabulary, foreign languages and non-word learning are also preserved. Converging evidence from different subject populations supports this view. Correlational studies in children have shown that the capacity of phonological memory is a main predictor of the subsequent acquisition of vocabulary, both in the native and in a second language. In normal adult subjects, the variables which disrupt immediate memory span (phonological similarity, item length, articulatory suppression) also impair the acquisition of non-words.

Polyglots have a greater capacity of phonological memory, compared to nonpolyglots, and a better ability to learn novel words. Phonological short-term memory may be considered as a learning device for the acquisition of novel phonological representations, and the building up of the phonological lexicon (Baddeley et al. 1998). A few observations in brain-damaged patients suggest a similar role for visuo-spatial short-term memory in the acquisition of new visual information, such as unfamiliar faces and objects.

4.2 Language Processing

The idea that short-term retention contributes to speech comprehension dates back to the 1960s. Phonological memory may withhold incoming auditory-verbal strings, while syntactic and lexical-semantic analyses are performed. Patients with defective phonological memory show a preserved comprehension of individual words, as well as many sentential materials, and a normal ability to decide whether or not sentences are grammatically correct. This may reflect, on the one hand, the operation of on-line lexical-semantic processes, heuristics and pragmatics, and, on the other, the complete independence of syntactic and lexical-semantic processes from phonological memory. Patients are however impaired by 'complex' sentences, where 'complexity' refers to a number of non-mutually-exclusive factors, such as: (a) a high speed of material presentation, which prevents the immediate build-up of an unambiguous cognitive representation; (b) word order conveying meaning-crucial information (e.g., in sentences in which a semantic anomaly is introduced by a change in the linear arrangement of words: 'The world divides the equator into two hemispheres, the southern and the northern'); (c) extralinguistic presuppositions biasing the interpretation of the spoken message.

Under such conditions, adequate interpretation may require backtracking to the verbatim (phonological) representation of the sentence, temporarily held in phonological memory. This provides a 'back-up' or 'mnemonic window' resource for performing supplementary cognitive operations necessary for comprehension (Vallar and Shallice 1990).

5. Neural Architecture of Short-term Memory

5.1 Phonological Short-term Memory

Anatomoclinical correlation studies in brain-damaged patients with a selective impairment of the auditory-verbal span indicate that the inferior parietal lobule (supramarginal gyrus) of the left hemisphere, at the temporoparietal junction, represents the main neural correlate of the 'store' component of phonological short-term memory (Vallar and Papagno 1995). The frontal premotor regions in the left hemisphere and other structures such as the insula are the major neural correlates of the 'rehearsal' component, even though the available anatomoclinical data are more limited (Vallar et al. 1997).

Functional neuroimaging studies in normal subjects concur with this pathological evidence, to suggest a left-hemisphere-based network. Activation in the left supramarginal gyrus [Brodmann's area (BA) 40] is associated with the 'store' component of short-term phonological memory, activation in the left frontal premotor BA 44 (Broca's area) and BA 6, and in the left insula, with the 'rehearsal' component (Paulesu et al. 1996). In these studies, in line with the behavioral evidence from normal subjects and patients, an immediate verbal span task activates both the inferior

parietal region (phonological short-term store) and the premotor cortex (rehearsal process) in the left hemisphere. Conversely, rhyme judgements selectively activate the left premotor regions, whose damage, in turn, disrupts the patients' ability to perform this task. These activation and lesion-based data support, from an anatomofunctional perspective, the behavioral distinction between a 'storage' component and a 'rehearsal process' in phonological short-term memory. Furthermore, they qualify 'rehearsal' as a process which makes use of components also concerned with the planning (i.e., programming in the left premotor cortex) of articulated speech. Seen in this perspective, phonological memory may be regarded as a component of the language system. Finally, connectionist modelling of this architecture is currently being developed (Burgess and Hitch 1999).

5.2 *Visual and Spatial Short-term Memory*

Studies in brain-damaged patients suggest an association between damage to the posterior regions of the right hemisphere and defective spatial short-term memory, as assessed by tasks requiring the reproduction of sequences of spatial locations. There is also evidence suggesting that damage to the posterior regions of the left hemisphere brings about disorders of visual short-term memory, such as defective immediate retention of sequences of visual stimuli, (e.g. lines), and impaired recognition of more than one visual stimulus at a time (defective simultaneous form perception) (Vallar and Papagno 1995).

Neuroimaging activation studies in humans support the distinction between primarily spatial (a location, a 'where' component) and visual short-term memory systems (Smith et al. 1995). The main neural correlates of spatial memory for locations include the occipital extra-striate, posterior–inferior parietal, dorsolateral premotor and prefrontal cortices. Short-term visual recognition memory is associated with activations in a network including the posterior–inferior parietal, temporal and ventral premotor and prefrontal cortices. Right hemisphere regions may play a more relevant role in spatial memory for locations, while left hemisphere regions are more involved in visual memory for objects. These patterns of activation indicate an association of the 'dorsal visual stream' with spatial short-term memory for locations and of the 'ventral visual stream' with short-term visual recognition memory.

6. *Conclusion*

Behavioural observations and neuroanatomical evidence from normal subjects and brain-injured patients concur in suggesting that 'short-term memory' should be conceived as a multiple-component system with specific functional properties and discrete neural correlates. These systems secure the retention of a limited amount of material in the time range of seconds and contribute to relevant aspects of cognition, such as long-term learning.

See also: Declarative Memory, Neural Basis of; Episodic and Autobiographical Memory: Psychological and Neural Aspects; Learning and Memory, Neural Basis ofMemory Retrieval; Memory: Synaptic Mechanisms; Recognition Memory (in Primates), Neural Basis of; Short-term Memory, Cognitive Psychology of; Working Memory, Neural Basis of

Bibliography

Atkinson R C, Shiffrin R M 1968 Human memory: a proposed system and its control processes. In: Spence K W, Taylor Spence J (eds.) *The Psychology of Learning and Motivation. Advances in Research and Theory*. Academic Press, New York

Baddeley A D 1976 *The Psychology of Memory*. Basic Books, New York

Baddeley A D 1986 *Working Memory*. Clarendon Press, Oxford, UK

Baddeley A D, Gathercole S, Papagno C 1998 The phonological loop as a language learning device. *Psychological Review*. **105**: 158–73

Burani C, Vallar G, Bottini G 1991 Articulatory coding and phonological judgements on written words and pictures: the role of the phonological output buffer. *European Journal of Cognitive Psychology* **3**: 379–98

Burgess N, Hitch G J 1999 Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review* **106**: 551–81

Della Sala S, Logie R H 1993 When working memory does not work: the role of working memory in neuropsychology. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*. Elsevier, Amsterdam

Glanzer M 1972 Storage mechanisms in recall. In: Bower G H (ed.) *The Psychology of Learning and Motivation. Advances in Research and Theory*. Academic Press, New York

James W 1895 *The Principles of Psychology*. Holt, New York

Miller G A 1956 The magic number seven, plus or minus two: some limits to our capacity for processing information. *Psychological Review* **63**: 81–93

Paulesu E, Frith U, Snowling M, Gallagher A, Morton J, Frackowiak R S J, Frith C D 1996 Is developmental dyslexia a disconnection syndrome? Evidence from PET scanning. *Brain* **119**: 143–57

Smith E E, Jonides J, Koeppe R A, Awh E, Schumacher E H, Minoshima S 1995 Spatial versus object working memory: PET investigations. *Journal of Cognitive Neuroscience* **7**: 337–56

Vallar G, Di Betta A M, Silveri M C 1997 The phonological short-term store-rehearsal system: Patterns of impairment and neural correlates. *Neuropsychologia* **35**: 795–812

Vallar G, Papagno C 1995 Neuropsychological impairments of short-term memory. In: Baddeley A D, Wilson B A, Watts F (eds.) *Handbook of Memory Disorders*. Wiley, Chichester, UK

Vallar G, Shallice T (eds.) 1990 *Neuropsychological Impairments of Short-term Memory*. Cambridge University Press, Cambridge, UK

G. Vallar

Shyness and Behavioral Inhibition

Since the 1980s, the study of shyness, behavioral inhibition, and social withdrawal in childhood has taken on a research trajectory that can best be described as voluminous. Yet, these related phenomena remain something of a mystery, carrying with them a variety of definitions and a number of very different perspectives concerning psychological significance. Because there appear to be several different characterizations of forms of shy, inhibited, or nonsocial behavior, the first goal of this article is to provide a definitional framework.

1. *Shyness, Behavioral Inhibition, and Social Withdrawal: Definitions*

In their efforts to identify the etiology of children's personalities and social behaviors, developmental scientists have attempted to determine the relevant dispositional dimensions of temperament that may underlie children's actions that are displayed consistently across situations, and continuously over time. One such dispositional construct is *behavioral inhibition* (Kagan 1997). Behavioral inhibition has been defined variously as (a) an inborn bias to respond to unfamiliar events by showing anxiety; (b) a specific vulnerability to the uncertainty all children feel when encountering unfamiliar events that cannot be assimilated easily; and (c) one end of a continuum of possible initial behavioral reactions to unfamiliar objects or challenging social situations. These definitions highlight some common elements: behavioral inhibition is (a) a pattern of responding or behaving, (b) possibly biologically determined, such that (c) when unfamiliar and/or challenging situations are encountered, (d) the child shows signs of anxiety, distress, or disorganization.

The term *shyness* has been used to refer to inhibition in response to novel *social* situations. In infancy and early childhood, shyness is elicited by feelings of anxiety and distress when confronted by unfamiliar people. To some, such behavior serves an adaptive purpose in that it removes children from situations they perceive as discomforting and dangerous (Buss 1986). But variability of the infantile shyness response is great; some children clearly demonstrate the behavior excessively.

In middle childhood, children have the cognitive skill to compare themselves with others and to understand that others can, and do, pass judgment on them. It is this understanding of social comparison and evaluation that can elicit shy behavior among older children, adolescents and adults (Buss 1986). That is, children may remove themselves from situations they believe will be discomforting because others will pass judgment on their skills, personae, and so on. Again, variability is great; some children display shyness that is clearly out-of-the ordinary.

A related construct is *social withdrawal*, a phenomenon that refers to the consistent (across situations and over time) display of solitary behavior when encountering *both* familiar and/or unfamiliar peers (Rubin and Stewart 1996). Thus, shyness is embedded within the construct of social withdrawal. The common denominator for inhibition, shyness, and social withdrawal is that the representative behavior is one that moves children *away* from their social worlds, that is, *solitude*.

2. *Different Forms of Inhibition, Shyness, and Solitude*

Inhibited behavior in the face of the unfamiliar typically has been assessed by observing very young children as they are confronted by unfamiliar adults, objects, or peers. Insofar as shyness is concerned, it is best exemplified behaviorally by the display of socially *reticent* behavior during the preschool and early childhood periods. Among unfamiliar others, reticent preschoolers hang back, watching peers from afar; they make few social initiations and back away from initiations made by others. When asked to speak up in groups of unfamiliar peers, reticent children choose not to do so at all, and if they do speak, it is uncomfortably and for a very short period of time. Further, when asked to collaborate with peers to solve problems, they spend their time watching others rather than providing the requested help.

Reticence is one of several forms of solitude demonstrated by children (Rubin and Asendorpf 1993). Some preschoolers appear to be satisfied playing with objects rather than people. These children play alone, constructing and exploring objects. The term used to describe such behavior is *solitary-passivity* (Coplan et al. 1994). When others initiate interaction with solitary-passive children, they do not back away; further, they participate fully and actively in-group problem-solving behavior. Thus unlike shy children who have a high social avoidance motivation, those whose solitary behavior is mainly of the solitary-passive ilk may be described as having a low social approach motivation.

Still yet other children demonstrate solitude that appears to reflect immaturity. These young children engage in solitary-sensorimotor activity (*solitary-active play*), repeating motor movements with or without objects (Coplan et al. 1994).

2.1 Developmental Differences and Change

Typically all forms of solitude decrease from early through middle and late childhood. As noted above, causes of shy behavior change from early wariness to the anxiety associated with being socially evaluated. While reticent behavior tends to decrease with age, it is also the case that the 'meanings' of different forms of solitude change as well. For example, as children come to cope with their social anxieties among unfamiliar peers, they become increasingly likely to display that form of solitude that had been viewed earlier as normal and adaptive—*solitary-exploration* and *construction*. Thus, with increasing age, such constructive solitude becomes increasingly associated with measures of social wariness, and physiological markers of anxiety and emotion dysregulation (see below).

3. Putative Causes of Inhibition, Shyness, and Social Withdrawal

3.1 Physiology

Behavioral inhibition has been thought to emanate from a physiological 'hard wiring' that evokes caution, wariness, and timidity in unfamiliar social and nonsocial situations. Inhibited infants and toddlers differ from their uninhibited counterparts in ways that imply variability in the threshold of excitability of the amygdala and its projections to the cortex, hypothalamus, sympathetic nervous system, corpus striatum, and central gray. That there is a physiological basis underpinning behavioral inhibition is drawn from numerous psychophysiological studies. For example, stable patterns of right frontal EEG asymmetries in infancy predict temperamental fearfulness and behavioral inhibition in early childhood (Calkins et al. 1996). The functional role of hemispheric asymmetries in the regulation of emotion may be understood in terms of an underlying motivational basis for emotional behavior, specifically along the *approach–withdrawal continuum*. Infants exhibiting greater relative right frontal asymmetry are more likely to withdraw from mild stress, whereas infants exhibiting the opposite pattern of activation are more likely to approach.

Another physiological entity that distinguishes wary from nonwary infants and toddlers is vagal tone, an index of the functional status or efficiency of the nervous system, marking both general reactivity and the ability to regulate one's level of arousal. Reliable associations have been found between vagal tone and inhibition in infants and toddlers: children with *lower vagal tone* (consistently high heart rate due to less parasympathetic influence) tend to be more behaviorally inhibited (Kagan et al. 1987).

In early childhood, *reticent* behavior is associated with the same physiological markers as is the case for behavioral inhibition in infancy and toddlerhood. Thus, in early childhood, reticent, fearful, solitary behavior is associated with greater relative right frontal EEG activation; but constructive solitude is not (Fox et al. 1996). Further, parents view children who have such EEG asymmetries as anxious. Among older, elementary school-age children, shy, reticent behavior among *familiar* peers (i.e., *social withdrawal*) has been associated positively with internalized negative emotions such as nervousness, distress, and upset; and negatively related to positive emotions such as enthusiasm and excitement.

3.2 Parent–Child Relationships

Attachment theorists maintain that the primary relationship develops during the first year of life, usually between the mother and the infant (Bowlby 1973). Maternal sensitivity and responsiveness influence whether the relationship will be secure or insecure. Researchers have shown that securely attached infants are likely to be well adjusted, socially competent, and successful at forming peer relationships in early and middle childhood whereas insecurely attached children may be less successful at social developmental tasks.

Researchers have proposed that those infants who are temperamentally reactive and who receive insensitive, unresponsive parenting come to develop an insecure-ambivalent ('C'-type) attachment relationship with their primary caregivers (e.g., Calkins and Fox 1992). In novel settings these 'C' babies maintain close proximity to the attachment figure (usually the mother). When the mother leaves the room briefly, these infants become quite unsettled. Upon reunion with the mother, these infants show angry, resistant behaviors interspersed with proximity—or contact-seeking behaviors. It is argued further that this constellation of infant emotional hyperarousability and insecure attachment may lead the child to display inhibited/wary behaviors as a toddler, and there are data supportive of this conjecture.

Given that the social behaviors of preschoolers and toddlers who have an insecure 'C'-type attachment history are thought to be guided largely by fear of rejection, it is unsurprising to find that when these children are observed in peer group settings, they appear to avoid rejection by demonstrating passive, adult-dependent behavior and withdrawal from social interaction. Lastly, 'C' babies lack confidence and assertiveness at age 4 years; then, at age 7 years they are seen as passively withdrawn (Renken et al. 1989).

3.3 Parenting

Recently, researchers have shown that parental influence and control maintain and exacerbate child

inhibition and social withdrawal. For example, mothers of extremely *inhibited* toddlers have been observed to display overly solicitous behaviors (i.e., intrusively controlling, unresponsive, physically affectionate). Mothers of *shy* preschoolers do not encourage independence and exploration. And mothers of *socially withdrawn* children tend to be highly controlling, overprotective, and poor reciprocators of their child's displays of positive behavior and positive affect. Lastly, researchers have shown that mothers of socially withdrawn children are more likely than those of normal children to use such forms of psychological control statements, as devaluation, criticism, and disapproval. Taken together, these parenting practices may attack the child's sense of self-worth (for a review, see Rubin et al. 1995).

4. *Correlates and Consequences of Inhibition, Shyness, and Social Withdrawal*

Researchers who have followed, longitudinally, the course of development for inhibited infants have found strong consistency of behavior over time. As a group, children identified as extremely inhibited are more likely to be socially wary with unfamiliar peers in both the laboratory and at school, and to exhibit signs of physiological stress during social interactions (Kagan et al. 1987). In a longitudinal study extending into adulthood, Caspi and Silva (1995) found that individuals identified as shy, fearful, and withdrawn at 3 years reported that they preferred to stick with safe activities, be cautious, submissive, and had little desire to influence others at 18 years. A subsequent follow-up at age 21 on interpersonal functioning showed that these same children were normally adjusted in both their work settings and their romantic relationships.

Social withdrawal appears to carry with it the risk of a child's developing negative thoughts and feelings about the *self*. Highlighting the potential long-term outcomes of social withdrawal is a recent report which showed that social withdrawal among familiar peers in school at age 7 years predicted negative self-perceived social competence, low self-worth, loneliness, and felt peer-group insecurity among adolescents aged 14 years (Rubin et al. 1995). These latter findings are augmented by related research findings. For example, Renshaw and Brown (1993) found that passive withdrawal at age 9 to12 years predicted loneliness assessed one year later. Ollendick et al. (1990) reported that 10-year-old socially withdrawn children were more likely to be perceived by peers as withdrawn and anxious, more disliked by peers, and more likely to have dropped out of school than their well-adjusted counterparts five years later. Finally, Morison and Masten (1991) indicated that children perceived by peers as withdrawn and isolated in middle childhood were more likely to think negatively of their social competencies and relationships in adolescence. In sum, it would appear as if early social withdrawal, or its relation to anxiety, represents a behavioral marker for psychological and interpersonal maladaptation in childhood and adolescence.

5. *Summary and Future Directions*

The study of inhibition and shyness garnered an enormous amount of attention in the 1990s. Most empirical research has focused on the contemporaneous and predictive correlates of social reticence, shyness, and withdrawal at different points in childhood and adolescence. These correlated variables include those of the biological, intrapersonal, interpersonal, and psychopathology ilk that have been chosen from conceptual frameworks pertaining to the etiology, stability, and outcomes of socially wary and withdrawn behaviors. Thus far, it appears that socially inhibited children have a biological disposition that fosters emotional dysregulation in the company of others. These children, if overly directed and protected by their primary caregiver, become reticent and withdrawn in the peer group. In turn, such behavior precludes the development of social skills and the initiation and maintenance of positive peer relationships. Yet again, this transactional experience seems to lead children to develop anxiety, loneliness, and negative self-perceptions of their relationships and social skills.

Despite these strong conclusions, however, it is important to recognize that the data bases upon which these conclusions rest are relatively few. Clearly, replication work is necessary. The extent to which dispositional factors interact with parenting styles and parent–child relationships to predict the consistent display of socially withdrawn behavior in familiar peer contexts still needs to be established. Further, the sex differences discussed above require additional attention.

Lastly, our knowledge of the developmental course of inhibition, shyness, and social withdrawal is constrained by the almost sole reliance on data gathered in Western cultures. Little is known about the developmental course of these phenomena in Eastern cultures such as those in China, Japan, or India; and even less is known in Southern cultures such as those found in South America, Africa, and southern Europe. It may well be that depending on the culture within which these phenomena are studied, the biological, interpersonal, and intrapersonal causes, concomitants, and consequences of inhibition, shyness, and social withdrawal may vary. In short, cross-cultural research is necessary, not only for the study of these phenomena, but also for most behaviors that are viewed as deviant or reflective of intrapsychic abnormalities in the West.

See also: Attachment Theory: Psychological; Emotional Inhibition and Health; Personality and Social Behavior; Personality Development and Temperament; Temperament and Human Development

K. H. Rubin

Bibliography

Bowlby J 1973 *Attachment and Loss. Attachment*. Basic Books, New York, Vol. 1

Buss A H 1986 A theory of shyness. In: Jones W H, Cheek J M, Briggs S R (eds.) *Shyness: Perspectives on Research and Treatment*. Plenum, New York

Calkins S D, Fox N A 1992 The relations among infant temperament, security of attachment and behavioral inhibition at 24 months. *Child Development* **63**: 1456–72

Calkins S D, Fox N A, Marshall T R 1996 Behavioral and physiological antecedents of inhibited and uninhibited behavior. *Child Development* **67**: 523–40

Caspi A, Silva P A 1995 Temperamental qualities at age three predict personality traits in young adulthood: Longitudinal evidence from a birth cohort. *Child Development* **66**: 486–98

Coplan R J, Rubin K H, Fox N A, Calkins S D, Stewart S L 1994 Being alone, playing alone, and acting alone: Distinguishing among reticence, and passive- and active-solitude in young children. *Child Development* **65**: 129–37

Fox N A, Schmidt L A, Calkins S D, Rubin K H, Coplan R J 1996 The role of frontal activation in the regulation and dysregulation of social behavior during the preschool years. *Development and Psychopathology* **8**: 89–102

Kagan J 1997 Temperament and the reactions to unfamiliarity. *Child Development* **68**: 139–43

Kagan J, Reznick J S, Snidman N 1987 The physiology and psychology of behavioral inhibition in children. *Child Development* **58**: 1459–73

Morison P, Masten A S 1991 Peer reputation in middle childhood as a predictor of adaptation in adolescence: A seven-year follow-up. *Child Development* **62**: 991–1007

Ollendick T H, Greene R W, Weist M D, Oswald D P 1990 The predictive validity of teacher nominations: A five-year follow-up of at risk youth. *Journal of Abnormal Child Psychology* **18**: 699–713

Renken B, Egeland B, Marvinney D, Mangelsdorf S, Sroufe L 1989 Early childhood antecedents of aggression and passive-withdrawal in early elementary school. *Journal of Personality* **57**: 257–81

Renshaw P D, Brown P J 1993 Loneliness in middle childhood: Concurrent and longitudinal predictors. *Child Development* **64**: 1271–84

Rubin K H, Asendorpf J 1993 Social withdrawal, inhibition and shyness in childhood: Conceptual and definitional issues. In: Rubin K H, Asendorpf J (eds.) *Social Withdrawal, Inhibition, and Shyness in Childhood*. L. Erlbaum Associates, Hillsdale, NJ

Rubin K H, Chen X, McDougall P, Bowker A, McKinnon J 1995 The Waterloo Longitudinal Project: Predicting internalizing and externalizing problems in adolescence. *Development and Psychopathology* **7**: 751–64

Rubin K H, Stewart S L 1996 Social withdrawal. In: Mash E J, Barkley R A (eds.) *Child Psychopathology*. Guilford, New York, pp. 277–307

Rubin K H, Stewart S, Chen X 1995 Parents of aggressive and withdrawn children. In: Bornstein M H (ed.) *Handbook of Parenting: Children and Parenting*. L. Erlbaum Associates, Mahwah, NJ, Vol. 1, pp. 225–84

Sibling-order Effects

Historically, sibling order has influenced important aspects of social, economic, and political life, and it continues to do so today in many traditional societies. Discriminatory inheritance laws and customs about royal succession that favor firstborns and eldest sons are just two examples. Sibling-order effects have also been documented for a wide variety of behavioral tendencies, although the magnitude and causal interpretation of these effects have been subject to debate. In the Darwinian process of competing for parental favor, siblings often employ strategies that are shaped by their order of birth within the family, and these strategies exert a lasting impact on personality. Radical revolutions are particularly likely to elicit differences in support by birth order. These behavioral patterns appear to be mediated by differences in personality as well as by differing degrees of identification with the family.

1. Social, Economic, and Political Consequences

Many societies—especially in past centuries and in non-Western parts of the world—have engaged in practices that favor one sibling position over another. For example, most traditional societies permit infanticide, especially when a child has a birth defect or when an immediately older infant is still breastfeeding. However, no society condones the killing of the elder of two siblings (Daly and Wilson 1988). An extensive survey of birth order and its relationship to social behavior in 39 non-Western societies found that the birth of a first child typically increases the status of parents and stabilizes the marriage. In addition, firstborns were generally found to receive more elaborate birth ceremonies, were given special privileges, and tended to exert authority over their younger siblings. In most of these 39 societies, firstborns maintained supremacy over their younger siblings even in adulthood. They also gained control of a greater share of parental property than laterborns, were more likely to become head of their local kin group, and tended to exert greater power in relationships with nonfamily members (Rosenblatt and Skoogberg 1974).

Previously, many Western societies employed sibling order as a means of deciding who inherits parental property or assumes political power. Primogeniture (the policy by which the eldest child or son automatically inherits property or political authority) was the most commonly employed mechanism, but other discriminatory inheritance practices have also been employed. For example, secondogeniture involves leaving the bulk of parental property to the second child or second son, and ultimogeniture involves leaving such property to the lastborn or youngest son.

Most variations in inheritance practices by sibling order can be understood by considering the advantages that accrue from such policies, given local economic circumstances (Hrdy and Judge 1993). For example, primogeniture has generally been practiced in societies where wealth is stable and based on limited land, and where talent does not matter much. Leaving the bulk of parental property to a single offspring avoids subdividing estates and hence reducing the social status of the family patronymic—something that was particularly important in past centuries among the landed aristocracy. However, in Renaissance Venice economic fortunes were generally based on speculative commerce rather than ownership of property, and parents typically subdivided their estates equally so as to maximize the chances of having one or more commercially successful offspring (Herlihy 1977). Ultimogeniture is a policy often found in societies that impose high death taxes on property. This inheritance practice has the consequence of maximizing the interval between episodes of taxation, thus reducing the overall tax burden.

Even in societies employing inheritance systems that favor one sibling over others, parents have commonly provided more or less equally for their offspring by requiring the child who inherits the family estate to pay a compensatory amount to each sibling. Primogeniture and related practices did not always mean disinheritance—a common misassumption. In medieval and early modern times, however, primogeniture among the landed aristocracy did mean that some younger sons and daughters faced difficult economic and social prospects. Among the medieval Portuguese nobility, for example, landless younger sons and daughters were significantly less likely to marry and leave surviving offspring (Boone 1986). Younger sons, for example, left 1.6 fewer children than did eldest sons. Younger sons were also nine times more likely than eldest sons to father a child out of wedlock. Because they posed a serious threat to political stability in their own country, younger sons were channeled by the state into expansionist military campaigns in faraway places such as India, where they often died in battle or from disease. Similarly, the Crusades can be seen, in part, as a church-state response to this domestic danger (Duby 1977). The surplus of landless younger daughters in the titled nobility was dealt with by sending them to nunneries.

2. *Personality*

Birth-order differences have long been claimed in the domain of personality, although these claims have remained controversial despite considerable research on this topic. Psychologists have investigated the consequences of birth order ever since Charles Darwin's cousin Francis Galton (1874) reported that eldest sons were overrepresented as members of the Royal Society. After breaking away from Sigmund Freud in 1910 to found a variant school of psychoanalysis, Alfred Adler (1927) focused on birth order in his own attempt to emphasize the importance of social factors in personality development. A secondborn, Adler considered firstborns to be 'power-hungry conservatives.' He characterized middleborns as competitive and lastborns as spoiled and lazy.

Since Adler speculated about birth order and its consequences for personality in 1927, psychologists have conducted more than 2000 studies on the subject. Critics of this extensive literature have argued that most of these studies are inadequately controlled for key covariates, such as social class and sibship size; that studies often conflict; and that birth-order differences in personality and IQ seem to have been overrated (Ernst and Angst 1983). Meta-analysis—a technique for aggregating findings in order to increase statistical power and reliability—suggests a different conclusion. If we consider only those well-designed studies controlling for important background variables that covary with birth order and can introduce spurious cross-correlations, a meta-analytic review reveals consistent birth-order differences for a wide range of personality traits (Sulloway 1995). For instance, firstborns are generally found to be more conscientious than laterborns, a difference that is exemplified by their being more responsible, ambitious, and self-disciplined. In addition, firstborns tend to be more conforming to authority and respectful of parents. Firstborns also tend to have higher IQs than their younger siblings—a reduction of about one IQ point is observed, on average, with each increment in birth rank (Zajonc and Mullally 1997).

These and other birth-order differences in personality can be usefully understood from a Darwinian perspective on family life (Sulloway 1996). Because they share, on average, only half of their genes, siblings will tend to compete unless the benefits of cooperating are greater than twice the costs (Hamilton 1964). In this Darwinian story about sibling competition, birth order does not exert a direct biological influence on personality. For example, there are no genes for being a firstborn or a laterborn. Instead, birth order is best seen as a proxy for various environmental influences, particularly disparities in age, physical size, power, and status within the family. These physical and mental differences lead siblings to pursue alternative strategies in their efforts to maximize parental investment (which includes emotional as well as physical

resources). This perspective on how sibling order shapes personality accords with research in behavioral genetics, which finds that most environmental influences on personality are not shared by siblings and hence belong to the 'nonshared environment' (Plomin and Daniels 1987).

Prior to the twentieth century, half of all children did not survive childhood. Even minor differences in parental favor would have increased an offspring's chances of getting out of childhood alive. Because eldest children have already survived the perilous early years of childhood, they are more likely than younger siblings to reach the age of reproduction and to pass on their parents' genes. Quite simply, a surviving eldest child was generally a better Darwinian bet than a younger child, which is why parents in most traditional societies tend to bias investment, consciously or unconsciously, toward older offspring. An exception to this generalization involves youngest children born toward the end of a woman's reproductive career. These children also tend to be favored, since they cannot be replaced (Salmon and Daly 1998).

Even when parents do not favor one offspring over another, siblings compete to prevent favoritism. Siblings do so, in part, by cultivating family niches that correspond closely with differences in birth order. Firstborns, for example, tend to act as surrogate parents toward their younger siblings, which is a good way of currying parental favor. As a consequence, firstborns tend to be more conservative and parent-identified than their younger siblings, which is one of the most robustly documented findings in the literature. Laterborns cannot compete as effectively as firstborns for this surrogate parent niche, since they cannot baby-sit themselves. As a consequence, laterborns seek alternative family niches that will help them to cultivate parental favor in different ways. To do so, they must often look within themselves for latent talents that can only be discovered through systematic experimentation. Toward this end, laterborns tend to be more open to experience—that is, more imaginative, prone to fantasies, and unconventional—propensities that offer an increased prospect of finding a valued and unoccupied family niche.

In addition to their efforts to cultivate alternative family niches, siblings diverge in personality and interests because they employ differing strategies in dealing with one another. These strategies are similar to those observed in mammalian dominance hierarchies. Because firstborns are physically bigger than their younger siblings, they are more likely to employ physical aggression and intimidation in dealing with rivals. Firstborns are the 'alpha males' of their sibling group, and they generally boss and dominate their younger brothers and sisters. For their own part, laterborns tend to employ low-power strategies to obtain what they want, including pleading, whining, cajoling, humor, social intelligence, and, whenever expedient, appealing to parents for assistance. Laterborns also tend to form coalitions with one another in an effort to circumvent the physical advantages enjoyed by the firstborn. Middle children are the most inclined to employ diplomatic and cooperative strategies. Some middle children are particularly adept at nonviolent methods of protest. Martin Luther King, Jr., the middle of three children, began his career as a champion of nonviolent reform by interceding in the episodes of merciless teasing that his younger brother inflicted upon their elder sister.

Only children represent a controlled experiment in birth-order research. They have no siblings and therefore experience no sibling rivalry. As a consequence, they are not driven to occupy a particular family niche. Although only children, like other firstborns, are generally ambitious and conform to parental authority, they are intermediate between firstborns and laterborns on most other personality traits. Because age spacing can affect functional sibling order, a firstborn whose next younger sibling is six or more years younger is effectively like an only child. Similarly, large age gaps can make some laterborns into functional firstborns or only children.

That brothers and sisters differ from one another for strategic reasons, rather than randomly, has been shown by studies involving more than one sibling from the same family (Schachter 1982). In one study, firstborns were found to be significantly different in personality and interests from secondborns, who were significantly different from thirdborns. By contrast, the first and third siblings were not as different as adjacent pairs, presumably because of less competition. This process of sibling differentiation has been termed 'deidentification' and extends to relationships with parents. For example, a firstborn child sometimes identifies more strongly with one parent than another. In such cases, the secondborn tends to identify more strongly with the parent that is not already preferred by the firstborn.

The most compelling evidence for birth-order effects in personality comes from studies in which siblings assess each other's personalities (Paulhus et al. 1999, Sulloway 1999). Such within-family designs control for the kinds of spurious correlations that can result from comparing individuals from different family backgrounds. Studies based on such direct sibling comparisons exhibit consistent birth-order effects in the expected direction. When the results are organized according to the Five-Factor Model of personality (Costa and McCrae 1992), firstborns tend to be more conscientious and slightly more neurotic than laterborns, whereas laterborns tend to be more agreeable, extraverted, and open to experience than firstborns.

Although reasonably consistent patterns for birth order and personality are observed according to the Five-Factor Model of personality, findings are significantly heterogeneous for three of the five personality dimensions. For example, firstborns tend to be more assertive than laterborns, which is an in-

dicator of extraversion, but laterborns tend to score higher on most other facets of extraversion, which include being more fun-loving, sociable, and excitement seeking. Similarly, laterborns tend to be more open to experience in the sense of being unconventional, whereas firstborns tend to be more open to experience in the sense of being intellectually oriented. Lastly, firstborns are more neurotic in the sense of being anxious about status, but laterborns are more neurotic in the sense of being more self-conscious.

As measured by direct sibling comparisons within the family, birth-order differences explain about four percent of the variance in personality, less than does gender and substantially more than do age, family size, or social class. In studies controlled for differences in age, sex, social class, and sibship size, siblings are about twice as likely to exhibit traits that are consistent with their sibling positions as to exhibit inconsistent traits. In short, not every laterborn has a laterborn personality (just as some firstborns deviate from the expected trend), but a reasonably consonant pattern is nevertheless present.

One still-unresolved question about birth order and personality is the extent to which within-family patterns of behavior transfer to behavior outside the family of origin. Recent studies suggest that birth-order effects observed in extrafamilial contexts are about one-third to one-half the magnitude of those manifested within the family (Sulloway 1999). Relative to firstborn spouses and roommates, for example, laterborn spouses and roommates are generally perceived to be more agreeable and extraverted, but less conscientious and neurotic. Outside the family of origin, birth-order effects seem to manifest themselves most strongly in intimate living relationships and in dominance hierarchies. These findings are not surprising, since they involve the kind of behavioral contexts in which sibling strategies were originally learned.

3. *Behavior During Radical Revolutions*

Some of the best evidence for sibling differences in behavior outside the family of origin comes from social and intellectual history. During socially radical revolutions, laterborns have generally been more likely than firstborns to question the status quo and to adopt a revolutionary alternative. This difference is observed even after controlling for the fact that laterborns, in past centuries at least, appear to have been more socially liberal than firstborns. As a rule, these effects are very context sensitive. During the Protestant Reformation, for example, laterborns were nine times more likely than firstborns to suffer martyrdom in support of the Reformed faith. (This statistic is corrected for the greater number of laterborns in the general population.) In countries that turned Protestant, such as England, firstborns were five times more likely than laterborns to suffer martyrdom for their refusal to abandon their allegiance to the Catholic faith. Thus laterborns gave their lives in the service of radical rebellion, whereas firstborns did so in an effort to preserve the waning orthodoxy. There was no generalized tendency, however, for laterborns to become 'martyrs' (Sulloway 1996).

Similarly, during the French Revolution firstborn deputies to the National Convention generally voted according to the expectations of their social class, whereas laterborn deputies were more likely to vote in the opposite manner. For this reason, there was only a modest tendency for firstborns and laterborns to vote in a particular manner, since voting was generally influenced by other important considerations that determined what it meant to conform or rebel in a rapidly changing political environment.

In response to radical conceptual changes in science one finds similar differences by birth order. Firstborns tend to be the guardians of what Thomas Kuhn (1970) has called 'normal' science, during which research is guided by the prevailing paradigms of the day. By contrast, laterborns tend to be the outlaws of science, sometimes flaunting its accepted methods and assumptions in the process of attempting radical conceptual change. During the early years of the Copernican revolution, which challenged church doctrine by claiming that the earth rotates around the sun, laterborns were five times more likely than firstborns to endorse this heretical theory. Copernicus himself was the youngest of four children. When Charles Darwin, the fifth of six children, became an evolutionist in the 1830s, he was 10 times more likely to do so than a firstborn. During other notable revolutions in science—including those led by laterborns such as Bacon, Descartes, Hutton, Semmelweis, and Heisenberg, and by occasional firstborns such as Newton and Einstein—younger siblings have typically been twice as likely as firstborns to endorse the new and radical viewpoint. Conversely, when new scientific doctrines, such as vitalism and eugenics, have appealed strongly to social conservatives, firstborns have been more likely than laterborns to pioneer and endorse such novel but ideologically conservative ideas. These birth-order effects typically fade over the course of scientific progress, as conceptual innovations accumulate supporting evidence and eventually convince the scientific majority.

In spite of their predilection for supporting radical innovations, laterborns have no monopoly on scientific truth. They were nine times more likely than firstborns to support Franz Joseph Gall's bogus theory of phrenology, which held that character and personality could be read by means of bumps on the head. In their willingness to question the status quo, laterborns run the risk of error through over-eager rebellion, just as firstborns sometimes err by resisting valid conceptual changes until the mounting evidence can no longer be denied.

During 'normal' science, firstborns possess a slight advantage over laterborns. Being more academically successful, firstborns are more likely than laterborns to become scientists. They also tend to win more Nobel prizes. This finding might seem surprising, but, according to the terms of Nobel's will, Nobel prizes have generally been given for 'discoveries' or creative puzzle solving in science, not for radical conceptual revolutions. Firstborn scientists who have innovated within the system include James Watson and Francis Crick, who together unraveled the structure of DNA, and Jonas Salk, who developed the polio vaccine. Such historical facts highlight another important point, namely, that firstborns and laterborns do not differ in overall levels of 'creativity.' Rather, firstborns and laterborns are preadapted to solving dissimilar kinds of problems by employing disparate kinds of creative strategies.

4. Family Sentiments

The kinds of birth-order effects that are observed during radical historical revolutions may depend as much on differences in 'family sentiments' as they do on personality. As Salmon and Daly (1998) have shown, firstborns (and to a lesser extent lastborns) are more strongly attached to the family system than are middle children. Historically, radical revolutions have tapped differences in family sentiments in two important ways. First, being considerably older than their offspring, parents were more likely to endorse the status quo, given that age is one of the best predictors of the acceptance of new and radical ideas (Hull et al. 1978). For this reason, endorsing a radical revolution has usually meant opposing parental values and authority, something that firstborns (and to a lesser extent lastborns) are less likely to do than are middleborns. Second, most radical revolutions, even in fields such as science, have tended to raise issues that are directly relevant to issues of parental investment and discrimination among offspring, filial loyalty, and overall identification with the family system. During the Reformation, for example, Protestant leaders such as Martin Luther strongly advocated the right to marriage by previously celibate clergymen and nuns, who tended to be younger sons and daughters. These Protestant leaders also advocated the egalitarian principles of partible inheritance, by which property—and even political rule—were subdivided equally among offspring (Fichtner 1989). During the height of the controversies raging over Darwin's theory of evolution, Darwin noted that discriminatory inheritance practices posed a serious impediment to human evolution, remarking to Alfred Russel Wallace: 'But oh, what a scheme is primogeniture for destroying natural selection!' (Sulloway 1996, p. 54).

Even when radical ideology does not play a central role in conceptual change, and when discriminatory inheritance systems are also not a factor, differences in identification with the family can impact on social and political attitudes. As Salmon (1998) has shown experimentally, political speeches appeal differentially to individuals by birth order depending on whether kinship terms are included in the speech. Firstborns and lastborns are more likely to react positively to speeches that employ kinship terms such as 'brother' and 'sister,' whereas middle children prefer speeches that employ references to 'friends.' To the extent that radical social movements make use of kinship terms, which they often do, birth-order differences in family sentiments will tend to influence how siblings react to radical change.

5. Conclusion

In past centuries birth order and functional sibling order have influenced such diverse social and political phenomena as royal succession, expansionist military campaigns, religious wars and crusades, geographic exploration, and inheritance laws. Even today, birth order and family niches more generally are among the environmental sources of personality because they cause siblings to experience the family environment in dissimilar ways. In particular, birth order introduces the need for differing strategies in dealing with sibling rivals as part of the universal quest for parental favor. This is a Darwinian story, albeit one with a marked environmental twist. Although siblings appear to be hard-wired to compete for parental favor, the specific niche in which they have grown up determines the particular strategies they adopt within their own family. Finally, because birth order and family niches underlie differences in family sentiments, including filial loyalty and self-conceptions about family identity, the family has historically supplied a powerful engine for revolutionary change.

See also: Family, Anthropology of; Family as Institution; Family Processes; Family Size Preferences; Family Systems and the Preferred Sex of Children; Galton, Sir Francis (1822–1911); Personality Development in Childhood; Property: Legal Aspects of Intergenerational Transmission; Sibling Relationships, Psychology of

Bibliography

Adler A 1927 *Understanding Human Nature*. Greenberg, New York

Boone J L 1986 Parental investment and elite family structure in preindustrial states: case study of late medieval-early modern Portuguese genealogies. *American Anthropologist* **88**: 859–78

Costa P T, McCrae R R 1992 *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, FL

Daly M, Wilson M 1988 *Homicide*. Aldine de Gruyter, Hawthorne, New York

Duby G 1977 *The Chivalrous Society*. Edward Arnold, London
Ernst C, Angst J 1983 *Birth Order: Its Influence on Personality*. Springer-Verlag, Berlin
Fichtner P S 1989 *Protestantism and Primogeniture in Early Modern Germany*. Yale University Press, New Haven, CT
Galton F 1874 *English Men of Science*. Macmillan, London
Hamilton W 1964 The genetical evolution of social behavior. Parts I and II. *Journal of Theoretical Biology* **7**: 1–52
Herlihy D 1977 Family and property in Renaissance Florence. In: Herlihy D, Udovitch A L (eds.) *The Medieval City*. Yale University Press, New Haven, CT, pp. 3–24
Hrdy S, Judge D S 1993 Darwin and the puzzle of primogeniture. *Human Nature* **4**: 1–45
Hull D L, Tessner P D, Diamond A M 1978 Planck's principle. *Science* **202**: 717–23
Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago
Paulhus D L, Chen D, Trapnell P D 1999 Birth order and personality within families. *Psychological Science* **10**: 482–8
Plomin R, Daniels D 1987 Why are children in the same family so different from one another. *Behavioral and Brain Sciences* **10**: 1–60
Rosenblatt P C, Skoogberg E L 1974 Birth order in cross-cultural perspective. *Developmental Psychology* **10**: 48–54
Salmon C 1998 The evocative nature of kin terminology in political rhetoric. *Politics and the Life Sciences* **17**: 51–7
Salmon C A, Daly M 1998 Birth order and familial sentiment: Middleborns are different. *Human Behavior and Evolution* **19**: 299–312
Schachter F F 1982 Sibling deidentification and split-parent identifications: A family tetrad. In: Lamb M E, Sutton-Smith B (eds.) *Sibling Relationships: Their Nature and Significance across the Lifespan*. Lawrence Erlbaum, Hillsdale, NJ, pp. 123–52
Sulloway F J 1995 Birth order and evolutionary psychology: A meta-analytic overview. *Psychological Inquiry* **6**: 75–80
Sulloway F J 1996 *Born to Rebel: Birth Order, Family Dynamics, and Creative Lives*. Pantheon, New York
Sulloway F J 1999 Birth order. In: Runco M A, Pritzker S (eds.) *Encyclopedia of Creativity* **1**: 189–202
Zajonc R B, Mullally P R 1997 Birth order: Reconciling conflicting effects. *American Psychologist* **52**: 685–99

F. J. Sulloway

Sibling Relationships, Psychology of

Siblings—brothers and sisters—have a key place in legends, history and literature throughout the world, from the era of the Egyptians and Greeks onward. The great majority of children (around 80 percent in Europe and the USA) grow up with siblings, and for most individuals their relationships with their siblings are the longest-lasting in their lives. Scientific study of the psychology of siblings is relatively recent, but is fast-growing, centering chiefly on studies of childhood and adolescence. The scientific interest of siblings lies, in particular, in the following domains: the nature and the potential influence of siblings on each other's development and adjustment, the illuminating perspective the study of siblings provides on developmental issues, and the challenge that siblings present to our understanding of how families influence development—why siblings differ notably in personality and adjustment even though they grow up within the same family, the significance of their shared and separate family experiences, and their genetic relatedness.

1. The Term 'Siblings'

This term is usually applied to 'full' siblings, brothers and sisters who are the offspring of the same mother and father, and share 50 percent of their genes; however, with the changes in family structure during the later decades of the twentieth century, increasing numbers of children have relationships with 'half siblings,' children with whom they share one biological parent, and 'step-siblings,' children who are unrelated biologically.

2. The Nature of Sibling Relationships

2.1 Characteristics of the Relationship

The relationship between siblings is one that is characterized by distinctive emotional power and intimacy from infancy onward. It is a relationship that offers children unique opportunities for learning about self and other, with considerable potential for affecting children's well-being, intimately linked as it is with each child's relationship with the parents. Clinicians and family systems theorists have, from early in the twentieth century, expressed interest in the part siblings play in family relationships, and in the adjustment of individuals. However, until the 1980s there was relatively little systematic research on siblings (with the notable exception of the classic studies of birth order by Koch 1954). In the 1980s and 1990s, research interest broadened greatly to include the investigation of sibling developmental influences, sources of individual differences between siblings, and links between family relationships (Boer and Dunn 1990, Brody 1996).

Individual differences in how siblings get along with each other are very marked from early infancy; siblings' feelings range from extreme hostility and rivalry to affection and support, are often ambivalent, and are expressed uninhibitedly. The relationship is also notable for its intimacy: siblings know each other very well, and this can be a source of both support and conflict. These characteristics increase the potential of the relationship for developmental influence. Because of the emotional intensity and familiarity of their relationships, the study of siblings provides an illuminating window on what children understand about each other, which has challenged and informed

conceptions of the development of early social understanding (Dunn 1992).

2.2 Developmental Changes in Sibling Relationships

As children's powers of understanding and communication develop, their sibling relationships change. Younger siblings play an increasingly active role in the relationship during the preschool years, and their older siblings take increasing interest in them. During middle childhood their relationships become more egalitarian; there is disagreement about how far this reflects an increase in the power of younger siblings over older siblings, or a decrease in the dominance both older and younger try to exert. A decrease in warmth between siblings during adolescence parallels the patterns of change found in the parent–child relationship as adolescents become increasingly involved with peers outside the family (Boer and Dunn 1990). There is a paucity of studies of siblings in adulthood; however, what information we have indicates that in the USA, most siblings maintain contact, communicate and share experiences until very late in life (Cicirelli 1996). During middle age, most adults describe feelings of closeness, rather than rivalry with their siblings, even when they have reduced contact, though family crises can evoke sibling conflict. Closeness and companionship become increasingly evident among older adults. Step- and half-siblings also continue to keep contact with each other, though they see each other less often than full siblings. Relationships with sisters appear to be particularly important in old age; this is generally attributed to women's emotional expressiveness and their traditional roles as nurturers. Little systematic research has focused on ethnic differences; however, in a national sample in the USA, sibling relationships in African-American, Hispanic, non-Hispanic whites, and Asian-American adult respondents were compared; the conclusion was that the similarities across groups in contact and social support far outweighed the differences (Riedmann and White 1966).

Siblings play a central role in adults' lives in many other cultures; the psychology of these relationships remains to be studied systematically.

2.3 Continuities in Individual Differences in Relationships

Do the marked individual differences in the affection or hostility that siblings show toward each other in early childhood show stability through middle childhood and adolescence? Research described in Brody (1996) indicates there is some continuity over time for both the positive and negative aspects of sibling relations, but there is also evidence for change. Contributors to change included new friendships children formed during the school years, which led to loss of warmth in the sibling relationship, or increased jealousy; developmental changes in the individuals; and life events (the majority of which contributed to increased intimacy and warmth, an exception being divorce or separation of parents). Research on adult siblings shows the relationship is not a static one during adulthood: life events, employment changes, etc. affect adult siblings' relations—some increasing intimacy and contact, others with negative effects (Cicirelli 1996).

2.4 Influences on Individual Differences

The temperament of both individuals in a sibling dyad is important in relation to conflict between them, and the match in their temperaments, too (Brody 1996). The effects of gender and age gap are less clear: Findings are inconsistent for young siblings, while research on children in middle childhood indicates that these 'family constellation' effects influence the relationship in complex ways; gender and socioeconomic status apparently increase in importance as influences on the sibling relationship during adolescence.

Most striking are the links between the quality of sibling relationships and other relationships within the family—the parents' relationships with each child, and the mother–partner or marital relationship. These connections are considered next.

3. Siblings, Parents, and Peers: Connections Between Relationships

To what extent and in what ways do parent–child relationships influence sibling relationships? Currently there is much debate and some inconsistency in the research findings. First, there is some evidence that the security of children's attachment to their parents is correlated with later sibling relationship quality, and that positive parent–child relations are correlated with positive sibling relations. Note that conclusions cannot be drawn about the direction of causal influence from these correlational data. There are also data that fit with a 'compensatory' model, in which intense supportive sibling relationships develop in families in which parents are uninvolved. This pattern may be characteristic of families at the extremes of stress and relationship difficulties.

More consistent is the evidence that differential parent–child relationships (in which more attention and affection, and less punishment is shown by a parent toward one sibling than another) are associated with more conflicted, hostile sibling relationships (Hetherington et al. 1994). These links are especially clear for families under stress, such as those who have experienced divorce, and in those with disabled or sick siblings. Again, note that the evidence is correlational.

Indirect links between parent–child and sibling relationships have been documented in studies of the arrival of a sibling; a number of different processes are implicated ranging from general emotional disturbance, through processes of increasing cognitive complexity—such as increased positivity between siblings in families in which mothers talked to the firstborn about the feelings and needs of the baby in early months. This evidence indicates that even with young siblings, processes of attribution and reflection may be implicated in the quality of their developing relationship.

The quality of the marital relationship is also linked to sibling relationships: mother–partner hostility is associated with increased negativity between siblings, while affection between adult partners is associated with positivity between siblings. Both direct pathways between marital and sibling relationships, and indirect pathways (via parent–child relationships) are implicated (Dunn et al. 1999).

4. *Developmental Influence*

4.1 *Influence on Adjustment*

Three adjustment outcome areas in which siblings are likely to exert influence are aggressive and externalizing behavior, internalizing problems, and self esteem. For example, hostile aggressive sibling behavior is correlated with increasing aggressive behavior by the other sibling. Patterson and his group have shown the shaping role that siblings play in this pattern (Patterson 1986), for both clinic and community samples. The arrival of a sibling is consistently found to be linked to increased problems: disturbance in bodily functions, withdrawal, aggressiveness, dependency, and anxiety. It is thought these changes are linked to parallel changes in interaction between the 'displaced' older siblings and the parents.

A growing literature links sibling influence to deviant behavior in adolescence. For example, frequent and problematic drinking by siblings increases adolescents' tendency to drink: siblings appear to have both a direct effect and an indirect later effect on other sibs' risks of becoming a drinker—through adolescents' selection of peers who drink. Siblings can also be an important source of support in times of stress, and act as therapists for siblings with some problems, such as eating disorders (Boer and Dunn 1990).

4.2 *Influence on Social Understanding*

The kinds of experience children have with their siblings are related to key aspects of their socio-cognitive development. For instance, positive cooperative experiences with older siblings are correlated with the development of greater powers of understanding emotion and others' mental states, both in early and middle childhood. Direction of effects in these associations is not clear: children who are good at understanding feelings and others' minds are more effective cooperative play companions, thus their early sophistication in social understanding may contribute to the development of cooperative play, which itself fosters further social understanding (Dunn 1992). Other aspects of prosocial, cooperative behavior, pretend play, and conflict management have all been reported to be associated with the experience of friendly sibling interactions.

While it appears plausible that experiences with siblings should 'generalize' to children's relationships with peers, the story appears more complex, and there is not consistent evidence for simple positive links. The emotional dynamics and demands of sibling and friend relationships are very different.

5. *Why Are Siblings So Different From One Another?*

Striking differences between siblings in personality, adjustment, and psychopathology have been documented in a wide range of studies. These present a challenge to conventional views of family influence, as the children are growing up within the same family. Extensive studies by behavior geneticists have now shown that the sources of environmental influence that make individuals different from one another work *within* rather than *between* families. The contribution of sibling studies to our understanding of the relative roles of genetics and environment in studies of socialization and development has been notable: the message here is not that family influence is unimportant, but that we need to document those experiences that are specific to each child in the family, and need to study more than one child in the family if we are to clarify what are the salient environmental influences on development (Dunn and Plomin 1990, Hetherington et al. 1994). The developmental research that documents the extreme sensitivity with which children monitor the interaction between other family members, and the significance of differential parent–child relationships combine here to clarify the social processes that are implicated in the development of differences between children growing up in the 'same' family.

6. *Methodological Issues*

For the study of siblings in childhood, a combination of naturalistic and structured observations, interviews with parents and siblings have proved most useful. Children are articulate and forthcoming about their relations with their siblings, and cope with interviews and questionnaires from an early age. As in the study of any relationship, it is important to get both participants' views on their relationship as these may

differ, and both are valid. Cicirelli (1996) points out some particular methodological problems with studying siblings in adults: incomplete data sets formed when one sibling is unwilling or unable to participate, choices over which dyad is picked, and limitations in sample representativeness.

7. Future Directions

Exciting future directions in sibling research include: clarification of the links between family relationships (including those involving step-relations); identification of the processes of family influence—for which we need studies with more than one child per family; further investigation of the role of genetics in individual development; and further exploration of significant sibling experiences for social understanding in middle childhood and adolescence.

See also: Family Processes; Genetic Studies of Personality; Infancy and Childhood: Emotional Development; Sibling-order Effects; Social Relationships in Adulthood

Bibliography

Boer F, Dunn J 1990 *Children's sibling relationships: Developmental and clinical issues*. Lawrence Erlbaum Associates, Mahwah, NJ

Brody G H 1996 *Sibling relationships their causes and consequences*. Ablex, Norwood, NJ

Cicirelli V G 1996 Sibling relationships in middle and old age. In: Brody G H (ed.) *Sibling relationships: Their causes and consequences*. Ablex, Norwood, NJ, pp. 47–73

Dunn J 1992 Siblings and development. *Current Directions in Psychological Science* **1**: 6–9

Dunn J, Deater-Deckard K, Pickering K, Beveridge M and the ALSPAC study team 1999 Siblings, parents and partners: Family relationships within a longitudinal community study. *Journal of Child Psychology and Psychiatry* **40**: 1025–37

Dunn J, Plomin R 1990 *Separate Lives: Why Siblings are so Different*. Basic Books, New York

Hetherington E M, Reiss D, Plomin R 1994 *Separate Social Worlds of Siblings: The Impact of Nonshared Environment on Development*. Lawrence Erlbaum Associates, Mahwah, NJ

Koch H L 1954 The relation of 'primary mental abilities' in five- and six-year-olds to sex of child and characteristics of his sibling. *Child Development* **25**: 209–23

Patterson G R 1986 The contribution of siblings to training for fighting: A microsocial analysis. In: Olweus D, Block J, Radke-Yarrow M (eds.) *Development of Antisocial and Prosocial Behavior: Research, Theories, and Issues*. Academic Press, Orlando, FL, pp. 235–61

Riedmann A, White L 1966 Adult sibling relationships: Racial and ethnic comparisons. In: Brody G H et al. (ed.) *Sibling Relationships: Their Causes and Consequences: Advances in Applied Developmental Psychology*. Ablex, Norwood, NJ, pp. 105–26

J. Dunn

Sign Language

Until recently, most of the scientific understanding of the human capacity for language has come from the study of spoken languages. It has been assumed that the organizational properties of language are connected inseparably with the sounds of speech, and the fact that language is normally spoken and heard determines the basic principles of grammar, as well as the organization of the brain for language. There is good evidence that structures involved in breathing and chewing have evolved into a versatile and efficient system for the production of sounds in humans. Studies of brain organization indicate that the left cerebral hemisphere is specialized for processing linguistic information in the auditory–vocal mode and that the major language-mediating areas of the brain are connected intimately with the auditory–vocal channel. It has even been argued that hearing and the development of speech are necessary precursors to this cerebral specialization for language. Thus, the link between biology and linguistic behavior has been identified with the particular sensory modality in which language has developed.

Although the path of human evolution has been in conjunction with the thousands of spoken languages the world over, recent research into signed languages has revealed the existence of primary linguistic systems that have developed naturally independent of spoken languages in a visual–manual mode. American Sign Language (ASL), for example, a sign language passed down from one generation to the next of deaf people, has all the complexity of spoken languages, and is as capable of expressing science, poetry, wit, historical change, and infinite nuances of meaning as are spoken languages. Importantly, ASL and other signed languages are not derived from the spoken language of the surrounding community: rather, they are autonomous languages with their own grammatical form and meaning. Although it was thought originally that signed languages were universal pantomime, or broken forms of spoken language on the hands, or loose collections of vague gestures, now scientists around the world have found that there are signed languages that spring up wherever there are communities and generations of deaf people (Klima and Bellugi 1988).

One can now specify the ways in which the formal properties of languages are shaped by their modalities of expression, sifting properties peculiar to a particular language mode from more general properties common to all languages. ASL, for example, exhibits formal structuring at the same levels as spoken languages (the internal structure of lexical units and the grammatical scaffolding underlying sentences) as well as the same kinds of organizational principles as spoken languages. Yet the form this grammatical structuring assumes in a visual–manual language is apparently

deeply influenced by the modality in which the language is cast (Bellugi 1980).

The existence of signed languages allows us to enquire about the determinants of language organization from a different perspective. What would language be like if its transmission were not based on the vocal tract and the ear? How is language organized when it is based instead on the hands moving in space and the eyes? Do these transmission channel differences result in any deeper differences? It is now clear that there are many different signed languages arising independently of one another and of spoken languages. At the core, spoken and signed languages are essentially the same in terms of organizational principles and rule systems. Nevertheless, on the surface, signed and spoken languages differ markedly. ASL and other signed languages display complex linguistic structure, but unlike spoken languages, convey much of their structure by manipulating spatial relations, making use of spatial contrasts at all linguistic levels (Bellugi et al. 1989).

1. *The Structure of Sign Language*

As already noted, the most striking surface difference between signed and spoken languages is the reliance on spatial contrasts, most evident in the grammar of the language. At the lexical level, signs can be separated from one another minimally by manual parameters (handshape, movement, location). The signs for *summer*, *ugly*, and *dry* are just the same in terms of handshape and movement, and differ only in the spatial location of the signs (forehead, nose, or chin). Instead of relying on linear order for grammatical morphology, as in English (*act*, *acting*, *acted*, *acts*), ASL grammatical processes nest sign stems in spatial patterns of considerable complexity (see Fig. 1), marking grammatical functions such as number, aspect, and person spatially. Grammatically complex forms can be nested spatially, one inside the other, with different orderings producing different hierarchically organized meanings.

Similarly, the syntactic structure specifying relations of signs to one another in sentences of ASL is also essentially organized spatially. Nominal signs may be associated with abstract points in a plane of signing space, and it is the direction of movement of verb signs between such endpoints that marks grammatical relations. Whereas in English, the sentences 'the cat bites the dog' and 'the dog bites the cat' are differentiated by the order of the words, in ASL these differences are signaled by the movement of the verb between points associated with the signs for *cat* and *dog* in a plane of signing space. Pronominal signs directed toward such previously established points or loci clearly function to refer back to nominals, even with many signs intervening (see Fig. 2). This spatial organization underlying syntax is a unique property of visual-gestural systems (Bellugi et al. 1989).

2. *The Acquisition of Sign Language by Deaf Children of Deaf Parents*

Findings revealing the special grammatical structuring of a language in a visual mode lead to questions about the acquisition of sign language, its effect on non-language visuospatial cognition, and its representation in the brain. Despite the dramatic surface differences between spoken and signed languages—simultaneity and nesting sign stems in complex co-occurring spatial patterns—the acquisition of sign language by deaf children of deaf parents shows a remarkable similarity to that of hearing children learning a spoken language. Grammatical processes in ASL are acquired at the same rate and by the same processing by deaf children as are grammatical processes by hearing children learning English, as if there were a biological timetable underlying language acquisition.

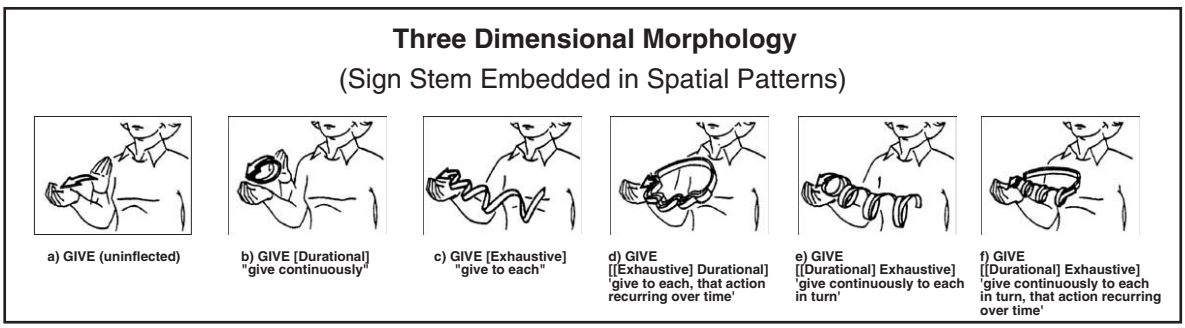

Figure 1
Spatially organized (three-dimensional) morphology in ASL

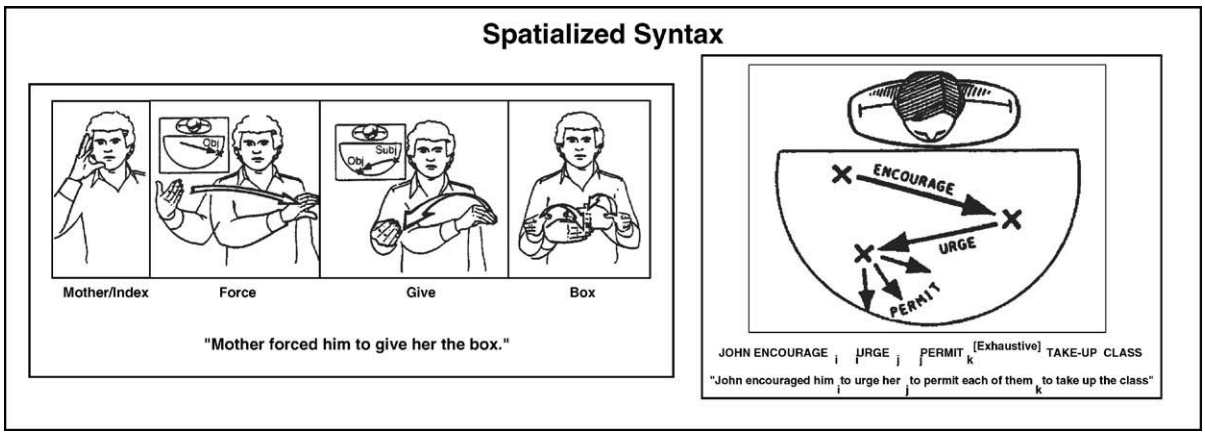

Figure 2
Spatialized syntax

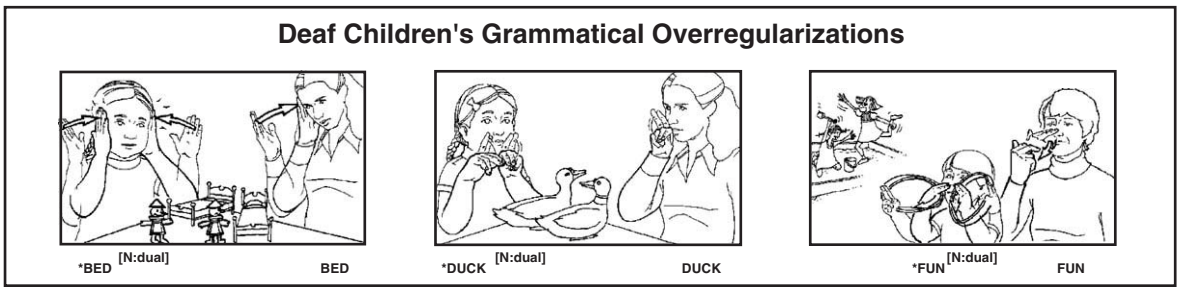

Figure 3
Deaf children's grammatical overregularizations

First words and first signs appear at around 12 months; combining two words or two signs in children, whether deaf or hearing, occurs by about 18–20 months, and the rapid expansion signaling the development of grammar (morphology and syntax) develops in both spoken and signed languages by about 3–3$\frac{1}{2}$ years. Just as hearing children show their mastery and discovery of grammatical regularities by producing overregularizations ('I goed there,' 'we helded the rabbit'), deaf children learning sign language show evidence of learning the spatial morphology signaling plural forms and aspectual forms by producing overregularizations in spatial patterns (see Fig. 3).

3. *Neural Systems Subserving Visuospatial Languages*

The differences between signed and spoken languages provide an especially powerful tool for understanding the neural systems subserving language. Consider the following: In hearing/speaking individuals, language processing is mediated generally by the left cerebral hemisphere, whereas visuospatial processing is mediated by the right cerebral hemisphere. But what about a language that is communicated using spatial contrasts rather than temporal contrasts? On the one hand, the fact that sign language has the same kind of complex linguistic structure as spoken languages and the same expressivity might lead one to expect left hemisphere mediation. On the other hand, the spatial medium so central to the linguistic structure of sign language clearly suggests right hemisphere or bilateral mediation. In fact, the answer to this question is dependent on the answer to another, deeper, question concerning the *basis* of the left hemisphere specialization for language.

Specifically, is the left hemisphere specialized for language processing *per se* (i.e., is there a brain basis for language as an independent entity)? Or is the left hemisphere's dominance generalized to process any type of information that is presented in terms of temporal contrasts? If the left hemisphere is indeed specialized for processing language itself, sign language processing should be mediated by the left

hemisphere, as is spoken language. If, however, the left hemisphere is specialized for processing fast temporal contrasts in general, one would expect sign language processing to be mediated by the right hemisphere. The study of sign languages in deaf signers permits us to pit the nature of the signal (auditory-temporal vs. visual-spatial) against the type of information (linguistic vs. nonlinguistic) that is encoded in that signal as a means of examining the neurobiological basis of language (Poizner et al. 1990).

One program of studies examines deaf lifelong signers with focal lesions to the left or the right cerebral hemisphere. Major areas, each focusing on a special property of the visual-gestural modality as it bears on the investigation of brain organization for language, are investigated. There are intensive studies of large groups of deaf signers with left or right hemisphere focal lesions in one program (Salk); all are highly skilled ASL signers, and all used sign as a primary form of communication throughout their lives. Individuals were examined with an extensive battery of experimental probes, including formal testing of ASL at all structural levels; spatial cognitive probes sensitive to right-hemisphere damage in hearing people; and new methods of brain imaging, including structural and functional magnetic resonance imaging (MRI, fMRI), event-related potentials (ERP), and positron emission tomography (PET). This large pool of well-studied and thoroughly characterized subjects, together with new methods of brain imaging and sensitive tests of signed as well as spoken language allows for a new perspective on the determinants of brain organization for language (Hickok and Bellugi 2000, Hickok et al. 1996).

3.1 Left Hemisphere Lesions and Sign language Grammar

The first major finding is that so far only deaf signers with damage to the left hemisphere show sign language aphasias. Marked impairment in sign language after left hemisphere lesions was found in the majority of the left hemisphere damaged (LHD) signers, but not in any of the right hemisphere damaged (RHD) signers, whose language profiles were much like matched controls. Figure 4 presents a comparison of LHD, RHD, and normal control profiles of sign characteristics from The Salk Sign Diagnostic Aphasia Examination—a measure of sign aphasia. The RHD signers showed no impairment at all in any aspect of ASL grammar; their signing was rich, complex, and without deficit, even in the spatial organization underlying sentences of ASL. By contrast, signers with LHD showed markedly contrasting profiles: one was agrammatic after her stroke, producing only nouns and a few verbs with none of the grammatical apparatus of ASL, another made frequent paraphasias at the sign internal level, and several showed many grammatical paraphasias, including neologisms, particularly in morphology.

Another deaf signer showed deficits in the spatially encoded grammatical operations which link signs in sentences, a remarkable failure in the spatially organized syntax of the language. Still another deaf signer with focal lesions to the left hemisphere reveal dissociations not found in spoken language: a dissociation between sign and gesture, with a cleavage between capacities for sign language (severely impaired) and manual pantomime (spared). In contrast, none of the RHD signers showed any within-sentence deficits; they were completely unimpaired in sign sentences and not one showed aphasia for sign language (in contrast to their marked nonlanguage spatial deficits, described below) (Hickok and Bellugi 2000, Hickok et al. 1998).

Moreover, there are dramatic differences in performance between left- and right-hemisphere damaged signers in formal experimental probes of sign competence. For example, a test of the equivalent of rhyming in ASL provides a probe of phonological processing. Two signs 'rhyme' if they are similar in all but one phonological parametric value such as handshape, location, or movement. To tap this aspect of phonological processing, subjects are presented with an array of pictured objects and asked to pick out the two objects whose signs rhyme (Fig. 5). The ASL signs for *key* and *apple* share the same handshape and movement, and differ only in location, and thus are like rhymed pairs. LHD signers are significantly impaired relative to RHD signers and controls on this test, another sign of the marked difference in effects of right- and left-hemisphere lesions on signing. On other tests of ASL processing at different structural levels, there are similar distinctions between left- and right-lesioned signers, with the right-lesioned signers much like the controls, but the signers with left hemisphere lesions significantly impaired in language processing. Moreover, studies have found that there can be differential breakdown of linguistic components of sign language (lexicon and grammar) with different left hemisphere lesions.

3.2 Right Hemisphere Lesions and Nonlanguage Spatial Processing

These results from language testing contrast sharply with results on tests of nonlanguage spatial cognition. RHD signers are significantly more impaired on a wide range of spatial cognitive tasks than LHD signers, who show little impairment. Drawings of many of the RHD signers (but not those with LHD) show severe spatial distortions, neglect of the left side of space, and lack of perspective. RHD deaf signers show lack of perspective, left neglect, and spatial disorganization on an array of spatial cognitive nonlanguage tests (block design, drawing, hierarchical

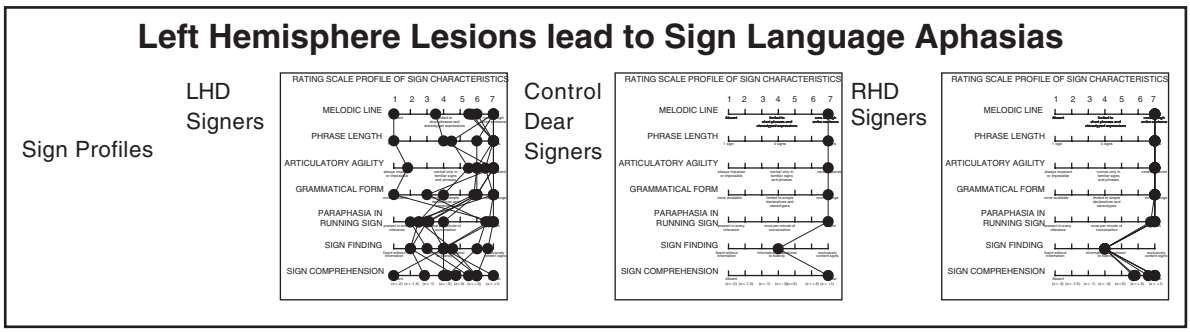

Figure 4
Left hemisphere lesions lead to sign language aphasias

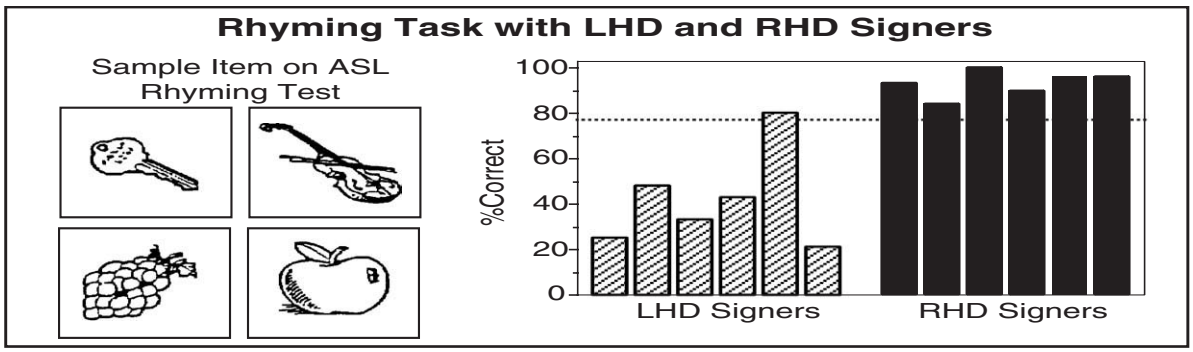

Figure 5
Rhyming task with LHD and RHD signers

processing), compared with LHD deaf signers. Yet, astonishingly, these severe spatial deficits among RHD signers do not affect their competence in a spatially nested language, ASL. The case of a signer with a right parietal lesion leading to severe left neglect is of special interest: Whereas his drawings show characteristic omissions on the left side of space, his signing (including the spatially organized syntax) is impeccable, with signs and spatially organized syntax perfectly maintained.

The finding that sign aphasia follows left hemisphere lesions but not right hemisphere lesions provides a strong case for a modality-independent *linguistic* basis for the left hemisphere specialization for language. These data suggest that the left hemisphere is predisposed biologically for language, independent of language modality. Thus, hearing and speech are not necessary for the development of hemisphere specialization—sound is not crucial. Furthermore, the finding of a dissociation between competence in a spatial language and competence in nonlinguistic spatial cognition demonstrates that it is the type of information that is encoded in a signal (i.e., linguistic vs. spatial information) rather than the nature of the signal itself (i.e., spatial vs. temporal) that determines the organization of the brain for higher cognitive functions.

4. *Language, Modality, and the Brain*

Analysis of patterns of breakdown in deaf signers provides new perspectives on the determinants of hemispheric specialization for language. First, the data show that hearing and speech are not necessary for the development of hemispheric specialization: sound is *not* crucial. Second, it is the left hemisphere that is dominant for sign language. Deaf signers with damage to the left hemisphere show marked sign language deficits but relatively intact capacity for processing nonlanguage visuospatial relations. Signers with damage to the right hemisphere show the reverse pattern. Thus, not only is there left hemisphere specialization for language functioning, there is also complementary specialization for nonlanguage spatial functioning. The fact that grammatical information in sign language is conveyed via spatial manipulation does not alter this complementary specialization.

Furthermore, components of sign language (lexicon and grammar) can be selectively impaired, reflecting

differential breakdown of sign language along linguistically relevant lines. These data suggest that the left hemisphere in humans may have an innate predisposition for language, regardless of the modality. Since sign language involves an interplay between visuospatial and linguistic relations, studies of sign language breakdown in deaf signers may, in the long run, bring us closer to the fundamental principles underlying hemispheric specialization.

See also: Evolution and Language: Overview; Language and Animal Competencies; Sign Language: Psychological and Neural Aspects

Bibliography

Bellugi U 1980 The structuring of language: Clues from the similarities between signed and spoken language. In: Bellugi U, Studdert-Kennedy M (eds.) *Signed and Spoken Language: Biological Constraints on Linguistic Form*. Dahlem Konferenzen. Weinheim/Deerfield Beach, FL, pp. 115–40

Bellugi U, Poizner H, Klima E S 1989 Language, modality and the brain. *Trends in Neurosciences* **10**: 380–8

Emmorey K, Kosslyn S M, Bellugi U 1993 Visual imagery and visual-spatial language: Enhanced imagery abilities in deaf and hearing ASL signers. *Cognition* **46**: 139–81

Hickok G, Bellugi U 2000 The signs of aphasia. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*, 2nd edn. Elsevier Science Publishers, Amsterdam, The Netherlands

Hickok G, Bellugi U, Klima E S 1996 The neurobiology of signed language and its implications for the neural organization of language. *Nature* **381**: 699–702

Hickok G, Bellugi U, Klima E S 1998 The basis of the neural organization for language: Evidence from sign language aphasia. *Reviews in the Neurosciences* **8**: 205–22

Klima E S, Bellugi U 1988 *The Signs of Language*. Harvard University Press, Cambridge, MA

Poizner H, Klima E S, Bellugi U 1990 *What the Hands Reveal About the Brain*. MIT Press/Bradford Books, Cambridge, MA

U. Bellugi

Sign Language: Psychological and Neural Aspects

Signed languages of the deaf are naturally occurring human languages. The existence of languages expressed in different modalities (i.e., oral–aural, manual–visual) provides a unique opportunity to explore and distinguish those properties shared by all human languages from those that arise in response to the modality in which the language is expressed. Despite the differences in language form, signed languages possess formal linguistic properties found in spoken languages. Sign language acquisition follows a developmental trajectory similar to spoken languages. Memory for signs exhibits patterns of interference and forgetting that are similar to those found for speech. Early use of sign language may enhance certain aspects of nonlanguage visual perception. Neuropsychological studies show that left hemisphere regions are important in both spoken and sign language processing.

1. Linguistic Principles of American Sign Language

1.1 Language and Deaf Culture

Sign languages are naturally-occurring manual languages that arise in communities of deaf individuals. These manual communication systems are fully expressive, systematic human languages and are not merely conventionalized systems of pantomime nor manual codifications of a spoken language. Many types of deafness are inheritable and it is not unusual to find isolated communities of deaf individuals who have developed complex manual languages (Groce 1985). The term Deaf Community has been used to describe a sociolinguistic entity that plays a crucial role in a deaf person's exposure to and acceptance of sign language (Padden and Humphries 1988). American Sign Language (ASL), used by members of the Deaf community in the USA and Canada, is only one of many sign languages of the world, but it is the one that has been studied most extensively .

1.2 Structure of American Sign Language

A sign consists of a hand configuration that travels about a movement path and is directed to or about a specific body location. Sign languages differ from one another in the inventories and compositions of hand-shapes, movements, and locations used to signal linguistic contrasts just as spoken language differ from one another in the selection of sounds used, and how those sounds are organized into words. Many sign languages incorporate a subsystem in which orthographic symbols used in the surrounding spoken language communities are represented manually on the hands. One example is the American English manual alphabet, which is produced on one hand and allows users of ASL to represent English lexical items.

All human languages, whether spoken or signed, exhibit levels of structure which govern the composition and formation of word forms and specify how words combine into sentences. In formal linguistic analyses, these structural levels are referred to as phonology, morphology and the syntax of the language. In this context, phonological organization refers to the patterning of the abstract formational

units of a natural language (Coulter and Anderson 1993). Compared to spoken languages in which contrastive units (for example, phonemes) are largely arranged in a linear fashion, signed languages exhibit simultaneous layering of information. For example, in a sign, a handshape will co-occur with, rather than follow sequentially, a distinct movement pattern.

ASL exhibits complex morphology. Morphological markings in ASL are expressed as dynamic movement patterns overlaid on a more basic sign form. These nested morphological forms stand in contrast to the linear suffixation common in spoken languages (Klima and Bellugi 1979). The prevalence of simultaneous layering of phonological content and the nested morphological devices observed across many different sign languages likely reflect an influence of modality on the realization of linguistic structure. Thus the shape of human languages reflects the constraints and affordances imposed by the articulator systems involved in transmission of the signal (i.e., oral versus manual) and the receptive mechanisms for decoding the signal (i.e., auditory versus visual).

A unique property of ASL linguistic structure is the reliance upon visuospatial mechanisms to signal linguistic contrasts and relations. One example concerns the use of facial expressions in ASL. In ASL, certain syntactic and adverbial constructions are marked by specific and obligatory facial expressions (Liddell 1980). These linguistic facial expressions differ significantly in appearance and execution of affective facial expressions. A second example concerns the use of inflectional morphology to express subject and object relationships. At the syntactic level, nominals introduced into the discourse are assigned arbitrary reference points along a horizontal plane in the signing space. Signs with pronominal function are directed toward these points, and the class of verbs which require subject/object agreement obligatorily move between these points (Lillo-Martin and Klima 1990). Thus, whereas many spoken languages represent grammatical functions through case marking or linear ordering, in ASL grammatical function is expressed through spatial mechanisms. This same system of spatial reference, when used across sentences, serves as a means of discourse cohesion (Winston 1995).

1.3 Sign Language Acquisition

Children exposed to signed languages from birth acquire these languages on a similar maturational timetable as children exposed to spoken languages (Meier 1991). Prelinguistic infants, whether normally hearing or deaf, engage in vocal play commonly known as babbling. Recent research has shown that prelinguistic gestural play referred to as manual babbling will accompany vocal babbling. There appear to be significant continuities between prelinguistic gesture and early signs in deaf children exposed to American Sign Language, though questions remain as to whether manual babbling is constrained by predominantly motoric or linguistic factors (Cheek et al. 2001, Petitto and Marantette 1991).

Between 10 and 12 months of age, children reared in signing homes begin to produce their first signs, with two-sign combinations appearing at approximately 18 months. Some research has suggested a precocious early vocabulary development in signing children. However, these reports are likely to be a reflection of parents' and experimenters' earlier recognition of signs compared to words rather than underlying differences in development of symbolic capacities of signing and speaking children.

Signing infants produce the same range of grammatical errors in signing that have been documented for spoken language, including phonological substitutions, morphological overregularizations, and anaphoric referencing confusions (Petitto 2000). For example, in the phonological domain a deaf child will often use only a subset of the handshapes found in the adult inventory, opting for a simpler set of 'unmarked' handshapes. This is similar to the restricted range of consonant usage common in children acquiring a spoken language.

2. Psycholinguistic Aspects of Sign Language

Psycholinguistic studies of American Sign Language have examined how the different signaling characteristics of the language impact transmission and recognition. Signs take longer to produce than comparable word forms. The average number of words per second in running speech is about 4 to 5, compared with 2 to 3 signs per second in fluent signing. However, despite differences in word transmission rate, the proposition rate for speech and sign is the same, roughly one proposition every 1 to 2 seconds. Compare, for example the English phrase 'I have already been to California' with the ASL equivalent, which can be succinctly signed using three monosyllabic signs, glossed as FINISH TOUCH CALIFORNIA.

2.1 Sign Recognition

Studies of sign recognition have examined how signs are recognized in time. Recognition appears to follow a systematic pattern in which information about the location of the sign is reliably identified first, followed by handshape information and finally the movement. Identification of the movement reliably leads to the identification of the sign. Interestingly, despite the slower articulation of a sign compared to a word, sign recognition appears to be faster than word recognition. Specifically, it has been observed that proportionally less of the signal needs to be processed in order to uniquely identify a sign compared to a spoken

word. For example, one study reports that only 240 msec. or 35 percent of a sign has to be seen before a sign is identified (Emmory and Corina 1990). In comparable studies of spoken English, Grosjean (1980) reports that 330 msec. or 83 percent of a word has to be heard before a word can be identified. This finding is due, in part, to the simultaneous patterning of phonological information in signs compared to the more linear patterning of phonological information characteristic of spoken languages. These structural differences in turn have implications for the organization of lexical neighborhoods. Spoken languages, such as English, may have many words that share in their initial phonological structures (tram, tramp, trampoline, etc.) leading to greater coactivation (and thus longer processing time) during word recognition. In contrast, ASL sign forms are often formationally quite distinct from one another, permitting quicker selection of a lexical unique entry.

2.2 Memory for Signs

Efforts to explore how the demands of sign language processing influence memory and attention have led to several significant findings. Early studies of memory for lists of signs report classic patterns of forgetting and interference, including serial position effects of primacy and recency (i.e., signs at the beginning and the end of a list are better remembered than items in the middle). Likewise, when sign lists are composed of phonologically similar sign forms, signers exhibit poorer recall (Klima and Bellugi 1979). This result is similar to what has been reported for users of spoken language, where subjects exhibit poorer memory for lists of similarly sounding words (Conrad and Hull 1964). These findings indicate that signs, like words, are encoded into short-term memory in a phonological or articulatory code rather than in a semantic code.

More recent work has assessed whether Baddeley's (1986) working memory model pertains to sign language processing. This model includes components that encode linguistic and visuospatial information as well as a central executive which mediates between immediate and long-term memory. Given the spatial nature of the sign signal, the question of which working memory component(s) is engaged during sign language processing is particularly interesting. Wilson and Emmorey (2000) have shown word-length effects in signs; it is easier to maintain a cohort of short signs (monosyllabic signs) than long signs (polysyllabic signs) in working memory. They also report effects of 'articulatory suppression.' In these studies, requiring a signer to produce repetitive, sign-like hand movements while encoding and maintaining signs in memory has a detrimental effect on aspects of recall. Once again, analogous effects are known to exist for spoken languages and these phenomena provide support for a multicomponent model of linguistic working memory that includes both an articulatory loop and a phonological store. Finally, under some circumstances deaf signers are able to utilize linguistically relevant spatial information to encode signs, suggesting the engagement of a visuospatial component of working memory.

2.3 Perception and Attention

Recent experiments with native users of signed languages have shown that experience with a visual sign language may improve or alter visual perception of nonlanguage stimuli. For example, compared to hearing nonsigners, deaf signers have been shown to possess enhanced or altered perception along several visual dimensions such as motion processing (Bosworth and Dobkins 1999), mental rotation, and processing of facial features (Emmorey 2001). For example, the ability to detect and attend selectively to peripheral, but not central, targets in vision is enhanced in signers (Neville 1991). Consistent with this finding is evidence for increased functional connectivity in neural areas mediating peripheral visual motion in the deaf (Bavelier et al. 2000).

As motion processing, mental rotation, and facial processing underlie aspects of sign language comprehension, these perceptual changes have been attributed to experience with sign language. Moreover, several of these studies have reported such enhancements in hearing signers raised in deaf signing households. These findings provide further evidence that these visual perception enhancements are related to the acquisition of a visual manual language, and are not due to compensatory mechanisms developed as a result of auditory deprivation.

3. Neural Organization of Signed Language

3.1 Hemispheric Specialization

One of the most significant findings in neuropsychology is that the two cerebral hemispheres show complementary functional specialization, whereby the left hemisphere mediates language behaviors while the right hemisphere mediates visuospatial abilities. As noted, signed languages make significant use of visuospatial mechanisms to convey linguistic information. Thus sign languages exhibit properties for which each of the cerebral hemispheres shows specialization. Neuropsychological studies of brain injured deaf signers and functional imaging studies of brain intact signers have provided insights into the neural systems underlying sign language processing.

3.2 Aphasia in Signed Language

Aphasia refers to acquired impairments in the use of language following damage to the perisylvian region

of the left hemisphere. Neuropsychological case studies of deaf signers convincingly demonstrate that aphasia in signed language is also found after left hemisphere perisylvian damage (Poizner et al. 1987). Moreover, within the left hemisphere, production and comprehension impairments follow the well-established anterior versus posterior dichotomy. Studies have documented that frontal anterior damage leads to Broca-like sign aphasia. In these cases, normal fluent signing is reduced to effortful, single-sign, 'telegraphic' output with little morphological complexity (such as verb agreement). Comprehension, however, is left largely intact. Wernicke-like sign aphasia following damage to the posterior third of the perisylvian region presents with fluent but often semantically opaque output, and comprehension also suffers. In cases of left hemisphere damage, the occurrence of sign language paraphasias may be observed. Paraphasia may be formationally disordered (for example, substituting an incorrect hand shape) or semantically disordered (producing a word that is semantically related to the intended form). These investigations serve to underscore the importance of left hemisphere structures for the mediation of signed languages and illustrate that sign language breakdown is not haphazard, but rather honors linguist boundaries (Corina 2000).

The effects of cortical damage to primary language areas that result in aphasic disturbance can be differentiated from more general impairments of movement. For example Parkinson's disease leads to errors involving timing, scope, and precision of general movements including, but not limited to, those involved in speech and signing. These errors produce phonetic disruptions in signing, rather than higher-level phonemic disruptions that are apparent in aphasic signing (Corina 1999).

Apraxia is defined as an impairment of the execution of a learned movement (e.g., saluting, knowing how to work a knife and fork). Lesions associated with the left inferior parietal lobe result in an inability to perform and comprehend gestures (Gonzalez Rothi and Heilman 1997). Convincing dissociations of sign language impairment with well-preserved praxic abilities have been reported. In one case, a subject with marked sign language aphasia affecting both production and comprehension produced unencumbered pantomime. Moreover, both comprehension and production of pantomime were found to be better preserved than was sign language. These data indicate that language impairments following left hemisphere damage are not attributable to undifferentiated symbolic impairments and demonstrate that ASL is not simply an elaborate pantomimic system.

3.3 Role of the Right Hemisphere in ASL

Studies of signers with right hemisphere damage (RHD) have reported significant visuospatial disruption in nonlinguistic domains (e.g., face processing, drawing, block construction, route finding, etc.) but have reported only minimal language disruption. Problems in the organization of discourse have been observed in RHD signers, as have also been reported in users of spoken languages. More controversial are the influences of visuospatial impairment in the use of highly spatialized components of the language such as complex verb agreement and the classifier system. Further work is needed to understand whether these infrequently reported impairments (both in comprehension and production) reflect core linguistic deficits or rather reflect secondary effects of impaired visuospatial processing.

3.4 Functional Imaging

Recent studies using functional brain imaging have explored the neural organization of language in users of signed languages. These studies have consistently found participation of classic left hemisphere perisylvian language areas in the mediation of sign language in profoundly deaf, lifelong signers. For example Positron Emission Tomography studies of production in British Sign Language (McGuire et al. 1997) and Langue des Signes Quebecoise (Petitto et al. 2000) reported deaf subjects activated left inferior frontal regions, regions similar to those that mediate speech in hearing subjects. Researchers using Functional Magnetic Resonance imaging techniques in investigating comprehension of ASL in deaf and hearing native users of signed language have shown significant activations in frontal opercular areas (including Broca's area and dorsolateral prefrontal cortex) as well as in posterior temporal areas (such as Wernicke's area and the angular gyrus) (Neville et al. 1998). Also reported was extensive activation in right hemisphere regions. Subsequent studies have conferred that aspects of the right posterior hemisphere activation appear to be unique to sign language processing and present only in signers who acquired sign language from birth (Newman et al. in press).

Investigations of psychological and neural aspects of signing reveal strong commonalities in the development and cognitive processing of signed and spoken languages despite major differences in the surface form of these languages. Early exposure to a sign language may lead to the enhancements in the specific neural systems underlying visual processing. There appears to be a strong biological predisposition for left hemisphere structures in the mediation of language, regardless of the modality of expression.

See also: Language Acquisition; Language Development, Neural Basis of; Sign Language

Bibliography

Baddeley A 1986 *Working Memory*. Oxford University Press, New York

Bavelier D, Tomann A, Hutton C, Mitchell T, Corina D, Liu G, Neville H 2000 Visual attention at the periphery is enhanced in congenitally deaf individuals. *Journal of Neuroscience* **20**: RC93

Bosworth R, Dobkins K 1999 Left-hemisphere dominance for motion processing in deaf signers. *Psychological Science* **10**: 256–62

Cheek A, Cormier K, Repp A, Meier R 2001 Prelinguistic gesture predicts mastery and error in the production of early signs. *Language*

Corina D 1999 Neural disorders of language and movement: Evidence from American Sign Language. In: Messing L, Campbell R (eds.) *Gesture, Speech and Sign*. Oxford University Press, New York

Corina D 2000 Some observations regarding paraphasia in American Sign Language. In: Emmorey K, Lane H (eds.) *The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima*. Lawrence Erlbaum Associates, Mahwah, NJ

Coulter G, Anderson S 1993 Introduction. In: Coulter G (ed.) *Phonetics and Phonology: Current Issues in ASL Phonology*. Academic Press, San Diego, CA

Conrad R, Hull A 1964 Information, acoustic confusion and memory span. *British Journal of Psychology* **55**: 429–32

Emmorey K 2001 *Language, Cognition, and the Brain: Insights From Sign Language Research*. Lawrence Erlbaum Associates, Mahwah, NJ

Emmorey K, Corina D 1990 Lexical recognition in sign language: Effects of phonetic structure and morphology. *Perceptual and Motor Skills* **71**: 1227–52

Gonzalez Rothi L, Heilman K 1997 Introduction to limb apraxia. In: Gonzalez Rothi L, Heilman K (eds.) *Apraxia: The Neuropsychology of Action*. Psychology, Hove, UK

Grosjean F 1980 Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics* **28**: 267–83

Groce J 1985 *Everyone Here Spoke Sign Language*. Harvard University Press, Cambridge, MA

Klima E, Bellugi U 1979 *The Signs of Language*. Harvard University Press, Cambridge, MA

Liddell S 1980 *American Sign Language Syntax*. Mouton, The Hague, The Netherlands

Lillo-Martin D, Klima E 1990 Pointing out differences: ASL pronouns in syntactic theory. In: Fisher S, Siple P (eds.) *Theoretical Issues in Sign Language Research I: Linguistics*. University of Chicago Press, Chicago

McGuire P, Robertson D, Thacker A, David A, Frackowiak R, Frith C 1997 Neural correlates of thinking in sign language. *Neuroreport* **8**: 695–8

Meier R 1991 Language acquisition by deaf children. *American Scientist* **79**: 60–70

Neville H 1991 Neurobiology of cognitive and language processing: Effects of early experience. In: Gibson K, Peterson A (eds.) *Brain Maturation and Cognitive Development: Comparative and Cross-cultural Perspectives*. Aldine de Gruyter Press, Hawthorne, NY

Neville H, Bavelier D, Corina D, Rauschecker J, Karni A, Lalwani A, Braun A, Clark V, Jezzard P, Turner R 1998 Cerebral organization for language in deaf and hearing subjects: Biological constraints and effects of experience. *Proceedings of the National Academy of Science* **90**: 922–9

Newman A, Corina D, Tomann A, Bavelier D, Jezzard P, BraunA, Clark V, Mitchell T, Neville H (submitted) Effects of age of acquisition on cortical organization for American Sign Language: an fMRI study. *Nature Neuroscience*

Padden C, Humphries T 1988 *Deaf in America: Voices from a Culture*. Harvard University Press, Cambridge, MA

Petitto L 2000 The acquisition of natural signed languages: Lessons in the nature of human language and its biological foundations. In: Chamberlain C, Morford J (eds.) *Language Acquisition by Eye*. Lawrence Erlbaum Associates, Mahwah, NJ

Petitto L, Marantette P 1991 Babbling in the manual mode: Evidence for the ontogeny of language. *Science* **251**: 1493–6

Petitto L, Zatorre R, Gauna K, Nikelski E, Dostie D, Evans A 2000 Speech-like cerebral activity in profoundly deaf people processing signed languages: Implications for the neural basis of human language. *Proceedings of the National Academy of Science* **97**: 13961–6

Poizner H, Klima E, Bellugi U 1987 *What the Hands Reveal About the Brain*. MIT Press, Cambridge, MA

Wilson M, Emmorey K 2000 When does modality matter? Evidence from ASL on the nature of working memory. In: Emmorey K, Lane H (eds.) *The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima*. Lawrence Erlbaum Associates, Mahwah, NJ

Winston E 1995 Spatial mapping in comparative discourse frames. In: Emmorey K, Reilly J (eds.) *Language, Gesture, and Space*. Lawrence Erlbaum Associates, Mahwah, NJ

D. P. Corina

Signal Detection Theory

Signal detection theory (SDT) is a framework for interpreting data from experiments in which accuracy is measured. It was developed in a military context (see *Signal Detection Theory, History of*), then applied to sensory studies of auditory and visual detection, and is now widely used in cognitive science, diagnostic medicine, and many other fields. The key tenets of SDT are that the internal representations of stimulus events include variability, and that perception (in an experiment, or in everyday life) incorporates a decision process. The theory characterizes both the representation and the decision rule.

1. Detection and Discrimination Experiments

Consider an auditory experiment to determine the ability of a listener to detect a weak sound. On some trials, the sound is presented, and the listener is instructed to say 'yes' (I heard it) or 'no' (I did not hear it). On other trials, there is no sound, discouraging the observer from responding 'yes' indiscriminately. The presence of variability in sensory systems leads us to call these latter presentations *noise* (N) trials, and the former *signal plus noise* ($S+N$) trials. Possible data based on 50 trials of each type are given in Table 1. Responses of 'yes' on signal trials are called *hits*, and

Table 1
Response frequencies in an auditory detection experiment

	'Yes'	'No'	Total
Signal + noise	42	8	50
Noise	15	35	50

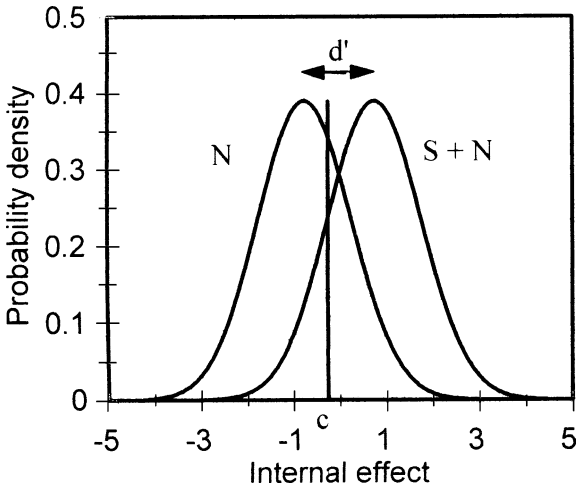

Figure 1
Distributions assumed by SDT to result from N and $S+N$. The horizontal axis is scaled in standard deviation units, and the difference between the means is d'. The observer responds 'yes' for values to the right of the criterion (vertical line), 'no' for those to the left

on no-signal trials are called *false alarms*. The observer's performance can be summarized by a *hit rate* (here 0.84) and a *false-alarm rate* (here 0.30).

The variability principle implies that repeated presentation of the signal leads to a distribution of internal effect, as does repeated presentation of no signal. In the most common SDT model, these distributions are assumed to be normal, and to differ only in mean, as shown in Fig. 1.

How does the observer reach a decision in this experiment, when any observation on the 'internal effect' axis is ambiguous with regard to source? The best decision rule is to establish a *criterion* value c on the axis, and respond 'yes' for values to the right of c, 'no' for values to the left. The location of the criterion can be determined from the data, for the area above it under the $S+N$ distribution must equal the hit rate. Consulting a table of the normal distribution reveals that in the example above the criterion must therefore be about one standard deviation below the mean of that distribution. Similarly, the area above the criterion under the N distribution must equal the false-alarm rate, so the criterion is about 0.5 standard deviations above the mean of that distribution.

2. *Sensitivity and Response Bias*

Figure 1 offers natural interpretations of the *sensitivity* of the observer in the task and of the *decision rule*. Sensitivity, usually denoted d', is reflected by the difference between the two distribution means; this characteristic of the representation is unaffected by the location of the criterion, and is thus free of response bias. The magnitude of d' is found easily by computing the distances between each mean and the criterion, found in the last section:

$$d' = z(\textit{hit rate}) - z(\textit{false-alarm rate}) \qquad (1)$$

where $z(p)$ is the point on a unit-normal distribution above which the area p can be found. The function z is negative for arguments less than 0.5 and positive for arguments above that value. In the example,

$$d' = z(0.84) - z(0.30) = 0.994 - (-0.524) = 1.518$$

Criterion location is one natural measure of response bias. If evaluated relative to the point at which the two distributions cross, it can be calculated by a formula analogous to Eqn. 1:

$$c = -\tfrac{1}{2}[z(\textit{hit rate}) + z(\textit{false-alarm rate})] \qquad (2)$$

In the example above

$$c = -\tfrac{1}{2}[z(0.84) + z(0.30)] = -0.235,$$

that is, the criterion is slightly to the left of the crossover point, as in fact is shown in Fig. 1.

3. *Receiver Operating Characteristic (ROC) Curves*

If sensitivity truly is unaffected by the location of the response criterion, then a shift in that criterion should leave d' unchanged. An observer's decision rule can be manipulated between experimental runs through instructions or payoffs. In the more efficient rating method, a rating response is required; the analysis assumes that each distinct rating corresponds to a different region along the axis of Fig. 1, and that each region is defined by a different criterion. The data are analyzed by calculating a hit and false-alarm rate for each criterion separately. A plot of hit rate vs. false alarm rate is called a *receiver-operating characteristic*, or *ROC*; the ROC that would be produced from the

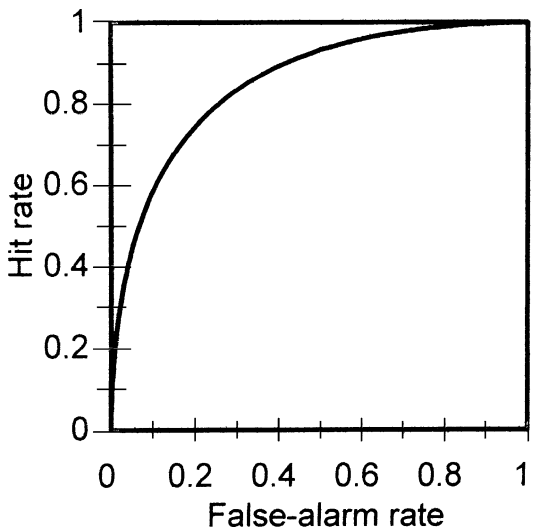

Figure 2
An ROC curve, the relation between hit and false-alarm rates, both of which increase as the criterion location moves (from right to left, in Fig. 1)

representation in Fig. 1 is shown in Fig. 2. Points at the lower end of the curve correspond to high criteria, and the curve is traced out from left to right as the criterion moves from right to left in Fig. 1.

The shape of the ROC is determined by the form of the distributions in the representation, and data largely support the normality assumption. However, the assumption that the $S+N$ and N distributions have the same variance is often incorrect (Swets 1986). In such cases, the area under the ROC is a good measure of accuracy (Swets and Pickett 1982); this index ranges from 0.5 for chance performance to 1.0 for perfect discrimination. An entire ROC (from a rating experiment, for example) is needed to estimate this area.

4. *Alternative Assumptions About Representation*

The essence of signal detection theory is the distinction between the representation and the decision rule, not the shape of the underlying distributions, and two types of nonnormal distributions have been important. One such distribution is the *logistic*, which has a shape very similar to the normal but is computationally more tractable. Logistic models have been used for complex experimental designs, and for statistical analysis. Another family of distributions, whose shape is *rectangular*, leads to rectilinear ROCs and can be interpreted as support for 'thresholds' that divide discrete perceptual states. Such curves are unusual, experimentally, but have been found in studies of recognition memory (Yonelinas and Jacoby 1994). Models using threshold assumptions have been developed for complex experimental situations, such as source monitoring in memory; in this context, they are called multinomial models (see *Discrete State Models of Information Processing*).

5. *Identification and Classification Along a Single Dimension*

Detection theory is extended easily to experiments with multiple stimuli differing on a single dimension. For example, a set of tones differing only in intensity may be presented one at a time for *identification* (the observer must identify the exact stimulus presented) or *classification* (the observer must sort the set into two or more categories). The representation is assumed to be a set of distributions along a decision axis, with $m-1$ criteria used to assign responses to m categories.

SDT analysis allows the calculation of d' (or a related measure, if equal variances are not assumed) for any stimulus pair, and an interesting question concerns the relation between this value and that found in a discrimination task using just those two stimuli. According to a theory of Braida and Durlach (1988), identification d' will be lower than discrimination d' by an amount that reflects memory constraints in the former task. The predicted relation is qualitatively correct for many stimulus dimensions, and can be quantitatively predicted for tones differing in intensity.

6. *Complex Discrimination Designs and Multidimensional Detection Theory*

The one-interval design for measuring discrimination is appropriate for some applications (it seems quite natural in recognition memory, for example, where the $S+N$ trials are studied items and the N trials are foils), but other methods are often preferred. One popular technique is *two-alternative forced-choice* (2AFC), in which *both* stimuli are presented on every trial and the observer must say whether N or $S+N$ occurred first. Another method is *same-different*, in which there are again two intervals, but each interval can contain either of the two stimuli and the observer must say whether the two stimuli are the same or different.

These and many other designs can be analyzed by assuming that each interval is represented by a value on a separate, independent perceptual dimension. Although d' is a characteristic of the stimuli and the observer, and therefore constant across paradigms, performance as measured by percent correct [$p(c)$] varies widely (Macmillan and Creelman 1991). For example, Sect. 2 showed that if $d' = 1.5$, $p(c) = 0.77$ in the one-interval task; but in 2AFC the same observer would be expected to score $p(c) = 0.86$, and in the same-different task as low as $p(c) = 0.61$. The relation between one-interval and 2AFC was one of the first

predictions of SDT to be tested (Green and Swets 1966), whereas that between same-different and other paradigms is a more recent topic of investigation. An interesting aspect of same-different performance is that two distinct decision strategies are available, and their use depends on aspects of the experimental design and stimulus set (Irwin and Francis 1995)

Many identification and classification experiments use stimulus sets that themselves vary in multiple dimensions perceptually, and SDT analysis can be extended to this more general case. Multidimensional detection theory has been particularly helpful in clarifying questions about independence (see Ashby 1992) (see *Signal Detection Theory: Multidimensional*).

7. *Comparing Observed with Ideal or Theoretically-derived Performance*

Detection theory has provided a forum for asking a slate of questions about the optimality of performance. One set of issues concerns the degree to which human behavior falls below the ideal level set by stimulus variability. Early auditory experiments using external noise revealed that human 'efficiency' was often quite high, though never perfect. A different type of application concerns the theoretical prediction of the ratio between the $S+N$ and N standard deviations, which can be derived from the shape of the ROC shape. For example, this ratio should be less than unity for detection experiments in which signal parameters are not known by the observer (Graham 1989). Ratcliff et al. (1994) showed that for an important class of recognition memory models, the ratio should be approximately 0.8, as it is empirically.

Some strong assumptions about the decision process that is at the heart of SDT have been tested. In general, these are well supported for one-dimensional representations, but complications sometimes arise for higher dimensionality. Do observers establish fixed criteria? Experiments requiring discrimination between *distributions* of stimuli have shown that they do when the variability is in one dimension and with adequate experience in two dimensions as well. Do observers set 'optimal' criteria? Actual criterion settings are often conservative compared to ideal ones predicted by Bayes's Theorem. In a multidimensional task, observers appear to adopt appropriate decision boundaries when those are simple in form (e.g., linear), but fall short of optimality when more complex rules are required. Does the representation remain fixed in the face of task variation? In one dimension, this amounts to the original, often-replicated datum that d' (or some sensitivity measure) is unchanged when criterion changes. In more than one dimension, attentional effects can alter the representation (Nosofsky 1986), and the question of what *is* invariant in these cases is still unresolved.

See also: Decision Theory: Classical; Perception: History of the Concept; Psychophysical Theory and Laws, History of; Signal Detection Theory, History of; Signal Detection Theory: Multidimensional; Time Perception Models

Bibliography

Ashby F G (ed.) 1992 *Multidimensional Models of Perception and Cognition*. Erlbaum, Hillsdale, NJ

Braida L D, Durlach N I 1988 Peripheral and central factors in intensity perception. In: Edelman G M, Gall W E, Cowan W M (eds.) *Auditory Function*. Wiley, New York, pp. 559–83

Graham N V 1989 *Visual Pattern Analyzers*. Oxford University Press, Oxford, UK

Green D M, Swets J A 1966 *Signal Detection Theory and Psychophysics*. Wiley, New York

Irwin R J, Francis M A 1995 Perception of simple and complex visual stimuli: Decision strategies and hemispheric differences in same-different judgments. *Perception* **24**: 787–809

Macmillan N A, Creelman C D 1991 *Detection Theory: A User's Guide*. Cambridge University Press, New York

Nosofsky R M 1986 Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* **115**: 39–57

Ratcliff R, McKoon G, Tindall M 1994 Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**: 763–85

Swets J A 1986 Form of empirical ROCs in discrimination and diagnostic tasks. *Psychological Bulletin* **99**: 181–98

Swets J A, Pickett R M 1982 *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York

Yonelinas A P, Jacoby L L 1994 Dissociations of processes in recognition memory: Effects of interference and of response speed. *Canadian Journal of Experimental Psychology* **48**: 516–34

N. A. Macmillan

Signal Detection Theory, History of

1. *Introduction and Scope*

Signal detection theory (SDT) sprouted from World War II research on radar into a probability-based theory in the early 1950s. It specifies the optimal observation and decision processes for detecting electronic signals against a background of random interference or noise. The engineering theory, culminating in the work of Wesley W. Peterson and Theodore G. Birdsall (Peterson et al. 1954), had foundations in mathematical developments for theories of statistical inference, beginning with those advanced by Jerzy

Neyman and E. S. Pearson (1933). SDT was taken into psychophysics, then a century-old branch of psychology, when the human observer's detection of weak signals, or discrimination between similar signals, was seen by psychologists as a problem of inference. In psychology, SDT is a model for a theory of how organisms make fine discriminations and it specifies model-based methods of data collection and analysis. Notably, through its analytical technique called the receiver operating characteristic (ROC), it separates sensory and decision factors and provides independent measures of them. SDT's approach is now used in many areas in which discrimination is studied in psychology, including cognitive as well as sensory processes. From psychology, SDT and the ROC came to be applied in a wide range of practical diagnostic tasks, in which a decision is made between two confusable alternatives (Swets 1996) (see also *Signal Detection Theory*).

2. *Electronic Detection and Statistical Inference*

Central to both electronics and statistics is a conception of a pair of overlapping, bell-shaped, probability (density) functions arrayed along a unidimensional variable, which is the weight of evidence derived from observation. In statistical theory, the probabilities are conditional on the null hypothesis or the alternative hypothesis, while in electronic detection theory they are observational probabilities conditional on noise-alone or signal-plus-noise. The greater the positive weight of evidence, the more likely a signal, or significant experimental effect, is present. A cutpoint must be set along the variable to identify those observed values above the cutpoint that would lead to rejection of the null hypothesis or acceptance of the presence of the designated signal. Errors are of commission or omission: type I and type II errors in statistics, and false alarms and misses in detection. Correct outcomes in detection are hits and correct rejections. Various decision rules specify optimal cutpoints for different decision goals. The weight of evidence is unidimensional because just two decision alternatives are considered. Optimal weights of evidence are monotone increasing with the *likelihood ratio*, the ratio of the two overlapping bell-shaped probability functions. It is this ratio, and not the shapes of the functions, that is important.

3. *Modern Signal Detection Theory*

In the early 1950s, Peterson and Birdsall, then graduate students in electrical engineering and mathematics, respectively, at the University of Michigan, developed the general mathematical theory of signal detection that is still current. In the process they devised the ROC—a graphical technique permitting measurement of two independent aspects of detection performance: (a) the location of the observer's cutpoint or decision criterion and (b) the observer's sensitivity, or ability to discriminate between signal-plus-noise and noise-alone, irrespective of any chosen cutpoint. The two measures characterize performance better than the single one often used, namely, the signal-to-noise ratio (SNR) expressed in energy terms that is necessary to provide a 50 percent hit (correct signal acceptance) probability at a fixed decision cutpoint, say, one yielding a conditional false-alarm probability of 0.05. A curve showing conditional hit probability versus SNR is familiar in detection theory and is equivalent to the power function of statistical theory. These curves can be derived as a special case of the ROCs for the SNR values of interest.

4. *The ROC*

The ROC is a plot of the conditional hit (true–positive) proportion against the conditional false-alarm (false–positive) proportion for all possible locations of the decision cutpoint. Peterson's and Birdsall's SDT specifies a form of ROC that begins at the lower left corner (where both proportions are zero) and rises with smoothly decreasing slope to the upper right corner (where both proportions are 1.0)—as the decision cutpoint varies from very strict (at the right end of the decision variable) to very lenient (at the left end). This is a 'proper' ROC, and specifically not a 'singular' ROC, or one intersecting axes at points between 0 and 1.0. The locus of the curve, or just the proportion of area beneath it, gives a measure of discrimination capacity. An index of a point along the curve, for example, the slope of the curve at the point, gives a measure of the decision cutpoint that yielded that point (the particular slope equals the criterial value of likelihood ratio). SDT specifies the discrimination capacity attainable by an 'ideal observer' for any SNR for various practical combinations of signal and noise parameters (e.g., signal specified exactly and signal specified statistically) and hence the human observer's efficiency can be calculated under various signal conditions. SDT also specifies the optimal cutoff as a function of the signal's prior probability and the benefits and costs of the decision outcomes.

5. *Psychology's Need for SDT and The ROC*

Wilson P. Tanner, Jr., and John A. Swets, graduate students in psychology at Michigan in the early 1950s, became aware of Peterson's and Birdsall's work on the same campus as it began. Tanner became acquainted through his interest in mathematical and electronic concepts as models for psychological and neural processes and Swets was attracted because of his interest in quantifying the 'instruction stimulus' in

psychophysical tasks, in order to control the observer's response criterion. As they pursued studies in sensory psychology, these psychologists had become dissatisfied with psychophysical and psychological theories based on overlapping bell-shaped functions that assumed fixed decision cutpoints and incorporated measures only of discrimination capacity, which were possibly confounded by attitudinal or decision processes. Prominent among such theories were those of Gustav Theodor Fechner (1860), Louis Leon Thurstone (1927), and H. Richard Blackwell (1963); see Swets (1996, Chap. 1).

Fechner and Thurstone conceived of a symmetrical cutpoint, where the two distributions cross, because they presented two stimuli on each trial with no particular valence; that is, both stimuli were 'signals' (lifted weights, samples of handwriting). Fechner also studied detection of single signals, but did not compare them to a noise-alone alternative and his theory was essentially one of variable representations of a signal compared to an invariant cutpoint. This cutpoint was viewed as a physiologically determined sensory threshold (akin to all-or-nothing nerve firing) and hence the values of the sensory variable beneath it were thought to be indistinguishable from one another.

Blackwell's task and model were explicitly of signals in noise, but the observer was thought to have a fixed cutpoint near the top of the noise function, for which the false-alarm probability was negligible, and, again, the values below that cutpoint were deemed indistinguishable. For both threshold theorists, the single measure of performance was the signal strength required to yield 50 percent correct positive responses, like the effective SNR of electronic detection theory, and it was taken as a statistical estimate of a sensory threshold. In psychology, the curve relating percentage of signals detected to signal strength is called the psychometric function. There had been some earlier concern in psychology for the effect of the observer's attitude on the measured threshold, e.g., Graham's (1951) call for quantification of instruction stimuli in the psychophysical equation, but no good way to deal with non-sensory factors had emerged.

In 1952, Tanner and Swets joined Peterson and Birdsall as staff in a laboratory of the electrical engineering department called the Electronic Defense Group. They were aware of then-new conceptions of neural functioning in which stimulus inputs found an already active nervous system and neurons were not lying quiescent till fired at full force. In short, neural noise as well as environmental noise was likely to be a factor in detection and then the observer's task is readily conceived as a choice between statistical hypotheses. Though a minority idea in the history of psychophysics (see Corso 1963), the possibility that the observer deliberately sets a cutpoint on a continuous variable (weight of evidence) seemed likely to Tanner and Swets. A budding cognitive psychology (e.g., 'new look' in perception) supported the notion of extrasensory determinants of perceptual phenomena, as represented in SDT by expectancies (prior probabilities) and motivation (benefits and costs).

6. *Psychophysical Experiments*

The new SDT was first tested in Swets's doctoral thesis in Blackwell's vision laboratory (Tanner and Swets 1954; Swets et al. 1961–see Swets 1964, Chap. 1). It was then tested with greater control of the physical signal and noise parameters in new facilities for auditory research in the electrical engineering laboratory. Sufficiently neat empirical ROCs were obtained in the form of the curved arc specified by SDT. ROCs found to be predicted by Blackwell's 'high-threshold' theory—straight lines extending from some point along the left axis, that depended on signal strength, to the upper right hand corner—clearly did not fit the data. Other threshold theories, e.g., 'low-threshold' theories, did not fare much better in experimental tests (Swets 1961–see Swets 1964, Chap. 4). It was recognized later that linear ROCs of slope = 1, intersecting left and upper edges of the graph symmetrically, were predicted by several other measures and their implicit models (Swets 1996, Chap. 3) and also gave very poor fits to sensory data (and other ROC data to be described; Swets 1996, Chap. 2). After a stint as observer in Swets's thesis studies, David M. Green joined the engineering laboratory as an undergraduate research assistant. He soon became a full partner in research, and co-authored a laboratory technical report with Tanner and Swets in 1956 that included the first auditory studies testing SDT. Birdsall has collaborated in research with the psychologists over the following decades and commented on a draft of this article.

The measure of discrimination performance used first was denoted d' ('d prime') and defined as the difference between the means of two implicit, overlapping, normal (Gaussian) functions for signal-plus-noise and noise-alone, of equal variance, divided by their common standard deviation. A value of d' can be calculated for each ROC point, using the so-called single-stimulus *yes–no* method of data collection. Similarly, a value of d' can be calculated for an observer's performance under two other methods of data collection: the multiple-stimulus *forced-choice* method (one stimulus being signal-plus noise and the rest noise-alone) and the single-stimulus *confidence-rating* method. Experimental tests showed that when the same signal and nonsignal stimuli were used with the three methods, essentially the same values of d' were obtained. The measures specified by threshold theories had not achieved that degree of consistency or internal validity.

For many, the most conclusive evidence favoring SDT's continuous decision variable and rejecting the high-threshold theory came from an experiment sug-

gested by Robert Z. Norman, a Michigan mathematics student. Carried out with visual stimuli, the critical finding was that a second choice made in a four-alternative forced-choice test, when the first choice was incorrect, was correct with probability greater than 1/3 and that the probability of it being correct increased with signal strength (Swets et al. 1961, Swets 1964, Chap. 1).

Validating the rating method, as mentioned above, had the important effect of providing an efficient method for obtaining empirical ROCs. Whereas under the yes–no method the observer is induced to set a different cutpoint in each of several observing sessions (each providing an ROC point), under the rating method the observer effectively maintains several decision cutpoints simultaneously (the boundaries of the rating categories) so that several empirical ROC points (enough to define a curve) can be obtained in one observing session.

7. *Dissemination of SDT in Psychology*

Despite the growing evidence for it, acceptance of SDT was not rapid or broad in psychophysical circles. This was partly due to threshold theory having been ingrained in the field for a century (as a concept and a collection of methods and measures), and probably because SDT arose in an engineering context (and retained the engineers' terminology), the early visual data were noisy, and the first article (Tanner and Swets 1954) was cryptic. Dissemination was assisted when J. C. R. Licklider brought Swets and Green to the Massachusetts Institute of Technology where they participated in Cambridge's hotbed of psychophysics and mathematical psychology, and where offering a special summer course for postgraduates led to a published collection of approximately 35 of the articles on SDT in psychology that had appeared by then, along with tables of d' (Swets 1964). Licklider also hired both of them part-time at the research firm of Bolt Beranek and Newman Inc., where they obtained contract support from the National Aeronautics and Space Administration to write a systematic textbook (Green and Swets 1966). Meanwhile, Tanner mentored a series of exceptional graduate students at Michigan. All were introduced by Licklider to the Acoustical Society of America where they enjoyed a feverishly intense venue for discussion of their ideas at biannual meetings and in the Society's journal. Another factor was the active collaboration with Tanner of James P. Egan and his students at Indiana University.

8. *Extensions of SDT in Psychology*

Egan recognized the potential for applying SDT in psychology beyond the traditional psychophysical tasks. He extended it to the less tightly defined vigilance task, that is, the practical military and industrial observing task of 'low-probability watch' (Egan et al. 1961–see Swets 1964, Chap. 15), and also to speech communication (Egan and Clarke 1957–see Swets 1964, Chap. 30). He also brought SDT to purely cognitive tasks with experiments in recognition memory (see Green and Swets 1966, Sect. 12.9). Applications were then made by others to those tasks, and also to tasks of conceptual judgment, animal discrimination and learning, word recognition, attention, visual imagery, manual control, and reaction time (see Swets 1996, Chap. 1). A broad sample of empirical ROCs obtained in these areas indicates their very similar forms (Swets 1996, Chap. 2). An Annual Review chapter reviews applications in clinical psychology, for example, to predicting acts of violence (McFall and Treat 1999).

9. *Applications in Diagnostics*

An analysis of performance measurement in information retrieval suggested that two-by-two contingency tables for any diagnostic task were grist for SDT's mill (Swets 1996, Chap. 9). This idea was advanced considerably by Lusted's (1968) application to medical diagnosis. A standard protocol for evaluating diagnostic performance via SDT methods, with an emphasis on medical images, was sponsored by the National Cancer Institute (Swets and Pickett 1982). By 2000, 'ROC' is specified as a key word in over 1000 medical articles each year, ranging from radiology to blood analysis. Other diagnostic applications are being made to aptitude testing, materials testing, weather forecasting, and polygraph lie detection (Swets 1996, Chap. 4).

A recent development is to use SDT to improve, as well as to evaluate, diagnostic accuracy. Observers' ratings of the relevant dimensions of a diagnostic task—e.g., perceptual features in an X-ray—are merged in a manner specified by the theory to yield an optimal estimate of the probability that a 'signal' is present (Swets 1996, Chap. 8). The latest work both in psychology and diagnostics has been described didactically (Swets 1998).

See also: Psychophysical Theory and Laws, History of; Signal Detection Theory; Signal Detection Theory: Multidimensional

Bibliography

Blackwell H R 1963 Neural theories of simple visual discriminations. *Journal of the Optical Society of America* **53**: 129–60

Corso J F 1963 A theoretico-historical review of the threshold concept. *Psychological Bulletin* **60**: 356–70

Fechner G T 1860 *Elemente der Psychophysik* Breitkopf & Hartel, Leipzig, Germany [English translation of Vol. 1 by Adler H E 1966. In: Howes D H, Boring E G (eds.) *Elements of Psychophysics*. Holt, Rinehart and Winston, New York]
Graham C H 1951 Visual perception. In: Stevens S S (ed.) *Handbook of Experimental Psychology*. Wiley, New York
Green D M, Swets J A 1966 *Signal Detection Theory and Psychophysics*. Wiley, New York. Reprinted 1988 by Peninsula, Los Altos Hills, CA
Lusted L B 1968 *Introduction to Medical Decision Making*. C. C. Thomas, Springfield, IL
McFall R M, Treat T A 1999 Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology* **50**: 215–41
Neyman J, Pearson E S 1933 On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London* **A231**: 289–311
Peterson W W, Birdsall T G, Fox W C 1954 The theory of signal detectability. *Transactions of the Institute of Radio Engineers Professional Group on Information Theory*, PGIT 4: 171–212. Also In: Luce R D, Bush R R, Galanter E 1963 (eds.) *Readings in Mathematical Psychology*. Wiley, New York, Vol. 1
Swets J A (ed.) 1964 *Signal Detection and Recognition by Human Observers*. Wiley, New York. Reprinted 1988 by Peninsula, Los Altos Hills, CA
Swets J A 1996 *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. L. Erlbaum Associates, Mahwah, NJ
Swets J A 1998 Separating discrimination and decision in detection, recognition, and matters of life and death. In: Osherson D (series ed.) Scarborough D, Sternberg S (vol. eds.) *An Invitation to Cognitive Science: Vol 4, Methods, Models, and Conceptual Issues*. MIT, Cambridge, MA
Swets J A, Pickett R M 1982 *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York
Tanner W P Jr, Swets J A 1954 A decision making theory of visual detection. *Psychological Review* **61**: 401–9
Thurstone L L 1927 A law of comparative judgment. *Psychological Review* **34**: 273–86

J. A. Swets

Signal Detection Theory: Multidimensional

Multidimensional signal detection theory (MSDT) is an extension of signal detection theory (SDT) to more than one dimension, with each dimension representing a different source of information along which the 'signal' is registered. The development of MSDT was motivated by psychologists who wanted to study information processing of more realistic, multidimensional stimuli, for example, the shape and color of geometric objects, or pitch, loudness, and timbre of tones. The typical questions of interest are: Are the dimensions (e.g., color and shape of an object) processed independently, or does the perception of one dimension (e.g., color) depend on the perception of the second (shape), and if so, what is the nature of their dependence? In MSDT, several types of independence of perceptual dimensions are defined and related to two sets (marginal and conditional) of sensitivity, d′, and response bias, β, parameters. This article gives the definitions, and outlines the uses in studying information processing of dimensions. MSDT is also known as General Recognition Theory (Ashby and Townsend 1986) or Decision Bound Theory (e.g., Ashby and Maddox 1994).

1. Introduction and Notation

Stimulus dimensions, A, B,…, represent different types of information about a class of stimuli, for example, in a two-dimensional (2-D) case, the shape (A) and color (B) of objects. The researcher specifies the levels on each dimension to study (e.g., square and rectangle are levels of shape, and blue, turquoise, and green are levels of color). A multidimensional 'signal' is stimulus A_iB_j, where the subscript indicates the specific level on the corresponding dimension (e.g., a green square is A_1B_3). For each physical stimulus dimension A (B,…), there is assumed a corresponding unique psychological or perceptual dimension X (Y,…).

The minimal stimulus set considered in this article consists of four 2-D stimuli, $\{A_1B_1, A_1B_2, A_2B_1, A_2B_2\}$, constructed by combining two levels on dimension A ($i = 1, 2$) with each of two levels on dimension B ($j = 1, 2$). For example, the set constructed from two levels of color (blue, green) and two levels of shape (square, rectangle) is {blue square, blue rectangle, green square, green rectangle}.

On a single trial, one stimulus is selected from the set at random, and it is presented to the observer whose task is to identify the stimulus. The presentation of a 2-D stimulus results in a percept, a point (x, y) in X-Y-perceptual space. The observer's response is based on the percept's location relative to the observer's placement of decision criterion bounds in the perceptual space, typically one bound per dimension, denoted c_A, c_B, … . These bounds divide the perceptual space into response regions such that if, for example, (x, y) falls above c_A and below c_B, the (2-D) response would be '21' (or any response that uniquely identifies it as A_2B_1).

As in unidimensional SDT, multiple presentations of a single stimulus are assumed to lead to a distribution of percepts, and one distribution (usually Gaussian) is assumed for each stimulus. An example with four 2-D stimuli is illustrated in Fig. 1. The bivariate joint densities, $f_{A_iB_j}(x, y)$, are represented more easily as equal-density contours (Fig. 1(b)), that are obtained by slicing off the tops of the joint densities in Fig. 1(a) at a constant height, and viewing the space

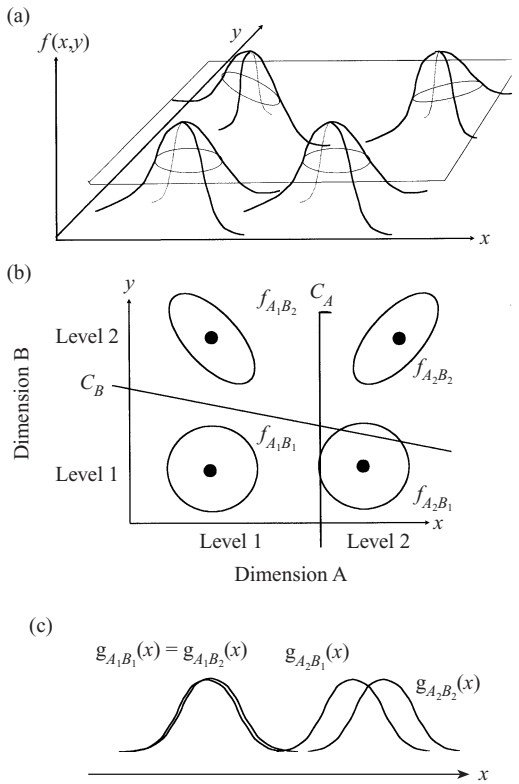

Figure 1
(a) The perceptual space showing four joint bivariate densities, $f_{A_iB_j}(x, y)$, that represent all possible (theoretical) perceptual effects for each of the four 2-D stimuli A_iB_j; (b) Equal density contours of the bivariate densities in (a) and the decision bounds c_A and c_B; (c) Marginal densities, $g_{A_iB_j}(x)$, representing perceptual effects for each stimulus A_iB_j on dimension A only. In this example, $g_{A_1B_1}(x) \neq g_{A_1B_2}(x)$ but $g_{A_2B_1}(x) g_{A_2B_2}(x)$

from 'above.' The joint density for each stimulus A_iB_j yields: (a) Marginal densities on each dimension, $g_{A_iB_j}(x)$ and $g_{A_iB_j}(y)$, obtained by integrating over all values of the second dimension (see Fig. 1(c) for marginal densities $g_{A_iB_j}(x)$ for dimension A); and (b) Conditional densities for each dimension (not shown), conditioning on the observer's response to the second dimension, and obtained by integrating only over values either *above* (or *below*) the decision bound on the second dimension, for example, $g_{A_iB_j}(x|y > c_B)$.

An observer's responses are tabulated in a 4×4 confusion matrix, a table of conditional probability estimates, $P(R_p|A_iB_j)$, that are proportions of trials that response R_p was emitted when stimulus A_iB_j was presented. These proportions estimate the volumes under each joint density in each response region. For example, the proportion of trials on which the observer responded '11' when in fact stimulus A_1B_2 was presented is

$$P(R_{11}|A_1B_2) = \int_{\infty}^{c_A} \int_{\infty}^{c_B} f_{A_1B_2}(x, y)\,\mathrm{d}x\,\mathrm{d}y \qquad (1)$$

2. *Definitions of Independence of Dimensions*

Several definitions of independence of dimensions are defined within MSDT. These are: perceptual separability (PS), decisional separability (DS), and perceptual independence (PI).

2.1 *Perceptual Separability (PS) and Decisional Separability (DS)*

Dimension A is PS from dimension B if the perceptual effects of A are equal across levels of dimension B, that is, $g_{A_iB_1}(x) = g_{A_iB_2}(x)$, for all values of x and both $i = 1$ and $i = 2$. Similarly, dimension B is perceptually separable from dimension A if $g_{A_1B_j}(x) = g_{A_2B_j}(x)$, for all y, and $j = 1, 2$.

DS is a similar type of independence of the *decisions* that the observer makes about the stimuli. DS holds for dimension A (or B) when the decision bound c_A (or c_B) is parallel to the Y- (or X-) perceptual axis.

PS and DS are global types of independence defined for perceptual dimensions across levels of the second dimension *for all stimuli in the set*. They can also be asymmetrical relations. In Fig. 1(b), for example, PS holds for dimension B across levels of dimension A, but PS fails for dimension A across levels of B; also, DS holds for dimension A but fails for dimension B.

2.2 *Perceptual Independence (PI)*

PI is distinguished from PS in that it is a local form of independence among dimensions *within a single stimulus*. PI holds within stimulus A_iB_j if

$$f_{A_iB_j}(x, y) = g_{A_iB_j}(x)\, g_{A_iB_j}(y)$$

for all (x, y). This is statistical independence, and it is a symmetric relation in a given stimulus.

The shape of an equal density contour describes the correlation among the perceived dimensions in one stimulus. A circle indicates PI in a stimulus with variances equal on the two dimensions. An ellipse with its major and minor axes parallel to the perceptual axes also indicates PI, but with unequal variances on the two dimensions. A tilted ellipse indicates failure of

PI, with the direction of the tilt (positive or negative slope of the major axis) indicating the direction of the dependence. In Fig. 1(b), stimuli A_1B_1 and A_2B_1 show PI, stimulus A_2B_2 shows a positive dependence, and stimulus A_1B_2 shows a negative dependence.

3. Marginal and Conditional d′s and βs

As in unidimensional SDT, sensitivity, d′, and response bias, β, are parameters that describe different characteristics of the observer. Two sets of these parameters are defined: (a) marginal d′s and βs, and (b) conditional d′s and βs. The marginal and conditional d′s and βs for the four-stimulus set discussed here can be generalized to any number of dimensions and any number of levels: for details, see Kadlec and Townsend (in Ashby 1992, Chap. 8).

3.1 Marginal d′s and βs

Marginal d′s and βs are defined using the marginal densities, $g_{A_iB_j}(x)$ and $g_{A_iB_j}(y)$. Let the mean vector and variance-covariance matrix of $f_{A_iB_j}(x, y)$, for stimulus $A_i\mathrm{B}_j$, be denoted by

$$\vec{\mu}_{(A_iB_j)} = \begin{bmatrix} \mu_{x(A_iB_j)} \\ \mu_{y(A_iB_j)} \end{bmatrix}$$

and

$$\Sigma_{(A_iB_j)} = \begin{bmatrix} \sigma^2_{x(A_iB_j)} & \sigma_{xy(A_iB_j)} \\ \sigma_{xy(A_iB_j)} & \sigma^2_{y(A_iB_j)} \end{bmatrix} \quad (2)$$

respectively. The marginal d′ for dimension A at level j of dimension B is

$$\mathrm{d}'(A \text{ at } B_j) = \frac{\mu_{x(A_2B_j)} - \mu_{x(A_1B_j)}}{\sigma_{x(A_1B_j)}} \quad \text{for} \quad j = 1, 2 \quad (3)$$

The subscript x indicates that for a d′ on dimension A, only the X-coordinates of the mean vectors and the standard deviation along the X-perceptual axis are employed. As in unidimensional SDT, the standard deviation of the 'noise' stimulus (here indicated as level 1 of dimension A) is used, and without loss of generality, it is assumed that $\mu_{(A_1B_1)} = [0\ 0]^T$ (^{T} denotes transpose) and $\Sigma_{(A_1B_1)} = \mathrm{I}$ (i.e., $\sigma^2_x = \sigma^2_y = 1$, and $\sigma_{xy} = 0$). Similarly, the marginal d′s for dimension B at level i of dimension A is

$$\mathrm{d}'(B \text{ at } A_i) = \frac{\mu_{y(A_iB_2)} - \mu_{y(A_iB_1)}}{\sigma_{y(A_iB_1)}} \quad \text{for} \quad i = 1, 2 \quad (4)$$

In a parallel fashion, marginal βs are defined as

$$\beta(A \text{ at } B_j) = \frac{g_{A_2B_j}(c_A)}{g_{A_1B_j}(c_A)} \quad \text{for} \quad j = 1, 2$$

and

$$\beta(B \text{ at } A_i) = \frac{g_{A_iB_2}(c_B)}{g_{A_iB_1}(c_B)} \quad \text{for} \quad i = 1, 2 \quad (5)$$

where c_A and c_B are the decision bounds for dimension A and B, respectively.

3.2 Conditional d′s and βs

In the four-stimulus set, two pairs of conditional d′s and βs are defined *for each dimension*, conditioning on the observer's response on the second. Using the terminology of unidimensional SDT for simplicity, with level 1 indicating the 'noise' stimulus and level 2 the 'signal' stimulus, one pair of conditional d′s for dimension A, with B presented at level 1, is: (a) d′(A|'correct rejection' on B)—conditioned on the observer's *correct* response that B was at level 1, defined using parameters of

$$g_{A_iB_1}(x|y < c_B),$$

i = 1 and 2 in Eqn. (3); and (b) d′(A|'false alarm' on B)—conditioned on the *incorrect* response that B was at level 1, using parameters of

$$g_{A_iB_1}(x|y > c_B)$$

in Eqn. (3). The second pair of conditional d′s for dimension A, with B presented at level 2, is: (c) d′(A|'hit' on B)—conditioned on the *correct* response that B was at level 2; and (d) d′(A|'miss' on B)—conditioned on the *incorrect* response that B was at level 1. Similar pairs are defined for conditional βs, and for d′s and βs for dimension B conditional on dimension A responses.

4. Using the d′s and βs to Test Dimensional Interactions

These d′ and β parameters are linked theoretically to PS, DS, and PI (Kadlec and Townsend 1992). Their estimates, obtained from the confusion matrix proportions (Eqn. (1)), can thus be used to draw tentative inferences about dimensional interactions (Kadlec 1999). DS for dimension A (or B) is inferred when c_A (c_B) estimates are equal at all levels of dimension B (A). Equal marginal d′s for a dimension across levels of the other are consistent with PS on that dimension. A stronger inference for PS on both dimensions can be drawn if the densities form a rectangle (but not a

parallelogram), and this condition is testable if DS and PI hold (Kadlec and Hicks 1998; see also Thomas 1999 for new theoretical developments). When DS holds, evidence for PI is gained if conditional d′s are found to be equal: Conversely, unequal conditional d′s indicate PI failure, and the direction of the inequality provides information about the direction of the dimensional dependencies in the stimuli. Exploring these relationships and looking for stronger tests of PS and PI is currently an active area of research.

MSDT is a relatively new development, and with its definitions of independence of dimensions, it is a valuable analytic and theoretical tool for studying interactions among stimulus dimensions. Current applications include studies of (a) human visual perception (of symmetry, Gestalt 'laws' of perceptual grouping, and faces); and (b) source monitoring in memory. It can be used in any research domain where the study of dimensional interactions is of interest.

See also: Memory Psychophysics; Multidimensional Scaling in Psychology; Psychometrics; Psychophysics; Signal Detection Theory; Signal Detection Theory, History of; Visual Perception, Neural Basis of

Bibliography

Ashby F G (ed.) 1992 *Multidimensional Models of Perception and Cognition*. Hillsdale, Erlbaum, NJ

Ashby F G, Maddox W T 1994 A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology* **38**: 423–66

Ashby F G, Townsend J T 1986 Varieties of perceptual independence. *Psychology Review* **93**: 154–79

Kadlec H 1999 MSDA-2: Updated version of software for multidimensional signal detection analyses. *Behavior Research Methods, Instruments & Computers* **31**: 384–5

Kadlec H, Hicks C L 1998 Invariance of perceptual spaces. *Journal of Experimental Psychology: Human Perception & Performance* **24**: 80–104

Kadlec H, Townsend J T 1992 Implications of marginal and conditional detection parameters for the separabilities and independence of perceptual dimensions. *Journal of Mathematical Psychology* **36**: 325–74

Thomas R D 1999 Assessing sensitivity in a multidimensional space: Some problems and a definition of a general d′. *Psychonomic Bulletin & Review* **6**: 224–38

H. Kadlec

Significance, Tests of

A recent article on the health benefits of herbal tea reported its use leading to a decreased incidence of insomnia in an experiment conducted at a sleep disorders clinic. Patients at the clinic were randomly assigned to daily consumption of herbal tea or a caffeine-free beverage of their choice, and were followed up for 10 months. The reported improvement was stated to be 'statistically significant ($p < 0.05$).' The implication intended was that the improvement should be attributed to the beneficial effects of the tea.

In fact, this article does not really exist, but examples of this sort are reported regularly in science articles in the press, and are very common in journal publications in a number of fields in science and social science. It has even been argued, in the press, that 'statistical significance' has become so widely used to give an *imprimatur* of scientific acceptability that it is in fact misused and should be abandoned (Matthews 1998).

A test of significance assesses the agreement between the data and an hypothesized statistical model. The magnitude of the agreement is expressed as an observed level of significance, or p-value, which is the probability of obtaining data as or more extreme than the observed data, if the hypothesized model were true. A very small p-value suggests that either the observed data are not compatible with the hypothesized model or an event of very small probability has been observed. A large p-value indicates that the observed data are compatible with the hypothesized model. Since the p-value is a probability, it must be between 0 and 1, and a very common convention is to declare values smaller than 0.05 as 'small' or 'statistically significant' and values larger than 0.05 as 'not statistically significant.' In this section we shall try to explain the historical rationale of this very arbitrary cut-off point.

To make these vague definitions more concrete, it is necessary to consider statistical models and the notion of a (simplifying) hypothesis within that model. The theory of this is outlined in Sect. 2 with some more advanced topics considered in Sect. 3. This section concludes with a highly idealized example to convey the idea of data being inconsistent with an hypothesized model.

Example 1. Students in a statistics class partake in an activity to assess their ability to distinguish between two competing brands of cola, and to identify their preferred brand from taste alone. Each of the 20 students expresses a preference for one brand or the other, but just one student claims to be able to discriminate perfectly between the two. Twenty cups of each brand are prepared by the instructor and labelled '1' and '2.' Each student is to record which label corresponds to Brand A. The result is that 12 students correctly the competing brands identify, although not the student who claimed a perfect ability to discriminate.

The labelling of the cups as 1 or 2 by the instructor was completely random, i.e., cup 1 was equally likely to contain Brand A or B. The students did not discuss their opinions with their classmates, and the taste testing was completed fairly quickly. Under these conditions, it is plausible that each student has a

Table 1
Probability of zero or one mistakes in n independent Bernoulli trials with probability of a mistake = 0.5

n	Probability
5	0.1875
6	0.1094
7	0.0625
8	0.0352
9	0.0195
10	0.0107
11	0.0059
12	0.0032
13	0.0017
14	0.0009
15	0.0002

probability of $\frac{1}{2}$ of identifying the brands correctly simply by guessing, so that about 10 students would identify the brands correctly with no discriminatory ability at all. That 12 students did so does not seem inconsistent with guesswork, and the p-value helps to quantify this. The probability of observing 12 or more correct results if one correct result has probability $\frac{1}{2}$ and the guesses are independent can be computed by the binomial formula as

$$\left\{\binom{20}{12}+\binom{20}{13}+\ldots+\binom{20}{20}\right\}(\tfrac{1}{2})^{20} = 0.34. \qquad (1)$$

This is not at all an unlikely event, so there is no evidence from these data that the number of correct answers could not have been obtained by guessing: in more statistical language, assuming a binomial model, the observed data is consistent with probability of success $\frac{1}{2}$.

The student who claimed to have perfect discrimination, but actually guessed incorrectly, argued that her abilities should not be dismissed on the basis of one taste test, so the class carried out some computations to see what the p-value for the same observed data would be if the number of trials was increased. The probability of one or zero mistakes in a set of n trials for various values of n, is given in Table 1. From this we see that, for example, one or no mistakes in five trials is consistent with guesswork, but the same result in 10 trials is much less so.

In both parts of this example we assumed a model of independent trials, each of which could result in a success or failure, with constant probability of success. Our calculations also assumed this constant probability of success was 0.5. This latter restriction on the model is often called a 'null hypothesis', and the test of significance is a test of this null hypothesis; the p-value measures the consistency of the data with this null hypothesis. In many applications the null hypothesis plays the role of a conservative position that the experimenter hopes to disprove, and one reason for requiring rather small p-values before declaring statistical significance is to raise the standard of proof required to replace a relatively simple working hypothesis with one that is possibly more complex and less well understood.

As formulated here, the hypothesis being tested is that the probability of a correct choice is 0.5, and not the other aspects of the model, such as independence of the trials, and unchanging probability of success. The number of observed successes does not measure such model features; it provides information; only on the probability of success. Functions of the data that do measure such model features can be constructed, and from these significance tests that assess the fit of an assumed model; these play an important role in statistical inference as well.

1. *Model Based Inference*

1.1 *Models and Null Hypothesis*

We assume that we have a statistical model for a random variable Y taking values in a sample space $\mathscr{Y}$, described by a parametric family of densities $\{f(y;\theta);\theta\in\Theta\}$. Tests of significance can in fact be constructed in more general settings, but this framework is useful for defining the main ideas. If Y is the total number of successes in n independent Bernoulli trials with constant probability of success, then

$$f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \qquad (2)$$

$\Theta = [0, 1]$ and $\mathscr{Y} = \{0, 1,\ldots, n\}$. If Y is a continuous random variable following a normal or bell curve distribution with mean θ_1 and variance θ_2^2, then

$$f(y;\theta) = \frac{1}{\sqrt{2\pi}\theta_2}\exp\left\{-\frac{1}{2\theta_2^2}(y-\theta_1)^2\right\} \qquad (3)$$

$\Theta = \Re\times\Re^+$, and $\mathscr{Y} = \Re$. The model for n independent observations from this distribution is

$$f(y_1,\ldots,y_n;\theta) = \frac{1}{(\sqrt{2\pi})^n\theta_2^n}\exp\left\{\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(y_i-\theta_1)^2\right\} \qquad (4)$$

$\Theta = \Re\times\Re^+$, and $\mathscr{Y} = \Re^n$. For further discussion of statistical models, see *Statistical Sufficiency*; *Distributions, Statistical: Special and Discrete*; *Distributions, Statistical: Approximations*; *Distributions, Statistical: Special and Continuous*.

As noted above, we assume the model is given, and our interest is in inference about the parameter θ. While this could take various forms, a test of significance starts with a so-called *null hypothesis* about θ, of the form

$$H_0 : \theta = \theta_0 \tag{5}$$

or

$$H_0 : \theta \in \Theta_0. \tag{6}$$

In Eqn. (5) the parameter θ is fully specified, and H_0 is called a *point* null hypothesis or a *simple* null hypothesis. If θ is not fully specified, as in Eqn. (6) H_0 is called a *composite* null hypothesis. In the taste-testing examples the simple null hypothesis was $\theta = 0.5$. In the normal model, Eqn. (3), a hypothesis about the mean, such as $H_0 : \theta_1 = 0$, is composite, since the variance is left unspecified. Another composite null hypothesis is $H_0 : \theta_2 = \theta_1$, which restricts the full parameter space to a one-dimensional curve in $\mathscr{R} \times \mathscr{R}^+$.

A test is constructed by choosing a *test statistic* which is a function of the data that in some natural way measures departure from what is expected under the null hypothesis, and which has been standardized so that its distribution is known either exactly or to a good approximation under the null hypothesis. Test statistics are usually constructed so that large values indicate a discrepancy from the hypothesis.

Example 2. In the binomial model (2), the distribution of Y is completely specified by the null hypothesis $\theta = 0.5$ as

$$f(y) = \binom{n}{y} 2^{-n} \tag{7}$$

and consistency of a given observed value y_0 of y, is measured by the p-value $\sum_{y=y_0}^{n} \binom{n}{y} 2^{-n}$, the probability of observing a value as or more extreme than y_0. If y_0 is quite a bit smaller than expected, then it would be more usual to compute the p-value as $\sum_{y=0}^{y_0} \binom{n}{y} 2^{-n}$. Each of these calculations was carried out in the discussion of taste testing in Sect. 1.

Example 3. In independent sampling from the normal distribution, given in Eqn. (4), we usually test the composite null hypothesis $H_0 : \theta_1 = \theta_{10}$, by constructing the t-statistic

$$T = \sqrt{n}(\bar{Y} - \theta_{10})/S \tag{8}$$

where $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ and $S^2 = (n-1)^{-1} \sum (Y_i - \bar{Y})^2$. Under H_0, T follows a t-distribution on $n-1$ degree of freedom, and the p-value is

$$\Pr\{T \geqslant \sqrt{n}(\bar{y} - \theta_{10})/s\}$$

where $\bar{y}$ and s are the values observed in the sample. This probability needs to be computed numerically from an expression for the cumulative distribution function of the t-distribution. Historically, tables of this distribution were provided for ready reference, typically by identifying a few critical values, such as $t_{0.10}$, $t_{0.05}$, and $t_{0.01}$ satisfying $\Pr\{T_\nu \geqslant t_\alpha\} = \alpha$, where T_ν is a random variable following a t-distribution on ν degrees of freedom. It was arguably the publication of these tables that led to a focus on the use of particular fixed levels for testing in applied work.

Example 4. Assume the model specifies that $Y_1, \ldots, Y_n$ are independent, identically distributed from a distribution with density $f(\cdot)$ on $\mathfrak{R}$ and that we are interested in testing whether or not $f(\cdot)$ is a normal density:

$$H_0 : f(y) = (\sqrt{2\pi})^{-1} \mathrm{e}^{-\frac{1}{2}y^2} \tag{9}$$

or

$$H_0 : f(y) = (\sqrt{2\pi})^{-1} \theta_2^{-1} \exp(-(1/2\theta_2^2)\{y - \theta_1\}^2) \tag{10}$$

the former is a simple and the latter is a composite null hypothesis. For this problem it is less obvious how to construct a test statistic or how to choose among alternative test statistics. Under Eqn. (9) we know the distribution of each observation (standard normal) and thus any function of the observations. The ordered values of Y, $Y_{(1)} \leqslant \cdots \leqslant Y_{(n)}$, could be compared to their expected values under Eqn. (9), for example by plotting one against the other, and deviation of this plot from a line with intercept 0 and slope 1 could be measured in various ways. In the case of the composite null hypothesis Eqn. (10), we could make use of the result that under H_0, $(Y_i - \bar{Y})/S$ has a distribution free of θ_1 and θ_2, and the vector of these residuals is independent of the pair $(\bar{Y}, S^2)$, and then for example compare the skewness $n^{-1} \sum \{(Y_1 - \bar{Y})/S\}^3$ with that expected under normality.

Example 5. Suppose we have a sample of independent observations $Y_1, \ldots, Y_n$ or a circle of radius 1 and our null hypothesis is that the observations are uniformly distributed on the circle. One choice of a test statistic is $T_1 = \sum_{i=1}^{n} \cos Y_i$, very large positive (or negative) values indicating a concentration of observations at angle 0 (or π). If, instead, we wish to detect clumps of observations at two angles differing by π, then $T_2 = \sum_{i=1}^{n} \{\cos(2Y_i) - 1\}$ would be more appropriate. The exact distribution of T_1 under H_0 is not available in closed form, but the mean and variance are readily computed as 0 and n, so a normal approximation might be used to compute the p-value.

In the examples described above, the test statistics are *ad hoc* choices likely to be large if the null hypothesis is not true; these are called pure tests of significance. Clearly, a more sensitive test can be constructed if we have more specific knowledge of the

likely form of departures from the null hypothesis. The theory of hypothesis testing formalizes this by setting up a null hypothesis and alternative hypothesis, and seeking to construct an optimal test for discriminating between them. See *Hypothesis Testing in Statistics*. In the remainder of this section we consider a less formal approach based on the likelihood function.

1.2 Significance Tests in Parametric Models

In parametric models, tests of significance are often constructed by using the likelihood function, and the p-value is computed by using an established approximation to the distribution of the test statistic. The likelihood function is proportional to the joint density of the data,

$$L(\theta; y) = c(y) f(y; \theta). \tag{11}$$

We first suppose that we are testing the simple null hypothesis $H_0 : \theta = \theta_0$ in the parametric model $f(y; \theta)$. Three test statistics often constructed from the likelihood function are the Wald or maximum likelihood statistic

$$w_e = (\theta - \theta_0)^T j(\hat{\theta})(\hat{\theta} - \theta_0), \tag{12}$$

the Rao or score statistic

$$w_u = U(\theta_0)^T \{ j(\hat{\theta}) \}^{-1} U(\theta_0), \tag{13}$$

and the likelihood ratio statistic

$$w = 2\{\ell(\hat{\theta}) - \ell(\theta_0)\} \tag{14}$$

where in Eqns. (12), (13), and (14) the following notation is used:

$$\sup_{\theta} L(\theta; y) = L(\hat{\theta}; y) \tag{15}$$

$$\ell(\theta) = \log L(\theta)$$

$$U(\theta) = \ell'(\theta) \tag{16}$$

$$j(\theta) = -\ell''(\theta). \tag{17}$$

The distributions of each of the statistics Eqns. (12), (13), and (14) can be approximated by a χ^2_k distribution, where k is the dimension of θ in the model. This relies on being able to apply a central limit theorem to $U(\theta)$, and to identify the maximum likelihood estimator $\hat{\theta}$ with the root of the equation $U(\theta) = 0$. The precise regularity conditions needed are somewhat elaborate; see, for example, Lehmann and Casella (1998, Chap. 6) and references therein. The important point is that under the simple null hypothesis the distributions of each of these test statistics is exactly known, and p-values readily computed.

In the case that the hypothesis is composite, a similar triple of test statistics computed from the likelihood function is available, but the notation needed to define them is more elaborate. The details can be found in, for example, Cox and Hinkley (1974, Chap. 9, Sect. 3) and the notation above follows theirs.

If θ is a one-dimensional parameter, then a one-sided version of the test statistics given in Eqns. (12), (13), and (14) can be used instead, since the square root of w_e, w_u, or w follows approximately a standard normal distribution.

It is rare that the exact distribution of test statistics can be computed, but the normal or chi-squared approximation can often be improved, and this has been a subject of much research in the theory of statistics over the past several years. A good book-length reference is Barndorff-Nielsen and Cox (1994). One result of this research is that among the three test statistics the square root of the likelihood-ratio statistic w is generally preferred on a number of grounds, including the accuracy of the normal approximation to its exact distribution. This is true for both simple and composite tests of a scalar parameter.

1.3 Significance Functions and Posterior Probabilities

We can also use a test of significance to consider the whole set or interval of values of θ that are consistent with the data. If θ is scalar, one of the simplest ways to do this is to compute $r(\theta) = \pm\sqrt{w(\theta)}$ as a function of θ, and tabulate or plot $\Phi\{r(\theta)\}$ against θ, choosing the negative root for $\theta < \hat{\theta}$, and the positive square root otherwise. This significance function will in regular models decrease from one to zero as θ ranges over an interval of values. The θ values for which $\Phi(r)$ is 0.975 and 0.025 provide the endpoints of an approximate 95 percent confidence interval for θ.

In a Bayesian approach to inference it is possible to make probability statements about the parameter or parameters in the model by constructing a posterior probability distribution for them. In a model with a scalar parameter θ based on a prior $\pi(\theta)$ and model $f(y; \theta)$ we compute a posterior density for θ as

$$\pi(\theta | y) \propto f(y; \theta)\pi(\theta) \tag{18}$$

and can assess any particular value θ_0 by computing

$$\int_{\theta_0}^{\infty} \pi(\theta | y) \mathrm{d}\theta$$

called the posterior probability of θ being larger than θ_0. This posterior probability is different from a p-

value: a p-value assesses the data in light of a fixed value of θ, and the posterior probability assesses a fixed value of θ in light of the probability distribution ascribed to the parameter. Many people find a posterior probability easier to understand, and indeed often interpret the p-value in this way. There is a literature on choosing priors called 'matching priors' to reconcile these two approaches to inference; recent developments are perhaps best approached from Kass and Wasserman (1996). See also *Bayesian Statistics*.

1.4 Hypothesis Testing

Little has been said here about the choice of a test statistic for carrying out a test of significance. The difficulty is that the theory of significance testing provides no guidance on this choice. The likelihood based test statistics described above have proved to be reasonably effective in parametric models, but in a more complicated problem, such as testing the goodness of fit of an hypothesized model, this approach is often not available. To make further progress in the choice of test statistics, the classical approach is to formulate a notion of a 'powerful' test statistic, i.e., one that will reliably led to small p-values when the null hypothesis is not correct. To do this in a systematic way requires specifying what model might hold if in fact the null hypothesis is incorrect. In parametric models where the null hypothesis is $H_0: \theta = \theta_0$ the alternative may well be $H_a: \theta \neq \theta_0$. In more general settings the null hypothesis might be H_0: 'the model is normal' and the alternative H_a: 'the model is not normal.' Even in the parametric setting if θ is a vector parameter it may be necessary to consider what direction in the parameter space away from θ_0 is of interest. The formalization of these ideas is the theory of hypothesis testing, which considers both null and alternative hypotheses, and optimal choices of test statistics. See *Hypothesis Testing in Statistics*; *Goodness of Fit: Overview*.

2. Further Topics

2.1 Combining Tests of Significance

The p-value is a function of the data, taking small values when the data are incompatible with the null hypothesis, and vice versa. As a function of y the p-value itself has a distribution under the model $f(y; \theta)$ and in particular under the null hypothesis H_0 has the uniform distribution on the interval (0, 1). In principle, then, if we have computed p-values from a number of different datasets the p-values can be compared to observations from a $U(0, 1)$ distribution with the objective of obtaining evidence of failure of the null hypothesis across the collection of datasets. This is one of the ideas behind meta analysis; see *Meta-analysis: Overview* and *Meta-analysis: Tools*. One difficulty is that the studies will nearly always differ in a number of respects that may mean they are not all measuring the same parameter, or measuring it in the same way. Another difficulty is that studies for which the p-value is not 'statistically significant' will not have been published, and thus are unavailable to be included in a meta analysis. This selection effect may seriously bias the results of the meta analysis.

2.2 Sample Size

The p-value is a decreasing function of the size of the sample, so that a very large study is more likely to show 'statistical significance' than a smaller study. This has led to considerable criticism of the p-value as a summary measure. Some statisticians have argued that, for this reason, posterior probabilities are a better measure of disagreement with the null hypothesis; see, for example, Berger and Sellke (1987).

To some extent the criticism can be countered by noting that the p-value is just one summary measure of a set of data, and excessive reliance on one measure is inappropriate. In a parametric setting it is nearly always advisable to provide, along with p-value for testing a particular value of the parameter of interest, an estimate of the observed effect and an indication of the precision of this estimate. This can be accomplished by reporting a significance function, if the parameter of interest is one-dimensional. At a more practical level, it should always be noted that a small p-value should be interpreted in the context of other aspects of the study. For example a p-value of less than 0.05 could be obtained by a very small difference in a study of 10,000 cases or a relatively larger difference in a study of 1,000. While a 1 percent reduction may be of substantial importance for some scientific contexts, this needs to be evaluated in its context, and not by relying on the fact that it is 'statistically significant.' Unfortunately, the notion that a study report is complete if and only if the p-value is found to be less than 0.05 is fairly widely ingrained in some disciplines, and indeed forms a part of the requirements of some government agencies for approving new treatments.

Another point of confusion in the evaluation of p-values for testing scalar parameters is the distinction sometimes made between one-sided and two-sided tests of significance. A reliable procedure is to compute the p-value as twice the smaller of the probabilities that the test statistic is larger than or smaller than the observed value, under the null hypothesis. This so-called two-sided p-value measures disagreement with the null hypothesis in two directions away from the null hypothesis, towards the alternative that the, say, new treatment is worse than the old treatment as well

Table 2
Hypothetical data from a small fictitious experiment on insomnia

	Decrease in insomnia reported	No decrease reported	No. of participants
Herbal tea	x	$10-x$	10
Other beverage	$13-x$	$x-3$	10
Total	13	7	20

as better. In the absence of very concrete *a priori* evidence that the alternative hypothesis is genuinely one-sided, this p-value is preferable.

In testing hypotheses about parameters of dimension larger than one, it can be difficult, as noted above, to define the relevant direction away from the null hypothesis. One solution is to identify the parameters of interest individually, and carry out separate tests on those parameters. This will usually be effective unless the dimension of the parameter of interest is unusually large, but in many applications it would be rare to have more than a dozen such parameters. Extremely large datasets such as arise in so-called data mining applications raise quite new problems of inference, however.

2.3 Fixed Level Testing

The problem of focusing on one or two so-called critical p-values is sometimes referred to as fixed-level testing. This was useful when computation of p-values was a very lengthy exercise, and it was usual to provide tables of critical values. It is now usually a very routine matter to compute the exact p-value, which is usually (and should be) reported along with other details such as sample size, estimated effect size, and details of the study design. There is still in some quarters a reliance on fixed level testing, with the result that studies for which the p-value is judged 'not statistically significant' may not be published. This is sometimes called the 'file drawer problem,' and a quantitative analysis was considered in Dawid and Dickey (1977).

More recently there has been a move to make the results of inconclusive studies available over the Internet, and if this becomes widespread practice will alleviate the file drawer problem. This issue is particularly important for meta analysis.

2.4 Achievable p-Values

In some problems where the distribution is concentrated on a discrete set, the number of available p-values will be relatively small. For example, Table 2 shows a hypothetical 2×2 table for the herbal tea experiment. The null hypothesis is that the two beverages are equally effective at reducing insomnia, and is here tested using Fisher's exact test. The relevant calculations are produced in Table 3, from which we see that the set of achievable p-values is relatively small.

Table 3
The set of achievable p-values from Table 2, as a function of x. The p-value for Table 2 is given by the formula $\sum_{s=x}^{\min(10,\,13-x)} p(s)$, where $p(s) = \binom{10}{s}\binom{10}{13-s}/\binom{20}{13}$. The mid p-value is given by $(1/2)p(x)+\sum_{s=x+1}^{\min(10,\,13-x)} p(s)$

	x			
	10	9	8	7
$p(x)$	0.0015	0.0271	0.1463	0.3251
p-value	0.0015	0.0286	0.1749	0.5000
Mid p-value	0.0007	0.0150	0.1018	0.3561

Some authors have argued that for such highly discrete situations a better assessment of the null hypothesis can be achieved by the use of Barnard's mid p-value, which is $(1/2)\Pr(X=x)+\Pr(X>x)$. See Agresti (1992) and references therein.

3. Conclusion

A test of statistical significance is a mathematical calculation based on a test statistic, a null hypothesis, and the distribution of the test statistic under the null hypothesis. The result of the test is to indicate whether the data are consistent with the null hypothesis: if they are not, then either we have observed an event of low probability, or the null hypothesis is not correct.

The choice of test statistic is in principle arbitrary, but in practice might be determined by convention in the field of application, by intuition in a relatively new setting, or by one or more considerations developed in statistical theory. It is convenient to use test statistics whose distributions can easily be calculated exactly or to a good approximation. It is useful to use a test statistic that is sensitive to the particular departures from the null hypothesis that are of particular interest in the application.

A test of statistical significance is just one component of the analysis of a set of data, and should be supplemented by estimates of effects of interest, considerations related to sample size, and a discussion of the validity of any assumptions of independence or underlying models that have been made in the analysis. A statistically significant result is not necessarily an important result in any particular analysis, but needs to be considered in the context of research in that field.

An eloquent introduction to tests of significance is given in Fisher (1935, Chap. 2). Kalbfleisch (1979,

Chap. 12) is a good textbook reference at an undergraduate level. The discussion here draws considerably from Cox and Hinkley (1974, Chap. 3), which is a good reference at a more advanced level. An excellent overview is given in Cox (1977). For a criticism of *p*-values see Schervish (1996) as well as Matthews (1998).

See also: Distributions, Statistical: Approximations; Frequentist Inference; Goodness of Fit: Overview; Hypothesis Testing in Statistics; Likelihood in Statistics; Meta-analysis: Overview; Resampling Methods of Estimation

Bibliography

Agresti A 1992 A survey of exact inference for contingency tables. *Statistical Science* **7**: 131–53

Barndorff-Nielsen O E, Cox D R 1994 *Inference and Asymptotics*. Chapman and Hall, London

Berger J O, Sellke T 1987 Testing a point null hypothesis: The irreconcilability of *p*-values and evidence. *Journal of the American Statistical Association* **82**: 112–22

Cox D R 1977 The role of significance tests. *Scandinavian Journal of Statistics* **4**: 49–70

Cox D R, Hinkley D V 1974 *Theoretical Statistics*. Chapman and Hall, London

Dawid A P, Dickey J M 1977 Properties of diagnostic data distributions. *Journal of the American Statistical Association* **72**: 845–50

Fisher R A 1935 *The Design of Experiments*. Oliver and Boyd, Edinburgh, UK

Kalbfleisch J G 1979 *Probability and Statistical Inference*. Springer, New York, Vol. 2

Kass R E, Wasserman L 1996 Formal rules for selecting prior distributions: A review and annotated bibliography. *Journal of the American Statistical Association* **91**: 1343–70

Lehmann E L, Casella G 1998 *Theory of Point Estimation*. Springer, New York

Matthews R 1998 The great health hoax. *Sunday Telegraph* 13 September. Reprinted at ourworld.compuserve.com/homepages/rajm/

Schervish M J 1996 *P* values: what they are and what they are not. *American Statistician* **96**: 203–6

N. Reid

Simmel, Georg (1858–1918)

Georg Simmel was born in the heart of Berlin on March 1, 1858. He was the youngest son of Flora (born Bodstein) and Eduard Simmel, who, although coming from Jewish families, had been baptized into Christianity (he as a Catholic, she as a Protestant). Following the early death of Simmel's father in 1874, the family suffered serious financial difficulties, which, where the young Georg was concerned, were overcome thanks to Julius Friedländer, a friend of the family (co-founder of the music publishing company 'Peters'). Friedländer felt a strong sympathy for the young Simmel, and indeed took him under his wing as his protégé. Thus, Georg could attend high school and then university in Berlin, where he studied philosophy, history, art history, and social psychology (*Völkerpsychologie*). Simmel received his degree as Doctor of Philosophy in 1881, but not without difficulty: his first attempt at a doctoral thesis, '*Psychological and Ethnological Studies on the Origins of Music*,' was not accepted, and he had instead to submit his previous work on Kant, *On the Essence of Matter—Das Wesen der Materie nach Kant's Physischer Monadologie*, which had earned him a prize (Köhnke 1996). The *Habilitation* (postdoctoral qualification to lecture) came next in 1885, for which he also encountered some controversy, and after this his academic career began immediately as a *Privatdozent* (external lecturer). An extraordinary (*außerordentliche*) professorship without salary followed in 1901, and indeed he had to wait until 1914 before he was offered a regular professorship at the University of Strasbourg, where he remained until his death on September 26, 1918, shortly before the end of the First World War.

During his lifetime, Simmel was a well-known figure in Berlin's cultural world. He did not restrict himself merely to scientific or academic matters, but consistently showed great interest in the politics of his time, including contemporary social problems and the world of the arts. He sought to be in the presence of, and in contact with, the intellectuals and artists of his day. He married the painter Gertrud Kinel, with whom he had a son, Hans, and maintained friendships with Rainer Maria Rilke, Stefan George, and Auguste Rodin, amongst many others. At his home in Berlin, he organized private meetings and seminars (Simmel's *privatissimo*), whose participants he would choose personally.

1. Georg Simmel and the Social Sciences

Simmel's contributions to the social sciences are immeasurable. Nevertheless, most of them remain misunderstood, or have been separated from the intentions of their creator, and, thus, their origins have been forgotten. From system theory to symbolic interactionism, almost all sociological theories need to rediscover Simmel as one of their main founding parents.

Simmel's interest in the social sciences, especially in sociology, can be traced to the very beginning of his academic career. After attending seminars offered by Moritz Lazarus and Heymann Steinthal (founders of the *Völkerpsychologie*) during his student years, he became a member of Gustav Schmoller's circle, where he

became acquainted with the debates concerning the national economy of the time, and for whom he held a conference on *The Psychology of Money* (GSG2 1989, pp. 49–65), which would later constitute the first pillar of one of his major works, *The Philosophy of Money* (Frisby and Köhnke 1989). In both of these circles Simmel became more aware of, and sensitized towards, social questions. Schmoller's engagement with social questions, together with Lazarus' and Steinthal's emphasis on the level of '*Überindividualität*' (supraindividuality), and their relativistic worldview and insistence that ethical principles are not of universal validity, as well as Simmel's own interest in Spencer's social theory, all helped to shape the contours of his sociological approach. These various influences were melded together with Simmel's philosophical orientation, particularly his interest in Kant, which yielded a new and rather far-reaching sociological perspective that would evolve extensively throughout his life.

For example, Simmel's first sociological work, *On Social Differentiation* (*Über Sociale Differenzierung*: GSG2 1989), written at the very beginning of his academic career, was deeply influenced by Herbert Spencer and the ideas of Gustav Schmoller and his circle. Slowly, however, as the 1890s drew on, the admiration Simmel had felt for Spencer's theories turned into rejection, and he distanced himself from an evolutionary-organicist approach to sociology. In this rejection of such theories, he brought his knowledge of Kant to his sociological thinking, and this would become the basis of his later contact and dialogue with the members of the southern-German neo-Kantian school, such as Rickert, Windelband, and with the sociologist who was closest to their ideas: Max Weber. This contact with neo-Kantianism influenced Simmel's approach to the social sciences at the end of the nineteenth century and during the first years of the 1900s.

From contemporary testimonies we know that Simmel was in fact one of the greatest lecturers at the University of Berlin, and that his seminars were attended by a great number of students (Gassen and Landmann 1993). It is thus difficult to understand why Simmel did not have a more successful academic career. We know from his correspondences and from Gassen and Landmann's attempts to reconstruct Simmel's life and *oeuvre* that Georg Jellinek engaged himself, though with no positive results, in seeking to obtain an ordinary professorship for Simmel at the University of Heidelberg in 1908 (with, following Jellinek's death, a further attempt by Alfred Weber in 1912, and another in 1915). His converted Jewish background (i.e., assimilated into German society) surely played a significant role in the lack of recognition Simmel received from the academic system; but also his peculiar and original understanding of science, which diverged greatly from established patterns, as well as his characteristic mode of writing, using essays instead of more 'academic' and standardized forms, contributed to his not being accepted into the rather formal and classical German academic milieu. Both attempts at obtaining a professorship for Simmel were blocked by the bureaucracy of the Grand Dukedom of Baden. In fact the letter of evaluation written by the Berlin historian Schäfer regarding a possible professorship for Simmel in Heidelberg in 1908 has as the primary argument for denying Simmel's potential ability to be a good professor his 'Jewishness,' which, according to Schäfer, too obviously tinged Simmel's character and intellectual efforts with a strong relativism and negativity, which could not be good for any student (Gassen and Landmann 1993, Köhnke 1996). Throughout his life Simmel had to fight against these kinds of accusation, and he endeavored to build a 'positive relativism' in an attempt to show that he did not question 'absolute pillars' and thus leave us with nothing, but sought instead to show that this sense of 'absoluteness' was also a product of human reciprocal actions and effects (*Wechselwirkungen*), and was not fundamentally absolute. Such an argument was too much for the society and scientific milieu of the time to accept and forgive, even when Simmel later radically rejected his Introduction to the Moral Science (*Einleitung in die Moralwissenschaften*: GSG3 1989 GSG4, 1991), referring to it as a sin of his youth. This was the work in which his relativism, in the critical, 'negative' sense, had been most fully in bloom (Köhnke 1996).

Notwithstanding these varying approaches to sociology, Simmel's interest in the discipline persisted right through his career. When Simmel, in his letter to Célestin Bouglé of March 2, 1908, wrote: 'At the moment I am occupied with printing my *Sociology*, which has finally come to its end,' and, sentences later, added that work on this book 'had dragged on for fifteen years' (Simmel Archive at the University of Bielefeld) he indicated that his engagement with sociology was a long-term project. Considering that *Sociology* (*Sociologie*)was first published in 1908 (and therefore his work on it must have begun around 1893), we can find its first seed in his article *The Problem of Sociology* (*Das Problem der Sociologie*) originally published in 1894. Simmel must have thought this article a significant contribution to sociology, since he endeavored to spread it abroad as much as possible. Hence, the French translation of *The Problem of Sociology* appeared, simultaneously with the original German version, in September 1894. The American translation appeared in the *Annals of the American Academy of Political and Social Science* a year later, and, by the end of the century, the Italian and Russian translations were also in print. The American translation is of particular significance, since Simmel emphasized therein, in a footnote, that sociology involved an empirical basis and research, and should not be thought of as an independent offshoot from philosophy, but as a science concerning

the social problems of the nineteenth century. In the Italian translation he delivered an updated version of the text, wherein he introduced clear references to his theoretical polemic with Emile Durkheim. From his letter to Célestin Bouglé of February 15, 1894, we know that Simmel was, after the completion of *The Problem of Sociology*, quite excited by this new discipline, and that he did not foresee a shift away to any other fields of inquiry in his immediate future. During those years he worked on most of the key sociological areas of research, thus articulating the key social problems of his time within a sociological framework: for example, workers' and women's movements, religion, the family, prostitution, medicine, and ethics, amongst many others. He seemed deeply interested in putting his new theoretical proposals and framework for the constitution of sociology into practice. He realised the institutionalisation of the new discipline would be reinforced by the establishment of journals for the discipline; hence his participation in: the '*Institut Internationale de Sociologie*,' for whom he became vice-president; the *American Journal of Sociology*, and, although only briefly (due to differences with Emile Durkheim), *l'Année sociologique* (Rammstedt 1992, p. 4). He also played with the idea of creating his own sociological magazine. Another means of solidifying the role of sociology within the scientific sphere was through the academia, and he engaged himself in organising sociological seminars, offering them uninterruptedly from 1893 until his death in 1918. On June 15, 1898, he wrote to Jellinek 'I am absolutely convinced that the problem, which I have presented in the *Sociology*, opens a new and important field of knowledge, and the teaching of the forms of sociation as such, in abstraction from their contents, truly represents a promising synthesis, a fruitful and immense task and understanding' (Simmel Archive at the University of Bielefeld) Despite his original intention, it is clear Simmel did not work continuously on the *Sociology* for 15 years as, from 1897 to 1900, he worked almost exclusively on *The Philosophy of Money* (*Philosophie des Geldes*), and also found time for the writing and publication of his *Kant* (1904), as well as a reprint of his *Introduction to the Moral Sciences* (1904), which he had intended to rewrite, for he no longer accepted most of the ideas he had presented therein when it was first published in 1892 (although he did not achieve this); the revised edition of *The Problems of the Philosophy of History* (*Probleme der Geschichtsphilosophie* 1905), (*The Philosophy of Fashion* (*Philosophie der Mode* 1905), and *Religion Die Religion* 1906/1912) were all worked on by Simmel during this period too (Rammstedt 1992b). In *The Problem of Sociology* he questioned for the first time the lack of a theoretically well-defined object of study for the emerging discipline, and sought to develop a specific sociological approach, which would entail a distinct object of study, in order to bestow legitimacy and scientific concreteness to a discipline, under attack from different, and more settled, lines of fire. Fearing his call had not been heard, Simmel endeavored to prove his point by writing a broader work, within which he attempted to put into practice the main guidelines he had suggested as being central to the newly emerging discipline. In this way the *Sociology*, almost one thousand pages long, was cobbled together from various bits and pieces, taken from several essays he had written between the publication of *The Problem of Sociology* in 1894 and its final completion in 1908. Simmel, as can be understood from his letters, was aware of the incompleteness of this work, but rescued it by saying that it was an attempt to realise that which he had suggested almost 15 years earlier, for it had not been noticed enough by the scientific community. Simmel did not wish, by that time, to be thought of as only a sociologist. This was due to sociology not being an established discipline within the academic world, which therefore did not allow him to obtain a professorship in the field (i.e., offering very little recognition). Nevertheless he never quite abandoned the field of sociology and continued to write about religion, women's issues, and the family. He merely broadened his scope to include *Lebensphilosophie* (*the philosophy of life*) and cultural studies in general, whilst participating at the same time in the founding of the German Sociological Society (*Deutsche Gesellschaft für Soziologie*), for whom he served as one of the chairmen until 1913.

Sociology marks the end of Simmel's strongly Kantian period, and represents a turning point in his interests, for, from this moment onwards, he refused to be labeled as a sociologist, seeking instead to return to philosophy. Indeed, following the publication of the 'big' *Sociology*, as it is usually called by those who work in Simmel studies, he did not publish any sociological papers for nine years (although he continued, as has already been mentioned, to offer sociological seminars until his death), instead devoting his efforts to philosophy, history and the philosophy of art. Hence, as part of Simmel's output from these years we find, amongst other works, *Kant und Goethe* (*Kant und Goethe* 1906/1916), *Goethe* (1913), *Kant* (1904/1913/1918), *The Principal Problems of Philosophy* (*Hauptprobleme der Philosophie* 1911), *The Philosophical Culture* (*Philosophische Kultur* 1911/1918), and *Rembrandt* (1916). This new direction was accompanied and partly motivated by Simmel's acknowledgment of Henri Bergson's *hoeuvre*, and his inclination towards *Lebensphilosophie* (the philosophy of life; see Fitzi 1999, GSG16 1999). Thus, 'life' became the primary focus of Simmel's theoretical work, and consequently 'society' was pushed into a secondary role. During these years Simmel occupied himself with the study of artists, of their production, and how the relation between life and its formal expression is crystallised in their work. It was as if Simmel had lost interest in sociology, as he did not write a single line

concerning it for years. Yet, unexpectedly in 1917, a year before his death, he wrote *Grundfragen der Soziologie* (*Main Questions of Sociology*), the 'small sociology,' as it is called by Simmel scholars (i.e., in contrast to his 1908 work). The impetus for writing this book came from a publisher (Sammlung Göschen), who intended to print an introductory work to sociology, which he invited Simmel to write, because of the success other works of his had enjoyed for the same publisher. If Simmel had indeed distanced himself from all sociological questions, it is likely he would have merely reached for the shelf to his previous works and rewritten them in shorter form. But this was not the case, because, although he utilised older material, Simmel rewrote and redefined his perspective, and in the *Main Questions of Sociology*, scarcely 100 pages long, presented the final stage of his sociological reflections, which melded together his previous study of forms of sociation with a perspective from the philosophy of life. This approach to sociology fell into neglect after his death, awaiting revitalisation, harbouring a broad scope and original perspective for new generations of sociologists to rediscover (Simmel's last work was a philosophical contribution to *Lebensphilosophie*, the *Lebensanschauung—View of Life—1918*, GSG16 1999).

2. *The Object of Sociology*

At a time when sociology was still far from being an established discipline, instead seeking to stand up and open its eyes for the first time, Simmel sought to release it from the burden of being 'the science of society.' According to him this burden was an impossible one for the new-born discipline to carry, since being the science of society meant having to compete with already settled and established disciplines for the legitimacy of its object of study; law and history, psychology and ethnology, all could argue the case that society was their object, thus leaving sociology with the mere pretence of including elements from them all. Viewed from this perspective, as an object society was an all-encompassing matter but, at the same time, it eluded any scientific investigation, just like sand falling through our fingers. Simmel, as Max Weber two decades later, contributed to the demystification of 'society' as some kind of essential entity (as it appeared in nineteenth-century sociology, and remained as such in the sociology of Durkheim and Tönnies), instead explaining it as a dynamic process, a continuous happening, a continuous becoming, which is nothing more than the mere sum of the existing forms of sociation.

Individuals as well as society/ies are not units in and of themselves, though they may appear as self-sufficient units depending on the distance the observer interposes between him/herself and them (as observed objects). However, Simmel argued in his *Main Questions of Sociology* (GSG16 1999, pp. 62–8) that if we merely take individuals and pretend, through adopting a significant distance, to approach 'society,' or a social phenomenon, we will not reach our goal. Therefore, when conducting a sociological inquiry, what we must seek to do is base ourselves on the forms of sociation, which are built in *Wechselwirkung*. Simmel used this concept when either addressing 'interactions' or following the Kantian definition as 'reciprocal actions and effects' (Simmel 1978). The difference between the two possible meanings of this concept, or, more precisely, the differentiation of the two different concepts hidden behind the same word, has been addressed in the English translation of *The Philosophy of Money*; so the usual mistake of translating *Wechselwirkung* as a direct synonym for interaction has been corrected. According to Simmel, what actually takes place between individuals will be seen to be that which constitutes the object of sociology, not merely the individuals by themselves, or society as a whole, for, as previously mentioned, society is nothing but the sum of forms of sociation, a continuous process in and for which these forms intertwine and combine themselves to form the whole (GSG11 1992, p. 19). Thus, Simmel defined society as the sum of the forms of reciprocal actions and effects.

3. *The Concept of 'Form'*

If we take reciprocal actions and effects as our starting point, we achieve only a perspective common to all social and human sciences: sociology as method. In order to construct sociology as an independent discipline, with a specific object of study, it is necessary to analytically differentiate between form and content, that is, 'forms' as means and patterns of interaction between individuals, social groups or institutions, and 'contents' as that which leads us to act, the emotions or goals of human beings. Thus, social forms were conceived as being the object of analysis for a scientific sociology, which would only be possible empirically. Contents, on the other hand, should be left to other disciplines, such as psychology or history, to analyze.

Simmel's sociological theory is orientated towards these 'forms of sociation' (defined as such in the subtitle of *Sociology*). For individuals to become social, they need to rely on such forms in order to channel their contents, as forms represent their only means of participation in social interaction. Forms are independent from contents, and when analyzing them, it is necessary that they are abstracted from individual, particular participation in concrete interactions: the question of social forms does not include that regarding the specific relationships between the participants involved in concrete interactions; we are dealing only with that which is *between* them, as for an

analysis of social forms the particular individuals involved in them are irrelevant. Human beings, with their particular 'contents,' only become social when they seek to realise these contents, and then acknowledge this is only possible in a social framework: via 'exteriorising,' via acting through forms. Hence forms are social objectivations, which impose themselves, with their norms, upon particular individuals, an imposition which could only be annulled by isolation. These impositions and constraints are part of what sociation is, a concept which plays a key role in Simmel's theory, for he actually maintained that human beings are not social beings by nature. Simmel placed sociation, as a product of reciprocal actions and effects, at the centre of his formal, or pure (*reine*), sociology. The form of competition can be used as an example of what Simmel actually meant by 'forms.' Competition does not imply any specific contents, and remains the same, independent of those who are competing, or what they are competing for. Competition forms 'contents' (such as job seeking, or looking for the attention of a beloved person, amongst many others), limits the boundaries of actions, and actually brings them into being via giving them a framework and shape in which they can appear in the arena of actions and effects.

This situation of having the choice between multiple forms for channelling one content is stressed appreciatively in Simmel's sociology, particularly in his *Lebensphilosophie*, when he no longer contrasted the concept of form with the concept of content, but with life. In this later period, forms are the crystallization of the unretainable flow of life, yet also the channels of expressing life, for life cannot be expressed as only itself. Simmel explained this apparent paradox by asserting that life is 'more-life' and also 'more-than-life.' The concept of 'more-life' merely implies that this continuous flow connects every complete moment with the next; 'more-than-life' implies that life is not life if it does not transcend its boundaries, and becomes crystallized in a form. So life becomes art, or science, for example; life becomes externalised and crystallised and hence expressed and fulfilled. For instance, we should understand art as one form of expressing (*äussern*) life through (artistic, aesthetic) forms. Actually uncovering how this perspective can be applied to sociology may appear to be a difficult task; the first step towards clarifying this was made by Simmel himself in his *Main Questions of Sociology*.

4. *'Me' and 'You'*

Deliberately distancing himself from the concept of 'alter ego,' Simmel emphasized the signific1ance the concept of 'you' should have to all sociological theory. He articulated and embedded this concept within his—sociological—theory of knowledge. In parallel with Kant's question on nature, Simmel formulated this question as 'How is society possible?' the first and well known digression of his *Sociology*. In order to answer this question he proposed three *a priori*. According to the first, 'Me' and 'You' would see each other as 'to some extent generalised' (GSG11 1992, p. 47), and he therefore assumed that each individual's perception of the other (i.e., 'you') would also be generalized to a certain degree. The second *a priori* affirms that each individual as an 'element of a group is not only a part of the society, but, on top of this, also something else' (GSG11 1992, p. 51) that derives its own uniqueness, its individuality from the duality of being/not being sociated. The third *a priori* is related to the assignment of each individual to a position within their society (in the sense of the sum of highly differentiated reciprocal actions and effects): 'that each individual is assigned, according to his qualities, to a particular position within his social milieu: that this position, which is ideally suited to him, actually exists within the social whole—this is the supposition according to which every individual lives his social life' (GSG11 1992, p. 59); this is what Simmel meant when he wrote about the 'general value of individuality.'

These sociological *a priori*, orientated towards role, type, and individual/society, also allow us to understand the central aspects of Simmel's methodology, which can be framed together with the concepts of 'differentiation,' in the sense of division and difference, and of 'dualism,' in the sense of an irreconcilable, tragic opposition of the elements of a social whole.

5. *A Proposal for Three Sociologies*

According to Simmel sociology needed to stand as an autonomous science in close relationship with 'the principal problem areas' (GSG16 1999, pp. 76–84) of social life, even when these areas were distant from each other. Already in 1895 he asserted that just which name to give to any particular group was quite unimportant, since the real question was to state problems and to solve them, and not at all to discuss the names which we should give to any particular groups (1895, *Annals of the American Academy of Political and Social Science* **6**: 420). He stressed this orientation into problem areas again in 1917, when he returned to theorizing on the *Main Questions of Sociology*. Thus, as an attempt to approach these main questions (i.e., which relationships exist between society and its elements, individuals) he concentrated on three different sets of problems with concrete examples, particularly in the third part of the first chapter.The first focuses on objectivity as a component of the social sphere of experience; the second includes the actual facts of life, in which, and through which, social groups are realised; and the third emphasises the significance of society (Fitzi and GSG16 1999, p. 84) for grasping and understanding attitudes towards the world and life. These three problem areas correspond to his three proposed sociologies: the 'general sociology,' which, wherever society exists, deals with the

central relationships between the individuals and the social constructions resulting from them. The aim of general sociology is to show which reciprocally orientated values these social constructions and individuals possess. He gave an example for general sociology in the second chapter of the *Main Questions of Sociology*, by illustrating the relationships which exist between the 'social and the individual level,' a perspective which, at the same time, sociologizes the masses theory. The 'pure or formal' sociology focuses upon the multiple forms individuals use to embody contents, that is, emotions, impulses, and goals, in reciprocal actions and effects with others, be it with other people, social groups, or social organizations, and thus they constitute society. According to him, each form wins, entwined in these social processes, 'an own life, a performance, which is free from all roots in contents' (GSG16 1999, p. 106). The example Simmel chose for his formal sociology was the form of 'sociability' (*Geselligkeit*), which he presented in the third chapter of the *Main Questions of Sociology*. Finally, the 'philosophical sociology' circles the boundary of the 'exact, orientated towards the immediate understanding of the factual' (GSG16 1999, p. 84) empirical sociology: on the one hand the theory of knowledge (*Erkenntnistheorie*), on the other, the attempts to complement, through hypothesis and speculation, the unavoidably fragmentary character of factual, empirical phenomena (*Empirie*), in order to build a complete whole (GSG16 1999, p. 85). In the fourth chapter, entitled 'The Individual and Society in Eighteenth and Nineteenth Century Views of Life', Simmel illustrated the abstract necessity of individual freedom, which should be understood as a reaction to the contemporary, increasing, social constraints and obligations. Thus, reference is made to general conflicts between individual and society, which derive from the irreconcilable conflict between the idea of society as a whole, which 'requires from its elements the one-sidedness of a partial function,' and the individual, who knows him/herself to be only partially socialized, and 'him/herself who wants to be whole' (GSG16 1999, p. 123).

6. *Simmel and Modern Sociology*

Simmel's theoretical significance to contemporary sociology resides in the various theories, which built on his sociology. As examples, it will be sufficient to name symbolic interactionism, conflict theory, functionalism, the sociology of small groups, and theories of modernity. Simmel also introduced into sociology the essay as an academic form of analysis, whilst his digressions within *Sociology* should be mentioned as well, for they have become generally recognized as classical texts; see, for example, his digressions on 'The Letter' (*Brief*), 'Faithfulness' (*Treue*), 'Gratefulness' (*Dankbarkeit*), and 'The Stranger' (*Fremde*). But above all sociology owes Simmel the freedom he gave it from the fixation on the 'individual and society' as an ontic object—in hindsight, a point of no return.

See also: Capitalism; Cities: Capital, Global, and World; Cities: Internal Structure; Differentiation: Social; Ethics and Values; Family, Anthropology of; Family as Institution; Fashion, Sociology of; Feminist Movements; Gender and Feminist Studies in Political Science; Groups, Sociology of; History and the Social Sciences; Individual/Society: History of the Concept; Individualism versus Collectivism: Philosophical Aspects; Interactionism: Symbolic; Kantian Ethics and Politics; Knowledge (Explicit and Implicit): Philosophical Aspects; Knowledge Representation; Knowledge, Sociology of; Labor Movements and Gender; Labor Movements, History of; Medicine, History of; Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects; Modernity; Modernity: History of the Concept; Modernization and Modernity in History; Modernization, Sociological Theories of; Money, Sociology of; Motivational Development, Systems Theory of; Personality and Social Behavior; Personality Structure; Personality Theories; Prostitution; Religion, Sociology of; Science and Religion; Self-knowledge: Philosophical Aspects; Social Movements and Gender; Social Movements, History of: General; Sociology, Epistemology of; Sociology, History of; Sociology: Overview; Symbolic Interaction: Methodology; Theory: Sociological; Urban Life and Health; Urban Sociology

Bibliography

Dahme H J, Rammstedt O (eds.) 1983 *Georg Simmel. Schriften zur Soziologie. Eine Auswahl*, Suhrkamp, Frankfurt am Main, Germany

Fitzi G 1999 *Henri Bergson und Georg Simmel: Ein Dialog zwischen Leben und krieg. Die persönliche Beziehung und der wissenschaftliche Austausch zweier Intellektuellen im deutsch-französischen Kontext vor dem Ersten Weltkrieg*. Doctoral thesis, University of Bielefeld

Gassen K Landmann M 1993 *Buch des Dankes an Georg Simmel. Briefe, Erinnerungen, Bibliographie. Zu seinem 100. Geburtstag am 1. März 1958*. Duncker & Humblot, Berlin

GSG 10: *Philosophie der Mode (1905). Die Religion (1906/1912). Kant und Goethe (1906/1916). Schopenhauer und Nietzsche (1907)*, ed. M Behr, V Krech, and G Schmidt, 1995

GSG 8: *Aufsätze und Abhandlungen 1901–1908, Band II*, ed. A Cavalli and V Krech 1993

GSG 2: *Aufsätze 1887–1890. Über sociale Differenzierung (1890). Die Probleme der Geschichtsphilosophie (1892)*, ed. H J Dahme, 1989

GSG 5: *Aufsätze und Abhandlungen 1894–1900*, ed. H J Dahme and D Frisby, 1992

GSG 16: *Der Krieg und die geistigen Entscheidungen (1917). Grundfragen der Soziologie (1917). Vom Wesen des historischen Verstehens (1918). Der Konflikt der modernen Kulur (1918). Lebsensanschauung (1918)*, ed. G Fitzi and O Rammstedt, 1999

GSG 6: *Philosophie des Geldes (1900/1907)*, ed. D Frisby and k C Köhnke, 1989
GSG 3: *Einleitung in die Moralwissenschaft. Eine Kritik der ethischen Grundbegriffe. Erster Band (1892/1904)*, ed. K C Köhnke, 1989
GSG 4: *Einleitung in die Moralwissenschaft. Eine Kritik der ethischen Grundbegriffe. Zweiter Band (1893)*, ed. K C Köhnke, 1991
GSG 1: *Das Wesen der Materie (1881). Abhandlungen 1882–1884. Rezensionen 1883–1901*, ed. K C Köhnke, 1999
GSG 7: *Aufsätze und Abhandlungen 1901–1908, Band I*, ed. R Kramme, A Rammstedt, and O Rammstedt, 1995
GSG 14: *Hauptpobleme der Philosophie (1910/1927). Philosophische Kultur (1911/1918)*, ed. R Kramme, and O Rammstedt, 1996
GSG 9: *Kant (1904/1913/1918). Die Probleme der Geschichtsphilosophie, 2 Fassung (1905/1907)*, ed. G Oakes, and K Röttgers, 1997
GSG 11: *Soziologie. Untersuchungen über die Formen der Vergesellschaftung (1908)*, ed. O Rammstedt, 1992
Köhnke K C 1996 *Der junge Simmel in Theoriebeziehungen und sozialen Bewegungen*, Suhrkamp, frankfurt am Main, Germany
Otthein R (ed.) 1992a *Georg Simmel Gesamtausgabe*. Suhrkamp, Frankfurt, Germany
Rammstedt O 1992 Programm und Voraussetzungen der Soziologie Simmels, *Simmel-Newsletter* **2**: 3–21
Simmel G 1978 *The Philosophy of Money*, trans. and ed. by D Frisby and T Bottomore, Routledge, London
Simmel G 1895 The Problem of Sociology. *Annals of the American Academy of Political and Social Science* **6**

O. Rammstedt and N. Cantó-Milà

Simulation and Training in Work Settings

Training is a set of activities planned to bring about learning in order to achieve goals often established using a training needs analysis. Methods of training include: behavior modeling, based on video or real activities; action training, which provides an opportunity to learn from errors; rule- and exemplar-based approaches that differ in terms of the extent to which theory is covered; and methods that aim to develop learning skills, the skills needed to learn in new situations. Simulation involves using a computer-based or other form of model for training, providing an opportunity for systematic exposure to a variety of experiences. Simulations are often used in training complex skills such as flying or command and control. Issues in training highlight the differing emphasis that can be placed on goals related to the process of learning or performance outcomes; motivation and self-management in training; and the transfer dilemma where effortful methods of training that enhance transfer may be less motivational. Finally organizational issues and training evaluation are discussed with a particular emphasis on the role of the environment and follow-up enhancing transfer of skills learned during training to the workplace.

1. Overview

Training is a set of planned activities organised to bring about learning needed to achieve organizational goals. Organizations undertake training as part of the socialization of newcomers. Even with the best selection systems a gap frequently remains between the knowledge, skills, and abilities (KSAs) required in a job, and those possessed by the individual. Training is used by organizations to improve the fit between what an individual has to offer and what is required. Training is also used when job requirements change such as with the introduction of new technology. Because of the frequency with which job requirements change, a major aim of training is to develop specific skills while also ensuring their transferability and adaptability. An ultimate aim is to increase learning skills where the emphasis in training is on learning to learn. Many methods of training are available (e.g. behavioral modeling, action, computer-based training), with an increasing trend toward incorporating evaluation as part of the training. Although simulations have been used in training for decades, the reduced cost and increased sophistication of simulations now make them more readily available for a wide range of situations. Simulations provide increased opportunities to train for transferable and adaptable skills because trainees can experiment, make errors, and learn from feedback on complex and dynamic real-time tasks. Evaluation of training programmes remains an important component of the training plan, and can be used to enhance transfer.

2. Training Needs Analysis

Most training starts with a training needs analysis, where the present and future tasks and jobs that people do are analyzed to determine the task and job requirements. The identified requirements are then compared with what the individual has to offer. This is done within a broader organizational context. Many different methods of job, task, and organizational analysis can be used, including observation, questionnaires, key people consultation, interviews, group discussion, using records, and work samples (that is, letting people perform a certain activity and checking what they need to know to perform well). A transfer of training needs analysis (Hesketh 1997a), places a particular emphasis on identifying the cognitive processes that must be practised during learning to ensure that the skills can be transferred to contexts and tasks beyond the learning environment.

3. The Training Plan and Methods

There has been an increase in the methods of training that can be used within a broader training plan that includes transfer and evaluation (Quinones and Ehrenstein 1997). The training plan provides a way of

combining the illustrative methods discussed below to ensure that the training needs identified in the analysis can be met.

3.1 Behavior Modeling

Behavior modeling is often combined with role-playing in training. A model is presented on video or in real life, and the rationale for the special behaviors of the model are discussed with the trainees. Trainees then role play the action or interaction, and receive feedback from the trainers and fellow trainees. This type of training is very effective as it provides an opportunity for practice with feedback (e.g. Latham and Saari 1979).

3.2 Action Training

Action training follows from action theory (Frese and Zapf 1994) and exploratory learning. Key aspects of action training involve active learning and exploration, often while doing a task. Action learning is particularly effective as a method of training (Smith et al. 1997). Another important aspect of action training involves obtaining a good mental model of the task and how it should be approached. A mental model is an abstraction or representation of the task or function. Trainees can be helped to acquire a mental model through the use of 'orientation posters' or advanced organizers, or through the provision of heuristic rules (rules of thumb) (Volpert et al. 1984). One of the advantages of action training is the opportunity to learn from feedback and errors. Feedback is particularly important in the early stages of learning, but fading the feedback at later stages of learning helps ensure that trainees develop their own self-assessment skills (Schmidt and Bjork 1992). Errors are central in action training since systematic exposure to errors during learning provides opportunities to correct faulty mental models while providing direct negative feedback. Although earlier learning theory approaches argued that there should be only positive feedback, active error training helps trainees develop a positive attitude toward errors because of their value in learning (Frese et al. 1991).

3.3 Rules versus Examples in Training

Although it has traditionally been assumed that rule-based training provides a sound basis for longer term transfer, recent research suggests that for complex nonlinear problems, exemplar training may be superior. Optimizing the combination of rules and examples may be critical. In order to facilitate transfer, examples should be chosen carefully to cover the typical areas of the problem. With only a few examples, the rule and the example are often confused. However, individual instances are difficult to recall if trainees are provided with too many examples (DeLosh et al. 1997), suggesting that care needs to be taken when deciding how many examples will be presented in training.

3.4 Developing Learning Skills: Learning to Learn

Downs and Perry (1984) offer a practical way of helping trainees develop learning skills such as knowing that there are different ways of learning (e.g., learning by Memorizing, Understanding and Doing, the MUD categories). Facts need to be memorized, concepts need to be understood and tasks such as driving a motor car need to be learned by doing. The approach stresses the importance of selecting the most appropriate method for the material to be learned so that learning skills can be trained while also developing content skills. Methods that typically encourage learning skills as well as content skills involve action learning, active questioning and discovery learning, rather than direct lecturing and instruction. The ideas in the learning to learn literature lead to issues such as developing self-management techniques and focusing on learning.

3.5 Simulation Training

Simulation involves developing a model of an on-the-job situation that can be used for training and other purposes. The advantages of using simulation for training include reduced cost, opportunity to learn from errors, and the potential to reduce complexity during the early stages of learning. Historically simulators have been used for training in military, industrial and transport industries, and in management training where business games are widespread.

Pilot training on simulators is well developed, to the extent that many pilots can transfer directly from a flight simulator to a real aircraft (Salas et al. 1998). Within the aviation industry, crew cockpit resource management training has also been undertaken using simulators. Driving simulators are not as well developed, and the transfer of skills learned on a driving simulator to the actual road remains to be established. Nevertheless, there is a view that attitudes and driving decisions can be trained on a simulator. Simulations of management decision situations and command and control in the military, police, and emergency personnel are well developed, and widely used for training. These simulations provide a miniaturized version of real crises, with key decision points extracted and shrunk

in time. They provide an opportunity for the trainee to experiment with decisions, make mistakes, and learn from these errors (Alluisi 1991). Simulation training has also been used to facilitate the acquisition of cross-cultural skills.

4. Issues in Training

4.1 Learning vs. Performance Goals

Whether the emphasis during learning is on performance or learning goals is used to explain differences in how people conceptualize their ability. Dweck and Leggett (1988) argue that some people conceptualize ability as increasing with learning (learning orientation), while others see ability as fixed (performance orientation). People with a learning orientation learn from mistakes and challenges. However, individuals with a performance orientation view mistakes as examples of poor performance and learn less from errors. Performance oriented people also tend to demonstrate a helpless response to problems and are therefore less likely to overcome challenges. Martocchio (1994) showed that ability self-conceptualization was related to computer anxiety and self-efficacy.

Motivation plays a key role in transfer of training (Baldwin and Ford 1988). Numerous motivation issues have been studied, the most important ones being self-efficacy, relapse prevention, perceived payoffs, goals, and the training contract. Self-efficacy is critical to transfer in that people will use a skill only if they believe that they can actually perform the appropriate behavior. Relapse prevention focuses on teaching solutions to those situations in which it may prove difficult to use the newly learned skills (Marx 1982). Trainees who receive relapse prevention training have been found to use their skills more often and perform their job better.

4.2 Self-management in Training

Self-management implies that one acquires the skills to deal with difficulties, reward oneself, and increase self-efficacy. Metacognitive strategies showing evidence of self-management and self-reward during training related to training performance. Furthermore, self-efficacy is related to increased post-training knowledge (Martocchio 1994) and transfer performance. Thus, self-efficacy functions as a predictor of both training and transfer performance.

4.3 Transfer Dilemma

Transfer is important to ensure that the skills learned in one context or on one task can be applied in a range of different contexts and tasks. For example fire fighters may learn how to fight fires in urban areas, but their skills must also transfer to the bush or rural areas where they are frequently required to work in emergencies. In the field of technology, training, in the use of one spreadsheet or database should lead to transfer to different spreadsheets and databases and a range of other software packages. Baldwin and Ford (1988) provide a model that highlights several factors that influence transfer, including the similarity of the training and transfer situation, the nature of the training methods used, and the extent to which environmental factors reinforce transfer. Annett and Sparrow (1986) explained that transfer would be best if the stimulus situations and behaviors trained shared identical elements with stimuli in the work environment and the behaviors required there. Because these are seldom identical it is important to use methods of training that encourage learners to bridge the gap and transfer their skills (Hesketh 1997b), and to create an environment that reinforces them for doing so. A transfer of training needs analysis can be used to discover how best to design training to increase transfer knowledge and skill (Hesketh 1997a). For example, a transfer of training needs analysis for a team leader in fire fighting would highlight the types of decisions that needed to be made in different fire incidents and the specific cues used in making the decisions. This information would be used to design training that emphasized practice of these decision processes in a range of systematically chosen contexts to facilitate transfer. This approach ensures that the cognitive skills required for transfer are practised during training, and that knowledge about transfer and likely barriers are discovered during training (Von Papstein and Frese 1988).

Druckman and Bjork (1992) has highlighted a dilemma in that the methods of training that trainees enjoy, and that often lead to better performance on the training task, are not necessarily the ones that enhance long term retention and transfer. Trainees enjoy methods of training that require less effortful cognitive processing. Yet to facilitate transfer, trainees need to engage in the problem solving and active processes that they will be required to do on transfer. This may create motivational difficulties. Designing the training to provide an appropriate level of challenge is important for both motivation and transfer.

4.4 Simulation and Transfer

Early debate about the appropriate level of physical and psychological fidelity of simulation for training remains important, but has been incorporated into the more general issue of transfer and generalization. A high fidelity simulation may facilitate transfer to a single context, but lower fidelity simulators may be more appropriate if transfer is required to a range of situations. Current research issues are addressing the

best ways of integrating simulation with other forms of training, and how to optimize the level of fidelity for the particular purpose and type of simulation. The debate is being informed by research on the best way of combining rules and examples for training. Simulators provide an ideal opportunity to structure systematic exposure to a carefully chosen set of exemplar situations.

Simulators have traditionally been used for training, but their potential use is much more widespread. Simulation may provide a more realistic selection task for situations that require dynamic decision-making. Simulations are also being used as a way of signing off competency levels. Questions remain about how to deal with re-testing with a simulator, e.g., whether the trainee should be given an opportunity to perform again on exactly the same sequence as used during an initial test or a transfer problem. Here the research on transfer of training is critical, and may be of use in resolving the reassessment debate. This illustrates the ways in which selection, assessment, and training are strongly related areas of research and practice.

4.5 Organizational Issues

Organizational characteristics often influence transfer success. Trainees develop expectations about whether or not it will pay off when they use what they have learned in training. Often, companies teach one thing and reward a completely different behavior. For example, trainees may learn to be cooperative in a training course, but may then be paid for their individual contribution in a highly competitive environment. In such situations there is no transfer. Trainees need to be reinforced for what they have learned, and to practise skills in circumstances where errors can be made without serious consequences. For example a practice niche can be created where a bank clerk who has learned a new program to calculate mortgages is provided with an opportunity to practise it first while answering written requests. Thus, the customer does not see all the mistakes the bank clerk makes when using the new program.

4.6 Evaluating Training

The importance of evaluation has always been emphasized in training, although until recently approaches were somewhat traditional and limited. The importance of training evaluation has been recognized because intuitive guesses about what works are often wrong. Druckman and Bjork (1991) concluded that many companies in the USA continue to use methods of training known to be suboptimal for transfer because they rely on short term reactions evaluation, rather than examining the longer term retention and transfer of skills. Training evaluation can also be used as a way of indicating the importance of the skills learned, and of providing an opportunity to practise. Integrating training evaluation into the training plan is the best way of achieving this. The more detailed understanding of ways in which knowledge structures change with skill acquisition has also provided a basis for evaluating training. For example, experts tend to have hierarchically organized knowledge structures, and are able to read off solutions to problems far more quickly. These ideas can be used to suggest innovative ways of evaluating training programs.

Bibliography

Alluisi E A 1991 The development of technology for collective training: SIMNET, a case history. *Human Factors* **33**: 343–62

Annett J, Sparrow J 1986 Transfer of training: A review of research and practical implications. *Programmed Learning and Educational Technology* (*PLET*) **22**(2): 116–24

Baldwin T T, Ford J K 1988 Transfer of training: A review and directions for future research. *Personnel Psychology* **41**: 63–105

DeLosh E L, Busemeyer J R, McDaniel M A 1997 Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **23**: 968–86

Downs S, Perry P 1984 Developing learning skills. *Journal of European Industrial Training* **8**: 21–6

Druckman D, Bjork R A 1991 *In the Mind's Eye: Enhancing Human Performance*. National Academy Press, Washington, DC

Dweck C S, Leggett E L 1988 A social-cognitive approach to motivation and personality. *Psychological Review* **95**: 256–73

Frese M, Brodbeck F C, Heinbokel T, Mooser C, Schleiffenbaum E, Thiemann P 1991 Errors in training computer skills: On the positive functions of errors. *Human–Computer Interaction* **6**: 77–93

Frese M, Zapf D 1994 Action as the core of work psychology: A German approach. In: Triandis H C, Dunnette M D, Hough L M (eds.) *Handbook of Industrial and Organizational Psychology*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA, Vol. 4, pp. 271–340

Hesketh B 1997a W(h)ither dilemmas in training for transfer. *Applied Psychology: An International Review* **46**: 380–6

Hesketh B 1997b Dilemmas in training for transfer and retention. *Applied Psychology: An International Review* **46**: 317–39

Latham G P, Saari L M 1979 Application of social-learning theory to training supervisors through behavioural modelling. *Journal of Applied Psychology* **64**: 239–46

Martocchio J J 1994 Effects of conceptions of ability on anxiety, self-efficacy, and learning in training. *Journal of Applied Psychology* **79**: 819–25

Marx R D 1982 Relapse prevention for managerial training: A model for maintenance of behavior change. *Academy of Management Review* **7**: 433–41

Quinones M A, Ehrenstein A 1997 *Training for a Rapidly Changing Workplace: Applications of Psychological Research*. APA, Washington, DC

Salas E, Bower C A, Rhodenizer L 1998 It is not how much you have but how you use it: toward a rational use of simulation to support aviation training. *International Journal of Aviation Psychology* **8**: 197–208

Schmidt R A, Bjork R A 1992 New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science* **3**: 207–17.

Smith E, Ford J K, Kozlowski S 1997 Building adaptive expertise: Implications for training design. In: Quinones M A, Ehrenstein A (eds.) *Training for a Rapidly Changing Workplace: Applications of Psychological Research*. APA, Washington, DC

Volpert W, Frommann R, Munzert J 1984 Die Wirkung heuristischer Regeln im Lernprozeß. *Zeitschrift für Arbeitswissenschaft* **38**: 235–40

Von Papstein P, Frese M 1988 Transferring skills from training to the actual work situation: The role of task application knowledge, action styles, and job decision latitude. In: Soloway E, Frye D, Shepard S B (eds.) *Human Factors in Computing Systems, ACM SIGCHI Proceedings, CHI' 88*, pp. 55–60

B. Hesketh and M. Frese

Simultaneous Equation Estimates (Exact and Approximate), Distribution of

A simple example of a system of linear simultaneous equations may consist of production and consumption functions of a nation: $Y = a + bK + cL +$ error, and $C = d + eY +$ error. The variables Y, K, L, and C represent the gross domestic product (GDP), the capital equipment, the labor input, and the consumption, respectively.

These variables are measures of the level of economic activity of a nation. In the production function, Y increases if the inputs K and/or L increase. C increases if Y increases in the consumption equation. Each equation is modeled to explain the variation in the left-hand side 'explained' variable by the variation in the right-hand side 'explanatory' variables. Error terms are added to analyze numerically the effect of the neglected factors from the right-hand side of the equation. These equations are different from the regression equations since the 'explained' variable Y is the 'explanatory' variable in the C equation, and Y and C are simultaneously determined by the two equations. Estimation of unknown coefficients and the properties of estimation methods are not straightforward compared with the ordinary least squares estimator.

In practice, this kind of simultaneous equation system is extended to include more than 100 equations, and regularly updated to measure the economic activities of a nation. It is indispensable to analyze numerically the effect of policy changes and public investments.

In this article, the statistical model and the estimation methods of all the equations are first explained, followed by the estimation methods of a single equation and their asymptotic distributions. Explained next are the exact distributions, the asymptotic expansions, and the higher order efficiency of the estimators.

1. *The System of Simultaneous Equations and Identification of the System*

We write the structural form of a system consisting of G simultaneous equations as

$$y_i = Y_i\,\beta_i + Z_i\,\gamma_i + u_i \equiv (Y_i\ Z_i)\delta_i + u_i, \qquad i = 1,\ldots, G \quad (1)$$

where y_i and Y_i are 1 and G_i subcolumns in the $T \times G$ matrix of whole endogenous variables $Y = (y_i, Y_i, Y_{ei})$, Z_i consists of K_i subcolumns in the $T \times K$ matrix of whole exogenous variables Z, β_i and γ_i are $G_i \times 1$ and $K_i \times 1$ column vectors of unknown coefficients, $\delta_i = (\beta_i', \gamma_i')'$, and u_i is the $T \times 1$ error term. This system of G equations with T observations is frequently summarized in a simple form

$$YB + Z\Gamma = U \quad (2)$$

the ith column of which is $y_i - Y_i\beta_i - Z_i\gamma_i = u_i$, i.e., Eqn. (1). The ith columns of B and Γ may be denoted as b_i and c_i where $(G - G_i - 1)$ and $(K - K_i)$ elements are zero so that $y_i - Y_i\beta_i - Z_i\gamma_i = Yb_i + Zc_i$. Zero elements are called zero restrictions. It is assumed that each row of U is independently distributed as $N(0, \Sigma)$. The reduced form of Eqn. (2) is

$$Y = Z\Pi + V \quad (3)$$

where the $K \times G$ reduced form coefficient matrix is $\Pi = -\Gamma B^{-1}$, and the $T \times G$ reduced form error term is $V = UB^{-1}$. Each row of V is assumed to be independently distributed as $N(0, \Omega)$, and then $\Sigma = B^{-1}, \Omega B^{-1}$.

The definition $\Pi = -\Gamma B^{-1}$, or $-\Pi B = \Gamma$ is the key to identify structural coefficients. Coefficients in β_i are identified if they can be uniquely determined by the equation $-\Pi b_i = c_i$ given Π. This equation is reduced to $-\Pi_0(1, \beta_i')' = 0$ denoting the $(K - K_i) \times (1 + G_i)$ submatrix in Π as Π_0. (Rows and columns are selected according to the zero elements in c_i and non-zero elements in b_i, respectively.) Given Π_0, this includes $(K - K_i)$ linear equations and G_i unknowns, and β_i is solvable if rank $(\Pi_0) = G_i$. This means $(K - K_i)$ must be at least G_i, or $L = K - K_i - G_i$ must be at least 0. If $L = 0$, β_i is uniquely determined. For the positive L, there are L linearly dependent rows in Π_0 since only G_i rows are necessary to determine β_i uniquely. Once β_i is determined, γ_i is determined by other K_i equations through $-\Pi b_i = c_i$. L is called the number of the degrees of overidentifiability of the ith equation.

Structural coefficients are not uniquely estimable if they are not identified (see *Statistical Identification and Estimability*).

2. *Estimation Methods of the Whole System and the Asymptotic Distribution*

The full information maximum likelihood (FIML) estimator of all nonzero structural coefficients δ_i, $i = 1,\ldots, G$, follows from Eqn. (3). Since it is in a linear regression form, the likelihood function can first be minimized with respect to Ω. Once Ω is replaced by the first-order condition, the likelihood function is concentrated where only B and Γ are unknown. The concentrated likelihood function is proportional to

$$\ln|\Omega_R|, \Omega_R = (Y-Z\Pi_R)'(Y-Z\Pi_R)/T, \quad \Pi_R = -\Gamma B^{-1}, \tag{4}$$

and all zero restrictions are included in B and Γ matrices. In the FIML estimation, it is necessary to minimize $|\Omega_R|$ with respect to all non-zero structural coefficients.

The FIML estimator is consistent, and the asymptotic distribution is derived by the central limit theorem. Stacking δ_i, $i = 1,\ldots, G$ in a column vector δ, the FIML estimator $\hat{\delta}$ asymptotically approaches $N(0, -I^{-1})$ as follows:

$$\sqrt{T}(\hat{\delta}-\delta)\xrightarrow{D}N(0, -I^{-1}), \quad I = \lim_{T\to\infty}\frac{1}{T}E\left(\frac{\partial^2\ln|\Omega_R|}{\partial\delta\,\partial\delta'}\right). \tag{5}$$

I is the limit of the average of the information matrix, i.e., $-I^{-1}$ is the asymptotic Cramer–Rao lower bound. Then the FIML estimator is the best among consistent and asymptotically normal (BCAN) estimators.

The right-hand side endogenous variable Y_i in (1) is defined by a set of G_i columns in (3) such as $Y_i = Z\Pi_i + V_i$. By the definition of V, Y_i or, equivalently, V_i is correlated with u_i since columns in U are correlated with each other. The least squares estimator applied to (1) is inconsistent because of the correlation between Y_i and u_i. Since Z is assumed to be not correlated with U in the limit, Z is used as K instruments in the instrumental variable method estimator. Premultiplying Z' to (1), it follows that

$$\begin{aligned} Z'y_i &= (Z'Y_i, Z'Z_i)\delta_i + Z'u_i \\ &= (0,\ldots, 0, Z'Y_i, Z'Z_i, 0,\ldots, 0)\delta + u_i^*, \\ &\qquad i = 1,\ldots, G \end{aligned} \tag{6}$$

where the $K\times 1$ transformed right-hand side variables $Z'Y_i$ is not correlated with u_i^* in the limit. Stacking all G transformed equations in a column form, the G equations are summarized as $w = X\delta + u^*$ where w and u^* stack $Z'y_i$ and u_i^*, $i = 1,\ldots, G$, respectively, and are $GK\times 1$. The covariance between u_i^* and u_j^* is $\sigma_{ij}(Z'Z)$ which is the ith row and jth column sub-block in the covariance matrix of u^*. (The whole covariance matrix can be written as $\Sigma\otimes(Z'Z)$ where $\otimes$ signifies the Kroneker product.) Once Σ is estimated consistently (by the 2SLS method explained in the next section), δ is efficiently estimated by the generalized least squares method

$$\hat{\delta}_{3\mathrm{SLS}} = \{X'[\hat{\Sigma}^{-1}\otimes(Z'Z)^{-1}]X\}^{-1}\{X'[\hat{\Sigma}^{-1}\otimes(Z'Z)^{-1}]w\}. \tag{7}$$

This is the three-stage least squares (3SLS) estimator by Zellner and Theil (1962). The assumption of the normal distribution error is not required in this estimation. The 3SLS estimator is consistent and is BCAN since it has the same asymptotic distribution as the FIML estimator.

3. *Estimation Methods of a Single Equation and the Asymptotic Distribution*

An alternative way of estimating structural coefficients is to pick up one structural equation, the ith, in a G-equation system, and estimate δ_i neglecting zero restrictions in other equations. Because of this, the other $(G-1)$ structural equations can be rewritten equivalently as $(G-1)$ reduced form equations. The limited information maximum likelihood (LIML) estimator by Anderson and Rubin (1949) applies the FIML method to a $(1+G_i)$-equation system consisting of (1) and $Y_i = Z\Pi_i + V_i$. This means the first column and the second G_i columns of B are $(1, -\beta_i')'$ and $(0, I)'$, respectively, and the first column and the second G_i columns of Γ are $(\gamma_i', 0')'$ and Π_i, respectively. (If we denote Y as (y_i, Y_i, Y_{ei}), Y_{ei} is weakly exogenous in estimating (1) or, equivalently, the structural and reduced form parameters of (y_i, Y_i) given Y_{ei} are variation-free from the reduced form coefficient of Y_{ei}. Then Y_{ei} is omitted from (2), (3), (4), and (8).) Using these limited information B and Γ matrices, we minimize $|\hat{\Omega}_R|$ with respect to δ_i and Π_i. Defining $P_F = F(F'F)^{-1}F'$ for any full column matrix F, $G = (y_i, Y_i)'(P_Z - P_{Z_i})(y_i, Y_i)$, and $C = (y_i, Y_i)'(I-P_Z)(y_i, Y_i)$, it turns out that β_i is estimated by minimizing the least variance ratio

$$\lambda(\beta_i) = \frac{(1, -\beta_i')G(1, -\beta_i')'}{(1, -\beta_i')C(1, -\beta_i')'}. \tag{8}$$

γ_i is estimated by the least squares method applied to (1) replacing β_i with the estimator.

The two stage least squares (2SLS) estimator is the generalized least squares estimator applied to Eqn. (6) using $Z'Z$ as the weight matrix. (See Eqn. (10), where k is set to 1.) The assumption of the normal distribution of the error term is not required in this estimation. Both LIML and 2SLS estimators are consistent, and the large sample distribution is

$$\sqrt{T}(\hat{\delta}_i - \delta_i)\xrightarrow{D}N(0, -I^{-1}(\delta_i, \delta_i)) \tag{9}$$

where I is calculated similarly to Eqn. (5) but Ω_R is defined using the limited information B and Γ matrices, and only the diagonal submatrix of I^{-1} which corresponds to δ_i is used. (Partial derivatives are calculated with respect to δ_i and columns in Π_i.) This asymptotic distribution is a particular case of Eqn. (5). Both estimators are consistent and BCAN under the zero restrictions imposed on Eqn. (1).

The k-class estimator $\hat{\delta}_i(k)$ unifies both LIML and 2SLS estimators (Theil 1961). It is

$$\hat{\delta}_i(k) = \begin{pmatrix} Y_i'P_Z Y_i - (k-1)Y_i(I-P_Z)Y_i & Y_i'Z_i \\ Z_i'Y_i & Z_i'Z_i \end{pmatrix}^{-1} \times \begin{pmatrix} Y_i'P_Z - (k-1)Y_i'(I-P_Z) \\ Z_i' \end{pmatrix} y_i. \quad (10)$$

This is the least squares estimator, the 2SLS estimator, and the LIML estimator when k is 0, 1, and $1+\lambda$, respectively. There are two important properties in the k-class estimator. It is consistent if $p\lim_{T\to\infty} k = 1$, and is BCAN if $p\lim_{T\to\infty}\sqrt{T}(k-1) = 0$. If k satisfies these conditions, the k-class estimator is consistent and BCAN even when $Y_i'(I-P_Z)Y_i$ and $Y_i'(I-P_Z)y_i$ are replaced with any matrix and a vector of order $O_P(T)$.

4. *Exact Distributions of the Single-equation Estimators*

Several early studies compared the bias and mean squared errors of OLS, LIML, 2SLS, FIML, and 3SLS estimators by the Monte Carlo simulations since all but OLS estimators are consistent and are indistinguishable. The OLS estimator was often found as reliable as other consistent estimators. Later, the studies went on to t ratios, and the real defect of OLS estimators was found: the deviation from the standard normal distribution is worse than any other simultaneous equation methods. See Cragg (1967) for related papers.

Drawing general qualitative comparisons from simulations is difficult since simulations require setting values of all population parameters. Simulation studies on the small sample properties led to the derivation of the exact distributions, which were expected to permit the drawing of general comparisons without depending on the particular parameter values.

If $n\times 1$ column vectors x_t, $t = 1,\ldots, T$ are independently distributed $N(m_t, \Omega)$, the density function of $\Sigma_{t=1,T}\, x_t x_t'$ is the non-central Wishart matrix denoted as $W_n(T, \Omega, M)$ where the non-centrality parameter is $M = \Sigma_{t=1,T}\, m_t m_t'$ (Anderson 1958, Chap. 13). The study of the exact distribution of the single-equation estimators started from the fact that $G = (G_{kl})$ and $C = (C_{kl})$, k, $l = 1,\ldots, 1+G_i$ in Eqn. (8) are the noncentral Wishart matrix $W_{1+G_i}(K-K_i, \Omega, M)$ with the noncentrality parameter matrix $M = \Pi_R' Z'(P_Z - P_Z)Z\Pi_R$, and the central Wishart matrix $W_{1+G_i}(T-K, \Omega, 0)$, respectively. The 2SLS, OLS, and LIML estimators of β_i are $G_{22}^{-1}\, G_{21}$, $(G_{22}+C_{22})^{-1}(G_{21}+C_{21})$, and $(G_{22}-\lambda C_{22})^{-1}(G_{21}-\lambda C_{21})$, respectively, and λ is the minimum root in the polynomial equation $|G-\lambda C| = 0$. Since all estimators are functions of elements in the G and C matrices, their distributions can be characterized by degrees of freedom of the two Wishart matrices, M and Ω matrices.

In deriving the exact density functions, the 2SLS and OLS estimators can be treated in a similar way. In the 2SLS estimator, the joint density of G is transformed into the joint density of $G_{22}^{-1}\, G_{21}$, $G_{11}-G_{12}G_{22}^{-1}\, G_{21}$, and G_{22}. Integrating out $G_{11}-G_{12}G_{22}^{-1}G_{21}$ and G_{22} results in the joint density of $G_{22}^{-1}G_{21}$. The resulting density function includes infinite terms, and zonal polynomials when G_i is greater than one. A pedagogical derivation is found in Press (1982, Chap. 5). In the LIML estimator, the joint density of G and C is transformed into that of characteristic roots and vectors. Since $\hat{\beta}$ is rewritten as a ratio of elements in a characteristic vector, the density function is derived by integrating out unnecessary random variables from the joint density function. However, the analytical operations are not easy when there are many endogenous variables. See Phillips (1983, 1985) to find comprehensive reviews on the 2SLS and LIML estimators, respectively.

It was somewhat fruitless to derive exact distributions because these include nuisance parameters and infinite terms. It was difficult to draw general conclusions on the qualitative properties of the estimators from the numerical evaluations of these distributions. See Anderson et al. (1982).

Qualitative properties of the estimators followed from the exact moments of estimators. Kinal (1980) proved that the (fixed) k-class estimator in a multiple endogenous variables case has moments up to $(T-K_i-G_i)$ if $0 \leqslant k < 1$, and up to L if $k = 1$. Mariano and Sawa (1972) proved that, in the $G_i = 1$ case, the mean and variance of the LIML estimator do not exit. (In Monte Carlo simulations, the bias of LIML estimators was often found to be smaller than that of others even though the exact mean is infinite. This showed the clear limitation of the simulation methods.)

5. *Asymptotic Expansions of the Distributions of the Single-equation Estimators*

Asymptotic expansion of the distribution was introduced as an analytical tool which is more accurate than the asymptotic distribution but is less complicated than the exact distributions. For instance, the

t ratio statistic, say X, which is commonly used in econometrics, has the density function $f(x) = c \cdot [1+(x^2/m)]^{-(1+m)/2}$ where c is a constant and m is the degree of freedom under conditions including the normally distributed error terms. Since the mean and variance of X is 0 and $m/(m-2)$, the standardized statistic $Z = \sqrt{(m-2)/m} \cdot X$ has the density $f(z) = c' \cdot [1+(z^2/(m-2))]^{-(1+m)/2}$ where c′ is a new constant. This density function is expanded to the third-order term as

$$f(z) = \phi(z)\{1+\frac{1}{4m}(z^4-6z^2+3)\}+o(m^{-1}) \quad (11)$$

where $\phi(z)$ is the standard normal density function, and the constant is adjusted so that the area under the curve is one. (Since the t distribution is symmetric around the origin, the $O(1/\sqrt{m})$ term does not appear in the right-hand side of the equation.) Rewriting in terms of X, the asymptotic expansion of the t statistic with m degrees of freedom is

$$f(x) = \phi(x)+\frac{1}{4m}(x^4-2x^2-1)\phi(x)+o(m^{-1}) \quad (12)$$

The first term in the right-hand side is the $n(0, 1)$ density function. The second term in the right-hand side converges to zero if the limit is taken with respect to m. This second term gives deviation of $f(x)$ from ϕ(x), and is called the third-order term. For finite m, the third-order term is expected to improve the standard normal approximation. The numerical evaluation of this expansion is easy.

The asymptotic expansions in the simultaneous equation estimators are long and include nuisance parameter matrices such as q below. See Phillips (1983) for a review and Phillips (1977) for the validity of expansion. The asymptotic expansion does not require the assumption of the normal distribution of the error term. Fujikoshi et al. (1982) gave the expansion of the joint density of the estimators of δ_i. In their study, the bias of the estimators is calculated from the asymptotic expansions as

$$AM(\sqrt{T}(\hat{\delta}_i-\delta_i)) = \frac{1}{\sqrt{T}}(dL-1)I^{-1}q+o(T^{-1}),$$

$$q = \frac{1}{\sigma^2}\begin{pmatrix} \varpi_{21}-\Omega_{22}\beta_i \\ 0 \end{pmatrix}, \quad (13)$$

where $\hat{\delta}$ is the estimator, d is 1 and 0 fo1r 2SLS and LIML estimators, respectively, and the Ω matrix is partitioned into submatrices conformable with the partitions of G and C matrices. $AM(\cdot)$ stands for the mean operator, but uses the asymptotic expansion for the density function. (Recall that the exact mean does not exist in the LIML estimator.) It is possible to compare the two estimators in terms of the calculated bias. For example, the bias of the 2SLS estimator is 0 when the degree of overidentifiability L is 1.

Further, the mean of the squared errors was calculated from the asymptotic expansions and used to compare estimators. It was proved that the mean squared error of the 2SLS estimator is smaller than that of the LIML estimator when L is less than or equal to 6. Historically, this kind of comparison of 'approximate' mean squared errors goes back to the 'Nagar expansion' by Nagar (1959) and the 'small disturbance expansion' by Kadane (1971). These qualitative comparisons gave researchers some guidance about the choice of estimators.

It was interesting to examine the accuracy of the approximations calculated by the asymptotic expansions of distributions. If the asymptotic expansions were accurate, calculation of the exact distributions could be avoided, and properties of estimators could be found easily from the approximations. However, the approximations were found not to be accurate enough to replace the exact distributions. There are many cases where the asymptotic expansion is accurate when the asymptotic distribution, the first term in the expansion, is already accurate; the asymptotic expansion is inaccurate when the asymptotic distribution is inaccurate. In particular, the asymptotic expansions are inaccurate when

(a) The value of asymptotic variance is small;
(b) The value of L is large; or
(c) The structural error term is highly correlated with the right-hand side endogenous variable.

It is noted, first, $Z'Z/T$ is assumed to converge a nonsingular fixed matrix in the asymptotic theory. Second, for an attempt to improve the accuracy of asymptotic distributions by incorporating large L values, see Morimune (1983). Third, the asymptotic expansions cannot trace closely the skewed exact distributions that happen particularly when the correlation is high.

6. *Higher-order Efficiency of the System and the Single-equation Estimators*

The asymptotic expansions of distributions can be used in comparing the probabilities of concentration of estimators about the true parameter values. One estimator is more desirable than another if its probability is greater than that of the other. This measure was used in comparing the single-equation estimators, and some qualitative results were derived.

Furthermore, the third-order efficiency criterion was brought in the comparisons. This criterion requires that estimators be adjusted to have the same asymptotic bias as in Eqn. (13). Then the adjusted estimators are compared, and the maximum likelihood

estimator is proved most efficient. It has the highest concentration about the true parameter values in terms of the asymptotic expansion of distribution to the third-order $O(1/T)$ terms. (Akahira and Takeuchi (1981), for example.) The adjusted maximum likelihood estimator has the smallest mean-squared error at the same time since the difference among estimators is found only in the mean-squared errors.

In whole system estimation, the FIML estimator is third-order efficient. The 3SLS estimator is less efficient than the FIML estimator in terms of the asymptotic probability of concentration once the bias of the two estimators is adjusted to be the same. Morimune and Sakata (1993) derived a simple adjustment of the 3SLS estimator so that the adjusted estimator has the same asymptotic expansion as the FIML estimator to the third-order $O(1/T)$ terms. This estimator is explained by modifying Eqn. (7). In Eqn. (6), Y_i is replaced by $\hat{Y}_i \equiv Z\,\hat{\Pi}_i = Z(Z'Z)^{-1} Z'Y_i$ so that the X matrix consists of $Z'\hat{Y}_i$ and $Z'Z_i$. In the modified estimator, we estimate Σ and Π by the first round 3SLS estimator and replace $\hat{Y}_i$ in X by $\tilde{Y}_i \equiv Z\tilde{\Pi}_i$ where $\tilde{\Pi}_i$ consists of proper subcolumns in $\tilde{\Pi} = -\hat{\Gamma}\hat{B}^{-1}$. The new X matrix is denoted as $\tilde{X}$. Finally, the new estimator is

$$\tilde{\delta}_{\mathrm{M3SLS}} = \{\tilde{X}'[\tilde{\Sigma}^{-1}\otimes(Z'Z)^{-1}]X\}^{-1}\{\tilde{X}'[\tilde{\Sigma}^{-1}\otimes(Z'Z)^{-1}]w\}. \quad (14)$$

This estimator has the same asymptotic expansion as the FIML estimator to the third-order terms and is third-order efficient.

The LIML and 2SLS estimators are simple cases of the FIML and 3SLS estimators, respectively. Then the modified 2SLS estimator which follows from Eqn. (14) has the same asymptotic expansion as the LIML estimator to the third-order term. The LIML estimator and the modified 2SLS estimator are third-order efficient in the single-equation estimation.

7. *Conclusion*

Laurence Klein received the 1980 Nobel prize for the creation of a macroeconometric model which is an empirical form of a simultaneous equation system, and for the application to the analysis of economic fluctuations and economic policies. The macroeconometric model became a standard tool to analyze the economies and policies of nations. Trygve Haavelmo received the 1989 Nobel prize for his contribution in the analysis of simultaneous structures. Haavelmo, together with other researchers at the Cowles Commission for Research in Economics, then at the University of Chicago, became the founders of simultaneous equation analysis in econometrics. Part of their research is collected in Hood and Koopmans (1953). Studies on the exact and approximate distributions of estimators came after the research conducted at the Cowles group, and helped to make econometrics rigorous.

Access to computers was the main concern when econometric model-building started spreading all over the world in the 1970s. Since then, computer facilities surrounding econometric model-building have changed greatly. Bulky mainframe computer have been replaced by personal computers. Computer programs were written individually, mostly in Fortran, and were used for regression analyses in model estimation as well as for simulation studies in econometrics theory. The packaged least squares programs replaced the individually written programs later in model estimation. They run on personal computers and have facilitated greatly the conducting of empirical studies.

See also: Simultaneous Equation Estimation: Overview

Bibliography

Akahira M, Takeuchi K 1981 *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*. Springer, New York

Anderson T W 1958 *An Introduction to Multivariate Statistical Analysis*. Wiley, New York

Anderson T W, Rubin H 1949 Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* **20**: 46–63

Anderson T W, Kunitomo N, Sawa T 1982 Evaluation of the distribution function of the limited information maximum likelihood estimator. *Econometrica* **50**: 1009–27

Basmann R L 1957 A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* **25**: 77–83

Cragg J G 1967 On the relative small sample properties of several structural equation estimators. *Econometrica* **35**: 89–110

Fujikoshi Y, Morimune K, Kunitomo N, Taniguchi M 1982 Asymptotic expansions of the distributions of the estimates of coefficients in a simultaneous equation system. *Journal of Econometrics* **18**: 191–205

Hood W C, Koopmans T C 1953 *Studies in Econometric Methods*. Wiley, New York

Kadane J B 1971 Comparison of k-class estimates when the disturbance is small. *Econometrica* **39**: 723–37

Kinal T W 1980 The existence of moments of k-class estimators. *Econometrica* **48**: 241–9

Mariano R S, Sawa T 1972 The exact finite sample distribution of the limited information maximum likelihood estimator in the case of two included endogenous variables. *Journal of the American Statistical Association* **67**: 159–63

Morimune K 1983 Approximate distributions of k-class estimators when the degree of over-identifiability is large compared with the sample size. *Econometrica* **51**: 821–41

Morimune K, Sakata S 1993 Modified three stage-least squares estimator which is third-order efficient. *Journal of Econometrics* **57**: 257–76

Nagar A L 1959 The bias and moment matrix of the general kappa-class estimators of the parameters in simultaneous equations. *Econometrica* **27**: 575–95
Phillips P C B 1977 The general theorem in the theory of asymptotic expansions as approximation to the finite sample distributions of econometric estimators. *Econometrica* **45**: 1517–34
Phillips P C B 1983 Exact Phillips. In: Griliches Z, Intriligator M D (eds.) *Handbook of Econometrics*. North-Holland, Amsterdam, Vol. 1, Chap. 8
Phillips P C B 1985 The exact distribution of LIML: 2. *International Economic Reviews* **25**(1): 249–61
Press S J 1982 *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference* [original edn. 1972, 2nd edn.] Robert E Krieger, Malabar, FL
Theil H 1961 *Economic Forecasts and Policy*, 2nd edn. North-Holland, New York, pp. 231–2, 334–6
Zellner A, Theil H 1962 Three-stage least squares: simultaneous estimation of simultaneous equations. *Econometrica* **30**: 54–78

K. Morimune

Simultaneous Equation Estimation: Overview

Simultaneous equations are important tools for understanding behavior when two or more variables are determined by the interaction of two or more relationships in such a way that causation is joint rather than unidirectional. Such situations abound in economics, but also occur elsewhere. Haavelmo (1943, 1944) began their modern econometric treatment.

The simplest economic example is the interaction of buyers and sellers in a competitive market which jointly determines the quantity sold and price. Another important example is the interaction of workers, consumers, investors, firms, and government in determining the economy's output, employment, price level, and interest rates, as in macroeconometric forecasting models. Even when one is interested only in a single equation, it often is best interpreted as one of a system of simultaneous equations.

Strotz and Wold (1960) argued that in principle every economic action is a response to a previous action of someone else, but even they agreed that simultaneous equations are useful when the data are yearly, quarterly, or monthly, because these periods are much longer than the typical market response time.

This article discusses the essentials of simultaneous equation estimation using a simple linear example.

1. *A Simple Supply–Demand Example*

Let q and p stand for the quantity sold and price of a good, y and w for the income and wealth of buyers (assumed independently determined), u_1 and u_2 for unobservable random shocks, and Greek letters for unknown constant parameters. Suppose that market equilibrium, in which the price is such that suppliers want to sell the same quantity that demanders what to buy, is described by these linear equations:

$$\text{supply:} \quad q = \gamma_1 + \beta_1 p + u_1, \quad \beta_1 > 0 \tag{1}$$

$$\text{demand:} \quad q = \gamma_2 + \delta_2 y + \varepsilon_2 w + \beta_2 p + u_2, \quad \beta_2 < 0 \tag{2}$$

Neither equation alone can determine either p or q, because each equation contains both: these are *simultaneous equations*. Solve them for p and q thus:

$$p = \pi_{11} + \pi_{12} y + \pi_{13} w + (u_2 - u_1)/\Delta \tag{3}$$

$$q = \pi_{21} + \pi_{22} y + \pi_{23} w + (\beta_1 u_2 - \beta_2 u_1)/\Delta \tag{4}$$

where

$$\pi_{11} = (\gamma_2 - \gamma_1)/\Delta, \quad \pi_{12} = \delta_2/\Delta, \quad \pi_{13} = \varepsilon_2/\Delta \tag{5}$$

$$\pi_{21} = (\beta_1 \gamma_2 - \beta_2 \gamma_1)/\Delta, \quad \pi_{22} = \beta_1 \delta_2/\Delta, \quad \pi_{23} = \beta_1 \varepsilon_2/\Delta \tag{6}$$

and

$$\Delta = \beta_1 - \beta_2 \tag{7}$$

2. *Types of Variables: Structural and Reduced Form Equations*

The variables p and q which are to be explained by the model are *endogenous*. Equations (1) and (2) are structural equations. Each of them describes the behavior of one of the building blocks of the model, and (as is typical) contains more than one endogenous variable. Equations (3) and (4) are reduced form equations. Each of them contains just one endogenous variable, and determines its value as a function of parameters, shocks, and explanatory variables (here y and w).

The explanatory variables are predetermined if the shocks for any given period are independent of the explanatory variables for that period and all previous periods; they are *exogenous* if the shocks for each period are independent of the explanatory variables for every period. Thus all exogenous variables are predetermined, but not conversely. For example, the value of an endogenous variable from a previous period cannot be exogenous, but it is predetermined if the shocks are serially independent.

3. *The Need for Estimation of Parameters*

One job of a reduced form equation is to tell how an endogenous variable responds to a change in any predetermined or exogenous variable. Typically, no single structural equation can do this. Another job of

a reduced form equation is to forecast the value of an endogenous variable in a future period, based on expected future values of the predetermined and exogenous variables and of the shocks (the latter can safely be set at zero in most cases). Typically, no single structural equation can do this either. If the reduced form equations are to do these jobs, numerical values of their parameters are needed (in this example, values of the πs in Eqns. (3) and (4)).

Numerical values of the structural parameters are needed as well (in this example, the βs, γs, δ_2 and ε_2 in Eqns. (1) and (2)). There are several reasons. First, one wants to understand each of the system's separate building blocks in its own right. Second, in overidentified cases (see Sect. 7) better estimates of the reduced form can be obtained by solving the estimated structure (as in Eqns. (5)–(7) than by estimating the reduced form directly. Third, if forecasts made by the reduced form are poor, one wants to know which of the structural equations fit the forecast-period's data poorly, and which (if any) fit well. This is discovered by inserting observed values of the variables into each estimated structural equation to obtain an estimate of its forecast-period shock. Then one can revise the poorly-fitting structural equation(s) and so try to improve the model's accuracy (see Sect. 11). Fourth, if one wants to find the new values of the reduced form parameters after a change in a structural parameter, one needs to know the old values of all the structural parameters, as well as the new value of the one that has changed. In the example, one can then use Eqns. (5)–(7) to compute the reduced form parameters. The critique of Robert Lucas (1976) provides a warning about pitfalls in doing this.

4. *Least Squares Estimators*

Least squares (LS) estimators (see *Linear Hypothesis*) of an equation's coefficients are biased if the shocks in each period are not independent of the explanatory variables in all periods. Clearly this is true of Eqns. (1) and (2), since Eqns. (3) and (4) show that both u_1 and u_2 influence both p and q. This is typically true of LS estimators of simultaneous structural equations, However, LS estimators often have small variances, so they may sometimes have acceptably small expected squared errors even when they are biased.

LS estimators of the π's in the reduced form, Eqns. (3) and (4) are unbiased if the shocks in each period have zero mean and constant variance, and are independent of y and w for all periods (so that y and w are exogenous). They have minimum variance among unbiased estimators and are consistent if in addition the shocks are uncorrelated across time. They remain consistent if the shocks are uncorrelated across time but y and w are predetermined rather than exogenous.

The generalized least squares (GLS) method is minimum variance unbiased if the explanatory variables are exogenous but the shocks are correlated across time. This method requires information about the variances and covariances of the shocks.

5. *The Identifiability of Structural Parameters*

Having estimates of the reduced form parameters, one can try to make them yield estimators of the structural parameters. In the example this means trying to solve the estimated version of Eqns. (5) and (7) for estimators of the γ's, β's, δ_2 and ε_2. Denote the LS estimators of the π's by $\hat{\pi}$. Equations (5) and (6) show that $\hat{\pi}_{22}/\hat{\pi}_{12}$ is one estimator of β_1, and $\hat{\pi}_{23}/\hat{\pi}_{13}$ is another. Either of them leads to an estimator of γ_1 in Eqn. (1). These are indirect least squares (ILS) estimators. They are not unbiased, but they are consistent if the LS reduced form estimators are consistent. The supply parameters β_1 and γ_1 are *identified*, meaning that the data and the model reveal their values (subject to sampling variation). Indeed, they are *overidentified*, because in small samples the two ways to estimate β_1 from $\hat{\pi}$'s yield different results. (If either δ_2 and ε_2 were zero, there would be only one way, and the supply equation would be *just identified*. If both δ_2 and ε_2 were zero, there would be no way, and the supply equation would be *unidentified*.) Equations (5)–(7) have an infinite number of solutions for the demand parameters β_2, γ_2, δ_2 and ε_2. Hence the data and the model do not reveal their values; they are unidentified.

Another way to see this is to imagine price and quantity data being generated by intersections of the supply and demand curves in the pq plane. Ignoring shocks, the supply curve is fixed, but the demand curve shifts when y or w changes. Hence the intersections reveal the slope and position of the supply curve as the demand curve shifts. But they reveal nothing about the slope of the demand curve.

The identifiability of parameters is crucial: they cannot be estimated if they are not identified. Reduced form parameters are usually identified, except in special cases. But the identifiability of structural parameters should be checked before one tries to estimate them. The least squares formula can be applied to an unidentified structural equation such as Eqn. (2), but the result is not an estimator of that equation. For a more detailed and general discussion, see *Statistical Identification and Estimability*.

6. *Simultaneous Equations Estimation Methods*

Many methods have been developed for estimating identifiable parameters of simultaneous structural equations without the bias and inconsistency of LS. Most of them exploit the assumed independence between shocks and predetermined variables. Some estimate one equation at a time, and others estimate the whole system at once.

7. Estimating One Structural Equation at a Time

The ILS method mentioned above is one way of estimating one equation that is part of a system. A second way is the instrumental variables (IV) method (Durbin 1954). LS and IV will be shown for Eqn. (1). Denote the deviation of each variable from its sample mean by an asterisk, for example, $p^* = p - \bar{p}$. Now rewrite Eqn. (1) in terms of deviations from sample means, thus eliminating the constant term γ_1; multiply it by p^*; sum it over all the sample observations; and divide the sum by $\sum p^{*2}$. The result is the LS estimator of β_1:

$$\hat{\beta}_1 = \sum q^* p^* / \sum p^{*2} = \beta_1 + \sum u_1^* p^* / \sum p^{*2} \qquad (8)$$

It is biased and inconsistent because the shock u_1 influences p via Eqn. (3) so that the error term, the last term in Eqn. (8), does not equal zero in expectation or in probability limit.

Now rewrite Eqn. (1) as before in terms of deviations from sample means, but this time multiply it by y^*, sum it over all observations, and divide by $\sum p^* y^*$. The result is an IV estimator of β_1:

$$\hat{\beta}_{1y} = \sum q^* y^* / \sum p^* y^* = \beta_1 + \sum u_1^* y^* / \sum p^* y^* \qquad (9)$$

If y is predetermined it is an *instrument*. In this case the IV estimator is consistent if u_1 has zero mean and constant variance and is serially independent, because then $\sum u_1^* y^*$ is zero in probability limit and $\sum p^* y^*$ is not. Similarly, if w is predetermined it is another instrument, and another IV estimator of β_1 is $\hat{\beta}_{1w} = \sum q^* w^* / \sum p^* w^*$. Clearly, these IV estimators are equivalent to the ILS estimators of β_1 obtained earlier. For example, the one based on the reduced form coefficients of y is

$$\hat{\pi}_{22}/\hat{\pi}_{12} = (\sum q^* y^* / \sum y^{*2}) / (\sum p^* y^* / \sum y^{*2}) = \sum q^* y^* / \sum p^* y^* = \hat{\beta}_{1y} \qquad (10)$$

Similarly, the other one is $\hat{\pi}_{23}/\hat{\pi}_{13} = \hat{\beta}_{1w}$. For a reduced form equation, IV is the same as LS, because the instruments are precisely the equation's predetermined variables.

The two-stage least squares (2SLS) method (Theil 1972, Basmann 1957) is another way of estimating one structural equation at a time. For an overidentified equation it is superior to ILS or IV because it results in one estimator rather than two or more; for a just identified equation it reduces to ILS and IV. It first computes the LS estimator of the reduced form equation for each endogenous variable that appears on the right side of the structural equation to be estimated (this is stage 1); it then replaces the observed data for those endogenous variables by the values calculated from the reduced form, and computes the LS estimator for the resulting equation (this is stage 2). For Eqn. (1), stage 1 is to compute the least squares estimators of the π's in the price equation (3) of the reduced form; the second stage is to compute $\hat{p} = \hat{\pi}_{11} + \hat{\pi}_{12} y + \hat{\pi}_{13} w$, substitute this $\hat{p}$ for p in (1), and compute the LS estimator $\sum q^* \hat{p}^* / \sum \hat{p}^{*2}$, which is the 2SLS estimator of β_1. 2SLS is an IV estimator, its instruments are the observed values of the equation's predetermined variables and the reduced-form-calculated values of the endogenous variables from the equation's right side. 2SLS is a *generalized* method of moments (GMOM) estimator (Hansen 1982). LS and IV are plain MOM estimators.

The k-class estimator (Nagar 1959) includes as special cases LS, 2SLS, and the limited information maximum likelihood estimator (LIML) (Anderson and Rubin 1949, 1950). LIML is similar to 2SLS (the two are the same for infinitely large samples), but it has largely been displaced by 2SLS because it requires iterative computations which 2SLS does not, and because it often has a larger variance and sometimes yields outlandish estimates (Kadane 1971, Theil 1972).

For an identified structural equation, ILS, IV, LIML, and 2SLS are consistent if the model's explanatory variables are predetermined and the shocks have zero means and constant variances and covariances and are independent across time. An advantage of these methods is that, unlike LS, if applied to an unidentified structural equation they fail to yield estimates.

The Bayesian method of moments (BMOM) method (Zellner 1998) obtains estimators of linear reduced form and structural equations without making assumptions about the likelihood function of the data. It assumes that the posterior expectations of the shocks given the data are uncorrelated with the predetermined variables, and that the differences between actual and estimated shocks have a covariance matrix of a particular form. Zellner shows how to find optimum BMOM estimators for several different loss functions, including a *precision* loss function which is a weighted sum of squares and cross-products of errors. For this loss function, the optimum BMOM estimator of an identified structural equation belongs to the k-class; it turns out to be LS if the ratio of the sample size to the number of predetermined variables in the model is 2, and it approaches 2SLS in the limit as this ratio grows without limit.

When a structural equation is overidentified, restrictions are placed on the reduced form parameters (such as $\pi_{22}/\pi_{21} = \pi_{32}/\pi_{31}$ in the example, because both ratios must equal β_1), but LS estimators of the reduced form ignore this information. Better estimators of the reduced form, because they use this information, can then be obtained by solving the estimated structure (see Sect. 3).

8. Estimating a Complete System

If in an identified complete simultaneous equations system the shocks are normally distributed and other suitable conditions are satisfied, an asymptotically efficient but computationally complex method of estimating all the parameters of a complete system is the full information maximum likelihood (FIML) method (Koopmans 1950, Durbin 1988). It involves maximizing the joint likelihood function of all the data and parameters, and requires iterative computations. It has largely been displaced by the *three stage least squares* (3SLS) method (Zellner and Theil 1962), which is also asymptotically efficient but is much easier to compute. 3SLS is an application of the GLS method. 3SLS gets the required information on the shocks' covariances from 2SLS estimates of the shocks; hence the name 3SLS.

9. Cross-section vs. Time Series Studies

So far the treatment has concerned time series data, where the observations describe the same country, city, family, firm, or what not, at successive periods of time. The same type of analysis is appropriate for cross-section data, where the observations describe different countries, firms, or what not, at a single point of time, with obvious modifications. The analysis has also been extended to panel data, that is, a time series of cross-sections.

10. The Decline of Simultaneous Equations Models

In recent years the econometric literature has paid diminishing attention to simultaneous equations models and their identifiability. Many modern textbooks give these topics little space, and that only near the end of the book (Davidson and MacKinnon 1993). Perhaps one reason is that modern computers can handle the estimation of nonlinear models, for which identification criteria are much more complicated and for which lack of identifiability is a much rarer problem.

11. The Problem of Choosing and Testing a Model

Thus far it has been presumed that the model being estimated is an accurate representation of the real-world process that actually generates the data. This is far from certain; at best it is likely to be only approximately true. Any model has been chosen by someone, perhaps with the aid of economic theory about the maximization of profit or utility, or perhaps because it was suggested by previously observed data. But there is no guarantee that it contains the right variables, or that their endogeneity or exogeneity is correctly stated, or that the correct number of lagged variables has been chosen, or that its eqns. have the right mathematical form, or that the assumed distribution of its shocks is correct.

One can perform diagnostic tests to see whether the estimated model fits past data well, and whether its calculated past shocks have constant variance and are free of any obvious systematic behavior. But this does not assure that it will do well with future data. In my view, the most stringent and most important test of a model is to expose it to data that were not available when the model was formulated. If it does not describe these data well, it leaves something to be desired. The more new data it can describe well, the more confidence one can have in it. But at any point the best that can be said about a model is that it has done a good job of describing the data that are available thus far.

See also: Instrumental Variables in Statistics and Econometrics; Linear Hypothesis; Simultaneous Equation Estimates (Exact and Approximate), Distribution of; Statistical Identification and Estimability

Bibliography

Anderson T W, Rubin H 1949 Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* **20**: 46–63

Anderson T W, Rubin H 1950 The asymptotic properties of estimates of parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* **21**: 570–82

Basmann R L 1957 A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* **25**: 77–83

Christ C F (ed.) 1994 *Simultaneous Equations Estimation*. Edward Elgar, Aldershot, UK

Davidson R, MacKinnon J G 1993 *Estimation and Inference in Econometrics*. Oxford University Press, New York

Durbin J 1954 Errors in variables. *Revue of the International Statistical Institute* **22**: 23–32

Durbin J 1988 Maximum likelihood estimation of the parameters of a system of simultaneous regression equations. *Econometric Theory* **4**: 159–70

Haavelmo T 1943 The statistical implications of a system of simultaneous equations. *Econometrica* **11**: 1–12

Haavelmo T 1944 The probability approach in econometrics. *Econometrica* **12**(suppl.): 1–115

Hansen L P 1982 Large sample properties of generalized method of moments estimators. *Econometrica* **50**: 1029–54

Hausman J A 1983 Specification and estimation of simultaneous equation models. In: Griliches Z, Intriligator M D (eds.) *Handbook of Econometrics*. North-Holland, Amsterdam, Vol. 1, pp. 391–448

Kadane J B 1971 Comparison of *k*-class estimators when the disturbances are small. *Econometrica* **39**: 723–7
Koopmans T C (ed.) 1950 *Statistical Inference in Dynamic Economic Models*. Cowles Commission Monograph 10. Wiley, New York
Lucas R E Jr 1976 Econometric policy evaluation: a critique. *Carnegie-Rochester Conference Series on Public Policy* **1**: 19–46
Nagar A L 1959 The bias and moment matrix of the general *k*-class estimators of the parameters in simultaneous equations. *Econometrica* **27**: 575–95
Strotz R H, Wold H O A 1960 A triptych on causal chain systems. *Econometrica* **28**: 417–63
Theil H 1972 *Principles of Econometrics*. Wiley, New York
Zellner A 1998 The finite sample properties of simultaneous equations' estimates and estimators Bayesian and non-Bayesian approaches. *Journal of Econometrics* **83**: 185–212
Zellner A, Theil H 1962 Three stage least squares: simultaneous estimation of simultaneous equations. *Econometrica* **30**: 54–78

C. F. Christ

Single-case Experimental Designs in Clinical Settings

1. *Differing Research Traditions*

Two major research traditions have advanced the behavioral sciences. One is based on the hypothetico-deductive approach to scientific reasoning where a hypothesis is constructed and tested to see if a phenomenon is an instance of a general principle. There is a presumed *a priori* understanding of the relationship between the variables of interest. These hypotheses are tested using empirical studies, generally using multiple subjects, and data are collected and analyzed using group experimental designs and inferential statistics. This tradition represents current mainstream research practice for many areas of behavioral science. However, there is another method of conducting research that focuses intensive study on an individual subject and makes use of inductive reasoning. In this approach, one generates hypotheses from a particular instance or the accumulation of instances in order to identify what might ultimately become a general principle.

In practice, most research involves elements of both traditions, but the extensive study of individual cases has led to some of the most important discoveries in the behavioral sciences and therefore has a special place in the history of scientific discovery. Single-case or single-subject design (also known as 'N of one') research is often employed when the researcher has limited access to a particular population and can therefore study only one or a few subjects. This method is also appropriate when one wishes to make an intensive study of a phenomenon in order to examine the conditions that maximize the strength of an effect. In clinical settings one is frequently interested in describing which variables affect a particular individual rather than in trying to infer what might be important from studying groups of subjects and assuming that the average effect for a group is the same as that observed in a particular subject. In the current era of increased accountability in clinical settings, single-case design can be used to demonstrate treatment effects for a particular clinical case (Hayes et al. 1999 for a discussion of the changing research climate in applied settings).

2. *Characteristics of Single-case Design*

Single-case designs study intensively the process of change by taking many measures on the same individual subject over a period of time. The degree of control in single-case design experiments can often lead to the identification of important principles of change or lead to a precise understanding of clinically relevant variables in a specific clinical context. One of the most commonly used approaches in single-case design research is the interrupted time-series. A time-series consists of many repeated measures of the same variable(s) on one subject, while measuring or characterizing those elements of the experimental context that are presumed to explain any observed change in behavior. An by examining characteristics of the data before and after the experimental manipulation to look for evidence that the independent variable alters the dependent variable because of the interruption.

3. *Specific Types of Single-case Designs*

Single-case designs frequently make use of a graphical representation of the data. Repeated measures on the dependent variable take place over time, so the abscissa (X-axis) on any graph represents some sort of across time scale. The dependent measure or behavior presumably being altered by the treatment is plotted on the ordinate (Y-axis). There are many variations of single-case designs that are well described (e.g., Franklin et al. 1997, Hersen and Barlow 1976, Kazdin 1982), but the following three approaches provide an overview of the general methodology.

3.1 *The A-B-A Design*

The classic design that is illustrative of this approach is called the A-B-A design (Hersen and Barlow 1976, pp. 167–97) where the letters refer to one or another

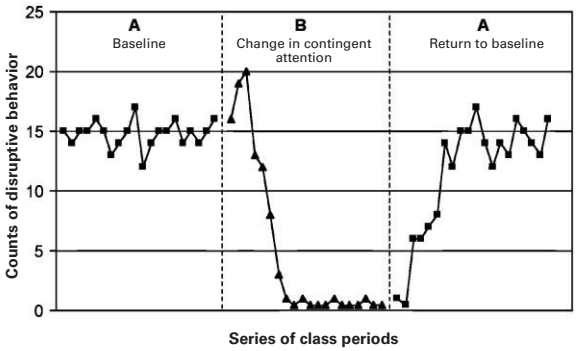

Figure 1
A-B-A design

experimental or naturally observed conditions presumed to be associated with the behavior of importance. A hypothetical example is shown in Fig. 1. By convention, A refers to a baseline or control condition and the B indicates a condition where the behavior is expected to change. The B condition can be controlled by the experimenter or alternatively can result from some naturally occurring change in the environment (though only the former are true experiments). The second A in the A-B-A design indicates a return to the original conditions and is the primary means by which one infers that the B condition was the causal variable associated with any observed change from baseline in the target or dependent variable. Without the second A phase, there are several other plausible explanations for changes observed during the B phase besides the experimental manipulation. These include common threats to internal validity such as maturation or intervening historical events coincidental to initiating B.

To illustrate the clinical use of such an approach, consider the data points in the first A to indicate repeated observations of a child under baseline conditions. The dependent variable of interest is the number of disruptive behaviors per class period that the child exhibits. A baseline of disruptive behaviors is observed and recorded over several class periods. During the baseline observations, the experimenter hypothesizes that the child's disruptive behavior is under the control of contingent attention by the teacher. That is, when the child is disruptive, the teacher pays attention (even negative attention) to the child that unintentionally reinforces the behavior, but when the child is sitting still and is task oriented, the teacher does not attend to the child. The experimenter then trains the teacher to ignore the disruptive behavior and contingently attend to the child (e.g., praise or give some token that is redeemable for later privileges) when the child is emitting behavior appropriate to the classroom task. After training the teacher to reinforce differentially appropriate classroom behavior, the teacher is instructed to implement these procedures and the B phase begins.

Observations are made and data are gathered and plotted. A decrease in the number of disruptive behaviors is apparent during the B phase. To be certain that the decrease in disruptive behaviors is not due to some extraneous factor, such as the child being disciplined at home or simply maturing, the original conditions are reinstated. That is, the teacher is instructed to discontinue contingently attending to task-appropriate behavior and return to giving corrective attention when the child is disruptive. This second A phase is a return to the original conditions. The plot indicates that the number of disruptions returns to baseline (i.e., returns to its previous high level). This return to baseline in the second A condition is evidence that the change observed from baseline condition A to the implementation of the intervention during the B phase was under the control of the teacher's change in contingent attention rather than some other factor. If another factor were responsible for the change, then the re-implementation of the original baseline conditions would not be expected to result in a return to the original high levels of disruptive behavior.

There are many variations of the A-B-A design. One could compare two or more interventions in an A-B-A-C-A design. Each A indicates some baseline condition, the B indicates one type of intervention, and the C indicates a different intervention. By examining differences in the patterns of responding for B and C, one can make treatment comparisons. The notation of BC together, as in A-B-A-C-A-BC-A, indicates that treatments B and C were combined for one phase of the study.

3.2 The Multiple-baseline Design

An A-B-A or A-B-A-C design assumes that the treatment (B or C) can be reversed during the subsequent A period. Sometimes it is impossible or unethical to re-institute the original baseline conditions (A). In these cases, other designs can be used. One such approach is a multiple-baseline design. In a multiple-baseline design, baseline data are gathered across several environments (or behaviors). Then a treatment is introduced in one environment. Data continue to be gathered in selected other environments. Subsequently the treatment is implemented in each of the other environments, one at a time, and changes in target behavior observed (Poling and Grossett 1986).

In the earlier example, if the child had been exhibiting disruptive behavior in math, spelling, and social studies classes, the same analysis of the problem might be applied. The teachers or the researcher might not be willing to reinstitute the baseline conditions if

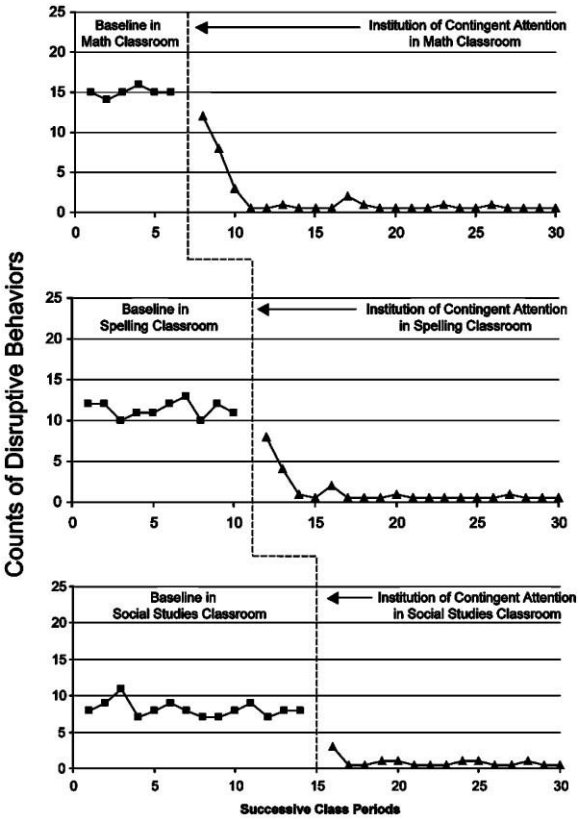

Figure 2
Example of multiple baseline design with same behaviour across different settings

improvements were noted because of its disruptive effects on the learning of others and because it is not in the best interests of the child. Figure 2 shows how a multiple-baseline design might be implemented and the data presented visually. Baseline data are shown for each of the three classroom environments.

While the initial level of disruptive behavior might be slightly different in each class, it is still high. The top graph in Fig. 2 shows the baseline number of disruptive behaviors in the math class. At the point where the vertical dotted line appears, the intervention is implemented and its effects appear to the right of dotted line. The middle and bottoms graphs show the same disruptive behaviors for the spelling and social studies class. No intervention has yet been implemented and the disruptive behavior remains high and relatively stable in each of these two classrooms. This suggests that the changes seen in the math class were not due to some other cause outside of the school environment or some general policy change within the school, since the baseline conditions and frequency of disruptive behaviors are unchanged in the spelling and social studies classes. After four more weeks of baseline observation, the same shift in attention treatment is implemented in the spelling class, but still not in the social studies class. The amount of disruptive behavior decreases following the treatment implementation in the spelling class, but not the social studies class.

This second change from baseline is a replication of the effect of treatment shown in the first classroom and is further evidence that the independent variable is the cause of the change. There is no change in the social studies class behavior. This observation provides additional evidence that the independent variable rather than some extraneous variable is responsible for the change in the behavior of interest. Finally, the treatment is implemented in the social studies class with a resulting change paralleling those occurring in the other two classes when they were changed. Rather than having to reverse the salutary effects of a successful treatment as one would in an A-B-A reversal design, a multiple baseline design allows for successively extending the effects to new contexts as a means of demonstrating causal control.

Multiple-baseline designs are used frequently in clinical settings because they do not require reversal of beneficial effects to demonstrate causality. Though Fig. 2 demonstrates control of the problem behavior by implementing changes sequentially across multiple settings, one could keep the environment constant and study the effects of a hypothesized controlling variable by targeting several individual behaviors. For example, a clinician could use this design to test whether disruptive behavior in an institutionalized patient could be reduced by sequentially reinforcing more constructive alternatives. A treatment team might target aggressive behavior, autistic speech, and odd motor behavior when the patient is in the day room. After establishing a baseline for each behavior, the team could intervene first on the aggressive behavior while keeping the preexisting contingencies in place for the other two behaviors. After the aggressive behavior has changed (or a predetermined time has elapsed), the autistic speech behavior in the same day room could be targeted, and so on for the odd motor behavior. The logic of the application of a multiple-baseline design is the same whether one studies the same behavior across different contexts or different behaviors across the same context. To demonstrate causal control, the behavior should change only when it is the target of a specific change strategy and other behaviors not targeted should remain at their previous baselines.

Interpretation can be difficult in multiple-baseline designs because it is not always easy to find behaviors that are functionally independent. Again, consider the disruptive classroom behavior described earlier. Even if one's analysis of teacher attention to disruptive behavior is reinforcing, it may be difficult to show change in only the targeted behavior in a specific classroom. The child may learn that alternative behaviors can be reinforced in the other classrooms and the child may

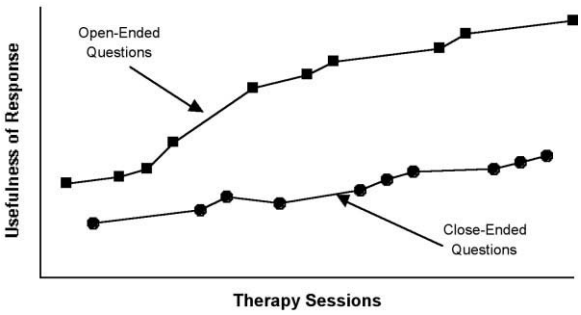

Figure 3
Example from an Alternating Treatment Design (ATD)

alter his or her behavior in ways that alter the teacher reactions, even though the teachers in the other baseline classes did not intend to change their behavior. Nevertheless, this design addresses many ethical and practical concerns about A-B-A (reversal) designs.

3.3 Alternating Treatment Design

The previous designs could be referred to as within-series designs because there are a series of observations within which the conditions are constant, an intervention with another series of observations where the new conditions are constant for the duration of series, and so on. The interpretation of results is based on what appears to be happening within a series of points compared with what is happening within another series. Another variation of the single-case design that is becoming increasingly recognized for its versatility is the alternating treatment design (ATD) (Barlow and Hayes 1979). The ATD can be thought of as a between-series design. It is a method of answering a number of important questions such as assessing the impact of two or more different treatment variations in a psychotherapy session. In an ATD, the experimenter rapidly and frequently alternates between two (or more) treatments on a random basis during a series and measures the response to the treatments on the dependent measure. This is referred to as a 'between-series' because the data are arranged by treatments first and order second rather than by order first, where treatments vary only when series change. There may be no series of one particular type of treatment, although there are many of the same treatments in a row.

To illustrate how an ATD might be applied, consider a situation where a therapist is trying to determine which of two ways of relating to a client produces more useful self-disclosure. One way of relating is to ask open-ended questions and the other is to ask direct questions. The dependent variable is a rating of clinically useful responsiveness from the client. The therapist would have a random order for asking one of the two types of questions. The name of the design does not mean that the questions alternate back and forth between two conditions on each subsequent therapist turn, but randomly change between treatments. One might observe the data in Fig. 3.

The X-axis is the familiar time variable and the Y-axis the rating of the usefulness of the client response. The treatment conditions are shown in separate lines. Note that points for each condition are plotted according to the order in which they were observed, but that the data from each condition are joined *as if* they were collected in one continuous series (even though the therapists asked the questions in the order of O-C-O-O-O-C-C-O …). The data in Fig. 3 indicate that both methods of asking questions produce a trend for the usefulness of the information to increase as the session goes on. However, the open-ended format produces an overall higher degree of utility.

4. Analyses

There are four characteristics of a particular series that can be compared with other series. The first is the level of the dependent measure in a series. One way in which multiple series may differ is with respect to the stable level of a variable. The second way in which multiple series may differ is in the trend each series shows. The third characteristic that can be observed is the shape of responses over time. This is also referred to as the *course* of the series and can include changes in cyclicity. The last characteristic that could vary is the variability of responses over time. Figure 4 shows examples of how these characteristics might show treatment effects. One might observe changes in more than one of these characteristics within and across series.

4.1 Visual Interpretation of Results

In single-case design research, there has been a long tradition to interpret the data visually once it has been graphed. In part, this preference emerged because those doing carefully controlled single-case research were looking for large effects that should be readily observable without the use of statistical approaches. Figure 4 depicts examples of clear effects. They demonstrate the optimal conditions for inferring a treatment effect, a stable baseline and clearly different responses during the post baseline phase. Suggestions for how to interpret graphical data have been described by Parsonson and Baer (1978).

However, there are many circumstances where patterns of results are not nearly so clear and errors in interpreting results are more common. When there are

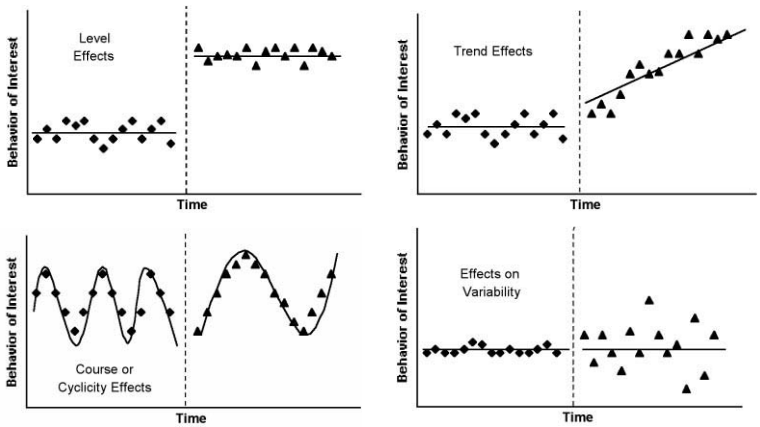

Figure 4
Examples of changes observed in single-case designs

outlier data points (points which fall unusually far away from the central tendency of other points in the phase), it becomes more difficult to determine reliably if an effect exists by visual inspection alone. If there are carry-over effects from one phase to another that the experimenter may not initially anticipate, results can be difficult to interpret. Carry-over is observed when the effects of one phase do not immediately disappear when the experiment moves into another phase. For example, the actions of a drug may continue long after a subject stops taking the drug, either because the body metabolizes the drug slowly or because some irreversible learning changes occurred as a result of the drug effects. It is also difficult to interpret graphical results when there is a naturally occurring cyclicity to the data. This might occur when circumstances not taken into account by the experimenter lead to systematic changes in behaviors that might be interpreted as treatment effects. A comprehensive discussion of issues of visual analysis of graphical displays is provided by Franklin et al. (1996).

4.2 Traditional Measurement Issues

Any particular observation in a series is influenced by three sources of variability. The experimenter intends to identify systematic variability due to the intervention. However, two other sources of variability in an observation must also be considered. These are variability attributable to measurement error and variability due to extraneous sources, such as how well the subject is feeling at the time of a particular observation. Measurement error refers to how well the score on the measurement procedure used as the dependent measure represents the construct of interest. A detailed discussion of the reliability of measurement entails some complexities that can be seen intuitively if one compares how accurately one could measure two different dependent measures. If the experimenter were measuring the occurrence of head-banging behavior in an institutionalized subject, frequency counts would be relatively easy to note accurately once there was some agreement by raters on what constituted an instance of the target behavior. There would still be some error, but it would be less than if one were trying to measure a dependent measure such as enthusiasm in a classroom. The construct of enthusiasm will be less easily characterized and have more error associated with its measurement and therefore be more likely to have higher variability.

As single-case designs are applied to new domains of dependent measures such as psychotherapy research, researchers are increasingly attentive to traditional psychometric issues, including internal and external validity as well as to reliability. Traditionally, if reliability was considered in early applications of single-case designs, the usual method was to report interrater reliability as the percent agreement between two raters rating the dependent measure (Page and Iwata 1986). Often these measures were dichotomously scored counts of behaviors (e.g., a behavior occurred or did not occur). Simple agreement rates can produce inflated estimates of reliability. More sophisticated measures of reliability are being applied to single-case design research including kappa (Cohen 1960) and variations of intraclass correlation coefficients (Shrout and Fleiss 1979) to name a few.

4.3 Statistical Interpretation of Results

While the traditional method of analyzing the results of single-case design research has been visual inspection of graphical representation of the data, there is a growing awareness of the utility of statistical verification of assertions of a treatment effect (Kruse and Gottman 1982). Research has shown that relying solely on visual inspection of results (particularly

when using response guided experimentation) may result in high Type I error rates (incorrectly concluding that a treatment effects exists when it does not).

However, exactly how to conduct statistical analyses of time-series data appropriately is controversial. If the data points in a particular series were individual observations from separate individuals, then traditional parametric statistics including the analysis of variance, t-tests, or multiple regression could all be used if the appropriate statistical assumptions were met (e.g., normally distributed scores with equal variances). All of these statistical methods are variations of the general linear model. The most significant assumption in the general linear model is that residual errors are normally distributed and independent. A residual score is the difference between a particular score and the score predicted from a regression equation for that particular value of the predictor variable. The simple correlation between a variable and its residual has an expected value of 0. In single-case designs, it can be the case that any particular score for an individual is related to (correlated with) the immediately preceding observation. If scores are correlated with their immediately preceding score (or following score), then they are said to be serially dependent or autocorrelated. Autocorrelation violates a fundamental principle of traditional parametric statistics.

An autocorrelation coefficient is a slightly modified version of the Pearson correlation coefficient (r). Normally r is calculated using pairs of observations (X_i, Y_i) for several subjects. The autocorrelation coefficient (ρ_k) is the correlation between a score or observation and the next observation (Y_i, Y_{i+1}). When this correlation is calculated between an observation and the next observation, it is called a lag 1 autocorrelation. The range of the autocorrelation coefficient is from -1.0 to $+1.0$. If the autocorrelation is 0, it indicates that there is no serial dependence in the data and the application of traditional statistical procedures can reasonably proceed. As ρ_k increases, there is increasing serial dependence. As ρ_k becomes negative, it indicates that there is a rapid cyclicity pattern. A lag can be extended out past 1 and interesting patterns of cyclicity may emerge. In economics, a large value of ρ_k at a lag 12 can indicate an annual pattern in the data. In physiological psychology, lags at 12 or 24 may indicate a systematic diurnal or circadian rhythm in the data.

Statistical problems emerge when there is an autocorrelation among the residual scores. As the value of ρ_k increases (i.e., there is a positive autocorrelation) for the residuals, the computed values of many traditional statistics are inflated, and type I errors occur at a higher than expected rate. Conversely, when residuals show a negative autocorrelation, the resulting statistics will be too small producing misleading significance testing in the opposite direction. Suggestions for how to avoid problems of autocorrelation include the use of distribution-free statistics such as randomization or permutation tests. However, if the residuals are autocorrelated even these approaches may not give fully accurate probability estimates (Gorman and Allison 1996, p. 172).

There are many proposals for addressing the autocorrelation problem. For identifying patterns of cyclicity, even where cycles are embedded within other cycles, researchers have made use of spectral analysis techniques, including Fourier analysis. These are useful where there may be daily, monthly, and seasonal patterns in data such as one might see in changes in mood over long periods of observation.

The most sophisticated of these approaches is called the autoregressive integrated moving average analysis or ARIMA method (Box and Jenkins 1976). This is a regression approach that tries to model the value of particular observation as a function of the parameters estimating the effects of past observations (ϕ), moving averages of the influence of residuals on the series (θ), and the removal of systematic trends throughout the entire time series. The computation of ARIMA models is a common feature of most statistical computer programs. Its primary limitation is that it requires many data points, often more than are practical in applied or clinical research settings. Additionally, there may be several ARIMA models that fit any particular data set, making results difficult to interpret. Alternative simplifications of regression-based analyses are being studied to help solve the interpretive problems of using visual inspection alone as a means of identifying treatment effects (Gorman and Allison 1996).

4.4 Generalization of Findings and Meta-analysis

Although single-case designs can allow for convincing demonstrations of experimental control in a particular case, it is difficult to know how to generalize from single-case designs to other contexts. Single-case design researchers are often interested in specific effects, in a specific context, for a specific subject. Group designs provide information about average effects, across a defined set of conditions, for an average subject. Inferential problems exist for both approaches. For the single-case design approach, the question is whether results apply to any other set of circumstances. For the group design approach, the question is whether the average result observed for a group applies to a particular individual in the future or even if anyone in the group itself showed an average effect.

In the last two decades of the twentieth century, meta-analysis has become used as a means of aggregating results across large numbers of studies to summarize the current state of knowledge about a particular topic. Most of the statistical work has focused on aggregating across group design studies.

More recently, meta-analysts have started to address how to aggregate the results from single-case design studies in order to try to increase the generalizability from this type of research. There are still significant problems in finding adequate ways to reflect changes in slopes (trends) between phases as an effect size statistic, as well as ways to correct for the effects of autocorrelation in meta-analysis just as there are for the primary studies upon which a meta-analysis is based. There is also discussion of how and if single-case and group design results could be combined (Panel on Statistical Issues and Opportunities for Research in the Combination of Information, 1992).

5. *Summary*

Technical issues aside, the difference between the goal of precision and control typified in single-case design versus the interest in estimates of overall effect sizes that are characteristic of those who favor group designs makes for a lively discussion about the importance of aggregation across studies for the purpose generalization. With analytic methods for single-case design research becoming more sophisticated, use of this research strategy to advance science is likely to expand as researchers search for large effects under well specified conditions.

See also: Behavior Analysis, Applied; Case Study: Logic; Case Study: Methods and Analysis; Case-oriented Research; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Experimenter and Subject Artifacts: Methodology; Hypothesis Testing: Methodology and Limitations; Laboratory Experiment: Methodology; Medical Experiments: Ethical Aspects; Psychotherapy: Case Study; Quasi-Experimental Designs; Reinforcement, Principle of; Single-subject Designs: Methodology

Bibliography

Barlow D H, Hayes S C 1979 Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis* **12**: 199–210

Box G E P, Jenkins G M 1976 *Time Series Analysis, Forecasting, and Control*. Holden-Day, San Francisco

Cohen J A 1960 A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**: 37–46

Franklin R D, Allison D B, Gorman B S 1997 *Design and Analysis of Single-case Research*. Lawrence Erlbaum Associates, Marwah NJ

Franklin R D, Gorman B S, Beasley T M, Allison D B 1996 Graphical display and visual analysis. In: Franklin R D, Allison D B, Gorman B S (eds.) *Design and Analysis of Single-case Research*. Lawrence Erlbaum Associates, Marwah NJ, pp. 119–58

Gorman B S, Allison D B 1996 Statistical alternatives for single-case designs. In: Franklin R D, Allison D B, Gorman B S (eds.) *Design and Analysis of Single-case Research*. Lawrence Erlbaum Associates, Marwah NJ, pp. 159–214

Hayes S C, Barlow D H, Nelson-Gray R O 1999 *The Scientist-practitioner: Research and Accountability in the Age of Managed Care*, 2nd edn. Allyn and Bacon, Boston

Hersen M, Barlow D H 1976 *Single Case Experimental Designs: Strategies for Studying Behavior Change*. Pergamon Press, New York

Kazdin A E 1982 *Single Case Research Designs: Methods for Clinical and Applied Settings*. Oxford University Press, New York

Kruse J A, Gottman J M 1982 Time series methodology in the study of sexual hormonal and behavioral cycles. *Archives of Sexual Behavior* **11**: 405–15

Page T J, Iwata B A 1986 Interobserver agreement: History, theory, and current methods. In: Poling A, Fuqua R W (eds.) *Research Methods in Applied Behavior Analysis: Issues and Advances*. Plenum, New York, pp. 99–126

Panel on Statistical Issues and Opportunities for Research in the Combination of Information 1992 *Contemporary Statistics: Statistical Issues and Opportunities for Research*. National Academy Press, Washington, DC, Vol. 1

Parsonson B S, Baer D M 1978 The analysis and presentation of graphic data. In: Kratochwill T R (ed.) *Single-subject Research: Strategies for Evaluating Change*. Academic Press, New York, pp. 101–65

Poling A, Grossett D 1986 Basic research designs in applied behavior analysis. In: Poling A, Fuqua R W (eds.) *Research Methods in Applied Behavior Analysis: Issues and Advances*. Plenum, New York

Shrout P E, Fleiss J L 1979 Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86**: 420–8

W. C. Follette

Single-subject Designs: Methodology

Science is about knowledge and understanding and is fundamentally a human enterprise. Adequate methodology is at the heart of scientific activity for it allows scientists to answer fundamental epistemic questions about what is known or knowable about their subject matter. This includes the validity of data-based conclusions, the generality of findings, and how knowledge may be used to achieve practical goals. Choice of methodology is critical in answering such questions with precision and scope, and the adequacy of a given methodology is conditional upon the purposes for which it is put. The purpose of this chapter is to provide a succinct description of a methodological approach well suited for work with single organisms, whether in basic research or applied contexts, and particularly where large N designs or inferential statistics are nonsensical and even inappropriate. This approach, referred to here as 'single-subject design' methodology, has a long history in the medical,

behavioral, and applied sciences and includes several design options, from the simple to the highly sophisticated and complex. As will be seen, all such designs involve a rigorous and intensive experimental analysis of a single case, and often replications across several individual cases. No attempt will be made here to describe all available design options, including available statistical tests, as excellent resources are available on both subjects (see Barlow and Hersen 1984, Busk and Marascuilo 1992, Hayes et al. 1999; Huitema 1986, Iversen 1991, Kazdin 1982, Parsonson and Baer 1992). Rather, the intent is to provide (a) an overview of single-subject design methodology and the rationale guiding its use in experimental and applied contexts; (b) a succinct description of the main varieties of single-subject designs, including basic guidelines for their application; and (c) to highlight recent trends in the use of such designs in applied contexts as a means to increase accountability.

1. *Single-subject Design: Overview and Rationale*

At the core, all scientists ultimately deal with particulars, whether cells, atoms, microbes, or the behavior of organisms. It is from such particulars that the facts of science and generality of findings are derived (Sidman 1960). This is true whether one aggregates a limited set of observations from a large sample of cases into groups that differ on one or more variables, or situations involving a few cases and a large number of observations of each over time (see also Hilliard 1993). The former is characteristic of hypothetico-deductive group design strategies that rely on inferential statistics to support scientific claims, whereas the latter idiographic and inductive approach is characteristic of single-subject methodology. As with other research designs, single-subject methodology is as much an approach to science as it is a set of methodological rules and procedures for answering scientific and practical questions. As an approach, single-subject designs have several distinguishing features.

1.1 *Subject Matter of Single-subject Designs*

Perhaps the most notable feature of single-subject methodology is its subject matter and sample characteristics; namely, the intensive study of the behavior of single (i.e., $N = 1$) organisms. This feature is also a source of great misunderstanding (Furedy 1999), particularly with regard to the generality and scientific utility of research based on an N of 1. It should be stressed that single-subject methodology gets its name not from sample size *per se*, but rather because the unit of analysis is behavior of individuals (Perone 1991), or what Gottman (1973) referred to as 'N-of-one-at-a-time' designs. Skinner (1966) described it this way: '… instead of studying 1000 rats for one hour each, or 100 rats for 10 hours each, the investigator is likely to study one rat for a 1000 hours' (p. 21); an approach that served as the foundation for operant learning (Skinner 1953, 1957) and one that, with the increasing popularity of behavior therapy, set the stage for a new approach to treatment development, treatment evaluation, and increased accountability (Barlow and Hersen 1984). Though single-subject research often involves more than one subject, the data are evaluated primarily on a case-by-case basis. The idiographic nature of this methodology is well suited for applied work where practitioners often attempt to influence and change problematic behavior(s) of individual clients (Hayes et al. 1999). Though not a requirement, the inclusion of more than one subject helps establish the robustness of observed effects across differentially imposed conditions, individuals, behaviors, settings, and/or time. Hence, direct or systematic replication serves to increase confidence in the observed effects and helps establish the reliability and generality of observed relations (Sidman 1960).

1.2 *Analytic Aims of Single-subject Designs: Prediction and Influence (Control)*

As with other experimental design strategies, the fundamental premise driving the use of single-subject methodology is to demonstrate control (i.e., influence) via the manipulation of independent variables and analysis of the effects of such variables on behavior (dependent variables). This emphasis on systematic manipulation of independent variables distinguishes single-subject designs from other small N research such as case reports, case studies, and the like which are often lucidly descriptive, but less controlled and systematic. With few exceptions (see Perone 1991), the general convention of single-subject design is to change one variable at a time and to evaluate such effects on behavior. This 'one-variable-at-a-time' rule positions the basic or applied researcher to make clear causal statements about the effects of independent variables compared with a previous state (e.g., naturally occurring behavior during baseline), to compare different independent variables across or within conditions, and only then to evaluate interaction effects produced by two or more independent variables together.

Indeed, single-subject methodology is unique in that it typically involves an intensive and rigorous experimental analysis of the behavior (i.e., thinking, emotional responses, overt motor acts) across time, and an equally rigorous attempt to isolate and control extraneous sources of variability (e.g., maturation, history) that may confound and/or mask any effects caused by systematic changes in the independent variable(s) across conditions and time (Sidman 1960). This point regarding the handling of extraneous sources of variability cannot be overstated. Iversen

(1991) and others (Barlow and Hersen 1984: Sidman 1960) have noted issues concerning how variability is handled in-group design research and single-subject methodology. With regard to such issues, Iverson put it this way:

> [in-group design research] the fit between theory and observation often is evaluated by means of statistical tests, and variability in data is considered a nuisance rather than an inspiration .... By averaging over *N* subjects the variability among the subjects is ignored *N* times. In other words, a data analysis restricted to only the mean generates *N* degrees of ignorance ... Single-subject designs are used because the data based on the individual subject are more accurate predictors of that individual's behavior than are data based on an averaging for a group of subjects (p. 194).

Though the issue of when and how to aggregate data (if at all) is controversial in single-subject research (Hilliard 1993), there is a consensus that variability is to be pinpointed and controlled to the extent possible. This is consistent with the view that variability is imposed and determined from without, and not an intrinsic property of the subject matter under study (Sidman 1960). Efforts are deliberately made to control such noise so that prediction and control can be achieved with precision and scope. Consequently, those employing experimental single-subject design methods are more likely to commit Type II errors (i.e., to deny a difference and claim that a variable is not exerting control over behavior, when it is) compared with more classic Type I errors (i.e., to claim that a variable is exerting control over behavior, when it is not; (see Baer 1977).

1.3 Single-subject Design: Measurement Issues

Single-subject design methodology also can be distinguished by the frequency with which behavior is sampled within conditions, and across differentially imposed conditions and time. It is common, for instance, for basic and applied researchers to include hundreds of samples of behavior across time (i.e., hours, days, weeks, and sessions) with individual subjects. Yet, a minimum of at least three data points is required to establish level, trend, and variability within a given phase or design element (Barlow and Hersen 1984, Hayes et al. 1999). Changes in level, trend, and at times variability from one condition to the next, and particularly changes that are robust, clear, immediate, and stable from behavior observed in a prior condition, help support causal inferences.

Here, stability is a relative term and should not be confused with something that is static or flat. Rather, stability provides a background by which to evaluate the reliability of changes produced by an independent variable across conditions and time. By convention, each condition is often continued until some semblance of stability is observed, at which point another condition is imposed, and behavior is evaluated relative to the previous condition or steady state (cf. Perone 1991). Such changes, in turn, are often evaluated visually and graphically relative to a prior less controlled state (i.e., baseline), or adjacent conditions involving another independent variable. In single-subject methodology, such stability can be evaluated within a condition (i.e., from one sample or observation to the next) and across conditions (i.e., changes in level, trend, and variability). The pragmatic appeal of such an approach, particularly in applied contexts, rests in allowing the practitioner to evaluate the effectiveness of imposed interventions in real time, and consequently to modify their intervention tactics as the data dictate. Indeed, unlike more formal group designs that are followed strictly once set in place, single-subject methodology is more flexible. Indeed, it is common for design elements to be added, dropped, and/or modified as the data dictate so as to meet scientific and applied goals. This feature has obvious parallels with how practitioners work with their clients in designing treatment interventions (see also Hayes et al. 1999).

In sum, all single-subject designs focus on individual organisms and aim to predict and influence behavior via: (a) repeated sampling and observation; (b) manipulation of one or more independent variables while isolating and controlling sources of extraneous variability to the extent possible; and (c) demonstration of stability within and across levels of imposed independent variables (see Perone 1991). Discussion will now turn to an enumeration of such features in the context of more popular single-subject designs.

2. Varieties of Single-subject Methodology

Most single-subject designs can be generally classified as representing two main types: within-series and between-series designs. Though each will be discussed briefly in turn, it is important to recognize that neither class of design precludes combining elements from the other (e.g., combined-series elements). In other words, a within-series design may, for some purposes, lead the investigator to add between-series elements and vice-versa, including other more complex elements (e.g., interaction effects, changing criterion elements).

2.1 Within-series Designs

The basic logic and structure of within-series designs are simple; namely, to evaluate changes within a series of data points across time on a single measure or set of related measures (see Hayes et al. 1999). Stability is judged for each data point relative to other data points that immediately precede and follow it. By convention, such designs include comparisons between a naturally occurring state of behavior (denoted by the letter A) and the effects of an imposed manipulation of an independent variable or intervention (denoted by different letters such as B, C, and so on). A simple A-B design involves sampling naturally occurring beha-

vior (A phase), followed by repeated assessment of responding when the independent variable is introduced (B phase). For example, suppose a researcher wanted to determine the effects of rational-emotive therapy on initiating social interactions for a particular client. If an A-B design were chosen, the rate of initiations prior to treatment would be compared with that following treatment in a manner analogous to popular pre-to-post comparisons in-group outcome research. It should be noted, however, that A-B designs are inherently weak in controlling for threats to internal validity (e.g., testing, maturation, history, regression to the mean). Withdrawal/reversal designs control for such threats, and hence provide a more convincing case of experimental control.

Withdrawal designs represent a replication of the basic A-B sequence a second time, with the term withdrawal representing the imposed removal of the active treatment or independent variable (i.e., a return to second baseline or A phase). For example, a simple A-B-A sequence allows one to evaluate treatment effects relative to baseline responding. If an effect due to treatment is present, then it should diminish once treatment is withdrawn and the subject or client is returned to an A phase. Other variations on withdrawal designs include B-A-B designs and A-B-A-B (reversal) designs (see Barlow and Hersen 1984, Hayes et al. 1999). A-B designs and withdrawal/reversal designs are typically used to compare the effects of a finite set of treatment variables with baseline response levels. Data regarding stable response trends are collected across several discrete periods (e.g., time, sessions), wherein the independent variable is either absent (baseline) or present (treatment phase). Phase shifts are data-driven, and response stability determines the next element added to the design, elements that may include other manipulations or treatments either alone (e.g., A-B-A-C-A-B-A) or in combination (e.g., A-B-A-B + C-A-B-A). As the manipulated behavior change is repeatedly demonstrated and replicated across increasing numbers of phase shifts and time, confidence in the role of the independent variable as a cause of such changes increases.

This basic logic of within-series single-subject methodology has been expanded in sophisticated and at times complex ways to meet basic and applied purposes. For instance, such designs can be used to test the differential effects of more than one treatment. Other reversal designs (i.e., B-C-B-C) involve the comparison of differential, yet consecutive, treatment interventions across multiple phase shifts. These designs are similar to the above reversals in how control is evaluated, but differ primarily in that baseline phases are not required. Designs are also available that combine features of withdrawal and reversal. For example, A-B-A-B-C-B designs allow one to compare A and B phases with each other (reversal; A-B-A-B), and a second treatment to B (reversal; e.g., B-C-B), or even to evaluate the extent to which behavior tracks a specified behavioral criterion (i.e., changing criterion designs; see Hayes et al. 1999). Changing criterion designs provide an alternative method for experimentally analyzing behavioral effects without subsequent treatment withdrawal. Criterion is set such that optimal levels given exposure can be met and then systematically increased (or decreased) to instill greater demand on acquiring new repertoires. For example, a child learning to add may be required to calculate 4 of 10 problems correctly to earn a prize. Once the child successfully and consistently demonstrates this level of responding, the demand increases to 6 of 10 correct problems. Changing criterion designs serve as a medium to demonstrate learning through achieving successive approximations of the end-state.

2.2 *Between-series Designs*

Between-series designs differ primarily from within-series designs in that data are first grouped and compared by condition, and then by time, whereas the reverse is true for within-series designs. Between-series designs need not contain phases, as evaluation of level, trend, and stability are organized by condition first, but not by time alone (Hayes et al. 1999, p.176). Designs within this category include 'alternating-treatment design' and 'simultaneous-treatment design.' The basic logic of both is the same, in that a minimum of two treatments are evaluated concurrently in an individual case. With alternating-treatments design, the concurrent evaluation is made between rapid and largely random alternations of two or more conditions (Barlow and Hayes 1979). Unlike within-series designs, alternating-treatment designs contain no phases (Hayes et al. 1999). The same is true for simultaneous-treatment designs; a design that is appropriate for situations where one wishes to evaluate the concurrent or simultaneous application of two or more treatments in a single case. Rapid or random alteration of treatment is not required with simultaneous-treatment design. What is necessary is that two or more conditions are simultaneously available, with the subject choosing between them. This particular design is not conducive for evaluating treatment outcome, but is appropriate for evaluating preference or choice (see Hayes et al. 1999). Alternating-treatment designs, by contrast, fit well with applied work, as therapists must often routinely target multiple problems concurrently, and thus need to switch rapidly between intervention tactics in the context of therapy.

2.3 *Combining Within- and Between-series Elements: Multiple-baseline Design*

Multiple-baseline designs build upon and integrate the basic logic and structure of within- and

between-series elements. The fundamental premise of multiple-baseline designs is to replicate phase change effects systematically in more than one series, with each subsequent uninterrupted series serving as a control condition for the preceding interrupted series. Series can be compared and arranged across behaviors, across settings, across individuals, or some combination of these (see Hayes et al. 1999). Such designs require that the series under consideration be independent (e.g., two functionally distinct behaviors), and that intervention be administered sequentially beginning with the first series, while others are left uninterrupted as controls. For example, a client might present with three distinct behavior problems all requiring exposure therapy. All behaviors would be monitored during baseline (A phase), and after some semblance of stability is reached the treatment (B phase) would be applied to the first behavior series, while the remaining two behaviors are continuously monitored in an extended baseline. Once changes in the first behavior resulting from treatment reach stability, the second series would be interrupted and treatment applied, while the first behavior continues in the B phase and the third behavior is monitored in an extended baseline. The procedure followed for the first two series elements is then repeated for the third behavior. This logic can be similarly applied to multiple-baseline designs across individuals or settings (see Barlow and Hersen 1984, Hall et al.). More complex within-series elements (e.g., A-B-A-B, or B-C-B, or counterbalanced phases such as B-A-B and A-B-A) can be evaluated across behaviors, settings, and individuals.

Note that with multiple-baseline designs the treatment is never withdrawn, rather, it is introduced systematically across a set of dependent variables. Independent variable effects are denoted from how reliably behavior change correlates with the onset of treatment for a particular dependent variable. For such reasons, multiple-baseline designs are quite popular, owing much to their ease of use, strength in ruling out threats to internal validity, built-in replication, and fit with the demands of applied practitioners working in therapy, schools, and institutions. Indeed, such designs can be useful when therapists are working with several clients presenting with similar problems, when clients begin therapy at different times, or in cases where targets for change in therapy occur sequentially.

3. *Recent Trends in the Application of Single-subject Methods*

Single-subject methodology has a long historic affiliation with basic and applied behavioral research, and the behavior therapy movement more generally (Barlow and Hersen 1984). Though the popularity of single-subject methods in published research appearing in mainstream behavior therapy journals appears to be on the decline relative to use of group design methodology (Forsyth et al. 1999), there does appear to be a resurgence of interest in single-subject methodology in applied work. There are several possible reasons for this renewed interest, but only two are mentioned here. First, trends in treatment-outcome research have increasingly relied on the now popular 'randomized clinical trial' group design methodology to establish empirical support for psychosocial therapies for specific behavioral disorders. Practitioners, who work predominantly with individual clients, are quick to point out that group-outcome data favoring a given intervention, however convincing statistically speaking, includes individuals in the group that did not respond to therapy. Moreover, the 'average' group response to treatment may not generalize to the specific response of an individual client to that treatment. Thus, practitioners have been skeptical of how such research informs their work with individual clients.

The second, and perhaps more important, reason for the renewed interest in single-subject methodology is driven by pragmatic concerns and the changing nature of the behavioral health care marketplace. Increasingly, third-party payers are requiring practitioners to demonstrate accountability. That is, to show that what they are doing with their clients is achieving the desired effect (i.e., good outcomes) in a cost-effective and lasting manner. Use of single-subject methodology has a place in assisting practitioners in making clinical decisions based on empirical data, in demonstrating accountability to third party payers and consumers. Further, such methodology, though requiring time and sufficient training to implement properly, is atheoretical and fits nicely with how practitioners from a variety of conceptual schools work routinely with their individual clients (Hayes et al. 1999). Thus, there is great potential in single-subject methodology for bridging the strained scientist–practitioner gap, and ultimately advancing the science of behavior change and an empirically-driven approach to practice and treatment innovation.

4. *Summary and Conclusions*

Methodology is a way of knowing. Single-subject methodology represents a unique way of knowing that has as its subject matter an experimental analysis of the behavior of individual organisms. Described here were the assumptions driving this approach and the main varieties of design options available to address basic experimental and applied questions. Perhaps the greatest asset of single-subject methodology rests with the flexibility with which such designs can be construc-

ted to meet the joint analytic goals of prediction and control over the behavior of individual organisms. This asset also represents one of the greatest liabilities of such designs in that flexibility requires knowledge of when and how to modify and/or add design elements to achieve analytic goals, and particularly skill in recognizing significant effects, sources or variability, and when one has demonstrated sufficient levels of prediction and control over the behavior of interest.

See also: Case Study: Logic; Case Study: Methods and Analysis; Psychotherapy: Case Study; Single-case Experimental Designs in Clinical Settings

Bibliography

Baer D M 1977 Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis* **10**: 167–72

Barlow D H, Hayes S C 1979 Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis* **12**: 199–210

Barlow D H, Hersen M 1984 *Single-Case Experimental Designs: Strategies for Studying Behavior Change*, 2nd edn. Allyn & Bacon, Boston, MA

Busk P L, Marascuilo L A 1992 Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In: Kratochwill T R, Levin J R (eds.) *Single-Case Research Design and Analysis*. Plenum, New York, pp. 159–85

Forsyth J P, Kollins S, Palav A, Duff K, Maher S 1999 Has behavior therapy drifted from its experimental roots? A survey of publication trends in mainstream behavioral journals. *Journal of Behavior Therapy and Experimental Psychiatry* **30**: 205–20

Furedy J J 1999 Commentary: On the limited role of the single-subject design in psychology: Hypothesis generating but not testing. *Journal of Behavior Therapy and Experimental Psychiatry* **30**: 21–2

Gottman J M 1973 N-of-one and N-of-two research in psychotherapy. *Psychological Bulletin* **80**: 93–105

Hall R V, Cristler C, Cranston S S, Tucker B 1970 Teachers and parents as researchers using multiple baseline designs. *Journal of Applied Behavior Analysis* **3**: 247–55

Hayes S C, Barlow D H, Nelson-Gray R O 1999 The *Scientist-Practitioner: Research and Accountability in the Age of Managed Care*. Allyn & Bacon, Boston, MA

Hilliard R B 1993 Single-case methodology in psychotherapy process and outcome research. *Journal of Consulting and Clinical Psychology* **61**: 373–80

Huitema B E 1986 Statistical analysis and single-subject designs: Some misunderstandings. In: Poling A, Fuqua R W (eds.) *Research Methods in Applied Behavior Analysis*. Plenum, New York, pp. 209–32

Iversen I H 1991 Methods of analyzing behavior patterns. In: Iversen I H, Lattal K A (eds.) *Experimental Analysis of Behavior*, pt 2. Elsevier, New York, pp. 193–241

Kazdin A E 1982 *Single-Case Research Designs: Methods for Clinical and Applied Settings*. Oxford University Press, New York

Parsonson B S, Baer D M 1992 The visual analysis of data, and current research into the stimuli controlling it. In: Kratochwill T R, Levin J R (eds.) *Single-Case Research Design and Analysis*. Plenum, New York, pp. 15–41

Perone M 1991 Experimental design in the analysis of free-operant behavior. In: Iversen I H, Lattal K A (eds.) *Experimental Analysis of Behavior*, pt 1. Elsevier, New York, pp. 135–71

Sidman M 1960 *Tactics of Scientific Research: Evaluating Experimental Data in Psychology*. Basic Books, New York

Skinner B F 1953 *Science and Human Behavior*. Macmillan, New York

Skinner B F 1957 The experimental analysis of behavior. *American Scientist* **45**: 343–71

Skinner B F 1966 Operant behavior. In: Honig W K (ed.) *Operant Behavior: Areas of Research and Application*. Appleton-Century-Crofts, New York, pp. 12–32

J. P. Forsyth and C. G. Finlay

Situated Cognition: Contemporary Developments

Situate analyses of cognition that draw on the substantive aspects of Vygotsky's and Leont'ev's work (see *Situated Cognition: Origins*) did not become prominent in Western Europe and North America until the 1980s. Contemporary situated perspectives on cognition can be divided into two broad groups. One of these, cultural historical activity theory (CHAT), has developed largely independently of mainstream Western psychology by drawing inspiration directly from the writings of Vygotsky and Leont'ev. The other group, which I will call *distributed cognition*, has developed in reaction to mainstream cognitive science and incorporates aspects of the Soviet work.

1. Cultural Historical Activity Theory

It is important to clarify that the term 'activity' in the acronym in CHAT refers to cultural activity or, in other words, to cultural practice. The reference to history indicates a focus on both the evolution of the cultural practices in which people participate and the development of their thinking as they participate in them. Olson's (1995) analysis of the historical development of writing systems is paradigmatic of investigations that focus on the evolution of cultural practices at what might be termed the macrolevel of history writ large. Olsen's goal was not merely to document the first appearance of and subsequent changes in writing systems. Instead, he sought to demonstrate that changes in writing systems precipitated changes in thought that in turn made possible further developments in both writing and thinking. In doing so, he elaborated Vygotsky's claim that it was not until reasonably sophisticated writing systems had

emerged that people became consciously aware of language as a means of monitoring and regulating thought. As Olsen has made clear, the findings of his analysis have significant implications for children's induction into literacy practices in school.

Saxe's (1991) investigation of the body parts counting system of the Oksapmin people of Papua New Guinea provides a useful point of contrast in that it was concerned with cultural change at a more local level. As Saxe reports, the Oksapmin's counting system has no base structure and no distinct terms for numbers. Instead, the Oksapmin count collections of items (e.g., sweet potatoes) by beginning with the left index finger and naming various body parts as they move up the left arm, across the head and shoulders, and down the right arm, ending with the right index finger.

At the time that Saxe conducted his fieldwork, a new technology was being introduced, a base-10 currency system. Saxe documents that the Oksapmin with the greatest experience in using the currency, the owners of indigenous trade stores, had developed relatively sophisticated reasoning strategies that were based on the body parts system but that privileged 10 (e.g., transforming the task of adding nine and seven into that of adding ten and six by 'moving' a body part). Saxe's finding is significant in that it illustrates a general claim made by adherents of CHAT, namely that changes in the artifacts people use, and thus the cultural practices in which they participate, serve to reorganize their reasoning.

The two CHAT studies discussed thus far, those of Olsen and Saxe, investigated the evolution of cultural practices. A second body of CHAT research is exemplified by investigations that have compared mathematical reasoning in school with that in various out-of-school settings. Following both Vygotsky and Leont'ev, these investigations can be interpreted as documenting the fusion of the forms of reasoning that people develop with the cultural practices in which they participate. A third line of CHAT research has focused on the changes that occur in people's reasoning as they move from relatively peripheral participation to increasingly substantial participation in the practices of particular communities. In their overview of this type of research, Lave and Wenger (1991) clarify that the cultural tools used by community members are viewed as carrying a substantial portion of a practice's intellectual heritage. As Lave and Wenger note, this implies that novices' opportunities for learning depend crucially on their access to these tools as they are used by the community's old-timers. Lave and Wenger also make it clear that in equating learning with increasingly substantial participation, they propose to dispense with traditional cognitive analyses entirely. In their view, someone's failure to learn should be explained in terms of the organization of the community and the person's opportunities for access to increasingly substantial forms of participation rather than in terms of cognitive deficits that are attributed to the person. This is clearly a relatively strong claim in that it equates an analysis of the conditions for the possibility of learning with an analysis of learning.

It should be apparent from this overview that CHAT research spans a wide range of problems and issues. One theme that cuts across the three lines of work is a focus on groups of people's reasoning, whether they be people using different writing systems in different historical epochs, Oksapmin trade store owners compared to Oksapmin who have less experience with the new currency, the mathematical reasoning of people in school and nonschool settings, or novices versus old-timers in a community of practice. Differences in the reasoning of people as they participate in the same practices are therefore accounted for in terms of either (a) experience in participating in the practice, (b) access to more substantial forms of participation, or (c) differences in the history of their participation in other practices. Each of these types of explanations instantiates Leont'ev's dictum that the individual-in-cultural-practice rather than the individual *per se* is the appropriate unit of analysis.

2. *Distributed Cognition*

Whereas CHAT research often involves comparisons of groups of people's reasoning, work conducted within the distributed cognition tradition typically focuses on the reasoning of individuals or small groups of people as they solve specific problems or complete specific tasks. Empirical studies conducted within this tradition therefore tend to involve detailed microanalysis of either an individual's or a small group's activity. Further, whereas CHAT researchers often frame people's reasoning as acts of participation in relatively broad systems of cultural practices, distributed cognition theorists typically restrict their focus to the immediate physical, social, and symbolic environment. This circumscription of analyses to people's reasoning in their immediate environments is one indication that this tradition has evolved from mainstream cognitive science. Several of the leading scholars in this tradition, such as John Seeley Brown, Alan Collins, and James Greeno, initially achieved prominence within the cognitive science community before substantially modifying their theoretical commitments, in the process contributing to the emergence of a distributed perspective on cognition.

The term 'distributed intelligence' or 'distributed cognition' is most closely associated with Roy Pea (1993). Pea coined this term to emphasize that, in his view, cognition is distributed across minds, persons, and symbolic and physical environments. As he and other distributed cognition theorists make clear in their writings, this perspective directly challenges a

foundational assumption of mainstream cognitive science. This is the assumption that cognition is bounded by the skin or by the skull and can be adequately accounted for solely in terms of people's construction of internal mental models of an external world. Distributed cognition theorists instead see cognition as extending out into the immediate environment such that the environment becomes a resource for reasoning.

In coming to this conclusion, distributed cognition theorists have been influenced by a number of studies conducted by CHAT researchers, one of the most frequently cited investigations being that of Scribner (1984). In this investigation, Scribner analyzed the reasoning of workers in a dairy as they filled orders by packing products into crates of different sizes. Her analysis revealed that the loaders did not perform purely mental calculations but instead used the structure of the crates as a resource in their reasoning. For example, if an order called for 10 units of a particular product and six units were already in a crate that held 12 units, experienced loaders rarely subtracted six from 10 to find how many additional units they needed. Instead, they might realize that an order of 10 units would leave two slots in the crate empty and just know immediately from looking at the partially filled crate that four additional units are needed. As part of her analysis, Scribner convincingly demonstrated that the loaders developed strategies of this type *in situ* as they went about their daily business of filling orders. For distributed cognition theorists, this indicated that the system that did the thinking was the loader in interaction with a crate. From the distributed perspective, the loaders' ways of knowing are therefore treated as emergent relations between them and the immediate environment in which they worked.

Part of the reason that distributed cognition theorists attribute such significance to Scribner's study and to other investigations conducted by CHAT researchers is that they capture what Hutchins (1995) refers to as 'cognition in the wild.' This focus on people's reasoning as they engage in both everyday and workplace activities contrasts sharply with the traditional school-like tasks that are often used in mainstream cognitive science investigations. In addition to questioning whether people's reasoning on school-like tasks constitutes a viable set of cases from which to develop general models of cognition, several distributed cognition theorists have also critiqued current school instruction. In doing so, they have broadened their focus beyond cognitive science's traditional emphasis on the structure of particular tasks by drawing attention to the nature of the classroom activities within which the tasks take on meaning and significance for students.

Brown et al. (1989) developed one such critique by observing that school instruction typically aims to teach students abstract concepts and general skills on the assumption that students will be able to apply them directly in a wide range of settings. In challenging this assumption, Brown et al. argue that the appropriate use of a concept or skill requires engagement in activities similar to those in which the concept or skill was developed and is actually used. In their view, the well-documented finding that most students do not develop widely applicable concepts and skills in school is attributable to the radical differences between classroom activities and those of both the disciplines and of everyday, out-of-school settings. They contend that successful students learn to meet the teacher's expectations by relying on specific features of classroom activities that are alien to activities in the other settings. In developing this explanation, Brown et al. treat the concepts and skills that students actually develop in school as relations between students and the material, social, and symbolic resources of the classroom environment.

It might be concluded from the two examples given thus far, those of the dairy workers and of students relying on what might be termed superficial cues in the classroom, that distributed cognition theorists do not address more sophisticated types of reasoning. Researchers working in this tradition have in fact analyzed a number of highly technical, work-related activities. The most noteworthy of these studies is, perhaps, Hutchins's (1995) analysis of the navigation team of a naval vessel as they brought their ship into San Diego harbor. In line with other investigations of this type, Hutchins argues that the entire navigation team and the artifacts it used constitutes the appropriate unit for a cognitive analysis. From the distributed perspective, it is this system of people and artifacts that did the navigating and over which cognition was distributed. In developing his analysis, Hutchins pays particular attention to the role of the artifacts as elements of this cognitive system. He argues, for example, that the cartographer has done much of the reasoning for the navigator who uses a map. This observation is characteristic of distributed analyses and implies that to understand a cognitive process, it is essential to understand how parts of that process have, in effect, been sedimented in tools and artifacts. Distributed cognition theorists therefore contend that the environments of human thinking are thoroughly artificial. In their view, it is by creating environments populated with cognitive resources that humans create the cognitive powers that they exercise in those environments. As a consequence, the claim that artifacts do not merely serve to amplify cognitive process but instead reorganize them is a core tenet of the distributed cognition perspective.

3. *Current Issues and Future Directions*

It is apparent that CHAT and distributed cognition share a number of common assumptions. For example, both situate people's reasoning within encom-

passing activities and both emphasize the crucial role of tools and artifacts in cognitive development. However, the illustrations presented in the preceding paragraphs also indicate that there are a number of subtle differences between the two traditions. One concerns the purview of the researcher in that distributed cognition theorists tend to focus on social, material, and symbolic resources within the immediate local environment whereas CHAT theorists frequently locate an individual's activity within a more encompassing system of cultural practices. A second difference concerns the way in which researchers working in the two traditions address the historical dimension of cognition. Distributed cognition theorists tend to focus on tools and artifacts *per se*, which are viewed as carrying the reasoning of their developers from prior generations. In contrast, CHAT theorists treat artifacts as one aspect of a cultural practice, albeit an important one. Consequently, they situate cognition historically by analyzing how systems of cultural practices have evolved. Thus, whereas distributed cognition theorists focus on the history of the cognitive resources available in the immediate environment, CHAT theorists contend that this environment is defined by a historically contingent system of cultural practices. Despite these differences, it is possible to identify several current issues that cut across the two traditions. The two issues I will focus on are those of transfer and of participation in multiple communities of practice.

3.1 Transfer

As noted in the article *Situated Cognition: Origins*, the notion that knowledge is acquired in one setting and then transferred to other settings is central to the cognition plus view as well as to mainstream cognitive science. To avoid confusion, it is useful to differentiate between this notion of transfer as a theoretical idea and what might be termed the phenomenon called transfer. This term refers to specific instances of behavior that a cognition plus theorist would account for in terms of the transfer of knowledge from one situation to another. As will become clear, CHAT and distributed cognition theorists both readily acknowledge the phenomenon of transfer but propose different ways of accounting for it. An analysis developed by Bransford and Schwartz (1999) serves to summarize concerns about the traditional idea of transfer that underpins many cognitive science investigations.

Writing from within the cognition plus view, Bransford and Schwartz (1999) observe that the theory underlying much cognitive science research characterizes transfer as the ability to apply previous learning directly to a new setting or problem. As they note, transfer investigations are designed to ensure that subjects do not have the opportunity to learn to solve new problems either by getting feedback or by using texts and colleagues as resources. Although Bransford and Schwartz indicate that they consider this traditional perspective valid, they also argue for a broader conception of transfer that includes a focus on whether people are prepared to learn to solve new problems. In making this proposal, they attempt to bring the active nature of transfer to the fore. As part of their rationale, they give numerous examples to illustrate that people often learn to operate in a new setting by actively changing the setting rather than by mapping previously acquired understandings directly on to it. This leads them to argue that cognitive scientists should change their focus by looking for evidence of useful future learning rather than of direct application.

Bransford and Schwartz's proposal deals primarily with issues of method in that it does not challenge transfer as a theoretical idea. Nonetheless, their emphasis on preparation for future learning is evident in Greeno and MMAP's (1998) distributed analysis of the phenomenon of transfer. Greeno and MMAP contend that the ways of knowing that people develop in particular settings emerge as relations between them and the immediate environment. This necessarily implies that transfer involves active adaptation to new circumstances. As part of their theoretical approach, Greeno and MMAP analyze specific environments in terms of the affordances and constraints that they provide for reasoning. In the case of a traditional mathematics classroom, for example, the affordances might include the organization of lessons and of the textbook. The constraints might include the need to produce correct answers, the limited time available to complete sets of exercises, and the lack of access to peers and other resources. Greeno and MMAP would characterize the process of learning to be a successful student in such a classroom as one of becoming attuned to these affordances and constraints. From this distributed perspective, the phenomenon called transfer is then explained in terms of the similarity of the constraints and affordances of different settings rather than in terms of the transportation of knowledge from one setting to another.

The focus on preparation for future learning advocated by Bransford and Schwartz (1999) is quite explicit in Beach's (1995) investigation of the transition between work and school in a Nepal village where formal schooling had been introduced during the last 20 years of the twentieth century. Beach worked within the CHAT tradition when he compared the arithmetical reasoning of high school students who were apprentice shopkeepers with the reasoning of shopkeepers who were attending adult education classes. His analysis revealed that the shopkeepers' arithmetical reasoning was more closely related in the two situations than was that of the students. In line with the basic tenets of CHAT, Beach accounted for this finding by framing the shopkeepers' and students' reasoning as acts of participation in relatively global

cultural practices, those of shopkeeping and of studying arithmetic in school. His explanation hinges on the observation that the students making the school-to-work transition initially defined themselves as students but subsequently defined themselves as shopkeepers when they worked in a shop. In contrast, the shopkeepers continued to define themselves as shopkeepers even when they participated in the adult education classes. Their goal was to develop arithmetical competencies that would enable them to increase the profits of their shops. In Beach's view, it is this relatively strong relationship between the shopkeepers' participation in the practices of schooling and shopkeeping that explains the close relationship between their arithmetical reasoning in the two settings. Thus, whereas Greeno and MMAP account for instances of the phenomenon called transfer in terms of similarities in the affordances and constraints of immediate environments, Beach does so in terms of the experienced commensurability of certain forms of participation. In the case at hand, the phenomenon called transfer occurred to a greater extent with the shopkeepers because they experienced participating in the practice of shopkeeping and schooling as more commensurable than did the students.

The contrast between Beach's and Greeno and MMAP's analyses illustrates how general differences between the distributed cognition and CHAT traditions can play out in explanations of the phenomenon called transfer. However, attempts to reconceptualize transfer within these two are still in their early stages and there is every indication that this will continue to be a major focus of CHAT and distributed cognition research in the coming years.

3.2 Participating in Multiple Communities of Practice

To date, the bulk of CHAT research has focused on the forms of reasoning that people develop as they participate in particular cultural practices. For their part, distributed cognition theorists have been primarily concerned with the forms of reasoning that emerge as people use the cognitive resources of the immediate environment. In both cases, the focus has been on participation in well-circumscribed communities of practice or engagement in local systems of activity. An issue that appears to be emerging as a major research question is that of understanding how people deal with the tensions they experience when the practices of the different communities in which they participate are in conflict. The potential significance of this issue is illustrated by an example taken from education.

A number of studies reveal that students' home communities can involve differing norms of participation, language, and communication, some of which might be in conflict with those that the teacher seeks to establish in the classroom. Pragmatically, these studies lead directly to concerns for equity and indicate the importance of viewing the diversity of the practices of students' home communities as an instructional resource rather than an obstacle to be overcome. Theoretically, these studies indicate the importance of coming to understand how students attempt to resolve ongoing tensions between home and school practices. The challenge is therefore to develop analytical approaches that treat students' activity in the classroom as situated not merely with respect to the immediate learning environment, but with the respect to their history of participation in the practices of particular out-of-school communities. The classroom would then be viewed as the immediate arena in which the students' participation in multiple communities of practice play out in face-to-face interaction. This same general orientation is also relevant when attempting to understand people's activity in a number of other specific settings. As this is very much an emerging issue, CHAT and distributed cognition research on participation in multiple communities of practice is still in its infancy. The most comprehensive theoretical exposition to date is perhaps that of Wenger (1998). There is every reason to predict that this issue will become an increasingly prominent focus of future research in both traditions.

4. Concluding Comments

It should be apparent from this overview of contemporary developments in situated cognition that whereas CHAT researchers draw directly on Vygotsky's and Leont'ev's theoretical insights, the relationship to the Soviet scholarship is less direct in the case of distributed cognition research. Inspired in large measure by the findings of CHAT researchers, this latter tradition has emerged as a reaction to perceived limitations of mainstream cognitive science. As a consequence, distributed cognition theorists tend to maintain a dialog with their mainstream colleagues. In contrast, the problems that CHAT researchers view as significant are typically far less influenced by mainstream considerations.

The issue addressed in the previous section of this overview was that of the possible future directions for the two research traditions. It is important to note that the focus on just two issues—transfer and participation in multiple communities—was necessarily selective. It would have been possible to highlight a number of other issues that include the learning of the core ideas of academic disciplines and the design of tools as cognitive resources. In both cases, it can legitimately be argued that the CHAT and distributed cognition traditions each have the potential to make significant contributions. More generally, it can legitimately be argued that both research traditions are in progressive phases at the present time.

See also: Cultural Psychology; Learning by Occasion Setting; Situated Cognition: Origins; Situated Knowledge: Feminist and Science and Technology Studies Perspectives; Vygotskij, Lev Semenovic (1896–1934); Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Beach K 1995 Activity as a mediator of sociocultural change and individual development: The case of school-work transition in Nepal. *Mind, Culture, and Activity* **2**: 285–302

Bransford J D, Schwartz D L 1999 Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education* **24**: 61–100

Brown J S, Collins A, Duguid P 1989 Situated cognition and the culture of learning. *Educational Researcher* **18**(1): 32–42

Greeno J G 1998 The situativity of knowing, learning, and research. *American Psychologist* **53**: 5–26

Hutchins E 1995 *Cognition in the Wild*. MIT Press, Cambridge, MA

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participating*. Cambridge University Press, Cambridge, UK

Olson D R 1995 Writing and the mind. In: Wertsch J V, del Rio P, Alvarez A (eds.) *Sociocultural Studies of Mind*. Cambridge University Press, New York, pp. 95–123

Pea R D 1993 Practices of distributed intelligence and designs for education. In: Salomon G (ed.) *Distributed Cognitions*: Psychological and Educational Considerations. Cambridge University Press, New York, pp. 47–87

Saxe G B 1991 *Culture and Cognitive Development: Studies in Mathematical Understanding*. Erlbaum, Hillsdale, NJ

Scribner S 1984 Studying working intelligence. In: Rogoff B, Lave J (eds.) *Everyday Cognition: Its Development in Social Context*. Harvard University Press, Cambridge, MA, pp. 9–40

Wenger E 1998 *Communities of Practice*. Cambridge University Press, New York and Cambridge, UK

P. Cobb

Situated Cognition: Origins

Situated cognition encompasses a range of theoretical positions that are united by the assumption that cognition is inherently tied to the social and cultural contexts in which it occurs. This initial definition serves to differentiate situated perspectives on cognition from what Lave (1991) terms the *cognition plus* view. This latter view follows mainstream cognitive science in characterizing cognition as the internal processing of information. However, proponents of this view also acknowledge that individual cognition is influenced both by the tools and artifacts that people use to accomplish goals and by their ongoing social interactions with others.

To accommodate these insights, cognition plus theorists analyze the social world as a network of factors that influence individual cognition. Thus, although the cognition plus view expands the conditions that must be taken into account when developing adequate explanations of cognition and learning, it does not reconceptualize the basic nature of cognition. In contrast, situated cognition theorists challenge the assumption that social process can be clearly partitioned off from cognitive processes and treated as external condition for them. These theorists instead view cognition as extending out into the world and as being social through and through. They therefore attempt to break down a distinction that is basic both to mainstream cognitive science and to the cognition plus view, that between the individual reasoner and the world reasoned about.

1. Situation and Context

I can further clarify this key difference between the situated and cognition plus viewpoints by focusing on the underlying metaphors that serve to orient adherents to each position as they frame questions and plan investigations. These metaphors are apparent in the different ways that adherents to the two positions use the key terms *situation* and *context* (Cobb and Bowers 1999, Sfard 1998). The meaning of these terms in situated theories of cognition can be traced to the notion of position as physical location. In everyday conversation we frequently elaborate this notion metaphorically when we describe ourselves and others as being positioned with respect to circumstances in the world of social affairs. This metaphor is apparent in expressions such as, 'My situation at work is pretty good at the moment.' In this and similar examples, the world of social affairs (e.g., work) in which individuals are considered to be situated is the metaphorical correlate of the physical space in which material objects are situated in relation to each other. Situated cognition theorists such as Lave (1988), Saxe (1991), Rogoff (1995), and Cole (1996) elaborate this notion theoretically by introducing the concept of participation in cultural practices (see Fig. 1). Crucially, this core construct of participation is not restricted to face-to-face interactions with others. Instead, all individual actions are viewed as elements or aspects of an

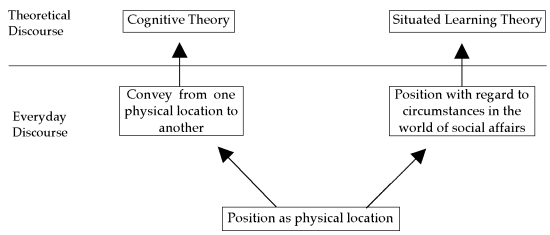

Figure 1
Metaphorical underpinnings of the cognition plus and situated cognition perspectives

encompassing system of cultural practices, and individuals are viewed as participating in cultural practices even when they are in physical isolation from others. Consequently, when situated cognition theorists speak of context, they are referring to a sociocultural context that is defined in terms of participation in a cultural practice.

This view of context is apparent in a number of investigations in which situated cognition theorists have compared mathematical reasoning in school with mathematical reasoning in various out-of-school settings such as grocery shopping (Lave 1988), packing crates in a dairy (Scribner 1984), selling candies on the street (Nunes et al. 1993, Saxe 1991), laying carpet (Masingila 1994), and growing sugar cane (DeAbreu 1995). These studies document significant differences in the forms of mathematical reasoning that arise in the context of different practices that involve the use of different tools and sign systems, and that are organized by different overall motives (e.g., learning mathematics as an end in itself in school vs. doing arithmetical calculations while selling candies on the street in order to survive economically). The findings of these studies have been influential, and challenge the view that mathematics is a universal form of reasoning that is free of the influence of culture.

The underlying metaphor of the cognition plus view also involves the notion of position. However, whereas the core metaphor of situated theories is that of position in the world of social circumstances, the central metaphor of the cognition plus view is that of the transportation of an item from one physical location to another (see Fig. 1). This metaphor supports the characterization of knowledge as an internal cognitive structure that is constructed in one setting and subsequently transferred to other settings in which it is applied. In contrast to this treatment of knowledge as an entity that is acquired in a particular setting, situated cognition theorists are more concerned with knowing as an activity or, in other words, with types of reasoning that emerge as people participate in particular cultural practices.

In line with the transfer metaphor, context, as it is defined by cognition plus theorists, consists of the task an individual is attempting to complete together with others' actions and available tools and artifacts. Crucially, whereas situated cognition theorists view both tools and others' actions as integral aspects of an encompassing practice that only have meaning in relation to each other, cognition plus theorists view them as aspects of what might be termed the stimulus environment that is external to internal cognitive processing. Consequently, from this latter perspective, context is under a researcher's control just as is a subject's physical location, and can therefore be systematically varied in experiments. Given this conception of context, cognition plus theorists can reasonably argue that cognition is not always tied to context because people do frequently use what they have learned in one setting as they reason in other settings. However, this claim is open to dispute if it is interpreted in terms of situated cognition theory where all activity is viewed as occurring in the context of a cultural practice. Further, situated cognition theorists would dispute the claim that cognition is partly context independent because, from their point of view, an act of reasoning is necessarily an act of participation in a cultural practice.

As an illustration, situated cognition theorists would argue that even a seemingly decontextualized form of reasoning such as research mathematics is situated in that mathematicians use common sign systems as they engage in the communal enterprise of producing results that are judged to be significant by adhering to agreed-upon standards of proof. For situated cognition theorists, mathematicians' reasoning cannot be adequately understood unless it is located within the context of their research communities. Further, these theorists would argue that to be fully adequate, an explanation of mathematicians' reasoning has to account for its development or genesis. To address this requirement, it would be necessary to analyze the process of the mathematicians' induction into their research communities both during their graduate education and during the initial phases of their academic careers.

This illustrative example is paradigmatic in that situated cognition theorists view learning as synonymous with changes in the ways that individuals participate in the practices of communities. In the case of research mathematics, these theorists would argue that mathematicians develop particular ways of reasoning as they become increasingly substantial participants in the practices of particular research communities. They would therefore characterize the mathematicians' apparently decontextualized reasoning as embedded in these practices and thus as being situated.

2. *Vygotsky's Contribution*

Just as situated cognition theorists argue that an adequate analysis of forms of reasoning requires that we understand their genesis, so an adequate account of contemporary situated perspectives requires that we trace their development. All current situated approaches owe a significant intellectual debt to the Russian psychologist Lev Vygotsky, who developed his cultural-historical theory of cognitive development in the period of intellectual ferment and social change that followed the Russian revolution. Vygotsky was profoundly influenced by Karl Marx's argument that it is the making and use of tools that serves to differentiate humans from other animal species. For Vygotsky, human history is the history of artifacts such as language, counting systems, and writing, that

are not invented anew by each generation but are instead passed on and constitute the intellectual bequest of one generation to the next. His enduring contribution to psychology was to develop an analogy between the use of physical tools and the use of intellectual tools such as sign systems (Kozulin 1990, van der Veer and Valsiner 1991). He argued that just as the use of a physical tool serves to reorganize activity by making new goals possible, so the use of sign systems serves to reorganize thought. From this perspective, culture can therefore be viewed as a repository of sign systems and other artifacts that are appropriated by children in the course of their intellectual development (Vygotsky 1978). It is important to understand that for Vygotsky, children's mastery of a counting system does merely enhance or amplify an already existing cognitive capability. Instead, children's ability to reason numerically is created as they appropriate the counting systems of their culture. This example illustrates Vygotsky's more general claim that children's minds are formed as they appropriate sign systems and other artifacts.

Vygotsky refined his thesis that children's cognitive development is situated with respect to the sign systems of their culture as he pursued several related lines of empirical inquiry. In his best known series of investigations, he attempted to demonstrate the crucial role of face-to-face interactions in which an adult or more knowledgeable peer supports the child's use of an intellectual tool such as a counting system (Vygotsky 1962). For example, one of the child's parents might engage the child in a play activity in the course of which parent and child count together. Vygotsky interpreted observations of such interactions as evidence that the use of sign systems initially appears in children's cognitive development on what he termed the 'intermental plane of social interaction.' He further observed that over time the adult gradually reduces the level of support until the child can eventually carry out what was previously a joint activity on his or her own. This observation supported his claim that the child's mind is created via a process of internalization from the intermental plane of social interaction to the intramental plane of individual thought.

A number of Western psychologists revived this line of work in the 1970s and 1980s by investigating how adults support or scaffold children's learning as they interact with them. However, in focusing almost exclusively on the moves or strategies of the adult in supporting the child, they tended to portray learning as a relatively passive process. This conflicts with the active role that Vygotsky attributed to the child in constructing what he referred to as the 'higher mental functions' such as reflective thought. In addition, these turn-by-turn analyses of adult–child interactions typically overlooked the broader theoretical orientation that informed Vygotsky's investigations. In focusing on social interaction, he attempted to demonstrate both that the environment in which the child learns is socially organized and that it is the primary determinant of the forms of thinking that the child develops. He therefore rejected descriptions of the child's learning environment that are cast in terms of absolute indices because such a characterization defines the environment in isolation from the child. He argued that the analyses should instead focus on what the environment means to the child. This, for him, involved analyzing the social situations in which the child participates and the cognitive processes that the child develops in the course of that participation.

Thus far, in discussing Vygotsky's work, I have emphasized the central role that he attributed to social interactions with more knowledgeable others. There is some indication that shortly before his death in 1934 at the age of 36, he was beginning to view the relation between social interaction and cognitive development as a special case of a more general relation between cultural practices and cognitive development (Davydov and Radzikhovskii 1985, Minick 1987). In doing so he came to see face-to-face interactions as located within an encompassing system of cultural practices. For example, he argued that it is not until children participate in the activities of formal schooling that they develop what he termed scientific concepts. In making this claim, he viewed classroom interactions between a teacher and his or her students as an aspect of the organized system of cultural practices that constituted formal schooling in the USSR at that time. A group of Soviet psychologists, the most prominent of whom was Alexei Leont'ev, developed this aspect of cultural-historical theory more fully after Vygotsky's death.

3. *Leont'ev's Contribution*

Two aspects of Leont'ev's work are particularly significant from the vantage point of contemporary situated perspectives. The first concerns his clarification of an appropriate unit of analysis when accounting for intellectual development. Although face-to-face interactions constitute the immediate social situation of the child's development, Leont'ev saw the encompassing cultural practices in which the child participates as constituting the broader context of his or her development (Leont'ev 1978, 1981). For example, Leont'ev might have viewed an interaction in which a parent engages a child in activities that involve counting as an instance of the child's initial, supported participation in cultural practices that involve dealing with quantities. Further, he argued that children's progressive participation in specific cultural practices underlies the development of their thinking. Intellectual development was, for him, synonymous with the process by which the child becomes a full participant in particular cultural practices. In other words, he viewed the development of children's minds and

their increasingly substantial participation in various cultural practices as two aspects of a single process. This is a strongly situated position in that the cognitive capabilities that children develop are not seen as distinct from the cultural practices that constitute the context of their development. From this perspective, the cognitive characteristics a child develops are characteristics of the child-in-cultural-practice in that they cannot be defined apart from the practices that constitute the child's world. For Leont'ev, this implied that the appropriate unit of analysis is the child-in-culture-practice rather than the child *per se*. As stated earlier, it is this rejection of the individual as a unit of analysis that separates contemporary situated perspectives from what I called the cognition plus viewpoint.

Leont'ev's second important contribution concerns his analysis of the external world of material objects and events. Although Vygotsky brought sign systems and other cultural tools to the fore, he largely ignored material reality. In building on the legacy of his mentor, Leont'ev argued that material objects as they come to be experienced by developing children are defined by the cultural practices in which they participate. For example, a pen becomes a writing instrument rather than a brute material object for the child as he or she participates in literacy practices. In Leont'ev's view, children do not come into contact with material reality directly, but are instead oriented to this reality as they participate in cultural practices. He therefore concluded that the meanings that material objects come to have are a product of their inclusion in specific practices. This in turn implied that these meanings cannot be defined independently of the practices. This thesis only served to underscore his argument that the individual-cultural-practice constitutes the appropriate analytical unit for psychology.

See also: Cultural Psychology; Learning by Occasion Setting; Situated Cognition: Contemporary Developments; Situated Knowledge: Feminist and Science and Technology Studies Perspectives; Situated Learning: Out of School and in the Classroom; Transfer of Learning, Cognitive Psychology of; Vygotskij, Lev Semenovic (1896–1934); Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Cobb P, Bowers J 1999 Cognitive and situated perspectives in theory and practice. *Educational Researcher* **28**(2): 4–15

Cole M 1996 *Cultural Psychology*. Harvard University Press, Cambridge, MA

Davydov V V, Radzikhovskii L A 1985 Vygotsky's theory and the activity-oriented approach in psychology. In: Wertsch J V (ed.) *Culture, Communication, and Cognition: Vygotskian Perspectives*. Cambridge University Press, New York, pp. 35–65

DeAbreu G 1995 Understanding how children experience the relationship between home and school mathematics. *Mind, Culture, and Activity* **2**: 119–42

Kozulin A 1990 *Vygotsky's Psychology: A Biography of Ideas*. Harvard University Press, Cambridge, MA

Lave J 1988 *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life*. Cambridge University Press, New York

Lave J 1991 Situating learning in communities of practice. In: Resnick L B, Levine J M, Teasley S D (eds.) *Perspectives on Socially Shared Cognition*. American Psychological Association, Washington, DC, pp. 63–82

Leont'ev A N 1978 *Activity, Consciousness, and Personality*. Prentice-Hall, Englewood Cliffs, NJ

Leont'ev A N 1981 The problem of activity in psychology. In: Wertsch J V (ed.) *The Concept of Activity in Soviet Psychology*. Sharpe, Armonk, NY, pp. 37–71

Masingila J O 1994 Mathematics practice in carpet laying. *Anthropology and Education Quarterly* **25**: 430–62

Minick N 1987 The development of Vygotsky's thought: An introduction. In: Rieber R W, Carton A S (eds.) *The Collected Works of Vygotsky, L.S.* (Vol. 1): *Problems of General Psychology*. Plenum, New York, pp. 17–38

Nunes T, Schliemann A D, Carraher D W 1993 *Street Mathematics and School Mathematics*. Cambridge University Press, Cambridge, UK

Rogoff B 1995 Observing sociocultural activity on three planes: Participatory appropriation, guided participation, and apprenticeship. In: Wertsch J V, del Rio P, Alvarez A (eds.) *Sociocultural Studies of Mind*. Cambridge University Press, New York, pp. 139–64

Saxe G B 1991 *Culture and Cognitive Development: Studies in Mathematical Understanding*. Erlbaum, Hillsdale, NJ

Scribner S 1984 Studying working intelligence. In: Rogoff B, Lave J (eds.) *Everyday Cognition: Its Development in Social Context*. Harvard Univesity Press, Cambridge, pp. 9–40

Sfard A 1998 On two metaphors for learning and the dangers of choosing just one. *Educational Research* **27**(21): 4–13

van der Veer R, Valsiner J 1991 *Understanding Vygotsky: A Quest for Synthesis*. Blackwell, Cambridge, MA

Vygotsky L S 1962 *Thought and Language*. MIT Press, Cambridge, MA

Vygotsky L S 1978 *Mind and Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

P. Cobb

Situated Knowledge: Feminist and Science and Technology Studies Perspectives

The expression 'situated knowledge,' especially in its plural form 'situated knowledges,' is associated with feminist epistemology, feminist philosophy of science, and science and technology studies. The term was introduced by historian of the life sciences and feminist science and technology studies scholar Donna Haraway in her landmark essay *Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective* (Haraway 1991, pp. 183–201). The

essay was a response to feminist philosopher of science Sandra Harding's discussion of the 'science question in feminism' (Harding 1986). In her analysis of the potential of modern science to contribute to the goals of feminism, Harding noted three different accounts of objective knowledge in feminist epistemology: feminist empiricism, feminist standpoint, and feminist postmodernism (Harding 1986, pp. 24–6). Feminist empiricism attempted to replace more biased with less biased science. The feminist standpoint, echoing the Marxist tradition from which it derived, stressed the relevance of the social positioning of the knower to the content of what is known. Feminist postmodernism accentuated the power dynamics underlying the use of the language of objectivity in science. Haraway, taking off from Harding, diagnosed a 'Scylla and Charybdis' of temptations between which feminists attempt to navigate on the question of objectivity: radical constructionism and feminist critical empiricism. As she put it, what feminists wanted from a theory of objectivity was 'enforceable, reliable accounts of things not reducible to power moves and agonistic, high status games of rhetoric or to scientistic, positivist arrogance' (Haraway 1991, p. 188). Her notion of situated knowledges was her attempt to provide just such a theory of objectivity.

Despite the single author provenance, the term resonated with attempts by other scholars in the history, sociology, and philosophy of science to address similar epistemological tensions. The concept was quickly and fruitfully taken up in science and technology studies and feminist theory, provoking a certain amount of reworking in its turn.

1. *'Situated Knowledge' and Objectivity: Constructed Yet Real*

The term 'situated knowledge' derives its theoretical importance from its seemingly oxymoronic character, particularly when applied to knowledge about the natural world. It is common to think of modern scientific knowledge as universal, so that it has the same content no matter who possesses it. It is also almost definitional to hold that objective knowledge is warranted by the fact that it captures reality as it really is, rather than being warranted by the situational circumstances out of which the knowledge was generated or discovered (see Shapin 1994, pp. 1–8 for discussion of, and citations relevant to, various manifestations of these points). Thus, if the law of gravity enables us to make reliable experimental predictions, it is because there is such a thing as gravity that is adequately captured by our scientific understanding; in short, the truth of the knowledge is its own warrant. It is only in the case of false or superseded knowledge that we typically explain what went wrong by reference to faulty assumptions, sloppy work, ill-calibrated equipment, the *Zeitgeist*, or other aspects of the context of discovery. The idea of 'situated knowledge' contests these supposed concomitants of objective knowledge. It suggests that objective knowledge, even our best scientific knowledge of the natural world, depends on the *partiality* of its material, technical, social, semiotic, and embodied means of being promulgated.

Haraway's notion thus has affinities with other feminist epistemologies which have noted that facts can differ in their content from one time, place, and knower to another (e.g., Collins 1989). It also has sympathies in common with sociologists of science and scholars of science and technology studies who have suggested that capturing 'reality as it really is' may be dependent on institutional, technical, and cultural norms (Kuhn (1962/1970), on practice (Clarke and Fujimura 1992, Pickering 1992), and attempts to witness, measure, comprehend, or command assent to it (Latour 1987, Shapin 1994, Shapin and Schaffer 1985). All these scholars share a search for the theoretical resources to do justice to the embeddedness of science and truth. These challenges to conventional views of objectivity bring situated knowledges into conversation with key debates in the philosophy of science around the theory-ladenness of facts (Hesse 1980, pp. 63–110). Additionally, the suspicion of transcendent universalism entrains an epistemological and political distrust of clear-cut distinctions between subject and object, and a blurring of the distinction between context and content of knowledge or discovery.

Situated knowledges are as hostile to relativism as they are to realism. Haraway describes relativism as 'being nowhere while claiming to be everywhere equally' (Haraway 1991, p. 191) and realism as 'seeing everything from nowhere' (Haraway 1991, p. 189), and conceives of them both as 'god-tricks' promising total, rather than partial, located, and embodied, vision. In contrast to realist or relativist epistemologies, Haraway sees the possibility of sustained and rational objective inquiry in the epistemology of partial perspectives. This requires, she maintains, reclaiming vision as a series of technological and organic embodiments, as and when and where and how vision is actually enabled. This crafting of a feminist epistemology of situated knowledges on the basis of vision and partial perspective is noteworthy. The links in the history of science to militarism, capitalism, colonialism, and male supremacy have been theorized around the masculinist gaze of the powerful but disembodied knower disciplining and subjugating the weak by means of a multitude of technologies of surveillance. Feminists have lamented the privilege granted to the visual as a sure basis of knowledge and bemoaned the sidelining in modernity of what some cast as more feminine and less intrinsically violent ways of knowing involving emotion, voice, touch, and listening (Gilligan 1982). Haraway is concerned that feminists not cede power to those whose practices they wish

critically to engage. It is in this spirit that she grounds her feminist solution in an embrace of science and vision, 'the real game in town, the one we must play' (Haraway 1991, p. 184).

2. *The Feminist Roots of the Dilemma*

A tension between emancipatory empiricism and its associated egalitarian or socialist politics, and feminist postmodern constructionism and its associated identity politics, resonates throughout contemporary Western feminist theory. It is a recent hallmark of those engaged in feminist philosophical and social studies of science that they seek to resolve one or another version of this tension. One horn of the feminist dilemma, according to Haraway, represents the good feminist reasons to be attracted to radical constructionism. Feminist postmodernists, and analysts of science and technology influenced by semiotics (including Haraway herself), helped develop and often appeal to 'a very strong social constructionist argument for *all* forms of knowledge claims' including scientific ones (Akrich and Latour 1992, Haraway 1991, p. 184). This position has the benefit of showing the links between power—such things as status, equipment, rhetorical privilege, funding, and so on—and the production of knowledge and credibility. The downside, from the point of view of feminists interested in arguing for a better world, is that the radical constructionist argument risks rendering all knowledges as fundamentally ideological, with no basis for choosing between more and less just ideas, more and less true versions of reality. As Haraway provocatively expressed it, embracing this temptation seemed to leave no room for 'those of us who would still like to talk about *reality* with more confidence than we allow the Christian right's discussion of the Second Coming and their being raptured out of the final destruction of the world' (Haraway 1991, p. 185).

The second horn of the dilemma, according to Haraway, involves 'holding out for a feminist version of objectivity' through materialism or empiricism (Haraway 1991, p. 186). Haraway briefly discusses both Marxist derived feminisms and feminist empiricism. Feminisms with Marxist inspirations are several, and their genealogy can be traced in a number of ways. Feminists have long criticized Marxist humanism for its premise that the self-realization of man is dependent on the domination of nature, and for its account of the historical progression of modes of production that grants no historical agency to domestic and unpaid labor (Hartmann 1981). Some have responded by developing feminist versions of historical materialism (Hartsock 1983). Feminist standpoint theorists appropriated the general insight, inherited from both Marxist thought and the sociology of knowledge, that one's social-structural position in society—such things as one's class or relation to the means of production, or one's gender or ethnonational characteristics—determine or affect how and what one knows (Smith 1990).

Likewise, the idea that some social structural positions confer epistemological privilege has been widely adopted by standpoint theorists and feminist epistemologists arguing for specifically feminine ways of knowing (Rose 1983). 'Seeing from below,' that is, from a position of subordination, has commonly been theorized by feminists as the position of epistemological privilege, on the grounds that those with little to gain from internalizing powerful ideologies would be able to see more clearly than those with an interest in reproducing the status quo. In Patricia Hill Collins' version of standpoint theory, for example, these insights are used both to validate the knowledges of the historically disenfranchised, and to reverse the hegemonic ranking of knowledge and authority, and claim epistemological privilege for African–American women (Collins 1989).

Psychoanalytic theory, particularly anglophone object relations theory, inspired some of the early writings on gender and science (Chodorow 1978, Keller 1985). Object relations theory attempted to explain the different relation of women and men to objectivity, abstract thought, and science in modern societies. To account for this difference, the theory posited gender-based differences in the socialization of sons and daughters in Western middle class heterosexual nuclear families. Boys, according to this theory, are socialized to separate from the primary caregiver who is the mother in this normative family scenario. They thus learn early and well by analogy with their emotional development that relational thinking is inappropriate for them; separating themselves from the object of knowledge, as from the object of love, is good. Girls, on the other hand, are supposedly socialized to be like their primary caregiver, so that they can reproduce mothering when their turn comes. Relationality and connectivity, not abstraction and separation, are the analogous ordering devices of girls' affective and epistemological worlds. As applied to objectivity and scientific knowledge, object relations theory seemed to explain to feminists, without resort to distasteful biological determinisms denying women scientific aptitude, why women were excluded from much of science and technology. It also suggested that there were (at least) two distinct ways of knowing, and that much might have been lost in the violence and separation of masculinist science that could be restored by a proper valuation of the feminine values of connection and empathy (Harding 1986, Keller 1983). Like Marxism, psychoanalytic approaches to objectivity gave feminists a means to show the relevance of one's social position to knowledge. Like feminist empiricism, they encouraged the belief in the possibility of an improved, feminist, objectivity (Harding 1992).

The feminist canon contains a number of empirical studies that have revealed the negative effects of such things as colonialism and stereotypes about race and gender on the production of reliable science (Fausto-Sterling 1995, Martin 1991/1996, Schiebinger 1989, Traweek 1988). Evelyn Fox Keller's call for 'dynamic objectivity' (Keller 1985, pp. 115–26) and Sandra Harding's demand for 'strong objectivity' (Harding 1992, p. 244) are exemplary of the aspirations of theoretical feminist empiricism. These projects seek to prescribe scientific methods capable of generating accounts of the world that would improve upon disembodied, masculinist portrayals of science because they would be alert to the practices of domination and oppression inherent in the creation, dissemination, and possession of knowledge. Feminist empiricism nonetheless remains problematic because of its reliance on the dichotomies of bias vs. objectivity, use vs. misuse, and science vs. pseudoscience. The feminist insight of the 'contestability of every layer of the onion of scientific and technological constructions' (Haraway 1991, p. 186) flies in the face of leaving these epistemological dichotomies intact.

3. *Subjects, Objects, and Agency*

Haraway's notion of situated knowledges problematizes both subject and object. Unlike standpoint theories which attribute epistemological privilege to subjugated knowers, and the sociology of knowledge which attributes espitemological privilege to those in the right structural position vis-à-vis a given mode of production, Haraway attributes privilege to partiality. This shift underscores that 'situated knowledge' is more dynamic and hybrid than other epistemologies that take the position of the knower seriously, and involves 'mobile positioning' (Haraway 1991, p. 192)

In situated knowledges based on embodied vision, neither subjects who experience, nor nature which is known, can be treated as straightforward, pre-theoretical entities, 'innocent and waiting outside the violations of language and culture' (Haraway 1991, p. 109). Haraway maintains that romanticizing, and thus homogenizing and objectifying, the perfect subjugated subject position is not the solution to the violence inherent in dominant epistemologies. As feminists from developing countries have also insisted, there is no innocent, perfectly subjugated feminist subject position conferring epistemological privilege; all positionings are open to critical re-examination (Mohanty 1984/1991). Subjectivity is instead performed in and through the materiality of knowledge and practice of many kinds (Butler 1990, pp. 1–34).

Conversely, the extraordinary range of objects in the physical, natural, social, political, biological, and human sciences about which institutionalized knowledge is produced should not be considered to be passive and inert. Haraway says that situated knowledges require thinking of the world in terms of the 'apparatus of bodily production.' The world cannot be reduced to a mere resource if subject and object are deeply interconnected. Bodies as objects of knowledge in the world should be thought of as 'material-semiotic generative nodes,' whose 'boundaries materialize in social interaction' (Haraway 1991, p. 201). The move to grant agency to material objects places the epistemology of situated knowledges at the center of recent scholarship in science and technology studies (Callon 1986, Latour 1987).

4. *Uptake and Critique*

Donna Haraway's essay ranks among the most highly cited essays in science and technology studies and has been anthologized. As stated above, situated knowledges is a provocative and rich methodological metaphor with resonances in many quarters. The dialogue between Harding and Haraway continued after the publication of *Situated Knowledges* (Harding 1992, pp. 119–63). Her epistemology has directly influenced, and has in turn been influenced by, the recent work of sociologists and anthropologists of science (Clarke and Montini 1993, Rapp 1999), feminist philosophers of science (Wylie 1999), and practicing feminist scientists (Barad 1996). In addition, 'situated knowledges' is used as a standard technical term of the field by more junior scholars.

Critics of situated knowledges have been few. Timothy Lenoir has pointed out that many of the epistemological ideas behind Haraway's situated knowledges are found not only in other major strands of science and technology studies, but also in the work of continental philosophers such as Nietzsche. He likewise critiqued the idea of situated knowledges for its dependence on the apparatus of semiotics (Lenoir 1999, pp. 290–301). Historian Londa Schiebinger, in her recent book summarizing the effects of a generation of feminist scholarship on the practice of science, places Haraway's situated knowledges together with Harding's strong objectivity as attempts to integrate social context into scientific analysis (Schiebinger 1999). Implicit critiques have been leveled against the limitations of the idea of being situated, for example, in the development of De Laet's and Mol's mobile epistemology (De Laet and Mol 2000). Sheila Jasanoff and her colleagues have argued for bringing differently spatialized entities such as the nation, the local, and the global, into the epistemology of science and technology studies, while retaining the insights gained by paying attention to practice, vision, and measurement. These critiques stand more as continuing conversations with, than rebuttals of situated knowledges, however. Overall, the idea of situated knowledges remains central to feminist epistemology and science studies and to attempts to understand the role of modern science in society.

See also: Cognitive Psychology: Overview; Contextual Studies: Methodology; Feminist Epistemology; Feminist Political Theory and Political Science; Feminist Theory; Feminist Theory: Postmodern; Knowledge, Anthropology of; Knowledge, Sociology of; Rationality and Feminist Thought; Science and Technology, Anthropology of; Scientific Knowledge, Sociology of; Situation Model: Psychological

Bibliography

Akrich M, Latour B 1992 A summary of a convenient vocabulary for the semiotics of human and nonhuman assumblies. In: Bijker W, Law J (eds.) *Shaping Technology/Building Society: Studies in Sociotechnical Change*. MIT Press, Cambridge, MA, pp. 259–64

Barad K 1996 Meeting the universe halfway: realism and social constructivism without contradiction. In: Hankinson Nelson L, Nelson J (eds.) *Feminism, Science, and the Philosophy of Science*. Kluwer, Dordrecht, The Netherlands, pp. 161–94

Butler J 1990 *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, London

Callon M 1986 Some elements of a sociology of translation: domestication of the scallops and fishermen of St. Brieuc Bay. In: J Law J (ed.) *Power, Action and Belief: A New Sociology of Knowledge*. Routledge, London, pp. 196–233

Chodorow N 1978 *The Reproduction of Mothering*. University of California Press, Berkeley

Clarke A, Fujimura J (eds.) 1992 *The Right Tools for the Job: At Work in Twentieth-Century Life Sciences*. Princeton University Press, Princeton, NJ

Clarke A, Montini T 1993 The many faces of RU486: Tales of situated knowledges and technological contestations. *Science, Technology and Human Values* **18**: 42–78

Collins P H 1989 The social construction of black feminist thought. *Signs* **14**(4): 745–73

De Laet M, Mol A 2000 The Zimbabwe bush pump. Mechanics of a fluid technology. *Social Studies of Science*

Fausto-Sterling A 1995 Gender, race, and nation: The comparative anatomy of 'Hottentot' Women in Europe, 1815–1817. In: Terry J, Urla J (eds.) *Deviant Bodies: Critical Perspectives on Difference in Science and Popular Culture*. Indiana University Press, Bloomington, IN, pp. 19–48

Gilligan C 1982 *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA

Haraway D 1991 *Simians, Cyborgs and Women: The Reinvention of Nature*. Routledge, New York

Harding S 1986 *The Science Question in Feminism*. Cornell University Press, Ithaca, NY

Harding S 1992 *Whose Science? Whose Knowledge? Thinking from Women's Lives*. Cornell University Press, Ithaca, NY

Hartmann H 1981 The unhappy marriage of Marxism and feminism. In: Sargent L (ed.) *Women and Revolution*. South End Press, Boston

Hartsock N 1983 The feminist standpoint: Developing the ground for a specifically feminist historical materialism. In: Harding S, Hintikka M (eds.) *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. Reidel, Dordrecht, The Netherlands, pp. 283–310

Hesse M 1980 *Revolutions and Reconstructions in the Philosophy of Science*. Indiana University Press, Bloomington, IN

Keller E F 1983 *A Feeling for the Organism: The Life and Work of Barbara McClintock*. W. H. Freeman and Company, New York

Keller E F 1985 *Reflections on Gender and Science*. Yale University Press, New Haven, CT

Kuhn T S 1962/1970 *The Structure of Scientific Revolutions*, 2nd edn. Chicago University Press, Chicago

Latour B 1987 *Science in Action: How to Follow Scientists and Engineers Through Society*. Open University Press, Philadelphia, PA

Lenoir T 1999 Was the last turn the right turn? The semiotic turn and A. J. Greimas. In: Biagioli M (ed.) *The Science Studies Reader*. Routledge, London, pp. 290–301

Martin E 1991/1996 The egg and the sperm: how science has constructed a romance based on stereotypical male–female roles. In: Laslett B, Kohlstedt G S, Longino H, Hammonds E (eds.) *Gender and Scientific Authority*. Chicago University Press, Chicago, pp. 323–39

Mohanty C T 1984/1991 Under Western eyes: Feminist scholarship and colonial discourses. In: Mohanty C, Russo A, Torres L (eds.) *Third World Women and the Politics of Feminism*. Indiana University Press, Bloomington, IN, pp. 51–80

Pickering A (ed.) 1992 *Science as Practice and Culture*. University of Chicago Press, Chicago

Rapp R 1999 *Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America*. Routledge, London

Rose H 1983 Hand, brain, and heart: A feminist epistemology for the natural sciences. *Signs* **9**(1): 73–90

Schiebinger L 1989 *The Mind Has No Sex? Women in the Origins of Modern Science*. Harvard University Press, Cambridge, MA

Schiebinger L 1999 *Has Feminism Changed Science*? Harvard University Press, Cambridge, MA

Shapin S 1994 *A Social History of Truth: Civility and Science in Seventeenth-century England*. University of Chicago Press, Chicago

Shapin S, Schaffer S 1985 *Leviathan and the Air-pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, Princeton, NJ

Smith D 1990 *The Conceptual Practices of Power; A Feminist Sociology of Knowledge*. University of Toronto Press, Toronto, ON

Traweek S 1988 *Beamtimes and Lifetimes: The World of High Energy Physicists*. Harvard University Press, Cambridge, MA

Wylie A 1999 The engendering of archaeology: Refiguring feminist science studies. In: Biagioli M (ed.) *The Science Studies Reader*. Routledge, London, pp. 553–68

C. M. Thompson

Situated Learning: Out of School and in the Classroom

Situated learning is not a unitary, well-defined concept. From an educational point of view, the core idea behind the different uses of this term is to create a situational context for learning that strongly resembles

possible application situations in order to assure that the learning experiences foster 'real-life' problem solving.

1. Situated Cognition–Situated Learning

Since 1985, the notion of situatedness of learning and knowing has become prominent in a variety of scientific disciplines such as psychology, anthropology, and computer science. In this entry, however, we focus only on situatedness approaches in education and educational psychology.

Since the late 1980s a particular educational problem has received much attention. In traditional forms of instruction, learners acquire knowledge that they can explicate, for example in examinations. When the learners are, however, confronted with complex 'real-life' problems, this knowledge is frequently not used, although it is relevant. Such knowledge is termed 'inert knowledge' (Renkl et al. 1996). This frequently found inertness phenomenon motivated some researchers to postulate that the whole notion of knowledge that is used in cognitive psychology as well as in everyday reasoning about educational issues is wrong (e.g., Lave 1988).

It has been argued that knowledge is not some 'entity' in a person's head that can be acquired in one situational context (e.g., classroom) and then be used in another context (e.g., workplace), but that instead it is context-bound. Hence, symbolic modeling of cognitive processes and structures is regarded as inappropriate. From a situatedness perspective, knowledge is generally constituted by the relation or interaction between an agent and the situational context they are acting in. Hence, it is proposed that we use the term 'knowing' instead of 'knowledge' in order to underline the process aspect of knowing/knowledge (Greeno et al. 1993). As a consequence of this conception of knowledge, learning must also be conceived as context-bound or situated. Thus, it is understandable that there is no transfer between such different contexts as the classroom and everyday life.

The theoretical assumption of the situatedness of knowing and learning also has consequences for research methodology. Laboratory experiments can no longer be seen as appropriate, because in this research strategy the phenomena are put out of their 'natural' context and their character is changed (e.g., Lave 1988). Situatedness proponents conduct primarily qualitative field studies.

Unfortunately, the term of situatedness is not well defined. One reason for the fuzziness of this construct is that its proponents differ in how far they depart from traditional cognitive concepts. Whereas Lave (1988), for example, radically rejects the notions of knowledge, transfer, and symbolic representations as not being sensible, other proponents, such as Greeno et al. (1993), hold a more modest position. Although they also stress the situatedness of knowing and learning, they aim, for example, to analyze conditions for transfer, or claim that representations can play a role in human activity. A factor that further contributes to the vagueness of the notion of situated learning is that it is used in two ways: descriptive and prescriptive. When descriptively used, situated learning means that learning is analyzed as a context-bound process. In the field of education, situated learning is, however, mostly used as prescriptive concept. It is argued that learning should be situated in a context that resembles the application contexts.

The prescriptive aspect of situated learning was quite appealing to the community of educational researchers in the 1990s. Hence, not only those who subscribe to the situated cognition perspective (i.e., rejection of cognitive concepts), but also people who still more or less remain within the traditional cognitive framework rely on this concept. The latter group often juxtapose situated learning with decontextualized (traditional) learning, in which concepts and principles are presented and acquired in an abstract way with little or no relation to 'real-world' problems. It is important to note that this 'assimilated' view, in principle, contradicts the more fundamental situatedness concept according to which there is no nonsituated learning. Even typical abstract school learning is situated in the very specific context of school culture, although this situatedness would usually be evaluated as unfavorable because of the differences between school and 'real-life' contexts. Hence, irrespective of the logical inconsistencies in the use of the notion of situated learning, the prescriptive notion of situated learning means that the learning and the application situations should be as similar as possible in order to assure that the learning experiences have positive effects on 'real-life' problem solving.

2. Learning and Problem Solving in School and Outside

From a situatedness perspective, typical school learning is not much help for problem solving in everyday or professional life because the contexts are too different. Based among others upon Resnick's (1987) analyses of typical differences between these contexts, the following main points are outlined:

(a) *Well-defined problems in school versus ill-defined problems outside*. For example, word problems in mathematics are usually well defined. It is clear what the problem is all about, all the necessary information is given in the problem formulation, there is usually only one way to arrive at the solution that is labeled as appropriate, and so on. Nontrivial, 'real-life' problems (e.g., improvement of an organization's communication structure) first often have to be defined more precisely for a productive solution: one has to decide whether one needs more information, one has to seek

information, one has to decide what is relevant and irrelevant information, there are multiple ways of solving a problem, and so on.

(b) *Content structured by theoretical systems in school versus structured by problems outside*. In traditional forms of instruction, content is structured according to theoretical systematizations (e.g., biological taxonomies). This helps learners to organize the content and to remember it. One very salient systematization of content is the distinction between different school subjects. When a 'real-life' problem (e.g., pollution in a city or evaluation of an Internet company's share) has to be solved, thinking within the boundaries of school subjects is often not helpful. Furthermore, the structure of concepts used in school is frequently not relevant to the problem at hand. In 'real life,' the nature of the problem to be solved determines what concepts and information are required and in which structure they are needed.

(c) *Individual cognition in school versus shared cognition outside*. In schools, the usual form of learning and performance is individualistic. Cooperation in examinations is even condemned. In professional or everyday life, in contrast, cooperation is valued and it is frequently necessary for solving problems.

(d) *Pure mentation in school versus tool manipulation outside*. In traditional instruction, pure 'thought' activities are dominating. Students should learn to perform without support of tools such as books, notes, calculators, etc. Especially in exams, tools are usually forbidden. In contrast, a very important skill in everyday or professional life is the competent use of tools.

(e) *Symbol manipulation in school versus contextualized reasoning outside*. Abstract manipulation of symbols is typical of traditional instruction. Students often fail to match symbols and symbolic processes to 'real-world' entities and processes. In ordinary life, on the other hand, not only are tools used, but reasoning processes are an integral part of activities that involve objects and other persons. 'Real-world' reasoning processes are typically situated in rich situational contexts.

(f) *Generalized learning in school versus situation-specific competencies outside*. One reason for the abstract character of traditional instruction is that it aims to teach general, widely usable skills and theoretical principles. Nobody can foresee what types of specific problem students will encounter in their later life. In everyday and professional life, in contrast, situation-specific skills must be acquired. Further learning is mostly aimed at competencies for specific demands (e.g., working with a new computer program).

On the one hand, this list is surely not complete and more differences could be outlined. On the other hand, this juxtaposition is somewhat oversimplifying. Nevertheless, there is a core of truth because *typical* learning in and out of school differs significantly. Given these differences, a consequence of situatedness assumptions is to claim that learning environments in school should strongly resemble application contexts or that learning should take place in the field (e.g., 'on the job'). A traditional model in which people acquire applicable skills in authentic contexts is apprenticeship learning.

3. *Apprenticeship Learning as Situated Learning*

The situatedness protagonists Lave and Wenger (1991) investigated out of school learning in the form of apprenticeship by means of qualitative analyses. They focused on people working in traditional skills, for example Indian midwives in Mexico, tailors in Liberia, and butchers in supermarkets. In these traditional apprenticeships, learners acquire mainly manual skills. In modern society, in contrast, 'cognitive' domains prevail (e.g., computer science, psychology), so that skilled activity is hardly 'visible.' Also in school learning, cognitive skills such as mathematical problem solving, reading, and writing dominate. Against this background, Collins et al. (1989) developed the instructional cognitive apprenticeship model. Herein the importance of explication or reification of cognitive processes (e.g., strategies, heuristics) during learning is stressed. Thus, cognitive processes can be approximately as explicit as the more manual skills trained in traditional apprenticeship.

The core of cognitive apprenticeship is a special instructional sequence and the employment of authentic learning tasks. Experts provide models in applying their knowledge in authentic situations. Thereby they externalize (verbalize) their reasoning. The learners then work on authentic tasks of increasing complexity and diversity. An expert or teacher is assigned an important role as a model and as a coach providing scaffolding. The learners are encouraged increasingly to take an active role, as the support by the expert is gradually faded out. Articulation is promoted so that normally internal processes are externalized and can be reflected. This means that one's own strategies can be compared with those of experts, are then open to feedback, and can be discussed. In addition, the student's own cognitive strategies can be compared with those of other students. In the course of interaction with experts and other learners, students can also get to know different perspectives on concepts and problems. As a result of this instructional sequence, the students increasingly work on their own (exploration) and may take over the role initially assumed by the expert. Lave and Wenger (1991) have characterized such a sequence as development from legitimate peripheral participation to full participation.

It is important to note that the type of apprenticeship learning that is envisioned by the proponents of the situatedness approaches implies much more than the acquisition of 'subject-matter knowledge.' It

is a process of enculturation. The learner gradually develops the competence to participate in a community of practice. Such participation presupposes more than the type of knowledge usually focused on in classroom learning. In addition, 'tricks of the trade' and knowledge of social norms, for example, are required.

4. Problem-based Learning as Situated Learning

Besides apprenticeship, problem-based learning is another possibility for implementing arrangements in accordance with a situatedness rationale. Learning should be motivated by a complex 'real-world' problem that is the starting point of a learning process. For example, Greeno and the Middle School Mathematics Through Applications Project Group (MMAP) (1998) designed learning arrangements in which mathematical reasoning is not triggered primarily in separate mathematics lessons, but within design activities in four domains: architecture, population biology, cryptography, and cartography. The design activities are supported by the employment of computer tools which are also typical of current situated learning arrangements. An instructional principle is to induce quantitative reasoning involving proportions, ratios, and rates during design activities that strongly resemble the activities of many everyday crafts and commercial practices. The mathematical reasoning within design activities can be quite sophisticated; however, it often remains implicit. The teachers' task is to uncover the mathematics the students are implicitly using. For this purpose, there are, among others, curricular materials for 'math extension' units (i.e., explicit mathematics lessons).

5. Common Critiques of the Situatedness Approach

The situatedness camp has criticized the fundamental assumptions of cognitively oriented educational research. Hence it is not surprising that the situatedness camp has also been heavily attacked. Three major objections against the situatedness approach and corresponding defending arguments should be outlined:

(a) *Faulty fundamental assumptions*: Anderson et al. (1996), in particular, have argued that the situatedness approach is based on wrong assumptions such as 'knowledge does not transfer between tasks' or 'training in abstraction is of little use.' Anderson et al. cited empirical studies that contradict these assumptions. In his reply, Greeno (1997) argues that the assumptions that Anderson et al. criticize are actually wrong, but that they are not the claims of the situativity approach. The core of a situativity theory is a different perspective on phenomena of learning. Instead of focusing on mental processes and structures, the situativity approach analyses '… the social and ecological interaction as its basis and builds toward a more comprehensive theory by … analyses of information structures in the contents of people's interactions' (Greeno 1997, p. 5).

(b) *Triviality*: It is also argued that there is nothing really new in the core arguments of the situatedness theories. For example, Vera and Simon (1993) argue that all findings of the situatedness research can be incorporated into the well-elaborated traditional framework of cognitive models (i.e., symbolic paradigm). From the situatedness perspective, however, analyzing symbolic structures and processes runs too short. They may just be a special case of activity (Greeno and Moore 1993). Another triviality argument is that many claims of the situatedness protagonists were already articulated long before, for example by Dewey, Piaget, and Vygotskji, so that they are hardly stating anything new (e.g., Klauer 1999). Renkl (2000) counters that it is to the merit of the situatedness protagonists that they have reactivated these classical ideas. Most of these ideas played only a very minor role in educational mainstream research before the situatedness approach emerged. Furthermore, situated learning approaches bring classical ideas together with new developments (e.g., learning with new technologies) to form new 'Gestalts.'

(c) *Weak methodology*: Klauer (1999) is one of many researchers who criticize that the situatedness protagonists employ purely qualitative research methods and that they often rely merely on anecdotes to support their claims of the situatedness of cognition and learning (e.g., the 'cottage cheese story'; cf. Lave 1988). Lave (1988), on the other hand, rejects the empirical-experimental paradigm as artificially decontextualizing phenomena in laboratory investigations. From a situatedness point of view, it is clear that results from laboratories are of questionable value for out-of-lab contexts. It is important to note that many researchers who merely assimilated the notion of situated learning, but do not subscribe to radical situatedness, keep on researching within the traditional empirical framework. Hence, the methodological critique aims at the more radical situatedness proponents.

6. Possible Futures of Situated Learning

On the one hand, at least some situatedness protagonists in the area of education make strong statements with respect to the advantage of their approach over the traditional one. On the other hand, the 'traditionalists' defend themselves. Against this background it is interesting to ask what the future will bring. Four main possibilities are discussed. (a) Critical traditionalists (Klauer 1999) argue that the situatedness approach and the discussion around it will disappear as other more or less fruitless debates (e.g., the person–situation debate in psychology) have done

before. (b) Others (e.g., Cobb and Bowers 1999) hope that the situatedness perspective will take the place of the cognitive paradigm in education, just as years ago the behavioral paradigm was driven out by the cognitive one. (c) Some 'observers' of the situatedness debate (e.g., Sfard 1998) argue that the two positions provide different metaphors for analyzing learning and that both are useful. Accordingly, they plea for a complementary coexistence. (d) Greeno and MMAP (1998) envision a very ambitious goal for their situativity approach; that is, to develop a situatedness approach that is a synthesis of the cognitive and the behavioral paradigm. Whatever possibility may become reality, there is at least consensus between the main opponents in the situatedness debate (cf. Anderson et al. 1997, Greeno 1997) on what the touchstone for the two approaches should be: the ability to improve education.

See also: Capitalism: Global; Piaget's Theory of Human Development and Education; School Learning for Transfer; Situated Cognition: Contemporary Developments; Situated Cognition: Origins

Bibliography

Anderson J R, Reder L M, Simon H A 1996 Situated learning and education. *Educational Researcher* **25**: 5–11

Anderson J R, Reder L M, Simon H A 1997 Situated versus cognitive perspectives: Form versus substance. *Educational Researcher* **26**: 18–21

Cobb P, Bowers J 1999 Cognitive and situated perspectives in theory and practice. *Educational Researcher* **28**: 4–15

Collins A, Brown J S, Newman S E 1989 Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In: Resnick L B (ed.) *Knowing, Learning, and Instruction. Essays in the Honor of Robert Glaser*. Erlbaum, Hillsdale, NJ, pp. 453–94

Greeno J 1997 On claims that answer the wrong questions. *Educational Researcher* **26**: 5–17

Greeno J G, Middle School Mathematics Through Applications Project Group 1998 The situativity of knowing, learning, and research. *American Psychologist* **53**: 5–26

Greeno J G, Moore J L 1993 Situativity and symbols: Response to Vera and Simon. *Cognitive Science* **17**: 49–59

Greeno J G, Smith D R, Moore J L 1993 Transfer of situated learning. In: Dettermann D K, Sternberg R J (eds.) *Transfer on Trial: Intelligence, Cognition, and Instruction*. Ablex, Norwood, NJ, pp. 99–167

Klauer K J 1999 Situated learning: Paradigmenwechsel oder alter Wein in neuen Schläuchen? *Zeitschrift für Pädagogische Psychologie* **13**: 117–21

Lave J 1988 *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life*. Cambridge University Press, Cambridge, UK

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK

Renkl A 2000 Weder Paradigmenwechsel noch alter Wein!—Eine Antwort auf Klauers 'Situated Learning: Paradigmenwechsel oder alter Wein in neuen Schläuchen?' *Zeitschrift für Pädagogische Psychologie* **14**: 5–7

Renkl A, Mandl H, Gruber H 1996 Inert knowledge: Analyses and remedies. *Educational Psychologist* **31**: 115–21

Resnick L B 1987 Learning in school and out. *Educational Researcher* **16**: 13–20

Sfard A 1998 On two metaphors for learning and the dangers of choosing just one. *Educational Researcher* **27**: 4–13

Vera A H, Simon H A 1993 Situated action: A symbolic interpretation. *Cognitive Science* **17**: 7–48

A. Renkl

Situation Model: Psychological

> For, it being once furnished with simple ideas, it [the mind] can put them together in several compositions, and so make variety of complex ideas, without examining whether they exist so together in nature (John Locke 1690 *An Essay Concerning Human Understanding*).

When we read a story, we combine the ideas derived from understanding words, clauses, and sentences into mental representations of events, people, objects, and their relations. These representations are called situation models. Thus, situation models are not representations of the text itself; rather, they could be viewed as mental microworlds. Constructing these microworlds is the essential feature of understanding. At the most basic level, situation models are mental representations of events. Aspects of events that are encoded in situation models are: what their nature is, where, when, and how they occur, and who and what is involved in them. Single-event models are integrated with models of related events constructed based on the preceding text, such that situation models may evolve into complex integrated representations of large numbers of related events. This is what occurs when we comprehend extended discourse, such as news articles, novels, or historical documents.

1. Background

Situation models were introduced to cognitive psychology by van Dijk and Kintsch (1983) and are based on earlier research in formal semantics and logic (e.g., Kripke 1963). The concept is primarily used in research on language and discourse comprehension. For a good understanding, it is important to distinguish situation models from similar concepts, such as mental models, scripts, and frames.

2. How Situation Models Differ

2.1 Mental Models

Originally proposed by Craik (1943) and elaborated and introduced to cognitive psychology by Johnson-Laird (1983), mental models (see *Mental Models,*

Psychology of) are mental representations of real, hypothetical, or imaginary situations. Situation models can be viewed as a special type of mental model. Situation models are mental models of specific events. They are bound in time and space, whereas mental models in general are not. For example, heart surgeons have mental models of our blood circulatory system, but they construct a situation model of the state of patient X's coronary artery at time (t).

2.2 Scripts and Frames

Originally proposed in the artificial intelligence literature by Schank and Abelson (1977) and Minsky (1975), respectively, scripts and frames (see *Schemas, Frames, and Scripts in Cognitive Psychology*) are representations of stereotypical sequences of events, such as going to a restaurant and spatial layouts, such as that of a living room or the interior of a church. Comprehenders use scripts and frames to construct situation models. For example, the restaurant script can be used to construct a mental representation of your friend's visit to a local Italian restaurant last night. This is accomplished by filling in the slots of the script. Thus, scripts and frames can be considered types, whereas situation models are tokens. Furthermore, scripts and frames are semantic memory representations, whereas situation models are episodic memory representations.

3. Other Representations Constructed During Comprehension

It is often assumed that readers construct multilevel mental representations during text comprehension. Although there is no complete consensus as to what other types of mental representations, besides situation models, are constructed during text comprehension, empirical and intuitive evidence for the role of the following representations has been put forth. The surface structure is a mental representation of the actual wording of the text. Surface–structure representations are typically short-lived in memory, except when they have a certain pragmatic relevance (for example in the case of insults or jokes) or when the surface structure is constrained by prosodic features, such as rhyme and meter, as in some poetry.

The textbase is a mental representation of the semantic meaning of what was explicitly stated in the text. The textbase usually decays rather rapidly. That is, within several days, comprehenders are unable to distinguish between what they read and what they inferred (see *Memory for Meaning and Surface Memory*). Some researchers view the textbase simply as that part of the situation model that was explicitly stated (rather than implied) and thus deny the textbase a special status. Analyses of naturalistic discourse suggest that comprehenders not only construct a model of the denoted situation, but also construct a model of the communicative context. For example, they make inferences about the attitudes of writers regarding the situation they describe. Van Dijk (1999) calls this type of representation 'context model' and argues that no account of discourse comprehension is complete without the inclusion of a context model.

4. Why Situation Models are Needed to Explain Language Use

A task analysis of text comprehension shows the need for situation models. For example, when we comprehend the instructions that come with a household appliance, we form mental representations of the actions needed to operate or repair the device. It would be of little use to construct a mental representation of the wording of instructions themselves. Similarly, when we read newspaper articles, our usual goal is to learn and be updated about some news event. For example, we want to know how the United States House of Representatives responded to the latest budget proposal by the President or why the latest peace negotiations in the Middle East came to a halt. In these cases, we construct mental representations of agents, events, goals, plans, and outcomes, rather than merely mental representations of clauses and words.

5. Components of Situation Models

Situation models are models of events. Events always occur at a certain time and place. In addition, events typically involve participants (agents and patients) and objects. Furthermore, events often entertain causal relations with other events or are part of a goal-plan structure. Thus, time, place, participants, objects, causes and effects, and goals and plans are components of situations, with time and place being obligatory. As linguists have observed, most of these components are routinely encoded in simple clauses. Verbs typically describe events (although some events can also be described by nouns, e.g., explosion), while nouns and pronouns denote participants and objects (although pronouns can also denote events) and prepositions primarily denote spatial relations (although spatial relations can also be denoted by, for instance, verbs, as in 'The nightstand supported a lamp' and prepositions can be used to indicate temporal relationships, as in 'In an hour'). Temporal information can be expressed lexically in a variety of ways, but is also encoded grammatically in the form of verb tense and aspect in languages such as English, German, and French.

Causation and intentionality are denoted lexically by verbs (e.g., 'caused,' 'decided to') or adverbs (e.g., 'therefore,' 'in order to'), but are often left to be inferred by the comprehender. For example, there is no explicitly stated causal connection between following two events 'John dropped a banana peel on the floor. The waiter slipped,' yet comprehenders can easily infer the connection. Participants and objects may have various kinds of properties, such as having blue eyes or a cylindrical shape, and temporary features, such as being sunburnt or overheated. Finally, participants and objects may be related in various ways (e.g., kinship, professional, ownership, part–whole, beginning–end, and so on).

6. *A Simple Example*

A simple example illustrates these points. Consider the following clause:

(1) John threw the bottle against the wall.

This clause describes an event (throw) that involves a male agent (John), and two objects (bottle, wall). The simple past tense indicates that the event occurred prior to the moment at which the sentence was uttered (e.g., as opposed to *John will throw the bottle against the wall*). John's location is not explicitly stated, but we infer it is in relative proximity to the wall. We may draw inferences as to the causal antecedent and consequent of this event. A plausible antecedent is that John was angry. A plausible consequent is that the bottle will break.

Comprehenders are more likely to infer causal antecedents than causal consequences. Note that there are many other inferences that could be drawn. For example, one might infer that John is a middle-aged man and that the bottle was a wine bottle and that it was green and half-filled. One might also infer that the wall was a brick wall, and so on. Comprehenders typically make relatively few of this kind of elaborative inferences. Rather, they are focused on the causal chain of events (Graesser et al. 1994).

7. *Situation Models Establish Coherence*

A major role of situation models in discourse comprehension is to help establish coherence. Suppose sentence (1) were followed in the text by (2):

(2) It shattered into a thousand pieces.

How would this sentence be comprehended? If the two sentences are part of a discourse, rather than read in isolation, they would have to be integrated in some fashion. They are about some set of events. Comprehenders assume by default that a new sentence will describe the next event in the chronological sequence. Thus, they will assume that the event described in (2) directly follows that in (1). This implies temporal contiguity between (1) and (2). The use of the simple past tense is consistent with this assumption. Event 2 occurs prior to the moment of utterance, but after event 1. Given that no passage of time was mentioned, temporal contiguity is assumed.

The pronoun *it* in sentence (2) is taken to refer to some entity already in the situation model. There are three potential referents, John, the bottle, and the wall. John is not appropriate, because animate male human beings require the pronoun *he*. However, the bottle and the wall are both inanimate and thus compatible with the pronoun. Because there is no linguistic way to select the proper referent, the comprehender will use background knowledge that bottles are typically made out of glass, which is breakable, and walls out of harder materials, such as brick or concrete, to infer that it was the bottle that broke. The absence of a time lapse and the presence of an object from the previous event will cause the comprehender to assume that the second event takes place in roughly the same spatial region as the first, given that the same object cannot be at two different places at the same time.

Thus, the two events can now be integrated into a single situation model in which an agent, a male individual named John, at t1 threw a breakable object, a bottle, against a wall, presumably out of anger, which immediately, at t2, caused the bottle to break into pieces. This is how situation models provide coherence in discourse. Discourse is more than a sequence of sentences. Rather, it is a coherent description of a sequence of events that are related on several dimensions. Situation model theory attempts to account for how these events are successively integrated into a coherent mental representation.

8. *Situation Models Integrate Textual Information with Background Knowledge*

Another function of situation models is that they allow for the integration of text-derived information with the comprehender's background knowledge. Consider the following examples (from Sanford and Garrod 1998):

(3) Harry put the wallpaper on the table. Then he put his mug of coffee on the paper.

It is rather straightforward to integrate these sentences. They call for a spatial arrangement in which the paper is on top of the table and the mug on top of the paper and the coffee inside the mug. However,

consider the following sentence pair, which differs from (3) by only one word:

(4) Harry put the wallpaper on the wall. Then he put his mug of coffee on the paper.

Many readers will have difficulty integrating these sentences, because this discourse snippet describes an impossible set of circumstances. Realizing this impossibility relies critically on the activation of background knowledge. In this case the knowledge that putting wallpaper on a wall produces a vertical surface which would not support a mug of coffee. Thus, even though (or, rather, because) there are linguistic cues in the second sentence affording integration with the first—the pronoun can be taken to refer to Harry, and paper to wallpaper—this integration of text-derived information does not produce the correct understanding that the described situation is impossible. It is necessary to activate the requisite background knowledge to make this determination. By extension, the requisite background knowledge will help the comprehender construct an adequate representation in (3).

9. Other Functions of Situation Models

Situation models are needed in translation. Word-for-word translations yield nonsense in most cases. In order to arrive at a proper translation, one has to construct a situation model based on the source language and then convey this situation model in the target language.

Situation models are needed to explain learning from text. When we read a newspaper article about a current event, for example a war, we update our situation model of this event. We learn what the status of the actors and objects and their relations is at a new time. We would not learn by simply storing a mental representation of the text itself in long-term memory. In fact, what we know about a current and historical political situation, is usually an amalgam of information obtained from various sources (TV news, newspapers, magazines, encyclopedias, conversations with friends, and so on).

10. The Representational Format of Situation Models

Situation models often have an almost perceptual quality. In the example about the bottle, the first thing we construct is an agent, next is his action, next is the instrument of the action, and subsequently, we see the consequence of the action. There is a debate regarding the perceptual nature of situation models. Traditionally, an amodal propositional format has been proposed for situation models (van Dijk and Kintsch 1983, Kintsch 1998). However, others have proposed perceptual symbols (Barsalou 1999, Glenberg 1997, Johnson-Laird 1983).

In an amodal propositional representation, there is no analog correspondence between the mental representation and its referent. In a perceptual–symbol representation, there is. As a result, perceptual symbols represent more of the perceptual qualities of the referent than do amodal symbols, which are an arbitrary code. One of the major challenges of situation-model research will be to gather empirical evidence that speaks to the distinction between amodal and perceptual symbol systems as the representational format for situation models.

Both amodal and perceptual symbol systems allow for the possibility that information constructed from a text and activated background knowledge from long-term memory have the same representational format, such that they can be integrated quite easily. This is an important quality, because the integration of text-derived information with background knowledge is an essential feature of comprehension.

11. Empirical Evidence

There is a wealth of empirical evidence supporting the notion of a situation model (see Zwaan and Radvansky 1998 for a review). Most of this evidence consists of reaction times collected while people are reading texts. In addition, there is an increasing amount of electrophysiologal evidence (see, for example, Münte et al. 1998). Analyses of reading times, as well as electrophysiological measures suggest that comprehenders have difficulty integrating a new event into the situation model when that event has different situational parameters from the previously described event, for instance when it occurs in a different time frame or involves a new participant. Evidence suggests that ease of integration of an event depends on its relatedness to the evolving situation model on each of the five situational dimensions.

Probe-recognition studies show that the activation levels of concept decrease after a change in situational parameters (e.g., a shift in time, space, participant, or goal structure). For example, people recognize the word 'checked' more quickly after having read 'He checked his watch. A moment later, the doorbell rang' than after 'He checked his watch. An hour later, the doorbell rang.' Thus, the 'here and now' of the narrated situation tends to be more accessible to comprehenders than other information. This, of course, mimics our everyday interaction with the world.

Analyses of memory-retrieval data shows that people tend to store events together in long-term memory based on the time and location at which they occur, the participants they involve, whether they are causally related and whether or not they are part of the same goal-plan structure. There is evidence that

memory retrieval is influenced by a combination of the link strengths between events on the five situational dimensions. It has furthermore been shown that people's long-term memory for text typically reflects the situation that was described; memory representations for the discourse itself are much less resistant to decay or interference.

12. Computational Models

Kintsch (1998) has developed a computational model of text comprehension, the construction-integration model (see also *Text Comprehension: Models in Psychology*). In this model, a network consisting of nodes representing the surface structure, the textbase, and the situation models is successively constructed and integrated in a sequence of cycles. The construction of the network is done by hand. The nodes are propositions (or surface structure elements) and the links between them can be based on a variety of criteria, such as argument overlap, causal relatedness, or other aspects of situational relatedness. Integration occurs computationally by way of a constraint-satisfaction mechanism. Simulations using the construction–integration model have been most successful in (quantitatively) predicting text recall. In addition, the model has been shown to provide qualitative fit with a variety of findings in text comprehension, such as anaphoric resolution and sentence recognition. Similar models have been proposed by other researchers. Like the construction-integration model, these models include aspects of situation-model construction. However, there currently exists no full-fledged computational model of situation-model construction.

13. Beyond Language

Situation models have significance beyond the domain of text comprehension. Researchers are beginning to apply this concept to other domains of cognitive psychology, such as the comprehension of visual media (e.g., movies) and autobiographical memory. In the first case, situation models are acquired vicariously, as in language comprehension, but, in part, on nonlinguistic visual and auditory information. In the second case, they are acquired via direct experience. The question of whether and how the mode of acquisition affects the nature of situation models is a fruitful one.

See also: Concept Learning and Representation: Models; Figurative Thought and Figurative Language, Cognitive Psychology of; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Language and Thought: The Modern Whorfian Hypothesis; Literary Texts: Comprehension and Memory; Mental Models, Psychology of; Mental Representations, Psychology of; Narrative Comprehension, Psychology of; Reasoning with Mental Models; Sentence Comprehension, Psychology of

Bibliography

Barsalou L W 1999 Perceptual symbol systems. *Behavioral and Brain Sciences* **22**: 577–660

Craik K 1943 *The Nature of Explanation*. Cambridge University Press, Cambridge, UK

Glenberg A M 1997 What memory is for. *Behavioral and Brain Sciences* **20**: 1–19

Graesser A C, Singer M, Trabasso T 1994 Constructing inferences during narrative text comprehension. *Psychological Review* **101**: 371–95

Johnson-Laird P N 1983 *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge, MA

Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, MA

Kripke S 1963 Semantical considerations on modal logics. *Acta Philosophica Fennica* **16**: 83–94

Minsky M 1975 A framework for representing knowledge. In: Winston P H (ed.), *The Psychology of Computer Vision*. McGraw-Hill, New York, pp. 211–77

Münte T F, Schiltz K, Kutas M 1998 When temporal terms belie conceptual order. *Nature* **395**: 71–3

Sanford A J, Garrod S C 1998 The role of scenario mapping in text comprehension. *Discourse Processes* **26**: 159–90

Schank R C, Abelson R 1977 *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Erlbaum, Hillsdale, NJ

van Dijk T A 1999 Context models in discourse processing. In: van Oostendorp H, Goldman S R (eds.), *The Construction of Mental Representations During Reading*. Erlbaum, Mahwah, NJ, pp. 123–48

van Dijk T A, Kintsch W 1983 *Strategies in Discourse Comprehension*. Academic, New York

Zwaan R A, Radvansky G A 1998 Situation models in language comprehension and memory. *Psychological Bulletin* **123**: 162–85

R. A. Zwaan

Skinner, Burrhus Frederick (1904–90)

After J. B. Watson, the founder of the behaviorist movement, Skinner has been the most influential, and also the most controversial figure of behaviorism. His contributions to behavioral sciences are manifold: he designed original laboratory techniques for the study of animal and human behavior, which he put to work to produce new empirical data in the field of learning, and to develop a theory of operant behavior that led him eventually to a general psychological theory; he further elaborated the behaviorist approach, both refining and extending it, in a version of behavioral science which he labeled radical behaviorism; he

formulated seminal proposals for applied fields such as education and therapy and pioneered in machine assisted learning; finally, building upon his conception of the causation of behavior, he ventured into social philosophy, questioning the traditional view of human nature and of the relation of humans to their physical and social environment. This part of his work has been the main source of sometimes violent controversy.

1. Biographical Landmarks

Skinner was born on March 20, 1904 in Susquehanna (Pennsylvania, USA) in a middle-class family and experienced the usual childhood and adolescence of provincial American life. He attended Hamilton College, which was not a particularly stimulant institution to him. He was first attracted by a literary career, which he soon gave up, after traveling to Europe. He turned to psychology, and was admitted at Harvard in 1928. He obtained his Ph.D. in 1931, with a theoretical thesis on the concept of reflex—a first landmark in his reflections on the causation of behavior, an issue he was to pursue throughout his scientific career. He stayed at Harvard five more years, beneficiary of an enviable fellowship, affiliated to the physiology laboratory headed by Crozier. In 1936, he was appointed professor at the University of Minnesota where he developed his conditioning chamber for the study of operant behavior in animals—which was to be known as the Skinner box—and wrote his first book *The Behavior of Organisms* (1938). In 1945, he moved to Indiana University, as Chairman of the Department of Psychology. In 1948, he was offered the prestigious Edgar Pierce Professorship in Psychology at Harvard University, where he was to stay until his death on August 18, 1990.

Skinner received during his lifetime the highest national awards an American psychologist could receive, and he was praised as one of the most prominent psychologists of the century, in spite of harsh attacks against some of his ideas, from the most opposite sides of scientific and lay-people circles.

Although most of Skinner's laboratory research was carried out with animal subjects, mainly rats and pigeons, using the so-called operant procedure that will be described hereafter, essentially he was interested, not in animal behavior proper, but in behavior at large, and more specifically in human behavior. After the tradition of other experimental psychologists before him, such as Pavlov and Thorndike, and of most of his colleagues in the behaviorist school of thought, such as Hull, Tolman, or Guthrie, he resorted to animals as more accessible subjects than humans for basic studies on behavior, just as physiologists used to do quite successfully. The relevance of extrapolating from animals to humans is of course an important issue in psychology. However, Skinner's main concern was obviously with humans, as evidenced by his literary writings in the field of social philosophy—namely the utopian novel *Walden Two* (1948) and the essay, *Beyond Freedom and Dignity* (1971)—as well as by his theoretical endeavors to account for human behavior (*Science and Human Behavior* 1953) and for verbal behavior (*Verbal Behavior* 1957).

2. Operant Conditioning and the Skinner Box

The operant conditioning chamber, often called the Skinner box, is a laboratory device derived from Thorndike's puzzle box and from the mazes familiar to students of learning in rats by the time Skinner started his career. In its most common form, it consists of a closed space in which the animal moves freely; it is equipped with some object that the subject can manipulate easily—be it a lever for rats, or a small illuminated disk upon which pigeons can peck—and with a food dispenser for delivering calibrated quantities of food. By exploring spontaneously this particular environment, eventually with the help of the experimenter in shaping progressively its behavior, the subject will eventually discover the basic relation between a defined response—pressing the lever or pecking the key—and the presentation of a reinforcing stimulus—a small food reward. The basic relation here is between an operant response (i.e., a response instrumental to produce some subsequent event) and its consequence (i.e., the reinforcement), rather than between a stimulus and a response elicited by it, as in Pavlovian or respondent conditioning. This simple situation may be made more complex either by introducing so-called discriminative stimuli, the function of which is not to trigger the response at the manner of a reflex, but to set additional conditions under which the response will be reinforced, or by changing the basic one response-one reinforcement link to some more complicated contingencies, for instance requiring a given number of responses to one reinforcement, or the passing of some defined delay. A wide variety of schedules of reinforcement have been so studied, be it for their own sake as sources of information on the lawfulness of behavior (for instance modern research has applied optimization models borrowed from economics to the study of operant behavior), or as efficient tools for other purposes (such as the analysis of sensory functions in animal psychophysics, of the effects of drugs acting upon the Central Nervous System in experimental psychopharmacology, or of cognitive capacities).

The operant technique presented two important features by the time Skinner developed it from the thirties to the fifties. It emphasized the study of individual subjects through a long period of time rather than groups of subjects for a few sessions, as

used to be the case in maze studies and the like. This interest in individual behavior would favor later applications to human subjects in educational and clinical settings. Second, the operations involved soon were automatized by resorting to electromechanical circuits, to be replaced later by online computer control. This led to a level of efficiency and precision unprecedented in psychological research.

3. *The Evolutionary Analogy*

Skinner captured the essence of operant behavior in the formula 'control of behavior by its consequences,' and very early he pointed to the analogy between the selection of the response by the subsequent event and the mechanism at work in biological evolution. An increasingly large part of his theoretical contributions were eventually devoted to elaborating the evolutionary analogy (Skinner 1987). The generalization of the selectionist model to behavior acquisition at the individual level, initially little more than a metaphoric figure, has recently gained credentials with the theses of neurobiologists, such as Changeux's *Generalised Darwinism* (1983) or Edelman's *Neural Darwinism* (1987), who both have substantiated in ontogeny selective processes previously reserved to phylogeny. One of the main tenets of Skinner's theory converges with contemporary views in neurosciences.

Skinner extended the selectionist explanation to cultural practices and achievements, joining some schools of thought in cultural anthropology and in the history of science, such as Karl Popper's selectionist account of scientific hypotheses.

4. *Radical Behaviorism*

As a behaviorist, Skinner viewed psychology as a branch of natural sciences, more explicitly as a branch of biology, which can deal with its subject matter using the same principles as in other fields of life sciences, be it with specific implementations as required by its particular level of analysis.

Skinner defended a brand of behaviorism quite distinct from the dominant view that prevailed in the second quarter of the century. His radical behaviorism was opposed to methodological behaviorism. For most psychologists, defining their science as the science of behavior, after Watson's recommendation, did not really mean that they had abandoned mental life as the main objective of their inquiry; rather, they had simply resigned themselves to study behavior, because they had to admit that they had no direct access to mental life. Such methodological behaviorism, in fact, remained basically dualistic. In contrast, radical behaviorism is definitely monist, and it rejects any distinction between what is called mental and behavioral. Skinner is, in this respect, closer to Watson's view than to the position of other influent neobehaviorists of his generation, although he developed a far more sophisticated view of behavior than Watson's. For instance, he rejected the simplistic claim that thought is nothing more than subvocal language, and admitted that not all behavior is directly observable. Part of human behavior is obviously private, or covert, and raises difficult methodological problems of accessibility; but this is no reason to give it different status in a scientific analysis.

Skinner vigorously denounced mentalism not so much because it refers to events that would occur in another space and be of a different substance than behavior, but because it offers pseudo-explanations which give the illusion of understanding what in fact remains to be accounted for. Mentalistic explanations, very common in daily life psychology, were still quite frequent in scientific psychology. For example, in the field of motivation, all sorts of behavior were assigned internal needs as causal agents: not only do we eat *because* we are hungry—a simple statement which, at any rate from a scientist's point of view, requires qualification—but we exhibit aggressive behavior because of some aggression drive, we interact with social partners because of a need for affiliation, we work to successful and creative outcomes because of a need for achievement, etc. For Skinner, such explanations dispense us from looking for the variables really responsible for the behavior; these variables are more often than not to be found in the environment, and can be traced to the history of the individuals interacting with their environment, social and physical.

Combining this epistemological conception and the results of his empirical research, Skinner developed a theory of human behavior emphasizing the determining role of environmental contingencies on human actions. However, along the lines of the evolutionary analogy, he suggested a model that would account equally well for novelty and creative behavior, as exhibited in art and science productions, and for stabilized, persistent habits adapted to unchanging conditions.

Skinner devoted special attention to verbal behavior, the importance of which in the human species justified special treatment. He presented his book *Verbal Behavior* as 'an essay in interpretation,' proposing a functional analysis of the verbal exchanges composing the global episode of speaker-listener communication. In spite of the violent criticisms expressed by the linguist Chomsky (1959), Skinner's analysis foreshadowed some aspects of the pragmatic approach adopted some years later by many psycholinguists, aware of the insufficiency of formal grammars to account for central features characterizing the use of language. He viewed verbal behavior as shaped by the linguistic community, and exerting a genuine type of control over an individual's behavior, distinct from the action of the physical environment. A large part of human behavior is, in his terms, rule governed,

that is, controlled by words, rather than shaped by contingencies, that is, through direct exposure to the physical world. Many behaviors can have one or the other origin: avoidance of fire flame can derive from direct experience with fire, or from warnings received during education. Once endowed with verbal behavior, individuals can use it to describe and anticipate their own behavior, or to develop it in its own right, as in literary composition. Were it not for the unfortunate use of the term rule, which makes for confusion with the word as used in formal linguistics and with the notion of coercive control, what Skinner was pointing to was akin to a distinction now familiar to contemporary psychology and neurosciences between top-down vs. bottom-up causation.

5. *Education*

Skinner's interests in applications covered three main areas: education, psychological treatment, and social practices at large. He dealt with the first two in a technical manner, developing principles and techniques for improving educational and therapeutic practices. His treatment of the third one is more akin to social philosophy than to scientific application, although he viewed it as very consistently rooted in his scientific thinking. At a time school education in the US was criticized for its deficiencies and for its inefficiency in competing with technological achievements of the Soviet Union, Skinner, as many other American scientists, inquired into the reasons of that state of affairs. Observing what was going on in any normal classroom, including in reputed schools, he concluded that it was violating blatantly most basic principles of learning derived from laboratory analysis of the learning process. Pupils and students were passively exposed to teachers' monologues rather than actively producing behaviors followed by feedback; there was no attempt to adjust the teacher's actions to individual level and rhythm of learning; negative evaluation based on mistakes and errors prevailed over positive evaluation pointing to progresses achieved; punitive controls, known to be poorly effective in shaping and maintaining complex behavior, was still widely used; general conditions and teaching practices were far from favorable to developing individual talents and creativity. Such criticisms had been made by others, but Skinner differed from them in the analysis of the causes and in the remedies proposed. He did not question the importance of endowing the students with basic knowledge and skills that they need if they are to engage in more complex and original activities. But he thought such skills could be mastered using more efficient methods than those currently in use in the classroom. This was the origin of teaching machines, a term that would raise strong objections on the ground that machines could not lead but to dehumanising teaching. In fact, Skinner's idea has been implemented since then in computer assisted learning, which is now widely accepted—with little reference to the pioneering projects of the behaviorist psychologist. The device he designed in the 1950s appears quite primitive compared with modern computers: it was an electromechanical machine—adapted from a record player—built in such a way that it would present in a window to the student successive small frames of the material to be learned, each frame requiring an active answer from the learner. The latter could learn at his or her own individual rhythm, ideally with no or few errors, and eventually reach the end of the program, with the guarantee that the subject matter had been mastered from end to end. Good programs would make exams useless, if exams are just a way to control that the material has been covered and understood. Most important, student and teachers would not waste the few hours they could work together in tasks easily fulfilled using teaching devices; they could devote the time so spared to more constructive activities requiring direct human contact.

In spite of numerous attacks in educational circles, Skinner's project inspired many applications, such as programmed instruction in book form, until modern computers would offer the elegant solution we know today. It also contributed to the development of individualized teaching approaches that favor methods allowing students to learn at their own pace in an autonomous and active way.

Teaching machines are but one aspect of Skinner's contribution to education. In a number of his writings, including his utopian novel, Skinner (1968, 1978) did express his reflections on educational issues. He was concerned especially with the disproportion between the resources devoted to education and the poor outcomes; with the tendency to level down individual differences; with the increasing distance between the school environment and real life; with violence in schools, and other matters which remain crucial issues today, with little improvement.

6. *Behavior Therapy*

Treatment of psychological disturbances was another field of application to which Skinner (1955, 1989) made influent contributions. By the middle of the twentieth century, psychopathology had elaborated refined descriptive systems of psychological disturbances and equally sophisticated explanatory models such as the psychoanalytic theory. Contrasting with these, methods for treatment were scarce and their results poor. Psychoanalytic treatment had limited indications and practical limitations. Rogers's non-directive therapy, though quite popular, did not provide convincing results. Psychopharmacology was still in the air.

Skinner did not question the classical categorization of mental illnesses, nor did he propose miracle remedies. He simply suggested to look at them as disturbances in behavior, rather than alterations of hypothetical mental structures, such as the psychic apparatus appealed to by psychoanalysis, of which abnormal behavior would be but observable indicators, or symptoms. Consequently, he proposed to attempt to change undesirable behavior by acting directly upon it, rather than upon underlying structures supposedly responsible for it. This approach was not totally new: behavior therapy had its origins in John Broadus Watson's attempts to treat fear in children by resorting to Pavlovian conditioning and in the theoretical work of some neobehaviorists aimed at transposing some psychoanalytical concepts into learning theory models. What Skinner added was his genuine theoretical elaboration, especially based on his antimentalist stand, and techniques of behavior modification drawn from the operant laboratory, supplementing Pavlovian techniques in use up to then. He also brought into the clinical field a sense of rigor transferred from the laboratory, perfectly compatible with the study of single cases.

Skinner did not practice behavior therapy himself. His influence in the field was indirect, by stimulating pioneering research on psychotic patients and mentally deficient people. He gave a decisive impulse to the development of the behavioral approach to treatment, which soon acquired a major position in clinical psychology and counseling, and which eventually merged, somewhat paradoxically, with cognitively oriented practices into what are labeled behavioral-cognitive therapies.

7. *Social Philosophy*

Skinner's social philosophy was based in a deep confidence that only science would help us in solving the problems we are facing in our modern societies. What is needed is a science of behavior, which in fact is now available, he thought, and could be applied if only we would be ready to abandon traditional views of human nature. He first expressed his ideas in the novel *Walden Two* (1948). Written two years after the end of the World War Two, the book describes a utopian community run after the principles derived from the psychological laboratory. It is by no means a totalitarian society, as some critics have claimed. Looked at retrospectively, it is surprisingly premonitory of social issues that are still largely unsolved half a century later. For instance, working schedules at *Walden Two* have been arranged so that all tasks needed for production of goods and good functioning of the community are distributed among members according to a credit system which results in an average amount of 24 hours per week, avoiding unemployment, abolishing any social discrimination between manual and intellectual work, and leaving many free hours for leisure activities such as sports, arts, and scientific research. Emphasis is put on active practice rather than passive watching, on co-operation rather than competition. The community is not isolated culturally: cultural products from outside such as books or records are of course welcome, but radio programs are filtered to eliminate publicity. Education is active; the school building symbolically has no door separating it from the life and work community; there are no age classes, no humiliating ranking; all learn at their own rhythm, in whatever orientation they feel appropriate, throughout their life time. Women enjoy complete equality with men. Waste of natural resources is avoided. Special charges in the management of the community are strictly limited in time, eliminating any risk of political career.

Similar themes, plus the frightening concerns with pollution, violence, uncontrolled population growth, nuclear weapons, and the like, are further elaborated in the essay *Beyond Freedom and Dignity* (1971) and a number of later articles. In an alarming tone, Skinner points to what he feels is the core of our inability to deal with these issues, that is our obstinacy in keeping a conception of human nature which scientific inquiry shows us to be wrong, and which bar any solution to the problems we are confronted with. We still stick to a view of humans as being the center of the universe, free and autonomous, dominating nature, while we are but one among many elements of nature. As a species, we are the product of biological evolution; as cultural groups, the result of our history; and as individuals the outcome of our interactions with the environment. Because we fail to admit this dependency, and draw the consequences of it, we might put in danger our own future. Freedom, autonomy, and merit are no absolute values: they were forged throughout history, and more often than not they are used to disguise insidious controls, the mechanisms of which should be elucidated if we want to develop counter-controls eventually allowing for the survival of our species.

Skinner viewed these various facets of his work as closely related, making for a highly consistent theory of human behavior, in which the critical analysis of social processes in modern society was deeply rooted in the experimental analysis of the behavior of animal subjects in the laboratory. So global an ambition has been criticised, and clearly various aspects of his contribution did not have the same fate. If the operant technique is now a widely used procedure to many purposes in experimental research in psychology and related fields, if his early attempts to build teaching machines appear now as ancestors of computer assisted learning and teaching, if a number of principles of contingencies analysis are now currently put in practice in behavior therapies, radical behaviorism has been seriously questioned and even shaken by the rise of the cognitivist approach in psychology, while

Skinner's social philosophy has been attacked from different fronts both on ideological and scientific grounds.

As most great theory builders of the twentieth century in psychology, from Sigmund Freud and Watson to Jean Piaget, Skinner may be blamed for having reduced the explanation of human nature to a very limited set of concepts and findings, namely those he had forged and observed in his own restricted field of research and reflection, ignoring even other concepts and facts in neighboring fields of psychology, leaving alone of other sciences. It is clear that Skinner has made no attempt at integrating, for example, contributions of developmental or of social psychology, nor those of sociology, cultural anthropology, or linguistics. Such neglects might have been deliberate, legitimated by the will to concentrate on what were, in Skinner's mind, essential points left out by other branches of psychology or other sciences dealing with human societies. However, they might appear as sectarianism to those who favor an integrative and plurisdisciplinary approach to the complex objects of human sciences. It cannot be decided whether his influence would have been larger or smaller had he adopted a less exclusive stand.

See also: Autonomic Classical and Operant Conditioning; Behavior Therapy: Psychiatric Aspects; Behavior Therapy: Psychological Perspectives; Behaviorism; Behaviorism, History of; Conditioning and Habit Formation, Psychology of; Darwinism: Social; Educational Learning Theory; Educational Philosophy: Historical Perspectives; Evolutionary Epistemology; Freud, Sigmund (1856–1939); Lashley, Karl Spencer (1890–1958); Mental Health and Normality; Operant Conditioning and Clinical Psychology; Pavlov, Ivan Petrovich (1849–1936); Piaget, Jean (1896–1980); Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychology: Historical and Cultural Perspectives; Thorndike, Edward Lee (1874–1949); Utopias: Social; Watson, John Broadus (1878–1958)

Bibliography

Bjork D W 1997 *B. F. Skinner, A Life*. American Psychological Association, Washington, DC

Changeux J-P 1983 *L'Homme neuronal*. Fayard, Paris (*The Neuronal Man*)

Chomsky N 1959 Review of Skinner B F, Verbal behavior. *Language* **4**: 16–49

Edelman G M 1987 *Neural Darwinism: The Theory of Neuronal Group Selection*. Basic Books, New York

Modgil S, Modgil C (eds.) 1987 *B. F. Skinner: Consensus and Controversy*. Falmer, New York

Richelle M 1993 *B. F. Skinner, A Reappraisal*. Erlbaum, Hove, London

Roales-Nieto J G, Luciano Soriano M C, Pérez Alvarez M (eds.) 1992 *Vigencia de la Obra de Skinner*. Universidad Granada Press, Granada (*Robustness of Skinner's Work*)

Skinner B F 1938 *The Behavior of Organisms*. Appleton Century Crofts, New York

Skinner B F 1948 *Walden Two*. Macmillan, New York

Skinner B F 1953 *Science and Human Behavior*. Macmillan, New York

Skinner B F 1955 What is psychotic behavior. In: Gildea F (ed.) *Theory and Treatment of the Psychoses: Some Newer Aspects*. Washington University Studies, St Louis, MO, pp. 77–99

Skinner B F 1957 *Verbal Behavior*. Appleton Century Crofts, New York

Skinner B F 1961 *Cumulative Record*. Appleton Century Crofts, New York

Skinner B F 1968 *The Technology of Teaching*. Appleton Century Crofts, New York

Skinner B F 1971 *Beyond Freedom and Dignity*, 1st edn. Knopf, New York

Skinner B F 1978 *Reflections on Behaviorism and Society*. Prentice-Hall, Englewood Cliffs, NJ

Skinner B F 1987 *Upon Further Reflection*. Prentice Hall, Englewood Cliffs, NJ

Skinner B F 1989 *Recent Issues in the Analysis of Behavior*. Merrill, Columbus, OH

M. N. Richelle

Slavery as Social Institution

Slavery is the most extreme form of the relations of domination. It has existed, at some time, in most parts of the world and at all levels of social development. This article examines five aspects of the institution: Its distinguishing features; the means by which persons were enslaved; the means by which owners acquired them; the treatment and condition of slaves; and manumission, or the release from slavery.

1. The Distinctive Features of Slavery

The traditional, and still conventional, approach is to define slavery in legal–economic terms, typically as 'the status or condition of a person over whom any or all the powers attaching to the right of ownership are exercised' (League of Nations 1938, Vol. 6). In this view, the slave is, quintessentially, a human chattel. This definition is problematic because it describes adequately mainly Western and modern, capitalistic systems of slavery. In many nonWestern parts of the world, several categories of persons who were clearly not slaves, such as unmarried women, concubines, debt bondsmen, indentured servants, sometimes serfs, and occasionally children, were bought and sold. Conversely, in many slave-holding societies certain categories of slaves, such as those born in the household, were not treated as chattels.

Slavery is a relation of domination that is distinctive in three respects. First, the power of the master was usually total, if not in law, almost always in practice. Violence was the basis of this power. Even where laws forbade the gratuitous killing of slaves, it was rare for masters to be prosecuted for murdering them, due to the universally recognized right of masters to punish their slaves, and to severe constraints placed on slaves in giving evidence in courts of law against their masters, or free persons generally.

The totality of the master's claims and powers in them meant that slaves could have no claims or powers in other persons or things, except with the master's permission. A major consequence of this was that slaves had no custodial claims in their children; they were genealogical isolates, lacking all recognized rights of ancestry and descent. From this flowed the hereditary nature of their condition. Another distinctive consequence of the master's total power is the fact that slaves were often treated as their surrogates, and hence could perform functions for them as if they were legally present, a valuable trait in premodern societies with advanced commodity production and long-distance trading, such as ancient Rome, where laws of agency, though badly needed, were nonexistent or poorly developed.

Second, slaves were universally considered as outsiders, this being the major difference between them and serfs. They were natally alienated persons, deracinated in the act of their, or their ancestors', enslavement, who were held not to belong to the societies in which they lived, even if they were born there. They lacked all legal or recognized status as independent members of a community. In kin-based societies, this was expressed in their definition as kinless persons; in more advanced, state-based societies, they lacked all claims and rights of citizenship. Because they belonged only to their master, they could not belong to the community; because they were bonded only to their master's household, they could share no recognized bond of loyalty and love with the community at large. The most ancient words for slaves in the Indo-European and several other families of languages translate to mean, 'those who do not belong,' or 'not among the beloved,' in contrast with free members of the community, who were 'among the beloved,' and 'those who belonged.'

Third, slaves were everywhere considered to be dishonored persons. They had no honor that a nonslave person need respect. Masters could violate all aspects of their slaves' lives with impunity, including raping them. In most slave-holding societies, injuries against slaves by third parties were prosecuted, if at all, as injuries against the person and honor of the master. Where an honor-price or *wergild* existed, as in Anglo-Saxon Britain and other Germanic lands, its payment usually went to the master rather than to the injured slave. Universally, slavery was considered the most extreme form of degradation, so much so that the slave's very humanity was often in question.

For all these reasons, there was a general tendency to conceive of slaves symbolically as socially dead persons. Their social death was often represented symbolically in ritual signs and acts of debasement, death and mourning: In clothing, hairstyles, naming practices, and other rituals of obeisance and nonbeing.

2. *The Modes of Enslavement*

Free persons became slaves in one of eight ways: capture in warfare; kidnapping; through tribute and taxation; indebtedness; punishment for crimes; abandonment and sale of children; self-enslavement; and birth. Capture in warfare is generally considered to have been the most important means of acquiring slaves, but this was true mainly of simpler, small-scale societies, and of certain volatile periods among politically centralized groups. Among even moderately advanced premodern societies, the logistics of warfare often made captivity a cumbersome and costly means of enslaving free persons.

Kidnapping differed from captivity in warfare mainly in the fact that enslavement was its main or sole objective, and that it was usually a private act rather than the by-product of communal conflict. Other than birth, kidnapping in the forms of piracy and abduction was perhaps the main form of enslavement in the ancient Near East and the Mediterranean during Greek and Roman times; and this was true also of free persons who were enslaved in the transSaharan and transatlantic slave trades.

Debt bondage, which was common in ancient Greece up to the end of the seventh century BC, the ancient Near East, and in South East Asia down to the twentieth century, could sometimes descend into real slavery, although nearly all societies in which it was practiced distinguished between the two institutions in at least three respects: Debt-bondage was nonhereditary; bondsmen remained members of their community, however diminished; and they maintained certain basic rights, both in relation to the bondholder and to their spouses and children.

Punishment for crimes was a major means of enslavement in small, kin-based societies; China, and to a lesser extent, Korea, were the only advanced societies in which it remained the primary way of becoming a slave. Nonetheless, it persisted as a minor means of enslavement in all slaveholding societies, and became important historically in Europe as the antecedent of imprisonment for the punishment of crimes.

The enslavement of foundlings was common in all politically centralized premodern societies, though rarely found in small-scale slaveholding communities. It was the humane alternative to infanticide and was especially important in China, India, European antiquity, and medieval Europe. It has been argued that

it ranked second to birth as a source of slaves in ancient Rome from as early as the first century CE until the end of the Western empire.

Self-enslavement was rare and was often the consequence of extreme penury or catastrophic loss.

In nearly all slave-holding societies where the institution was of any significance, birth rapidly became the most important means by which persons became slaves, and by which slaves were acquired. Contrary to a common misconception, this was true even of slave societies in which the slave population did not reproduce itself naturally. The fact that births failed to compensate for deaths, or to meet the increased demand for slaves—which was true of most of the slave societies of the New World up to the last decades of the eighteenth century—does not mean that birth did not remain the main source of slaves.

However, the role of birth as a source of slaves was strongly mediated by the rules of status inheritance, which took account of the complications caused by mixed unions between slaves and free persons. There were four main rules.

(a) The child's status was determined by the mother's status only, regardless of the father's status. This was true of most modern and premodern Western slave societies, and of nearly all premodern non-Western groups with matrilineal rules of descent.

(b) Status was determined by the father only, regardless of the mother's status. This unusual pattern was found mainly among certain rigidly patrilineal groups, especially in Africa, where it was the practice among groups such as the Migiurtini Somali, the Margi of northern Nigeria, and among certain Ibo tribes. The practice, however, was not unknown in the West. It was the custom in Homeric Greece and was the norm during the seventeenth century in a few of the North American colonies such as Maryland and Virginia, and in South Africa and the French Antilles up to the 1680s.

(c) Status was determined by the principle of *deterior condicio*, that is, by the mother or father, whoever had the lower status. This was the harshest inheritance rule and was the practice in China from the period of the Han dynasties up to the reforms of the thirteenth and fourteenth centuries. It found its most extreme application in Korea, where it was the dominant mode down to 1731. The rule also applied in Visigothic Spain, and medieval and early modern Tuscany. The only known case in the New World was that of South Carolina in the early eighteenth century.

(d) The fourth principle of slave inheritance, that of *melior condicio*, is in direct contrast with the last mentioned, in that the child inherited the status of the free parent, whatever the gender of the parent, as long as the father acknowledged his paternity. This is the earliest known rule of slave inheritance and may have been the most widely distributed. It was the norm in the ancient Near East and, with the exception of the Tuareg, it became the practice among nearly all Islamic societies. The rule was supported among Muslims by another Koranic prescription and practice: The injunction that a slave woman was to be freed, along with her children, as soon as she bore a son for her master.

The only known cases in Europe of the *melior condicio* rule both emerged during the thirteenth century. In Sweden, it was codified in the laws of Ostergotland and Svealand as part of a general pattern of reforms. In Spain, religion was the decisive factor in the appearance of a modified version of the rule. Baptized children of a Saracen or Jewish owned slave and a Christian were immediately freed. Throughout Latin America, although the legal rule was of the first type—the children of slave women were to become slaves—the widespread custom of concubinage with slave women, coupled with the tendency to recognize and manumit the offspring of such unions, meant that, in practice, a modified version of the *melior condicio* rule prevailed where the free person was the father, which was usually the case with mixed unions.

3. *The Acquisition of Slaves*

Slaves not born to, or inherited by, an owner were acquired mainly through external and internal trading systems; as part of bride and dowry payments; as money; and as gifts. There were five major slave trading systems in world history. The Indian Ocean trade was the oldest, with records dating back to 1580 BC, and persisted down to the twentieth century AD. Slaves from subSaharan Africa were transported to the Middle and Near East as well as Southern Europe. It has been estimated that between the years 800 AD and 1800 approximately 3 million slaves were traded on this route, and over two million were traded during the nineteenth century.

The Black Sea and Mediterranean slave trade supplied slaves to the ancient European empires and flourished from the seventh century BC through the end of the Middle Ages. Over a quarter of a million slaves may have been traded in this system during the first century of our era. The European slave trade prospered from the early ninth century AD to the middle of the twelfth, and was dominated by the Vikings. One of the two main trading routes ran westward across the North Sea; the other ran eastward. Celtic peoples, especially the Irish, were raided and sold in Scandinavia. Most slaves traded on the eastern routes were of Slavic ancestry. It was the Viking raiding, and wide distribution, of Slavic slaves throughout Europe that accounts for the common linguistic root of the term 'slave' in all major European languages.

The transSaharan trade has persisted from the mid-seventh century AD down to the twentieth and involved the trading of subSaharan Africans throughout North Africa and parts of Europe. It has been

estimated that some 6.85 million persons were traded in this system up to the end of the nineteenth century. Although it declined substantially during the twentieth century, largely under European pressure, significant numbers of Africans are still being traded in Sudan and Mauritania.

The transatlantic slave trade was the largest in size and certainly the most extensive of all these systems. The most recent evidence suggests that between the years 1500 and 1870, some 11 million Africans were taken to the Americas. Of the 10.5 million who were forced from Africa between 1650 and 1870, 8.9 million survived the crossing. Although all the maritime West European nations engaged in this trade, the main traders were the British, Portuguese, and French. Four regions account for 80 percent of all slaves going to the New World: the Gold Coast (Ghana), the Bights of Benin and Biafra, and West-Central Africa. Forty percent of all slaves landed in Brazil; and 47 percent in the Caribbean. Although only 7 percent of all slaves who left Africa landed in North America, by 1810 the United States had, nonetheless, one of the largest slave populations due to its unusual success in the reproduction of its slave populations. For the entire three and a half centuries of the Atlantic slave trade, approximately 15 percent of those forced from Africa died on the Atlantic crossing (some 1.5 million) with losses averaging between 6000 and 8000 per year during the peak period between the years 1760 and 1810.

4. *The Treatment of Slaves*

It is difficult to generalize about the treatment of slaves, since this not only varied considerably between societies but also within them. There was no simple correlation of favorable factors. Furthermore, the same factor may operate in favor of slaves in one situation, but against them in the next. Thus, in many small, kin-based societies, slaves were relatively well treated and regarded as junior members of the master's family, but could nonetheless be sacrificed brutally on special occasions. In advanced premodern societies such as Greece and Rome, as well as modern slave societies such as Brazil, slaves in the mines or latifundia suffered horribly short lives, while skilled urban slaves were often virtually free and sometimes even pampered. In the Caribbean, the provision ground system, by which slaves supported themselves, led to high levels of malnutrition compared to the US South, where masters provided nearly all the slaves' provision. Nonetheless, Caribbean slaves cherished the provision ground system for the periods of self-determination and escape from the master's direct control that it offered.

In general, the most important factors influencing the condition of slaves were the uses to which they were put, their mode of acquisition, their location—whether urban or rural—absenteeism of owners, proximity to the master, and the personal characteristics of the slaves. Slaves were acquired for purely economic, prestige, political, administrative, sexual, and ritual purposes. Slaves who worked in mines or in gangs on highly organized farming systems were often far worse off than those who were employed in some skilled craft in urban regions, especially where the latter were allowed to hire themselves out independently. Slaves acquired for military or administrative purposes, as was often the case in the Islamic world—the Janissaries and Mameluks being the classic examples—were clearly at an advantage when compared with lowly field hands or concubines. Newly bought slaves, especially those who grew up as free persons and were new to their masters' society, usually led more wretched lives than those born to their owners. High levels of absenteeism among owners—which was true of the owners of slave latifundia in ancient Rome as well as the Caribbean slave societies and some parts of Latin America—often meant ill-usage by managers and overseers paid on a commission basis.

Proximity to the master cut both ways with respect to the treatment of slaves. Slaves in the household were usually materially better off, and in some cases close ties developed between masters and these slaves, such as those between masters and their former nannies, or with a favored concubine. However, proximity meant more sustained and direct supervision, which might easily become brutal.

Ethnic and somatic or perceived racial differences between masters and slaves operated in complex ways. Intra-ethnic slavery was uncommon in world history, although by no means nonexistent, while that between peoples of different perceived 'racial' groups was frequent. The common view that New World slavery was distinctive in that masters and slaves belonged to different perceived races is incorrect. Where there were somatic differences, the treatment of the slave depended on how these differences were perceived and, independently of this, how attractive the slave was in the eyes of the master. Scandinavian women were prized in the slave harems of many Islamic Sultans, but so were attractive Ethiopian and subSaharan women. Furthermore, in Muslim India and eighteenth-century England and France, dark-skinned slaves were the most favored, especially as young pages. In the New World, on the other hand, mulatto and other light-skinned female slaves were often better treated than their more African-looking counterparts.

Two other factors should be mentioned in considering the treatment of slaves: Laws and religion. Slightly more than a half of all slave societies on which data exist had some kind of slave laws, in some cases elaborate servile codes, the oldest known being those of ancient Mesopotamia. Slightly less than a half had none. Laws did make a difference; a much higher proportion of societies without any slave codes tended to treat slaves harshly. Nonetheless, the effectiveness

of laws were mediated by other factors such as the relative size of the slave population, and religion.

The degree to which religion, especially Islam and Christianity, influenced the treatment of slaves is a controversial subject. Islam had explicit injunctions for the treatment of slaves, and these were sometimes influential, especially those relating to manumission. Although racism and strong preference for light complexion were found throughout the Islamic lands, it is nonetheless the case that Islam as a creed has been more assimilative and universalist than any other of the world religions, and has rarely been implicated in egregiously racist movements similar to those that have tarnished the history of Christianity such as apartheid, the Ku Klux Klan, Southern Christianity during the era of segregation, and the ethnic cleansing of Eastern Europe. However, Islam never developed any movement for abolition, and, in general, strongly supported the institution of slavery, especially as a means of winning converts.

For most of its history up to the end of the eighteenth century, Christianity simply took slavery for granted, had little to say about the treatment of slaves, and generally urged slaves to obey their masters. This changed radically with the rise of evangelical Christianity in the late eighteenth and early nineteenth centuries, during which it played a critical role in the abolition of both the slave trade and of slavery in Europe and the Americas. Throughout the world, Christianity appealed strongly to slaves and ex-slave populations, and in the Americas their descendants are among the most devout Christians. While Christianity may have had conservative influences on converted slaves, it is also the case that nearly all the major slave revolts from the late eighteenth century were influenced strongly by rebel leaders, such as Daddy Sharp in Jamaica and Nat Turner in America, who interpreted the faith in radical terms, or by leaders of syncretic Afro-Christian religions.

5. *Manumission as an Integral Element of Slavery*

With a few notable exceptions, manumission, the release from slavery, was an integral and necessary element of the institution wherever it became important. The reason for this is that it solved the incentive problem implicit in slavery. The promise of redemption proved to be the most important way of motivating the slaves to work assiduously on their masters' behalf.

Slaves were manumitted in a wide variety of ways, the most important being: Self-purchase or purchase by others, usually free relatives; through the post-mortem or testamentary mode; through cohabitation or sexual relations; through adoption; through political means or by the state; and by various ritual or sacral means. Self-purchase or purchase by relatives or friends of the slave into freedom was by far the most important means in the advanced slave economies of Greece and Rome, and in the modern capitalistic slave regimes. However, it was not the most widespread in other kinds of slave systems, and in parts of the world such as Africa it was uncommon. Post-mortem manumission by wills and other means was common in Islamic lands and in many parts of Africa. This form of manumission was usually intimately linked to religious practices and with expectations of religious returns for what was often defined as an act of piety.

As indicated earlier, in many slave societies slave concubines sometimes gained their freedom for themselves and their children from the master. Manumission by the state for acts of heroism or for military action was an important, though episodic, form of manumission not only in the ancient world, but in many New World slave societies. Thousands of slaves gained their freedom in this way, not only in Latin America, but also in North America during the American war of independence and in wars against Spain in the southern USA. Manumission by adoption was unusual, but in certain societies, such as ancient Rome, it constituted the most complete form of release from slavery.

Slaves were sometimes manumitted for ritual or religious reasons, or on special celebratory occasions. Although thousands of slaves were manumitted at Delphi, ostensibly by being sold to Apollo, such manumissions had become merely legal formalism by the second century BC, although the practice may have harked back to an earlier era when they were genuinely religious in character. In other societies, religious or ritual manumissions were often substitutes for earlier practices in which slaves were sacrificed. Since manumission meant the negation of slavery, for many peoples, freeing slaves was symbolically identical to killing them. Such practices were common in some parts of Africa and the Pacific islands, and among some indigenous tribes of the Northwest Coast of America slaves were either killed or given away in potlatch ceremonies. There is an extension of this primitive symbolic logic in Christianity, where Christ's sacrificial death is interpreted as a substitute for the redemption of mankind from enslavement to sin and eternal death, 'redemption' (from Latin *redemptio*) literally meaning 'to purchase someone out of slavery.'

In all slave societies, certain categories of slave were more likely to be manumitted than others. The most important factors explaining the incidence of manumission were: Gender, the status of parents, age, skill, residence and location, the means of acquisition and, where relevant, skin color. These factors are similar to those influencing the treatment of slaves and will not be discussed further.

They were also important in explaining varying rates between societies. Thus societies with relatively higher proportions of skilled slaves, a greater location

of slaves in urban areas, higher ratios of female to male slaves, and higher rates of concubinage between masters and slaves were more likely to have higher rates of manumission than those with lower levels of these attributes. Added to this is another critical variable: The availability of slaves, either internally or externally, to replace manumitted slaves. As long as such sources existed and the replacement value of the manumitted slave was less than the price of manumission, it suited slave-owners to manumit slaves, especially when the manumitted slaves were nearing, or had already reached, the end of their useful life.

However, on rare occasions the supply of slaves was cut off when demand remained high or was on the increase. This always resulted in very low rates of manumission. The most striking such case in the history of slavery was the US South, where the rise of cotton-based, capitalistic slavery in the early nineteenth century came within a few years of the termination of the Atlantic slave trade to America. Planters responded by reducing the manumission rate to near zero. A similar situation, though not as extreme, developed in the Spanish islands of the Caribbean during the nineteenth century, when the plantation system developed in the context of British enforcement of the abolition of the slave trade in the region. The result was that previously high rates of manumission plunged to very low rates.

A final point concerns the status of freed persons. This varied considerably across slave societies and bore little relation to the rate of manumission. Thus manumission rates were high in the Dutch colonies of Curaçao and in nineteenth-century Louisiana, but the condition of freedmen was wretched. Conversely, in the British Caribbean, where manumission rates were low, the condition of freedmen was relatively good, some groups achieving full civil liberties before the end of slavery. The main factor explaining the difference in treatment was the availability of economic opportunities for freedmen, and the extent to which the dominant planter class needed them as allies against the slaves. In the Caribbean, the small proportion of slaveholders and Europeans and the existence of a vast and rebellious slave population gave much political leverage to the small but important freed population. No such conditions existed in the United States, where the free and white population greatly outnumbered the slaves, and conditions for rebellion were severely restricted. In the ancient world, ethnic barriers meant generally low status and few opportunities for the manumitted in the Greek states, in contrast with Rome, where cultural and economic factors favored the growth and prosperity of a large freedmen class, a class that eventually came to dominate Rome demographically and culturally, with major implications for Western civilization.

See also: Caribbean: Sociocultural Aspects; Inequality; Inequality: Comparative Aspects; Property Rights; Slavery: Comparative Aspects; Slaves/Slavery, History of; Subaltern History; West Africa: Sociocultural Aspects

Bibliography

Blackburn R 1997 *The Making of New World Slavery*. Verso, London

Cohen D W, Greene J P (eds.) 1972 *Neither Slave Nor Free*. Johns Hopkins University Press, Baltimore, MD

Davis D B 1966 *The Problem of Slavery in Western Culture*. Cornell University Press, Ithaca, NY

Drescher S, Engerman S L (eds.) 1998 *A Historical Guide to World Slavery*. Oxford University Press, New York

Engerman S 1973 Some considerations relating to property rights in man. *Journal of Economic History* **33**: 43–65

Engerman S L (ed.) 1999 *Terms of Labor: Slavery, Serfdom, and Free Labor*. Stanford University Press, Stanford, CA

Engerman S, Genovese E (eds.) 1975 *Race and Slavery in the Western Hemisphere*. Princeton University Press, Princeton, NJ

Eltis D, Richardson D (eds.) 1997 *Routes to Slavery*. Frank Cass, London

Findlay R 1975 Slavery, incentives and manumission: A theoretical model. *The Journal of Political Economy* **83**(5): 923–34

Finley M I (ed.) 1960 *Slavery in Classical Antiquity: Views and Controversies*. Heffer, Cambridge, UK

Fogel R W 1989 *Without Consent or Contract: The Rise and Fall of American Slavery*. Norton, New York

Garlan Y 1995 *Les Esclaves en Greece Ancienne*. Editions de la Decouvert, Paris

Kirschenbaum A 1987 *Sons, Slaves and Freedmen in Roman Commerce*. Catholic University Press, Washington, DC

Landers J (ed.) 1996 Against the odds. *Slavery and Abolition*, Special issue **17**(1)

League of Nations 1938 *Report to the League of Nations Advisory Committee of Experts on Slavery*, Vol. 6. League of Nations, Geneva, Switzerland

Lovejoy P E 2000 *Transformations in Slavery: A History of Slavery in Africa*. Cambridge University Press, New York

Meillassoux C 1991 *The Anthropology of Slavery: The Womb of Iron and Gold*, trans. A Dasnois. Athlone, London

Miers S, Kopytoff I (eds.) 1977 *Slavery in Africa*. University of Wisconsin Press, Madison, WI

Miller J C 1993 *Slavery and Slaving in World History: A Bibliography, 1900–1991*. Kraus International, Millwood, NY

Patterson O 1967 *The Sociology of Slavery: Jamaica, 1655–1838*. McGibbon & Kee, London

Patterson O 1982 *Slavery and Social Death*. Harvard University Press, Cambridge, MA

de Queiros Mattoso K M 1986 *To Be a Slave in Brazil, 1550–1888*, trans. A Goldhammer. Rutgers University Press, New Brunswick, NJ

Reid A (ed.) 1983 *Slavery, Bondage and Dependency in Southeast Asia*. University of Queensland Press, St. Lucia, Queensland

Rodriguez J P (ed.) 1977 *The Historical Encyclopedia of World Slavery*. ABC-CLIO, Santa Barbara, CA

Shepherd V, Beckles H (eds.) 2000 *Caribbean Slavery in the Atlantic World*. Ian Randle Publishers, Kingston, Jamaica

Watson J (ed.) 1980 *Asian and African Systems of Slavery*. Blackwell, Oxford, UK

O. Patterson

Slavery: Comparative Aspects

In the comparative study of slavery it is important to distinguish between slave holding societies and large-scale, or what Moses Finley called 'genuine,' slave societies (Finley 1968). The former refers to any society in which slavery exists as a recognized institution, regardless of its structural significance. Genuine slave societies belong to that subset of slave holding societies in which important groups and social processes become heavily dependent on the institution.

The institution of slavery goes back to the dawn of human history. It remained important down to the late nineteenth century, and persisted as a significant mode of labor exploitation in some Islamic lands as late as the second half of the twentieth century. Remnants of it are still to be found in the twenty-first century in a few areas. Yet it was only in a minority of cases that it metastasized socially into genuine slave societies, though in far more than the five cases erroneously claimed by Keith Hopkins (1978). Thus, the institution existed throughout the ancient Mediterranean, but only in Greece and Rome (and, possibly, Carthage) did genuine slave societies emerge. It was found in every advanced precapitalist society of Asia, but only in Korea did it develop into large-scale slavery. All Islamic societies had slavery, but in only a few did there emerge structural dependence on the institution. In all, there were approximately 50 cases of large-scale slavery in the precapitalist world. With the rise of the modern world, slavery became the basis of a brutally efficient variant of capitalism. There were at least 40 such cases, counting the Caribbean slave colonies as separate units (for the most complete list, see Patterson 1982, App. C)

This article examines two sets of problems. The first concerns those factors associated with the presence of institutionalized slaveholding. The critical question here is: why is the institution present in some societies yet not in other, apparently similar, ones? The second set of problems begins with the assumption that slavery exists, and attempts to account for the growth in significance of slavery. More specifically, such studies attempt to explain the origins, structure, dynamics, and consequences of genuine slave societies.

1. Comparative Approaches to Slaveholding Societies

There is now a vast and growing body of literature on slavery (Patterson 1977a, Miller 1993, Miller and Holloran 2000). Yet, with the notable exception of certain anthropological studies (to be considered shortly) relatively few recent works are truly comparative in that they aim to arrive at general conclusions about the incidence, nature, and dynamics of slavery. Even those few works that compare two slaveholding societies tend to be concerned more with highlighting, through contrast, the distinctive features of the societies under consideration. This highly particularistic trend marks a regrettable departure from earlier studies of slavery.

The evolutionists of the nineteenth century were the first to offer explanations for the presence of slaveholding. Their basic proposition—that there was a close relationship between stages of socioeconomic development and the rise of slavery—has received no support from modern comparative studies (Pryor 1977). However, some of their less grandiose views have survived later scrutiny. One was their emphasis on warfare, and the demand for labor at certain crucial points in their scales of development as the important factors (Biot 1840, Westermarck 1908). The other was their finding that the socioeconomic role of women was critical. It was claimed, for example, that the subjection of women provided both a social and an economic model for the enslavement of men (Biot 1840, Tourmagne 1880).

The most important of the early-twentieth-century theorists was H. J. Nieboer (1910), who broke with the evolutionists with his open-resource theory. His work, which is still influential, was unusual also in its reliance on statistical data. His main hypothesis was that slavery existed to a significant degree only where land or some other crucial resource existed in abundance relative to labor. In such situations, free persons cannot be induced to work for others, so they must be forced to do so. The theory has had lasting appeal, especially for economic historians, since it is both testable and consistent with marginal utility theory. It was revived by Baks (1966) and by the MIT economist, Domar (1970). However, the theory has been shown to have little empirical support from modern cross-cultural data (Patterson 1977b, Pryor 1977); and Engerman (1973) has questioned seriously its theoretical consistency.

In the course of criticizing Nieboer, Siegel (1945) proffered a functionalist theory which claimed that chattel slavery may be expected to occur in those societies where there is a tendency to reinforce autocratic rule by means of wealth symbols, and where the process results in a rather strongly demarcated class structure. There is little empirical support for this theory, and it verges on circularity. In many premodern hierarchical societies with wealth symbols, it was slavery which was the main cause of increased stratification.

The best recent comparative work on slavery has come from historical anthropologists studying Africa. Meillassoux (1975, 1991) and his associates have

demonstrated ably from their studies of West Africa and the Sahel just how valuable a nondogmatic Marxian approach can be in understanding the dynamics of trade, ethnicity, status, and mode of production in complex lineage-based societies. The anthropologist Jack Goody (1980), drawing on Baks (1966), began by arguing that 'slavery involves external as well as internal inequality, an unequal balance of power between peoples.' From this unpromising start he enriches his analysis with both the ethnohistorical data on Africa and the cross-cultural statistical data from the Murdock Ethnographic Atlas. Central to his analysis is the role of women. The complex facts of slavery cannot be explained, he argues, 'except by seeing the role of slaves as related to sex and reproduction as well as to farm and production, and in the case of eunuchs, to power and its non-proliferation.' While being valuable, Goody's arguments are confined to Africa, and even with respect to this continent they are insufficient to explain why it was that some African societies came to rely so heavily on the institution while closely related neighboring groups did not.

The economist Frederic Pryor (1977) has come closest to formulating a general theory of premodern slavery using modern statistical techniques applied to cross-cultural data. He distinguishes between social and economic slavery, and argues that different factors explain the presence of one or the other. Central to his theory is the nineteenth-century idea that there is a correspondence or 'homologism' between male domination of women and masters' domination of slaves. He tried to demonstrate that economic slavery was most likely to occur 'in societies where women normally perform most of the work so that the slave and the wife act as substitutes for each other,' whereas social slavery was related to 'the role of the wife in a polygynous situation.' The theory is interesting and robustly argued but is problematic. Where women dominated the labor force there was no need for men to acquire slaves for economic purposes. On the contrary, it was precisely in such situations that slaves, when used at all, were acquired for social purposes. The presumed correspondence between wives and slaves is also questionable. There were profound differences between these two roles. Wives everywhere belonged to their communities, and were intimately kin-bound, whereas slaves everywhere were deracinated and kinless. Wives always had some rights, some power, at least in the protection of their kinsmen, in relation to their husbands and could rarely be killed with impunity, which was not true of slaves. Wives everywhere had honor, while slaves had none.

Far more comparative work is needed for an understanding of the institutionalization of slavery. The main variables involved in any explanation are now well known. They are: The economic and social role of women; polygyny; internal and external warfare; the mode of subsistence (mainly whether pastoral or agricultural); and the mode of political succession. However, the causal role and direction of each of these variables is complex, and their interaction with each other even more so. To take the role of women as an example, in some cases it is their role as producers that is important; in others, their role as reproducers. It is also difficult using static cross-cultural data to ascertain whether low female participation in production is the result of slavery or its cause.

2. *Approaches to the Study of Genuine Slave Societies*

Marxist scholars were the first to take seriously the problem of the origins, nature, and dynamics of large-scale slavery (for a review, see Patterson 1977a). Engels's (1962) periodization view that large-scale slavery constituted an inevitable stage in the development of all human societies was merely one version of nineteenth-century evolutionism, discussed earlier. It dominated East European thought until the deStalinization movement, and is still the orthodox view in mainland China.

More sophisticated, but no less problematic, have been attempts of modern Marxists to formulate a 'slave mode of production.' The most empirically grounded of these attempts is that of Perry Anderson (1974) who argued that 'the slave mode of production was the decisive invention of the Graeco-Roman world, which provided the ultimate basis both of its accomplishments and its eclipse.' There is now no longer any doubt that slavery was foundational for both Athens and Rome, and that these were the first societies in which large-scale or genuine slavery emerged. However, the concept of the 'slave mode of production' overemphasizes the materialistic aspects of slavery, making it of limited value for the comparative study of slavery. Slaves were indeed sometimes used to generate new economic systems—as in Rome and the modern plantation economies—but there are many cases where, even when used on a large scale, there was nothing innovative or distinctive about the economic structure. The narrow materialist focus not only leads to a misunderstanding of the relationship between slavery and technology in the ancient world, but more seriously, it fails to identify major differences between the Greek and Roman cases, and it is of no value in the study of genuine slave societies in which the structurally important role of slaves was noneconomic, as was true of most of the Islamic slave systems.

Several non-Marxist historical sociologists, drawing also on the experience of ancient Europe, have made important contributions to our understanding of genuine slave societies. According to Weber (1964) slavery on a large scale was possible only under three conditions: '(a) where it has been possible to maintain slaves very cheaply; (b) where there has been an

opportunity for regular recruitment through a well-supplied slave market; (c) in agricultural production on a large scale of the plantation type, or in very simple industrial processes.' However suggestive, there is little support for these generalizations from the comparative data. Medieval Korea (Salem 1978) and the US South (Fogel and Engerman 1974) disprove (a) and (b). And work on urban slavery in the modern world as well as the relationship between slavery and technology in the ancient world disprove (c) (Finley 1973).

Finley (1960, 1973, 1981) was the first scholar to grapple seriously with the problem of defining genuine slave societies, which he explicitly did not confine to those in which the slaves were economically important. He also cautioned correctly against too great a reliance on numbers in accounting for such societies. His emphasis on the slave as a deracinated outsider led him to the conclusion that what was most advantageous about slaves was their flexibility, and their potential as tools of change for the slaveholder class. Finlay also offered valuable pointers in his analyses of the relationship between slave and other forms of involuntary labor. And in criticizing Keith Hopkins's (1978) conquest theory of the emergence of genuine slave societies, persuasively he encouraged an emphasis on demand, as opposed to supply factors in the rise of slave society. Romans, he argued, captured many thousands of slaves during the Italian and Punic wars because a demand for them already existed and 'not the other way around.' He postulated three conditions for the existence of this demand: private ownership of land, and some concentration of holdings; 'a sufficient development of commodity production of markets'; and 'the unavailability of an internal labor supply.'

3. *A Framework for the Study of Slave Societies*

It is useful to approach the comparative study of slave society with an understanding of what structural dependence means. There are three fundamental questions. First, what was the *nature* of the dependence on slavery; second, what was the *degree* of dependence; and third, what was the *direction* of dependence. Answers to these three questions together determine what may be called the *modes of articulation* of slavery.

The nature of dependence on slavery may have been primarily economic or social, political or militaristic, or a combination of these. Economic dependence was frequently the case, especially in ancient Rome and in the modern capitalistic slave systems. Here the critical questions are: What were the costs and benefits of imposing and maintaining an economy based on slave labor? Stanley Engerman (1973) has written authoritatively on this subject. He distinguishes between the costs of imposition, of enforcement, and of worker productivity. Lower maintenance costs, more constant supply of labor, greater output due to the neglect of the non-pecuniary costs of labor, higher participation rates and greater labor intensity, and economies of scale, are the main factors proposed by Engerman in explaining the shift to slave labor. They apply as much to ancient Rome as they do to the modern capitalistic slave systems of America and the Caribbean.

Military and bureaucratic dependence were the main noneconomic forms, and they could be as decisive for societies as was economic dependence. The rise of Islam was made possible by the reliance on slave soldiers, and the Abbasid Caliphate, along with other Islamic states, was maintained by slave and ex-slave soldiers and bureaucrats. Although there was little economic dependence on slaves in most Islamic states—with the notable exception of ninth-century Iraq—many of them qualify as genuine slave societies as a result of the politico-military dependence of these regimes on slavery, and the ways in which slaves influenced the character of their cultures, many of the literati also being slaves or descendants of slaves (see Pipes 1981, Crone 1981, Marmon 1999).

The degree of dependence must be taken into account, although it is sometimes difficult to quantify. There has been a tendency to emphasize the size of the slave population over other variables, and while demography is important it can sometimes be misleading. Thus, it has been noted that no more than one in three adults were slaves in Athens at the height of its slave system in the late fifth century BCE as a way of playing down the importance of slavery. But as M. I. Finley liked to point out, the same was true of most parts of the American Slave South, and no one has ever questioned the fact that it was a large-scale slave system.

Finally, there is the direction of dependence. Even where a society had a high functional dependence on slavery it was not necessarily the case that slavery played an active, causal role in the development of its distinctive character. The institution, though important, was not structurally transformative. In ancient Rome, the Sokoto Caliphate and the slave societies of the Caribbean, the American South and Brazil, slavery was *actively articulated* and transformative. In other cases, however, it was *passively articulated*, the classic instance being Korea during the Koryo and Yi periods where, although a majority of the rural population were at times slaves, there was no significant transformation of the economy, and no impact on the regime's government and culture. The same was true of several of the modern Spanish colonies of Central and South America. During the sixteenth and seventeenth centuries there was marked structural dependence on slavery in Mexico and Central America as well as Peru, and the urban areas of Chile and Argentina, but the institution was not determinative in the course of development of these societies and, as in Korea, when slavery ended it left hardly a trace, either

cultural or social (Mellafe 1975, Palmer 1976, Klein 1986, Blackburn 1997).

These three factors together determined a finite set of modes of articulation, which is a sociologically more useful construct than that of the so-called 'slave mode of production.' Space permits only the most cursory mention of some of the most important such modes of articulation.

The *lineage mode of articulation* refers to those kin-based societies in which large-scale slavery was related critically to the rise to dominance of certain lineages in the process of class and state formation. In some cases, slavery originally served primarily economic ends; in others, mainly social and political ones; but in the majority the institution became multifunctional. This kind of genuine slave society was most commonly found in western and west-central Africa. Warfare, combined with some critical internal factor such as demographic change accounted for the rise of slavery (see Miller 1977 for the Kongo kingdom). The ideal case of this mode of articulation was the Asante state (Wilks 1975, Klein 1981). Slaves were originally incorporated as a means of expanding the number of dependents, a tendency reinforced by the matrilineal principle of descent. The growing number of slaves enhanced the lineage heads who owned them, and facilitated the process of lineage hierarchy and state formation. Later, the role of slaves was greatly expanded to include a wide range of economic activities, including mining.

The *predatory circulation mode* refers to those slave societies in which warfare and raiding mainly for slaves were the chief occupations of a highly predatory elite. The warrior class was usually assisted by a commercial class which traded heavily in slaves. There was usually a high rate of manumission of slaves, who not only contributed to the production of goods, but in their role as freedmen soldiers helped the ruling class to produce more slaves. Thus there was a continuous circulation of persons in and out of slave status as outsiders were incorporated as loyal freedmen retainers, creating a constant need for more enslaved outsiders. Slaves and freedmen often played key roles in, and sometimes even dominated, the palatine service and elite executive jobs.

The contiguous existence of pastoral and sedentary agricultural peoples with different levels of military might was the major factor in the development of this mode of articulation of slavery. The mode is strongly associated with Islam, and there are many examples of it in the Sahel and North Africa. The west–east spread of the Fulani over an area of some 3,000 miles over a period of 800 years provides one of the best cases of this mode, on which there is an abundance of historical and anthropological data (Lovejoy 2000, Pipes 1981, Meillassoux 1975, 1991).

The *embedded demesne mode* embraces those patrimonial systems dominated by large landed estates in which slaves were incorporated on a substantial scale to cultivate the demesne land of the lords. Serf or tenant laborers continued, in most cases, to be the primary producers of food for the society, and their rents or appropriated surpluses remained a major source of wealth for the ruling class. However, slaves were found to be a more productive form of labor on the home farms of the lords for the cost–benefit reasons analyzed by Engerman, mentioned above. The landowners got the best of both types of labor exploitation. This was a particularly valuable arrangement where a landed aristocracy needed to change to a new crop but was not prepared to contest the technical conservatism of serfs; where there was a supply of cheap labor across the border; or where there was a high level of internal absenteeism among the landed aristocracy.

This is the least understood or recognized form of advanced slave systems, perhaps because its mode of articulation was usually passive. Many of them were to be found in medieval Europe, especially in France, Spain, and parts of Scandinavia (Dockes 1982, Verlinden 1955, 1977, Bonassie 1985, Anderson 1974, Patterson 1991). The slave systems of the western 'states' of eleventh-century England, where slave populations were as high as 20 percent of the total are likely examples. So, possibly, was Viking Iceland, which may well have had similar proportions of slaves (Williams 1937, Foote and Wilson 1970). However, the ideal case of embedded demesne slavery was to be found in Korea, especially during the Koryo and early Yi periods. Here the slave population sometimes exceeded that of other forms of bonded labor (Salem 1978, Wagner 1974, Hong 1979).

The *urban–industrial modes* of articulation were those in which the urban elites came to rely heavily on slaves for support. Slaves played a relatively minor role in agriculture, although they may well have dominated the 'home farms' of certain segments of the ruling urban elites. Slave labor was concentrated in urban craft industries which produced goods for local consumption and exports, as well as the mining sector where it existed.

Slavery emerged on a large scale in such systems as a result of a combination of factors among which were the changing nature and frequency of warfare, conquest of foreign lands, the changing tastes of the ruling class, crises in the internal supply of labor, shifts in food staples, growing commercial links with the outside world, and demographic changes.

This mode of articulation could be either passive or active. To the extent that the character of the urban civilization depended on its urban economy, and to the extent that the economy depended on slave laborers, both manual and technical, to that degree were these systems active in their articulation. The classic case of the active mode of slave articulation were the ancient Greek slave systems, especially Athens of the fifth and fourth centuries BC (Finley 1981, Garlan 1982, De Ste. Croix 1981). Typical of the

passive mode of urban–industrial articulation were several of the Spanish slave systems of Central and South America during the sixteenth and seventeenth centuries, mentioned above.

The *Roman or urban–latifundic mode*: ancient Roman slavery stands in a class by itself, having no real parallels in either the ancient, medieval, or modern worlds. It came the closest to a system of total slavery in world history. It was distinctive, first, in the sheer magnitude of its imperial power and the degree of dependence on slavery, both at the imperial center and in its major colonial sectors. Second, Rome was unique in the extent of its reliance on slavery in both its rural and urban–industrial sectors. Third, the articulation of slavery was more actively transformative than in any other system, entailing what Hopkins (1978) called an 'extrusion' of free, small farmers and their replacement by slaves organized in gangs on large latifundi. Rome was unusual too, not only for its high levels of manumission, but for the extent to which slaves and slavery came to influence all aspects of its culture.

The *capitalist plantation mode*: contrary to the views of early economic theorists such as Adam Smith, and of Marxist scholars until fairly recently (Genovese 1965), modern plantation slavery was in no way incompatible with capitalism. Indeed, the rise of capitalism was intimately bound up with this mode of articulation of slavery, and at its height in nineteenth-century America was as profitable as the most advanced industrial factories of Europe or the northern United States (Fogel and Engerman 1974). Plantation slavery constituted one version of the worldwide systemic spread of capitalism, in which capital accumulation was advanced through the use of slaves in the peripheral colonial regions, complementing the use of so-called free labor in the metropolitan centers, and of serf and other forms of dependent labor in the semi-peripheral areas of Eastern Europe and Latin America (Wallerstein 1974, Blackburn 1997).

While plantation slavery bore some organizational resemblance to the ancient slave latifundia, and indeed can be traced historically to late medieval and early modern variants of the ancient model (see Solow 1991), it was distinctive in its complex, transnational system of financial support, in its production for international export, its heavy reliance on a single crop, its advanced organizational structure, which in the case of sugar involved innovative agri-industrial farms, in the vast distances from which slaves were imported, entailing a complex transoceanic slave-trading system, and in its reliance on slaves from one major geographic region who differed sharply from the slaveholding class in ethnosomatic terms.

However, it is important to recognize that there existed concurrently with the capitalistic plantation mode several other modes of slave articulation that were either precapitalist or at best protocapitalist. As has already been noted, many of the Spanish colonial systems in the Americas relied on forms of slavery that were nonaccumulative and distinctly premodern in their articulation, for example the urban–industrial modes of Mexico and parts of South America which were focused heavily on mining. In the Spanish Caribbean up to the third quarter of the eighteenth century a peculiar premodern form of agri-pastoral slavery prevailed which had little in common with the highly capitalistic slave plantation systems of the neighboring French and British islands, and which had to be dismantled at considerable political and socioeconomic cost in order to make way for the belated capitalistic slave systems that were imposed forcefully during the nineteenth century (Knight 1970).

While slavery was largely abolished in the Americas through the course of the nineteenth century (Blackburn 1988, Davis 1975), pockets of the institution persist into the twenty-first century in northern Africa and parts of Asia (Stearman 1999). With the possible exception of Mauritania, however, in none of these societies do we find genuine slave systems. It can be claimed, with cautious optimism, that this most inhuman form of societal organization has vanished from the world.

Bibliography

Anderson P 1974 *Passages from Antiquity to Feudalism*. London

Ayalon D 1951 *L'Esclavage du Mamelouk*. Jerusalem

Baks C J et al 1966 Slavery as a system of production in tribal societies. *Bijdrgen tot de Taal-, Land- en Volkenkunde* **122**

Biot E 1840 *De l'abolition de l'esclavage ancien en Occident*. Paris

Blackburn R 1988 *The Overthrow of Colonial Slavery, 1776–1848*. London

Blackburn R 1997 *The Making of New World Slavery*. London

Bonassie P 1985 Survie et extinction due regime esclavagiste dans l'Occident du haut moyet age. *Cahiers de Civilisation Medievale* **28**

Crone P 1981 *Slaves on Horses*. New York

Davis D B 1966 *The Problem of Slavery in Western Culture*. Ithaca, NY

Davis D B 1975 *The Problem of Slavery in the Age of Revolution, 1770–1823*. Ithaca, NY

De Ste Croix G E M 1981 *The Class Struggles in the Ancient Greek World*. Ithaca, NY

Dockes P 1982 *Medieval Slavery and Liberation*. Chicago

Domar E 1970 The causes of slavery or serfdom. *Journal of Economic History* **30**

Engels F 1962 *Anti-Duhring*. Moscow

Engerman S 1973 Some considerations relating to property rights in man. *Journal of Economic History* **33**

Finley M I (ed.) 1960 *Slavery in Classical Antiquity*. Cambridge

Finley M I 1968 Slavery. *International Encyclopedia of the Social Sciences*, Vol. 14

Finley M I 1973 *The Ancient Economy*. London

Finley M I 1981 *Economy and Society in Ancient Greece*. In: Shaw B D, Saller R P (eds.) , New York

Fogel R W, Engerman S 1974 *Time On the Cross*. Boston, Vols. 1 and 2

Foote P, Wilson D 1970 *The Viking Achievement*. London
Garlan Y 1982 *Les esclaves en grece ancienne*. Paris
Genovese E D 1965 *The Political Economy of Slavery*. New York
Goody J 1980 Slavery in time and space. In: Watson J L (ed.) *Asian and African Systems of Slavery*. Oxford
Hong S H 1979 The legal status of the private slaves in the Koryo Dynasty (in Korean). *Han'gukakpo* **12**
Hopkins K 1978 *Conquerors and Slaves*. Cambridge
Klein A N 1981 The two Asantes. In: Lovejoy P (ed.) *The Ideology of Slavery in Africa*. Beverly Hills
Klein H S 1986 *African Slavery in Latin America and the Caribbean*. New York
Knight F W 1970 *Slave Society in Cuba During the 19th Century*. Madison, WI
Lovejoy P E 2000 *Transformations in Slavery: A History of Slavery in Africa*. New York
Marmon S E (ed.) 1999 *Slavery in the Islamic Middle East*. Princeton, NJ
Meillassoux C (ed.) 1975 *L'esclavage en Afrique precoloniale*. Paris
Meillassoux C 1991 *The Anthropology of Slavery: The Womb of Iron and Gold*. (Trans Dasnois A). London
Mellafe R 1975 *Negro Slavery in Latin America*. Berkeley, CA
Miller J C 1977 Imbangala lineage slavery. In: Miers S, Kopytoff I (eds.) *Slavery in Africa*. Madison, WI
Miller J C (ed.) 1993 *Slavery and Slaving in World History: A Bibliography*. Millwood
Miller J C, Holloran J R (eds.) 2000 Slavery: Annual bibliographical supplement. *Slavery and Abolition* **21**(3)
Nieboer H J 1910 *Slavery as an Industrial System*. Rotterdam, The Netherlands
Palmer C 1976 *Slaves of the White God: Blacks in Mexico, 1570–1650*. Cambridge, MA
Patterson O 1977a Slavery. *Annual Review of Sociology* **3**: 407–49
Patterson O 1977b The structural origins of slavery: A critique of the Nieboer–Domar hypothesis. *Annals of the New York Academy of Science* **292**: 12–34
Patterson O 1982 *Slavery and Social Death*. Cambridge, MA
Patterson O 1991 *Freedom in the Making of Western Culture*. New York
Pipes D 1981 *Slave Soldiers and Islam*. New Haven, CT
Pryor F L 1977 A comparative study of slave societies. *Journal of Comparative Economics* **1**
Salem H 1978 *Slavery in Medieval Korea*. Ph.D. dissertation, Columbia University
Siegel B J 1945 Some methodological considerations for a comparative study of slavery. *American Anthropologist* **7**
Solow B (ed.) 1991 *Slavery and the Rise of the Atlantic System*. Cambridge, MA
Stearman K 1999 *Slavery Today*. Hove
Tourmagne A 1880 *Histoire de l'esclavage encien et moderne*. Paris
Verlinden C *L'Esclavage dans l'Europe Medievale*. Vols. 1 and 2 (Bruges, 1995; Ghent, 1977)
Wagner E 1974 Social stratification in 17th century Korea. *Occasional Papers on Korea* **1**
Wallerstein I 1974 *The Modern World System*. New York
Weber M 1964 *The Theory of Social and Economic Organization*. New York
Westermarck E 1908 *The Origins and Development of the Moral Ideas*. New York
Wilks I 1975 *Asante in the Nineteenth Century*. Cambridge
Williams C O 1937 *Thralldom in Ancient Iceland*. Chicago, IL

O. Patterson

Slaves/Slavery, History of

Slavery comes in many culturally specific forms. Common to all these different forms is the fact that slaves are denied the most basic rights. Slaves remain outside the social contract that binds the other members of a given society; they are people without honor and kin, subject to violent domination. They are objects of the law, not its subjects, considered as property, or chattel. Hence, we may describe slavery with Orlando Patterson as a form of 'social death,' and slaves as outsiders kept in a position of institutionalized marginality. In many languages, the term for foreigner also denoted slaves. Nowadays slavery is considered a most inhumane practice. The United Nations Universal Declaration of Human Rights proclaims: 'No one shall be held in slavery or servitude; slavery and the slave trade shall be prohibited in all their forms.' Signing the slavery conventions of 1926 and 1956, 122 states have committed themselves to abolish slavery and similar practices such as debt bondage, serfdom, servile marriage, and child labor.

In historical times, however, slavery as an institution was accepted in many societies with a minimum of social stratification all over the world, from early Mesopotamia and Pharaonic Egypt to ancient China, Korea, India, medieval Germany, and Muscovite Russia, from the Hebrews to the Aztecs and Cherokees. Yet, according to Moses Finley there were only five full-fledged slave societies with slaves constituting the majority of people and their work determining the whole economy. All five were located in what we tend to call the West, two in antiquity: Athenian Greece and Rome between the second century BC and the fourth century AD, three in modern times: Brazil, the Caribbean, and the southern US. The Ottoman Empire was notorious for its slave soldiers, the janissary, taken from its Slavic neighbors to the north and from Africa. The most exploitative form of chattel slavery emerged in the plantation economies of the New World under the aegis of European colonial rule with the UK playing a dominant role. Interestingly, this happened at the time when the medieval concept of labor as a common community resource was progressively replaced by a free labor regime in the UK. Ultimately, and for the first time in history, employer and employed in Europe were equal before the law, while slavery with its extreme status differential prevailed in the new world. The ensuing contradictions eventually led to the rise of an abolitionist movement in Europe and to the progressive banning of slavery in the nineteenth and twentieth centuries.

The last to make slavery illegal were the governments on the Arabian Peninsula. They did so in 1962.

Where there are slaves, there are people arguing about slavery, some rationalizing it, others denouncing it, and yet others to better demarcate the duties and rights of those involved. There is an extensive Islamic literature about slavery, paying particular attention to the question of who may, and who may not, be enslaved, to the duties of slave owners and the terms of manumission. Many classical and Christian authors dealt with the same problems. The abolitionist movement generated a literature of its own, stressing the horrors of slavery as in 'Oronooko,' the justly famous play by Aphra Behn, and the life narrative of Olaudah Equiano, who was taken captive as a child in his native Igbo village in what is now Nigeria. A common aspect of all these contributions written by contemporaries is their normative approach, based either in religion or moral philosophy, their focus on institutional and legal matters, and not the least on their reliance on anecdotal evidence and personal experience. In ancient Greece and in Rome, slaves were perceived as captives of war even when bought and defined as things until the Code of Justinian defined slaves as persons. Christian and Islamic thought highlighted religious difference. Slavery was primarily the fate of non-believers taken captive in 'just' wars. Yet, neither Christianity nor Islam prevented the enslavement of coreligionists, and if conversion obliged Islamic slave owners to manumit their slaves, it did not change the slaves' status in Christian societies. Later authors rather emphasized property relations. The same line of argument informs the 1926 League of Nations slavery convention which defines slavery as the 'status or condition of a person over whom any or all the powers attaching to the right of ownership are exercised.'

European and American scholars pioneered the historical study of slavery. Taking their cues from Greece, Rome, and medieval European experiences, they tended to interpret slavery as an intermediate stage in the general process of historical development. G. W. F. Hegel construed slavery as a sort of development aid for the benefit of Africans. Classical social sciences from Adam Smith to Max Weber identified slavery with backwardness. Marxist thinkers were even more explicit, positing primitive society, slavery, feudalism, capitalism, and communism as five consecutive stages of world history. Thus, slavery and capitalism, slavery and markets, and slavery and technological innovations were considered utterly incompatible. How then can one explain the rise of racial slavery in the plantation economies of the New World? From about 1500 to the late nineteenth century Portuguese, British, Dutch, French, Danish, Spanish, and American vessels carried some 12 million plus Africans across the Atlantic to be used mainly as field hands for the production of sugar, coffee, tobacco, rice, indigo, and cotton. Others were put to work in the gold and silver mines of Spanish America. Together they constituted by far the largest forced migration the world has known. Moreover, how should one interpret slavery in the South of the US, the first modern nation, if it is in opposition to capitalist logic?

Studies of US and West Indian slavery have dominated the field of slave studies from its very beginning. They set the agenda for research into other domains, such as: the transatlantic slave trade and its precedents in the Mediterranean and the Black Sea during the Middle Ages and the Renaissance. In the Italian city-states slaves were common until the fourteenth century, on the Iberian Peninsula up to the sixteenth century, many of them Moslems. Christian sailors, captured on the seas, ended up in North African slavery. Later, researchers shifted their attention to plantation slavery in other areas of the New World, and finally to slavery in various non-Western societies.

In the beginning, scholars studying Southern slavery followed an institutional approach, describing plantation slavery as a closed system of generalized violence and totalitarian control with dehumanizing effects similar to Nazi concentration camps. Slaves in these pioneering interpretations were victims, devoid of any power and will of their own. All the power was said to be vested in the slave owners who treated their slaves as mere chattel; whipping them into submission and keeping them at starvation level, even working them to death; denying them any family life and breeding them like cattle, buying and selling the slaves at their will and pleasure. Then came Eugene D. Genovese's monumental historical anthropology of plantation life in the American South. A Marxist himself, Genovese held on to the idea that the planters constituted a pre-capitalist class. Yet he showed that the slaves, although people without rights or hardly any rights, were not people without will. Quite to the contrary, they were even able to shape the structures of the system as a whole. Using cunning and other weapons of the weak in the daily transactions at the work place, they moved their owners to accept their humanity and some minimal duties. The outcome was a paternalist system with some give and take between slave-owner and slave.

With this new paradigm slaves got their agency back. Slavery as such came to be seen as a negotiated relationship and as an intrinsically unstable process, the terms of which varied with time and place. This process spanned different stages from enslavement and transporation to life in slavery and possible, yet by no means inevitable social re-integration. There were different ways to overcome the outsider status: resistance and flight, manumission or the purchase of freedom. Prompted by the American civil rights movement, slave lives, slave culture, and slave resistance developed into privileged subjects of study. Herbert G. Gutman showed that slaves against all odds were living in families, creating enduring networks of kinship and self help. Religion proved to be

another major sphere where slaves gained some cultural autonomy. Integrating Christian teachings and African beliefs they developed a richly textured religious life with dance and song as integral parts of worship.

As to resistance, scholars had to contend with the fact that apart from the Haitian Revolution, slave rebels were nowhere able to end slavery. Studies of slave rebellions nonetheless brought to the fore numerous instances of organised resistance. A more salient aspect of these studies pioneered by John Hope Franklin was the insight into the complexities of resistance which ranged from mocking tales and mimicking dances to the deliberate cultivation of African customs, to go slow tactics and flight and the establishment of so-called Maroon societies. These societies developed into refuges for other slaves. Most of them had only a limited life span, some, however, such as the seventeenth century kingdom of Palmares in the hinterland of Pernambuco evolved into autonomous centers of power at the margins of the colonial system, often cultivating their own forms of slavery. Most successful were the rebels in relatively weak colonial states with open frontiers such as Brazil, Spanish America, and Surinam where the Saramaka fought a successful war of liberation against the Dutch, lasting more than 100 years.

A major advance in the study of slavery was achieved when historians shifted their focus from the interpretation of anecdotal literary evidence to the analysis of statistical material and other mass data with the help of economic modeling techniques. Philip D. Curtin was the first to attempt a census of the transatlantic slave trade using customs records from the Americas. His calculations have later been complemented and refined, yet his general insights withstood any later scrutiny and still hold true at the beginning of the twenty-first century. Others tried to extend his census to the much longer established Trans-Saharan, Indian Ocean, and Red Sea trades although there is no comparable database for these regions. Current estimates stand at 12 million plus Africans transported along these routes from the seventh to the twentieth century. Others reconstructed slave prices at the points of purchase and sale, carefully tracing changes over time, and studying all the other aspects of slave trading and the plantation business. These endeavors engendered seminal new insights. Scholars were able to prove that the length of a voyage was the single most important factor determining mortality during the voyage across the sea, the ghastly Middle Passage, with death rates ranging from 10 to 20 percent and more. They showed that pricing was competitive and that supply was responsive to changes in price and that the transatlantic slave trade continued well into the nineteenth century. It ended when Brazil and Cuba emancipated their slaves in the 1880s, the last to do so in the Western hemisphere. Scholars also discovered that markets shifted eastwards and southwards along the West African coast, from Senegambia and the Upper Guinea coast to the Bight of Biafra and northwestern Angola. Imports into Africa varied over time and from place to place with cotton goods, brandy, guns, and iron tools in high demand everywhere along the coast. With the recent publication of more than 27,000 Atlantic slave trading voyages in one single data set, even non-specialists have access to parts of the evidence used in the econometric computations.

Taken together, the studies of the new economic history show that the Marxist thesis of an inherent contradiction between slavery and capitalism is untenable is untenable as both slave trading and plantation economies were capitalist to the core, with prices set by the laws of supply and demand. In a major reinterpretation of slavery in the antebellum South Robert William Fogel and Stanley L. Engerman also convincingly demonstrated that during the reign of 'King Cotton,' the planters acted as capitalist entrepreneurs who were ready to innovate when it paid to do so and who rather used incentives than the whip to drive their slaves. They even ventured the thesis that in material terms (diet, clothing, housing, and life expectancy) slaves in the South were better off than many peasants and workers in Europe and the northern US. Moreover, the system was well and flourishing to the end with high levels of productivity. Parallel quantitative studies of the plantation economies in the West Indies arrived at similar conclusions: abolition came not in the wake of economic decline, but rather in spite of economic success. Seymour Drescher even coined the term econocide.

A major contentious issue among scholars has been the significance of slavery for the Industrial Revolution. In 1947 Eric Williams argued that the profits from the 'triangular trade' had financed the Industrial Revolution in the UK, adding that abolition was a consequence of economic decline. This double-pronged 'Williams thesis' became a central tenet of dependency theory. A close look at hundreds and thousands of slave voyages, however, showed that average profits were smaller than scholars had previously assumed and could never explain why the UK achieved the stage of self-sustained growth in the latter half of the eighteenth century. New data on capital formation in that period points in the same direction. Nevertheless, slavery and the plantation economies contributed to British economic growth and the advent of modernity. Slavery reduced the cost of sugar and tobacco for European consumers and it made great fortunes; it created markets for cheap industrial mass-products, and sugar plantations with their gang-labor system may even be interpreted as 'factories in the fields' with a work discipline prefiguring that of modern industrial plants. The impact of slavery was also felt in the realm of ideas and ideology. It furthered a racially informed negative perception of all things African in the West and set the stage for racism, a most

tragic legacy of more than 400 years of slave trading, that the West has not yet come to terms with. The rise of racial slavery in the Atlantic system is a reminder of the fact that economic self-interest and markets, when left to themselves, can produce the most immoral and humanly destructive institutions. The term 'racial slavery' refers to the fact that almost all slaves employed in the Americas hailed from Africa as the Spanish had banned the enslavement of Amerindians already by the 1540s.

The impact of slavery on Africa is also very controversially debated. Yet, once again a change of perspective has led to a shift of emphasis. An earlier generation of scholars with first-hand experience of colonial domination saw African societies mainly as victims of external forces. They singled out the transatlantic slave trade as a first step towards underdevelopment as it depopulated whole areas, or so ran the argument, and induced African rulers to engage in wars for the sake of making prisoners for sale in exchange for guns, which then triggered new wars. This set of ideas informed dependency theory as well as world-system analysis and the early nationalist African historiography. Walter Rodney even argued that African slavery on the Guinea coast was a result of external demand.

Quantitative studies shattered some of these assumptions. The gun-slave-cycle could not be substantiated. In addition, historical demography, although a risky business, points to demographic stagnation from 1750 to 1850, with population losses in the areas directly affected by slaving when other continents were beginning to experience demographic growth. The causal link between the slave trade and Africa's economic backwardness as posited by Joseph E. Inikori is a contested issue. A better understanding of the complexities of African politics has furthermore led scholars to argue that most African wars of the period were independent of the external demand for slaves, although the slave trade generated income, strengthening some polities while weakening others. Asante, Dahomey, and the Niger Delta city-states were among the main beneficiaries while the kingdom of Benin was one of those who abstained from participating in the slave trade. Slave sales shifted with the vagaries of African politics, not to forget the cycles of drought with the ensuing threats of starvation, which also led people into slavery. In other cases, slavery was a punishment for the crimes people had committed.

To stay in business, European and American slave traders had to adapt to local circumstances. They were at the mercy of local merchants and local power holders. They hardly ever ventured beyond the ports of call and the forts they built along the West African coast. They also bought women and children, although the planters considered them second choice to men, and people from other areas than those that the planters preferred. Women constituted up to a third of those sent to the Americas. The female ratio among slaves was higher than among early European immigrants. It ensured an enduring African physical presence. It is also a further proof that African cultural and political parameters had the power to shape the Atlantic system, an argument best developed by David Eltis.

Studies of African slavery have furthered the understanding of these cultural parameters. Slavery was widespread in precolonial Africa and expanded in the nineteenth century despite, or rather because of, the abolition of the transatlantic slave trade by the UK in 1807. Some scholars even argue that West African societies then developed into fully-fledged slave societies with 50 percent and more of their population kept in bondage. First among these were the Sudanic societies from the Futa Jallon to Sokoto and Adamawa, where Islamic revolutions and state-building initiated more intense slave raiding than ever before. Yet, African slavery, often referred to as 'household slavery,' markedly differed from chattel slavery. As elsewhere, slaves had to do the hardest work and their owners treated them as outsiders, the first to be sacrificed in public rituals and put to death at the whim of their owner. However, it was common to find slaves in positions of albeit borrowed authority, such as traders, officers, and court officials. Elsewhere slave and slave owner worked side by side. Alternatively, slaves lived in villages of their own. They owned property and in some cases, even slaves of their own. More importantly, the slave status was not a fixed status; rather it changed over time along a slave to kinship continuum, with people born into slavery exempt from further sale. Recent scholarship has stressed that women were the pre-eminent victims of slavery in Africa as they were valued for their labor power and for the offspring they might give birth to. Moreover, one should never forget that power and prestige in African societies rather depended on the number of followers someone could count on, than the control of land which was not a scarce resource yet, or not everywhere. According to H. J. Nieboer agricultural slavery developed wherever land was plentiful and the productivity of agricultural labour was low.

The slave to kinship paradigm as expounded most forcefully by Igor Kopytoff and Suzanne Miers has lost some of its appeal when scholars discovered that in many places a slave past carries a social stigma to this day. Yet Africa is a continent with a wide spectrum of cultures. Hence, what prevails at one place may be contradicted at another. Nonetheless, the general rule of slavery getting more exploitative and more strictly closed when used for the production of export staples and in societies with strong state structures holds true for Africa as well as for the Americas. A case in point is slavery on the East African coast and the island of Zanzibar during the clove boom of the mid-nineteenth century. In West Africa too, slavery turned more adversarial when set to the production of palm oil in

the aftermath of the abolition of the transatlantic slave trade, called the period of legitimate trade. The less centralized power in a society was, the easier it was for slaves to overcome their outsider status. This, however, meant integration into a land-holding kin group as access to women and land, the keys to reproduction, was generally controlled by these. In modern Europe, on the other hand, property rights in human labor, one's own or others, were vested in the individual. Hence freedom in Africa, and other non-Western societies, meant belonging, but independence in Europe.

Colonialism also helped to keep the memory of slavery alive. In the early nineteenth century, the UK had done its best to end the slave trade. It even stationed a naval squadron for this purpose in West African waters and concluded anti-slave trade treaties with African rulers who had trouble seeing the benefits of abolition. The European public later tended to take the imperialist scramble for Africa as an anti-slavery crusade, but colonial administrations, while forcefully suppressing the slave trade, were reluctant to fight slavery as a social system; rather they closed their eyes and did their best to freeze existing social relations while claiming to bring progress to Africa. When slavery finally ended in Africa, it did so not so much as the result of conscious acts of emancipation but rather because of the emergence of a colonial labor market, with the socially defined older mechanisms for the integration of outsiders easing the transition. The men and women held in bondage eagerly took up alternatives wherever they were viable.

The cultural turn in social sciences has deeply influenced slave studies. There is no end to quantitative analyses. However, scholars have discovered that cultural parameters are important variables affecting what people do, even when they operate under the impress of a market system. Even *Homo oeconomicus* has to consider cultural values. Gender and slave memories have also attracted much more attention than before. The 41 volumes of interviews with former slaves conducted in the 1920s and 1930s in the US remain unrivalled, yet cf. the splendid collection of peasant narratives from Niger presented by Jean Pierre Olivier de Sardan. Archeologists have started to investigate slave material culture. Others have begun to retrace the construction of different Creole identities and cultures in the Americas so rich in imaginative creativity. They all encounter the same problem: as slaves by definition were people without voice, what they experienced and what they thought is to a great extent buried in the documents written by the perpetrators of slavery. To give the slaves their voice back is one of the nobler tasks of historical scholarship, but it is immensely difficult.

While studies of slavery become ever more detailed and localized, historians have come to realize that the Middle Passage is more a bridge than an abyss or a one-way street, binding different cultures together, a point first made by the anthropologist Melville J. Herskovits. Slaves came with a history of their own to the Americas, returning Africans and children of former slaves such as Olaudah Equiano, Wilmot Blyden, and James Africanus Horton were among the first to consciously define themselves as Africans. Hailing from different regions of Africa, yet sharing a common plight, they developed an awareness of a common African identity. The books they wrote helped to lay the foundations for Panafricanism, a strand of ideas and a political movement central to the freedom struggle of Africans in the twentieth century.

David Brion Davis noted a similar correlation between slavery and freedom in Western religious, legal, and philosophical discourses. Hence the double paradox and legacy of slavery: it inflicted death and hardship on millions and millions of people of mostly African descent for the benefits of a few; and while it was instrumental for the rise of racial prejudice, it also shaped the notion of freedom and equality as we know and cherish it today. To consider these global dimensions, is the challenge of any further research into the history of slavery even when dealing with very specific local aspects of the problem. But this is easier said than done because it requires historians to venture beyond the deeply rooted traditions of privileging the history of a particular nation-state, usually their own, into the open sea of a new comparative history. The fight against slavery is not won so long as debt bondage, servile marriage, forced labor, child labor, trafficking of women, forced prostitution, and other forms of bondage exist in many societies, including the very rich countries of the West.

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Human Rights, History of; Property, Philosophy of; Slavery as Social Institution; Slavery: Comparative Aspects; Trade and Exchange, Archaeology of

Bibliography

Bales K 1999 *Disposable People: New Slavery in the Global Economy*. University of California Press, Berkeley

Berlin I 1998 *Many Thousands Gone. The First two Centuries of Slavery in North America*. The Belknap Press, Cambridge, MA

Blackburn R 1997 *The Making of New World Slavery. From the Baroque to the Modern, 1492–1800*. Verso, London

Clarence-Smith W G 1989 *The Economics of the Indian Ocean Slave Trade in the Nineteenth Century*. Frank Cass, London

Craton M 1978 *Searching for the Invisible Man: Slaves and Plantation Life in Jamaica*. Harvard University Press, Cambridge, MA

Curtin P D 1967 *African Remembered; Narratives by West Africans from the Era of the Slave Trade*. University of Wisconsin Press, Madison, WI

Curtin P D 1969 *The Atlantic Slave Trade: A Census*. University of Wisconsin Press, Madison, WI

Davis D B 1966 *The Problem of Slavery in Western Culture*. Cornell University Press, Ithaca, NY
Davis D B 1984 *Slavery and Human Progress*. Oxford University Press, New York
Drescher S 1977 *Econocide: British Slavery in the Era of Abolition*. University of Pittsburgh Press, Pittsburgh, PA
Drescher S, Engerman S L 1998 *A Historical Guide to World Slavery*. Oxford University Press, New York
Elkins S M 1968 *Slavery; a Problem in American Institutional and Intellectual Life*, 2nd edn. University of Chicago Press, Chicago
Eltis D (ed.) 1999 *The Trans-Atlantic Slave Trade: a Database on CD-ROM*. Cambridge University Press, Cambridge, UK
Eltis D 2000 *The Rise of African Slavery in the Americas*. Cambridge University Press, Cambridge, UK
Finley M I 1987 *Classical Slavery*. Frank Cass, London
Fogel R W 1989 *Without Consent or Contract. The Rise and Fall of American Slavery*, 1st edn. Norton, New York
Fogel R W, Engerman S L 1974 *Time on the Cross; the Economics of American Negro Slavery*. Little, Brown, Boston
Franklin J H, Moss A A 2000 *From Slavery to Freedom: a History of African Americans*, 8th edn. A. A Knopf, New York
Genovese E D 1974 *Roll, Jordan, Roll: The World the Slaves Made*, 1st edn. Pantheon Books, New York
Gutman H G 1976 *The Black Family in Slavery and Freedom, 1750–1925*, 1st edn. New York Pantheon Books, New York
Hellie R 1982 *Slavery in Russia, 1450–1725*. University of Chicago Press, Chicago
Herskovits M J 1941 *The Myth of the Negro Past*. London Harper & Brothers, New York
Inikori J E, Engerman S L (eds.) 1992 *The Atlantic Slave Trade: Effects on Economies, Societies, and Peoples in Africa, the Americas, and Europe*. Duke University Press, Durham, NC
James C L R 1938 *The Black Jacobins: Toussaint l'Ouverture and the San Domingo Revolution*. The Dial Press, New York
Klein H S 1986 *African Slavery in Latin America and the Caribbean*. Oxford University Press, New York
Knight F W 1997 *The Slave Societies of the Caribbean. General History of the Caribbean*. UNESCO Pub, London
Kopytoff S M, Miers I 1977 African slavery as an institution of marginality. In: Kopytoff S M (ed.) *Slavery in Africa: Historical and Anthropological Perspectives*. University of Wisconsin Press, Madison, WI
Lovejoy P E 1983 *Transformations in Slavery: a History of Slavery in Africa*. Cambridge University Press, Cambridge, UK
Manning P 1990 *Slavery and African Life. Occidental, Oriental, and African Slave Trades*. Cambridge University Press, Cambridge, UK
Mattoso K M, de Queirós 1986 *To be a Slave in Brazil, 1550–1888*. Trans. Goldhammer A. Rutgers University Press, New Brunswick, NJ
Meillassoux C 1991 *The Anthropology of Slavery: The Womb of Iron and Gold*. University of Chicago Press, Chicago
Miller J C 1993 *Slavery and Slaving in World History: A Bibliography, 1900–1991*. Knaus International Publications, Millwood, NY
Morrissey M 1989 *Slave Women in the New World: Gender Stratification in the Caribbean*. University Press of Kansas, Lawrence, KA
Nieboer H J 1910 *Slavery as an Industrial System; Ethnological Researches*, 2nd rev. edn. M Nijhoff, The Hague, The Netherlands
Olivier de Sardan J P 1976 *Quand nos ancêtres étaient captifs. Récits paysans du Niger*. Paris
Patterson O 1982 *Slavery and Social Death. A Comparative Study*. Harvard University Press, Cambridge, MA
Phillips U B 1918 *American Negro Slavery: A Survey of the Supply, Employment and Control of Negro Labor as Determined by the Plantation Regime*. D Appleton, New York
Price R 1973 *Maroon Societies: Rebel Slave Communities in the Americas*, 1st edn. Anchor Press, Garden City, CA
Raboteau A J 1978 *Slave Religion: The 'Invisible Institution' in the Antebellum South*. Oxford University Press, Oxford, UK
Rawick G P, Federal Writers' Project 1972 *The American Slave: a Composite Autobiography*. Contributions in Afro-American and African studies; no. 11. Greenwood, Westport, CT
Reid A, Brewster J 1983 *Slavery, Bondage, and Dependency in Southeast Asia*. St Martin's Press, New York
Robertson C C, Klein M A 1983 *Women and Slavery in Africa*. University of Wisconsin Press, Madison, WI
Rodney W 1982 African slavery and other forms of social oppression on the Upper Guinea Coast in the context of the African slave trade. In: Inikori J (ed.) *Forced Migration. The Impact of the Export Slave Trade on African Societies*. Africana , New York
Schwartz S B 1985 *Sugar Plantations in the Formation of Brazilian Society: Bahia, 1550–1835*. Cambridge University Press, Cambridge, UK
Sheriff A 1987 *Slaves, Spices, and Ivory in Zanzibar: Integration of an East African Commercial Empire into the World Economy, 1770–1873*. J. Currey, London
1980 *Slavery & Abolition*. Cass, London, Vol. 1
Stampp K M 1956 *The Peculiar Institution: Slavery in the Ante-Bellum South*, 1st edn. Knopf, New York
Thornton J 1992 *Africa and Africans in the Making of the Atlantic world, 1400–1680*. Cambridge University Press, Cambridge, UK
Toledano E R 1998 *Slavery and Abolition in the Ottoman Middle East*. University of Washington Press, Seattle, WA
Verger P F 1982 *Orisha: les Dieux Yorouba en Afrique et au Nouveau Monde*. Métaillé, Paris
Verlinden, C 1955 *L'esclavage dans l'Europe Médiévale*. De Tempel, Bruges, Belgium
Watson J L (ed.) 1980 *Asian and African Systems of Slavery*. University of California Press, Berkeley, CA
Williams E E 1994 *Capitalism and Slavery*, 1st edn. University of North California Press, Chapel Hill, NC
Willis J R 1985 *Slaves and Slavery in Muslim Africa*. Frank Cass, London
Wirz A 1984 *Sklaverei und kapitalistisches Weltsystem*. Suhrkamp, Frankfurt am Main, Germany

A. Wirz

Sleep and Health

1. Sleep Physiology

Sleep is a ubiquitous mammalian phenomenon which is associated with diminished responsiveness to external stimuli and a familiar and delightful sense of restoration under the circumstances of a normal night of sleep. The importance of sleep in daily living can

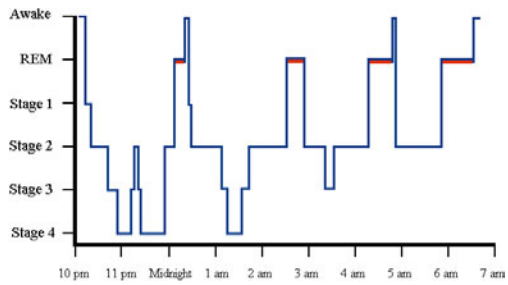

Figure 1
The distribution of sleep stages over the course of a normal night of sleep

easily be discerned when an individual has even a modest restriction in a normal night of sleep. This has been well documented to be associated with performance decrement and alterations in mood. The mysteries of sleep are many, and how it produces its restorative effects remain vague, but it is clear that it is at the core of mental and physical well being. Biological functioning cannot be thoroughly appreciated without integrating sleep and sleep biology into our understanding of waking biology. It is obvious now to the layman and professional sleep researcher both that sleeping behavior effects waking behavior, and waking behavior effects sleeping behavior.

In the mid-1950s, some of the mysteries of sleep began to unravel by a series of discoveries by Nathaniel Kleitman, William Dement, and Eugene Aserinsky at the University of Chicago (Dement 1990). These individuals produced a series of studies which delineated the brain wave patterns associated with different stages of sleep, and the unique stage of sleep which was termed rapid eye movement (REM) sleep, shown to be associated with dreaming. These stages of sleep were determined by the concomitant recording of the electroencephalogram (EEG), electro-octulogram (EOG), and the electromyogram (EMG). The stages of sleep are characterized by a pattern of activity in each of these variables. Brain wave activity changes dramatically from waking to the deep stages of sleep (stages 3 and 4) with high voltage slow waves becoming increasingly dominant. These stages are distributed throughout the night in a well-defined pattern of activity associated with REM periods approximately every 90 minutes (Fig. 1). REM periods are interspersed with stages 2, 3, and 4 such that stages 3 and 4 are identified largely in the first third of the night, with relatively little noted in the last third of the night. REM sleep is accumulated with episodes occurring every 90 minutes as noted, but as the night progresses, the REM periods become progressively longer, thus producing very long REM periods in the last third of the sleeping interval. Thus, REM sleep is identified primarily in the last third of the night, with relatively little in the first third of the night (Carskadon and Dement 1994).

Of interest, is the fact that REM sleep is associated with some unique physiologic changes. Thermoregulatory and respiratory functioning are markedly compromised during REM sleep. For example, the vasomotor mechanisms necessary to respond appropriately to increases and decreases in ambient temperature in order to maintain a normal core body temperature are suspended during REM sleep. This renders the mammalian organism poikilothermic, or cold blooded, during REM sleep. In addition, appropriate respiratory compensatory responses to both hypoxemia and hypercapnia are substantially blunted during REM sleep. These data combine to suggest that even under normal circumstances, REM sleep is a period of physiologic risk woven into the fabric of a normal night of sleep.

Other phenomena associated with REM sleep are of considerable interest. For example, it is known that REM sleep is associated with a skeletal muscle paralysis. This has been substantiated by the documentation of hyperpolarization of motor neurons during REM sleep in cats. In addition, REM sleep is generally associated with thematic dreaming, and it is felt that the inhibition of skeletal muscles prevents the 'acting out' of dreams. Also during each REM period penile tumescence occurs, resulting in a normal full erection in males (Carskaden and Dement 1994).

Non-REM (NREM) sleep, alternatively, is associated with a slowing of physiological processes to include heart rate, blood pressure, and general metabolic rate compared to the waking state. During the first episode of slow wave sleep (stages 3 and 4) there is a dramatic increase in the secretion of growth hormone. This has been shown to be specifically linked to the first episode of slow wave sleep, rather than a circadian phenomenon. No other hormone is quite so precisely identified with secretion during a specific stage of sleep.

2. *Subjective Aspects of Sleep*

The quality of sleep, and the physiologic parameters which produce 'good sleep' or a feeling of restoration in the morning are not understood. Deprivation of REM sleep as opposed to non-REM sleep does not appear to differentially affect subsequent mood or performance. It is well established that older adults have characteristic alterations in their sleep patterns, i.e., a marked diminution in stages 3 and 4 sleep, which are quite characteristically associated with complaints of nonrestorative, or poor sleep. Determining the physiologic characteristics of the subjective elements of 'good sleep' is difficult in that it is well known that the subjective and physiologic aspects of sleep can be

easily disassociated. It is not unusual, for example, for individuals with significant complaints of insomnia or poor sleep to have essentially normal physiological sleep patterns. Sleep is well known to be easily disturbed by mood and psychological state. Sleep is clearly disturbed by anxiety and in patients with generalized anxiety disorder. Characteristically this is associated with prolonged sleep onset latency and multiple awakenings subsequent to sleep onset. In addition, a well-recognized prodrome to a clinically significant depressive episode is an alteration in sleep pattern which includes an early onset of REM sleep and early morning awakenings with difficulty falling back to sleep. There are also data which document the fact that REM deprivation in depressed patients can produce a significant antidepressant effect (see *Depression*).

Good sleep is commonly associated with good health and a sense of well being. Measures of overall functional status have been known to be significantly correlated with both subjective and objective measures of daytime sleepiness. Other studies have shown that sleep disordered breathing is associated with lower general health status, with appropriate controls for body mass index, age, smoking status, and a history of cardiovascular conditions. Even very mild degrees of sleep disordered breathing have been shown to be associated with subjective decrements in measures of health status which are comparable to those individuals with chronic disease such as diabetes, arthritis, and hypertension.

Complaints of poor sleep and/or insomnia, and daytime sleepiness and fatigue are common. A recent Gallop survey in the USA indicated that approximately one third of Americans have insomnia. Thirty-six percent of American adults indicated some type of sleep problem. Approximately 27 percent reported occasional insomnia, while 9 percent indicated that their sleep-related difficulty occurs on a regular, chronic basis. In a study of a large sample of Australian workers, the prevalence of a significant elevation in the Epworth Sleepiness Scale was just under 11 percent. These data were not related significantly to age, sex, obesity, or the use of hypnotic drugs. Perhaps the most common cause of daytime fatigue and sleepiness, aside from self-imposed sleep restriction, is obstructive sleep apnea. The *sine qua non* of this sleep disorder is persistent sonorous snoring. Snoring itself is the most obvious manifestation of an increase in upper airway resistance, and as it progresses to more significant levels, the upper airway gradually diminishes in cross-sectional diameter and may produce a complete occlusion. Sonorous snoring can exist as a purely social nuisance or it can be associated with multiple episodes of partial and complete upper airway obstruction during sleep associated with dangerously low levels of oxygen saturation. This sleep-related breathing disorder is also associated with complaints of daytime fatigue and sleepiness that may be only minimally obvious to some individuals, but can be quite severe and debilitating in others (Orr 1997).

3. *Sleep Disorders*

There are a variety of documented sleep disorders which have been described in a diagnostic manual which can be obtained through the American Academy of Sleep Medicine. This manual describes the diagnostic criteria of a plethora of sleep disorders ranging from disorders which manifest themselves primarily as symptoms, i.e., insomnia, narcolepsy, to physiologic disorders which can be defined only through a polygraphic sleep study such as sleep apnea, periodic limb movements during sleep, and nocturnal gastroesophageal reflux. The prevalence of these disorders varies from narcolepsy (5–7 cases per 10,000) to OSAS (2–4 cases per 100) to insomnia which may be as high as 25 percent of the general population who report occasional problems with this disorder. Space constraints do not permit a discussion of all of these disorders, therefore we will touch only on the most commonly encountered sleep disorders.

Clearly, sleep disorders have well-documented consequences with regard not only to health but to daytime functioning. Perhaps the most common of all relates to the behavioral consequences of sleep fragmentation or sleep restriction. Either as a result of anxiety, or a physiologic sleep disturbance, sleep restriction and sleep fragmentation can produce documentable declines in performance, and increases in daytime sleepiness. The effects of even minimal sleep restrictions are cumulative across nights, but the effects can be quickly reversed in a single night of normal sleep. Sleep restriction is commonly noted in shift workers; it is estimated that approximately 30 percent of the American work force works rotating shifts, or a permanent night shift. Even permanent night shift workers rarely obtain 'normal' sleep in the daytime, and complaints of inadequate or non-restorative sleep are persistent in this group of workers. Studies have shown that accidents are 20 percent higher on the night shift, and 20 percent of night shift workers report falling asleep on the job. The demands of our increasingly complex technologic society have created a workforce that is under greater stress, obtaining less sleep, and clearly increasing prevalence of sleep complaints and sleep disorders.

An epidemiological study has estimated the prevalence of sleep disordered breathing (defined as five obstructive events per hour or greater) was 9 percent for women and 24 percent for men (Young et al. 1993). It was estimated that 2 percent of women and 4 percent of middle-aged men meet the minimal diagnostic criteria for sleep apnea, which includes five obstructive events per hour with the concomitant symptom of daytime sleepiness. Although, obstructive

sleep apnea syndrome (OSAS) is felt to be a predominantly male phenomenon, the incidence rises sharply among postmenopausal women. Persistent, loud snoring, and the concomitant upper airway obstruction, has been shown to carry with it a variety of medical and behavioral risks. Various studies have shown that snoring is a significant predictor of hypertension, angina, and cerebral vascular accidents independent from other known risk factors. One study has shown in a multiple regression analysis that snoring was the only independent risk factor which differentiated stroke occurring during sleep and stroke occurring at other times of the day. Other studies have documented a higher mortality rate in patients with moderate to severe OSAS compared to those with a less severe manifestation of this disorder. Furthermore, other studies have shown that in individuals with curtailed life expectancy secondary to OSAS, the most common cause of death was myocardial infarction. Other studies have confirmed that more aggressive treatment of OSAS reduces the incidence of vascular mortality (Hla et al. 1994).

The issue of the relationship between snoring, OSAS, and hypertension is somewhat controversial. The frequent association of significant OSAS with male gender, and obesity make it difficult to determine the relative contribution of each variable to hypertension. One large population study of 836 males from a general medical practice in England revealed a significant correlation between overnight hypoxemia and systemic blood pressure, but this could not be determined to be independent of age, obesity, and alcohol consumption (Stradling and Crosby 1990). No significant relationship was found with snoring. Alternatively, another excellent study which actually utilized overnight polysomnography and 24-hour ambulatory blood pressure monitoring did find an association between hypertension and sleep apnea which is independent of obesity, age, and sex in a nonselected community based adult population (Hla et al. 1994). Clinically, it is recognized that appropriate treatment of OSAS will often result in a notable reduction in blood pressure, often independent of weight loss. In a recent National Institute of Health supported project on sleep and cardiovascular functioning, preliminary data have shown that 22 to 48 percent of hypertensive patients have been observed with significant OSA, and 50 to 90 percent of sleep apnea patients have been documented to have hypertension.

Perhaps the most well-known 'sleeping disorder' or 'sleep sickness' is narcolepsy. Long assumed to be synonymous with uncontrollable daytime sleepiness, narcolepsy is now known to be a neurological disorder, and genetic abnormalities have recently been described in a canine model of narcolepsy. This disorder is associated with an extraordinary degree of daytime sleepiness but that alone does not define the syndrome. Many individuals with OSAS also have an equivalent degree of daytime sleepiness. Narcolepsy is identified by other ancillary symptoms. Perhaps the most common and dramatic is cataplexy in which individuals have a partial or complete loss of skeletal muscle control in the face of emotional stimuli such as fear, anger, or laughter. In the sleep laboratory, narcolepsy is associated with a unique pattern of sleep where an onset of REM sleep is noted usually within 10 to 15 minutes after sleep onset. This is considered to be a pathognomonic sign of narcolepsy. Unfortunately, there is no cure for narcolepsy at the present time, and treatment is limited to the symptomatic improvement of daytime sleepiness with stimulant medication, and the control of the ancillary symptoms primarily via tricyclic antidepressant medication.

The results of the Gallop survey indicated that 36 percent of American adults suffer from some type of sleep problem. Approximately 25 percent reported occasional insomnia, while 9 percent said that their sleep-related difficulty occurs on a regular or chronic basis. In addition, the survey documented that insomniacs were 2.5 times more likely than noninsomniacs to report vehicle accidents in which fatigue was a factor. In addition, compared to noninsomniacs, insomniac patients reported a significantly impaired ability to concentrate during the day. The well-documented excessive daytime sleepiness in patients with obstructive sleep apnea has been documented in several studies to result in a significant increase in traffic accidents, as well as at-fault accidents (Findley et al. 1989). The extreme sleepiness noted in patients with OSAS is almost certainly the result of an extreme degree of sleep fragmentation secondary to arousals from sleep by repeated events of obstructed breathing. Furthermore, it is well established that the sleepiness experienced by patients with OSAS is completely reversible by appropriate therapy which resolves the sleep related breathing disorder. The most common approach to therapy currently is the application of a positive nasal airway pressure via a technique referred to as continuous positive airway pressure (CPAP).

The effects of sleep deprivation, either as a result of willful sleep restriction, poor sleep secondary to insomnia, or fragmented sleep secondary to OSAS, has substantial effects on physical and behavioral health as noted above. Studies on chronic sleep deprivation have shown minimal physiologic effects, but have documented dose related effects in performance decrements and decreased alertness. More recently, however, an extraordinary set of studies has demonstrated that long-term sleep deprivation does have very significant effects in rats, and can ultimately result in death. The most likely explanation of death appears to be the result of a profound loss of thermal regulation (Rechtschaffen 1998). Also of interest, are recent studies which suggest alterations in immune function secondary to chronic sleep deprivation in humans.

In conclusion, the intuitive notion that sleep has important consequences with regard to one's health

appears to be clearly documented. Alterations in sleep producing a fragmentation of sleep or restriction in the normal duration of sleep have been shown to have significant consequences with regard to waking behavior. In addition, sleep disorders which fragment sleep such as obstructive sleep apnea produce effects that relate not only to waking behavior, but also have significant consequences with regard to both cardiovascular and cerebral vascular complications as well as producing a significant increase in mortality. Perhaps, the most important message to be gleaned from the remarkable increase in our knowledge of sleep is that sleeping behavior affects waking behavior and waking behavior affects sleeping behavior.

See also: Stress and Health Research

Bibliography

Carskadon M, Dement W 1994 Normal human sleep: An overview. In: Kryger M, Roth T, Dement W (eds.) *Principles and Practice of Sleep Medicine*. Saunders, Philadelphia, PA

Dement W C 1990 A personal history of sleep disorders medicine. *Journal of Clinical Neurophysiology* **7**: 17–47

Findley L J, Fabrizio M, Thommi G, Suratt P M 1989 Severity of sleep apnea and automobile crashes. *New England Journal of Medicine* **320**: 868–9

Hla K M, Young T B, Bidwell T, et al 1994 Sleep apnea and hypertension. *Annals of Internal Medicine* **120**: 382–8

Orr W 1997 Obstructive sleep apnea: Natural history and varieties of the clinical presentation. In: Pressman M, Orr W (eds.) *Understanding Sleep*. American Psychological Association, Washington, DC

Rechtschaffen A 1998 Current perspectives on the function of sleep. *Perspectives in Biology and Medicine* **41**: 359–90

Stradling J R, Crosby J H 1990 Relation between systemic hypertension and sleep hypoxaemia or snoring: Analysis in 748 men drawn from general practice. *British Medical Journal* **300**: 75

Young T, Palta M, Dempsey J, et al 1993 The occurrence of sleep-disordered breathing among middle-aged adults. *The New England Journal of Medicine* **328**: 1230–5

W. C. Orr

Sleep Disorders: Psychiatric Aspects

Complaints of too little sleep (insomnia) or too much sleep (hypersomnia), or of sleep that is not restorative enough, are termed dyssomnias. This chapter is intended to give an overview on the frequency and risk factors of dyssomnias, and on the psychiatric conditions that may be associated with them, and the physiopathological concepts related to sleep changes in depression. Of the greater than 30 percent of the general population who complain of insomnia during the course of one year, about 17 percent report that it is 'serious.' Insomnia is encountered more commonly in women than men and its prevalence increases with age. Insomnia represents one of the major features of depression and appears to be one type of the prodromal symptoms and risk factors for the development of major depression. Since chronic dyssomnia most often occurs as a comorbid disturbance of psychiatric and physical conditions, a thorough evaluation of the patient and his/her sleep complaints is needed to lay the foundation for accurate diagnosis and effective treatment.

The diagnosis of insomnia is based upon the subjective complaint of sleeping too little. Patients report difficulties in initiating or maintaining sleep or of non-restorative sleep, i.e., not feeling well rested after sleep that is apparently adequate in amount, and tiredness during the day. Insomnia may occur as primary or as secondary disorder due to other psychiatric conditions, due to general medical conditions, and/or due to substance misuse. Compared to secondary insomnia, the prevalence of primary insomnia is relatively small.

Hypersomnia includes complaints of excessive sleepiness characterized either by prolonged sleep episodes and/or excessive sleepiness during the day. These symptoms may interfere with social, occupational, and/or other areas of functioning. Like insomnia, hypersomnia may occur as a primary disorder or as a secondary disorder due to psychiatric and/or medical conditions, and/or due to substance misuse. Especially among patients encountered in psychiatric practices and hospitals, secondary sleep disturbances are more common than primary sleep disturbances. This is particularly important to bear in mind when evaluating a patient with sleep complaints. Dyssomnia is particularly often associated with psychiatric disorders such as depression, schizophrenia, anxiety disorders, or personality disorders, and with misuse of drugs and/or alcohol. Whenever possible, it is clinically useful to assess which is the primary and which is the secondary disturbance. This will facilitate clinical treatment and management and improve preventive measures.

1. Frequency and Risk Factors

Somnipathies are among the most frequent complaints a general practitioner must deal with. According to epidemiological studies, 19–46 percent of the population report sleep problems. Of these, 13 percent suffer from moderate or severe disturbances. In terms of the narrow definition of diagnostic criteria, that is, initial insomnia, interrupted sleep and disturbance of daily well being, 1.3 percent of the population suffers from somnipathies.

Risk factors for somnipathies are psychological stress or psychiatric illness. A high prevalence of

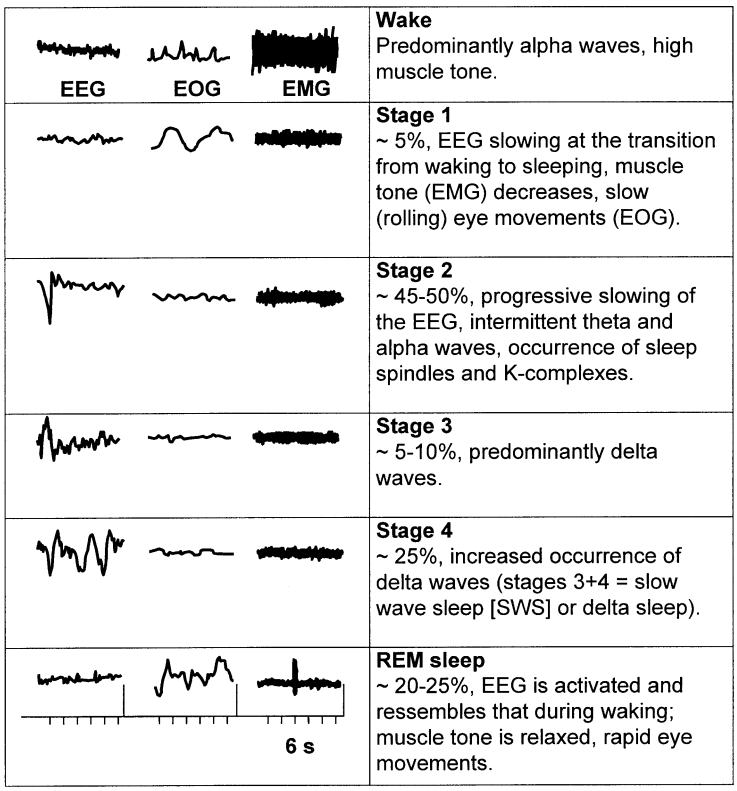

	Wake Predominantly alpha waves, high muscle tone.
	Stage 1 ~ 5%, EEG slowing at the transition from waking to sleeping, muscle tone (EMG) decreases, slow (rolling) eye movements (EOG).
	Stage 2 ~ 45-50%, progressive slowing of the EEG, intermittent theta and alpha waves, occurrence of sleep spindles and K-complexes.
	Stage 3 ~ 5-10%, predominantly delta waves.
	Stage 4 ~ 25%, increased occurrence of delta waves (stages 3+4 = slow wave sleep [SWS] or delta sleep).
	REM sleep ~ 20-25%, EEG is activated and ressembles that during waking; muscle tone is relaxed, rapid eye movements.

Figure 1
Sleep stages according to Rechtschaffen and Kales (1968)

somnipathies was reported by volunteers suffering from stress, tension, solitude, or depression. More severe sleep problems were found to be clearly related to psychiatric illness such as depression and anxiety disorders as well as substance misuse.

A recent World Health Organization (WHO) collaborative study in 15 different countries found that insomnia is more common in females than males and increases with age. In fact, it is thought that half of the population over 65 years of age suffers from chronic sleep disturbances. The WHO study also found that 51 percent of people with an insomnia complaint had a well defined *International Classification of Diseases-10* (*ICD-10*) mental disorder (mainly depression, anxiety, alcohol problems). Although many somnipathies develop intermittently, 70–80 percent of the surveyed subjects had been suffering from sleeping problems for more than one year. Disturbances of concentration and memory, difficulties in getting on with daily activities, and depressive mood changes were among the problems most often related to insomnia. The pressure imposed by their ailment leads many patients to seek help in alcohol and drugs.

Insomnia has also costly financial consequences. The economical impact of insomnia can be divided into direct and indirect costs. Direct costs of insomnia include outpatient visits, sleep-recordings, and medications directly devoted to insomnia. There is too little knowledge about the exact figures. However, the direct costs of insomnia in the US has been estimated in 1990 to be $10.9 billion (with $1.1 billion devoted to substances used to promote sleep and $9.8 billion associated with nursing home care for elderly subjects with sleep problems). The direct costs related to the evaluation of sleep disorders by practitioners seem to be a small part of the total costs of insomnia. The indirect costs of insomnia include the presumed secondary consequences of insomnia such as health and professional problems, and accidents. The exact quantification of these costs is, however, controversial. It is often not known if sleep disorders are the cause or the consequence of various medical or psychiatric diseases. For instance, it has been observed that insomniacs report more medical problems than good sleepers do and have twice as many doctor visits per year than good sleepers do. Furthermore, subjects with severe insomnia appear to be hospitalized about twice as often as good sleepers. It has also been observed that insomniacs consume more medication for various problems than good sleepers do.

These results confirm previous observations showing that insomnia is statistically linked with a worse health status than individuals with good sleep. Again, it can not be established whether insomnia is the cause or the result of this worse status. For instance, one could reasonably hypothesize that insomnia promotes fatigue that could increase the risk of some diseases, or more simply decrease the threshold of others that could more easily develop.

2. *Regulation of Normal Sleep*

The quantity and quality of sleep is measured with polysomnographic recordings (PSG) that include the electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG). Human sleep consists of two major states, the Rapid-Eye-Movement (REM) sleep, and the nonREM (nonREM) sleep (Fig. 1). NonREM sleep is characterized by increasing EEG synchronization and is divided into stages 1–4. Stages 3 and 4 are also termed slow wave sleep (SWS) or delta sleep ('delta waves'). REM sleep has also been termed paradoxical sleep because of its wake-like desynchronized EEG pattern combined with an increased arousal threshold. Typical nocturnal sleep is characterized by 3-5 cycles of non-REM and REM sleep phases of about 60–100 minutes duration. At the beginning of the night, the cycles contain more non-REM sleep, particularly more SWS. Towards the end of the night, the amounts of REM sleep increases and the amount of SWS decreases. The following standard terms are used to describe quality and quantity of sleep: sleep continuity refers to the balance between sleep and wakefulness (i.e., initiating and maintaining sleep), and sleep architecture refers to the amount, distribution, and sequencing of specific sleep stages. Measures of sleep continuity include sleep latency (usually defined as the duration between lights out and the first occurrence of stage 2 sleep), intermittent wake after sleep onset, and sleep efficiency (ratio of the time spent asleep to total time spent in bed). REM latency refers to the elapsed time between sleep onset and the first occurrence of REM sleep. The amount of each sleep stage is quantified by its percentage of the total sleep time. REM density is a measure of eye movements during REM sleep; typically this is low in early REM sleep periods and increases in intensity with successive REM sleep periods. In normal sleep, the cycle of non-REM and REM sleep lasts approximately 60–100 minutes. Sleep usually begins with the non-REM sleep stage 1 and progresses to stage 4 before the appearance of the first REM period.

The duration of the REM sleep episodes and REM density usually increases throughout subsequent sleep cycles. As shown in Fig. 2, the pattern of nocturnal growth hormone secretion is associated with the development of SWS during the first non-REM sleep period. Cortisol secretion, on the other hand, appears to be associated with increased amount of REM sleep. Several non-REM and REM sleep measures including the amount of SWS, REM latency and density may be particularly altered in affective and schizophrenic disorders, in aging, and with the administration of certain drugs.

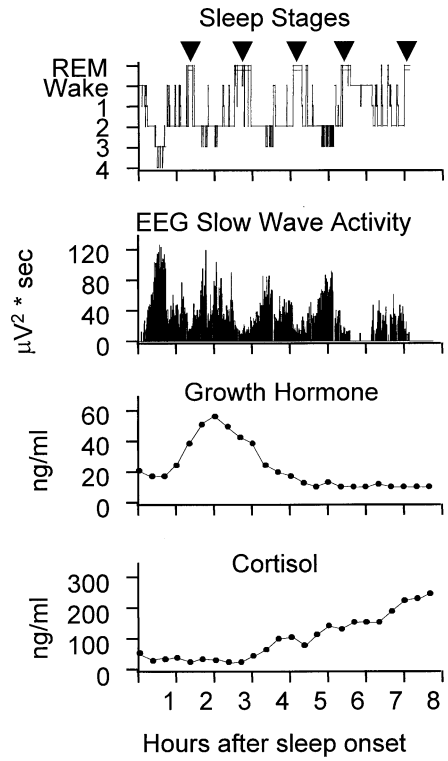

Figure 2
Bidirectional interaction between sleep EEG and nocturnal hormone secretion

At least three major processes are involved in the regulation of normal sleep. According to the Two-Process-Model of sleep regulation, sleep and wakefulness are influenced by both a homeostatic and a circadian process that interact in a complex way. The homeostatic Process S (related to SWS) increases with the duration of wakefulness prior to sleep onset and augments sleep propensity. The circadian Process C, driven by the internal clock located in the suprachiasmatic nuclei (SCN), describes the daily cycle of sleepiness and wakefulness and is also related to REM sleep. The third process is the ultradian rhythm of non-REM and REM sleep. The electrophysiological measures are associated with endocrinological and other physiological events.

Wakefulness, non-REM sleep, and REM sleep appear to be controlled by interacting neuronal networks, rather than by unique necessary and suffic-

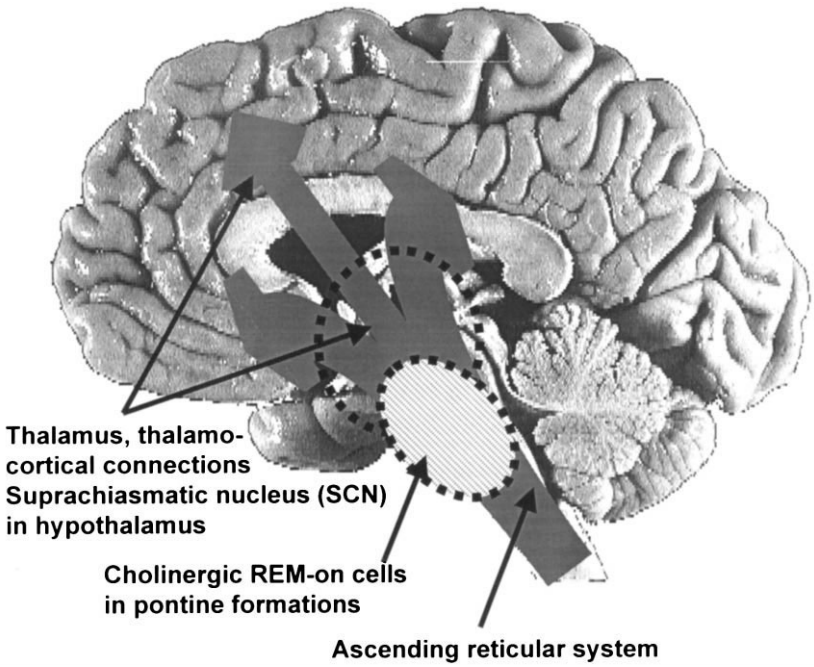

Figure 3
Neuroanatomy of sleep–wake regulation

ient centers. A simplified neuroanatomy of sleep–wakefulness is shown in Fig. 3. The non-REM–REM sleep cycle is regulated within the brainstem. According to current concepts, REM sleep is initiated and maintained by cholinergic neurons originating within the lateral dorsal tegmental and pedunculopontine nuclei in the dorsal tegmentum, and is inhibited by noradrenergic and serotonergic neurons originating in the locus coeruleus and dorsal raphe nuclei. Human pharmacological data are consistent with the neurophysiological concepts of the control of REM and non-REM sleep.

In contrast to the brain-activated state of REM sleep, non-REM sleep is characterized by synchronized, rhythmic inhibitory–excitatory potentials in large numbers of neurons in cortical and thalamic regions. Other areas implicated in the control of non-REM sleep include cholinergic and noncholinergic neurons in the basal forebrain, the hypothalamus, and the area of the solitary tract.

In addition to neuronal mechanisms in the control of sleep, more than 30 different endogenous substances have been reported to be somnogenic. These include delta sleep-inducing peptide, prostaglandin D_2, vasoactive intestinal peptide, adenosine, growth hormone-releasing hormone, and several cytokines including interleukins, tumor necrosis factor-α, and interferons. The significance of endogenous sleep factors in normal sleep physiology remains to be proven, but the possibility has interesting implications. As the Two Process Model of sleep regulation suggests, increased sleep propensity with progressive duration of wakefulness might be associated with the accumulation of a sleep factor and the homeostatic function of sleep or at least delta sleep might reflect its 'catabolism.'

3. *Sleep and Psychiatric Disorders*

Polygraph sleep research helped paving the way to the modern age of scientific, biological psychiatry. Much of the early sleep research beginning in the mid-1960s was descriptive mapping of events taking place during sleep and laid out the currently rich picture of polysomnographic features in the different psychiatric disorders. Although many of the objective sleep abnormalities associated with psychiatric disorders appear not to be diagnostically specific, these studies have also established sleep measures as neurobiological windows into the underlying pathophysiology associated with psychiatric illnesses.

Psychiatric disorders are among the most common primary causes of secondary sleep complaints, particularly of insomnia. Sleep abnormalities may be caused by central nervous system abnormalities associated with psychiatric illnesses as well as by accompanying

	Sleep continuity	SWS	REM sleep	Hyper-somnia
Depression/ dysthymia	↓↓↓	↓↓	↑↑↑	↕
Schizophrenia	↓↓↓	↓↓	↑↑	↕
Anxiety disorders	↓	↔	↔	↔
Alcohol dependence	↓↓	↓↓↓	↑	↔
Personality				
Borderline disorders	↓	↔	↑	↔
Dementias	↓↓↓	↓↓↓	↔	↕
Eating disorders	↓	↔	↔	↔

Table 1
Typical EEG sleep findings in psychiatric disorders (meta-analysis by Benca et al. 1992)

behavioral disturbances. Patients with depression, anxiety disorders, misuse of alcohol or drugs, schizophrenia and personality disorders may complain of difficulty falling asleep, staying asleep, or of inadequate sleep. Although specific sleep patterns are not necessarily diagnostic of particular psychiatric disorders, there are relationships between certain sleep abnormalities and categories of psychiatric disorders (Table 1).

A review of the literature on sleep in psychiatric disorders showed that EEG sleep in patients with affective disorders differed most frequently from those of normal control subjects.

4. *Sleep Abnormalities in Depression*

Sleep abnormalities in patients with major depressive disorders, as assessed by laboratory studies, can be classified as difficulties of sleep continuity, abnormal sleep architecture, and disruptions in the timing of REM sleep. Sleep initiation and maintenance difficulties include prolonged sleep latency (sleep onset insomnia), intermittent wakefulness and sleep fragmentation during the night, early morning awakenings with an inability to return to sleep, reduced sleep efficiency, and decreased total sleep time. With regard to sleep architecture, abnormalities have been reported in the amounts and distribution of nonREM sleep stages across the night and include increased shallow stage 1 sleep and reductions in the amount of deep, slow-wave (stages 3 + 4) sleep. REM sleep disturbances in depression include a short REM latency ($<$ 65 minutes), a prolonged first REM sleep period, and an increased total REM sleep time, particularly in the first half of the night.

Sleep disturbances are generally more prevalent in depressed inpatients, whereas only 40–60 percent of outpatients show sleep abnormalities. Moreover, a recent meta-analysis indicated that no single polysomnographic variable might reliably distinguish depressed patients from healthy control subjects or from patients with other psychiatric disorders. Table 1 gives an overview of the typical EEG sleep changes in major psychiatric conditions. This prompted some researchers to conclude that clusters or combinations of sleep variables better describe the nature of sleep disturbances in depression.

Although there is some disagreement as to which specific sleep EEG variables best characterize depressed patients, the importance of sleep to depression is clear. Persistent sleep disturbance is associated with significant risk of both relapse and recurrence, and increased risk of suicide. Sleep variables such as REM latency has also been shown to predict treatment response and clinical course of illness in at least some studies. It has also recently been suggested that the nature of the sleep disturbance at initial clinical presentation may be relevant to the choice of antidepressant medications and the likelihood of experiencing treatment-emergent side effects.

One of the most sensitive parameter for discrimination of patients with major depression from patients with other psychiatric disorders and healthy subjects is REM density which is substantially elevated only in depressed patients. The persistence of a depression-like sleep pattern in fully remitted depressed patients suggests that the pattern is a trait characteristic of the sleep measurements. However, in the past, subjects have undergone investigations only after the onset of the disorder, and therefore the altered sleep pattern may merely represent a biological scar. The answer to the question 'trait or scar' lies in the investigation of potential patients before the onset of the disorder. The EEG sleep patterns of subjects without a personal history but with a strong family history of an affective disorder differed from those of the controls without any personal history of family history of psychiatric disorders showing a depression-like sleep pattern with diminished SWS and increased REM density. Follow-up studies will determine whether the sleep pattern indeed represents a trait marker indicating vulnerability.

The importance of sleep in depression is also shown in other ways. Many well-documented studies show that total and partial sleep deprivation or selective REM sleep deprivation has antidepressant effects. Additionally, following total or partial sleep deprivation, patients with depression appear to be uniquely susceptible to clinical mood changes when they return to sleep. Patients who have shown a clinically significant antidepressant response to sleep deprivation are

at risk of awakening depressed again, even after very short naps (see *Depression*; *Antidepressant Drugs*).

5. *Neurobiology of Sleep: Relevance to Psychiatric Disorders*

Inspired by the growing knowledge of the underlying neurobiology of sleep, investigators proposed theories into the pathophysiology of psychiatric disorders. Because depression has been studied more than any other psychiatric syndrome in recent decades, the models have attempted to explain features of sleep in depression, such as short REM latency, decreased delta sleep, and the antidepressant effects of sleep deprivation in depressed patients.

Among the most prominent models of sleep changes in depression are the cholinergic–aminergic imbalance hypothesis for depression, the Two-Process-Model of sleep regulation, the Phase Advance Hypothesis, the Overarousal hypothesis, and the REM sleep hypothesis.

The 'cholinergic–aminergic imbalance hypothesis' for depression postulates that depression arises from an increased ratio of cholinergic to aminergic neurotransmission in critical central synapses. Because various features of the sleep of depressed patients (decreased REM latency and delta sleep) have been simulated in normal volunteers by pharmacological probes, the reciprocal interaction hypothesis from basic sleep research and the cholinergic–aminergic imbalance model from clinical psychiatry have been correlated. The reciprocal interaction model assumes that the cycling alternating pattern of non-REM and REM sleep is under the control of noradrenergic/serotonergic and cholinergic neuronal networks. Linking these concepts it is suggested if depression results from diminished noradrenergic and serotonergic neurotransmission, cholinergic activity would be expected to increase and lead therefore to the sleep disturbances of depression.

In recent years the role of serotonin in the regulation of sleep has afforded increased attention. Among other neurotransmitters, serotonin plays a role in the pathophysiology of depression and its treatment; chiefly all antidepressants ultimately lead to an enhancement of serotonergic neurotransmission that is believed to be associated with clinical improvement. Depression is associated with a disinhibition of REM sleep (shortened REM latency, increased REM density) and serotonin leads to a suppression of REM sleep. Furthermore, there is evidence that the antidepressant effect of sleep deprivation is related to a modification of serotonergic neurotransmission, thus sleep regulation and depression share common physiopathological mechanisms at the serotonergic level.

According to the 'Two-Process-Model of sleep regulation,' depression is thought to result from, or is associated with, a deficiency of Process S. This model was supported by evidence that EEG power density in the delta frequencies was decreased during sleep in depressed patients compared with controls and by the fact that sleep deprivation increases both Process S and mood. In the 'Extended Two-Process-Model of sleep regulation' the interaction of hormones with EEG features has been integrated. Preclinical investigations and studies in young and elderly normal controls and in patients with depression demonstrate that neuropeptides play a key role in sleep regulation. As an example, growth hormone releasing hormone (GHRH) is a common stimulus of SWS and growth hormone release, whereas corticotropin-releasing hormone (CRH) exerts opposite effects. It is suggested that an imbalance of these peptides in favor of CRH contribute to changes in sleep EEG and endocrine activity during depression. Based on the findings that the hypothalamic–pituitary–adrenocortical (HPA) axis is disregulated in depression, and that CRH produces EEG sleep changes reminding to depression on one hand (SWS reduction, REM sleep disinhibition), and that the somatotrophic system reciprocally interacts with the HPA system (decreased GHRH and SWS), it has been postulated that deficient Process S is associated with deficient GHRH and an overdrive of CRH. Although abnormally high values of cortisol secretory activity normalize after recovery from depression, growth hormone release and several characteristic disturbances of the sleep EEG may remain unchanged. These putative trait-dependent alterations suggest that strategies aiming at restoration of these sleep changes are worthy of exploration for their potential antidepressant effect.

The depressed state appears not only to be associated with a disturbance of sleep-wake homeostasis but also of circadian function or of their interaction. The 'Phase Advance Hypothesis' suggests that the phase position of the underlying circadian oscillator is 'phase advanced' relative to the external clock time. This is supported by studies showing that short REM latency could be simulated in normal controls by appropriate phase shifts of hours in bed.

As mentioned above, sleep deprivation has potent antidepressant effects in more than half of all depressed patients. This observation has prompted the hypothesis that depressed patients are 'overaroused.' Later, the 'Overarousal Hypothesis' has been forwarded and was supported by psychological self-rating studies suggesting that clinical improvement to sleep deprivation in depression may be associated simultaneously with subjective feelings of more energy (arousal), less tension, and more calmness (dearousal). Other data consistent with this hypothesis include the short, shallow, and fragmented sleep patterns, lowered arousal thresholds, and elevated nocturnal core body temperature often seen in depressed patients. This hypothesis has been tested by means of studies of localized cerebral glucose metabolism with ^{18}F-deoxyglucose positron emission tom-

ography in separate studies in depression, sleep deprivation, and the first non-REM period of the night. In these studies it was found that elevated brain metabolism prior to sleep deprivation predicted clinical benefits in depressed patients and that normalization of these measures is associated with clinical improvement. It was also found that local cerebral glucose metabolism in cingulate and amygdala at baseline was significantly higher in clinical responders than in nonresponders or normal controls. Furthermore, it was shown that glucose metabolic rate was increased during the first non-REM period in depressed patients compared with normal controls. Moreover, these studies demonstrated significant 'hypofrontality,' that is; reduced ratio of frontal to occipital activity compared with normal controls.

The 'REM sleep hypothesis' of depression has been based on the findings that REM sleep is enhanced in depression and that virtually all antidepressant drugs suppress REM sleep. Early, however not replicated, studies have shown that a sustained remission from depression may be achieved by selective REM sleep deprivation carried out repeatedly every night during about two weeks. This treatment modality was not followed because the long-term REM sleep deprivation was too exhausting for the patients. Recently, researchers have proposed a treatment modality that combined some of the above hypotheses. The so-called 'Sleep-Phase-Advance' protocol, that is scheduling sleeping times in a way that minimizes the occurrence of REM sleep, has been shown to produce a sustained antidepressant effect. However, also this treatment modality is demanding for both the patients and the institutional staff.

6. *Future Perspectives*

We have outlined just some aspects of sleep research that are relevant to depression. Linking basic and clinical approaches has been one of the 'royal roads' to the neurobiological underpinnings of psychiatric diseases and their treatment in the past, and to narrow the gap between bench and bedside. One of the most impressing links between sleep and depression is the fact that sleep deprivation alleviates depressive symptoms within hours and that sleep may restore the initial symptoms within hours or even minutes. Given the fact that the core symptoms of only few psychiatric and medical conditions may be 'switched' on and off, one promising lead of future research in depression is the quest to understanding the mechanisms underlying the neurobiological processes associated with sleep and wakefulness. In the future it is hoped that the rapidly evolving progress in basic neuroscience including recent molecular biology with systematic screening of gene expression and functional brain imaging techniques will help us in progressively uncovering the mystery of sleep, and ultimately improving treatment strategies of depression.

Bibliography

Benca R M, Obermayer W H, Thisted R A, Gillin J C 1992 Sleep and psychiatric disorders: A meta-analysis. *Arch. Gen. Psychiatry* **49**: 651–68

Borbély A A, Achermann P 1999 Sleep homeostasis and models of sleep regulation. *Journal of Biological Rhythms* **14**(6): 557–68

Gillin J C, Seifritz E, Zoltoski R, Salin-Pascual R J 2000 Basic science of sleep. In: Sadock B J, Sadock V A (eds.) *Kaplan & Sadock's Comprehensive Textbook of Psychiatry—VII*. Lippincott Williams and Wilkins, Philadelphia, pp. 199–208

Holsboer-Trachsler E, Kocher R 1996 Somnipathies: new recommendations for their diagnosis and treatment. *Drugs of Today* **32**(6): 477–82

Kupfer D J 1999 Pathophysiology and management of insomnia during depression. *Annals of Clinical Psychiatry* **11**(4): 267–76

Rechtschaffen A, Kales A 1968 *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Department of Health, Education and Welfare, Neurological Information Network, Bethesda

Silva J, Chase M, Sartorius N, Roth T 1996 Special report from a symposium held by the World Health Organization and the World Federation of Sleep Research Societies: an overview of insomnia's and related disorders—recognition, epidemiology, and rational management. *Sleep* **19**(5): 412–6

Steiger A, Holsboer F 1997 Neuropeptides and human sleep. *Sleep* **20**: 1038–52

Tononi G, Cirelli C 1999 The frontiers of sleep. *Trends in Neuroscience* **22**(10): 417–8

E. Holsboer-Trachsler and E. Seifritz

Sleep Disorders: Psychological Aspects

1. *Definition*

Good sleep is usually defined by its consequences. It is the amount and quality of sleep that results in the ability to feel and function well the next day. In contrast, a sleep disorder is a disturbance in the quantity or quality of sleep that interferes with waking performance and/or feelings of well-being. Psychological factors play a role in both the etiology and maintenance of many of the 84 sleep disorders recognized in the *The International Classification of Sleep Disorders Diagnostic Coding Manual* (1997). These in turn have both short-term and long-term effects on the psychological functioning of patients.

2. *Background*

Historically, sleep has not been an area of concern in psychology. Aside from psychoanalytic theory, which placed a good deal of emphasis on the role of unconscious motivation in explaining behavior, most psychological theory was based on behavior that was objectively observable. With the discovery in the early 1950s of rapid eye movement (REM) sleep, and its

close association with the distinctive mental activity, dreaming, a good deal of work was undertaken to explore whether and how this interacted with waking behavior. The question of whether dreaming has some unique psychological function was driven by the observation that this phenomenon was universal, regularly recurring, and persistent when experimentally suppressed.

Initially it was speculated that the study of dreaming might lead to better understanding of the psychoses. Was the hallucination of mental illness a misplaced dream? Was the high degree of brain activation in REM sleep conducive to memory consolidation of new learning? Did dreaming play a role in emotional adaptation? When, after some 25 years of experimentation, no clear consequences for waking behavior could be attributed to the suppression of REM sleep, interest in this area faded. Loss of REM sleep did not appear to interfere with memory consolidation, nor did it promote waking hallucinations. Its role in emotional adaptation remained speculative. Foulkes's (1982) study of children's dream development concluded that dream construction was not something special. It could be explained by the periodic internal activation of the brain stimulating bits of sensory memory material represented at the same level of cognitive sophistication as the child was capable of achieving in waking thought. Adult dreams were further diminished in importance by the Activation-Synthesis hypothesis of Hobson and McCarley (1977), whose work pinpointed the area of the brain where REM sleep turns on as, not where the higher mental processes take place, but in the lower brain stem. In this theory dreams only acquire meaning after the fact, by association to what are initially, unplanned, random sensory stimuli.

3. *Current Research*

Interest in the 24-hour mind, the interaction of waking–sleeping–waking behavior came back into focus with the development of the field of sleep disorder medicine. The impact of disordered sleep on psychological functioning is documented most convincingly through large epidemiological studies of insomnia and hypersomnia that involved extensive follow up. The findings of Ford and Kamerow (1989) that both insomnia and hypersomnia are significant risk factors for new psychiatric disorders have now been replicated in several large studies. Further work has established that the onset of a new episode of waking major depression can be predicted from the presence of two weeks of persistent insomnia (Perlis et al. 1997). These findings have sparked the development of psychological interventions for the control of insomnia in an effort to prevent the development of these psychiatric disorders. Some of these are behavioral programs that work directly to manipulate sleep to improve its continuity, others are psychotherapeutic in nature such as interpersonal therapy and cognitive behavioral therapy which address the interaction patterns, emotional and cognitive styles that are dysfunctional in depressed persons.

Again, it was the observation that morning mood on awakening in major depression was frequently low, that suggested an investigation of the REM sleep and dreaming of the depressed person to test whether defects in this system prevented overnight mood regulation or emotional adaptation. The work of Kupfer and his colleagues (Reynolds and Kupfer 1987) established that there are several REM sleep deviations associated with major depression not seen in normal individuals. This leads to a variety of manipulations in an attempt to correct these. Vogel's (Vogel et al. 1975) studies of extensive REM deprivation interspersed with recovery nights were most successful in improving waking mood and in bringing about remission without further treatment. Cartwright's work (Cartwright et al. 1998) on the effect on morning mood of various within-night dream affect patterns, showed that a 'working through' pattern, from dreams expressing negative mood to those with predominantly positive affect, predicted later remission from depression.

The recurring nightmares characteristic of post-traumatic stress disorder (PTSD) has stimulated renewed interest in addressing dreams directly in treatment programs, especially for those with long-lasting symptoms of this disorder. These methods typically involve active rehearsal of mastery strategies for the control of the disturbing dream scenario.

A good deal of effort is now being addressed to discovering those attitudes, beliefs, and habitual behaviors that are implicated in maintaining poor sleep patterns since the correction of these may help to restore normal sleep and abort the development of a major psychiatric disorder.

Studies of the psychological profiles of those exhibiting various sleep disorders have most often employed the Minnesota Multiphasic Personality Inventory (MMPI). These show insomnia patients have more scale elevations above the norms especially of those indicating neurotic personality characteristic, such as phobias and obsessive compulsive disorders, than do matched controls. Generally, insomnia patients appear to internalize tensions rather than express them and may often somaticize these and express them as pain syndromes. In terms of demographic variables most studies report more women than men complain of insomnia. Rates are higher in those who are separated or widowed rather than single or married, and among the unemployed. In addition, insomnia is more common in those of middle to lower socioeconomic status than in the highest class. This picture suggests that a loss of the time structure of work and of a love relationship may be precipitating psychological factors in disrupting sleep/wake rhythms.

Other sleep disorders, such as periodic limb movements of sleep, and sleep-related breathing disorders, are not related to personality, but do have an impact on relationships, due to an inability to share the bed at night, and on waking performance. Severe levels of any sleep disorder, whether insomnia or hypersomnia, limits work efficiency, cognitive clarity, and emotional stability.

4. *Methodological Issues*

Technological problems still hamper the progress in this field. Sleep laboratory studies are expensive and in many ways unnatural. Home studies while possible are unable to make repairs or adjustments as needed. Subjects must be awakened to obtain reports of their ongoing mental activity to investigate dream content. This aborts the end of each REM period, thus truncating the dream story. Dream reports are also limited by the subject's ability to translate his/her sensory experience during sleep into waking verbal terms.

Recent studies using positive emission tomography (PET) to image the brain activity during REM sleep have established that dreaming sleep differs from waking and from non-REM in the activation of the limbic and paralimbic systems in the presence of lower activity in the dorsal–lateral frontal areas (Maquet et al. 1996, Nofzinger et al. 1997). This is interpreted as confirming that dreaming engages the emotional and drive-related memory systems in the absence of higher systems of planning and executive control. This gives new impetus to the study of dreams as a unique mental activity whose place in the 24-hour psychology may now move ahead on firmer ground.

5. *Probable Future Directions*

Work initiated by Solms (1997) on the effect of localized brain lesions on dream characteristics suggests a new line of investigation in mapping how the brain constructs dreams. This strategy may give more power to understand some of the dream disorders currently less well understood; perhaps even the hallucinations of psychoses.

See also: Sleep: Neural Systems; Sleep States and Somatomotor Activity

Bibliography

American Sleep Disorders Association 1997 *The International Classification of Sleep Disorders Diagnostic and Coding Manual*. Rochester, MN

Cartwright R, Young M, Mercer P, Bears M 1998 The role of REM sleep and dream variables in the prediction of remission from depression. *Psychiatry Research* **80**: 249–55

Ford D E, Kamerow D B 1989 Epidemiologic study of sleep disturbances and psychiatric disorders: An opportunity for prevention? *Journal of the American Medical Association* **262**: 1479–84

Foulkes D 1982 *Children's Dreams: A Longitudinal Study*. Wiley and Sons, New York

Hobson J A, McCarley R W 1977 The brain as a dream state generator: An activation-synthesis hypothesis of the dream process. *American Journal of Psychiatry* **134**: 1335–48

Maquet P, Peters J-M, Aerts J, Delfiore G, Degueldre C, Luxen A, Franck G 1996 Functional neuroanatomy of human rapid-eye-movement sleep and dreaming. *Nature* **383**: 163–6

Nofzinger E A, Mintun M A, Wiseman M B, Kupfer D, Moore R 1997 Forebrain activation in REM sleep: an FDG PET study. *Brain Research* **770**: 192–201

Perlis M L, Giles D E, Buysse D J, Tu X, Kupfer D 1997 Self-reported sleep disturbance as a prodromal symptom in recurrent depression. *Journal of Affective Disorders* **42**: 209–12

Reynolds C F, Kupfer D J 1987 Sleep research in affective illness: state of the art circa 1987. *Sleep* **10**: 199–215

Solms M 1997 *The Neuropsychology of Dreams: A Clinico-anatomical Study*. L. Erlbaum Associates, Mahwah, NJ

Vogel G, Thurmond A, Gibbsons P, Sloan K, Boyd M, Walker M 1975 REM sleep reduction effects on depressed syndromes. *Archives of General Psychiatry* **32**: 765–7

R. D. Cartwright

Sleep States and Somatomotor Activity

The drive to sleep has an awesome power over our lives. If we go without sleep or drastically reduce it, the desire to sleep quickly becomes more important than life itself. The willingness to go to extraordinary efforts in order to obtain even a little sleep demonstrates vividly that sleep is a vital and necessary behavior. In fact, we often allow ourselves to be placed in life-threatening situations in order to satisfy the need to sleep.

But what is sleep, really? All of us feel, at some basic level, that we really do understand what sleep is all about. We certainly know how it feels, and the 'meaning' of the word is generally accepted in ordinary conversation. But, in point of fact, no one knows what sleep really is. Once we proceed beyond a simple description of an apparently quiet state that is somehow different from wakefulness, we find that the following major questions about the nature of sleep are largely unanswered: Why do we sleep?; What is (are) the function(s) of sleep?; What are the mechanisms that initiate and sustain sleep?; and, How, when and why do we wake up?

In order to understand 'sleep,' we must clarify the key processes and mechanisms that initiate and maintain this state. From a purely descriptive point of view, we simply need to know what is going on in the brain

(and body) during sleep. But first, we must agree on how to describe and/or define those behaviors that we intuitively understand as constituting the sleep states.

1. NREM and REM Sleep

Traditionally, three physiological measures are employed for describing the states of sleep in controlled laboratory situations. They are the EEG (electroencephalogram, which reflects the electrical activity of the cerebral cortex), the EOG (electrooculogram, which is a recording of eye movements), and the EMG (electromyogram, which is an indication of the degree of muscle activity, i.e., contraction). Together, these measures are used to define and differentiate sleep in mammals, which is comprised of two very different states. These states are called nonrapid eye movement (NREM) sleep and rapid eye movement (REM) sleep: each is nearly as different from the other as both are distinct from wakefulness.

Generally speaking, the EEG during NREM sleep consists of slow large-amplitude waves. Involuntary, slow, rolling, eye movements occur during the transition from drowsy wakefulness to NREM sleep; otherwise, NREM essentially lacks eye movements. Few motor events occur during NREM sleep; however, body repositioning and occasionally some motor behavior, such as sleepwalking, talking, eating or cooking, take place during NREM sleep. In general, motor processes that occur during NREM sleep are comparable to those present during very relaxed wakefulness.

The cortical EEG during REM sleep closely resembles the EEG of active wakefulness, i.e., it is of low voltage and high frequency. This is a surprising finding, considering that these two states are so dramatically different from a behavioral point of view. Bursts of rapid eye movements occur phasically during REM sleep. These eye movements are similar to those which would occur during wakefulness if one were looking at an object that went from right to left and then from left to right, rapidly and continuously, for a couple of seconds at a time. During REM sleep there is also a nonreciprocal flaccid paralysis, i.e., atonia, of the major muscle groups (with the principal exception being certain muscles used in respiration). However, at intervals that usually coincide with the phasic periods of rapid eye movements, there are brief muscle twitches that involve principally, but not exclusively, the distal muscles (e.g., the muscles of the fingers, toes, and face).

REM sleep is subdivided into 'phasic' and 'tonic' periods. During phasic REM sleep periods there are brief episodes of rapid eye movements and muscle twitches as well as muscle atonia (Fig. 1). Tonic REM

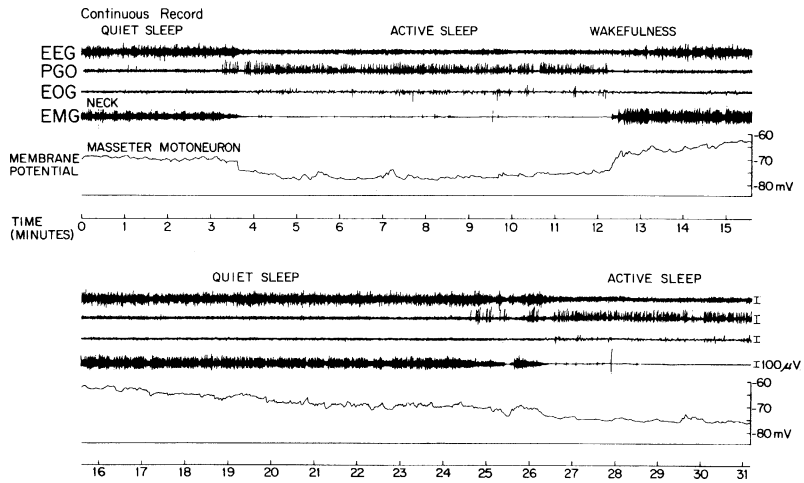

Figure 1
Intracellular recording from a trigeminal jaw-closer motoneuron: correlation of membrane potential and state changes. The membrane potential increased rather abruptly at 3.5 min in conjunction with the decrease in neck muscle tone and transition from quiet (NREM) to active (REM) sleep. At 12.5 min, the membrane depolarized and the animal awakened. After the animal passed into quiet sleep again, a brief, aborted episode of active sleep occurred at 25.5 min that was accompanied by a phasic period of hyperpolarization. A minute later the animal once again entered active sleep, and the membrane potential increased. EEG trace, marginal cortex, membrane potential band pass on polygraphic record, DC to 0.1 Hz. PGO, ponto-geniculo-occipital potential (reprinted from Chase MH, Chandler SH, Nakamura Y 1980 Intracellular determination of membrane potential of trigeminal motoneurons during sleep and wakefulness. *Journal of Neurophysiology* **44**: 349–58)

periods occur when the preceding phasically occurring eye and muscle twitches are absent, but there is still persistent muscle atonia.

2. *Somatomotor Activity During REM Sleep*

Because there is very little motor behavior during NREM sleep, the present overview of somatomotor activity during sleep consists mainly of a description of motor events which take place during REM sleep and entails: (a) an exploration of the mechanisms that control muscle activity during this sleep state, and (b) a description of the 'executive' mechanisms that are responsible for these REM sleep-related patterns of motor control (Chase and Morales 1994).

In order to understand the mechanisms responsible for the control of motor activity during REM sleep, it is important to first describe the changes in muscle fiber activity that occur during this state. The passage from wakefulness to NREM sleep is accompanied by a decrease in the degree of contraction of somatic muscle fibers, that is, there is a decrease in muscle tone, i.e., hypotonia. Surprisingly, during the state of REM sleep there occurs somatomotor atonia, or the complete lack of tone of many somatic muscles (Fig. 1). In order to understand how atonia of the somatic musculature during REM sleep is achieved, it is necessary to understand how the activity of muscle fibers are, in general, controlled.

First of all, muscles (i.e., muscle fibers) do one thing, they contract/shorten. When muscle fibers contract, they do so because command signals, which are really trains of action potentials, travel from the cell bodies of motoneurons along their axons to activate muscle fibers by changing the permeability of the muscle fiber membrane. A great many muscle groups are constantly contracting, at least to some degree, when we maintain a specific posture or perform a movement when we are awake. Thus, *tone* or some degree of muscle fiber contraction is dependent on the asynchronous, sustained discharge of motoneurons and the action potentials that travel down their axons to initiate the contraction of muscle fibers.

There is a gradual, but slight decline in muscle tone during NREM sleep compared with wakefulness. During REM sleep, there is such a strikingly potent suppression of muscle activity and motoneuron discharge. This results in a complete loss of muscle tone, or atonia. Thus, the key to understanding atonia during REM sleep resides in understanding the manner in which the activity of motoneurons is reduced or eliminated during this state.

There are two basic mechanisms that could, theoretically, be responsible for the elimination of motoneuron discharge during REM sleep: one is the direct inhibition of motoneurons (i.e., postsynaptic inhibition) and the second is a cessation of excitatory input to motoneurons (i.e., disfacilitation). Experimentally, these processes have been explored in the laboratory by recording, intracellularly, from individual motoneurons during REM sleep.

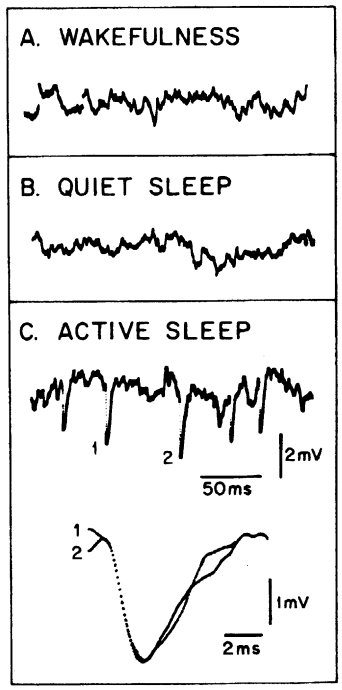

Figure 2
High-gain intracellular recording of the membrane potential activity of a tibial motoneuron during wakefulness (A), quiet (NREM) sleep (B), and active (REM) sleep (C). Note the appearance, *de novo* during active sleep, of large-amplitude, repetitively occurring inhibitory postsynaptic potentials. Two representative potentials, which were aligned by their origins, are shown at higher gain and at an expanded time base (C1, 2). These potentials were photographed from the screen of a digital oscilloscope. The analog to digital conversion rate was 50 μs/min. During these recordings the membrane potential during active sleep was −67.0 mV; the antidromic action potential was 78.5 mV (reprinted from Morales FR, Boxer P, Chase MH 1987 Behavioral state-specific inhibitory postsynaptic potentials impinge on cat lumbar motoneurons during active sleep. *Experimental Neurology* **98**: 418–35)

During REM sleep, it has been found that motoneurons become hyperpolarized, and as a consequence, they become relatively unresponsive to excitatory input. Analyses of the various membrane properties of these motoneurons indicate that their lack of responsiveness is due primarily to postsynaptic inhibition (Fig. 1). Thus, it is clear that postsynaptic

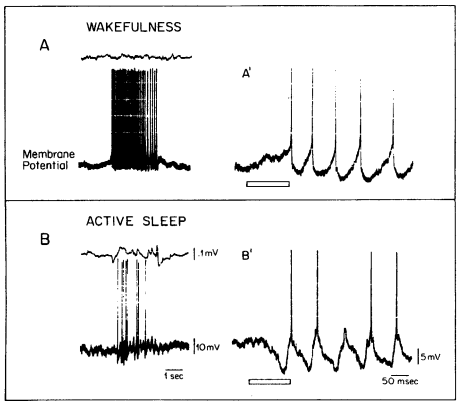

Figure 3
Action potential generation during wakefulness (A) and a rapid eye movement period of active (REM) sleep (B). Gradual membrane depolarization (bar in A′) preceded the development of action potentials during wakefulness, whereas a strong hyperpolarizing drive was evident during a comparable period of active sleep (bar in B′). This difference was also observed preceding the subsequent generation of each action potential during the rapid eye movement episode. Action potentials in A′ and B′ are truncated owing to the high gain of the recording. Resting membrane potentials are −58 mV in (A) and −63 mV in (B). Data are unfiltered; records were obtained from a single tibial motoneuron (reprinted from Chase MH, Morales FR 1982 Phasic changes in motoneuron membrane potential during REM periods of active sleep. *Neuroscience Letters* **34**: 177–82)

inhibition is the primary mechanism that is responsible for atonia of the somatic musculature during this state.

Postsynaptic inhibition of motoneurons during REM sleep occurs when certain neurons which are located in the brainstem liberate glycine at their point of contact with motoneurons, which is called the synapse. When glycine is liberated, motoneurons respond by generating changes in their membrane properties and their ability to discharge, which are recorded as 'inhibitory' postsynaptic potentials.

In an effort to elucidate the bases for motoneuron inhibition (which is viewed behaviorally as atonia) during REM sleep, spontaneous postsynaptic potentials, recorded from motoneurons, were examined. A unique pattern of sleep-specific inhibitory postsynaptic potentials was found to bombard motoneurons during REM sleep; they were of very large amplitude and occurred in abundant numbers (Fig. 2). Thus, it is clear that there is a unique set of presynaptic cells that are responsible for the REM sleep-specific inhibitory postsynaptic potentials that bombard motoneurons. For brainstem motoneurons (and likely for spinal cord motoneurons as well), these presynaptic inhibitory neurons are located in the ventromedial medulla.

In summary, postsynaptic inhibition is the principal process that is responsible for the atonia of the somatic musculature during REM sleep. This postsynaptic process is dependent on the presence of REM sleep-specific inhibitory potentials, which arise because glycine (an inhibitory neurotransmitter) is released by presynaptic neurons onto the surface of postsynaptic motoneurons.

All of the inhibitory phenomena that have been described in the previous sections are not only present but also enhanced during the phasic rapid eye movement periods of REM sleep. How, then, could there be twitches and jerks of the eyes and limbs during phasic rapid eye movement periods when there is also enhanced inhibition? The answer is simple: most of these periods are accompanied not only by increased motoneuron inhibition but also by potent motor excitatory drives that impinge on motoneurons. Thus, during these phasic periods of REM sleep, even though there is a suppression of the activity of motoneurons, there occurs, paradoxically, the concurrent excitation of motoneurons whose discharge results in the activation of muscle fibers (which leads to twitches and jerks, mainly of the limbs and fingers) (Fig. 3). These patterns of activation reflect descending excitatory activity emanating from different nuclei in the pons and possibly from the forebrain as well.

Consequently, from time to time, and for reasons as yet unknown, during REM sleep excitatory drives overpower the enhanced inhibitory drives. When this occurs, motoneurons discharge and the muscle fibers that they innervate contract. Thus, there is an increase in excitatory drives that is actually accompanied by an increase in inhibitory drives (Fig. 3). When motoneurons discharge during the REM periods of active sleep, contraction of the muscles that they innervate is unusual because the resultant movements are abrupt, twitchy, and jerky; they are also without apparent purpose.

Thus, from the perspective of motoneurons, REM sleep can be characterized as a state abundant in the availability of strikingly potent patterns of postsynaptic inhibition and, during REM periods, by new periods of excitation and also by enhanced postsynaptic inhibition.

3. The Executive Control Mechanism for Somatomotor Activity During REM Sleep

The preceding description of the neural control of somatomotor control during REM sleep can also be viewed from the perspective of the central nervous system executive mechanisms that are responsible this physiological pattern of activity. A critical structure

that is involved in the generation and/or maintenance of REM sleep, and especially, the motor inhibitory aspects of this state, is the nucleus pontis oralis, which is located in the rostral portion of the pontine reticular tegmentum. This area is thought to be activated by cholinergic fibers that arise from cell bodies that are located in the laterodorsal pontine tegmentum and the pedunculo pontine tegmentum.

There is evidence that a neuronal mechanism that resides within or is part of the nucleus pontis oralis is involved in the generation of wakefulness (and motor excitation) as well as REM sleep (and motor inhibition). For example, studies of the nucleus pontis oralis have shown that it is central to the phenomenon of reticular response reversal. This is a phenomenon wherein stimulation of this pontine nucleus results in an increase in somatic reflex activity during wakefulness, but remarkably, the identical stimulus during REM sleep yields potent postsynaptic inhibition of the same somatic reflexes.

We have suggested the existence of a neuronal switch, based on the phenomena of reticular response reversal, that is responsible for controlling the animal's state and somatomotor inhibitory processes during wakefulness as well as REM sleep. The existence of this switch is based on the hypothesis that wakefulness occurs when inhibitory (GABAergic) neurons in the nucleus pontis oralis suppress the activity of REM sleep neurons that are also located in, or in the vicinity of, the nucleus pontis oralis (Xi et al. 1999). Thus, when REM sleep-controlling neurons discharge in this and related areas the result is the generation of this state and its attendant patterns of somatomotor inhibition.

In support of the preceding hypothesis, we have found that when the inhibitory neurotransmitter GABA is placed in the nucleus pontis oralis, prolonged periods of wakefulness and heightened motor activity are elicited in cats. Conversely, the application of bicuculline, a $GABA_A$ antagonist, results in the occurrence of episodes of REM sleep and somatomotor inhibition of very long duration. We therefore conclude that a pontine GABAergic system plays a critical role in the control of wakefulness and REM sleep.

There are also recent data which demonstrates that when cells of the nucleus pontis oralis discharge, they initiate a cascade of events, the first one being the excitation of a group of premotor inhibitory neurons in the lower part of the brainstem, i.e., in the medulla (Morales et al. 1999). Fibers from these medullary neurons release glycine onto motoneurons, which results in atonia of the somatic musculature.

When the motor inhibitory mechanisms of REM sleep cease to function properly, and there is still ongoing motoneuron excitation (and/or enhanced motor excitation during the phasic periods of REM sleep), a syndrome occurs in cats that is called 'REM without atonia.' In humans, a comparable syndrome is called 'REM Behavior Disorder.' When humans and cats with this syndrome enter REM sleep, they begin to twitch violently, jump about and appear to act out their dreams; their sleep is quite disrupted and they are certainly a threat both to themselves and, of course, to others.

It is clear that the motor inhibition that occurs during REM sleep plays a critical role in the maintenance of this state. There is also no doubt that the simple release of glycine onto motoneurons during REM sleep, which results in atonia, i.e., a lack of muscle activity, has importance and implications of far-reaching and widespread proportions.

Bibliography

Chase M H, Chandler S H, Nakamura Y 1980 Intracellular determination of membrane potential of trigeminal motoneurons during sleep and wakefulness. *Journal of Neurophysiology* **44**: 349–58

Chase M H, Morales F R 1982 Phasic changes in motoneuron membrane potential during REM periods of active sleep. *Neuroscience Letters* **34**: 177–82

Chase M H, Morales F R 1994 The control of motoneurons during sleep. In: Kryger M H, Roth T, Dement W C (eds.) *Principles and Practice of Sleep Medicine*, 2nd edn. W.B. Saunders, Philadelphia, PA, pp. 163–75

Morales F R, Boxer P, Chase M H 1987 Behavioral state-specific inhibitory postsynaptic potentials impinge on cat lumbar motoneurons during active sleep. *Experimental Neurology* **98**: 418–35

Morales F R, Sampogna S, Yamuy J, Chase M H 1999 *c-fos* expression in brainstem premotor interneurons during cholinergically-induced active sleep in the cat. *Journal of Neuroscience* **19**: 9508–18

Xi M-C, Morales F R, Chase M H 1999 Evidence that wakefulness and REM sleep are controlled by a GABAergic pontine mechanism. *Journal of Neurophysiology* **82**: 2015–19

M. H. Chase, J. Yamuy and F. R. Morales

Sleep: Neural Systems

Since we spend roughly one third of our lives asleep, it is remarkable that so little attention has been paid to the capacity of sleep to organize the social behavior of animals including humans. To grasp this point, try to imagine life without sleep: no bedrooms and no beds; no time out for the weary parent of a newborn infant; and nothing to do all through the long, cold, dark winter nights. Besides demonstrating the poverty of the sociobiology of sleep, this thought experiment serves to cast the biological details which follow in the broader terms of social adaptation. Family life, reproduction, childrearing, and even economic commerce all owe their temporal organization to the power of sleep. The dependence of these aspects of our

lives on sleep also underlines its importance to deeper aspects of biological adaptation—like energy regulation—that are only beginning to be understood by modern scientists.

1. Behavioral Physiology of Sleep

Sleep is a behavioral state of homeothermic vertebrate mammals defined by: (a) characteristic relaxation of posture; (b) raised sensory thresholds; and (c) distinctive electrographic signs.

Sleep tends to occur at certain times of day and becomes more probable as sleeplessness is prolonged. These two organizing principles are the pillars of Alex Borbely's two-factor model: factor one captures the temporal or 'circadian' aspect while factor two captures the energetic or 'homeostatic' aspect (see *Sleep Behavior*). Sleep is usually associated with a marked diminution of motor activity and with the assumption of recumbent postures. Typically the eyes close and the somatic musculature becomes hypotonic. The threshold to external stimulation increases and animals become progressively less responsive to external stimuli as sleep deepens.

The differentiation of sleep from states of torpor in animals that cannot regulate core body temperature has an important phylogenetic correlation with the neural structures mediating the electrographic signs of sleep. These include the cerebral cortex and thalamus whose complex evolution underlies the distinctive electroencephalographic features of sleep in the higher vertebrate mammals. Sleep also constitutes the state of entry to and exit from hibernation in mammalian species that regulate temperature at lower levels during winter.

In humans, who can report upon the subjective concomitants of these outwardly observable signs of sleep, it is now clear that mental activity undergoes a progressive and systematic reorganization throughout sleep. On first falling asleep, individuals may progressively lose awareness of the outside world and experience microhallucinations and illusions of movement of the body in space; after sleep onset, mental activity persists but is described as thoughtlike and perseverative, if it can be recalled at all upon awakening. (For a description of the subsequent changes in mental activity, see the entry on dreaming.)

2. Electro-physiological Aspects of Sleep

There is a complex organization of behavioral, physiological, and psychological events within each sleep bout. To detect this organization, it is necessary to record the electroencephalogram (EEG) from the surface of the head (or directly from the cortical structures of the brain), to record the movement of the eyes by means of the electrooculogram (EOG), and to record muscle tone by means of the electromyogram (EMG). These three electrographic parameters allow one to distinguish sleep from waking and to distinguish two distinctive and cyclically recurrent phases within sleep: NREM (non-rapid eye movement) and REM (rapid eye movement) sleep. NREM, also known as synchronized or quiet sleep, is characterized by a change in the EEG from a low-amplitude, high-frequency to a high-amplitude, low-frequency pattern (see Fig. 1a).

The degree to which the EEG is progressively synchronized (that is, of high voltage and low frequency) can be subdivided into four stages in humans: In stage one, the EEG slows to the theta frequency range (4–7 cycles per second or cps) and is of low voltage (< 50 mV and arrhythmic). Stage two is characterized by the evolution of distinctive sleep spindles composed of augmenting and decrementing waves at a frequency of 12–15 cps and peak amplitudes of 100 mV. Stage three is demarcated by the addition to the spindling pattern of high voltage (> 100 mV) slow waves (1–4 cps), with no more than 50 percent of the record occupied by the latter. In stage four, the record is dominated by high-voltage (150–250 mV) slow waves (1–3 cps). At the same time that the EEG frequency is decreasing and the voltage increasing, muscle tone progressively declines and may be lost in most of the somatic musculature. Slow rolling eye movements first replace rapid saccadic eye movements of waking and then also subside, with the eyes finally assuming a divergent upward gaze (Fig. 1a).

After varying amounts of time, the progressive set of changes in the EEG reverses itself and the EEG finally resumes the low-voltage, fast character previously seen in waking. Instead of waking, however, behavioral sleep persists; muscle tone, at first passively decreased, is now actively inhibited; and there arise in the electrooculogram stereotyped bursts of saccadic eye movement called rapid eye movements (the REMs, which give this sleep state the name REM sleep). Major body movements are quite likely to occur during the NREM–REM transition. The REM phase of sleep has also been called activated sleep (to signal the EEG desynchronization) and paradoxical sleep (to signal the maintenance of increased threshold to arousal in spite of the activated brain). Human consciousness is associated with the low-voltage, fast EEG activity of waking and REM sleep but its unique character in dreaming depends upon other aspects of REM neurophysiology (see Fig. 1b).

In all mammals (including aquatic, arboreal, and flying species) sleep is organized in this cyclic fashion: sleep is initiated by NREM and punctuated by REM at regular intervals. Most animals compose a sleep bout out of three or more such cycles, and in mature humans the average nocturnal sleep period consists of four to five such cycles, each of 90–100 minutes

(a)

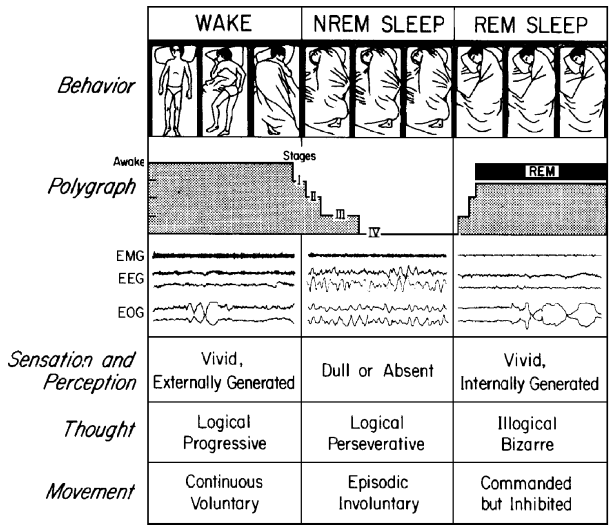

(b)

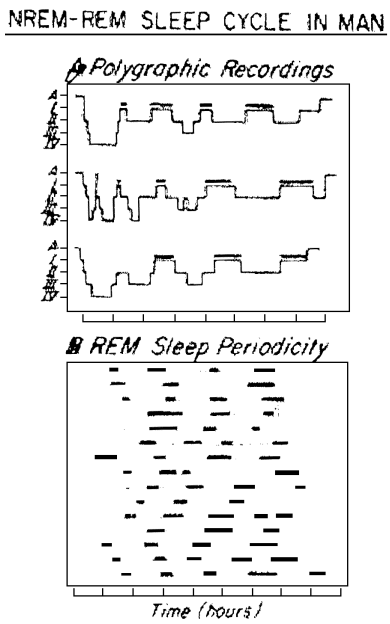

Figure 1
(a) Behavioral states in humans. The states of waking, NREM, and REM sleep have behavioral, polygraphic, and psychological manifestations which are depicted here. In the behavioral channel, posture shifts- detectable by time lapse photography or video- can be seen to occur during waking and in concert with phase changes of the sleep cycle. Two different mechanisms account for sleep immobility: disfacilitation (during stages I–IV of NREM sleep) and inhibition (during REM sleep). In dreams, we imagine that we move, but we do not. The sequence of these stages is schematically represented in the polygraph channel and sample tracings are also shown. Three variables are used to distinguish these states: the electromyogram (EMG) which is highest in waking, intermediate in NREM sleep, and lowest in REM sleep; the electroencephalogram (EEG) and electrooculogram (EOG) which are both activated in waking and REM sleep, and inactivated in NREM sleep. Each sample record is about 20 s long. Other subjective and objective state variables are described in the three lower channels. (*b*) Sample records of human sleep (Snyder and Hobson)

duration. After a prolonged period of wake activity (as in humans) the first cycles are characterized by a preponderance of high-voltage, slow wave activity (i.e., the NREM phase is enhanced early in the sleep bout) while the last cycles show more low-voltage, fast wave activity (i.e., the REM phase is enhanced late in the sleep bout). The period length is fixed across any and all sleep periods in a given individual but is shorter in immature and smaller animals indicating a correlation of NREM–REM cycle length with brain size.

3. *Neural Mechanisms of Sleep*

In all animals, sleep is one of a number of circadian behavioral functions controlled by an oscillator in the suprachiasmatic nucleus of the hypothalamus. It is the interaction of the intrinsic propensity to sleep at about 24 hour (i.e., circadian) intervals with the daily cycles of extrinsic light and temperature that gives sleep its diurnal organization. The exact mechanism by which hypothalamic structures act to potentiate sleep at one or another diurnal period is unknown, but increasing evidence suggests that peptide hormones are involved in regulating this coordination.

The hypothesis that sleep is hormonally regulated is an old one and is linked to the subjective impression of fatigue with its implied humoral basis. The fact that the sleep drive increases with the duration of time spent awake in both normal and sleep-deprived animals clearly indicates the homeostatic regulation function of sleep, and experiments begun by Pieron at the beginning of the twentieth century suggested that a circulating humoral factor might mediate this propensity. Subsequent efforts to test the humoral hypothesis have involved the isolation of an S-muramyl peptide as a slow wave sleep-enhancing factor found both in the cerebrospinal fluid of sleep-deprived animals and in normal human urine. It remains to be established that this so-called factor S is involved in physiological sleep mediation; this is most important since factor S is known to be a product of bacterial cell walls but is not produced by the brain.

Whatever its mediating mechanism, the onset of sleep has long been linked by physiologists to the concept of deafferentation of the brain. In early articulations of the deafferentation hypothesis it was thought that the brain essentially shut down when deprived of external inputs. This idea was put forth by Frederic Bremer to explain the results of midcollicular brain stem transection after which the forebrain persistently manifested the EEG synchronized pattern of slow wave sleep. Bremer thought that the so-called *cerveau isolé* (or isolated forebrain) preparation was asleep because it had been deprived of its afferent input. That interpretation was upset by Bremer's own subsequent finding that transection of the neuraxis at C-1 produced intense EEG desynchronization and hypervigilance in the so-called *encephale isolé* (or isolated brain) preparation.

The possibility that active neural mediation of arousal by brain stem structures situated somewhere between the midcollicular and high spinal levels was thus suggested. The faint possibility that afferent input (via the trigeminal nerve) might account for the difference between the two preparations was eliminated by Giuseppe Moruzzi's finding that the mediopontine pretrigeminal transected preparation was also hyperalert. Thus the notion of sleep mediation via passive deafferentation was gradually replaced by the idea of sleep onset control via the active intervention of brain stem neural structures. This new idea does not preclude a contribution from deafferentation. And indeed, they appear to be complementary processes.

The clear articulation of the concept of active brain stem control of the states of sleep and waking came from the 1949 finding by Giuseppe Moruzzi and Horace Magoun that high-frequency stimulation of the mid-brain reticular formation produced immediate desynchronization of the electroencephalogram and behavioral arousal in drowsy or sleeping animals. The most effective part of the reticular activating system was its rostral pole in the diencephalon where the circadian clock is located.

These data indicated that wakefulness was actively maintained by the tonic firing of neurons in an extensive reticular domain. Arousal effects could be obtained following high-frequency stimulation anywhere in the reticular formation from the medullary to the pontine to the midbrain and diencephalic levels. The activating effect was thought to be mediated by neurons in the thalamus, which in turn relayed tonic activation to the entire neocortex. This concept has been recently confirmed at the cellular level by Mircea Steriade, following up on the earlier work of Dominic Purpura which had indicated that the spindles and slow waves of NREM sleep were a function of a thalamocortical neuronal interaction that appeared whenever brain stem activation subsided.

Once Moruzzi and Magoun had substantiated the concept of active neural control of the states of waking and sleep, other evidence supporting this notion promptly followed. For example, it was found by Barry Sterman and Carmine Clemente that high-frequency stimulation of the basal forebrain could produce slow wave sleep in waking animals. When the same region was lesioned by Walle Nauta, arousal was enhanced. The cellular basis of these basal forebrain effects and their possible link to both the circadian oscillator of the hypothalamus and the arousal system of the reticular activating system are currently under intense investigation.

The second dramatic demonstration that sleep states were under active neural control was Michel Jouvet's lesion and transection work suggesting a role of the pontine brain stem in the timing and triggering of the REM phase of sleep. The pontine cat preparation

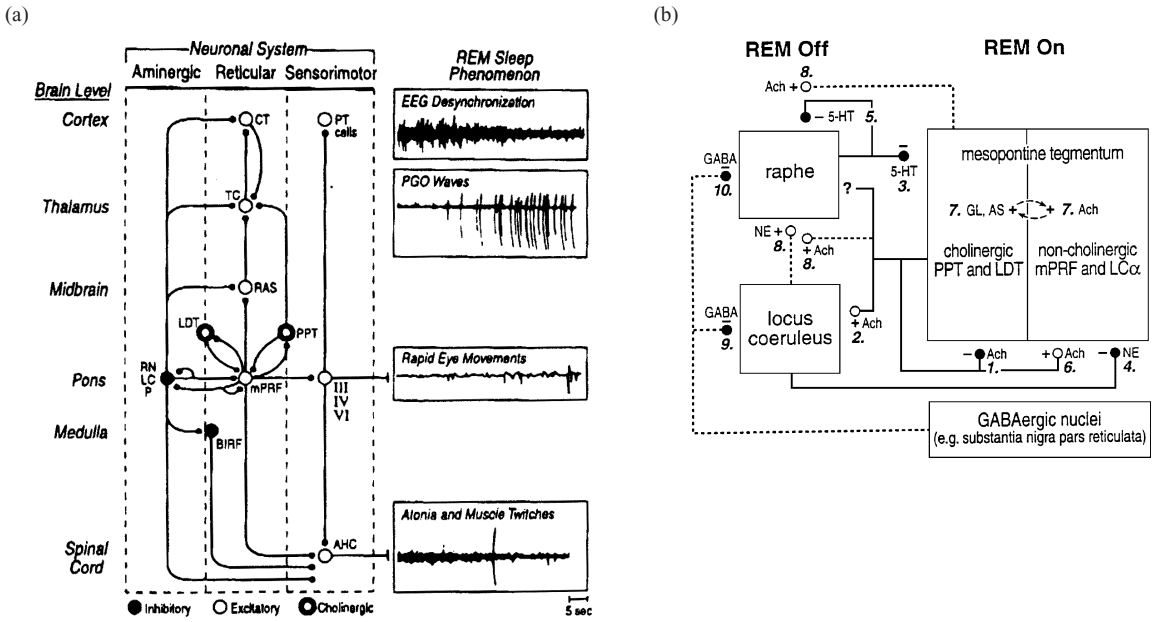

Figure 2
(a) Schematic representation of the REM sleep generation process. A distributed network involves cells at many brain levels (left). The network is represented as comprising 3 neuronal systems (center) that mediate REM sleep electrographic phenomena (right). Postulated inhibitory connections are shown as solid circles; postulated excitatory connections as open circles. In this diagram no distinction is made between neurotransmission and neuromodulatory functions of the depicted neurons. It should be noted that the actual synaptic signs of many of the aminergic and reticular pathways remain to be demonstrated, and, in many cases, the neuronal architecture is known to be far more complex than indicated here (e.g., the thalamus and cortex). Two additive effects of the marked reduction in firing rate by aminergic neurons at REM sleep onset are postulated: disinhibition (through removal of negative restraint) and facilitation (through positive feedback). The net result is strong tonic and phasic activation of reticular and sensorimotor neurons in REM sleep. REM sleep phenomena are postulated to be mediated as follows: EEG desynchronization results from a net tonic increase in reticular, thalamocortical, and cortical neuronal firing rates. PGO waves are the result of tonic disinhibition and phasic excitation orf burst cells in the lateral pontomescencephalic tegmentum. Rapid eye movements are the consequence of phasic firing by reticular and vestibular cells; the latter (not shown) directly excite oculomotor neurons. Muscular atonia is the consequence of tonic postsynaptic inhibition of spinal anterior horn cells by the pontomedullary reticular formation. Muscle twitches occur when excitation by reticular and pyramidal tract motorneurons phasically overcomes the tonic inhibition of the anterior horn cells. Anatomical abbreviations: RN, raphe nuclei; LC, locus coeruleus; P, peribrachial region; FTG, gigantocellular tegmental field; FTC, central tegmental field; FTP, parovocellular tegmental field; FTM, magnocellular tegmental field; TC, thalamocortical; CT, cortical; PT cell, pyramidal cell; III, oculomotor; IV, trochlear; V, trigmenial motor nuclei; AHC, anterior horn cell. (Modified from Hobson et al. 1986)
(b) Synaptic modifications of the original reciprocal interaction model based upon recent findings. Reported data from animal (cat and rodent) are shown as solid lines and some of the recently proposed putative dynamic relationships are shown as dotted lines. The exponential magnification of cholinergic output predicted by the original model can also occur in this model with mutually excitatory cholinergic-non-cholinergic interactions (7.) taking the place of the previously postulated, mutually excitatory cholinergic-cholinergic interactions. The additional synaptic details can be superimposed on this revised reciprocal interaction model without altering the basic effects of aminergic and cholinergic influences on the REM sleep cycle. For example: i. Excitatory cholinergic-non-cholinergic interactions utilizing ACh and the excitatory amino acid transmitters enhance firing of REM-on cells (6., 7.) while inhibitory noradrenergic (4.), serotonergic (3.) and autoreceptor cholinergic (1.) interactions suppress REM-on cells. ii. Cholinergic effects upon aminergic neurons are both excitatory (2.), as hypothesized in the original reciprocal interaction model and may also operate via presynaptic influences on noradrenergic-serotonergic as well as serotonergic-serotonergic circuits (8.). iii. Inhibitory cholinergic autoreceptors (1.) could contribute to the inhibition of LDT and PPT cholinergic neurons which is also caused by noradrenergic (4.) and serotonergic (3.) inputs. iv. GABAergic influences (9., 10.) as well as other neurotransmitters such as adenosine and

(with ablation of all forebrain structures above the level of the pontine tegmentum) continued to show periodic abolition of muscle tone and rapid eye movements at a frequency identical to the REM periods of normal sleep. The implications were (a) that the necessary and sufficient neurons for timing and triggering by REM were in the pons and (b) that this system became free-running when the pontine generator was relieved of circadian restraint from the hypothalamus.

Small lesions placed in the pontine tegmentum between the locus coeruleus and the pontine reticular formation resulted in periods of REM sleep without atonia: the cats exhibited all the manifestations of REM sleep except the abolition of muscle tone, indicating that intrapontine connections were essential to mediate and coordinate different aspects of the REM sleep period.

4. *A Cellular and Molecular Model of the Sleep–Wake Cycle*

Microelectrode studies designed to determine the cellular mechanism of these effects have indicated that REM sleep is actively triggered by an extensive set of executive neurons ultimately involving at least the oculomotor, the vestibular, and the midbrain, pontine, and medullary reticular nuclei (see Fig. 2). The intrinsic excitability of this neuronal system appears to be lowest in waking, to increase progressively throughout NREM sleep, and to reach its peak during REM sleep. Cells in this system have thus been designated REM-on cells. In contrast, neurons in the aminergic brain stem nuclei (the serotoninergic dorsal raphe and the catacholaminergic locus coeruleus and peribrachial region) have reciprocally opposite activity curves. These aminergic REM-off cells fire maximally in waking, decrease activity in NREM sleep, and reach a low point at REM onset.

The reciprocal interaction model of state control ascribes the progressive excitability of the executive (REM-on) cell population to disinhibition by the modulation (REM-off) cell population. How the aminergic systems are turned off is unknown, but Cliff Saper has recently suggested that GABA-ergic inhibition may arise from neurones in the hypothalamic circadian control system.

5. *Cholinergic REM Sleep Mediation*

Confirming the early studies of Jouvet and Raul Hernandez-Peon, numerous investigations have found prompt and sustained increases in REM sleep signs when cholinergic agonist drugs are microinjected into the pontine brain stem of cats. The behavioral syndrome produced by this treatment is indistinguishable from the physiological state of REM sleep, except that it is precipitated directly from waking with no intervening NREM sleep. The cats can be aroused but revert to sleep immediately when stimulation abates. This drug-induced state has all the physiological signs of REM sleep (such as the low voltage fast EEG, EMG atonia, REM, and ponto-geniculo-occipital (PGO) waves), suggesting that it is a valid experimental model for physiological REM. Humans who are administered cholinergic agonist by intravenous injection during the first NREM cycle also show potentiation of REM sleep; this cholinergically enhanced REM sleep is associated, as usual, with dreaming.

Mixed nicotinic and muscarinic agonists (e.g., carbachol), pure muscarinic agonists (e.g., bethanechol) and dioxylane are equally effective potentiators of REM sleep and recent results suggest that activation of the M2 acetylcholine receptor suffices to trigger REM sleep behavior. The inference that acetylcholine induces REM sleep is suggested by Helen Baghdoyan's finding that neostigmine, (an acetylcholinesterase inhibitor which prevents the breakdown of endogenously released acetylcholine) also potentiates sleep after some delay. All of these effects are dose-dependent and are competitively inhibited by the acetylcholine antagonist atropine.

REM sleep induction is obtained only by injection of cholinergic agonists into the pons: Midbrain and medullary injections produce intense arousal with incessant turning and barrel-rolling behavior, respectively. Within the pons, the response pattern differs according to the site of injection, with maximal effects obtained from a region anterodorsal to the pontine reticular formation and bounded by the dorsal raphe and the locus coeruleus nuclei.

In the peribrachial pons, where both aminergic and cholinergic neurons are intermingled, carbachol injection produces continuous PGO waves by activating the cholinergic PGO burst cells of the region. But agonist induction of these drug-induced waves is state-independent and REM sleep is not potentiated in the

Figure 2
(continued)
nitric oxide (see text) may contribute to the modulation of these interactions. Abbreviations: open circles, excitatory postsynaptic potentials; closed circles, inhibitory postsynaptic potentials; mPRF, medial pontine reticular formation; PPT, pedunculopontine tegmental nucleus; LDT, laterodorsal tegmental nucleus; LCα, peri-locus coeruleus alpha; 5HT, serotonin; NE, norepinephrine; ACh, acetylcholine; glut, glutamate; AS, aspartate; GABA, gamma-aminobutyric acid

first 24 hours following drug injection. It is only later that REM sleep increases reaching its three- to fourfold peak at 48–72 hours and thereafter remaining elevated for 6–10 days. By injection of carbachol into the sub- and peri-locus coeruleus regions of the pontine reticular formation, where tonic, neuronal activity can be recorded during REM sleep, atonia may be generated, while waking persists, suggesting a possible animal model of human cataplexy.

Cholinergic cells of the PGO wave generator region of the peribrachial pons are known to project to both the lateral geniculate nucleus and the perigeniculate sectors of the thalamic reticular nucleus, where they are probably responsible for the phasic excitation of neurons in REM sleep. This supposition is supported by the finding that the PGO waves of the LGN are blocked by nicotinic antagonists.

Wolf Singer's recent work suggests the functional significance of these internally generated signals. Singer proposes that the resulting resonance contributes to the use-dependent plasticity of visual cortical networks. Besides providing evidence that REM sleep signs are cholinergically mediated, these recent pharmacological results provide neurobiologists with a powerful experimental tool, since a REM-sleeplike state can be produced at will. Thus Francisco Morales and Michael Chase have shown that the atonia produced by carbachol injection produces the same electrophysiological changes in lumbar motoneurons as does REM sleep: a decrease in input resistance and membrane time constant, and a reduction of excitability associated with discrete IPSPs. These findings facilitate the exploration of other neurotransmitters involved in the motorneuronal inhibition during REM sleep which is mediated, at least in part, by glycine.

Another important advance is the intracellular recording of pontine reticular neurons in a slice preparation by Robert McCarley's group. Neuronal networks can be activated cholinergically, producing the same electrophysiological changes as those found in REM sleep of intact animals. A low threshold calcium spike has been identified as the mediator of firing pattern alterations in brain-stem slice preparations.

The activation process within the pons itself has also invited study, using carbachol as an inducer of REM sleep. Using the moveable microwire technique, Peter Shiromani and Dennis McGinty showed that many neurons normally activated in REM sleep either were not activated or were inactivated by the drug. Ken-Ichi Yamamoto confirmed this surprising finding and found that the proportion of neurons showing REM-sleeplike behavior was greatest at the most sensitive injection sites.

To achieve greater anatomical precision in localizing the carbachol injection site, James Quattrochi conjugated the cholinergic agonist carbachol to fluorescent microspheres, thereby reducing the rate of diffusion tenfold and allowing all neurons projecting to the injection site to be identified by retrograde labeling. Surprisingly, the behavioral effects are no less potent despite the reduced diffusion rate of the agonist. They are maximized by injection into the same anterodorsal site described by Helen Baghdoyan. That site receives input from all the major nuclei implicated in control of the sleep cycle; the pontine gigantocellular tegmental field (glutamatergic), the dorsal raphe nuclei (serotonergic), the locus coeruleus (noradrenergic), and the dorsolateral tegmental and pedunculo-pontine nuclei (cholinergic). Thus the sensitive zone appears to be a point of convergence of the neurons postulated to interact reciprocally in order to generate the NREM–REM sleep cycle. Whether this is also a focal point of reciprocal projection back to those input sites can now be investigated.

6. *Conclusions*

Because sleep occurs in a social context, and because sleep actively determines many aspects of that context, its study constitutes fertile ground for integration across domains of inquiry. At the basic science level, the cellular and molecular mechanisms emphasized here reach inexorably down, to the level of the genome. The upward extension of sleep psychophysiology to the psychology of individual consciousness, to dreaming, and hence to the mythology that guides cultures is also advancing rapidly. Altogether absent is a sound bridge to the social and behavioral realms where sleep has been ignored almost as if it were a non-behavior and hence devoid of social significance. Yet when, where, and especially with whom we sleep roots us profoundly in our domestic and interpersonal worlds.

See also: Circadian Rhythms; Dreaming, Neural Basis of; Hypothalamus; Sleep and Health; Sleep Disorders: Psychiatric Aspects; Sleep Disorders: Psychological Aspects; Sleep States and Somatomotor Activity

Bibliography

Hobson J A 1999 *Consciousness*. Scientific American Library

Datta S, Calvo J, Quattrochi J, Hobson J A 1991 Long-term enhancement of REM sleep following cholinergic stimulation. *NeuroReport* **2**: 619–22

Jones B E 1991 Paradoxical sleep and its chemical/structural substrates in the brain. *Neuroscience* **40**: 637–56

Gerber U, Stevens D R, McCarley R W, Greene R W 1991 Muscarinic agonists activate an inwardly rectifying potassium conductance in medial pontine reticular formation neurons of the rat *in vitro*. *Journal of Neuroscience* **11**: 3861–7

Steriade M, Dossi R C, Nunez A 1991 Network modulation of a slow intrinsic oscillation of cat thalamocortical neurons implicated in sleep delta waves: Cortically induced synchroniz-

ation and brainstem cholinergic suppression. *Journal of Neuroscience* **11**: 3200–17
Steriade M, McCarley R W 1990 *Brainstem Control of Wakefulness and Sleep*. Plenum Press, New York

J. A. Hobson

Small-group Interaction and Gender

When people interact together in groups that are small enough to allow person-to-person contact (2 to 20 people), regular patterns of behavior develop that organize their relations. The study of gender in this context examines how these patterns of behavior are affected by members' social locations as men or women and the consequences this has for beliefs about gender differences and for gender inequality in society.

1. *The Emergence of the Field*

The systematic, empirical study of small group interaction and gender (hereafter, gender and interaction) developed during the 1940s and 1950s out of the confluence of a general social scientific interest in small groups and structural-functionalist theorizing about the origin of differentiated gender roles for men and women. Parsons and Bales (1955) argued that small groups, like all social systems, must manage instrumental functions of adaptation and goal attainment while also attending to expressive functions of group integration and the well-being of members. Differentiated instrumental and expressive gender roles develop to solve this problem in the small group system of the family. Children internalize these functionally specialized roles as personality traits that shape their behavior in all groups.

Although eventually discredited on logical and empirical grounds, the functional account of gender roles stimulated broader empirical attention to gender and interaction, not only within the family but also in task-oriented groups such as committees and work groups. Evidence accumulated that gender's effects on interaction are complex and quite context specific (see Aries 1996; Deaux and LaFrance 1998). This evidence is inconsistent with the view that gender is best understood as stable individual traits that affect men and women's behavior in a consistent manner across situations.

These findings on interaction contributed to a gradual transformation of social scientific approaches to gender. From an earlier view of gender as a matter of individual personality and family relations, social science has increasingly approached gender as a broad system of social difference and inequality that is best studied as part of social stratification as well as individual development and family organization.

Considering gender in a broader context draws attention to how the interactional patterns it entails are both similar to and different from those that characterize other systems of difference and inequality such as those based on race, ethnicity, or wealth. While interaction occurs between the advantaged and the disadvantaged on each of these forms of social difference, the rate of interaction across the gender divide is distinctively high. Gender divides the population into social categories of nearly equal size; it cross-cuts kin and households and is central for reproduction. Each of these factors increases the frequency with which men and women interact and the intimacy of the terms on which they do so. Furthermore, research on social cognition has shown that people automatically sex categorize (i.e., label as male or female) any concrete person with whom they interact. As a consequence, gender is potentially at play in all interaction and interactional events are likely to be important for the maintenance or change of a society's cultural beliefs and practices about gender. West and Zimmerman (1987) argue that for gender to persist as a social phenomenon, people must continually 'do gender' by presenting themselves in ways that allow others to relate to them as men or women as that is culturally defined by society.

Recognition of these distinctive aspects of gender has increased attention to the role that gendered patterns of interaction play in gender inequality. Interaction mediates the process by which people form social bonds, are socially evaluated by others, gain influence, and are directed towards or away from positions of power and valued social rewards. Interaction also provides the contexts in which people develop identities of competence and sociality. To the extent that gender moderates these interaction processes, it shapes the outcomes of men and women as well as shared beliefs about gender.

2. *Current Theories*

Four theoretical approaches predominate. Two, social role theory and expectation states theory, single out a society's cultural beliefs about the nature and social value of men's and women's traits and competencies as primary factors that create gendered patterns of interaction in that society. The theories conceptualize these beliefs in slightly different terms (i.e., as gender stereotypes or gender status beliefs) but are in substantial agreement about their explanatory impact on interaction through the expectations for behavior that develop when these beliefs are evoked by the group situation. The remaining theories take somewhat different approaches.

2.1 *Social Role Theory*

Eagly's (1987) social role theory argues that widely shared gender stereotypes develop from the gender

division of labor that characterizes a society. In western societies, men's greater participation in paid positions of higher power and status and the disproportionate assignment of nurturant roles to women have created stereotypes that associate agency with men and communion with women. In addition, the gendered division of labor gives men and women differentiated skills. When gender stereotypes are salient in a group because of a mixed sex membership or a task or context that is culturally associated with one gender, stereotypes shape behavior directly through the expectations members form for one another's behavior. When group members enact social roles that are more tightly linked to the context than gender, such as manager and employee in the workplace, these more proximate roles control their behavior rather than gender stereotypes. Even in situations where gender stereotypes do not control behavior, however, men and women may still act slightly differently due to their gender differentiated skills.

Social role theory has a broad scope that applies to interaction in all contexts and addresses assertive, power related behaviors as well as supportive or feeling related behaviors (called socioemotional behaviors). The explanations offered by the theory are not highly specific or detailed, however. The theory predicts that women will generally act more communally and less instrumentally than men in the same context, that these differences will be greatest when gender is highly salient in the situation, and that gender differences will be weak or absent when people enact formal, institutional roles.

2.2 *Expectations States Theory*

Another major approach, Berger and co-worker's expectation states theory, offers more detailed explanations within a narrower scope. The theory addresses the hierarchies of influence and esteem that develop among group members in goal-oriented contexts and makes predictions about when and how gender will shape these hierarchies due to the status value gender carries in society (see Ridgeway 1993; Wagner and Berger 1997). It does not address socioemotional behavior. Gender status beliefs are cultural beliefs that one gender (men) is more status worthy and generally more competent than the other (women) in addition to each having gender specific competencies. When gender status beliefs become salient due to the group's mixed sex or gender associated context, they create implicit expectations in both men and women about the likely competence of goal oriented suggestions from a man compared to those from a similar woman. These often unconscious expectations shape men and women's propensity to offer their ideas to the group, to stick with those ideas when others disagree, to positively evaluate the ideas of others, and to accept or resist influence from others, creating a behavioral influence hierarchy that usually advantages men over women in the group.

In mixed sex groups with a gender-neutral task, the theory predicts that men will participate more assertively and be more influential than women. If the group task or context is culturally linked to men, their influence advantage over women will be stronger. If the task or context is associated with women's culturally expected competencies, however, the theory predicts that women will be somewhat more assertive and influential than men. There should be no gender differences in assertive influence behavior between men and women in same sex groups with a gender-neutral task, since gender status beliefs should not be salient.

2.3 *Structural Identity Theories*

A set of symbolic interactionist theories, including Heise and Smith-Lovin's affect control theory and Burke's identity theory forms the structural identity approach (see Ridgeway and Smith-Lovin (1999) for a review). It, too, emphasizes shared cultural meanings about gender but focuses on the identity standards those beliefs create for individuals in groups. People learn cultural meanings about what it is to be masculine or feminine and these meanings become a personal gender identity standard that they seek to maintain through their actions. Identity standards act like control systems that shape behavior. If the context of interaction causes a person to seem more masculine or feminine than his or her gender identity standard, the person reacts with compensatory behaviors (e.g., warm behaviors to correct a too masculine impression). Consequently, different actions serve to express and maintain gender identities in different situational contexts.

Since people automatically sex categorize one another, this approach assumes that gender identity standards affect behavior in all interaction, although the extent of their impact varies with gender's salience in the context. Gender is often a background identity that modifies other, more situationally prominent identities, such as woman judge. Unlike the other theories, the predictions of structural identity theories focus primarily on the behavioral reactions gender produces to events in small groups.

2.4 *Two-cultures Theory*

Maltz and Borker's (1982) two cultures theory, popularized by Tannen (1990), takes a different approach. It limits its scope to informal, friendly interaction. People learn rules for friendly conversation from peers in childhood, it argues. Since these peer groups tend to

be sex-segregated and because children exaggerate gender differences in the process of learning gender roles, boys and girls groups develop separate cultures that are gender-typed. Girls learn to use language to form bonds of closeness and equality, to criticize in nonchallenging ways, and to accurately interpret the intentions of others. Boys learn to use speech to compete for attention and assert positions of dominance. In adult mixed sex groups, these rules can cause miscommunication because men and women have learned to attribute different meanings to the same behavior. Men and women's efforts to accommodate each other in mixed-sex interaction, however, modifies their behavior slightly, reducing gender differences. In same sex interaction, gendered styles of interaction are reinforced. Thus, two cultures theory predicts greater gender differences in behavior between men and women in same-sex groups than in mixed sex groups. The theory has been criticized for ignoring status and power differences between men and women and oversimplifying childhood interaction patterns (see Aries 1996).

3. *Research Findings*

The body of systematic evidence about men and women's behaviors in small group interaction is large and growing. Several methodological concerns must be kept in mind in order to interpret this evidence and infer general patterns.

3.1 *Methodological Issues*

Interaction in small groups is an inherently local phenomenon that is embedded within larger socio-cultural structures and affected by many aspects of those structures besides gender. Three methodological problems result. First, care is required to ascertain that behavioral differences between men and women in a situation are indeed due to their gender and not to differences in their other roles, power, or statuses in the situation. Second, reasonably large samples of specific interactional behaviors are necessary to infer gendered patterns in their use. Third, attention must be paid to the specific cultural context within in which the group is interacting. At present, almost all systematic research has been based on small groups in the US composed predominately of white, middle class people. Since several theories emphasize the importance of cultural beliefs about gender in shaping interaction, researchers must be alert to subcultural and cross-cultural variations in these beliefs and appropriately condition their empirical generalizations about gendered patterns of interaction. The available studies that compare US populations such African–Americans whose gender beliefs are less polarized than the dominant beliefs find that gender differences in interaction are also less for these populations (Filardo 1996).

3.2 *Empirical Patterns in North American Groups*

Taking these methodological concerns into account, narrative and meta-analytic reviews of research suggest several provisional conclusions about gender and interaction in groups governed by the dominant culture of North American society. Aries (1996), Deaux and La France (1998), and Ridgeway and Smith-Lovin (1999 provide reviews of the research on which these conclusions are based.

Gender differences in behavior do not always occur in small groups and vary greatly by context. Behavioral expectations associated with the specific focus and institutional context of the small group (e.g., the workplace, a committee, a friendship group, a student group) generally are more powerful determinants of both men's and women's behavior than gender. When gender differences occur, they tend to be small or moderate in effect size, meaning that there is usually at least a 70 percent overlap in the distributions of men's and women's behavior.

When men and women are in formal, prescribed roles with the same power and status, there are few if any differences in their behavior. Research has shown that men and women in equivalent leadership or managerial roles interact similarly with subordinates of either sex. On the other hand, when women are gender atypical occupants of positions of power, they are sometimes perceived by others as less legitimate in those roles and elicit more negative evaluations when they behave in a highly directive, autocratic way than do equivalent men. These findings are in accord with the predictions of social role theory and expectations states theory.

Influence over others and assertive, goal-directed behavior such as participation rates, task suggestions, and assertive patterns of gestures and eye gaze are associated with power and leadership in small groups. In mixed sex groups with a gender-neutral task, men have moderately higher rates of assertive behaviors and influence than do women who are otherwise their peers. When the group task or context is culturally linked to men, this gender difference increases. When the task or context is one associated with women, however, women's rates of assertive behaviors and influence are slightly higher than men's. When performance information clearly demonstrates that women in a mixed sex group are as competent as the men, gender differences in assertiveness and influence disappear. In same sex groups, there are no differences between men's and women's rates of assertive behaviors or influence levels. These patterns closely match the predictions of expectations states theory, are

consistent with social role theory, and inconsistent with two cultures theory. They suggest that gender status beliefs in society and the expectations for competence in the situation that they create are an important determinant of gender differences in power and assertiveness in groups, independent of men and women's personalities or skills. They indicate as well that both men and women act assertively or deferentially depending on the situational context.

Men, like women, show higher rates of socioemotional behaviors when they are in subordinate rather than superordinate positions in groups. These are verbal and nonverbal behaviors that support the speech of others, express solidarity, and show active, attentive listenership. In mixed sex groups, women engage in slightly more socioemotional behavior than men. However, women engage in the highest rates of socioemotional behaviors, and men the lowest, in same sex groups. The latter findings are the only ones in partial accord with the two-cultures theory. That theory, however, does not account for the partial association of socioemotional behaviors with lower status positions.

Status factors alone, however, do not explain women's increased socioemotional behaviors in female groups. Assertive, instrumental behaviors appear to reflect power, competence, and status equally for men and women and, thus, do not reliably mark gender identity for the actor. To the extent that people signal gender identity consistently across interaction contexts, they appear to do so primarily through socioemotional behaviors that are less associated with instrumental outcomes.

4. Conclusions

Both the gender division of labor and gender inequality in a society depend on its cultural beliefs about the nature and social value of gender differences in competencies and traits. Such taken for granted beliefs allow actors to be reliably categorized as men and women in all contexts and understood as more or less appropriate candidates for different roles and positions in society. For such cultural beliefs to persist, people's everyday interactions must be organized to support them. The empirical evidence from North America suggests that unequal role and status relationships produce many differences in interactional behavior that are commonly attributed to gender. Network research suggests that most interactions between men and women actually occur within the structural context of unequal role or status relations (see Ridgeway and Smith-Lovin 1999). These points together may account for the fact that people perceive gender differences to be pervasive in interaction, while studies of actual interaction show few behavioral differences between men and women of equal status and power. Small group interaction is an arena in which the appearance of gender differences is continually constructed through power and status relations and identity marking in the socioemotional realm.

Theory and research on gender and interaction have focused on the way cultural beliefs about gender and structural roles shape interaction in ways that confirm the cultural beliefs. New approaches investigate the ways that interactional processes may perpetuate or undermine gender inequality in a society as that society undergoes economic change. If the cultural beliefs about gender that shape interaction change more slowly than economic arrangements, people interacting in gendered ways may rewrite gender inequality into newly emerging forms of socioeconomic organization in society. On the other hand, rapidly changing socioeconomic conditions may change the constraints on interaction between men and women in many contexts so that people's experiences undermine consensual beliefs about gender and alter them over time.

See also: Androgyny; Cultural Variations in Interpersonal Relationships; Feminist Theory; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Language: Cultural Concerns; Gender Differences in Personality and Social Behavior; Gender Ideology: Cross-cultural Aspects; Gender-related Development; Groups, Sociology of; Interactionism: Symbolic; Interpersonal Attraction, Psychology of; Language and Gender; Male Dominance; Masculinities and Femininities; Social Networks and Gender; Social Psychology: Sociological; Social Psychology, Theories of; Social Relationships in Adulthood; Stereotypes, Social Psychology of

Bibliography

Aries E 1996 *Men and Women in Interaction: Reconsidering the Differences*. Oxford University Press, New York

Deaux K, LaFrance M 1998 Gender. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, Vol. 1, pp. 788–827

Eagly A H 1987 *Sex Differences in Social Behavior: A Social-Role Interpretation*. Earlbaum, Hillsdale, NJ

Filardo E K 1996 Gender patterns in African-American and white adolescents' social interactions in same-race, mixed-sex groups. *Journal of Personal Social Psychology* **71**: 71–82

Maltz D N, Borker R A 1982 A cultural approach to male-female miscommunication. In: Gumperz J J (ed.) *Language and Social Identity*. Cambridge University Press, Cambridge, UK, pp. 196–256

Parsons T, Bales R F 1955 *Family, Socialization, and Interaction Process*. Free Press, Glencoe, IL

Ridgeway C L 1993 Gender, status, and the social psychology of expectations. In: England P (ed.) *Theory on Gender/Feminism on Theory*. A de Gruyter, New York

Ridgeway C L, Smith-Lovin L 1999 The gender system and interaction. *Annual Review of Sociology* **25**: 191–216
Tannen D 1990 *You Just Don't Understand: Women and Men in Conversation*. Morrow, New York
Wagner D G, Berger J 1997 Gender and interpersonal task behaviors: Status expectation accounts. *Sociological Perspectives* **40**: 1–32
West C, Zimmerman D 1987 Doing gender. *Gender and Society* **1**: 125–51

C. L. Ridgeway

Smith, Adam (1723–90)

Adam Smith was born in Kirkcaldy, in the County of Fife, and baptized on June 5, 1723 (the date of birth is unknown). He was the son of Adam Smith, Clerk to the Court-Martial, later Comptroller of Customs in the town, and of Margaret Douglas of Strathendry.

1. *Early Years*

Smith entered Glasgow University in 1737—at the not uncommon age of 14. He was fortunate in his mother's choice of University and in the period of his attendance. The old, nonspecialised, system of 'regenting' had been replaced in the late 1720s by a new arrangement whereby individuals professed a single subject. There is little doubt that Smith benefited from the teaching of Alexander Dunlop (Greek) and of Robert Dick (Physics).

But two men in particular are worthy of note in view of Smith's later interests. The first is Robert Simson (1687–1768) whom Smith later described as one of the 'greatest mathematicians that I have ever had the honour to be known to, and I believe one of the two greatest that have lived in my time' (TMS, III.2.20). Smith could well have acquired from Simson and Matthew Stewart (father of Dugald) his early and continuing interest in mathematics. Dugald Stewart recalled in his memoir (Stewart 1977) that Smith's favorite pursuits while at university were mathematics and natural philosophy.

Campbell and Skinner noted that 'Simson's interests were shared generally in Scotland. From their stress on Greek geometry the Scots built up a reputation for their philosophical elucidation of Newtonian fluxions, notably in the *Treatise on Fluxions* (1724) by Colin Maclaurin (1698–1746), another pupil of Simson's who held chairs of mathematics in Aberdeen and Edinburgh (1982, p. 20).

But important as it undoubtedly was Simson's influence upon Smith pales by comparison with that exerted by Francis Hutcheson (1694–1746). A student of Gerschom Carmichael, the first Professor of Moral Philosophy, Hutcheson succeeded him in 1729. A great stylist, Hutcheson lectured in English (rather than Latin) 3 days a week on classical sources and 5 days on Natural Religion, Morals, Jurisprudence, and Government. As Dugald Stewart was to observe, Hutcheson's lectures contributed to diffuse, in Scotland, the taste for analytical discussion, and a spirit of liberal enquiry.

It is believed that Smith graduated from Glasgow in 1740 and known that he was elected to the Snell Exhibition (scholarship) in the same year. He matriculated in Balliol College on July 7, and did not return to Scotland until 1746—the year that ended the Jacobite Rebellion and saw the death of Hutcheson. The 6 years spent in Oxford were often unhappy. Smith had a retiring personality, his health was poor, and the College was pro-Jacobite and 'anti-Scottish.' Some members were also 'unenlightened'; a fact confirmed by the confiscation by one tutor of David Hume's *Treatise on Human Nature*.

And yet Smith was to write to the Principal of Glasgow University (Archibald Davidson), on the occasion of his election as Lord Rector (letter 274, dated November 16, 1787):

> No man can owe greater obligations than I do to the University of Glasgow. They educated me, they sent me to Oxford, soon after my return to Scotland they elected me one of their members, and afterwards preferred me to another office, to which the abilities and virtues of the never to be forgotten Dr Hutcheson had given a superior degree of illustration.

The reference to Oxford was no gilded memory.

Balliol had one of the best libraries in Oxford. Smith's occasional ill-health may be explained by his enthusiastic pursuit of its riches. It has been assumed that Smith developed an interest in Rhetoric and *Belles Lettres* during this period, although it now seems likely, in view of his later career, that he also further developed longer standing interests in literature, science, ethics, and jurisprudence.

2. *Professor in Glasgow*

Smith returned to Kirkcaldy in 1746 without any fixed plan. But his wide-ranging interests must have been known to his friends, three of whom arranged a program of public lectures which were delivered in Edinburgh between 1748 and 1751. The three friends were Robert Craigie of Glendoick, James Oswald of Dunnikier, and Henry Home, Lord Kames.

The lectures were of an extramural nature and delivered to a 'respectable auditory.' They probably also included material on the history of science (astronomy), jurisprudence, and economics.

The success of Smith's courses in Edinburgh no doubt led to his appointment to the Glasgow Chair of Logic and Rhetoric in 1751—where once again he

enjoyed the support of Henry Home, Lord Kames. Evidence gathered from former pupils confirms that Smith did lecture on rhetoric but also that he continued to deploy a wide range of interests, leading to the conclusion that he continued to lecture on the history of philosophy and science in the Glasgow years (Mizuta 2000, p. 101).

Smith attached a great deal of importance to his essay on the history of Astronomy (Corr, letter 137) the major part of which may well have been completed soon after leaving Oxford (Ross 1995, p. 101).

Adam Smith was translated to Hutcheson's old chair of Moral Philosophy in 1752. As John Millar recalled to Dugald Stewart, the course was divided into four parts: natural theology, ethics, jurisprudence, and expediency (economics). Millar also confirmed that the substance of the course on ethics reappeared in *The Theory of Moral Sentiments* (1759) and that the last part of the course featured in the *Wealth of Nations* (1776). We know from Smith's own words that he hoped to complete his wider plan by writing a 'sort of Philosophical History of all the different branches of literature, of philosophy poetry and eloquence' together with 'a sort of theory and history of law and government' (Corr, letter 248).

Smith returned to this theme in the advertisement to the sixth and final edition of TMS in noting that TMS and WN were parts of a wider study. 'What remains the theory of jurisprudence, which I have long projected, I have hitherto been hindered from executing, by the same occupations which had till now prevented me from revising the present work.' But the outlines of the projected study are probably revealed by the content of LJ(A) and LJ(B), and by those passages in WN which can now be recognized as being derived from them (e.g., WN, III, and V.i.a.b).

The links between the parts of the great plan are many and various. The TMS, for example, may be regarded as an exercise in social philosophy, which was designed in part to show the way in which so self-regarding a creature as man, erects (by natural as distinct from artificial means) barriers against his own passions, thus explaining the observed fact that he is always found in 'troops and companies.' The argument places a good deal of emphasis on the importance of general rules of behaviour which are related to experience and which may thus vary in content, together with the need for some system of government as a precondition of social order.

The historical analysis, with its four socioeconomic stages, complements this argument by formally considering the origin of government and by explaining to some extent the forces which cause variations in accepted standards of behavior over time. Both are related in turn to Smith's treatment of political economy.

The most polished accounts of the emergent economy and of the psychology of the 'economic man' are to be found, respectively, in the third book of WN and in Part VI of TMS that was added in 1790. Yet both areas of analysis are old and their substance would have been communicated to Smith's students and understood by them to be what they possibly were: a preface to the treatment of political economy.

From an analytical point of view, Smith's treatment of economics in the final part of LJ was not lacking in sophistication, which is hardly surprising in view of his debts to Francis Hutcheson, and through him, to Gerschom Carmichael, and thus to Pufendorf.

3. *The French Connection*

Despite his anxieties, the TMS proved to be successful (Corr, letter 31) and attracted the attention of Charles Townshend, the statesman, who set out to persuade Smith to become tutor to the young Duke of Buccleuch. Smith eventually agreed, and left Glasgow with two young charges to begin a stay of almost 2 years in France. Smith resigned his Chair in February 1764 (Corr, letter 87). The party spent many months in Toulouse before visiting Avignon and Geneva (where Smith met the much-admired Voltaire) prior to reaching Paris to begin a stay of some 10 months.

From an intellectual point of view, the visit was a resounding success and arguably influential in the sense that Smith was able to meet, amongst many others, Francois Quesnay and A. R. J. Turgot. Quesnay's economic model dates from the late 1750s but it is noteworthy that he was working on a new version of the *Tableau*, the *Analyse*, during the course of Smith's visit. It is also known that Smith met Turgot at this time and that the latter was at work on the *Reflections*.

While Smith wrote a detailed commentary on Physiocratic teaching (WN IV.ix), his lectures on economics, as delivered in the closing months of 1763, did not include a model of the kind which he later associated with the French Economists—thus suggesting that he must have found a great deal to think about in the course of 1766. But there were more immediate concerns. The young Duke was seriously ill in the summer, leading Smith to call upon the professional services of Quesnay. Smith's distress was further compounded by the illness and death of the younger brother, Hew Scott. The visit to Paris was ended abruptly. Smith spent the winter of 1766 in London and returned to Kirkcaldy in the following Spring to begin a stay of more than 6 years during which he worked upon successive drafts of the WN.

4. *London 1773–76*

Smith left Kirkcaldy in the spring of 1773 to begin a stay of some 3 years in London. David Hume assumed that publication of WN was imminent, but in fact, the

book did not appear until March 1776. The reason may well be that Smith was engaged in systematic revision of a complex work. But he was also concerned with two sophisticated areas of analysis.

The first of these questions related to Smith's interest in the optimal organization of public services, notably education, and features in a long letter addressed to William Cullen dated September 1774. The issue of public services was one of the topics that Smith addressed during his stay in London. Cullen had written to Smith seeking his opinion on proposals from the Royal College of Physicians in Edinburgh. The petition suggested that doctors should be graduates, that they should have attended College for at least 2 years, and that their competence be confirmed by examination. Smith opposed this position, arguing that universities should not have a monopoly of provision and in particular that the performance of academic staff should be an important factor in determining salaries. In sum, an efficient system of higher education requires: a state of free competition between universities and private teachers; a capacity effectively to compete in the market for men of letters; freedom of choice for students as between teachers, courses, and colleges, together with the capacity to be sensitive to market forces, even if those forces were not always themselves sufficient to ensure the provision of the basic infrastructure. Smith also argued that education should be paid for through a combination of private and public funding (WN, V.i.i.5). Smith was arguing in favour of the doctrine of 'induced efficiency' and applied the principles involved to the whole range of public services (Skinner 1996, chap. 8).

At the same time, Smith also addressed another complex question, which was considered at great length in WN (Book IV, vii), namely the developing tensions with the American Colonies. Indeed, Hume believed that Smith's growing preoccupation with the Colonial Question was the cause of the delay in the publication of the book (Corr, letter 139).

Smith's long analysis of the issues involved provides the centerpiece of his critique of the 'mercantile' system of regulation (WN, IV, v.a.24). On Smith's account the Regulating Acts of Trade and Navigation in effect confined the Colonies to the production of primary products, while the 'mother country' concentrated on more refined manufacturers—thus creating a system of complementary markets which benefited both parties. But it was Smith's contention that the British economy could not sustain indefinitely the cost of colonial administration and that in the long run the rate of growth in America would come into conflict with restrictions currently imposed.

Smith, therefore, argued that Great Britain should dismantle the Acts voluntarily with a view to creating a single free trade area, in effect an Atlantic Economic Community, with a harmonized system of taxation, possessing all the advantages of a common language and culture.

Smith accepted the principle that there should be no taxation without representation. In this connection, he noted that:

> In the course of little more than a century perhaps the produce of American might exceed that of British taxation. The seat of empire would then naturally remove itself to that part of the empire which contributed most to the general defence of the whole (WN IV, vii.c.79).

Hugh Blair, Professor of Rhetoric in Edinburgh, objected to Smith's inclusion of this material on the grounds that 'it is too much like a publication for the present moment' (Corr, 151). Others were more perceptive, recognizing that Smith was applying principles drawn from his wider system to the specific case of America (e.g., Thomas Pownall, Corr, ap.A). Pownall, a former Governor of Massachusetts, was an acute critic of Smith's position (Skinner 1996, chap. 8) but recognized that Smith had sought to produce 'an institute of the Principia of *those laws of motion*, by which the operations of the community are directed and regulated, and by which they should be examined' (Corr p. 354).

5. *The Wealth of Nations*

Dugald Stewart, Professor of Moral Philosophy in Edinburgh, noted that: 'It may be doubted, with respect to Mr Smith's Inquiry, if there exists any book beyond the circle of the mathematical and physical sciences, which is at once so agreeable in its arrangement to the rules of a sound logic and so accessible to the examination of ordinary readers' (Stewart 1977, IV22).

This is a complement which Smith would have appreciated, conscious as he was of the 'beauty of a systematical arrangement of different observations connected by a few common principles' (WN, V.i.f. 25).

But there are really two systems in WN. The first is a model of 'conceptualised reality' (Jensen 1984) which provides a description of a modern economy, which owed much to the Physiocrats. The second system is analytical and builds upon the first.

5.1 *A Model of Conceptualized Reality*

If the *Theory of Moral Sentiment* provides an account of the way in which people erect barriers against their own passions, thus meeting a basic precondition for economic activity, it also provided an account of the psychological judgments on which that activity depends. The historical argument on the other hand explains the origins and nature of the modern state and provides the reader with the means of understanding the essential nature of the exchange economy. For Smith: 'the great commerce of every civilised society, is that carried on between the inhabitants of

the town and those of the country ... The gains of both are mutual and reciprocal, and the division of labour is in this, as in all other cases, advantageous to all the different persons employed in the various occupations into which it is subdivided' (WN, III.i.1).

The concept of an economy involving a flow of goods and services, and the appreciation of the importance of intersectoral dependencies, were familiar in the eighteenth century. Such themes are dominant features of the work done, for example, by Sir James Steuart and David Hume. But what is distinctive about Smith's work, at least as compared to his Scottish contemporaries, is the emphasis given to the importance of three distinct factors of production (land, labor, capital) and to the three categories of return (rent, wage, profit) which correspond to them. What is distinctive to the modern eye is the way in which Smith deployed these concepts in providing an account of the flow of goods and services between the sectors involved and between the different socio-economic groups (proprietors of land, capitalists, and wage-labor). The approach is also of interest in that Smith, following the lead of the French Economists, worked in terms of period analysis— typically the year was chosen, so that the working of the economy is examined within a significant time dimension as well as over a series of time periods. Both versions of the argument emphasise the importance of capital, both fixed and circulating.

5.2 A Conceptual Analytical System

The 'conceptual' model which Smith had in mind when writing the *Wealth of Nations* is instructive and also helps to illustrate the series of separate, but interrelated problems, which economists must address if they are to attain the end which Smith proposed, namely an understanding of the full range of problems which have to be encountered. Smith, in fact, addressed a series of areas of analysis which began with the problem of value, before proceeding to the discussion of the determinants of price, the allocation of resources between competing uses, and, finally, an analysis of the forces which determine the distribution of income in any one time period and over time.

The analysis offered in the first book enabled Smith to proceed directly to the treatment of macroeconomic issues and especially to a theory of growth which provides one of the dominant features of the work as a whole (c.f. Skinner 1996, chap. 7). The idea of a single, all-embracing conceptual system, whose parts should be mutually consistent is not an ideal which is so easily attainable in an age where the division of labor has increased significantly the quantity of science through specialization. But Smith becomes even more informative when we map the content of the 'conceptual (analytical) system' against a model of the economy, which is essentially descriptive.

Perhaps the most significant feature of Smith's vision of the 'economic process,' to use Blaug's phrase, lies in the fact that it has a significant time dimension. For example, in dealing with the problems of value in exchange, Smith, following Hutcheson, made due allowance for the fact that the process involves judgments with regard to the utility of the commodities to be received, and the disutility involved in creating the commodities to be exchanged. In the manner of his predecessors, Smith was aware of the distinction between utility (and disutility) anticipated and realized, and, therefore, of the process of adjustment which would take place though time. Young (1997, p. 61) has emphasized that the process of exchange may itself be a source of pleasure (utility).

In an argument which bears upon the analysis of the TMS, Smith also noted that choices made by the 'rational' individual may be constrained by the reaction of the spectator of his conduct—a much more complex situation than that which more modern approaches may suggest. Smith makes much of the point in his discussion of Mandeville's 'licentious' doctrine that private vices are public benefits, in suggesting that the gratification of desire is perfectly consistent with observance of the rules of propriety as defined by the 'spectator,' that is, by an external agency. In an interesting variant on this theme, Etzioni (1988, pp. 21–4) noted the need to recognize 'at least two irreducible sources of valuation or utility; pleasure and morality.' He added that modern utility theory 'does not recognise the distinct standing of morality as a major, distinct, source of valuations' and hence as an explanation of 'behaviour' before going on to suggest that his own 'deontological multi-utility model' is closer to Smith than other modern approaches.

Smith's theory of price, which allows for a wide range of changes in taste, is also distinctive in that it allows for competition among and between buyers and sellers, while presenting the allocative mechanism as one which involves simultaneous and interrelated adjustments in both factor and commodity markets.

As befits a writer who was concerned to address the problems of change and adjustment, Smith's position was also distinctive in that he was not directly concerned with the problem of equilibrium. For him the 'natural' (supply) price was: 'as it were, the central price, to which the prices of all commodities are continually gravitating ... whatever may be the obstacles which hinder them from settling in this center of repose and continuance, they are constantly tending towards it' (WN, I.vii.15).

The picture was further refined in the sense that Smith introduced into this discussion the doctrine of 'net advantages' (WN, I.x.a.1). This technical area is familiar to labor economists, but in Smith's case it becomes even more interesting in the sense that it provides a further link with the TMS, and with the discussion of constrained choice. It was Smith's contention that men would only be prepared to

embark on professions that attracted the disapprobation of the spectator if they could be suitably compensated (Skinner 1996, p. 155) in terms of monetary reward.

But perhaps the most intriguing feature of the macroeconomic model is to be found in the way in which it was specified. As noted earlier, Smith argued that incomes are generated as a result of productive activity, thus making it possible for commodities to be withdrawn from the 'circulating' capital of society. As he pointed out, the consumption of goods withdrawn from the existing stock may be used up in the present period, or added to the stock reserved for immediate consumption, or used to replace more durable goods which had reached the end of their lives in the current period. In a similar manner, undertakers and merchants may add to their stock of materials, or to their holdings of fixed capital while replacing the plant which had reached the end of its operational life. It is equally obvious that undertakers and merchants may add to, or reduce, their inventories in ways that will reflect the changing patterns of demand for consumption and investment goods, and their past and current levels of production.

Smith's emphasis upon the point that different 'goods' have different life-cycles means that the pattern of purchase and replacement may vary continuously as the economy moves through different time periods, and in ways which reflect the various age profiles of particular products as well as the pattern of demand for them. If Smith's model of the 'circular flow' is to be seen as a spiral, rather than a circle, it soon becomes evident that this spiral is likely to expand (and contract) through time at variable rates.

It is perhaps this total vision of the complex working of the economy that led Mark Blaug to comment on Smith's sophisticated grasp of the economic process and to distinguish this from his contribution to particular areas of economic analysis (c.f. Jensen 1984, Jeck 1994, Ranadive 1984).

What Smith had produced was a model of conceptualized reality, which is essentially descriptive, and which was further illuminated by an analytical system which was so organized as to meet the requirement of the Newtonian model (Skinner 1996, chap.7). Smith's model(s) and the way in which they were specified confirmed his earlier claim that government's ought not to interfere with the economy—a theme stated in the 'manifesto of 1775' (Stewart 1977, IV.25), repeated in LJ, confirmed by Turgot, and even more eloquently defended in WN (IV.ix.51).

Smith would no doubt be gratified that Hume had lived to see the publication of WN and pleased with the assessment that it has 'Depth and Solidity and Acuteness' (Corr, letter 150). Hume died in the summer of 1776. Two years later Smith was asked by a former pupil, Alexander Wedderburn, Solicitor-General in Lord North's administration, to advise on the options open to the British Government in the aftermath of Bourgoyne's surrender at Saratoga. Smith returned to his old theme of Union, but recognized that the most likely outcome was military defeat (Corr, app B). In February of the same year, Smith was appointed Commissioner of Customs and of the Salt Duties that gave him an income of £600pa to be added to the £300 which he still received from the Duke.

Smith then moved to Edinburgh where he lived with his mother and a cousin. He died in 1790 after instructing his executors, Joseph Black and James Hutton, to burn the bulk of his papers. We may believe that the pieces which survived (which include the *Astronomy*) and which were later published in the *Essays on Philosophical Subjects*, were all specified by Smith.

6. *Influence*

Adam Smith's influence upon his successors is a subject worthy of a book rather than a few concluding paragraphs. But there are some obvious points to be made even if the list can hardly be exhaustive.

If we look at the issues involved from the standpoint of economic theory and policy, the task becomes a little simpler.

On the subject of policy, Teichgraeber has noted that Smith's advocacy of free trade or economic liberalism did not 'register any significant victories during his life-time' (1987, p. 338). Indeed, Tribe has argued that 'until the final decade of the eighteenth century, Sir James Steuart's Inquiry was better known than Smith's *The Wealth of Nations*' (1988, p. 133). The reason is that Steuart's extensive and unique knowledge of conditions in Europe, gained as a result of exile, made him acutely aware of problems of immediate relevance; problems such as unemployment, regional imbalance, underdeveloped economies and the difficulties which were presented in international trade as a result of variations in rates of growth (Skinner 1996, chap. 11). In view of later events, it is ironic to note that Alexander Hamilton considered Steuart's policy of protection for infant industries to be more relevant to the interests of the young American Republic than the 'fuzzy philosophy of Smith' (Stevens 1975, pp. 215–7).

But the situation was soon to change. In the course of a review of the way in which WN had been received, Black noted that:

> On the side of policy, the general impression left by the historical evidence is that by 1826 not only the economists but a great many other influential public men were prepared to give assent and support to the system of natural liberty and the consequent doctrine of free trade set out by Adam Smith' (1976, p. 47).

Black recorded that the system of natural liberty attracted attention on the occasion of every anni-

versary. But a cautionary note was struck by Viner (1928). Having reviewed Smith's treatment of the function of the state in *Adam Smith and Laisser-Faire*, a seminal article, Viner concluded:

> Adam Smith was not a doctrinaire advocate of laisser-faire. He saw a wide and elastic range of activity for government, and was prepared to extend it even further if government, by improving its standards of competence, honesty and public spirit, showed itself entitled to wider responsibilities' (Wood 1984, i.164).

But this sophisticated view, which is now quite general, does not qualify Robbins' point that Smith developed an important argument to the effect that economic freedom 'rested on a two fold basis: belief in the desirability of freedom of choice for the consumer and belief in the effectiveness, in meeting this choice, of freedom on the part of the producers' (Robbins 1953, p. 12). Smith added a dynamic dimension to this theme in his discussion of the Corn Laws (WN, IV.v.b). The thesis has proved to be enduringly attractive.

Analytically, the situation is also intriguing. Teichgraeber's research revealed that there 'is no evidence to show that many people exploited his arguments with great care before the first two decades of the nineteenth century (1987, p. 339). He concluded: 'It would seem at the time of his death that Smith was widely known and admired as the author of the Wealth of Nations. Yet it should be noted too that only a handful of his contemporaries had come to see his book as uniquely influential' (1987, p. 363).

Black has suggested that for Smith's early nineteenth-century successors, the WN was 'not so much a classical monument to be inspected, but as a structure to be examined and improved where necessary' (1984, p. 44). There were ambiguities in Smith's treatment of value, interest, rent, and population theory. These ambiguities were reduced by the work of Ricardo, Malthus, James Mill, and J. B. Say, making it possible to think of a classical system dominated by short-run self-equilibrating mechanisms and a long-run theory of growth.

But there was one result of which Smith would not have approved in that the classical orthodoxy made it possible to think of economics as quite separate from ethics and history. In a telling passage reflecting upon the order in which Smith developed his argument (ethics, history, economics), Hutchison concluded that Smith was unwittingly led, as if by an Invisible Hand, to promote an end which was no part of his intention, that 'of establishing political economy as a separate autonomous discipline' (1988, p. 355).

But the economic content of WN did, after all, provide the basis of classical economics in the form of a coherent, all-embracing account of 'general interconnexions' (Robbins 1953, p. 172). As Viner had earlier pointed out, the source of Smith's originality lies in his 'detailed and elaborate application to the wilderness of economic phenomena of the unifying concept of a co-ordinated and mutually interdependent system of cause and effect relationships which philosophers and theologians had already applied to the world in general' (Wood 1984, i.143).

Down the years it is the idea of system which has attracted sustained attention perhaps because it is now virtually impossible to duplicate a style of thinking which becomes more informative the further we are removed from it. No one who is familiar with the Smithian edifice can fail to notice that he thought mathematically and in a manner which reflects his early interest in related disciplines, including the life sciences—all mechanistic, evolutionary, static, and dynamic, which so profoundly affected the shape assumed by his *System of Social Science*.

Works of Adam Smith	(OUP, 1976–1983)
Corr	Correspondence, eds E C Mossner and I S Ross, (1977)
Early Draft of WN	in W R Scott, Adam Smith as Student and Professor, (Jacksons, Glasgow, 1937)
EPS	Essays on Philosophical Subjects, incl. the Astronomy, General eds D D Raphael and A S Skinner (1980)
Letter	Letter to Authors of the Edinburgh Review (1756), in EPS
LJ(A) and LJ(B)	Lectures on Jurisprudence ed R L Meek, P G Stein and D D Raphael (1978)
LRBL	Lectures on Rhetoric and Belles Letters, ed J C Bryce (1983)
TMS	The Theory of Moral Sentiments, ed D D Raphael and A L Macfie (1976)
WN	The Wealth of Nations, eds R H Campbell, A S Skinner and W B Todd (1976)
Stewart	Account of the Life and Writings of Adam Smith (1977) in Corr.

Bibliography

Black R D C 1976 Smith's contribution in historical perspective. In: Wilson T, Skinner A S (eds.) *The Market and the State*. Oxford University Press, Oxford, UK

Campbell R H, Skinner A S 1982 *Adam Smith*. Croom-Helm, London

Etzioni A 1988 *The Moral Dimension: Towards a New Economics*. Macmillan, London

Groenwegen P 1969b Turgot and Adam Smith. *Scottish Journal of Political Economy* **16**

Hutchison T 1988 *Before Adam Smith*. Blackwell, Oxford, UK
Jeck A 1994 The macro-structure of Adam Smith's theoretical system. *European Journal of the History of Economic Thought*. **3**: 551–76
Jensen H E 1984 Sources and Contours in Adam Smith's Conceptualised Reality. Wood, ii.194
Meek R L 1962 *The Economics of Physiocracy*. Allen and Unwin, London
Meek R L 1973 *Turgot on Progress, Sociology and Economics*. Cambridge, UK
Mizuta H 2000 *Adam Smith: Critical Responses*. Routledge, London
Ranadive K R 1984 The wealth of nations: the vision and conceptualisation. In: Wood J C (ed.) *Adam Smith: Critical Assessments*, ii. 244
Robbins L 1953 *The Theory of Economic Policy in English Classical Political Economy*. Macmillan, London
Ross I S 1995 *Life of Adam Smith*. Oxford University Press, Oxford, UK
Skinner A S 1996 *A System of Social Science: Papers Relating to Adam Smith*, 2nd edn. Oxford University Press, Oxford, UK
Stevens D 1975 Adam Smith and the colonial disturbances. In: Skinner A S, Wilson T (eds.) *Essays on Adam Smith*. Oxford University Press, Oxford, UK
Teichgraeber R 1987 Less abused than I had reason to expect: The reception of the *Wealth of Nations* in Britain 1776–1790. *Historical Journal* **80**: 337–66
Tribe K P 1988 *Governing Economy: The Reformation of German Economic Discourse, 1750–1840*. Cambridge University Press, Cambridge, UK
Viner J 1928 Adam Smith and Laissez-Faire. *Journal of Political Economy* **35**: 143–67
Wood J C 1984 *Adam Smith: Critical Assessments*. Croom-Helm, Beckenham, UK
Young J 1977 *Economics as a Moral Science: The Political Economy of Adam Smith*. Edward Elgar, Cheltenham, UK

A. S. Skinner

Smoking and Health

1. Introduction

Worldwide, 1.1 billion people use tobacco products, primarily cigarettes. An estimated four million die annually as a consequence. By the year 2030, an estimated 1.6 billion will consume tobacco and tobacco's death toll will have skyrocketed to 10 million per year, making tobacco the leading behavioral cause of preventable premature death throughout the world. Tobacco already claims that dubious distinction in the world's developed countries (World Health Organization 1997).

The causes of this pandemic of avoidable disease and death are a complex web of physiological, psychological, social, marketing, and policy factors. Social and behavioral scientists of all disciplinary stripes have contributed to identifying and disentangling these factors, although no single model has yet adequately characterized the interactions of these determinants. Social and behavioral scientists have also elucidated important mechanisms by which both private sector and governmental intervention can discourage the initiation of smoking and encourage quitting.

After a brief description of the burden of smoking and how epidemiological science unearthed the principal tobacco-related disease connections, the causes of tobacco consumption are summarized. The efforts of social and behavioral scientists to improve the ability of smokers to quit smoking are then described, including better theoretical characterization of the determinants of quitting success and the development of more effective cessation interventions. Next examined is how social and behavioral scientists have analyzed the effects of selected important tobacco control measures, and thereby contributed to the formulation and implementation of tobacco control programs and policies. Finally, the future contribution of the social and behavioral sciences to dealing with a potentially even more cataclysmic crisis in world health during the twenty-first century is considered.

Traditional methods of youth smoking prevention, such as school health education and enforcement of minimum age-of-purchase laws, are only touched on due in part to limited scientific understanding of how initiation can be effectively discouraged, and due to its coverage elsewhere (Lantz et al. 2000; see *Substance Abuse in Adolescents, Prevention of*; *Smoking Prevention and Smoking Cessation*; *Health Promotion in Schools*).

2. Cigarette Smoking and Disease: Making the Connection

In the major industrialized nations, smoking causes from a sixth to a fifth of all mortality. The two major smoking-related causes of death are lung cancer and coronary heart disease (CHD). In the United States, epidemiologists estimate that smoking accounts for approximately 90 percent of all lung cancer deaths, with lung cancer the leading cancer cause of death in both men and women. Smoking is credited with better than a fifth of CHD deaths. In addition, smoking causes four-fifths of all chronic obstructive pulmonary disease mortality and just under a fifth of all stroke deaths (US Department of Health and Human Services 1989). The exposure of nonsmokers to environmental tobacco smoke is also a cause of disease and death (American Council on Science and Health 1999).

The toll of smoking is proportionately smaller in developing countries, reflecting the more recent rise and lesser intensity of smoking to this point. However, projections indicate a future chronic disease epidemic quite comparable to that now experienced in the world's more affluent nations (World Health Organization 1997).

Although the health hazards of tobacco smoking have been suspected for centuries, serious interest is a twentieth-century phenomenon, coincident with the advent of cigarette smoking. Prior to the twentieth century, tobacco was smoked primarily in pipes and cigars, chewed, or used nasally as snuff. Harsh tobaccos made deep inhalation of tobacco smoke difficult. Tobacco likely exacted a modest toll through the nineteenth century.

In the US, four factors combined in the early twentieth century to make cigarette smoking the most popular, and lethal, form of tobacco consumption. Perfection of the Bonsack cigarette rolling machine in 1884 introduced relatively inexpensive and neater cigarettes to the market. Discovery of the American blend of tobaccos (combining more flavorful tobaccos from Turkey and Egypt with milder American tobaccos) made deep inhalation of cigarette smoke feasible. The development and marketing of Camel cigarettes in 1913, the first American blend cigarette, introduced this new-generation product to the American public with an advertising campaign often credited as inaugurating modern advertising. Finally, cigarettes were included in soldiers' rations during World War I, permitting soldiers a quick and convenient battlefield tobacco break. Considered effeminate and unattractive prior to the war, cigarette smoking came home a 'manly' behavior. Since then, cigarette smoking has dominated all other forms of tobacco use by far in the US and in most countries of the world.

Lung cancer was virtually unknown at the beginning of the twentieth century. By 1930, it had become the sixth leading cause of cancer death in men in the United States. Less than a quarter of a century later, lung cancer surpassed colorectal cancer to become the leading cause of male cancer death. It achieved the same status among American women in the mid-1980s, surpassing breast cancer (US Department of Health and Human Services 1989).

The rapidly growing rate of lung cancer in the early decades of the century spurred a number of epidemiologists and laboratory scientists to begin investigating the relationship between smoking and cancer. In 1950, Wynder and Graham published a now classic retrospective epidemiologic analysis that strongly linked the behavior to the disease. Throughout the decade, a series of articles documented similar findings from major prospective (cohort) mortality studies in the US and the UK. Scientific groups in both countries soon thereafter published seminal public health documents indicating smoking as the major cause of lung cancer and a cause of multiple other diseases (Royal College of Physicians of London 1962, US Public Health Service 1964, US Department of Health and Human Services 1989).

By the end of the twentieth century, approximately 70,000 studies in English alone associated cigarette smoking with a wide variety of malignant neoplasms and cardiovascular and pulmonary diseases, as well as numerous other disorders. The modern plague of tobacco-produced disease is now rapidly metastasizing—both figuratively and literally—to the world's poorer nations, where increasing affluence and western image-making have combined to place cigarettes in the mouths of a sizable majority of men and a growing minority of women.

3. *The Causes of Smoking*

Smoking affords users a very mild 'high' and addresses a number of often seemingly competing physical and psychological needs, for relaxation or stimulation, distraction or concentration, for example. The crucial ingredient in the sustained use of cigarettes is the addictiveness of the nicotine in tobacco smoke. Once addicted, as most regular smokers are, smoking serves the physiological purpose of avoiding nicotine withdrawal. Nicotine affects brain receptors in much the same manner as other addictive substances, and withdrawal shares many of the same unpleasant characteristics (US Department of Health and Human Services 1988).

In many countries, smoking is initiated during the teen and even pre-teen years. At the time of initiation, new smokers typically do not appreciate the nature of addiction, much less the addictiveness of nicotine. Further, they tend to be short sighted, unconcerned about potential distant adverse health consequences (US Department of Health and Human Services 1994). The combination translates into a large pool of new smokers who have adopted the behavior, and become addicted, without considering either the danger or addictiveness of smoking, a situation that most will come to regret.

The physical effects of nicotine notwithstanding, history teaches that virtually all drug use is socially conditioned; tobacco smoking is no exception. The use of tobacco in the Americas prior to the arrival of Europeans demonstrates this vividly. Although tobacco played a prominent role in most native societies, in some tobacco smoking was restricted exclusively to the shaman, who used it for medicinal and relig-ious purposes. In other societies, tobacco smoking anointed official nonsectarian functions of tribal leaders (such as the famous 'peace pipe'). In still others, nearly all males smoked tobacco frequently, for personal and social reasons (Goodman 1993).

In contemporary society, the initiation of cigarette smoking is often viewed as a rite of passage to adulthood. In many poorer societies, smoking is seen as a symbol of affluence. In virtually all cases, smoking results from role modeling, of parents, prominent celebrities, or youthful peers.

Many tobacco control advocates attribute much of smoking's holding power to sophisticated tobacco industry marketing campaigns. In the United States alone, the industry spends over $5 billion per year on

advertising and other forms of marketing. The industry insists that the purpose of its advertising is solely to vie for shares of the existing market of adult smokers, a claim greeted with derision by the public health community. With US smokers exhibiting strong brand loyalty and two companies controlling three-quarters of the US market, marketing directed at brand switching (or its defensive analog, maintaining brand loyalty) would appear to be a relatively fruitless investment. Tobacco control advocates thus believe that much of the industry's marketing effort is directed toward attracting new smokers, primarily children but also groups of first- and second-generation American adults not yet fully acculturated into American society (e.g., Hispanic females, who have a very low smoking prevalence).

Similarly, the introduction of aggressive Western cigarette marketing into Asian societies and Eastern European and African countries has been attacked by the public health community as spurring the growth of smoking among children and other traditionally non-smoking segments of society (e.g., Asian females). In many of these countries, cigarette advertising was nonexistent or restricted to modest use of print media prior to the introduction of advertising by the major multinational tobacco companies. Before dissolution of the Soviet Union, there was virtually no advertising of the state-produced cigarettes in the Eastern European countries. That ended with a flourish when multinational tobacco companies entered the newly liberated economies and, in some instances, took over cigarette production from the inefficient state-run enterprises. In Japan, the state tobacco monopoly had never bothered to advertise on television prior to the entry of the multinationals. In short order, competition from the advertising of Western cigarette brands led to cigarettes becoming the second most advertised product on Japanese television. In Africa, the advertising media portray smoking as the indulgence of the much admired affluent set, with billboards in desperately poor villages depicting high-society Africans smoking and smiling in convivial social settings (McGinn 1997).

Although many tobacco control advocates find advertising a convenient villain to attack, widespread smoking has often preceded formal marketing. Tobacco smoking spread like wildfire through much of Europe in the sixteenth and seventeenth centuries. Similarly, male smoking rates in Russia and the Eastern European countries were very high prior to the fall of Communism and the advent of modern cigarette marketing. The effects of advertising are further examined below in Sect. 5.

4. *The Art and Science of Smoking Cessation*

In countries in which the hazards of smoking have been well publicized, surveys find that most smokers would like to quit and most have made at least one serious quit attempt, yet relatively few of those who try to quit succeed on any given attempt. In the US, for example, approximately three-quarters of smokers report they want to stop and as many as a third try each year. Yet only 2.5–3 percent succeed in quitting each year. Combined with persistent cessation efforts by many smokers, this modest quit rate has created a population in the US in which there are as many former smokers as current smokers (US Department of Health and Human Services 1989, 1990). In the aggregate, thus, quitting has significantly reduced the toll of smoking, but the toll remains stubbornly high due to the difficulty of quitting.

Although most former smokers quit without the aid of formal programs or products, the widespread desire to quit, paired with its difficulty, has created a small but thriving market for smoking cessation. Formal cessation interventions range from the distribution of how-to-quit booklets, to mass media cessation campaigns, to individual and group counseling programs, to self-administered nicotine replacement therapy (NRT) products, to use of cessation pharmaceuticals combined with clinical counseling. The efficacy of interventions ranges from sustained quit rates of about 5–7 percent for the most general and least resource-intensive interventions (e.g., generic cessation booklets) to 30 percent or more for the most resource-intensive programs that combine sophisticated counseling with the use of pharmaceuticals (Agency for Health Care Policy and Research 1996).

Over the years, behavioral scientists have helped refine smoking cessation techniques by evaluating interventions and developing theory applied to smoking cessation. Pre-eminent in the domain of theory have been models that characterize how smokers progress, through a series of 'stages,' to contemplate, attempt, and eventually succeed or fail in quitting, with cessation maintenance also examined (Prochaska and DiClemente 1983). Relating cessation advice to smokers' stages of readiness to change constitutes one way of 'tailoring' cessation messages. Another involves tying cessation advice to the specific motivations of individual smokers to quit and their specific concerns. For example, consider a smoker motivated to quit primarily by the high cost of smoking but also concerned about gaining weight. Armed with this knowledge, a cessation counselor can develop specific information on the financial savings the smoker can expect once he or she quits, while suggesting specific strategies to avoid weight gain (or to deal with it if it occurs).

Tailored cessation messages have the obvious virtue of meeting the idiosyncratic needs of individual smokers. In the absence of computer technology, however, they would entail the substantial cost of collecting information on individuals' needs and concerns and then developing individualized (tailored) cessation advice for them. Computers, however, per-

mit simple and inexpensive collection and translation of information into tailored cessation materials, such as individualized advice booklets and calendars with specific daily advice and reminders (Strecher 1999). Health informatn kiosks placed in malls and other public locations can provide instant tailored suggestions to help people quit smoking. The concept of tailoring holds great potential to enhance quit rates both in individual counseling situations and in low-cost nontreatment settings, such as these kiosks.

Social and behavioral scientists have also played a central role in defining appropriate medical treatment of smokers. The US Agency for Health Care Policy and Research (1996) clinical guideline for smoking cessation was produced by an expert committee including social and behavioral scientists who had worked on smoking cessation as service providers or developers or evaluators of interventions. The guideline urges physicians to regularly counsel their smoking patients about the implications of the behavior and to encourage them to quit. It concludes that a highly effective cessation approach involves physician counseling to quit, supplemented with use of cessation pharmaceuticals and maintenance advice through follow-up contacts. A cost-effectiveness analysis of the guideline added economics to the social sciences contributing to understanding optimal smoking cessation therapy (Cromwell et al. 1997).

Despite substantial improvements in the efficacy of cessation treatments, in any given year only a small fraction of the smokers who say they want to quit succeed in doing so, and only a small fraction of these have employed professional or programmatic assistance. As such, in any short-run period of time, the contribution of smoking cessation interventions to reducing the health toll of smoking is modest at best. This source of frustration has led a subset of smoking cessation professionals to explore methods of achieving 'harm reduction' that do not depend on smokers completely overcoming their addictions to nicotine. Harm reduction techniques range from helping smokers to reduce their daily consumption of cigarettes to encouraging consideration of the long-term use of low-risk nicotine-delivery devices, such as nicotine 'gum' or the patch. Though fraught with problems, harm reduction may be an idea whose time has come (Warner 2001).

5. *Analysis of the Effects of Tobacco Control Policies*

Another way to grapple with the toll created by smoking is to develop policies that discourage the initiation or continuation of smoking, or that restrict it to areas in which it will not impose risk on nonsmokers. Social and behavioral scientists have devoted substantial effort to studying the impacts of policies, as well as the processes by which such policies come to be adopted. This section examines the former. (For the latter, see, e.g., Fritschler and Hoefler 1996, Kagan and Vogel 1993.)

The World Health Organization and the World Bank have described and evaluated a wide array of tobacco control policies in countries around the globe (Roemer 1993, Prabhat and Chaloupka 1999). Three that have commanded the greatest amount of research attention are taxation, restricting or banning advertising and promotion, and limiting smoking in public places.

5.1 *Taxation*

Research performed primarily by economists has established that taxing cigarettes is among the most effective measures to decrease smoking (Chaloupka and Warner 2000, Prabhat and Chaloupka 1999). Because the quantity of cigarettes consumed declines by an amount proportionately smaller than the associated tax rise, taxation increases government revenues at the same time that it decreases smoking. In developed countries, economists find that a 10 percent increase in cigarette price induces approximately a 4 percent decrease in quantity demanded. In developing countries, the demand impact may be twice as large. Effective dissemination of research findings has made taxation one of the central tenets of a comprehensive tobacco control program.

Although the 'bottom line' about taxation is well established, research by economists and others points to concerns that remain unanswered. For example, does taxation discourage the initiation of smoking? Although evidence preponderantly suggests that it does, recent studies challenge the conventional view (Chaloupka and Warner 2000, Chaloupka 1999). 'Side effects,' or subtle unanticipated impacts of taxation, warrant additional attention as well. Notably, cigarette tax increases in the US cause some smokers to switch to higher nicotine cigarettes to get their customary dose of nicotine from fewer cigarettes, particularly among younger smokers (Evans and Farrelly 1998).

In the emerging field of 'behavioral economics,' a small cadre of psychologists is studying experimentally how smoking responds to incentives. For example, investigators give subjects a daily 'budget' (e.g., play money) with which to buy cigarettes, food, and other commodities, at prices established by the investigators, and evaluate the effects of 'taxation' by raising the price of cigarettes. Findings generally have been quite consistent with those in the mainstream economics literature (Chaloupka et al. 1999).

Behavioral economics can investigate questions not addressable using existing real world data, but also has decided limitations in predicting how people will

respond to price changes in that real world. By learning from each other, the two fields can enrich the evidentiary base for future tobacco tax policy.

5.2 Advertising

A diverse group of sociologists, psychologists, economists, marketing specialists, and public health scholars has studied the effects of marketing on smoking. Research by several scholars concludes that modern Western-style advertizing creates an imagery that many people, including large proportions of children, find easily recognizable and attractive. There is a strong correlation between children's interest in cigarette marketing campaigns and their subsequent smoking behavior. Temporally, advertising campaigns directed at specific segments of a population have often been followed by significant growth in smoking in the targeted groups, including women in the US in the 1960s, young women in Asia in the 1990s, and children in many countries (Warner 2000 McGinn 1997).

Whether any of these associations constitutes a causal relationship is the essential question. The same unidentified factor that makes cigarette advertising attractive to certain children could account for their subsequent smoking, independent of the advertising *per se*. Similarly, cigarette marketers might foresee an (independently occurring) expansion of a market segment and dive into the advertising void to compete for shares of that new market.

In the absence of the ability to run randomized controlled trials, empirical analysis of the relationship between cigarette advertising and consumption has been unable to prove a causal connection or to estimate its likely extent (US Department of Health and Human Services 1989, Warner 2000). However, strong new evidence supporting causality comes from a study re-examining data on the relationship between countries' policies with regard to cigarette advertising (ranging from no restrictions to complete bans) and levels of smoking within those societies (Saffer and Chaloupka 1999). Blending marketing theory with empirical analysis, this study concluded that a complete ban on cigarette advertising would decrease smoking by about 6 percent, while partial bans (e.g., banning cigarette ads on the broadcast media) would be unlikely to have an impact on cigarette consumption.

Combined with the previously existing evidence, the new research leads to the most plausible interpretation of the relationship between advertising and cigarette consumption. It is a conclusion that likely will satisfy neither tobacco industry defenders of the 'brand-share only' argument, nor tobacco control advocates who condemn marketing efforts as a principal cause of smoking. Cigarette advertising and other forms of marketing likely do increase the amount of smoking in a statistically significant manner, possibly accounting for as much as 10 percent of cigarette consumption. (This figure is likely to vary among societies, depending on the maturity of the smoking market and on familiarity with large-scale marketing campaigns.) The converse, of course, is that advertising and marketing almost certainly do not account for the majority of cigarettes consumed. For these, one must turn to other influences, all less policy tractable than advertising, including role modeling, peer behavior, and the addictiveness of nicotine and smoking.

5.3 Restrictions on Smoking in Public Places

'Clean indoor air laws,' which restrict or prohibit smoking in public places and workplaces, grew out of concerns that the exposure of nonsmokers to environmental tobacco smoke (ETS) could create a risk of disease, leading in the US to rapid diffusion of state laws, beginning in 1973. A decade later, similar laws emerged at the local level of government, where most of the legislative action in the US has remained since then (Brownson et al. 1997).

The scientific knowledge base actually followed early diffusion of legislation, with a number of studies of the relationship between ETS exposure and the risk of lung cancer published in the 1980s (Brownson et al. 1997). By the 1990s, the research base had become sufficiently strong that the US Environmental Protection Agency (EPA) declared ETS a 'Class A Carcinogen,' a proved environmental cause of cancer in nonsmoking humans. The EPA estimated that ETS caused approximately 3,000 lung cancer cases annually in the US, and also detailed nonfatal respiratory disease effects of ETS, particularly in children (US Environmental Protection Agency 1992). More recently, research has implicated ETS in heart disease deaths in adult nonsmokers, possibly an order of magnitude greater than the lung cancer impact (American Council on Science and Health 1999).

Social and behavioral scientists have informed the debate by studying the process of the emergence and diffusion of clean indoor air laws and, subsequently, the effects of such laws on nonsmokers' exposure to ETS and on smokers' smoking behavior. Studies have found generally high levels of compliance with the laws, even in bars (where low compliance was anticipated by many) (Warner 2000), and reductions in employees' exposure to ETS, as measured both by self-reports, air sampling, and examination of employees' body burdens of cotinine, a nicotine derivative.

Less obvious is whether clean indoor air laws discourage smoking or merely 'reallocate' it to times and places in which smoking is permitted. According to several studies, the laws do discourage smoking

among workers in regulated workplaces, producing increased quit rates and lower daily consumption among continuing smokers (Brownson et al. 1997).

5.4 Interaction of Policies

As policy research becomes more sophisticated, the relationships among policies and their impacts will come to be better understood. Illustrative is research examining the joint effects of tax increases and the adoption of clean indoor air laws. For example, some of the smoking decline credited to tax increases might be associated with a citizenry more interested in reducing smoking. In turn, the latter could be reflected in the adoption of clean indoor air laws. Examining the joint effects of the two policies confirmed researchers' suspicions, thereby suggesting a reduced (but still quite substantial) impact of taxation on smoking. Similarly, disentangling the impacts of state-level tobacco control programs that mix tax increases with media antismoking campaigns is important but analytically challenging. Recent research paves the way toward better evaluation of multiple-component policies (Chaloupka 1999, Chaloupka and Warner 2000).

6. The Future of Social and Behavioral Science Contributions to Tobacco Control

Social and behavioral science have contributed greatly toward understanding how and why the epidemic of smoking has evolved throughout the twentieth century. They have also elucidated a set of tools that can help society to extricate itself from the tenacious grip of this public health disaster.

With innovations in information and pharmacological technology combining with better insights into human behavior, further improvements in assisting smokers to quit appear to be virtually certain. Although the long-range goal must focus on preventing future generations of children from starting to smoke, helping current adult smokers to quit remains critical to reducing the tobacco-attributable burden of disease and mortality over the next three decades. As such, the role of social and behavioral scientists in smoking cessation will likely be increasingly productive.

The effectiveness of efforts to encourage smokers to quit depends to a significant degree on the environment in which smoking occurs. If smoking is increasingly viewed as antisocial, quitting smoking will become easier, and more urgent, for current smokers. Policy making is society's best means of intentionally altering the environment; and better understanding of the effects of policy interventions, and of how they come to be adopted, will be crucial to shaping the environment with regard to smoking in the coming years.

The multidimensional problems associated with tobacco use vividly illustrate the need for scientists of all disciplinary persuasions to work together. The next great challenge confronting the diverse fields of social and behavioral science in tobacco control is how to combine the methods and insights of the various disciplines into an integrated whole (Chaloupka 1999).

See also: Drug Addiction; Drug Addiction: Sociological Aspects; Drug Use and Abuse: Cultural Concerns; Drug Use and Abuse: Psychosocial Aspects; Smoking Prevention and Smoking Cessation; Substance Abuse in Adolescents, Prevention of

Bibliography

Agency for Health Care Policy and Research 1996 *Smoking Cessation: Clinical Practice Guideline, No. 18, Information for Specialists*. US Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research (AHCPR Publication No. 96-0694), Rockville, MD

American Council on Science and Health 1999 *Environmental Tobacco Smoke: Health Risk or Health Hype?* American Council on Science and Health, New York

Brownson R C, Eriksen M P, Davis R M, Warner K E 1997 Environmental tobacco smoke: Health effects and policies to reduce exposure. *Annual Review of Public Health* **18**: 163–85

Chaloupka F J 1999 Macro-social influences: The effects of prices and tobacco control policies on the demand for tobacco products. *Nicotine & Tobacco Research* **1**(51): 105–9

Chaloupka F J, Grossman M, Bickel W K, Saffer H (eds.) 1999 *The Economic Analysis of Substance Use and Abuse: An Integration of Econometric and Behavioral Economic Research*. University of Chicago Press, Chicago

Chaloupka F J, Laixuthai A 1996 US trade policy and cigarette smoking in Asia. National Bureau of Economic Research Working Paper No. 5543. NBER, Cambridge, MA

Chaloupka F J, Warner K E 2000 The economics of smoking. In: Culyer A J, Newhouse J P (eds.) *Handbook of Health Economics*. Elsevier, Amsterdam

Cromwell J, Bartosch W J, Fiore M C, Hasselblad V, Baker T 1997 Cost-effectiveness of the clinical practice recommendations in the AHCPR guideline for smoking cessation. *Journal of the American Medical Association* **278**: 1759–66

Evans W N, Farrelly M C 1998 The compensating behavior of smokers: taxes, tar, and nicotine. *RAND Journal of Economics* **29**: 578–95

Fritschler A L, Hoefler J M 1996 *Smoking and Politics: Policy Making and the Federal Bureaucracy*, 5th edn. Prentice Hall, Upper Saddle River, NJ

Goodman J 1993 *Tobacco in History: The Cultures of Dependence*. Routledge, New York

Kagan R A, Vogel D 1993 The politics of smoking regulation: Canada, France, and the United States. In: Rabin R L, Sugarman S D (eds.) *Smoking Policy: Law, Politics, & Culture*. Oxford University Press, New York

Lantz P M, Jacobson P D, Warner K E, Wasserman J, Pollack H A, Berson J, Ahlstrom A 2000 Investing in youth

tobacco control: a review of smoking prevention and control strategies. *Tobacco Control* **9**: 47–63
McGinn A P 1997 The nicotine cartel. *World Watch* **10**(4): 18–27
Prabhat J, Chaloupka F J 1999 *Curbing the Epidemic: Governments and the Economics of Tobacco Control*. World Bank, Washington, DC
Prochaska J O, DiClemente C C 1983 Stages and processes of self-change of smoking: toward an integrative model of change. *Journal of Consulting and Clinical Psychology* **51**: 390–5
Roemer R 1993 *Legislative Action to Combat the World Tobacco Epidemic*, 2nd edn. World Health Organization, Geneva, Switzerland
Royal College of Physicians of London 1962 *Smoking and Health: A Report on Smoking in Relation to Cancer of the Lung and other Diseases*. Pitman Publishing Co., London
Saffer H, Chaloupka F J 1999 Tobacco advertising: economic theory and international evidence. National Bureau of Economic Research Working Paper No. 6958. NBER, Cambridge, MA
Strecher V J 1999 Computer-tailored smoking cessation materials: A review and discussion. *Patient Education and Counseling* **36**: 107–17
US Department of Health and Human Services 1988 *The Health Consequences of Smoking: Nicotine Addiction*. (DHHS Publication No. (CDC) 88-8406), Government Printing Office, Washington, DC
US Department of Health and Human Services 1989 *Reducing the Health Consequences of Smoking: 25 Years of Progress. A Report of the Surgeon General*. US Department of Health and Human Services, Public Health Service, Centers for Disease control, National Center for Chronic Disease Prevention and Health Promotion, Office of Smoking and Health (DHHS Publication No. (CDC) 89-8411), Rockville, MD
US Department of Health and Human Services 1990 *The Health Benefits of Smoking Cessation. A Report of the Surgeon General*. US Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Chronic Disease Prevention and Health Promotion, Office of Smoking and Health (DHHS Publication NO. (CDC) 90-8416), Rockville, MD
US Department of Health and Human Services 1994 *Preventing Tobacco Use Among Young People: A Report of the Surgeon General*. US Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. US Government Printing Office, Washington, DC
US Environmental Protection Agency 1992 *Respiratory Health Effects of Passive Smoking: Lung Cancer and Other Disorders*. US Environmental Protection Agency, Washington, DC
US Public Health Service 1964 *Smoking and Health. Report of the Advisory Committee to the Surgeon General of the Public Health Service*. US Department of Health, Education, and Welfare, Public Health Service, Center for Disease Control. PHS Publication No. 1103, Washington, DC
Warner K E 2000 The economics of tobacco: Myths and realities. *Tobacco Control* **9**: 78–89
Warner K E 2001 Reducing harms to smokers: Methods, their effectiveness, and the role of policy. In: Rabin R L, Sugarman S D (eds.) *Regulating Tobacco: Premises and Policy Options*. Oxford University Press, New York
World Health Organization 1997 *Tobacco or Health: A Global Status Report*. World Health Organization, Geneva, Switzerland
Wynder E L, Graham E A 1950 *Journal of the American Medical Association* **143**: 329–96

K. E. Warner

Smoking Prevention and Smoking Cessation

Cigarettes became popular after 1915. Today, approximately 1.1 billion people aged 15 years and over smoke: about one-third of the global population (see *Smoking and Health*). This article provides an overview of health effects, and methods for smoking prevention and cessation.

1. *Health Effects*

In the late 1940s, epidemiologists noticed that annual death rates due to lung cancer had increased fifteenfold between 1922 and 1947 in several countries. Since the middle of the twentieth century, tobacco products contributed to more than 60 million deaths in developed countries. The estimated annual mortality is 540,000 in the European Union, 461,000 in the USA, and 457,000 in the former USSR (Peto et al. 1994). Tobacco is held responsible for three and a half million deaths worldwide yearly: about seven percent of all deaths per year. This figure will grow to ten million deaths per year by the 2020s; about 18 percent of all deaths in developed countries and 11 percent of all deaths in developing countries. Half a billion people now alive will then be killed by tobacco products; these products will have killed more people than any other single disease (WHO 1998).

More than 40 chemicals in tobacco smoke cause cancer. Tobacco is a known or probable cause of about 25 diseases. Tobacco is recognized as the most important cause of lung cancer, but it kills even more people through many other diseases, including cancers at other sites, heart disease, stroke, emphysema, and other chronic lung diseases. Smokeless tobacco and cigars also have deadly consequences, including lung, larynx, esophageal, and oral cancer (USDHHS 1994). Lifetime smokers have a 50 percent chance of dying from tobacco. Half of these will die in middle age, before age 70, losing 22 years of normal life expectancy.

Exposure to environmental tobacco smoke (ETS) has been found to be an established cause of lung cancer, ischemic heart disease, and chronic respiratory disease in adults. Reported effects for children are sudden infant death syndrome, bronchial hyperresponsiveness, atopy, asthma, respiratory diseases,

reduced lung function, and middle ear disease. Barnes and Bero (1998) demonstrated from 106 reviews that conclusions on ETS were associated with the affiliations of the researchers. Overall, 37 percent concluded that passive smoking was not harmful to health. However, 74 percent of these were written by researchers with tobacco industry affiliations. The only factor associated with concluding that passive smoking is not harmful was whether the author was affiliated with the tobacco industry.

Declining consumption in developed countries has been counterbalanced by increasing consumption in developing countries. Globally, 47 percent of men and 12 percent of women smoke. In developing countries 48 percent of men and 7 percent of women smoke, while in developed countries 42 percent of men smoke as do 24 percent of women. Tobacco use is regarded as the single most important public health issue in industrialized countries.

2. *Smoking Prevention*

The process of becoming a smoker can be divided into several stages: preparation (never smoking), initial smoking, experimental or occasional (monthly) smoking, and regular (weekly and daily) smoking. In the preparatory stage attitudes towards smoking are formed. While at least 90 percent of the population has ever smoked a cigarette, the likelihood of becoming a regular smoker increases if initial smoking is repeated several times. In the third stage, a child learns how to smoke and the perceived advantages of smoking start to outweigh the disadvantages. In the fourth stage smoking becomes a routine. Most onset of smoking takes place during adolescence between age 10 and 20. Smoking onset as well as cessation are influenced by a variety of cultural (e.g., availability, litigation, smoke free places), biological (addiction), demographic (e.g., socioeconomic status, parent education), social factors (e.g., parental and peer smoking, parental and peer pressure, social bonding), and psychological factors (e.g., self-esteem, attitudes, self-efficacy expectations).

Various efforts have been undertaken to prevent youngsters to start to smoke (see e.g., Hansen 1992, Reid et al. 1995, USDHHS 1994).

2.1 *School-based Prevention Programs*

Three types of school programs can be distinguished (see *Health Promotion in Schools*). Knowledge programs were not effective. The social influence approaches result in reduced onset ranging from 25 to 60 percent; effects may persist up to four years. Long term effects are found with programs embedded in more comprehensive community approaches. The method includes five to ten lessons, emphasizes short-term consequences of smoking, discusses social (mostly peer) pressures, and includes refusal skills training. Life skills approaches focus on the training of generic life skills. Their effects are less strong than those of the social influence approaches.

2.2 *Out-of-school Approaches*

Youngsters can also be reached out of school. Mass media approaches and nonsmoking clubs are popular methods. They attract attention to the subject and may influence attitudes. A review of 63 studies about the effectiveness of mass media found small effects on behavior (Sowden and Arblaster 1999). Targeting smoking parents is important as well; children are almost three times as likely to smoke if their parents do. Helping parents to quit smoking may prevent their children from starting to smoke and may encourage adolescent cessation.

2.3 *Policies*

A range of policy interventions can be used to stimulate the prevention of smoking.

(a) *Price policies* can have preventive effects. Higher prices encourage cessation among current smokers and discourage initiation among young smokers. A price elasticity of -0.05 implies that a ten percent increase in price reduces consumption or prevalence by five percent. Price elasticity ranges between -0.20 to -0.50.

(b) *School policies* can stimulate smoking prevention. When examining the impact of school smoking policies over 4,000 adolescents in 23 Californian schools, it was found that schools with many smoking prevention policies had significantly lower smoking rates than schools with fewer policies and less emphasis on smoking prevention.

(c) *Public smoking restrictions* can contribute to adolescents' beliefs that nonsmoking is normative and that smoking creates health problems. Smoking regulations are more effective in preventing teenagers from starting to smoke than in reducing their consumption. Restriction of sales to minors often results in noncompliance: between 60 to 90 percent of the adolescents succeeded to buy tobacco products in situations where this was not allowed.

3. *Smoking Cessation*

Physical addiction is caused by the pharmacological effects of nicotine. Psychological addiction occurs because smoking becomes linked with numerous daily circumstances and activities (e.g., eating, drinking) and emotional and stressful events. A person becomes motivated to quit if he has a positive attitude, encounters social support, and has high self-efficacy expectations towards quitting (De Vries and Mudde 1998).

Cessation is a process: a smoker in precontemplation is not considering to quit within six months; a smoker in contemplation is, but not within a month; a smoker in preparation is considering to quit within a month; a person in action has quit recently; a person in maintenance has quit for more than six months (Velicer and Prochaska 1999).

Three outcome measures can be used to assess smoking cessation: point prevalence (having smoked during the preceding seven days), prolonged abstinence (not smoked during six or twelve months), and continuous abstinence (not smoked at all since the time of intervention). In longitudinal experimental designs more smokers than quitters may drop out, thus resulting in too optimistic estimates of successes of treatments. To correct for this bias, dropouts are coded as smokers; this is referred to as the 'intention to treat procedure.' This procedure may, however, result in conservative effect estimates. The desirability of biochemical validation is still controversial. Misreporting seldom exceeds five percent. Detection of occasional smoking in youngsters is difficult and expensive. Biochemical validation of self-reports should be considered when high demand situations are involved. A random subsample can be used to estimate bias and correct reported cessation rates. Cotinine has emerged as the measure of choice, with carbon monoxide as a cheaper but less sensitive alternative (Velicer et al. 1992).

Comparing the results of cessation studies is hindered by differences in outcome measures, populations (e.g., the percentage of precontemplators), and follow-up periods. This overview includes methods that have evidence for success. The numerous studies assessing and reviewing the efficacy of smoking cessation interventions provide different success rates. Hence, the figures reported below are estimates derived from various studies that are reported below, as well as from inspection of other publications.

3.1 Pharmacotherapy

Pharmacotherapeutic interventions increase quit rates approximately 1.5- to 2-fold. The absolute probability of not smoking at 6–12 months is greater when additional high intensity support is provided (Hughes et al. 1999, Silagy et al. 1999).

Nicotine Replacement Therapy (NRT) products are available in a number of forms, including gum, transdermal patch, nasal spray, lozenge, and inhaler. NRT is recommended to be part of the core treatment package offered to all smokers. There are few instances in which the use of NRT is contraindicated. A meta analysis including 49 trials found significant effects when applying gum, patches, nasal spray, inhaled nicotine, and sublingual tablet. These effects were largely independent of the intensity of additional support provided or the setting in which the NRT was offered. The efficacy of NRT appears also to be largely independent of other elements of treatment although absolute success rates are higher with more intensive behavioral support. The NRT cessation effects range from 6 to 35 percent, mostly ranging between 15 and 25 percent.

Bupropion is an atypical antidepressant that has both dopaminergic and adrenergic actions. Unlike NRT, smokers begin bupropion treatment one week prior to cessation. The recent evidence shows that cessation rates are comparable to those by NRT, between 15 and 25 percent (Hurt et al. 1997, Jorenby et al. 1999). One study found that the nicotine patch was less effective than bupropion (Jorenby et al. 1999). The drug appears to work equally well in smokers with and without a past history of depression.

3.2 Motivational Strategies

Various studies report on the effectiveness of group courses, self-help materials, computer tailoring, and competitions (e.g., Eriksen and Gottlieb 1998, Fisher et al. 1993, Matson et al. 1993, Skaar et al. 1997, Strecher 1999, Velicer and Prochaska 1999).

Group courses can be effective. Studies suggest 10 to 30 percent abstinence rates. The disadvantage is that they attract mostly smokers who are highly motivated to quit, and that many smokers want to quit on their own.

Self-help cessation interventions use different formats such as brochures, cassettes, and self-help guides. Point prevalence quit rates at one-year follow-up range between 9 and 15 percent.

Competitions were effective in three out of five workplace studies. No study showed enhanced smoking cessation past six months. Net effects for cessation rate of competition plus group cessation program over group program alone in three studies varied between 1 to 25 percent.

Computer tailoring results in personalized materials for smokers which are based on the (cognitive) characteristics of a person that were measured by a questionnaire. Large segments of smokers can be reached (De Vries and Brug 1999). A review of ten randomized trials found significant effects up to 25–30 percent.

3.3 Policy Interventions

Policy interventions can effect cessation (Cummings et al. 1997, Eriksen and Gottlieb 1998, Glantz et al. 1997, Meier and Licari 1997, Zimring and Nelson 1995).

Price policies, as with adolescents, influence cessation rates. Price elasticity rates for adults range between -0.3 and -0.55. Increases in cigarette prices may place a greater burden on those with lower incomes who tend to have greater difficulty in stopping smoking.

Smoke-free areas can be effective as well, although evidence is not unequivocal. Smoking bans in workplaces resulted in reduced tobacco consumption or cessation at work, results ranging from 12 percent to 39 percent. The findings on reductions on prevalence were not consistent.

Advertising, including tobacco sponsorship of events, is believed to stimulate the uptake of smoking and to reinforce the habit. This finding was supported by results from the COMMIT trial. The use of health warnings, however, has been found to be able to reduce tobacco consumption. Litigation by states and other groups may change the way tobacco is advertised and sold.

3.4 Special Settings for Cessation

Multiple smoking cessation interventions can be applied in specific settings. The goal is to attract large segments of smokers, or particular segments of smokers.

Advice by health intermediaries (physicians, nurses, midwives) is effective. A review describing the efficacy of 188 randomized controlled trials reported a two percent cessation rate resulting from a single routine consultation by physicians. While modest, the results were cost-effective. The effects are most salient among special risk groups, such as pregnant women and patients with cardiac diseases, although the additional cessation rates range from approximately 5 to 30 percent (Law and Tang 1995). In workplace settings cessation rates were four percent for a simple warning, eight percent for short counseling. They were highest among high-risk groups: 24 percent among workers at high risk for cardiovascular problems and 29 percent among asbestos exposed workers (Eriksen and Gottlieb 1998).

Community interventions have the potential to reach all segments of a community. They combine different methods, such as mass media approaches, counseling, and self-help materials. Both the broader cardiovascular risk reduction interventions and those focusing solely on smoking had very small effects (1–3 percent cessation rates), although somewhat larger effects may be reached when combining them with mass media approaches (Fisher et al. 1993, Velicer and Prochaska 1999).

Workplace programs also often combine methods. They can reach large and diverse groups of smokers and may produce an average long-term quit rate of 13 percent (ranging from 1.5 percent to 37 percent) at an average of 12 month's follow-up regardless of the intervention methods (Eriksen and Gottlieb 1998). Group programs were more effective than minimal programs, although less intensive treatments when combined with high participation rates can influence the total population as well.

4. Conclusions

School-based smoking prevention can prevent or delay smoking onset. It is unrealistic to expect long-term effects with short programs when prosmoking norms are communicated through multiple channels. Consequently, at least ten lessons are needed during adolescence in several grades. Outside of school, programs are needed, since adolescents with the highest rates of tobacco use are least likely to be reached through school based programs (Stanton et al. 1995). Innovative cessation interventions are needed as well, because youth cessation programs show no long-term success. Since multiple interventions are found to be most effective, priority should be given to broad-based interventions aimed at both the youngsters, the schools, the parents, and the community as a whole, including mass campaigns for all age groups, fiscal policies, restrictions on smoking, and bans on advertising.

Pharmacotherapy, motivational strategies, and policies are effective tools to promote cessation. The potential of other antidepressants (e.g., nortriptyline) on smoking cessation need attention as well. Innovative media strategies, such as computer tailoring, may reach less motivated smokers. Specific attention for underprivileged populations (e.g., ethnic populations, low socioeconomic groups) is needed as well. The combination of multiple methods given by multiple health care providers on multiple occasions will result in the greatest impact. Future research should assess the effectiveness of programs in smokers differing in motivational stages, since the most effective methods mostly attract smokers who are motivated to quit.

Effective methods that only reach a small segment of motivated smokers have a low impact. Less effective methods that reach large segments of the population have a higher impact. Velicer and Prochaska (1999) suggest the greatest impact for computer-tailored programs. Finally, cost-effectiveness studies reveal that cessation methods are most cost-effective when aimed at a particular subgroup, multiple interventions, and at maintaining abstinence.

See also: Health Behaviors; Health Education and Health Promotion; Health: Self-regulation

Bibliography

Barnes D E, Bero L A 1998 Why review articles on the health effects of passive smoking reach different conclusions? *Journal of the American Medical Association* **279**: 1566–70

Cummings K M, Hyland A, Pechacek T, Orlandi M, Lynn W R 1997 Comparison of recent trends in adolescent and adult cigarette smoking behaviour and brand preferences. *Tobacco Control* **6:** suppl: 31–7

De Vries H, Brug H 1999 Computer-tailored interventions motivating people to adopt health promoting behaviours; introduction to a new approach. *Patient Education and Counseling* **36**: 99–105
De Vries H, Mudde A 1998 Predicting stage transitions for smoking cessation applying the Attitude–Social influence–Efficacy model. *Psychology & Health* **13**: 369–85
Eriksen M P, Gottlieb N H 1998 A review of the health impact of smoking control at the workplace. *American Journal of Health Promotion* **13**: 83–104
Fisher B E, Lichtenstein E, Haire-Joshu D, Morgan G D, Rehberg H R 1993 Methods, successes, and failures of smoking cessation programs. *Annual Reviews of Medicine* **44**: 481–513
Glantz S A, Fox B J, Lightwood J M 1997 Tobacco litigation. Issues for public health and public policy. *Journal of the American Medical Association* **277**: 751–3
Hansen W B 1992 School-based substance abuse prevention: a review of the state of art in curriculum, 1980–1990. *Health Education Research* **7**: 403–30
Hughes J R, Goldstein M G, Hurt R D, Shiffman S 1999 Recent advances in the pharmacotherapy of smoking. *Journal of the American Medical Association* **281**: 72–6
Hurt R D, Sachs D P L, Glover E D et al. 1997 A comparision of sustained-release bupropion and placebo for smoking cessation. *New England Journal of Medicine* **337**: 1195–202
Jorenby D E, Leishow S J, Nides M A, Rennard S I, Johnston J A, Hughes A R, Smith S S, Muramoto M L, Daughton D M, Doan K et al. 1999 A controlled trial of sustained-release bupropion, a nicotine patch or both for smoking cessation. *New England Journal of Medicine* **340**: 685–91
Law M, Tang J L 1995 An analysis of the effectiveness of interventions intended to help people stop smoking [see comments]. *Archives of International Medicine* **155** (18): 1933–41
Matson D M, Lee J W, Hopp J 1993 The impact of incentives and competitions on participation and quit rates in worksite smoking cessation programs. *American Journal of Health Promotion* **7**: 270–80
Meier K J, Licari M J 1997 The effect of cigarette taxes on cigarette consumption 1995–1994. *American Journal of Public Health* **87**: 1126–30
Peto R, Lopez A D, Boreham J, Thun M, Heath C 1994 *Mortality from Smoking in Developed Countries 1950–2000*. Oxford University Press, Oxford, UK
Reid D J, McNeill A, Glynn T 1995 Reducing the prevalence of smoking in youth in Western countries: An international review. *Tobacco Control* **4**: 266–77
Skaar K L, Tsoh J Y, McClure J B, Cinciripini P, Friedman K, Wetter D W, Gritz E 1997 Smoking cessation 1: An overview of research. *Behavioral Medicine* **23**: 5–13
Silagy C, Mant D, Fowler G, Lancaster T 1999 Nicotine replacement therapy for smoking cessation (Cochrane Review). In: *The Cochrane Library*, Issue 4. Oxford, UK. Update Software
Sowden A J, Arblaster L 1999 Mass media interventions for preventing smoking in young people (Cochrane Review). In *The Cochrane Library*, Issue 1. Oxford, UK. Update Software
Stanton W R, Gillespie A M, Lowe J R 1995 Reviewing the needs of unemployed youth in smoking intervention programmes. *Drug and Alcohol Review* **14**: 101–8
Strecher V J 1999 Computer-tailored smoking cessation materials: A review and discussion. *Patient Education and Counseling* **36**: 107–17
USDHHS 1994 *Preventing Tobacco Use Among Young People*; *A Report of the Surgeon General*. US Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA
Velicer W, Prochaska J 1999 An expert system intervention for smoking cessation. *Patient Education and Counselling* **36**: 119–29
Velicer W, Prochaska J, Rossi J, Snow M 1992 Assessing outcome in smoking cessation studies. *Psychological Bulletin* **111**: 23–41
WHO 1998 Tobacco epidemic: Health dimensions. *Fact Sheet* 154. World Health Organisation, Geneva, Switzerland
Zimring F E, Nelson W 1995 Cigarette taxes as cigarette policy. *Tobacco Control* **4:** suppl: 25–33

H. de Vries

Soap Opera/Telenovela

'Soap opera' is a pejorative term coined in the 1930s by the entertainment trade press of the United States to designate the daytime dramatic serials that were broadcast on radio and aimed primarily at women. The term 'soap' was due to the sponsorship of the programs, which usually were produced by the advertising agencies of the soap and toiletry industries (see *Advertising Agencies*). The term 'opera' refers ironically to the dramatic character of the genre and its 'low cultural quality.' Soap operas are daytime serials with a dramatic content aimed primarily at a female audience. The genre emerged in the early period of the radio industry and has been broadcast by television stations since the 1950s. 'Telenovela' is a term used throughout Latin America to designate the melodramatic serials that became the most popular programs of the television industries of the region. They have been exported to every continent in the world. Although telenovelas have a similar history and common features when compared to soap operas, they are two distinct genres (see *Television: Genres*).

1. Soap Opera

The soap opera emerged in the US radio industry of the early 1930s. The genre developed as a response to a fundamental problem for the commercial radio stations and the advertising agencies: how to achieve the largest audience of potential consumers of certain products. The advertising agencies of the soap, toiletry, and foodstuff industries took on the role of developing programs that could attract the members of the general audience that were the main potential consumers: women between the ages of 18 and 49. When searching for formulas that could do this job, the agencies found in serials a type of narrative that

proved to be suitable in creating a faithful audience. By the late 1920s, a program that relied upon narrative and fictional characters, *Amos 'n' Andy*, established a crucial precedent by demonstrating the appeal of the radio serial form. The program presented the adventures of two Southern black men living in the city of Chicago and achieved an estimated daily audience of 40 million listeners by 1929.

The first radio programs that can be classified as soap operas appeared in the early 1930s and were based on these early successful experiments. The serial *Painted Dreams*, first broadcast in October 1930 by the station WGN, told the story of an Irish woman, her household, and daughter. The program was written by a school teacher turned radio actress, Irna Phillips, who has been credited with 'inventing' the genre soap opera. Phillips wrote many successful plots, for example, *The Guiding Light*, the longest running soap ever. The serial was introduced in January 1937, and lasted for 19 years on radio. It is still broadcast on television by CBS. Frank Hummert and his wife, Anne, were other key figures in the birth of the soap opera. Their advertising agency created one of the first daytime serials ever produced, *Stolen Husband*, in 1931. Although the serial failed, it helped the Hummerts develop the formula that proved to be successful in attracting large audiences.

Between 1933 and 1937, soap operas consolidated their dominant position in radio daytime programming and became the primary advertising vehicle for the soap, toiletry, and foodstuff industries. Although the soap opera faced a decline in broadcasting time and in number after this period of expansion, it remained one of the most popular genres in the history of radio. By 1940, the 64 serials being broadcast each day constituted 92 percent of all sponsored daytime broadcast hours and the 10 highest-rated daytime programs were all soap operas (Allen 1985). It is estimated that in the same year, one of the best in the history of daytime serials, about 20 million female viewers, approximately half the women at home during the day, listened to two or more serials daily (Cantor and Pingree 1983) (see also *Radio as Medium*).

The radio era of the soap opera in the US was terminated by the advent of a new technology, television, which would soon become the most important advertising medium. By 1955, radio soap operas were practically eliminated and the last ones were discontinued in the 1960–61 season. When the last daytime serials left radio in the early 1960s, they were already an established form of television programming and advertisers continued to produce soap operas after the transition to television. Proctor and Gamble, for example, produced the first television network soap opera, *The First Hundred Years*, first broadcast by CBS in December 1950. The reason for the longevity of the daytime serials and their successful transition to television was the suitability of fulfilling one key demand of the commercial broadcasting system: the need to attract the audience of heavy consumers in a cost-effective manner.

The main feature of the narrative text offered by the soap operas is 'openness.' Soaps are characterized by the absence of ultimate closure, since the plots never begin or end, staying on the air for several years or decades. The central elements of the serials' content usually gravitate towards issues such as love, family, intimate relationships, and other domestic concerns, although the genre has been characterized by a process of diversification. As we have seen, the history of the soap opera can be traced to the US. Nevertheless, the dramatic serials have become a common programming form all over the world, with specific national peculiarities (e.g., Allen 1995). In Europe, for example, an analysis of the network of family and romantic relationships portrayed by soap operas has shown that European serials are not simply an American genre which was imported. Europe has developed distinctive subtypes of the genre, most notably the 'community model' in which loves and family conflicts are colored by more or less pedestrian hardships of sickness, unemployment, and teenage drug habits (Liebes and Livingstone 1998).

By definition, soap operas are daytime serials. Nevertheless, in the case of the US, some prime-time dramatic serials, such as *Dallas* and *Dynasty*, have been broadcast during prime-time hours and have achieved a huge success both on the national and the international level. These prime-time serials frequently have been defined as soap operas, but they differ from daytime serials in several aspects. Prime-time serials demand a more expensive production structure and have fewer episodes. They also have a faster pace that leads to more resolution of conflicts when compared to soap operas (Cantor and Pingree 1983).

Despite the fact that soap operas have been one of the most popular and central genres in the history of television (see *Television: History*), there were very few published works on television serials until the early 1970s. This 'invisibility' of the soap opera was due, to a great extent, to the low position it occupies in the hierarchy of taste in the cultural field (see *Taste and Taste Culture*) and to the unequal gender relations that constrain it. Media, academic, and public discourses about soap operas have been marked by the fact these serials are 'woman's forms.' It was feminism that has transformed soap opera into a field for academic inquiry, although some feminists have also responded in a hostile manner toward the genre. Initially, feminist scholars used textual analysis to compare 'real' women with their stereotypical and unrealistic images (the sex object and the housewife), criticizing soap operas' tendency to confirm female subordination. Later on, feminists used more varied methodological approaches, including qualitative audience research, to affirm and authenticate the pleasures offered by the genre to its female audiences. The feminist notion that 'the personal is political' helped turn media research in

the 1970s and 1980s away from news and current affairs toward 'softer' programs. Ensuing studies found soap operas' focus on the domestic sphere of women's everyday lives to provide meaningful, rather than escapist, entertainment; they tended to validate, rather than condemn, the role that soap opera viewing plays in women's lives (see Brunsdon 1997, Brunsdon et al. 1997).

2. *Telenovela*

Telenovelas are melodramatic serials produced in Latin America. They have become the most popular programs of the television industries of the region and a successful phenomenon in the international market. Although telenovelas have features in common with soap operas, they are two distinct genres. Soap operas are broadcast during the day, usually in the afternoon, while telenovelas dominate the prime-time slots of Latin American TV networks. As a result, soap operas are watched primarily by women while telenovelas attract a broader audience in terms of age and gender. If soaps are 'open' forms, broadcast for years or decades, telenovelas are 'closed,' having a clear beginning and an end, and lasting approximately 180–200 episodes or 6–8 months. Finally, if soaps are characterized by authorial anonymity, telenovela writers in Latin America, particularly in Brazil, have personal styles that get recognized by the public (see Vink 1988).

As with their American counterparts, telenovelas have their roots in the daytime serials that were broadcast by the commercial radio stations, the *radionovelas*. Early on, Cuba consolidated a commercial broadcasting system, which became an important production center of radio melodramatic serials. By 1930, the city of Havana had proportionally more radio stations than New York and in 1932 Cuba was the third country in the world in terms of the number of radio receivers (Ortiz et al. 1989). A key figure in the development of the radio serials in Latin America was the Cuban writer Felix Caignet. After the first *radionovelas* were broadcast in Cuba around 1935, Caignet started writing several melodramatic texts which became popular not only in Cuba, but all over Latin America. For example, the story *El Derecho de Nacer* (*The Right to Be Born*) has been broadcast by radio stations all over the region for several decades and was adapted with great success for television in the 1960s.

After the introduction of the first television stations in Mexico and Brazil in 1950, telenovelas gradually became the most popular programming form of the television industries of the continent. Since the late 1960s, telenovelas have dominated prime-time and commanded the highest advertising rates. Globo Network in Brazil and Televisa in Mexico became giant media corporations and their successful programming strategies have been based largely on the local production of telenovelas. In the late 1970s, after becoming virtual monopolies in their national markets, these companies began exporting telenovelas with great success. Globo pioneered the conquest of the international market in the early 1980s by exporting the telenovela *Escrava Izaura* (*Izaura, the Slave*) that became a national obsession in countries like Poland, China, and Cuba. On the other hand, Mexico concentrated first on Latin America and Hispanic audiences in the US. Later on, some of its telenovelas, as *Los Ricos También Lloran* (*The Rich also Weep*), were successful phenomena in countries like Italy and Russia. Latin American telenovelas have been viewed in more than 120 countries and besides Globo and Televisa, the telenovelas from Venevision Network of Venezuela have also been successful in the world market (see Mattelart and Mattelart 1990, Lopez 1995, Martin-Barbero 1995).

The Latin American telenovelas focus on melodramatic conflicts, passion, tragic suffering, and moral conflicts. There are, nevertheless, important national peculiarities in the development of the genre. Mexican telenovelas are known for their weepiness and lack of specific historical references, while Brazilian telenovelas are more 'realistic' in their depiction of specific historical and political events of the country (see Lopez 1995), although more recently Mexican telenovelas such as *Nada Personal* (*Nothing Personal*) and Colombian serials like *Por estas Calles* (*In These Streets*) have also dealt with political and social contexts in a more explicit way. Latin American telenovelas have, therefore, played a growing role in representing the cultural, social and political conflicts that have shaped the region.

See also: Adolescents: Leisure-time Activities; Cultural Policy; Mass Communication: Normative Frameworks; Mass Media and Cultural Identity; Mass Society: History of the Concept; Media and History: Cultural Concerns; Media and Social Movements; Media Imperialism; Media, Uses of; Popular Culture; Television: Genres; Television: History; Television: Industry

Bibliography

Allen R C 1985 *Speaking of Soap Operas*. University of North Carolina Press, Chapel Hill, NC and London

Allen R C 1995 (ed.) *To Be Continued… Soap Operas Around The World*. Routledge, London and New York

Brunsdon C 1997 *Screen Tastes: Soap Opera to Satellite Dishes*. Routledge, London and New York

Brunsdon S, D'Acci J, Spigel L (eds.) 1997 *Feminist Television Criticism: A Reader*. Clarendon Press, Oxford, UK

Cantor M G, Pingree S 1983 *The Soap Opera*. Sage, Beverly Hills, CA

Liebes T, Livingstone S 1998 European soap operas: The diversification of a genre. *European Journal of Communication* **13**: 147–80

Lopez A 1995 Our welcomed guests: Telenovelas in Latin America. In: Allen R C (ed.) *To Be Continued… Soap Operas Around The World.* Routledge, London and New York

Martin-Barbero J 1995 Memory and form in the Latin American soap opera. In: Allen R C (ed.) *To Be Continued… Soap Operas Around The World.* Routledge, London and New York

Mattelart M, Mattelart A 1990 *The Carnival of Images: Brazilian TV Fiction.* Bergin & Garvey, New York

Ortiz R, Borelli S, Ramos J 1989 *Telenovela: História e Produção*, 1st edn. (Telenovela: History and Production). Brasiliense, São Paulo

Vink N 1988 *The Telenovela and Emancipation: A Study on Television and Social Change in Brazil.* Royal Tropical Institute, Amsterdam

M. P. Porto

Sociability in History

Sociability is a notion used by historians to apprehend different forms of social relationships, in particular interpersonal bonds that are initiated either consciously or unconsciously in a given context. General distinctions are made, according to their social character, between bourgeois sociability and popular sociability, and, according to the shape it assumes, between formal sociability and informal sociability.

1. *Sociological Background*

Before historians adopted the term sociability, sociologists had used it in various ways. The first to cite is Georg Simmel, who made the notion of *Geselligkeit* (sociability) a key concept of his formal sociology. In his *Grundfragen der Soziologie* (1917), Chapter III is entitled *Die Geselligkeit. Beispiel der Reinen oder formalen Soziologie*. According to Simmel, sociability is a form of socialization, engendered by interactions exercised among individuals on the basis of reciprocity and equality. This new sociological approach made it possible to shift the emphasis of the axis of a social analysis from the content to the form of the social games.

Another sociologist for whom the notion of sociability was fundamental was Georges Gurvitch. For him, sociability comes under what he calls microsociology. In *La vocation actuelle de la sociologie* (1950), a chapter is devoted to the microsociological scale where different forms of sociability are dealt with. One significant point is his distinction between organized sociability and spontaneous sociability. This corresponds by and large with the historians' distinction between formal sociability and informal sociability.

2. *Adoption by Historians of the Notion of Sociability*

Agulhon was the first to use this notion as a key concept in his historical analysis of Provençal society. Agulhon was not inspired directly by the sociological theories mentioned above; their interpenetration came about later little by little. The starting point was his dissertation of 1966, *La sociabilité méridionale: Confréries et associations en Provence orientale dans la deuxième moitié du XVIIIe siècle*. In this work, Agulhon threw into relief an extremely dense network of social bonds, given concrete expression at the popular level in the shape of confraternities of penitents, and at the higher level, masonic lodges. He recognized in these phenomena certain characteristics of Provençal society, and called them *sociabilité méridionale*, or *southern sociability*. He later extended his research beyond the French Revolution to the mid-nineteenth century, and discerned at the base of republican movements developing in the villages a strong tradition of southern sociability. In the first half of the nineteenth century it is to be seen at the popular level in the increasing number of associations known as *chambrées*, and at the higher level in the proliferation of *cercles* (Agulhon 1970).

These pioneering works by Agulhon reverberated widely among social historians and opened up a new direction in the field of social history. The two following points represent the especial novelty of the notion. (a) Until then, the fundamental categories that constituted the very basis of social history had been the notions of social class and nation. Agulhon introduced a new dimension by adopting an ethnological perspective aimed at penetrating the intimacy of everyday life and discerning there the original social bonds that govern the society in question. (b) Social history in France came into being through criticizing traditional political history—so much so that there was a certain tendency to separate politics from social history. That was also the case with the Annales school until the 1970s. Agulhon, on the other hand, succeeded in reintegrating politics and combining it closely with social studies. In this way, his work inaugurated a new political history, closely linked with political anthropology.

The role played by Agulhon is, therefore, very important, but that does not mean that previous or parallel research in the same field had been lacking, even though they did not make use of the notion of sociability. To begin with, the very early case of Georges Lefebvre must be cited. In his 1932 presentation on the revolutionary crowd to the *Semaine de*

Synthèse presided over by Henri Berr (*Foules révolutionnaires* in the Proceedings of the colloquium, 1932), he identified three levels of social grouping: (a) simple (involuntary) crowd; (b) semivoluntary gathering; and (c) (voluntary) assembly.

While the first level corresponds to an amorphous, involuntary crowd, and the third to a body that acts collectively with a common objective and a precise consciousness, like the Revolutionary *sans-culottes*, the second intermediate level represents people who form a certain community in their everyday lives through activities such as communal work in the fields, Sunday mass, evening meetings, markets or fairs, local festivals, or drinking in taverns, and who are thus imbued with their own collective mentality. By identifying these three levels, Georges Lefebre especially underlined the importance of this second level for correctly apprehending the popular movements of the Revolutionary period. It must be noted that what he here calls 'semi-voluntary gathering' is no other than informal sociability.

A similar preoccupation is also to be found among historians contemporary with Agulhon. Yves Castan, for example, succeeded in discerning the social bonds that affect ordinary people in southwest France, by discovering the systems of codes that determine their behavior (*Honnêteté et relations sociales en Languedoc au XVIIIe siècle*, 1974). Without actually using the word *sociability*, and without any prior contact with the work of Agulhon, this study by Castan was inspired by the same concerns and remains a classic of the history of popular sociability.

For sociability in the upper levels of society, one of the best examples is presented by Roche. In his doctoral dissertation, he studied some 2,500 provincial academics who constituted a cultural network of their own. Here is another instance of a classic study of sociability without actually using the term (Roche 1978).

3. *Acceptance of the Notion of Sociability in French Historiography*

The term sociability proposed by Agulhon in his first book was favorably accepted by historians such as Philippe Ariès, Emmanuel Le Roy Ladurie, and Daniel Roche, exceeding even the writer's expectations. Vovelle, who compiled in 1980 a survey of the previous ten years' research on southern sociability, concluded: 'Sociability, a notion contrived or rediscovered over ten years ago, has entered the ranks of those supporting concepts needed by the history of mentalities in any attempt to define the collective realities on which it faces' (Vovelle 1982). Agulhon himself revised and refined this theme on various occasions. Other historians also undertook parallel studies, first within the framework of the Mediterranean, and later crossing those frontiers that had initially been assumed. Witness is borne by Guitton, whose book extended the horizons of investigation to villages throughout France, and discovered in northern France, too, intensely real-life social bonds that would constitute a village sociability (Guitton 1979).

As fast as research proceeded, discussion arose around Agulhon's initial hypothesis. Above all, criticism concerned the following two points. (a) While Agulhon considered the density and vivacity of associative life as one of the special features of Provençal life or at most the regions of the south, other researchers such as Guitton, Michel Bozon, or Jean-Luc Marais demonstrated that the phenomenon exceeded the boundaries of Provence, and that the notion of sociability was also applicable to other regions. (b) While distinguishing between the two levels of sociability, namely informal (spontaneous) sociability and formal (organized) sociability, Agulhon attached importance to the evolution from the former to the latter, this latter being the very basis of associative life. Vovelle, however, believes that informal sociability retains a vague but persistent existence, without necessarily taking on any institutional form. Agulhon himself, in his later articles, revised his initial position. He admitted in part that the notion of sociability was operative on a general plan beyond regional temperament. He also acknowledged that informal sociability had a being of its own and played an important role, especially at the popular level (Agulhon 1976, 1986).

Thus, since the 1970s, sociability has been an honored bastion of social history, alongside the history of mentalities or the history of everyday life (see the excellent survey in François and Reichardt 1987). Increasing numbers of monographs have appeared on a wide variety of themes. At the level of informal sociability, different social activities have been studied, both for villages and for urban environments, for example, collective work in the fields, community's assembly, taverns, evening meetings (*veillées*), rough music (*charivari*), carnivals or feasts, youth groups, markets and fairs, pilgrimages, games, and divertissements. All of these activities are supposed to favor a spontaneous sociability, not organized but having organic life (J-P. Guitton, R. Muchembled, N. Pellegrin, M. Bozon). In the urban framework, neighborhoods and the social life in a street or a city quarter have been brought into relief, rather than the municipal institutions that used to be studied by classical urban historians (see Farge 1979). Confraternities and *chambrées*, which Agulhon showed to be so important, now constitute a major theme of research in popular sociability.

As far as formal sociability itself is concerned, what counts above all is the sociability of the bourgeois middle classes or of the intellectual elite. Apart from freemasonry and the provincial academies mentioned earlier, objects of study include local activities of literary or other erudite people, societies of all kinds,

as well as less closed locations frequented by less elevated social strata: cafés or reading clubs (Roche, Jean-Pierre Chaline, Eric Saunier, D. Goodman).

For research as a whole, the Association for Research in Sociability (ARS) at the University of Rouen has made a valuable contribution in its wide perspective of current research. The themes chosen for colloquia that have been organized there display a wide variety of work in progress (ARS 1987, 1989, 1992, 1997).

4. *Beyond French Historiography: Influence or Coincidence*

For Germany and Switzerland, it is a case more of coincidence than of influence. In 1983, François organized an important Franco–German colloquium on sociability. According to him, 'The study of the facts of sociability and of associative life has become one of the most fertile fields for study in the new history,' but 'this parallel discovery—or re-discovery—of sociability as an object of historical research and the resulting profusion of studies have come about in France, as in Germany, in almost complete ignorance of inquiries being carried out at the same instant by historians in the neighboring country' (François 1986). A stinging comment. He also points out that German historians were concentrating on institutionalized associations (*Vereinswesen*), while their French counterparts were interested rather in all forms of sociability (*Geselligkeit*), in particular the informal sociability. This contrast between two neighboring cultural zones is also confirmed by Otto Dann in his presentation to the Franco–German colloquium. He attributes the contrast not to national temperaments, but to the difference in historical conditions experienced by the two countries in the eighteenth and nineteenth centuries. In any case he insists that henceforth for both Germany and France there is a primordial necessity to carry out parallel and joint research on these two complementary dimensions of sociability (François 1986).

In Germany in the eighteenth century, a process of transition in upper-class sociability took place from the aristocracy or the court to the rising middle classes. There was a clear increase in numbers of associations such as academies, scientific societies, musical or literary societies, reading circles, masonic lodges, student associations. Also in the nineteenth century a transference is to be observed from *bourgeois* sociability to the working-class sociability. Workers' educational associations were set up on the initiative of middle-class leaders, and working-class movements for political mobilization came into being. The end of the nineteenth century, with the rise in the consciousness of Germany as a single nation, saw the formation on a national scale of new organizations for sport, for soldiers, and so on. One feature of these associations is that women were excluded. As a result, from 1848 onward, women began setting up their own associations to campaign for their rights to education, and indeed the right to vote.

Because of its importance in the country's cultural, social, and even political life, this kind of institutionalized sociability (*Vereinswesen*) has long been a focus of interest for German historians (Dann 1984, Siebert 1993). But recently they have begun to show interest in another kind of sociability, which is rather less formal and less rigidly organized, as in family or neighborhood ties. Family celebrations, dinner parties, balls, coffee house circles, and family concerts are in their turn being studied from the point of view of sociability. Historians are also examining the blurred boundaries between public and private, and including in their work the problems posed by gender history. It is in this direction that the current of the history of everyday life is flowing (see Lüdtke 1989). Here this new orientation in German historiography can be seen converging with the preoccupations of the French historical anthropology.

The contribution of Swiss historians is important, too. The work of Im Hof not only for Switzerland but also for Europe in the Enlightenment is well known (Im Hof 1982). In his presentation to the Franco-German colloquium mentioned previously, he stresses the liveliness and the density of associative life in Switzerland. During the first half of the nineteenth century there was a blossoming of associations of widely varied types: not only artists, musicians, historians, or students, but also gymnasts, marksmen, and singers banded together in associations. He points out, too, a boom in explicitly democratic sociability, but with a strong patriotic and national inclination—a predominant Swiss characteristic (François 1986).

In the Anglo-Saxon world, the term sociability as a historical notion is not very frequent, but there are certain currents in research which have clearly been inspired by the history of sociability. At the level of informal sociability, neighborhood has become an appealing theme for urban historians: the London suburbs of the seventeenth century and the local community in eighteenth-century Paris have been admirably well analyzed (Boulton 1987, Garrioch 1986). The bonds of sociability in taverns are also considered an essential element of popular culture (Brennan 1988). In this way the approaches of English social history concur with those of French historians of sociability. On the other hand, at the level of the history of ideas, a perceptive study by Gordon points up the correlations between egalitarian-minded associations and the formation of the public sphere (Gordon 1994). This important problem of the public sphere should be further examined in close connection with social practices.

In the Mediterranean world, the influence of French historiography is more strongly felt. In Italy, Gemelli

and Malatesta compiled a volume of extracts from 10 or so major articles in the history of sociability, with a substantial Introduction which succeeds in locating the 'adventures' of sociability in the current of contemporary French historiography (Gemelli and Malatesta 1982). It should be noted that Italian *microstoria* and the new French history share many of their concerns. The works of Levi (1985) or Cerutti (1990) are good examples of this. Moreover, for analyzing urban societies in the Middle Ages and the Renaissance, Italian historians propose the notions of *parenti* (family), *amici* (friends), and *vicini* (neighbors) as elements of the social network. These are very suggestive of the point of view of sociability.

Similarly, outside the Western world, new currents in historical research are to be seen. In Japan, for example, the immediate postwar years saw the birth of a powerful current of social science history which, strongly influenced by Marxism, tried to analyze Japanese society based almost exclusively on the notions of social class and nation-state. For criticizing the nationalistic historians, who had been dominant during the War, or the traditional pure positivists, these concepts worked well and made it possible to propose a structural vision of the history of Japan. Nevertheless, with the passage of time, this unidimensional viewpoint proved too categorical and not capable of explaining the sociocultural background of Japanese history. And the rapid, unexpected transformation of Japanese society after the War made it necessary to reconsider the framework of social analysis. Therefore, since the 1970s, a new current in historical research is to be observed which, endowed with a multidimensional viewpoint, tries first to define history as lived by the people in their everyday lives and from there to approach economic or political aspects in order to reformulate the mechanisms by which Japanese society functions. It is in these new circumstances that notions of mentalities and of sociability have been integrated into Japanese historical research. A colloquium organized in Tokyo in 1994 by Hiroyuki Ninomiya on 'Different forms of social ties, or to what extent can the notion of sociability operate' was one outcome of these efforts (see Ninomiya 1995). In this way the notion of sociability has become operative even in the fields of Japanese history. The same endeavors are being followed up for other Asian regions.

5. *Perspectives*

The notion of sociability remains ambiguous. But despite its ambiguity, or rather because of its vagueness, it has been able to play a heuristic role, just like the notion of mentalities (see Le Goff 1974). Thus, it has opened up a new perspective in social history. To be sure, there are some important gaps to be filled in actual research.

(a) In general, formal (organized) and informal (spontaneous) sociability are treated separately even now. The connection and the interference between these two aspects of sociability must be elucidated.

(b) There is a certain tendency to consider sociability in isolation. Originally, sociability was about openness, exchange, communication. It would be interesting to look at it from the point of view of network and strategy theory, and to exchange problems.

(c) The relations between sociability and the creation of public sphere should be reconsidered not only at the level of bourgeois sociability and its associations, but also at the level of popular sociability, without detaching the public sphere from everyday life.

See also: Civil Society, Concept and History of; Communication and Democracy; Friendship, Anthropology of; Social Capital; Voluntary Associations, Sociology of

Bibliography

Agulhon M 1966 *La sociabilité méridionale: Confréries et associations dans la vie collective en Provence orientale à la fin du XVIIIe siècle dans la deuxième moitié du XVIIIe siècle*, 2 Vols. La Pensée universitaire, Aix-en-Provence [2nd edn. under a modified title *Pénitents et Francs-maçons de l'ancienne Provence: Essai sur la sociabilité méridionale*. Fayard, Paris]

Agulhon M 1970 *La République au village. Les populations du Var de la Révolution à la IIe République*. Plon, Paris

Agulhon M 1976 La sociabilité, la sociologie et l'histoire. *L'Arc* **65** [Reprinted in *Le cercle dans la France bourgeoise 1810–1848: Étude d'une mutation de sociabilité*. Armand Colin, Paris, 1977]

Agulhon M 1979 Sociabilité populaire et sociabilité bourgeoise au XIXe siècle. In: Poujol G, Labourie R (eds.) *Les Cultures Populaires*. Privat, Toulouse

Agulhon M 1986 La sociabilité est-elle objet d'histoire? In: François E (ed.) *Sociabilité et société bourgeoise en France, en Allemagne et en Suisse (1750–1850)*. Éditions Recherche sur les Civilisations, Paris

ARS (Association de la Recherche sur la Sociabilité de l'Université de Rouen) 1987 *Sociabilité, pouvoirs et société, Actes due colloque de Rouen (1983)*. Rouen University, France

ARS 1989 *Aux sources de la puissance: Sociabilité et parenté, Actes du colloque de Rouen (1987)*. Rouen University, France

ARS 1992 *La sociabilité à table: Commensalité et convivialité à travers les âges, Actes du colloque de Rouen (1990)*. Rouen University, France

ARS 1997 *La rue, lieu de sociabilité? Rencontres de la rue, Actes du colloque de Rouen (1994)*. Rouen University, France

Boulton J 1987 *Neighbourhood and Society: A London Suburb in the Seventeenth Century*. Cambridge University Press, Cambridge, UK

Brennan T E 1988 *Public Drinking and Popular Culture in Eighteenth Century Paris*. Princeton University Press, Princeton, NJ

Cerutti S 1990 *La ville et les métiers. Naissance d'un langage corporatif (Turin, XVIIe–XVIIIe siècle)*. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris

Dann O (ed.) 1984 *Vereinswesen und bürgerliche Gesellschaft in Deutschland*. Oldenbourg, München

Farge A 1979 *Vivre dans la rue au XVIIIe siècle*. Gallimard, Paris
François E (ed.) 1986 *Sociabilité et société bourgeoise en France, an Allemagne et en Suisse (1750–1850)/Geselligkeit, Vereinswesen und bürgerliche Gesellschaft in Frankreich, Deutschland und der Schweiz (1750–1850)*. Editions Recherche sur les Civilisations, Paris
François E, Reichardt R 1987 Les formes de sociabilité en France, du milieu du XVIIIe siècle au milieu du XIXe siècle. *Revue d'histoire moderne et contemporaine* **34**: 3
Garrioch D 1986 *Neighbourhood and Community in Paris, 1740–1790*. Cambridge University Press, Cambridge, UK
Gemelli G, Malatesta M 1982 *Forme di Sociabilità nella Storiografia Francese Contemporanea*. Feltrinelli, Milano
Goodman D 1999 Sociabilité. In: Ferrone V, Roche D (eds.) *Le Monde des Lumières*. Fayard, Paris
Gordon D 1994 *Citizens without Sovereignty: Equality and Sociability in French Thought, 1670–1789*. Princeton University Press, Princeton, NJ
Guitton J-P 1979 *La sociabilité villageoise dans l'ancienne France: Solidarités et voisinages du XVIe au XVIIIe siècle*. Hachette, Paris
Gurvitch G 1950 *La vocation actuelle de la sociologie*. Presses Universitaires de France, Paris
Im Hof U 1982 *Das gesellige Jahrhundert: Gesellschaft und Gesellschaften im Zeitalter der Aufklärung*. Beck, München
Le Goff J 1974 Les mentalités: une histoire ambigué. In: Le Goff J, Nora P (eds.) *Faire de l'histoire, Vol. 3: Nouveaux objets*. Gallimard, Paris
Levi G 1985 *L'eredita immateriale. Carriera di un esorcista nel Piemonte del seicento*. Giulio Einaudi, Turin, Italy
Lüdtke A (ed.) 1989 *Alltagsgeschichte*. Campus Verlag, Frankfurt, Germany
Muchembled R 1989 *La violence au village. Sociabilité et comportements populaires en Artois du XVe au XIXe siècle*. Editions Brepols, Turnhout
Ninomiya H (ed.) 1995 *Musubiau Kataihi: Soshiabirite-ron no Shatei*. Yamakawa-shuppansha, Tokyo
Roche D 1978 *Le siècle des Lumières en province: Académies et Académiciens provinciaux, 1680–1789*, 2 Vols. Mouton, The Hague
Roche D 1993 *La France des Lumières*. Fayard, Paris
Siebert E 1993 *Der literarische Salon: Literatur und Geselligkeit zwischen Aufklärung und Vormärz*. Metzler, Stuttgart
Vovelle M 1982 Dix ans de sociabilité méridionale. In: Vouelle M (ed.) *Idéologies et mentalités*. François Maspero, Paris

H. Ninomiya

Social Behavior (Emergent), Computer Models of

By the term 'emergent social behavior' we mean those patterns of social behavior that arise automatically without relevant specifications being in-built in the individual. Under what conditions such patterns may emerge, and of what kind they are, is unpredictable in itself, but can be learned from 'bottom-up' models. In such models, a so-called 'process-oriented' approach is used. An artificial system is constructed in which agents are equipped with behavioral rules and mechanisms. While these agents interact with their environment and other agents, their behavior is studied in the same way as in ethological studies. Results show that in this way simple agents generate complex social structures, and that the same set of rules may lead to different patterns depending on the past experiences of the individuals, the demography of their population or the distribution of their food.

Studies of such complex systems were, of course, made possible by the development of computers, which have led to new modeling techniques. Initially, these studies dealt with emergent phenomena in chemical and physical research (Gleick 1987). In this article, the major issues and characteristics of these models (called 'individual-oriented,' 'individual-based,' or artificial life models) are explained by examples. Their connection with views on evolution and their general usage is discussed.

1. Major Issues of Self-Organized Social Behavior

Computer models in question deal with the emergence of:

(a) the formation of groups;
(b) collective 'decisions' (e.g., as regards the choice of a food source);
(c) the distribution of tasks over individuals (such as brood care and foraging);
(d) a reproductive 'clock';
(e) spatial structure of groups; and
(f) social organization of groups.

1.1 Group Formation

Many animal species live in groups which vary in characteristics, such as composition, size, permanency, and cohesion. One of the most cohesive and largest societies is found in army ants (with populations of up to 20,000,000 ants). Remarkably, these ants are practically blind and rely on a pheromonal system with which they mark their paths and by which they follow paths taken by others. Preying upon insect colonies, they raid over the forest in phalanxes of about 200,000 workers. Different species display different swarming patterns. To explain the differences among species in swarming, it is usually assumed that there are corresponding differences in the underlying behavioral motivations. Deneubourg and co-authors (1989), however, have shown that such markedly different swarming patterns may arise from one and the same system of laying and following pheromone

trails, when ants are raiding in different environments with different distributions of food. To show this, the authors used a simulation in which 'artificial ants' moved in a discrete network of points, and marked their path with pheromones. Further, when choosing between left and right, they preferred the more strongly marked direction. By introducing different distributions of food, different swarming patterns arose from the interaction between the flow of ants heading away from the nest to collect food, and the different spatial distributions of the foragers returning with the food. These different swarm types were remarkably similar to those of certain army ants.

1.2 Collective 'Decisions'

Further, the use of trail pheromones is also a mechanism of collective 'decision' making. Experiments with real ants combined with modeling have shown that trail laying often causes ants to choose the food source that is nearest to the nest (Deneubourg and Goss 1989). This observation seems to suggest that individual ants possess considerable cognitive capacities, but Deneubourg and Goss show in a computer model that ants may accomplish this remarkable feat without comparing distances to various food sources. When ants mark their path by pheromones and follow the more strongly marked branch at crossroads, then, by returning to the nest more quickly when the road to and from the food source is shorter, they obviously imprint the shorter path more heavily with pheromones. Thus, positive feedback is the result: as the shorter path attracts more ants, it also receives stronger marking, etc.

In subsequent extensions of this model (Detrain et al. 1999), path marking and following are shown to imply automatically risk avoidance (which agrees with observations in some species of ants), as follows: a hostile confrontation delays the forager for a certain time before returning to the nest. Thus the path to the dangerous food source automatically is marked more slowly and less strongly. Consequently, the safest food source is usually 'preferred.'

1.3 Distribution of Tasks

Such a division of labor occurs in colonies of many species of social insects. Remarkably, such division is flexible, i.e., the ratios of workers performing different tasks may vary according to the changing needs of the colony. The earliest model on the emergence of task division is that designed for bumblebees (*Bombus terrestris*) by Hogeweg and Hesper (1983, 1985). In an earlier experimental study, it was discovered that, during the growth of the colony, workers developed into two types, 'common' and 'elite' workers. The activities carried out by the two types differ remarkably: whereas the 'common' workers mainly forage and take rests, the 'elite' workers are more active, feed the brood, interact often with each other and with the queen, and sometimes lay eggs. To study the conditions for the formation of the two types of workers, Hogeweg and Hesper (1983, 1985) set up a so-called 'Mirror' model. This kind of model is complex in the number of variables included, but simple as regards behavioral processes. It contains biological data concerning the time of the development of eggs, larvae, pupae, etc. Space in the model is divided into two parts, peripheral (where inactive common workers doze part of the time) and central (where the brood is, and where all interactions take place). The rules of the artificial adult bumblebees operate 'locally' in that their behavior is triggered by what they meet. What they meet is picked out in the model randomly from what is available in the space in which the bumblebee finds itself. For instance, if an adult bumblebee meets a larva, it feeds it, and if it meets a pupa of the proper age, it starts building a cell in which a new egg can be laid, etc. All workers start with the same dominance value after hatching, with only the queen starting with a much higher rank. When one adult meets another, a dominance interaction takes place, the outcome of which (victory or defeat) is self-reinforcing. The positive feedback is 'damped,' however, because expected outcomes reinforce or diminish the dominance values of both partners only slightly, whereas unexpected outcomes give rise to relatively large changes in dominance value. Dominance values of the artificial bumblebees influence almost all their behavioral activities; for instance, individuals of high rank are less likely to forage.

It appears that this model generates two stable classes automatically, the 'elites' and 'commons,' with their typical dominance and behavioral conduct. For this result, the nest has to be divided into a center and a periphery (as found in real nests).

The flexibility of the distribution of tasks becomes obvious upon halving the worker force in the model. In line with observations in similar experiments on real bumblebees, this increases markedly each individual's tendency to forage. This arises in the model as follows: the lowered numbers of workers reduces the number of encounters among them, and thus increases the frequency of encounters between bumblebees and brood. Encountering the brood more often induces workers to collect food more frequently.

A different approach to the effect of social interactions on task allocation was taken by Pacala et al. (1996). In their model, individuals will stick to a certain task as long as it can be executed successfully, but otherwise give up (e.g., stop foraging when food is depleted). Then, as soon as they encounter another individual that is performing a task successfully, they will switch to perform that task too.

Another approach to the emergence of task division in social insects is described by Bonabeau et al. (1996). Their basic assumption is that different types of workers have different, genetically determined response thresholds for behavioral tasks. This is particularly relevant for dimorphic species, such as the ant *Pheidole megacephala*, in which two physically different castes, the so-called majors and minors, exist. Majors generally have higher thresholds than minors.

This fixed threshold model has been extended to variable response-thresholds by Theraulaz et al. (1998). They added learning in the form of a reinforcement process. This implies that performing tasks lowers the threshold of task execution and not-performing them heightens it.

1.4 Behavioral 'Clock'

In real bumblebees, the queen switches at the end of the season from producing sterile offspring to fertile ones, after which she is either chased away or killed in spite of the fact that she is still as dominant as before. The time of departure/death of the queen is important for the 'fitness' (i.e., number of new queens) of the colony.

After the demise of the queen, new queens and males (so-called generative offspring) are reared. It is assumed that the switch should take place some weeks before the end of the season, because at that time the colony is at its largest and can take care of the largest number of offspring. But it is very difficult to think of a possible external signal that could trigger such a switch. However, Hogeweg and Hesper (1983) show in their Bumblebee model that no external signal is needed, and that the switch may occur as a socially regulated 'clock' involving a fixed threshold, as follows. During the development of the colony, the queen is able to inhibit worker ovipositions by producing a certain pheromone that delays their production of eggs. Just before she is killed, she can no longer suppress the 'elite' workers, because, due to the large size of the colony, individual workers are less subjected to the dominance of the queen. So they start to lay unfertilized (drone) eggs. Further, the stress on the queen increases each time she has to perform dominance interactions with workers during her egg laying. When the 'stress' of the queen is at a low fixed threshold value, she switches to producing male eggs (drones), and when it increases to a high fixed threshold, she is killed or chased away.

The notion of this switch, which functions as a kind of social clock, has been challenged, because generative offspring are sometimes also found in small nests. However, in response to this objection, Hogeweg and Hesper have shown that the switch in small nests created in the Bumblebee model by reducing the growth speed, appears to occur at the same time as in large nests (1985). This is due to a complicated feedback process.

1.5 Spatial Structure

In many animal groups, there is a spatial structure in which dominants are in the center and subordinates at the periphery. This spatial structure is usually explained by the well-known 'selfish herd' theory of Hamilton (1971). The basic assumption of this theory is that individuals in the center of a group are best protected against predators, and therefore have evolved a preference for positions in which group members are between them and the predators, the so-called 'centripetal instinct.' However, such a spatial structure with dominants in the center also emerges in a model designed by Hemelrijk (2000), called DomWorld, in which such a preference for the center is lacking. The model consists of a homogeneous virtual world inhabited by agents that are provided with only two tendencies: (a) to group, and (b) to perform dominance interactions. Dominance interactions reflect competition for resources (such as food and mates) and the agent's capacity to be victorious (i.e., its dominance) depends, in accordance with ample empirical evidence, on the self-reinforcing effects of winning (and losing).

In this model, the spatial configuration arises without any 'centripetal instinct,' but is the result of the mutually reinforcing effects of spatial structure and strong hierarchical differentiation. It can be explained as follows. Pronounced rank differentiation causes low-ranking agents to be chased away by many, and therefore to end up at the periphery. Automatically, this leaves dominants in the center. Besides, this spatial structure causes agents to interact mainly with partners close in rank; therefore, if a rank change occurs, it is only a minor one. Thus, the spatial structure stabilizes the hierarchy and maintains its differentiation.

Spatial structure in human organizations is, for instance, evident in ghetto formation and in the formation of same-sex groups at parties. The Harvard economist, Thomas Schelling (1971), has shown how a stronger segregation emerges than is wanted by the individuals—who have only a slight preference for their own type—due to a positive feedback. Once individuals obtain minority status in a subgroup, because they are surrounded by an insufficient number of individuals of their own type (race or sex), they leave the subgroup, and this induces others of the same type to leave their former subgroup too. Their arrival in another subgroup may in its turn induce individuals of the other type to leave. Accordingly, Schelling notes that, even if the preference to have some individuals of one's own type close by is slight, its behavioral consequences may increase at a group level, leading to unexpected 'macropatterns.'

1.6 Social Organization of Groups

Two types of society are generally distinguished in the study of animal behavior: egalitarian and despotic ones. In certain primate species, such as macaques, a steep dominance or power hierarchy characterizes a despotic society, and a weak hierarchy an egalitarian one. Furthermore, both societies differ in a number of other characteristics, such as bidirectionality of aggression and cohesion of grouping. For each of these (and other unmentioned) differences a separate, adaptive (i.e., evolutionary) explanation is usually given, but DomWorld provides us with a simpler explanation for many of them simultaneously (Hemelrijk 1999): by simply increasing the value of one parameter, namely, intensity of aggression (in which egalitarian and despotic macaques differ—see Thierry 1990), the artificial society switches from a typical egalitarian to a despotic society with a cascade of consequences that resemble characteristics observed in the corresponding society of macaques. Such resemblances imply that differences between egalitarian and despotic societies of macaques may be entirely due to one trait only, intensity of aggression. On an evolutionary timescale, one may imagine that species-specific differences in aggression intensity arose, for example, because in some populations of a common ancestor, individuals suffered more food shortages, and consequently only those with intense aggression could survive.

Apart from variation among individuals in power, human societies are also characterized by a distribution of wealth. Epstein and Axtell (1996) show how, under many circumstances, a very skewed distribution of wealth emerges among heterogeneous agents (for instance, differing in vision and metabolism) in a model called 'Sugarscape.' Such a skewed distribution of wealth appears to be a general outcome of many specifications of agents that are extracting resources from a landscape of fixed capacity.

2. Characteristics of Models for Emergent Social Patterns

Typical of models for self-organized social behavior are the following characteristics of some or all of the above examples: (a) nearby perception and activation, (b) positive (instead of the more commonly used negative) feedback processes, (c) spatial distribution, (d) the possibility of studying complex behavior, and (e) 'open-endedness'.

(a) *Nearby perception and activation*. Individual agents have exclusive information about what is happening nearby. Further, timing of acts of agents is not ruled by a central clock, but nearby agents activate each other. Otherwise, agents are activated randomly.

(b) *Positive feedback processes*. Examples of these are the trail marking (with pheromones) and tracking system of ants, winning and losing of dominance interactions and a mutual reinforcement between spatial structure and the development of the hierarchy among group-living agents.

(c) *Spatial distribution*. The inclusion of space in the model creates the possibility for spatial structures to emerge. Such a spatial structure may affect behavior of individuals.

(d) *The possibility of studying complex behavior*. In these models, behavior is recorded in behavioral units similar to those used by ethologists for studying the behavior of real animals. Data are analyzed to detect new behavioral activities and patterns.

(e) '*Open-endedness*'—that is, at the time when the model is designed, it is unknown what patterns may emerge.

3. Emergence and Evolution

Because in these models behavioral mechanisms are implemented without continuous reference to pay-off benefits, this procedure leads to a parsimonious or restrictive view of evolution in the sense that the behavioral patterns observed need not always be based on trait-specific genes, and are not willed expressly by the individuals. Besides, many of the behavioral patterns usually attributed to long-term evolution may instead be explained as direct, short-term effects of other mechanisms. In these respects, these models agree with the social genetic and evolutionary explanations of, for instance, Lewontin, Gould, and Eldredge (Depew and Weber 1995), and they contrast with the thermodynamic and gene-focused views of Fisher and of sociobiology. In sociobiology, it is not the 'processes' that are of interest, but questions regarding adaptivity of behavior, and regarding evolutionarily stability of strategies. These issues are studied in the sociobiologically-oriented game-theoretic models. In contrast to models for emergence, behavioral strategies (such as 'hawks and doves,' 'cooperators and defectors') are defined in terms of pay-off without taking into account the underlying mechanisms or behavioral and environmental features; entities with different strategies are pitted against each other in order to examine which maintain themselves and which disappear, or in what proportions they may continue to exist in combination; new strategies that are not included from the start cannot arise in these game-theoretic models.

4. The Benefits of Models for Emergent Social Behavior

Models of self-organization are never to be regarded as a proof of how behavior arose. They are, however, useful for several reasons. First, they may suggest explanations that cannot be reached by theorizing alone. Take, for instance, Camazine's model on the structuring of the comb of honey bees (Camazine

1991). The comb of honey bees is structured with brood in the center, and pollen and honey in layers around it. This organization is usually attributed to a cognitive blueprint in the honey bees. Camazine's model shows, however, that the structuring of the comb may also result as a side-effect from the feeding activities of the brood, and from the relative quantities of pollen and honey imported or removed.

Second, these models can be used as 'search indicators' for empirical studies. If models are firmly based on observed biological mechanisms (as in many of the above mentioned examples), and if the emerging patterns agree with the patterns observed in real groups of organisms, then additional patterns observed in the model, but not yet noted in the real world, may be looked for in the real world. In the same way, in a model that has been shown to be relevant for real animal behavior, consequences of certain parameters may also be studied, e.g., the effects of different degrees of cohesion of grouping. To carry out such experiments in computer simulation is very easy. Usually, similar experiments are much more difficult in the real world.

Third, the insight that certain patterns of social behavior may be emergent often leads to new, parsimonious hypotheses about their evolution.

Fourth, such models can also be used in teaching pupils how to think in terms of self-organization in general as is done, for instance, by Resnick (1994).

Fifth, certain kinds of insight obtained from models may be used to solve real-world-problems (e.g., Bonabeau and Theraulaz 2000). For instance, the ant models by Deneubourg have been extended by Dorigo to solve the famous 'traveling salesman problem.' In this problem, a person must find the shortest route by which to pay a single visit to a number of cities. This problem is very difficult to solve, because the possibilities easily run into the millions.

See also: Artificial Social Agents; Emergent Properties; Evolutionary Social Psychology; Social Psychology; Social Psychology: Sociological; Social Psychology, Theories of; Social Simulation: Computational Approaches; Sociobiology: Overview; Sociobiology: Philosophical Aspects

Bibliography

Bonabeau E, Theraulaz G 2000 Swarm smarts. *Scientific American* **282**: 72–80

Bonabeau E, Theraulaz G, Deneubourg J-L 1996 Quantitative study of the fixed threshold model for the regulation of division of labour in insect societies. *Proceedings of the Royal Society London B: Biological Sciences* **263**: 1565–9

Camazine S 1991 Self-organizing pattern formation on the combs of honey bee colonies. *Behavioural Ecology and Sociobiology* **28**: 61–76

Deneubourg J L, Goss S 1989 Collective patterns and decision-making. *Ethology, Ecology and Evolution* **1**: 295–311

Deneubourg J L, Goss S, Franks N, Pasteels J M 1989 The blind leading the blind: modeling chemically mediated army ant raid patterns. *Journal of Insect Behavior* **2**: 719–25

Depew D J, Weber B H 1995 *Darwinism Evolving. Systems Dynamics and the Genealogy of Natural Selection*. MIT Press, Cambridge, MA

Detrain C, Deneubourg J-L, Pasteels J M 1999 Decision-making in foraging by social insects. In: Detrain C, Deneubourg J-L, Pasteels J M (eds.) *Information Processing in Social Insects*. Birkhäuser Verlag, Basle, Switzerland, pp. 331–54

Epstein J M, Axtell R 1996 *Growing Artificial Societies. Social Science from Bottom Up*. MIT Press, Cambridge, MA

Gleick J 1987 *Chaos: Making A New Science*. Viking, New York

Hamilton W D 1971 Geometry for the selfish herd. *Journal of Theoretical Biology* **31**: 295–311

Hemelrijk C K 1999 An individual-oriented model on the emergence of despotic and egalitarian societies. *Proceedings of the Royal Society London B: Biological Sciences* **266**: 361–9

Hemelrijk C K 2000 Towards the integration of social dominance and spatial structure. *Animal Behaviour* **59**: 1035–48

Hogeweg P, Hesper B 1983 The ontogeny of interaction structure in bumble bee colonies: a MIRROR model. *Behavioral Ecology and Sociobiology* **12**: 271–83

Hogeweg P, Hesper B 1985 Socioinformatic processes: MIRROR modelling methodology. *Journal of Theoretical Biology* **113**: 311–30

Pacala S W, Gordon D M, Godfray H J C 1996 Effects of social group size on information transfer and task allocation. *Evolutionary Ecology* **10**: 127–65

Resnick M 1994 *Turtles, Termites and Traffic Jams. Explorations in Massively Parallel Microworlds*. MIT Press, Cambridge, MA

Schelling T C 1971 Dynamic models of segregation. *Journal of Mathematical Sociology* **1**: 143–86

Theraulaz G, Bonabeau E, Deneubourg J-L 1998 Threshold reinforcement and the regulation of division of labour in social insects. *Proceedings of Royal Society London B: Biological Sciences* **265**: 327–33

Thierry B 1990 Feedback loop between kinship and dominance: The macaque model. *Journal of Theoretical Biology* **145**: 511–21

C. K. Hemelrijk

Social Capital

With roots in de Tocqueville, Durkheim, Weber, and others, current approaches to social capital were formulated most clearly by Coleman (1990), who distinguished the term from physical and human capital. Physical capital identifies investment in tools and other tangible productive equipment; human capital refers to less tangible investments in skill and knowledge by individuals (manifested specially by educational attainment). Social capital is even less tangible, and refers to relations among persons that facilitate action, embodied in the collective norms of

communities that extend beyond immediate family members and the trustworthiness of the social environment on which obligations and expectations depend.

Most ensuing discussions of social capital emphasize levels of trust and patterns of social networks within communities. Putnam takes social capital to reflect 'features of social organization, such as trust, norms, and networks, that can improve the efficiency of society by facilitating coordinated actions' (1993). For Fukuyama, social capital is 'a capability that arises from the prevalence of trust in a society or certain parts of it' (1995). More complete surveys of the term are available in Jackman and Miller (1998) and Portes (1998).

While it is an intangible that affects the behavior of individuals, social capital itself is cast as a collective or group-level phenomenon that identifies the informal social contexts that constrain and influence the behavior of individuals. Because it is relational and inheres in social networks, social capital minimally involves dyads and typically groups of larger size: communities, neighborhoods, ethnic and national groups. In high social capital environments, individuals are networked into a variety of larger informal groups that emphasize norms of trust and reciprocity from which those individuals benefit.

While the concept is intangible, the underlying variable is clearly consequential. There is broad agreement that effective social networks and pervasive interpersonal trust enhance efficiency in social relations and lower transaction costs, both of which have direct implications for social, economic, and political behavior. Some have further suggested that high levels of social capital foster cooperative and perhaps altruistic behavior, a theme with communitarian origins. The tendency to couple social capital and cooperative behavior is specially pronounced among scholars uncomfortable with the premise that human behavior is grounded in the egoistic optimization of individual utility. However, the line between cooperation and collusion is fine.

Much has been written about social capital recently. In a noted analysis, Putnam laments the political consequences of a decline of social capital in the US during the 1980s and 1990s, a trend he claims has led increasingly to a society of go-it-alone bowlers, isolated both from each other and from broader currents of community political life (Putnam 1995). Yet the evidence for any such pattern is mixed. Voter turnout in the US has declined since around 1960, and there has also been a decline in reported levels of political trust on some survey responses (Levi and Stoker 2000). These trends appear to be the exception rather than the rule, however, and their meaning is unclear. A broad review of the available evidence since the 1960s finds no reduction in social capital, indexed by general levels of trust or group membership (Ladd 1999). Shrinking membership rates in some organizations (e.g., labor unions) have been more than offset by increases in others (e.g., environmental groups). Further, there is no evidence of a decline in social activities associated with group membership. Associational membership in the US thus appears live and well. Data on trends in group membership in other countries are less complete, but do not suggest a decline.

A more fundamental issue hinges on the causal significance attached to social capital, and here there is considerable disagreement. The bowling alone thesis attracted attention because of the consequences it imputes to social capital. Reduced social capital becomes noteworthy because it heralds increasing problems of governability and other social pathologies. The idea that social capital affects other outcomes is, indeed, a recurring theme in recent studies. For example, Putnam (1993) concludes that the effectiveness of Italy's regional governments stems from durable regional differences in patterns of social networks (civic associations): his statistical analyses explain current performance as a function of social networks in place a century earlier, and at various points, regional differences are traced back almost 1,000 years. Similarly, Fukuyama (1995) argues that the key to economic growth is social capital in the form of a supporting culture of trust, or 'spontaneous sociability.' This theme is echoed in Harrison's (1992) conclusion that political and economic developments hinge critically on values involving trust, ethical codes, and orientations to work and risk-taking. Inglehart (1997) offers a minor variation by suggesting that membership in voluntary associations and levels of trust explain stable democracy and patterns of democratization in recent years. In such ways, social capital has been linked to a variety of political and social outcomes.

These claims are more than a simple statement that social capital is important. They instead imply that social capital reflects enduring cultural norms passed across generations by socialization, norms that cannot be explained in terms of reasonable responses to current circumstances. Further, these cultural norms are durable: allowing for slight short-term forces, they reflect the long haul. Finally, these norms are taken to be the key exogenous factors in generating economic and governmental performance. Social capital is, in this sense, more crucial than objective conditions embodied in institutions, and endures in the face of institutional change. Social capital is critical in a second sense. Different levels of social capital lead people to process the same information differently, even in face of the same institutional constraints and incentives.

Such claims parallel earlier arguments linking political culture to development and governmental performance. For example, Banfield (1958) argued that the poverty of a small southern Italian town stemmed from a dominant ethos of 'amoral familism.' Residents displayed no concern for collective issues facing the

community, operating instead according to the axiom, 'Maximize the material, short-run advantage of the nuclear family: assume that all others will do likewise.' In current terms, social and political pathologies in the town were attributed to low levels of social capital. Similarly, Almond and Verba (1963) claim that variations in democratic performance reflect national differences in civic orientations as a function of social capital.

An alternative approach identifies conditions that enhance levels of mutual trust. This is the strategy suggested by Coleman (1990) and others, one that parallels recent institutional approaches to politics that specify how institutions constrain behavior through the incentives they embody. Social capital (reflected in trust and group membership) then becomes a function of the political and economic performance induced by effective institutional arrangements, not the other way round. Consider the following examples.

The Mafia is often seen as a product of distinctive enduring Sicilian subcultural norms in which trust is at a premium. Since the norms are seen as both exogenous and durable, the Mafia should persist indefinitely. In contrast, Gambetta (1993) takes the Sicilian Mafia to be a loose cartel of families that produces and sells private protection. Gambetta treats this product as a surrogate for the enforcement of property rights by the state, so that protection is an imperfect substitute for mutual trust, and demand for protection is highest when the state does not enforce contracts. While state enforcement of property rights constitutes a public good, however, Gambetta shows that the distinctive feature of the Mafia's product is that the protection offered is private. Traders can distinguish insiders (other traders with Mafia protection) from outsiders (traders without it). The incentives for both insiders and the Mafia favor cheating outsiders, thereby increasing distrust. Distrust is thus endogenized, a product of the Mafia, not its source.

Recasting the argument this way makes a difference on at least two counts. First, by identifying the sources of low social capital in the form of distrust, it offers an analysis of ways in which levels of trust might be enhanced. In this case, one focuses on ways in which the state might enforce contracts even-handedly, as opposed to centering attention on a deep-seated subcultural ethos. Second, it offers a general approach that can be applied in different settings. For example, the rapid increase in Mafia activities in Russia after the collapse of the Soviet Union also appears to have origins in the state's inability or unwillingness to enforce contracts (Varese 1994). Clearly it involves a stretch to attribute such changes to entrenched and distinctive subcultural values, whether from Sicily or elsewhere.

Ethnically divided societies constitute another significant phenomenon, since ethnic markers have so often been taken to reflect the boundaries of 'primordial' groups (the quintessential form of exogenously-defined cultural groups) that pose particular threats to states. However, despite numerous ethnically divided societies, ethnic conflict itself is relatively unusual, which implies that such conflict is not an automatic fuction of the ethnic differences themselves. A principal source of ethnic conflict in divided societies is fear of victimization in the form of aggression from rivals (Weingast 1998). When a group fears victimization and can employ the state against another, it has an incentive to mount a preemptive strike against the other. Since being an aggressor is preferable to being a victim, and because promises not to aggress by individual groups are insufficient to deter conflict, ethnic conflict is likely.

It is, thus, crucial to construct political institutions that raise the costs for any group to use the state this way. Effective institutions modify incentives to make toleration self-enforcing, credibly committing each group to honor its promises and thereby minimizing the odds of victimization for any group. Such institutions provide incentives for groups and their leaders to conclude that neither they nor their rivals have an interest in striking first. They thereby help induce trust and co-operation. The advantage of this perspective is that it allows variations in conflict across divided societies to be specified. It also focuses more directly on the temporal ebbs and flows of ethnic conflict that occur with institutional collapse and rebuilding within divided societies.

Levels of trust and patterns of group membership within communities, however intangible, are clearly of intrinsic social and political consequence, and recent work on social capital has restored these issues to center stage. Fully understanding the sources and distribution of social capital is an important issue on the current research agenda.

See also: Altruism and Self-interest; Coleman, James Samuel (1926–95); Human Capital: Educational Aspects; Mafia; Network Analysis; Networks and Linkages: Cultural Aspects; Networks: Social; Political Culture; Trust: Cultural Concerns; Trust: Philosophical Aspects; Trust, Sociology of

Bibliography

Almond G A, Verba S 1963 *The Civic Culture; Political Attitudes and Democracy in Five Nations*. Princeton University Press, Princeton, NJ

Banfield E C 1958 *The Moral Basis of a Backward Society*. Free Press, New York

Coleman J S 1990 *Foundations of Social Theory*. Harvard University Press, Cambridge, MA

Fukuyama F 1995 *Trust: The Social Virtues and the Creation of Prosperity*. Free Press, New York

Gambetta D 1993 *The Sicilian Mafia; The Business of Private Protection*. Harvard University Press, Cambridge, MA

Harrison L E 1992 *Who Prospers? The Business of Private Protection* Basic Books, New York
Inglehart R 1997 *Modernization and Postmodernization: Cultural, Economic and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ
Jackman R W, Miller R A 1998 Social capital and politics. *Annual Review of Political Science* **1**: 47–73
Ladd E C 1999 *The Ladd Report*. Free Press, New York
Levi M, Stoker L 2000 Political trust and trustworthiness. *Annual Review of Political Science* **3**: 475–507
Portes A 1998 Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology* **24**: 1–24
Putnam R D 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ
Putnam R D 1995 Bowling alone: America's declining social capital. *Journal of Democracy* **6**: 65–78
Varese F 1994 Is Sicily the future of Russia? Private protection and the rise of the Russian Mafia. *European Journal of Sociology* **35**: 224–58
Weingast B R 1998 Constructing trust: The political and economic roots of ethnic and regional conflict. In: Soltan K, Uslaner E M, Haufler V (eds.) *Institutions and Social Order*. University of Michigan Press, Ann Arbor, MI

R. W. Jackman

Social Categorization, Psychology of

Humans experience the world through a filter of categorization. They distinguish among natural kind categories such as plants, animals, and minerals, and among artifact categories such as tools, machines, and works of art. Still other categories are abstract (e.g., peace, love, happiness) or *ad hoc* (e.g., things I need to pack for a trip). Social categorization is mainly concerned with the way people categorize themselves and others. Like any other kind of categorization, social categorization raises two fundamental questions. How do categories come to be coherent, and how do they affect human perception and action (Sternberg and Ben-Zeev 2001)

Before addressing these questions, one may wonder what would happen if there were no social categories. People would either be lumped together or considered individually. Though intuitively appealing, both options are impractical. Neither an all-inclusive human category nor all-exclusive person categories allow precise inferences about individuals. Such inferences are important because of their adaptive functions. People need to be able to predict—beyond chance levels—what others are like and what they are likely to do. Social categorization facilitates and constrains two kinds of inductive inferences. Downward induction occurs when the features of a category are attributed to individual members; upward induction occurs when the features of individuals are generalized to the category at large. Both inferences become more precise as categories become smaller and more homogeneous (Holland et al. 1986). Knowing that Jacques is French, for example, suggests more precise attributions than merely knowing that he is European. These attributions, in turn, may change when even more exclusive categories (e.g., Parisian) are used (Nelson and Miller 1995). Similarly, the discovery that Jacques is a connoisseur of fine wine is more readily generalized to the French than to Europeans, and it is even more readily generalized to other people from Jacques' region.

Precision and economy are two inversely related aspects of the quality of inductive inferences. The use of exclusive and homogeneous social categories benefits precision, but it is not economical. Any given inference, be it attribution or generalization, can reach only few individuals. Conversely, large (and heterogeneous) categories permit attributions and generalizations to many members, but these inferences lack precision. Like object categories, social categories form hierarchies in which large heterogeneous categories comprise smaller and more homogeneous ones (Sternberg and Ben-Zeev 2001). In a hierarchy of religious categories, for example, the distinction between Christians and Muslims lies at a highly inclusive level. Further distinctions between Catholics and Protestants, and between Sunnis and Shiites, lie at more exclusive levels.

Social categorization is extraordinarily flexible because any particular category falls within a web of overlapping and conflicting hierarchies (Wattenmaker 1995). When Jacques is considered Parisian, French, and Catholic, three hierarchies intersect. Most Parisians are French, and most French are Catholic, but no category is contained fully within another. Parisians can also be classed with other city-dwellers (e.g., Romans) as urbanites, and the French can also be classed with other nations (e.g., the Italians) as Europeans. The availability of alternative groupings permits strategic choices among equally valid depictions of the social world. Whether the choice of a social category is apt depends heavily on the context in which it is made. Natural kind categories are more resistant to changes in context. Most of the time, a furry companion is simply a dog.

1. *Coherence*

Categories are coherent if induction effectively differentiates among them. Downward induction is effective if the features of a category can be attributed to its own members at a higher rate than to the members of other categories. Upward induction is effective if the features observed in an individual can be generalized to other members of this category at a higher rate than they can be generalized to members of other categories.

Inspired by Aristotle's beliefs about the structure of nature, the classic theory postulated the existence of features that are individually necessary and jointly sufficient for categorization (Sternberg and Ben-Zeev 2001). When there are defining features, inferences are certain and deductive instead of probabilistic and inductive. Because, for example, all geometric shapes with three sides and angles summing to 180 degrees are known to be triangles, no generalization needs to occur. Incidental features, such as the color of the shapes, have little inductive interest. Most social categories lack defining features. To illustrate, Allport (1954) noted that the Jews 'cannot be classified as racial, ethnic, national, religious, or as any other single sociological type.' The lack of classic definitions in the social domain creates a psychological challenge. How can we perceive and judge people by categories that we are unable to define?

The idea of essences protects the faith in classic definitions of social categories. Essences are invisible—or at least so far undetected—defining features. No matter how much a person's surface features may change, the philosophy of essences assumes the existence of an immutable underlying nature (Rothbart and Taylor 1992). Essentialist beliefs thus obscure the social origin of many categories. Even during the nineteenth century, some scientists believed that differences observed between human races were actually differences between species (Gould 1996). Since essences, by definition, are postulated rather than discovered, they tend to follow rather than precede social categorization. Once a set of social categories—be it by cultural convention or political expediency—is established, people tend to accept it as natural. They tend to believe that there must be defining features even if they are unable to name them.

Wittgenstein (1958), who was more interested in language than in nature, refuted the classic theory when he asked what games are. Although they have little in common, chess and bingo can be grouped together because other games share features with either one or the other (e.g., poker requires both skill and luck). Family resemblance makes categories cohere without requiring that a single feature is shared by all members (Rosch and Mervis 1975). The term family resemblance aptly describes what many vintage portraits of extended families reveal. The patriarch seated in the center is obviously related to everyone, but the cousins on either side have no apparent similarities.

Categorization by family resemblance is effective because it captures the lumpiness of the social world. Family resemblance categories comprise members of varying typicality, they have unstable and fuzzy boundaries, and their features tend to be correlated with one another (Rosch and Mervis 1975). Consider the social category of 'actors.' Typical members, or prototypes (e.g., Humphrey Bogart), have more features in common with members of the same category than with members of other categories. Prototypes come to mind faster and are recognized more easily than less typical members. Atypical members (e.g., Ronald Reagan) are not categorized effectively. Minor changes in their feature sets, or even changes in the perceiver's attention to some features, can lead to categorization elsewhere. Crossing a fuzzy boundary, Ronald Reagan went from being a moderately typical actor to being a prototypical politician. His example illustrates how some features (e.g., the ability to portray sincerity) can contribute to separate family resemblance categories. Individual features tend to be correlated across categories. Feathered animals also tend to fly, and people who prefer Beaujolais over beer tend say *santé* instead of *prosit*. Correlations among features and across categories provide useful, though partly redundant, information. They increase the perceiver's confidence that categorization has been correct and they allow inductive inferences from one feature to another.

2. *Consequences*

Unless categorization is random (see discussion of 'thin categories' below), most individuals are more similar to other members of the same group than to members of other groups. Both attribution and generalization inferences benefit from this basic similarity relation. If there is no specific information about a group member, any trait can be attributed to him or her with a probability corresponding to the perceived prevalence of the trait in the group. To make such trait attributions is to stereotype. Stereotyping may be questioned on ethical grounds—especially when the expectations are incorrect (Banaji and Bhaskar 2000)—but as an inductive inference, stereotyping is necessary if beliefs about categories and their members are to be coherent (Krueger and Rothbart 1988). Conversely, traits observed in individuals may be generalized to the group. If there are many observations, the prevalence of a feature in a sample is the best estimate for its prevalence in the group at large. But is it possible to justify a generalization to the category if only a single category member has been observed?

Mill (1843/1974) noted that people sometimes generalize from single instances, whereas at other times they do not. He realized that this fundamental problem of induction can only be understood by taking into account *a priori* expectations about the homogeneity of the group. In the simple case of dichotomous features (i.e., a person either does or does not possess a given trait), the group becomes more homogeneous as the prevalence of the trait approaches either zero percent or 100 percent. If traits are thought to be either present in all group members or in none, a single observation of a particular trait can

be generalized to everyone. If the group is thought to be more heterogeneous, generalizations are 'regressive,' which means they shift further towards the 50 percent mark.

Suppose half of the traits are presumed common (with $p = 0.8$) and half are presumed rare ($p = 0.2$). Before there is relevant information about an individual, any particular trait may be expected with the average of these probabilities (i.e., $p = 0.5$). When the person turns out to be friendly, the trait of friendliness is more likely among the common ones than among the rare ones. The probability that friendliness is a common trait (0.8) may then be multiplied with the probability of encountering a friendly person if friendliness is indeed common (also 0.8). An analogous calculation is done for the possibility that the trait is rare (0.2×0.2), and the products are summed and multiplied by 100. The result is an estimate for the prevalence of friendliness (68 percent), which, although it does not match either of the assumed base rates, minimizes the expected error in the long run (Dawes 1990). Generalizations allow the formation and the change of stereotypes. Because generalizations, unlike attributions, are regressive, stereotypes are harder to change than to apply.

Categories are coherent inasmuch as similarities within categories are greater than similarities between categories. This simple rule suggests that similarity relationships are there to be detected, and that social categorization is free from errors if this detection task is mastered. Somewhat surprisingly, however, no theory of similarity can fully explain why people use certain categories and not others. Before similarity can be estimated, a finite number of features must be selected from the infinite pool of possible features (Sternberg and Ben-Zeev 2001). These selections partly depend on value judgments or other (social) intuitions that may have little to do with similarity itself. In other words, categorization depends heavily on the perceiver's theories of how to construe the world and what features to ignore. Some social categories are so 'thin' that they are held together by little more than the label itself. Thin categories are superimposed on social reality rather than extracted from it (Krueger and Clement 1994). Labels create quasi-Aristotelian categories because they do double duty as defining features. Labels create patterns of inclusion and exclusion that might as well be different. The US Census 2000, for example, distinguishes between Hispanics and non-Hispanic Blacks, not between Blacks and non-Black Hispanics. Despite meager underlying similarity relations, these categories organize perception and thought, and thus provide a platform for prejudice, stereotyping, and intergroup relations.

The perceiver's perspective on social categories depends critically on their own sense of identity (Oakes et al. 1994). Some categories include the self, whereas others exclude it. The egocentric distinction between ingroups and outgroups, which is peculiar to social categorization, has two major consequences: ingroup favoritism and perceptions of outgroup homogeneity. Both phenomena are related reciprocally to the thinness (or arbitrariness) of categorization and to inductive reasoning.

3. *Ingroup Favoritism*

Typically, a group is judged more favorably by people who are in it than by people who are not. In real social categories, ingroup favoritism has many causes, such as competition for scarce resources, a history of ideological or symbolic conflict, or unrepresentative interactions among group members. Therefore, it is important to note that the same bias emerges in the so-called minimal group situation. Minimal groups are very thin categories created in research laboratories. Group membership is completely arbitrary; often it depends merely on the luck of the draw in a lottery. There is no contact among members, no information about the attributes of others, and no dependence on their actions. Still, people discriminate (Tajfel et al. 1971). Compared with outgroup members, ingroup members give away more money or symbolic rewards, they are more interested in making friends, and they expect to see and are prepared to remember more positive personality traits.

These differences point to the crucial role of self. When people find themselves in an unfamiliar group, self-knowledge is the only information available to support judgments about others, and social categorization provides the only framework for the generalization of this knowledge. Indeed, people 'project' their own features to others with whom they know they have something salient in common, namely group membership (Krueger and Clement 1994). Because most people have favorable self-images, generalization creates a favorable description of the group. Outgroup members do not generalize, believing that their own features do not represent the group. This lack of induction results in a more neutral description of the group. Whose judgments are mistaken? The interplay of induction and categorization suggests that the judgments made by outgroup members are biased. When a population is arbitrarily broken up into different categories, these categories inherit the same features. Therefore, any individual is similar to both ingroup and outgroups. Ingroup favoritism primarily reflects a failure to generalize one's own favorable features to outgroups, rather than any excessive tendency to ascribe desirable features to the ingroup.

Although egocentric and asymmetric induction patterns go a long way to explain ingroup favoritism, other motives also come into play (Abrams and Hogg 1999). Being able to discriminate against an outgroup

can strengthen a sense of social identity and heighten self-esteem while reducing feelings of uncertainty. These motives can be so strong that people forfeit possible gains for the ingroup as long as the ingroup is better off than the outgroup.

4. Perceptions of Outgroup Homogeneity

The mere act of categorization suggests that the categorized instances can be treated as equivalents. In social categorization, the assimilation of instances (people) to the group average is asymmetric. People who do not belong to the group usually perceive it to be even less variable than group members do (Park and Rothbart 1982). Italians, for example, may appear to be more homogeneous to Americans than to Italians themselves. This homogeneity bias is expressed in several ways. Individual group members are easily confused with one another; individual traits appear to be either extremely common or extremely rare; and judgments about individual group members on a given trait dimension do not depart much from the average of these judgments.

The homogeneity bias is not a mere by-product of ingroup favoritism because it affects both positive and negative attributes, and because it is absent in minimal groups where ingroup favoritism is common. The psychological need to differentiate ingroups from outgroups—which contributes to ingroup favoritism—would lead people to perceive the ingroup as homogeneous. Moreover, the homogeneity bias does not simply reflect differences in induction. Although outgroup members may have less information than ingroup members do, differences in familiarity do not necessarily produce a bias. Statistically, sample size is unrelated to sample variance (unless the samples are very small). Finally, the selective induction from the self to ingroups works against the outgroup homogeneity bias. Induction reduces perceptions of variability and thus uncertainty. Not using self-related information to make inferences to outgroups thus supports perceptions of heterogeneity.

The main source of perceived outgroup homogeneity are processes of categorization themselves. People who belong to a group are more likely to perceive differentiated subgroups (Park and Rothbart 1982). Italians more than Americans, for example, further categorize Italians by geographic origin, distinguishing between Tuscans, Romans, Calabrians, and so forth. To ingroup members, the group is thereby located at a higher level in the hierarchy than it is to outgroup members. The homogeneity bias then results from the assumption that categories not divided any further are more homogeneous than divisible categories. The error is the failure to realize that members of other groups consider their own categories to be as divisible as we consider our own.

5. Conclusion

Frameworks of social categorization may change across time and place, but they are functional in that they enable people to reason inductively and to situate themselves within the social world. Making predictions about the self and others is potentially more adaptive than withholding judgment altogether (Krueger 1998). In the process, however, social categorization also enables the persistence of prejudicial beliefs and discriminatory behavior. The tension between the adaptive and the biasing effects of categorization highlights the importance of this topic.

See also: Stereotypes, Social Psychology of

Bibliography

Abrams D, Hogg M A 1999 *Social Identity and Social Cognition*. Blackwell, Oxford

Allport G W 1954 *On the Nature of Prejudice*. Addison-Wesley, Reading, MA

Banaji M R, Bhaskar R 2000 Implicit stereotypes and memory: the bounded rationality of social beliefs. In: Schacter D L, Scarry E (eds.) *Memory, Brain and Belief*. Cambridge University Press, Cambridge, UK, pp. 139–75

Dawes R M 1990 The potential nonfalsity of the false consensus effect. In: Hogarth R M (ed.) *Insights in Decision Making: A Tribute to Hillel J. Einhorn*. University of Chicago Press, Chicago, pp. 179–99

Gould S J 1996 *The Mismeasure of Man*. Norton, New York

Holland J H, Holyoak K J, Nisbett R E, Thagard P R 1986 *Induction: Processes of Inference, Learning, and Discovery*. MIT Press, Cambridge, MA

Krueger J 1998 On the perception of social consensus. *Advances in Experimental Social Psychology* **30**: 163–240

Krueger J, Clement R W 1994 Memory-based judgments about multiple categories: A revision and extension of Tajfel's accentuation theory. *Journal of Personality and Social Psychology* **67**: 35–47

Krueger J, Clement R W 1996 Inferring category characteristics from sample characteristics: Inductive reasoning and social projection. *Journal of Experimental Psychology: General* **128**: 52–68

Krueger J, Rothbart M 1988 Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology* **55**: 187–95

Mill J S 1974/1843 *A System of Logic Ratiocinative and Inductive*. University of Toronto Press, Toronto, ON

Nelson L J, Miller D T 1995 The distinctiveness effect in social categorization. *Psychological Science* **6**: 246–9

Oakes P J, Haslam S A, Turner J C 1994 *Stereotyping and Social Reality*. Blackwell, Oxford, UK

Park B, Rothbart M 1982 Perceptions of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes on in-group and out-group members. *Journal of Personality and Social Psychology* **42**: 1051–68

Rosch E, Mervis C B 1975 Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* **7**: 573–605

Rothbart M, Taylor M 1992 Category labels and social reality: Do we view social categories as natural kings? In: Semin G R, Fiedler K (eds.) *Language, Interaction and Social Cognition*. Sage, London, pp. 11–36

Sternberg R J, Ben-Zeev T 2001 *Complex Cognition: The Psychology of Human Thought*. Oxford University Press, New York

Tajfel H, Flament C, Billig M, Bundy R 1971 Social categorization and intergroup behaviour. *European Journal of Social Psychology* **1**: 149–78

Wattenmaker W D 1995 Knowledge structures and linear separability: Integrating information in object and social categorization. *Cognitive Psychology* **28**: 274–328

Wittgenstein L 1958 *Philosophical Investigations*, 2nd edn. Blackwell, Oxford

J. Krueger

Social Change: Types

Two sets of assumptions must be made explicit to identify different types of social change. The actor's attributes are needed to understand what they do and why. The social structure—options and payoffs—within which the actors choose and act are needed to explain what happens and why. The two sets of assumptions are needed to construct a logical narrative of systemic effects.

Changes can consist in either

(a) *the actors changing*: what they want, know, do, obtain, or are burdened with, i.e., actor change in cognition, attitudes, sentiments, attributes, actions; or

(b) *the structure changing*: i.e., change of states, their correlatives or distributional aspects. These face the actors as facts, options, or structured costs and benefits, such as property rights, vacancies, socially available marriage partners, health hazards, or language boundaries.

The structure may change the actors, their hearts and minds. The actors may change the structure, its positions and possibilities. And the actors may react to conditions they have brought forth—by an echo effect, so to speak, from an edifice of their own making.

From answering two questions: (a) do the actors change, and (b) does the structure change, we may identify four kinds of models of social change, as shown in Fig. 1. Examples from classical texts below illustrate these types.

		Does the structure change?	
		No	**Yes**
Do the actors change?	**No**	Aggregate effects	Structural effects
	Yes	Actor effects	Dialectical effects

Figure 1
Types of models of social change

1. *Models of Aggregate Effects*

An example of a model of 'aggregate effects' is found in microeconomic theory. Under 'pure barter,' each actor is assumed to prefer certain combinations of goods, to know who commands what goods, etc. Among the structural assumptions are that actors have rights to goods, that goods can be exchanged freely without transaction costs, etc. From such assumptions it follows that the actors will barter goods they like less for goods they want more. The driving force is the initial gap between what they have and what they want—between what they actually prefer and what they initially control. Hence, a new distribution of goods will arise among the actors from which nobody will want to move because none can get more satisfaction from another exchange. The actors themselves remain unchanged, as does the value of their goods. The model is one of change only in the sense than that the actors' bundles of goods at the end are different from the outset. Hence economists call this genre 'models of static equilibrium,' though a process runs until this equilibrium is reached. Distributional characteristics may be modified—e.g., the degree of concentration of goods. But they have no further motivational effects in these models—the actors' choices shape the final allotment of goods, but they do not react to it once equilibrium is reached. The distribution of final control is just an aggregate outcome, not input for further action. Economics has many variations on this theme. It is a separate, empirical question to what extent such a model agrees with a barter process in the real world.

There are even more stripped down models of aggregate effects. A stable population model assumes nothing about what the actors want or know—only that they have constant probabilities of offspring and death at specific ages. Initially the population pyramid may deviate from its final shape. Having arrived there, it will remain unchanged. The stable proportions in the different age/sex categories are equilibrium states which represent the 'latent structural propensities of the process characterized by the [parameters]' (Ryder 1964, p. 456)—the values that proportions move toward when parameter values remain constant. For some, even important, phenomena, cut down actor models (no intentions or knowledge) are powerful. Simple assumptions can take you a long way in making a logical narrative.

2. *Models of Actor Effects*

The second type of model is illustrated by 'learning by doing.' In *Capital*, Marx describes how a specialist (a 'detail laborer') acquires skills within manufacture

production. A 'collective laborer' is the organized combination of such specialists. Manufacture ripens the virtuosity of the 'detail laborer' in performing his specialized function by repetition:

> A labourer who all his life performs one and the same simple operation, converts his whole body into the automatic, specialized implement of that operation. Consequently, he takes less time doing it, than the artificer who performs a whole series of operations in succession. But the collective labourer, who constitutes the living mechanism of manufacture, is made up solely of such specialized detail labourers. Hence in comparison with the independent handicraft, more is produced in a given time, or the productive power of labour is increased. Moreover, when once this fractional work is established as the exclusive function of one person, the methods he employs become perfected. The workman's continued repetition of the same simple act, and the concentration of his attention on it, teach him by experience how to attain the desired effect with the minimum of exertion. (Marx [1864] 1967, p. 339)

The first logical element of this argument is actor assumptions:

(a) Actors are purposive—i.e., they seek desired effects. Indeed, Marx was no strict determinist, but assumed purposive action explicitly:

> A spider conducts operations that resemble those of a weaver, and a bee puts to shame many an architect in the construction of her cells. But what distinguishes the worst architect from the best of bees is this, that the architect raises his structure in imagination before he erects it in reality. And at the end of every labor-process, we get a result that already existed in the imagination of the labourer at its commencement. He not only effects a change of form in the material on which he works, but he also realizes a purpose of his own. (Marx [1864] 1967, p. 178)

(b) The actors are satisficing, i.e., they start pursuing their goals by procedures that may not be the best (which provides room for improvement).

(c) They recognize deviations from present satisfactory performance.

(d) They evaluate deviations as either positive or negative by a criterion such as the 'principle of least effort' (to attain the desired effect with the minimum of exertion).

(e) There are two sources of deviation from current routines. The actors may search for better solutions, i.e., purposively probing for improvements by 'trial and error.' Or actors make mistakes—a stochastic element producing random departures from current routines. Accidentally hitting better solutions could be called 'spot and snatch.' Hence the model has two sources of new alternatives—deliberate exploration and random variation—one sought *ex ante*, the other sighted *ex post*.

In short, behavior is result controlled: when the outcome of an action deviates from the intended result, actors may modify their conduct. Negative deviations in terms of an actor's purposes will be avoided and result in a return to the standard routine. Positive deviations will be retained.

The second logical element is the structure assumptions of Marx' selection model:

(a) The actors are set to repetitive tasks.

(b) There are alternative ways of carrying out the tasks.

(c) The choice of alternative has consequences for the actors, being rewarding or penalizing in terms of their goals. These consequences may be due to laws of nature (e.g., efforts required or difficulties encountered) or they can be socially organized (e.g., reducing costs or increasing payoffs).

Such are the assumptions of Marx' argument. He constructed a selection model which at the individual level works like this: the probability of mistakes is cut with every attempt, while methods giving better results are retained. Initially satisfactory but unwieldy procedures are abandoned—more effective ones are kept when captured, e.g., by the principle of least effort. Hence the laborer becomes a special tool for the task at hand—more adept and adroit. The best form found solidifies into a routine. And the collective worker as the sum of detail laborers can reach a perfection beyond any single individual carrying out all operations alone. Manufacture begets handicraft industries that push the skills required for carrying out a specific task in the same direction. The progression toward a best response pattern hence leads to homogenization of actions.

In short, on the assumptions of purposive action and selection of beneficial deviations, Marx explains individual adaptation as well as uniforming selection—i.e., increased dexterity in repetitive tasks until the form best fitted to the task is found. Note that in this model what changes is not actors' preferences but their skills—not what they want, but what they know. Their bias for less effort is stable.

There are other models where actors change without this being part of their intention or recognition—effects may occur behind their backs. For example, a contagious disease may immunize a group against further outbreaks even if neither the contagion nor its effects are understood. If insights are won, actors may change practice, as when humans started to infect themselves with cowpox as protection against smallpox.

3. *Models of Structural Effects*

The brilliant Norwegian sociologist Eilert Sundt (1817–75) provides an example of structural change inspired by the logic of mutations and selection found in *The Origin of Species*. 'Learning by doing' can modify not just behavior, but also material forms. Sundt had a keen eye for the logic of both error-based 'spot and snatch' and the more deliberate 'trial and error' mentioned above. In this quote, both the stochastic component in generating variation and the

purposive element in discovery and selection of slight beneficial deviations are exposed lucidly:

> A boat constructor may be skilled, and yet he will never get two boats exactly alike, even if he exerts himself to this end. The deviations occurring may be called *accidental*. But even a very small deviation usually is noticeable during navigation, but it is *not accidental* that the sailors come to *notice* that the boat that has become improved or more convenient for their purpose, and that they should recommend it to be *chosen* for *imitation* … One may think that each of these boats is perfected in its way, since it has reached perfection through one-sided development in one particular direction. Each kind of improvement has progressed to the point where further development would entail defects that would more than offset the advantage … Hence I conceive of the process in the following way: when the idea of new and improved forms had first been aroused, then a *long series of prudent experiments* each involving extremely small changes, could lead to the happy result that from the boat constructor's shed there emerged a boat whose like all would desire. (Sundt [1862] 1976)

In this case, then, the material structure of society is modified via learning by doing. Note also that Sundt argued that the recognition of accidental discoveries ('spot and snatch') itself leads to the discovery of the usefulness of experiments ('trial and error'), i.e., to second order learning. Marx used a similar line of argument to explain the differentiation of tools adapted to different tasks, and second order learning when science is used to improve industrial means of production.

Before turning to the fourth type, dialectical change, the analyses of Marx and Sundt may highlight how the same basic mechanism can be used to explain constancy as well as change, i.e., the evolution of social habits, tools, and organizational forms toward an optimal equilibrium pattern. Or, put differently, how continual trial and error leads to optimal responses which are then crystallized as customs, standard operating procedures, or best practices. The skills required are pushed in the same direction and the progression toward a best response pattern leads to homogenization of actions, i.e., to uniforming selection. Likewise, adaptation for divergent purposes generates differentiation. The essence of ordinary work is repetition—i.e., variations on the same theme. But by chance variations and serendipity, skills can be acquired, tools developed, and workgroups rearranged. This is how Marx explains changes in organizations and in aggregates at the macro level.

The first organizational consequence is that craftspeople, the best performers, may become locked into lifetime careers. The second organizational consequence is that, on the basis of skill and mastery, industry may sprout into different occupations—a structural change arising from the social organization of individual skills. That is, a particular task selects a best response that becomes fixed as a routine, different tasks select different types of best responses: 'On that groundwork, each separate branch of production acquires empirically the form that is technically suited to it, slowly perfects it, and as soon as a given degree of maturity has been reached, rapidly crystallizes that form' (Marx [1864] 1967, p. 485). Marx uses his selection model to explain structural differentiation. The resulting structure may be no part of any individual's or group's intentions, though it is a result of their choices.

By the same logic Marx explains change in the social structure of society as the gradual adaptation of different groups of workers. What is the best relative size of groups of specialized workers in different functions? Marx argues that there is a constant tendency toward an equilibrium distribution, forced by competition and the technical interdependence of the various stages of production:

> The result of the labour of one is the starting-point for the labour of the other. The one workman therefore gives occupation directly to the other. The labour-time necessary in each partial process, for attaining the desired effect, is learnt by experience; and the mechanism of Manufacture, as a whole, is based on the assumption that a given result will be obtained in a given time. It is only on this assumption that the various supplementary labour-processes can proceed uninterruptedly, simultaneously, and side by side. It is clear that this direct dependence of the operations, and therefore of the labourers, on each other, compels each one of them to spend on his work no more than the necessary time, and thus a continuity, uniformity, regularity, order, and even intensity or labour, of quite a different kind, is begotten than is to be found in an independent handicraft or even in simple cooperation. The rule, that the labour-time expended on a commodity should not exceed that which is socially necessary for its production, appears, in the production of commodities generally, to be established by the mere effect of competition; since, to express ourselves superficially, each single producer is obliged to sell his commodity at its market price. In Manufacture, on the contrary, the turning out of a given quantum of product in a given time is a technical law of the process of production itself. (Marx [1854] 1967, p. 345)

There are two types of selection pressures in this argument: *structural dependence*, or integration that compels each to spend on his or her work no more than the necessary time, and economic *competition*. In short, the argument is based not just on purposive actors but on their location in a particular social macrostructure (competition) resulting in modification of another part of the structure (occupational composition). This argument also uses a conception of bottlenecks—a condition or situation that retards a process and thwarts the purposes of actors involved. This mode of explanation is often used by social scientists, e.g., for the logic of the industrial revolution: the invention of the Spinning Jenny made weaving a bottleneck and hence led to improved shuttles and machine looms requiring greater power to drive them, favoring the use of the steam engine, etc.

Learning in this case consists in gradual adaptation of group sizes to obtain the most efficient production.

That definite numbers of workers are subjected to definite functions Marx calls 'the iron law of proportionality' (Marx [1854] 1967, p. 355). From firms, Marx moves on to the societal level: adjusting different work groups within the manufacture generates a division of labor at the aggregate or macro level.

Third, the basic selection model is used to explain not only change—i.e., the succession of steps towards the best suited response—but also *constancy*, i.e., the locking in of a structurally best response. 'History shows how the division of labour peculiar to manufacture, strictly so called, acquires the best adapted form at first by experience, as it were behind the backs of the actors, and then, like the guild handicrafts, strives to hold on to that form when once found, and here and there succeeds in keeping it for centuries' (Marx [1854] 1967, p. 363).

The fourth point made by both Sundt and Marx is a conceptual one: learning may take place not just by doing, but also through a second form of learning, namely, *imitation*. That is, those who have found the knack can be emulated, and, as Sundt argued, tools improved by accidents are chosen for imitation. Marx wrote: 'since there are always several generations of labourers living at one time, and working together at the manufacture of a given article, the technical skill, the tricks of the trade thus acquired, become established, and are accumulated and handed down' (Marx [1854] 1967, p. 339). In other words: socially organized purposive actors may learn vicariously. They recognize and copy the improved solutions others find to their own problems.

Basically then, Marx and Sundt argued that a constant technology advances the development of an equilibrium pattern of behavior, corresponding to the best possible practice, which also congeals social forms at the macro level. This takes place through the selection of the most adept methods by individuals, which, when aggregated, moves a group or even a mass of people from one distribution to another. They used selection mechanisms to explain a wide range of phenomena: development of individual dexterity, the evolution of lifetime careers, the homogenization of practices and the differentiation of tools, the evolution of occupations—indeed, the drift of large aggregates as well as equilibrium patterns and constancy.

The fifth point is that they considered new technology as a driving force which may require readaptation of behavioral responses (cf. Arrow 1962). From the notion of change moving toward an equilibrium or constancy Marx introduces the notion of the *constancy of change*. This happens when the random nature of learning by doing yields to the systematic search for new knowledge by applied science:

> The principle, carried out in the factory system, of analysing the process or production into its constituent phases, and of solving the problems thus proposed by the application of mechanics, of chemistry, and the whole range of the natural sciences, becomes the determining principle everywhere. Hence machinery squeezes itself into the manufacturing industries first for one detail process, then for another. Thus the solid crystal of their organisation, based on the old division of labor, becomes dissolved, and makes way for constant changes. (Marx [1854] 1967, p. 461)

Marx' explicit logic, therefore, was not just one of learning by doing and second order learning by out-and-out search and science. He went even further in analyzing how the introduction of new implements required new learning by the users of those implements, and indeed, led to a change in attitudes as well. In other words, learning made people change their own conditions—by adapting to those conditions people changed themselves by learning to learn. This is the double logic of dialectical models.

4. *Models of Dialectical Effects*

Dialectical models analyze structures which motivate actors, by accident or design, to bring forth structural changes which then change the actors' own attributes. In shorthand, such models look like this:

Structure → Actoral Change → Structural Change → Actoral Change

This is, of course, a very important category for social analysis—the more so the longer the time perspective. People have not shaped the society into which they are born—they are molded by a society before they can remake it. But humans can remake themselves indirectly by changing the structure or culture within which they act. A classical text, Max Weber's (1904–5) *The Protestant Ethic and the Spirit of Capitalism*, can serve to illustrate such models.

Weber first identified a specific tendency among Protestants to develop economic rationalism not observed to the same extent among Catholics (Hernes 1989). Weber's paradoxical thesis is that the supposed conflict between otherworldliness or asceticism on the one hand, and participation in capitalistic acquisition on the other, is in fact an intimate relationship.

Weber's model departs from what he calls traditionalism, populated by two types of actors: traditionalistic workers, and traditionalistic entrepreneurs. The key attribute of the former is that increasing their rewards does not lead to an increase in their output, but rather makes them decrease the amount of work they put in. People 'by nature' do not want to earn as much as possible: traditionalistic workers adjust their supply of labor to satisfy their traditional needs. This 'immense stubborn resistance' is the 'leading trait of pre-capitalist labor.' Likewise, traditional entrepreneurs adjust their efforts to maintain the traditional rate of profit, the traditional relations with labor, and the traditional circle of customers.

However, what these actors want is not hard-wired—their character can be changed. Weber wants to explain this change from the tranquil state of traditionalism to the bustling dynamism of capitalism. The first part of his argument is about how the actors' temperaments are changed: how a *homo novus*—or rather two—is brought forth. The second part of the argument is about how they trap themselves and everyone else in the new conditions or structure they create. The new entrepreneurs came to see money-making as an end in itself. The new workers came to see labor as an end in itself. The argument runs briefly as follows.

In the beginning there was logic. Conceptual activity worked on beliefs to turn doctrines of faith into maxims for behavior. At the outset traditionalists had faith in God. He was omniscient and omnipotent. By pure logic this led to predestination: an omniscient God had to know ahead of time who were among the chosen; an omnipotent God must have decided already and must know the outcome of his decisions. Humans could do nothing to bring about their salvation by their own actions. Fates were preordained by God. Those who had salvation could not lose it and those without it could not gain it. What is more, neither group could know who had it. So for every true believer one question must have forced all others into the background: 'Am I one of the elect?' At first there seemed to be no answer; however, all the faithful were to act *as if* they were among the chosen. Doubts were temptations of the Devil—a sign of failing faith, hence of imperfect grace. And the faithful should act in the world to serve God's glory to the best of their abilities. Those unwilling to work therefore had the symptoms of lack of grace. Hence there were *some* signs.

But the logic of this argument resulted in a peculiar psychology of signs—a mental reversal of conditional probability notions. The original argument ran: if you are among the elect, you work in a calling. The psychological reversal became: if you work in a calling, you have a sign that you are among the elect. Moreover, this sign— working—could be affected. By choosing to work hard, you attain a sign that you are among the chosen:

> In practice this means that God helps them who helps themselves. Thus the Calvinist, as it is sometimes put, himself creates his own salvation, or, as would be more correct, the conviction of it. But this creation cannot, as in Catholicism, consist in a gradual accumulation of individual good works to one's credit, but rather in a systematic self-control which at every moment stands before the inexorable alternative, chosen or damned. (Weber 1930, p. 115)

This is the first part of Weber's analysis of actor change: the moral construction of the new actors—new workers and new entrepreneurs working in a calling for the glory of God—driven to restless work by their fierce quest for peace of mind.

However, this actoral change is not a sufficient condition to explain the structural change, i.e., the establishment of a new economic system—capitalism. Hence the next part of Weber's logical chain is the following. Working in a calling was not enough. For even if many were called, few were believed chosen. Identifying who they were—and if one was among them—could only be done by comparisons. 'Success reveals the blessings of God' (Weber 1930, p. 271, n. 58). But success can only be measured in relative terms: 'But God requires *social achievement* of the Christian because He wills that social life shall be organized according to His commandments, in accordance with that purpose' (Weber 1930, p. 108, emphasis in original). Each Christian had, so to speak, to prove that he or she was 'holier than thou.' Moreover, each of them had to assess their success relative to others, and moreover, they had to do it in terms of economic achievement. The sign of salvation is not what you did unto others—it was whether you outdid the others.

Here is the nexus between actor change and structural change. From this mode of thinking, the individualistic Protestants were, so to speak, connected behind their backs. Religious attainment became measured by private profit. The market united those that theology separated. From being religious isolates the Christians became rivals—assessed by the relative fortunes they amassed. But then, to hold their own, they forced each other to work harder. In effect they set up a new social structure for themselves by their thinking and by their actions. Their religious mindset produced the new social setup. By taking not a good day's work, but the *relative* fruits of labor in a calling, as a measure of success, they in effect caught themselves in a prisoner's dilemma.

Since capital accumulated but not capital consumed was measured, the logic of asceticism 'turned with all its force against one thing: the spontaneous enjoyment of life and all it had to offer' (Weber 1930, p. 166). Hence assets could also serve as a relative measure of success. And this had the 'psychological effect of freeing the acquisition of goods from the inhibitions of traditionalistic ethic' (p. 174). The double-barreled effect of ascetic Protestantism, in short, was to join the compulsion to work with the obligation to reinvest its output.

But the prisoner's dilemma that the Protestants established by their particular comparative, even compulsive, work ethic not only captured their creators. By their logic of relative success they also dragged everybody else into the vortex of the market—they changed the social structure in a decisive way because it also entrapped all others:

> The Puritan wanted to work in a calling; we are forced to do so. For when asceticism was carried out of the monastic cells into everyday life, and began to dominate worldly morality, it did its part in building the tremendous cosmos of the modern

> economic order. This order is now bound to the technical and economic conditions of machine production which today determine the lives of all individuals who are born into this mechanism, not only those who are directly concerned with economic acquisition, with irresistible force. (Weber 1930, p. 181)

The structure that the actors created turned on themselves with a compelling might—their internal convictions transformed into external forces, so to speak. Moreover, not only are people compelled to react to conditions of their own making. The new structure of the competitive market in turn creates yet new types of actors without the religious motivation from which it sprang:

> The capitalistic economy of the present day is an immense cosmos into which the individual is born, and which presents itself to him, at least as an individual, as an unalterable order of things in which he must live. It forces the individual, in so far as he is involved in the system of market relationships, to conform to capitalistic rules of action. The manufacturer who in the long run acts counter to these norms, will just as inevitably be eliminated from the economic scene as the worker who cannot or will not adapt himself to them will be thrown into the street without a job. Thus the capitalism of today, which has come to dominate economic life, educates and selects the economic subjects which it needs through a process of economic survival of the fittest. (Weber 1930, p. 54 f.)

By measuring success in relative terms, the Puritans established a prisoner's dilemma in which they not only ambushed themselves and ensnared non-Puritans, but in addition, once it had been socially established, in turn liberated capitalism from its connection to their special *Weltanschauung*. It no longer needed the support of any religious forces: 'Whoever does not adapt his manner of life to the conditions of capitalistic success must go under. But these are phenomena of a time in which modern capitalism has become dominant and has become emancipated from its old supports' (Weber 1930, p. 72).

Hence, having explained the generation of the capitalist system, Weber takes competition to be source of its own reproduction, i.e., for the constancy of the new system. The paradox is that the religious belief that set people free from restraints on acquisitive activity, in turn liberated the resulting economic system from its creators and their beliefs. Once the new system is in place, it selects and shapes the actors needed to maintain it. The Protestant entrepreneurs who established the new system thereby in turn called forth yet another *homo novus*—industrialists and managers and workaholics infused not with Protestantism, but with the spirit of capitalism.

The short version of Weber's analysis is this: religious conversion led to economic comparison; and the importance of relative success as a religious sign led to competition. This was a systemic transformation that in turn led not only to accumulation and innovation, but also to selection and socialization of new actors. A system brought into being by actors who feared eternal damnation bred actors who maintained it from concern for their daily survival and without religious qualms.

Weber's analysis provides a very good example of dialectical change in that it explains how a new type of actor, with different beliefs and knowledge, modifies its own environment out of existence and in turn creates the conditions for yet another type of actor that is different from themselves. As Garrett Hardin (1969, p. 284) has written of ecological succession: 'It is not only true that environments select organisms; organisms make new selective environments' (cf. Hernes 1976, p. 533).

5. *Conclusion*

The analysis of change is at the heart of social theory because it attempts to answer this question: How do humans cope with the effects of their own creativity? To carry out such analysis one needs simple, yet potent, tools. They are embedded in the classical works of social science. When they are identified, we can emulate the best practitioners in the field.

Bibliography

Arrow K 1962 The economic implications of learning by doing. *Review of Economic Studies* **9**(June): 155–73

Hardin G 1969 The cybernetics of competition: A biologist's view of society. In: Shephard P, McKinley D (eds.) *The Subversive Science: Essays Towards an Ecology of Man*. Houghton Mifflin, Boston, pp. 275–95

Hernes G 1976 Structural change in social processes. *American Journal of Sociology* **38**(3): 513–46

Hernes G 1989 The logic of 'The Protestant Ethic.' *Rationality and Society* **1**: 123–62

Marx K [1864] 1967 *Capital*. International Publishers, New York

Ryder N 1964 Notes on the concept of a population. *American Journal of Sociology* **69**(March): 447–63

Sundt E [1862] 1976 Om havet. In: *Samlede verker*. Gyldendal, Oslo, Vol. 7

Weber M [1904–5] 1930 *The Protestant Ethic and the Spirit of Capitalism*. Translated by Talcott Parsons, Allen & Unwin, London

G. Hernes

Social Changes: Models

There is an amazing multiplicity of social forms: rites, tools, languages, and rules. How has this variety come about and why is it forever changing? Yet most of the

time humans behave like those around them. How has this conformity developed? And why is it maintained?

Models of social change address these questions: What generates change, and is it differentiating or uniforming? What maintains established structures, whether they are simple or complex?

Classics, such as Durkheim, Marx, Sundt and Weber, provided models of the transformation as well as the preservation of social forms, and analyzed diversification as well as homogenization. This they did not so much by general theoretical statements as by laying bare the social logic driving specific historical passages. They can be used to highlight models of social change and provide paragons that present-day studies can be appraised against.

1. *Individual Actions and Social Facts*

Durkheim, in *Suicide* ([1897] 1951) sought to explain both constancy and variety. He asked: why are suicide rates so stable—indeed, more stable than ordinary death rates—and yet vary systematically between groups? Why do they take on a new stubborn stability after marked shifts? Durkheim held stable, but differing suicide rates to be social facts. They exist relatively independent of the individuals who compose the group on which they are measured. Social facts—e.g., the language we speak, rites we share, or streets we walk—endure, Durkheim argued, *sui generis*.

Social facts have an existence of their own, like a swarm of moths which can change its shape or shift its position—yet, of course, there are no social phenomena without individuals being part of them: no swarm without the ephemeral movements of moths. Put paradoxically: social facts—both change and constancy—are relatively independent of those on whom they depend. Individuals are held in place by their fellows and take their cues from each other. This is important in study of social change: it is always mediated through individuals who are shaped by their physical and social environment before they can act on it, and who change it by reacting to circumstances they have brought forth or are transmitted from the past.

Social change is at the core of sociological analysis. Yet, few studies cover the whole cycle of analyzing 'how men react to conditions of their own making and in so doing change these conditions' (Hernes 1976, p. 513). Rather they present snapshots of a condition, such as a poll of opinions or their causal determinants. The reasons are many: collecting data as a process unfolds may be too costly or time-consuming. Data from other sources do not easily lend themselves to analysis over time. Survey techniques often result in a kind of aggregate social psychology rather than in studies of structural change, and when it restricts itself to 'enumerating individual characteristics, [it] treats the individual as if he were detached from his environment and hence as an abstraction' (Boudon 1971, p. 48). Indeed, the methodological successes of survey analysis can lead to conceptual failures. So most social scientists do not go into deep analysis of social change even if it is at the core of social analysis.

2. *Requirements of Theories of Social Change*

From the above observations, three logical requirements of theories of social change can be identified (Hernes 1976). First, it should be possible to use the same basic approach in studying both constancy and change. This implies that a stable structure should be seen as a process in (temporary) equilibrium and that the theory must include an explanation of why the forces or parameters governing a process themselves are regenerated. Structural change must be described in terms of the processes that generate the change. Structural stability must be explained in terms of the processes that not only maintain the stability, but also maintain the processes that maintain the stability.

Some accounts are static, e.g., 'percentage unemployed.' Yet, from a model point of view, it not only makes better sense, but also for more interesting analysis to conceptualize even 'static' phenomena in terms of processes. For example, if the number unemployed remains stable, this does not imply that the same individuals remain out of work—only that the flows of individuals in and out of employment cancel out. Hence stability can be viewed as a process in equilibrium (Coleman 1964).

Second, models of social change should incorporate intrinsic sources of change. Changes in social systems may be induced from the outside, but all changes cannot be thus induced (Elster 1971). Even when the impetus for change is imported, the mechanisms transmitting external impacts should be identified. Indeed, when an item is taken from the outside, it should be possible to employ the same kinds of explanations for its adoption as if it had been invented at home. Moreover, the sources of change should also be intrinsic to the theory in the sense that effects of previous actions provide or change premises for further choices of action.

Lauriston Sharp (1952) gives an example in 'Steel axes for Stone Age Australians.' He describes how the aboriginal Yir Yoront group as late as the 1930s lived an essentially isolated and independent economic existence based on Stone Age techniques. However, 'their polished stone axes were disappearing fast and being replaced by steel axes which came to them in considerable numbers, directly or indirectly, from various European sources to the south' (Sharp 1951, p. 17). Sharp then discusses what changes in the lives of the group resulted from the replacement of stone axes around which their culture was built. For example, stone axes were an item of trade hence important in intertribal relations, belief systems, celebrations, and totemic ceremonies. Men crafted the stone axes, which were a symbol of their masculinity, yet they were lent

to women who used them in their heavy chores. Furthermore, 'The shift from stone to steel provided no major technological difficulties' (Sharp 1952, p. 20). Sharp then analyzes how this change of tools changed trade relations, gender interactions, and authority structures.

> The most disturbing effects of the steel axe, operating in conjunction with other elements also being introduced from the white man's several sub-cultures, developed in the realms of traditional ideas, sentiments, and values. These were undermined at a rapidly mounting rate, with no new conceptions being defined to replace them. The result was the erection of a mental and moral void which foreshadowed the collapse and destruction of all Yir Yoront cultures, if not, indeed, the extinction of the biological group itself. (Sharp 1952, p. 21)

Here the new technology came from the outside—the impetus for change was exogenous. But its adoption is explained in terms of simple principles at the level of actors: steel axes were nearly free good tools, simple to use, and requiring less effort—the same types of explanation we would use if they had been invented locally. The effects of this *deus ex machina* changed premises for further choices of action—in this case with devastating aggregate effects: it reduced women's dependence on men, it undermined belief systems, it wrecked havoc with traditional trade relations and their social embellishments.

Finally, as just seen, social change is mediated through individual actors. Hence the third requirement of theories of social change: they must show how macrovariables or structural characteristics affect individual motives, capacities, and choices and how the choices of actors in turn change the macrovariables. For the Yir Yoront, adopting a new tool changed their social structure in ways that was no part of their intention.

3. *Elements of Models of Social Change*

Next we can identify the locuses of change. Basically they can be of two kinds: (a) the actors can change, and (b) the structure within which they operate can change. Models must be constructed from these two sets of abstract elements—actors and structures—to mirror what drives the changes in them (Hernes 1989, 1999).

Hence, the first element to specify in a model of social change is the types of actors assumed to operate. In the language of the theater, specifying the actors corresponds to the casting of a play. For example, in microeconomic theory two types of actors are distinguished: consumers maximizing utility by spending their resources to obtain goods, and producers maximizing profit by the quantities they market.

The second element to specify is the structure the actors are lodged in—the staging of a play. For example, in microeconomic theory three main market structures are distinguished: perfect competition where the actors are so many and small that none of them can affect the price; oligopoly, where the producers are so few that they react to each others' decisions; and monopoly, where a single producer can set the price.

One reason for combining actors and the structure within which they operate, is to construct a model so that one can logically work out the systemic effects of actors' choices within the structure—the plotting of the play. The point is to answer questions about what happens to the actors or to the structure as a consequence of the actors reacting to the initial conditions—i.e., the logical sequence of consequences that flow from the assumptions made. In other words, we construct models of actors within a structure in order to understand what the actors are doing and why, and to explain what results from their actions and why, whether the results are intended or not. In other words, a model of change is a logical narrative. A logical narrative may have logical flaws and hence be theoretically unsound. And one such logical narrative may exclude others: it simultaneously empowers our reasoning and limits our comprehension—both our understanding of the actors and the explanation of the results of their interaction.

A second purpose is to compare the logical outcomes of the model to a real life sequence of events. A compelling argument may be empirically false—the logical narrative may not match the historical record. Hence, if the model is to provide a valid understanding or a tenable explanation, there must be a rough goodness of fit between what the model logically leads us to expect about what will transpire and what actually takes place in the real world. The model cannot just be plausible in theory—it must at the end of the day be admissible by the facts: the observed must agree with the inferred. If the model is ruled out by the facts, it must be discarded and another constructed from different assumptions that provide a more realistic understanding and credible explanation.

For example, the simple model of perfect competition implies that profits will tend towards zero. If we observe that profits in fact are high, we may modify the model to better account for this fact by changing the assumptions, e.g., by allowing producers to change consumers' attitudes (advertising) or by allowing for monopolization—i.e., discarding the model of perfect competition.

4. *Specifying Actor Assumptions and Structure Assumptions*

To give a logical portrait of the types of actors involved in a social process, the following kinds of questions must be answered:

(a) What do they *want*—what are their preferences or purposes?

(b) What do they *know*—what are their beliefs, their potential for changing them or for learning?

(c) What can they *do*—what are their capacities or skills?

Sometimes answers are only rudimentarily specified, e.g., in models of contagion where contracting a disease clearly is not something sought but something caught.

Actor assumptions allow us to understand what actors can do, what they actually do and why, or how others react to them and why.

A similar set of questions must be answered to specify assumptions about the structure:

(a) What *alternatives* or *constraints* do the actors face? These may span from opportunity for migration to household budgets. They include collective properties such as laws and norms and physical barriers such as roads or rivers.

(b) What *social positions* can the actors hold—e.g., organizational roles? What are the *correlates of positions*—e.g., their status or the hierarchy of wages linked to a given division of labor? Rewards, rights, and burdens are structural correlates of positions—they also define, fix, or freeze the alternatives actors face.

(c) What are the *distributional characteristics* of actors, as distinct from their individual manifestations? For example, the number of other actors may affect options as in perfect competition, the age distribution may affect outlooks for marriage. The sex ratio and the population pyramid are structural characteristics.

Specifications of the actors' attributes are needed to understand what they do and why. Specifications of the structure—options and payoffs—within which they choose and act are needed to explain what happens and why. The two sets of assumptions are needed to construct the logical narrative of systemic effects.

5. *Intentions vs. Effects*

A key point in analyzing social change is the nexus between intentions and effects. Purposive action may have unanticipated consequences (Merton [1936] 1976). That is why understanding the actors is not sufficient—explaining the interworkings of their actions beyond their intentions is necessary. The net outcome of rational choices may be counterintuitive. The way to hell may be paved with good intentions and simultaneous choices based on the best of motives may be counterproductive by some hidden hex. Private vices may give public benefits (Mandeville [1714] 1989), self-serving interest may yield benefits as if guided by an invisible hand. Sometimes actors are caught by what they have wrought; sometimes they do not master the world they have made. Such unintended effects, counterproductive actions and counterintuitive results which flow from the actors' choices, not their intentions, often produce engaging logical narratives.

6. *Expected vs. Desired Consequences*

The distinction between actors and structure is useful also when what happens deviates from what was anticipated.

In 1936 Robert K. Merton wrote a famous article on 'The unanticipated consequences of purposive social action' (Merton [1936] 1976). A key point is that the effects of actors' choices may interact to produce effects that are no part of the intentions. This notion is very useful for describing models or mechanisms of social change.

The simplest models of change are of course when purposive actors produce intended results. What appears is what was planned, what materializes is what was sought. However, sometimes what turns up is not what the actors had in mind. Sometimes their best intentions interact to produce self-defeating results, as when a stampede to get out blocks the exit. Sometimes selfish motives produce positive overall effects over and beyond their expectations, with improved welfare coming as a by-product, as if an invisible hand guides self-interest to efficient use of resources. As the argument runs in Adam Smith's classic text:

> He [every individual] generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. By preferring the support of domestic to that of foreign industry, he intends only his own security; and by directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for the society that it was no part of it. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it. I have never known much good done by those who affected to trade for the public good. It is an affectation, indeed, not very common among merchants, and very few words need be employed in dissuading them from it. (Smith [1776] 1937, Book 4, Chap. 2)

A key point for the analysis of social change is that what results, whether it appears as the compelling influence from a hidden hex or an invisible hand, flows from the actors' choices, not from their motives. That is why understanding the actors is not sufficient—explanation of interworkings of their actions beyond their intention, expectation, or belief is necessary. Simultaneous choices based on the best of motives may be counterproductive. The net outcome of rational choices may be counterintuitive. Silly choices

Are the consequences expected?	Are the consequences desired? Yes	Are the consequences desired? No
Yes	Intention = consequence	Suboptimality
No	Invisible hand	Counterfinality

Figure 1
Actions and consequences

may prove lucky strikes, bad motives may result in good fortunes. The way to hell may be paved with good intentions just as private vices may give public benefits (Mandeville [1714] 1989).

In fact, we can classify types of explanations from the answers to two questions: were the consequences expected or not, and were they desired or not? If the results coincide with what was desired, the actors have produced the intended results.

If the results were not expected, but when they appear are desired, we have what could be called 'invisible hand' explanations, since they are of the kind illustrated by the quote from Adam Smith above.

The case where the consequences are expected, but actors do not like them when they turn up, can be termed *suboptimality* (cf. Elster 1978): the actors know that when they act in their own interest, they nevertheless collectively end up in an undesired state—the joint outcome of their choices make them worse off. 'The Tragedy of the commons' illustrates the logic. A 'commons' is any resource that can be used as though it belongs to all. However, when rational agents use it, such a shared resource is doomed, though this is no part of the users' intent. Everyone gains from using it, none gains anything from not taking out what they can when others do. When utilized near carrying capacity, additional use will degrade its value to all. 'Ruin is the destination toward which all men rush, each pursuing his own best interest in a society that believes in the freedom of the commons' (Hardin 1968). The inexorable working out of the resource's ruin is Hardin's tragedy of the commons. Social change is the consequence of the sum of self-interested actions. And all could be better off by restraints on their own action. If such collective restraints are imposed by the actors themselves, this represents a change of political strutcture. This latter logic and model of structural change was first suggested by Thomas Hobbes ([1651] 1982) who called the commons 'the state of nature' and 'war of all against all,' whereas the joint establishment of an authority to keep all from self-defeating actions he called Leviathan (cf. Hernes 1999).

The fourth case is where the actors misjudge what others will do. Intentions that could be realized under constant conditions are rendered impossible when everyone seeks to realize them and change the conditions. Hog cycles is an economic example: producers act on the same assumptions and overinvest or underinvest in concert—the results of their decisions appear too late to become premises for their decisions. Erosion is another example—the irreversible destruction of alternatives. With a term Elster (1978) borrows from Sartre, this could be called *counterfinality*.

The four kinds of models are summarized in Fig. 1. Sometimes the actors get caught by what they have wrought; sometimes they do not master the world they have made. But independent actions can also turn out to be blessings in disguise.

7. *Models vs. Types of Social Changes*

Some key aspects of models of social change have been identified. A question of equal import is whether it is possible to identify basic types of social change. To answer this question requires a separate analysis.

See also: Action, Collective; Action, Theories of Social; Durkheim, Emile (1858–1917); Marx, Karl (1818–89); Organizations: Unintended Consequences; Social Change: Types; Sociology: Overview; Structuralism, Theories of; Structure: Social; Theory: Sociological; Weber, Max (1864–1920)

Bibliography

Boudon R 1971 *The Uses of Structuralism*. Heinemann, London
Coleman J 1964 *Introduction to Mathematical Sociology*. The Free Press, Glencoe, IL
Durkheim E [1897] 1951 *Suicide*. The Free Press, Glencoe, IL

Elster J 1971 *Nytt syn på økonomisk historie*. Pax, Oslo
Elster J 1978 *Logic and Society. Contradictions and Possible Worlds*. Wiley, Chichester, UK
Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–48
Hernes G 1976 Structural change in social processes. *American Journal of Sociology* **38**(3): 513–46
Hernes G 1989 The logic of 'The Protestant Ethic'. *Rationality and Society* **1**: 123–62
Hernes G 1999 Real virtuality. In: Hedström P, Swedberg R (eds.) *Social Mechanisms: An Analytic Approach to Social Theory*. Cambridge University Press, Cambridge, UK, pp. 74–101
Hobbes T [1651] 1982 *Leviathan*. Penguin, London
Mandeville B [1714] 1989 *The Fable of the Bees*. Penguin, London
Merton R K [1936] 1976 The unanticipated consequences of purposive social action. In: Merton R K (ed.) *Sociological Ambivalence*. The Free Press, Glencoe, IL
Sharp L 1952 Steel axes for Stone Age Australians. *Human Organization* **2**: 17–22
Smith A [1776] 1937 *The Wealth of Nations*. The Modern Library, New York

G. Hernes

Social Class and Gender

Both 'class' and 'gender' are contested concepts. Debates around gender and class have focused on employment-based approaches to 'class analysis.' It is argued that these have reached an *impasse*. It is, however, still of importance to retain a focus on the *processes* of gender and class structuring.

1. *Concepts and Definitions*

Three different meanings (or dimensions) of the class concept may be identified:

(a) 'Class' as prestige, status, culture, or 'lifestyles.'

(b) 'Class' as structured inequality (related to the possession of economic and power resources).

(c) 'Classes' as actual or potential social and political actors.

The use of 'class'—often as an adjective—to describe prestige or lifestyle is probably the most common use of the concept in everyday speech. Social scientists have also carried out descriptive investigations of social differentiation (e.g., Warner 1963), and market research companies have devised occupational classifications in order to give an indication of lifestyles and consumption patterns.

Following Weber, many sociologists would insist on the distinction between this use of 'class' to indicate prestige, or consumption patterns, and the use of 'class' in the second sense above, that is to describe groups with differential access to economic and power resources (these could be the ownership of capital or productive resources, skills and qualifications, networks of contacts, etc.) that generate structures of material inequality. Such unequally rewarded groups are often described as 'classes.' The identification of these groups summarizes the outcome, in material terms, of the competition for resources in capitalist market societies.

The term 'class,' however, has not just been used to describe levels of material inequalities or social prestige. 'Classes' have also been identified as actual or potential social forces, which have the capacity to transform society. Marx considered the struggle between classes to be the major motive force in human history. From the French Revolution and before, 'classes'—particularly the lower classes—have been regarded as a possible threat to the established order. Thus 'class' is also a term with significant *political* overtones.

The use of the single word 'class,' therefore, may describe hierarchical rankings of social prestige or lifestyles, patterns of material inequalities, as well as revolutionary or conservative social actors. It is a concept that is not the particular preserve of any individual branch of social science—unlike, for example, the concept of marginal utility in economics. It is not possible to identify a 'correct' sociological perspective, or an agreed use of the term.

In contrast to 'class,' the concept of gender might seem to be unproblematic, given its overt physical manifestations in all spheres of human life. However, its meaning has also been disputed. It is commonplace to draw a distinction between sex and gender, where sex denotes anatomical and physiological differences, and gender the social construction of masculinity and femininity. This is not universal (i.e., there is no 'essence' of masculinity or femininity), but created and interpreted differently by different cultures. However, recent theoretical developments within poststructuralist and postmodernist feminism have argued that sex itself is a social construct rather than biologically given, and the categories 'men' and 'women' *discursively* constructed through 'performance.' Thus, 'gender' has become overlaid with 'sexuality.' This has resulted in a focus on the sexual identity of the individual, and the discourses through which these identities are constructed, rather than on gender as a set of social relations.

However, an emphasis on the discursive construction of gender (or sex) effectively detracts from the investigation of the systematic structuring of inequality between men and women (through, for example, the gender division of labor). The question of sexuality has a focus on the *individual* and their associated identity—albeit an individual that is constructed, in process, undetermined, etc.—whereas gender is inevitably a *relational* concept. However, although gender, like class, is a social relationship of

power, there is a crucial difference between the two phenomena. Were private property to be abolished overnight, and class differentiation processes outlawed, there would be no 'classes.' However, were the differentiation processes of 'gender' as we know it to cease tomorrow, approximately half of the human population would nevertheless continue to have breasts and wombs. To recognize the biological roots of gender, however, is not to collapse into a biological determinism or essentialism. While gender differentiation is universal and persistent, what constitutes this concrete 'difference' will be variable in its nature and extent, and subject to constant construction and reconstruction.

2. Approaches to Class Analysis

The three different meanings of the class concept identified above are widely employed in a range of different approaches to the analysis of class and stratification. These approaches may be grouped within four areas (Crompton 1993, 1998):

(a) The study of the *processes* of the emergence and perpetuation of advantaged and disadvantaged groups or 'classes' within society. This may be described as 'class analysis' as a 'dependent variable' enterprise, in which the 'class' (or classes) itself is the object of study. All three meanings of 'class' may be employed in this approach, and indeed, these different dimensions are often seen as being intertwined with each other—as in, for example, E. P. Thompson's *The Making of the English Working Class* (1968).

(b) The study of the concrete consequences (or *outcomes*) of class location. This may be described as 'class' as an 'independent variable' specialty, in which 'class' is the constant factor and its contribution to other factors—e.g. voting behavior—is that which has to be explained (Wright 1997). Occupational 'class' may be related to consumption, inequalities, and political preferences, thus all three meanings of 'class' may be drawn upon in empirical investigations. However, the methodology of 'employment aggregate' class analysis means that these different dimensions of class have to be treated as discrete variables. Thus, for example, its practitioners argue that it is important to establish that occupational classifications do not correspond to a 'status' scale. In practice, however, economic and cultural dimensions are closely intertwined (particularly in the case of gender), and this separation is problematic.

(c) The discussion of the significance of 'class,' and class processes, for macro theories of societal change and development. These discussions mainly relate to the third meaning of class identified above. In Marx's work, 'classes' are seen as having the causal capacity to shape societies and generate social change (Sayer and Walker 1992; it should be noted that these claims are not only made by Marxists). Much of the debate with 'reflexive' or 'post' modernist critics of 'class analysis' has centered on this issue, i.e., whether, with increasing 'individuation,' 'classes' are still relevant social forces (Pakulski and Waters 1996, Clark and Lipset 1991). These preoccupations lead directly into a fourth approach to 'class' within sociology:

(d) The development of class and status cultures and identities. These discussions have tended to have a major focus on the first meaning of class identified above. However, Bourdieu's (1973) influential work combines an analysis of both class cultures and class inequalities, and focuses on both the processes of class formation as well as their outcomes.

3. Processes and Outcomes

The notion of class *processes*, therefore, may be used as a general term to identify the patterns of change, tensions, and struggle that generate the differentiated groupings that we call 'classes'—although, as we have stressed above, there would be no general agreement on the composition of these 'classes.' We have to distinguish between the processes of the formation and emergence of advantaged and disadvantaged, more powerful and less powerful, groups (or 'classes') in society, and the concrete consequences or outcomes of these processes. It has been suggested that this distinction broadly corresponds to two rather different approaches to 'class analysis' ((a) and (b) above). In a similar vein, it may be argued that it is important to distinguish between the processes of differentiation between men and women—that is, the construction of gender as a social relation—and the concrete consequences of this construction, that is, measurable differences between the sexes when taken in aggregate. Thus measurable 'sex differences' may be seen as outcomes of gender differentiation processes.

4. Measuring Social Class

The occupational structure is often used to generate 'class' groupings. This approach to the measurement of class may be described as the 'employment aggregate' approach, and is so widespread that the occupational structure and the class structure are frequently referred to as if they were synonymous.

From the 1970s leading practitioners of 'class analysis' argued that the classifications used by leading researchers of the time—(e.g., Blau and Duncan 1967)—did not measure 'class' in a truly sociological sense. Albeit from rather different theoretical perspectives, both Wright and Goldthorpe argued that the occupational measures previously used had in fact been hierarchical ('gradational') status scales rather than 'relational' class schemes. Thus throughout the

1980s, two major cross-national projects, both of which developed their own, employment-based class schemes, were established. The International Class Project, directed by Wright (1985, 1997), was explicitly Marxist in its inspiration and the scheme(s) he devised classified jobs according to a Marxist analysis of relations of domination and exploitation in production. The Comparative Analysis of Social Mobility in Industrial Societies (CASMIN) project used an occupational classification initially derived from Goldthorpe's study of social mobility.

Even in these improved versions, however, there are a number of problems with the assumption that the job or occupational structure may be divided so as to generate 'classes.' The occupational structure does not give any indication of wealth or property holdings. Not all adult persons have an occupation and there are therefore problems in allocating an individual 'class situation'—this has been particularly problematic in relation to the 'economically inactive,' such as housewives. The occupational structure also bears the imprint of other major stratifying factors—notably gender, race, and age—which are difficult to disentangle from those of 'class.'

We will expand our discussion of feminist criticisms of 'class analysis' in the next section, but for the moment, it is important to reiterate the point that occupational class schemes are only a proxy for 'classes.' Occupation gives a reasonable indication of 'life chances,' but no accurate measure of property or wealth. Employment aggregates do not correspond to politically conscious collectivities. Rather, sociological class schemes may be seen as an approximate measure (probably the best we have) of the outcomes of the *processes* of class differentiation. Wright and Goldthorpe take different positions regarding the identification of these processes. Wright's location of jobs into particular classes is guided by Marxist class theory. However, Goldthorpe emphasizes that no particular theory has contributed to his class scheme which must be judged by its empirical adequacy—'it is consequences, not antecedents, that matter.' Nevertheless, the schemes intercorrelate quite highly and indeed Wright suggests that his scheme can be interpreted 'in a Weberian or hybrid manner.'

5. *Feminist Criticisms of 'Class Analysis'*

In contrast to both Goldthorpe and Wright, feminists have argued that the processes of class formation and emergence—the divisions of capital and labor which led to the development of the bourgeoisie and mass proletariat—were intimately bound up with parallel processes of gender differentiation (e.g., Bradley 1989). They have described the processes whereby the sexual division of labor which culminated in that stage of modern capitalism we may loosely describe as 'Fordist' was characterized by the 'male breadwinner' model of the division of labor, around which 'masculine' and 'feminine' gender blocs were crystallized.

Thus feminist criticisms of employment aggregate class analysis derive from the observation that, because the primacy of women's family responsibilities has been explicitly or implicitly treated as 'natural' and men have been dominant in the employment sphere, this approach has effectively excluded women from any systematic consideration in 'class analysis.' It is argued:

(a) That the primary focus on paid employment does not take into account the unpaid domestic labor of women. Thus, women's contribution to production is not examined or analyzed (as in, for example, the debate around 'domestic labor,' see Molyneux 1979).

(b) The expansion of married women's paid employment has, apparently, rendered problematic the practice of taking the 'male breadwinner's' occupation as a proxy for the 'class situation' of the household.

(c) That gendered occupational segregation makes it difficult to construct universalistic 'class schemes' (that is, classifications equally applicable to men and women). The crowding of women into lower-level occupations, as well as the stereotypical or cultural 'gendering' of particular occupations (such as nursing, for example), gives very different 'class structures' for men and women when the same scheme is applied to both. Even more problematic is the fact that the same occupation (or 'class situation') may be associated with very different 'life chances' as far as men and women are concerned (for example, clerical work).

This third criticism emphasizes the *de facto* intertwining of class and gender within the employment structure (Crompton and Mann 1994). As noted above, the occupational structure emerging in many industrial societies in the nineteenth and early twentieth centuries was grounded in a division of labor in which women took the primary responsibility for domestic work, while male 'breadwinners' specialized in market work. Today this is changing in that married women have taken up market work, and this has had very important consequences for employment aggregate class analysis.

5.1 *Response of Leading Practitioners*

When Goldthorpe's major national investigation of the British class structure was published (1980/1987), it was subjected to extensive criticism on the grounds that it focused entirely on men, women only being included as wives. However, Goldthorpe argued that as the family is the unit of 'class analysis,' then the 'class position' of the family can be taken to be that of the head of the household, who will usually be a male. Thus were the assumptions of the male breadwinner model incorporated into sociological class analysis in

Britain. To incorporate women's employment into 'class analysis' on the same terms as men's, he argued, would lead to confusion. Many women work in lower-level white-collar jobs ('Intermediate' class locations in Goldthorpe's original class scheme), and to include their occupations would result in the generation of 'excessive' amounts of spurious social mobility.

However, he subsequently modified his original position in adopting, with Erikson, a 'dominance' strategy, in which the class position of the household is taken as that of the 'dominant' occupation in material terms—whether a man or a woman holds this occupation. Furthermore, although Erikson and Goldthorpe (1992) still insist that the unit of class analysis is the household, their class scheme has been modified in its application to women as individuals. Class IIIb (routine nonmanual) has been categorized as 'intermediate' for men, but 'labor contract' for women. Besides these modifications of his original approach, Goldthorpe has demonstrated that as measured by the Goldthorpe class scheme, the pattern of women's relative rates of social mobility closely parallels that of men. Thus although, because of the effects of occupational segregation, the distribution of men and women within occupational 'class' schemes is different, when considered separately, the impact of 'class' on mobility chances is similar for men and women.

Wright's analysis generally takes the individual, rather than the household, to be the unit of class analysis. However, in respect of economically inactive housewives, Wright (1997) employs a similar strategy to that of Goldthorpe. He introduces the notion of a 'derived' class location, which provides a 'mediated' linkage to the class structure via the class location of others. Wright is sensitive to the issue of gender, and the fact that gender is a major sorting mechanism within the occupational structure as well as reciprocally interacting with class. Nevertheless, he argues that while gender is indeed highly relevant for understanding and explaining the concrete lived experiences of people, it does not follow that gender should be *incorporated* into the abstract concept of 'class.' Thus in his empirical work, 'class' and 'gender' are maintained as separate factors.

We can see, therefore, that in responding to feminist criticisms, both Goldthorpe and Wright insist that class and gender should be considered as *distinct* causal processes. This analytical separation of class and gender may be seen as part of a more general strategy within the employment aggregate approach in which the continuing relevance of 'class' is demonstrated by the empirical evidence of 'class effects' (Goldthorpe and Marshall 1992). Although, therefore, Goldthorpe and Wright have apparently developed very different approaches to 'class analysis,' their underlying approach to the articulation of gender with class is in fact the same. It may be suggested that this stems from the similarity of the empirical techniques used by the CASMIN and International Class Projects: that is, the large-scale, cross-nationally comparative, sample survey. This kind of research proceeds by isolating a particular variable—in this case, employment class—and measuring its effects.

Feminist criticisms, therefore, were important in making explicit the fact that the 'employment-aggregate' approach within 'class analysis' is largely concerned with the *outcomes* of employment structuring (via its analysis of occupational aggregates or 'classes'), rather than the *processes* of this structuring. To paraphrase Goldthorpe and Marshall, employment aggregate 'class analysis' now appears as a rather less ambitious project than it once appeared to be. In a parallel fashion, it may be suggested that a major weakness of Wright's class project (not specific to the gender question) is that the linkage between Marxist theory and his 'class' categories has not been successfully achieved.

Although important issues have been clarified, therefore, this debate can go no further, and Wright argues that it is necessary to get on with '… the messy business of empirically examining the way class and gender intersect.' While one may be in broad agreement with this sentiment, it may be noted that Wright's preferred methodology focuses only on the association between job categories and biological sex. However, developments within the employment structures of the advanced service economies suggest the need for an approach that recognizes the complex processes of the structuring of both class and gender.

6. *The Restructuring of Gender and Class: the Shift to Services*

In today's advanced service economies, employment is still a significant, although not the only, determinant of 'life chances' and access to them. The last 20 years, however, have seen two major, and related, developments. These are, first, the shift to service work, and second, the growth and development of women's employment.

The growth of male-dominated industrial employment was paralleled by the increasing development of social protections associated with the welfare state. Welfare state institutions, such as occupational pensions and other social benefits, were in many countries explicitly created on the assumption that the male breadwinner model was the norm, and women received social benefits via their 'breadwinner.' In countries without extensive state welfare protections, such as the US, there has been an extensive development of marketized personal services. The expansion of the welfare state (as well as the expansion of other services such as education and health) was a major source of employment growth for women. Similarly, marketized personal services are also female dominated. Thus the basis of the male breadwinner model was being eroded even as its principles were

being consolidated in national institutions and policies.

Other factors were also leading to the growth of women's employment, including 'push' factors such as rising levels of education, effective fertility controls, and the growth of 'second wave' feminism, as well as 'pull' factors including the buoyant labor markets of the 1950s and 1960s. Service expansion was fueled not only by state expenditure and job creation, but also the growth of financial, leisure, and business services.

In the West, much of the growth of women's employment in the decades after World War II was in low-level and subordinate jobs. This reflected not only the directly discriminatory employment practices that were a major target of feminist action and protest in the 1960s and 1970s, but also the relatively low level of formal employment experience among women entering the labor force at this time. However, over the last 10 years, women have been steadily increasing their representation in managerial and professional occupations.

7. *The Growth of Women in Management and the Professions*

In all Western countries, the proportion of women in professional and managerial occupations is increasing at a faster rate than the proportion of women in the labor force as a whole. The level of increase of women's representation in some heavily feminized professions, such as teaching, is relatively low, suggesting that these professions might be nearing female saturation. In other professions such as engineering, the increase of female representation is again only modest, but in these instances, from a low base. Well above average rates of increase in the numbers and proportions of women are to be found in public service, financial and legal professions, in 'people centered' management such as personnel, marketing, and advertising, and above average rates in previously male dominated medical professions such as medicine. These trends support Esping-Andersen's (1993) assertion that the postindustrial (or service) job hierarchy is becoming increasingly feminized.

The feminization of the middle classes would appear to be a universal trend. This means that growing numbers of women are earning sufficient amounts to enable them to support a household. There has been an increase in the proportion of 'dual breadwinner' households. One outcome of this trend has been growing social polarization at the level of the household, as the gap between two income and one income households widens. Occupational polarization is also occurring with the growth of a new servant class. Another major outcome of the increase of women in managerial and professional occupations has been a decline in fertility levels, particularly among well-educated women. These demographic trends serve to illustrate, in a rather dramatic fashion, the *de facto* intertwining of employment and domestic life, which was one of the major points emphasized by 'second-wave' feminism.

8. *Discussion*

Changes in the gender division of labor are leading to increasing tensions between the demands of employment and caring responsibilities (Hochschild 1997), and some have argued that gender conflicts are replacing class conflicts (Beck and Beck-Gernsheim 1995). We have suggested, however, that rather than viewing 'gender' and 'class' as *alternative* causal processes, we should rather focus on their interaction. With the mass entry of women into the labor force, it is increasingly problematic to regard paid and unpaid labor as residing in 'separate spheres,' as class analysis conventionally has done. Demands originating from the domestic realm—for example, for time to carry out caring responsibilities or unpaid labor (both men and women may make these demands)—might potentially have a substantial impact on the structuring of paid employment and thus on a significant aspect of class relations. Thus although debates between feminist critics and 'employment aggregate' practitioners may have ended in stalemate, the actual need for a refocus on the interaction of gender and class *processes* is, arguably, greater than ever.

See also: Class Consciousness; Class: Social; Feminist Theory: Marxist and Socialist; Gender and Feminist Studies in Sociology; Gender, Class, Race, and Ethnicity, Social Construction of; Labor Movements and Gender; Right-wing Movements in the United States: Women and Gender; Sex Differences in Pay; Sex Segregation at Work; Social Movements and Gender

Bibliography

Beck U, Beck-Gernsheim E 1995 *The Normal Chaos of Love*. Polity Press, Cambridge, UK

Blau P, Duncan O D 1967 *The American Occupational Structure*. John Wiley, New York

Bourdieu P 1973 *Distinction: A Social Critique of the Judgement of Taste*. Routledge, London/New York

Bradley H 1989 *Men's Work, Women's Work*. Polity Press, Cambridge, UK

Clark T, Lipset S M 1991 Are social classes dying? *International Sociology* **6**(4): 397–410

Crompton R 1993/1998 *Class and Stratification*. Polity Press, Cambridge, UK

Crompton R, Mann M (eds.) 1986/1996 *Gender and Stratification*. Polity Press, Cambridge, UK

Erikson R, Goldthorpe J H 1992 *The Constant Flux*. Clarendon Press, Oxford, UK

Esping-Andersen G (ed.) 1993 *Changing Classes: Stratification and Mobility in Post-Industrial Societies*. Sage, London
Goldthorpe J H 1980/1987 *Social Mobility and Class Structure in Modern Britain*. Clarendon Press, Oxford, UK
Goldthorpe J H, Marshall G 1992 The promising future of class analysis. *Sociology* **26**(3): 381–400
Hochschild A R 1997 *The Time Bind*. Metropolitan Books, New York
Molyneux M 1979 Beyond the domestic labour debate. *New Left Review* **116**
Pakulski J, Waters M 1996 The reshaping and dissolution of social class in advanced society. *Theory and Society* **25**: 667–91
Sayer A, Walker R 1992 *The New Social Economy*. Blackwell Publishers, Oxford, UK
Thompson E P 1968 *The Making of the English Working Class*. Penguin, Harmondsworth, UK
Warner L 1963 *Yankee City*. Yale University Press, New Haven, CT
Wright E O 1985 *Classes*. Verso, London
Wright E O 1997 *Class Counts*. Cambridge University Press, Cambridge, UK

R. Crompton

Social Cognition and Affect, Psychology of

The role of affect in social cognition has fascinated writers and philosophers since the dawn of antiquity. Starting with Plato, many theorists saw affect as a potentially dangerous, invasive force that subverts rational thinking. However, recent advances in social cognition, neuroanatomy, and psychophysiology suggest that affect is often a useful and even essential component of adaptive responding to social situations. The 1980s and 1990s produced something like an 'affective revolution' in psychology. Indeed, most of what we know about the role of affect in social thinking has been discovered since the early 1980s. As a result, we are now in a better position than ever before to pursue the age-old quest to understand the relationship between the rational and the emotional aspects of human nature. This article will consider early research on affect and social thinking, followed by a survey of contemporary theories and substantive research areas. The behavioral consequences of affect, and integrative theories of affect and social cognition, will also be discussed.

1. *Background*

Although affect has long been recognized as having a critical influence on thinking and judgments, empirical research on this link is quite recent. Psychologists traditionally studied the three fundamental faculties of the human mind—affect, cognition, and conation—as independent, isolated entities. Radical behaviorism consigned affect into the 'black box' of irrelevant mental phenomena, and the objective of most cognitive research was the study of cold, affectless ideation. The idea that affect is an integral and often adaptive part of how information about the social world is processed was not seriously entertained until the early 1980s. In a radical departure, Robert Zajonc argued that affective reactions often constitute the primary response to social stimuli and may influence subsequent attitudes and behaviors even in the absence of any cognitive memory trace (e.g., Zajonc 2000). Affective reactions such as feelings of anxiety, intimacy, or pleasure are also critical to cognitive representations of social experiences. More importantly, affect has a dynamic influence on memory, thinking, and judgments. Such 'affect infusion' was explained initially in terms of either psychodynamic or associationist principles.

Psychoanalytic theories suggested that affect has an invasive quality and can 'take over' thinking and judgments unless adequate psychological control is exerted. Attempts to suppress affect may increase the 'pressure' and could cause affect infusion into unrelated thoughts and judgments. In contrast, conditioning theories maintained that reactions to previously neutral social objects can be conditioned by simultaneous exposure to emotion-arousing stimuli. Consistent with this idea, investigators found that people who feel good or bad report significantly more negative or positive attitudes towards social stimuli. Byrne and Clore (1970) relied on this principle to explain the influence of affect on interpersonal attitudes and attraction.

2. *Contemporary Cognitive Theories of Affect and Social Cognition*

In contrast to psychoanalytic and conditioning accounts, contemporary cognitive theories focus on the information-processing mechanisms that link affect and cognition. Two kinds of mechanisms have been proposed: memory-based accounts (e.g., the affect-priming model; see Bower 1981, Forgas 1995), and inferential models (e.g., the affect-as-information model; see Clore et al. 1994). In addition, several theories highlight the influence of affect on information-processing strategies (Bless 2000, Fiedler 2000).

2.1 *Memory Models*

Affect may influence access to positively or negatively valenced ideas, producing more affect-congruent thoughts and judgments by happy or sad people. The associative network model proposed by Bower (1981) suggested that affect and cognition are integrally linked within an associative network of mental repre-

sentations. Affect will prime associated thoughts and ideas that are more likely to be used in constructive cognitive tasks. However, affect priming is subject to important boundary conditions. Mood-state dependence in memory is more likely when the affective state is strong, salient, and self-relevant, and the task involves the active generation and elaboration of information (Eich and Macauley 2000).

2.2 Inferential Models

Alternatively, rather than computing a judgment on the basis of recalled features of a target, individuals may simply ask themselves, 'How do I feel about it?' and in doing so, may use their mood to infer a reaction to the target. The 'how-do-I-feel-about-it' heuristic thus involves an inferential error: People misattribute their prevailing affect to an incorrect source. This theory is similar to conditioning approaches (Byrne and Clore 1970), as well as more recent work on misattribution, and research on judgmental heuristics. Accumulating evidence suggests that affect is most likely to be used as a heuristic cue when people are unfamiliar with the task, have no prior responses to fall back on, involvement is low, and cognitive resources are limited (Clore et al. 1994, Forgas 1995). Calling attention to the source of affect can reduce or even eliminate this effect. The affect-as-information model assumes that affect has invariant informational properties. Critics argue, however, that the informational value of affective states is not given but depends on the situation. The affect-as-information model has little to say about how affect and the stimulus information, and internal knowledge structures, are combined in producing a response. In that sense, this is more a theory of nonjudgment or aborted judgment rather than a theory of social judgment. Social cognitive tasks that require some degree of processing are more likely to be influenced by affect-priming rather than affect-as-information mechanisms.

2.3 Affect and Information Processing

In addition to informational effects (influencing *what* people think), affect can also influence the process of cognition, that is, *how* people think. It has been suggested that positive affect promotes less effortful and more superficial processing strategies, while negative affect triggers a more systematic and vigilant processing style. However, recent studies show that positive affect can also produce processing advantages. Happy people seem to adopt more creative and constructive thinking styles, use broader categories, show greater mental flexibility and perform better on secondary tasks (Bless 2000, Fiedler 2000). Explanations of mood effects on processing style at first emphasized the influence of affect on processing capacity. However, evidence for such accounts remains inconsistent. Alternative, motivational accounts suggest that people experiencing positive mood may try to maintain this pleasant state by refraining from any effortful activity. In contrast, negative affect may motivate people to engage in vigilant, effortful processing. Such a 'cognitive tuning' account seems consistent with evolutionary ideas about the adaptive functions of affect.

More recently, Bless (2000) and Fiedler (2000) argued that positive and negative affective states trigger equally effortful, but qualitatively different processing styles. Thus, positive affect promotes a more assimilative, schema-based, top-down processing style, where pre-existing ideas, attitudes and representations dominate information processing. In contrast, negative affect produces a more accommodative, bottom-up and externally focused processing strategy where attention to situational information drives thinking.

3. The Empirical Evidence

A growing number of empirical studies illustrate the role of affect in thinking and judgments in such areas as the perception of self and other people, the use of stereotypes, intergroup attitudes, attitude change, and persuasion.

3.1 Affect and Social Perception

Affect has a significant influence on how people respond to others (Forgas 2001). According to memory-based theories, affect infusion into social judgments is due to greater availability of affectively primed information when interpreting ambiguous social behaviors. Thus, a smile that may be seen as 'friendly' in a good mood could be judged as 'awkward' or 'condescending' by an observer experiencing negative affect. Surprisingly, these mood effects seem to be greater when people need to think more extensively about a more complex or unusual person, as confirmed by studies measuring reaction times and processing latencies. The more ambiguous the target, the more people will need to rely on their stored knowledge about the world to make sense of these stimuli. Such affect infusion effects also occur in realistic interpersonal judgments when people evaluate their real-life interpersonal relationships (Forgas 1995, 2001).

3.2 Affect and the Self

Affect has a particularly strong influence on self-related attitudes and judgments (Sedikides 1995): Positive affect improves, and negative affect impairs, the valence of self-conceptions. Thus, students in a negative mood tend to blame themselves more for failing, and those in a positive mood claim more credit for success in an exam (Forgas 1995). More recent

studies indicate that affect congruence in self-related thinking is subject to a number of boundary conditions. Affect infusion into self-judgments is more likely when the task requires open, constructive thinking. Compared with peripheral self-conceptions, central and familiar self-conceptions that require less elaboration also show reduced affect sensitivity (Sedikides 1995). Low self-esteem persons generally also have less certain and stable self-conceptions, and affect appears to have less of an influence on their self-judgments. Affect intensity and motivational factors also moderate mood congruency effects. Overall, affect seems to have a strong mood-congruent influence on many self-related attitudes and judgments, but only when some degree of open and constructive processing is required, and there are no motivational forces to override affect congruence.

Positive affect may also serve as a resource that allows people to overcome defensiveness and deal with potentially threatening information about themselves. People in a positive mood appear to be more willing to expose themselves voluntarily to threatening but diagnostic information. It seems that positive mood functions as a buffer, enabling people to handle the affective costs of receiving negative information, as long as the negative feedback is seen as useful (Trope et al. 2001).

3.3 *Affect and Stereotyping*

Early theories based on the frustration–aggression hypothesis and psychoanalytic notions of projection implied that negative affect may directly produce negative responses to disliked groups. Conditioning theories offered an alternative explanation of such effects, and also suggested that increased contact with outgroup members in positive situations should reduce aversive feelings and improve intergroup relations, a prediction that received considerable support (the so-called 'contact hypothesis'). Positive affect may also promote more inclusive cognitive categorizations of social groups, thus reducing intergroup distinctions (Bodenhausen and Moreno 2001). Individual differences such as trait anxiety can also significantly moderate the influence of negative affect on intergroup judgments; paradoxically, highly anxious people seem to show reduced intergroup discrimination when in an aversive mood. Affect is thus likely to influence intergroup judgments both by influencing the information processing strategies adopted, and the way positive and negative information about outgroups is selected and used.

3.4 *Affect in Persuasion and Attitude Change*

Writers on rhetoric have long assumed that producing an emotional response in an audience is an important prerequisite for effective persuasion. Recent dual-process theories of persuasion suggest that response to a persuasive message depends on the particular information-processing strategy people use, partly influenced by mood (Petty et al. 2001). For example, college students are significantly more likely to agree with persuasive messages advocating comprehensive exams when they are interviewed on a pleasant, sunny day rather than an unpleasant, rainy day. In general, affect should selectively prime mood-congruent thoughts and should influence responses to persuasion when a systematic, central-route strategy is adopted. However, people in a positive mood often process persuasive messages more superficially, and respond to both low-quality and high-quality messages the same way (Petty et al. 2001). Sad persons, in contrast, seem to be more persuaded by high-quality messages, consistent with their more externally focused and bottom-up information-processing style (Bless 2000).

3.5 *Affect and Cognitive Dissonance*

The experience of dissonance between attitudes and behaviors is a potent mechanism that produces attitude change. Cognitive dissonance involves negative affect because discrepancy among cognitions undermines our clear and certain knowledge about the world, and thus our ability to engage in effective action (Harmon-Jones 2001). Studies suggest that positive affect decreases, and negative affect increases, dissonance reduction and attitude change. Once consonance is restored, affective state also tends to improve. However, much work remains to be done in discovering the precise cognitive mechanisms responsible for these effects.

3.6 *Affect, Cognition and Social Behavior*

Affect can influence not only social cognition but also subsequent social behaviors. People need to evaluate and plan their behaviors in inherently complex and uncertain social situations. Positive affect may thus prime positive information and produce more confident, friendly, and cooperative 'approach' attitudes and behaviors, whereas negative affect should prime negative memories and produce avoidant, defensive, or unfriendly attitudes and behaviors. Research shows that people in a happy mood respond more positively when approached with an impromptu request, and produce more confident and more direct requests themselves in a variety of social situations. These effects also influence real-life interpersonal behaviors. Further, affect infusion into attitudes and behaviors is greater when more elaborate, substantive processing is required to produce a strategic response in more complex social situations. Affective states also play an important role in elaborately planned interpersonal encounters such as the planning and performance of complex negotiating encounters (Forgas 2001). Why do these effects occur? When people face an

uncertain and unpredictable social encounter, such as a negotiating task, they need to rely on open, constructive processing in order to formulate new plans to guide their interpersonal behaviors. Affect can selectively prime more affect-congruent thoughts and associations, and these ideas will ultimately influence people's plans and behaviors. However, affective influences on social behaviors are highly process-dependent: affect infusion is increased or reduced depending on just how much open, constructive processing is required to deal with a more or less demanding interpersonal task.

3.7 Integrative Theories: The Affect Infusion Model

Affective states thus have powerful informational and processing effects on the way people perceive, interpret, and represent social information. It is also clear, however, that affective influences on social cognition are highly context-sensitive. A recent integrative theory, the Affect Infusion Model (AIM; Forgas 1995) argued that affect infusion should only occur in circumstances that promote an open, constructive processing style. Constructive or generative processing may be defined here as genuinely open-ended, generative thinking that involves the selective search for, and on-line elaboration of, external information, and the constructive combination of stimulus details with internal knowledge structures that results in the creation of a new response. In contrast, merely reconstructive processing involves the retrieval and use of a previously computed response, with no on-line elaboration or stimulus details (Fiedler 2000).

The AIM identifies four alternative processing strategies: direct access, motivated, heuristic, and substantive processing. These four strategies differ in terms of two basic dimensions: the degree of effort exerted, and the degree of openness or constructiveness of the information search strategy. The combination of these two processing features, quantity (effort), and quality (openness), produces four distinct processing styles: substantive processing (high effort/open, constructive), motivated processing (high effort/closed), heuristic processing (low effort/open, constructive), and direct access processing (low effort/closed).

Direct access processing occurs when a pre-existing, stored response is directly accessed and used without any on-line elaboration. Motivated processing is driven by a strong pre-existing motivational objective that determines information search strategies and thus limits the likelihood that incidentally primed, affective information will be incorporated into the response. Heuristic processing, although a constructive process, is based on the limited consideration of information, and the use of mental shortcuts and heuristics (such as the affect-as-information heuristics) whenever possible. It is only substantive processing that involves the full and elaborate processing of stimulus information and the constructive integration of stimulus details with primed memories and associations. According to the AIM, mood congruence and affect infusion is most likely when constructive processing is used, such as substantive or heuristic processing. In contrast, affect is unlikely to affect the outcome of closed, merely reconstructive tasks involving motivated or direct access processing (see also Fiedler 2000).

4. Conclusion

The evidence shows that mild everyday affective states can have a highly significant influence on the way people perceive and interpret social situations. Further, several of the experiments discussed here show that different information-processing strategies play a key role in moderating these effects. Affect infusion influences not only attitudes and judgments but also interpersonal behaviors, such as the monitoring and interpretation of observed encounters; the formulation of, and responses to, requests; and the planning and execution of strategic negotiations. In contrast, affect infusion is absent whenever a social cognitive task could be performed using a simple, well-rehearsed direct access strategy, or a highly motivated strategy. Obviously a great deal more research is needed before we can fully understand the multiple influences that affect has on attitudes, judgments, and interpersonal behavior.

See also: Adulthood: Emotional Development; Automaticity of Action, Psychology of; Emotion in Cognition; Social Cognitive Theory and Clinical Psychology

Bibliography

Bless H 2000 The interplay of affect and cognition. In: Forgas J P (ed.) *Feeling and Thinking: The Role of Affect in Social Cognition* (pp. 201–22). Cambridge University Press, New York

Bodenhausen G V, Moreno K N 2001 How do I feel about them? The role of affective reactions in intergroup perception. In: Bless H, Forgas J P (eds.) *The Role of Subjective States in Social Cognition and Behavior*. Psychology Press, Philadelphia, pp. 283–303

Bower G H 1981 Mood and memory. *American Psychologist* **36**: 129–48

Byrne D, Clore G L 1970 A reinforcement model of evaluation responses. *Personality: An International Journal* **1**: 103–28

Clore G L, Schwarz N, Conway M 1994 Affective causes and consequences of social information processing. In: Wyer R S, Srull T K (eds.) *Handbook of Social Cognition* (2nd edn.). Erlbaum, Hillsdale, NJ, pp. 323–417

Eich E, Macauley D 2000 Fundamental factors in mood-dependent memory. In: Forgas J P (ed.) *Feeling and Thinking: The Role of Affect in Social Cognition* (pp. 109–30). Cambridge University Press, New York

Fiedler K 2000 Towards an integrative account of affect and cognition phenomena using the BIAS computer algorithm. In:

Forgas J P (ed.) *Feeling and Thinking: The Role of Affect in Social Cognition*. Cambridge University Press, New York, pp. 223–52
Forgas J P 1995 Mood and judgment: The affect infusion model. *Psychological Bulletin* **117**: 39–66
Forgas J P (ed.) 2000 *Feeling and Thinking: The Role of Affect in Social Cognition*. Cambridge University Press, New York
Forgas J P (ed.) 2001 *The Handbook of Affect and Social Cognition*. Erlbaum, Mahwah, NJ
Harmon-Jones E 2001 The role of affect in cognitive dissonance processes. In: Forgas J P (ed.) *The Handbook of Affect and Social Cognition*. Erlbaum, Mahwah, NJ, pp. 237–56
Petty R E, DeSteno D, Rucker D 2000 The role of affect in attitude change. In: Forgas J P (ed.) *The Handbook of Affect and Social Cognition*. Erlbaum, Mahwah, NJ
Sedikides C 1995 Central and peripheral self-conceptions are differentially influenced by mood: Tests of the differential sensitivity hypothesis. *Journal of Personality and Social Psychology* **69**, 4: 759–77
Trope Y, Ferguson M, Ragunathon R 2001 Mood as a resource in processing self-relevant information. In: Forgas J P (ed.) *The Handbook of Affect and Social Cognition*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 257–89
Zajonc R B 2000 Feeling and thinking: Closing the debate over the independence of affect. In: Forgas J P (ed.) *Feeling and Thinking: The Role of Affect in Social Cognition* (pp. 31–58). Cambridge University Press, New York

J. P. Forgas

Social Cognition and Aging

There is a rich and abundant history of research in cognitive aging. However, studies typically have ignored the influence of social factors on cognition with one exception, metacognition. More specifically, there has been considerable research on self and other beliefs about memory such as personal control and memory self-efficacy. In the 1990s social cognition and aging research has grown considerably in both breadth and scope. This movement reflects a broader interface between mainstream social cognition and cognitive aging to better understand how cognition in the aging adult operates in a social context.

There are pervasive findings in the cognitive aging research that demonstrate age-related decline in cognitive performance such as working memory, abstract reasoning, and speed of processing. However, taking into consideration the pragmatics of cognition and social-contextual factors may help broaden our understanding of cognitive functioning in older adults beyond linear decline. Accordingly, a recent wave of studies has begun to examine social cognitive functioning in an effort to better understand the effects of cognitive aging and associated adaptive developmental processes. These studies raise important issues in relation to cognitive functioning such as age-related increases in experience, pragmatic knowledge, social expertise, beliefs about aging and social values, and social contextual variation. In addition, they highlight the importance of considering both the basic cognitive architecture of the aging adult and the functional architecture of everyday cognition in a social context. In the latter case, even if certain basic cognitive mechanisms have declined, older adults may possess appropriate knowledge and skills in a given context to function adequately through adaptation and compensation. Before we entertain these issues and findings in the extant literature, let us first consider a definition of the social cognitive perspective as it is applied to aging research.

1. *Definition of Social Cognition*

The basic goal of a social cognitive perspective is to understand how individuals make sense of themselves, others, and events in everyday life. These goals are reflected in the study of (a) cognitive mechanisms that may explain social cognitive processes (as illustrated above); (b) the content and structure of social knowledge; and (c) the influence of social situational and individual demands on social information processes and social knowledge.

To illustrate these points within an adult development and aging perspective, first consider the well-documented finding that age-related decline in processing capacity correlates with age-related decline in cognitive performance. The question is whether age-related changes in working memory, attentional capacity, and speed of processing also relate to and/or explain age changes in social cognitive processes (Hess 1999, Schwarz et al. 1999). Researchers have suggested that age-related decline in processing capacity could explain the finding that older adults rely more on easily accessible trait-based information when making social judgments as opposed to more elaborative processing (Hess 1999).

However, from the second goal perspective listed above, age-related change in content and structure of social knowledge, accessibility of this knowledge held in memory, and related motivation and goals of the individual in the context of concurrent processing resources may offer competing explanations of age differences in social reasoning. Finally, personal and situational factors associated with aging could also explain a reliance on easily accessible social knowledge and beliefs. For example, older adulthood has been associated with a decrease in personal characteristics such as tolerance for ambiguity and attitudinal flexibility (Blanchard-Fields 1999). It may be the case that such decreases in cognitive flexibility may lead to social processing biases (Hess 1999). Alternatively, older adulthood is associated with changing situational factors, which influence social goals such as an increase in the importance of emotional considerations. In this case such an increase in emotional

salience influences how information is processed (see Turk-Charles and Carstensen 1999 for a review). Let us now turn to each of these goal perspectives in more detail.

2. Cognitive Mechanisms and Social Cognition

There is a history of social cognition research that posits information-processing models for making social judgments (e.g., Gilbert and Malone 1995). For example, Gilbert and Malone have shown that the ability to make nonbiased attributional judgments depends upon the cognitive demand accompanying those judgments. In this case individuals are more likely to rely on their initial dispositional attribution (attributing the cause of an outcome to personal characteristics of the main character involved in producing the outcome), ignoring compelling situational causes. Accordingly, in a number of studies, Blanchard-Fields and co-workers (see Blanchard-Fields 1999 for a review) have found that older adults consistently attributed the cause of problem dilemmas to dispositions about the primary protagonist more than younger age groups. However, this was only demonstrated when events reflected negative interpersonal content. It may be the case that older adults relied more heavily on easily accessible dispositional attributions given the demands on cognitive resources to deliberate about situational factors. However, keep in mind that these findings were content-specific (i.e., only found when situations were negative and interpersonally oriented).

Hess and co-workers (see Hess 1999 for a review) have demonstrated that the use of diagnostic trait information (such as immoral behavior) in making initial impressions of an individual varied with age. They find age differences in the use of negative information in modifying impressions. When first presented with a positive portrayal of an individual (e.g., honest) followed by negative information (e.g., dishonest behavior) older adults were more willing to modify their impression from an initially positive to a more negative one. However, older adults did not modify their first impression when it was an initially negative portrayal followed by positive information. Younger adults did not exhibit this pattern of impression change. Instead, they engaged in effortful processing of the text in their concern with the situational consistency of the new information presented. Hess suggests that older adults may rely more on life experiences and social rules of behavior in interpreting these types of situations, whereas young adults were more concerned with situational consistency of the new information presented. Again, it is suggested that older adults do not correct their initial impressions because they were affected by the availability of negative information or a negativity bias. Such influences of processing biases suggest that decline in cognitive functioning limits the ability of older adults to override the impact of more easily accessible information such as the negativity bias. However, Hess also suggests that the use of such information appears to be related to age differences in social experience and changes in the salience of specific behavioral realms.

Finally, another example of a processing resource approach to social cognition and aging is illustrated in the work of Schwarz et al. (1999). They found that manipulating the salience of the information given, and thus its accessibility, can modify social judgments in young and older adults. However, in some circumstances, older adults' judgments show smaller context effects than young adults. One explanation of such findings is that reduced processing resources constrain the processing of contextual information by older adults.

Based on the above research, it appears that both processing resource limitations as well as the social content of memory play an important role in understanding how older adults process social information. However, it is also the case that the extent to which social information is accessible may operate independently of a processing resource limitation to influence social judgments. Next we will explore the issue of knowledge-based views of social cognitive processing.

3. Social Knowledge and Social Cognition

There is a history of social cognition research that suggests that strongly held and easily accessible social beliefs and knowledge are activated automatically and can influence social judgments in general, and promote social judgment biases in particular (e.g., Gilbert and Malone 1995). The prepotent influence of such beliefs and knowledge on social judgments may operate even when individuals are not constrained by high cognitive demands and possess ample resources to engage in elaborative processing. This brings into play alternative explanations for older adults' tendencies to rely on easily accessible information in making social judgments. First, as alluded to earlier, older adults may exhibit a cognitive style that forecloses further consideration of information such as intolerance for ambiguity or a need for structure. As indicated above, age-related decreases in attitudinal flexibility may lead to biases in social judgment. Second, strong attitudinal beliefs and values may be evoked by the presenting situation. For example, Blanchard-Fields (1999) suggests that particular types of situations such as negative interpersonal relationship situations trigger relatively automatic trait- and rule-based social rules or schemas pertinent to the actor in the problem. In this case the individual may not engage in intentional, effortful processing of detailed information. Instead, the older adults may have relied more on activated social beliefs resulting in dispositional biases. Thus, it is not a matter of insufficient processing capacity to

engage in elaborative social inference, but overriding considerations (e.g., irrepressible beliefs or values).

Accordingly, Chen and Blanchard-Fields (reviewed in Blanchard-Fields 1999) explored both a processing explanation and a belief content hypothesis to explain dispositional biases in older adults' attributions about negative relationship situations. They manipulated the timing of when dispositional ratings were collected and obtained belief statements regarding appropriate behaviors in the social situations presented. They found that older adults produced more dispositional biases only when time was constrained. However, they also found that older adults made more evaluative rule statements about the main character than younger adults and this rendered all age differences in dispositional biases nonsignificant.

Such findings suggest that a knowledge-based explanation of social judgments is a viable alternative to resource limitations. This approach is also evidenced in the research examining the influence of social situational factors and cognitive functioning in older adults. In this case, the research focus is on one's changing life context and adaptation or social competence. We turn to this approach next.

4. *Social Situational Factors, Social Competency, and Social Cognition and Aging*

The social cognition paradigm has offered an enriched understanding of competency in older adulthood. From this perspective, social competence is seen as an important and valid dimension of cognition and intelligence. Social cognitive performance, from this perspective, incorporates principles derived from the lifespan contextual perspective and can be explained in terms of adaptation to particular life contexts. For example, research exploring such social cognitive constructs as possible selves (one's hopes and fears for the future) indicates that they change throughout the latter half of the lifespan (Hooker 1999). For example, health-related selves achieve greater importance in older adulthood and relate to adaptive self-regulatory health behaviors.

This is most evident in areas such as wisdom, practical and social problem-solving, and postformal reasoning. For example, Staudinger et al. (1992) argue that wisdom involves an expert body of knowledge and strategies related to life composition and life insight. These strategies are used to produce adaptive outcomes in managing and composing one's life. Examples of such knowledge and strategies include the ability to select appropriate abilities and strategies, compensate for declining abilities, and optimize one's performance, e.g., as in increased flexibility in coping behavior. Similarly, Turk-Charles and Carstensen's (1999) work suggests that changes in the relative importance of social cognitive goals across the lifespan profoundly influence the interpretation of social information and the selection and maintenance of intimate relationships. In this latter case, important new variables affecting social information processing emerge. Carstensen and co-workers argue that perception of time is responsible for this change in social motivation. By considering such changes, age-related decreases in social interaction can be recast in terms of its adaptive significance.

In another domain, Dixon (1999) examines cognition in social collaborative contexts. The primary thrust of this work examines age differences in the benefits or costs of collaboration in cognitive performance (e.g., memory, problem solving). He argues that collaborative cognitive functioning characterizes another aspect of social competence. Finally, we cannot ignore the negative role that social situational factors can play in the social cognitive functioning of older adults. This is reflected in the work on social stereotypes and aging. A number of researchers have examined adult developmental changes in the content and structure of stereotypes (e.g., Hummert 1999). For example, Hummert finds that older and young adults hold quite similar age stereotypes. However, she and other researchers find that older persons have more complex representations of the category 'older adult' as opposed to young adults who, in addition, hold more negative representations. However, unlike the social cognitive research reviewed above, these studies primarily look at age differences in the content of stereotypes. The explicit social cognitive mechanisms that account for these differences are not revealed clearly.

Relative to the knowledge-based framework discussed above, the content of social knowledge, attitudes, and beliefs such as stereotypes is important. However, the content must be accessed or activated in order to influence behavior. An important question is how does stereotype activation encourage or discourage the individual to be a more effective information processor? With respect to older adults, how do the stereotypes of aging provide expectations of mental deficiencies? In her research Levy (1996) suggests that such automatically activated (that is subliminally primed) stereotypes about aging and memory adversely affect the cognitive performance of older adults. This is, of course, a provocative finding and in need of replication. However, the point here is that it provokes issues for further research. The issue of directly manipulating threatening and/or facilitative and nonfacilitative social contexts is only beginning to undergo examination.

5. *Future Directions*

In conclusion, the social cognition and aging research highlights the importance of considering social factors in explanations of social cognitive functioning. Such factors as the communicative context, social beliefs

and values, cognitive style, stereotypes, motivational goals, among others influence social information processing in important ways. Thus, it is important not to limit explanations of social cognitive change to cognitive processing variables. There are important social factors influencing how and when older adults will attend to specific information and when this information will influence social cognitive functioning.

However, current research in social cognition and aging has only begun to address important questions regarding older adults' social cognitive functioning. In order to understand more fully the role and interplay of social cognitive structures and mechanisms in social cognition and aging, more research is needed. For example, are there qualitative changes in underlying mechanisms influencing social cognitive functioning? Age differences in the nature of social beliefs, values, and goals activated in social reasoning need to be explored. How and under what conditions do age differences in the content and activation of social knowledge and social belief systems account for social information processing? Can factors that influence variability as to when social judgment biases occur in young adults account for age-related variability? Under what conditions does working memory load compromise social information processing and under what conditions is it irrelevant? We can be optimistic in the growth of the social cognitive approach in research on aging. Although many of the cognitive and socio-emotional aspects of behavior and aging have been addressed, we have only begun to scratch the surface.

See also: Adult Cognitive Development: Post-Piagetian Perspectives; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Cognitive Aging; Cognitive Development in Childhood and Adolescence; Lifespan Development, Theory of; Lifespan Theories of Cognitive Development; Social Learning, Cognition, and Personality Development; Wisdom, Psychology of

Bibliography

Blanchard-Fields F 1999 Social schematicity and causal attributions. In: Hess T M, Blanchard-Fields F (eds.) *Social Cognition and Aging*. Academic Press, San Diego, CA

Dixon R 1999 Exploring cognition in interactive situations: The aging of N + 1 minds. In: Hess T M, Blanchard-Fields F (eds.) *Social Cognition and Aging*. Academic Press, San Diego, CA

Gilbert D T, Malone P S 1995 The correspondence bias. *Psychological Bulletin* **117**: 21–38

Hess T M 1999 Cognitive and knowledge-based influences on social representations. In: Hess T M, Blanchard-Fields F (eds.) *Social Cognition and Aging*. Academic Press, San Diego, CA

Hooker K 1999 Possible selves in adulthood: Incorporating teleonomic relevance into studies of the self. In: Hess T M, Blanchard-Fields F (eds.) *Social Cognition and Aging*. Academic Press, San Diego, CA

Hummert M L 1999 A social cognitive perspective on age stereotypes. In: Hess T M, Blanchard-Fields F (eds.) *Social Cognition and Aging*. Academic Press, San Diego, CA

Levy B 1996 Improving memory in old age through implicit stereotyping. *Journal of Personality and Social Psychology* **71**: 1092–107

Schwarz N, Park D, Knauper B, Sudman S 1999 *Aging, Cognition, and Self-reports*. Psychology Press, Philadelphia, PA

Staudinger U M, Smith J, Baltes P B 1992 Wisdom-related knowledge in a life review task: Age differences and the role of professional specialization. *Psychology and Aging* **7**: 271–81

Turk-Charles S, Carstensen L L 1999 The role of time in the setting of social goals across the life span. In: Hess T M, Blanchard-Fields F (eds.) *Social Cognition and Aging*. Academic Press, San Diego, CA

F. Blanchard-Fields

Social Cognition in Childhood

'Social cognition' refers in very general terms to thinking about people, knowledge of people, and the social world, as well as to social processes in cognitive development. That is, it is a term that is used to refer both to the nature or content of children's understanding of others and themselves, and to theories and explanations of how social experiences are involved in the development of understanding more broadly considered—the role of social factors in cognitive growth. The study of social cognition then is at the border of cognitive and social developmental psychology (Butterworth 1982). It brings together very different theories of the growth of knowledge, and is a growing point for new ideas on the nature of children's understanding of mind and emotion, and on the relation of this understanding to children's prosocial and moral development, and their social relationships.

1. Changes in Ideas on Social Cognition

Five themes in the changing ideas on the development of social cognition over recent decades are evident. One concerns the question of how social experiences are implicated in sociocognitive development, in terms of the processes by which cognitive change takes place. The second concerns ideas and evidence on the nature of children's understanding of emotion, mind, intention, and the connections between these inner states and people's actions—a domain which has come to dominate cognitive developmental research since the early 1990s. The third theme is the rapidly growing interest in individual differences in social understand-

ing, and the implications of these differences for a wide range of children's outcomes. The fourth is the role of social experiences in contributing to these individual differences in children's social understanding. The fifth is the recent interest in bridging the gap between the study of cognitive and of social development through a focus on how cognitive development is implicated in social relationships, moral and prosocial development. These themes will be considered in turn.

1.1 Individual and Social Approaches to Cognitive Development

Piaget's ideas on cognitive development centered on both the process of understanding, and on the nature of knowledge, and a key notion was that the individual child actively constructs knowledge by selecting and interpreting information in the environment. The processes he proposed for the growth of intelligence were seen as applying to both the inanimate and the social world. His investigations of children's understanding of the mind began in the 1920s (e.g., Piaget 1926); after these first books he focused chiefly upon children's growing knowledge of the physical world, and the development of logicomathematical thought. The emergence of social knowledge was not seen as essentially different from other forms of knowledge, though social interactions (especially those with peers) were seen as contributing to children's growing understanding (see Sect. 1.4 below).

In contrast to Piaget's focus on the individual actively constructing knowledge as an individual, there is a strong tradition within psychology that emphasizes the social nature of cognitive processes, beginning as early as the 1890s with Baldwin's seminal writings (e.g., Baldwin 1895). In Russia, Vygotsky argued that cognitive growth was socially determined—that cognitive development proceeded from the social plane to the individual (see Wertsch 1985). He famously proposed that new cognitive abilities were first seen when a child was interacting with another person, and only later became internalized. Individual development was, he argued, embedded in a social context. Vygotsky emphasized the difference between a child's level of potential ability, and his or her actual ability.

The difference between these two levels he termed the 'zone of proximal development'—abilities that were just beyond what a child could achieve alone, but was able to achieve when interacting with a cooperative, supportive other person (adult or peer). Thus to Vygotsky, the scaffolding or facilitation provided in social interaction (especially interaction between adult and child, or expert and novice) was key for the development of new abilities. Within this broad theoretical framework, researchers have examined the significance of 'scaffolding' and facilitation in cognitive performance, and the 'expert–novice' interaction (Rogoff 1990). A whole new view of cognitive growth in infancy has flourished, in which the context in which new abilities are fostered is seen as essentially social (Hala 1997). So for example, recent research on the development of understanding of emotion takes off from Darwin's observations of the development of emotions (which highlight the biological foundation and function of emotional expression) to emphasize the sociocultural contributions to the development of both expression and understanding of emotions.

The idea that cognitive growth takes place through social interaction has also been experimentally investigated within a different research tradition, that of Doise and his colleagues in Geneva, with a series of studies that began in the 1970s (e.g., Doise and Mugny 1984). Rather than emphasizing scaffolding and supportive cooperation, they argued that cognitive conflict between individuals who may view a problem differently, or have different levels of ability can stimulate cognitive development.

These theoretical positions have all been concerned with the normative processes by which cognitive changes take place, rather than with the role of social experiences as contributors to individual differences in social cognition; this is considered in Sect. 1.4 below.

1.2 The Nature of Children's Developing Social Understanding

After Piaget's seminal work in the 1920s, it was not until the mid-1980s that interest in children's understanding of mental life, and their perspective on why people behave the way they do, began to revive; since the early 1990s there has been a great increase in developmental research on children's understanding of emotion, mind, and intention. The focus has been chiefly on normative developmental changes in children's understanding of emotional and mental life and their relation to people's behavior, beginning with the work on infancy that highlights the extent to which babies are 'prepared to learn about people' from very early in infancy. For example, babies discriminate different emotional expressions very early (a key first step for understanding emotions); by the time they are nine months old they already act in ways that indicate they have considerable grasp of affective communication, and are able to communicate their desires—clearly, these are efforts at symbolic communication. At the end of the first year, the emergence of 'social referencing' is a key development: in situations of ambiguity for the baby (such as when a strange person or object appears), the baby looks at a familiar adult (in most experimental settings, the mother), and apparently is guided by her emotional expression, showing different behavior if she is expressing happiness or fear, for instance. Such differential behavior is taken as evidence that the baby 'understands' that

the mother's emotional expression provides important information about the ambiguous situation.

During the second year, babies begin to show pride, shame, and embarrassment—the 'social emotions,' emotions that highlight how closely tuned the children are to other people's reaction to them. They also clearly show their understanding of others' feelings in their attempts to comfort, tease, and joke, and, with the acquisition of language, in their talk about feelings (Bartsch and Wellman 1995).

The marked changes between two and five years old in children's ability to distinguish thoughts and things; their increasing interest in talking about feeling, thinking, and knowing; their curiosity about why people behave the way they do; their ability to play with hypothetical social events, to share narratives in pretend play, and to deliberately deceive others, have been documented in a burst of research since the early 1990s (Astington 1993). There has been considerable controversy about the details of the timing of these normative changes (conclusions about children's abilities depend very much on the methods used and the setting in which children are studied; see Sect. 2), and competing theoretical accounts of these cognitive changes have been articulated. However the broad outline of development in what has been termed children's 'theory of mind' has been mapped, as follows.

Children's understanding of people's feelings and mental states, and their use of this understanding to explain and predict people's behavior begins in infancy, as the evidence of their intention to communicate shows. It develops in the toddler years with their growing ability to talk first about what they and other people see, want, and feel, then later in their talk about what they think and know. The critical understanding that people behave in the world as they see it or believe it to be, even when those beliefs or perceptions may be inaccurate, is reached by most children around four years. Note that much of the most recent research emphasizes the substantial competence of children as young as three, and the gradual nature of developmental change. A key tool in this research has been the demonstration that children understand how a 'false belief' can influence how someone will behave.

Other key developments in children's social understanding include children's growing understanding that people are different, their grasp of understanding the self, and their moral understanding. On the first of these, children's understanding and attribution of traits, a clear developmental pattern was described from work in the 1970s and 1980s, with marked change around six to eight years in children's appreciation of trait terms; it is usually presumed that this reflects age-related cognitive changes. Recently the 'theory of mind' research has been integrated with this 'traditional' approach, and there is increasing empirical support for the idea that both cognitive development and sociocultural influence play a role in the development of trait attribution (Hala 1997).

On the development of moral understanding, there has been a parallel shift from a focus primarily on the cognitive underpinnings of moral development (e.g., Kohlberg 1984) to a recognition that cognition, emotion, and moral action are closely interrelated (Eisenberg et al. 1997). Thus while we now know that developmental changes in prosocial behavior are complex, and influenced by the methodology employed (Eisenberg and Fabes 1998), it is recognized that the relations between perspective-taking abilities and prosocial behavior are moderated by affective factors, by the particular emotional relationship between the child and the other person involved, and by the child's interpretation of the situation, and attributions about the individuals involved.

1.3 Individual Differences in Social Cognition

Individual differences in sociocognitive abilities have been shown to be importantly linked to children's adjustment and 'social competence' (skills, deficits) in middle childhood; the wealth of research on social cognition and children's peer relations—and specifically to problems in those relations—illustrates these associations. One line of research has focused on information-processing models (as in Dodge's 1986 social information-processing model), which highlight children's attributions about their peers, their goals, and strategies in peer relations. These models conceptualize an ordered series of steps in social information processing that includes encoding and representational processes, search and decision-making strategies. Differences in processing patterns may underlie the form of aggressive behavior that children show towards their peers. Experimental studies with aggressive children show, for instance, that they did not attend to all relevant cues, they attributed hostile intent to a protagonist in an ambiguous setting, and failed to generate enough viable alternative strategies. The original social information-processing models were criticized on the grounds that the linear model was too constrained, and that no role of emotion was included. Revised models now include emotional states as influences on what is perceived, how it is processed, and what attributions are made.

A second line of evidence links children's thoughts about themselves, such as their perceptions of their own competence, to individual differences in peer relations (Ladd 1999). Thus children who are both aggressive and rejected by their peers, compared to other children, overestimate their own social skills and competence, and underestimate the degree to which their peers dislike them. Other research links individual differences in early social understanding to later differences in moral sensibility, and to the ways in which children adjust to school. Such findings high-

light the importance of clarifying what are the key processes implicated in the development of such individual differences in social understanding.

1.4 Social Experience as Contributor to Individual Differences in Social Understanding

To understand these individual differences in children's social understanding, it is increasingly clear that their social experiences must be examined. Research that includes a focus on social experiences not only reveals children's abilities vividly, but also suggests some key processes implicated in their development (Dunn 1996). The interactive events that have been highlighted include sharing a pretend world with another person (especially a peer or sibling), participation in conversations about inner states (Bartsch and Wellman 1995), joint narratives and causal discourse about the social world, engagement in certain kinds of 'constructive' conflict with others. For example, several studies have shown that children who grew up in families in which they participated in discourse about feelings, and causal talk about why people behave the way they do, were, in later assessment of social understanding, more successful than those who had not engaged in such talk; other research has shown that those who had experienced much shared, cooperative pretend play with a sibling were also in later assessments particularly mature in their social understanding.

Three general principles that emerge from this empirical research are the following. First, the emotional context in which such interactions take place may be key to what is learned; it is not simply the child's exposure to new information or another's point of view that matters, but the emotions and pragmatics of the relationship between the interactants that are important. Second, how children use their social understanding in real-life interactions differs in different social relationships, depending on the quality of that relationship. Relationship quality depends on both partners, and is unlikely to be linked simply to the sociocognitive skills of either partner. Third, different kinds of social relationship appear to play different roles in influencing the development of social understanding.

Two recent lines of evidence illustrate this point. First, studies of child–child relationships show these peer interactions to be of special significance in the development of understanding feelings and other minds, and of moral sensibility (as originally proposed by Piaget [1932] 1965). Second, recent work on attachment quality links the security of children's attachments to their later mind-reading abilities—and it is inferred that the key process implicated here is mothers' sensitivity to their children's current levels of understanding and their use of mental state terms in interactions with them, a propensity that Meins (1997) refers to as 'mind-mindedness.' The role of family processes, more broadly conceived, in influencing children's social cognitive abilities during early and middle childhood—both directly and indirectly—continues to be a focus of much research interest. It should be noted that inferences about the direction of effects in such research present particular problems.

1.5 The Significance of Cognitive Processes in Social Development

The advances in our understanding of children's cognitive development have important implications for how we view and study children's social development and their social relationships—both the nature of social development and the processes implicated in it. First, what has been learned from cognitive research about the development of understanding of emotion and mind in childhood, means not only that dimensions of empathy and understanding are part of their early close relationships, but that aspects of intimacy (previously thought only relevant to the relations between much older children) may also be present in the relationships between preschoolers. Thus observational research shows that three year olds who are close friends engage in shared imaginative play in which their cooperation, the smooth co-construction of narratives, reflect an impressive 'meeting of minds'—a mutual understanding of how the other thinks the play might develop, and what are the exciting next moves in the narrative. Second, the cognitive research alerts us to the importance of looking beyond the 'security' aspects that have dominated conceptions of parent–child relationships. Manipulation of control and power within a dyadic relationship, deception, teasing, and shared humor all depend on a grasp of the other's motivation, beliefs, desires, and capacities. These dimensions may well be relevant for a full account of young children's relationships. Third, to understand developmental changes in children's social relationships, the evidence for developments in metacognitive skills is clearly relevant. Perhaps most important of all are the implications of cognitive research for children with difficulties in social relations, for example children with problems such as autism or hyperactivity. Whether these children's social problems are antecedent to or sequels of their cognitive problems remains a question of central social and clinical importance, as well as one of theoretical interest.

Two other lessons from cognitive research have major implications for the study of children's sociocognitive development. The first concerns the developmental impact of relationships between other people. Cognitive research has shown us that children are interested in, and reflect on, others' intentions,

thoughts, and feelings, reminding us to pay attention to the developmental impact of what is happening between the other people in children's social worlds. Three lines of evidence illustrate this significance: the evidence for the negative impact of marital conflict on children's adjustment; the research on co-parenting, which shows the poor outcome of children whose parents disagree and fail to support one another over parenting; and third, the accumulating evidence that differential treatment of siblings within the family is associated with later problems in children's development.

The second lesson is that the evidence for children's interest in the inner states of others suggests that children may be sensitive to the responses, opinions, and relationships between other members of their peer groups to an extent that is only recently being appreciated. Recognizing that children's problems in peer group relations are a key marker for later adjustment problems suggests that further attention to this sensitivity to the group is important.

2. *Methodological Issues and Problems*

In research on social cognition there is a new concern with the contextual validity of experimental tasks used, particularly in terms of the social situation in which children are placed in such tasks; their relationship with the experimenter plays a role in their task performance (Dunn 1996). More generally there is a concern to expand the social context of such studies to more 'real-life settings': research within a Vygotskean framework, for example, has chiefly been limited to adult–child encounters of a didactic nature, which are affectively neutral—unlike the interactional settings within the family in which children are likely to make advances in sociocognitive ability (Goodnow 1990).

Research on children in real-life social settings—as opposed to controlled experimental settings—highlights the capacities that even young children can show in the domain of understanding other people's feelings, beliefs, and intentions. Studies of natural discourse have been centrally important here (Bartsch and Wellman 1995). There is a tension here in current developmental research, as the naturalistic data suggest children's capacities for social understanding are evident at considerably younger ages than would be expected on the basis of the standard experimental assessments of sociocognitive development; the various studies of deception provide a good example (Dunn 1996). Such studies not only show that children's abilities may be misrepresented by their limited abilities in more formal settings; they have also contributed importantly to ideas on the social processes that may influence the development of individual differences in social cognition.

Different methodologies can importantly affect the inferences we make about children's prosocial behavior (Eisenberg and Fabes 1998), as well as about their mind-reading and emotion understanding. Moreover, children use their sociocognitive abilities differently within their various relationships: one lesson from research that focuses on children in different relationships is that we should move away from the conception of sociocognitive capacity as a characteristic of a child that will be evident in all social situations (Dunn 1996).

3. *Future Directions*

Among the many challenges posed by the research to date, four stand out as key. First, to make progress in understanding the development of social understanding, we need to address the difficult issue of how affect is linked to cognitive change, and to bridge the divide between research on social and cognitive development. The developmental work that documents children's explanations of people's actions first in terms of emotions and desires, and then later incorporating the notion of belief, has led for instance to the provocative argument that an understanding of cognitive states arises from an earlier understanding of emotional states (Bartsch and Estes 1996).

Second, we need to investigate the differentiation of different aspects of social understanding—too often combined into global categories at present—and to further explore the relation of language and general intelligence to the various aspects of understanding the social world.

Third, the challenging question of how developmental accounts based on children within the normal range of individual differences relate to the development of children at the extremes remains both theoretically and clinically important. Fourth, more research on implications of cultural and class differences, and more diversity in the children studied is needed. Finally at a methodological level, researchers need to include both experimental and naturalistic strategies, and to study children in different social settings, to use as informants both children and others in their social worlds.

See also: Cognitive Development: Child Education; Family Bargaining; Infancy and Childhood: Emotional Development; Piaget's Theory of Child Development; Social Learning, Cognition, and Personality Development

Bibliography

Astington J W 1993 *The Child's Discovery of the Mind*. Harvard University Press, Cambridge, MA

Baldwin J M 1895 *Mental Development in the Child and the Race*. Macmillan, New York
Bartsch K, Estes D 1996 Individual differences in children's developing theory of mind and implications for metacognition. *Learning and Individual Differences* **8**: 281–304
Bartsch K, Wellman H M 1995 *Children Talk About the Mind*. Oxford University Press, Oxford, UK
Butterworth G 1982 A brief account of the conflict between the individual and the social in models of cognitive growth. In: Butterworth G, Light P (eds.) *Social Cognition: Studies of the Development of Understanding*. The Harvester Press, Brighton, UK, pp. 3–16
Dodge K 1986 A social information processing model of social competence in children. In: Perlmutter M (ed.) *Minnesota Symposium on Child Psychology*. Erlbaum, Hillsdale, NJ, Vol. 18, pp. 77–125
Doise W, Mugny G 1984 *The Social Development of the Intellect*. Pergamon, Oxford, UK
Dunn J 1996 The Emanuel Miller Memorial Lecture 1995: Children's relationships: Bridging the divide between cognitive and social development. *Journal of Child Psychology and Psychiatry and Allied Disciplines* **37**: 507–18
Eisenberg N, Fabes R 1998 Prosocial development. In: Damon W (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. 3, pp. 701–78
Eisenberg N, Losoya S, Guthrie I K 1997 Social cognition and prosocial development. In: Hala S (ed.) *The Development of Social Cognition*. Psychology Press, East Sussex, UK, pp. 329–63
Goodnow J J 1990 The socialization of cognition: What's involved? In: Stigler J W, Shweder R A, Herdt G (eds.) *Cultural Psychology*. Cambridge University Press, Cambridge, UK, pp. 259–86
Hala S 1997 *The Development of Social Cognition*. Psychology Press, East Sussex, UK
Kohlberg L 1984 *The Psychology of Moral Development*. Harper and Row, San Francisco, CA, Vol. 2
Ladd G W 1999 Peer relationships and social competence during early and middle childhood. *Annual Review of Psychology* **50**: 333–59
Meins E 1997 *Security of Attachment and the Social Development of Cognition*. Psychology Press, East Sussex, UK
Piaget J 1926 *The Language and Thought of the Child*. Kegan Paul, London
Piaget J [1932] 1965 *The Moral Judgement of the Child*. Free Press, Glencoe, IL
Rogoff B 1990 *Apprenticeship in Thinking*. Oxford University Press, Oxford, UK
Wertsch J V 1985 *Vygotsky and the Social Formation of Mind*. Harvard University Press, Cambridge, MA

J. F. Dunn

Social Cognitive Theory and Clinical Psychology

1. Social Cognitive Theory

Human behavior has often been explained in terms of one-sided causation. In these unidirectional models, behavior is depicted as being predominantly shaped and controlled by environmental influences, or impelled by inner drives and dispositions. Social cognitive theory explains human functioning in terms of triadic reciprocal causation (Bandura 1986). In this model of reciprocal determinism personal determinants in the form of cognitive, biological, and emotional factors, behavior patterns, and environmental events all operate as interacting determinants that influence each other bidirectionally.

2. Agentic Perspective

Social cognitive theory is rooted in an agentic perspective. People are self-organizing, proactive, self-reflecting, and self-regulating, not just reactive organisms shaped and shepherded by external events. Human adaptation and change are rooted in social systems. Therefore, personal agency operates within a broad network of sociostructural influences. In these agentic transactions, people are producers as well as products of social systems. Sociostructural and personal determinants are treated as co-factors within a unified causal structure rather than as rival conceptions of human behavior.

3. Fundamental Human Capabilities

In social cognitive theory, people are characterized in terms of a number of fundamental capabilities. These are enlisted in effecting personal and social change.

3.1 Symbolizing Capability

The extraordinary capacity to represent events and their conditional relations in symbolic form provides humans with a powerful tool for comprehending their environment and for creating and managing environmental conditions that touch virtually every aspect of their lives. Most environmental events exert their effects indirectly, through cognitive processing, rather than directly. Cognitive factors partly determine which environmental events are observed, what meaning is conferred on them, what emotional impact and motivating power they have, and how the information they convey is organized and preserved for future use. Through the medium of symbols, people transform transient experiences into cognitive models that serve as guides for reasoning and action. By symbolizing their experiences, people give structure, meaning, and continuity to their lives. The sociocognitive methods for promoting personal and social change draw heavily on this symbolizing capability.

If put to faulty use, the capacity for symbolization can cause distress. Indeed, many human dysfunctions and torments stem from problems of thought. People live in a psychic environment largely of their own making. In their thoughts, they often dwell on painful

pasts and on perturbing futures of their own invention. They burden themselves with stressful arousal through apprehensive rumination, debilitate their efforts by self-impeding ideation, drive themselves to despondency by harsh self-evaluation and dejecting modes of thinking, and often act on misconceptions that get them into trouble. Thought can thus be a source of human failings and distress as well as a source of human accomplishments. Faulty styles of thinking can be changed to some extent by verbal analysis of faulty inferences from personal experiences. However, corrective and enabling mastery experiences are more persuasive than talk alone in altering faulty beliefs and dysfunctional styles of thinking (Bandura 1986, 1997, Rehm 1988, Williams 1990).

3.2 *Vicarious Capability*

There are two basic modes of learning. People learn by experiencing the effects of their actions, and through the power of social modeling. Trial and error learning is a tedious and hazardous process. Fortunately, this process can be short cut by social modeling. Humans have evolved an advanced capacity for observational learning that enables them to expand their knowledge and competencies rapidly through the information conveyed by the rich variety of models (Bandura 1986, Rosenthal and Zimmerman 1978).

Modeling is not simply a process of response mimicry as commonly believed. Modeled activities convey rules for generative behavior. Once observers extract the rules underlying the modeled activities they can generate new patterns of behavior that go beyond what they have seen or heard. Self-regulatory and other cognitive skills can also be developed by using models to verbalize plans, action strategies, and self-guidance to counteract self-debilitating thought patterns, as solutions to problems are worked through (Meichenbaum 1984). Deficient and faulty habits of thought are corrected as participants adopt and enact the verbal self-guidance.

In addition to cultivating new competencies, modeling influences can alter incentive motivation, emotional proclivities, and value systems (Bandura 1986). Seeing others achieve desired outcomes by their efforts can instill motivating outcome expectations in observers that they can secure similar benefits for comparable performances; seeing others punished for engaging in certain activities can instill negative outcome expectations that serve as disincentives. Observers can also acquire lasting attitudes and emotional and behavioral proclivities toward persons, places, or things that have been associated with modeled emotional experiences. Observers learn to fear the things that frightened models, to dislike what repulsed them, and to like what gratified them.

During the course of their daily lives, people have direct contact with only a small sector of the physical and social environment. As a result, their conceptions of social reality are greatly influenced by modeled representations of society, mainly by the mass media (Gerbner 1972). Video and computer systems feeding off telecommunications satellites are now rapidly diffusing new ideas, values, and styles of conduct worldwide. At the societal level, symbolic modeling is transforming how social systems operate and serving as a major vehicle for sociopolitical change (Bandura 1997). As Braithwaite (1994) has shown, the speed with which Eastern European rulers and regimes were toppled was greatly accelerated by televised modeling of successful mass action.

The vast body of knowledge on modeling processes is being widely applied for personal development, therapeutic purposes, and social change (Bandura 1997, Bandura and Rosenthal 1978, Rosenthal and Steffek 1991). For human problems that stem from sociocognitive deficits, the guided mastery approach that is highly effective in cultivating personal efficacy and psychosocial competencies combines three components. First, the appropriate competencies are modeled in a stepwise fashion to convey the basic rules and strategies. Second, the learners receive guided practice under simulated conditions to develop proficiency in the skills. Third, they are provided with a graduated transfer program that helps them to apply their newly learned skills in their everyday lives in ways that will bring them success. In ameliorating anxiety and phobic dysfunction, the modeling provides coping strategies for managing threats. One can eventually overcome fears without the aid of modeling through repeated unscathed contact with threats, although it takes longer and is more stressful. However, modeling of cognitive and behavioral skills is a key ingredient in the development of complex competencies.

Sociocognitive approaches rely on mastery experiences as the principal vehicle of change. The enabling power of social modeling is enhanced by guided mastery enactments. When people avoid what they dread, they lose touch with the reality they shun. Guided mastery treatment quickly restores reality testing in two ways. It provides disconfirming tests of phobic beliefs by persuasive demonstrations that what phobics dread is safe. Even more importantly, it provides confirmatory tests that phobics can exercise control over what they fear. Intractable phobics, of course, are not about to do what they dread. Therapists must, therefore, create environmental conditions that enable phobics to succeed despite themselves. This is achieved by enlisting a variety of performance-mastery aids. Threatening activities are repeatedly modeled to demonstrate coping strategies and to disconfirm people's worst fears. Intimidating tasks are reduced to graduated subtasks of easily mastered steps. Joint performance with the therapist enables frightened people to do things they would refuse to do on their own. Another method for overcoming re-

sistance is for phobics to perform the feared activity for only a short time. As they become bolder the length of involvement is extended. Protective conditions can be introduced to weaken resistance that retards change. After effective functioning is fully restored, increasingly challenging self-directed mastery experiences are then arranged to strengthen and generalize the sense of coping efficacy. Varied mastery experiences build resiliency by reducing vulnerability to the negative effects of adverse experiences. This is a powerful treatment that eliminates phobias, anxiety arousal, biochemical stress reactions, and wipes out recurrent nightmares and intrusive rumination (Bandura 1997, Williams 1990).

Symbolic modeling lends itself readily for society-wide applications through creative use of the electronic media. For example, the soaring population growth and the environmental devastation it produces is the most urgent global problem. Worldwide use of enabling and motivating dramatic serials is raising people's efficacy to exercise control over their family lives, enhancing the status of women to have some say in how they live their lives, and lowering the rates of childbearing (Singhal and Rogers 1999, Vaughan et al. 1995).

3.3 Self-regulatory Capability

People are self-reactors with a capacity to motivate, guide, and regulate their activities through the anticipative mechanism of forethought. People set goals for themselves, anticipate the likely consequences of prospective actions, and plan courses of action that are likely to produce desired outcomes and avoid detrimental ones. The projected future is brought into the present through forethought. A forethoughtful perspective provides direction, coherence, and meaning to one's life.

Much human motivation and behavior is regulated anticipatorily by the material and social outcomes expected for given courses of action. Social cognitive theory broadens this functionalism to include self-evaluative outcomes. People do things that give them satisfaction and a sense of self-worth, and refrain from actions that evoke self-devaluative reactions. Incentive systems are enlisted, if needed, to promote and help sustain personal and social change (Bandura 1986, 1998, O'Leary and Wilson 1987).

In keeping with the agentic perspective of social cognitive theory, the greatest benefits that psychological treatments can bestow are not specific remedies for particular problems, but the self-regulatory capabilities needed to deal effectively with whatever situations might arise. To the extent that treatment equips people to exercise influence over events in their lives, it initiates an ongoing process of self-regulative change.

Personal standards for judging and guiding one's actions play a major role in self-motivation and in the exercise of self-directedness (Bandura 1991, Locke and Latham 1990). Self-regulatory control is achieved by creating incentives for one's own actions and by anticipative reactions to one's own behavior depending on how it measures up to personal standards. Sociocognitive principles of self-regulation provide explicit guidelines for self-motivation, personal development, and modification of detrimental styles of behavior.

A favorable self-regulatory system provides a continuing source of motivation, self-directedness, and personal satisfaction, but a dysfunctional self-system can breed much human misery. For people who adopt stringent personal standards, most of their accomplishments bring them a sense of failure and self-disparagement (Bandura 1997, Rehm 1988). In its more extreme forms, harsh standards of self-evaluation give rise to despondency, chronic discouragement, and feelings of worthlessness and lack of purposefulness. For example, Hemingway, who took his own life, imposed upon himself demands that were unattainable throughout his life, pushed himself to extraordinary feats, and constantly demeaned his accomplishments.

Effective treatments for despondency arising from stringent self-imposed standards remedy each of the dysfunctional aspects of the self-system (Rehm 1981). They correct self-belittling interpretative biases that minimize one's successes and accentuate one's failures. To increase self-satisfaction and a sense of personal accomplishment, participants are taught how to set themselves attainable subgoals in activities of personal significance and to focus their efforts and self-evaluation on progress toward their aspirations. Through proximal structuring, goals become motivating rather than demoralizing. Depressed individuals tend to be less self-rewarding for successes and more denying and self-punishing for failures than the nondepressed for similar performances. They are taught how to be more self-rewarding for their progressive personal attainments.

Deficient or deviant standards also create problems, although the distress is inflicted on others rather than on oneself. Unprincipled individuals who pursue an ethic of expediency, and those who pride themselves in antisocial activities readily engage in conduct that is socially injurious (Bandura 1991). Antisocial and transgressive styles of conduct have diverse sources requiring multifaceted approaches to personal change. Successful treatments promote the development of competencies and prosocial self-regulatory standards and styles of behavior sufficiently attractive to supplant antisocial ones (Bandura 1973, Reid and Patterson 1991, Silbert 1984). Education provides the best escape from crime and poverty. Promoting educational development is, therefore, a key factor in reducing crime and delinquency.

3.4 Self-reflective Capability

Among the mechanisms of personal agency none is more central or pervasive than people's beliefs in their capability to exercise control over their own functioning and over environmental events (Bandura 1997). Efficacy beliefs are the foundation of human agency. Unless people believe that they can produce desired results by their actions, they have little incentive to act or to persevere in the face of difficulties. Meta-analyses, which combine the findings of numerous studies, attest to the influential role played by efficacy beliefs in human adaptation (Holden 1991, Holden et al. 1990, Multon et al. 1991, Stajkovic and Luthans 1998).

Self-efficacy beliefs regulate human functioning through their impact on cognitive, motivational, emotional, and choice processes. They determine whether people think pessimistically or optimistically, and in self-enhancing or self-debilitating ways. Efficacy beliefs play a central role in the self-regulation of motivation through goal challenges and outcome-expectations. Whether people act on the outcomes they expect prospective performances to produce depends on their beliefs about whether or not they can produce those performances. It is partly on the basis of efficacy beliefs that people choose what challenges to undertake, how much effort to expend in the endeavor, how long to persevere in the face of obstacles and failures, and whether failures are motivating or demoralizing. Efficacy beliefs play a key role in shaping the courses lives take by influencing the types of activities and environments people choose to enter. By choosing their environments, people can have a hand in what they become.

People's beliefs in their coping capabilities also play a pivotal role in the self-regulation of affective states (Bandura 1997). Efficacy beliefs influence how potential threats are perceived and cognitively processed. If people believe they can manage threats they are not distressed by them. But if they believe they cannot control potential threats they experience high anxiety. Through inefficacious thinking they distress themselves and constrain and impair their functioning.

Many human distresses result from failures of thought control. It is not the sheer frequency of disturbing cognitions, but the perceived inability to turn them off that is the major source of distress. Similarly, in obsessional disorders, it is not the ruminations, *per se*, but the perceived inefficacy to stop them that is perturbing.

Phobics display high anxiety and physiological stress to threats they believe they cannot control. After their perceived efficacy is raised to the maximal level by guided mastery experiences, they manage the same threats with equanimity.

It was widely assumed that phobic behavior was controlled by anxiety. Many therapeutic procedures were, therefore, keyed to extinguishing anxiety arousal. The anxiety control theory has not withstood experimental scrutiny, however. A low sense of coping efficacy produces both anxiety and phobic behavior. People often perform activities even though highly anxious, as long as they believe they can master them. Intense stage fright, for example, does not stop actors from going on stage. Conversely, people avoid situations they believe exceed their coping capabilities without waiting for visceral arousal to tell them to do so. Williams (1992) has shown that people base their actions on efficacy beliefs in situations they regard as risky not on anxiety arousal.

A low sense of efficacy to exercise control over things one values can give rise to feelings of futility and despondency through several pathways. In one pathway, a low sense of efficacy to fulfill personal standards of worth gives rise to self-devaluation and depression. A second pathway occurs through a low sense of social efficacy to develop social relationships that bring satisfaction to people's lives and enable them to manage chronic stressors. Another is through the exercise of control over depressing thoughts themselves. Low efficacy to regulate ruminative thought contributes to the occurrence of depressive episodes, how long they last, and how often they recur.

People's beliefs about their efficacy are constructed from four principal sources of information. The most effective way of instilling a strong sense of efficacy is through mastery experiences. Successes build a robust belief in one's personal efficacy. Failures undermine it. Development of resilient self-efficacy requires experiences in overcoming obstacles through perseverant effort. The second method is by social modeling. Models serve as sources of competencies and motivation. Seeing people similar to oneself succeed by perseverant effort raises observers' beliefs in their own capabilities. Social persuasion is the third mode of influence. The fourth way of altering self-efficacy beliefs is to enhance physical strength and stamina and alter mood states on which people partly judge their capabilities.

These different modes of influence create and strengthen beliefs of personal efficacy across diverse spheres of functioning. Such beliefs predict the level, scope, and durability of behavioral changes (Bandura 1997). Microanalyses of how given treatments work, reveal that the efficacy belief system is a common pathway through which different types of treatment produce their effects. Knowledge of the determinants and processes governing the formation of efficacy beliefs provides explicit guidelines on how best to structure programs to achieve desired change.

4. Sociostructural Change

Many human problems are sociostructural, not simply individual. Common problems require social solutions through collective action. Social cognitive theory

extends the conception of human agency to collective agency (Bandura 1997). People's shared belief in their collective power to produce desired results is a key ingredient of collective agency. The stronger the people's perceived collective efficacy the higher their aspirations and motivational investment in their undertakings, the stronger their staying power in the face of impediments and setbacks, the higher their morale and resilience to adversity, and the greater their accomplishments.

Efforts to improve the human condition must, therefore, be directed not only at treating the casualties of adverse social practices, but also at altering the social practices producing the causalities. The models of social change derived from social cognitive theory (Bandura 1986, 1997), draw heavily on knowledge of modeling, motivational, regulatory, and efficacy mechanisms operating at the collective level.

See also: Self-efficacy; Self-regulated Learning

Bibliography

Bandura A 1973 *Aggression: A Social Learning Analysis*. Prentice-Hall, Englewood Cliffs, NJ

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ

Bandura A 1991 Social cognitive theory of moral thought and action. In: Kurtines W M, Gewirtz J L (eds.) *Handbook of Moral Behavior and Development*. Erlbaum, Hillsdale, NJ, Vol. 1, pp. 45–103

Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman W H, New York

Bandura A, Rosenthal T L 1978 Psychological modeling: Theory and practice. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 2nd edn. Wiley, New York, pp. 621–58

Braithwaite J 1994 A sociology of modeling and the politics of empowerment. *British Journal of Sociology* **45**: 445–79

Gerbner G 1972 Communication and social environment. *Scientific American* **227**: 152–60

Holden G 1991 The relationship of self-efficacy appraisals to subsequent health related outcomes: A meta-analysis. *Social Work in Health Care* **16**: 53–93

Holden G, Moncher M S, Schinke S P, Barker K M 1990 Self-efficacy of children and adolescents: A meta-analysis. *Psychological Reports* **66**: 1044–6

Locke E A, Latham G P 1990 *A Theory of Goal Setting and Task Performance*. Prentice-Hall, Englewood Cliffs, NJ

Meichenbaum D 1984 Teaching thinking: a cognitive-behavioral perspective. In: Glaser R, Chipman S, Segal J (eds.) *Thinking and Learning Skills: Research and Open Questions*. Erlbaum, Hillsdale, NJ, Vol. 2, pp. 407–26

Multon K D, Brown S D, Lent R W 1991 Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology* **38**: 30–8

O'Leary K D, Wilson G T 1987 *Behavior Therapy: Application and Outcome*. 2nd edn., Prentice-Hall, Englewood Cliffs, NJ

Rehm L P 1981 A self-control therapy program for treatment of depression. In: Clarkin J F, Glazer H I (eds.) *Depression: Behavioral and Directive Treatment Strategies*. Garland STPM Press, New York, pp. 68–110

Rehm L P 1988 Self-management and cognitive processes in depression. In: Alloy L B (ed.) *Cognitive Processes in Depression*, Guilford Press, New York, pp. 143–76

Reid J B, Patterson G R 1991 Early prevention and intervention with conduct problems: a social interactional model for the integration of research and practice. In: Stoner G, Shinn M, Walker H (eds.) *Interventions for Achievements and Behavior Problems*. National Association of School Psychologists, Silver Spring, MD, pp. 715–39

Rosenthal T L, Steffek B D 1991 Modeling applications. In: Kanfer F H, Goldstein A P (eds.) *Helping People Change*, 4th edn. Pergamon, New York, pp. 70–121

Rosenthal T L, Zimmerman B J 1978 *Social Learning and Cognition*. Academic Press, New York

Silbert M H 1984 Delancy Street Foundation—process of mutual restitution. In: Reissman F (ed.) *Community Psychology Series*, Human Sciences, New York, Vol. 10, pp. 41–52

Singhal A, Rogers E M 1999 *Entertainment-education: A Communications Strategy for Social Change*. Erlbaum, Mahwah, NJ

Stajkovic A D, Luthans F 1998 Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin* **124**: 240–61

Vaughan P W, Rogers E M, Swalehe R M A 1995 *The effects of "Twende Na Wakati," an entertainment-education radio soap opera for family planning and HIV/AIDS prevention in Tanzania*. Unpublished manuscript. University of New Mexico, Albuquerque, NM

Williams S L 1990 Guided mastery treatment of agoraphobia: beyond stimulus exposure. In: Hersen M, Eisler R M, Miller P M (eds.) *Progress in Behavior Modification*. Sage, Newbury Park, CA, Vol. 26, pp. 89–121

Williams S L 1992 Perceived self-efficacy and phobic disability. In: Schwarzer R (ed.) *Self-efficacy: Thought Control of Action*. Hemisphere, Washington, DC, pp. 149–76

A. Bandura

Social Comparison, Psychology of

Social comparison occurs when an individual compares himself or herself to another person. The dimension on which the comparison is made can be almost anything—attractiveness, tennis skill, wealth, etc. The comparison might be deliberately sought, or it might occur by chance. It is not social comparison when one compares one person to another, as in comparing one neighbor to another; the individual making the comparison must be included in the comparison for it to fit the commonly accepted definition.

1. The Original Theory

Leon Festinger's (1954) paper, *A Theory of Social Comparison Processes*, still serves as the basic formulation, although there have been many additions

and revisions. Festinger argued that it is necessary for people to know what they are capable of doing and to hold correct opinions about the world. Sometimes people can perform a physical reality test on their abilities and opinions, but that could be dangerous, inconvenient, or impossible. For example, if individuals want to know whether they can swim across a swift river, they could simply try it, but that could be dangerous. If it is a person's opinion that the Russian Revolution would not have occurred without V. I. Lenin, there is simply no physical reality test of the truth of that opinion. Consequently, people often seek *social reality* rather than physical reality for their abilities and opinions, and that social reality is obtained through social comparison.

1.1 Similarity

A cornerstone of Festinger's theory is that comparison with others who are similar provides better information than comparison with dissimilar others. The theory was ambiguous, however, about the concept of similarity, and critics argued that if an individual already knows that someone is similar, that implies that a comparison has already been made and that further comparison with that person would be redundant. Goethals and Darley (1977) combined *attribution theory* with social comparison theory and suggested that similarity should refer to similarity on 'related attributes,' attributes correlated with and predictive of the particular ability or opinion to be evaluated. To answer the question, 'How good a tennis player am I?' a person should compare with someone of the same age, gender, physical condition, and tennis experience. The person should logically expect to perform equally to that comparison target and, if this turns out to be so, would probably feel that he or she is a fairly good player ('as good as I ought to be'). If the comparison is with someone advantaged (or disadvantaged) on related attributes, the person would expect to lose (or to win) and, if this turns out to be so, wouldn't be able to draw any clear conclusion about his or her ability level. The person wouldn't know whether the loss or win was due to ability or to the related attributes. 'Did I win because I am a good tennis player, or because I am in better physical condition?'

1.2 The Unidirectional Drive Upward

An important difference between opinions and abilities in Festinger's theory is a *unidirectional drive upward* for abilities but not for opinions. Regardless of the ability, people would generally like to have more of it than other people have. With opinions, however, people feel confident when they have the same opinions as others. Thus, all the members of a group can be satisfied when they have the same opinion, but not when they have exactly the same level of ability.

2. Methods for Studying Social Comparison

The three major methods for studying social comparison are 'selection' approaches, which examine characteristics of the people with whom individuals choose to compare; 'reaction' approaches, which examine individuals' emotional or cognitive reactions to comparison information; and 'narration' approaches, which examine individuals' self-reports of comparisons made (Wood 1996). Each of these will be examined below.

2.1 The Selection Approach

Individuals are given the choice of two or more comparison targets that differ in theoretically important ways. The choices they make (and related data) provide clues concerning social comparison motives. The first notable example, known as the 'rank order paradigm,' was introduced to study the unidirectional drive upward. Experimental participants were tested on a (fictitious) desirable or undesirable personality trait such as 'intellectual flexibility' or 'intellectual rigidity' and were told that they occupied the middle rank in a group of several other participants tested at the same time. Participants were given their own (fictitious) score on the test, as well as the rank of each of the other participants, and were told that they could be shown the score of one other person in the group. The rank of the person participants chose to see could be of someone better off than themselves (known as an 'upward comparison') or worse off than themselves (a 'downward comparison'). In most of the experiments using this method, participants made upward comparisons. That is, they wanted to compare their scores to better scores, either more flexible or less rigid (Wheeler et al. 1969). They believed that they were similar to and almost as good as those better off than themselves and chose to compare with these others in order to confirm that similarity.

2.2 The Reaction Approach

Individuals are presented with social comparison information, and their responses to that information are measured. In what is known as the Mr. Clean/Mr. Dirty study, job applicants casually encountered another applicant whose appearance was either socially desirable (dark suit, well-groomed, and confident)

or socially undesirable (torn and smelly clothes, poorly-groomed, confused). Those participants who encountered Mr. Clean showed a decrease in self-esteem measured a few minutes later, whereas those who encountered Mr. Dirty increased in self-esteem (Morse and Gergen 1970). In this research, participants contrasted themselves against the other job applicant, but this is not always what happens. For example, Van der Zee et al. (1998) exposed cancer patients to a bogus interview with another patient who was doing better than participants (upward comparison) or worse (downward comparison). Patients responded with more positive affect to upward than to downward comparisons, particularly those patients who scored high on a measure of identification with the upward comparison patient. Collins (1996) called this 'upward assimilation,' the perception that one is similar to superior others. This is opposed to 'upward contrast,' in which comparing with someone who is better off (like Mr. Clean) makes people feel worse about themselves.

2.3 The Narrative Approach

This approach attempts to measure the comparisons that people make or have made in their everyday lives. At the most basic level, participants are sometimes asked who they have compared themselves with in the past. This places too great a burden on memory, however, and a better method is the *Social Comparison Record*, a diary recording of all comparisons as they occur over a period of two weeks, including the dimension of comparison (e.g., appearance, happiness), the direction (e.g., upward, downward), relationship to the target (e.g., friend, stranger), and mood before and after the comparison (Wheeler and Miyake 1992).

Free response measures are also narrative in approach. These are spontaneous statements, usually made during the course of an interview or conversation, implying comparisons. Wood et al. (1985) reported that breast cancer patients made many free response comparison statements, most of which seemed to indicate downward comparison, such as 'I think I am coping a little better than these other women.'

3. Downward Comparison Theory

The rank-order experiments drew attention to the difference between upward and downward comparisons, and one of the experiments (Hakmiller 1966) demonstrated downward comparison. Participants who were threatened with having a very high score on the undesirable personality trait 'hostility toward one's parents' chose to compare with others who had even more of the trait. This was a downward comparison because it was with someone worse off than themselves. By verifying that someone had even more of the negative trait than themselves, participants could defend their self-esteem. Influenced by this research, Wills (1981) proposed a *theory of downward comparison*—that persons experiencing threat or negative affect will enhance their subjective well-being through comparison with someone worse off than themselves, thereby reducing distress. This prediction received support from studies of cancer and other medical patients, particularly those studies using free response measures (see Buunk and Gibbons 1997).

3.1 Challenges to Downward Comparison Theory

There is substantial controversy about the empirical support for downward comparison theory. The theory makes the intuitively appealing prediction that when people are experiencing distress they will make a downward comparison to reduce the distress. Others have questioned whether people experiencing distress are psychologically capable of making a downward comparison; their attention may be so focused on their distress that they are unable to see others as being worse off than themselves. In research using the Social Comparison Record (see above), Wheeler and Miyake (1992) found that when participants were experiencing negative affect, they made upward comparisons rather than the downward comparisons predicted by downward comparison theory. In contrast, when the same participants were experiencing positive affect, they made downward comparisons. Such a result is not consistent with downward comparison theory but is consistent with *affect-cognition priming* models in which affect primes (makes ready and available) cognitions about the self that are congruent with the affect. Experiencing negative affect primes a person to have negative thoughts about the self and thus to see others as superior to the self, leading to a contrastive upward comparison and further feelings of inferiority and negative affect.

4. A Return to the Concerns of the Original Theory

4.1 Abilities

It should be apparent that Festinger's (1954) original theory is very different from most of the subsequent research and theory. Festinger was concerned with social comparison providing accurate evaluation of one's opinions and abilities, whereas much of the later work was concerned with how social comparison relates to how people feel about themselves. Some recent work has returned to Festinger's original con-

cerns. For example, one of the questions addressed by Festinger's theory was 'Can I do X?' Can I swim across the river? Can I complete a college degree? Can I have both a family and a career? One way to answer such questions is to compare oneself to someone—called a proxy—who has already succeeded at the task (Martin 2000). If people seem to have the same amount or more of the ability or abilities required as the proxy does, they could conclude with some confidence that they could also perform the task. There are two ways that people can be confident that they have as much of the underlying ability as the proxy: (a) they have competed directly against the proxy and have performed as well as the proxy when the proxy was performing to maximum, or (b) they have observed the proxy performing, are unsure whether the proxy was performing to maximum, but are similar to the proxy on attributes related to the ability. In both cases they have established with some degree of certainty that they are similar in ability to the proxy and thus have the same action possibilities.

4.2 Opinions and Attitudes

Suls (2000) proposed that the social comparison of opinions is best considered in terms of three different evaluative situations: preference assessment (i.e., 'Do I like X?'), belief assessment (i.e., 'Is X correct?'), and preference prediction (i.e., 'Will I like X?'). Each evaluative question is associated with a different comparison dynamic. (a) If the question is 'Do I like X?,' persons similar in related attributes have special importance. Thus, in deciding whether one likes a new political candidate, a person should prefer comparison with someone with similar political values and goals. Such a comparison target should share one's assessment of the candidate, and if not, something is amiss, and the situation should be re-examined. (b) If the question is 'Is X correct?,' comparisons with persons of more advantaged status (or 'expert') are most meaningful, although they also should hold certain basic values in common (the 'similar expert'). Thus, to determine the truth of the belief that guns should not be in private hands, comparison with the opinion of an expert on the relevant statistics would be useful, but only if the expert shared one's values about human life, freedom, etc. (c) If the question is 'Will I like X?,' the most meaningful comparisons are with persons who have already experienced 'X' and who share a consistent history of preferences. Thus, in deciding whether a movie would be enjoyable, a person would consult with someone who has seen the movie and who has in the past usually agreed or usually disagreed with the person. If the other has usually agreed, the person would expect to have the same reaction to the movie; if the other has usually disagreed, the person would expect to have an opposed reaction to the movie.

5. Prospects

There is more interest in social comparison theory now than any time in its almost half-century history because it has become clearer that comparison with other people is central to the human experience. The theory is being applied to new problems and is increasingly influencing and being influenced by other areas within psychology.

See also: Attributional Processes: Psychological; Cognitive Dissonance; Personality and Conceptions of the Self; Self-efficacy; Self-evaluative Process, Psychology of; Self-monitoring, Psychology of

Bibliography

Buunk B P, Gibbons F X (eds.) 1997 *Health, Coping, and Well-Being: Perspectives from Social Comparison Theory*. Lawrence Erlbaum Associates, Mahwah, NJ

Collins R L 1996 For better or worse: The impact of upward social comparison on self-evaluations. *Psychology Bulletin* **119**(1): 51–69

Festinger L 1954 A theory of social comparison processes. *Human Relations* **7**: 117–40

Goethals R R, Darley J M 1977 Social comparison theory: an attributional approach. In: Suls J M, Miller R L (eds.) *Social Comparison Processes: Theoretical and Empirical Perspectives*. Hemisphere, Washington, DC

Hakmiller K L 1966 Threat as a determinant of downward comparison. *Journal of Experimental Social Psychology* **2**(1): 32–9

Martin R 2000 'Can I do X?': Using the proxy comparison model to predict performance. In: Suls J, Wheeler L (eds.) *Handbook of Social Comparison: Theory and Research*. Plenum, New York

Morse S, Gergen K J 1970 Social comparison, self-consistency, and the concept of self. *Journal of Personality and Social Psychology* **16**: 148–56

Suls J 2000 Opinion comparison: The role of corroborator, expert, and proxy in social influence. In: Suls J, Wheeler L (eds.) *Handbook of Social Comparison: Theory and Research*. Kluwer Academic/Plenum, New York

Van der Zee K, Buunk B, Sanderman R 1998 Neuroticism and reactions to social comparison information among cancer patients. *Journal of Personality* **66**(2): 175–94

Wheeler L, Shaver K G, Jones R A, Goethals G R, Cooper J, Robinson J E, Gruder C L, Butzine K 1969 Factors determining choice of a comparison other. *Journal of Experimental Social Psychology* **5**: 219–32

Wheeler L, Miyake K 1992 Social comparison in everyday life. *Journal of Personality and Social Psychology* **62**(5): 760–73

Wills T A 1981 Downward comparison principles in social psychology. *Psychology Bulletin* **90**: 245–71

Wood J V 1996 What is social comparison and how should we study it? *Personality and Social Psychology Bulletin* **22**(5): 520–37

Wood J V, Taylor S E, Lichtman R R 1985 Social comparison in adjustment to breast cancer. *Journal of Personality and Social Psychology* **49**: 1169–83

L. Wheeler, J. Suls, and R. Martin

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Social Competence: Childhood and Adolescence

1. Definitions

Social competence refers to the effectiveness of a person's functioning as an *individual*, in *dyadic relationships* and in *groups*. Like other general constructs used by psychologists (e.g., *risk*, *stress* and *vulnerability*), competence refers to a general class or category of phenomena. In this respect, *competence* brings together a set of ideas or concepts regarding individuals' capacities to function in their environments. Insofar as competence refers to a property of the relationship between the individual and the environment, one cannot think of individuals as being competent without knowing the contexts in which they are engaged. Competence is also a metatheoretical construct in the sense that it is not linked to a single theoretical perspective. Instead, it serves as an organizing principle for research programs and it allows scientists from diverse areas of study to see the links and commonalities among them.

1.1 Initial Conceptualizations

An etymological analysis of the word *competence* shows that it refers to the engagement of an individual in activity oriented toward the achievement of a state or goal (Bukowski et al. 1998). Therefore, a competent person is not a person who possesses a particular skill or ability but is a person who is able to engage in the activities needed to achieve particular goals or outcomes. This exact view—that competence is a process—serves as the centerpiece of White's (1959) seminal conceptualization of competence as an organism's orientation toward an efficient interaction with the environment. For White (1959), this orientation was a profound and basic form of motivation, as critical as any fundamental drive. According to his view, competence is not a form of motivated behavior that serves other drives, but is a fundamental form of motivation in and of itself. To support this argument, he referred to the high rates of exploratory behavior seen in many animals and the infant's instinctual curiosity in novelty. Both of these phenomena show that organisms are inherently oriented toward interaction with the environment. He found evidence of this 'instinct to master' in the observations and interpretations provided by neo-analysts (e.g., Fenichel 1945, Hendrick 1942). Hendrick, for example, went so far as to write that individuals have an 'inborn drive to do and to learn to do.' Again, White was careful to point out that according to these neo-analysts, such activity is not in the service of another goal (e.g., anxiety reduction), but is instead a basic human motivation by itself. White noted that this proposal was central also to the ideas of Erik Erikson. In Erikson's (1953) description of the latency years, he emphasized that this developmental period is a time in which competence is the main developmental struggle that children confront. Erikson wrote that 'children need a sense of being able to make things and to make them well and even perfectly: this is what I call a sense of industry.'

1.2 Current Definitions

The impact of White's views is, in part, apparent in the definitions of competence that subsequent writers have offered. These include the following: 'the effectiveness or adequacy with which an individual is capable of responding to various problematic situations which confront him' (Goldfried and D'Zurilla 1969 p. 161); 'an individual's everyday effectiveness in dealing with his environment' (Zigler 1973); 'a judgment by another that an individual has behaved effectively' (McFall 1982 p. 1); 'attainment of relevant social goals in specified social contexts, using appropriate means and resulting in positive developmental outcomes' (Ford 1982 p. 324); the ability 'to make use of environmental and personal resources to achieve a good developmental outcome' (Waters and Sroufe 1983 p. 81); and 'the ability to engage effectively in complex interpersonal interaction and to use and understand people effectively' (Oppenheimer 1989 p. 45). In a review of these definitions, Rubin et al. (1998), pointed out that they share at least two points: (a) an emphasis on effectiveness, and (b) an understanding of the environment so as to achieve one's own needs or goals. One has to presume that the dynamism that stood at the center of White is implicit in these definitions even if it is not stated explicitly.

Using definitions noted above, as well as from others found in the work of Attili (1989), Dodge (1986), Gresham (1981), Strayer (1989) as a point of departure, Rubin and Rose-Krasnor (1992, 1986) have defined social competence as *the ability to achieve personal goals in social interaction while simultaneously maintaining positive relationships with others over time and across situations*. A significant feature of this definition is its implicit recognition of the importance of individual and social goals. This emphasis reflects essential duality of self and other, placing the individual within a social and personal context. The conceptualization of self and other as indistinguishable is an important feature of theory from several domains including personality theory (Bakan 1966, Sullivan 1953), gender role theory (Leaper 1994), feminist psychology (Jordan et al. 1991), and philosophy (Harré 1979, 1984). Accordingly, this definition is valuable as it points to the complex goals that persons confront as individuals and as members of

groups. It is in this regard that individuals, as well as social scientists, have struggled with the tension between agentic and communal orientations.

2. The Socially Competent Child and Adolescent

Children's competence within social contexts depends upon their social skills. These skills are both behavioral and cognitive. The cognitive skills include the ability to understand and predict the internal states (i.e., the thoughts, feelings, and intentions) of other children and to understand the meaning of these states within the present context. Using this information, a competent child is able to develop plans for interacting with others to achieve desired outcomes and to evaluate the flow of the interaction as it unfolds. Part of this evaluation concerns the consequences of one's actions as well as those of the partner, on both instrumental and moral grounds, so as to maintain positive interaction, and either repair or terminate negative interactions. In the behavioral domain, a competent child is able to act prosocially and is capable of displaying appropriate positive affect while avoiding or inhibiting negative affect and is able to communicate verbally and nonverbally so as to facilitate the partner's understanding of the interaction. (These issues are discussed further in chapters found in Asher and Coie (1990) and Schneider et al. (1989).)

2.1 Competence as Process

Recent approaches to social competence have gone beyond the emphasis on skills *per se* and have argued that social competence needs to be understood as a hierarchical process. Two models of this sort have been proposed by Rubin and Rose-Krasnor (1992) and Crick and Dodge (1994). In parallel to previous process-oriented approaches to social cognition and behavior (e.g., Miller et al. 1960) these models include a series of steps that are presumed to underlie social interaction. Rubin and Rose-Krasnor adopt an interpersonal problem solving model of social competence which is largely predicated on the view that most social interactions appear to be routine, reflecting an automaticity in thinking and a scriptedness in social interchanges (Abelson 1981, Nelson 1981, Schank and Abelson 1977). They note, for instance, that children quickly learn many rudimentary scripted forms of social behavior, such as greeting and leave-taking sequences, as well as acquiring an understanding that many social relationships 'slot' members into particular roles that are played out in highly predictable ways (e.g., Strayer 1989). Such scripts and role understandings allow children easy access to sequences of actions corresponding to highly familiar social situations thus facilitating social competence. Indeed, for both competent and less-than-competent children, familiar social situations easily elicit routine, script-driven behavior. In the absence of such familiarity, when appropriate and routine eliciting conditions are not present, this sort of script driven enactment of social behavior is precluded.

2.2 Rubin and Rose-Krasnor Model

Rubin and Rose-Krasnor propose that in such novel conditions, social competence derives from the following sequentially organized steps. First, a goal needs to be chosen. Often several goals are apparent, and they can often conflict with each other. For this reason Dodge et al. (1989) argued that the successful *coordination* of goals is more likely to be a challenge to an individual's competence than is the achievement of a particular goal. Second, the chosen goals need to be understood according to contextual factors, such as whom they are interacting with (e.g., stranger, acquaintance, friend) and this person's characteristics, and the features of the context (e.g., school, home, or playground), population density, and visibility (public versus private encounters). Next, strategies for achieving that goal need to be chosen. To do this, a child may retrieve one or more strategies from long-term memory or new strategies may be constructed from those already available. Once selected, the child must produce the strategy in its relevant context and the results need to be evaluated. If the strategy is judged to be *successful*, the problem-solving process ends and relevant information concerning the social interchange is retained in long-term memory. If a strategy has been judged as *partially successful*, the actor may accept the outcome as 'successful enough' and proceed as if the outcome was a success; alternately, the individual may decide to judge the outcome as a failure.

If this case, the child may stop the sequence and leave the goal unattained or may choose to either repeat the original strategy or choose to alter the previous strategy in some way while maintaining the same goal.

2.3 Crick and Dodge Model

Another approach to understanding the process of social competence is seen in a model proposed by Crick and Dodge (1994). They also emphasize the importance of information processing as the basis of social competence. Their model involves six stages, each of which involves a form of information processing that draws upon currently available information as well as a data base of rules, schemas, and social knowledge. The six steps in their model are as follows.

First the child *encodes* social cues by attending to relevant social information such as who is in the social milieu (a friend, a stranger, a popular child, an unpopular child), how it is that this person is expressing him or herself (affectively happy, angry, or sad), and so on. Second, the child must *interpret* the encoded cues. This requires that the child search his or her memory store in an effort to understand the meaning of the encoded information. For example, one of the social cues encoded at Step 1 might be a facial expression denoting anger. Interpreting the facial expression correctly may lead to the assumption that the social situation involves hostility and anger. Third, the child *develops a goal or set of goals* based on his/her interpretation of the encoded cues.

Fourth, the child *accesses or constructs potential responses* to achieve the selected goals. The fundamental question asked here is 'What are all the things I can do in this situation?' Next, she or he must *decide on a particular response* to be enacted. This selection depends on various considerations including the expected outcomes of particular responses, his/her confidence that the selected strategy can be enacted, and their evaluation of the response itself. For example, the individual might consider the potential interpersonal consequences ('Will she still like me?'), the instrumental consequences ('Will I get what I want?'), and the moral value of the strategy. Finally, the child *enacts* the chosen response.

The models of Rubin and Rose-Krasnor (1986) and Crick and Dodge (1994) are similar in that they both emphasize (a) that the processing underlying social competence occurs rapidly in real time, (b) that the steps of information processing follow a particular sequence, (c) that they are dynamically interrelated, yet separable, and (d) that processing can proceed automatically at the unconscious level. Most importantly they regard social competence as a plan of action by which social goals are achieved.

3. Competence at Different Levels of Social Complexity

Thoreau, in his description of his cabin near Walden Pond said 'I have three chairs—one for solitude, two for friendship, and three for society.' This multilevel scheme applies equally well to *social competence*. As we state above, competence is not a feature of a single domain of functioning but instead can be seen at multiple levels of social complexity, specifically, the *individual*, the *dyad* and the *group*. This recognition that competence needs to be assessed at multiple levels of experience is a central theme in the writings of Bronfenbrenner and Crouter (1983) and Hinde (1979, 1987, 1995). Our point is that these different levels of complexity provide different contexts for competence. One goal of research on peer relations has been to identify competence at different levels and to understand how these levels are related to each other. In the following sections we describe briefly the ways that competence is manifested at these three levels of analysis.

3.1 The Individual

The level of the individual refers to the behavioral, affective, and cognitive characteristics and dispositions that a child brings to social interaction. These characteristics include behavioral tendencies, either derived from experience or temperament, cognitions about self and other, motivations and goals, relationship history, and affective states and traits. The level of the individual is the most basic level of social complexity. While it might appear that the level of the individual has no social complexity, such complexity can exist in a child's representations of self and other (e.g., in internal working models) as well as in their planning for social interactions and relationships. Certainly, characteristics from the level of the individual can be interpreted as indicating social competence. The two features from the level of individual that have been studied most broadly concern individuals' general social behaviors and their social cognitions. In parallel to psychology's general concern with the processes of (a) moving against others, (b) moving toward others, and (c) moving away from others, research on social behavior and competence has been largely organized around the constructs of aggression, sociability and helpfulness, and withdrawal. This interest in how the child's individual behaviors influence experiences with peers has comprised a large portion of research on the peer system during the past 15 years. In fact, the very large literature on children's popularity has been typically focused on behaviors at the level of the individual. Information about the correlates of popularity is important as it reveals which characteristics from the level of the individual are related to effectiveness in establishing a general place within the peer group.

3.1.1 Individuals and behavior. In a meta-analysis of this large body of literature regarding individual behaviors and popularity, Newcomb et al. (1993) were able to reach conclusions about differences between individuals from different popularity groups, especially, the popular, rejected, and neglected groups. They were able to interpret these differences as a function of what they reveal about competence. Specifically, they reported that popular children (i.e., those who were liked by many children and disliked by very few) are more likely to be helpful, to interact actively with other children, to show leadership skills, and to engage in constructive play (see Newcomb et

al. 1993). These findings show that popular children are capable of pursuing and achieving their own goals while simultaneously allowing others to do the same. In other words, they not only engage in 'competent' action themselves, but they are able to promote it in others. One critical finding in regard to the behavior of popular children is that there are no differences between them and other children on all aspects of aggression. Newcomb et al. differentiated assertive behaviors and disruptiveness. Popular children did not differ from others on the dimension of aggression but they were significantly less disruptive. These findings show that although popular children engage in assertive and agonistic behavior, these assertive actions are unlikely to interfere with the actions and goals of others. Consistent with other reports (Olweus 1977, Pepler and Rubin 1991), it would appear as if some forms of aggression may be compatible with being popular and with the concept of social competence. As Hawley (1999) has argued, children's strategies for controlling resources do not necessarily involve coercive or agonistic actions. In this way children can be dominant and competent. In contrast with the popular children, Newcomb et al. reported that rejected children not only showed high levels of disruptive behavior but also showed little evidence of engagement in activities with peers. In other words, their activities disrupted their peers' effective engagement with the environment, and showed little effective engagement themselves. This latter point is true of the neglected group also.

3.1.2 Individuals and social cognition. A second phenomenon from the level of the individual is that of social cognition. Social cognition includes a broad array of processes, including the way that an individual interprets and understands the behaviors of one's self and of others. Akin to research on the link between behavior and experiences with peers, much of the research on social cognition has focused on how an individual's use of social information is linked to the person's experiences with peers at the group level. Indeed, research on social cognition has also examined how an individual's understanding and interpretation of experience may partially determine the impact that experiences with peers have on development. This sort of social understanding is likely to be a prerequisite for competence. Without knowing how others will act, and without having a means of understanding the motives of others, individuals do not have a 'ground' on which to base their action (Harré 1974). Moreover, in order to effectively interact with the environment, one needs to possess the 'local knowledge' about how this environment functions (Geertz 1983). Considering this, it is not surprising that most models of social competence place a strong emphasis on social understanding as a powerful antecedent to competent social action (Crick and Dodge 1994, Rubin and Rose-Krasnor 1986).

3.2 The Dyad

Peer experiences at the level of the dyad can be conceptualized as reflecting either interactions or relationships. According to Hinde (e.g., 1979, 1987, 1995) interactions are the series of interchanges that occur between individuals over a limited time span. Interactions are defined as interdependencies in behavior (Hinde 1979). That is, two persons are interacting when their behaviors are mutually responsive. If a person addresses someone (e.g., says 'Have a nice day') and the other person responds ('Don't tell me what to do') they have had an interaction, albeit a rather primitive one. Relationships are based on these patterns of interaction. The relationship consists of the cognitions, emotions, internalized expectations, and qualifications that the relationship partners construct as a result of their interactions with each other.

3.2.1 Basic aspects of friendship. Researchers can examine dyadic relationships, such as friendship, according to (a) what the partners do together (i.e., the content of the relationship); (b) the number of different activities in which the partners engage (i.e., the breadth of their interactions); and (c) the quality of the interactions within the relationship (e.g., reciprocal, complementary, positive, negative). As in the research on popularity in which associations between behavior and popularity were used as the basis of conclusions regarding competence, research on friendship, a primary form of relationship experience for children, informs us of the processes that are central to engagement at the dyadic level. In research on friendship, different characteristics of relationships can be contrasted empirically, for example, between friends and acquaintances or disliked others.

3.2.2 Conflict. One domain of interaction, which nicely illustrates the nature of competence in friendships, is that of conflict. Conflict is one of the few dimensions of interaction in which there are no differences between friends and non-friends. Researchers have shown repeatedly that during childhood and adolescence, pairs of friends engage in the same amount of conflict as pairs of non-friends (Hartup 1992, Laursen et al. 1996, Newcomb and Bagwell 1995). There is, however, a major difference in the conflict resolution strategies that friends and non-friends adopt. In a large meta-analysis of research on friendship, Newcomb and Bagwell showed that

friends were more concerned about achieving an equitable resolution to conflicts than achieving outcomes that would favor one partner. More specifically, Laursen et al. and Hartup reported that friends are more likely than nonfriends to resolve conflicts in a way that will preserve or promote the continuity of their relationship. In this respect, competence in friendship appears to require the ability of the two partners to find a balance between individual and communal goals.

3.2.3 Self and other. A significant feature of this interpretation of these findings is its implicit recognition that competence in friendship is based on achieving a balance between personal desires against social consequences. This emphasis reflects an essential duality of self and other, placing the individual within a social and personal context. The conceptualization of self and other as interdependent is an important feature of theory from several domains including personality theory (Bakan 1966, Sullivan 1953), gender role theory (Leaper 1994), feminist psychology (Jordan et al. 1991) and philosophy (Harré 1979, 1984). This view of the self as inseparable from the other underscores the complex goals that persons confront as individuals and as members of dyads. In this way competence in friendship involves resolving the tension between agentic/individual and communal orientations.

3.3 The Group

Group experiences refer to experiences among a set of individuals who have been organized by either formal or informal means. Group phenomena do not refer to individuals but instead refer to (a) the links that exist among the persons in the group, and (b) the features that characterize the group. Groups can be measured according to (a) their structure and size (see Benenson 1990), and (b) the themes or the content around which groups are organized (see Brown 1989). Many of the best known studies of children and their peers were concerned with the group *per se*, including Lewin et al. (1938) study of group climate, and Sherif et al. (1961) study of intragroup loyalty and intergroup conflict.

3.3.1 Studying the group. In spite of the apparent importance of phenomena at the level of the group, group measures have received far less attention than measures conceptualized at the level of the dyad and the individual. This is surprising because persons generally refer to the peer system as a group phenomenon. Indeed references to experiences with peers invariably refer to the 'peer group.' Moreover, as Hartup (1983) noted, early theory and research on the peer system was organized around the idea that group experiences were powerful determinants of behavior. Nevertheless, research on the factors related to the peer group *per se* has been relatively sparse. In spite of a relative lack of research, the studies regarding peer group behavior generally converge on the same conclusion. This conclusion is that groups that show the most effective engagement with the environment (e.g., those that can complete a task most efficiently) are those that have the strongest sense of cohesion and shared purpose (e.g., Sherif et al. 1961) or that have the most engaged leadership (Lewin et al. 1938). These studies show that a factor that may be critical for an individual to be competent within a group is the shared knowledge or understanding of a task or ritual necessary for a group to function toward a common goal.

4. Conclusion and Future Directions

Social competence refers to an individual's capacity to achieve goals at the individual, dyadic, and group levels. Being competent means that one can balance one's own goals with those of one's partners and group members. Competent behavior is predicated on social cognitive processes underlying the encoding and interpretation of social problems and the development, execution and evaluation of plans to solve problems. The socially competent child and adolescent achieves his/her own goals while promoting the goal-oriented actions of their friends and companions. In this regard they have enduring positive relationships with others. Research on the development of competence needs to be focused on how children resolve the demands required for functioning at the individual, dyadic, and group levels of social complexity. Consistent with definitions of competence, researchers need to treat competence as a process, rather than as a static individual entity. An emphasis on action and process is necessary to reveal the integrative mechanisms that underlie a child's 'competence' in social and personal domains.

See also: Friendship: Development in Childhood and Adolescence; Intrinsic Motivation, Psychology of; Motivation and Actions, Psychology of; Shyness and Behavioral Inhibition

Bibliography

Abelson R P 1981 The psychological status of the script concept. *The American Psychologist* **36**: 715–29

Attili G 1989 Social competence versus emotional security: The link between home relationships and behavior problems in preschool. In: Schneider B H, Attili G, Nadel J, Weissberg R P (eds.) *Social Competence in Developmental Perspective*. Kluwer Academic Publishers, Boston

Asher S R, Coie J D (eds.) 1990 *Peer Rejection in Childhood*. Cambridge University Press, New York

Bakan D 1966 *The Duality of Human Existence*. Rand McNally, Chicago

Benenson J F 1990 Gender differences in social networks. *Journal of Early Adoloescence* **10**: 472–95

Bronfenbrenner U, Crouter A C 1983 The evolution of environmental models in developmental research. In: Hetherington E M (ed.) *Handbook of Child Psychology: Vol. 1. History, Theories and Methods,* 4th edn. Wiley, New York, pp. 357–414

Brown B B 1989 The role of peer groups in adolescents' adjustment to secondary school. In: Berndt T J, Ladd G W (eds.) *Peer Relationships in Child Development*. Wiley, New York, pp. 188–216

Bukowski W M, Bergevin T, Sabongui A, Serbin L 1998 Competence: The short history of the future of an idea. In: Pushkar D, Bukowski W, Schwartzman A, Stack D, White D (eds.) *Improving Competence Across the Life Span*. Plenum, New York, pp. 91–100

Crick N R, Dodge K A 1994 A review and reformulation of social information processing mechanisms in children's social adjustment. *Psychological Bulletin* **115**: 74–101

Dodge K A 1986 A social information processing model of social competence in children. In: Perlmutter M (ed.) *Minnesota Symposium on Child Psychology*. L. Erlbaum Associates, Hillsdale, NJ, Vol. 18, pp. 77–125

Dodge K A, Asher S R, Parkhurst J T 1989 Social life as a goal coordination task. In: Ames C, Ames R (eds.) *Research on Motivation in Education*. Academic Press, San Diego, CA, Vol. 3, 107–35

Erikson E H 1953 Growth and crises of the healthy personality. In: Kluckhorn C, Murray H A, Schneider D M (eds.) *Personality in Nature, Society, and Culture*, 2nd rev. and enl. edn. Knopf, New York, pp. 185–250

Fenichel O 1945 *The Psychoanalytic Theory of Neurosis*. W. W. Norton, New York

Ford M E 1982 Social cognition and social competence in adolescence. *Developmental Psychology* **18**: 323–40

Geertz C 1983 *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York

Goldfried M R, D'Zurilla T J 1969 A behavior analytic model for assessing competence. In: Spielberger C D (ed.) *Current Issues in Clinical and Community Psychology*. Academic Press, New York, pp. 151–98

Gresham F M 1981 Validity of social skills measures for assessing social competence in low status children: A multivariate investigation. *Developmental Psychology* **17**: 390–8

Harré R 1974 The conditions for a social psychology of childhood. In: Richards M P M (ed.) *The Integration of a Child into a Social World*. Cambridge University Press, Cambridge, UK

Harré R 1979 *Social Being*. Rowman and Littlefield, Totowa, NJ

Harré R 1984 *Personal Being*. Harvard University Press, Cambridge, MA

Hartup W W 1983 Peer relations. In: Hetherington E M (ed.) *Handbook of Child Psychology: Vol. 4. Socialization, Personality and Social Development,* 4th edn. Wiley, New York, pp. 103–96

Hartup W W 1992 Conflict and friendship relations. In: Shantz C U, Hartup W W (eds.) *Conflict in Child and Adolescent Development*. Cambridge University Press, Cambridge, UK, pp. 185–215

Hawley P H 1999 The ontogenesis of social dominance: A strategy-based evolutionary perspective. *Developmental Review* **19**: 97–132

Hendrick I 1942 Work and the pleasure principle. *Psychoanalytic Quarterly* **11**: 33–58

Hinde R A 1979 *Towards Understanding Relationships*. Academic Press, London

Hinde R A 1987 *Individuals, Relationships and Culture*. Cambridge University Press, Cambridge, UK

Hinde R A 1995 A suggested structure for a science of relationships. *Personal Relationships* **2**: 1–15

Jordan J V, Kaplan A, Miller J B, Steiver I, Surrey J 1991 *Women's Growth in Connection: Writings from the Stone Center*. Guilford Press, New York

Laursen B, Hartup W W, Koplas A L 1996 Toward understanding peer conflict. *Merrill-Palmer Quarterly* **42**: 76–102

Leaper C 1994 Exploring the consequences of gender segregation on social relationships. In: Leaper C (ed.) *Childhood Gender Segregation: Causes and Consequences*. Jossey-Bass, San Francisco, CA, pp. 67–86

Lewin K, Lippit R, White R K 1938 Patterns of aggressive behavior in experimentally created 'social climates.' *Journal of Social Psychology* **10**: 271–99

McFall R M 1982 A review and reformulation of the concept of social skills. *Behavioral Assessment* **4**: 1–33

Miller G A, Galanter E, Pribram K H 1960 *Plans and the Structure of Behavior*. Henry Holt, New York

Nelson K 1981 Social cognition in a script format. In: Flavel J H, Ross L (eds.) *Social Cognitive Development*. Cambridge University Press, New York, pp. 97–118

Newcomb A F, Bagwell C L 1995 Children's friendship relations: A meta-analytic review. *Psychological Bulletin* **117**: 306–47

Newcomb A F, Bukowski W M, Pattee L 1993 Children's peer relations: A meta-analyic review of popular rejected neglected controversial and average sociometric status. *Psychological Bulletin* **113**: 99–128

Olweus D 1977 Aggression and peer acceptance in adolescent boys: Two short-term longitudinal studies of ratings. *Child Development* **48**: 1301–13

Oppenheimer L 1989 The nature of social action: Social competence versus social conformism. In: Schneider B H, Attili G, Nadel J, Weissberg R P (eds.) *Social Competence in Developmental Perspective*. Kluwer Academic Publishers, Boston, pp. 40–70

Pepler D J, Rubin K H (eds.) 1991 *The Development and Treatment of Childhood Aggression*. L. Erlbaum Associates, Hillsdale, NJ

Rubin K H, Buowski W M, Parker J G 1998 The peer system: Interactions, relationships and groups. In: Damon W (series ed.), Eisenberg N (Vol. ed.) *The Handbook of Child Psychology*, 5th edn. J. Wiley, New York

Rubin K H, Rose-Krasnor L 1986 Social cognitive and social behavioral perspectives on problem solving. In: Perlmutter M (ed.) *Minnesota Symposium on Child Psychology*. L. Erlbaum Associates, Hillsdale, NJ, Vol. 18, pp. 1–68

Rubin K H, Rose-Krasnor L 1992 Interpersonal problem solving. In: Van Hassett V B, Hersen M (eds.) *Handbook of Social Development*. Plenum, New York, pp. 283–323

Schank R C, Abelson R P 1977 *Scripts, Plans, Goals, and Understanding*. Erlbaum, Hillsdale, NJ

Schneider B H, Attili G, Nadel J, Weissberg R P (eds.) 1989 *Social Competence in Developmental Perspective*. Kluwer Academic Publishers, Boston

Sherif M, Harvey O J, White B J, Hood W R, Sherif C W 1961 *Inter-group Conflict and Cooperation: The Robbers Cave Experiment*. University of Oklahoma Press, Norman, OK
Strayer F F 1989 Co-adaptation within the early peer group: A psychobiological study of social competence. In: Schneider B H, Attili G, Nadel J, Weissberg R P (eds.) *Social Competence in Developmental Perspective*. Kluwer Academic Publishers, Boston, pp. 145–74
Sullivan H S 1953 *The Interpersonal Theory of Psychiatry*, 1st edn, Norton, New York
Waters E, Sroufe L A 1983 Social competence as a developmental construct. *Developmental Review* **3**: 79–97
White R 1959 Motivation reconsidered: The concept of competence. *Psychological Review* **66**: 297–333
Zigler E 1973 Project Head Start: Success or failure? *Learning* **1**: 43–7

W. M. Bukowski, K. H. Rubin, and J. G. Parker

Social Constructivism

In its original version, social constructivism is a view about the social nature of science. It rests on the methodological assumption that a sociological analysis of science and scientific knowledge can be empirically fruitful and epistemologically illuminating. This approach has generated detailed empirical studies of scientific practices (for instance, of what is going on in laboratories on a day-to-day basis). According to social constructivism, these studies show that it does not depend exclusively on the objective external world which scientific beliefs are held to be true or false, and thus, what are the scientific facts, but rather also (or even mainly or exclusively) on social arrangements resulting from negotiations between scientists taking place in the course of scientific practices. It is in this sense that scientific knowledge and scientific facts are supposed to be socially constructed. Social constructivism is not a unique specified doctrine, however, but rather a bunch of related studies representing different versions of the general approach.

1. *Historical Background*

One of the guiding ideas of social constructivism is that science, scientific knowledge, and scientific practices are *socially determined*. Historically, this idea is rooted in the historical materialism developed by Marx. Historical materialism is best understood as a research program, roughly in the sense of the methodology of research programs proposed by Lakatos; the core of historical materialism consists of the claim that the economic base of a society determines its superstructure. This core is supposed to generate testable historical theories or hypotheses about what are, in various societies, the base and the superstructure, and about the specific ways in which they are related. The strongest, but not the only possible, interpretation is that the superstructure is causally determined by the base. This approach was taken up by sociologies of science worked out during the twentieth century. Thus, for instance, Karl Mannheim wanted to show how humanities and theories of society are socially determined; and one of the leading American sociologists of science, Robert Merton, suggested that it is primarily the rise and success of science that is determined by social forces (Merton 1973).

Another important intellectual source of social constructivism is the historical turn of twentieth century philosophy of science (Kuhn 1962, Bachelard 1937, Canguilhem 1966, Foucault 1969, Hacking 1990). This movement is vigorously opposed to scientific realism and rejects the claim, so crucial for the traditional philosophy of science, that there are universal rational scientific methods and rules such that the history of science can be rationally reconstructed as an continuous intellectual enterprise constantly improving our scientific knowledge by applying these methods. Instead, proponents of the movement see the history of science and scientific beliefs shattered by contingent breaks and dependent on normative attitudes and hidden non-scientific assumptions that an adequate philosophy of science must aspire to reveal.

A third theoretical background of social constructivism is the program of naturalizing epistemology initiated by Quine (Quine 1969). The crucial idea of this programme is that traditional epistemology, centering around the question of the justification of claims to (scientific) knowledge, should be replaced by a scientific study of causal processes of belief formation and information. In this way, epistemology is supposed to be transformed into a natural science.

In sum, social constructivism puts together three ideas, borrowed from a broadly Marxist sociology of science, from the historical turn of twentieth century philosophy of science, and from the program of naturalizing epistemology, that the development of scientific knowledge is (a) determined by social forces, (b) essentially contingent and independent of rational methods, and should (c) be analyzed in terms of causal processes of belief formation.

2. *The Edinburgh School of the Sociology of Science*

The first prominent version of social constructivism is the Edinburgh school of the sociology of science. The proponents of this school start out by criticizing the Marxist sociology of science for the exclusion of mathematics and natural sciences as a proper subject of the sociology of science. The extended claim is that mathematics and natural sciences are socially determined too, and it is this extension that is fun-

damental to the further claim of many social constructivists that, in a sense, the world as such is socially constructed (Bloor 1991; Knorr-Cetina 1981; Latour and Woolgar 1986). Specifically, it is, according to the Edinburgh sociologists of science, not only the historical development of science, its rise and success, that is influenced by social forces; rather, it is the content of accepted scientific beliefs that is determined by social factors or by the social interests involved in scientific practices.

Edinburgh sociologists of science conclude that there is no definite unique set of rational methods that guide scientific practices which can be referred to in order to explain how scientific results and beliefs are established. Another result is that it is not helpful to make a sharp distinction between the context of discovery and the context of justification in sociological investigations of scientific practices. In this way the Edinburgh school rejects fundamental assumptions held by the traditional philosophy of science.

If scientific practices and the development of science cannot be rationally reconstructed, then the sociological analysis of scientific practice must basically be a causal explanation of belief *formation* and belief *validation*; it is supposed to show in detail how specific scientific beliefs are established as a result of a causal process proceeding from social conditions, social interests, and negotiations between scientists guiding the specific scientific practice under consideration. Therefore, social constructivism is, in the view of the Edinburgh school, itself a kind of natural science. The domain of this sort of science is the set of all scientific beliefs that get accepted for the time being, independently of whether they are true or false and whether they ultimately prove to be successful or not.

Some constructivists focus not on belief formation, but rather on scientific facts: they want to know in which way such facts are hardened, i.e., get accepted as facts by all researchers in the field (Latour and Woolgar 1986). However, focusing in this way on scientific facts instead of scientific beliefs does not make much difference if scientific facts are supposed to be what accepted scientific beliefs are about—roughly the content of statements showing up in established scientific textbooks on the field (Quine 1969). Nevertheless, it is precisely by shifting the focus from beliefs to facts that some social constructivists are inclined to entertain an *ontological* version of social constructivism that sees scientists as essentially contributing to the existence of scientific facts. Obviously, this claim has idealistic implications only if scientific facts are understood in a strictly realistic fashion—roughly as the way things are independently of the way we take them to be.

Social constructivism, as proposed by the Edinburgh sociologists, can also be formulated as a methodological claim mandating a naturalistic approach to scientific practices; as such, social constructivism is best understood as being part of a naturalized epistemology concentrating on the investigation of the social causes of belief formation. In the methodological version, social constructivism does not seem to be committed to ontological claims; however, it remains committed at least to the epistemological claim that scientific knowledge cannot simply be seen as a good representation of the external objective world, but must rather be taken as a result of an extremely complex process involving mainly, and most importantly, social causes. In this way, social constructivism is not only a form of anti-rationalism, but also a form of anti-realism and in this twofold sense, a form of relativism.

3. *The Actor-network Theory*

The actor-network theory (Latour and Woolgar 1986, Latour 1987) is a form of constructivism that rejects the idea of a social determination of scientific knowledge, prominent in the Edinburgh school, mainly for the reason that the social is barely better understood than the natural. The leading thought is that scientific knowledge is an effect of established relations between objects, animals, and humans engaged in scientific practices. An *actor* is, according to this theory, everything that in some causal way affects the production of scientific statements and theories: not only scientists, but also, for instance, background assumptions, methodologies, techniques, social rules and institutions, routines, experiments, measurements and the appropriate instruments, scientific texts and, last but not least, external objects. For an entity to be an actor in this sense it is obviously not required to have contentful mental states, but to be able to perform actions as a kind of behavior describable under some intention. Thus, there can be many sorts of relations and interactions between actors; in particular, some actors can transform other actors (these transformations are sometimes called translations). A *network* is a set of actors such that there are relations and translations between the actors that are stable, in this way determining the place and functions of the actors within the network. Once a network has been established it implies a sort of closure that prevents other actors or relations from entering the network, thereby opening the possibility of the accumulation of scientific knowledge that is taken to be the result of translations within the network. Establishing a scientific belief, theory, or facts comes down, from the point of view of the actor-network theory, to placing these actors in a stable network. In this sense, scientific beliefs, knowledge, theories, and facts are taken to be constructed by translations taking place in established networks.

The actor-network theory shares a number of basic assumptions with social constructivism as conceived in the Edinburgh school. Thus, both approaches entertain a naturalistic account of scientific practices,

do not presuppose a distinction between true or successful and false or unsuccessful scientific beliefs, and reject the possibility of a rational reconstruction of scientific practices and their outcomes. But constructivism in the sense of the actor-network theory is social not in the strong sense that social forces that are presupposed to exist largely independently of scientific practices have a causal impact on these practices; but rather in the extremely weak sense that as a *result* of processes taking place in networks, a scientific claim can eventually be developed about a distinction between the natural and the social, and consequently also about the function of the social for scientific practices (Pickering 1992).

Social constructivism usually does not hold that, in the course of scientific practices, scientific facts of the external world are literally constructed of some other entities. The crucial idea of a (social) construction of scientific knowledge and scientific facts is, rather that an analysis of the process and history of scientific belief formation will not be able to show that the methods of science continuously increase the probability that scientific beliefs will be good representations of an independent external world, and should not even try. Instead, scientific belief formation should be modeled in terms of very different factors, mainly social ones like rules, techniques, institutions, power relations, and negotiations, which affect scientific belief formation in a causal way that can be studied empirically and sociologically

One of the most debated applications of social constructivism is the claim held by many feminists that gender is socially constructed (see *Comparable Worth in Gender Studies*, *Feminist Political Theory and Political Science*). Originally, the most prominent feminist theories introduced a distinction between a naturalistic notion of sex and a social notion of gender to denounce all attempts of deriving properties of gender from properties of sex as mere constructions of gender that do not refer to reality. More recently, in some postmodern feminist accounts of sex and gender, it is also the very distinction between sex and gender, and thus, sex itself, that are supposed to be socially constructed.

4. Social Constructivism About the Social

In one of the most recent developments of social constructivism, leading proponents of the Edinburgh school try to exploit the thought that scientific knowledge is, after all, something like an institution and that institutions are constructed by humans (Bloor 1996). It seems to follow that scientific knowledge is, because constructed by humans, socially constructed. This idea is part of what is often called the intentionalist program of social ontology, the aim of which is to clarify the ontological status of social entities, i.e., groups and of institutions. The fundamental claim is that it is on the basis of intentional mental states and associated actions that social entities come into existence. That is to say that human beings construct social entities by having specific contentful thoughts and by performing intentional actions.

It is obvious that this approach is a sort of social constructivism in the sense that it is a theory about the construction of the social; this is held to be so insofar as it consists largely of normative interactions between humans. It must be emphasized that this sort of constructivism is restricted to the realm of the social, that it presupposes a sharp distinction between the natural and the social, and that it has nothing to say about the construction of natural facts or physical entities. Most importantly, though, the intentionalist program relies on an intentional vocabulary. In all these respects the program differs fundamentally from traditional versions of social constructivism.

Basically the intentionalist program of social ontology is an individualistic account of the social; at the same time, however, one of the crucial theoretical moves is to introduce notions of collective intentionality and of collective actions. The first step is to assume that for some persons to be socially related in a specific way is to be treated by most of the other members of their society as being related in this way. Social relations are taken to imply rules as well as certain rights. In general, therefore, to treat some persons as socially related in way S is to treat them as being entitled or committed to perform actions according to the rules and rights derivable from S.

One of the next important moves is to introduce the notion of a collective attitude (or, as it is sometimes called, a we-attitude and, in particular, the notion of a collective intention to perform a specific action. Roughly speaking a person P as member of a set S of persons has a contentful collective attitude of the sort A iff (a) P has the A-attitude herself, (b) P believes that every other member has the A-attitude, and (c) P believes that every other member M of S believes that every member of S other than M has the A-attitude. In particular, if a number of persons P_i intend to perform action A together or collectively, then it must be the case, in addition to (a)–(c), that every P_i intends to perform A by performing a specific action A_i as her specific contribution to the performance of A such that P_i thinks that A_i is necessary for A and results, together with the other specific actions A_j intended to be performed by the other P_j, in performing A, and that P_i believes that the other P_j intend to do the same thing and believe themselves that every other person intends to do the same thing.

On the basis of these notions it can be asserted that a number of persons P_i have the collective intention to perform action A iff each P_i has the collective intention to do A by performing a specific A_i, in the sense just explained; and P_i perform a collective action A iff A can be described under some collective intention of the P_i. It might be suggested that a number of

persons P_i represent a social group iff they share a collective attitude. Finally, let MR be a set of sematically consistent rules R_i such that a number of persons collectively treat an X as a Y by following R_i, then MR is, in the most basic sense, an institution. In particular, if MR is the set of rules according to which scientific results and scientific knowledge are established, and if the persons in question are trained scientists, then MR is an institution of epistemic or scientific practices, and in this sense, scientific knowledge might be called an institution that is constructed by intentional acts of human beings (Searle 1995, Tuomela 1995).

This line of argument is a rough sketch that by no means exhausts the theoretical complexity of the intentionalist program. For instance, something has to be said about the mechanisms of the formation of collective attitudes which account for the stability of social groups and institutions; at this point, a theory of consensus and a theory of power has to be incorporated into the general account. In any case, even a rough sketch of this sort of social ontology shows how, to put it very generally, the mind can be thought of as bringing into existence, and keeping in existence (and thus to construct) social entities in a fairly literal manner. It also becomes clear how and why traditional social constructivists might be inclined to render their position more precise by exploiting the intentionalist program of social ontology.

This program is a philosophically sound approach; it makes evident, however, that human beings do not construct social facts as they construct natural facts. Traditional versions of social constructivism tend to downplay this distinction, which might be thought of as a fundamental flaw. Another problem with these versions is that they seem to commit themselves to relying exclusively on a causal, physicalist vocabulary. While it certainly makes sense to examine whether a purely naturalistic account of epistemic practices might be fruitful, it does not follow from this strategy, even if it proves fruitful, that intentionalist strategies must be excluded. On the contrary, it is doubtful whether epistemic practices can be adequately described without also using intentionalistic and normative vocabulary. Finally, it is hard to see how traditional social constructivists can overcome Davidson's point that, according to interpretation theory a basic form of rationality is constitutive for mastering a natural language and must therefore be the same in all epistemic practices (Davidson 1973); obviously, this insight is inconsistent with the unqualified contingency claim about the development of scientific knowledge which is so central to traditional social constructivism.

See also: Foucault, Michel (1926–84); History of Science: Constructivist Perspectives; Interactionism: Symbolic; Scientific Knowledge, Sociology of; Technology, Social Construction of

Bibliography

Bachelard G 1937 *Le nouvel esprit scientifique*. Alcan, Paris
Bloor D 1991 *Knowledge and Social Imagery*. University of Chicago Press, Chicago
Bloor D 1996 *Idealism and the Sociology of Knowledge*. *Social Studies of Science* **26**: 839–56
Canguilhem G 1966 *Le Normale et le Pathologique*. PUF, Paris
Davidson D 1973 Radical interpretation. *Dialectica* **27**: 313–28
Foucault M 1969 *L'archéologie de Savoir*. Gallimard, Paris
Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK
Knorr-Cetina K D 1981 *The Manufacture of Knowledge: An Essay on the Constructivist Contexualist Nature of Science*. Pergamon Press, Oxford
Knorr-Cetina K D (ed.) 1982 *Science in Context*. Sage, London
Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
Latour B 1987 *Science in Action: How to follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge, MA
Latour B, Woolgar S 1986 *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, NJ
Merton R K 1973 *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press, Chicago
Pickering A (ed.) 1992 *Science as Practice and Culture*. University of Chicago Press, Chicago
Quine W V 1969 *Ontological Relativity and Other Essays*. Columbia University Press, New York
Searle J R 1995 *The Construction of Social Reality*. The Free Press, New York
Tuomela R 1995 *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford University Press, Stanford, CA

W. Detel

Social Democracy

The term 'social democracy' can refer either to a form of societal organization, an ideology, a set of public policies, a political party or group of parties, or a broad social and political movement centered on such parties and their allies within different sectors of society. The meaning of 'social democracy' has varied enormously since its original adoption by a small, vaguely-socialist group within the left wing of the French democratic-republican opposition prior to the Revolutions of 1848. At various times and places, the social-democratic label has been adopted by revolutionary Marxist, democratic-socialist, liberal/center-left, and even conservative-authoritarian political parties and tendencies, usually but not always with closer or more distant ties to blue-collar labor unions. Most commonly, the term 'social democracy' is used either in connection with political parties that have belonged to the contemporary Socialist Inter-

national or its predecessor, the Second International of 1889–1914, or with reference to policies and ideologies close to such parties.

As parties calling themselves social-democratic have grown steadily more moderate over time, gradually becoming less and less anticapitalist and antiliberal, usage of the term 'social democracy' has evolved in a similar fashion. At the turn of the twenty-first century, 'social democracy' generally refers either to the programs and policies of the largely center-left political parties that belong to the Socialist International, or to a vague set of essentially liberal, center-left ideological and policy orientations that are often associated with these parties. While specific policies may vary, such center-left 'social-democratic' approaches share a commitment to a relatively free-market-based capitalist economy, a fairly developed welfare-state, a liberal-democratic ('polyarchic') political system, and a relatively internationalist approach to questions of foreign policy, security, and globalization. All over the world, parties and ideologies of this type are currently among the most important in the vast majority of countries with relatively competitive elections. The term 'social democracy' is rarely any longer used as a self-description by Marxist or other anticapitalist and antiliberal tendencies.

1. *The History and Significance of Social Democracy*

Ever since the Revolutions of 1848, which are generally considered to mark the birth of the social-democratic movement, the history of social democracy has been characterized by a profound ideological and tactical struggle between relatively radical, anticapitalist tendencies and relatively moderate, reformist orientations. In this dialectical contest, the radicals have generally adhered to some form of State socialism, usually of a Marxist (but anti-Leninist) variety, and/or some version of decentralized and radical-democratic cooperative socialism ('self-management,' 'economic democracy'); the comparatively moderate reformists, by contrast, have increasingly made their peace with multiclass ('catch-all') parliamentary politics, liberal ideology, and the capitalist system. Over the decades, this internal struggle has interacted with the often complex relationships of social-democratic movements with other political, social, and economic forces within particular societies to give rise to a wide variety of more or less 'socialist' ideological visions and public policies, often very effective and/or highly original, which have directly or indirectly exerted a tremendous influence on virtually all aspects of political, social, economic, and intellectual life in the modern world, particularly in countries with relatively capitalist economic systems and relatively liberal-democratic political systems. Although social-democratic movements and governments have frequently been frustrated in their reformist efforts by powerful domestic or international pressures, the resulting compromises and modifications have often marked significant new developments in the history of their respective countries, and of the world as a whole. Indeed, even when confined to the ranks of the political opposition in a particular country, social democrats have sometimes played a major role in bringing about political, social, or economic reforms, by means of political pressure, foreign example, or intellectual influence.

Among the most important of these social-democratic contributions to the development of the 'welfare-capitalist' synthesis that has gradually come to dominate an increasing portion of the world since the 1880s, have been the following: (a) the striking (albeit gradual) democratization and humanization of political, social, and economic life; (b) the enormous growth and modernization of both the state and the economy, particularly in the realms of social policy, industrial policy, macroeconomic policy, industrial and labor relations (the much-celebrated 'democratic corporatism' of various Northern European countries, now largely defunct in a more market-oriented age), and recently also environmental policy; (c) the more or less successful promotion of social harmony and ideological consensus within previously much more polarized societies; and (d) strong support for international integration and world peace (although the temptation to join in nationalistic ventures, on the one hand, or in pacifistic crusades, on the other hand, has been a frequent source of tension within the social-democratic movement ever since its inception, most notably during the First World War).

Although self-styled 'social-democratic' parties and ideas have played a leading role in the history of an increasing number of countries around the world since 1880s, the concept of 'social democracy' is generally couched in terms based on the experience of Northern European Social Democratic, Socialist, and Labor parties (particularly those of Sweden and the other Scandinavian countries, as well as Germany, Austria, Great Britain, the Netherlands, and Belgium), because they have usually been the leading forces associated with the 'social-democratic' label, whether in terms of ideological and tactical development, policy-innovation, electoral support, governmental influence, or international prestige.

To the extent that world history since 1848 has been the story of the grand struggle between capitalism and socialism, social-democratic movements and ideas have often played a decisive role in finding a more or less successful middle-ground between the two (frequently under the slogan of some sort of 'Third Way' between unbridled capitalism and state-socialism), particularly in Sweden, Norway, and the other countries of Northern Europe. And to the extent that world history during the same period has been the story of the grand struggle between democracy and

dictatorship, social-democratic movements and ideas have contributed greatly to both sides: on the one hand, social democracy spawned radical, antiliberal offshoots that quickly evolved into the rival totalitarian movements of communism and fascism during and after the First World War; on the other hand, social democrats have often played a decisive role in resisting and defeating these totalitarian movements, as well as other antidemocratic political, social, economic, and intellectual forces.

From the 1945 onwards social-democratic parties all over the world have grown steadily more dependent on the success of the welfare-capitalist social-market economy and the support of middle-class voters and relatively privileged elites in both the public and private sectors. This has been the result of two major interrelated processes: first, the continuation of these parties' aforementioned gradual movement toward the liberal center of the ideological spectrum; and second, the search for new bases of support and new reasons for existence under the impact of the gradual economic and cultural disappearance and disaffection of these parties' traditional blue-collar industrial ('proletarian') base. The first of these processes, that of increasing ideological moderation, has largely been due to the combination of mass prosperity, the rise of the welfare-state, profound political and social democratization, the failure of Marxist ideas and State socialist regimes ('real-existing socialism'), the spread of international peace and integration, American political and cultural hegemony, and the all-but-inevitable habituation of social-democratic elites to existing power-structures. The second process, namely the disappearance and disaffection of social democracy's old blue-collar base, has stemmed primarily from the combination of four interrelated developments that are associated with relentless economic and social modernization: first, much greater levels of social and economic mobility and opportunity, partly as a result of successful social-democratic policies; second, the accelerated disintegration and dissolution of inherited social and cultural solidarities and cleavages, including class-consciousness ('individualization,' 'desegregation,' 'depillarization'); third, the drastic weakening of inherited attitudes, beliefs, and values ('postmodernization,' the rise of 'postmaterialist' or 'left-libertarian' values), particularly among highly-educated elites; and fourth, the transition from a predominantly 'Fordist' socioeconomic system based on mass production in 'smokestack' industries, to an increasingly 'postindustrial' economy based on much more sophisticated technology, a much older, more highly-skilled, and more highly-educated labor-force, a burgeoning service sector, much more widespread share-ownership, much higher levels of flexibility in all areas of economic and social life, much greater concern for environmental protection, and much higher levels of international interdependence, integration, and homogenization ('globalization').

2. *The Challenges Facing Social Democracy Today*

In the twentieth century, social-democratic parties and elites increasingly staked their claim to rule less on utopian visions of 'socialism' than on their ability to manage and modernize the welfare-capitalist social-market economy more fairly, more humanely, more democratically, and, above all, more competently and more efficiently, than their rivals on right and left. Yet, although egalitarian and statist alternatives to free markets and a fairly rapacious capitalism were in the year 2000 weaker and more uncertain than ever before, challenges from the populist and often vehemently antiliberal and antimodern right, which since the 1980s has increasingly drawn the support of traditionally social-democratic (or communist or Christian Democratic) working-class voters, are often close at hand. Thus social democracy is faced with the difficult task of reconciling the interests and aspirations of an economically and culturally ever-more diverse potential constituency, a task which may prove extremely difficult once the worldwide economic boom at the turn of the twenty-first century dissipates. The danger, then, is that social democracy may become caught between the conflicting demands of relatively successful or privileged and relatively unsuccessful or underprivileged sectors within increasingly unequal and *anomic* postindustrial societies, with the attendant risk that important groups within either or both sectors may gravitate toward antiliberal and antimodern movements on the left or, more likely, the right.

See also: Industrial Policy; Liberalism; Liberalism: Historical Aspects; Political Parties, History of; Socialism; Socialism: Historical Aspects; Welfare State

Bibliography

Anderson P, Camiller P (eds.) 1994 *Mapping the West-European Left*. Verso, London

Berman S 1998 *The Social Democratic Moment: Ideas and Politics in the Making of Interwar Europe*. Harvard University Press, Cambridge

Esping-Andersen G 1985 *Politics Against Markets: The Social Democratic Road to Power*. Princeton University Press, Princeton NJ

Gamble A, Wright T 1999 *The New Social Democracy*. Basil Blackwell, Oxford

Gay P 1962 (1952) *The Dilemma of Democratic Socialism: Eduard Bernstein's Challenge to Marx*. Collier, New York

Giddens A 2000 *The Third Way and Its Critics*. Polity Press, Cambridge UK

Gillespie R, Paterson W (eds.) 1993 *Rethinking Social Democracy in Western Europe*. Frank Cass, London

Kitschelt H 1994 *The Transformation of European Social Democracy*. Cambridge University Press, Cambridge UK

Lindemann A 1983 *A History of European Socialism*. Yale University Press, New Haven
Marx K 1978 (1875) Critique of the Gotha Program. In: Tucker R (ed.) *The Marx–Engels Reader*. W.W. Norton, New York
Merkel W 1993 *Ende der Sozialdemokratie? Machtressourcen und Regierungspolitik im westeuropäischen Vergleich*. Campus, Frankfurt am Main, Germany
Notermans T 1999 *Money, Markets, and the State: Social Democratic Economic Policies Since 1918*. Cambridge University Press, Cambridge
Padgett S, Paterson W 1991 *A History of Social Democracy in Postwar Europe*. Longman, London
Piven F F (ed.) 1991 *Labor Parties in Postindustrial Societies*. Polity Press, Cambridge
Przeworski A 1985 *Capitalism and Social Democracy*. Cambridge University Press, Cambridge
Sassoon D 1996 *One Hundred Years of Socialism: The West-European Left in the Twentieth Century*. The New Press, New York
Scharpf F 1991 (1987) *Crisis and Choice in European Social Democracy*. Cornell University Press, Ithaca

T. P. Harsanyi

Social Demography: Role of Inequality and Social Differentiation

The description and analysis of inequality and social differentiation are fundamental components of social demography. In this context, a socially differentiated population is one characterized by a variety of subgroups that are distinct according to generally persistent social criteria such as race, ethnicity, religion, language, and so on. Inequality, on the other hand, is a hierarchical concept and is used here to refer to differences in the actual possession of resources of various types or in access to such resources. Resources are in the form of political power or influence, income, schooling, health care services, etc. Differentiated groups may also be extremely unequal but the idea of socially different need not imply inequality. This article briefly reviews several of the reasons why demographers study differentiation and inequality and then goes on to discuss how four different theoretical perspectives have been applied to the problem. The first emphasizes cultural explanations; competing somewhat with it are resource-based explanations which focus particularly on differences in opportunities and constraints. Cutting across both of these are reference group theory and life-course analysis.

1. *Introduction*

A perennial demographic question is the extent to which diverse or unequal groups exhibit marked differences in demographic behavior. Are there substantial race, ethnic, or religious differences in marital or fertility behavior, in mortality, or in patterns of migration? One reason it is important to study such differentials is that they indicate the extent to which the overall demographic performance of a population is a function of subgroup diversity. How big an impact a subgroup's behavior can have on overall population measures depends on the distinctiveness of that group's behavior weighted by its relative size. Although size counts for much, if the group's behavior is unusual enough, it can still have a substantial impact on overall measures. For example, using never married women, aged 30–34, as a rough indicator of marriage delays in the US, 22 percent of all women of that age were never married in 1998. However, the proportion was much higher for African American than white women—47 as opposed to 17 percent. The differences were so large that even though African American women were only 14 percent of all women in this age group while whites were 80 percent, the proportion never married was pulled up by five points—or almost 30 percent—for all women as compared to whites (US Bureau of the Census 1998).

Documenting subgroup differences in demographic behavior also provides important insights into the demographic concomitants of inequality and a socially differentiated population. It also poses analytical questions about the underlying determinants of these differences since group membership is usually an indicator of a variety of more proximate causes of an observed demographic pattern—e.g., the role of unequal access to medical care in ethnic group differences in mortality. A classic example in social demography is the influence of marriage timing on fertility in early industrializing populations. Given women's circumscribed reproductive period, delayed marriage in a noncontracepting population typically leads to smaller completed family size, while an earlier marriage promotes higher fertility. Thus, in rural areas in many parts of Western Europe, where inheritance of the farm was a precondition for marriage, marriages were often greatly delayed. On the other hand, factory workers in the cities usually married much earlier as their economic fortunes were tied to the developing wage economy rather than the traditional inheritance system (Anderson 1971).

2. *Varying Theoretical Perspectives*

The remainder of this article will focus on the impact of inequality and social differentiation on demographic behavior by examining four theoretical approaches to the problem. Together these underlie most research in social demography; hence the discussion will begin by a brief introduction to the general thrust of each perspective before applying them to issues of social differentiation and inequality. Two distinct and often competing theoretical approaches

have been particularly influential in social demography; these are cultural compared to resource explanations of demographic behavior. Cross-cutting them are two additional perspectives—reference group theory and life-course analysis.

2.1 Socioeconomic Constraints and Opportunities

Resource based explanations focus on how differences in resources affect demographic behavior—e.g., differences in social or economic constraints and opportunities. Thus, a major feature of the Theory of the Demographic Transition is the argument that economic development raises the cost of children and reduces their economic value to families, thereby promoting the transition from the relatively high fertility characteristic of most preindustrial societies to the typically low fertility of industrialized societies (see *Fertility Transition*: *Economic Explanations*). By focusing on differences in resources, this approach also lends itself well to analysis of the determinants and consequences of inequality.

In the resource-based approach, differences or changes in goals or preferences among individuals or groups are frequently not examined in any great detail. In fact, neoclassical economists have typically argued that preferences are fixed across time and do not systematically vary among individuals, thereby permitting changes in demographic behavior to be unambiguously attributed to changes in constraints. However, for most social demographers, such an assumption is generally too strong and variations in preferences are usually considered an important, if often understudied, source of differences in people's behavior.

2.2 Cultural Factors

A second approach views 'culture' as the most important factor influencing demographic behavior. The nature of cultural factors being considered are highly variable, ranging from the narrowly to the broadly conceived—from norms regarding the permissibility of divorce or the number of children a couple should raise all the way to more complex and sometimes highly abstract constructs such as religious or basic value systems. For example, it has been argued that both earlier and more recent European fertility transitions are largely the result of a major change in the 'ideational' system marked by the growth of secular individualism among other value changes. While some recognition is given to the influence of economic factors, their importance is often minimized by scholars operating in this tradition (see *Fertility Transition: Cultural Explanations*; *Demographic Transition, Second*; *Family Theory: Role of Changing Values*). The cultural approach tends to emphasize *differences* in goals and values and hence seems more oriented towards explaining socially differentiated behavior rather than behavioral differences based on an inequality of resources. However, it is often recognized that certain cultural patterns arise out of conditions of inequality and may even contribute to its persistence (Parsons and Goldin 1989).

2.3 Reference Group Models

Cutting across these general perspectives are two additional approaches that have been influential in explanations of demographic behavior. The first is the use of reference group models. The idea here is that people use selected others whose behavior or characteristics provide positive—or sometimes negative—models for themselves. How this approach is implemented partly depends on whether the researcher's primary orientation is cultural or resource based. If cultural, the emphasis is more likely to be on reference groups as normative models to emulate. If the investigator is more resource oriented, the emphasis is likely to be on relative economic status concerns (see *Social Networks and Fertility*; *Human Ecology: Insights on Demographic Behavior*).

2.4 Life Course Approach

A second influential cross-cutting perspective has been the lifecycle/life course approach. Lifecycle analysis is the pioneering version of this approach and focuses particularly on age-related characteristics and behavior. It can be applied either to an individual—an individual's career or family cycle, for example—or to a group—the family's developmental cycle. Because of the scarcity of life history data, lifecycle analysis has often had to restrict itself to age patterns measured at one or more points in time (using cross-sectional data on the behavior of different age groups, i.e., 'synthetic' cohorts) rather than being able to follow individual life histories over real time. The later developing life course perspective adds a particular emphasis on such life histories, although it too has sometimes had to settle for cross-sectional data. The availability of longitudinal data has increased in recent years, permitting more extensive analyses of life histories. Some life course researchers have also placed great emphasis on normative issues as well, a major one being that there are norms governing the sequencing of life course transitions—for example, that there is a norm against marrying before completing school (Elder 1977, Hogan 1978).

While all these perspectives, singly and in combination, have been generally important in demography, the discussion below will focus on their use in analyses of the role of socioeconomic inequality and differentiation in demographic behavior. Although often viewed as competing, neither the resource nor cultural perspective alone can do an entirely satisfactory job of explaining patterns or changes in demographic behavior. Both approaches profit by incorporating elements from the other.

3. *Cultural Differences and Demographic Behavior*

Probably the most extensive research project which came to espouse a cultural explanation of demographic behavior was the Princeton European Fertility Project which examined the transition from high to low fertility occurring over much of late nineteenth and early twentieth century Europe. By using provincial-level data, the project set out to test several propositions of Demographic Transition Theory, one of which was that fertility declines were a consequence of economic development. A major conclusion of this body of research was that the observed fertility declines appeared to be only weakly related to conventional measures of socioeconomic development, such as the proportions in a province who were literate, employed in agriculture, lived in large cities, etc. In fact, in most European countries the fertility transition appeared to have begun at nearly the same time, despite considerable geographical variability in economic and social conditions. Hence, this too seemed to be a strong indication that the fertility transition was not an adaptive response to the changing conditions brought about by economic development. However, areas which shared similar *cultural* characteristics, such as language and religion, exhibited different rates of fertility change compared to areas at a similar economic level but with different cultural characteristics (Coale and Watkins 1986).

To explain these findings, an innovation and diffusion hypothesis was invoked, involving elements of reference group as well as cultural theories. The argument is that the deliberate control of completed marital fertility was generally unknown in Europe throughout most of its history although large families were not highly desired. However, late in the nineteenth century certain influential opinion leaders in local areas not only became innovators in the use of contraception but also in promoting the view that family size could and should be legitimately controlled. Their views and practices provided a behavioral model and were, accordingly, rapidly adopted by others. Such changes were more likely to occur within cultural groups, defined, for example, by language and religion, partly because of the ease of within-group communication. In addition, shared norms and values increased the influence innovators were likely to have within the group, thereby fostering a rapid diffusion of innovations. Moreover, some cultural groups were more open to change than others. In Belgium, for example, traditional Catholic doctrines were much more powerful in the Flemish-speaking areas than in the French which were more secularized. Throughout much of Europe, France being a notable exception, fertility changes were more rapid in Protestant than Catholic areas (see *Fertility Transition: Cultural Explanations*).

The Princeton Project's contention that the European fertility transition was only weakly related to economic development has been subject to a variety of criticisms. The argument heavily depends on two empirical generalizations—first, that the adoption of fertility control was an innovative behavior and, second, that the fertility transition began at roughly the same time over most of Europe despite considerable regional variations in the pace of industrialization. However, because the kind of data collected during this historical period were severely limited, indirect statistical measures had to be developed as indicators of these behavioral patterns. Recent work on the validity of these measures, some by an original contributor to one of the measures (Trussell), indicates that they will not reliably detect the start of the transition until it is well underway and hence arguments about the simultaneity of transitions all over Europe, regardless of the level of economic development, are not reliable. Moreover, the statistical measure of the proportion of couples who were fertility controllers requires a fairly large number of them in order to be detected by the measure, casting doubt on the ability to determine whether fertility control was actually an innovation at the times indicated (Guinnane et al. 1994). In addition, recent analyses of German data, employing more sophisticated multivariate statistical methods than typically used in the Princeton project, have found that indicators of economic development were strongly related to declines in fertility during this period; however, although a cultural factor, such as religion, was important in predicting fertility *levels*, it did not have an effect on *changes* in fertility (Galloway et al. 1998).

While there is increasing doubt that cultural factors alone can explain the European fertility transition, the continuing social policy goal of reducing fertility in currently developing societies has kept alive a strong interest in more rigorously developing an innovation and diffusion model. However, recent work in this area has placed a greater emphasis on the value of the process as a mechanism for explaining changing fertility patterns in *conjunction* with changing socioeconomic constraints rather than *instead* of them (Montgomery and Casterline 1996). Moreover, in contrast to the provincial-level data the Princeton project was necessarily limited to, detailed microlevel research involving network analysis is being carried

out in several less developed countries to determine the particular social interaction and networking mechanisms involved in promoting the diffusion of contraceptive use and declines in fertility (see *Social Networks and Fertility*; *Human Ecology: Insights on Demographic Behavior*).

In general, although there is little doubt regarding the importance of investigating the effect of cultural differences on demographic behavior, in practice it is not always so easy to pin down the causal role of cultural factors. The easiest but probably also the least sophisticated approach is to assume that if there is a simple correlation observed this constitutes reliable evidence of a causal relationship between demographic behavior and membership in socially differentiated groups such as those defined by religion, ethnicity, social class, metropolitan residence, etc. Sometimes, the existence of cultural determinants is deduced as the residual explanation when other variables, such as socioeconomic ones, do not explain much of the between-group differences. However, to do so makes the unrealistic assumption that all other relevant variables are included in the analysis and are also properly measured. Moreover, given the preliminary nature of some of the theories under investigation as well as the uneven quality of demographic data and weaknesses in current measurement capabilities, attaching specific meanings to unexplained variance is always a risky procedure.

Another problem involved in opting too quickly for purely cultural explanations is that cultural identities are usually related to economic characteristics. Hence multivariate analyses are essential for separating out the role of economic constraints from cultural characteristics. Even then, the nature of the causal process is not always clear. For example, education is more highly stressed among Chinese than Latin American immigrant populations in the United States, suggesting a causal sequence between ethnicity, educational attainment, and socioeconomic status rather than one of these factors alone dominating the process. This may indicate a culturally related higher value placed on education among the Chinese—because of a Confucian tradition, for example—but it may also partly reflect differences in the socioeconomic origins of the two immigrant groups—one having a greater representation of professionals while the other coming from a much poorer socioeconomic background. Hence, sorting out the nature of the causal process is often complex and the data available to accomplish this are not always readily available.

The magnitude of the effect of cultural differentiation on demographic behavior can also be quite different depending on whether one is discussing *levels* or *changes* in behavior. For example, religion usually has historically had a consistently sizable impact on fertility in the cross-section but if the religious composition of an area does not change markedly over time, and the effect of any given religious identification remains constant, religion's role in *changes* in demographic behavior may be minimal. However, if different cultural groups vary in their openness to change introduced from the outside, as is sometimes argued, then cultural differences could indeed have a substantial impact on change.

4. *Socioeconomic Constraints and Opportunities*

Social scientists who primarily focus on the role of resources and constraints in demographic behavior have often ignored how institutional and normative commitments may constrain individual choices, both in terms of affecting preferences and also by limiting what are considered the legitimate means for realizing these preferences. In this respect, neoclassical economists have often chastised sociologists for their apparent willingness to posit a plethora of attitudes and norms to explain every observed behavior, thereby trivializing the analytical process. Nevertheless, it is possible to accept the idea of systematic variations and changes in preferences as a reasonable analytical tool and to apply it in a disciplined fashion in conjunction with resource constraints—that is, without positing a different or changing norm or value to explain every observed variation in demographic behavior. Hence the view that one must choose either a cultural approach *or* a resource constraint approach is unnecessarily confining intellectually.

4.1 *Relative Economic Status Models*

A major example of the integration of the idea of using varying lifestyle aspirations in conjunction with varying resource constraints to explain demographic behavior is found in Richard Easterlin's work on relative cohort size in the US. Easterlin hypothesized that intergenerational shifts in relative cohort size produce intergenerational inequalities in people's economic position which, in turn, affect age at marriage, fertility, and a number of other demographic responses. In addition, he maintained that lifestyle aspirations were learned in the parental household so that children raised in affluent households would develop higher-level consumption goals than those raised in poorer households. He combined these two views by arguing that cohorts who were entering young adulthood in the early postwar period were small due to low fertility during the Great Depression and were also raised in less affluent households, resulting in relatively modest consumption aspirations. However, because of this depression cohort's small size, they were in a very favorable labor market position as adults which encouraged early marriage and childbearing, as well as larger families—in short, they produced the postwar baby boom. The baby boomers, in turn, were raised in these affluent house-

holds and acquired high consumption aspirations; however, because of their large size, their labor-market position was not as good as that of their parents and this led to delayed marriage and lower fertility, resulting in the baby bust. In short, the Easterlin hypothesis posited that a feedback effect was produced by this process and created self-generating demographic cycles. Due to intergenerational shifts in relative economic status and consumption aspirations, small cohorts begat large ones which, in turn, produced small cohorts (Easterlin 1987).

Empirically, the status of Easterlin's theory has been criticized on a number of counts (see *Baby Booms and Baby Busts in the Twentieth Century*). A major one is that the postwar demographic changes were primarily driven by 'period' rather than 'cohort' factors. For example, annual birth, marriage, and divorce rates have tended to fluctuate together among larger and smaller cohorts alike, suggesting that there was something about particular historical periods that influenced behavior rather than cohort size. Nevertheless, the relative economic status aspect of the model remains interesting because the existence of inequality can provide a motivation for adaptive behavioral changes to occur; thus it is an intrinsically dynamic type of model.

Easterlin's hypothesis focused only on *inter*-generational inequalities; however, reference group and relative economic status models can also be applied with*in* generations. One example comes from the rapid postwar rise in US married women's paid employment so that by 1990 around 70 percent of wives, aged 25–54, were in the labor force compared to less then 25 percent in 1950. There is little evidence that this transformation was precipitated by major changes in norms or values. While opinion poles and other attitudinal studies indicate that approval of married women's employment became increasingly positive with time, attitudinal changes did not precede the changes in employment behavior but rather lagged behind them. A far more influential factor appears to have been the postwar economic expansion which greatly increased job opportunities for married women because the demand for women workers in female dominated occupations far exceeded the size of the population group that had traditionally been the backbone of the female labor force—young single women (see Oppenheimer 1970).

In addition to substantial changes in labor-market conditions, a type of innovation and diffusion model combined with a relative economic status approach might further clarify why the employment shift has taken place so rapidly. The rationale is that the rise in married women's employment changes how families are stratified by reducing some previous socioeconomic inequalities and creating others. When married women's employment outside the home was relatively rare, a family's socioeconomic position was primarily dependent on the husband's economic status, although earlier in the twentieth century the work of the family's adolescent and young adult children was also important at certain periods in the family's developmental cycle. As wives' employment became more common in the postwar period, however, many families within the same socioeconomic group, but without working wives, could increasingly be put at a relative economic disadvantage, thereby generating inequalities within groups which had previously been at roughly the same socioeconomic level, based on husbands' economic position alone. On the other hand, wives' employment can *reduce* the relative economic advantage of some families who might previously have been at a somewhat higher socioeconomic level. Hence, wives' employment in some families but not others creates both an objective and subjective instability in how people experience the stratification system. But it also creates a model of how to deal with these instabilities—the entry of other wives into the labor force. Wives already working also provide a source of information on the nature of the job market for others contemplating entering the labor force, as well as providing models of strategies for how wives and mothers could cope with work and family responsibilities.

The impact of this process should, if anything, have increased over time. As long as there were few wives employed, their work would have relatively little impact on both intra- and intergroup reference comparisons and the easier it would be to dismiss their behavior as 'deviant' or the 'exception that proves the rule.' However, as the proportion of wives working rose and became increasingly full-time, then their example would become more salient, thereby increasing its destabilizing effect on reference group comparisons. Moreover, the more wives who worked, the fewer nonworking women there would be around to socialize with, share childcare responsibilities with, etc., hence, the lifestyles as well as the relative socioeconomic position of families with nonworking wives would be increasingly affected. All this should have contributed to the continued rapid expansion of married women's employment in the postwar era in addition to any changes in labor market conditions (Oppenheimer 1982).

Although not easy to operationalize, due to data limitations, relative economic status models have attractive theoretical features for analyzing change. They posit changes in lifestyle aspirations which affect behavior but which nevertheless remain embedded in a stable preference structure—maintaining one's socioeconomic position in the community. They also involve feedback effects that foster a continuation of the diffusion process whose momentum may even increase rather than peter out over time. Furthermore, the existence of feedback effects means that not all those changing their behavior need experience the same initial motivating factors that the earlier innovators were exposed to; whatever its original cause,

the behavior of innovators itself creates an additional set of pressures that influences others, in turn, to shift their behavior and attitudes as well. Nor need such an amalgam of a relative economic status model with an innovation and diffusion model operate independently of social norms. For example, in the American case, although there continued to be a *general* disapproval of married women's employment long after it had already substantially increased, even early on in the process there was evidence of a considerable flexibility in attitudes, depending on the circumstances. Thus attitudes were very negative in the 1940s if it was specified there was only a limited number of jobs but much more permissive if a wife worked in order to facilitate a marriage (Oppenheimer 1970). Hence, the conditionality of many norms permits a certain amount of innovative behavior which may ultimately modify the normative structure more generally. In this case, changing socioeconomic conditions encouraged an increase in wives' employment which was facilitated by norms viewing working to achieve family goals as a legitimate exception to the general disapproval of married women's employment. The ultimate result, however, was not only a permanent change in employment behavior but a radical reduction in the *general* disapproval of wives' working as well.

4.2 Socioeconomic Inequalities and the Life Course

While cross-sectional analyses are major sources of information on socioeconomic inequality and its demographic concomitants, they often miss an important dimension of inequality: the variation in socioeconomic well-being over the life course and how this may systematically differ among socioeconomic groups as well as change over time. However, there is a substantial body of work—primarily in historical demography—that directly addresses these problems within a life course context.

An inescapable fact of human biology is that individuals' economic productivity varies over the lifecycle. For example, very young children and the very old have few productive capabilities. The nature of the economic system also influences these variations and, along with it, the individual's earnings position. If work involves considerable physical effort or health risks, a reduction in productivity may happen at a relatively young age so that the most economically productive part of the career cycle occurs fairly early in adulthood. On the other hand, if skill is the major job characteristic required, relatively high productivity may last well into old age; however, in this case, lengthy training requirements during an individual's youth are likely to reduce productivity then. Research on middle-class English, European, and American families and on the working classes around the turn of the nineteenth century not only reveals the substantial general economic inequality between these groups but the considerable differences in how relative affluence or deprivation was experienced over the life course, along with the resulting differences in demographic responses. In the case of the middle classes, while earnings might ultimately be substantial, early in the career cycle they were usually inadequate to permit a young man to marry and set up a household in the style considered appropriate to his social station in life. The common pattern was therefore for men to delay marriage until they were economically established and had proved themselves capable of 'properly' supporting a wife and family (Banks 1954, Alter 1988).

In contrast to the middle classes, the career-cycle pattern of earnings was far different for industrial workers, both in Britain and many other nineteenth century industrializing populations. In addition to their generally much lower earnings, the lifecycle period where earnings were greatest and most assured was in young adulthood rather than middle age; hard physical labor, the risk of debilitating disease or accidental maiming or death in the far more dangerous work settings of that era all raised the risk that the middle and older years would be ones of declining or lost productivity. Several adaptive strategies seemed to have been extremely common in such populations. One was a major reliance on the earnings of children to help maintain or improve the family's economic stability over its developmental cycle. Another was an early timing to the start of both marriage and childbearing, thus placing these transitions during the period when working class men's earnings were at their greatest; in addition, this increased the likelihood that at least some children would be old enough to go out to work and contribute to the family's income as the father aged and entered into a higher risk period. However, the downside to this strategy was a frequent pattern of life-cycle poverty or, at the very least, of relative deprivation—for example, when there were several children in the household but none old enough yet to go out and work: this problem was particularly acute if the earnings of the father were lost or much reduced. A different kind of life-cycle poverty typically occurred after all the children had left home and the parents were too old to engage in extensive employment. As a consequence of this life-cycle poverty pattern, the economic situation of individuals at any single point in time was a poor indicator of the frequency with which poverty or relative economic deprivation might be experienced over their entire life course (Anderson 1971, Haines 1979, Rowntree 1922).

While such strategies might well be characterized as rational adaptations to the life course circumstances faced by many working class families, they were also cultural in nature. Historical research indicates that they had a major normative component; children were placed under a strong obligation to work and help improve the family's economic well-being and they

were often allowed to keep very little of their earnings (Parsons and Goldin 1989). Although their work might also improve their own well-being while in the household, such a pattern could nevertheless work against children's long-term self-interest because it tended to reduce the amount of schooling they received and hence their adult socioeconomic status. Moreover, delaying or even foregoing marriage was not entirely uncommon because of the obligations adult children had to help support the family and/or help care for younger siblings, especially if one or both parents were lost (Hareven 1982).

Work on contemporary populations shows that different socioeconomic groups have also been experiencing quite contrasting career development patterns and that this can have significant demographic consequences. For example, consistent with their traditionally important and normatively prescribed economic role in the family, it was still true in the 1980s that American men were much less likely to marry while in an unstable labor market position with low earnings. How large a cumulative impact this had on their marriage timing, however, depended on the length of time a young man remained in such an 'immature' career state and this varied considerably among racial and schooling groups. Compared to men with one or more years of college, a much higher proportion of high school graduates, as well as dropouts, were experiencing a far more difficult and prolonged transition to a relatively stable career position, leading to substantial educational differences in marriage timing. African American males were having particular difficulties and their marriages were much more delayed than that of whites, as indicated by the never married figures cited earlier. These findings also suggest that, given the considerable rise in economic inequality in the US starting in the mid-1970s, increasing career-entry difficulties may have played a significant role in the overall rise in the age at marriage, not only for those reaching adulthood during the 1980s but for earlier and later cohorts as well (Oppenheimer et al. 1997; see *Family Theory: Competing Perspectives in Social Demography*).

See also: Differentiation: Social; Equality and Inequality: Legal Aspects; Inequality; Inequality: Comparative Aspects; Resource Geography

Bibliography

Alter G 1988 *Family and the Female Life Course*. University of Wisconsin Press, Madison, WI

Anderson M 1971 *Family Structure in Nineteenth Century Lancashire*. Cambridge University Press, Cambridge, UK

Banks J A 1954 *Prosperity and Parenthood*. Routledge & K. Paul, London

Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ

Easterlin R A 1987 *Birth and Fortune*. University of Chicago Press, Chicago

Elder G H 1977 Family history and the life course. *Journal of Family History* **2**: 279–304

Galloway P R, Lee R D, Hammel E A 1998 Urban versus rural: fertility decline in the cities and rural districts of Prussia: 1875 to 1910. *European Journal of Population* **14**: 209–64

Guinnane T W, Okun B S, Trussell J 1994 What do we know about the timing of fertility transitions in Europe? *Demography* **31**: 1–20

Haines M 1979 Industrial work and the family life cycle, 1889–1890. *Research in Economic History* **4**: 449–95

Hareven T K 1982 *Family Time and Industrial Time*. Cambridge University Press, Cambridge, UK

Hogan D P 1978 The variable order of events in the life course. *American Sociological Review* **43**: 573–86

Montgomery M R, Casterline J B 1996 Social learning, social influence, and new models of fertility. *Population and Development Review* **22** (Suppl. 5): 151–75

Oppenheimer V K 1970 *The Female Labor Force in the United States*. University of California, Berkeley, CA

Oppenheimer V K 1982 *Work and the Family: A Study in Social Demography*. Academic Press, New York

Oppenheimer V K, Kalmijn M, Lim N 1997 Men's career development and marriage timing during a period of rising inequality. *Demography* **34**: 311–30

Parsons D O, Goldin C 1989 Parental altruism and self-interest: child labor among late nineteenth-century American families. *Economic Inquiry* **27**: 637–59

Rowntree B S 1922 *Poverty: A Study of Town Life*, new edn. Longmans, Green, London

US Bureau of the Census 1998 Marital Status and Living Arrangements: March 1998. *Current Population Reports, P20-514*.

V. K. Oppenheimer

Social Dilemmas, Psychology of

The functioning of societies, groups, and relationships is strongly challenged by social dilemmas, situations in which private interests are at odds with collective interests. Social dilemmas are potentially disruptive for functioning of collectives because the well-being of these larger units is threatened when most or all individuals act in a self-serving manner. For example, during periods of water shortage, it is tempting not to worry too much about how much water to use; but if most people exercise no restraint, this may create severe societal problems. Also, meetings at work would be considerably more productive if all people invest sufficient time and effort in preparing for these meetings. And close partners too, face social dilemmas, such as whether or not one should spend less time with one's independent friends. In a sense, social dilemmas are so pervasive in everyday life that one can go so far as to claim that the most challenging task governments, organizations, and even partners in

a relationship face is to successfully solve social dilemmas. After a brief discussion of the history of social dilemmas, as well as their methodological underpinnings, the basic question of how cooperation between people can be promoted will be addressed.

1. Social Dilemmas: An Historical and Methodological Perspective

The intellectual parents of social dilemmas can be found in early formulations of game theory concerned with the strategic analysis of conflicts of interests (e.g., Luce and Raiffa 1957). Game theory captures decisions in terms of experimental games, that is, situations in which each of the participants has to choose one of two (or more) alternatives, which affect one's own outcomes as well as the other's outcomes. The most famous game following from this work is the Prisoner's Dilemma Game, which represents a true conflict of private interests and collective interests. The prisoner's dilemma derives its name from an anecdote about two prisoners who were accused of robbing a bank. The district attorney, unable to prove that they were guilty, confronted the prisoners with two options: either confess to the crime (option D), or not confess to it (option C). Paradoxically, option D stands for the Defecting option and option C for the Cooperative option, for reasons which will be apparent by the consequences these choices had for both prisoners. If both confess, each will receive a five-year sentence, whereas if neither confesses each of them receives a one-month sentence. Thus, each prisoner would be better off when neither confesses than when both confess. However, the decision is complicated by the fact that each prisoner has a clear temptation to confess. If one confesses and the other does not, then the prisoner who does not confess will receive a 10-year sentence, while the other will be set free. Thus, in a prisoner's dilemma, two individuals are interdependent in that each person is better off by behaving noncooperatively, regardless of the other's behavior (individual rationality). At the same time, both individuals are better of when both behave cooperatively, rather than when both behave noncooperatively (collective rationality).

The interdependence structure underlying prisoner's dilemmas is characteristic of many dyadic interactions. For example, two neighbors living in apartments with thin walls may wish to play their favorite music at high volume, but they together would probably be better off by playing it at a somewhat lower volume. The conflict between individual and collective rationality apparent in the prisoner's dilemma can be extended to situations involving many individuals, which are often referred to as N-person prisoner's dilemmas. For example, trade unions can function considerably better to the extent that a greater number of individuals become members; however, membership requires fees and time and thus is a costly option from an individual perspective. Social dilemmas is a generic concept, embodying both two-person and N-person prisoner's dilemmas, as well as some situations that are quite similar to prisoner's dilemmas, and that do represent a conflict between self-interest and collective interest. It is also noteworthy that in everyday life social dilemmas tend to take the form of either a resource dilemma or a public good dilemma. A resource dilemma calls for a decision whether or not to take (or how much to take) from a public resource. An example is the degree to which one uses water during a period of water shortage. A public good dilemma calls for a decision whether or not contribute (or how much to contribute) to a public good. An example is the degree to which one contributes to the maintenance or improvement of public television. Public good dilemmas pose the problem of free-riding, not contributing but still enjoying the benefits of the public good.

In actual research on social dilemmas, researchers often use experimental games, a technique is that closely linked to study of conflict, cooperation and competition, as well as intergroup relations (e.g., conflict and conflict resolution; cooperation and competition, intergroup relations). The experimental game approach has a number of advantages, in that allows researchers to confront participants with a real dilemma, especially when the 'outcomes' represent real outcomes (e.g., money). Thus, this approach focuses on actual behavior in a context characterized by high levels of experimental realism—it is an involving experimental task. At the same time, to enhance mundane realism researchers have become increasingly interested in examining simulations of social dilemmas as they appear in the real world. Examples are well-designed simulations of the resource dilemma or the public good dilemma. And there is increasing research which examines social dilemmas directly in the real world, which involves correlational survey research, observational research, and experimental field research (see Komorita and Parks 1995). These multitudes of techniques and procedures are used by not only psychologists, but also by anthropologists, economists, sociologists, political scientists, and even biologists. Many or most of the findings reviewed below are observed in various research paradigms.

2. What Factors Promote Cooperation in Social Dilemmas?

How can one promote cooperation in social dilemmas? One important factor derives from the structure itself; that is, the degree to which self-interest and collective interest are conflicting. Cooperation can be enhanced by both decreasing the incentive associated with

noncooperative behavior and increasing the incentive associated with cooperative behavior (Kelley and Grzelak 1972). Generally, the influence of the structure itself has been shown to be strong not only in experimental work on social dilemmas, but also in field studies. For example, it has been found that monetary reward for electricity conservation was one of the most effective means to attaining electricity conservation.

A second factor concerns the possibility of communication. Indeed, one interesting and potentially important way to escape from massive noncooperation is to coordinate the social dilemma in a manner such that people feel committed, or actually promise to contribute to the collective welfare (Orbell et al. 1988). However, opportunities for verbal communication tend to be effective only when individuals indeed discuss their actions relevant to the dilemma at hand. One of the explanations for the fruitful effects of communication is that it enhances trust in others' cooperative choices, and that it is difficult for people to break a promise.

A third factor derives from the personalities involved, in particular differences among three types of social value orientation: prosocial orientation (enhancement of collective outcomes as well as equality in outcomes), individualistic orientation (enhancement of own outcomes with no or very little regard for others' outcomes), and competitive orientation (enhancement of relative advantage over others; Messick and McClintock 1968; Van Lange 1999). People with a prosocial orientation exhibit greater cooperation, expect others to behave more cooperatively, are more trusting of others' intentions and motivations, and interpret the dilemma more strongly as a moral issue and less of an intelligence task than do individualists and competitors. Also, such personality differences tend to be a function of social interaction experiences, in that, for example, prosocials were raised in larger families, having more sisters in particular than individualists and competitors, and people over time tend to become more prosocial and less individualistic and competitive (Van Lange et al. 1997; see also *Personality Development in Childhood*; *Personality Development in Adulthood*).

3. *From Two-person Social Dilemmas to N-person Social Dilemmas*

In two-person social dilemmas, individuals may to some degree be able to promote cooperation in the other. In this regard, one of the most effective strategies is the tit-for-tat strategy (Pruitt and Kimmel 1977), a strategy which starts with a cooperative choice, and subsequently mimics the other's prior choice. Thus, tit-for-tat's willingness to behave cooperatively is conditional upon the other's cooperative behavior. Several studies have revealed that tit-for-tat is more effective in eliciting cooperation in the other than a strategy which always makes cooperative choices irrespective of other's choices (perfect cooperation), which is more effective than a strategy which always makes noncooperative choices irrespective of other's choices (perfect noncooperation). Tit-for-tat is assumed to be effective because it is nice, retaliatory, forgiving, and clear (Axelrod 1984).

One potential limitation of tit-for-tat is that it can only be effectively used in the context of dyadic relationships or small groups. How is one to pursue tit-for-tat in large groups? Apart from the strategies that people use, there is typically a decline in cooperation as groups become larger in size. More specifically, cooperation typically declines as groups become larger up to about seven or eight persons. However, if groups become larger than seven or eight persons then the level of cooperation is not strongly influenced by a further increase in group size. How can one account for the group-size effect? It has been suggested that, with increasing group size, anonymity is greater, individuals become more pessimistic about the efficacy of their efforts to promote collective outcomes (Kerr 1996), people tend to feel less personally responsible for good collective outcomes. An effective solution might be a structural solution, by which one effectively changes the underlying structure, by implementing a system that rewards cooperation or sanctions noncooperation (Yamagishi 1986). The benefits need not necessarily be financial or even material. For example, reduction of car use (so as to reduce smog or congestion) might be promoted to some degree by establishing carpool priority lanes. Other solutions might derive from strengthening social norms,facilitating effective communication, or enhancing feelings of identity with the group or a sense of 'we-ness.' (see also *Social Categorization, Psychology of*)

4. *From Interpersonal to Intergroup Relations*

Interestingly, interactions between groups are characterized by lower levels of cooperation and higher levels of competition than are interactions between individuals (Insko et al. 1998). This individual-group discontinuity effect is accounted for by, first, lower levels of trust in intergroup relations, eliciting fear of being exploited by the other group. Second, individuals within a group are able to support each other in their pursuit of the interest of their own group (and each member's self-interest) at the expense of the other group. Third, responsibility for self-centered behavior is often shared in intergroup settings but interpersonal settings one is individually responsible for his or her actions. One way to enhance cooperation between groups is when one of both groups follows the tit-for-

tat strategy, which yields a level of cooperation comparable to that found among interactions between individuals when one of them follows the tit-for-tat strategy (Insko et al. 1998).

5. *Issues for Future Research*

Recent approaches to social dilemmas examine the ways in which social-cognitive mechanisms (e.g., construal of social dilemmas) might be associated with behavior in social dilemmas, examine how individuals themselves seek to change features of the social dilemma to enhance cooperation (e.g., excluding members from the group) or to escape from the dilemma (e.g., seeking greater independence), and examine the functionality and evolutionary stability of various strategies in large-scale social dilemmas through computer simulation techniques. These issues are examined by various social and behavioral scientists, and these interdisciplinary efforts should make important contributions to theory regarding cooperation and competition, social dilemmas, and interdependence as well as to providing effective solutions to various complex social dilemmas we face in the real world.

See also: Altruism and Self-interest; Cooperation and Competition, Psychology of; Cooperation: Sociological Aspects; Decision-making Systems: Personal and Collective; Personality Development in Adulthood; Personality Development in Childhood; Prisoner's Dilemma, One-shot and Iterated

Bibliography

Axelrod R 1984 *The Evolution of Cooperation*. Basic Books, New York

Insko C A, Schopler J, Pemberton M B, Wieselquist J, Mcllraith S A, et al. 1998 Long-term outcome maximization and the reduction of interindividual-intergroup discontinuity. *Journal of Personality and Social Psychology* **75**: 695–711

Kelley H H, Grzelak J 1972 Conflict between individual and common interests in an n-person relationship. *Journal of Personality and Social Psychology* **21**: 190–7

Kerr N L 1996 'Does my contribution really matter?': Efficacy in social dilemmas. In: Stroebe W, Hewstone H (eds.) *European Review of Social Psychology*. Wiley, Chichester, UK, Vol. 7 pp. 209–40

Komorita S S, Parks C D 1995 Interpersonal relations: Mixed-motive interaction. *Annual Review of Psychology* **46**: 183–207

Luce R D, Raiffa H 1957 *Games and Decisions: Introduction and Critical Survey*. Wiley, New York

Messick D M, McClintock C G 1968 Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology* **4**: 1–25

Orbell J M, Van der Kragt A J, Dawes R M 1988 Explaining discussion-induced cooperation. *Journal of Personality and Social Psychology* **54**: 811–19

Pruitt D G, Kimmel M J 1977 Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology* **28**: 363–92

Van Lange P A 1999 The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology* **77**: 337–49

Van Lange P A, Otten W, De Bruin E M, Joireman J A 1997 Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology* **73**: 733–46

Yamagishi T 1986 The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* **51**: 110–16

P. A. M. Van Lange

Social Evolution: Overview

1. *Concepts and their Definitions*

Social evolution: a process or set of processes of directional social change.

Unilinear evolution: the movement of societies along a single path of change.

Universal evolution: Steward's name for the revival of unilinear evolutionism by Childe and White.

Multilinear evolution: Steward's name for the movement of societies along many different paths of change.

General evolution: Sahlins's name for the overall movement of society and culture along a major path.

Specific evolution: Sahlins's name for the divergent evolutionary paths taken by different societies and cultures.

Progress: an improvement in the overall functional efficiency of society and the quality of life of the members of a society.

Evolutionary theory: any theory that attempts to describe and explain a sequence or sequences of directional social change.

Evolutionary transformations: see social evolution.

Survivals: structures or processes that have been carried by the force of custom or habit beyond the stage of society in which they originated.

Neolithic Revolution: the first great social transformation in human history associated with the domestication of plants and animals and the emergence of settled village life.

Urban Revolution: the name given by Childe to the second great social transformation in human history. It was characterized by the emergence of cities and states.

Environmental circumscription: blockage of the movement of peoples out of a geographic region by such natural barriers as mountain ranges or large bodies of water.

Social stratification: the existence within a society of distinct social strata, or groups distinguished by unequal levels of power, privilege, and prestige.

Economic surplus: a quantity of goods above and beyond what is needed by individuals for survival.

Adaptation: the process whereby individuals adjust themselves to their physical and social environments and solve the basic problems of human living.

Environmental depletion: the diminution of resources within an environment to a point whereby the standard of living is lowered.

Evolutionary materialism: the name given by Sanderson to the general theoretical framework developed by him on the basis of Harris's evolutionary theory.

Evolutionary universal: the name given by Parsons to any new social structure or process that allows a society to achieve a higher level of adaptation or functional efficiency.

Evolutionary typology: a set of stages through which a society or one of its sectors is said to pass.

Differentiation: the emergence of increasing complexity within a society or one of its subsystems.

Militant society: Spencer's name for a society that is highly specialized for warfare.

Industrial society: (a) a type of society characterized by urban life, the factory system, and the mass production of goods; (b) Spencer's name for a type of society specialized for economic activity rather than warfare.

Savagery: Morgan's name for the earliest and most primitive type of human society, one based on hunting and gathering, and a rudimentary technology.

Barbarism: Morgan's name for an intermediate type of society organized around simple forms of agriculture.

Civilization: (a) a type of society characterized by elaborate social stratification, urbanism and city life, occupational differentiation and craft specialization, trade and markets, monumental architecture, writing and record keeping, and the presence of a state; (b) Morgan's name for the highest stage of social evolution, one characterized by the phonetic alphabet and writing.

Societas: a stage of social evolution identified by Morgan that is organized on the basis of kinship and that is normally highly egalitarian and democratic.

Civitas: a stage of social evolution identified by Morgan that is organized along the lines of property and territory rather than kinship.

Mode of production: in Marxian theory, the economic base of society consisting of the level of technological development and the social relationships through which people carry on economic activity.

Primitive communism: in Marxian theory, the first stage of human society characterized by a rudimentary technology and common ownership of the resources of nature.

Slavery: in Marxian theory, the second stage of human society characterized by social and economic relations between a large class of slaves and a small class of slave owners.

Ancient mode of production: see *slavery* (above).

Feudalism: in Marxian theory, the third stage of human society characterized by social and economic relations between a large class of peasants or serfs and a small class of landlords.

Capitalism: in Marxian theory, the fourth stage of human society characterized by social and economic relations between a large working class and a small class of capitalists. Capitalism is based on the production of goods for sale in a market for profit and the accumulation of capital over time.

Hunting and gathering: the simplest stage of social evolution characterized by an economy based on the hunting of wild animals and the collection of fruits, nuts, vegetables, and other plant materials.

Simple horticulture: a stage of social evolution characterized by societies in which people cultivate gardens through the use of digging sticks.

Advanced horticulture: a stage of social evolution characterized by societies in which people cultivate gardens through the use of metal hoes.

Agrarianism: a stage of social evolution characterized by societies in which people cultivate the land with plows and draft animals.

Simple agrarian society: a society that cultivates the land with plows and draft animals but that lacks iron tools and weapons.

Advanced agrarian society: a society that cultivates the land with plows and draft animals and that possesses iron tools and weapons.

Industrialism: see *Industrial Society/Post-industrial Society: History of the Concept*.

Industrial Revolution: a technological and economic transformation, beginning in England in the second half of the eighteenth century, characterized by the development of the factory system and the substitution of machinery for hand power.

Band: a residential association of nuclear families, usually numbering less than 100 individuals, who govern themselves by purely informal mechanisms.

Tribe: a population, numbering in the thousands or tens of thousands, sharing a common culture and language and living in politically autonomous and economically self-sufficient villages of a hundred to several hundred people.

Chiefdom: a hierarchical sociopolitical structure consisting of a chief and various subchiefs who administer an integrated set of villages.

State: a form of political organization characterized by a monopoly over the use of force within a society.

Capitalist world-economy: the name given by Wallerstein to the large-scale intersocietal system that began to develop in Europe in the sixteenth century.

The capitalist world-economy was associated with production of goods for the market and high levels of economic and geographical specialization.

Social evolution is a term that has been used in a number of different ways, but most often it refers to a form of social change that exhibits some sort of directional sequence, a passage from less of something to more of something, or from one type of social arrangement to another type. Usually the term connotes slow and gradual change, and most often it carries with it the notion that the change in question is progressive, i.e., that it leads to some sort of improvement or enhancement. However, social evolution can be fast and even sudden, and it need not lead to improvement or enhancement.

Erik Olin Wright (1983) has offered a precise set of criteria for identifying an evolutionary theory of human society. According to him, a theory is evolutionary if it has four characteristics: (a) it proposes a typology of social forms with potential directionality; (b) the order or arrangement of the social forms is based on the assumption that the probability of remaining at the same stage exceeds the probability of regressing; (c) it asserts a probability of transition, even if only weak, from one stage to another; and (d) it proposes a mechanism or set of mechanisms that will explain movement from one stage to another. One of the benefits of Wright's characterization is that it makes no assumption that social evolution is some sort of automatic or inevitable process. Regression is permitted (although unlikely), and much of the time societies may be undergoing little or no change. Nor need it be assumed that social evolution always proceeds through some sort of rigid sequence.

2. *Historical Development of Theories of Social Evolution*

Theories of social evolution go back to the very beginnings of the social sciences and even earlier. The second half of the nineteenth century was a sort of 'golden age' of evolutionary thinking in sociology and anthropology, the fields that have contributed by far the most to the study of social evolution. There were many well-known evolutionary theorists during this time, but the three greatest were Herbert Spencer (1972, Peel 1971), Lewis Henry Morgan (1974), and Edmund Burnett Tylor (1871). Spencer was a British social theorist who contributed to philosophy, psychology, biology, and sociology. He saw evolution as a process that occurred in the entirety of the natural and social world. All phenomena, he said, had an inherent tendency to change from a state of 'incoherent homogeneity' to a state of 'coherent heterogeneity.' In simpler terms and with respect to human societies, they grew and expanded over time, their structures and functions became increasingly differentiated, and these increasingly differentiated wholes achieved higher levels of unification or integration. Morgan, an American anthropologist, saw social evolution as rooted in technological change, and in his great work *Ancient Society* (1974) he focused mainly on the evolution of government, property, and the family. In terms of the institutions of property and government, Morgan traced out a process whereby highly democratic and egalitarian societies were gradually replaced by societies marked by differential property ownership, social inequality, and the domination of the many by the few. Tylor was a British anthropologist who was much less concerned with the evolution of technology, economy, and political life and much more interested in what anthropologists regard as ideational and symbolic culture, i.e., with myth, language, customs, rituals, and, in particular, religion. Tylor is famous for his use of the concept of survivals, which are features of culture that have been carried by the force of custom or habit into types of societies beyond the one in which they initially appeared. For Tylor, survivals were excellent evidence of a process of social evolution. Completely outside of the disciplines of sociology and anthropology, Karl Marx and Friedrich Engels (1964) developed their own evolutionary model of society. This model focused on the economic evolution of societies over thousands of years and made use of a revised version of the Hegelian dialectic as the mechanism of evolution. Within the economic infrastructures of societies, contradictions or tensions built up that eventually led to explosive situations and the transition to a new stage of society. Engels developed his own highly abstract theory of social evolution in his book *Anti-Duhring* (1939). He also formulated a much less abstract theory in *The Origin of the Family, Private Property, and the State* (1970), a book that depended heavily on the ideas of Morgan.

After about 1890 the golden age of social evolutionary thought came to an end and was succeeded by what has been called 'the antievolutionary reaction' (Harris 1968, Sanderson 1990). Under the leadership of Franz Boas (1940), anthropologists in large numbers turned against the evolutionary theories of Spencer, Tylor, and Morgan. Evolutionary thought was kept alive to some extent through the work of such thinkers as L. T. Hobhouse (Hobhouse et al. 1965) and the celebrated sociologist William Graham Sumner (Sumner and Keller 1927). However, the intellectual atmosphere was for the most part unfriendly to it. But by the 1930s this atmosphere began to change and evolutionism revived. Three thinkers were most responsible for the revival. In 1936 V. Gordon Childe, an Australian archaeologist who taught in Scotland, wrote a famous book entitled *Man Makes Himself*. In this book Childe identified two great evolutionary transformations in human history. The first he called the Neolithic Revolution, which was associated with the domestication of plants and animals and led to larger and more settled societies. The second evolutionary transformation Childe called the Urban

Revolution, which was made possible by the invention of the plow. The Urban Revolution was marked by the emergence of the city, the development of sharp class divisions, and of the form of political organization known as the state.

Following closely on the heels of Childe was the American anthropologist Leslie White. Throughout the 1940s White wrote a series of articles in which he attacked Boas and other antievolutionists for what he saw as their distortions of the claims of the nineteenth-century evolutionists. White (1943, 1959) also developed his own conception of the evolution of culture, which he expressed in terms of a law of evolution. This law stated that culture developed in direct proportion to the amount of energy per capita that societies had harnessed. Societies harnessed energy as they moved from one technological stage to another or as they made their existing technology more efficient, and thus for White technological change was what was driving the evolutionary process. The third major figure in the evolutionary revival was Julian Steward (1955). Steward introduced the concept of multilinear evolution, which he opposed to unilinear and universal evolution. According to him, the nineteenth-century evolutionists had all been guilty of seeing societies as moving through a fixed series of stages (unilinear evolution), and thus greatly oversimplified and distorted the process of evolutionary change. The resurrection of this idea by Childe and White Steward called universal evolution, and it was even more problematic than unilinear evolution. For Steward, evolution was largely multilinear, i.e., it moved along a series of paths rather than one grand path.

By 1960 the evolutionary revival had proceeded to the point of putting evolutionary thinking back into a dominant position, at least in anthropology. After 1960 a new generation of anthropologists (and at least one sociologist) built on this revival and solidified it (Sanderson 1990). Marshall Sahlins (1960) developed an important distinction between general and specific evolution. This paralleled the distinction Steward had made between unilinear or universal evolution and multilinear evolution, except that Sahlins saw both general (unilinear, universal) and specific (multilinear) evolution as real social processes that anthropologists should be studying. Despite being unique, specific evolutionary changes also involved an overall movement of social and cultural life from one stage of evolutionary development to another.

Robert Carneiro (1970), like Sahlins a student of Leslie White, developed a theory of the evolution of the state that would become one of the most famous theories of this phenomenon. The critical factors in Carneiro's theory were population growth, warfare, and what Carneiro called environmental circumscription. Circumscribed environments are those in which areas of fertile land are surrounded by natural barriers that prevent or impede the movement of people out of the area. These barriers include such things as mountain ranges or large bodies of water. Warfare is the result of population growth and resource scarcity, and when land is plentiful people respond to war by simply moving away to new land. But in circumscribed environments this process can only be taken so far. Land is eventually filled up, and the solution to more population pressure and resource scarcity becomes political conquest, a process that eventually results in the type of political society known as the state.

About the same time Gerhard Lenski (1966), a sociologist, was using the ideas of Childe and White to develop a theory of the evolution of social stratification. For Lenski the key to the rise of stratification was technological advancement and increasing economic productivity. As societies advance technologically, their economic productivity rises to the point where they can begin to produce an economic surplus, or a quantity of economic goods above and beyond what people need to survive. When economic surpluses arise competition and conflict over their control emerge, and as these surpluses grow larger the struggles between individuals and groups intensify and stratification systems become more elaborate and polarized.

One of the most important theories of social evolution of this period was that of Marvin Harris (1977). He broke ranks with such theorists as Childe, White, Sahlins, and Lenski on a critical point. These other theorists saw social evolution as a generally adaptive and progressive process that was driven by technological change. When new technologies became available, people readily adopted them because they saw their potential for improving the quality of life. According to Harris, technological change came about for a very different reason: the tendency of societies to deplete their environments as the result of population growth. When populations grew, pressure on resources intensified and standards of living dropped. At some point people had no choice but to advance their technologies so as to make their economies more productive. Thus farming replaced hunting and gathering, and farming with the use of the plow replaced farming with the use of simple hand tools. But technological change itself leads to further population growth and greater environmental depletion, and so a new wave of technological change will soon become necessary.

For Harris, social evolution, at least in preindustrial societies, is a process in which people have been running as hard as they can just to avoid falling farther and farther behind. Sanderson (1994a, 1994b, 1995) has built on the evolutionary ideas of Harris. He has formalized, extended, and to some extent modified them by developing a comprehensive theory that he calls evolutionary materialism. He uses evolutionary materialism as a general framework within which to understand three great evolutionary transformations: the Neolithic Revolution, the rise of the state and civilization, and the emergence of the modern capi-

talist world-economy. Like Harris's, Sanderson's view of social evolution is nonprogressivist and in some respects even antiprogressivist.

About this same time a very different kind of evolutionary theory was developed by the sociologist Talcott Parsons (1966, 1971). Parsons concentrated not on the evolution of technology and other features of material life but on the evolution of ideas and social institutions. Parsons formulated the concept of the evolutionary universal, which is a new kind of social structure or process that allows a society to achieve a new stage of evolutionary adaptation and thereby improve its level of functional efficiency. An early evolutionary universal was social stratification, which allowed a more efficient form of societal leadership to emerge. Later such evolutionary universals as administrative bureaucracy and money and markets emerged. An important evolutionary universal in more recent times, Parsons claimed, was the democratic association, which gave the use of power a broad societal consensus.

3. *Typologies of Evolutionary Stages*

Most theories of social evolution have involved the delineation of evolutionary stages. This was no less true in the nineteenth century than it is today. Spencer used two different evolutionary typologies. One of these was based explicitly on Spencer's notion of evolution as increasing differentiation: simple, compound, doubly-compound, and trebly-compound societies. At one end of the scale of differentiation we have simple societies, which are politically headless or have only rudimentary forms of headship. At the other end, that of trebly-compound societies, we have civilizations such as the Egyptian empire, ancient Mexico, the Roman empire, or the contemporary United States. Spencer's other typology distinguished between what he called militant and industrial societies. Military societies are those in which individuals are subordinated to the will of society and forced to obey its dictates. They are produced by the needs associated with warfare and the preparation for it. Industrial societies, on the other hand, are characterized by an emphasis on economic activities rather than warfare, and are notable for their much greater individualism and overall freedom. Although Spencer did not work out in any precise way just how these two typologies corresponded to each other, just as he saw evolution as a process of increasing differentiation he also saw it as a generalized movement from militant to industrial societies. For Spencer the epitome of industrial societies was his own nineteenth-century Britain.

Morgan divided societies into three major stages, which he called 'ethnical periods': savagery, barbarism, and civilization. The first two of these are divided into lower, middle, and upper substages. Societies in the stage of savagery are the oldest and most primitive. They survive by hunting and gathering and use rudimentary tools, such as spears or bows and arrows. Societies have reached the stage of barbarism when they have begun to domesticate plants and animals and live by agriculture rather than by hunting and collecting. Civilization is achieved with the invention of the phonetic alphabet and writing. Tylor made use of a very similar typology, as did other nineteenth-century thinkers. However, Morgan also employed another typology, which involved the distinction between *societas* and *civitas*. Societies at the stage of *societas* were organized on the basis of kinship and had relatively egalitarian and democratic social and political arrangements. Societies having reached the stage of *civitas*, by contrast, rested on property and territory as organizational devices. Kinship declined greatly in importance and was replaced by the state as the integrator of society. Economic inequalities became prominent, and the democratic arrangements of earlier times gave way to one or another form of despotism.

Marx and Engels introduced a typology of stages that was to become one of the most famous of all time. For them a society rested upon one of four possible modes of production: primitive communism, slavery, feudalism, or capitalism. The earliest human societies were at the stage of primitive communism. Here technology was very primitive, and people subsisted by hunting and gathering, by simple forms of agriculture, or by animal herding. Private property had not yet emerged, nor had class divisions. People lived in harmony and equality. Slavery, sometimes called the ancient mode of production, was characteristic of such earlier civilizations as Egypt and Mesopotamia and ancient Greece and Rome. Technology had undergone major advances since the stage of primitive communism, but as a result private property and class divisions had emerged. A small class of slave-owning masters confronted a large class of slaves. Slavery was followed in due course by feudalism, which characterized Europe during the Middle Ages. Slavery had largely given way to serfdom, and the major class division in society was that between landlords and serfs. The transition to capitalism was associated with another major upheaval in social life. Landlords and peasants were replaced by capitalists and workers as societies industrialized and urbanized. In due time the capitalist mode of production would fill with internal contradictions and eventually burst asunder, giving way to a socialist society without class divisions.

In recent decades very useful typologies have been produced by Gerhard Lenski (1966, 1970) and Elman Service (1971). Lenski has distinguished five major stages based on the level of technological development: hunting and gathering, simple horticulture, advanced horticulture, agrarianism, and industrialism. Hunting and gathering societies are the simplest and oldest. Here people survive by living on what nature provides. Men hunt and women collect fruits, nuts, vegetables,

and other plant matter. The transition to simple horticulture was made with the Neolithic Revolution that began in some parts of the world about 10,000 years ago. Simple horticulturalists subsist by a rudimentary form of agriculture that is based on the cultivation of gardens with the use of digging sticks. They often use a form of cultivation known as 'slash-and-burn' cultivation. This involves cutting down a section of forest, burning off the accumulated debris, and then spreading the ashes over the garden area as a form of fertilization. Usually gardens can only be used for a few years because soil fertility is quickly lost, so they will be abandoned in favor of other areas, only to be returned to many years hence. Advanced horticulturalists use the same basic form of cultivation but till the land with metal hoes rather than digging sticks, which give them a technological advantage. Agrarian societies are based on the plow and the harnessing of animal energy for plowing. Fields are completely denuded of shrubs and trees and slash-and-burn is abandoned. Industrial societies began to emerge with the Industrial Revolution of the late eighteenth century. They are based on industry, the factory system, and urban life.

Service's typology is based on the form of sociopolitical organization rather than the mode of technology, and distinguishes four stages: bands, tribes, chiefdoms, and states. Bands and tribes are similar in that they lack any formal mechanisms of social control. Leaders exist, but they have no real power or authority to compel anyone to do anything. They lead by example and they advise and suggest, but their decisions have no binding force. A band is usually a residential association of nuclear families, with a population anywhere from about 25 persons to 100. A tribe may be thought of as a larger collection of bands. Often the term tribe is used to designate a population, numbering in the thousands or tens of thousands, that shares a common culture and language. This cultural and linguistic whole is composed of politically autonomous and economically self-sufficient villages, which may consist of anywhere from a hundred to several hundred persons. Chiefdoms develop when the villages of the tribe come to be subordinated to an administrative structure of the whole. This structure is hierarchical and is dominated by a chief and at least a few subchiefs, each of whom has the capacity to compel the actions of others. Chiefdoms rest on at least some degree of coercion, but by definition this coercion is not sufficient to prevail when there is a rebellion from below. When such a degree of coercion has been achieved, a state has emerged. States are thus large-scale, highly centralized political systems in which rulers have a monopoly over the use of force.

Other typologies exist, and all have their strengths and weaknesses. Morton Fried (1967) developed a typology based on the degree of social inequality: egalitarian, rank, and stratified societies. Eric Wolf (1982) formulated a Marxist typology of stages based on the mode of production: the kin-based mode, the tributary mode, and the capitalist mode. And Talcott Parsons (1966) formulated a typology based on three major stages, which he called primitive, intermediate, and modern societies. There is no one way to create a typology of evolutionary stages. What is important is that the typology describes a real directional sequence or process based upon a clearly identified mechanism of change. All of the typologies identified here do that to one extent or another.

4. *Social Evolution: Its Main Outlines*

The discussion now moves from typologies and theories of social evolution to the actual process of social evolution itself. There is considerable disagreement about what the most important dimensions and driving forces of this process are, but the following sketch probably would not produce too much controversy among social evolutionists (Sanderson 1995).

Virtually all social evolutionists would agree that the first great social transformation was the Neolithic Revolution, which introduced plant and animal domestication into the world. The most striking thing about the Neolithic from an evolutionary perspective is that it occurred in remarkably parallel form in several major regions of the world. Beginning about 10,000 years ago in southwest Asia, the transition to communities based on agriculture rather than hunting and gathering occurred at later times in other parts of the Old World—southeast Asia, China, Africa, and Europe—and three parts of the New World: Mesoamerica, South America, and North America. It is now generally understood that most if not all of these Neolithic transitions occurred independently of one another, a phenomenon that can only be explained in evolutionary terms. The effects in these various regions were much the same too. The transition to agriculture led to settled and more densely populated communities that for a while remained relatively egalitarian but eventually gave way to stratified societies organized into chiefdoms.

By around 5,000 years ago in several parts of the world societies that had evolved into chiefdoms were beginning to make the transition to a state level of political organization, or to what many scholars have called civilizations. This occurred first in Mesopotamia and Egypt, and then later in China, the Indus valley in northern India, parts of Europe, and Mesoamerica and Peru in the New World. Civilizations are societies with highly centralized political institutions involving a monopoly over the use of force, elaborate social stratification, urbanism and city life, occupational differentiation and craft specialization, trade and markets, monumental architecture, and writing and record keeping. Most civilizations have been agrarian societies, thus cultivating the land with plows and draft animals and intensively fertilizing the soil. Like

the Neolithic Revolution, the transition to civilization and the state was a process of independent parallel evolution in several parts of the world.

Social evolution occurred rapidly throughout the world between 10,000 and 5,000 years ago, but once the stage of civilization had been reached there was a considerable slowing of the evolutionary process. Social change continued in that societies became bigger and more densely populated, technology advanced, trade and markets expanded, and empires grew larger and more powerful. However, there was no qualitative leap to a new mode of social organization in the sense that there had been with the transition from hunter–gatherer to agricultural communities or to civilization and the state. Lenski (1970) has distinguished between simple and advanced agrarian societies on the basis of the possession by the latter of iron tools and weapons. He has suggested that advanced agrarian societies were in many ways so different from simple agrarian societies that it is difficult to include them in the same category. Yet both types were based on the same technological and economic foundations and the differences between them were more quantitative than qualitative. It was to take another several thousand years before there was a shift to a qualitatively new mode of social organization.

What was this shift and when did it occur? This is where sociologists enter the picture. Most sociologists have taken the view that it was the Industrial Revolution of the late eighteenth century that introduced a qualitatively new form of social life, industrial society. However, in recent years some sociologists, along with social scientists from other fields, have moved this qualitative transformation back in time to the sixteenth century (Wallerstein 1974). The qualitative shift is then considered to be the transition to a capitalist world economy. Capitalism—selling goods in a market in order to earn a profit—in some form or another had existed for thousands of years, but after the sixteenth century it began to get the upper hand and to replace earlier, precapitalist forms of social life. From this perspective, the Industrial Revolution was simply part of the logic inherent in the advance of capitalism. Most scholars have treated the rise of capitalism as a unique feature of European society and Europe's decisive contribution to the world. However, at about the same point in history a society at the other end of the world, Japan, began to undergo a capitalist transition of its own, and it is possible to identify a number of striking similarities between Europe and Japan that may have facilitated their transition to the new economic system (Sanderson 1994b).

Once the transition to capitalism had been made and the Industrial Revolution had commenced, strikingly similar changes began to occur among rapidly industrializing societies. Today's industrial capitalist societies all have very similar occupational structures, governments based on parliamentary democracy, systems of mass education that are closely linked to the work world, similar stratification systems with relatively high levels of mobility, and rationalization in the form of advanced science and technology. Some sociologists and most historians will protest that recognition of these similarities should not prevent us from seeing the many important differences among industrial capitalist societies. That is certainly true, but from the perspective of the evolutionist, it is the similarities that are more compelling.

5. *Conclusions*

It was noted above that evolutionary theories of society have ebbed and flowed in their popularity ever since they were introduced. They were extremely prominent between 1850 and 1890, fell to a low point between 1890 and 1930, and rose to prominence again around the middle of the twentieth century. But the bubble burst again, and beginning in the 1970s evolutionary theories fell from grace. They have been criticized on numerous grounds (Nisbet 1969, Sztompka 1993). Among other things, it has been charged:

(a) that they overestimate the amount of directionality in history;

(b) that they rely on teleological explanations, which are illegitimate;

(c) that they have a strong endogenous bias and ignore, or at least insufficiently attend to, the external relations among societies;

(d) that they employ a specious concept of adaptation; and

(e) that their progressivist view of history is much too optimistic.

All of these criticisms have been answered at length (Sanderson 1990, 1997). Precedent suggests that evolutionism will once again become prominent in sociology and anthropology. This type of theorizing has fared well in optimistic periods of history and poorly in more pessimistic periods. Western society and Western social science have long been in a pessimistic period, but this cannot last. With the eventual turn toward greater optimism, evolutionism will again assume its rightful place as a major perspective on social life and its historical changes.

See also: Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Industrial Society/Post-industrial Society: History of the Concept; Inequality; Slavery as Social Institution

Bibliography

Boas F 1940 *Race, Language, and Culture*. Macmillan, New York

Carneiro R L 1970 A theory of the origin of the state. *Science* **169**: 733–8

Childe V G 1936 *Man Makes Himself*. Watts, London
Engels F 1939 *Herr Eugen Duhring's Revolution in Science* (*Anti-Duhring*). 3rd edn. International Publishers, New York
Engels F 1970 *The Origin of the Family, Private Property, and the State*. International Publishers, New York
Fried M H 1967 *The Evolution of Political Society*. Random House, New York
Harris M 1968 *The Rise of Anthropological Theory*. Crowell, New York
Harris M 1977 *Cannibals and Kings: The Origins of Cultures*. Random House, New York
Hobhouse L T, Wheeler G C, Ginsberg M 1965 *The Material Culture and Social Institutions of the Simpler Peoples*. Routledge and Kegan Paul, London
Lenski G 1966 *Power and Privilege: A Theory of Social Stratification*. McGraw-Hill, New York
Lenski G 1970 *Human Societies: A Macro-level Introduction to Sociology*. McGraw-Hill, New York
Marx K, Engels F 1964 *The German Ideology*. Progress Publishers, Moscow
Morgan L H 1974 *Ancient Society, or Researches in the Line of Human Progress from Savagery through Barbarism to Civilization*. Peter Smith, Gloucester, MA
Nisbet R A 1969 *Social Change and History: Aspects of the Western Theory of Development*. Oxford University Press, New York
Parsons T 1966 *Societies: Evolutionary and Comparative Perspectives*. Prentice-Hall, Englewood Cliffs, NJ
Parsons T 1971 *The System of Modern Societies*. Prentice-Hall, Englewood Cliffs, NJ
Peel J D Y 1971 *Herbert Spencer: The Evolution of a Sociologist*. Basic Books, New York
Sahlins M D 1960 Evolution: Specific and general. In: Sahlins M D, Service E R (eds.) *Evolution and Culture*. University of Michigan Press, Ann Arbor, MI
Sanderson S K 1990 *Social Evolutionism: A Critical History*. Basil Blackwell, Oxford, UK
Sanderson S K 1994a Evolutionary materialism: A theoretical strategy for the study of social evolution. *Sociological Perspectives* **37**: 47–74
Sanderson S K 1994b The transition from feudalism to capitalism: The theoretical significance of the Japanese case. *Review* **17**: 15–55
Sanderson S K 1995 *Social Transformations: A General Theory of Historical Development*. Blackwell, Oxford, UK
Sanderson S K 1997 Evolutionism and its critics. *Journal of World-Systems Research* **3**: 94–114
Service E R 1971 *Primitive Social Organization: An Evolutionary Perspective*. 2nd edn. Random House, New York
Spencer H 1972 *Herbert Spencer on Social Evolution*. Peel J D Y (ed.) University of Chicago Press, Chicago
Steward J H 1955 *Theory of Culture Change: The Methodology of Multilinear Evolution*. University of Illinois Press, Urbana, IL
Sumner W G, Keller A G 1927 *The Science of Society*. 4 Vols. Yale University Press, New Haven, CT
Sztompka P 1993 *The Sociology of Social Change*. Blackwell, Oxford, UK
Tylor E B 1871 *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art, and Custom*. 2 Vols. John Murray, London
Wallerstein I 1974 *The Modern World-system: Capitalist Agriculture and the Origins of the European World-economy in the Sixteenth Century*. Academic Press, New York
White L A 1943 Energy and the evolution of culture. *American Anthropologist* **45**: 335–56
White L A 1959 *The Evolution of Culture*. McGraw-Hill, New York
Wolf E R 1982 *Europe and the People Without History*. University of California Press, Berkeley, CA
Wright E O 1983 Giddens's critique of Marxism. *New Left Review* **138**: 11–35

S. K. Sanderson

Social Evolution, Sociology of

The term social evolution describes sociological theories that study the changes and development of societies as a whole, by adopting a long-term perspective. These theories have alternately emphasized: (a) the characteristics of successive social orders; (b) the mechanisms and laws that articulate this succession; (c) the *direction* of this evolution; and (d) its *progressive* character. Although explicit references to social evolution were certainly not lacking in the reflections of philosophers who dealt with the whole history of human evolution (Plato, Vico, Bossuet), actual theories on social evolution were only established in the eighteenth century (Mandeville, Hume, Ferguson, Smith, Savigny), and nineteenth century (Saint-Simon, Comte, Marx, Spencer, Menger, Durkheim, Tönnies). Even though they are radically different, these theories all identify a tendency to progress from relatively simple and uniform social organizations characterized by a small or even nonexistent division of labor, to social organizations that are marked by a growing division of labor and a considerable increase of population.

In spite of their shared minimalist presuppositions, two currents of sociological theories have been established, each interpreting social evolution in a radically different way: (a) theories that try to understand society's inevitable *laws of evolution*; and (b) theories that interpret social evolution as an unintended process generated by the spontaneous combination of individual actions.

1. *Social Evolution as an Ineluctable Development*

Many philosophers, sociologists and economists have tried to single out the laws that clearly explain the succession of different social phases. They have tried to identify *evolutionary criteria* that would allow them to look back at changes that have already taken place, and foresee future changes. According to Saint-Simon, a *law of progress* supports the *process* of evolution and this *progress* is not linear, since 'organic periods' of progress necessarily alternate with 'critical periods' of temporary and partial regression. However, the

industrial revolution and scientific progress inevitably lead to a more evolved organic period that was organized according to the principles of 'positive science' which puts power in the hands of scientists. According to Comte (Comte 1875), 'the development of civilization follows a necessary law', which individuals 'are subject to by invariable necessity': the 'law of the three phases'. This law establishes that every society goes through 'theological, metaphysical and positive phases' which follow a fixed and predetermined scheme. The positive phase characterizes the most advanced stage of social progress, dominated by a new and higher form of science, 'social physics', that permits the organization and management of society as a whole. Marx considers 'dialectic materialism' the progressive law of all society, marking the inevitable evolution from 'primitive communism' to the 'feudal' phase that leads to the 'capitalistic' phase, which in turn leads to the inevitable success of 'communism.'

These theories, based on a *holistic* view of society that considers social evolution a *teleocratic* process, have been subjected to a great deal of methodological criticisms. The impossibility of foreseeing how human knowledge will develop and the inevitable emergence of unintended consequences are sufficient criticism to the determinism of these theories. There is also a fundamental epistemological objection to the claim these theories have of giving a scientific explanation to the entire social process: in order to explain a social phenomenon scientifically, universal laws are necessary, i.e., laws that are not empirically assessed only on the phenomenon itself. Where social evolution is concerned this is not possible, because it is a historically unique process. Consequently, there are no scientific 'laws of evolution' to be discovered, but only 'tendencies' that can and must be explained by laws (Popper 1957). In a more general sense, these evolutionary theories view society as a *closed system* where it is possible to determined the evolutionary dynamics. Marx, on the basis of a situational analysis, considered social evolution an *emerging effect*, and attributed a necessary and universal quality to mere tendencies, simply because he did not allow for the fact that a social system is necessarily *opened* by, and to, new information from the outside.

2. *Social Evolution as a Spontaneous Order*

Individualistic theories on social evolution are formulated in a completely different way. They do not seek *laws of evolution*, but try to define a *principle of evolution* that is determined as the spontaneous combination of subjectively rational, individual actions. Social order and its evolution are conceived mostly as the unintended result of the interaction between individuals with limited knowledge, pursuing their private interests through an ultimately positive *exchange* with others.

B. De Mandeville and *Scottish* economists and philosophers of the eighteenth century (Hume, Ferguson, Millar, Smith) are among the first supporters of this theory. They proposed an *invisible hand explanation* for the origin and development of the most important social institutions. Even before, Smith, Mandeville explained the affirmation of the division of labor as a spontaneous result generated by the necessity of satisfying the needs of the individual in an increasingly efficient way. The division of labor, like other social institutions is therefore, according to Ferguson's famous definition, the 'result of human action but not of human planning.' Among the *Scottish* philosophers, Adam Smith particularly drew his conclusions at a more completely macro-sociological level, identifying trade as the mechanism behind social order, since it spontaneously *combines* single actions. Smith believed that in the exchange with others, individuals are driven 'by an *invisible hand* to support objectives that were not part of their original intentions. [---] By pursuing their own interests, they advance society's interests more efficiently than was deliberately intended.'

Owing to Burke's influence on the German historical school (Savigny, Neibuhr), the evolutionary approach was established in linguistics (Humboldt) and law in the early nineteenth century. It was thus adopted first of all in the human and social sciences, and only later in the natural sciences (for example, with Lyell in geology and Darwin in biology). In fact, Hayek had defined these theories of evolutionism as 'Darwinian prior to Darwin.' As the evolutionary theory was increasingly accepted in biology, theories on social evolution were given new impetus. This was also thanks to Spencer, who should be considered one of the most significant theoreticians of social evolutionism based on individualism, which was later developed by the economists of the Austrian school. Spencer considered social evolution as a process generated by a combination of individual actions, which tend to organize spontaneously, establishing rules and social organizations that are selected on the basis of their fitness to perform the basic functions of human society (survival of the species, production of riches). The 'obligatory co-operation' at the base of 'military society' and the 'spontaneous co-operation' at the base of 'industrial society' are the two most important 'types' of social orders that have selectively evolved. Spencer believed that societies where spontaneous cooperation was established had an *evolutionary advantage* because they maximised the *problem solving capacity* by safeguarding individual liberty and allowing the greatest possible safeguarding individual liberty and allowing the greatest possible number of compatible individual plans to be accomplished. For this reason, Spencer gives 'spontaneous co-operation' an ethical value because it is the best solution to reproduction of

the species and communal welfare, which he considers the two most important issues of humanity.

The Austrian Economic School is probably the strongest current of theorists of social evolution with an individualistic foundation. It spanned the nineteenth and twentieth centuries with Menger, Böhm-Bawerk, Mises and Hayek and the neo Austrian Americans (Rothbard, Kirzner, Stigler). By emphasizing the primacy of the theory in the social sciences and suggesting that they be refounded on an individualistic basis, Menger harshly criticises the 'pragmatist' theories that consider the genesis and evolution of social order as effects of intentional plans. In his opinion, 'law, language, the State, currency, trade [---], in their various expressions and incessant changes, are largely the spontaneous result of social evolution.' Mises developed a momentous theory of rationality (praxeology), providing an epistemological defence of social order as produced by market, and demonstrating the impossibility of making an 'economic calculus' in a planned economy.

But, it was above all F. A. von Hayek who suggested the strong link between methodological individualism and social evolution. Hayek harshly criticised the reification of the *Kollektivebegriffe* starting from the 'social system', and based his methodological individualism on an *ontological individualism*. He stated that there are only individuals whose intentional actions (and their unintentional consequences) tend to combine spontaneously, giving rise to a 'spontaneous order' meaning *open systems* that are self-generated and self-regulated. In Hayek's view, the great 'misunderstanding' of evolutionists is confusing the study of social evolution with the impossible search for the 'laws of evolution' on one hand, and on the other, with studying the selection process between single individuals (social Darwinism). Social evolution is actually the result of 'group selection', meaning the competition between groups organized according to different rules, that are selected on the basis of their functional adaptation. Hayek, like Spencer, demonstrates the *evolutionary superiority* of the 'spontaneous order' (*Cosmos*) by distinguishing it from 'contrived order' (*Taxis*). This superiority is generated by the fact that the spontaneous order allows each individual to use the generalized knowledge of others to reach their goals. *Competition* between different types of organizations and individual plans is therefore the means of *propagation* and *retention* of evolutionary selection results. It is the method for exploring the *unknown* and as such, is the driving force of social evolution.

These individualistic theories of social evolution are based on a fallibilistic view of knowledge, and as such they harshly criticise the rationalistic and 'constructionist' claims of scholars who try to discover the inevitable laws of social evolution and then, on the basis of this discovery, plan social order or at least a significant part of it. Social order should be seen instead as the unintended result of the spontaneous combination of rational, subjective actions that relate to situations in an *adaptive* way. This nomocratic order is an autopoietic system that incorporates the principle of its own evolution: a spontaneous cooperation between free individuals. As such, it is teleonomic, since it has a tendency to self-organization, and is not teleological, because it is not guided by any previously determined, predictable, and inevitable goal.

3. *The Evolution of Social Systems*

Several other important theories of social evolution exist beside these two currents of thought. Although they do not attempt to define the inevitable laws of development, they study the characteristics and sequential succession of cultural structures, systems and social phases, considering them separate and above the order of the single individuals of which they are composed. Durkheim and Parsons are the most significant exponents of these theories.

Even though Durkheim's works do contain interesting examples of the *invisible hand explanation* (such as the explanation of the genesis of the family in *De la division du travail social*), he also harshly criticises Smith and the 'economists' who interpreted social evolution as a result of the interaction between single individuals pursuing their private interests. Society cannot be 'deduced' from individual behavior because 'the agents in exchanges remain mutually estranged', since a subjective interest is 'the least constant thing in the world.' Durkheim considers society a '*sui generis*' reality, and therefore 'collective life is not the result of individual life; on the contrary, the latter is the result of the former.' Society exercises 'moral authority' over individuals and has a 'collective conscience.' In archaic societies, which are characterized by a 'mechanical solidarity', the collective conscience entirely replaces individual conscience. In more advanced societies characterized by an 'organic solidarity', a progressive individualization occurs that is not however sufficient to free the individual from collective *tutorship*. Evolution that leads to modern society is produced by the division of labor, which in turn, according to Durkheim (influenced by Spencer in this case), 'varies in direct proportion to the volume and density' of society. Having given up trying to define a *law of development*, Durkheim finally explained the division of labor in a way that is similar to the invisible hand explanation. He writes in fact that 'the economic services which the division of labor can provide are insignificant compared to the moral effect it produces, and its true function is to achieve solidarity between two or more individuals.'

Parsons' 'neo-evolutionism' has a *culturalistic* imprint. After the Second World War, he clearly revised the views he expressed in *The Structure of Social*

Action. Parsons defines three 'levels' of social evolution: 'primitive society' (centered on family relations), 'intermediate society' (regulated by religious institutions), and 'modern society' (established by the industrial revolution). The passage from one social stage to another can be explained by *the laws of general evolution* and the *law of cybernetic hierarchy*. The general laws of evolution (for which Parsons is much indebted to Spencer), explains how societies evolve through the 'increase of adaptive capacity' namely through their capacity to 'transform the environment in order to satisfy their needs (Parsons 1971).' Successive social phases, under the pressure of functional adaptation, are therefore characterized by a progressive 'differentiation' (of roles, functions and sub-systems) and by their 'integration.' The *law of cybernetic hierarchy* explains that the 'social system', works and evolves by the restraints placed by 'cultural structures' (i.e., systems of rules and values) on the 'social system' and on individual actions, which Parsons places on the lower levels of the cybernetic scale. Parsons describes society as a '*sui generis reality*' explicitly following Durkheim's theories. Social evolution is therefore the result of the 'evolutionary change' of cultural variables. These variables have the 'function' of 'legitimizing the rules of a society', providing a view of society in its 'totality.'

4. *Evolution and Social Progress*

A dogma of theories on social evolution in the eighteenth and nineteenth century, based on the Enlightenment and positivism, was that evolution coincided with *progress*. Saint-Simon, Comte, Marx and Condorcet before them, all believed that the *laws of evolution* coincided with the *laws of progress*, meaning that social evolution occurs in a predictable way that is equated with a necessary human progress. This attempt to make *historic development* and *human progress* coincide is invalidated by ' Hume's law', which states that it is not *logically* possible to derive *prescriptions* from *descriptions*. The idea of progress cannot therefore be *rationally founded*. This does not prevent social evolution from being defined as *positive* or *progressive*, or being either optimistic (Spencer, Hobhouse, Childe, Withe, Parsons) or pessimistic (Malthus, Speigler), for these are *evaluations* based on subjective value choices. As such, they can be debated and critically discussed, but they cannot be considered as describing a necessity.

It is possible to avoid the *naturalistic fallacy* of confusing *goodness* with *evolution* by following ′Hume's law.' In order to assess the results of evolutionary adaptation, it is necessary to move from a historical-sociological plane to an ethical one. Individualistic evolutionists, like Spencer and Hayek, separated evolution from progress and—fearing that social evolution might alter the nature of the spontaneous order—gave progress a moral value by turning it into an ideal to be defended and developed should certain value choices be accepted (individual liberty, the increase of wealth and knowledge, etc.). The purpose of evolutionary ethics is precisely to demonstrate that spontaneous cooperation based on free individual actions is the best means to achieve these values.

5. *Adaptation and Innovation*

The *variation/selection* mechanism is the *hard core* of both biological and social evolution. As neo-darwinian models were increasingly accepted in biology and theories of self-organized complex systems were increasingly accepted in physics, social sciences also began to study this evolution-generating *mechanism*. However, the dynamics of this mechanism can be interpreted in different ways that once again oppose individualism to collectivism. It is possible to have a 'structural' view of innovation, by seeing it as a deterministic product of *environmental instruction* according to Lamarck's theories. Variations would then be the manifestation of *functional needs* that arise from, and are selected by, the need to guarantee the social system's equilibrium. Social changes would therefore be essentially conceived as 'endogenous', i.e., part of the social structure and necessary to its reproduction. Individualistic evolutionists, like Weber and Schumpter proposed a completely different view of social innovation. They emphasized the individual, unpredictable and *strategic* character of innovation by focusing respectively on the innovative role of the charismatic *élites* and of entrepreneurs. Innovation does not answer a question posed by the environment, it is the result of individual strategies, and strategies between individuals, who both condition and are conditioned by the environment in which they develop. *Variations* cannot be explained by strict Lamarkian theories, because the social organism's ingrained environmental needs and tendencies do not deterministically generate them. On the other hand, they also cannot be interpreted using an orthodox darwinian model, since they are not *random* or *random-like*.

These two pure types of innovation correlate to two *pure* kinds of adaptation. Parsons reduces adaptation almost exclusively to its functional dimension, defining it as a process that allows society as a whole (as well as its subsystems) to respond to some essential, functional needs. Another view of adaptation considers it closely linked to a biological model (especially Spencer and Hayek). Adaptation is seen as an *endless order generating process* that reduces the variety of configurations produced by variations through a *rule and group selection*. This selection essentially refers to groups that spontaneously adopt different rules and organizations. Among these groups, only those which are better equipped to address the most important social needs will survive.

On another level, a similar and complimentary process of *selection through competition* takes place between individuals belonging to the same group, and this process determines the subjective actions and projects that are most adaptable to environmental circumstances and particularly to the rules of the group itself. However, the fact that emphasis is placed on *group selection* does not mean that this evolutionary prospect denies methodological individualism (as stated by D T Campbell 1995). Organized groups are the spontaneous result of the fact that individuals follow specific rules of conduct that are conveyed and spread by *learning through imitation*. Following these guidelines, one eventually considers tradition as a necessary element of evolution since it characterizes the process of retention of kinds of actions and dispositions.

6. Social Evolution and Evolutionary Epistemology

D T Campbell has used the expression 'evolutionary epistemology' to describe epistemological theories (from Poincaré and Mach to Popper) that saw the evolution of human knowledge as a process that progresses by *trial and error-elimination*, creating new hypotheses and selectively eliminating false ones. Evolutionary epistemology also helps to better understand the social evolution process, which is also a *discovery procedure* that *lato sensu* has a cognitive nature. Social innovations are nothing more than attempts to resolve problems (intentionally proposed or occurring spontaneously), that are selected on the basis of their capacity to deal with *problematic situations*. Comparing the two evolutionary processes, it can be said that: (a) *fitness* is nothing more than the elimination of errors; (b) the transmission of selected solutions (in social and cognitive contexts) means *learning by error*; and (c) the increase of complexity in social systems is simply an increase of *background knowledge*. Like the evolution of knowledge, social evolution will never reach a state of *equilibrium* because new *instructions* will result in an environmental change that will be the *background* of new variations. Finally, an important distinction should be made between social evolution and the evolution of the knowledge. Even though the former cannot be reduced to a Lamarckian model, its *environmental instruction* is more significant while the evolution of knowledge (if one excludes the collectivistic theories of some contemporary science sociologists such as Bloor, Barnes and others), is articulated by *creative inventions* that, unlike social innovation, are much less affected by *external selection*.

See also: Comte, Auguste (1798–1857); Cycles: Social; Durkheim, Emile (1858–1917); Hayek, Friedrich A von (1899–1996); Marx, Karl (1818–89); Parsons, Talcott (1902–79); Social Evolution: Overview; Spencer, Herbert (1820–1903); Theory: Sociological

Bibliography

Boudon R 1986 *Theories of Social Change*. Basil Blackwell/Polity Press, London

Campbell D T, Heylighen F 1995 *Selection of Organization at the Social Level: Obstacle and Facilitators of Metasystem Transition. World Future: the Journal of General Evolution* **45**: 181–212

Childe V G 1951 *Social Evolution*. H Schumann, London

Comte A 1875 *The Positive Philosophy*. Trubner, London

Di Nuoscio E 2000 *Epistemologie dillatione e ondine spontaneo*. Rubbeltino, Sovena Mannelli

Durkheim E (1893) 1964 *The Division of Labor in Society*. Free Press of Glencoe, New York

Hayek F A von 1973–79 *Law, Legislation and Liberty*. Routledge, London, 3 Vols.

Hayek F A von 1988 *The Fatal Conceit*. Routledge, London

Heylighen F, Aerts A 1999 (eds.)*The Evolution of Complexity*. Kluwer, Dordrecht, Germany

Infantino L 1998 *Individualism in Modern Thought*. Routledge, London

Parsons T 1971 *The System of Modern Societies*. Englewood Cliffs, NJ

Popper K R 1957 *The Poverty of Historicism*. Routledge, London

Sanderson S K 1990 *Social Evolutionism: a Critical History*. Cambridge, MA

Spencer H 1882–1896 *Principles of Sociology*. Williams and Norgate, London, 3 Vols.

Turchin V F 1977 *The Phenomenon of Science. A Cybernetic Approach to Human Evolution*. Columbia University Press, New York

White L A 1959 *The Evolution of Culture*. McGraw-Hill, New York

E. Di Nuoscio

Social Facilitation, Psychology of

Social facilitation refers to the effect of the presence of others on the individual. This phenomenon is of psychological interest to scholars who regard the presence of others as a minimal condition for social influence.

1. History of the Term

The term *social facilitation* was coined by the American Floyd Allport in an early textbook of social psychology (Allport 1924). Allport's was a behavioral treatment of the field. As part of his *Social Psychology*, Allport maintained that a response called forth by a nonsocial stimulus could be augmented by an incidental social stimulus. He termed this contributory influence *social facilitation*. Allport reported demonstrations of the social facilitation of students' task performance from the sight and sound of other students performing the same task. Allport's research participants cancelled vowels more quickly, multiplied

numbers more quickly, and generated word associations more quickly when they were working alongside peers than when they were not.

When an individual is in the presence of others, the others will invariably assume some role. Sometimes they are watching the individual, as members of an audience. Sometimes they occupy the same role as the individual, and all are engaged in the same behavior. In this latter case, the others are said to be *co-actors*. Audience and co-action researchers seek to infer social facilitation from experimental data, by comparing individuals who are in the presence of others with individuals who are alone. However, it is difficult to collect valid evidence of social facilitation, because the monitoring that the experiments require may compromise research participants' sense of solitude. In such cases, no one is truly 'alone,' and the impact of others' presence is difficult to isolate.

Social facilitation was first reported in an 1898 journal article by Norman Triplett. Studying records of bicycle races, the sports psychologist noted a tendency for cyclists to race more quickly when riding alongside competitors than when riding alone against the clock. Triplett developed a laboratory-based analogue of this social phenomenon, when he had school children turn a fishing reel alone and in the presence of coacting classmates. The presence of these young coactors produced faster reel-turning, an effect that Triplett attributed to a 'dynamogenic' increase in motivation. Years later, Triplett's 1898 study of others' presence was dubbed the first experiment in social psychology, although its claim to this title has been disputed.

2. *Effects of the Presence of Others*

Social facilitations have been observed in a number of species. Chickens peck at food more quickly when other chickens are pecking; rats press a bar faster in the presence of other rats; cockroaches run with greater speed when running alongside other cockroaches. To the biologically oriented psychologist (e.g., Clayton 1978), these demonstrations of animal social facilitation hold profound interest.

People can be affected in many ways by the presence of others. Physiologically inclined researchers report social facilitations of human heart rate, blood pressure, and electrodermal activity. If often the presence of others increases an individual's physiological arousal, sometimes affiliation with others can reduce high levels of arousal (Mullen et al. 1997). Psychological investigators have documented social facilitations of various behaviors (Kent 1994). In the presence of others, people eat large meals, express conventional judgments, and show a tendency to smile. Gender and cultural differences in social facilitation are sometimes observed. In the presence of others, males (but not females) display a heightened tolerance for pain. In the presence of others, Japanese (but not American) students conceal facial signs of distress.

Since Triplett's pioneering investigation, psychologists have reported hundreds of experiments on the social facilitation of human task performance. The evidence that they have amassed indicates that the term social facilitation may be a misnomer. If sometimes the presence of others facilitates a person's task performance, often it has the opposite effect. In the presence of others, people have trouble unscrambling anagrams; in the presence of others, people lack motor coordination; in the presence of others, people do poorly on memory tasks.

When is a person's task performance facilitated by the presence of others, and when it is impaired? Psychologists had been pondering this question for decades before they found a convincing answer. In 1965, social psychologist Robert Zajonc published an article in the prestigious journal *Science*. There, Zajonc contended that the impact of the presence of others on task performance depends on the complexity of the performer's task. The presence of others serves to facilitate the performance of simple tasks and to impair the performance of complex tasks, Zajonc claimed. Subsequent evidence verifies that the impact of the presence of others does, indeed, depend on task complexity. At the same time, it indicates that social impairments of complex performance are stronger than social facilitations of simple performance. People perform simple tasks more *quickly* in the presence of others. However, there is little evidence that others' presence has any effect on simple performance *quality*, a quantitative review of 241 'social facilitation' studies concluded (Bond and Titus 1983).

3. *Explanations of Social Facilitation*

Psychologists have attempted to explain social effects on task performance. In his 1965 article, Robert Zajonc proposed the most influential explanation. According to Zajonc, the presence of others increases an individual's level of generalized drive. Generalized drive, learning psychologists Clark Hull and Kurt Spence hypothesized, functions to energize an individual's dominant response tendency toward any stimulus compound. Generalized drive enhances this dominant response tendency, to the exclusion of any weaker tendencies. According to Zajonc's drive theory of social facilitation, the impact of the presence of others on the performance of a task should hinge on the performer's dominant response tendency when attempting the task. On simple tasks, where the dominant tendency is toward a correct response, the presence of others should facilitate performance. On complex tasks, where the dominant tendency is to

make a mistake, the presence of others should enhance this error, thereby impairing performance.

Zajonc's theory intensified interest in the study of social facilitation. Many experiments were designed to test this drive theory of social facilitation, and the theory was often corroborated. When the presence of others caused errors in athletic performance, undermined spatial maze learning, and enhanced response tendencies that had been trained to dominance in the laboratory, researchers claimed evidence for the operation of a socially induced drive. The presence of others was shown to have different effects on task performance at different stages in the learning of the task—retarding individuals' acquisition of novel skills, then facilitating execution of those same skills once performance reached a plateau. Over training, drive theorists reason, correct response tendencies come to dominate over errors; thus dominant-response enhancements can explain these opposing effects.

Although for a time theorists agreed that the presence of others functioned as a source of generalized drive, there were debates about the origin of the drive. Zajonc attributed drive to the unpredictability of species mates, and an individual's need, in the presence of others, to be alert to the unexpected. Other social psychologists (notably Nickolas Cottrell) believed that drive reflected a learned tendency to associate the presence of others with positive and negative outcomes. Only when others were in a position to confer such outcomes would they have any effects on the individual, Cottrell implied.

Positive and negative outcomes are often associated with the presence of others, all psychologists agreed. Many wondered, however, whether the *mere* presence of others would be sufficient to influence the individual, independent of any associative processes. In search of effects of the mere presence of others, experimenters conducted some colorful studies. They fed rats in the presence of anaesthetized rats; confronted chickens with a dead companion; and had undergraduates trace spatial mazes in front of blindfolded mannequins. When results indicated that individuals can be affected by others who are merely present, many psychologists acted as though Cottrell's analysis of social facilitation could be discarded. More recently, the theoretical significance of mere presence effects has been revisited. Robert S. Baron argues that if an individual learns to associate positive and negative outcomes with the presence of others, then others who are merely present should become a conditioned stimulus for those outcomes. From Baron's perspective, mere presence effects are simply conditioned responses, and have no bearing on the origin of any socially induced drive. No doubt, there will be continuing discussion about the theoretical significance of the effects of the mere presence of others. In the meantime, it is apparent that others' mere presence can have some effects.

As drive theorists engaged in intramural disputes, alternative theories of the effects of the presence of others were being developed (Geen 1989). A self-presentational analysis attributed the social facilitation of task performance to the performer's regulation of a public image, and the social impairment of performance to embarrassment from a loss of public esteem. Theoretically, performers lose esteem when they must attempt complex tasks in public. Complex tasks, theorists note, are ones on which the individual makes many mistakes. Self-awareness theories held that the presence of others prompts individuals to compare their task performance to a performance standard. These self-evaluations may inspire performance facilitation or occasion a withdrawal from the task, depending on the outcome of the comparison. Cybernetic models of the self-comparison process were constructed. Performers must divide their attention between a task and any others who are present, an influential distraction-conflict theory emphasized. This theory made the novel prediction that social distractions would facililtate simple task performance, and inspired evidence of an audience-induced attentional overload (Baron 1986). The inability to visually monitor others' reactions arouses uncertainty, and this is the key to understanding mere presence effects, writes the Australian psychologist Bernard Guerin (1993).

4. *Related Phenomena*

Alongside social facilitation from the presence of others are broader facets of the psychology of public performance. Performances, for example, take place in front of audiences of greater or lesser size. Mathematical equations have been developed to relate the number of people in the audience to the audience's impact on the performer. Data have been fit to the proposed equations, and reveal a negatively accelerating impact, as an audience grows in size. An audience may be supportive of the performer, or it may be hostile. Traditionally it was assumed that audience support would benefit the performer, and create a home-field advantage in sports. However, audience support may intensify pressure and make the performer choke, some paradoxical data suggest. Performers may bring positive expectations or negative expectations to a task. The presence of others benefits performers who expect to succeed, and impedes performers who expect to fail, the self-efficacy theorist Lawrence Sanna writes. The presence of co-actors may obscure the individual's contributions to a group effort and provide an incentive to loaf, a related research tradition emphasizes. Psychologists continue to pursue these and other possibilities as part of their century-long study of social facilitation.

See also: Adolescent Development, Theories of; Adult Development, Psychology of; Deindividuation, Psychology of; Developmental Psychology; Group Decision Making, Social Psychology of; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Infant and Child Development, Theories of; Lifespan Development, Theory of

Bibliography

Allport F H 1924 *Social Psychology*. Houghton Mifflin, Boston

Baron R S 1986 Distraction-conflict theory: progress and problems. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 19, pp. 1–40

Bond C F, Titus L J 1983 Social facilitation: a meta-analysis of 241 studies. *Psychological Bulletin* **94**: 265–92

Clayton D A 1978 Socially facilitated behavior. *The Quarterly Review of Biology* **53**: 373–92

Cottrell N B 1972 Social facilitation. In: McClintock C G (ed.) *Experimental Social Psychology*. Holt, New York, pp. 185–236

Geen R G 1989 Alternative conceptions of social facilitation. In: Paulus P B (ed.) *Psychology of Group Influence*, 2nd edn. L. Erlbaum Associates, Hillsdale, NJ, pp. 15–51

Guerin B 1993 *Social Facilitation*. Cambridge University Press, Cambridge

Kent M V 1994 The presence of others. In: Hare A P, Blumberg H H, Davies M F, Kent M V (eds.) *Small Group Research: A Handbook*. Ablex, Norwood, NJ, pp. 81–105

Mullen B, Bryant B, Driskell J E 1997 Presence of others and arousal: an integration. *Group Dynamics: Theory, Research, and Practice* **1**: 52–64

Zajonc R B 1980 Compresence. In: Paulus P B (ed.) *Psychology of Group Influence*. L. Erlbaum Associates, Hillsdale, NJ, pp. 35–60

C. F. Bond, Jr.

Social Geography

1. Introduction—When and What

Social geography emerged as a significant sub-discipline during the 1960s. Although earlier references to 'social geography' can be found, Johnston (1987) notes that it is barely mentioned in reviews of geography published in the 1950s. The development of the new sub-discipline was part of exciting theoretical and methodological change that geography experienced in the 1960s. In the 1970s and 1980s, vigorous theoretical debates led to a variety of conflicting approaches to research in social geography. By the early 1980s, the sub-discipline was of major importance within human geography. Indeed, a recent discussion of the area claimed that it became 'virtually synonymous with the whole field of human geography' (Johnston et al. 1994, p. 562). Nevertheless, the growing prominence of cultural geography towards the end of the 1980s and into the 1990s had a major impact on social geography. This was neatly symbolized by the decision of the 'Social Geography Study Group' of the Institute of British Geographers (which had been formed in the early 1970s) to change its title to 'The Social and Cultural Geography Study Group' in the 1990s.

What is social geography and what do social geographers do? The rapid growth of the sub-discipline and the significant theoretical debates within it have involved shifts in its definition and practice. In 1965, Pahl suggested that it was about 'the theoretical location of social groups and social characteristics, often within an urban setting' (p. 82). In 1975, Emrys Jones argued that 'social geography involves the understanding of the patterns which arise from the use social groups make of space as they see it, and of the processes involved in making and changing such patterns' (p. 7). The emphasis on social groups evident in Pahl's earlier definition remains of particular importance, in that what is held to make social geography 'social' is its concern not with individuals but with people as members of groups. However, there are some significant changes between the two definitions. First, Jones' definition introduces the idea that social processes produce spatial patterns and vice versa. This understanding of social geography remained dominant into the 1980s. Second, Jones' definition includes the phrase 'the use social groups make of space *as they see it*' (author's emphasis). This signals an increasing interest in people's subjective experiences in specific places—an interest that has become increasingly prominent in research designs.

The understanding of the scope and aims of social geography implied in Jones' definition endured until the mid-1980s when it came under increasing attack. These attacks and their implications for the development of social geography make better sense in the context of the account of the type of work that social geographers did from the 1960s to the mid 1980s, presented below.

2. Engaging with Positivism

2.1 Describing the Social Characteristics of Urban Areas

A major focus of research during the late 1950s and 1960s was the statistical description and analysis of the characteristics of the populations of urban areas. Central to this project was the identification of systematic patterns in the location of different social groups. Analysts examined which groups were likely to be segregated from other groups and whether there

was evidence for particular patterns of location for different groups within cities.

At this time, there were also important moves within human geography to apply positivist philosophy and methodology to the pursuit of geographical knowledge. These moves encouraged, on the one hand, attention to clearly specified theoretical statements and predictions and, on the other, the use of quantitative data and statistical techniques that were presented as allowing rigorous, scientific testing of hypotheses. Also associated with the growing commitment to positivism, was a desire for predictive models of the location of human activity and people's 'spatial behavior,' often drawing on neoclassical economics. For geographers interested in social groups and in cities this implied a need to develop predictive theories of the spatial form of cities, including theories of the location of social groups within urban areas.

Geographers looking for theoretical and methodological inspiration turned to the work of urban sociologists. First, they drew upon the work of the Chicago School—including Louis Wirth, Robert Park, and Ernest Burgess—who carried out extensive studies of Chicago in the first half of the twentieth century. They used maps of a wide range of statistical information in order to identify 'natural areas'—residential areas with distinctive physical and cultural characteristics—that the Chicago School sociologists argued developed from unplanned processes of competition and cooperation amongst urban residents. Burgess had developed a general model of urban residential land use as a series of concentric zones focused around the Central Business District.

Second, urban geographers were drawn to the work of Eshref Shevky and Wendell Bell who had studied characteristics of the populations of Los Angeles and San Francisco. Shevky and Bell had argued that the 'increasing scale' of society resulting from the increasing complexity of economic organization and the associated growth of urban areas could be linked to three dimensions of urban residential differentiation: 'social rank' or 'economic status'; 'urbanization' or 'family status'; and 'segregation' or 'ethnic status.' A key feature of their work was the development of 'social area analysis' using data on census tracts. They selected variables to represent the three dimensions and then used these to categorize census tracts in terms of 'high' or 'low' economic, family, or ethnic status. Although the theoretical links between increasing societal scale and the three dimensions of economic, family, and ethnic status were poorly specified and strongly criticized, geographers enthusiastically embraced the use of census data to characterize areas of the city.

During the 1960s, computers that were more powerful allowed geographers to use sophisticated statistical techniques to analyze data on census tracts. Factorial ecology was particularly popular—the use of factor analysis or principal component analysis to identify a few general constructs which underlie a large number of variables. Shevky and Bell's three status dimensions often emerged from the census data. The analyst could also classify census tracts in terms of the underlying constructs and map the results to see whether any systematic spatial patterns in characteristics of census tracts could be identified. These maps were compared with Burgess' concentric model of urban form and comparisons between factorial ecologies of different cities also made. Similar techniques were also used to look for relationships between residential differentiation, location, and patterns of crime, deviance, and mental health.

The availability of census data for many cities (mostly, in the 'developed' world) and increasing access to computers allowed social geographers to indulge in an orgy of analysis. However, there were some clear limitations to this work. First, the nature of the constructs identified was limited by the available census data: they could be as much a reflection of the data available as of the social 'reality' of urban areas. Second, because the data related to areas rather than individuals, false associations between variables may have been made (the 'ecological fallacy'). Moreover, differences in the sizes and rules for creating census tracts between countries limited the comparability of the results. Third, although the practitioners of these analyses aspired to work within a positivist framework much of the research was inductive description rather than tests of clearly specified hypotheses linked to well identified theoretical propositions. Nevertheless, this body of work provided a considerable body of descriptive information on residential differentiation that has become part of the taken for granted knowledge of geography.

2.2 *The Behavioral Challenge*

The descriptive analysis of urban residential characteristics was not the only type of work in social geography done during the 1960s and early 1970s. There were those who, despite their hopes for a positivist approach to the development of theory within geography, were dissatisfied with the approaches to studying human behavior that were used widely at that time.

In the 1960s, much research in human geography drew its inspiration from models in economic geography that posited that people possessed perfect information and were rational and optimizing in their behavior. Dissatisfaction with these assumptions prompted some researchers to turn to psychological theory concerned with cognition and decision-making in order to investigate what information people had available to use in spatial decision-making and how they used it. The principal method employed in these studies was the questionnaire survey, frequently incorporating some variety of attitudinal and personality

tests drawn from social psychology. The research carried out within this broad framework is known as 'behavioral geography.'

Work in behavioral geography was concerned with the locational decision-making of firms, individuals' decisions over where to live, people's everyday movements, and their information and ideas about space and place. Gould's work on mental maps was particularly influential (Gould and White 1974). His initial research examined American undergraduates' views of each US state as a place in which to live after graduation. The results showed that individuals had a strong preference for their home area and that this preference was superimposed upon an image of a national space with marked variations in preferences. The idea of an incomplete, biased and personal vision of the spatial world conjured up by the term 'mental map' became part of geographers' ideas about how people made decisions about both everyday movement and migration. Information about space and place was also recognized as important within studies of residential mobility and of people's everyday travel behavior. In the former, researchers were influenced by the early work of the sociologist Peter Rossi whose 1955 book *Why Families Move* was subtitled *A study in the social psychology of residential mobility*. A large number of studies of residential mobility in different cities and countries resulted. In the latter, diary data on travel behavior were used to develop inductive theories about the way people choose and order travel destinations. Travel behavior and migration research explored how people searched for information in space and developed strategies for choosing locations.

Research using the perspectives of behavioral geography—while part of social geography—remains marginal to the field because of its focus on the individual rather than on people as interacting with and partly constituted by their social milieu. However, the impossibility of such an individualist approach is evident in much of the work actually carried out. Researchers often examined contrasts between individuals drawn from different social groups (for example, different income, class, age and, occasionally, gender groups). Others examined the influence of the resources available to individuals on their spatial cognition and so ended up investigating differences between social groups. The research is also significant to social geography because criticisms of its individualistic approach were important to the development of alternatives.

3. Competing Approaches

During the early 1970s, many geographers became dissatisfied not only with the philosophical approach of positivism but also with the lack of critical political engagement that they detected in contemporary research. These concerns grew out of the radical ideas promoted in America and Europe as part of opposition to the American war in Vietnam, anti-nuclear campaigns (especially in Britain and European countries), civil rights (particularly in America), and the growing 'rediscovery' of poverty in the prosperous countries of the West. What was needed, it was felt, was research that could suggest how to change society, not merely how to understand it. Positivism was criticized for suggesting that an objective approach to social science was necessary, feasible, and desirable and that it was possible to find 'universal laws' of human society. Diverse competing 'radical' approaches were proposed; they all shared the desire to challenge the inequalities and injustices of contemporary society although some sought this through changing our valuation of human beings, some through institutional change and some through revolutionary political action. A sketch of some of the main approaches follows. Although here they are separated into demarcated 'boxes,' in reality their boundaries were fuzzy.

3.1 Humanistic Geography

The humanistic geography that emerged in the 1970s was animated by a desire to present an alternative to positivism. It shared the behavioral geographers' critique of some positivist approaches as dehumanized—insofar as they did not engage with the world as experienced by individuals—but added to this the more fundamental criticism that positivist social scientists made false claims to objectivity—by failing to recognize the subjectivity of both the observed and the observer. Here, humanistic geography made a decisive break with positivism and turned to approaches which require empathy with people's own perception of the values and meanings which inform their action (Jackson and Smith 1984). Two other important features of humanistic approaches were the insistence that people were active agents shaping their own worlds and a fascination with the links between place and social identity.

Although links have been made between the humanistic geography of the 1970s and a variety of early work in geography, the most immediate influence was the Chicago School. Humanistic geographers turned to the wealth of studies of the 'cultural life' of 'natural areas' produced by the Chicago School, as well as drawing on the 'community studies' tradition in sociology which had itself been partially inspired by the Chicago School. One of the major strands of work focused on understanding the way of life of particular social groups in particular areas of a city (although a few studies of rural areas were also carried out). It used ethnographic methods including participant observation, firsthand observation and interviews as well as

documentary information and census materials. Other researchers used ethnographic methods to explore features of crime and deviance in selected neighborhoods, often using factorial ecology to identify suitable areas for study.

Humanistic geographers also engaged with what were for 1970s geographers the new philosophical approaches of phenomenology and existentialism. There were two important consequences. First, a move towards thinking of the individual as constituted through the social interactions in which they take part. These social interactions and the everyday behaviors in which people engage create a taken-for-granted world which the social scientist can attempt to deconstruct. Second, there was an attempt to avoid prior theorization about the values and meanings of those people who are studied and to be open to understanding the world through their eyes. However, this still allows the findings of such studies to be interpreted within a theoretical or practical context and for disagreement about such interpretations to be fierce.

Another focus was on understanding the cultural meanings of, and values attached to particular landscapes and built forms. This work employed the methods of hermeneutics (for example, Meinig 1979). In the 1970s, this work was considered to be on the margins of social geography and soon became seen as part of the revival of cultural geography. As such, its investigation of the multiplicity and instability of meanings, their communication and reformulation began a tradition of work that has had important implications for contemporary social geography.

3.2 Welfare and Weber

David Smith's *Human Geography: A Welfare Approach* (1977) addressed what had begun to emerge as a new approach to human geography of particular significance to social geography. This approach focused on 'who gets what, where, and how'—that is, on issues of social justice and inequality. It therefore involved both arguments about what constitutes social (and often, territorial) justice and attempts to measure empirically various types of social inequality. An early focus was on developing social indicators to measure inequality between social groups living in different areas at a variety of scales ranging from the city to the world. This research built on the expertise developed in earlier descriptive analyses of census data on cities. It also included measurement of people's spatial access to services (for example health care or education) or proximity to negative factors such as pollution. However, although this answered questions of where inequality or deprivation might be found and who suffered such problems, it did not answer questions of how this inequality arose. Consequently, researchers developed a range of analyses of the operation of particular agencies and agents whose actions seemed significant to the distribution of resources.

A substantial body of work thus developed around state and private institutions' management of the distribution of resources, drawing on the theories of Max Weber. The concept of 'managerialism' was prominent. It was first proposed by Pahl (1975), who suggested that within both the state and private sector there were 'gatekeepers' whose decisions were crucial to access to resources and so to people's quality of life. Within social geography, the ideas of managerialism stimulated a wide range of studies of access to both public and private sector housing. These were counterposed to the earlier studies of housing search and housing choice to suggest that they had paid too little attention to the constraints on people's choices in the housing market. The theme of choice and constraint reappears within this body of work applied not only to housing but also to resources like health care, education, services for the elderly, parks, and open space and delivery services in rural areas. However, the ideas of Weber also introduced an increasing interest in conflict between and among institutions and groups. Here social geographers drew upon the ideas of the sociologist John Rex (1968) on housing classes, as well as on Weberian ideas about the operation of bureaucracies and professional groups, to explore conflicts over the distribution of resources.

A different approach to access used the ideas of time geography developed by the Swedish geographer Hagerstrand (1975). He suggested that time and space are resources which people must use to realize the 'projects' which they wish to carry out. Since traversing space requires time and many activities require the co-presence of individuals or presence at a particular location, not all projects can be realized. People's ability to carry out particular projects is limited by capacity, coupling, and authority constraints. Despite the enthusiasm of some geographers and a few empirical studies, the potential of these ideas remained regrettably underdeveloped. Most empirical use of Hagerstrand's ideas focused on the physical limits to access and did not develop the possibilities of a more fully realized social theory of human behavior in space and time.

The majority of welfare-focused research was concerned with cities in 'developed' countries; however, attention was also paid to inequalities in the rural areas of such countries and these perspectives have been used in studies of developing countries, Eastern Europe and the former USSR. There were overlaps with other radical approaches. The study of conflict and constraint often employed ethnographic methods to explore the values and patterns of behavior of gatekeepers and the social groups attempting to access resources. Attempts to answer the how element of welfare geographers' problematic often drew on research using Marxist theory which sought to situate

explanation of issues such as housing access within an analysis of the operation of capitalist society.

3.3 Marxist Approaches

The book that best illustrates the thinking that drew geographers to Marxist theory was Harvey's *Social Justice and the City* (1973). A key event also was the founding of *Antipode: a Radical Journal of Geography* in 1969 at Clark University in the USA. Harvey's book charts and explains a shift in his thinking from liberal to Marxist. At the start of the 1970s, very few geographers knew anything about Marxism, and Marxist analyses of the political economy of cities and regions were rare. By the end of the decade, there had been a major expansion in the use of Marxist approaches sufficient to create a new orthodoxy. This had implications across all geography while in social geography Marxist approaches were embraced with particular enthusiasm.

The analysis of residential differentiation and change had been a central part of social geography since its inception. The new Marxist theory offered ways of understanding these processes through ideas of class struggle and capital accumulation and large-scale patterns of investment, and disinvestment could be interpreted in terms of crises in processes of accumulation. Such analyses were applied to gentrification and suburbanization. Processes of socialization within residential areas could also be theorized as involving the reproduction of particular social subgroups (class fractions). These early formulations were challenged by an important new body of socialist-feminist work that developed in the early 1980s. While this incorporated many of the postulates of Marxism it criticized the conceptual and practical neglect of gendered power relations and the failure to identify and explore the links between gender and class relations. Critical empirical research on both gentrification and suburbanization sought to explore the ways in which gender relations were involved in the construction of the spatial structure of the city and the implications of this for women's oppression in society. In general, this socialist-feminist work was more sensitive than other Marxist approaches to the role of human agency and to humanist claims concerning the partiality of knowledge.

The approach to analyzing access to services was also strongly affected by Marxist analysis. Castells' work on collective consumption and social movements was highly influential. Castells defines collective consumption as the provision of services that are produced and managed collectively, and distributed through non-market allocation processes. He argued that these were provided through the State apparatus because they were necessary to the reproduction of labor power but would not be provided directly by the capitalist market because of a lower than average profit rate. Castells made a highly contentious definition of the city as 'a unit of collective consumption corresponding more or less to the daily organization of a section of labour power' (1976, p. 148). More importantly to research in social geography, he also spelt out how a variety of social movements might arise in response to inadequacies in the provision of collective consumption. Such social movements might fragment class solidarity and reduce the possibility of radical political change or enhance it. Castells' ideas stimulated extensive conceptual and empirical research in both urban sociology and in social geography.

3.4 Critiques and Cross-fertilization

By the early 1980s, social geographers had become better versed in social theory and more critical and sophisticated in its application. There were debate and cross-fertilization of ideas amongst researchers working in the three areas identified above and increasing awareness of their shortcomings as well as their strengths. There were strong critiques from humanists and feminists of the neglect of human agency in Marxist approaches in geography and this encouraged geographers' interest in the sociologist Anthony Giddens' 'structuration' theory (Giddens 1981). This attempted to show how human agency and structures were mutually constitutive within a time-space framework and drew on the earlier ideas of Hagerstrand. These critiques encouraged attempts to develop a more 'humanistic' Marxism sensitive to both human agency and contingency. At the same time, those concerned with issues of welfare and social justice were increasingly using some form of political economy approach deriving from Marxism to frame their research. All three approaches, as well as earlier work in social geography, were criticized by feminist geographers for their neglect of gender (Bowlby and McDowell 1987) and a major new body of research by feminist social geographers began to develop. This involved ethnographic empirical study of the everyday life of women and incorporated ideas drawn from welfare geography (in relation to issues of access and inequality), humanistic geography (in relation to women's experience of and social relations within particular places) and, as indicated above, Marxist geography. Some of this feminist work was linked to another major development within geography—the explosive growth of cultural geography during the late 1980s with its insistence that because meaning is constantly created and changed through shared but contested discourses, both individual and group self-definitions are unstable and contested.

This new cultural geography drew on earlier humanistic social geographical research as well as on feminist and postmodern critiques of modernist theory. In particular, modernist theory was criticized for assuming a white, bourgeois, and masculine subject and for its attempts to discover universal truth. The

impact of these critiques, and of the new empirical work done in the 1980s and 1990s in both cultural and economic geography, has been to call into question both earlier definitions of social geography and the coherence of the sub-discipline.

4. *What is 'the Social' in Social Geography?*

The later 1980s and the 1990s saw a progressive destruction of the main features of Emrys Jones's 1975 definition of social geography. That definition involved the separation of social process and spatial pattern and a focus on the use of space and creation of spatial patterns by social groups.

4.1 *Social Process and Spatial Pattern or 'Spatiality'?*

The 1980s saw growing dissatisfaction with the idea of social processes as distinct from spatial patterns, on the grounds that, because all social interactions necessarily take place in and through space, spatial pattern cannot be isolated from social process. These attacks derived from Marxist, humanistic, and cultural geographers' critiques of positivist analyses of abstract space as an empty container of social processes. Instead, Marxists argued that space was produced through capitalist social processes and through processes of class and other groups' opposition to capitalism. Humanistic and cultural geographers were also using phenomenological theory and empirical research to argue that our understandings of space are produced through our everyday social activities and do not draw on prior concepts of abstract space.

4.2 *The Social vs. the Economic*

In the early work, people's participation in the life of social groups made social geography social and distinguished it from approaches that focused on the individual. The existence of social groups and the salience of characteristics such as ethnicity or income were largely taken for granted. Questions about which social groups to study, and why, have had a number of implications for social geography.

First, increasing familiarity with Marxist and Weberian social theory made social geographers more aware of the need to define what were significant social groups in relation to the major constellations of power in society. Since economic relations are clearly of central importance to social power this required social geographers to consider the relations between the social and the economic. Were social relations to be seen as simply derivative of more 'fundamental' economic relations? If so, was social geography of less significance than economic geography?

Second, social geography had always concerned itself with issues of social reproduction—with housing, education, health care, and crime. However, as the new feminist work of the 1980s soon established, they had neglected some central aspects of social reproduction—notably those to do with domestic labor and the home. This feminist work also showed through analyses of the links between the gendering of the 'home' and the 'workplace' how untenable was the division between social reproduction and production. Thus, they claimed that 'production and reproduction—the conventionally accepted subjects of economic and social geography respectively—cannot analytically be separated' (Bowlby and McDowell 1987, p. 305). At the same time, researchers in economic geography were arguing for analyzing economies and economic relations as social constructs. During the 1990s, the importance of examining the social relations embedded in economic exchange was re-emphasized by a number of researchers. In particular, feminist research on labor markets showed clearly the interpenetration of the social and the economic (Hanson and Pratt 1995). Thus, the term 'social' no longer clearly delimits the subject matter for social geographers.

4.3 *Discourses and Social Groups*

Meanwhile a resurgence of cultural geography and associated growth of interest in postmodern theory has called into question social geographers' accustomed ways of thinking. On the one hand, these developments have drawn attention to the power relations inherent in the definition of social groups and to the political problems of representation this poses. They have also stressed the fluidity and contested nature of such definitions. On the other hand, postmodern approaches have questioned the validity of modernist 'grand theory' such as that of Marx and Weber that has been part of the stock in trade of social geographers from 1975 to 2000. Cultural geography has thus required social geographers to examine processes of representation and communication more closely and to question the idea that there is a single 'social reality' to be uncovered through research.

4.4 *What Future for Social Geography?*

There are significant vulnerabilities in the current situation of social geography. It now lacks a distinctive focus. In the 1990s, the definition of social geography given in the *Dictionary of Human Geography* was 'The study of social relations in space and the spatial structures that underpin those relations' (Johnson et al. 1994, p. 562). This is so broad that it is difficult to see where social geography ends and other sub-disciplines—for example, economic, political, or cultural geography—start.

It is also arguable that the term has become shorthand for research that concentrates on issues of 'social relations in space' only in Western societies. Despite the existence of much research in developing countries which, if carried out in a Western society,

would be termed 'social geography' there has been too little interaction between those researchers working on developing and developed countries. There are, however, some indications that this is changing with greater appreciation of the implications of the globalization of communication and economic relationships and the significance of diasporic cultures.

For many social geographers the growth of cultural geography presents an opportunity to combine many of the themes of the two sub-disciplines in fruitful investigation of social inequality and difference and to focus on and destabilize definitions of a wider range of 'social groups'—such as disabled people and gay and lesbian people. Social geographers' tradition of political engagement and ethnographic research can combine positively with cultural geographers' sensitivity to discourse and symbolic expression of difference. The rapprochements between social, cultural, and development geographers seem likely to produce valuable research. Nevertheless, the term 'social geography' on its own—as opposed to 'social and cultural geography' or 'social and development geography'—seems likely to revert to its earlier obscurity.

See also: Access: Geographical; Behavioral Geography; Crime, Geography of; Critical Realism in Geography; Cultural Geography; Disability, Geography of; Gender and Feminist Studies in Geography; Humanistic Geography; Marxist Geography; Postcolonial Geography; Postmodern Urbanism; Postmodernism in Geography; Qualitative Methods in Geography; Residential Segregation: Geographic Aspects; Rural Geography; Sexuality and Geography; Space and Social Theory in Geography; Spatial Analysis in Geography; Time-geography; Urban Geography; Urban System in Geography

Bibliography

Bowlby S R, McDowell L 1987 The feminist challenge to social geography. In: Pacione M (ed.) *Social Geography: Progress and Prospect*. Croom Helm, London

Castells M 1976 Theoretical propositions for an experimental study of urban social movements. In: Pickvance C (ed.) *Urban Sociology: Critical essays*. Methuen, London

Giddens A 1981 *A Contemporary Critique of Historical Materialism*. Macmillan, London

Gould P, White R 1974 *Mental Maps*. Penguin, London

Hagerstrand T 1975 Space, time and human conditions. In: Karlqvist A, Lundqvist L, Snickars F (eds.) *Dynamic Allocation of Urban Space*. Saxon House, Farnborough, UK, pp. 3–14

Hanson S, Pratt G 1995 *Gender, Work and Space*. Routledge, New York

Harvey D 1973 *Social Justice and the City*. Edward Arnold, London

Jackson P, Smith S 1984 *Exploring Social Geography*. Allen and Unwin, London

Jones E 1975 *Readings in Social Geography*. Oxford University Press, Oxford, UK

Johnston R J 1987 Theory and methodology in social geography. In: Pacione M (ed.) *Social Geography: Progress and Prospect*. Croom Helm, London

Johnston R L, Gregory D, Smith D M 1994 *The Dictionary of Human Geography*, 3rd edn. Blackwell, Oxford, UK

Meinig D (ed.) 1979 *The Interpretation of Ordinary Landscapes*. Oxford University Press, New York

Pahl R 1965 Trends in social geography. In: Chorley R J, Haggett P (eds.) *Frontiers in Geographical Teaching*. Methuen, London

Pahl 1975 *Whose City?* 2nd edn. Penguin, Harmondsworth, UK

Rex J 1968 The sociology of a zone in transition. In: Pahl R E (ed.) *Readings in Urban Sociology*. Pergamon Press, Oxford, UK

Smith D 1977 *Human Geography: A Welfare Approach*. Edward Arnold, New York

Shevky E, Bell W 1955 *Social Area Analysis*. Stanford University Press, Stanford, CA

S. Bowlby

Social History

The term 'social history' refers to a subdiscipline of the historical sciences on the one hand and to a general approach to history that focuses on society at large on the other hand. In both manifestations social history developed from marginal and tentative origins at the end of the nineteenth and the beginning of the twentieth centuries and experienced a triumphant expansion from the 1950s to the 1980s. Throughout, social history can be best defined in terms of what it wants not to be or against that to which it proposes an alternative.

On one hand it distinguished itself from political history, which had been dominant during most of the nineteenth and the early twentieth centuries. Consequently, the 'social' in social history meant dealing with the structures of societies and social change, social movements, groups and classes, conditions of work and ways of life, families, households, local communities, urbanization, mobility, ethnic groups, etc. On the other hand social history challenged dominant historical narratives which were constructed around the history of politics and the state or around the history of ideas by stressing instead social change as a core dimension around which historical synthesis and diagnosis of the contemporary world should be organized. With these aims in mind, social historians (including historical sociologists and economic historians) sought to uncover the relationships between economic, demographic and social processes and structures, as well as their impact on political institutions, the distribution of resources, social movements, shared worldviews, and forms of public and

private behavior. For the comprehensive variant of 'social history,' the term 'history of society' (*Gesellschaftsgeschichte*) has gained currency. Not only in Western Europe and North America but also in other parts of the world, the establishment, expansion, and specialization of social history belong among the significant trends of interwar and particularly postwar intellectual and academic life. Nevertheless, in recent years, it shares with other fields of the humanities and social sciences a 'time of doubts' (Chartier 1993), a period of challenges from interpretative, constructivist, and discursive approaches (Palmer 1990, Bonnel and Hunt 1999).

1. *Early Institutionalization (Late Nineteenth–Early Twentieth Century)*

Like other generations of historians in the twentieth century that have at some point used the label 'new' for themselves, the claim to innovation by social historians in the 1960s and 1970s obscured the rich diversity of predecessors and competitors, which goes back to the second half of the nineteenth century (Kocka 1986). In that period, the academic professionalization and disciplinary differentiation of economics, sociology, and political science on the one hand and history on the other prevailed.

The mutual contact zones or, by contrast, the strength of boundaries were marked by significant national differences. The resulting interactions and hierarchies between these 'fields of knowledge' (Ringer 1992) differed in respect to the amount of space they provided for historical approaches. The 'Historical School' of political economy in Germany, dominant in the last third of the nineteenth century, is a case in point. It was part of economics and did not gain much recognition in the discipline of history. Some professional historians took part in a process of specialization in 'social and economic history' that kept a critical distance as much from this strand of economics as from the traditional history of great men, events, and ideas. The development of the combined field 'social and economic history' marked a double opposition, on the one hand to the history of political events and on the other hand to an increasingly ahistorical economic science.

The creation of a specialized journal in Germany allows one to give a precise date to the development of this subdiscipline of historiography. In 1893 the journal for social and economic history was founded, which was continued after ten years (and still exists today) as the *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte* (VSWG), under the direction of the Austrian socialist Ludo Moritz Hartmann and the conservative medievalist Georg von Below. That journal served as a direct reference for Lucien Febvre and Marc Bloch (see *Bloch, Mark Léopold Benjamin (1886–1944)*) when they founded the *Annales* in 1929 in Strasbourg. Not surprisingly, their journal was first called *Annales d'histoire économique et sociale* (1929–1939). The international exchanges and mutual observation between some of the protagonists of historiographic innovations in the early decades of the twentieth century were intense. Thus, only two years after the French flagship, the Polish journal *Roczniki dziejow spolecznych i gospodarczych* (Yearbooks for social and economic history) appeared. These early journals shared a further characteristic which distinguished them from the mainstream political history around them, namely a keen interest for 'universal,' non-nationalistic history.

In Germany, social and economic history remained a specialized branch without much impact on general history. In England, economic history maintained close links to social themes. After World War II, the French *Annales* historians tended to move their journal away from its earlier identification with this 'subdiscipline.' Similarly, the program represented by Ernest Labrousse for a history of class relations taking as its starting point the economic 'basis' broadened out more and more into the areas of urban history, quantitative analysis of social strata and the integral description of regional societies. Where 'social and economic history' continued to be an institutionalized specialty, it did not necessarily figure as an area of major innovations, as was later the case with the 'new' history of the 1960s and 1970s.

These early efforts to establish a specialized field and an alternative view of history with 'society' as the main object and organizing principle of analysis were paralleled by similar movements in various countries. In the USA, the *New History* of the turn of the century; in France, Henri Berr with his *Revue de Synthèse* (*Historique*) starting in 1900 and the reception of Durkheimian sociology by historians; in Germany, the development of a historical sociology by Max Weber and others; and in Great Britain in the first half of the twentieth century the impulses of George Unwin and R. H. Tawney for the development of an economic history that included social and cultural factors, were all among the most influential strands of socio-historical thinking (Iggers 1975).

Two characteristics of these debates and innovations deserve emphasis: the challenge of Marxism in a broad, nonorthodox sense provoked both emulation and animosity. Secondly, it is interesting to note in light of current discussions that especially in the German context the concept of 'culture' rather than that of 'society' was conceived as the vantage point for the synthetic organization of knowledge in the social sciences. In the 1950s the representatives of French 'new history' used the term 'history of civilization' to define the broadest possible framework of historical synthesis. Even Trevelyan's 1942 definition of social history as 'the history of a people with the politics left out' which is cited as often as it is refuted provides a surprisingly wide room for the 'social,' implicitly

including culture, mentalities, everyday life. (For a detailed treatment of the above topics see *Historiography and Historical Thought: Indigenous Cultures in the Americas*; *History and the Social Sciences*).

In recent times a further root of social history has attracted wider attention. It lies in the antimodern nationalist currents, critical of industrialization, which in particular in the 1920s and 1930s presented a special attraction for the younger social and human scientists. The interest in rural life, population, and race combined particularly in Germany with aggressive nationalism and criticism of the new political borders (Raphael 1997). This historical writing opposed conventional research subjects and forms; it turned away from the state and classical political history, looked for alternatives to the traditional focus on elites, and took as its example the methods which had already seen much progress in demography, geography, and anthropology. In the 1930s and 1940s a number of representatives of this so-called *Volksgeschichte* (history of the people) moved closer to National Socialism and became active as wartime political consultants. As some of these historians and social scientists played a significant role in the modernizing of the human sciences in the German Federal Republic, a debate arose as to the extent to which lines of continuity led from 'history of the people' to social history, and concerning the relation between its representatives' rightist views and their proposed 'structural history' of modern industrial society (Schulze and Oexle 1999, Conrad 1999).

2. *Postwar Expansion (1950s to 1970s)*

Not only in Western Europe and North America but also in other parts of the world, the establishment, expansion, and specialization of social history belong among the significant trends of interwar and particularly postwar intellectual and academic life. The spread of the 'new' histories since the 1960s occurred at a time when systems of higher education in Europe and North America were dramatically expanding. The general increase in academic jobs, student enrolments and Ph.D. candidates, as well as the massive growth in book publications and journals, allowed the various strands of social history to find ample space without replacing or pushing aside more traditional ways of writing and teaching history.

The institutional growth of social history followed different national directions. But it can generally be said that in the 1970s a high point was reached as far as the establishment within universities and among the larger public was concerned. The most impressive indicator is the founding of new journals concentrated around the year 1975: *Social Science History* (USA, 1975), *Geschichte und Gesellschaft* (West Germany, 1976), *History Workshop Journal* (UK, 1976); *Social History* (UK, 1976). Along with the *Journal of Social History* (USA) founded in 1967 they represented the new 'historical social science,' complementing the older 'school-forming' reviews such as above all *Annales ESC* (France, refounded 1946) and *Past and Present* (UK, 1952).

2.1 *National Variations and International Circulation of Ideas*

The triumph of social history in the 1960s and 1970s was a transnational phenomenon. The themes, questions, and methods were astonishingly similar in different countries; the international specialist public was interconnected through increased contacts. In many places the attractiveness and legitimacy of new approaches were aided by models coming from other countries. Particularly the French '*nouvelle histoire*,' often described by others as the *Annales* school, developed into a veritable export hit. Already at the international historians' congresses in Paris in 1950 and in Rome in 1955, its French representatives and Italian or English allies very much set the tone (Erdmann 1987). In the following decades the international circulation of concepts and approaches took on greater strength through the translation of books. A few historiographic models or theoretical influences (such as for instance the work of Fernand Braudel) became widespread in the countries of southern Europe and South America, while others won international recognition via their reception and dissemination on the North American market.

Yet for all that the national differences remained considerable, even in retrospect and after several decades of approaches and exchanges. The strengths of the French school remained—with a few exceptions—limited to the Middle Ages and the early modern period. Not the genesis of industrial society but the explanation of the revolution of 1789 was and remains the dominant theme for French historians. By contrast in Great Britain, an early market society and 'first industrial nation,' historians gave particular importance to these factors. The central approach which English social historians shared with social reformers and social scientists was however focused on the persistence and dynamism of class society.

It is by no means a coincidence that E. P. Thompson in his pioneering work *The Making of the English Working Class* (1963) made decisive strides for social history in this field. And the historical self-examination in other countries can also—though certainly not entirely—be characterized through central motives in their social development: the USA as an immigration and frontier society, in which ethnic cleavages are given more weight than class conflicts; Sweden with its experiences of mass emigration and social crises at the end of the nineteenth and beginning of the twentieth centuries; and finally Japan, which after

World War II looked to describe its latest developments in terms of Marxian-inspired modernization models (Conrad 1999). The political breaches and catastrophes of the twentieth century also formed the node of the new German social history, which was especially influenced by the political social history by the 'Bielefeld School' historians. The German *Sonderweg*, the German divergence from the West, and the social conditions for the national-socialist rule were the central preoccupations of a whole generation of historians who broke with 'historicist' (see *Historicism*) and individualizing explanatory models (Kocka 1999, Welskopp 1998).

2.2 Topics, Neighbors, Models

The place of social history in the intellectual landscape of the historical and social sciences of the post-World War II era can be understood by tracing the main trends with respect to its preferred research topics and questions, interdisciplinary alliances, methods, and explanatory models. These trends were most conspicuous in the period from the late 1950s to the 1970s, which witnessed the most pronounced change toward the 'new' social history. The reorientation effected by the spread of social history is most obvious with respect to the subjects chosen for monographs, in the programs of congresses or in the contents of major journals. The history of social classes and social stratification in general entered into the historical mainstream.

In researching social mobility and standards of life in the past, historians in Europe and North America tended to prolong the prehistory of their contemporary societies into the nineteenth century (and later even further back). Labor history (*Labor History*) left its subaltern and ideological niche and expanded its focus to the realities of work, family, and everyday life. The economic interests and social position of political actors were taken seriously and entered into the 'social interpretation' of great processes like the American and French revolutions, colonialism and imperialism, the development of political parties and social movements, or the causes of war. The fluctuation of harvests, prices and wages, the ebb and flow of population movements and land use, even microbes and the climate were considered with growing technical sophistication and by an increasing numbers of istorians.

Population history with the more micro-oriented historical demography became a leading sector of the new historiography: migration, demographic behaviour, household structure were reconstructed by large research projects. The Cambridge Group for the History of Population and Social Structure founded in 1964 by Peter Laslett, E. A. Wrigley, and Roger Schofield as well as the historical department of the French *Institut National d'Etudes Demographiques* represented the effort of making history 'scientific,' and influenced similar enterprises in Germany and Austria, the Scandinavian countries and North America among others. Similar aims were held by historically minded sociologists, political scientists, and economists who joined historians in the USA to create a discussion forum under the heading of 'Social Science History' (with conferences since the 1960s and a journal of the same name since 1975).

The flourishing subfields, each with their own journals, congresses, and experts, read like a repertoire of social history themes (Stearns 1994): labor, demographic, family, urban, rural, ethnic, women's, and gender histories emerged and became aspects of a more encompassing social history (*Social History*). Internal sequences and echoes were to be observed in the establishment of subdisciplines and the proliferation of new topics: the growing emphasis on mentality, agency, subjectivity, and experience since the late 1970s was itself a response to the establishment of abstract, often quantitative studies on standard of living, family and demography, social and geographic mobility, or political interest groups. Women's historians criticized the unspoken gender assumptions of much of historical demography and family studies, as well as of the history of labor markets and welfare states. The very meaning of the term 'social' was itself reconceived; it lost its direct association with 'socialist' movements or 'social policy' and embraced instead the whole abundance of the life-world. The enlarged thematic spectrum of the Amsterdam-based *International Review of Social History* (1956), the Parisian *Le Mouvement Social* (1960) or the German journal *Archiv fuer Sozialgeschichte* (1961), are examples of this transition which had gained considerable ground by the mid-1970s.

The shift in interdisciplinary alliances mirrored and underpinned topical interests. The nearest neighbors to history were no longer philology, philosophy, or law, but economics, political science, and sociology. For the French *Annales* School the close relationship with geography and demography was particularly important. The West German focus on a 'political social history' of the late nineteenth and early twentieth centuries led to an alliance with historical sociology and political science. Various strands of descriptive, often quantitative economics, social surveys, and sociography are found among the more practical roots of historians who embarked on studies of living standards, family formation, or business cycles. Equally it should not be forgotten that the enormous growth of sociology in the postwar era held a particular attraction for historians.

The varieties of social and cultural anthropology, or ethnology as it is called in some countries, served as models for a minority of social historians, especially those working on early modern history. The journal *Comparative Studies in Society and History* (founded in 1958) became an important forum for historically

minded anthropologists and anthropologically interested historians and social scientists. The relationship between history and anthropology—a discipline which has also changed dramatically during the last three or four decades—is a good indicator for the underlying change of emphasis from 'society' to 'culture' within social history. Natalie Z. Davis's influential collection of essays *Society and Culture in Early Modern France* (Davis 1975) or the wide impact of Clifford Geertz's writings, especially his 1973 collection *The Interpretation of Cultures*, are cases in point.

Parallel to these shifts, the protagonists and agents of the new social history reflected on the mechanisms of scientific change as such. What appeared to be happening under their own eyes and with their help in the humanities and social sciences seemed to resemble a 'paradigm shift,' as Thomas S. Kuhn had described it for physics in his influential book *The Structure of Scientific Revolutions* of 1962. Notwithstanding Kuhn's own doubts as to whether his model could be applied to the social sciences, the 'paradigm shift' became a battle cry in the nascent struggle to establish social history over and against the history of politics, great men, or ideas. The implications of Kuhn's work for the history of science were also welcomed by social historians. Indeed, the emphasis on social and external factors in the development and application of knowledge clearly corresponded to the efforts of social historians to stress context, interests, and structural constraints when dealing with ideas and ideologies (Appleby et al. 1994).

Although at the time many were fascinated by the idea of a 'paradigm shift,' in retrospect we see a more subtly differentiated, less abrupt transition. It is certainly true that in individual national academic traditions there were in part intense debates as to the 'right' sort of history. And in most cases a political dimension was also present, in which social history was defined as being more to the left, and political history more to the right. However rather than superseding an old paradigm, social historians could work on several levels (institutional, methodological, and thematic) and go about broadening and innovating their field within quite different time frames. If one sees the success story of social history as a pluralistic project fed from many different sources, it becomes easier to understand its expansion and, later, its reorientation in the direction of cultural history. In fact the many social-political and intra-academic forces which were involved in the rise of social history have continued to develop independently, and gradually cast doubt upon the plausibility of their own basic assumptions.

These definitional clarifications should not, however, obscure the impact of the new social history on the traditions of historiography. The historicist principles of the historians' craft were challenged by analytical, theory-driven approaches. The borrowing of concepts and even the adaptation of 'middle range theories' (R. K. Merton) as tools in historical research and writing has become a trademark of social historians as compared to traditional 'narrative historians' (Le Goff and Nora 1985, Fairburn 1999). The use of comparative methods gained wide currency (Tilly 1984). Theories of modernization (*Modernization, Sociological Theories of*) became more attractive, above all for historians of the nineteenth and twentieth centuries. Even if the explicit use of theories remained in general limited, process-oriented terms such as industrialization, urbanization, secularization, and professionalization became ever more widely used in historiography. And they were by no means limited to social history in the more restricted sense. Modernization, industrial society, even democratization, all were part of the meta narrative governing a good part of social history writ large—albeit with national differences and often heavy criticism from alternative political currents within social history (see for example Judt 1979). Yet despite evident tension the height of the social history wave was marked by a good deal of common ground.

The explanatory models proposed by the 'new' histories of the 1950s, 1960s, and 1970s shared the aim of explaining collective behavior and social relations by linking them to the structures of the economy and society. Although most opposed the rigidity of official Marxism-Leninism, many historians conceived of society as an architecture which resembled a vaguely materialist interaction between 'basis' and 'superstructure.' In her work on the French revolution Lynn Hunt explicitly rejected the widespread partition of monographs in three levels of economy/demography, society, and politics or mentality by simply turning it on its head (Hunt 1990, pp. 102–3). Interests, world views, political stances were largely understood as collective responses to objective social positions. Since the 1970s, social history has experienced shifts in its basic assumptions about causality that have led to a much less homogenous situation. But there seems to be a widespread refocusing first on agency and practices, secondly on meaning and identity, and thirdly on microprocesses.

2.3 *The Will to Synthesis*

The social-historical approach to history set other requirements on the writing of an overall historical account as conventional political history. The narrative structure of political events, the periodization by means of wars or revolutions, and the intentionality of a historiography oriented towards the 'big players' could not be simply transposed on to social history. In this new field syntheses therefore had to contain—whether one liked it or not—a reflection on the character of 'society.' The will to synthesis is for this reason part and parcel of social history's emphasis on a comprehensive approach to history. Here social

history sets itself off most clearly from the study of social and economic phenomena within a context which continues to be defined in political terms.

The authors of large syntheses in social history are for that reason also more or less explicitly protagonists of a 'history of society,' or of a *histoire totale* in the French tradition. Mikhail Rostovtzeff's *Social and Economic History of the Roman Empire* (1926) and Marc Bloch's *Société féodale* (1939–1940) are both early and distinguished examples. Fernand Braudel, Eric Hobsbawm, and Hans-Ulrich Wehler emphasized in their own way this requirement with monumental general histories. But they could do so only by consciously borrowing concepts and models from the neighboring social sciences, and by setting out theoretical considerations on social structure and on the motors of social change. Whether these involved models on temporal and geographic structures such as Braudel's, or a flexible use of Marxist theories such as Hobsbawm's, or an explicit use of modernization theories such as Wehler's—in each case this type of general history distinguished itself from the conventional historiography also in its explicit theoretical orientation. Yet a clear difference in the degree to which theory governs the historical narrative or the display of arguments remains between these works and those of historical sociologists such as Immanuel Wallerstein, Theda Skocpol, Charles Tilly, or Michael Mann—authors who nevertheless share common interests with the historians.

But also at a more humble geographical and temporal level social historians have continuously endeavored to reconstruct the object of their investigations, society, as a coherent if multifaceted system, sometimes even in a literary way. Above all however they continue writing within their respective national historical frameworks (Wehler 1987–1995). Often this takes the form of collections of articles in which various authors describe the individual aspects of society (a recent example is Thompson 1990). But even the short, broad-based handbook giving an overview of the state of research belongs in this category. In part this reflects the tendency to specific research on social classes and strata which leaves out further societal, cultural, and political contexts (Charle 1994). Although the emergence of the newer directions of gender, experience, and cultural history has increased the fear that social history would become even more fragmented, José Harris (1993), for example, has explicitly endeavored to integrate these very perspectives into a general historical account.

3. *The Revenge of Success*

However, social history's expansion in quantitative scale and conceptual scope was large enough to produce serious side-effects. First, specialization increased dramatically. Greater demand for technical expertise such as quantification, the creation of such subfields as historical demography, or the tendency of methodological approaches like oral history to take on a life of their own contributed to the differentiation of the field. Second, the proliferation of topics as well as of local or regional units of analysis deemed worthy of a monograph concurred to create a very colorful map of approaches and subjects. In the late 1980s and later, representatives of social history looking back on the success story of their field could made out the first signs of self-doubt already in the second half of the 1970s, at the highpoint of social history's intellectual and institutional glory (see contributions in Kocka 1989). More than any other approach to history, social history gradually came to be marked by fragmentation and dispersion. Third, social history as a general approach gained access to other historiographical periods and to other disciplines in the humanities. In literary criticism and modern languages the social history of literature became a trademark in the 1970s and early 1980s and, since then, has met with growing skepticism and disillusionment.

Certainly it must be said that these developments represent a tribute to the success of social history. In the course of differentiation and specialization, the new approaches and subfields which had established themselves under the heading of social history gradually changed their character and underlying priorities (Burke 1991). What is often described at the end of the twentieth century as a cultural 'turn' in fact emerged over the years. Indeed, disillusionment with the social history of the 1970s can best be understood not as a 'paradigm shift' but rather as an incremental, self-reinforcing, sometimes conflictual transition on different, non-unified, levels of historical work: research methods, preferred topics and sources, examples taken from other disciplines, new 'buzz words' and concepts, more hidden understandings of causality, agency, and evolution. Without question, the more recent set of shifts, from social history to the 'new cultural history,' makes us think harder about the establishment of social history and its concomitant changes of the societal and political context in earlier decades.

The recent 'cultural turn' can first of all be understood immanently, that is intra-academically. In its various manifestations, social history continued in the tradition of an oppositional approach against the historical mainstream. Small wonder that it has attracted attacks from every new wave of critical and revisionist historians as well as the revival of older (biographic, narrative) forms of historical writing. The concept of gender has in many countries been at the forefront of debates around structure and agency, macrosociological categories and deconstructivist criticism. Particularly in Germany, a fierce debate with aspects of a generation conflict divided the 'history of society' embodied in the Bielefeld school of *Gesellschaftsgeschichte* (Iggers 1975, Kocka 1986,

Wehler 1987–1995) and the 'history of everyday life' (*Alltagsgeschichte*) represented by grassroots history workshops and a number of younger academics at the margin of the profession (Ludtke 1995).

Yet one must also consider the external, extra-academic factors which in the last decades have changed the relationship of the social sciences to society. The social and political movements of the postwar years made their mark on social history. Especially in the USA the expansion of themes and problems in social history since the 1960s had been given a strong push by the corresponding conflicts in American society: the movements of Afro-Americans, of students, of women, and of gays and lesbians added to the traditional input from the workers' movement and radical intellectuals. Since the 1980s, the already proverbial trinity of race, class, and gender reflected these social forces in an academic discourse. In Europe, too, a similar, by no means linear elective affinity could be observed between social movements (the alternative left, the women's, ecological and peace movements) and fresh ideas in history (gender, everyday, and microhistory).

The 1980s and 1990s experienced a strong move toward independence and pluralization of identity politics and the associated attempts at claiming one's own history for oneself. Taken together these trends have aided the transition away from a history of anonymous social processes and structures. This reorientation became even stronger with growing critique of the ethnocentric dominance of industrial societies in social history leading up to this time. In recent years, the slow inroads which postcolonial discourses have also made into historical science have on the one hand furthered the significance of classical social history, this time in the countries of Asia, Africa, and Latin America. On the other hand the established concepts of modernization, development, progress, and culture were subject to a thoroughgoing critique and redefinition. Finally, the breakdown of actually existing socialism in 1989/1990 and the resulting political and economic transformation led to reflection on the basics of politics and the economy (Kocka 2000).

The answers to these challenges and interrogations are by no means clear. The voices which speak of a 'crisis' are not restricted to social history, but are directed at the historical sciences, if not the humanities in general. Two trends can be made out: beyond all methodological differences the self-reflection of the academic historian has moved to center stage. The idea that one can produce reliable knowledge about society—contemporary or past—was in 2000 taken far less for granted than a few decades previously (Bonnell and Hunt 1999). On the other hand the historicization of all forms of knowledge has become, as it were, a patent remedy in all debates. In opening up towards this sociocultural direction, it is true that social history may have lost in unity, perhaps in recognizability, but it has gained in intellectual dynamism.

See also: Business History; Cultural History; Economic History; Family and Kinship, History of; Gender and Feminist Studies in History; Gender History; Historical Demography; History and the Social Sciences; Labor History; Life Course in History; Modernization and Modernity in History; Social Movements, History of: General; Urban History; Work and Labor: History of the Concept; Work, History of; Working Classes, History of

Bibliography

Appleby J, Hunt L, Jacob M 1994 *Telling the Truth about History*. Norton, New York

Bonnell V E, Hunt L (eds.) 1999 *Beyond the Cultural Turn. New Directions in the Study of Society and Culture*. University of California Press, Berkeley, CA

Burke P (ed.) 1991 *New Perspectives on Historical Writing*. Cambridge University Press, Cambridge, UK

Charle, C 1994 *A Social History of France in the Nineteenth Century*. Berg, Oxford, UK

Chartier, R 1993 Le temps des doutes. *Le Monde*. Paris, March **18**: vi–vii

Conrad, S 1999 *Auf der Suche nach der verlorenen Nation. Geschichtsschreibung in Westdeutschland und Japan 1945–1960*. Vandenhoeck & Ruprecht, Göttingen, Germany

Davis N Z 1975 *Society and Culture in Early Modern France*. Stanford University Press, Stanford, CA

Erdmann K E 1987 *Die Ökumene der Historiker. Geschichte der Internationalen Historikerkongresse und des Comité International des Sciences Historiques*. Vandenhoeck & Ruprecht, Göttingen, Germany

Fairburn M 1999 *Social History. Problems, Strategies and Methods*. Macmillan, London

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Harris J 1993 *Private Lives, Public Spirit: Britain 1870–1914* (The Penguin Social History of Britain). Oxford University Press, Oxford, UK

Hobsbawm E J 1972 From social history to the history of society. In: Gilbert F, Graubard R (eds.) *Historical Studies Today*. Norton, New York

Hunt L 1990 History beyond social theory. In: Carroll D (ed.) *The States of 'Theory'. History, Art, and Critical Discourse*. Columbia University Press, New York

Iggers G G 1975 *New Directions in European Historiography*. Wesleyan University Press, Middletown, CT

Iggers G G 1997 *Historiography in the Twentieth Century: From Scientific Objectivity to the Postmodern Challenge*. Wesleyan University Press, Middletown, CT

Judt T 1979 A clown in regal purple: Social history and the historians. *History Workshop Journal* **7**: 66–94

Kocka J 1986 *Sozialgeschichte. Begriff, Entwicklung, Probleme*, 2nd edn. Vandenhoeck & Ruprecht, Göttingen, Germany

Kocka J (ed.) 1989 *Sozialgeschichte im internationalen Überblick. Ergebnisse und Tendenzen der Forschung*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany

Kocka J 1999 Asymmetrical historical comparison: the case of the German *Sonderweg*. *History and Theory* **38**: 45–50

Kocka J 2000 Historische Sozialwissenschaft heute. In: Nolte P, Hettling M et al. (eds.) *Perspektiven der Gesellschaftsgeschichte*. Beck, Munich, Germany
Le Goff J, Nora P (eds.) 1985 *Constructing the Past: Essays in Historical Methodology*. Cambridge University Press, Cambridge, UK
Ludtke A 1995 (ed.), *The History of Everyday Life*. Princeton University Press, Princeton, NJ
Palmer B D 1990 *Descent into Discourse. The Reification of Language and the Writing of Social History*. Temple University Press, Philadelphia, PA
Raphael L 1997 Die 'neue Geschichte' – Umbrüche und neue Wege der Geschichtsschreibung in internationaler Perspektive (1880–1940). In: Küttler W et al. (eds.) *Geschichtsdiskurs*, Vol. 4. Fischer, Frankfurt/Main, Germany
Ringer F 1992 *Fields of Knowledge. French Academic Culture in Comparative Perspective, 1890–1920*. Cambridge University Press, Cambridge, UK
Schulze W, Oexle O G (eds.) 1999 *Deutsche Historiker im Nationalsozialismus*. Fischer, Frankfurt/Main, Germany
Stearns P N (ed.) 1994 *Encyclopedia of Social History*. Garland, New York
Thompson F M L 1990 (ed.) *The Cambridge Social History of Britain*, 3 Vols. Cambridge University Press, Cambridge, UK
Tilly C 1984 *Big Structures, Large Processes, Huge Comparisons*. Russell Sage Foundation, New York
Wehler H-U 1987–95 *Deutsche Gesellschaftsgeschichte*, 3 Vols. Beck, München, Germany
Welskopp T 1999 Westbindung auf dem 'Sonderweg'. Die deutsche Sozialgeschichte vom Appendix der Wirtschaftsgeschichte zur historischen Sozialwissenschaft. In: Küttler W et al. (eds.) *Geschichtsdiskurs*. Fischer, Frankfurt/Main, Germany, Vol. 5

C. Conrad

Social Identity, Psychology of

1. Social Identity

Social identity is those aspects of the self-concept that are derived from belonging to social groups and categories. Social identity theory was developed by Henri Tajfel and colleagues at the University of Bristol in the 1970s and 1980s. Social identity is a concept that continues to be used widely in the social psychology of intergroup relations, social influence, and the self. This article describes social identity theory and associated work on intergroup relations and group processes. It points to the diverse areas of research to which the theory has contributed and concludes with issues that are currently being explored.

2. Minimal Conditions for Intergroup Bias

In the 1960s and 1970s social psychology underwent a crisis of confidence that generated a powerful critique of individualistic theorizing and research. In response, European social psychologists began to develop theories that concerned the relationship between individuals and society. Social identity connected cognitive processes of categorization with the social structure of society. Research on social identity therefore ranges from the detailed analysis of cognitive processes underlying stereotyping and prejudice to wide-ranging questions such as reactions to group disadvantage. Tajfel's early work on social categorization included development of the minimal group paradigm. In this widely used paradigm people are randomly allocated to categories that have little meaning (e.g., A group and B group, or people who prefer Klee vs. Kandinsky). Research participants are unaware of which other individuals belong to the category, or what their characteristics might be. Evidence shows that minimal group membership is a sufficient basis for people to allocate points or money preferentially to anonymous members of their own category. They also evaluate in-group members more positively, although it seems that reward allocation and evaluations are not highly correlated. If a trivial categorization is sufficient to generate in-group bias, it can be reasoned that more significant and meaningful categorizations have substantial real-word consequences. An important theoretical implication of in-group bias in the minimal group paradigm is that in-group bias does not necessarily have its roots in objective or realistic intergroup conflict or in personality differences or interpersonal dependency.

3. Social Identity Theory

Social identity theory (Tajfel and Turner 1979) offers an explanation for minimal intergroup bias, and also a broader statement of how relationships between real-world groups relate to social identity. According to social identity theory, there is a continuum on which identity and relationships can be depicted as purely personal and interpersonal at one extreme, and purely social, or intergroup at the other. Personal identity is derived from personality, idiosyncratic traits and features, and particular personal relationships with other individuals. A special emphasis of social identity theory is that an effective account of group and intergroup phenomena is not possible if the theory only considers personal identity and relationships. Social identity is the individual's knowledge that he or she belongs to certain social groups together with the emotional and value significance of the group memberships. Social identity affects and is affected by intergroup relationships. The theory aims to explain the uniformity and coherence of group and intergroup behavior as mediated by social identity (see Hogg and Abrams 1988 for a detailed account).

3.1 Categorization

Categorization leads to accentuation of differences among objects in different categories, and accentu-

ation of similarities of objects within categories. When social categorizations are salient, accentuation results in stereotypical perceptions of both categories. When people categorize one another they generally include the self in one of the categories. Attributes that are stereotypical for the in-group may then be ascribed to self, a process labeled self-stereotyping. As a result of self-stereotyping the self is 'depersonalized.' This means that the self and other in-group members are seen as interchangeable in their relationship to out-group members.

3.2 Self-categorization

Self-categorization theory (Turner et al. 1987) extends social identity theory. It proposes that the salience of self-categorization depends on a combination of the relative accessibility of the categories (i.e., the perceiver's readiness to make use of them) and the comparative and normative fit between the social stimuli (i.e., people being perceived) and the available categories. The more that the targets are classifiable by a clear criterion (such as gender) and the more that their behavior seems to fit the category normatively (e.g., being domineering vs. considerate), the more readily the categorization will be applied. The categories will be represented in terms of prototypes based on a meta-contrast ratio that minimizes the differences within categories and maximizes the differences between categories on criterial dimensions. Categorizing self and others involves perceiving them in terms of their closeness to the relevant group prototype, rather than in terms of their unique personal characteristics. According to self-categorization theory, stereotyping of in-group members, out-group members, and the self is a normal functional process that responds flexibly to the particular categories that are being compared. The theory therefore challenges traditional models of stereotyping as a form of bias, and of the self-concept as a stable psychological structure (see Abrams and Hogg 2001).

3.2.1 Social influence. Depersonalization is thought to underlie social influence within groups. When social identity is salient, in-group prototypes serve as norms or reference points. For example, attraction to group members is affected more strongly by shared social identity than by interpersonal similarities (Hogg 1992). The coordinated and consistent behavior of crowd members also seems to be more explicable in terms of a shared social identity than in terms of characteristics of the individuals, resulting from a process of referent informational influence (Turner 1991). In classic social influence paradigms such as Asch's conformity experiments, or Sherif's norm formation experiments, people are more strongly influenced when they share a category with the source of influence than when they do not. Moreover, reactions to group deviants are much more hostile if they are deviating toward an out-group norm (Marques et al. 1998).

3.3 Motivation

3.3.1 Positive distinctiveness. Social identity theory offers a motivational explanation for in-group bias. First, judgments about self as a group member are held to be associated with the outcome of social comparisons between the in-group and relevant out-groups. Second, it is assumed that people desire a satisfactory self-image, and positive self-esteem. Positive self-evaluation as a group member can be achieved by ensuring that the in-group is positively distinctive from the out-group. Usually group members will engage in social competition with out-groups to try to make the in-group positively distinctive. For example, in minimal group experiments, people show a consistent bias both towards maximizing in-group profit and toward maximizing differential profit in favor of the in-group, even when the total in-group profit suffers. The theory does not argue that material considerations are unimportant, but that the symbolic meaning of the group's position relative to other groups is a powerful motivating consideration.

3.3.2 The self-esteem hypothesis. The motivational component of the theory, sometimes described as the 'self-esteem hypothesis,' has been scrutinized in some detail (Abrams and Hogg 2001). Two corollaries of the hypothesis are that expression of in-group bias should maintain or increase self-esteem, and that lowered self-esteem should motivate the expression of in-group bias. These hypotheses only relate to the specific evaluations of self as a group member immediately before and after judgments or allocations have been made toward the in-group and the out-group. However, few empirical tests have adhered to these requirements, and the evidence supporting them is weak or mixed. Evidence does suggest that people with higher personal self-esteem are more likely to show in-group bias and favoritism. As a consequence of theoretical and practical difficulties of the self-esteem hypothesis, other motivational processes have been suggested. Brewer (1981) has proposed that people seek optimal distinctiveness, such that they have some assimilation to a group, but the group has some distinctiveness from other groups. Hogg and Abrams (1993) have suggested that people are motivated to reduce uncertainty, and that clarification of group membership, adherence to group norms, and associating positive group features with

the self are ways to achieve this. However, apart from conformity to pre-existing biased in-group norms, there seems to be no strong alternative candidate to the self-esteem hypothesis as an explanation for in-group bias.

3.4 Social Structure

A second component of social identity theory explicitly addresses the status relationships among groups. To maintain or enhance social identity, people may engage in a variety of strategies. The choice of strategy will depend on perceptions of the relationship between their own and other social groups. The boundaries between groups may be viewed as permeable or impermeable, the status differences as legitimate or illegitimate, and the nature of these differences as being stable or unstable. When status differences are seen as legitimate and stable, the differences between groups are secure. If group boundaries are permeable, the individuals may have a social mobility belief structure. In this situation they may strive to improve their identity by moving to the higher status group. For example, it is possible to go from school to university by passing the relevant entrance examinations. Social mobility belief structures are strongly endorsed by Western individualistic democracies, in which individual effort and ability are viewed as the basis for the allocation of social goods and resources. When boundaries between groups are viewed as impermeable (e.g., that between men and women, or between blacks and whites), members consider that their social identity can only be improved if the social relationship between the groups changes (a social change belief structure). When the social structure is viewed as illegitimate and/or unstable, the intergroup relationship is 'insecure,' and it becomes possible to imagine cognitive alternatives to the status quo. Members of lower status groups will engage in social competition with higher status groups to evaluate and reward their own group more positively. In the absence of cognitive alternatives, group members may still find ways of conferring positive distinctiveness on the in-group, using a social creativity strategy. Social creativity strategies include comparing the groups on new (in-group favoring) dimensions, trying to re-value the existing dimensions of comparison so recasting in-group attributes positively, and engaging in comparisons with other, lower status, out-groups. There is considerable evidence from experimental tests of the theory that these strategies are adopted (e.g., Ellemers et al. 1999).

3.5 Language

Social identity has been studied widely in natural intergroup settings. One important area of research is language use and maintenance. Language is an important marker of social identity. For example, use of the matched guise technique, in which the same speaker delivers a passage using different accents or dialects, has revealed that judgments of in-group and out-group speakers reveal clear in-group favoritism. Moreover, speakers' accents converge or diverge depending on whether they view the relationship positively. The ability of a group to sustain its language may reflect its ethnolinguistic vitality, which is a function of the perceived status, demography, and institutional support that it has (Giles and Johnson 1987). Finally, there seems to be a linguistic intergroup bias in which people describe positive in-group and out-group attributes in more abstract terms using more specific terminology (Semin and Fiedler 1991).

3.6 Collective Behavior

A second significant area of research is collective behavior and protest. Traditional theories, such as relative deprivation or frustration–aggression theory, have depicted mass protest, rioting, and revolution as products of frustration or dissatisfaction felt by individuals. However, research has shown that social protest and mobilization is predicted by the perception of deprivation of the group relative to other groups. In turn, collective action is more likely when group members feel there is no legitimate alternative, no possibility of social mobility, and when they identify strongly with their group. Some evidence suggests that collective action is usually a last resort, but other evidence suggests that participation in collective action is likely when people identify strongly with their group (Kelly and Breinlinger 1996).

3.7 Organizational Behavior

Social identity is also highly relevant to organizational behavior. Commitment to organizational goals and willingness to contribute to organizational activities, as well as lowered absenteeism and turnover intentions, have all been shown to be associated with higher identification with organizations (Ashforth and Mael 1989, Tyler and Smith 1998). Moreover, people's endorsement of leaders seems to be affected at least as much by the leader's perceived prototypicality (representativeness of the specific in-group norm) as by how effective or stereotypically leader-like the person is.

4. Critical Comments and Future Directions

The concept of social identity is widely accepted in social psychology and has contributed to a large volume of research (Robinson 1996). Current work is moving away from the idea that group members are

primarily motivated by self-enhancement, and examines the idea that people may be motivated to reduce uncertainty or find meaning from their group memberships. In addition, identification may be manifested in different ways (Deaux 1996), and there are individual differences that affect people's behavior, such as differences in how strongly they identify with particular groups, differences in their social dominance orientation, and differences in their willingness to express prejudice. It seems that the strength with which people identify with groups, and the degree of threat to social identity, can moderate the relationship between group status and intergroup bias (Ellemers et al. 1999). Current work is also examining the cognitive processes that affect the self–group relationship. For example, some researchers argue that the self is the primary psychological unit and that it simply extends to include in-groups, in contrast to the idea that the self can be redefined by salient self-categorizations. Recent research also shows evidence of a rapprochement between social cognition approaches to intergroup relations and the social identity perspective (Abrams and Hogg 1999).

See also: Ethnic Identity/Ethnicity and Archaeology; Ethnic Identity, Psychology of; Identity and Identification: Philosophical Aspects; Identity in Anthropology; Identity in Childhood and Adolescence; Identity: Social; Multiculturalism and Identity Politics: Cultural Concerns; Personal Identity: Philosophical Aspects; Regional Identity

Bibliography

Abrams D, Hogg M A (eds.) 1999 *Social Identity and Social Cognition*. Blackwell, Oxford, UK

Abrams D, Hogg M A 2001 Collective identity: group membership and self-conception. In: Hogg M A, Tindale R S (eds.) *Handbook of Social Psychology: Group Processes*. Blackwell, Oxford, UK, pp. 425–60

Ashforth B E, Mael F 1989 Social identity theory and the organization. *Academy of Management Review* **14**: 20–39

Brewer M B 1991 The social self: on being the same and different at the same time. *Personality and Social Psychology Bulletin* **17**: 475–82

Deaux K 1996 Social identification. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology: Handbook of Basic Principles*. Guildford Press, New York, pp. 777–98

Ellemers N, Spears R, Doojse B 1999 *Social Identity*. Blackwell, Oxford, UK

Giles H, Johnson P 1987 Ethnolinguistic identity theory: a social psychological approach to language maintenance. *International Journal of Sociol Language* **68**: 256–69

Hogg M A 1992 *The Social Psychology of Group Cohesiveness*. Harvester Wheatsheaf, Hemel Hempstead, UK

Hogg M A, Abrams D 1988 *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*. Routledge, London

Hogg M A, Abrams D 1993 *Group Motivation: Social Psychological Perspectives*. Harvester Wheatsheaf, Hemel Hempstead, UK

Kelly C, Breinlinger S 1996 *The Social Psychology of Collective Action: Identity, Injustice and Gender*. Taylor and Francis, London

Marques J M, Abrams D, Martinez-Taboada C, Paez D 1998 The role of categorization and in-group norms in judgments of groups and their members. *Journal of Personality and Social Psychology* **75**: 976–88

Robinson P (ed.) 1996 *Social Groups and Identities: Developing the Legacy of Henri Tajfel*. Butterworth-Heinemann, Oxford, UK

Semin G R, Fiedler K 1990 The linguistic category model: its bases, applications and range. In: Stroebe W, Hewstone M (eds.) *European Review of Social Psychology*. J. Wiley, Chichester, UK, Vol. 2, pp. 1–30

Tajfel H, Turner J C 1979 An integrative theory of intergroup conflict. In: Austin W G, Worchel S (eds.) *The Social Psychology of Intergroup Relations*. Brooks-Cole, Monterey, CA

Turner J C 1991 *Social Influence*. Open University Press, Milton Keynes, UK

Turner J C, Hogg M A, Oakes P J, Reicher S D, Wetherell M 1987 *Rediscovering the Social Group: A Self-categorization Theory*. Blackwell, Oxford, UK

Tyler T R, Smith H J 1998 Social justice and social movements. In: Gilbert D, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, pp. 595–629

D. Abrams

Social Inequality and Schooling

Do systems of education in modern societies help to break down the inheritance of privilege from one generation to the next, thereby providing opportunities for mobility for members of less privileged groups, or do schools largely reinforce and legitimate the inequalities among individuals evident in the larger society? This article shows that social background affects educational attainment; discusses mechanisms through which family privilege is translated into advantaged educational opportunities and achievement; identifies unequal school resources and opportunities as determinants of inequality in outcomes; and shows that egalitarian-based school reforms have generally failed to reduce social inequalities in education significantly. Unequal educational outcomes derive both from inequality in educational opportunities and inequalities in family environment and resources.

1. Social Background and Educational Inequality

1.1 Basic Relationships

A relative constant about education in the modern world is that social background has a major impact on educational performance and on educational attain-

ment, even net of ability: on average, students from more advantaged class and/or status backgrounds perform better in school and go further in the educational system, thereby obtaining higher-level educational credentials (Ishida et al. 1995). The association between family background and educational outcomes is even stronger in the developing world than in advanced industrial societies and is less strong in the USA than in the many European countries in which more explicit class and gender barriers in access to schooling (e.g., in the form of fees for secondary schooling and gender-segregated schools, respectively) have existed historically (see Brint 1998). Given that educational performance and credentials are important determinants of adult success, education thereby plays an important role in the reproduction of social inequality. Some theorists propose that education is a form of 'social capital' that has replaced property as a means of intergenerational status transmission (Bourdieu 1973). Whereas elite families once ensured their children's future positions by passing on property, they now ensure their children's futures by helping them obtain the 'best' educational credentials. Nonetheless, even the strongest proponents of class reproduction theory recognize that education does not perfectly reproduce the class/status structure of modern societies. It also provides opportunities for some talented individuals from less advantaged backgrounds to succeed.

Social scientists have for decades tried to determine why class and status inequalities in educational outcomes occur. Part of the explanation, as discussed more fully below, is or has been due to unequal educational opportunities. Further, the growth of mass education throughout the world has nominally equalized educational opportunities at some levels (primary education in the developing world, secondary education in the developed world) while displacing the level of education that has real consequence for adult success to a higher level—one to which elites retain privileged access (e.g., the university level in developed countries). School expansion, therefore, has not reduced the educational advantage that elite children enjoy (Walters 2000, pp. 254–7).

But even when educational access is formally equal, inequalities in outcomes still occur. Some of this is due to class or status segregation within schools or school systems: Disadvantaged children tend to go to schools with largely disadvantaged student bodies (i.e., there is between-school segregation), or be in streams or tracks within schools that are fairly socially homogeneous (i.e., there is within-school segregation). That is, the proximate schooling experience is class- and status-segregated to a large extent in many societies, and this segregation has an adverse effect on disadvantaged children's achievement (Oakes et al. 1992).

Some of the inequality in educational outcomes is also due to the direct effects that families and family background have on learning, achievement, and educational experiences, aside from wealthy parents' ability to buy private schooling for their children or, in some developing countries at present, poor families' dependence on child labor, which interferes with schooling. Children from more advantaged families, for example, are more likely to start school with strong academic skills, be helped in their school work by their parents, have parents who actively intervene in their schooling careers, exhibit linguistic patterns preferred by teachers, have resources to buy learning materials or supplemental help, and in general live in a family environment that supports learning and academic accomplishment (Brint 1998, pp. 208–14).

In many societies, educational programs have been established to help disadvantaged children make up for their family-based learning 'deficits' (one notable example is Project Head Start in the USA). With the exception of far-going educational reforms in some socialist states, however (see Sect. 2.3), reforms have not attempted to deprive children from advantaged families of most of their sources of educational advantage.

1.2 Bases of Social Inequality

The particular social groups that are educationally 'advantaged' and 'disadvantaged' vary by places and times, depending on the stratification system in place in the larger society. Generally, those groups that are advantaged in society at large are advantaged with respect to educational opportunities and educational outcomes. In most societies in the modern world, social class position confers tremendous educational advantage or disadvantage. Children from advantaged social classes have better educational opportunities, have higher levels of 'achievement' as officially measured and recognized by the school, and receive on average 'better' educational credentials. In capitalist societies, upper-class and upper-middle-class children reap these educational class advantages. In socialist societies, children of party elites reaped educational class advantages, official policy notwithstanding (see Shavit and Blossfeld 1993). Class is arguably the most important source of educational advantage or disadvantage, and the source that has been most consistent in recent decades.

Another common social distinction that has important educational consequences is race and ethnicity. The degree of educational disadvantage associated with being a member of a disadvantaged racial or ethnic group has varied historically. The USA and South Africa, for example, historically had public educational systems that segregated blacks into separate schools and offered them far fewer educational opportunities; in both cases, changes in state policy eliminated the explicit racial barriers to school access although some forms of racial inequality in education persist. There is also significant cross-

cultural variation in the racial or ethnic groups that are educationally disadvantaged: For example, Koreans and other Asians perform better than whites in American schools but are educationally disadvantaged in Japan (Brint 1998, p. 216).

The final major basis of social inequality in education is gender. Gender inequality in education—both in terms of access and outcomes—has decreased dramatically in recent decades in most advanced industrial countries. Whereas girls were once commonly barred from certain types or levels of education, gender barriers in access are insignificant in most industrial countries at present but remain in those less developed countries in which women are still largely restricted to the private (family) sphere. Similarly, in most industrial countries schools offer similar if not identical curricular opportunities at present to boys and girls, and girls outperform boys in some important respects (e.g., in the USA girls earn better grades on average than boys and have higher rates of graduation from high school and college). The significance of gender as a basis of social inequality in education has declined significantly, then, in many parts of the world but remained high in societies in which gender remains a significant basis of differentiation in the larger society.

2. *Social Inequality in School Access and Resources*

2.1 *Translating Advantaged Family Background to Educational Advantage*

The ways in which children from advantaged families gain access to advantaged schools or schooling experiences vary from society to society. Most European societies have long been known for their class-based systems of education (which are also often ethnic-based), characterized by a clear differentiation into educational streams at relatively early ages. Children from advantaged backgrounds were more likely to enter streams leading to university-level education at the ages of 10 or 12, whereas at the same time working-class children or racial and ethnic minorities were more likely to enter streams leading to earlier school exit and vocational training. At present, some European states retain a system of early branching (e.g., Germany and Austria), albeit in modified form from prior decades. Others, however (e.g., France, England, and Sweden), have eliminated early branching, have combined vocational and academic training in the same secondary school buildings, and allow easier movement among streams at the high-school level—in part because of the explicit recognition that early and relatively permanent branching maximizes the effect of family background on schooling outcomes (see Brint 1998, pp. 29–64).

But even in the USA—a society whose educational system is regarded as more open, in which most students attend comprehensive high schools, in which most students get second, third, or more chances for access to high-level education, and in which a private school secondary education is less important for access to the top universities—inequality in access to 'good' educational opportunities is still evident. In part this is due to the ability of advantaged parents to use their personal resources to send their children to expensive private schools if they so choose. But largely it is due to wide variation in the USA among states, within states, and among school districts in the quality of public schools. This variation favors those states and localities in which advantaged families predominate, because of the historic reliance in the USA on local property taxes to fund public education. Unlike the financing basis of state-supported schooling in many other industrialized countries, this system allows wide variation in per-student school funding.

This reliance on local taxes in the USA has produced a system where the 'best' schools—the ones with the most expensive facilities, the better-trained and most highly paid teachers, the most advanced curricular opportunities that 'produce' students with the highest standardized test scores, and send the highest proportion of their graduates to college—are located in those local areas with the highest concentrations of wealth, largely very wealthy suburbs of large metropolitan areas (Kozol 1987). This system produces sharp inequalities in educational access by class and race, the latter because African-Americans are concentrated in central cities, which have crumbling and fiscally strapped public-school systems.

2.2 *The Importance of School Resources*

Starting with the publication called the Coleman Report in the United States (Coleman et al. 1966), numerous studies by social scientists have been interpreted to show that schools—or, more precisely, resources devoted to schools—do not matter. Coleman found that family socioeconomic background is a much more important determinant of educational achievement as measured by standardized test scores than spending on schools. One implication of this finding is that any observed differences in public moneys spent on schools attended by the elite versus schools attended by the disadvantaged is irrelevant with respect to class or status differences in educational outcomes. In other words, families are implicated in the creation of inequality in educational outcomes, but schools are not.

As politically popular as this argument has proven to be, at least in the USA, a number of recent meta-analyses show that the existing studies offer stronger proof of a positive effect of resources on student test scores than an absence of an effect (e.g., Hedges et al.

1994; but for a counter interpretation see Hanushek 1994). Further, an important longitudinal experimental study in Tennessee, in which students were randomly assigned to different-sized classes, demonstrated that placement in relatively small classes, an important type of school resource, had a significant effect on students' performance on standardized tests in the first and subsequent years of the study (Krueger 1999).

There are other problems with the argument that inequality in school resources is inconsequential with respect to educational outcomes. One, Coleman's (and many others') measure of educational outcome was ability, not achievement, years of schooling completed, entry to college, or the like, all of which are far more consequential for adult success than one's score on an ability test. A number of studies have found positive effects of school resources on educational attainment and on adult earnings (see review in Card and Krueger 1996). Two, aggregate school-level family social economic status and school funding co-vary, making it hard to disentangle the effects of family background and school spending. Three, what families 'know' themselves about the importance of going to a well-funded school belies the argument that school spending doesn't matter: when given the opportunity, families go to extraordinary lengths to get their own children placed in resource-rich schools.

2.3 Egalitarian School Reforms

Although it is doubtful that eliminating class, race, gender, and other bases of inequality in access to schooling or the quality of schooling would result in equality in educational outcomes (because of the strong effects family background have on school performance and persistence), it is clear that the presence of class, race, gender or other barriers to access to school or to particular levels of school promotes social inequality in education and in adult life. Thus, in the past several decades many countries have instituted school reforms intended to eliminate historic patterns of unequal access to education. In those places where racial minorities were allowed only separate—and unequal—educations, desegregation was an important step toward equalization of opportunity. In societies in which class barriers to secondary education existed in the form of school fees and the like, such as Ireland, elimination of fees was intended to equalize class access to secondary education (Raftery and Hout 1993). Some socialist societies went much further, instituting systems that gave clear preference in access to education to children from formerly disadvantaged backgrounds (see Shavit and Blossfeld 1993). Interestingly, in very few cases did these reforms, even the most far-reaching ones, succeed in eliminating former class or status disadvantages in schooling. Social differentials in educational attainments, for example, have proven remarkably resistant to egalitarian school reform (Shavit and Blossfeld 1993). One major reason for this is that increasing the access of formerly disadvantaged groups to a particular level of schooling—say secondary education—increases the proportion of advantaged children attending as well *and* generally does not significantly alter class or status effects on the likelihood of entering that level of education (Raftery and Hout 1993, p. 60). Further, once secondary education, for example, is readily available to formerly disadvantaged groups, higher education becomes the significant arena in which class and status advantages are preserved (Walters 2000). Gender is an exception: in some Western societies, women are now more likely than men to obtain baccalaureate degrees, although not Ph.Ds.

It appears that the only way to significantly reduce class or status inequalities in educational outcomes is to combine egalitarian school reform with a variety of social policies that significantly reduce inequalities in living conditions, family-based resources, and occupational opportunities. Of the 13 countries covered by Shavit and Blossfeld's (1993) important edited book, for example—including a range of countries from capitalist to state socialist—only in Sweden and the Netherlands were egalitarian school reforms coupled with social welfare policies that substantially reduced inequalities in family resources and did educational inequality significantly decline. Further, the substantial decline in gender inequality in education witnessed in some societies is due to school-based reforms *as well as* a significant decline in gender stratification in the public sphere.

The failure of sweeping school reform in a number of socialist states, including 'reverse discrimination' policies intended to favor children of working-class or peasant backgrounds, is particularly illuminating. To an important degree, new elites took the place of old elites (e.g., party elites replaced old class elites) but children of elites still enjoyed educational advantages. Further, none of these reforms closed off all possible means by which even former elites could ensure their children's educational advantage: for example, bribes, favors based on personal ties, passing on various forms of cultural capital that helped children overcome formal barriers to their access to privileged schooling, and the lingering effect of family financial capital (see summary in Walters 2000, p. 255).

3. Summary

Social inequality in education is an enduring feature of the twentieth century. There have been changes in which social groups are advantaged and disadvantaged, but groups that are disadvantaged in the larger society are generally—though not always—disadvan-

taged with respect to educational achievement and attainment, if not with respect to educational opportunities. Many societies have implemented changes that make access to education formally more open to disadvantaged groups, but these changes and reforms have not succeeded in creating equality of educational outcomes. Inequality in family-based resources and patterns of stratification and inequality in the larger society continue to be determinants of social inequality in education.

See also: Academic Achievement: Cultural and Social Influences; Economic Status, Inheritance of: Education, Class, and Genetics; Education and Gender: Historical Perspectives; Educational Sociology; Equality and Inequality: Legal Aspects; Equality of Opportunity; Gender, Class, Race, and Ethnicity, Social Construction of; Inequality: Comparative Aspects; Social Inequality in History (Stratification and Classes)

Bibliography

Bourdieu P 1973 Cultural reproduction and social reproduction. In: Brown R (ed.) *Knowledge, Education and Cultural Change*. Tavistock, London, pp. 71–112

Brint S 1998 *Schools and Societies*. Pine Forge Press, Thousand Oaks, CA

Card D, Krueger A B 1996 School resources and student outcomes: An overview of the literature and new evidence from North and South Carolina. *Journal of Economic Perspectives* **10**: 31–50

Coleman J S et al. 1966 *Equality of Educational Opportunity*. Government Printing Office, Washington, DC

Hanushek E A 1994 *Making Schools Work: Improving Performance and Controlling Costs*. Brookings Institution, Washington, DC

Hedges L V, Laine R, Greenwald R 1994 Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher* **23**: 5–14

Ishida H, Muller W, Ridge J M 1995 Class origin, class destination, and education: A cross-national study of ten industrial societies. *American Journal of Sociology* **101**: 145–93

Kozol J 1987 *Savage Inequalities: Children in America's Schools*. HarperCollins, New York

Krueger A B 1999 Experimental estimates of education production functions. *Quarterly Journal of Economics* **114**: 497–532

Oakes J, Gamoran A, Page R N 1992 Curriculum and differentiation: Opportunities, outcomes and meanings. In: Jackson P W (ed.) *Handbook of Research on Curriculum*. American Educational Research Association, Washington, DC, pp. 570–608

Raftery A E, Hout M 1993 Maximally maintained inequality: Expansion, reform, and opportunity in Irish education, 1921–75. *Sociology of Education* **66**: 41–62

Shavit Y, Blossfeld H-P (eds.) 1993 *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*. Westview Press, Boulder, CO

Walters P B 2000 The limits of growth: School expansion and school reform in historical perspective. In: Hallinan M (ed.) *Handbook of the Sociology of Education*. Kluwer Academic/Plenum Publishers, New York, pp. 241–64

P. Barnhouse Walters

Social Inequality in History (Stratification and Classes)

As a universal fact of all human civilization, social inequality has been shaping advanced societies in a most fundamental way for many thousands of years, especially since the rise of more complex cultures and political systems in the wake of the 'Neolithic Revolution.' The concept of social inequality refers to an unequal distribution of positions in society, a distribution which is determined by structures of power (e.g., the modern state and its administrative elites), of technology and economy (e.g., industrial technology and capitalism), and of cultural expression and perception (e.g., Christian religion and its attempt to define and explain a 'holy' order of worldly society).

In modern societies, social inequality is often framed as a differentiation of society in distinct groups of similar positions conceivable as 'layers' of the social order on a vertical scale. A general term for this type of social inequality structure is 'stratification.' In the Western tradition, the concept of stratification often—in a more or less explicit way—refers to a ranking of social positions according to the distribution of economic resources; the conflicting groups resulting from this are frequently called 'classes' (in a broad sense). The notion of a society divided into classes, which through their collective action fundamentally shape the dynamic processes of society, has been most widely applied to the specific historical case of Western industrial societies with their market economy, but has also been transferred from here to the description of agrarian societies in ancient or medieval times.

Since around 1970, concepts of stratification and class have been criticized for their neglect of other dimensions of social differentiation and inequality such as gender, race, or ethnicity. Important as this recent restructuring of the debate on historical (and contemporary) modes of inequality is, stratification and class have been essential concepts for social history in the past (and appear to remain so in the future, if in a modified way). Indeed, it may be said that the development of social history as a field of historical research is hardly conceivable without the analytical concepts of stratification and class. In considering social inequality, stratification, and class in history, two perspectives may be taken. First,

stratification is an empirical problem, a problem of historical subject matter. In this regard, one can follow the development of systems of social inequality throughout history, from antiquity to the twenty-first century. Second, stratification refers to an analytical and theoretical problem in (social) history; it appears as a problem of conceptualizing historical processes and structures. In this vein, one can look at concepts, approaches, and methods in research on stratification and class, and consider their development, especially since the nineteenth century, and their usage in different historiographical traditions.

This article will concentrate on the second, the conceptual level. However, both perspectives are intimately linked, since there is no empirical description of inequality without an acute awareness of concepts and language. However, for historians (as opposed to social theorists), all methodological discussion eventually points to empirical research in order to better understand systems of social inequality in the past, and therefore, in the present.

1. *Theories of Social Stratification and Class Formation: Historical Development Through the 1960s*

Even before the rise of the modern social sciences with their more elaborate theories of social structure, the explanation (and justification) of inequality was an important part of intellectual thought in advanced, literate societies, often within a general philosophical or theological context. In ancient Athens, Aristotle regarded society as consisting of three hierarchically ordered groups, among which he favored the middle, or '*mesor*,' as being most important for the stability of the *polis*. Aristotle's scheme of 'stratification' remains influential in attempts at defining society in terms of 'upper,' 'middle,' and 'lower' classes or strata, among which the middle class is held in highest regard (e.g., in political culture and public discourse in the USA.

In feudalist Western Europe during the Middle Ages, a different kind of tripartite scheme of inequality gained prominence: the idea of a functional differentiation of society into the three orders of Christian clergymen, noblemen, and peasants. With the rise of cities and the emergence of town burghers as an influential group, the functionalist scheme was adapted to new social realities, but it remained a central concept for the interpretation of inequality.

While these theories are themselves a subject of historical inquiry because they tell a lot about the framing and self-understanding of pre-modern societies (Duby 1978), a major shift toward modern concepts of stratification and class only occurred in the late eighteenth and nineteenth centuries. In the historical and philosophical works of the Scottish enlightenment (e.g., Adam Smith, Adam Ferguson, and John Millar), principal stages of civilizational development were associated with particular systems of social inequality and stratified class structure. The language of social inequality that is still in use, with its typical metaphors of 'class' and 'strata,' is heavily indebted to scientific discourses of the eighteenth century, and especially to the enlightenment's quest for 'classification' (Karl v. Linné) and order in the realms of both nature and culture. In the following decades, with Adam Smith and David Ricardo providing important linkages, attention in the theory of stratification increasingly shifted towards a search for the origins of inequality in economic systems of production.

The most powerful, and historically most influential, theory of class as a product of basic economic forces was for a long time Karl Marx' (and Friedrich Engels') concept of Historical Materialism, a theory according to which at each stage of society, an oppressed lower class rises against the economic and political dominance of a ruling class, and eventually overturns the old system through class struggle and revolution. Whatever critical may be said of Marx' theories and its political consequences, it nevertheless became the major reference point for most theories of inequality and stratification until the middle of the twentieth century, and possibly also thereafter. For social history in particular, it provided an extremely suggestive framework for understanding historical change in terms of economic progress, class formation, and societal conflict, instead of referring to diplomacy, politics, and the 'great men' as driving forces of history, as was the case in much traditional nineteenth and early twentieth century history. After Marx, theories of inequality based on functional differentiation and the division of labor played an important role for a while (e.g., with Herbert Spencer in the UK, Gustav Schmoller in Germany, and Émile Durkheim in France), but they too were primarily grounded in the idea of economic foundations of particular class systems.

At the beginning of the twentieth century, German sociologist Max Weber in some ways moved class analysis back to Marxian scope and intentions, while on the other hand he developed, if in a fragmentary manner not easy to grasp for later scholars, a broader, historically more flexible, and less dogmatic, framework for the interpretation of social inequality and class structures. Although Weber, like most of his contemporaries, under the impression of industrial capitalism's triumphant advance, acknowledged the enormous impact of economic forces on the molding of society, he was also prepared to systematically consider other dimensions of stratification. Political power, in Weber's view, often played a crucial role in determining positions in a class system; and also inequality for him appeared as fundamentally shaped by 'culture': by general interpretations of the world and the social order as, for example, provided by religion (*Weltbilder*), as well as by the honor, prestige,

or status attributed to an individual (or to members of a larger group, such as a specific profession) within the normative setting of a society (see Weber 1978).

After a brief period of neglect, Weber's concepts of social inequality and social class exerted a strong influence on both American and European social science after 1945. In postwar American sociology and political science in particular, stratification became one of the dominant topics in the field, a topic that all major 'schools' and theoretical currents had to reckon with (see Barber et al. 1968). Despite an enormous variety of approaches and methods, some common features of stratification research in the 1950s and early 1960s stand out, making them attractive for historians. Often, they were not just present-minded, but embedded in a macro-sociological and historical framework of thought. Typically, they were interested in linkages between class and political power structures, and in the cultural dimensions of inequality as expressed in neo-Weberian concepts like 'status' or 'prestige' (Bendix and Lipset 1953, Lenski 1966).

Since the late 1950s, Weberian and functionalist research in stratification began to overlap with a renaissance of critical theories of society and inequality that were attempting to benefit from the Marxian tradition in an undogmatic manner. Authors like Ralf Dahrendorf or Anthony Giddens were widely read for their innovative intellectual combinations of Marxism and liberalism, or Marxian and Weberian traditions (see Dahrendorf 1959, Giddens 1973). In Western social science, the revival of the Marxian tradition proved short-lived and mostly faded away in the late 1970s, leaving some strands of cultural Marxism as survivors into the then emerging new age of culturalism in the social sciences. However, Weber's influence, because of the historical depth of his work and a rare combination of flexibility and analytical precision in his categories, never quite vanished, and his work continuously remained inspiring for historians doing research on stratification and class. At the same time, Weber proved remarkably resistant against changing intellectual moods since the late 1970s and 1980s, when classical concepts of stratification and class came under attack. He thus continued to be a starting-point in some of the most important recent approaches in the sociology and social history of inequality and class.

2. *Stratification and Social Classes in Historiography: Trends and Examples*

With the rise of social history to intellectual dominance in the 1960s and 1970s in most Western countries, research on the history of stratification and class, mostly in modern societies since the Industrial Revolution of the nineteenth century, became a central historiographical theme, and even a topic defining what the 'New Social History' was all about. In the general political and academic climate of the 1960s, with its social-democratic air and its resurgence of Neo-Marxian theories, interest for the less privileged groups of industrial society—and especially for the working class—spilled over into history and not only generated a host of special studies on stratification, mobility, and class formation, but also influenced synthetic approaches in social history and the general discourse of historians in that field. While this was a pervasive international tendency, it rendered different results within the national traditions of social history. Therefore, Germany and the US will be considered as two important examples.

2.1 *Germany*

With its strong methodological roots in nineteenth-century historicism and its political conservatism in the historical profession, Germany was a late-comer in the development of postwar social history, and in many respects remained dependent on the reception of theoretical innovation from other countries such as France, the UK, and the USA. However, due to its strong traditions in social theory, and the theory of class in particular (Marx, Weber), and also because of the great historical significance of class structures in late nineteenth- and early twentieth-century Germany, topics like stratification and class formation quickly took center stage in West German social history during the late 1960s and 1970s. To a larger degree than in other countries, stratification and class emerged as the paradigmatic categories of research in social history (see, e.g., Kocka 1975), a tendency that came under attack in the 1990s as competing concepts such as gender and ethnicity by many historians were regarded as at least equally important.

For the history of stratification and class in the 1970s and 1980s, two approaches deserve special mention. First, as was the case elsewhere, the history of classes in the beginning concentrated on the less privileged groups and lower strata of society. Here, a particular focus was on the formation of a modern, industrial working class in the nineteenth century, and on the problem of the homogenization of that class from the origins of a multitude of pre-industrial, agrarian, and artisanal occupations and interests. In what he termed the 'Weberian application of a Marxian model,' historian Jürgen Kocka studied the formation of a German proletariat, which he conceived as a Marxian class in its basic socio-economic determination, and at the same time as a Weberian ideal type with which reality can never fully correspond (see Kocka 1980, 1987, see also: *Labor History*; *Working Classes*, *History of*). Class formation in this approach was considered an historical process progressing through different levels. It proceeded—

ideally, though not necessarily in historical reality—from economic homogenization in the workplace towards increasing social and cultural coherence and might end with class consciousness and the political organization of the working class. Since the 1980s, class formation models have also been applied to the upper strata of German society, particularly the bourgeoisie and the professional middle classes, but it appeared that in this case, 'culture' seemed to operate as a much more influential determinant of class than a common economic position.

Second, German traditions in social theory fostered an unusually strong interest in the historical conceptualization of the overall class structure and stratification system, again with an emphasis on the period of German modernization and industrialization from the late eighteenth century onwards. In conceiving social history as the 'history of society,' and the history of society as grounded in structures and processes of social inequality, stratification, and class, German historian Hans-Ulrich Wehler in a very influential project attempted a synthesis of modern German history in terms of a history of inequality and class. His *Gesellschaftsgeschichte* in a systematic way asks for the transformation of the estate system of preindustrial society into an industrial society comprised of social classes, with the bourgeoisie and proletariat in the center, but also giving attention to the nobility and to agrarian classes such as farmers or farmworkers (Wehler 1987–95). Typical for much German social history in the 1970s and 1980s, Wehler's approach is methodologically under the heavy influence of Max Weber's categories; and it politically adheres to a vision of an increasingly equal, less classrifted (if not classless) society.

While most other German authors are not as ambitious and comprehensive as Wehler in their concern with social structures, they have, however, been more inclined to expand the explanatory framework of social inequality into the realm of culture. A new focus of research is on the complicated relationship between social structure on the one hand, and cultural perceptions and practices on the other hand.

2.2 United States

In its self-image, the US has been born as a country of equality, defined by a lack of class and stratification in the European sense, and by the predominance of a broad middle class. The fact that the US has never known an estate system of social order has indeed, in combination with the country's more pragmatist tradition in social thought, exerted a deep influence on social scientific and historical approaches toward class structure and socio-economic inequality in America. On the other hand, critical strands in social theory and historiography, which were able to identify and discuss inequality of class, appeared earlier than in Germany and remained stronger in the first half of the twentieth century—with the 'Progressive historians' such as Charles Beard being the most important example.

During the 1960s, the blossoming of the 'New Social History,' together with a more self-critical mood in American society in general, prompted a new interest in the history of social class and inequality in America, much in the same way as in other Western countries at about the same time. With regard to approach and method, historians often concentrated on less privileged groups, such as workers and immigrants, and they often did so in case studies of particular cities and through the study of social mobility among the lower strata of urban society (Thernstrom 1964; see also *Social Mobility, History of*). For some time around 1970, Neo-Marxist approaches to the study of class and 'class struggle' in American society drew some attention, but on the whole, even leftist historians of class and inequality undertook much less theoretical effort in conceptualizing their topics than in Germany.

Similar to trends in European social history, however, quantitative methods (as in the study of social mobility) since the late 1970s lost much of their earlier appeal. Instead, the problem of class was more often approached through qualitative, even hermeneutical methods, and through the lenses of class culture and class experience. Also, outside the somewhat narrowing field of working class and labor history, the study of social classes shifted its main focus from the lower classes to the middle classes and their origins in nineteenth-century social change (e.g., Blumin 1989, Zunz 1990). This shift may be understood as an attempt to deconstruct the pervasive myth of the American middle class, and at the same time to historically explain the middle class orientation of twentieth-century America.

In a broader perspective, the study of stratification and class as such after the 1970s retreated from the center of concern for social historians much more quickly, and more pervasively, than in Europe. Obviously, the most important reason for this is the existence of other, competing dimensions of inequality in American society, especially the inequality of race and ethnicity. Class formation in the US rarely occurred in ethnically homogeneous milieus, and often racial and ethnic identity turned out to be more important, and certainly of a much more durable character, than class identity and economic position.

For the same reason, synthetic efforts with generalized views on American society from the perspective of stratification and class were hardly undertaken by American historians. The one large project that probably comes closest to such a synthesis is the two-volume textbook 'Who Built America?,' collectively conceived and written by the American Social History Project under the intellectual guidance of labor historian Herbert Gutman (See, American Social History Project 1989–92). The book attempts a synthesis of

American history from colonial times to the present by centering its narrative around the development of social classes—particularly the industrial working class—and the struggle of less privileged groups for participation in the economic and political resources of 'mainstream' America. At the same time, the authors' efforts demonstrate that a narrative of American social history from the viewpoint of stratification and class alone cannot make sense; it needs to be expanded by, and connected with, additional categories of inequality such as race, ethnicity, and gender.

3. *The Challenge of Culture in the History of Stratification and Class*

The discussion of the two national examples, especially the case of the US, has already hinted at important new trends in historical research which have begun to change the field in fundamental ways since the 1980s and will probably continue to do so in the future. Although many of the recent tendencies stem from similar motives—e.g., the increasing dissatisfaction with interpretations of class based on the structures of male industrial work—and therefore often overlap in historiographical practice, it is useful to distinguish two horizons of criticism and expansion of the classical stratification and inequality paradigms. Some scholars, for different reasons, fundamentally question the relevance of concepts such as 'class' and 'stratification' and try to develop alternative concepts, or alternative theoretical and methodological viewpoints on the problem of inequality in history (see Sect. 4). Others consider 'class' and 'stratification' as still important, or even indispensable notions for social history, yet argue that they have to be redefined in order to meet the challenges of social reality and social science at the turn of the twenty-first century. This ongoing redefinition of class is in most cases centered around the idea of 'culture.' Culture, especially in the forms of 'language' and 'experience,' is now widely believed to be the prime factor in determining stratification systems and class structures. However, at a closer look, the challenge of culture differentiates into a number of theoretical and empirical approaches.

3.1 *Class as Social Action, Experience, and Culture*

In 1963, the English Marxist historian E. P. Thompson published his two-volume 'Making of the English Working Class,' which is arguably among the dozen most influential works in twentieth-century social history (Thompson 1963). At that time, classical approaches to class formation grounded in systematic sociology or political science were still widely in use, and in some countries just beginning to influence the expanding field of social history. Thompson's massive study was rather conventional in the sense of its practical empiricism and its Marxian notion of a working-class winning identity and togetherness in the struggle against its bourgeois class enemy. What was new, however, was Thompson's insistence that 'class' was not a quasi-objective structure determined by certain economic forces, but a very fluid result of practical action and social experience on the part of the lower classes in the decades around 1800. As Thompson largely focused on pre-industrial work and living experience, he discovered the roots of class in traditions and re-inventions of popular culture. In many respects, Thompson's book was ahead of his time. Reception and debate in Germany, for example, only took off in the 1980s and turned his pioneering work into a major stimulant for the then emerging culturalist turn in social history (see *Cultural History*).

3.2 *Class Experience: From Production to Consumption*

Classical approaches to the history of stratification, from the enlightenment theories of a division of labor to Marx and beyond, have based their identification of certain classes, and their description of the overall inequality structure of society, on the sphere of production, and on the relative position an individual (mostly seen as a male adult and family head) acquired in that society's system of work and employment. As the nineteenth-century fascination with factory work and industrial production faded and the twentieth century increasingly developed the self-image of a consumer economy and society, it seemed appropriate to search for social spheres of class formation outside the workplace experience: at home, in leisure activities, and especially in consumption. In American and in British historiography in particular, the history of consumption has been one of the most significant discoveries for the field of social and cultural history in the 1990s (see *Consumption, History of*), and several attempts have been made to conceptually link consumption and class, some of which build on much earlier efforts such as Thorstein Veblen's 'Theory of the Leisure Class' with its ideas about the role of 'conspicuous consumption' for the life of the upper classes.

The notion of consumption can play a key role because of its ability to connect the perspective of economy—the appropriation of goods by groups with different economic resources—with the perspective of culture—insofar goods serve as symbols and as expressions of (individual or collective) identity and lifestyle. It also allows for a more adequate consideration of female and private-life contributions to the formation of social inequality and class. In an important empirical study on Chicago in the 1920s and 1930s,

historian Lizabeth Cohen investigated the formation of class identity among ethnic workers and their families through their shared experience of consumption and modern mass culture institutions like radio (Cohen 1990). As the history of consumption still flourishes at the turn of the twenty-first century, it is likely that its impact on the history of inequality and class will grow even larger in the near future.

3.3 Pierre Bourdieu: Class and Cultural Distinction

The first two examples for the cultural expansion of the history of class represent developments within the historical profession, yet trends and new concepts in other social scientific disciplines have also had their impact on the topic. While national traditions of thought remain influential—with Max Weber, for example, still being of prime importance in Germany, and anthropologist Clifford Geertz stimulating American debates—in international perspective one scholar's work in particular has received near unanimous attention and acclaim by social historians during the 1980s and 1990s, namely French sociologist Pierre Bourdieu's theory of class and cultural distinction as developed in his major work *La distinction* (Bourdieu 1979).

Arguing from his research on contemporary French society, Bourdieu secures a central place in social research for the problem of inequality and even sticks to the concept of class. Yet in a brilliant way he enlarges that concept by understanding class as the result of a complex ensemble of not just economic, but also social and cultural resources or types of 'capital.' Similar to the category of 'prestige' in 1950s American sociology, Bourdieu defines an individual's position in society in terms of his or her 'distinction' and 'habitus,' a concept referring to socio-cultural outlook and self-stylization. Though obviously fitting the self-oriented consumer and leisure society of the later twentieth century especially well, the appeal of Bourdieu's approach also extends to the analysis of more traditional societies in the early modern era, and even to non-Western societies, not least because of Bourdieu's own intellectual origins as a social anthropologist. As Bourdieu at the same time draws on 'classical' authors of stratification research such as Weber and Marx, his work is extremely important in providing a bridge between the classical theoretical tradition and the more recent approaches that center around language, experience, culture, and consumption.

4. Beyond Class, Beyond Stratification?

While the 'culturalist turn' in the humanities has substantially expanded and reframed perspectives on stratification and class in history since around 1980 other scholars have gone a step further in arguing that the conceptual framework of stratification and class might better be abandoned altogether and replaced by alternative concepts of social order. However, that does not mean that they would deny the existence, and empirical relevance, of stratification systems and class formation under certain historical circumstances. Rather, the argument is that the category of class itself is a historical product of the nineteenth century, and that therefore the continuing conceptual primacy of class endangers the ability of scholars to recognize different lines of social division and inequality.

4.1 Gender, Race, and Ethnicity

This point has been most persuasively made by feminist social scientists and historians who have argued that the predominance of class in social history has in fact only been able to describe the male half of the social world—and even that in an inadequate manner, because class formation among male individuals (or, household heads) cannot be conceived of independent from the family and gender systems out of which the sphere of production and patriarchy grows (see Canning 1992). In a similar way, historians of race and ethnicity have maintained that the stratification of a society in terms of individuals' economic positions never functions as an independent variable in multiracial societies such as the US, Brazil, or South Africa—nor even in European societies which are increasingly being shaped by their immigrant populations. It has hence become almost a standard method in American, and now also in British social history, to frame the problem of social inequality in the trias of 'race, class, and gender.' In the view of many feminist historians, however, this solution is a dissatisfactory compromise, because it implicitly affirms a conceptual primacy of class by molding the two other concepts after the example of class, or as mere expansions of it (see Scott 1986).

Therefore, the history of gender is in need of developing specific theoretical foundations out of twentieth-century social theory in order to establish itself as an independent approach in social studies, and not just a part of the 'litany of class, race, and gender' (Scott 1986, p. 1055). In the near future, we will perhaps see more theoretical efforts, and empirical historical studies, going in that direction, not just in gender history, but also in the history of ethnicity.

4.2 From Classes to Life-Styles, 'Milieus,' and Radical Individualism

A completely different line of thought, which also calls traditional concepts of stratification and class into question, has emerged in recent sociological work on contemporary western societies in their 'postindustrial' stage. As industrial work, and the labor and class relations so intimately connected with it, loses its central status for the definition of modern societies,

these societies may no longer be adequately described as stratified class societies. The large, socially coherent and even politically self-conscious groups such as the 'working-class' are considered to be rapidly dissolving, without new, equally coherent group structures building up. Instead, social relations in modern society may be described as based on radical individualization, and a similarity of social status at best finds expression in rather diffuse social 'milieus' (see Beck 1986, Hradil 1987). Western societies, in this perspective, are not becoming completely leveled societies with citizens of equal material resources, but the absence of material wants, and a re-definition of the value system towards post-material orientations, dramatically diminish the relevance of occupational situation, income, and even class-specific consumption patterns.

As 'social inequality' (in the form of social strata and classes) was the defining element of nineteenth and early twentieth century societies, these authors hold, categories like 'risk' or 'information' are defining societies on the verge of the twenty-first century. Although not yet widely received in social history, these new developments in theoretical and empirical sociology will probably have a considerable impact on historical views on inequality and social structure in the future. They offer important insights into the now rapidly expanding field of twentieth century social history, but it may also be asked whether societies before the industrial age can be interpreted beyond the classical thought system of social estate, class, and stratification.

4.3 Constructing Class, Constructing Inequality: From Economic Realism to Cultural Relativism

With the class structure of industrial societies appearing more as an historical exception rather than historical normalcy, and with the interest in the socio-economic foundations of inequality and class diminishing, historians have become more sensitive to the relativist or constructivist character of social structure in general and of 'class' in particular. Especially in British social history, questions have been asked about the 'language of class,' that is, about the use of concepts such as class by the historical actors themselves, who were trying to express collective self-identities by communicating new notions of social bonding. With regard to the working classes, the main thrust of this research was on political organization and political action (e.g., in English Chartism), and the objective nature of class was hardly fundamentally doubted (see Stedman Jones 1983).

Especially in the 1990s, however, and with empirical interest somewhat shifting to the middle classes, the language approach was radicalized under the double influence of cultural and linguistic theory on the one hand, and of 'radical constructivism' in epistemology and the sociology of knowledge on the other hand. Stratification and class were now analysed as 'imaginary' realms of communication and culture; they figured as notions that expressed 'visions' of the social order as perceived by the contemporaries, rather than realities of economic and power structure as described by social scientists and historians (see Corfield 1991, Joyce 1991, Nolte 2000, Wahrman 1995). Some of these approaches fundamentally undermine any 'realistic' concept of social inequality and class, and therefore sound like a triumphant signal for the final abandonment of any social scientific or historical study in stratification and class structures. Yet at the same time it can be argued that the language-oriented and constructivist approaches have introduced important expansions to the more traditional methods of social history. They have provided bridges between social and intellectual history and have induced historians to a more careful usage of language and concepts. Also, ironically, in devoting so much energy to the problem of class in historical language and self-consciousness, this work has also underscored the importance of inequality and class in past societies.

5. Conclusion

After a long period of undisputed reign as a kind of master concept in the study of present and past social structures, theories of stratification and class are now undergoing a period of substantial challenge and criticism, redefinition, and even rejection. The question may thus arise whether theoretical and empirical interest in social inequality will persist in the future, especially in the discipline of history, which is now characterized by a very pronounced shift away from problems of socio-economic structure in favor of a history of culture, symbolism, and subjectivity.

It is almost certain that inequality, stratification, and class, will lose (and already have lost) some of their original relevance as defining ground for social history as such, when compared with the 1950s to 1970s boom of those topics in the social sciences. In addition, the emerging intercultural debates on the history of non-Western societies may prove that concepts such as 'class' and 'social strata,' which are obviously deeply rooted in the Western tradition of thought and culture since classical antiquity, are not suitable instruments for analyzing Indian, or Central African, societies, be it past or present. As 'post-colonial theories' in the social and cultural sciences describe the 'hybrid' character of such societies, the hybridity of class and competing principles of social relations may also be a distinctive mark of their inequality structures.

On the other hand, three points may be made which serve as a reminder for the continuing importance, not only of the general notion of social inequality, but also of more specific concepts like stratification and class. First, social inequality—whatever it is grounded on, and in whatever way it is communicated—is an anthropological feature of all human civilization, not

only in the West. Second, inequality has often taken the form of a hierarchical order, or 'stratification,' of distinct groups according to their access to power, their rights, and material well-being. Moreover, the hierarchical mode of social differentiation is a widespread mark of the less advanced, less technologically and culturally complex societies in which historians are often especially interested. Third, it is hard not to acknowledge the success, and the astonishing flexibility, of the concept of 'class' for describing systems of inequality. As new and intellectually promising approaches have been developed since around 1980 that link 'culture' with 'class,' and as these approaches are now applied in empirical research, there can be little doubt about the fact that the history of inequality, stratification, and class will retain a prominent place in the field of social and cultural history.

See also: Class: Social; Equality and Inequality: Legal Aspects; Inequality; Inequality: Comparative Aspects; Marx, Karl (1818–89); Mobility: Social; Social Stratification; Weber, Max (1864–1920)

Bibliography

Barber B, Lipset S M, Hodge R W, Siegel P M, Stinchcombe A L, Rodman H 1968 Stratification, Social. In: Sills D (ed.) *International Encyclopedia of the Social Sciences*. Macmillan and Free Press, New York, Vol. 15

Beck U 1986 *Risikogesellschaft. Auf dem Weg in eine andere Moderne*. Suhrkamp, Frankfurt, Germany

Bendix R, Lipset S M (eds.) 1953 *Class Status, and Power: Social Stratification in Comparative Perspective*. Free Press, New York

Blumin S M 1989 *The Emergence of the Middle Class. Social Experience in the American City, 1760–1900*. Cambridge University Press, New York

Bourdieu P 1979 *La distinction. Critique sociale du jugement*. Editions de Minuit, Paris

Canning K 1992 Gender and the Politics of Class Formation: Rethinking German Labor History. *American Historical Review* **97**: 736–68

Cohen L 1990 *Making a New Deal: Industrial Workers in Chicago, 1919–1939*. Cambridge University Press, Cambridge, UK

Corfield P J (ed.) 1991 *Language, History, and Class*. Blackwell, Oxford

Dahrendorf R 1959 *Class and Class Conflict in Industrial Society*. Stanford University Press, Stanford, CA

Duby G 1978 *Les Trois Ordres ou l'Imaginaire du Féodalisme*. Gallimard, Paris

Giddens A 1973 *The Class Structure of the Advanced Societies*. Hutchinson, London

Hradil S 1987 *Sozialstrukturanalyse in einer fortgeschrittenen Gesellschaft. Von Klassen und Schichten zu Lagen und Milieus*. Leske und Budrich, Opladen, Germany

Joyce P 1991 *Visions of the People: Industrial England and the Question of Class 1848–1914*. Cambridge University Press, Cambridge, UK

Kocka J 1975 Theorien in der Sozial- und Gesellschaftsgeschichte. Vorschläge zur historischen Schichtungsanalyse. *Geschichte und Gesellschaft* **1**: 9–42

Kocka J 1980 The Study of Social Mobility and the Formation of the Working Class in the 19th Century. *Le Mouvement Social* **111**: 97–117

Kocka J 1987 *Lohnarbeit und Klassenbildung. Arbeiter und Arbeiterbewegung in Deutschland 1800–1875*. J H W Dietz, Berlin, Germany

Lenski G E 1966 *Power and Privilege: A Theory of Social Stratificaton*. McGraw Hill, New York

Levine B et al. American Social History Project 1989–92 *Who Built America? Working People and the Nation's Economy, Politics, Culture, and Society*. 2 Vols. Pantheon, New York

Nolte P 2000 *Die Ordnung der deutschen Gesellschaft: Selbstentwurf und Selbstbeschreibung im 20. Jahrhundert*. Beck, Munich, Germany

Scott J W 1986 Gender: A Useful Category of Historical Analysis. *American Historical Review* **91**: 1053–5

Stedman Jones G 1983 *Languages of Class: Studies in English Working Class History, 1832–1982*. Cambridge University Press, Cambridge, UK

Thernstrom S 1964 *Poverty and Progress: Social Mobility in a Nineteenth Century City*. Harvard University Press, Cambridge, MA

Thompson E P 1963 *The Making of the English Working Class*. Penguin, Harmondsworth, UK

Wahrman D 1995 *Imagining the Middle Class. The Political Representation of Class in Britain, 1780–1840*. Cambridge University Press, Cambridge, UK

Weber M 1978 *Economy and Society. An Outline of Interpretive Sociology*. In: Roth G (ed.) 2 Vols. University of California Press, Berkeley, CA

Wehler H-U 1987–95 *Deutsche Gesellschaftsgeschichte*, 3 Vols. C H Beck, Munich, Germany

Zunz O 1990 *Making America Corporate 1870–1920*. University of Chicago Press, Chicago

P. Nolte

Social Influence, Psychology of

Social influence refers to the ways people alter the attitudes or behavior of others. Typically social influence results from a specific action, command, or request, but people also alter their attitudes and behaviors in response to what they perceive others might do or think. Social psychologists often divide social influence phenomena into three categories—conformity, compliance, and obedience—although the distinction between the categories is often difficult to discern. *Conformity* generally refers to changes in opinion or behavior to match the perceived group norm. These changes are voluntary, although the individual may not always be aware that he or she is conforming. *Compliance* refers to changes in behavior that are elicited by a direct request. These requests

usually call upon the individual to buy a product, do a favor, or donate money or services. *Obedience* refers to responses to direct commands or demands from an authority figure. Although individuals are not physically coerced into these responses, they often feel as if they have little choice but to obey the command.

1. *Conformity*

People exhibit conformity when they change their attitudes or behavior to resemble what they believe most people like them would think or do. Everyday examples of conformity include selecting clothes to match what other people are wearing and recycling newspapers because it appears that most people support recycling. Psychologists have identified two primary motives to explain conformity behavior—*informational influence* and *normative influence* (Deutsch and Gerard 1955). When informational influence is operating, individuals conform because they want to be accurate. People agree with the perceived norm because they lack confidence in their own judgment and assume that the typical judgment is correct. When conformity reflects normative influence, the individual's primary concern is to gain social approval and to avoid the consequences of appearing deviant. People go along with the perceived norm because they hope their conformity will lead to acceptance from the group or because they hope to avoid the criticism, ridicule, and possible rejection that may follow a failure to conform. In practice, informational influence and normative influence often operate hand in hand. For example, a conforming individual may hope to gain acceptance from a group and may also believe that the majority of group members are correct in their judgment.

The first demonstration of information influence in a conformity study was reported by Muzafir Sherif (1936). This classic investigation placed male participants in a totally darkened room. Periodically a small dot of light appeared on a wall approximately 15 feet away, and the participants' task was to report how much, if any, they saw the light move. In fact, the light did not move. Sherif took advantage of a perceptual illusion known as the *autokinetic effect*—the tendency to see a stationary point of light move in a dark environment. When three participants sat together and reported their perceived movement aloud, Sherif found that over time the estimates each participant gave tended to resemble those of the others in the room. In other words, each group established its own norm for the amount of movement they perceived, and the group members tended to see an amount of movement close to that norm. These participants typically were unaware that they had been influenced by the group, and later investigations found participants continued to rely on the group norm even after the other individuals were no longer in the room.

A second set of early conformity studies demonstrated the power of normative influence. Solomon Asch (1951) asked participants to engage in a simple line judgment task. In the standard procedure, several individuals at a time were presented with a picture of a 'standard' line accompanied by three comparison lines. The task was to identify which of the comparison lines was the same length as the standard line. The individuals gave their responses aloud for the experimenter to record. Because the length of the lines differed by noticeable amounts, the task was an easy one. However, only one of the individuals in the study was a real participant. The others were confederates who, beginning with the third line judgment task, began to give wrong answers (before the real participant's turn to respond). On 12 of the remaining 16 trails, the real participant heard each of the others select a comparison line that clearly was incorrect. The question was whether the participants would give what they knew to be the correct answer, or if they would go along with the response given by the group.Typically, participants went along with what their senses told them. However, on 37 percent of the trials the participants conformed to the norm and gave the same response as the rest of the group. Moreover, approximately three-quarters of the participants went along with the majority at least once.

Most of the conforming participants in Asch's studies were motivated by normative influence. That is, these individuals were concerned about how they would look in the eyes of the other participants if they were to express their true judgments. When participants in one study were allowed to give their responses in private, the rate of conformity dropped dramatically (Asch 1956). However, at least some informational influence also appeared to be operating in the Asch studies. That is, some people came to believe that the shared perception of the group must be more accurate than their own visual perception.

1.1 *Variables Affecting Conformity*

Although numerous replications of Asch's findings provide persuasive evidence for the power of conformity, not all people conform to the same degree in the same situation. For example, cultural variables have been found to play a role in the conformity process. Bond and Smith (1996) reviewed 133 studies using the Asch procedure conducted in 17 countries. They found participants from collectivist cultures (see *Cultural Psychology*) were more likely to conform to the group norm than people from individualistic cultures. These investigators also found a trend toward less conformity since the 1950s in studies with US participants. Thus, changes in societal standards, such as the value placed on independence and autonomy, may affect the extent to which members of a given culture conform. Age also appears to play a role in this

process. Consistent with conventional wisdom, researchers find that young adolescents are particularly prone to conforming with the behavior of peers. Finally, several investigations find that personality plays a role in conformity. People who have a high need to feel in control are less likely to conform, whereas those who are highly concerned with the impression they make on others are more likely to go along with the perceived norm.

Psychologists have not always agreed on the extent to which gender influences conformity behavior (Cooper 1979, Eagly and Chravala 1986). Although some disagreement remains about the size of the effect, it does appear that women are more likely to conform than men. Moreover, this tendency to conform seems to be especially likely in face-to-face situations like the Asch procedure. Explanations for this difference include men's need to appear independent and women's tendency to promote smooth social interactions.

Finally, the number of other people in the situation affects conformity behavior. Asch (1956) varied the number of confederates in his procedure, creating conditions with one to fifteen confederates. He found that conformity increased noticeably as the number of confederates grew, but that this effect leveled off after three or four. In other words, five or fifteen confederates giving an incorrect response to the line judgment task generated no more tendency to conform to the norm than four confederates. Other researchers have developed elaborate models to account for the effects of crowd size on conformity (Latane 1981, Tanford and Penrod 1984). These models consider such variables as characteristics of the individuals in the group and physical proximity of the group members.

1.2 Perceived Norms

The impact of norms is not limited to situations like the Asch paradigm, in which people observe the behavior of others and face immediate social consequences. Rather, many daily decisions are influenced by what individuals believe to be the normative behavior for a given situation. Whether people give money to a charitable cause, yield the right of way to pedestrians, or take an extra copy of the paper from a newspaper box depends to some degree on what they perceive to be the normative behavior in that situation.

Researchers have identified two kinds of perceived norms (Cialdini et al. 1991). *Injunctive norms* refer to correct behavior, that is, what society says one should do in the situation. Injunctive norms tell people they should give to charitable causes and that they should not take a newspaper they have not paid for. *Descriptive norms* refer to what most people actually would do in the situation. Although people typically do not have adequate data to make an accurate assessment of what most people would do, they commonly rely on whatever information they have to estimate the descriptive norm, however incorrect they may be.

Injunctive and descriptive norms often are very similar and lead us to the same decision about our behavior. However, psychologists also point to some notable exceptions. For example, a widely accepted injunctive norm states that people should not litter in public settings. But a parking garage filled with bits of litter suggests to passersby that the descriptive norm—what people actually do—is quite different. Whether people respond to the injunctive or descriptive norm in this situation may depend on which of these two is brought to their awareness. Researchers in one set of studies placed useless flyers on windshields in a parking garage (Reno et al. 1993). When the drivers' attention was drawn to the litter that others had apparently tossed to the ground (descriptive norm), they tended to also drop the flyer on the ground. When these drivers were made aware of the injunctive norm, such as when they saw another person pick up a piece of litter to keep the environment clean, most decided against littering.

2. Compliance

People demonstrate compliance when they agree to an explicit request, such as a request to buy a product or to volunteer their time. Much of the research on compliance has examined the effectiveness of sequential-request procedures. By presenting participants with certain kinds of request in specific sequences, investigators have identified many of the psychological processes that lead people to accept or refuse a request. Four of these procedures are described here.

The most widely researched of the sequential-requests procedures is the *foot-in-the-door* technique (Freedman and Fraser 1966). The procedure begins with a small request that virtually everyone agrees to. At some later point, the same requester or a different individual returns and presents a larger request. This second request is known as the target request, because getting participants to agree with the larger request is the real purpose of the procedure. If the foot-in-the-door manipulation is successful, participants will comply with the target request at a higher rate than if they had been presented only with the target request.

Numerous investigations demonstrate that the foot-in-the-door procedure can increase compliance to the request. However, reviews of this literature find that many studies fail to produce the effect and that the overall size of the effect is not large (Burger 1999). Part of the problem with this research is that investigators have used a variety of procedures to create a foot-in-

the-door manipulation, some of which are not effective. For example, the foot-in-the-door procedure is more likely to increase compliance when participants are allowed to perform the initial request, when the initial request requires more than a minimal amount of effort, and when the second request appears to be a continuation of the initial request (Burger 1999).

Researchers disagree on the psychological processes underlying the foot-in-the-door phenomenon. The most widely cited interpretation for the effect is the self-perception explanation. According to this account, agreeing with the initial request causes people to alter the way they think about themselves. When later presented with the target request, these individuals are said to use their compliance to the small request as an indicator of their attitude toward such causes or toward participating in these kinds of activities. Because they agreed to the small request, they are more likely than control participants to agree to the target request.

Another common sequential-request procedure begins with an initial request so large that most people refuse. These individuals then are presented immediately with a smaller (target) request. If this *door-in-the-face* technique is successful, participants will agree to the target request more often than if they had not first rejected the large request (Cialdini et al. 1975). For example, participants in one study were asked to volunteer for a two-year placement working as a youth counselor. After they refused, experimenters asked participants if they would chaperone some children on a two-hour trip to the zoo. These participants agreed at a higher rate than those asked only the target request (Cialdini et al. 1975).

The door-in-the-face procedure is said to be effective because the requester takes advantage of a widely practiced social rule called the *reciprocity norm*. The reciprocity norm governs many social exchanges and requires that people return favors and acts of kindness (Gouldner 1960). Most people feel a strong sense of obligation to someone who has done them a favor, even when that favor was not requested. Moreover, studies find that people often give back more than they received in order to relieve themselves of their obligation. When applied to the door-in-the-face situation, the requester 'gives' the individual what appears to be a concession. That is, the requester seems to give up something he or she wants, and the individual reciprocates by giving something to the requester, i.e., the target request.

Another sequential-request procedure begins when the participant agrees to a request at a given price. Requesters using the *low-ball* procedure then raise the cost of the request slightly. The procedure leads to more compliance at the higher price than when the price is presented at the outset (Cialdini et al. 1978). For example, undergraduates in one study agreed to participate in a psychology experiment. The students then were informed that the experiment was to be held at 7.00 a.m., thus making participation more costly than they probably had imagined. Nonetheless, these students showed up for the early morning experiment more often than those told about the time of the experiment at the beginning (Cialdini et al. 1978). The effectiveness of the low-ball procedure is said to result from a sense of commitment that develops when the individual agrees to the initial request. The individuals feel committed to perform the agreed-upon behavior, or at least to do something for the requester (Burger and Petty 1981). This sense of commitment causes people to continue to agree to the request even at the higher price.

In contrast to the small-to-large price progression used in the low-ball technique, the *that's-not-all* procedure uses a large-to-small price progression. Requesters using this procedure present the request at a given price, then improve the deal before the individual can respond (Burger 1986). For example, participants in one study were told the price of a cupcake was one dollar. Before participants could respond, the salesperson said the price really was 75 cents. The participants in this condition were more likely to buy the cupcake than those told at the outset that the price was 75 cents (Burger 1986). The that's-not-all procedure is said to work in part because, as in the door-in-the-face procedure, the requester appears to be making a concession. The procedure also is effective because it appears to alter the 'anchor point' the individual uses to decide if the final price is a good or fair one. That is, participants are said to use the initial price as a reference point when deciding whether to agree with the final price. Because the final price is lower, participants are more likely to see it as reasonable than when they are presented only with the lower price.

3. *Obedience*

Obedience refers to an individual's response to a command from an authority figure. Implicit in this description is the notion that the recipient of the command is reluctant to engage in the behavior and probably would not unless give direct orders to do so. Consequently, psychologists' ability to conduct controlled laboratory studies of obedience has been limited by concerns for the welfare of participants who are torn between obeying commands and following their conscience. As a result of heightened concerns for the ethical treatment of human participants in recent years, most of the insights psychologist have about the obedience process come from a series of demonstrations conducted in the 1960s by Stanley Milgram (1974).

The basic procedure in Milgram's studies used one real participant, one confederate posing as a participant, and an experimenter. The study was described to participants as a learning experiment designed to

test the effects of punishment. The participant's task was to teach a list of word associations to the confederate by punishing the confederate for incorrect answers. The method of punishment was electric shock, with the participant instructed to give increasingly severe shocks to the confederate for each wrong answer. The participant's dilemma occurred when the confederate complained about the shocks and demanded to be released from the shock apparatus. In most versions of the procedure the participant could hear the confederate screaming in pain through the wall that separated them. Although it was apparent that the confederate was suffering and was probably receiving dangerous levels of electric shock, the experimenter commanded the participant to continue giving shocks. The procedure continued until the participant refused to go on, or until the participant pressed the highest button on the apparatus (labeled 450 volt) three times, for a total of 33 shocks.

The results of the study were unexpected. Prior to conducting the investigation, Milgram asked students, middle-class adults, and psychiatrists to predict the results. Everyone agreed that the typical participant would stop very early in the shock sequence and that virtually no one would continue to the 450 volt level. However, in the basic procedure Milgram found that 65 percent of the participants continued to follow the experimenter's commands all the way to the end.

Milgram (1974) conducted a series of variations on the basic procedure in an effort to explain the participants' high level of obedience. He found that proximity of the participant and confederate appeared to play a role. Obedience dropped to 40 percent when the participant was in the same room as the confederate, and dropped to 30 percent when the participant was required to touch the confederate to administer the shock. Although the original studies were conducted with male participants, a similar pattern was found when Milgram used women as participants. The level of obedience did not diminish when the experimenter gave his commands in a meek manner, but dropped to 20 percent when the commands were given by another confederate posing as a participant. Milgram concluded from this last finding that obedience does not depend on the authority figure's personality or charismatic style, but rather that obedience requires only that the individual be recognized as a legitimate authority.

Most social psychologists point to Milgram's research as an example of the often-unrecognized power of the situation to influence behavior. Although personality characteristics of the participants can affect whether they obey or refuse the command (Blass 1991), the discrepancy between people's predictions and the actual outcome of the Milgram studies suggests that the situation had a greater influence than most people recognize. Other psychologists point out that Milgram's procedure also took advantage of conformity and compliance processes. For example, by moving from small to progressively larger shocks, Milgram placed participants in a situation similar to that used in foot-in-the-door demonstrations. Milgram also found a drop in obedience when participants had information about descriptive norms. When participants saw two other participants (actually confederates) refuse the command, only 10 percent obeyed the experimenter's commands to the end.

4. Future Directions

Although researchers have identified many social influence phenomena, questions remain about the psychological processes underlying some of these behaviors. Future research is likely to examine these processes, with particular attention to the way multiple processes combine to produce conformity and compliance. Investigators also may focus on variables affecting social influence. In particular, cultural differences in conformity and compliance behavior are likely to be explored. However, future research on obedience probably will be limited by concerns for the ethical treatment of human participants. For many years now, these concerns have prevented researchers from replicating the procedures used by Milgram in his obedience studies. The challenge for investigators is to develop procedures that create a realistic obedience situation without exposing participants to excessive emotional stress.

See also: Compliance and Obedience: Legal; Conformity: Sociological Aspects; Experimental Design: Compliance; Impression Management, Psychology of; Obedience: Social Psychological Perspectives

Bibliography

Asch S E 1951 Effects of group pressure upon the modification and distortion of judgments. In: Guetzkow H (ed.) *Groups, Leadership, and Men*. Carnegie Press, Pittsburgh, PA

Asch S E 1956 Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs* **70** (9): 416.

Blass T 1991 Understanding behavior in the Milgram obedience experiment: The role of personality, situations, and their interactions. *Journal of Personality and Social Psychology* **60**: 398–413

Bond R, Smith P B 1996 Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin* **119**: 111–37

Burger J M 1986 Increasing compliance by improving the deal: The that's-not-all technique. *Journal of Personality and Social Psychology* **51**: 277–83

Burger J M 1999 The foot-in-the-door compliance procedure: A multiple-process analysis and review. *Personality and Social Psychology Review* **3**: 303–25

Burger J M, Petty R E 1981 The low-ball compliance technique: Task or person commitment? *Journal of Personality and Social Psychology* **40**: 492–500

Cialdini R B, Cacioppo J T, Bassett R, Miller J A 1978 The low-ball procedure for producing compliance: Commitment then cost. *Journal of Personality and Social Psychology* **36**: 463–76
Cialdini R B, Kallgren C A, Reno R R 1991 A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology* **24**: 201–34
Cialdini R B, Vincent J E, Lewis S K, Catalan J, Wheeler D, Darby B L 1975 Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique. *Journal of Personality and Social Psychology* **31**: 206–15
Cooper H M 1979 Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology* **37**: 131–46
Deutsch M, Gerard H B 1955 A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology* **51**: 629–36
Eagly A H, Chravala C 1986 Sex differences in conformity: Status and gender-role interpretation. *Psychology of Women Quarterly* **10**: 203–20
Freedman J L, Fraser S C 1966 Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology* **4**: 195–202
Gouldner A W 1960 The norm of reciprocity: A preliminary statement. *American Sociological Review* **25**: 161–78
Latane B 1981 The psychology of social impact. *American Psychologist* **36**: 343–56
Milgram S 1974 *Obedience to Authority*. Harper, New York
Reno R R, Cialdini R B, Kallgren C A 1993 The transsituational influence of social norms. *Journal of Personality and Social Psychology* **64**: 104–12
Sherif M 1936 *The Psychology of Social Norms*. Harper, New York
Tanford S, Penrod S 1984 Social influence model: A formal integration of research on majority and minority influence processes. *Psychological Bulletin* **95**: 189–225

J. M. Burger

Social Insurance: Legal Aspects

1. Social Insurance and Private Insurance

The notion of insurance is widespread and clear: insurance is a way of providing for the consequences of negative future events (risks or contingencies) in a collective manner with a special technique of financing. The distinction between private and social insurance is less clear: social insurance is organized by public authorities through legislation, and is usually compulsory, while private insurance is usually voluntary. Social insurance seeks to provide coverage against the consequences of risks like sickness, industrial accident, invalidity, old age, and unemployment for a large part of the population of a nation. To attain these goals, a compulsory form of insurance is considered a necessity. Bismarck's social insurance in Germany is in some ways the prototype of social insurance. However, his social insurance covered only workers rather than the general population (Koehler and Zacher 1982).

The difference between social and private insurance can be found in the goals, the techniques of protection, the legal form of the insurance bodies as private or public organizations, and the way of financing the benefits. But there is no clear distinction between them. Sometimes, the goals of social insurance are attained by compulsory private insurance and, in some countries, the bodies responsible for social insurance are private bodies.

One of the main distinctions between social and private insurance concerns the ways in which risk is calculated and financed. In the field of private insurance, the basis of calculation of the premiums is the risk. In the field of social insurance, the risk is less important, and social aims (such as health and welfare) prevail against the economic criterions.

Private insurance, moreover, distributes the burden of an individual risk across a group. Social insurance has the same effects, but it also effects income redistribution between those who are healthy and sick, richer and poorer, younger and older people, or households without and with family members.

2. Two Characteristic Items of Social Insurance: Financing and Organization

Thus, two characteristic items of social insurance remain, which allow us to distinguish between social insurance and other techniques of social protection. Those interlinked items are the ways of financing social insurance and the organization of social insurance.

2.1 Financing

Social insurance benefits were originally financed by contributions levied on earnings and divided between employee and employer, thus expressing the shared responsibility between employer and employee with regard to work-related social risks. State subsidies completed the financing. However, this scheme changed over time. Some countries introduced increasing state subsidies, while others changed the proportion of the contributions burden payable by employer and employee. In Germany, for example, the employees had to relinquish a public holiday to compensate employers for the new financial burden created by the introduction of long-term care insurance. Currently, there are ongoing discussions about changing the financial basis from the earnings of employees to the surplus value of enterprises.

2.2 Organization

Originally, social insurance bodies were special bodies that were not directly integrated in the administrative organization of a state. Sometimes, they are co-

administered by representatives of employers and employees. This kind of self-government is more prevalent in countries where social insurance bodies are private organizations. In countries with special public authorities for social insurance, but without self-governing or autonomous elements of administration, the ideal image of social insurance is left behind.

3. Social Insurance and Social Security

The concept of social insurance should not be confused with the concept of social security (cf. UN Declaration of Human Rights, art. 22), which is an international goal of social policy. Social insurance is considered a special, nevertheless widespread technique of social protection. It combines the scope and character of a compulsory insurance on the one hand, and the financing of the benefits by obligatory contributions on the other.

From the perspective of international law, social insurance does not entail special legal problems. Nevertheless, international law scholars have taken a special interest in social insurance as a mechanism of social protection. Thus, many international treaties and other international or regional instruments, especially involving Europe, address social insurance.

In the field of social security, international law has two main dimensions: (1) coordinating different national social security schemes to guarantee freedom of movement, and (2) defining social standards and aims (e.g., minimum standards, standards to attain, principles).

The coordination task was originally primarily due to the migration of workers, while the standard defining task began with the treaty of Versailles, which can be considered the basis of the development of an international social policy (Zacher 1976).

4. Fulfillment of International Requirements (International Standards of Social Insurance)

Two kinds of requirements prevail internationally: attainment of minimum standards and achievement of higher levels than a given minimum.

4.1 Minimum Standards

The best known and most basic international instrument to set minimum standards of social security is Convention 102 of the International Labor Organization (ILO) (1952). This convention, ratified by 40 states, contains minimum standards regarding medical care and benefits in the case of sickness, unemployment, old age, employment injury, maternity, invalidity, and for survivors. This convention applies to insurance schemes as well as other schemes of protection.

4.2 Higher Levels of Social Standards

The European Social Charter (1961) of the Council of Europe, a regional instrument for the European countries, serves as a social standard instrument for the European Union nations. It is now ratified by 24 states. In the frame of the Council of Europe, other social security conventions (e.g., Code of Social Security) also aim at higher social standards.

5. Social Insurance Benefits

5.1 Range of Benefits

The range of risks covered by social insurance schemes, labeled as 'traditional' social risks or contingencies, is described in ILO Convention 102 on minimal standards of social security (see Sect. 4.1) (von Maydell and Hohnerlein 1994). As new social risks develop, it is not clear whether protection will be provided by social insurance schemes or by other means (van Langendonck 1997, Pieters 1998).

Social insurance bodies do not provide benefits unique to social insurance. Other bodies of social protection and even private insurance bodies provide comparable or identical benefits. However, there are typical links between social insurance and the benefit provided, especially for those cash benefits designed to replace former earnings, e.g., sickness benefits or old-age pensions. As the insured person pays contributions based on his/her earnings (usually a certain percentage of the income), the benefit replacing the earnings is calculated as a function of the former earnings and is not provided as a flat-rate sum.

Whereas private insurance is almost always paid in cash, social insurance benefits can be either benefits in kind or in cash.

5.2 Guarantee of Benefits

In some countries, entitlement and even qualifying periods for social insurance entitlement are protected by the constitution. In the German constitutional doctrine, contributions to the old-age pension scheme are protected by the constitution as property. The new Ukrainian Constitution gives another kind of guarantee which may apply to social insurance benefits: it requires that the content and scope of existing rights and freedoms shall not be diminished in the adoption of new laws or in the amendment of laws that are in force.

5.3 Public/Private Mix of Benefits

Initially, social insurance benefits were the only way of providing a subsistence base in the case of risk occurrence. Nowadays, social insurance benefits are increasingly complemented by other means of protection. Mainly in the field of old-age pensions, the interplay of public, occupational, and private protection increasingly serves as a current mode of protection (International Bank for Reconstruction and Development 1994).

6. Future Developments in the Field of Social Insurance

To reach the goals of social security the technique of social insurance was one of the most important tools in the evolution of the welfare states. Social insurance, in the sense of Bismarck's social insurance, has undergone major changes. The initial role of social insurance, the social protection of workers, has lost its importance as the personal scope of social insurance has widened. Other forms of social protection, especially state-financed social programs for families, unemployed people, or for persons in need are increasingly prevalent.

Today, there is no unique mode of using social insurance to address social problems. Social insurance schemes rely mainly on financing by contributions and are thus, as regards their financing basis, dependent on the evolution of earnings. This reliance is considered a critical factor in the future financial stability of social insurance schemes and is broadly under discussion (Pieters 1999b).

See also: Health Insurance: Economic and Risk Aspects; Insurance; Insurance and the Law; Law and Risk; Legal Insurance; National Health Care and Insurance Systems; Social Security

Bibliography

Comparative Tables of Social Security Schemes in Council of Europe Member States not Members of the European Union, in Australia, in Canada, in Armenia and in New Zealand 1999 (9th ed., situation at 1 July 1998) Council of Europe Publishers, Strasbourg

International Bank for Reconstruction and Development 1994 *Averting the Old Age Crises: Policies to Protect the Old and Promote Growth*. Oxford University Press, New York

International Social Security Association 1993 *Social Security Systems in Central and Eastern Europe*. International Social Security Association, Geneva, Switzerland

International Social Security Association 1998 *Social Security Issues in English-speaking African Countries*. International Social Security Association, Geneva, Switzerland

International Social Security Association 1998 *Social Security in Asia and the Pacific: the Road Ahead*. International Social Security Association. Manila, Philippines

Koehler P A, Zacher H F (eds.) 1982 *The Evolution of Social Insurance 1881–1981. Studies of Germany, France, Great Britain, Austria and Switzerland*. Pinter, London

Langendonck J van (ed.) 1997 *The New Social Risks*. Kluwer Law International, London

Maydell B von, Hohnerlein E M (eds.) 1994 *The Transformation of Social Security Systems in Central and Eastern Europe*. Peeters Press, Leuven, Belgium

Myers R J 1985 *Social Security*. 3rd edn., published for McCahan Foundation, Bryn Mawr, PA by R D Irwin, Homewood, IL

Pennings F 1998 *Introduction to European Social Security Law*. 2nd edn., Kluwer Law International, The Hague, Netherlands

Pieters D 1993 *Introduction into the Basic Principles of Social Security*. Kluwer, Deventer, Netherlands

Pieters D (ed.) 1998 *Social Protection of the Next Generation in Europe*. Kluwer Law International, London

Pieters D (ed.) 1999a *International Impact upon Social Security*. Kluwer Law International, London

Pieters D (ed.) 1999b *Changing Work Patterns and Social Security*. Kluwer Law International, London

Swedish National Social Insurance Board – European Commission 1997 *25 Years of Regulation (EEC) No. 1408/71 on Social Security for Migrant Workers*. Stockholm

Zacher H F 1976 *Internationales und Europäisches Sozialrecht*. Verlag R. S. Schulz, Percha/Kempenhausen am Starnberger See, Germany

G. Igl

Social Integration, Social Networks, and Health

Since the mid-1970s there have been dozens of articles, and now books, on issues related to social networks and social support. It is now recognized widely that social relationships and affiliation have powerful effects on physical and mental health for a number of reasons (House et al. 1988, Berkman and Glass 2000, Cohen et al. 2000).

When investigators write about the impact of social relationships on health, many terms are used loosely and interchangeably, including social networks, social support, social ties, and social integration. The aim of this article is to discuss: (a) theoretical orientations from diverse disciplines which are fundamental to advancing research in this area; (b) findings related to mortality; (c) a set of definitions of networks and aspects of networks and support; and (d) an overarching model which integrates multilevel phenomena.

1. Theoretical Orientations

There are several sets of theories that form the bedrock for the empirical investigation of social relationships and their influence on health. The earliest theories

came from sociologists such as Emile Durkheim, as well as from psychoanalysts such as John Bowlby, who first formulated attachment theory. A major wave of conceptual development also came from anthropologists including Bott and Mitchell, as well as quantitative sociologists such as Burt, Laumann, and Wellman who, along with others, have developed social network analysis. This eclectic mix of theoretical approaches coupled with the contributions of epidemiologists form the foundation of research on social ties and health.

Durkheim's contribution to the study of the relationship between society and health is immeasurable. Perhaps most important is the contribution he has made to the understanding of how social integration and cohesion influence suicide (Durkheim 1897). Durkheim's primary aim was to explain how individual pathology was a function of social dynamics. In light of recent attention to 'upstream' determinants of health, Durkheim's work re-emerges with great relevance today.

Bowlby (1969), one of the most important psychiatrists of the twentieth century, proposed theories suggesting that the environment, especially in early childhood, played a critical role in the genesis of neurosis. Bowlby proposed that there is a universal human need to form close affectional bonds. Attachment theory, proposed by Bowlby, contends that the attached figure creates a secure base from which an infant or toddler can explore and venture forth. The strength of Bowlby's theory lies in its articulation of an individual's need for secure attachment for its own sake, for the love and reliability it provides, and for its own 'safe haven.' Primary attachment promotes a sense of security and self-esteem that ultimately provides the basis on which the individual will form lasting, secure and loving relationships in adult life.

1.1 Social Network Theory: A New Way of Looking at Social Structure and Community

During the mid-1950s, a number of British anthropologists found it increasingly difficult to understand the behavior of either individuals or groups on the basis of traditional categories such as kin groups, tribes, or villages. Bott (1957) and Mitchell (1969) developed the concept of 'social networks' to analyze ties that cut across traditional kinship, residential, and class groups to explain behaviors they observed such as access to jobs, political activity, or marital roles. The development of social network models provided a way to view the structural properties of relationships among people.

Network analysis 'focuses on the characteristic patterns of ties between actors in a social system rather than on characteristics of the individual actors themselves and use these descriptions to study how these social structures constrain network members' behavior' (Hall and Wellman 1985, p. 26). Network analysis focuses on the structure and composition of the network, and the contents or specific resources that flow through those networks. The strength of social network theory rests on the testable assumption that the social structure of the network itself is largely responsible for determining individual behavior and attitudes, by shaping the flow of resources which determine access to opportunities and constraints on behavior.

2. Health, Social Networks, and Integration

From the mid-1970s through the present, there has been a series of studies showing consistently that the lack of social ties or social networks predicted mortality from almost every cause of death. These studies have been done in the USA, Europe, and Asia and most often captured numbers of close friends and relatives, marital status, and affiliation or membership in religious and voluntary associations. These measures were conceptualized in any number of ways as assessments of social networks or ties, social connectedness, integration, activity, or embeddedness. Whatever they were named, they defined embeddedness or integration uniformly as involvement with ties spanning the range from intimate to extended.

In the first of these studies, from Alameda County (Berkman and Syme 1979), men and women who lacked ties to others (in this case, based on an index assessing contacts with friends and relatives, marital status, and church and group membership) were 1.9 to 3.1 times more likely to die in a 9-year follow-up period than those who had many more contacts.

Another study in Tecumseh, Michigan (House et al. 1982) shows a similar strength of positive association for men, but not for women, between social connectedness/social participation and mortality risk over a 10–12 year period. An additional strength of this study was the ability to control for some biomedical predictors assessed from physical examination (e.g., cholesterol, blood pressure, and respiratory function).

Similar results from several more studies have been reported, one from a study in the United States and three from Scandinavia. Using data from Evans County, Georgia, Schoenbach et al. (1986) found risks to be significant in older white men and women even when controlling for biomedical and sociodemographic risk factors, although some racial and gender differences were observed. In Sweden, two studies (Welin et al. 1985, Orth-Gomer and Johnson 1987) report significantly increased risks among socially isolated adults. Finally, in a study of 13,301 men and women in Eastern Finland, Kaplan et al. (1988) have shown that an index of social connections predicts

mortality risk for men but not for women, independent of standard cardiovascular risk factors.

Studies of older men and women confirm the continued importance of these relationships into late life (Seeman et al. 1993a). Furthermore, two studies of large cohorts of men and women in a large Health Maintenance Organization (HMO) (Vogt et al. 1992), and 32,000 male health professionals (Kawachi et al. 1996) suggest that social networks are, in general, more strongly related to mortality than to the incidence or onset of disease.

Two studies, in Danish men (Pennix et al. 1997), and Japanese men and women (Sugiswa et al. 1994), further indicate that aspects of social isolation or social support are related to mortality. Virtually all of these studies find that people who are socially isolated or disconnected from others have between two and five times the risk of dying from all causes compared to those who maintain strong ties to friends, family, and community.

Social networks and support have been found to predict a very broad array of other health outcomes from survival post-myocardial infarction to disease progression, functioning and the onset and course of infectious diseases. The reader is referred to several lengthy reviews mentioned earlier for additional information.

3. *A Conceptual Model Linking Social Networks to Health*

Although the power of measures of networks or social integration to predict health outcomes is indisputable, the interpretation of what the measures actually assess has been open to much debate. Hall and Wellman (1985) have commented appropriately that much of the work in social epidemiology has used the term social networks metaphorically, since rarely have investigators conformed to more standard assessments used in network analysis, according to calls to develop a second generation of network measures (Berkman 1986, House et al. 1988).

A second wave of research developed in reaction to this early work, and as an outgrowth of work in health psychology, that turned the orientation of the field in several ways (House and Kahn 1985, Sarason et al. 1990). These social scientists focused on the qualitative aspects of social relations (i.e., their provision of social support or, conversely, detrimental aspects of relationships) rather than on the elaboration of the structural aspects of social networks.

Most of these investigators follow an assumption that what is most important about networks is the support function they provide. While social support is among the primary pathways by which social networks may influence physical and mental health status, it is not the only critical pathway (Berkman and Glass 2000). Moreover, the exclusive study of more proximal pathways detracts from the need to focus on the social context and structural underpinnings that may have an important influence on the types and extent of social support provided.

In order to have a comprehensive framework in which to explain these phenomena, it is helpful to move 'upstream' and return to a more Durkheimian orientation to network structure and social context. It is critical to maintain a view of social networks as lodged within those larger social and cultural contexts which shape the structure of networks. In fact, some of the most interesting recent work in the field relates social affiliation to social status, and social and economic inequality (Kawachi et al. 1997).

Conceptually, social networks are embedded in a macrosocial environment in which large-scale social forces may influence network structure, which in turn influences a cascading causal process beginning with the macrosocial to psychobiological processes to impact health. Serious consideration of the larger macrosocial context in which networks form and are sustained has been lacking in all but a small number of studies, and is almost completely absent in studies of social network influences on health.

Networks may operate at the behavioral level through at least four primary pathways: (a) provision of social support, (b) social influence, (c) social engagement and attachment, and (d) access to resources and material goods. These psychosocial and behavioral processes may influence even more proximate pathways to health status, including: (a) direct physiological responses, (b) psychological states including self-esteem, self-efficacy, depression, (c) health-damaging behaviors such as tobacco consumption or high-risk sexual activity, and health promoting behavior such as appropriate health service utilization or exercise, and (d) exposure to infectious disease agents such as HIV, other sexually transmitted diseases (STDs), or tuberculosis. The four primary psychosocial or behavioral pathways by which networks may influence health will be reviewed briefly. The reader is referred to Berkman and Glass (2000) for a lengthier review.

3.1 *The Assessment of Social Networks*

Social networks might be defined as the web of social relationships that surround an individual and the characteristics of those ties. Burt (1982) has defined network models as describing 'the structure of one or more networks of relations within a system of actors.' Network characteristics cover:

(a) *range* or *size* (number of network members),

(b) *density* (the extent to which the members are connected to each other),

(c) *boundedness* (the degree to which they are defined on the basis of traditional structures such as kin, work, neighborhood), and

(d) *homogeneity* (the extent to which individuals are similar to each other in a network).

Related to network structure, characteristics of individual ties include:

(a) *frequency of contact*,

(b) *multiplexity* (the number of types of transactions or support flowing through ties),

(c) *duration* (the length of time an individual knows another), and

(d) *reciprocity* (the extent to which exchanges are reciprocal).

3.2 Downstream Social and Behavioral Pathways

3.2.1 Social Support. Moving downstream, discussion follows of the mediating pathways by which networks might influence health status. Most obviously the structure of network ties influences health via the provision of many kinds of support. This framework acknowledges immediately that not all ties are supportive, and that there is variation in the type, frequency, intensity, and extent of support provided. For example, some ties provide several types of support while other ties are specialized and provide only one type. Social support typically is divided into subtypes which include emotional, instrumental, appraisal, and informational support (House 1981). Emotional support is related to the amount of 'love and caring, sympathy and understanding and/or esteem or value available from others.' Emotional support is most often provided by a confidant or intimate other, although less intimate ties can provide such support under circumscribed conditions.

Instrumental support refers to help, aid, or assistance with tangible needs such as getting groceries, getting to appointments, phoning, cooking, cleaning, or paying bills. House identifies instrumental support as aid in kind, money, or labor. Appraisal support often defined as the third type of support, relates to help in decision-making, giving appropriate feedback, or help deciding which course of action to take. Informational support is related to the provision of advice or information in the service of particular needs. Emotional, appraisal, and informational support are often difficult to disaggregate and have various other definitions (e.g., self-esteem support).

Perhaps even deeper than support are the ways in which social relationships provide a basis for intimacy and attachment. Intimacy and attachment have meaning not only for relationships that we think of traditionally as intimate (e.g., between partners, parents, and children) but for more extended ties. For instance, when relationships are solid at a community level, individuals feel strong bonds and attachment to places (e.g., neighborhood) and organizations (e.g., voluntary and religious).

3.2.2 Social influence. Networks may influence health via several other pathways. One pathway that is often ignored is based on social influence. Shared norms around health behaviors (e.g., alcohol and cigarette consumption, healthcare utilization) might be powerful sources of social influence with direct consequences for the behaviors of network members, quite apart from the provision of social support taking place within the network concurrently. For instance, cigarette smoking by peers is among the best predictors of smoking for adolescents (Landrine et al. 1994). The social influence which extends from the network's values and norms constitutes an important and under-appreciated pathway through which networks impact health.

3.2.3 Social engagement. A third and more difficult to define pathway by which networks may influence health status is by promoting social participation and social engagement. Participation and engagement result from the enactment of potential ties in real life activity. Getting together with friends, attending social functions, participating in occupational or social roles, group recreation, and church attendance are all instances of social engagement. Thus, through opportunities for engagement, social networks define and reinforce meaningful social roles including parental, familial, occupational, and community roles, which in turn provide a sense of value, belonging, and attachment. Several recent studies suggest that social engagement is critical in maintaining cognitive ability (Bassuk et al. 1999) and reducing mortality (Glass et al. 1999).

In addition, network participation provides opportunities for companionship and sociability. Rook (1990) argues that these behaviors and attitudes are not the result of the provision of support *per se*, but are the consequence of participation in a meaningful social context in and of itself. One reason measures of social integration or 'connectedness' may be such powerful predictors of mortality over long periods of follow-up is that these ties give meaning to individuals' lives by virtue of enabling them to participate in it fully, to be obligated (in fact, often to be providers of support), and to feel attached to their community.

3.2.4 Person-to-person contact. Another behavioral pathway by which networks influence disease is by restricting or promoting exposure to infectious disease agents. In this regard, the methodological links between epidemiology and networks are striking. What is perhaps most remarkable is that the same network characteristics that can be health-promoting can at the same time be health-damaging if they serve as vectors for the spread of infectious disease.

The contribution of social network analysis to the modeling of disease transmission is the understanding that in many, if not most, cases, disease transmission is not spread randomly throughout a population. Social network analysis is well suited to the development of models in which exposure between individuals is not random but rather is based on geographic location, sociodemographic characteristics (age, race, gender), or other important characteristics of the individual such as socioeconomic position, occupation, or sexual orientation (Laumann et al. 1989). Furthermore, because social network analysis focuses on characteristics of the network rather than on characteristics of the individual, it is suited ideally to the study of diffusion of transmissible diseases through populations via bridging ties between networks, or uncovering characteristics of ego-centered networks that promote the spread of disease.

3.2.5 Access to material resources. Surprisingly little research has sought to examine differential access to material goods, resources, and services as a mechanism through which social networks might operate. This is unfortunate given the work of sociologists showing that social networks operate by regulating an individual's access to life-opportunities by virtue of the extent to which networks overlap each other. In this way networks operate to provide access or to restrict opportunities in much the same way that social status works. Perhaps the most important among the studies exploring this tie is Granovetter's classic study of the power of 'weak ties' that, on the one hand lack intimacy, but on the other facilitate the diffusion of influence and information, and provide opportunities for mobility (Granovetter 1973).

There are five mechanisms by which the structure of social networks might influence disease patterns. While social support is the mechanism most commonly invoked, social networks also influence health through additional behavioral mechanisms including (a) forces of social influence, (b) levels of social engagement and participation, (c) the regulation of contact with infectious disease, and (d) access to material goods and resources. To date, the evidence linking aspects of social relationships to health outcomes is strongest for general measures of social integration, social support, and social engagement. However, it should be noted that these mechanisms are not mutually exclusive. In fact, it is most likely that in many cases they operate simultaneously.

4. Conclusion

The aim in this review was to integrate some classical theoretical work in sociology, anthropology, and psychiatry with the current empirical research underway on social networks, social integration, and social support. Rather than review the vast amount of work on health outcomes which is the subject of several excellent recent papers, a conceptual framework that guides work in the future has been developed.

With the development of this framework, two issues of profound importance stand out. The first is the 'upstream' question of identifying those conditions that influence the development and structure of social networks. Such questions have been the substantive focus of much of social network research, especially in relationship to urbanization, social stratification, and culture change. Yet little of this work becomes integrated with health issues in a way that might guide us in the development of policies or intervention to improve the health of the public. Of particular interest would be more cross-cultural work comparing countries with different values regarding social relationships, community, or sense of obligation. The same might be true to specific areas within countries, or specific cultural or ethnic groups with clearly defined values.

The second major issue relates to the 'downstream' question. Many investigators have assumed that networks influence health via social support functions. The framework in this article makes clear that this is but one pathway linking networks to health outcomes. Furthermore, the work on conflicts and stress points out that not only are not all relationships positive in valence, but that some of the most powerful impacts on health that social relationships may have are through acts of abuse, violence, and trauma (Rook 1992). Fully elucidating these downstream experiences, and how they are linked to health via which biological mechanisms, remains a major challenge in the field.

See also: Illness: Dyadic and Collective Coping; Integration: Social; Networks: Social; Social Network Models: Statistical; Social Networks and Fertility; Social Support and Health; Social Support and Recovery from Disease and Medical Procedures; Social Support and Stress; Social Support, Psychology of

Bibliography

Bassuk S, Glass T, Berkman L 1999 Social disengagement and incident cognitive decline in community-dwelling elderly persons. *Annals of Internal Medicine* **131**: 165–73

Berkman L 1986 Social networks, support and health: Taking the next step forward. *American Journal of Epidemiology* **123**(4): 559–62

Berkman L F, Glass T 2000 Social integration, social networks, social support and health. In: Berkman L F, Kawachi I (eds.) *Social Epidemiology*. Oxford University Press, New York, pp. 137–73

Berkman L F, Syme S 1979 Social networks, host resistance, and mortality: A nine-year follow-up of Alameda County residents. *American Journal of Epidemiology* **109**: 186–204

Bott E 1957 *Family and Social Network*. Tavistock Press, London

Bowlby J 1969 *Attachment and Loss*. Hogarth Press, London

Burt R S 1982 *Toward a Structural Theory of Action*. Academic Press, New York
Cohen S, Underwood S, Gottlieb B 2000 *Social Support Measures and Intervention*. Oxford University Press, New York
Durkheim E 1897 *Suicide*. Free Press, New York
Glass T, Mendes de Leon C, Marottoli R, Berkman L F 1999 Population based study of social and productive activities as predictors of survival among elderly Americans. *British Medical Journal* **319**: 478–83
Granovetter M 1973 The strength of weak ties. *American Journal of Sociology* **78**: 1360–80
Hall A, Wellman B 1985 Social networks and social support. In: Cohen S, Syme S L (eds.) *Social Support and Health*. Academic Press, Orlando, FL, pp. 23–41
House J S 1981 *Work Stress and Social Support*. Addison-Wesley, Reading, MA
House J S, Kahn R 1985 Measures and concepts of social support. In: Cohen S, Syme S L (eds.) *Social Support and Health*. Academic Press, Orlando, FL, pp. 79–108
House J S, Landis K R, Umberson D 1988 Social relationships and health. *Science* **241**: 540–5
House J S, Robbins C, Metzner H 1982 The association of social relationships and activities with mortality: Prospective evidence from the Tecumseh Community Health Study. *American Journal of Epidemiology* **116**: 123–40
Kaplan G, Salonen J, Cohen R, Brand R, Syme S, Puska P 1988 Social connections and mortality from all causes and cardiovascular disease: Prospective evidence from eastern Finland. *American Journal of Epidemiology* **128**: 370–80
Kawachi I, Colditz G A, Ascherio A, Rimm E B, Giovannucci E, Stampfer M J et al. 1996 A prospective study of social networks in relation to total mortality and cardiovascular disease in men in the U.S.A. *Journal of Epidemiological Community Health* **50**: 245–51
Kawachi I, Kennedy B, Lochner K, Prothrow-Stith D 1997 Social capital, income inequality, and mortality. *American Journal of Public Health* **87**: 1491–8
Landrine H, Richardson J K, Klondoff E A, Flay B 1994 Cultural diversity in the predictors of adolescent cigarette smoking: the relative influence of peers. *Journal of Behavioral Medicine* **17**: 331–6
Laumann E O, Gagnon J H, Michaels S, Michael R T, Coleman J S 1989 Monitoring the AIDS epidemic in the US: A network approach. *Science* **244**: 1186–9
Mitchell J C 1969 *The Concept and Use of Social Networks*. Manchester University Press, Manchester, UK
Orth-Gomer K, Johnson J 1987 Social network interaction and mortality: A six year follow-up study of a random sample of the Swedish population. *Journal of Chronic Diseases* **40**: 949–57
Pennix B W, van Tilburg T, Kriegsman D M, Deeg D J, Boeke A J, van Eijk J T 1997 Effects of social support and personal coping resources on mortality in older age: The Longitudinal Aging Study, Amsterdam. *American Journal of Epidemiology* **146**: 510–9
Rook K S 1990 Social relationships as a source of companionship: Implications for older adults' psychological well being. In: Sarason B R, Sarason T G, Pierce G R (eds.) *Social Support: An Interactional View*. Wiley, New York, pp. 221–50
Rook K S 1992 Detrimental aspects of social relationships: Taking stock of an emerging literature. In: Viel H O, Baumann U (eds.) *The Meaning and Measurement of Social Support*. Hemisphere, New York, pp. 157–69
Sarason B R, Sarason I G, Pierce G (eds.) 1990 *Social Support: An Interactional View*. Wiley, New York
Schoenbach V, Kaplan B, Freedman L, Kleinbaum D 1986 Social ties and mortality in Evans County, Georgia. *American Journal of Epidemiology* **123**: 577–91
Seeman T, Berkman L, Kohout F, LaCRoix A, Glynn R, Blazer D 1993 Intercommunity variation in the association between social ties and mortality in the elderly: A comparative analysis of three communities. *Annals of Epidemiology* **3**: 325–35
Sugisawa H, Liang J, Liu X 1994 Social networks, social support and mortality among older people in Japan. *Journal of Gerontology* **49**: S3–S13
Vogt T M, Mullooly J P, Ernst D, Pope C R, Hollis J F 1992 Social networks as predictors of ischemic heart disease, cancer, stroke, and hypertension: Incidence, survival and mortality. *Journal of Clinical Epidemiology* **45**: 659–66
Welin L, Tibblin G, Svardsudd K, Tibblin B, Ander-Peciva S, Larsson B et al. 1985 Prospective study of social influences on mortality: The study of men born in 1913 and 1923. *Lancet* **1**: 915–8

L. F. Berkman

Social Intervention in Psychiatric Disorders

As early as the nineteenth century psychiatrists already suspected that harmful influences from the social environment might promote the appearance of mental illnesses or have a deleterious effect on their course. In that era, the isolation of the mentally ill in the 'healthful' milieu of institutions far removed from their original living environment was thought to be an appropriate means of shielding the patients from the disease-producing influences of their surroundings. The objective of psychiatric reforms since the mid to late 1950s was to resettle chronically mentally ill persons from large custodial institutions to community settings. The idea behind this was to (re-)integrate patients in their normal living environment as rapidly as possible, because it was thought that their natural social network would make a major supporting contribution to the healing process (Rössler 2001).

This reform has changed the perception of mental illness. Enabling disabled persons to live a normal life in the community caused a shift away from a focus on an illness model toward a model of functional disability (Grove 1994). As such, other outcome measures aside from clinical conditions became relevant. Since then social role functioning including social relationships, work, and leisure, as well as quality of life and the family burden, became of major interest for affected individuals living in the community as well as for the health professionals involved in treatment and care of those persons. Most of the social intervention

strategies as applied by health professionals in the community can be integrated within the concept of psychiatric rehabilitation (Rössler 2000).

In current scientific discourse, environmental and psychosocial factors are often relegated to the status of a single variable of merely peripheral interest, without further specification. Biological science has indeed made enormous progress toward an understanding of mental disorders, but it should not be forgotten that research in social sciences also yielded a considerable body of knowledge with respect to the psychosocial and environmental factors that influence it (Rössler 2001).

The relevance of psychosocial and environmental problems is also reflected in the DSM IV and ICD-10. Axis 4 of the DSM (American Psychiatric Association 1994) and codes Z 55–Z 65 and Z 73 of the ICD-10 (WHO 1993) are assigned for reporting psychosocial and environmental problems that may affect the diagnosis, treatment, and prognosis of mental disorders.

1. *Target Population of Social Interventions*

Although the majority of the chronically mentally ill have the diagnosis of schizophrenic disorders, other patient groups with psychotic and nonpsychotic disorders are targeted by social interventions. All patients suffering from severe mental illness (SMI) now require social interventions. The core group is drawn from patients with:

- persistent psychopathology,
- marked instability characterized by frequent relapses, and
- social maladaptation (Royal College of Psychiatrists 1996).

There are other definitions currently used to characterize the chronically mentally ill, but they all share some common elements, i.e. a diagnosis of mental illness, prolonged duration, and role incapacity.

Furthermore, up to 50 percent of persons with SMI carry dual diagnoses, especially in combination with substance abuse (Cuffel 1996). Young adult chronic patients constitute an additional category that is diagnostically more complicated (Schwartz et al. 1983). These patients present complex pictures of symptomatology difficult to categorize within our diagnostic and classification systems. Many of them also have a history of suicide attempts. All in all, they represent a patient population that is mostly difficult to treat.

2. *Models of Vulnerability*

There are several vulnerability–stress models forming the theoretical basis into which biological and social intervention approaches of psychiatric treatment, rehabilitation, and care can be integrated. These models were developed mainly for persons with schizophrenic disorders, but they apply essentially to all individuals with SMI.

In the *vulnerability–stress* model it is presumed that socioenvironmental stressors superimposed on an underlying and enduring biological vulnerability lead to abnormalities in central nervous system functioning (Zubin and Spring 1977). In the *vulnerability–stress–competence* model (Nuechterlein and Dawson 1984), individuals, families, and natural support systems can cope with stressors and have the competence to create a higher threshold for symptomatic illness. In an *interactive–developmental* model (Strauss et al. 1981), the persons affected do not remain in a passive role but actively interact with their environment. They modulate and select the kind of help they get from other people, they choose to comply with the treatment, and they seek or avoid social support and stressful situations.

The interactive–developmental model is the basis for the consumer movement in mental health care. People with mental disorders and their caregivers prefer to see themselves as consumers of mental health services with an active interest in learning about psychiatric disorders and in selecting the respective treatment approaches. Consumerism makes it possible to take up the affected person's perspective and to consider seriously courses of action relevant to them (Kopelowicz and Liberman 1995).

3. *Conceptual Framework of Social Interventions*

The overall philosophy of social interventions in psychiatric disorders comprises two intervention strategies: the first strategy is individual-centered and aims at developing the patient's skill in interacting with a stressful environment. There is no consensus among researchers as to what this individual-centered approach actually does accomplish. Some understand developing patient's skill as an approach to help disabled persons to compensate for impairments and to function optimally with the deficiencies they have. Others assume that different forms of individual-centered interventions help with recovery from the disorder itself, while the contributing factors to the healing process are not clear (Strauss 1986).

The second strategy is an ecological approach directed toward developing environmental resources to reduce potential stressors. This approach includes social interventions concerning housing and occupation as well as measures to build natural networks, as many chronically mentally ill and disabled persons lose close and stable relationships in the course of the disease. Most disabled persons need a combination of both approaches.

In contrast to acute treatment there are mostly no legal powers to enforce social interventions in long-term care. Thus, the sufferer's autonomy concerning treatment decisions should be respected. Within this

frame, therapeutic alliance plays a crucially important role in engaging the affected persons in their own care planning. As such, the therapist should acknowledge that disagreement about the illness between patient and therapist is not always the result of the illness process (Eichenberger and Rössler 2000).

4. *The International Classification of Impairments, Disabilities, and Handicaps*

The long-term consequences of major mental disorders targeted by social interventions might be described on different levels. A useful framework is provided by the international classification of impairment, disability, and handicaps (WHO 1980). In addition to acute symptoms, chronically mentally ill patients, in particular individuals with schizophrenia, have impairments in the most basic cognitive processes concerning information input, memory, and abstraction. These cognitive impairments are present even when all symptoms have subsided, and they seem to be enduring markers of a vulnerability to the illness. Persistent and severe impairments lead to disability, i.e., affected persons will be limited in performing certain roles in their social and occupational environment. Finally, handicaps occur when disabilities place the individual at a disadvantage relative to others in society.

On more specified levels concerning impairments, medication and cognitive training are provided. Effective pharmacotherapy and cognitive training stabilize affected persons so that they are more responsive to learning from their environment. Interventions concerning disability encompass mainly social skills training and family intervention. Furthermore, social interventions emphasize vocational training suitable to the extent of the handicap.

4.1 *Reduction of Impairments*

4.1.1 Medication. Even when taken regularly, medication is only able to reduce rather than abolish symptoms, and delay rather than prevent relapse. In addition to that, many of the traditional drugs have adverse effects and can weaken patients' ability to perform their social roles or can impair their social and vocational functioning. As such, it is not surprising that noncompliance with medication is one of the most serious problems in the long-term care of persons with SMI.

Several treatment strategies have been developed for improving compliance with medication. For patients with cognitive deficits, the complexity of the medication regimen should be reduced (Gaebel 1997). Some patients need to be given the medication by a professional service on a regular basis. Depot medication (i.e., intramuscular injections that release the medication slowly into the bloodstream) might be especially useful for patients with fluctuating motivation or insight. It is also useful to involve families and caregivers in increasing compliance with medication.

Yet, many patients living in the community have to take responsibility for their medication themselves. Training in self-management of medication (Eckman et al. 1990) emphasizes the patient's autonomy and increases acceptance of and responsibility for treatment. This also includes variation of medication without consultation within certain limits.

4.1.2 Cognitive training. With respect to cognitive deficits, consistent results emerged from most studies in so far as remediation of specific deficits through repeated practice or related techniques is successful. But it is not clear if and how recovered cognitive deficits should facilitate the remediation of functional outcome and community functioning. Addressing these limitations, Brenner et al. (1992) developed a comprehensive cognitive training program which they called 'integrated psychological therapy' (IPT). IPT comprises five stages: cognitive differentiation, social perception, communication skills, interpersonal problem solving, and social skills training. It is presented to groups of five to seven patients for 30 to 60 minutes two or three times a week. The hierarchically arranged program is aimed at facilitating acquisition and maintenance of more complex skills by remediating cognitive deficits. Findings indicate that IPT reduces psychotic disorganization and improves social, cognitive, and problem-solving skills (Penn and Mueser 1996).

4.2 *Remediation of Disabilities*

4.2.1 Skills training. In recent years, social skills training in psychiatric rehabilitation has become very popular and has been disseminated widely. The most prominent proponent of skills training is Robert Liberman, who has designed systematic and structured skills training since the mid-1970s. Liberman and his colleagues packaged the training in the form of modules which focus on different topics such as medication management, symptom management, substance abuse management, basic conversational skills, interpersonal problem solving, friendship and intimacy, recreation and leisure, workplace fundamentals, community (re-)entry, and family involvement. Each module is composed of skill areas. The symptom management module, for example, encompasses skills areas such as identifying warning signs of relapse, developing a relapse prevention plan, coping with persistent symptoms, and avoiding alcohol and drugs. The skills areas are taught in exercises with demonstration videos, role play and problem-solving exercises, and *in vivo* and homework assignments (Liberman 1988).

The results of several controlled studies suggest that disabled individuals can be taught a wide range of social skills. Social and community functioning improve when the trained skills are relevant for the affected person's daily life, and the environment perceives and reinforces the changed behavior. Unlike medication effects, benefits from skills training occur more slowly. Furthermore, long-term training has to be provided for positive effects (Penn and Mueser 1996).

4.2.2 Family intervention. As a consequence of deinstitutionalization, the burden of care has fallen increasingly on the relatives of the mentally ill. It is estimated that up to 65 percent of mentally disabled individuals live with their families (Goldmann et al. 1981). This is a task many families do not choose voluntarily. Additionally, not all families are equally capable of giving full support to their disabled member, and may not be willing to replace an insufficient health care system (Johnson 1990). But families also represent support systems which provide natural settings for context-dependent learning important for recovery of functioning (Schooler 1995). As such there has been a growing interest in helping families since the beginnings of care reforms (Strachan 1986).

One area of interest dealt with the expectations of relatives concerning the provision of care. Relatives quite often feel ignored, not taken seriously, and insufficiently informed by health professionals. They also may feel that their contribution to care is not appreciated, or that they will be blamed for the patient's problems. It is certainly not surprising that there is a lot of frustration and resentment among relatives related to the extent of the physical, financial, and emotional family burden. As important research has demonstrated a relationship between a negative affective family atmosphere and an unfavorable course of a mental disorder, the need for helping families and patients with family intervention programs became clear.

All family intervention programs include an educational aspect covering basic information on the etiology, treatment, and prognosis of schizophrenia. Although family interventions differ in the treatment focus, all family interventions are organized around the principal goal of providing family members with more information about the disorder and strategies for managing common problems. There are more common features: family interventions are concrete, practical, and focus on everyday coping.

Family intervention programs have produced the most promising results. Family intervention is effective for lowering relapse rate and improving outcome (e.g., social functioning). Possibly, family intervention can reduce family burden. Furthermore, the treatment gains are fairly stable.

However, it is not clear what the effective components of the different models are. Additionally, family interventions differ in frequency and length of treatment. There are also no criteria for the minimum amount of treatment necessary.

Finally, it should be noted that most of the familiy interventions were developed in the context of Western societies during deinstitutionalization. Family caregiving might be quite different in a different cultural context. This refers to other cultures in total as well as to minority groups in Western societies (Guarnaccia 1998).

4.2.3 Reduction of handicaps by vocational rehabilitation. The beneficial effects of work for mental health have been known for centuries (Harding et al. 1987). Vocational rehabilitation therefore has been a core element of psychiatric reforms since its beginnings.

Vocational rehabilitation is based on the assumption that work not only improves activity, social contacts, etc., but may also promote gains in related areas such as self-esteem and quality of life, as work and employment are a step away from dependency toward integration into society. Enhanced self-esteem in turn improves adherence to long-term care of individuals with impaired insight (McElroy 1987).

Vocational rehabilitation originated in psychiatric institutions, where the lack of activity and stimulation led to apathy and withdrawal among their inpatients. Long before the introduction of medication, occupational and work therapy contributed to sustainable improvements in long-stay inpatients. Today occupational and work therapy are no longer hospital-based, but represent the starting point for a wide variety of rehabilitative techniques teaching vocational skills (Royal College of Psychiatrists/College of Occupational Therapists 1992).

Vocational rehabilitation programs in the community provide a series of graded steps to promote job (re-)entry. For less disabled persons, brief and focused techniques are used to teach them how to find a job, fill out applications, and conduct employment interviews (Jacobs et al. 1988). In transitional employment a temporary work environment is provided to teach vocational skills that should enable the affected person to move on to competitive employment. But too often the gap between transitional and competitive employment is so large that the mentally disabled individuals remain in a temporary work environment. Sheltered workshops providing prevocational training also quite often prove to be a dead end for disabled persons.

Today, the most promising vocational rehabilitation model is supported employment. In this model, disabled persons are placed in competitive employment as soon as possible and receive all the support needed to maintain their position. Skills are taught in

the actual environment, tailored to the job and the individual. If necessary, additional advice is given to the employer. As the course of SMI is unpredictable and job requirements often change, the support provided is neither time-limited nor delivered on a fixed schedule (Wallace 1998). Participation in supported employment programs is followed by an increase in the ability to find and keep employment (Baronet and Gerber 1998). But most of the research does not allow evaluation of the impact of vocational rehabilitation programs on nonvocational domains.

Although findings regarding supported employment are encouraging, many critical issues remain unanswered. Most individuals in supported employment obtain unskilled part-time jobs. Therefore, the high drop-out rate (up to 40%) is not surprising. Since most studies only evaluated short (12–18 months) follow-up periods, the long-term impact remains unclear. At the time of writing, it is not known which individuals benefit from supported employment and which do not (Mueser et al. 1998).

It must be acknowledged that the integration into the labor market by no means depends only on the ability of the persons affected to fulfil a work role and on the provision of sophisticated vocational training and support techniques, but also on the willingness of society to integrate its most disabled members.

One result of the difficulties of integrating mentally disabled individuals into the common labor market has been the steady growth of cooperatives which operate commercially with disabled and nondisabled staff working together on equal terms and sharing in management. The mental health professionals work in the background providing support and expertise (Grove 1994). Cooperatives serve as a bridge between sheltered workshops and competitive jobs on the common labor market.

4.2.4 Developing environmental resources. Effective community psychiatry requires individualized and specialized treatment which has to be embedded in a comprehensive and coordinated system of community services. But even when a variety of services is available, they are poorly linked in many cases, and costly duplication may occur.

While developing community support systems it became obvious that there was a need to coordinate and integrate the services provided as each professional involved concentrates on different aspects of the same patient. Therefore, as a key coordinating and integrating mechanism, the concept of case management (CM) originated. CM focuses on all aspects of the physical and social environment. The core elements of CM are the assessment of patient needs, the development of comprehensive service plans for the patients, and arrangement of service delivery (Rössler et al. 1992).

Since the 1980s, a range of different models of CM have been developed which exceed the original idea that CM intends mainly to link the patient to needed services, and to coordinate those services. Today most clinical case managers also provide direct services in the patient's natural environment. This model is called intensive case management (ICM). ICM is difficult to distinguish from assertive community treatment (ACT).

The basic compounds of ACT were developed by Stein and Test in the 1970s (Stein and Test 1980). The original program was designed as a community-based alternative to hospital treatment for persons with severe mental illnesses. A comprehensive range of treatment, rehabilitation, and support services in the community is provided through a multidisciplinary team. ACT is characterized by an assertive outreach approach, i.e., interventions are provided mainly in the natural environment of the disabled individuals (Scott and Dixon 1995). Research on CM and ACT indicates that these models can reduce time in hospital, but have only small or moderate effects on improving symptomatology and social functioning (Mueser et al. 1998).

5. *Conclusions*

The goal of social intervention embedded in the broader concept of psychiatric rehabilitation is to help disabled individuals to establish the emotional, social, and intellectual skills needed to live, learn, and work in the community with the least amount of professional support (Anthony 1979). The refinement of all these approaches as described has achieved a level where they should be made readily available for every person with SMI.

But we have to be aware that it is a long way to translate research into practice. In 1992, the National Institute of Mental Health, for example, funded the 'schizophrenia patient outcome research team' (PORT) to develop and disseminate recommendations for the treatment of schizophrenia based on existing scientific evidence. The recommendations address in essence all interventions as presented above (Lehman and Steinwachs 1998). Lehman and colleagues assessed the patterns of usual care for schizophrenic patients and examined the conformance rate with the PORT treatment recommendations. The conformance rate was modest, generally below 50 percent. As such, it is obvious that current care practice has to be improved substantially in the light of the outcome research available (Lehman et al. 1995).

Serious mental illness can devastate the lives of people who suffer from it. These persons have a right to be treated according to current knowledge. But mental illnesses are associated with a significant amount of stigma, resulting in social isolation and discrimination in housing, education, and employ-

ment opportunities. Mentally ill people are even disadvantaged in receiving appropriate treatment and care, so stigma and discrimination further increase the burden on affected persons and their relatives. Yet, social interventions aim at reintegrating the mentally ill into the community and improving their quality of life. For these reasons a responsive and open-minded environment is indispensable. Therefore, translating research into practice always encompasses combating stigma and discrimination of mental illnesses as well (World Psychiatric Association 1999).

See also: Health Interventions: Community-based; Mental Health: Community Interventions; Mental Illness, Epidemiology of; Mental Illness: Family and Patient Organizations; Public Psychiatry and Public Mental Health Systems; Vulnerability and Perceived Susceptibility, Psychology of

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders* (*DSM IV*), 4th edn. American Psychiatric Association, Washington, DC

Anthony C W 1979 *The Principles of Psychiatric Rehabilitation*. University Park Press, Baltimore

Baronet A M, Gerber G 1998 Psychiatric rehabilitation: Efficacy of four models. *Clinical Psychology Review* **18**: 189–228

Brenner H D, Hodel B, Genner R, Roder V, Corrigan P W 1992 Biological and cognitive vulnerability factors in schizophrenia: Implications for treatment. *British Journal of Psychiatry* **161**(suppl. 18): 154–63

Cuffel B 1996 Comorbid substances use disorder: Prevalence, patterns of use, and course. In: Drake R E, Mueser K T (eds.) *Dual Diagnosis of Major Mental Illness and Substance Abuse: Recent Research and Clinical Implications*. Jossey-Bass, San Francisco, Vol. 2, pp. 93–105

Eckman T, Liberman R, Phipps C, Blair K 1990 Teaching medication management skills to schizophrenic patients. *Journal of Clinical Psychopharmacology* **10**: 33–8

Eichenberger A, Rössler W 2000 Comparison of self-ratings and therapist ratings of outpatients' psychosocial status. *Journal of Nervous and Mental Disease* **188**: 297–300

Gaebel W 1997 Towards the improvement of compliance: The significance of psychoeducation and new antipsychotic drugs. *International Clinical Psychopharmacology* **12**(suppl. 1): 37–42

Goldmann H H, Gattozzi A A, Taube C A 1981 Defining and counting the chronically mentally ill. *Hospital and Community Psychiatry* **32**: 21–7

Grove B 1994 Reform of mental health care in Europe. Progress and change in the last decade. *British Journal of Psychiatry* **165**: 431–3

Guarnaccia P J 1998 Multicultural experiences of family caregiving: A study of African American, European American, and Hispanic American families. *New Directions for Mental Health Services* **77**: 45–61

Harding C, Strauss J, Hafez H, Lieberman P 1987 Work and mental illness: I. Toward an integration of the rehabilitation process. *Journal of Nervous and Mental Disease* **175**: 317–26

Jacobs H E, Kardashian S, Kreinbring R K, Ponder R, Simpson A 1988 A skills-oriented model for facilitating employment in psychiatrically disabled persons. *Rehabilitation Counseling Bulletin* **27**: 87–96

Johnson D 1990 The family's experience of living with mental illness. In: Lefley H, Johnson D (eds.) *Families as Allies in Treatment of the Mentally Ill: New Directions for Mental Health Professionals*. American Psychiatric Press, Washington, DC, pp. 31–63

Kopelowicz A, Liberman R P 1995 Biobehavioral treatment and rehabilitation of schizophrenia. *Harvard Review of Psychiatry* **3**: 55–64

Lehman A F, Carpenter W T J, Goldman H H, Steinwachs D M 1995 Treatment outcomes in schizophrenia: Implications for practice, policy, and research. *Schizophrenia Bulletin* **21**: 669–75

Lehman A F, Steinwachs D M 1998 Survey Co-Investigators of the PORT Project. At issue: Translating research into practice: The Schizophrenia Patient Outcomes Research Team (PORT) Treatment Recommendations; and Patterns of usual care for schizophrenia: Initial results from the Schizophrenia Patient Outcomes Research Team (PORT) Client Survey. *Schizophrenia Bulletin* **24**: 1–10 , 11–20

Liberman R (ed.) 1988 *Psychiatric Rehabilitation of Chronic Mental Patients*. American Psychiatric Press, Washington DC

McElroy E 1987 Sources of distress among families of the hospitalized mentally ill. *New Directions for Mental Health Services* **34**: 61–72

Mueser K T, Bond G R, Drake R E, Resnick S G 1998 Models of community care for severe mental illness: A review of research on case management. *Schizophrenia Bulletin* **24**: 37–74

Nuechterlein K H, Dawson M E 1984 A heuristic vulnerability/stress model of schizophrenic episodes. *Schizophrenia Bulletin* **10**(2): 300–12

Penn D L, Mueser K T 1996 Research update on the psychosocial treatment of schizophrenia. *American Journal of Psychiatry* **153**: 607–17

Rössler W 2000 Rehabilitation techniques. In: Gelder M G, Lopez-Ibor J J (eds.) *New Oxford Textbook of Psychiatry*. Oxford University Press, Oxford, pp. 1141–95

Rössler W 2001 Schizophrenia: Psychosocial factors. In: Henn F, Sartorius N, Helmchen H, Lauter H (eds.) *Contemporary Psychiatry, Vol. 3, Part 1*. Springer, Berlin, pp. 121–8

Rössler W, Fätkenheuer B, Löffler W, Riecher-Rössler A 1992 Does case management reduce the rehospitalization rate? *Acta Psychiatrica Scandinavica* **86**: 445–9

Royal College of Psychiatrists 1996 *Psychiatric Rehabilitation*, rev. edn. Gaskell, London

Royal College of Psychiatrists/College of Occupational Therapists 1992 *Occupational Therapy and Mental Disorders. A Consensus Statement by the Royal College of Psychiatrists and the College of Occupational Therapists*. Royal College of Psychiatrists, London

Schooler N 1995 Integration of family and drug treatment strategies in the treatment of schizophrenia: A selective review. *International Clinical Psychopharmacology* **10**(suppl. 3): 73–80

Schwartz S, Goldfinger S, Ratener M, Cutler D 1983 The young adult patient and the care system: Fragmentation prototypes. *New Directions for Mental Health Services* **19**: 23–36

Scott J E, Dixon L B 1995 Assertive community treatment and case management for schizophrenia. *Schizophrenia Bulletin* **21**: 657–68

Stein L I, Test M A 1980 Alternative to mental hospital treatment: I. Conceptual model, treatment program, and clinical evaluation. *Archives of General Psychiatry* **37**: 392–7
Strachan A 1986 Family intervention for the rehabilitation of schizophrenia: Toward protection and coping. *Schizophrenia Bulletin* **4**: 678–98
Strauss J S 1986 What does rehabilitation accomplish? *Schizophrenia Bulletin* **12**: 720–3
Strauss J, Bartko J, Carpenter W T Jr 1981 New directions in diagnosis: The longitudinal processes of schizophrenia. *American Journal of Psychiatry* **138**: 954–8
Wallace C J 1998 Social skills training in psychiatric rehabilitation: Recent findings. *International Review of Psychiatry* **19**: 9–19
World Health Organization (WHO) 1980 *International Classification of Impairments, Disabilities and Handicaps*. WHO, Geneva
World Health Organization (WHO) 1993 *International Statistical Classification of Diseases and Related Health Problems: Tenth Revision*. WHO, Geneva
World Psychiatric Association 1999 *The WPA Global Programme Against Stigma and Discrimination Because of Schizophrenia*. WPA, Geneva
Zubin J, Spring B 1977 Vulnerability—a new view of schizophrenia. *Journal of Abnormal Psychology* **86**: 103–26

W. Rössler

Social Justice

Social justice is a subcategory of the wider term justice and concerns justice as it refers to the distribution of valued goods and necessary burdens. Such distributions can take place at all levels of societal aggregation: at the micro level of face-to-face interaction, at the meso level of intermediate institutions and organizations, and at the macro level of a society's basic structure. The ways in which these various levels provide (or deny) access to social positions, in which they regulate the allocation of rights and duties as well as of scarce goods and necessary burdens, and in which they interact through an encompassing framework of institutionalized rules and procedures, form the subject of social justice.

1. *The Levels of Social Justice*

1.1 *The Macro Level*

The macro level is itself subdivided into several layers. The highest layer is that of the constitutional order of a society: its polity determining the (permitted) forms of government and the ground rules, as well as scope, of 'normal' politics; its economic regime, its social and legal systems, and the integration of its major institutions into one large scheme. The next lower layer is that of concrete politics. Here, different social groups compete for influence in the designation of particular policies concerning all aspects of social life which can become a matter of political and legal regulation under the chosen constitutional frame. And finally, there are the policies themselves, e.g., particular fiscal, economic, educational, welfare, etc. policies. The initial design of such policies can have far-reaching implications for future policies because once a particular policy structure or pattern is established in a given field, it tends to shape the public's expectations towards it and to lead to a certain degree of inertia, making radical changes difficult and limiting the options available to future policy makers. In many ways, these policies also affect what happens at the meso level.

1.2 *The Meso Level*

This is the level of organizations. When considering organizations from the point of view of social justice, it is useful to distinguish between types of organizations according to the societal sector or social subsystem into which they fall, e.g., business organizations belonging to the economic system, political parties belonging to the political system, hospitals belonging to the health system, schools belonging to the educational system, etc., because the various subsystems follow different logics of operation, subjecting their organizations to separate 'local' rationalities which in turn affect the ways in which these organizations (and their representatives) conceive of and deal with 'local' justice problems. Local justice problems concern the allocation of various in-kind (and mostly non- or not fully marketable) goods such as university places, jobs, public housing, and scarce medical resources; in short, the rules for inclusion into or exclusion from organizations and for the distribution of some of their main 'products' and services. The problems are typically dealt with and resolved by the respective organizations themselves. However, both the degree of local autonomy and of local scarcity are subject to the influence of macro level decisions. Thus, the state can mandate particular principles that must be observed in their handling (e.g., anti-discrimination laws) or it can refrain from doing so; it can issue rules for some such problems but not for all of them, and these rules can be more or less determinate. As for the degree of scarcity, on the other hand, the state's influence is particularly evident in the health sphere. The lower the budget assigned to healthcare as a whole (macro allocation level), the greater the need for rationing in the various branches of the health system (micro allocation level).

1.3 *The Micro Level*

This level is not to be confused with the above micro allocation level referring to problems arising within organizations. Instead it concerns a separate societal

level, namely that of interactions between individuals in relatively unorganized social contexts. An example are families and the ways in which they allocate the burdens facing them in their everyday life. It is obvious that at least theoretically this level of decision making leaves the widest scope for individual choice. However, in reality this scope is multiply constrained by cultural traditions framing the perception of appropriate choices, even though the traditions themselves may undergo constant changes. Moreover, higher-level decisions can have a significant impact on this micro level as well. Thus, if a welfare state exists which provides generous old age pensions, then this will obviously relieve the youngers' (immediate) duties against their parents. Likewise, if such a welfare state implicitly favors a pattern of lifelong, uninterrupted full-time employment, then, given a cultural background in which it is 'normal' for women to bear the bulk of child-rearing duties coupled with a tendency for work organizations to allocate better paid jobs to men, this will weaken women's bargaining position in the negotiation of the distribution of household chores.

2. *Theories of Social Justice*

Most theories of social justice focus on the macro level. Some, like the famous Theory of Justice by John Rawls (1971), view this level, or more precisely, its highest layer, as the primary, if not only subject of a theory of social justice (see also Barry 1995). Other conceptions, such as Michael Walzer's (1983), include a much wider range of issues.

Many theories of the first type deliberately restrict themselves to the formulation of principles for a society's basic structure. Once these principles are met, they are willing to leave the 'rest' to the political process. If the rules by which political decisions are made are fair, then, according to these theories, the outcomes must be accepted as legitimate by everyone because they have been brought about in a manner against which no one can reasonably object. Hence, the obligation to comply with policies and norms even if one objects to their content. The nature of these theories is thus predominantly procedural (e.g., Habermas 1983, Röhl and Machura 1997), although some, like the Rawlsian theory, also formulate one or more substantive principles. There are also macro level theories which go beyond the basic structure and address issues of normal politics as well, e.g., welfare policies relating to public education, medical care, or unemployment benefits, as in the case of David Miller (1976).

The second group of theories is much broader in scope, covering the whole range from micro to macro issues. More recently, there has been a development towards more specialization on particular types of justice problems arising within single subsystems of society. An example of this is the proliferating work in the so-called applied ethics, especially in the fields of medical and bioethics. Some of these theories also cover issues of distributive justice, with special attention devoted to problems of medical triage. Another recent development is the aim to cut across the various social subsystems and to cover, in a systematic way, the whole range of micro allocation problems facing intermediate institutions and organizations. This is the local justice approach developed by Jon Elster (1992). And finally, there is a growing body of feminist literature (e.g., Okin 1989). Much of this work starts from relative disadvantages that women encounter at the micro and meso levels and looks at the ways in which macro level institutions or policies account for them.

One pervasive problem in the theoretical literature on social justice is a widespread lack of awareness of differences in the subject matters of the available conceptions. The result is that the participants in the pertinent debates often talk past each other. They will, for instance, argue, that a principle suggested by a particular theory is unsuitable for a particular problem, but ignore that the principle was never meant to regulate problems of that sort because its suggested scope is much more limited than the critics impute. Or they will argue that a theory fails to give due weight to a particular principle, but fail to realize that the principle in question may be appropriate for some situations, but not necessarily for the ones forming the subject of the criticized theory. A favorite target of this sort of criticism is John Rawls' theory.

A second problem is that much of the theorizing in the philosophical literature lacks a proper understanding of the structure and functioning conditions of modern society. But political theories (as they often conceive themselves) that do not 'know' the society whose operations they seek to guide (see Luhmann 1980 for this criticism) are unlikely to be politically very instructive. The social theory underlying most normative theories of social justice is often rather simplistic, variously modeled after the Greek polis, a medieval village community, the market place or a handful of stranded islanders having to regulate their common affairs. The complexity of modern society is mostly ignored by this kind of theorizing—although its principles are proposed for the operations of real societies, not fictitious ones.

A third problem is that the available theories of justice are anything but clear about the meaning of social justice. Some associate it with a particular understanding of equality, others equate it with some notion of fairness, and still others use the term in still another way. There is a plethora of theories promoting different principles of justice, some favoring desert (or equity) as the first principle, others rights, and still others needs, but even those choosing the same principle may interpret it very differently. For example, how is one to determine, measure and

remunerate a person's deserts (and in which contexts are they relevant)? No consensus exists on any of these questions in the pertinent literature. Moreover, there are not only various single-principle conceptions, but also proposals for pluralist theories, some covering all three of the above principles, others allowing for as many principles as there are 'spheres' of justice. In the latter case, the meaning ascribed to social justice becomes largely a function of the problem at hand. Taken together, these (and further) difficulties leave the impression of a great deal of messiness and ambiguity in the (mostly normative) theorizing of social justice.

3. *Empirical Work on Social Justice*

Not much different is the situation in the social science literature on social justice. Here, the starting point is often some ideal of equality drawn from the enlightenment tradition and forming a standard against which the performance of a society's institutions is judged (e.g., Kluegel and Smith 1986). The ideal is rarely openly expressed or explicitly defended, but nevertheless drives ongoing attempts at laying bare inequalities in all areas of life and the social mechanisms underlying and reproducing them, with the implicit message being that the contradiction between ideal and reality calls for social reform. One problem with this literature is its hidden normativity, its unquestioned support for a certain notion of equality whose rightness is presumed (and presupposed) rather than argued for. Another problem is that it possesses no measure for admissible degrees of inequality or for prescribed degrees of equality. This problem is discussed with great vigor and sophistication in philosophy (see, e.g., Dworkin 1981a, 1981b) and (welfare) economics (e.g., Sen 1992). The advances made in this literature, however, have so far remained both unattended and unmatched in the social scientific work on inequality.

In a related way, this problem also comes up in a second body of social science literature on social justice, that on judgments of justice. Its primary subject is that of income (in)justice and (in)equality (see, e.g., Alves and Rossi 1978, Kluegel et al. 1995). To simplify somewhat, a recurrent finding across societies and social categories is that most people tend to favor some degree of inequality over absolute equality (although they disagree about the proper extent and about the fields in which inequalities may prevail or should exist). They also tend to underrate the factually existing income (and hence, wealth) differences. Still, they would mostly prefer a narrowing of the differences that they believe to exist, even though most people are roughly satisfied with their own earnings. Furthermore, there is a widespread belief that people's earnings roughly reflect their economic desert and that this is also what they should reflect. At the same time, market economies are generally preferred over planned economies, and their outcomes are largely taken to be desert-based. However, market allocations are anything but that; they are, as has been noted repeatedly, in principle unprincipled (Hirsch 1977). And this is by no means the only inconsistency; the one impression that inevitably emerges from the literature on people's judgments is that these judgments are multiply confused. So what is one to make of them?

A third type of empirical work on social justice is Elster's local justice approach. This work is concerned less with judgments of justice than with the ways in which justice problems arising in organizations are actually resolved and with the weight that justice considerations are accorded in such situations (see Elster 1995, Schmidt and Hartmann 1997). The aim is to supplement the normative literature with factual knowledge about the fields into which it seeks to intervene, the underlying premise being that normative theories need empirical foundations in order to render them applicable to a social reality which is much more complex than philosophical theories usually allow for.

Finally, in a series of articles Raymond Boudon (1997, 1999; see also Boudon and Betton 1999) has suggested a rationalist and cognitivist conception of justice and morality more generally, according to which there are objective normative truths just as there are objective scientific truths, the basis for accepting them being in both cases their grounding in strong reasons. This conception, which is directed primarily against relativistic and culturalistic positions, seems very promising but also raises a number of methodological questions: how are we to determine, by sociological means, what these truths are, how can we 'find' and accurately identify them, and what are the techniques needed to separate them from those normative statements and proposals which only claim to be true? Given the novelty and as yet mainly programmatic nature of Boudon's approach, it seems too early for a final assessment of its merits.

4. *Desiderata*

The concept of social justice is used very differently in the available literature, so if one is not already committed to a particular definition, it is hard to distil a common core out of its many usages. To some extent this has to do with the term's indiscriminate application to almost any kind of distribution problem without consideration of the different quality of problems arising at different levels and in different subsystems of society. A first desideratum is therefore that these differences be adequately acknowledged in attempts at theorizing about the concept: social justice in the design of a society's basic structure may not mean the same as social justice in the determination of wage levels by firms or in the distribution of household

duties. To the extent that there are overlaps between the term's meaning and links between single justice problems and/or levels of societal aggregation, they must be carefully examined and extracted. But if the term social is to have any serious meaning at all, then it is also important to develop a clear sense of the types of problems it is to be applied to, of their differences, their relative interdependence and independence. Or, to put it differently, what is needed is greater clarity about the locus of regulation of different categories of problems. Some such problems need to be regulated at the macro level, others are better left to lower-level instances. And it is quite likely that even a well-ordered basic structure will leave many problems of social justice unresolved and hence cannot prevent all injustices in the spheres for which it provides no determinate answers. If this is the case, then it is all the more important that the theoretical discussion be more focused and less diffuse.

Second, there is a need for more cross-disciplinary exchange, allowing the empirical and normative approaches to enrich each other reciprocally (see Miller 1999, Schmidt 1994, 1999). The philosophical literature needs an improved understanding of the real world because elegance is no guarantee for relevance; in fact, it frequently is an impediment because simplification can easily result in simplistic ideas about real world problems and their suitable regulation. The social science literature, on the other hand, needs more terminological, methodological and conceptual sophistication, and a better understanding of the categorical and epistemological differences between 'is' and 'ought' questions, in order to find its proper place in the analysis of normativity. A familiarization with the pertinent philosophical literature can be of great benefit in this regard.

See also: Deprivation: Relative; Institutions; Legitimacy; Legitimacy, Sociology of; Moral Reasoning in Psychology; Moral Sentiments in Society; Norms; Values, Sociology of

Bibliography

Alves W, Rossi P 1978 Who should get what? Fairness judgments of the distribution of earnings. *American Journal of Sociology* **84**: 541–64

Barry B 1995 *Justice as Impartiality*. Clarendon Press, Oxford, UK

Boudon R 1997 The moral sense. *International Sociology* **12**: 5–24

Boudon R 1999 Multiculturalism and value relativism. In: Honneger C, Hradil S, Traxler F (eds.) *Grenzenlose Gesellschaft?* Leske and Budrich, Opladen, Germany, pp. 97–118

Boudon R, Betton E 1999 Explaining the feelings of justice. *Ethical Theory and Moral Practice* **2**: 365–98

Dworkin R 1981a What is equality? Part 1: Equality of welfare. *Philosophy and Public Affairs* **10**: 185–246

Dworkin R 1981b What is equality? Part 2: Equality of resources. *Philosophy and Public Affairs* **10**: 283–345

Elster J 1992 *Local Justice. How Institutions Allocate Scarce Goods and Necessary Burdens*. Russell Sage Foundation, New York

Elster J 1995 (ed.) *Local Justice in America*. Russell Sage Foundation, New York

Habermas J 1983 *Moralbewußtsein und kommunikatives Handeln*. Suhrkamp, Frankfurt, Germany

Hirsch F 1977 *The Social Limits to Growth*. Routledge & Kegan Paul, London

Kluegel J R, Smith E R 1986 *Beliefs About Inequality. Americans' Beliefs of What Is and What Ought to Be*. Aldine de Gruyter, New York

Kluegel J R, Mason B, Wegener B (eds.) 1995 *Social Justice and Political Change: Public Opinion in Capitalist and Post-communist States*. Aldine de Gruyter, New York

Luhmann N 1980 *Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie der modernen Gesellschaft*. Suhrkamp, Frankfurt, Germany

Miller D 1976 *Social Justice*. Clarendon Press, Oxford, UK

Miller D 1999 *Principles of Social Justice*. Harvard University Press, Cambridge, MA

Okin S 1989 *Justice, Gender and the Family*. Basic Books, New York

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Röhl K F, Machura S (eds.) 1997 *Procedural Justice*. Ashgate, Aldershot, UK

Schmidt V H 1994 Bounded Justice. *Social Science Information* **33**: 305–33

Schmidt V H (eds.) 1999 *Justice in Philosophy and Social Science*. Special issue of *Ethical Theory and Moral Practice*. Kluwer, Dordrecht, The Netherlands

Schmidt V H, Hartmann B K 1997 *Lokale Gerechtigkeit in Deutschland. Studien zur Verteilung knapper Bildungs-, Arbeits- und Gesundheitsgüter*. Westdeutscher Verlag, Opladen, Germany

Sen A 1992 *Inequality Re-examined*. Russell Sage Foundation, New York

Walzer M 1983 *Spheres of Justice. A Defense of Pluralism and Equality*. Basic Books, New York

V. H. Schmidt

Social Learning, Cognition, and Personality Development

Social learning theory views the course of human development on the basis of children's socialization experiences and acquisition of self-regulation. Children's development of personality characteristics, such as dependency and aggression, as well as their skill in academics, sports, arts, or professions, are assumed to emerge from learning experiences embedded within the social milieu of their family, peers, gender, and culture. This article describes a social learning perspective and describes the intellectual context from which it emerged, changes in researchers' emphasis over time, and current emphases, methodological issues related to the construct, and directions for future research.

1. Defining a Social Learning Perspective

Social learning theory defines children's socialization in terms of specific social learning experiences, such as modeling, tuition, and reinforcement, and the cognitions, emotions, and behavior that emerge from these formative experiences. Children's personal reactions were explained historically using personality constructs, such as identification, conditioning, or drive reduction, but are explained contemporaneously in cognitive terms, such as personal agency beliefs and self-regulatory processes. Social learning experiences are responsible for children's development of self-regulatory competence to adapt to changing settings, including those associated with their age or level of physical development. Self-regulation is essential to children's development because socialization involves giving up immediately pleasurable activities or familiar methods of coping to achieve delayed benefits.

2. Intellectual Context for Social Learning Research

Social learning research grew initially out of a seminar offered at Yale University during the late 1930s by Clark Hull. Among the members of that historic class of students were John Dollard, Neal Miller, Robert Sears, and John Whiting. The goal of the seminar was to provide learning explanations for key aspects of children's personality development that had been discussed by Freud, such as dependency, aggression, identification, conscience formation, and defense mechanisms. Of particular interest to class members was the psychoanalytic notion of children's identification with powerful models, such as parents or other adults, which Freud used to explain a wide range of personality characteristics, such as, dependency, aggression, and even psychosexual development. Identification was singled out by Miller and Dollard (1941) in a series of experimental studies of imitative learning. They found that a general tendency to imitate could be learned through reinforcement and that rewards could be structured to teach counter-imitation (i.e., the opposite choice of a model) as well as imitation. They concluded that imitation was a learned class of instrumental behavior but was not a learning process in its own right. Sears studied the impact of parents and their relationship with their children on the youngsters' personality development, especially socialization processes that could explain how children internalize the values, attitudes, and behavior of the culture in which they were raised. According to Sears' identification theory, social learning experiences lead to the internalization of self-sustaining drives to emulate their parents and other adult figures. Sears was particularly interested in children's development of self-control over aggressive behavior. Whiting conducted cross-cultural studies of differences in children's socialization to determine whether the severity of these experiences could lead to cultural differences in anxiety, guilt, and the formation of conscience.

3. Changes in Emphasis Over Time

Albert Bandura's first book with Richard Walters on adolescent aggression tested Sears' identification theory as a basis for internalization of controls over children's aggressive behavior. This research ultimately led these researchers to reject the psychoanalytic theoretical basis for social learning in a second book (Bandura and Walters 1963). They disproved the Freudian hypothesis that modeled violence would reduce observers' aggressive drives through identification and catharsis by demonstrating that such modeling actually taught or facilitated aggression, especially when modeling was accompanied by vicarious rewards. In other research, Bandura and his colleagues showed that through modeling and vicarious reward experiences children could learn new responses without actually performing them or receiving rewards. This disproved earlier instrumental conditioning accounts of modeling and imitation, and led Bandura to distinguish the cognitive effects of learning due to modeling from the motivational effects of rewards on imitative performance. To explain these two aspects of imitative learning, Bandura distinguished four subprocesses: two that affect vicarious learning—attention and retention—and two that affect imitative performance—production and motivation. To learn a new skill socially, a learner must attend to and encode a model's actions, and to perform a skill imitatively, a learner must be motivated and motorically able to produce the response. Bandura used the label *observational learning* to describe these distinctive influences of modeling without reference to Freudian identification, Hullian acquired drives, or instrumental conditioning.

During the late 1960s and early 1970s, social learning researchers began to study vicarious learning of abstract properties of a model's performance, such as grammar or conceptual classes underlying spoken language. They discovered that children could acquire a variety of abstract rules with minimal mimicry of a model's specific behaviors, such as imitating a model's style of questioning but not the model's exact words (Rosenthal et al. 1970). These demonstrations of abstract modeling on children's concepts, rules, and reasoning attracted adherents looking for alternatives to stage views of children's development. Many learning researchers, including those interested in social learning processes, questioned the validity of claims of an invariant sequence of qualitative stages that are resistant to direct training and are transituational in their impact. Instead they sought to demonstrate that children could be taught advanced

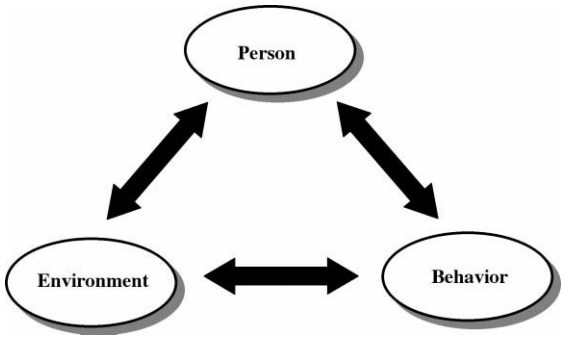

Figure 1
Reciprocal causation among triadic classes of social learning determinants

stage Piagetian concepts, Chomskian grammatical rules, and Kohlbergian or Piagetian moral judgments and that task and situational variations could cause cross-stage variations in these measures. Mary Harris, Robert Liebert, Ted Rosenthal, James Sherman, Grover Whitehurst, and Barry Zimmerman, among many others, used abstract modeling techniques to teach advanced stage functioning among children of a variety of ages, and the results were chronicled in the book on social learning and cognition (Rosenthal and Zimmerman 1978). Although age-related shifts in children's functioning were of interest to social learning researchers, they sought to explain these outcomes more on the basis of changes in social learning experience, hierarchies of goals and knowledge, and level of motoric competence rather than simply cognitive growth and development. From a social learning perspective, shifts in children's personal functioning at the approximate ages of 2, 7, and 12 years by numerous stage theories are better predicted by social learning experiences and accomplishments, such as the acquisition of speech and mobility, entrance into school, and the onset of puberty. When considering later lifespan developmental issues, social learning researchers have focused on adult life-path experiences associated with employment, marriage, child rearing, and physical declines in functioning rather than cognitive stage indices.

During the late 1970s, social learning researchers began to shift their interest from children's initial acquisition of important personal skills via modeling to the other end of the developmental spectrum—the point at which they can regulate these skills on their own. To explain children's development and use of self-regulation, Bandura (1977) depicted social learning experiences in triadic terms involving reciprocal causal relations between personal (cognitive-affective), behavioral, and environmental sources of influence (see Fig. 1). Each of the three components of his theory influences the other components bidirectionally. Cognitive events, although central to learning and performance, were constrained by specific features of social environmental contexts and by specific performance outcomes. The close bidirectional linkage of cognitive-affective processes to performance outcomes and situational contexts enabled Bandura's triadic model to explain dynamic features of children's development, such as variations in children's aggression across settings, as well as enduring individual differences. Walter Mischel also used a triadic social learning formulation to explain children's development of self-regulatory capability to delay gratification. He found that both modeling experiences and cognitive encoding enhanced children's willingness to defer a small reward for a later larger one. Michael Mahoney and Carl Thoresen studied adults' self-control of a range of clinical problems triadically in terms of covert personal forms of self-control as well as environmental and behavior forms. Bandura (1977) undertook a series of clinical training studies using highly effective (or mastery) models to improve the self-control of patients plagued with disabling fears, such as snake and dog phobias. These modeling studies were designed to enable patients' to overcome their phobia by showing the effectiveness of specific handling techniques with the feared animal. Bandura discovered that although these patients were confident about the effectiveness of a model's techniques to handle the feared animal, they were not confident about their own ability to learn and use those techniques, and this cognitive and affective deficit greatly affected their willingness to change. Bandura defined these personal perceptions of capability to learn or perform as self-efficacy beliefs.

4. *Emphases in Current Theory and Research*

Increasing evidence of the importance of cognitive aspects of self-regulation, especially self-efficacy beliefs, led Bandura to rename his approach as social cognitive theory (Bandura 1986). His book inspired numerous researchers, such as Gail Hackett, Robert Lent, Alan Marlatt, Ann O'Leary, Frank Pajares, Dale Schunk, Ralf Swartzer, Anita Woolfolk Hoy, Robert Wood, and Barry Zimmerman, among others, to begin studying the role of self-efficacy and self-regulatory processes in a wide variety of areas, such as academic functioning, physical health and medicine, sports, business and management, mental health, and the arts. A social cognitive model was used very extensively to explain how students acquire the sense of self-efficacy and the self-regulatory skill to assume responsibility for their academic learning (Schunk and Zimmerman 1994). Much of this research focused on how modeling experiences could lead school-aged children to acquire self-regulatory control of their academic performance and pursue desired careers.

During the 1990s, Bandura undertook a series of studies of the psychosocial influences of families, peers,

and schools on children's development of self-efficacy to self-regulate their academic and social attainments. In research with Claudio Barbarelli, Gian Vittorio Caprara, Manuel Martinez-Pons, Concetta Pastorelli, and Barry Zimmerman, Bandura found that self-efficacy beliefs and goals of parents for their children significantly affected the youngster's self-efficacy beliefs, aspirations, level of depression, and adherence to moral codes of conduct. Children's self-efficacy was assessed in a variety of areas of functioning, such as perceived social efficacy and efficacy to manage peer pressure for detrimental conduct, both of which were found to contribute to the youngsters' academic attainments. In 1997, Bandura published a 20-year review of research on self-efficacy involving hundreds of studies that spanned such tasks as mental and physical health, sport, academics, and business and management (Bandura 1997). Self-efficacy proved to be highly predictive of such motivational measures as free choice, persistence, and effort, of measures of self-regulation, and of social and academic learning outcomes.

5. *Methodological Issues*

In order to capture the triadic interdependence of person-related processes, such as children's self-efficacy beliefs, with their behavior and environmental setting, Bandura has advocated situationally specific forms of assessment and microanalyses of cognitive and behavioral functioning. By assessing children's, peers,' and adults' functioning in this specific manner before, during, and after efforts to perform behaviorally, researchers can test causal relations among social and self-regulatory processes. There is a body of evidence indicating that task-specific measures, such as self-efficacy judgments, are more predictive of actual performance efforts and outcomes than global trait measures of self-belief (Bandura 1997). The greater predictive power of task-specific measures was also reported by Mischel (1968) when he reviewed the literature on a wide range of personality traits, such as friendliness or conscientiousness. He found higher correlations when situationally specific measures than when trait measures of personal processes were employed, and he concluded that context-free measures of personal belief or competence were unable to explain variations in behavioral functioning across contexts.

6. *Future Directions for Research and Theory*

To explain better the sequential cognitive and behavioral aspects of self-regulation, Zimmerman (1998) developed a three-phase, cyclical self-regulation model. Forethought phase processes anticipate efforts to learn and include self-motivational beliefs such as self-efficacy. Performance phase processes seek to optimize learning efforts and include the use of imagery, self-verbalization, and self-observation. Self-reflection processes follow efforts to learn and provide understanding of the personal implications of outcomes. They include self-judgment processes such as attributions and self-reaction processes such as self-satisfaction, defensiveness, or adaptation. These self-reactions are hypothesized to influence forethought processes cyclically regarding future efforts to perform. It is also expected that as children develop skill in an area of functioning, they will show stronger cyclical links among these self-regulatory processes.

Schunk and Zimmerman (1997) identified four successive social learning levels associated with children's development of a particular skill, such as writing or golf. A first level, observational learning, occurs when a learner induces the major features of skill or strategy (for complex skills) from watching a model learn or perform, whereas a second level, emulative learning, is attained when a learner's motoric performance approximates the general strategic form of a model's. Emulative accuracy can be improved further when a model assumes a teaching role and provides guidance, feedback, and social reinforcement during performance efforts. A third level, self-controlled learning, is attained when a learner masters the use of a skill or strategy in a structured setting outside the presence of a model by relying on a previously abstracted standard from a model's performance. A fourth level, self-regulated learning, is attained when learners can systematically adapt their performance to changing personal and contextual conditions without reference to a model. It is not assumed that this is a stage model; thus, learners do not have to advance through the four levels in sequence or that the highest level of self-regulation will be used continuously without regard to motivation or context. Rather, it is hypothesized that students who master skills or subskills (in the case of complex tasks) according to this social learning sequence will display greater motivation and achievement than students using self-discovery methods. This multilevel formulation integrates the findings of modeling research with those of self-regulation research within a sequential model of children's socialization.

See also: Family Bargaining; Infancy and Childhood: Emotional Development; Personality and Marriage; Self-regulation in Childhood; Social Cognition in Childhood; Social Competence: Childhood and Adolescence; Socialization and Education: Theoretical Perspectives

Bibliography

Bandura A 1977 *Social Learning Theory*. Prentice-Hall, Englewood Cliffs, NJ

Bandura A 1986 *Social Foundations of Thought and Action: a Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ
Bandura A 1997 *Self-efficacy: the Exercise of Control*. Freeman, New York
Bandura A, Walters R H 1963 *Social Learning and Personality Development*. Holt, Rinehart and Winston, New York
Miller N, Dollard J 1941 *Social Learning and Imitation*. Yale University Press, New Haven, CT
Mischel W 1968 *Personality and Assessment*. Wiley, New York
Rosenthal T L, Zimmerman B J 1978 *Social Learning and Cognition*. Academic Press, New York
Rosenthal T L, Zimmerman B J, Durning K 1970 Observationally induced changes in children's interrogative classes. *Journal of Personality and Social Psychology*. **16**: 631–88
Schunk D H, Zimmerman B J (eds.) 1994 *Self-regulation of Learning and Performance: Issues and Educational Applications*. Erlbaum, Hillsdale, NJ
Schunk D H, Zimmerman B J 1997 Social origins of self-regulatory competence. *Educational Psychologist* **32**: 195–208
Zimmerman B J 1998 Developing self-fulfilling cycles of academic regulation: an analysis of exemplary instructional models. In: Schunk D H, Zimmerman B J (eds.) *Self-regulated Learning: from Teaching to Self-reflective Practice*. Guilford Press, New York pp. 1–19

B. J. Zimmerman

Social Mobility, History of

1. General Questions

Social mobility was one of the important themes of social history during its beginnings in the 1960s and the 1970s. Historians explore social mobility for two reasons. First, they want to study the general question of the history of equality of social opportunities. The question whether modern societies enlarged or reduced chances of social ascent for men as well as for women was attractive for historians. The rising openness or reinforced exclusiveness of modern elites, the rising or declining opportunities of upward social mobility for men and women from the lower classes, immigrants, poor families, and ethnic groups, the broadening or reduced access to channels of social ascent such as education, careers in business or public bureaucracies, and politics, and the role of family networks were central topics. Second, historians discuss the history of social mobility in a comparative view. European and American historians explore the myth of the opportunities for social ascent in America, which were unique compared to Europe. Historians also investigated the myth of unblocked social mobility in communist countries before 1989/91 compared to Western Europe. Social mobility was a historical theme to which major contributions did not only come from historians, but also from sociologists and political scientists.

2. Definitions and Sources

Studying the history of social mobility, historians usually investigate only the social mobility of individuals. They usually do not explore under the heading of social mobility the grading up or grading down of entire social groups or entire social classes. Moreover, the study of social mobility primarily focuses upon mobility in social structures and hierarchies rather than upon the geographical mobility of individuals as the term might suggest. It also mostly concentrates upon occupational mobility. If investigating career mobility (or intra-generational mobility), historians trace the mobility of individuals between different occupational positions or their persistence in the same occupation throughout the life. If they investigate mobility between generations (or intergenerational mobility), they compare the occupation of the father, the mother or the ancestors at specific points of their life with the occupations of a historical individual. Occupation is usually seen as the crucial indicator of the situation of an individual in a historical society. To be sure, historians are fully aware of the fact that the activity of an individual in history, more often than today, might comprise a variety of occupations at the same point of time or might include professions that still are in the making. Historians of social mobility work to link various historical sources and to trace individual persons through marriage license files, census materials, tax files, last wills, records of churches and public administrations, and autobiographical sources. The competence of historians in linking various sources has grown distinctly.

During recent years, the study of social mobility has become highly sophisticated. Individual careers are more frequently explored in micro studies of as many details as possible. These studies investigate a few richly documented individual cases rather than all members of a local society. They often use autobiographical materials difficult to analyze quantitatively. This type of micro study rarely concentrates only on social mobility, but covers a large variety of aspects. At the same time, the international and interregional comparative study of the history of social mobility has become somewhat more frequent, using the rich results of about 30 years of historical research in this field.

3. Three Main Debates

A first debate on social mobility by historians covers the increase of social mobility from 1800–2000. In this debate the rise of social mobility has various meanings. It sometimes means a more meritocratic recruitment, especially for the few most prestigious, most powerful, and best paid positions. It sometimes means more

mobility between occupations, upward mobility as well as downward mobility, and job mobility in the same occupation as well as mobility between occupations in the same social class. Sometimes, the increase of social mobility includes the chances of both genders and of minorities. Sometimes it means a clear increase of the opportunities of the lower classes compared to the opportunities of the upper and middle classes and not just more social mobility for everybody. One has to make sure what meaning is used by individual authors.

The advocates of the rise of social mobility usually think of a general increase of mobile people, but often also of a rising number of upwardly mobile persons since industrialization. They argue that various major social changes should have led to more social mobility and to more social ascent: the general decline of the fertility rate during the late nineteenth and early twentieth centuries made it possible for parents not only to invest more in individual help and education of their children, but also to promote their own professional careers. The rapid expansion of secondary and higher education especially since the end of the nineteenth century enlarged the chances for better training enormously. The rapid increase of geographical mobility since the second half of the nineteenth century led to a widening of the labor market and to a greater variety of new chances. The fundamental changes of the active population from the predominance of agrarian work up to the nineteenth and early twentieth centuries to the predominance of service work especially since the 1970s generated substantial social mobility between occupations. The distinct increase of the sheer number of occupations in all modern societies since the Industrial Revolution also must have led to more social mobility. The general change of mentalities, the weakening of the emotional identification with specific professions, social milieus, and specific local milieus, and the rising readiness for job mobility and for lifelong training further enlarged the number of socially mobile persons. The rise of the welfare state, the mitigation of individual life crisis, and the guarantee of individual social security clearly improved the chances for further training and for the purposeful use of occupational chances. Deliberate government policies of enhancing educational and occupational opportunities for the lower classes, for women, ethnic and religious minorities, and immigrants should also have had an impact on social mobility. In sum, a substantial list of factors in favor of an increase of social mobility from 1800–2000 can be put forward from this side of the debate.

The advocates of stability or even the decline in social mobility are a heterogeneous group. Arguments stem from very different ideas of social developments. It is sometimes argued that nineteenth and early twentieth century industrialization not only led to a rising number and a fundamental change of occupations, but also to a class society in which the major social classes, the middle class, the lower middle class, the working class, the peasants, and in some societies also the aristocracy tended to reinforce the demarcation lines to other social classes and, hence, to reduce rather than to enlarge the number of mobile persons. Other advocates of the skeptical view argue that the fundamental upheaval of modern societies during industrialization led to a unique rise of social mobility, of upward as well as downward mobility, and that modern societies thereafter became more closed since the generation of pioneers in business ended, as most occupational careers became more formalized and more dependent on formal education, as modern bureaucracies emerged and as mentalities adapted to the modern, highly regulated job markets. Other advocates argue for the stability of social mobility rates in a different and much more narrow sense. They argue that the long-term change of social mobility from the Industrial Revolution until the present was mostly structural, that is, it depended almost exclusively upon the redefinition of the active population rather than on the reduction of social, cultural, and political barriers. In this view social mobility remained stable if one abstracts from the changes simply induced by alternations in occupational structure (as peasants, for example, became workers, a real change, but not necessarily a case of upward mobility). Still other advocates of the long-term stability of social mobility posit a stable inequality of educational and occupational chances of lower classes, women, and minorities in comparison with the educational and occupational chances of the middle and upper class, the male population, or the ethnic majority, respectively.

This long debate has led since the beginning of quantitative studies of social mobility after World War II to a large number of historical studies of social mobility and to a wide range of results. To sum up briefly, three main results ought to be mentioned. First, only in very rare cases could a clear decline of social mobility rates to be found. Most studies show either stable or increasing rates of social mobility, depending upon the type of community and country and the generation and the period under investigation. There is no overwhelming overall evidence, either for the stability or for an increase, of social mobility rates. Second, changes of overall social mobility rates do in fact depend to a large degree on changes in occupational and educational structure. So one can say that modern societies became more mobile to a large degree because education expanded so much and because occupational change became so frequent and normal. Finally, there is much evidence that the educational and social mobility of the lower classes and women did not increase to the detriment of the educational and occupational chances of the middle and upper classes and men. Except for the Eastern European countries in some specific periods, social mobility was usually not a zero numbers game.

A second debate covers the more advanced social mobility in the United States. This old debate dates at least from the early nineteenth century, when the French social scientist Alexis de Tocqueville argued that American society offered more opportunities of social ascent than did European.

Various pieces of evidence was presented in favor of the American lead. Some empirical studies by sociologists demonstrated that in some crucial aspects a clear American lead could be shown. This was especially true for mobility into the professions. Higher education was more extensive and offered more chances than in Europe. Hence, the social ascent from the lower classes into the professions that are based on higher education was clearly more frequent than in Europe. In addition, comparative historical studies on the late nineteenth and early twentieth century American and European cities showed that in a special sense a modest American lead existed during that period: unskilled workers in fact moved up into white-collar positions in American cities somewhat more frequently than in European cities. Finally, historians demonstrated that the important difference between American and European societies could be found in the American creed in more opportunities in America than in closed European societies.

Other recent comparative studies tend to argue that a general lead of the American society no longer exists in comparison with Europe. According to this view industrialization and modernization led to about the same opportunities everywhere. Various international comparisons of social mobility rates support this view and demonstrate that the overall rates were not distinctly higher in the US than in old societies such as Western Europe or Japan. As these comparisons almost exclusively deal with the second half of the twentieth century, the results might be due to the fundamental social changes in Europe and Japan since World War II.

The lead of communist countries in social opportunities especially in the USSR during the 1920s and 1930s and in the Eastern European communist countries in the late 1940s and 1950s is the topic of a less intensive, but important debate. Some historical studies of social mobility demonstrate that during these periods rates of upward social mobility into the higher ranks of the social hierarchy were substantial compared to Western European societies. This was true partly because of the rapid expansion of higher education, the communist abolition of the middle class, and the rapid change of the employment structure due to rapid industrialization. However, the rise of social opportunities in communist countries, if it rose at all, was largely limited to the period of the system upheaval. Most comparative studies of the 1970s and 1980s show that rates of social mobility were not distinctly higher in Eastern Europe compared to Western Europe. This was partly because the communist political and administrative elite became exclusive and gentrified, because social change slowed down, and because the expansion of higher education was reduced in several communist countries.

Gender contrasts in social mobility were explored only by a small number of studies by historians. A debate among historians has not yet started. But gender contrasts undoubtedly will add new important aspects to the general debate about long-term trends in social mobility. Four conclusions can be expected from the few studies. First, in a more radical sense than in the study of male mobility, female mobility raises the question whether social mobility should in fact be centered around occupational mobility or other factors such as marriage and unpaid or partially paid work in emerging professions are to be taken in account much more than so far. Second, the question of a rise of downward mobility during the transition to modern society, i.e., during the rise of female activity outside the family sphere, is to be explored. A study of female social mobility in twentieth century Berlin demonstrates a high rate of intergenerational declassement of active women during the early parts of the twentieth century. Third, the study of the social mobility of women demonstrates much more clearly than that of males the effects of economic crises and fundamental transitions on social mobility. Opportunities for women seem to have depended strongly on economic prosperity, on long-term social stability. In periods of economic crisis and rapid transitions such as the upheaval of 1989/91 women more than men belonged to the losers. Here again the study of female mobility might draw the attention of the historians to a more general aspect of mobility that was not investigated enough. Finally, the social mobility of women also demonstrates that definite changes of social opportunities can only be reached in a long-term perspective. It was shown that even when important channels of upward social mobility such as education offered equal chances to women, they did not lead to a parallel improvement of occupational chances for women. Besides the study of institutions and policies, the historical study of the experience of social mobility and the perception of social mobility will become crucial.

4. *The Future Historical Research on Social Mobility*

During 1980–2000 social mobility was much less frequently investigated by historians. The declining interest in social mobility has partly to do with the rising research standards that are more and more difficult to match by projects limited in time and financial resources. Moreover, the initial questions asked in this field lost the former attraction because it became widely accepted that social mobility rates were about the same in most societies and did not rise distinctly during in the nineteenth and twentieth

industrialization and modernization centuries. Finally, the general thematic trends in historical research made social mobility look unlike a modern theme.

The future of the study of social mobility is that of a normal theme among many others in history rather than a top theme in an expanding branch of history as in the 1960s and 1970s. One can hope for four sorts of studies of neglected aspects of social mobility. A first future aspect is the historical study of social mobility in Eastern Europe, but also outside Europe, and thereafter the international comparisons reaching beyond Western Europe and the US. The question of the particularities of social mobility in Europe might then be answered differently. A second future aspect of social mobility is gender contrasts of social mobility. We need a well-reflected number of case studies of contrasting countries, different female activities, contrasting general conditions such as prosperity and economic depression, peace and war periods, stability and transitions. A third future theme is specific factors of social mobility such as religion, types of family, immigration, unemployment, and poverty and more generally the impact of background social milieu. It seems that historians will explore these contexts of social mobility often in looking at a limited number of individuals and that they will include also the subjective experience of social mobility. The fourth future topic is to be the history of the debate on social mobility, as an aspect of the history of identities, e.g., the European or American identity. The debate among historians was inspired by this debate and this topic will help us to understand better why historians became interested in this theme in the 1960s.

See also: Bourgeoisie/Middle Classes, History of; Class Consciousness; Class: Social; Elites: Sociological Aspects; Mobility: Social; Political Elites: Recruitment and Careers; Social Stratification; Working Classes, History of.

Bibliography

Boyle S C 1989 *Social Mobility in the United States, Historiography and Methods*. Garland, New York

Charle C 1987 *Les élites de la République (1880–1900)*. Fayard, Paris

Erickson R, Goldthorpe J H 1992 *The Constant Flux, A Study of Class Mobility in Industrial Societies*. Clarendon, Oxford, UK

Fitzpatrick 1979 *Education and Social Mobility in the Soviet Union 1921–1934*. Cambridge University Press, Cambridge, UK

Goldthorpe J H 1980 *Social Mobility and Class Structure in Modern Britain*. Clarendon, Oxford

Grusky D B, Hauser R M 1984 Comparative social mobility revisited: Models of convergence and divergence in 16 countries. *American Sociological Review* **49**: 19–38

Haller M (ed.) 1990 *Class Structure in Europe: New Findings from East–West Comparisons of Social Structure and Mobility*. Sharpe, Armonk, NY

Hout M 1989 *Following in Father's Footsteps, Social Mobility in Ireland*. Harvard University Press, Cambridge, MA

Kaelble H 1985 *Social Mobility in the 19th and 20th Centuries: Europe and America in Comparative Perspective*. Berg, Dover, NH

Kocka J, Ditt K, Mooser J, Reif H, Schüren R 1980 *Familie und soziale Plazierung*. Westdeutscher, Opladen, Germany

Kocka J, Ditt K, Mooser J, Reif H, Schüren R 1984 *Journal of Social History* 411–434

König W, Müller W 1986 Educational systems and labour markets as determinants of worklife mobility in France and West Germany: A comparison of men's career mobility, 1965–1970. *European Sociological Review* **2**: 73

Lipset S M, Bendix R 1959 *Social Mobility in Industrial Society*. University of California Press, Berkeley, CA

Mach B W, Mayer K U, Pohoski M 1994 Job changes in the Federal Republic of Germany and Poland: A longitudinal assessment of the impact of welfare-capitalist and state-socialist labor-market segmentation. *European Sociological Review* **10**

Miles A 1999 *Social Mobility in 19th and Early 20th Century England*. Saint Martin's, New York

Pinol J L 1991 *Les mobilités de la grande ville, Lyon fin XIXe-début XXe siècle*. Presses de la fondation nationale des sciences politiques, Paris

Sewell W H Jr 1985 *Structure and Mobility. The Men and Women of Marseille, 1820–1870*. Cambridge University Press, Cambridge, UK

Sorokin P A 1959 *Social and Cultural Mobility*. Free Press, Glencoe, IL

Thernstrom S 1973 *The Other Bostonians*. Harvard University Press, Cambridge, MA

Vincent D, Miles A (eds.) 1993 *Building European Society, Occupational Change and Social Mobility in Europe, 1840–1940*. Manchester University Press, Manchester, UK

H. Kaelble

Social Movements and Gender

Social movements can be defined as collectivities acting outside institutionalized channels to promote or resist change in an institution, society, or the world order. Although there are a diversity of theoretical approaches to the study of social movements, until recently mainstream theory and research in the fields of social movements and political sociology have ignored the influence of gender on social protest. Beginning in the 1990s, scholars trained in the gender tradition that flourished in the social sciences during the 1980s turned their attention to questions pertaining to the gender dynamics of a variety of historical and contemporary movements. As a result, a thriving body of research demonstrates that gender hierarchy is so persistent that, even in movements that do not evoke the language of gender conflict or explicitly embrace gender change, the mobilization, leadership patterns, strategies, ideologies, and outcomes of social movements are influenced by gender and gender differences. This article delineates the way gender relations and hierarchy influence various social movements, including feminist, antifeminist, men's, labor, civil rights,

gay and lesbian, nationalist, rightwing, and fascist movements. It ends by discussing the broader implications of integrating gender to mainstream theories of social movements.

1. Linking Gender and Social Movement Theories

Notions of femininity and masculinity, the gender division of labor in private and public spheres, and the numerous institutions, practices, and structures that reinforce male dominance are not expressions of natural differences between women and men. Rather, the gender order is constructed socially and gender is a pervasive feature of social life. Although gender scholars fairly universally adopt social constructionist perspectives to understand gender relations, there is considerable variation in the epistemologies, contexts, and levels of analyses of the various theories that have developed to explain gender hierarchy. The most sophisticated conceptualizations treat gender as an element of social relationships that, like race and class, operates at multiple levels. Gender categorizes and distinguishes the sexes and ranks men above women of the same race and class (Lorber 1994). Multilevel frameworks seek to avoid universal generalizations that are impervious to the historicity and contextualized nature of perceived differences between the sexes by conceptualizing gender as operating at three levels. Multilevel approaches, first, view gender as organizing social life at the interactional level through the processes of socialization and sex categorization that encourage gender-appropriate identities and behavior in everyday interactions. Second, multilevel theories emphasize the complex ways gender operates at the structural level, so that gender distinctions and hierarchy serve as a basis for socioeconomic arrangements, the division of labor in families, the organization of sexual expression and emotions, work and organizational hierarchies, and state policies. Finally, multilevel theories attend to the cultural level, specifically to the way gender distinctions and hierarchy are expressed in ideology and cultural practices, such as sexuality and language, as well as in the symbolic codes and practices of institutions such as art, religion, medicine, law, and the mass media.

Recent writings on social movements treat the social construction of gender as taking place at these three respective levels. The significance of gender is more apparent for some movements than for others. Its relevance is obvious in the case of women's movements targeting discriminatory policies and practices pertaining to employment, family, reproduction, and sexuality. Nevertheless, gender is also constructed in movements that do not explicitly evoke the language of gender conflict. One example is McAdam's (1992) landmark study of the US 1964 Freedom Summer campaign, when white college students spent the summer in the South participating in the US civil rights movement. The traditional sexual double standard meant that women were less likely to be recruited to the Mississippi project than men because of the taboos associated with interracial sexual relationships, especially those between Black men and white women. Those women who managed to make it to Mississippi were more likely to be slotted into traditional female roles—specifically, clerical work and teaching rather than canvassing and voter registration. Furthermore, the biographical consequences of participation differed for men and women. Men were more likely to remain activists over the course of their lives than women, and women who participated in Freedom Summer were less likely to marry than men. To provide another example, masculinity is integral to nationalist politics in the contemporary world (Nagel 1998). Gender hierarchy is expressed in forms such as the construction of patriotic manhood and exalted motherhood as icons of nationalist ideology. The relevance of gender can also be seen in the domination of masculine interests in the ideology of nationalist movements and sexualized militarism that simultaneously constructs the male enemy as over-sexed and under-sexed (as rapists and wimps) and the female enemy as promiscuous (sluts and whores).

A systematic incorporation of gender into an understanding of social movements requires that we examine the way gender processes are embedded in the factors that trigger and sustain social movements. The most influential theoretical perspectives on social movements combine the insights of classical collective behavior theory, resource mobilization theory, and new social movements theory (McAdam et al. 1996). They emphasize three factors that explain the emergence and development of any movement: shifting political and social opportunities and constraints; the forms of organization used by groups seeking to mobilize; and variations in the ways challenging groups frame and interpret their grievances and identities. Let us examine the way gender comes into play in each of these three broad sets of conditions that are relevant to the emergence and operation of social movements.

1.1 Political and Cultural Opportunities and the Gender Regime of Institutions

One of the major premises of social movement theory is that changes in the larger political and cultural context can stimulate or discourage people from participating in collective action. Shifts in gender differentiation and gender stratification in the particular sociohistorical context have been important facilitating conditions for the emergence and operation of feminist movements. In modern welfare states, public policy with respect to pregnancy and mother-

hood historically has been formulated either by treating women and men the same with gender-neutral policies or by recognizing women's special needs as child bearers and mothers and providing female-specific benefits. In the US, during the early twentieth century, government support of maternity often took the form of 'protective legislation' that regulated women's hours, working conditions, and prohibited women from working in occupations thought to be dangerous. While nineteenth century feminists, by and large, supported special female legislation and other policies consistent with women's primary identity as mothers, by the late 1960s the feminist campaign to formulate public policy switched to the sameness side of the sameness/difference divide, reflected in campaigns for policies such as the Equal Rights Amendment and the Family and Medical Leave Act.

Feminist movements in the US historically have targeted gender inequality in a wide range of institutions, including medicine, religion, education, politics and the law, the military, and the media. For instance, feminists criticized the medical system's bias against women in medical research, the labeling and classification of medical conditions, and doctor/patient relationships. The entry of larger numbers of women into the medical field beginning in the 1970s and 1980s facilitated the growth and expansion of feminist women's health centers and women's self-help movements focused on conditions unique to women, such as rape, battering, breast cancer, childbirth and pregnancy, and postpartum illness.

1.2 Gendered Mobilizing Structures

Expanding political and cultural opportunities create possibilities for collective action, but the emergence of a movement depends upon whether aggrieved groups are able to develop the organization and solidarity necessary to get a challenge underway. Gender differences and relations figure heavily into the pre-existing mobilizing structures or nonmovement ties that exist among individuals prior to mobilization and the organizational form a movement takes once it is underway. Gender, racial, class, ethnic, and other inequalities are often embedded in the informal networks, clubs, and formal organizations that provide the incentives for people to take the risks associated with participating in protest groups. In her study of African American women's participation in the US civil rights movement, Robnett (1996) identifies a distinct form of grassroots behind-the-scenes leadership carried out by women who were prevented from occupying formal leadership positions by the exclusionary practices of the Black church. Gender divisions can also facilitate mobilization. For example, sex-segregated women's colleges, clubs, associations, and informal networks historically have been fertile grounds for the growth of feminism. Blee's (1991) research on women in the 1920s Ku Klux Klan provides further evidence of the importance of sex-segregated networks. She reveals how women drew upon family ties and traditions of church suppers and family reunions to mobilize. She describes the way Klan women deployed women's gossip in the form of 'poison squads' to spread a message of racial and religious hatred.

The connective structures that link leaders and followers and different parts of a movement, allowing them to persist, can also be gendered. Some scholars have argued that there is no particular structure essential to a feminist organization. Others, however, find that some feminists' beliefs in fundamental differences between female and male values have led to feminist organizations with distinctive emotion cultures that dictate open displays of emotion, empathy, and attention to participants' biographies. By contrast, the large-scale all-male meetings of the Promise Keepers, a mostly white, middle-class, and heterosexual men's rights movement that arose in the late 1990s to reassert essentialist views of women's and men's roles in the family, were steeped in an aura of powerful masculinity, holding meetings in stadiums associated with organized sports events (Messner 1997). Gender can also be a subtext in the organizational logic, beliefs, and practices of movements that do not make explicit gender claims. Einwohner (1999) found that prevailing ideas about gender and gender differences are a driving force behind the largely female composition, emotion-laden strategies, and negative reactions of the public to the contemporary animal rights movement. The well known decentralized grassroots Alinsky model that influenced early social welfare activism since the early 1970s, according to Stall and Stoecker (1998), is organized along an essentially masculine logic that views community organizing as a means to power, in contrast with 'the women-centered model,' which treats community building as a goal of organizing. Brown (1992) implicates gendered organization in the decline of the Black Panther Party. She recounts the way the male power rituals of the Party—specifically its military structure, use of violence and aggression, and promotion of sexual rivalries—not only excluded women from the Black power struggle but planted the seeds of its destruction.

As this last example illustrates, gender inequality often persists within social movement organizations as Fonow (1998) reports in her research on women's participation in the 1985 Wheeling–Pittsburgh steel strike. In this case, major strike activities—picket duty and meal preparation—reflected a gender division of labor. While working kitchen duty helped mobilize the support of the larger community by bringing together the wives and families of workers and nonworkers, the exclusion of women steelworkers from the picket lines—seen as an expression of 'virile unionism'—

undermined solidarity between men and women steelworkers.

1.3 Gender and Framing Processes

Social movements often appropriate gender ideology to legitimate and inspire collective action and to identify a challenging group's commonalities. Contemporary scholars of social movements rely upon two concepts for analyzing the unique frames of understanding that people use to define their shared grievances: collective action frames and collective identities. To the extent that gender operates as a constitutive element of social interaction and relationships, it is not surprising that images of masculinity and femininity appear in the language and ideas social movement activists use to frame movement demands and injustices. Gender dualism, for example, was woven tightly into German National Socialism. Not only was masculinity glorified and male bonding considered the foundation of the Nazi state, but gender-linked oppositions served as the metaphor for the polarization of Jews as weak and emotional and Aryans as strong and rational. The language of gender difference and power is pervasive in contemporary women's self-help movements and serves as a major framework for understanding the problems that trouble women. In the postpartum self-help movement, gender plays a pivotal role in mobilizing participants. Activists linked postpartum illness to women's responsibilities for parenting, to ideas that mothers (not fathers) have to be self-sacrificing, and to the gender bias of the largely male medical establishment (Taylor 1999). Social movement scholars use the concept of collective identity to research the question of how challenging groups define and make sense of the question 'who we are' and the complex and changing process collective actors use to draw the circles that separate 'us from them.' Gender difference in virtually every culture is key to the way actors identify themselves as persons, organize social relations, and symbolize meaningful social events and processes. That means that the identities that social movements of all kinds create and recreate are bound to be molded, at least in part, by gender meanings and practices. For example, Taylor and Whittier (1992) argue that in the lesbian feminist movement of the 1970s and 1980s, the drawing of boundaries between male and female promoted the kind of oppositional consciousness necessary for organizing one's life around feminism. Similarly, some branches of the contemporary breast cancer movement frame women's commonalities in gender terms by connecting the experiences of breast cancer survivors to the societal emphasis on women's breasts as crucial to women's attractiveness to men. At the same time, such activists point out, the medical establishment has provided less funding and attention to breast cancer than to diseases that affect men more than women (Klawiter 1999). If gender symbolism plays an important role in the socially constructed solidarities that mobilize collective action, the development of gender solidarity is never automatic, even in women's movements. Analyzing African American women's autobiographies written before 1970, Brush (1999) shows that the messages that came from the civil rights movement encouraged them to interpret their grievances in terms of race rather than gender, while the women's movement made gender seem irrelevant to Black women's experiences. The tendency to see the generic 'Black' as male and the generic 'woman' as white shows that using gendered meanings to construct a group's commonalities can exclude some people, since deciding who 'we' are also entails deciding who 'we' are not (Collins 1990).

Gender as a set of cultural beliefs and practices figures heavily into many of the identity movements that have swept the political landscape in recent years. In attempting to privilege white racial identity, the contemporary white supremacist movement not only presents a range of racial enemies but deploys gender metaphors—e.g., treating the Black male body as dangerous and animal-like and the white female body as always at risk and in need of protection—to appeal to the growing number of disillusioned white males who feel threatened by what they perceive to be the successes of social movements for gender and racial equality (Ferber 1998). Finally, gender symbolism and ideas about masculinity and femininity figure heavily in the modern gay and lesbian movement, which challenges the notion that women are feminine and men masculine and, further, that erotic attraction must be between two people of different sexes and genders (Gamson 1997).

2. The Significance of Integrating the Study of Social Movements and Gender

Feminist scholars interested in the study of social movements recently have called attention to the gendering of social movement processes and social movement theory. The role of gender in the emergence and dynamics of social movements, even those seemingly not about gender, has been obscured through the gender-neutral discourse that characterizes prevailing theories of social movements. Treating gender as an analytic category illuminates a range of new questions pertaining to the political opportunities, mobilizing structures, and framing processes of movements. A synthesis of the theoretical literatures on gender and social movements also has the potential to advance the field of gender. Theories of gender emphasize the maintenance and reproduction of gender inequality and neglect countervailing processes of resistance, challenge, conflict, and change. The literature on social movements spotlights the role of human agency in the

making and remaking of femininity, masculinity, and the existing gender system.

See also: Collective Behavior, Sociology of; Feminist Movements; Gender and Feminist Studies in Sociology; Social Movements, History of: General; Social Protests, History of

Bibliography

Blee K M 1991 *Women of the Klan: Racism and Gender in the 1920s*. University of California Press, Berkeley, CA

Brown E 1992 *A Taste of Power: A Black Woman's Story*. Pantheon Books, New York

Brush P S 1999 The influence of social movements on articulations of race and gender in Black women's autobiographies. *Gender & Society* **13**: 120–37

Collins P H 1990 *Black Feminist Thought*. Unwin Hyman, Boston

Einwohner R L 1999 Gender, class, and social movement outcomes: Identity and effectiveness in two animal rights campaigns. *Gender & Society* **13**: 56–76

Ferber A L 1998 *White Man Falling: Race, Gender, and White Supremacy*. Rowman and Littlefield, Lanham, MD

Fonow M M 1998 Protest engendered: The participation of women steelworkers in the Wheeling–Pittsburgh steel strike of 1985. *Gender & Society* **12**: 710–28

Gamson J 1997 Messages of exclusion: Gender, movements, and symbolic boundaries. *Gender & Society* **11**: 178–99

Klawiter M E 1999 Racing for the cure, walking women, and toxic touring: Mapping cultures of action with the bay area terrain of breast cancer. *Social Problems* **46**: 104–26

Lorber J 1994 *Paradoxes of Gender*. Yale University Press, New Haven, CT

McAdam D 1992 Gender as a mediator of the activist experience: the case of freedom summer. *American Journal of Sociology* **97**: 1211–40

McAdam D, McCarthy J D, Mayer N Z 1996 *Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings*. Cambridge University Press, New York

Messner M A 1997 *Politics of Masculinities: Men in Movements*. Sage, Thousand Oaks, CA

Nagel J 1998 Masculinity and nationalism: Gender and sexuality in the making of nations. *Journal of Ethnic Racial Studies* **21**: 242–69

Robnett B 1996 African–American women in the civil rights movement, 1954–1965: Gender, leadership, and micromobilization. *American Journal of Sociology* **101**: 1661–93

Stall S, Stoecker R 1998 Community organizing or organizing community? Gender and the crafts of empowerment. *Gender & Society* **12**: 759–56

Taylor V 1999 Gender and social movements: Gender processes in women's self-help movements. *Gender & Society* **13**: 8–33

Taylor V, Whittier N 1992 Collective identity and lesbian feminist mobilization. In: Morris A D, Mueller C M (eds.) *Frontiers in Social Movement Theory*. Yale University Press, New Haven, CT, pp. 104–29

V. Taylor

Social Movements: Environmental Movements

Since World War II environmental issues have increasingly entered Western political discourses. From the 1992 UN-conference in Rio de Janeiro onwards, environmental issues appeared on the government agendas of the South and of reforming countries and were promoted by international programs under the label of sustainable development. This contribution examines the ways in which environmental issues were taken up by non-formal collective political action—often referred to as the 'New Social Movements.' It looks at the informal actors participating in the movements and how they organize themselves and identify with their demands. Furthermore, it focuses on their strategies for engaging with power, which have redrawn the lines of what is regarded as 'political' and 'politics.'

1. Social Movements and New Social Movements: Theoretical Perspectives

Research on social movements—as the term is understood today—dates back to the 1930s when US sociologists, who were confronted with the turmoil of the emergence of fascism and communism in Europe, concentrated on ideological aspects of collective behavior, and perceived collective political action as a threat to the established order (Eyerman and Jamison 1991, p. 10). Herbert Blumer from the Chicago School was the first to outline the positive potential of the new forms of social interaction. In contrast to this social-psychological focus emphasizing the individual, Talcott Parsons took a structuralist-functionalist approach drawing on the classic work of Max Weber and Emile Durkheim. The combination of the two views resulted in a framework which has come to be known as the collective behavior approach, balancing the micro-oriented approach with the macro-oriented.

This perspective dominated research on social movements until the 1960s (Eyerman and Jamison 1991, p. 11f.). However, the civil rights movements and the students' demonstrations of the late 1960s led American sociologists to criticize the functionalist logic of the collective behavior theory, which devalued action as reactive behavior without strategic capability. US sociologists directed their attention instead toward the costs and benefits of taking part in social movements, grounding their concepts in actors as 'movement "entrepreneurs"' who make rational choices, using the material and symbolic resources available. Critics have pointed out that this emphasis on resources normally held by influential social groups leads to the destruction of the resistance due to marginalized actors. Moreover the exclusive focus on

rational choices does not take adequate account of emotions (Della Porta and Diani 1999, pp. 7–9).

As a response to this criticism the field of political research has directed its attention toward interactions between political establishment insiders and outsiders, and toward the outsiders' less conventional forms of action. This framework allowed social movements to be viewed as part of political debates and dialogs rather than as excluded from the political system. Despite their insights, however, analysts of political processes have failed to generate an explanation for the innovative nature of social movements and the new spaces created and occupied by contemporary movements of youth groups, of women, and of homosexual or minority ethnic groups (Della Porta and Diani 1999, p. 10). Della Porta and Diani draw upon concepts developed by European scholars, whose criticism of the deterministic Marxist models shed new light on the 'New Social Movements' in Europe. In particular, Alain Touraine's so-called post-industrial approach has become influential. It posits that the ruling and popular classes remain major antagonists and forces of social change. The New Social Movements are characterized by their decentralized and participatory organizational structures as well as their critical attitudes towards ideology and solidarity structures. As Alberto Melucci states, material claims are—in contrast to the workers' movement—no longer central to the Movements' agenda. Guaranties of security and state welfare are no longer demanded. Instead, the intrusion of the political-administrative system into daily life is something to be resisted (Della Porta and Diani 1999, p. 12f.).

Although the exchange between European and American scholars has intensified and research on social movements taken up by a variety of disciplines, Calderon criticizes the still predominant structuralist research paradigm, arguing that a structuralist view cannot provide sufficient insight into the social expression brought about by the movements (Calderon 1992, p. 23). Referring to Melucci, David Slater (1997, p. 259) uses the metaphor of the 'nomads of the present' for the post 1989 movements, capturing the fluidity and territorial flexibility of today's mobilization. Intensified cooperation between the disciplines and the theoretical schools has led to regarding manifestations of resistance and opposition as something increasingly independent from universalist discourses. Connected in terms of space, these new mobilizations stick together as 'archipelagos of resistance' characterized both by their global reach and their local embedding and specificities (Slater 1997, p. 259). This framework seems particularly helpful for understanding the emergence of social movements with an environmental focus. Before turning to environmental issues, however, it is necessary to consider the state's changing role in confronting collective political action.

Thirty years ago collective action was strongly oriented toward the state and state politics. Today, collective practices create their own cultural identities and specific spaces for social expression, taking advantage of and contributing to the severe crisis faced by nation states worldwide (Calderón et al. 1992, p. 23). Globalization processes are undermining what used to be the territorial state's significant spaces of action. Furthermore, the state is faced with the fragmentation of civil society: segments of society create their own spaces, often with a self-determined distance from the state which—in its modernist conception—has traditionally furthered integration by excluding certain social categories (Calderón et al. 1992, p. 25). Thus the state has lost its central role as the focus of attention for social movements, i.e., for collective action concerning ethnic or ecological questions. In general, both emerging forms of social and political participation as well as highly influential economic players are challenging the legitimacy of the state and its political representatives on environmental issues.

Taking his analysis of resistance even further, Steven Pile (1997, p. 3) criticizes social movements research that operates exclusively with the dual model of 'authority' on the one hand and 'resistance' on the other. He perceives resistant political subjects as also influenced by driving forces '…such as desire and anger, capacity and ability, happiness and fear, dreaming and forgetting,' which mobilize forces not simply based on the antagonism of established power and opposition. This analytical frame seems useful for understanding social movements organized around issues of identity and natural resources.

2. *Characteristics of Social Movements with an Environmental Focus*

The first of the New Social Movements started in 1968 with the student revolts in Western Europe. Independent of established political institutions, these new forms of social and political participation contrasted sharply with the traditional workers' movements. Starting out as student riots supported by feminist activists and intellectuals, this movement soon grew, also becoming more critical of transformations occurring in industrial society. The New Social Movements questioned modernism and economic progress, focusing on the costs of social transformation. In the 1970s these issues were extended to environmental conflicts in Western Europe. Protest groups not only questioned the values of industrial society and the price of such things as economic progress, security, and sustainability, but also the political process, its parties, and politicians. Touraine identified such central conflicts in post-industrial society as the struggle for the control of symbolic production.

In addition the organizational structures developed by the New Social Movements are different from those

of traditional political parties: they are more decentralized and follow participatory organizational principles. They usually involve people who have not been interested in traditional politics or who have not been integrated into traditional political parties before. In general these movements want to experience new forms of democracy and focus on environmental and non-profit issues.

Another characteristic feature of the New Social Movements—especially of those concerned with environmental issues—is the high involvement of women, minorities, young people, and people from the South. Seager (1996, p. 271) in her writing on 'hysterical housewives' points to the high percentage of women participating in environmentalist and conservationist organizations. Women around the world would, according to Seager, support more restrictive environmental laws and spend more money on ecological projects. However, they often face inflexible bureaucracies which adhere to fixed beliefs about economic growth and success. The potential of the politics of resistance lies in the interstices of self-determination, interpersonal and international solidarity as well as solidarity with future generations and the environment.

According to Della Porta and Diani (1999, p. 14f), the New Social Movements evince at least four characteristic features:

(a) Informal interaction networks. The movements can be regarded as informal interaction networks between a plurality of individuals, groups and/or organizations with loose and dispersed links. Such networks promote the circulation of essential resources for action (such as information or material resources), creating and preserving a system of meaning through adequate organizational resources. In many cases the networks intend to create a new model of democracy based on greater participation and decentralization than in traditional democracies.

(b) Shared beliefs and solidarity. Another characteristic feature of the New Social Movements is the set of beliefs and the sense of belonging they share. Social movements help to find new approaches to existing environmental conflicts, to raise new public issues and to create visions. By contributing to a symbolic redefinition of reality and possible future perspectives they construct new collective identities and value systems. In many cases it is up to the leaders of movements to create appropriate ideological representations.

(c) Collective action focusing on conflicts. The actors of social and environmental movements are usually engaged in political and cultural conflict. Very often they set their value system against that of a powerful establishment. The actors involved in the movements share common interests, values, collective experiences, and a feeling of belonging to certain places and environments.

(d) Use of protest. The movements make use of unusual patterns and deviating forms of political behavior and public protest, e.g., civil disobedience, collective action, and the mobilization of public opinion by spectacular activities (Della Porta and Diani 1999, p. 14f).

3. *New Actors and Alliances in the Political Struggle on Environmental Issues*

According to Calderón et al. (1992, p. 24) the reason for the emergence of social movements lies in the tension between social demands and the inability of the existing political institutions to hear these demands, not to mention give in to them. The protagonists of traditional social movements were rather large social entities easily identified by one common social denominator, e.g., their identity as workers or peasants. Unlike the protagonists of earlier social movements today's activists are not held together by a political force, i.e., by an aggregation of interests and actions. Although the New Social Movements are led by seemingly unimportant social actors, i.e., people in the lower ranks of the social hierarchy, such as black people, mothers, indigenous or disabled people, they leave a considerable mark on public platforms (Calderón et al. 1992, p. 26). Their highly specific interests and demands enable them to challenge the established structures more effectively. As Steven Pile (1997, p. 1) describes it, they perform 'acts of resistance [that] take place (…) in the spaces under the noses of the oppressor … .' This framework requires a more complex understanding of power and how it is distributed. The simple model of authority in the upper rank and its victims among the lower ranks has to be dropped.

In many of the New Social Movements feminist and ecological activists have joined together for collective action. There are several reasons for this (Rocheleau et al. 1996, p. 15ff): declining ecological and economic conditions which activate women because they threaten the survival of families; the impact of structural adjustment policies and the 'retreat of the state' from support of public services, social welfare, and environmental regulations; consciousness raising and political awareness; the political marginality of most women; the expansion of the women's movement to address converging issues of gender, race, class and culture. They focus on policy and environmental management issues, access to and distribution of resources under conditions of environmental decline, resource scarcity and political change, and environmental sustainability.

The unequal 'Geometries of Power' (Massey 1993) are responsible for generating social movements with an environmental focus and have led to a shift in scale between locally anchored movements and international organisations with global impact. In many

cases small indigenous groups succeeded by involving Non Governmental Organisations (NGOs) from Western countries in their struggles and were able to achieve their goals by globalizing the conflict. Some small and spontaneous local movements have expanded to become influential global players in environmental and human rights arenas. Several international NGOs, e.g., Friends of the Earth International, the WWF, Greenpeace, Global 2000, Peace Brigades International, have become global social movements with highly sophisticated organizational structures and professional fund raising.

In some cases local social movements with a specific environmental focus obtained wider social support and political influence, and were finally transformed into political parties (see below). Several Green Parties in Europe have developed along this line (Wastl-Walter 1996).

4. *Issues*

4.1 *Preservation of Natural Diversity*

Of the world's biodiversity, 98 percent exists on lands controlled by indigenous peoples. The land uses that sustain this diversity are intimately tied to the survival of the people in question, who often struggle to preserve their 'space' and lifeways against exogenous forces (Della Porta and Diani 1999, p. 12). Power relations are usually tilted against indigenous peoples, who have little political and social capital. These relations are further blurred by values that differ from those of the West. For some indigenous peoples, land is not only life, it is sacred. Hence many conflicts are also cultural in nature, and can only be won if indigenous groups obtain the support of Western social movements, which can contribute their strategies and resources for successful confrontation with global economic powers. A well known example is the conflict of interest dating from 1988 concerning the construction of the world's biggest dam on land belonging to the Kayapo Indians on the Rio Xingœ in Brazil. If the project were realised, 7.6 million ha would be flooded and 11 indigenous peoples would be forced to leave their land. Furthermore a unique ecosystem would be destroyed. The actors involved are the Government of Brazil, the World Bank, the Kayapo leaders and people, the OAB (Ordem de Abogados do Brasil), the NWF (National Wildlife Federation), the EDF (Environmental Defense Fund), Amnesty International, Survival International, Cultural Survival, and the International Society of Ethnobiology. The example shows that the actors involved range from local groups to groups with an international dimension. The groups differ vastly in terms of interests. Since various peoples would lose their home and hence their existence, human rights issues—and therefore the possibility of new alliances—should be considered as well.

4.2 *Preservation of Resources and the Right of Access to Environmental Resources*

A central aim of social movements with an environmental focus is the preservation of and right of access to environmental resources. Natural resources are endangered by the rich world's resource-intensive, wasteful, and polluting habits of production and consumption as well as by the environmentally destructive practices of the poor, who are forced to intensify their production under inappropriate conditions leading, for example, to depletion and degradation of large areas. Similar conflicts could be mentioned: in terms of land use, this means food crops vs. cash crops or sustainable farming vs. industrialization. One of the first famous examples of conflicts of this kind was the Kalinga women's struggle against the Chico Dam project (Philippines) from 1967 until 1987. The women linked the sociopolitical issue of emancipation with their ecological interests, questioning the right of access to environmental resources as such. The dam project, which was financed by the IWF/World bank, was their first that had to be cancelled due to the opposition of the indigenous population.

In Europe the first conflicts where social movements were concerned with environmental issues were the anti-nuclear power movements of the seventies. Again the main actors in terms of protest and the key speakers were women and mothers (*Mütter gegen Wackersdorf* in Germany). They were worried about the preservation of a healthy environment for their children and spoke against further economic and technological development at any price. They also questioned the legitimacy of technical experts.

5. *Methods and Strategies of the New Social and Environmental Movements*

These examples illustrate how the New Social Movements are forced to apply differing strategies and methods to succeed against the political establishment and the economic powers that be. Steven Pile draws our attention to the specific geographies around which acts of resistance unfold: 'on the streets, outside military bases, and so on, or, further, around specific geographical entities such as the nation, or 'our land,' or world peace, or the rainforests; or, over other kinds of geographies, such as riots in urban places or revolts by peasants in the countryside or jamming government web sites in cyberspace' (Pile 1997, p. 1).

The groups involved are often poorly resourced and act on a local level. In order to succeed they go public thus extending their reach, i.e., allying themselves in 'Joint Ventures' with global-players such as NGOs. For that reason, as Jayanta Bandyopadhyay states, social movements with an environmental focus cannot only develop fighting strategies; they also have to invest in public relations in order to become a trend-setting movement (Bandyopadhyay 1999, p. 880).

The movements are forced to use reassessed indigenous knowledge to overcome the arguments of modern scientific knowledge systems. And these alternative sources of knowledge need to be fed into channels of modern communication technology in order to reach a broad audience.

Local movements brought about by so called 'disposable people' (Ekins 1993) need to create a myth about themselves which triggers activists and solidarity movements among the more privileged. In many cases these groups have charismatic leaders or 'movement poets' (as they are called in the Chipko movement in India), who by personalizing the conflicts manage to obtain access to the mass media. Several such people have gained an international reputation in the course of their commitment, e.g., Medha Patkar, the leader of the movement against the Sardar Sarovar Dam in India, who was awarded the alternative Nobel prize.

The history of the Green party in Germany reveals a different developmental pattern: the party come to represent the 'Eco-Establishment' (Seager 1993, p. 170). A social movement has become an official public agent. The debate between 'fundamentalists' and 'realists' was about whether to promote social change from within the system (realists) or from the outside (fundamentalist). Spontaneous demonstrations relying on a very affective approach have yielded to pragmatic and committed politics using the official channels of public life. Professional respectability has replaced passionate involvement. A change in essence and style towards the perspectives of the establishment can be noted, with an acceptance of free market economy criteria, an emphasis on practicability, efficiency, and scientific arguments (Seager 1993, p. 201).

6. *Social and Political Relevance of Environmental Movements*

Environmental problems are caused by conflicting interests regarding natural resources and the preservation of the environment. Apart from those problems social movements also view environmental conflicts as a crisis of democracy and participation and as a cultural and ethical crisis because of the uneven power relations and conflicting systems of values and norms involved. With new actors entering the political field, traditional national politics and even states themselves are called into question by environmental movements. Traditional political parties lose members and voters while green parties or more spontaneous political groups gain influence.

In most cases it is impossible to solve the conflicts at either the local, regional, or national level. The issues at stake are manifold, each interest group portraying its own perspective. It is the interstices between self-determination and solidarity with human and non-human agents which are likely to create a substantial mobilizing potential. Hence new networks, operating at an international level and able to consider the differing interests and to mediate the uneven power relations, have to be built up in order for social movements with little formal political power to be heard.

See also: Environmental Justice; Environmental Security; Indigenous Knowledge and Technology; Moral Economy and Economy of Affection

Bibliography

Bandyopadhyay J 1999 Of floated myths and floated realities. *Economic and Political Weekly* **34**: 880

Bula-at L 1996 The struggle of the Kalinga women against the Chipco Dam project. *idoc internazionale* **3**(July–Sept): 24–7

Calderón F, Pscitelli A, Reyna J L 1992 Social movements: actors. Theories, expectations. In: Escobar A, Alvarez S E (eds.) *The Making of Social Movements in Latin America. Identity, Strategy and Democracy*. Westview Press, Boulder, pp. 19–36

Della Porta D 1995 *Social Movements, Political Violence, and the State*. Cambridge University Press, Cambridge, UK

Della Porta D, Diani M 1999 *Social Movements. An Introduction*. Blackwell, Oxford, UK

Ekins P 1992 *A New World Order. Grassroots Movements for Global Change*. Routledge, London

Eyerman R, Jamison A 1991 *Social Movements. A Cognitive Approach*. Pennsylvania State University Press, University Park, PA

Massey D 1993 Power geometry and a progressive sense of place. In: Bird B et al. (ed.) *Mapping the Futures: Local Cultures. Global Change*. Routledge, London, pp. 59–69

McCormick J 1989 *The Global Environmental Movement. Reclaiming Paradise*. Belhaven Press, London

Mohanty M, Tõrnquist O (eds.) 1998 *People's Rights. Social Movements and the State in the Third World*. Sage, Thousand Oaks, CA

Pile S 1997 Opposition, political identities and spaces of resistance. In: Pile S, Keith M (eds.) 1997 *Geographies of Resistance*. Routledge, London, pp. 1–32

Posey D 1995 High noon am rio xingœ. In: Rasper Martin (ed.) *Landräuber: Gier und Macht—Bodenschätze contra Menschenrechte*. Focus, Giessen, pp. 164–75

Rocheleau D, Thomas-Slayter B, Wangari E (eds.) 1996 *Feminist Political Ecology: Global Issues and Local Experiences*. Routledge, London

Seager J 1993 *Earth Follies: Coming to Terms with the Global Environment Crisis*. Routledge, New York

Seager J 1996 Hysterical housewives and other mad women—grassroot environmental organizing in the United States.

In: Rocheleau D, Thomas-Slayter B, Wangari E (eds.) 1996 *Feminist Political Ecology: Global Issues and Local Experiences*. Routledge, London, pp. 271–86
Slater D 1997 Spatial politics/social movements. Questions of (b)orders and resistance in global times. In: Pile S, Keith M (eds.) *Geographies of Resistance*. Routledge, London, pp. 258–76
Staggenborg S 1998 *Gender, Family and Social Movements*. Pine Forge Press, Thousand Oaks, CA
Wastl-Walter D 1996 Protecting the environment against state policy in Austria: From women's participation in protest to new voices in parliament. In: Rocheleau D, Thomas-Slayter B, Wangari E (eds.) *Feminist Political Ecology: Global Issues and Local Experiences*. Routledge, London, pp. 86–104

D. Wastl-Walter

Social Movements, Geography of

Social movements, like all social relations, are geographically constituted. What one means by 'social movements,' however, is open to a wide range of interpretations. Here, social movements are defined broadly as collective mobilizations of people seeking social or political change.

While social movements necessarily take place somewhere, their geography involves considerably more than an inventory of their locations. Geography is not a background onto which aspatial social processes are mapped. Rather, it is a fundamental aspect of the constitution of all social processes, affecting their evolution and outcome directly. It is part of the explanation of social processes, as well as an outcome (Massey and Allen 1984).

While a variety of geographic concepts can be employed in the analysis of social movements, three core geographic concepts are particularly important: space, place, and scale. Each of these concepts has its own evolution, which can be seen in the ways in which analyses of the geography of social movements have changed over time.

1. *Evolving Geographic Concepts and the Analysis of Social Movements*

1.1 *Spatial-analytic Analysis of Social Movements*

The spatial-analytic paradigm that dominated geographical analysis in the 1960s and 1970s worked from a conception of space that was based, by and large, on measures of distance. In this paradigm, social phenomena are described and explained in terms of absolute (e.g., with reference to a fixed grid such as latitude and longitude) or relative (e.g., transportation costs or travel time) location. Virtually all social phenomena have been shown to possess a 'distance decay' function as interaction costs increase with distance from the activity center. The spatial diffusion of information, as well as barriers to such diffusion, became a common theme in much spatial-analytic research, including that addressing social movements (see *Diffusion: Geographical Aspects*). The growth and mobilization of social movements is clearly dependent upon flows of information, as demonstrated by Sharp's (1973) analysis of the spatial diffusion of US postal strikes, and Hedstrom's (1994) work on the spatial diffusion of Swedish trade unions. Most spatial-analytic work on social movements, however, examines the co-presence of specific social conditions within clearly delimited places. Adams (1973), for instance, examines the co-location of factors promoting urban protests, while Cutter et al. (1986) consider how characteristics measured at the state level are associated with support for nuclear weapons freeze referenda (see *Peace Movements, History of*). In the 1970s, attempts to explain the occurrence of social phenomena in terms of distance or co-location came under heavy criticism for ignoring power relations, and in particular, the relation between 'the social' and 'the spatial.'

1.2 *Contemporary Conceptions of Space and the Analysis of Social Movements*

Influenced by the work of Henri Lefebvre and others, contemporary conceptions of space emphasize the mutual constitution of 'the social' and 'the spatial.' Space is seen increasingly as a necessary component of the constitution of power relations. Equally significant, space is no longer conceptualized primarily in terms of the physical location of material objects. Instead, both material spatial practices, as well as symbolic or representational spaces, are examined.

As political economy approaches swept through the social sciences in the 1970s and 1980s, the spatial constitution of material interests was linked explicitly to the geography of social movements. The relationship between the geography of political and economic conditions, and the incidence of protest is demonstrated in detail in Charlesworth's (1983) atlas of rural protest in the UK. Bennett and Earle (1983) show that the failure of socialism in the USA can be linked directly to geographic patterns of wage and skill differentiation; where differentiation was substantial, working class solidarity was undermined (see *Labor Movements, History of*). Similarly, Massey (1984) argues that the spatial division of labor affects directly the material interests that underlie political movements. Markusen (1989) makes a somewhat different point: a region's stage in the profit cycle shapes the types of movements that are likely to emerge in it. Environmental movements are more likely to emerge

in economically booming regions where rapid growth places stress on the environment, while pro-growth politics are common in regions of decline (see *Social Movements: Environmental Movements*).

Geographically differentiated state systems also shape the geography of social movements. Kitschelt's (1986) comparison of the political structures of France, Sweden, West Germany, and the USA shows that differences in central state structures are related to national-level variations in anti-nuclear mobilization. Other studies link local-level variation in social movement mobilization and success to differences in political opportunity structures among local states (Duncan et al. 1988, McAdam 1996, Tarrow 1996). In the USA many social movements began as highly localized movements in politically opportune places, then spread nationally (see *Citizen Participation*).

Symbolic or representational spaces affect the geography of social movements as well. The symbolic codings of spaces, i.e., the often subtle and unacknowledged signs that shape people's reactions to spaces, may indicate belonging for some groups and exclusion for others, empowerment for some groups and repression for others, safety for some groups and danger for others. Such codings can foster or hinder social movement mobilization. The creation of symbolic space is intertwined with the process of collective identity formation, as Marston (1988) shows for ethnic identity, Rose (1993) demonstrates for gendered identity, and Ruddick (1996) illustrates for interlocking 'race,' class, and gender identities (see *Social Movements and Gender*). Such identities are crucial to the formation of solidarity necessary for movement mobilization (see *Social Movements, History of: General*). The dominant symbolic codings of space are open to contestation, and counter-representations are often constructed by insurgent groups. Representational spaces, and their internalization, relate directly to the ways in which social movement messages resonate in some places, and not in others.

1.3 Place and the Analysis of Social Movements

Many geographic analyses emphasize the place-based constitution of social processes. Place is best understood as a multidimensional concept entailing 'locale,' i.e., the geographical setting of everyday social relations, 'location,' i.e., geographical position vis-a-vis wider scale interactions; and 'sense of place,' i.e., the esthetic and emotional bonds to locale (Agnew 1987). The concept of place stresses the unique ways in which social processes may interact in specific milieus, calling our attention to the 'situatedness' of all human action and institutions. At the same time, place is not to be equated with 'the local.' Place implies the articulation of localized processes, with other processes operating at broader scales.

Much of the geographical literature on social movements stresses the ways in which identities and allegiances are shaped in place. Thrift and Williams (1987) analyze the ways in which consciousness and identity are produced through place-specific social practices and the time–space paths of individuals' everyday lives. Katznelson's (1981) work can be interpreted as a place-sensitive analysis of US workplace and community politics, stressing how distinct places shape identity in distinct ways. Savage's (1987) analysis of British working-class politics demonstrates how the interactions of workplace and lifeworld spaces have effects on identity that vary from place to place. Recent work on identity construction stresses the place-specific interplay of spaces of production, reproduction, and the state that produce multiple, fragmented, and sometimes contradictory identities, all with direct implications for social movement mobilization (Ruddick 1996, Cope 1996, Brown 1997). Place itself can be a primary basis for collective identity construction. The defense of places associated with specific collective identities is a common theme of many new social movements (Melucci 1994). Increasingly, cities find themselves in competition with each other for capital investment; in response, place-based identification and coalition building has often taken precedence over class and other axes of social identity (Hudson and Sadler 1986, Swyndegouw 1989).

Identity is not all that is shaped in place. Agnew (1987, p. 59) observes that 'different places produce different degrees and orientations of organizational capacity.' Routledge (1993), for example, shows that social and environmental movements in India faced opportunities for, and constraints to, mobilization that relate to place-specific social networks, systems of meaning, and patterns of political repression. Miller (2000) shows how a variety of place-specific characteristics, including political opportunity structures, available resources, material interests, and attachments to place, shaped the strategies and effectiveness of peace-organizing differentially in several Boston area municipalities.

1.4 Scale and the Analysis of Social Movements

The concept of place clearly carries within it issues of geographic scale. Geographic scale has commonly been thought of as a fairly straightforward technical issue relating to the telescoping frame of geographic analysis (from local to global), but such a notion of scale as pre-given or unproblematic is increasingly challenged.

Most analyses of social movements assume that the appropriate scale of analysis is that of the nation state, but such an assumption overlooks the often critical subnational, as well as international, constitution of

social movement processes. Moreover, rather than being seen as an 'external fact awaiting discovery,' scale should be viewed as a 'way of framing conceptions of reality' (Delaney and Leitner 1997, pp. 94–5) that is employed by social movement organizations and their opponents as a way of gaining political advantage. As Smith (1993, p. 90) explains:

> The construction of scale is not simply a spatial solidification of contested social forces and processes; the corollary also holds. Scale is an active progenitor of specific social processes. In a literal as much as metaphorical way, scale both contains social activity and at the same time provides an already partitioned geography within which social activity takes place. Scale demarcates the sites of social contest, the object as well as the resolution of contest.

2. *Geographic Tactics and Strategies of Social Movements*

Given that all social movements are geographically constituted, all social movement organizations must make decisions regarding the geographic tactics and strategies they employ. Sometimes these decisions are explicitly considered, often they are not.

Social movements frame their messages in particular ways that they hope will resonate with particular groups. Given a highly differentiated geography of culture, symbolization, and interests, the resonance of any given frame will depend in substantial measure on the geographic targeting of the message. Similarly, the recruitment of social movement members is intertwined inextricably intertwined with the geography of mobilization activities (Harvey 1985, Miller 2000).

While identities constructed in place and place-based identities frequently are central to social movement solidarity, the absence of such bases of solidarity can be rectified through efforts to construct new places of group interaction and identity. Castells (1983), for example, details how the growth of gay bars, businesses, and gathering places was crucial to the construction of shared identity in the San Francisco gay community. Castells also shows how, in the case of the Madrid citizens' movement, the construction of a community center in the Orcasitas shantytown was essential to breaking down social walls and building broad, place-based bonds around which the movement organized.

The tactics and strategies that social movements employ to challenge the existing social order are rooted in an often-unexamined normative geography. Cresswell (1996) makes a compelling case that spatial structures and systems of places provide ideologically-loaded schemes of perception. Dominant conceptions of what is 'normal' and 'appropriate' is defined 'to a significant degree, geographically, and deviance from this normality is also shot through with geographical assumptions concerning what and who belong where' (p. 27). He illustrates his argument with several examples, including that of the Greenham Common women's peace camp that challenged British participation in the nuclear arms race through 'out of place' behavior that 'denaturalized the dominant order [and] showed people that what seemed natural, could, in fact, be otherwise' (p. 125).

Indeed, social movements can be mobilized and social conditions can be changed through a re-symbolization of existing spaces or through a re-territorialization of existing symbolization (Pile and Keith 1997). Harvey (1989, p. 233) makes a case for the re-symbolization of space when he argues:

> If workers can be persuaded … that space is an open field of play for capital but a closed terrain for themselves, then a crucial advantage accrues to the capitalists … power in the realms of representation may end up being as important as power over the materiality of spatial organization itself.

An example of the re-territorialization of existing symbolization can be found in the recent work that has been done in several US metropolitan areas to recast the ongoing political battles between central cities and suburbs. In several metropolitan areas this common conception of territorial interests is being replaced with one that aligns the central cities with the inner ring suburbs, creating a far more powerful territorial alliance behind the movement for a more equitable metropolitan distribution of affordable housing, school funding, and tax revenue (Orfield 1997).

Many of the efforts of social movements to redefine the spaces of struggle involve attempts to 'jump' to a more favorable geographic scale (Smith 1993). In addition to changing symbolic geography, altering the scale of material processes can be central to social movement strategy. Examples include shifting from regional to national-scale labor bargaining (Herod 1998), and seeking out more favorable political opportunity structures at a different level of the state, e.g., local rather than central (Miller 2000).

The geographic tactics and strategies adopted by social movements and their opponents play a powerful role in the nature of social movement mobilization and the degree of success or failure social movements achieve.

See also: Community Power Structure; Interest Groups, History of; Peace Movements; Political Protest and Civil Disobedience

Bibliography

Adams J 1973 The geography of riots and civil disorders in the 1960s. In: Albaum M (ed.) *Geography and Contemporary Issues*. Wiley, NewYork, pp. 542–64

Agnew J A 1987 *Place and Politics*. Allen & Unwin, Boston, MA
Bennett S, Earle C 1983 Socialism in America: a geographical interpretation of its failure. *Political Geography Quarterly* **2**(1): 31–55
Brown M 1997 Radical politics out of place? In: Pile S, Keith M (eds.) *Geographies of Resistance*. Routledge, New York, pp. 152–67
Castells M 1983 *The City and the Grassroots*. University of California Press, Berkeley, CA
Charlesworth A (ed.) 1983 *An Atlas of Rural Protest in Britain, 1548–1900*. University of Pennsylvania Press, Philadelphia, PA
Cope M 1996 Weaving the everyday: identity, space, and power in Lawrence, Massachusetts, 1920–1939. *Urban Geography* **17**(2): 179–204
Cresswell T 1996 *In Place/Out of Place*. University of Minnesota Press, Minneapolis, MN
Cutter S L, Holcomb H B, Shatin D 1986 Spatial patterns of support for a nuclear weapons freeze. *The Professional Geographer* **38**(1): 42–52
Delaney D, Leitner H 1997 The political construction of scale. *Political Geography* **16**(2): 93–97
Duncan S, Goodwin M, Halford S 1988 Policy variations in local states: Uneven development and local social relations. *International Journal of Urban and Regional Research* **12**(1): 107–28
Harvey D 1985 *Consciousness and the Urban Experience*. Johns Hopkins University Press, Baltimore
Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Cambridge, MA
Hedstrom P 1994 Contagious collectivities: On the spatial diffusion of Swedish trade unions, 1890–1940. *American Journal of Sociology* **99**(5): 1157–79
Herod A (ed.) 1998 *Organizing the Landscape: Geographical Perspectives on Labor Unionism*. University of Minnesota Press, Minneapolis, MN
Hudson R, Sadler D 1986 Contesting work closures in Western Europe's old industrial regions: Defending place or betraying class? In: Scott A J, Storper M (eds.) *Production, Work, Territory*. Allen & Unwin, Boston, pp. 172–94
Katznelson I 1981 *City Trenches*. University of Chicago Press, Chicago
Kitschelt H 1986 Political opportunity structures and political protest: Anti-nuclear movements in Four Democracies. *American Sociological Review* **49**: 770–83
Markusen A 1989 Industrial restructuring and regional politics. In: Beauregard R (ed.) *Economic Restructuring and Regional Politics*. Sage, Newbury Park, CA, pp. 115–48
Marston S A 1988 Neighborhood and politics: Irish ethnicity in nineteenth century Lowell, Massachusetts. *Annals of the Association of American Geographers* **78**: 414–32
Massey D 1984 *Spatial Divisions of Labor*. Methuen, New York
Massey D, Allen J 1984 *Geography Matters*. Cambridge University Press, New York
McAdam D 1996 Conceptual origins, current problems, future directions. In: McAdam D, McCarthy J D, Zald M N (eds.) *Comparative Perspectives on Social Movements*. Cambridge University Press, New York, pp. 23–40
Melucci A 1994 A strange kind of newness: what's 'new' in new social movements? In: Larana E, Johnston H, Gusfield J R (eds.) *New Social Movements: From Ideology to Identity*. Temple University Press, Philadelphia, pp. 101–30
Miller B A 2000 *Geography and Social Movements*. University of Minnesota Press, Minneapolis, MN
Orfield M 1997 *Metropolitics: A Regional Agenda for Community and Stability*. The Brookings Institution, Washington, DC
Pile S, Keith M 1997 *Geographies of Resistance*. Routledge, New York
Rose G 1993 *Feminism and Geography*. University of Minnesota Press, Minneapolis, MN
Routledge P 1993 *Terrains of Resistance: Nonviolent Social Movements and the Contestation of Place in India*. Praeger, Weatport, CT
Ruddick S 1996 Constructing difference in public spaces: Race, class, and gender as interlocking systems. *Urban Geography* **17**(2): 132–51
Savage M 1987 *The Dynamics of Working Class Politics: The Labour Movement in Preston, 1880–1940*. Cambridge University Press, Cambridge, UK
Sharp V 1973 The 1970 postal strikes: The behavioral element in spatial diffusion. In: Albaum M (ed.) *Geography and Contemporary Issues*. Wiley, New York, pp. 523–32
Smith N 1993 Homeless/global: scaling places. In: Bird J, Curtis B, Putnam T, Robertson G, Tickner L (eds.) *Mapping the Futures: Local Cultures, Global Change*. Routledge, New York, pp. 87–119
Swyndegouw E 1989 The Heart of the Place: The Resurrection of Locality in an Age of Hyperspace. *Geografiska Annaler* **71B**: 31–42
Tarrow S 1996 States and opportunities: The political structuring of social movements. In: McAdam D, McCarthy J D, Zald M (eds.) *Comparative Perspectives on Social Movements*. Cambridge University Press, New York, pp. 41–61
Thrift N, Williams P 1987 The geography of class formation. In: Thrift N, Williams P (eds.) *Class and Space: The Making of Urban Society*. Routledge & Kegan Paul, New York, pp. 1–22

B. Miller

Social Movements, History of: General

Social movements, according to sociologist Sidney Tarrow's definition, are 'collective challenges, based on common purposes and social solidarities in sustained interaction with elites, opponents and authorities' (Tarrow 1998, p. 4). This definition, derived from studies of both historical and contemporary social movements, locates the earliest movements in the late eighteenth and early nineteenth centuries. The emergence of social movements was contemporaneous with: (a) the increased centralization of national states that accompanied war and revolution in western Europe and North America; (b) the rise of the popular press and the circulation of notions of natural rights and liberty put forward by Enlightenment and revolutionary thinkers; and (c) particularly in England and the English-speaking colonies of North America, the rise of evangelical and nonconforming religions in which literacy and reading were cultivated and principles of human rights formulated. From that period and place, social movements have proliferated to the world.

The chief focus of this article is the history of social movements in the West that have attracted most scholarly attention; it draws particularly on studies done in the years 1975–2000—a period in which the concept of social movement has been redefined and developed—for its theoretical framework and empirical base. The British movement for the abolition of slavery—an emblematic, and by some criteria the first social movement—serves as a concrete example of an early crusade that embodied many of the central features of the phenomenon.

1. The Abolition of Slavery in Great Britain

The eighteenth and nineteenth century British abolitionist movement was distinctive in that it was aimed at an institution which had only a minor presence in Great Britain, whereas most social movements have been concerned with issues close to hand. Slavery nevertheless involved many Britons: British ships sailing from Liverpool were active in the slave trade carrying slavers to Africa and slaves to the Americas, and the 'peculiar institution' was common in Britain's North American and universal in its Caribbean colonies. The question of the legality of slavery in Britain was raised as early as the 1720s, but it was not until 1772 that Lord Chief Justice Mansfield ruled that slavery was illegal under English law. English members of the Society of Friends had developed the earliest critique of slavery as contrary to Christian teaching, and local Quaker meetings first excluded slave traders in 1761 and later slave owners from membership.

Quakers then turned to political action to end British involvement in the slave trade, establishing in 1787 the Society for the Abolition of the Slave Trade, commonly known as the London Committee. This organization welcomed non-Friends as members, and recruited both middle and working-class adherents. The political technique utilized by the Committee in this and later campaigns was mass petitions to Parliament. Thomas Clarkson, a gifted organizer, and other activists worked outward from London and local Friends meetings to form regional organizations, recruit members, distribute leaflets, and collect signatures on petitions to Parliament calling for the end of the slave trade. (Clarkson and his associates began their crusade with an exclusive focus on the slave trade, as less contentious than abolition.) Visual images portraying the horrors of the slave trade such as a print of the cross section of a slave ship with its tightly packed human cargo and Josiah Wedgwood's medallion of a kneeling chained slave, inscribed with the motto 'Am I not a Man and Brother?' were highly effective in attracting supporters.

There was an outpouring of support, with about 60,000 male signers on the first petitions, including 11,000 men from Manchester, the center of the burgeoning cotton spinning industry. Parliamentary sympathy for the cause seemed strong up to 1792, when 519 petitions with up to 400,000 signers were collected. However, legislative support dropped with the slave revolt in Haiti, the contemporaneous radicalization of the French Revolution, and the outbreak of war against France. In 1794, the House of Lords decisively rejected the abolitionist resolutions presented to it.

Political opportunities for action on abolition reappeared in 1804, after Britain's victories in the Napoleonic Wars and its seizure of the French Caribbean colonies. Clarkson and other organizers again canvassed the country seeking signatures; he also sought support from members of Parliament, in particular William Wilberforce, an Evangelical MP from Yorkshire who spoke eloquently about the evils of the trade and Samuel Whitbread, a nonconformist Protestant already known as a social reformer. Clarkson's powers of persuasion and his Parliamentary allies combined to produce in 1807 a majority vote in Parliament to forbid British ships' participation in the slave trade to both British possessions and foreign slave-holding colonies. The African Institution, founded in the same year, took as its task ascertaining the effectiveness of the British legislation.

Although British diplomats signed agreements eschewing the slave trade with other European powers, and indeed paid compensation when they did, the results were uneven. British activists resolved to continue their efforts, founding the Society for the Amelioration and Gradual Abolition of Slavery (known familiarly as the Anti-Slavery Society) in 1823 and the Agency Committee (1831), which demanded immediate and unconditional slave emancipation. The Agency's professional organizers traveled around England, promoting local petition drives and pressure on candidates in the Parliamentary elections to declare their support of full abolition. Members of the House of Commons elected under the 1832 Reform Act were ready to vote for the abolition of slavery in British possessions, but the House of Lords was not. The outcome was a compromise Emancipation Act of 1833 that set August 1, 1834 as the date on which slavery would end, but stipulated a complex gradual process to complete freedom.

The former slaves would become 'apprentices,' working for pay three-quarters of their time for their ex-masters; one-quarter of the ex-slaves' time was to be reserved for them to cultivate their newly-granted plots. Slave owners would be compensated by the state for the loss of their 'property.' The story did not end there, however. Disagreements about strategy between the Anti-Slavery Society (which had accepted the compromise Emancipation Act) and the Agency (which had not) continued, and the latter established yet another organization (the Central Negro Emancipation Committee) and sought to persuade Parliament that oppressive conditions still prevailed. In

the end, the complex system based on the 1833 Act proved difficult to manage, and the planters themselves ended apprenticeship in 1838.

The international abolition movement matured in the 1830s, and began to hold annual congresses (which some American abolitionists attended) in 1840. The United States Civil War (1861–5) was the outcome of a series of legal and political disputes about slavery and state's rights. In 1863 President Abraham Lincoln's Emancipation Proclamation ended slavery in the Confederacy, and during the postwar Reconstruction the Thirteenth, Fourteenth, and Fifteenth Constitutional amendments (which completed abolition, made the ex-slaves citizens, and guaranteed their right to vote) were ratified. The abolition of slavery in Cuba (1886) and Brazil (1888) ended legal slavery in the West.

2. *The Social Character of Movements*

Over the period 1975–2000, sociologists have begun to define 'contentious politics'—confrontations of ordinary people with authorities and/or elites—as the building blocks of social movements. Sidney Tarrow defines the conditions under which contentious politics is triggered:

> When changing political opportunities and constraints create incentives for social actors who lack resources on their own, they contend through known repertoires of contention and expand them by creating innovations at their margins. When backed by dense social networks and galvanized by culturally resonant, action-oriented symbols, contentious politics leads to sustained interaction with opponents. The result is the social movement (Tarrow 1998, p. 2).

The case of the British movement for the abolition of the slave trade and slavery illustrates this definition:

(a) the movement was collective and interactive, between authorities and the populace; it was conducted early on through organizations established for the purpose and spread by movement professionals;

(b) it was aimed at political authorities (in this case Parliament) which could take action to achieve the movement's goal;

(c) each phase of activism on abolition began with an opening of political opportunity and ended when new political constraints closed off forward movement;

(d) the network of Quaker meetings was the first activated in the cause, but other church groups and local branches of the anti-slavery societies joined the abolitionist network over the period;

(e) interaction was contentious and drew on symbols that helped promote action;

(f) and participation in the movement came to be perceived by activists as constitutive of their identities.

3. *Woman's Suffrage in the United States*

The formation of the United States woman's suffrage movement was closely connected to the social groups and organizations that promoted the abolition of slavery there. For example, American woman's suffrage activists Lucretia Mott and Elizabeth Cady Stanton first met each other at the London World Anti-Slavery Convention of 1840, where they discussed the possibility of organizing an organization to promote women's rights. Mott, a New Englander then living in Philadelphia and the senior of the two had founded the first Female Anti-Slavery Society in the United States; her home was a stop on the 'underground railroad' through which slaves escaped to Canada and freedom. The well-educated daughter of a judge, Stanton was born in western New York State, married an abolitionist leader, and eventually settled with her family in Seneca Falls, NY. Stanton and Mott renewed their acquaintance in 1848, when they and three like-minded women issued a call to a women's rights conference in the summer. To their surprise, some 300 women and men came, applauded Stanton's speech echoing the Declaration of Independence and passed a resolution on women's suffrage: 'Resolved, that it is the duty of the women of this country to secure to themselves their sacred right to the elective franchise.'

The movement developed slowly through annual conferences, spreading primarily to the Middle West and East coast. In 1851, Stanton met Susan B. Anthony, then a paid organizer in the temperance movement. Anthony brought her organizational skills to women's rights, beginning a drive in 1854 based on organizers in each county of the state of New York who recruited local women. These in turn collected signatures for a petition calling for women's control of their own earnings, mothers' guardianship of children in divorce, and women's suffrage. The first concrete accomplishment—a New York state law granting women property rights—was passed in 1860.

With the outbreak of the Civil War in April 1861, organizing efforts were temporarily suspended. Not for long, however, as Anthony and Stanton reactivated their abolitionist and women's rights networks in 1863, issuing a call to a mobilization meeting in New York. The outcome was the National Women's Loyal League that set to work gathering close to 400,000 signatures on a petition for the Thirteenth constitutional amendment that would free the slaves. Women also served on the Civil War Sanitary Commission, which raised funds and organized and administered nursing and support services for the military hospitals and convalescent centers.

The campaign for ratification of the Fourteenth and Fifteenth amendments (the first made male ex-slaves citizens and the second granted them the vote) split the former abolitionist movement colleagues between those who supported rights for the ex-slaves as a first

priority, and those who supported universal suffrage. In 1869, two new organizations dedicated to women's suffrage emerged from the split: the National Woman Suffrage Association founded by Anthony and Stanton which objected to the Fourteenth and Fifteenth Amendments because they specified male ex-slaves; and the American Woman Suffrage Association, founded by Lucy Stone, her husband Henry B. Blackwell, and Mary Livermore. These associations pursued the cause by different methods and separately for 20 years during which the westward movement was completed, the country industrialized, and new women's associations such as the Women's Christian Temperance Union (1874) and the National Association of Colored Women (1895) were founded. In 1890, the National and the American Associations merged to form the National American Woman Suffrage Association, with Elizabeth Cady Stanton as its first president; she was succeeded two years later by Susan B. Anthony who served to 1900.

Led after 1900 by Carrie Chapman Catt, a gifted organizer the National American Association continued its strategy: seeking constitutional change state by state. Catt resigned in 1904 to play an active role in founding the International Woman Suffrage Alliance launched that year. There was little activity around the woman suffrage ('Susan B. Anthony') amendment to the United States Constitution for over 10 years. New tactics appeared in 1913, devised by the more militant new movement recruits Alice Paul and Lucy Burns who were appointed to the National American Suffrage Association's Congressional Committee. They promptly organized a demonstration of some 5,000 women who marched in Washington in support of a constitutional amendment the day before Woodrow Wilson's inauguration for a second term. Despite the organization's permit for the parade, the police were unprepared for the crowd reaction and the women were mobbed and jeered. Later that year Paul and Burns established the Congressional Union, and their organization eventually became independent of the National American, which continued to pursue its state-by-state strategy, with only slow progress.

In 1915, Carrie Chapman Catt agreed to lead the campaign for women's suffrage in New York State, and once again became president of the National American Suffrage Association with a newly energetic board committed to her plan for a renewed campaign for a national constitutional amendment. (The Congressional Union and its successor the Woman's Party did not join the effort. The Woman's Party nevertheless contributed to the process by courting arrest through disruptive and provocative action, which made the National American's actions look reasonable.) At its convention in late 1917 the latter warned that if Congress did not approve the suffrage amendment before the next election, it would adopt the British suffragette method: opposing the election of Congressional candidates who opposed the federal amendment, no matter what their party. The amendment passed in the House of Representatives by the required two-thirds majority in January 1918, the Senate in 1919, and ratification by the requisite proportion of states was completed in 1920.

4. *Twentieth Century Social Movements: Italian Fascism*

In the early twentieth century, several movements demonstrated unequivocally the adaptability of the social movement to regressive ends. The Italian Fascist movement serves as an example.

After the end of World War One late in 1918, Italy was swept by waves of contentious politics: demobilized troops demanded jobs; peasants and agricultural laborers agitated for land; consumers called for price controls; and workers struck for higher wages. Among the groups engaging in collective action were the *fasci di combattimento* founded by Benito Mussolini (formerly a revolutionary socialist) who convened the first *fascio* in Milan on March 21, 1919. At a public rally two days later the national program was only vaguely described (it was not fully formulated and publicized until months later). The program called for integration of all Italian 'national' territory into the nation, and pledged to oppose all neutralist politics no matter the party involved; other radical social and political demands included universal (male) suffrage, a constituent assembly, a republic with local autonomy, the eight-hour day, worker participation in factory management, land to the peasants, elimination of the Senate and all noble titles, and an end to conscription, expropriative taxes, speculation, and monopoly.

The Fascists' first public action (April 15, 1919) was a counter-demonstration together with students, World War veterans, and revolutionary syndicalists which attacked an anarchist/socialist march protesting repression in Milan. It ended with the sacking and burning of the headquarters of the socialist newspaper *Avanti*! In the next two years most Italian popular contention was dominated by socialists and organized social Catholics (represented by the newly-founded Popular Party) who both sought to control and lead the strikes in northern industries and large-scale agriculture in the Po Valley, the land occupations, and eventually in fall, 1920, the factory occupations. These struggles all occurred in northern Italy, where socialist and Popolari strength was concentrated. The Fascist counterattack from its urban bases (again in the north) began in the same period: physically assaulting unionized farm workers on strike, the headquarters of peasant leagues, cooperatives, and unions, burning left wing newspapers' offices and printing plants, and breaking up meetings of the socialist communal councils as well.

Socialist and Popular Party organizations crumbled as Fascist membership and attacks both increased.

The Fascists had become a political force to be reckoned with and their success went hand in hand with the demobilization of the left. In August 1921, Mussolini signed the so-called Pacification Pact with the socialists and Popolari, promising to eschew violence, but his followers forced him to back away from the agreement and he publicly renounced it later in the year. He also announced the formation of a political party, which he described as a rival to the state, outside traditional politics. Violent clashes between socialists and Fascists continued into 1922, when the Fascist threat to seize the state forced the cancellation of a general strike planned by an ad hoc group of unions; by then the entire Po Valley, and much of North and Central Italy was in Fascist hands.

As the constitutional parties called for a new government, Mussolini threatened to march on Rome to seize power. The march never happened; instead, Fascist Black Shirts occupied provincial government and party headquarters, post offices, and police stations (again in the north, where socialist power had been concentrated) sometimes without resistance. Fascist threats and his own concerns led the king to refuse to sign the declaration of martial law which the prime minister had drawn up, and Mussolini's first government (based on a coalition of center and right parties) was formally constituted on October 31, 1922. Mussolini used nationalism wreathed in patriotism as a weapon against the left, and he continued to be a master of language and image once in power.

Although there had been social movements in the European colonies (and in China, as well) before the outbreak of World War I in 1914, there were many more non-Western movements in the twentieth century; they ranged in principal concerns from nationalism to widow burning and female circumcision (in India and Kenya respectively). Nationalists sometimes rejected Western notions of modernity and connected customs that seemed unacceptable to Westerners to colonial cultural authenticity. Two examples are Mahatma Gandhi in his representation of himself with a spinning wheel and in his advocacy of *khadi* (handspun and woven cotton cloth which he himself wore in his later years), and Kenyan nationalist Jomo Kenyatta who defended female circumcision on behalf of the Kikuyu Central Association, a pre-independence nationalist organization. More recently, televised news reports and rapid communication by telephone, fax, and e-mail have made possible the almost instantaneous transmission of news not only in print, but of visual images as well. To what extent can we speak of transnational social movements?

5. *Transnational Social Movements*

The above discussion of anti-slavery and women's rights as social movements notes the international dimensions that were integral to social movements even in the late eighteenth century; similarities and connections of social movements across borders are not new. Sidney Tarrow offers a more restrictive definition of a transnational social movement:

> sustained contentious interactions with opponents—national or nonnational—by connected networks of challengers organized across national boundaries … What is important … is that the challengers themselves be both rooted in domestic social networks and connected to one another more than episodically through common ways of seeing the world, or through informal or organizational ties, and that their challenges be contentious in deed as well as in word (1998, p. 184).

Tarrow discusses three movements that fit his definition: the contemporary ecological and environmental movement, the European and American peace movement of the 1980s, and Islamic fundamentalism.

The ecological/environmental movement, which mounted coordinated worldwide Earth Day demonstrations in the late 1970s and early 1980s, provides the example discussed here. Some of its social movement organizations and their coordinated activities across borders do qualify as transnational organizations, as Tarrow demonstrates. Greenpeace, for example, has been a genuinely transnational social movement organization, with professional organizers and offices in many countries, and supporters who share a common world view and are willing to support its advocacy. Founded in Vancouver, Canada, according to its 1992–3 report it had over five million supporters (contributors), 43 offices, and over a thousand paid employees. It developed out of domestic movements that had similar goals, organized advocacy networks, and conducted contentious politics that cross national boundaries. Its members and sympathizers demonstrated against French nuclear testing in the Pacific and against abusive fishing for whales by Japanese and Soviet whalers.

Nevertheless, most international connections are less sustained and less professionalized than Greenpeace. Not all issues can attract support across the varied conditions of world polities, and most political, social, and economic international movements continue to work within states and national-level private institutions like large business organizations or labor unions rather than through transnational voluntary associations which represent individuals' world views directly.

6. *Overview: Social Movements from the Eighteenth to the Twentieth Century*

Sidney Tarrow's definition of the social movement (quoted in the opening sentence of this article) posits its applicability to the entire period since the first appearance of social movements in the late eighteenth century up to the present. His definition is based on the premise that although the issues and tactics of social

movements have changed over the two centuries in which they have proliferated the defining form of social movements—their collective nature, common purposes, and shared base in social solidarities with roots in the social structure in which they occur—have remained relatively unchanged. Not all social movement scholars agree with this premise; among European sociologists and political scientists in particular the concept of 'new social movements' has been developed since 1985.

Kriesi et al. (1995) argue that a new type of social movements which address similar issues yet vary significantly in form according to the national political context has emerged in Western Europe since 1968. Kriesi and his colleagues emphasize contemporary cross-national variation in domestic politics and corresponding differences in the timing and other characteristics of social movements. The examples that they discuss under the rubric of new social movements include 'the ecology movement (with its antinuclear energy branch), the solidarity movement (with the Third World), the contemporary women's movement, the squatters' movement, as well as various other movements mobilizing for the rights of discriminated-against minorities' such as gays (1995, p. xii). The new type of social movements discussed by Kriesi et al. (1995) is classified according to the chief constituencies of the movements and the connections among them. Tarrow's definition of social movements leaves open the characteristics of their members and their goals and stresses the movements' longstanding form. Two alternatives present themselves: in terms of their political concerns and context then, recent social movements may herald changes in some aspects of movement politics and require redefinition of the concept. Alternatively, the broad definition with which this article begins argues for stability in the form of social movements even as the issues that bring them into the public arena change. Whether either one of these alternatives will contribute to understand contemporary and future social movements better is an empirical question that can only be determined with a longer span of observations.

Bibliography

Anderson B 1991 *Imagined Communities*. Verso, London

Costain A R, McFarland A S (eds.) *Social Movements and American Political Institutions*. Rowman and Littlefield, Boulder, CO

Keck M E, Sikkink K 1998 *Activists Beyond Borders: Advocacy Networks in International Politics*. Cornell University Press, Ithaca, NY

Kriesi H, Koopmans R, Duyvendak J W, Giugni M G (eds.) 1995 *New Social Movements in Western Europe: A Comparative Analysis*. University of Minnesota Press, Minneapolis, MN

McAdam D, McCarthy J D, Zald M N (eds.) 1996 *Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings*. Cambridge University Press, Cambridge, UK

Melucci A 1985 The symbolic challenge of contemporary movements. *Social Research* **52**(4): 789–816

Offe C 1985 New social movements: Challenging the boundaries of institutional politics. *Social Research* **52**(4): 817–67

Tarrow S 1998 *Power in Movement: Social Movements and Contentious Politics*, 2nd rev. edn. Cambridge University Press, Cambridge, UK

Tilly C 1995 *Popular Contention in Great Britain, 1758–1834*. Harvard University Press, Cambridge, MA

Touraine A 1985 An introduction to the study of social movements. *Social Research* **52**(4): 749–87

L. Tilly

Social Movements: Psychological Perspectives

Social movements are sustained actions undertaken by organized groups of people seeking social or political change. The groups involved in social movements are structured, have goals and ideologies, and pursue their goals over time. The psychology of social movements focuses on the motives that lead people to join, work on behalf of, and stay within social movements.

One reason for people to join social movements is because of the desire to adopt a favorable identity and/or engage in positive actions. Social identity theory notes that people use groups to define themselves, so joining a group is one way to create a favorable identity (Tajfel 1982, see *Social Identity, Psychology of*). When people identify with a group, they take on the values and lifestyle of that group, seeing themselves as defined through the attributes that define the group. People joining a sports team, for example, will come to think about themselves in terms of their athletic ability—a prototypical attribute of the group.

In addition to giving people values or attributes to define themselves, identifying with groups provides people with information about their self-worth. Since people value positive self-worth, they are more likely to identify with and join a group when the group they join has a favorable reputation. Of course, people can also work collectively to better the reputation of their group, as in the gay rights movement, the civil rights movement, or other social movements motivated by the desire to enhance the stature of low status groups (Kelly and Kelly 1994, Simon et al. 1998). Social identity theory suggests one motivation for joining social movements—to create a favorable identity and to work on behalf of the values and lifestyle that identity represents. Although this is an important motivation for social movements, this positive motivation has not been the focus of most of the

psychological research on social movement participation. Instead, the focus has been on protest movements, which people join in response to perceived injustice. That motivation is explored in several theories, the earliest being the theory of relative deprivation.

1. Relative Deprivation

The factors that shape the generation of the grievances that lead to social movements are first addressed by the psychological literature on relative deprivation. This literature argues that people's feelings about their social conditions are only loosely connected to their objective situations. Instead of developing directly from people's objective circumstances, subjective feelings of satisfaction and dissatisfaction are linked to the comparisons that people make of their own outcomes to other outcomes. This work makes the important point that it is the interpretation of grievances that is key to their role in shaping social movements (see Klandermans 1989).

One type of comparison involves temporal comparisons of one's current outcomes to one's outcomes at some other point in time. Temporal comparisons are the basis for several of the best known models of the psychology underlying the development of social movements (Davies 1962, Gurr 1970). Gurr (1970) distinguishes among several different patterns of relative deprivation, each of which can create the necessary conditions for social movements. The best known pattern is progressive deprivation, which describes a situation in which both prospects and expectations are rising, but expectations are rising higher (Davies's 'j-curve'). In this case, the discrepancy between expectations and prospects increases, leading to dissatisfaction and support for social movements.

A second type of comparison involves social comparison to the outcomes of others. Again the key point is that people's satisfaction is not directly linked to the objective quality of their outcomes. Instead, it is comparative. The main question is how people decide to whom to compare the outcomes they receive (Suls and Wills 1991).

Studies suggest that when people are comparing their outcomes to standards, they apply principles of entitlement or deservingness (see *Justice: Social Psychological Perspectives*). Use of these justice-based principles shows that people's feelings about outcomes are linked to their assessments of their fairness. The importance of evaluations of outcome fairness ('distributive justice') has been compellingly demonstrated in the literature on the distribution of rewards in work settings, where equity is a central principle of justice. This literature shows that people are most satisfied when they receive equitable outcomes (Walster et al. 1978).

Justice theories also suggest that people react to the way that outcomes are distributed, that is, to the process of decision making, and to their treatment during that process ('procedural justice,' Thibaut and Walker 1975). Studies demonstrate that people's assessments of the fairness of procedures shape their satisfaction with their outcomes and with authorities and institutions (Tyler et al. 1997, Tyler and Smith 1997).

These findings link people's motivation to join social movements to their views about the fairness of social outcomes and procedures. In particular, they suggest that people react negatively to experiencing unfair procedures. However, the experience of injustice can potentially lead to either individual or collective actions: people can respond to perceived injustice by acting as an individual, or they can join a group. Only those actions that involve a collective response to injustice are relevant to the issue of social movements.

2. When do People Act Collectively?

Early research on both distributive and procedural justice focused primarily upon the feelings of the individuals who experienced justice or injustice. On this level, there was widespread evidence that justice played a central role in shaping people's feelings and behaviors. Further, people both reacted to their own experiences of justice or injustice and acted to provide justice to other people when they observed injustice. However, most of this research did not acknowledge that justice can also be conceptualized and responded to on the level of groups, not individuals. People can think about injustice in terms of unfairness to groups and they can respond to injustice on a group level. The perception of group level injustice is the key to the development of social movements.

The first crucial insight is that people can think about injustice in terms of unfairness to groups. This insight is contained in the early distinction between egoistical and fraternal deprivation (Runciman 1966). Instead of being concerned with personally receiving fairness, people can judge the fairness received by groups. Psychologists have increasingly recognized that group memberships are an important element in people's self-definitions (Hogg and Abrams 1988) and this has heightened attention to people's reactions to the experiences of the groups to which they belong, as well as to their own personal experiences.

The key issue is how people interpret experiences, that is, their view about the reason for the injustices they experience or observe. If people feel that the injustice is occurring to someone as an individual person, either themselves or someone else, and is due to their personal actions, they will respond personally. If they feel that the injustice is occurring to someone due to their membership in particular groups, or to all the people who share a common group membership, they will respond collectively.

Social identity theory argues that the way people interpret their experience is determined by how they

think about people. In particular, people are influenced by the degree to which they construct their sense of self in terms of the groups to which they belong (Hogg and Abrams 1988). To the degree that people think of themselves or others in personal terms, they interpret experiences as reflecting unique personal characteristics and behaviors, and they think about fairness in personal terms. To the degree that they think of people in terms of the group(s) to which they belong (their 'social' self), they interpret experiences as reflecting treatment as a member of groups and think about justice in group terms.

Studies confirm that those with strong social-selves are more likely to interpret experiences as being shaped by others attitudes toward the groups to which they and others belong, and to respond to injustice by engaging in collective behavior (Grant and Brown 1995, Kelly and Kelly 1994). So, thinking about injustice in collective terms leads to acting collectively in response to injustice, for example, by joining social movements.

Decisions about whether to respond to injustice individually or collectively are not only shaped by judgments about why injustice is occurring. Choices among possible behavioral responses are also influenced by people's judgments about the intergroup situation (Ellemers 1993). For example, people are influenced by their assessments of the permeability of group boundaries. If people believe that it is possible for individuals to move from low status groups to higher status groups, they are more likely to act as individuals. If people believe that the group's boundaries are not permeable, they are more likely to act collectively to raise the status of their group. Interestingly, studies suggest that very few low status group members need to be successful for people to view the group's boundaries as sufficiently permeable to justify individual, as opposed to collective, action (Wright et al. 1990).

People are also influenced by the perceived stability of group status. When they believe that group status can change, people are more likely to act collectively on behalf of the group. People are similarly affected by their views about the legitimacy of the status of existing groups. If people view current social arrangements as illegitimate, they are more likely to engage in collective action to change them (Major 1994).

Finally, people are influenced by pragmatic concerns. They respond to their sense of the probable gains and losses associated with various types of actions (Klandermans and Oegema 1987). When people decide that injustice has occurred, their general behavioral reaction is to do nothing. This inaction reflects the real risks associated with confronting powerful others, and an objective recognition of the low likelihood of success. On the other hand, risk is not enough of an explanation for people's behavior. For example, although the willingness to engage in collective action is influenced by the likelihood of success (Klandermans 1997), studies of collective action suggest that people often engage in collective actions against injustice even when the likelihood of success is small.

Similar factors shape people's motivation to engage in groups for positive, identity based, reasons. If people feel that their identity is firmly rooted in their group membership (stability), they will act to try to build the stature of their group through collective action. An example is ethnicity, which is often linked to physical appearance. If people feel that their status is not changeable, they are more likely to identify with that group, and to work to improve its status. Of course, such identification can also be strong when physical appearance is not involved. For example, many people identify strongly with being gay, even though they could 'pass for straight.' However, when people are members of a low status group they are more strongly tempted to leave the group as an individual and join a higher status group, instead of defining themselves in terms of group membership.

3. *Types of Collective Responses to Injustice*

Two dimensions distinguish among collective responses to injustice. One dimension is sustained vs. temporary action. The other dimension is normative vs. nonnormative.

Many of the most dramatic and widely seen collective responses to injustice are temporary events. Riots have this quality. They erupt quickly in response to some instance of injustice that reflects underlying issues of injustice. Riots typically spread rapidly and widely, and end quickly. They are not sustained and have little formal structure, or even a clear agenda beyond responding to an injustice. In contrast, social movements have leadership structures, political agendas, and long-term plans. Such movements may engage in actions such as violent protests, but these actions are typically planned and in the service of long-term goals.

The second dimension of collective movements distinguishes normative vs. non-normative responses to injustice. People can work for groups within the context of social norms in the political process to get their candidate elected. They can also step outside of social norms to engage in illegal and even violent social protests.

4. *Summary*

The psychological perspective on social movements examines the factors that encourage people to join and participate in social movements. Studies suggest that the experience of injustice is key to the development of a personal sense of grievance. When injustice is viewed as being linked to group membership, it is likely to

lead to participation in collective actions, such as social movements.

See also: Attitudes and Behavior; Civil Rights Movement, The; Conflict and Conflict Resolution, Social Psychology of; Ethnic and Racial Social Movements; Feminist Movements; Gay/Lesbian Movements; Intergroup Relations, Social Psychology of; Justice: Social Psychological Perspectives; Labor Movements and Gender; Labor Movements, History of; Media and Social Movements; National Liberation Movements; New Religious Movements; Parties/Movements: Extreme Right; Peace Movements; Peace Movements, History of; Right-wing Movements in the United States: Women and Gender; Science and Social Movements; Social Categorization, Psychology of; Social Identity, Psychology of; Social Movements and Gender; Social Movements: Environmental Movements; Social Movements, Geography of; Social Movements, History of: General; Social Movements: Resource Mobilization Theory; Social Movements, Sociology of; Social Psychology; Stigma, Social Psychology of; Youth Movements; Youth Movements, History of

Bibliography

Davies J C 1962 Toward a theory of revolution. *American Sociological Review* **27**: 5–9

Ellemers N 1993 The influence of socio-structural variables on identity management strategies. *European Review of Social Psychology* **4**

Grant P R, Brown R 1995 From ethnocentrism to collective protest. *Social Psychology Quarterly* **58**: 195–212

Gurr T R 1970 *Why Men Rebel*. Princeton University Press, Princeton, NJ

Hogg M A, Abrams D 1988 *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*. Routledge, New York

Kelly C, Kelly J 1994 Who gets involved in collective action?. *Human Relations* **47**: 63–88

Klandermans B 1989 Grievance interpretation and success expectations. *Social Behavior* **4**: 113–25

Klandermans B 1997 *The Social Psychology of Protest*. Blackwell, Cambridge, UK

Klandermans B, Oegema D 1987 Potentials, networks, motivations, and barriers: Steps toward participation in social movements. *American Sociological Review* **52**: 519–31

Major B 1994 From social inequality to personal entitlement: The role of social comparisons, legitimacy appraisals, and group membership. *Advances in Experimental Social Psychology* **26**: 293–355

Runciman W G 1966 *Relative Deprivation and Social Justice*. University of California Press, Berkeley, CA

Simon B, Loewy M, Sturmer S, Wever U, Freytey P, Habig C, Kempmeier C, Spahlinger P 1998 Collective identification and social movement participation. *Journal of Personality and Social Psychology* **74**: 646–58

Suls J, Wills T A 1991 *Social Comparison*. Erlbaum, Hillsdale, NJ

Tajfel H 1982 *Human Groups and Social Categories*. Cambridge University Press, New York

Thibaut J, Walker L 1975 *Procedural Justice*. Erlbaum, Hillsdale, NJ

Tyler T R, Boeckmann R J, Smith H J, Huo Y J 1997 *Social Justice in a Diverse Society*. Westview, Boulder, CO

Tyler T R, Smith H J 1997 Social justice and social movements. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*. Vol. 2, pp. 595–632

Walster E, Walster G W, Berscheid E 1978 *Equity: Theory and Research*. Allyn and Bacon, Boston

Wright S C, Taylor D M, Moghaddam F M 1990 Responding to membership in a disadvantaged group. *Journal of Personality and Social Psychology* **58**: 994–1003

T. Tyler

Social Movements: Resource Mobilization Theory

Resource mobilization theory (RMT) is premised on the idea that the central factor shaping the rise, development, and outcome of social movements is resources. 'Resource' here is taken broadly to mean any social, political, or economic asset or capacity that can contribute to collective action. Thus, solidarities and cultural outlooks of groups that enable them to act collectively as well as individual resources, such as communication skills, discretionary time-schedules, and independence from negative constraints, as well as organizational resources (e.g., leadership, meeting and communication facilities, and an *esprit de corps*) are relevant. Drawing on a rational choice approach to collective action, RMT views social movements as collective attempts to bring about change in social institutions and the distribution of socially valued goods. Typically hailed as the dominant paradigm in contemporary social movement research, some have contended that it is opposed by a rival 'new social movement' approach (Dalton and Kuechler 1990) but recent developments (discussed below) have shown that most arguments about the formation of collective identities and the role of values and social networks in collective action can be integrated into RMT. Research informed by RMT has focused on three major issues: (a) the microprocesses giving rise to individual participation; (b) organizational processes shaping mobilization; and (c) the political opportunities that guide social movement development and outcomes.

1. Collective Goods and Individual Participation

Early RMT argued that grievances stemming from structural strains and relative deprivation are secondary or possibly irrelevant to movement partici-

pation and that what changes, giving rise to social movements, is the availability of resources (McCarthy and Zald 1987, Jenkins and Perrow 1977). This assumption, however, has been questioned as researchers have shown that grievances in the sense of articulated collective claims about social injustice, including suddenly imposed grievances (Walsh and Warland 1983), are critical to individual participation. Drawing on rational choice assumptions, some point to the utility calculations of participants, arguing that they respond to perceived costs and benefits (Opp 1989), including the incentives of solidarity and self-esteem flowing from acting on principle. A major focus of debate has been Olson's (1971) theory of collective action according to which selective incentives (i.e. individual benefits) are critical to participation while collective goods *per se* create free riding. Arguing against this selective incentive thesis, Klandermans (1995) shows that most free riding is due to ignorance, high perceived costs in terms of work and family obligations, weak integration into interpersonal networks which provide both information and solidary incentives, and the perceived irrelevance of participation. Marwell and Oliver (1993) argue that free riding is an option only when individuals perceive that their contributions to the collective good follows a decelerating production function (i.e., each contribution makes others' subsequent contributions less worthwhile). Mass mobilization follows an accelerating production function (i.e., each contribution makes the next one more worthwhile). Hence, there may be large start-up costs to collective action but there is no point of diminishing returns. Thus, the key questions are expectations about the number of likely participants and the likelihood that the goal will be achieved.

RMT argues that collective identity and solidarity stemming from interpersonal networks provide for participation. Marwell and Oliver (1993) contend that organizers (or movement entrepreneurs) pay much of the start-up costs by building solidary networks, developing injustice frames (Snow et al. 1986), and creating collective perceptions of an accelerating production function. Klandermans (1997) shows that both selective and collective incentives matter with some movements depending more on selective incentives and others on collective commitments to movement goals. Thus, the women's movement and unions rely more on selective incentives while purposive commitments are paramount to the peace movement.

2. *Organizational Processes*

Zald and McCarthy (1987) argue that contemporary movements are professionalized, led and maintained by paid full-time staff with minimal 'grassroots' support. 'Conscience constituents' in the sense of sympathetic bystanders, wealthy patrons, and institutional sponsors are the major contributors to these movements. 'Speaking for' rather than mobilizing the aggrieved, grassroots grievances are seen as secondary if not irrelevant to these movements. Mass media (including direct mail solicitation and the Internet) has lowered organizing costs and the development of professional advocacy careers and training schools provide a ready supply of movement entrepreneurs. Thus, there has been a proliferation of professional movements, as in the environmental movement, sections of the women's movement, the peace and human rights campaigns, and other 'public interest' causes (Berry 1997). Drawing on organizational ecology theory, the increased density of social movement organizations (or SMOs) creates organizational survival problems with increased SMO mortality (Minkoff 1995). Similarly, issue attention cycles limit the ability of professional movement entrepreneurs to mobilize external support, forcing them to migrate among causes as public attention.

This argument, however, has been criticized for overstating the importance of professionalization and misdiagnosing its origins. Many contemporary movements are based indigenously on the voluntary labor of direct beneficiaries who are integrated by loosely structured networks, strong collective identities, and volunteer organizers. Morris (1984) shows that the Southern civil rights movement was centered organizationally in African–American churches and organized by a loosely knit coalition of ministers and college students. While external institutional resources may be central to the professionalization of movements, these contributions are often spurred by indigenous protests. Thus, the civil rights protests mobilized by the more militant groups spurred foundations and government officials to donate to the more moderate SMOs, creating a 'radical flank' effect (Haines 1986). Professionalization did not, however, demobilize these movements but rather channeled them into more moderate institutional activities, including the implementation of public policy gains (Jenkins and Eckert 1986). Thus, the professional model is most relevant for post-movement SMOs that have become institutionalized interest associations and 'public interest' movements that lack cohesive constituencies, such as the environmental, consumer, and general human rights movements.

3. *Political Opportunities*

A core rational choice idea is that participants respond to perceived costs and the likelihood of success. Thus, movement development should be shaped by political opportunities. Drawing on Piven and Cloward's (1977) thesis that elite divisions legitimize movement demands and thus facilitate protest and Tilly's (1978)

arguments about elite-challenger coalitions that lead to revolutionary situations, McAdam (1982) showed that competition between Democratic and Republican Presidential candidates for the African–American vote facilitated the rise of the civil rights movement. Jenkins (1985) showed that a dominant left-center governing coalition removed political obstacles to farm worker unionization and strengthened the institutional sponsors of the United Farm Worker Union, thus giving it more organizing space. Kriesi et al. (1995) show in a comparison of Western Europe that decentralized federal states, left-center governments, and strong alliances with unions facilitate the rise of the environmental and women's movements but opportunities mattered little for the more culturally-oriented homosexual and counter-cultural movements. Distinguishing between 'early riser' and 'late riser' movements that develop within a general protest cycle, Tarrow (1998) argues that 'late risers' are more successful because they can benefit by the policy precedents and resources created by 'early riser' challenges.

In a widely cited analysis, Gamson (1975/1990) argues that opportunities along with strategies are central to movement outcomes. Movements that 'think small' in terms of narrow incremental gains, use formal organization and thus avoid schisms, have allies, use unruly and violent tactics, and organize during political crisis periods are more likely to secure political acceptance and tangible benefits. By contrast, movements that attempt to overthrow their targets, rely on decentralized informal organization, lack political allies, use conventional tactics, and organize during periods of stability are less successful.

Subsequent research has focused largely on small-scale policy changes, such as the passage of favorable legislation and public expenditure showing that the resources under movement control (especially disruptive protest) combined with political opportunities are central to bringing about favorable policy outcomes (Jenkins 1985, Amenta 1998). Cress and Snow (2000) show that combinations of favorable strategies and opportunities are critical to the policy victories of homeless movements in US cities. Burstein (1985) argues that favorable public opinion is the key factor in equal employment legislation with nonviolent protest serving merely as a catalyst for greater public support. However, violent and highly disruptive protest often creates public opinion backlashes, making it more difficult for movements to gain favorable policies and raising the possibility of political repression.

It is also important to examine the broader impact of movements on culture and social institutions. Thus, expanding on Gamson's (1975/1990) typology of movement 'acceptance' and 'tangible gains,' movements also create new cultural codes and collective identities, such as in the women's and counter-cultural youth movements. The women's movement made it more legitimate for women to pursue full-time careers, to treat motherhood as optional, and promoted changes in social institutions from schools to collegiate sports and family practices that insured broader social inclusiveness. Movement participation also has profound biographical consequences, creating 'bridge-burning' experiences that channel activists into specific adult careers and transform their personal lives. Thus, McAdam (1988) shows that the adult lives of the white 'summer of 1964' student activists who went South to aid the civil rights movement were changed dramatically, leading most to embark on careers in the law, academe, and social work that allowed them to act on their political activist beliefs and leading to life course experiences, such as lower marital stability, geographic mobility, and lower income.

A major problem in inferring the effects of social movements is distinguishing between authentic movement impacts vs. ongoing social changes that are contributing simultaneously to both movement mobilization and changes reflecting the proclaimed goals of the movement. Most research is 'movement-centered' in that it takes the activities of the movement as central and focuses on the ostensive impact (e.g., public policy changes) of movement activity. It thus neglects the broader context in which movements emerge. Some contend, however, that these changes would have occurred without the movement or that, by working through movement activity, these deeper social changes are more important. Thus, for example, the changes in work and family typically attributed to the women's movement are also rooted in deeper cultural changes emphasizing individual autonomy and in the structural growth of service sector employment where most women have found work. Thus, while the women's movement may be central in these changes, evaluating how important it is relative to broader social changes that have also facilitated the women's movement is complex and difficult. Another complexity is that movement impact is often indirect and unintended. Thus, the women's movement encouraged more women to go into professional careers, which reinforced the number of feminist activists and legislators, which in turn strengthened pressures for feminist legislation, and so on. Similarly, social revolutions rarely have the outcomes intended by revolutionary leaders but, nonetheless, may have major unintended impacts on people's lives and the societies in which they live. Thus, RMT currently is being expanded by research on movement impact as well as by efforts to assess the social and cultural processes involved in mobilization.

See also: Action, Collective; Collective Behavior, Sociology of; Leadership in Organizations, Sociology of; Organizations, Sociology of; Political Sociology; Social Movements, Sociology of

Bibliography

Amenta E A 1998 *Bold Relief*. Princeton University Press, Princeton, NJ

Berry J M 1997 *The Interest Group Society*, 3rd edn. Longman, New York
Burstein P 1985 *Discrimination, Jobs and Politics*. University of Chicago Press, Chicago
Cress D M, Snow D A 2000 The outcomes of homeless mobilization. *American Journal of Sociology* **105**: 1063–1104
Dalton R J, Kuechler M 1990 *Challenging the Political Order*. Oxford University Press, New York
Gamson W A 1975/1990 *The Strategy of Social Protest*. Dorsey, Homewood, IL
Haines H H 1986 *Black Radicals and the Civil Rights Mainstream, 1954–1970*. University of Tennessee Press, Knoxville, TN
Jenkins J C 1985 *The Politics of Insurgency: The Farm Worker Movement in the 1960s*. Columbia University Press, New York
Jenkins J C, Eckert C M 1986 Channeling black insurgency: Elite patronage and the channeling of social protest. *American Sociological Review* **51**: 812–29
Jenkins J C, Perrow C 1977 Insurgency of powerless: Farm worker movements. *American Sociological Review* **42**: 249–68
Klandermans B 1995 *The Social Psychology of Protest*. Blackwell, Cambridge, MA
Kriesi H P, Koopmans R, Duyvendak J W, Giugni M G 1995 *New Social Movements in Western Europe: A Comparative Analysis*. University of Minnesota Press, Minneapolis, MI
Marwell G, Oliver P 1993 *The Critical Mass*. Cambridge University Press, Cambridge, UK
McAdam D 1982 *Political Process and the Development of Black Insurgency, 1930–1970*. University of Chicago Press, Chicago
McAdam D 1988 *Freedom Summer*. Oxford University Press, New York
Minkoff D C 1995 *Organizing for Equality*. Rutgers University Press, New Brunswick, NJ
Morris A D 1984 *The Origins of the Civil Rights Movement*. Oxford University Press, Oxford, UK
Olson M 1971 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
Opp K D 1989 *The Rationality of Political Protest*. Westview, Boulder, CO
Piven F F, Cloward A 1977 *Poor People's Movements*. Pantheon, New York
Snow D A, Rochford B, Worclen S, Benford R D 1986 Frame alignment processes, micromobilization and movement participation. *American Sociological Review* **51**: 464–81
Tarrow S G 1998 *Power in Movement*. Cambridge University Press, Cambridge, UK
Tilly C 1978 *From Mobilization to Revolution*. Addison-Wesley, Reading, MA
Walsh E J, Warland R H 1983 Social movement involvement in the wake of a nuclear accident. *American Sociological Review* **48**: 764–80
Zald M N, McCarthy J D 1987 *Social Movements in an Organizational Society*. Transaction, New Brunswick, NJ

J. C. Jenkins

Social Movements, Sociology of

For the whole set of actions and events subsumed under the heading of social movements, the problem of boundary demarcation is not to be taken lightly. Social movements unquestionably are, as Hanspeter Kriesi (1988, p. 350) puts it, 'elusive phenomena with unclear boundaries in time and space.' Furthermore, the question is not a mere empirical one to be solved, in each case, by deciding when a social movement begins and when it ends, or by circumscribing the social field in which it takes place and evolves. Social movements are also hard to grasp conceptually. First of all, even if a social movement as such cannot be equated with an organized group (or several ones), it has distinctively collective features: in this respect, Heberle (1951) had grounds for defining it 'as a kind of social collective.' Second, the scholar must be aware of two pitfalls: on the one hand, it is mistaken to consider a social movement to be a self-perpetuating crowd, for a crowd totally lacks mechanisms for sustaining association between people; on the other hand, it is an unduly narrow view of social movements to conceive them as necessarily linked to class-based actions, although the history of the concept can, as we will see, account for this usage. Finally, drawing the line between social movements and other close phenomena is, as a general rule, difficult, and that is why Turner and Killian ([1957] 1987) characterized these phenomena as 'quasi movements.' A classical instance of these borderline cases is to be found in some religious movements, such as messianic and millenarian sects: at the beginning of the twenty-first century, it is generally agreed upon that they both fall within the ambit of social movements when they set up meetings for voicing protest and aim at establishing a new type of social order.

What emerges from this view is perhaps the main criterion for defining a social movement, that is, its critical relation to social change. Herbert Blumer (1946, p. 199) aptly dwelt on this point when he suggested that social movements should be viewed 'as collective enterprises to establish a new order of life,' and his concise wording may still be used as a starting-point. But social movements may differ both in the direction and the nature of the change aimed at: some act 'to promote a change,' some others 'to resist' it, in the words of Turner and Killian ([1957] 1987); and the change claimed sometimes is a partial one, and sometimes implies a transformation of the whole social order.

A second distinctive feature of social movements is the use of uninstitutionalized means by participants, at least in some occasions and situations. A third one is also worth noting: the actions of a social movement are protest-oriented and even contest-oriented; therefore they have political relevance. We are now in a position to state an elementary, although tentative, definition of a social movement: it is 'a collective enterprise of protest and contention which aims at bringing about changes of varying importance in the social and/or political structure by the frequent but not necessarily exclusive use of uninstitutionalized means' (Chazel 1992, p. 270).

1. From the Formulation of the Idea to the Development of a Self-contained Field

The term *soziale Bewegung* was first used and coined by the German scholar, Lorenz von Stein ([1850] 1964) in his pioneer study, first published in German as *Geschichte der sozialen Bewegung in Frankreich von 1789 bis* auf *unsere Tage*. Rejecting Hegel's idealism, he did not study socialism and communism as forms of social thought, but regarded them as expressing the strivings of the industrial proletariat toward economic and political power; in his view, they were an essential development of the period and it is why, for him, the social movement meant the efforts of the rising industrial working class at organizing and sustaining mass action. Although Marx and Engels shared many interests with Stein, they did not themselves use the term. But Stein's conception remained influential until the beginning of the twentieth century. In Werner Sombart's ([1896] 1909, p. 2) wording, the social movement is still defined as 'the conception of all the attempts at emancipation on the part of the proletariat.'

Rudolf Heberle is to be credited with the discarding of this exclusive identification of the notion with proletarian movements in industrial society: the definition must be enlarged to take account of such important phenomena as peasant movements, Fascism, and National Socialism, or independence uprisings against colonial powers. It is to be noted that Heberle is still under the restraining influence of the traditional conception, when he states that 'the main criterion of a social movement is [to] aim [at] bring[ing] about fundamental changes in the social order, especially in the basic institutions of property and labor relationships' (1951, p. 6) or when he draws a somewhat arbitrary distinction between 'genuine social movements' and more or less short-lived 'protest movements.' Yet Heberle's work must be acknowledged as a landmark in the field, paving the way for the systematic study of social movements: he also points out their political implications and relevance, and therefore subtitles his book, *Social Movements*, as *An Introduction to Political Sociology*.

However original his work was, Heberle, belonging to a generation of political exiles, was the heir to European social theory, and research in social movements mostly developed along other lines in the USA. Social movements were to be conceived as one major form of collective behavior which Robert Park, who coined the term, defined as 'the behavior of individuals under the influence of an impulse that is common and collective, an impulse, in other words, that is the result of social interaction' ([1921] 1969). This notion has broad applicability and, while rather denoting for Park an approach to social phenomena, it quickly became a label for a distinctive field of investigation.

Some valuable studies and surveys of social movements, such as those, already mentioned, of Blumer, Turner, and Killian, have been worked out within this framework. But Neil Smelser's (1962) *Theory of Collective Behavior*, which is the most ambitious attempt at building up an elaborate theory, cannot be categorized in such a straightforward way, because he approaches the empirical phenomena subsumed under the field of collective behavior from a quite different theoretical perspective, the Parsonian hierarchy of 'components of social action.' By using criteria derived from this conceptual frame of reference, Smelser draws a demarcation line between 'panics,' 'crazes,' and 'hostile outbursts' on the one hand and social movements on the other, while making a further distinction between 'norm-oriented movements' and 'value-oriented movements.' This latter distinction, because it proceeds from an analytical basis, is more significant than purely descriptive ones, and it may be quite appropriate for differentiating, within a large movement, between a reform-oriented trend and a more radical one. Yet participants to social movements often account for their normative commitments by referring to widespread values. Discriminated as they are by their relation to a distinctive component of social action, these different kinds of collective behavior share common attributes, as Smelser makes clear in his formal characterization. All episodes of collective behavior imply 'an uninstitutionalized mobilization for action in order to modify one or more kinds of strain on the basis of a generalized [belief]' (Smelser 1962, p. 71).

Two major features of this conception are to be pointed out: the sharp contrast between the uninstitutionalized character of collective behavior and established behavior and Smelser's emphasis on beliefs of a very peculiar type, because they entail a 'process of short-circuiting,' that is, 'the jump from extremely high levels of generality to specific, concrete situations,' (p. 82) and these are the very points which will later on give rise to much controversy.

The irony of an unexpected outcome also is worth noticing: students of collective behavior were so keen on bringing out the salient characteristics of social movements that this topic gradually became a teaching and research area in itself. A quite significant instance of this development can be found in Robert Faris's *Handbook of Modern Sociology* (1964), in which the chapter on 'Social Movements,' written by Lewis Killian, is next to the one on 'Collective Behavior' by Ralph Turner.

2. New Emphases in Theory and Research

The turbulent 1960s aroused much interest in social movements, and many scholars did not agree with Smelser's statement that social movements are, as any other kind of collective behavior, 'the action of the

impatient.' During this period important developments internal to the social sciences were also under way, with systematic attempts at tackling sociological problems in terms of the economic paradigm. Facing troubled times that they had not predicted, as often happens, and grappling with theoretical puzzles of a new kind, sociologists and political scientists were thus induced to take a fresh and finally inventive approach to social movements.

In this process, Mancur Olson's *The Logic of Collective Action* (1965) was a seminal step. Olson drew attention to a vexing paradox, unnoticed in traditional theory: 'it does *not* follow, because all of the individuals in a group would gain if they achieved their group objective, that they would act to achieve that objective, even if they were all rational and self-interested' (p. 2); indeed, in very large groups individuals will not as a rule act to obtain a collective good. In Olson's view, there is only one way of preventing rational actors in such 'latent' groups from not mobilizing to provide a collective good: it consists of an incentive which discriminately operates upon some individuals in the group and which, by offering positive inducements to them or by coercing them, leads these individuals to act in the common interest. Actually, Olson's theoretical importance lies less in 'selective incentives,' which have been appraised by several scholars as a quite limited remedy, than in his skill in throwing light on the social dilemma of movement participation itself: mobilization never is to be taken for granted.

Thus the true significance of Olson's book rested in its setting a theoretical puzzle to a new generation of scholars: how and when are social actors able to overcome the dilemma of movement participation? These attempts at solving the dilemma resulted in what is now known as 'resource mobilization' theory. Yet, because there are different emphases within this theoretical approach, we can differentiate three distinctive ways of overcoming the dilemma, all of which ascribe weight to organization. The first one, which may be regarded as classic, has been elaborated by Anthony Oberschall in his book *Social Conflict and Social Movements* (1973). As Oberschall stresses in his 'sociological theory of mobilization,' collective protest is more likely to be present in a collectivity which has a strong organizational base, whether it is of a communal or of an associational kind, and which also is cut off from the rest of society, that is, is 'segmented.' A second way of getting over the dilemma was put forward by John McCarthy and Mayer Zald, who coined the convenient label, 'resource mobilization': they assert that particular attention must be paid to outside support, funding, and leadership; consequently, they dwell on the prominent part of 'conscience constituents' and 'adherents' on the one hand, of political entrepreneurs on the other, arguing that they 'turn Olson on his head' (1977). It is a theoretical point worth noticing, although it may be observed that Zald and McCarthy overrate the aggregate import of 'professional social movements' during the 1960s.

The third way of solving the dilemma perhaps is not so closely connected with resource mobilization theory; but the major questions at issue are the same. Charles Tilly (1978) puts as much emphasis as Oberschall or Zald and McCarthy on organization and interests, yet he stresses the political context in which mobilizations take place. Therefore he calls attention to emerging coalitions, of which a quite significant kind brings together 'members' of the 'polity' and 'challengers' aiming at entering it; and he takes account of the role of the government in this process, which can facilitate as well as repress an emergent mobilization.

Taken as a whole, this stream of thinking is to be considered 'revisionist,' as William Gamson aptly called it in his 'Introduction' to the collected essays by Zald and McCarthy (1987), and it is so, on two basic points. First, the advocates of the new approach assume that participation in social movements is to be explained in terms of rationality, as opposed to the prevailing view, which was exemplified by Smelser's emphasis on 'generalized beliefs.' Second, they are anxious to downgrade the psychosocial approach to social movements, as it was used and developed by collective behavior theorists. When at a later date Doug McAdam (1982) boldly asserts that social movements must be regarded as political phenomena, no more as psychosocial ones, he is in full agreement with this line of thought.

The theoretical import of McAdam's (1982) book has to be pointed out. In many respects he brings resource mobilization theory to completion, by joining the strand derived from Oberschall (although the relative weight of 'indigenous organizational strength' vs. outside support in the civil rights movement is a bone of contention between them), and the one derived from Tilly whose term, 'political process,' he took up. But on an important point, which resource mobilization theorists did not tackle, he makes a significant advance by drawing attention to potential participants' cognitive activity. Thus McAdam paves the way for attempts at multidimensional synthesis, in which, somewhat ironically, the psychosocial viewpoint will be awarded due consideration. For both of these achievements, his book can be regarded as a landmark.

3. *Contemporary Advances and Issues*

Any comment on the latest trends is open to question and correction but it is worth trying to outline the main line of development and, as already mentioned, the keynote of the times seems to lie in attempts at multidimensional synthesis.

Yet the path has gradually to be mapped out, and there are heuristic blind alleys, which are often linked

with some overemphasis on one analytical dimension or empirical type of social movements. Truly, social movements are in many respects significant political phenomena and, as mentioned, Tilly and McAdam opened up new vistas. But it does not follow that the whole approach must center in the concept of 'political opportunity structure,' which then is unduly stretched. As William Gamson and David Meyer aptly put it:

> the concept of political opportunity structure is ... in danger of becoming a sponge that soaks up virtually every aspect of the social movement environment ... Used to explain so much, it may ultimately explain nothing at all (1996, p. 275).

It is well known that with new waves of contention different problems are to be put on the research agenda. Thus the 'new social movement' school was quite justified in investigating the distinctive features of this type of movement. But the wrong inference that these movements were unparalleled was not always resisted: 'new social movements' perhaps are not so new as they appeared to be. Nevertheless this approach, by stressing the search for a collective identity, revives interest in the cultural dimension of social movements, as opposed to the resource mobilization stream of thinking, which had little regard for it; therefore the reconciliation of these quite divergent viewpoints is urgently needed.

The abundance of books on social movements during the 1990s, such as those edited by Aldon Morris and Carol Mueller (1992), Hank Johnston and Bert Kandermans (1995), McAdam et al. (1996), or the individual works by Sidney Tarrow (1994) and Klandermans (1997), along with the publishing of a specialized journal, *Mobilization*, first testifies to the institutionalization of an academic field. But their theoretical significance lies mainly in systematic attempts at integrating mobilizing structures and organizations, political opportunities, and 'framing' processes. A true synthesis seems to be under way, although it is not an easy job to articulate distinctive dimensions with one another.

A last point is worth noting: the clarifying of cognitive and discursive processes, whether in terms of causal attribution, as by Myra Ferree and Frederick Miller (1985), or in terms of 'frame alignment' and 'frame resonance,' as by David Snow and his associates (1986, 1988), is part and parcel of this development. This new interest in collective beliefs can be construed as Smelser's revenge, but it is an ambiguous one: as opposed to 'generalized beliefs,' the beliefs which are now to be investigated are reasonable ones and can be interpreted in terms of a rational, though not simply rationalistic, model.

See also: Action, Collective; Civil Rights Movement, The; Collective Behavior, Sociology of; Conflict Sociology; Feminist Movements; Gay/Lesbian Movements; Identity Movements; Participation: Political; Peace Movements, History of; Revolutions, Sociology of; Revolutions, Theories of; Social Change: Types; Social Movements: Environmental Movements; Social Movements, Geography of; Social Movements, History of: General; Social Movements: Psychological Perspectives; Social Movements: Resource Mobilization Theory; Youth Movements, History of

Bibliography

Blumer H 1946 Collective behavior. In: Lee A M (ed.) *New Outline of the Principles of Sociology*. Barnes and Noble, New York, pp. 166–222

Chazel F 1992 Mouvements sociaux. In: Boudon R (ed.) *Traité de sociologie*. Presses Universitaires de France, Paris, pp. 263–312

Chazel F 1997 Les ajustements cognitifs dans les mobilisations collectives. In: Boudon R, Bouvier A, Chazel F (eds.) *Cognition et sciences sociales*. Presses Universitaires de France, Paris, pp. 193–206

Faris R E L (ed.) 1964 *Handbook of Modern Sociology*. Rand McNally, Chicago

Ferree M M, Miller F D 1985 Mobilization and Meaning: Toward an integration of social psychological and resource perspectives on social movements. *Sociological Inquiry* **55**: 38–61

Gamson W, Meyer D 1996 Framing political opportunity. In: McAdam D, McCarthy J D, Zald M N (eds.) *Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures and Cultural Framings*. Cambridge University Press, Cambridge, UK, pp. 275–90

Gusfield J R 1969 *Symbolic Crusade*. University of Illinois Press, Urbana, IL

Heberle R 1951 *Social Movements: An Introduction to Political Sociology*. Appleton-Century-Crofts, New York

Hirschman A O 1970 *Exit, Voice and Loyalty*. Harvard University Press, Cambridge, MA

Johnston H, Klandermans B (eds.) 1995 *Culture and Social Movements*. University of Minnesota Press, Minneapolis, MN

Killian L M 1964 Social movements. In: Faris R E L (ed.) *Handbook of Modern Sociology*. Rand McNally, Chicago, pp. 426–55

Klandermans B 1997 *The Social Psychology of Protest*. Blackwell, Oxford

Kriesi H 1988 The interdependence of structure and action: Some reflections on the state of the art. In: Klandermans B, Kriesi H, Tarrow S (eds.) *From Structure to Action: Comparing Social Movement Research across Cultures*. JAI Press, Greenwich, CT, pp. 349–68

McAdam D 1982 *Political Process and the Development of Black Insurgency*. University of Chicago Press, Chicago

McAdam D, McCarthy J D, Zald M N (eds.) 1996 *Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures and Cultural Framings*. Cambridge University Press, Cambridge, UK

McCarthy J D, Zald M N 1987 Resource mobilization and social movement: a partial theory. In: Zald M N, McCarthy J D (eds.) *Social Movements in an Organizational Society: Collected Essays*. Transaction Books, New Brunswick, NJ

Morris A D, Mueller C M (eds.) 1992 *Frontiers in Social Movement Theory*. Yale University Press, New Haven, CT
Oberschall A 1973 *Social Conflict and Social Movements*. Prentice-Hall, Englewood Cliffs, NJ
Oberschall A 1993 *Social Movements: Ideologies, Interests and Identities*. Transaction Books, New Brunswick, NJ
Olson M 1965 *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge, MA
Park R [1921] 1969 Collective behavior. In: Park R, Burgess E W (eds.) *Introduction to the Science of Sociology*. Greenwood Press, New York, pp. 865–78
Smelser N J 1962 *Theory of Collective Behavior*. Free Press, New York
Snow D, Benford R D 1988 Ideology, frame resonance and participant mobilization. In: Klandermans B, Kriesi H, Tarrow S (eds.) *From Structure to Action: Comparing Social Movement Research across Cultures*. JAI Press, Greenwich, CT, pp. 197–217
Snow D, Rochford E B, Worden S K, Benford R D 1986 Frame alignment processes, micro-mobilization and movement participation. *American Sociological Review* **51**: 464–81
Sombart W [1896] 1909 *Socialism and the Social Movement*. Dent, London
Stein L von [1850] 1964 *The History of the Social Movement in France 1789–1850* [trans. K. Mengelberg]. Bedminster Press, Totowa, NJ
Tarrow S 1994 *Power in Movement: Social Movements and Contentious Politics*. Cambridge University Press, Cambridge, UK
Tilly C 1978 *From Mobilization to Revolution*. Addison-Wesley, Reading, MA
Touraine A 1988 *Return of the Actor: Social Theory in Postindustrial society*. University of Minnesota Press, Minneapolis, MN
Turner R H, Killian L M [1957] 1987 *Collective Behavior*, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ
Zald M N, McCarthy J D (eds.) 1987 *Social Movements in an Organizational Society: Collected Essays*. Transaction Books, New Brunswick, NJ

F. Chazel

Social Network Models: Statistical

Late twentieth-century developments in statistical models for social networks reflect an increasing theoretical focus in the social and behavioral sciences on the interdependence of social actors in dynamic, network-based social settings. As a result, a growing importance has been accorded the problem of modeling the dynamic and complex interdependencies among network ties and the actions of the individuals whom they link. The early focus of statistical network modeling on the mathematical and statistical properties of Bernoulli and dyad-independent random graph distributions has been replaced by efforts to construct theoretically and empirically plausible parametric models for structural network phenomena and their changes over time. These models in this article are based on graphs where the nodes are individual actors, which are 'linked' by lines representing relational ties (see *Graphical Models: Overview*).

1. Statistical Models for Social Networks

1.1 Notation

In most of the literature on statistical models for networks, the set $N = \{1, 2,\ldots, n\}$ of network nodes is regarded as fixed and the edges, or ties, between nodes are assumed to be random. The tie linking node i to node j $(i, j \in N)$ may be denoted by the random variable X_{ij}. In the simplest case of binary-valued random variables, X_{ij} takes the value 1 if the tie is present and 0 otherwise. Other cases of interest include:

(a) valued networks where X_{ij} is assumed to take values in the set $\{0, 1,\ldots, C-1\}$;

(b) multiple relational or multivariate networks, where the variable X_{ijk} represents the possible tie of type k from node i to node j (with $k \in R = \{1, 2,\ldots, r\}$, a fixed set of types of tie); and

(c) time-dependent networks, where X_{ijt} represents the tie from node i to node j at time t.

In each case, network ties may be *directed* (e.g., X_{ij} and X_{ji} are distinct random variables) or *nondirected* (e.g., X_{ij} and X_{ji} are not distinguished). The $n \times n$ array $\boldsymbol{X} = [X_{ij}]$ of random variables can be regarded as the adjacency matrix of a random (directed) graph on N; the state space of all possible realizations of these arrays may be denoted by Ω_n. The array $\boldsymbol{x} = [x_{ij}]$ denotes a realization of $\boldsymbol{X}$, with $x_{ij} = 1$ if there is an observed tie from node i to node j, and $x_{ij} = 0$ otherwise. In some cases, models may also refer to variables measured on actor attributes: for instance, the mth attribute $Z_i^{[m]}$ of actor i gives rise to a random vector $\boldsymbol{Z}^{[m]}$ with realization $\boldsymbol{z}^{[m]}$.

The adjacency matrix of an illustrative nondirected binary network is shown in Table 1. The data are from a study of interorganizational networks amongst mental health agencies conducted by Morrisey et al. (1994). The nodes of the network are 37 mental health agencies in one of the US cities in the Morrissey study; an edge links two agencies if they engage in regular and mutual exchange of information. The graphical representation of these data is shown in Fig. 1. We use this network to illustrate several of the distributions and models described below.

1.2 Bernoulli Graphs

A simple statistical model for a (directed) graph assumes a Bernoulli distribution, in which each edge,

Table 1
Adjacency matrix for a nondirected binary network.

```
0000000101100000010000000001000001110
0000000000100000000000000001000000000
0000000000000000000100000000000000000
0000000000000000000000000000000000000
0000000000000000100100000000010000010
0000000000000000000000000000000000000
0000000000000000000010111111000000000
1000000000100001010000000000000000100
0000000000000000000000000000000000000
1000000000101001000000000000000000100
1100000101000010000100000001100000000
0000000000001000000000000000000000000
0000000001010000000001000000000001000
0000000000000000000001001001000100000
0000000000100000000001101001000100000
0000000101000000000000000000000000000
0000100000000000000100000000010000000
1000000100000000000010000000000000100
0000000000000000000000000000000001000
0010100000100000110000001010000000100
0000001000000000000000101010000000000
0000000000001110000000111110000000100
0000001000000010000011001011010001100
0000001000000000000001001011010001000
0000001000000110000111110111000001100
0000001000000000000001001011000000000
0000001000000000000111111101000001100
1100001000100110000000111110000000000
0000000000100000000000000000000000000
0000100000000000100000110000000000000
0000000000000000000000000000000000000
0000000000000110000000000000000000000
0000000000000000000000000000000000000
1000000000001000001000111010000000000
1000000101000000010101101010000000000
1000100000000000000000000000000000000
0000000000000000000000000000000000000
```

or tie, is statistically independent of all others and governed by a theoretical probability P_{ij}. In addition to edge independence, simplified versions also assume equal probabilities across ties; other versions allow the probabilities to depend on structural parameters (Frank and Nowicki 1993). These distributions have often been used as models (Bollobas 1985), but are of questionable utility due to the independence assumption. Bernoulli graph distributions are described at length by Erdos and Renyi (1960) and Frank (1981).

1.3 Dyadic Structure in Networks

Statistical models for social network phenomena have been developed from their edge-independent beginnings in a number of major ways. In one important series of developments, the p_1 model introduced by Holland and Leinhardt (1981), and developed by Fienberg and Wasserman (1981), recognized the theoretical and empirical importance of dyadic structure in social networks, that is, of the interdependence of the variables X_{ij} and X_{ji}. This class of Bernoulli dyad distributions and their generalization to valued, multivariate, and time-dependent forms, gave parametric expression to ideas of reciprocity and exchange in dyads and their development over time. The model assumes that each dyad (X_{ij}, X_{ji}) is independent of every other and, in a commonly constrained form, specifies:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \prod_{i<j} \exp\left\{\lambda_{ij} + \theta\left(\sum_{i \neq j} x_{ij}\right) + \rho\left(\sum_{i<j} x_{ij}x_{ji}\right) + \alpha_i\left(\sum_j x_{ij}\right) + \beta_j\left(\sum_i x_{ij}\right)\right\} \quad (1)$$

where θ is a density parameter; ρ is a reciprocity parameter; the parameters α_i and β reflect individual differences in expansiveness and popularity; and λ_{ij} ensures that probabilities for each dyad sum to 1. This model is log-linear (see *Multivariate Analysis: Discrete Variables (Loglinear Models)*).

An important elaboration of the p_1 model permits the dependence of density, popularity and expansiveness parameters on a partition of the nodes into subsets that can either be specified *a priori* (e.g., from hypotheses about social position) or can be identified *a posteriori* from observed patterns of node interrelationships (e.g., Wasserman and Faust 1994). The resulting stochastic blockmodels represent hypotheses about the interdependence of social positions and the patterning of network ties (as originally elaborated in the notion of blockmodel by White et al. 1976). A potentially valuable variant of these models regards the allocation of nodes to blocks as a latent classification; Snijders and Nowicki (1997) have provided Bayesian estimation methods for the case of two latent classes. Another valuable extension of the p_1 model is the so-called p_2 model (Lazega and van Duijn 1997). Like the p_1 model, the p_2 model assumes dyad independence, but it adds the possibility of arbitrary node and dyad covariates for the parameters of the p_1 model, and models the unexplained parts of the popularity and expansiveness parameters as random rather than fixed effects.

1.4 Null Models for Networks

Despite the useful parameterization of dyadic effects associated with the p_1 model and its generalizations, the assumption of dyadic independence has attracted

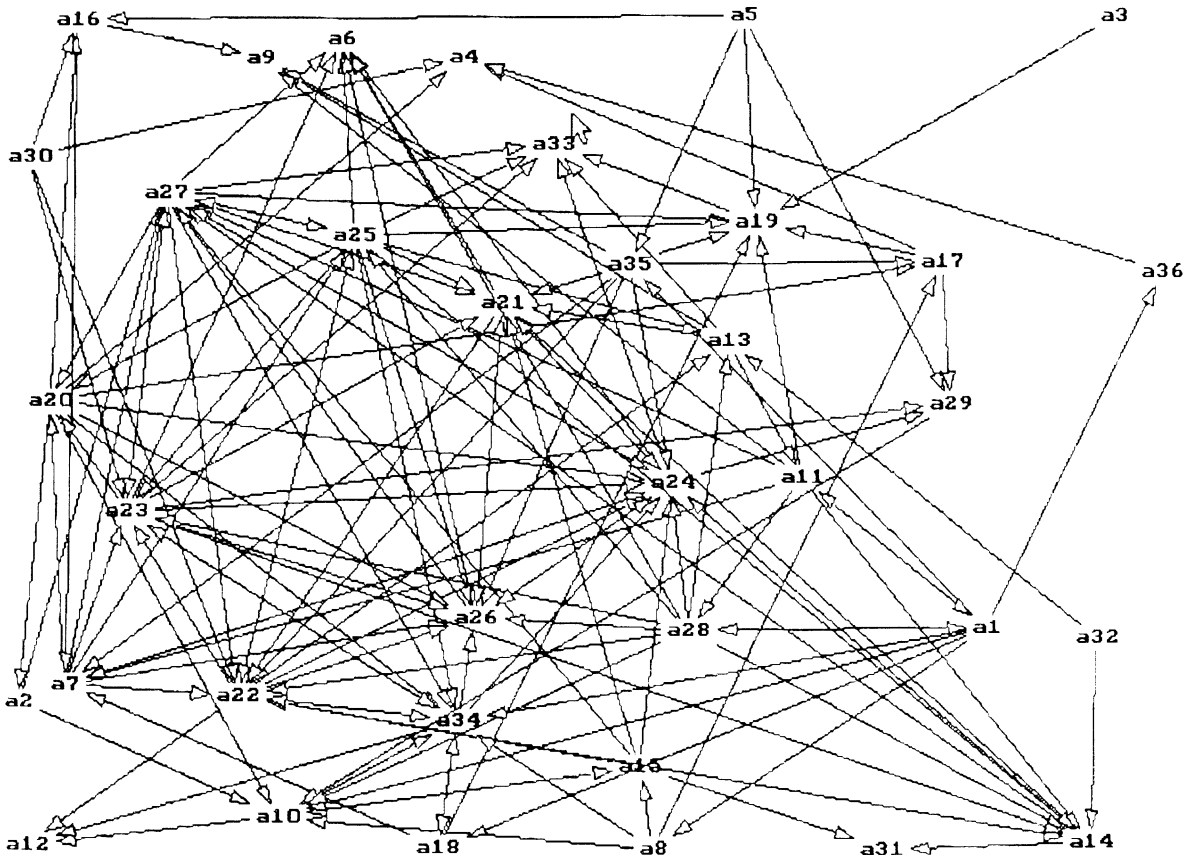

Figure 1
Graphical representation of interorganizational networks amongst mental health agencies in one US city (data from Morrissey et al. (1994)).

sustained criticism. Thus, another series of developments has been motivated by the problem of assessing the degree and nature of departures from simple structural assumptions like dyadic independence. Following a general approach exemplified by early work of Rapoport (e.g., Rapoport 1949), a number of conditional uniform random graph distributions were introduced as null models for exploring the structural features of social networks (see Wasserman and Faust 1994 for a historical account). These distributions, denoted by $U \mid Q$, are defined over subsets Q of the state space Ω_n of directed graphs, and assign equal probability to each member of Q. The subset Q is usually chosen to have some specified set of properties (e.g., a fixed number of mutual, asymmetric, and null dyads, as in the $U \mid MAN$ distribution of Holland and Leinhardt 1975). When Q is equal to Ω_n the distribution is referred to as the uniform (di)graph distribution, and is equivalent to a Bernoulli distribution with homogeneous tie probabilities. Enumeration of the members of Q and simulation of $U \mid Q$ is often straightforward (e.g., see Wasserman and Pattison, in press), although certain cases, such as the distribution that is conditional on the indegree $\left(\sum_j x_{ji}\right)$ and outdegreee $\left(\sum_j x_{ij},\right)$ of each node i in the network, require more complicated approaches (Snijders 1991, Wasserman 1977).

A typical application of these distributions is to assess whether the occurrence of certain higher-order (e.g., triadic) features in an observed network is unusual, given the assumption that the data arose from a uniform distribution that is conditional on plausible lower-order (e.g., dyadic) features (e.g., Holland and Leinhardt 1975). For example, the network displayed in Table 1 has 84 edges, and 333 and 69 triads with exactly 2 and 3 edges, respectively. If Q denotes the subset of complete Ω_{37}, with 84 edges, the expected number of triads of each kind is 250.1 and 50.8, and the observed values lie at the 0.003 and 0.005 percentiles of the $u \mid x_{++} = 84$ distribution, respectively. This general approach has also been developed for the analysis of multivariate social networks. The best-known example of this approach is probably the

quadratic assignment procedure or model (*QAP*) for networks (Hubert and Baker 1978). In this case, the association between two graphs defined on the same set of nodes is assessed using a uniform multivariate graph distribution that is conditional on the unlabelled graph structure of each univariate graph. In a more general framework for the evaluation of structural properties in multivariate graphs, Pattison et al. (in press) also introduced various multigraph distributions associated with stochastic blockmodel hypotheses.

1.5 Extra-dyadic Local Structure in Networks

A significant step in the development of parametric statistical models for social networks was taken by Frank and Strauss (1986) with the introduction of the class of Markov random graphs. This class of models permitted the parameterization of extra-dyadic local structural forms and so allowed a more explicit link between some important network conceptualizations and statistical network models. Frank and Strauss (1986) adapted research originally developed for modeling spatial dependencies (see *Spatial Statistical Methods*) among observations to graphs. They observed that the Hammersley-Clifford theorem provides a general probability distribution for $\boldsymbol{X}$ from a specification of which pairs (X_{ij}, X_{kl}) of edge random variables are conditionally dependent, given the values of all other random variables (see *Graphical Models: Overview*).

Define a dependence graph $\mathbf{D}$ with node set $N(\mathbf{D}) = \{(X_{ij}: i, j \in N, i \neq j\}$ and edge set $E(\mathbf{D}) = \{(X_{ij}, X_{kl}): X_{ij}$ and X_{kl} are assumed to be conditionally dependent, given the rest of $\boldsymbol{X}\}$. Frank and Strauss used $\mathbf{D}$ to obtain a model for $\Pr(\boldsymbol{X} = \boldsymbol{x})$, denoted p^* by Wasserman and Pattison (1996), in terms of parameters and substructures corresponding to cliques of $\mathbf{D}$. The model has the form

$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = p^*(\boldsymbol{x}) = (1/c)\exp\left\{\sum_{\mathbf{P} \subset \mathbf{N(D)}} \alpha_{\mathrm{P}} z_{\mathrm{P}}(\boldsymbol{x}\}\right\} \qquad (2)$$

where:

(a) the summation is over all cliques P of $\mathbf{D}$ (with a clique of $\mathbf{D}$ defined as a nonempty subset P of $N(\mathbf{D})$ such that $|P| = 1$ or $(X_{ij}, X_{kl}) \in E(\mathbf{D})$ for all $X_{ij}, X_{kl} \in P$);

(b) $z_{\mathrm{p}}(\boldsymbol{x}) = \prod_{\mathrm{x}ij \in \rho}^{\mathrm{x}_{ij}} x_{ij}$ is the (observed) network statistic corresponding to the clique ρ of $\mathbf{D}$; and

(c) $c = \Sigma_{\mathrm{x}} \exp\{\Sigma_{\mathrm{p}} \alpha_{\mathrm{p}} z_{\mathrm{p}}(\boldsymbol{x})$ is a normalizing quantity.

specification of a dependence graph raises a deep theoretical issue for network analysis: what constitutes an appropriate set of dependence assumptions? Frank and Strauss (1986) proposed a Markov assumption, in which $(X_{ij}, X_{kl}) \in E(\mathbf{D})$ whenever $\{i, j\} \cap \{k, l\} \neq \phi$. This assumption implies that the occurrence of a network tie from one node to another is conditionally dependent on the presence or absence of other ties in a local neighborhood of the tie. A Markovian local neighborhood for X_{ij} comprises all possible ties involving i and/or j. Any network tie is seen both as dependent on its neighboring ties and as participating in the neighborhood of other ties, so the resulting model embodies the assumption that social networks have a local self-organizing quality. In the case of Markovian neighborhoods, cliques in the dependence graph correspond to triadic and star-like network configurations. A homogeneity assumption, that model parameters do not depend on the specific identities of nodes but only on the network configurations to which they correspond, yields a model that expresses the probability of the network in terms of propensities for subgraphs of triadic and star-like structures to occur.

For example, a homogeneous Markov model for the network of Table 1 with parameters corresponding to edges, 2-stars, and triangles appears to be a substantial improvement in fit over a Bernoulli model with just an edge parameter, since the mean absolute residual for fitted edge probabilities drops from .220 to .175. Pseudolikelihood estimates for model parameters (Wasserman and Pattison 1996) are –3.31, 0.056, and 0.616 for edges, 2-stars, and triangles, respectively, suggesting that graphs with a large number of triangles are relatively more likely in the assumed homogeneous Markov p^* distribution than those with few.

p^* random graph models permit the parameterization of many important ideas about local structure in univariate social networks, including transitivity, local clustering, degree variability, and centralization. Valued (Robins et al. 1999) and multivariate (Pattison and Wasserman 1999) generalizations also lead to parameterizations of substantively interesting multirelational concepts, such as those associated with balance and clusterability, generalized transitivity and exchange, and the strength of weak ties.

1.6 Prospects for Structural Models

The p^* family of random graph models poses many interesting questions, both substantive and statistical. A major empirical question is the sufficiency of these local structural descriptions: do they suffice, or do they need to accommodate other characterizations of local dependence, more global considerations, other types of constraints (e.g., temporal and spatial), or variability in local structural dependencies? On the statistical side, major questions are raised by the complexity and rich parameterization of models that are deemed to be theoretically plausible. For instance, does a Markov chain Monte Carlo (see *Markov Chain Monte Carlo Methods*) maximum likelihood approach to parameter estimation offer a viable alternative to approximate methods such as pseudolikelihood esti-

mation (Strauss and Ikeda 1990)? How can these distributions best be simulated, and what are the convergence properties of various approaches, such as the Metropolis-Hastings algorithm?

2. Dynamic Models

Empirically and theoretically defensible parametric models for networks constitute one challenge for statistical modeling, but an arguably more significant challenge is to develop models for the emergence of network phenomena, including the evolution of networks and the unfolding of individual actions (e.g., voting, attitude change, decision-making) and interpersonal transactions (e.g., patterns of communication or interpersonal exchange) in the context of longer-standing relational ties. Early attempts to model the evolution of networks in either discrete or continuous time assumed dyad independence and Markov processes in time.

A step towards continuous time Markov chain models for network evolution that relaxes the assumption of dyad independence has been taken by Snijders (1996). These models incorporate both random change and change arising from individuals' attempts to optimize features of their local network environment (such as the extent to which reciprocity and balance prevail) and use the computer package *SIENA* for fitting. A tension function is proposed to represent the function that actors wish to minimize and Robbins-Munro procedures are used to estimate free parameters. The approach illustrates the potentially valuable role of simulation techniques for models that make empirically plausible assumptions; clearly, such methods provide a promising focus for future development.

The value of simulation has also been demonstrated by Watts and Strogatz (1998) and Watts (1999) in their analysis of the so-called small-world phenomenon (the tendency for individuals to be linked by short acquaintanceship network paths). They considered a probabilistic model for network change in which the probability of a new tie was determined, in part, by existing patterns of neighboring ties (a clustering component) and, in part, by a small random component. They used simulations to establish that only a small random component was needed for the simulated network to be likely to possess the small-world property.

3. Models Bridging Analytic Levels

A recurring theme among social theorists is the need for models that integrate the dynamic, interdependent, interpersonal, and cultural contexts in which social action occurs. The network modeling challenge posed by such framing of the nature of social processes is the development of models that represent interdependencies across several analytic categories (e.g., individuals, relational ties, settings, groups).

Snijder's (1996) model for network change represents one step in this direction; the class of social influence models represents another (Friedkin 1998). These models are exemplified by the network effect model for the mth attribute variable:

$$\mathbf{z}^{[m]} = \alpha \boldsymbol{x}\boldsymbol{z}^{[m]} + \boldsymbol{y}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where α is a network effects parameter, $\boldsymbol{y}$ is a matrix of exogenous attribute variables with corresponding parameter vector $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ is a vector of residuals.

A further step in the direction of modeling interdependent node attributes and network ties has been to generalize the dependence graph approach to p^* network models (Sect. 1.5) so as to include both node attribute variables ($\boldsymbol{Z}^{[m]}$) and relational variables ($\boldsymbol{X}$). This step is facilitated if directed dependencies are permitted in the dependence graph (i.e., dependencies in which one random variable is assumed to be conditionally dependent on a second, but the second not on the first). Robins et al. (in press) drew on graphical modeling techniques to construct models with directed dependencies among network ties and individual attributes. If the dependencies are directed from network ties to node attributes, models for social influence are obtained; with dependencies oriented from node attributes to network ties, social selection models result. Indeed, this generalization introduces substantial flexibility in model construction, with potential applications to joint influence and selection models, discrete time models for network change, network path models, and group-level attribute models.

The dependence approach of Sect. 1.5 has also been used to construct principled models for more general multiway relational data structures representing interdependent analytic categories. Examples to date have included bipartite (Faust and Skvoretz 1999) and k-partite (Pattison et al. 1998) graphs and illustrate how this general distribution has the capacity for a systematic and coherent approach to a variety of major network modeling challenges.

4. Conclusion

In addition to some of the specific future prospects already noted, a more general expectation is that the current classes of models are likely to be the foundation of substantial empirical and model-building activity in the early twenty-first century. From their early beginnings, statistical approaches to network modeling have

now reached a point where sustained interaction is likely to be empirical research that allows the development of greatly improved models for processes occurring on networks, with important applications in many fields of social and behavioral science, including epidemiology, organizational studies, social psychology, and politics.

For more information on these approaches to social networks analysis, please see Wasserman and Faust (1994) and Wasserman and Pattison (1996, in press).

See also: Graphical Models: Overview; Multivariate Analysis: Discrete Variables (Loglinear Models); Network Analysis; Network Models of Tasks; Networks and Linkages: Cultural Aspects; Networks: Social; Spatial Statistical Methods

Bibliography

Bollobas B 1985 *Random Graphs*. Academic Press, London

Erdos P, Renyi A 1960 On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**: 17–61

Faust K, Skvoretz J 1999 Logit models for affiliation networks. In: Becker M, Sobel M (eds.) *Sociological Methodology 1999*. Blackwell, New York, pp. 253–80

Fienberg S, Wasserman S 1981 Categorical data analysis of single sociometric relations. In: Leinhardt S (ed.) *Sociological Methodology 1981*. Jossey-Bass, San Francisco, pp. 156–92

Frank O 1981 A survey of statistical methods for graph analysis. In: Leinhardt S (ed.) *Sociological Methodology 1981*. Jossey-Bass, San Francisco, pp. 110–55

Frank O, Nowicki K 1993 Exploratory statistical analysis of networks. In: Gimbel J, Kennedy J W, Quintas L V (eds.) *Quo Vadis Graph Theory? A Source Book for Challenges and Directions*. North-Holland, Amsterdam

Frank O, Strauss D 1986 Markov graphs. *Journal of the American Statistical Association* **81**: 832–42

Friedkin N 1998 *A Structural Theory of Social Influence*. Cambridge University Press, New York

Holland P W, Leinhardt S 1975 The statistical analysis of local structure in social networks. In: Heise D R (ed.) *Sociological Methodology 1976*. Jossey-Bass, San Francisco, pp. 1–45

Holland P W, Leinhardt S 1981 An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association* **76**: 33–50

Hubert L J, Baker F B 1978 Evaluating the conformity of sociometric measurements. *Psychometrika* **43**: 31–41

Lazega E, van Duijn M 1997 Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data. *Social Networks* **19**: 375–97

Morrisey J P, Calloway M 1994 Local mental health authorities and service system change: Evidence from the Robert Wood Johnson Foundation Program on chronic mental illness. *The Milbank Quarterly* **72**(1): 49–80

Pattison P, Mische A, Robins G L 1998 The plurality of social relations: *k*-partite representations of interdependent social forms. Keynote address, Conference on Ordinal and Symbolic Data Analysis, Amherst, MA, Sept. 28–30

Pattison P E, Wasserman S 1999 Logit models and logistic regressions for social networks, II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology* **52**: 169–194

Pattison P, Wasserman S, Robins G L, Kanfer A M 2000 Statistical evaluation of algebraic constraints for social networks. *Journal of Mathematical Psychology* **44**: 536–68

Rapoport A 1949 Outline of a probabilistic approach to animal sociology, I. *Bulletin of Mathematical Biophysics* **11**: 183–196

Robins G L, Pattison P, Elliott P 2001 Network models for social influence processes. *Psychometrika* **66**: 161–190

Robins G L, Pattison P, Wasserman S 1999 Logit models and logistic regressions for social networks, III. Valued relations. *Psychometrika* **64**: 371–94

Snijders T A B 1991 Enumeration and simulation methods for 0–1 matrices with given marginals. *Psychometrika* **56**: 397–417

Snijders T A B 1996 Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology* **21**: 149–72

Snijders T A B, Nowicki K 1997 Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification* **14**: 75–100

Strauss D, Ikeda M 1990 Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association* **85**: 204–12

Wasserman S 1977 Random directed graph distributions and the triad census in social networks. *Journal of Mathematical Sociology* **5**: 61–86

Wasserman S, Faust K 1994 *Social Network Analysis: Methods and Application*. Cambridge University Press, New York

Wasserman S, Pattison P E 1996 Logit models and logistic regressions for social networks, I. An introduction to Markov random graphs and *p**. *Psychometrika* **60**: 401–25

Wasserman S, Pattison P E in press *Multivariate Random Graph Distributions*. Springer Lecture Note Series in Statistics, Springer-Verlag, New York

Watts D 1999 Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology* **105**: 493–527

Watts D, Strogatz S 1998 Collective dynamics of 'small-world' networks. *Nature* **393**: 440–2

White H C, Boorman S, Breigen R L 1976 Social structure from multiple networks, I. Blockmodels of roles and positions. *American Journal of Sociology* **81**: 730–80

P. Pattison and S. Wasserman

Social Networks and Fertility

Explanations offered for fertility decline in both developed and developing countries underwent fundamental changes in the last decades of the twentieth century. In addition to more refined economic reasoning and the introduction of cultural aspects, theories of the fertility transition now routinely reserve a place for diffusion effects. Such effects arise because individuals are themselves members of larger groups. The information that is known to group members, the choices they make, and the norms and values they adhere to, can all exert a powerful influence on

individual incentives to innovate. In settings in which fertility has been high, for example, innovation might take the form of adopting modern contraception and fertility limitation. Social interaction with group members represents one important avenue along which such innovative demographic behavior can diffuse: peers, community members or relatives may provide relevant information to potential adopters of family planning, and they also represent the institutional and social structures that may facilitate or constrain individual fertility or family planning decisions.

Theories of diffusion usually encompass two central elements. First, they emphasize a particular set of social interactions that are associated with the adoption of innovations, such as social networks, opinion leaders, public media, etc. Second, diffusion models have an explicit dynamic dimension in which the adoption of an innovation is often associated with 'contagion' or 'tipping point' phenomena, which can accelerate social change. Because fertility transitions are often rapid, irreversible, and only weakly associated with the specific socioeconomic conditions in a population, such diffusionist explanations have appealed to demographers ever since the European Fertility Project was carried out (Coale and Watkins 1986)

At the same time, diffusionist approaches have been criticized for several key reasons, such as the unclear definition of what diffuses, the unclear connection with economic models of behavior, the mostly indirect empirical evidence, and the insufficient knowledge about the preconditions for the onset of the diffusion process. Several of these potential problems can be overcome by investigating the diffusion process with the tools of social network analysis.

1. Why Social Networks Matter for Understanding Fertility

The theoretical perspective of social networks rests on one central insight: human action is not carried out by individuals who make their decisions in isolation and independence, but by socialized actors who are deeply rooted in their social environments. These environments consist of other individuals with whom a focal actor stays in various relationships such as kinship, love, friendship, competition, or business transactions. Within these relationships, norms are negotiated and enforced, and information and goods are exchanged (Mitchell 1973). On the one hand this implies that a network member has access to resources owned by other network partners. In addition, a network enables its members to use the resources of community members to whom they are tied only indirectly. Social networks therefore constitute a pool of resources which form opportunity structures for action (Lin 1999). The kind and quality of these resources are decisive for the goals that are attainable for the network members. On the other hand, social relationships create, maintain, and modify attitudes (Erickson 1988). People feel uncomfortable if the attitudes of their network partners are too different from their own. Social networks therefore bring the attitudes of their members in accordance with each other. Alternatively, if the network fails to bring attitudes into conformity, the relationships become less intensive or are abandoned.

The resources that are available in a network, and the tendencies towards similar attitudes, depend on the characteristics of individuals' direct relationships with one another as well as on the structure of the entire network. One basic characteristic of network relations is their strength (Granovetter 1973). Strong relationships, i.e., relationships that are characterized by a high frequency of contact or by a high level of emotional intensity, can be influential because they tend to transfer more valuable goods, more binding norms and more trustworthy—but also more redundant—information. On the other hand, a fundamental dimension of the network structure is its density. A dense network, i.e., a network in which a high proportion of the individuals are in direct contact with one another, can lead to a high degree of normative pressure on its members and it can provide intensive support for members in situations when help is needed. Less dense networks are characterized by a greater degree of heterogeneity. Such networks provide more opportunities for individual development, and they are the source of more diverse and heterogeneous information.

The mechanisms by which social networks affect the diffusion process can be summarized under the headings 'social learning,' 'joint evaluation,' and 'social influence' (Bongaarts and Watkins 1996, Montgomery and Casterline 1996). At early stages, potential adopters need to be informed about the existence of new methods of fertility control, their availability, costs and benefits, and possible side effects. 'Social learning' is defined as the process by which individuals (a) learn about the existence and technical details of new contraceptive techniques, and (b) reduce the uncertainties associated with the adoption of these methods by drawing on the experience of network partners.

The spread of information about modern contraceptives is always accompanied by the spread of new ideas about the social role of women, the status of the family, and the desired number of children. 'Joint evaluation' refers to the fact that social networks (a) facilitate the transformation of these new ideas into terms that are meaningful in the local context, and (b) assist in the evaluation of the respective social changes. A woman's decision to adopt a modern form of contraception, however, rests not only on information, conviction, and discursive evaluation, but also on the normative pressures of dominant groups or actors. 'Social influence' describes the process by

which an individual's preferences and attitudes are influenced by the prevailing norms and values of the social environment. The direction of this social influence can differ in different stages of fertility decline. In the early stages, social influence may constrain the adoption of contraception, which demographers may interpret as being in a woman's best interest, because deliberate control of fertility is still considered deviant. Once a noticeable fertility decline has set in, however, the direction of social influence can shift. As low fertility becomes commonplace, the interaction among women or couples is likely to encourage a decline in the demand for children and behaviors that lead to a limiting of fertility.

2. *Empirical Evidence*

The earliest systematic evidence for the influence of diffusion processes and social networks on fertility decisions is found in two seminal analyses of demographic change. First, the European Fertility Project (Coale and Watkins 1986) showed that areas which are culturally similar, in language or ethnicity for example, often exhibit similar fertility patterns despite substantially different socioeconomic conditions and development levels. Thus, new ideas about fertility as well as new contraceptive methods seem to have diffused within cultural and linguistic boundaries. More sophisticated approaches employing similar methodological ideas have used the correlation pattern in pooled time series to show support for the role of social interactions in the adoption of family planning over time (Montgomery and Casterline 1993). Second, the Taichung family planning experiment in Taiwan (Freedman and Takeshita 1969) used a randomized experiment and longitudinal follow-up to show how direct interventions of a family planning program targeted to some individuals spread to others who were not directly contacted by the program. The study thus suggests that family planning interventions extend beyond the direct effect on the contacted individuals and that interaction in social networks constitutes an important aspect in the diffusion of new contraceptive methods and ideas.

The above evidence, although intriguing, is far from conclusive. In all of these data, the evidence is indirect and based on correlated fertility developments in geographically and/or culturally close populations. Support for the relevance of social networks in fertility transitions requires more individual-level data, gathered either in structured surveys or ethnographic studies. The work of Watkins and collaborators in South Nyanza, Kenya, for example, represents an important step in this direction. Qualitative group interviews show that casual conversations about family planning occur frequently as women go about their daily activities. The network partners in these conversations are usually geographically and socially close to one another and are of a similar age or education. Frequently they also know each other and share multiple relations, e.g., stemming from money lending, help in the household, or family ties.

These qualitative impressions are further supported by survey data that include individuals and their social networks. Until recently, only one effort had been made to assemble such data, and this was by E. Rogers and colleagues in their pioneering work in rural Korea (Rogers and Kincaid 1981). Currently there are several related research projects being conducted in Kenya and Malawi (Watkins and colleagues), Thailand (Entwisle and colleagues), and Ghana (Casterline and colleagues). In all of these projects socioeconomic panel information on women and couples is being collected along with data on social networks that may have influenced contraceptive decisions. Research using these data has consistently shown that the prevalence of contraceptive use in social networks is positively associated with the probability of women using family planning, and that the use of family planning by network partners tends to increase both the knowledge about modern methods and the willingness to use them (Entwisle and Godley 1998, Kohler et al. 2001, Montgomery and Chung 1998). Although this association is robust across different socioeconomic settings and studies, a causal interpretation is not necessarily warranted. In particular, it has been challenging for researchers to distinguish between true network effects on fertility decisions on one side, and other unobserved factors which render the behavior of network members similar on the other. Nevertheless, additional work that fully exploits the panel nature of the above studies and improves the temporal measures of interactions and the adoption of family planning has the potential to overcome this limitation.

A further criticism of existing empirical work is that the link between the theoretical arguments of why social interaction matters and the respective empirical implementation is weak in many analyses. One possible way to overcome this limitation is to use additional information about the structure of the social network, e.g., the density. Dense networks impose more behavioral constraints on the adoption or nonadoption of family planning than do sparse networks, which are more open to innovative information and behavior. Thus, density can be used as an indicator for the mechanisms of 'social influence' and 'social learning.' If the prevalence of contraceptive use in dense networks is more relevant for the decision to adopt contraception than it is in sparse networks, then social influence is the most relevant aspect of social networks.

If it is the other way around, however, then it is social learning that is the dominant component. In the South Nyanza district of Kenya, for instance, evidence for both pathways of influence can be found. Moreover, the relative importance of these two mechanisms depends on the socioeconomic context of the regions.

Social learning seems to be relatively more important in areas with higher levels of market activity (Kohler et al. 2001).

Although the empirical evidence supporting the argument that social networks play an important role in fertility decisions is accumulating from various sources, it remains inconclusive in important aspects, e.g., as regards the empirical identification of causal network effects. Nevertheless, the extensive efforts currently in progress to collect data, combined with both an improved theoretical understanding of social interaction and a more sophisticated empirical methodology, are likely to reveal further insights.

3. *Implications of Social Networks For Fertility Change*

Current research on social networks and fertility is in part stimulated by the fact that the dynamic implications of fertility models that include social interactions seem to be more consistent with the empirical evidence on fertility transitions than individualistic models. However, specific theoretical analyses of social interactions and fertility decisions are required as further support for this conclusion. Early formal models treat the diffusion of low fertility in a similar manner to the way the spread of contagious diseases is treated in epidemiological models (Rosero-Bixby and Casterline 1993). More recent developments have used stochastic processes or simulation, and they explicitly consider the interplay between social networks, women's fertility decisions, and economic incentives on a microeconomic level (Kohler 2000a, 2000b).

The main consequence of interactions in social networks is that fertility decisions of women or couples in a population become interdependent and exert mutual influences on each other. The implications of this interdependence can be illustrated in a simple nonlinear model of social interaction. In particular, assume that the probability that a woman ever adopts modern contraception is given by

$$\mathrm{P}\,(\textit{woman adopts modern contraception}\,|\,x,\bar{y}) = F(\alpha(\bar{y}-0.5)+\beta x+\gamma) \qquad (1)$$

where α and β are positive parameters, γ represents individual characteristics, x is the level of family planning effort, F is the cumulative logistic function, and $\bar{y}$ is the proportion of women in the population who use modern contraception in the reference period. Hence, the social utility term $\alpha(\bar{y}-0.5)$ captures the dependence of a woman's contraceptive choice on the fertility behavior in her network, where we assume for the sake of simplicity that this network includes all members of the population. The model implies that social effects are absent in a population in which half of the members use modern contraception and the other half do not. If the proportion of contraceptive users exceeds (is below) 0.5, then the social network increases (decreases) the probability of using contraception. This simple model already reveals that the aggregate properties of fertility dynamics in the presence of social interaction are distinct from the fertility dynamics in a purely individualistic model. In particular, the relevance of social interaction arises from three factors: social multiplier effects, multiple equilibria, and dense vs. sparse networks.

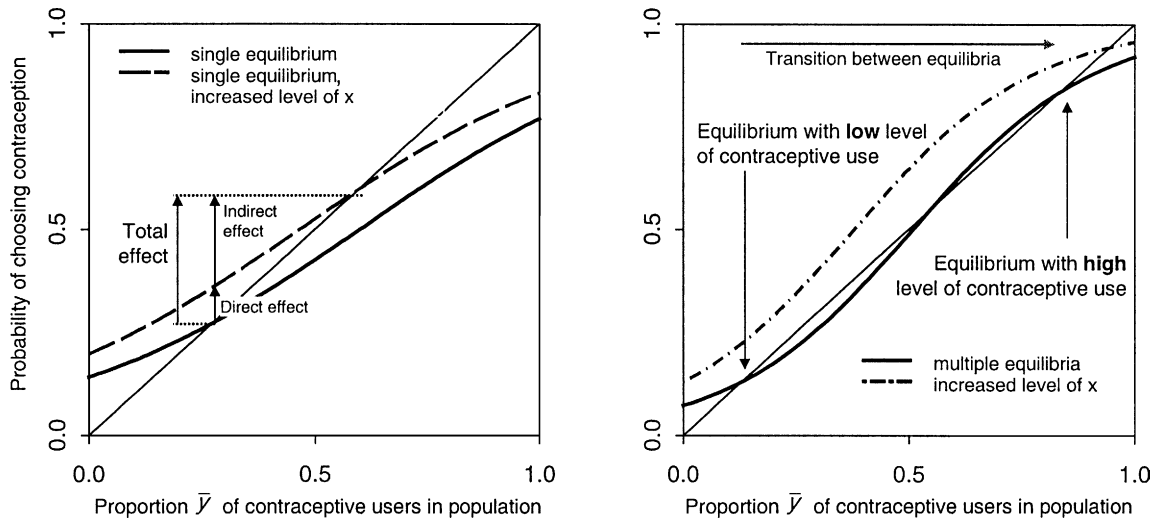

Figure 1
Implications of social networks for population dynamics. In the left-hand panel, social networks exert a relatively weak influence on individual contraceptive decisions ($\alpha = 3$), while in the right-hand panel the social network effect is relatively strong ($\alpha = 6$)

3.1 Social Multiplier Effects

Social interaction implies the presence of multiplier effects which change how the fertility level adjusts to changes in family planning programs. An increase in family planning effort x, for instance, increases the probability of choosing modern contraception in two distinct ways. The direct effect reflects the increase in the propensity to choose modern contraception while holding the population prevalence $\bar{y}$ constant. However, because the increased level of x affects all community members, the prevalence of modern contraception in the population tends to increase as well. This change in the community has a feedback or social multiplier effect on individual behavior in future periods via the social utility term in Eqn. (1). This means that the total change in the prevalence of family planning exceeds the direct effect. The left-hand panel of Fig. 1 shows how this total effect is composed of a direct effect and an indirect effect through social interaction. The interaction in social networks thus enhances the changes in contraceptive behavior that follow changes in family planning efforts or other socioeconomic changes.

3.2 Multiple Equilibria

Many formal fertility models assume that there is only one aggregate fertility level, denoted as 'equilibrium,' associated with any given level of socioeconomic conditions. When the interdependence of fertility decisions in social networks is sufficiently strong, however, multiple equilibria can emerge (right-hand panel of Fig. 1). This possibility of there being multiple equilibria is relevant because in this case a population can be 'stuck' at a low level of contraceptive use, even though a second equilibrium exists with a higher level of contraceptive use and higher individual utility levels. In this case, increases in the policy level x (or other socioeconomic changes) can induce a transition from the low contraceptive use to the high contraceptive use equilibrium. These transitions between equilibria are often thought to occur at a rapid pace, resulting in significant changes in fertility behavior within a relatively short space of time. Moreover, both the high and low contraceptive use equilibria are stable, so that a transitory increase in the family planning effort can yield sustainable long-term changes in fertility levels.

3.3 Populations With Dense Versus Sparse Networks

An important question in diffusion models is whether 'more' interaction in networks always leads to larger multiplier effects and more rapid diffusion. It can be shown that already in simple models such as Eqn. (1) social networks can be status-quo enforcing: the long-term effect associated with changes in the program effort can decrease with an increase in parameter α, which reflects the importance of the network influences in individual fertility decisions. The changes in contraceptive use resulting from changes in the program effort can therefore be smaller for very densely knit societies than for more sparse societies.

Bibliography

Bongaarts J, Watkins S C 1996 Social interactions and contemporary fertility transitions. *Population and Development Review* **22**: 639–82

Coale A J, Watkins S C 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ

Entwisle B, Godley J 1998 Village network and patterns of contraceptive choice. Paper presented at a meeting of the Population Committee of the National Academy of Sciences, January 29–30, Washington, DC

Erickson B H 1988 The relational basis of attitudes. In: Wellman B, Berkowitz S D (eds.) *Social Structures: A Network Approach*. Cambridge University Press, Cambridge, pp. 99–121

Freedman R, Takeshita J Y 1969 *Family Planning in Taiwan: An Experiment in Social Change*. Princeton University Press, Princeton, NJ

Granovetter M S 1973 The strength of weak ties. *American Journal of Sociology* **73**: 1361–80

Kohler H-P 2000a Fertility decline as a coordination problem. *Journal of Development Economics* **63**: 231–63

Kohler H-P 2000b Social interaction and fluctuations in birth rates. *Population Studies* **54**: 223–38

Kohler H-P, Behrman J R, Watkins S C 2001 The density of social networks and fertility decisions: Evidence from South Nyanza District, Kenya. *Demography* **38**: 43–58

Lin N 1999 Building a network theory of social capital. *Connections* **22**: 28–51

Mitchell J C 1973 Networks, norms, and institutions. In: Boissevain J, Mitchell J C (eds.) *Network Analysis: Studies in Human Interaction*. Mouton, The Hague, Holland pp. 2–35

Montgomery M R, Casterline J B 1993 The diffusion of fertility control in Taiwan: Evidence from pooled cross-section time-series models. *Population Studies* **47**: 457–79

Montgomery M R, Casterline J B 1996 Social learning, social influence learning and new models of fertility. *Population and Development Review* **22**: S151–75

Montgomery M R, Chung W S 1998 Social networks and the diffusion of fertility control in the Republic of Korea. In: Leete R (ed.) *The Dynamics of Values in Fertility Change*. Oxford University Press, Oxford

Rogers E M, Kincaid D L 1981 *Communication Networks: Toward a New Paradigm for Research*. Free Press, New York

Rosero-Bixby L, Casterline J B 1993 Modelling diffusion effects in fertility transition. *Population Studies* **47**: 147–67

H. P. Kohler and C. Bühler

Social Networks and Gender

One of the most frequently cited causes for inequalities in the distribution of income, power, and status to men and women in the workplace is gender differences in

social networks (Baron and Pfeffer 1994). The term network refers to the set of relationships defined by an individual and his or her direct contacts with others (see Ibarra 1993 for various definitions). Networks shape careers by awarding access to jobs, providing mentoring and sponsorship, channeling the flow of information and referrals, augmenting power and reputations, and increasing the likelihood and speed of promotion (e.g., Brass 1985, Burt 1992, Ibarra 1995, Kanter 1977). This article reviews the last two decades of research on men's and women's work networks and proposes a social structural perspective that can integrate findings about gender differences in network characteristics and their consequences. A social structural view underscores the embeddedness of social relationships within structural contexts that constrain the availability and ease of developing various kinds of social contacts. But such a perspective also gives prominence to the interaction patterns that intervene between causal processes at the macrostructural level of socioeconomic organizations and those operating at the individual level.

Three broad sets of factors moderate the relationship between gender and work-related networks: sex stratification, interaction dynamics, and gender beliefs and identities. Women are under-represented in executive positions and men and women are segregated into different jobs, occupations, firms, and industries such that women are over-represented in lower paying jobs (Baron and Pfeffer 1994, Reskin 1993). Men and women interact on a daily basis in social life, but the structure of gender roles outside the workplace is such that they engage in different activities and form part of different social circles outside of work (Ross 1987). Voluntary group activities, for example, mirror and reinforce the gender segregation found in the workplace (McPherson and Smith Lovin 1986). Furthermore, as in the workplace, women's social roles tend to be lower in status and prestige than men's; as a result, although men and women interact frequently, only a minority of these cross-gender interactions occur between people of roughly equivalent power and status (Ridgeway and Smith-Lovin 1999).

Kanter (1977) established the effects of differences in power, opportunity, and numbers on interaction dynamics and individual attitudes. Interaction dynamics, which are both cause and consequence of stratification inside and outside the workplace, further reinforce differences in networks via several mechanisms. Well-documented preferences for homophily, i.e., same-gender interaction, lead men and women to organize discretionary interaction within different social circles (Smith-Lovin and McPherson 1993). Preferences for interaction with high status individuals—those who are powerful and upwardly mobile—influence the dynamics of interpersonal attraction and network formation (Ibarra 1992). But, 'female' is a lower ascribed status than 'male' (Ridgeway 1991). Thus, for women, preferences for homophily and preferences for status conflict, while these preferences are congruent for men. This serves to further segregate networks and aids in the perpetuation of stereotypic beliefs about gender and work competency (Ridgeway and Smith-Lovin 1999). Gender beliefs and identities, which are themselves a product of interaction patterns, affect how people interact with the constraints of their structural contexts, shaping what contacts men and women seek and how they negotiate their interactions (Ridgeway and Smith-Lovin 1999).

In sum, the literature on work networks suggests that social structures, social interaction, and social identities intertwine to differentiate women's and men's networks. Macrostructural demographic and job stratification patterns interact with the microdynamics of interpersonal exchange and identity construction to reinforce gender differences in networks as well as in the types of networks that each need in order to further their attainment. Below is reviewed the research evidence in each of these areas.

1. *Differences in Networks*

Structuralists have typically studied how men and women enter the occupational world with different human capital and how they are distributed in different jobs, hierarchical levels, firms, and even industries. These structural factors account for most of the differences observed in men and women's networks and related mobility outcomes. When men and women hold similar jobs, occupations, and hierarchical levels, they have similar networks and attain equally central informal positions (Brass 1985, Ibarra 1992, Miller 1986, Burt 1992, Aldrich et al. 1989, Moore 1990).

Regardless of position, however, organizational demography affects networks by regulating the availability of same-gender contacts that are also influential (Ibarra 1992). Compared to men, therefore, women typically have fewer, high-status, same-gender contacts available to them for work-related interaction (Ibarra 1992). As a result, women working in male-dominated industries, firms, or jobs tend to develop 'functionally segregated' networks with personal and professional contacts relegated to different, rarely overlapping networks. This differs from the networks of their male counterparts, who tend to develop more 'multiplex' network relationships, i.e., relationships that have both a personal and professional dimension, rather than just one. The trust and loyalty associated with coaching and advocacy for promotion is more likely to develop in relationships defined by multiple dimensions (Kanter 1977).

Outside the workplace, family and life-course factors also produce similar structural effects. Kinship ties tend to account for a greater proportion of women's ties and this effect is heightened with child rearing when the effects of occupation are held

constant (Campbell 1988, Fischer and Oliker 1983, Moore 1990). Sex segregation of activities inside and outside the workplace, therefore, limits the status of contacts and multiplexity of women's work relationships and thus the career instrumentality of their network.

2. *Differences in the Relationship Between Networks and Professional Outcomes*

More recent research suggests that, due to the structural factors outlined above, women must use different network means to achieve the same career goals as men. Many studies, reviewed in more detail below, find that an interaction between gender and human capital or structural position—rather than main effects for either—explains network characteristics.

Men appear to be better able than women to convert positional and human capital resources such as hierarchical rank, educational attainment, and external professional contacts into access to central network positions (Ibarra 1992, Miller 1986, Miller et al. 1981). Burt (1992) found that men furthered their career mobility by creating networks rich in 'bridging ties' to people who were not directly connected to one another. Women, by contrast, required strong ties to strategic partners to signal their legitimacy and thus contribute to their advancement. He argues that networks dominated by strong ties, i.e., relationships high in emotional intensity, intimacy, or frequency of interaction (Granovetter 1982), while detrimental for men are beneficial for women by serving to compensate for their lower status and legitimacy in the managerial world. In a study of Fortune 500 middle managers, Ibarra (1997) found that fast-track men reported networks much like those of Burt's highly mobile managers, while the fast-track women built networks that were higher in both the tie strength and range of contacts. These findings are consistent with Granovetter's (1982) argument that weak ties are less advantageous for people in socially or economically insecure positions.

Women may also develop different network routes in order to compensate for the few women available to them in the workplace. Men's workplace ties are primarily with men, regardless of the type of relationship involved; for women, by contrast, contacts with men provide certain types of instrumental access, while relationships with a mix of men and women serve as sources of social support and friendship (Ibarra 1992, Lincoln and Miller 1979). Brass (1985) found that homophilous workplace networks were more detrimental for women than for men because those who control the promotion process tend to be male. Ibarra (1997), by contrast, found that same-gender ties help managerial women advance. When women are under-represented in managerial and professional ranks, homophilous ties play a critical role in helping those in the minority gain advice from others who have had similar experiences. A beneficial side effect is that reaching out to other women typically requires women to extend their network links beyond the immediate workgroups or functional areas, adding breadth and range to their networks.

3. *Effects of Identity on Networks*

Network behavior is not only driven by instrumental motives such as career advancement. People initiate and maintain relationships as expressions of valued social identities, and participate in activities and institutions that are congruent with those identities (Stryker and Serpe 1982). Gender identity, therefore, may affect a person's choice of social contexts (within which different types of potential network contacts are more or less available) as well as the discretionary contacts that are developed within those contexts. But, the simple existence of homophily on some trait is not sufficient motive for tie formation (Ibarra 1997).

Women vary in the strength of their identification with their multiple identities, and this fact may explain large within-gender group differences in network homophily. Chatman and Brown (1996) found that social identity mediates the effects of demographic similarity at the dyadic level on friendship formation in business school student networks. Thus, identification with one's gender group explains variance in the formation and nature of network ties over and above that explained by demography. Demography, however, is still expected to play an important role because ascribed categories such as gender acquire their meaning in the context of social interaction (Wharton 1992). In work environments in which the distribution of power and status favors men numerically, the extent to which gender is a salient identity will affect women's interaction choices. Ely (1994), for example, found a correlation between the organizational demography of law firms and the quality of peer and hierarchical relationships among women in those firms. She argued that this correlation was mediated by the extent to which the women in those firms identified with each other on the basis of gender.

People have multiple identities (e.g., consultant, wife, mother, tennis player, Protestant) that vary in their importance to the individual and in the degree to which they are reinforced or embedded in one's network (Stryker and Serpe 1982). Stryker and Serpe argue that a person's 'commitment' to any given identity is a function of the number, affective importance, and multiplexity of network ties that are formed by the person when enacting the identity. As discussed above, the demographics of the workplace and social groups outside work make it easier for men's personal and professional contacts to work

synergistically; women's are more likely to be uniplex, or valuable on a single dimension.

This pattern is exacerbated by life course and family factors because gender interacts statistically with major family transitions to change network patterns (Elder and O'Rand 1995). Munch et al. (1997) found that childbirth leads women to drop their job and career-related ties in order to limit the overall size of their networks, while men substitute family ties for male social friends, but without losing job contacts. Changes in family status, therefore, are associated with changes in patterns of social relations, which shape and give meaning to women's and men's experiences, both in their families and jobs. The network configurations created by major family transitions, in turn, will shape the hierarchical ordering of social identities and the cultural assumptions that women and men make about the appropriate balance of job- and family-related work. Changes in network characteristics associated with family transitions may explain cohort and individual differences in women's responses to job–family demands.

4. Conclusion

The existing evidence suggests that different network configurations are effective in promoting the careers of men and women. In most organizations men and women differ in the jobs they hold, in their opportunity for informal interaction with high-status, same-gender others, and the expectations that govern their interactions with others. Outside the workplace, men and women are subject to different role demands. They come to develop different types of networks because they are, in fact, operating in different social contexts that require different network configurations to accomplish similar objectives. These findings raise critical questions for future research on the interaction between gender and networks as it affects the demographics of the workplace and course of careers. They suggest that future research should shift focus from comparing men's and women's networks towards exploring why different networks work for women and men, and what factors affect the formation of networks with such characteristics.

See also: Gender and Feminist Studies in Sociology; Language and Gender; Networks: Social; Sex Segregation at Work; Small-group Interaction and Gender

Bibliography

Aldrich H, Reese P R, Dubini P 1989 Women on the verge of a breakthrough: Networking among entrepreneurs in the United States and Italy. *Entrepreneurship and Regional Development* **1**: 339–56

Baron J, Pfeffer J 1994 The social psychology of organizations and inequality. *Social Psychology Quarterly* **57**: 190–209

Brass D J 1985 Men's and women's networks: A study of interaction patterns and influence in an organization. *Academy of Management Journal* **28**: 327–43

Burt R S 1992 *Structural Holes*. Harvard University Press, Cambridge, MA

Campbell K E 1988 Gender differences in job-related networks. In: *Work and Occupations* **15**: 179–200

Chatman J A, Brown R A 1996 It takes two to tango: Demographic similarity, social identity and friendship. Paper presented at the Stanford Conference on Power, Politics and Influence

Ely R J 1994 The effects of organizational demographics and social identity on relationships among professional women. *Administrative Science Quarterly* **39**: 203–38

Fischer C, Oliker S 1983 A research note on friendship, gender, and the life cycle. *Social Forces* **62**: 124–32

Granovetter M 1982 The strength of weak ties: A network theory revisited. In: Marsden P V, Lin N (eds.) *Social Structure and Network Analysis*. Sage, Beverly Hills, CA

Ibarra H 1992 Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative Science Quarterly* **37**: 422–47

Ibarra H 1993 Personal networks of women and minorities in management: A conceptual framework. *Academy of Management Review* **18**(1): 56–87

Ibarra H 1997 Paving an alternate route: Gender differences in managerial networks for career development. *Social Psychology Quarterly* **60**: 91–102

Ibarra H, Smith-Lovin L 1997 New directions in social network research on gender and careers. In: Jackson S, Cooper C (eds.) *Futures of Organizations*. John Wiley & Sons, New York, pp. 359–84

Kanter R M 1977 *Men and Women of the Corporation*. Basic Books, New York

McPherson J M, Smith-Lovin L 1986 Sex segregation in voluntary associations. *American Sociological Review* **51**: 61–79

McPherson J M, Smith-Lovin L 1987 Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Journal of Sociology* **52**: 370–79

Miller J 1986 *Pathways in the Workplace*. Cambridge University Press, Cambridge, UK

Moore G 1990 Structural determinants of men's and women's personal networks. *American Sociological Review* **55**: 726–35

Munch A, Miller McPherson J, Smith-Lovin J 1997 Gender, children and social context: A research note on the effects of childrearing for men and women. *American Sociological Review* **62**: 509–20

Reskin B 1993 Sex segregation in the workplace. *Annual Review of Sociology* **19**: 241–70

Ridgeway C 1991 The social construction of status value: Gender and other nominal characteristics. *Social Forces* **70**: 367–86

Ridgeway C L, Smith-Lovin L 1999 The gender system and interaction. *Annual Review of Sociology* **25**: 191–216

Ross C E 1987 The division of labor at home. *Social Forces* **65**: 816–33

Smith-Lovin L, McPherson M J 1993 You are who you know: A network approach to gender. In: England P (ed.) *Theory on Gender/Feminism on Theory*. Aldine, New York

Stryker S, Serpe R T 1982 Commitment, identity salience and role behavior. In: Ickes W, Knowles E (eds.) *Personality, Roles and Social Behavior*. Springer-Verlag, New York, pp. 199–218

Wharton A S 1992 The social construction of gender and race in organizations: A social identity and group mobilization perspective. In: Tolbert P, Bacharach S (eds.) *Research in the Sociology of Organizations*. JAI Press, Greenwich, CT, Vol. 10, pp. 55–84

H. Ibarra

Social Neuroscience

Social neuroscience refers to the study of the relationship between neural and social processes. A fundamental assumption in the science of psychology is that neurophysiological processes underlie social and psychological phenomena. The early American psychologist William James (1890/1950) was among the first to articulate this assumption; to recognize that developmental, environmental, and sociocultural factors influence the neurophysiological processes underlying psychological and social phenomena; and to acknowledge that these influences could be studied as neurophysiological transactions. James also asserted that unnecessary diseconomies and conundrums would result if neurophysiological events were the exclusive focus of psychological investigations.

Despite this early recognition that molecular processes and mechanisms bear on molar social psychological processes and vice versa, the early identification of biology with innate mechanisms and the identification of social psychology with self-report data contributed to an antipathy that endured long after each field became more sophisticated. Accordingly, neuroscientific analyses tended to be limited to behavioral phenomena occurring within an organism (e.g., attention, learning, memory), whereas social psychological analyses traditionally eschewed biological data in favor of self-report and behavioral data.

As neuroscientific approaches were applied to the study of diseases and to elementary cognitive and behavioral processes, subtler medical and behavioral phenomena succumbed to neuroscientific inquiry. The 1990s were declared by the US Congress to be the 'decade of the brain,' reflecting the interest in and importance of the neurosciences in theory of and research on cognition, emotion, behavior, and health.

1. *Doctrine of Multilevel Analyses*

The decade of the brain has led to a realization that a comprehensive understanding of the brain cannot be achieved by a focus on neural mechanisms alone. The brain, the organ of the mind, is a fundamental but interacting component of a developing, aging individual who is a mere actor in the larger theater of life. This theater is undeniably social, beginning with prenatal care, caregiver–infant attachment, and early childhood experiences and ending with feelings of loneliness or embeddedness and with familial or societal decisions about care for the elderly. Mental disorders such as depression, anxiety, schizophrenia, and phobia are both determined by and are determinants of social processes; and disorders such as substance abuse, prejudice and stigmatization, family discord, worker dissatisfaction and productivity, and the spread of acquired immunodeficiency syndrome are quintessentially social as well as neurophysiological processes.

Indeed, humans are such social animals that a basic 'need to belong' has been posited (Baumeister and Leary 1995, Gardner et al. in press). People form associations and connections with others from the moment they are born. The very survival of newborns depends on their attachment to and nurturance by others over an extended period of time. Accordingly, evolution has sculpted the human genome to be sensitive to and succoring of contact and relationships with others. For instance, caregiving and attachment have hormonal (Uvnas-Mosberg 1997) and neurophysiological substrates (Carter et al. 1997). Communication, the bedrock of complex social interaction, is universal and ubiquitous in humans. In the rare instances in which human language is not modeled or taught, language develops nevertheless (Goldin-Meadow and Mylander 1983, 1984).

The reciprocal influences between social and biological levels of organization do not stop at infancy. Affiliation and nurturant social relationships, for instance, are essential for physical and psychological well-being across the lifespan. Disruptions of social connections, whether through ridicule, separation, divorce, or bereavement, are among the most stressful events people must endure (Gardner et al., 2000). Berkman and Syme (1979) operationalized social connections as marriage, contacts with friends and extended family members, church membership, and other group affiliations. They found that adults with fewer social connections suffered higher rates of mortality over the succeeding nine years even after accounting for self-reports of physical health, socioeconomic status, smoking, alcohol consumption, obesity, race, life satisfaction, physical activity, and preventive health service usage. House et al. (1982) replicated these findings using physical examinations to assess health status. In their review of five prospective studies, House et al. (1988) concluded that social isolation was a major risk factor for morbidity and mortality from widely varying causes. This relationship was evident even after statistically controlling for known biological risk factors, social status, and baseline measures of health. The negative health consequences of social isolation were particularly strong among some of the fastest growing segments of

the population: the elderly, the poor, and minorities such as African Americans. Astonishingly, the strength of social isolation as a risk factor was found to be comparable to high blood pressure, obesity, and sedentary lifestyles.

Social isolation and loneliness are associated with poorer mental as well as physical well-being (e.g., Ernst and Cacioppo 1999, Gupta and Korte 1994, Perkins 1991). People who report having contact with five or more intimate friends in the prior six months are 60 percent more likely to report that their lives are 'very happy,' as compared to those who do not report such contact (Burt 1986). People appear to be cognizant of the importance of social relationships. When asked 'what is necessary for happiness?' most rated relationships with friends and family as being the most important factor (Berscheid 1985).

Initially, studies of the neural structures and processes associated with psychological and social events were limited to animal models, post-mortem examinations, multiple determined peripheral assessments, and observations of the occasional unfortunate individual who suffered trauma to or disorders of localized areas of the brain. Developments in electrophysiological recording, functional brain imaging, neurochemical techniques, neuroimmunologic measures, and ambulatory recording procedures have increasingly made it possible to investigate the role of neural systems and processes in intact humans. These developments fostered multilevel integrative analyses of the relationship between neural and social processes. Cacioppo and Berntson (1992) coined the term social neuroscience and outlined several principles for spanning molar and molecular levels of organization.

1.1 The Principle of Multiple Determinism

The principle of multiple determinism specifies that a target event at one level of organization, but particularly at molar (e.g., social) levels of organization, may have multiple antecedents within or across levels of organization. For example, consider the multiple factors that contribute to drug abuse. On the biological level, researchers identified the contribution of individual differences in the susceptibility of the endogenous opiod receptor system while on the social level investigators have noted the important role of social context. Both operate, and our understanding of drug abuse is incomplete if either perspective is excluded. A corollary to this principle is that the mapping between elements across levels of organization becomes more complex (e.g., many-to-many) as the number of intervening levels of organization increases. One implication is that the likelihood of complex and potentially obscure mappings increases as one skips levels of organizations.

1.2 The Principle of Nonadditive Determinism

The principle of nonadditive determinism specifies that properties of the whole are not always readily predictable from the properties of the parts. Consider an illustrative study by Haber and Barchas (1983), who investigated the effects of amphetamine on primate behavior. The behavior of nonhuman primates was examined following the administration of amphetamine or placebo. No clear pattern emerged between the drug and placebo conditions until each primate's position in the social hierarchy was considered. When this social factor was taken into account, amphetamines were found to increase dominant behavior in primates high in the social hierarchy and to increase submissive behavior in primates low in the social hierarchy. The importance of this study derives from its demonstration of how the effects of physiological changes on social behavior can appear unreliable until the analysis is extended across levels of organization. A strictly physiological (or social) analysis, regardless of the sophistication of the measurement technology, may not have unraveled the orderly relationship that existed.

1.3 The Principle of Reciprocal Determinism

The principle of reciprocal determinism specifies that there can be mutual influences between microscopic (e.g., biological) and macroscopic (e.g., social) factors in determining behavior. For example, not only has the level of testosterone in nonhuman male primates been shown to promote sexual behavior, but the availability of receptive females influences the level of testosterone in nonhuman primates (Bernstein et al. 1983, Rose et al. 1972). Within social psychology, Zillmann (1994) has demonstrated that exposure to violent and erotic materials influences the level of physiological arousal in males, and that the level of physiological arousal has a reciprocal influence on the perceptions of and tendencies toward sex and aggression. Accordingly, comprehensive accounts of these behaviors cannot be achieved if either biological or social levels of organization are considered unnecessary or irrelevant.

2. Insights Fostered by Multilevel Analyses

Recent research has provided growing evidence that multilevel analyses spanning neural and social perspectives can foster more comprehensive accounts of cognition, emotion, behavior, and health (e.g., Anderson 1998).

First, important inroads to the logic of social processes have emerged from theory and research in

the neurosciences (e.g., brain organization and localization of function, genetic determinants of behavior). Research in the neurosciences, for instance, has revealed that early ideas about memory, in which knowledge and skills were thought to be served by the same underlying mechanism, are incorrect. Instead, separable neural systems underlie the memory of people, episodes and ideas (declarative, episodic memory), and the memory of procedures, interactions, or skills (procedural, implicit memory).

Similarly, research in the neurosciences has influenced what are thought to be the processes underlying social attitudes and decisions. Mechanisms for differentiating hostile from hospitable environmental stimuli are imperative for the survival of species and for the formation and maintenance of social units. The human brain and body therefore can be conceptualized as being shaped by natural selection to calculate utility and to respond accordingly. Evaluative decisions and responses are so critical that organisms have rudimentary reflexes for categorizing and approaching or withdrawing from certain classes of stimuli, and for providing metabolic support for these actions (Lang et al. 1990). A remarkable feature of humans is the extent to which the evaluation of utility is shaped by learning and cognition.

The processes subserving the evaluation of hostile and hospitable events are also physiological processes and cannot be understood fully without considering the structural and functional aspects of the physical substrates. Noninvasive investigations of the physiological operations associated with evaluative processes provide an important window through which to view these processes without perturbing them. For instance, the neural circuitry involved in computing the utility of a stimulus (i.e., evaluative processing) diverges from the circuitry involved in identification and discrimination (i.e., nonevaluative processing). Furthermore, evaluative discriminations by humans (e.g., attitudes) have traditionally been conceptualized as being bipolar (hostile–hospitable) and have been measured using bipolar scales to gauge the net affective predisposition toward a stimulus. Such an approach treats positive and negative evaluative processes (and the resulting affective states and judgments of utility) as equivalent directions on a single scale, reciprocally activated, and interchangeable in the sense that increasing one is equivalent to decreasing the other. Physical constraints may restrict behavioral manifestations of utility calculations to bivalent actions (approach–withdrawal), but research in the behavioral and social neurosciences now suggests that approach and withdrawal are behavioral manifestations that come from distinguishable motivational substrates. This work has led to evolutionary models of affect and emotion in which a stimulus may vary in terms of the strength of positive evaluative activation (i.e., positivity) and the strength of negative evaluative activation (i.e., negativity) it evokes. Furthermore, the underlying positive and negative evaluative processes are distinguishable, are characterized by distinct activation functions, and have distinguishable antecedents (Cacioppo and Berntson 1994).

Second, the study of social processes has challenged existing theories in the neurosciences, resulting in refinements, extensions, or complete revolutions in neuroscientific theory and research. Classically, immune functions were considered to reflect specific and nonspecific physiological responses to pathogens or tissue damage. It is now clear that immune responses are heavily influenced by central nervous processes that are affected by social interactions and processes. For instance, the effects of social context now appear to be among the most powerful determinants of the expression of immune reactions (Glaser and Kiecolt-Glaser 1994). It is clear that an understanding of immunocompetence will be inadequate in the absence of considerations of psychosocial factors. Research on these interactions was activated by investigations demonstrating the direct and moderating effects of psychosocial factors (e.g., conditioned stimuli, bereavement, social support, major life events) on immune competence (Kiecolt-Glaser et al. 1984). Thus, major advances in the neurosciences can derive from increasing the scope of the analysis to include the contributions of social factors and processes.

Third, the social environment shapes neural structures and processes and vice versa. Meaney and colleagues (Liu et al. 1997, Meaney et al. 1993), for instance, have found that experimental manipulations of maternal care influence the development of individual differences in neuroendocrine responses to stress in rats. As adults, the offspring of mothers that exhibited more licking and grooming of pups during the first ten days of life were also characterized by reduced adrenocorticotropic hormone and corticosterone responses to acute stress. As mothers, these rats also tended to lick and groom their pups.

Finally, deciphering the structure and function of the brain is fostered by sophisticated social psychological theories in which the elementary operations underlying complex social behaviors are explicated, and by experimental paradigms that allow these social psychological operations to be studied in isolation using neuroscientific methods. Brain imaging studies of the neural bases of emotion have contrasted positive and negative emotions based on the assumption that these emotions are served by the same neural structures. Research and theory in the social and behavioral sciences now suggests that the processes underlying positive and negative emotions are separable, and paradigms for eliciting positive and negative emotions have been developed. Accordingly, recent brain imaging studies in which positive and neutral emotions are contrasted, and negative and neutral emotions are contrasted, have revealed the activation of distinguishable neural structures during positive and negative emotions.

In sum, social neuroscience is an interdisciplinary field that is helping to illuminate questions ranging from the social sciences to the neurosciences by examining how organismic processes are shaped, modulated, and modified by social factors and vice versa.

See also: Behavioral Neuroscience; Comparative Neuroscience; Psychophysiology

Bibliography

Anderson N B 1998 Levels of analysis in health science. A framework for integrating sociobehavioral and biomedical research. *Annals of the New York Academy of Sciences* **840**: 563–76

Baumeister R F, Leary M R 1995 The need to belong: Desire for interpersonal attachment as a fundamental human motivation. *Psychological Bulletin* **117**: 497–529

Berkman L F, Syme S L 1979 Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda County residents. *American Journal of Epidemiology* **109**: 186–204

Bernstein I S, Gordon T P, Rose R M 1983 The interaction of hormones, behavior, and social context in nonhuman primates. In: Svare B B (ed.) *Hormones and Aggressive Behavior*. Plenum Press, New York pp. 535–61

Berscheid E 1985 Interpersonal attraction. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*. Random House, New York

Burt R S 1986 Strangers, friends, and happiness. *GSS Technical Report* **72**: National Opinion Research Center, University of Chicago, Chicago

Cacioppo J T, Berntson G G 1992 Social psychological contributions to the decade of the brain: The doctrine of multilevel analysis. *American Psychologist* **47**: 1019–28

Cacioppo J T, Berntson G G 1994 Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin* **115**: 401–23

Carter C S, Lederhendler I I, Kirkpatrick B 1997 *The Integrative Neurobiology of Affiliation*. New York Academy of Sciences, New York

Ernst J M, Cacioppo J T 1999 Lonely hearts: Psychological perspectives on loneliness. *Applied and Preventive Psychology* **8**: 1–22

Gardner W L, Gabriel S, Diekman A B 2000 Interpersonal processes. In: Cacioppo J T, Tassinary L G, Berntson G G (eds.) *Handbook of Psychophysiology*, 2nd edn, Cambridge University Press, New York, pp. 643–64

George M S, Ketter T A, Parekh P I, Horwitz B, Herscovitch P, Post R M 1995 Brain activity during transient sadness and happiness in healthy women. *American Journal of Psychiatry* **152**: 341–51

Glaser R, Kiecolt-Glaser J K 1994 *Handbook of Human Stress and Immunity*. Academic Press, San Diego, CA

Goldin-Meadow S, Mylander C 1983 Gestural communication in deaf children: Noneffect of parental input on language development. *Science* **221**: 372–4

Goldin-Meadow S, Mylander C 1984 Gestural communication in deaf children: The effects and noneffects of parental input on early language development. *Monographs of the Society for Research in Child Development* **49**: 1–121

Gupta V, Korte C 1994 The effects of a confidant and a peer group on the well-being of single elders. *International Journal of Aging and Human Development* **39**: 293–302

Haber S N, Barchas P R 1983 The regulatory effect of social rank on behavior after amphetamine administration. In: Barchas P R (ed.) *Social Hierarchies: Essays Toward a Sociophysiological Perspective*. Greenwood Press, Westport, CT, pp. 119–32

House J S, Landis K R, Umberson D 1988 Social relationships and health. *Science* **241**: 540–5

House J S, Robbins C, Metzner H L 1982 The association of social relationships and activities with mortality: Prospective evidence from the Tecumseh Community Health Study. *American Journal of Epidemiology* **116**: 123–40

James W 1890/1950 *Principles of Psychology*. Dover, New York

Kiecolt-Glaser J K, Garner W, Speicher C E, Penn G, Glaser R 1984 Psychosocial modifiers of immunocompetence in medical students. *Psychosomatic Medicine* **46**: 7–14

Lang P J, Bradley M M, Cuthbert B N 1990 Emotion, attention, and the startle reflex. *Psychological Review* **97**: 377–95

Liu D, Diorio J, Tannenbaum B, Caldji C, Francis D, Freedman A, Sharma S, Pearson D, Plotsky P M, Meaney M J 1997 Maternal care, hippocampal glucocorticoid receptors, and hypothalamic-pituitary-adrenal responses to stress. *Science* **277**: 1659–62

Meaney M J, Bhatnagar S, Diorio J, Larocque S, Francis D, O'Donnell D, Shanks N, Sharma S, Smythe J, Viau V 1993 Molecular basis for the development of individual differences in the hypothalamic pituitary adrenal stress response. *Cellular and Molecular Neurobiology* **13**: 321–47

Perkins H W 1991 Religious commitment, 'Yuppie' values, and well-being in post collegiate life. *Review of Religious Research* **32**: 244–51

Rose R M, Bernstein I S, Gordon T P 1972 Plasma testosterone levels in the male rhesus: Influences of sexual and social stimuli. *Science* **178**: 643–5

Uvnas-Mosberg K 1997 Physiological and endocrine effects of social contact. *Annals of the New York Academy of Sciences* **807**: 146–63

Zillmann D 1994 Cognition-excitation interdependencies in the escalation of anger and angry aggression. In: Potegal M, Knutson J F et al. (eds.) *The Dynamics of Aggression: Biological and Social Processes in Dyads and Groups*. Erlbaum, Hillsdale, NJ, pp. 45–71

J. T. Cacioppo and G. G. Bernston

Social Properties (Facts and Entities): Philosophical Aspects

One area of philosophical concern, dating from the very beginning of philosophy itself, is ontology: what is there?; or, what kinds of things are there? Many questions about existence are straightforwardly em-

pirical, and hence of no interest to the ontologist. Are there unicorns, wombats, planets circling stars other than our Sun? But other existence questions are not empirical, and these are the province of the ontologist. Are there any physical objects, ideas, forms, numbers, sets, propositions, minds, concepts, a Deity, social wholes? To answer these questions is to engage in producing a list of the metaphysical kinds of things that one believes exist.

'Existence' in what sense? Some philosophers have held that each metaphysically different kind of thing exists in its own, special sense of 'existence'; or they may distinguish between being, existence, subsistence, and so on.

A more plausible view is that 'existence' is univocal, that whatever exists exists in the same sense, and that differences in the 'mode' of existence, say, of mind and body, arise not because *existence* differs in the two cases but because of the differences between minds and bodies themselves. Bodies have a spatial 'mode' of existence, it might be thought, and minds do not, but, if so, that is because of the differences between them, not because they exist in different senses of that word. Besides, if existence had a special sense for each metaphysical kind of thing, ontology could become somewhat trivial. To the question of whether, say, propositions really exist, it would be anodyne simply to say that they do exist, but only in the special proposition-sense of existence. Or that minds exist, but only in the special mental sense of existence. Or that possibilia subsist rather than exist (since that is what possible things do) whereas actual things exist rather than subsist. Ontology would entail no honest labor, and therefore no real rewards.

1. *Ontological Commitments and How to Wriggle Out of Them*

When we speak, indeed even whenever we simply think, we typically take ourselves to be speaking or thinking about various things. It is not easy to capture this notion of aboutness, but here I rely on the readers' intuitive grasp of the idea.

I do not always think or speak about anything, whenever I speak or think: in (a) 'whoever is a bachelor is neurotic,' I am not trying to speak or think about any particular thing, or even about any set of things. In (a), I am merely linking up two properties: the property of being a bachelor and the property of being neurotic, but I am not speaking about those properties. If I want to speak about the set of bachelors and the set of neurotic persons, I could say: (b) 'The set of bachelors is a proper subset of the set of neurotic persons.' (b) is certainly not equivalent to (a), since (b) has ontological commitments of which (a) is wholly innocent.

I might be thinking or speaking about Socrates or the man I met yesterday or the tallest building in New York. Let us call whatever I think or speak about 'entities.' There are a number of ways in which things can go wrong when I speak about entities. First, I can make mistakes, e.g., when I take myself to be thinking about something that does not in fact exist: Atlantis, Vulcan, witches, or phlogiston. These are cases in which we really are *intending* to think about something, but fail. Let us call these sorts of cases, 'cases of elimination.' There simply are no such things, and if we have at first placed them on our list of what there is, on finding that they do not exist, we remove or eliminate them from the list. Some philosophers have said this about minds, or mental states. In cases of elimination, nothing can be salvaged from talk about such things as it stands. Such talk is plain false (see Lycan and Pappas 1972).

Second, there are cases which, on first appearance, seem to be ones in which I am speaking or thinking about something, but closer inspection dispels the appearance. Not all cases of apparently thinking about something turn out, on analysis, really to be such. I may not be really intending to talk about the thing in question at all. I may, for example, talk about the average family, but such talk or thought is not really about what it might seem to be about. There is no such thing as *the* average family about which I am talking.

Average-family talk or thought is merely shorthand for talking about the number of people in the country divided by the number of households (or something like this anyway). There is no such thing as the average family, nothing that should be taken ontologically seriously, beyond taking seriously households and people.

Let us call these cases, 'cases of replacement.' The original talk, r_1 ('the average family') has apparent ontological commitments of some sort. The replacing talk, r_2 ('households' and 'persons') does not have those same apparent ontological commitments (although of course it has some of its own). The burden of proof, in such cases, falls on the person wishing to show that some of the ontological commitments of the original talk are apparent but not real. He or she must indicate the replacing talk that lacks them. In these cases, our original talk gets salvaged, because it is true, if the replacing sentence without the same apparent ontological commitments is true.

What must be the case for the replacing talk, r_2, to count as an adequate replacement for the original talk, r_1? The two talks, r_1 and r_2, must convey the same information, and part of that requirement is that r_1 if and only if r_2. No two sentences can convey the same information unless they at least have the same truth values. Beyond this, matters get controversial, but the same-truth-value condition will suffice for our purposes.

It is not required that the replacement can actually be carried out in practice. Actually replacing r_1 by r_2 might be cumbersome, complicated, long-winded, or whatever. The replacement must be theoretically

possible, and the supporter of the replacement needs to show how the replacement is possible in principle, but not necessarily in practice.

Third, and closely related to the above, are cases in which our talk or thought really and not just apparently is about some entities, but they turn out *not* to be *sui generis*, since they can be reductively identified as or with another thing. These are, appropriately, cases of reductive identification.

This is the reductive identification which claims so much of philosophers' energies. There are, it may be said, physical objects, but they are really only sets of sense data; there are mental states, but they are identical with brain states; there are numbers but they are merely the same as sets of sets. Such attempts at reductive identification allow the philosophical cake to be both had and eaten. Along with common sense, it can be asserted that there are such things. But, along with some ideal of philosophical acceptability, they can be shown to be reductively identified as merely another thing after all.

Insofar as philosophical reductive identification succeeds, it allows us to shorten our ontological list, while not denying the claims of common sense. It is for this reason that the term 'reductive' is appropriate; these identifications permit a reduction in the number of basic ontological commitments. On a materialist reductive analysis, for example, there are minds, but they do not have to be taken ontologically seriously, beyond taking bodily or brain states or whatever seriously. Or, according to Berkeleyian idealism, there are tables, chairs, and so on, but they are only ideas in the mind. Once ideas in the mind are on the ontological list, adding tables and chairs would be superfluous. Tables and chairs are already on the list, if ideas in the mind are.

Again, in cases of reductive identification, our talk gets salvaged, since the original talk is true if the analysis containing the reductive identifier is (see *Reduction, Varieties of*).

2. *Entities, Properties, Facts*

Socrates or the man I met yesterday or the tallest building in New York are concrete objects. Sets and numbers and propositions, if there are such, are abstract objects, which are neither spatial nor temporal. Other things we think or speak about include spatio-temporal nonconcreta: events and actions, such as the eruption of Mount Vesuvius and the first running of a mile under 4 min; processes such as the erosion of the soil in some specific place over a specific period; and states such as the lake's remaining calm on such-and-such a night or the stillness of the air on a summer's eve or the presence of helium in a balloon.

Thinking or talking about any sort of thing is caught by reference; we refer to something, namely, to that entity our thought or language is about.This entry confines itself to discussion of facts, entities, and properties, as such referents of our thought, but no discussion of social ontology could be complete unless it brought to the fore actions, and in particular social actions (see *Joint Action* and *Action, Collective*).

Typically, when I think about something, I also ascribe some feature or characteristic to it. For example, when I say of Socrates that he is snub-nosed, or of the man I met yesterday that he is a spy, or of the erosion of the soil that it was accidental, I ascribe such a feature or characteristic. Let us call these features or characteristics 'properties.' The duality of reference and ascription is captured grammatically by the distinction between subject and predicate, and metaphysically by the distinction between entity and property. (All of this needs serious qualification, and extension to other sorts of sentences, but I forego that pleasure here.)

Are there, ontologically speaking, properties as well as entities? There are at least two ways in which to handle this question. First, there is the austere point of view that says that ascribing properties is not enough to be ontologically committed to them. There are properties, only if we *talk or think about* them, *refer to* them, and if such talk or thought cannot be eliminated or replaced by other talk or thought which makes no such ontological commitment (see Quine 1963).

There are, though, cases in which we do appear to refer to properties: (a) 'The color crimson is a shade of the color red.' It is easy to show that (b) 'The set of crimson things is a subset of the set of red things' will not do as a replacement for (a). And if no such replacement or elimination is possible, then there are properties, although they may not be *sui generis*, if any reductive identification of them is possible.

Second, there is the more relaxed approach, which says that we are ontologically committed to properties if they are actually instantiated (or perhaps even if only well defined), whether or not we refer to them or talk about them. Ascription alone, without reference, may still involve us in ontological commitment. If so, 'this visual patch is red,' which ascribes the property of being red to something, but does not refer to the property, commits us to the property of being red as well as to a visual patch, since the sentence insures that redness has at least one actual instance.

Finally, and much more controversially, some philosophers have believed that our account of what there is must include not only entities and properties, but also facts, for example the fact that some entity *e* has some property *P*. Facts have been enlisted for many purposes, one such being a correspondence theory of truth, and another being causal analysis. Are sentences true when they correspond to the facts? Is it the fact that the match was struck that caused it to light? These two particular uses of facts have come in for some devastating criticism. But even so, there may well be facts, even if they are useless in theories of truth and for analyses of causality.

If it is a fact that Socrates is snub-nosed, then is not there at least that fact? At least on first appearance, it seems that there are facts. If so, there are not only Socrates and the property of being snub-nosed, but also Socrates' being snub-nosed (a state) and the fact that he is snub-nosed. If there are facts, then they count as a kind of entity, if they can be referred to and occupy subject position in sentences. 'The fact that Fiorello LaGuardia was elected disturbed many of his opponents,' was no doubt true, so we do seem to refer to facts. Without mere replacement, or elimination, or reductive identification, we have facts on our ontological list, as well as entities and properties, states, and events.

Needless to say, philosophers have disputed whether all of these 'reifications' are really required, and whether some can be understood in terms of the others. Perhaps numbers are just sets. Perhaps propositions are just sets of similar sentences. Perhaps minds are just brains. Perhaps there is the fact that Socrates is snub-nosed, or perhaps there is no such thing. And even if there is this fact, maybe it is identical with the state, Socrates' being snub-nosed. But none of that need be decided here.

3. *Some Social Ontology*

Thought and discourse about society, whether lay or by the social scientist, raises these ontological questions too. Just as I can think about Socrates or the tallest building in New York, so too I can think about France or the Icelandic working class or the commodity market or the National Football League. Our very thought about society appears to commit us to believing in nation states, governments, classes, social structures, tribes, clubs, associations, and so on, just as our thought about Socrates and the tallest building in New York appears to commit us to believing in persons and buildings. And we are really committed to these social entities, unless we can show that these turn out to be cases in which either we can, without loss, eliminate discourse about social entities entirely, or we can replace such talk with other talk that makes no similar ontological commitment, or there is some reductive identification available for the putative social entities.

What does 'social' connote? The basic application of the term relates to properties (for more about social properties, see below). What makes a property a social property? This is, I think, a difficult question. Some entity's having a property may have a social cause or social effect, without making the property itself social. For example, the soil's being eroded may be caused by inept bureaucracy, and may result in a stock market plunge, but none of that makes the property of being eroded a social property. Whether a property is social depends on something intrinsic about the property itself, and not on the causes or effects of its being instantiated.

A property is social (roughly, without some refinements unnecessary here) if and only if from the fact that it is instantiated, it follows that at least two people exist and have an interlocking system of beliefs and expectations about one another's thoughts or actions. Intuitively, I think this gets our idea of sociality about right, although one might just think that the requisite number should be greater than two. For example, the property of jointly carrying something is not social on this account, and rightly so, because two men might jointly carry something without either being aware of what the other was doing, oblivious to each other's very existence. Jointly carrying requires two people, but makes no demands on their beliefs and expectations about one another. Purchasing a stone is, on the other hand, a social property on this account, and rightly so. No two people could be engaged in such a transaction without having a whole host of beliefs and expectations about the actions of the other.

Social entities, if such there be, are entities such that they have at least one social property essentially. France, for example, is essentially a nation; the Icelandic working class is essentially a class, a working class, and Icelandic. Once the idea of a social property is to hand, we can use it to introduce the idea of sociality as applied to other ontological categories. So our idea of the social fans outwards, as it were, from our idea of a social property.

Social entities, if such there be, are never concrete, in the way in which earlier examples of entities were. Concrete entities, like Socrates and the tallest building in New York, are exclusive space occupiers, no two of which can occupy the same space at the same time.

Social entities do, in general, appear to be space occupiers; France occupies a certain region of physical space, the Icelandic working class is physically in Iceland. But social entities are not exclusive space occupiers. Suppose a diocese of the Catholic Church, call it 'Sancta Gallia,' occupies exactly the same geographical area as does France. If so, two distinct social entities, France and Sancta Gallia, occupy the same physical space at the same time.

What, then, of the three options we have available, for dealing with these apparent ontological commitments of our social discourse? Elimination of discourse about social entities would seem to involve an intellectual loss; we could not even in principle say things we would want to say without the discourse. (Imagine where we would be without sociology, anthropology, and a part of economics, for example.) The issue here is not just one of linguistic economy. It is not that we could not nearly as easily say what we want without social discourse but could only do so in an altogether more baroque fashion; rather, the claim is that we would have no way in principle of conveying what we want to say without that social discourse.

On the other hand, replacement does seem possible for some very simple examples. A case which has been mentioned in the literature is 'The Jewish race is cohesive,' which appears to commit the speaker or thinker to an entity, the Jewish race. But, it is alleged, this is replaceable by 'Jews tend to marry other Jews,' etc., and the latter makes no such ontological commitment to the existence of a race. 'The Jewish race is cohesive' behaves very much like 'The average family has 3.2 members.' (See J. W. N. Watkins, 'Ideal Types and Historical Explanation' and 'Historical Explanation in the Social Sciences,' the first of which is reprinted in Ryan (1973), and both of which are reprinted in O'Neill (1973).)

I do not wish to claim that these replacements are never available; perhaps they do work for the case just mentioned. But it seems hard to see what replacement strategy would work for many other examples. Consider 'France is a charter member of the United Nations.' What sentence or sentences could replace this, yet convey the same information, at least in the sense of having the same truth value, and yet make no ontic commitments to a nation state and a world organization? I can here only invite the reader to suggest candidates; I myself know of none. Replacement does not work as a general strategy.

However, it is one thing to admit that there are such entities as France, the United Nations, and the Icelandic working class, but quite another to insist that they cannot be reductively identified with something more respectable. Many philosophers are drawn to reductive strategies of one kind or another: there are mental states, but they are (merely) identical with brain states; there are numbers, but they are (merely) identical with sets of sets; there are physical objects, but they are (merely) identical with sets of sense data.

So too various reductive strategies suggest themselves for social entities: France is (merely) a geographical area, a set of persons, etc. No one can 'prove' that no reductive strategy could work. What must be done, I think, is to evaluate each reductive strategy that suggests itself, one-by-one.

There are some general lessons one can draw, though. The vague thought, 'Nations are identical with the people who make them up,' is ill-formed. In any reductive identification, names must flank both sides of the identity sign. So, for instance, 'France = ⋯.' Well, identical with what? Name fillers for the space on the right-hand side of the identity sign will either be extensional or nonextensional entities, in the following sense. Sets and aggregates and land masses are extensional, since, if and only if a and b are sets or aggregates or land masses, $a = b$ if and only if a and b have exactly the same members, or components, or parts.

But no extensional entity in this sense could be identical with France, for example, since their identity conditions will differ. France can remain numerically one and the same, in spite of variation in its citizenry or residents, or land occupancy. Even if one considered the set of all actual citizens of France or pieces of land that France ever occupied at any time, past, present, and future, there would be counterfactual differences.

France could have occupied a different piece of land than it did, or had different citizens or residents than it did, but this could not be true of the set or aggregate of its citizens, residents, or lands. France is nonextensional; sets, aggregates, and land masses are extensional, and no nonextensional entity could be identical with an extensional one. (In fact, this little argument which relies on counterfactual differences assumes that 'France' and 'the set of ⋯.' both rigidly designate, since the identity required is identity across all possible worlds. But I would argue that this assumption is true.)

If France is a nonextensional entity and no nonextensional entity can be identical with an extensional one, then what about some nonextensional entity for the reductive identification of France? Examples that come to mind are things like groups. Is France to be identified with the *group* of French people? The problem with this suggestion is that a group seems itself to be a social entity, and hence hardly a candidate for the reductive identification of France with a nonsocial entity. What the determined reductivist needs to find is a nonextensional entity which is also nonsocial, for the reductive identification of France, and I am aware of no such obvious candidate.

So much for social entities. What sort of properties get ascribed to things when we think or talk socially? We can think or speak of either social entities or nonsocial ones like Socrates or the tallest building in New York and, in either case, ascribe social properties to them: 'France is a charter member of the United Nations' ascribes to a pair of social entities, France and the United Nations, the relational social property of being a charter member of; 'Socrates was mayor of Athens' ascribes (falsely, as it happens) to a nonsocial entity, Socrates, the relational social property of being a mayor of.

From the fact that social properties figure in our discourse, it does not yet follow that we refer to or think about social properties. The relational social property, being the mayor of, certainly figures in 'Fiorello LaGuardia was mayor of New York,' but the sentence is not *about* that social property. It is only about Fiorello LaGuardia and New York.

On the relaxed criterion, the sentence commits us to the existence of three things: a person, a city, and the social property of being a mayor of. On the austere criterion, since the sentence is only about Fiorello LaGuardia and New York, the sentence commits us only to the existence of a person and a city, but not to the existence of a social property.

But even were we to adopt the austere view, there are a host of arguments, modeled on ones about properties generally, which do move us to talking about social properties, and so to the existence of

social properties. Here are terse statements of two such arguments: (a) If Iceland is capitalist, and if France is capitalist, then there is some property that Iceland and France share; (b) If France is a charter member of the United Nations, then it follows that France is a member of the United Nations. The inference in (b) requires taking properties ontologically seriously.The first argument parallels one by Hilary Putnam (1980) on properties generally; the second parallels one by Donald Davidson (1980) on action.

So we talk and think about social properties, and I do not see how we could simply eliminate, do without, such discourse. Nor do I know of any discourse which could replace discourse about social properties, and which both lacked the ontological commitment to social properties but allowed us to convey the same information or messages. So, on the austere view, there seem to be social properties, as well as nonsocial ones. *A fortiori*, on the relaxed view, we get properties, social and nonsocial, just insofar as we ascribe them to actually existent things, or just in case they are well defined.

Could each social property be reductively identified either with some specific nonsocial property, or with some finite Boolean construction of such? I doubt whether this could be so. Let me give an indication of why I think this. Some social properties are both nonvariable and weak, in the following senses. A social property P is nonvariable if and only if P is instantiated, then some specific set s of interlocking beliefs and expectations exists. For example, the property of participating in the custom of drinking tea at breakfast is nonvariable, since the beliefs and expectations associated with it must be the set of beliefs and expectations about drinking tea at breakfast. A social property P is weak if and only if P is instantiated, then the interlocking beliefs and expectations are about types of events which are not themselves social. The social property of participating in the custom of drinking tea at breakfast is also weak, since the beliefs and expectations associated with it concern drinking tea at breakfast, which is itself a nonsocial event or activity.

But the most characteristic sorts of social property are neither nonvariable nor weak; they are variable and strong. Consider the property of being a mayor. If that property is instantiated, it is true that some set of interlocking beliefs and expectations must exist, but there is no one particular set that must exist. There is an indefinitely wide range of things that a mayor might be expected to do in various different societies or social settings. So the property is variable, in the sense of being indefinitely variably instantiable. Moreover the property is strong, since what a mayor is expected to do involves events and activities like receiving the ceremonial key to the city, opening formal events, representing the city at certain social or political functions, and so on, and these things are themselves social events or activities.

It seems to me that these two features of typically social properties make the reductive identification of social properties with nonsocial ones exceedingly implausible. One might have thought that a social property could be reductively identified with the nested and interlocking set of beliefs and expectations with which it is associated. But since there is an indefinitely long list of such sets associated with each typically social property, and these beliefs and expectations themselves require the existence of *social* events or activities, the reductive identification appears to fail.

I do not have the space here to discuss the issue at length of whether it is necessary to admit facts to our ontology. Frequently, the case for or against reduction of social science to physical science has been put in terms of facts; supervenience claims in social science typically enlist social and nonsocial facts as the terms of the debate. (See Maurice Mandelbaum, 'Societal facts,' reprinted in Ryan (1973) and O'Neill (1973).)

But what does seem plain is that *if* there are facts, then, in the light of what we have already said above about social properties and social entities, there must be social facts. A fact would qualify as social if either it was a fact about a social entity, or it was a fact about any kind of entity, social or nonsocial, and attributed a social property to that entity. So the fact that the United Nations contributed to relief in Kosovo is a social fact, since it is about a social entity, and the fact that LaGuardia was mayor of New York is a social fact, since it attributes a social property to a nonsocial entity, namely to a person.

Often, the positions described in the relevant literature about social ontology are posed in terms of individualism and holism, sometimes with the qualifying adjective, 'methodological' added (see *Methodological Individualism: Philosophical Aspects*). Sometimes, the positions are expressed in terms of facts, and at other times, in terms of entities. Supervenience claims for the social ('the social supervenes on the nonsocial') are sometimes posed in terms of facts, at other times in terms of properties. It might be claimed that there are no irreducible social facts, or that there are no irreducible social entities or that there are no irreducible social properties. As can be seen from the discussion above, the positions are not equivalent. For example, as I have claimed, some social facts ascribe social properties to nonsocial entities.

Rarely, if at all, is the distinction drawn between the existence of entities, properties, and facts. Does individualism in the social sciences eschew social entities but admit irreducible social properties? Or should it eschew both? What should it say about facts? Might there be distinct versions of individualism, so that it comes in stronger and weaker forms?

The lesson is that these terms, 'holism' and 'individualism,' do not name unambiguous positions unless and until various distinctions are drawn, and the ontological claims of each position are clearly

specified. (See S. Lukes, 'Methodological Individualism Reconsidered,' reprinted in Ryan (1973).)

See also: Atomism and Holism: Philosophical Aspects; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Identity and Identification: Philosophical Aspects; Individualism versus Collectivism: Philosophical Aspects; Shared Belief

Bibliography

Currie G 1984 Individualism and global supervenience. *British Journal for the Philosophy of Science* **35**: 345–58

Davidson D 1980 The logical form of action sentences. Reprinted in Davidson D *Actions and Events*. Clarendon Press, Oxford

Garfinkel A 1981 *Forms of Explanation: Rethinking the Questions in Social Theory*. Yale University Press, New Haven, CT

Gilbert M 1989 *On Social Facts*. Routledge, London

Goodman N 1961 About. *Mind* **70**: 1–24

Lycan W, Pappas G 1972 What is eliminative materialism? *Australasian Journal of Philosophy* **50**: 149–59

McKeon R (ed.) 1966 *The Basic Works of Aristotle*. Random House, New York.

Mellor D H 1982 The reduction of society. *Philosophy* **57**: 51–75

O'Neill J (ed.) 1973 *Modes of Individualism and Collectivism*. Heinemann Educational Books, London

Pettit P 1993 *The Common Mind*. Oxford University Press, Oxford, UK, Chaps. 3, 4

Putnam H 1980 On properties. Reprinted in Putnam H *Philosophical Papers*, Cambridge University Press, Cambridge, UK, Vol. 1

Quine W V O 1963 On what there is. In: *From A Logical Point of View*. Harper and Row, New York, pp. 1–19.

Quinton A 1975–6 Social objects. *Proceedings of the Aristotelian Society* **76**: 1–27

Ruben D-H 1985 *The Metaphysics of the Social World*. Routledge and Kegan Paul, London

Ruben D-H 1990 *Explaining Explanation*. Routledge, London, Chap. V

Ruben D-H 1998 The philosophy of the social sciences. In: Grayling A C (ed.) *Philosophy 2: Further Through the Subject*. Oxford University Press, Oxford, UK, pp. 420–69

Ryan A (ed.) 1973 *The Philosophy of Social Explanation*. Oxford University Press, London

Taylor B 1985 *Modes of Occurrence*. Blackwell, Oxford, UK

D.-H. Ruben

Social Protests, History of

Social protests have been integral to collective life since the first complex communities were formed. One of the oldest references to protest was found on an inscription in a tomb in Egypt's Valley of the Kings from the reign (1198–66 BC) of Rameses III in the New Kingdom translated as follows: 'Today the gang of workmen have passed by the walls of the royal tomb saying: we are hungry, eighteen days have gone by, and they sat down behind the funerary temple of Tuthmosis III … the workmen remained in the same place all day' (Romer 1982, pp. 193–5).

Social protest is defined here as contentious action undertaken collectively in response to perceived injustice or unfair action on the part of those who hold legitimate political and economic power. It seeks to achieve social (as opposed to political and economic) ends, or alternatively to restore or return to earlier ways of life. Examples of such aims are a more equitable distribution of privilege or wealth, reducing inequality among persons or groups, changing or restoring religious beliefs and/or cultural practices, and reversing cultural change. This article examines social protest in Western and non-Western societies starting in Antiquity and through the nineteenth century.

1. Protest in Imperial Rome

The authoritarian regime of Imperial Rome did not offer many opportunities for peaceful communication of ordinary people's grievances or wishes; despite the city's heavy policing, however, the large agglomerations of people in the Colosseum, Circus, and theaters offered the possibility of mass (and more or less anonymous) expressions of popular opinion. In these settings, crowds could shout demands in the Emperor's presence for the disciplining of a hated official, or object to an unpopular war. The streets too were a locus for popular protest, especially that about the shortage or high price of grain. Street disorder of this type contributed to the fall of Nero. Crowds of ordinary citizens might also call for the downfall of a usurper of the Imperial throne or express their dissatisfaction with a prefect or his policy.

The grain riot or angry complaints about the high price of bread continued to be a feature of popular protest in many parts of the world, right into the twentieth century. Most medieval and early modern European states sought to regulate the grain trade and concomitantly the price of bread at retail. These controls began to be modified and eventually removed as trade was liberalized in Western Europe, and newly formulated laissez-faire theories were put into effect in the late seventeenth century. The result was an upsurge of two types of protest: (a) crowd blockage of grain being moved on roads or waterways out of a growing region with a surplus to one with a shortage, or a city (especially an administrative city or the capital) with a stronger claim on adequate supplies; and (b) the market riot with confiscation of grain, its sale by the

crowd, and often payment (at lower than the asking price, of course) to the owner for his grain. Market riots were more likely to occur in cities or towns, sometimes along with attacks on bakers and looting of their shops.

2. Protest about Food Supply and Taxation in Old Regime France

Looking more closely at the relationship of food supply and public order in *ancien regime* France, widespread food riots beginning in the late winter or early spring (when the price of grain began to rise if the fall's crop had been small) and involving large numbers of people occurred in cycles beginning in the seventeenth century and ended only in the early 1850s. The timing of the upsurge suggests a relationship with political centralization, including policy decisions concerning economic matters and the modification of longstanding paternalist policies, especially those concerning the production and marketing of grain. Several of the modern period's great revolutions (French, 1789 and Russian, 1917) began with contention about food prices, and went on to attack—and overthrow—monarchies.

Tax revolts are another common form of protest that span the centuries of European history; the well-documented case of France is again illustrative. Increased taxation to finance the dynastic ambitions and consequent wars of the Bourbon kings made heavy demands on ordinary French men and women, to which they responded with protest. In a political system in which noble status (and ecclesiastical office) exempted the upper classes from most taxation, the fiscal burden fell disproportionately on peasants and commoners. Seventeenth century disputes over the legitimacy of new direct taxes, including attacks on collectors, were usually community affairs, but sometimes transcended single communities when the tocsin was sounded at a tax collector's first stop to arouse a whole region. New indirect taxes were farmed out to individuals (tax farmers) who had the responsibility for collecting them through private agents; both tax farmers and agents were subject to attack. Although at first, punishment was more exemplary than extensive, widespread repression became common over the seventeenth century; the label 'croquant' (a countryman armed with a stick) was first applied to peasant rebels in 1594 in southwest France, Yves Berce tells us, and continued to be used in recurrent rebellions until 1752. The titles of other seventeenth and early eighteenth-century rebellions likewise suggest the commonalty: Sabotiers (shoemakers, 1658), Bonnets Rouges (red caps, 1675), and Camisards (white shirts, intermittently from 1685 to 1710). These rebellions occurred in off-the-beaten-track areas of France that had formerly been only lightly taxed or exempt from particular taxes, so that inhabitants took extreme umbrage at the imposition of new taxes or stricter enforcement of the old.

3. Early Modern Religious Protest in Europe

The revolt of the Camisards illustrates religious protest, another type of social protest which can be found across Europe in the sixteenth to eighteenth centuries. In France, the Edict of Nantes (1598, promulgated after decades of religious wars by Henry IV) had granted freedom of conscience to Protestants all over France, permitted public worship by Protestants in designated towns, and guaranteed their civil rights; in 1685, however, Louis XIV revoked the edict, thus capping several decades of decrees limiting the effect of the policy of tolerance, and increasing the pressure on Protestants. They nevertheless continued their forbidden forms of worship (often outdoors, since they had no churches), especially in the southwest province of Languedoc; there Protestant Camisard bands conducted a guerilla war in the hills, clashing over 25 years with the forces of order. The military repression was expensive, and the rebels reduced tax revenues by attacking tax collectors and destroying property, but royal forces finally succeeded in ending Protestant protest in 1710. Over the last years of the Old Regime, a more tolerant policy emerged unevenly, alternating with periods of intensified repression.

Outside France, but in Western and Central Europe, the sixteenth century was also a period of extensive social protest connected with the Protestant Reformation. The German Peasant War has been the object of controversy since it began and Catholic authorities promptly claimed that Luther's revolt against the church had promoted peasant rebellion. Historians of the rebellion have also disagreed, with Marxists—starting with Friedrich Engels—claiming that the Peasant War was socioeconomic class warfare and non- or anti-Marxists interpreting it as a conservative movement seeking to restore Catholic principles of right and justice. More recently, Peter Blickle (1981) has again emphasized the religious underpinnings of the conflict. He argues that contrasts between urban and rural revolt are not clear cut; both urban and rural ordinary people emphasized justification through scripture. And there was only one outcome: the rising ended with the defeat of the rebels and its leaders executed.

In 1536, the Pilgrimage of Grace challenged Henry the VIII's dissolution of the monasteries in England, with a brief aftermath early in 1537. Like the revolt of the Camisards, the Pilgrimage (the term includes several more or less coordinated local risings) mixed religious issues with political and economic ones; unlike the Camisards, however, the Pilgrims were protesting Henry and his minister Thomas Cromwell's

rejection of aspects of Roman Catholic doctrine. The Pilgrims' banners portrayed the Five Wounds of Christ and they were often led by priests. Among the significant political events which preceded the revolt in 1536 were the passage of parliamentary acts which dissolved the 'lesser monasteries,' and permitted the king to name his own successor; the clergy of England then rejected the doctrine of purgatory, and accepted Henry's Ten Articles making (what seemed to some) modest changes in religious doctrine, a compromise document.

The Pilgrimage began with the rising of nine 'hosts' (popular armies, including members of the gentry, clergymen, town dwellers, and peasants) starting in Lincolnshire, followed by eight more raised in Yorkshire and elsewhere in the north. Attracting new recruits on their way to take the city of York without a battle, the rebels united around the term 'pilgrimage' to designate their undertaking to ask the king's grace in preserving 'Christ's church' and their sworn oath. Robert Aske, a member of the gentry became their spokesperson. First a truce, and then a general pardon from the king were bargained and granted (it contained promises of a parliament in York and that the abbeys which had been closed under the dissolution decree would be reopened) which led to class divisions within the hosts and in January 1537, renewed rebellion, which was put down militarily. Compared to the German Peasant War, the Pilgrims of 1536 included gentry, peasants, and townsmen, but like most early modern large-scale protests, it failed to achieve its backward-looking goals.

4. *England's Glorious Revolution and its American Colonies*

England's seventeenth century Stuart monarchs struggled with Parliament over taxation and royal prerogative, a struggle that ended in civil war when Charles I ordered the arrest of his opponents in the House of Commons. Charles I was beheaded after he was defeated in battle but refused to abandon his opposition to parliamentary claims. When the Puritan Republic ended after Oliver Cromwell's death with the restoration of the Stuarts—Charles II (1660–85) and James II (1660–88)—the struggle between kings and Parliament resumed. The struggle was finally resolved in the Glorious Revolution of 1689, when James II abdicated to be replaced by William and Mary; the outcome was a greater role in English governance for Parliament.

Wars consumed the West European mainland until the mid-eighteenth century, when monarchical governments reached a new great power balance in which Spain played only a peripheral role. France and Great Britain (following the unification under the same monarch of England and Scotland) faced each other as rivals not only in Europe but also in North America and South Asia where both had colonial interests. The following paragraphs look at social protest in other world regions, both colonized and free, starting with the British colonies in North America, continuing with two Asian examples.

During the Restoration period, the British kings sought to increase their control over their North American colonies by passing Navigation Acts, based on mercantilist economic theory which would limit colonial trade and manufacturing in order to promote English trade and manufacturing. In pursuing this goal James II replaced the proprietary governments of New Hampshire, Massachusetts, and the Carolinas with royal governors and suspended their elected assemblies. The colonists resisted these interventionist efforts and overthrew the governors of New York and Massachusetts, and replaced the proprietor of Maryland. The interests of colonial populations had diverged from those of the royal governments in England, including that of William and Mary after the Glorious Revolution (1689) in which James II was forced into exile. The 13 North American British colonies increasingly became another locus for the rivalry of France and Britain that was also being played out in Europe.

With the end in 1763 of the Seven Years War, Britain took over most of the former French colonies, while domestically, it staggered under an enormous war debt. The government searched for a solution to its liquidity crisis in the colonies, attempting to contain settlement south of Canada to the area east of the Appalachians, impose new taxes, and limit self-government in North America. These heavy-handed British policies only increased colonist dissatisfaction, and to make matters worse they were followed by new revenue generating laws: the Sugar Act (to improve enforcement of the duties on molasses), the Currency Act (which forbade the colonies from issuing paper money), and the Stamp Act (which taxed the paper used in legal documents and publications). Public protest erupted in several of the colonies, with coordinated resistance in Virginia and Massachusetts, and a meeting (the Stamp Act Congress) of delegates from nine colonies in New York in 1765 to plan further protest. Colonists boycotted imports from Britain and formed the Sons of Liberty who sponsored more public meetings; attacks on tax collectors spread. Although the Stamp Act was repealed, new taxes were passed including a duty on tea; to enforce compliance special courts were established and British troops were transferred from the frontiers to eastern cities. The dispute again escalated as colonists began new boycotts and renewed attacks on tax collectors.

The tea grievance did not disappear even though the new duties were again repealed, for Parliament granted the British East India Company a monopoly (another mercantilist policy) for the importation of tea into the colonies. This time patriots disguised as Native Americans dumped a large quantity of tea into Boston

Harbor. British authorities placed the Massachusetts colony under military rule, and open rebellion followed, starting in 1776. By 1781, the colonists had defeated Britain with the help of France and a peace treaty was signed which granted unconditional independence to the 13 colonies. Spanish America also had its colonial protest and revolutions, beginning when Spain was distracted by Napoleon's invasion in 1808 and most of its colonies in the Americas achieved independence.

5. *France and Germany before and after 1789*

Protest in continental Europe in the eighteenth century fell into the patterns described for early modern England. In France, protest about grain shortfalls or bread prices continued to dominate even as the monarchy sought to improve the circulation of grain so that regional shortages would not arouse public ire. In fact, the methods that the French state followed to reduce protest—removing market and price controls—merely agitated local populations because they could no longer count on controlled grain markets. The 'free trade' solution for France's economic problems had the effect of making matters worse from the point of view of consumers, and both Napoleon and restored monarchy after 1815 reinstated some controls in difficult times.

After the establishment of the Third Republic in 1871 both the paternalist policies and old forms of protest disappeared. Expanded political incorporation via universal male suffrage—passed early in the Revolution of 1848 but manipulated under Napoleon III to produce needed electoral majorities—was finally implemented in the 1870s. Better harvests and the more efficient transportation that the railroads provided for moving grain to regions with poor harvests helped to reduce protest as well. Worker unionization became legal in 1864 and with time, strikes became a routinized form of collective action in France as well as Britain. This does not mean that violence disappeared, for violence is an interaction between police (or the military in cases of martial law) and protesters. The point is that both more professional methods on the part of the authorities handling strikes and protest and greater formalization and nationalization of political parties and workers' organizations led to the development of routinized policies for dealing with collective protest, thus reducing its impact (Tilly 1971).

The late nineteenth and early twentieth century was a watershed in typical forms of collective protest which the German states also crossed. Reports from German states such as Saxony in 1842 detailed how rural craftsmen who were dependent on wages to buy bread had mobbed grain dealers who had bought and were moving grain out of the region to more distant urban areas (Tilly et al. 1975, p. 192). By the early twentieth century the industrialization of production, unionization of laborers, and urban populations had increased disproportionately in Prussia. (Unified Germany was a federation of formerly independent states, the most powerful being Prussia, and urban public order was the responsibility of the individual states.) In 1910 there was a massive disciplined demonstration in Berlin in which the Socialist Workers Party (SPD) leaders called for a political response: reform of the Prussian three-class electoral laws. The demonstration was then still a comparatively recently adopted form of protest through which ordinary people could bring their dissatisfaction with public policies to the attention of the government.

6. *Social Protest in Nineteenth-century Asia*

We turn now to Asia, to examine two non-European cases of nineteenth-century protest: in the Dutch East Indies and China. The Dutch East India Company (founded in 1602) had displaced the Portuguese in Southeast Asia by seizing Malacca at the southern end of the Malay Peninsula and conquering the kingdom of Acheh on Sumatra as well as the western part of Java by mid-century. The Dutch built their administrative capital, Batavia in western Java, leaving the rest of the island for some years to the Javanese kingdom of Mataram. A series of civil wars and regional revolts involving the Dutch and local princes ended in the 1750s with the division of Mataram into four new princely states, but lower level war continued. By then, Chinese immigrants controlled local and inland trade and the Dutch ran plantations and long distance trade in coffee and teak. By the late nineteenth century, Michael Adas (1979) shows, Dutch control over Javanese society had profoundly undermined the island's formerly independent economy. The Dutch system for the collection of tolls on roads and canals (using Chinese subcontractors) was similar to the seventeenth and eighteenth century French tax farming and aroused similar objections. Moreover, the Dutch could not win militarily because they lacked knowledge of Java's interior and in the end, they chose expediency and limited goals, preventing further unrest and raising new revenues to pay the costs of the war. This was done by forcing Javanese peasants to dedicate part of their time to cultivating cash crops to be sold overseas by the government. Although peace was restored, Dutch agricultural policy severely distorted Javanese social and demographic patterns, and drained its economy as Clifford Geertz (1963) demonstrates in his study of 'agricultural involution.'

China was never fully colonized, but Western intervention in its economic and political affairs beginning in the eighteenth century nevertheless had a great impact on its social and economic development. The Taiping Rebellion of the mid-1800s serves as an example here. Jesuit missionaries had been active for some time in the area around Guanzhou (Canton), as were British merchants buying tea and looking for customers to buy their products, with little success.

The immediate result was a negative trade balance for the British; to remedy this European merchants together with Chinese partners imported opium from India into China through Canton, a profitable commerce that grew rapidly in the early years of the nineteenth century. When the Qing court belatedly addressed the problem in 1839 and sought to cut off the trade altogether, the British easily established their superiority in gunboats and arms and forced the Chinese to sign a treaty that opened five ports to them with very low tariffs. Soon after, a Chinese treaty with the United States legalized the importation of opium by foreigners. Chinese resentment of both missionary efforts to convert them and Westerners' privileges increased.

The study of C. K. Yang (1975) of 'mass action incidents' (group actions involving public issues) that he links with others to constitute 'events' shows an upsurge of incidents in the years 1846 to 1875. The best known of these events is the Taiping Rebellion, which started in the poor and backward region of Guanxi, to the west of Guanzhou. Its originator and leader was a disappointed office seeker who encountered Protestant missionaries in Guangxi and—inspired by the Christianity they preached to him—developed his own messianic version of the message. Believing himself to be the brother of Jesus Christ, he would found a new Heavenly Kingdom (Taiping) that would bring peace on earth.

The Qing government sent a military force to Thistle Mountain, where Hong had settled, to arrest the Taiping leaders, but the mission failed, and the Taiping continued to preach. New converts extended the membership beyond the Taiping. Life among the Taipings was highly regimented, with sex-segregated military work teams; women were forbidden to bind their feet and were welcomed as workers and soldiers. The Qing forces of order were able to chase down the Taipings who had crossed the Yanzi River and in 1853 occupied Nanjing and held it for more than 10 years. The rebellion was only crushed after the French and British, who had invaded Beijing in 1860 to punish the Qing for not following up the treaties ending the Opium War, joined the Qing army in crushing the Taipings. Frederic Wakeman Jr. (1966) concludes that the Taiping Rebellion earned the unhappy record as the world's bloodiest civil war, in which tens of millions were killed in fighting and executions. As social protest, it combined many incidents and lasted for years, but nevertheless is considered one event because of the numerous links between the incidents.

7. *Twentieth-century Protest in the West*

The twentieth century has seen new or reinvented forms of social protest such as boycotts, sit-ins, and more formalized social movements (q.v.). Citizens of most national states have by and large received rights to participate fully in the political process, but the extent to which governments respect those rights continue to differ. Whether citizens can influence policy formally (through elections and representative government) or informally (through protest) is still disputed between citizens and states in many parts of the world. What has changed is greater consensus between protestors and authorities that has largely formalized protest and how it is expressed. Even in the West, however, social protest may become violent either through design (for example, the anarchist protest at the World Trade Organization's 1999 meetings in Seattle), or through the breakdown of policing procedures that have been developed in Europe and North America to handle such situations. The contours of social protest continue to shift in many countries, especially in those in which channels for making citizens' opinions known are very limited or non-existent and where governments seek to limit or even prevent popular participation. Social protest has not disappeared: it continues to occur in both authoritarian states and those with long-established formalization of procedures as a means for expressing dissent or the desire for change.

See also: Right-wing Movements in the United States: Women and Gender; Social Movements and Gender

Bibliography

Adas M 1979 *Prophets of Rebellion: Millenarian Protest Movements Against the European Colonial Order*. University of North Carolina Press, Chapel Hill, NC

Africa T 1971 Urban violence in imperial Rome. *Journal of Interdisciplinary History* **2**: 3–19

Berce Y-M 1990 *History of Peasant Revolts: The Social Origins of Rebellion in Early Modern France* (Trans. Whitmore A). Cornell University Press, Ithaca, NY

Blickle P 1981 *The Revolution of 1525: The German Peasant War from a New Perspective* (Trans. Brady T A, Midelfort H C E). Johns Hopkins Press, Baltimore

Bush M 1996 *The Pilgrimage of Grace: A Study of the Rebel Armies of October 1536*. Manchester University Press, Manchester, UK

Davies C S L 1968 The pilgrimage of grace reconsidered. *Past and Present* **41**: 54–76

Geertz C 1963 *Agricultural Involution: The Process of Ecological Change in Indonesia*. Association of Asian Studies, University of California Press, Berkeley, CA

Romer J 1982 *People of the Nile: Everyday Life in Ancient Egypt*. Crown, New York

Scribner B, Benecke G (eds.) 1979 *The German Peasant War of 1525: New Viewpoints*. George Allen and Unwin, London

Tilly C 1986 *The Contentious French*. Belknap Press of Harvard University Press, Cambridge, MA

Tilly C, Tilly L A, Tilly R 1975 *The Rebellious Century*. Harvard University Press, Cambridge, MA

Tilly L A 1971 The food riot as a form of political conflict in France. *Journal of Interdisciplinary History* **2**: 23–57

Wakeman F 1966 *Strangers at the Gate: Social Disorder in South China, 1839–1861*. University of California Press, Berkeley, CA

Weller R P 1994 *Resistance, Chaos and Control in China: Taiping Rebels, Taiwanese Ghosts and Tiananmen*. University of Washington Press, Seattle

Yang C K 1975 Some preliminary statistical patterns of mass actions in nineteenth-century China. In: Wakeman F, Grant C (eds.) *Conflict and Control in Late Imperial China*. University of California Press, Berkeley, CA, pp. 174–210

L. Tilly

Social Psychology

Social psychology may be defined as the scientific study of how people perceive, influence, and relate to other people. Explicit in this definition are the field's modus operandus—empirical, data-based investigations of theories and hypotheses derived from those theories—and its goal of understanding the processes by which people think about and interact with others. Traditionally, this very broad umbrella incorporates diverse specific topics, including, but not limited to, attitudes, social cognition, decision-making, social motivation, prejudice and stereotyping, social identity, persuasion, altruism, aggression, interpersonal attraction, close relationships, conformity, group dynamics, group productivity, and conflict resolution, all of which are discussed in more detail elsewhere in this Encyclopedia.

Because social interaction is fundamental to most human behavior, it is useful to consider how the perspective of social psychology fits with that of the other behavioral and social sciences. Although in its early development social psychology was closely aligned with sociology, primarily because of the shared interest in social interaction, since the 1960s these fields have become largely independent of each other. The social psychological perspective is fundamentally psychological; it tends to focus on the thoughts, feelings, and behavior of individuals, whereas the sociological perspective is more immediately concerned with the analysis of social structure and the operation of social and societal systems. Similarly, social psychology tends toward a relatively individual-centered analysis of certain questions that other disciplines such as anthropology, economics, and political science generally address at more systemic levels of analysis (although the recent emergence in those fields of such specialties as game theory and behavioral economics suggest growing commonalities).

In contrast, social psychology's interest in situations (discussed below) and normative patterns of behavior differs from that of personality psychology (see *Personality Psychology*), which tends to be concerned with individual differences and their ramifications. Whereas personality psychologists seek to understand consistencies of a single individual's behavior across varying situations, social psychologists seek to understand consistencies across persons in their responses to particular situations. Thus, the social psychological level of analysis may be regarded as a mid-station between those theoretical systems that conceptualize behavior on the one hand in terms of the individual's unique internal properties and on the other hand in terms of the external social structure.

1. Core Principles of Social Psychology

1.1 The Situational Context of Behavior

Social psychology traditionally has emphasized the latter term in Kurt Lewin's classic dictum, $B = f(P,E)$—behavior is a function of the person and the environment—to reflect the fact that behavior varies, often profoundly so, depending on the social environment in which it occurs (see also: *Lewin, Kurt (1890–1947)*). The term social environment typically is used in a very broad manner, encompassing not only patently interpersonal examples (e.g., whether there is conflict or commonality of interest between partners, whether or not a new acquaintance is a member of one's in-group, and whether performance feedback is supportive or controlling) but also its relatively more impersonal features (e.g., whether persuasive appeals are based on style or substance, whether induced behavior is logically consistent or inconsistent with personal beliefs, and whether circumstances encourage deliberate or automatic processing of social information). Regardless of which variables are included under the heading of situational context, however, social psychology's published literature is replete with abundant evidence demonstrating that context is responsible for clear, substantial, and theoretically informative differences in behavior.

Although situationism is sometimes positioned conceptually as if it were the logical antithesis of the influence of personality, in reality nearly all contemporary social psychologists believe not only that both the situational context and individual differences matter, but that each affects the other interactively and continuously. For example, situations must be perceived and interpreted by the individual to possess relevant features in order for influence to occur; these interpretations are often shaped significantly by person factors (which are not limited to traits but may also include goals and motives). Similarly, personality traits are likely to be influential only to the extent that the situation affords relevant opportunities for their expression (e.g., social anxiety may have relatively little impact during interaction with close friends, but substantial effects with strangers). Individual differences also play an important role in determining which

situations the individual enters or avoids. Thus, it is most appropriate to view social psychology as emphasizing the complex interplay of forces within the person and in the outside situation in casting an individual's interpretation of, and response to, situational contexts.

1.2 A Process Approach is Most Informative

Phenomena may be regarded in terms of their distinctive and unique features or as the product of more general processes. Social psychologists, because they believe that certain general principles underlie many specific behaviors observed in everyday social life, typically focus their attention on these processes shared in common. For example, a social psychologist interested in prejudice likely would consider psychological mechanisms responsible for negative affect toward members of outgroups, rather than the details of prejudice toward one or another particular outgroup (except, of course, insofar as those details might be informative about the general process). Similarly, social psychologists study interaction between spouses less so to understand marriage *per se* and more so to discern how high degrees of interdependence influence social interaction. This emphasis does not imply that social psychologists are disinterested in the relevance of their research to particular instances; rather it denotes their belief that theoretical analysis grounded in broadly applicable principles is likely to be most revealing.

The field's emphasis on general processes is evident in the concerted attention routinely invested to identify causal mechanisms responsible for a given phenomenon. That is, once a phenomenon has been demonstrated, researchers almost invariably seek to explain its operation by ascertaining the relevant underlying mechanisms—in other words, to know when, why, and how something occurs. Typically such investigations address questions of moderation—establishing conditions under which a phenomenon is augmented or diminished—and of mediation—determining the specific step-by-step operations by which one variable produces changes in another variable. In recent years, the field has been particularly attentive to mediation by cognitive and affective processes, as described below. Considerable attention also is devoted toward ruling out alternative explanations; that is, toward conducting fine-grained experimental tests capable of distinguishing among alternative, often somewhat similar, explanations for a given phenomenon.

1.3 Basic Theory and its Application

Kurt Lewin's other classic dictum was that 'there is nothing so practical as a good theory.' Many social psychologists, in fact, initially were attracted to the discipline because of its potential for generating solutions to the pressing social problems of the day. Thus, the oft-cited distinction between 'basic' and 'applied' research is to many social psychologists somewhat misleading in the sense that basic research is often triggered by contemporary events and societal needs, and in turn may produce important evidence not only for better understanding these problems but also for generating interventions. The credo that sound theory and rigorous empiricism provide the firmest possible foundation for social application underlies Lewin's conviction, one that is shared by many current researchers. Effective applications require good theory, and good theories have real-world relevance.

As the social psychological literature has grown, so, too, has the dissemination of its principles and findings into the culture at large. As a result, the impact of social psychology is sometimes difficult to trace, although numerous examples are readily apparent, ranging from concrete interventions derived explicitly from experimental findings to more amorphous influences on prevailing public opinions. Documented impact of social psychological research can be found, for example, on such contemporary social issues as prejudice and intergroup relations, sex discrimination, healthcare, the operation of the legal system, and family violence.

2. A Brief History of Social Psychology

Social psychological thinking has deep roots. Rudimentary versions of contemporary theories are evident in many ancient sources, such as the Bible and the writings of Aristotle, Plato, and the other Greek philosophers. But it was not until the turn of twentieth century, with the field's first experimental studies and publication of the first textbooks explicitly bearing the name, *Social Psychology*, that the scientific study of human social interaction began to cultivate its own disciplinary identity. Although the field's development has been more or less continuous since then, two distinct and noteworthy trends mark its intellectual progression.

Much of the field's most influential early work was concerned with social influence, namely the processes by which the real or imagined presence of others affects belief and behavior. Research by Muzafer Sherif and Solomon Asch are considered landmarks, but the most durable and widespread impact came from Lewin's elegant Field Theory, which conceived of social relations in terms of interdependent force fields, such that each organism (i.e., person) acts on and responds to the other organisms in its space. As compelling as Lewin's conceptual model was his emphasis on experimentation as a formal method for evaluating theoretical propositions, an emphasis that became increasingly popular during the heady post-World War II expansion of psychological science. Field theory, especially in its application to investigating longstanding questions about group dynamics and

social influence, and its accompanying experimental method quickly became a dominant theme in the field's growth.

In the late 1950s and 1960s, social psychology's purview began to widen, such that social influences on the mental life of individuals came to occupy center-stage. Notable in this trend were studies of the impact of social feedback on self-appraisals, including Leon Festinger's theory of social comparison processes, Stanley Schachter's studies of self-attribution of emotion, and Festinger's subsequent theory of cognitive dissonance, which concerned self-perceived consistencies and inconsistencies between one's beliefs and behavior. In several respects, this movement set the stage for the so-called cognitive revolution of the 1970s and 1980s, in which principles of cognitive psychology were applied to social stimuli (i.e., other people, groups) and phenomena. Fundamental to the original conception of social cognition, as this subdiscipline came to be called, was the idea that many cognitive processes are designed to enable people to contend with the basic interpersonal tasks of everyday life. In their processing of social information, it was argued, people often function as naive scientists, seeking, within human limits, to comprehend cause-and-effect associations among the many pieces of social information encountered (see *Heuristics in Social Cognition*). The earliest emphasis in social cognition research was on understanding people's attributions of causality, led by the influential theories of E. E. Jones and Harold Kelley. Subsequent work has broadened this purview considerably.

More recent trends in social psychology balance the emphasis on cognitive processes with a functionalist perspective, or as Fiske (1992) put it, 'thinking is for doing.' For social psychology, the doing, of course, is relating to others in the social world. Social cognition is thus best considered as a mediating process between the individual's interpersonal goals and circumstances on the one hand, and his or her social behavior on the other hand. Embodying this perspective, contemporary research has increasingly emphasized affect in mediating the individual-social behavior link as well as the importance of the self and its motives. The logic of this latter work assumes that because action is functional and purposive, reactions to particular situations may help reveal the social motives operative in that context. Also, there is increasing recognition of the nature of ongoing relationships as part of the interpersonal context of behavior.

3. *Methods of Social Psychological Research*

It is no accident that social psychologists pride themselves on being rigorous, broad, and clever methodologists. The field's domain is wide—ranging from society-wide intergroup relations to intimate dyads—and its principled emphasis on identifying cognitive and affective mediators—ranging from cultural norms to idiosyncratic emotions—dictates that its research methods be flexible and diverse, yet suitable to uncovering and disentangling often subtle phenomena. One reason why social psychologists are in demand across a variety of disciplines and applied settings is their deftness in balancing the demands of experimental rigor, the necessity of adapting methods to diverse questions, and the various methodological issues inherent in studying human behavior.

Social psychology's favored method is the laboratory experiment, in which research participants are randomly assigned to one of several conditions defined by an experimenter so as to permit evaluation of a prespecified hypothesis about cause-and-effect. This emphasis follows directly from the high priority ascribed to establishing an effect's internal validity: laboratory manipulation allows experimenters to construct conditions whose carefully controlled differences explicitly contrast alternative explanations for a phenomenon while ruling out experimental artifacts. Contemporary social psychology has myriad other strategies and techniques in its toolkit, reflecting the realization that many questions are more appropriately addressed by different methods and the further realization that, because all methods are limited in one way or another, the validity of a research program ultimately depends on its methodological diversity. Other popular methods include quasi-experiments (when manipulation is not possible), surveys and questionnaires, daily event recording, behavioral observation, and meta-analysis.

Technological innovation, both from within the field and by expropriating advances in other disciplines, has facilitated researchers' ability to ask more sophisticated questions of their data, and to answer those questions with a greater degree of precision. Examples of these include computerized methods for investigating cognitive or affective mediation of basic social processes, especially when such mediation occurs outside of the individual's conscious awareness (e.g., priming and response latency methods), methods for assessing psychophysiological systems that underlie social behavior, internet-based data collection, and videotape-based protocols for fine-grained analysis of observable behavior in dyads and groups. The social psychologist's readiness to capitalize on new methodologies that provide fresh perspectives on the field's longstanding conceptual interests is further demonstrated by the eager acceptance of new statistical procedures in the literature. Recent and noteworthy examples include structural equation modelling, multilevel models, and methods for analyzing data in interdependent groups and dyads.

4. *The Future of Social Psychology*

Because an understanding of social relations is in one way or another fundamental across all of the social and behavioral sciences, social psychology is a cross-

roads discipline. Managing that position, that is, acquiring theoretical insights and new methods from neighboring disciplines while exporting its own theories, findings, and tools to those disciplines, represents an important challenge for social psychology. Several boundaries seem particularly promising at this juncture: neuroscience, with its ability to identify markers of social psychological phenomena in the brain; cognitive psychology, which enhances the field's understanding of mental representation of social entities; evolutionary psychology and behavioral genetics, with particular reference to identifying and understanding the specialized mechanisms created by evolution in making humans the most social of animals; and the study of culture, a contextual variable whose specific cognitive, affective, and interpersonal manifestations ultimately influence all social behavior.

At the same time, in its zeal to embrace related disciplines, social psychology must remain focused on its core phenomena. Despite the numerous important empirical findings and theoretical insights of the past century, many more questions remain to be asked and answered. The reasons underlying the importance of building bridges with related disciplines are one and the same as the reasons why social psychology must maintain its commitment to continued investigation of the field's core processes: that many of the most important unresolved questions about human functioning and well-being concern the processes by which people perceive, influence, and relate to other people.

See also: Attitude Change: Psychological; Attitude Formation: Function and Structure; Attitudes and Behavior; Attributional Processes: Psychological; Automaticity of Action, Psychology of; Cognitive Dissonance; Communication and Social Psychology; Conflict and Conflict Resolution, Social Psychology of; Cultural Variations in Interpersonal Relationships; Deception, Psychology of; Group Decision Making, Social Psychology of; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Interactionism and Personality; Intergroup Relations, Social Psychology of; Interpersonal Attraction, Psychology of; Justice: Social Psychological Perspectives; Love and Intimacy, Psychology of; Mental Representation of Persons, Psychology of; Motivation and Actions, Psychology of; Obedience: Social Psychological Perspectives; Person Perception, Accuracy of; Reactance, Psychology of; Self-conscious Emotions, Psychology of; Social Categorization, Psychology of; Social Cognition and Affect, Psychology of; Social Comparison, Psychology of; Social Dilemmas, Psychology of; Social Facilitation, Psychology of; Social Psychology, Theories of; Social Relationships in Adulthood; Stereotypes, Social Psychology of; Stigma, Social Psychology of

Bibliography

Fiske S T 1992 Thinking is for doing: Portraits of social cognition from daguerreotype to laser photo. *Journal of Personality and Social Psychology* **63**: 877–89

Gilbert D T, Fiske S T, Lindzey G (eds.) 1998 *The Handbook of Social Psychology*. McGraw-Hill, New York

Hewstone M, Manstead A S R (eds.) *Blackwell Encyclopedia of Social Psychology*. Blackwell, Oxford, UK

Jones E E 1998 Major developments in five decades of social psychology. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, Vol. 1, pp. 3–57

Kruglanski A, Higgins E T (eds.) 1996 *Social Psychology: Handbook of Basic Principles*. Guilford, New York

Reis H T, Judd C M (eds.) 2000 *Handbook of Research Methods in Social and Personality Psychology*. Cambridge University Press, New York

Ross L, Nisbett R E 1991 *The Person and the Situation: Perspectives of Social Psychology*. McGraw-Hill, New York

Smith E R, Mackie D M 2000 *Social Psychology*. Psychology Press, Philadelphia, PA

Taylor S E 1998 The social being in social psychology. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York, Vol. 1, pp. 58–95

H. T. Reis

Social Psychology: Research Methods

Social psychological research methods involve methods in use throughout the behavioral and social sciences. The goal of social psychology is to understand human behavior as it is affected by and affects social contexts. Accordingly, social psychologists attempt to measure and explain characteristics of individuals as social actors and characteristics of the social environments in which they act. Importantly, they also attempt to understand the interactive effects of both individual and environmental variables. Social psychological methods are fundamentally empirical methods, using observation and data to guide theoretical development. Ultimately, however, those methods are simply tools in the development of theoretical insights into the nature of human social behavior.

1. Research Validities

Methods of research in social psychology are useful only in so far as they permit one to critically evaluate and develop theoretical claims. Accordingly, a set of research validities that represent criteria by which any given piece of research, and any program of research, can be evaluated is defined. Ultimately research methods that permit the accomplishment of these

validities permit the gathering of data that will be maximally informative for theory construction. Importantly, however, as will be discussed, there are no perfect methods that simultaneously maximize all of the research validities. Ultimately, convergence on theoretical insights is only gained through the use of multiple methods of research, triangulating across individual studies and individual methods, to maximize the validity of an entire program of research (for general discussions of research validities, see Brewer 2000, Cook and Campbell 1979).

1.1 Construct Validity

Theoretical claims in social psychology typically involve statements about abstract and poorly defined constructs. So, for instance, small group research in social psychology has been interested in the role of alternative leadership styles on the successful functioning of the group. Researchers in intergroup relations have explored how levels of group cohesion affect the derogation of outgroups. Research in person perception has explored how information that is inconsistent with prior beliefs about someone is cognitively integrated into an overall impression of what that person is like. In each of these examples, the phenomena that social psychologists are interested in understanding are quite vague constructs, amenable to multiple alternative definitions.

To gather data that are informative for theory development, these constructs must be measured. The success of that measurement determines the construct validity of any piece of research. In other words, construct validity concerns the extent to which the constructs of theoretical interest have been successfully captured by the variables actually measured in research. The fundamental tenant of construct validity is that multiple operationalizations of any construct are mandatory. Any one measure is only an imperfect indicator of the construct it is supposed to measure. Confidence in measurement is produced only if multiple operationalizations tend to converge or triangulate upon the same answer.

1.2 Internal Validity

Theoretical claims in social psychology typically involve claims about the causal impact of constructs upon each other (e.g., social norms cause conformity, persuasive communications cause attitude change). Once the constructs have been operationalized, then research that is informative for theory construction and evaluation is research in which one can argue not only that one variable is associated with another, but also that one variable (the independent variable) is causally responsible for variation in another (the dependent variable). Research that is more internally valid is research where causal claims are more defensible.

Internal validity is affected by the nature of the research design that is employed. Ultimately, confidence in causal claims is only possible in the case of what are called randomized experimental designs, in which the units of observation (typically persons) have been randomly assigned to levels of the independent variable. For instance, to examine whether group sizes (the independent variable) influences the quality of a group decision (the dependent variable), people would be randomly assigned to groups of different sizes. Random assignment means that the researcher typically must control and manipulate the independent variable. Random assignment is importantly not the same thing as haphazard or uncontrolled assignment. It involves careful planning so that the probability of any unit occurring in a particular level of the independent variable is constant across all experimental units.

For many research questions in social psychology, randomized experiments are not feasible. Accordingly a variety of quasi-experimental designs can be employed. As is later described, these involve the collection of longitudinal data, without random assignment of units to levels of the independent variable. Finally, from the point of view of internal validity, cross-sectional studies (also known as correlational) studies are the least valid, although data from such studies can be quite informative for other reasons. Additionally, the absence of covariation between the independent and dependent variables in a correlational study would certainly lead one to doubt the existence of a causal relationship between the two.

1.3 External Validity

Data are always collected from specific units in specific settings. Obviously, theoretical conclusions require generalization from those units and settings to broader populations of theoretical interest. The external validity of any given piece of research indexes the extent to which this sort of generalization is feasible.

In general, generalization from specific samples to populations of interest is only possible if probability-sampling methods have been used to select the units and settings from which to collect data. Although widely used in social psychology, nonprobability-sampling methods (e.g., gathering data from participants who are given experimental credit in introductory psychology classes) pose serious threats to external validity.

1.4 Tradeoffs Among Research Validities

It is exceedingly likely that there are tradeoffs among these validities in any particular study. For instance,

to effectively manipulate an independent variable in order to maximize internal validity by the use of a randomized experimental design, the construct validity of the manipulated independent variable may be seriously compromised. Historically, much of the research reported in the leading social psychological journals has relied upon randomized experiments conducted in laboratory settings with participants recruited from introductory psychology pools. Implicitly, such research has assigned internal validity paramount importance while compromising the other two validities. Although tradeoffs among the validities may be inevitable in any particular study or data collection effort, it is important to employ a variety of different research strategies and designs across a line of research, moving from the laboratory to the field, from experimental designs to correlational ones, from manipulated independent variables to carefully measured ones. In this way, across a series of studies, it becomes possible to maximize all three sorts of research validities.

2. *Research Designs*

As just defined, randomized experimental designs, in wide use in social psychology, involve the random assignment of units to levels of the independent variables. Most frequently, multiple independent variables are included in crossed experimental designs, so that their separate, as well as interactive effects, can be assessed simultaneously. The interactive effects of multiple independent variables have been of great theoretical importance in social psychology, as they suggest, for instance that the impact of situational or environmental factors may depend on characteristics of the units being assessed (i.e., person by situation interactions).

If the effects of independent variables are considered to be short-lived, then a within-subjects research design may be employed to maximize power. In other words, each experimental unit occurs in each level of the independent variable (or variables). Random assignment in this case means that the order of the levels of the independent variables to which units are exposed is counterbalanced, with units randomly assigned to the different possible orders.

Many experimental designs in social psychology involve some mixture of between-subject and within-subject independent variables, and these are typically crossed so that their interactions can be assessed. The typical analytic approach to such data involves the use of analysis of variance.

Quasi-experimental designs involve the longitudinal assessment of units on the dependent measure(s) both before and after exposure to levels of the independent variable. The classic quasi-experimental design is the non-equivalent control group design in which units are assessed on the dependent variable twice, once as a pretest and once as a post-test. Some units are exposed to the experimental treatment and others to the control condition (i.e., the two levels of the independent variable), but the assignment process is typically an uncontrolled one in which the assignment rule, rather than being random, is unknown. Analyses typically involve the analysis of differences between conditions on the post-test, attempting to remove or adjust for any pre-existing differences on the pretest (through an analysis of covariance or some other adjustment procedure).

A more elaborate version of the nonequivalent control group design uses the same basic approach except that multiple pretests and multiple post-tests are assessed. Such designs permit greater confidence about treatment effects because differences in units in spontaneous growth can be modeled, and these differences can then be ruled out as competing explanations for any condition differences.

A particularly strong version of this quasi-experimental design would be one in which the treatment variable is turned on or off at different points along the time sequence during which the dependent variable is being assessed. There may be multiple measures of the dependent variable over time and at some time during that sequence the treatment gets turned on, but the time point at which that occurs varies between units (with some units serving as control units where the treatment never gets turned on).

The regression—discontinuity quasi-experimental design is one in which a known, rather than a random assignment rule is used to assign units to either the treatment or control conditions. Although seldom used in basic social psychology, this design has received attention in more applied research endeavors (e.g., deciding who will receive a new and scarce medical treatment based on the severity of symptoms manifested by patients).

And finally, as defined above, correlational designs involve the cross-sectional measurement of both the independent and dependent variables. (See Cook and Campbell 1979, Judd and Kenny 1981, Smith 2000, and West et al. 2000 for discussions of alternative research designs and their implications.)

3. *Measurement*

As discussed under Sect. 1.1, measured variables in social psychology are assumed to be only imperfect indicators of the constructs that are of theoretical interest. Classic measurement theory has focused on random errors of measurement, assuming that variation in measured variables derives from two sources: true score variation and random error variation. According to this classic approach, the reliability of a variable is defined as the proportion of variation in the

variable that is true score variation. A variety of procedures for assessing reliability have been developed from this classic model, the most widely used being coefficient alpha which assesses the degree to which a set of items share true score variation.

More recent approaches to measurement take a more fine-grain approach, assuming those errors of measurement can involve systematic as well as random errors of measurement. Accordingly variation in any variable can be decomposed into three sources: variation due to the construct of interest; variation due to other constructs that represent sources of systematic error; and random error variation. Convergent validity assesses the contribution of the construct of interest; discriminant validity assesses the contribution of systematic sources of error variation; and reliability assesses the contribution of random errors of measurement. Procedures for disentangling these sources of error variation have been developed by collecting data according to the multitrait-multimethod matrix, where multiple measures of multiple constructs are assessed, with various measures sharing systematic sources of error, particularly due to measurement methods (Campbell and Fiske 1959).

In addition to these approaches to measurement, based on the presumption that measurement adequacy is best assessed by patterns of covariation across multiple alternative measures, some have argued for a more axiomatic approach to measurement (Dawes and Smith 1985). This approach is based on the presumption that regularities of individual responses should follow certain axiomatic principles deduced from the constructs that are assessed (e.g., transitivity). In general, this approach has not proved fruitful for empirical work since random error in responses often prohibits strong tests of these axiomatic principles. Nevertheless, a few examples, such as Guttman and Thurstone scaling, are of historical importance in the domain of attitude measurement (see Judd and McClelland 1998 for a general treatment of measurement in social psychology).

4. *Analysis*

As discussed above, historically the dominant empirical approach in social psychology has involved the use of experimental designs, with independent variables orthogonally manipulated either within or between participants. Accordingly, the dominant analytic approach to data has been the analysis of variance. As researchers have increasingly extended their bag of research tools to include field studies and other sorts of research designs, more general approaches to data analysis, based on multiple regression and the general linear model, have become more prevalent. Increasingly, researchers are employing multiple regression as a general data analytic approach, incorporating analysis of variance through the use of contrast-coded predictors (Cohen and Cohen 1983, Judd and McClelland 1989).

Longitudinal data structures as well as more sophisticated measurement models have resulted in increasing use of latent variable structural equation modeling procedures (Bollen 1989, Kline 1998). In general these procedures allow researchers to examine and test models of data that incorporate both systematic and random measurement errors, estimating the linear relations among latent constructs that are each assessed with multiple indicators. Although sample size requirements and other restrictions have placed limits on the use of these procedures in social psychology, the general data analytic approach that they represent is particularly promising.

Data analysis strategies that allow for correlated errors of measurement are particularly promising in the subfield of social psychology devoted to the study of dyadic and small group interaction. Traditional inferential procedures normally assume that observations are independent of each other and this assumption is nearly always violated in dyadic and small group research where individual persons interact and are jointly influenced by situational factors. Accordingly, considerable work has been devoted in recent years to procedures for the analysis of dependent data structures. The most general of these approaches involve multilevel or hierarchical linear models, where individual observations are nested under factors whose levels are themselves treated as random observations. Traditional analysis of variance and regression procedures do not allow for this sort of hierarchical data structure with random factors (see Bryk and Raudenbush 1992, Kenny et al. 1998).

Beyond the domain of small group and dyadic research, these hierarchical models are also particularly useful in longitudinal data structures, where individual observations are repeatedly taken from participants over time and one is interested in modeling growth or change in response to various situational factors that also vary across time. Data from various quasi-experimental designs are particularly likely to exhibit this sort of structure.

5. *Conclusion*

Although experimental laboratory research procedures have been the dominant approach to research in social psychology, increasingly researchers are using a wide range of research designs, measurement procedures, and data analytic techniques to address the multifaceted research questions of social psychology. Certainly there will remain an important role for laboratory experimental work in social psychology. However, the wide range of methods now used throughout social psychology offers tremendous po-

tential for addressing the diversity of theoretical questions addressed by social psychology in very rich ways.

See also: Evolutionary Social Psychology; Experimental Design: Randomization and Social Experiments; Experimental Laboratories: Social/Pyschological; Experimentation in Psychology, History of; Observational Methods in Human Development Research; Small-group Interaction and Gender; Social Psychology; Social Psychology: Sociological; Social Psychology, Theories of; Structural Equation Modeling

Bibliography

Bollen K A 1989 *Structural Equations with Latent Variables*. Wiley, New York

Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park, CA

Brewer M B 2000 Research design and issues of validity. In: Reis H T, Judd C M (eds.) *Handbook of Research Methods in Social and Personality Psychology*. Cambridge University Press, New York

Campbell D T, Fiske D W 1959 Convergent and discriminant validition by the multitrait-multimethod matrix. *Psychological Bulletin* **56**: 81–105

Cohen J, Cohen P 1983 *Applied Multiple Regression/Correlation for the Behavioral Sciences*, 2nd edn. Erlbaum, Hillsdale, NJ

Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand McNally College Pub. Co., Chicago

Dawes R M, Smith T L 1985 Attitude and opinion measurement. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*, 3rd edn. Random House, New York

Judd C M, Kenny D A 1981 *Estimating the Effects of Social Interventions*. Cambridge University Press, New York

Judd C M, McClelland G H 1989 *Data analysis: A Model-Comparison Approach*. Harcourt, Brace, Jovanovich, San Diego, CA

Judd C M, McClelland G H 1998 Measurement. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston

Kenny D A, Kashy D A, Bolger N 1998 Data analysis in social psychology. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, MA

Kline R B 1998 *Principles and Practice of Structural Equation Modeling*. Guilford, New York

Smith E R 2000 Research design. In: Reis H T, Judd C M (eds.) *Handbook of Research Methods in Social and Personality Psychology*. Cambridge University Press, New York

West S G, Biesanz J C, Pitts S C 2000 Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In: Reis H T, Judd C M (eds.) *Handbook of Research Methods in Social and Personality Psychology*. Cambridge University Press, New York

C. M. Judd

Social Psychology, Sociological

Social psychology is concerned with the interplay of society and individual. In this specification, 'society' is a gloss for all relatively stable social relationships and patterns of social interaction, and the close examination of microsociological processes is included in social psychology. The major contributors to this enterprise are psychologists and sociologists.

1. The Two Social Psychologies

In practice, there are two social psychologies. Psychological social psychology assumes that 'in the beginning there is the individual' and focuses on individuals' social cognitions. Sociological social psychology assumes that 'in the beginning there is society'; its distinguishing charge is to locate interactional processes in their social structural context. This difference leads to differences in problems studied, conceptualizations of problems, and research findings. Marcus and Zajonc (1985) observe that cognitive social psychology starts and ends with persons' cognitions; sociologists note that unless the rootedness of cognitions in social structure is recognized, more causal power than warranted is accorded cognitions. Cognitive social psychologists recognize that cognitions reflect experience, but experience does not enter theoretical formulations or research designs; for sociologists how experience is organized is central to both theory and research. The former developed largely through experimental research; the latter—more eclectic in methods and exhibiting greater willingness now than earlier to use experimental methods—has preferred studying social behavior *in situ* using observational procedures or survey and questionnaire methods locating persons in social milieus. The two social psychologies share topical interests: e.g., attitudes, self, and identity. They also differ in topics, as a glance at textbooks for each will evidence.

2. Focusing on Sociological Social Psychology (SSP)

The concern here is with social psychological work reflecting sociological interests and contributing to sociology as an intellectual enterprise. This does not imply that psychological social psychology is 'wrong' or irrelevant to SSP; sociologists can learn much from their psychological counterparts. Yet, SSP has its relatively distinct approaches, interests, and objectives.

Sociologists have offered versions of social psychology as alternatives to a sociology built on social

structural concepts; Herbert Blumer's (1969) advocacy of symbolic interactionism is an important example. Today, sociological social psychologists typically understand their work as a specialization within sociology, focusing on persons as social beings and their interaction that contributes to sociological analyses in general; this view is adopted here. Most sociologists doing social psychology share a set of core ideas, largely taken for granted. They also significantly differ in approaches taken. The variations are attended to first.

3. *Variations in Perspectives within SSP*

Three perspectives are considered—interactionist, group processes, and social structure–personality—encompassing most SSP; discussion of each is selective. These perspectives are not as independent as separate treatment suggests. Interactionists incorporate fundamentals of the social structure–personality frame by recognizing structural constraints on personal development and interaction. The social structure–personality formula recognizes the mediating role of interaction processes and microstructures. And group processes work builds upon the social definitional emphases of the interactional frame and the macrostructural contexts of small group structures and processes. Nevertheless, the classificatory scheme allows the recognition of significant variation in perspectives within SSP.

3.1 *The Interactionist Perspective*

Interactionism developed under the early influence of the Scottish moral philosophers, George Herbert Mead, Charles Horton Cooley, W. I. Thomas. More recent influences include Blumer, Everett Hughes, Erving Goffman, Anselm Strauss, and Ralph Turner. Basic principles are ideas drawing on Mead. Subjectively held meanings are central to explaining or understanding social behavior. Meanings are products of persons' communication as they seek solutions to collective problems, and humans' capacity for symbolization. Of particular importance are meanings attached to self, others with whom persons interact, and situations of interaction; these guide interaction, and are altered in interaction. Implied is that who persons interact with under what circumstances is critical to the development of, and changes in, meanings and thus to interaction processes.

The centrality of meanings underwrites a basic methodological principle of interactionism: actors' definitions must be taken into account in examining their behavior. For many, this principle leads to viewing everyday life as the preferred arena of research; use of observational methods, ethnography, case histories, or intensive interviewing as preferred means of gathering data; and qualitative methods as preferred analysis procedures. Other interactionists, accepting the methodological principle involved, find these preferences unnecessarily limiting in their implied rejection of quantitative analyses of large-scale data sets for research stemming from interactionist premises.

Given these ideas, it is unsurprising that many interactionists attend to self, its dimensions or components, variability or stability, content under varying structural conditions, authenticity, efficacy, self-change, the social construction and interactional consequences of affect, the interactional sources and social behavioral resultants of self-esteem, interactional and symbolic processes in defense of self, the relation of self to physical and mental health, labeling, stigmatization, deviance, and a host of other topics. A related focus is on socialization, for the shaping of self is key to entrance into roles, normal and deviant. Attention is paid to socialization processes within institutional settings—family, school, work, etc., as well as settings that are not fully institutionalized, like child and adolescent play, informal friendship of young and old, or developing social movements, and noninstitutional settings like homelessness.

Responding to the insight that socialization is a lifelong process, attention is accorded life-course related socialization processes. Interaction itself, of both intimates and nonintimates, is the focus of other interactionists, whose studies examine (among other topics) teasing among adolescents, language use both as and in action and interaction; variations in the implicit rules governing interaction and interaction strategies in various settings (Erving Goffman has been especially influential here and with reference to the preceding topic); role relationships within both formal and informal settings, and the impact of status and power inequalities on interaction.

Such long-standing interests of interactionists continue. A comparatively recent turn is an emphasis on emotions. Present in varying degrees in earlier interactionist social psychology, the expanded interest emphasizes emotion as basic to all interaction, expressed for example in David Heise's affect control theory; the role of affect in self-processes; the social construction and interactional consequences of emotion in general and particular emotions.

Other recent developments can only be mentioned. Contemporary sociological interest in culture is seen in work on the generation of culture through interaction. A 'new look' in social movement research makes framing processes and collective identity formation of particular consequence and has reinvigorated interactionist interests in that topic. A vision of self as comprised of multiple group or network based identities has opened up work on the consequences of consonant or conflicting identities that carries theory and analyses beyond traditional role-conflict studies.

3.2 The Group Processes Perspective

Group processes work reflects diverse substantive interests with a common focus on interaction in social groups or social networks. The interaction processes attended to include: cooperation; competition; conflict; conflict resolution; social exchange; inequality; bargaining; power, status, and influence processes; procedural and distributive justice and injustice; the resolution of social dilemmas; the emergence of social structures from interaction; and the reproduction of interaction processes by social structures deriving from them. According to Karen Cook, a social interdependence theme integrates these, as do ideas about how group processes work relate to the general sociological enterprise: group and network interaction processes provide the foundations of macrosociological theory; theories developed on interacting individuals apply to corporate or collective actors; group- and network-based interaction mediates the relation of individual to society. Also, group process students have a common lineage and hold similar beliefs about how to do research and build theory.

That lineage includes Georg Simmel on implications of social forms for interaction; George Homans' exchange theory built upon principles of economics and operant conditioning psychology; Robert F. Bale's research on groups resolving tasks; 'small groups' research; theoretical as well as experimental economics work on bargaining and reward allocation; and game-theoretic inspired research in political science and psychology. Richard Emerson is especially important to this lineage, developing a social exchange theory escaping the limitations of Homans' strict behaviorism and extending exchange theory to deal with exchange networks.

This lineage led to experimentation and mathematical modeling of interaction processes. The focus on experimentation enabled programmatic research, an ideal carried out more in practice by group process researchers than by other SSP researchers. Group process researchers have provided models of such research, building small theories to explain given facts, testing the theories experimentally, discovering flaws in the fit of theory, revising theory to accommodate new data, in an ongoing process of theory building and theory testing.

Further attention is limited to two topics, social exchange and status structures, epitomizing theoretical themes of the group processes perspective and its programmatic research emphasis. Contemporary social exchange theory starts with actors' interdependency: people need others for, and provide others with, things they value and so engage in social exchange. Exchange occurs in situations of mutual dependence; persons act to increase positively valued outcomes and decrease negatively valued outcomes, exchange relations with specific others are repeated, and satiation and marginal utility apply to valued outcomes. Exchange networks—chains of exchange relations—are basic to sociological exchange theory and SSP because they bridge to larger social structures. From here, network theory and research proliferate, extending to corporate actors, the role of power in facilitating or impeding exchanges across networks, the impact on exchanges of negative (competitive) or positive (noncompetitive) network connections, the emergence of norms in exchange networks, etc.

Current work on status structures builds on expectation states theory, developed by Joseph Berger and colleagues. This theory argues that power and prestige orderings in groups arise from members' expected contributions to problem solutions. These expectations shape group interaction confirming the expectations and stabilizing the orderings. The theory further argues that in the absence of information directly linking members' abilities to tasks at hand, group members draw inferences about those abilities from the stratification system of the larger social context, thus reproducing that system. On this foundation, work on expectation states theory has expanded, systematically refining and extending the theory and examining its implications. Of particular interest under contemporary circumstances is work showing that new information about group members' task-relevant abilities can negate the effects on group interaction of stereotypes drawn from the existing stratification order, potentially changing the stratification order of the larger social context. Also of interest is the way ideas underlying this theory parallel ideas fundamental to the interactionist perspective in SSP.

3.3 The Social Structure–Personality Perspective

Described by Melvin Kohn, a major proponent, as the quintessentially sociological approach to social psychology, the social structure–personality perspective examines the reciprocal impact of macrosociological structure and person. It traces to the classical sociological work of Emile Durkheim and others, tying features of persons to types of societies. More proximately, it emerges from cultural anthropological work seeking to understand basic personality structures of society's members through analyses of culture. Influenced by psychoanalysis, this work invoked family and kinship structures as vehicles sustaining and transmitting culture, evidencing the relevance of social structure to persons produced. Social structure replaces culture in a social structure–personality perspective, bringing to the fore the economy, stratification, and work institutions as basic to personality.

An important part of social structure–personality work has been comparative, with nation as context, examining the hypothesis that similar work structures have similar psychological effects. Inkeles and Smith (1974), in a six-nation study, proposed that an

attitudinal-behavioral syndrome, 'individual modernity,' resulted from working in organizations produced by industrialization. This work was criticized on ideological and methodological grounds, and its failure to specify the empirical linkage of structure and person and how mechanisms accounting for that linkage affected subsequent work. The evolving research program initiated by Kohn provides evidence that complex jobs allowing self-direction link societal-level class and stratification structures to intellectual flexibility and self-directedness. Longitudinal analyses suggest the linkages are reciprocal. These propositions hold in comparative contexts, and under conditions of radical social change, arguing for generality of structural effects independent of cultural variation. The operative psychological mechanism underlying these relationships is taken to be learning generalization.

The social structure–personality perspective exists in work on many topics filling the structural 'space' between nation and person, for example, status attainment and mobility processes, gender, work, the life course, and both physical and mental health. Further comment, necessarily brief, on two of these topics provides both insight into the utility of SSP for sociology and why SSP is not in itself an alternative sociology.

Painted large, life course research focuses on the interaction of biography and historical social structures over time in order to understand adult developmental processes. While the complexity of this enterprise defies easy description or summary, it involves a vision of persons living their lives in the context of a continuously changing society, experiencing important historical events at varying ages and points in their lives, and undergoing (or not undergoing) a series of important age-related transitions involving status and role changes; of people's existence within multiple institutional contexts and multiple networks of social relationships, some of which are operative over lengthy portions of their lives; of the life course as implicating multiple developmental trajectories necessitating coordination. A consequence of the contextual complexity and multistrandedness of life is what Elder and O'Rand (1995) term a loose coupling between social transitions and age-gradings. Loose coupling opens up room for choices reflecting meanings and interpretations (including reinterpretations) of events and of prior experience. Persons are selected by societal allocation processes for entry into social positions and roles but also select themselves: they choose to marry, have children, take or leave jobs, divorce, etc., and these choices are consequential. In short, persons are active agents importantly constructing their own lives.

Research from a social structure–personality perspective on status attainment processes also provides a contrary moral. That research indicates that structures of relationships and interaction built on differences in socioeconomic status, race, and gender impact people's skills, motivations, and aspirations for success in schools and work. It also shows that skills, motivations, and aspirations relate to educational and occupational success, i.e., to individual mobility. At the same time, research shows that the structure of schools, jobs, and labor markets can severely restrict the impact of personal characteristics on social mobility and that even voluminous individual mobility need not alter the overall character of a societal-level stratification system. In short, there are limits on the impact of agency on social structures.

4. *Commonalities*

Underlying most work in SSP are ideas reflecting the assumption that 'in the beginning there is society':

(a) Human experience is socially organized. Societies are not invariably coherent; they are, rather, typically complex and shifting mosaics of frequently conflicting (but sometimes not) role relationships, social networks, groups, organizations, institutions, communities, and strata, sometimes related but sometimes isolated from one another. What people experience is importantly a function of just what relationships, networks, groups, communities, and strata they enter, remain in, and leave, how they enter these social structures, and how these social structures relate to one another.

(b) The forms and content of social life are social constructions, the product of the collective activities of persons as they join one another in the course of developing solutions to problems in their daily lives. To make this claim, it is neither necessary nor in accord with accumulated evidence of SSP and sociology itself to argue that processes of social construction are ephemeral or infinitely mutable, or that they occur without constraints imposed by the natural world or prior constructions.

(c) Human beings are active agents who respond selectively to social and nonsocial environments impinging on them, who initiate transactions with, and sometimes alter, those environments. Implied by the idea of social construction, action produced at the initiative of an actor presupposes the symbolic and reflexive capacities of humans, in short, presupposes a 'self,' but does not deny the power of conditioning or of normative demands. In general, all social behavior is some blend of action and reaction.

(d) The world on which humans act and to which they react is a symbolized world specified by meanings attached to the objects comprising it.

(e) There is both constraint and choice in personal and social life. Whether one, the other, or both are operative depends on particulars of the patterns of people's relationships to others in social groups and networks, and on the patterned relationships of those groups and networks to one another.

(f) A final idea within SSP can serve as a summary of the previous five and an appropriate concluding

statement: the basic responsibility of SSP is to deal with ways in which social structures impinge on people's social behavior and the reciprocal impact of that behavior on social structures. While each of the three perspectives within SSP reviewed tends to emphasize different levels of structure—the interactionist perspective stresses proximate structures closest to interacting persons (e.g., families, friendship networks); the implicit structural contexts of interaction for the small group perspective are middle-level structures (e.g., work organizations); and the social structure–personality perspective focuses on large-scale structures (e.g., the nation, stratification systems)—an encouraging development is the increasingly common recognition that structures on all levels are important to the social behaviors studied from each perspective.

See also: Age, Sociology of; Conflict and Conflict Resolution, Social Psychology of; Emotions, Sociology of; Ethnomethodology: General; Exchange: Social; Goffman, Erving (1921–82); Group Processes, Social Psychology of; Groups, Sociology of; Interactionism: Symbolic; Life Course: Sociological Aspects; Mead, George Herbert (1863–1931); Personality and Conceptions of the Self; Personality and Social Behavior; Self: History of the Concept; Small-group Interaction and Gender; Social Psychology; Social Psychology, Theories of; Socialization, Sociology of; Status and Role, Social Psychology of; Symbolic Interaction: Methodology

Bibliography

Berger J, Fisek M H, Norman R Z, Zelditch M Jr 1977 *Status Characteristics in Social Interaction: An Expectation States Approach*. Elsevier, New York

Blumer H 1969 *Symbolic Interactionism: Perspective and Method*. Prentice-Hall, Englewood Cliffs, NJ

Bryson G 1945 *Man and Society: The Scottish Inquiry of the Eighteenth Century*. Princeton University Press, Princeton, NJ

Cook K S 1977 *Social Exchange Theory*. Sage: Newbury Park, NJ

Cook K S, Fine G A, House J S 1995 (eds.) *Sociological Perspectives on Social Psychology*. Allyn and Bacon, Boston

Elder G H Jr, O'Rand A 1995 Adult lives in a changing society. In: Cook K S, Fine G A, House J S (eds.) *Sociological Perspectives on Social Psychology*. Allyn & Bacon, Boston, pp. 452–75

Emerson R M 1962 Power-dependence relations. *American Sociological Review* **27**: 31–41

Gerth H, Mills C W 1953 *Character and Social Structure*. Harcourt Brace, New York

Goffman E 1963 *Behavior in Public Places: Notes on the Social Organization of Gatherings*. Free Press, New York

Homans G C 1961 *Social Behavior: Its Elementary Forms*. Harcourt, New York

Inkeles A, Smith D H 1974 *Becoming Modern: Individual Change in Six Developing Countries*. Harvard University Press, Cambridge, MA

Joas H 1985 *G.H. Mead: A Contemporary Re-examination of his Thought*. MIT Press, Cambridge, MA

Kerckhoff A C 1995 Social stratification and mobility processes: Interaction between individuals and social structures. In: Cook K S, Fine G A, House J S (eds.) *Sociological Perspectives on Social Psychology*. Allyn and Bacon, Boston, pp. 476–96

Kohn M L, Schooler C 1973 Occupational experience and psychological functioning: An assessment of reciprocal effects. *American Sociological Review* **38**: 97–118

Marcus H, Zajonc R B 1985 The cognitive perspective in social psychology. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*. Random House, New York, pp. 137–230

Mead G H 1934 *Mind, Self, and Society*. University of Chicago Press, Chicago

Ridgeway C 1991 The social construction of status value. *Social Forces* **70**: 367–86

Strauss A L 1978 *Negotiations: Varieties, Contexts, Processes, and Social Order*. Jossey-Bass, San Francisco

Stryker S 1980 *Symbolic Interactionism: A Social Structural Version*. Benjamin/Cummings, Menlo Park, CA

S. Stryker

Social Psychology, Theories of

Social psychology studies how individual people's thoughts, feelings, and behavior are influenced by other people (Allport 1954). Social psychology's theories account for how, why, and when people affect other people. Social psychological theories provide networks of propositions, testable by scientific observations, to describe, explain, and predict systematic phenomena in people's reactions to others. This article addresses mid-range theories, not general systems (i.e., Lewin's entire field theory) and not single phenomena (i.e., the reliable finding that first impressions matter). Because social psychology specializes in mid-range theories, the review imposes a larger framework, providing mere pointers to the theories flagged (see the cited books reviewing theory for detail).

1. Organization of this Review

Social psychology's theories each tend to center on one of a few major types of social motivation, describing the social person as propelled by particular kinds of general needs and specific goals. Most reviews acknowledge these motivational roots by reference to broad traditions within general psychology or sociology: role theories, cognitive and gestalt theories, learning and reinforcement theories, and psychoanalytic or self-theories.

As a variant, this review takes a more integrated but compatible premise, based on people's evolution in a social niche: to survive and thrive, people need other people. Several core social motives arguably result

from this perspective. Over the twentieth century, social and personality psychologists frequently have identified the same five or so core social motives, which should enhance social survival (Stevens and Fiske 1995). *Belonging* reflects people's motive to be with other people, especially to participate in groups. *Understanding* constitutes people's motive for shared social accounts of themselves, others, and surroundings. *Controlling* describes people's motive to function effectively, with reliable contingencies between actions and outcomes. *Self-enhancing* comprises people's tendencies to affirm the self. *Trusting* concerns people's motives to see others (at least own-group others) positively. While these motives are not absolute (other reviewers would generate other taxonomies), not invariant (people can survive without them), nor distinct (they overlap), they do arguably facilitate social life, and they serve the present expository purpose. Within each core social motive, distinct levels of analysis address social psychological processes primarily within the individual, between two individuals, and within groups.

2. Theories of Belonging: Role Theories

People's core social motive to participate in group life generates theories that emphasize the place of the individual in the group. Role theories define people's behavior as a function of their *roles*, their position in social structures and the expectations attached to that position, all in the context of group *norms*, the unwritten, informal rules that govern group life.

2.1 Belonging, Within the Individual

Emphatically sociological, role theories describe the effects of group roles and norms on individual self-presentation and self-understanding. In the theater of social life, so aptly described in the *dramaturgical theory* of Goffman, self-presenting people define the scene, take on roles, and follow cultural scripts. Here, *symbolic interaction* with generalized others forms the self (Mead), or others' reflected appraisals construct the *looking glass self* (Cooley).

People's motivation to fit into groups sparks strategic *impression management* (Schlenker) and *self-presentation* (Jones). Wanting to operate successfully in groups provides goals to convey desired (but not necessarily positive) self-images that are personally beneficial and socially believable, for a given audience. For example, people may ingratiate, presenting themselves as likable; self-promote, presenting themselves as competent; or supplicate, presenting themselves as helpless. When people feel under public or private scrutiny, they become *objectively self-aware*, in Wicklund's theory; this awareness motivates matching to internalized social standards. People and situations differ in motivating people to *self-monitor*; that is, to adjust their behavior to the demands of the social group (Snyder).

Belonging motivates not only selves, but also attitudes and values. Merton's *reference group* theory explained how people derive their values from belonging to particular groups or desiring to belong to them. Through face-to-face interaction, reference groups theoretically serve a *normative function*, defining acceptable standards for group membership (Deutsch and Gerard, Kelley). Some attitudes thus serve a *social adjustive* function (Katz, Kelman, Smith, Bruner, and White); that is, attitudes and values facilitate belonging.

2.2 Belonging, Between Individuals

Motives to belong and get along with other people appear in theories describing reciprocity: expectating something in return. Cialdini's *compliance principles* include reciprocity; people help other people who help them (see Gouldner on *reciprocity norms*). *Exchange theories* (see Sect. 4.2) posit that people expect fair outcomes (outcomes proportional to inputs), but later theories proposed different norms, depending on the relationship. C*ommunal-exchange theory* (Clark and Mills) describes exchange relationships (equitable outcomes) versus communal relationships (need-based outcomes). A. Fiske's *social relations model* theorizes four relationship types and norms: equality matching, communal sharing, authority ranking, and market pricing.

People's adherence to relationships over merely self-serving outcomes appears in the *dual-concern model* (Pruitt and Carnevalle): bargaining elicits both own-concern and other-concern. People also evaluate outcomes by *procedural justice*; that is, whether the process itself was fair (Thibaut and Walker, Lind and Tyler). Institutions and authorities especially are evaluated according to procedural justice, thereby eliciting loyalty or anger.

Belonging also appears in leadership theories. Milgram's theoretical analysis of destructive *obedience to authority* includes sociocultural contexts of obedience, subtle and progressive commitment, and denial of personal responsibility—all in the service of the group. More benign forms of authority appear in the *contingency theory of leadership* (Fiedler), describing how leaders influence members most successfully through relationship-oriented, rather than merely task-oriented leadership, though other contingency models (e.g., Vroom) note that effective leadership style depends on context. *Transactional leadership* (Burns) adheres to norms of leader–member reciprocity, whereas *transformational leadership* changes the demands of the followers.

Regardless, individuals gain from belonging. *Social support theories* explain how relationships benefit

health and well-being. Main effects models (e.g., attachment, see Sect. 6) suggest a monotonic relationship between relationships and health. *Buffering theories* (Cohen, House, Sarason) posit that social support undermines the ill effects of stress, if the support fits the stressor.

2.3 Belonging, In Groups

Crowds turn into groups, in Turner's *emergent norm theory*, when people cue each other's behavior. Theoretically, extreme cases result in *deindividuation* (Diener, Zimbardo); the self lost in the group.

Less radically, people join groups because they are attracted to them, according to *group cohesion theory* (Hogg); solidarity transforms aggregates into groups (see Moreno's *sociometry*). Networks of *interaction processes* (Bales) define the group structure along task, social, and activity dimensions. Groups valuing cohesion over task effectiveness suffer Janis's *groupthink*, faulty decision-making that results from conformity pressures. French and Raven's *referent power* and Kelman's *identification* both influence through desire to belong. Group-level influence processes depend on the strength, number, and immediacy of people attempting influence (Latané's *social impact theory*; Tanford and Penrod's *social impact model*). These theories elaborate Zajonc's *social facilitation theory*, which describes the presence of others facilitating an individual's dominant response, benefiting well-practiced skills and undermining less-developed performances.

In organizations, effective performance meets reviewers' standards, maintains working relationships, and satisfies individual needs, in Hackman's *normative model of group effectiveness*. Relatedly, Steiner describes *process losses* that endanger group productivity, when coordination or motivation fails. *Social loafing* diminishes motivation in groups (Latané), but *social compensation* (Williams and Karau) can enhance motivation in groups.

The predominantly positive effects of group belonging all pertain to own group (in-group). Nevertheless, in-group favoritism (rewarding one's own), not out-group deprivation, poisons intergroup relations. Tajfel's *social identity theory* describes how in-group favoritism stems from self-esteem and categorization, the latter expanded in Turner's *self-categorization theory*. These theories hold that categorization into in-group and out-group differs from other kinds of categorization (see Sect. 3) because the perceiver belongs to one of the categories: Mere categorization into groups promotes conflict. In contrast, Sherif's *realistic group conflict theory* blames intergroup hostility on competition over scarce resources. If single in-group belonging promotes intergroup conflict, multiple belonging diminishes it, in the theory of *crossed categorization* (Dovidio and Gaertner).

3. Theories of Understanding: Gestalt and Cognitive Theories

People's core social motive to hold a coherent, socially shared understanding has inspired theories that account for various social cognitive processes. *Gestalt theories* in social psychology emphasize harmonious fit among elements that constitute a coherent whole, hence their influence on *consistency* theories of attitudes and on *schema* theories of social perception (see Sect. 3.1). Their descendants, modern social cognition theories, most often focus on dual processes, one *automatic* and based on superficial, peripheral *heuristics* or *categories*, the other *controlled* and based on systematic, central, deliberation or *individuation* (see Chaiken and Trope 1999).

3.1 Understanding, Within Individuals

Derived from gestalt psychology, three types of theory focus on processes within the social perceiver: attribution, impression formation, and consistency theories. Other attitude theories and self theories build indirectly on these origins, but still emphasize understanding as primary.

Heider's theories of social perception focused on harmonious, coherent wholes: *invariance* in perceived personality. Heider's social perceiver, portrayed as a *naive scientist*, searches for consistencies in behavior, to make coherent *dispositional attributions* (inferring stable, personal causes). Other *attribution theories* developed: Jones's *theory of correspondent inference* describes how perceivers impute dispositions that fit an actor's behavior, attributions increased by a behavior's unique ('noncommon') effects and low social desirability. Kelley's *covariation theory* likewise notes a behavior's distinct (i.e., unique) target and its degree of consensus (i.e., desirability) across actors, but adds the observation of consistency over time and circumstances. Extending fundamental theories, *stage models* (Quattrone, Gilbert, Trope) converged on automatic categorization or identification of behavior, followed by dispositional anchoring or characterization, followed by controlled situational correction or adjustment. These recent theories bring dual process perspectives to attribution theories that emphasized controlled processes.

A second line of person perception theories also originated in gestalt ideas and eventuated in dual-process models. Asch proposed a *holistic* theory of impression formation, in which the parts (most often personality traits) interact and change meaning with context. The alternative, an algebraic model that merely summed the traits' separate evaluations, matured in Anderson's later averaging model of *information integration*. Asch's most immediate heirs were modern *schema theories*, examining impression formation as a function of social categories that cue organized prior knowledge. Following Taylor and

Fiske's *cognitive miser* perspective, social perceivers were viewed as taking various mental shortcuts (below), schemas among them. Eventually, a more balanced perspective emerged, describing perceivers as *motivated tacticians*, who sometimes use shortcuts and sometimes think more carefully, depending on their goals. One such dual-process model, the *continuum model* (Fiske and Neuberg), holds that perceivers begin with automatic categories (e.g., stereotypes), but with motivated attention to additional information, perceivers may individuate instead. Brewer's *dual-process* model posits automatic identification processes and controlled categorization, individuation, and personalization.

Other mental shortcuts emphasize relatively *automatic processes*, as defined by Bargh to include being unintentional, effortless, unconscious, and unstoppable. Illustrative theories address influences by stimuli that are arbitrarily *salient* in the environment (Taylor and Fiske) or *accessible* in mind (Higgins, Bargh). In Kahneman and Tversky's *heuristics*, people estimate probabilities by irrelevant but convenient processes: ease of generating examples (*availability*), ease of generating scenarios (*simulation*), and ease of moving from an initial estimate (*anchoring and adjustment*). *Norm theory* (Kahnman and Miller) posits that people retrospectively estimate probabilities via the simulation heuristic, bringing to mind similar but *counterfactual* scenarios. Although not a theory, a catalog of *inferential errors and biases* follows from people's limited capacity (Nisbett and Ross; see Sect. 4.1).

A third line of theories follows from gestalt approaches. Heider's *balance theory* posits that perceivers prefer similarly evaluated people and things also to belong together. This emphasis fits other *consistency theories*. Most prominently, Festinger's *cognitive dissonance theory* holds that people seek consistency among the cognitions relevant to their attitudes, including their cognitions about their own behavior. Attitudes change more easily than behavior, especially behavior based on counter-attitudinal advocacy, forced compliance, free choice, unjustified effort, and insufficient justification (see updates emphasizing *self-esteem* (Aronson), accepting responsibility for an aversive event (Cooper and Fazio's *new look*), and *self-affirmation* (Steele). Other consistency theories hold that people seek harmony within and between their attitudes.

Some attitude theories focus on understanding dual processes, downplaying consistency motives. The *elaboration likelihood model* (Cacioppo and Petty) describes the object-appraisal (understanding and evaluating) function of attitudes, but depicts two modes, depending on motivation and capacity: The *peripheral mode* processes persuasive communications based on superficial, message-irrelevant cues (such as communicator, context, or format), whereas the *central mode* processes message content, generating cognitive responses pro and con, which predict persuasion. Chaiken's *heuristic–systematic model* similarly proposes a rapid, simple process that contrasts with a more deliberate, in-depth process of attitude change.

Focusing on understanding via deliberate control, far from gestalt perspectives, two *subjective expected utility models* predict attitude–behavior relations. Fishbein and Ajzen's *theory of reasoned action* posits that behavior results from intention, which in turn results from attitudes toward a behavior (evaluating the behavior's consequences, weighted by likelihood) and from subjective norms. Ajzen's updated *theory of planned behavior* adds a third component to predict intentions, namely perceived behavioral control.

Like theories of attitudes and social perception, theories of self-perception emphasize coherence. *Self-schema theory* (Markus) describes few, core dimensions for efficiently organizing self-understanding. Self-concepts may be more or less elaborate, resulting in respectively more stable and moderate or volatile and extreme self-evaluations (Linville's *complexity–extremity theory*).

People learn about themselves partly by looking at others. Festinger's *social comparison theory* describes how people strive to evaluate themselves accurately, by comparing their standing relative to similar others, on matters of ability or opinion. Schachter extended this work in exploring the *fear-affiliation* hypothesis, whereby people under threat affiliate with similar others, perhaps to gauge the appropriateness of their emotional reactions. From this came Schachter's subsequent *arousal-cognition theory* of emotion, positing that unexplained physiological arousal elicits cognitive labels from the social context. Although the outside world may anchor self-understanding, other theories describe how the self anchors understanding of the world outside, as in Sherif and Hovland's *social judgment theory*, explaining how people assimilate nearby attitudes within a latitude of acceptance and contrast far-off attitudes within a latitude of rejection.

3.2 Understanding, Between Individuals

Moving to understanding in actual interactions between people: although generally these theories do not take an explicit dual-process approach, one set of theories focuses on the role of old, familiar, prepackaged information, whereas another set focuses on the role of newly processed information.

One of the fundamental principles of attraction is familiarity. The Zajonc *mere exposure theory* describes people favoring people and objects encountered frequently. Similarly, people are attracted to people in *propinquity* (close proximity) (Festinger, Schachter, and Back), perhaps because of exposure. Although explained in non-cognitive terms, familiarity effects fit the principle of liking those who seem comfortable,

safe, and easily understood. Consistent with this view, Newcomb developed the *similarity principle* of attraction, derived from balance theory (above), to explain the clear attraction to others of shared background, attitudes, etc. Similar others also maximize social influence, in Cialdini's *compliance principles*.

Just as theories describe how familiar others are liked, perhaps because easily understood, other theories emphasize dislike of dissimilar others, as in *theories of stigma* offered separately by Goffman, Jones, and Katz. *Category-based theories* of impression formation (Sect. 3.1) describe how negative outgroup categories stimulate negative understanding. Moreover, perceivers behaviorally confirm their negative categories of out-groupers, who sometimes affirm the stereotype (see Merton's theory of *self-fulfilling prophecies*, Rosenthal's theory of nonverbally communicated *expectancies*, Snyder's theory of *behavioral confirmation*, Hilton and Darley's theory of *expectancy-confirmation processes*). Mere knowledge of the stereotypes about one's group creates *stereotype threat* (Steele), undermining the performance of distracted, frustrated, targets of stereotyping.

Although familiar information tends to reproduce itself, new information has an impact, according to many theories. In noting the *informational influence* of others, Kelley, as well as Deutsch and Gerard, noted that reference groups can serve a *informative or comparative function* defining objective accuracy from the group's perspective. In Darley and Latané's *cognitive model of helping*, people consult social cues at each step of the process of noticing, interpreting, taking responsibility, deciding, and implementing help. Weiner analyzes *attributions of responsibility* as stemming from information about degrees of control and its internal or external locus. People acquire new information through *social learning* (Sect. 4.2) that encourages them to help (Berkowitz, Rushton, Staub) or to aggress (Bandura). As people learn by doing, they are socially influenced by self-*consistency and commitment*, according to another of Cialdini's compliance principles.

3.3 Understanding, In Groups

In groups, the familiar prior knowledge about own group and other groups is captured by *social identity theory* and *self-categorization theory* (Sect. 2.3). In addition, group-level schemas and categories at the societal level take the form of Moscovici's *social representations*; shared social understandings make the unfamiliar into the familiar by reference to old socially shared concepts.

People in groups also acquire new information. Some individuals hold power by virtue of *expertise or information* (French and Raven), evoking *internalization* (privately held beliefs; Kelman) in those they influence. More generally, *persuasive arguments theory* (Burnstein) describes how groups polarize shared opinions, compared to individuals, when they receive novel supporting information. Moreover, *minority influence* (Moscovici) within groups operates partly through minorities' conviction provoking majorities to systematically process their arguments.

4. Theories of Controlling: Reinforcement Theories

The core social motive to perceive some contingency between one's actions and one's outcomes is a sense of *effectance*, White's term, or control, in other formulations, drives a set of social psychology theories concerned with people's desire to effect certain outcomes. These theories often predict people's behavior from the economics of *reinforcements* (rewards and costs). People aware of situational contingencies act to maximize or at least improve their outcomes, in this view.

4.1 Controlling, Within Individuals

Attribution theories that focused on rational understanding by the *naive scientist* (Sect. 3.1) gave way to theories and mini-theories documenting the *shortcut errors and biases* of the *cognitive miser* (Sect. 3.1). Several of these errors share a focus on belief in human agency. The *fundamental attribution error* (Ross), *correspondence bias* (Jones), and the observer part of the *actor–observer effect* (Jones and Nisbett) all suggest that people over-rely on personality or *dispositional explanations* for other people's behavior. People making attributions about another's behavior underestimate the role of social situations, a phenomenon first noted by Heider. Theoretically, people explain their own behavior situationally, the actor part of the actor–observer effect.

Nevertheless, other theories emphasize people's need to believe in their own agency. For example, Brehm's *reactance theory* describes people's resistence to limits on their behavioral freedom and attraction to threatened or constrained behavioral choices. Bandura's *social cognitive theory* and *self-efficacy motivation* emphasize individual human agency. Bem theorized that people are as balanced as outside observers; his radical behaviorist *self-perception theory* suggests that people observe the contingencies controlling their own behavior and infer their attitudes from the pattern of freely chosen behavior. Indeed, people's *intrinsic motivation* (Deci and Ryan, Lepper, Greene, and Nisbett) depends on free choice, autonomy, and interest; it declines when under extrinsic reward (*over-justification effect*); that is, excessive external control. When people lack a sense of control,

theories hold that they may exhibit *learned helplessness* (Seligman), leading to depression, and *mindlessness* (Langer), leading to unintelligent behavior, but they may also increase *control motivation* (Pittman).

4.2 Controlling, Between Individuals

Various social exchange theories concern outcome-oriented standards for enacting and judging relationships. Homans's *distributive justice* applied elementary principles of operant (Skinnerian) behaviorism to social interdependence, holding that individuals expect rewards proportionate to costs. From this came *equity theory* (Adams, Walster (later Hatfield), Walster, and Berscheid) that argued that people seek fair ratios of outcome to investment. Although couched in reward–cost terms, inequity theoretically relates to dissonance (Sect. 3.1), creating a drive to reduce it. The exchange idea in relationships developed into *interdependence theory* (Kelley and Thibaut), that posits that human interactions follow from degrees, symmetries, bases, and kinds of dependence. In Levinger's *stage theory of relationships*, people begin with a cost–benefit analysis. Whether close relationships switch (Sect. 2.2) or not (Sect. 4.2) to a nonexchange (i.e., communal) orientation, people in relationships do control their own and the other's outcomes, addressed by theories of intent attribution (Sect. 3.1), *emotion in relationships* (Berscheid), and *accommodation* to disruption (Rusbult). Even outside close relationships, outcome dependence motivates *individuation* (Fiske), undercutting stereotypes. Control over one's outcomes appears in a *cost–reward model* (Dovidio, Piliavin) of helping.

Although motives to control one's outcomes theoretically underlie most positive relationships, they theoretically also underlie aggression. Excessive control over another person, to obtain desired outcomes, motivates *instrumental aggression* (Geen), following the *frustration–aggression hypothesis* (Dollard, Miller, Doob, Mowrer, and Sears). Other formulations carry an explicitly cognitive *neoassociationist analysis* (Berkowitz) of aggressive patterns. Lack of control, brought on by *environmental stressors* (Anderson) enhances aggression, as does *social learning* (Bandura), whereby people are socialized to imitate successful aggressive acts.

Incentives motivated the earliest frameworks for *persuasive communication* (Hovland, Janis, and Kelley), as well as more recent principles of compliance involving *scarcity* (Cialdini) as a threat to control over outcomes.

4.3 Controlling, In Groups

Group members theoretically analyze costs and benefits at role transitions during *group socialization* (Moreland and Levine). *Interdependence* of outcomes defines group life (Deutsch, Kelley and Thibaut, Sherif). Some people control outcomes more than others, placing them in positions of power, defined as the capacity to influence others (French and Raven); although only some *bases of power* (reward, coercive) explicitly relate to reinforcements, even the others (Sects. 2.3 and 3.2) could be viewed as incentive-based. Pure incentive orientation, however, induces only *compliance* (public performance, not private belief) (Kelman).

Belief in the power of human agency to control outcomes underlies several theories of group-level world views. Lerner's *just world theory* posits individual differences in the belief that people and groups get what they deserve. Sidanius and Pratto's *social dominance theory* posits individual differences in the beliefs that group hierarchy is inevitable; individuals high in social dominance orientation seek hierarchy-maintaining careers and exhibit high levels of intergroup prejudice; societies high in social hierarchy exhibit highly controlling social practices, including violent oppression, toward subordinate groups. *Evolutionary social psychology*, coming out of *sociobiology* (Trivers, Wilson), holds that individuals are motivated to dominate other individuals, to enhance their own reproductive outcomes (Buss, Kenrick).

5. Enhancing Self: Psychoanalytic and Self-theories

People seem motivated to have a special affection for self, and in the West this often takes the form of self-enhancement, whereas in East Asia, it may take the form of self-improvement, but with a certain sympathy for the self's difficulties. Theories of the special role of the self, especially defense of self-esteem, influenced social psychology first through *psychoanalytic perspectives*, especially from Freud, of course, but also Erikson, Maslow, and Sullivan. Subsequently, the importance of *self-esteem* became a non-Freudian truism driving a range of theories at individual, interpersonal, and groups levels.

5.1 Self-enhancing, Within Individuals

The clearest influence of psychoanalytic theory, the *authoritarian personality theory* (Adorno, Frenkel-Brunswik, Levinson, and Sanford) proposed that rigid, punitive, status-conscious child-rearing practices reinforce unquestioning, duty-bound, hierarchical-oriented obedience. To maintain safety, the child idealizes the parents, leaving no outlet for the unacceptable primitive impulses of sex and aggression. According to the theory, these are displaced onto those of lower status, notably outgroups, leading to lifelong prejudice. Altemeyer's recent *right-wing authoritarianism* has revived interest in prejudice as a self-enhancing motivational construct, absent the Freudian backdrop.

More broadly, several Western theories pick up the theme that people think, interpret, and judge in ways that elevate the self. Taylor and Brown suggest that *positive illusions* about the self, possibly inaccurate, benefit mental and physical health. Kunda describes *motivated reasoning* that protects self-esteem. Even people with stigmas buffer self-esteem by taking advantage of *attributional ambiguity* (Crocker and Major), attributing negative outcomes to prejudice instead of personal failings.

More generally, Steele's *self-affirmation theory* posits that people need to feel worthy, so a threat to one aspect of self can be counteracted by affirming another self-aspect. *Terror management theory* (Greenberg, Pyszczynski, and Solomon) holds that people specifically feel threatened by their own mortality, so to allay their anxiety, they subscribe to meaningful world-views that allow them to feel enduring self-worth. Similarly, the *ego-defensive function* (Katz, Smith, Bruner, and White) of attitudes protects the self from threat, whereas the *value-expressive function* of attitudes (Katz) more broadly affirms the self-concept.

The *self-discrepancy theory* of Higgins argues self regulation aims to attain positive self-feelings and avoid negative self-feelings. Discrepancies between the *actual self* and the desired *ideal self* cause depression, at the lack of the positive outcome. Discrepancies between the actual self and the *ought self* cause anxiety, at the failure of not living up to standards.

5.2 Self-enhancing, Between Individuals

Various theories argue that people use relationships to enhance the self. Aron and Aron's *theory of self-in-relationship* posits that relationship and self-concept merge, presumably affecting self-esteem according to relationship quality. Not all positive features of a close other boost self-esteem. According to Tesser's *self-evaluation maintenance theory*, people feel threatened by the success of close others in areas of self-identified competence, but threat diminishes with decreased closeness or identification.

Even interpersonal altruism has been hypothesized to result from egoistic motives, such as *negative state relief* (Cialdini) that puts an end to the perceiver's own guilt or sadness at another's distress. Certainly aggression has been hypothesized to result from self-protective motives. Indeed, Baumeister argues that *inflated self-esteem* underlies much aggression and violence. More broadly, *affective aggression* (vs. instrumental aggression, Sect. 4.3) theoretically results from perceived threats to self, as in *attributions of hostile intent* (Dodge).

5.3 Self-enhancing, In Groups

According to sociometer theory (Baumeister and Leary), people base their self-esteem on their standing in the group; from a broad perspective, this fits the belonging motive (Sect. 2), but the theory also specifically addresses the process of self-enhancing in the context of a group. People's sense of positive identity theoretically results from group membership (Sect. 2.1). Nevertheless, people balance group identity and personal identity, according to Brewer's optimal distinctiveness theory, to achieve a positive view of self. To the extent that people incorporate the self-in-the-group (Smith), people will experience emotions on behalf of the group, presumably including group-based self-esteem. Traditional theories of relative deprivation (Stouffer, Runciman, Crosby) distinguish egoistic deprivation, which assumes that people base their feelings about their own esteem on their standing compared to others, versus fraternal deprivation, which provides the parallel for one's group standing compared to other groups. In each case, a form of self-esteem depends on gaps within or from a group.

The group as a basis for self-esteem theoretically differs in more *individualistic vs. collectivistic cultures* (Triandis) or more *independent or interdependent cultures* (Markus and Kitayama). In relatively individualistic, independent cultures, the self is autonomous from the group, and self-esteem is personal; most theories of self-enhancement result from this North American and European cultural perspective. The autonomous self moves in and out of many groups, with multiple group identities all affecting self-esteem. In relatively collectivistic, interdependent cultures, the self is subordinate to the group, and strives to enhance the group by improving the less important self. The interdependent self remains loyal to few primary groups, which sustain the self.

6. Trusting

Just as people have a special positive feeling for the self, they also have a positive feeling for others, at least in the in-group, all else being equal. Bowlby's *attachment theory* argues for the survival value of children having secure relationships, a principle social psychologists extend to adult relationships.

6.1 Trusting, Within Individuals

Matlin and Stang's *Pollyanna hypothesis* holds that people perceive and expect positivity across a variety of domains. People especially perceive individual people more positively than comparable groups or entities, in the *person-positivity hypothesis* (Sears). Given this positive baseline, negative information theoretically differs in *salience* (Fiske), *diagnosticity* (Skowronski and Carlston), and *motivated vigilance* (Peeters and Czapinski). People are risk-averse, according to *prospect theory* (Kahneman and Tversky). Taylor's *mobilization-minimization hypothesis* describes how people react strongly to negative events, but quickly readjust to a moderately positive baseline.

Gilbert and Wilson's *durability bias* describes people's ignorance of their own readjustment to baseline after unusually negative (or positive) events.

6.2 Trusting, Between Individuals

All relationships, but especially close ones, depend on trust, according to *appraisal process theory* (Holmes). A*ttachment theory*, as applied to close relationships (Hazan and Shaver), describes how people vary in trust for others and value for self, resulting in a typology of *relationship orientations*. Trust benefits relationship quality and longevity. Trust, in an *exchange* form (Sect. 4.2) constitutes *reciprocity* (Sect. 2.2); in a *communal* form (Sect. 2.2), trust constitutes response to another's needs. When people vicariously experience each other's emotions, a strong form of communal sharing, they are more likely to help others, in accord with the *empathy–altruism hypothesis* (Batson, Eisenberg).

In *game theory* (Luce and Raiffa), trust entails cooperative instead of competitive behavior in mixed motive games, in which outcomes conflict but need not sum to zero. Structural features determine cooperation, according to Deutsch's *theory of trust and suspicion*; when people can communicate intent, they trust more. Larger groups enable less trust, according to Yamagishi and Sato's *theory of trust*.

6.3 Trusting, In Groups

Shared goals promote trust both within and between groups, in an *interdependence theory* (Sect. 4.3) analysis. But shared goals also contribute to the *identity* of the group (Sect. 2.3). Another theoretical account is that *superordinate goals* require collaborative effort (Sherif and Sherif) that creates trust and overcomes intergroup divisions. Under optimal conditions, Allport's *contact hypothesis* specifies that equal status, cooperative, pleasant, sanctioned intergroup contact reduces prejudice and increases trust between groups. Members of distinct groups may come to view themselves as *one-group* (Gaertner and Dovidio).

7. Conclusion and Disclaimer

Theories flourish in social psychology, usually accompanied by rigorous programs of research. This review has not evaluated the empirical success of the noted theories, merely recorded some prominent ones. At this count, the 32-volume American *Advances in Experimental Social Psychology*, the 10-volume *European Review of Social Psychology*, the 4-volume *Personality and Social Psychology Review*, as well as social psychological theories appearing in *Psychological Bulletin* and *Psychological Review* (by computerized search) amount to just short of 600 separate theories, not to mention the ones appearing only in 'empirical' journals or books. The present article comprises a series of pointers to theories that reflect both past and current thinking, in a social evolutionary framework of five (plus or minus five) core social motives that facilitate social life.

See also: Attitude Change: Psychological; Attitude Formation: Function and Structure; Attitudes and Behavior; Attributional Processes: Psychological; Cognitive Dissonance; Conflict and Conflict Resolution, Social Psychology of; Group Decision Making, Social Psychology of; Heuristics in Social Cognition; Impression Management, Psychology of; Intergroup Relations, Social Psychology of; Justice: Social Psychological Perspectives; Leadership, Psychology of; Love and Intimacy, Psychology of; Mental Representation of Persons, Psychology of; Motivation and Actions, Psychology of; Obedience: Social Psychological Perspectives; Person Perception, Accuracy of; Reactance, Psychology of; Schemas, Social Psychology of; Self-evaluative Process, Psychology of; Small-group Interaction and Gender; Social Categorization, Psychology of; Social Cognition and Affect, Psychology of; Social Comparison, Psychology of; Social Dilemmas, Psychology of; Social Identity, Psychology of; Social Influence, Psychology of; Social Psychology; Social Relationships in Adulthood; Stereotypes, Social Psychology of

Bibliography

Allport G W 1954 The historical background of modern social psychology. In: Lindzey G (ed.) *The Handbook of Social Psychology*, 1st edn. Addison-Wesley, Reading, MA, Vol. 1

Berkowitz L (eds.) 1964/1989 *Advances in Experimental Social Psychology*, Vols.1–22. Academic Press, New York

Chaiken S, Trope Y (eds.) 1999 *Dual-process Theories in Social Psychology*. Guilford, New York

Deutsch M, Krauss R M 1965 *Theories in Social Psychology*. Basic Books, New York

Gilbert D T, Fiske S T, Lindzey G (eds.) 1998 *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York

Higgins E T, Kruglanski A W (eds.) 1996 *Social Psychology: Handbook of Basic Principles*. Guilford, New York

Jones E E 1985 Major developments in social psychology during the past five decades. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*, 3rd edn. Vol 1. Random House, New York

Jones E E, Gerard H B 1967 *Foundations of Social Psychology*. Wiley, New York

Manstead A S R, Hewstone M (eds.) 1995 *The Blackwell Encyclopedia of Social Psychology*. Blackwell, Oxford, UK

1997/2000 *Personality and Social Psychology Review* 1–4

Ross L R, Nisbett R E 1991 *The Person and the Situation: Perspectives of Social Psychology*. McGraw-Hill, New York
Shaw M E, Costanzo P R 1982 *Theories of Social Psychology*, 2nd edn. McGraw-Hill, New York
Stevens L E, Fiske S T 1995 Motivation and cognition in social life: A social survival perspective. *Social Cognition* **13**: 189–214
Stroebe W, Hewstone M (eds.) 1990/1998 *European Review of Social Psychology*, Vols. 1–8. Wiley, Chichester, UK
Taylor S E 1998 The social being in social psychology. In: Gilbert D, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. Vol. 1. McGraw-Hill, New York
West S G, Wicklund R A 1980 *A Primer of Social Psychological Theories*. Brooks/Cole, Monterey, CA
Zanna M P (eds.) 1990/2000 *Advances in Experimental Social Psychology*, Vols. 23–32. Academic Press, New York

S. T. Fiske

Social Question: Impact on Social Thought

A modern social science not only originates within an intellectual project of widening rational knowledge, but also as a means to govern what Simmel (1908) called 'the material of social processes.' The making of what Comte (1848) described as 'the preponderance of a social point of view' implied that there were problems that no established scientific approach could successfully address. Even more, the individualistic premises of modern legal and economic sciences themselves opened new problems from the vantagepoint of social organization. If the major trend of modernity has been toward the legal emancipation of the individual and the economic emancipation of labor from the burdens of traditional society, the emergence of social thought was to make evident that this could not suffice to organize a stable modern society. There was a need for models of organization and social integration, i.e., for a 'science of social order.' For such a science, the first object of analysis has been the 'social question', that is, the mass pauperism growing alongside the rise of modern liberalism that social thought regarded as a danger for social stability. This became particularly clear in France where revolutions brought the problem of popular poverty to the forefront of the political arena.

1. *The Experience of the Revolution*

The idea that the French Revolution played an important role in the rise of sociology has been established by important sociologists such as Reinhardt Bendix and Robert Nisbet. The very term 'social science' became official during the French Revolution through authors like Condorcet (Baker 1964, McDonald 1993). Less common is acknowledging that its development was linked to the question of poverty and the new political importance it assumed. The Revolution was more than an intellectual advance in rationalism; it was a dramatic experience cutting bridges with most of what was known and abruptly revealing the depth of the social crisis. The new awareness of the crisis, more than the rationalist critique, was crucial to social thought.

Nisbet (1943), stressing the impact of the French Revolution over the making of sociological concepts, suggested that the experience of the crisis focused them on the 'social group', as their own response to eighteenth century social theory centered on the opposition between the state and the individual. Social thought developed by establishing the group as an object of science, not only with an ethical implication, but also with a primary theoretical value in the study of man.

Why the group? Because the impact of the Revolution upon traditional social groups had driven the crisis until the destruction of the whole institutional frame, eroding the very bases of social belonging. The 'rational state,' as Nisbet put it, that the eighteenth century social theory had been regarded as the solution to the crisis of traditional bonds (religion, family, guilds, and associations) and the correlation to individuals' emancipation, no longer seemed able to nourish a real sense of social belonging. The group raised a new problem that could not be faced in economic or in juridical terms: to rebuild society required new foundations for the social bond. The experience of the crisis and the condition of insecurity due to the destabilization of order opened the way to a science that would fill the task of laying down the proper conditions for a new social order.

2. *The Autonomy of 'Society'*

The 'social question' was the most dramatic evidence for both the crisis and the need for a new social order. Poverty could no longer be regarded as a mere individual condition to be dealt with by charitable support and police repression, as in the Old Regime treatment of mendicancy and vagrancy. The political emancipation gained through the Revolution, as well as the economic emancipation guaranteed by a free labor, made the increase of poverty among the urban popular masses an issue of public concern (Simmel 1908). How could the poor play a role in the new social order, be inscribed in the political arena of universal citizenship and in the liberalized market of productive labor? As Hannah Arendt (1963) stressed, the social question broke the social unanimity around the revolutionary project—and from then onwards, it has been the stumbling block of any revolution.

The modern idea of rights, making the poor just citizens the same as everyone else, and the modern economic imperative to include them into productive labor, did converge in giving a new visibility to the social question. In fact, poverty challenged both the achievement of individual autonomy as well as the generalization of free labor. It was therefore important to act against the social causes of poverty, and therefore to know them, what neither legal nor economic sciences could do. The importance of the poverty question grew together with the political awareness that a stable society demanded an end to the revolution. Both social and political problems contributed to a realization that social change could not be understood through individual actions, but rather as the result of relatively impersonal processes. Society appeared to have autonomous conditions to its own development and an autonomous ability to resist changing, both impossible to curb by law or through the market.

The social question has been the first object of application where the autonomy of social thought's premises, concepts, and methods was at stake. It was thus the arena for a gradual autonomization of social thought from economic and legal categories, founding the problem of order as specifically social (Procacci 1993). In this way, it represented a crucial step toward the construction of the very idea of society as a theoretical concept and at the same time a practical field of investigation.

3. *The Science of Social Order*

The analysis of the social question had a crucial impact over the making of positive sociology. With Saint-Simon and Comte, the search for a new social order was firmly located within antisocial effects of industrial transformation: how was it possible to organize inequality in a way that would not prevent social stability? Their foundation concepts were marked by their source in the analysis of social inequality.

First of all, they carried a vision of society as an organic entity, essentially concerned with the cohesion of all its parts—hence the importance of the metaphor of social organism. Most importantly, this located sociology on the side of the whole social body, the common interest, the collective being—apart from the opposition opened by liberalism between individual and collective points of view.

Secondly, all dangers for social order were interpreted in the perspective of social pathology, calling for some kind of therapeutics. Poverty was a pathological element within the social body—the task of social thought was to identify the means for restoring a normal society. Hence the importance of medical knowledge, with the normative character intrinsic in the idea of restoring a normal society. Methods of inquiry were also drawn out of medical practice: observation of symptoms, collection of cases, and statistical recurrences. This shaped concepts animated fundamentally by a will to reform. There was no longer a pure science; the programme of positive sociology was a science able to found the criteria for acting in order to improve human conditions on positive knowledge.

Thirdly, the foundation concepts of positive sociology inherited the moral element that had characterized the philanthropic approach to poverty. Particularly with Comte, the problem of order became a question of moral consistence of society, impossible to solve through legal or economic reforms. His main theoretical questions run about the position of bourgeoisie and the working class in the new social order (Elias 1978). For this reason, once the revolution exploded again in 1848 and the popular movement claimed political and economic reforms to solve the social question, he felt the need to engage into the public debate. His 'Positivist Society', created a few days after the February insurrection, had the explicit task to oppose the positivist view of the social question to socialists. Insufficient labor did not depend upon capitalist economy, but upon an irrational organization that scientific development would overcome. The emphasis was rather on society's moral engagement to fight against poverty, a problem more difficult to solve: it required to put 'social sentiment' in the forefront, by promoting a sociability that the Comtian concept of 'duty' was to organize as the cement for new social bonds.

4. *Democracy and Socialism*

A different perspective is the one opened by Tocqueville's analysis of democracy. Elected in 1848 to the parliament by the first universal manhood suffrage, he was directly involved in writing the new Constitution which was to put an end to the revolution. As such, he politically intervened against the reforms claimed by the popular movement. Political organization of democracy above all demanded fighting its tendency to ally with a centralized power. The social question implied such a danger, strengthening the state that was expected to solve it; therefore nothing could be granted, except an exclusively moral engagement of society to reduce poverty. Distinguishing moral duties (*la morale*) from legal rights (*le droit*) was strategic in preventing the state from becoming either an economic agent providing jobs or a regulator reviving old protectionism. A political solution could only come from organising democracy: pushing people to act on collective issues and promoting public virtue instead of private freedom, were the only means for undermining the spectre of state despotism. The perspective of self-government led to rehabilitate associations as the arena where people would exert their public responsibility and resist the state.

The 1848 revolution was a crucial moment for social thought consolidation through its analysis of the social question. The work of Lorenz von Stein (1850) provided important evidence and was a crucial source for Marx. Indeed it would be difficult not to mention the influence of social question over socialist social thought, as the revolution of 1848 exemplified. The experience of French revolutions was also central for Marx's historical concept of class, conceptualizing the fundamental inequality producing poverty in a capitalist economy. But class conflict was for Marx (Marx and Engels 1848) the key to development, and not a pathology to be cured; there could not be any other solution to it but the revolution. The strategical programme of Marxian social thought deeply differed from the one of sociology: it was not a question of reanimating society through moral duties and intermediary associations. Under this light, there is more continuity with Durkheim's theory of social solidarity.

In his study on socialism, Durkheim (1895) stressed its affinity with sociology, given the central role played by organization and a set of social rules controlling the economic sphere. Since societies autonomously tend toward solidarity, the antagonism between capital and labor was a danger, and could only result from disorganization. The social question was a problem of socialization; even unions might work as a means of social integration. The concept of organic solidarity combined the characteristics of positive sociology and the attempt to enhance democratic strategies: it provided criteria for techniques reorganizing social bonds, such as the division of labor and the resurgence of intermediary groups.

In conclusion, the analysis of poverty contributed to strengthen a critique of economic and legal individualism towards a new science of social solidarity. The impact of the social question tied the fate of social science to the development of an administrative state, and engaged it into the political arena; thus promoting the re-establishment within the social body of those groups that liberalism had strategically eliminated.

See also: Civil Rights; Democracy; Garbage Can Model of Behavior; Inequality; Institutions; Liberalism; Modernization, Sociological Theories of; Poverty, Sociology of; Revolutions, Sociology of; Socialism

Bibliography

Arendt H 1963 *On Revolution*. Viking Press, New York

Baker K 1964 The early history of the term 'social science'. *Annals of Science* **20**: 211–25

Comte A 1848/1957 The social aspect of positivism. In: *A General View of Positivism* [trans. by Bridges J H]. R. Speller, New York, Chap. 2

Durkheim E 1895 *Le Socialisme*, 1971. PUF, Paris

Elias N 1978 *What is Sociology?* [trans by Mennell S, Morrissey G]. Hutchinson, London

Foucault M 1972 *L'Histoire de la Folie à l'Age Classique, Suivi de mon corps, ce papier, ce feu et la folie, l'absence d'oevre*, 2nd edn. Gallimard, Paris

Marx K, Engels F 1848/1964 *The Communist Manifesto*. Washington Square Press, New York

McDonald L 1993 *The Early Origins of the Social Sciences*. McGill-Queen's University Press, Montreal

Nisbet R 1943 The French Revolution and the rise of sociology in France. *American Journal of Sociology* **49**: 156–64

Procacci G 1993 *Gouverner la Misère. La Question Sociale en France*. Seuil, Paris.

Simmel G 1908 *Sociology, Untersuchungen über die Formen der Vergesellschaftung*. Leipzig, Germany

von Stein L 1850/1921 *Geschichte der Soziale Bewegung in Frankreich* Salomon G (ed.) Munich, Germany

Tocqueville A 1851/1978 *Souvenirs*. Gallimard, Paris, [*Recollections*.]

G. Procacci

Social Relationships in Adulthood

1. Basics About Relationships

1.1 What is a Relationship?

What constitutes a relationship? The answer, many relationship researchers will say, is simple. A relationship exists between two people when each person influences the other's thoughts, feelings, and/or behavior. In other words, a relationship exists when people are at least minimally interdependent (Kelley et al. 1983). As the frequency, diversity, impact, and/or time span over which people influence one another increase, the relationship becomes increasingly interdependent or, according to Berscheid et al. (1989), closer.

1.2 Types of Relationships

Interdependence takes many forms, a fact reflected in the diversity of terms we use to refer to relationships—friends, romantic partners, sibling relationships, business partners, and physician–patient relationships. Most past, and many current, relationship researchers have chosen one of these types of relationships to describe in detail. This has resulted in separate literatures on friendships, romantic relationships, marriages, and sibling relationships, and these literatures continue to expand.

Increasingly, however, researchers are focusing upon conceptual differences between relationships.

Some have identified conceptual dimensions along which relationships vary. For instance, Wish et al. (1976) find evidence for four dimensions underlying people's ratings of relationships: cooperative/friendly versus competitive/hostile; equal versus unequal; intense versus superficial; and socioemotional/informal versus task-oriented/formal. Blumstein and Kollock (1988) suggest that relationships vary according to whether they do or do not involve kin, sexual/romantic aspects, cohabitation, differences in status, and cross-sex participants.

Still others have proposed taxonomies including sets of qualitatively distinct categories. For instance, Fiske (1992) posits that four basic rules direct individuals' relations with others in all social interactions: people may share a common identity and treat one another equally (communal sharing); people may attend to status differences and give benefits accordingly (authority ranking); people may weight the utility of their behavior in achieving desired outcomes for themselves (market pricing); and finally, people may follow a quid-pro-quo principle (equality matching). Clark and Mills (1979, 1993) distinguish between: communal relationships, wherein benefits are given, non-contingently, on the basis of the recipient's needs; exchange relationships, wherein benefits are given with the expectation of receiving comparable benefits in return; and relationships in which people are simply concerned about their own self-interest. They also emphasize a quantitative dimension relevant to communal relationships—the degree of responsibility that one person feels toward another. The stronger the communal relationship, the greater the costs (in time, effort, money, and foregone opportunities) a person will incur to respond to the other's needs. Most recently, Bugental (2000) has proposed grouping relationships within an attachment domain, a hierarchical power domain, a coalitional group domain, or a reciprocity domain. In addition to thinking more about the conceptual dimensions and qualities characterizing different relationships, many researchers now emphasize understanding the interpersonal processes that influence relationships.

1.3 On What Types of Relationships is Most Psychological Research Currently Focused?

Despite recognition that a variety of conceptually distinct relationships exist, research on processes occurring within relationships is not evenly distributed among these types. Most research has been conducted on intimate, or potentially intimate, relationships (e.g., friendships, romantic partnerships, and marriages). These relationships tend to fall in attachment or reciprocity domains (Bugental 2000). Further, people expect intimate relationships to be cooperative, friendly, equal, intense, social-emotional (Wish et al. 1976), and communal in nature (Clark and Mills 1979, 1993). Increasingly these relationships are being studied longitudinally. A much smaller corpus of research exists on interpersonal processes relevant to competitive (Wish et al. 1976), market pricing (Fiske 1992), authority ranking (Fiske 1992), or hierarchical (Bugental 2000) relationships, and little of it is longitudinal in nature. Hence, just a few comments are devoted to research on competitive and hierarchical relationships here. More extensive comments are included relevant to interpersonal processes within friendships, romantic relationships, and family relationships.

2. Research on Non-intimate Relationships

2.1 Hierarchical Relationships

Many social relationships exist in which one member has power or authority over the other, and psychologists have reported some interesting findings relevant to such relationships. Fiske (1993), for instance, found that one's position in a hierarchical relationship influences one's tendency to view one's partner in stereotypical terms. Those with less power tend to pay careful attention to information about powerful people, especially information that contradicts stereotypes of these individuals. In contrast, in thinking about their underlings, those in more powerful positions rely on stereotypes. Presumably underlings have more time to attend to details about those who have power over them and are motivated to do so by a desire to gain control over the situation. Those in power who oversee many people have less time and, given the control their power affords them, less need to attend to details regarding underlings.

Other researchers have found disturbing links between some men's ideas of power and their tendencies to view women as attractive. Bargh et al. (1995) used a questionnaire to measure males' propensity to sexual aggression. Later, these men participated in a study with a woman. Some men were primed to think about power (by sitting behind a large desk), while others were not. Among those prone to sexual aggression, being primed to think about power was associated with seeing the woman as being significantly more attractive—a link about which these men were not consciously aware.

2.2 Relationships Between Negotiators

A great deal of work exists on interpersonal processes occurring between people who must negotiate with one another to come to a solution to a problem, most often involving a division of resources. Much early work focused on how individual differences in personality and gender influence negotiating, yet it yielded little in the way of findings. Other early work on how situational factors influence negotiations was more

fruitful. For example, the presence of observers has been shown to lead negotiators to express greater advocacy for previously announced positions, whether or not the observers have a vested interest in the outcome. Later work, falling within a 'behavioral decision perspective,' has explored how negotiators deviate from rationality in their decision-making. Findings suggest that negotiators are too influenced by whether risks are framed positively or negatively, rely too heavily on the most available information, and are too confident about the likelihood of attaining outcomes that favor themselves. The nature of preexisting relationships between negotiators has also been shown to influence the process of negotiation. For instance, negotiators prefer equal payoffs when the relationship is positive or neutral. In contrast, when the relationship is negative, they prefer inequality—favoring themselves. Other findings show that when relationships are close there is more information sharing, less coercive behavior, less demanding initial offers, and quicker agreements. Surprisingly, better joint outcomes are not superior overall, however. (See Bazerman et al. 2000 for a summary of this and other work on relationships between negotiators.)

3. *Research on Intimate Relationships*

3.1 *A Very Active Area With Some Well-established Findings*

Psychological research on intimate relationships is flourishing. Twenty-five years ago, the relevant literature was small and limited almost entirely to studies of initial attraction between people. The research was primarily experimental and conducted within the confines of a laboratory. When ongoing relationships were examined, the dependent measure of interest was, most often, global satisfaction. Today the relationship literature focuses largely on naturally occurring, ongoing relationships and on specific interpersonal processes relevant to forming, building, repairing, maintaining, and (sometimes) ending relationships. Although some research remains experimental and laboratory-based, field research often involving longitudinal surveys and diary studies of ongoing relationships has taken center stage. The field has advanced sufficiently that there is clear consensus on some facts about close relationships. These topics are discussed next.

3.2 *Having Relationships is Good for You*

One fact that highlights the importance of research in this area is simply that, overall, having relationships is associated with good mental and physical health. Having friends and family members with whom one maintains regular contact, and/or a romantic relationship, is associated with greater overall reports of happiness and life satisfaction (e.g., Diener 1984, Myers 1992). Researchers using self-report measures have found that people report feeling best when they are with other people. Being socially involved (as indicated by such diverse indices as one's network size, the density of that network, simply being married or not (especially for men), and perceiving that one has social support available if needed) is linked to better physical health and a lower risk of death (e.g., Berkman and Syme 1979, Cohen 1988). Indeed, many studies have related social involvement to lower susceptibility to the common cold (Cohen et al. 1997), cancer (Glanz and Lerman 1992), and cardiovascular disease (Berkman et al. 1993).

That having social relationships is linked to better mental and physical health is clear. However, why this is so remains unclear. Personality attributes may predict both good relationships and good health. Thus, having relationships may not necessarily be causally linked to good health. However, most researchers in this area believe that it is. Perhaps our relationship partners encourage good health practices. Perhaps they help us cope with and reduce stress. Perhaps just knowing that our partners are there to support us lessens our anxiety (and its physiological concomitants). Such issues are the subject of current research.

3.3 *We Can Predict Who Forms Relationships With Whom*

People tend to form relationships with those who are physically or functionally proximal to them. However, we come into proximity with many others. With which of these people will we form relationships? One factor that influences us is what members of our already existing relationships think. We tend to form relationships with people of whom our present relationship partners approve (Parks et al. 1983). Another factor is similarity. The evidence is clear that 'birds of a feather flock together.' People form relationships with those who are similar in attitudes, age, race, and educational background. In the case of romantic relationships, we also tend to form relationships with others whose physical attractiveness matches our own. This may be because we are drawn to situations which attract similar others. Alternatively, we may actively seek out similar others. Evidence suggests we generally assume people are similar to us and that the discovery of dissimilarities may be what drives people apart (Rosenbaum 1986). With one notable exception, there is very little evidence for complementariness in attraction. Opposites do not seem to attract. Even work on

masculinity and femininity does not suggest that couples including a masculine male paired with a feminine female get along particularly well. Rather, it appears that compatibility is greater when each member of a couple possess both masculine and feminine traits (Ickes and Barnes 1978, Zammichieli et al. 1988).

Performance domains appear to be the one exception to the rule that 'birds of a feather flock together.' Specifically, there is evidence that when an area of performance is central to one's self-definition—say, being good at academic subjects—a close other's good performance in the same area can be threatening to one's self-esteem. That, in turn, can drive the relationship apart. In contrast, partners excelling in different performance domains can promote the relationship. That way each person can bask in the other's reflected glory, and neither suffers from social comparison with the other (Erber and Tesser 1994). Still another important determinant of initial attraction is whether a potential partner likes you. Backman and Secord (1959) have demonstrated this experimentally, and Kenny has demonstrated this in studies of ongoing relationships (Kenny and La Voie 1982).

Finally, we note that there is ample evidence that physical attractiveness is a powerful determinant of initial attraction (e.g., Berschied et al. 1971). Although both males and females seem to value attractiveness in a relationship partner, it seems especially important to males when judging females. Evolutionary theorists suggest that this has resulted from attractiveness being a signal of the health and presumed fertility of that potential mate (Buss 1989). Evolutionary theorists' work on mate preferences also suggests that women especially value males with good resources (presumably because they can provide for offspring). At the same time, these theorists have found evidence of women placing high value on resources in males and males placing high value on attractiveness in females, they have simultaneously found that both males and females value such attributes as kindness and honesty in one another. Indeed, most people rate these attributes as being more important than resources or attractiveness.

3.4 Responsiveness to Needs is a Process that is Important to Good Relationship Functioning

Social psychologists have made considerable strides in identifying interpersonal processes that are important to good relationship functioning once relationships have been established. Responsiveness to one another's needs is one such process. It has long been recognized, initially by reinforcement theorists, that receiving rewards from one's partner is important to relationship satisfaction (e.g., Byrne and Clore 1970). However, recent theory and empirical research has demonstrated that the manner in which people meet (or fail to meet) one another's needs is very important to a person's sense that his or her partner is truly concerned for his or her welfare. This work suggests that a person needs to receive more than benefits from another person to feel intimate with that person. The person needs to believe that the relationship partner understands, validates, and, ultimately, cares for him or her. This appears to emerge from such things as the other responding appropriately to one's statements of needs, self-disclosures, and emotional expressions by listening carefully, responding with understanding and acceptance, and, ultimately, by actually helping (Reis and Patrick 1996) (see *Love and Intimacy, Psychology of*). It seems likely that these processes, in turn, are facilitated by having a partner who is high in empathic accuracy—that is, by having a partner who accurately perceives one's thoughts and emotions (Ickes 1997).

Feeling cared for also appears to be enhanced when partners do not seem to be giving benefits to get benefits. For instance, forgoing self-interest to benefit a partner enhances trust in the partner's caring more than giving a benefit that aids both the recipient and the giver (Holmes and Rempel 1989). People led to desire a close relationship with another want that relationship to be characterized by noncontingent responsiveness to needs (Grote and Clark 1998). That is, they react more positively to benefits given without strings attached than to those given with the expectation of repayment (Clark and Mills 1979), and they do not wish to keep track of just who has contributed what to the relationship in order to allocate benefits (Clark 1984). Giving benefits with strings attached or in exchange for comparable benefits calls the fundamental caring component of relationships into question.

For relationships to be characterized by responsiveness to needs, not only is it important for each person to try to meet the other person's needs, it is also important for each person to convey their needs to their partner. This greatly facilitates the partner's ability to respond to those needs. Thus, willingness to self-disclose has been shown to promote relationships (Cohen et al. 1986). So, too, does it appear that freely expressing one's emotions—which convey needs—is good for relationships (Feeney 1995).

From a personality perspective, it is noteworthy that researchers have found that individual, chronic, 'attachment styles' are closely related to perceptions that others will meet one's needs and hence to a wide variety of behaviors relevant to meeting needs in close relationships. People with a secure attachment style (presumably resulting from having had appropriately responsive caregivers early on) are more effective support-givers and -seekers than are those with insecure attachment styles (e.g., Simpson et al. 1992). Indeed, research on attachment styles has been the fastest-growing area of relationship research over the past two decades (Cassidy and Shaver 1999) (see *Adult Psychological Development: Attachment*).

3.5 Other Processes Associated With High- vs. Low-quality Close Relationships

Responsiveness to needs is not the only interpersonal process important for maintaining high-quality intimate relationships. A number of studies suggest that viewing one's partner positively and reconstruing a partner's faults as virtues is also beneficial for relationships (Murray and Holmes 1993). (Perhaps this aids in maintaining faith that one's partner really does care about one's needs.) These same researchers have also found evidence that when people who are satisfied with their dating relationships do see faults in those relationships, the negative thoughts tend to be linked in memory to positive thoughts about their partners in what they call a 'yes but' manner. For instance, a member of such a couple might say, 'I think his laziness is actually kind of funny; it gives us a reason to laugh.' Less-satisfied couples tend to link negative thoughts about the partner with other negative thoughts about the partner and the relationship (Murray and Holmes 1999). Seeing one's partner in a more positive light than one's partner sees him- or herself also seems good for one's relationship and appears, over time, actually to result in one's partner living up to one's ideals (Murray et al. 1996).

Reflection and social comparison are two more processes that appear to influence the quality of intimate relationships. Reflection refers to feeling that a partner's performance reflects on you, either positively or negatively. Social comparison refers to the process of comparing one's own performance in a given domain with a partner's performance and discovering that one has performed better or worse than that partner. Each process can benefit or harm a relationship. When a performance domain is irrelevant to one's identity, it appears that a partner's good performance in that domain enhances one's own self-esteem (and helps the relationship), whereas a partner's poor performance decreases one's self-esteem (and harms the relationship). The closer the relationship, the greater the effects (Erber and Tesser 1994).

Still other processes that appear to influence the quality of relationships have been identified. For example, research based on Aron and Aron's self-expansion model (1996) has demonstrated that when we fall in love our conception of ourselves expands to include aspects of the other person and that doing exciting and novel activities with one another can increase the quality of a couple's relationship (see *Interpersonal Attraction, Psychology of*). To give another example of processes that appear to increase the quality of relationships, Rusbult has reported research showing that highly committed couples are especially likely to respond constructively (i.e., to accommodate) in the face of a partner's poor behavior and that such couples are also likely to engage in such relationship-protective processes as seeing attractive (and therefore threatening) alternative partners as significantly less attractive than their less committed peers do (Johnson and Rusbult 1989; Rusbult and Van Lange 1996).

Interdependence theory has been particularly useful in pinpointing factors independent of ongoing interpersonal processes occurring within the relationship that influence both satisfaction and commitment to a relationship. For instance, although interdependence theorists agree with reinforcement theorists and with those emphasizing responsiveness to needs that rewards received (and costs avoided) are important determinants of relationship satisfaction, they add the important point that the expectations people bring with them to any particular relationship will influence satisfaction. The higher a person's expectations (comparison level) are for the relationship, the tougher their standards will be and, potentially, the lower the satisfaction with any given relationship.

Interdependence theorists also emphasize that satisfaction is not the only, nor necessarily even the most important, determinant of the stability of a relationship. Instead, they have identified a separate variable—commitment to a relationship—as a stronger determinant of relationship stability. Satisfaction does contribute to higher commitment to a relationship, but there are other important determinants of commitment as well. One is 'comparison level for alternatives.' The better one's alternatives, the less committed one will be to the current relationship (independent of satisfaction). This means that one could lack commitment to a relationship which is, in and of itself, satisfying or be committed to a relationship which is, in and of itself, unsatisfying (but more satisfying than one's alternatives). Investments (anything put into the relationship that cannot be easily retrieved should one exit the relationship) also influence commitment. Having investments such as joint friends, joint memories, children, a joint house, well-established routines, and so forth all increase one's sense of having invested in a relationship and, hence, one's commitment to that relationship. Finally, outside prescriptives, either coming from oneself or one's social circle, can influence commitment which, as we have already noted, is an important predictor of relationship stability as well as of partners' willingness to accommodate one another (Rusbult and Van Lange 1996).

4. Future Directions

It is always difficult to predict the future. However, it seems likely that research on the interpersonal processes that occur in, and affect the quality of, friendships, romantic relationships, and family relationships will continue to occupy center stage for a while. We can also expect to see increasing collaboration between epidemiologists and health psychologists, who have demonstrated links between having relationships and health, and social and personality psychologists, who

have investigated interpersonal processes that contribute to high-quality intimate relationships. Many researchers currently are calling for such collaboration, in the hope that the fallout will be an understanding of how and why relationships are beneficial for mental and physical health.

Further, with the increasing emphasis on studying relationships longitudinally, it may be expected that theories and empirical work on interpersonal processes occurring in relationships over time will become more frequent. Right now, we know lots about first impressions and some about the very initial stages of relationship deepening. We also know a bit about the dissolution of relationships. However, we know little about long-term trajectories of happy and unhappy relationships. We will know much more in 10 years.

It also seems likely that relationships other than those emphasized in recent research (and hence in this entry) will become the focus of more research—relationships such as teacher–student and physician–patient relationships, relationships between co-workers, and relationships between siblings.

See also: Adult Psychological Development: Attachment; Friendship, Anthropology of; Personality and Marriage; Sibling Relationships, Psychology of; Social Support and Health

Bibliography

Aron A, Aron E 1996 Self and self expansion in relationships. In: Fletcher G J O, Fitness J (eds.) *Knowledge Structures in Close Relationships: A Social Psychological Approach*. Erlbaum, Hillsdale, NJ, pp. 325–44

Backman C W, Secord P F 1959 The effect of perceived liking on interpersonal attraction. *Human Relations* **12**: 379–84

Bargh J A, Raymond P, Pryor J B, Strack F 1995 Attractiveness of the underling: An automatic power–sex association and its consequences for sexual harassment and aggression. *Journal of Personality and Social Psychology* **68**: 768–81

Berkman L F, Syme S L 1979 Social networks, host resistance, and mortality: A nine year follow-up study of Alameda County residents. *American Journal of Epidemiology* **100**: 186–204

Berkman L F, Vaccarino V, Seeman T 1993 Gender differences in cardiovascular morbidity and mortality: The contribution of social networks and support. *Annals of Behavioral Medicine* **1**: 112–8

Berscheid E, Dion K, Walster E, Walster G W 1971 Physical attractiveness and dating choice: A test of the matching hypothesis. *Journal of Experimental Social Psychology* **7**: 173–89

Berscheid E, Snyder M, Omoto A M 1989 The Relationship Closeness Inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology* **57**: 792–807

Blumstein P, Kollock P 1988 Personal relationships. *Annual Review of Sociology* **14**: 467–90

Bugental D B 2000 Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin* **126**: 187–219

Buss D M 1989 Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences* **12**: 1–49

Byrne D, Clore G 1970 A reinforcement model of evaluative responses. *Personality: An International Journal* **1**: 103–28

Cassidy J, Shaver P (eds.) 1999 *Handbook of Attachment: Theory, Research, and Clinical Applications*. Guilford Press, New York

Clark M S 1984 Record keeping in two types of relationships. *Journal of Personality and Social Psychology* **47**: 549–57

Clark M S, Mills J 1979 Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology* **37**: 12–24

Clark M S, Mills J 1993 The difference between communal and exchange relationships: What it is and is not. *Personality and Social Psychology Bulletin* **19**: 684–91

Cohen S 1988 Psychosocial models of the role of social support in the etiology of physical disease. *Health Psychology* **7**: 269–97

Cohen S, Doyle W J, Skoner D, Rabin B S, Gwaltney J M 1997 Social ties and susceptibility to the common cold. *Journal of the American Medical Association* **277**: 1940–44

Cohen S, Sherrod, D R, Clark M S 1986 Social skills and the stress-protective role of social support. *Journal of Personality and Social Psychology* **50**: 963–73

Diener E 1984 Subjective well-being. *Psychological Bulletin* **95**: 542–75

Erber R, Tesser A 1994 Self-evaluation maintenance: A social psychological approach to interpersonal relationships. In: Erber R, Gilmour R (eds.) *Theoretical Frameworks for Personal Relationships*. Erlbaum, Hillsdale, NJ, pp. 211–34

Feeney J A 1995 Adult attachment and emotional control. *Personal Relationships* **2**: 143–60

Fiske A P 1992 The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review* **99**: 689–723

Fiske S 1993 Controlling other people: The impact of power on stereotyping. *American Psychologist* **48**: 621–8

Glanz K, Lerman C 1992 Psychosocial impact of breast cancer: A critical review. *Annals of Behavioral Medicine* **14**: 204–12

Grote N, Clark M S 1998 Why aren't indices of relationship costs always negatively related to indices of relationship quality? *Personality and Social Psychology Review* **2**: 2–17

Holmes J G, Rempel J K 1989 Trust in close relationships. In: Hendrick C (ed.) *Close Relationships: Review of Personality and Social Psychology*. Sage, London, Vol. 10, pp. 187–220

Ickes W 1997 *Empathic Accuracy*. Guilford, New York

Ickes W, Barnes R D 1978 Boys and girls together and alienated: On enacting stereotyped sex roles in mixed sex dyads. *Journal of Personality and Social Psychology* **36**: 669–83

Kelley H H, Berscheid E, Christensen A, Harvey, J H, Huston T L, Levinger G, McClintock E, Peplau L A, Peterson D 1983 *Close Relationships*. Freeman, New York

Kelley H H, Thibaut J 1978 *Interpersonal Relations: A Theory of Interdependence*. Wiley, New York

Kenny D, LaVoie L 1982 Reciprocity of attraction: A confirmed hypothesis. *Social Psychology Quarterly* **45**: 54–8

Murray S L, Holmes J G 1993 Seeing virtues in faults: Negativity and the transformation of interpersonal narratives in close relationships. *Journal of Personality and Social Psychology* **65**: 701–22

Murray S L, Holmes J G 1999 The (mental) ties that bind: Cognitive structures that predict relationship resilience. *Journal of Personality and Social Psychology* **77**: 1228–44

Murray S L, Holmes J G, Griffin D W 1996 The benefits of positive illusions: Idealization and the construction of sat-

Myers D G 1992 *The Pursuit of Happiness: Who is Happy—and Why?* Marrow, New York
Myers D G, Diener E 1995 Who is happy? *Psychological Science* **6**: 10–19
Parks M R, Stan C M, Eggert L L 1983 Romantic involvement and social network involvement. *Social Psychology Quarterly* **46**: 116–31
Reis H T, Patrick B C 1996 Attachment and intimacy: Component processes. In: Higgins E T, Kruglanski A (eds.) *Social Psychology: Handbook of Basic Principles*. Guilford, New York, pp. 523–63
Rosenbaum M E 1986 The repulsion hypothesis: On the nondevelopment of relationships. *Journal of Personality and Social Psychology* **51**: 1156–66
Simpson J A, Rholes W S, Nelligan J 1992 Support-seeking and support-giving within couples in an anxiety provoking situation: The role of attachment styles. *Journal of Personality and Social Psychology* **62**: 434–46
Wish M, Deutsch M, Kaplan S J 1976 Perceived dimensions of interpersonal relations. *Journal of Personality and Social Psychology* **33**: 409–20
Zammichieli M E, Gilroy F D, Sherman M F 1988 Relation between sex-role orientation and marital satisfaction. *Personality and Social Psychology Bulletin* **14**: 747–54

M. S. Clark

Social Science, the Idea of

Can human studies, e.g., sociology and psychology, be properly scientific? Since the mid-nineteenth century, philosophers and social scientists have debated this question more or less continuously. It has periodically assumed considerable urgency, indeed become a focus of attention within the disciplines, as new influences on these subjects were felt or as new models of science itself were developed. Indeed, the complexity and potential obscurity of the issue is already evident. There is an implied equation, '*human studies* are/are not *science*,' in at least two variables. There has been considerable debate, at least since the 1850s, about what kinds of role human studies might play. Likewise, what is required for a discipline to be scientific has also been a matter of continuing controversy. As Peter Manicas put it (1987, p. 168), 'neither "science" nor, *a fortiori*, the branches of the social sciences are "natural kinds".' Notice, furthermore, that the variables are not entirely independent of one another. What human studies practitioners have aspired to has itself been influenced by what they think about the desirability and the possibility of making their enterprises genuinely scientific. Similarly, though perhaps to a lesser extent, what has counted as scientific has reflected the actual practices of, *inter alia*, students of sociology and psychology. To approach this question, a suitably sophisticated taxonomy of possibilities is therefore required.

1. Some Taxonomic Distinctions and a Definition

If human studies are to count as scientific, then what it takes to make a science must first be considered.

1.1 Modality

First, *empirical* and *normative* approaches to this question need to be distinguished—roughly, 'What are the sciences like?' and 'What should the sciences be like?' Of course, these two questions cannot be treated in isolation, especially in view of the widely accepted maxim that '*Ought* implies *can*.' This maxim means, in this instance, that those subjects whose scientific status is being assessed cannot sensibly be required to attain some excellence unless, of course, *paradigms* of scientific achievement have themselves attained this excellence. With these points in mind, the question about human studies can be understood in this way: given what human studies *are* like, do they or can they conceivably measure up to whatever normative standards of excellence the best scientific disciplines *do* measure up to?

1.2 Focus

Of course, there are other issues about what it takes to make a science. The question of focus is perhaps the most important. Science and nonscience can be distinguished either in terms of their *products* or in terms of their *processes*. This fundamental distinction divides pre- and post-Kuhnian conceptions of science. Whereas *before* Thomas Kuhn's epochal work *The Structure of Scientific Revolutions* (1970), science was distinguished from nonscience by the kinds of *statements* the two discourses produced, or, more importantly, by the status of these statements relative to epistemic standards of confirmation, generality, or objectivity, *after* this work, science was increasingly distinguished from nonscience in terms of the activities of its practitioners. Obviously, it will make a difference which of these criteria is deemed more appropriate. On the product conception, it might count against the scientific pretensions of sociology and psychology, for instance, that the more robust results in these subjects lack the kind of generality that robust results typically display in paradigm, 'hard' sciences such as physics and chemistry. But perhaps this lack of generality is irrelevant if a process criterion is adopted. Perhaps sociologists and psychologists obtain their less general findings by the same technical means and methods that natural scientists use in producing more general or more stable results.

1.3 Scope

Finally, there is the issue of scope. Roughly speaking, *wide* and *narrow* understandings of the sciences can be

distinguished. On the narrow reading, it is the concrete and specific particularities of the paradigm sciences that provide criteria for judging other candidate sciences. Typically, the paradigm sciences are laboratory based, involve heavy reliance on statistical techniques of data analysis, and produce results in the form of highly confirmed differential equations relating measurable variables. On the narrow criterion, then, to count as a science, any other discipline must also exhibit these specific features. On the other hand, a wide approach to the issue requires of candidate sciences merely that they exhibit some abstract and high-level virtue which is shared by paradigm scientific achievements—that they permit the attainment of objectivity, for instance. A wide approach to the issue is probably indicated; at the concrete and specific level of technique and theory, there is enough variation *within* the paradigm sciences to rule out any narrow approach.

In considering whether human studies are scientific, an approach which is normative, process oriented, and wide in its scope should therefore be adopted. The question will be, in other words, whether it is possible that human studies be conducted in such a way that they might achieve whatever levels of objectivity in their procedures as are attainable in the paradigm sciences. (The issue is phrased in this way to permit the rather complicated judgment (likely to be favored by 'postmodernists') that human studies can achieve the same *degree* of objectivity as paradigm sciences, but that neither can achieve the *kind* of 'objectivity' that was previously thought to be characteristic of paradigm, natural sciences such as physics.)

1.4 Objectivity

To assess the scientific status of disciplines such as sociology and psychology, the objectivity of their procedures must, on this account, be assessed. But what is objectivity? Of course, there may be no more determinacy on this matter than on some others which have already been alluded to. Just as 'science' does not designate any natural kind, neither does 'objectivity'; it will be a matter of persistent debate what counts as objective. Some preliminary observations can nevertheless be offered.

First of all, transcendental and nontranscendental accounts of objectivity can be distinguished. In the transcendental account, objectivity in investigative activities has been attained when and only when all that is specific to the situation in which they are undertaken, including the personalities of the investigators, has been eliminated from them. Individuals who are otherwise dissimilar must, on this account, conduct themselves in some common way when they examine the objects of their investigations. (When scientists have managed to specify methods for the replication of results, some approximation to this ideal may have been achieved.) They know how to tell others who differ from them how to negate the effects of these differences and thus achieve the same results as they have. Objectivity, on this account, consists of transcending local circumstances by eliminating all factors which vary from one situation to another.

Of course, philosophers and others working under the rubric of 'postmodernism' have widely repudiated this ideal. True transcendence of local circumstances is not possible, it is not desirable, and, indeed, it is not even necessary for a form of objectivity (that is worthy of being pursued) to be attained. Objectivity requires not that local differences be eliminated, but, rather, that they be exploited in order to discover, where this is possible, a point of convergence of otherwise very different perspectives. According to this account, objectivity is attained when an individual is able to affirm from his or her own perspective a system of propositions that can likewise be affirmed from the otherwise different perspectives of such other individuals whose agreement matters (see Rorty 1989).

2. Why the Issue Matters

Why should it matter whether human studies are or are not scientific, especially if there is little agreement about the concepts in terms of which the issue is defined. Discussing the so-called 'interpretive turn' in human studies, Geertz (1993, p. 8) put the point this way: 'Whether this [development] is making the social sciences less scientific or humanistic study more so (or, as I believe, altering our view, never very stable anyway, of what counts as science) is not altogether clear and perhaps not altogether important.' There are, in fact, at least three reasons why it matters whether human studies are considered scientific.

2.1 Institutionally

The scient*ism* of its public culture, i.e., the privilege that is granted to scientific disciplines and their methods and results, is one prominent characteristic of modernity. Science has immense, though perhaps no longer undisputed, cognitive authority, and plays an influential role in defining as well as creating the world which human beings inhabit. If human studies succeed in gaining recognition as scientific subjects, then they too could hope to exercise this kind of authority. Foucault, typically, puts the matter the other way around (1980, p. 85): 'What types of knowledge do you want to disqualify in the very instant of your demand: "Is it a science?" Which speaking, discoursing subjects of experience and knowledge do you want to "diminish" when you say: "I who conduct this discourse am conducting a scientific discourse,

and I am a scientist?" Which theoretical-political *avant garde* do you want to enthrone in order to isolate it from all the discontinuous forms of knowledge that circulate about it?'

2.2 Heuristically

Whether sociologists and psychologists see themselves and are seen by others as scientists will, of course, influence the conduct of their enquiries and the kinds of products they aspire to deliver. As Geertz (1993, p. 8) put it, the issue is 'who borrows what from whom.' Do human studies borrow techniques of observation, experimentation, and data manipulation from 'hard' sciences such as chemistry and biology, or do they, instead, look to other disciplines for their techniques or their standards of objectivity?

Indeed, this issue is part of a long-standing debate. In the late nineteenth century, the crucial questions were about the *Geisteswissenschaften*—are they concerned with understanding or with explanation? Can they articulate general psychic and/or social laws or must they, instead, confine themselves to the accurate description and contextualization of concrete and specific incidents (Windelband's *nomothetic* vs. *idiographic* conceptions)? In the middle of the twentieth century, the key issue was the cognitive status of social scientific generalizations and, indeed, their very possibility in the face of innumerable obstacles and points of difference from paradigm 'hard' sciences. At the end of the twentieth century, the question is whether interpretation, with all its vicissitudes, constitutes a more central method for human studies than law-based explanation, and, indeed, whether interpretation itself is possible in a sufficiently 'objective' manner to permit the judgment that human studies are, perhaps in some extended sense, scientific.

2.3 Morally

How human studies are conceptualized may have profound moral significance. This point is dependent on the previous two. Suppose human studies are modeled on the 'hard' sciences and that success in doing so is conspicuous enough so that these subjects gain the kind of cognitive authority that some of their practitioners have long aspired to. In this case, ideas about the human situation that are moral and cultural foundations of our longest-standing and most deeply embedded traditions are at risk of being undermined. Putnam (1978, p. 63) put it this way: '[S]uppose our functional organization became transparent to us. Suppose we had a theory of it, and we could actually *use* this theory in a significant class of cases. What would happen to us? … Would it be possible even to think of *oneself* as a *person*? … [T]he development of *that* sort of knowledge of ourselves and each other would modify our natures in ways that we cannot predict at all. Every institution we now have: arts, politics, religion—even science—would be changed beyond all recognition.'

2.4 Some Positions

Whether human studies can be scientific is therefore too important a question simply to ignore. And, of course, it has not been ignored. Philosophers, social scientists, and others have, in fact, taken a number of positions on this question. Before proceeding with more substantive matters, a threefold primary distinction among these positions should be considered.

In the middle of the twentieth century, the most common position, though not so designated until later, was *naturalism*. According to this doctrine, there is nothing methodologically or theoretically distinctive about human studies in relation to paradigm, natural sciences. In particular, social and natural scientists legitimately share both aspirations (to discover and confirm general laws covering specific phenomena) and techniques (operationalization of variables, controlled experimentation, and statistical analysis of its results).

Both earlier and later than this period, however, a very common, perhaps even the most common position was *interpretivism*. According to this doctrine, human studies are distinguished from paradigm, 'hard' sciences both in their aspirations and in their techniques. Social scientists should not try to emulate the paradigm sciences narrowly construed (see Sect. 1.3); they should strive, instead, to interpret human social and psychic phenomena, and should employ the techniques of hermeneutics, broadly construed, to achieve this.

Finally, towards the end of the twentieth century, though also but to lesser degrees at earlier periods, 'scepticism' about the social sciences has been an increasingly influential position. On this account, laws of human being cannot be discovered, nor can properly objective interpretations of what human beings do be provided. Although the results of sociological or psychological research are often presented in the form of statements of knowledge, they are not entitled to this status, and must be understood in some other, less exhaulted way.

Sceptical positions may be either local or global. Global scepticism, often associated with the work of Jacques Derrida, repudiates all claims to determinate and methodologically adequate knowledge of human phenomena. Local scepticism adopts such a dismissive attitude only to certain claims, e.g., those which have been derived in some specific way or from some specific social position. (For instance, Marxists might be skeptical in this local sense about the claims

produced by mainstream, and hence in their view 'bourgeois' sociology or psychology. Similarly, feminists are often skeptical about the claims of mainstream and hence in their view 'patriarchal' social science.)

3. Impediments to Human Studies Being Scientific

That the legitimacy of human studies has so long been a matter of discussion is *prima facie* evidence that the investigation of human phenomena is not straightforward. What, then, are some of the reasons why it might not be possible to have a genuinely scientific approach to human social and psychic phenomena?

3.1 Preinterpretation

Theorists otherwise as diverse as Peter Winch, R. G. Collingwood, and Alfred Schütz have pointed, in distinguishing human studies from science *per se*, to the fact that students of human phenomena work, whereas paradigm scientists do not work, with materials that already have meaning and significance (see, e.g., Winch 1958, p. 87). There are strong and weak versions of this point. On the weak version, students of human phenomena must always rely, methodologically, on the objects of their research in ways in which paradigm scientists need not. Saying *what* is going on is preliminary to saying *why* it is happening, and hence to any scientific explanation. But what is going on, socially or psychically, must be understood, in Geertz's phrase, 'from the native's point of view.' How the subject understands what is going on determines the description of what is going on. On the strong version, once scientists have characterized 'the native's point of view,' they have discovered all that they need to or can know about the objects of their investigation. In either case, what paradigm scientists do and what human studies practitioners do is very different. The latter seem forced to rely on the subjective perspectives of the objects of their research.

3.2 Instability

There are three, inter-related factors which make it unlikely that sociologists and psychologists will be able to develop stable, objective accounts of the phenomena they take an interest in.

First, there is the *historicity* of both the subjects and the objects of human studies research. When Gadamer (1975, p. 204) asks 'Is not the fact that consciousness is historically conditioned inevitably an insuperable barrier to reaching perfect fulfilment in historical knowledge?' he articulates this fundamental difficulty: neither scientist nor subject 'stands still,' and hence what the one should say about the other changes as they themselves change. Paradigm, natural sciences, on the other hand, typically deal with statements of 'timeless' generality.

Second, in human studies relations of representation are subject to *reflexivity* in a way that is not common in the sciences *per se*. What practitioners say about human phenomena influences how their subjects enact or react to these phenomena, and hence often undermines the claim, necessary if objectivity is sought, that scientists have accurately *represented* the behavior of their subjects. (Practitioners have, instead, caused or persuaded subjects to act so that their behavior conforms to the practitioners' descriptions of it.) To mid-twentieth century philosophers, this possibility was best known under the heading 'reflexive predictions' (self-defeating and self-confirming 'prophecies'), but the crucial points have since been generalized extensively. Of course, as Lyotard (1984, p. 57) rightly points out, reflexivity must not be understood in too one-sided a way. It is not just that sociologists and psychologists can create the objects of their investigations; it is, as well, that these objects can, *strategically*, resist the characterizations which practitioners attempt to fit them with. This too is an impediment to the achievement of objectivity as it is usually conceived.

Finally, there is radical *indeterminacy* about the objects of investigation. This point is an ontological not an epistemological one. It is not that, in some or perhaps many cases, scientists cannot tell when they have characterized some phenomenon accurately. It is, rather, that there may be no sense, in these cases, to the very idea of accurate characterization. Typically, this means that there are, in the human domain, no fixed objects with well-defined characteristics for scientists to (try to) represent. In its more familiar forms, this point is, of course, an obvious development of Dilthey's observations about the 'hermeneutic circle.' It has been much emphasized by Derrida (1978) and others of his school. Specifically, if interpretation takes place against and in terms of a background that is not 'given' but must itself be an object of interpretation, then every interpretation has the form, not of an objective (if possibly incorrect) description, but, rather, of a (performative) proposal to see the situation in one of the many ways in which it could (as) legitimately be seen. There is on this account insufficient interpretive stability to warrant attributions of objectivity.

3.3 Complexity

The economist Hayek (1978, 26f) has long emphasized the peculiar kind of complexity which human studies

practitioners must try to deal with. Such 'organized complexity' resists reduction either to deterministic or to statistical models that might be appropriated from the 'hard' sciences. The behavior of a social aggregate often cannot, in other words, be reduced to the sum or product of the behavior of its individual parts, nor can it be adequately understood by characterizing its statistical properties (as the behavior of an aggregate of gas molecules might be understood, for instance). Notice, furthermore, that there may be moral and prudential (as well as practical) reasons for refusing to reduce this complexity in ways standard in the 'hard' sciences, i.e., by 'experimentalizing' it. So, while scientists might be able to 'standardize' individual agents in an experimental setting so that their interactions *there* can be understood deterministically or statistically, there are moral objections to exporting these conditions to *extra*-experimental settings. Indeed, there may be prudential reasons for not using these techniques of standardization in 'naturalistic' settings—as is implied, of course, by Hayek's own analysis of the market. In particular, adequate functioning of a complexly organized whole may depend precisely on its individual elements *not* being standardized. In either case, there are thus limits to the objectivity (actually the 'ecological validity') of the results that might be obtained in laboratory settings.

3.4 Interim Conclusion

All these impediments manifest themselves, of course, in a simple *fact*: there is, especially compared with paradigm 'hard' sciences, a conspicuous lack of progress in human studies. In particular, multiple and opposed paradigms persist in human studies, especially in any area of investigation which is likely to be of broadly 'political' interest and concern. As Geertz (1993, p. 4) has said, 'calls for "a general theory" of just about anything social sound increasingly hollow, and claims to have one megalomanic. Whether this is because it is too soon to hope for unified science or too late to believe in it is, I suppose, debatable. But it has never seemed further away, harder to imagine, or less certainly desirable than it does now.'

4. Alternative Conceptions of Human Studies

It may not make sense to compare sociology and psychology to the 'hard' sciences. How, then, are they to be conceptualized … and hence institutionalized? There are a number of possibilities. (Manicas's cautionary note still applies, of course; if 'science' isn't a natural kind, then no more is 'literature' or any of the other disciplines with which human studies might be compared.)

4.1 Literature

If human studies cannot properly be compared to the 'hard' sciences, perhaps a comparison with literature and other forms of potentially edifying narrative or practical discourses is more appropriate. This, certainly, is the opinion of a wide range of late-twentieth century theorists, including Geertz, Charles Taylor, Gadamer, and others. Geertz, for instance, advocates (1993, p. 6) that we 'turn from trying to explain social phenomena by weaving them into grand textures of cause and effect to trying to explain them by placing them in local frames of awareness … .' Of course, it might seem, to practitioners, that such a heuristic reorientation (see Sect. 2.2) would reduce the likelihood that human studies might have significant institutional influence (see Sect. 2.1). However, it needs to be borne in mind that the institutional influence of sociology and psychology is rather patchy (this is made clear in utilization studies), and that literature still forms the basis for highly influential mass-media representations and analyses of the human situation.

Another related possibility is this. While sociology and psychology do not or cannot provide an objective account of human phenomena, they can and do provide a variety of potentially important *perspectives* on these phenomena, and thus enable human beings to orient themselves more productively to them, especially by giving them frameworks for understanding these phenomena. In this vein, Rorty (1982, p. 247) draws a contrast between scientistic goals of prediction and control, on the one hand, and, on the other, practical goals of developing 'descriptions which help one decide what to do.' Since deciding what to do depends on values and other normative elements, and since there is widespread and apparently irreducible diversity in relation to these elements, the kind of 'practical knowledge' that nonscientistic forms of psychology and sociology might provide will not, of course, satisfy the most demanding requirements of objectivity.

4.2 Critique

There is a subtle difference, but an important one between the view that human studies offer edifying perspectives on or frameworks for the interpretation of human phenomena (see Sect. 4.1) and the view that they offer a basis for critique and the advocacy of change. In the latter half of the twentieth century, critical theorists (e.g., Habermas), genealogists (e.g., Foucault), feminists, and advocates for other marginalized groups (see Seidman 1994) proposed and in some cases made progress in achieving a 'takeover' of human studies activities and institutions. Across this spectrum techniques which have some claim to being 'antiscientific' (Foucault 1980, p. 83) are widely employed. In particular, practitioners of critique shift

the focus from showing how what is, socially, is an intelligible consequence of its causal antecedents to showing how what is, socially, represents the highly contingent realization of practices and structures to which there are clear, achievable, and desirable alternatives. As Foucault says (Rabinow 1984, p. 46): '[T]his critique will be genealogical in the sense that it will not deduce from the form of what we are, what is impossible for us to do and to know; but it will separate out, from the contingency that has made us what we are, the possibility of no longer being, doing, thinking, what we are, do, or think. It is not seeking to make possible a metaphysics that has finally become a science; it is seeking to give new impetus, as far and wide as possible, to the undefined work of freedom.' It is not, then, a matter of objective representation, but, rather, of politically effective analysis of the possibilities for change. (This is, of course, an extension of Karl Marx's 11th thesis on Feuerbach: 'Philosophers have only *interpreted* the world in various ways; the point is to *change* it.')

4.3 Disciplines or Ideologies

There are (see Sect. 3.3) moral and prudential disincentives to investigating human phenomena experimentally and, *a fortiori*, to 'exporting' into 'naturalistic' settings those conditions in which objective, reproducible experimental effects have been obtained. This has not, of course, stopped such things from happening and, indeed, it is the main conclusion of Foucault and others of his school that human studies have played a key role in this process, and, in particular, that they have contributed to the production of a 'disciplinary society' in which individuals are governed via their consent, i.e., through the ways in which they are made into 'subjects.' Although Jürgen Habermas disputed much that Foucault wrote and did, he was at one with him on this point. Habermas (1994, p. 54) says, for instance: 'It is no accident that these sciences, especially clinical psychology, but also pedagogy, sociology, political science, and cultural anthropology, can, as it were, frictionlessly intermesh in the overall technology of power that finds its architectural expression in the closed institution. They are translated into therapies and social techniques, and so form the most effective medium of the new, disciplinary violence that dominates modernity.'

That human studies are 'disciplines' is one way of capturing their importance and their negative impact on the lives of ordinary people; that they have functioned ideologically is another way of making this point. Perhaps the most important issue here is the way in which human studies have been recruited by other powerful institutions to serve the specifically political interests of those institutions while seeming to provide a nonpolitical analysis, a value-neutral account of contemporary social conditions. The idea of expertise has played an important role in this development. Insofar as social and psychic problems can be modeled on purely technical problems, solutions can be offered and consumed on a purely technical basis, and the reality of their dependence on fundamentally political matters can remain inaccessible to consciousness.

5. Conclusion

Should human studies plausibly aspire to scientific status? This remains questionable, as it has long been. Can they realistically aspire to such a status? This too is questionable. Perhaps Rorty (see Sect. 4.1) is right to suggest that human studies have aspired to too much, that their two main goals are disjoint and hence cannot both be achieved under the same methodological or institutional auspices. In this case, practitioners would perhaps be wise to disentangle their various objectives, and adjust their 'aspiration levels' (Simon 1996, p. 30) to more realistic goals. Or perhaps an even more radical adjustment of aspirations and disciplinary cultures is called for. Certainly Toulmin thinks so. Toulmin (1992, p. 193) says: 'The task is not to build new, more comprehensive systems of theory with universal and timeless relevance, but to limit the scope of even the best-framed theories, and fight the intellectual reductionism that became entrenched during the ascendancy of rationalism … [thus reinstating] respect for the pragmatic methods appropriate in dealing with concrete human problems.' On this account, human studies are to be understood more on the model of *ad hoc* tools for a variety of perhaps incommensurable human purposes, and not, as in their ascendency in the mid-twentieth century, as sciences with all the cognitive and institutional entitlements which are associated with that sometimes honorific term.

See also: Atomism and Holism: Philosophical Aspects; Causes and Laws: Philosophical Aspects; Economics, Philosophy of; Idealization, Abstraction, and Ideal Types; Individualism versus Collectivism: Philosophical Aspects; Interpretation and Translation: Philosophical Aspects; Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects

Bibliography

Derrida J 1978 *Writing and Difference*. University of Chicago Press, Chicago

Foucault M 1980 *Power/Knowledge*. Pantheon, New York

Gadamer H-G 1975 *Truth and Method*. Seabury Press, New York

Geertz C 1993 *Local Knowledge*. Fontana Press, London

Habermas J 1994 The critique of reason as an unmasking of the human sciences. In: Kelly M 1994 *Critique and Power*. MIT Press, Cambridge, MA
Hayek F A 1978 *New Studies in Philosophy, Politics, Economics and the History of Ideas*. Routledge and Kegan Paul, London
Kuhn T S 1970 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
Lyotard J-F 1984 *The Postmodern Condition*. University of Minnesota Press, Minneapolis, MN
Manicas P T 1987 *A History and Philosophy of the Social Sciences*. Blackwell, Oxford, UK
Putnam H 1978 *Meaning and the Moral Sciences*. Routledge and Kegan Paul, London
Rabinow P 1984 *The Foucault Reader*. Penguin, London
Rorty R 1982 *Consequences of Pragmatism*. University of Minnesota Press, Minneapolis, MN
Rorty R 1989 *Contingency, Irony, and Solidarity*. Cambridge University Press, Cambridge, UK
Seidman S 1994 *Contested Knowledge*. Blackwell, Oxford, UK
Simon H A 1996 *The Sciences of the Artificial*. MIT Press, Cambridge, MA
Toulmin S 1992 *Cosmopolis*. University of Chicago Press, Chicago
Winch P 1958 *The Idea of a Social Science*. Routledge and Kegan Paul, London

F. D'Agostino

Social Security

'Social Security' denotes both a complex of public institutions to protect individuals against common life risks and the programmatic idea that a commonwealth should provide all of its members such protection that they may feel secure within that commonwealth. This concept is broader than 'social insurance' and narrower than 'welfare state,' but relates to similar problems.

1. *The Emergence of the Concept*

'Security' (Latin *securitas*) has a long-standing tradition in the political rhetoric in Europe since the Roman Empire. In early modern times it was often used together with 'welfare' and 'felicity' in order to paraphrase the duties of the prince to care for his subjects. Thomas Hobbes made 'the safety of the people' a central task for his *Leviathan*, and this idea became seminal for modern political theory and constitutional practice. Wilhelm von Humboldt (1792) was the first to distinguish clearly between 'safety/security' and 'welfare': he restricted the task of the state to the provision of 'safety/security,' i.e. on the one hand the protection of citizens' freedom from external threats, and on the other hand the internal protection of citizens' commitments and rights via the judiciary and police. 'Welfare' by contrast was now assigned to the private realm as the individual pursuit of happiness from which the state was to abstain from interfering. In the nineteenth century the idea of public safety and security became restricted to the defense and protection of individual rights, the latter becoming differentiated between security within the public realm and the reliability of legal protection. But in the twentieth century the idea of security became paramount and was related to psychology and psychoanalysis, to the theory of cognition, to industrial and technical safety as well as to social security.

The specific context for the emergence of the term 'Social Security' was the United States and the Great Depression. In American pragmatism 'the quest for certainty' was abandoned and superseded by 'the search for security' (John Dewey). Inspired by the 'four wishes' of William I Thomas, the 'wish for security' became a basic assumption about human motives and 'insecurity' a commonplace attribute of the *Zeitgeist*.

The invention of the expression 'Social Security' is commonly attributed to Franklin Delano Roosevelt or to his entourage. Roosevelt used it first in a Message to Congress on 30 September 1934 after having created a 'Committee on Economic Security' to prepare the 'Social Security Act' on 29 June 1934. However, a voluntary association promoting 'Old Age Security' renamed itself 'American Association for Social Security' in 1933 and changed the title of its journal from *Old Age Security Herald* to *Social Security* because the program of this association had been expanded from old-age pensions to all branches of social insurance (Rubinow 1934, pp. iii, 279).

In any case the expression quickly gained steam: it coined first the US 'Social Security Act' of 1935 (SSA). It then appeared in the Atlantic Charter of 1941, became the catchword for the Declaration of Philadelphia (1944) of the International Labour Organization (ILO) and found its way in 1948 into the Universal Declaration on Human Rights of the United Nations (Art. 22). After the Second World War 'Social Security' had become a leading term internationally in the programs for the development of welfare states. Its specific meaning remained vague, however.

2. *Institutional Developments*

At the time when the SSA was drafted, a public commitment to income maintenance for people in need was by no means a new idea. At the end of the nineteenth century two models were competing in Europe: public subsidizing of mutual benefit associations (called the 'System of Gent'), and social insurance along the lines of the German ('Bismarck') Model. The British social reforms brought two new principles: payment of old-age pensions for deserving poors from the budget (1908), and flat-rate contributions and benefits in unemployment insurance (1911). In 1913 Sweden created the first universal system for securing income in old age. After the

Russian Revolution the Bolsheviks drafted a comprehensive system of social protection against all major contingencies of life, following a concept which Lenin had already propagated in 1912. Starting in France (1932), hitherto voluntary child allowances of employers were made obligatory, and this ushered in a new model of administration: calling upon private bodies for the attainment of public social policy ends. Thus social insurance for dependent workers was already competing with other models of social protection. The victory of the term 'social security' over 'social insurance' was a consequence of its vague content which allowed the inclusion of all kinds of public social protection. In 1947 the 'International Social Insurance Conference' changed its name to 'International Social Security Association' and opened membership to state-administered systems (such as those of Britain and the US) which until then had been excluded.

Thus, the *American* SSA brought a new name but hardly new ideas. It had a problem, however, that was not present in Europe, namely federalism. The Supreme Court had long held attempts by the Federal Government to introduce measures of social protection to be unconstitutional. And it was only after a positive ruling by the Supreme Court on the SSA in 1937 that federal social policies took shape, including policies to combat the risks of the death of a breadwinner and of disability. Besides a federal system of Old-Age Insurance the SSA introduced a new federal instrument into American politics: matching grants to the states in order to induce them to abolish the old forms of the poor law, to introduce or to improve a system of unemployment compensation, and to provide better care to the blind as well as to mothers and children. Thus the hitherto strictly separated domains of the federal and state government became linked for the first time. This remained a continual subject of controversy in the American polity with lasting consequences for the structure of social protection: there exists a two-tiered system consisting of 'social security' on the federal and of 'welfare' on the state level. The first program, protecting most of the employed population, is well administered, highly stable, and positively valued. The second, providing help as a last resort, has remained an ever-contested object of repeated reforms. It is unevenly administered at the local level and its recipients remain stigmatized.

That which was innovative in the concept of social security was developed first in a report to the British Government entitled 'Social Insurance and Allied Services' in 1942, by a commission chaired by William Beveridge. It is *the program of a comprehensive and unified system, covering the whole population against all basic risks of life*—an idea that Beveridge had already formulated in 1924. The Beveridge report met with unprecedented enthusiasm in the British population and became the blueprint for social legislation after the war.

The Beveridge report proved seminal also for the creation of the *French* 'Sécurité Sociale' (1946/48). But contrary to the case in the United Kingdom, the principle of administrative unification of the system met with strong resistance in France from various social groups which preferred to keep or create separate insurances meeting their specific needs and interests. In *Germany* only the term was received without changing the structure of the existing system. Moreover, a semantic distinction was made between the institutional (*Soziale Sicherung*) and the normative (*Soziale Sicherheit*) aspect of social security. In *Scandinavia* the term 'Social Security' was hardly received although the systems of social protection underwent substantial improvements.

Although many industrialized countries had developed measures of social protection against certain risks *before* World War II, their impact remained generally modest. Some groups, e.g. public servants or veterans, were privileged by public budgets, whereas industrial workers received only modest benefits out of contributions. With the exception of Scandinavia, the rural population remained mostly outside of existing systems. In most countries there existed a multitude of funds respectively designed for specific risks, groups, and localities. Repeated economic crises culminating in the Great Depression often impaired the reliability of expected benefits.

After the Second World War the new ideal of 'Social Security' was received differently in different countries. Almost everywhere coverage provided by existing institutions was extended or new institutions were created to protect all or most of the population against certain risks. Most countries assimilated the regulatory frameworks of different systems which covered the same risks, or integrated the systems. Finally, the risks covered by schemes underwent an international standardization. The Convention No. 102 of the ILO on 'Minimum Standards of Social Security' (1951) distinguishes nine forms of benefit: old-age benefits, survivor benefits, disability benefits, family allowances, unemployment benefits, sickness benefits, medical care, and maternity benefits. To be sure, most advanced industrial countries' social protection institutions do not correspond one-for-one with that list of benefits, but today they indeed do tend to cover all of these risks, albeit with varying degrees of coverage and benefits.

From a *structural perspective* the institutions of Social Security form a delimited, though not everywhere equally defined system. Issues of delimitation concern the forms of benefit (in cash or in kind), of administration (centralized or decentralized; public, parapublic, or private), and of entitlement (public provision, insurance, relief), as well as the groups included (privileged or underprivileged groups are sometimes excluded) and the risks covered (family allowances, for example, are often excluded from the concept).

3. Security and Social Security as Value-concepts

The administrative codification of Social Security as a complex of institutions which provide income replacement or benefits in kind should not obscure the fact that the success of the concept is due less to these institutions than to the value connotations of the concept. President Roosevelt evinced a keen sense of these connotations when he characterized social security as 'freedom from want' (6 January 1941), and when he described social security as a functional equivalent of earlier security deriving from 'the interdependence of members of families upon each other and of the families within a small community upon each other' (8 June 1934).

But what does security mean? And why did the concept become so prominent and pervasive in the most advanced industrial countries whose populations were more remote from serious want than any generations heretofore? Security relates to human attitudes towards the future and not the present. Insecurity does not mean an imminent danger but the possibility of future damages whose probability remains uncertain. The quest for security is a correlate to the growing complexity of society and the ensuing acceleration of social change. Thus security means not only protection, but also a subjective perception of the reliability of protection and the ensuing feeling of peace of mind. But in this psychological sense, security as self-assuredness can also mean the subjective capacity to bear the insecurity of risk. Hence on the individual level the concept of security remains highly ambivalent, and has indeed been abandoned by the discipline of psychology.

Security was also a longstanding pillar of political order which had already been incorporated into many constitutional documents and became transposed by Roosevelt into a new economic and social context. Just as government had the duty to care for external and internal security it was now to provide a framework to guarantee the means of existence for all, through full employment, social insurance, or other forms of benefits or services. 'Social Security' thus became the umbrella word for economic, social, and cultural rights as expressed in Article 22 of the United Nations Universal Declaration of Human Rights: 'Everyone as a member of society, has the right to social security and is entitled to realization through national effort and international co-operation and in accordance with the organization and resources of each state, of the economic, social and cultural rights indispensable for his dignity and the free development of his personality.' These rights are then specified as the right to work (Art. 23), the right to recreation from work (Art. 24), the right to health, well-being, and social protection, with special emphasis on mothers and children (Art. 25), the right to education (Art. 26), and the right to participation in culture, science, and the arts (Art. 27).

Thus Social Security became another name for welfare in its political-philosophical sense. It became a key term in the international evolution of the welfare state after World War II. It was adopted differently on the national level, however. In most countries the institutional aspect has long dominated. In the actual context of 'crisis' or 'reconstruction' of the welfare state 'social security' shows again its normative aspect: to what extent are politics free to dispose about political regulations which aim at stabilizing the future economic protection of the population or defined parts of it? 'Social Security' means the right of everybody who belongs to a certain commonwealth to be included into the basic provisions against destitution and want. And it suggests moreover that such provisions should take a coherent and clearly accessible form for everyone, so that people are not only legally protected but also have easy and reliable access to such protection.

4. Controversial Issues

4.1 General Problems

From the perspective of the Universal Declaration of Human Rights as well as William Beveridge or Pierre Laroque, the 'father of the French *sécurité sociale*,' Social Security was not restricted to publicly organized redistribution within populations at risk but also included policies of full employment and minimum wages. There was a clear awareness of the trade-off between full employment and the possibilities for sufficiently funding social protection. *Social security as the goal of guaranteeing to the entire population the opportunity to participate in the economic and cultural life of its society implies education and employment as primary means and compensatory social protection only as a 'second-best' solution*. The reification of 'Social Security' has obscured this original perspective and focuses public attention on processes of redistribution only. The current trend to substitute 'work for welfare' in the case of the able-bodied is quite in line with the original intention, provided that a decent minimum living standard is attainable at any rate. There is a strong divergence, however, with regard to whether the role of labor markets should be exclusive (as in the Anglo-Saxon world) or whether there should be a subsidiary public role (as in the Scandinavian and German-speaking countries) in employing the less productive parts of the active population.

From the perspective of the 'Chicago School' of economics, the existing systems of Social Security are overly costly, they disguise the relationship between an individual's contributions and benefits, and they benefit the middle classes more than those who are truly in need. A negative income tax would better serve the poor, they maintain. Such a 'rationalist' economic

perspective does not take into account the existing trust in public systems of reciprocal security in most countries. The 'deal' which pools the risks of the whole population in one scheme and thus makes basic economic security a matter of democratic politics seems to find high acceptance in most countries, although many people are aware of the likelihood of being among the net payers to the system.

Given the natural inequality of abilities and the social inequalites resulting from birth, education, labor markets, and competition in general, all political endeavors to foster social security—be it in the wider or in the narrower sense—face the problem of *distributive effects*. These relate not only to the amount of individual benefit and the forms of public financing but to the institutional arrangements as well. Strong political conflicts emerged in many countries around the issue of creating a single comprehensive system for certain provisions or for maintaining a fragmented system differentiating among classes of populations at risk. A somewhat related problem concerns the issue of securing only minimal standards or advanced standards of benefits through public provision. Finally, there is an obvious political struggle in all countries concerning the level of redistribution and who should bear the costs of social protection.

Comparative research has focused mainly on the explanation of national differences in institutional design and distributive effects. Path dependency upon early national approaches seems to be an important explanatory factor for these variations, beside political power relations and cultural orientations. An interesting finding is that systems which provide high levels of protection tend to protect minimal standards better as well.

4.2 Sectoral Aspects

Most political controversies do not arise with regard to the system of Social Security as a whole but concerning particular issues. As the functional organization of social protection is different from country to country, it remains difficult to generalize about this. One may identify some broad issues, however.

The paramount problem in the contemporary debate—particularly as a consequence of demographic aging—is *the security of income in old age*. A three-tiered system consisting of universal and state-administered protection of basic standards (first tier), employment-related (and often private) protection of advanced standards (second tier), and complementary forms of personal provision (third tier) seems to be a promising model to resist the demographic and economic challenges of the decades to come. In order to ensure income security in old age for all, some public regulation and supervision of the second and third tiers remains important, however.

The second complex of risks is related to *illness, industrial diseases, and long-term disability*. Here, benefits in cash and in kind are necessary, and coverage for these risks is organized quite differently in different countries. Essentially, there are two different institutional models: the National Health Service and protection through social insurance. Both systems face the problem of containing the explosion of costs which has resulted from the compounding of a wide range of medical, technological, demographic, and economic factors.

The third complex of Social Security is related to the *protection of the family*. Since child labor has been forbidden and education made compulsory, children are no longer an economic asset for their parents, but rather a strong source of social inequality. It is still contested to what extent parents and especially mothers should be supported by public means, beyond the generally accepted subsidy in cases of manifest poverty. The low birth rate in most European populations now gives more political weight to demands to improve the living conditions and social protection of persons who are raising children or nursing their permanently incapacitated parents.

The most contested part of Social Security is *public provision for the unemployed*. Here, the trade-offs with labor-market policies and full employment are obvious, and convictions vary about how to deal with this problem.

The issue of poverty cuts across such functional distinctions. The extent to which it emerges as a separate problem which must be addressed by specific measures depends to a large extent on the institutional approaches which a country has adopted in the above-mentioned realms of social security.

The issues discussed here are those generally attributed to the concept of social security. The concept of the welfare state covers additional services such as housing, education, and personal services.

See also: Insurance; Social Welfare Policy: Comparisons; Welfare; Welfare: Philosophical Aspects; Welfare Programs, Economics of; Welfare State; Welfare State, History of

Bibliography

Alber J 1982 *Vom Armenhaus zum Wohlfahrtsstaat. Analysen zur Entwicklung von Sozialversicherung in Westeuropa* (*From the Poor-House to the Welfare State. The Emergence of Social Insurance in Western Europe*). Campus, Frankfurt-am-Main, Germany

Arnold D R, Graetz M J, Munell A H (eds.) 1998 *Framing the Social Security Debate. Values, Politics, and Economics.* National Academy of Social Insurance, Washington, DC

Atkinson A B 1989 *Poverty and Social Security*. Harvester Wheatsheaf, New York

Baldwin S, Falkingham J (eds.) 1994 *Social Security and Social Change. New Challenges to the Beveridge Model.* Harvester Wheatsheaf, New York

Burns E M 1956 *Social Security and Public Policy*. McGraw-Hill, New York

Cohen W J, Friedman M 1972 *Social Security: Universal or Selective?* American Enterprise Institute, Washington, DC
Committee on Economic Security 1937 *Social Security in America. The Factual Background of the Social Security Act as Summarized from Staff Reports to the Committee on Economic Security*. Published by the Social Security Board. Government Printing Office, Washington, DC
Galant H 1955 *Histoire politique de la sécurité sociale francaise 1945–1952* (*A Political History of French Social Security 1945–1952*). Armand Colin, Paris
George V 1968 *Social Security: Beveridge and After*. Routledge & Kegan Paul, London
Heclo H 1974 *Modern Social Politics in Britain and Sweden: From Relief to Income Maintenance*. Yale University Press, New Haven, CT
Kaufmann F-X 1973 *Sicherheit als soziologisches und sozialpolitisches Problem. Untersuchungen zu einer Wertidee hochdifferenzierter Gesellschaften* (*Security as a Problem for Sociology and Social Policy. An Inquiry into a Universal Value of Differentiated Societies*), 2nd edn. Enke/Lucius u. Lucius, Stuttgart, Germany
Lowe R 1993 *The Welfare State in Britain Since 1945*. Macmillan, London
Noble C 1997 *Welfare As We Knew It. A Political History of the American Welfare State*. Oxford University Press, New York
Robbers G 1987 *Sicherheit als Menschenrecht* (*Security as a Human Right*). Nomo, Baden-Baden, Germany
Rubinow I M 1934 *The Quest for Security*. Henry Holt, New York
Social Insurance and Allied Services 1942 Report by Sir William Beveridge. HMSO, London
Weir M, Orloff A S, Skocpol T (eds.) 1988 *The Politics of Social Policy in the United States*. Princeton University Press, Princeton, NJ

F.-X. Kaufmann

Social Simulation: Computational Approaches

Computational social simulation involves the use of computer algorithms to model social processes. These algorithms require greater precision and logical rigor than natural language yet lack the generality of closed-form mathematical equations. The use of computer simulation by social scientists has increased in recent years, due to growing interest in modeling nonlinear dynamical processes using graphical interfaces that allow highly intuitive visual representations of the results. Social simulation has also undergone a qualitative change as the emphasis has shifted from prediction to exploration.

This article reviews three 'waves' of innovation: *dynamical systems, microsimulation, and adaptive agent* models (Gilbert and Troitzsch 1999). These three approaches will be summarized as follows.

(1) 1960s: Dynamical systems simulations were empirically grounded, inductive, holistic, and functionalist.

(2) 1970s: Microsimulation introduced the use of individuals as the units of analysis but retained the earlier emphasis on empirically based macro-level prediction.

(3) 1980s: Adaptive agent models revolutionized computational social science by simulating actors not factors (system attributes). These 'bottom-up' models explored interactions among purposive decision-makers.

Although these three 'waves' overlap (e.g., economists continue to use holistic simulation of productive factors, microsimulation was invented in 1957, and nascent agent-based models appeared in the late 1960s), most of the recent excitement in computational modeling in the social and behavioral sciences has centered on agent-based simulation. So too has much of the controversy.

1. Dynamical Systems

The first wave of computer simulations in social science occurred in the 1960s (see, for example, Cyert and March 1963). These studies used computers to simulate dynamical systems, such as control and feedback processes in organizations, industries, cities, and even global populations. The models typically consist of sets of differential equations that describe changes in system attributes as a function of other systemic changes. Applications included the flow of raw materials in a factory, inventory control in a warehouse, urban traffic, military supply lines, demographic changes in a world system, and ecological limits to growth (Forrester 1971; Meadows et al. 1974).

Models of dynamical systems are functional and holistic (Gilbert and Troitzsch 1999). *Functional* means that theoretical interest centers on system *equilibrium* or balance among interdependent input–output units. *Wholistic* means that the system is modeled as an irreducible and indivisible entity. Although these models allow nested organizational levels, at any given level, they assume that sets of related attributes (such as economic growth, suburban migration, and traffic congestion) are causally linked at the aggregate level.

2. Microsimulation

The second wave of computational modeling in the social sciences developed in the late 1970s. Known as microsimulation, it represents the first step in the progression from factors to actors. In striking contrast to the earlier wholistic approach,

> microsimulation is a 'bottom-up' strategy for modeling the interacting behavior of decision-makers (such as individuals, families, and firms) within a larger system. This modeling strategy utilizes data on representative samples of decision-

> makers, along with equations and algorithms representing behavioral processes, to simulate the evolution through time of each decision-maker, and hence of the entire population of decision-makers. (Caldwell 1997)

For example, Caldwell and Keister use CORSIM, a large-scale dynamic microsimulation model of the US population of individuals and families, to integrate individual and family-level wealth behavior with aggregate-level stratification outcomes. This microanalytic approach allows researchers to avoid a serious limitation in the earlier generation of simulations: population homogeneity. For example, the relationship between wealth and age may not be uniform across ethnic or religious subcultures, geographic regions, or birth cohorts. Complex interactions across subpopulations are lost when the focus is on the correlation among aggregate factors. This precludes the ability to predict the effects of policy changes that impact only certain groups.

Microsimulation solves this problem by making individuals, not populations, the unit of analysis. The database structure consists of individual records for a representative sample of a population, with a set of attributes measured at multiple points in time. These individuals are then 'aged' by updating their attributes, based on a set of empirically derived state-transition probabilities (e.g., an individual's move from work to retirement at age 65). These probabilities are then used to predict new states of each individual, and these are then aggregated to make population projections.

3. *Agent-based Models and Complex Adaptive Systems*

The third wave, agent-based models, attracted widespread interest beginning in the 1980s. These models extend the microanalytical approach by deriving global states from individual *behaviors*, not individual attributes. In microsimulation models, individuals are inert in two ways: (a) the state of an individual is a numerical function of other traits, and (b) individuals do not interact (Gilbert and Troitzsch 1999, p. 8). Agent-based models assume an individual has intentions or goals and makes choices that affect other agents whose choices in turn affect that individual. These models impose three key assumptions.

(a) Agents interact with little or no central authority or direction. Global patterns emerge from the bottom up, determined not by a centralized authority but by local interactions among autonomous decision-makers.

(b) Decision-makers are adaptive rather than optimizing, with decisions based on heuristics, not on calculations of the most efficient action (Holland 1995, p. 43). These heuristics include norms, habits, protocols, rituals, conventions, customs, and routines. They evolve at two levels, the individual and the population. Individual learning alters the probability distribution of rules competing for attention, through processes like reinforcement, Bayesian updating, or the back-propagation of error in artificial neural networks. Population learning alters the frequency distribution of rules competing for reproduction through processes of selection, imitation, and social influence (Latané 1996). Genetic algorithms (Holland 1995) are widely used to model adaptation at the population level.

(c) Decision-makers are strategically interdependent. Strategic interdependence means that the consequences of each agent's decisions depend in part on the choices of others. When strategically interdependent agents are also adaptive, the focal agent's decisions influence the behavior of other agents who in turn influence the focal agent, generating a complex adaptive system (Holland 1995, p. 10).

Thomas Schelling's (1971) 'neighborhood segregation' model was one of the earliest agent-based social simulations using *cellular automata*, a method invented by Von Neumann and Ulam in the 1940s. Consider a residential area that is highly segregated, such that the number of neighbors with different cultural markers (such as ethnicity) is at a minimum. If the aggregate pattern were assumed to reflect the attitudes of the constituent individuals, one might conclude from the distribution that the population was highly parochial and intolerant of diversity. Yet Schelling's simulation shows that this need not be the case. His 'tipping' model shows that highly segregated neighborhoods can form even in a population that prefers diversity. The aggregate pattern of segregation is an emergent property that does not reflect the underlying attitudes of the constituent individuals.

Another example of emergence is provided by Latané's (1996) 'social impact model.' Like Schelling, Latané also studies a cellular world populated by agents who live on a two-dimensional lattice. However, rather than moving, these agents adapt to those around them, based on a rule to mimic one's neighbors. From a random start, a population of mimics might be expected to converge inexorably on a single profile, leading to the conclusion that cultural diversity is imposed by factors that counteract the effects of conformist tendencies. However, the surprising result was that 'the system achieved stable diversity. The minority was able to survive, contrary to the belief that social influence inexorably leads to uniformity' (Latané 1996, p. 294). Other researchers, including Axelrod (1997), Carley (1991), and Kitts et al. (1999) have used computational models that couple homophily (likes attract) and social influence. If local similarity facilitates interaction which in turn facilitates imitation, it is indeed curious that the global outcome is diversity, not uniformity. However, this outcome depends on interactions that are locally transitive, as in a spatially distributed population.

Schelling and Latané modeled agents with identical and fixed behavioral rules. Many agent-based models

assume heterogeneous populations with the ability to learn new rules. The 'genetic algorithm' invented by John Holland is a simple and elegant way to model strategies that can progressively improve performance by building on partial solutions. Each strategy consists of a string of symbols that code behavioral instructions, analogous to a chromosome containing multiple genes. The string's instructions affect the agent's reproductive fitness and hence the probability that the strategy will propagate. Propagation occurs when two or more mated strategies recombine. If different rules are each effective, but in different ways, recombination allows them to create an entirely new strategy that may integrate the best abilities of each 'parent' and thus eventually displace the parent rules in the population of strategies.

Axelrod (1997, pp. 14–29] used a genetic algorithm to study the evolution of cooperation in a 'prisoner's dilemma,' a game in which choices that are individually rational aggregate into collective outcomes that everyone would prefer to avoid. The results showed that strategies based on reciprocity (such as 'tit for tat') tend to be very successful. The secret of their success is that they perform well against copies of themselves. Axelrod's result has been challenged by Nowak and Sigmund (1993), by Binmore (1998), and by Macy (1995). Working independently, these researchers found that 'tit for tat,' a strategy that teaches its partner a lesson, can be supplanted by 'Pavlov' (aka. 'win–stay, lose–shift' and 'tat for tit'), a strategy that learns (Binmore 1998). The ability to learn appears to be at least as important for emergent social order as the ability to teach.

Learning involves adaptation at the level of the individual rather than the population. Artificial neural networks are simple self-programmable devices that model agents who learn through reinforcement. Like genetic algorithms, neural nets have a biological analog, in this case, the nerve systems of living organisms. The device consists of a web of neuron-like units (or neurodes) that fire when triggered by impulses of sufficient strength, and in turn stimulate other units when fired. The magnitude of an impulse depends on the strength of the connection (or synapses) between the two neurodes. The network learns by modifying these path coefficients in response to environmental feedback about its performance. Neural nets have been used to study the evolution of religion (Bainbridge 1995), kin altruism (Parisi et al. 1995), the emergence of status inequality (Vakas-Duong and Reilly 1995), group dynamics (Nowak and Vallacher 1997), and social deviance (Kitts et al. 1999).

4. *Criticisms of Simulation*

Agent-based simulations have been criticized as unrealistic, tautological, and unreliable.

(a) Simulations, like mathematical models, tell us about a highly stylized and abstract world, not about the world in which we live. In particular, agent-based simulations rely on highly simplified models of rule-based human behavior that neglect the complexity of human cognition.

(b) Like mathematical models, simulations cannot tell us anything that we do not already know, given the assumptions built into the model and the inability of a computer to do anything other than execute the instructions given to it by the program.

(c) Unlike mathematical models, simulations are numerical and therefore cannot establish lawful regularities or generalizations.

The defense against these criticisms centers on the principles of complexity and emergence. How can simple agent-based models explain the rich complexity of social life? In complex systems, very simple rules of interaction can produce highly complex global patterns. 'Human beings,' Simon contends (1998, p. 53), 'viewed as behaving systems, are quite simple.' We follow rules, in the form of norms, conventions, protocols, moral and social habits, and heuristics. Although the rules may be quite simple, they can produce global patterns that may not be at all obvious and are very difficult to understand. Hence, 'the apparent complexity of our behavior is largely a reflection of the complexity of the environment,' including the complexity of interactions among strategically interdependent, adaptive agents. The simulation of artificial worlds allow us to explore the complexity of the social environment by removing the cognitive complexity (and idiosyncrasy) of constituent individuals.

4.1 *The Exploration of Artificial Worlds*

If Simon, Axelrod, and the complexity theorists are right, then the artificiality of agent-based models is a virtue, not a vice. When simulation is used to make predictions or for training personnel (e.g., flight simulators), the assumptions need to be highly realistic, which usually means they will also be highly complicated (Axelrod 1997, p. 5). 'But if the goal is to deepen our understanding of some fundamental process,' Axelrod continues, 'then simplicity of the assumptions is important and realistic representation of all the details of a particular setting is not.' As such, the purpose of these models is to generate hypotheses, not to test them (Prietula et al. 1998, p. xv).

Holland (1995, p.146) offers a classic example of the problem of building theories that too closely resemble actuality: Aristotle's mistaken conclusion that all bodies come to rest, based on observation of a world that happens to be infested with friction. Had these observations been made in a frictionless world, Aristotle would have come to the same conclusion

reached by Newton and formulated as the principle that bodies in motion persist in that motion unless perturbed. Newton avoided Aristotle's error by studying what happens in an artificial world in which friction had been assumed away. Ironically, 'Aristotle's model, though closer to everyday observations, clouded studies of the natural world for almost two millennia' (Holland 1995, p. 146). More generally, it is often necessary to step back from our world in order to see it more clearly.

For example, Schelling's 'neighborhood segregation' model made no effort to imitate actuality or to resemble an observed city. Rather, the 'residents' live in a highly abstract cellular world without real streets, rivers, train tracks, zoning laws, red-lining, housing markets, etc. This artificial world shows that all neighborhoods in which residents are able to move have an underlying tendency towards segregation even when residents are moderately tolerant. This hypothesis can then be tested in observed neighborhoods, but even if observations do not confirm the predicted segregation, this does not detract from the value of the simulation in suggesting empirical conditions that were not present in the model that might account for the discrepancy with the observed outcome.

In short, agent-based models 'have a role similar to mathematical theory, shearing away detail and illuminating crucial features in a rigorous context' (Holland 1995, p. 100). Nevertheless, Holland reminds us that we must be careful. Unlike laboratory experiments, computational 'thought experiments' are not constrained by physical actuality and thus 'can be as fanciful as desired or accidentally permitted.'

4.2 *Unwrapping and Emergence*

A second criticism of agent-based modeling addresses a problem known as 'unwrapping.' Unwrapping occurs 'when the 'solution' is explicitly built into the program [such that] the simulation reveals little that is new or unexpected' (Holland 1995, p. 137). Critics charge that all computer simulations share this limitation, due to the inability of computers to act in any way other than how they were programmed to behave.

This criticism overlooks the distinction between the micro-level programming of the agents and the macro-level patterns of interaction that emerge. These patterns are 'built in' from the outset yet the exercise is valuable because the logical implications of behavioral assumptions are not always readily apparent. Indeed, a properly designed computational experiment can yield highly counter-intuitive and surprising results. This theoretical possibility is based on the principle of *emergence* in complex systems. Biological examples of emergence include life that emerges from non-living organic compounds and intelligence and consciousness that emerges from dense networks of very simple switch-like neurons. Schelling's (1971) neighborhood model provides a compelling example from social life: segregation is an emergent property of social interaction that is not reducible to individual intolerance.

4.3 *Rationality and Adaptive Behavior*

A third criticism is that simulation models, unlike mathematical models, are numerical, not deductive, and therefore cannot be used to form generalizations. Worse still, simulations of evolutionary processes can be highly sensitive to the basin of attraction in which the parameters are arbitrarily initialized (Binmore 1998). However, game-theoretic mathematical models pay a high price for the ability to generate deductive conclusions: multiple equilibria that preclude a uniquely rational solution. Equilibrium selection requires constraints on the perfect rationality of the agents. Simon appreciates the paradox: 'Game theory's most valuable contribution has been to show that *rationality* is effectively undefinable when competitive actors have unlimited computational capabilities for outguessing each other, but that problem does not arise as acutely in a world, like the real world, of *bounded rationality*.' (1998, p. 38) But bounded rationality often makes analytical solutions mathematically intractable. 'When the agents use adaptive rather than optimizing strategies, deducing the consequences is often impossible: simulation becomes necessary.' (Axelrod 1997, p. 4) In sum, simulation requires sensitivity analysis and tests of robustness, but properly implemented, it not only offers a solution to the problem of equilibrium selection but can also tell us something about how evolutionary systems move from one equilibrium to another.

5. *Conclusion*

From a narrowly positivist perspective, agent-based artificial worlds might seem unrealistic, unfalsifiable, and unreliable, especially when contrasted with earlier approaches to simulation that were data driven, predictive, and highly realistic (e.g., flight simulators). However, agent-based models may be essential for the study of open and complex adaptive systems, whether social, psychological, or biological. When used as an exploratory tool, 'bottom-up' simulation does not test theory against observation. Rather, these thought experiments suggest possible mechanisms that may generate puzzling empirical patterns.

A key theme that runs through these puzzles is that it is often very difficult to recognize the underlying causal mechanisms using conventional data-analytic methods. As Holland concludes (1995, p. 195), 'I do not think we will understand morphogenesis, or the

emergence of organizations like Adam Smith's pin factory, or the richness of interactions in a tropical forest, without the help of such models.'

See also: Artificial Social Agents; Functional Equations in Behavioral and Social Sciences; Markov Models and Social Analysis; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Social Behavior (Emergent), Computer Models of; Social Network Models: Statistical

Bibliography

Axelrod R 1997 *The Complexity of Cooperation*. Princeton University Press, Princeton, NJ

Bainbridge W 1995 Neural network models of religious belief. *Sociological Perspectives* **38**(4): 483–96

Binmore K 1998 Axelrod's 'The Complexity of Cooperation'. *The Journal of Artificial Societies and Social Simulation* **1**: 1

Caldwell S 1997 *Dynamic Microsimulation and the Corsim 3.0 Model*. Strategic Forecasting, Ithaca, NY

Carley K 1991 A theory of group stability. *American Sociological Review* **56**: 331–54

Cyert R, March J G 1963 *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ

Forrester J W 1971 *World Dynamics*. MIT Press, Cambridge, MA

Gilbert N, Troitzsch K 1999 *Simulation for the Social Scientist*. Open University Press, Buckingham, UK

Holland J 1995 *Hidden Order: How Adaptation Builds Complexity*. Perseus, Reading, MA

Kitts J, Macy M, Flache A 1999 Structural learning: Attraction and conformity in task-oriented groups. *Computational and Mathematical Organization Theory* **5**: 129–45

Latané B 1996 Dynamic social impact: Robust predictions from simple theory. In: Hegselmann R, Mueller U, Troitzsch K (eds.) *Modeling and Simulation in the Social Sciences from a Philosophy of Science Point of View*. Kluwer Dorderecht, Boston, pp. 287–310

Macy M 1995 Natural selection and social learning in prisoner's dilemma: Co-adaptation with genetic algorithms and artificial neural networks. *Sociological Methods and Research* **25**: 103–37

Meadows D L, Behrens W W III, Meadows D H, Naill R F, Randers J, Zahn E K 1974 *The Dynamics of Growth in a Finite World*. MIT Press, Cambridge, MA

Nowak A, Vallacher R 1997 Computational social psychology: Cellular automata and neural network models of interpersonal dynamics. In: Read S, Miller L (eds.) *Connectionist Models of Social Reasoning and Social Behavior*. Erlbaum, Mahwah, NJ

Nowak M, Sigmund K 1993 A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoners' dilemma game. *Nature* **364**: 56–7

Parisi D, Cecconi F, Cerini A 1995 Kin-directed altruism and attachment behaviour in an evoking population of neural networks. In: Gilbert, Coute R C (eds.) *Artificial Societies*. University College London Press, London

Prietula M, Carley K, Gasser L 1998 *Simulating Organizations: Computational Models of Institutions and Groups*. MIT Press, Cambridge, MA

Schelling T 1971 Dynamic models of segregation. *Journal of Mathematical Sociology* **1**: 143–86

Simon H 1998 *The Sciences of the Artificial*. MIT Press, Cambridge, MA

Vakas-Duong D, Reilly K 1995 A system of IAC neural networks as the basis for self-organization in a sociological dynamical system simulation. *Behavioral Science* **40**: 275–303

M. W. Macy

Social Stratification

In all complex societies, the total stock of valued goods is distributed unequally, with the most privileged individuals and families enjoying a disproportionate share of income, power, and other valued resources. The term 'stratification system' refers to the complex of social institutions that generate observed inequalities of this sort. The key components of such systems are (a) the institutional processes that define certain types of goods as valuable and desirable; (b) the rules of allocation that distribute these goods across various positions in the division of labor (e.g., doctor, farmer, 'housewife'); and (c) the mobility mechanisms that link individuals to positions and thereby generate unequal control over valued resources. It follows that inequality is produced by two types of matching processes: The social positions in society are first matched to 'reward packages' of unequal value, and members of society are then allocated to the positions so defined and rewarded.

There are of course many types of rewards that come to be attached to social positions (see Table 1). Given this complexity, one might expect stratification scholars to adopt a multidimensional approach, with the objective being to describe and explain the full distribution of goods listed in Table 1. Although some scholars have indeed advocated an approach of this sort, most have instead opted to characterize stratification systems in terms of discrete classes or socioeconomic strata whose members are endowed with similar amounts and types of resources. The goal of stratification research has thus reduced to (a) describing the major forms of class inequality in human history, (b) specifying the structure of contemporary classes and strata, (c) modeling the processes by which individuals move between class and socioeconomic positions, and (d) examining the effects of race, ethnicity, and gender on such mobility processes.

1. Basic Concepts

The foregoing lines of inquiry cannot be adequately reviewed without first defining some of the core

Table 1
Types of assets, resources, and valued goods underlying stratification systems

Asset group	Selected examples	Relevant scholars
1. Economic	Ownership of land, farms, factories, professional practices, businesses, liquid assets, humans (i.e., slaves), labor power (e.g., serfs)	Karl Marx; Erik Wright
2. Political	Household authority (e.g., head of household); workplace authority (e.g., manager); party and societal authority (e.g., legislator); charismatic leader	Max Weber; Ralf Dahrendorf
3. Cultural	High-status consumption practices; 'good manners'; privileged life-style	Pierre Bourdieu; Paul DiMaggio
4. Social	Access to high-status social networks, social ties, associations and clubs, union memberships	W. Lloyd Warner; James Coleman
5. Honorific	Prestige; 'good reputation'; fame; deference and derogation; ethnic and religious purity	Edward Shils; Donald Treiman
6. Civil	Rights of property, contract, franchise, and membership in elective assemblies; freedom of association and speech	T. M. Marshall; Rogers Brubaker
7. Human	Skills; expertise; on-the-job training; experience; formal education; knowledge	Kaare Svalastoga; Gary Becker

concepts in the field. The following definitions are especially relevant for our purposes:

(a) The degree of *inequality* in a given reward or asset depends, of course, on its dispersion or concentration across the population. If some types of assets (e.g., civil rights) are distributed more equally than others (e.g., political power), then the level of inequality obviously cannot be adequately characterized with a single parameter.

(b) The *rigidity* of a stratification system is indexed by the continuity (over time) in the social standing of its members. The stratification system is said to be highly rigid, for example, if the current income, power, or prestige of individuals can be accurately predicted on the basis of their prior statuses or those of their parents.

(c) The stratification system rests on *ascriptive processes* to the extent that traits present at birth (e.g., sex, race, ethnicity, parental wealth) influence the subsequent social standing of individuals. In modern societies, ascription of all kinds is usually seen as undesirable or discriminatory, and much governmental policy is therefore directed toward fashioning a stratification system in which individuals acquire resources solely by virtue of their achievements.

(d) The degree of *status crystallization* is indexed by the correlations among the assets in Table 1. If these correlations are strong, then the same individuals (i.e., the 'upper class') will consistently appear at the top of all status hierarchies, while other individuals (i.e., the 'lower class') will consistently appear at the bottom of the stratification system.

These four variables can be used to characterize differences across societies in the underlying structure of stratification (see Table 2). As the following discussion shall reveal, there is much cross-societal variability not merely in the extent of inequality, but also in the underlying shape of the class structure and the processes by which classes are reproduced.

2. *Forms of Stratification*

It is useful to begin with the purely descriptive task of classifying the various types of stratification systems that have appeared in past and present societies. The first panel of Table 2 pertains to the 'primitive' tribal systems that dominated human society from the very beginning of human evolution until the Neolithic revolution of some 10,000 years ago. Although tribal societies have of course assumed various forms, the total size of the distributable surplus was in all cases quite limited; and this cap on the surplus placed corresponding limits on the overall level of economic inequality. Indeed, some observers have treated tribal societies as examples of 'primitive communism,' since the means of production (e.g., tools) were owned collectively and other types of property were typically distributed evenly among tribal members. Moreover, insofar as positions of power emerged (e.g., shamans), these were never inherited but instead were secured by demonstrating superior skills in the relevant tasks. While meritocratic criteria are often seen as prototypically modern, they were in fact present in incipient form at quite early stages of societal development, no doubt because the surplus was too small to permit the luxury of less adaptive forms of allocation.

With the emergence of agrarian forms of production, the economic surplus became large enough to support more complex and less meritocratic systems of stratification. The 'Asiatic mode,' which some commentators regard as a precursor of advanced agrarianism, is characterized by a poorly developed proprietary

Table 2
Basic parameters of stratification for eight ideal-typical systems

System center (1)	Principal assets (2)	Major strata or classes (3)	Inequality (4)	Rigidity (5)	Crystallization (6)	Justifying ideology (7)
A. Hunting and gathering society						
1. Tribalism	Human (hunting and magic skills)	Chiefs, shamans, and other tribe members	Low	Low	High	Meritocratic selection
B. Horticultural and agrarian society						
2. Asiatic mode	Political (i.e., incumbency of state office)	Office-holders and peasants	High	Medium	High	Tradition and religious doctrine
3. Feudalism	Economic (land and labor power)	Nobility, clergy, and commoners	High	Medium–high	High	Tradition and Roman Catholic doctrine
4. Slavery	Economic (human property)	Slave owners, slaves, 'free men'	High	Medium–high	High	Doctrine of natural and social inferiority (of slaves)
5. Caste society	Honorific and cultural (ethnic purity and 'pure' lifestyles)	Castes and subcastes	High	High	High	Tradition and Hindu religious doctrine
C. Industrial society						
6. Class system	Economic (means of production)	Capitalists and workers	Medium–high	Medium	High	Classical liberalism
7. State socialism	Political (party and workplace authority)	Managers and managed	Low–medium	Low–medium	High	Marxism and Leninism
8. 'Advanced' industrialism	Human (i.e., education, expertise)	Skill-based occupational groupings	Medium	Low–medium	Medium	Classical liberalism

class and a powerful state elite that extracted surplus agricultural production through rents and taxes (see line B2 in Table 2). This mode provides the conventional example of how a 'dictatorship of officialdom' can flourish in the absence of institutionalized private property. Whereas political assets were thus dominant in the Asiatic mode, the ruling class under Western feudalism was, by contrast, very much a propertied one. The distinctive feature of feudalism was that the nobility not only owned large estates or manors but also held legal title to the labor power of its serfs (see line B3). If a serf fled to the city, this was considered a form of theft: The serf was stealing that portion of his or her labor power owned by the lord. With this interpretation, the statuses of serf and slave differ only in degree, and slavery thereby constitutes a limiting case in which workers lose all control over their own labor power (see line B4; also see *Slavery as Social Institution*).

The historical record makes it clear that agrarian stratification systems were not always based on strictly hereditary forms of social closure. To be sure, the era of classical feudalism (i.e., post-twelfth century) was characterized by a rigid stratification of classes, but there was far greater permeability during the period prior to the institutionalization of the manorial system and the associated transformation of the nobility into a legal class. The most extreme example of agrarian closure can of course be found in caste societies (see line B5). The Indian caste system, for example, is based on (a) a hierarchy of status groupings (i.e., castes) that are ranked by ethnic purity, wealth, and access to goods or services; (b) a corresponding set of 'closure rules' that restrict all forms of intercaste marriage or mobility and thereby make caste membership both hereditary and permanent; (c) a high degree of physical and occupational segregation enforced by elaborate rules and rituals governing intercaste contact; and (d) a justifying ideology (i.e., Hinduism) that induces the population to regard such extreme forms of inequality as legitimate and appropriate. What makes this system so distinctive, then, is not merely its well-developed closure rules but also the fundamentally honorific (and noneconomic) character of the underlying social hierarchy.

The defining feature of the industrial era (see panel C) has been the emergence of egalitarian ideologies and the consequent 'delegitimation' of the extreme forms of stratification found in caste, feudal, and slave systems. This can be seen, for example, in the European revolutions of the eighteenth and nineteenth centuries that pitted the egalitarian ideals of the Enlightenment against the privileges of rank and the political power of the nobility. In the end, these struggles eliminated the last residue of feudal privilege, but they also made new types of inequality and stratification possible. Under the class system that ultimately emerged (see line C6), the estates of the feudal era were replaced by purely economic groups (i.e., 'classes'), and closure rules based on heredity were likewise supplanted by (formally) meritocratic processes. The resulting classes were neither legal entities nor closed status groupings, and the associated class-based inequalities could therefore be represented and justified as the natural outcome of competition among individuals with differing abilities, motivation, or moral character (i.e., 'classical liberalism'). As indicated in line C6, the class structure of early industrialism had a clear economic base, so much so that Marx ([1894] 1972) defined classes in terms of their relationship to the means of economic production. The precise contours of the industrial class structure are nonetheless a matter of continuing debate (see below); for example, a simple Marxian model focuses on the cleavage between capitalists and workers, while more elaborate Marxian and neo-Marxian models identify additional intervening or 'contradictory' classes (e.g., Wright 1997), and yet other (non-Marxian) approaches represent the class structure as a continuous gradation of income, prestige, or socioeconomic status.

Whatever the relative merits of these models might be, the ideology underlying the socialist revolutions of the nineteenth and twentieth centuries was of course explicitly Marxist (see line C7). The intellectual heritage of these revolutions and their legitimating ideologies can again be traced to the Enlightenment, but the rhetoric of equality that emerged in this period was now directed against the economic power of the capitalist class rather than the status and honorific privileges of the nobility. The evidence from Eastern Europe and elsewhere suggests that these egalitarian ideals were only partially realized. In the immediate postrevolutionary period, factories and farms were indeed collectivized or socialized, and various fiscal and economic reforms were instituted for the express purpose of reducing income inequality and wage differentials among manual and nonmanual workers. Although these egalitarian policies were subsequently weakened through the reform efforts of Stalin and others, inequality on the scale of pre-revolutionary society was never re-established among rank-and-file workers (see Lenski 2001). There nonetheless remained substantial inequalities in power and authority; most notably, the socialization of production did not have the intended effect of empowering workers, as the capitalist class was replaced by a 'new class' of party officials and managers who continued to control the means of production and to allocate the resulting social surplus. This class has been variously identified with intellectuals or intelligentsia, bureaucrats or managers, and party officials or appointees (see Gouldner 1979). Regardless of the formulation adopted, the presumption is that the working class ultimately lost out in contemporary socialist revolutions, just as it did in the so-called bourgeois revolutions of the eighteenth and nineteenth centuries (e.g., see *Postsocialist Societies*).

Whereas the means of production were socialized in the revolutions of Eastern Europe and the former USSR, the capitalist class remained intact throughout the process of industrialization in the West, even as ownership and control separated and a distinct managerial class emerged (see Dahrendorf 1959). The capitalist class may nonetheless be weakened by the structural changes of postindustrialism, with the most important of these being the rise of a service economy and the consequent emergence of technical expertise, educational degrees, and training certificates as new forms of property (see line C8). By this formulation, a dominant class of cultural elites may be emerging in the West, much as the transition to state socialism (allegedly) generated a new class of intellectuals in the East. This is not to suggest that all theorists of advanced industrialism posit a grand divide between the cultural elite and an undifferentiated working mass. In fact, some commentators (e.g., Dahrendorf 1959) have argued that skill-based cleavages are crystallizing throughout the occupational structure, with the result being a finely differentiated class system made up of discrete occupations (Grusky and Sørensen 1998) or a continuous gradation of socioeconomic status (e.g., Hauser and Warren 1997).

3. *The Structure of Contemporary Stratification*

The history of stratification theory is in large part a history of such debates about the emerging structure of contemporary inequality. Moreso than usual, political and intellectual goals are conflated in this literature and, accordingly, the relevant research is infused with much scholarly contention. These debates are complex and wide-ranging, but it will suffice to review four approaches that have proven to be especially popular.

3.1 *Reductionism*

Among contemporary interpreters of inequality, the prevailing approach has long been to claim that only one of the 'asset groups' in Table 1 is truly fundamental, whereas all others are somehow epiphenomenal. There are nearly as many claims of this sort as there are dimensions in Table 1. To be sure, Marx is most commonly criticized (with some justification) for placing 'almost exclusive emphasis on economic factors as determinants of social class' (Lipset 1968, p. 300), but in fact much of what passes for stratification theorizing amounts to reductionism of one form or another. Among non-Marxist scholars, inequalities in honor or power are frequently regarded as the most fundamental sources of class formation, while the distribution of economic assets is seen as purely secondary. For example, Dahrendorf (1959) argues that 'differential authority in associations is the ultimate 'cause' of the formation of conflict groups' (p. 172), while Shils (1968) suggests that 'without the intervention of considerations of deference position the ... inequalities in the distribution of any particular facility or reward would not be grouped into a relatively small number of vaguely bounded strata' (p. 130). These extreme forms of reductionism have been less popular of late; indeed, even neo-Marxian scholars now typically recognize several stratification dimensions, with the social classes of interest then being defined as particular combinations of scores on the selected variables (e.g., Wright 1997).

3.2 *Synthesizing Approaches*

There is an equally long tradition of research based on synthetic measures that simultaneously tap a wide range of assets and resources. As noted above, many of the rewards in Table 1 (e.g., income, honor) are principally allocated through the jobs or social roles that individuals occupy, and one can therefore measure the standing of individuals by classifying them in terms of their social positions. In this context, Parkin (1971) has referred to the occupational structure as the 'backbone of the entire reward system of modern Western society' (p. 18), while Hauser and Featherman (1977) argue that studies 'framed in terms of occupational mobility … yield information simultaneously (albeit, indirectly) on status power, economic power, and political power' (p. 4). The most recent representatives of this position, Grusky and Sørensen (1998), have argued that detailed occupations are not only the main conduits through which valued goods are disbursed but are also deeply institutionalized categories that are salient to workers, constitute meaningful social communities and reference groups, and serve as fundamental units of collective action. Although occupations continue, then, to be the preferred classificatory tool within this tradition, other scholars have pursued the same synthesizing objective by simply asking community members to locate their peers in a hierarchy of social strata (e.g., Warner et al. 1949). Under the latter approach, a synthetic classification is no longer secured by ranking and sorting occupations in terms of the bundles of rewards attached to them, but rather by passing the raw data of inequality through the fulcrum of individual judgment.

3.3 *Aggregation Exercises*

Regardless of whether a reductionist or synthesizing approach is taken, most scholars adopt the final simplifying step of defining a relatively small number of discrete classes (e.g., see *Class: Social*). For example, Parkin (1971) argues for six occupational classes with the principal 'cleavage falling between the manual and non-manual categories' (p. 25), whereas

Dahrendorf (1959) argues for a two-class solution with a 'clear line drawn between those who participate in the exercise [of authority] ... and those who are subject to the authoritative commands of others' (p. 170). While close variants of the Parkin scheme continue to be used, the emerging convention among quantitative stratification scholars is to apply either the 12-category neo-Marxian scheme fashioned by Wright (1997) or the 11-category neo-Weberian scheme devised by Erikson and Goldthorpe (1994). At the same time, new classification schemes continue to be regularly proposed, with the impetus for such efforts typically being the continuing expansion of the service sector and the associated growth of contingent work relations (e.g., Esping-Andersen 1999). The question that necessarily arises for all contemporary schemes is whether the constituent categories are purely nominal entities or are truly meaningful to the individuals involved. If the categories are intended to be meaningful, one would expect class members not only to be aware of their membership (i.e., 'class awareness') but also to identify with their class (i.e., 'class identification') and occasionally act in its behalf (i.e., 'class action').

3.4 Gradationalism

The prior approaches involve mapping individuals or families into mutually exclusive and exhaustive categories. By contrast, the implicit claim underlying gradational approaches is that such 'classes' are largely the construction of overzealous sociologists, and that the underlying structure of modern stratification can, in fact, be more closely approximated with gradational measures of income, status, or prestige. Although there is some sociological precedent for treating income as a gradational indicator of class, most sociologists opt for either prestige scales based on popular evaluations of occupational standing or socioeconomic scales constructed as weighted averages of occupational income and education. In recent years, such gradationalism has been subjected to criticism on various fronts, with the main objections being that (a) the fluidity and openness of the stratification system is overstated by conventional prestige and socioeconomic scales (Hauser and Warren 1997); (b) the desirability of jobs is best indexed by directly measuring such job-level attributes as earnings, promotion opportunities, complexity, or autonomy (Jencks et al. 1988); and (c) the convention of converting discrete occupations into scales strips away precisely those *gemeinschaftlich* features of discrete occupations that make them analytically so powerful (Grusky and Sørensen 1998). There is of course no guarantee that any of these criticisms will take hold; indeed, because socioeconomic scales have dominated sociological research since the 1960s, inertial forces are quite strong and will not be easily overcome.

4. Generating Stratification

The foregoing debates may be seen, then, as disputes over which data reduction techniques best reveal the fundamental features of the distribution of rewards. If one or more of these simplifying assumptions is applied and a mapping of inequality is thereby developed, it becomes possible to examine the flow of individuals between the various categories of this mapping. The language of stratification theory thus makes a sharp distinction between the distribution of social rewards and the distribution of opportunities for securing these rewards. The latter distribution has come to determine popular judgments about the legitimacy of stratification; that is, substantial inequalities in power, wealth, or honor are typically seen as tolerable (and even desirable) provided that the opportunities for securing these social goods are distributed equally. Whatever the wisdom of this popular logic might be, stratification researchers have long sought to explore its factual underpinnings by monitoring and describing the structure of mobility chances. The relevant literature is vast, but of course broad classes of inquiry can be distinguished, as indicated below.

4.1 Mobility Analyses

The conventional starting point for mobility scholars has been to analyze bivariate 'mobility tables' formed by cross-classifying the class origins and destinations of individuals (e.g., see *Social Mobility, History of*). The tables constructed can be used to estimate densities of inheritance, to map the social distances between classes and their constituent occupations, and to examine differences across subpopulations in the amount and patterning of fluidity and opportunity. Moreover, when comparable mobility tables are assembled from several countries, it becomes possible to address fundamental debates about the underlying contours of cross-national variation in stratification systems (e.g., Erikson and Goldthorpe 1994). This long-standing line of analysis, while still underway, has nonetheless declined of late, perhaps because past research has been so definitive as to undercut further efforts. The focus has thus shifted to studies of income mobility, with the twofold impetus for this development being (a) concerns that poverty may be increasingly difficult to escape and that a permanent underclass may be forming, and (b) the obverse hypothesis that growing income inequality may be counterbalanced by increases in the rate of mobility between income groups.

4.2 The Process of Stratification

It is by now a sociological truism that Blau and Duncan (1967) and their colleagues (e.g., Sewell et al.

1969) revolutionized the field with their formal 'path models' of stratification. These models were intended to represent, if only partially, the processes by which background advantages could be converted into socioeconomic status through the mediating variables of schooling, aspirations, and parental encouragement. Under formulations of this kind, the main sociological objective was to show that socioeconomic outcomes were structured not only by ability and family origins, but also by various intervening variables (e.g., schooling) that were themselves only partly determined by origins and other ascriptive forces. The picture of modern stratification that emerged suggested that market outcomes depend in large part on unmeasured career contingencies (i.e., 'individual luck') rather than influences of a more structural sort (Jencks et al. 1972). This line of research, which fell out of favor by the mid-1980s, has been recently reinvigorated as stratification scholars react to the controversial claim (i.e., Herrnstein and Murray 1994) that inherited intelligence is increasingly determinative of stratification outcomes (e.g., Fischer et al. 1996). In a related development, contemporary scholars have also turned their attention to the effects of family structure on attainment, given that new nontraditional family arrangements (e.g., female-headed households) may in some cases reduce the influence of biological parents and otherwise complicate the reproduction of class.

4.3 *Structural Analyses*

The preceding models are frequently criticized for failing to attend to the social structural constraints that operate on the stratification process independently of individual-level traits. The structuralist accounts that ultimately emerged from these critiques initially amounted, in most cases, to refurbished versions of dual economy and market segmentation models that were introduced and popularized many decades ago by institutional economists. When these models were redeployed by sociologists in the early 1980s, the usual objective was to demonstrate that women and minorities were disadvantaged not merely by virtue of deficient human capital investments (e.g., inadequate schooling and experience), but also by their consignment to secondary labor markets that, on average, paid out lower wages and offered fewer opportunities for promotion or advancement. At the same time, more deeply sociological forms of structuralism have also appeared, both in the form of (a) mesolevel accounts of the effects of social networks and 'social capital' on attainment, and (b) macrolevel accounts of the effects of institutional context (e.g., welfare regimes) on mobility processes and outcomes.

The history of these research traditions is arguably marked more by statistical and methodological signposts than substantive ones. Although there is, then, no grand theory that unifies seemingly disparate models, the field has long relied on middle-range theorizing addressing such issues as the forces making for discrimination, the processes through which class-based advantage is reproduced, and the effects of industrialism, capitalism, and socialism on mobility and attainment. The main contenders, at present, for a 'grand theory' of mobility are various forms of rational action analysis that allow middle-range accounts to be recast in terms of individual-level incentives and purposive behavior (e.g., see *Rational Choice Theory in Sociology*). Indeed, just as the assumption of utility maximization underlies labor economics, so too a theory of purposive behavior might ultimately organize much, albeit not all, sociological theory on social mobility and attainment.

5. *Ascriptive Solidarities*

To this point, the discussion has focused on class and socioeconomic inequalities, while considerations of race, ethnicity, and gender have been brought in only indirectly. This omission reflects, however improperly, the historical development of the field; that is, status groups were once regarded as secondary forms of affiliation, whereas class-based ties were seen as more fundamental and decisive determinants of social and political action. The first step in the intellectual breakdown of this approach was the fashioning of multidimensional models of stratification (e.g., Lipset and Bendix 1959) that moved beyond a strict class analytic emphasis on economic inequalities. The early multidimensionalists thus emphasized that social behavior could only be understood by taking into account all status group memberships and the complex ways in which these interacted with one another. There was much concern, in particular, that 'inconsistent' statuses (e.g., upper class and African American) might generate personal stress that would ultimately be resolved through extremist political action.

Although this old variant of multidimensionalism has fallen out of fashion, new variants have been pursued on two intellectual fronts, neither of which similarly emphasize the stressfulness of multiple or inconsistent statuses. In the postmodernist variant, the essentialism of conventional class analytic theorizing is again rejected forcefully, with the claim being that class affiliations are by no means fundamental and that various nonclass statuses are more salient in many situations. Under some postmodern formulations, even the categories of race, ethnicity, and gender have no privileged position, and instead the presumption is that individuals are simply a congeries of manifold, situationally invoked statuses. As the British sociologist Saunders (1989) puts it, 'On holiday in Spain we feel British, waiting for a child outside the school gates we are parents, shopping in Marks and Spencer we are consumers, and answering questions, framed

by sociologists with class on the brain, we are working class' (pp. 4–5).

This approach, which has not yet informed much empirical research, might be usefully contrasted to the 'intersectionist' accounts of Collins (1990), where the categories of race, class, and gender are seen as master statuses that form a new holy trinity of stratification analysis. The theoretical framework motivating this approach is not well-developed, but the implicit claim seems to be that the most fundamental subcultures in the modern context are defined by the *intersection* of race, class, and gender categories (e.g., black working-class women, white middle-class men). The social spaces defined by particular combinations of these statuses thus shape the experiences, lifestyles, and life chances of individuals and define the settings in which interests typically emerge. The obvious effect of this approach is to invert the traditional post-Weberian perspective on status groupings; that is, whereas orthodox multidimensionalists described the stress experienced by individuals in inconsistent statuses, these new multidimensionalists emphasize the shared interests and cultures generated within commonly encountered status sets.

Among stratification scholars, the attention paid to issues of race, ethnicity, and gender has thus burgeoned of late, and entirely new literatures within these subfields have of course emerged (e.g., see *Feminist Political Theory and Political Science*, *Ethnic Conflicts*, *Racial Relations*). In organizing these literatures, one might usefully distinguish between (a) macro-level research addressing the structure of ascriptive solidarities and their relationship to class formation, and (b) attainment research exploring the effects of race, ethnicity, and gender on individual life chances. At the macrolevel, scholars examine such issues as the social processes by which ascriptive categories (e.g., 'white,' 'black') are constructed, the sources and causes of ethnic conflict and solidarity, and the relationship between patriarchy, racism, and class-based forms of organization. The microlevel research tradition emphasizes, by contrast, such topics as the size and causes of the gender gap in income, the sources of occupational sex segregation, and the effectiveness of various social interventions (e.g., affirmative action, comparable worth initiatives) in reducing ascriptive inequalities.

6. *The Future of Stratification*

It is instructive to conclude by contrasting structural and cultural accounts of stratificatory change. Although there are many variants of structuralism, a common starting point is the claim that human and political capital are replacing economic capital as the principal stratifying forces in advanced industrial society. In the most extreme versions of this claim, the old class of moneyed capital is represented as a dying force, and a new class of intellectuals, managers, or party bureaucrats is assumed to be on the road to power (e.g., Gouldner 1979). There is of course much criticism of 'new class' interpretations of this sort. The (orthodox) Marxist stance is that 'news of the demise of the capitalist class is ... somewhat premature' (Zeitlin 1982, p. 216), whereas the contrasting position taken by Bell (1973) is that neither the old capitalist class nor the so-called new class will have unfettered power in the postindustrial future. To be sure, there is widespread agreement among postindustrial theorists that human capital is becoming a dominant form of property, yet this need not imply that 'the amorphous bloc designated as the knowledge stratum has sufficient community of interest to form a class' (Bell 1987, p. 464).

As is well-known, Bell (1973) also argues that human capital (e.g., education) will become the main determinant of life chances, if only because the expansion of the professional and technical sectors serves to upgrade job skills and thus requires a well-trained workforce. Although the returns to education are indeed increasing as predicted, the occupational structure is not upgrading quite as straightforwardly as Bell (1973) suggested; and various 'pessimistic versions' of postindustrialism have accordingly emerged. In the US variant of such pessimism, the main concern is that postindustrialism leads to a 'declining middle' and consequent polarization, as manufacturing jobs are either rendered technologically obsolete or exported to less developed countries where labor costs are lower. These losses may be compensated by the predicted growth in the service sector, yet the types of service jobs that have emerged are quite often low skill, routinized, and accordingly less desirable than Bell imagined. In Europe, the same low-skill service jobs are less commonly found, with the resulting occupational structure more closely approximating the highly professionalized world that Bell envisaged. The European pessimists are nonetheless troubled by the rise of mass unemployment and the associated emergence of 'outsider classes' that bear disproportionately the burden of unemployment. In both the European and US cases, the less-skilled classes are therefore losing out in the market, either by virtue of unemployment and exclusion (i.e., Europe) or low pay and poor prospects for advancement (i.e., the US). The new pessimists thus anticipate a 'resurgent proletarian underclass and, in its wake, a menacing set of new class correlates' (Esping-Andersen 1999, p. 95).

The driving force behind these accounts is, of course, structural change of the sort conventionally described by such terms as industrialism, post-industrialism, and post-Fordism. By contrast, cultural accounts of change tend to de-emphasize these forces or to cast them as epiphenomenal, with the focus thus shifting to the independent role of ideologies, social

movements, and cultural practices in changing stratification forms. This approach underlies, for example, all forms of postmodernism that seek to represent 'new social movements' (e.g., feminism, ethnic and peace movements, environmentalism) as the vanguard force behind future stratificatory change. As argued by Beck (1999), the labor movement can be seen as a fading enterprise rooted in the old conflicts of the workplace and industrial capitalism, whereas new social movements provide a more appealing call for collective action by virtue of their emphasis on issues of lifestyle, personal identity, and normative change. Under this formulation, the proletariat is stripped of its privileged status as a universal class, and cultural movements emerge as an alternative force 'shaping the future of modern societies' (Haferkamp and Smelser 1992, p. 17). Although no self-respecting postmodernist will offer up a fresh 'grand narrative' to replace that of discredited Marxism, new social movements are nonetheless represented within this subtradition as a potential source of change, albeit one that plays out in fundamentally unpredictable ways.

The final, and more prosaic, question that might be posed is whether changes of the preceding sort presage a general decline in the field of stratification itself. It could well be argued that Marxian and neo-Marxian models of class will decline in popularity with the rise of postmodern stratification systems and the associated uncoupling of class from lifestyles, consumption patterns, and political behavior. This line of reasoning is not without merit, but it is worth noting that (a) past predictions of this sort have generated protracted debates that, if anything, have re-energized the field; (b) the massive facts of economic, political, and honorific inequality will still be with us even if narrowly conceived models of class ultimately lose out in such debates; and (c) the continuing diffusion of egalitarian values suggests that all departures from equality, no matter how small, will be the object of considerable interest among sociologists and the lay public alike. The latter sensitivity to all things unequal bodes well for the future of the field even in the (unlikely) event of a long-term secular movement toward diminishing inequality.

See also: Class: Social; Democracy; Democratic Theory; Egalitarianism: Political; Egalitarianism, Sociology of; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Income Distribution; Income Distribution: Demographic Aspects; Inequality; Inequality: Comparative Aspects; Mobility: Social; Poverty: Measurement and Analysis; Poverty, Sociology of; Racism, Sociology of; Social Inequality in History (Stratification and Classes); Social Mobility, History of; Socialism; Socialism: Historical Aspects; Urban Geography; Urban Poverty in Neighborhoods; Wealth Distribution

Bibliography

Beck U 1999 *World Risk Society*. Polity Press, Cambridge, UK

Bell D 1973 *The Coming of Post-industrial Society*. Basic Books, New York

Bell D 1987 The new class: A muddled concept. In: Heller C S (ed.) *Structured Social Inequality*, 2nd edn. Macmillan, New York, pp. 455–68

Blau P M, Duncan O D 1967 *The American Occupational Structure*. Wiley, New York

Collins P H 1990 *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Unwin Hyman, Boston

Dahrendorf R 1959 *Class and Class Conflict in Industrial Society*. Stanford University Press, Stanford, CA

Erikson R, Goldthorpe J H 1994 *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Clarendon Press, Oxford, UK

Esping-Andersen G 1999 *Social Foundations of Postindustrial Economies*. Oxford University Press, Oxford

Fischer C S, Hout M, Sánchez Jankowski M, Lucas S R, Swidler A, Voss K 1996 *Inequality by Design: Cracking the Bell Curve Myth*. Princeton University Press, Princeton, NJ

Gouldner A 1979 *The Future of Intellectuals and the Rise of the New Class*. Seabury, New York

Grusky D B, Sørensen J B 1998 Can class analysis be salvaged? *American Journal of Sociology* **103**: 1187–234

Haferkamp H, Smelser N J 1992 Introduction. In: Haferkamp H, Smelser N J (eds.) *Social Change and Modernity*. University of California Press, Berkeley, CA, pp. 1–23

Hauser R M, Featherman D L 1977 *The Process of Stratification: Trends and Analyses*. Academic Press, New York

Hauser R M, Warren J R 1997 Socioeconomic indexes of occupational status: A review, update, and critique. In: Raftery A (ed.) *Sociological Methodology, 1997*. Blackwell, Cambridge, MA, pp. 177–298

Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York

Jencks C, Perman L, Rainwater L 1988 What is a good job? A new measure of labor-market success. *American Journal of Sociology* **93**: 1322–57

Jencks C, Smith M, Acland H, Bane M J, Cohen D, Gintis H, Heyns B, Michelson S 1972 *Inequality: A Reassessment of the Effect of Family and Schooling in America*. Basic Books, New York

Lenski G E 2001 New light on old issues: The relevance of 'really existing socialist societies' for stratification theory. In: Grusky D B (ed.) *Social Stratification: Class, Race, and Gender in Sociological Perspective*, 2nd edn. Westview, Boulder, CO, 77–84

Lipset S M 1968 Social class. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, pp. 296–316

Lipset S M, Bendix R 1959 *Social Mobility in Industrial Society*. University of California Press, Berkeley, CA

Marx K [1894] 1972 *Capital*, 3 vols. Lawrence and Wishart, London

Parkin F 1971 *Class Inequality and Political Order: Social Stratification in Capitalist and Communist Societies*. Praeger, New York

Saunders P R 1989 Left write in sociology. *Network* **44**: 3–4

Sewell W H, Haller A O, Portes A 1969 The educational and early occupational attainment process. *American Sociological Review* **34**: 82–92

Shils E 1968 Deference. In: Jackson J A (ed.) *Social Stratification*. Cambridge University Press, New York, pp. 104–32
Warner W L, Meeker M, Eels K 1949 *Social Class in America*. Science Research Associates, Chicago
Wright E O 1997 *Class Counts: Comparative Studies in Class Analysis*. Cambridge University Press, Cambridge, UK
Zeitlin M 1982 Corporate ownership and control: The large corporation and the capitalist class. In: Giddens A, Held D (eds.) *Classes, Power, and Conflict*. University of California Press, Berkeley, CA, pp. 196–223

D. B. Grusky

Social Support and Health

The concept of social support as an important factor in health and illness is not new. 'Friendship is a basic human need along with food, shelter, and clothing. We naturally desire to love other human beings and to be loved by them. A totally loveless life—a life without friends of any sort—is a life deprived of much needed good.' These thoughts were expressed by Aristotle, 350 years BC (Alcalay 1983). During and toward the end of the Middle Ages the concept was even addressed as a possible remedy for illness. In 1599 Paracelsus, a pharmacist and natural scientist, actually prescribed 'love as the best possible cure for several diseases.' That love and affection 'influence life and death in both animals and humans' was further observed by a Russian ethologist (Kropotkin 1908).

In this article some of the empirical evidence that 'love and affection influence life and death' will be reviewed. The distinctions between social networks and social support are explained; the definition, concepts, and functions of social support are further described; and their impact on health and disease are demonstrated and discussed. With a focus on coronary heart disease (CHD), the most common cause of death among men and women in the developed world, the hazards of lack of social support are disentangled, and differences and similarities in men and women are discussed. Finally, implications for therapeutic interventions are considered.

Systematic and conclusive empirical studies of social support and health, however, did not begin until the 1970s, when this field emerged as a new area of research, making a bold link between people's social ties and social networks on the one hand and population morbidity and mortality on the other. In the pioneering work by Berkman and Syme on residents of Alameda County, California, modern epidemiological tools were applied to demonstrate what had long been sensed to be true: social ties are protective against ill health.

Soon a number of population-based studies mainly from the USA and Scandinavia confirmed the findings of Berkman and Syme. These findings were systematically interpreted and discussed by House et al. (1988) who concluded that the impact of lack of social ties was similar to that of smoking. With such emphasis and visibility of the concept, a new research field was established.

A general definition of social support was agreed upon, which was broad enough to incorporate most of the research traditions in the field: 'Social support is the resource provided by other persons. By viewing social support in terms of resources—potentially useful information or things—we allow for the possibility that support may have negative as well as positive effects on health and well-being' (Cohen and Syme 1985).

Soon, however, researchers began to ask why and how social ties might influence mortality and morbidity; in particular investigators were focusing on the function and contents of social ties and their assessment (Cohen and Syme 1985). On the one hand, large epidemiological studies of populations typically would find that mortality increased with a decreasing number and frequency of social contacts. On the other hand, the more sophisticated measures of social context, social anchorage, social environment, and social support were limited to specific indicators of disease such as blood pressure in smaller selected groups of persons or patients. Thus two types of research traditions developed within this area—social network epidemiology (see *Social Integration, Social Networks, and Health*) and social support psychology. Attempts to unite these traditions have not always been successful. Where the first type of research mostly examined all causes of mortality or general well-being as end-points, the latter has looked more closely at specific disease end-points, such as cardiovascular disease. The latter has become a model for research on social support and disease, and will be the focus here, because heart diseases are common, develop gradually, and are found in all populations.

Coronary heart disease (CHD) is the most common of cardiovascular diseases and the most common cause of death in highly developed and industrialized countries, and it is rapidly increasing in developing nations of the third world. In many countries it is also the most common cause of hospitalizations and an important factor in work absenteeism. The clinical manifestations of CHD are dependent on the extent of underlying atherosclerotic changes of the coronary arteries. With more advanced atherosclerosis, the risk of an acute myocardial infarction (AMI) or sudden cardiac death (SCD) is increased. The occurrence of both atherosclerosis and clinical manifestations of CHD increases with increasing age, in men as in women.

Below the age of 65, clinical manifestations of CHD, including myocardial infarction and sudden death are rare in women. In particular before menopause at age 50, women seem to be protected. Women are also relatively spared in the following decade, between 50

and 60 years of age. By the age of 60, sex differences are beginning to vanish. This may be one of the reasons why research on CHD has been so remarkably focused on men in their middle ages, under age 65. Most research on coronary risk factors and factors that are protective of CHD has been carried out on middle-aged men. Therefore knowledge about heart disease in women is less evidence-based than that of men. This is particularly true as concerns the psychosocial influences on heart disease and its risk factors.

1. *The Social Environment and the Heart*

Lifestyle and daily habits—the way we live in the broadest sense—strongly influence the atherosclerotic process. If we smoke, lead a sedentary life, eat too much fat and too little fiber, the risk of premature atherosclerosis and early heart disease is increased. It is also well known that many social environments may enhance or prevent the improvement of lifestyle and behavioral risk factors. Therefore, psychosocial influences, including the concepts of stress and social support should be recognized as relevant, both in cardiology and in general health care.

Support from social networks may help the individual to deal with stress in many ways. Network members may provide help to recognize and evaluate stressors in the environment, and they may also be helpful in identifying solutions to overcome the stressor (information/appraisal support). This means that they may provide tools to improve problem-oriented coping, and thereby enhance stress reduction. Network members may also provide emotional support that may counteract the individual's feeling of loneliness and hopelessness, and thus help him or her to deal more efficiently with problems. Supportive actions may also relieve stress by reducing tension and anxiety, and counteracting depressive reactions (emotional support). Furthermore, supportive actions may also be helpful as they provide practical help (tangible support). Therefore, the empirical evidence of social support effects may be found at various stages of the pathogenic processes leading to heart disease. An epidemiological as well as a behavioral and psychophysiological view of the relationship is scientifically sound (see *Social Support and Stress*).

2. *Social Support and Cardiovascular Health in Men*

Large epidemiologic studies of social support and social networks have mainly focused on mortality as an endpoint. Scarce networks and few contacts are generally found to increase mortality risk. The incidence of acute cardiac events such as myocardial infarction or sudden cardiac death, however, has rarely been investigated in relation to social support. Furthermore, the most commonly studied aspect of social support is marital status. In men, marriage or cohabiting seems to improve cardiovascular health and to decrease mortality (Case et al. 1992). In women, as we shall see below, the picture is more complex.

Several large-scale epidemiological studies have been conducted in Sweden and Scandinavia. Scandinavian countries have a relatively homogenous, easily assessed population, living under good material conditions, where poverty, in its concrete sense, is rare. This does not mean that loneliness and lack of social support are irrelevant. In contrast, where the basic physical needs are met and a decent standard of living provided by public or governmental services, social ties and relations with other people tend to become diminished. In addition, the cold and dark climate, characteristic of these countries, does not promote social interaction, but rather may prevent naturally occurring social encounters.

In one of the early population-based studies of risk factors for CHD, the study of 50-year-old men born in Gothenburg (men born in 1913, in 1923, and in 1933), the impact of poor social networks and lack of social support was first assessed. Scarce social ties and social isolation increased both total and cardiovascular mortality (Welin et al. 1985). In further studies of men born in 1914 in the city of Malmö in south Sweden this observation was confirmed. The provision of satisfactory social ties and adequate social and psychological anchorage was essential to good health and longevity in men (Hanson 1989).

In a subsequent assessment of the Gothenburg study group, a random sample of 776 men, who were found healthy on initial examination, were followed for incidence of AMI and death from CHD. Lack of social support, as measured by the Interview Schedule for Social Interaction (ISSI), an instrument which was modified for population based studies in our laboratory, was found to be an independent predictor of AMI and death from CHD, including sudden cardiac death. This instrument is also described in Rosengren et al. (1993). These effects were maintained even after controlling for the effects of standard cardiovascular risk factors (hypertension, hyperlipidemia, smoking, diabetes, lack of exercise, etc.). The effect of lack of support was comparable in magnitude to that of smoking (Rosengren et al. 1993).

All other risk factors, including elevated cholesterol, or fibrinogen levels, hypertension or obesity were less strong predictors. When smokers and non-smokers were analyzed separately, risk gradients between men with high and men with low support were similar in magnitude. Male smokers with little social support were at highest risk for CHD whereas male non-smokers with large networks and good support were at lowest risk (7.1 percent vs. 0.8 percent) (Rosengren et al. 1993). These and other studies have demonstrated that lack of social support and its various functions are

predictors of disease incidence, and exert an effect that is independent of physiological and lifestyle risk factors.

However, once signs and symptoms of chronic disease have occurred one would expect them to override the impact of 'soft' risk factors such as social support. Therefore the study of patients suffering from chronic disease, such as coronary disease, is crucial.

In an early landmark study of over two thousand male survivors of acute myocardial infarction, Ruberman et al. (1984) showed that, even in the presence of clinical coronary disease, social isolation and stress contributed significantly to the increase in mortality risk. These risk factors were more common in men with low education. This early finding of a social gradient in health and disease has subsequently been frequently discussed and confirmed in both population- and patient-based studies (Haan et al. 1987).

In general, findings in men support the notion that one way in which the health hazards of social disadvantage are expressed is through the lack of adequate social networks and support. In order to provide appraisal and tangible support in times of hardships, men with high education and high job positions usually have a vast network of friends and acquaintances from professional and other organizations or groups, which they can draw upon for help in difficult situations. In low job positions, networks are known to be smaller and less effective; also social activities and social interactions are known to be more centered on family and family members. Therefore, the specific functions of information and concrete practical help, which are unevenly distributed across social strata and found to be crucial for maintenance of good health (Hanson 1989) may explain why cardiovascular health and health in general is better in higher social classes.

3. *Social Support and Emotions*

Social supports are provided to the individual by his/her social environment in an ongoing interactive process. This interactive process depends not only on the availability of supportive structures, but also on individual personal characteristics, such as the coronary prone behavior and/or emotional reactive patterns, such as depressive and anxious feelings. Whether the individual is able to receive and to provide social supports makes a difference for health outcomes. This includes the frequently studied coronary prone and hostile behaviors, which particularly in men were predictive of new incidence of coronary attacks (Brummett et al. 1998).

If considered in the presence of supporting social structures some of these findings are intuitively easier to understand. In a Swedish study of 150 male cardiac patients who were followed for ten years, evidence of such interactive effects on disease risk was found, in addition to clinical cardiac prognostic factors. The psychosocial factors assessed in these patients included social integration and social support as well as the Type A behavior pattern, using the original Structured Interview procedure. In accordance with other studies at the time, Type A behavior was not found to worsen prognosis in cardiac patients (Ragland and Brand 1988): 24 percent of men with Type A as compared to 22 percent of men with Type B died from CHD during follow-up. In contrast, social isolation was a prominent risk factor for poor prognosis. Men who said they practically never engaged in social activities with others for leisure were defined as socially isolated. They had about three times the mortality risk of men who said they were socially active at least once a month (Orth-Gomér et al. 1988). The worst prognosis, however, was found in Type A men who were socially isolated, their ten-year mortality risk being 69 percent, whereas, in socially active men, the behavior type made no difference (21 percent in type B and 17 percent in Type A men) (Orth-Gomér et al. 1988).

These results suggest that investigating either social environmental or individual behavioral factors alone is insufficient. Instead an attempt to integrate social and physical environment with individual psychological reactions appears to be more meaningful. Such models have been successfully applied and tested. An important example is the role of depressive reactions and their relation to outcome in patients with coronary disease.

In several patient-based studies it has been have shown that depressed symptoms and feelings increase mortality risk in patients post AMI (Barefoot et al. 1996, Frasure-Smith et al. 1995). These depressive reactions are tied into a social context that is much influenced by the individual and collective capacity for coping with stressors (Holahan et al. 1995). Such coping capacity will in its turn depend on the provision of social supports. Therefore the more recent studies of psychosocial factors in women have been conceptualized taking these integrative aspects into consideration.

4. *Social Support in Relation to CHD in Women*

Women have rarely been studied in regard to psychosocial factors and CHD. This is true for both social contextual factors, such as stress and social support, and the more individually based aspects of depression, coping, mastery, etc. It is interesting to note that many studies of coronary risk profiles did include women—but usually in too small numbers to be conclusive about their risk profile or about gender differences. Nevertheless findings are often generalized to both men and women.

This was one of several reasons for initiating the Stockholm Female Coronary Risk Study (FemCor-

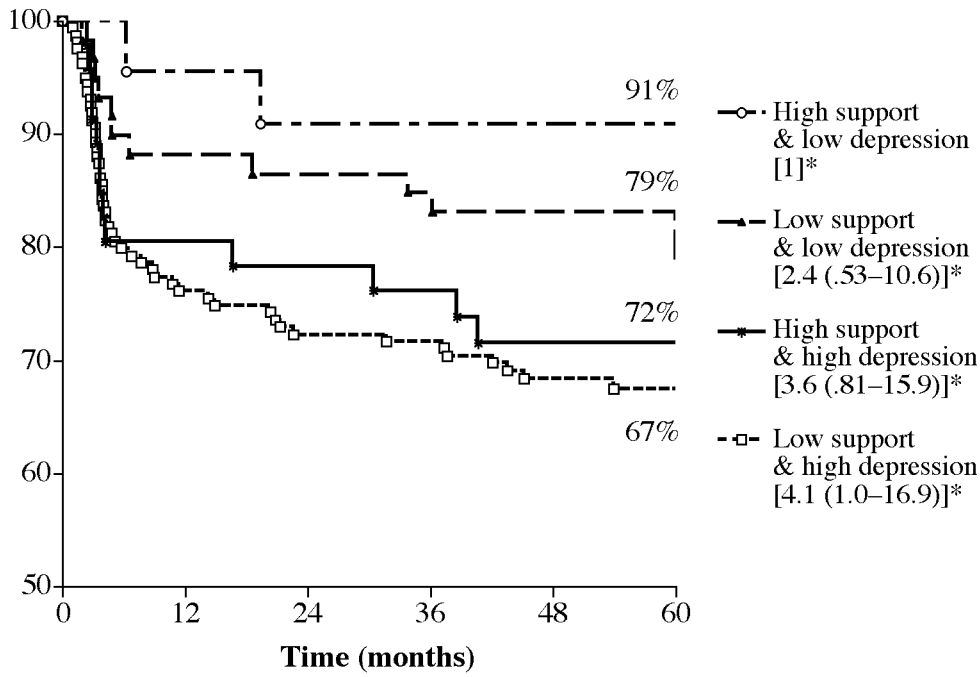

*Hazard ratio (95 percent CI), adjusted for age, P for trend = 0.01.

Figure 1
Percentage of women surviving free of cardiac events according to social support and depressive symptoms

Risk Study). This is a community-based case-control study, comprising 300 women below age 65, who were admitted to the coronary care units with acute coronary syndromes in greater Stockholm during a three-year period. Women with CHD were compared to a random sample of 300 age-matched women from the normal population, obtained via the official census registry of the same area. Psychosocial and biological risk factors were examined using interview, survey, psychosocial, and physical examination techniques. The core study objectives were:

(a) To examine psychosocial risk factor effects on CHD in women;

(b) To examine psychosocial determinants of lifestyle and the psychobiological mechanisms involved;

(c) To describe protective factors and mechanisms which buffer against noxious psychosocial influences. Several different strategies for analyses of the results have been applied, including case-control comparison, cross-sectional analyses among healthy women, prospective follow-up of women with cardiac disease, etc.

Lack of social support and social isolation was found to be one of the major stressors identified in this study group. As the sick role and the biological disease process itself may influence both social activities and social support in most people, we relied on follow-up rather than cross-sectional comparisons. Case-control comparisons were used to adjust for risk factors such as education, occupation, etc. Separate analyses of underlying coronary artery disease were undertaken, with patients being examined with quantitative coronary angiography (QCA).

QCA was carried out in 131 of the original 292 patients, based on available space at the angiographic laboratory. Patients who did and did not undergo angiography were compared, but differences in coronary risk factor profile were not found. Measures of severity (presence and number of stenoses greater than 50 percent) and measures of extent (number of stenoses greater than 20 percent) of coronary artery disease were evaluated.

After adjustment for age alone, lack of social support was associated with both measures. After adjustment for lifestyle and physiological risk factors, the risk ratio for stenosis greater than 50 percent in women with poor as compared to strong social support was 2.5 (Orth-Gomér et al. 1998). Women with poor social support also had a greater number of less hazardous stenoses, obstructing the coronary lumen by only 20 percent. After multivariate control for all other risk factors, however, this association was of only borderline significance. These associations were

not explained by severity of anginal symptoms, self-rated subjective health status, or other clinical indicators of poor prognosis.

Subsequently, the impact of social isolation and depressed affect was examined in a five-year follow-up of recurrent events in all women patients (Horsten et al. 2000). They were followed for a median of five years after admission for the acute coronary event. By means of the centralized healthcare system in Sweden, complete follow-up information for recurrent events (hospitalizations and deaths) was obtained (Alfredsson et al. 1997). A total of 81 patients (27 percent) either died from cardiovascular causes or had a recurrent AMI or a revascularization procedure.

Women patients, who at initial examination showed signs of depressed affect (two or more symptoms of depression as measured by a short instrument by Pearlin et al. 1981) had twice the risk of recurrent events as compared to women who were not depressed. When social isolation was added to the model a further increase in risk was noted. As shown in Fig. 1, 33 percent of those who showed both depressed affect and social isolation had new events as compared to 9 percent of those who were free from both these problems, an almost fourfold difference. Prognostic risk factors including age, menopausal status, diagnosis at index event, symptoms of heart failure, diabetes, systolic blood pressure, smoking, body mass index, exercise, total cholesterol, triglycerides, HDL, therapy of beta-blockers, or ACE-inhibitors did not explain the effects of psychosocial determinants (Horsten et al. 2000).

5. *Work Stress, Family Stress, and CHD in Women*

Psychosocial stress at work has been consistently studied and shown to increase the risk of CHD (Karasek and Theorell 1990). Using the concept of high demand and low control presented by Karasek and Theorell an increased CHD risk ratio of about two has been found in men workers. Although women have not been as extensively studied, it is suggested that they have about the same CHD incidence rates associated with work stress as men do (Schnall et al. 1994).

When analyzing the contribution of high demand and low control in the case-control comparison of the FemCorRisk Study an odds ratio of about two was found. A further—and stronger—source of stress was family problems. Based on the concept presented by Frankenhaeuser et al. (1989), exposure to both stressors at work and those that originate from outside work was analyzed in a structured interview procedure. The predictive values of the two different work stress measures were almost identical. In the case-control comparison the odds ratio for work stress was 2.3 whereas that for family stress originating from a dyadic relationship was 2.9. In addition both work stress and family stress were associated with low SES. As expected, women with unskilled jobs had lower control over their jobs, but they also had increased family stress, in particular from separations.

Furthermore the effects of these stressors on event-free survival in women with CHD were examined. The risk of a recurrent event in women patients with severe marital stress as compared to no or little stress was 2.8 after control for age, menopausal status, family history of CHD, symptoms of chest pain and heart failure, diabetes, smoking, and levels of triglycerides and HDL cholesterol. The actuarial five-year probability of being free from cardiac recurrent events in women without marital stress was 85 percent. In women with severe marital problems it was 65 percent (Orth-Gomér et al. 2000).

This effect was modified by individual capacity to cope with stressors, as assessed by a short instrument describing degree of mastery, obtained from Pearlin et al. (1981). This self-rated scale attempts to quantify the extent to which the individual perceives that he/she can deal with and master problems and stressors in everyday life.

In women patients with poor mastery the experience of severe marital stress increased the risk of recurrent events by a factor of eight. In women with good mastery there was no significant effect of marital stress on risk of recurrences. Furthermore, the ability to master or cope with problems was related to socioeconomic status (women with low SES having poorer mastery scores), depressive symptoms, and social isolation (Wamala 2001).

These findings may be summarized as follows: women patients who experienced marital stress, when assessed after an acute coronary event, had a worsened prognosis if they also had a low education or an unskilled job, if they had depressive symptoms, if they were socially isolated, and poor at mastering difficulties and problems of everyday life. Numbers in subgroups were too small to analyze risk ratios in these different categories, but it is probable that all these factors are interrelated and overlap. This knowledge can be used to implement and improve preventive measures aimed at secondary prevention in women. Such interventions are ongoing, both in the US (ENRICHD) and in Sweden (Healthier Female Hearts in Stockholm).

6. *Gender Differences in Cardiovascular Effects*

To our knowledge, the harmful effects of family stress, depressed affect, poor mastery, and social isolation have not been investigated in a similar way in men. However, in an earlier study of Swedish men with CHD, the question was asked which life domain had generated the most severe stress experiences, work,

family, education, or other. A majority of men (80 percent) ascribed their worst stress periods to the work situation, and only 10 percent to the family and spousal problems. This apparent gender difference is further supported by recent data from two large surveys of men and women of the general population in Massachusetts (Shumaker and Hill 1991).

When asked which person was the most important provider of social support, men were more than twice as likely to name their spouse, whereas women were most likely to name a friend or relative, usually female, as their primary supporter. In addition, women were more likely than men to report that they give more than they receive in dyadic relationships. In the FemCorRisk Study, 89 percent of the women said they were their spouse's closest confidants, whereas only 75 percent of women named their spouse as their closest confidant. In contrast to previous studies in men (see *Gender and Cardiovascular Health*), being married or cohabiting did not in itself provide any extra protection from cardiovascular ill health. Instead, for these women, the strain from a problematic spousal relationship significantly contributed to a poor prognosis over and above the effect of clinical cardiac predictors.

Therefore, the determinants, the protective health effects, and the mechanisms of social support may differ profoundly between men and women. Consequently they need to be separately conceptualized, examined, measured and analyzed, and implemented for use in clinical and community-based interventions and preventive efforts. In both men and women, however, the protective effects of social support and the health hazards of social isolation are strong and complex. Further research on processes and pathways as well as methods to implement this knowledge in preventive efforts is highly needed.

See also: Family Health; Gender and Cardiovascular Health; Gender and Physical Health; Illness: Dyadic and Collective Coping; Social Integration, Social Networks, and Health; Social Support and Recovery from Disease and Medical Procedures; Social Support and Stress; Social Support, Psychology of; Stress at Work; Support and Self-help Groups and Health

Bibliography

Alcalay R 1983 . In: Alcalay R (ed.) *Ethics*. Elsevier, Amsterdam, pp. 71–88

Alfredsson L, Hammar N, Hodell A, Spetz C-L, Åkesson L-O, Kahan T, Ysberg A-S 1997 Värdering av diagnoskvaliteten för akut hjärtinfarkt i tre svenska län 1995 [Validation of the quality of diagnosis for acute myocardial infarction in three Swedish districts]. *Socialstyrelsen* 84–8

Barefoot J C, Helms M J, Mark D B, Blumenthal J A, Califf R M, Haney T L, O'Connor C M, Siegler I G, Williams R B 1996 Depression and long-term mortality risk in patients with coronary artery disease. *American Journal of Cardiology* **78**: 613–17

Brummett B H, Babyak M A, Barefoot J C, Bosworth H B, Clapp-Channing N E, Siegler I C, Williams R B, Mark D B 1998 Social support and hostility as predictors of depressive symptoms in cardiac patients one month following hospitalization: A prospective study. *Psychosomatic Medicine* **60**: 707–13

Case B R, Moss A J, Case N 1992 Living alone after myocardial infarction. *Journal of the American Medical Association* **267**: 520–4

Cohen S, Syme S L 1985 *Social Support and Health*. Academic Press, New York

Frankenhaeuser M, Lundberg U, Fredriksson M, Melin B, Tuomisto M, Myrsten A-L, Hedman M, Bergman-Losman B, Wallin L 1989 Stress on and off the job as related to sex and occupational status in white-collar workers. *Journal of Organizational Behavior* **10**: 321–46

Frasure-Smith N, Lesperance F, Talajic M 1995 Depression and 18-month prognosis after myocardial infarction. *Circulation* **91**: 999–1005

Haan M, Kaplan G A, Camacho T 1987 Poverty and health: Prospective evidence from the Alameda County Study. *American Journal of Epidemiology* **125**: 989–98

Hanson B S 1989 Social network and social support influence mortality in elderly men. The prospective population study of 'Men born in 1914' Malmo, Sweden. *American Journal of Epidemiology* **130**(1): 100–11

Holahan C J, Holahan C K, Moos R H, Brennan P L 1995 Social support, coping, and depressive symptoms in a late-middle-aged sample of patients reporting cardiac illness. *Health Psychology* **14**: 152–63

Horsten M, Mittleman M, Wamala S P, Schenck-Gustafsson K, Orth-Gomér K 2000 Social isolation and depression in relation to prognosis of coronary heart disease in women. *European Heart Journal* **21**: 1072–80

House J S, Landis K R, Umberson D 1988 Social relationships and health. *Science* **241**: 540–5

Karasek R, Theorell T 1990 *Healthy Work: Stress, Productivity, and the Reconstruction of Working Life*. Basic Books, New York

Kropotkin P 1908 *Gegenseitige Hilfe in der Tier- und Menschenwelt*. Thomas, Leipzig, Germany

Orth-Gomér K, Horsten, M, Wamala S P, Mittleman M A, Kirkeeide R, Svane B, Rydén L, Schenck-Gustafsson K 1998 Social relations and extent and severity of coronary artery disease. The Stockholm Female coronary risk study. *European Heart Journal* **19**: 1648–56

Orth-Gomér K, Undén A-L, Edwards M-E 1988 Social isolation and mortality in ischemic heart disease. A 10-year follow-up study of 150 middle-aged men. *Acta Medica Scandinavica* **224**: 205–15

Orth-Gomér K, Wamala S P, Horsten M, Schenck-Gustafsson K, Schneiderman N, Mittleman M A 2000 Marital stress worsens prognosis in women with coronary heart disease. The Stockholm female coronary risk study. *Journal of the American Medical Association* **284**(23): 3008–14

Paracelsus [Theophrastus Bombastus von Hohenheim] 1599 *Labyrinthus medicorum errantium*. Hanoviae

Pearlin L L, Menagham E G, Lieberman M A, Mullan J T 1981 The stress process. *Journal for Health Social Behavior* **22**: 337–56

Ragland D R, Brand R J 1988 Type A behavior and mortality from coronary heart disease. *New England Journal of Medicine* **318**: 65–9

Rosengren A, Orth-Gomér K, Wedel H, Wilhelmsen L 1993 Stressful life events, social support, and mortality in men born in 1933. *British Medical Journal* **307**: 1102–5

Ruberman W, Weinblatt E, Goldberg J D 1984 Psychosocial influences on mortality after myocardial infarction. *New England Journal of Medicine* **311**: 552–9

Schnall P L, Landsbergis P A, Baker D 1994 Job strain and cardiovascular disease. *Annual Review of Public Health* **15**: 381–411

Shumaker S A, Hill D R 1991 Gender differences in social support and physical health. *Health Psychology* **10**: 102–11

Wamala, S P 2001 Stora sociala skillnader bakom dvinnors risk för kranskärlssjukdom. Okvalificerat jobb och slitningar i familjen avgörande faktorer [Large social inequalities in coronary heart disease risk among women. Low occupational status and family stress are crucial factors]. *Läkartidningen* [Journal of the Swedish Medical Association] **98**(3): 177–81

Welin L, Svärdsudd K, Ander-Perciva S, Tibblin G, Tibblin B, Larsson B, Wilhelmsen L 1985 Prospective study of social influences on mortality. The study of men born in 1913 and 1923. *The Lancet* I : 915–8

Williams R B, Barefoot J C, Califf R M, Haney T L, Saunders W B, Pryor D B, Hlatky M A, Siegler I C, Mark D B 1992 Prognostic importance of social and economic resources among medically treated patients with angiographically documented coronary artery disease. *Journal of the American Medical Association* **267**: 520–4

K. Orth-Gomér

Social Support and Recovery from Disease and Medical Procedures

1. Social Support and Recovery from Disease

Evidence examining the effect of social support on recovery from illness comes from a variety of health areas. These include studies on infectious disease, as well as cancer, heart disease and other chronic illnesses. Under most circumstances, social support has been shown to moderate the negative effects of illness and improve recovery. The effect of social support, however, is dependent to a large degree on the characteristic of the illness and the match between the type of support offered and the needs of the patient.

A recent study looked at the susceptibility to developing the common cold in volunteer subjects given nasal drops containing one or two cold viruses (Cohen et al. 1997). The results showed that those volunteers with more diverse social networks had a fourfold lower susceptibility to developing upper respiratory illness than those with a smaller number of social ties. In this study, social contact *per se* was not as important as the diversity of the individual's social relationships. The more an individual had regular contact with a variety of social groups such as family, workmates, neighbors and other social networks, the more resistant he or she was to developing a cold. Another study examining the role of psychosocial factors on progression to AIDS, found stress and social support to be significant predictors of HIV disease advancement (Leserman et al. 1999). In a follow-up period of five-and-a-half years, those participants who were more satisfied with the level of social support offered by their friends and families at study entry were twice as likely to be free from AIDS than participants who were dissatisfied with their available level of social support.

As well as infectious diseases, social support has been shown to have a positive effect on prognosis and recovery in a range of other medical conditions and illnesses. A study of low-income pregnant women found social support was associated with fewer complications during pregnancy and a higher infant birth weight. Research has found women who receive higher quality social support prior to delivery give birth to healthier babies and experience less post-partum depression. In chronic illnesses, Penninx et al. (1998) found positive effects on depression for having a partner and having many close relationships, although the strength of the effect differed across illnesses. Other work has shown social support to be associated with lower levels of disability and depression in diabetics, and higher levels of compliance in patients on renal dialysis.

In the cardiovascular area, the effect of having a close confidant has been shown to have an important effect on future mortality (see *Coronary Heart Disease (CHD), Coping with*). Williams et al. (1992) followed 1,965 patients with coronary artery disease after angioplasty. After accounting for known medical prognostic factors such as cigarette smoking and family history of heart disease, the most important socio–economic variable in predicting time to death was the presence or absence of a spouse or confidant. This finding is consistent with other studies of heart attack patients which have found a spouse to play a protective role in outcome.

A study of elderly myocardial infarction (MI) patients found those with low levels of social support were more likely to die in hospital and within six months of their heart attack even after controlling for severity of MI and other risk factors (Berkman et al. 1992). The available evidence suggests that the positive effect of social support seems to be stronger for male than female MI patients. For example, a 10-year follow-up study of MI patients admitted to Baltimore hospitals found higher survival rates in married compared to unmarried patients. The positive effect for marriage on both in-hospital and follow-up survival was stronger for males than females (Chandra et al. 1983).

Recent data suggest the more similar the spouse and patient view the illness the greater likelihood of effective adjustment and recovery. In work with MI patients, beliefs about the cause of the heart attack

were associated with positive changes in health behavior six months following MI. The belief by patients and spouse that the MI was caused by a faulty lifestyle was significantly related to overall improvements in diet and to an increase in the frequency of strenuous exercise. Attributions about how the illness developed are very common following MI and show a high degree of agreement between patient and spouse. They provide important guideposts for patients and their spouses by directing coping towards controlling a future myocardial infarction. Patient and spousal beliefs that the MI was caused by a faulty lifestyle were precursors to making changes in the types of foods eaten at home and participation in regular exercise. However, the belief that the MI was caused by stress, which is the most common attribution made by patients, had no relationship with later lifestyle changes (Weinman et al. 2000).

A large number of studies have examined the effect of social support on adjustment to cancer but fewer have studied the effect of social support on mortality. The literature shows social support interventions to generally have a positive effect on emotional adjustment to cancer. For example, a follow-up study of 294 people with breast, lung, or colorectal cancer found emotional support reduced distress and promoted adjustment to the disease (Ell et al. 1992). This study also found some survival benefits for social support in women with localized breast cancer.

Some researchers have directly investigated the value of social support by using it in controlled trials with chronic illness populations. Two randomized trials seem to show some health benefits for social support interventions. Spiegel et al. (1989) randomized women with metastatic breast cancer to participate in a one-year support group experience or usual care. There were no differences at four or eight months, but at one year the interventions group reported greater levels of adjustment than the control group. In a 10-year follow-up, the researchers reported that the women who participated in the support group had a significantly longer survival time. In another study of patients with Stage I and II malignant melanoma, patients were randomized to an intervention group that combined informational and education support or to a control condition. At six years, the intervention group showed decreased recurrence and increased survival six years later (Fawzy et al. 1993).

While these findings are encouraging for social support having health benefits in some chronic illnesses, it is not clear from the research to date what part of the interventions are the most effective or in which way they may influence health outcomes. It is still equivocal whether social support *per se* is the active ingredient that results in the positive outcome from these trials. As well as providing emotional support, the research interventions also provided education and information. Moreover, some later studies that have included social support interventions have not found the same effect on health outcomes (e.g., Cunningham et al. 1998).

A popular and relatively recent source of social support for patients with chronic illness are informal support groups. Here, patients meet together to discuss their condition and the difficulties and successes they have experienced in coping with their illness. As well as sharing personal illness experiences, these groups offer the advantage of providing information about utilizing health services, normalizing experiences and concerns and providing emotional support to group members. An analysis of social support groups found that the experience of physical illness was the most common reason for participation in such groups (Lieberman and Snowden 1993). Even after excluding substance abuse groups, this American survey found coping with physical illness made up 42 percent of the self-help group population. This study also suggested that these groups were more effectively accessed by the middle class populations who often have good access to other support services. One limit of such social support groups is that participants have to be well enough to be physically able to attend meetings.

This pattern may be changing with the advent of the Internet. In the past five years there has been a dramatic increase in the use of the Internet in most Western households. This has been accompanied by an increasing number of discussion and support groups for chronic illness. Now without leaving home, chronic illness sufferers can read postings from other sufferers and participate in discussions about aspects of their condition. Recent research on these illness newsgroups suggests that certain diseases have a higher number of postings than others and these rates differ markedly from their level of morbidity in the community (Davidson and Pennebaker 1997). For example, patients with breast cancer and chronic fatigue syndrome have a high Internet usage, while postings from heart disease and prostate cancer patients are relatively under-represented. The value of Internet support groups for patients with chronic illness has yet to be evaluated but the use of this type of support is likely to rise markedly in the future.

2. *Social Support and Recovery from Medical Procedures*

Although not as extensively studied as recovery from illness, there is considerable evidence that social support also has a positive influence on recovery from hospitalization and medical procedures. In a large study of 40,820 patients admitted to a University Hospital in Ohio, researchers investigated the effect of marital status on patient outcome (Gordon and Rosenthal 1995). They found that unmarried patients presented to hospital later than married patients. Unmarried surgical patients also had a higher risk of

dying in hospital than married surgical patients, even after controlling for severity of illness, diagnosis and other demographic factors. Furthermore, unmarried patients tended to stay longer in hospital than married patients. The differences found in this study were greater in patients who had never married when compared with patients who were widowed, divorced, or separated. This study suggests that being unmarried and in particular, never having married, increases your risk of negative outcomes during hospitalization.

The effect of different types of social support on recovery has been examined in the context of coronary artery surgery. A study examining the effect of different types of social support on recovery following cardiac surgery found that the perceived availability of social support was related to emotional and functional outcomes up to a year following the operation. In particular, esteem support, or feedback that one is valued and respected by others, was consistently related to outcome over the follow-up period (King et al. 1993). It is not clear why esteem-related social support is associated with positive outcome, but it is likely that such support may give the individual in this situation greater resilience to setbacks and help instill more optimistic beliefs that they have the ability to recover from the surgery.

One fascinating aspect in the study of social support and recovery from medical procedures is the effect of the patient's hospital roommate on recovery. Research in this area demonstrates that whether a patient has a pre- or postoperative roommate can influence outcome. Patients about to undergo surgery who are also assigned to a preoperative roommate show greater levels of anxiety and a slower recovery than those assigned a roommate who has already been through surgery. As well as lower levels of anxiety prior to surgery, being placed with a roommate who had been through surgery resulted in getting out of bed more quickly following the operation and a faster discharge from hospital (Kulik et al. 1993).

It seems likely then that the positive benefits resulting from a postoperative roommate are partly due to the natural relief of seeing someone who has survived the same operation that the preoperative patient is about to undergo, and by normalizing the feelings and problems following the operation so these are less likely to be viewed as unusual or abnormal by the patient. Thus a roommate who has been through the operation or procedure can provide a practical role model for preparing patients for the post-operative pain, physical sensations and sequence of recovery following the operation.

Actions by others designed to aid or assist adjustment with medical problems is not always seen as helpful by the recipient. Negative reactions to intended helpful actions are actually quite common. One study of elderly chronic illness patients found half to receive support from their spouse they did not need and 28 percent reported not receiving assistance for a daily activity when needed. This mismatch between the need and provision of support has been the focus of study in the general social support area and is very relevant to the area of recovery from medical illness. Here, the evidence suggests that different types of support are differentially valued depending on the relationship the provider has with the patient.

In an important study with cancer patients, Dakof and Taylor (1990) found emotional support was most likely to be valued when it came from intimate others, whereas informational support was more valued when it came from a physician or another cancer patient. This study found when there was a mismatch, such as when a patient needed emotional support from a spouse but received advice or information, then these actions were more likely to be seen as unhelpful. Unhelpful behaviors noted by patients were minimizing the problem, forced cheerfulness, medical care delivered without any emotional support, and being told not to worry. A later study using patients with the chronic non-life-threatening illnesses of irritable bowel syndrome and chronic headache, found these patients to put greater value on tangible and practical assistance and less on emotional support than the cancer patients in Dakof and Taylor's sample (Martin et al. 1994).

These studies highlight how the characteristics of the illness may strongly influence the type of social support needed by patients. In some life-threatening conditions, where patients are frightened or daunted by the demand of treatment and prognosis, emotional needs are more prominent. In other illnesses, where the illness forms a major hurdle to be met on a more routine daily basis without the threat of disfigurement or death, then more practical issues such as assistance with childcare and transport become more valued.

A disparity in the intensity of social support can also cause difficulties for patients. Sometimes an aversion or fear of physical illness can cause family to avoid contact with the patient. While emotional support is the type of support most received by patients it is also the type of support likely to be perceived as most inadequate. However, overly intrusive offers of assistance and help can also cause a negative reaction and increased stress. These findings highlight the importance of finding a balance between the needs of the patient and the type and level of social support.

3. *Summary*

The available evidence suggests social support has a positive influence on individuals' susceptibility to infectious disease and recovery and adjustment to a variety of chronic illnesses. In the cardiac area there seem to be particularly strong benefits for having a spouse or close confidant and these effects seem to be even stronger for men. Social support as an intervention in chronic illness seems to hold considerable

promise but further trials are needed to confirm early findings and to identify the necessary components of successful intervention programmes. With the rise of the Internet, electronic social support groups are going to provide an important source of information and peer support for patients with chronic illness.

There is strong evidence that the lack of social support is a risk factor for negative outcomes following hospitalization. Patients low in social support are more likely to present with more advanced illness and to be more likely to die while in hospital. However, it should be noted that many of the studies in the area are correlational, and the mechanisms that link social support to recovery from illness are still largely unknown. Recovery from surgery is an example where action by others that conveys to patients that they are respected and valued has a positive effect on outcome. The type of roommate can also set up a more positive outcome from the operation by providing a constructive role model by minimizing catastrophizing and demonstrating effective coping strategies.

A critical factor running through social support and the recovery from illness and medical procedures is the match between the needs of the patient and the type of social support provided. Different types of illness and procedures cause differing needs in terms of whether they require more emotional or tangible support. Emotional support is valued highest when it comes from a significant other. A mismatch in the type or intensity of support can lead to difficulties and hamper overall recovery.

See also: Chronic Illness, Psychosocial Coping with; Social Integration, Social Networks, and Health; Social Support and Health; Social Support and Stress; Social Support, Psychology of; Stressful Medical Procedures, Coping with

Bibliography

Berkman L F, Leo-Summers L, Horwitz R I 1992 Emotional support and survival after myocardial infarction: A prospective, population-based study of the elderly. *Annals of Internal Medicine* **117**: 1003–9

Chandra V, Szklo M, Goldberg R, Tonascia J 1983 The impact of marital status on survival after an acute myocardial infarction: a population-based study. *American Journal of Epidemiology* **117**: 320–5

Cohen S, Doyle W J, Skoner D P, Rabin B S, Gwaltney J M 1997 Social ties and susceptibility to the common cold. *Journal of the American Medical Association* **227**: 1940–4

Cunningham A J, Edmonds C V I, Jenkins G P, Pollack H, Lockward G A, Ward D 1998 A randomized controlled trial of the effects of group psychological therapy on survival in women with metastatic breast cancer. *Psycho-Oncology* **7**: 508–17

Dakof G A, Taylor S E 1990 Victims' perceptions of social support: What is helpful from whom? *Journal of Personality and Social Psychology* **58**: 80–9

Davidson K P, Pennebaker J W 1997 Virtual narratives: Illness representations in online support groups. In: Petrie K J, Weinman J A (eds.) *Perceptions of Health and Illness: Current Research and Applications*. Harwood Academic Publishers, Amsterdam, pp. 463–86

Ell K, Nishimoto R, Mediansky L, Mantell J E, Hamovitch M B 1992 Social relations, social support and survival among patients with cancer. *Journal of Psychosomatic Research* **36**: 531–41

Fawzy F I, Fawzy N W, Hyun C S, Elashoff R, Guthrie D, Fahey J L, Morton D L 1993 Malignant melanoma: Effects of an early structured psychiatric intervention, coping, and affective state on recurrence and survival 6 years later. *Archives of General Psychiatry* **50**: 681–9

Gordon H S, Rosenthal G E 1995 Impact of marital status on outcomes in hospitalized patients: Evidence from an academic medical center. *Archives of Internal Medicine* **155**: 2465–71

King K B, Reis H T, Porter L A, Norsen L H 1993 Social support and long-term recovery from coronary artery surgery: Effects on patients and spouses. *Health Psychology* **12**: 56–63

Kulik J A, Moore P J, Mahler H I M 1993 Stress and affiliation: hospital roommate effects on preoperative anxiety and social interaction. *Health Psychology* **12**: 118–24

Leserman J, Jackson E D, Petitto J M, Golden R N, Silva S G, Perkins D O, Cai J, Folds J D, Evans D L 1999 Progression to AIDS: The effects of stress, depressive symptoms and social support. *Psychosomatic Medicine* **61**: 397–406

Lieberman M A, Snowden L R 1993 Problems in assessing prevalence and membership characteristics of self-help group participants. *Journal of Applied Behavioral Science* **29**: 166–80

Martin R, Davis G M, Baron R S, Suls J, Blanchard E B 1994 Specificity in social support: Perceptions of helpful and unhelpful provider behaviors among irritable bowel, headache and cancer patients. *Health Psychology* **13**: 432–9

Penninx B W J H, van Tilburg T, Boeke A J P, Deeg D J H, Kriegsman D M W, van Eijk J T M 1998 Effects of social support and personal coping resources on depressive symptoms: Different for different diseases? *Health Psychology* **17**: 551–8

Spiegel D, Kraemer H C, Bloom J R, Gottheil E 1989 Effect of psychosocial treatment on survival of patients with metastatic breast cancer. *Lancet* **2**: 888–91

Weinman J, Petrie K J, Sharpe N, Walker S 2000 Causal attributions in patients and spouses following a first-time myocardial infarction and subsequent lifestyle changes. *British Journal of Health Psychology* **5:** 263–73

Williams R B, Barefoot J C, Califf R M, Haney T L, Saunders W B, Pryor D B, Hlatky M A, Siegler I C, Mark D B 1992 Prognostic importance of social and economic resources among medically treated patients with angiographically documented coronary artery disease. *Journal of the American Medical Association* **267**: 520–4

K. J. Petrie

Social Support and Stress

Social support is a dynamic process of transactions between people whereby assistance is received, especially during periods of stressful demands. Assist-

ance may take the form of advice, material aid, help performing tasks, or emotional support. Indeed, emotional support may be the most active ingredient in the support process. People who have a history of being supported by others when they need it tend to feel a strong sense of attachment to a group that is caring or loving and to which they feel they belong. This perception of being closely attached to others who may help when needed is a major aspect of why social support may have health benefits. That is, social support consists both of actual receipt of help and the perception that help will be forthcoming if needed.

1. *History of the Study of the Beneficial Effects of Supportive Ties*

Although the concept of social support was introduced in the social sciences in the 1970s, the idea that social relations may have a health-enhancing function is long-standing. Moreover, medicine has often spoken of the curative aspects of love, friendship, and positive social attachments. The influence of positive social interactions on mental health were first introduced in psychiatry by the French physician, Pinel, in the early nineteenth century and the English and American physicians, Connoly and Rush, in the late eighteenth to mid-nineteenth centuries. Having witnessed the madhouses of the time at Bedlam and the Bictre, they found that maniacs were treated with great cruelty. They offered a radical alternative to treatment. They suggested instead, that the mentally ill be treated gently, with patience, and with the use of 'kindness' and 'community' as the active ingredients of treatment.

Sociological influences into the importance of positive social attachments can be traced to the seminal work of Durkheim (1897), who found suicide to be related to social alienation from others. Later, the Chicago School of Sociology represented by Park et al. (1926) in their classic book *The City*, proposed that health and social accomplishment vs. sickness and deviancy were consequences of the extent that social groups were integrated into and supported by the broader community. Their theory was ahead of its time, and clearly challenged the then current explanations that genetic or moral inadequacy was responsible for deviancy. Such genetic and morality explanations were the predominant psychological view of the time held by such eminent scholars as Lewis Terman of Stanford and Robert Yerkes of Harvard, who argued that Jewish and other 'brunette' immigrant groups were genetically inferior and should be prevented from immigrating (Kamin 1974). Moving to more recent empirical investigations, researchers in the 1970s examined the relationship between social attachments and wellbeing. They found an association between socioenvironmental conditions and infant mortality, tuberculosis, and mental illness. They noted that social disorganization and negative social conditions contributed to increased illness on psychological and physical levels.

In a pioneering investigation, Berkman (1977) examined the social relationships and health of a random sample of adults in Alameda County, California. Those individuals who had ongoing social relationships, such as marriage, close friendships and relatives, church membership, and informal and formal group associations, had the lowest mortality rates. She interpreted her findings to mean that social isolation and lack of social and community ties disrupted people's disease resistance and that the presence of positive social attachments, in contrast, acted as a resistance resource (see *Social Integration, Social Networks, and Health*). But how are social relationships translated into physical or emotional health? Sarason (1974) suggested that people obtained a psychological sense of community, and theorized that this sense of community contributed to people's wellbeing. The healthy community guards its residents by facilitating healthy behavior (caring), offering people a positive sense of who they are (self-esteem), provides for safety and mutual support, and opens access to resources. He defined psychological sense of community as 'the sense that one was part of a readily available, mutually supportive network of relationship upon which one could depend … (who are) part of the structure of one's everyday living … (and) available to one in a "give and get" way.' The work of Caplan (1974), a pioneering psychiatrist, was perhaps the first to identify the concept of social support and its importance in the stress process as the active ingredient that these other theorists were noting. Caplan identified the nature of the process by which social relationships provided psychological sustenance, acting 'to mobilize … psychological resources and master … emotional burdens; share tasks; and supply (needed resources)' (Caplan 1974). Caplan further emphasized that support provision supplied and activated resources, but also contributed to a positive sense of another critical stress resistance resource—sense of mastery.

2. *Social Support's Influence on Mental and Physical Health*

When people encounter stressful conditions they may benefit from the assistance of others. The people who they naturally might rely on are termed their social support network. The social support network consists of the constellation of individuals who have a relationship with one another that is characterized by helping. How strong the social support network is depends on how often social support occurs between

network members, the kinds of support that is communicated through the network, and the strength of the ties. Different kinds of social support networks also perform different function. To find a job, a large, loosely connected network may be of most assistance, even though the relationships are not close. The greater number of contacts may maximize the chance of learning about job openings and obtaining interviews. In contrast, many studies have shown that when major life stressors occur, such as death, illness, or divorce, then a small intimate network might prove most helpful. In such a close network, people are more likely to share their sorrows and receive consistent, ongoing help and emotional succor (see *Gender Role Stress and Health*). Moreover, the help received from intimate others has greater personal meaning (Sarason et al. 1986). Social support contributes both to mental and physical health. Social support is related to lower psychological distress in the form of anxiety and depression. Comprehensive reviews of the literature found that social support contributes to mental health both through the positive feelings that are engendered by having close attachments with others and because help is mobilized in times of need (Cohen and Wills 1985, Schwarzer and Leppin 1989). The attachment aspects of social support are seen as contributing to wellbeing whether or not stressful events are present. This has been called support's direct effect. In contrast, the contribution of the actual receipt of support and knowing support will be forthcoming if needed may prove most beneficial during high stress periods. In this regard, those who have access to social support do better in particular when challenged by stressful circumstances. This has been called support's stress buffering effect.

Research has also examined social support's positive impact on physical health. Comprehensive reviews of this literature found that social support contributes to better clinical health outcomes (i.e., less disease) (House et al. 1988) and more positive stress responding on a physiological level that is subclinical (Uchino et al. 1996). Social support is related to lower morbidity and mortality across a wide array of diseases including heart disease, cancer, and the common cold. Those who are ill but who obtain social support also have been found to have less severe symptoms including less pain, less physical disability, and more rapid recovery. On a physiological level that underlies health, social support is related to beneficial influences on cardiovascular, endocrine, and immune systems. Social support may, in part, help by encouraging positive health behavior but this does not seem to be the major channel through which its positive influence occurs. Rather, social support has direct benefits on lowering cardiovascular reactivity, improved immune responding, and more physiologically adaptive responding of stress-related hormones. As in the mental health studies, support from intimates seems most important for sustaining physical health.

In the excitement and groundswell of interest in social support, researchers and theorists almost entirely forgot that interpersonal relationships could also be a critical source of stress. Rook (1984) shattered this honeymoon with the social support concept by reminding the field that interpersonal conflict was also one of the main contributors to stress. Moreover, it is notable that positive and negative social interactions often emanate from the same people. Rook also showed that when positive and negative interactions cooccur, that the negative ones are more disruptive of wellbeing than the positive ones are health enhancing. In work with women during the Israel–Lebanon War, researchers further found that even positive social interactions could have a negative impact when people shared a crisis. They called this the pressure cooker effect, whereby stress contagion occurs among people who find their resources commonly drained and their problems accentuated by having to deal with the same severe problem. This finding is especially critical because many of life's problems impact families or groups of people in a shared way, such as the case when there is a death in the family or in the face of disaster.

3. *Social Support and the Exchange of Resources*

The process and value of social support is placed in broader ecological context by Hobfoll's (1998) Conservation of Resources (COR) theory. COR theory's basic tenet is that people strive to obtain, retain, protect, and foster that which they value. These things that they value are termed resources. Resources are either valued in their own right or serve as a means of obtaining valued ends. There are four basic kinds of resources: objects, conditions, personal characteristics, and energies. Object resources are physical entities that are valued such as transportation, a house, or a diamond ring. Condition resources are social circumstances that avail people to other resources, such as love, money, status, or shelter. They include such conditions as marriage, tenure, and employment. Personal characteristics include skills or personality attributes that enable an individual to better withstand stressful conditions, achieve desired goals, or obtain other resources. They include personal attributes such as sense of mastery, self-esteem, and optimism and skills such as job skills or social skills. Finally, energy resources are resources that can be used to obtain other resources, but that may become valued in and of themselves. They include money, credit, and knowledge.

COR theory proposes that stress occurs when people (a) are threatened with resource loss, (b) actually lose resources, or (c) fail to gain resources following resource investment. Further, people use resources in order to limit such losses or to gain resources. For example, people use self-esteem in order to bolster their self-confidence after doing poorly on

an examination. Similarly, people invest money to purchase insurance in order to offset potential financial loss. Because resources are often hard to obtain and maintain, resource loss according to COR theory is considered to be more salient and of greater impact than resource gain. Resource gain, in turn, becomes more important in the face of resource loss.

COR theory helps explain the importance of social support. People have limited resources and, especially when under stress, they may find their resources inadequate. Through social support people can rely on others to offer the resources they lack, bolster their flagging resources, or remove them from the stressful circumstances so that they can regain resources or the ability to use their resources. This may take the form of offering object resources, such as shelter or transportation, and energy resources, such as money or information. It may also take the form of bolstering personal resources that the stressful conditions have depleted. For example, divorce may compromise individuals' self-confidence, hope, and optimism. Emotional support can act to replenish these diminished resources by reminding individuals that they are loved, important, and likely to succeed in the future.

COR theory also explains the potential costs of social support. First, the resources offered may be of poor quality. Supportive messages may contain submessages that are actually undermining, such as when people offer support but imply that they have the upper hand. Second, using social support means incurring social obligations that may be costly in the future. This is the case because people often refrain from wanting to incur obligations, and because when they need to return the favors bestowed upon them they may have to risk valuable resources at an inopportune time. This may differ in communal vs. individualistic cultures, but even in communal cultures where the transport of support is more balanced, incurring obligations has social costs.

A third cost of obtaining social support is that other resources often have to be expended to call on and employ help. Reaching out to others requires time, a critical resources that may already be taxed during stressful conditions. Calling on help may also force people to admit to some inadequacy, which they may not wish to acknowledge. Particularly in Anglo-Saxon and Germanic cultures the concept of 'standing on one's own two feet' is a central value. This, in turn, may diminish self-esteem, self-confidence, and sense of mastery.

COR theory further emphasizes that social support is not just a matter of receiving resources from others when needed. Building a supportive social network requires extensive, long-term outlays of resources in order to build and maintain supportive ties. This means that to have support available requires a history of resource investment. In some relationships this may be one-sided, as in parent–child relations. Similarly, different relationships are characterized by varying histories of give and take. These patterns will set the stage for the commerce of social support when a crisis occurs.

4. *Fitting of Social Support to the Type of Stress*

The help offered by social support may or may not fit the demands of any given stressful situation. This relationship between the nature of the demand and the adequacy of social support in meeting those demands has been termed social support fit. Emotional support has been found to fit many kinds of demands, and is the most robust, wide-reaching kind of support. However, even emotional support may not be appropriate for a given demand. When advice is needed it is important that the right kind of advice be available. Similarly, when support is needed for tasks some kinds of help may hinder more than they help.

An important aspect of social support fit is highlighted by COR theory. Specifically, many stressful circumstances are chronic. COR theory suggests that resource reservoirs are finite and that even if initial support has excellent fit with demand, this does not mean that the right kinds of support will be available for the ongoing demands that ensue (Norris and Kaniasty 1996). People, however loving and helpful, may need to return to their own immediate family or may need to shelter their resources for the needs of other members of the family. How this is the case with money is readily understood, but love too takes time to communicate. Even a parent may have to withdraw resources from a needy child so as not to sacrifice the needs of other children at some point. Chronic stressors will deplete individuals' resource reservoirs and those of their support system (see *Stress and Coping Theories*).

5. *Future Research Directions*

Future research on social support is likely to take a number of directions. On the biological level, researchers are only beginning to uncover the relationships between social support and physiological reactivity and disease. Research will examine more specific diseases and the interaction of social support with the time element of disease. Hence, researchers will be examining social support's impact on premorbid states, disease onset, and recovery. On the physiological level, research will be looking at how social support influences physiological reactivity that precedes disease or sustains health.

On the more social psychological level, research will be investigating developmental aspects of social support in terms of how it comes about and how it is sustained. This will include evaluation of how culture

plays a role in social support interactions and the different meanings of social support in different cultures. Social support's impact will continue to be examined in every imaginable stressful circumstance, from university examinations, to the workplace, to disasters. With an aging population, how social support impacts the process of adaptation among older adults will be an expanding area of interest. It is clear that social support is multifaceted and can be looked at on multiple levels and interest in the topic is likely to not only continue but to actually increase in scope.

See also: Chronic Pain: Models and Treatment Approaches; Social Integration, Social Networks, and Health; Social Support and Health; Social Support and Recovery from Disease and Medical Procedures; Social Support, Psychology of; Stress and Health Research

Bibliography

Berkman L F 1977 Social networks, host resistance and mortality: A follow-up study of Alameda county residents. Ph.D. thesis. University of California, Berkeley, CA

Caplan G 1974 *Support Systems and Community Mental Health*. Behavioral Publications, New York

Cohen S, Wills T A 1985 Stress, social support, and the buffering hypothesis. *Psychology Bulletin* **98**: 310–57

Conolly J 1856 *The Treatment of the Insane Without Mechanical Restraints*. Smith, Elder, & Co, London

Durkheim E 1897/1952 *Suicide, a Study in Sociology*. Routledge & K. Paul, London

Hobfoll S E 1998 *Stress, Culture, and Community: The Psychology and Philosophy of Stress*. Plenum, New York

House J S, Landis K R, Unberson D 1988 Social relationships and health. *Science* **241**: 540–5

Kamin L J 1974 *The Science and Politics of I.Q*. L. Erlbaum Associates, Potomac, MD

Norris F H, Kaniasty K 1996 Received and perceived social support in times of stress: A test of the social support deterioration deterrence model. *Journal of Personality and Social Psychology* **71**: 498–511

Park R E, Burgess E W, McKenzie R D 1925 *The City*. University of Chicago Press, Chicago

Rook K S 1984 The negative side of social interaction: Impact on psychological well-being. *Journal of Personality and Social Psychology* **46**: 1097–1108

Sarason I G, Sarason B R, Shearin E N 1986 Social support as an individual difference variable: Its stability, origins, and relational aspects. *Journal of Personality and Social Psychology* **5**: 845–55

Sarason S B 1974 *The Psychological Sense of Community: Prospects for a Community Psychology*, 1st edn. Jossey-Bass, San Francisco, CA

Schwarzer R, Leppin A 1989 Social support and health: A meta-analysis. *Psychology and Health* **3**: 1–15

Uchino B N, Cacioppo J T, Kiecolt-Glaser J K 1996 The relationship between social support and physiological processes: A review with emphasis on underlying mechanisms and implications for health. *Psychology Bulletin* **119**: 488–531

S. E. Hobfoll

Social Support, Psychology of

The concept of social support has both scientific and popular appeal. The topic has received attention as the important influence of psychological and interpersonal factors on the health and well-being of individuals has increasingly been recognized.

1. History

Early research in this field was characterized by vague definitions, poor and inadequate samples, and inappropriate generalizations. Psychologists have long recognized the importance of the association between social support and health generally, and interpersonal factors and individual mental health specifically. It was only when the importance of these factors on physical health was documented that the research became of interest to a broader array of scientists, researchers, and clinicians. This turning point can perhaps be best identified with the early work of Berkman (Berkman and Syme 1979). Using the Alameda county longitudinal epidemiological data which followed a large representative sample of residents from the California county, Berkman and Syme were able to document that people reporting social ties with friends and organizations were more likely than people without such ties to be alive nine years later. Other studies essentially replicated these findings with respect to mortality in Tecumseh, Michigan (House et al. 1982) and Evans County, Georgia (Blazer 1982, see also Bowling and Grundy 1998 for a recent review). This began a series of studies documenting the association between social support and numerous aspects of physical and mental health, including depression, recovery from disease and medical procedures, cancer, heart disease, experience of pain, and adherence to a medical regimen (Antonucci 1990, Berkman 1985, Holahan et al. 1995).

2. Definitions

While the term social support was initially used as a nonspecific, generalized term, experts have now come to recognize the importance of providing more detailed definitions and, more importantly, of distinguishing between different related concepts. We

propose the generalized term social relations to describe the broad array of factors and interpersonal interactions that characterize social exchanges and social support between people. Under this rubric we find it is useful to distinguish three critical terms: social network, social support, and support satisfaction.

Social networks can best be understood as the objective characteristics that describe the people with whom an individual maintains interpersonal relations. Thus, one can describe one's social network members in terms of age, gender, role relationship, years known, residential proximity, frequency of contact, and the like. Describing a social network tells you about the people with whom an individual has relationships but not about the exchange that takes place within those social relations. To describe this aspect of social relations, the term social support is used.

Social support refers to the actual exchange of support. Social support has been described as an interpersonal transaction involving the exchange of one of three key elements: aid, affect, and affirmation (Kahn and Antonucci 1980). Aid refers to instrumental or tangible support such as lending money, helping with chores, or providing sick care. Affect refers to emotional support such as love, affection, and caring. Affirmation refers to agreement or acknowledgment of similarities or appropriateness of one's values or point of view. Other terms have also been used, e.g. informational and appraisal support.

The third term often used in this literature is evaluative support or support satisfaction. This term refers to the assessment individuals make about the support they receive. It is an evaluative reflection of a combination of both social network and social support characteristics as well as more idiosyncratic characteristics such as perceived and expected support. Since individuals are fundamentally psychological beings, every individual does not experience the objective receipt of support in the same way. Thus, a support provider might say to a support recipient 'I love you', which might be heard by one person as a pledge of lifelong romantic love and by another as a passing sign of affection—an indication of platonic, brotherly love, or even a sarcastic statement of disdain. Support satisfaction refers to how individuals evaluate or assess the support they receive. They might feel support is adequate, that it meets their needs. Some individuals will be satisfied with the support they receive, i.e. positively evaluate the support they receive, and others will not—even when an observer would report that the two individuals have received the same support.

3. *Measurement*

Recognition of these different concepts and the distinctly different definitions associated with them leads directly to the recognition of the problems associated with measuring these constructs. Measurement has evolved significantly over the last several years. Different types of measures have been used such as open-ended questions and structured interviews. Also interesting is the fact that many different methods have now been used to study social support specifically and social relations more generally. These include in-depth ethnographic studies, laboratory studies, as well as large representative, epidemiological studies. When assessing primary sources in this field it is important to be cognizant of the specific definitions and consequent measures which have been used to assess the social relations characteristics of interest since these facts fundamentally influence the kind of information obtained.

4. *Theoretical Perspectives*

Several theoretical conceptualizations relating to social support have been offered such as attachment across the lifespan, socioemotional development, family systems, and developmental aspects of friendships (see Levitt 2000 for an overview). It is now generally recognized that a lifespan perspective provides a critical underlying base upon which to study social relations. The most important exchanges of social support and social relations usually involve long-lasting, close relationships such as those between parent and child, siblings, husband and wife. Interpersonal relationships are lifespan in nature either because they have existed across the lifespan or because they are built on interpersonal relationships that have been experienced over the individual's lifetime. To understand how an individual experiences a relationship at any one point in time, it is most useful to understand the history of that specific relationship as well as other relationships in the person's life. Thus, the study of a 50-year-old woman and her 70-year-old mother is best conceptualized as a relationship that has continued from when the 50-year-old was an infant through her early childhood, adolescence, young adulthood into the current state of middle age. Similarly, her mother was at one time a young 20-year-old mother who loved and cared for her child then and throughout the next 50 years. The resultant mother–child relationship is likely to be understood very differently depending on these lifetime-accumulated experiences. One might even argue that a specific negative experience could and would be forgiven because of a lifetime of positive experiences. Or, on the contrary, a contemporary supportive interaction may be unable to make up for a lifetime of disappointment or neglect.

4.1 *Convoy Model*

The convoy model of social support is actually a model of social relations more generally (Antonucci

1990, Kahn and Antonucci 1980). It conceptualizes the longitudinal nature of these relations: the convoy is a structural concept shaped by personal (age, gender, personality) and situational (role expectations, resources, demands) factors which influence social support and support satisfaction. Changes in the convoy over time critically affect individual health and well-being.

4.2 Theoretical Work

Recent theoretical work has focused on explaining how social support positively affects the health and well-being of the individual. Antonucci and Jackson (1987) offer the support/efficacy model. This model suggests that it is not only the exchange of specific support that accomplishes these effects but the cumulative expression by one or several individuals to another that communicates to the target person that he or she is an able, worthy, capable person. Assuming that the support recipient perceives the support provided as accurate and altruistically motivated, the support recipient will come to internalize this same belief communicated by the supportive other. Thus, with multiple and repeated exchanges of this type, the supported person accumulates a belief in his or her own ability which will enable that person to face and succeed in the multiple goals and challenges of life, i.e., to cope with problems across the lifespan. Not all people receive support from those who surround them and not all support is positive (Rook 1992). Thus, some people, instead of being told that they are competent, capable, and worthy, may be led to believe that they are incompetent, incapable, and unworthy. Just as the cumulative effect of positive exchanges can have a positive affect on an individual's health and well-being, so the opposite can have a devastating and cumulative negative effect on health and well-being. Some support networks can have a negative effect by supporting or encouraging behaviors detrimental to health and well-being. Drug buddies, drinking buddies, eating buddies, and buddies who discourage exercise are all examples of people who help or 'support' behaviors that simply are not good for you.

Other theoretical perspectives have been offered. For example, some scholars (e.g., Thoits 1995, Pearlin et al. 1996) consider social support a coping mechanism. They argue that as people are confronted with stress, the receipt of social support from important close others is a critical and significant coping strategy. Sarason et al. (1990) have taken a more personality-based perspective, arguing that individuals with specific personality characteristics will seek and benefit from different types of social relations including specific types of social support exchanges. Carstensen et al. (1999) have suggested an interweaving of social relationships, emotional development, and lifespan experiences. Although not technically a theory of social support, Carstensen has offered a perspective which argues that at younger ages people are more likely to seek out and be advantaged by numerous exchanges and social contacts, whereas as they age people reduce the number of exchanges in which they engage to only those which are perceived as most close. Finally, two other perspectives on social relationships have focused on intergenerational solidarity (Bengtson et al. 1996) and adult friendships (Blieszner and Adams 1992). Bengtson has documented the unique perspective of each generation towards intergenerational support. Older people of the more senior generations are likely to report that they provide a great deal of instrumental and affective support to their younger, junior generational family members. Younger people, while often agreeing to feelings of solidarity with older generations, report fewer support exchanges than their senior generation relatives. Bengtson et al. explained this as important differences in developmental stake in family, generational, and close social relationships. Finally, Blieszner and Adams note the unique supportive nature of adult relationships, recognizing that adult friendships serve an important function because, unlike family relationships, they are nonobligatory, most often among peers, and therefore on an equal or nonhierarchical basis.

5. Support Findings

While there is still much to learn about the way in which social relations and, specifically social support, operates, there is quite a bit that is already known. There are both sociodemographic and cultural similarities and differences in social relations (Jackson and Antonucci 1992). It is known, for example, that most people report immediate family, mother, father, spouse, child, and siblings as their closest relationship (e.g., Antonucci and Akiyama 1995). It is known that people differentiate their relationships from very close to much less close. While all these relationships can be important, it is clear that the closest relationships have the biggest effects in most circumstances. However, it is worth noting that there can be special circumstances where a person generally not considered very close can have a significant effect on the individual. A supervisor not considered a friend but who goes out of his/her way to extend praise or deliver criticism is an example.

It is also known that social relations vary by age, gender, socioeconomic status (SES), race, ethnicity, culture, and other sociodemographic characteristics. The number of close social relationships and amount of emotional support is relatively stable across the lifespan until very old age (Due et al. 1999), when a significant decline occurs (Baltes and Mayer 1999). This seems true in most countries for which we

currently have data. Although men and women seem to have the same general structure to their social networks and very little difference in the size of their networks, it does seem to be the case that men have fewer close relationships than women. Men seem to maintain a significant, close, and high-quality relationship with their spouse, while women maintain this type of relationship with several people. On the other hand, men seem to be less burdened by their relationships and to feel less responsible for solving the problems of people in their networks. They also have fewer people they feel they can turn to in times of trouble and feel less comfortable seeking and obtaining support from others.

The data on SES indicates that the social networks, support exchanges, and support satisfaction of people from different SES levels vary considerably. To put it succinctly, people at lower SES levels have smaller networks which tend to be sex segregated and to consist mainly of family members. They exchange support with fewer people and are often, though not always, less satisfied with the support they receive. Race, ethnicity, and culture have been shown to affect social relationships although there appears to be some fundamental characteristics of social relationships that transcend group membership. Thus, for example, most support networks consist of family members but different groups have different customs and sometimes different characteristics, which influence support relationships. So, for example, the church is a formal group that provides significant support to most members of the African-American community. On the other hand, Hispanics place a particularly high value on family relationships which are characterized by large, densely knit support groups. In the United States, the daughter is often the most significant support provider and caregiver for older parents. However, in Japan, daughters-in-law more often provide support than daughters because the traditional living arrangements in Japan require daughters-in-law to live with her husband's parents. It is nevertheless important to remember that structure and culture do not always predict perception, expectation, and/or quality of support. A recent study (Tilburg et al. 1998) comparing older people in the Netherlands who often live alone with older people in Italy who often live with their children, found the Italians reporting less social integration and more loneliness than the Dutch. It may be that Italians expect more, perceive less, and therefore are more likely to downgrade the quality of the support they receive. These are important distinctions of which we are aware but about which we understand little. An additional perspective is offered by research now available which demonstrates, at the psychoimmunological level, that people who are lonely or have more hostile interactions with others are much more likely to become ill, develop heart disease, cope less well with stress, and take longer to recover from health problems (Uchino et al. 1996).

6. *Future Directions*

Future programs designed to improve social relations and social support need to be sensitive to individuals and yet not encourage or induce dependence (Silverstein et al. 1996). Social support can be most useful when it instills in the individual a feeling of being valued and competent, of being worthy and capable. We are just beginning to understand how this might happen. Acknowledged possibilities include the exchange of support, which involves both receiving and providing support, reciprocity in social relations, and the convergence of expectations, perceptions, and receipt of support. Simply stated, identify ways to maximize positive support and minimize negative support. The next important challenge in this area is finding ways to incorporate what we know about social support specifically and social relations more generally into prevention and intervention programs that will improve the physical and mental health of people of all ages.

See also: Adult Psychological Development: Attachment; Coping across the Lifespan; Social Relationships in Adulthood; Social Support and Health; Social Support and Recovery from Disease and Medical Procedures; Social Support and Stress

Bibliography

Antonucci T C 1990 Social supports and social relationships. In: Binstock R H, George L K (eds.) *The Handbook of Aging and the Social Sciences*, 3rd edn. Academic Press, San Diego, CA, pp. 205–26

Antonucci T C, Akiyama H 1995 Convoys of social relations: Family and friendships within a lifespan context. In: Blieszner R, Bedford V H (eds.) *Handbook of Aging and the Family*. Greenwood Press Westport CT, pp. 355B–372

Antonucci T C, Jackson J S 1987 Social support, interpersonal efficacy, and health. In: Carstensen L L, Edelstein B A (eds.) *Handbook of Clinical Gerontology*. Pergamon Press, New York, pp. 291–311

Baltes P B, Mayer K U (eds.) 1999 *The Berlin Aging Study: Aging from 70 to 100*. Cambridge University Press, New York

Bengston V L, Rosenthal C, Burton L 1996 Paradoxes of families and aging. In: Binstock R H, George L K (eds.) *Handbook of Aging and the Social Sciences*. Academic Press, San Diego, CA, pp. 253–82

Berkman L F 1985 The relationship of social network and social support to morbidity and mortality. In: Cohen S, Syme L S (eds.) *Social Support and Health*. Academic Press, New York, pp. 241–62

Berkman L F, Syme S L 1979 Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda county residents. *American Journal of Epidemiology* **109**: 186–204

Blazer D G 1982 Social support and mortality in an elderly community population. *American Journal of Epidemiology* **115**: 684–94

Blieszner R, Adams R G 1992 *Adult Friendship*. Sage Publications, Newbury Park, CA

Bowling A, Grundy E 1998 The association between social

networks and mortality in later life. *Reviews in Clinical Gerontology* **8**: 353–61
Carstensen L L, Isaacowitz D, Charles S T 1999 Taking time seriously: A theory of socioemotional selectivity. *American Psychologist* **54**: 165–81
Due P, Holstein B, Lund R, Modvig J, Avlund K 1999 Social relations: Network, support and relational strain. *Social Science and Medicine* **48**: 661–73
Holahan C J, Moos R H, Rudolf H, Holahan C K, Brennan P L 1995 Social support, coping and depressive symptoms in a late-middle-aged sample of patients reporting cardiac illness. *Health Psychology* **14**: 152–63
House J S, Robbins C, Metzner H L 1982 The association of social relationships and activities with mortality: Prospective evidence from the Tecumseh community health study. *American Journal of Epidemiology* **116**: 123–40
Jackson J S, Antonucci T C 1992 Social support processes in the health and effective functioning of the elderly. In: Wykle M L (ed.) *Stress and Health Among the Elderly*. Springer, New York, pp. 72–95
Kahn R L, Antonucci T C 1980 Convoys over the life course: Attachment, roles, and social support. In: Baltes P B, Brim O (eds.) *Life Span Development and Behavior*. Academic Press, New York, Vol. 3, pp. 253–86
Levitt M J 2000 Social relations across the life span: In search of unified models. *International Journal of Aging and Human Development* **51**(1): 71–84
Pearlin L I, Aneshensel C S, Mullan J T, Whitlach C J 1996 Caregiving and its social support. In: Binstock R H, George L K (eds.) *Handbook of Aging and the Social Sciences*. Academic Press, San Diego, CA, pp. 283–302
Rook K S 1992 Detrimental aspects of social relationships: Taking stock of an emerging literature. In: Rook K S (ed.) *The Meaning and Measurement of Social Support*. Hemisphere Publishing, New York, xiv, 327, pp. 157–69
Sarason B R, Sarason I G, Pierce G R 1990 *Social Support: An Interactional View*. John Wiley and Sons, New York
Silverstein M, Chen X, Heller K 1996 Too much of a good thing? Intergenerational social support and the psychological well-being of older parents. *Journal of Marriage and the Family* **58**: 970–82
Thoits P A 1995 Stress, coping, and social support processes: Where are we? What next? *Journal of Health and Social Behavior*. Extra Issue : 53–79
Tilburg T, de Jong Gierveld J, Lecchini L, Marsiglia D 1998 Social integration and loneliness: A comparative study among older adults in the Netherlands and Tuscany, Italy. *Journal of Social and Personality Relationships* **15**: 740–54
Uchino B N, Cacioppo J T, Kiecolt-Glaser J K 1996 The relationship between social support and physiological processes: A review with emphasis on underlying mechanisms and implications for health. *Psychological Bulletin* **119**: 488–531

T. C. Antonucci

Social Survey, History of

The social survey grew out of concerns about the social consequences in Western Europe and North America of rapid industrialization and urbanization, as well as from the desire to investigate society.

The origins of the investigation of the condition of the working or laboring classes in modern times may be traced to social investigations in the UK in the late eighteenth and early nineteenth century. Such social investigations were not surveys, but are part of the immediate prehistory of the social survey. The social survey as a tool of scientific inquiry is not much more than 110 years old.

1. *The Survey Defined*

Several specific characteristics distinguish the social survey from the modes of social investigation that preceded it. A social survey involved field work, the collection of data at first hand by a social investigator rather than reliance upon reports by others or on pre-existing data. Surveys attempted to achieve comprehensive rather than haphazard coverage, albeit initially within a local rather than a national area. The data in surveys related to individuals, families, and households rather than aggregates, and were analyzed accordingly. Survey research involved the attempt, however primitive, at counting and quantifying the phenomena with which it was concerned. And the social survey developed in close relationship with public policy and social reform.

Throughout the nineteenth century there were investigations into public health, housing, family life, and employment, particularly as these affected the working classes. Some, such as the studies by Frédéric Le Play of working-class family budgets, had wider scientific aspirations. The investigations were carried out by private individuals (many of them, in the later nineteenth century, women of some social position), certain professions (associated with, in particular, medicine), members of voluntary associations concerned with social welfare, a few journalists and (by the end of the century) one or two academic scholars. Although the methods of inquiry varied, there was some common ground. Indeed, for a period in the USA what became known as the Social Survey Movement flourished, engaging significant numbers of private, government, and academic researchers. These inquiries signified increasing upper-class and middle-class interest in the condition of the working classes as well as a desire to intervene—a desire both to remedy want and disease through voluntary or state action and to achieve a greater degree of social control through the use of scientific expertise.

There was thus a dual involvement of both social scientists and social reformers in the history of the social survey, and an interplay between them. It is important to stress that before 1940 the social survey was not associated particularly closely with academic social science, which was in any case itself quite small in scale. The modem scientific sample survey, established by academics such as Rensis Likert and Paul

Lazarsfeld, government statisticians such as P. C. Mahalanobis and Louis Moss, and market researchers such as Henry Durant and Mark Abrams, has developed since that time. Before World War II, a variety of individuals and groups, only a minority of them academics, contributed to the establishment of the social survey. Occasionally surveys have originated in an abstract desire for more knowledge about the structure and workings of society; more frequently, however, they have been carried out as an indispensable first step in measuring the dimensions of a social problem, ascertaining its causes, and then deciding upon remedial action.

2. *Studies of Local Social Conditions*

The history of the social survey can be conveniently schematized into two periods, the early studies of local social conditions—particularly poverty—and the post-1940 probability sample survey. The former were mainly local, the latter frequently national and more recently international in scope and design. However, both types of survey shared features in common. A fundamental change in the concepts guiding the acquisition of social knowledge accompanied the rise of the social survey. A comparison between Henry Mayhew's examination of the London poor in the middle of the nineteenth century and Charles Booth's surveys three decades later is instructive. Booth conceptualized the problem differently, conducted a larger-scale inquiry, attempted to measure the phenomena with which he was concerned, and collected data from multiple observers rather than relying upon a single observer. The emergence of the interview, however, was a gradual process.

The history of the social survey is usually dated in Britain to the 1880s when, as economic depression heightened social tensions, social and political groups that had competed in prosperity descended into conflict. One manifestation of these changing circumstances was sporadic social unrest. One particular question which exercised opinion in the middle and upper classes was the extent of poverty. In the 1880s, with more trying economic conditions, the problem of poverty became more insistent and the discrepancy between the numbers in receipt of poor relief and the proportion of the population having inadequate incomes due to underemployment, unemployment, family breakdown, ill-health, old age, and other misfortunes was more apparent. But how great was the difference and who would provide an answer to the question?

The absence of an adequate institutional base from which to undertake such inquiries was a major obstacle. This base was not provided by the political parties. Government departments did little at this period to gather information on social conditions. With the exception of one or two positions in economics, social science was not institutionalized in universities.

Existing methods of collecting social data about poverty were patchy and often lacking in reliability. Henry Mayhew's observational approach emphasized vivid insight and detailed description over extensive enumeration. Statistics compiled by central and local government, such as they were, related to administrative function and were neither designed nor collected by people with professional training. The interrogatory methods of parliamentary committee and Royal Commissions as devices for investigating problems were unlikely to lead to a satisfactory overall picture of social conditions, producing impressionistic results which varied between individuals. There was a marked lack of clarity in defining the phenomenon.

3. *Charles Booth*

The pivotal figure of Charles Booth looms large in the history of poverty surveys. Charles Booth was a wealthy Liverpool shipowner who was annoyed by the lack of a factual basis for the accounts of urban poverty of the time and lack of clarity about three questions to which he sought answers: How many were poor, why were they poor, and what should be done to alleviate poverty? Through careful empirical study Booth sought answers for the East End of London to the first two questions. *The Life and Labour of the People in London*, published in 17 volumes between 1889 and 1903, is the first great empirical study in the social survey tradition. The publication of the first volume of Booth's study of poverty in London in 1889 created a sensation.

Booth's originality lay in his methods of inquiry, not just in how he collected and analyzed his data, but in the way he conceptualized the problem of poverty. Booth attempted to introduce precision to the concept of the 'poverty line' where hitherto there had been vagueness, a characteristic which justifies including him as one of the founding fathers of social science. Booth sought to measure the numbers of the population above and below the poverty line with some precision, even though the classification categories that he used lacked adequate justification. This too differentiated his work from previous studies, and foreshadowed the later use of the social survey for estimating population values with some exactitude. Furthermore, although Booth's study, like most of the early social surveys, was a local study, its coverage was comprehensive—he aimed to gather information on as large a proportion of the population as was possible and especially on all families with children in the parts of London covered by the survey.

Booth was also unusual in using a team of investigators. First-hand data about the household circum-

stances of poor families were not collected directly by his research team but by what one of them, Beatrice Webb, called the method of the 'wholesale interview.' School board visitors, who entered the homes of children attending London schools, were interviewed by Booth to gather information about the circumstances of each family. Booth had printed data collection notebooks that categorized the information gained in interviews with the visitors. Great care was taken to achieve uniformity in the data collection between the different interviewers, and the information obtained was checked against other sources in an attempt to increase its reliability. The quantitative and qualitative data from the notebooks were analyzed by Booth and his staff. The published results were aggregations of the measurable and distillations of the unquantifiable. The findings had direct policy implications and contributed to movements toward social reform and old age pensions in particular.

Booth's research had important effects on the development of the social survey, independent of any influence his findings may have had, and popularized survey research. Several early US studies were derivative of Booth. And Booth's work served as a template for a large number of early social researchers.

4. *Seebohm Rowntree and A. L. Bowley*

The next major survey in Britain, Seebohm Rowntree's study of York, was conducted in 1899 and published in 1901. This further refined both the conceptualization and measurement of the phenomenon of poverty. Most important, perhaps, it used 'retail' rather than 'wholesale' interviewing; that is to say, people were interviewed directly in their own homes by members of the research team, not indirectly through reports by middle-class professionals with knowledge of individual families. This transition from reliance upon informants who could report on the circumstances of working-class people to reliance upon the response of the people themselves was a most important one. Rowntree also refined the conceptual understanding of poverty by introducing the concept of the 'life cycle,' and by showing how poverty varied between different phases of the life cycle. There were a considerable number of other local poverty studies in Britain prior to 1940, but the original studies by Booth and Rowntree, and restudies conducted in York by Rowntree in 1936 and 1951 were the most influential.

The first social surveyor to use sampling methods was Arthur Lyon Bowley in his studies of poverty in five towns, published as *Livelihood and Poverty* in 1915. This innovation, which was slow in being adopted, eventually undercut the local social survey, replacing it with the modem probability sample survey. The use of sampling in practice only emerged slowly and was first seriously taken up by US market researchers between the wars. It was used in social survey research for government and academic research only during and after World War II.

5. *Early Surveys in the United States*

Developments in the USA shared many features of this British experience. In both societies the social survey was shaped by the salience of voluntary effort, of liberal political traditions that restricted government initiatives, and of values promoting 'responsible individualism.' Thus we can speak of an Anglo-American tradition of social survey movements that shared many fundamental features. In both countries the settlement house movement provided a base for fostering social investigation. The influence of their ideas and the publicity they gave to the existence of social conditions created a more thoughtful climate for the reception of social surveys.

At the same time, the US context contained challenges not found in the British milieu. More isolated from political power, unable to draw on traditional criticisms of industrialization, and less likely to work closely with organized labor, reformers had to make the most of their middle-class resources. They also stood in a different relationship to emerging social science. By the end of the century, there was a different relationship between professional expertise and the developing skills of social investigation than in the UK, one which also had repercussions for conceptions of objectivity and of what constituted social science. In these circumstances, the social survey became even more important to middle-class reformers. For the objective data provided by the surveys were uniquely effective in mobilizing 'public opinion'—that amorphous combination of middle-class and working-class attitudes that activated politicians and labor leaders alike within the US political environment. Surveys could get action (especially when linked with sensationalist press reports) even when 'monster' meetings and petitions could not.

The social survey in the USA in the period between 1880 and 1940, while resembling and differing from its British counterpart, directly imitated Booth's work in certain respects. Florence Kelley's work on the Hull-House Maps and Papers was one of the earliest US surveys and very much in the Booth tradition. Other studies followed. The role of the settlement house movement in publicizing social conditions was not confined to poverty. Indeed issues such as child labor, the treatment of juveniles in the court system, sweated labor and poor housing conditions loomed as large if not larger in their concerns. At the state level some notable innovations were effected, for example, the establishment of the Illinois Juvenile Court in 1899. W. E. B. DuBois' study of *The Philadelphia Negro* (1899) by a young black sociologist was notable for its thoroughness and detachment.

6. The Social Survey Movement

The social survey movement, which was best expressed in the Pittsburgh Survey of 1906–9, was different. It emerged from the charity movement, and the belief among significant numbers of social workers that the roots of social problems were to be found in society rather than in the individual. The Pittsburgh Survey, funded by the newly created Russell Sage Foundation, was more in the nature of investigative journalism, with much less emphasis upon quantification. More emphasis was put on influencing public opinion. When published, a great effort was made to feed back the results of the research to the locality in which it had been carried out, by means of publicity and exhibits. This gave surveys of this type a completely different character, for they became exercises in community self-study more than scientific examinations of social questions. Indeed, its success was followed by the establishment of the Department of Survey and Exhibits at the Russell Sage Foundation, which between 1912 and 1931 supported a very large number of local surveys. The US social survey movement had run out of steam by 1930 and disappeared shortly thereafter. There were other developments, such as 'crime surveys' as an aid to administrative reform, and the country life movement and the church survey movement, but these were of little lasting methodological significance.

The geographically local social survey between 1880 and 1940 in the UK and the USA led to distinctive changes in social investigation. What distinguished Booth from Henry Mayhew, W. E. B.DuBois from Ray Stannard Baker, A. L. Bowley from R. H. Tawney, was the urge to present systematic and defensible numerical statements about the problem studied based on a large-scale data collection exercise derived from questioning individuals. Such an appeal to 'objectivity' and 'science' is one which became increasingly prominent as the twentieth century advanced. The local survey was closer to social reform and amelioration, to political action and to nonacademic institutions.

7. The Probability Sample Survey

After 1920, the social survey developed more rapidly and achieved more outstanding results in the USA than in the UK, and at the same time underwent a change in form. It is true that at the theoretical level, Britain maintained its preeminence for longer. The traffic to sit at the feet of statisticians Karl Pearson, R. A. Fisher, G. Udny Yule, and A. L. Bowley in the University of London continued until well into the 1930s. Leading US survey researchers such as Samuel Stouffer learned advanced inferential statistics from them and returned to apply those methods to good effect in the USA. At a practical level, however, the main implementation of probability sampling methods which came gradually to be used, giving the survey as a method very much greater power, took place in the USA. After Kiaer and Bowley's pioneering work, the application of sampling in national studies to provide a representative national picture may be traced to market research and opinion polling in the 1930s. From 1928, George Gallup had been successfully using probability sampling for marketing and political research in the USA. In 1936 Henry Durant established the British counterpart of Gallup's organization, the British Institute of Public Opinion (BIPO). Gallup's triumph over the *Literary Digest* in 1936 in correctly predicting Roosevelt's victory from a much smaller sample was a clear marker of the future potential of the method.

At the same time, efforts to improve social measurement were being made. During the 1920s, political scientists and psychologists, such as Harold Gosnell and L. L. Thurstone, made advances in the systematic measurement of voting and of social attitudes. Research units in government departments, one of them led by Rensis Likert, were established. Paul Lazarsfeld came to the USA from Austria as a Rockefeller Fellow, and remained there to commence survey research on mass media audiences in the mid-1930s. By the late 1930s, the basis for the development of the modern probability social survey was in place. With the outbreak of World War II and in particular the entry of the USA into the war in 1941, significant activity took place in government. The American Soldier study directed by Stouffer was important in its own right theoretically and methodologically, and for the stimulus it gave to postwar research at Harvard and Columbia. Likert's group in the Department of Agriculture became the nucleus of the postwar development of the Survey Research Center at the University of Michigan. Lazarsfeld independently widened his interest in the media to voting and the effects of mass communications in elections.

With the increases in federal government support for social science post-1945, survey research took off, becoming established in the academic world as well as in bureaucratic social survey organizations outside it. There was particularly strong cross-fertilization at centers at the National Opinion Research Center at the University of Chicago, the Survey Research Center at the University of Michigan and the Bureau of Applied Social Research at Columbia. In the UK, government was important, with the Government Social Survey remaining the main center for survey research. Other major survey research organizations and centers in Western Europe followed later. Since the mid-twentieth century, the intellectual leadership in modem sample survey research has lain in the USA, which has tended to remain in the forefront of intellectual fertility, analytic inventiveness, and methodological innovation.

See also: Archives and Historical Databases; Databases, Core: Sociology; Demographic Data Regimes; Demographic Techniques: Data Adjustment and Correction; Economic Panel Data; Longitudinal Data; Panel Surveys: Uses and Applications; Sample Surveys, History of; Survey Research: National Centers; Surveys and Polling: Ethical Aspects

Bibliography

Bulmer M, Sklar K K, Bales K (eds.) 1991 *The Social Survey in Historical Perspective 1880–1940*. Cambridge University Press, Cambridge, UK

Converse J 1987 *Survey Research in the United States: Roots and Emergence, 1890–1960*. University of California Press, Berkeley, CA

M. Bulmer

Social Trauma

Social trauma is a new concept in the social sciences. It transfers a notion from the clinical psychology of individuals to the sociology of groups and cultural communities. On the social level it denotes a collectivity's response to an event that is considered to be an overwhelming and unexpected threat to its cultural identity and social order.

1. History of the Concept

Although the term trauma had been used in medical contexts before, its origin as a distinctly psychological concept is commonly attributed to Sigmund Freud. In their studies on hysteria Freud and Breuer characterized the memory of the psychic trauma as 'a foreign body which long after its entry must continue to be regarded as an agent that is still at work' (Freud and Breuer 1955 (1893–1895), p. 6). Very early Freud noticed the time delay and the possible disparity between the cause and the traumatic response: 'In traumatic neurosis the operative cause of the illness is not the trifling physical injury but the effect of fright' (Freud and Breuer 1955 (1893–1895), pp. 5–6). Freud did not yet conceive traumatic memory as a social or collective response. Instead, he refers to processes within the psychic system, in particular to the fundamental tension between instinctive drives and defensive controls. An overwhelming and threatening experience results in an indelible anxiety that is coped with by various defense mechanisms like denial, reversal of cause and effect, or projection of the evil to others.

Psychoanalysis prepared the ground, but the rise of the notion to its current use is due to its relation to Post-Traumatic Stress Disorders (PTSD) that were clinically studied and professionally treated after the First World War (shell shock) and, increasingly, after the Second World War (Holocaust survivors, war experiences, rape, violence). During the 50 years since the wars the range of possible traumatizing experiences as well as the symptomatology of trauma vastly expanded —turning 'trauma' into a key concept for the analysis of individual and collective phenomena that before had been studied under the conceptual labels of 'crisis,' 'neurosis,' 'anomie,' and 'disorder.'

2. Social Trauma as Collective Response to the Breakdown of Order

Genocide and the outburst of violence, military defeat and foreign occupation, sudden economic crisis and mass poverty, forced migration and expulsion from home—but also natural disaster and the spread of mortal plagues—are experienced as shocking events that not only cause great individual suffering, but may also result in the breakdown of social order, that is, in the collapse of the most basic 'taken for granted' social expectations. As these 'evil' events affect more than a few members of a social community, they cannot be considered as exceptional and individual misfortunes, but rather as events that affect the validity and stability of the social order itself. As a result, they erode basic trust in social institutions, in the solidarity that creates the boundaries of family, neighborhood, and country. The predictability of everyday life is suspended; ordinary time seems to stop (Neal 1998); and the social fabric is disrupted. Under the impact of the shocking event, feelings of hopelessness, apathy, fear, and disorientation spread in the community. Participation in general politics and public debates as well as economic investments and consumption decrease sharply, and crime rates increase. This sharp decline of trust and the rise of apathy and disorientation can be called a social trauma.

In distinction to individual trauma, the social trauma even extends to those members of the community who, as individuals, did not suffer from the traumatizing condition. Because social trust typically extends beyond immediate others, the distrust produced by collective trauma tends to be copied by other members of the community who are devoid of a strong personal experience of breakdown and crisis.

In this conception, social trauma is seen as a response to the sudden breakdown of legitimate social expectations and the rapid decay of social institutions and social structure. The trauma is seen as an 'adequate' or natural response to the collapse of social order, and this response can occur without a time delay, immediately after the experience of the traumatizing condition. Although there are mediating institutions and cultural patterns that reinforce and support the traumatic experience, this experience is

seen as caused by an external event or condition that is directly observable and objectively accessible. An event becomes a social trauma by its very catastrophic objectivity, that intrudes into ordinary life and resists any attempt to ignore it or to accommodate to it.

3. Social Trauma as a Construction of Collective Memory

An alternative theoretical paradigm defines social trauma as the sudden collapse of the culturally generated web of meaning that supports individual or collective identity. Because the event that disrupts the web of meaning threatens the identity of an individual or of a collectivity, it may often pass unnoticed when it happens: the mind—individual or collective—is unable to comprehend the moment when its own death may occur. Therefore, social traumas are rarely witnessed directly and immediately, but are crystallized and become available for experience only in the reconstruction of the past. It is the individual or collective memory that strives, voluntarily or forced by others, to remember and to cope with an intrusive event that crushed the social and cultural basis of identity.

This ruminating memory of a past that suddenly confronts a community with the abyss of incomprehensibility is called a social trauma. Because social traumas refer to incomprehensible or horrible occurrences, they are hard to put into narration or to represent by common images or patterns. As individual or collective identities do not refer to objects but to the subjectivity of agents, they can be named, but they escape general definitions and descriptions and so do the traumatic challenges and threats to these identities. In this respect the traumatic memory profoundly differs from the narrative memory that remembers past episodes of meaningful action. However, the narrative memory repeats the past event because it fits into the cultural patterns of the present; the ruminations of traumatic memory result from the very fact that it cannot be narrated and represented by the cultural forms at hand—it occupies our mind because it resists being integrated into our perspective on the world. The strong disturbance and disruption of the individual or collective consciousness produces ambivalent responses and fosters hysteria—obsessive anxiety and repression by denial, the fear of conspiracies and the feeling of guilt, the persecution of scapegoats and hostility towards the victims (Smelser 2001).

4. Individual and Collective Trauma

The culturally constructed trauma can appear on the level of the individual or on the level of the collective memory or on both. If individuals remember the same kind of personal traumatic memories they can—in principle at least—communicate their common or similar experiences and thus develop a collective traumatic memory as members of a generation, as a community of victims, etc. Thus the victims of rape and violence, of racist attacks or forced migration can develop a distinctive collective identity based on common traumatic experiences. But not all personal traumatic memories can be translated into the collective memory of a community. Some remain enclosed in the individual mind and body—because the individual is unable to talk about it, because no other person is expected to understand it, or because the individual trauma is framed by a collective triumph, like the victory of the nation that tends to forget the personal trauma of its soldiers. In contrast to these private and individual traumas, there are social traumas that are largely disconnected from individual experience and memory, but exist, after some time at least, only on the level of collective communication and public culture. Thus the traumatic origins of a nation, a religious community, a generation or an ethnic group can be remembered by public rituals, memorial days, and official monuments, although there may be no surviving witnesses who could claim a personal memory of the traumatizing event. Here, the continuation and reproduction of the trauma is shifted entirely from personal memory to the collective memory of a community, its social institutions, and cultural traditions. Sometimes the collective remembering of the trauma is even impeded by haunting personal memories that cannot be spoken of. After the social trauma is decoupled from personal memories, and transferred to social carriers devoid of these personal memories it can be remembered by public rituals and literature, and it has to be remembered in this way if it is not to be passed into oblivion.

Although there is hardly a traumatic memory that cannot refer to any individual suffering, the cultural construction of a traumatic past does not directly correspond to the amount of individual suffering and the actual number of victims. A victorious war may not be remembered as a trauma, although the number of casualties may exceed those produced by a war that resulted in a military defeat and that, therefore, is remembered as a social trauma. Viewed from a culturalist perspective, the social trauma is not the plain and natural result of individual suffering but its reconstruction in collective consciousness. Whoever partakes in the communicative reconstruction of the trauma can regard him- or herself to be a member of the traumatized community, regardless of whether he or she can claim personal experience of the traumatizing event.

5. Modes of Coping with the Trauma

Most analysts of social trauma agree that there are different modes of coping with traumatic memory, but not all consent to the idea that the traumatized

community has to go through a fixed sequential order in coping with the memory. Following Freud we can distinguish between the period of latency and the period of speaking out and working through, or between different responses like denial and uncoupling, reversal, and projection. In the period of latency the community is unable to speak publicly about the trauma because the personal memories are still too vivid to be soothed by public rituals. The trauma is denied or silenced in public discourse or official social representations. If the issue cannot be avoided in communication, it is decoupled from the traumatized community: the time and circumstances of victimization are regarded as being radically different from today's situation; the victims are even seen as partly responsible for their fate (reversal); the evil is projected on scapegoats or on a few unquestionably responsible individuals.

Only when some time has passed, when a new generation has entered the stage, when a large and influential group of impartial but interested observers takes part in the public debates, can the trauma be publicly addressed by political representatives and even be ritually remembered by memorial days, monuments, and museums. Collective traumas like the Irish famine or the Holocaust of the European Jews, the slavery of African Americans, or the American civil war usually take more than the timespan of a generation to be publicly remembered and recognized. Even if the social trauma can be spoken of and is publicly recognized, its representation remains mostly an issue of contestation and conflict between different generations, political camps, religious or ethnic groups. Questions as to who is to blame for the trauma and who was the most important victim imply, even for succeeding generations, closeness or distance to the symbolic center of the community and are therefore a matter of identity and interest for collective actors within this community. In parallel with these public debates and conflicts, the trauma can be transferred to specialized institutional arenas like historical research and museums, literature, or movies, where it is worked through according to the special logic of the respective institution. Finally, the trauma may be accepted by a majority of a community and become an integral part of its construction of collective identity and history. Not all communities remembering a social trauma will pass through all stages of the sequence and, frequently, different modes of coping coexist and are carried by different social groups and institutional arenas.

6. *Victims, Perpetrators and the Public Audience*

In its common usage, the concept of social trauma refers to a community of victims. The trauma of the victims resulted from a profound and unexpected denial of their claims on dignity and autonomy as subjects and cohumans. Instead, they have been treated as objects that could be traded, used, deported, oppressed, and killed. Like objects, they were considered as cases of a category deprived of a name, a face, and a place in the community. Slavery and genocide, displacement, and mass violence are the paradigm cases of this dehumanizing treatment of victims. Immediately after the traumatizing event the surviving victims, because of shame and shock, can hardly report on their suffering—they are muted and sometimes they are even isolated from the regular members of their community in special camps or asylums. It is only from a temporal or social distance that the suffering of victims is remembered, that they are recognized as subjects and symbolically reintegrated into the community. This recognition, remembering, and reintegration cannot be achieved by the descendants of the victims alone. Instead, it requires the participation of an impartial public or of a noninvolved third party who consents to this process and agrees with the cultural self-definition of victims. If a nation or an ethnic group claims the identity of traumatized victim, it remains a contingent question as to whether other nations or groups recognize this claim in a public debate.

In contrast to a vast range of research on the trauma of victims, relatively few studies center upon the trauma of perpetrators. The trauma of perpetrators results from the fact that they violated fundamental moral principles, ensuring not only the victims' status as subjects, but also their own status. Because of their shame and their ignorance, perpetrators tend to be silent and to deny the moral crisis that they have caused. In this case, even more than in the case of victims, a strong third party is required to publicly debate the victimization and insist on stigmatizing the perpetrators. Particular institutions like courts of justice, or professional experts like medical doctors, can also take this position of a noninvolved third party. Only in exceptional cases, however, is the trauma of perpetrators accepted by those who were directly and actively involved in creating the traumatizing event. Instead, it is mainly the succeeding generation and other noninvolved members of the community who can speak of the collective trauma and admit publicly to the collective guilt. Like the social trauma of victims, the trauma of perpetrators also leads to public contestations and conflicts about the question of how the boundaries of the traumatized community of victims or of perpetrators will be demarcated. It also determines who has to be included, even if he or she was not directly concerned as an individual person. In the course of this debate the social trauma of victims and perpetrators is turned into the core of a public discourse about the collective identity of the nation or of the other communities involved. As such, it tends to replace the traditional public representation of national identity that showed the nation as a triumphant or tragic hero resisting its

adversaries. This demise of heroism and the rise of social trauma as the core reference for national identity is reflected in public monuments, museums, and movies, but also by the rhetoric and the rituals of political representatives who remember the victims of the past in public.

See also: Collective Behavior, Sociology of; Collective Memory, Anthropology of; Collective Memory, Psychology of; Genocide: Anthropological Aspects; Genocide: Historical Aspects; Mass Killings and Genocide, Psychology of

Bibliography

Eyerman R 2001 *Cultural Trauma*. University of California Press, New York, CA

Amato J A 1990 *Victims and Values. A History and a Theory of Suffering*. Praeger, New York

Assmann A, Assmann J 2002 *Pits and Pyramids. Dimensions and Dynamics of Cultural Memory*. University of Chicago Press, Chicago

Caruth C 1996 *Unclaimed Experiences. Trauma, Narrative and History*. Johns Hopkins University Press, Baltimore, MD

Caruth C (ed.) 1995 *Trauma. Explorations in Memory*. Johns Hopkins University Press, Baltimore, MD

Erikson K 1995 Notes on trauma and community. In: Caruth C (ed.) *Trauma. Explorations in Memory*. Johns Hopkins University Press, Baltimore, MD, pp. 183–99

Freud S 1964 [1939 (1934–1938)] Moses and monotheism. In: Strachey J (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London, Vol. XXIII, pp. 7–237

Freud S, Breuer J 1955 (1893–1995) Studies in hysteria. In: Strachey J (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London, Vol. II

Giesen B 2001 Lost paradises, failed revolutions, remembered victims. In: Giesen B, Eyerman R, Smelser N, Sztompka, P. *Cultural Trauma*. University of California Press, Berkeley, CA

Giesen B 2002 *Triumph and Trauma*. Cambridge University Press, New York

Giesen B 2000 National identity as trauma: The German case. In: Strath B (ed.) *Myth and Memory in the Construction of Community. Historical Patterns in Europe and Beyond*. P I E-P Lang, Brussels, Belgium, pp. 227–47

Kolk B A van der, McFarlane A C, Weisaeth L (eds.) 1996 *Traumatic Stress: The Effects of Overwhelming Experience on Mind, Body, and Society*. Guilford Press, New York

La Capra D 1998 *History and Memory after Auschwitz*. Cornell University Press, Ithaca, NY

La Capra D 1994 *Representing the Holocaust. History, Theory, Trauma*. Cornell University Press, Ithaca, NY

Neal A 1998 *National Trauma and Collective Memory. Mayor Events in the American Century*. M E Sharpe, New York

Smelser N J 2001 Psychological Trauma and Cultural Trauma. In: Giesen B, Smelser N, Eyerman R, Sztompka P (eds.) *Cultural Trauma*. Cambridge University Press, New York, CA

Vries de M 1996 Trauma in cultural perspective. In: Kolk B A van der, McFarlane A C, Weisaeth L (eds.) *Traumatic Stress: The Effects of Overwhelming Experience on Mind, Body, and Society*. The Guilford Press, New York, pp. 398–413

Weatherill R 1994 *Cultural Collapse*. Free Association Books, London

B. Giesen

Social Welfare Policies and Gender

In conceptualizing and defining social welfare policies, much discussion has revolved around the residual/institutional or the marginal/comprehensive distinction. This distinction is important in determining what policies are included in the term social welfare policies. Narrow definitions frequently confine social welfare policies to income maintenance programs—assistance and social insurance, and sometimes tax benefits. Broad definitions expand the policies to encompass provision of housing, healthcare, education, and an array of other social services.

The rise of the women's movement in the late 1960s and the development of women's studies prompted a re-examination of social welfare policies that focused on gender. In bringing gender into the analysis of social welfare policies, two strategies were pursued. First, the dearth of information about the impact of policies on women and differing outcomes for women and men led to an approach that primarily centered on women. A byproduct of this approach has been to make women visible in the analysis. The second strategy has stressed gender as a relational category where the inquiry explicitly deals with both women and men. Instead of making women visible, this strategy has shown how gender is involved in policies where it had previously been thought to be completely irrelevant.

A number of trends in research on gender and social welfare policies are discernible during the 1980s and 1990s. The initial emphasis on studying the impact of social welfare policies on women has been supplanted by a concern with how gender structures policies and how policies affect gender relations (O'Connor 1996). Generalizations based on the policies of a single country have given way to theory building grounded in comparative analyses (Lewis 1992; Orloff 1993; Sainsbury 1996; O'Connor et al. 1999). The first wave of theorizing was informed largely by the social welfare policies of the English-speaking countries (e.g., Wilson 1977), while a second wave reflected the experiences of the Scandinavian countries (e.g., Hernes 1987), and in the 1990s the range of countries has further expanded (e.g., Sainsbury 1999).

Comparative research has charted and attempted to explain major variations in policies across nations. The variations revealed by this research have challenged earlier generalizations derived from the experi-

ences of one nation, especially those that held specific features and outcomes were intrinsic to social welfare policies. Views on the nature of welfare policies have also shifted. In the 1970s a widespread view was that welfare policies were mainly instruments of patriarchal oppression; eventually more consideration has been devoted to the emancipatory potential of policies. This reconsideration has focused on both the unintentional consequences of policies that have benefited women or have undermined traditional gender relations (Pateman 1988) and the implications of different policy constructions (Sainsbury 1996). A parallel shift has occurred in theorizing about women's relationships to welfare policies. Previously women were regarded as the objects of policy, while in the 1990s research explored women's agency in shaping welfare policies (Skocpol 1992; O'Connor et al. 1999; Sainsbury 1999). Finally, a weakness of many earlier studies was their exclusive preoccupation with gender. Gradually, however, more attention has been paid to how gender intersects with race, ethnicity, and class in determining social entitlements (e.g., Williams 1989; Quadagno 1994).

Employing the lens of gender, scholars have called in question the definition of social welfare policies. They have been critical of narrow definitions of social welfare policies that focus solely on income maintenance programs. Feminist researchers have objected to the pivotal position of paid work in the analysis of social welfare policies. The emphasis on paid work takes men as the norm, and it obscures the importance of unpaid labor. This focus has also downplayed services and care. The incorporation of gender into the analysis has also substantially altered the research agenda on social welfare policies. Priority has been assigned to four areas of investigation: (a) the gender division of welfare, (b) the gendered construction of welfare policies, (c) welfare policies as an ordering force in gender relations, and (d) gender and explanations of welfare policies.

1. *The Gender Division of Welfare*

Social welfare policies—especially assistance and social insurance benefits—traditionally have been conceived as instruments of social protection and redistribution. At a minimum, social welfare policies should protect individuals from poverty and relative deprivation. More ambitiously, the policies aim to promote the general welfare and well being of the population. Mainstream studies evaluating the redistributive effects of social welfare policies largely have been gender blind. The units of analysis have been social class or occupational groups, households, generations, regions within a country, or nations. By contrast, feminist research has focused on the gender division of welfare and how the construction of policies has produced differences in the benefits received by women and men. The differences concern both access to benefits and benefit levels.

With this focus, studies have mapped out the gender composition of beneficiaries of social welfare policies. Early research found that a disproportionate number of women were beneficiaries, but these studies examined the English-speaking countries. When other countries were included in the examination, a different pattern emerged: men predominated as beneficiaries. In these countries social insurance schemes whose eligibility was based on labor market participation were central to social provision; men also collected family allowances as part of these schemes. Even when the allowance has been paid to the married mother, eligibility has been determined by the father's rights as an insured employee. Although men predominate as beneficiaries in some countries, households headed by women are more reliant on social welfare benefits than male headed households because men command a larger share of earnings. In many cases, especially when benefits are earnings related, the benefit levels of men are higher than those of women.

Studies have also documented disparities in the poverty rates of women and men. An initial finding was the feminization of poverty in the US in the mid-1970s. Although the poverty rates of women and men in the US had dropped substantially since the 1950s, the composition of the poor had changed dramatically so that women were overwhelmingly in the majority. In many countries the poverty rates of women are higher than those of men. However, among the industrialized countries, the poverty rate of women in the US ranks as highest, and the gender poverty gap is also the largest (Casper et al. 1994) (see *Poverty and Gender in Affluent Nations*). Nonetheless, there are countries—such as Italy, The Netherlands, and the Scandinavian countries (Sainsbury 1999)—where differences in the gender poverty gap are small or women have lower poverty rates than men. In summary, empirical research on the gender composition of beneficiaries of income maintenance programs in the US revealed an anomaly: women outnumber men as beneficiaries but their poverty rates were much higher than men's. This anomaly, together with cross-national variations, spurred theorizing about the sources of gender inequalities in benefits and the gendered construction of social welfare policies.

2. *The Gendered Construction of Social Welfare Policies*

Much analysis has concentrated on how gender norms and meanings are encoded in social welfare policies. One line of theorizing emphasizes that the gendered construction of social welfare policies results in dual welfare. This involves a two tiered system of social benefits, consisting of a 'masculine' set of programs geared to individuals who make claims as earned

rights based on labor market participation and a 'feminine' set of programs oriented to households that claim benefits to compensate for family failures. Underlying the two-tier benefit system are gender norms that define the home as a female sphere and outside work is a male sphere. Women thus claim benefits on the basis of tasks in the home and their benefits are familialized, while men's claims are primarily labor market related and individualized (Fraser 1989). Another variant of the thesis puts more weight on the dual labor market as the source of dual welfare (Nelson 1984). In any event, according to the dual welfare thesis, women and men are segregated into different types of social welfare programs: women rely more heavily on means-tested or income-tested benefits, and men on social insurance schemes. Assistance programs and social insurance are further differentiated in terms of benefit levels, political legitimacy and administrative intervention in private lives so that the quality of women's and men's social rights diverge greatly.

The importance of the dual welfare thesis is that it calls attention to the gender differentiation of social provision and specifies the differing implications for women and men. However, the thesis has been criticized on several grounds. The thesis is a generalization based primarily on social welfare policies in the US, but gender differentiation in entitlements varies in different ways across countries. Even with regard to the social welfare policies of the US the dual welfare thesis is misleading in one significant respect. The emphasis on predominance of women claiming means-tested benefits overlooks that many more women receive social insurance benefits on the basis of their husband's entitlements in the US (Orloff 1993; O'Connor 1996).

In studying the gendered construction of income maintenance policies, Orloff (1993) and O'Connor et al. (1999) distinguishes between gender differentiation and gender inequality. Gender differentiation is related to separate programs for labor market and family needs, shaping claims to benefits: women claim benefits as wives and caregivers, while men make their claims as family providers and earners. Gender inequality pertains to the differences in the benefit levels of women and men. The legislation of most countries ascribes less value to wifely and motherly labor than to paid work, resulting in gender inequalities in benefit levels. Spouse benefits and additional allowances for dependents have generally been a percentage—often around half—of the regular benefits received by men. Furthermore, the traditional division of labor between the sexes assigns domestic work and care to women, which has limited their ability to qualify for more lucrative benefits as earners.

Another important starting point has been an examination of the inscription of male breadwinner family model in social security legislation, taxation, and employment policies. As distinct from the discussions on dual welfare and gender differentiation, which look at income maintenance schemes, the male breadwinner model has been applied to a wider range of policies. Lewis and Ostner (1991) argue that this model has been built into social provision in varying degrees, and they classify countries as strong, moderate, and weak male breadwinner states. Subsequently, the weak male breadwinner category has more appropriately been termed the dual breadwinner model (Lewis 1992) or the dual earner model. A major contribution of this typology was to emphasize that the strength of the male breadwinner model as reflected in social welfare policies varied across countries. Nonetheless, the typology does assume a single basic underlying dimension, and that social welfare policies are informed by the male breadwinner model. A major criticism of this analytic construct is that it is based on two principles of entitlement: (a) as breadwinner or earner and (b) as the dependent of the breadwinner. Thus, it overlooks other bases of entitlement and it conflates women's entitlements as wives and mothers (Sainsbury 1996). The problematic nature of the typology is also revealed by its inadequacy in analyzing the entitlements of single mothers. Hobson (1994) has documented substantial variation across strong breadwinner states—such as The Netherlands, Germany, and Britain—in terms of social welfare policies providing mother only families with an acceptable standard of living.

Jenson (1997) has criticized Lewis's emphasis on the nexus between paid work, unpaid labor, and welfare, stressing that unpaid work is not a synonym for care. Jenson calls for making care central in theorizing about the gendered construction of welfare policies. More specifically, this would involve an analysis of 'the gender division of labor among caregivers, gender differences in the capacity or need to pay, and the gender consequences of different institutional arrangements for provision' (1997, p. 187).

Sainsbury (1996) has developed a framework based on two ideal types: the male breadwinner model and the individual model. This framework formalizes and expands the dimensions of variation implicit in Lewis's typology. The dimensions are: the nature of the familial ideology, the basis and unit of entitlement, the taxation system, employment and wage policies, and the organization of care work. It also clarifies an alternative policy model to the male breadwinner model. A major drawback of this analytic scheme is that it limits the models to two contrasting ideal types, but its dimensions of variation have proved to be a useful point of departure for empirical analysis and theory building.

A final development has been to theorize the gendered construction of social welfare policies in terms of gender regimes or gender policy regimes. More precisely, a regime consists of values, norms and rules, thus providing a normative or regulatory framework that influences behavior. A policy regime re-

directs the analysis from behavior to policies. Analogous to the notion of a social policy regime that emphasizes that a given state–economy organization provides a logic of it own, a gender policy regime is a given organization of gender relations associated with a specific policy logic. The organization of gender relations is determined by principles and norms (gender ideologies and practices) that specify the tasks, obligations, and rights to the two sexes (e.g., Sainsbury 1999).

All of these frameworks have furnished insights into the gendered construction of social welfare policies. Studies employing the frameworks have often challenged earlier classifications of countries' policies that neglected gender, leading to revisions and new understandings. Analyzing Britain, The Netherlands, Sweden, and the US, Sainsbury (1996) found strong divergences in the policy patterns of the four countries, and that the clustering of countries was quite different from how countries are bracketed together in analyses using mainstream typologies. In several mainstream studies Sweden and The Netherlands have been grouped together as countries with generous social welfare policies, but the gendered construction of Dutch and Swedish policies represented opposite poles, especially prior to the mid-1980s. Similarly, a study comparing the inscription of gender in the social welfare policies of Denmark, Finland, Norway, and Sweden revealed sharp differences between Norway and Sweden—the two countries that most closely approximate the social democratic welfare regime in mainstream literature (Sainsbury 1999). Orloff's examination of income maintenance policies in Australia, Britain, Canada, and the US led her to conclude that the way gender norms are encoded in these programs undermines the coherence of the liberal welfare regime (O'Connor et al. 1999).

3. *The Impact of Social Welfare Policies on Gender Relations*

Conceiving of social welfare policies as an ordering force in gender relations rivets attention on the implications of policies for reinforcing or undermining the traditional division of labor among women and men in the family and society. Theorizing on how traditional gender relations have been inscribed in policy frameworks has underlined that policies and benefit structures perpetuate existing relations and the subordination of the female sex.

Whether policies actually have these effects, however, is a matter for empirical research. In other words, it is necessary to distinguish between the gendered framework of policies and policy outcomes. Several discussions have conflated policy models and behavior. Women's labor market participation rates have been taken as evidence of a particular policy model but without examining whether policies are in place to enable or limit women's employment. For example, married women in the US have entered the labor force in massive numbers, despite income tax policy and social security measures that are biased heavily in favor of the traditional family with a single earner.

Both the principles underlying entitlement and the redistributive effects of policies have major implications for gender relations. The basis of entitlement is a crucial aspect in determining the impact of welfare policies on gender relations and whether policies buttress women's dependency on men or enhance their autonomy (Sainsbury 1996). At one extreme, women's entitlements as wives and means-tested benefits reinforce married women's dependency. If women are entitled to benefits and services as wives, i.e., their entitlement derives from their husband's rights, their dependency within the family is exacerbated. In this situation a woman's rights are jeopardized by divorce or may be contingent upon a marriage test—a specified number of years of marriage that guarantees entitlement. In means tested programs, the family is usually the unit of benefit; and need is established on the basis of family income. In these instances, benefits are familialized. Benefits based on the principle of care occupy a middle position. Although these benefits may buttress traditional gender relations, they recognize the importance of care and thus alter notions of deservingness. The principle of care also erodes the importance of marriage. Compared to entitlements as wives, benefits based on the principle of care enhance autonomy. Benefits whose eligibility is based on labor market participation or citizenship/residence have greater potential to increase personal autonomy. Even if women workers are likely to receive lower benefits compared to men workers, benefits based on labor market participation are an independent source of income and they can lessen married women's dependency on benefits based on their husband's rights. Finally, benefits based on citizenship or residence accord equal social rights to women and men. This basis of entitlement neutralizes marital status as a criterion of eligibility. It not only equalizes benefits between women and men but also between unmarried and married persons—and husband and wife. Likewise, citizenship/residence as a basis of entitlement makes no distinction between paid and unpaid work, which enhances the social rights of persons with weak attachments to the labor market.

Regardless of the basis of entitlement, welfare policies have redistributive effects that can alter the resources of women vis-à-vis men. In turn, changing resources between the sexes can affect gender relations. In countries where married women receive their own retirement pension, even though entitlement was based on their husband's rights, wives' contribution to family earnings has on average been higher compared to other couples. An even more striking example is that income maintenance programs, public provision of

affordable child care services, and/or tax benefits enable women to leave an unhappy relationship or marriage and strike out on their own. In most industrialized countries lone parent families are increasing, and social welfare policies help to make this family form a viable option (see *Lone Mothers in Affluent Nations*).

Policies that promote women's access to paid work are also vital to changing gender relations (Orloff 1993; O'Connor et al. 1999; Sainsbury 1999). Policy measures enabling employment provide both earnings and individualized social benefits, weakening or eliminating dependence on the male family provider. Women with high earnings have substantial bargaining power within the family, and this leverage can alter how partners divide unpaid domestic work among themselves.

While most analysts have empirically investigated the outcomes of policies on gender relations, Fraser (1994) adopts a normative stance and asks how can social welfare policies promote gender equity. She presents three policy models and evaluates the strengths and weaknesses of each model in terms of achieving gender equity. The first is the universal breadwinner model where policies promote women's employment so that the breadwinner role becomes universal—and is no longer reserved for men. In the caregiver parity model, public policies support informal care so that the role of caregiver is on a par with the role of breadwinner. The policies of the earner–carer model support both roles but simultaneously aim at eliminating gender differentiation. This involves policies that aid women in becoming earners and men in becoming carers (cf. Sainsbury 1996, 1999).

4. Gender and Explanations of Social Welfare Policies

In studies emphasizing the gendered construction of policies, gender relations themselves or the prevailing gender ideology are often a primary explanation of the particular form that social welfare policies take. Kessler-Harris (2000), e.g., conceives of gender as a historical agent, as an ideational force that configures and sustains systems of cultural representation. She has examined how gender mediates policy discussions between interest groups in providing or denying legitimacy to their proposals in the areas of social insurance, taxation and labor legislation.

Other studies have emphasized the role of women's political mobilization and the institutional setting in shaping social welfare policies. In attempting to explain the late introduction of social insurance schemes and the early adoption of maternal welfare measures in the US, Skocpol (1992) underlines the importance of the state in politicizing gender identities that lead groups to mobilize, the correspondence between the form of political organizing and the structure of the state that allows or thwarts a group's influence on policies, and policy legacies.

Along similar lines, O'Connor et al. (1999) suggest an explanatory framework made up of the following elements: social movement mobilization and orientation, political party configuration, the political opportunity structure, and the institutional context and legacy. The movement's orientation involves its political strategy to achieve gender equality, specifically the movement's attitude toward the state and its gender ideology, whether emphasis is on sameness or differences between the sexes. Political party configuration refers to the strength of left, center and right parties. The political opportunity structure is defined as access to the state institutions, the stability or fluidity of political alignments and relationships to allies and support groups. The institutional context and legacies encompasses the structure of the state, bureaucratic capacity, and the industrial relations framework (see also Sainsbury 1999). In short, many scholars have abandoned the view that women are simply the objects of policy, or 'policy takers' (Hernes 1987, p. 35). Instead research has focused on, first, women's power resources—organizational forms, discourses and strategies—in a variety of national and historical contexts and, second, how these contexts limit or promote women's capacities to influence social welfare policies.

See also: Lone Mothers in Affluent Nations; Motherhood: Economic Aspects; Poverty and Gender in Affluent Nations; Poverty, Sociology of; Sex Differences in Pay; Sex Segregation at Work; Social Welfare Policy: Comparisons; Underclass; Welfare Programs, Economics of; Welfare State, History of

Bibliography

Casper L M, McLanahan S S, Garfinkel I 1994 The gender-poverty gap: what we can learn from other countries. *American Sociological Review* **59**: 594–605

Fraser N 1989 *Unruly Practices: Power, Discourses and Gender in Contemporary Social Theory*. University of Minnesota Press, Minneapolis, MN

Fraser N 1994 After the family wage: gender equity and the welfare state. *Political Theory* **22**: 591–618

Hernes H M 1987 *Welfare State and Woman Power*. Norwegian University Press, Oslo, Norway

Hobson B 1994 Solo mothers, social policy regimes, and the logics of gender. In: Sainsbury D (ed.) *Gendering Welfare States*. Sage, London

Jenson J 1997 Who cares? Gender and welfare regimes. *Social Politics* **4**: 182–7

Kessler-Harris A 2001 *The Pursuit of Equity: How Gender Shaped American Economic Citizenship*. Oxford University Press, New York

Lewis J 1992 Gender and the development of welfare regimes. *Journal of European Social Policy* **2**: 159–73

Lewis J, Ostner I 1991 Gender and the evolution of European social policies. Paper presented at the CES workshop on

Emergent Supranational Social Policy, Center for European Studies, Harvard University, Cambridge, MA
Nelson B 1984 Women's poverty and women's citizenship. *Signs* **10**: 209–31
O'Connor J S 1996 From women in the welfare state to gendering welfare state regimes. *Current Sociology* **44**(3): 1–130
O'Connor J S, Orloff A S, Shaver S 1999 *States, Markets, Families: Gender, Liberalism and Social Policies in Australia, Canada, Great Britain and the United States*. Cambridge University Press, Cambridge, UK
Orloff A S 1993 Gender and the social rights of citizenship: the comparative analysis of gender relations and welfare states. *American Sociological Review* **58**: 303–28
Pateman C 1988 The patriarchal welfare state. In: Gutmann A (ed.) *Democracy and the Welfare State*. Princeton University Press, Princeton, NJ
Quadagno J S 1994 *The Color of Welfare*. Oxford University Press, New York
Sainsbury D 1996 *Gender, Equality and Welfare States*. Cambridge University Press, Cambridge, UK
Sainsbury D (ed.) 1999 *Gender and Welfare State Regimes*. Oxford University Press, New York
Skocpol T 1992 *Protecting Soldiers and Mothers:The Political Origins of Social Policy in the United States*. The Belknap Press of Harvard University Press, Cambridge, MA
Williams F 1989 *Social Policy*. Polity Press, Cambridge, UK
Wilson E 1977 *Women and the Welfare State*. Tavistock, London

D. Sainsbury

Social Welfare Policy: Comparisons

Welfare policy comparisons have been a prominent feature of macro-level social science research since the 1970s. Dominated by political scientists and sociologists, most studies focus on the advanced industrial democracies. The literature is guided by two overarching questions: One, how do we explain the rise of welfare states and their diversity? Two, how do variations in policy design affect welfare distributions? The overview that follows will treat the two questions separately.

It has been difficult to arrive at clear-cut answers to either question because of statistical overdetermination: the number of rival explanations often exceeds the number of cases that can be studied (see also *Quantitative Cross-national Research Methods*). There are, moreover, conceptual difficulties because scholars do not always compare and explain the same phenomenon. Some compare social policies (say pensions), others welfare states, and still others welfare regimes, and the difference is rarely made clear. The presence of a set of social policies does not automatically imply a welfare state which, following Marshall (1950) implies social entitlements as a matter of citizenship, comprehensive risk coverage, commitments to full employment, and active reduction of inequalities. The welfare state is also an ideological construct, a means to inculcate broad solidarities and national integration (Briggs 1961, Flora 1986). Rather than speaking of more-or-less welfare statism, the literature has moved towards typologies of kinds of welfare states. Many scholars use the concept of welfare regimes to denote the broader complex, or mix, of social welfare provision between state, market, and families (for a discussion, see Esping-Andersen 1999).

1. Conceptualizing Welfare Policy Models

Early quantitative comparisons, such as Cutright (1965) and Wilensky (1975), compared welfare policy differences through national social spending levels. This was challenged on validity grounds (spending may not reflect welfare effort, and it tells us little about what kind of welfare is being produced and for whom). As a result, research in the 1980s and 1990s has been greatly concerned with the precise meaning and content of social policies and the welfare state. Most conceptualizations begin with Marshall's (1950) emphasis on the rights of social citizenship, and programs are then compared in terms of how strong are rights, to what extent do they lessen market dependency ('decommodification'), and to what degree do they segment or unify citizens (Flora and Alber 1981, Korpi 1983, Baldwin 1990, Esping-Andersen 1990). More recently, reflecting the rise of feminist scholarship, new dimensions, such as the inherent gender bias, or the degree of women 'friendliness,' have been added (Hernes 1987, Sainsbury 1996, Orloff 1993, O'Connor et al. 1999).

The greater conceptual attention to policy content and objectives has led to a keener appreciation of the qualitative differences between national welfare systems. Drawing upon Titmuss (1958, 1974), there emerged a voluminous literature on welfare state typologies, initially focused on welfare *state* models, such as in Furniss and Tilton (1979). As research expanded to the interplay between state, market, and family welfare provision, emerged classifications of welfare *regimes*, such as the liberal, conservative, and social democratic trichotomy in Esping-Andersen (1990) and further refinements, such as Castles' (1993) notion of a unique Antipodean 'wage earners' model,' the identification of a specific, unusually familialistic and patronage-based Mediterranean model (Castles 1995, 1998, Esping-Andersen 1999). Some feminist writings have proposed typologies based on the strength of the conventional male breadwinner assumption (Lewis 1993). With the exception of the latter, most typologies classify nations rather similarly: a Nordic group, distinguished by its universalism, comprehensive coverage, generosity, and a strong pro-women bias; an Anglo-Saxon group, characteristic for

its residual public provision, its targeting of benefits, and its encouragement of private welfare; a Continental European group, which stands out for its corporativistic status-segmented social insurance; and, more controversially, a Mediterranean (and, possibly, East Asian) group primarily distinguished by its strong accent on familial welfare provision. As Castles' (1993) points out, these typologies resemble 'families of nations,' nation-clusters with a deep common historical tradition.

2. Explaining Policy Divergence

The great vitality of comparative welfare state research comes from its address to central theoretical issues in the social sciences. Existing reviews of the literature usually distinguish two broad explanatory thrusts: the convergence thesis that derives from theories of industrialism and modernization, and the 'politics matter' thesis that derives from theories of conflict and power mobilization (Shalev 1983, Esping-Andersen and van Kersbergen 1992, O'Connor and Olsen 1998). In addition, some theorists emphasize the process of nation building and democratization (Flora 1986, Flora and Alber 1981), nations' position in, and response to, the world economy (Cameron 1978, Katzenstein 1985), and also the policy-making role that state bureaucracies exert (Skocpol and Amenta 1986).

2.1 The Modernization Thesis

The earliest systematic welfare state comparisons were chiefly inspired by the theory that industrialization (accompanied by urbanization, the rise of efficient bureaucracy, and by demographic aging) erodes traditional social protection mechanisms, such as the family, and makes necessary public responsibilities. As nations industrialize, they will also converge because social risks become more similar, and because the functional necessity of public social security becomes pressing. Main exponents of this thesis include Rimlinger's (1971) as yet unrivaled historiography of the roots of welfare policy in Europe, America, and Russia, and Wilensky's (1975) quantitative demonstration that rich countries gradually converge in terms of the share of gross domestic product (GDP) spent on welfare. Wilensky provided no direct measure of political variables, and his conclusions were therefore challenged by adherents of the 'politics matter' position.

2.2 The 'Politics Matter' Thesis

The resurgence of conflict theory in the 1970s provoked two antitheses to the modernization paradigm. One was Marxian and quite functionalist, arguing that the modern welfare state was not an egalitarian response to new needs, but rather a mechanism, however contradictory, of sustaining the legitimacy of capitalist economies (Gough 1979, Offe 1972, O'Connor 1973). This perspective has produced little systematic empirical comparison.

The second antithesis argues explicitly that 'politics *does* matter' in the sense that the nature, scope, and objectives of a welfare state depend primarily on the balance of political power between 'left' and 'right.' Since it is well established that the roots of welfare policy are generally conservative, as in the case of Bismarck, this position is mainly concerned with welfare state variations in mature industrial democracies. The premise is that the collective resources of wage earners can, through parliamentary democratic means, be mobilized so as to establish a kind of welfare state that effectively circumscribes the distributional logic of markets and the decisional power of capital (Korpi 1983, Stephens 1979, Castles 1982, Hicks 2000; for an overview, see O'Connor and Olsen 1998). This thesis lies close to classical social democratic theory and its postulate that a parliamentarian transition from capitalism to democratic socialism is possible (Stephens 1979, Esping-Andersen 1990). For this reason, it is often labeled the 'social democratic' model of the welfare state. Since Sweden is the epitome of social democratic power and achievement, the thesis has been criticized for its inherent 'Swedo-centrism' (Shalev 1983).

The power mobilization hypothesis has given rise to a very large but also conceptually and methodologically heterogenous literature which, cumulatively, makes a strong case in favor of the underlying thesis (the implications of methodological design are examined in Janoski and Hicks 1994). Whether studying individual social programs, such as pensions, sickness benefits, or labor market policies (Myles 1984, Palme 1990, Korpi 1989, 1991, Kangas 1991), or more broadly 'welfare states' (Stephens 1979, Korpi 1983, Esping-Andersen 1990, Huber et al. 1993, Hicks 2000), the impact of the political strength of the left is quite robust.

Still, the 'left power' explanation is questioned by several authors working within the 'politics matter' tradition. Some argue against a simple, linear left-power effect, arguing that real politics are premised on enduring party coalitions. Postwar Christian Democracy, for example, has been seen as a functional equivalent to labor movements, producing similarly generous welfare states (Wilensky 1981). Esping-Andersen (1990, 1999), Huber et al. (1993), Castles (1994) and, especially, van Kersbergen (1995) demonstrate, however, that Christian Democracy produces welfare states that may be generous, but which are also strongly corporativistic in design, very 'familialistic,' and transfer-biased. Also Baldwin (1990) questions the simple left-power thesis because what

are usually imputed to be social democratic achievements (such as universalism) arose from entirely different political quarters (agrarians in particular). In a unique synthesis of the modernization and politics-matter theses, Pampel and Williamson (1989) suggest that all governments, left or right, will favor welfare expansion because of population and, thus, median-voter aging.

2.3 The Impact of Globalization

Since there is a strong coincidence between a country's size, its degree of economic openness, and the scope of its welfare state, the correlation between left power and welfare statism may be causally spurious (Cameron 1978, 1984, Katzenstein 1985). The point is that small, export dependent economies are prone to international shocks and need to be adaptive. This favors the emergence of centralized, consensual social pacts between the social partners, but also strong social guarantees to workers in order to assure their compliance.

Put in this way, the causal link between power and welfare state outcomes becomes ambiguous. It may be that powerful labor movements *and* generous welfare states both derive from global economic pressures. It is difficult to sort the real causal order because the number of countries we can compare is so few.

2.4 Nation-building and Democratization

The puzzle that any 'politics-matter' explanation faces is that the first social reformist steps were usually taken by conservative, sometimes anti-democratic forces, often with the explicit intent of containing working class mobilization (Rimlinger 1971, Flora and Alber 1981). Social policy *per se* cannot therefore be a robust indicator of left-political achievements. The issue is really one of historical sequence. Most would agree that the historical roots of social policy are closely connected to late nineteenth-century nation building, a means to sustain the established order in the face of emerging social, linguistic, ethnic, or religious cleavages. The 'politics-matter' paradigm, in contrast, is concerned with the epoch of actual welfare state construction in the postwar era of full democracy.

2.5 The State-centered Thesis

Modernization theory stresses the functional necessity of efficient bureaucracy for welfare state development because, without it, effective taxation and administration of complex distributional programs is simply impossible. Bureaucracy, once erected, may also come to play an active role in decision-making and thus determine policy evolution. Heclo (1974) maintained that Swedish and British welfare policy, although originally rather similar, eventually parted ways because of inherently different bureaucratic policy-making styles and traditions. The independent powers of public officials in framing key decisions have also been emphasized especially by Skocpol (1992). This, so-called state-centered thesis has proven difficult to test empirically against rival arguments. As a result, empirical applications are limited to select case studies (such as Skocpol 1992) that are difficult to generalize, and a handful more systematic, quantitative studies (Korpi 1989).

3. The Consequences of Welfare States

The literature on the consequences of welfare state differences is, possibly, as huge as the former but the protagonists tend to be different. For one, there is a longstanding tradition in economics to study the micro-economic effects of social policies on work incentives as well as its macro-economic effects on inflation, savings, unemployment, or economic growth. This literature, which, in any case, is rarely comparative, will not be reviewed here (for an overview, see Danziger et al. 1981, Atkinson and Mogensen 1993, and Barr 1993).

Closer to our concerns are studies that examine the impact of welfare state design on distributional outcomes. Most scholars agree that the focus should be on life chance dynamics, but this has been thwarted by data scarcity. The literature has therefore concentrated on cross-sectional analyses of income inequalities, poverty and, in recent years, also on gender equality.

3.1 Poverty and Income Distribution

The conventional assumption has been that equality is positively and linearly related to welfare state generosity. This was implicit in Wilensky (1975), and some early attempts at empirical verification produced positive and significant correlations (Hewitt 1977). Until the mid-1980s, however, robust comparisons were hampered by severe problems of cross-national data comparability. Only with the advent of the Luxembourg Income Study (LIS) and, later, European Union data collection, did genuinely comparable income data become available.

Most studies of the redistributive effects of welfare states (or single social programs) are based on a comparison of pre- and post-transfer/tax distributions (a representative example is Smeeding et al. 1990; an extensive review of the literature is provided by Gottschalk and Smeeding 1997). A few studies attempt to study more explicitly the impact of *types* of welfare states or social policies on distributional outcomes (Ringen 1987, McFate et al. 1995).

Quantitative comparisons come to a surprisingly consensual view: the larger the welfare state, the greater the degree of equality in net disposable income,

and the lower the rates of poverty. A less obvious finding is that welfare states targeted to the truly needy do not result in more poverty reduction than does the universalistic approach to social protection (for a review, see Gottschalk and Smeeding 1997). Reflecting the paucity of comparable panel data, studies on poverty and inequality *dynamics* are, so far, few. Preliminary evidence does, however, suggest that comprehensive, generous welfare states are substantially more able to guarantee against long-term entrapment in poverty or welfare dependency (Burkhauser and Poupore 1993, Duncan et al. 1993). Most studies conclude that targeted, means-tested policies create strong poverty traps.

The finding that comprehensive, universalistic welfare states are more redistributive than those favoring heavy targeting has provoked debate about the precise mechanisms involved. Some argue that income tests create stigma and lower the take-up rate. With their notion of the 'paradox of redistribution,' Korpi and Palme (1998) argue that the basic reason lies in crowding out of private welfare that ensues from comprehensive, universalistic policy: limited welfare states nurture private welfare plans which are extremely inequality producing.

3.2 Gender Inequality

The study of welfare policies and gender inequalities is primarily focused on family-related social policies, such as caring services and leave schemes; what Bradshaw et al. (1993) call the family benefit package. Specific policy variations are linked to either welfare state typologies (as in the case of Gornick et al. 1997, Gustafsson 1994), or to male-breadwinner typologies (Lewis 1993, Mclanahan et al. 1995, Sainsbury 1996), and welfare systems are often ranked in terms of their degree of 'women friendliness' (for an overview, see O'Connor 1996). A number of rather clear-cut empirical conclusions have emerged. One, day care provision is a powerful precondition for women's economic independence, for their capacity to harmonize careers and family, and also for safeguarding against poverty in single parent families (Gornick et al. 1997, Gustafsson 1994). A second finding is that welfare systems that grant women independent entitlements both lessen dependency and augment gender equality (Mclanahan et al. 1995, Sainsbury 1996; O'Connor et al. 1999).

See also: Policy History: State and Economy; Social Security; Social Welfare Policies and Gender; Welfare; Welfare: Philosophical Aspects; Welfare Programs, Economics of; Welfare State; Welfare State, Geography of; Welfare State, History of

Bibliography

Atkinson A, Mogensen G V (eds.) 1993 *Welfare and Work Incentives*. Clarendon Press, Oxford, UK

Baldwin P 1990 *The Politics of Social Solidarity. Class Bases of the European Welfare State 1875–1975*. Cambridge University Press, Cambridge, UK

Barr N 1993 *The Economics of the Welfare State*. Stanford University Press, Palo Alto, CA

Briggs A 1961 The welfare state in historical perspective. *Archives Europeennes de Sociologie* **2**: 221–58

Burkhauser R, Poupore J 1993 A cross-national comparison of permanent inequality in the United States and Germany. *Cross-National Studies in Ageing*. Program Project Paper no 10. State University of New York, Syracuse, NY

Cameron D 1978 The expansion of the public economy: A comparative analysis. *American Political Science Review* **72**: 1243–61

Castles F G 1982 The impact of parties on public expenditure. In: Castles F G (ed.) *The Impact of Parties*. Sage, London, pp. 21–96

Castles F G 1993 *Families of Nations: Patterns of Public Policy in Western Democracies*. Aldershot, Dartmouth, UK

Castles F G 1994 On religion and public policy: Does Catholicism make a difference? *European Journal of Political Research* **25**: 19–40

Castles F G 1995 Welfare state development in Southern Europe. *West European Politics* **18**: 291–313

Castles F G 1998 *Comparative Public Policy*. Elgar, Cheltenham, UK

Cutright P 1965 Political structure, economic development and national social security programs. *American Journal of Sociology* **70**: 537–50

Danziger S, Haveman R, Plotnik R 1981 How income transfers affect work, savings, and income distribution. *Journal of Economic Literature* 19

Duncan G, Gustavsson B, Hauser R 1993 Poverty dynamics in eight countries. *Journal of Population Economics* **6**: 215–34

Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Polity Press, Cambridge, UK

Esping-Andersen G 1999 *Social Foundations of Postindustrial Economies*. Oxford University Press, Oxford, UK

Esping-Andersen G, van Kersbergen K 1992 Contemporary research on social democracy. *Annual Review of Sociology* **18**: 187–208

Flora P 1986 Introduction. In: Flora P (ed.) *Growth to Limits, Vol. 1*. De Gruyter, Berlin

Flora P, Alber J 1981 Modernization, democratization and the development of welfare states in Western Europe. In: Flora P, Heidenheimer A J (eds.) *The Development of Welfare States in Europe and America*. Transaction Books, New Brunswick, NJ

Furniss N, Tilton T 1979 *The Case for the Welfare State*. Indiana University Press, Bloomington, IN

Gauthier A 1996 *The State and the Family*. Clarendon Press, Oxford, UK

Gornick A, Meyers M, Ross K E 1997 Supporting the employment of mothers. *Journal of European Social Policy* **7**: 45–70

Gottschalk P, Smeeding T 1997 Cross-national comparisons of earnings and income inequality. *Journal of Economic Literature* **35**: 633–81

Gough I 1979 *The Political Economy of the Welfare State*. Macmillan, London

Gustafsson S 1994 Childcare and types of welfare states. In: Sainsbury D (ed.) *Gendering Welfare States*. Sage, London
Heclo H 1974 *Modern Social Policies in Britain and Sweden*. Yale University Press, New Haven, CT
Hernes H 1987 *The Welfare State and Women Power*. Norwegian University Press, Oslo, Norway
Hewitt C 1977 The effect of political democracy and social democracy on equality in industrial societies. *American Sociological Review* 42
Hicks A 2000 *Social Democracy and Welfare Capitalism*. Cornell University Press, Ithaca, NY
Hicks A, Swank D 1992 Politics, institutions, and welfare spending, 1960–1982. *American Political Science Review* **86**(3): 658–74
Huber E, Stephens J, Ragin C 1993 Social democracy, christian democracy, constitutional structure and the welfare state. *American Journal of Sociology* **99**: 711–449
Janoski T, Hicks A (eds.) 1994 *The Political Economy of the Welfare State*. Cambridge University Press, Cambridge, UK
Kangas O 1991 *The Politics of Social Rights*. Swedish Institute for Social Research, Stockholm, Sweden
Katzenstein P 1985 *Small States in World Markets*. Cornell University Press, Ithaca, NY
Korpi W 1983 *The Democratic Class Struggle*. Routledge, London
Korpi W 1989 Power, politics and state autonomy in the development of social citizenship. *American Sociological Review* **54**: 309–28
Korpi W 1991 Political and economic explanations for unemployment. *British Journal of Political Science* **21**: 315–48
Korpi W, Palme J 1998 Redistribution and strategies of equality in Western countries. *American Sociological Review* **5**: 661–87
Lewis J 1993 *Women and Social Policies in Europe*. Elgar, Aldershot, UK
McFate K, Lawson R, Wilson W J 1995 *Poverty, Inequality, and the Future of Social Policy*. Russell Sage, New York
Mclanahan S, Casper L, Sorensen A M 1995 Women's roles and women's poverty. In: Oppenheimer V, Jensen A (eds.) *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK
Marshall T H 1950 *Citizenship and Social Class*. Oxford University Press, Oxford, UK
Myles J 1984 *Old Age in the Welfare State*. Little, Brown, Boston
O'Connor J 1973 *The Fiscal Crisis of the State*. St Martin's Press, New York
O'Connor J 1996 From women in the welfare state to gendering welfare state regimes. *Current Sociology (special issue)* **44**
O'Connor J, Olsen G 1998 *Power Resources Theory and the Welfare State*. University of Toronto Press, Toronto
O'Connor J, Orloff A, Shaver S 1999 *States, Markets, Families*. Cambridge University Press, Cambridge, UK
Offe C 1972 Advanced capitalism and the welfare state. *Politics and Society* **4**: 479–88
Orloff A 1993 Gender and the social rights of citizenship. *American Sociological Review* **58**: 303–28
Palme J 1990 *Pension Rights in Welfare Capitalism*. Swedish Institute for Social Research, Stockholm, Sweden
Pampel F, Williamson J 1989 *Age, Class, Politics and the Welfare State*. Cambridge University Press, Cambridge, UK
Rein M, Rainwater L 1986 *The Public-Private Mix in Social Protection*. M E Sharpe, Armonck, New York
Rimlinger G 1971 *Welfare and Industrialization in Europe, America, and Russia*. Wiley, New York
Ringen S 1987 *The Politics of Possibility: A Study in the Political Economy of the Welfare State*. Clarendon Press, Oxford, UK
Sainsbury D 1996 *Gender, Equality and Welfare States*. Cambridge University Press, Cambridge, UK
Shalev M 1983 The social democratic model and beyond: Two generations of comparative research on welfare states. *Comparative Social Research* **6**: 315–51
Skocpol T 1992 *Protecting Soldiers and Mothers*. Harvard University Press, Cambridge, MA
Skocpol T, Amenta E 1986 States and social policies.*Annual Review of Sociology* **12**
Smeeding T, O'Higgins M, Rainwater L 1990 *Poverty, Inequality and Income Distribution in Comparative Perspective*. Harvester/Wheatsheaf, New York
Stephens J 1979 *The Transition from Capitalism to Socialism*. Macmillan, London
Titmuss R 1958 *Essays on the Welfare State*. Allen and Unwin, London
Titmuss R 1974 *Social Policy*. Allen and Unwin, London
Van Kersbergen K 1995 *Social Capitalism*. Routledge, London
Wilensky H 1975 *The Welfare State and Equality*. University of California Press, Berkeley, CA
Wilensky H 1981 Leftism, catholicism and democratic corporatism. In: Flora P, Heidenheimer A (eds.) *The Development of Welfare States in Europe and America*. Transaction Press, New Brunswick, NJ

G. Esping-Andersen

Socialism

A workable definition of Socialism, subject to the qualifications offered below, would be this: 'Socialism is an economic organization of society in which the material means of production are owned by the whole community and operated by organizations that are representative of, and responsible to, the community according to a general economic plan, all members of the community being entitled to benefit from the results of such socialized planned production on the basis of equal rights' (Lichtheim 1971, p. 113).

Leszek Kolakowski's survey (1978) calls attention to 'the surprising diversity of views expressed by Marxists in regard to Marx's so-called historical determinism.' However, what is 'surprising,' is not this diversity—which becomes much wider if we examine not Marxism in particular, but socialism in general. The surprising thing is Kolakowski's pontifical belief that faced with it he could 'schematize with precision the trends of twentieth century Marxism' (1978, Vol. I, pp. 6–7). His failure to attain such precision serves as a salutary warning to anyone attempting to characterize the broader and much more 'diverse' terrain of socialism, of which Marxism—nineteenth and twentieth century Marxism—is itself but a 'main current.' In what follows, which will (in Wittgenstein's words) 'strive *not* after exactness, but after a synoptic view,' attention will be paid to Cole's admonition: that 'the most that can be attempted is the discovery of some

central core of meaning, present with varying additions in all or most of the manifold uses of the word (socialism)' (Cole 1967, Vol. I, p. 1). This entails sensitivity to the shifts in meaning socialism has undergone since its inception; only beneath these shifts can a substratum of meaning emerge.

1. Capitalism, Socialism and Democracy

The word 'socialism' first appeared in Robert Owen's *Co-operative Magazine* in 1827, '*le socialisme*' in the Saint-Simonian journal *Le Globe* in 1832. It meant a projected alternative sociopolitical system—this was still offered as a definition of 'socialism' in the 1933 OED—not a political movement or demand. Early socialist doctrine contained little about the proletariat, the class system, labor, or revolution. While John Stuart Mill's *Principles of Political Economy* accordingly distinguished socialism and communism as theories, not political movements, Marx and Engels's *Manifesto of the Communist Party* of the same year (1848) identified communism as a movement that would render socialism, so defined, beside the point because it was merely a theory—utopian, that is because it was impracticable. It was communism, not socialism, that carried with it the idea of revolutionary struggle and human agency; a new and better society could not be wished or legislated into existence by benevolent doctrinaires, but had to await the advent of a politically conscious, class-based labor movement. This was a key shift in the coordinates, not just of communism, but of socialism. It too had now to be, not just thinkable, but feasible. After the failure of the 1848 revolutions, communism ceased for many years to be practical politics, yet some time was to pass before Marx and Engels consented to having their cause described as 'socialist' (Lichtheim 1969, p. 209). They did so because socialism had become not just a doctrine but a movement. Thereafter, 'Marxism created that distinctively German socialism which was soon to assume an ideological dominance over most of the continent, driving the older forms of socialism before it as chaff before the wind' (Cole 1967, Vol. I, p. 222).

The second key shift in the coordinates of European socialism occurred in 1917, when, against all the odds, communism in its Leninist form arose as a clear alternative to democratic socialism—the same democratic socialism that had itself, against all previous odds, survived World War One. Both authoritarian communism and democratic socialism laid claim to a lineage deriving from Marx; but they had little else in common. Twentieth-century communists showed no reluctance to clothe their particular doctrines in the broader mantle of socialism. The locutions, 'Union of Soviet Socialist Republics,' 'socialism in one country,' 'socialist realism,' 'socialist internationalism' (in its twentieth-century Soviet version, which meant subservience to the foreign policy of the USSR), the fact that after World War Two the East German Communist Party was officially called the 'Socialist Unity Party'—all these are instances of synecdoche, if not the sheerest verbal trickery. We should not allow them to addle us into thinking, languidly, that all twentieth-century communists were socialists, whereas not all socialists were communists. There were in the twentieth century not just non- but also anti-communist socialists aplenty, socialists actively and vociferously opposed to communism, from the very moment when such opposition first became possible, at the Berne International in February 1919, when adherence to democracy was reasserted as a prerequisite for socialism (Cole 1967, vol. IV, pp. 292–9). The socialism that survived World War One to become part of the European political landscape had had as its prerequisite not democracy but capitalism. Socialists had long been in the forefront of the struggle for democracy. (Indeed, socialism was weakest in those countries, the US and the UK, where it was a less powerful force in the struggle for democracy than it was on the European continent.) In Cole's words, Liebknecht, Bebel, and Kautsky in Germany, Guesde in France, indeed, most socialist stalwarts in Europe before World War One 'felt that they had no *right* to make the revolution without the backing, or at least the assent, of the people. They thought of the proletariat as being, or as in the process of becoming, the majority of the people; and they envisaged the mass conversion of the proletariat to the socialist cause as a necessary preliminary to the revolution' (Cole 1967, Vol. III, pp. 946–7). This democratic perspective was not necessarily reformist or purely parliamentary; democratization was also meant to encompass work relations and trade union structures, and could scarcely have avoided doing so. What the Second International socialists rightly feared was war; the solitary success of Russian Bolshevism in 1917 was as little expected or forseen as its eventual collapse was to be in 1989.

2. Marx, Marxism and Socialism

Marx casts a longer shadow over the history of socialism than anyone else—not because every socialist agreed (or was even very familiar) with what he said, but because, all the same, his presence proved impossible to ignore. It is in this sense that Marxism can be called the *lingua franca* of socialism (Cole 1967, vol. I, p. 379). Without Marx, we would still have had revolution as a word and as a concept, as the French and the Americans had put revolutionary change on the political agenda during the eighteenth century, and we would have had capitalism, socialism, communism, and (thanks to Babeuf) the proletariat too. The

celebrated opening line of the *Manifesto of the Communist Party* was designed to call the reader's attention to 'the specter haunting Europe,' to invoke, not invent, communism. It is no doubt easier to imagine a world without Marx than a world without revolution, capitalism, communism, and socialism. But in the world we actually inhabit, all these still have to be seen through Marx. He may not have coined these terms, but he decisively set his seal on them. They still cannot properly be discussed without bringing him in. Marx contributed enormously to the vocabulary of socialism (class, class struggle, class war, class consciousness, ideology, alienation, the commodity, forces, relations, and modes of production, exploitation, contradiction, and crisis). Even though 'socialist and communist movements (took) up some aspects of Marx's theory and discarded others, even this selectivity (was) carried out in Marx's name' (Gilbert 1981, p. 270).

Yet much that was carried out 'in Marx's name' was attained without much knowledge of what he had written. The revisionism debate (at the turn of the twentieth century) that almost tore apart German Social Democracy and the Second International, was about Marxist 'orthodoxy,' but was conducted in an astounding ignorance of Marx's writings. While practicing politicians commonly regard themselves as having more immediately pressing concerns than questions of textual fidelity, these same questions were never ignored among Marxists—although the theoretical corpus to which socialist stalwarts appealed was not necessarily made up of Marx's writings. The problem was not that a canon of basic texts failed to emerge. (To the contrary, textual fidelity was adhered to tirelessly—and, at times, tiresomely.) It was that the canon as it developed was not so much derived from what Marx had written as it was constructed around what little of Marx was available in published form. Historical materialism—it still amazes people when they learn that this term was not Marx's own—had its basic texts long before the first appearance in any language (during the 1930s) of so basic a text as Marx and Engels's *The German Ideology*. These texts included Franz Mehring's *On Historical Materialism* (1893), G. V. Plekhanov's *The Development of the Monist View of History* (1895), Antonio Labriola's *Essays on the Materialist Interpretation of History* (1896) and Karl Kautsky's *The Materialist Interpretation of History* (1927). Even though Marx in some sense inspired these books and others like them, what he had to say eventually needed to be disentangled from their arguments. It could in no way be directly encountered through them or explained by them (Thomas 1991, pp. 32–3).

3. *The Varieties of Socialist Experience*

Even if Marx was more known about than known, the fact remains that by the end of the nineteenth century a politically-oriented, class-conscious European socialist movement had emerged in his name; and by this time the word socialist meant what it still means today, a social and political doctrine and a political movement or system. In view of this latitude of meaning it is scarcely surprising that 'socialism' has come down to us in social science literature as a noun qualified by an adjective—as in 'utopian socialism,' 'scientific socialism,' 'state socialism,' 'revolutionary socialism,' 'evolutionary socialism,' or (until recently) as 'actually existing socialism,' or again (today) as 'market socialism.' Similarly, the adjective 'socialist' can be employed as a pendant qualifying or characterizing an abstract noun, as in 'socialist internationalism,' 'socialist economics,' 'socialist realism,' or 'socialist feminism.'

The above-listed collocations—the inventory is illustrative, not exhaustive—are not just convenient subdivisions, but serial attempts at spot-welding that can readily be arrayed in historical sequence. 'Utopian socialism,' as we have seen, dates from the 1830s and 1840s, 'Fabian' or 'evolutionary socialism' dates from the late nineteenth and early twentieth centuries, 'Guild socialism' dates from the period between the two World Wars, and 'actually existing socialism' dates from the latter days of Soviet communism. 'Socialist internationalism' peaked, then bottomed out dramatically, in the period before 1914; 'socialist realism' (like 'socialism in one country') belongs to the Stalinist period in the Soviet Union; 'socialist feminism' to Western scholarship of a later date. To array these composite terms historically in order to arrive at the meaning of their least common denominator, socialism, is not to pigeonhole them arbitrarily, nor yet to proceed reductively, if it is borne in mind throughout that each instance of spot-welding affected the coordinates of its successors.

More specifically, anarchism, syndicalism and their offshoots (whether revolutionary or not) are kinds of socialism that were defined and elaborated against state socialism (Thomas 1980); Engels in 1875 contrasted scientific with utopian socialism (even if the latter was largely a spent force, the former was to enjoy some shelf-life well into the twentieth century); Lenin defined Bolshevism, then communism as he understood the term, against the parliamentary, reformist socialism of the pre-1914 German SPD (and by extension the Second International); this reformist socialism had itself been elaborated in contradistinction to the revolutionary socialism that appeared to have had its come-uppance at the time of the Paris Commune (1871). Lenin's successors proceeded to counterpose socialism in one country against the socialist internationalism the First and Second Internationals had espoused. Overall, these developments do not add up to a tidy, formulaic picture. Yet there is a subtle patterning in the carpet. Positions and tendencies throughout the history of socialism were debated, articulated and defined in relation to other, pre-existing socialist positions and tendencies.

See also: Marx, Karl (1818–89); Marxian Economic Thought; Marxism and Law; Marxism/Leninism; Marxist Social Thought, History of; Social Democracy; Social Security; Social Welfare Policy: Comparisons; Socialism: Historical Aspects

Bibliography

Bottomore T B 1993 Socialism. In: Outhwaite W, Bottomore T B (eds.) *The Blackwell Dictionary of Twentieth Century Political Thought*. Blackwell, Oxford, pp. 617–9

Braunthal J 1966 *History of the International 1864–1914*, 2 Vols. Collins H, Mitchell K (trans.). Nelson, London

Cole G D H 1967 *A History of Socialist Thought*, 5 Vols. Macmillan, London

Gilbert A 1981 *Marx's Politics. Communists and Citizens*. Rutgers University Press, New Brunswick, NJ

Kolakowski L 1978 *Main Currents in Marxism*, 3 Vols. Falla P S (trans.). Clarendon Press, Oxford, UK

Landauer C 1959 *European Socialism: A History of Ideas and Movements from the Industrial Revolution to Hitler's Seizure of Power*, 2 Vols. University of California Press, Berkeley, CA

Lichtheim G 1961 *Marxism: An Historical and Critical Study*. Praeger, New York

Lichtheim G 1969 *The Origins of Socialism*. Praeger, New York

Lichtheim G 1971 *A Short History of Socialism*. Praeger, New York

Schorske C E 1955 *German Social Democracy 1905–1917. The Development of the Great Schism*. Harvard University Press, Cambridge, MA

Thomas P 1980 *Karl Marx and the Anarchists*. Routledge and Kegan Paul, London

Thomas P 1991 Critical reception: Marx then and now. In: Carver T (ed.) *The Cambridge Companion to Marx*. Cambridge University Press, Cambridge, UK

P. Thomas

Socialism: Historical Aspects

1. The Meaning of a Group of Ideas

The term socialism designates a set of doctrines which originated in the nineteenth century and intended to criticize the actual social organization and values (capitalism, bourgeois society, individualism) in terms of equality. Socialist ideologies contributed to the formation of a cultural world, linked mainly with the lower classes, composed of political parties and other associations which had as their goals projects of future societies and programs calling for collective action. In fact, the term socialism must comprehend a certain number of doctrines, political opinions, ideological vocations, and feelings of rebellion and emancipation. Today, the way of considering this problem is far from the acceptance of a hierarchy and of a teleological vision, which was established essentially by the Marxist tradition and defined the history of socialism as a successive progression from 'precursors' and 'utopians' to a more and more 'mature' conception of socialism.

Socialism, in its origins, both from a linguistic and conceptual point of view, was part of that rich patrimony of ideas which originated with the French revolution and at the same time was linked to the social experiences evolving from the industrial revolution. During the nineteenth century, in continental Europe, the term was also used to define social and economic doctrines advocating an intervention of the State in the economy and/or a change in the regime of property from the individual to the collective property of the means of production. In the twentieth century, the socialist movement was widely permeated by political conflicts on how to achieve a socialist society (reforms or revolution) and also regarding what kind of model such a society should be based on. Marxism was the most influential doctrine in the constellation of socialism, both for the role it played in the history of socialist culture and in the contribution it gave to the creation of socialist political movements. The Russian revolution claimed to have created 'socialism,' and in spite of the criticism regarding the lack of democracy in this realized socialism, the idea continued to spread from Europe to other continents. Between 1917 and 1989, the term was also used to designate an actual social and economic system. According to the American Marxist Paul M. Sweezy (1949), 'the term "socialism" means a social system which is differentiated from other social systems by the character of its property relations. Thus socialism properly belongs to the same species as capitalism and feudalism.'

There is a general agreement regarding the difficulty of defining and the complexity of the term socialism. This depends on the fact that the word has always designated an overall group of very complex ideas, that these ideas have been subjected to changes in time and space, and that by placing the history of the word in a crossroads of intense ideological polemics it has often changed its meaning. From the word used by Pierre Leroux in France during the 1830s to define an anti-individualistic political and social doctrine close to saintsimonism, up until the socialist ideology of the New Left in the 1970s, the road this term and concept have followed has been long and tortuous.

2. The Origins of Socialism

When, between 1847 and 1848, Marx and Engels wrote the Communist Manifesto, the majority of their contemporaries would not have been able to precisely define the difference between socialism, communism and democracy. Even though it was the democracy which smelled 'of blood and barricades' (Rosenberg 1962), the upper classes perceived the different socialist

'schools' as a threat to property and order. These 'schools' had a common root in the political thought and in the social experience derived from the French and the industrial revolutions. At the same time, however, they were their most severe critic. In France and England, the term socialism was used for the first time in the 1820s and the 1830s essentially to designate economic orientations contrary to individualism. In the period preceding 1848 various political and cultural centers appeared which could be traced to similar trends of thought.

In France, the first equalitarian ideas after the Revolution came from the Jacobin left and are those of F. N. Babeuf, which are known to us above all through the literary works of F. Buonarroti. Babeuf tried to achieve his ideals (elimination of private property and the establishment of a society based on community property) by means of an attempted insurrection which failed. Later on, in the midst of political reflections after the French revolution and the experiences of the Napoleonic era, C. H. de Saint-Simon imagined a future society based on science and work. His ideals influenced utopian French socialism affirming an optimistic confidence in technical and scientific progress. Fourier, who was also a critic of contemporary society, proposed the construction of a harmonious society organized around the Phalanstères which were kinds of self-supporting communities promoted by free and happy men, who were able to achieve equality between the sexes and to satisfy any material necessity.

In England, differently from in France, the new doctrines were diffused in a framework of relative legitimacy. Robert Owen, who was at the same time a utopian socialist and a social reformer, tried to build communities based on the collective property of the means of production and on noncompetitive social relationships. He also succeeded in creating workers' organizations based on the same ideals, which contributed to the formation of cooperatives and unions. It was the first time—and this experience will eventually have consequences in the Chartist movement—that socialist ideas became integrated with real workers.

Both in France and in England, the 1830s marked a change in the political climate and in the panorama of social conflicts. The artisans, impoverished by the processes of industrialization, and the alienated wage workers living on the borderline of a survival level were sensitive to the ideas of socialism which had circulated before the revolutions of 1848. It was a type of socialism containing strong elements of solidarity, cooperation, emancipation, and social justice. The experiences of the previous years, thanks to the influence of these first groups of organized workers, were able to put together a first set of common ideas: unjust profit, equal salary, the presence of two worlds in opposition or—as Fourier said—of an 'upside-down world' in which a small number of privileged people were opposed to a multitude of those condemned to unrewarding work.

The spontaneous struggles of the workers, which in France both preceded and followed the July Monarchy, led to a politicization which exploited the socialist theories. Auguste Blanqui was a central figure in this context. Influenced by the theories of Saint-Simon, Fourier, and Babeuf, Blanqui was above all a conspirator: he not only created several secret societies of republican-socialist inspiration, but he was also a tireless revolutionary activist. From certain points of view, Blanqui embodies the prototype of the 'professional revolutionary' who believed that the victory of the proletariat could be reached by a determined minority ready to conquer the power and use dictatorial methods. At the same time Proudhon was one of the first to underline the libertarian aspects in the criticism of capitalism, and in his essay *What is Property* (1840), he condemned in particular the oppressive function of the state and private property. In England, between 1837 and 1848, the Chartist movement developed as a reaction to the electoral reform of 1832 and the enforcement of the Poor Law which had further worsened the living standards of the working classes. The People's Charter, which constituted the basic document of Chartism, claimed the right to universal male suffrage and considered a just electoral reform as the principal instrument for the achievement of social reforms in favor of the workers. Though Chartism was harshly repressed and completely disappeared after 1848, it constituted an important experience because of the connection which was established between social struggles and politics in that decade.

This connection had its most profound roots in the changes which took place in the lives of the lower classes between the end of the eighteenth and the beginning of the nineteenth century. In France during the Revolution the *sansculottes* were mostly concerned with the problems of subsistence, instead now there was a different attention mainly focused on the social dimensions of work, wages, and on the opposition between the bourgeoisie and the proletariat. At the same time in England the social changes brought about by the industrial revolution stimulated a connection with the radical traditions of the past in the new social reformers. In both cases ideas of emancipation, equalitarianism, and restoration of a usurped legality were diffused among the lower classes; these popular ideas coexisted and at times integrated extremely well with the works of intellectuals and political leaders. However, socialism in its entire existence was not able to offer a unanimous solution to the problem regarding the ways through which a more just society could be achieved. In fact, socialists in the period preceding 1848 were often romantic and bizarre figures (Hobsbawm) who relied more on the intrinsic rationality of their ideas rather than on a collective action to bring about the realization of their projects.

3. Marx and Engels

The revolutions of 1848 constituted a very important chronological fracture in the history of socialism. In the first place because the so called 'social question' was brought to the attention of the public due to the intense participation of socialist workers and artisans in the social struggles in the spring of that year, and particularly in the 'June Days' in France. This situation contributed to transforming the social question into a political one and it began a process in which socialists and communists became more and more different from democrats and republicans. In the second place, a shift began which later would be called a swing of the center of gravity in the development of socialism from England and France towards Germany. In preceding years the German culture had not been completely absent from the discussions regarding socialism; however, differently from other countries, intellectual reflections were isolated from the experiences of the working class.

Now, between 1847 and 1848, along a tortuous road which brought them from the study of German classical philosophy to the Bund der Kommunisten, Marx and Engels published a brief pamphlet which would come to represent a milestone in the history of socialism, the *Communist Manifesto*. In this pamphlet, political activism was combined with an analysis of the social relationships created by the development of capitalism. The *Manifesto* defined the ways through which the bourgeoisie and proletariat became social classes irremediably opposed not only on the basis of moral choices, but above all on the basis of economic and social laws which dominated the history and the existence of mankind. Communists were gifted with the scientific awareness of the development of human society and were able to guide the proletarians in acquiring their class consciousness and in achieving a classless society in which the exploitation of men against men would be eliminated. Marx and Engels not only affirmed a 'scientific' conception of socialism, they also intended to close the doors on that world of socialist ideas which had preceded them. In the third part of the *Manifesto*, they made a distinction between reactionary socialism, conservative or bourgeois socialism, and critical-utopian socialism and communism. What Marx and Engels accomplished was the creation of an arrangement of other nuclei of the socialist culture in a historical perspective in which their 'scientific' doctrine overcame the preceding socialist traditions and ended up by expressing a negative criticism on half a century of socialist culture. Instead, the importance of the old socialism actually consisted in its capacity to exploit other cultural and political trends circulating among the lower classes: anticonformism, dissent, radical jacobinism, and democracy. Without this cultural atmosphere, Marx and Engels probably wouldn't have had such a large public following. If it's true, according to an intuition of Durkheim, that socialism was not only a 'miniature sociology' but also a 'cry of pain,' it's just as true that its roots should be sought for not only in the traditions of 'scientific socialism' but also in the culture—e.g. in the beliefs, in the feelings, in the religion—of the lower classes and of the rank and file militants.

In the 20 or 30 years following 1848, the doctrine of Marx and Engels became 'marxism.' Marx published the first book of the '*Capital*' in 1867, and both Engels and he were engaged in an intense journalistic activity which accompanied the development of a first international nucleus of an organized workers movement. It's the so-called First International, inside of or next to which, new groups and doctrines came about in conflict with those of Marx and Engels. One of these groups were the anarchists, followers of Bakunin in the rural areas of Latin and Slavic Europe; others were the Russian populists, and both showed the extent of the influence exerted by the International Association of Workers; and their activism also showed the richness and complexity of socialism. In March 1871, groups of internationalists with different political origins gave birth to the Paris Commune, which Marx and Engels considered a first experience of the proletarian dictatorship. Around that date, even though a bloody repression concluded the experience of the Paris Commune, nuclei of political subjects were formed in all of Europe and through a complicated game of conflicts and unification would eventually give origin to the different socialist parties. Principally in France and Germany the working class movements will become efficient bearers of the socialist culture and will stimulate emulation in groups and organizations in other countries.

4. The Diffusion of Socialism

Beginning in 1889, the process of the formation and expansion of socialist parties as well as an incredibly varied body of organizations, associations, unions, and cooperatives which accompany them assumed a rapid and intense rhythm. A type of socialism developed in which Marxism consolidated its central role. However, it was a type of Marxism remote from the cold reasoning of Marx and Engels. Though Karl Kautsky, principal exponent in this phase of the expansion of Marxism contributed in defining it as 'scientific socialism,' it was an ideology not far from the mechanistic features of positivistic common sense. The idea of 'natural necessity' (*Naturnotwendigkeit*) so widely present in the Marxist texts and documents which exalted the automatism in the transformation of the means of production, influenced the mentality of rank and file giving them the certainty of the breakdown of the capitalistic system and the birth of socialism. The 'scientific' aspect of this conception of the past, present, and future of the society ended up by creating a certain 'rationality of history.' Thanks to

this faith in the future, the rank and file—according to Gramsci—think that they are only 'momentarily defeated,' but that 'the force of things will work for them in the long run.' Socialism became the ideology of the socialist parties and the mixture of its components became even more enriched. According to H.-J. Steinberg, along with Marxism other components must be considered as a part of the socialist ideology of the German Social Democracy: these were, pragmatism, anti-intellectualism, social Darwinism, the idea of the predictability of the final victory, and other ideas which could be found on uncertain grounds between the common sense of the lower classes and the vulgarization of Marxism.

Socialism ended up by assuming a worldwide dimension. The Socialist International (which Lenin later called the *Second* International, when he was creating the *Third*) was not only the association where general political questions were discussed or agreed upon (participation or non participation in left wing bourgeois government coalitions, to call or not to call for a general strike, etc.), but it was also the place where other aspects of the socialist mentality and culture such as internationalism, pacifism, and solidarity were expressed. It is difficult to distinguish between the diffusion of socialist ideas and the force of attraction exerted by the major socialist parties. Certainly the socialist or social democratic parties were the most efficient means for the circulation of socialist ideas in the period of the Second international. But also the ways political cultures and ideologies reacted against socialism might have played a role in the process of the diffusion of socialism and Marxism. It was in this period that the integration of socialist parties in the European political systems took place and at the same time a complex cultural effort aimed at legitimizing the socialist presence developed. Precisely because the historical dimension of the socialist culture was an integral part of its way of confronting the present, it produced a noteworthy effort directed towards the ennoblement of its past in some traditions of the European culture (from Plato to Campanella, from Thomas More to Thomas Müntzer). More in general, the study of the past of the working class movement and socialism was inextricably connected with the political passion in the works of Jaurès and Mehring, of Compère-Morel and Grünberg. Socialism influenced the historical culture also in the academic field and at the end of the nineteenth century there were many interpretations regarding the end of slavery in the Ancient world or the class struggles in medieval towns which echoed Marxist canons. The legitimation of the socialist parties provoked tensions and oppositions. However, in general at the beginning of the twentieth century it became an accomplished fact.

In reality, besides the brilliant electoral successes which consolidated the idea that the final victory was guaranteed by a slow and continuous increase of votes, socialism had at the same time stimulated and won over the intellectuals. From the American populist writers at the beginning of the century (the so called muckrakers) such as Jack London or Upton Sinclair to those writers such as the Russian Maksim Gorki, there are many artists who look towards socialism with sympathy: the old Emile Zola or the young Italian painter Carlo Carrà whose first work in 1901 was a portrait of Engels. But above all, socialism involved intellectuals and scholars in the field of the social sciences. They were challenged to interpret this key aspect in the life of contemporary societies and sometimes looked at it with sympathy. From Vilfredo Pareto to Benedetto Croce, from Werner Sombart to Emile Durkheim, each had to take a stand regarding socialism, and this meant that in a period of increasing alphabetization, newspapers, and public opinion attributed remarkable importance to this problem.

It is at this time that the political and cultural panorama of socialism became more complicated. The more socialist parties became actors in the political systems, the more intense the discussion within the party regarding the use of parliament, the role of social reform, the ways of passage toward a socialist society became. In general, during the period of the Second international, socialist thought tried to define the principal problems of the contemporary society. It radically criticized liberalism, unwillingly supported parliamentary democracy, tried to interpret imperialism, organized and coordinated the social protest, and created or unified complex labor organizations. The socialist world, however, was only partially accepted as a legitimate subject in the society: it had developed an attitude of 'negative integration,' and produced a 'subculture' closed within itself (Roth 1963). At the beginning of the century, the nonsocialist public opinion let out a sigh of relief upon seeing the relative availability of moderate socialists to enter and take part in coalition governments oriented toward social reform. However, even the most hopeful among them couldn't have predicted that in August 1914 internationalism, one of the milestones of the socialist culture, would have found itself in the midst of a crisis because of the outbreak of the war and that this crisis would have created a new irremediable rift opposing socialists against socialists.

5. *Socialisms in the Twentieth Century*

During the war, the conferences of Zimmerwald and Kienthal sanctioned the formation of socialist groups which criticized the failure of internationalism. Among them, the Bolsheviks played a particularly active role. After the outbreak of the October revolution the Bolsheviks promoted the formation of a new international organization which they called the Communist International. From that time on, the relationship between socialists and communists be-

came in certain extreme situations one of violent confrontation. Paradoxically right when the Bolsheviks were building 'socialism,' they identified 'socialists' as their principal adversaries. The development of 'socialism in one country' accompanied a progressive erosion of the democracy in Russia and caused social conflicts which were resolved only with the use of repression. In the period of the '*entre deux guerres*' socialists and social democrats (the two names coexist, the first in Latin Europe, the second in Germany and Scandinavia) strongly criticized communism, affirming the refusal of violence and exploiting the role of democratic institutions. After the experience of fighting together against fascism, socialists and communists split again during the cold war when the formation of 'socialist' or 'popular' democracies accompanied once again the implementation of repressive policies in Eastern Europe. In the East and West two opposing socialist worlds were formed: in the West, the German social democracy modified its identity by eliminating Marxism from its own political culture; in the East, Marxism was subjected to further dogmatic changes and was definitely transformed into Marxism–Leninism. After the Second World War the geography of socialism underwent a complex transformation. The crisis of colonialism and the anticolonialist tradition of the socialist movement influenced the formation of currents of socialist thought in several non-European countries. The Communist international had already exerted a worldwide influence outside the European boundaries, from Mexico to China, from Peru to Bengal but it was a simple center–periphery relationship. Starting with the beginning of the 1950s, autonomous socialist nuclei appeared especially in the Near East and in some colonies of sub-Saharan Africa. One of the examples was the so called 'African socialism' which had as its essential component the idea of a noncompetitive and nonindividualistic development with the objective of integrating religious, ethnic, and racial differences. In the 1960s and 1970s, the two different models of socialism continued to clash; on the one hand, democratic socialism, represented by political parties and experiences in coalition governments in Western Europe, on the other hand, the Soviet model. At the end of the 1980s, following an internal economic, political, and social crisis of enormous proportions, 'realized socialism' in the countries of eastern Europe ceased to exist, and with it the whole idea of socialism in the rest of the world underwent a crisis.

In the same way that it was so difficult at the beginning of the nineteenth century, it is just as difficult to offer a satisfying definition of socialism two centuries later. Anti-authoritarian opinions, nonindustrialist conceptions of nature, the desire to defend great humanitarian causes and civil rights have come back from a remote past to be seen again in the foreground, while the passion for equalitarianism and social justice shows continuity with the socialist traditions. The perspective from which these traditions were looked at has long been fairly ambiguous: often, in fact, the historians of socialism were also militant socialists, and political options have influenced the reconstruction of the past. The integration of the history of socialism in the academic environment has allowed scholars to look back at this subject with a certain detachment and has directed a special interest towards the place which socialist ideas have had in the culture of the lower classes and towards the role socialist doctrines and movements have played in the formation of contemporary political systems. Today an international socialist organization exists, socialist parties exist, and policies considered socialist exist. Their identity, however, no longer consists of a general conception of the world and of history, nor in allegiances or beliefs. Instead it consists of general cultural orientations aimed to control public expenditures in complex societies, directing resources towards citizens with lower income; in other words to extend the welfare state. Pragmatic options have substituted secular religions.

See also: Capitalism; Communism; Communism, History of; Democracy, History of; Dictatorship; Eastern European Studies: History; Eastern European Studies: Politics; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Labor Movements, History of; Marx, Karl (1818–89); Psychohistory; Right-wing Movements in the United States: Women and Gender; Russian Revolution, The; Sociability in History; Social Democracy; Social Movements and Gender; Social Movements, History of: General; Socialist Societies: Anthropological Aspects; Soviet Studies: Politics; Working Classes, History of

Bibliography

Cole G D H 1953–60 *A History of Socialist Thought*. 5 vols. St. Martin's Press, New York

Haupt G 1986 *Aspects of International Socialism, 1871–1914*. Cambridge University Press, Cambridge, UK

Histoire Générale du Socialisme 1972–78 [Publiée sous la direction de Jacques Droz.] 4 vols. Presses Universitaires de France, Paris

Hobsbawm E, Haupt G, Marek, F et al. (eds.) 1978–82 *Storia del marxismo*. 4 vols. Einaudi, Torino, Italy

Rosenberg A 1962 (1939) *Demokratie und Sozialismus. Zur politischen Geschichte der letzten 150 Jahre*. Europäische Verlagsanstalt, Frankfurt am Main, Germany

Roth G 1963 *The Social Democrats in Imperial Germany; a Study in Working-Class Isolation and National Integration*. [Pref. by R. Bendix.] Bedminster Press, Totowa, NJ

Steinberg H-J 1976 *Sozialismus und deutsche Sozialdemokratie. Zur Ideologie der Partei vor dem ersten Weltkrieg*. Dietz, Berlin

Sweezy P M 1949 *Socialism*. McGraw-Hill, New York

F. Andreucci

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Socialist Law

Socialist law, as an actual legal order, was a major legal family whose members possessed similar characteristics. The common features of the legal systems of the Soviet Union and communist East-Central Europe resulted from the use of socialist law in maintaining communist-party dominance over state and society. Yet the law in the various socialist states differed, too, not only because of modifications in the forms of domination but also because Soviet legal solutions were imposed on diverse cultures and societies.

The ideological legacy of Marxism–Leninism is expressed in the concept of socialist law. According to Marx, law was the will of the ruling class provided with statutory validity. Certain socialist (socialistic) theories of law have emphasized the importance of law in class domination. These theories served to legitimate communist rule. Other socialist (mostly Western) theories envisioned a normative system that reflected social equality and solidarity, that is, a theory of law based on social justice.

Socialist law as the positive legal system in communist countries was understood by its proponents as the expression of the will of the working class (or, when class struggle was less emphasized, of the 'toilers'), and it was an instrument of class struggle as well as a tool for building communist society. In the early years of the Bolshevik Revolution, workers' revolutionary legal consciousness was deemed the source of applicable law. Given the totalitarian roots of the legal system, to what extent the norms of the communist state fit into any concept of law remains a fundamental problem. If law is a social phenomenon or institution that differs from its environment, is socialist law to any extent autonomous? And, is it law at all?

1. *Stages in the Development of Socialist Law*

Socialist law changed over time. It served totalitarian regimes whose efforts were not always successful. State socialism was totalitarian in the sense that it attempted to construct all-embracing state control over every sphere of social life in the service of a single goal. In the early years of the socialist revolution and during certain periods of Stalin's rule, physical control and repression through brute force played the crucial role in this attempt to achieve total social control. It followed that law as a tool of social management or one that limits government action in accordance with its own specific logic could not exist. Law was not a serious consideration for the revolutionary mindset. This disregard was facilitated by Marx's belief that there was no place for state and law in the classless society. But Marx also believed that the transition leading to the classless society would be long; and he expected that law, in socialism, would contribute to the extension of equality under the law. Lenin argued for a dictatorship of the proletariat 'based directly upon force and unrestricted by any laws.' In the postrevolutionary period, the communist-party position was uncertain regarding the future of law. Yevgeny Pashukanis, the leading authority on communist legal theory, wrote in 1928 that law was a temporary system dictated by commodity relations. Once socialism had eradicated commodity relations, legal forms would also disappear.

In the 1930s, law (not only criminal law) was structured as an unrestrained authorization, and even a command, to intervene in anything and everything. Such calls to uncircumscribed, coercive intervention transgressed the limits of legal forms. The 1930 joint decree of the Central Executive Committee and the Council of People's Commissars of the USSR, which empowered local authorities to 'take all necessary measures … to fight kulaks,' would hardly satisfy any definition of law. But legal forms continued to exist even after the planned economy was fully developed. And even during the oppressive Stalinist regime, legal forms were employed at least for show-trial and propaganda purposes. Law remained an important internal and external tool of legitimation as exemplified in the Soviet Constitution of 1936 (known as the Stalin Constitution). Law served total repression, in the 1930s, in the sense that it reminded people that they could be brutally handled for no apparent reason and without any protection from other state institutions.

As communist-party oppression neared anarchy, unlimited power turned against the elite and their supporters. Totalitarian disregard for the specific needs of the economy and other areas of social activities drifted into dysfunction. Stalin himself 'brought law back in.' But the promised legal stability had 'nothing to do with the legal restraint on power, and it coincided with the period of greatest repression.' The system moved 'from lawless repression to repressive law' (Krygier 1994). Officially socialist law was proclaimed the most developed form of law in history as it served the most developed social formation (communism). It provided socialist legality, a qualitatively superior alternative to bourgeois rule of law. In reality it often served show-trial and other propaganda purposes only, as exemplified by the Stalin Constitution.

Laws enabled the direct repression of Stalin's Great Terror, although the mass-scale extermination (the purges, the gulag, etc.) was based on action that presupposed a disregard of promulgated legal norms (trial-procedure rules, evidence). This is not to claim that the 'disregard' was not officially sanctioned and ordered: disregard of the then-existing official law was a systemic component of the totalitarian use of power. The resulting destruction of social and individual resistance to communist authority, the irrevocable stamp of fear in almost every citizen, served in the

remaining half of the twentieth century as the foundation of the repressive order, which would afford to make use of less coercive laws.

After Stalin's death, in 1953, the communist elite reverted to a certain level of legal formalism, by and large for their own protection. This was achieved by a duplication, of sorts, of the normative system. The legal authorities had strictly observed secret rules regarding actions affecting members of the communist *nomenklatura*: to proceed against one of its members required authorization from the proper body within the *nomenklatura*. The *nomenklatura* followed its own 'rule of law', which did not include the right to a hearing of the concerned. The legalistic turn had a certain spillover effect on the rest of the population. Increasingly, law became less directly coercive and punitive and, at least in some of the satellite states, in matters of secondary importance it no longer depended fully on politics. On the other hand, law was recognized as a system that provided structure to a bureaucratic state capable of economic centralization. The copying of Soviet solutions resulted in less damage to legal forms in the East-Central European states than in the Soviet Union because the use of brute, Stalinist force lasted for shorter periods in these states and because Western legal traditions existed in some of them.

The relative autonomy of law was officially recognized, beginning in the 1970s. For reasons of regime legitimation it was overemphasized to disguise a poorer reality of limited autonomy. Nevertheless, the formal characteristics of modern law (abstractness, general nature of commands, etc.) survived, although in simplified and distorted forms. The instrumental use of law served a bureaucratically ossified institutional politics. The official doctrine emphasized that the autonomy of law could be relative only vis-à-vis the economy, which determined all forms of social life. 'Legislation, whether political or civil, never does more than proclaim, express in words, the will of economic relations' (Marx 1976). In reality neither the economy nor other social relations were respected, and only specific concerns of power and domination mattered, as understood by the communist leadership. Law remained a tool of the planned economy, and it served the deliberate changes of the planning system.

Law was declared the engineering device that helped organize people in their building of communism. Revolutionary law, though an instrument of class struggle, had an element of internal legitimacy as it served social (class) justice (through murderous revenge, expropriation, and redistribution). In the instrumental understanding of socialist law moral considerations became irrelevant. Instrumentalism was also distorted, however, as 'relative autonomy' allowed only very limited respect for the characteristics of the tool. Legal dogmatics or consistency requirements were easily skirted. Yet there were actual elements of independence in law simply because the political system could not generate a unified will. There was competition within the communist political elite, which contributed to at least superficial respect for legal forms. Moreover, law-enforcement organizations had a vested interest in legalism, and to the limited extent that they played a role in party politics, law became a point of consideration. To the extent that the communists needed efficient state machinery they needed law as an instrument of control over the state bureaucracy (see Sect. 4). Law was important insofar as centralization required uniformity and unity of action.

2. *The Structure of Socialist Law*

Socialist law was understood as an ideological superstructure, and its provision were subject to ideological scrutiny. Law (its provisions and understanding) had to be politically correct as understood by whatever ideological line the communist party touted. The ideological roots and functions of socialist law led many commentators to the conclusion that it was a mutation of natural law. It was argued that it originated in a set of doctrines or expectations that were above the law. Even the vulgar Marxist claims that law as superstructure reflects economic realities might have been interpreted as being in the natural-law tradition: it was the mystical economy with its 'objective law' that determined law. Socialistic traditions of social justice also show structural elements of natural law. Max Weber criticized these socialistic tendencies to material justice as detrimental to the formal, rational qualities of law.

Socialist legal systems, as they existed in state-socialist regimes, were positivistic and functioned as positive legal orders. The prevailing definition of law (legal system) was that it was the hierarchically structured sum total of state-made norms. The profound departure from promulgated norms and principles occurred not because of the observance of some kind of higher law. On the contrary, the departure was the result of changing secret orders, commands, and expectations emanating from the communist-party leadership and secret services. Of course, once the external political intervention becomes routine or even institutionalized, one can argue that there is no more law in the sense of normative, predictable expectations; hence socialist law (at its worst) does not fit into any notion of law.

Socialist legal systems, in principle, were strictly hierarchic. Statutes (parliamentary enactments) were of absolute supremacy, while constitutions offered only guiding perspective and basic rules of state organization. Law had to originate from the people's representatives; therefore, judicial lawmaking was ruled out as ideologically unacceptable, after the early revolutionary period. Statutes were supreme because they were enacted by parliament, which was held to be

the supreme state body. The rigid position on legal origins reflected the dominant nineteenth century German and French views and a simplification of Rousseau. Each socialist regime institutionalized some kind of small and reliable body within parliament (often called a presidium). The presidium had full power to act on behalf of parliament, which convened infrequently. The presidium had the power to amend existing laws. Decrees were subordinated to statutes. In theory, decrees were to serve exclusively for the execution of statutes. But in reality, substantive issues were often determined at the level of executive decrees. Officially courts were restricted to pure application of the law. However, a number of practical issues were determined by guiding principles of supreme courts. *Stare decisis* was unknown in this system; lower courts were bound to follow interpretations of higher courts. This was imperative given the informal control over the judiciary (see Sect. 4).

3. *Substantive and Procedural Institutions*

The centralized socialist management of the economy and social life was suspicious of any expression of autonomy. The priority of 'social interests' (or 'society's interests') defined the legal system. No autonomy (in the sense of self-determination) was permitted, and legal structures enabling or protecting autonomy were pruned even from private law. The traditional continental divide between public and private law was refuted as bourgeois legality. As Lenin stressed: 'We acknowledge nothing as "private." For us *everything* in the province of economics is in the domain of *public law* and not of private law … . Hence our task is … to broaden the right of the state to abrogate "private" contracts.' In practical terms this meant not only the denial of privacy and sanctity of contract but also the primacy and special protection of state property. Depending on the period, private property and private initiative were either prosecuted, prohibited, or greatly restricted, at best. Property consisting of the 'means of production' was restricted even where it was permitted (as in the German Democratic Republic (GDR), or, as far as land was concerned, in Poland). 'Social interests,' determined by the state, were favored above all other claims. Restrictions on personal freedom went beyond etatist concerns. Freedom of movement and residence was limited, and, at least in the Soviet Union, where the czarist tradition survived, freedom of internal movement was subject to administrative approval.

Speech and press freedoms were permitted only in the interest of workers, which resulted in institutionalized or informal but total censorship. Personal choices and alternatives were limited or prohibited (including, in education, health, profession, sexual identity, and the like). State control over the individual was extended to the intimate, private sphere. Decisions regarding family relations were subject to intervention by the *Prokuratura* (State Attorney's Office). Abortion was not an issue of self-determination but a matter of demographic or workforce politics, although certain social compromises were observed in this context. In Hungary, the nearly unlimited access to state-funded abortion was the result of an official repudiation of coerced demographic policies. In other countries, like Russia, abortion as the primary form of family planning derived from a total disregard for human dignity.

The legal restrictions imposed on the general population went hand in hand with the award of privileges for supporters of the regime, although only certain privileges were codified (e.g., access to higher education to family members of committed regime supporters). Privileges were not absolute. They were the reward for loyalty to the system. But the ordinary, loyal citizen (contrary to the privileges of the *nomenklatura*) had no title or claim to privileges, which, when bestowed, were an act of grace and goodwill. The discretionary distributive powers of the public administration were used to favor the politically loyal (or, at least, to punish the suspect). This was the case, for instance, in the assignment of public housing. During the less-dictatorial periods of state socialism, these discretionary methods of allocating goods and services were the primary tools of social control. Social control through the discretionary application of law put society in a condition of permanent dependence.

4. *Administration of Justice*

The contribution of the administration of justice to the restriction of liberty varied historically. During the Stalinist terror, courts served as theaters of repression and injustice. In less dramatic times, the administration of justice promoted etatism and protected the interests of the communist party. Most administrative decisions could not be appealed in court, although in Hungary and Poland, as a rule, there was the possibility of appeal against administrative decisions, except those involving the more sensitive issues of military and police administration.

As to judges, in the formative revolutionary years of socialist power, the legally trained personnel of the previous regime were replaced with incompetent but politically trustworthy laymen. The basis of revolutionary justice was the revolutionary consciousness of worker-judges. After World War II, in the Soviet-dominated satellite states the 'democratic attitude' of lay assessors played a role in politically charged procedures. In more settled times, judges received formal legal training but worked in personal and organizational dependence on court presidents. Courts were technically under the control of the

minister of justice. These courts were not structured to be independent, and the existing political control further increased the likelihood of 'telephone justice'.

Socialist legal theory emphasized the key role of the *Prokuratura* in the administration of justice. The organizational design of the Soviet *Prokuratura* was inherited from the czarist administration. It was a select, centralized body; procurators served in total subordination to their superiors, in what was a quasimilitary organization. They were much better paid than judges. The *Prokuratura* was in charge of criminal prosecution, representing the state in court, but it also routinely supervised the activities of the state administration. Lenin considered the *Prokuratura* the bulwark of socialist legality. He knew that a centralized state power could not afford any particularism, and all local, particular normative systems represented serious deviance. Normative particularism was especially dangerous if it originated in state bodies. As Lenin wrote to Stalin 'legality cannot be legality of Kaluga or the legality of Kazan but must be one single all-Russian legality.' Thus the task of the *Prokuratura* was to 'see to the establishment of a truly uniform understanding of legality.' The supervisory powers of the *Prokuratura* included the authority to ask for ordinary and extraordinary review of court decisions, even when it was not a party in the case it sought to review. Ironically, the fear of localism resulted in minimal protection for the population: the central bodies had an interest in mobilizing local populations against local authorities, and they encouraged denunciation of local (nonpolitical) abuses to the extent that they deviated from centrally authorized injustice. The dependence of citizens on local services was attenuated in the name of central law. Rules of procedure aimed to protect the material and political interests of the state, and allowed for paternalistic intervention in private litigation. The procedures followed simplified continental models that had traditionally existed in the given country. The inquisitorial nature of the traditional continental criminal procedure was further aggravated by Soviet solutions regarding the admissibility of police-gathered evidence.

In civil matters judges were required to establish 'objective justice' and were not bound by parties' motions. The paternalistic nature of continental civil procedure dramatically increased, resulting in further dependence, as far as the parties were concerned. On the other hand, the simplifications and limited access to court yielded cheap and relatively swift administration of justice (where courts were available at all). Civil litigation rates were stable, and the most common type of litigation was divorce related. Criminal litigation (except when criminal justice was used as a tool of class struggle, in the 1930s and 1950s, stabilized at low level. Criminality was limited because of the efficient social control exercised by the oppressive state mechanism.

See also: Law and Society: Sociolegal Studies; Marxism/Leninism; Marxist Social Thought, History of; Socialism; Socialism: Historical Aspects

Bibliography

Collins H 1982 *Marxism and Law*. Oxford University Press, Oxford, UK

David R, Brierley J E C 1985 *Major Legal Systems in the World Today: An Introduction to the Comparative Study of Law*. Stevens and Sons, London

Eörsi Gy 1979 *Comparative Civil* (*Private*) *Law*. Akadémiai Kiadó, Budapest, Hungary

Feifer G 1964 *Justice in Moscow*. Simon and Schuster, New York

Hazard J N 1969 *Communists and Their Law*. University of Chicago Press, Chicago

Krygier M 1994 Marxism, ideology and law. In: Krygier M (ed.) *Marxism and Communism: Posthumous Reflections on Politics, Society, and Law*. Rodopi, Amsterdam

Marcuse H 1971 *Soviet Marxism: A Critical Analysis*. Penguin, Harmondsworth, UK

Marx K, Engels F 1976 *The Poverty of Philosophy, Collected Works*. Lawrence and Wishart, London, Vol. VI

Podgorecki A, Olgiati V (eds.) 1996 *Totalitarian and Post-totalitarian Law*. Dartmouth, Aldershot, UK

Thompson E P 1977 *Whigs and Hunters: The Origin of the Black Act*. Penguin, Harmondsworth, UK

Vyshinskii A I A (ed.) 1948 *The Law of the Soviet State*. Macmillan, New York

Zweigert K, Kötz H 1987 *Introduction to Comparative Law*. Clarendon Press, Oxford, UK

A. Sajo

Socialist Societies: Anthropological Aspects

Socialist (also known as communist) societies constitute a class of twentieth-century societies sharing two distinctive features: the political dominance of a revolutionary—usually a Communist—Party, and widespread nationalization of means of production, with consequent preponderance of state and collective property. This definition excludes societies governed by socialist or social-democratic parties in multiparty systems, such as the Scandinavian welfare states. It includes the Soviet Union, the People's Republic of China, and (for some part of their histories) the East European countries, Mongolia, North Korea, Vietnam, South Yemen, Cuba, Nicaragua, Ethiopia, and Mozambique.

Socialist societies came into being with the 1917 Bolshevik Revolution and the establishment of the Soviet Union. Others founded later by various means (revolution, conquest, or annexation) all reflected the

influence—both negative and positive—of the Soviet model created under Lenin and Stalin. All employed extensive coercion and central control, yet they increased living standards, access to housing, employment, education, and medical care. Although they instituted new hierarchies, they also promoted gender equality and improved life-chances for many disadvantaged people.

1. Research on Socialist Societies

For much of the twentieth century, social science understanding of these societies was politically circumscribed. Their own scholars were under strict Party control; nonetheless, invaluable knowledge came from dissidents (e.g., Bahro 1978, Casals 1980) and from scholars in countries with strong reform movements (e.g., Kornai 1980, Wesolowski 1966). In Western Europe and the US, political and economic hostility to communist principles distorted thinking about socialist societies as epitomized in the concept of totalitarianism, which overstressed the importance of terror and top-down control. In all Western social science disciplines, impediments to on-site fieldwork hindered understanding. This situation improved in the late 1960s, for Eastern Europe, and the 1980s, for the Soviet Union and China, as *détente* and the impetus for technology transfer produced an infrastructure for intellectual exchange. The result was analyses that displaced totalitarian models to explore the societies' inner workings, with greater attention to phenomena such as kinship, ritual, the mobilization of labor, and the exchange of gifts and favors, as well as more subtle treatments of political economy (e.g., Humphrey 1983).

2. Political Economy and Social Organization

Although they varied widely in economic development (from industrialized Czechoslovakia, to agricultural China and Cuba, to pastoral Mongolia), most societies that became socialist lagged behind the advanced industrial countries economically; all embarked on programs of state-led development through central planning. The following description of their political economies emphasizes 'classic' socialism prior to the reforms introduced at various times, and draws on research in both agricultural and industrial settings in China, the Soviet Union, and Eastern Europe.

2.1 Redistributive Bureaucracies and Reciprocal Exchanges

Owing to the role of exchange in legitimizing socialism, it is appropriate to begin with Karl Polanyi's three basic modes of exchange: reciprocity, redistribution, and markets. Socialist systems suppressed the market principle, instituting redistribution as the chief exchange modality and the system's *raison d'être*: everyone would have the right to work, and the wealth they produced would be redistributed as welfare for all. Centralized redistribution required social ownership and control of productive resources, together with centralized appropriation of the product for reallocation along lines set by the Party. The instrument Party leaders developed to orchestrate production and distribution was the plan. Fulfilling it, however, depended on relations of reciprocity, both clientelistic and egalitarian.

Although these were highly bureaucratized societies, they were not Max Weber's impersonal modern bureaucracies dedicated to procedural rationality. Dense social networks sustained reciprocal exchanges that, far from constituting resistance to bureaucratic control, were intrinsic to its functioning: horizontal exchanges enabled socialist managers to do their jobs, and clientelism facilitated controlling both labor and Party cadres in seemingly noncoercive ways. Personalistic ties extended far beyond the bureaucracy as well, linking ordinary citizens with one another and with bureaucrats. The warp and woof of socialist societies, then, consisted of vertical and horizontal relations of patronage, loyalty, and exchanges of goods, favors, and gifts. Although the forms of these relations differed according to the various societies' prerevolutionary histories and cultures, (e.g., gift relations in China, *compadrazgo* in Cuba or Nicaragua, or specific kinship structures of Soviet minorities), they were widely prevalent in all. In each socialist society words indicating 'connections' (Russian *blat*, Chinese *guanxi*, Romanian *pile*, Hungarian *protekció*, Cuban *compañerismo*, Polish *dojscie*) appeared often in daily speech.

2.2 Organized Shortage

Personal relations were integral to socialism in part because of certain pervasive features of planned economies: tension between Party planners and firm managers, bargaining and soft budget constraints, and barter amid shortage. Socialist property could be productive only if the center's ownership/control rights were disaggregated and devolved on to lower-level managers charged with realizing production. Because planning could never foresee all contingencies, these managers had considerable discretionary power. Expected to achieve Party directives with inadequate wage funds and to exceed their plan targets with inadequate raw materials, they bore responsibility far greater than the means allotted them for fulfilling it. Therefore, they strove constantly to expand their local power base and their effective capacities. In doing so, they posed a threat to their superiors, who thus were motivated to encourage subordinates' loyalty with favors and patronage.

In redistributive systems governed not by demand for products but by the Party's planned allocation, materials for production could not simply be bought on a market; their availability depended on the supplies budgeted in plans and on often-inefficient central distribution. Managers, therefore, requested more supplies than they needed, hoping to obtain enough materials to fulfill and exceed their targets. Because the planning mechanism required firms to produce regardless of profitability—they operated under soft budget constraints and were rescued rather than bankrupted if they lost money—local managers could with impunity overstate their needs for materials and investments and then hoard any excess. They also strove to bargain their plan targets downward, making it easier to fill these and have goods left over. Comparable processes occurred in both industry and agriculture, as cadres everywhere manipulated information, under-reported production, and engaged in illicit trade to benefit their firm or locale.

The result of bargaining and hoarding was endemic scarcity of the materials necessary for production; thus, classic socialist societies were economies of shortage (Kornai 1980). Shortage caused competition among firms but also widespread exchanges, managers supplying from their hoard today the materials needed by others who would return the favor tomorrow. Thus, firms hoarded materials not only to cover emergencies in their own production but to backstop the supplies needed by others in their network. Shortage affected materials for production and also consumption goods, generating the queues characteristic of many socialist societies. It also affected two other crucial resources: information and labor.

Party planners could implement top-down planning only if they had detailed and accurate information. This happened rarely, for lower-level cadres had many incentives to falsify information so as to build careers by appearing successful, evade responsibility for disasters, inflate their materials requests, provide cushions against production shortfalls, overfulfill their targets to receive bonuses, etc. From these discrepant needs came a perpetual struggle over information. Higher authorities hoarded information as a source of power and engaged in continual information gathering at all levels of society (Horváth and Szakolczai 1992). Censorship and careful management of information at the top provoked information hunger below, generating a culture of rumor, gossip, selective secrecy, and conspiratorial explanations of events, as well as producing a citizenry skilled at reading between the lines. Classic socialist societies were thus a special kind of information society.

2.3 Labor Shortage and Rights in People

An additional resource subject to shortage was labor. Relatively underdeveloped socialist economies had initial problems mobilizing skilled workers where there was effectively no 'working class,' and they lacked the resources for either paying labor well or substituting capital-intensive methods. Moreover, the shortage economy encouraged firms to employ excess workers so they could complete monthly plans once sufficient materials were on hand (in a frenzied effort known as 'storming'); managers, therefore, padded their workforce, exacerbating shortage. Labor shortage, thus, had both macro and micro dimensions. Means for stabilizing the labor supply included making workplaces the locus of benefits such as daycare, housing, vacation permits, pensions, medical care, etc. Certain cities were closed to immigration and construction of urban infrastructures was limited (Szelényi 1983), thus compelling millions to become village-based 'peasant-workers.'

Within each industrial and agricultural workplace, managers had to ensure a labor supply adequate to both normal production and periods of 'storming.' Workers, for their part, might bargain for better conditions by withholding labor, through intentional slowdowns or time off for household tasks and moonlighting—that is, labor shortage gave workers structural leverage. Conflicting demands for labor put a premium on ways of accumulating rights in people. For managers, these included extending patronage and favors (e.g., access to schooling, houses, or building sites), overlooking petty illegalities, and securing loose plans that enabled them to produce a surplus they might use to create labor-securing debt and exchange relations (cf. Humphrey 1983). Both managers and others needing labor might appeal to kinship idioms, emphasize ethnic identities, participate in special rituals, and expand networks of reciprocity through gift-giving. Thus, the quest for labor further encouraged personal ties and reinforced particularistic identities.

2.4 Second Economies and Socialist Reform

Believing they alone could best determine how social wealth should be redistributed, Party leaders initially opposed economic activity not encompassed within plans. The inability of planning to cover social needs at the given level of technological endowment, however, compelled officials to permit and even legalize some small-scale private effort, known as the 'second' (or informal, unofficial, or shadow) economy. Among its forms were food production on small plots, after-hours repair work or construction, typing, tutoring, unofficial taxi services, etc. Because these activities overlapped with semilegal and illegal ones, their situation was everywhere precarious, with authorities persecuting them more in some times and places than in others. The second economy was largest in Hungary after 1968, for example, small in Cuba until Castro's 'Rectification' of 1986, harassed in Romania throughout the 1980s; it burgeoned in post-1978 China; extreme forms are reported for the Soviet Union,

where entire factories ran illegal production after hours. Crucial to understanding the second economy is that it nearly always utilized materials from the first (or formal, official) economy; its much-noted high rates of productivity were subsidized, then, by state firms. The prevalence of second-economic activity both indicated popular resistance to the Party's definition of needs and helped to fill those needs by voluntarily lengthening the working day.

Pressure from the second economy was but one of many signs of difficulty in socialist planning that led to repeated efforts at reform, initiated from both within and outside the Party. Beginning with Lenin's New Economic Policy, every socialist society experienced cycles of reform and retrenchment, devolution and recentralization (cf. Skinner and Winckler 1969, Nee and Stark 1989). Most exhibited a trend toward less stringent planning and the introduction of market mechanisms, heightened material incentives, and mixed property forms. System-wide experimentation began with Khrushchev's 1956 'Secret Speech' criticizing Stalin and increased as each society moved from 'extensive' development (mobilizing resources) to the 'intensive' phase (attention to productivity). Hungary and Yugoslavia introduced the most durable early reforms; those in the Soviet Union ended in the collapse of the Soviet bloc, while comparable reforms continued in China, North Vietnam, and Cuba. As they reformed, socialist societies increasingly diverged not only from the Stalinist model but from one another, introducing path-dependent differences that became ever more marked.

Despite these differences, the reforms everywhere redefined basic units of activity (e.g., revitalizing villages or households at the expense of collective farms, teams, or brigades); altered gender relations (usually in favor of men and patriarchal authority) and increased other inequalities; affected networks of reciprocity (expanding horizontal over vertical connections); dismantled at least some socialist property (as with China's decollectivization, begun in 1978); shifted the locus of authority (usually downward, provoking reactions from higher-level bureaucrats); and entailed new, more intimate forms of state penetration (implied, e.g., in Chinese rituals that no longer imagined gods or ancestors as inhabiting a nether world but found them immediately present).

3. *Social Engineering and Resistant Personhood*

All Communist Parties promoted massive projects of social engineering. They accorded knowledge and expertise a privileged place, under Party monopoly. Convinced of the creative power of language, Party leaders both used and modified language in commanding ways. They spoke of things that had yet to be created—the working class, the proletarian dictatorship, the new socialist man—as if those things already existed, and they developed hieratic speech with reduced vocabularies, clusters of noun phrases, and few (often passive) verbs, creating a limited, static verbal world. Party cadres intervened exhaustively in all facets of life, aspiring to create new moralities and to control populations in myriad ways. They attacked religion as superstition, instituted new socialist rituals, and expanded educational access to create enlightened citizens indifferent to religious belief. They filled the air with moral exhortations aimed at instilling a new, puritan, socialist ethic. Toward better management of resources, they introduced massive population programs, from Romania's enforced pronatalism to China's stringent birth control, insinuating themselves into people's most intimate lives, often to devastating effect (e.g., Kligman 1998).

Dedicated to building an alternative to the world of bourgeois capitalism, communist leaders announced an assault on all forms of inequality, including those based in kinship status, gender, class, and ethnonational difference. Society would be homogenized, lineage-based kinship structures broken up, gender differences minimized and authority relations in families altered, classes and private property abolished, national minorities given equal chances. New social entities emerged—state and collective farms, teams, and brigades, for instance, to take the place of kin-groups and village structures. When pre-existing social relations proved difficult to abolish, they would be co-opted and turned to the Party's purposes. From all these efforts would result the 'new socialist man,' a new kind of person, and the 'people-as-One,' a new kind of human community, bearing an exalted socialist consciousness.

The results of these interventions were decidedly mixed. To begin with, the Party failed to speak with one voice, as ministries vied for resources rather than cooperating to remold society, and as cadres built their careers by under- or over-executing central directives and engaging in rampant corruption. In addition, the annexation of pre-existing forms inevitably modified the purposes they were intended to serve. Exhibiting greater capacity to formulate goals than to implement them, socialist governments were plagued continually by the unforeseen consequences of their policies. To solve conflicts among nationalities they institutionalized ethnonational difference in ways that strengthened it; they embraced nationalist idioms that then overtook socialist ones (Verdery 1991, 1996). Parties created working classes only to find themselves opposed by their creation, as with the Polish authorities and Solidarity.

Beyond this, resistance to social engineering proliferated even as people adopted the new forms. Constant surveillance politicized behavior: small acts of joke telling became willful anarchy in the eyes of both authorities and joke tellers. Dissidents sought support internally and abroad for their critiques of the system, launching projects to build 'civil society' against the state. Less visibly, peasant refusal to accept starvation

from China's Great Leap Forward or Eastern Europe's delivery quotas ultimately forced the authorities to renounce these policies (Yang 1996, Rév 1987), while popular insistence on adequate living standards pressed the parties toward reform.

Highly complex political effects centered on the making of persons and political subjects. Redistribution and socialist paternalism tended to create subjects disposed to dependency and passive expectation; early campaigns taught the poor to see themselves as victims of oppression (rather than simply of fate) and offered them the Party's instrument of vengeance, in the form of denunciations. Thus arousing people's complicity, the Party also instilled self-denigration. At the same time, however, many adopted dissimulation as a mode of being: apparent compliance covered inner resistance. The resultant 'social schizophrenia,' 'doubling,' duplicity, or split self is described for every socialist society. It accompanied a pervasive dichotomization of the world into 'Us' (the people) and 'Them' (the authorities). More significantly, it encouraged subtle forms of self-making in people's own terms: defiantly they consumed forbidden Western goods, created self-respect through diligent second-economy work while loafing on their formal jobs, participated in ethnic- or kin-based identities and rituals constitutive of self, and gave gifts not just to secure advantage but to confirm their sociality as persons and human beings.

The collapse of the Soviet bloc hinders a balanced assessment of socialism, which once represented great hope for countless people. Undeniably, socialist societies diminished certain forms of oppression, yet they created others. In many of these societies, an early élan for the expressed goals gave way to cynicism and a feeling of betrayal, as the means used to achieve them sabotaged their fulfillment. Admirable gains in education, medical care, and social well-being counted for less against the devastating purges, labor camps, Party-induced famines, environmental degradation, and abuses of human rights. Top-down engineering foundered as socialism's leaders resisted their citizens' efforts to educate them about social relationships, self and human feeling, the value of ritual practices, and the limits to imposed change. The lessons of these citizens' refusals, their silent revolutions from below, should guide future attempts to create alternatives to the world of global capitalism.

See also: Mao Zedong; People's Republic of China; Soviet Union; Stalin; Cold War, The; Communism; Communism, History of; Informal and Underground Economics; Marxism/Leninism; Socialism; Socialism: Historical Aspects; Soviet Studies: Culture; Totalitarianism

Bibliography

Bahro R 1978 *The Alternative in Eastern Europe*. Verso, London

Casals F G 1980 *The Syncretic Society*. M. E. Sharpe, White Plains, NY

Horváth A, Szakolczai A 1992 *The Dissolution of Communist Power: The Case of Hungary*. Routledge, New York

Humphrey C 1983 *Karl Marx Collective: Economy, Society and Religion in a Siberian Collective Farm*. Cambridge University Press, Cambridge, UK

Judd E R 1994 *Gender and Power in Rural North China*. Stanford University Press, Stanford, CA

Kligman G 1998 *The Politics of Duplicity: Controlling Reproduction in Ceauşescu's Romania*. University of California Press, Berkeley, CA

Kornai J 1980 *Economics of Shortage*. North-Holland, Amsterdam

Kornai J 1992 *The Socialist System*. Princeton University Press, Princeton, NJ

Nee V, Stark D (eds.) 1989 *Remaking the Economic Institutions of Socialism: China and Eastern Europe*. Stanford University Press, Stanford, CA

Rév I 1987 The advantages of being atomized. *Dissent* **34**: 335–50

Skinner G W, Winckler E A 1969 Compliance succession in rural Communist China: A cyclical theory. In: Etzioni A (ed.) *A Sociological Reader on Complex Organizations*, 2nd edn. Holt, Rinehart and Winston, New York

Szelényi I 1983 *Urban Inequalities under State Socialism*. Oxford University Press, Oxford, UK

Verdery K 1991 *National Ideology under Socialism: Identity and Cultural Politics in Ceauşescu's Romania*. University of California Press, Berkeley, CA

Verdery K 1996 *What Was Socialism, and What Comes Next?* Princeton University Press, Princeton, NJ

Walder A G 1986 *Communist Neo-traditionalism: Work and Authority in Chinese Industry*. University of California Press, Berkeley, CA

Wesolowski W 1966 *Classes, Strata, and Power*. Routledge & Kegan Paul, London

Yang D 1996 *Calamity and Reform in China: State, Rural Society, and Institutional Change since the Great Leap Famine*. Stanford University Press, Stanford, CA

K. Verdery

Sociality: Anthropological Aspects

1. The Problem of Culture

Sociality is the capacity for social behavior, a capacity particularly marked among human beings, as evidenced by the intensity, complexity, and endless variety of human social life. This definition echoes the earlier use of sociality in philosophy and psychology, where it is used to denote the fact of human social interdependence and to correct a tendency in those disciplines to explain human behavior in individualistic terms alone. Sociality is used differently among evolutionary biologists, where it denotes the actually observed social behavior of a species, such as its mating system or dominance hierarchy; sociality so described is then subjected to an explanation from

natural selection (Wilson 1975). But human social behavior is so varied and so open-ended that no such simple description or explanation can be given. We must, therefore, describe human sociality as a potential, one which results in many different social systems.

To understand human sociality we need to turn first to the concept of culture. For it is culture which has borne most of the weight of anthropological efforts to think about human nature in general and about human potential, and indeed one could regard ideas about culture as anthropology's *first approximation* to a picture of human nature. This was sketched in by theorists and fieldworkers before World War II and confirmed in the great expansion of fieldwork and theory after that war.

In this first approximation, human beings are dependent upon the possession of some particular culture, both for their survival and for their recognizably human character. For culture comprises everything that we might consider knowledge, from knowledge of language and other practical techniques, through knowledge about the immediate natural world and the larger cosmos, to knowledge of social customs and social relations. Geertz captured the encompassing import of culture when he remarked that 'man's nervous system does not merely enable him to acquire culture, it positively demands that he do so if it is going to function at all' (Geertz 1973, p. 73). On this view, human beings have evolved to have culture, and each individual human being begins as a more or less blank slate upon which the knowledge of culture can be written, or as a new computer into which a program must be loaded. Moreover, the saliency of culture, as opposed to raw genetic inheritance, in human nature is certified by the tremendous variety in ways of life across the world. Humanity presents a patchwork of sharply discrete societies, each with its own social organization, its own language, and its own view of the world. It is the nature of these cultures, according to the first approximation, to be deeply traditional and resistant both to change and to mutual understanding.

Yet anthropologists' confidence in this first approximation has given way since the 1980s to sharp doubts. For such a conception of human nature is still, for all its ambition and universal scope, a narrow and partial one. It misses out the change, incompleteness, ambiguity and potential—that is the historicity—that are so prominent in human society and experience (Peel 1987, Fox 1985). Wolf (1982) showed that there have existed no peoples beyond the reach of historicity, change, and interchange between peoples, no primary and untouched cultures. Human beings do create new roles and new institutions and ways of doing things—such as the joint stock company, which has played such a great part in world history in recent centuries. And human beings do, as a matter of course, make effective contact and create enduring social relationships beyond the apparent limitations of our natal culture, as evidenced in networks of trade, in long-distance marriage networks (such as those of India), in the practice of ethnographic fieldwork, and in all the other occasions, such as migration for work, for safety, and even for conquest, that create a continual sifting and reblending of people and cultures across the globe. (For a sense of relatedness beyond ostensible boundaries, see Hannerz 1996.) Anthropologists have begun to look to human sociality to fashion a second approximation of human nature that will allow for such flexibility (see Ingold 1986, Carrithers 1992).

2. *Sociality as Intersubjectivity and Social Intelligence*

It is perhaps easiest to capture sociality to work in infancy, when the infant has not yet acquired the tools of a particular culture. Human infants already show a responsiveness to others which goes well beyond that found among any other species, including the great apes. Even a newborn can imitate facial expressions and respond to the emotions of others. This is primary intersubjectivity, in which the baby already has a dual 'self + other' orientation (Trevarthen and Logotheti 1989). By the age of one year, infants have developed the ability to share purposes and intentions with others, even if they cannot yet formulate these clearly in speech.

So the infant possesses an innate and growing social intelligence that operates with the feelings and doings of others. Upon the basis of primary intersubjectivity, for example, the infant will be able to go on in the first two years of life to achieve joint attention with others, following a carer's gaze or gesture to an object, and will learn to draw a carer's attention to an object. The intellectual achievements of intersubjectivity and joint attention form what Bruner (1983) has called the Language Acquisition Support System, the scaffolding of abilities upon which the joint capacity of speech, as well as the other achievements of culture, will be constructed. Later, at the age of three years or so, the child will begin to differentiate what they know and feel from what others around know and feel; they will then begin to simulate the knowledge and feelings of others, and on that basis will be able to experiment with a growing variety of social relationships, both in play and in earnest. Already at this age, that is, the child has begun to 'take the role of the other,' in George Herbert Mead's famous phrase, the facility upon which all distinctively human socializing is built.

It is useful to think of sociality as social intelligence, and indeed it seems quite likely that our technical and symbolic intelligence is an evolutionary by-product of the formidable abilities developed in social interaction. But there is one implication of the term 'intelligence' that must be stripped away, and that is the assumption that sociality is a strictly individual capacity. On the

contrary, it comprises an intellectual, cognitive, and emotional openness to others which runs throughout human experience, in childhood or adulthood. The child matures, but into a matured form of intersubjectivity, of self + other orientation, not into sealed isolation. This means, for example, that most human abilities and knowledge are *distributed*, a term drawn from computing to indicate the distribution of a computing problem across more than one computer (or central processing unit). One sees this on a smaller scale in the coordination of persons and sharing of information in subsistence activities among hunter–gatherers, or on a larger scale in the division of labor that brings this encyclopedia to the reader: at no point need, or can, all the necessary skills and information reside in one individual, but must be distributed across all participants. The knowledge and skills exist only with the scaffolding of persons-in-relationship who exercise the knowledge. To return to Geertz's dictum: it is not that the individual nervous system acquires culture, but that the socially intelligent nervous system completes itself with others to acquire culture jointly with them. It is a good deal more difficult for researchers to keep this more complex, more nuanced image in mind, but it is far more faithful to the character of human life.

2.1 Sociality as Mind-reading in Talk

So one achievement of sociality is to make human knowledge and skills much more than the sum of individuals, and to that extent, more powerful. But there is a cost as well, for the continual self + other orientation of people requires constant attention and a great deal of energy. We are socially calculating beings. We must be able to decide not only what to do, but to calculate the consequences of our actions on others, their probable reaction to those consequences, and the significance of those reactions for ourselves.

The calculating ability thus required can be called *higher-order intentionality*, the ability to think complexly about one's own and others' states of mind and reactions. To put this abstractly: one represents to oneself others' states of mind (first-order intentionality), or one represents to oneself others' representations of one's own state of mind (second-order intentionality), or one even represents to oneself others' representations of one's own representations of their state of mind (third-order intentionality). Here is an example of third-order intentionality: I *believe* that Amy *recognizes* that I *accept* what she is saying. One can go further: I *believe* that you *recognize* that I *think* you *understand* the general idea of higher-order intentionality. In thus, spelling out the idea of higher-order intentionality, we can see how complicated it is, and therefore how much intelligence it requires. Yet these are also the everyday, unremarked, and swiftly-deployed abilities that a mature human being needs to conduct even the most workaday talk. Whiten (1991) has called this ability 'mind-reading,' but of course it is not mind-reading as some mystical skill of seeing directly into the little grey cells, but is rather based on interpretation of the clues of glance, tone, timing, and nuance that people emanate continually in their interactions.

To weigh the significance of these central abilities in human sociality, we need first to note the cumulative effect of people's ability to conduct such everyday talk. For through such talk, we are able to create and sustain complex skeins of joint action, including the joint action of provisioning and sustaining ourselves. As Maurice Godelier remarked, 'human beings … do not just live in society, they produce society in order to live' (Godelier 1986, p. 1). On that account, sociality is a resounding practical success, opening creatively onto many forms of joint social action.

But such practical success, leading to some reliability and permanence in social relations, is only partial. For, as one can already see from the nature of mind-reading, the imputation of beliefs and attitudes to others can hardly be infallible. Slips are frequent, misunderstandings are sometimes repaired and sometimes not, and the interpretation of what one says can never fully be determined by the words themselves, as discourse and conversation analysts have so meticulously documented (see e.g., Moerman 1988). The cumulative direction of conversation cannot be controlled fully by any single participant, and all participants' interpretations and interests contribute to the process through implicit or explicit negotiation (Heritage 1984). So the seeming sureness of everyday talk is actually a laborious and uncertain business, a matter of minute-by-minute, indeed second-by-second, interpretation, with all the fallibility, but also creativity, inherent in the idea of interpretation. Max Weber's famous phrase, 'unintended consequences,' is usually used of large social phenomena, but unintended consequences begin in the fine grain of everyday life.

2.2 Sociality in Narrative Thought

So human sociality involves a noisy eventfulness that resembles that of other primate groups, but which also has notably human features. We often do things, not only because of what others do, but also because of what they might think or feel: I may *apologise* because she was *angry*, I *explain* myself because someone might *think* I am inconsiderate. So an action can be not only practical, but also have an interactive meaning with social effect: I built that fence to keep out stray cattle, but also to spite my neighbor. Moreover people can, and do, weave such actions and reactions together over a long series, linking cause and effect, action and reaction. This may be called a plot line or narrative, and indeed this larger scale mutual awareness may be called narrative thought (Bruner

1986, Carrithers 1992). Narrative thought is richly evidenced in gossip, story, history, and myth, that is, in all the ways in which people are able to represent complex skeins of action to one another. To understand the story of Oedipus, for example, the audience must be able to bear in mind the relationships between Jocasta, Tiresias, and Oedipus as they develop through a complex play of events and changes in the characters' states of mind. Moreover people are able, through narrative thought, to come to understand the underlying propensities, attitudes, beliefs—in a word, the character—of those involved in the skein of action. But narrative thought is much more than a passive competence, because it is only through understanding one's place in the various skeins of action and characters that comprise one's own life that one is able to act sensibly and appropriately to others. Societies vary greatly in their descriptions of characters and their notions of plots, but the basic theme of narrative thought, that of placing oneself with others in mutual awareness of larger time frame and a larger society, going beyond immediate events and those immediately present, is characteristic of all human life.

Moreover, this larger view over events concerns not only the past, but the future as well. Goody (1995) has written of 'anticipatory interactive planning' as a characteristic of human sociality, and Cole (1996) has shown how people use their narrative understanding of the past to project the future. He calls this *prolepsis*, 'looking ahead.' He uses the example (drawn from the observations of an (obstetrician) of a mother who said, as she held her newborn daughter in the hospital for the first time, 'I shall be worried to death when she's 18.' The remark is a sort of miniature of the complexities of human sociality. It shows the engagement of the mother with the daughter, and looks forward to a projected, predicted life cycle for the daughter. And of course the remark looks back as well, to the experience of the mother and her peers and forebears: the mother is (re)collecting her own and others' past to predict her daughter's future.

3. Sociality and Culture

In regarding human nature from the point of view of sociality, we bring human agency, human intersubjectivity, and historicity to the fore, linking human interaction and human creativity. What then of culture? It is difficult to see how culture-as-a-computer-program could fit here, but there is a fruitful alternative, namely to regard culture as a collection of artifacts. Through such artefacts human beings mediate interactions between each other, and between themselves and the natural world (Cole 1996). Actual tools are of course included, and we can see a good deal of how they mediate between people when we think of writing, books, telephones, or computers. But language itself, with all its panoply of different tools, is also a collection of artefacts for mediating between people. Thus, for example, South Asian languages tend to possess an extensive set of second person pronouns, each denoting a degree of rank difference between the speaker and the interlocutor. Such tools allow people to set a relationship … or to dispute it, by choosing a pronoun different from the one the interlocutor may wish to hear. Kinship systems are another set of such tools. There are also corporeal artefacts, such as the gesture of obeisance in South Asia known as offering *puja* or worship, which establishes a relationship of great respect to some real or spiritual other. Other artefacts include what have been called 'scripts,' the little slices of action through which people accomplish routine tasks. There are scripts, for example, for greeting people politely throughout the world, politeness being a marked feature of human sociality (Brown and Levinson 1987, Strecker 1988). These scripts may be played practically, to get the work of setting a relationship or conversation going, but also strategically, to attain some end of the speaker or interlocutor. And there are also stories, or story templates, that guide people over longer stretches of action. Many have noted, for example, how the story of the good stranger arriving in town to deal with evil-doers has often influenced American foreign policy.

It is characteristic of human ingenuity that these artefacts may not only be used strategically, but changed to achieve new ends or to meet new circumstances. Perhaps the most spectacular examples of this are the creation of writing from speech, and later books from writing, thus creating new, mediated, forms of interaction, 'exploded interactions' (Carrithers 1992), between people separated in time and space.

On the first approximation of human nature, it was difficult to see how cultures could change, or indeed how the great variety of cultures had come into being. But by regarding culture as a collection of artefacts and humans as possessing sociality beyond any prescriptions of culture, we can begin to see that any culture itself has a history, and that no one culture circumscribes finally the potential of those who use it. In this second approximation of human nature, both historicity and the creation of new cultural forms are already woven into the human fabric.

Bibliography

Brown P, Levinson S C 1987 *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK

Bruner J S 1983 *Child's Talk: Learning to Use Language*. W N Norton, New York

Carrithers M 1992 *Why Humans have Cultures: Explaining Anthropology and Social Diversity*. Oxford University Press, Oxford, UK

Cole M 1996 *Cultural Psychology: A Once Future Discipline*. Harvard University Press, Cambridge, MA

Fox R G 1985 *Lions of the Punjab: Culture in the Making*. University of California Press, Berkeley, CA
Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York
Godelier M 1986 *The Mental and the Material: Thought, Economy, and Society*. Verso, London
Goody E N (ed.) 1995 *Social Intelligence and Interaction: Expressions and Implications of the Social Bias in Human Intelligence*. Cambridge University Press, Cambridge, UK
Hannerz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London
Heritage J 1984 *Garfinkel and Ethnomethodology*. Polity Press, Cambridge, UK
Ingold T 1986 *Evolution and Social Life*. Cambridge University Press, Cambridge, UK
Moerman M 1988 *Talking Culture: Ethnography and Conversation Analysis*. University of Pennsylvania Press, Philadelphia, PA
Peel J D Y 1987 History, culture, and the comparative method. In: Holy L (ed.) *Comparative Anthropology*. Blackwell, Oxford, UK
Strecker I A 1988 *The Social Practice of Symbolisation: An Anthropological Analysis*. Athlone, London
Trevarthen C, Logotheti K 1989 Child in society, society in children. In: Howell S, Willis R (eds.) *Societies at Peace: Anthropological Perspectives*. Routledge, London
Whiten A (ed.) 1991 *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*. Blackwell, Oxford, UK
Wilson E O 1975 *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA
Wolf E R 1982 *Europe and the People without History*. University of California Press, Berkeley, CA

M. Carrithers

Sociality, Evolution of

Sociality involves multiple individuals of the same species engaging in a common task. Such tasks may involve any sort of behavior, but most research on the evolution of sociality has focused on the evolution of social cooperation in the production and rearing of offspring. The main goal of this research program is to explain and predict the occurrence of different social systems of multiple individuals engaged in reproduction, from aspects of the genetic, phenotypic, and ecological traits present in different species or populations. Pursuing this goal requires definition and classification of social systems.

1. *Defining Social Systems*

The purpose of defining social systems is twofold: clear communication between researchers, and classificatory structure for comparative evolutionary analysis of the causes of social evolution. To the extent that social system classifications are not based solely on phylogenetic relationships of species, they are, however, artificial constructs of humans. As such, they are more or less useful in answering specific questions, rather than more or less true and realistic.

1.1 *Social System Classification*

Reproductive social systems can be classified into three categories, based on (1) the presence and absence of behaviorally distinct sets of individuals within a social group, one of which specializes in reproduction and one of which specializes in helping the more-reproductive individuals, and (2) permanence of these behaviorally distinct sets, such that in societies with permanently distinct groups (called 'castes') some individuals become irreversibly phenotypically distinct prior to adulthood (Choe and Crespi 1997). 'Communal' societies exhibit cooperation in reproduction, but they lack behavioral specialization into 'reproductives' and 'helpers'. 'Cooperatively breeding' societies show such behavioral specialization, but it is not irreversible: some individuals can and do switch between the reproductive and helper roles as adults. 'Eusocial' ('truly social') societies exhibit castes. These three categories refer to societies rather than to species, because some species exhibit social variability across geographic regions or across the life cycle.

1.2 *Social Taxa*

Taxa exhibiting communal social systems include some bees, beetles, mites, spiders, birds, rodents, and primates. By contrast, cooperative breeding has been described in many birds, some bees and wasps, and various mammals (especially carnivores, viverrids, and rodents).

Taxa exhibiting eusociality include many Hymenoptera (ants, bees and wasps), some gall-inducing aphids and thrips, sponge-dwelling, shrimp, termites, encyrtid wasps, one species of bark beetle, naked mole rats, and damaraland mole rats (Choe and Crespi 1997). The challenge for students of social evolution is to make sense of the sporadic and idiosyncratic taxonomic distributions of these social systems and to find the commonalities and differences between groups that are responsible for their convergent and divergent social systems. The main determinants of social system evolution include genetic relatedness and aspects of ecology.

2. *Genetic Relatedness and Altruism*

Much cooperation is mutualistic in that all of the interacting individuals gain from the behavior. However, many forms of cooperation involve altruistic behavior, whereby one individual provides benefits to another at a cost to itself. The presence of such behavior appears paradoxical to evolutionary theory

Table 1
The three mechanisms leading to social cooperation

1. Benefits gained by cooperators
   a. mutualism or altruism – all benefit equally from cooperation, no incentive to cheat (i.e., exploit resources created by cooperators)
   b. mutualism or altruism – unequal benefits from cooperation, incentive to cheat, success of cheaters limited in some way (e.g., is frequency-dependent)
2. Costs imposed on non-cooperators
   a. manipulation – dominant reduces independent breeding prospects of subordinate(s), favoring helping
   b. punishment – dominant individual reduces fitness of non-cooperators
   c. policing – less-reproductive individuals prevent one another from engaging in reproductive cheating
3. Forced cooperation: dominant individual takes control of behavior away, such that subordinate individual must cooperate

because reduced fitness cannot evolve. Hamilton (1964) proposed a resolution to this paradox: that genes for altruism can increase in frequency and be maintained because they may also be present in the recipient of the benefits. Thus, by helping genetic relatives, an individual can maximize its 'inclusive fitness,' the sum of its own reproduction (descendant kin), plus its reproduction via its positive effects on collateral kin devalued by its genetic relatedness to them. Altruistic behavior can be selected for under such 'kin selection' if $rb-c>0$, where r is the relatedness of actor to recipient (the probability that a focal gene in the actor is present in the recipient), b is benefit to the recipient, and c is cost to the actor. Inclusive fitness theory predicts that genetic relatedness should be high among individuals in social systems with altruism. In accordance with this prediction, substantial levels of genetic relatednesses have been found between interacting individuals in the overwhelming majority of cooperatively-breeding and eusocial species thus far studied (Choe and Crespi 1997). By contrast, the few communal species for which such data are available exhibit colonies comprised of non-relatives or distantly-related individuals. All told, inclusive fitness theory has gained tremendous empirical support and is one of the cornerstones of social evolution theory.

Altruism can also evolve if it is reciprocal, such that the altruistic act of one individual to another is later reciprocated (Trivers 1971) (Table 1). By this mechanism, individuals need not be relatives, but there must be some means whereby cheating is thwarted lest the system break down. Such means may include recognition of individuals or the building of 'reputations' and as such reciprocal altruism may be limited to species with highly-developed cognitive abilities, such as primates.

3. *Ecology*

The ecology of a species comprises its habitat, food sources, demography, competitors, predators, parasites, and mutualists. The ecological niche of a species, and the nature of its interspecific and intraspecific interactions, form the selective context and selective agents for the evolution of social behavior (Evans 1977, Strassmann and Queller 1989).

3.1 *Habitat*

Cooperation in reproduction is concentrated in habitats that select for juveniles to stay at home after reaching adulthood, because of benefits to philopatry and costs to dispersal and independent reproduction. Among social vertebrates, the preponderance of social species are territorial and live in environments where independent breeding opportunities appear more limited than in related non-social taxa (Brown 1974, Emlen 1982). Indeed, many studies have shown a positive association between the degree of 'ecological constraint' on independent reproduction (measured in terms of some aspect of habitat availability) and the extent to which the offspring of a set of parents stay at home to serve as 'helpers,' forming a cooperative-breeding social system. By staying at home, such offspring may be able to increase their inclusive fitness via altruism towards relatives, they may eventually inherit the territory, or they may simply gain a favorable environment in which to wait for a breeding opportunity to arise elsewhere.

Among social invertebrates, sociality is more or less restricted to species that live in improvable, expansible, or defensible microhabitats that serve either as nurseries or as combined nurseries and food-rich sites (Alexander et al. 1991). For example, social aphids and thrips live in galls (hollow modified leaves or stems), social encyrtid wasps live in the body cavities of their hosts, social bark beetles live in the heartwood of trees, social snapping shrimp live in sponges, and most social Hymenoptera live in nests in the ground, in stems, or in wood (Choe and Crespi 1997). These habitats are especially valuable and tend to select for offspring philopatry and helping behavior, in the same way as do the territories of social vertebrates (Strassmann and Queller 1989). The two species of social mole rat provide a striking link between the invertebrates and vertebrates: both live in colonies underground in expansible burrows like bees, ants, or

termites, and, like social thrips, aphids, shrimp, and some termites, their food is available at the nest site.

Although cooperation is clearly associated with particular sorts of habitat in both vertebrates and invertebrates, few studies have yet to compare social taxa to closely-related non-social taxa, with regard to the specific characteristics of habitats believed to be important. Moreover, researchers are only beginning to link differences in habitat characteristics to differences in social system, to discern what aspects of habitat tend to select for communal, cooperative breeding, or eusocial behavior. The most obvious pattern to date is that virtually all species living in habitats that combine food and shelter, such that the colony members need not leave their domicile during the period of rearing offspring, exhibit eusociality, if they also have predators and the ability to defend against them (Crespi 1994).

3.2 Natural Enemies

Because the habitats of social animals are so rich in resources such as food, helpless juveniles, and shelter, they are prime targets for attack by a multitude of natural enemies, including predators, competitors and parasites (Evans 1977, Strassmann and Queller 1989). These enemies represent strong selective forces for cooperation because they can decimate entire social groups of invertebrates and because defense against them often involves the evolution of behavioral and morphological specializations that trade off with reproduction. Almost all social species have some means of defense, such as the stings of Hymenoptera, the mouth parts or modified forelegs of soldier aphids, the armed forelegs of soldier thrips, the incisive teeth of mole rats, the claws of shrimp, or the diversity of chemical and physical defenses of soldier termites (Choe and Crespi 1997). Some of these defensive traits apparently served as pre-adaptations for the origin of sociality (e.g., stings, which evolved in a non-social context and may have facilitated the origin of sociality), whereas other traits, such as those exhibited by the soldiers of termites, clearly evolved in the context of social defense.

In contrast to invertebrates, vertebrates have few of the types of natural enemies that select for sociality, except in some taxa (e.g., mongooses) that are highly vulnerable to attack by other vertebrates (Alexander et al. 1991). For most vertebrates, intraspecific competition for resources, both between and within social groups, appears to be a more important selective force for cooperation.

3.3 Demography

Social behavior evolves within the context of life history and aspects of demography, such as birth and death schedules and lifespans, have come under increasing scrutiny as factors that may influence social systems and be influenced by them. One demographic benefit of sociality is that an individual may help to rear dependent relatives when it is young, before it would normally reproduce on its own. If mortality rates of adults are sufficiently high, relative to the duration of offspring dependence, then such inclusive fitness 'head starts' may greatly increase the benefits of helping (Queller 1989). Moreover, extrinsic mortality rates of helpers are expected to be higher than those of reproductives, which results in faster senescence of helpers, extended lifespans of the reproductives, and an accelerated divergence of castes in behavior and morphology (Alexander et al. 1991). Indeed, in many Hymenoptera species, reproductives can live for many years, while helpers only survive for a few weeks or months.

In some invertebrates (i.e., gall thrips and gall aphids), sociality appears to be associated with a longer duration of the habitat (the gall) than in closely-related non-social forms (Choe and Crespi 1997). Moreover, cooperatively-breeding bird species appear to be longer lived than their non-social relatives (Arnold and Owens 1998). In both cases, the direction of casuation is unclear. In birds, cooperation may be favored by longer lifespans leading to stronger ecological constraints, or longer lifespans may result from delayed breeding. Similarly, longer-lived galls may be more vulnerable to attack, leading to the evolution of soldier castes, or having soldiers may allow for an extended gall duration. Because demography is so closely tied to fitness, assessment of its role in social evolution is one of the main imperatives of current research.

4. Mechanisms of Cooperation

Social behavior can arise not only via benefits gained by cooperators, through mutualism, altruism, or reciprocity, but also via costs imposed on non-cooperators, through manipulation, punishment, or policing, or via dominant individuals forcing others to cooperate (Table 1). Determining under what ecological and genetic-relatedness circumstances these different mechanisms evolve is critical to further progress in the study of social evolution.

5. Synthesis of Relatedness, Ecology and Mechanism

Explaining and predicting the evolution of social systems requires the development of models that combine relatedness, ecology, and mechanism into a common framework. Models of 'reproductive skew' (i.e., variation in reproductive success) parameterize aspects of ecology into two components: (1) the degree of ecological constraint on independent reproduction, and (2) the productivity benefits of cooperation (i.e.,

the benefits of cooperation and division of labor in terms of offspring production) (Vehrencamp 1983, Johnstone 2000). These two variables, plus genetic relatednesses, enter into models that seek to predict under what genetic-relatedness and ecological conditions a pair of individuals will form a stable cooperative group and to what degree reproduction in the group is skewed in favor of the dominant. A critical assumption of most skew models is the mechanism of reproductive control: the dominant individual controls the extent of reproduction of the subordinate and can thus concede reproduction to the subordinate to motivate it to stay and help, rather than leave and attempt independent reproduction. The degree to which the assumption of dominant control over reproduction is satisfied in nature is currently a matter of considerable dispute (Johnstone 2000).

Skew models have predicted a number of observed broad-scale patterns, such as low relatedness and low ecological constraint in communal social systems versus high relatedness and constraint under cooperative breeding. However, their usefulness is still limited, because analysis of their assumptions, the ecological and genetic variables involved, and their predictions is empirically so challenging. Future development of skew models requires (1) expansion of their domain of applicability to include both intersexual and intrasexual interactions, (2) enlarging their range of assumptions regarding control of reproduction, (3) consideration of additional mechanisms of cooperation, such as manipulation (Alexander 1974), policing, eviction, and punishment (Table 1), (4) incorporation of demography and an explicit element of time into the models, and (5) feedback between the findings of empirical studies and the assumptions of the models. Cooperation can arise in a wide variety of ways and a hierarchical set of skew models, encompassing this diversity of mechanisms, will spur further progress in understanding how sociality evolves.

See also: Altruism and Prosocial Behavior, Sociology of; Collective Behavior, Sociology of; Evolutionary Social Psychology; Group Processes, Social Psychology of; Groups, Sociology of; Human Behavioral Ecology; Primate Socioecology; Primates, Social Behavior of; Social Evolution, Sociology of; Social Psychology; Social Psychology, Theories of; Sociality: Anthropological Aspects

Bibliography

Alexander R D 1974 The evolution of social behavior. *Annual Review of Ecology and Systematics* **5**: 325–83

Alexander R D, Noonan K M, Crespi B J 1991 The evolution of eusociality. In: Sherman P W, Jarvis J U M, Alexander R D (eds.) *The Biology of the Naked Mole Rat*. Princeton University Press, Princeton NJ, pp. 3–44

Arnold K E, Owens I P F 1998 Cooperative breeding in birds: A comparative test of the life history hypothesis. *Proceedings of the Royal Society of London Series B* **265**: 739–45

Brown J L 1974 Alternative routes to sociality in jays – with a theory for the evolution of altruism and communal breeding. *American Zoologist* **14**: 63–80

Choe J C, Crespi B J 1997 *The Evolution of Social Behavior in Insects and Arachnids*. Cambridge University Press, Cambridge UK

Crespi B J 1994 Three conditions for the evolution of eusociality: are they sufficient? *Insectes Sociaux* **41**: 395–400

Emlen S T 1982 The evolution of helping. I. An ecological constraints model. *American Naturalist* **119**: 29–39

Evans H E 1977 Extrinsic versus intrinsic factors in the evolution of insect sociality. *BioScience* **27**: 613–7

Hamilton W D 1964 The genetical evolution of social behavior, I and II. *Journal of Theoretical Biology* **7**: 1–52

Johnstone R A 2000 Models of reproductive skew: a review and synthesis. *Ethology* **106**: 5–26

Queller D C 1989 The evolution of eusociality: Reproductive head starts of workers. *Proceedings of the National Academy of Sciences of the USA* **86**: 3224–6

Strassmann J E, Queller D C 1989 Ecological determinants of social evolution. In: Breed M D, Page R E Jr (eds.) *The Genetics of Social Evolution*. Westview Press, Boulder CO, pp. 81–101

Trivers R L 1971 The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**: 35–57

Vehrencamp S L 1983 A model for the evolution of despotic versus egalitarian societies. *Animal Behavior* **31**: 667–82

B. Crespi

Socialization and Education: Theoretical Perspectives

Socialization refers to all processes by which an individual learns the ways, ideas, beliefs, values, and norms of his particular culture and adapts them as part of his own personality in order to function in his given culture. Education, on the other hand, deals explicitly with intentional attempts at influencing an individual's development course. In the following, some core anthropological assumptions guiding the thinking and theorizing about socialization and education will be described. Second, major units of analysis, in particular phases of individual development, topics of socialization, socializing institutions and cultural variations will be briefly addressed. Third, an overview of three groups of theoretical approaches to socialization and educational processes, i.e., biological, psychological and sociological theories, will be presented. Finally, some elements of an integrative theoretical framework for studying socialization and educational processes will be outlined.

1. Core Anthropological Assumptions

At the heart of all theorizing about socialization and education are central prototheoretical beliefs concerning the nature and development of human beings. In fact, as Montada (1998) has pointed out, there are at least four distinguishable models of individual development depending on whether the individual and/or his environment are conceived of playing an active or passive part. Besides endogenistic, exogenistic, and self-creation theories exists a fourth group of theories which views both person and environment as active agents.

The latter group of theories rests on the common assumption that individual development is a dynamic interactional process which evolves in a sequence of reciprocally patterned person–environment transactions. Besides this basic assumption, there are at least six additional aspects that are central to a transactionals conceptualization of human development, i.e., (a) learning capacity, (b) social dependence, (c) experience formation, (d) internal representations of experiences, (e) action competence, and (f) freedom of choice. Action competence and freedom of choice, in particular, are indispensable prerequisites for innovation and change concerning individual and sociocultural conditions of life.

2. Perspectives in Studying Socialization and Education

When looking at socialization and educational processes from an empirical point of view there are at least four different, albeit interrelated, approaches that can be distinguished, i.e., a chronological, topical, institutional, and cultural approach.

In a *chronological perspective*, the course of individual development can be divided in different age-graded phases such as infancy, childhood, adolescence, and early, middle, and late adulthood. The sequence of phases and corresponding transitions is structured by individual developmental tasks (Havighurst 1953) which, on one hand, are based on physical and social cognitive maturation and, on the other, on normative expectations prevalent in a given society.

From a *topical perspective*, socialization processes have a strong impact on certain life domains such as (a) physical growth and health, (b) cognitive, social, moral, and spiritual development, (c) sex and gender differences, and (d) self and personality development.

From the vantage point of *institutions*, socialization takes place in different settings which, in some instances, are especially designed to advance the development of a broad spectrum of individual dispositions and competencies, either by formal or informal learning processes, i.e., families, schools, peer-groups, work settings, and so on.

Finally, from a *cultural* point of view, an individual's development depends a great deal on the values and beliefs that are typical of the culture to which he or she belongs. It has been shown, for example, that the prevalence of an independent-autonomous conception of the self within individualistic cultures vs. an interdependent–relational self within collectivistic cultures is closely related to distinct socialization goals and practices stressing either autonomy or relatedness (Triandis 1995).

3. Major Theoretical Approaches to Socialization and Education

Although in the present context it is not possible to give a full account of the vast array of theories dealing with the structural and dynamic aspects of socialization and education, three major classes of theories, i.e., biological, psychological, and sociological, will be briefly reviewed.

3.1 Biological Theories

3.1.1 Sociobiology and evolutionary psychology. Based on Darwin's groundbreaking ideas concerning the phylogenetic evolution of the species, sociobiology and, somewhat later, evolutionary psychology, contended that ultimate explanations of human behavior must take into account the superior goals of securing and expanding the successful reproduction of the human species. Especially mate selection and parental investment have been associated with universally evolved psychological mechanisms that are rooted in the human genotype as a consequence of phylogenetic adaptation. For example, Belsky et al. (1991) have proposed and also presented some research evidence for an evolutionary theory of socialization relating different kinds of reproductive strategies to specific childhood experiences, which in turn entail specific outcomes of an individual's psychological and somatic development. More specifically, these authors argue for two distinct developmental pathways of reproductive strategy. One is characterized by stressful contextual and rearing environments, insecure attachments, early pubertal development and onset of sexual activity within the context of short-term, unstable pair bonds, which in turn predispose to higher fertility rates and limited parental investment. In contrast, the other developmental trajectory shows the opposite pattern of rearing conditions and interpersonal relationships.

3.1.2 Behavior genetics. While evolutionary theories focus on universal adaptive strategies, behavior genetic theories concentrate on explaining the variability of individual differences that can be accounted for by genetic and/or environmental influences

(Plomin 1994). Behavior geneticists, in particular, have shown that individual differences in cognitive and socio-emotional dispositions are not only substantially influenced by the individuals' genotype but also by genetic components of the environment (e.g., via the parents' genotype impacting on their children's environment). Moreover, besides the conceptually and empirically useful distinction between shared and non-shared environments, various kinds of genotype-environment covariation and interaction have been distinguished to account for genetically informed processes of individual development. In this vein, empirical analyses of cross-sectional and longitudinal data have led some authors to challenge the environmentalist position according to which, for example, familial influences such as parenting style or family climate have a strong impact on the development of individual differences in the filial generation.

3.2 Psychological Theories

There is a wide range of psychological theories that have been applied to describing, explaining, and controlling socialization and educational processes. In fact, all theories dealing with personality and social development have much to contribute to this goal. In the following, four theoretical approaches will be briefly described with a special emphasis on the processes of socialization and education (Schneewind and Pekrun 1994).

3.2.1 Psychoanalytical theories. In classical Freudian psychoanalytical theory the operation of defense mechanisms such as identification or sublimation plays a central role in socializing the growing individual's initially untamed libidinal and aggressive drives. Especially, coping with the oedipal situation via identification with the aggressor is supposed to be a developmental milestone, because it allows the young boy or girl to enormously expand his or her learning experiences, thus leading to a gradually more realistic and differentiated satisfaction of formerly more primitive expressions of his or her drive impulses. Later on, sublimination is the most elaborate and mature defense mechanism, which provides the individual with new ways in his private or professional life to follow through with his otherwise unsatisfied libidinal and aggressive needs at a socially acceptable level.

Fixation is another important psychoanalytical concept that has been invoked to explain character formation at different stages of an individual's psychosexual development, depending on either a lack, inconsistency, or abundance of gratifications at a particular developmental stage. Thus, distinct patterns of socialization practices at different psychosexual stages have been related to particular outcomes of children's personality development. Cultural anthropologists, in particular, have followed this line of reasoning in contrasting more lenient vs. strict socialization practices in various cultures to demonstrate, albeit with only partial success, their impact on character formation (e.g., Whiting and Child 1953).

While the classical Freudian psychoanalytical model is based on the transformation of primitive into more differentiated ways of need satisfaction, neoanalytic theories, especially object relations theories, approach the issue of socialization and personality development quite differently. In object relations theories the central dynamic principle is a person's need for interpersonal relatedness (Greenberg and Mitchell 1983). In a developmental perspective, the individual acquires internal representations of the self, the other, and the relationship between self and other, which tend to become more and more differentiated. Thus, the quality of relationships that an individual is engaged in appears to be vitally important with respect to his internal representations of self and other. Empirical studies, however, are scarce and have mostly been limited to the pathology of object relations

3.2.2 Attachment theory. Although object relations theories bear some resemblance to attachment theory, the empirical status of the latter is much more elaborated. Attachment theory which combines concepts drawn from ethology, psychoanalysis, cognitive psychology, and systems theory has sparked off a multitude of theoretical, methodological, and empirical refinements (Cassidy and Shaver 1999). Of particular importance is the discovery of various attachment types (i.e., secure, resistant, avoidant and disorganized/disoriented attachment in infancy; secure, preoccupied, anxious, and dismissing attachment in adulthood) and corresponding internal working models. Moreover, it has been shown that specific aspects of socialization are related to different attachment types in theoretically meaningful ways. For example, several aspects of caregiving that promote secure mother–infant attachments have been found. These include characteristics such as sensitivity, positive attitude, synchrony, mutuality, support and stimulation. Similarly, a number of distinct interpersonal behaviors such as mutual need for intimacy, exchange of positive affect, and constructive conflict resolution have been observed in adult relationships belonging to the secure type.

3.2.3 Epigenetic theories. The common feature of epigenetic theories is that individual development is conceived as a sequence of qualitatively distinct phases concerning the structure of experience and behavior. A prominent example is Piaget's (1954) theory of cognitive development, comprising four

major periods from the sensorimotor period to that of formal operations. These are associated with specific cognitive schemes that are built on the interplay of assimilation and accommodation processes. The epigenetic principle has influenced a series of theoretical conceptualizations concerning other developmental domains including moral reasoning, distributive justice, social perspective taking, or ego development. While Piaget emphasized the epigenetic sequence of development as an inherent process of self-creation, Vygotsky (1978) put a stronger emphasis on the cultural influences on development. In particular, his concept of the zone of proximal development clearly stresses the importance of socialization and education in the progression of epigentically patterned levels of development.

3.2.4 Learning theories. This class of theories covers a wide range of theoretical approaches spanning from orthodox behaviorism characteristic of Watson and Skinner to social–cognitive and action–theoretic conceptualizations of learning. Although the principles of classical and operant conditioning are by no means outdated, the so-called cognitive revolution in psychology led to a paradigmatic shift in learning theories. One of the earliest contributions to this new type of theories was Rotter's (1972) social learning theory of personality which combines basic principles of learning within an expectancy–value framework. In particular, Rotter's concept of generalized expectancies for internal vs. external control of reinforcement and the social conditions shaping these expectancies, especially within the family and in the school, stimulated a tremendous amount of research, presumably because of the affinity of this construct to other socially relevant concepts such as autonomy, personal control, helplessness, or alienation stemming from quite different theoretical positions.

More recently, and in a similar vein, Bandura's (1999) social cognitive learning theory of personality has contributed much to the analysis and theory-guided application of specific procedures fostering the development of the self system by socialization and self-directed action. Besides his influential theorizing and research on the processes of observational learning and modeling Bandura has focused on the construct of self-efficacy, i.e., an individual's belief that he or she possesses the necessary competencies to attain specific goals. In addition, Bandura has shown that the concept of self-efficacy can also be extended to collective efficacy beliefs of groups or even societies.

3.3 Sociological Theories

Not unlike psychological theories there exist quite a few theoretical approaches that have been created within sociology to account for the processes of socialization and education. In the following, three will be briefly described because of their historical and contemporary importance.

3.3.1 Structural functionalism. The theory of structural functionalism is closely associated with the work of Parsons who has written more than 150 articles and books on this topic. The main question that Parsons addresses in his theory refers to processes of how individuals become members of a given society in order to guarantee the survival and maintenance of the social system (see Parsons and Bales 1955). According to Parsons, society is a global social system which is based on an integrated value system. The individual person participates in the social system by interacting with others in accordance with the various roles and positions he or she holds in that system. The global social system itself consists of hierarchically ordered subsystems which are characterized by corresponding institutionalized norms. On the one hand, these norms are supposed to be congruent with society's integrated value system and, on the other hand, they determine the expectations and rules attached to specific positions and roles. These are further specified with respect to a set of pattern variables comprising, for example, particularistic vs. universalistic values. In this respect, the family plays a particularly salient role (Parsons and Bales 1955). Drawing heavily on Freudian psychoanalytic theory Parsons contends that via identification with their same-sex parents boys and girls become socialized into specific male and female role patterns. More precisely, gender-related role specialization is characterized by instrumental or task-oriented roles for males and expressive or person-oriented roles for females. Again, these gender-specific complementary roles are supposed to maintain the equilibrium of the family system as well as the social system as a whole.

3.3.2 Symbolic interactionism. While Parsons' theory of structural functionalism focused on the conditions ensuring the conservation and stability of the social system, symbolic interactionism called into question whether the expectations and behaviors associated with the enactment of roles can be determined as precisely as Parsons suggested. Instead, Mead (1934), who is credited as the originator of the symbolic interaction framework, maintained that social roles are acquired by imparting and sharing the meaning of symbol systems, particularly that provided by language, in order to ensure an appropriate coordination of the actions of two or more interacting persons. The central vehicle by which this can be achieved is the process of communication which makes role-taking a necessary prerequisite. In essence, the shared meaning of symbols is a co-constructive process which is an outgrowth of interpersonal communication. Therefore, the notion of

socialization is crucial in symbolic interactionism although Mead himself never used this term. Instead he talked about the 'importation' of social symbols into a person's mind. For the developing child the process of importation is marked by two stages, i.e., a play and a game stage. In the play stage, the child is an actor with his own needs and interpretation of the situation (which in Mead's terms represents the self as subject or 'I') while in the game stage the child is an actor who is confronted with the needs and social expectations of other actors (which refers to Mead's self as object or 'Me'). In certain times and situations, however, balancing the I and Me aspects of the self may become quite conflictual, thus calling for appropriate means of intra-and interpersonal conflict resolution.

3.3.3 Conflict theories. Generally speaking, conflict within and between social groups is inevitable because the needs and interests of individuals and in-groups are more or less incompatible with the needs and interests of other individuals or groups. At the core of conflict theories is the phenomenon of power, or more precisely the imbalance of power, in particular when the powerful determine the lives and resources of the underprivileged, and at the same time ignore to a great extent their needs and interests. Conflict theories as, for example, Marx' and Engels' dialectical theory of history, try to understand the conditions leading to social inequality and injustice. However, Marxism not only provides an explanatory framework for the understanding of social inequality and conflict; it also has an emancipatory element inasmuch as it urges political action to dissolve inequality by endowing the underprivileged with more power and control to allow for a more self-determined conduct of life. Marx and Engels used their theory to explain the differences and conflict between social classes. In principle, however, Marxist theorizing can be applied to all social constellations where inequality and lack of justice cause conflict as, for instance, in feminist theories that deal with social gender inequalities.

Unresolved conflict and the plea for emancipation call for appropriate means of conflict resolution which cover a wide range from physical aggression to negotiation between conflict parties. In order to ensure that basic human rights are not violated, socialization for negotiation and constructive conflict resolution is a vital prerequisite. At a philosophical level, Habermas (1981), a prominent contemporary proponent of critical theory, has suggested to develop the art of 'communicative competence' to handle conflicts in a so-called 'ideal discourse situation.' This means that all participants in discourse are granted equal opportunity to express their own point of view and to comment critically on the contributions of their fellow participants, keeping in mind that the ultimate goal is to find a solution that all participants can live with. At a more practical level, specific procedures have been worked out to enhance constructive communication and conflict resolution skills, particularly for the family, school, and work context.

4. *Elements of an Integrative Framework for Studying Socialization and Education*

Based on the core anthropological assumptions and theoretical approaches described in the preceding sections, an integrative framework for the study of socialization and educational processes must at least cover the following five aspects.

4.1 *Units of Environmental Influence*

Since the developing person is a part of different and progressively expanding environments, the units of analysis should comprise all settings that may exert direct or indirect influences on the individual. Bronfenbrenner's (1979) ecological systems approach to human development provides an excellent framework for this purpose. By distinguishing between four systemically interrelated levels of developmental con-t-exts, i.e., the micro-, meso-, exo-, and macrosystem, the scope of the framework is sufficiently broad to encompass all possible socializing influences that an individual might have to deal with. More specifically, Super and Harkness (1994) have introduced the concept of 'developmental niche' to describe the conditions that constitute the developmental context of a particular person, especially a particular child. According to these authors, there are three major components of the developmental niche, i.e. (a) the physical and social contexts of everyday life, (b) culturally regulated customs of child care and child rearing, and (c) particular psychological characteristics of a child's parents. Together these three components constitute the proximal context which determines the child's experiential field. On the whole, broader ecological systems as well as more specific aspects of the developmental niche form the opportunity structures of the developing person.

4.2 *Characteristics of the Individuals to be Socialized*

Within systems of interacting persons the characteristics of the individuals to be socialized play an important role in shaping the processes and outcome of socialization. Particularly salient individual characteristics that moderate and mediate socializing influences are (a) basic personal characteristics (e.g., genotype, health, and maturational status of cognitive, motivational, and temperamental dispositions), (b)

characteristic adaptational strategies (e.g., information processing, coping and defense mechanisms, and volitional and emotion regulation processes), and (c) typical forms of experiencing the self and the world (e.g., self-concepts, experience of personal identity and coherence, and internal representations of relationships, situations, and environments).

4.3 Proximal Processes

To combine the characteristics of a person and his or her environment in a meaningful way, proximal processes, i.e., reciprocal interactions between the human organism and the persons, objects, and symbols in its environment (Bronfenbrenner 1995), are presumed to produce specific socialization outcomes. More specifically, Caspi (1997) has called attention to three kinds of person–environment transactions explaining developmental diversity and continuity. They are (a) reactive person–environment transactions which occur when different individuals express different forms of experience and behavior when exposed to the same situation; (b) evocative person–environment transactions, i.e., when an individual's specific personality characteristics elicit particular responses from other persons; and (c) proactive person–environment transactions which refer to the selection and creation of environments by the individuals themselves.

4.4 Goals and Values

In the behavioral and social sciences there is disagreement as to whether or not goals and values governing socialization and education should be incorporated as normative statements in corresponding theories, as is clearly the case in conflict theories of socialization. Although goals and values are undoubtedly objects of socialization theories and research, this controversy remains unresolved. However, even if the originators of socialization theories deliberately restrict themselves to descriptive and explanatory statements, such theories can nevertheless contribute to a great extent to the clarification of socialization goals, and to a better understanding of their antecedents and consequences. Thus, theories of socialization, and, particularly, corresponding research findings, provide valuable information which can be used to make informed choices with respect to one's own beliefs.

4.5 Self-socialization

Based on the anthropological assumptions presented above, socialization consists not only, as Mead (1934) contended, of the 'importation' of social symbols into an individual's mind, but deals also with the capacity to select and create personally and socially meaningful goals, due to the individual's progressively increasing action competence and freedom of choice. The latter process refers to the principle of self-socialization within the limits of personal and socio-cultural opportunity structures. In this context, human beings are creating themselves and the culture in which they live. Moreover, self-socialization enables them to work on the life-long project of designing their lives in a personally and communally rewarding way. Thus, theories of socialization and education are not only an outgrowth of humans' cultural activity but serve also as a critical guide for making well-informed choices to optimize the process of self-socialization.

See also: Behavioral Genetics: Psychological Perspectives; Cultural Diversity, Human Development, and Education; Educational Sociology; Learning and Instruction: Social-cognitive Perspectives; Learning Theories and Educational Paradigms; Social Inequality and Schooling; Socialization in Adolescence; Socialization, Sociology of

Bibliography

Bandura A 1999 Social cognitive theory of personality. In: Cervone D, Shoda Y (eds.) *The Coherence of Personality*. Guilford Press, New York

Belsky J, Steinberg L, Draper P 1991 Childhood experience, interpersonal development, and reproductive strategy: An evolutionary theory of socialization. *Child Development* **62**: 647–70

Bronfenbrenner U 1979 *The Ecology of Human Development*. Harvard University Press, Cambridge, MA

Bronfenbrenner U 1995 Developmental ecology through space and time: A future perspective. In: Moen P, Elder G H Jr, Lüscher K (eds.) *Examining Lives in Context*. American Psychological Association, Washington, DC

Caspi A 1997 Personality development across the life course. In: Eisenberg N (ed.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, Vol. 3

Cassidy J, Shaver R P (eds.) 1999 *Handbook of Attachment: Theory, Research, and Clinical Applications*. Guilford Press, New York

Greenberg J R, Mitchell S A 1983 *Object Relations in Psychoanalytic Theory*. Harvard University Press, Cambridge, MA

Habermas J 1981 *Theorie des kommunikativen Handelns*. Suhrkamp, Frankfurt

Havighurst R J 1953 *Human Development and Education*. Longmans Green, New York

Mead G H 1934 *Mind, Self and Society*. University of Chicago Press, Chicago

Montada L 1998 Fragen, Konzepte, Perspektiven. In: Oerter R, Montada L (eds.) *Entwicklungspsychologie*. Psychologie Verlags Union, Weinheim

Parsons T, Bales R F 1955 *Family Socialization and Interaction Process*. Free Press, New York
Piaget J 1954 *The Construction of Reality in the Child*. Basic, New York
Plomin R 1994 *Genetics and Experience: The Interplay Between Nature and Nurture*. Sage, Thousand Oaks, CA
Rotter J B 1972 An introduction to social learning theory. In: Rotter J B, Chance J E, Phares E J (eds.) *Application of a Social Learning Theory of Personality*. Holt, Rinehart and Winston, New York
Schneewind K A, Pekrun R 1994 Theorien der Erziehungs- und Sozialisationspsychologie. In: Schneewind K A (ed.) *Psychologie der Erziehung und Sozialisation*. Hogrefe, Göttingen
Super C M, Harkness S 1994 The developmental niche. In: Lonner W J, Malpass R (eds.) *Psychology and Culture*. Allyn and Bacon, Boston
Triandis H C 1995 *Individualism and Collectivism*. Westview, Boulder, CO
Vygotsky L S 1978 *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA
Whiting J W M, Child I L 1953 *Child Training and Personality: A Cross-cultural Study*. Yale University Press, New Haven, CT

K. A. Schneewind

Socialization in Adolescence

Socialization is the process through which individuals' behavior, attitudes, and values are shaped by the individuals and social institutions to which they are exposed. Because adolescence is a transitional time during which individuals are prepared for the roles they will occupy as adults, this period of development has received a great deal of attention from researchers interested in socialization. Among the issues most often researched are influences on academic achievement, problem behavior, and social competence.

Although we often think of the family as the main agent of socialization in children's development, young people are influenced by forces outside the family as well, including peers, schools, and broader cultural institutions, such as the mass media. The influence of nonfamilial agents of socialization is especially important in adolescence, as individuals begin to spend increasingly more time away from their parents and exposed to the potential influence of other socialization agents. Not surprisingly, much attention has been devoted in the study of socialization in adolescence to questions concerning the ability of parents to influence their children in the face of these other forces.

Research on socialization in adolescence has focused mainly on three broad questions. First, to what extent are adolescents influenced by different agents of socialization such as the family or peer group? Second, which aspects of adolescent development (e.g., school performance, drug use, prosocial behavior) are most influenced by these various socialization agents? Finally, what characteristics of individuals (e.g., age, gender, upbringing) make them more or less susceptible to different sources of influence?

1. *A Brief History of Research on Adolescent Socialization*

Early studies of socialization in adolescence concentrated on the influence of parents. Building on the prior work of scholars who had studied socialization during early and middle childhood, students of adolescent socialization examined different forms of parental discipline and their impact on the adolescent's character development. The dramatic increase in the relative size of the adolescent population during the 1960s, however, as well as growing concerns about the segregation of adolescents from adults created by the growth of secondary and postsecondary education, fueled scholarly interest in the study of adolescent peer groups and the potential influence of adolescents on each other. The publication of James Coleman's book, *The Adolescent Society* (1961) popularized the notion that young people were growing up in a separate world in which they were exposed to attitudes and values not shared by adults. Warnings about the development of a 'counterculture' provided the impetus for much of the socialization research conducted during the 1960s and 1970s, on such topics as peer pressure and the generation gap.

The spread of popular culture during the latter part of the twentieth century broadened the study of adolescent socialization further, as society began to question the impact of such forces as television, popular music, and film. The mass media were seen as responsible for a variety of social ills afflicting adolescents, such as sexual promiscuity, drug use, and violence. As was the case in earlier studies of the peer group, the mass media were cast as evil influences on adolescent behavior, attitudes, and values, and parents were portrayed as struggling to overcome this insidious force.

2. *General Challenges in Socialization Research*

Although the goal of understanding how adolescents are influenced by the individuals and institutions to which they are exposed is straightforward, answering questions about the strength of different socialization agents has proven very difficult, for a number of reasons. One substantial challenge, particularly in the study of socialization in the family, involves separating genetic from environmental influences on the adolescent. Adolescents may behave in ways that reflect their parents' interests and values, but similarity between children and their parents can be due both to

the influence of parents on children as well as to the fact that parents and children share certain genes in common. Toward the end of the twentieth century, advances in the study of behavioral genetics challenged socialization researchers to demonstrate that what they had described as parental influence was actually an effect of socialization, and not merely a reflection of shared genetic background (Plomin 1990). Although today this issue remains far from resolved, it is clear that genetic factors play a far more important role in adolescent personality and social development than had been recognized, and that much of what had been viewed as parental influence is attributable to heredity rather than socialization (Reiss et al. 2000).

A different but no less formidable challenge faces researchers interested in the influence of peers, mass media, and other nonfamilial socialization agents (Harris 1995). Because adolescents choose their friends and select the media they are exposed to, it is difficult to know whether the apparent impact of these agents on adolescent development is actually due to their influence, rather than merely indicative of the choices that adolescents make about how they spend their time. Thus, for example, similarity between adolescents and their friends could be due to the impact that friends have on each other, but it may also be due to the fact that adolescents choose as their friends individuals who share their interests. Separating out selection from socialization can only be accomplished through experimental research or longitudinal studies that track individuals over time.

3. *Adolescent Socialization in the Family*

More is known about adolescent socialization in the family than about socialization in other contexts, although far more is known about the influence of parents than about other family members, such as siblings. Research on parental socialization has traditionally examined the links between variations in adolescent functioning and variations in parenting practices (the specific things that parents do to influence their child in one way or another, such as close monitoring of the child's whereabouts) or parenting styles (the general affective climate of the parent–child relationship). Over the past 25 years, several broad conclusions about socialization in the family context have emerged (Steinberg 1990).

First, most research indicates that adolescents fare better developmentally when their parents combine firm and consistent discipline with affection and involvement in the adolescent's life. This combination of 'demandingness' and 'responsiveness' is referred to as *authoritative parenting*. In many studies, adolescents raised by authoritative parents have been compared with their peers who come from homes that are *authoritarian* (demanding but relatively less responsive), *indulgent* (responsive but relatively less demanding), or *neglectful* (neither demanding nor responsive). Generally speaking, adolescents raised in authoritative homes score higher on measures of psychological adjustment and academic achievement and lower on measures of psychological and behavior problems than adolescents from other households. Adolescents from neglectful homes, in contrast, score lowest on measures of adjustment and highest on measures of maladjustment. Adolescents from authoritarian or indulgent homes generally score somewhere between the two extremes.

Second, research indicates that adolescent development is more closely linked to the nature of their family relationships than to the composition of their household. Studies of adolescents from single-parent homes or stepfamilies, for example, find only small differences between adolescents from these sorts of families and those who live with their two biological parents. Similarly small differences emerge in studies that compare adolescents whose mother works outside the home with those whose mother do not. For the most part, the benefits of authoritative parenting transcend variations in household composition. Most, but not all, research indicates that the benefits of authoritative parenting also cut across ethnic and socioeconomic groups.

Finally, research on socialization in the family indicates that the broader context within which the family lives affects the ways in which parents behave and, as a consequence, the impact that parents have on their adolescent's development. Authoritative parenting is more likely to occur in families with greater financial and social resources, and who live in neighborhoods populated with other well-functioning families. In contrast, studies of families living under stressful conditions, such as poverty, neighborhood deterioration, or high rates of crime, indicate that such conditions tend to foster parenting that is hostile, inconsistent, or uninvolved. As a result of these variations in the broader context of socialization, adolescents growing up under different circumstances often exhibit different levels of psychological adjustment.

A different line of research has examined changes in family relationships over the course of adolescence. This work has pointed to the importance of puberty as a period of significant transformation in family dynamics, although it is not clear whether these changes in relationships are triggered by the physical or hormonal changes of puberty *per se* or by other, co-occurring events (e.g., the transition out of elementary school, the development of abstract thinking). Whatever the causes, during puberty, adolescents often become more assertive toward their parents, and parent–child conflict, especially over issues concerning the adolescent's independence, tends to increase. Most families report that this phase is temporary, however, with relationships becoming closer and less conflicted toward the middle adolescent years (Steinberg 1990).

4. Adolescent Socialization in the Peer Group

Researchers have studied the socialization influence of peers at multiple levels of analysis, including close friends and their impact on one another, cliques and social networks, and crowds, the large amorphous groups that define the social terrain of most high schools. Scholars of adolescent peer relations maintain that it is important to distinguish among friends, cliques, and crowds, because these different levels of peer influence operate in different ways (Brown 1990). It is also important to distinguish among different arenas of socialization; peers are more influential over day-to-day activities than over enduring values, for example.

As a general rule, adolescents tend to be similar to the peers with whom they affiliate, a phenomenon known as 'homophily.' Although much of this similarity is due merely to the tendency for like-minded individuals to select each other as friends, there have been several longitudinal studies that indicate that friends do in fact socialize each other. The most widely documented effects have been in the areas of drug and alcohol use, delinquency, and academic achievement.

Less is known about the influence of *cliques* (groups of between six and 10 close friends) on adolescent development, but a number of studies have examined changes in individuals' susceptibility to peer pressure over the course of adolescence. This research shows that children's susceptibility to peer influence increases steadily between the ages of 9 and 14, then gradually declines as adolescents move into the later years of high school. Boys, on average, are more susceptible to peer pressure than girls, especially when the pressure is to engage in antisocial behavior. It is important to keep in mind, however, that most peer pressure is not antisocial, and that adolescents' cliques often pressure their members to avoid drugs and alcohol and perform well in school.

Studies of adolescents in Western industrialized cultures have also examined the role of *crowds* in adolescent socialization. Crowds are large, reputation-based groups that are used to locate adolescents within the school's social world; most schools have been found to have different crowds defined by athletics, academic commitment, antisocial activities, or social adeptness. By setting norms and establishing standards for such things as attire, tastes in popular culture, and leisure pursuits, crowds socialize their members to adopt some behaviors and attitudes and to eschew others.

Contemporary studies of socialization in the peer group have, like studies of the family, adopted a more ecological perspective on adolescent development. Instead of asking whether adolescents are more influenced by peers than by parents, for example, recent research has examined how parental and peer influence interact. Although there are exceptions to the rule, most studies in this vein indicate that parental and peer influence are synergistic, not antagonistic. More often than not, adolescents select friends whose values and attitudes reinforce, rather than undermine, the values and attitudes of the family. In addition, parents may influence the process of peer socialization by directing their children toward some peer groups and away from others.

5. Adolescent Socialization and Other Institutions

Although most work on adolescent socialization has focused on the family and the peer group, some research has examined the influence of social institutions, such as school, the workplace, the neighborhood, and the mass media. One problem in drawing conclusions from this body of work is that there is considerable variability among schools, work settings, neighborhoods, and the media to which adolescents are exposed. Thus, it is difficult to draw conclusions about the impact of these institutions on adolescent socialization without examining the specific features of the institution to which an individual adolescent is exposed. Unfortunately, research on adolescent socialization in these contexts has lagged behind that on the family and peer group in both quantity and sophistication.

The study of schools mainly has focused, as one would expect, on characteristics of schools or teachers that foster classroom engagement, achievement motivation, and scholastic success, and special attention has been paid to the need to make schools more sensitive and responsive to the particular psychological needs of adolescents. Much of this research indicates that effective schools and teachers share much in common with effective parents. Specifically, students are more strongly oriented to school and interested in achievement when they attend schools that are orderly and consistent in their goals, and staffed by teachers and administrators who are firm but supportive. These characteristics are reminiscent of the attributes associated with authoritative parenting described earlier (Entwisle 1990).

Research on school programs designed to socialize certain attitudes or behaviors among adolescents (e.g., programs designed to prevent drug use, aggression, or unsafe sexual activity) has not been especially encouraging, however. Most evaluations of such deliberate educational efforts indicate that they are modestly successful in transmitting information but largely unsuccessful in changing students' behavior. Most experts agree that school-based efforts of this sort need to be accompanied by programs aimed at parents, peers, and other community institutions, and that classroom programs in isolation from other sorts of interventions have little lasting influence over adolescents' socialization.

Several investigators have examined the role of the workplace in the socialization of adolescents, especial-

ly among adolescents who hold part-time jobs during the school year (Steinberg and Cauffman 1995). Because rates of student employment are so much higher in the United States than in other countries, almost all of this research has involved American samples. We do not know if these findings are generalizable to adolescents in other countries, because the nature of work and its relation to schooling varies considerably from country to country.

The results of research on student employment in the United States have indicated that the impact of work on adolescent socialization is to a great extent dependent on the amount of time adolescents spend on the job. In moderate amounts—10 or fewer hours per week—a part-time job can be a useful means of exposing adolescents to the world of work and helping them learn to manage their time more responsibly. When adolescents work longer hours, however, the costs of employment may overshadow the benefits. More specifically, studies indicate that employment during the school year in excess of 20 hours per week is associated with diminished performance in and commitment to school, and increased drug and alcohol use, and some research links long work hours to increased delinquency. These negative correlates of working are thought to accrue because long work hours interfere with school and diminish parents' ability to monitor their teenagers, both of which have been shown to be associated with poor school performance and increased problem behavior. In addition, adolescents who work long hours have large amounts of discretionary income, which may pull them away from school and toward more recreational pursuits.

Research on the impact of neighborhoods on adolescent socialization is relatively new. This work has examined how adolescents attitudes and behavior are shaped by variations in the communities in which they live along such dimensions as the neighborhood's income, employment rate, crime, and demography (Brooks-Gunn et al. 1997). Most research to date has found relatively few direct effects of neighborhood conditions on adolescent socialization, however. Rather, studies suggest that variations in neighborhoods affect adolescent development primarily through their impact on the family and peer group. For example, adolescents living in disorganized neighborhoods are themselves more likely to be involved in antisocial activity because neighborhood disorganization is associated with ineffective parenting and with insufficient monitoring of peer group activities. This sort of finding is consistent with the general view that socialization is mediated primarily through interactions with significant others.

Despite widespread popular belief in the deleterious effects of sex and violence in the mass media on the socialization of adolescents, scientific evidence on the impact of exposure to violent or sexual imagery—in television, film, music, or computer games—is sparse and inconsistent. There is some research indicating that exposure to televised violence may increase aggression among younger children, but this issue has not been adequately studied among adolescents. The study of the influence of other media (e.g., computer games, music) or of other types of media content (e.g., sexual imagery) on adolescent socialization is virtually nonexistent.

See also: Adolescence, Sociology of; Adolescent Behavior: Demographic; Adolescent Development, Theories of; Identity in Childhood and Adolescence; Social Learning, Cognition, and Personality Development; Socialization and Education: Theoretical Perspectives; Socialization in Infancy and Childhood; Socialization, Sociology of; Substance Abuse in Adolescents, Prevention of

Bibliography

Brooks-Gunn J, Duncan G J, Aber J L (eds.) 1997*Neighborhood Poverty*. Russell Sage Foundation, New York

Brown B 1990 Peer groups. In: Feldman S S, Elliott G R (eds.) *At the Threshold: The Developing Adolescent*. Harvard University Press, Cambridge, MA, pp. 171–96

Coleman J S 1961 *The Adolescent Society: the Social Life of the Teenager and its Impact on Education*. Free Press of Glencoe, New York

Entwisle D 1990 Schools and the adolescent. In: Feldman S S, Elliott G R (eds.) *At the Threshold: The Developing Adolescent*. Harvard University Press, Cambridge, MA, pp. 197–224

Harris J R 1995 Where is the child's environment? A group socialization theory of development. *Psychological Review* **102**: 458–89

Plomin R 1990 *Nature and Nurture: An Introduction to Human Behavioral Genetics*. Brooks Cole, Pacific Grove, CA

Reiss D et al. 2000 *The Relationship Code: Deciphering Genetic and Social Patterns in Adolescent Development*. Harvard University Press, Cambridge, MA

Steinberg L 1990 Autonomy, conflict, and harmony in the family relationship. In: Feldman S S, Elliot G R (eds.) *At The Threshold: The Developing Adolescent*. Harvard University Press, Cambridge, MA, pp. 255–76

Steinberg L, Cauffman E 1995 The impact of employment on adolescent development. In: Vasta R (ed.) *Annals of Child Development*. Jessica Kingsley Publishers, London, Vol. 11, pp. 131–66

L. Steinberg

Socialization in Infancy and Childhood

Socialization is the process whereby an individual's standards, skills, motives, attitudes, and behaviors change to conform to those regarded as desirable and appropriate for his or her present and future role in any particular society. Many agents play a role in the

socialization process including family, peers, schools, and the media. It is assumed that these various agents function together rather than independently. Families have been recognized as an early pervasive and highly influential context for socialization. Children are dependent on families for nurturance and support from an early age, which accounts, in part, for their prominence as a socialization agent.

1. Contemporary Perspectives on Family Socialization Theory

Several themes are evident in current theoretical approaches to socialization. First, the rise of systems theory (Sameroff 1994) has transformed the study of socialization from a parent–child focus to an emphasis on the family as a social system. To understand fully the nature of family relationships, it is necessary to recognize the interdependence among the roles and functions of all family members. It is being increasingly recognized that families are best viewed as social systems. Consequently, to understand the behavior of one member of a family, the complementary behaviors of other members also need to be recognized and assessed. For example, as men's roles in families shift, changes in women's roles in families must also be monitored.

A closely related theme involves the recognition of the importance of the historical time period in which the family interaction is taking place (Elder 1998). Historical time periods provide the social conditions for individual and family transitions: examples include the 1930s (the Great Depression) or the 1980s (Farm Belt Depression). Across these historical time periods, family interaction may, in fact, be quite different due to the unique conditions of the particular era. In order to understand the nature of parent–child relationships within families, a multilevel and dynamic approach is required. Multiple levels of analysis are necessary in order to capture the individual, dyadic, and family unit aspects of operation within the family itself as well as to reflect the embeddedness of families within a variety of extrafamilial social systems. The dynamic quality reflects the multiple developmental trajectories that warrant consideration in understanding the nature of families in infancy. Distinctions among different developmental trajectories, as well as social change and historical period effects, are important because these different forms of change do not always harmonize. For example, a family event such as the birth of a child may have more effects on a man who has just begun a career than on one who has advanced to a stable occupational position. Moreover, individual and family developmental trajectories are embedded within both the social conditions and the values of the historical time in which they exist. The role of parents as socialization agents is responsive to such fluctuations.

Consistent with a family systems viewpoint, recent research has focused on a variety of subsystems, including parent–child, marital, and sib–sib systems. We employ a tripartite model to understand the parent–child subsystem and the relation between this parent–child subsystem and children's social adaptation. The most common pathway is the parent–child relationship. Parents may influence their children through a second pathway namely in their role as direct instructor, educator, or consultant. In this role parents may explicitly set out to educate their children concerning appropriate norms, rules, and mores of the culture. In a third role, parents function as managers of their children's social lives and serve as regulators of opportunities for social contacts and cognitive experiences (Parke et al. 1994). In this article, we will focus on each of these pathways, as well as marital and sibling subsystems as contexts for socialization. Finally, we will examine the determinants of parental socialization strategies.

2. Quantitative and Qualitative Assessments of Mother and Father Involvement in Intact Families

In spite of current shifts in cultural attitudes concerning the appropriateness and desirability of shared roles and equal levels of participation in routine caregiving and interaction for mothers and fathers, the shifts are more apparent than real in the majority of intact families. Fathers spend less time with their infants, children, and adolescents than mothers not only in the USA but in other countries such as Great Britain, Australia, France, and Belgium (see Lamb 1987).

Fathers participate less than mothers in caregiving but spend a greater percentage of the time available for interaction in play activities than mothers do. The quality of play across mothers and fathers differs too. With infants and toddlers, fathers play more physically arousing games than mothers. In contrast, mothers play more conventional motor games or toy-mediated activities, and are more verbal and didactic (Parke 1996), although fathers in several other cultures, such as Sweden, India, and Central Africa, do not show this physical play style. As children develop, fathers become more involved in physical/outdoor play interactions and fixing things around the house and garden, where mothers are more actively involved in caregiving and household tasks, in school work, reading, playing with toys, and helping with arts and crafts. In adolescence, the quality of maternal and paternal involvement continues to differ. Just as in earlier developmental periods mothers and fathers may complement each other and provide models that reflect the tasks of adolescence—connectedness and separateness. Recent evidence suggests that fathers may help adolescents develop their own sense of identity and autonomy by being more 'peer-like' and more playful

(joking and teasing) which is likely to promote more equal and egalitarian exchanges (Larson and Richards 1994).

Why do mothers and fathers play differently? Both biological and environment factors probably play a role. Experience with infants, the amount of time spent with infants, the usual kinds of responsibilities that a parent assumes—all of these factors influence the parents' style of play. The fact that fathers spend less time with infants and children than mothers may contribute as well. Fathers may use their distinctive arousing style as a way to increase their salience in spite of more limited time. Biological factors cannot be ignored in light of the fact that male monkeys show the same rough and tumble physical style of play as American human fathers. Perhaps predisposing biological differences between males and females may play a role in the play patterns of mothers and fathers. At the same time, the cross-cultural data underscore the ways in which cultural and environmental contexts shape play patterns of mothers and fathers and remind us of the high degree of plasticity of human social behaviors.

3. *Assessing Parent–Child Interaction: Two Approaches*

Two approaches to this issue of the impact of parent–child interaction on children's socialization outcomes have been utilized: (a) a typological approach which focused on styles of child-rearing practices and (b) a social interaction approach which focused on the nature of the interchanges between parent and child.

3.1 *The Typological Approach*

The most influential typology has been offered by Baumrind (1973) who distinguished between three types of parental child-rearing typologies: authoritative, authoritarian, and permissive. Authoritative parents were not intrusive and did permit their children considerable freedom within reasonable limits, but were firm and willing to impose restrictions in areas in which they had greater knowledge or insight. In general, high warmth and moderate restrictiveness are associated with the development of self-esteem, adaptability, and social competence. In contrast, the authoritarian parents were rigid, power-assertive, harsh, and unresponsive to the children's needs. This results in the unhappy, conflicted, neurotic behavior often found in these children. Finally, in spite of the permissive parents' reasonably affectionate relationship with their children, their excessively lax and inconsistent discipline and encouragement of the free expression of their children's impulses were associated with the development of uncontrolled, impulsive behavior in their children.

Baumrind has followed these types of parents and their children from the preschool period through adolescence (Baumrind 1991). She found that authoritative parenting continued to be associated with positive outcomes for adolescents as with younger children and that responsive, firm parent–child relationships were especially important in the development of competence in sons. Moreover, authoritarian childrearing had more negative long-term outcomes for boys than for girls.

Maccoby and Martin (1983) extended the Baumrind typology and included a fourth type of parenting style which is characterized by neglect and lack of involvement. These are disengaged parents who are 'motivated to do whatever is necessary to minimize the costs in time and effort of interaction with the child' (Maccoby and Martin 1983). In infants such a lack of parental involvement is associated with disruptions in attachment; in older children it is associated with impulsivity, aggression, noncompliance, and low self-esteem.

The approach has been challenged on several fronts. First, questions remain concerning the processes that contribute to the relative effectiveness of these different styles. Second, it is unclear whether parenting styles are, in part, in response to the child's behavior. Placing the typology work in a transactional framework would argue that children with certain temperaments and/or behavioral characteristics would determine the nature of the parental style. A third concern is the universality of the typological scheme. Recent studies have raised serious questions about the generalizability of these styles across either socioeconomic status (SES) or ethnic/cultural groups. In lower SES families, parents are more likely to use an authoritarian as opposed to an authoritative style but this style is often an adaptation to the ecological conditions such as increased danger and threat that may characterize the lives of poor families (Furstenberg 1993). A second challenge to the presumed universal advantage of authoritative child-rearing styles comes from cross-ethnic studies. In Chinese families, authoritarian styles of child-rearing are more common and some (Chao 1994) have argued that the application of these stylistic categories to Chinese parents may be 'ethnocentric and misleading' (Chao 1994, p. 1111) since these child-rearing types represent an American perspective which emphasizes an individualistic view of childhood socialization and development. Contextual and cultural considerations need to be given more attention in typological approaches to child-rearing.

3.2 *The Parent–Child Interactional Approach to Socialization*

Research in this tradition is based on the assumption that face-to-face interaction with parents may provide the opportunity to learn, rehearse, and refine social skills that are common to successful social interaction

with other social partners. The style of the interaction between parent and child is linked to a variety of social outcomes including aggression, achievement, and moral development. Parents who are responsive, warm, and engaging are more likely to have children who are more socially competent. In contrast, parents who are hostile and controlling have children who experience more difficulty with age-mates. Moreover, these findings are evident in the preschool period, middle childhood, and adolescence (Parke and Buriel 1998). Although there is an overlap between mothers and fathers, fathers make a unique and independent contribution to their children's social development. Although father involvement is quantitatively less than mother involvement, fathers have an important impact on their offspring's development. Quality rather than quantity of parent–child interaction is the important predictor of cognitive and social development.

4. *Parental Instruction, Advice Giving, and Consultation*

Learning about relationships through interaction with parents can be viewed as an indirect pathway since the goal is often not explicitly to influence children's social relationships with extrafamilial partners such as peers. In contrast, parents may influence children's relationships directly in their role as a direct instructor, educator, or advisor. In this role, parents may explicitly set out to educate their children concerning appropriate ways of initiating and maintaining social relationships and learning social and moral rules.

The quality of advice that mothers provided their children prior to entry into an ongoing play dyad varied as a function of children's sociometric status (Russell and Finnie 1990). Mothers of well-accepted children were more specific and helpful in the quality of advice that they provided. In contrast, mothers of poorly accepted children provided relatively ineffective kinds of verbal guidance, such as 'have fun' or 'stay out of trouble.' The advice was too general to be of value to the children in their subsequent interactions.

As children develop, the forms of management shift from direct involvement or supervision of the ongoing activities of children and their peers to a less public form of management, involving advice or consultation concerning appropriate ways of handling social problems.

5. *Parents as Managers of Children's Opportunities*

Parents influence their children's social relationships not only through their direct interactions with their children, but also function as managers of their children's social lives and serve as regulators of opportunities for social contact with extrafamilial social partners. This parental role is of theoretical importance in light of the recent claims that parents' impact on children's development is limited and peer group level processes account for major socialization outcomes. In contrast to this view, we conceptualize the parental management of access to peers as a further pathway through which parents influence their children's development (Parke et al. 1994). From infancy through middle childhood, mothers are more likely to assume the managerial role than fathers. In infancy, this means setting boundaries for play, taking the child to the doctor, or arranging daycare. In middle childhood, it was found that mothers continue to assume more managerial responsibility (e.g., directing the child to have a bath, to eat a meal, or to put away toys).

6. *Beyond the Parent–Child Dyad: The Marital Subsystem as a Contributor to Children's Socialization*

Children's experiences in families extend beyond their interactions with parents. Evidence suggests that children's understanding of relationships is also shaped through their active participation in other family subsystems (e.g., child–sibling) as well as through exposure to the interactions of other dyadic subsystems (e.g., parent–parent).

6.1 *Influence of Marital Discord on Child Outcomes*

Considerable evidence indicates that marital functioning is related to children's short-term coping and long-term adjustment. Children exposed to marital discord are likely to have poorer quality of interpersonal relationships, including internalizing and externalizing behavior problems, and changes in cognitions, emotions, and physiology (see Fincham 1998). Two alternatives, but not mutually exclusive models, have been proposed to account for the impact of marital relations on children's developmental outcomes. One theoretical framework conceptualized marital discord as an indirect influence on children's adjustment that operated through its effect on family functioning and the quality of parenting. A second model focuses on the direct effects of witnessed marital conflict on children's outcomes rather than on the indirect effects. Both of these models have received empirical support (see *Marital Interaction: Effects on Child Development*).

7. *The Sibling System as a Contributor to Children's Socialization*

Siblings play a critical role in the socialization of children. Most children are likely to spend more time in direct interaction with siblings than parents and this

array of interactions between siblings have been found to be typified by greater emotional intensity than the behavioral exchanges that characterize other relationships. Sibling relationships contribute to children's socialization in a number of significant ways. Through their interactions with siblings, children develop specific interaction patterns and social understanding skills that generalize to relationships with other children. Relationships with siblings may also provide a context in which children can practice the skills and interaction styles that have been learned from parents or others. Older siblings function as tutors, managers, or supervisors for their younger siblings. Also paralleling the indirect influence that the observation of parent–parent interaction has on children, a second avenue of influence on children's development is their observation of parents interacting with siblings. These interactions may serve as an important context in which children deal with issues of differential treatment and learn about complex social emotions such as rivalry and jealousy (Dunn 1993).

8. *Determinants of Family Socialization Strategies*

One of the major advances in the field has been recognition of the importance of understanding the determinants of parenting behavior. Belsky (1984) proposed a three-domain model of the determinants of parenting, which included personal resources of the parents, characteristics of the child, and contextual sources of stress and support.

8.1 *Child Characteristics*

Child characteristics take two forms: universal predispositions that are shared by all children and individual differences in particular characteristics. Infants are biologically prepared for social, cognitive, and perceptual challenges and these prepared responses play a significant role in facilitating children's adaptation to their environment. Under the influence of recent advances in behavior genetics, there is increasing recognition of the role of individual differences in temperament on parenting behavior (Rothbart and Bates 1998). Although debates about the relative contributions of genetic and experiential factors to the emergence of individual differences in temperament continue, temperament clearly is a determinant of parental socialization tactics. For example, infants with difficult temperaments elicit more arousal and distress from caregivers than less difficult infants. The impact of these individual differences on parental socialization behavior are not independent of environmental conditions. Crockenberg (1981) showed that the impact of a difficult infant temperament on the parent–infant attachment relationship varied as function of the degree of social support available to the mother, which underscores the potential modifiability of temperament-based influences.

8.2 *Personal Resources*

A variety of studies support the prediction that personal resources—conceptualized as knowledge, ability, and motivation to be a responsible caregiver—alter parenting behaviors (Belsky 1984). Particularly striking are recent studies of how parental psychopathology such as depression will alter parenting behavior. When interacting with their infants, depressed mothers show flat affect and provide less contingent stimulation than nondepressed mothers; in turn, their infants showed less attentiveness, more fussiness, and lower activity levels. Differences are particularly evident when depression is protracted and not merely transient (Campbell et al. 1995).

8.3 *Social Support and Family Socialization Strategies*

Recognition of the role of the community and community agents as modifiers of family interaction is necessary for an adequate theory of socialization. Recognition of the embeddedness of families in a set of broader social systems such as community and culture is only a first step. The next task is to articulate the ways in which these other levels of social organization affect family functioning and explore the ways in which these influence processes take place. Community influence can be either positive or negative, since a high degree of connectedness with community resources is not necessarily positive. In addition, the relationship between communities and families is bidirectional and varies across development. Moreover, the influence of support systems on families may either be direct or indirect in its effects. Finally, both availability and utilization need to be separately considered. Families may have friends, relatives, and neighbors available, but fail to utilize these members of their informal social network in times of stress or crises or on a day-to-day basis.

8.4 *Socioeconomic Status as a Determinant of Family Socialization Strategies*

There is a long history of research concerning the links between socioeconomic status and/or social class and parenting beliefs and practices, in contrast to traditional assumptions that SES is a static state, SES is

a dynamic concept. Over the course of childhood and adolescence, families change social class and change is greatest in the youngest ages. Over 50 percent of American children change social class prior to entering school. In spite of the controversies surrounding this variable, there are SES differences in parental socialization practices and beliefs. Lower SES parents are more authoritarian and more punitive and controlling than higher SES families. Second, there are more SES differences on language measures than on nonverbal measure with higher SES mothers being more verbal than low SES mothers. Some SES differences are independent of race and poverty. In China, where there are relatively small differences in income across groups who vary in terms of education, less educated parents used more imperatives with their toddlers than better educated mothers (Tardif 1993). Similarly, studies of cognitive socialization found clear SES differences in African-American lower class and middle class families (Parke and Buriel 1998).

8.5 Ethnicity as a Determinant of Family Socialization Strategies

Recent cross-cultural and intracultural theories have emphasized the importance of socialization goals, values, and beliefs as organizing principles for understanding cultural variations (Harkness and Super 1995). In contrast to the older cultural deficit models of socialization, the more recent models emphasize how ecological demands shape values and goals. In the past, cultural deficit models were popular explanations for the socialization and child outcome differences observed between ethnic minorities and Euro-Americans. The focus on ethnic minority families has shifted away from majority–minority differences in developmental outcomes and more toward an understanding of the adaptive strategies ethnic minorities develop in response to both majority and minority cultural influences on their development. The parents' individual history of interaction with the larger sociocultural context, including their awareness of their ethnic group's history within the larger society, affect the manner in which they socialize their children. An important dimension of socialization in ethnic minority families is teaching children how to interact effectively in dual cultural contexts; the context of their ethnic group and the context of the larger Euro-American society. Harrison et al. (1990) have adopted an ecological orientation to explain the diverse environmental influences that contribute to the socialization of ethnic minority children. They conceptualize the socialization of ethnic minority children in terms of the interconnectedness between the status of ethnic minority families, adaptive strategies, socialization goals, and child outcomes. Emerging out of the adaptive strategies of adults are the socialization goals that they endeavor to inculcate in children to help them meet the ecological challenges they will face as ethnic minorities in a class- and race-conscious society. Ethnic pride and interdependence are two important socialization goals that enable ethnic minority children to function competently as members of both their minority culture and the larger society (see Parke and Buriel 1998).

Research needs to take into account the acculturation level of parents and children in recent immigrant families and the effects it has on family processes and child outcomes. Intergenerational differences in acculturation can create role strains between parents and children that have implications for childrearing styles, disciplinary practices, and overall parent–child relations. Together with acculturation, recognition of biculturalism as both an adaptation strategy and socialization goal is important. The effects of prejudice and discrimination on ethnic minorities, in such areas as social and emotional development, ethnic identity, and achievement motivation, deserve more attention. Language development research should also give greater attention to second language acquisition (usually English) and bilingualism and its relation to cognitive development and school achievement. More attention must also be given to the role of ethnic minority fathers, grandparents, and extended family members in the socialization of children (see also *Parenting in Ethnic Minority Families: United States*).

9. The Impact of Social Change on Family Socialization

Families are not static but dynamic and are continuously confronted by challenges, changes, and opportunities. A number of society-wide changes have produced a variety of shifts in the nature of family relationships. Fertility rates and family size have decreased, the percentage of women in the workforce has increased, the timing of onset of parenthood has shifted, divorce rates have risen, and the number of single parent families have increased (Hernandez 1993). These social trends provide an opportunity to explore how families adapt and change in response to these shifting circumstances and represent 'natural experiments' in family coping and adaptation. Moreover, they challenge our traditional assumptions that families can be studied at a single point in historical time since the historical contexts are constantly shifting. Our task is to establish how socialization processes operate similarly or differently under varying historical circumstances. In both the United States and other parts of the world a variety of changes, including the timing of parenthood, increases in women's employment, and increases in rates of divorce and remarriage, have taken place. These social changes have a major impact on children's socialization (Parke and Buriel 1998). To date, societal changes such as shifts in the timing of parenting, work participation,

or divorce have been treated relatively independently, but, in fact, these events co-occur rather than operate in any singular fashion. Multivariate designs which capture the simultaneous impact of multiple events on family socialization strategies are necessary.

10. Conclusion

Socialization is a multiply determined process; while our focus has been on families as central socialization agents, it is critical to recognize the roles of other social agents such as schools, peers, and the mass media in this process. A fuller understanding of socialization will come from more attention to the interplay among these diverse agents and how cultural and ethnic background influence socialization practices and outcomes.

See also: Attachment Theory: Psychological; Child Care and Child Development; Divorce and Children's Social Development; Identity in Childhood and Adolescence; Infant and Child Development, Theories of; Marital Interaction: Effects on Child Development; Parenting: Attitudes and Beliefs; Parenting in Ethnic Minority Families: United States; Self-development in Childhood; Sibling Relationships, Psychology of; Social Learning, Cognition, and Personality Development; Socialization in Adolescence; Socialization, Sociology of

Bibliography

Baumrind D 1973 The development of instrumental competence through socialization. In: Pick A D (ed.) *Minnesota Symposium on Child Psychology*. University of Minnesota Press, Minneapolis, MN, Vol. 7, pp. 3–46

Baumrind D 1991 Parenting styles and adolescent development. In: Brooks-Gunn J, Lerner R M, Peterson A C (eds.) *The Encyclopedia of Adolescence*. Garland, New York, pp. 746–58

Belsky J 1984 The determinants of parenting: A process model. *Child Development* **55**: 83–96

Campbell S B, Cohn J F, Meyers T 1995 Depression in first time mothers: Mother-infant interaction and depression chronicity. *Developmental Psychology* **31**: 349–57

Chao R K 1994 Beyond parental control and authoritarian parenting style: Understanding Chinese parenting through the cultural notion of training. *Child Development* **65**: 1111–19

Crockenberg S B 1981 Infant irritability, mother responsiveness, and social support influences on the security of infant–mother attachment. *Child Development* **52**: 857–65

Dunn J 1993 Young Children's Close Relationships: Beyond Attachment. Sage, Newbury Park, CA

Elder G H 1998 The life-course and human development. In: Damon W (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. I

Fincham F 1998 Child development and marital relations. *Child Development* **69**: 543–74

Furstenberg F F 1993 How families manage risk and opportunity in dangerous neighborhoods. In: Wilson W J (eds.) *Sociology and the Public Agenda*. Sage, Newbury Park, CA, pp. 231–58

Furstenberg F F, Harris K M 1993 When and why fathers matter: Impacts of father involvement on children of adolescent mothers. In: Lerman R I, Ooms T (eds.) *Young Unwed Fathers: Changing Roles and Emerging Policies*. Temple University Press, Philadelphia, PA, pp. 117–38

Harkness S, Super C M 1995 Culture and parenting. In: Bornstein M (ed.) *Handbook of Parenting*. Erlbaum, Mahwah, NJ, Vol. 2

Harrison A O, Wilson M N, Pine C J, Chan S Q, Buriel R 1990 Family ecologies of ethnic minority children. *Child Development* **61**: 347–62

Hernandez D J (ed.) 1993 *America's Children: Resources from Family, Government, and the Economy*. Russell Sage Foundation, New York

Lamb M E (ed.) 1987 *The Father's Role: Cross-Cultural Perspectives*. Erlbaum, Hillsdale, NJ

Larson R, Richards M H 1994 *Divergent Realities: The Emotional Lives of Mothers, Fathers and Adolescents*. Basic Books, New York

Maccoby E E, Martin J A 1983 Socialization in the context of the family: Parent–child interaction. In: Hetherington E M (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. 4

Parke R D 1996 *Fatherhood*. Cambridge, MA: Harvard University Press

Parke R D, Buriel R 1998 Socialization in the family: Ecological and ethnic perspectives. In: Damon W (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. 3, pp. 463–552

Parke R D, Burks V, Carson J, Neville B, Boyum L 1994 Family–peer relationships: A tripartite model. In: Parke R D, Kellam S (eds.) *Advances in Family Research, Vol. 4: Exploring Family Relationships with Other Social Contexts*. Erlbaum, Hillsdale, NJ, pp. 115–45

Rothbart M K, Bates J E 1998 Temperament. In: Damon W (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. 3, pp. 105–76

Russell A, Finnie V 1990 Preschool children's social status and maternal instructions to assist group entry. *Developmental Psychology* **26**(4): 603–11

Sameroff A J 1994 Developmental systems and family functioning. In: Parke R D, Kellam S G (eds.) *Exploring Family Relationships with Other Social Contexts*. Erlbaum, Hillsdale, NJ, pp. 199–214

Tardif T 1993 *Adult-To-Child Speech and Language Acquisition in Mandarin Chinese*. Yale University Press, New Haven, CT

R. D. Parke and R. Buriel

Socialization: Political

In its broadest sense political socialization refers to the content, processes, and dynamics of political learning across time and space. Related, more restricted terms include civic education and citizenship training. Because of the importance commonly attached to the consequences of early socialization, scholarship in this area has focused on the pre-adult years though not to the exclusion of later periods as well. Political socialization figures heavily in discussions of political

culture, the relation of the individual to the political system, and political system stability and change. This article covers the major foci of the field, methodological approaches, and the field's evolution and development since the 1950s.

1. Major Topics and Questions

Some scholars, political scientists in particular, are concerned about the implications of political socialization for how political institutions function. Other scholars focus more on what socialization means for individuals and how they cope with their political environment.

Scholars have tended to borrow from other disciplines rather than developing original perspectives. Applications of social learning and cognitive growth theories predominate. Learning about politics occurs as a result of intentional and direct efforts as well as unintentional and indirect efforts.

The range extends from broad topics such as national identity and consensus political norms to narrower ones such as partisanship and issue preferences. The broader topics are especially important in terms of how political culture is preserved, while the narrower ones help explain individual level behavior and attitudes.

The 'agents of socialization' is one of the most explored topics, with the role of the family, schools, and political events receiving heaviest scrutiny. Complex research designs are often employed here so that the individuals being socialized can be matched up with the agents providing the socialization cues.

There are intracountry variations. Political socialization may proceed in different ways and generate different outcomes according to individual and contextual characteristics. Most attention has been paid to the role of social class, gender, race, region, religion, and birth cohort.

On the other hand cross-country variations exist. Commonalties based on individual developmental processes and differences based on institutional arrangements have been observed. Some of these conclusions are based on systematic, multicountry investigations whereas others are derived from a comparison of independent, one-country studies.

Although much emphasis has been placed on transitions from early childhood up through young adulthood, the study of political socialization increasingly is being viewed from a lifelong perspective in order to comprehend the residues of earlier socialization as well as the impact of adult life stage transitions. For many nonconsensus political beliefs and practices the formative years appear to be young adulthood, after which beliefs and practices tend to solidify. For beliefs and values of a more integral and widely shared nature, the formative years occur earlier.

Of interest to a wide range of scholars, generational and historical effects may reflect either gradual changes in the social composition of succeeding cohorts or sharper discontinuities due to secular events. Rising education levels and affluence exemplify generational effects due to social composition. Wars, depressions, revolutions, and student protest movements are examples of more abrupt, event-driven effects.

2. Methodology

The majority of the literature rests on a foundation of empirical methodology. Attention has focused more on individuals than on institutions and systems as units of analysis.

There are various modes of data gathering. Surveys of probability and nonprobability samples compose the major mode. Three variants have been used. One consists of self-administered questionnaires of elementary and secondary school students. The use of such instruments for young children has been questioned. A second variant consists of personal interviews, either face to face or by telephone. Some of these surveys employ a structured format while others are more semi-structured. Investigators using either mode typically employ statistical analysis. Other modes of data collection include intensive interviews, sometimes combined with participant observation and projective techniques, with a small number of subjects (Coles 1987). Analysis in this instance is usually qualitative in nature. Although not typically classified under the heading of political socialization, analytic biographies of political figures—employing a variety of data sources—also constitute a mode of data gathering. Studies based on experiments are relatively rare.

Most investigations are cross-sectional, that is, the subjects were studied only once. However, a few overtime studies of the same individuals exist, thereby enabling more insight into developmental patterns. Some studies stretch from pre to full adulthood and involve either special populations or more general ones. Other longitudinal projects, especially those assessing the impact of educational institutions and political events, cover a shorter period of time.

Replicated surveys of the same populations provide historical trend data. Longitudinal studies of preadult populations are now available in a number of countries via national and international archives. Similarly, replicated surveys of adult birth cohorts are widely available. These surveys facilitate analyses devoted to assessing the probable impact of preadult experiences as cohorts age.

3. Initial Development and Ascendance of the Field

The phenomenon of political socialization has a long history. The preparation of individuals for their roles in the political world is as old as political life

itself. Both political authorities and ordinary citizens have been subject to the practices and outcomes of socialization regardless of political regime type. By contrast, the systematic study of political socialization has a relatively short history. Around the turn of the twentieth century educators in the United States conducted studies of school children that included choices of exemplars, national figures most admired by the children (Greenstein 1965). These 'surveys' and occasional other investigations in a variety of settings marked the first empirical studies of what would later be called political socialization.

It was not until the post World War II era that more sustained, disciplined research emerged. The term itself surfaced in the mid-1950s. By the early 1960s the label had achieved wide academic currency among political scientists in the United States and quickly spread to other western countries. Early work in the Unites States, the initial center of the empirical investigations, focused on pre-adolescents and emphasized the positive and relatively benign processes and outcomes of political socialization. Although more realistic perceptions and understandings set in with advancing age and cognitive development, the overall effect was viewed as salutary both for the successful functioning of the political system and the citizenry (Easton and Dennis 1969).

Following in the wake of these pioneering studies came a flood of projects in the United States and elsewhere. These inquiries tended to elaborate upon, modify, and in some instances to reject the conclusions of the earlier work. Variations according to social class, race, and location were identified. Political institutions conditioned the content, processes, and outcomes in fundamental ways. Considerable attention was devoted to the agents of socialization. Although the family appeared to be the key agent, it became clear that many other forces—including individuals themselves—also contribute to the formation and maintenance of political orientations. In general the more concrete, affect-laden, and reinforced the attribute in question, the more likely it was to be reproduced in the person being socialized and to persist over time.

4. *Waning Interest*

Scholarly inquiry in this area began to decline after the mid-1970s, especially in the United States. What had been a bull market turned bearish. Two factors appeared to be at work beyond the fading effect often associated with scholarly trends. One source lay in the dwindling novelty of studying school children. Political scientists in particular were accustomed to dealing with the 'real world' of politics, one occupied and ruled by adults. A second likely source rested in the appearance of youth movements throughout much of the western world in the late 1960s and early 1970s. This development not only shifted attention away from pre-adults but also appeared to undermine the proposition that the trusting orientations observed in many of the earlier studies would persist over time. Indeed, the question of persistence posed a serious threat to the field.

Ironically, even as fresh research efforts tailed off, the concept and the term itself became embedded in any number of subfields, including public opinion, electoral behavior, the politics of race and gender, political culture, and regime stability. Political socialization came to be assumed, rather than observed. Still, research continued on in a number of countries (Percheron 1993, Ichilov 1990) and longitudinal studies began to appear (Alwin et al. 1991, Jennings and Niemi 1981).

5. *Resurgence and Future Prospects*

After a prolonged lull, scholarly inquiry began to increase in the 1990s and has continued onward. Three developments helped lead to a resurgence of interest. One development was the crumbling of the Soviet Union and the appearance of transitional regimes and new democracies around the globe. These changes generated numerous surveys of adults that emphasize generational differences and studies of how and with what effect pre-adults are being socialized under the new regimes (e.g., Flanagan and Sherrod 1998). New, large-scale immigration patterns also presented opportunities for studies of adult resocialization. Thus, dramatic events in the political world have provided a natural laboratory for examining the processes and outcomes of political socialization. A second development springs from the decline in social capital and civic virtue among upcoming cohorts said to characterize most Western countries. In an effort to understand what is happening and to propose possible solutions, a variety of institutions and scholars have turned to the question of education and training of the young. National and cross-national assessments dealing with the character and impact of curriculum and school characteristics are being employed (Smith 1999). Much of this work is being carried out under the banner of civic education inquiries (Torney-Purta et al. in press).

A third development consists of renewed interest in the question of persistence and the dynamics of learning. Results from the long-term panel studies form part of this renewal as do the shorter-term studies of children and adolescents. The accumulation of replicated surveys as represented by such projects as national election studies, the General Social Surveys, and the Eurobarometer Surveys are also relevant because they permit the tracing of birth cohorts over extended periods of time. In doing so, they enable

informed interpretations drawing upon the conditions under which pre-adult socialization occurred and how these cohorts respond to later developments. These new developments augur well for continued scholarly inquiry. At the same time, the prospects remain bleak with respect to the study of young children, a major focus of the early research efforts.

See also: Citizen Participation; Citizenship: Political; Civic Culture; Educational Sociology; Generations: Political; Opinion Formation; Participation: Political; Political Communication; Political Culture; Political Discourse; Political Elites: Recruitment and Careers; Political Protest and Civil Disobedience; Socialization and Education: Theoretical Perspectives; Socialization, Sociology of

Bibliography

Alwin D F, Cohen R L, Newcomb T M 1991 *Political Attitudes Over the Life Span: the Bennington Women after Fifty Years*. University of Wisconsin Press, Madison, WI

Coles R 1987 *The Political Lives of Children*. Houghton Mifflin, Boston

Easton D, Dennis J 1969 *Children in the Political System*. McGraw Hill, New York

Flanagan C A, Sherrod L R (eds.) 1998 Political development: Youth growing up in a global community. *Journal of Social Issues* **54**: 447–627

Greenstein F I 1965 *Children and Politics*. Yale University Press, New Haven, CT

Ichilov O (ed.) 1990 *Political Socialization, Citizenship Education, and Democracy*. Columbia University, New York

Jennings M K, Niemi R G 1981 *Generations and Politics: A Panel Study of Young Adults and their Parents*. Princeton University Press, Princeton, NJ

Percheron A 1993 *La Socialization Politique* (*Political Socialization*). Armand Colin, Paris

Sears D O, Valentino N A 1997 Politics matters: Political events as catalysts for preadult socialization. *American Political Science Review* **91**: 45–65

Smith E S 1999 The effects of investments in the social capital of youth on political and civic behavior in young adulthood: A longitudinal analysis. *Political Psychology* **20**: 553–80

Torney-Purta J, Lehmann R, Oswald H, Schulz W in press *Citizenship and Education in Twenty-eight Countries: The Civic Knowledge and Engagement of Adolescents*. International Association for the Evaluation of Educational Achievement, Amsterdam

M. K. Jennings

Socialization, Sociology of

Socialization generally refers to the process of social influence through which a person acquires the culture or subculture of his or her group, and in the course of acquiring these cultural elements the individual's self and personality are shaped. Socialization, therefore, addresses two important problems of social life: the problem of societal continuity and the problem of human development. Because of its broad scope and importance, socialization is claimed as a major process by most social science disciplines. Different disciplines, however, have emphasized different aspects of this process. Anthropology tends to view socialization primarily as cultural transmission from one generation to the next. Anthropological interest in socialization coincides with the emergence of the culture and personality orientation in the late 1920s and 1930s. Psychologists are more likely to emphasize various aspects of the individual's development. For developmental psychologists, socialization is largely a matter of cognitive development that typically is viewed as a combination of social influences and maturation. For behavioral psychologists, socialization is synonymous with learning patterns of behavior, primarily through environmental reinforcement or through vicarious processes such as modeling. For clinical psychologists, socialization is viewed as the establishment of personality traits, usually in the context of early childhood experiences. The subfield of child development is associated most closely with the topic of socialization within psychology, where much of the focus is on child rearing (see *Socialization in Infancy and Childhood*; *Socialization in Adolescence*; *Socialization and Education: Theoretical Perspectives*).

1. Sociological Perspectives on Socialization

Within sociology there have been two major orientations to socialization and several minor ones. One of the oldest, associated primarily with structure-functionalism, considers socialization mainly as the learning of social roles. Parsons (1955), a major architect of this theoretical perspective, maintained that individuals become integrated members of society by learning and internalizing the relevant roles and statuses of the groups to which they belong. Since roles are important elements of social and cultural systems, comprising normative and value components, socialization into roles becomes a major means of integrating individuals into social systems and, thereby, enabling the continuity of these systems. This integration is viewed as functional for individuals and for society, although Parsons's interest was more in the latter than the former. Merton (1957), another key figure in American functionalism, also viewed socialization as the learning of social roles, but conceived of the process as more problematic. His concept of 'latent function' points to the possibility of unintended and undesirable consequences of normative and institutional behavior, such as the perpetuation of social inequalities or the development of individual

dysfunction. Functionalism was the dominant theoretical perspective in sociology in the 1950s and 1960s. Its view of socialization as the learning of social roles largely characterized sociology's contribution to 'socialization' in the previous edition of *The International Encyclopedia of the Social Sciences* (Brim 1968). But it has declined since then. Role learning is still an important focus of sociological inquiry into socialization, but it is no longer the most important focus, and no longer from a primarily functionalist perspective. The other, and more prevalent, sociological orientation views socialization mainly as self-concept formation. This view of socialization is associated closely with the symbolic interactionist perspective, a pragmatically based constructivist theory of self and society that places major emphasis on the construction of meanings through symbolic (mostly linguistic) interaction. In Mead's (1934) influential writings, the self is a reflexive phenomenon (i.e., both subject and object) that emerges out of symbolic interaction. Language enables the development of role taking whereby the individual is able to view himself or herself from the perspective of another. Role taking is the means by which culture (in the form of values, role identities, norms, etc.) becomes internalized as elements of the self-concept. This internalization, in turn, reproduces society. From the interactionist perspective, both self and society depend on the same process of social interaction through which realities are created, maintained, and negotiated.

For contemporary interactionists as well, socialization is distinguished from other types of learning and other forms of social influence by its relevance for self-conceptions. As such, socialization is not merely the process of learning rules or norms or behavior patterns. It involves learning these things, to be sure, but only to the extent that they become part of the way in which we think of ourselves, because the mark of successful socialization is the transformation of social control (via agents of socialization) into self-control (in the form of conscience and other self-regulatory mechanisms). This is accomplished largely through the development of identities, the various labels and characteristics attributed to the self, and guided by various self-evaluative processes (especially self-esteem, self-efficacy, and sense of authenticity). Commitment to identities, especially those based on important roles and values, is a source of motivation for individuals to act in accordance with the behavioral expectations implied by these identities (Gecas 1986, Stryker 1980). The focus on identity also emphasizes the membership component of socialization: to be socialized is to belong to some group.

Socialization as self-concept formation occurs through various processes of learning and social influence, such as, modeling, reinforcement, role-playing, social comparisons, direct instruction, and reflected appraisals. For sociologists, the most important of these is reflected appraisals. Based on Cooley's (1902) concept of the 'looking-glass self' and Mead's (1934) ideas on role taking, reflected appraisals refers to our perceptions of how others see and evaluate us. The reflected appraisals proposition suggests that we come to see ourselves as we think others see us, especially significant others. Much of the research on the development of self-esteem, as well as programs to enhance self-esteem, is based on this idea. An important variant of this proposition is labeling theory, which was the dominant sociological theory of deviance in the 1960s and 1970s. The theory maintained that a key causal factor in the etiology of deviance is the imposition of a (usually negative or stigmatized) identity or label on a person. Individuals may come to think of themselves in terms of the imposed label, even if there was no or little initial basis for it, because of the responses of others toward them (Schur 1971). It is a form of self-fulfilling prophecy, used to explain not only the development of deviant or criminal careers but also to address classroom dynamics regarding the consequences of teachers' differential expectations of students.

Reflected appraisals, in its several forms, are an important process for self-concept formation, but it has been subject to overstatement and uncritical application. Research has found only modest support, at best, for the proposition that our self-conceptions are a function of how we think others see us (Felson 1985). Some of this is due to the effects of situational contingencies (e.g., lack of specification of the nature of the relationship between self and other, the nature of the feedback given) and methodological problems (poor measures, cross-sectional research designs). But an even more important factor, increasingly recognized as such, is the agency and influence of the self in affecting its own socialization. The self does not passively receive input from the environment. Rather, it is actively engaged 'selecting, distorting, ignoring, influencing others' (Rosenberg 1979). Even very young children try to shape and manipulate their social environments by, for example, trying to influence their parents or resisting influence attempts from their parents. An important source of the self's agency are self-motives. By virtue of having a self, an individual is motivated to evaluate it favorably (self-esteem motive), perceive it as efficacious (self-efficacy motive), and experience it as meaningful or real (authenticity motive). It is uncertain how much of the self's manipulations, distortions, and attempts to control its environment are due to these self-motives, but probably quite a bit, and more so for some people than for others. They are a source of resistance and distortion in the reflected appraisals process. But they also have broader implications for socialization. Within socialization contexts (as well as other settings), individuals will try to maximize their self-esteem, self-efficacy, and sense of authenticity. To the extent that agents of socialization can tap into one or more of these self-motives and utilize them toward socializa-

tion goals, the socialization effort will be more effective. Conversely, these same self-motives can serve as major sources of resistance to socialization pressures (Gecas 1986). In other words, the most effective socialization is self-activated, in which the individual is a willing accomplice in his or her own socialization.

The picture of socialization that emerges from the symbolic interactionist perspective is of a highly reciprocal process between agents and subjects of socialization, in which self-conceptions are affected by reflected appraisals and other forms of social influence, but that selves exert an influence in their own right on the process and its outcomes. Society and culture do provide the stage and a general script for the novice, but typically there is plenty of room for interpretation and creativity.

2. *Research on Socialization*

Much of the sociological research on socialization, past and present, is context specific, that is, it studies socialization processes within specific contexts of interaction, such as family, peers, workplace, day care, boot camp, etc. (Gecas 1981). Some of the studies guided by symbolic interaction theory use ethnographic or field methods to record in rich detail the patterns of interaction, negotiation of meanings, the normative constraints of social settings, and their socializing consequences for the participants. Becker's (1961) study of the processes by which medical students become doctors, Fine's (1987) field studies of the development of gender identities among preadolescent boys on a little league team, and Corsaro's (1997) observations on the emergence of peer group subcultures among preschoolers in day care, are excellent examples of this genre. While these studies provide rich and insightful observations of socialization processes as they occur in their natural settings, they are relatively rare because of the time and effort involved in this methodology. Most of the research on socialization, by symbolic interactionists and other sociologists, utilizes survey methods (e.g., interviews, questionnaires). This research tends to be quantitative and narrower in focus, for example, focusing on the effects of a socializing agent's behavior on some aspect of the socializee's development.

2.1 *The Family Context*

Even though socialization is not limited to experiences in childhood but occurs throughout one's life, most of the research attention has been on children (although less so in sociology than psychology). Therefore, the most important context of socialization is the family, where the initial or primary socialization of children takes place. Studies of the consequences of parent–child interaction for the development of children are numerous, particularly along the dimensions of parental support and control. Research indicates that parents are most effective as agents of socialization when they express high levels of support or nurturance combined with the use of inductive control. Under these conditions, children are most likely to identify with parents, internalize parental values and expectations, use parents as role-models, be receptive to parents' influence attempts, and develop positive self-conceptions and a strong conscience. Conversely, low parental support, combined with reliance on coercive control, is associated with unfavorable socialization outcomes (Peterson and Rollins 1987). These associations typically are due, to some extent, to reciprocal influences: not only are these child outcomes affected by these parental behaviors, but the characteristics of the child (e.g., compliance vs. rebellion, high self-esteem vs. low self-esteem) may elicit differential parental behavior. After all, children do effect parents as well as the reverse.

Besides parental support and control, there are many other important variables in the family setting affecting socialization processes and outcomes, such as political and religious values expressed by parents, performance expectations, family structure regarding number of parents and children, birth order and sibling interaction, and personality characteristics of parent and child. Sociological interest in the effects of family structure, especially single-parent vs. two-parent families, on child socialization has been increasing, as the prevalence of single-parent families continues to increase in America and other western countries. This is one of the hotly contested and political issues in the family socialization area. While the research evidence seems to suggest that child socialization in single-parent families is disadvantageous for children, compared to socialization in two-parent families (McLanahan 1999), there are numerous factors that affect this relationship (such as economic level, quality of husband–wife relationship), leaving room for ambiguity and need for further research.

2.2 *The School Context*

The school as a context of child socialization has generated almost as much public concern as has the family and almost as much research. Like the family, the school is a social institution directly involved in socializing children, although its mandate is more narrowly defined as providing formal education and developing children's knowledge and skills. In this sense, schools are less involved in primary socialization (i.e., the development of basic values, motivations, and conceptions of self) and more involved in sec ondary socialization (i.e., the development of knowledge and skills). This is not a very precise distinction, however, since a good deal of primary socialization occurs in schools as well, involving values, motivations, and self-concepts. Many school activities have

implications for the child's self-concept. One of the most important involves evaluation of the student's academic performance, and this evaluation is more public than evaluations by parents or peers. Success in academic performance is good for self-esteem and self-efficacy. But failure is not, and public failure is worse. Besides its implications for self-esteem, performance evaluations result in the categorization or labeling of students, by teachers as well as classmates, as 'smart,' 'dumb,' 'lazy,' etc. Such labels affect the way others respond to the person and, thereby, shape the person's self-concept. But students, like other socializees, are not passive recipients of the pressures they experience. Some resist negative labels, some try to change them by working harder, and some adopt failure-avoiding strategies (such as withdrawal or procrastination) which may enable maintenance of self-esteem but at the expense of academic performance (Covington and Beery 1976).

American public schools, concerned with this problem, began to develop strategies to enhance children's self-esteem. By the mid-1980s, self-esteem enhancement became part of the curriculum in many public (especially elementary) schools. Grading and public evaluations of students were discouraged, negative feedback of any kind was also discouraged, positive feedback and affirmations of self-worth were encouraged. This emphasis on self-esteem in the school curriculum had its own set of undesirable socialization consequences (Hewitt 1998) and began to decline by the 1990s.

2.3 Peer Groups

The third most important context for the socialization of children is the peer group. In terms of structure and function, the peer group is very different from family and school as a context of socialization. Peer groups are voluntary associations of status equals and are based on friendship bonds. Children's peer groups typically are segregated by sex and differ in organizational patterns: girls' peer groups tend to be closely knit and egalitarian friendships, whereas boys' peer groups tend to be loosely-knit, larger groups, with status hierarchies (Thorne 1990). An important socialization consequence of intensive association with same-sex peers and involvement in sex-typed activities is that it strongly reinforces identification and sense of belonging with members of the same sex. It also contributes to the development of stereotypical attitudes concerning members of the opposite sex. The development of gender identities and much of sexual socialization during childhood occurs in the context of peer rather than family associations.

Peer interactions have a wide range of consequences for the child's developing sense of self. In his field studies of pre-adolescent boys, Fine (1987) observed that friendships are especially appropriate for the mastering of self-presentation and impression-management skills, since inept displays usually will be ignored or corrected without severe loss of face. Friendships characterized by mutual trust and tolerance are important for exploring the boundaries of the self. A wider latitude of behavior is allowed, and even encouraged, since friends, unlike parents, are not responsible for changing or shaping the child's behavior. Mutual understanding, role taking, and empathy—important to the development of self in relation to others—develops best in the context of friendship bonds.

Peers provide an alternate reference group for children, as well as an alternate source of self-esteem and identity, and they provide a context for the exercise of independence from adult control. As such, the peer group has been described as a socialization context in which subcultures of opposition or resistance to the dominant culture develop, such as the adolescent subcultures described by Coleman (1961) and in much of the literature on juvenile delinquency. The existence of adolescent subcultures, with values, norms, and tastes distinct from and often in opposition to the mainstream culture, has tended to polarize parents and peers as agents of socialization and make them competitors for the youth's attention and loyalty (see *Youth Culture, Sociology of*).

2.4 Adult Socialization

The socialization experienced by adults generally falls into the category of secondary socialization, building on the socialization experiences of childhood. Much of this is role specific (Brim 1968), that is, learning the knowledge and skills required for the performance of specific adult roles (e.g., occupation, marriage, parenthood). As individuals become committed to the roles they play, they come to identify themselves and think of themselves in terms of these role identities (Stryker 1980).

Since work is a dominant activity and context for most men and women in modern societies, much of adult socialization involves preparation for an occupation, which usually takes place in specialized schools or training programs (e.g., medical school, college). The work setting itself can have a substantial socializing effect on workers, affecting more than just their knowledge and skills. Much of the socialization that occurs in work settings is incidental to the main activities on the job and often unintentional, involving the workers adaptations to the conditions of work. These work conditions vary considerably with regard to structural features, power relations, presence of co-workers, degree of complexity and autonomy, etc., and are consequential for many aspects of self and personality. Kohn and Schooler (1983) have found certain work conditions (e.g., complexity, autonomy, routinization) affect the development of workers' values and personality. Kanter (1977) found that the nature of work relations, especially the structure of

opportunity on the job, affects the workers' attitudes and behavior. Workers' adaptations to their work conditions do not lead necessarily to commitment to the job or self-investment in the occupational role-identity. On the contrary, a prevalent theme in the sociological literature on work, especially from a Marxist perspective, deals with the alienating consequences of work in capitalist societies (Hochschild 1983).

2.5 Resocialization Contexts

Many other contexts have socializing consequences for adults: family, church, recreation settings, and various voluntary organizations. The socialization that takes place in these contexts typically is a continuation and expansion of past socialization experiences. Resocialization, however, refers to socialization experiences constituting a more radical change in the person. Resocialization contexts (such as mental hospitals, reform schools, political indoctrination camps, religious conversion settings, some therapy groups) have as their explicit goal the transformation of the individual, because the individual is perceived by self and/or others to be fundamentally dysfunctional, misguided, sinful, or deficient. The key task in resocialization is the replacement of the person's previous set of values, beliefs, and self-conceptions with a new set based on a new ideology or worldview, that is, the 'death' of the old self and the birth of a new self. Typically, this is accomplished through intense small group interaction and pressure to change, in which the physical and symbolic environments are highly controlled by the agents of socialization. The actual techniques used to bring about change may vary considerably in resocialization contexts depending, for example, on whether the socializee is there voluntarily or not (Ofshe 1992). Resocialization contexts have received more attention from the popular press than they have from social scientists. They deserve much more research attention, since they constitute an important and more drastic type of adult socialization. Also, the prevalence of resocialization contexts in the form of religious and therapeutic groups seems to be increasing in modern societies.

3. Current Trends and Future Directions

Interest in socialization tends to increase in troubled times, when social institutions (especially family systems and public schools) are perceived as breaking down or working badly, or when the rate of social change creates stress, difficulties in self-maintenance, and anxiety about the future. These are such times in most modern societies. Consequently, current academic and public interest in socialization is high. It is also marked by controversy. Much of the controversy centers on family issues and family trends: increasing divorce rates, number of single-parent families, births to unmarried women, use of day-care centers, dual-career families, etc. It is the consequences of these trends for child socialization and wellbeing that is at issue. Current opinion (academic and public) is divided strongly on the implications of these changes for child socialization. More research is needed and will be forthcoming, since funding agencies are very interested in these issues as well.

Socialization theory and research typically lags behind societal changes and developments. It is slow in acknowledging and addressing social change. What effect will the rapidly evolving computer technology have on the socialization of children and adults? The ubiquity of computers in homes, classrooms, workplaces, and recreation centers is quite obvious in modern societies. Children are becoming involved with this technology at very early ages—even preschoolers are becoming computer-literate. Unlike television, computers are an interactive technology, more likely to engage and develop the user's cognitive and motor skills. As such, there is reason to hope that the influence of computers on children's socialization will be beneficial, as some scholars have suggested (Turkle 1995). Computers and information technologies are transforming modern societies. Surely they are also affecting processes and outcomes of socialization. But just how, when, and what these processes and outcomes might be remains to be studied.

Socialization theory in sociology has long had a primarily cognitive emphasis, especially evident in symbolic interactionism. The interactionist conception of the self as a reflexive process, that develops out of role taking (a cognitive process) and language, left little room for emotions. This started to change in the mid-1970s. Sociologists and social psychologists interested in socialization rediscovered emotions and have been trying to include them in theorizing about socialization. Special attention is given to self-relevant or reflexive emotions, such as shame and guilt, and to their importance in the development of the moral self (Shott 1979). Research on emotions has lagged far behind these theoretical developments, but is increasing and will continue to increase in the sociology of socialization.

A perennial issue regarding human development is the relative influence of biological factors vs. social factors (i.e., nature vs. nurture) in the development of mind, self, and personality. Sociology has long been squarely on the 'nurture' side of the issue, resisting any claims of biological determinism of any part of the socialization domain. There are signs that a more conciliatory stance is developing between sociology and biology. One reason is sociology's renewed interest in emotions, which increases the relevance of physiological factors. Another reason is the dramatic advances in biology (e.g., genetics, cell biology, physiology of the brain). These have potential implications

for a wide range of socialization concerns. How do biological factors interact with environmental influences to produce socialization outcomes? Research at the intersection of biology and sociology could be very important for the development of socialization theory.

See also: Adult Education and Training: Cognitive Aspects; Child Care and Child Development; Childhood: Anthropological Aspects; Coleman, James Samuel (1926–95); Erikson, Erik Homburger (1902–94); Family as Institution; Family Processes; Freud, Sigmund (1856–1939); Identity: Social; Integration: Social; Mead, George Herbert (1863–1931); Self-regulation in Adulthood; Self-regulation in Childhood; Socialization: Political; Status and Role, Social Psychology of

Bibliography

Becker H S 1961 *Boys in White: Student Culture in Medical School*. University of Chicago Press, Chicago

Brim Jr. O G 1968 Adult socialization. In: Sills D L (eds.) *International Encyclopedia of the Social Sciences*. Macmillan, New York

Coleman J S 1961 *The Adolescent Society*. Free Press of Glencoe, New York

Cooley C H 1902 *Human Nature and the Social Order*. Scribner's Sons, New York

Corsaro W A 1997 *The Sociology of Childhood*. Pine Forge Press, Thousand Oaks, CA

Corsaro W A, Eder D 1990 Children's peer cultures. *Annual Review of Sociology* **16**: 197–220

Covington M V, Beery R G 1976 *Self-worth and School Learning*. Holt, Rinehart and Winston, New York

Felson R B 1985 Reflected appraisals and the development of the self. *Social Psychology Quarterly* **48**: 71–8

Fine G A 1987 *With the Boys: Little League Baseball and Preadolescent Culture*. University of Chicago Press, Chicago

Gecas V 1981 Contexts of socialization. In: Rosenberg M, Turner R H (eds.) *Social Psychology: Sociological Perspectives*. Basic Books, New York

Gecas V 1986 The motivational significance of self-concept for socialization theory. In: Lawler E J (eds.) *Advances in Group Process*. JAI Press, Greenwich, CT, Vol. 3

Hewitt J P 1998 *The Myth of Self-esteem*. St. Martin's Press, New York

Hochschild A R 1983 *The Managed Heart: Commercialization of Human Feeling*. University of California Press, Berkeley, CA

Kanter R M 1977 *Men and Women of the Corporation*. Basic Books, New York

Kohn M L, Schooler C 1983 *Work and Personality: An Inquiry into the Impact of Social Stratification*. Ablex, Norwood, NJ

McLanahan S S 1999 Father absence and the welfare of children. In: Hetherington E M (eds.) *Coping with Divorce, Single Parenting, and Remarriage*. Erlbaum, Mahwah, NJ

Mead G H 1934 *Mind, Self, and Society*. University of Chicago Press, Chicago

Merton R K 1957 *Social Theory and Social Structure*. Free Press, Glencoe, IL

Ofshe R 1992 Coercive persuasion and attitude change. In: Borgatta E F, Borgatta M L (eds.) *Encyclopedia of Sociology*. Macmillan, New York

Parsons T 1955 Family structure and the socialization of the child. In: Parsons T, Bales R F (eds.) *Family, Socialization, and Interaction Process*. Free Press, Glencoe, IL

Peterson G W, Rollins B C 1987 Parent-child socialization: A review of research and applications of symbolic interaction concepts. In: Sussman M B, Steinmetz S K (eds.) *Handbook of Marriage and the Family*. Plenum Press, New York

Rosenberg M 1979 *Conceiving the Self*. Basic Books, New York

Schur E M 1971 *Labeling Deviant Behavior*. Harper & Row, New York

Shott S 1979 Emotions and social life: A symbolic interactionist analysis. *American Journal of Sociology*. **84**: 1317–34

Stryker S 1980 *Symbolic Interactionism: A Social Structural Version*. Benjamin/Cummings, Menlo Park, CA

Thorne B 1990 Children and gender: Construction of difference. In: Rhode D L (eds.) *Theoretical Perspectives on Sex Differences*. Yale University Press, New Haven, CT

Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon and Shuster, New York

V. Gecas

Societies, Types of

While the concept of society is to be found in most sociological writings, it remains ambiguous and relatively ill-defined. Like most of the scientific concepts that are also used in common speech, that of society seems to need no introduction and to reflect reality in a rather straightforward, transparent way. Yet although it is one of the commonest, this notion is also one of the most ambiguous; the question has recently been raised whether it should not be considered obsolete or even harmful (Ingold 1990). It is not the purpose here to go that far; nevertheless, its ambiguities cannot be ignored, and it can be argued that the concept of society has probably been more a means of constructing reality than a representation of it. This is also the case with typologies that have been constructed around it, and that can be considered as useful only within quite well-defined boundaries.

1. From Society to Societies

One of the first ambiguities of the concept of society lies in the fact that it is used, with quite different meanings, in the plural as well as in the singular. In sociology, the use of the singular refers to social reality in a general way, which can be both abstract and empirical. In social anthropology, on the other hand, it is more commonly used in the plural, and it is very widely used. Like most anthropological concepts, it is

rather loosely defined, and is often presented almost as a material entity. In other words, it is used as if it were an observable fact, as if it had empirical status: a society is something that exists and can be observed; in other words, it is a clearly observable object. Consequently, social anthropologists have used the concept of society indiscriminately, and have transformed any little village into a society. The theoretical, and indeed the political, consequences of such a use are important, as shall be seen below. At this stage, it needs only to be noted that the omnipresence of the concept of society, in sociology as well as in social anthropology, does not prevent its being a rather blurred category, sometimes meaningless or at least fluid. A quick glance at different textbooks and dictionaries suffices to indicate that, although it is referred to on almost every page and seems pivotal, it does not always merit its own entry. It might, at first glance, seem useful, as there is no term more appropriate for referring to a well-knit social group differing from the more political and neighboring concepts of state, nation, or ethnic group. But this advantage soon turns out to be a handicap as one does not really know to what it precisely refers; yet it subsumes the social classes and oppositions that characterize any social reality.

It is undoubtedly legitimate, and probably inevitable, to classify social realities into types. Yet one must avoid the temptation of considering societies as well-defined objects. On the contrary, it is not the concept of society that entails comparison, but comparison that entails the concept of society; in other words, the need for typologies creates the elements to be classified (Amselle 1999, p. 21). A typology can only be provisional and of limited use. That is surely the reason why no classification can be considered as definitive. That is particularly true of the perpetually recurrent dichotomies such as those that oppose 'modern' and 'traditional' societies, which are inevitable sources of confusion. A 'type' is always the outcome of a process of classification, and can never be taken as revealing an 'essence' (Boudon 1998, p. 104).

2. *A Dichotomic Outlook*

If humans are naturally social beings, predisposed to social life, the task of sociology is to understand the mechanisms of life in society. Nevertheless, sociologists very soon stressed the limitations of a purely abstract conception of society, and attempted to distinguish between various fundamental types of society. To some extent, it could be said that the very idea of society automatically opens up into the discussion of its diversity. Furthermore, society is hardly conceivable aside of its specific instantiations. One of the most influential definitions was coined by the German sociologist Ferdinand Tönnies (1855–1936). It is interesting to point out that his definition includes a comparative perspective, since, according to him, the concept of society (*Gesellschaft*) can only be understood in reference to the concept of community (*Gemeinschaft*): community is characterized by the proximity, both affective and spatial, between individuals. Within the community, the 'we' prevails over the 'I'; the individual does not exist outside the collective. Society, on the other hand, characterizes the more modern forms of sociability and is founded on individual interests. These latter ground social relations, which are oriented toward the market and the contract. One finds traces of German Romanticism in Tönnies' conception of society, which can be considered as a theory of *Heimat*.

Tönnies' opposition between *Gesellschaft* and *Gemeinschaft* was to be taken up by Durkheim, according to whom two types of solidarities can be distinguished: mechanical solidarity and organic solidarity. The first expresses the similarity and the community of feelings that unite the members of a group. The second, on the contrary, consists of interdependence, in a division of labor, as each member contributes to the whole according to his or her capabilities. Although it has sometimes remained unnoticed, the place of the individual is central to Durkheim's conception of solidarity: according to him, the crucial element in mechanical solidarity is that individuals do not exist as such, they are completely subsumed in the collectivity (Durkheim 1978, pp. 100, 170). Organic solidarity, on the other hand, presupposes the existence of the individual; here, all individuals have a 'sphere of action' and their own personality. The individuals here are necessarily different from one another (p. 101). Durkheim also considers that mechanical solidarity is more present in 'inferior societies' and, more generally, he related the type of solidarity to the degree of division of labor: the more rudimentary the division of labor, the more mechanical the society and, in such circumstances, the individuals of a group look very much alike. Their character lacks any kind of originality; there is no psychological individuality. Conversely, among civilized people, the division of labor is at its peak, and individuals are very different from one another.

Organic solidarity replaces, little by little, mechanical solidarity, which is more present in clan-based societies, which are made up of similar entities. As society becomes more complex, solidarity becomes organic (Durkheim 1978, p. 159). Certainly, traces of mechanical solidarity remain here and there, but there is a global process of substitution of the one type for the other. In *La Division du travail social*, Durkheim seems to realize that he has validated a liberal conception of social life in modern societies, and he therefore nuances his thought by reasserting that altruism is essential to any kind of social life: in order to be able to live together, people must make mutual sacrifices (Durkheim 1978, pp. 207–9). He then suggests that mechanical solidarity never completely disap-

pears. Later on, he asks for more social justice and more equality (p. 311) and he reminds us that society cannot be conceived without some solidarity.

Durkheim did not really draw any conclusion from these last remarks that might have led him to reduce the gap between 'inferior' and 'civilized' societies. If he sees the necessity of some collective feeling in industrial societies, he admits no doubt of the idea that the individual is basically absent from mechanical solidarity. Yet one of the main difficulties of the opposition between *Gemeinschaft* and *Gesellschaft*, and indeed of other dichotomic approaches, lies in the 'annihilation' of individuals, and in the equally problematic sense of communality that is meant to be typical of preindustrial societies. Yet such a radical opposition between 'us' and 'them' is by no means limited to these classical examples. It is also to be found in more contemporary writers such as Lévi-Strauss, who opposes 'cold' and 'hot' societies on the grounds of their degree of 'historicity' (Lévi-Strauss 1973, p. 40). The French anthropologist was wise enough to recognize that such a distinction is purely theoretical and does not correspond to any concrete entity; however, one can wonder whether his warning is not purely rhetorical, as he (and others) continued to use the same kind of dichotomous categories based on the relative impact of history and change.

It is no exaggeration to say that such views have been very influential and that they have directed the way we conceptualize preindustrial societies. It is, then, important to point out that those hunting and gatherering groups that have been taken as living examples of preneolithic societies do not live in autarky, and have been in contact with more 'advanced' groups for a long time; while others in this category have in fact adopted hunting quite recently. Furthermore, as has been suggested by Lévi-Strauss himself, the passage from the type to the reality remains delicate, and it is often difficult to decide which group belongs to one or the other pole.

Those typologies often seem to derive from theoretical prejudices, such as the deeply grounded ideas either that some peoples are resistant to change or even of the absence of the individual in some groups. This latter idea was also at the basis of the opposition between holism and individualism, which, in Dumont's theory, amounts to two types of societies opposing each other. Whereas Dumont sometimes considers holism and individualism as 'ideologies,' he does not take this specification into account when he argues that they also represent two 'types of societies': 'Where the individual is the supreme value, I speak of individualism; in the opposite case, where the value lies in the society as a whole, I speak of holism' (Dumont 1983, p. 37). Elsewhere, he assimilates this dichotomy to the opposition between 'traditional' and 'modern' societies (Dumont 1966, p. 23). In his view, India then becomes the archetype of traditional societies, which might look somewhat paradoxical.

If the distinction between individualism and holism is debatable in itself, it becomes even more contestable as soon as one seeks for some empirical data to exemplify it. Indian studies, for instance, are increasingly ill at ease with the idea that the individual is absent from the Indian social system (see, e.g., Mines 1994). To consider the 'renouncer' (the ascetic) as the sole agent of social change in India (Dumont 1983) amounts to reducing India to a purely religious order in which there is no room for the military, the merchant, the king, or even the householder: in other words, a world without politics or economics.

More generally, the same criticism applies to any dichotomous representation of the world. The main *raison d'être* of typologies of this kind lies precisely in their deficiency: they soon appear as alternatives, radical and exclusive oppositions that reveal an 'essence.' They are also linked to a kind of sociological illusion that consists of dissolving individuals into a system by denying them any kind of action. They also transform the world into blocs that are said to be homogeneous (e.g., East and West, North and South). What is the relevance of this kind of dichotomy, which seems to impoverish our analyses? Recent changes in the contemporary world make this kind of view even more hazardous: globalization, migration, and the new technologies make any search for essence even more dubious. Furthermore, modern societies are far from having given up any kind of 'holism,' as is shown by some modern movements of ethnic renaissance or religious fundamentalism, or indeed in US multiculturalism, all of which promote the need for collective recognition at the expense of individual differences. Conversely, the absence of a division of labor does not entail any psychological conformity among individuals, and no structural constraint has ever succeeded in inhibiting every kind of individuality. In short, the classical distinction between 'them' and 'us' is always much less radical than is suggested by some classifications.

3. *Economic Categories*

Beyond this rather Manichean use of typologies, one can find some more elaborate attempts that base themselves on a variety of criteria in order to compare economic, kinship, or religious systems. Their main interest is allowing the possibility of comparison. However, these systems also fail to avoid the trap of reducing social realities, complex by nature, to a single criterion. One sees categories such as 'matrilineal societies' or 'polyandric societies,' which seem to imply that a specific feature can qualify a complex whole. One could argue that these classes can only be useful within very precise limits, in this case a comparison between 'marriage systems'; however, an examination of their use in practice shows that they have inevitably been used in a much broader sense, even though the

selected criterion had little relevance to the questions at issue. That is what Leach deplored : 'Comparison is a matter of butterfly collecting—of classification, of the arrangement of things according to their types and subtypes' (Leach 1968, p. 2). The comparative anthropologist, Leach complains, is a mere butterfly collector, and the types and subtypes proposed in these systems are so problematic that they lead to nothing. This is particularly the case with categories based on unilineal descent, which might be without any sociological significance: 'We must be prepared to consider the possibility that these type of categories have no sociological significance whatsoever. It may be that to create a class labelled matrilineal societies is as irrelevant for our understanding of social structure as a class of blue butterflies is irrelevant for the understanding of the anatomical structure of lepidoptera' (Leach 1968, p. 4).

The classification of societies according to economic criteria is one of the commonest forms, perhaps because it gives the impression of being based on objective elements, in a materialistic perspective that has always been influential in the social sciences. Lewis Henry Morgan was among the first to take technical inventions as the foundation for his evolutionary stages. According to him, the first two stages of humankind—savagery and barbarism—were subdivided into inferior, middle, and superior degrees, whereas the third stage—civilization—started with the invention of writing. The first degree of savagery is characterized by fruit-gathering, and the middle degree by fishing, whereas hunting only appears at the superior level. Barbarism starts with pottery, and proceeds to the domestication of animals, and ends with metallurgy. There is no need to dwell upon the deficiencies of such a model, which does not even take agriculture as constituting a significant stage. In an evolutionary perspective, comparison was above all diachronic, and aimed at stressing the sense of history. What is being compared is not 'social systems' but different stages in the development of humankind. In some ways, it is more a philosophy of history than a typology that is at stake here.

Morgan had a deep influence on Marx and Engels who also considered humankind in an evolutionary perspective: social wholes are defined from the relations of production between the various social classes, at a given level of technology. One calls a 'mode of production' a special arrangement of those elements at a specific moment of history. In a strict sense, the mode of production denotes a few specific types of relations of production. In the years 1960–70, however, the concept took on a more elaborated meaning, especially among those scholars who followed Althusser in an attempt to combine Marxism and structuralism. The mode of production was then conceived as a model or structure, that encompassed the totality of the social relations. It consisted of a special arrangement within the 'economic infrastructure' (the 'base') that determined the superstructure (politics, law, and ideology). One could say that the superstructure necessarily reflects the infrastructure; in other words, the economic base determines all social and jural relations. The mode of production became a key concept of Marxist theory, and numerous efforts were mobilized toward the delineation of the most characteristic modes of production: besides the classical Marxian major types (Asiatic, feudal, and capitalist modes of production), Marxists invented new types, such as the domestic mode of production or the lineage mode of production.

The weakening of Marxism and structuralism put an end to this line of research, which had not brought any convincing results. One of its major shortcomings was to consider social realities as systems that are closed and self-sufficient. It was quite legitimate to stress the importance of economic realities, but the overall principle of determination was somewhat simplistic, and apart from a few remarks about the 'relative autonomy' of the superstructure, it was unable to account for the influence of ideas and values in social formations. The very existence of the socialist 'democracies,' with their Orwellian omnipresence of the state, could serve to illustrate the limitations of a purely economic determinism. With the decline of Marxism in the 1980s, the mode of production became a rather obsolete concept.

Nevertheless, the type of production remained one of the major criteria used to differentiate societies. Numerous anthropological studies attempted to characterize societies on the base of hunting, pastoralism, and agriculture. Hunting societies have always been a major source of fascination among anthropologists, partly because they can be seen as survivors of the Palaeolithic era. The first anthropologists took hunting societies as living fossils surviving from prehistoric times. Such a position is hardly acceptable today since we know that most contemporary hunting societies have been in contact with more 'advanced' groups for a long time while others have only taken up to hunting quite recently. Furthermore, few peoples have lived exclusively on hunting. If the contemporary hunter can no longer be considered as typical of the first stages of humankind, hunting remains a special case of the relationship between humans and nature.

In *Femmes, greniers et capitaux*, Meillassoux (1974), a French anthropologist, showed that hunting and agriculture entailed different types of social relations. According to him, there was no rupture between production and reproduction: thus kinship relations are also determined by productive forces. One can distinguish between two economic types according to whether land is an 'object' or a 'means' of labor. Hunting groups use land as an object of labor: it is used directly, without any preliminary investment of human energy. Therefore one can speak of an economy of *ponction* (withdrawal of capital). Here, the

cycle of reproduction of the labor force (human energy) is short: there is no possible accumulation of wealth, and what is being produced must be consumed. Social relations are affected by such an organization: the producing groups are precarious, and prevent the appearance of any form of authority. The membership of a horde is constantly changing, and people move freely from one to another. This free mobility of adults between the groups is, according to Meillassoux, the main mechanism of social reproduction. According to him, the horde is not made up of kinship relations but of 'adhesion'. In other words, the social status of individuals is not fixed at birth through descent rules, but depends on the voluntary choice of the individual. These productive constraints determine a special type of social relationship.

However, it is difficult to follow Meillassoux when he argues that kinship ties are reduced to a minimum in the hunting band: we have no example to support his claim that children are easily given up at an early age, that marriages are short-lived, or even that weddings and funeral ceremonies are less elaborate than the rituals welcoming new members. The classical example of the !Kung goes against Meillassoux's assertions. His mistake was to consider that relations of adhesion necessarily contradicted kinship ties. It is now well known that the elementary family is common among most social groups, whatever their degree of complexity (see for instance Kuper 1994, p. 178). Yet, if some hunting groups have a quite elaborate social organization, the majority of them have almost no internal differentiations, except those based on age and sex. They have no state organization and have an economy that is largely based on exchange.

The concept of 'peasant society' is another interesting attempt to analyze the differences between hunters and agriculturalists. It was also a means of considering the latter in a comparative perspective. The works of Eric Wolf, Robert Redfield, or Marshall Sahlins (who spoke of a 'domestic mode of production') all tend in that direction. The problems of the 'peasant society' as a concept can be seen at first glance, since it comprehends such a vast and disparate ensemble. One could wonder what all agriculturists have in common, apart from working on the land. Yet the concept was quite fecund, and gave rise to interesting developments: Robert Redfield (1956), for instance, stressed the difference between the great and little traditions, and saw peasant communities as 'partial societies' that belong to larger entities. In France, Henri Mendras (1995) also coined a concept of peasant society based on three main features: the relative autonomy of the community, the importance of the domestic group, a relative self-sufficiency, the existence of a local collectivity, and the role of mediators between the local group and the main society.

Sociologists have always been preoccupied with the different forms of contemporary societies, and Max Weber was undoubtedly a precursor in this field. According to Weber, the rational organization of free labor is one of the main features of Western capitalism. Raymond Aron was much inspired by Weber in his definition of industrial society by five main criteria: separation between family and industry, an original type of production and division of labor, the accumulation of capital, rational economic calculation, and the concentration of the workers in the factory. Two main types of industrial societies could then be distinguished according to the role of the state in the distribution of resources and the type of property: the socialist and the capitalist society. Aron considered this distinction as typological in the sense that it had a merely instrumental nature, and therefore did not match reality. It was above all a tool. This is surely true of all typologies, which are useful inasmuch as they remain 'ideal-types,' in the Weberian sense of the term.

4. *Political Systems*

Political organization is another common means of classifying societies. Here, too, the first attempts tried to oppose 'modern' and 'traditional' societies, as the works of Rousseau or Hobbes show. In the political context, this opposition took the shape of a dichotomy between state and stateless societies. In *Primitive Society*, Robert Lowie argued that there is no radical difference between the two categories: some sort of political organization is present in any kind of group, and it is therefore possible to study its growing manifestations. However, the culturalist and relativist foundation of modern anthropology was more prone to stress the differences than the continuities, and anthropologists increasingly emphasized the gap between 'them' and 'us.' One of the most influential books in the field was undoubtedly *African Political Systems*, edited by Edward Evans-Pritchard and Meyer Fortes. Relying on their African examples, the contributors to this volume made a distinction between societies that have autonomous political institutions and those that have not. In the latter case, politics is expressed through kinship and the lineage system. The African examples were taken as having a universal value and social anthropologists saw lineages everywhere, and went as far as speaking of 'lineage societies'.

In a classical study, *The Nuer*, Evans-Pritchard (1940) asked himself how a 'society' could function in the absence of any political institutions and state organization. He discovered that while the Nuer have no state, they have a segmentary structure, in other words a division into segments of a similar nature, which merge or fight according to the circumstances. Evans-Pritchard's study became very influential, not only in social anthropology, but also in the political sciences, where *The Nuer* is considered as a case study of preindustrial society. In recent years, however,

social anthropologists became increasingly critical of the picture constructed by Evans-Pritchard. The model was undoubtedly elegant, but it contrasted with a rather chaotic reality on the ground. It also appeared that social organization was not solely characterized by the ties of kinship but that, on the contrary, the bands included many people who were not related by blood ties. In other words, the contrast between the Nuer and more advanced societies was perhaps less striking than was stated by the anthropologist. This is probably typical of the phenomenon of the insistence on the difference: in the Australian context, for instance, anthropologists have labeled as 'Kariera' and 'Aranda systems' societies that are divided into four or eight sections. Yet those very peoples also in fact live in nuclear families.

Clastres's opposition between society and the state is not very convincing either and reminds us of the dichotomies that were mentioned above. According to Clastres, primitive societies are stateless societies, and differ radically from state societies. According to him, the real revolution in history was not economic (the invention of agriculture), but political. The advent of the state, of power and domination, radically altered the mode of existence of societies and provoked a breach in human history (Clastres 1974, p. 172). The appearance of the state meant the death of 'society,' in the narrow sense of the term. Clastres was attacking Marxist economism, but he fell into a simplistic eulogy of primitivity, and amalgamated different realities under a single category.

Social anthropologists then sought to improve their typologies and spoke of bands or chiefdoms, or tribes without rulers. These categories may be useful, though only to a certain extent.

5. *From Structure to Relativism*

In sociology, the work of Talcott Parsons became one of the most interesting attempts to classify social groups according to their 'structure.' In his structuro-functionalist perspective, he identified four models of society. Radcliffe-Brown took up a similar viewpoint, and defined social anthropology as a comparative sociology, the aim of which was to compare social structures.

In social anthropology, however, his influence was undermined by a very strong reluctance to search for general laws. The studies of Ruth Benedict, for instance, put the stress on the irreducibility of every culture. By considering participant observation (or experience) as the main source of knowledge, social anthropology encouraged a kind of scepticism. Although she entitled her first book *Patterns of Culture*, Benedict spoke of the plasticity of human nature, and marveled at the infinite diversity of cultures. As for society, culture became plural. The US culturalists themselves never became hostile to comparison; but they had sown the seeds of a relativist ideology that invaded the social sciences and rejected any kind of generalization.

Relativism itself is an ideology that is historically determined. In a sense, it is true that all groups are different from one another (as indeed are all individuals), but the social sciences will have to go beyond this if they wish to prolong their existence. In that case, they will have to resume their elaboration of models and types, since classification and comparison remain elementary scientific activities. At the same time, typologies will never succeed in replacing analyses, and will always remain intellectually unsatisfactory. Moreover, when applied to societies, they also tend to reinforce the idea of homogeneous, well-defined entities, an idea that is problematic, both scientifically and politically.

See also: Community/Society: History of the Concept; Industrial Society/Post-industrial Society: History of the Concept; Peasants and Rural Societies in History (Agricultural History); Society/People: History of the Concept; Sociology: Overview; Structure: Social; System: Social

Bibliography

Amselle J-L 1999 *Logiques métisses: anthropologie de l'identité en Afrique et ailleurs*. Payot et Rivages, Paris

Aron R 1962 *Dix-huit leçons sur la société industrielle*. Gallimard, Paris

Boudon R 1998 *La Place du désordre: critique des théories du changement social*. Presses Universitaires de France, Paris

Clastres P 1974 *La Société contre l'État*. Minuit, Paris

Dumont L 1966 *Homo Hierarchicus: essai sur le système des castes*. Gallimard, Paris

Dumont L 1983 *Essais sur l'individualisme: une perspective anthropologique sur l'idéologie moderne*. Le Seuil, Paris

Durkheim E [1930] 1978 *De la division du travail social*. Presses Universitaires de France, Paris

Evans-Pritchard E 1940 *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People*. Clarendon Press, Oxford, UK

Ingold T (ed.) 1990 *The Concept of Society is Theoretically Obsolete*. Group for Debates in Anthropological Theory, London

Kuper A 1994 *The Chosen Primate: Human Nature and Cultural Diversity*. Harvard University Press, Cambridge, MA

Leach E 1968 *Rethinking Anthropology*. The Athlone Press, London

Lévi-Strauss C 1973 *Anthropologie structurale deux*. Plon, Paris

Lévy-Bruhl L 1951 *Les Fonctions mentales dans les sociétés inférieures*. Presses Universitaires de France, Paris

Meillassoux C 1974 *Femmes, greniers et capitaux*. Maspéro, Paris

Mendras H 1995 *Les Sociétés paysannes*. Gallimard, Paris

Mines M 1994 *Public Faces, Private Voices: Community and Individuality in South India*. University of California Press, Berkeley, CA

Morgan L H [1877] 1971 *La Société archaïque*. Anthropos, Paris
Parsons T 1951 *The Social System*. The Free Press, New York
Redfield R 1956 *Peasant Society and Culture: An Anthropological Approach to Civilization*. Chicago University Press, Chicago
Schnapper D 1999 *La Compréhension sociologique: démarche de l'analyse typologique*. Presses Universitaires de France, Paris
Service E 1966 *The Hunters*. Prentice-Hall, Englewoods Cliff, NJ
Tönnies F [1887] 1977 *Communauté et société: catégories fondamentales de la sociologie pure*. Retz, Paris

R. Deliège

Society/People: History of the Concept

'People' is a concept that has come to acquire two related, yet different meanings. On the one hand, there is the democratic, political conception of 'the people' — the people as *demos*—and on the other there is the ethnic, cultural conception of 'a people'—the people as *ethnos*. In either sense of the word, the crystallization of the modern concept of the people can be traced back to the latter part of the eighteenth century when the idea of the 'sovereignty of the people' became central to democratic political theory, and the nation-state emerged as the paradigmatic form of state institution, informed by the principle that ideally each state contained one nation/people, and that each nation/people constituted one state. In this article each meaning will be discussed, as well as their relationship to each other.

1. The 'People' and the Democratic and National Imaginaries

Prior to the eighteenth century the concept of the people was of relatively minor importance. Present from the outset were two basic meanings of the term, a general one, closely related to 'nation,' 'community,' and ' society,' and a particular one, which opposed the people to the elite. Take for example Samuel Johnson's *A Dictionary of the English Language* from 1755 which juxtaposes the following definitions: (a) 'A Nation, those who compose a community,' (b) 'The Vulgar,' and (c) 'The Commonality; not the princes or nobles.' Similarly, Abbé Fuertière's *Dictionaire universel* from 1690, list both a positive, general meaning—'the mass of persons who live in one country, who compose a nation'—and a negative one, defined 'by opposition to those who are noble, wealthy, and educated.'

During the eighteenth century the concept, which until then had carried largely negative connotations, associated with the vulgar, the lowly, and the beastly—what the Romans referred to as *plebs*—underwent a dramatic change, and 'the people' was valorized and reformulated in distinctly positive terms, as the source of both mortal virtue, political legitimacy, and cultural authenticity, a sentiment captured in the increasingly common saying, 'the voice of the people is the voice of God.' By around 1800 the 'people' was thus constituted as what Reinhart Koselleck calls a 'key-concept' (see Brunner et al. 1972–92). In dictionaries published after 1800 a veritable explosion of neologisms derived from the root-words 'people,' '*Volk*,' '*peuple*,' '*narod*,' 'folk,' and so on, can be observed, reflecting what has been called the 'discovery of the people' (Burke 1994).

This production of new words had two origins. On the one hand there were the terms derived from the vocabulary of the English, French, and American revolutions, to which the principle of the 'sovereignty of the people' was the central tenet. In opposition to the Old Regime, with its feudal, hierarchical society with the sovereign king at the top of the social and legal pyramid, a new idiom was elaborated that spoke in terms of citizenship, civil liberties, and legal equality. In England the primacy of Parliament and especially the House of Commons gave expression to this new political reality, which also brought with it a new sense of common, popular English nationality. In France the Third Estate—which contained the now celebrated people—became synonymous with the nation, while the nobility was excluded from such membership, associated, as Sieyès put it, with 'incurable idleness,' 'collapse of morality,' and a variety of exclusive and insulting social and political rights, privileges, and prerogatives. And in the USA, the Declaration of Independence spoke of the rights of 'we the people' to assert its sovereignty against the tyrannical and illegitimate powers of the English king.

On the other hand, there were Romantic, ethnographic terms inspired by the German philosopher Johann Gottfried von Herder and his followers who ventured into the countryside to explore the very popular culture that hitherto had been the objects of scorn and contempt. In 1774 Herder coined the word *Volkslied*, or 'folksong,' as the title of a collection of songs he published, and a host of similar words pertaining to 'folklore' and popular culture soon became commonplace in most European languages. As in the realm of politics, 'the people' in this ethnographic sense was now to be rescued from oblivion and condescension and instead to be presented as the authentic bearer of national culture. True poetry, Herder suggested, gives expression to the 'soul of the people,' and according to Jacob Grimm such authentic, national poems 'belong to the whole people.' For Herder, the people was first and foremost a matter of history and culture, and he harbored an organic vision of the concept. Each people was thus conceived as a collective individual that possessed its own national character, its own language, its own mores, its own morality, and its own sense of self. In contrast to the political definition of the people as *demos*, Herder thus stood in the forefront on an

elaboration of the people as *ethnos*. Indeed, Herder tended to perceive the emerging modern state with a great deal of suspicion, as a 'machine,' an 'artificial form of society' that threatens the organic, authentic culture of a people.

From a wider perspective these new terms belonged to a more general phenomenon of linguistic innovation during this period. As Lynn Hunt has observed in the context of the French revolution, 'words came in torrents' as the old political, social, and economic order broke down and attempts were made to semantically capture a new one (Hunt 1984). As the rate of change accelerated, experience could no longer guide action in the present. A new interpretative language was required to make sense of the world. Thus attempts had to be made to formulate a vocabulary that no longer simply referred to what Koselleck calls 'the space of experience,' but instead reached into the future, creating a 'horizon of expectation,' terms through which prognostic political goals could be hatched (Koselleck 1985). In other words, the concept of 'the people' acquired a utopian character. As such it came to occupy a central position in the 'democratic imaginary' (Joyce 1994), as a concept that did not merely serve to describe an existing state of affairs, but also prefigured the future. By elaborating visions of popular sovereignty, a politically potent expectation was created that has come to define the context for politics ever since. Similarly, the people imagined as a form of community—the nation—was a vision that came to inform European and global politics as nationalism became a fundamental political force, erupting in the revolutions of 1848— 'the springtime of the peoples'—in the unification of Italy and Germany, in the collapse of the Hapsburg Empire after World War I in the shadow of President Wilson's ideology of the right to national self-determination, and in the politics of anticolonial national liberation movements during most of the twentieth century.

2. *Structural Relations: Who Belongs to the People?*

In structural terms the difference between these two conceptions is one of how membership in 'the people' is delimited. In the one case the inclusion/exclusion is made in inside/outside terms. Here the emphasis is on setting limits against other peoples by defining the peculiar character of one's own people and culture in relation to those of others. In its most extreme and lethal form, such as the one practiced by the National Socialists in Hitler's Germany, the people may be defined in terms of race and blood, and citizenship be determined according to the strictures of *jus sanguinis*. Alternatively, membership may be determined through participation in what Rousseau called 'civil religion,' allegiance to a national constitution and to the law, a process by which individuals become active citizens, devoted to the nation, and to the practice of civic virtue. In this context, ethnicity tends to play a less defining role, and rights to citizenship may, as in France or the USA, be conferred automatically on anyone born in the country (*jus soli*) (Brubaker 1992). However, in either case, citizenship and nationhood is by definition simultaneously a matter of inclusion and exclusion, however benign and civic the criteria for membership in the national community may be. Thus the imagined community of the people/nation is always both sovereign and limited, surrounded by and defined against other 'peoples' who likewise are bounded and sovereign within those boundaries (Anderson 1983).

This emphasis on the people as a community built on affective ties of solidarity and comradeship furthermore requires a minimum of cultural if not ethnic homogeneity, and thus nation-states have consistently engaged in efforts at 'nationalizing the masses' through public institutions such as the schools and the military, creating a nationally minded people where there previously existed diverse and culturally heterogeneous populations, divided by all manner of differences in language, custom, and law, turning 'peasants into Frenchmen,' as the historian Eugen Weber memorably put it in the case of France (Hobsbawm 1983, Mosse 1975, Weber 1976).

In the other case, associated with the democratic definition of the people, the delimitation is made along a vertical axis where 'the people' is set against segments of the population at large that are defined as not belonging to the people. This can be done by excluding groups either at the top or the bottom. So, for example, the people can be defined as the entire population except the government, or the broad segments of the population vs. the ruling elite. It has also been done by excluding groups at the bottom. Thus, to take the German case, the working class could be seen as *vaterlandslose Gesellen* whose primary identification was with the international proletariat and not with the German nation. The words *Pöbel* (from French *peuple*) and *Masse* were deployed in similar ways, to distinguish the 'real,' respectable *Volk* from the degenerate, urban underclass. Indeed, the delimitation could be made in both directions simultaneously; the Social Democrats, for example, could claim the noble title of the *Volk* defined against both the capitalist elites above and the *Lumpenproletariat* below.

In the socialist vocabulary, the concept of the people had to compete with related notions that figured prominently in the Marxist idiom, particularly 'class' and 'proletariat,' especially since it was often closely linked to the concept of the 'nation,' an idea that sat uneasily within Marxist theory. Eventually, however, 'the people' retained a largely positive status, as evidenced in the common usage of the concept of 'the people's republic' in referring to a variety of communist regimes. In the context of the liberal-democratic political system of the West, the utopian

impulse at the heart of the democratic definition of the people, one that envisions not only the sovereignty of the people but also that all individuals within a sovereign state should belong to the people/*demos*, has led, at least in theory, to an increasing democratization of society and to the abolition of traditional obstacles to citizenship on the basis of categories such as class, creed, race, or gender.

3. *National Variations: Neighboring Concepts and Political Ramifications*

While the concept of the people has played a central role throughout the Western world and beyond, the impact of this key concept has varied considerably from country to country. This is true both when it comes to the specific political impact and how the 'people' came to be understood in relation to neighboring key concepts such as state, society, community, class, proletariat, and individual rights and liberties.

In the French case the ideas of the people and the nation became strongly linked to the state and the political notion of the general will. Thus the French nineteenth century historian Jules Michelet, in his influential book *Le Peuple* (1846), focused on the role of the Revolution and democratic politics as constitutive of the people/nation, and similarly Ernest Renan, in his essay, 'What is a Nation' (1882), emphasized that the nation should not be seen as first and foremost a matter of blood, but as an expression of political will, 'a daily plebiscite.' As Renan put it, when it comes to nationality, 'the people's wish is after all the only justifiable criterion, to which we must always come back.' Here Renan was significantly arguing against fellow German historians over the status of Alsace-Lorraine, the provinces that lay in the borderland separating Germany from France, an area annexed by the Germans in the wake of the Franco-Prussian war of 1870. By contrast, the German conception of the *Volk* and the nation was, pace Herder, at heart a cultural conception that at best saw the state as a protective shell safeguarding the *Volk/Nation*, or, as in the case of Herder himself, saw the state as an artificial and even destructive form of society. At the same time, the French and the German conceptualizations of the people were similar in that they both were collectivist. *Le peuple* and *das Volk* were both formulations that suggested a collective, be it a national, ethnically/culturally defined community, or a collectivity that expressed the democratic, general will of the people.

In England the idea of the people came, as in France, to be strongly associated with political institutions and practices. However, whereas in France the idea of the people/nation tended to be associated with the collective, communal notion of the general will, which gave priority to the nation and to the social over the rights of the individual, it was in England more strongly associated with the assertion of parliamentary power, individual liberty, and the limitation of state power. Thus English national identity itself came to be linked to political as well as economic liberalism, and to a conception of the general good that took as the basic unit the freeborn individual Englishman.

These tendencies were even more fully realized in the USA, where the separation of powers and the protection of individual rights were enshrined in the constitution itself. The beginning point was the stress on the American colonists' entitlement to 'English liberties' and civil rights, rights that, it was claimed, had been trampled upon. The next step was to separate from this close association to Englishness and instead to refer, as was done in the Declaration of Independence, to the 'inalienable rights' ultimately grounded in universal Natural Law and 'the rights of man.' However, in spite of this emphasis on universality, the Constitution—and the rights and liberties it guaranteed—soon became the basis for the elaboration of an American civic nationalism, based on the idea that the American people, formed by its liberal-democratic institutions, was exceptional in its commitment to individual liberty and equality. Thus in the USA, the idea of 'We the People' as the ultimate source of legitimacy and sovereignty was informed by a pervasive individualism and an enduring suspicious of centralized government. If in the French case the primacy of the people as a unified collectivity went hand in hand with strong central government, national standards, and a distaste for local difference, it was conversely true that in the US case the primacy of the people as a collection of free individuals informed an opposite tendency.

At the same time, Herderian notions of national character also came to infuse US self-understanding of the essential character of the American people, which revolved around the idea that America as a people and a nation was more mature and evolved than other peoples by virtue of its deep-seated capacity for democratic self-government. A similar fusion of an ethnic and a civic notion of the people occurred in Sweden, where leading Social Democrats after World War I came to embrace the idea that Swedes were endowed with a peculiar propensity for democracy and freedom, amounting to what one of them referred to as having 'democracy in the blood.' Citing a long, unbroken legacy of self-government that linked modern-day Swedes to the ancient yeoman peasants, the existence of a tradition was claimed, in which 'democracy is rooted not merely in the constitution, but also in our traditions and in the disposition of the folk,' as the Swedish prime minister Per Albin Hansson put it in 1935.

On this basis, Swedish Social Democrats conjured up a vision of a *folkhem*, a people's home, which usefully can be contrasted with the contemporary German slogan, *Volksgemeinschaft*, or 'people's community,' championed by Hitler and the National

Socialists. While both concepts celebrate national community, one prefigured the Nazi racial state, while the other became the organizing slogan for the Swedish welfare state. In the Swedish case, the concept of the *folk* connoted democratic sentiment rooted in a particular, common history; in the German one, the corresponding concept of the *Volk* referred to a racial community. This is particularly clear in the adjectival forms of *folk*/*Volk*, both equally difficult to translate into English. The German word, *völkisch*, means 'national' but with a distinct emphasis on race, and specifically with an anti-Semitic thrust. The Swedish word, *folklig*, conversely, is better translated as 'popular/populist/democratic,' with an association to the democratic legitimacy of claims made by those who credibly speak in the name of 'the people,' first and foremost the nonelite, that is, workers and peasants.

Thus the precise relationship between the concept of the people and such neighboring concepts as state, society, culture, race, and the individual become absolutely central if we are to understand the political implications of the concept of the people. In one case the proximity of *Volk*/nation/culture/race/collectivism is associated with racist nationalism; in another the linkage between *peuple*/nation/state/society connotes a collectivist, if democratic, conception of popular sovereignty; in a third the association with the people/individual liberty suggests a liberal and individual rights-based political order; while in a fourth one the correlation of folk/home/solidarity/state informs the vision of a welfare state.

See also: Citizenship, Historical Development of; Citizenship: Sociological Aspects; Civilization, Concept and History of; Civilizations; Community/Society: History of the Concept; Democracy; Democratic Theory; Ethnic Identity, Psychology of; Identity and Identification: Philosophical Aspects; Identity in Anthropology; Identity Movements; Identity: Social; Individual/Society: History of the Concept; Multiculturalism and Identity Politics: Cultural Concerns; Nationalism: General; Nationalism, Sociology of; Nations and Nation-states in History; Personal Identity: Philosophical Aspects; Social History; Social Identity, Psychology of; Societies, Types of; State and Society; State, History of; State, Sociology of the

Bibliography

Anderson B 1983 *Imagined Communities*. Verso, New York

Brubaker R 1992 *Nationhood and Citizenship in Germany and France*. Harvard University Press, Cambridge, MA

Brunner O, Conze W, Koselleck R (eds.) 1972–92 *Geschichtlische Grundbegriffe*, Vols. 1–7. Klett, Stuttgart, Germany

Burke P 1994 *Popular Culture in Early Modern Europe*. Scolar, Aldershot, UK

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Hunt L 1984 *Politics, Culture, and Class in the French Revolution*. University of California Press, Berkeley, CA

Joyce P 1994 *Democratic Subjects*. Cambridge

Koselleck R 1985 *Futures Past*. MIT Press, Cambridge, MA

Mosse G 1975 *The Nationalization of the Masses*. Cornell University Press, Ithaca, NY

Weber E 1976 *Peasants into Frenchmen*. Stanford University Press, Stanford, CA

L. Trägårdh

Sociobiology: Overview

'Sociobiology' refers to the study of social phenomena within the conceptual framework of biological science. More specifically, it usually refers to an explicitly neo-Darwinian approach that grew out of the theoretical contributions of W. D. Hamilton (1964) and G. C. Williams (1966). The hallmark of this approach is the interpretation of social behavior, and its causal mechanisms and processes, as products of the Darwinian processes of natural and sexual selection (see *Darwinism*), and hence in terms of adaptive function or 'design' (Dawkins 1986).

The term 'sociobiology' was apparently coined by J. P. Scott, a student of the genetics and development of mammalian social behavior, who began organizing paper sessions on animal behavior and sociobiology at scientific meetings around 1950. In 1956, the Ecological Society of America established a new section of animal behavior and sociobiology, the forerunner of today's Animal Behavior Society, with Scott as its chairman (Dewsbury 2000). However, the word 'sociobiology' was not widely used before 1975, when Harvard entomologist Edward O. Wilson made it the title of an ambitious volume reviewing and synthesizing information and ideas about the evolution of social behavior across the animal kingdom.

Wilson (1975, p. 4) defined sociobiology as 'the systematic study of the biological basis of all social behavior,' but although this definition is frequently quoted, it does not convey the word's meaning in actual usage. In the broadest sense, since 'biology' is 'the science of life and life processes, including the structure, functioning, growth, origin, evolution, and distribution of living organisms' (the *American Heritage Dictionary*), and since only living organisms have 'social behavior' in any interesting sense of the word 'social,' Wilson's definition could be construed as inclusive of all social science. However, 'sociobiology' is obviously not used as a synonym for 'social science' and most social scientists do not consider themselves biologists. A narrower construction of the meaning of

'biology' does not help Wilson's definition conform to usage: for most scientists, the phrase 'biological basis' connotes the genetic, endocrine, and brain mechanisms underlying social behavior, but these 'proximate causal' factors are not the primary focus of sociobiologists, and most work in these areas would not be considered sociobiological. In practice, both Wilson's tome and subsequent sociobiological work have been primarily concerned with applying theories from population biology, genetics, and ecology to the problem of explaining the evolution of social phenomena and their diversity both among and within animal species (see *Comparative Method, in Evolutionary Studies*), in much the same way as one might endeavor to understand the evolution and diversity of anatomical structures.

Like anatomists and physiologists, whose research is guided by their interpretations of the component parts and subsystems of organisms as devices 'for' digestion, respiration, vision, circulation of the blood, and so forth, sociobiologists are 'adaptationists' (see *Adaptation, Fitness, and Evolution*). That is, sociobiologists assume that social phenomena—such as mating preferences, offspring sex ratios, warning calls, systematically discriminative treatment of young by parents, situationally contingent gregariousness, seasonality of territorial defense, etc.—have a functional significance for the behaving animal that can be illuminated by appropriate hypothesis testing. Moreover, sociobiologists are 'selectionists' (see *Natural Selection*). That is, their hypotheses about the adaptive functions of social phenomena generally entail the assumption that the social attributes under consideration have been shaped by a history of Darwinian selection to be effectively fitness-promoting in ancestral environments (see *Adaptation, Fitness, and Evolution*). Indeed, effective contribution to fitness (the replicative success of the focal individual's genes relative to their alleles, and hence genetic posterity) is considered the superordinate criterion of adaptive function; more immediate aims, such as foraging efficiently, attracting mates, and protecting one's young from predators, are interpreted as having evolved as means to the end of fitness.

Sociobiological theory and research began to flourish only after the fallacy of naïve group-level adaptationism or 'greater goodism' (Cronin 1991) was dispelled. For over a hundred years, Darwin's (1859) concept of 'natural selection' was widely misunderstood, even by evolutionary biologists, who often uncritically interpreted evolved attributes as having been selected to promote 'the reproduction of the species' (see *Evolutionary Selection, Levels of: Group versus Individual*). Williams (1966) dissected and demolished this fallacy, persuading a majority of evolutionists that the capacity of Darwinian selection to create adaptive organization resides primarily in the differential reproductive success of alternative 'designs' within species, that the attributes that evolve therefore tend to be functionally organized to help individual organisms to outreproduce conspecific rivals, and that any ensuing benefits or harms to a whole population or species are incidental byproducts of organismic designs rather than what the designs are organized to achieve. This critique inspired a view of organisms as 'evolved reproductive strategists' whose attributes must be understood as tactics for effective reproductive competition against others of their own kind. (It is important to note that the 'strategy' metaphor carries no implication of foresight or intent. Indeed, the concept applies to brainless organisms as fully as to those with complex cognition. Thus, a sociobiologist might refer to a plant's 'reproductive strategy,' which is analyzed as consisting of numerous special-purpose 'tactics,' such as germinating in response to some degree of abrasion of the seed husk, flowering in response to a threshold air temperature, avoiding self-fertilization by asynchronous maturation of male and female flower parts, and so forth.)

The second main conceptual development leading to the flowering of sociobiological theory and research in the final third of the twentieth century was Hamilton's (1964) formulation of 'inclusive fitness' (or 'kin selection') theory (see *Adaptation, Fitness, and Evolution*). Classical Darwinism had construed the fitness that selection tends to maximize as personal reproductive success, but Hamilton noted that selection will favor any phenotype (see *Genotype and Phenotype*) that effectively promotes the replicative success of copies of the focal organism's genes, relative to their alleles (alternative forms of the same gene), regardless of whether those genes reside in descendants or in collateral relatives. Thus, for example, the complex brood care and colony maintenance behaviors of sterile worker ants can be favorably selected if these acts promote the reproduction of a queen who is closely related to the workers. Hamilton proposed that the quantity that selection tends to maximize is not personal reproduction (sometimes called 'direct fitness'), but the focal individual's 'inclusive fitness effect,' which is its total impact on the proliferation of its alleles via both direct fitness and any effects that the focal individual may have on the reproductive success of collateral kin ('indirect fitness'). Hamilton's analysis implied that organisms are not just evolved 'reproductive strategists,' but evolved 'nepotistic strategists' (see *Social Evolution: Overview*). In this context, 'nepotism' means directing benefits to one's genetic relatives, and the implication of Hamilton's revision of the fitness concept was, to quote Alexander (1979, p. 46), that living creatures 'should have evolved to be exceedingly effective nepotists, and we should have *evolved* to be nothing else at all.' (Note that Alexander, like most evolutionary biologists, uses the word 'nepotism' to encompass acts that promote both the direct and indirect components of fitness; confusingly, however, some authors use this term to refer only to acts that promote indirect fitness, excluding parental

care from its domain. Even more confusing is the common misuse of the term 'kin selection' as a substitute for 'nepotism,' such that an individual who is helping relatives is said to be exhibiting or engaging in kin selection. Used correctly, 'kin selection' refers to a population genetical process whose expected phenotypic consequence, on an evolutionary time scale, is nepotism, which is often but not necessarily manifested as selectivity in social response. Many sociobiologists, however, including Hamilton, have eschewed this popular phrase, both because it has so often been misused and because even its correct usage is often misinterpreted as logically parallel to, and theoretically alternative to, 'individual selection' or 'group selection'; see *Evolutionary Selection, Levels of: Group versus Individual*).

By the late 1980s, the word 'sociobiology' had begun to fall from favor. Most practitioners now refer to their field as 'behavioral ecology,' a term with slightly broader scope, in that it encompasses the adaptationist, selectionist approach not only to social phenomena but to asocial behavior as well, especially foraging. Thus, although one might suppose from its title that the journal *Behavioral Ecology and Sociobiology*, established in 1976, would cover a broader domain than that of *Behavioral Ecology*, established in 1990, the scope of the two journals is identical. Declining use of the word 'sociobiology' should not be mistaken for a decline in the paradigm's popularity. Indeed, during the 1980s, the sociobiological approach came to dominate the study of animal social behavior, as it still does (Alcock 2000).

1. *The 'Problem of Altruism'*

A central concern of sociobiological theory and research has been the 'problem of altruism': if Darwinian selection is a process that favors and refines only those attributes that contribute to their bearers' fitness relative to the fitness of conspecific rivals, then why is nature not more consistently 'red in tooth and claw'? In other words, how can such prosocial phenomena as restraint in aggressive conflict or active cooperation have evolved? In this connection, sociobiologists define 'altruism' as action whose average consequence is a reduction in the actor's direct fitness (expected reproductive success) and an increase in the direct fitness of someone else. (It should be noted that although this definition encompasses much of what might be deemed 'altruistic' in an ordinary English sense, since its essential feature is self-sacrifice that benefits others, it differs from ordinary usage in the fact that other-regarding intent is irrelevant.) At first glance, it would appear that selection must always penalize action that is altruistic in this technical sense, and yet animals often perform acts that clearly qualify, such as working to help raise others' young, or paying some energetic or predation risk cost in order to warn neighbors of danger. How these phenomena came into existence and are able to persist in a Darwinian world was a mystery before the 1960s, although the unexamined greater-goodism of many biologists kept them from recognizing that there was something in need of explanation.

Hamilton's inclusive fitness theory provided the most important answer. If the beneficiary of an altruistic act is a relative of the actor, and if that relative's gain is sufficiently large relative to the actor's sacrifice, then the net inclusive fitness effect for the actor can be positive and the altruistic tendency can increase in its relative frequency. More specifically, Hamilton theorized that the necessary condition for a heritable altruistic tendency to increase in prevalence under selection (often called 'Hamilton's rule') is $c/b < r$, where c is the cost (in units of expected direct fitness) incurred by the actor, b is the benefit (in the same currency) enjoyed by the altruistic act's beneficiary, and r is a 'coefficient of relatedness' between the actor and the beneficiary, indicative of the extent to which the two share identical genes by virtue of descent from recent common ancestors. Subsequently, more complex population genetical models confirmed the robustness and generality of Hamilton's rule, at least as a close approximation, making it a cornerstone of sociobiological theory (see, e.g., Frank 1998).

A great deal of sociobiological research has been concerned with testing whether altruistic behavior in nature can be understood as nepotism in accordance with Hamilton's rule. Supportive evidence has mostly taken the form of demonstrations that animals behave differently as a function of their relatedness to their interactants or neighbors; animals such as colonial ground squirrels and prairie dogs, for example, have been found to call out warnings of predator presence facultatively, according to the potential caller's degrees of relatedness to others in the vicinity. Progress in genetic methods of kinship assessment, especially 'DNA fingerprinting,' has fuelled an explosion of research on the ways in which social behavior in the field is affected by relatedness. There has also been considerable interest in elucidating how animals 'recognize' kin when they behave nepotistically; in some cases, they respond to cues that are under more or less direct genotypic control, such as certain odors, and in others to reliable circumstantial cues (e.g., those who share my mother's womb with me must be close kin), while in still other cases there is no evident discrimination among one's actual interactants at all but a 'rule of thumb' that does not require true recognition (e.g., behave cooperatively until I emigrate at maturity, at which time my average level of relatedness to those that I encounter will plummet).

The direct fitness costs incurred by altruists or nepotists and the fitness benefits that they bestow are usually harder to measure than relatedness, since they often depend on small changes in the risk of infrequent outcomes such as death by predation. However, one

area in which relatively precise measurement and quantitative hypothesis testing have been substantial is that of 'cooperative breeding,' where individuals other than a parent or pair of parents participate in the care and rearing of dependent young. The two taxonomic groups that have been most studied are the social insects, which manifest complex consequences of some unusual twists on genetic relatedness (Crozier and Pamilo 1996), and cooperatively breeding birds, in which helpers are usually, but not always, close kin of those they assist (Stacey and Koenig 1990). Recent theory and research on 'cooperative breeding' have focused especially on factors affecting the degree to which reproduction is monopolized or shared (e.g., Reeve 2000).

Theoretical sociobiologists have also sought nonnepotistic solutions to the problem of altruism, with particular interest in the circumstances that permit reciprocity ('you scratch my back and I'll scratch yours') to be evolutionarily stable. This issue has been addressed primarily by game theoretical models (see *Game Theory*), which have highlighted the necessity that nonreciprocators ('cheaters') be detected and denied the benefits provided to reciprocators, as well as the difficulty of establishing the requisite punishment of cheaters in multi-party games, when punishing is itself a sort of altruism (i.e., the provision of a 'public good'). Empirical investigations of nonnepotistic reciprocity in nonhuman animals have been few, however, perhaps indicating that the complex social cognition that is necessary for all but the most limited forms of reciprocation-in-kind is restricted to human beings and our near relatives.

2. *Sexual Selection and Sexual Conflict*

Another major focus of sociobiological theory and research has been sexual reproduction. Problems include when and why organisms reproduce sexually (i.e., biparentally) rather than uniparentally, and, in the event of 'dioecy' (the existence of two sexes with distinct adaptations), exactly how and why females and males differ. In 1871, Darwin supplemented his 1859 theory of evolution under the influence of 'natural selection,' by proposing a distinct evolutionary force that he called 'sexual selection' (see *Darwinism*). Whereas natural selection referred to the nonrandom differential reproduction of types as a result of differential success in coping with such exigencies as finding food, avoiding predators, and enduring climatic factors, sexual selection referred to the nonrandom differential reproduction of types as a result of differential access to mates and their reproductive labors. Darwin distinguished the two because they could sometimes act in opposition, as for example when sexual selection favors the evolution of conspicuous courtship structures, like the peacock's tail, that enhance mating success but apparently also increase mortality risk.

Sexual selection was the object of scant theoretical or empirical work for a century after Darwin proposed the concept. Williams (1966) and Trivers (1972) resurrected it as a focal topic in sociobiology, proposing that the extent to which sexual selection operates differently on females and males, and thereby induces the evolution of sex differences in anatomy, physiology, psychology, and behavior, was determined by the extent to which the reproductive efforts (especially parental care) provided by one sex constituted the principal resource limiting the fitness of the less investing sex (usually, but by no means always, males), and hence the object of competition among members of the less investing sex. More recent work has focused on the criteria and consequences of mating preferences, with particular interest in differentiating among theories that attribute the evolution of such preferences primarily: (a) to their utility for recruiting 'good genes' for one's offspring; or (b) to their utility for garnering material resources and/or minimizing such costs of mating as lost time or energy and elevated disease transmission risk; or (c) to 'sensory biases' whose primary function resides in contexts other than mate choice; or (d) to a self-reinforcing dynamic ('Fisherian runaway selection') that tends to produce display features and preferences that are arbitrary with respect to all criteria of functionality outside the mating context (see Cronin 1991, Andersson 1994).

Recent theory and research have also focused on conflicts of interest between mates (or potential mates) as well as those between same-sex rivals. Work in this area is related to problems in the evolution of communicative signals between agents whose commonality of interests is only partial (see *Game Theory*).

3. *'Selfish Genes' and Intragenomic Conflict*

An area of great contemporary interest concerns the evolutionary consequences of the fact that selection can operate differentially on different components of the genome. Because the male sex-determining Y-chromosome is transmitted only from father to son in such animals as fruit flies and human beings, for example, mutations on the Y that have the effect of biasing a progeny towards male production have been observed to increase in prevalence over generations, even when they reduce the inclusive fitness of the host individual (see *Y-chromosomes and Evolution*). Conversely, mitochondrial DNA and other cytoplasmic inheritance factors are transmitted only through daughters, and mutations in these factors that bias sex allocation towards females may similarly proliferate despite negative inclusive fitness effects (see

Mitochondria and Evolution). The consequence is 'intragenomic conflict': simultaneous selection at different genetic loci, within the same population, in favor of alleles with antagonistic phenotypic effects (Haig 1997).

The recent discovery of 'genetic imprinting,' a process whereby certain genes are differentially activated according to whether they were inherited from one's father or mother, has greatly extended the known potential for such intragenomic conflict, as for example when genes of paternal origin are selected to produce fetal and infantile phenotypes that effectively 'discount' the value of the mother's potential future offspring (who may have distinct paternity) more steeply than do genes of maternal origin, leading to the evolution of directly antagonistic phenotypic effects of maternally and paternally imprinted genes (Haig 1993). In a sense, these discoveries challenge the orthodox sociobiological view of organisms as nepotistic strategists with integrity of purpose subsidiary to the function of maximizing inclusive fitness, since selection clearly can and does favor the proliferation of genes that promote their own replication at the expense of 'their' organisms' inclusive fitness. However, the individual-level focus of sociobiological research continues to be productive (Alcock 2001), and is unlikely to be abandoned soon.

See also: Adaptation, Fitness, and Evolution; Comparative Method, in Evolutionary Studies; Darwin, Charles Robert (1809–82); Darwinism; Evolution and Language: Overview; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolution: Self-organization Theory; Evolutionary Epistemology; Evolutionary Selection, Levels of: Group versus Individual; Evolutionary Social Psychology; Game Theory; Genotype and Phenotype; Mitochondria and Evolution; Natural Selection; Social Evolution: Overview; Social Evolution, Sociology of; Sociality, Evolution of; Sociobiology: Philosophical Aspects

Bibliography

Alcock J 2001 *The Triumph of Sociobiology*. Oxford University Press, New York

Alexander R D 1979 *Darwinism and Human Affairs*. University of Washington Press, Seattle, WA

Andersson M 1994 *Sexual Selection*. Princeton University Press, Princeton, NJ

Cronin H 1991 *The Ant and the Peacock: Altruism and Sexual Selection from Darwin to Today*. Cambridge University Press, Cambridge, UK

Crozier R H, Pamilo P 1996 *Evolution of Social Insect Colonies: Sex Allocation and Kin Selection*. Oxford University Press, Oxford, UK

Darwin C 1859 *On the Origin of Species by Means of Natural Selection*. John Murray, London

Dawkins R 1986 *The Blind Watchmaker*. Longman, Harlow, UK

Dewsbury D 2000 Obituary: John Paul Scott (1909–2000). *Animal Behavior Society Newsletter* **45**(2): 3–5

Frank S A 1998 *Foundations of Social Evolution*. Princeton University Press, Princeton, NJ

Haig D 1993 Genetic conflicts in human pregnancy. *Quarterly Review of Biology* **68**: 495–532

Haig D 1997 The social gene. In: Krebs J R, Davies N B (eds.) *Behavioural Ecology*, 4th edn. Blackwell Scientific, Oxford, UK, pp. 284–304

Hamilton W D 1964 The genetical evolution of social behaviour, I and II. *Journal of Theoretical Biology* **7**: 1–16, 17–52

Reeve H K 2000 A transactional theory of within-group conflict. *American Naturalist* **155**: 365–82

Stacey P W, Koenig W D 1990 *Cooperative Breeding in Birds*. Cambridge University Press, Cambridge, UK

Trivers R L 1972 Parental investment and sexual selection. In: Campbell B (ed.) *Sexual Selection and the Descent of Man 1871–1971*. Aldine, Chicago, pp. 136–79

Williams G C 1966 *Adaptation and Natural Selection: a Critique of Some Current Evolutionary Thought*. Princeton University Press, Princeton, NJ

Wilson E O 1975 *Sociobiology: The New Synthesis*. Belknap Press, Cambridge, MA

M. Daly and M. Wilson

Sociobiology: Philosophical Aspects

Wilson (1975) published an overview of the evolution and adaptive significance of social behavior. It focused on the social behavior of nonhuman animals. But the final chapter applied evolutionary reasoning to human action; arguing that certain common patterns in human behavior were adaptations. That is, they exist because they conferred fitness advantages on our ancestors. Indeed, our ancestors were ancestors in part because they did, and others did not, manifest these behavioral dispositions. As a consequence, 'sociobiology' is a term often used narrowly, as a term for the evolutionary analysis of human social behavior, and that will be the focus of this article. Human sociobiology has been a very controversial enterprise (Kitcher 1985, Hull and Ruse 1998). Yet on the face of it, that is puzzling. We are evolved organisms. Human behavior has an evolutionary history; a history that must at least partially explain that behavior.

None of sociobiology's chief critics deny this basic fact. Instead, they have developed a minimization strategy, accepting the basic fact of human evolution, but arguing that human evolutionary history is largely irrelevant to an explanation of specific features of human social behavior. One idea is that human evolutionary history can explain only what is individually fixed or innate, and what is common across cultures. But the most striking features of human

experience are individual developmental plasticity and cultural variability. Thus, within the social sciences it is common to suppose that our evolved 'human nature' places only the broadest constraints on our cultural life. If humans reproduced asexually, or if human infants were independent at 18 months of age, human cultures would be very different. But human evolutionary history makes possible many human cultures and social organizations. Since every human group has a similar set of biological resources, the massive differences between groups must be explained in terms of differing cultural resources. Difference explains difference. Human evolutionary history explains how hominids developed the ability to transmit culture and the plasticity to be shaped by that culture. But that is all it explains (Lewontin et al. 1984, Levins and Lewontin 1985).

Sociobiologists resist this line of argument. In particular, Tooby and Cosmides argue that human cultural diversity and individual developmental plasticity is more bounded than the minimisation strategy suggests. And they argue that diversity itself may have an evolutionary explanation. Organisms are adapted to behave differently in different circumstances. The differences between castes of social insects are often profound. Yet these differences are different developmental expressions of a single adaptation. Equally, a single mechanism of mate choice might generate one behavior in a New York singles bar, and quite a different one in an Amish church (Tooby and Cosmides 1992). Sects. 2 and 3 outline of the development of human sociobiology since 1975. The article then concludes with three short sections introducing key conceptual and methodological issues raised by contemporary sociobiology.

1. *The Wilson Program*

Wilson and his co-workers began with the insight that behavioral differences are just as subject to selection as any other feature of an organism. Animals within a population are as apt to differ in behavior as they are in any other aspect of their phenotype. These differences make a difference to fitness. Moreover, there is good reason to expect that behavioral differences are heritable. Evolutionary histories reconstructed from behavioral traits agree well with those from morphological and genetic ones (see, e.g., Kennedy et al. 1996). So selection can work. The mound building behavior of the orange-footed scrubfowl—a bird that lays its eggs in a mound that functions as an incubator—is just as likely to be an adaptation as its large powerful feet.

Wilson simply extended this idea from animal to human behaviors. He had in mind such behaviors as incest avoidance, male sexual promiscuity and female coyness, infanticide, rape, and hostility to strangers. Just as the scrubfowl has been selected to build, guard, and monitor the temperature of its mound, perhaps humans have been selected to act with hostility towards strangers or to use infanticide to manipulate the sex ratio of their offspring (Wilson 1975, Wilson 1978).

Adaptationist analyses of behavior can seem quite plausible. Consider sex role differentiation. In some bird species, the male takes responsibility for brooding the eggs and caring for the chicks. But, in general females, and especially female mammals, have a higher initial investment than males in any reproductive act. Eggs are much larger, and hence more biologically expensive, than sperm. Second, the costs of pregnancy and postparental care are very significant. By accepting a sexual partner, a female mammal commits a serious fraction of her total life reproductive resources. That need not be true of males: they do not bear the costs of pregnancy and lactation. Sex is not without cost to males. Copulation exposes the male to predators, parasites, and pathogens. But the female also bears these costs together with those of pregnancy. So mating costs do not fall evenly on the sexes, and this suggests different optimal strategies for males and females. Males are likely to attempt to be more promiscuous; females, more coy. We might also expect some gender differentiation in caring decisions. If her young are to live, a female mammal has no choice but to engage in childcare. Not so the male, who is unconstrained by physiology. For him, all options are possible. These range from outright desertion of the pregnant female, through diversion of some of his resources to other mates and their young, to full participation in childcare.

Adaptive hypotheses of this kind about human behavior are all very controversial. This is not surprising, for they are hard to test. A standing problem for human sociobiology is to move beyond plausibility considerations to clear empirical tests, and it is not clear how that is to be done. However, the most fundamental problem with the Wilson program is its grain of analysis. Lewontin has pointed out that an adaptive explanation of a trait presupposes that the trait can, over evolutionary time, vary independently of others (Lewontin 1978). It is, thus, very unlikely that there is any selective explanation for the color vision properties of the human left eye. The color vision capacities of the left retina do not vary (heritably) independently of those of the right retina. Its receptor sensitivities are not adaptive or evolutionary atoms. Human skin color, on the other hand, almost certainly is such an atom.

Behavioral traits are sometimes atoms. The scrubfowl's mound building and mound maintaining behavior probably has evolved independently, and operates independently, of the scrubfowl's feeding, roosting, and predator avoiding activities. The human behavioral repertoire, however, is not an aggregation of independent units. The realization that human action depends on a small number of interacting psycho-

logical mechanisms propelled psychology from behaviorism to cognitivism. Our behavior is produced by mental mechanisms that play a role in many different behaviors. Some of the mental mechanisms used in hunting are also used in storytelling. Rape, xenophobia, child abuse, homosexuality, and other human behaviors very likely do not vary independently of other human actions. They could only be altered by altering the underlying mental mechanisms, and since these mechanisms are used for many different purposes, any change would have many other consequences. Particular behavior patterns are unlikely to have independent selective histories.

2. *Darwinian Psychology*

The idea that we should not be looking for adaptive hypotheses about specific behaviors has gradually become part of the accepted wisdom of human sociobiology. The psychological mechanisms that generate behavior are the proper focus of evolutionary theorizing. Sociobiologists have turned from the idea that each behavior evolved because it was selected to the idea that many different behaviors are caused by a relatively small number of cognitive mechanisms. It is these mechanisms that have evolved (Barkow et al. 1992, Plotkin 1997).

Evolutionary psychologists defend a general schema of the evolved nature of the human mind, and they have developed a discovery procedure for filling out that schema. They reject the metaphor of human psychology as a general-purpose computer programmed differently by different cultures. They replace this vision with a modular theory of mind. The mind is a cluster of evolved information-processing mechanisms. The main empirical program of evolutionary psychology is to characterize these 'Darwinian algorithms.' For example, Buss and Symons think that there are Darwinian algorithms of sexual attraction. These result in the tendency of human males to find those females attractive that bear the cultural marks of youth, and of women to find attractive those that bear the cultural marks of high status (Symons 1979, Buss 1994). Baron-Cohen and Sperber have argued that we have a specific mental mechanism, a 'mind reading' device, designed to enable us to anticipate others' behavior (Baron-Cohen 1995, Sperber 1996). Darwinian algorithms are supposed to be mental 'modules' in the sense defined by Fodor: domain-specific, mandatory, opaque, and informationally encapsulated (Fodor 1983). They are domain-specific because they deal with a specific class of situations in which the organism finds itself. The mind reader is not used to predict the outcome of physical interactions. Darwinian algorithms are mandatory. We do not choose to interpret a particular smile as false. Rather, our mechanisms of emotion recognition are engaged automatically once we recognize a particular stimulus as a face. Darwinian algorithms are opaque because their internal processes are not consciously accessible. If Cosmides and Tooby are right, we have cognitive mechanisms specifically adapted to detect cheating in social exchange (Cosmides and Tooby 1992). But we cannot tell by introspection whether we have such mechanisms or how they operate. Finally, Darwinian algorithms are informationally encapsulated because they do not make use of the information stored elsewhere in the cognitive system.

Chomsky looms large in the development of evolutionary psychology. If he is right, the differences between human languages are not profound. There are many very important features common to all human languages. The class of possible human languages is constrained quite tightly by the nature of an innate language acquisition device. Moreover, that device contains conditional elements whose different settings explain many of the differences between languages. Language thus illustrates two themes of evolutionary psychology. Diversity is less profound than it appears, and it can be explained by a single mechanism; one operating differently in different circumstances.

Evolutionary psychologists take the human mind to be a complex of modules, each adapted to solve a specific problem or cluster of problems that was central to the life of our ancestors. But how are we to identify and characterise these modules? They hope to solve this problem by the strategy of 'adaptive thinking.' Adaptive thinking infers the solution, the adaptation, from the problem, the ecological context in which the organism evolved. So the first task of the evolutionary psychologist is to identify the adaptive problems our ancestors confronted in their environments. For example, in complex and unstable social groups, our ancestors would have needed social expertise; they would have needed to be good judges of others' likely actions (Whiten and Byrne 1997). The second task is to discover the stable correlations between those aspects of the environment humans are equipped to sense, and those aspects they need to know about. We can expect natural selection to engineer into task-specific devices implicit knowledge of these correlations. So if, for example, aggressive behavior or defecting from an agreement is usually preceded by a characteristic stance or facial expression, the evolutionary psychologist would expect this to be engineered into those mechanisms specialized for managing social interaction. The third task of the evolutionary psychologist is to construct an information-processing design that could solve the adaptive problem using the available cues. The fourth and final task is to experimentally test for the existence of the hypothesized mechanisms. For many potentially useful adaptations will not actually be engineered into us.

Human sociobiology has developed from a theory of the adaptive significance of particular behavior

patterns to a theory of the architecture of the human mind, and of the adaptive significance of the components of that architecture. Much remains controversial about this program, both in its detail and its overall orientation. So this article will finish by outlining three critical issues that are still open.

3. Testing Sociobiological Hypotheses

One central problem for sociobiology is to move beyond plausibility stories to develop clear empirical tests of specific sociobiological hypotheses. It is not obvious that this problem can be solved at all. The resources of comparative biology offer one of the most powerful methods of testing adaptationist hypotheses (Harvey and Pagel 1991). For example, suppose we suspect that a certain leaf shape is an adaptation to arid conditions. If we show that within a particular tree family (e.g., the eucalypts) that there is a strong covariation between long, narrow, waxy leaves and distribution through an arid habitat, we could reasonably regard that hypothesis as well confirmed. It would be confirmed especially well if the ancestor of our (putative) arid-adapted trees did not have long and narrow leaves, for we could then rule out the hypothesis that the leaf shape was no more than an inheritance from a common ancestor. If we found the same pattern in other tree species, our confidence in our adaptive hypothesis would be stronger still. Unfortunately, we cannot apply the methods of comparative biology to test the most interesting adaptive hypotheses about human psychological mechanisms. Our closest relatives are extinct. So many of these mechanisms are found only in our own species. Even where they are shared with our closest living relatives (e.g., the facial expression of emotion) they exist in a different form. Moreover, the human lineage—and even the human and great ape lineage combined—is not bushy; the eucalypt, in contrast, boasts hundreds of species. The comparative method can tell us nothing at all about the adaptive significance of language and very little about tool making, true imitation, and many other important human cognitive adaptations.

The main alternative to comparative biology is the development of quantitative evolutionary models. If there is a good quantitative fit between model and data, we can regard the model as passing an important test (Orzack and Sober 1994, Orzack and Sober 1996). But here too the prospects are dim. First, it is often hard to actually measure the impact of behavior on human fitness. Economic resources typically are used as a measure of fitness benefits. But even in hunter–gatherer societies this is probably too crude. For fitness probably does not vary as a linear function of economic resources. Second, in developing an evolutionary model, we need to define a space of possibilities. For the adaptationist hypothesis is that the fittest behavior within that space will become established by selection in the population. Yet how do we develop a possibility space for human behavior? To recycle on example from Kitcher, what were the realistic alternatives of, say, a man in an avunculate society? Could he defect from the avunculate, directing his resources to his wife's children, rather than to his nieces and nephews (Alexander 1979, Kitcher 1985)? Finally, and most importantly, we face the problem that humans now act in a very different world from that in which human psychological mechanisms evolved. So measuring the impact of behavior on current fitness seems beside the point. Yet, of course, we have no access to the fitness consequences of behaviors in those ancestral environments.

So the problem of testing sociobiological hypotheses is very serious. The shift in theoretical interest from behavior to the mechanism that produces that behavior in one way makes the problem even more intractable. The Wilson program took the unit of adaptationist analysis to be a behavioral pattern. Such patterns are observable, though if they are patterns in sexual behavior, violence to outsiders, or infanticide, they will not be easy to observe. The program of evolutionary psychology adds to the existing methodological problems yet another: inferring cognitive mechanism from behavior. This is known to be difficult. But in one important respect, evolutionary psychology improves the prospects for testing sociobiological hypotheses. To the extent that evolutionary psychology makes specific architectural claims about human cognition, these are subject to direct empirical test. For example, Frank has argued that emotions function as commitment devices; they bind agents to future courses of action that will not then be in the agent's interest, and thus tend to maximise the agent's overall long run welfare (Frank 1988). It is a consequence of this theory both that emotions are not under central control, and that other agents are good at using emotional signals to predict future behavior in commitment situations. These consequences are directly testable by experimental psychology.

However, a positive test does not directly vindicate the adaptive hypothesis that is alleged to explain the existence of the proximate mechanism. Moreover, the relationship between adaptive function and proximate mechanism is often less direct than this example suggests. Sober and Wilson point out this in their discussion of altruism. They argue that there are good reasons to believe that humans are a group-selected species, and hence that human behavior will often be altruistic in the evolutionary sense. Its function will often be to promote the fitness of the group at some cost to the agent's fitness. But Sober and Wilson point out that while this constrains the proximate mechanism, it does not fix it. Altruistic agents might act altruistically because of irreducibly altruistic motivations, or because selection has built in internal rewards.

Altruistic behavior could be engineered either way (Sober and Wilson 1998). In short, then, despite the shift from adaptive theories of behavior to adaptive theories of psychological mechanism, the problem of the testability of sociobiology remains very serious.

4. Evolutionary Psychology and Rational Agency

Suppose the basic hypothesis of evolutionary psychology is right. Human minds are ensembles of adapted modules. What follows for folk psychology, i.e., belief/desire theories of human behavior? Many models in the social sciences depend on folk psychology, either directly or on its formalized cousin, the view that humans are rational agents that maximize their expected utility. Does evolutionary psychology drive out rational choice theory? The idea that humans are rational agents is compatible with a partially modular theory of mind, to one along the lines of Fodor's 1983 model. For he retains a central processor with the critical role of maintaining an overall model of the agent's environment, and using that model to guide rational action. A Fodorian agent could well be a rational maximizer; indeed, a rational economic maximizer. But evolutionary psychology in its standard formulation seems inconsistent with rational choice theory understood as a descriptive claim about human behavior. If there is no general-purpose device—if there is nothing like a central processor—it is hard to see how a rational actor model could accurately describe the pattern of human action. When humans respond to problems that were important in the environments in which our cognitive mechanisms evolved, and when they do so in environments relevantly like those ancestral environments, and when fitness consequences correlate well with individual welfare, then of course human behavior might maximize expected utility. But much human behavior is not directed on problems that were critical to our ancestors, and humans often act in environments very different from those in which psychologically modern humans evolved.

One defender of evolutionary psychology to confront this problem is Sperber. He defends the view that the mind is wholly an ensemble of modules. But he argues that one of these modules, the metarepresentation module, can act as a surrogate general-purpose device. The function of the metarepresentation module is to guide our behavior with respect to other humans. For we can estimate their actions best by estimating their representation of the world. Since the proper function of the metarepresentation module is to represent representations, we have a derived, second hand capacity to represent anything, or at least anything that anyone else can represent (Sperber 1996). However, this suggestion does not address the issue of integration and the effective use of information to guide behavior with respect to novel aspects of our physical environment. So it is not clear that the behavioral abilities of a Sperber agent would match those of a Fodor agent.

So we may have to choose between a wholly modular theory of the mind and rational agent psychology. Perhaps, though, evolutionary psychologists should retreat from their commitment to a wholly modular theory of mental organization. For there are good reasons for thinking that a commitment to a modular view of the mind is premature. Adaptive thinking suggests modular mechanisms have a downside: they are vulnerable to exploitation in a malign world. For if they are encapsulated, they use only a fraction of the information potentially available to an agent. Yet, if our minds are the result of an arms race, they evolved in a hostile world, not merely an indifferent one. Game theoretic models of the evolution of language have a strong cooperative element; they are close to one end of a mathematical spectrum from games of pure cooperation to games of pure conflict (Skyrms 1996). Both parties in a linguistic interaction benefit from successfully communicating their intended meaning. Even if they have further, incompatible agendas, neither will succeed unless the utterance is understood. A module for resource sharing might well be manipulated to gain a better share of resources. Similarly, if there are mate choice modules that work on inflexible, prewired cues, these two will be vulnerable to mimics.

Perceptual systems exemplify a surprising truth about human mentality. The information-processing tasks implicit in much human action are much more complex and difficult than one would intuitively expect. The outputs of the visual systems are reliable representations of what is seen, yet the signals that reach the retina are typically fragmentary and equivocal. That supports the view that those tasks could only be carried out by mental organs specifically adapted for those very tasks. This realization has generated a series of 'poverty of the stimulus' arguments in many branches of cognitive psychology. These attempt to show that human cognitive skills are too complex, and develop from too narrow an information basis, to be directed by a general learning mechanism.

However, the 'poverty of the stimulus' argument does not sustain some of the central hypotheses of evolutionary psychology. For example, Cosmides and Tooby argue that we have a module of social exchange (Cosmides 1989, Cosmides and Tooby 1989, 1992). But their reasoning is the inverse of 'poverty of the stimulus' reasoning; it depends on the fact that we usually find certain computationally trivial reasoning tasks extraordinarily difficult, but when those tasks concern social exchange we find them easier. So humans do an easy task in one domain with less difficulty than the same task in other domains. A poverty of the stimulus argument applies in the reverse situations, those where we find a computationally complex task effortless. Determining motion from

changes in apparent shape and size is computationally extraordinarily complex, yet we do it easily.

Moreover, many important problems cannot be solved by modular mechanisms. Fodor has argued convincingly that the pragmatics of language cannot be handled by a specialist device (Fodor 1983). It is one thing to know what a sentence means; it is another to know the intentions that lie behind its utterance. The latter problem is not solvable by shortcuts from a restricted database i.e., by an encapsulated device. Everything the hearer knows is potentially relevant and potentially used in decoding the speaker's intent. The same problem seems to arise for many of domains of interest to evolutionary psychology. Could an encapsulated mechanism deliver reliable judgments about the probability of a prospective partner's cheating? It is not at all obvious that there are cues that are reliable and are stable across generations. Both conditions must be met, if selection is to build a specialist mechanism to solve such problems. So the most plausible version of evolutionary psychology might be somewhat closer to Fodorian architecture; a complex of domain general and domain specific devices. If that is so, there may be ways of integrating evolutionary psychology and rational choice theory.

5. *Individual Behavior and Collective Action*

One central problem in the social sciences concerns the relationship between the explanation of individual action, and the explanation of collectives: of the nature and activity of cultures, classes, institutions, and the like. It is uncontroversial that there is a minimum constraint of the individual on the collective. Social facts supervene on individual facts, for social facts cannot change unless individual facts do, but not vice versa. A somewhat stronger claim is also relatively uncontroversial: social or collective causation requires individual microfoundations. So, for example, a theory of working-class collective action that depends on class consciousness owes us an account of the thoughts and actions of the individual agents whose pattern of individual action constitute that collective action (Elster 1989, Wright et al. 1993).

So one important issue about evolutionary psychology is concerned with the microfoundations it makes possible, and those that it rules out. Here the single most central issue concerns cooperation. Human life is pervasively cooperative, and this is a fact that is notoriously hard to explain consistently with the assumption that humans are rational economic maximizers. For the prisoner's dilemma and its relatives threaten to subvert cooperation between such agents, even though all would be better off under cooperative regimes. An array of solutions to this problem have been developed around reciprocation and iterated interaction (Axelrod 1984). These retain the assumption that individual agents act to maximize their own economic welfare, but it is far from obvious that these models are robust or realistic. For example, the famous 'tit for tat' strategy is by no means robust against error. If either player mistakes cooperation for defection, cooperation collapses (Sigmund 1993). So one important role for sociobiology might be to expand the space of microfoundational options for the social sciences. If Frank or Skyrms is right, in some circumstances evolved individual agents might reach the cooperate–cooperate option, even though agents that maximize their own expected utility would not (Frank 1988, Skyrms 1996). If Sober and Wilson are right, a proper regard for evolution does not commit us to the view that individual agents are designed to maximize their individual fitness (Sober and Wilson 1998). A theory of individual agency based on evolutionary psychology might, therefore, be significantly different from microfoundations based on an economic version of rational agency. However, the nature and importance of these differences remain to be elucidated.

See also: Evolution, Natural and Social: Philosophical Aspects; Evolutionary Social Psychology; Evolutionary Theory, Structure of; Evolutionism, Including Social Darwinism; Psychological Development: Ethological and Evolutionary Approaches; Sociobiology: Overview

Bibliography

Alexander R D 1979 *Darwinism and Human Affairs*. Washington University Press, Seattle, WA

Axelrod R 1984 *The Evolution of Cooperation*. Basic Books, New York

Barkow J H, Cosmides L, Tooby J (eds.) 1992 *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, Oxford, UK

Baron-Cohen S 1995 *Mindblindness: An Essay on Autism and the Theory of Mind*. MIT Press, Cambridge, MA

Buss D M 1994 *The Evolution of Desire: Strategies of Human Mating*. Basic Books, New York

Cosmides L 1989 The logic of social exchange: has natural selection shaped how humans reason? Studies with the wason selection task. *Cognition* **31**: 187–276

Cosmides L, Tooby J 1989 Evolutionary theory and the generation of culture, part II. Case study: A computational theory of social exchange. *Ethology and Sociobiology* **10**: 51–97

Cosmides L, Tooby J 1992 Cognitive adaptations for social exchange. In: Barkow J H, Cosmides L, Tooby J (eds.) *The Adapted Mind*. Oxford University Press, Oxford, UK, pp. 163–227

Elster J 1989 Marxism, functionalism and game theory: The case for methodological individualism. In: Callinicos A (ed.) *Marxist Theory*. Oxford University Press, Oxford, UK

Fodor J A 1983 *The Modularity of Mind*. MIT Press, Cambridge, MA

Frank P 1988 *Passion Within Reason: The Strategic Role of the Emotions*. W W Norton, New York

Harvey P H, Pagel M D 1991 *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, UK
Hull D L, Ruse M (eds.) 1998 *Philosophy of Biology*. Oxford University Press, Oxford, UK
Kennedy M, Spencer H G, Gray R D 1996 Hop, step, and gape: Do the social displays of the Pelecaniformes reflect phylogeny?*Animal Behavior* **51**: 273–91
Kitcher P 1985 *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. MIT Press, Cambridge, MA
Levins R, Lewontin R C 1985 *The Dialectical Biologist*. Harvard University Press, Cambridge, MA
Lewontin R C 1978 Adaptation. *Scientific American* **239**: 156–69
Lewontin R C, Rose S, Kamin L J 1984 *Not In Our Genes*. 1st ed. Penguin, London
Orzack S H, Sober E 1994 Optimality models and the test of adaptationism. *American Naturalist* **143**: 361–80
Orzack S H, Sober E 1996 How to formulate and test adaptationism. *The American Naturalist* **148**: 202–10
Plotkin H 1997 *Evolution in Mind: An Introduction to Evolutionary Psychology*. Penguin, London
Sigmund K 1993 *Games of Life: Explorations in Ecology, Evolution and Behavior*. Oxford University Press, Oxford, UK
Skyrms B 1996 *The Evolution of the Social Contract*. Cambridge University Press, Cambridge, UK
Sober E, Wilson D S 1998 *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, Cambridge, MA
Sperber D 1996 *Explaining Culture: A Naturalistic Approach*. Blackwell, Oxford, UK
Symons D 1979 *The Evolution of Human Sexuality*. Oxford University Press, Oxford, UK
Tooby J, Cosmides L 1992 The psychological foundations of culture. In: Barkow J H, Cosmides L, Tooby J (eds.) *The Adapted Mind*. Oxford University Press, Oxford, UK, pp. 19–136
Whiten A W, Byrne R W (eds.) 1997 *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge University Press, Cambridge, UK
Wilson E O 1975 *Sociobiology: The New Synthesis*. Belknap Press of Harvard University Press, Cambridge, MA
Wilson E O 1978 *On Human Nature*. Bantam Books, Toronto, ON
Wright E O, Levine A, Sober E 1992 *Reconstructing Marxism*. Verso Press, London

K. Sterelny

Sociocybernetics

The term cybernetics derives from the Greek word for steersman. It can be translated roughly as the art of steering, and refers here to a set of related approaches, including especially general systems theory, and first- and second-order cybernetics. It can offer some interesting concepts and models to social science, and may also in that sense have a steering function.

Sociocybernetics consists of the applications of first-order and especially second-order cybernetics to the social sciences, and the further development of the latter (Buckley 1988). First-order cybernetics was developed in the 1940s, outside of the social sciences, whereas second-order cybernetics was developed to remedy the shortcomings of first-order cybernetics when applied to biology and the social sciences. For reasons of space, the essentials of first-order cybernetics and its main concepts—system boundaries; the distinction between systems, subsystems, and supra-systems; circular causality; positive and negative feedback loops; simulation—cannot be dealt with here (see Geyer and van der Zouwen 1991).

1. *Modern (Second-order) Cybernetics*

Second-order cybernetics originated in the early 1970s. The term was coined by Heinz von Foerster (1970) in 'Cybernetics of cybernetics.' He defined first-order cybernetics as the cybernetics of *observed* systems, and second-order cybernetics as the cybernetics of *observing* systems.

Indeed, second-order cybernetics explicitly includes the observer(s) in the systems to be studied. These are generally living systems, ranging from simple cells to human beings, rather than the control systems for inanimate technological devices studied by most first-order cybernetics. It could be said to have a biological approach, or at least a biological basis.

Apart from von Foerster, several other authors haven given concise definitions of the differences between first-order and second-order cybernetics: the purpose of a model vs. the purpose of the modeler (Pask), controlled systems vs. autonomous systems (Varela), the interaction among variables in a system vs. interaction between the observer and the system observed (Umpleby), and theories of social systems vs. theories of the interaction between ideas and society (Umpleby). This latter difference reflects Parsons's approach as that of a first-order systems theorist interested in system stability and system maintenance, and Luhmann's as that of a second-order cybernetician interested more in change and morphogenesis.

2. *The Usefulness of Second-order Cybernetic Concepts For Sociology*

The main concepts of second-order cybernetics—self-organization, self-reference, self-steering, autocatalysis, and autopoiesis—all start with 'self' (or 'auto'), possibly reflecting accelerating social change, whereas first-order cybernetics concepts do not point specifically to either the system or its environment.

2.1 *Self-organization*

Contemporary scientific interest in many disciplines centers on the emergent, self-organizing properties of complex aggregates, especially in biology and social science, and has stimulated second-order cybernetics. First-order cybernetics, though certainly aware of self-

organizing systems, did not pursue this line of inquiry. Ashby (1952, 1956), for example, also stressed self-organization. Recent developments in cognitive science—an interdisciplinary effort of neuroscience, artificial intelligence (AI), linguistics, philosophy, and cognitive psychology—demonstrate the importance of self-organization. Though the science only began in the 1970s, one can already distinguish two schools: cognitivism and connectionism.

Cognitivism and cognitive technology furthered the development of AI. Cognitivism conceived mental processes as manipulations according to the rules of deductive logic, as localized and serial information processing of symbols supposedly representing external reality (Varela et al. 1993, pp. 4–5). Thus, it seems to be close to first-order cybernetics. It was the mainstream paradigm in cognitive science until its drawbacks became more apparent in the 1970s.

(a) At some stage, symbolic information processing encounters the 'Von Neumann bottleneck': sequentially applied rules become problematic in case of large numbers of sequential operations, e.g., weather forecasting or image analysis.

(b) Symbolic information processing is not distributed but localized. Local malfunctioning of some symbols or rules therefore tends to result in overall malfunctioning, not having the resilience towards disturbances of distributed processing.

Connectionism, contrary to cognitivism, incorporates many second-order cybernetics views: It is explicitly bottom-up rather than top-down, eschews abstract symbolic descriptions, stresses emergence and self-organization, and assumes simple, 'stupid,' neuron-like components that develop cognitive capacities when appropriately connected. The intelligence is in the structure, in the connections made—hence connectionism—and not in the components. Hebb's 1949 rule occupies a central place. It states that learning is based on postulated changes in the brain resulting from correlated activity between neurons. If two neurons are active together, their connection is strengthened; if not, it is weakened. Thus, history—always a history of transformations—is incorporated. Simulations of a simple network of neurons according to Hebb's rule demonstrate that pattern recognition is possible, after a learning phase in which some of the connections are strengthened or weakened, provided the number of patterns presented does not exceed about 15 percent of the participating neurons.

The components of neural networks do not need an external processing unit which guides their operations. They work locally according to local rules, but global cooperation emerges when the states of all participating neurons reach a mutually satisfactory state. This passage from local rules to global coherence is the essence of self-organization—also denoted as emergent or global properties, network dynamics, nonlinear networks, complex systems, and so forth. Once one looks for it, emergence abounds. In different domains, like vortices and lasers, chemical oscillations, genetic networks, population genetics, immune networks, ecology and geophysics, self-organizing networks give rise to new properties.

The implications of these developments in cognitive science are potentially interesting for sociology, because of the possible analogies with self-organization in human societies:

(a) Autonomous systems, though computer-simulated in connectionism, give meaning to their interactions on the basis of their own history, rather than on the basis of the intentions of the programmer—or should one say a manipulating environment?

(b) Neural networks, and possibly human networks as well, produce emergent phenomena resulting from both simultaneous processes (the emergent pattern itself arises as a whole) and sequential ones (participating components have to engage in back-and-forth activity to produce the pattern).

2.2 *Self-reference*

Self-reference is typical of human beings, and possibly apes, both on the individual and the group level. It is not a concept in first-order cybernetics, which—as Norbert Wiener ([1948] 1961, 1956) stressed—concerns itself with the commonalities rather than the differences between humans, animals, and machines.

Self-reference is defined here in its second-order cybernetics meaning, where the (individual or social) system collects information about its own functioning, which in turn can influence that functioning. Minimal requirements are self-observation, self-reflection, and some degree of freedom of action. One of the main characteristics of individual and social systems is indeed their potential for self-referentiality in this sense. The self-knowledge accumulated by the system not only affects the structure and operation of that system, but also implies that (self-referential) feedback loops exist between parts of external reality and models and theories about these parts of reality.

Whenever social scientists systematically accumulate and publish new knowledge about the structure and functions of their society or subgroups within it—in principle available also for those to whom that knowledge pertains—such knowledge will often be invalidated, because the research subjects may react to this knowledge in ways that falsify the analyses or forecasts made. In this respect, social systems are different from many other systems, including most biological ones: There is a clearly two-sided relationship between the system's self-knowledge and its behavior and structure.

While biological systems also show goal-oriented behavior of actors, self-organization, self-reproduction, adaptation, and learning, only psychological and social systems arrive systematically, through experiment and reflection, at knowledge about their

own structure and operating procedures, with the obvious aim to improve these—both on the microlevel of the individual, as in psychoanalysis or other self-referential activities, and on the macrolevel of world society, as in planning international trade or optimal distribution of available resources.

The usual social science examples of self-referential behavior consist of self-fulfilling and self-defeating prophecies. Henshel (1990), for example, has studied *serial* self-fulfilling prophecies, where the accuracy of earlier predictions, themselves influenced by the self-fulfilling mechanism, impacts upon the accuracy of the subsequent predictions. However, in most empirical research, self-referential behavior does not loom large, perhaps because it is largely survey research, which offers less opportunity for self-referential behavior than depth interviews. These concentrate more on the 'why' than the 'what' of people's opinions, and thus tend to elicit self-referential remarks.

2.3 Self-steering

The futility of large-scale planning efforts has led to the realization that individuals and organizations are largely self-steering. While recognizing the usefulness of efforts to steer societies, a cost–benefit analysis of intensive large-scale steering efforts often turns out to be negative: Intensive steering implies intensive social change, i.e., a long period of increased uncertainty, with also an increased chance for changing planning preferences and for conflicts between different emerging planning paradigms during such a period.

Aulin (1982, 1988) tried to answer two basic questions in this respect.

(a) Should one opt for the 'katascopic' or the 'anascopic' view of society, for top-down planning of individual and group behavior to maximize societal survival chances, or should the insight of actors at every level, including the bottom one, be increased and therewith their competence to handle their environment more effectively and engage more successfully in goal-seeking behavior?

(b) What should be the role of science, especially the social sciences, in view of this choice: should it deliver useful knowledge for an improved centralized steering of the behavior of social systems and individuals, or should it strive to improve the competence of actors at the grass-roots level, so that these actors can steer themselves and their own environment with better results?

Aulin followed a cybernetic line of reasoning that argues for nonhierarchical forms of steering. Ashby's Law of Requisite Variety indeed implies a law of requisite hierarchy if only the survival of the system is considered, i.e. if the regulatory ability of the regulators is assumed constant. However, the need for hierarchy decreases if this regulatory ability itself improves—which is indeed the case in advanced industrial societies, with their well-developed productive forces and correspondingly advanced distribution apparatus (the market mechanism). Since human societies are not simply self-regulating systems, but self-steering systems aiming at an enlargement of their domain of self-steering, advanced industrial societies have the possibility nowadays to combine societal governability with ever less control, less centralized planning, and less concentration of power. When moving from a work-dominated society to an information-dominated one, less centralized planning is even a prerequisite. The intellectual processes dealing with information are self-steering, not only self-regulating, and cannot be externally steered by definition. Consequently, there should be no excessive top-down planning, and science should help individuals in their self-steering efforts, and certainly should not get involved in the maintenance of hierarchical power systems.

Of course, there are types of systems within a society that can indeed be planned, governed, and steered, simply because such systems have been specifically designed to exemplify the concept of the control paradigm of first-order cybernetics. However, even in first-order cybernetics control does not necessarily imply hierarchy, as even the simple case of the thermostat demonstrates. Thus, although there certainly still is a limited place for organizations that exemplify a more or less hierarchical control paradigm, modern, complex, multigroup society in its entirety, conceptualized as a matrix in which such systems grow and thrive, can never be of this type.

If one investigates a certain system with a research methodology based on the control paradigm, the results necessarily have a conservative bias. Changes of the system as such are almost prevented by definition. De Zeeuw (1988) considers a different methodological paradigm necessary if one wants to support fundamental social change and prevent 'post-solution' problems. It is based on a multiple-actor design, does not strive towards isolation of the phenomena to be studied, and likewise does not demand a separation between a value-dependent and a value-independent part of the research outcomes.

2.4 Autocatalysis and Cross-catalysis

Laszlo (1987, 1988) describes two varieties of (chemical) catalytic cycles: auto-catalytic cycles, in which the product of a reaction catalyzes its own synthesis, and cross-catalytic cycles, where two different products or groups of products catalyze each other's synthesis. An example of a model of a cross-catalytic cycle, developed by Prigogine and colleagues, is the Brusselator:

(1) $A \rightarrow X$
(2) $B + X \rightarrow Y + D$
(3) $2X + Y \rightarrow 3X$
(4) $X \rightarrow E$

With X and Y as intermediate molecules, there is an overall sequence in which A and B become D and E. Step 3 can be seen as autocatalysis, while steps 2 and 3 in combination describe cross-catalysis: X helps to produce Y, while Y helps to produce more X. Autocatalytic sets ($A \rightarrow B \rightarrow C \rightarrow \cdots \rightarrow Z \rightarrow A$) bootstrap their own evolution, provided the complexity of interactions is rich enough. The system then changes from a subcritical to a supercritical state, and autocatalysis follows. Kaufmann (1991, 1992) even uses the concept to explain the origins of life from a 'primal soup' of simple chemical elements as an inevitable production of order, rather than as a unique and extremely unlikely historical accident. Simple chemical laws coupled with the presence of a sufficient number of frequently interacting elements produce ever more complex elements. These have new characteristics, and often turn out to be part of new catalytic processes at higher levels of molecular complexity—processes which in turn boost the emergence of still higher levels of complexity. Along similar lines, Swenson (1991) likewise maintains that—in spite of the Second Law of Thermodynamics—'the world is in the order production business.'

The economist Arthur (1990), collaborating closely with Kaufmann at the Santa Fe Institute, applied the concept of autocatalytic sets to the economy—which also bootstraps its own evolution, as it grows more complex over time. Beyond a certain critical threshold, phase transitions occur. Stagnant developing countries can enter the take-off stage when their economy has diversified sufficiently; increased trade between two countries in a subcritical state can similarly produce a more complex and interwoven economy which becomes supercritical and explodes outward. Catalytic effects may also operate in 'negative' phase transitions, where critical thresholds of violence are passed as in Northern Ireland or Bosnia.

2.5 *Autopoiesis*

Autopoiesis, or 'self-production,' is a concept introduced in the 1970s by the biologists Maturana and Varela to differentiate the living from the nonliving. An autopoietic system was defined as a network of inter-related component-producing processes such that the components in interaction generate the same network that produced them. Although they considered the concept applicable only in biology, and not in the social sciences, an interesting 'theory transfer' was made by Luhmann (1988). He defended the quite novel thesis that, while social systems are self-organizing and self-reproducing systems, they do not consist of individuals—or roles or even acts, as commonly conceptualized—but exclusively of communications. When generalizing the usages of autopoiesis, developed while studying biological systems, to make it also applicable to social systems, the biology-based theory of autopoiesis should therefore be expanded into a more general theory of self-referential autopoietic systems. Social and psychic systems are based upon another type of autopoietic organization than living systems: namely on communication and consciousness, respectively, as modes of meaning-based reproduction.

While Luhmann thus views communications instead of actions as the elementary units of social systems, the concept of action admittedly remains necessary to ascribe certain communications to certain actors. Thus, the chain of communications can be viewed as a chain of actions. This enables social systems to communicate about their own communications and to choose their new communications, i.e., to be active in an autopoietic way. Such a general theory of autopoiesis has important consequences for the epistemology of the social sciences: It draws a clear distinction between autopoiesis and observation, but also acknowledges that observing systems are themselves autopoietic systems, subject to the same conditions of autopoietic self-reproduction as the systems they are studying.

The theory of autopoiesis thus belongs to the class of global theories, i.e. theories that point to a collection of objects to which they themselves belong. Classical logic cannot really deal with this problem, and it will therefore be the task of a new systems-oriented epistemology to develop and combine two fundamental distinctions: between autopoiesis and observation, and between external and internal (self-)observation. Classical epistemology searches for the conditions under which external observers arrive at the same results, and does not deal with self-observation. Consequently, societies cannot be viewed, in this autopoietic perspective, as either observing or observable. Within a society, all observations are by definition self-observations.

3. *The Science of Complexity: A Convergence of Paradigms*

Compared with simpler societies, complex modern ones are highly dynamic and interactive. Consequently, they change at accelerated rates, and are generally in a far-from-equilibrium situation. According to Prigogine and Stengers (1984) nonlinear relationships obtain in systems that are far from equilibrium, where relatively small inputs can trigger massive consequences. At such 'revolutionary moments' or bifurcation points, chance influences determinism, and the direction of change is inherently impossible to predict: a disintegration into chaos, or a 'spontaneous' leap to a higher level of order or organization—a so-called 'dissipative structure,' which requires more energy to sustain it, as compared with the simpler structure it replaces.

In stressing this possibility for self-organization, for 'order out of chaos,' Prigogine comes close to the

concept of autopoiesis. In modern societies, the mechanistic and deterministic Newtonian world view—emphasizing stability, order, uniformity, equilibrium, and linear relationships between or within closed systems—is being replaced by a new paradigm, more in line with today's accelerated social change, which stresses disorder, instability, diversity, disequilibrium, nonlinear relationships between open systems, morphogenesis, and temporality. Social scientists, often still thinking in terms of linear causality, would be well-advised to try out the explanatory powers of Prigogine's conceptual vocabulary—fluctuations, feedback amplification, dissipative structures, bifurcations, (ir)reversibility, auto-and cross-catalysis, self-organization, and so forth—on the phenomena they study, as well as the concepts and methods of second-order cybernetics in general.

It is reassuring in this novel and therefore risky field of research that there seems to be a convergence of paradigms, subsumed here under the name of second-order cybernetics. However, it might also be designated by other names like cognitive science, general systems theory, artificial intelligence, artificial life, or perhaps indeed most aptly, as Prigogine suggested, 'the emerging sciences of complexity.'

4. *Methodological Problems of Empirical Research in First- and Second-order Cybernetics: A Bridge Too Far?*

If one realizes that the social sciences mainly study self-organizing, self-referential, autopoietic systems, with their own strategies and expectations, with intertwining processes of emergence and adaptation, then one is confronted with one of the core problems of sociology, economics, and other social sciences: how to make a science out of studying of imperfectly smart agents exploring their way into an essentially infinite space of possibilities which they—let alone the social scientists researching them—are not even fully aware of.

There is indeed quite a methodological problem here. It is already very difficult to apply the principles and methods (e.g., feedbacks and nonlinearities) of first-order cybernetics to empirical social research. It is nearly impossible to incorporate a second-order cybernetics approach in one's research design. Second-order cybernetics may be a bridge too far, given the research methodology and the mathematics presently available.

Applying first-order cybernetics principles in empirical research already poses heavy demands on the data sets and the methods of analysis: Every feedback ($X_t \rightarrow Y \rightarrow X_{t+1}$), every interaction between variables [$Z \rightarrow (X \rightarrow Y)$], and every nonlinear equation ($Y = cX^2 + bX + a$), let alone nonlinear differential equation ($Y' = cY^2 + bY + a$), demands extra parameters to be estimated, and quickly exhausts the information embedded in the data set. Admitting on top of that the second-order notions that the research subjects can change by investigating them, and may even reorganize themselves on the basis of knowledge acquired by them during the research, exceeds the powers of analysis of even the most sophisticated methodologists: it equals the effort to solve an equation with at least three unknowns.

How can one obtain reliable data within such a framework, where nothing is constant and everything is on the move, let alone formulate policy-relevant decisions? How can one still forecast developments when at best retrospective analysis of how a new level of complexity has emerged seems possible? These problems are far from solved, and much work lies ahead before hypotheses derivable from second-order cybernetics will be fully testable, at least in a classical Popperian sense. However, testing hypotheses pertaining to the nonlinear dynamical regimes stressed by complexity theory and chaos studies may demand different criteria, invalidating the classical concept of a theory as a mathematical model yielding stable, consistent and complete predictions of a system's behavior. Nevertheless, the opportunities offered by the second-order cybernetic paradigm to present a truly realistic analysis of the complex adaptive behavior of interacting groups of agents seems to good to pass up. There is certainly a challenge here, for theorists and methodologists alike.

See also: Functionalism, History of; Functionalism in Sociology; Structure: Social; System: Social; Theory: Sociological

Bibliography

Arthur W B 1990 Positive feedbacks in the economy. *Scientific American* **262**: 92–9

Ashby W R 1952 *Design for a Brain—The Origin of Adaptive Behavior*. Wiley, New York

Ashby W R 1956 *An Introduction to Cybernetics*. Chapman & Hall, London

Aulin A 1982 *The Cybernetic Laws of Social Progress: Towards a Critical Social Philosophy and a Criticism of Marxism*. Pergamon, Oxford, UK

Aulin A 1988 Notes on the concept of self-steering. In: Geyer F, van der Zouwen J (eds.) *Sociocybernetic Paradoxes—Observation, Control and Evolution of Self-steering Systems*. Sage, London, pp. 100–18

Buckley W 1988 *Society—A Complex Adaptive System*. Gordon and Breach, London

De Zeeuw G 1988 Social change and the design of enquiry. In: Geyer F, van der Zouwen J (eds.) *Sociocybernetic Paradoxes—Observation, Control and Evolution of Self-steering Systems*. Sage, London, pp. 131–44

Geyer F, van der Zouwen J 1991 Cybernetics and social science: Theories and research in sociocybernetics. *Kybernetes* **20**(6): 81–92

Hebb D O 1949 *Organization of Behavior*. Wiley, New York

Henshel R L 1990 Credibility and confidence loops in social prediction. In: Geyer F, van der Zouwen J (eds.) *Self-referencing in Social Systems*. Intersystems Publications Salinas, CA pp. 31–58
Kauffman S A 1991 Antichaos and adaptation. *Scientific American* **265**: 78–84
Kaufmann S A 1992 *Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press, Oxford, UK
Laszlo E 1987 *Evolution—The Grand Synthesis*. Shambala, Boston
Laszlo E 1988 Systems and societies: The basic cybernetics of social evolution. In: Geyer F, van der Zouwen J (eds.) *Sociocybernetic Paradoxes—Observation, Control and Evolution of Self-steering Systems*. Sage, London, pp. 145–72
Luhmann N 1988 The autopoiesis of social systems. In: Geyer F, van der Zouwen J (eds.) *Sociocybernetic Paradoxes—Observation, Control and Evolution of Self-steering Systems*. Sage, London, pp. 172–92
Prigogine I, Stengers I 1984 *Order out of Chaos—Man's New Dialogue with Nature*. Flamingo, London
Swenson R 1991 End-directed physics and evolutionary ordering: Obviating the problem of the population of one. In: Geyer F (ed.) *The Cybernetics of Complex Systems—Self-Organization, Evolution, and Social Change*. Intersystems Publications, Salinas, CA, pp. 41–60
Varela F, Thompson E, Rosch E 1993 *The Embodied Mind—Cognitive Science and Human Experience*, 3rd edn. MIT Press, Cambridge, MA
Von Foerster H 1970 Cybernetics of cybernetics. Paper delivered at 1970 annual meeting of the American Society for Cybernetics. Washington, DC
Von Foerster H 1974 On constructing a reality. In: von Foerster H (ed.) *Observing Systems*. Intersystems Publications, Salinas, CA
Wiener N 1956 *The Human Use of Human Beings—Cybernetics and Society*. Doubleday, Garden City, NY
Wiener N [1948] 1961 *Cybernetics, or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, MA

F. Geyer

Socioeconomic Status and Health

Observation of an inverse association between socioeconomic status (SES) and risk of poor health and death are not new, extending back to the twelfth century (Antonovsky 1967). In medieval Europe, Paracelsus noted unusually high rates of disease in miners. In 1790, Johan Peter Frank, one of the founders of the public health movement, wrote 'The people's misery, mother of disease.' Daniel Defoe in his *Journal of the Plague Year* echoed similar sentiments when he opined that '… the misery of that time lay upon the poor who, being infected, had neither food or physic, neither physician or apothecary to assist them, or nurse to attend.'

By the nineteenth century, systematic investigations were being conducted by Villermè into the relationship between rent levels of areas and mortality in Paris (Susser et al. 1985). In 1848, Virchow reported on the relationship between poor living conditions and typhus in Upper Silesia (Rather 1988). In England, Farr examined differences in mortality by occupation (Rosen 1993), while Engels (1848) deplored the impact of new working conditions on the health of the poor as a result of the industrial revolution in England.

Since 1850, a rich scientific literature has developed that further extends these earlier observations that link socioeconomic position with health, and in recent years there has been an exponential increase in scientific work on this topic. As a prelude to summarizing this literature, it is first necessary to briefly discuss the ways in which socioeconomic status or position has been conceptualized in this research tradition. A variety of terms have been used in the epidemiologic literature in this area, including social class, social stratification, social inequality, social status, and socioeconomic status (Krieger et al. 1997). To a large extent these terms reflect different historical, conceptual, and disciplinary roots. More extensive discussions of the details and the potential advantages and disadvantages of various approaches to measurement have been presented in reviews by various authors (cf. Krieger et al. 1997). It should be noted that in Europe, and particularly the United Kingdom, occupational measures of social class have predominated, whereas in North America there has been more focus on individual or aggregate measures of socioeconomic status derived from education, income, or occupation. In what follows, the term 'socioeconomic position' (SEP) will be used generically to refer to the various measures that have been used.

1. The Relationship Between Socioeconomic Position and Health

Figure 1 shows the association between relative risk of death and household income, after adjustment for age, sex, and household size (Wolfson et al. 1999) in a large sample of households in the United States. The figure also shows the relative distribution of levels of household income in the population. There are several important aspects of this relationship. First, risk of death, relative to mean household income, decreases monotonically with increasing income. Second, while risk of death increases monotonically with decreasing income, the greatest elevations of risk relative to mean income are found in approximately the bottom third of the household income distribution. This monotonic but nonlinear relationship between socioeconomic position and mortality risk or life expectancy has been found using many different measures of SEP and in many different countries. While a great deal of attention has been devoted to the monotonic nature of this risk relationship, much less attention has been devoted to the striking nonlinearity in this association.

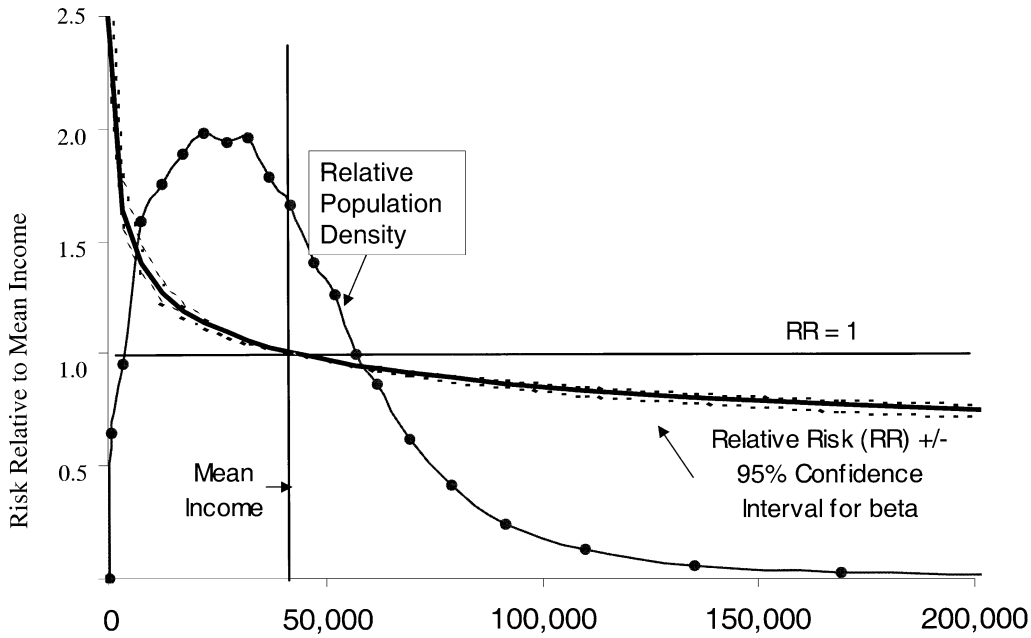

Figure 1
Plot of risk relative to mean income (modified from Wolfson et al. 1999)

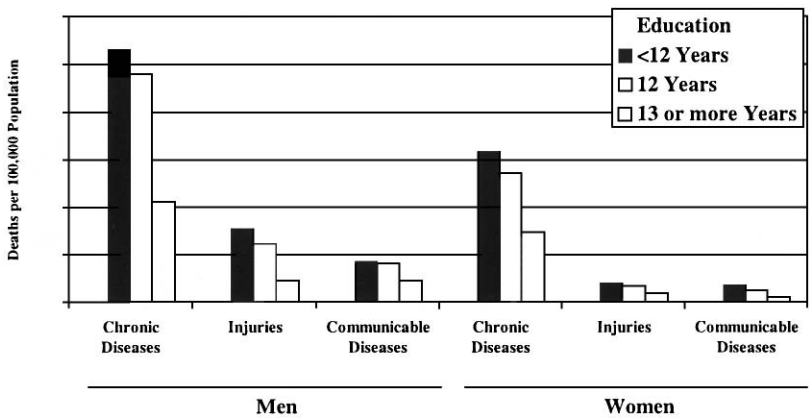

Figure 2
Age-adjusted death rates for selected causes for adults 25–64 years of age, by education level and sex: US 1995.
Note: Injuries include homicides, suicides, unintentional injuries and deaths resulting from adverse effects of medical procedures. Source: Pamuk et al. 1998

The association between SEP and health outcomes is not restricted to any particular age group, although some evidence suggests it is somewhat attenuated during late adolescence, early adulthood, and later old age. For example, Ericson et al. (1984) found in Sweden that lower parental SEP was associated with higher rates of perinatal mortality, prematurity, low birth weight, and being born small for gestational age; and rates of 'stunting' (below the 5th percentile in height) were elevated in children born in families living below the poverty level in the United States (USDHHS 1986). Wolfson et al. (1993), using Canadian data, showed important differences in survival after age 65 based on income levels just before retirement, and the effects of educational level on other health outcomes continue into old age (Guralnik et al. 1993).

Lower SEP is associated with increased risk for most diseases and health outcomes. Figure 2 shows the

relationship between level of education and rates of death from chronic diseases, injuries, and communicable disease for men and women in the United States (Pamuk et al. 1998). In Susser et al.'s (1985) compilation of findings from the Registrar General's Decennial Supplement for England and Wales, they noted that lower occupational class was related to higher rates of death from malignant neoplasms, endocrine, nutritional and metabolic diseases, diseases of the blood-forming organs, infective and parasitic diseases, diseases of the nervous system and sense organs, diseases of the circulatory system, diseases of the respiratory system, diseases of the digestive system, diseases of the genitourinary tract, diseases of the musculoskeletal system and connective tissue, congenital anomalies, and accidents, poisoning, and violence. For those outcomes where there is a positive relationship between SEP and disease risk, such as breast cancer, survival is generally worse for those at lower levels of SEP.

The association between SEP and health outcomes in not fixed in time, with some evidence for increasing stratification of health by SEP (Black et al. 1980, Pappas et al. 1993). For some health outcomes, there have also been changes in the direction of the association (Morgenstern 1980). These observations of changes in the magnitude and direction of the SEP–health association, as well as the substantial international variation (Kunst et al. 1998), are important because they suggest that informed interventions could possibly reduce socioeconomic inequalities in health.

2. *Explanations of Inequalities in Health*

Typical explanations for SEP-related variations in health status focus on the role of medical care, behavioral risk factors such as smoking or alcohol consumption, material standards of living, or the impact of poorer health on SEP (Black et al. 1980). While there are likely to be variations in access to and quality of medical care by SEP, with the magnitude of the difference varying widely by country, medical care is not likely to be a major cause of variations in inequality by SEP. The argument in support of this assertion is based, among other things, on observations that the introduction of widespread availability of medical care in some countries has not lowered SEP variation in health and that variations in health status by SEP are found for conditions that are not 'amenable to medical care.'

The differential distribution of high-risk behaviors among lower SEP populations also does not seem to fully explain the SEP–health association. Studies that have information on such individual-level risk factors have generally been unsuccessful in completely explaining SEP gradients in health (Marmot et al. 1978). Furthermore, even if statistical adjustment for individual higher risk behaviors did reduce or eliminate SEP-related gradients in health, SEP-related factors are likely to be antecedent, causal determinants of high-risk behavior.

Some have interpreted SEP-related gradients in health as reflecting the impact of poor health on income and wealth. In this view, the poor end up less healthy because they were, first, less wealthy. While it is undoubtedly true that chronic illness is in most contexts associated with declines in income and wealth, most researchers have concluded that these selection effects account for a relatively small amount of the association between SEP and health. For example, downward selection effects cannot easily account for the pervasive impact of level of education on health outcomes that occur long after the completion of education, nor the impact of SEP of one family member on the health of another, nor Wolfson's finding of strong variations in postretirement survival by pre-retirement income in Canadian men who had no declines in income during the years before they retired (Wolfson et al. 1993).

SEP is generally related to material standard of living, while the magnitude of the association clearly varies between countries, and over time. Nonetheless, lower SEP is generally associated with poorer housing, less nutritious diet, greater exposure to adverse environmental substances, poorer working conditions, greater exposure to crime and dangerous environments, and a wide variety of other 'material' and psychosocial factors. It seems likely that any comprehensive understanding of the relationship between lower SEP and poorer health will have to simultaneously consider a wide variety of these pathways, but the dominant analytic strategy has been to focus on the primacy of one pathway or another.

3. *Future Directions*

It is unlikely that any single set of explanations will account for the observed relationships between SEP and health across time, places, and health outcomes. Undoubtedly, any full understanding will involve many of the above pathways, and perhaps others. Current efforts are focused on a number of activities aimed at clarifying our understanding of this complex area of investigation.

3.1 *Individual vs. Area SEP*

Many studies that lack individual measures of SEP have used measures that reflect the SEP of the area in which these individuals live, e.g. median level of income, percent in poverty, or area deprivation scores. In many cases the motivation is to use these as proxy

measures of individual information. However, it is increasingly being recognized that such measures reflect properties of the social and economic environment in which people live and that these measures are not identical to individual measures of SEP. What is more, a number of studies indicate that such area measures are importantly associated with health outcomes beyond individual (Macintyre et al. 1993). Considerable work remains to clarify under what situations these area measures are related to health outcomes, their overlap with neighborhood and community phenomena, and the economic, social, and behavioral pathways that link area characteristics with health outcomes in individuals.

3.2 Inequality in Income Distribution and Health

A growing literature indicates that inequalities in income distribution are importantly related to population health outcomes beyond individual incomes (Wilkinson 1996). For example, in the United States, the share of income received by the least well-off half of the population was correlated with age-adjusted mortality rates for states and metropolitan areas, independent of measures of level of income (Kaplan et al. 1996). Greater equality in the distribution of income was associated with lower mortality rates and other health outcomes. While controversial, these findings have led to a lively debate concerning the relative role of individual income, psychosocial pathways, and material conditions in individual and population health (Lynch et al. 2000, Wilkinson 1996).

3.3 Life Course Effects

Most studies have examined SEP at only one time and few have considered either variability in SEP over the short- or long-term. However, there is considerable variation beginning in childhood, based on macroeconomic and individual pathways, on parental SEP and adult educational, occupational, or income trajectories. Some evidence suggests that income trajectories are important (Blane et al. 1996), while others suggest that adult destinations are more important (Lynch et al. 1994). A growing literature is devoted to documenting life-course effects, many of which involve variations in SEP on adult health outcomes (Kuh and Ben-Schlomo 1997).

3.4 Biological Pathways

Evidence is rapidly accumulating on the proximal biological pathways that mediate the association between SEP and health outcomes, which form much of the evidence relating to cardiovascular disease (Brunner 1997). For example, measures of SEP have been found to be associated with the prevalence and progression of subclincial atherosclerosis (Lynch et al. 1997), with levels of cholesterol and other lipids, blood pressure, and hemostatic factors (Kaplan and Keil 1993), and with the clustering of cardiovascular risk factors (Helmert et al. 1989). A growing focus is on the cumulative physiologic effects of chronic stress and adaptive efforts and their link with SEP (McEwen 1998).

4. Conclusion

The pervasive associations observed between various measures of SEP and many health outcomes, and the risk factors for these outcomes, supports the fundamental role of SEP as a determinant of health. Understanding these associations requires elucidating the variety of multilevel determinants and pathways that link SEP and individual health outcomes as well as determining the societal and individual forces that stratify these determinants and pathways along socioeconomic dimensions. An appreciation of the breadth and magnitude of the imprint of SEP on the health of individuals and populations suggests that this is a scientific area with potentially great importance for public policy.

See also: Ethnicity, Race, and Health; Health Care Delivery Services; Health Care, Rationing of; Health Care Systems: Comparative; Health Economics; Minority Access to Health Services: United States; Public Health; Public Health as a Social Science; Socioeconomic Status and Health Care

Bibliography

Antonovsky A 1967 Social class, life expectancy, and overall mortality. *Milbank Memorial Fund Quarterly* **45**: 31–73

Black D, Morris J, Smith C, Townsend P 1980 *Inequalities in Health: Report of a Research Working Group*. Department of Health and Social Security, London

Blane D, Hart C L, Davey Smith G, Gillis C R, Hole D J, Hawthorne V M 1996 Association of cardiovascular disease risk factors with socioeconomic position during childhood and during adulthood. *British Medical Journal* **313**: 1434–8

Brunner E 1997 Stress and the biology of inequality. *British Medical Journal* **314**(7092): 1472–6

Davey Smith G, Neaton J D, Wentworth D, Stamler R, Stamler J 1996 Socioeconomic differentials in mortality risk among men screened for the Multiple Risk Factor Intervention Trial: 1. White men. *American Journal of Public Health* **86**: 486–96

Ericson A, Eriksson M, Westerholm P, Zetterstrom R 1984 Pregnancy outcome and social indicators in Sweden. *Acta Paediatrica Scandinavica* **73**(1): 69–74

Guralnik J M, Land K C, Blazer D, Fillenbaum G G, Branch L G 1993 Educational status and active life expectancy among older blacks and whites [see comments]. *New England Journal of Medicine* **392**: 110–16

Helmert U, Herman B, Joeckel K H, Greiser E, Madans J 1989 Social class and risk factors for coronary heart diseasae in the Federal Republic of Germany. Results of the baseline survey

of the German Cardiovascular Prevention Study (GCP). *Journal of Epidemiology and Community Health* **43**(1): 37–42
Kaplan G A, Keil J E 1993 Socioeconomic factors and cardiovascular disease: a review of the literature. *Circulation* **88**(4): 1973–98
Kaplan G A, Pamuk E, Lynch J W, Cohen R D, Balfour J L 1996 Inequality in income and mortality in the United States: analysis of mortality and potential pathways. *British Medical Journal* **312**: 999–1003
Krieger N, Williams D R, Moss N 1997 Measuring social class in U.S. public health research: concepts, methodologies, and guidelines. *Annual Review of Public Health* **13**: 341–78
Kuh D, Ben-Schlomo Y 1997 *A Life Course Approach to Chronic Disease Epidemiology*. Oxford University Press, Oxford, UK
Kunst A E, Groenhof F, Mackenbach J P, Health E W 1998 Occupational class and cause specific mortality in middle aged men in 11 European countries: comparison of population-based studies. EU Working Group on Socioeconomic Inequalities in Health. *British Medical Journal* **316**(7145): 1636–42
Lynch J W, Kaplan G A, Cohen R D, Kauhanen J, Wilson T W, Smith N L, Salonen J T 1994 Childhood and adult socioeconomic status as predictors of mortality in Finland. *Lancet* **343**: 524–7
Lynch J W, Kaplan G A, Salonen R, Salonen J T 1997 Socioeconomic status and progression of carotid atherosclerosis: prospective evidence from the Kuopio Ischemic Heart Disease Risk Factor Study. *Arteriosclerosis, Thrombosis and Vascular Biology* **17**: 513–9
Lynch J W, Davey Smith G, Kaplan G A, House J S 2000 Income inequality and mortality: importance to health of individual income, psychosocial environment, or material conditions. *British Medical Journal* **320**: 1200–4
Macintyre S, MacIver S, Sooman A 1993 Area, class and health: should we be focusing on places or people? *Journal of Social Policy* **22**: 213–34
Marmot M G, Rose G, Shipley M, Hamilton P J S 1978 Employment grade and coronary heart disease in British civil servants. *Journal of Epidemiology and Community Health* **32**: 244–9
McEwen B S 1998 Protective and damaging effects of stress mediators. *New England Journal of Medicine* **338**(3): 171–9
Morgenstern H 1980 The changing association between social status and coronary heart disease in a rural population. *Social Science and Medicine [Medical Psychology and Medical Sociology]* **14A**(3): 191–201
Pamuk E, Makuc D, Heck K, Reuben C, Lochner K 1998 *Socioeconomic Status and Health Chartbook. Health, United States. 1998*. National Center for Health Statistics, Hyattsville, MD
Pappas G, Queen S, Hadden W, Fisher G 1993 The increasing disparity in mortality between socioeconomic groups in the United States, 1960 and 1986 (published erratum appears in *New England Journal of Medicine* **329**(15)). *New England Journal of Medicine* **329**(2): 103–9
Rather L J 1988 Report on the typhus epidemic in Upper Silesia. In: Rather L J (ed.) *Rudolph Virchow: Collected Essays on Public Health and Epidemiology*. Science History, Canton, MA, Vol. 1, pp. 205–20
Rosen G 1993 *A History of Public Health*, expanded edn. Johns Hopkins University Press, Baltimore, MD
Susser M, Watson W, Hopper K 1985 *Sociology in Medicine*. Oxford University Press, New York
Wilkinson R G 1996 *Unhealthy Societies—The Afflictions of Inequality*. Routledge, New York
Wolfson M, Kaplan G A, Lynch J W, Ross N, Backlund E 1999 Relation between income inequality and mortality: empirical demonstration. *British Medical Journal* **319**: 953–7
Wolfson M, Rowe G, Gentleman J F, Tomiak M 1993 Career earnings and death: a longitudinal analysis of older Canadian men. *Journal of Gerontology, Social Sciences* **48**: S167–79

G. A. Kaplan

Socioeconomic Status and Health Care

Although many factors determine the availability and use of health care, there is substantial evidence across countries in that markers of socioeconomic status (SES), including income, education, and occupational status, are significantly associated with health care utilization. This includes health insurance coverage and the use of personal health services such as visits to primary care providers and specialists, hospitalizations, dental care, vision care, mental health services, prescription-drug use, and home health care. Although there is ample evidence that people of higher socioeconomic position use more health care services and tend to receive higher-quality care, the reasons behind these disparities are not fully understood. Below, the relationship between SES and health services utilization is described, and explanations for this relationship are summarized.

1. The Relationship between SES and Health Care Utilization

Numerous studies have found a positive graded relationship between SES and most types of health-services utilization, measured by whether or not a service type was used in the past year, or as the mean number of encounters/visits for the service during a specified time period (National Center for Health Statistics 1998, Fiscella et al. 2000, Whitehead 1999). Although there are some exceptions to these general findings, the bulk of the empirical findings to date suggest that the association is positive; that is, higher SES is associated with a higher rate of health-care utilization (see Table 1 for some examples from the US). Higher SES is also associated with lower rates of untreated disease and unmet medical needs. That SES has a positive impact on health care use has been observed for both men and women, and for children and adults.

The relationship between SES and health care is especially apparent in countries without universal health insurance, where there is a strong relationship between SES and health insurance status (Bongers et al. 1997, Whitehead 1999). Those with lower levels of

Table 1
Health care utilization by poverty status for selected measures, United States 1994–1995. Data from National Center for Health Statistics (1998)

Health service	Poor	Near poor	Middle income	High income
No health insurance				
Children (< 18 years)	22.0 percent	22.8 percent	8.6 percent	4.2 percent
Adults (18–64 years)	44.1 percent	35.3 percent	13.8 percent	6.1 percent
No physician contact in the past year				
Children (< 6 years)	11.2 percent	10.1 percent	6.9 percent	4.1 percent
Adults (18–64 years)	21.2 percent	19.1 percent	14.9 percent	11.4 percent
Unmet need for health care				
Adults 18–64	32.5 percent	29.0 percent	16.0 percent	7.0 percent
Number of physician contacts in past year, age-adjusted by health status and gender	Poor	Near poor	Middle/high income	
Good to excellent health				
Males	3.9	3.8	4.6	
Females	5.4	5.0	5.8	
Fair or poor health				
Males	13.7	14.7	16.3	
Females	16.2	17.3	22.4	

income and education are more likely to be uninsured, or to have insurance coverage that only covers catastrophic health events. In turn, those who are uninsured are less likely to use health services, and to delay or forego care for serious medical symptoms (Fiscella et al. 2000, Brown et al. 1998). Importantly, SES is also related to health care use in countries with universal health insurance (Black et al. 1982; Whitehead 1999, Kephart et al. 1998). Thus, even in countries with a long history of universal coverage, SES disparities in health-care utilization are strong and apparent.

While the association between SES and health-care utilization is positive and graded in most studies, other patterns have been documented as well. An inverse association between SES and use of physician services was noted in Canada, along with the hypothesis that excess use by those of lower income and education levels is attributable to SES differences in health status (Kephart et al. 1998). In the US, those without health insurance, those living in low-income areas, and racial/ethnic minorities are significantly more likely to receive potentially-avoidable hospitalizations and medical procedures (Pappas et al. 1997, Fiscella et al. 2000).

In summary, people from lower economic and educational strata tend to have fewer physician visits for both primary and specialty care, and to use fewer mental health, dental, vision and preventive services. They also are less likely to have a usual source of care, and have significantly higher rates of unmet health needs. In some analyses, lower SES is associated with greater physician and hospital use (especially for avoidable causes of injury or illness), although it appears that these findings are primarily driven by higher rates of poor health status or medical needs in disadvantaged populations.

SES and its proxies are not the only factors related to health care use. Utilization is also patterned by gender, geography, age, culture, and other factors. While it is difficult to sort out the independent effects of SES and race/ethnicity on health care, it appears that minority status has an important independent effect. In the US, certain racial and ethnic minorities (most notably, African Americans, Native Americans and Hispanics) have been found to have lower rates of utilization in general, to receive less aggressive care for certain diseases, and to report less satisfaction with, and more experiences with discrimination regarding, health care (Fiscella et al. 2000). Even so, SES appears to be a major social predictor of health-care use. The relationship is strong across different types of health services, and is robust across a variety of study designs from a number of different countries.

2. *Explanations for the Relationship between SES and Health Care Utilization*

While the evidence for a relationship between SES and health care use is strong and compelling, there are many methodological challenges in measuring this association and its causes (Brown et al. 1998). First, the relationship is confounded by health status, which

is related to both SES and the need for services. Thus, it is important to measure health-services utilization within the context of medical need. Second, it is difficult to have all necessary variables in a single population-based dataset, including valid and reliable measures of SES, health insurance, medical need, and other confounding factors. Third, there are many different types of health services, and relationships are not uniform in magnitude or patterns across these different types. Fourth, simple measures of utilization do not provide a very deep understanding of differences that may exist in terms of the quality, content, or continuity of care.

Despite the methodological challenges for this type of research, several factors have been investigated to explain why people of lower SES generally use fewer health services than those of higher social position, and why their level of use does not appear to match their need. Dutton (1978) reported that explanations for this phenomenon generally fall into three categories: (a) explanations that focus on financial access or affordability differences (e.g., the poor use fewer services because they have fewer financial means); (b) explanations that focus on knowledge and attitudinal differences regarding care-seeking (e.g., the poor use fewer services because they value and/or understand the benefits of health care less than other people); and (c) explanations that focus on health care system factors (e.g., the poor use fewer services because of how the health care system is organized or responds to them). This set of explanations still reflects the primary reasons offered for the strong association between SES and health care utilization.

2.1 *Financial or Affordability Explanations*

Since being uninsured is strongly correlated with income and education, it is often assumed that the strong relationship between SES and health-care use operates primarily through the more proximate variable of health insurance. However, the conclusion of numerous studies is that health insurance explains only a small part of the socioeconomic differences in the use of most types of health services (Bongers et al. 1997). In addition, even among people with health insurance, lower levels of income and education are associated with lower overall health-care use (Fiscella et al. 2000).

Although access to health care is often equated with financial access, access is in fact multidimensional, and other aspects of access (such as the availability, convenience, and acceptability of providers and services) are of equal importance (Gold 1998). Also, as mentioned above, there is ample evidence from countries with universal health insurance that significant SES differences in utilization exist even within systems where there are few or no financial access barriers to services. Thus, SES differences in health care use are not fully explained by economic or financial factors.

2.2 *Knowledge, Attitudinal, and Cultural Explanations*

Another type of explanation for the lower use of health services among disadvantaged populations is that a lack of attention to health is part of a larger way of life or subculture related to poverty. Underuse of health care services is viewed as stemming in part from a lack of understanding regarding appropriate care-seeking behavior for specific symptoms or conditions. In addition, lower use of health services has been viewed as derivative of a larger set of attitudes that tend to accompany poverty and deprivation. This includes a 'crisis-oriented' approach to life, a tendency to overvalue immediate gains and rewards and discount longer-term outcomes, decreased self-reliance and self-efficacy, and increased skepticism regarding the potential benefits of medical care (Riessman 1994).

Although culture of poverty explanations for SES differences in health care use have been criticized as 'classist' and pejorative, this type of explanation continues to be used explicitly and implicitly by many researchers (Riessman 1994, Dutton 1978). Such explanations focus on social class differences in knowledge, attitudes, and psychosocial predispositions; and, as a result, programmatic and policy interventions that focus on these individual traits are not uncommon (Adler 1993, Reilly et al. 1998). While these individual characteristics have been shown to be related to health-care use, they are most often taken into account outside of the social and cultural contexts in which they were produced and the system context in which health care is provided. Thus, they provide a limited or incomplete explanation for observed SES disparities in health care use (Dutton 1978, Riessman 1994, Lantz et al. 1997).

2.3 *Health Care System Explanations*

Empirical evidence from a number of countries suggests that health care system factors, especially non-financial components of access, offer the greatest explanatory power regarding SES differences in utilization. It appears that deficits, inequities and inefficiencies in the health care system rather than patient economic or psychosocial characteristics are essential to understanding SES differences in the use of health services (Dutton 1978, Riessman 1994, Pappas et al. 1997). Research has shown that SES differences in health-care utilization are driven much more by structural elements of the health-care system than by individual factors such as health insurance or other

financial variables (Rosenbach et al. 1999). These system-level factors include cultural deterrents to care, the unwieldy nature of large medical bureaucracies, fragmentation in care, racism and other forms of provider bias, transportation barriers, and non-financial dimensions of access.

There are many ways in which the design of health-care systems leads people to feel estranged, dissatisfied, or distrustful, especially those people who are from marginalized populations because of social class, race, or ethnicity (Baker 1999). Quality of care and patient perceptions of care experiences are important in system explanations for SES disparities in health-services use. Not only is SES related to health-care utilization, it is also related to health-care quality (Fiscella et al. 2000). For example, the uninsured are more likely to receive substandard medical care and to suffer a medical injury as a result (Burstin et al. 1992).

While individual characteristics (such as financial means or attitudes regarding care) clearly play a role, it is important to view these individual characteristics within the broader context or system of how health-care services are organized, delivered, and financed. The policy implications of such findings are important; for they suggest that providing universal coverage or offering free health services to the poor is not a panacea for reducing social inequalities in health-care utilization. In the absence of significant changes in the ways in which health care is organized and delivered, improvements regarding financial access and personal beliefs will likely not do much to reduce SES disparities in health-services utilization.

3. *The Role of Health Care in Explaining SES Disparities in Health Status*

There are many explanations for the strong relationship between SES and health status, as outlined in the accompanying chapter on this topic (see *Socioeconomic Status and Health*). This includes explanations that focus on different material resources, of which health care is viewed as an important type (MacIntyre 1997). Research from a number of countries suggests that, for specific health conditions, the receipt of clinical preventive services and/or timely therapeutic care play an important role in health outcomes, and thus may account for some portion of socioeconomic disparities in health status. Therefore, it has been argued that efforts to increase the use of health services among the socioeconomically disadvantaged—primarily by reducing financial barriers to health care or health insurance—have great potential for reducing socioeconomic disparities in health status (Andrulis 1998, Fiscella et al. 2000).

Research at the end of the twentieth century, however, also suggests that the contribution of SES differences in health-care use to overall SES disparities in health status is likely to be small (Adler 1993, MacIntyre 1997). Evidence for this comes from a variety of sources, including—as described above—countries with universal health insurance. For example, the Black report demonstrated that socioeconomic inequalities in mortality in England and Wales increased rather than attenuated in the decades after the British National Health Services was established (Black et al. 1982). It has also been shown that, in developed countries, there is no relationship between health indicators (such as life expectancy) and the proportion of economic resources spent on health care (Roberts and House 2000). In addition, a cross-national study found that health-care expenditures did not explain variation in infant mortality rates above and beyond the contribution of socioeconomic resources (Kim and Moody 1992). While it is important not to overgeneralize from these ecological studies, a wide variety of research strongly suggests that differential access to and use of personal health services is but one of myriad factors contributing to social inequalities in health status (Roberts and House 2000, MacIntyre 1997).

With respect to developing countries, it is clear that economic development, growth in formal education, and public-health measures such as clean water and basic nutrition are of greater significance to population health than the delivery of health care. While the importance of personal health services (such as immunizations or oral rehydration therapy) in less-developed countries should not be underestimated, broad public health and economic development advances are believed to afford the most potential for making significant improvements in health.

This is not to suggest that health care has no impact on individual health or social disparities in health status. There are indeed some diseases and conditions for which health services can significantly reduce morbidity and mortality, and improve physical functioning and quality of life, as demonstrated in clinical trials and other types of studies (Andrulis 1998, Reilly et al. 1998, Fiscella 2000). The bulk of the knowledge generated to date, however, suggests that while reducing SES disparities in health care may help to reduce socioeconomic and racial disparities for some specific diseases or health issues, significant inequalities in health status would be likely to remain. Environmental, community, socioeconomic, and psychosocial exposures to both health risks and health-promoting resources would still exist, creating disparities in the incidence and prevalence of a number of health problems. Furthermore, many health conditions and problems are not amenable to medical interventions or to prevention practices delivered through health-care providers.

It is also important to emphasize that socioeconomic disparities in health care are not immutable (Fiscella et al. 2000). Thus, many researchers, clinicians, advocates and policymakers alike believe that efforts should continue to be made across and

within countries to reduce SES differentials in health-care access and utilization by focusing on the financial, attitudinal, and—most importantly—health system barriers that contribute to these significant disparities. The policy goal should be a health care delivery system in which all dimensions of access and the content and quality of care delivered are the same across social strata.

See also: Health Insurance: Economic and Risk Aspects; Socioeconomic Status and Health; Unemployment and Mental Health

Bibliography

Adler N E, Boyce W T, Chesney M A, Folkman S, Syme L 1993 Socioeconomic inequalities in health: No easy solution. *Journal of the American Medical Association* **269**: 3140–5

Andrulis D P 1998 Access to care is the centerpiece in the elimination of socioeconomic disparities in health. *Annals of Internal Medicine* **129**: 412–6

Baker R 1999 Minority distrust of medicine: a historical perspective. *The Mount Sinai Journal of Medicine* **66**: 212–22

Black D, Morris J N, Smith C, Townsend P 1982 *Inequalities in Health: The Black Report*. Penguin, New York

Bongers J M B, Van der Meer J B W, Van den Bos J, Mackenbach J P 1997 Socio-economic differences in general practitioner and outpatient specialist care in the Netherlands: A matter of health insurance? *Social Science and Medicine* **44**: 1161–8

Brown M E, Bindman A B, Lurie N 1998 Monitoring the consequences of uninsurance: A review of methodologies. *Medical Care Research and Review* **55**: 177–210

Burstin H R, Lipsitz S R, Brennan T A 1992 Socioeconomic status and risk for substandard medical care. *Journal of the American Medical Association* **268**: 2383–7

Dutton D B 1978 Explaining low use of health services by the poor: Costs, attitudes, or delivery systems? *American Sociological Review* **43**: 348–68

Fiscella K, Franks P, Gold M R, Clancy C M 2000 Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care. *Journal of the American Medical Association* **283**: 2579–84

Gold M 1998 Beyond coverage and supply: measuring access to healthcare in today's market. *Health Services Research* **33**: 625–52

Kephart G, Thomas V S, MacLean D R 1998 Socioeconomic differences in the use of physician services in Nova Scotia. *American Journal of Public Health* **88**: 800–3

Kim K K, Moody P M 1992 More resources better health? A cross-national perspective. *Social Science and Medicine* **34**: 837–42

Lantz P M, Weigers M E, House J S 1997 Education and income differentials in breast and cervical cancer screening. *Medical Care* **35**: 219–36

MacIntyre S 1997 The Black Report and beyond: What are the issues?*Social Science and Medicine* **44**: 723–45

National Center for Health Statistics 1998 *Health, United States, 1998 With Socioeconomic Status and Health Chartbook*. US Department of Health and Social Services, Hyattsville, Maryland

Pappas G, Hadden W C, Kozak L J, Fischer G F 1997 Potentially avoidable hospitalizations: Inequalities in rates between US socioeconomic groups. *American Journal of Public Health* **87**: 811–16

Reilly B M, Schiff G, Conway T 1998 Part II: Primary care for the medically underserved: Challenges and opportunities. *Disease-A-Month* **44**: 320–46

Riessman C K 1994 Improving the health experience of low income patients. In: Conrad P, Kern R (eds.) *The Sociology of Health and Illness*. St Martin's Press, New York, pp. 434–37

Roberts S A, House J S 2000 Socioeconomic inequalities in health: an enduring sociological problem. In: Bird C E, Conrad P, Fremont A M (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice Hall, Upper Saddle River, NJ, pp. 79–97

Rosenbach M L, Irvin C, Coulam R F 1999 Access for low-income children: Is health insurance enough? *Pediatrics* **103**: 1167–74

Whitehead M M 1999 Where do we stand? Research and policy issues concerning inequalities in health and in healthcare. *Acta Oncologica* **38**: 41–50

P. M. Lantz

Sociolinguistics

Sociolinguistics is the study of the relationship between language and society. The field of sociolinguistics is arguably the best known and most firmly established of the various hyphenated varieties of linguistics (e.g., psycholinguistics, neurolinguistics, sociolinguistics) that emerged during the second half of the twentieth century. Its roots can be located in the field research of American anthropological linguists such as Dell Hymes, who noted the interesting range of ways of talking and the variety of functions of talk among native American tribes whose languages they described, and of dialectologists, such as Bill Bright, who drew attention to the social bases of much of the linguistic diversity they documented in multilingual speech communities.

Although it is a well-established field, there is still some controversy about what counts as sociolinguistics. William Labov, a pioneering and very influential American social dialectologist, has argued since the 1960s that his research constitutes core linguistics, and contributes directly to the development of linguistic theory. In the 1970s, Peter Trudgill, a leading British social dialectologist, noted a distinction between research in social dialectology, which he argued constituted the core of sociolinguistics, and research in areas such as multilingualism and language shift, which Joshua Fishman had labeled 'the sociology of language,' an area Trudgill proposed to exclude from sociolinguistics proper. While that debate has effectively been resolved—the current leading

sociolinguistics journals include research in both these areas—discussion over the precise boundaries of the term sociolinguistics continues.

This article takes an inclusive approach. The sections which follow provide a brief description of the concerns of areas which uncontroversially constitute core sociolinguistics, namely the sociology of language, and social dialectology, as well as areas which some would regard as more peripheral, such as discourse analysis (DA) and stylistics, language awareness, and applied sociolinguistics.

1. *Multilingualism; Diglossia; Bilingualism*

Sociolinguistic research in the area of multilingualism is concerned with identifying the relationship between social factors and choice of language or 'code' in specific situations. In Zaire, for example, people often use five or six languages, or more accurately 'codes' or 'varieties,' in their everyday interactions. The terms 'code' and 'variety' refer to any set of linguistic forms which are socially patterned, including different accents, different linguistic styles, different dialects, and different languages which contrast with each other for social reasons. The linguistic repertoire of a typical Zairean teenager might include three varieties of Swahili (standard Zairean, local Swahili, and Indoubil, a teenage slang variety), and two varieties of a tribal language such as Shi (a formal and a casual style). How the young person draws on and makes use of this repertoire in interaction is a matter of prime sociolinguistic interest (Goyvaerts 1988).

Sociolinguistic research aims to establish the range of social factors relevant to code choice in such multilingual contexts, and their relative weight. The use of one code rather than another in any specific situation tends to reflect such factors as the relationship between speaker and addressee(s), the formality of the context, and the function of the interaction. In Paraguay, for example, the indigenous language, Guaraní, is used in interactions with friends in the pub, while Spanish is the appropriate language to use with your doctor and the bureaucrats (Rubin 1985). Code choice may also indicate more abstract influences such as orientation to an urban rather than a rural lifestyle. More recently, researchers have begun to analyze code choice as a more dynamic activity, examining ways in which speakers change the emphasis in an interaction through linguistic initiatives, and how code choice actively contributes to the construction of particular types of interaction (see *Code Switching*: *Linguistic*).

Diglossia is a term used to describe a particular kind of societal bilingualism. In his classic article on the topic, Ferguson (1959) defined diglossia as a situation where two varieties of the same language, a H(igh) variety and a L(ow) variety, are used in a community for complementary functions, a linguistic division of labor. The H variety is restricted to formal, religious, symbolic, and literary functions. The L variety is used for everyday conversations in the home and in all informal contexts. The term proved so useful it was extended to cover a much wider variety of bilingual and multilingual situations where different languages were associated with different domains or social contexts. The term 'polyglossia' was coined to take account of communities such as Singapore with more than one H language as well as several L varieties and even some M(id) varieties falling between the H and L in terms of prestige (Platt 1977). Thus, sociolinguists have steadily modified the scope of the diglossia concept to handle more complex patterns of societal multilingualism. Moreover, further research on the four classic cases Ferguson used for exemplification (German-speaking Switzerland, Haiti, Greece, and Arabic-speaking countries) suggests that even these do not fit the original framework as neatly as was once thought.

1.1 *Code-switching*

A related area of sociolinguistic research is the study of *code-switching*. Situational code switches have been defined as responses to changes in observable features of the context, such as the participants or the setting, and they typically occur at clearly demarcated discourse boundaries. Recent research has increasingly focused on interpreting the social meaning of conversational code-switching, sometimes called code-mixing, the much less predictable rapid code-switching which typically occurs among bilinguals in informal situations (e.g., Auer 1998). The development of theoretical models that can account for the linguistic constraints on switches has been one focus of this research. Another has been the social and pragmatic meanings of switches. Analyzing code-choice in Kenya, for instance, Myers-Scotton (1997) discusses the sophisticated affective messages conveyed through complex and rapid alternations between codes, often within a very short exchange. So, for example, through subtle switching of codes, rather than any explicit refusal, a store-keeper politely rejects his sister's attempts to obtain cut-price goods from him, thus preserving their relationship. This strand of research considers how switches actively contribute to constructing a particular type of interaction, and to the dynamics of identity construction. Code-switching is one means by which bilinguals and multilinguals make creative use of their linguistic repertoires to convey social meaning.

2. *Language Maintenance and Shift; Language Endangerment/Revival*

While diglossia is defined as a society-wide phenomenon, individual bilingualism is often a transitional phase for new immigrants in a society as they shift

from the ethnic language to the dominant language of the wider society. This process of language shift is typically complete in three generations. Fishman's research dominates the area of language maintenance and shift, with studies of patterns of language use among ethnic minorities in the USA dating back to the early 1960s. A range of theoretical models has been developed to account for the relative survival rates of different ethnic languages in different countries, including measures of 'ethnolinguistic vitality' (Giles 1977), and a scale describing the steps involved in taking a language from near language death through to language revival (Fishman 2000).

A language which is spoken only in a restricted geographical area and which is under threat from a colonial or world language is typically regarded as endangered. The relative chances of survival of different *endangered languages* have been the topic of a good deal of sociolinguistic analysis. Research on the numbers of users, social contexts of use, and linguistic attitudes has been undertaken on languages as diverse as Welsh, Cree, Maori, Dyirbal, and Basque, with some disagreement between experts on the crucial conditions for preventing language death.

3. *Language Contact; Pidgins and Creoles*

While the development of bilingualism, or a shift to the majority language, are possible responses to language contact, another common development is the emergence of a pidgin language. Pidgins typically develop in situations where a *lingua franca* is needed for communication between groups who speak different languages, for example, for trade purposes, or between workers from different tribes on a crop plantation. They are structurally very simple and functionally extremely restricted. They are always ancillary languages developed for a specific and restricted purpose.

Creoles develop from pidgins when they begin to be used as first languages and therefore expand functionally and structurally into adequate means of communication for the group using them. Typically they draw on the vocabulary of the language of the culturally dominant group, while their grammar is based on local vernacular languages which contribute the substrate component. A number of official and standard languages such as Swahili, Afrikaans, and Tok Pisin had creole origins.

Current debates among creolists center mainly around the issue of the origins and development of creole languages. Do they develop according to universal linguistic principles and according to universal linguistic constraints, or are the remarkable similarities between creoles due to the fact that they share sources of input from European languages such as English and Portuguese? Another current area of interest is the decreolization process (whereby creoles gradually develop in the direction of the standard variety) which is occurring in places such as Guyana (Romaine 1994).

4. *Social Dialectology*

Social dialectologists are concerned with tracing the relationships between linguistic variation and social factors. The pioneering research of William Labov determined the direction of subsequent research in the area for several decades, both in terms of the research questions he considered worth addressing, and in relation to the innovative methodology and quantitative ('variationist') analytical techniques he developed. His influential survey of the New York East Side (Labov 1972a) demonstrated that linguistic variation was not random or 'free' as previous dialectologists had tended to assume, but rather patterned systematically according to factors such as the social class, gender, age, and ethnicity of the speakers, and the formality of the style of speech they were using (see *Language and Social Class*; *Language and Ethnicity*).

The basic patterns which Labov identified in New York were replicated in many other urban speech communities including cities in Canada, England, Australia, and New Zealand. In Europe too, variationist approaches gradually displaced those of traditional regional dialectologists, and were adopted in urban dialectology research with similar results. Researchers found, for instance, that stable variables, which typically included grammatical variables, such as multiple negation, and phonological variables, such as final *-ing* (e.g., *he don't know nothin'*) in English, tended to divide the community clearly into two major social groups. Linguistic variables involved in change, however, were typically more finely stratified with a gradient pattern reflecting a continuum from the lowest to the highest social groups. In the great majority of surveys, women lead sound change whether towards the standard or towards the vernacular variety. In New Zealand, for example, women are leading the rapid spread of the high rising terminal, as well as the merging of the vowels in words such as EAR and AIR. Recent variationist studies indicate that in specific speech communities variables such as class, ethnicity, social role, and style often interact to modify such general patterns.

Labov formalized the idea that linguistic change could be studied in progress; he analyzed 'apparent time' data demonstrating the spread of postvocalic (r) in words like *car* and *card* through different generations of the New York speech community. This approach to the study of language change has since been used by sociolinguists in speech communities worldwide.

Labov also used quantitative methods to collect people's subjective evaluations of different social accents he played to them, a method which has been

widely used since in language attitude research. He demonstrated that, despite wide variations in their linguistic production, there was a startling uniformity of evaluative norms and attitudes to language among members of the same speech community. The New York accent, for example, was universally disparaged, a pattern replicated in many other urban speech communities. On the basis of this analysis, Labov suggested that the most useful defining characteristic of a speech community was precisely such shared linguistics norms. Others have suggested shared rules of use, density of patterns of interaction, or a shared pattern of variation. More recently, some sociolinguists have begun to explore the concept of the 'community of practice' as a useful concept in sociolinguistic research which moves the focus to the ways in which communities are constructed through shared goals, activities, and verbal resources (see Holmes and Meyerhoff 1999).

Labov set the methodological standard in social dialectology, influencing all subsequent research. He used an extensive interview schedule which included word lists, reading passages, questions skilfully designed to elicit casual speech, and innovative attitude elicitation techniques. Later refinements include the use of a network approach by Lesley Milroy and James Milroy in the study of working class Belfast. Instead of using social survey results and consequently interviewing strangers, they used networking (the 'friend of a friend' approach) to contact potential contributors. They also operationalized the network concept in terms of density and multiplexity, demonstrating that those who were most closely integrated into the local community used more vernacular forms (Milroy and Margrain 1980).

Current variationist research can be found in the journal *Language Variation and Change* with material ranging from the finest vowel and consonant distinctions to the broadest features of interactional discourse. For instance, the use of quotative *be like* (as in *so she was like you aren't really going out wearing that thing and he was like why not*) provides one example of an interesting, relatively new discourse marker which is being well documented as it spreads from American English into other varieties (e.g., Tagliamonti and Hudson 1999).

5. *Social Constructionist Approaches*

Another important development over the last decade has been the influence of social constructionism on sociolinguistic research. Rather than regarding social categories, such as social class, ethnicity, and gender, as primary, with linguistic behavior reflecting membership, a social constructionist approach shifts the emphasis to language as a dynamic resource used to construct particular aspects of social identity at different points in an interaction. Social categories are not fixed but are subject to constant change; talk itself actively creates different styles and constructs different social contexts and social identities as it proceeds. So, for example, at certain times during a romantic dinner, a woman may select linguistic forms which contribute to the construction of a more 'feminine' identity, while in a Board meeting she is chairing she may linguistically construct a statusful and powerful identity. Interacting with her children or perhaps with certain subordinates, she may linguistically construct a 'maternal' identity. Even within one interaction she may use language to perform a range of social identities as she responds to the development of the discourse in context.

6. *Discourse and Style*

6.1 *Style*

The influence of social constructionism is also evident in recent research on the sociolinguistics of style. As mentioned above, Labov elicited a range of styles within the interview context by varying the interviewee's task. Subsequent research focused on the influence of the addressee (e.g., Giles and Smith 1979) and/or the 'audience' (Bell 1984). Speech Accommodation Theory, later relabeled Communication Accommodation Theory (CAT), focuses on the mutual influence of participants on a wide range of aspects of communication: so, for example, people may 'accommodate' to each other's nonverbal behavior, speech volume, rate of speaking, prosodic patterns, vowel realizations, discourse markers, and even language. Collecting social dialect data, for instance, Trudgill (1988) found he converged towards his interviewees' levels of final glottal stops for /t/ in words such as *not*. Similarly in New Zealand, convergence in levels of the use of the discourse marker *eh* are common (Meyerhoff 1994). Divergence also occurs when interlocuters wish to differentiate themselves. Giles and Smith (1979) describe, for example, a switch to Welsh as one response to outsiders' disapproval of efforts to maintain the language, and, in Australia, Eisikovits (1989) explains adolescent boys' use of vernacular forms in an interview by suggesting they are deliberately diverging from the standard speech of their middle-class female interviewer. More recently there has been some exploration of Rampton's concept of 'crossing,' where speakers apparently use out-group linguistic styles for specific stylistic effects, or code switch into varieties not generally associated with their group in order to challenge or problematize group boundaries (Rampton 1999).

Currently, then, many sociolinguists argue that style is not so much a matter of self-monitoring (as Labov suggested) or of responding to others as CAT theorists argue, but rather a matter of the way an individual chooses to present herself in everyday interaction. Style emerges from 'practice,' the individual's ongoing performance of her complex of social

identities in particular sociocultural contexts. Eckert's analysis of American adolescents' stylistic 'performance' in the school context, for instance, demonstrates that style is not just a matter of speech, but also involves 'dress, attitude and action' (2000, 211ff). Their choice of linguistic variables, however, affects every utterance. They are constantly manifesting a particular style (e.g., 'flamboyant' or conservative'), just as they are ongoingly involved in the performance of their particular social identity in the school (e.g., as 'jock' or 'burn-out').

Few would argue at the inclusion of such research within sociolinguistics. However, turning to the closely related area of pragmatics, the borderlines become much less clear.

6.2 Pragmatics

Pragmatics is concerned with the analysis of language in use. At the more linguistic end of pragmatics, researchers analyze how people disambiguate utterances such as *I watched them standing up*. Social context becomes increasingly relevant in analyzing the complex ways in which meaning is conveyed through such an apparently simple utterance as *you didn't get a paper then*? which can be interpreted as a request for information, a complaint, or even a rebuke, depending on the social relationship between participants. Similarly, *I can hear someone talking* can convey very different meanings depending on the relevant social context and participant relationships: consider, for instance, a classroom vs. police officers investigating an 'empty' house.

Closely related is sociopragmatic research on politeness, with analysis of the ways in which people express rapport and solidarity (positive politeness), as well as respect and social distance (negative politeness) (see Brown and Levinson 1987, Thomas 1995). Linguistic politeness devices include choice of pronoun (e.g., *tu/vous*, *du/Sie etc*), endearment terms (e.g., *love*, *honey*, *dear*), hedges (e.g., tag questions, epistemic modal verbs, modal particles, such as *perhaps*, *possibly*), pragmatic particles (e.g., *you know*, *of course*, *sort of*, *I mean*), and many more, some of which are very subtle.

Sociolinguists examine the patterning of such pragmatic devices in the usage of different social groups. Gender differences have been identified, for instance (e.g., Coates 1996), as well as ethnic and class differences (e.g., Goodwin 1980). More recently, researchers have examined the ways in which people draw on these devices in constructing different aspects of their social identity. Such devices are important resources in constructing ethnic identity, for example, or a 'feminine,' compliant, accommodating identity (e.g., Holmes 1997).

Sociopragmatic research has also been developed by applied sociolinguists who have examined cross-cultural differences in the way politeness is expressed, and in the realizations of such speech functions as apologies, complaints, compliments, directives, refusals, and disagreements (e.g., Gass and Neu 1996). The applied value of such work in second language teaching and learning is obvious.

6.3 Discourse Analysis

Not all sociolinguists would regard DA as appropriately included within sociolinguistics. Introductory courses and textbooks vary in whether they have sections introducing DA. Just a few approaches to DA are outlined here. In its broadest sense DA includes any analysis of spoken or written discourse. In practice, sociolinguists who adopt a DA approach to their materials generally do so in order to identify the ways in which social phenomena emerge from or are instantiated in spoken interaction. The way an oral narrative is constructed, for instance, may provide evidence of the salience for the interlocuters of a particular social or ethnic or gender identity (e.g., Labov 1972b, Cheshire 2000). Discourse markers, such as end tags (*and stuff*, *and everything*), can be used in ways which reflect or express social group membership.

Conversational analysis (CA) is one distinctive approach to the analysis of discourse based on a model of communication as jointly organized activity (Sacks 1984): like dancing, or joint musical performance, it rejects the typical linguistic model of communication as sending and receiving messages. CA is concerned with how participants work together to jointly achieve conversational closings, storytelling, disputes, medical diagnoses, the mutually dependent roles of interviewer and interviewee, and so on. CA focuses on how interaction unfolds as a sequence of actions by different participants, with the significance of an utterance or gesture highly dependent on its position in a sequence, as well as being jointly negotiated.

Others put more emphasis on the social context in which interaction takes place, and the relevant backgrounds, knowledge, and experience of the participants. This more ethnographic approach, generally identified as interactional sociolinguistics, is associated with John Gumperz and his associates. One example of recent work in this area is the exploration of workplace interactions between managers and workers, with a particular focus on areas of miscommunication and potential misunderstanding (e.g., Roberts et al. 1991).

Another approach, generally known as critical discourse analysis (CDA), has developed largely within the European tradition. CDA is distinguished by its critical focus, its broad scope, and its 'overtly political agenda' (Kress 1990, pp. 84–5). CD analysts aim to reveal connections between language, power, and ideology, and to describe the way power and dominance are produced and reproduced in the discourse structures of generally unremarkable interactions.

7. Language, Culture, and Ideology

Another important strand of sociolinguistic research which can be traced to the influence of American anthropological linguists is the quest for a solution to the conundrum of the relationship between language, culture, and thought. Edward Sapir and his pupil Benjamin Lee Whorf developed the hypothesis that language influences thought rather than the reverse. The strong form of the Sapir–Whorf hypothesis claims that people from different cultures think differently *because of* differences in their languages. So, native speakers of Hopi perceive reality differently from native speakers of English because they use different languages, Whorf claimed. Few sociolinguists would accept such a strong claim, but most accept the weaker claim of linguistic relativity, that language influences perceptions, thought, and, at least potentially, behavior.

More recently, sociolinguists such as Fairclough (1995) and applied linguists such as Pennycook (1994) have in quite different ways developed this approach. Fairclough's research uses a CDA approach to identify and expose the ways in which ideology and power are constantly instantiated and enacted in the familiar discourse of the media and of everyday interactions. Others have examined such issues as the continuing and increasingly subtle ways in which sexism, in the form of the derogation of women, or racist suppression or misrepresentation of ethnic minorities, are sustained through the ways language is used in the media or in public documents. Pennycook's controversial research also emphasizes critical awareness; he argues that the imperialist expansion of the English language has been at the cost of local languages in many countries. Not surprisingly, his views have been contested by those who regard the development of a wide variety of Englishes around the world as evidence that English no longer belongs to the English—nor even to the Americans.

8. Applied Sociolinguistics

The implications and applications of sociolinguistic research have always concerned sociolinguistic researchers. Indeed, many began their research efforts in an attempt to answer a 'real world' question such as

(a) does language play a role in accounting for differences in the educational performance of different sociocultural groups?

(b) what role can education play in the maintenance of a minority language?

(c) what is the role of the sociolinguist in the area of language policy?

(d) how can sociolinguistic information improve the workings of the legal system?

(e) what sociolinguistic competence is required by immigrants in the workplace?

As the breadth of these examples suggests, this is an extensive area of research and no more than a few illustrative examples can be provided.

Sociolinguistic research has always had important social repercussions in the area of education. The early work of the late British sociologist Basil Bernstein (1971, [1973]) was hugely influential, especially in Germany, where deficit theories of various sorts dominated some areas of sociolinguistic research for decades. Bernstein suggested that a child's social class determined their access to linguistic 'codes,' with potential cognitive consequences. His term 'restricted code' and the linguistically naive assumptions underlying some of his claims were challenged by social dialectologists, leading to the extended debates of the 'deficit–difference' controversy (e.g., Labov 1969. Its most recent manifestations range from arguments about the place of Ebonics or African American Vernacular English in the American school curriculum to debates on the linguistic rights of ethnic minorities such as Aboriginal children in Australia, British Black children in England, and the various Gastarbeiter communities in European countries.

Another area where sociolinguistic advice is often sought is in the areas of language policy and language planning. Sociolinguistic descriptions provide crucial information for planning educational policy in multilingual contexts. In Vanuatu, for example, with more than 80 vernacular languages spoken by no more than 200,000 people in total, the government's proposed policy of education in the vernacular requires extensive preparatory sociolinguistic and linguistic research (Crowley 2000). Similarly, attempts to revive a 'dying' language must start from an accurate sociolinguistic description of current numbers of users and contexts of use. Bilingual education is more likely to succeed if the sociolinguistic context in which the program is to be developed is taken into account in selecting an appropriate model. Accurate information about the way people talk can usefully counter misleading gender stereotypes which may restrict people's job opportunities. Plans for sexist language reform also benefit from accurate information about the strength of attitudes in the relevant speech community.

Forensic linguistics provides a final example where applied sociolinguistics has made a contribution. Sociolinguists have demonstrated ways in which the legal system discriminates against those who use nonstandard dialects, and those whose sociocultural rules for language use differ from those of the dominant group. In communities where Aboriginal English is used, for instance, direct questions are an inappropriate way of eliciting information (Eades 1996). Culturally conditioned responses to a lawyer's direct questions (such as silence or a formulaic *I don't remember*) may severely disadvantage Aboriginal defendants. The assumption that the norms of the Western legal system are universal regularly result in misinterpretation of the linguistic behavior of those

from different sociocultural backgrounds. Many applied socolinguists orient their research to highlighting such areas of potential injustice.

9. Conclusion

This article has necessarily been selective. The aim has been to provide a broad overview within which the range of activities of current sociolinguists can be located. Many of the topics are discussed in more detail in separate articles. Recent introductory textbooks (e.g., Romaine 1994, Wardhaugh 1998, Holmes 2001) and collections of readings in sociolinguistics (e.g., Coupland and Jaworksi 1997) provide further resources for those wishing to pursue the subject further.

How will sociolinguistics develop in the twenty-first century? The search for robust and explanatory sociolinguistic theories currently occupies many sociolinguists and seems likely to continue for some decades. The steady integration of macrosociolinguistic ('sociology of language') approaches with variationist and microsociolinguistic (interactional/DA) approaches seems a likely direction of development. Increasingly researchers using different theoretical approaches or from different disciplinary backgrounds are attempting to communicate with each other in theory development. Finally, it is safe to predict that sound descriptive sociolinguistic research will continue to provide the crucial underpinning for language policy developments in areas such as minority groups, immigration, education, and second language teaching.

See also: Language and Gender; Language and Social Class; Language and Social Inferences; Language and Society: Cultural Concerns; Language Policy: Linguistic Perspectives; Language Policy: Public Policy Perspectives; Pidgin and Creole Languages; Politeness and Language

Bibliography

Auer P 1998 *Code-Switching in Conversation*. Routledge, London

Bell A 1984 Language style as audience design. *Language in Society*. **13**(2): 145–204

Bernstein B 1971 [1973] *Class, Codes and Control, Vols. 1 and 2: Theoretical Studies Towards a Sociology of Language*. Routledge and Kegan Paul, London

Brown P, Levinson P 1987 *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK

Cheshire J 2000 The telling or the tale? Narratives and gender in adolescent friendship networks. *Journal of Sociolinguistics* **4**(2): 234–62

Coates J 1996 *Women Talk*. Blackwell, Oxford, UK

Coupland N, Jaworski A (eds.) 1997 *Sociolinguistics; A Reader and Coursebook*. Macmillan, London

Crowley T 2000 *Vernaculars in Education in Vanuatu*. Report to World Bank and Vanuatu Ministry of Education, Youth and Sport

Eades D 1996 Legal recognition of cultural differences in communication: The case of Robyn Kina. *Language and Communication* **16**(3): 215–27

Eckert P 2000 *Linguistic Variation as Social Practice*. Blackwell, Oxford, UK

Eisikovits E 1989 Girl-talk/boy-talk: Sex differences in adolescent speech. In: Collins P, Blair D (eds.) *Australian English*. University of Queensland Press, St Lucia, Qld, pp. 35–54

Fairclough N L 1995 *Critical Discourse Analysis: Papers in the Critical Study of Language*. Longman, London

Ferguson C A 1959 Diglossia. *Word* **15**: 325–40

Fishman J A (ed.) 2001 *Can Threatened Languages be Saved? Reversing Language Shift, Revisited: A 21st Century Perspective*. Multilingual Matters, Clevedon, UK

Gass S, Neu J (eds.) 1996 *Speech Acts Across Cultures: Challenges to Communication in a Second Language*. Mouton de Gruyter, Berlin

Giles H (ed.) 1977 *Language, Ethnicity and Intergroup Relations*. Academic Press, London and New York

Giles H, Smith P 1979 Accommodation theory: Optimal levels of convergence. In: Howard G, St. Clair R (eds.) *Language and Social Psychology*. Blackwell, Oxford, UK, pp. 45–65

Gwin M H 1980 Directive-response speech sequences in boys and girls talk activities. In: McConnell-Ginet S, Borker R, Furman N (eds.) *Women and Language in Literature and Society*. Praeger, New York, pp. 157–73

Goyvaerts D L 1988 Indoubil: A Swahili hybrid in Bukavu. *Language in Society*. **17**(2): 231–42

Holmes J 1997 Women, language and identity. *Journal of Sociolinguistics* **2**(1): 195–223

Holmes J 2001 *Introduction to Sociolinguistics*, 2nd edn. Longman, London

Holmes J, Meyerhoff M 1999 The community of practice: theories and methodologies in language and gender research. *Language in Society: Special Issue: Communities of Practice in Language and Gender Research* **28**(2): 173–83

Kress G 1990 Critical discourse analysis. *Annual Review of Applied Linguistics* **11**: 84–99

Labov W 1969 The logic of non-standard English. *Georgetown Monographs on Language and Linguistics*. Georgetown University Press, Georgetown, Vol. 22

Labov W 1972a *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, PA

Labov W 1972b The transformation of experience in narrative syntax. In: Labov W (ed.) *Language in the Inner City*. University of Pennsylvania Press, Philadelphia, PA, pp. 354–96

Meyerhoff M 1994 Sounds pretty ethnic eh? A pragmatic particle in New Zealand English. *Language in Society* **23**(3): 367–89

Milroy L, Margrain S 1980 Vernacular language loyalty and social network. *Language in Society* **3**: 43–70

Myers-Scotton C 1997 *Duelling Languages: Grammatical Structure in Code-Switching*. Clarendon Press, Oxford, UK

Pennycook A 1994 *The Cultural Politics of English as an International Language*. Longman, London

Platt J 1977 A model for polyglossia and multilingualism (with special reference to Singapore and Malaysia). *Language in Society* **6**(3): 361–78

Rampton B 1999 *Journal of Sociolinguistics* **4**(3): 421–585
Roberts C, Jupp T, Davies E 1991 *Language and Discrimination*. Longman, London
Romaine S 1994 *Language in Society*. Oxford University Press, Oxford, UK
Rubin J 1985 The special relation of Guaraní and Spanish in Paraguay. In: Wolfson N, Manes J (eds.) *Language of Inequality*. Mouton, Berlin, pp. 111–20
Sacks H 1984 Notes on methodology. In: Maxwell Atkinson J, Heritage J (eds.) *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, Cambridge, UK, pp. 21–7
Tagliamonte S, Hudson R 1999 Be *like* et al. beyond America: The quotative system in British and Canadian youth. *Journal of Sociolinguistics* **3**(2): 147–72
Thomas J 1995 *Meaning in Interaction*. Longman, London
Trudgill P 1988 Norwich revisited: Recent linguistic changes in an English urban dialect. *English World-Wide* **9**(1): 33–49
Wardhaugh R 1998 *An Introduction to Sociolinguistics*, 3rd edn. Blackwell, Oxford, UK

J. Holmes

Sociology, Epistemology of

The word 'epistemology' will be taken here in its etymological acceptation, which is the one kept in the Romance languages, namely as philosophy of science. Like any other special branch of the philosophy of science, the epistemology of sociology may be construed as having three main components: ontology (or metaphysics), epistemology in the narrow sense (theory of knowledge), and methodology (or normative epistemology). This order is not arbitrary, for every research strategy depends upon the type of knowledge being sought, which in turn depends on the nature of the thing studied. Thus, if the object of study is concrete, like a photon or a social group, and if what is sought is objective knowledge of it, then the scientific method will be adopted. If, by contrast, the object of knowledge is assumed to be spiritual, and if things spiritual are deemed to be beyond the reach of science, then an alternative method will be preferred—such as revelation, intuition, interpretation (Verstehen), unchecked speculation, or narrative. In this article the ontology, epistemology and methodology of sociology will be glanced at.

1. *Ontology: Nature of the Social*

1.1 *The Idealism–Materialism Dilemma*

Most sociologists adopt the vulgar view that the human being is a composite of body and mind (or spirit or soul), and do not ask how these two components are kept together. Hence they do not characterize social facts as either material or spiritual, and do not worry about the place of sociology in the system of human knowledge: they are dualists. But some sociologists do worry about these questions, and argue for or against one of two monistic worldviews: idealism or materialism. Idealism is the view that everything is either ideal or ultimately ruled by ideas, whereas materialism holds that all existents are material or concrete.

In social studies, the idealism/materialism cleavage appears notably in the place and relative weight assigned to the so-called ideal (cultural and political) superstructure vis à vis the material (biological and economic) infrastructure. This very distinction betrays a dualistic stand, even if the pre-eminence or priority of either 'structure' (system) is postulated.

Consistent idealists will assert that everything social is spiritual (or cultural), whence sociology is a *Geisteswissenschaft* (science of the spiritual). This is the view of hermeneutics, symbolic interactionism, phenomenological sociology, and ethnomethodology. By contrast, materialists assert that persons and social groups are material, though not necessarily confined to the physical or even the biological level. And, unlike idealists, they claim that all cultural and political activities are influenced or even determined by environment, reproduction, and material production. While it is undeniable that the reproductive patterns and means of subsistence condition culture and politics, it is no less true that the latter in turn help shape the former. In a systemic perspective, changes in the environment and in any of the subsystems of society (the biological, economic, political and cultural) may initiate processes altering any of the others. People may work or fight, live or die, for King and country, no less than for food and family. There is no 'last analysis.'

Marxism is the oldest, and still the most influential, of the materialist schools in social science. However, given its proposed division between the material infrastructure and the ideal superstructure, it is questionable that Marxism is consistent. Moreover, philosophical Marxism is nowadays split into a large number of different schools, neither of which is flourishing. By contrast, some of Marx's contributions to social science, particularly his emphasis on production, on the economic roots of the power elites, and on the economic causes of many an international conflict, have survived. Even some of Weber's studies are materialist, particularly those of the congruence between Hinduism and the Indian caste system, and on the causes of the decline of slavery in ancient Rome. Likewise the *Annales* historical school owes much to Marx.

Sociobiology is the newest and most popular materialist school in social studies. Sociobiologists hold that everything social is ultimately biological, for having evolved as an adaptation to the natural environment. This view overlooks all the types of self-

destructive and antisocial behavior, from suicide to military aggression to environmental degradation to gullibility and ideological zealotry. And it ignores all social inventions (like formal education) and social revolutions (like the information revolution)—hence it is basically conservative. The sociobiological speculations on the origin of mental faculties are even more far-fetched and they lack empirical support. Still, something does remain of the attempt to reduce social science to biology, namely, the reminder that sociologists cannot afford to ignore human physiology and ecology. For example, they should not ignore that malnutrition stunts development and limits productivity; that plagues, foreign invasions and wars are bound to eliminate some genes and change the immune system; and that a few more centuries of uncontrolled environmental degradation is likely to make our planet uninhabitable.

1.2 The Individualism–Holism–Systemism Trilemma

There are essentially three consistent views on social structure: that it does not exist, that it is self-existing, and that it is a property of a system composed of individuals. The first, or individualism, identifies 'society' with 'crowd.' The second view, or holism, regards society as an organism, and the individual as wholly subordinated to it. The third, or systemism, views society as a system made up of individuals linked to one another by ties of different kinds and strengths. Individualism stresses agency at the expense of structure, and tends to explain action in terms of 'rational' (self-interested) decisions. Holism emphasizes collective traits, which it claims to be unexplainable, and it explains individuals by wholes. Lastly, systemism admits the novelty of systemic properties, as well as the possibility of explaining them in terms of interactions among individuals and systems.

Individualists rightly see in individual action the root of everything social. And, even if they acknowledge the importance of the social context or situation, they resist the very idea of a social system (family, village, firm, etc.) Radical individualism cannot work in sociology if only because the idea of an individualist social science makes as much sense as the idea of a collectivist psychology. After all, sociology studies social facts and, although these result from individual actions, they have nonindividual traits. Families, businesses, schools, armies, religious congregations, and other social systems are structured groups, and a structure is a set of relations, among them the bonds holding the system together. Likewise, social processes, such as trade, governance, and war, are impersonal affairs. This is why there are no radical individualist sociologists, with the possible exception of Homans. Many who call themselves methodological individualists, from Weber and Pareto to Coleman and Boudon, admit the peculiarity of the social, and only attempt to explain it as emerging from individual actions: they behave more like systemists than like individualists.

Holists rightly emphasize the centrality of connections, but they exaggerate in claiming that social structure precedes and dominates individuals. This view is logically untenable, because there are no relations without relata. In reality there are only interrelated individuals: relations, like their relata, are only the products of abstraction. For instance, there is no such thing as 'married' in itself, any more than there are unrelated spouses. Hence relations cannot act upon individuals, but interrelated individuals can alter or even sever some relationships—though never all of them at the same time except through suicide.

Systemism, often confused with holism, is a sort of synthesis of the latter with individualism. It holds that, because individuals are inter-related, individual behavior can only be understood in terms of both intentions and social constraints and stimuli. Thus, agency and structure are mutually inseparable. And social change, though brought about by individual action, consists in structural change—that is, change in the nature and strength of the bonds that hold social systems together. A social system may be characterized by its composition, environment (natural and social), structure, and mechanism. A change in any of these features may modify the other three. However, some features may resist changes in the others. For example, social order may survive demographic turnover and environmental catastrophe. Still, all four aspects coexist and, in the long run, important changes in any are bound to alter the other three. Yet individualism ignores structure and underrates environment, whereas holism ignores composition and mechanism, and underrates the natural environment. Only systemism accounts for all four, and is thus likely to favor realistic models of social groups and effective social policies.

2. Epistemology: Knowledge of the Social

There are as many views on the nature of social studies as general views on the nature of knowledge: skepticism, apriorism, empiricism, and realism.

2.1 Skepticism

Skepticism comes in two strengths: moderate (or methodological) and radical (or systematic). The moderate skeptic doubts a thesis as long as no solid evidence for it is available. By contrast, the radical skeptic questions the very possibility of knowledge of some kind. Whereas moderate skepticism is part of the

scientific attitude, radical skepticism is destructive and even self-destructive for, if everything is doubtful, why not skepticism itself? The variety of radical skepticism most common concerning the social is cognitive relativism, or the view that social facts cannot be known objectively. This view has been falsified by the growing pile of reliable social data and approximately true sociological hypotheses. Since every research project sets out to find out the truth about something, it presupposes the ability to find it.

2.2 *Apriorism*

Apriorism is speculation without reality checks. The 'grand social theories' of old were apriorist. So are the sociobiological theories on the alleged genetic roots of all social traits, from the incest taboo to social stratification to ideals of human beauty. However, nowadays the most popular *a priori* theories in social studies are the rational choice models. All of these assume that we always act in self-interest and moreover 'rationally,' that is, striving to maximize our expected utilities. These theories are *a priori* because they are seldom if ever checked against empirical data. In particular, they postulate that agents have certain preference rankings (or utility functions), that are seldom checked. Moreover, some rational choice models are untestable. Thus, if agents behave in an obviously irrational fashion, they are said to be subjectively rational—as when they play lottery. And if people are altruistic, it will be argued that altruism is nothing but enlightened self-interest. Being impregnable to empirical data, such models are unscientific.

2.3 *Empiricism*

Empiricism is the view that only empirical data count: that they are the source, content, and final judges of ideas. Note that the emphasis is on data (reports on observable facts), not on objective facts. So much so, that the neopositivists disallow all talk of objective facts as metaphysical, in going beyond experience. However, empiricism is not limited to positivist philosophers. Behaviorism is a prime example of empiricism, since it deliberately avoids the neural springs of behavior and it eschews theorizing. Ethnomethodology, though ostensibly tributary to the anti-empiricist philosophies of Dilthey and Husserl, is another instance of empiricism, since all its practitioners do is to record overt everyday behavior.

Empiricism condemns students to the shallowness characteristic of common sense, because the most interesting facts of any kind are invisible, whence they must be guessed. For example, we can see a status indicator but not status, let alone class; and we can hear political talks, but not the interests that they manifest or disguise. Only sociological theory can suggest the search for social data that indicate social process, in particular the mechanisms behind appearances. By banning theory, or restricting it to correlating commonsense data, empiricism curtails experience. Moreover, its limitation has elicited some of the postmodern attacks upon science.

2.4 *Scientific realism*

Scientific realism, or objectivism, has two components: the ontological thesis that the world outside the knower exists on its own, and the epistemological thesis that it can be known. As applied to sociology, scientific realism is the view that social facts are just as real as physical facts, and that they can be known by blending theory with observation, measurement, and (whenever possible) experiment. Such knowledge is objective, but partial and gradual. In the case of social facts, research must include not only the facts themselves but also the way the actors perceive them—for example, the fact that people gauge social inequalities, and react to them, by reference to the adjoining social group.

Although scientific realism demands that theories be ultimately justified by data, it does not stifle the sociological imagination the way empiricism does. A realist concedes that, just as some data trigger the curiosity of the theorist, at other times untested theories motivate the search for pertinent empirical data. And, on still other occasions, advances come from the confrontation of either rival theories or mutually incompatible data. In short, according to scientific realism, theorizing and empirical research should go hand in hand, now matching, now challenging one another.

On the critical side, scientific realists will ask interpretivists and other armchair students to exhibit empirical evidence for their 'interpretations' (hypotheses). They will ask rational choice theorists to specify and justify the utility functions they impute to people. They will remind Marxists and other 'economic imperialists' that real people engage in some activities that are economically unproductive or even counterproductive, and that many a social change—such as the diffusion of the computer and the revival of religious fundamentalism—has come about with little or no class struggle.

3. *Methodology: Social Research Strategy*

Methodology is the normative or prescriptive branch of epistemology: it attempts to lay down, justify, refine, or criticize research rules and procedures. Philosophers are interested only in general or cross-disciplinary strategic principles: they leave special

techniques to specialists. Here is a tiny random sample of the class of general methodological problems: How are causation and statistical correlation related? What is a social indicator? How to infer intention from behavior? Can every social feature be quantified, or are there irreducibly qualitative properties? Is an ideographic (or particularistic) science possible? Here we shall only review three rival research strategies.

3.1 Dataism: Collect Data

In the Anglo-Saxon sociological tradition, research is equated with data hunting, gathering, and processing, whereas theory is regarded as mere data compression (as in curve fitting), or even as parasitic upon data collection. To be sure, there is no factual (or empirical) science without empirical data. But data alone a science does not make, because they are what calls for explanation. To understand a bit of the world is to explain it. And explaining involves guessing (and then checking) mechanisms—causal, random, intentional, or hybrid. Hence, theorizing is just as important as data collecting, whether in sociology or in any other factual science. Moreover, only theory, by suggesting interesting variables, can spare us the barrenness and boredom of mindless data gathering.

3.2 Interpretivism: Impute Intention

The interpretivist school holds, rightly, that we can explain an agent's behavior if we know his intentions. But it overlooks the fact that we only have direct access to behavior, not to intentions. That is, the interpretivist is faced with the inverse problem of guessing intentions from actions. (Typically, inverse problems are insoluble or have multiple solutions. Think, for example, of the problem of decomposing a given even number into prime numbers, as opposed to the problem of adding prime numbers.) The interpretivist is entitled to guess anything, but not to present his guesses (hypotheses) as true unless they have passed the suitable empirical tests. The imputation of intentions is not only dicey: it is also insufficient, because intentions, even if known, may explain success but not failure. The social constraints upon agency, and the reactions of other people to one's actions, must be envisaged too. That is, individuals must be placed in their systems.

3.3 Scientism: Gather Data and Craft Theory Consistent with Data

Scientism is the thesis that the scientific approach is the best of all approaches to cognitive problems of all kinds in all fields. Scientism should not be mistaken for naturalism, or the thesis that the social world is a fragment of nature, and consequently social research should be conducted in exactly the same manner as the study of nature. One may admit that everything social is made, while holding that, once made, a social item is out there, along with rivers and butterflies. That is, social reality is constructed, but not by the knower. And it is changeable, but not because the knower changes their viewpoint. Scientism does not involve the denial of the specificity and irreducibility of the social. It only holds that the scientific method is applicable across all the disciplinary barriers, hence that it is advisable to trespass them whenever they are artificial. In the case of sociology, scientism favors close relations with all its sister disciplines, particularly economics, politology, and historiography. Scientism also promotes the proliferation of social indicators, the production and statistical processing of data, and the crafting of empirically testable theories—in particular mathematical models.

Hundreds of social indicators have been invented, measured and used since the mid-1960s. However, their very nature is still somewhat unclear, due to insufficient philosophical investigation. A social indicator bridges the observed to the unobserved, effects to causes, symptoms to mechanisms, and data to hypotheses. However, it does so ambiguously unless it can be shown that no other unobserved trait or event could possibly account for the data. For example, the popularity of low-cost merchandise may measure either consumer satisfaction or poverty; and high church attendance may measure either religiosity or the need for social integration (or for respectability). Ambiguity is typical of empirical indicators. By contrast, the indicators used in the more advanced sciences are unambiguous because they are backed by theories. For example, pH is a reliable measure of acidity because it is an index of proton concentration, and theory defines acids as proton donors. The moral for sociology is that the reliability of a social indicator can be increased either by grounding it in a theory, or by using it jointly with other indicators of the same hidden trait.

Values and morals would seem to defy scientific realism. After all, Weber held that science ought to be value-free and morally neutral because he wished to preserve objectivity as well as the difference between science and ideology. Now, it is certainly correct to distinguish value judgments from factual statements. But not all values are subjective: some, such as peace, welfare, and truth, are not a matter of taste. Moreover, it is possible to investigate scientifically how values and moral norms emerge, and why some of them have become obsolete. It may even be argued that there are moral facts, such as misery and torture, and consequently that some moral judgments can be justified scientifically. Example: 'Anomie is wrong because it is personally and socially destructive.' It is also correct to distinguish science from ideology. Social scientists endeavor to understand social reality, not to alter it.

However, some of their findings can be used by social activists, politicians and managers to alter social conditions in a rational fashion rather than through costly improvisation. That is, a scientific ideology—one based on social science and empirically testable—is not out of the question.

4. The Sociology–Epistemology Connection

The traditional view of the relation between science and philosophy is that they are disjoint or even mutually incompatible. This view is mistaken, because science teems with philosophical concepts, such as those of reality, law, and knowledge, and because scientific research presupposes the reality and knowability of the world—a philosophical thesis. In short, philosophy and science, in particular social science, overlap partially. This overlap is only partial because only philosophers analyze philosophical constructs, and only scientists are equipped to study social facts.

However, not all philosophies are congenial with science. Whereas some philosophies are extraneous to science, others are relevant to it. For example, phenomenology is extraneous to science because it rejects the scientific method and consists in egology, or the self-study of the self. Existentialism is even more alien, because it rejects rationality and does not deal at all with problems of knowledge. The philosophies congenial with science examine, among others, the problems, concepts, hypotheses, and methods common to science and epistemology. Hence, they can be judged to be adequate to the extent that they propose true accounts of actual scientific practice and help advance science. For example, Popper's philosophy is congenial with science because of its rationalism and objectivism. But his devaluation of confirmation is incompatible with the search for truth, since a hypothesis can only be said to be true to some extent if it has been well confirmed.

5. Sociology of Epistemology

The sociology of philosophy is only embryonic, and mostly speculative. In particular, little is known about the social roots of the various epistemologies of sociology. For one thing, it is not true that they are motivated by class interests. Nor is it true that all idealists are reactionary and all materialists progressive. For example, Hegel's followers split into a left-wing and a right-wing. The second Comte inspired both the Mexican conservatives led by Porfirio Díaz, and their contemporary Brazilian progressives.

What is true is that, once crafted for whatever reasons, some philosophical views favor certain social groups rather than others. For instance, radical skepticism, even while revolting against the establishment, objectively favors the latter in that it discourages scientific social research and therefore reasoned and well-organized opposition. Individualist philosophies favor anarchism and liberalism. By contrast, holistic philosophies jibe with communitarianism and totalitarianism. In short, some philosophies, particularly if fuzzy, can be craftily interpreted to suit anyone's social interests. But the more abstract philosophical ideas—in particular the logical and semantical ones—are socially neutral. Thus, neither the logical calculi nor the theories of (semantic) meaning and truth refer to social facts, whence they cannot support any social ideology. However, they are useful to examine the cogency of any views whatsoever, and to this extent they have a subversive potential.

See also: Action, Theories of Social; Empiricism, History of; Grounded Theory: Methodology and Theory Construction; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Macrosociology–Microsociology; Marx, Karl (1818–89); Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects; Neutrality: Axiological; Phenomenology in Sociology; Positivism: Sociological; Rational Choice Theory in Sociology; Sociobiology: Overview; Sociology, History of; Structure: Social; System: Social; Theory: Sociological; Weber, Max (1864–1920)

Bibliography

Albert H 1994 *Kritik der reinen Hermeneutik: Der Antirealismus und das Problem des Verstehens*. Mohr, Tübingen

Barnes B 1977 *Interests and the Growth of Knowledge*. Routledge & Kegan Paul, London

Boudon R 1980 *The Crisis in Sociology*. Columbia University Press, New York

Boudon R 1994 *The Art of Self-Persuasion*. Polity, Cambridge

Bunge M A 1996 *Finding Philosophy in Social Science*. Yale University Press, New Haven, CT

Bunge M A 1998 *Social Science under Debate*. University of Toronto Press, Toronto

Bunge M A 1999 *The Sociology–Philosophy Connection*. Transaction Publishers, New Brunswick, NJ

Caplan A L 1978 *The Sociobiology Debate*. Harper & Row, New York

Dilthey W 1959 Einleitung in die Geisteswissenschaften. In: *Gesammelte Schriften*, Vol. 1. Teubner, Stuttgart

Durkheim E 1895 *Les règles de la méthode sociologique*. Flammarion, Paris

Fay B 1996 *Contemporary Philosophy of Social Science*. Blackwell, Oxford

Fox R 1989 *The Search for Society: Quest for Biosocial Science and Morality*. Rutgers University Press, New Brunswick, NJ

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Gellner E 1973 *Cause and Meaning in the Social Sciences*. Routledge & Kegan Paul, London

Menger C 1968 Untersuchungen über die Methode der Sozialwissenschaften, und der politischen Oekonomie insbesondere. In: *Gesammelte Werke*, Vol. 2, 2nd edn., Mohr, Tübingen

Merton R K 1968 *Social Theory and Social Structure*, rev. and enl. edn., The Free Press, New York

O'Neill J (ed.) 1973 *Modes of Individualism and Collectivism*. Heinemann, London
Pareto V 1963 *A Treatise on General Sociology*, 4 Vols. Dover, New York
Popper K R 1957 *The Poverty of Historicism*. Routledge & Kegan Paul, London
Popper K R 1945 *The Open Society and its Enemies*, 2 Vols. Routledge & Kegan Paul, London
Weber M 1968 *Gesammelte Aufsätze zur Wissenschaftslehre*. Mohr, Tübingen
Winch P 1858 *The Idea of a Social Science*. Routledge & Kegan Paul, London

M. Bunge

Sociology, History of

The standard way in which surveys of the history of sociology are written and taught still lags behind actual developments in the field itself. The writing and teaching of the history of sociology continue to be strongly dominated by the tradition of the history of ideas. Only recently have some major attempts been made to view the history of sociology from a sociological angle, using sociological concepts, methods, and theories.

In the absence of a standard sociological survey of the development of sociology, we shall in this outline mainly follow the familiar textbook trajectory. This approach inevitably implies distortions. It suggests a pattern of Whig history, in which the history of sociology is made into a neat succession of successes.

To balance that bias, the reader should bear in mind that the development of sociology itself has been a social and cultural process, consisting of a multitude of short-term planned actions and interactions with aggregate results that in the long term have not been planned by any single individual. This process took place, moreover, as a part of wider social and cultural developments.

The context was primarily European and, since the second half of the nineteenth century, North American; yet, as Europe and North America belonged to a more encompassing global constellation, we should also consider the relevance of the processes of colonization and decolonization affecting nineteenth and twentieth century society and culture all over the world (Connell 1997, Collins 1997).

Sociology, furthermore, was one of the social sciences amidst other intellectual disciplines: academic disciplines ranging from physics and biology to history and philosophy, as well as nonacademic disciplines including journalism, political debate, and literature. Sociology took on an articulate form in continuous dialog with those other disciplines, sometimes borrowing ideas from them, sometimes opposing them. Most of the controversies within sociology itself reflected its relations to the larger field of intellectual activities. This would be true of the moral or political views informing sociological theories as well as of the preference for either quantitative, 'scientific' methods or a qualitative, 'hermeneutic' approach (Lepenies 1988).

In addition to the differentiation into (a) theoretical and methodological orientations, sociology has also become differentiated according to (b) empirical specializations, and (c) national traditions. These various forms of differentiation are respectively related to (a) the prevailing orientation to other intellectual disciplines including science, philosophy, literature, and journalism; (b) the general process of differentiation of functions in society at large; and (c) international relations. While the variety of empirical specializations and national traditions is easily visible, this is not so clearly the case with the way the national traditions are 'nested' in an evolving international constellation characterized by specific hierarchies, loyalties, and corresponding intellectual interests.

We shall not continuously refer to the more general conditions affecting sociology in the following account which, because of its encyclopedic format, is written in accordance with standard practice. The reader should realize, however, that these conditions are relevant to each of the five major stages into which the development of sociology can conveniently be divided:

(a) A long 'predisciplinary' stage up to 1830, marked by what in retrospect may be recognized as 'protosociologies.'

(b) The formation of an intellectual discipline, 1830–1890.

(c) The formation of an academic discipline with diverging national traditions, 1890–1930.

(d) The establishment as a fully fledged academic discipline with autonomous degrees, departments, and research facilities and an emerging international hierarchy, 1930–1970.

(e) A period of crisis, fragmentation, and attempts at new synthesis, 1970–2000.

In discussing the successive stages we shall have to observe a basic rule of 'phaseology' (Goudsblom 1996): when a later phase begins, new elements are added to the previous constellation; but most of the elements which were already present in the earlier phase will continue to exist, albeit in a modified guise.

1. *The Predisciplinary Stage*

By a discipline we mean a 'unit of teaching, research and professional organization' (Heilbron 1995). There is a body of knowledge, recognized as such under a generally accepted name, set down in textbooks and discussed in professional journals by practitioners.

All of these elements were absent from the field of sociology prior to the nineteenth century. And yet we can list many names of people who may, arguably, be regarded as predecessors, 'protosociologists': people who reflected about social phenomena, collected data

about social life, or tried to find patterns in human history. Some individual contributors to these traditions have gained great renown in their own cultures. Thus, from ancient Greece, the names of such men as Herodotus, Plato, Thucydides, and Aristotle stand out. Although they have been 'appropriated' as founding fathers by the practitioners of the established academic disciplines of philosophy and history, many of the problems they addressed lay in the area of what was later to become sociology. Chinese, Indian, and Arab culture have similar intellectual traditions; here, too, a few individuals have become the famous exponents of those traditions (Collins 1998).

In this brief survey, we shall be able to pursue only the Western line leading up to the emergence of sociology as an intellectual discipline in Western Europe and North America in the nineteenth and twentieth centuries. Since the Middle Ages, there was, on the one hand, a strong tradition of philosophizing about human social life, mainly in theological and moral-legal terms, and, on the other hand, an increasingly strong tradition of collecting facts, most often related to state matters and focused on knowledge of the population and its resources, both at home and abroad (as reported by explorers and travellers). By and large, the two traditions stood apart.

The transition from the predisciplinary stage to the actual conception of sociology as an intellectual discipline in its own right occurred in the eighteenth and nineteenth centuries. The shift consisted partly in conscious attempts to bridge the gap between theory and practice, between philosophical reflection and the collection of facts about social life. The leading intellectual model, furthermore, was sought no longer in theology or metaphysics but in empirical science. And, in so far as the focus had been either on the political realm or on more 'private' moral issues, it now moved to the structure and development of 'society' at large. This threefold shift in orientation was pioneered by the groups which have become known as the French *philosophes* and the Scottish moral philosophers. Their work was developed in different directions during the last decades of the eighteenth and the first decades of the nineteenth century.

Among the *philosophes*, Montesquieu was one of the first and most influential. He demonstrated that types of government are to be understood in relation to their moral and physical infrastructure. For an understanding of such connections, a recognition of the 'general spirit' of a nation is central. This 'spirit' or 'common character' results from the combined effect of various causes, including climate, religion, and commerce. Instead of drawing deductions from an original principle such as a political or social contract, Montesquieu thus started to unravel the complex interdependencies of human society.

While Montesquieu pioneered the study of what Auguste Comte later was to call 'social statics,' his younger contemporary, the future statesman Turgot, pointed to the possibility of also viewing human society from a developmental perspective, in other words, in terms of 'social dynamics.' That idea was further elaborated in theories of successive stages or phases by such writers as Condorcet and Henri de Saint-Simon.

In a kindred spirit, the Scottish moral philosopher David Hume proposed introducing 'the experimental method of reasoning into moral subjects.' Hume and his younger friend Adam Smith rejected theories based on an assumed state of nature. Such reasoning had characterized the natural law tradition—the predominant framework for early modern theories of government and moral obligation (Grotius, Hobbes, Locke, Pufendorf). For Hume, the state of nature was a hypothetical construct, incompatible with the precepts of empirical science. To consider states as the result of an original voluntary act of agreement was illusory and contrary to scientific procedure. Social institutions, in Adam Ferguson's famous phrase, 'are the result of human actions, but not the execution of any human design.' From this perspective, morality would have to be studied historically and comparatively, tracing its development through the successive stages of hunting, shepherding, agriculture, and commerce.

As the idea of a general social science gained ground, various attempts were made to connect that idea to existing scholarly disciplines. Some groups, particularly in France, pursued the ideal of a social science modeled after the natural sciences. Thus, by applying probability theory to voting procedures and judicial decision making, mathematicians such as Condorcet and Laplace proposed a 'social mathematics.'

British liberal thinkers such as Jeremy Bentham and James Mill somewhat similarly tried to build a social and political theory upon the idea of a moral arithmetic. Individuals and governments alike should promote the amount of happiness and reduce the amount of pain. Their 'felicific calculus,' aimed at the greatest happiness of the greatest number, provided the means for assessing the utility of public institutions. Others, such as Saint-Simon, couched their analysis of early industrial society in a medical idiom, advocating a 'social physiology.' Conservative social theorists such as Edmund Burke, Joseph de Maistre and Louis de Bonald, by contrast, while recognizing the reality of 'social relations' and 'society,' opposed the notion of a natural science of society, and rather sought inspiration in theology, philosophy, and history.

2. *The Formation of an Intellectual Discipline, 1830–90*

In the course of the nineteenth century empirical social research became a regular and organized endeavor. Most of it was centered on the 'social question' and intimately linked with governmental agencies and

reform movements. Besides a multitude of local initiatives, national associations arose for the coordination of these efforts: the French Académie des sciences morales et politiques (1832), the British National Association for the Promotion of Social Science (1857), the US Social Science Association (1867), the German Verein für Socialpolitik (1873). More theoretically informed work made use of the research findings, but was carried out mainly by individual scholars who either possessed private means or made a living by writing and lecturing. Thus Auguste Comte, John Stuart Mill and Herbert Spencer were for most of their professional lives publicists, men of letters with a strong interest in the sciences. All three contributed significantly to the formation of sociology as an emerging discipline, but none of them was a regular academic or was actively involved in social research.

Other writers, such as Alexis de Tocqueville, Karl Marx and Adolphe Quetelet, shared similar interests, but consciously avoided the term sociology. Marx and Tocqueville have been incorporated in the canon of the discipline only after 1945. While Tocqueville called for a 'new political science for a new world,' Marx's critique of political economy resulted in a theory of capitalism and class domination. The Belgian astronomer and statistical entrepeneur Quetelet continued the efforts of Condorcet and Laplace. He considered the ever growing body of statistics to be the proper basis for a 'social physics.'

When Comte introduced the term 'sociology,' he was likewise concerned with the idea of a fundamental science of human society. At the time, Comte had no interest in applied knowledge, rejected premature specialization, and wanted to construct an appropriate foundation for social science. The term sociology captured all of these ambitions quite well. It conveyed the idea of a positive science, suggesting both a high level of generality and a great degree of independence from the natural sciences. For Comte sociology was to become an uncompromising positive science which, however, should not be conceived of in terms of either physics or physiology. Sociology was to biology what biology was to physics: a relatively autonomous science with its own subject matter and methods.

In the *Cours de philosophie positive* (1830–42) Comte explained that different methods prevail in the various sciences: the experimental method in physics, the comparative method in biology, the historical method in sociology. Sociology had to be founded on a historical law. Since human beings specifically have the capacity to learn, the development of knowledge is the core of human development, and Comte considered his law of the three stages in the development of knowledge as the proper beginning of sociology.

Comte's plea for sociology as the general social science was well received in the UK, first in the circle of John Stuart Mill, then by Herbert Spencer. Mill rejected Comte's criticisms of psychology and political economy, but admired his philosophy of science and embraced the idea of sociology. Yet, like the French followers of Comte, Mill never wrote a properly sociological study. One of the first sociological treatises appeared in the 1870s and was written by Spencer—the most widely read sociologist of the nineteenth century.

For Spencer, progressive change was the common denominator of all natural processes. From the maturation of an embryo to the evolution of life in general, all living things evolve from the simple to the complex through successive differentiation and integration. Evolution is the natural process of change from incoherent homogeneity to coherent heterogeneity. Human societies are no exception: they too develop through differentiation, a process steadily accompanied by higher levels of integration and coordination.

Spencer's view of evolution was thus broader in scope than both Comte's sociological and Darwin's biological theory. Combining a laissez-faire view borrowed from economics with an embryological model of growth, Spencer gave the evolutionary principle the status of a universal law and made it the core of his all-embracing system of synthetic philosophy. As stated in his influential essay 'The Social Organism' (1860), Spencer's sociology was cast in the organic idiom, while carefully maintaining an individualist perspective. In *Principles of Sociology* (1876–97) he elaborated his analysis, emphasizing the passage from 'militant' to 'industrial society.' In the latter, social integration is no longer imposed from a controlling center, but the spontaneous result of individuals who cooperate on the basis of a division of labor. Spencer strongly favored a laissez-faire stance, opposing state politics (in matters of welfare as well as in colonial expansion), and inspired what was—inaccurately—called social Darwinism.

His position as a freelance writer and lecturer allowed Spencer to make virtually unlimited claims for the scope of sociology. While he was most outspoken and influential in articulating the idea of sociology as an intellectual superstructure sustained by an evolutionary perspective and overarching all the other social sciences, that idea was shared by several other leading advocates of sociology in Europe and the USA.

3. *The Formation of an Academic Discipline, 1890–1930*

Sociology as an academic discipline emerged between 1890 and 1930. The first courses and chairs were established, sociological journals appeared as the primary outlet for research, and sociological associations were founded for furthering intellectual exchange and professional interest. France, the UK, Germany, and the USA were the core countries. In all

of these countries, distinct national traditions took shape, with the curious effect that some of the academic pioneers, most notably Durkheim and Weber, never referred to each other.

Emile Durkheim was the pivotal figure in French sociology. He taught his first course in 1887, wrote classic works on the division of labor, suicide, and the rules of sociological method in the 1890s, and organized a network of highly productive scholars around his journal, the *Année sociologique* (1898–1912). By proclaiming that social facts had to be explained by other social facts, Durkheim broke away from the predominant biologistic and purely psychological approaches. His explanation for variations in the suicide rate was a paradigmatic example. The social realm was a reality *sui generis* with its own structure and regularities. Systems of collective representations are an essential part of social life, not merely when studying religion or morality, but also for understanding the economy and political institutions. Following this line of argument, Durkheim and his collaborators (Marcel Mauss, Henri Hubert, Marcel Granet, Maurice Halbwachs, François Simiand) worked on a wide range of topics, contributing to many fields outside of sociology proper.

In the UK sociology did not gain a proper entry into the university system until well after World War II. The rich tradition of factual inquiry was continued by government officials and wealthy individuals (Charles Booth, B. S. Rowntree), but remained almost entirely separated from academia. Despite the existence of a sociological journal and a professional association, the British academic establishment did not show much interest in sociology. The only chair for a long time had been established in 1907 at the London School of Economics for L. T. Hobhouse, who, in 1929, was succeeded by his pupil Morris Ginsburg. Both men maintained a cautious evolutionary perspective but neither one built up a network of productive scholars. Whereas British academics showed little interest in studying their own society, they had no such inhibition about studying the exotic regions of the British Empire. Social anthropology became a respected academic discipline and partly served as a substitute for sociological concerns.

Sociology in Germany emerged later than in France and England; a professional association was founded in 1910, a first chair in 1917, a journal in 1921. The academic pioneers in Germany, men like Ferdinand Tönnies, Georg Simmel and Max Weber, faced two powerful indigenous traditions. The oldest was that of the *Staatswissenschaften*. The proponents of these disciplines (law, economics, administration) consistently subsumed civil society to the various functions of the state. This mode of thinking had its roots in the eighteenth-century system of centralized political administration and economic policy known as cameralism and was refashioned in the nineteenth century by idealist philosophers such as Hegel. After the founding of the German Empire, the members of the *Verein für Socialpolitik* maintained much of this conception of social science. They were opposed by liberal economists like Carl Menger during the first *Methodenstreit*, and by Max Weber and others during the second *Methodenstreit*. An important aim in Weber's argument about the value-freedom of scientific inquiry was to create a greater distance from the practice of the state sciences.

The other tradition with which German sociologists grappled radically opposed the natural to the cultural sciences: the first were supposed to be concerned with causal explanation, the second with interpretative understanding. Early sociologists such as Comte and Spencer had often been rejected in Germany as representatives of an inadequately naturalistic conception of the human sciences. Reflecting on these epistemological issues, Max Weber proposed to transcend the dichotomy by ingeniously defining sociology as the science of social action, which by interpretative understanding seeks to explain its course and consequences.

Weber was nominated for a chair in sociology shortly before his death in 1920. His achievements and those of others of his generation (Simmel, Alfred Vierkandt, Werner Sombart, Franz Oppenheimer) contributed to the recognition of sociology as an academic discipline during the Weimar Republic (1919–33). Sociology expanded rapidly, opening opportunities for a generation of young scholars including Max Horkheimer, Theodor Adorno, Theodor Geiger, Karl Mannheim, and Norbert Elias. Although these men had initially received a training in different academic fields, they were ready for brilliant careers in sociology. However, the seizure of state power by the National Socialist Party in 1933 made an abrupt end to the favorable outlook for German sociology, forcing all these men, along with many of their colleagues, into exile.

Sociology's most rapid expansion took place in the USA. The first courses were taught in the 1870s and 1880s, while the expansion of higher education in the 1890s produced rapid institutionalization. Within two decades sociology was taught in most colleges and universities; textbooks standardized the concepts and methods of the new discipline. US sociology developed primarily on the basis of empirical research into social problems. The main topics were the issues which were widely debated by politicians, religious leaders, and social reformers: massive immigration, urban problems, industrialization, local communities, social (dis)integration, ethnic and race relations. Empirical studies proliferated, based either on statistical material or on actual fieldwork.

By the 1920s the sociology department in Chicago, founded by Albion Small in 1893, had become the predominant center. Small also played a key-role in the *American Journal of Sociology* (1895) and—together with George E. Vincent—wrote the most widely

used textbook, *An Introduction to the Study of Society* (1894). The work of the Chicago School was founded on qualitative case studies, using observations, interviews, and life histories. Exemplary were William I. Thomas and Florian Znaniecki's *The Polish Peasant* (1918) and the collective volume on *The City* (1925).

4. *Establishment as an Academic Discipline, 1930–70*

Between 1930 and 1970, sociology became established as an academic field at most universities in the Western world. The number of practitioners and students greatly increased, and there was a proliferation of specialisms in research and teaching. Far from being an even process of growth all over, however, these trends were marked by far-reaching structural changes: a forcible break in the development of sociology in Europe, and a concomitant shift of intellectual leadership in the field from Europe to the USA.

That shift, which strongly affected the development of sociology, was illustrated strikingly by two historical introductions to the discipline written toward the end of this period: *The Sociological Tradition* by Robert A. Nisbet (1966) and *The Origins of Scientific Sociology* by John Madge (1962). Nisbet's book celebrated the formative era of the second and third stages, which he designated as 'the golden age of sociology.' Here was a US scholar revering the 'titans' of nineteenth-century sociology: a small pantheon of authors whose highly personal visions provided the basic ideas and concepts for their more pedestrian successors in the twentieth century. Almost inevitably, all the heroes of Nisbet's great sociological tradition were Europeans.

Madge's book radiated a very different atmosphere. Madge was an Englishman, proudly hailing the achievements of scientific sociology in the twentieth century. He discussed 12 pieces of empirical research; with the only exception of Durkheim's *Suicide*, they all belonged to the fourth stage in the development of sociology. Many of the studies selected by Madge were published under joint authorship, and, again except for *Suicide*, they were all carried out in the USA.

The leaders of the first generation of US academic sociologists all acknowledged an intellectual debt to European masters. Even in the early 1920s, it was still not uncommon for promising US social scientists to start their career with a European apprenticeship, as did Talcott Parsons who went to London and Heidelberg. At the time, however, the westward exodus of sociologists from Eastern Europe was already underway, as the communist regime in the USSR and the fascist regime in Hungary tightened their grip, greatly restricting the margins for sociological inquiry and reflection. 1933 brought the decisive rupture: the seizure of power by the Nazis in Germany put an end to the flourishing sociology of the Weimar Republic, and forced a great number of prominent and promising sociologists into exile, mostly to the USA.

In contrast to Europe, sociology in the USA suffered no drastic interruptions in the years between 1933 and 1945. It remained a subject in most college curricula, with textbooks such as Park and Burgess's *Introduction to the Science of Society* (1924) continuing to supply an authoritative framework of concepts and ideas. At the same time, practically oriented research in areas of social problems and social policy was stimulated by increasingly generous funding from both government and private foundations (Platt 1996). Empirical research received a powerful boost after 1941, when social scientists were recruited for the war effort and commissioned to study factors fostering the morale of the Armed Forces.

After 1945, the USA emerged as the major political, economic, and cultural power in the world. In the social sciences, including sociology, US examples were virtually unrivaled in setting an international standard of competence in method and theory. An impoverished Europe lacked the resources for challenging the US ascendency. The relatively autonomous sociological traditions in Germany and Italy were virtually destroyed. Sociology in the UK had only a minor position at the London School of Economics. In France the Durkheimian tradition continued to exert considerable intellectual influence among anthropologists and historians, but the actual institutional basis for sociology as an academic discipline was very weak.

Thus the scene was set for US dominance. The sociology departments at Harvard and Columbia University in particular gained international reknown. At Harvard, Talcott Parsons developed a new synthesis of sociological theory, mainly derived from European predecessors, which he presented as a 'general theory of action'—although in its worldwide reception it became known as 'structural functionalism' or simply 'functionalism.' Still, Parsons' generally recognized status as the world's leading sociological theorist did not imply that his ideas put a mark on the entire discipline. There were some influential rival theorists, including his Harvard colleague George Homans who advocated a strictly scientistic 'exchange theory' inspired by the model of *homo economicus*. At Columbia, Parsons' former student Robert Merton propagated a more modest ideal of 'theories of the middle range,' which proved to be very well compatible with the work of his colleague Paul Lazarsfeld, an immigrant from Austria who specialized in the organizational and methodological aspects of social research.

The years between 1955 and 1975 in particular spelled a boom period for American sociology (Turner and Turner 1990). The successful launching of Sputnik

by the USSR in 1957 inaugurated an era of unprecededly high government funding of scientific research in the United States, which also benefited sociology. In the 1960s, a combination of demographic trends (the postwar baby boom) and political discontent made for a dramatic increase in student enrollments. Professional self-confidence, which had been clearly mounting since the early 1940s, surged to new heights. The auspices were favorable for manifold activities in the realm of academic sociology: systematic elaboration of theoretical ideas in the Parsonian fashion; exploration of the possibilities of rival theories presented under such labels as exchange theory or symbolic interactionism; empirical research, conducted in many forms from intensive 'qualitative' case studies to extensive quantitative surveys, and applied to a broad range of social phenomena. Sociologists in most other capitalist-democratic countries outside the USA adopted similarly diverse programs for teaching and research. As the meetings of the International Sociological Association testified, by the end of the 1960s sociology was a blooming pluralistic discipline, with a tacitly accepted hierarchy of prestige headed by US sociologists whose ideas set the agenda for the discipline at large. Unintentionally, the publication of the *International Encyclopedia of the Social Sciences* in 1968 spelled the ending of that era.

5. *Crisis, Fragmentation and Attempts at New Synthesis, from 1970*

If the years around 1970 formed a watershed in the development of sociology, the tendencies that then became manifest had been at work for some time. By the time that readers all over the world pored over Alvin Gouldner's (1970) *The Coming Crisis of Western Sociology*, that crisis actually was in full swing. As Gouldner pointed out, mainstream academic sociology as represented by the combination of functionalist theory and survey research was rapidly losing its appeal to a new and vociferously critical generation. With much programmatic gusto, new 'paradigms' were launched, highly diverse in their implications, but all radiating the same radical spirit. Inspired by political protest movements, a culture of dissent arose in sociology, criticizing the intellectual canon and the professional stance of the discipline. The short-term result of this turmoil was the demise of functionalism as the leading theoretical perspective, and a flurry of controversies about a mixture of ideological, epistemological, and methodological issues. In the longer term, after the controversies receded, a state of fragmentation ensued to which most practitioners tacitly accomodated.

Fragmentation took many forms. The most obvious and least contested form was the proliferation of subdisciplines around an unsystematic array of specific research areas such as the sociology of the family, religion, organizations, etc. Then, equally noticeable but far more controversial, there was a plurality of theoretical orientations, often with corresponding methodological presumptions. Some theories were primarily attuned to specific forms of microsociology, marked by an affinity to various philosophic schools ranging from phenomenology (embraced by ethnomethodologists), through pragmatism (preferred by symbolic interactionists), to neopositivism (for rational choice theorists). At the level of macrosociology, functionalism as a perspective gave way to new approaches influenced by Marxism and emphasizing social conflict and long-term social transformations. Fragmentation was further enhanced by a weakening of the hierarchy in the international sociological community and an increased articulation of national traditions.

None of these changes came out of the blue, but together they produced an almost revolutionary shock to the world of sociology. Whereas the most famous sociologists of the previous generation had all been from the USA, now European names again came to the fore: Jürgen Habermas, Pierre Bourdieu, Anthony Giddens. During the early 1970s, while many older sociologists deplored the patent end to the illusion of consensus in the discipline, there was also a widespread sense of excitement over new intellectual possibilities and career opportunities. Ideological debates around many current topics, from race relations to the reform of psychiatry, were saturated with sociological references. In his critique of sociology, Alvin Gouldner (1970) did not present himself as a prophet of doom, but rather as a pioneer in 'reflexive sociology' who was able to point at new ways of fulfilling what C. Wright Mills earlier called 'the promise of sociology.'

The later 1970s saw a dramatic reversal in this mood of confidence. General public interest turned away from sociology, student enrollments dropped rapidly, research funds shrank–the only sediment of the earlier high tide that remained was internal fragmentation, pluralism. The general stagnation at first reinforced the tendencies toward fragmentation, as established subdisciplines tried to maintain their position in the face of diminishing institutional support for, and intellectual vigor of the sociological discipline as a whole. However, the 1980s and 1990s also brought new attempts towards a synthesis. Thus, in historical and comparative sociology the orthodox Marxist approach tended to be softened by combining it with a more liberal Weberian perspective. 'World systems' theory as originally developed by Andre Gunder Frank and Immanuel Wallerstein helped to break away from the implicit nation-centeredness that characterized most sociological work during the fourth stage. Making the international order, conceived either as a world system, or alternatively as 'transnational society,' or in terms of a process of

globalization, into the centerpiece for sociological study gave counterweight to the tendencies toward national parochalism.

Other ventures in historical and comparative sociology sought synthesis in a different direction. During the 1980s and 1990s, Randall Collins (1997) valiantly subjected a wide range of long-term cultural trends (including religion and philosophy as well as technology) to sociological analysis, while Jack Goldstone and Charles Tilly brought sophisticated techniques of statistical analysis to bear on huge sets of historical data. Meanwhile Norbert Elias's work, aiming at a thorough integration of sociology with history and psychoanalysis, was gaining influence inside and beyond the discipline.

These promising signs cannot hide the fact, however, that after the decades of rapid growth in the 1960s and 1970s, sociology had entered a period of far less favorable circumstances. The welfare state came under siege in many Western countries, and neoliberal policies in favor of deregulating and extending markets gained the upper hand. Even if the collapse of the USSR and the termination of the Cold War have not spelled 'the end of history,' these events have reduced interest in ideological matters and theories of society. A theoretical perspective that commands much public interest seems to be sociobiology—which, as the anthropologist Marvin Harris observes, may well constitute the most provocative challenge to the social sciences.

At the universities, the established social science disciplines (economics, sociology, anthropology, social science) have faced increasing competition from new, often multidisciplinary fields of study which are more immediately focused on policy needs (organizational studies, management, business) or the demands of particular student groups ('Black studies', gender studies). The strains of the present situation are also conducive to a revival of the old polarity between 'scientistic' and 'humanistic' approaches. The scientistic posture has given a great deal of prestige to 'rational choice' models by means of which some sociologists are applying assumptions and methods derived from economics to a wide array of non-economic relations. Others fear that the economic models thus adopted may have the effect of a Trojan horse, undermining sociology at its very center. Ironically, the humanistic approach is posing a similar risk: embracing culturalistic and postmodern ideas, drawn mainly from current anthropology, literary studies, and hermeneutic philosophy, may also detract sociologists from the core of their discipline—the study of social processes and social structures.

The opposing social currents at work in the world today, as well as the more directly felt institutional pressures and intellectual challenges are confronting sociology with a variety of contradictory claims. Recognizing the many sources of confusion is a first requisite in order to arrive at a balanced response to this situation. Equally important is a continuous reconsideration of the history of the discipline.

See also: Comte, Auguste (1798–1857); Disciplines, History of, in the Social Sciences; Durkheim, Emile (1858–1917); Marx, Karl (1818–89); Paradigms in the Social Sciences; Sociology: Overview; Spencer, Herbert (1820–1903); Universities, in the History of the Social Sciences; Utilitarian Social Thought, History of; Weber, Max (1864–1920)

Bibliography

Bottomore T, Nisbet R (eds.) 1978 *A History of Sociological Analysis*. Basic Books, New York

Boudon R, Cherkaoui M, Alexander J (eds.) 1997 *The Classical Tradition in Sociology*, 8 Vols. Sage, London

Collins R 1997 A sociological guilt trip: Comment on Connell. *American Journal of Sociology* **102**(6): 1558–64

Collins R 1998 *The Sociology of Philosophies*. Belknap Press of Harvard University Press, Cambridge, MA

Connell R W 1997 Why is classical theory classical? *American Journal of Sociology* **102**: 1511–57.

Goudsblom J 1977 *Sociology in the Balance*. Blackwell, Oxford, UK

Goudsblom J 1996 Human history and long-term social processes: towards a synthesis of chronology and 'phaseology'. In: Goudsblom J, Jones E, Mennell S *The Course of Human History*. M. E. Sharpe, Armonk, NY

Goudsblom J, Jones E, Mennell S 1996 *The Course of Human History*. M. E. Sharpe, Armonk, NY

Gouldner A W 1970 *The Coming Crisis of Western Sociology*. Basic Books, New York

Heilbron J 1995 *The Rise of Social Theory*. Polity Press, Cambridge, UK

Kilminster R 1998 *The Sociological Revolution. From the Enlightenment to the Golden Age*. Routledge, London

Lepenies W (ed.) 1981 *Geschichte der Soziologie*, 4 Vols. Suhrkamp, Frankfurt, Germany

Lepenies W 1988 *Between Literature and Science: The Rise of Sociology*. Cambridge University Press, Cambridge, UK

Levine D N 1995 *Visions of the Sociological Tradition*. Chicago University Press, Chicago

Madge J 1962 *The Origins of Scientific Sociology*. Free Press of Glencoe, New York

Mohan R P, Wilke A S (eds.) 1994 *International Handbook of Contemporary Developments in Sociology*. Greenwood Press, Westport, CT

Nisbet R A 1966 *The Sociological Tradition*. Basic Books, New York

Platt J 1996 *A History of Sociological Research Methods in America 1920–1960*. Cambridge University Press, Cambridge, UK

Turner S P, Turner J H 1990 *The Impossible Science. An Institutional Analysis of American Sociology*. Sage, Newbury Park, CA

Wagner P 1990 *Sozialwissenschaften und Staat: Frankreich, Italien, Deutschland 1870–1980*. Campus Verlag, Frankfurt, Germany

J. Goudsblom and J. Heilbron

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Sociology: Overview

The institutionalization of sociology has followed a long and nonlinear path.

The word sociology was created by Auguste Comte and became established once it was institutionalized in periodicals. Meanwhile, the existence of sociology was recognized in an implicit fashion. When Tocqueville ([1856], 1986) writes in the very first sentence of his book on *The Old Regime and the Revolution* 'This book is not a book of history,' he meant that it dealt with what we would call today 'comparative sociology.' But he could not use a word which was not yet common and probably considered Comte as an eccentric figure. Max Weber expressed doubts as to whether his work should be qualified as belonging to sociology. The institutionalization of the word and of the discipline owes much to Durkheim and later to the Chicago school of sociology.

1. Sociology: A Unique or Multiple Discipline?

Sociology, commonly defined as the scientific study of social relations,social institutions, and societies, has nevertheless displayed a great diversity of ways of conceiving its objectives, uses, styles, and methods. Durkheim never quoted Weber nor Weber Durkheim, although they were contemporary and could read one another. Nisbet (1966) maintains though that, beyond the differences, a common inspiration can be detected in the works notably of Tocqueville, Marx, Simmel, Weber, Pareto, or Durkheim. All would have seen that the rational utopian theory of society developed by the Enlightenment was unacceptable. For this reason, they all insisted on the social functions of religion, of the hierarchical distinctions between social statuses, on the role of traditions, or on the distinction between 'community' and 'society.'

Describing the identity of classical, and still more of modern, sociology can be harder though than Nisbet states, for both classical and modern sociologists endorse diverse main views as to the uses of sociology and to the ways it should be conducted.

1.1 Three Types of Objectives

A first type of sociologist sees sociology as having a mainly *informative* function: Its role would consist in producing data and proposing analyses implicitly and explicitly oriented toward decision makers. Thus, Le Play in his *Les ouvriers européens* (1855) produced a survey of the conditions of life of European workers, with the final objective of improving the social policies of the time. While workers themselves know the way they live, this local knowledge is not directly available at the level of societal decision makers. Collecting statistical data on crime is another illustration of this informative function: Such data are crucial to the determination of crime policy. Tarde became a sociologist because, as a judge and later as the head of the French Bureau of Statistics on Crime, he was concerned with the increase of crime rates. Schumpeter qualified this informative dimension of the social sciences as 'cameral' (*Kameralwissenschaften*). In the seventies, Lazarsfeld et al. (1967) produced an overview of the production of this 'cameral' sociology in their *Uses of Sociology*.

A second type of sociologist sees the function of sociology rather as *critical*: Sociologists should mainly aim at identifying the defects of actual societies and devising remedies to correct them. Saint-Simon, Proudhon, or Owen can be considered as examples among the pioneers of this type of sociology. The members of the Frankfurt School provide a modern illustration of 'critical' sociology.

A third type of sociologist sees sociology as having the same main objective as other scientific disciplines: explaining social phenomena which appear at first glance as enigmatic. Thus, Max Weber wonders why peasants appear as more attracted by polytheistic than monotheistic religions and Durkheim why all religions include the notion of 'soul' in one form or another. In such cases, the interests of sociologists are essentially of a *cognitive* character.

These three basic orientations can appear in combination. Marx is concerned both with explaining social mechanisms and with changing society. Durkheim gathered data on suicide with the objective of explaining the variations of suicide rates, but he considered them also as indicators of the general state of modern society. But Tocqueville, Weber, or Durkheim have reached the status of 'classics' essentially because they have proposed scientifically valid explanations of puzzling social phenomena, rather than by their diagnoses of the society of their time. Durkheim considered that sociology could be socially useful only if it was scientifically solid, in other words that good criticism can derive only from valid explanation. By contrast with Durkheim or Weber, the Frankfurt School is known more for its criticism of modern societies than because it would have proposed convincing explanations of puzzling phenomena.

The three orientations characterize modern and contemporary as well as classical sociology. Foucault's *Supervise and Punish* is generally recognized rather as a 'critical' than as a 'scientific' contribution (Foucault 1975). The 'labeling theory' teaches that some features of individual A can be interpreted as a 'label' by observer B and can structure B's behavior and expectations toward A. This has been known for ever, though, so the theory does not bring additional knowledge. Its 'critical' function is more undebatable than its cognitive function. So, the 'labeling theory' is not a theory in the scientific sense. But, as Foucault's

theory, it notably had an influence on our views and on the views of public officials as to the way prisoners should be considered and treated.

1.2 Several Types of Orientations

But sociology is perceived as a discipline with a weak identity, not only because of the existence of these three basic distinct goals, but also because the sociologists who consider that the main goal of sociology is to produce valid new knowledge on social phenomena endorse a variety of methodological and theoretical orientations. Thus, Weber's work is considered as illustrating a particular style of sociology, 'comprehensive' sociology, which is currently opposed to the 'positivist' conception of sociology. The former proposes to analyze social phenomena as the outcome of comprehensible individual actions; the latter to analyze the relations between social phenomena. Weber himself saw his own version of 'comprehensive' sociology as very different from others, and talked currently of 'comprehensive sociology in my sense (*in meinem Sinne*).' In the 1960s, 'quantitative' and 'qualitative' sociology were perceived as irreconcilable competitors: the former saw collecting standardized data able to be subjected to statistical analysis as the most effective research procedure, while the latter insisted that it would produce biased findings and proposed to use much more flexible methods. Merton's 'middle range theory,' that aimed at explaining specific social phenomena, was seen as the proper alternative to the type of general theory illustrated by Parsons, which C. Wright Mills had polemically called 'Grand Theory,' the object of which was 'society' as a whole. These notions label some of the main conceptions of sociology that can be distinguished and appear as more or less permanent.

This diversity in the basic theoretical orientations of sociology is also a lasting feature of sociology, observable in contemporary as well as classical sociology. At the beginning of the twentieth century, a first 'methodological quarrel' opposed those who thought that sociology should follow the path of other scientific disciplines, to those who, like Rickert, maintained that, while the goal of natural sciences is 'explanation,' the goal of the social sciences would be 'interpretation.' In the 1960s, the quarrel was reopened, opposing the 'positivists,' as their challengers called them, to those—the 'hermeneuticians'—who, once again, saw sociology as in charge of disentangling the 'meaning' of social phenomena—of 'interpreting' them—rather than identifying their causes, i.e. 'explaining' them. Adorno and Popper were respectively the main protagonists of the 'antipositivist' and of the 'positivist' group (Adorno et al. [1969] 1972). In various forms, this 'quarrel' reappears almost permanently. Thus, not long ago, ethnomethodologists claimed that the idea of producing an objective social knowledge would be self-contradictory, for social phenomena would be understandable only from the inside of a social group: one could interpret, not explain them.

Weber's sociology is particularly important, for it defines a third way between 'positivism' and 'antipositivism.' Explaining a phenomenon means to Weber finding its causes. The causes of any social phenomenon reside in the set of individual actions, beliefs, or attitudes of individual people responsible for it. So the sociologist has to make these individuals actions, beliefs, or attitudes 'understandable.' But understanding an individual action means finding out the meaning of the action for the actor, in other words the motivations and reasons that have inspired it. For Weber this operation amounts to building a theory of the motivations of social actors subject to the ordinary criteria any scientific theory should meet. Thus, when I see somebody cutting wood in his yard, I would reject the theory according to which he would cut wood because he is cold, if this action takes place on a hot summer day. On the whole, to Weber, 'understanding' is a crucial moment in the 'explanation' of any social phenomenon.

So, the sociologists mainly interested in the 'cognitive' function of their discipline, namely in its capacity to produce valid new knowledge, can be classified along a first main dimension, according to whether they see it as able to produce '*explanations*,' as in other scientific disciplines, or as rather entitled to produce '*interpretations*' with an unavoidably subjective character. But a secondary distinction should be introduced. It opposes those who, like the Durkheim of the *Rules*, believe that the subjectivity of social actors should be ignored by sociologists (Durkheim 1895) and those who, like Weber (1951) consider on the contrary that it is essential to sociological knowledge and in no way incompatible with a scientific ambition.

The latter distinction is correlated with another one: to some sociologists, individuals are considered as the atoms of any sociological analysis, while, to others, sociology should ignore individuals and rather be concerned with analyzing the associations between social phenomena and variables. The former follow the principles that Schumpeter called '*methodological individualism*,' the latter '*methodological holism*.' Weber declared explicitly that, to him, individuals are the atoms of any sociological analysis and that sociology should be 'individualistic as far as its method is concerned' (Mommsen 1965). Durkheim, by contrast, saw sociology rather as devoted to the analysis of the associations between observable social variables. He defends this viewpoint theoretically in his *Rules* (1895) and practically, for instance, in his *Suicide* (Durkheim 1897).

It suffices to evoke a few modern sociological schools to see that the above distinctions, along which classical sociological works can be classified, can also be used in the case of modern works. Thus, 'struc-

turalism' can be ranked on the side of positivism: It proposes to identify the stable relations prevailing between social variables; it claims to be hard science; on the other hand, it is holistic, since it treats individuals as being merely the products or reflections of 'structures.' 'Rational choice theorists' for their part propose to analyze social phenomena as the effects of individual actions; they believe firmly that sociology can be as objective and scientific as any other discipline, but under the condition that it respects the postulate that social phenomena are the outcome of individual actions led by individual interests. 'Ethnomethodologists' believe that sociology is an interpretative rather than explanatory discipline, but are mainly concerned with the interpretations given by individuals of the social setting they are embedded in. Holistic sociologists of history, such as Elias or Wallerstein, following the footsteps of a Burckhardt or a Braudel, are concerned with detecting diffuse orientations in the course of historical and social processes. They worry rather about interpreting than explaining social phenomena and are interested mainly in identifying global macroscopic trends.

The fact that the various conceptions of sociology emerging from the above typologies are lasting explains why sociology appears as a less unitary discipline than others. Moreover, the relative popularity of each of these conceptions tends to follow an 'oscillatory' process, as Pareto would have said. In many circles, structuralism was considered in the 1960s as the orientation which would make the social sciences genuinely scientific. But it was sharply criticized and became obsolete. Being 'positivist' and 'holistic,' it was replaced by antipositivist and individualistic currents, such as ethnomethodology or 'phenomenology' in the sociological sense.

2. *A Few Genuine Achievements*

The most effective way to refute the skepticism toward sociology that develops as a consequence of its diversity is to mention a few genuine achievements in some of the branches of the typologies outlined above.

Durkheim's *Suicide* (1897) is a classical example of the holistic-positivist type of sociology. The work consists in a systematic analysis of correlations between rates of suicides and a number of independent variables, such as sex, living in cities vs. living in the countryside, religious affiliation (Protestant, Catholic, Jew), and so forth. Thus, Durkheim found that, when anomie grows, rates of suicide become higher. Other things being equal, professionals are for instance more exposed to anomie than peasants, so that the rates of suicide of the former are higher. Anomie is greater in periods of economic boom, for expectations tend to be overoptimistic then, opening the way to disillusion and making the rates of suicide higher. Durkheim's *Suicide* provides a methodological model which has inspired many writers. Cusson (1990) wonders for instance why the rates of criminality are characterized by similar trends in most Western countries, with the two exceptions of Japan and Switzerland. He maintains that this exception is due to the fact that primary social control is more developed in these two countries than in others, because, even in large cities, Japanese and Swiss societies have, for historical reasons, maintained networks of interpersonal relations characteristic of traditional societies.

It should be noted, however, that, while such analyses are explicitly holistic, they implicitly consider individual motivations. Thus, in his *Suicide*, Durkheim contends that periods of political crises should be associated with a decline in the rates of suicide and that, other things being equal, they were effectively lower during the major political crises in France in the last two decades of the nineteenth century, as well as during the crisis between Prussia and Austria in the mid-1860s. As elsewhere in *Suicide*, the theory predicts an association between two variables defined at the societal level, and is in that sense holistic. However, the theory rests implicitly on the 'psychological' assumption that in periods of serious political crises, people cannot feel exclusively concerned with their own private worries. The same could be said of most analyses developed in Durkheim's *Suicide*. Thus, the activity of professionals is much less routinized than the activity of peasants; they cannot be content with following day after day the same rules of behavior, and must on the contrary always devise new rules, objectives, procedures, or ideas. This higher uncertainty generates a higher anxiety. So, implicitly, in spite of his official opposition to any psychological explanation of social facts and of the recommendation formulated in the *Rules* always to 'explain social facts by social facts,' 'understanding' in Weber's sense is implicitly conducted in the analyses of *Suicide*. Thus, in the nineteenth century, women had lower suicide rates than men because, being more exposed to social control, they were, according to Durkheim, more guided by routinized rules and norms and thus less exposed to the anxiety generated by uncertainty.

So, the contrast between the antipsychological holistic sociology of Durkheim and Weber's individualistic approach is less marked than is sometimes contended. Weber is as interested as Durkheim in macroscopic relations. Thus, in his *Essays in the Sociology of Religion*, he notes that Mithraism developed essentially among Roman civil servants before reaching other strata (Weber 1920–1). His explanation is that Mithraism was more palatable to civil servants than the old traditional Roman religion, because it could be perceived as a symbolic expression of the political philosophy which grounded the Roman bureaucracy. Roman civil servants were hierarchized in a number of grades associated with specific functions; admission to the bureaucracy and promotion from one grade to the higher was determined by

impersonal procedures; the whole system was placed under the authority of the Emperor, but the latter was held rather as the incarnation of the impersonal rules codified in the Roman Right. Now, Mithraism was a religion placed under the auspices of a God symbolizing impersonal rules; as in the case of Freemasonry, the believers were admitted and promoted through a scale of grades with the help of impersonal procedures. Here, Weber solves a macroscopic question—why are such and such strata more receptive toward a religious doctrine?— in an explicitly 'individualistic' fashion, by showing that Roman civil servants had strong reasons for perceiving Mithraism as more attractive than the old Roman religion, which was born in the context of a purely agrarian society. For the same type of reasons, Freemasonry in Prussia has attracted civil servants more than other strata.

The Weberian vein is noticeable in many modern analyses. Thus, Merton's classical analysis of the anti-Black attitudes of the white American workers in the time of the Great Depression of the 1930s rests upon the assumption that white unionists had strong reasons to believe that black workers, coming from the paternalist big farms of the South, would be less prepared to be faithful to the unions than the white workers from the North (Merton 1949). Bergeron (1999) wondered why methadone programs were rejected in France, while they were adopted in the Netherlands and Switzerland. His answer is that Lacanian theories, very influent among French psychiatrists, led psychiatrists to the diagnosis that drug addiction results from a disruption of social relations. Hence they drew the conclusion that drug addicts should be cured by a methodical restoration of their social relations. By contrast, the Swiss or Dutch psychiatrists did not think they had at their disposal a theory from which effective remedies against drug addiction could be drawn. As they did not feel able to cure drug addiction, they believed that their main duty was, to protect people, notably against the threat of AIDS involved in drug addiction. The French had strong reasons to stress that their role was to cure; the Dutch and Swiss to believe that their role was to protect, since they did not know how to cure. In an analysis such as this one, one retrieves the very flexible theory of rationality used by individualist sociologists following the Weberian tradition.

2.1 A Particular Version of the Individualistic Tradition: The Rational Choice Model

Contemporary sociology has developed a particular version of the individualistic tradition. Instead of the Weberian open theory of rationality, it proposes to use the theory of rationality developed by economics. According to this theory, social actors are supposed to be moved essentially by their interests and to be able to pick up the best means to follow their ends. This model—called the 'rational choice model'—the inspiration of which goes back to Bentham, is very useful. It has produced many fruitful theories in several fields: notably the sociology of crime, of education, of organizations. These achievements have motivated some sociologists, such as James Coleman, to contend that the 'rational choice model' could play in sociology the role of a general theory able to integrate heterogeneous findings. But this model cannot be considered as a realistic theory of behavior in all circumstances. Olson (1965) has stressed that his own theory of collective action, which applies literally the postulates of neoclassical economics, can explain that collective action does not occur as long as its axioms can be considered as realistic. But, he adds, in many circumstances, they cannot, simply because people can be moved and mobilized, not only by their interests, but also by their beliefs. Now, beliefs can seldom be satisfactorily explained merely by the interests of social actors. On the whole, the 'rational choice model' has produced many convincing theories, but it cannot be considered as a general theory, neither actually nor potentially.

The styles of research illustrated by the previous examples have in common that they provide explanations of social phenomena that follow the criteria generally used in all scientific disciplines. In that sense, they belong to the 'positivist' category of the previous typology.

3. Interpretation vs. Explanation

In his book on the methodology of the social sciences, *Problems of the Philosophy of History*, Simmel (1892) made the important point that history, as well as the social sciences, raises naturally two types of questions which respectively can and cannot be given a unique scientifically controlled answer. In Rickert's vocabulary, some questions raised by sociologists aim at 'explaining,' others at 'interpreting' social phenomena. Thus, Simmel (1900) himself proposed in his *Philosophy of Money* an *explanatory* theory of the decline of feudalism: From the moment when the peasants of the Middle Ages could pay their rents in money rather than in kind, they could decide what they would grow, they started to be in a position to create and sell surpluses, and were eager to participate in collective actions against landowners, so that on the whole this apparently minor institutional innovation generated a chain reaction which powerfully eroded the feudal organization of society.

But other questions call for '*interpretation*' (or '*description*') rather than 'explanation': what are the main features of modernity? Is the twenty-first century going to be characterized by a religious revival? What does it mean to be poor in an affluent society? How do

people live in poor suburban areas? Of course, such questions cannot be answered only in one way. There are many ways of describing the conditions of life in a poor suburban area. In other words, the answers to such questions are doomed—by the very nature of the questions—to be for one part subjective. Actually, as Simmel states, such questions can be answered in an interesting fashion, only if analysts are able to inject their own sensibility in their description or interpretation. Answering such questions requires on the whole a type of ability that makes sociologists close to artists. A good biographer, for instance, cannot follow exclusively the criteria defining a scientific work. He or she must also be able to give a unity to the numerous observations which have been collected by witnesses on the biography's subject, to that person's behavior and decisions in innumerable circumstances. A number of sociological works are faced with the same type of questions. Giddens's (1984) work on postmodern society, as earlier Burckhardt's work on the *Kultur der Renaissance in Italien* ([1860], 1952), are examples of works aiming at characterizing the main features of historical eras or identifying broad historical trends. Other works dealing, not with societies as wholes, but with local phenomena, belong also to the interpretative-descriptive category rather than to the explanatory one; though they deal with social reality, they evoke the procedures of novelists. Oscar Lewis' *The Children of Sanchez* (1961) is a classical illustration of this 'literary' sociology. The anthropologist Clifford Geertz (1973) has coined the expression 'thick description' to characterize the methodology of this type of work. Runciman (1983–97) has developed methodological considerations which can be considered as elaborating on Simmel's insights.

4. *A Complex Identity*

In summary, all the types of sociology which can be derived by the combination of the three main ideal-typical objectives (explaining, informing, criticizing) and of the four types generated by the combination of the binary variables holism vs. individualism and 'positivism' vs. 'hermeneutics' can be illustrated as easily by examples taken from contemporary as from classical sociology. This shows perhaps that all the types are legitimate and fruitful. The cost of this heterogeneity lies in the difficulty of perceiving the identity of sociology, its benefit in the fact that it makes sociology able to fill the gaps that other social sciences, such as economics, more devoted to a single paradigm, are unable to fill.

See also: Action, Theories of Social; Critical Theory: Contemporary; Critical Theory: Frankfurt School; Durkheim, Emile (1858–1917); Hermeneutics, Including Critical Theory; Historical Sociology; Marx, Karl (1818–89); Methodological Individualism in Sociology; Parsons, Talcott (1902–79); Phenomenology in Sociology; Positivism: Sociological; Rational Choice Theory in Sociology; Sociology, Epistemology of; Sociology, History of; System: Social; Theory: Sociological; Weber, Max (1864–1920)

Bibliography

Adorno T W, Dahrendorf R, Albert H, Popper K R [1969] 1972 *Der Positivismusstreit in der deutschen Soziologie*. Luchterhand, Neuwied, Germany

Bergeron H 1999 *Soigner ou prendre soin des Toxicomanes: Anatomie d'une Croyance Collective*. Presses Universitaires de France, Paris

Burckhardt J [1860] 1952 *Die Kultur der Renaissance in Italien*. Kröner, Stuttgart, Germany

Cusson M 1990 *Croissance et Décroissance du Crime*. Presses Universitaires de France, Paris

Durkheim E 1895 *Les Règles de la Méthode Sociologique*. Alcan, Paris

Durkheim E 1897 *Le Suicide, Étude Sociologique*. Alcan, Paris

Foucault M 1975 *Surveiller et Punir; Naissance de la Prison*. Gallimard, Paris

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Giddens A 1984 *The Constitution of Society*. Polity Press, Cambridge, UK

Lazarsfeld P F, Sewell W H, Wilensky H L (eds.) 1967 *The Uses of Sociology*. Basic Books, New York

Le Play M F 1855 *Les Ouvriers Européens*, 2nd edn. Imprimerie Imperiale, Paris

Lewis O 1961 *The Children of Sanchez, Autobiography of a Mexican Family*. Random House, New York

Merton R K 1949 *Social Theory and Social Structure*. Free Press, Glencoe, IL

Mommsen W 1965 Max Weber's political sociology and his philosophy of world history. *International Social Science Journal* **17**(1): 44

Nisbet R A 1966 *The Sociological Tradition*. Basic Books, New York

Olson M Jr 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA

Runciman W G 1983–97 *A Treatise on Social Theory*. Cambridge University Press, Cambridge, UK

Simmel G 1892 *Die Probleme der Geschichtsphilosophie*. Duncker & Humblot, Leipzig, Germany

Simmel G 1900 *Philosophie des Geldes*. Duncker & Humblot, Leipzig, Germany

Tocqueville A de [1856] 1986 *L'Ancien Régime et la Révolution*. Levy, Paris

Weber M 1920–1 *Gesammelte Aufsätze zur Religionssoziologie*. Mohr, Tübingen, Germany

Weber M 1951 *Aufsätze zur Wissenschaftslehre*. Mohr, Tübingen, Germany

R. Boudon

Solidarity: History of the Concept

The concept of solidarity epitomizes the total process through which sociology established itself among the social sciences. An originally juridical concept, per-

taining to the character of legal obligations, was given a new social and political meaning which made it tantamount to the theoretical conditions of possibility for the constitution of sociology as such. In order to do this, the concept of solidarity had to take a distance from other close concepts, such as compassion and fraternity, and to develop the specific features which would ensure its role. Throughout the nineteenth century, it expressed the search for a new articulation between individualism and collectivism which was at the origins of crucial transformations, in social theory as well as in social policies.

1. Compassion, Fraternity, Solidarity

Compassion had inspired the treatment of the poor in the Old Regime society, through private charity mixed with police repression. The inadequacy of such a treatment to a modern liberal order, from the vantage point of equality of rights as well as of a new way of production requiring the active participation of everybody, became clear in the early stages of the French Revolution (Procacci 1993). Against the fortuitous nature of private charity, solidarity openly marked its own public vocation. Instead of relying upon an occasional feeling or the practice of a religious piety, it was to become the principle of a secular organization framing the competition of individual interests according to the needs of the common good.

'Fraternity' was one of the key-words of the revolutionary tradition in France, together with equality and liberty, but proved to be much more difficult to fix in normative terms. Moreover, during the new 1848 revolution, fraternity inspired popular claims for a social and economic reform, engaging with the socialist tradition. Its revolutionary implications made more and more clear that, although fraternity was an integral part of the new legal system, it could not offer guidelines for public action—a role rather played by solidarity (Borgetto 1992). Solidarity presented the advantage of being a more 'scientific,' positive notion, interpreting more effectively a concern about social order. Fraternity had offered the first foundation to solidarity, well before the latter, toward the end of the nineteenth century, met the idea of social interdependence developed by sociology. Nevertheless, in French public law it eventually became restricted to the link among members of the nation, while solidarity was rather entrusted to regulate a system of mutual rights and duties organized by the state.

As a matter of fact, the concept of solidarity inherited the same political space covered by fraternity, having to respond to the need of governing the persistence of inequality within a society of equals. Solidarity took its social meaning within a general framework deeply marked by the establishment of modern individual rights—the process of liberal individualism. Individual rights entering into the new social reality, through a twofold process of emancipation, juridical and economic, did not offer criteria for organizing the life of the collectivity. Freed from the revolutionary echo resounding through fraternity, solidarity pointed at that sort of third way between liberal individualism and socialist collectivism that offered the proper room for the rise of sociology. The concept promised to reconcile individual independence and social cohesion, and on this promise built its own fortune.

2. Rights and Duties

Robert Nisbet (1966) correctly highlighted the contribution of conservative antirevolutionary authors to the social concept of solidarity, through the emphasis they put upon communitarian ties. De Bonald, for instance, denounced the revolution for having destroyed any collective order and therefore any possibility for solidarity. While sharing his insistence on the need for a systematic doctrine about the bases of social order, Saint-Simon fundamentally disagreed with his attempt to ground solidarity within religious thought. The recovery of social unity in a modern society required a new attitude, and foundations were to be searched within science. In fact, with Saint-Simon and Comte the reflection about solidarity clearly engaged in a criticism of the difficulty that legal and economic individualism opened to solidarity. From this point of view, the rise of sociology marked a discontinuity with respect to patterns of solidarity belonging to the past.

For Auguste Comte ([1830–42] 1975) the specific task of a social science was to oppose to the destructive strength of individualism a principle of social organization based on the authority of science. The revolution was the result of individualism's incapacity to provide criteria useful to organize any collective order; the only way to stop the revolution was to be able to remove its causes. To problems raised by the emancipation of individual rights, sociology must respond by providing concrete principles for social reorganization. Comte proposed to replace the emphasis on individual rights with a system of mutual duties, linking everybody to each other, therefore able to promote the solidarity of all parts within the social system. The concept of duty appeared to him directly social, in other words an organizer: It put at the forefront 'the intimate feeling of social solidarity,' the priority of common interest over individual ones. It was thus to become the key for a new politics, focused upon the search for the best way to combine the efforts of everybody, rather than upon the competition of interests. This was the meaning of 'order and liberty,' the motto of the *Société Positiviste* that Comte created as a response to the 1848 revolution. The imperative was to reorganize society, the essence of social

phenomena was organic; the concept of duty was to become the principle of an active morality, overcoming the opposition between individual and collective.

Along these lines, throughout the nineteenth century solidarity deepened its distance from fraternity. By the end of the century, the sociological definition of solidarity found the support of scientific discoveries: Microbiology and microphysics confirmed the interdependence of all parts of an organism under the form of a pathological contamination, giving to an old idea the new strength of a scientific evidence. If each single element depends upon the others, then their solidarity with the whole becomes a fact. Solidarity only refers to the whole, it concerns the community, and does not distinguish among categories of members. As such, solidarity designates a twofold bond, through the space within the community, and over time between generations.

3. *Organic Solidarity*

After the dramatic resurgence of revolution with the 1871 Commune, the concept of solidarity was at the core of the Third Republic political response to the conflict between two opposite views of the role of the state proposed by liberalism and socialism (Donzelot 1984). In the name of market economy, liberals opposed any intervention of the state in the solution of social problems, even at the risk of an objective alliance with conservatives' defense of the role of family and traditional groups. Socialists, in turn, regarded state intervention as necessary. Republicans stayed in between, but without any clear theory about the role of the state were forced to a negative position; solidarity offered them a positive principle for their politics, to the point of becoming an 'official philosophy' (Scott 1951).

It expressed a peculiar conception of social relations at the basis of a new social contract founded not merely on individual rights, but on the need for a collective reparation of social injustice. Alfred Fouillée (1885), the founder of the so-called *solidarist philosophy*, insisted on the Comtian description of society as a system of mutual obligations, as the source for an 'active repairing justice' more suitable to a democratic society than a liberal negative justice. His idea of a *contractual organism* found its theoretical foundation in the concept of solidarity and at the same time the key for a new set of concrete political measures transforming the system of social protection (Bec 1998).

Solidarity was not only the source of a philosophy, but became the core of scientific sociology with the seminal work of Emile Durkheim ([1893] 1978) on the division of labor. His attempt was to give a scientific foundation to the specific nature of social phenomena; solidarity conceptualized the principle of cohesion which only could transform a set of individuals into something like a society. Individual consensus could not explain society; it depended upon an objective cause underlying the characters of the social environment. Through his distinction between a mechanic, indifferentiated solidarity, and an *organic solidarity*, Durkheim defined the bases for a social organization where individual rights and state's ends did not exclude each other. From one model to another, the nature of solidarity changed, but it still worked as the foundational law of any society. A modern democratic society depended upon the level of division of labor: Individual functions must be specialized, but conversely their cooperation for a final end ought to be ensured. The division of labor was the pattern of the only kind of solidarity suitable to a differentiated society. Far deeper than individual freedom of contract, the organic solidarity organized functional interdependence of all parts with respect to the whole.

Durkheim was interested in the moral value of the division of labor and what consequences this would have from the point of view of society and individuals. It increased social dynamics and mutual dependence, but would it produce solidarity? If labor was badly distributed, then the priority of the whole over the individual would not be ensured. The intensity of social conflicts depended upon a lack of perception of solidarity: The problem thus was not in the structure, but in the representation of internal social bonds, and so was ultimately a political problem. This is the role of the modern democratic state: to organize solidarity through a right balance of social cohesion and individual liberty, constantly enhancing its perception.

The theory of solidarity established that societies live and change according to their own laws, they could neither be frozen in their past forms, nor be violated in their natural course. To acknowledge these laws became thus an indispensable step for politics, while promoting sociology's role. As such, solidarity offered the basis for grounding political reforms on the scientific knowledge of these laws, intentionally adopted by a *politics of solidarity*. From the 'natural fact' of interdependence, the principle of solidarity implied that everybody had to be guaranteed from some common risks, and reciprocally that all must do their part in financing common protection according to their means. On the idea of shared social risks, a whole set of solidarity techniques were developed, shaped above all on an insurance model (Ewald 1986). Welfare and pension laws followed at the beginning of the twentieth century in most European countries, based on the idea of a social debt toward the less privileged and their corresponding right to support. Techniques of solidarity were at the origins of modern welfare systems and of more general political strategies of socialization.

The very nature of solidarity, a concept born at the crossing point between liberal individualism and socialism, makes it particularly relevant today (Chevallier 1992), when we face again the need to

define a new balance between those two principles, not so differently from what gave substance more than a century ago to the concept of solidarity. The whole debate nowadays within social sciences on the future of welfare is there to prove it.

See also: Collectivism: Cultural Concerns; Communitarianism: Political Theory; Communitarianism, Sociology of; Comte, Auguste (1798–1857); Durkheim, Emile (1858–1917); Individualism versus Collectivism: Philosophical Aspects; Integration: Social; Liberalism: Historical Aspects; Marxist Social Thought, History of; Political Thought, History of; Socialism: Historical Aspects; Solidarity, Sociology of

Bibliography

Bec C 1998 *L'assistance en démocratie*. Belin, Paris
Borgetto M 1992 *La notion de fraternité en droit public français*. Librairie générale de droit, Paris
Chevallier J (ed.) 1992 *La solidarité: un sentiment républicain?* Presses Universitaires de France, Paris
Comte A [1830–42]1975 In: Enthoven J-P (ed.) *Physique sociale. Cours de philosophie positive*, 2 Vols. Hermann, Paris
Donzelot J 1984 *L'invention du social*. Grasset, Paris
Durkheim E [1893]1978 *De la division du travail social*, 10th edn. Presses Universitaires de France, Paris
Ewald F 1986 *L'Etat providence*. Fayard, Paris
Fouillée A 1885 *La science sociale contemporaine*, 2nd edn. Hachette, Paris
Nisbet R 1966 *The Sociological Tradition*. Basic Books, New York
Procacci G 1993 *Gouverner la misère. The social question in France*. Seuil, Paris
Scott J A 1951 *Republican Ideas and the Liberal Tradition in France, 1870–1914*. Columbia University Press, New York

G. Procacci

Solidarity, Sociology of

Solidarity refers to the binding, or melding, of individuals into a cohesive group or collectivity. As such, it is an emergent attribute of groups that facilitates social coordination and social order, and is a precondition to all nonspontanteous collective action.

Since social order and collective action are the alpha and omega of classical sociological theory, solidarity has always been a central concept in sociology. The concept came to the fore following the changes wrought by the development of industry and the rise of market economies. Although they were unjustified in doing so, most classical social theorists regarded the cohesion of agrarian communities as unproblematic. In their stylized view, such communities were technologically and demographically stagnant and uninvolved in long-distance trade.

Given these assumptions and barring exceptional circumstances—such as those resulting from wars or massive epidemics—social mobility would be minimal and most children would be destined to recapitulate the social lives of their parents. As agrarian communities tended to offer relatively little scope for individualism, the attainment of solidarity was viewed as relatively unproblematic. According to Ferdinand Tönnies (1887), such communities (which he termed *gemeinschaften*) were breeding grounds for social relations based on deep emotional, quasifamilial commitments. Emile Durkheim (Durkheim 1893) referred to the solidarity characteristic of this kind of society as 'mechanical,' implying thereby that it had a certain automatic quality. Mechanical solidarity rests on a set of common values, internalized norms, and beliefs (a *conscience collective*) engendered by physical co-presence, a common focus of attention, and a sense of mutual awareness. This kind of solidarity is often expressed in a religious credo.

Mechanical solidarity appeared to be threatened, however, by the rise of industry and the expansion of market economies in Western Europe, North America, Australasia and Japan beginning in the late eighteenth century. Among their many other effects, these transformations increased the size and scope of social networks, thereby affording many individuals greater options, ranging from marriage partners to occupations, in their daily lives. The growth of individualism and the concomitant decline of a *conscience collective* was the inevitable result (Simmel 1922). It was far from evident how groups, communities, and societies could maintain mechanical solidarity in the wake of this increasing individualism. Durkheim, the foremost early theorist of solidarity, initially argued that industrial societies were held together by individuals' mutual functional interdependence—a form of solidarity that he termed 'organic.' At the end of his life, however, he came to regard common values and norms as the basis for solidarity in all kinds of groups and societies (Parsons 1937).

Durkheim's insistence that solidarity requires compliance to a set of collective obligations based on particular values and norms has greatly influenced subsequent research. For example, a religious group is solidary only to the degree that its adherents comply fully with its obligations, both sacred (rites) and secular (tithing). Likewise, a working class is solidary only to the degree that individual workers regard themselves as members of a collectivity (e.g., the working class) with its own obligations, and refrain from engaging in activities—such as strikebreaking—that compromise or countermand these obligations.

Yet compliance to collective obligations can owe to causes other than solidarity. Consider the mobilization of labor in a profit-making firm. To attain maximum profit, firms demand that their employees perform to

the best of their ability and refrain from shirking. Employees usually comply with the firm's obligations, not out of a sense of solidarity but because they are compensated for doing so in wages or salaries. In contrast, the members of a religious group or a working class are not given pecuniary compensation to comply with their corporate obligations. On this account, it is likely that compliance in such groups is due to solidarity. Since compliance with corporate obligations can be attained via mechanisms other than solidarity, conducting empirical research on solidarity is inherently difficult.

This difficulty has not inhibited theorists, however. Theoretical perspectives on solidarity may usefully be divided into three distinct types. *Normative* perspectives suggest that solidarity arises from common values and norms, *structural* perspectives argue that it arises from common material interests, and *rational choice* perspectives hold that it arises from mutual interdependence.

1. *The Normative Perspective*

Normative theorists claim that solidarity is unlikely to result merely from the interaction of rational, self-interested individuals (Parsons 1951). This is because it is never in the interest of rational egoists to adhere to collective obligations if they can profit by ignoring them. In a world of rational egoists, the only means of assuring the requisite compliance is by sanctioning. Normative theorists hold that sanctioning is both too costly and too unreliable a means of assuring high levels of compliance, especially among large numbers of individuals. For them, values and internalized norms make individuals fit for social life by removing the conflict between the individual's interest and that of the relevant collectivity. If individuals adhere to the values and norms that underlie corporate obligations, they will tend to engage in the obligatory actions without the carrot of rewards or the stick of sanctions. According to normative theorists, socialization is the key to the attainment of solidarity. Solidarity therefore is most likely to emerge in socially homogeneous collectivities. It should be noted that the emergence of values and norms is exogenous to the theory, as are socialization processes. Moreover, although casual observation confirms that compliance with values and norms is never automatic, normative theory has been mute about the conditions under which the requisite compliance is likely to occur.

2. *The Structuralist Perspective*

Structural theorists typically, if implicitly, regard solidarity as an outcome of individually rational action. On their view, rational individuals form solidary collectivities principally by virtue of sharing common material interests. They discover their common interests in the course of mutual interaction, especially when they perceive themselves to be threatened by powerful antagonists (Marx 1972 [1845–46], p. 179). When mutual interaction occurs repeatedly, this may set the stage for the development of sentiments, or affective feelings, that further strengthen compliance to collective obligations (Collins and Hanneman 1998). However, the weaknesses of structural theories of solidarity are notable. Both commonality of interest and repeated interaction are insufficient causes of solidarity. Solidarity varies widely among workers and married couples, for example. Sometimes workers refuse to support striking colleagues, and the incidence of divorce has risen dramatically in nearly all industrial societies. Neither of these trends is easily accounted for by structural theories of solidarity. Further, if the individuals who share interests are rational egoists, as most structural theorists contend, then why would they ever join forces to pursue their common ends?

3. *The Rational Choice Perspective*

This question is the point of departure for rational choice perspectives of solidarity. Truly rational actors will *not* join a group to pursue common ends when, without participating, they can reap the benefit of other people's activity in obtaining them. Instead of contributing to the common weal, rational actors will free ride (Olson 1965) and behave individualistically when it is expedient. The upshot is minimal solidarity despite an extensive commonality of interest. The rational choice theory of solidarity (Hechter 1987) hinges on precluding free riding. In this theory, rational individuals form groups to obtain goods that they either cannot produce on their own, or that they cannot attain as economically. The ultimate motivation for group formation is held to be the individual's desire to gain access to jointly produced goods. These goods can only be produced if prospective members comply with the rules and obligations that are necessary to ensure production of the given joint goods. The more dependent individuals are on the group for access to valued goods, the greater its potential power over them. This dependence increases when the given goods are not readily available elsewhere, when members lack information about alternatives, when the costs of leaving the group are high, and when personal ties to other members are strong. Moreover, the greater the dependence of members on a group, the greater their willingness to comply with highly extensive obligations. Hence, dependence creates conditions not just for the emergence of corporate obligations, but for obligations that guide and regulate behavior to a high degree.

Yet dependence is an insufficient cause of compliance. Children tend to be highly dependent on their parents, but this is no guarantee that they will routinely perform their chores, or study diligently in school. Dependence by itself is unlikely to resolve the free rider problem. Free riding can only be precluded when the given group has a high control capacity. This control capacity, in turn, is a function of monitoring and sanctioning. Monitoring is the process of detecting noncompliance to corporate obligations; sanctioning is the use of rewards and punishments to induce compliance. The theory therefore holds that compliance to corporate obligations—an indirect measure of solidarity in noncompensatory groups—is jointly a function of dependence and control mechanisms. That said, this theory remains incomplete because the sources of individual dependence are exogenous to it.

4. *Future Research*

The empirical study of solidarity has been hamstrung by a lack of agreement on definitions of the concept. Defining solidarity as an intersubjective state (as normative theorists do) poses a severe challenge for empirical research. This is because there are as yet no consensually accepted measures of individual values and internalized norms in social science; direct measures of the sentiments or feelings that may hold within collectivities perforce are unavailable. Most empirical research on solidarity therefore focuses on explaining variable compliance with corporate obligations. When this research fails to pay sufficient heed to the pitfalls of inferring an intersubjective state from behavior (see above), its results are questionable.

Future research on solidarity is likely to proceed along both methodological and substantive fronts. Methodologically, new measurement instruments some day may make it possible to gauge solidarity directly as an intersubjective state, rather than using behavioral compliance as a proxy. Substantively, there is much to learn about the interaction of solidary groups in wider social systems. Since solidarity is built around a set of normative obligations and enforcement mechanisms, it follows that the normative content of any given solidary group may be more or less congruent with that of its global social environment. Thus minority ethnic groups often espouse norms at variance with their host society. Adolescent peer groups evolve norms that are mutually inconsistent, and generally inconsistent with school norms. Factory workers can embrace output restriction norms that are at variance with their employers' expectations, and inner city gangs and organized crime groups establish norms that are at variance with those of the society at large. The effects of solidarity in such countercultural groups are a matter of dispute among both social scientists and political decision makers (see *Social Capital*). Further research is needed to clarify the mechanisms that are responsible for the emergence of such countercultural groups and the consequences these groups have on social order.

After a long period of dormancy, interest in solidarity has increased in several social science disciplines. The rational choice theory of solidarity has been used to explain variations in legislative party voting, the membership rules of friendly societies and mutual benefit associations, the survival of US intentional communities, cross-national crime rates (Hechter and Kanazawa 1993), social order in Japan and the United States (Miller and Kanazawa 2000), and the prevalence of nationalism in world history (Hechter 2000). Whether group solidarity incites or inhibits intergroup conflict is now the subject of lively debate (Bhavnani and Backer 2000, Fearon and Laitin 1996, Gould 1999). Legal theorists (Ellickson 1991, Posner 2000) are exploring group solidarity as an institutional alternative to law as a means of dispute resolution. Finally, a variety of mathematical models of solidarity recently have been proposed by a group of social and physical scientists (Doreian and Fararo 1998).

See also: Action, Collective; Communities of Practice; Community, Expression of; Community, Social Contexts of; Community Sociology; Community Studies: Anthropological; Durkheim, Emile (1858–1917); Integration: Social; Networks and Linkages: Cultural Aspects; Rational Choice Theory in Sociology; Societies, Types of; Solidarity: History of the Concept; Theory: Sociological

Bibliography

Bhavnani R, Backer D 2000 Localized ethnic conflict and genocide: Accounting for differences in Rwanda and Burundi. *Journal of Conflict Resolution* **44**: 283–306

Collins R, Hanneman R 1998 Modelling the interaction ritual theory of solidary. In: Doreian P, Fararo T J (eds.) *The Problem of Solidarity: Theories and Models*. Gordon and Breach, Amsterdam, pp. 213–38

Doreian P, Fararo T J (eds.) 1998 *The Problem of Solidarity: Theories and Models*. Gordon and Breach, Amsterdam

Durkheim E 1893 *The Division of Labor in Society*. Free Press, Glencoe, IL

Ellickson R C 1991 *Order Without Law: How Neighbors Settle Disputes*. Harvard University Press, Cambridge, MA

Fearon J D, Laitin D D 1996 Explaining interethnic cooperation. *American Political Science Review* **90**: 715–35

Gould R V 1999 Collective violence and group solidarity: Evidence from a feuding society. *American Sociological Review* **64**: 356–80

Hechter M 1987 *Principles of Group Solidarity*. University of California Press, Berkeley, CA

Hechter M 2000 *Containing Nationalism*. Oxford University Press, Oxford, UK

Hechter M, Kanazawa S 1993 Group solidarity and social order in Japan. *Journal of Theoretical Politics* **3**: 455–93

Marx K 1845–46, 1978 The German ideology. In: Tucker R *The Marx–Engels Reader*. 2nd edn, W W Norton, NY, pp. 146–200
Miller A S, Kanazawa S 2000 *Order by Accident: The Origins and Consequences of Conformity in Contemporary Japan*. Westview Press, Boulder, CO
Olson M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA
Parsons T 1937 *The Structure of Social Action*. McGraw-Hill, New York
Parsons T 1951 *The Social System*. Free Press, Glencoe, IL
Posner E A 2000 *Law and Social Norms*. Harvard University Press, Cambridge, MA
Simmel G 1922 The web of group affiliations. In: Simmel G (ed.) *Conflict and the Web of Group Affiliations*. Free Press, New York pp. 125–95
Tönnies F 1887 *Community and Society*. Transaction Books, New Brunswick, NJ

M. Hechter

Somatoform Disorders

The history of somatization phenomena, the 'borderland between medicine and psychiatry' (Lipowski 1988), or in terms of neuroscience the problem of the 'body–brain-interface' (Guggenheim 2000), can be traced back to ancient Greek medicine, where the term 'hysteria' was used to describe patients' somatic complaints without the presence of somatic disease. Hysteria (from hystera = womb) was understood as a psychological problem in women, whose intrapsychic conflicts, resulting from the unfulfilled wish to bear a child, lead to somatic complaints (see *Hysteria*). Similar clinical observations by different European clinicians since the mid-nineteenth century (Briquet, Charcot, Janet, Breuer) were antecedents of the Freudian concept of conversion disorder. The psychodynamic etiology and psychoanalytic treatment approach to patients with conversion disorders, psychosomatic complaints, and hysteria was most influential to the 1970s. Nowadays, the term somatoform disorders replaces the traditionally used, but not precisely defined terms of hysteria, 'functional symptoms,' and 'masked depression,' etc., that were mainly confounded with etiological (psychoanalytical) assumptions. The current concept of somatoform disorders encompasses a heterogeneous group of psychiatric disorders that are characterized by four core features: (a) patients suffer from enduring or recurrent bodily complaints, symptoms and/or pain (e.g., palpitations, headache, fatigue, and dizziness), which are experienced as indicators of somatic illness, (b) mainly these complaints are not due to physical or organ dysfunction, (c) psychological and social factors—usually refused as influences by the patients—play a predominant role in the development and course of these disorders, and (d) patients experience a general preoccupation with bodily symptoms and health concerns.

This article describes the various somatoform disorders, their prevalence, and etiology. Critical aspects regarding diagnosis, treatment, and management of somatoform disorders in healthcare systems will be discussed.

1. *Classification and Epidemiology*

From a nosological perspective, somatoform disorders were grouped together in the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III*) in 1980 for the first time. However, this clustering was primarily based on clinical criteria and not on empirically derived or laboratory findings. The tenth edition of the *International Classification of Diseases* (*ICD-10*) includes most of the somatoform disorders of the recent *DSM-IV* (American Psychiatric Association 1994), except for conversion disorders that are grouped with the dissociative disorders. The main feature of these disorders is the 'repeated presentation of physical symptoms with persistent request for medical investigations, in spite of repeated negative findings and reassurances by doctors that these symptoms have no physical basis. If any somatic disorders are present, they do not explain the nature and the extent of the symptoms or the distress and preoccupations of the patient' (WHO 1991). For the diagnosis of all somatoform disorders the patients' symptoms have to be of sufficient severity to warrant clinical attention. The *ICD-10* differentiates between four main nosological entities (seven different diagnoses) which were characterized by different diagnostic criteria (see Table 1). Recurrent, multiple somatic complaints that are not associated with any physical disorder are the essential feature of somatization disorder. The diagnosis requires a history of many bodily symptoms of several years' duration with onset before the age of 30. In somatoform autonomic dysfunctions symptoms of autonomic arousal (e.g., palpitations, sweating, and stomach discomfort) are the main complaints of the patients. Chronic pain syndromes (duration > 6 months) like tension headache, lower back pain, and atypical facial are labeled as persistent somatoform pain disorder when emotional conflicts or psychosocial problems are involved as main influences. Psychological factors can play either a major role within a general medical condition (e.g., rheumatoid arthritis) or may be the only determinants of pain (e.g., not otherwise specified lower back pain). The pain in one or more anatomical sites is the main focus of clinical presentation.

Conversion disorder is mostly a transient disturbance of physical functions (typically occurring in stressful situations) which do not conform to current

Table 1
Somatoform disorders and their characteristics

Subtypes	Characteristics and (selected) main criteria (*ICD-10*)
Somatization disorder	• a history of at least 2 years of multiple and variable complaints which cannot be better explained by somatic factors or other psychiatric disorders (A) • preoccupation with the symptoms causes persistent distress and leads to seek repeated medical consultations (B) • persistent refusal to accept medical reassurance that there is no adequate physical cause (C) • a total of six or more enduring autonomous symptoms from different groups (gastrointestinal, cardiovascular, genitourinary system, skin and pain symptoms) (D)
Hypochondriasis	• a persistent belief of the presence of a serious physical disease (A1) • a persistent preoccupation with a presumed deformity or disfigurement (body dysmorphic disorder) (A2)
Somatoform autonomic dysfunction	• Symptoms of autonomic arousal attributed by the patient to a physical disorder (heart and cardiovascular system/upper and lower gastrointestinal tract/respiratory or genitourinary system) (A) • Two or more autonomic symptoms like palpitations, sweating, dry mouth, flushing or blushing, epigastric discomfort etc. (B)
Persistent somatoform pain disorder	• Persistence of severe and distressing pain for at least 6 months which cannot be explained adequately by a physiological process or physical disorder (A)

concepts of the anatomy and physiology of the nervous system. Many disorders seem to be neurological disorders (e.g., motor disorders or disturbances of sensory function), but do not show the usual pathological signs. Hypochondriasis is not exclusively a secondary phenomenon within affective, anxiety, or psychotic disorders but a diagnostic subgroup of its own. Patients are preoccupied with their belief of already suffering from a severe illness or the fear of becoming ill with a life threatening disease (e.g., cancer, stroke, or AIDS). The diagnostic congregation of body dysmorphic disorder—which is rarely seen as a full-blown syndrome—is different in the nosological systems. In the *ICD-10* it is a subgroup of hypochondriasis, with the preoccupation on a presumed deformity or body abnormality. Patients subjectively feel unattractive, ugly, and disfigured despite lacking evidence for major defects in appearance or stigmata. Positive feedback from significant others about normality in appearance has no corrective effects.

Somatoform disorders show high prevalence rates in the community, especially in the primary health care system. In industrialized societies up to 4–12 percent of all people suffer from somatoform disorders (Escobar et al. 1987, Simon and von Korff 1991). In medical outpatient and inpatient treatment, patients with the probable diagnosis of somatoform disorders are estimated at up to 25 to 30 percent. The onset is in adolescence or young adulthood, and the course in untreated cases is usually chronic with a waxing and waning of symptoms. Women, persons from lower social classes, and ethnic subgroups are more likely to develop somatoform disorder. Furthermore, somatization is more often seen after separation and divorce. Somatoform disorders are often associated with other psychiatric disorders, especially depression, anxiety disorders, substance abuse disorders, and personality disorders (see *Depression*, *Anxiety and Anxiety Disorders*, and *Personality Disorders*). Patients with somatoform disorders are often not correctly recognized by their doctors and therefore insufficiently treated. Also, the latency between onset of the disorder, the appropriate diagnosis, and the referral to adequate treatment is usually high. These factors lead to largely extended costs for medical and psychosocial services (nine-fold costs of care than media costs of all patients).

2. *Etiology and Course*

No single factor can explain the pathogenesis of somatoform disorders. Comparable with other psychiatric disorders, biological, psychological, and social factors interact in complex ways (see *Mental Illness, Etiology of*). Their impact probably varies over time within the same individual and between different patients resulting in a common final pathway. There is no general etiological model for somatization phenomena (Pilowsky 1997). However, interesting preliminary data demonstrate that patients with somatization disorder may have an abnormality in cortical and neuropsychological functioning, e.g., bifrontal impairment of the hemispheres, abnormal

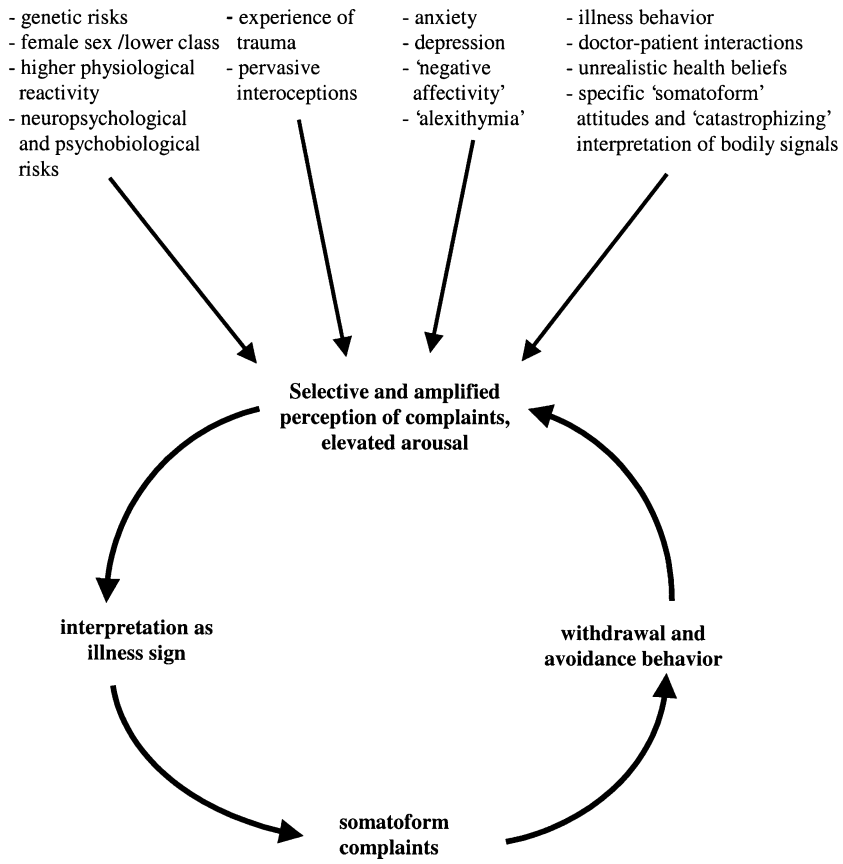

Figure 1
Risk model for somatoform disorders (adapted from Rief and Hiller 1998)

auditory-evoked potentials, and elevated levels of cortisol (Guggenheim 2000, Rief and Nanke 1999). Several etiological conditions play a prominent role, especially in maintaining the process and can be considered as empirically derived risk factors for somatoform disorders (see Fig. 1).

(a) Genetic risks, e.g., higher rates of somatoform disorders for monozygot than for dizygot twins, and in first-degree relatives.

(b) Higher physiological reactivity which can lead to pervasive interoception of physiological signs. Unspecific bodily sensations are perceived and interpreted as symptoms of disease. The consequence is the intensification of the complaints and elevated attention to them.

(c) Biographic vulnerability, a high proportion of physical, sexual abuse (within the family), and traumatic experiences like war and natural catastrophes are over-represented in these patients. Furthermore, chronically ill parents or siblings, and learning by model within the family may influence the localization and types of somatoform symptoms.

(d) Personality traits, e.g., difficulties in the perception and expression of emotions with an inverse relationship of reduced emotional expressions and heightened psychophysiological arousal in stressful situations.

(e) Unrealistic health beliefs with a low tolerance for normal complaints, 'catastrophizing' ideas about physiological processes, and exaggerated expectations regarding effects of modern medicine to all kinds of complaints.

(f) Chronification factors like multiple diagnostic procedures, the overvaluation of normal variants through the caring physician (iatrogen factors) and positive (e.g., attention from others) or negative reinforcement (e.g., defense mechanism for negative emotions, advantages of the sick-role, withdrawal from responsibilities).

Which of these potential factors are important, and how much impact they have in the individual patient have to be determined within the diagnostic and psychotherapeutic process (see *Mental Illness, Etiology of*).

3. Diagnostic Process

The diagnostic process follows a sequence starting with a careful monitoring of all presented bodily complaints and ending up with a reliable and valid diagnosis of a special subtype of somatoform disorder and its severity. Knowledge of base rates for somatoform disorders and the potential medical disorders which have to be ruled out, age of the patients and their presentation style, history of other bodily complaints and excessive visits to other doctors might be hints for the assumption of the diagnosis.

The physician is confronted with the dilemma of possible false decisions: (a) they might diagnose somatoform disorder although the symptoms are better explained by a somatic disease which has to be treated by medical guidelines (type-1 error, 'false positive'); (b) the diagnosis of a somatic disease/dysfunction is assumed although the patient suffers from a somatoform disorder (type-2 error; 'false negative'). In most cases doctors and patients judge the risk of missing the correct somatic diagnosis as more adverse and therefore try to minimize it. However, this is one of the major factors maintaining somatoform disorders. A pragmatic rule for diagnosis is 'to have somatic diagnostic efforts as much as necessary but as less as possible.' Thorough, but not excessive or indefinite, medical examinations have to rule out organic disease and dysfunctions that might be causing the symptoms (e.g., hyperparathyroidism, multiple sclerosis, or brain tumor). It has to be supported by adequate and plausible procedures (e.g., physical and neurological examination, laboratory examination). Results and documents from other doctors should be revised in order to prevent repeated examination without relevance. A precise mental health examination for a complete overview of previous severe psychiatric disorders and minor complaints and the current psychopathological status is necessary. In single cases it can be difficult to decide whether the symptoms can better be explained by affective, anxiety, and personality disorders or whether they are real comorbid states. The sequence over time and the severity of symptoms are helpful in finding the correct hierarchy. Symptoms, pathological illness behavior, course, and consequence of the complaints will allow one to find the adequate nosological diagnoses along with the diagnostic criteria of *DSM-IV* or *ICD-10*. This diagnosis within the group of somatoform disorder has the status of a working hypothesis and is helpful for further treatment planning. If the patient visits a mental health specialist he or she will gather specific information regarding nosological diagnosis. A sophisticated analysis of all relevant information concerning the individual development, learning and predisposing factors, the onset, and further course of the symptoms, which is based on a life-chart interview, and information about actual psychosocial conditions, conflicts and traumata is inevitable. Of course it partly depends on the theoretical assumptions and models of the different psychotherapeutic schools (see *Differential Diagnosis in Psychiatry*).

4. Treatment

Since psychological factors are most influential in the development and course of somatoform disorders they have to be addressed in every effective treatment. Of course the setting, intensity, and content of psychological interventions can vary depending on factors like severity of symptoms, comorbidity, insight, and motivation of the patient and availability of psychotherapy. In a step-by-step approach the primary care physician should start with short-term verbal interventions. Acute and isolated somatoform symptoms may respond to psycho-education and counseling within the usual medical visits. Counseling and advice for general and easy-to-administer methods of symptom monitoring and management can be helpful. The physician should know the pitfalls in the treatment of patients with somatoform disorders and avoid procedures, which might reinforce the patients' pathological illness behaviors (e.g., no medical tests or laboratory examinations without clear indication). Information about and motivation for more specific forms of psychotherapy are essential if these basic interventions within the primary care system fail. For the more severe patients a referral to mental health experts is needed, for chronic and therapy-refractory cases even psychiatric or psychosomatic inpatient treatment may be necessary. Unspecific factors, which are important in any psychotherapy (e.g., therapeutic alliance, expectations of being helped, enhancing the sense of mastery, etc.) can be differentiated from specific therapeutic techniques and be adapted to relevant pathological processes in somatoform disorders (see below). Multimodal and multidisciplinary treatment approaches seem to be more effective than single treatment approaches.

A validating, accepting, and (especially in the beginning) nonconfronting patient–therapist relationship is a prerequisite for active participation in therapy. It is also the basis for the development of an alternative model of illness and the motivation to change. Self-monitoring diaries aim at a precise identification of symptoms in relation to external or interoceptive stimuli, events, and their positive and negative consequences in concrete situations. Psychoeducation and cognitive therapy are essential for a critical discussion and revision of inadequate concepts about health and disease and a better understanding of ongoing mechanisms. For example, the hypochondriac patient may realize the effects of a feedback loop leading to a vicious circle as is the case in increased autonomic arousal through anxious thoughts and worries about illness. In a collaborative way the therapist tries to convey a more realistic and helpful

paradigm of body–mind interactions which aims at a better understanding and mastery of symptoms. Relaxation training is provided as a method for an active reduction of tension, arousal, and apprehension. The patient is motivated to reduce maladaptive illness behaviors like body scanning and checking, excessive medical visits, seeking reassurance, and social withdrawal. Cooperation with physicians and family members is warranted so that they can avoid inadequate reactions to the patients' illness behaviors. Alternative behaviors, e.g., exercises, sports, and positive activities have to replace the symptomatic ones and can help to re-experience bodily functions in a more positive way. Stress management, social competence, and problem-solving interventions can be used optionally to improve coping with stress, conflicts, and interpersonal problems. Furthermore, therapy has to prevent relapses and enhance a more balanced life style. As some of the symptoms might resist the described interventions or take a longer time to be effective, patients have to learn how to 'live with the residual complaints as best as possible' and to regain a better quality of life instead of waiting for complete remission.

Up to now there is no empirically validated form of specific pharmacological therapy. Clinical experience shows that in some cases tricyclic antidepressants and selective-serotonin-inhibitors might be helpful, especially in patients with comorbid depression, anxiety disorders, and pain. However, without concomitant psychotherapy the risk of relapse is usually high. Benzodiazepines and low-potency neuroleptics should not be given because of the risks of long-term intake, especially abuse and dependency in benzodiazepines and tardive dyskinesia in neuroleptics.

Compared to other psychiatric disorders the empirical evidence for effective treatments is low in somatoform disorders. But there is a growing number of controlled outcome studies since the syndromes were clearly defined in 1980. These studies have to face the fact that most of these patients are not treated within mental health services but in the primary care system or in consultation and liaison services. Additional difficulties arise from the complexity of recent approaches, which can only be evaluated as 'total treatment packages' (see *Antidepressant Drugs*; *Behavior Therapy: Psychiatric Aspects*).

5. *Summary and Outlook*

The clear operationalization of diagnoses within *DSM* and *ICD* classification systems stimulated a lot of empirical studies concerning the epidemiology, etiology, pathogenesis, and treatment of somatoform disorders in the 1980s and 1990s. Medical and mental health services have realized the clinical relevance of these disorders and the need for early recognition and multidisciplinary intervention programs. Clinical observations and outcome studies have challenged the former statement, that patients with somatoform disorders in general have a poor prognosis. The acceptance of psychotherapeutic and psychiatric interventions can be enhanced if their positive effects are transparent to the patient. Convincing models for a better explanation and understanding of symptoms, the definition of limited and concrete goals, and the provision of successful experiences in symptom reduction and improving coping abilities may have the function of a 'positive vicious circle' and convey a sense of mastery.

However, patients and their caregivers have to overcome a series of obstacles:

(a) The first step is the correct recognition and early diagnosis by the primary care physician.

(b) Divergent health belief systems of patient and physician may burden the relationship and cooperation between them, so that it is difficult to realize psychological interventions.

(c) Although psychological treatment is indicated, physicians often avoid time-consuming verbal interventions or are not adequately trained for their application. However, it is not clear whether psychotherapeutic interventions, which are proven within a psychotherapeutic context, can be adopted by general practitioners without diminishing their efficacy.

(d) If physicians try to refer to mental health experts a lot of patients refuse this referral or are not motivated for psychotherapy.

(e) Since empirical validation of different treatment approaches for somatoform disorders is still rare, the selection of therapeutic methods mainly depends on local availability and personal preferences rather than on evidence-based criteria.

Each of these items is a challenge for clinical and research efforts as well as for a better dissemination of existing knowledge and competence to different health professionals who are involved in the treatment of somatoform patients.

There is also an ongoing debate concerning the psychiatric nosological congregation in somatoform, related disorders and their diagnostic criteria (Rief and Hiller 1999). High rates of mutual comorbidity between anxiety, affective, substance abuse, and somatoform disorders have challenged the usefulness of the concept of discrete and stable nosological entities and different treatment needs. Some clinicians propose a concept of spectrum disorders with more transient borderlines between different, but obviously linked, clusters of psychopathology and their courses in a long-term perspective. Specifically the relevance of subclinical forms of disorders as precursors and remaining psychopathological states after partial remission is stressed. Furthermore, the congregation within the somatoform disorders is questioned, e.g., for hypochondriasis, where parallels to anxiety disorders are obvious. Etiologic and pathogenetic models of panic and obsessive compulsive disorder seem to be

of value for a better understanding and effective modification of health anxieties in somatization syndromes, too. Some new candidates are thought to be potentially affiliated as psychiatric diseases with apparent parallels to somatoform disorders. Chronic fatigue syndrome, fibromyalgia, and multiple chemistry sensitivity, etc. are syndromes with bodily complaints of unknown etiology and pathogenesis. Their treatment might profit from adopting interventions, which have been successfully applied in somatoform disorders.

See also: Fear: Psychological and Neural Aspects; Freud, Sigmund (1856–1939); Hysteria; Janet, Pierre (1859–1947); Mental Health and Normality; Psychoanalysis in Clinical Psychology; Psychoanalysis: Overview; Psychosomatic Medicine; Stress and Health Research; Stress, Neural Basis of

Bibliography

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders*: *DSM-IV*, 4th edn. American Psychiatric Association, Washington, DC

Barsky A J, Klerman G L 1983 Overview: Hypochondriasis, bodily complaints and somatic styles. *American Journal of Psychiatry* **140**: 273–81

Bass C M, Murphy M R 1990 Somatization disorder: Critique of the concept and suggestions for future research. In: Bass C (ed.) *Somatization: Physical Symptoms and Psychological Illness*. Blackwell, Oxford, UK

Escobar J I, Burnam M A, Karno M, Forsythe A, Golding J M 1987 Somatization in the community. *Archives of General Psychiatry* **44**: 713–8

Guggenheim F G 2000 Somatoform disorders. In: Saddock B J, Saddock V A (eds.) *Comprehensive Textbook of Psychiatry*. Lippincott Williams and Wilkins, Philadelphia, PA, pp. 1504–32

Kellner R 1985 Functional somatic symptoms and hypochondriasis. *Archives of General Psychiatry* **42**: 821–33

Lipowski Z J 1988 Somatization. The concept and its clinical application. *American Journal of Psychiatry* **145**: 1358–68

Mayou R, Creed F, Sharpe M 1995 *Treatment of Functional Physical Symptoms*. Oxford University Press, Oxford, UK

Pilowsky I 1997 *Abnormal Illness Behaviour*. Wiley, Chichester, UK

Rief W, Hiller W 1998 *Somatisierungsstörung und Hypochondrie* [*Somatization Disorder and Hypochondriasis*]. Hogrefe, Göttingen

Rief W, Hiller W 1999 Toward empirically based criteria for the classification of somatoform disorders. *Journal of Psychosomatic Research* **46**: 507–18

Rief W, Nanke A 1999 Somatization disorder from a cognitive-psychobiological perspective. *Current Opinion in Psychiatry* **12**: 733–8

Simon G E, von Korff M 1991 Somatization and psychiatric disorder in the NIMH Epidemiologic Catchment Area study. *American Journal of Psychiatry* **148**: 1494–1500

Warwick H M C, Salkovskis P M 1980 Hypochondriasis. *Behavioral Research and Therapy* **28**: 105–17

World Health Organization 1991 *International Classification of Diseases* (*ICD-10*). WHO, Geneva

J. Angenendt and M. Härter

Somatosensation

The somatosensory system provides both some of our most immediate and compelling sensations—e.g., toothache, orgasm, and the urge to urinate—and those which, almost unnoticed, guide our movements as we write a letter, button a shirt, knead bread dough, use a screwdriver, or simply walk across the room. Under the rubric of somatosensation we include not a single modality, but several distinct submodalities: touch, proprioception, thermoception, and nociception, each served by characteristic receptor types, and conveyed to the brain by distinct neural pathways. Unlike vision or hearing, receptors for somatosensation are not localized in one place—like the retina or cochlea—but scattered throughout the body: in skin, muscle, and viscera (Gardner et al. 2000).

The sensitivity of those receptor systems is astonishing. Using active touch of the fingertip, we can distinguish a raised dot .006 mm high and 0.04 mm wide (smaller than Braille) from a smooth surface. This high degree of spatial resolution is matched by a temporal resolution that enables a Braille reader to process up to 100 characters per minute (Phillips et al. 1990). Conversely, as a probe is moved stepwise across the skin, we can identify, with great precision, not just the location of the probe, but the direction and speed of its movement and the duration of its contact. Similarly, variations in the angle of the knee joint can be detected with a resolution of 0.5°. Note that all of these feats of discriminative somatosensation involve movement: either movement of a limb with respect to the body, movement of a body surface with respect to the stimulus, or of a stimulus with respect to the body surface. These examples highlight several key features of the somatosensory system; its multimodal nature, its close association with movement—in proprioception and active touch—and the topographic (i.e., somatotopic) mapping of sensory inputs with respect to the body image.

Although the somatosensory system processes both pain and temperature, this review will focus on its role in mechanoreception. Progress in this area over the past few decades has involved a continuing interaction between studies of somatosensory psychophysics in human and animals, especially primates (Talbot et al. 1968). The development of methods for single-unit recording and controlled presentation of tactile stimuli to awake primate preparations has made it possible to

study, at the cellular level, neuronal activity correlated with performance on somatosensory discrimination tasks (Gardner and Kandel 2000).

As with any sensory system, we begin with the initial elements of the system—the receptors and primary afferent fibers—characterizing their receptive field properties, and considering their putative contribution to the encoding of specific sensations or sensory attributes. We identify the ascending neural pathways that convey the encoded information to the spinal and cranial divisions of the CNS (central nervous system), and thence to the thalamus and somatosensory cortex. The somatotopic and physiological organization of somatosensory cortical regions is reviewed, with a focus on the manner in which information about modality and topography are integrated at the cortical level to provide functionally significant outputs. In the final section we examine processing within somatosensory cortex and between the 'primary' sensory areas and adjacent cortical 'association' and 'motor' areas involved in the generation of 'percepts' and the control of movement.

1. *Receptors, Receptive Field Properties and Encoding of Somatosensory Properties*

Human skin contains several classes of mechanoreceptors—Merkel disks, Meissner's, Ruffini, and Paccinian corpuscles—whose differential stimulus selectivity is associated with a distinctive morphology. All four are capable of transducing stimulus energy into peripheral neural activity via the generation of receptor potentials. Their associated afferent fibers may be classified with respect to receptor location (cutaneous vs. deep), receptive field size (small vs. large), and adaptation to maintained stimulation (slowly vs. rapidly adapting). In humans, a given receptor type is uniquely associated with a specific type of conscious sensation (e.g., light touch\pressure, flutter\vibration, stimulus movement). In addition to these tactile receptors, muscles and joints are endowed with a variety of proprioceptors (e.g., muscle spindles, Golgi tendon organs) activated by limb displacement. This ensemble of mechanoreceptors and their associated afferents mediate the extraordinarily rich array of sensations that accompany contact with the skin and movement of the limbs.

By combining analysis of single primary afferent responses to controlled mechanical stimulation in monkeys, with psychophysical measurements in humans it has been possible to assess the differential contribution of different receptor classes to the encoding of tactile stimulus properties. Such studies suggest that spatial properties of tactile stimuli (form, texture) are best preserved in the discharge of slowly adapting cutaneous receptors; their temporal properties (flutter, vibration, slip) by the discharge of rapidly adapting afferents. However, natural stimuli do not engage a single receptor, but a population of receptors of different types. Thus, complex movements like grasping, lifting, and manipulating objects are guided by a continuous flow of afference from slowly and rapidly adapting receptors of several types, while information about the static position and movements of limbs involves a combination of muscle, joint, and cutaneous inputs (Johnson and Hsiao 1992, Valbo et al. 1984).

2. *Ascending Neural Pathways: Modality Segregation and Somatotopic Organization in the Lemniscal System*

Each of the receptor types discussed above is associated with a primary afferent fiber arising from a peripheral ganglion cell in the spinal or cranial divisions of the somatosensory system and conveying inputs from, respectively the body or head, to higher levels of the system. The ascending somatosensory pathways are organized according to the principle of modality segregation, which insures that information from each type of receptor is conveyed along a separate 'labeled' line, to terminate in a unique collection of cells in the CNS. This insures that, at least at early stages of processing, information about e.g., light touch, vibration, or proprioception is kept separate and processed in parallel within a 'lemniscal' pathway. Moreover, pain and temperature inputs are conveyed and processed by a distinctly different—anterolateral—pathway.

Information from the different mechanoreceptor types projects to 'higher' brain regions via a series of sensory 'relay' nuclei at brainstem and thalamic levels. The organization of this 'lemniscal' projection reflects not only modality segregation but an initial dissociation of inputs from the body (spinal) and head (trigeminal) divisions into dorsal column-lemniscal and trigeminal lemniscal pathways, respectively. At each level of the pathway, the representations of head and body within a nucleus are somatotopically organized. For example, at lower brainstem levels (dorsal column nuclei: gracilis and cuneatus) the body representation (which is ipsilateral) is mapped so as to form a 'homunculus,' with the lower trunk, legs, and feet medial in the medulla, and the upper trunk, hands, and arms, lateral. The body and head representations travel in the medial and trigeminal lemnisci, which cross at medullary levels so that in the thalamic 'homunculus' the contralateral head and body are represented in medial, and lateral posterior subnuclei (VPM, VPL) respectively. The ventrobasal complex, in turn, relays this information to two cortical somatosensory areas: SI, located on the postcentral gyrus of the parietal lobe, and SII, on the banks of the lateral fissure. In addition to sharing a thalamic input, the two regions are linked by reciprocal connections within and between the two hemispheres (Gardner and Kandel 2000).

3. Somatosensory Cortex: Brain Maps and Sensory Function. Anatomical, Physiological and Functional Organization

SI occupies a cortical region divisible, by cytoarchitectonic criteria, into Brodmann's areas (3a, 3b, 1, and 2) and lying along an antero-posterior axis bounded rostrally by motor cortex and caudally by posterior parietal regions. Within the area defined as SI there are four somatotopically similar but physiologically distinct representations of the body surface (Nelson et al. 1980). These somatosensory maps are 'functional' rather than 'veridical.' All parts of the body are not equally represented since representation of a body part reflects its significance for the animal, not the actual space it occupies. The 'significance' of a body part is highly correlated with density of its receptor innervation. Thus the tail of the spider monkey, the whiskers of the rat, and the hand and tongue of humans occupy disproportionate amounts of the cortical map of those species. Moreover, the cortical map preserves the modality segregation inherent in the 'labeled line' processing of receptor inputs. The representation in each of the four areas is dominated by input from one mechanoreceptor submodality: in 3a it is primarily muscle stretch receptors; in 3b input from cutaneous receptors, in area 2 the dominant input is from deep pressure receptors, while in area 1 rapidly adapting cutaneous receptors predominate. Despite these differences in their predominant inputs, the various somatosensory representations within SI share a common physiological organization, evident in the results of single-unit studies. As an electrode advances downward from the cortical surface along a radial traverse through successive layers, all the neurons encountered in its path will respond to the same type of stimuli (cutaneous or deep, rapidly or slowly adapting, muscle or skin) originating from the same part of the body. This principle of 'columnar organization,' first encountered in SI, has proven to be an important principle of cortical organization across other sensory systems. For the somatosensory system, the mapping process results in an orthogonal relationship between modality and topography, such that a recording from any point on the map will generate information about both the nature of the peripheral sensory message and the part of the body from which it originates

4. Somatosensory Plasticity and the Phantom Limb Syndrome

The complexity and stereotypy of cortical maps suggested that such maps are a fixed and relatively immutable property of mammalian somatosensory cortex. However, it is now evident that the somatotopic maps defined by recording studies are modifiable, even in adult primates, by a variety of sensory and experiential manipulations, including deafferentation and behavioral training. Differential use of specific digits or portions of a single digit (in monkeys, or human Braille readers) may result in expansion of the map area devoted to the stimulated region (Recanzone et al. 1992). Deafferentation of specific portions of the map of the hand in primates produces a sequence of stages, in which the denervated area may initially be unresponsive, but then become gradually responsive to inputs, not from the originally innervated region, but from regions adjacent to it. It is as if the areas devoted to the adjacent regions expand into the region deprived of its original input by the deafferentation. This interpretation is supported by an experiment in which removal of a single middle digit (D3) is followed by expansion into its representation by the adjacent digits (D4 and D2). Similarly, after surgical fusion of two contiguous digits (D3, D4) the receptive fields of the two digits become continuous across the area of fusion, and this continuity remains even after the fusion is removed (Clark et al. 1988, Jones 2000). However, the most striking examples of such somatosensory 'plasticity' are seen after limb amputations, which result in expansion of the adjacent face representation into the area originally activated by inputs from the limb. Stimulation of the face in such animals activated neurons, not only in the face, but also in the denervated hand or arm areas. Anatomical tracing experiments have shown that these new patterns of excitation are correlated with the emergence (over time) of a widespread expansion of lateral connections in the affected cortical regions. These findings have prompted a reexamination of the 'phantom limb' phenomenon of human amputees, and the suggestion that it may reflect reorganization processes similar to those seen in denervated monkeys (Florence et al. 1997, Kaas et al. 1983, Ramachadran 1993).

5. From 'Somatosensation' to Somatosensorimotor Functions

Ablation studies of SI and SII have demonstrated a rough correlation between the physiological properties of anatomically discrete somatosensory cortical areas and the differential function of those subdivisions. Ablation studies of SI in primates are consistent with the physiological data. The differential effects of lesions of SI subdivisions upon tactile discriminations (texture, size, and shape) are correlated with the stimulus properties processed by that area. Lesions of SII, in addition to their effects upon texture and size discriminations, also disrupt the ability to transfer a tactile discrimination task from one hand to the other, an observation consistent with the presence, in SII, of a bilateral somatosensory representation, mediated by callosal connections (Carlson 1980).

However, the parallel processing of peripheral inputs to SI may represent only a first stage in the

transition from 'sensation' to perception and movement, i.e., from the identification of specific sensory attributes to the recognition of, or interaction with objects. The possibility of a serial, possibly 'hierarchical' mode of processing is suggested both by the intracortical connectivity patterns within SI and the increasing complexity of the receptive field properties of neurons encountered in penetrations along the anterior-posterior axis of the postcentral gyrus. The initial thalamic input to SI terminates primarily within layers III and IV of area 3b, which project, in turn, upon layer IV of more caudal areas 1 and 2. This connectivity pattern is accompanied by an increase in both the size and complexity of somatosensory receptive fields along the anterior posterior axis. Thus neurons recorded in area 3b tend to have relatively small, punctate fields involving only a single digit, and comparable with those recorded from primary afferent fibers. The receptive fields of posterior neurons (areas 2 and 1) are responsive across larger areas, including the palm and several digits, and respond to moving stimuli with distinct preferences for the direction or orientation of movement (Gardner et al. 1989). Such neurons would be relatively unresponsive to passive contact by small probe tips, or active contact of a smooth sheet of paper, but show increased discharge rate to edges and three-dimensional shapes. Such observations suggest a special role for areas 1 and 2 in the digital coordination involved in grasping and manipulation of objects. Reversible (pharmacological) disruptions of cortical activity centered on area 2 produced a marked impairment in such manipulative abilities in primates. Interestingly, ablations of specific finger regions of area 3a and 3b deactivate homologous digit areas in area 1, but not the reverse (Iwamura et al. 1991). Finally, lesion data also suggests that SI lesions which disrupt complex discriminative functions (such as categorizing the relative speed of a tactile stimulus) do not impair a simple stimulus detection task. These findings are consistent with serial processing mechanisms involving the convergence of processed inputs from areas 3b, 3a and 1 upon area 2. Evidence for an additional level of serial processing involving SII comes from observations that neurons in this region may have still more complex receptive fields, including bilateral representations, involvement in active touch and modulation (enhancement or suppression) of neuronal activity by attentional processes (Burton et al. 1995). The participation of such processes should not be surprising. In nature, animals seek out, attend to, and actively engage objects, using 'mobile sensors,' such as the human fingers, to generate movement patterns directed at surfaces and three-dimensional objects (Lederman and Klatsky 1987). Such 'haptic' scanning patterns are among the most complex movements mediated by the somatosensorimotor system. Their generation, sensory control and contribution to object recognition are likely to involve the interaction of SI and SII with contiguous frontal and parietal 'association' cortex (Gardner and Kandel 2000).

See also: Audition and Hearing; Olfactory System; Pain, Neural Basis of; Phantom Limbs; Vestibular System

Bibliography

Burton H, Fabri M, Alloway K D 1995 Cortical areas within the lateral sulcus connected to cutaneous representations in areas 3b and 1: A revised interpretation of the second somatosensory area in macaque monkeys. *The Journal of Comparative Neurology* **355**: 539–62

Clark S A, Allard T, Jenkins W, Merzenich M M 1988 Receptive fields in the body surface map in adult cortex defined by temporally correlated inputs. *Nature* **332**: 444–5

Carlson M 1980 Characteristics of sensory deficits following lesions of Brodmann's areas 1 and 2 in the postcentral gyrus of *Macacca Mulata*. *Brain Research* **204**: 424–30

Florence S, Jain N, Kaas J 1997 Plasticity of somatosensory cortex in primates. *Seminars in Neuroscience* **9**: 3–12

Gardner E P, Kandel E 2000 Touch. In: Kandel E., Schwartz J, Jessel T (eds.) *Principles of Neural Science*. McGraw-Hill, New York, Ch. 23, pp. 451–71

Gardner E P, Martin J H, Jessell T 2000 The bodily senses. In: Kandel E, Schwartz J, Jessel T (eds.) *Principles of Neural Science*. McGraw-Hill, New York, Ch. 22, pp. 430–50

Gardner E P, Hamalainen H A, Palmer C I, Warren S 1989 Touching the outside world: Representation of motion and direction within primary somatosensory cortex. In: Lund J S (ed.) *Sensory Processing in the Mammalian Brain: Neural Substrates and Experimental Strategies*. Oxford University Press, New York, pp. 49–96

Iwamura Y, Tanaka M 1991 Organization of the first somatosensory cortex for manipulation of objects: an analysis of behavioral changes induced by muscimol injection into identified cortical loci of awake monkeys. In: Franzen O, Westman J (eds.) *Information Processing in the Somatosensory System*. Macmillan, London, pp. 371–80

Johnson K O, Hsiao S S 1992 Neural mechanisms of tactual form and texture perception. *Annual Review of Neuroscience* **15**: 227–50

Jones E G 2000 Cortical and subcortical contributions to activity-dependent plasticity in primate somatosensory cortex. *Annual Review of Neuroscience* **23**: 1–37

Kaas J, Merzenich M, Killackey H 1983 The reorganization of somatosensory cortex following peripheral nerve damage in adult and developing animals. *Annual Review of Neuroscience* **6**: 325–56

Lederman S J, Klatzky R 1987 Hand movements: a window into haptic object recognition. *Applied Cognitive Psychology* **19**: 342–68

Nelson R J, Sur M, Fellerman D J, Kaas J H 1980 Representations of the body surface in postcentral parietal cortex of *Macacca fascularis*. *The Journal of Comparative Neurology* 611–43

Phillips J R, Johansson R S, Johnson K O 1990 Representation of Braille characters in human nerve fibers. *Experimental Brain Research* **81**: 589–92

Ramachadran V S 1993 Behavioral and magnetoencephalographic correlates of plasticity in the adult human brain. *Proceedings of the National Academy of Sciences* **90**: 10413–20

Recanzone G H, Merzenich M M, Schreiner C E 1992 Changes in the distributed temporal response properties of S1 cortical neurons reflect improvements in performance on a temporally based tactile discrimination task. *Journal of Clinical and Experimental Neurophysiology* **67**: 1071–91

Talbot W H, Darian-Smith I, Kornhuber H H, Mountcastle V B 1968 The sense of flutter vibration: Comparison of the human capacity with the response patterns of mechanoreceptive afferents from the monkey hand. *Journal of Clinical and Experimental Neurophysiology* **31**: 301–34

Valbo A B, Olsson K A, Wetsberg K G, Clark F J 1984 Microstimulation of single tactile afferents from the human hand: Sensory attributes related to unit type and properties of receptive fields. *Brain* **107**: 727–49

P. Zeigler

Somatotherapy, History of

1. Definition and Roots

In psychiatry, somatotherapy refers to a variety of treatments of mental illness by physical or chemical means. It also includes psychosurgery and operations practiced in the late nineteenth and early twentieth century, such as castration, or removal of the clitoris, tonsils, teeth, etc. (Kalinowsky and Hippius 1969).

The roots of physical therapies are in bloodletting, and in the use of emetics, and purgatives. The roots of chemical therapies are in soporifics which have been in use since ancient medicine. Poppy (opium) and Indian hemp (*Cannabis indica*) were known to the Egyptians, and mandrake (*Atropa mandragora*) to the Babylonians and Hebrews. Henbane (*Hyoscyamus*) and dewtry (*Datura stramonium*) were used by the Orientals and the Greeks. The roots of psychosurgery are in 'decompressive trephining' that had been practiced by the neolithic Gauls and the Bohemians (Garrison 1929).

2. Pharmacotherapy: Behavior Control

The developments, which led to pharmacotherapy in psychiatry, began with the isolation of morphine from opium in Germany by Serturner in 1806; the separation of bromine from the ashes of the seaweed in France by Balard in 1826; and the synthesis of chloral hydrate in Germany by Liebig in 1832. Morphine, for controlling agitation and aggression, and potassium bromide for relieving restlessness and anxiety, were introduced for clinical use in the late 1850s. Chloral hydrate for controlling insomnia became available about 10 years later. The judicious use of these three drugs provided the necessary means for day- and night-time sedation. It also allowed the replacement of physical restraint by pharmacological control (Lehmann and Ban 1970).

Isolation of hyoscine (scopolamine) from hyoscyamine and the preparation of paraldehyde during the 1880s provided alternative treatments. Subcutaneously administered combinations of morphine (or apomorphine) and scopolamine (hyoscine) became the prevailing treatment modality of excitement and agitation in psychiatry by the dawn of the twentieth century (Ban 1969).

The synthesis of malonylurea (barbituric acid) by von Baeyer in 1864 opened the path for the development of barbital (Veronal), the first barbiturate hypnotic introduced in 1903. Barbital, together with many other long-, intermediate-, short-, and ultra-short-acting barbiturates, dominated the treatment of insomnia for about 50 years (Lehmann 1993).

3. Beyond Behavioral Control

The first treatments in psychiatry which aimed beyond behavioral control were fever therapy and prolonged sleep (Shorter 1997).

3.1 Fever Therapy

The development, which led to fever therapy, began in the mid-1890s. It was triggered by Wagner-Jauregg's observation that one of his female patients, who contracted erysipelas, recovered from her psychosis after she developed a high fever.

Instrumental to further development was the preparation of the tuberculin vaccine in 1890 by Robert Koch and the demonstration of the presence of *Treponema* (Spirocheta) *Pallidum*in the brains of patients with syphilitic general paresis in 1905 by Moore. The recognition of the heat sensitivity of spirochetes was also important.

To follow up his observation, Wagner-Jauregg induced fever with the employment of the tuberculin vaccine in several of his patients with general paresis at the University of Vienna. In spite of the favorable results, he replaced the vaccine in 1917 by blood from patients with tertian malarial attacks, because malaria could be reliably terminated by quinine.

Treatment with malaria-induced fever was an overwhelming success. Six of the first nine patients treated with 'fever therapy' improved; three fully recovered (Wagner-Jauregg 1919). Wagner-Jauregg received the Nobel Prize in 1927 for introducing the first effective treatment for neurosyphilis.

Fever therapy had also been tried in psychiatric disorders other than general paresis, e.g., depression,

schizophrenia; and fever had been induced by means other than the tuberculin vaccine and malarial blood, e.g., typhoid vaccine, killed *Escherichia coli* suspension, sodium nucleinate, sulfur in oil, hot bath. But only in general paresis did 'malaria therapy' have a demonstrable therapeutic effect. Even in general paresis, only in one-third of the patients did it yield 'satisfactory' results (Pichot 1983).

3.2 Sleep Therapy

Sleep therapy can be defined as pharmacologically induced sleep for about 20 hours a day for an extended period, e.g., two weeks. It was first introduced in Shanghai with the employment of bromides ('bromide sleep') by Macleod for mania in 1900. Sleep therapy was also used by Wolff, with the employment of methylsulfonal one year later ('Trionalkur').

Sleep therapy was reintroduced as 'prolonged sleep' (*'l'ipnosi prolongata'*) with long acting barbiturates by Epifaneo in 1915 at the University of Turin. It was adopted by Klaesi for the treatment of schizophrenia in 1922 at the Burgholzli psychiatric clinic in Zurich. Klaesi referred to sleep therapy (prolonged sleep) as 'continuous narcosis' (*Dauernarkose*) (Klaesi 1922).

For about three decades, during the 1930s, 40s and 50s, 'sleep therapy' was used in conditions ranging from manic and catatonic excitement, through severe anxiety and depression, to various forms of schizophrenia. Various drug combinations had been employed to induce and sustain sleep. Klaesi himself used Somnifen, a barbiturate combination that contains in equal amounts diethyl and allylisopropyl barbituric acid. The Cloetta mixture was another frequently used combination. It contains paraldehyde, amylhydrate, chloral hydrate, alcohol, isopropyl-allyl-barbituric acid, digalene, and ephedrine.

Sleep therapy was of marginal benefit, but of considerable risk. In Klaesi's original sample of 26 schizophrenic patients, the improvement rate was about 30 percent, that is about 10 percent higher than the 20 percent spontaneous remission rate in a similar population. This relatively small gain was offset by a higher than 5 percent fatality from pneumonia and/or circulatory collapse. Nevertheless, 'sleep therapy' was lingering on, even after the introduction of new psychotropics—with the incorporation of drugs like chlorpromazine in some of the 'sleep mixtures'—before fading away in the 1960s.

In 1971, Pflug and Tolle, found sleep deprivation therapeutic in endogenous depression (Pflug and Tolle 1971). The findings are in keeping with neuropharmacological theory about the action mechanism of tricyclic antidepressants. Since the early 1970s sleep deprivation has been employed with success in the facilitation of the therapeutic effect of these drugs (see *Schizophrenia, Treatment of*).

4. Physical Therapies

Insulin-induced coma and electrically-induced convulsions were the first 'physical therapies' of schizophrenia with demonstrable therapeutic effectiveness without considerable risk (Sargant and Slater 1954).

4.1 Insulin Coma

Insulin was isolated from the pancreas by Banting and Best in 1921. Insulin is the hormone that regulates sugar metabolism and since 1922 it has been used as replacement therapy in diabetes mellitus, a disease caused by its deficiency. The first report on the successful treatment of symptomatic depression with insulin in diabetic patients was published in 1923.

Insulin, was introduced into psychiatry for stimulation of appetite and sedation in the late 1920s. Its indications were extended to the alleviation of the discomfort in morphine withdrawal in the early 1930s. Manfred Sakel's report in 1930 was one of the first on the therapeutic effect of insulin in morphine withdrawal.

The origin of insulin-coma therapy in schizophrenia was in Sakel's observation that some of his restless and agitated addicts who inadvertently slipped into coma became tranquil and accessible. By following up in schizophrenics, at the neuropsychiatric clinic of the University of Vienna, the observations he made in addicts at the Lichterfelde Sanatorium in the suburb of Berlin, Sakel established the effectiveness of insulin coma therapy in schizophrenia. He published his findings in a series of 13 papers in the *Wiener medizinische Wochenschrift* between November 1934 and February 1935 (Sakel 1935).

Insulin coma therapy fared favorably to all other therapies used in the treatment of schizophrenia in the 1940s. It was especially effective in recent cases. The spontaneous remission rate was doubled if the treatment was used within the first year of the illness. The risks with insulin-coma therapy were low. Of the 90 deaths reported out of the 12,000 patients treated with insulin coma therapy in the United States, only half could be attributed to the treatment. One British expert had only one fatality in over 1,000 cases (Mayer-Gross et al. 1960). Nevertheless, insulin-coma therapy was short-lived in the Western world. It was promptly abandoned after the introduction of new psychotropic drugs. This was not the case in the Soviet Union and China, where it remained one of the major forms of treatment of schizophrenia until the late 1980s.

4.2 Electroshock

Developments, which led to the introduction of electroconvulsive therapy in psychiatry, began in the

late 1920s with Nyiro and Jablonsky's clinical observation that epileptic patients improved if they developed schizophrenia. It continued with Laszlo Meduna's findings in autopsies that the neuroglia, which is the supporting tissue of the nerve cells, was very thick in epilepsy and very thin in schizophrenia, and postulation of an antagonism between epilepsy and schizophrenia.

To pursue matters further, Meduna induced convulsions for the first time in a schizophrenic patient by the intramuscular administration of camphor on January 23, 1934 at the National Psychiatric Hospital for Nervous and Mental Diseases in Budapest. Adopting the treatment schedule used in malarial therapy, he repeated the injection every third day. After the fifth injection he found the patient markedly improved.

In spite of the favorable findings in several patients, Meduna replaced camphor by intravenously administered Cardiazol (Metrazol, pentetrazol) because of the pain caused by the intramuscular injection, the delay in the onset of the convulsions, and the agonizing fear patients experienced during the delay (Meduna 1937).

Pharmacologically induced convulsions were replaced by Ugo Cerletti and Lucio Bini by electrically induced convulsions in the late 1930s. Electroshock (EST), originally referred to as electroconvulsive therapy (ECT), was first given to a schizophrenic patient on April 18, 1938 at the neuropsychiatric clinic of the University of Rome. It was administered through electrodes placed on both temples that Bini found safe in dogs. Cerletti and Bini reported their findings with EST to the Italian Academy of Medicine in Rome on May 28, 1938 (Cerletti and Bini 1938). The treatment was therapeutically effective and safe; the only drawbacks it had were fractures and memory disturbance.

The risk for memory disturbance with EST was greatly reduced by the use of 'brief-pulse currents' and moving the path of the current in the brain away from the memory centers by locating the electrodes on one side, or over the very front of the head. An added advantage of bifrontal electrode placement is that seizures are elicited at near threshold frequencies.

The risk for fractures was greatly reduced by pretreatment with a neuromuscular blocker. Bennett in the United States was first in 1940 to employ curare to relax the muscle (Bennett 1940). Curare, a substance that competes with acetylcholine at the neuromuscular junction, was replaced by succinylcholine, a safer drug, which prevents the action of acetylcholine by depolarizing the muscle end-plates.

To prevent the panic which is caused by not being able to breath, succinylcholine administration is preceded by the administration of a short acting barbiturate. ECT, administered under general anesthesia with continuous oxygenation and monitoring of the body's physiology during treatment, is the safest among the somatic therapies today. It offers specific advantages for the elderly, in pregnancy, and for patients with cardiac disease (Fink 2000, Kaplan and Sadock 1988).

The indications of ECT have been extended over the years from schizophrenia to include mania and depression. However, in mania and schizophrenia, it is used only in treatment failures or in patients with specific contraindications to drugs. In depression ECT is also used as primary treatment in delusional-psychotic patients and in patients perceived as suicidal risks (Abrams 1997, Fink 1979).

In the mid-1990s a new technique, high-energy transcranial magnetic stimulation (rTMS) of the brain, was introduced. Expectations were high that it could eventually replace ECT. This, however, was not to be the case. Compared with ECT, the physiological effects of rTMS are rather modest.

5. *Psychosurgery*

Development of psychosurgery began in the late nineteenth century in Switzerland with the removal of parts of the brain (topectomy, craniotomy) in six psychiatric patients by Burckhardt. The first operations were of questionable success; one patient died, one improved, two became 'quieter,' and two remained unchanged (Burckhardt 1891).

Progress in psychosurgery received a great impetus in 1935 by Jacobsen and Fulton's demonstration that removal (ablation) of the frontal lobe has a calming effect in the monkey. It was their report which prompted Egas Moniz, a Portuguese neurologist, to develop transcortical leukotomy, a method in which the white matter between the frontal lobes and the thalamus in the brain is severed through burr-holes cut in the cranium. The first transcortical leukotomies were conducted in 1935 by Moniz and Almeida Lima, a neurosurgeon, in schizophrenic patients from the Bombarda asylum, at the Santa Naria Hospital in Lisbon. The operation was a success. It reduced the severity of tension and psychotic symptoms in over 60 percent of the patients. (Moniz 1937). Moniz received the Nobel Prize in 1949 for his pioneering work in the new field and for his role in the development of psychosurgical techniques.

Psychosurgery was introduced in the United States by Walter Freeman, a Washington neurologist, in collaboration with James Watts, a neurosurgeon. The first prefrontal lobotomy was performed in Washington on a patient with agitated depression at George Washington University in 1936 (Freeman and Watts 1936). About 10 years later, in 1946, Freeman introduced transorbital lobotomy by approaching the brain through the eye socket with an ice pick.

Psychosurgery fell into disrepute during the 1950s and 1960s, because it deprived patients from their judgement and social skills. Demonstration of its effectiveness and safety with the introduction of improved stereotactic techniques, however, led to its

revival in the 1970s and 1980s. The new techniques allow the placement of discrete lesions in the brain, and also to make the actual lesions by radioactive implants, cyroprobes, electrical coagulation, proton beans, and ultrasound waves (Lehmann and Ban 1994).

Today, psychosurgery covers a wide range of operations which aim to reduce the severity of psychopathology by destroying specific regions (lobotomy) or connecting tracts (leukotomy, tractotomy) in the brain. Among the most frequently employed operations are cingulectomy in the treatment of refractory obsessive compulsives, and thalamotomy in the treatment of refractory depressions.

6. *Pharmacotherapy in the 1930s*

There was a steadily growing interest in pharmacotherapy during the 1930s. One of the targets was catatonic inhibition which by the mid-1930s had been transiently removed in some patients by the inhalation of a 30 percent carbondioxide and 70 percent oxygen mixture, and the administration of sodium amobarbital and apomorphine (Loevenhart et al. 1929).

Racemic amphetamine, a powerful stimulant, was introduced by Prinzmetal and Bloomberg in the treatment of narcolepsy in 1935 (Prinzmetal and Bloomberg 1935). Two years later, dextroamphetamine was first used by Charles Bradley for calming hyperexcitable–hyperactive children. These were the first neuropsychiatric applications of the amphetamines. The amphetamines are substances with a phenylpropylamine structure synthesized by Edeleano in 1887.

Amphetamines were also tried in the treatment of depression, but with little success. Although treatment with hematoporphyrin, a photosensitizing substance, and dinitrile succinate, seemed to offer some promise in the treatment of depression, tincture of opium, introduced during the first quarter of the century, remained the prevailing pharmacological treatment modality in depressive illness during the 1930s, 1940s, and early 1950s (Ban 1981).

7. *Causal Treatments: The 1940s*

The developments which led to causal treatments in phychiatry during the 1940s began with the isolation of nicotinic acid (vitamin B-3) by Funk in 1911, and the demonstration of its therapeutic effect in pellagra by Fouts and his associates in 1937. Pellagra is a disease which may cause delirium in the acute, depression in the subacute, and dementia in the chronic stages of the illness (Ban 1971). It continued with the discovery of penicillin by Fleming in 1929, and the establishment of its therapeutic effect in syphilitic general paralysis by Stokes and his associates in 1944. It culminated with the recognition of the link between the selective memory disturbance of Wernicke's encephalopathy, first described in 1881, and thiamine deficiency by de Wardener and Lennox in 1947.

The diagnostic distribution of patients in psychiatric hospitals drastically changed with the introduction of nicotinic acid, penicillin, and thiamin into psychiatry during the 1940s. Psychoses due to cerebral pellagra and dementia due to syphilitic general paralysis virtually disappeared by the 1950s in the western world, and the prevalence of dysmnesias markedly decreased. This was the state of art in the somatotherapy of mental illness immediately prior to the introduction of new psychotropic drugs.

8. *Psychotropic Drugs: 1950–2000*

Introduction of therapeutically effective psychotropic drugs in the treatment of mental illness began in 1949 and has continued since. The term psychotropic was introduced by Ralph Gerard, an American neurophysiologist, in 1957 for drugs with an effect of mental activity and human behavior.

The history of psychotropic drugs can be separated into three distinct periods. The first corresponds with the introduction of prototype drugs during the 1950s. The second, with the introduction of the first-generation psychotropics, identified on the basis of structural and/or pharmacological similarities to the prototypes, during the 1960s and 1970s. The third, ongoing period, with the introduction of the second generation of psychotropics, developed with consideration of theories about the pathomechanism of mental illness, derived from studies on the action mechanism of therapeutically effective drugs.

The first set of psychotropic drugs included lithium, chlorpromazine, reserpine, meprobamate, iproniazid, and imipramine. In a broad sense it also included chlorprothixene, haloperidol, and chlordiazepoxide. It consisted of four antipsychotics (neuroleptics, major tranquilizers), i.e., chlorpromazine, reserpine, chlorprothixene, and haloperidol; two antidepressants, i.e., imipramine and iproniazid; two anxiolytic sedatives, i.e., meprobamate and chlordiazepoxide; and one mood stabilizer, i.e., lithium. Antipsychotics are effective in controlling excitement and psychomotor agitation, in the treatment of schizophrenia and mania, and in the management of psychosis. Antidepressants are effective in the treatment of depressive illness and also in panic disorder and obsessive compulsive disease. Mood stabilizers are effective in both the manic and the depressive phases of manic depressive illness, and in the acute, maintenance, and prophylactic treatment of bipolar disease. Anxiolytic sedatives are effective in the treatment of generalized anxiety disorder. Anxiolytics alleviate anxiety and relieve tension.

The first set of psychotropic drugs was introduced within a period of about 10 years. Lithium, in 1949

was first, followed by chlorpromazine in 1952, reserpine in 1954, meprobamate in 1956, imipramine and iproniazid in 1957, chlorprothixene in 1958, haloperidol in 1959, and chlordiazepoxide in 1960 (see *Antidepressant Drugs*).

The introduction of each of these drugs into psychiatry has its own story. The discovery of the therapeutic effect of lithium in mania was the result of John Cade's recognition that the substance protected guinea pigs from the lethal effects of urea, responsible for the toxicity of manic urine. The discovery of the therapeutic effect of iproniazid in depression was triggered by clinical observations on the euphorizing effect of the substance when used in the treatment of tuberculosis. The origin of chlorpromazine, the first antipsychotic phenothiazine, was in research with antihistaminics. Its introduction into the treatment of schizophrenia and mania by Delay, Deniker, and Harl followed within two years of its synthesis by Paul Charpentier in 1950. The origin of reserpine was in research with Rauwolfia alkaloids. Its introduction in the treatment of psychosis by Delay, Deniker, Tardieu, and Lemperiere followed within two years of its isolation from the 'snake-root' plant by Muller, Schlittler, and Bein in 1952. Development of meprobamate, the first widely used propanediol in the treatment of anxiety, was in Frank Berger's research that aimed at developing antibiotics for Gram negative bacteria originally. Development of haloperidol, the first widely used antipsychotic butyrophenone, was in Paul Janssen's research which aimed at developing analgesics which are more potent than pethidine (meperidine) originally. Imipramine was the result of the search for structurally related drugs to chlorpromazine at Geigy; chlordiazepoxide, of the screening for pharmacologically related drugs to meprobamate at Roche; and chlorprothixene, of the effort to synthesize structurally and pharmacologically related drugs to chlorpromazine at Lundbeck (Ayd and Blackwell 1970, Ban et al. 1998, Ban and Ray 1996, Healy 1996, 1998).

Some of the prototype drugs had a major impact on psychotropic drug development and by the end of the 1970s, the number of psychotropic drugs available for clinical use in the treatment of mental illness increased to about 50. Today there are well over 20 antipsychotics, antidepressants, and anxiolytic sedatives, and at least three mood stabilizers employed in the treatment of mental illness. In addition, there are also a couple of cholinesterase inhibitors in clinical use for the treatment of Alzheimer type of senile dementia (see *Alzheimer's Disease: Antidementive Drugs*).

9. *Neuropsychopharmacology in Perspective*

The introduction of therapeutically effective psychotropic drugs and the spectrophotofluorimeter during the 1950s led to the development of neuropsychopharmacology.

The spectrophotofluorimeter is an instrument with the capability (resolution power) to measure the concentration of cerebral monoamines, and their metabolites, involved in neuronal transmission at the synaptic cleft. Prior to its introduction, research dealing with central nervous system acting drugs was restricted to behavioral pharmacology and neurophysiological measures. Spectrophotofluorimetry provided direct access to the detection of biochemical changes responsible for behavioral effects.

Employment of spectrophotofluorimetry in the mid-1950s led to the demonstration that iproniazid increased, whereas reserpine decreased, cerebral monoamines, i.e., serotonin (5-HT) and norepinephrine (NE). Since iproniazid, a monoamine oxidase inhibitor (MAOI), induced euphoria in some tubercular patients, and reserpine, dysphoria in some hypertensives, the possibility was raised that the mood changes were mediated by the changes in 5-HT and NE concentrations in the brain (Brodie et al. 1956). The postulation of a relationship between reserpine-induced depletion of monoamines and depression led to the introduction of the reserpine reversal test in the screening for potential antidepressants which, in turn, led to a substantial increase in imipramine-type antidepressants by the end of the 1970s. It also provided the rationale for the research which, with the employment of radioactive isotope labeling techniques, led to the finding that tricyclic antidepressants block the neuronal reuptake of NE and 5-HT (Glowinski and Axelrod 1964).

Employment of spectrophotofluorimetry in the early 1960s led to the demonstration that nialamid, an MAOI, induced an accumulation of the 3-*O*-methylated metabolites of dopamine and NE in chlorpromazine and haloperidol treated rats. The increase in the accumulation of 3-methoxytyramine and normetanephrine could be explained by a compensatory increase of tyrosine hydroxylase activity to the drug-induced blockade of catecholamine receptors (Carlsson and Lindquist 1963). Postulation of a relationship between dopamine receptor blockade and antipsychotic effect led to the introduction of tests based on dopamine antagonism in the screening for potential antipsychotics which in turn led to a substantial increase in the number of chlorpromazine-type antipsychotics by the end of the 1970s. It also provided the rationale for the research which, with the employment of X-ray crystallography led to the demonstration of dopamine receptor blockade with antipsychotics (Creese et al. 1975).

Development of receptor binding assays during the 1970s opened the path for the determination of NE and 5-HT reuptake blocking potency of antidepressants. Identification of receptor subtypes during the 1980s led to the determination of the dopamine-D2 and serotonin-5-HT2A blocking potency of antipsychotics, and to the delineation of the receptor profile of psychotropic drugs.

Separation of receptor subtypes was further refined during the 1980s by the introduction of genetic technology. With the employment of receptor cloning by the late 1990s five subtypes of the dopamine receptor, 14 subtypes of the serotonin receptor, and five subtypes of the muscarinic-cholinergic receptor had been identified (Lucas and Hen 1995, Richelson 1999).

The new genetic technology rendered the tailoring of psychotropic drugs to receptor affinities a distinct possibility. It opened up a new perspective in the somatotherapy of mental illness.

See also: Peptides and Psychiatry; Psychopharmacotherapy: Side Effects

Bibliography

Abrams R 1997 *Electroconvulsive Therapy*. 3rd edn. Oxford University Press, New York

Ayd Jr F J, Blackwell B (eds.) 1970 *Discoveries in Biological Psychiatry*. Lippincott, Philadelphia

Ban T A 1969 *Psychopharmacology*. Williams & Wilkins, Baltimore, MD

Ban T A 1971 Nicotinic acid and psychiatry. *Canadian Psychiatric Association Journal* **16**: 413–31

Ban T A 1981 *Psychopharmacology of Depression*. Karger, Basel, Switzerland Munchen, Paris, London, New York, Sydney Australia

Ban T A, Ray O S (eds.) 1996 *History of CINP*. JM Productions, Brentwood, TN

Ban T A, Healy D, Shorter E 1998 *The Rise of Psychopharmacology and the Story of CINP*. Animula, Budapest, Hungary

Bennett A E 1940 Preventing traumatic complications in shock therapy by curare. *Journal of the American Medical Association* **141**: 322–4

Brodie B B, Shore P A, Pletscher A 1956 Serotonin-releasing activity limited to Rauwolfia alkaloids with tranquilizing action. *Science* **123**: 992–3

Burckhardt G 1891 Ueber Rindenexcisionen, als Beitrag zur operativen Therapie der Psychosen. *Allgemeine Zeitschrift fur Psychiatrie* **47**: 463–58

Carlsson A, Lindquist M 1963 Effect of chlorpromazine or haloperidol on formation of 3-methoxytyramine and normetanephrine in mouse brain. *Acta Pharmacology Toxicology* **20**: 140–4

Cerletti U, Bini L 1938 L'Ellettroshock. *Archivio Generale Di Neurologia* **19**: 266–8

Creese I, Burt D R, Snyder S H 1975 Dopamine receptor binding: Differentiation of agonist and antagonist states with 3H-dopamine and 3H-haloperidol. *Life Sciences* **17**: 993–1001

Fink M 1979 *Convulsive Therapy: Theory and Practice*. Raven Press, New York

Fink M 2000 Electroshock revisited. *American Scientist* **88**: 162–7

Freeman W, Watts J 1936 Prefrontal lobotomy in agitated depression: report of a case. *Medical Annals District of Columbia* **5**: 326–8

Garrison F H 1929 *An Introduction to the History of Medicine*, 4th rev. and enl. edn., WB Saunders Company, Philadelphia

Glowinsky J, Axelrod J 1964 Inhibition of uptake of tritiated nonadrenaline in the intact rat brain by imipramine and structurally related compounds. *Nature* **204**: 1318–9

Healy D 1996 *The Psychopharmacologists*. Chapman and Hall, London

Healy D 1998 *The Psychopharmacologists II*. Chapman and Hall, London

Kalinowsky L B, Hippius H 1969 *Pharmacological, Convulsive and Other Somatic Treatments in Psychiatry*. Grune & Stratton, New York

Kaplan H I, Sadock B J 1988 *Synopsis of Psychiatry*, 5th edn. Williams and Wilkins, Baltimore, MD, pp. 528–9

Klaesi J 1922 Ueber die therapeutische Anwendung der 'Dauernarkose' mittels Somnifens bei Schizophrenen. Zeitschrift fur die gesamte. *Neurologie und Psychiatrie* **74**: 557–92

Lehmann H E 1993 Before they called it psychopharmacology. *Neuropsychopharmacology* **8**: 291–303

Lehmann H E, Ban T A 1970 *Pharmacotherapy of Tension and Anxiety*. Thomas, Springfield, IL

Lehmann H E, Ban T A 1994 Psychiatric research in North America. In: Oldham J M, Riba M B (eds.) *Review of Psychiatry*. American Psychiatric Press, Washington, DC, pp. 27–54

Loevenhart A S, Lorenz W F, Waters R M 1929 Cerebral stimulation. *Journal of the American Medical Association* **92**: 880–3

Lucas J J, Hen R 1995 New players in the 5-HT receptor field; genes and knockouts. *Trends in Pharmacological Sciences* **16**: 246–52

Mayer-Gross W, Slater E, Roth M 1960 *Clinical Psychiatry*, 2nd edn. Cassell and Company Ltd., London, pp. 296–7

Meduna L 1937 *Die Konvulsionstherapie der Schizophrenie*. Marhold, Halle, Germany

Moniz E 1937 Prefrontal leukotomy in the treatment of mental disorders. *American Psychiatric Association Journal* **93**: 1379–85

Pflug B, Tolle R 1971 Disturbance of the 24-hour rhythm in endogenous depression and the treatment of endogenous depression by sleep deprivation. *International Pharmacopsychiatry* **6**: 187–96

Pichot P 1983 *A Century of Psychiatry*. Roger Dacosta, Paris

Prinzmetal M, Bloomberg W 1935 The use of benzedrine in the treatment of narcolepsy. *Journal of the American Medical Association* **105**: 2051–3

Richelson E 1999 Receptor pharmacology of neuroleptics: relation to clinical effects. *Journal of Clinical Psychiatry* **60**(10): 5–14

Sakel M 1935 *Neue behandlung der Schizophrenie*. Perles, Vienna

Sargant W, Slater E 1954 *An Introduction to Physical Methods of Treatment*. Livingstone, Edinburgh, UK

Shorter E 1997 *History of Psychiatry*. Wiley, New York

Wagner-Jauregg J 1919 Ueber die Einwirkung der Malaria auf die progressive Paralyse. *Psychiatrisch-Neurologische Wochenschrift* **20**: 251–5

T. A. Ban

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Sou

South America: Sociocultural Aspects

South America, the worlds fourth largest continent, is composed of 12 nation-states and one territory. The significant differences among these 13 countries are matched by great internal sociocultural and ethnic diversity. Contemporary neoliberal and authoritarian governments strive to centralize and homogenize their populations. Internal cultural, social, and political movements heighten diversity. The military often plays a strong role in centralizing efforts. The Roman Catholicism of European colonialism historically and culturally shapes pan-national and internal cultural unity of the 10 Iberian-derived nations. Diversity is especially strengthened by the striking polarization between the small wealthy minority and the large poor majority; by indigenous and Afro-Latin American movements of cultural and ethnic self-affirmation; by the presence of significant ethnic blocs resulting from European migrations since the two World Wars; by the struggles among leftist, rightist, and populist political parties; by the challenge of the twentieth century spread of protestantism; and by the liberal Catholicism of recent reforms.

1. Human and Topographic Diversity

The rapidly expanding population contains as wide a spectrum of phenotypic features found anywhere in the world. Its modern nation-states include Venezuela, Colombia, Ecuador, Peru, Bolivia, Chile, Uruguay, Paraguay, and Argentina, where Spanish is the official language; Brazil, where Portuguese dominates; Guyana where English is the national language; Suriname, where English, Dutch, and Sranan are prevalent; and Guyane, an administrative district (*département*) of France, where French is dominant. Many other languages are spoken in most of these countries.

Western South America is dominated by the mighty Andes mountains, with snow-coned active volcanoes that rise to 20,000 feet. They extend southward from the Caribbean Island of Trinidad to terminate in southern Argentina at Tierra del Fuego. In this region Venezuela, Colombia, and Ecuador are known as the northern Andes (and the former two as the Spanish Main); Peru and Bolivia constitute the central Andes, and Chile, Paraguay, Uruguay, and Argentina are called the Southern Cone. Eastern South America features the great Amazon Basin, epitomized by the Amazon itself, the 'Ocean Sea.' Toward the headwaters of the huge feeder rivers of the Amazon lie the mountainous Guiana Shield in the north, the Brazilian Shield in the south, and the Andean foothills or *montaña*, rimmed by the *ceja*, 'eyebrow,' in the west. Northwest and northeast of the Amazon regions, and west of the Guiana Shield, are the plains (*llanos*) of Colombia and Venezuela, which phase into the Orinoco Basin, and west of the Brazilian Shield are the *llanos de Mojos* of Bolivia. A subregion that combines elements of Amazonian and Andean systems is that of the North Pacific coast. From northern Ecuador through western Panama an Amazonian ecosystem prevails. From northern Peru southward, high aridity is characteristic.

2. Precolumbian People: Their Lands, Products, Societies, and Languages

Native South American cultures prior to 1492 represented a dynamic mosaic of languages, cultures, social systems, rituals, and power relationships. long-distance trade—sometimes extending for thousands of miles—combined with bi- and multilingualism and multiculturality. Such trade traversed some of the most rugged landscapes in the world, creating networks of native knowledge and imagery crosscutting great geographical differences. Antiquity of South American native peoples is highly controversial. In the east, and thence up the Amazon and Orinoco (the latter has its head waters in the Parima and Pacarima mountains of the Guiana Shield and flows west, then north, then east out into the Atlantic), evidence suggests antiquity of 12,000 years ago or earlier, but the routes of arrival are obscure. Neither transatlantic nor transpacific routes seem credible to scholars in 1999, though sporadic contacts in South America by the first Americans with Asian-Pacific and Euro-African peoples remain subjects of possibility. Plant domestication seems to begin in Amazonia and in the Andes by about 8,000 years ago.

The period beginning about 6,000 years ago and continuing until the European conquest and colonization marks the florescence of fine ceramics in Amazonia, and the development of textiles, gold, and silver work in the Andes and adjacent coastal regions. At the mouth of the Amazon, ceramics manufactured on Marajó Island, an area the size of Connecticut in

the USA, rivaled the quality of any ceramic wares found anywhere in the world in terms of their beauty of form and decoration.

By 4500 years ago great migrations from the lower and central Amazon region were underway, and regional consolidation in the Andes was occurring. People from coastal, Andean, and Amazonian regions were in contact with one another, and with the people of the Caribbean Islands, Central America, and Mesoamerica. People who belong to three language families—Arawak, Carib, and later Tupi—spread from the eastern and central Amazon to the Orinoco drainage and on to the Caribbean (Arawak), and to the base along the rim of the *montaña*. Other Amazonian language families of prominence include Tukano, Tukuna, Jívaro, Pano, and Gê; these languages spread in a general arc from northwest of the Amazon around the central Amazonian rim to the southwest. Andean languages were also highly diverse.

The Inca empire, radiating north and south out of its capital of Cuzco, in what is now Andean Peru, emerged in the fifteenth century to severely alter the cultural mosaic in the Andes, and set up political-economic and cultural shock waves that radiated eastward. This empire was in full conquest fervor by the time word arrived from the Caribbean and Central America of the pending Spanish advances down the Pacific Coast and up into the Andes. At the height of their expansion, the Inca controlled all Andean systems from central Chile to southern Colombia, and their north-south 3,416-mile parallel cobble stone roadway (one coastal, one Andean) through this vast territory (385,000 square miles) amazed the conquering Spaniards. Although the Inca never dominated the adjacent indigenous people of Upper Amazonia, there is ample evidence of extensive trade between Amazonia and Inca-dominated Andes and Coast. Quechua (a variant of which is Quichua) was the lingua franca used by the Inca in its conquest, and became the language of 12 million people throughout the empire, and beyond. Other prominent Andean languages include Chibcha in the north Andes, Aymara in the central Andes and Mapuche in the southern cone.

In the greater Amazonian areas east of the Andes people depended on maize and manioc cultivation for caloric intake, and on intensive fishing and reptile and amphibian collecting for protein and amino-acid balance. Hunting in the various ecosystems of the flooded forest, the upland forest, and the flood plains and riverine systems complemented the balance of fishing and horticulture. As time progressed manioc (*Manihot esculenta* Crantz) came to be the staple of the agricultural system, with maize sporadically and regionally grown and consumed. Fishing in rivers, estuaries, streams, and coast was of paramount importance, as were techniques for corralling turtles and other amphibians. Polities varied from well-developed chiefdoms and relatively small-scale states to more dispersed systems. Intermarriage between peoples at war with one another, and speaking different languages, seems fairly common.

In the Andean regions, the potato was the supreme crop, with beans, squash, maize, quinoa, and other grains providing a balanced diet when combined with animal protein. Approximately 240 species of potato were grown, harvested, and prepared in many ways. Llama (one of three Andean camilids), guinea pig, and much later the hairless dog, were the primary animals raised for food, and the llama was used as a beast of burden. Polities ranged from small-scale systems to states and the expansive Inca Empire.

3. *The European Conquest of Native Lands and People*

A new cultural system based on conquest and capitalism was violently created by the overseas empires of the Spanish and Portuguese after 1492 as they sought huge profits in gold, silver, spices, and slaves. In the Treaty of Tordesillas, signed in 1494, the Spanish and the Portuguese divided up their colonies long before they knew the vast extent of their respective territorial 'holdings.' From the terms of this treaty the Portuguese would lay claim to Brazil, and Spain would take the rest, including all of the Pacific regions from South America to the Philippines and Japan, all of the Andes, and a parts of Amazonia, a territory itself nearly the size of the USA. Portugal's interests first lay in the route around the horn of Africa to India, and only a bit later in its vast new colonies. In 1499, sailing (and/or writing) for Spain, Amerigo Vespucci signed his map of the 'New World' thereby creating the currently favored name for the entire continent. During the early to mid 1500s Spain set out to locate, conquer, and colonize its Amazonian, Andean, and coastal regions, and Brazil set out on the conquest of its Amazonian, coastal, and mountainous regions. The Spanish first brought sugar to the Americas and the Portuguese first brought the banana from India. Both of these overseas empires introduced cattle, sheep, goats, pigs, and chickens.

Everywhere they went the conquerors encountered native people. They subdued them by enslavement, trickery and trade, and they spread European and African diseases for which native resistance was low or nonexistent. Where possible they enlisted them in slaving activities, and they killed indigenous people with firearms, swords, spears, and arrows with steel points. In areas feasible for such warfare they sent their warhorses and war dogs to maim and kill them in their radical recolonization of the subcontinent. This brutal transformation resulted in 90 percent or greater depopulation of the native inhabitants throughout the continent. Trade and warfare, now with an escalating colonial need for indigenous chattel slaves in the lowlands, and indigenous landed slaves in the highlands, created a persistent poverty that endures into

the twenty-first century. Nonetheless, cultural creativity endured with local-level technical competence promoting subsistence and adaptations.

4. Blackness

The Spanish and Portuguese brought Black slaves from Africa and from Iberia with them. Many such people escaped to mountainous, forested, inaccessible areas, and created a myriad of free, fortified sites where Native American and African American people often collaborated, intermarried, and established powerful political systems on the fringe of the dominant Spanish and Portuguese territories. From these refuge sites, often called *palenques* in Spanish and *quilombos* in Portuguese, self-liberated residents established powerful political systems on the fringe of the dominant Iberian and colonial territories. They threatened the Crown political economies. In some regions such as western Ecuador, southwestern Colombia, western Venezuela, and the north of Brazil, active trade between runaway people, often called *cimarrones*, and the colonial planters ensued.

5. Mestizaje, Cultural Practices and Ideologies of Racial Blending

The Spanish initiated an ideology of ethnic blending, called *mestizaje* (miscegenation) that excluded both indigenous people and people of African descent. Spanish administrators from the republic of Spaniards demarcated a republic of 'indios,' so named by Christopher Columbus who thought he was in India. There was no legal space for Black people, except what they themselves created. Those excluded from the Spanish-Mestizaje realm were sometimes referred to as *la mancha*, 'the (racial) stain.' From this viewpoint the dominant, European concept of hybridity prevailed in this hegemonic model of racial mixing: the White, civilizing genes and cultures were to mix with the savage or barbarian 'Indian' and 'Black' to form such intermediate categories as *mestizo* and *mulato* in Spanish. In Portuguese, *mestiço*, and *mulato* referred to White-Black hybridity.

Those people with hybridized 'racial' features, according to European and later Creole (New World born) ideologues, ranged up and down the scale from white-civilized-Iberian to dark-civilizing-Creole. These diverse *mestizos* were known as the *castas*. In some times in some places the number of their racialized categories ranged from a dozen named 'types' to hundreds. Indeed, there could be five or more such 'types' among the sons and daughters of one couple. They were constituted as those people somewhere betwixt and between the savage and the civilized, and legal and colloquial labels for some of them included such racialized concepts as 'wolf,' 'coyote,' 'throw back,' 'up in the air,' and even 'there you are!'

Black people (including slave, *bozal*, and free, *ladino*) and native people (usually divided into tame, *manso*, and fierce or wild, *bravo*), fell outside of the scheme of *castas*. Their own intermixture, often rendered in Spanish by the term *zambo*, was (and is) a category filled with racist ideas about danger and power, for here was a mixture without the civilizing affects of European-sponsored hybridity. An indigenous artist using European techniques in 1599 did the first oil painting in South America. His subject was that of the 'Zambos' of western Ecuador who, adorned in European princely clothes, wearing indigenous gold jewelry, and carrying spears with iron tips characteristic of Africa and Europe, walked from their free Republic on the coast of Ecuador to Quito to swear alliance to the Spanish Crown in the face of English encroachments.

Before the rise of nation-states, movements of self-liberation by Black people and indigenous people were common, and some were quite successful. Early revolts in the sixteenth century resulted in ethnic regions that endured into the era of nation-state emergence. In 1579 the Quijo and the Jívaro of Ecuadorian *montaña* revolted in attempts to unite Amazonian and Andean people. In Peru and Bolivia the revolts of Tupac Amaru and Tupac Katari later did the same. In the forested and mountainous coastal zone of Alagoas, Pernambuco, Brazil, Black and indigenous people formed the state of Palmares in the 1640s, and fought Portuguese and Dutch soldiers for a century. Black people in Suriname revolted in the mid-seventeenth century, and initiated a 100 year war against the Dutch and British, winning independence for six nations—Saramaka, Paramaka, Kwinti, Ndujka, Matawai, and Aluku—in the interior of this region. In neighboring Venezuela, the Carib-speaking Yekuana expelled the Spanish in 1767, and maintained their complete independence for 100 years until a brutal nationalist regime militarized the area for the purposes of exploiting wild rubber.

6. The Transformation from Colonies to Nation-states

Wars of nationalist Independence from Spanish and Portuguese colonial rule began in 1810 in areas of vast distances from one another, from what is now Mexico through northern South America to the southern cone. Those who achieved positions at the pinnacle of power in South America were *criollos*, 'Creoles,' who outnumbered those of Spanish descent. In turn, they were vastly outnumbered by their darker compatriots, those especially of African descent who fought in wars led by Simón Bolívar in the north, and by José San

Martín in the south. As the Creole nationalist wars moved through slave territories, slavery was abolished, but blackness and indigenousness were kept under rigid control by fears that the revolution would become an 'ethnic' one. The ethnically marginalized continued to be excluded from dominions of power. Both Bolívar and San Martín envisioned a united Spanish-speaking Creole nation, and lightness of phenotype within the ideology of *mestizaje* was clearly ascendant.

Such a continental nation did not materialize and, one by one, contemporary nations rapidly emerged from their previously liberated colonial regimes. By 1822 Spain had lost all of its territories in the Americas, and the Portuguese King lost his colony to a new Brazilian monarchy with its own New World Empire. By 1830 all of the republics of Spanish-speaking South America emerged, only to begin border wars, geopolitical vestiges of which last to the present day. Peru, for example, invaded Ecuador in 1941, claiming half the latter's territory, and these countries fought a war over a very small sector of disputed territory as recently as 1981 and 1995, both of them in indigenous Jivaroan-occupied sectors long serving as refuge areas due to the rugged, mountainous, Upper Amazonian topography.

African chattel slavery continued in the new states side by side with Black freedom, as various systems of oppression of indigenous people continued side by side with native revolts. The last country to abolish slavery, Brazil, did not do so until 1888, and in some states illegal slavery existed side by side with manumission legalities. Not until 1966 (Guyana) and 1975 (Suriname) would two of the three Guianas gain nation status from the British and the Dutch, respectively.

7. *The Fusion of Cultures*

The powers of Spanish and Portuguese Catholicism came to South America in the heyday of the Spanish Inquisition. Powerful priests and others sought out indigenous and African systems of thought and knowledge to be stigmatized as idolatry and Devil worship. An underlying fear of converted Jews permeated some middle ranks of conquest and colonialism. Those of humble Iberian origin sought alternative sources of inspiration for religion and magic from indigenous and Black people, thereby heightening the symbolic powers of the oppressed. Fusions of Iberian, African, and Indigenous belief systems and ways of knowing and 'seeing' became legion. All this occurred at micro-levels in macro-systems of Iberian Catholic missionization that gives almost all of South America its special 'Latin American' sociocultural character.

In the area of language, pidgins emerged out of African, Indigenous, and European imperfect syntheses, to become Creole languages. In medicine the humoral system from Mediterranean classical antiquity fused with African curing programs and indigenous shamanism. The result was hundreds, perhaps thousands, of highly regional systems of health and curing which, in turn, served as vehicles of communication with people in other regions who had experienced similar cultural fusions. Religious syncretisms became the *modus operandi* of colonizing curates and of indigenous and African specialists in esoteric and numinous thought and reflection. Even horrors of the conquest were syncretized into indigenous and Afro-Latin American cultures.

For example, The battle cry of the Spanish conquistadors, '*Santiago*!' came from Santiago de Matamoros (Santiago the Moor Killer) as the Moors were defeated during the Castillan *reconquista* that was concluded in 1492. It was used to mark Spanish territories in the Americas. After the reading of the 'Requirement' document informing all 'Indians' of their loss of rights to the Spanish, the conquerors yelled 'Santiago!' and attacked with firearms, war horses, and war dogs, often feeding parts of the dead natives to the dogs, and rendering fat from the same dead natives to rub on wounds of dogs and horses. Santiago underwent a ritual reversal in the Central Andes where this Saint became the patron icon of indigenous, Quechua-speaking peoples, and was called forth to resist the Spanish. Also in this region, and in parts of Amazonia, indigenous people say that some outsiders, whom they call *pishtaco* in Peru, seek 'Indian fat' for nefarious purposes. This may well be a modern reference to practices of the Spanish conquerors—an indigenous dimension of historical consciousness—though it is usually treated as an item of strange and curious native lore.

8. *Cultural Continuities*

In the face of radical transformations and syncretisms, continuities were also abundant. Amazonian languages such as Tupi, Arawak, Tukano, Tukuna, Jívaro, and Pano continued in their respective indigenous areas, with trade between different people also continuing, in spite of the efforts of the Catholic curates, first the Jesuits and Franciscans, and then others, to 'reduce' the populations to nucleated areas. Tupi itself, as Neenghatú, became the Lengua Geral of rural Brazil until replaced by Portuguese in the late nineteenth century, and the same language, as Guaraní, became the national language of Paraguay, where citizens spoke it along with Spanish. Quechua spread in the Andes to Argentina and in Upper Amazonia where it may have served as a language of trade. Today at least 13 million people speak Quechua, making it by far the greatest indigenous language family in the Americas. Many languages that defy classification with other known indigenous languages

continue to be spoken today, especially in the Andean nation-states with substantial Amazonian regions and in the eastern countries of Brazil and the Guianas.

Rich realms of story telling, often called mythology, are legion among native peoples, Afro-Latin American peoples, and people of 'mixed' descent, where they constitute sources of spirituality and historicity embedded in more than 100 languages. They convey rich information ranging from medicinal plants to healing procedures, to ways of thinking about humanity and society that confront the models of hierarchically organized and exploitative nation-state bureaucracies. Such telling of pasts motivates symbolically the development of political movements such as that of Pachakutik in Ecuador in the mid-1990s, and the Ecuadorian indigenous uprisings in that nation during the same decade.

Within the socioeconomic classes oriented toward national politics, which have long been articulated to the global economy, the figure of the *caudillo* has loomed large since the end of the colonial era. A caudillo is a powerful figure, usually a man but sometimes a woman, such as the late Eva Perón in Argentina, who emerges in a time of political-economic or sociocultural crisis, attracts followers from many sectors of national life, including radically opposed political parties and classes. One of the greatest caudillos of all time was the liberator Simón Bolívar. The caudillo rallies such followers to his or her own personal benefit, thereby often losing considerable influence, and even office, when the crises are over. It is to the benefit of the caudillo, then, to maintain an aura of crisis, which in South America today, and back through its tumultuous history, is not difficult to do. The caudillo is matched in the indigenous sector by the *cacique* (sometimes called *curaca*), who shares the qualities just described. (Cacique is also used for leaders in organized crime.) Both caudillos and caciques, as leaders in peripheral nations, depend in large part on economic gains to come from industrialized, wealthy nations. The cultural form of dependence of South American countries is closely tied to the developmental policies of the center nations, whose funding for externally designed projects often support *caudillismo* and *caciquismo*.

9. *Modernities and Alternative Modernities*

Modernity among the elite and middle classes of South America is tied to urbanity, whiteness, and Euro-North Americanized consumer culture. The doctrine of *mestizaje* serves these interests in the sense that the identity of mestizo underscores the processes of social whitening (*blanqueamiento*). Similarly, the representations as mestizo of those below the elite or middle classes refers to the darker complexions and stigmatized and racialized vulgarity of the common people, who in many nations constitute the vast majority. For people of indigenous and Afro-Latin American descent, and for others classed by the elite as 'mestizos,' modernity consists of passing back and forth among different systems of thought and behavior.

Passages themselves give rise to new dimensions of power politics in South American nations, just as they gave rise to the Creole imagined nationalities of the nineteenth century. The alternative modernities in the Euro-American dominated global economy that spawned earlier state nationalism are now sewing the seeds of ethnic nationalism, manifest in many countries by the self-conscious rise of new nationalities out of the symbolic transformations and continuities of syncretic formations. Public displays of collective selfhood by indigenous people tap significant financial resources through Non Governmental Organizations (NGOs) of Europe and North America, giving such social formations and movements a new articulation to economic resources. Black selfhood, however, does not attract significant financial resources.

The nations of South America vary greatly one from the other, and the salient characteristics of their respective modernities also vary. Twentieth century migration has changed the complexion of many countries, as Europeans (Jewish, Catholic, and Protestant), Middle Easterners (Jewish, Christian, and Islamic), and Asians (Chinese, Japanese, Korean, and Vietnamese) have become prominent actors in various class sectors. In the southern cone nations of Argentina, Chile, and Paraguay, live significant populations of German immigrants. The ex-President of Peru, Alberto Fujimori, is the son of Japanese immigrants, and the Ecuadorian ex-President, Jamil Mahuad Witt, is the son of Lebanese and German immigrants.

South American nationality is often expressed through public performances as crystalizations of 'authentic' local-level or regional cultures. Concepts of cultural hybridity and alternative modernity fold ethnic and cultural diversity into a unified national pride. Festive centralization is often matched by local-level and regional divergence, as people in many walks of life appropriate nationalist stereotypes about authenticity into their own self-presentations.

For example, indigenous people of Andean Ecuador have long performed individualized constructions of the European festival of Corpus Christi. Cultural images of rain-forest inhabitants are evoked in these performances. In response to both the politicized indigenous movements and the craving of foreign tourists for 'authentic' Incaic performative reproductions, some of these festivals have been renamed '*Inti Raymi*,' solstice of the sun. Performance of them becomes transformed to a nationalist, 'mestizo' rendition of Incaic authenticity and also an indigenous rendition of their increasing cultural spaces in the modern nation state.

In the late twentieth century, nationalist governments use their expanding military powers to stress

centralization and homogenization, and pan-Roman Catholicism continues to be a centripetal force at the beginning of the twenty-first century. Nonetheless, indigenous and Black movements of South America together with labor movements, feminist movements, and various cultural emphases promoting counter-hegemonic diversity work against the exclusive imagery of ethnic blending in their respective nation states and across national boundaries. New and dynamic cultural mosaics are constantly created out of the interplay of secular-sacred centralization, and ethnic-cultural diversity. The dynamic, transformative balance between centralization and diversity that began in early colonial times continues into the beginning of the twenty-first century.

See also: Central America: Sociocultural Aspects; Historiography and Historical Thought: Indigenous Cultures in the Americas; Latin American Studies: Economics; Latin American Studies: Geography; Latin American Studies: Politics; Latin American Studies: Religion; Latin American Studies: Society

Bibliography

Ades D 1989 *Art in Latin America: The Modern Era, 1820–1980*. Yale University Press, New Haven, CT

Anderson B 1991 (1983) *Imagined Communities*, 2nd edn. Verso, London

Collier S, Blakemore H, Skidmore T E (eds.) 1985 *The Cambridge Encyclopedia of Latin America and the Caribbean*. Cambridge University Press, Cambridge, UK

Denevan W M (ed.) 1976 *The Native Population of the Americas in 1492*. University of Wisconsin Press. Madison, WI

Dillehay T D 2000 *The Settlement of the Americas: A New Prehistory*. Basic Books, New York

Fuentes C 1992 *The Buried Mirror: Reflections on Spain and the New World*. Houghton Mifflin, New York

Goodman E J 1992 (1972) *The Explorers of South America*. University of Oklahoma Press, Norman, OK

Guss D M 1999 *The Festive State: Race, Ethnicity, and Nationalism as Cultural Performance*. University of California Press, Berkeley, CA

Hill J (ed.) 1988 *Rethinking History and Myth: Indigenous South American Perspectives on the Past*. University of Illinois Press, Urbana, IL

Richardson J B III 1994 *People of the Andes*. St Remy Press, Montreal, Canada and Smithsonian Press, Washington, DC

Stern S J (ed.) 1987 *Resistance, Rebellion, and Consciousness in the Andean Peasant World*. University of Wisconsin Press, Madison, WI

Van Cott D L (ed.) 1995 *Indigenous Peoples and Democracy in Latin America*. St Martin's Press, New York

Whitten N E Jr, Torres A (eds.) 1998 *Blackness in Latin America and the Caribbean: Social Dynamics and Cultural Transformations*, 2 Vols. Indiana University Press, Bloomington, IN

Williams J M, Lewis R E (eds.) 1993 *Early Images of the Americas: Transfer and Invention*. University of Arizona Press, Tucson, AZ

N. E. Whitten, Jr

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

South Asia, Archaeology of

The archaeology of South Asia, or the Indian subcontinent, is a very substantial topic, since there is a deep history in this land of diverse peoples and cultural traditions. Moreover, archaeological fieldwork is conducted on a substantial scale, with a dozen or more excavations each year, and widespread exploration. Since it is impossible to be comprehensive in an article such as this, the article will focus on four topics (the Palaeolithic, Early Village Farming Communities, the Indus Civilization, and the Early Iron Age) that are at least representative of the principal problems on which contemporary archaelogy has focused.

1. Palaeolithic

Unexpectedly early stone tools were found at Riwat, in the Soan Valley, about 6.2 miles (10 km) south-east of Rawalpindi between 1983 and 1988 (Allchin 1995) (see Fig. 1). Twenty-five pieces of flaked quartzite were recovered. Of these, only three are undeniably artifacts; that is, objects shaped by humans. Other pieces from Riwat might be the products of human activity, but they are so simple that this cannot be determined with certainty. Two of these artifacts are cores, the third is a flake. They are generally similar to the Oldowan industry of Africa, which dates to between two and three million years ago. The geological deposits of the artifacts at Riwat have reversed polarity and are therefore associated with the Matuyama Chron, between 780,000 and 2.60 million years ago. Volcanic ash immediately above the Riwat artifacts dated to c. 1.8 million years ago. The artifacts are therefore older than that; how much older is somewhat problematic. An excavation was undertaken in 1990 in the Pabbi Hills, to the south-east of Riwat. This excavation yielded tools in a deposit that was palaeomagnetically dated to 1.2 to 1.4 million years ago.

The Lower Palaeolithic (c. 1.5 million to 200,000 years ago) is represented in virtually all regions of India, but not Sri Lanka. It is a period of two major traditions of early tool making, the western core biface (hand axe/cleaver) tradition and the eastern chopper/chopping tool tradition. The line defined by Professor Hallam Movius that follows the foothills of the Himalaya mountains down the Ganges Valley, dividing these two traditions, is a concept that has returned favor. The distribution of these tool-making traditions

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

centralization and homogenization, and pan-Roman Catholicism continues to be a centripetal force at the beginning of the twenty-first century. Nonetheless, indigenous and Black movements of South America together with labor movements, feminist movements, and various cultural emphases promoting counter-hegemonic diversity work against the exclusive imagery of ethnic blending in their respective nation states and across national boundaries. New and dynamic cultural mosaics are constantly created out of the interplay of secular-sacred centralization, and ethnic-cultural diversity. The dynamic, transformative balance between centralization and diversity that began in early colonial times continues into the beginning of the twenty-first century.

See also: Central America: Sociocultural Aspects; Historiography and Historical Thought: Indigenous Cultures in the Americas; Latin American Studies: Economics; Latin American Studies: Geography; Latin American Studies: Politics; Latin American Studies: Religion; Latin American Studies: Society

Bibliography

Ades D 1989 *Art in Latin America: The Modern Era, 1820–1980*. Yale University Press, New Haven, CT

Anderson B 1991 (1983) *Imagined Communities*, 2nd edn. Verso, London

Collier S, Blakemore H, Skidmore T E (eds.) 1985 *The Cambridge Encyclopedia of Latin America and the Caribbean*. Cambridge University Press, Cambridge, UK

Denevan W M (ed.) 1976 *The Native Population of the Americas in 1492*. University of Wisconsin Press. Madison, WI

Dillehay T D 2000 *The Settlement of the Americas: A New Prehistory*. Basic Books, New York

Fuentes C 1992 *The Buried Mirror: Reflections on Spain and the New World*. Houghton Mifflin, New York

Goodman E J 1992 (1972) *The Explorers of South America*. University of Oklahoma Press, Norman, OK

Guss D M 1999 *The Festive State: Race, Ethnicity, and Nationalism as Cultural Performance*. University of California Press, Berkeley, CA

Hill J (ed.) 1988 *Rethinking History and Myth: Indigenous South American Perspectives on the Past*. University of Illinois Press, Urbana, IL

Richardson J B III 1994 *People of the Andes*. St Remy Press, Montreal, Canada and Smithsonian Press, Washington, DC

Stern S J (ed.) 1987 *Resistance, Rebellion, and Consciousness in the Andean Peasant World*. University of Wisconsin Press, Madison, WI

Van Cott D L (ed.) 1995 *Indigenous Peoples and Democracy in Latin America*. St Martin's Press, New York

Whitten N E Jr, Torres A (eds.) 1998 *Blackness in Latin America and the Caribbean: Social Dynamics and Cultural Transformations*, 2 Vols. Indiana University Press, Bloomington, IN

Williams J M, Lewis R E (eds.) 1993 *Early Images of the Americas: Transfer and Invention*. University of Arizona Press, Tucson, AZ

N. E. Whitten, Jr

South Asia, Archaeology of

The archaeology of South Asia, or the Indian subcontinent, is a very substantial topic, since there is a deep history in this land of diverse peoples and cultural traditions. Moreover, archaeological fieldwork is conducted on a substantial scale, with a dozen or more excavations each year, and widespread exploration. Since it is impossible to be comprehensive in an article such as this, the article will focus on four topics (the Palaeolithic, Early Village Farming Communities, the Indus Civilization, and the Early Iron Age) that are at least representative of the principal problems on which contemporary archaelogy has focused.

1. Palaeolithic

Unexpectedly early stone tools were found at Riwat, in the Soan Valley, about 6.2 miles (10 km) south-east of Rawalpindi between 1983 and 1988 (Allchin 1995) (see Fig. 1). Twenty-five pieces of flaked quartzite were recovered. Of these, only three are undeniably artifacts; that is, objects shaped by humans. Other pieces from Riwat might be the products of human activity, but they are so simple that this cannot be determined with certainty. Two of these artifacts are cores, the third is a flake. They are generally similar to the Oldowan industry of Africa, which dates to between two and three million years ago. The geological deposits of the artifacts at Riwat have reversed polarity and are therefore associated with the Matuyama Chron, between 780,000 and 2.60 million years ago. Volcanic ash immediately above the Riwat artifacts dated to c. 1.8 million years ago. The artifacts are therefore older than that; how much older is somewhat problematic. An excavation was undertaken in 1990 in the Pabbi Hills, to the south-east of Riwat. This excavation yielded tools in a deposit that was palaeomagnetically dated to 1.2 to 1.4 million years ago.

The Lower Palaeolithic (c. 1.5 million to 200,000 years ago) is represented in virtually all regions of India, but not Sri Lanka. It is a period of two major traditions of early tool making, the western core biface (hand axe/cleaver) tradition and the eastern chopper/chopping tool tradition. The line defined by Professor Hallam Movius that follows the foothills of the Himalaya mountains down the Ganges Valley, dividing these two traditions, is a concept that has returned favor. The distribution of these tool-making traditions

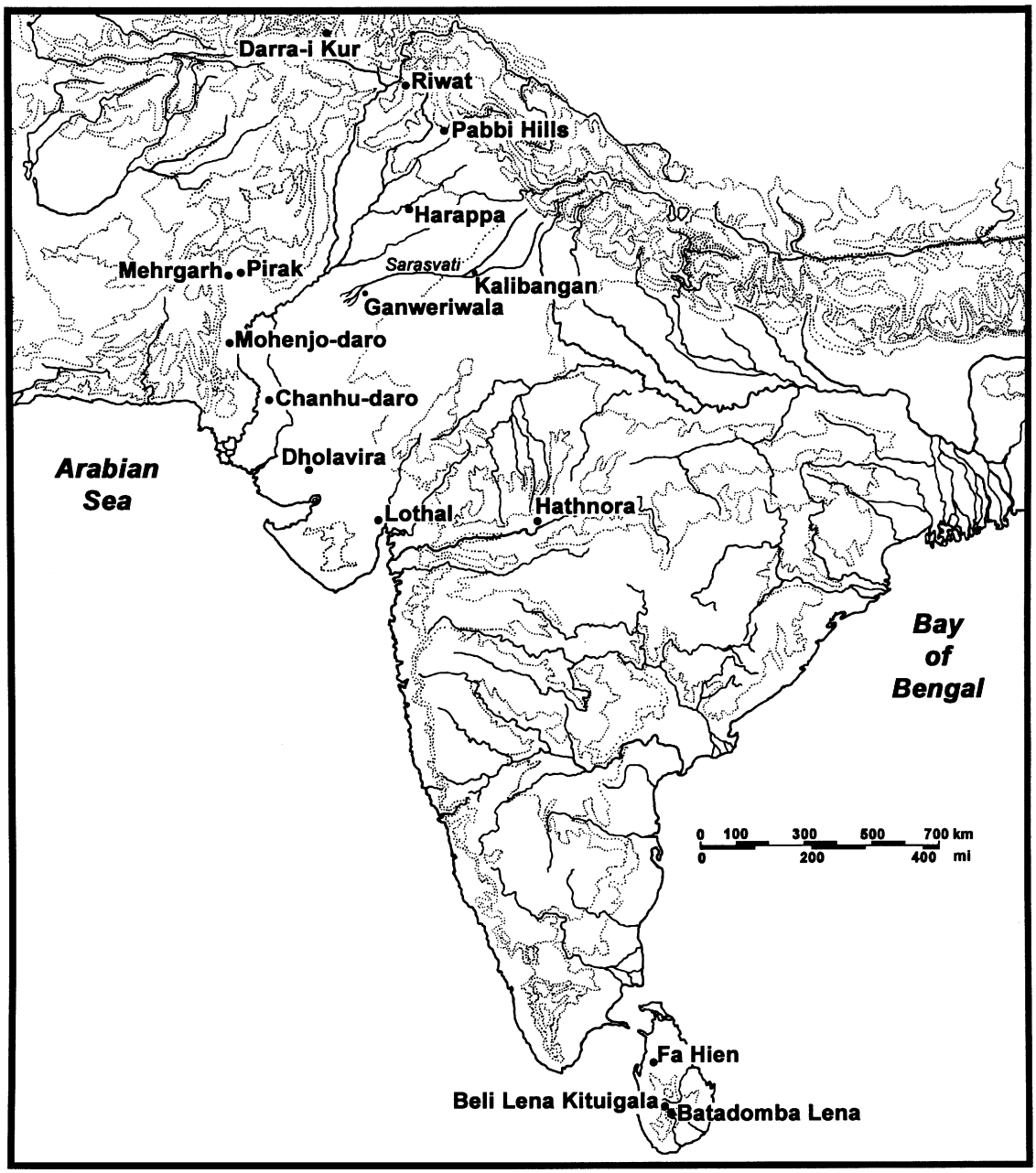

Figure 1
Sites and places mentioned in the text

seems to be coincident with the distribution of bamboo, which could have been a major source of tools in the Lower Palaeolithic.

A Middle Palaeolithic (250,000–30,000 years ago) stone-flake tool industry is widely distributed in India, Pakistan and Afghanistan. Some sites have evidence of the 'Levallois,' technique. The cave of Darra-i Kur in north-eastern Afghanistan produced a fossil that is anatomically a modern *Homo sapiens* dated to c. 30,000 BP (Kennedy 2000).

Sites with Upper Palaeolithic blade and burin tools are not well represented in the Indian subcontinent.

Establishing a chronology for such a tool-making tradition, and documenting its distribution and relationship with the Middle Palaeolithic and the development of microlithic technology, is one of the major challenges in the prehistory of South Asia.

The oldest fossil hominid in South Asia is an archaic *Homo sapiens* from Hathnora on the Narmada River, near the famous palaeolithic site at Hoshangabad (Kennedy 2000, pp. 1173–80). This find is not associated directly with stone tools, but the deposits from which it was derived are Upper Pleistocene and date from about 200,000 to 250,000 years ago. Representatives of anatomically modern *Homo sapiens* have been found in Sri Lanka in the caves of Fa Hien, Batadomba Lena and Beli Lena Kituigala (Kennedy 2000, pp. 180–8). Radiocarbon dates have shown that the Fa Hien specimens date to c. 33,000 BP and are associated with a non-geometric microlithic tool industry. The fossil humans from Batadomba Lena date to c. 28,000 BP and are associated with microliths, bone tools, and faunal and floral remains.

2. *Early Village Farming Communities*

Insights into the beginnings of food production in South Asia have been found at Mehrgarh (Jarrige et al. 1995), on the Kachi Plain of the Indus Valley. Period I is aceramic, with small rectangular houses, and interesting multiroomed buildings thought to have been storehouses or granaries. A new periodization of Mehrgarh emerged from the 1998 excavations there. Period I now has three phases: IA and IB, with habitation, and IC, during which the excavated area was used as a cemetery (Meadow 1998, p. 13).

The date for Period I at Mehrgarh has to be inferred, somewhat unsatisfactorily, from eight radiocarbon samples that range from c. 4000 BC to older than 8000 BC. The early ceramic Neolithic of Period IIA is reasonably well dated to c. 5300–4700 BC, by seven calibrated determinations, thus the beginnings of Mehrgarh IA at 7000 BC, is a reasonable estimate, given the 11 building levels there. More than 90 percent of the food grains recovered from Period I are naked six-row barley, along with domestic hulled six-row barley, and wild and domestic hulled two-row barley (Meadow 1998). The naked barley is probably cultivated and on its way toward full domestication in Period I. There are also three forms of domesticated wheat, including emmer, einkorn and a third, free-threshing, form.

The animals from Period I include wild forms of sheep, goats and gazelle, along with other ungulates that were never domesticated. Two burials of Period IC contain the remains of five kids in each, taken to indicate their domesticated status. This inference is corroborated by the remains of relatively small subadult or adult goats in trash deposits. After gazelle, goats were found to be the most common animal in the Period IA. Over the course of Period I, sheep and cattle rise in proportion to other animals; and by the end of Period IB cattle bones comprise over 50 percent of the remains. It is thought that increasing representation and decreasing body size 'strongly supports a hypothesis of local domestication. In addition, both osteological and figurine evidence indicate that zebu (humped) cattle (*Bos indicus*) were present and likely to have been the dominant form' (Meadow 1998, p. 16).

It is clear from the Mehrgarh data that cattle, sheep, and goats were domesticated locally. Wild barley has been found in the hills above Mehrgarh, and it too would therefore seem to have been part of the same process. The one remaining resource, wheat, has never been found in the wild on the eastern edge of the Iranian Plateau, but suggesting that it was 'imported' as part of an otherwise local process of domestication creates an untenable awkwardness in the interpretation of the Mehrgarh sequence.

It is also apparent that Mehrgarh does not document the beginnings of domestication, since both the plants and animals in the earliest levels there are either domesticated, or in the process of being so. The early beginnings of this process should be in the hills above Kachi, where the sheep, goats, and cattle as well as barley (and wheat?) would be found in the wild.

Pursuing this line of inference is one of the most exciting challenges of contemporary South Asian archaeology.

It was sites like Mehrgarh and their predecessors that laid the foundations in subsistence productivity that led to the beginnings of urbanization in the Indian subcontinent.

3. *The Indus Civilizaiton*

3.1 *Introduction*

South Asia's earliest cities were a part of the Indus, or Harappan, Civilization, which arose on the plains of the Greater Indus Valley of Pakistan and northwestern India in the middle of the third millennium (c. 2500–1900 BC). The Harappan peoples were, for the most part, farmers, herders, and skilled craftspeople who lived in the cities of Mohenjo-daro, Harappa, and Ganweriwala. The Harappan Civilization covers about one million square kilometers (Possehl 1999, Allchin and Allchin 1997; see Possehl 1999, pp. 38–154 for the story of its discovery).

The Indus Civilization can be defined by its ideology, which emerges as a grand act of creation. Some parts of older traditions survived, since it is clear that there was little change in subsistence, and in the sociocultural, economic and political balances between farmer and pastoralists. But Harappan ideology was based in nihilism. The new social order was founded in urbanization and sociocultural complexity.

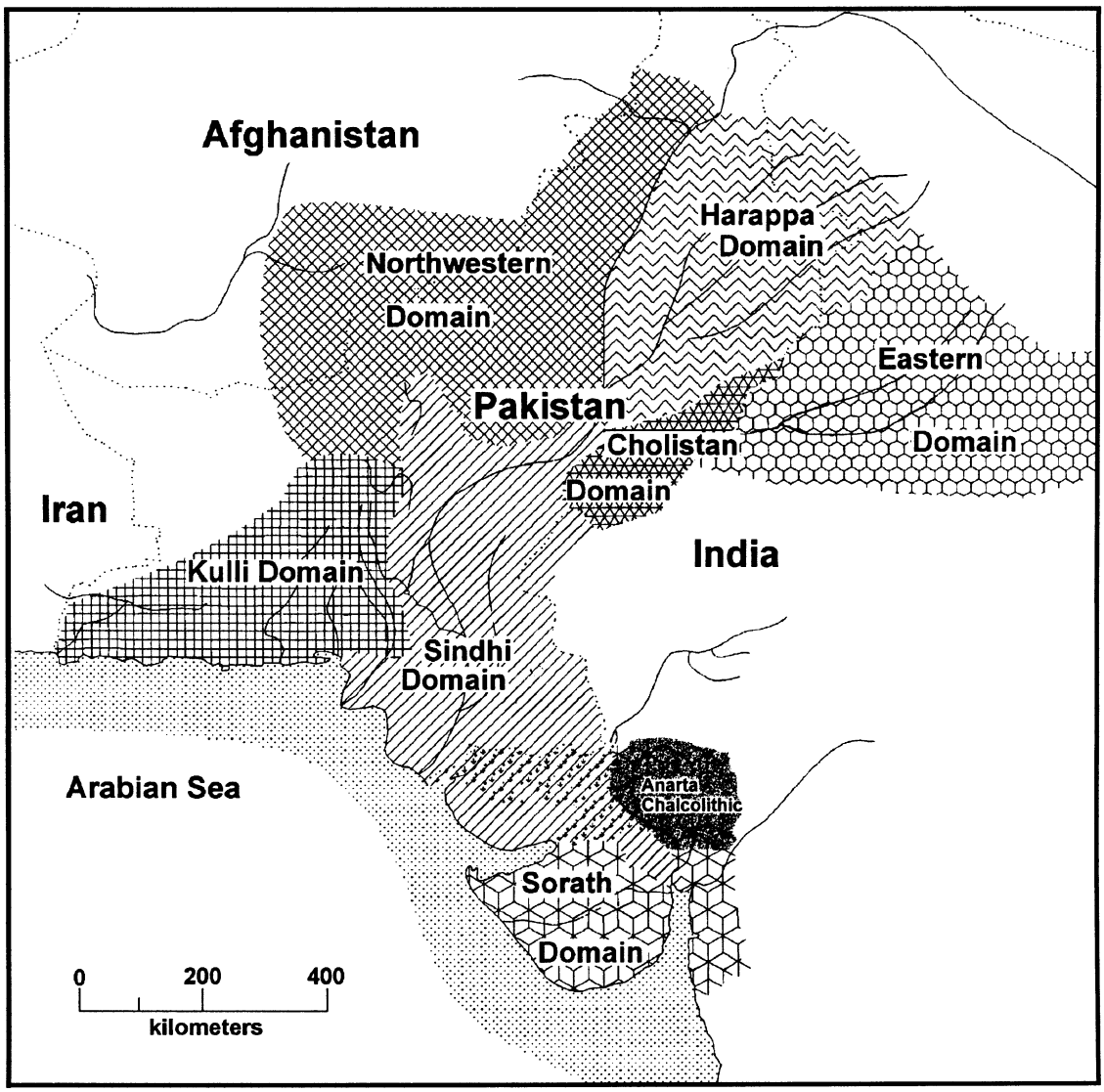

Figure 2
Domains of the Harappan civilization

Water and 'cleanliness' were powerful forces in the Mature Harappan ideology. The Mature Harappans were technological innovators.

3.2 Geography

The ancient cities of the Indus were built in riverine settings, in the lower Indus River, the Ravi of the Punjab, and the ancient Sarasvati. The land was quite different in the third millennium, since the now-dry Sarasvati, or Ghaggar-Hakra, was an active stream, and the rolling uplands between the rivers of the Punjab were unirrigated scrub grasslands. Harappan peoples are evidenced from the Iranian border to the Aravalli Mountains, and the Arabian Sea to the Himalayas. There is a Harappan outlier site on the Amu Darya in northern Afghanistan known as Shortughai. In an effort to deal with the cultural geography of the Indus Civilization, their territory has been divided into 'domains'—geographical and cultural regions, possibly with political significance (see Fig. 2).

3.3 Urban Origins

The beginnings of the Harappan Civilization are still not well understood. There was an Early Harappan

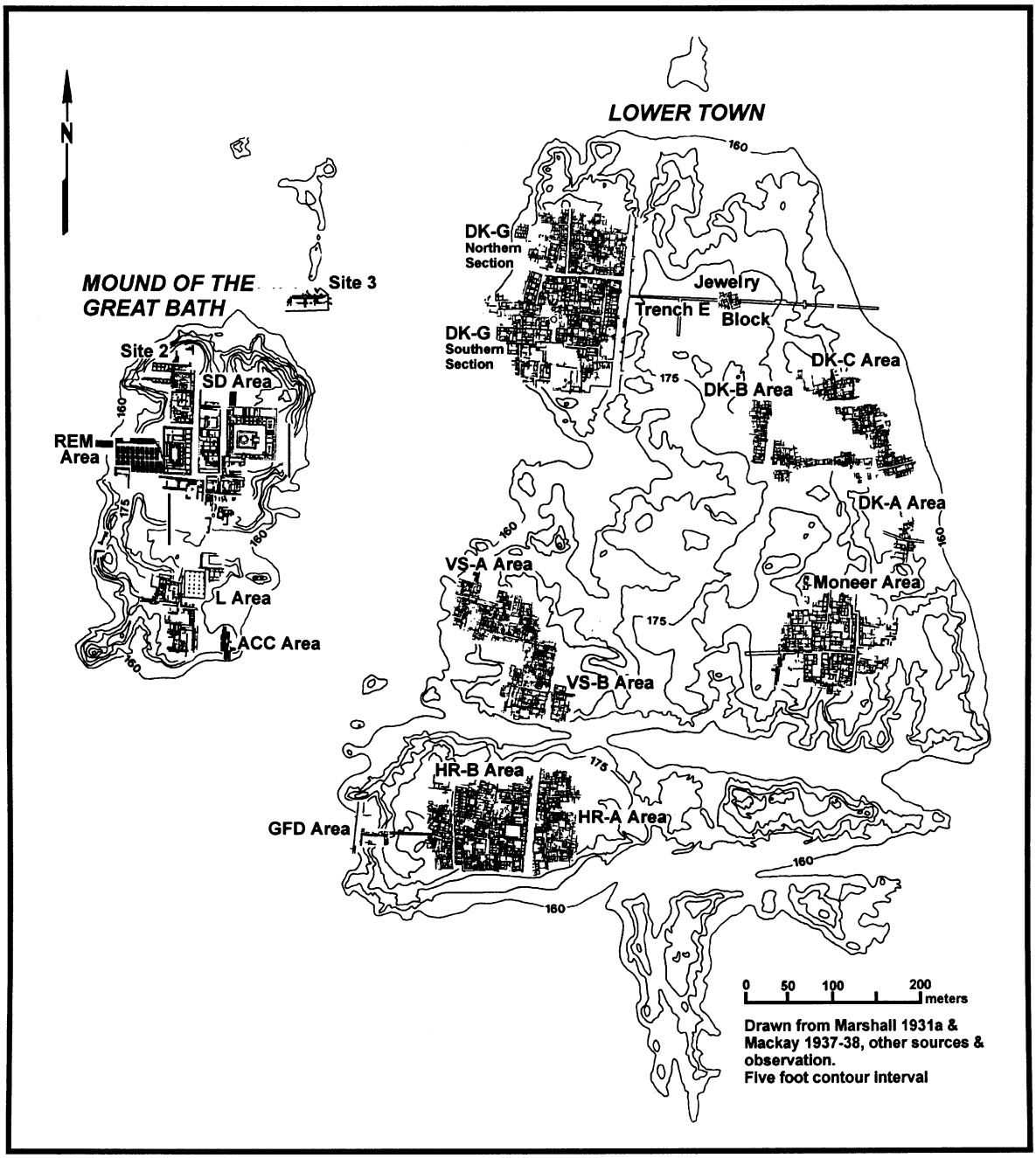

Figure 3
Plan of Mohenjo-daro

period (c. 3200–2600 BC) of four roughly contemporaneous phases (Kot Dijian, Amri-Nal, Damb Sadaat, and Sothi-Siswal) without large settlements or strong signs of social stratification. Craft specialization during the Early Harappan was not developed to a marked degree. There was a Transitional Stage between the Early Harappan and Mature Harappan at about 2600–2500 BC, during which most of the complex sociocultural institutions of the Indus Civilization came together (Possehl 1999, pp. 713–23).

3.4 Major Settlements

The major excavated settlements of the Harappan Civilization are Mohenjo-daro and Harappa, each of about 100 hectares, with estimated ancient populations of about 20,000. The third city, Ganweriwala, in Cholistan, is unexcavated.

Mohenjo-daro is the best preserved of the Indus cities (see Fig. 3). It was a planned settlement from the start, and was probably built in the middle centuries of the third millennium BC. The majority of the buildings were constructed of burnt brick, which was served to preserve the site in an extraordinary way. A Lower Town was situated to the west, laid out in a grid, with major thoroughfares running north–south and east–west. Multiroom, and probably multistoried buildings line these streets. The private houses were generally provided with bathing facilities, which were connected to an integrated civic drainage system. These bathing facilities were not privies. To the west of the Lower Town was the Mound of the Great Bath, an artificial hill with a series of public edifices on it. The most important of these were a Great Bath, probably used by some portion of the elite population as a place of ritual ablution, and a warehouse, once thought of as a State Granary.

Other major sites are:

Kalibangan, a regional center of the Indus Civilization near the present border between India and Pakistan on the ancient Sarasvati. It documents the succession from the Early to the Mature Harappan (Thapar 1973).

Chanhu-daro, a former Harappan town, located 70 miles south of Mohenjo-daro. It is important for the presence of a bead- and seal-making workshop (Mackay 1943).

Dholavira, located in Kutch, is midway between Chanhu-daro and Lothal. This was an important way station between major Mature Harappan regions, and was 60 hectares in size. It was fortified, with important water storage and drainage facilities. There is a stratigraphic succession from the Early Harappan to the Mature Harappan and succeeding cultures (Bisht 1999).

Lothal, located in Gujarat at the head of the Gulf of Cambay, provides an understanding of Mature Harappan trade and craftsmanship, and the stratigraphic relationship between the Mature Harappan and the Posturban Harappan in the region. The large brick-lined enclosure at the site, once described as a dockyard, was probably a water storage facility (Rao 1979, 1985).

3.5 Writing

The Indus peoples mastered the art of writing, which was rendered in pictographic form on a range of physical objects. The best known of these are the Harappan stamp seals with writing and an animal device (see Fig. 4). In spite of many claims to the contrary, the Indus script remains undeciphered (Possehl 1996) and the names we use for their settlements are not likely to be those of the Indus peoples themselves.

3.6 Subsistence

Most of the Harappan peoples were farmers and herders who also hunted and fished, with some gathering mixed in. The chief food grains were barley and wheat. They also cultivated at least two forms of gram, the chickpea, field pea, mustard, sesame cotton, and dates. The evidence for the cultivation of rice during the Mature Harappan is ambiguous, but possible. The Harappans used grapes.

The peoples of the Indus Civilization were cattle keepers. Cattle remains are consistently above 50 percent of the faunal assemblage, and often much more. Cattle imagery is also prominent, and makes it clear that this was the most important animal in their culture as well as probably the principal form of wealth. The Harappans also kept water buffalo, sheep, goats, and pigs. Fish, both marine and freshwater, were traded over wide areas.

3.7 Craft and Trade

The Indus craftsmen plied their trades with great skill and had a mastery of pyrotechnology that was a model for its age. Their pottery was fashioned using a number of forming techniques, often in multiple parts that were welded together prior to firing. Bead making was an Indus forte, especially their long, barrel-shaped carnelian examples, sometimes etched. They were accomplished smiths, who smelted copper and tin, as well as arsenic to make bronze. They used gold, silver, electrum, lead, and antimony. Maritime shells, especially the conch, were formed into bangles, ladles, inlays, and other objects. Bangles and other objects were formed also from stoneware and faience, another of the Harappan specialties.

The Mature Harappan peoples were participants in the Middle Asian Interaction Sphere which linked the regions from the Indus to the Mediterranean and Bactria/Central Asia to the Arabian Gulf. This interaction sphere has deep roots and began much earlier than the Mature Harappan, but was at its peak during the second half of the third millennium BC. Maritime trade with Mesopotamia is well documented, both textually and archaeologically. Mature Harappan peoples were also in the Arabian Gulf. The merchandise that the Indus peoples supplied to Mesopotamia, according to the cuneiform texts, are: carnelian, lapis lazuli, pearls, woods, fresh dates, a bird, a form of dog, cats, copper, and gold. The products traded to Meluhha are not as clearly documented, but

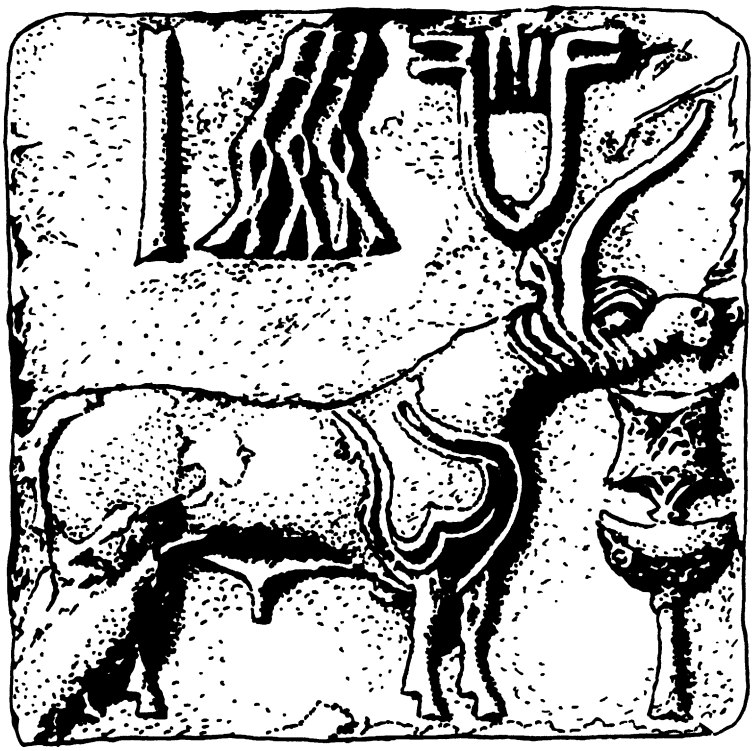

Figure 4
A Harappan stamp seal

included food products, oils, cloth, and the like. There are many Harappan artifacts in Mesopotamia, including seals, etched carnelian beads, and ceramics. The Mesopotamian products found in Harappan contexts at Indus sites are very few and there is a considerable disparity in the archaeological record of this subject.

3.8 Religion

The central theme of Harappan religion is a male deity usually shown with bull or water-buffalo horns (see Fig. 5), matched by a number of different female images (Possehl in press). At times, notions of gender may be merged. The plant and animal imagery of the Indus Civilization can be seen as specific aspects of the great duality of the Harappan Great Tradition. The multiheaded animals, 'unicorns' with elephant trunks, perhaps 'unicorns' themselves, as in the terracotta figurines from Chanhu-daro, are all themes appropriate to zoolatry. So are the tigers with bull horns and half human-half quadrupeds seen on the seals, such as the cylinder from Kalibangan (see Thappar 1973, p. 29) (see Fig. 6).

There is widespread evidence for the place of animals in Harappan ideology. Two Indus sealings, in fact, show small animal figures carried above the crowd in a procession. This is just another minor observation that supports the idea that animals were a source of wealth for these people, and entered their ideological world as well. There are many portrayals of composite animals, some with three heads on one body, or tigers entwined, human torsos on four-legged bodies, 'minotaur-like,' tigers with horns, 'unicorns' with elephant trunks and 'unicorns' growing out of trees. There is no good evidence for fire-worship.

Two iconographic themes that appear on Harappan seals have parallels in Mesopotamian mythology, both related to the Gilgamesh epic. The first is shown on a seal from Mohenjo-daro with a half-human female half-bull monster attacking a horned tiger (see Fig. 7). This Harappan seal could be seen as portraying the story of the Mesopotamian goddess Aruru, who created Ebani or Enkidu as a bull–man monster to do combat with Gilgamesh, but who ultimately became his ally and fought with Gilgamesh against the wild animals. The second motif is the well documented Mesopotamian combat scene, with Gilgamesh fighting off rampant animals on either side.

F. R. Allchin sees in a seal from Chanhu-daro a representation of Heaven, (the Bull), who is at once the consort and the father of Earth; and of Earth, who is at the same time the consort of the Bull (Heaven),

Figure 5
The great Harappan male god, once called the 'Proto-Shiva'

Figure 6
Kalibangan cyclinder seal

Figure 7
A possible representation of the Mesopotamia goddess Aruru in combat with a tiger

and the mother of the Bull, her calf; and that these themes can be understood by reference to the Creation myths found in the *Rgveda* and *Atharvaveda* (Allchin 1985, p. 381). This is perhaps our best evidence for a connection between a belief in the Bronze Age being carried forward into historical India.

3.9 The Eclipse of the Ancient Cities of the Indus

Mohenjo-daro was largely abandoned at the end of the Mature Harappan (c. 1900 BC). Similar evidence is available at Harappa, although there is a small occupation there of the so-called 'Cemetery H people.' The situation at Ganweriwala is less clear, since there has been no excavation there, but surface prospecting indicates that the occupation was limited to Mature Harappan times.

Older theories which hold that the cities and civilization were destroyed by invading Aryan tribes, as depicted in the Rig Veda, make very little sense. This is in part because there is no evidence for the sacking of any of the Mature Harappan settlements. Nor is there good chronological agreement between the date of the Vedic texts and the changes seen so graphically at Mohenjo-daro and Harappa. The proposition that a natural dam formed across the Indus River in Sind and flooded out the civilization has been critiqued widely and is not a viable proposition (Raikes 1964, Wasson 1987).

The evidence from the three urban centers, as well as regional surveys, indicates that Sind and West Punjab experienced the widespread abandonment of Mature Harappan settlements at the opening of the second millennium BC. A summary table of the substantive data comparing Mature Harappan and Posturban archaeological assemblages by regions is given in Table 1.

These figures indicate that the 'eclipse of the Harappan Civilization' might hold for Sindh and Baluchistan, but in other regions, notably in the east and to a lesser degree in Gujarat, the course of culture change was different. In these areas there were strong lines of continuity through the early centuries of the second millennium with little, if any, of the 'trauma' that effected Sindh. The stark image one has for Baluchistan in the second millennium BC represents a clear challenge for field archaeology, because it would not seem reasonable to presume that the entire area was deserted at that time. Excavation and exploration are needed to give a more realistic sense of the culture history at that time (Possehl 1997).

Archaeological documentation supports the notion that the locus of culture change for the transformation of the Indus Civilization was urban, those physical settings and institutions most closely associated with sociocultural complexity, and the Mature Harappan ideology. In some places thought of as a part of the Harappan Civilization, this transformation went almost undocumented, as in Saurashtra and the Eastern Domain. There are also strong lines of cultural continuity that link the Harappan Civilization with the early Iron Age and historical South Asia. The

Table 1
Site counts and settled area for the mature and posturban Harappan

	Site count	Average site size (hectares)	Settled area (hectares)
Sindh			
Mature Harappan	86	8.0	688
Posturban	6	5.6	34
Cholistan			
Mature Harappan	174	5.6	974
Posturban	41	5.1	209
Baluchistan			
Kulli-Quetta/Harappan	129	5.8	748
Posturban	0	0	0
Saurashtra			
Sorath Harappan	310	5.4	1674
Posturban	198	4.3	815
East			
Mature Harappan	218	13.5	2943
Posturban	853	3.5	2985

reasons behind the transformation of the Harappan Civilization are still not understood, but it is quite possible that the ideology on which it was based had lost its strength. This might be documented archaeologically by the abandonment of the Great Bath at Mohenjo-daro 200, or even 300, years prior to the abandonment of the city by 1900 BC. It should also be noted that the locus of change during the transformation was urban, and thus centered on Mature Harappan ideology.

4. *The Early Iron Age*

Four archaeological assemblages document the Early Iron Age in South Asia: the Painted Grey Ware Assemblage, the Gandharan Grave Culture, the Pirak Assemblage, and the Megalithic Complex (Possehl and Gullapalli 1999).

4.1 *The Painted Grey Ware Assemblage*

The Painted Grey Ware Assemblage is quite different from that of the Harappan and Post-urban Harappan that preceded it in Northern India and Pakistan. This difference, and the association with iron, tended to set the Painted Grey Ware apart, causing some archaeologists to think of it as intrusive, even foreign to the Indian subcontinent. But archaeologists from the Archaeological Survey of India have done much to dispel this impression through their explorations and excavations in the Indian Punjab. Their excavations at Bhagwanpura and Dadheri have found evidence for Post-urban Harappan occupations followed by an overlap between the Post-urban Harappan and the Painted Grey Ware (Joshi 1993). Thus the culture-historical gap that once separated the Early Iron Age from the preceding Bronze Age in Northern India and Pakistan has now been closed and there is good evidence for cultural continuity between the two periods. Radiocarbon dates indicate that Painted Grey Ware begins around 1000 BC and persists to the middle of the first millennium BC, possibly to 300 BC.

Iron is widely associated with Painted Grey Ware, at times in considerable quantity. At Atranjikhera in the Ganges Valley, iron objects occurred in profusion, along with furnaces, slag, and tools used by smiths, suggesting that iron goods were manufactured at the site, along with smelting. About 500 Painted Grey Ware sites have been discovered. They extend from the Bahawalpur region of Pakistan east across the Punjab into Uttar Pradesh in India. The largest Painted Grey Ware site is Satwadi in Pakistani Cholistan at 13.7 hectares, far from what one would expect for a city. Most of these sites are in the one- to two-hectare range. Since iron is not mentioned in the Rig Veda, the earliest text of South Asia, there is little or no chance that the peoples who occupied the Painted Grey Ware sites were the Vedic Aryans.

4.2 *The Gandharan Grave Culture*

A series of cemeteries and associated habitation sites in northern Pakistan have produced evidence for Late Bronze Age and Early Iron Age life in the eastern portion of ancient Gandhara. Most of the excavations have been conducted at the cemetery site of Timargarha and the settlement of Balambat in northern Pakistan, and in Swat at places like Aligrama, Loebanr I, and Katelai. Other sites are distributed through the southern valleys of the Hindu Kush, and possibly Ladakh.

The earliest of the graves, Period 1 of the Gandharan Grave Culture can be assigned to the Late Bronze Age (c. 1500–1000 BC) and consist of pits lined and covered by stones. Flexed burials with hand-made red and gray pottery, and bronze artifacts were found at Timargarha, followed by evidence for cremation in Period II (1000–800 BC). The latest series of graves, Period III of the Gandharan Grave Culture (800–500 BC), contain iron as well as artifacts of other metals. At Timargarha, seven objects of iron are reported, including a spoon, spear head, and a nail, and, interestingly enough, an iron cheek piece from a snaffle bit that dated between the tenth and the sixth centuries BC.

4.3 The Pirak Assemblage

The site of Pirak is on the Kachi Plain 20 kilometers east of Mehrgarh. There are only four sites of this assemblage, with a few other small sherd scatters in Kachi. Excavations at Pirak took place between 1968 and 1974 (Jarrige and Santoni 1979, Enault 1979). Period I (c. 1750–1200 BC). The site has the characteristic Pirak pottery and well built houses. Stray sherds of Wet Ware, Faiz Mohammad Grey Ware, and other third millennium wares mark some continuity with the Harappan cultural tradition. A compartmented copper stamp seal of Period II does the same. The subsistence regime was based on rice, millets, and other cultivars. There is evidence for the Bactrian camel and an equid, possibly the domesticated horse, in the form of figurines.

Period II (c. 1200–1100 BC) has continuity with Period I, and is defined on the basis of new ceramic shapes and fabrics. Copper/bronze tools continue along with the original subsistence regime. Iron is introduced in Period III (1100–900 BC). Most of the iron was found in a large, basically quadrangular set of buildings constructed directly upon the remains of Period II. The iron artifacts, mostly points and 'bars' are complemented by the appearance of crucibles and iron slag, as well as copper/bronze, lead, silver, and gold artifacts. The appearance of certain gray wares in Period III is important in the definition of this phase of occupation.

4.4 The Megalithic Complex

The Megaliths of South Asia are an immense field of study, far too large to be covered adequately in this short survey. Burial monuments incorporating large stone constructions, associated with habitation sites, are known from prehistoric times in many regions of the Indian subcontinent, principally the north-western mountainous areas of Kashmir and Almora, the Vindhyas, the Deccan Plateau, and the deep south of Peninsular India. Aspects of this tradition may still be present in the subcontinent, as among the Nagas and Khasis of the north-east, and other tribes of the Chota Nagpur, and Bastar regions. The connections have not yet been sorted out for this complex, deep set of historical relationships. Archaeological stratigraphy, radiocarbon dates, and the study of associated materials indicates that the beginnings of Megaliths in south Asia is in the very late second millennium BC, and comes with the mass production of iron. The earliest iron implements associated with the Peninsular Indian Megalithic are simple things like arrow heads, daggers, and domestic vessels. They are also found in the transitional times between this cultural complex and the preceding South Indian Neolithic.

Megalithic burial monuments in South India include dolmenoid cists, slabbed cists, shallow-pit burials, deep-pit burials, umbrella-stones and hat-stones, hood-stones, multiple hood-stones, and menhirs. In excess of 1400 Megalithic sites are known in the Indian subcontinent, with 1116 of these in Peninsular India, about 240 of which have been excavated in some way. Most of these monuments contain human remains and associated burial furniture, sometimes in great quantity. Horses and vehicles are present, along with pottery and metal artifacts including tools, horse trappings, Roman coins, and other artifacts. Iron occurs in abundance. Human remains from the burials have been studied by Kennedy (2000, pp. 326–57).

See also: Hunter–Gatherer Societies, Archaeology of; South Asian Studies: Culture

Bibliography

Allchin B 1995 The Potwar Project 1981 to 1993: A concluding report on the British Archaeological Mission to Pakistan's investigations into hominid and early Human cultures and environments in the northern Punjab, Pakistan. *South Asian Studies* **11**: 149–56

Allchin B, Allchin R 1997 *Origins of a Civilization: The Prehistory and Early Archaeology of South Asia*. Viking, New Delhi

Allchin F R 1985 The interpretation of a seal from Chanhu-daro and its significance for the religion of the Indus Valley. In: Schotsmans J, Taddei M (eds.) *South Asian Archaeology 1983*. Instituto Universitario Orientale, Dipartimento di Studi Asiatici, Naples; Italy, Series Minor 23, pp. 369–84

Bisht R S 1999 Dholavira and Banawali: Two different paradigms of the Harappan urbis forma. *Puratattva* **29**: 14–37

Enault J-F 1979 *Fouilles du Pirak*, 2 Vols. Fouilles du Pakistan 2(2). Publications de la Commission des Fouilles Archaeologique, Paris

Jarrige J-F, Santoni M 1979 *Fouilles de Pirak*, 2 Vols. Fouilles du Pakistan 2(1). Publications de la Commission des Fouilles Archaeologique, Paris

Jarrige C, Jarrige J-F, Meadow R H, Quivron G 1995 *Mehrgarh: Field Reports 1974–1985, from Neolithic Times to the Indus Civilization*. Department of Culture and Tourism of Sindh, Pakistan, Department of Archaeology and Museums, French Ministry of Foreign Affairs, Karachi, Pakistan

Joshi J P 1993 *Excavation at Bhagwanpura 1975–76 and Other Explorations and Excavations 1975–81 in Haryana, Jammu & Kashmir and Punjab*. Memoirs of the Archaeological Survey of India, 89, Delhi
Kennedy K A R 2000 *Gods-Apes and Fossil Men*. University of Michigan Press, Ann Arbor, MI
Mackay E J H 1943 *Chanhu-daro Excavations 1935–36*. American Oriental Series 20. American Oriental Society, New Haven; CT
Meadow R H 1998 Pre- and proto-historic agricultural and pastoral transformations in northwestern South Asia. *Review of Archaeology* **19**(2): 12–21
Possehl G L 1996 *Indus Age: The Writing System*. University of Pennsylvania Press, Philadelphia, PA
Possehl G L 1997 The transformation of the Indus Civilization. *Journal of World Prehistory* **11**(4): 425–72
Possehl G L 1999 *Indus Age: The Beginnings*. University of Pennsylvania Press, Philadelphia, PA
Possehl G L in press The religion of the Indus civilization. In: Hinnells J R (ed.) *Penguin Encyclopaedia of Ancient Religions*. Penguin Books, London
Possehl G L, Gullapalli P 1999 The Early Iron Age in South Asia. In: Pigott V (ed.) *The Archaeometallurgy of the Asian Old World*. University Museum Monograph 89, MASCA Research Papers in Science and Archaeology Vol. 16. The University Museum, University of Pennsylvania, Philadelphia, PA, pp. 153–75
Raikes R L 1964 The end of the ancient cities of the Indus. *American Anthropologist* **66**(2): 284–99
Rao S R 1979 *Lothal: A Harappan Port Town, 1955–62*. Memoirs of the Archaeological Survey of India 78(1), Delhi
Rao S R 1985 *Lothal: A Harappan Port Town, 1955–62*. Memoirs of the Archaeological Survey of India 78(2), Delhi
Thappar B K 1973 New traits of the Indus Civilization at Kalibangan: An appraisal. In: Hammond N (ed.) *South Asian Archaeology*. Noyes Press; Park Ridge, 11, pp. 85–104
Wasson R J 1987 The sedimentological basis of the Mohenjo-daro flood hypothesis – a further comment. *Man and Environment* **11**: 122–23

G. L. Possehl

South Asia: Sociocultural Aspects

1. Area and Population

The term 'South Asia' has come to stand for a group of seven countries: India, Pakistan, Bangladesh, Nepal, Bhutan, Sri Lanka, and the Maldives; and it is in that sense that it will be used in this article. These countries differ greatly in area and population, and in geographical aspect and location. India, with over a billion inhabitants, has the second largest population in the world; the Maldives have a population of only a few hundred thousand people. Nepal and Bhutan are entirely landlocked; Sri Lanka is an island in the Indian Ocean; and the Maldives comprise a chain of more than 1000 islands, most of them uninhabited.

The people of these countries have a shared past by which many of their present social and cultural characteristics have been shaped. Demographically, economically, and politically, India, Pakistan, and Bangladesh are the three most important countries of the region. Until 1947 they were one country. It was partitioned in that year, when India and Pakistan became separate and independent nations. Pakistan was in turn partitioned in 1971, leading to the creation of Bangladesh as a separate country. The first partition was occasioned by differences of religion, and the second by differences of language. In each case the partition was attended by considerable movements of population.

The Indian subcontinent, comprising India, Pakistan, and Bangladesh, was under British rule for a long time, as was also Sri Lanka. The two landlocked kingdoms of Nepal and Bhutan never came directly under British rule and remained relatively isolated from the economic and political forces released by it. Although many social and cultural changes have taken place in the region since the end of colonial rule in the middle of the twentieth century, large parts of it still remain economically backward.

The subcontinent is the home of a very ancient and distinct civilization that may be described as the 'pan-Indian' or the 'Indic' civilization. This civilization had its origin in the Indo–Gangetic plain in the second millennium BC, and it antedates the birth of both Christianity and Islam. Its characteristic bearers today are the Hindus. It has undergone many changes, and its influence has not only permeated throughout India and its neighboring countries but also extended to lands beyond the seas such as Indonesia, Cambodia, and Thailand. India is not only the home of Hinduism, it also gave birth to Buddhism which is the official religion of Bhutan and the predominant religion of Sri Lanka.

2. Religion and Language

Although Hinduism has had the longest and most pervasive influence on society and culture in the subcontinent, Islam too has a strong presence in it, including present-day India. Pakistan is an Islamic country; Bangladesh has a predominantly Muslim population; and although the Muslims are a minority in India, they are a very large minority. The medieval phase of Indian history was dominated by Muslim sultanates and empires; and the Muslim population of Bangladesh, Pakistan, and India taken together make up a considerable part of the Muslim population of the world as a whole.

In South Asia, the divisions of religion cut across those of state and nation, and the region as a whole may be viewed as a mosaic of religious communities, sects, and denominations. India provides an extreme example of co-existence of diverse religious faiths and practices. Although the Hindus predominate in the

country as a whole, there are districts in which Sikhs, Muslims, or Christians predominate. Again, Nepal is a predominantly Hindu country, but there are areas in which Buddhists predominate; in Sri Lanka it is the other way around.

South Asia is the home not only of many religions but also of innumerable languages, and, whether on the plane of cultural values or of social identity, language is as important as religion. Like the divisions of religion, those of language also cut across the boundaries of state and nation. The Punjabi-speaking community is divided between India and Pakistan; the Bengali-speaking community between India and Bangladesh; and the Tamil-speaking community between India and Sri Lanka. Sometimes, as in Sri Lanka, where the Sinhala-speakers are predominantly Buddhist and the Tamil-speakers predominantly Hindu, the divisions of language and religion follow the same line. But this is not always the case. It is common in the subcontinent, if not characteristic, for the divisions of language and religion to cut across each other; there are Muslims, Hindus, and Sikhs among the Punjabi-speakers in India and Pakistan, and Hindus and Muslims among the Bengali-speakers in India and Bangladesh.

Because of historical and demographic reasons India has the largest variety of languages among South Asian countries: it is indeed the most polyglot country in the world. There are more than a dozen literary languages, divided between the Indo–Aryan and the Dravidian families of language. In addition, there are numerous tribal languages and dialects belonging to the Austro–Asiatic, the Tibeto–Burman, and other families. In India, as in the subcontinent generally, the divisions of language correspond more closely to geographical divisions than do the divisions of religion; but here too, population movements through the centuries have led to a certain dispersal of language communities.

If the civilization of South Asia had a design, it was most clearly visible in the Indian subcontinent until its first partition in 1947. Visitors to the area from far and near have been for centuries struck by the distinctness and continuity of that design, and have commented on its influence on society and culture in the different countries of the region. Despite the changes and upheavals of the twentieth century, its distinctive features are still visible, most plainly in India, but elsewhere as well.

3. *Accommodation of Diversity*

A brief sketch of the design of the traditional civilization of the subcontinent will be provided as an aid to the understanding of the multitude of beliefs, practices, habits, customs, and manners prevalent in the region to this day. The first striking feature of this design is the accommodation of diversity. The rich ethnographic record available from the middle of the nineteenth century onward shows an endless, almost inexhaustible, variety in material culture, social arrangements, and religious beliefs and practices. Some of the variation in material traits, in matters such as food, dress, and habitation, may be explained by variations in geographical environment and natural resources. But the cultural variation is far in excess of what would be required by variations of natural environment. For instance, even where its basic ingredients are the same, food might be prepared and served differently among the different castes and communities in the same village.

Then there is the great diversity of social arrangements in matters relating to family, marriage, descent, inheritance, residence, and so on. What needs to be stressed is not simply the diversity of practices but also the diversity of rules; they were the despair of the colonial administrators of the nineteenth century who sought to codify the laws and customs of the subcontinent. There is finally the profusion of religious beliefs, doctrines, rituals, customs, and usages which have been mentioned.

The presence of diversity has been more than a matter of mere existence. In Indian society in particular and South Asian societies in general, it has been sustained by law, custom, and popular religion. Accommodation without assimilation has been a core value of Indian civilization until modern times. It has led to the co-existence of a large multiplicity of beliefs and practices as well as of social groups. Adding new components has not meant discarding old ones, and new and old components of the most heterogeneous kinds have existed cheek-by-jowl to a greater extent than in other civilizations. The Indian subcontinent has been and still largely remains a vast mosaic of tribes, clans, castes, sects, and communities, each with its own identity.

4. *Collective Identities*

The second distinctive feature of South Asian societies is the predominance of collective identities. In the past, what counted was not so much the individual as the group of which he was a member by birth, and this continues to a large extent even now. It has often been said that traditional Indian society was a society not of individuals but of castes and communities. As has been noted, there was a great multiplicity of groups, and this multiplicity was a feature even of small local communities. The members of a single village might observe different religious practices, speak more than one language, and belong to a dozen or more different castes and subcastes.

Collective identities were maintained by fidelity to the customs of one's ancestors and by rules for the regulation of marriage. In course of time, subtle differences developed in craft practices, dress, and

ornamentation, marriage rules and ritual observances between groups having the same origin. Each group and subgroup maintained its distinctive style of life and transmitted it from generation to generation. The group preserved a sense of its own identity despite changes in its social position. The high value placed on the accommodation of diversity allowed and indeed encouraged each group to retain a strong sense of its own identity.

The persistence of collective identities may be illustrated with two examples from the subcontinent. The first relates to what has been called 'the Hindu method of tribal absorption.' In the past, tribal groups which were relatively isolated from the wider society sometimes established relations with that society, discarding some old practices and acquiring some new ones. In this way, a tribe was often transformed into a caste but, despite a change of name and to some extent also of social position, it preserved its old collective identity. The second example relates to the survival of caste identities despite changes of religion. The census reports of British India show that in the Punjab, on both sides of the present international frontier, the same castes, such as Ahir, Jat, and Gujjar were present among Hindus, Muslims, and Sikhs. When a section of a Hindu caste embraced Islam or Sikhism, it adopted a new religion but did not abandon its old and distinct collective identity.

5. *Hierarchical Organization*

The accommodation of diversity did not mean that the tribes, clans, castes, sects, and communities that were allowed or even encouraged to maintain their distinctive identities were all placed on the same plane of equality. Diversity was organized on the basis of hierarchy and not of equality, and that is the third distinctive feature of the traditional civilization of South Asia. It is true that all communities were allowed or even encouraged to maintain their distinctive styles of life; but they were not all equally esteemed. While all premodern civilizations were hierarchical to a greater or lesser extent, the principle and practice of hierarchy were carried furthest on the Indian subcontinent, particularly among the Hindus.

Hierarchy is not simply a matter of inequality in the distribution of material resources or of life chances; it is above all a system of values according to which the division of the world into social groups deemed to be of unequal worth is a part of the natural scheme of things. The hierarchical conception of society found its fullest expression in the classical Hindu texts known as the *Dharmashastra* which provided the framework for Hindu law. Although the law has changed radically in the direction of equality, the bias of custom is still often in favor of hierarchy.

The Hindu caste system has been represented as the prototype of a hierarchical social order. But hierarchy extends beyond caste and permeates social relations and social institutions of the most diverse kinds among Hindus, Muslims, Buddhists, and others throughout the region. Despite the doctrine of equality characteristic of Islam, agrarian relations and gender relations are cast in a deeply hierarchical mould in Pakistan and to some extent also in Bangladesh.

6. *The Village*

Colonial administrators throughout the nineteenth century saw the subcontinent as essentially a land of villages. They have left behind memorable accounts of the unity, the self-sufficiency, and the continuity of the Indian village. A large number of village studies have been conducted in India by professional anthropologists and others since the country became independent in 1947. Similar studies have been conducted in Sri Lanka, Pakistan, Bangladesh, and Nepal. In the paragraphs below there follows a brief account of the broad social characteristics of the Indian village, many of which may also be found in villages in the other countries just mentioned.

In the past, the Indian village had a distinct physical identity which became particularly prominent during floods, famines, and other natural calamities. Most villages were nucleated although there were also dispersed villages, depending in part on the terrain. The life of the village was based on an economy of land and grain, and its rhythm was, and to a large extent still is, governed by the annual cycle of agricultural activities. The village also has an annual cycle of religious ceremonies and festivals, and the two cycles are related closely. With the development of transport and communication in the twentieth century, and particularly since the 1950s, the isolation of the village is becoming progressively reduced, although not at the same pace in all parts of the country or the region. The less isolated villages are now becoming urbanized, though in a highly uneven manner.

The social morphology of the village shows considerable variation in terms of size and internal differentiation. Irrigated villages in river valleys and close to the ancient and medieval centers of civilization tend to have an elaborate structure; 'dry' villages in the remote areas tend to be smaller and simpler in their structure. Only the latter may be described as communities of peasants. The former are often highly differentiated and stratified. They include noncultivating landowners, owner-cultivators, cultivating tenants, sharecroppers, and agricultural laborers, in addition to groups of families associated with various specialized crafts and services, usually on a hereditary or quasi-hereditary basis. Agrarian reform and the expansion of the market are changing some of this, although many of the older features remain.

Its relative isolation and self-sufficiency and its well-knit division of labor between agriculture, crafts, and

services gave the Indian village a certain unity. In the nineteenth century—and even later—it was often described as a 'little republic.' But it was not a republic of equals. In the larger irrigated villages the social boundaries between the superior, intermediate, and inferior sections of the community—landowners, tenants, agricultural laborers, priests, potters, and scavengers—were clearly maintained. If there was a certain unity in such villages, as indeed there was, it is best described as a 'vertical unity.' The unity of the village has been weakened in recent decades, but that does not mean that the inequalities have all disappeared.

While it is true that South Asia is a land of villages, one cannot ignore the presence of towns and cities in it. The subcontinent has cities going back to ancient and medieval times, and India had urban centers before most of Europe did. In the past, towns and cities grew around places of worship, the courts of kings and princes, and commercial and trading centers. Colonial rule led to the creation of a new kind of city, mainly along the coast, such as Calcutta, Madras, Colombo, Bombay, and Karachi. In the years since the end of colonial rule, industrialization has added to the population of existing cities and led to the creation of new ones. The urban middle class and the industrial working class are important factors in the economic and political life of contemporary India.

Returning briefly to the Indian village, a prominent place was occupied in it by caste. Indeed, caste may be viewed as the social basis of the division of labor in it, and the differentiation and stratification in the Indian village cannot be understood without taking caste into account.

7. *Caste*

Caste is an institution of great antiquity, and it has shown resilience in the face of the economic and political changes of the twentieth century. In its origin and fullest development, it is an institution of the Hindus, but near analogs of the Hindu caste system may be found among Sikhs, Muslims, and Christians on the subcontinent and among Buddhists in Sri Lanka; and it is of course very important in the predominantly Hindu population of Nepal. On the doctrinal plane, Hinduism, Buddhism, and Islam are different from each other; on the plane of everyday life at the local level, social divisions among Hindus, Muslims, and Buddhists act in broadly similar ways in most parts of South Asia.

The caste, the subcaste, and even the subsubcaste, to each of which the term *jati* may be applied, maintained its identity and continuity through the regulation of marriage. The most general rule was the rule of endogamy, and among Hindus this was supplemented by the rule of hypergamy. In the past, Hindu law prohibited intercaste marriage except under very specific conditions; the law has now changed, but the bias of custom is still in favor of endogamy among the Hindus, and to some extent also among the others.

Every caste or *jati* was in principle associated with a specific occupation, although agriculture was open to most if not all castes. The association between caste and occupation was stronger among the Hindus than among the others, but from the end of the nineteenth century, it began to weaken even among the former. A very important factor behind the weakening of the association has been the emergence, since the middle of the nineteenth century, of a large and increasing number of occupations that are 'caste-free' in the sense that they have grown outside the framework of the old division of labor based on caste.

The different castes are socially ranked; indeed, it is in the caste system that the principle of social ranking finds its most extreme and elaborate expression. In South Asia, castes are ranked among all religious groups, but the ranking is not as rigid or as exhaustive among the non-Hindus as it is among the Hindus. Caste ranking among the Hindus was expressed in the idiom of purity and pollution, but that idiom is now less pervasive than before, and new criteria of social ranking, having little to do with caste in the traditional sense, are becoming increasingly salient, particularly in the urban areas.

8. *Family, Kinship, and Marriage*

Given the great significance of the rules regulating marriage and in particular the rule of endogamy, it is not surprising that there is a close association between caste and kinship. Through repeated intermarriage from generation to generation, all the members of the local section of a caste or subcaste become related to each other by real or putative ties of kinship. This is generally true of all the major religious and linguistic divisions of South Asian society.

Much has been written about the prevalence of the joint family in the subcontinent and the changes it is undergoing as a result of urbanization, industrialization, and modernization. The Hindu undivided family had certain legal characteristics peculiar to itself, and some of these are still recognized by the law in India. One important feature of the family in the subcontinent, whether among Hindus, Sikhs, Muslims, or Christians, is that marriages are still to a large extent arranged by parents and older relatives, and the new couple generally lives in the parental home immediately after marriage even though it may set up a separate household after one or more children are born.

Reference has already been made to wide variations in the rules of descent, inheritance, and marriage across the many and diverse communities that inhabit the region. But through all these variations, kinship has a significant place in social life everywhere and

among all communities. Relatives are numerous and various, and the obligations of kinship are cherished and cannot be denied easily. The rules of endogamy act against the formation of real ties of kinship between members of different castes; but they do not prevent the formation of fictive ties of kinship between them. The idiom of kinship is used extensively in interpersonal relations across castes and communities in villages throughout the region.

Changes are undoubtedly taking place in the size and composition of the household, particularly in the rapidly growing cities. But even where nuclear households outnumber extended ones, the matrix of kinship within which they operate gives them a different character from their counterparts in the West. The boundaries between households are porous, and no household is secure against intrusion by cousins, uncles, aunts, nephews, nieces, and grandparents.

In discussing caste, kinship, and family it is impossible to ignore the subject of gender, although only a passing reference to it is possible here. Despite variations according to language, religion, and caste, and between patrilineal and matrilineal communities, women have since time immemorial occupied a subordinate position in South Asian societies. Polygamy was sanctioned by law among both Hindus and Muslims. In India it is now prohibited by law among Hindus but not among Muslims; it is permitted by law in Pakistan, and attempts to regulate polygamy by law have had little success among Muslims. The position of women is changing in all countries and among all communities in South Asia, although the change is highly uneven. Here, more important than the variations of language, religion and caste are those that arise from education, occupation, and income.

9. *New Institutions and Classes*

The first impediment against attempting a concise account of society and culture in South Asia arises from the problem of variation; the second impediment arises from the problem of change. Two important developments in the subcontinent since the middle of the nineteenth century have to be noted. These are, first, the growth of open and secular institutions such as schools, colleges, universities, libraries, laboratories, hospitals, law courts, and newspapers; and, second, the emergence of a new middle class based on a new educational system and a new occupational system. These two related developments have taken place everywhere, but most conspicuously in India where they are now a significant feature of the social and cultural landscape.

Open and secular institutions are sustained by norms and values that are very different from the norms and values by which the traditional institutions of South Asian societies have been sustained. The new institutions are often at odds with the old ones which exert various kinds of pressures on them and sometimes threaten their very survival. At the same time, and particularly in a country like India, they have grown to an extent that makes it difficult to dispense with them or to devise alternatives to them that are fully in harmony with the institutions of the past. As a result, the conflict of norms and values has become endemic in these societies.

The new middle class is both the creation and the creator of the open and secular institutions just referred to. The origins of the Indian middle class go back to the first half of the nineteenth century. Its first beginnings took place in the presidency capitals of Calcutta, Bombay, and Madras in each of which a new university was established in 1857. The graduates of these universities and of the others which came up in course of time became administrators, managers, lawyers, doctors, scientists, teachers, and clerks. The Indian middle class is now very large, at well over 100 million persons. Although it is still a relatively small proportion of the population of the country, it is among the largest and the most differentiated of the middle classes in the world.

The emergence of a new and increasingly dynamic middle class has not led to the effacement of the old identities and distinctions based on language, religion, and caste. Those still remain, although they are now coming to be organized in a different way. The middle class in turn is differentiated not only in terms of occupation, income, and education, but also in terms of language, religion, and caste. Caste and class reinforce each other in some respects and cut across each other in others. While there is some individual mobility, it does not lead to the equal distribution of all castes and communities among the different economic strata. The social mosaic is, thus, becoming in some sense even more complex than in the past.

See also: Area and International Studies: Development in South Asia; Caste; Colonization and Colonialism, History of; Communalism; South Asian Studies: Culture; South Asian Studies: Politics

Bibliography

Alavi H, Harriss J (eds.) 1989 *South Asia*. Macmillan, Basingstoke, UK

Béteille A 1991 *Society and Politics in India*. Athlone Press, London

Breman J, Kloos P, Saith A (eds.) 1997 *The Village in Asia Revisited*. Oxford University Press, Delhi, India

Dumont L 1980 *Homo Hierarchious*. University of Chicago Press, Chicago

Fuller C J (ed.) 1996 *Caste Today*. Oxford University Press, Delhi, India

Madan T N (ed.) 1976 *Muslim Communities of South Asia*. Vikas, New Delhi, India

Mandelbaum D G 1970 *Society in India*, 2 Vols. University of California Press, Berkeley, CA

Shah A M 1998 *The Family in India*. Orient Longman, New Delhi, India
Srinivas M N 1976 *The Remembered Village*. Oxford University Press, Delhi, India

A. Béteille

South Asian Studies: Culture

This article discusses studies of culture in South Asia, from colonial times to the present, focusing on the changing meanings of the term, the contribution of various disciplines to research in culture, and recent developments in cultural studies proper. In the South Asian context, culture can mean the cultural heritage of the nation–states of the region, cultural practices or ways of life, community identities, as well as culture in the aesthetic sense, including literature and the arts.

1. Colonial Beginnings

The first elaboration of the idea of an Indian culture occurs in the colonial era, through the combined efforts of European Indologists and Indian nationalists. The Indologists' construction of an Indian tradition, based largely on textual sources, was picked up by the nationalists, who divided the cultural realm into 'two domains, the material and the spiritual,' conceding the West's superiority in the former, while claiming sovereignty over the 'spiritual' domain, which bore the 'essential' marks of 'cultural identity' (Chatterjee 1993). This realm was to be out of bounds for colonial reformers, but at the same time, as Chatterjee has argued, the nationalists had their own project 'to fashion a ''modern'' national culture that is nevertheless not Western.'

At this stage, however, terms like tradition and civilization were more prevalent, and the meaning of culture was not fixed. In 1910, during the era of *swadeshi* (the movement for the promotion of native industry), Ananda K. Coomaraswamy, a cultural nationalist and art historian who included even present-day Sri Lanka in his map of Indian culture, describes culture as a 'capacity for immediate and instinctive discrimination between good and bad workmanship' and a 'view of life essentially balanced.' Here culture is understood as a historically developed human attribute, an assimilated refinement of taste that goes with a certain settled, rooted way of life ('restlessness is essentially uncultured'), recalling the ideas of William Morris (Coomaraswamy 1994). It is this meaning that seems to have prevailed when the nationalists translated culture as *sanskriti*, a word now used in most of the major Indian languages, while Rabindranath Tagore's suggestion of *krishti*, which is closer to the English word in its derivation from *krishi* or cultivation, never gained acceptance even in his native Bengal, in spite of his being the most influential cultural figure of his time.

In its substantive definition, the national culture included the classical heritage in the arts, traditions of education (the *guru-sishya parampara*), family structure (the joint family was celebrated as quintessentially Indian), and the deep-rooted customs and practices of village India. The twentieth century witnessed a widespread campaign for the reform, rediscovery, and revival of the classical arts. Thus a traditional temple dance around which a system of 'courtesanship' had developed was taken and 'purified' to create 'Bharatanatyam,' one of the currently widely practiced national dance forms. The textual tradition was revisited in the light of Orientalist scholarship and selectively annexed to the national cause, with the *Bhagavadgita* emerging as a sort of national text, embodying the spiritual distinction of Indian civilization. The search for a 'living tradition' that would supplement the classical heritage led to the celebration and appropriation of folk arts and village crafts (Guha-Thakurta 1992). Early nationalist constructions of India's cultural heritage tended to focus exclusively on Hindu achievements, ignoring the Islamic heritage, on the basis of an ideological negation of Muslim rule as the cause of decline of Hindu civilization.

Intellectuals like Rabindranath Tagore and the progressives in the national movement on the other hand tried to construct a more inclusive cultural history, locating themselves in the modern present and acknowledging the irreversible re-making of Indian culture and society by colonial intervention. Jawaharlal Nehru, who became India's first prime minister, was a key figure in this project but it was a sociologist, D. P. Mukherjee, who produced the first extended reflection on the idea of a 'modern Indian culture.' In the 1940s, when Mukherjee wrote his book, the cultural climate seems not to have been very hospitable to such an idea, given the widespread preoccupation with the revival and preservation of the disappearing cultural heritage of the nation. Mukherjee's project in the book is to reflect on the contemporary cultural situation in the soon-to-be-independent nation, to inventory the cultural, social, and intellectual heritage and its effectiveness in the present, as well as to produce a concept of the present moment as constituted by a diversity of forces, traditions, and processes. Uncharacteristically for his time, Mukherjee, a partisan of a socialist future for India, distances himself from any approach to culture that privileges nationalism, and insists on a sociological account of culture as 'the whole social process' (Mukherjee 1979). Rejecting the idea of culture as heritage, he locates modern culture in a society marked by 'the artifice of an unreal class-structure.' He attacks the idea of India as a land prone to the mystic and the

spiritual, and is, throughout, preoccupied with the most pressing issue of the time: that of the coexistence of Hindus and Muslims, and other minorities, within a modern nation–state.

2. Tradition and Modernity

The 1940s and 1950s are a crucial period for the emergence of culture as an object of study. In this period we see the triumph of social anthropology over sociology as the disciplinary home of culture studies, and in consequence, the decline of Mukherjee's idea of a modern Indian culture as defined by contemporary struggles between social forces, whether traditional, entrenched, emergent, or imposed. The overwhelming sense of the contemporary, which favored a strictly sociological approach, was soon replaced by a more historicist approach, as the dualism of tradition and modernity, by far the most influential paradigm in South Asian cultural studies, took hold.

The key figures in this shift were M. N. Srinivas and Milton Singer, an associate of Robert Redfield. In 1952, Srinivas published *Religion and Society Among the Coorgs of South India*, a work which, Singer asserts, demonstrated how the social anthropological method could be applied to a Great tradition. It was Redfield who proposed, in his project for the study of civilizations at Chicago (Srinivas had been in California), the fundamental distinction between Great and Little Traditions, roughly equivalent to 'higher' and 'lower' orders of cultural practice, the former more reflective, more systematic, and textually elaborated, while the latter is considered to be more spontaneous, fragmented, primitive. Until then the anthropological method had only been employed in the study of the so-called primitive societies, but Redfield was proposing a research project of global sweep to study the cultural heritage of humanity. Singer undertook the Indian portion of the study, concentrating on the south Indian city of Madras, later published under the title *When a Great Tradition Modernizes: An Anthropological Approach to Indian Civilization*, with a foreword by M. N. Srinivas. Singer 1972, argued that the Indian Great tradition (the use of the singular was to attract criticism) 'was culturally continuous with the Little Traditions to be found in the diverse regions, villages, castes and tribes' and that therefore, 'even the acceptance of "modernizing" and "progress" ideologies does not result in linear forms of social and cultural change but may result in the "traditionalizing" of apparently "modern" innovations' (Singer 1972). The most significant element in this formulation is the suggestion that a civilization with a tradition evolved over the *longue duree* acquired the strength to assimilate ideas and changes coming from outside, and to convert them into organic elements of its own make-up.

One of the most influential and controversial concepts in this new disciplinary thrust was Srinivas's 'Sanskritization' that, together with 'Westernization' served to explain social change in modern India. Sanskritization, a process by which the lower orders of traditional caste society aspire for a higher social status by adopting the customs and manners of the upper castes, was seen as one of the ways in which the continuity of Little and Great Traditions was maintained. Singer's concept of 'cultural performance' illustrates both the notion of cultural continuity between Great and Little Traditions and that of the absorption of modern influences. Singer defines a cultural performance in the broadest possible manner, including within its ambit plays, concerts, and lectures, and the cinema and radio, as well as 'prayers, ritual readings and recitations, rites and ceremonies, festivals, and all those things we usually classify under religion and ritual rather than with the cultural and artistic' (Singer 1972). In other words, the anthropologist's, the Indologist's, and the aesthetician's definitions of culture have here been fused into one, to constitute a seamless continuum of culture, object of the new social anthropology. The disruptions and displacements brought about by colonial modernity, which were foregrounded in Mukherjee's sociology of the present, are now located as challenges which the Great Tradition takes in its stride. A. K. Ramanujan, a pioneer of Indian folklore studies, also based in Chicago, rejected the Great and Little Tradition dichotomy and asserted that 'cultural traditions in India are indissolubly plural' and organized according to the principles of context-sensitivity and reflexivity (Ramanujan 1999).

3. Culture and Development

Around the 1960s, the dualism 'culture and development/modernization' begins to vie for space with its elder cousin Tradition/Modernity. Culture and development acquired wide currency at a more grassroots level, as development projects, undertaken by the nation–state and international agencies, began to transform the territory. While the Tradition/Modernity paradigm was prevalent among Indologists, anthropologists, and the nationalist intelligentsia, 'culture and development' rallies a wide range of social science disciplines including economics, political science, sociology, gender studies, as well as environmentalists and other grassroots activists and NGOs. 'Culture and development' the decolonizing, modernizing nation–state's version of Tradition/Modernity. Culture in this framework can be a hindrance to development, as in Amartya Sen's famous formulation about the 'missing women,' who are the victims of cultural constraints on women's access to food. Superstitions and prejudices nourished by entrenched cultural practices can come in the way of educating illiterate people in family planning, health, education,

hygiene, and other developmental concerns. Culture can also be a resource: traditional cultural forms can be usefully employed to spread developmental messages. Ecological debates have thrown up notions such as 'masculinist forestry' and turned to women as good agents in preserving the environment (Dietrich in Menon 1999). Thirdly, there is also the question of 'cultural survival,' cultural rights of minorities, tribes, and other groups, which come under threat from development's blind onward march and the imposition from above, of Western models of linear progress and development.

The most sustained critique of development and modernization from a point of view that affirmed the validity and continuity of Indian traditions was undertaken by Ashis Nandy. Recouping a Gandhian 'critical traditionalism,' Nandy attacked the deracinating effects of Western rationality, individualism, and other ideologies adopted by the Nehruvian state and the middle-class intelligentsia in its developmental campaign. Against the Western tendency to emphasize rupture as the precondition of change, Nandy, following Gandhi, emphasizes continuity. Against the rigid separation between male and female, between individual and individual, Nandy argues that fluid identities and ambiguous selves are more characteristically Indian. The imperative of cultural difference and diversity, of multiple rationalities, also gives rise to a critique of Western science and its hegemonic universalism, in the work of Nandy, Tariq Banuri, and others. The assertion of cultural specificity occurs in other fields of knowledge as well. Thus Sudhir Kakar, a psychoanalyst, has argued for a culture-sensitive psychoanalytic practice and has elaborated his own analytic picture of the Hindu psyche. There is a substantial body of psychoanalytic readings of South Asian culture, including the work of Girindrashekhar Bose, the first Indian analyst, Philip Spratt, Erik Erikson, Gananath Obeyesekere, and Nandy (see Vaidyanathan and Kripal 1999). There is sometimes a tendency to culturalism reduction in these writings, producing a domesticated psychoanalysis from which the fundamental alienation that psychoanalysis posits at the threshold of human subjectivity is wished away. More recently, Lacanian psychoanalysis has found favor among film studies scholars.

Nandy's rise to eminence as an ideologue of decolonization roughly coincides with the emergence of 'postcolonial studies,' galvanized by the publication of Edward Said's *Orientalism* in 1978. In India, Saidian postcolonial studies as well as studies of development undertaken in the social sciences were influenced by Nandy's discourse. Postcolonial theorists like Gayatri Chakravorty Spivak and Homi Bhabha also shared this space of the critique of colonial reason. Apart from this, another important development was the emergence of *Subaltern Studies*, a series of volumes to which mainly historians contributed, in which colonial and nationalist historiography was critiqued and a 'subaltern history,' of ordinary people, of tribal, and lower caste groups, during colonial rule and after, was undertaken. Culture plays a very important role in this project, especially when there is an emphasis on the spontaneity of 'peasant insurgency' and tribal uprising, and the question of the subaltern mentality. Postcolonial studies, the *Subaltern Studies* project, and the critique of development and modernization by Nandy and others, together constitute the legacy of cultural studies in the 1970s and 1980s.

4. *Globalization and Local Cultures*

In the 1990s, the terms shift again, as globalization arrives on the scene. In 'globalization and local cultures' we have the third and most recent version of the Tradition/Modernity paradigm, where again the emphasis is on questions of cultural survival in the face of globalization, the resilience of local cultures, and their ability to 'consume modernity' on their own terms. Arjun Appadurai has offered a comprehensive theory of globalization and the emergence of what he terms 'public culture.' Appadurai treats culture as 'the dimension of difference,' of identity based on difference, emerging after the rupture of globalization, in a world where he considers nation–states to be on their last legs. Global relations and movements, or 'flows,' have a more decisive bearing on human lives today than national identification. Others are less sanguine about the effects of globalization, and more skeptical about the nation–state's imminent demise. Appadurai is confident about the ability of societies to assimilate modernity, which is what globalization transports: although we have traveled far, we are still within Singer's paradigm where traditional societies respond to and assimilate modernity in their own ways.

The three variants of a paradigm that have been examined so far all share one thing in common: they approach the question of Indian culture on an international plane. In each case, one term in the opposition refers to a force, a process—modernity, development/modernization, globalization—which is of extraneous provenance, while the other term indicates the culture that is at the receiving end. None of them was developed with specific reference to India, which is only one of the sites to which they are applied. The concept of culture employed in all three variants is also predominantly anthropologically defined.

5. *National Culture*

Within India, other paradigms of cultural analysis have devoted themselves to reading and analyzing the stuff of national culture. Some of them emphasize a distinctive native culture with its own rationality, its sense of self, and strategies of survival. There are also attempts to forge an indigenous conceptual series for

the study of Indian culture, to reconnect with indigenous intellectual traditions after the 'amnesia' of colonialism. Beginning in the mid-1980s and throughout the 1990s, as questions of globalization assumed importance, the domestic politics of communalism, which resulted in popular mobilization and widespread violence against minorities, led to a fresh attempt to reexamine the past. As the Hindu nationalists put forth versions of history that supported their political activities, liberal and left intellectuals, particular historians, revisited the past to reassert the plurality and ineluctable syncretism and hybridity of India's cultural heritage (Thapar 2000). The question of multiculturalism, and of cultural rights, also became more urgent as cultural and ethnic conflict was seen as threatening the secular-democratic fabric of the nation–state.

Social anthropologists have noted the difficulty in the Indian instance, of separating a sphere of culture from those of religion, community, caste, etc. Caste has been a central category for understanding Indian society, from Dumont's notion of *homo hierarchicus* to more recent studies of caste conflict in a contemporary setting, where caste has more to do with cultural and political identity, than with social position. This shift to caste as identity brings it into the realm of culture, and various studies have asserted, against the idea of a universal dynamic of Sanskritization, that caste groups enjoy a measure of cultural autonomy and strive to maintain their cultural identity through the formation of networks across regional and language barriers. *Dalits'* (literally, 'the oppressed,' a term now used for the lowest castes in the caste hierarchy, especially the 'untouchables') struggles for social justice have included literary movements and other forms of cultural expression, as well as attempts, under the broad rubric of folklore studies, to record and study the traditional cultural forms prevalent among these groups.

Feminist scholarship has engaged with questions of culture at many levels (see Menon 1999, Thapan 1997). For nationalists woman was the guarantor of cultural identity and continuity. In the confrontations between the nationalists and the colonial government, and after independence, in the confrontations that arose between religious groups and the nation–state, women became the object of reformist attention and patriarchal protectionism. Two recent debates arising out of events in the 1980s that have had a lasting impact on the character of the national polity have posed a challenge to feminist scholarship, raising questions of the competing claims of women and minority cultural rights and the law, and of female agency in traditional practices that are offensive to a modern, secular outlook.

These relate to an incident of *sati* (self-immolation by a widow on her husband's funeral pyre, a practice that was thought to have died out) in a village of Rajasthan, and the case of Shah Bano, a divorced Muslim woman who filed a suit for maintenance from her former husband, leading to a debate about the competing claims of the state and the community's own institutions to handle such disputes. Using these incidents from the 1980s and the legislative moves that followed each, the anthropologist Veena Das has investigated how 'a web of creative or destructive tensions in the matter of cultural rights' determines relations between communities, the state, and the individual (in Menon 1999). Feminist scholarship has shown how communities deploy culture as a means to assert the supremacy of community rights over women's rights as citizens in a democratic polity. At the same time, on the Uniform Civil Code issue, which has been intensely debated in recent years, feminists have become more sensitive to community laws which, in certain instances, may be more beneficial to women, in opposition to the Hindu nationalist deployment of the discourse of citizenship to press for a civil code that will prevail over all other culturally specific laws (see Menon 1999).

The realm of Indian politics is also continuous with that of contemporary popular culture, in particular the culture of popular cinema. South Indian cinema culture has been the breeding ground of some of the most powerful political leaders to emerge in the states of Tamil Nadu, Andhra Pradesh, and Karnataka. Sociologists have studied the ways in which the unique film cultures of these regions have functioned as a platform for electoral politics. The fan clubs of these popular film stars are today an important feature of everyday political and social life, in some cases emerging as militant frontrunners in campaigns for a linguistic national identity.

Meanwhile, the impact of the Birmingham school of cultural studies and a reconstructed anthropology's search for new objects of research have spurred projects for cultural studies in India, notably studies of communities of reception for the popular film, popular cultures of photography, music, etc. Film has come to be seen as the emblematic cultural institution of modern India, and in recent years film studies have acquired legitimacy as a discipline, with a multidisciplinary resource base. The importance of globalization notwithstanding, there is a distinct national sphere of cultural studies, which draws upon many of the ideas and paradigms mentioned above, but locates its concerns within the national space. One instance is the study of aesthetic modernism. Postcolonial societies attract development and modernization approaches, where cultural issues are subsumed under sociological, economic, and political questions. Studies of modernism, on the other hand, are concerned with the emergence of artistic practices and ideologies that participate in the international modernist movement while speaking from within (though not only about) their own national space. Modern literature, cinema, theatre, architecture, painting, and other fine arts have been the object of this critical appreciation

(Kapur 2000), with the *Journal of Arts and Ideas* serving as an important forum for the theorization of an Indian modernism.

In the contemporary field of cultural studies, the traditional disciplines like anthropology are joined by new discourses like postcolonial studies, feminism, psychoanalysis, semiotics, post-structuralism, and postmodernism (see Niranjana et al. 1993, Spivak 1988). Ethnographies of cultural communities, including traditional ones as well as newer consumerist types of community, feminist critiques of gendered cultural forms and expressions, and of patriarchal ideologies, film and television studies, and studies of other modern, technology-dependent cultural forms, caste politics, popular religion, histories of modern and popular art, and political cultures are some of the types of cultural study being currently undertaken in India, by scholars belonging to a range of social science and humanities disciplines (see Manuel 1993, Niranjana et al. 1993, Thapan 1997, Vasudevan 2000). The concept of culture has undergone a definite shift of emphasis from ancient heritage and primitive ways of life to the more unstable and complex practices and processes of contemporary existence. The past coexists with the present, as it does in any social formation, but not in a historicist time-space. Cultural studies increasingly regard the present moment as a synchronic dimension, where all constituent elements, whether ancient or recent, foreign or indigenous, constitute a symbolic network with its own unique properties, making it radically contemporary (Dhareshwar 1995). Forging the tools for analyzing this complex cultural space is the task of the future.

See also: Area and International Studies: Development in Southeast Asia; Buddhism; Buddhism and Gender; South Asia: Sociocultural Aspects

Bibliography

Appadurai A 1997 *Modernity at Large: Cultural Dimensions of Globalization*. Oxford University Press, Delhi, India

Chatterjee P 1993 *The Nation and its Fragments: Colonial and Post-colonial Histories*. Oxford University Press, Delhi, India

Coomaraswamy A K 1994 *Art and Swadeshi*. Munshiram Manoharlal, New Delhi, India

Dhareshwar V 1995 Our time: History, sovereignty and politics. *Economic and Political Weekly* **30**(6)

Guha-Thakurta T 1992 *The Making of a New 'Indian' Art: Artists, Aesthetics and Nationalism in Bengal, c. 1850–1920*. Cambridge University Press, Cambridge, UK

Kapur G 2000 *When was Modernism: Essays on Contemporary Cultural Practice in India*. Tulika, New Delhi, India

Manuel P 1993 *Cassette Culture: Popular Music and Technology in North India*. University of Chicago Press, Chicago

Menon N (ed.) 1999 *Gender and Politics in India*. Oxford University Press, Delhi, India

Mukherjee D P 1979 [1948] *Modern Indian Culture*, 2nd edn. [Alternative title *Sociology of Indian Culture*] Rawat Publications, Jaipur, India

Nandy A 1998 *Exiled at Home*. Oxford University Press, Delhi, India

Niranjana T, Sudhir P, Dhareshwar V (eds.) 1993 *Interrogating Modernity: Culture and Colonialism in India*. Seagull Books, Calcutta, India

Obeyesekere G 1981 *Medusa's Hair: An Essay on Personal Symbols and Religious Experience*. University of Chicago Press, Chicago

Ramanujan A K 1999 *The Collected Essays of A. K. Ramanujan*. Oxford University Press, Delhi, India

Said E 1978 *Orientalism*. Routledge, London

Singer M 1972 *When a Great Tradition Modernizes: An Anthropological Approach to Indian Civilization*. Praeger, New York

Spivak G C 1988 *In Other Worlds: Essays in Cultural Politics*. Routledge, New York

Srinivas M N 1952 *Religion and Society Among the Coorgs of South India*. Asia Publishing House, Bombay, India

Srinivas M N 1966 *Social Change in Modern India*. University of California Press, Berkeley, CA

Thapan M 1997 *Embodiment: Essays on Gender and Identity*. Oxford University Press, Delhi, India

Thapar R 2000 *History and Beyond*. Oxford University Press, Delhi, India

Vaidyanathan T G, Kripal J J (eds.) 1999 *Vishnu on Freud's Desk: A Reader in Psychoanalysis and Hinduism*. Oxford University Press, Delhi, India

Vasudevan R (ed.) 2000 *Making Meaning in Indian Cinema*. Oxford University Press, Delhi, India

M. M. Prasad

South Asian Studies: Economics

South Asia is the sister that was all dressed up, but did not make it to the ball. At the time of independence at the close of the first half of the twentieth century, India, Pakistan and Sri Lanka (Ceylon at the time), the larger countries of the region, all seemed poised for successful development. Sri Lanka's per capita income was much higher than Korea's in 1950, and India and Pakistan were at income levels comparable to those of Korea. These South Asian countries had inherited a reasonable infrastructure, both physical and administrative, and, freed of colonial shackles, all truly seemed ready for the 'tryst with destiny' that Jawaharlal Nehru articulated for India.

Fifty years later, South Asia is far behind East and Southeast Asia in standards of living. While the region has made progress, the East Asian 'miracle' puts it to shame. The damning comparison is not just with the most successful developing countries;

> South Asia is fast emerging as the poorest, the most illiterate, the most malnourished the least gender-sensitive, indeed, the most deprived region in the world (Haq 1997, p.2).

What went wrong? This piece examines the economic development experience of South Asia and the lost

opportunities for the region's people. It touches on long-standing issues of growth vs. development, state vs. market, and agriculture vs. industry, as well as the more recent themes of political economy, decentralization, liberalization and reform. It discusses the role of economic ideas and of political ideologies and compulsions in shaping the course of economic policies and region's economies. The experience of the region's economies stands as a mirror of those ideas and ideologies: South Asia has been a laboratory for economic research on development.

1. India

We begin with objectives. Newly independent nations unambiguously wanted to catch up with the West that had ruled over them, to develop, to modernize, to industrialize—three somewhat different goals. Industrialization was the clear winner out of these. In India, the government was to occupy the 'commanding heights' of the economy, and engineer economic growth led by import-substituting heavy industrialization. The Soviet Union seemed an attractive model at the time, and formal planning exercises were launched in India and in Pakistan soon after independence. A Marxist view of an imperialist world economy, Fabian socialism, a colonial-trained elite, and an economy shaped by wartime controls, all met on common ground to define how India proceeded. The initial intellectual framework for Indian planning came from non-economists, but, at the time, the general approach of a vast, centralized resource optimization exercise was a common one in global thinking on economic development. Dissenting views to this model, accepted by most development researchers at the time, were rare, though not absent.

Through the nineteen-fifties, this model was reasonably successful. There were no purges, and no mass starvation (as economist Amartya Sen has suggested, these facts are connected by the constraints of a democratic polity). India did not break apart. And the economy grew steadily, at a rate much faster than in colonial times, 3.5 percent a year (later caricatured as 'the Hindu rate of growth'). Agriculture was not neglected, if not particularly favored. Land reform was legislated, but not really implemented, though it seemed to be called for on grounds of efficiency as well as equity. Agricultural pricing and food distribution also received attention from academics and policymakers, with the goals of increasing marketable surpluses and providing adequate food to the poor. Innumerable studies were performed on the productivity of Indian agriculture, using data collected through extensive surveys. What really wrought lasting change in agriculture, however, was the introduction of high-yielding varieties of wheat and rice. This 'green revolution' first took place in Punjab, where land holdings were relatively equal, and where the state government had created an infrastructure of market and rural roads. Central government irrigation projects also played a key role.

The Indian planning exercise was a peculiar one, in that the mechanisms for implementation were weak: this was not Soviet-style command planning. At the same time, the functioning of markets as resource allocation mechanisms was incredibly constrained, almost stifled: prices (for internal and external trade) as well as investment and production decisions were bureaucratically controlled and greatly distorted. Theoretically, these distortions were meant to promote equity (e.g., rationing basic foodstuffs) and efficiency (where social costs and benefits supposedly diverged from private ones). In practice, they did a little for the former, and impeded the latter. The government's 'license-permit-quota *raj*' epitomized the 'rent-seeking society,' where incentives for activities aimed at capturing a share of the current economic pie dominated those that would increase the pie itself. In fact, the experience of India seemed to inspire the coinage of the term 'rent-seeking society' by an economist, Anne Krueger (1974).

Compromise and consensus were viewed as necessary for India's continued unity, and this worsened the inefficiencies of the system. The nostalgic Gandhian ideal of a village-based *Ram Rajya* that rejected industrialization was accommodated through subsidies and protections for handlooms and small-scale industries in general: 'big business' could not encroach on their domain. A related populism that grew in the 1960s, as political competition greatly intensified, led to further restrictive polices. War, drought, and commodity price shocks worsened the performance of an economy hobbled by policies that led to inefficient capital use, despite increasing rates of savings and investment. Academic economists—a scattered few through the 1970s (Jagdish Bhagwati, Padma Desai, and T. N. Srinivasan being among the most prominent), then many more in the 1980s—pressed for economic reform, but ideology and vested interests in the economy and polity (as elegantly analyzed by Bardhan, 1984) engendered delay. A more pragmatic approach to economic policy-making made a halting entrance in the 1980s, and Indian economic growth picked up, to 5 percent along with a documented increase in productivity growth (Ahluwalia 1985, 1991, Goldar 1986). However, this took place in a manner that made India more vulnerable to external shocks. The balance of payments crisis of 1991 therefore ushered in a new wave of reform. The collapse of the Soviet Union and the success of East Asia demolished ideological barriers, and India considerably liberalized government controls on domestic and trade-related economic activities. The result has been a further improvement in India's economic growth rate.

Despite its relative economic failure, India has succeeded in building resilient political structures

during the five decades after independence. Its constitution was drafted expeditiously, and has stood the test of time. Fiscal and monetary institutions such as government budget processes and the central bank have remained transparent, accountable, and reasonably effective. In particular, the central bank has excelled in monetary management within its given constraints. Quasi-government bodies such as the Finance Commission that governs central-state fiscal transfers have developed expertise and reputation, and a slew of new regulatory bodies, such as the Securities and Exchange Board of India, promise to improve the quality of economic governance. Gandhian-inspired decentralization, after many half-hearted experiments that met with limited success, has taken root in a constitutional amendment strengthening local governments. If properly implemented, this decentralization may at last be the vehicle for effectively delivering health, education and social services to the bulk of India's poor, something that the planning process ultimately failed to do. The institutional developments in India again parallel broader research on the economic role of governance institutions (e.g., Williamson 1996): ideas help shape policy.

2. *Pakistan*

If consensus to the point of gridlock has often seemed to characterize India's development path, the word that describes Pakistan might be 'contradictions.' In many respects, policies followed in Pakistan were quite similar to those in India: import-substituting industrialization, controlled markets and prices, relative neglect of broad-based agricultural development and of health and education for the rural masses, and half-hearted attempts at land reform. This was part of the same economic development research consensus that shaped India's post-independence policies. At the same time, Pakistan diverged from India in significant ways. On the economic front, Pakistan was earlier in allowing the private sector freedom for industrial growth. It was able to squeeze resources for industrialization from the agricultural sector more effectively than India. It relied much more heavily on foreign aid and private remittances from overseas workers, and its domestic savings rate has been much lower than India's. Overall, as researchers would agree, the result was effectively a more efficient use of capital, and, up to the last decade or so, a higher growth rate than in India.

The starkest contrast with India, according to many analyses of Pakistan's economic performance, has been in the country's governance. Whereas India has evolved a robust, if still flawed, democratic system, Pakistan has essentially been an authoritarian state, where interludes of populism with the trappings of democracy can not disguise the essential nature of the system as a contest between three overlapping and related governing elites: the bureaucracy, the military, and the large landowners, with a small number of wealthy industrialist families acting as a background influence. Furthermore, ethnic divisions have sharpened this contest. While this sounds similar to Pranab Bardhan's (1984) analysis of India, the degrees of inequality and rent capture, and lack of political checks and balances in Pakistan have differentiated the two countries. When Pakistan was created in two parts, East and West, the East was more populous, but the West contained the bulk of the elite. Differences in language, culture and economic structure all stood against religion as the only common point. This cleavage heavily influenced the nation's political and economic development. Historians of the political economy of Pakistan, such as Omar Noman (1990) argue that real democracy was impossible for the Western elites to accept. Policies that squeezed the East's agriculture to benefit the West's industrial development exacerbated the divide. From this perspective, the country's split in 1971 was almost inevitable.

After the creation of Bangladesh from East Pakistan, the remainder of Pakistan faced further internal conflicts, reflecting a mix of class and ethnic divisions. The loss of the eastern wing sharpened the focus of these western conflicts. Society was further destabilized by the Soviet invasion of Afghanistan in 1979. This made Pakistan strategically more important, and led to greater foreign assistance, economic and military. This helped the economy in terms of overall growth, but exacerbated problems of governance, with a huge influx of arms into the region. When the Soviets withdrew, the problems for Pakistan remained, and persist. Around the same time, the decade-long oil boom in Persian Gulf wound down, curtailing the remittances from Pakistanis working there, and having a significant negative impact on the economy.

If we have described Pakistan's development path by its relation to that of India, this only reflects the consciousness of the nation. While those who govern India and determine its policies certainly are aware of Pakistan, India's larger neighbor to the north, China, is often more of a mental presence. The most popular view, typical for such a large nation, has probably been just one of insularity: 'India is different.' For Pakistan, however, India is the large neighbor, that helped to dismember it in 1971, and that, even after three wars and numerous skirmishes, maintains a firm grip on Muslim-majority Kashmir. Pakistan spends a much larger proportion of its GNP on defense than does India—as much as a third of its budget—and always must be aware of its giant neighbor.

However, Pakistan does not necessarily follow India in policy-making. Import substitution, bureaucratic control, and heavy industrialization at the expense of agriculture, all were chosen in Pakistan and in India because of their common history and a common understanding of the history and the possibilities of

development. Pakistan embarked on a path of economic reform shortly before India, in 1990, because of external pressures resulting from economic crisis—partly caused by external shocks, but ultimately reflecting the failure of the previous direction of policy. However, Pakistan may have a harder road ahead than India: its lead in per capita income has disappeared over the last few years, and it lags dramatically in human development indicators such as literacy and school enrollment, particularly for girls. Its population, too, is growing much more rapidly. Large inequalities and poor governance structures exacerbate the tensions inherent in its situation. Economists and political scientists would agree that this is not a promising situation for further economic development and growth.

3. *Bangladesh*

If Pakistan's prospects have dimmed recently, those of Bangladesh have grown brighter. When the nation was born, it had suffered from close to 25 years of economic stagnation as East Pakistan, and a devastating war of independence. Some observers dismissed it as basket case. It underwent the same cycles of military rule and authoritarian populism as Pakistan did after 1971. It also struggled with government-led development via Five Year Plans. While Bangladesh still trails India and Pakistan in measures of income, poverty and human development, it has shown progress in recent years, and has begun its demographic transition to a lower population growth rate, at income levels much lower than had been predicted by demographic research at that time.

Some researchers suggest that what distinguishes Bangladesh from its larger South Asian cousins is its relative societal homogeneity. Therefore, after initial conflicts among the elite, the nation seems to have settled into an approach that is delivering some broader-based human development. Agricultural production has improved substantially since independence, and the countryside has proved to be an impressive laboratory for nongovernmental organizations such as the Grameen Bank (rural microcredit) and the Bangladesh Rural Advancement Committee (community development for the poorest). This approach, drawing on the strengths of the nation's civil society, contrasts favorably with more centralized, government-led approaches in India and Pakistan. Of course each of these countries does have its own success stories, and building on them in the future will be important for sustained development. Not surprisingly, theoretical and empirical analyses of rural microcredit and targeted community development in South Asia have been very common (as documented, e.g., in the World Bank's annual World Development Reports).

Bangladesh also shares institutional problems with India and Pakistan: lack of systems that support widespread entrepreneurial attitudes, grossly inefficient formal financial sectors, and outmoded systems of taxation and tax administration. In each of these Bangladesh lags behind the other two nations, and its economic reforms in the last few years, where these institutional weaknesses are critical, make the reform process seem like pushing on a string, according to economists such as M. G. Quibria (1997) One thing that can potentially help Bangladesh is India's external liberalization—international trade within the region has been very low in general, despite the absence of the hostility that has limited trade between India and Pakistan. Bangladesh may also be in a position to move into export niches vacated by Southeast and East Asian countries moving up to more sophisticated manufacture. This possibility, and the importance of international trade in general, has been widely recognized by economists. In agriculture, also, India may matter: Bangladesh is a vast delta for rivers that flow from India. Controlling and using that water requires international cooperation that (as documented and analyzed by Crow and Singh, 2000) has only recently begun to develop.

4. *The Himalayan Nations*

International trade and water resource development also loom large for India's Himalayan neighbors, Nepal and Bhutan. The economic development of these small, land-locked nations has been, and will be, heavily influenced by India's own economic success, the openness of its economic regime, and its strategic concerns with respect to China. Nepal has had its own economic experiments with five year plans, and its own struggles with developing democracy. Its infrastructure is poor, but, in a more open regional economy, the potential of its natural resources is high.

5. *Sri Lanka*

Sri Lanka and the tiny Maldives are the island nations of South Asia. They are outliers in other respects too, with higher incomes and higher human development indices than their continental cousins. Yet Sri Lanka has undergone some of the same problems as the others. At independence, its economy was somewhat depleted by the costs imposed by World War II, but it was still relatively rich, with a strong base in exported cash crops. Post-independence policies that nurtured the population in terms of health, education and access to food gave the country a persistent lead over the continental region in broader human welfare, and all would agree with the position emphasized by Amartya Sen (1981) that these are significant achievements. Yet growth has been slow, and the lack of adequate economic growth has cost Sri Lanka heavily, by exacerbating ethnic conflict that has, proportionate to the nation's size, been the worst in the region.

The reasons for Sri Lanka's relative failure, as in the case of India and Pakistan, involve some themes that are familiar in analyses of the development of the region: populism, government policies overly restrictive of private entrepreneurship, direct government control through nationalization, failure to attract enough foreign investment bringing the right mix of capital and technological know-how, and severe distortions in the foreign trade regime. Both in the failure of its economic policies and of management of ethnic tensions, Sri Lanka's most striking comparison and contrast is with Malaysia, rather than the larger nations of South Asia. The economic policy mistakes of the 1950s and 1960s fed into austerity measures in the 1970s and the explosion of conflict between Tamils and Sinhalese in the1980s. This meant that, while reforms began earlier than in India or Pakistan, they have not been undertaken in the best conditions. Sri Lanka has managed to grow somewhat more rapidly recently, and to reduce the role of primary products by establishing labor-intensive manufacturing exports as a significant part of the economy, but the story remains one of lost time and lost opportunity.

6. *Prospects*

If all the major nations of South Asia, despite numerous differences in initial conditions and trajectories, made similar economic policy mistakes over the last five decades, can we lay the blame on common causes? The discipline of economics and its practitioners share some responsibility. The understanding of the workings of markets and entrepreneurship, and the roles played by information, incentives and innovation, was certainly incomplete when these countries embarked on their attempts at guided development. Economists who were skeptical of the abilities of government to be a guide or director of the economy were in a minority initially. Yet by the time economists began to be more forceful in their criticisms, policies had created entrenched interests, and reform has not come easily.

This is not to argue that the state has no role. Currently, economists agree that the issue is not simply 'state vs. market,' or the degree of state intervention along some one-dimensional conceptual continuum. There are many dimensions. The state does some things better than others. It may do nothing very well, but it may still be better than the market in some spheres of the economy. And, even if it does not 'govern the market' (to use Robert Wade's (1990) characterization of East Asia's experience), it certainly 'enables the market,' through the maintenance of law and order, and through a judicial system that enforces contracts. The state also provides safety nets for the unfortunate, physical infrastructure, and, sometimes, social-order-preserving redistribution, all through a system of taxation and expenditures. In all this, the state is itself the agent of interest groups and constituencies. Part of the lesson of South Asia's experiences of economic development is therefore a practical one: what does and does not work in economic policy, and the proper roles of state and market.

Returning to common causes for South Asia's relative failure, ideology, society and history all have their place—though not the catchall scapegoat of culture. The ideology of Marxism and its Soviet incarnation cast a long shadow, especially over India. Societal divisions have certainly made the tasks of governance and development enormously more difficult throughout the region, right from the partition that gave birth to Pakistan. Vertical cleavages have been as pernicious as horizontal ones, and the neglect of human development in South Asia may ultimately, perhaps, be blamed on the rule of an elite very different from any that exists in the social structures of Southeast and East Asia. Finally, history always matters, whether one begins with partition, or colonization, or even further back. While we have presented what might be termed the consensus of research on particular aspects of South Asia's experience, it is difficult to argue that overall agreement exists on the relative importance of the different factors that have shaped the region's development in the past half-century.

While history is always important, though, it sometimes becomes a fondly remembered golden past. For nationalistic politicians, 'In Search of Lost Time' has that Proustian relevance. In this case, however, the lost time is really that of midnight's children, the generations of South Asians who failed to realize the full promise of independence. Luckily, though, the ball does not end at midnight, and midnight's grandchildren may do much better than the previous generation. Perhaps economists have learned enough in the past few decades to help them.

Bibliography

Ahluwalia I J 1985 *Industrial Growth in India*. Oxford University Press, New Delhi, India

Ahluwalia I J 1991 *Productivity and Growth in Indian Manufacturing*. Oxford University Press, New Delhi, India

Bardhan P 1984 *The Political Economy of Development in India*. Blackwell, Oxford, UK

Bhagwati J 1993 *India in Transition: Freeing the Economy*. Oxford University Press, New York

Bhagwati J N, Desai P 1970 *India: Planning For Industrialization: Industrialization and Trade Policies Since 1951*. Oxford University Press, New York

Bhagwati J, Srinivasan T N 1975 *Foreign Trade Regimes and Economic Development: India*. National Bureau of Economic Research, New York

Crow B, Singh N 2000 Impediments and innovation in international rivers: The waters of South Asia. *World Development* **28**: 1907–25.

Drèze J, Sen A 1995 *India: Economic Development and Social Opportunity*. Oxford University Press, Delhi, India

Goldar B N 1986 *Productivity Growth in Indian Industry*. Allied Publishers, New Delhi, India

Haq M 1997 *Human Development in South Asia 1997*. Oxford University Press, Karachi, Pakistan
Hossain M, Islam I, Kibria R 1999 *South Asian Economic Development: Transformation, Opportunities and Challenges*. Routledge, London
Husain I 1999 *Pakistan: The Economy of an Elitist State*. Oxford University Press, Karachi, Pakistan
Karunatilake H N S 1987 *The Economy of Sri Lanka*, 1st edn. Centre for Demographic and Socio-economic Studies, Colombo, Sri Lanka
Khan A R, Hossain M 1989 *The Strategy of Development in Bangladesh*. Macmillan, Basingstoke, UK
Krueger A O 1974 The political economy of the rent seeking society. *American Economic Review* **64**: 291–303
Noman O 1990 *Pakistan: A Political and Economic History Since 1947*. Kegan Paul, London
Quibria M G 1997 *The Bangladesh Economy in Transition*. Oxford University Press, Delhi, India
Sen A K 1981 *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford University Press, New York
Wade R 1990 *Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization*. Princeton University Press, Princeton, NJ
Williamson O E 1996 *The Mechanisms of Governance*. Oxford University Press, New York

N. Singh

South Asian Studies: Environment

Cultural traditions of ecological prudence evolve in sedentary societies inhabiting stable environments, when populations are close to resource saturation and technology is stagnant (Gadgil and Guha 1992). Accordingly, the emergence, maintenance, erosion, and restoration of ecologically prudent behavior in the south Asian region is examined here. Following sections sketch colonization of the subcontinent, evolution of agricultural society and resource partitioning amongst occupation-based caste groups, and environmental erosion triggered by the colonial conquest and the resultant industrial society. The concluding section discusses the implications of ongoing globalization and emergent environmental conflicts. Differing research perspectives on important issues, both within and beyond the subcontinent, are also examined.

1. *Colonization*

1.1 *Waves of Immigration*

The first major migration into south Asia was that of Austric-speaking, i.e., Australasian, peoples from the northeast about 65,000 years ago (Mountain et al. 1995). Subsequent numerous waves of immigrations were propelled by technological innovations outside the subcontinent, as underlined by Western-trained subcontinental researchers (Gadgil and Guha 1992). The farming of wheat, barley, cattle, and pigs in the Middle East around 10,000 years ago propelled several immigrations into south Asia of Dravidian-speakers approximately 6,000 years ago. Similarly, the farming of rice, buffalo, etc. in southeast Asia 8,000 years ago drove immigrations of Sino-Tibetan-speakers into the south Asian subcontinent about 6,000 years ago. Indo-Europeans dominated the latest immigrations around 4,000 years ago, following domestication of the horse in central Asia 6,000 years ago. Lastly, Europeans, pushed by resource scarcity following the 'Little Ice Age' and armed with superior ships and cannons, invaded India in the sixteenth century to control the spice trade.

These manifold invasions have shaped, genetically and culturally, the most diverse society of the world, as Western researchers have revealed (Mountain et al. 1995). This persistence of over 50,000 tribe-like endogamous groups, characterized by hunting–gathering/shifting cultivation cultures, in a complex agrarian and later industrial society, is very unusual. This situation has resulted from the peculiar regional tradition of subjugation and isolation, in contrast with the worldwide practice of elimination or assimilation of subordinated communities by dominant groups. This diversity is accompanied by extreme specialization of occupation, as subcontinental researchers emphasize (Gadgil and Guha 1992). The diversity of environmental regimes throughout the subcontinent, ranging from warm coastal mudflats to snow-clad mountain peaks, also nurtured the cultural diversity.

1.2 *Conservation Ethic*

In the relatively predictable tropical and subtropical environments, the primitive hunter–gatherer society evolved largely into sedentary, territorial tribes with populations close to carrying capacity. This promoted the evolution of cultural practices of ecological prudence, rationalized within a religious framework. Such traditions could have changed a little when hunters–gatherers advanced to the level of shifting cultivators, as is evident from the age-old practice of maintaining sacred groves (tree clumps protected against harvests by local people because of their religious significance), still followed in areas of shifting cultivation tracts such as the Himalayas. Harvests from such relict forest patches are prohibited or rigorously regulated by the local community. Even these communities showed some further occupational specialization by age and sex; such as shamans and traders of honey, shells, feathers, etc.

2. *Agricultural Civilization*

2.1 *Agricultural Evolution*

A sedentary agricultural civilization evolved gradually in favorable situations, employing paddy cultivation

in the river valleys and maintaining large cattle herds. The resultant surplus promoted the evolution of a predatory warrior social group, specializing in physical protection of the surplus and its usurpation as tax, as well as the expansion of territories. This accompanied the so-called Aryan invasion from west and central Asia, with horse chariots and iron axes easily clearing forest tracts for cultivation. Consequent deforestation in the Gangetic plains began around 1100 BC, proceeding to the southeastern end of the subcontinent by 600 BC (Gadgil and Guha 1992).

This new agricultural–pastoral society organized into sedentary, largely self-sufficient village communities seemingly nurtured many tribal practices of ecological prudence. The generation of a large surplus by agriculturists and herdsmen promoted the expansion of trade, including that in luxury goods such as spices, fine cloth and precious metal, which were exported to other continents. This strengthened the caste of traders, while tribal shamans evolved into the priestly caste claiming a monopoly over codified knowledge. The peasants and artisans constituted the lowest groups.

2.2 *Religious Evolution*

The hierarchical religion then practiced called for extensive sacrifice of animals, especially oxen. Such sacrificial ceremonies also necessitated extensive tree cutting for the holy fire. As the Gangetic plains became saturated with cultivation, woodlands and pastures shrank and oxen became a critical source of agricultural power and trade. As a result there was a strong protest from the peasants and traders against animal sacrifice and tree cutting, reflected in two new religions, Buddhism and Jainism, by the sixth century BC (Gadgil and Guha 1992). Respect for all forms of life and nonviolence was taken to extremes by Jainism, where one sect wears no clothes but for gauze over the nose to avoid trapping and killing insects. Jainism is strongest in the northwestern subcontinent, a region with a particularly fragile environment, where the dangers of overexploitation are acute.

2.3 *Resource Impoverishment*

Kautilya's Arthashastra, a manual of statecraft dating from the third century BC, near the peak of imperial expansion in the region, provides insights into the ecologically prudent practices of the rulers, such as maintaining royal hunting preserves (Gadgil and Guha 1992). Poaching wild elephants from forests along the boundary maintained for military purposes attracted the death penalty. Large empires began to disintegrate after the seventh century AD, probably due to the gradual decline of agricultural fertility over a millennium. This eroded the dominance of traders and their main religions Buddhism and Jainism by the tenth century. This resulted partly from a revival of Hinduism, through transformations such as the incorporation of several tribal deities, and the abandonment of ritual animal sacrifice and consequent meat consumption. This revival consolidated Hindu caste society, which now manifested rigid barriers of endogamy such as those found among the tribes.

2.4 *Resource Partitioning by Castes*

However, unlike the tribespeople, the Hindu castes do not occupy exclusive territories and have a specialized mode of subsistence, which used to be hereditary. In a small region, one caste specialized in catching fish, a second in herding buffaloes, a third in making salt from sea water. A fourth caste may have tapped liquor from palm trees, a fifth may have cultivated paddy, and so on. Thus, about 10 to 50 sedentary castes would coexist nearly independently within an area of about 1,000 square kilometers. The same area would also be visited by dozens of other nomadic castes of artisans, pastoralists, and entertainers. These nomadic castes also ranged over fixed areas. Traditionally, all these castes carried out barter with one another.

The most remarkable feature of the caste society was the regulation of competition for limited resources through the diversification of resource use (Gadgil and Guha 1992) with a biological perspective, such as the partitioning of food resources for organisms with differing ecological roles in an ecosystem. Resource regulations included intracaste and intercaste territoriality manifested in land ownership, livestock category, or grazing seasons. Besides resource use diversification, this caste society also exhibited traditions of regulated harvests, some of which continue even today. An example is the setting up of sacred ponds upstream to protect fish breeding pools and confining their extraction during, a particular season, so ensuring good harvests.

Besides religious motivation, villagers are also often aware of the direct benefits of sustainability practices, such as the utility of bird guano from protected village heronries as farm manure. Such understanding has manifested folk conservation practices beyond focal habitats to species protected by specific communities. For instance, the Bishnois, a sect of Hindus in the semiarid northwestern subcontinent, sacrificed several lives by hugging their sacred trees to prevent the soldiers of the local king from cutting down the trees for a lime kiln for the king's palace. Such strict religious protection has meant that even today Bishnoi villages are green islands with freely roaming peacocks, blackbucks, and blue bulls, amidst an otherwise desolate landscape. Incidentally, while the sacred trees are very useful to the Bishnois, their animals are protected and fed with stored grain despite the crop depredation they inflict.

Conservation practices amongst rulers continued, such as the moratorium on felling green trees, imposed by King Shivaji in western India in the seventeenth century. This king's navy established the first plantations of teak trees long before the celebrated British plantations in Malabar two centuries later. Conservation traditions were not confined to tribal peoples or Hindus, as evident from sacred ponds around Muslim shrines in Bangladesh (Gadgil and Guha 1992).

3. *Colonial Conquest*

While the subcontinental society responded to environmental constraints by resource specialization and cultural prudence, Europeans responded to resource scarcity by exploring new lands and technologies around the fifteenth century. The ensuing industrial revolution tilted the balance of power from landlords to traders and industrialists. The British government gradually whisked away the common property resources in the subcontinent and actively eroded the traditions of prudent communal management such as the sacred groves, condemned by Westerners as superstitious or obstacles to state control (Brandis 1884).

Subsequent to the British invasion, reckless hunting and timber harvesting took place. This denuded the forests and drove animals such as the Asiatic lion and cheetah nearly to extinction. The realization of the harm gradually promoted so-called scientific forestry, which nevertheless had to hire continental foresters, as Britain had no tradition of scientific forest management. Eventually the British granted protection to royal hunting preserves and established some new game preserves. Meanwhile the continued breakdown of the traditional economy and the solidarity of village communities forced the indiscriminate use of common property resources, now owned by the government but with usufruct privileges to local people. Indiscriminate grazing, tree cutting, fishing, hunting, and other destructive activities increased, further eroding the impoverished resource base. The forests reserved for the exclusive use of the government, for example to provide timber for the emerging railway network, did not fare well either as local people agitated for cultivation on this forest land and failed to cooperate in its management. Thus, beyond the decimation of the environment, the British period marked an erosion of the conditions that had fostered an ethos of ecological prudence.

4. *Industrialize and Perish?*

4.1 *Subsidized Profligacy*

The British looked at the subcontinent initially as a source of cheap raw material and, later, as a market for manufactured goods. In consequence, they deliberately destroyed traditional industries such as handicrafts and prevented industrialization from gathering pace. When several countries in the region became independent after World War II, they vigorously promoted mechanized industrialization as a remedy for the maladies of an impoverished resource base and growing population. The philosophy 'industrialize or perish' was quickly embraced by the urban elite, while farmers and artisans could hardly participate.

Unbridled industrialization opened the gates to mass exploitation of the natural resources at throw-away prices to industry. Highly subsidized land, water, and power, and unregulated pollution soon severely eroded the environment and resources. However, the resultant phenomenal profits posted by the industries far exceeded the output of the agricultural sector, allowing industries to shift the resource base. Such immunity promoted ecological profligacy and prudence was thrown out of the window. Concomitant elitist conservation initiatives were confined to the conversion and expansion of hunting reserves into wildlife sanctuaries and national parks, further cornering the resources of the dependent masses, triggering fresh ecological conflicts that haunt most large projects such as irrigation and power schemes even today.

4.2 *Grassroots Conservation*

This profligacy of the social groups that have a wide resource base and import goods from all over the world, termed 'biosphere people,' contrasts with the prudence inherited by the masses, termed as 'ecosystem people' (because they are dependent on the consumption and trade of immediately available resources from predominantly natural ecosystems). The latter group is dependent on protecting its neighboring resource catchment (Gadgil and Guha 1995). This is contradictory to the Western, Malthusian approach which attributes environmental degradation to population explosion of poor people (Ehrlich 1969). Regional researchers retort that the much higher resource consumption levels of the Western elite cause greater environmental degradation. Imposing the costs not only of environmental degradation but also of conservation by the elite on the masses was soon met with fierce resistance.

Ensuing social conflicts prompted many fresh initiatives of ecological prudence on the part of the masses, who were left with no option but to protect their resource base. For instance, northern Himalayan people were agitated to discover that besides encouraging ecological insecurity, the government was starving their cooperative forest-based industry for the benefit of distant private sector industries. This led to a tree-hugging protest by local people along the lines of the Bishnoi, mentioned above, with women in the forefront, against industrial tree felling. It became famous as the Chipko ('hugging') movement.

Chipko triggered unprecedented environmental awareness and mass conservation movements. The repercussions are socializing the government-sponsored afforestation drive, though enhancing the stake of villagers in managing existing forests is proving much slower.

Fishermen along the coast are agitating to prevent mechanized fishing trawlers from overfishing and destroying the spawning beds, ultimately eroding the fish stocks. This causes immense harm to traditional fishermen who avoid fishing during the breeding season and use nets with a large mesh size so that small fish can escape. The traditional fishermen are also opposing the aquaculture industry, as it promotes prawn cultures by landlords with heavy chemical inputs that ultimately pollute and destroy the estuarine fisheries which sustain small fishermen. In places, these conflicts have even assumed bloody dimensions. Often, multinational corporations (MNC) provide the finance for such environmentally damaging and socially inequitable projects.

5. *Globalization*

The global dimensions of even localized conservation conflicts became more evident after the recent ban by the USA of marine food imports from India. The USA has appealed to the World Trade Organization (WTO) to declare a multilateral ban. The ban was justified on the grounds that it would foster conservation, as the Indian trawlers lack expensive turtle-excluding devices, primarily used and sold by the developed countries. The WTO has ruled out a multilateral ban, acceding to India's request that environmental disputes be addressed at more appropriate platforms such as the international Convention on Biological Diversity (CBD). The threat to mass livelihoods from the global elite still looms large under the guise of environmental protection, with developed countries unsuccessfully trying to impose environment-related trade restrictions. This trend was seen at the Seattle millennium conference of the WTO in 2000. If the trawler ban was ever implemented, it would not only affect Indian trawler owners but also conservation-minded masses such as the traditional fishermen who eke out a living from fish exports.

Another example of an environmental conflict with the global elite is the transgenic cotton controversy in southern India (Gadgil and Utkarsh 1999). Here, cotton farmers were faced with disaster after crop failure due to extensive pest attacks in 1998. Not long afterwards, a multinational agro-corporation announced that it was carrying out ongoing field experiments prior to the likely release of a pest-resistant transgenic cotton variety. This prompted widespread protest from environmentalists and farmers, who feared not only that the transgenic strain might pose environmental hazards, but that if the MNC had a monopoly over the strain they might end up dependent upon it. Further, such transgenic crops might accelerate the erosion of traditional cultivars, thus undermining environmental efficiency and farmers' rights to their seeds.

The controversy about large dams on the River Narmada and consequent unfair widespread displacement is a rare case where unrelenting protest by the masses resulted in the withdrawal of an international investor, the World Bank. However, transnational investors do not generally display such restraint, especially if they are wooed by local governments unmindful of the environmental hazards and related protests by local people.

6. *The Future*

Future environmental conflicts may be glimpsed from the ongoing legislative reforms taking place in the south Asian region (Gadgil and Utkarsh 1999). For instance, the Indian biological diversity bill and patent act amendments intend to compel even foreign agencies to pay royalties for commercial applications based on Indian biological diversity and related folk knowledge. This might prevent cases like patenting of innovations based on Indian biodiversity and knowledge abroad, as in the infamous cases of Basmati rice and Neem-based pesticide where patents were obtained abroad on improved rice variety and process to stabilize volatile Neem extract, without sharing any profits with the original South Asian farmers. The bill proposes spending such funds on incentives for promoting grassroots conservation. It also buttresses people's resource rights, in theory, by requiring the consent of relevant village councils for any development or conservation project. Implementation of such intentions is awaited, but many national and international projects are already encouraging villagers' involvement in sustainable management of even government property, enthused by the success of village forest protection committees in Nepal and India or water users' associations in India. Other initiatives drawing on substantial information industry strengths in the region include registration of people's knowledge of biodiversity use, to prevent unfair patenting.

While social systems within the region are slowly adapting to resource conflicts, new international conflicts are emerging, such as that over the CBD biosafety protocol. This has implications for the competition between local cultivars and the largely foreign-controlled genetically modified organisms (GMOs). GMOs are feared for their likely environmental and social hazards.

Debate between the countries of the region and developed countries also concerns country contributions to the causes and mitigation of the global climate change. While developed country agencies fix the blame on deforestation and rice cultivation in the region (Anon 1992), regional researchers claim the opposite

(Achanta 1993). Regional cooperation unfortunately seems to be lacking, unlike the case in other tropical regions. For instance, SAARC (the South Asian Association for Regional Cooperation), is hardly cohesive and vocal in pushing its case at the CBD and climate change negotiations, whereas its east Asian, African and South American counterparts are very vocal and united, probably due to people's pressure in those regions.

The partitioning of the people's stake in the environment forms the focus of ongoing environmental research debate in the south Asian subcontinent. Those claiming to represent grassroots communities shun all that is modern and Western as environmentally destructive (Shiva 1993). Other Western-influenced regional researchers reject such categorization of conservation efforts (Gadgil and Utkarsh 1999), and advocate strategies to cope with globalization rather than to denounce it and presume that it is avoidable.

7. *Conclusion*

In future, conservation and livelihood conflicts might escalate both nationally and internationally. The evolution of cultural practices of ecological prudence has characterized much of the subcontinental society until now (Kothari et al. 2000). However, the emerging dominant social groups such as corporate industries are unlikely to inherit this traditional ecological prudence or to evolve an alternative paradigm. This stems from their ability to harness resources from many alternative locations, employing evolving technologies and measuring benefits solely in monetary terms. This profligate behavior is rationalized in the framework of economic growth and technological progress. The article emphasizes that ecological prudence might be maintained only through an innovative blend of the emergent corporate culture and the masses that inherit the conservation traditions.

See also: Caste; Colonization and Colonialism, History of; Environmentalism: Preservation and Conservation; Globalization: Geographical Aspects; South Asia: Sociocultural Aspects; South Asian Studies: Culture; South Asian Studies: Economics; South Asian Studies: Geography; South Asian Studies: Politics

Bibliography

Achanta A N (ed.) 1993 *The Climate Change Agenda: An Indian Perspective*. Tata Energy Research Institute, New Delhi, India

Anonymous 1992 *World Resources 1992–93* World Resource Institute, Washington/Oxford University Press, New York

Brandis D 1884 *Progress of Forestry in India*. McFarlane and Erskine, Edinburgh, Scotland

Ehrlich P 1969 *The Population Bomb*. Ballantine Books, New York

Gadgil M, Guha R 1992 *This Fissured Land: An Ecological History of India*. Oxford University Press, New Delhi, India/University of California Press, Berkeley, CA

Gadgil M, Guha R 1995 *Ecology and Equity: Use and Abuse of Nature in Contemporary India*. Penguin Books, New Delhi, India/Routledge, London

Gadgil M, Utkarsh G 1999 Intellectual property rights and agricultural technology: Linking the micro- and the macro-scales. *Indian Journal of Agricultural Economics*. **54**(3): 191–210

Kothari A, Pathak N, Vania F 2000 *Where Communities Care: Community Based Wildlife and Ecosystem Management in South Asia*. International Institute for Environment and Development, London/Kalpavriksh, Pune, India

Mountain J L, Hebert J M, Bhattacharya S, Underhill P A, Ottolenghi C, Gadgil M, Cavalli-Sforza L L 1995 Demographic history of India and mtDNA-sequence diversity. *American Journal of Human Genetics* **56**: 979–92

Shiva V 1993 *Monocultures of the Mind: Perspectives on Biodiversity and Biotechnology*. Third World Network, Penang, Malaysia

G. Utkarsh

South Asian Studies: Geography

The landscape of south Asia contains tremendous natural and cultural diversity, and the geographic study of the region exhibits an equally wide range of themes. During the last quarter of the twentieth century, geographers shifted focus from a regional studies approach, which provides general overviews of key geographic characteristics of the region, to a problem-oriented approach that highlights contemporary social and environmental issues confronting it. This article briefly introduces the regional geography of south Asia and then examines in detail the themes that inform contemporary geographic study of the region.

1. *Regional Geography of South Asia*

Many of the early regional geographies of Asia contain sections on south Asia and provide useful summaries of background information about society and the physical environment across the Indian subcontinent. An exclusive focus on the region developed in Asian regional geography after the 1947 partition of the subcontinent and the emergence of the newly independent countries of India and Pakistan. The early regional studies offered comprehensive overviews with selected coverage of specific geographic details, notably population and economy. In the 1960s, important studies of individual countries in the region appeared, which provided greater depth and attention to local scale geography (Karan 1963).

The regional approach to south Asian geography gained in sophistication during the 1970s, when geographers blended the study of regional patterns with a critical focus on process and on changes in society and the landscape. A study of the historical evolution of south Asian geopolitics appeared in the form of the monumental *A Historical Atlas of South Asia*, which offered new insights into the ebb and flow of imperial power and the design of regional landscapes (Schwartzberg 1978). The historical geography approach not only highlighted the magnificent periods of civilization in the region, but also showed how the contemporary organization of society and land owes its spatial origin in great part to the political past. The connection between history and modern times occupies a nexus of geographic study about south Asia. The historical perspective in regional geography was advanced further in the 1970s with studies that interpreted the spatial organization of the subcontinental economy according to the impact of colonialism. This work called specific attention to the role of the British in fundamentally altering not only the globalization of the regional economy, but also the administration of south Asian society and the usage of land and other natural resources. The regional geographies, including atlases of the subcontinent that have appeared since the 1970s, blend the comprehensive and historical approaches (Farmer 1983). Much of the contemporary study of south Asian geography, however, favors a problem-oriented approach over the regional perspective.

2. *Contemporary Themes of South Asian Geography*

The general issues of modern society drive the current study of south Asian geography. Prominent among these are the place of tradition in modern life, the function of the peasant economy amid globalization, the dilemma of development, territorial movements and ethnic conflict, and environmental change. The coverage of these thematic areas is distributed at varying scales across the region, from village-level study to international inquiry.

2.1 *The Place of Tradition in Modern Society*

The wealth of tradition in south Asia is one of its defining features and the subject of much study on the cultural geography of the region. The geography of India, for example, is commonly defined according to cultural practices, and the spatial organization of the landscape results in part from such complex traditions of culture as pilgrimage and sacred places (Bhardwaj 1973). The maintenance of traditional practices in contemporary life reflects their continued importance even while the material conditions of life rapidly change. Among the various roles of culture in the region is the importance of tradition in determining how people use the land to obtain their basic needs. Environmental outlooks, organized according to indigenous knowledge systems, remain significant features for the geographic study of local conservation strategies and people-based environmental movements. Modern agriculture and forestry initiatives in south Asia owe a great deal to the traditional systems of environmental management that obtain in specific localities, and, indeed, the latter define in many cases the prospectus for sustainable development in the region.

2.2 *The Function of the Peasant Economy*

The geographic study of peasant agriculture illuminates the place of culture in nature, as well as the social structures that organize village production systems. Farming systems research throughout southern Asia informs our knowledge about such farmer practices as crop type and rotation strategies, gender divisions of work, and the application of technologies in village economic systems. The harmonization of village agriculture with the vagaries of climate is the focus of much study, notably on the impacts of severe weather events and flooding on rural production systems in places such as Bangladesh, where natural events combine with human settlements to create potential hazards. Droughts and famine compelled early geographical studies on the nature of village farming systems, but the persistence of poverty within the peasant economy of south Asia is what drives much contemporary geographical inquiry into the function of the peasant economy. The contradictory nature of the green revolution and the political economic structures that determine how farmers participate in agricultural development programs are the focus of some important geographical studies that employ a social theory perspective.

Geographers working in the cultural ecology and political ecology traditions have successfully utilized a systems perspective to examine the linkages between environment and society within the peasant economy. In this perspective, the household is the nexus for material and energy exchanges between forests, pastures, and farm fields. The main contribution of this approach to peasant studies is its ability to analyze the interplay of a diverse set of environmental and societal arrangements, which collectively define the function of the village economy. Generally, the systems approach is utilized at the micro or village level, but it also has been applied successfully to the regional setting, for example in the Punjab, with the application of agricultural land-use models.

2.3 *The Dilemma of Development*

Geographical inquiry into south Asian development processes has a strong backing among both the early

regional geographies and the more recent thematic studies. In the former case, the various development indices (both economic and social) are applied on a country basis to map patterns of development, but generally such efforts are flawed because too little attention is given to explaining the geographic inequities that appear on the maps. The more recent development studies in south Asian geography take a very different tack. They examine foremost the processes that contribute to the overall low level of development in the region and to the unequal distribution of development successes within it.

The efforts of British geographers working with the Overseas Development Group at the University of East Anglia have led the way toward a political–economic explication of poverty in Nepal (Blaikie et al. 1980). Their work, guided by center–periphery relations theory, has contributed a great deal to our understanding of underdevelopment not only in the Himalayan kingdom of Nepal but throughout the subcontinent. The role of infrastructure in the development of spatial economies, the significance of land tenure relations for resource entitlement, and the geographic variation in wealth and poverty related to structural problems within national economies are the major foci for such studies.

Most recent south Asian development studies by geographers have adopted a critical perspective that views poverty as a social construction rather than a product of ineffective economic policy. It is argued in this case that poverty is a result of social histories wherein some segments of society (or indeed entire countries) are made dependent upon other societal elements (or countries). In this vein, the role of foreign assistance as a development vehicle is challenged on the basis that it mainly serves Western interests and simply promotes further dependency among the poor nations of south Asia (Shrestha 1997). Furthermore, the geography of development in south Asia is shown to reflect inequities in the distribution of foreign assistance as a result of poor planning, fiscal mismanagement, and official corruption. The social theory and critical perspectives found among recent development geographies of the region lend sophistication to the conventional, neoliberal interpretations of development that describe most early regional geographies of the region.

Development geography in south Asia pays considerable attention as well to urban and regional planning. Detailed analyses of urban systems show the role of cities as growth nodes for regional economic development and modernization. Micro-scale studies of south Asian cities examine such phenomena as urban migration and the development of fringe squatter settlements, the role of the informal economic sector in meeting the needs of the urban poor, and the environmental impacts of urban growth (Dutt 1994). In rural south Asia, especially in the northern mountains, geographers have examined closely infrastructures associated with regional development, mainly road alignments, and have shown how these make formerly remote areas more accessible to the influences of globalization and modernity. The rural development studies also consider the geographic role of tourism, which is shown to widely transform regions throughout south Asia with its impacts on society and the natural environment (Zurick 1995).

2.4 *Territorial Movements and Ethnic Conflict*

The fractious nature of the south Asian landscape derives in part from the tremendous cultural diversity of the region and the competing claims on territory by ethnic groups. The study of ethno-geographic zones and territorial aspirations of ethnic groups contributes to an understanding of peace and conflict in the region. Ongoing separatist movements are found throughout the subcontinent, from Kashmir in the northwest to Assam in the northeast, through the India plains where Sikh, Gorkhali, *adivasi*, and other ethnic groups claim regional autonomy or independence, and on to Sri Lanka, where Tamils and Sinhalese are embroiled in a civil war. The conflict potential of ethnic groups is viewed to be a function of such motivating conditions as government encroachment onto ethnic territories and socioeconomic discrimination, and such enabling conditions as group cohesiveness, population expansion, and access to infrastructure, foreign support and resources. Geographical analyses of territory-based ethnic conflict consider these factors as well as other historical circumstances of the conflict regions.

2.5 *Environmental Change*

The causes and consequences of environmental approaches used to study environmental change are diverse and include geo-technical monitoring in the Eastern Ghats of southern India, landscape assessment in Rajasthan, irrigation and water resource development in the Gangetic plain, and studies of the relations among subsistence economies, the natural environment, and political economy in the Himalaya (Zurick and Karan 1999). The geographic research on environmental trends generally subscribes to a fieldwork-based research model and is theoretically informed by cultural ecology, ecosystems research, and political ecology. Much environmental geography in south Asia centers on the Himalaya, where international research efforts span the mountains from Afghanistan to Burma.

Environmental research in the Himalaya is informed in part by a proposed 'Theory of Himalayan Environmental Degradation', which links land degradation to population growth and overuse of resources, and by various theories of global environmental and

social change. The work of German geographers in the upper Indus river valley in Pakistan follows the geoecology tradition, which centers upon the human exploitation of altitudinal-based environmental zones, and examines human alterations of the environment for subsistence purposes as well as for modernization along the Karakoram Highway. The UNESCO Man and Biosphere Project contains the work of geographers in Nepal who have examined such issues as land use change, hill slope stability, and natural hazards in the middle mountains region. Much of the contemporary geography in the Himalaya seeks to understand the human underpinnings of environmental change, distinguishing societal causes from natural ones, and to resolve environmental problems according to human needs (Ives and Messerli 1989).

The applied aspects of environmental geography in south Asia include such issues as forest regeneration, sustainable agriculture, geographic information systems (GIS)-based monitoring of environmental trends, and the reconciliation of societal aspirations with mountain carrying capacities. A great deal of applied geography is conducted under the auspices of the International Center for Integrated Mountain Development (ICIMOD) based in Kathmandu, Nepal. ICIMOD's critical program areas include the impacts of globalization on fragile environments, livelihood security, and natural hazards management. These efforts apply broadly across south Asia, linking the environmental and social conditions in the mountains to those in the densely settled plains.

Future geographic studies on the region will continue to engage scholars in the pressing issues of society and the natural environment. The regional perspective will benefit by considering the Indian subcontinent in terms of globalization trends and in respect to its potential role in the geographic alliances emerging within Asia. Among a set of related topics are the impacts of ethnic fragmentation on the evolution of national boundaries and on international relations among the south Asian countries. Attention on culture and environment will continue to consider the place of tradition as a guiding force toward sustainable practices related to economic development and the use of the natural environment. The overall economic transformation of the region will necessitate new geographic appraisals of the peasant economy, societal equity, and land management, linking these within new conceptions of environmental conservation, cultural survival, and sustainable development.

See also: Colonialism, Anthropology of; Ethnic Conflicts; Peasants and Rural Societies in History (Agricultural History); Peasants in Anthropology; Regional Geography; South Asia, Archaeology of; South Asian Studies: Culture; South Asian Studies: Economics; South Asian Studies: Environment; Tradition, Anthropology of.

Bibliography

Bhardwaj S M 1973 *Hindu Places of Pilgrimage in India*. University of California Press, Berkeley, Los Angeles, and London

Blaikie P, Cameron J, Seddon D 1980 *Nepal in Crisis: Growth and Stagnation in the Periphery*. Oxford University Press, Oxford

Dutt A 1994 *The Asian City*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Farmer B H 1983 *An Introduction to South Asia*. Methuen, London/New York

Ives J D, Messerli B 1989 *The Himalayan Dilemma*. Routledge, London/New York

Karan P P 1963 *The Himalayan Kingdoms: Bhutan, Sikkim, and Nepal*. Van Nostrand, Princeton, NJ

Schwartzberg J 1978 *A Historical Atlas of South Asia*. University of Chicago Press, Chicago/London

Shrestha N R 1997 *In the Name of Development*. University Press of America, Lanham, MD

Zurick D 1995 *Errant Journeys: Adventure Travel in a Modern Age*. University of Texas Press, Austin, TX

Zurick D, Karan P P 1999 *Himalaya: Life on the Edge of the World*. Johns Hopkins University Press, Baltimore/London

D. N. Zurick

South Asian Studies: Health

1. Background

1.1 Common Heritage

For a proper appreciation of the status of the study of health in south Asia, it is important to underline certain facts pertaining to the character of the region and to health as a field of study. The seven countries in the region share a common cultural heritage as well as political history. Three of these countries—Bangladesh, India, and Pakistan—accounting for nearly 97 percent of the population of the region, were one country until 1947. The cultural and religious ties between India, Nepal, Bhutan, and Sri Lanka predate recorded history. This affinity and closeness have not only produced a similar mindset but also a largely common history of medicine and health.

Further, the story of health research in this region is almost as long as human history itself. This history is divided into four periods, each of which is marked by distinct characteristics regarding the nature and extent of health research.

2. Study of Health

2.1 The Pre-Islamic Period

The history of health in South Asia starts with the *Atharvaveda*, the last of the four Vedas—ancient

collection of sacred Hindu hymns, pages, and liturgical formulae. The *Atharvaveda*, which is said to date back to the second millennium BC, is the seminal source of Indian medicine. From it emerged the better known *ayurveda*, credited to Dhanvantari who supposedly received it from Brahma, the God of creation in the Hindu Trinity. The period between 800 BC and AD 1000 is regarded as the golden age of Indian medicine. During this period there was great advancement in knowledge about diseases, techniques of surgery, and the concept of health. Two very well-known medical treatises, *Charak Samhita* by Charak, a physician, and *Susruta Samhita* by a surgeon, Susruta, were also produced during this period. Although the exact dates of their writing are debatable, both these works are considered far older than any counterpart work produced in Greek medicine, giving rise to a claim for the priority of Indian scientific medicine. During this era, a sophisticated concept of health emerged, postulating interactive unity among soma, psyche, and soul (not too different from the modern concept of holistic medicine), an identification of water, food, physical and occupational environment, sanitation, hygiene and lifestyle, biological and karmic inheritance, and sociopolitical order as primary determinants of health.

Because of religious taboos on cutting bodies and killing animals for experimentation, deductive reasoning drove the study of medicine to a large extent. In this pursuit, the thinking about human anatomy and physiology was greatly influenced by the prevalent concept of universe; the human body was perceived as a reflection of the cosmos. However, there is some evidence to indicate a knowledge of germs and bacteria, disease etiology, and the use of empirical observations to determine the effectiveness of various treatments. There was detailed knowledge about a large number of diseases; the extensive *materia medica* included at least 760 medicinal plants, herbs, and spices, as well as a large number of minerals and animal remedies based on the milk of various animals, bones, dung, and urine. The list of surgical instruments invented and in use was equally impressive (Kutumbiah 1962, Zimmer 1948). But in later years there was a noticeable slowdown in the growth of medicine in this region, probably due to increased emphasis on the theory of 'karma' and wider prevalence of a medieval ethos characterized by feudatory attitudes and parochial concerns.

2.2 Islamic Period

The arrival of the Turks (around AD 1000) brought Indian medicine to a standstill because of the deterioration of traditional patronage. With the strengthening of Muslim hegemony, *unani* medicine (a mixture of Greek, Arab, Turkish, and Iranian medicines) gained the favor of the ruling class, and spread with the Muslim conquests. However, little documentary evidence is available regarding the state of medicine and health during this period, beyond the prevalence of epidemics and famines. Scholarship during this period seems to be concerned primarily with the importation of medical know-how from Iran, Turkey, and Arab countries.

2.3 The Colonial Period

The colonial period started on November 24, 1510, when Alfonso de Albuquerque captured Goa, and ended in 1948 when the British withdrew from Sri Lanka, after having withdrawn from India, Pakistan, and Bangladesh a year earlier.

During this period, the government by and large limited its attention to the healthcare of its employees, the military, and plantation workers. For this purpose, government health clinics and hospitals were built at most of the district headquarters, major cantonments, primary railway depots, and on tea and coffee plantations. (Due to the greater concentration of tea plantations in Sri Lanka, a proportionately larger health infrastructure was instituted there.)

Christian missionaries also built hospitals in major cities and established clinics in the tribal and backward areas. To staff these facilities doctors and nurses were initially imported from Britain, but later several medical colleges (the first of which was established in Madras in 1835) were set up to train local manpower. In 1921 a School of Tropical Medicine was established in Calcutta to train personnel for public health work. In 1932 it was upgraded to become the All India Institute of Hygiene and Public Health (Kumar 1998). A system of licensing the practitioners of allopathic medicine was introduced. However, the private practitioners of *ayurvedic* and *unani* medicine, who were not licensed, provided most of the medical services. Scholarship on health was largely limited to reports based on the decennial census introduced in 1872, and on vital statistics, although the latter were inaccurate due to a widespread failure to register births, deaths, and other vital events. Most of these studies, produced by Westerners, were part of larger sociopolitical concerns. These included official reports, travelogs, diaries, and the occasional book. (*Mother India* is an example of such a book, written by an American, Katherine Mayo, which became highly controversial due to its political overtones.)

In 1943, towards the end of the British rule in the region, an official committee under the chairmanship of Sir Joseph Bhore was appointed to produce a detailed report on health conditions in India (Pakistan and Bangladesh were then parts of India) and to make recommendations for improvement. The committee was called the Health Survey and Development Committee, and is popularly known as the Bhore Committee. This committee submitted its report in 1946, which is considered a seminal work on health-

care in this region insofar as it laid the basic foundation of scholarship as well as for political actions regarding the importance of health and the role of government in the provision of health services.

2.4 The Postcolonial Period

The basic character of the health system in South Asia was formed during colonial rule: the supremacy of allopathic medicine; an ambivalent concept of healthcare (human service vs. economic necessity); a high degree of dependence on the West for basic knowledge and know-how in all matters pertaining to health; a largely reactive role for the government, and a dependence on the government for health statistics.

Further, the deep-rooted cultural tendency to rely on deductive logic for problem-solving continued to be the hallmark of scholarship, placing high reliance on creative thinking over empirical proof. This is reflected in the fact that until recently the primary landmarks in the history of scholarship on healthcare in this region were a series of expert reports, starting with the Bhore Report.

The Bhore Committee, reversing the minimalist approach of the British government in India to the provision of healthcare, advocated that the government should take responsibility for ensuring universal access to preventive and curative health services because health was too important to be left to the vagaries of circumstances; it recommended the development of a nationwide network of primary health centers, the provision of integrated preventive and curative health services to all regardless of ability to pay, extensive health education to modernize the knowledge, beliefs and practices of the people, the training of needed health manpower, and the construction of necessary health facilities.

These recommendations reflected the mood of the time: the success of the Soviet Union in improving the health status of its population through central planning, the release of the Beveridge report in 1942 in Great Britain (recommending universal healthcare) which led to the passage of the National Health Services Act in 1946, and the high emphasis placed on social welfare by the leaders of the independence movement in the South Asian countries.

The Bhore Committee recommendations were incorporated in the first Five Year Plan of India (1951–6), and as a result a foundation was laid for the development of a nationwide network of primary health centers, as well as for the expansion of the training facilities to provide the necessary manpower. This foundation was built upon in each of the successive Five Year Plans, guided by a series of other expert committee reports: Health Survey and Development Committee (chaired by A. L. Mudaliar) 1961; committees on the integration of health services and family planning, 1963, 1968, and 1970; committee on multipurpose health workers, 1973; Joint Committee of Indian Council of Social Science Research and Indian Council of Medical Research, 1978, and so on. Other countries in the region also followed more or less a similar path and during this period expert committee reports remained the dominant focus of scholarship pertaining to health.

3. Recent Developments

3.1 Health Research

Since the early 1970s the study of health in this region has received increasing attention from international organizations (WHO, UNICEF, World Bank and USAID), other donor organizations, and American and British universities. This heightened attention is a result of a significant flow of foreign assistance to contain the burgeoning population and high saliency health problems. However, in-region scholarship is still in an early stage of development.

Earlier, it was noted that in this region deductive reasoning is a dominant scholarly tradition. This tradition persists despite the fact that the educational system is Western. But unlike those in the West, the institutions of higher learning in south Asia do not emphasize and reward empirical research. Instead, importance is given to the writing of treatises. This is reflected in the fact that while a Medline search covering the years 1980–2000 produced only 483 articles (around 25 articles per year, on average)—the majority of which reported biomedical research, advocated some policy posture or were descriptive in nature—the Sage India's summer 2000 list of books (covering the period 1998–2000) included 24 titles on various aspects of health and an additional 26 titles related to health. There are very few refereed journals published in south Asia that are concerned primarily with the study of health systems.

The *National Medical Journal of India* of the All India Institute of Medical Sciences, New Delhi, has been published since 1987. In recent years, three new journals have been established: *Journal of Health and Population in Developing Countries* (originally published in India by the University of North Carolina, but now moved to the USA), *Journal of Health Management* (published by the Indian Institute of Health Management and Research, Jaipur) and *Journal of Health, Population and Nutrition* (published by the International Centre for Diarrheal Diseases Research, Bangladesh). Funds for research also tend to be very scarce. The major sources of these funds are donor organizations, but they fund only those studies that meet their own needs. As a result, the researcher-initiated studies tend to be limited to doctoral dissertations, few of which are published.

The studies on health may be grouped as follows.

Table 1
Disability adjusted life expectancy (DALE) in South Asian countries, 1997

Country	DALE (years)	Ranking (184 countries)
Bangladesh	49.9	140
Bhutan	51.8	138
India	53.2	134
Maldives	53.9	130
Nepal	49.5	142
Pakistan	55.9	124
Sri Lanka	62.8	76

Source: WHO *The World Health Report* 2000, Annex Table 5 (for DALE scores); Annex Table 1 (for international ranking)
Note: DALE is calculated by subtracting the average number of expected years of disability due to illness and injury throughout life from the average expected years of life at birth

Table 2
Ranking of South Asian countries for health system performance, 1997 (191 countries)

Country	Performance on health level	Overall performance
Bangladesh	103	88
Bhutan	73	124
India	118	112
Maldives	147	147
Nepal	98	150
Pakistan	85	122
Sri Lanka	66	76

Source: WHO *The World Health Report* 2000, Annex Table 10
Note: Performance on health level is calculated by adjusting the DALE score for the efficiency of the health systems. Overall performance is a composite measure of achievement in the level of health, the distribution of health, responsiveness to people's concerns, fairness of financial contributions and managerial efficiency. For details, see *The World Health Report* 2000, explanatory notes, pp. 141–51

3.2 Health System Achievements

After 50 years of independence, the goal of universal access to preventive and curative health services in South Asia is far from fulfilled. This perception has been a catalyst for a number of studies typified by Qadeer and Sen (1998) and Wysocki et al. (1990).

A large number of studies prepared by the governments in this region describe progress made. For example, various reports produced by the Ministry of Health in India point out that a network of health centers, each serving 5,000–10,000 persons and staffed by auxiliary health workers, covers the country. In 1947, there were 26 medical colleges in India; by 2000 there were 146. Further, India is largely self-sufficient in drugs and medical supplies. A significant research infrastructure has been developed (some of the more prominent bodies include the All India Institute of Medical Sciences, National Institute of Communicable Diseases, National Institute of Health and Family Welfare, Haffkine Institute, Pasteur Institute, National Institute of Virology, and International Institute of Population Sciences). Two major social insurance programs provide medical care to targeted populations: the Employees State Insurance Corporation, established in 1948, provides healthcare to the employees of private industry and their dependents (nearly 3.5% of India's population); and the Central Government Health Insurance Scheme, which covers the employees of the central government and their families (more than 3 million people).

India's average life expectancy at birth rose from 41 years in 1951 to more than 60 years by 1995, and infant mortality rate fell from 146 infant deaths per 1,000 live births in 1951 to 71 per 1,000 live births in 1997. Other countries publish similar reports of progress.

These reports by the governments are often incorporated in the reports of international and other donor organizations; very few of them collect independent data. However, in 1993 the World Bank departed from this tradition by publishing a report, *World Development Report: Investing in Health*, based largely on independently collected data. Now this initiative has been carried forward by WHO in its *The World Health Report* 2000, in which countries are rated for their various achievements. According to this report, the South Asia region fares poorly both on the health status of its population (see Table 1) as well as on the overall performance of health systems (see Table 2).

3.3 Factors Hindering Progress

Increasing efforts are being made to understand the reasons for the poor performance of health systems in South Asia. The donors, mainly the WHO, the World Bank, UNICEF, and USAID, lead these efforts. The WHO has asserted that 'health is fundamentally different from other things that people want' (WHO 2000, p. 4) and, citing Jonathan Miller (Miller 1978) about its importance, has argued that health requires priority not only in funding decisions but also in ensuring effective access to the services. However, its analysis indicates that this argument seems to have had little effect on resource allocation decisions in South Asia. Countries in this region rank among the lowest for average per capita total expenditure on health as well as for public expenditure (Table 3). It is argued that the low priority given to health in this region is typical for developing countries. Generally, there is a strong positive correlation between the wealth of countries and the amount spent on health. The poorer the country, the lower the proportion of its wealth devoted to health (World Bank 1993, WHO 2000). It may be because the poor, after spending their income on food and shelter, are left with little for other things. But it may also be because the governments in the developing countries, even after the end of colonial rule, continue to perceive spending on health as 'expenditure,' not as 'investment.'

Table 3
Health expenditure in South Asian countries, 1997

Country	Total per capita expenditure in international dollars (1)	Total expenditure as % of GDP (2)	Public expenditure as % of total expenditure on health (3)	Public expenditure on health as % of total public expenditure (4)	Ranking for per capita total expenditure on health (184 countries) (5)
Bangladesh	70	4.9	46.0	9.1	144
Bhutan	82	7.0	46.2	10.1	135
India	84	5.2	13.0	3.9	133
Maldives	284	8.2	63.9	11.0	76
Nepal	41	3.2	26.0	5.3	170
Pakistan	71	4.0	22.9	2.9	142
Sri Lanka	77	3.0	45.4	5.2	138

Source: WHO *The World Health Report* 2000, Annex Table 8 for columns 1–4, and Annex Table 1 for column 5

Table 4
Ranking of South Asian countries for equality of access to health services and fairness of financial contributions (191 countries)

Country	Equality of access	Fairness of financial contributions
Bangladesh	125	51
Bhutan	158	89
India	153	42
Maldives	134	51
Nepal	161	186
Pakistan	183	62
Sri Lanka	80	76

Source: WHO *The World Health Report* 2000, Annex Table 5 for equality of access; Annex Table 7 for fairness of financial contributions

Several studies have focused on the issues of access to health services and financial burden (see Table 4), and have argued that the prevalent inequality is caused by uneven distribution of services between urban and rural areas, poor means of transportation and communication, gender bias and a high prevalence of for-profit health services (reliance on the private sector is 87% in India, 79% in Pakistan to 54% in Bangladesh). Further, it is observed that the poor carry a disproportionately heavy financial burden for obtaining health services, because of their lack of political power: not having effective access to state-supported health services, they are forced to purchase them. Even after paying for them, the services they get tend to be of poor quality. According to one estimate, more than 75 percent of all healthcare in India was delivered by 'private' physicians, most of whom lacked proper training in either allopathic or other healing traditions (Garrett 2000). The situation is not very different in most of the other countries in the region.

Other factors for the poor performance that have been identified include the scholarly tradition, the role of donors, prevalent work ethics and the unreliability of government-generated data.

It is argued (Jain et al. 1999) that the dominance of scholarship based on deductive reasoning generates interesting debates, but little action. Most of the energy in such debates is dissipated in derecognizing those from other groups—a notable characteristic of highly populated societies in which people face acute competition for survival and advancement. This not only makes decision-making difficult but, more importantly, it frustrates implementation. The countries in this region are known for the loftiness of their goals, progressive policies and legislation, and noticeably poor implementation due to the absence of what is described variously as 3Cs—concern with the issues, commitment to resolve them, and competence needed to resolve them—and also as 'organizational capital' (Brynjolfsson et al. 2000).

The scholarly tradition is only partly to be blamed for poor implementation. The prevalent work culture and work ethics have been found to be equally powerful factors. The dominant work culture is feudal and colonial, which denigrates 'work' and has a high tolerance for avoidance of accountability. As a result, 'delinquent' (absentee) and 'ghost' (fictitious) employees, toothless mandatory program evaluation, and noticeable non-performance, have become important features of the health systems in these countries (Jain et al. 1995). In combination with pervasive corruption (Chadda 2000, World Bank 1998, Jain et al. 1996, Jain and Nath 1996, Jain et al. 1997), the work culture saps the energy of the health system in many different ways.

3.4 The Role of Foreign Assistance

Foreign donors, including international organizations, bilateral bodies, foundations, non-governmental organizations (NGOs), churches, and charities, make significant financial contributions to the health

Table 5
Foreign assistance for health, 1990

Country	Total aid (million $)	Aid per capita ($)	Aid as a percentage of total health expenditure
Bangladesh	128	1.2	17.9
Bhutan	NA	NA	NA
India	286	0.3	1.6
Maldives	NA	NA	NA
Nepal	33	1.8	23.6
Pakistan	76	0.7	5.4
Sri Lanka	26	1.5	7.4

Source: World Bank *World Development Report* 1993, Table A-9

budgets of the countries in the region (Table 5). But this contribution is perceived increasingly as a mixed blessing, and as a result, a new body of literature is emerging which identifies the pros and cons of foreign assistance and proposes new models for making this assistance more effective.

Most of this literature is not health-specific, but the findings are generally applicable to this field. Following publication of the very popular book *The Ugly American* (Lederer and Burdick 1958), *The Ugly Indian*—a column written by the well-known Indian journalist, Kuldip Nayar (May 2000)—highlighted the excesses of Indian aid to Nepal. Dennis Merrill in his book *Bread and Ballot* (1990) reviewed US assistance for economic development during 1947–63. Choudhary (1991) produced a similar study on Canadian aid to countries in South Asia. Ghai (1991) examined the role of the International Monetary Fund in developing countries; and Rao (1999) and Priya and Qadeer (1998) examined the impact of the World Bank's conditional assistance for healthcare programs in South Asia. A number of other treatises also appeared (Ghosh 1984, Kanbur et al. 1999) examining prevalent models of developmental assistance and proposing new ones. In essence, these studies observed that while the countries received much-needed resources, the process of foreign assistance compromised their freedom of action, especially in those countries in which the foreign assistance accounts for large proportions of the health budgets.

This happened in several different ways. Each donor is led by its values, priorities, style, needs and power position. Consequently donors, whether individually or as a group, are rarely interested in the overall improvement in the health system of the country, but only in those parts that interest them. For instance, the WHO has been promoting primary healthcare under the flag of 'Health for All by Year 2000' (Wysocki et al. 1990); the US Agency for International Development (USAID) is interested mainly in the provision of family planning services through NGOs (Misra 2000); The World Bank is focused on 'structural adjustments,' privatization, and cost recovery (World Bank 1993, Rao 1999); and UNICEF on immunization of children. Each donor has different rules for project development, implementation and monitoring; and therefore makes different demands on the governments. More importantly, using their power position, the larger donors have tended to usurp the primary initiative for the stewardship of their health systems from the countries. As a result, governments have tended to spend a significant amount of their energy in addressing the interests and concerns of the donors, instead of providing the needed leadership to their health systems.

This incapacity for effective leadership is further compounded by the fact that the top health leadership in the governments tends to suffer from a high rate of turnover. One study found that the average tenure of health ministers and health secretaries was less than two years, while the senior staff of the donor organizations stayed in their posts an average of four years (Jain et al. 1995).

While the operating style of the donor organizations and their representatives has hurt the morale of the health leadership in the region, it has also contributed to the spread of corruption. However, this issue has not received focal attention in the studies on foreign assistance in South Asia, because until recently 'corruption' was a taboo term. Although everybody talked about it in informal settings, nobody wrote about it. Recourse was taken to euphemisms such as 'work ethics' and 'unhealthy conduct' when referring to corruption. This inhibition has prevented serious study of this grave problem. In this connection, special attention needs to be given to the staff evaluation policies of donor organizations. Most donor organizations, in the evaluation of the performance of their staff, place heavy emphasis to their ability to disburse aid funds. This policy places some pressure on their staff to make them resort to an 'end-justifies-means' approach (Jain et al. 1996, Jain and Nath 1996, Chadda 2000).

3.5 Health Statistics

South Asia is known for the superiority of its know-how in mathematics, statistics, demography, and computer sciences. Several world-class institutions such as the Indian Statistical Institute and International Institute of Population Sciences are located here. It has a long history of good quality census. Despite all these positives, the overall quality of health statistics in the region is fair to poor. This phenomenon has attracted the attention of major donors who have a large appetite for good data. Therefore, they have been probing this issue to find out the causes so that they could facilitate the needed correction. Most of these studies are in the form of consultancy reports, and they have identified a diverse group of factors

responsible for the poor quality of health statistics in the region. These include: health systems in these countries are dominated by physicians with little training in data management (Jain 2000), health data are generated primarily for the donors, health managers in these countries are more inclined to base their decisions on intuition and situational expediency than on data (Jain et al. 1999), the overall system of health data management is so out of date and inefficient that the data needed for decision-making are often not readily available, most program managers do not know how to use data in their work, target-oriented program planning generates pressure for falsification, and corruption needs the darkness of non-working information systems.

To compensate for the poor reliability of official health statistics, USAID supports an independent sample survey to monitor progress in health and population programs, the National Family Health Survey (NFHS), which is conducted periodically and covers several countries in this region as well as many other developing countries. In India, the autonomous National Sample Survey Organization (NSSO) was established in 1950. This organization, which has a reputation for producing high-quality data, was originally intended to collect data on the state of the economy but now also gives attention to health and other social sectors. However, the health statistics produced by NSSO and NFHS are fairly limited in scope. Improving health requires improved stewardship of the health system, and this, in turn, requires a sound information system. There is a strong need for active research to improve the quality of health statistics in the region.

3.6 Emerging Issues

To improve the health of their people, the countries in South Asia need to surmount the problems of high rate of population growth, infant mortality, maternal mortality, work injuries, and road accidents; contain malaria, TB and other old diseases as well as the new threat of HIV/AIDS; address the problems of a rapidly growing elderly population, rampant urbanization, and deteriorating environment; improve access to and quality of health services; deal with new pressures caused by new international forces such as the World Trade Organization (WTO) and Intellectual Property Rights (IPR), and get ready for new battles over linking trade to child labor and the environment, and foreign aid to privatization, structural adjustments, and corruption (Priya and Qadeer 1998). This poses a huge agenda of research for the social scientists interested in South Asian health.

See also: AIDS, Geography of; Area and International Studies: Development in South Asia; Health: Anthropological Aspects; South Asian Studies: Politics.

Bibliography

Brynjolfsson E, Hitt L, Yang S 2000 Intangible assets: how use of information technology and organizational structure affects market valuations. *MIT Working Paper*, July 2000

Chadda M 2000 *Building Democracy in South Asia*. Sage, New Delhi, India

Choudhary N K 1991 *Canada and South Asian Development*. Brill Academic, New York

Garrett L 2000 *Betrayal of Trust: The Collapse of Global Public Health*. Hyperion, New York

Ghai D P 1991 *The IMF and the South*. St. Martin's Press, New York

Ghosh P K 1984 *Foreign Aid and Third World Development*. Greenwood Press, Westport, CT

Health Survey and Development Committee, India, 1946 *Report of the Health Survey and Development Committee*. Manager of Publications, New Delhi, India

Jain S C, Gupta S D, Srinivasan P 1995 *Improving Health System Management in Rajasthan*. Carolina Consulting Corporation, Chapel Hill, NC

Jain S C, Nath L M, Ahmed J S 1996 *End of Project Evaluation: Health and Family Welfare Area Project, Rajasthan, 1989–96*. UNFPA, New Delhi, India

Jain S C, Nath L M 1996 *End of Project Evaluation: Family Welfare Area Project II, Himachal Pradesh, 1990–96*. UNFPA, New Delhi, India

Jain S C, Nath L M, Assifi N 1997 *End of Project Evaluation: Family Welfare Area Project II, Maharashtra, 1990–96*. UNFPA, New Delhi, India

Jain S C, Barkat A, Lundeen K, Faisel A J, Ahmed T, Islam M S 1999 Improving family planning program performance through management training: the 3Cs paradigm. *Journal of Health and Population in Developing Countries* **2**: 1

Jain S C 2000 Education and training: capacity building. *Journal of Health and Population in Developing Countries* **3**: 1 (Special Issue on WHO (SEARO) Regional Conference on Public Health in south-east Asia in the twenty-first century)

Kanbur R S M, Sandler T, Morrison K M 1999 *The Future of Development Assistance*. Overseas Development Council, Washington, DC

Kumar A 1998 *Medicine and the Raj: British Medical Policy in India 1835–1911*. Sage, New Delhi, India

Kutumbiah P 1962 *Ancient Indian Medicine*. Orient Longman, Bombay, India

Lederer W, Burdick E 1958 *The Ugly American*. Norton, New York

Lele U J, Nebi I 1991 *Transition in Development*. ICS Press, Oakland, CA

Merrill D 1990 *Bread and the Ballot*. University of North Carolina Press, Chapel Hill, NC

Miller J 1978 *The Body in Question*. Random House, New York

Misra S 2000 *Voluntary Action in Health and Population*. Sage, New Delhi, India

Priya R, Qadeer I 1998 Impact of structural adjustment policies on health in south Asia. *The National Medical Journal of India* **11**(3): 140–2

Qadeer I, Sen K 1998 Public health debacle in south Asia: a reflection of the crisis in welfarism. *Journal of Public Health Medicine* **20**: 1

Rao M 1999 *Disinvesting in Health: The World Bank Prescriptions for Health*. Sage, New Delhi, India

World Bank 1993 *World Development Report: Investing in Health*. Oxford University Press, New York

World Bank 1998 *Assessing Aid: What Works, What Doesn't and Why*. Oxford University Press, New York
World Health Organization 2000 *The World Health Report 2000, Health Systems: Improving Performance*. WHO, Geneva, Switzerland
Wysocki M J, Krishnamurthi C R, Orzeszyna S 1990 Monitoring the progress of health-for-all strategies: the situation in south-East Asia. *World Health Statistics Quarterly* **43**:
Zimmer H R 1948 *Hindu Medicine*. Johns Hopkins Press, Baltimore, MD

S. C. Jain and N. Johri

South Asian Studies: Politics

The countries of South Asia that matter most have a colonial legacy in common. The British gradually established universities teaching social sciences all over the Indian subcontinent. This policy prepared the ground for the development of political science in most South Asian countries. The pattern is different in Nepal, which had eluded the British Raj. In this Himalayan kingdom social sciences remained centered on the sole department of political science, at Tribhuvan University, that was founded in 1961. It gained momentum gradually, not only because of the democratization process but also because Katmandu became the center for South Asian associations—including NGOs (nongovernmental organizations) and journals—dealing with education or human rights issues, like Himal and the South Asia Forum for Human Rights.

The most common, pattern, however, based on British-built institutions is found in India, Pakistan, and Sri Lanka. In the latter country, Colombo is the most active place, not only because of the university but also because of research centers such as the Regional Centre for Strategic Studies. Like Nepal, Sri Lanka can use its relatively neutral position vis-à-vis the Indo-Pakistani conflict for developing South Asian studies. However, political science remains rather underdeveloped. Interestingly, at the University of Colombo, political scientists are affiliated to the Department of History and Political Science (which was created in 1963). Besides, scholars tend to focus on one political issue, that is ethnic conflicts. (In 1982, an International Centre for Ethnic Studies was launched in Colombo). Often, the scholars who approached the national question either were not political scientists (Tambiah 1986, an anthropologist by training) or did not stay in Sri Lanka (Wilson 1974, 1988, occupied first the Chair of Political Science at the University of Ceylon but was appointed Professor of Political Science at the University of New Brunswick in 1972). This 'extraversion' is one of the specific features of the political science of India and Pakistan, the two countries on which we shall focus in this article.

1. *The Institutional Bases: Building Upon the British Legacy*

1.1 *The Indian Efflorescence*

Political science has flourished in India whether in the examination of the political system *per se* and its links to society or in the study of international relations. The British contributed to these postcolonial developments from two points of view. First, they enabled Indian students to study politics in England, notably at the London School of Economics. Second, they introduced courses of political science in Indian university, more especially in Calcutta (where U. Goshal became lecturer in Comparative Politics in the 1920s) and in Allahabad, where Beni Prasad (who was to found the Indian Association for Political Science and the *Indian Journal for Political Science*) was Reader in Civics and Politics in the 1920s too. However, these two cities receded in the background gradually as political institutions were transferred to Delhi.

With its twin educational and research institutions, the Jawaharlal Nehru University (JNU) and the Delhi University, India's State capital has quite logically been at the heart of the whole process. The School of Social Sciences and the School of International Relations at the JNU and the Faculty of Political Sciences at the University of Delhi were both founded in 1955. Partly, because of the influence of Marxist thought, until the late 1980s, the historiographical tradition was strong, particularly in the JNU, where the study of politics was above all seen in interlinked terms of class and political economy.

In spite of the domineering presence of these two rival institutions, some other research centers based in Delhi such as the Centre for the Study of Developing Societies (CSDS) founded in 1963 and the Institute for Defense Studies and Analyses (IDSA), founded in 1965, have managed to grow outside their fold and have been able to contribute new approaches in their own, very specialized, domains. The Centre for Policy Research (CPR), initially established in 1973 as an informal think-tank has also grown into a major center for studying public policies as well as foreign affairs and the political sociology of India.

In the same manner, political science departments have blossomed throughout India and can be found in most, if not all, the state universities. To name but a few, Calcutta, Chandigarh, Hyderabad, Jaipur but also Banaras and even smaller places like Meerut are endowed with universities that have their own departments of political science and in some cases also of

international relations, where research and teaching are being conducted with a focus on regional and local issues. This has in its times led to the development of journals and associations that are as many proofs to the vitality of the subject. The Indian Association for Political Science—which brings together large contingents of political scientists to the conferences of the International Political Science Association—edits the *Indian Journal of Political Science*. The *Indian Political Science Review* is another such journal based in Meerut University.

1.2 Outward-looking Pakistan

Political science is not as well established in Pakistan as in India but international relations tend to be studied by a comparatively larger number of institutions. The gap may be explained from three points of view. First, the political context was not very conducive to the development of this discipline: the prolonged phases of authoritarianism under Ayub Khan (1959–70), Zia ul Haq (1977–88), and Pervez Musharraf (1999–) did not allow much freedom of investigation and the instability of the regime did not help the conceptualization of what is the Pakistani political system. The phases of civil rule were not conducive to research work either. For instance, G. W. Chaudhury, who had published two reference books in the 1950s–1960s, was forced to exile under Zulfikar Ali Bhutto because of his account of the war of Bangladesh. Second, political scientists were dominated by history, the most influential social science because they took refuge in history for the reasons mentioned above, and, like in India, the historiographical tradition overpowered other disciplines. Third, political science—like many other social sciences suffered much from the weaknesses of the university system: the textbooks are substandard and the teaching of the discipline is especially bad because in most universities and colleges it is taught in Urdu—a language in which there are no quality standard textbooks (Shafqat 1989).

In stark contrast to the situation prevailing in India, there is no association of political scientists and no 'Journal of Political Science' in Pakistan. Certainly there are departments of political science in many Pakistani Universities—including Karachi and Lahore—but not in some of the largest ones. On the other hand, international relations are taught in many places including the departments of international relations of the Qaïd-i-Azam University (Islamabad), founded in 1973, the University of Karachi (1958), and the University of Sindh (1972). Besides these departments, there are many more specific bodies dealing with international affairs—such as the Institute of Regional Studies (1982) and the Institute of Strategic Studies (1973)—or with area studies, such as the Centre for South Asian Studies of the University of the Punjab (1973) and the Central Asia Centre of the University of Peshawar (1976). This list reflects the traditional outward looking tendency of Pakistan, a country at the crossroad of South Asia, Central Asia, and the Middle East and affiliated to associations of regional cooperation and defense since the 1950s. However, India and Pakistan have more in common as far as the relationship of their political scientists with the outside intellectual world is concerned.

1.3 The Study of South Asian Politics Abroad

The command of English by South Asian political scientists partly explains their sometimes-strong affinities with Anglo-Saxon specialists in the field. First, many of them studied abroad and settled down there. Among the founding fathers of 'Pakistani political science,' one finds several expatriates including Anwar Hussain Syed, Khalin Bin Sayyed, and S. M. M. Qureshi. On the Indian side, the list is much longer, especially in the United States where—possibly because of political correctness—the universities have recruited an ever-larger number of Indian scholars who found the American environment increasingly attractive in terms of its intellectual climate as well as its working conditions.

Second, American and British universities have well established traditions of South Asian studies in political science. In England, while Oxford and Cambridge tended to specialize in history, the School of Oriental and African Studies, the London School of Economics, and the University of Sussex have developed an early interest in South Asian politics. They were joined subsequently by the Institute of Development Studies (Brighton), Hull University, and Bristol University. While the other European countries have been less active, with the possible exception of Sciences-Po (Paris), the Südasien Institut (Heidelberg), and the Amsterdam School of Social Science Research, American universities have played a major role in this field.

The University of Pennsylvania's Department of South Asia Regional Studies (Philadelphia) was founded in 1948 by W. Norman Brown. Brown was not a political scientist. He had became professor of Sanskrit at Pennsylvania in 1926 (and retired in 1966 in this capacity) but he was interested in Indian current affairs and trained political scientists, such as Richard Park who, in 1951 became the first director of the Berkeley South Asia Program and then of the Center for Southern Asian Studies in Michigan (Crane 1985). Nicholas Dirks points out that Norman Brown's life 'helps to explain why area studies at the University of Pennsylvania, and elsewhere, privileged a combination of classical Indological scholarship and

modern political and economic concern in the early history of the field' (Dirks in press). Pennsylvania University, where Francine Frankel developed the Center for the Advanced Studies of Asia in the 1990s, continued to play a leading role in addition to Berkeley and Michigan. In Chicago, the study of South Asia was dominated by anthropology—Milton Singer played a major part there in the 1950s—but political science asserted itself in the 1960s with Susan Hoeber Rudolph and Lloyd Rudolph. The Massachusetts Institute of Technology, where Myron Weiner was the moving spirit of South Asian Studies between 1961 and his death in 1999, was one of the other key institutions. Every year, American social scientists—including political scientists—meet at the University of Wisconsin (Madison) for the Annual Conference on South Asia, an event which is incomparably more important for political scientists than the biyearly meeting of the European Association for the Study of Modern South Asia, which is still in its infancy.

The American institutional base, the presence of many South Asian political scientists in the United States, and the strong influence of this country over political science—a young discipline in which there was no well-established European tradition—largely explain that the field was initially dominated by American-born methodologies and theories (Brass 1995).

2. *Approaches in the Study of South Asian Politics*

2.1 *From Developmentalism to Political Sociology*

The first researches into the politics of South Asia were developed after World War II within the then most influential framework of the theories of development and modernization. Myron Weiner, whose pioneering work, *Party Politics in India* was published in 1957, played a leading role in this enterprise. He 'represented' India, for instance, in 'classics' of this Princeton University-dominated school of thought, such as Almond and Coleman (1960) and Binder et al. (1971). His books—more especially *The Indian National Congress* (1967)—however, did not suffer from the limitations of this developmentalist paradigm because they relied on intensive fieldwork and an empiricist methodology.

Contemporaries of Weiner, including the Rudolphs and Morris-Jones, however, initiated a different approach. They rejected the theories that postulated the replication in South Asia of the political trajectories that had occurred in the West. W. H. Morris-Jones underlined the specificities of the Indian political cultures in his *The Government and Politics of India* (1964) and the Rudolphs made a similar argument by focusing on the case of Gandhi in *The Modernity of Tradition* (1967).

Both trends—Weiner's empiricism cum developmentalism and the school emphasizing the indigenous dimension of the modernization process—had an echo in India thanks to the towering figure of Rajni Kothari, the real founder of political science in India. In the 1960s, Kothari developed the notion of 'the Congress system' in an often quoted article to describe the political domination of the Congress party along lines that were very similar to Morris-Jones' analysis (Kothari 1964b, 1970a). Secondly, he took up the issue of modernization in a famous article (that was also published in an Anglo-Saxon journal, Kothari 1964a) that criticized the approach of the Rudolphs but pointed towards the same direction. Thirdly, Kothari worked with M. Weiner (Kothari and Weiner 1965). He drew some of his inspiration from the developmentalist approach—as evident from the name of the research institution he founded—The Centre for the Study of Developing Societies (CSDS), and was favorably inclined towards empiricism. As a result, the CSDS played a pioneering role in electoral studies; its Data Unit organized the first opinion polls in 1967 and 1971 under the guidance of Bashiruddin Ahmed who published with Samuel Eldersveld a book entitled *Citizens and Politics. Mass Political Behavior in India* (1978).

Right from the beginning, R. Kothari and his center introduced a *modus operandi* combining two major features: openness to the Anglo-Saxon approaches (stemming in collaborative works) and, at the same time, an attempt at Indianizing the concepts of analysis (this is most obvious in Kothari's 1970b study of the functional role of caste in the Indian democracy). While these two characteristics virtually contradicted each other, the contradiction disappeared when the American brand of developmentalist receded in the background and let the empiricist approach result in research works paying more attention to Indian political cultures and practices. M. Weiner himself challenged 'modernization' theory that predicted erosion of religious, caste, linguistic, and tribal identities, as developing countries modernized in *Sons of the Soil* (1978), a book he wrote with Mary Katzenstein.

Paul Brass' initial work exemplifies this approach. Brass' initial interest in the politics of Uttar Pradesh had led him to question the relevance of parties as the basic unit of the state Indian politics; he magisterially demonstrated that factions—partly related to castes—played a more important role in his book, *Factional Politics in an Indian State. The Congress Party in Uttar Pradesh* (1965). His approach prepared the ground for a political anthropology, which materialized gradually with the growing interest of anthropologists as Harold Gould in politics. However, Brass is best known for his study of ethnic politics. In his seminal work, *Language, Religion and Politics* (1974), he analyzed

'communal' movements in Punjab, Uttar Pradesh, and Bihar in an instrumentalist perspective, explaining that such identity politics results from the conflicts of interest of ethnic groups. This anti-primordialist paradigm was well in tune with the approach of a large number of Indian scholars.

2.2 Political Economy and the Marxist School

Besides the initial—and sustained—interaction as well as cross-fertilization relationship between Indian and Western—mostly American—political scientists, another characteristic of the study of South Asia politics lay in the strong emphasis on political economy. The main object of study here is the State and the issue at stake is: why did the State fail in its efforts at implementing its socialist agenda during the Nehru years and after? F. Frankel argues that the dependence of the Congress Party on privileged classes, landlords, capitalists, and other notables delivering votes or funds, prevented the government from pursuing its intended policies (1978). Marxist scholars, like C. P. Bhambri, offer a similar analysis but emphasize the weaknesses of class-consciousness among the poor, especially the peasantry. However, the Rudolphs (1978) and Atul Kohli (1990) suggest that the State has gained some autonomy.

Indian Marxists have often drawn their inspiration from Gramsci and his theories about social hegemony and passive revolution. Pranab Bardhan has used this approach to analyze Indian society's economic as well as political relations. In *The Political Economy of Development in India* (1984) he shows how three 'proprietary classes,' the political establishment, the landlords, and the capitalists share power in India, the first group making concessions to the others in order to remain in power: the Congress party gave up substantial land reform in order to get the support of rural notables at the time of elections and did not implement its socialist agenda fully, so much so that the business community gave financial support to the party.

Parallel to this class-oriented approach of social tensions, political scientists have more and more emphasized the role of caste. Eleanor Zelliot (1992) and Owen Lynch (1969) focused on the mobilization of the Untouchables (and more especially on the movement of B. R. Ambedkar). In the wake of these pioneering works, Marc Galanter (1984) analyzed the impact of the affirmative action programs on the lower castes, Gail Omvedt, the rise of Dalit politics (1994), and F. Frankel and M. S. A. Rao the 'decline of a social order' that was based on the traditional caste system (1989). The issue of caste, class, and power, originating from the 1960s works of the Marxist scholars, has remained central to the study of Indian politics not only because the State has failed in its endeavor to bring the benefits of development to the masses but also because of the interaction between economic, social, and political power in India.

2.3 New Developments in 'Domestic Political Science'

In the 1990s, the development of communalism and the increased political use of religion India was witnessing, led many authors to scrutinize a subject that had become a non-issue: secularism. In an attempt to understand the polarization of communities along religious lines, the real impact of secularism on Indian society was reexamined by sociologists as well as political scientists (Nandy 1985, Bhargava 1996, 1998). The State was also a central focus of the research undertaken, the question being why it had failed to anchor the secularist agenda in Indian society (Basu and Kohli 1998). The historical and the Marxist traditions again played an important encompassing part in delimiting the approach. This can be seen in the works of Ashgar Ali Engineer, for instance, whose interpretation of communalism as the result of economic rivalry (1984) had a large scientific impact. It is also evident as prominent historians like Gyanendra Pandey had a large share of the study of communalism from the late 1980s onwards (1993).

Running parallel to the study of communalism was the new examination of the political system to account for the rise of the parties relying on religious identities for support. Bruce Graham's work on the Bharatiya Jana Sangh, covered both the fields of the study of opposition politics (organization, support bases, etc.) and that of the political use of religion (1990). Generally speaking, however, the study of the political parties has been one of the declining themes. While the Congress, when it was a ruling party, had been the central focus of works by Kothari, Morris-Jones, Weiner, Brass, and Stanley Kochanek (1968), its decline has not been studied very closely. Similarly, the communist parties, the socialist movement, and the peasant-based parties such as the Lok Dal have remained largely ignored. The State institutions, which were scrutinized when they were set up after independence Austin 1966, did not generate the same amount of work in the 1990s. Scholars like James Manor (who edited a book on the Indian Prime Minister), however, argue that institutional decay is not an irreversible trend and that the resurgence of the judiciary and the Election Commission, for instance, are a good reflection of this tendency (1994).

Similarly, State politics has not continued to attract the same interest. In the 1960s–70s Myron Weiner and Iqbal Narain edited volumes studying State politics, almost state by state (1974–7). Paradoxically, such comparative studies have become less and less numerous (a notable exception being Zoya Hasan's 1998

book on Uttar Pradesh), precisely at the moment when the dismantlement of the one-party system favors the emergence of an even larger number of regional parties. In the 1990s, the regionalization of politics resulted in the multiplication of political parties in Parliament. There were about 40 in the late 1990s.

On the other hand, election studies gained momentum. They have always been a very active sector of political science in, or about, India. Myron Weiner, once again, published several volumes untitled *India at the Polls*. Before him, R. Park and S. Kogekar had covered the first general elections (1956) and Iqbal Narain had coordinated similar surveys (Varma and Narain 1968). In a manner very different from the early years of Indian political science, the political assertion of new actors—including low castes, communal forces, and regional parties—and the tremendous transformations of the party system it triggered is now also being assessed through empirical studies of the voting pattern including exit polls. The Centre for the Study of Developing Societies again plays an essential part. Its Data Unit under the direction of Yogendra Yadav conceives and conducts nationwide polls before and after most elections since 1993. Focusing on socioeconomic criteria, they provide an exceptional corpus of material for the understanding of Indian political behavior. Yogendra Yadav's articles on the 1993–5 elections have cleared the ground for the analysis of post-Congress politics.

The current revival of Indian political science is driven by the analysis of the major and somewhat brutal changes undergone by Indian society since the late 1980s. One of the consequences is the relative decline of the historical tradition and the assertion of a more 'political' school of analysis. The main English-medium scientific press, the Oxford University Press, had launched a new series in 1997 that could be seen as an indication of the change taking place. This 'Themes in Politics' series co-edited by Rajiv Bhargava and Partha Chatterjee explored three subjects: the State, secularism, and gender, the three having been at the core of the then evolution of Indian society. The series shifted in 2000 to a new publishing house, Permanent Black, founded by Rukund Advani, a former senior commissioning editor with OUP.

2.4 International Relations: Nonalignment and Indian Neighbors

In the early 1960s, the study of international relations ran parallel with the setting of independent India's foreign policy and focused chiefly on the expectations rising from the Nonaligned Movement and the part India played in it. From the 1970s onwards, the development of strong links with the USSR provided a new focus of research. Soviet foreign policy became one of the main themes of research, in particular in relation to the USA and Eastern European countries. It therefore had a strong regional taste to it while conceptual issues were less prominent.

While it focused on countries with which India was having close interactions, Indian political science hardly covered countries with which there were few links. The case of the Chinese school of study provides an interesting example of the close relationship between research and India's foreign policy. With the 1962 war between India and China, the good relationship the two countries had enjoyed until the mid-1950s collapsed completely. With diplomatic relations down to the minimum, too little scientific research was encouraged either on China's foreign policy or later on the Chinese way to economic development. The Institute of Chinese Studies started publishing *China Report* in 1964, but it is rather later that China's evolution into a major player on the international scene helped to realize India's lack of knowledge of its chief neighbor.

Western Europe and the USA were not central to the study of international relations in India, and were chiefly investigated through their links to the USSR. Central Asia, in spite of its proximity to India, was also hardly studied at all. The same goes also for East and Southeast Asia. The new dynamism of research on the region owes a lot to India's 'Look East policy' since the early 1990s and the growing economic links between India and the various Southeast Asian countries.

Apart from the interest in regional studies, Indian schools of international relations also developed strategic analysis. Partly due to the country's size and its tendency to look inward, strategic culture had eluded India. With India's growth on the South Asian regional scene and its claims to being a major international player, a school of strategic thinking started to emerge.

The Institute for Defence Studies and Analysis (IDSA) played an essential part in this new school of thought. The center was founded in 1965 and has since focused on India's national security and orientations to defense. Through its long established influence on the Indian government, the IDSA has been instrumental in establishing India's naval and regional policy. The Centre's main publications are the journals *Strategic Analysis* and *Strategic Digest*, and the yearly *Asian Strategic Review*. The latter has been conceived as a reply to the Stockholm Institute for Peace Research's *Yearbook* and London's International Institute for Strategic Studies' *Military Balance*. Arguing against the Western perception that India's military policy was a risk to regional stability, the IDSA has argued in favor of the nuclear deterrence and of the necessity of developing a security establishment of adequate size. Its former director, K. Subrahmanyam, and director in 1999 Jasjit Singh, wield considerable influence over the Indian government as far as security issues are concerned.

Indicating the change of scale in the Indian current foreign policy, the relation to China, the nuclear doctrine, and the question of 'multipolarity' are nowadays central issues in international relation studies. The IDSA and the Delhi-based Centre for Policy Research concentrate on those subjects which can be seen as a major shift in Indian strategic thinking.

2.5 How to Study the Politics of Pakistan?

The study of Pakistan's politics, which is much less developed than its Indian counterpart has been dominated by two themes, the question of the regime and the issue of national unity. Historically, the latter came first: Pakistan was created to give a nation-state to the Muslims of British India but Islam did not turn out to be a sufficiently strong cementing force with East Pakistan fighting—successfully—for the establishment of a separate country (Bangladesh), the Pathans showing signs of irredentism vis-à-vis Afghanistan, the Baluchis displaying centrifugal forces too, and Sindh being gradually the framework for a major conflict between Mohajirs and Sindhis. Hence there are a large number of studies dealing with ethnic issues. Some authors focus on language as the main line of cleavage (Rahman 1998), while others (Amin 1988) look at ethnic groups (with their other characteristics such as religious affiliations). But most of the authors reject the primordial paradigm. They rather interpret ethnic conflicts as the outcome of socio-economic tensions or political rivalries. However, emphasizing that cultural features are instrumentalized for promoting collective interests does not mean that culture is ignored. This reading applies to studies of religious groups too. For instance, Vali Nasr underlines that the Jama'at-i-Islami had some affinities with Sufi orders (1994).

The question of the regime has been looked at very closely after the first military coup terminating Pakistan's first attempt at establishing democratic institutions. In his 'classic,' *The Political System of Pakistan*, Khalid B. Sayeed coined the expression 'constitutional autocracy' (1967) to describe Ayub Khan's regime. Pakistan's incapacity to escape the cyclical alternation of civil rule and praetorian politics is also central to another classic by Mohammad Waseem, *Politics and the State in Pakistan* (1994). Unsurprisingly, the periods of military rule have been studied in greater details. Ayesha Jalal and Mohammad Waseem have scrutinized respectively the reigns of Ayub Khan and that of Zia-ul-Haq in two complementary perspectives. While the former emphasized the political economy of defense (1990), the latter put a stress on the authoritarian overtone of the political culture of Pakistan—the viceregal legacy of colonial rule—and its affinities with the ethos of the army (1987).

Pakistani political scientists however often point out that the army's domination over society and politics—even after 1988 when the military personnel continued to exert a strong influence directly or via the President—is not in tune with the country's political culture. Saeed Shafqat refers to 'a case of disharmony between democratic creed and autocratic reality' (1998).

Political science in and about South Asia shares some features irrespective of the country one looks at. On the one hand, in most cases it is derived from the British tradition and this starting point generated rather strong affinities with the Anglo-Saxon academic world. On the other hand, India was especially studied by political scientists (either in India or abroad) from the point of view of its domestic scene whereas scholars working on Pakistan tended to pay more attention to its relations with neighbors and other countries. This differentiation is evident from a clear unevenness in terms of institutional developments. This contrast can be explained by the size of India–a world in itself—and its unique, democratic trajectory, on the one hand, and by the extraversion of Pakistan, a country where political surveys have never been easy, especially when it was under military rule, on the other hand. The alternation of civil governments and military coups probably explains the rather limited development of political science in Bangladesh, the only country of this area we have not mentioned yet. However, the universities of Dhaka, Chittagong, and Jahangirnagar have departments of political science and the resumption of the democratization process may contribute to their consolidation.

Bibliography

Ahmed B, Eldersveld S 1978 *Citizens and Politics. Mass Political Behavior in India.* Chicago, IL

Almond G A, Coleman J S (eds.) 1960 *The Politics of the Developing Areas.* Princeton, NJ

Amin T 1988 *Ethno-national Movements of Pakistan. Domestic and International Factors.* Institute of Policy Studies, Islamabad, Pakistan

Austin G 1966 *The Indian Constitution: Cornerstone of a Nation.* Clarendon Press, Oxford, UK

Bardhan P 1984 *The Political Economy of Development in India.* Delhi, India

Basu A, Kohli A (eds.) 1998 *Community and the State in South Asia.* Oxford University Press

Bhargava R 1996 How not to defend secularism. In: Bidwai P, Mukhia H, Vanaik A (eds.) *Religion, Religiosity and Communalism.* Manohar, Delhi, India

Bhargava R (ed.) 1998 *Secularism and His Critics.* Oxford University Press, Delhi, India

Binder L et al. 1971 *Crises and Sequences in Political Development.* Princeton, NJ

Bose S 1994 *States, Nations, Sovereignty. Sri Lanka, India and the Tamil Eelam Movement.* Sage, New Delhi, India

Brass P 1965 *Factional Politics in an Indian State. The Congress Party in Uttar Pradesh.* Berkeley, CA

Brass P 1974 *Language, Religion and Politics.* Cambridge

Brass P 1995 American Political Science and South Asian Studies: Virtue unrewarded. *Economic and Political Weekly* Sept.: 2257–62
Chaudhury G W 1959 *Constitutional Development in Pakistan*. London
Chaudhury G W 1963 *Democracy in Pakistan*. Dacca, Bangladesh
Chaudhury G W 1974 *The Last Days of Pakistan*. Bloomington, IN
Crane R 1985 Preface on Richard L. Park. In: Wallace P (ed.) *Region and Nation in India*. IBH. Publishing and Co., New Delhi, India
Dirks N B in press South Asian Studies: *Futures Past*
Engineer A A 1984 *Communal Riots in Post-Independence India*. Delhi, India
Frankel F 1978 *India's Political Economy, 1947–1977*. Princeton University Press, Princeton, NJ
Frankel F, Rao M S A (eds.) 1989 *Dominance and State Power in Modern India, 2 Vols*. Oxford University Press
Galanter M 1984 *Competing Equalities. Law and the Backward Classes in India*. Oxford University Press, Delhi, India
Graham B D 1990 *Hindu Nationalism and Indian Politics: The Origins and Development of the Bharatiya Jana Sangh*. Cambridge University Press, Cambridge, UK
Hasan Z 1998 *Quest for Power. Oppositional Movements and Post-Congress Politics in Uttar Pradesh*. Oxford University Press, Delhi, India
Jalal A 1990 *The State of Martial Rule. The Origins of Pakistan's Political Economy of Defence*. Cambridge University Press, Cambridge, UK
Kochanek S 1968 *The Congress Party of India*. Princeton University Press, Princeton, NJ
Kogekar S V, Park R L (eds.) 1956 *Reports on the Indian General Elections. 1951–52*. Popular Book Depot, Bombay, India
Kohli A 1990 *Democracy and Discontent: India's Growing Crisis of Governability*. Cambridge,
Kothari R 1964a *Tradition and Modernity Revisited. Government and Opposition* Summer
Kothari R 1964b The Congress System. *Asian Survey* (Dec.): 116–73
Kothari R 1970a *Politics in India*. Orient Longmans, Hyderabad, India
Kothari R (ed.) 1970b *Caste in Indian Politics*. Orient Longman, New Delhi, India
Kothari R Weiner M 1965 *Indian Voting Behavior*. Calcutta, India
Lynch O 1969 *The Politics of Untouchability*. Columbia University Press, New York
Manor J 1994 *Nehru to the Nineties. The Changing Office of Prime Minister in India*. Viking, Delhi, India
Morris-Jones W H 1964 *The Government and Politics of India*. London
Nandy A 1985 An anti-secular manifesto. *Seminar* Oct.
Nasr S V R 1994 *The Vanguard of the Islamic Revolution. The Jama'at-i Islami of Pakistan*. University of California Press, Berkeley, CA
Omvedt G 1994 *Dalits and the Democratic Revolution*. Sage, New Delhi, India
Panandikar V A P, S 1983 *A Changing Political Representation in India*. Uppal, New Delhi, India
Panandikar V A P, S 1992 *Integration of Population, Environment and Development Policies*. CPR, New Delhi, India
Pandey G (ed.) 1993 *Hindus and Others*. Viking, Delhi, India
Rahman T 1998 *Language and Politics in Pakistan*. Oxford University Press, Karachi, Pakistan
Rudolph L I, Rudolph S H 1967 *The Modernity of Tradition*. Chicago, IL
Rudolph L I, Rudolph S H 1978 *In Pursuit of Lakshmi: The Political Economy of the Indian State*. Chicago, IL
Sayeed K B 1967 *The Political System of Pakistan*. Houghton Mifflin, Boston, MA
Shafqat S 1989 Political science problems, prospects and scope in Pakistan. In: Hashimi S H (ed.) *The State of Social Sciences in Pakistan*, Quaid-i-Azam University, Islamabad, Pakistan, pp. 159–68
Shafqat S 1998 Political culture of Pakistan: A case of disharmony between democratic creed and autocratic reality. In: Shafqat S (ed.) *Contemporary Issues in Pakistan Studies*, Azad, Lahore, Pakistan
Tambiah S J 1986 *Sri Lanka: Ethnic Fratricide and the Dismantling of Democracy*. Oxford University Press, Delhi, India
Varma S P, Narain I (eds.) 1968 *Fourth General Elections in India*. Orient Longmans, Bombay, India
Wasem M 1987 *Pakistan Under Martial Law, 1977–1985*. Vanguard, Lahore, Pakistan
Waseem M 1994 *Politics and the State in Pakistan*. Islamabad, Pakistan
Weiner M 1957 *Party Politics in India*. Princeton, NJ
Weiner M 1967 *The Indian National Congress*. Chicago, IL
Weiner M, Field O J (eds.) 1974–77 *Electoral Politics in the Indian States. 4 Vols*. Manohar, Delhi, India
Weiner M, Katzenstein M 1978 *Sons of the Soil*. Princeton, NJ
Wilson A J 1974 *Politics in Sri Lanka, 1947–1973*.
Wilson A J 1988 *The Break-up of Sri Lanka*. Hurst, London
Yadav Y Reconfiguration in Indian Politics. State Assembly Elections, 1993–1995. *Economic and Political Weekly*
Zelliot E 1992 *From Untouchable to Dalit. Essays on Ambedkar Movement*. Manohar, Delhi, India

C. Jaffrelot and J. Zerinini-Brotel

Southeast Asia, Archaeology of

Southeast Asia, a term which entered into common use and academic discourse during and after World War II, refers to the region covered by the modern nation-states of Burma, Thailand, Laos, Cambodia, Vietnam, Malaysia, Singapore, Indonesia, East Timor, Brunei, and the Philippines (although archaeologists also include southern China and Taiwan in studying aspects of the region's prehistory). Divided into mainland and island areas, this region extends across tropical latitudes for approximately 5,000 km southeastwards from Burma to eastern Indonesia.

Scholarly traditions of archaeological research in Southeast Asia are generally divided according to their focus on either prehistoric or historic periods. Prehistory begins with the Pleistocene artifactual record of early hominids and ends with the appearance of written records, the timing of which varies by area (e.g., in Indonesia stone inscriptions in Sanskrit appear in the fourth century AD whereas extant written

accounts of the Philippines date from the tenth century AD onward). Historical archaeology is subdivided into, for example, early, classical, Islamic, and protohistoric, depending upon the area. While the archaeology of European colonialism and the spread of capitalism are often used to define historical archaeology elsewhere in the world, in Southeast Asia these constitute more recent topics of research. European colonialism of Southeast Asia, however, was the context within which archaeology developed, and its legacy continues to affect contemporary archaeological practice. This history and the current understanding of Southeast Asia's prehistoric sequence (as based on English-language sources and primarily the better known archaeological records, at least for certain periods, of Thailand, Vietnam, Malaysia, Indonesia, and the Philippines) are summarized here, followed by a discussion of some of the problems and future directions in the archaeology of this region.

1. *Development of Southeast Asian Archaeology*

The study of archaeological remains in Southeast Asia has a long history, predating even that conducted by Europeans who are credited with introducing archaeology. Thai and Burmese rulers, for example, had previously initiated studies of stone inscriptions to reconstruct political-economic aspects of earlier states. The practice of archaeology as a Western scientific discipline began with European colonial expansion into Southeast Asia. Following an initial period of antiquarian interest, which in places such as Indonesia lasted for more than a century, during which Dutch and British recorded monumental sites such as Prambanan, collected artifacts (including specifically for museums), and published descriptions, the first archaeological institutions were established. In Burma, a research officer in archaeology was appointed in 1881 under the Indian Archaeology Department established by the British; in Indonesia, the Dutch established an archaeological society in 1885 to coordinate scientific exploration of ancient monuments in Java; in Indo-China (Cambodia, Laos, Vietnam), the French established the École Française d'Extrême Orient in 1898, which directed excavations, recorded and restored sites, and published reports; in Thailand, the Siam Society was founded in 1904 as a forum for those interested in the country's past, followed by the establishment of the Archaeological Service of Siam in 1924; in the Philippines, systematic explorations and collections were initiated in 1881, with archaeology established during the US colonial period; and in Malaysia, archaeological excavations began in the early 1930s conducted by the Raffles Museum in Singapore. Early investigations tended to focus on monumental sites, often in association with the study of the growing body of stone inscriptions, where these existed; however, excavations of prehistoric sites were also undertaken. Unlike Thailand, where from the beginning archaeological and art historical research was conducted by Thais, in the colonies this was not the case; thus, for example, Vietnamese, Khmer, and Laotians were neither trained in archaeology nor participated as professionals until after independence.

Beginning in the 1960s, there was a significant rise in archaeological research focused on prehistoric sites (although this did not happen everywhere; for example, no archaeological research was conducted in Laos between 1945 and 1975). Some were joint projects between foreign and local institutions, which included training of local students (a first in some countries such as Thailand). One outcome of the various research projects, some of which claimed early developments of agriculture and metallurgy, was that by the 1970s Southeast Asian archaeology was no longer viewed as a branch of Indian, Chinese, or Far Eastern studies, nor as a marginal zone in prehistory that was a backward appendage of the more advanced cultures of India and China. At the same time, diffusion and migration explanations of culture change began to be replaced by frameworks emphasizing local processes of innovation and adaptation. Also since the 1970s, the majority of archaeological research in the region has been conducted by Southeast Asian archaeologists (Miksic 1995) employing a range of perspectives from culture-historical to Marxist.

2. *From* Homo erectus *to Early States*

2.1 *Early Hominid Occupation of Southeast Asia*

The earliest known hominid occupation of Southeast Asia was by *Homo erectus*. It occurred by at least one million years ago, although redating of fossil-bearing deposits in Java suggests it may have been as early as 1.8 million years ago (Swisher et al. 1994). The majority of evidence is paleontological and primarily from sites on Java, although *Homo erectus* fossils are also known from Burma and Vietnam. Stone tool assemblages, composed of large pebble and flake tools (which differ from the contemporary Acheulian hand ax industries of Africa, Europe, and India: see *Sub-Saharan Africa, Archaeology of*), and generally not from primary contexts, have not yet been found in association with hominid fossils. The interpretation of much of the archaeological record assigned Middle Pleistocene dates thus continues to be the subject of debate. *Homo erectus*'s technical achievements may have included maritime technology, as suggested by the recovery of lithic artifacts from excavated sites on Flores dated to the Early Pleistocene (c. 800,000 year ago). Occupation of this island would have entailed crossing deep sea channels since it was not part of Sundaland (the land mass connecting the mainland with Sumatra, Java, Bali, Borneo, and Palawan, exposed at various times during the Pleistocene) and suggests *Homo erectus* possessed voyaging skills and watercraft (Morwood et al. 1999).

2.2 Late Pleistocene–Early Holocene Hunter-gatherers

The oldest evidence for occupation by anatomically modern humans dates to approximately 40,000 years ago. Only a few archaeological sites provide evidence of Late Pleistocene occupation: Lang Rongrien rock shelter, Lang Kamnan cave, Son Vi sites, Niah Cave, Tabon Cave, and Leang Burung 2 (see, for example, Anderson 1990, Ha Van Tan 1997, Shoocongdej 2000). The hunter-gatherer groups who occupied these caves and rock shelters are known primarily from their flake and pebble/cobble tool assemblages, although hearths and remains of hunted and collected animals (e.g., land tortoises, rodents, wild pig, monkeys, orangutan, shells) are also found. The majority of their material culture was probably made of bamboo and wood, as suggested by the generalized nature of the stone tools and their probable use in manufacturing activities.

By the early Holocene, hunter-gatherers appear to have occupied much of Southeast Asia, with some areas, such as the Malay Peninsula, reoccupied following the expansion of the postglacial equatorial rainforest. Occupation of the upland areas of the mainland is better known in contrast to the lowlands, where rising sea levels and subsequent landscape changes after the last glacial maximum destroyed or buried most sites. Holocene hunter-gatherers are known primarily from their stone tool assemblages, which are dominated by pebble tools. Other types of tool included grinding stones, bone points, and antler wedges, although the majority of their material culture was presumably made of perishable materials. Caves and rock shelters were used for occupation as well as for burial; individuals were buried in either flexed or extended positions and accompanied by stone artifacts, animal bones, and red ochre. Coastal areas were also occupied/exploited, as indicated by shell midden sites. Mobility strategies may have varied with the wet and dry seasons, entailing different types of settlement (including in open areas) and exploitation of different food resources (Shoocongdej 2000). Diet varied by location and included a wide variety of terrestrial (e.g., pig, deer, macaque, rhinoceros), riverine and marine shell species. Holocene hunter-gatherers also exploited a wide range of plants; their role in plant domestication, a topic of long-standing interest in Southeast Asia, remains little known. In part, this reflects current emphasis on rice—the subsistence base of later complex societies—especially the timing of its domestication, adoption and spread. (See *Hunter–Gatherer Societies, Archaeology of*.)

2.3 Early Rice-cultivating Communities

The earliest known domesticated rice comes from middle Yangtzi valley sites and dates to c. 6000 BC (see *Food Production, Origins of*); Southeast Asian sites have not yet yielded direct evidence of rice this early. Consequently, this evidence in conjunction with linguistic data, underlies modern models (e.g., Bellwood 1997, Higham 1996) which propose that in the third millennium BC early rice agriculture was introduced into mainland Southeast Asia by expanding Austroasiatic-speaking rice-farming communities that originated in the Yangtzi river area, and into island Southeast Asia by expanding Austronesian-speaking communities from Taiwan or southern China (see *Pacific Islands, Archaeology of*).

Archaeologically, early mainland rice-farming communities appear in northern Vietnam in the third millennium BC. Village settlements with associated cemeteries (of extended burials and infant-jar burials) were established in inland areas, and were engaged in rice cultivation, the raising of domesticated animals (dogs, cattle, pigs), the manufacture of decorated pottery, and the use of stone ornaments. In northeastern and central Thailand, farming villages are known from habitation remains and primarily from cemeteries of extended burials; some mortuary evidence suggests social status differences based on achievements in fine pottery crafting and exchange activities (Higham 1996). These contexts have yielded decorated pottery (some tempered with rice chaff), polished stone adzes, marine shell beads and bracelets, stone reaping knives, bone tools, spindle whorls, and remains of domesticated cattle, pigs, chickens, and dogs. These mainland communities also continued to hunt and collect wild resources, and they engaged in regional systems of exchange involving stone and marine shell. Interestingly, in northern coastal Vietnam, shell middens and other sites dated from the fifth millennium BC have yielded pottery and polished stone adzes, material culture often associated elsewhere with 'Neolithic' communities. Similarly, in Thailand, pottery (appearing as early as 6000 BC) and stone adzes, often from cave burials, predate evidence considered indicative of rice-farming villages. Whether such evidence indicates 'affluent foragers' or early farming activities of rice or other plants remains unknown.

Within the islands and the Malay Peninsula, early rice-farming communities are known almost entirely from occupation and burial remains in caves and rock shelters, along with a few shell midden sites (Bellwood 1997). In the Philippines, by the early third millennium BC, such sites contain various items including pottery, stone and shell adzes, spindle whorls, stone and shell jewelry, and tattooing chisels. Within the Indonesian archipelago, pollen evidence suggests minor forest clearance on Sumatra around 4500 BC and permanent clearance after 1000 BC, presumably for agriculture; other islands (again primarily from cave and rock shelter sites) provide evidence of pottery and rice around 2500 BC, and the human introduction (via watercraft) of pigs and nondomesticated animals

between 2500 and 2000 BC. During the second millennium BC, caves along the Malay Peninsula were used for burial rather than habitation, and contain graves of extended individuals accompanied by items such as stone adzes, bone points, shell beads and bracelets, and pottery.

2.4 Early Metal Crafting

By the early to mid-second millennium BC, autonomous village communities on the mainland (known primarily from mortuary deposits) became involved in the production and consumption of copper/bronze craft goods, with stone and pottery crafting also continuing (see *Intensification and Specialization, Archaeology of*). This technological change may not have been accompanied by changes in sociopolitical complexity, unlike Bronze Age periods elsewhere in the world. Bronze objects appear in sites in northern Vietnam prior to 1500 BC, in association with exotic jades from China. Sites in central and northeast Thailand, with initial use in the early mid-second millennium BC, provide the earliest evidence for the mining, crushing, smelting, and casting of copper into ingots (e.g., Pigott and Natapintu 1988), activities which appear not to have required full-time specialists (White and Pigott 1996). After 1500 BC, bronze-casting technology, which involved the use of ceramic crucibles and stone or ceramic molds, appears to have spread quickly along riverine and coastal exchange routes (Higham 1996). Copper ingots entered into existing exchange networks, thus enabling communities situated far from ore sources to engage in bronze casting. Prior to production of iron craft goods (c. 500 BC), bronzes were cast into ornaments (bracelets, ear ornaments) and utilitarian objects (e.g., socketed ax heads and other implements, chisels, arrow and spear points, fishhooks, sickles). Bronze objects and copper ingots (along with other items such as pottery vessels, animal remains, stone tools, and shell ornaments) served as grave accompaniments, although whether their value in mortuary contexts related to distinguishing status differences and/or other social identities remains unknown.

2.5 The Development of Chiefly Polities and Early States

Around the mid-first millennium BC, significant social changes entailing systems of hereditary ranking and hierarchical organization appear to have been under way, which archaeologists interpret as indicative of complex societies such as chiefdoms (see *Chiefdoms, Archaeology of*). In mainland Southeast Asia, larger settlements (some with moats, particularly in northeastern Thailand) integrated into regional hierarchies appeared, as well as urban settlements (Co Loa in northern Vietnam and Beikthano in Burma). Other changes, occurring variably within the region, included the development of iron production (of, for example, ornaments and tools in localized forms), with smelting and forging first appearing in the Chao Phraya and middle to lower Mekong river valleys; increased scale of production and variety of bronzes crafted (e.g., various vessel shapes, bells, large drums, axes, spearheads, jewelry), as exemplified by Dong Son bronzes; greater use of bronzes in mortuary ritual; glass working (in southern Vietnam); intensification of wet rice cultivation using water buffalo and iron plows; participation in maritime and long-distance exchange relations between the mainland and island areas (Bellwood 1997) and India (the earliest evidence dating to the fourth century BC; see, for example, Glover 1996), with exotic goods appearing in mortuary contexts; and the incorporation of the Dong Son polity (northern Vietnam) into the Han empire in 111 BC.

The complex societies that were developing during the last few centuries BC in areas of island Southeast Asia acquired bronze prestige/status items (such as Dong Son drums) and iron goods via exchange with the mainland, which was soon followed by the establishment of local metalworking centers. Island polities also engaged in exchange networks involving China and India, with Bali yielding the earliest dated Indian materials at around two thousand years ago (Ardika and Bellwood 1991). As with the mainland, the majority of information on island societies assigned to an 'early metal phase' (Bellwood 1997) derives from burials, in open sites and caves, including the distinctive jar burial, slab grave (Sumatra, Java and Peninsular Malaysia), and carved stone sarcophagi (Java and Bali) forms of interment. While these sites have yielded a wealth of material culture (e.g., decorated pottery; bronze socketed tools, bowls, drums; glass and carnelian beads; iron tools) and information on mortuary rituals, much remains to be learned of the social, political, economic, and ideological aspects of this period, which preceded the formation of state polities in various areas of the Indo-Malaysian archipelago.

States in Southeast Asia are traditionally thought to have appeared on the mainland during the early centuries AD. This dating reflects, in part, the privileging of written sources in the identification of state-level polities, and is exemplified by the case of 'Funan,' which is generally viewed as the earliest state. Written accounts, the earliest dating to the third century AD by visiting Chinese emissaries, describe Funan's origin story, dynastic lineage, raiding activities, vassal polities, the walled settlements which contained a palace, the taxation system (paid in gold, silver, perfumes, and pearls), and the presence of a representative of the Indian Murunda king, among other things. Oc Eo, a 450-hectare rectangular site (enclosed by several ramparts and moats) of approximately the second to sixth

century, is located in the lower Mekong valley, the area in which 'Funan' is thought to have been situated. It has yielded evidence of participation in extensive exchange networks involving Iranian, Roman, Indian (some engraved in Indian script), and Chinese goods from the second century AD onwards; and of glass, stone (precious and semiprecious), and metalcraft production (Higham 1996). An extensive system of canals (not yet dated) connects Oc Eo with several other large sites including Ta Keo (the probable port) and Angkor Borei. The latter was continuously occupied from the late centuries BC onwards, and was contemporary with and shared similarities in material culture to Oc Eo, which suggests the two were part of the same polity (Stark et al. 1999). The lower Mekong delta state, however, was not an isolated polity in its early years, as other contemporary maritime-oriented states appear to have existed on the west coast of the Malay Peninsula and in west Java (Wisseman 1995).

From the mid-first millennium AD, more historic Southeast Asian states are known (e.g., Chenla, Dvaravati, Champa, Kedah, Srivijaya), and these exhibit a shared incorporation of Indian legal, political, and religious ideas and institutions, including the use of Sanskrit names by rulers, as seen in stone inscriptions (first in south Indian and then indigenous scripts), and in the layout and styles of religious architecture and carvings. While 'Indianization' (the view that statecraft was brought to Southeast Asia by Indians) is no longer accepted, relatively little is known about indigenous sociopolitical processes in polities which immediately preceded the early historical states, including whether they already represented emerging states.

State polities may have existed earlier in northern Vietnam, prior to Han expansion, and this is an issue which requires further consideration and investigation. Several research projects aimed at understanding state development on the mainland have been initiated in central coastal Vietnam, the area of the Cham state (Glover and Yamagata 1995); the lower Mekong delta, an area associated with 'Funan' (Stark et al. 1999); and northeastern Thailand (Higham and Thosarat 2000). Throughout Southeast Asia, fundamental theoretical and empirical issues related to state formation remain, including the definition of a 'state,' the relevance of Western theoretical conceptualizations of states and of indigenous political theory, the dating of the earliest states, and the interplay of ideological, social, economic, and other factors in the transformation to states (see *States and Civilizations, Archaeology of*).

3. *Issues of Culture and Chronology*

'Diversity' is often used to characterize Southeast Asia (e.g., linguistically, ethnically, religiously), and this certainly applies to the premodern and prehistoric periods, in which, for example, at any point in time during the last two thousand plus years there existed various social formations from small, mobile hunter-gatherer groups to large, regional state polities. Consequently, the application of cultural frameworks derived from elsewhere in the world has been criticized, and attempts have been made to develop locally specific ones (e.g., Bayard 1984). Nevertheless, the challenge remains to develop cultural-chronological frameworks that do not rely on assumptions of simultaneous 'progressive' regional transformations. Related to this problem is the paucity of well-dated regional archaeological sequences, and the continued reliance on a few 'key' sites or artifact 'types' for dating, interpreting, and explaining Southeast Asia's pre- and early history; both reflect a need for more institutional and economic resources for archaeology in this region.

4. *Future Research Directions of Southeast Asian and Western Archaeologists*

The establishment of chronometrically anchored local sequences will continue to be an aim of much of the archaeological research conducted in the near future. Within the community of Southeast Asian archaeologists, future directions (as expressed at the first International Colloquium on Archaeology in Southeast Asia in the 3rd Millennium, 1998) include more collaborative projects and comparative research; development of theoretical- and problem-oriented research; theoretical contributions to world archaeology; the application of anthropological (also in combination with historical) approaches to archaeological investigation and interpretation; regional as opposed to single-site investigations; and analysis of entire assemblages (particularly in the case of stone tool studies). In addition, Southeast Asian and Western archaeologists will continue research on issues now beginning to be raised, such as local processes of domestication and plant cultivation; the impact and contesting of European colonialism by various social groups; pre- and early historic gender systems and ethnicity; the role of foreign religions (Hinduism, Buddhism) in indigenous social processes; and how nationalism and colonialism have impacted archaeological practice in Southeast Asia.

See also: Area and International Studies: Development in Southeast Asia; Southeast Asian Studies: Geography

Bibliography

Anderson D 1990 *Lang Rongrien Rockshelter: A Pleistocene-Early Holocene Archaeological Site from Krabi, Southwestern Thailand.* University of Pennsylvania Museum, Philadelphia

Ardika I W, Bellwood P 1991 Sembiran: The beginnings of Indian contact with Bali. *Antiquity* **65**: 221–32
Bacus E A, Lucero L J (eds.) 1999 *Complex Polities in the Ancient Tropical World*. American Anthropological Association, Arlington, VA
Bacus E A 2000 *World Archaeology* Archaeology in Southeast Asia **32**(1) (special issue)
Bayard D 1984 A tentative regional phase chronology far Northeast Thailand. In: Bayard D (ed.) *Southeast Asian Archaeology at the XV Pacific Science Congress*. Department of Anthropology. University of Otago, Dunedin, pp. 161–8
Bellwood P 1997 *Prehistory of the Indo-Malaysian Archipelago*. University of Hawaii Press, Honolulu, HI
Glover I 1996 Recent archaeological evidence for early maritime contacts between India and Southeast Asia. In: Ray H P, Salles J F (eds.) *Tradition and Archaeology: Early Maritime Contacts in the Indian Ocean*. Manohar Publishers, New Delhi, India, pp. 129–58
Glover I, Yamagato M 1995 The origins of Cham civilization: Indigenous Chinese and Indian influences in central Vietnam as revealed by excavations at Tra Kieu, Vietnam, 1990 and 1995. In: Yeung C, Li H-L (eds.) *Conference Papers on Archaeology in Southeast Asia* (*1995: Hong Kong*). Hsiang-Kang ta hsueh mei shu po wu kwan, Hsiang-Kang
Ha Van Tan 1997 The Hoabinhian and before. In: Bellwood P, Tillotson D, Soejono R P, Welch D (eds.) *Indo Pacific Prehistory: The Chiang Mai Papers Vol. 3, Bulletin of the Indo-Pacific Prehistory Association* **16**: 35–41
Higham C 1989 *The Archaeology of Mainland Southeast Asia*. Cambridge University Press, Cambridge, UK
Higham C 1996 *The Bronze Age of Southeast Asia*. Cambridge University Press, Cambridge, UK
Higham C, Thosarat R 2000 The origins of the civilization of Angkor. *Antiquity* **74**: 27–8
Junker L L 1999 *Raiding, Trading, and Feasting: The Political Economy of Philippine Chiefdoms*. University of Hawaii Press, Honolulu, HI
Miksic J N 1995 Evolving archaeological perspectives on Southeast Asia, 1970–95. *Journal of Southeast Asian Studies* **26**: 46–62
Morwood M J, Aziz F, O'Sullivan P, Nasruddin, Hobbs D R, Raza A 1999 Archaeological and palaeontological research in central Flores, east Indonesia: Results of fieldwork 1997–98. *Antiquity* **73**: 273–86
Pigott V C, Natapintu S 1988 Archaeological investigations into prehistoric copper production: The Thailand archaeometallurgy project 1984–1986. In: Maddin R (ed.) *The Beginning of the Use of Metals and Alloys*. MIT Press, Cambridge, MA, pp. 156–62
Shoocongdej R 2000 Forager mobility organization in seasonal tropical environments of western Thailand. *World Archaeology* **32**: 14–40
Stark M T, Griffin P B, Chuch P, Ledgerwood J, Dega M, Mortland C, Dowling N, Bayman J M, Bong S, Tea V, Chhan C, Latinis K 1999 Results of the 1995–1996 archaeological field investigations at Angkor Borei, Cambodia. *Asian Perspectives* **38**: 7–36
Stark M T Allen S J (eds) 1998 *International Journal of Historical Archaeology* (*The Transition to History in Southeast Asia*) **2**(3 and 4)
Swisher C C, Curtis G H, Jacob T, Getty A G, Suprijo A, Widiasmoro 1994 Age of the earliest known hominids in Java, Indonesia. *Science* **263**: 1118–21
White J C, Pigott V C 1996 From community craft to regional specialization: Intensification of copper production in prestate Thailand. In: Wailes B (ed.) *Craft Specialization and Social Evolution: In Memory of V. Gordon Childe*. The University Museum of Archaeology and Anthropology, University of Pennsylvania Press, Philadelphia, PA, pp. 151–75
Wisseman C J 1995 State formation in early maritime Southeast Asia. *Bijdragen* **151**: 235–88

E. A. Bacus

Southeast Asia: Sociocultural Aspects

Southeast Asia includes the ASEAN states of Brunei, Indonesia, Malaysia, the Philippines, Singapore, Myanmar, Thailand, Cambodia, Laos, Vietnam, and the territory of East Timor. Despite its extreme diversity in terms of population density, ethnic groupings, religion, politics, and economic development the area has common characteristics, many of them constructed by social scientists and politicians, that are increasingly recognized. In addition to the nation-building efforts of individual governments a common Southeast Asian identity has emerged, at least in the urban middle classes.

1. The Sociocultural Construction of Southeast Asia

1.1 Conceptualizing Southeast Asia

Southeast Asia has undergone a multitude of crises, economic, political, cultural, and social. Most of these have resulted in a fresh look at what course Southeast Asian societies have taken, and what concepts and theories can be used to analyze new developments. As a matter of fact, some of the new concepts and paradigms have been exported to other world regions, utilized and generalized and entered into the realm of general theory, while their procreators and their region of origin have often been forgotten. Dualism, plural societies, involution, strategic groups, or loosely structured social systems are by now standard terms in social science research, though originally they were empirically-based concepts used to describe Southeast Asian social conditions.

What we now call Southeast Asia has gone through a cycle of the social and political construction of itself. As Said (1978, p. 5) has rightly asserted 'geographical and cultural entities … are man-made.' Localities are not given, but socially constructed. To mention just a few terms that over time have signified the different external and internal efforts to construct a more or less unified image of the area: the 'Golden Chersonese' of Ptolemaeus, Further-India, Hinterindien, Extrême

Table 1
Conceptualizing Southeast Asia

Early concepts	Contemporary concepts	Future directed concepts
Social structure	Deconstruction	Transformation
Indianization, islamization, dualism	Moral economy of the peasant and of trade	Transnational business networks
Plural societies, loosely structured social systems, and involution	Deconstructing village society and ethnicity	Social transformation
Big men and theater states	Deconstructing class and the state: strategic group formation	Democratization and civil society
Modernization	Modernization of modernity	Globalization
Religious revivalism	Postmodernism, life styling, consumerism, and the new middle classes	Epistemic culture
Urbanization	Globalization and locality	Cultural construction of space

Orient, Nusantara, the Greater co-prosperity sphere, Pacific Asia, Southeast Asia, and, more recently, ASEAN.

Though empirical social science research has been carried out on many divergent topics, a limited number of theoretical concepts have emerged that show a certain consistency over time. The following table (Table 1) proposes a rough guide to the flow of sociocultural research on Southeast Asian societies. It also provides a framework in which to locate major concepts used in anthropological and sociological research on Southeast Asia. Some of these concepts will be explained below.

1.2 Indianization and Islamization

The term Southeast Asia became common use only during and after World War II. Before the area was seen as culturally colonized by its large neighbors India and China. Coedès (1968, pp. 15–16) described the classical states as *États Hindouisés* or in its translation as 'indianized states.' 'Indianization must be understood as the expansion of an organized culture that was founded upon the Indian conception of royalty, was characterized by Hinduist or Buddhist cults, the mythology of the *Purānas*, the observance of the *Dharmaśāstras*, and expressed itself in the Sanskrit language. Far from being destroyed by the conquerors, the native peoples of Southeast Asia found in Indian society, transplanted and modified, a framework within which their society could be integrated and developed.' This marriage of Austronesian and other local cultures with Indian patterns of sociocultural organization has remained a common background, on which later Chinese, Muslim, and European influences have been grafted. The Balinese and several smaller societies, such as the Tengger, maintain the Hindu tradition.

Islam came to Southeast Asia probably as early as the eighth century AD, but became a major force of social and political change only from the fourteenth century onwards. The creation first of a trading and religious, then of a political elite through conversion introduced a new element into the coastal states of Southeast Asia along the straits of Malacca, the Northern coast of Java, Sulawesi, and the Moluccas. The process of Islamization is a still ongoing process. Islamic revivalism is strong in Malaysia and is gaining strength in Indonesia, which has become the largest Muslim country worldwide.

1.3 Religious Diversity and Religious Revivalism

In addition to animistic beliefs and related belief systems, most major world religions are found in Southeast Asia. Indonesia, Brunei, and Malaysia together have the largest Muslim population worldwide. The majority of the population of Myanmar, Thailand, Cambodia, and Laos adheres to Theravada Buddhism, the Philippines has a majority of Christians, and Singapore and Vietnam are characterized by a multitude of religious beliefs often cutting across family and kinship lines, though Confucian ideas are also widespread. The Buddhist monkhood in Thailand (Tambiah 1976), Islam in Indonesia (Geertz 1960), and in Malaysia have been studied mainly in relation to local religious forms.

During the second half of the twentieth century, religious revivalist movements have swept through most of Southeast Asia. In Malaysia a missionary movement (*dakwah*) has led to a resurgence of Islamic practices, especially in the Eastcoast states of Kelantan and Trengganu. In Thailand, middle-class-based religious organizations have preached a return to Buddhist values. The Javanese mystical tradition has influenced Muslim brotherhoods or led to the foun-

dation of mystical sects: aliran kebatinan, Subud, and others (Geels 1997). The Singapore government has propagated Confucian values, while Christian sects have proliferated and found new adherents among students and young upwardly mobile employees.

1.4 Ethnic Diversity

Boundaries between ethnic groups are blurred, and there are differences between self-interpretation and ethnic classifications imposed by others. In Singapore, the population officially is classified into Chinese, Indians, Malays, and others (usually identified as Eurasians and Caucasians), though there are many more self-ascribed categories. The Malaysian government distinguishes between *bumiputra* (sons of the soil) and immigrants, but there are also *Orang Asli* (aborigines), Indians, and Chinese. People of Arab descent no longer are seen officially as a separate group, but in Indonesia the culture, legal system, and language of larger ethnic groups as Javanese, Balinese, Batak, Minangkabau, and many others are recognized and expressed in cultural, though not political terms. In Thailand, Laos, Myanmar, and Vietnam the lowland population is distinguished from 'hill tribes,' an ethnically and linguistically diverse population (Kunstadter 1967). The difference between a lowland, rice-growing peasant society and nonsedentary 'tribal groups' practicing shifting agriculture has been emphasized but is disappearing slowly due to assimilation or replacement.

The peoples of Southeast Asia have also been classified along linguistic lines. Despite recent criticism, the classification into Sino-Tibetan, Austro-Asiatic, Tai-Kadai, and Malayo-Polynesian is still widely used (LeBar et al. 1964, 1972).

In most countries or regions we can distinguish one dominant lowland and several upland satellite societies. Anthropological and sociological research has enriched our knowledge of the major societies that have been politically dominant in the formation of pre- and postcolonial states, but also the large number of smaller societies from hunter and gatherer societies, like the Orang Asli groups in Malaysia, to larger groups, like the Shan and Kachin in Myanmar (Leach 1954), the various 'hill tribes' of Thailand, Laos, and Vietnam (Kunstadter 1967), or the matrilineal Minangkabau of West Sumatra (Kahn 1993).

1.5 Plural Societies

The in-migration of Arab, Chinese, and South Asian migrants in the wake of the expansion of a colonial economy has created what Furnivall termed a plural society, 'with different sections of the community living side by side, but separately, within the same political unit' (Furnivall 1948, p. 304). This ethnic pluralism was transcribed into colonial and postcolonial administrative practice, giving rise to a classification of the population into Chinese, Malay, Indian, and others in colonial and contemporarary Singapore or into indigenous Malays as 'sons of the soil' (*bumiputra*) and migrant communities in Malaysia. Thailand, in contrast, has followed a policy of assimilation, in which all minorities were classified as Thai, whereas in Indonesia the Javanese dominance was balanced precariously by allowing the construction of ethnic, but not political, pluralism in the form of a cultural symbolism of local art forms and dances. Except for Singapore, one common national language has been introduced and enforced with varying degree of acceptance and use in all Southeast Asian states.

1.6 Moral Economy

Geertz and others have underestimated the dynamics of Southeast Asian peasant societies. Rebellions and peasant uprisings have continued well into the twenty-first century (East Timor, Aceh). The 'moral economy of the peasant,' illustrated by Scott (1976) with reference to Vietnamese and Burmese societies, indicates that a subsistence ethic governs peasant behavior. Only if the self-defined subsistence level is threatened have peasants rebelled.

The moral economy also governs petty trade in the informal sector of most of Southeast Asia. The moral obligation of sharing work (*gotong royong* in Indonesia) and profits puts a burden on small traders, mostly women, who find it difficult to accumulate profits to sustain and expand their trading enterprises. There are, however, solutions to this 'trader's dilemma' (Evers and Shrader 1994): where indigenous traders fail, immigrant-trading minorities like the Chinese take over the business. Otherwise, the indigenous traders themselves improve their moral status by forming or joining religious fundamentalist movements, like the Santri traders of Indonesia (Geertz 1965), to escape the pressure of moral obligations. Often trade is carried out on a subsistence level like by female petty traders where the obligation to share meagre profits does not arise because no cash is accumulated.

The informal sector of small handicraft production, petty trade, and other services is widespread in Southeast Asia, has increased in Vietnam during the transformation from a planned socialist economy to a more market-oriented economy, but has declined in Singapore and Malaysia. This sector is playing an important role in cushioning the effects of land shortage in the rice growing areas of Southeast Asia, of rural–urban migration and a general lack of social security provisions.

1.7 Big Men and Theater States

The Austronesian/Melanesian/Polynesian content of Southeast Asian culture becomes evident, the more one proceeds Southeast through Indonesia and the

Philippines. Wolters (1982) has drawn attention to the 'big man' phenomenon, which he locates in Polynesia. The current emphasis on personal prowess, on the power of the oratory and the admiration for superiors and leaders is put into this context. The state and its aristocratic bureaucracy is admired as much in Thailand as in Malaysia or Indonesia.

The state in these countries is highly ritualized and has been named 'theatre state,' following Geertz's analysis of the precolonial Balinese states (Geertz 1980). The state was largely a performance with little administrative or military power. It is debatable whether this analysis withstands the test of historical data, but at least the theme of vagueness, looseness, involution, and theatricality in some Southeast Asian societies pervades the anthropological literature.

1.8 Loosely Structured Social Systems and Involution

John Embree once classified Thai society as 'loosely structured,' thus producing a long-drawn controversy among social scientists (Evers 1980, pp. 163–98). Thai society has a cognatic kinship system, a shallow recognition of descent, and a high degree of individualism, but a highly regulated 'bureaucratic polity' and a tendency to form cliques and patronage. Another, equally unconvincing attempt at using single cultural traits or selected ethnographic details has stressed dualism as a common characteristic of Southeast Asian societies. There was a greater acceptance of the work of Geertz, who classified Javanese society as 'involuted,' by 'overdriving of an established form in such a way that it becomes rigid through inward over elaboration of detail' (Geertz 1963, p. 82). This process applied mainly to wet rice agriculture and its tenancy system, but was seen at work by other scholars also in other regions like the Philippines or Thailand, or in the relatively slow process of Southeast Asian urbanization up to the middle of the twentieth century (Evers and Korff 2000).

2. Contemporary Southeast Asian Studies

2.1 Diasporas and Business Networks

Migrant communities, like the Overseas Chinese originating from Southern China, the South Indian Chettiar, or the Hadramaut Arabs, have been classified as trading minorities in view of their economic position in Southeast Asian societies. Others have migrated to Southeast Asia in large numbers at the beginning of the twentieth century to work in plantations, mines, and the construction sector. Chinese business has been studied in great detail. Singapore has become the centre of the Chinese business community (Chan and Chiang 1994, Menkhoff and Gerke 2001). Confucian ethics, patrilineal kinship-based networks, and the patriarchal organization of family firms have been held responsible for economic success and the predominance of the Nanyang Chinese. Prosecution, discrimination, or conflict have strained interethnic relations in Malaysia, Indonesia, and Vietnam.

2.2 Social Class and Strategic Group Formation

Colonial Southeast Asian societies were 'moulded on racial principles: belonging to the dominant white upper caste provided one with prestige and power largely independent of one's personal capabilities. A strict ritual was introduced and maintained, by force when necessary, to preserve the white caste from contacts with Asiatics on a basis of equality and to maintain the former's prestige as the dominant group' (Wertheim 1968, p. 432).

In the course of decolonization and modernization Southeast Asian societies have undergone a process of class formation (Evers 1973, pp. 108–31). The colonial upper classes were replaced by ethnically differentiated elites, in which indigenous aristocrats, Chinese and European business people, and the upper ranks of a fast-growing bureaucracy and military officers played a major role. The concept of strategic group can be used to analyze this new situation more adequately. The pattern of strategic group formation differed from country to country. In Thailand, which had escaped the fate of colonialism, the growing bureaucracy, backed by sections of the military, revolted against the aristocracy and replaced the absolute monarchy in 1932 with a new regime, in which the strategic group of the military gained the upper hand, until a growing group of professionals, university lecturers, and students managed to establish a democratic government. Strategic group conflict continued, mediated by the electoral process, and put big business into power in 2001. Similar strategic group conflicts, rather than class struggle, signified the situation in most countries, except perhaps in the socialist states of Vietnam and Laos, were party cadres, another strategic group, maintained power after the communist victory in the Vietnam War. The partial reprivatisation of land and the deregulation of the economy introduced, however, new social processes from rural–urban migration to corruption of government officials and the formation of a group of newly rich businessmen.

Class formation moved ahead and the growth of a middle class became a much researched topic in Southeast Asia (Robison and Goodman 1996, Chua 2000). The growth of a new middle class was seen as a basic requirement for democratization and the building of a civil society. Whereas in Thailand, the Philippines, and in Indonesia a civil society with a proliferation of nongovernment organizations (NGOs) developed and pressed for reformation and

the introduction of democracy, the prodemocracy movement was crushed in Myanmar, while the strategic group of military officers stayed in power.

Middle classes have developed a special life-style of their own that blends Western consumerism with elements of Asian modes of living. Hawker centers, food courts, and markets exist alongside big department stores, pizza parlors, and hamburger outlets. The anthropology of Southeast Asian food has to be concerned with syncretism and life-styling as much as religious symbolism.

Along with the demand for consumer goods, the demand for knowledge also increased. In fact, the ASEAN region has undergone a tremendous expansion of higher education. In the core states Thailand, Malaysia, Singapore, the Philippines, and Indonesia the number of students and teachers, and of secondary schools, colleges, and universities, has risen dramatically. Malaysia and Singapore, in particular, are on the way to become knowledge societies, where economic and social development is increasingly knowledge driven. Most ASEAN leaders and governments have developed visions of the future that are targeting a knowledge society as a way to achieve parity with the Western world. This vision is but one in a long line of ideas on how Southeast Asia should be seen, interpreted, and eventually constructed.

2.3 Family and Gender Relations

The small nuclear family is widespread, or has become the norm more recently as in Vietnam (Pham Van Bich 1999) or among Singapore Chinese. The position of women has been regarded as strong throughout Southeast Asia; in particular in matrilineal societies like the Minangkabau (Kahn 1993). Demographic changes and the decline of fertility due to family planning has strengthened the position of women in Thailand, Vietnam, and Indonesia, giving them greater control over their own lives (Gerke 1992). Industrialization and the employment of female factory workers, especially in the garment and electronics industry has led to migration and chances for educational attainment but also to exploitation and harassment of women (Heyzer 1985, Sen and Stivens 1998).

2.4 Globalization, Urbanization, and Locality

Right into the 1980s Southeast Asia, except the city state of Singapore, showed only relatively low rates of urbanization. Since then urbanization has taken off, creating large urban conglomerates like the Greater Bangkok area, Jabotabek (Jakarta, Bogor, Tanggerang, Bekasi), the Manila Metropolitan Area, and the Klang Valley in Malaysia including the federal territory of Kuala Lumpur, the new administrative city of Putrajaya, and the high-technology center Cyberjaya. Southeast Asian urbanism has its roots in the Chinese-influenced walled city form, like Bangkok or Mandalay, the Malay or Muslim ports like Malacca, Banten, or Makassar, the court cities like Yogyakarta, and the colonial cities like Yangoon, Singapore, or Manila (Evers and Korff 2000). Next to the city centers with their culturally globalized population, we find localities of migrant groups, still maintaining a sociocultural flavor of their own. The various China towns, 'Little India' in Singapore (Siddique and Shotam 1982), or localities in Greater Manila (Berner 1997) should not be seen as remnants of traditional culture, but newly constructed forms of urban society.

2.5 Cultural Analysis

Southeast Asian cultures as systems of beliefs, values, symbols, and presentations are as diverse as the ethnic plurality of the local and migrant population. This diversity was analyzed in detail in earlier anthropological studies (Koentjaraningrat 1985, Boon 1977, Lombard 1990, and many others). There are, however, some underlying trends that have been used to construct national cultures or an ASEAN identity. The family and its values, reaching decisions through consensus rather the majority vote and a religiously grounded work ethic have been stressed in a debate on Asian values. Social scientists have taken up the cultural analysis of modernity (Hooker 1995). In his study of Thai, Indonesian, and Philippine middle-class culture, Mulder (1992) has stressed the reinterpretation of classical concepts for the construction of modernity.

The knowledge on Southeast Asian societies is increasing fast (Halib and Huxley 1996). Most of the relevant books and articles are still produced by foreign scholars, affiliated to universities or research institutions around the globe, but an increasing proportion of works are written by Southeast Asian nationals or by scholars, attached to Southeast Asian universities. An epistemic culture of local Southeast Asianists is emerging, contributing to a further construction of Southeast Asia as a sociocultural unit.

See also: Big Man, Anthropology of; Diaspora; Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity: Anthropological Aspects; Plural Societies; Southeast Asian Studies: Culture; Southeast Asian Studies: Economics; Southeast Asian Studies: Gender; Southeast Asian Studies: Geography; Southeast Asian Studies: Politics; Southeast Asian Studies: Religion; Southeast Asian Studies: Society

Bibliography

Barry K (ed.) 1996 *Vietnam's Women in Transition*. St. Martin's Press, New York

Berner E 1997 *Defending a Place in the City. Localities and the Struggle for Urban Land in Metro Manila*. Ateneo de Manila University Press, Quezon City
Boon J A 1977 *The Anthropological Romance of Bali 1597–1972. Dynamic Perspectives in Marriage and Caste, Politics and Religion*. Cambridge University Press, Cambridge, UK
Chan K B, Chiang C 1994 *Stepping Out. The Making of Chinese Entrepreneurs*. Prentice Hall, New York
Chua B-H 1995 *Communitarian Ideology and Democracy in Singapore*. Routledge, London
Chua B-H (ed.) 2000 *Consumption in Asia. Lifestyles and Identities*. Routledge, London
Coedès G 1968 *The Indianized States of Southeast Asia*. East-West Center Press, Honolulu
Condominas G 1977 *We Have Eaten the Forest. The Story of a Montagnard Village in the Central Highlands of Vietnam*. Penguin, London
Evers H-D (ed.) 1973 *Modernization in Southeast Asia*. Oxford University Press, Singapore
Evers H-D (ed.) 1980 *Sociology of Southeast Asia*. Oxford University Press, Singapore
Evers H-D, Korff R 2000 *Southeast Asian Urbanism. The Meaning and Power of Social Space*. LIT Verlag, Münster; St. Martin's Press, New York
Evers H-D, Schrader H (eds.) 1994 *The Moral Economy of Trade*. Routledge, London
Furnivall J S 1948 *Colonial Policy and Practice*. Cambridge University Press, London
Geels A 1997 *Subud and the Javanese Mystical Tradition*. Curzon, Richmond
Geertz C 1960 *The Religion of Java*. Free Press, New York
Geertz C 1963 *Agricultural Involution*. University of California Press, Berkeley, CA
Geertz C 1965 *The Social History of an Indonesian Town*. MIT Press, Cambridge, MA
Geertz C 1980 *Negara, the Theatre State in Nineteenth-century Bali*. Princeton University Press, Princeton, NJ
Gerke S 1992 *Social Change and Life Planning of Javanese Women*. Breitenbach, Saarbrücken and Fort Lauderdale, TX
Gomez E T 1999 *Chinese Business in Malaysia*. Curzon, Richmond
Halib M, Huxley T (eds.) 1996 *An Introduction to Southeast Asian Studies*. Tauris, London
Heyzer N 1985 *Working Women in Southeast Asia: Development, Subordination and Emancipation*. The Open University Press, Reading, UK
Hooker V M (ed.) 1995 *Culture and Society in New Order Indonesia*. Oxford University Press, Kuala Lumpur
Kahn J 1993 *Constituting the Minangkabau. Peasants, Culture, and Modernity in Colonial Indonesia*. Berg, Oxford, UK
Koentjaraningrat R M 1985 *Javanese Culture*. Oxford University Press, Singapore
Kunstadter P (ed.) 1967 *Southeast Asian Tribes, Minorities, and Nations*. 2 vols. Princeton University Press, Princeton, NJ
Leach E 1954 *Political System of Highland Burma*. London School of Economics, London
LeBar F M, Gerald H G, Musgrave J K (eds.) 1964 *Ethnic Groups of Mainland Southeast Asia*. HRAF, New Haven, CT
LeBar F M, Gerald H G, Musgrave J K (eds.) 1972 *Ethnic Groups of Insular Southeast Asia*. HRAF, New Haven, CT
Lombard D 1990 *Le Carrefour javanais: essai d'histoire globale* (3 vols.). École des Hautes Études en Sciences Sociales, Paris
Menkhoff T, Gerke S 2000 *Chinese Entrepreneurship and Asian Business Networks*. Curzon, Richmond, VA
Mulder N 1992 *Inside Southeast Asia: Thai, Javanese and Filipino Interpretation of Everday Life*. Editions Duan Kamol, Bangkok
Pham Van B 1999 *The Vietnamese Family in Change. The Case of the Red River Delta*. Curzon, Richmond, VA
Robison R, Goodman D S G (eds.) 1996 *The New Rich in Asia. Mobile Phones, McDonald's and Middle-class Revolution*. Routledge, London
Said E W 1978 *Orientalism*. Pantheon, New York
Scott J 1976 *The Moral Economy of the Peasant*. Yale University Press, New Haven, CT
Sen K, Stivens M (eds.) 1998 *Gender and Power in Affluent Asia*. Routledge, London
Siddique S, Shotam N P 1982 *Singapore's Little India: Past, Present and Future*. ISEAS, Singapore
Tambiah S J 1976 *World Conquerer and World Renouncer: A Study of Buddhism and Polity in Thailand Against a Historical Background*. Cambridge University Press, Cambridge, UK
Wertheim W F 1968 Asian society: Southeast Asia. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan and The Free Press, New York, Vol. 1, pp. 423–38
Wolters O W 1982 *History, Culture, and Region in Southeast Asian Perspectives*. ISEAS, Singapore

H.-D. Evers

Southeast Asian Studies: Culture

1. *Southeast Asia as a Cultural Region*

The study of culture in Southeast Asia was from the beginning influenced by debates among Western scholars as to just what comprised the region. Before and after World War II, there was a tendency to merge the study of Southeast Asian culture with that of India, as evident in references to the region as 'greater India.' This approach highlighted the extensive borrowing that had taken place between parts of the region and India in the pre-modern era. But the characterization was not helpful for understanding the many local populations less affected by these exchanges, including Muslims, Southeast Asian Chinese, and nonstate peoples living in the region's vast hinterlands. The characterization also failed to do justice to the selective ways in which pre-modern Indian influences have been integrated into the mosaic of modern Southeast Asia.

In the 1950s and 1960s, the institutionalization of Southeast Asian studies in Western universities led some scholars to attempt to devise a less derivative list of cultural traits that might allow researchers to visualize Southeast Asia as a culture area in its own right. Early iterations of this approach identified a number of traits as prototypically Southeast Asian:

(a) the high status of women, especially by contrast with China and India;

(b) court ceremony and high arts, both with strong Indian influences;

(c) the resilience of ancestral and spirit cult traditions alongside world religions;

(d) the prevalence of nuclear families and cognatic (or bilateral) kinship as opposed to extended families and unilineal descent systems; and

(e) a marked contrast between the egalitarian cultures of upland dry farmers and the hierarchical ways of the wet-rice lowlands.

Although this effort to bundle traits into a convincing cultural typology helped to distinguish the region from India and China, it also suffered from conceptual blind spots. Some Southeast Asian societies display few of these traits. Equally serious, typological approaches of this sort provide few clues as to how the elements that comprise any specific culture are integrated into a meaningful whole. In the 1970s, then, bundled typologies gave way to two new approaches to the study of Southeast Asian cultures: historical-dynamic approaches that examine the diffusion and integration of diverse elements into local cultures over time, and the 'thick description' of individual cultures in specific social settings.

1.1 Historical-dynamic Approaches to Cultural Variation

Dynamic approaches to Southeast Asian culture recognized that Southeast Asia lacked the cultural integration characteristic of China and India, but attempted to demonstrate that there was a unity to Southeast Asia in the way in which different populations integrated incoming elements into a unique cultural whole. In religion, for example, Southeast Asia today is home to large numbers of Muslims, Buddhists, and Christians, as well as small populations of Hindus (immigrant and indigenous). Alongside each of these religious communities, there are also strong traditions of shamanism, ancestral veneration, monistic mysticism, and guardian-spirit veneration. Rather than seeing this religious diversity as evidence of cultural incoherence, historical-dynamic approaches to Southeast Asian culture emphasize that local populations have selectively appropriated external cultural influences in a manner that reflects their own exigencies and constraints.

This selective appropriation of cultural elements appears to go far back in time. Recent studies have demonstrated, for example, that from the first centuries of the Common Era to the thirteenth and fourteenth centuries, Saivite Hinduism was the predominant cultic tradition at courts across Southeast Asia. Just as Greco-Roman civilization provided a model of cultural excellence for courts and high-status people in the pre-Christian Mediterranean, Hindu civilization provided the symbols of distinction in law, literature, the performing arts, state ceremony, and religion. Even today, the aesthetic traditions and ceremonial customs of many Southeast Asian societies show strong 'Indic' (Indian-influenced) inflections. Vietnam, the Philippines, and Southeast Asia's non-state hinterlands were not much involved in this cultural exchange. However, even in regions exposed to the full force of Hindu influence, there was clear selectivity in the appropriation of foreign elements. As with Hinduism in contemporary Bali (Warren 1993), bits and pieces of India's caste traditions were incorporated into religious ceremony and the social hierarchy, but no population in Southeast Asia ever implemented the full range of regulations operative among Indian Hindus in matters of diet, marriage, and everyday commensality. Similarly, the restrictions on women's autonomy seen among Indian Hindus (or, when making the comparison for Islam, Middle Eastern Muslims) exercised less influence in Southeast Asia, where women's public roles remained relatively high. Rather than 'failed' Hinduization, this resistance to caste and restrictive women's roles showed the continuing influence of Southeast Asian notions of commensality and female autonomy.

The dynamic approach to the study of Southeast Asian culture has also been used to examine key aspects of political and economic culture in Southeast Asia. In the 1960s and 1970s, political scientists and anthropologists used their understandings of local culture to assess the impact of nationalism and nation building on Southeast Asian societies. As with Clifford Geertz's (1973) reflections on the 'integrative revolution' in postcolonial Indonesia, this research was initially concerned with the prospects for peace and integration in an unstable region. The nationalist project in Southeast Asia was seen as particularly daunting owing to the deep ethnic and religious divides in most countries. The concern for nation building has remained a key theme in analysis of Southeast Asian politics to this day. Rather than linking their research to questions of national stability, however, recent studies have examined the cultural terms through which the nation is 'imagined,' and highlighted the role of elites in manipulating the 'invention' of national tradition for their own ends (Keyes et al. 1994, Tonneson and Antlov 1996, Winichakul 1994).

As these examples illustrate, one of the distinctive features of cultural analysis in Southeast Asia has long been that it is applied, not just to the usual domains of religion and the arts, but also to politics and economics. In Thailand, the political scientist Fred W. Riggs (1966) drew on anthropological studies of Thai society in assessing the prospects for sustained economic growth. Riggs characterized the Thai state as a 'bureaucratic polity,' suggesting that it was so thoroughly preoccupied with parasitic patronage that any sustained economic initiative was unlikely. Bureaucrats regarded Sino-Thai businesspeople as 'pariah entrepreneurs,' and restricted their operations accordingly. As Ruth McVey (1992) has observed, Riggs was right to emphasize the tensions between

political and economic culture in Thai society, and to imply that corruption might remain a brake on economic progress. However, Riggs failed to anticipate the speed with which capitalist culture took hold in Thai society in the 1970s and 1980s. Popular society and bureaucratic culture both proved more open to new styles of production and new modes of conspicuous consumption than cultural analysts had earlier predicted.

In a related study of indigenous enterprise in Java and Bali, the anthropologist Clifford Geertz (1963) reached a different conclusion than Riggs. Geertz argued that there were extensive cultural resources for market development in both of these Indonesian societies. In Java, Geertz observed, a culture of enterprise was taking shape among Muslim entrepreneurs marginal to the local (court-based) status hierarchy. In Bali, by contrast, the carriers of the market revolution were aristocrats intent on displaying their distinction in market enterprise to neutralize challenges from upstart rivals.

Although both applied cultural analysis to the study of economic change, Riggs and Geertz's conclusions were, curiously, inverse images of each other. Riggs attributed too great a hold to culture, and thus failed to anticipate the flexibility Thai society would show in the face of new economic initiatives. By the 1980s, capitalist culture in Thailand showed an 'increasing intimacy and equality between business and political leadership' (McVey 1992, p. 22). If Riggs' preoccupation with politics led him to exaggerate the conservatism of Thai culture, Geertz was too optimistic precisely because insufficiently attentive to the political. His gaze focused on local entrepreneurs, Geertz failed to notice the looming structural impediments to indigenous enterprise. Already in the late 1950s, the Indonesian military had become a major player in the national economy. Concerned that an independent business class might challenge their power, after 1965 military and civilian elites channeled contracts and concessions away from indigenous entrepreneurs toward wealthy Chinese and cronies reluctant to protest against mandatory kick-backs and sweetheart deals. The native entrepreneurial class that Geertz had lionized declined precipitously.

The Thai and Indonesian examples remind us of the dangers of seeing culture in too-static terms, or isolating it from the social and political worlds that constrain its development. Recent studies of the culture of capitalism in Southeast Asia have adopted just such an interactive approach to culture and sociopolitical structure in their efforts to understand the economic boom of the 1980s and early 1990s. Although attentive to the legacy of ethnic and religious culture, these studies also emphasize that culture is not a finished script or program, but an open body of knowledge continually reconstructed in the face of changing social and political realities (Hefner 1997, Li 1989).

Historical-dynamic approaches remain a key instrument in the study of Southeast Asian culture; they also serve to counteract the tendency to think of culture in purely local terms. Two historians, Denys Lombard (1990) and Anthony Reid (1988 and 1993), have both made especially significant contributions to the study of Southeast Asian culture on such a grand scale. Tracing cultural developments in the arts, economy, politics, and religion since the early modern era, both authors show that Southeast Asian culture has long been affected by globalizing forces and transnational flows. Historical linguists (Collins 1996) have recently used a similar methodology to examine the diffusion of Malay language across vast stretches of island Southeast Asia in the early modern era. Malay's spread, these studies demonstrate, was intertwined with the expansion of Islam and Malay-based commerce. In these and other examples, historical studies reveal that the cultural variation seen in Southeast Asia is the result, not of slumbering societies lost in time, but of vibrant regional networks and globalizations with deep historical precedents. Southeast Asian culture was 'globalized' well before the late twentieth century.

1.2 Thick Descriptive Culture

Dynamic approaches to the study of culture arose through the efforts of historians and others to explain Southeast Asia's diversity and to illuminate the creativity of Southeast Asians in appropriating external influences. Locally oriented, 'thick' descriptions of culture, by contrast, had less to do with Southeast Asian studies as such than they did a methodological shift in the human sciences in the 1970s and 1980s. The shift was premised on a new skepticism toward general theory, especially the modernization variant, and a heightened interest in local lifeworlds and meanings.

The single most influential writer on Southeast Asian culture, Clifford Geertz's intellectual biography illustrates this subtle shift in the terms of cultural analysis. In his early years, Geertz had been a student of modernization theory, especially that version promoted by his two mentors, Talcott Parsons and Edward Shils. As in his *Religion of Java* (1960) and his, *The Interpretation of Cultures* (1973), Geertz's concept of culture at first owed much to the Parsonian model of culture as a system of meanings in ongoing interaction with institutions and patterned action (the 'social system'). According to this model, rapid social change causes serious instability when it runs so far ahead of the cultural system as to make difficult the development of new systems of meaning capable of orienting collective behavior. In his analysis of religion in Java, Geertz identified three subtraditions (an Islamic, a syncretic, and an elite-aristocratic) and speculated that Indonesian social structure was changing more rapidly than these cultural traditions could accommodate. The resulting discontinuity between

structure and meaning, Geertz argued, was likely to give rise to serious religious conflict and political instability.

Even in his early works, however, Geertz used other methodologies that showed great sensitivity to the nuances of local context and meaning. In the late 1960s and early 1970s, as interest in modernization theory waned, Geertz shifted his attention away from general theory and broad analyses of modernization toward the 'thick description' of local lifeworlds, performances, and meanings. If commented on at all, regional movements and transcultural flows were distant concerns at best. This was to be a hallmark of cultural research in Southeast Asia for the next 20 years.

One salutary effect of this heightened skepticism toward general models and macrohistorical generalizations was that students of Southeast Asia were able to explore a range of topics neglected in earlier theories. There was, for example, an efflorescence of research on the interaction between local culture and great religious traditions. With its split gaze on Buddhism and indigenous religion, S. J. Tambiah's (1970) study of Buddhism and spirit cults in northeast Thailand represents an early and paradigmatic example of this new approach. Rather than portraying local religion as a 'little' tradition playing catch up to modernized forms, the ever-unfinished interaction of the two traditions was seen as the key to religion's vitality. Students of religion in Muslim Southeast Asia were slower to apply a similarly bifocal approach to Islam. As John Bowen (1993) has observed, they often neglected to examine the textual and normative bases of Islam, concentrating instead on syncretisms and exotic local customs. This neglect of the normative reflected the lack of training in Arabic and Islamic textual traditions among Southeast Asian specialists of Islam. Some argue that it also reflected the conviction among certain scholars that, unlike Buddhism, Islam was only a nominal influence on Southeast Asian culture as a whole. In the 1990s, however, as thick description began to be complemented by cross-cultural comparison, researchers made the balanced examination of transregional Islam and local tradition a hallmark of revitalized Islamic studies (Bowen 1993).

Another area of cultural research that showed new vitality in the aftermath of modernization theory's demise was the study of women and gender. Because of the unusually public status of women even in traditional societies, gender had from early on been a focus of attention in Southeast Asian cultural studies, as illustrated in the pioneering works of Hildred Geertz (1961) and Rosemary Firth (1966). Influenced by feminist studies and social constructionist understandings of gender in anthropology, however, in the 1980s and 1990s gender issues moved into a far more central position in Southeast Asian studies as a whole. Whereas earlier studies had examined gender from the perspective of the household and family, the new generation used gender as a point of entry to a broader and often more political array of concerns. Aihwa Ong (1987) looked at female factory workers to examine the dynamics of gender, capitalism, and Islamic resurgence in Malaysia. Pham Van Bich (1999) used changes in family and gender roles to provide a unique perspective on the achievements and shortcomings of the Vietnamese revolution. More generally, the study of gender in Southeast Asia has been used to challenge Western-derived understandings of the public and private spheres (Peletz 1996).

With this heightened sensitivity to gender, and with the growing influence of subaltern studies in the new field of cultural studies, there was growing interest in the early 1990s in the study of marginal groups and in the cultural construction of 'marginality' itself (Tsing 1993). Some proponents of this research opposed all forms of comparative and translocal research, arguing that broad comparisons only did violence to local life worlds. As this research moved forward, however, it inevitably gave rise to its own comparative questions. Gender inequalities in one society raise questions concerning the conditions that facilitate gender equity in others. Marginality raises questions concerning citizenship and the processes by which local actors might control the terms of their participation in the public sphere. As these examples hint, the fine-grained thick description of the 1980s had, by the late 1990s, given rise to a renewed interest in cultural comparison, translocal communication, and globalization.

2. *The Return of the Translocal*

The effort to apply in-depth local cultural research to translocal and comparative issues is not a new one in Southeast Asian studies. As noted above, it was always a feature of cultural history in the region. Similarly, literary research has long shown a lively interest in the changing uses and meanings of texts as they move across time and social space (Sweeney 1987). In the late 1970s, one anthropologist attempted to carry out comparative ethnography, examining reformist Muslim psychology in Malaysia, Singapore, and Indonesia (Peacock 1978). It would be years, however, before other anthropologists would follow his model. Finally, and very important, the study of Chinese culture in Southeast Asia has always been a great exception to the tendency to ignore comparison or focus simply on the local. In both anthropology and history, studies of Southeast Asian Chinese have always examined the movement of people and ideas across national borders and sought to understand the way in which local environments inflect Chinese culture. Recent research shows that this concern remains as vibrant as ever (Dobbin 1996, Reid 1996).

Another important precedent for comparative and translocal research in Southeast Asian studies has

been the study of the specifically cultural dimensions of nationalism. With its attention to the impact of vernacular education, colonial culture, religious reform, and new forms of public association on Malay political identity, William R. Roff's (1967) study of the origins of Malay nationalism was an early and exemplary example of just such a line of research. Anthony Milner's (1994) more recent study begins from a similar historical base, but shows an even more deliberate concern with the question of whether Southeast Asian modernity is unique or reflects political challenges general to the modern world. With its comparisons to Western debates on liberalism and communitarianism, Beng-Huat Chua's (1995) study of culture and politics in Singapore shows a similarly refined balance between the study of local cultures and comparative reflection on the modern condition as a whole.

As researchers look to the future, it seems inevitable that efforts like these to balance local research with regional and cross-cultural comparison will become more common. Because many Southeast Asian societies have always been deeply multicultural, it has been difficult here to ignore comparisons or, conversely, dissolve specific cultures into a general analytic brew. At the same time, Southeast Asia's position as a locus of civilizational interaction has helped to sensitize scholars to the fact that cultural 'globalization' is by no means a new phenomenon. Whether in the study of politics, gender, or economic culture, Southeast Asia also shows that trans-regional flows and globalization are not necessarily part of a long march toward a drearily homogeneous modernity. In this crossroads of civilizations, the most enduring cultural globalizations have always been those able to take root and 'localize.'

Bibliography

Bich P V 1999 *The Vietnamese Family in Change: The Case of the Red River Delta*. Curzon, Richmond, UK

Bowen J R 1993 *Muslims Through Discourse: Religion and Ritual in Gayo Society*. Princeton University Press, Princeton, NJ

Chua B-H 1995 *Communitarian Ideology and Democracy in Singapore*. Routledge, London

Collins J T 1996 *Malay, World Language: A Short History*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia

Dobbin C 1996 *Asian Entrepreneurial Minorities: Conjoint Communities in the Making of the World Economy, 1570–1940*. Curzon, Richmond, UK

Firth R 1966 *Housekeeping Among Malay Peasants*, 2nd edn. Athlone Press, London

Geertz C 1960 *Religion of Java*. Free Press, New York

Geertz C 1963 *Peddlers and Princes: Social Development and Economic Change in Two Indonesian Towns*. University of Chicago Press, Chicago

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York

Geertz H 1961 *The Javanese Family: A Study of Kinship and Socialization*. Free Press, New York

Hefner R W (ed.) 1997 *Market Cultures: Society and Morality in the New Asian Capitalisms*. Westview, Boulder, CO

Keyes C F, Kendall L, Hardacre H 1994 *Asian Visions of Authority: Religion and the Modern States of East and Southeast Asia*. University of Hawaii Press, Honolulu, HI

Li T 1989 *Malays in Singapore: Culture, Economy, Ideology*. Oxford University Press, Singapore

Lombard D 1990 *Le carrefour javanais: Essai d'histoire globale*, 3 Vols. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris

McVey R (ed.) 1992 *Southeast Asian Capitalists*. Southeast Asia Program, Cornell University, Ithaca, NY

Mead M 1949 *Male and Female: A Study of Sexes in a Changing World*. Morrow, New York

Milner A 1994 *The Invention of Politics in Colonial Malaya: Contesting Nationalism and the Expansion of the Public Sphere*. Cambridge University Press, Cambridge, UK

Ong A 1987 *Spirits of Resistance and Capitalist Discipline: Factory Women in Malaysia*. State University of New York Press, Albany, NY

Peacock J L 1978 *Muslim Puritans: Reformist Psychology in Southeast Asian Islam*. University of California Press, Berkeley, CA

Peletz M G 1996 *Reason and Passion: Representations of Gender in a Malay Society*. University of California Press, Berkeley, CA

Reid A 1988 and 1993 *Southeast Asia in the Age of Commerce, 1450–1680*, 2 Vols. Yale University Press, New Haven, CT

Reid A 1996 *Sojourners and Settlers: Histories of Southeast Asia and the Chinese*. Allen & Unwin, St. Leonard's, Australia

Riggs F W 1966 *Thailand: The Modernization of a Bureaucratic Polity*. East-West Center Press, Honolulu, HI

Roff W R 1967 *The Origins of Malay Nationalism*. Oxford University Press, Kuala Lumpur, Malaysia

Sweeney A 1987 *A Full Hearing: Orality and Literacy in the Malay World*. University of California Press, Berkeley, CA

Tambiah S J 1970 *Buddhism and the Spirit Cults in North-east Thailand*. Cambridge University Press, Cambridge, UK

Tonneson S, Antlov H 1996 *Asian Forms of the Nation*. Curzon, Richmond, UK

Tsing A L 1993 *In the Realm of the Diamond Queen: Marginality in an Out-of-the-way Place*. Princeton University Press, Princeton, NJ

Warren C 1993 *Adat and Dinas: Balinese Communities in the Indonesian State*. Oxford University Press, Kuala Lumpur, Malaysia

Winichakul T 1994 *Siam Mapped: A History of the Geo-body of a Nation*. University of Hawaii Press, Honolulu, HI

R. W. Hefner

Southeast Asian Studies: Economics

The ten countries of Southeast Asia—Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar (also still known as Burma), the Philippines, Singapore, Thailand, and Vietnam—soon to become eleven when East Timor obtains its independence, have a population of approximately 503 million persons, and an aggregate gross domestic product (GDP) of some $578 billion.

Their population is similar to that of the entire Latin American and Caribbean (LAC) region, about 80 percent of that of Sub-Saharan Africa (SSA), and 80 percent larger than that of the Middle East and North Africa (MENA). Their combined economies are similar to totals for South Asia and MENA, about 90 percent larger than SSA, about 60 percent of that of China, and about 30 percent of the LAC total.

With an average per capita GDP of $1,150 in 1998, the region is quite poor—about 50 percent above that of China, and almost treble that of India, but only about one-quarter of that of the large Latin American countries, Brazil and Mexico. The average, of course, conceals enormous intraregional differences. Indeed, apart from geographic proximity and membership of a regional association (ASEAN—the Association of Southeast Asian Nations), the countries share very few common characteristics. For example:

(a) They include countries both very rich (Singapore, with a per capita income of more than $30,000) and extremely poor (the four mainland Southeast Asian states, other than Thailand, have per capita incomes of less than $400, as does East Timor).

(b) They include economies which have been consistently among the most open in the world (Singapore, and to a lesser extent, Malaysia and Thailand) and those which, until quite recently, have been among the most closed in the world (the four mainland states again).

(c) There are mineral-rich economies, most notably the oil sultanate of Brunei, and also Indonesia and Malaysia, alongside the city state of Singapore, whose only real natural resource is its strategic location.

(d) Four of the ten countries have achieved rapid rates of economic growth since the 1960s, for at least three decades—Indonesia, Malaysia, Singapore, and Thailand. Three have generally performed very poorly—Cambodia, Laos, and Myanmar. The Philippines and Vietnam have an indifferent record over this period, taken as a whole, though for certain subperiods their performance has been quite good.

(e) The impact of the East Asian economic crisis, 1997–8, has been very uneven: Indonesia was very badly affected, as were Malaysia and Thailand, whereas the others, although affected to varying degrees, avoided economic contraction.

(f) One country is very large—Indonesia has a population of about 210 million people, the fourth largest in the world; three are sizable—the Philippines, Thailand and Vietnam each have 60–80 million people; while two (Brunei and East Timor) have fewer than one million inhabitants.

(g) There are pronounced subnational differences in all the larger countries: Java and the 'Outer Islands', north and south Vietnam, and East and West Malaysia. Indonesia and the Philippines are the world's largest archipelagic states. Bangkok is often mentioned as one of the most significant examples of a primate city.

This article draws attention to some of the stylized facts of the Southeast Asian economies, the lessons learned from past experiences, and some major ongoing analytical and policy debates.

1. *The Record, Lessons Learnt*

Economic performance and policy regimes in these countries have been episodic in varying degrees. Even high-growth Singapore experienced a mild recession in the mid-1980s, and a slowdown in 1998. The Philippines performed quite well until the late 1970s and during the mid-1990s but, like much of Latin America (with which it is often compared), it experienced a 'lost decade' during the 1980s. In the late 1980s Vietnam instituted significant liberalizing economic reforms under its 'Doi Moi' (renovation) banner, and in consequence its economic performance improved markedly. Indonesia swung from Sukarno's dirigisme, nationalizations and hyperinflation in the first half of the 1960s to economic orthodoxy for most of the Suharto rule (1966–98). Even Myanmar effectively renounced its one-time 'Burmese Road to Socialism' and isolation in the late 1980s, although, owing to its poor human rights record, it remains an international pariah state.

In such a heterogeneous region, generalizations are risky. Nevertheless, the following observations apply to many of the countries most of the time. First, rapid economic growth has been associated with profound structural change. As incomes have risen, the share of agriculture in output and employment has declined, at first as a proportion of the aggregate and at higher incomes in absolute terms. In Malaysia and Thailand, agriculture now contributes only about one-tenth of GDP, whereas in Cambodia and Laos its share is still about one-half. The slower the growth, the slower the rate of structural change. In the Philippines, for example, the share of manufacturing in GDP has not risen appreciably since the late 1960s.

Second, sustained economic growth is associated almost invariably with improved living standards and reduced poverty. In all the high-growth economies, there have been pronounced social improvements: education enrolments and life expectancy have risen, infant mortality and poverty incidence have fallen. Nevertheless, the key to poverty alleviation has been 'growth plus'. While there is less agreement over just what constitutes the 'plus,' the core elements almost certainly include the provision of basic public goods (e.g., primary education and health), an 'inclusive' labor-intensive growth path (especially one which combines outward orientation), and a reasonably equitable distribution of wealth and social opportunity.

There are, of course, marked differences. The Philippines had a head start in educational standards,

but over time the gap has narrowed. Regional differences also persist: Indonesia's eastern provinces are mostly poorer and have grown more slowly; Thailand has been concerned about its lagging northeast region; the northeastern states of West Malaysia are poorer than the more prosperous west and south; the southern regions have been more dynamic than the north in Vietnam. Particularly in Indonesia and Malaysia, much of the focus on distribution has an ethnic dimension, an issue to which we return below.

Third, we have at least a proximate understanding of the causes of economic growth, in Southeast Asia and elsewhere. Economic growth has been associated positively with—and by implication caused by—at least the following factors:

(a) macroeconomic orthodoxy, resulting from an appropriate mix of monetary, fiscal and exchange rate policies, has achieved a predictable and low inflation environment;

(b) openness to foreign trade, investment, and technology (and in some cases, labor);

(c) financial development, including effective prudential supervision;

(d) reasonable distribution outcomes, and a community perception that the benefits of growth have been 'fair';

(e) a set of political and legal institutions which, while not necessarily enshrining well-developed democratic principles, has at least ensured some measure of responsiveness to community concerns, including judicial redress and protection of property rights; and

(f) the public provision, or the coordinated private provision of, physical infrastructure at world competitive prices.

The particular 'fit' of these propositions to each country has, of course, varied considerably. But the higher a country has 'scored' according to these criteria, the better its long-term performance has been. Naturally, these factors do not take account of major insurrections (which plagued much of Indo-China until the late 1970s), serious political disturbances (a factor in Indonesia in 1965–6 and 1997–8 and the Philippines in the decade 1982–92), major seismological events (the Philippines again), or the systematic destruction of physical infrastructure (East Timor in late 1999).

Fourth, to varying degrees, these have been flexible economies, both in their policy regimes and their factor markets. Adverse exogenous shocks, such as a deterioration in the terms of trade, have resulted in compensating exchange rate and/or labor market adjustments. Indonesia in the early and mid-1980s is a good example of this proposition. When a strategy has not worked, or has been overdone—e.g., Singapore's high-wage policy in the early 1980s, Malaysia's high-tech ambitions in the 1980s—governments have introduced remedial measures. The presence of large numbers of migrant workers, often illegal, has facilitated labor market adjustment in Malaysia, Singapore, and Thailand. However, it is possible that some of these economies are becoming less flexible over time, as was evident in the crisis of 1997–8. Long-lived regimes in Indonesia and Malaysia were either unable or unwilling to introduce some of the bold reforms required by the international donor community. A similar problem afflicted the late Marcos period in the Philippines. Labor market regulations in Indonesia and the Philippines in the 1990s resulted in less flexibility. Vietnam's reforms seemed to have run out of steam by the mid–late 1990s. However, these economies generally remain more open and flexible than those of Latin America and South Asia.

Fifth, globalization, regional integration, and national economic integration have all proceeded more or less hand-in-hand throughout the region, to varying degrees, since at least the late 1980s, and a good deal earlier in some countries. Each has been propelled by different sets of factors. More open policy regimes have been introduced as suspicion towards markets and fear of Western commercial domination have waned. Initial reform successes have spawned coalitions favoring continued reforms. International institutions such as the IMF and the WTO have to some degree underpinned the process. Differences within the region remain very significant, however. Singapore has removed practically all obstacles to merchandise trade, but not services. Malaysia too has a very open goods sector, though in 1998 it introduced controversial controls on short-term capital movements. By contrast, the four poor economies of mainland Southeast Asia have barely established formal tariff codification provisions, nontariff barriers are often ill-defined (and are sometimes applied arbitrarily), and foreign investors have limited judicial protection.

Regional economic cooperation has been consistent with this general thrust toward increasing openness. ASEAN has wisely avoided EU-style bureaucratic concentration in its headquarters, and discriminatory trade policy liberalization. The Association was, however, unable to play an effective role during the region's recent economic crisis (or, for that matter, East Timor's political crisis of 1999).

Finally, there are numerous caveats which might be attached to the growth record, but none invalidates fundamentally the propositions that growth is good for a country's citizens, and that people are generally better off in richer than in poorer countries. Admittedly, the quality of urban amenities in Southeast Asia generally has deteriorated over the past three decades. But, through careful planning, ample resources, and strict controls over private automobile use, rich Singapore has one of the world's highest-quality urban environments. (Although, as illustrated dramatically during Indonesia's forest fires of 1997, it cannot control imported pollution.) Moreover, the most serious urban amenity problems seem to be

evident where rapid population growth and slow economic growth intersect, as in the case of Manila.

It is also probably the case that the quality of rural environments has deteriorated, as topsoils have been eroded, forests poorly (and corruptly) managed, mine wastes disposed of recklessly, and waterways polluted. Here also, economic growth may initially be a negative influence, but ultimately a positive factor, as incomes rise, and with them the resources to improve environmental standards. In addition, some exercises in 'green accounting' suggest that the economic growth in Indonesia and Malaysia may have been overstated by a percentage point or two once account is taken of the exploitation of nonrenewable natural resources, particularly oil and gas. Putting aside the heroic assumptions underpinning such calculations, here too it needs to be emphasized that economic growth in these countries was not explained primarily by resource depletion. Indeed, from the mid-1980s, growth accelerated in both Indonesia and Malaysia even as the share of the natural resource sector in the economy declined sharply.

2. *Some Debates and Puzzles*

While the general picture of economic development and its policy underpinnings may be clear enough, major analytical and policy debates abound, and there remain significant lacunae in our knowledge of the development process. The following are illustrative examples, by no means complete, of some of these gaps and intellectual challenges.

2.1 *The Crisis of 1997–8: Origins and Implications*

This crisis led to an unprecedented decline in output, of about 14 percent in Indonesia, 10 percent in Thailand, and 7 percent in Malaysia in 1998. It caused much social dislocation and hardship, it was instrumental in toppling a seemingly impregnable regime (Suharto's Indonesia), and it shook the region's self-confidence. It also caught governments, international development agencies, investors, and scholars off guard. Just when these groups were variously trumpeting the virtues of 'Asian style' development, making personal fortunes, or attempting to develop an intellectual framework for understanding the unprecedented high levels of growth, there was a sudden collapse. The crisis was not foreseen by anybody, notwithstanding much debate concerning the need for reform and the ills of corruption. There were some early warning indicators in Thailand, manifested in flight out of the capital market and the currency. Perhaps the biggest analytical conundrum was Indonesia: its economy remained buoyant right through to the general onset of the crisis in July 1997, and yet it suffered the most dramatic reversal of fortune in the region, from a precrisis growth rate of about 8 percent to a contraction of 14 percent.

There is still no consensus regarding the causes of the crisis among the development policy community and in its wake, and there are differences of emphases concerning future policy directions. On the causes, technical explanations focus on: (a) the dangers of a fixed but adjustable exchange rate regime which, in the presence of marked differentials between domestic and international interest rates, led to massive unhedged international borrowings; (b) rapidly rising short-term external debt and, more generally, a large stock of highly mobile capital; and (c) a poorly regulated and supervised domestic financial sector, especially in the presence of an open international capital account. Political and institutional explanations emphasize the widespread corruption and cronyism found in much of the region, and the weak state of the legal system, and 'civil society' more generally. A third set of explanations draws attention to international factors, including the rapid build-up of volatile, short-maturity international capital flows, the weak state of the Japanese economy throughout the 1990s, and the heavy-handed conditionality imposed by the International Monetary Fund (IMF) once the crisis had hit. Extending the latter point, a fourth dimension to the literature seeks an explanation not so much in the precrisis 'vulnerability' indicators as in mismanagement following its onset: Apart from inappropriate IMF medicine, this includes governments intent on protecting vested interests rather than reform (Indonesia and Thailand), inflammatory attacks on the international financial industry (Malaysia's prime minister), and loss of control of monetary policy (Indonesia).

Future policy directions are also the subject of much debate. Should countries allow their currencies to float? Will currency boards become attractive, as they were briefly to Indonesia's Suharto in early 1998? Or is Malaysia's fixed exchange rate a preferred option? The latter will normally entail some sort of supervision over short-term capital controls, which Malaysia instituted for 12 months in September 1998, and which continue to find favor among some intellectual circles in the region. Should the financial sector be open to foreign banks, and can central banks be trusted to supervise this sector? How can effective legal systems be developed in countries with poorly paid and demoralized civil services?

2.2 *The Philippines: Why an Outlier?*

Just as Indonesia's sudden economic collapse in 1998 is difficult to explain, so too is the Philippines' poor economic performance longer-term. In the 1950s, this was the country of hope in the region. It was richer than Korea and Thailand. It had the best education record in East Asia after Japan, and a large English-

speaking community. It had a functioning democracy, an independent judiciary, and a vigorous press. Its special relationship with the USA conferred upon it both security protection and privileged access to the US market. Yet, first the NIEs, then Thailand, and even Indonesia for a period, overtook it. Its real per capita income in 2000 was about the same as it was in 1980.

What went wrong? As with the crisis, there is no simple or single explanation, and a smorgasbord of offerings is available. These include:

(a) Technical explanations, which focus particularly on poor macroeconomic management (manifested in consistently above-average inflation rates) and inward-looking trade policies, including high levels of protection and a disequilibrium exchange rate.

(b) Cultural factors, among them the stereotype of a 'Latin American country located in East Asia.'

(c) Corruption and the 'Marcos factor,' particularly the nature of his regime in the last 6–8 years of his 20-year rule (1966–86); sometimes these arguments are extended to include the control by the alleged '20' (or '40', or some other number) ruling families, many of whom are viewed as nondevelopmental in some sense.

(d) Persistent insurrections, especially in the southern island of Mindanao, and the breakdown of law and order, resulting in frequent kidnappings of Filipino Chinese and foreign businessmen.

(e) Bad luck: In the mid-1970s, Marcos's strategy of external borrowing during OPEC I, sanctioned by the World Bank, came unstuck at the end of that decade, not only because the funds were used unwisely, but also because of the unfortunate conjunction of rising world interest rates, further increases in oil prices, and collapsing export prices (first sugar, then coconuts). In the late 1980s, the precarious post-Marcos recovery was jeopardized by a succession of coups and natural disasters, and serious power shortages. Ultimately, all these factors contribute to an explanation, but the puzzle remains, especially as some of these negative factors were also present in better-performing neighboring countries.

2.3 State vs. Market

Attempts to explain East Asian's economic success, beginning with Japan from the late 1940s, have to confront the fact that rapid growth has been associated with extensive, micro level government intervention. There continues to be a major debate over whether, and if so, how, this intervention has been growth promoting. The issue here is not concerned with government investments in physical infrastructure and the provision of social goods, which have usually generated high rates of return and promoted social cohesion. (Although large infrastructure projects have often been corruption-prone.) Rather, it is whether governments have demonstrated the capacity to 'pick winners,' to anticipate better than markets the industries of the future, and to manipulate interindustry (and sometimes interfirm) incentives accordingly.

The empirical evidence in support of this proposition in Southeast Asia is weak. The rates of return on expensive high-tech programs, commencing in the region mostly around the early 1980s, have generally been low, or even negative, even allowing for long gestation periods and substantial learning exercises. The externalities—broader economy-wide benefits—have been minimal. This applies to virtually all such initiatives—Indonesia's strategic industries (especially former President Habibie's aircraft plant), Malaysia's heavy industry program, and the Philippines' major industrial projects. The problem with these programs was not so much corruption, though there was plenty of that, but a failure to appreciate the evolutionary nature of technological innovation and industrialization, and the fact that 'great leaps forward' in a bureaucratically-driven environment rarely succeed. More generally, there is very little support for the proposition that selective industry assistance—via tariffs, fiscal incentives, credit subsidies, and so on—has paid off. Indeed, very often, the most assisted industries have performed the worst.

This is not to say that all forms of selective intervention in Southeast Asia have been unsuccessful. In Singapore, with its very open economy, well-paid bureaucracy, and fierce anti-corruption campaigns, it may well be that certain selective interventions have paid off—such as the goal of establishing the country as a major electronics base, and pushing the industry to upgrade quickly. Agricultural extension programs in several countries have induced farmers to innovate successfully, especially in the adoption of higher yielding varieties. The real challenge for the policy and research community in the region is to identify genuine cases of market failure, and to develop programs which do not exceed domestic bureaucratic capacities. The weaker these capacities—and most in Southeast Asia at the micro level are weak—the more likely the problem of 'government failure'.

2.4 Managing Ethnic and Regional Disparities

Southeast Asia is a region of great diversity, and one of its challenges is managing this diversity while maintaining the development momentum. Two dimensions are especially noteworthy—both long-standing but also painfully evident in Indonesia since 1997.

One is ethnic diversity. One of the colonial legacies was that disproportionate wealth was in the hands of ethnic minorities, particularly in Indonesia and Malaysia where, because of religious differences, ethnic assimilation has been very limited. In both countries, the minority ethnic Chinese community (less than 30 percent of Malaysia's population, and about 3 percent in Indonesia) traditionally has dominated the private, domestic, 'modern sector' economy.

Both countries have sought to maintain economic growth while experimenting with various affirmative action programs. Malaysia's approach, following the serious ethnic disturbances of May 1969, was to develop explicit goals for asset redistribution, employment, and university entrance. Indonesia adopted a more informal style, which included the promotion of state enterprises (especially during the 1970s oil boom), a preference for *pribumi* (indigenous) groups in government procurement, credit, and cooperatives, and occasional presidential cajoling of Chinese businesses to redistribute some of their assets.

The results have been mixed. There has been a major redistribution of assets in Malaysia, although the rising *bumiputera* share has been achieved mainly at the expense of foreign firms. Indonesia under Suharto witnessed the rise of several large *pribumi* conglomerates, mostly 'palace-connected,' and an expansion of *pribumi* managerial employment. It is sometimes argued that Malaysia avoided Indonesia's traumatic ethnic tensions of 1997–8 precisely because it had earlier and comprehensively instituted these programs. But there have been costs—the well-connected have benefited disproportionately, a new class of rent-seekers has emerged in both countries, and it is not obvious that the above approaches have been the most effective policy instruments. Good quality, widely accessible public education is probably the most effective long-term redistributional strategy. Here, both countries score well, though Indonesia much less so on quality grounds. Much more could be done in the way of enforcing a more progressive tax structure, again especially in Indonesia. The fierce resentment toward Indonesia's Chinese community during the recent crisis underlines the fact that governments need a strategy, and need to be seen to be doing something, but there are no 'quick fixes' in remedying such deep-seated disparities.

The second dimension, regional diversity, is equally complex. The record of regional development in the resource-rich countries of Indonesia and Malaysia has been fairly good, in the sense that some quite effective programs have been introduced to ameliorate sharp regional disparities. But in an era where nation states are under challenge as never before, regions with a rich natural resource base are asking why they have to surrender such a large proportion of revenue to the central government, especially when their own social indicators are sometimes lagging behind the national average. The problem has been particularly serious in cases where there is already a long-standing grievance against the central government, as in the Indonesian provinces of Aceh and Irian Jaya (the latter recently renamed Papua Barat, in an attempt to mollify local sentiment). East Timor was a special case, given its unhappy history under both Portuguese colonialism and heavy-handed Indonesian rule. But it is quite likely that other secessionist attempts will be mounted in Indonesia in coming years. Meanwhile, there will be continuing tension through the region between central governments' desire to maintain territorial integrity and national unity, and pressure from below for decentralization and greater autonomy.

2.5 Economic Growth and Political/Institutional Development

A major puzzle is why and how some Southeast Asian countries—especially, but not only, Indonesia—were able to grow so fast for so long in the presence of extensive corruption, much bureaucratic incompetence, and a weak legal system. As a corollary, the literature extolling the virtues of a strong and able civil service, and invoking it as a key explanatory factor in East Asian growth, has been of little relevance to Southeast Asia, with the exception of Singapore, and to some extent Malaysia.

Part of the explanation is that the high growth Southeast Asian economies with these problems had competent macroeconomic management, at least until the first half of the 1990s. If governments get a few key policy variables more or less 'right'—fiscal, monetary, and exchange rate policy; reasonably open trade and investment regimes; a basic national education program; and a predictable policy environment—economies can 'carry' major deficiencies elsewhere.

Another part of the explanation is that corruption is as much a distributional as a growth issue. Most civil service salaries in poorer Southeast Asian countries are below those in the private sector, especially for middle and senior officials, and thus small-scale corruption is a means, however morally reprehensible, of redressing the imbalance. Even grander corruption—for example, much of Indonesia's huge forestry royalties have tended to accrue to well-connected individuals rather than the state—may not necessarily retard growth if the funds are largely recycled through the domestic economy, as in the main they were. (Of course, in consequence, public programs were starved of resources, with short-term distributional and long-term growth implications.) Indeed, there is even a school of thought which, in seeking to explain the pronounced Indian–Indonesian growth differentials in the presence of similar levels of corruption, maintains that the latter's corruption was more conducive to growth because the recipients were more likely to 'deliver.'

But it is clear now that the past regime and practices are no longer sustainable. Perhaps corruption crosses some sort of 'threshold level,' when it so sours public support for a government (for example, the late Marcos and Suharto periods), that there are macro political ramifications. The presence of highly mobile international capital, which can suddenly leave the country if a political crisis develops, imposes a constraint on these governments which was hardly present in the 1980s and before. Continued IMF financial support for the crisis economies is also

conditional upon major institutional reform. Moreover, weak legal and financial institutions may not be a problem so long as the growth momentum is maintained. But they become a problem in times of crisis: bankruptcy laws have to function in order that debt workouts may proceed, for example.

Wide-ranging institutional reform is a complex and lengthy process. Democratization may hasten the process but it is not a simple cure-all. If indeed substantial reform is a necessary prerequisite for a durable recovery, the poorer countries of Southeast Asia may be in for a prolonged period of slower growth.

3. Summing Up

The Southeast Asian countries are important because they are large in aggregate, strategically located, exceptionally diverse, and intellectually challenging. Some analytical and policy issues in the region, as elsewhere, are reasonably settled: social progress follows economic development, and a few factors are absolutely critical to successful development. But region-wide generalizations for Southeast Asia beyond these basic propositions are hazardous. Rich, high-tech Singapore is a world away from the grinding poverty of East Timor and Myanmar. Some countries in the transition from communist to capitalist development have yet to establish the institutions which underpin a modern state that participates in the global economy. Indeed, some of the these states, particularly those on mainland Southeast Asia, are unsure how, and how far, to embrace globalization. The region's largest state, Indonesia, is now grappling with its territorial integrity and possibly even its survival.

For 30 years, ASEAN was the envy of the developing world. Now, quite suddenly, its economy is again growing slowly, there are serious social problems, and the region's seemingly powerful cohesion has faltered. The age of uncertainty has returned to Southeast Asia, as illustrated not only by its short-term economic prospects but also by the diversity of opinion on the many important social and economic challenges it now faces.

4. Further Reading

There is a huge literature on most aspects of the Southeast Asian economies. The bibliography contains some recent titles, which also include references to many earlier volumes. A 4-volume collection of leading articles on the subject will appear in Hill (2001).

See also: Area and International Studies: Development in Southeast Asia; Colonialism: Political Aspects; Colonization and Colonialism, History of; Development, Economics of; Development: Rural Development Strategies; Development: Social-anthropological Aspects; Quantification in History; Southeast Asian Studies: Geography; Southeast Asian Studies: Politics; Southeast Asian Studies: Society

Bibliography

Booth A 1998 *The Indonesian Economy in the Nineteenth and Twentieth Centuries: A History of Missed Opportunities*. Macmillan, London

Gomez E T, Jomo K S 1997 *Malaysia's Political Economy*. Cambridge University Press, Cambridge, UK

Hill H 2000 *The Indonesian Economy*, 2nd edn. Cambridge University Press, Cambridge, UK

Hill H (ed.) 2001 *Southeast Asian Economic Development*. Edward Elgar, Cheltenham, UK

Huff W G 1994 *The Economic Growth of Singapore: Trade and Development in the Twentieth Century*. Cambridge University Press, Cambridge, UK

Kyi K M et al. 2000 *Economic Development of Burma: A Vision and a Strategy*. Singapore University Press, Singapore

Leung S (ed.) 1999 *Vietnam and the East Asia Crisis*. Edward Elgar, Cheltenham, UK

Lucas R E B, Verry D 1999 *Restructuring the Malaysian Economy: Development and Human Resources*. Macmillan, London

McLeod R H, Garnaut R (eds.) 1998 *East Asia in Crisis: From Being a Miracle to Needing One?* Routledge, London

Phongpaichit P, Baker C 1995 *Thailand: Economy and Politics*. Oxford University Press, Kuala Lumpur, Malaysia

Vos R, Yap J T 1996 *The Philippine Economy: East Asia's Stray Cat? Structure, Finance and Adjustment*. Macmillan, London

Warr P (ed.) 1994 *The Thai Economy in Transition*. Cambridge University Press, Cambridge, UK

Woo W T, Sachs J D, Schwad K (eds.) *The Asian Financial Crisis: Lessons for a Resilient Asia*. MIT Press, Cambridge, MA

H. Hill

Southeast Asian Studies: Gender

Social science research on Southeast Asia only began to make its mark after World War II. The study of gender (the cultural practices and symbols which construct and assign certain roles for men and women) is very recent, and reflects its origins in Women's Studies and Women in Development programs.

1. Southeast Asia and the High Status of Women

The term 'Southeast Asia' (including Myanmar, Thailand, Laos, Cambodia, Vietnam, the Philippines, Singapore, Brunei, Malaysia, and Indonesia) can be dated to World War II, when it was used to refer to the theater of operations between China and India. Even in very early times, travelers saw this region as distinct from its larger neighbors, despite its cultural diversity. In 1944, in his *Histoire ancienne des états hindouisés d'extrême-orient*, the French scholar

Georges Cœdès (1886–1969) identified a shared corpus of cultural features, among which was 'the importance of the role conferred on women and of relationships in the maternal line.' Since the post-war development of Southeast Asia studies, the high status of women has been frequently invoked as a unifying characteristic of the region.

A number of reasons have been proposed to explain the relative equality of male–female relations in Southeast Asia. Among those most commonly cited are the importance of women in food production, especially rice cultivation; their prominence in marketing and other economic activities; complementary gender roles in ritual, and female prominence as healers and spirit mediums; the prevalence of bilateral kinship patterns, matrilocal residence and bride-wealth; the fact that women commonly inherit family property and maintain their own source of income; low population densities prior to the nineteenth century, which placed a high value on women's work and female fertility; the absence of a strong state structure and a low level of urbanization.

Individuals who incorporated both female and male elements were also important in indigenous belief systems, although this aspect of gender has received less scholarly attention.

2. *Gender in the Social Sciences Before World War II*

The foundations of social science research in Southeast Asia lie in the late eighteenth and early nineteenth century, when Europeans began to compile 'scientific' studies of the region. Apart from Thailand, all Southeast Asia was colonized by Europeans, whose research naturally focused on the areas under their control. By the late nineteenth century work on contemporary Myanmar, Singapore, Malaysia, and Brunei was almost solely the preserve of British scholars, with the Dutch concentrating on Indonesia, the French on Indochina (Laos, Vietnam, and Cambodia); and the Spanish and, after 1900, the Americans on the Philippines. Produced primarily by missionaries and colonial administrators, this research has provided valuable data for understanding gender relations, although that was rarely its focus. Colonial governments, by contrast, did produce periodic reports related to women, notably in regard to community health, literacy, and sex ratios, especially in countries that had sizeable migrant populations. The development of the new field of anthropology in the latter part of the nineteenth century prompted greater interest in local cultures and beliefs. Other social science fields considered the effects of capitalism and colonialism on indigenous society, but provided little room for the discussion of male and female relationships.

Until the 1930s ethnographic literature on Southeast Asia was largely confined to studies of smaller societies or tribal areas, but although the egalitarian nature of male–female relations was usually noted, gender issues were addressed only in passing. This began to change with the growing influence of structural anthropology. In the Netherlands, F. A. E. van Wouden's 1936 dissertation on social structure in eastern Indonesia gave gender relationships a prominent place because of his focus on cross-cousin marriage, the exchange of women, and the connections between matrilineal and patrilineal descent groups. In mainland Southeast Asia Edmund Leach's pre-war research on highland groups in Myanmar (published in 1954), strongly influenced by the work of Claude Lévi-Strauss, discussed the influence of cross-cousin marriages on constructions of hierarchy and equality in social relationships.

In the 1930s the expanding fields of social and cultural anthropology also coincided with a new generation of academic women whose research focused on the family and the domestic economy. Working in Bali, Margaret Mead, one of the foremost anthropologists of her time, gave specific attention to child-rearing, education, adolescence, and sexuality. Another American anthropologist, Jane Belo, also analyzed family relationships within Balinese culture. A different approach was taken by Rosemary Firth in her investigation of Malay village housekeeping in 1939–40. Her socioeconomic study emphasized what has become a truism in descriptions of Southeast Asian societies—women are dominant in marketing and small trading, have a strong entrepreneurial tradition, and control domestic finances.

3. *Gender and Southeast Asian Studies After World War II*

World War II represents a divide in the development of Southeast Asian studies. Not only did it witness the end of the colonial regimes; it also saw the escalation of Cold War hostilities, which enhanced the strategic importance of Southeast Asia. In the United States the government injected considerable funds into area studies, which led to a marked rise in social science research, especially on Thailand. Usually community-focused, this research encompassed a range of topics—nutritional intake, agricultural practices, child-rearing, the practice of magic, cultural values, etc.—which all necessitated an understanding of gender roles. Although comments on the relative equality between men and women in village societies became standard, specialists were cautious about generalizing beyond specific linguistic or ethnic groups. At most, they attempted to identify cultural features considered to be emblematic of a particular society—patron-client relationships in the Philippines, Thailand's loosely-structured social system—but these were rarely broken down in terms of gender.

During the 1950s and 1960s the disciplinary study of Southeast Asia was also introduced to (largely American) universities, but an interest in women's issues was not evident until the 1970s. As the region became more prosperous, local governments began to appreciate the economic importance of Southeast Asian women as well as the need to involve them in rural development programs. In the West social science work on Southeast Asia was influenced by the women's liberation movement and the development of feminist anthropology. The foundation issue of a new feminist journal included an article which identified the major social science research areas relevant to women in Southeast Asia (Papanek 1975), and it is worth noting that Michelle Rosaldo, co-editor of the classic collection, *Woman, Culture and Society* (1974) was a specialist on the Philippines.

The United Nations Decade of Women in 1975–85 was a catalyst for work on women. The combination of development agendas, feminist-inspired research and international concern with the effects of economic change in the Third World gave rise to a new academic field, Women in Development (WID). Pioneering studies focused on Asian women (Ward 1963, Boserup 1970) reasserted the economic autonomy and the high status of Southeast Asian women. The demand for work with practical applications for social policy and human resource management legitimized new research and teaching programs on women in Southeast Asia itself. While primarily concerned with empirical observation and evaluation, the work of Southeast Asian sociologists and developmental theorists has been a salutary reminder that analytical frameworks developed in the West do not necessarily apply elsewhere.

4. *Contemporary Trends*

Contemporary work on women and gender in Southeast Asia is unevenly distributed. The bulk of research has been carried out in Indonesia, where anthropology has traditionally been strong, and in Thailand, the Philippines, Malaysia, and Singapore. Work elsewhere is limited, although research on Vietnam is expanding. The persistence of country and local studies and the lack of comparative material makes generalizations about trends in scholarship difficult. In moving away from macro theories of gender to concentrate on specific groups, social scientists are demonstrating the need to place all research in historicized contexts that take into account factors like ethnicity and class. Historians have also warned against the tendency to place gender relations in a context that sets an exploitative present against an unduly idealized past. By pointing to similarities between Southeast Asian societies and those of southern India and southern China, scholars have questioned the received wisdom that contrasts an essentialized Southeast Asia with China and India. A recognition that the notion of 'Southeast Asian women' is a chimera has led on to the understanding that women's speech does not necessarily reveal a woman's point of view and that women do not always speak from the gender identity of women. The shift to the analysis of gender has made it clear that in order to understand women it is necessary to think about men in different kinds of ways, and be alert to the strategies by which masculinity is culturally created. Research on gay male identity in Thailand and the Philippines, for instance, has shown that male homosexuality, far from being an imported western borrowing imposed on traditional sexual or gender norms, involves redefinitions that are located within a specific sex/gender system.

A dominant paradigm in gender theory since the 1970s, the effects of globalized industrialization on the sexual division of labor, has been particularly relevant in Southeast Asia. In order to establish a manufacturing base Southeast Asian governments have used the lure of a cheap and docile work force to attract capital from the industrialized West. Much of the factory labor, especially in textiles, clothing, and electronics, has been provided by women, usually young and unmarried. From the 1980s a growing body of research, focused primarily on Indonesia, Thailand, Malaysia, and the Philippines, has explored the effects of economic globalization on women's work. Social scientists have not reached a consensus regarding the extent to which Southeast Asian women have benefited from the new employment opportunities. Some have argued that although the nature of work has changed, gender hierarchies have been transferred to the factory floor and social norms which make women subject to family and patriarchal controls have been retained. On the other hand, analyses that depict women as victims of patriarchy and capitalism have been challenged by research that places greater stress on female agency. A common theme has been patterns of resistance that assert female worker solidarity against a male management, and often result in small workplace triumphs. Less researched, however, are the expanding numbers of female entrepreneurs and employers among the middle class.

The study of gender in contemporary Southeast Asian societies has reaffirmed the problem of drawing a divide between the domestic economy and the formal work sector. As in the past, the average woman maintains her own finances independent of her husband, and indigenous economic networks mean that work is often an extension of home-based activities. Again, no Southeast Asian model has been established, in part because of the economic variation across the region, even within countries. In considering gender issues and industrialization, for example, it may be more useful to compare contemporary developments on Java with Gambia or Taiwan than with another part of Indonesia where economic development has not occurred.

Widespread industrialization and rural migration to cities has drawn attention to urban poverty and the tendency of unskilled and unemployed women to drift into prostitution. But in Southeast Asia, as elsewhere, social scientists have argued that the presumed universality of a concept like prostitution requires interrogation if it is to be used in transcultural contexts. Women involved in the sex trade can certainly be seen as victims, but it is also possible that poor women see prostitution as a path for advancement, especially when the customers are white men or there is the possibility of becoming 'temporary wives.' The phenomenon of 'mail-order brides' and the international recruitment of female workers, especially from the Philippines and Indonesia, have also attracted research interest, often in an effort to garner public and official support for greater legal protection of victimized women. In affluent and urbanized Singapore, by contrast, sociological research on gender is concerned less with poverty than with examining how different ethnic groups deal with smaller families and the care of the aged.

The rapid pace of economic change in combination with government goals for development has also stimulated studies of the effects of modernization on rural society, most notably in Indonesia, Malaysia, the Philippines, and Thailand. The literature can be roughly divided into officially sponsored projects of obviously practical application, and more critical and theoretical work, most commonly by anthropologists. Both categories have generated studies of women's responses to family planning and population control, and examinations of the extent to which women's lives have been changed by health and education. In rural Southeast Asia, where it is generally accepted that premodern gender relations were relatively equal, some scholars have contended that increased mechanization and rural development programs have privileged men and disadvantaged women. Similarly, it has been alleged that the increasing value placed on formal educational qualifications and the spread of state health services undermines traditionally female areas such as midwifery and spirit healing.

The argument that globalization has adversely affected women's traditional status has led to some re-examination of cultural constructs of gender, especially in regard to religious teaching. The acknowledged place of sexuality and the male–female relationship in indigenous belief systems has encouraged greater attention to the effects of the male-dominated world religions on gender relations in Southeast Asian societies, and the extent to which imported ideas were localized to suit existing gender relationships. The debate has been most pronounced in Thailand, where textual statements denigrating women's status in Buddhism can be contrasted with the prominence of women in the actual practice of religion. In Muslim societies the emergence of a more fundamental form of Islam that seeks a stricter application of *shari'ah* law has obvious implications for gender relations, especially in Malaysia where Islam is the state religion. In the context of Islam's adaptation to local custom, the combination of matrilineality, economic dynamism and Islamic commitment found among the Minangkabau in Sumatra and Negeri Sembilan in the Malay peninsula continues to attract researchers.

Historians of Southeast Asia are now giving greater attention to gender considerations and are providing social scientists with a stronger basis for discussions of change. Research on the nineteenth and early twentieth century, for example, has explored the interactions between race and gender in the hierarchies created by the colonial state. Another line of inquiry concerns the position of women in Southeast Asian nationalist movements prior to World War II, since it is evident that the call for female education and marriage reform reiterated by women's organizations were rarely accorded priority by male nationalist leaders. These issues were not resolved with independence, and it is evident that the colonial heritage often has a direct bearing on the way modern governments conceptualize gender roles. State involvement in education and the economy has perpetuated notions of gender difference and hierarchy by socializing men and women into different roles. For instance, there has been little encouragement for women to enter political life, except in support of male activities. Though the activities of women's branches of political parties have attracted some research (e.g., in Malaysia), this work has made no impact on mainstream descriptions of political systems. By contrast, the involvement of women in non-Government Organizations (NGOs) holds out the promise of theoretically stronger work, as researchers attempt to discern the extent to which NGOs replicate existing patterns of social hierarchy and gender, or act as agents of change.

Despite a general acceptance of gender complementarity in Southeast Asian societies, gender studies is still developing, and despite a growing interest in constructions of masculinity and gay identity, is currently dominated by work on women. Because of the array of cultures, languages, economies, religions, and modes of government, scholars of Southeast Asia will need to judge the extent to which they can move from case studies to larger generalizations. Specialists recognize that it is necessary to understand how gender operates within a particular social and cultural context before regional and global connections can be drawn. The similarities and contradictions that such investigations uncover may themselves be revealing, and should be actively identified rather than ignored.

See also: Fertility Transition: Southeast Asia; Islam and Gender; Land Rights and Gender; Southeast Asian Studies: Culture; Southeast Asian Studies: Politics; Southeast Asian Studies: Religion; Southeast Asian Studies: Society

Bibliography

Andaya L Y 2000 The Bissu: A Study of a Third Gender in Indonesia. In: Andaya B W (ed.) *Other Pasts: Women, Gender and History in Early Modern Southeast Asia*. Center for Southeast Asian Studies. University of Hawaii, Honolulu, HI

Atkinson J M, Errington S (eds.) 1990 *Power and Difference: Gender in Island Southeast Asia*. Stanford University Press, Stanford, CA

Boserup E 1970 *Woman's Role in Economic Development*. St. Martin's Press, New York

Eberhardt N (ed.) 1988 *Gender, Power and the Construction of the Moral Order: Studies from the Thai Periphery*. Center for Southeast Asian Studies, University of Wisconsin, Madison, WI

Firth R 1943 *Housekeeping among Malay Peasants*. Athlone Press, London

Hicks D 1984 *A Maternal Religion: The Role of Women in Tetum Myth and Ritual*. Center for Southeast Asian Studies, Northern Illinois University, DeKalb, IL

Karim W J (ed.) 1995 *'Male' and 'Female' in Developing Southeast Asia*. Berg, Oxford, UK

Keyes C F 1984 Mother or mistress but never a monk: Buddhist notions of female gender in rural Thailand. *American Ethnologist* **11**(2): 223–41

Manderson L, Jolly M (eds.) 1997 *Sites of Desire, Economies of Pleasure: Sexualities in Asia and the Pacific*. University of Chicago Press, Chicago

Murray A J 1991 *No Money, No Honey: A Study of Street Traders and Prostitutes in Jakarta*. Oxford University Press, Singapore

Ong A, Peletz M G (eds.) 1995 *Bewitching Women, Pious Men: Gender and Body Politics in Southeast Asia*. University of California Press, Berkeley, CA

Papanek H 1975 Women in South and Southeast Asia: Issues and Research. SIGNS **1**(1): 193–214

Rosaldo M Z, Lamphere L (eds.) 1974 *Woman, Culture and Society*. Stanford University Press, Stanford, CA

Van Esterik P (ed.) 1996 *Women of Southeast Asia*. Center for Southeast Asian Studies, Northern Illinois University, DeKalb, IL

Ward B E (ed.) 1963 *Women in the New Asia*. UNESCO, Paris

B. W. Andaya

Southeast Asian Studies: Geography

As with the classical roots of the discipline of geography, geographical interest in Southeast Asia (in the Western intellectual tradition) can be traced back to the era of discovery and exploration when the mystical Orient first became exposed to European curiosity and conquest. From the beginning of the sixteenth century the expanding European presence in the region was paralleled by a growing need for information about the *terra incognito* of the Orient, and erstwhile geographers were at the forefront of the collection and classification of such information: indeed, to all intents and purposes their activities at the time were geography. The functional need for knowledge and understanding to support colonial commerce and administration constituted what has been called the 'encyclopedic trend' in classical geography.

Towards the end of the nineteenth century geography started to emerge as an educational and academic discipline, and this led to the systematic study by professional academics of regions such as Southeast Asia both in Europe and within the colonies as part of the colonial educational system. Thus, well before Southeast Asia emerged as a recognized region with the creation of the South-East Asia Command during World War II, the geographical study of phenomena and processes therein was already well established.

Southeast Asian geography may be found at several points along a continuum between area studies and 'pure' geography, reflecting scholars' degree of areal specialization. Thus the field includes Southeast Asianists who happen also to be geographers, scholars in mainstream geography departments who take a specialist interest in Southeast Asia, and geographers whose work (e.g., on globalization, industrialization, environmental degradation) includes Southeast Asia. Whilst the inclusion of such a range of scholars makes any categorical overview of Southeast Asian geography difficult, it is equally unhelpful to exclude people simply on the basis of either their institutional location or their degree of disciplinary/area specialization.

Before World War II there would have been less of a dilemma in this regard. Up until the early 1960s the study of 'regions' and 'areal differentiation' constituted an important cornerstone of the geographer's approach, and thus regional geographers with a specialist interest in Southeast Asia were in many regards 'mainstream' geographers. Furthermore, until approximately this time there were no specialist area studies centers in which Southeast Asian geographers plied their trade (see Halib and Huxley 1996, pp. 2–5). In the United States, for instance, the first Southeast Asia program was initiated in Cornell University (and also Yale University) in 1950 by an endowment from the Rockefeller Foundation, but geography was not a prominent discipline in this program (as is still the case today). Similarly, although the School of Oriental Studies was established in London in 1917 (becoming the School of Oriental and African Studies (SOAS) in 1938), it was not until Charles Fisher was appointed the first Professor of Geography with reference to Asia at SOAS, in 1965, that Southeast Asian geography became established there. Similar situations appertained in continental Europe with geography being either absent or peripheral from such prominent Asia-focused institutions as the Koninklijk Instituut voor Taal-, Land- en Volkenkunde (KITLV) in Leiden (specializing on linguistics and anthropology), the École Française d'Extrême-Orient in Paris (anthropology and history), and the Institut für Asienkunde in Hamburg (political, economic, and social developments in Asia).

In the early period, some of the most Southeast Asia-focused geographers were those who were associated with the colonial presence in the region and, later, the strategic imperatives of the Pacific War. A number of Dutch geographers (e.g., S. van Valkenberg, H. Th. Verstappen, F. J. Ormeling) received their first exposure to the region as colonial officers. The French geographer Robert Garry was originally a member of the Colonial Service in French Indo-China. Robert Pendleton was attached to the Philippine Bureau of Plant Industry from 1923–35 during the period of US colonial administration, and later went to the Royal Department of Agriculture and Fisheries in Siam. Joseph Spencer joined the US Office of Strategic Services, and was later assigned to the Far Eastern Division as Assistant Chief. Karl Pelzer became part of a large US team concerned with the development of overseas territories.

Some of the people who were to become leading scholars of the geography of Southeast Asia first became associated with the region as staff members of local universities, including the University of Rangoon, Raffles College Singapore (later to form the University of Malaya), the University of Hanoi, and the University of Phnom Penh. From newly established geography departments within these institutions geographers embarked on fieldwork, sometimes in conjunction with undergraduate students, or otherwise developed their research interests in their host territory. Prominent scholars in this regard included Pierre Gourou, Dudley Stamp, O. H. K. Spate, E. H. G. Dobby, and B. W. Hodder. Dutch geographical scholarship on the East Indies was a little more restricted by virtue of the relatively late (postwar) establishment of departments of geography in universities on Java. Meanwhile, despite lacking a colonial presence in the region, German scholarship on Southeast Asia (e.g., Wilhelm Credner, Albert Kolb) dates back to the late nineteenth century and is distinctive for its ability to transcend national boundaries because of its lack of association with colonial territorialization. Another German scholar (later a naturalized American), Karl Pelzer, was one of the first to use the designation *Sudestasien* (Southeast Asia) in his work, some time before it became more popularly used to describe Mountbatten's South-East Asia Command.

In line with the requirements of war and colonial administration, and reflecting the prevailing orthodoxy in the discipline of geography, a central preoccupation of Southeast Asian geographers at the time was with gathering and presenting information, set within a defined regional or areal context and typically employing various cartographic techniques. A further preoccupation was the drawing together of the physical and human realms of life.

Among the 'area specialists,' two particularly stood out. The American geographer Joseph Spencer specialized on the cultural geography of the region, and was strongly influenced by the classical American geographer, Carl Sauer. Spencer also had wide experience of Asia, and this clearly influenced his identity and approach. The distinguished British Southeast Asian geographer, Charles Fisher, also combined an engagement of contemporary concepts and processes with the insight and 'feel' of the area specialist. He wrote several articles on matters appertaining to political development and international relations within Pacific Asia, but his most significant contribution was his classic text *South-East Asia: A Social, Economic and Political Geography* (1964). Fisher had the ability to cover in equal strength matters appertaining to history, physical geography, politics, society, and economy, a talent few of his more specialized successors possess.

From the 1960s the discipline of geography entered a positivist era as it underwent a 'quantitative revolution' and remodeled itself as a 'spatial science.' Regional geography was pushed to the periphery of the discipline, although it was not so much the emphasis on area, space, and place that was responsible for its marginalization as the predominant use of descriptive and unsophisticated methods. Regional geographers and area specialists continued to be dogged by the tag of 'chorology' (the identification and description of physical forms and human activities): regional geography was no longer seen as the highest form of the geographer's art.

During the 1970s spatial science itself became problematic, not least because of the relegation of context and human values in geographical analysis. Accordingly, in the 1980s a 'new regional geography' emerged which recognized the importance of area, space, place, and locale in the understanding of problems and processes in contemporary society. A new approach to the study of regions emerged which was much more broadly based in the social sciences and in social theory. One of the champions of this new regional geography, Nigel Thrift, was at the time engaged in research on Southeast Asia (especially Vietnam). Southeast Asian geographers have since made a full contribution to this new regional geography, and at the same time have regained a place and respectability within the discipline. This has been aided to a considerable extent by the recent transformation of Southeast Asia, which has increased its prominence within the global system and, at the same time, the level of academic interest. Geographers have been at the forefront of recent analyses of such processes as globalization and the global-local interface, resource exploitation and environmental degradation, regional integration, urbanization, migration, postcolonialism, and cultural change. Southeast Asia thus provides a rich domain for what Doreen Massey has described as the drawing into geography of the traditional realms of other disciplines. Rebecca Elmhirst's article in *Political Geography* (1999) on transmigration in Indonesia is a good example of how geographers are drawing on, and helping to advance,

concepts and ideas that have their roots in other social science disciplines.

The large majority of Southeast Asian geographers today can be found in mainstream geography departments, with many fewer in multidisciplinary area studies centers dedicated to the study of Southeast Asia. Most Southeast Asian geographers therefore have thematic specialisms that draw in Southeast Asia to varying degrees. By and large these are geographers rather than area specialists, the latter being identified by their greater degree of immersion in the region, linguistic capability, cultural awareness, post-Eurocentrism, etc., and their typical location in area studies departments. A perusal of the World Wide Web will reveal a large number of geographers in mainstream departments whose research and teaching draw on Southeast Asian examples; there are by comparison appreciably fewer geographers in dedicated Southeast Asian studies centers, typically seldom more than one, and quite frequently none. There are also a small number of distinctive 'clusters' of Southeast Asian geographers in disciplinary departments in several parts of the world. The following survey aims to convey a sense of the global structure of Southeast Asian geography, but is of necessity selective and thus far from exhaustive.

In Europe, the largest number of Southeast Asian geographers (approximately 30) is found in the United Kingdom, although only a handful are located in the country's two Southeast Asian area studies institutes (SOAS and Hull). The remainder are scattered across the length and breadth of the country, with small clusters in Manchester, Liverpool, and the University of London as a whole. In general, the Higher Education Funding Council for England (HEFCE) provides funding for these posts, whilst the majority of funding for research derives from the Economic and Social Research Council (ESRC) and the Natural Environmental Research Council (NERC). The British Academy Committee for South-East Asian Studies also provides dedicated funding for research in and on the region. Elsewhere in Europe the main loci for Southeast Asian geographers (mirroring the situation almost a century ago) are in the Netherlands (with a major cluster in the University of Utrecht, and individuals in the Universities of Nijmegen, Amsterdam, and Eindhoven), France (with a major cluster in the Centre National de la Recherche Scientifique), and Germany (Hamburg). There has also been a recent growth in Southeast Asian geography in Scandinavia, with the field being particularly well represented in the Universities of Lund and Göteborg in Sweden, and Copenhagen University in Denmark. Within Europe there are also a significant number of Southeast Asia-focused scholars whose work in, for example, demography, urban studies, planning, tourism, and ecology clearly overlaps with that of mainstream geographers. For more information on Southeast Asian geography in Europe, see van Dijk (1998).

The situation in North America is similar to that in Europe, with the vast majority of Southeast Asian geographers being found in disciplinary departments and with negligible representation in some of the long-established Southeast Asian studies programs (e.g., Cornell, Northern Illinois, the East-West Center). Southeast Asian geographers are quite dispersed in the United States (e.g., small numbers at the University of California at Berkeley, the University of Michigan), but there are larger clusters in Canada, most notably the University of Victoria and the University of British Columbia, underpinned by strong financial support for research on Southeast Asia by the Canadian government.

In Australasia there are significant clusters of Southeast Asian geographers at the University of Sydney (where there is a strong emphasis on Australian government-funded research on the Mekong subregion), Flinders University, and Monash University. Meanwhile, the principal area studies center, the Research School of Pacific and Asian Studies (together with the Department of Geography) at the Australian National University contains one of the world's most prominent groupings of geographers specializing on Southeast Asia. In New Zealand there are a number of Southeast Asian geographers in the universities of Canterbury, Wellington, Victoria, Massey, and Waikato.

One of the most important clusters of Southeast Asian geographers is found in the region itself, and it is likely to become even more prominent in the future as a result of changes that are currently taking place in the scholarly world, particularly in connection with the study of 'other' societies. A Southeast Asia Geography Association was formed in 1990 in order to create a forum for academic collaboration and integration within the region. There are several quite large geography departments in the region (e.g., in the University of Malaya, Universiti Kebangsaan Malaysia, Universitas Gadjah Mada, Universiti of Brunei Darussalam, University of the Philippines, and the National University of Singapore), and research-active staff members most typically engage in research work in and on the Southeast Asian region. Historically, much of the earlier work undertaken by Southeast Asian scholars was somewhat 'imitative of the Western academic tradition' (McGee 1995), largely because they received their initial training from the Western scholars who staffed the region's main university geography departments until as recently as the 1970s in many cases, and still today in the case of Singapore. However in recent years, in the National University of Singapore in particular, there has been a strong resurgence of innovative and stimulating indigenous scholarship by local staff who have looked critically at the nature and consequences of, for instance, 'globalization' and 'global-local interaction.'

Meanwhile, postcolonial tendencies in contemporary scholarship in the humanities and social sciences

are forcing Western scholars to look critically at the stifling and arrogant Eurocentrism that has characterized a significant amount of work on Southeast Asia in the past—an accusation from which geographers are far from absolved—and are currently leading to a reappraisal of established scholarship and the search for indigenous theories and concepts. Work by geographers on urbanization, rural-urban interaction, regional integration, and cultural resilience/resurgence has been particularly prominent in this regard, with several Southeast Asian scholars being at the forefront of recent developments (see Savage et al. 1993, Forbes 1997 McGee 1995, Rigg 1998). A 1999 joint publication by Singaporean and Western scholars on the contested territories of globalization in Pacific Asia (Olds et al. 1999) is indicative both of the current direction of geographical scholarship on Southeast Asia and of the fusing of Asian and Western geographical minds.

Southeast Asian geography has therefore experienced fluctuating fortunes as the region itself has waxed and waned in its global prominence over the last century or so. From being at the disciplinary forefront during the colonial era, courtesy of the imperatives of colonial administration and economic transformation, it experienced a largely self-inflicted hiatus during the immediate postindependence era. The development boom of the 1980s, and associated processes of global/regional integration and environmental degradation, rekindled geographers' interest in the region, and at the same time provided a powerful boost for local scholarship and institutions. The post-1997 economic crisis extended the post-boom shelf life of geographical research in and on Southeast Asia, and continuing processes of rapid economic, social, political, and environmental change should provide plenty of ammunition for future geographical enquiry. Whether the enquiring geographical minds will be those of 'area specialists' or 'area incidentalists' is difficult to anticipate: the disciplinary mainstream is once again disparaging the parochialism of area studies. Meanwhile, educational reforms in a neoliberal environment are further limiting the space and scope for 'areal dedication.'

Bibliography

Dijk K van 1998 *European Directory of South-East Asian Studies*. IIAS, Leiden, The Netherlands

Elmhirst R 1999 Space, identity politics and resource control in Indonesia's transmigration programme. *Political Geography* **18**(7): 813–35

Fisher C A 1964 *South-East Asia: A Social, Economic and Political Geography*. Methuen, London

Forbes D 1997 Metropolis and megaurban region in Pacific Asia. *Tijdschrift voor Economische en Sociale Geografie* **88**(5): 457–68

Halib M, Huxley T (eds.) 1996 *An Introduction to Southeast Asian Studies*. Tauris, London

McGee T 1995 Eurocentrism and geography: Reflections on Asian urbanization. In: Crush J (ed.) *Power of Development*. Routledge, London, pp. 192–210

Olds K, Dicken P, Kelly P F, Kong L, Yeung H W (eds.) 1999 *Globalisation and the Asia-Pacific: Contested Territories*. Warwick Studies on Globalisation Series, No. 1, Routledge, London

Rigg J 1998 *Southeast Asia: The Human Landscape of Modernization and Development*. Routledge, London

Savage V T, Kong L, Yeoh B S A 1993 The human geography of Southeast Asia: An analysis of post-war developments. *Singapore Journal of Tropical Geography* **14**(2): 229–49

M. J. G. Parnwell

Southeast Asian Studies: Health

Scientific medicine (biomedicine) has had a predominantly mechanistic, individual-oriented perspective, especially since the development of germ theory at the end of the nineteenth century. Health and illness are defined in terms of the presence or absence of pathogens in the body, and treatment relies on specific curative agents. An alternative approach, with an affinity to public health concerns, relates health and illness to the wider social and economic environment. It has its origin in the emergence of the science of political economy at the end of the eighteenth century and the 'consideration of disease as a political and economic problem of social collectivities' (Foucault 1980, p.166). The periodized discussion below of health in Southeast Asia draws upon multidisciplinary scholarship that reflects this social health perspective.

1. Precolonial Period

Anthony Reid (1988) argues that the health status of people in the precolonial period was generally better than their contemporaries in the West, and attributes their good health to good diet and hygiene (due to the abundance of water and habit of frequent bathing, and to the open and elevated houses). The region was by no means free of the afflictions of disease—leprosy, syphilis, and smallpox were especially feared. However, he suggests there was only 'a mild epidemic cycle.' From the fifteenth to the seventeenth centuries, Southeast Asia was very open to world commerce but the habit of washing and the pattern of dispersed village-type dwellings probably limited the spread of the worst urban infections of Europe and India (e.g., plague and typhoid).

Southeast Asian theories of disease causation (etiologies), diagnosis, and healing were characterized by their syncretism. Etiologies were either personalistic (disease caused by the intervention of supernatural beings) or naturalistic (disease caused by metaphysical

imbalance) or a complex amalgam of both. Personalistic theories derived from a prehistoric and indigenous substratum of animistic beliefs. Some naturalistic theories were also indigenous, such as the belief in the need to balance 'hot' and 'cold' substances. Others were adapted by Southeast Asians from Chinese, Indian (Ayurvedic) or Arabic (Unanic) humoral medicine (Owen 1987).

Modes of healing were similarly diverse and there was a great array of healers (astrologers, herbalists, bonesetters, masseurs and masseuses, midwives, exorcists, mediums, etc.). In seeking cures, Southeast Asians were pragmatic, using different modes of treatment in sequence or even concurrently (Owen 1987). How effective were these treatments? Certainly, Europeans were impressed by the variety and efficacy of herbal medicines and the relief provided by massage for rheumatism and soreness. Reid claims that the ubiquitous chewing of betel contributed to the prevention of tooth decay, indigestion, and dysentery (Reid 1988).

Despite the relatively good health of Southeast Asians in this period, population was very low, with an estimated total of a little more than 22 million people in the year 1600 and a population density of only 5.5 persons per square kilometer. Reid's theory (Reid 1988) is that this was due to endemic warfare caused by dynastic instability and succession wars. Battle casualties were actually minimal but crop destruction (resulting in famine) and the transportation of war captives led to great loss of life. Fertility was also probably low due to the constant need for flight during wars, with avoidance of further births until older children were able to run by themselves.

2. *Colonial Period*

In the nineteenth century, colonial governments became increasingly concerned about the heavy toll of death and disease in the colonies. However, initially this concern was limited largely to the high mortality and morbidity of the military. With the expansion of the colonial empires in the second half of the century medical services (including public health measures) were extended to the growing number of colonial administrators and other European expatriates. As colonial economic interests expanded (in the form of plantation, mines, and railways) concern grew for the health of the native populations, which was in large measure due to the realization that poor health affected the productivity of native workers. Thus, the government of British Malaya justified public health measures to control malaria on rubber estates in terms of the economic cost of the disease to planters. Also, the introduction of maternal and child health services in Malaya in the early twentieth century was based on an awareness of 'the importance of biological reproduction to ensure the supply of labor' (Manderson 1989). However, the extension of medical and health services to rural areas had ideological as well as economic functions. The apparent benefits of European 'scientific' medicine and public health served to legitimize the colonial state and to give it moral authority (Manderson 1996).

In Europe, the development of public health accorded doctors a powerful role in society as hygienists and led to a number of authoritarian medical interventions (Foucault 1980). Medical policing was no less evident in the European colonies of Southeast Asia, especially in the twentieth century. In Indonesia, Dutch health authorities adopted draconian measures to fight outbreaks of the plague in Java in the first three decades of the century (Hull 1987). In Malaya, public health surveillance intruded increasingly into the domestic and personal domains to search for vermin, whitewash walls, and disinfect premises as well as to check infant feeding and hygiene (Manderson 1989, 1996).

A noteworthy feature of the period of colonial rule was the rapid and prolonged population growth. The demographic growth began in about the mid eighteenth century and accelerated in the nineteenth century, reaching rates of natural increase of 1–2 per cent per annum. One theory is that the colonial *pax imperica* established an uneasy peace in the region, put a stop to internecine warfare and thereby reduced mortality. Other theories posit an increase in fertility arising from colonial demands for labor (in Java) and an increased demand for children. However, the evidence for these theories is inconclusive. Population growth in this period was definitely not matched by improved nutrition—there is some evidence of an actual decline in per capita food consumption—or reduced morbidity (Owen 1987). Morbidity and mortality were very high in colonial cities in the nineteenth century (at least for the predominantly non-European populations) (see Abeyesekere 1987 on Batavia, and Manderson 1996 on Kuala Lumpur).

Only in the twentieth century did the health situation improve noticeably. Despite some major plague and influenza epidemics, the overall frequency of epidemics declined. Public health measures (e.g., water supply, sewerage, swamp drainage) led to a reduction in cholera, typhoid, and malaria. Also, improved transportation enabled governments to ship food to areas hit by crop failures and epidemics (Owen 1987).

3. *Postcolonial Period*

By the late 1950s almost all the countries of Southeast Asia had achieved independence. Most of these countries became closely allied with the West. Postwar theories of modernization (e.g., Rostow's *Stages of Economic Growth* 1960) attributed the underdevelopment of Third World countries to the lack of diffusion of Western values, capital, skills, knowledge, and technology. Vicente Navarro (1982), writing from the

Table 1
Child (under-5) mortality (deaths per 1,000 live births)

	Per capita GNP ($US) (1995)	1960	1980	1997
Higher income countries				
Singapore	27,730	40	13	6
Brunei	11,990[a]	87	n.a.	12
Malaysia	3,890	105	42	21
Thailand	2,740	146	61	36
Philippines	1,050	102	70	42
Indonesia	980	216	128	59
Lower income countries				
Myanmar (Burma)	220[a]	237	146	90
Vietnam	240	219	105	51
Laos	350	233	190	140
Cambodia	270	217	330	131
Australia	18,720	24	13	8
Japan	39,640	40	11	6

a 1992
Source: UNICEF (1994); World Health Report (1998)

perspective of Marxist-oriented 'dependency theory,' has argued that the underdevelopment of Latin American countries, on the contrary, was caused by excessive cultural and economic dependence on the West due to the alliances between national urban elites and their Western metropolitan counterparts. From the health perspective this translates into self-serving, urban and elite-centered 'enclave' medicine that is reliant on Western medical education and technology.

In Southeast Asia, Thailand provides an apt example of such dependency. Donaldson (1982) has highlighted the key role of the Rockefeller Foundation in the dominance of Western scientific medical education in this country and the way it stifled attempts in the 1930s to introduce a low-cost, rural oriented Junior Doctor scheme. Others (e.g., Cohen 1989) have traced the subsequent growth of an elitist, urban centered and entrepreneurial medical profession that has accompanied the expansion of the capital sector in Thailand.

Socialism provided the philosophical basis for alternative health policies. In Vietnam, radical changes to health care accompanied the overthrow of French colonial rule in 1954, land reform, and agricultural collectivization. A socialist health policy was instituted based on principles of preventative medicine, combined use of Western and traditional medicine, and popular participation. The health system was highly decentralized and focused on the commune health center. Decentralization was consistent with socialist principles but was also a practical response to three decades of war (1945–75), including heavy US bombing in the last decade.

It was the success of Vietnamese and Chinese innovatory public health policies that served as models for the World Health Organization's primary health care (PHC) approach and goal of 'Health for All by Year 2000' enunciated in the Alma Ata Declaration of 1978. Since then all Southeast Asian countries have instituted some kind of PHC care system, and at least formal support for PHC principles of equity, popular participation, and affordable and culturally appropriate health care. This support was influenced in part by political unrest during the 1970s and by demands to redistribute wealth and income to satisfy the 'basic needs' of the poor (including improved living conditions and health). However, the level of political commitment to PHC throughout the region has been weak, except in Vietnam.

Nevertheless, there has been a trend toward better health in Southeast Asia (except for a period in Cambodia, which experienced economic decline and massive loss of life during the rule of Pol Pot between 1975 and 1978). At the most basic level, this general improvement in health has been a consequence of economic growth. This is demonstrated in Table 1, which relates per capita gross national product to child mortality (the principal indicator used by UNICEF to measure human and economic progress). In the lower income countries, Vietnam has achieved faster reductions in child mortality despite its low level of income—a consequence of its strong political commitment to PHC and very successful PHC initiatives.

The indicators also highlight the fact that there remain very significant differences in health between countries in the region, with Singapore and Brunei matching the industrialized nations of Australia and Japan, and with Myanmar, Laos, and Cambodia comparable in health status to the poor countries of Africa. However, national statistics for the higher income countries may disguise regions within coun-

tries with health status well below the national average, e.g., West Nusa Tenggara (Lombok, Sumbawa) in Indonesia, and Northeast Thailand.

The health status of urban areas in Southeast Asia is generally better than rural areas, in stark contrast to the pathogenic cities of the nineteenth century. In the postcolonial period, urban health has improved more rapidly than that of rural areas due to better access to safe water, sanitation, and health services. However, national statistics that compare urban and rural areas obscure sizeable pockets of urban poverty, and the existence of urban slums and squatter settlements. Here, appalling living conditions are ideal for the spread of mosquito-borne malaria and dengue fever, and for high levels of infectious and communicable diseases. The growth of urban slums and squatter settlements has been particularly pronounced in Malaysia and Thailand, which have experienced rapid economic growth based on urban manufacturing and rural–urban migration (see Cohen and Purcal 1995).

There are two diseases that have created serious health problems in Southeast Asia in recent years and threaten to reverse the general trend toward improved health in the region. These are malaria and HIV/AIDS; the one is resurgent and the other is emergent. Both are closely connected to poverty and human population movements.

In the 1950s and 1960s malaria was in decline in the region due to DDT spraying and highly effective drugs such as chloroquine. However, since then there has been a significant resurgence of the disease. For example, in 1998, Thailand had 125,013 confirmed cases of malaria. This resurgence in Southeast Asia can be partly attributed to insecticide-resistant mosquito vectors and drug-resistant strains of malaria parasites. Other causes can be found in the changing political economy of the region; for example, population pressure on land resources has forced land-poor and nonimmune families into circular migration in search of supplementary income in forested areas where they were exposed to malaria vectors (Singhanetra-Renard (1993).

The first case of HIV infection in the region was reported in Thailand in 1984. Within a few years the disease had reached epidemic proportions, fueled by international domestic sex industries (exploiting young women from the poor northern and northeastern regions) and by a large population of injecting drug users. By 1997, 780,000 people were HIV infected in Thailand. Here there is some evidence that the rate of growth had been slowed down by 1994 by government campaigns promoting public awareness and condom use. By this time, however, the disease was becoming rampant in other Southeast Asian countries, especially in Myanmar (440,000 cases of HIV infection) and in Cambodia (130,000 cases). The massive scale of population movement in the region, due to the integrative effect of global economic forces, has contributed greatly to the spread of the disease. In 1996 some 2.6 million migrant workers were circulating within the Asian region. Growth zones have encouraged the flow of people, goods, and ideas across international frontiers. For example, the Economic Quadrangle has established commercial and transport links between parts of Thailand, Laos, China, and Myanmar, and has led to an unprecedented movement across borders of poor people who are vulnerable to HIV infection—commercial sex workers, truckers, migrant workers, and ethnic minorities—in a way that greatly increases the chances of HIV transmission (Porter and Bennoun 1997).

Against a backdrop of general trends in health in Southeast Asia—some positive, some negative—I have presented an overview of health studies from a broad range of disciplines (history, anthropology, sociology, demography, and geography). What these studies have in common is a social health perspective. Theoretically, the work of some of these scholars can be identified with Marxist or Foucaultian political economy; others are less obviously associated with a particular theoretical approach but, nevertheless, relate health to political and economic changes in the region.

See also: AIDS (Acquired Immune-deficiency Syndrome); AIDS, Geography of; Colonialism, Anthropology of; Colonization and Colonialism, History of; Health: Anthropological Aspects; Public Health; Southeast Asian Studies: Society

Bibliography

Abeyesekere S 1987 Death and disease in nineteenth century Batavia. In: Owen N (ed.) *Death and Disease in Southeast Asia: Explorations in Social, Medical and Demographic History*. Oxford University Press, Oxford, UK

Cohen P 1989 The politics of primary health care in Thailand, with special reference to non-government organizations. In: Cohen P, Purcal J (eds.) *The Political Economy of Primary Health Care in Southeast Asia*. Australian Development Studies Network/ASEAN Training Centre for Primary Health Care Development, Canberra

Cohen P T, Purcal J T 1995 Urbanisation and health in Southeast Asia. In: Cohen P, Purcal J (eds.) *Health and Development in Southeast Asia*. Australian Development Studies Network, Australian National University, Canberra

Donaldson P J 1982 Foreign intervention in medical intervention: A case study of the Rockefeller Foundation's involvement in a Thai medical school. In: Navarro V (ed.) *Imperialism, Health and Medicine*. Pluto Press, London

Foucault M 1980 The politics of health in the eighteenth century. In: Gordon C (ed.) *Power/Knowledge: Selected Interviews and Other Writings 1972–1977*. Harvester Press, Brighton, UK

Hull T 1987 The plague in Java. In: Owen N (ed.) *Death and Disease in Southeast Asia: Explorations in Social, Medical and Demographic History*. Oxford University Press, Oxford, UK

Manderson L 1989 Political economy and the politics of gender: Maternal and child health in colonial Malaya. In: Cohen P, Purcal J (eds.) *The Political Economy of Primary Health Care*

in Southeast Asia. Australian Development Studies Network/ASEAN Training Centre for Primary Health Care Development, Australian National Univeristy, Canberra
Manderson L 1996 *Sickness and the State: Health and Illness in Colonial Malaya*. Cambridge University Press, Cambridge, UK
Navarro V 1982 The underdevelopment of health or the health of underdevelopment: An analysis of the distribution of human health resources in Latin America. In: Navarro V (ed.) *Imperialism, Health and Medicine*. Pluto Press, London
Owen N 1987 Towards a history of health in Southeast Asia. In: Owen N (ed.) *Death and Disease in Southeast Asia: Explorations in Social, Medical and Demographic History*. Oxford University Press, Oxford, UK
Porter D, Bennoun R 1997 On the borders of research, policy and practice: Outline of an agenda. In: Linge G, Porter D (eds.) *No Place for Borders: The HIV/AIDS Epidemic and Development in Asia and the Pacific*. St. Martin's Press, New York
Reid A 1987 Low population growth and its causes in pre-colonial Southeast Asia. In: Owen N G (ed.) *Death and Disease in Southeast Asia: Explorations in Social, Medical and Demographic History*. Oxford University Press, Oxford, UK
Reid A 1988 *Southeast Asia in the Age of Commerce, 1450–1680*. Yale University Press, New Haven, CT
Rostow W W 1960 *The Stages of Economic Growth*. Cambridge University Press, Cambridge, UK
Singhanetra-Renard A 1993 Malaria and mobility in Thailand. *Social Sciences in Medicine* **37**(9): 1147–54
UNICEF (United Nations Children's Fund) 1999 *The State of the World's Children*. Oxford University Press, Oxford, UK
World Health Report 1998 *Life in the 21st Century: A Vision for All*. World Health Organisation, Geneva World Health Organisation, Switzerland

P. T. Cohen

Southeast Asian Studies: Politics

The study of the politics of Southeast Asia, a subfield of comparative politics, developed as an area of academic research following the regaining of independence of the former colonies of the region after World War II. Driven initially by a desire by outsiders to understand the political dynamics of societies with different cultural and historical origins to those of the West, the study has now become embedded in some national academies, with more work being pursued by local scholars wishing to apprehend the political forces at work in their own and neighboring polities.

The development of the discipline has often been driven and shaped by the often global external security and economic interests of foreign scholars and governments, but more recently local concerns for democratization and the advancement of human liberties have emerged as a catalyst to research. Though a field of comparative politics, little of the published work has been truly comparative but rather has concentrated on the issues of political power, authoritative institutions, and constitutional issues within particular countries. The great political heterogeneity of the region, combined with its religious, cultural, and modern historical diversity, have posed severe constraints on any attempts to develop a truly comparative political theory of Southeast Asia. In several countries, the discipline has not developed to any meaningful extent because of the political constraints that are placed on independent analysis and public debate of governmental actions and political behavior.

1. *Nationalism*

The rise of nationalism as the colonial powers of Southeast Asia lost their desire and capacity to continue to rule in British controlled Burma, Malaya, Singapore, French-run Vietnam, Laos, and Cambodia, the American Philippines, and the Dutch East Indies (Indonesia), stimulated an awareness that these societies had internal political dynamics which were helping to shape the postwar international scene. The existence of powerful left-wing political forces in most of these nationalist movements, emerging at the same time as the United States and its European allies became concerned about the establishment of communist states beyond the Soviet Union, provided a political as well as an academic incentive for the development of Southeast Asian political studies. In the early 1950s, the first American university-based area studies centers devoted to the region began to study the origins and development of nationalism and communism.

These early studies were predominantly the result of research and publication by American and American-trained scholars. They turned, though, in most cases for their historical and cultural understanding of these new polities to the research done by colonial officials and scholars of the pre-war era. The notion that the apparently unique and distinctive cultures of these foreign lands were overwhelmingly important in shaping their politics was very influential at this time. Earlier research, summarizing the cosmological basis of Southeast Asian kingship, such as Robert Heine-Geldern's (1942) essay *Conceptions of State and Kingship in Southeast Asia*, became frequently cited as sources of theories about the nature of the pre-colonial state which, by analogy, had applicability to the then period.

Though nationalism was the predominant theme of these early studies, such as George McT. Kahin's path-breaking *Nationalism and Revolution in Indonesia* (1952), religion and ethnicity were seen as key components of that political phenomenon. Anthropological studies of Southeast Asian societies, which also began to emerge at about this time, had a significant impact on early political studies because of their insights into peasant political values and attitudes. Culture became the leitmotif of most political studies and the discipline as applied to the region tended to develop away from the mainstream political science

topics of industrialized societies. The fact that almost all Southeast Asian economies in the 1950s and 1960s were predominantly agrarian in nature, and that the peasantry formed the largest social category, provided a material basis for this view.

The American historical sympathy with nationalism, following from the dominant understanding of their own country's historical evolution and morally justified struggle against colonialism, made these studies naturally sympathetic to the nationalist cause. In those countries where the left, and especially communist political parties, had a major influence, studies attempted to differentiate between the indigenous, anti-colonial roots of nationalism and the allegedly foreign inspiration and guidance of left wing nationalist movements. Nationalism was not initially analyzed in broader ideological terms, and there was little attempt to distinguish among the often-contradictory origins and purposes of rival organizations and ideological predispositions within nationalist movements.

The religious, ethnic, and geographical diversity of Southeast Asian societies, divisions that frequently paralleled economic and administrative boundaries, were often obscured in the national level analyses that drove the nationalist paradigm. This was shared not only by most foreign scholars but also by national governments and capital-city based researchers. The findings of the postwar pioneers of Southeast Asian political research were summarized in the first, and still useful, major comprehensive review of the politics of Southeast Asia, the multi-author volume edited by George McT. Kahin, *Government and Politics of Southeast Asia* (1959, 1964). This highly influential work, with each country of the region discreetly analyzed with no attempt at any comparative analysis with other states either in or out of Southeast Asia, became the model for most of the next two decades.

2. *Behaviorism and the Collapse of Constitutional Democracy*

As the newly independent states of the region regained their independence between 1945 and 1965, each initially established a form of constitutional democracy, often modeled on the political institutions of the former colonial power. These political systems quickly failed to achieve the idealistic intentions of their formulators. The issue of the failure of constitutional democracy in the politics of Southeast Asia thus became an early issue in political research.

The rise of military and authoritarian regimes, even those which maintained the democratic facade of elections, became the next major research issue for students of the region. Herbert Feith's *The Decline of Constitutional Democracy in Indonesia* (1962) was amongst the most influential of these works. The political failure of constitutional democracy was often attributed to the belief that somehow the cultures of these countries were incompatible with the give and take of democratic politics. In peasant-based societies organized through strong patron-client ties controlling limited resources, the tolerance of criticism and debate that is part of democratic politics was argued to be missing but for cultural, rather than structural, reasons.

Thus while political studies in Western societies tended to focus on interests and institutions of power and authority, political studies of Southeast Asia generally looked to cultural forces and historical patterns. The very variety and complexity of Southeast Asian societies, which themselves posed severe issues of political unity and control on the often uncertain and weak postcolonial governments of the region, were consequently somewhat overlooked or dismissed. The postcolonial states were attempting to hold together what in many cases were artificially created colonial entities containing sharply contrasting communities with diverse and not necessarily coherent interests with fewer financial, administrative, and military resources than the former colonial administrations could have mobilized.

The development of authoritarian, and usually army-led or army-backed, governments in many nations in Asia, Africa, and Latin America during the height of the Cold War in the 1960s, led to a spate of studies on the role of the military in politics. Thailand, Indonesia, Burma, and South Vietnam became cases that were sometimes cited in the wider comparative literature at this time. Such studies tended to be highly descriptive and revealed little of the internal political dynamics of the societies in question.

The divorce between Southeast Asian political studies and the dominant trends in the discipline in the United States and Europe during the 1960s was enhanced by the emergence of behaviorism as an increasingly dominant mode of analysis in Western political science. During that period many American, European, and Asian scholars were trained in these methods that, however, had little applicability to conducting research in Southeast Asian societies. The quantitative methods of behaviorism were difficult to apply to societies with different cultures, little studied histories and institutions, almost non-existent social statistics, and limited political rights and freedom of expression. The failure of behaviorism in Southeast Asian conditions inevitably drove the discipline toward other issues after the early dominance of nationalism as an analytical theme.

3. *New Themes in Southeast Asian Political Research*

The research interests and themes of scholars of Southeast Asian politics have broadened since the 1960s. The political roles of the bureaucracy, lead-

ership, ideology, and political parties, as well as non-governmental organizations, economic policies and forces, the law, and electoral behavior have emerged as themes which have helped to bring the study of Southeast Asian politics into a more comparative position with the study of politics and political economy elsewhere, particularly other areas of Asia. Most of this research has, however, remained country specific, with explanatory theories often rooted in broad concepts of 'modernization' or political culture and development.

Some of the new themes in Southeast Asian political research began to point to possible theories that were thought possibly applicable to other states in the region. One of the earliest and most influential was the concept of Thailand as a 'bureaucratic polity.' Fred W. Riggs 1966, in his *Thailand: The modernization of a bureaucratic polity*, argued that the bureaucracy, including the army in that country, rather than being an instrument of the government as assumed in American and European studies, was the government. The fact that his hypothesis could trace itself back to the pre-modern paradigms of the early Thai monarchy, produced ideas of an indigenous, Southeast Asian political form which might be applicable to other authoritarian states in the region.

The development of research on patron-client relations, showing the political dependence of peasants and poorer urban elements on powerful 'big' men, and in turn the dependence of the patrons on maintaining the support of their clients in order to amass political power, reinforced many of the themes in Riggs' 'bureaucratic polity' model. What was becoming increasingly clear in research of this kind across the region was that political power in Southeast Asian societies was usually derived from wealth, status, and office, rather than from the meritocratic, democratic, or egalitarian assumptions of most Western political theories. Benedict J. Kerkvliet's *The Huk Rebellion: A Study of Peasant Revolt in the Philippines* (1977) and James C. Scott's *Weapons of the Weak: Everyday Forms of Peasant Resistance* (1985) were particularly influential in this genre.

3.1 Political Economy

Two developments within the economies of some of the countries of the region led to the emergence of theories of politics derived from political economy following the rapid, if partial, industrialization of first Singapore, and then Malaysia, Thailand, and Indonesia from the 1970s. Not only did this economic growth draw people from the countryside into the cities, but it also created stronger political, economic, and cultural linkages with the advanced industrial West as well as Japan, South Korea, and Taiwan. These studies were often criticized, as were earlier explanatory theories, for applying Western-derived, neo-classical assumptions to situations that were still in many ways pre-market, as the bureaucratic dominance of Southeast Asian states distorted 'normal' political forces.

A more fruitful line of research, however, was developed by those scholars who began to study the relationship between economic resources and political power. Anek Laothamatas (1992), for example, in his *Business Associations and the New Political Economy of Thailand: From bureaucratic polity to liberal corporatism*, demonstrated how the rise of commerce and industry in the Thai economy was undermining the earlier dominance of the bureaucracy in Thai politics. Elite transformations, as a consequence of economic development, became a theme for research as illustrated in *The Philippine State and the Marcos Regime* by Gary Hawes (1987).

Considerable interest was attracted in the late 1980s and early 1990s, before the Asian financial crisis of 1997, on the so-called 'developmental state' theories which grew from attempts to understand the rapid industrialization of several East Asian economies. Some arguments of political economists drew inspiration from neo-Marxist dependency theories while others looked to varieties of Wallerstein's 'world system' model. To most students of Southeast Asian politics, however, these theories were unacceptable because of their unwillingness to recognize the dynamics of domestic political forces as well as their Euro-centrism. The rise of arguments of political economy was fueled not only by changes in the region itself, but by intellectual trends in much of social science at this time to explain the 'Asian economic miracle.'

The fact that this phenomenon affected only a minority of the polities of the region meant that the study of states with the least developed economies tended to stagnate. This was paralleled by a relative lack of interest in their politics as well as strong political inhibitions against independent research in the poorest Southeast Asian societies. Thus while the indigenous study of Southeast Asian politics grew stronger in Singapore, Thailand, Malaysia and, to a lesser extent, Indonesia, the discipline remained stagnant, if not nonexistent, in Burma, Laos, Cambodia, Vietnam, and Brunei.

3.2 Democratization

The symbiotic relationship between political and economic change and the growth of an indigenous political science can be seen in the development of research on the requirements for the establishment of democratic regimes in more recent Southeast Asian political studies. Just as the global political and intellectual preoccupations of Western students of Southeast Asian politics in the 1950s and 1960s shaped the early development of the discipline, the re-emerg-

ence of democracy as a global political value in the eyes of the West has shaped recent political research.

Some of this research has been often shaped by externally generated optimistic assessments of the rapid speed at which societies can adapt to new political forms and expectations subsequent to earlier economic and social structural changes. At the same time, the existence of a larger core of local students of politics has ensured that external themes and preoccupations have not driven research away from local realities as happened in earlier years.

The democratization theme has often concentrated on creating dichotomies of political forms (military/civilian, democratic/authoritarian, free market/planned economy) that can tend to obscure an analysis of the actual political forces and interests at work in a society. The volume *The Politics of Elections in Southeast Asia* Taylor R H (ed) (1995) brought together a large amount of research which demonstrated the contradictory realities between the widespread practice of elections in Southeast Asia since independence, if not earlier, and the ability of governments to use democratic institutions to assert a powerful hold by the state over the population and other semi-dependent institutions.

3.3 Public Policy and National Security

Because of the threat of criticism of the political behavior of rulers which is often implicit in studies which examine how the politics of societies actually work, even in the most open of Southeast Asian polities, political scientists have often been forced to shape their findings to fit with what is politically acceptable. While very good research is often conducted, such as the useful volumes on Singapore (Quah et al. 1985, *Government and Politics of Singapore*), and Thailand (Xuto 1987, *Government and Politics of Thailand*) much such work tends to be historically oriented and essentially descriptive of institutions. Official documents and speeches by government leaders are often accepted at face value as the justification of particular policies and courses of action. The structural bases of politics are largely ignored and there is little discussion about where power lies in society and with what consequences.

The first American studies of Southeast Asian politics emerged from area studies centers established at major universities with government and foundation support in order to improve the understanding of the region's politics so as to develop more effective internal and external security policy options. Similarly, there has developed in Southeast Asia since the 1960s a set of national political 'think tanks' from which much political research has been conducted and published.

The first such institution to be established was the Institute of Southeast Asian Studies (ISEAS) in Singapore. Though largely funded and led in its first years by Western foundations and scholars, its role in studying the security issues of Singapore and the region has brought it more closely in contact with government interests. During the Cold War, much of the political research of ISEAS and similar centers that were subsequently established in Kuala Lumpur, Jakarta, and Bangkok, centered on issues of relations with China, other communist powers, and internal security topics.

With the ending of the Cold War, and the rapid industrialization of some of the region's economies, economic policy issues have also come under scrutiny by students of politics in the region. In the style of the Association of Southeast Asian Nations (ASEAN), the region's leading international organization, the scholarship emanating from these institutions refrains from criticizing the policies and politics of neighboring countries.

The strains on regional relations arising from the Asian economic crisis of 1997, as well as the establishment of governments with regard for human rights in the Philippines and Thailand in the latter half of the 1990s, placed further strains on this shared policy of self restraint in research and publication. With the expansion of ASEAN to include all ten countries of Southeast Asia, similar think tanks have been established in Vietnam, Cambodia, Laos, and Myanmar, but the political constraints on their activities means that no genuinely independent research by indigenous scholars on these polities has been published within their countries.

3.4 International Politics

International politics and international relations studies have not been central to the development of Southeast Asian political studies. From the earliest post-independence period, most studies of the region have centered around the themes of the foreign policies of particular states, the relations of one or more of the former superpowers with the region as a whole, or the development of regional with the formation of ASEAN as a long standing regional body. Most studies of each country's foreign policy making and purposes have relied heavily on an analysis of the motivations and interests of national leaders. Foreign policy has been less seen as a national political process, or as the pursuit of national interests, and more often as an extension of the personality of one of the region's long dominant leaders.

4. Discourse Analysis

The diversity of Southeast Asia has made the development of a coherent discipline of Southeast Asian political studies impossible. The field is not a large one and there are few scholars. In the approximately 50 years since the first postcolonial studies began to

emerge, however, much progress has been made. While the early studies tended to be highly descriptive and to be the work of scholars who saw themselves as clearly outside their object of study, more recent and increasingly sophisticated studies by both foreign and indigenous scholars have begun to analyze the languages of politics and their diverse meanings in particular circumstances. The depth of cultural and linguistic knowledge required for such work has required Southeast Asian political studies to become in many ways multidisciplinary. Historians, anthropologists, and students of the region's various literatures have contributed as much to this mode of political apprehension as have political scientists.

5. *Conclusion*

Southeast Asian political studies are still in their infancy despite 50 years of development. From the early descriptive and largely derivative studies of the late 1940s, to the more in-depth, locally aware studies of the 1990s, the field has moved away from a vague and amorphous concept of culture as a catchall for political explanation to an understanding which grows out of the region itself. As the field has become increasingly rooted in South East Asia, it has become increasingly a field of study with implications for the development of the discipline more widely.

Bibliography

Feith H 1962 The *Decline of Constitutional Democracy in Indonesia*. Cornell University Press, Ithaca, NY

Hawes G 1987 *The Philippine State and the Marcos Regime*. University of Hawaii Press, Honolulu, HI

Heine-Geldern R 1942 Conceptions of state and kingship in southeast Asia. *Far Eastern Quarterly* **2**(1): 15–39

Kahin G McT 1952 *Nationalism and Revolution in Indonesia*. Cornell University Press, Ithaca, NY

Kahin G M (ed.) 1964 *Government and Politics in Southeast Asia*, 2nd edn. Cornell University Press, Ithaca, NY

Kerkvliet B J 1977 *The Huk Rebellion: A Study of Peasant Revolt in the Philippines*. University of California Press, Berkeley, CA

Laothamatas A 1992 *Business Associations and the New Political Economy of Thailand: From Bureaucratic Polity to Liberal Corporatism*. Westview Press, Boulder, CO

Quah J S T, Chee C H, Meow S C 1985 *Government and Politics of Singapore*. Oxford University Press, Singapore

Riggs F W 1966 *Thailand: The Modernization of a Bureaucratic Polity*. East-West Center Press, Honolulu

Scott J C 1985 *Weapons of the Weak: Everyday Forms of Peasant Resistance*. Yale University, New Haven, CT

Taylor R H 1992 Political science and south East Asian studies. *South East Asia Research* **1**(1): 5–26

Taylor R H 1995 *The Politics of Elections in Southeast Asia*. Cambridge University Press, New York

Uchia T (ed.) 1984 *Political Science in Asia and the Pacific: State Reports on Teaching and Research in Ten Countries*. UNESCO Regional Office for Education in Asia and the Pacific, Bangkok, Thailand

Xuto S (ed.) 1987 *Government and Politics of Thailand*. Oxford University Press, Singapore

R. H. Taylor

Southeast Asian Studies: Religion

From their first appearances in Southeast Asia, the world religions have been intimately linked to states and statecraft, and so they remain. This organizing thesis will be used to demonstrate different disciplinary approaches to the study of religion. Archaeological and historical research on the world religions in this region reveal, at the same time, the dawn here of state-level polities. Historical research shows the modern nation-states of the region emerging in the context of a colonial order in which the politics of religion were often formative. The predominant approach to the study of religion in contemporary Southeast Asia, however, combines ethnographic with textual studies. Scholars in religious studies, who once might have focused mostly on textual exegesis, increasingly couple this with contextual participation and observations. Anthropologists, who once considered participant observation the only method they needed to master, now realize that they have to be conversant, too, in the textual legacy of a particular faith if they hope to understand people's religious practice.

The renowned diversity of Southeast Asia owes a great deal to religious differences. On the mainland portion, Buddhism predominates, divided between the Mahayana form found in Vietnam (where it both blends and vies with a Confucian tradition), and the Theravada form practiced in Burma, Thailand, Laos, and Cambodia. Insular Southeast Asia is home to the world's largest Muslim country—Indonesia—and Islam holds a commanding presence in Malaysia and the southern Philippines as well. Christianity is the majority faith in the Philippines, however, but many minorities throughout Southeast Asia, especially those located in the mountainous uplands, also follow this faith. Nowhere in the region have these world religions completely displaced older traces of Hindu cosmology or religious practices, older still, based on the veneration of ancestors and the propitiation of local spirits.

The relationship between states and world religions has been reciprocally advantageous: states or rulers gained legitimacy by association with religious symbols; religious institutions benefited materially from the state's largess, gaining the resources for additional propagation or reproduction of the faith. The world religions in Southeast Asia have comprised the primary criterion of civilization, their absence always marking the terrain of the primitive. Textually based, these world religions demanded literacy among at least

some members of a society and entailed systems of formal education. The distribution of the world religions has thus signaled both the limits of state power and of cultural domestication. The animistic religions of upland or relatively isolated peoples have, therefore, been in retreat for two millennia, as states have grown increasingly capable of penetrating daily life at their peripheries as well as in their centers. Religion in Southeast Asia surely does not reduce to politics, but the patchwork distribution of religions in this region and the dynamics of religious change in the present are hardly comprehensible apart from issues of power, especially state power.

1. *Theocratic Formations in Early Southeast Asian States*

The climatic conditions over most of Southeast Asia hamper archaeological preservation. Despite that obstacle to methods based on excavation, inferences from stone architecture and sculptures, as well as inscriptions associated with these, reveal a close relationship between politics and piety (Fontein 1990). One of the most innovative recent attempts to mine architectural evidence is the exhaustive and exacting measurements undertaken at Angkor Wat, measurements which confirm the close relationship between gods, kings, and cosmology in the classic Khmer world (Mannikka 1996).

The earliest polities deserving the term 'state' in this region began to take shape around the first century CE. They have been termed Indic, owing to the predominance of Indian influence in their architecture and religious statuary, yet this was not an instance of conquest or political colonization. Rather, it appears that kingdoms such as Srivijaya, Champa, Khmer, Pagan, Majapahit, and others appropriated Hindu and Buddhist symbolism via contacts that grew along the trade routes that crossed Southeast Asia linking the circum-Mediterranean world to India and China. Religious monuments such as Angkor Wat, the Borobudur, and Prambanan represent only the most spectacular architectural traces of Indianization. Smaller temples and ritual complexes, such as those scattered throughout East and Central Java and along the Vietnam littoral, bear mute witness as well to the spread of these world faiths. Rulers sometimes adopted one variant in contrast to the religions of neighboring or past regimes, a devotion to Shiva, Vishnu, or the Buddha characterizing different Khmer rulers, for example. At other times, even Buddhism and Hinduism were regarded as non-exclusive. Both permitted avenues for deifying the ruler as a god-king.

These early states have been termed mandalas or galactic polities, connoting a focus on an exemplary center where humanly constructed spaces and rituals mirrored cosmological and astronomical orders. Angkor Wat's highest peak represents the *axis mundi*, and its various measurements inside and out encode astronomical cycles, mythical events, and cosmological relationships, sometimes simultaneously (Manikka 1996). Although measurements such as these would have been esoteric, the bas reliefs and statuary, plainly representational and didactic, would have instructed even illiterate persons who wanted to learn or remember stories about the gods. Amassing the materials and especially the labor to accomplish these massive constructions indexed at once a ruler's wealth and power as well as his devotion and piety. More than mere devotees of the gods, however, kings united mystically with them, becoming themselves earthly manifestations of Vishnu or Shiva.

Buddhist rulers harkened back to King Asoka as exemplar, an Indian monarch (reigned 265–238 BC) whose support of the monastic order and construction of reliquaries in the form of stupas set the standard for monarchs building their legitimacy around Buddhism (Swearer 1995, pp. 64–8). Conceiving of the king's position as one accruing from past merit, Buddhists in Pagan (present-day Burma) saw the king also as a *bodhisattva*, one who helps others achieve salvation through his patronage of the religion and through the righteousness of his rule. As such, he was both *devaraja* and *dhammaraja*, god-king and just-king (Aung-Thwin 1985).

Breaking this pattern of fusing ruler and god, Islam filtered into Southeast Asia via India as well as the Arab world, following paths of transoceanic commerce in the tenth century to take root in coastal locations along the region's archipelago and peninsular portions (Chaudhuri 1985). Since Muslims largely shun religious representational art, the archaeology of Islam's advance derives mostly from gravestone inscriptions that reveal the deceased as Muslim (Drewes 1968). Conversion would have facilitated relationships with the Muslim traders who began to dominate the Indian Ocean trade in this era. To have equated ruler with God would have compromised Islam's insistent monotheism, yet the mutually beneficial relationship between state and religion continued to characterize the Muslim sultanates and rajadoms of the Malay-speaking world of insular Southeast Asia. Malay rajas were often described as the 'Shadow of God on Earth,' or his deputy, and legends sometimes relate that when a ruler converted he commanded his followers to do likewise (Milner 1985). Given the Hindu-Buddhist environment, it is not surprising that Islamic Sufism, a mystical tradition, initially had the greatest success in this region (Woodward 1989).

2. *Colonizing the Spirit: Historical and Ethnographic Qualifications*

Spain and Portugal, in competition with each other to gain access to trade with the East, sacralized their exploration and conquests of new worlds as a religious

competition against Protestants (Chadwick 1972, p. 321). In the Philippines, the Church and the colonial state operated through 'an interpenetration of functions' (Steinberg 1982, p. 64) whereby the colonial church enjoyed virtually full financial support from the government, and Spanish monarchs approved or disapproved the placement of religious personnel. With its vast land holdings, ownership of banks, demands for local labor, and increasingly racist limitations on recruiting clergy, the Church was targeted by the *mestizo* revolutionaries who first came to imagine themselves as independent Filipinos. While nationalism became a secular alternative to religion for some leaders, others sought to indigenize or nationalize the Church. Today, 'the Roman Catholic experience' remains important to a sense of Philippine identity, raising a conundrum for the primarily Muslim inhabitants of the southern islands, peoples still termed as the Spanish labeled them 'Moros.' Militant Moro separatism has smoldered in the southern islands for decades.

However, historical approaches have recently qualified this model of conversion and conquest in the Philippines. Tagalog speakers 'translated' Christianity to leave largely intact a distinctly Filipino worldview (Rafael 1988). Fannela Cannell (1999), an anthropologist, found a similar process in Bicol. The records most accessible to the historian, those of the Church and state, are perhaps the most likely to elide this insight. But by including myth as part of historical consciousness, and by ethnographic attention to beliefs about magic and spirits, Aguilar's (1998) history of labor conflicts on sugar haciendas shows how we might read between the lines of the institutional records to discover how Filipinos localized Christianity and resisted colonial domination.

3. *Religion and Nation*

The trading companies of the Dutch and British that eventually displaced Spain and Portugal in the rest of insular Southeast Asia did not perceive their mission to be one of converting the natives. Toward the end of the eighteenth century, as these mercantile bodies died and European states began to engage directly in colonization, independent Protestant missionary organizations also emerged, following European flags wherever these were planted. Protestant missionaries found Islam and Buddhism, deeply entrenched with institutions of governance and leadership, their major competitors throughout much of the region. Missionaries targeted, instead, the tribal peoples who had remained autonomous from the lowland court centers and resistant to their religions. In some cases, colonial governments limited missionaries to working among tribal peoples, cautious of irritating the majority population and its leaders. Indeed, early nationalistic resistance throughout the region drew strength from Islamic and Buddhist organizations (von der Mehden 1963). But perhaps because in the Dutch and British colonies the state was ostensibly (if not always in practice) neutral toward religion, nationalist opposition to these regimes was far less anticlerical than in the Philippines, moving increasingly toward secular-socialist forms.

Thailand was the only country to have remained free from colonial rule, although it was forced to concede considerable territory to France and substantial commercial rights to Britain. The idea of Thailand as a modern nation, an equal player in a world of nations, indicated a shift from conceptualizing the polity as an exemplary center, the edges of which were of little import, to one where mapped boundaries defined precisely the compass of the state's reach (Tongchai 1994).

The Chakri dynasty, selecting Bangkok for its capital, led Thailand during its transition into nationhood. The fourth of the Chakri kings, Mongkut, lived as a Buddhist monk for 27 years before ascending the throne, and advocated reforms to rationalize and purify the *sangha*, the clerical body of Buddhism. His son, Chulalongkorn, carried out extensive governmental and social reforms, but it was the sixth Chakri ruler, Vajiravudh who brought the nation, the monarchy, and Buddhism together in the basic form that they still hold to this day:

> Vajiravudh's Thai nation (and he was the first to popularize this phrase) was founded on the basic triad of 'nation-religion-monarch' (*chat-satsana-phramahakasat*), a trinitarian mystery in which all three elements were inextricably bound together. Allegiance to any one of the three meant loyalty to all three; disloyalty or disobedience or disrespect toward one meant disrespect toward all. (Some sophistries were engaged in to secure the allegiance of Siam's considerable Muslim and Christian minorities Wyatt 1984, p. 229)

Precisely because the state has a vested interest in melding Buddhism with the nation, scholars have not readily seen the extent to which Thai thought and practice is not Buddhist, or perhaps is nominally Buddhist but substantially indigenous. Nicola Tannenbaum, an anthropologist, discovered this in her research on the Shan (1995, 1999). Like other anthropologists, she looked less at Buddhism as textual tradition than at Buddhists and their practices. In contrast, the methods of scholars in religious studies tend to focus more on the canonical aspects of religion than on the everyday 'heresies' of actual practitioners. This methodological difference between religious studies and anthropology is less and less firm, however. Donald Swearer, for example, a scholar in religious studies who specializes in Buddhism in Southeast Asia, approaches his subject ethnographically as well as historically (Swearer 1995).

As Thai monarchs in the past made merit by building or renovating temples and stupas, so the

current monarch, and the Thai state, remain prominent patrons of Buddhism and the *sangha*. One of the more recent royal projects of this type is the large and well-manicured park, Buddha Monthon, located on the edge of Bangkok. Dominated by a gigantic standing image of the Buddha and a large golden stupa, the park includes a lake, fountains, and quiet trails: a serene and inspirational space for public celebrations and high-level conferences or meetings.

During the 1970s, Thailand allied itself with the United States in attempting to halt the spread of communism in Southeast Asia. Fearing its own communist insurgency from within, the government of Thailand utilized the Village Scout Movement to mobilize sentiment and patriotism in support of nation–religion–monarch. Initiation into this movement that attracted mainly teenagers and young adults began with isolating the initiates for several days in a camp-like setting, then depriving them of sleep while they participated in lectures, skits, songs, and other group activities. At the ritual's climax, when initiates received the scarves that officially symbolized their membership in the Scouts, many were sobbing, some so overcome with patriotic emotion they had to be removed on stretchers (Bowie 1997).

At the same moment, Thailand's neighbor, Laos, was taking a path going in the diametrically opposite direction. With the ascendancy of the Lao People's Revolutionary Party (LPRP) in 1975, Laos attempted to forget its monarchy and to harness Buddhism for socialist ends (Evans 1998). Forming the Lao United Buddhist Association, the LPRP hoped to purge the religion of 'superstition' while supporting the state's reforms and programs:

> For many monks and for many laymen this political reorientation of the Lao *sangha* was unacceptable. As a result some left the order and returned to ordinary life, some fled to Thailand, and among the flood of Lao refugees to Thailand from 1975 until the early 1980s, many cited dissatisfaction with changes to Lao Buddhism among their reasons for leaving (Evans 1998, pp. 62–3).

With the reforms that have swept the communist world since 1989, Laos has also instituted a more market-based economy. What socialism and Lao identity mean in this new era have come under new scrutiny, and 'the regime has turned increasingly to Buddhism in its search of new ideologies of legitimation, and of a reformulated Lao nationalism' (Evans 1998, p. 67). Members of the ageing politburo who have died in recent years have been buried with full Buddhist funeral rites, and the upgraded constitution of 1991 gives more explicit attention to Buddhism than its predecessor did.

Islamic Southeast Asia presents examples of two different solutions to the question of religion and nation. When Britain granted independence to Malaysia, Chinese Malaysians, who made up some 30 percent of the population, controlled most of the urban economy. The indigenous Malays, most of whom lived by agriculture or fishing, comprised the scant majority of the population but were granted constitutional privileges of governorship in the new nation. Added to this ethnic mix was a South Asian population comprising about 10 percent of the population, and a small scattering of tribal aborigines in the mountains. Ethnic tensions erupted in 1969 during two weeks of rioting that caused widespread property damage and hundreds of casualties, mostly of Chinese and Indian Malaysians. Responding to this tragedy, the legislature passed economic policies that included affirmative action for Malays in education and business.

These politics of ethnicity have much to do with Islam's position in Malaysia (Shamsul 1994). The constitution, while providing citizens the right to practice and profess any religion, also establishes that 'Islam is the religion of the nation' (Bunge 1985, p. 239). Specifically, Islam is the religion of the Malays. Islam is so centrally part of Malay identity that, historically, throughout much of the Malay-speaking world, converting to Islam was often termed 'becoming Malay' (*masuk Melayu*). The government's promoting Islam in this context of ethnic politics, therefore, promotes 'Malayness.' This explains why so much public architecture in Malaysia, such as the striking railway station in downtown Kuala Lumpur, looks distinctly Islamic.

The relationship of Islam and nation in Indonesia has been quite different from the situation in Malaysia. Numerically speaking, Islam appears to be even more predominant here than in Malaysia, where but 53 percent were Muslim in 1980. Muslims comprise 88 percent of Indonesia's huge population, yet Indonesia's constitution does not name Islam as the religion of the nation. When the new republic was founded, some voices did call for the formation of an Islamic state, but the adherents of minority faiths, as well as those Muslims who feared the power of the state to enforce an orthodoxy they did not want to practice, opposed this concept. To mollify proponents of an Islamic state, however, the deliberative body drawing up the lineaments of a new nation determined at least to avoid a purely secular stance. Thus, the first of the five ideological 'pillars' (*Pancasila*) underlying the nation is the belief in the Almighty God. Significantly, too, the religiously neutral term *Tuhan* (Lord) is used to denote God in this statement rather than the term Allah. A government ministry to support religion was established at the same time, with different branches to promote the interests of each of five officially recognized religions: Islam, Hinduism, Buddhism, Protestant, and Catholic Christianity (Boland 1971).

Because Suharto's New Order came to power as the result of crushing an attempted coup by the Indonesian Communist Party, communism remained the state's 'enemy number one' long after an estimated 500,000

were killed in recrimination for the coup, a bloodbath that effectively ended organized communism in the country. In this political climate, religion remained a significant tool in Indonesian politics (Kipp 1993). Religious affiliation became a cloak of respectability that shielded one from suspicion in that time of terror that ushered in the New Order, and later, that vouched safe one's claim to be a solid and progressive citizen. Adherence to one of these five official religions was required of all government employees, such as teachers and bureaucrats, as proof of one's immunity to communism. Tribal peoples whose animistic religions were not accorded protection and respect found themselves consigned to categories such as 'backward,' or 'isolated.' While adherents of the five approved religions were protected by law from proselytization, the practitioners of animism were opened to Christian and Muslim missions. Some of these tribal peoples yielded to this pressure by converting; others reconstituted their traditional practices as 'Hinduism,' allying themselves with the Balinese, whose religious beliefs and practices bear a family resemblance to Indian religion (Geertz 1972). While Islam was not set up as the religion of state, and Indonesians have been free to be religious in one of five approved ways, they have been far less free to claim allegiance to other kinds of religions, or to claim no religion at all.

The ethnographic study of Islam in Indonesia was long hampered by the assumption that the rituals and practices of Muslims were universal in the Muslim world, and thus already known (Bowen 1993). Until recently, anthropologists working in Muslim Indonesia attended to those religious practices and beliefs that they presumed were non-Islamic or pre-Islamic, and in the process, underestimated the depth of Islamic faith and Muslim identities (Woodward 1989). Now scholars recognize that Islamic practice—everywhere the heart of this faith—deserves careful study, often leading us to the debates and differences that animate and divide Muslim communities.

Indonesian Muslims whose practices incorporate substantial amounts of local precedent, following revered teachers, are termed Traditionalists. They have been gradually losing discursive ground to those who promulgate more modernist, standardized forms of Islam (Bowen 1993). The political economy of this shift can be attributed to the different positions of power occupied by those who espouse the Modernist variants as compared to the largely rural and less well educated Traditionalists. The state has also helped to tip the balance in this discourse. Religious instruction is a mandated part of the curriculum in all public schools through the high-school level, and the Department of Religion has established a nationwide system of teacher-training schools to prepare teachers of religion. Thus, schoolchildren receive a standardized, textbook version of Islam that may contradict the particulars of religious practice in their communities (Hefner 1994).

4. *Both More and Less than an Instrument of State*

As rulers in Southeast Asia from ancient times to the present have attempted to harness the political power of religion, certain forms of religion acquire the privileges of official recognition and financial patronage, while others are at best ignored if not censured. When 'modern state Buddhism' developed under the patronage of the Chakri dynasty in the late nineteenth and early twentieth centuries, it overpowered many regional monastic traditions (Kamala 1997). A tradition of wandering asceticism, among other local Buddhist traditions, has been all but elided from history. Official Thai Buddhism was shaped by the wish to rationalize the faith in relation to a textual canon, whereas localized Thai Buddhism had been rooted in devotional practices, such as meditation, rather than in textual exegesis, and relied more on firsthand relationships between teachers and students than on the literate transmission of knowledge. The official Thai Buddhism also elevated some peculiarly local elements of central Thai experience as the national standard:

> The system—still in use today—rests on degrees, examinations, and ranks in the *sangha* hierarchy. It defines the ideal Buddhist monk as one who observes strict monastic rules, has mastered Wachirayan's printed texts, teaches in Bangkok Thai, carries out administrative duties, observes holy days, and performs religious rituals based on Bangkok customs (Kamala 1997, p. 9).

The considerable numbers of Thai who speak Lao or some other language may thus feel less than comfortable with the orthodoxy deriving from Central Thai traditions.

Religious faith and expression always go well beyond the designs of people in high places. The deepening Islamization of Indonesia and Malaysia in the closing decades of the twentieth century was not wholly an engineered artifact of the Cold War and internal ethnic politics. Muslims worldwide were experiencing a resurgence of faith and confidence, and as travel and communication have reduced the world's distances, Southeast Asia's Muslims easily participate in discourse within the worldwide *ummah*, the community of Islam. What Indonesians and Malaysians do with or make of their faith is not fully under any government's control. A small handful of Muslim leaders in Indonesia, authorities worry, inspire a 'fanaticism' among their followers that could threaten the peace and security of the regime. Other prominent Muslim leaders want to depoliticize Islam. They have become wary of the government's patronage of Islam, seeing there an avenue for political cooptation and control. They disavow the goal of an Islamic state, calling instead for the deep transformation of society through individuals' adherence to the practices of Islam: a revolution of the spirit rather than of the political structure.

Buddhism in Southeast Asia has experienced similar revivals and reinventions, some of which go well beyond what the state intends. In Singapore, for example, with an explicitly secular constitution and state ideology, Buddhism replaced Taoism as the leading religion according to the 1990 census. Over one-third of Singaporeans profess Buddhism, and it attracts especially university-educated adherents who attend lectures and youth camps, and who utilize the Internet to study Buddhist philosophy. Likewise, urban and educated Thai have become increasingly skeptical of the Buddhism purveyed as a civil religion. New, unorthodox religious movements based on Buddhism have attracted many in the Thai middle class, and, what is more, some monks have dared to oppose openly, and at the risk of arrest, the state's development policies that destroy the environment (Taylor 1993).

Archaeology and art history reveal that the world religions were tied intimately to statecraft in the region's first states. Historical research tells us that religious imperialism was implicated in colonial expansion, and that the dawn of nationalistic movements sometimes drew energy from religious identities forged in opposition to the colonials. Modern research on the world religions in Southeast Asia, whether undertaken by those in religious studies, anthropology, or history takes seriously both the literate tradition of a faith as well as the imperative to find out what the faithful actually do and say. These hybrid methods add nuance to histories of religion, showing us, for example, how people translate religious concepts into locally meaningful forms, resisting religious and political imperialism in the process. At the same time, such hybrid approaches demand that anthropologists investigate the supposedly universal rituals of the world faiths and then place their knowledge of local practices against a more expansive background of regional history and religious literature.

See also: East Asia, Religions of; India, Religions of; Religion: Definition and Explanation; Religion: Evolution and Development; Religion: Family and Kinship; Religion, History of; Religion: Nationalism and Identity; Religion, Sociology of; Southeast Asian Studies: Culture; Southeast Asian Studies: Politics; Southeast Asian Studies: Society

Bibliography

Aguilar F 1998 *Clash of Spirits: History of Power and Sugar Planter Hegemony on a Visayan Island*. University of Hawaii Press, Honolulu, HI

Aung-Thwin M 1985 *Pagan: The Origins of Modern Burma*. University of Hawaii Press, Honolulu, HI

Boland B J 1971 *The Struggle of Islam in Modern Indonesia*. Nijhoff, The Hague, The Netherlands

Bowen J R 1993 *Muslims Through Discourse*. Princeton University Press, Princeton, NJ

Bowie K A 1997 *Rituals of National Loyalty: An Anthropology of the State and the Village Scout Movement in Thailand*. Columbia University Press, New York

Bunge F M 1985 *Malaysia: A Country Study*. United States Government, Washington, DC

Cannell F 1999 *Power and Intimacy in the Christian Philippines*. Cambridge University Press, Cambridge, UK

Chadwick O 1972 *The Reformation*. Penguin Books, Harmondsworth, UK

Chaudhuri K N 1985 *Trade and Civilization in the Indian Ocean: An Economic History from the Rise of Islam to 1750*. Cambridge University Press, Cambridge, UK

Drewes G W J 1968 New light on the coming of Islam to Indonesia? *Bijdragen tot de Taal-, Land- en Volkenkunde* **124**(4): 433–59

Evans G 1998 *The Politics of Ritual and Remembrance: Laos Since 1975*. Silkworm Books, Chiangmai, Thailand

Fontein J 1990 *The Sculpture of Indonesia*. National Gallery of Art, Washington, DC

Geertz C 1972 'Internal Conversion' in contemporary Bali. In: Geertz C (ed.) *The Interpretation of Cultures*. Basic Books, New York

Hefner R W 1994 Reimagined community: A social history of Muslim education in Pasuruan, East Java. In: Keyes C F, Kendall L, Hardacre H (eds.) *Asian Visions of Authority: Religion and the Modern States of East and Southeast Asia*. University of Hawaii Press, Honolulu, HA

Kamala T 1997 *Forest Recollections: Wandering Monks in Twentieth-century Thailand*. University of Hawaii Press, Honolulu, HA

Kipp R S 1993 *Dissociated Identities: Ethnicity, Religion, and Class in an Indonesian Society*. University of Michigan Press, Ann Arbor, MI

Mannikka E 1996 *Angkor Wat: Time, Space, and Kingship*. University of Hawaii Press, Honolulu, HI

Milner A R 1985 Islam and Malay kinship. In: Ibrahim A, Siddique S, Hussain Y (eds.) *Readings on Islam in Southeast Asia*. Institute of Southeast Asian Studies, Singapore

Rafael V 1988 *Contracting Colonialism: Translation and Christian Conversion in Tagalog Society Under Early Spanish Rule*. Ateneo de Manila University Press, Manila, The Philippines

Shamsul A B 1994 Religion and ethnic politics in Malaysia: The significance of the Islamic resurgence phenomenon. In: Keyes C F, Kendall L, Hardacre H (eds.) *Asian Visions of Authority: Religion and the Modern States of East and Southeast Asia*. University of Hawaii Press, Honolulu, HI

Steinberg D J 1982 *The Philippines: A Singular and a Plural Place*. Westview Press, Boulder, CO

Swearer D K 1995 *The Buddhist World of Southeast Asia*. State University of New York Press, Albany, NY

Tannenbaum N 1995 *Who Can Compete Against the World? Power-protection in Shan Worldview*. Association for Asian Studies, Inc., Ann Arbor, MI

Tannenbaum N 1999 Buddhism, prostitution, and sex: Limits on the academic discourse on gender in Thailand. In: Jackson P, Cook N (eds.) *Genders and Sexualities in Modern Thailand*. Silkworm Books, Chiangmai, Thailand

Taylor J 1993 Buddhist revitalization, modernization, and social change in contemporary Thailand. *Sojourn* **8**(1): 62–91

Tongchai W 1994 *Siam Mapped: A History of the Geo-body of a Nation*. University of Hawaii Press, Honolulu, HI

Von der Mehden F R 1963 *Religion and Nationalism in Southeast Asia*. University of Wisconsin Press, Madison, WI
Woodward M R 1989 *Islam in Java: Normative Piety and Mysticism in the Sultanate of Yogyakarta*. University of Arizona Press, Tucson, AZ
Wyatt D K 1984 *Thailand: A Short History*. Silkworm Books, Chiangmai, Thailand

R. S. Kipp

Southeast Asian Studies: Society

1. *Analyzing Southeast Asia as a Form of Knowledge*

Society is both real and imagined. It is real through face-to-face contact and imagined when the idea of its existence is mediated through mediums such as printed materials and electronic images. So, the term society refers simultaneously to a microunit that we could observe and to a macro one that we could only partially engaged with. We, therefore, have observable 'societies' within a macro imagined 'society,' so to speak. Southeast Asia, like other regions in the world, has both. But it is the way that both of these components have been weaved into an enduring complex whole, which seemed to have made Southeast Asia and Southeast Asians thrive and survive even under adverse conditions, such as the recent financial–economic crisis, that has become the source of endless intellectual attraction and academic inquiry to both scholars and others, hence the birth, growth, and flourishing of Southeast Asian studies.

Thus, Southeast Asian studies, dominated by humanities and the social sciences, have been about the study of the 'society' and 'societies' in the region, in their various dimensions, in the past and at present. The complex plurality of these 'society' and 'societies,' or societal forms, that do indeed co-exist, endure, and enjoy some functional stability, have made it imperative for researchers to apply an equally diverse set of approaches, some discipline-based (anthropology, sociology, geography, history, political science, etc.) and others thematically-oriented (development studies, gender studies, cultural studies, etc.) in studying Southeast Asian society. In some cases, it even involved disciplines from the natural as well as applied sciences.

The greatest challenge in Southeast Asian studies, and to its experts, has been to keep pace with the major changes that have affected the 'society' and/or 'societies' and then narrate, explain, and analyze these changes and present the analysis in a way that is accessible to everyone within and outside the region. Therefore, framing the analysis is critical in understanding as to how Southeast Asian studies constitute and reproduce through the study of 'society' and 'societies' within Southeast Asia. The 'knowledge baseline' approach is useful in making sense of the said framing process.

2. *The 'Knowledge Baseline' in Southeast Asian Studies*

Social scientific knowledge (humanities included) on Southeast Asia has a clear knowledge baseline, meaning a continuous and interrelated intellectual-cum-conceptual basis, which emerged from its own history and has, in turn, inspired the construction, organization and consumption process of this knowledge. The two popular concepts that have been used frequently to characterize Southeast Asia are 'plurality' and 'plural society,' both of which are social scientific constructs that emerged from empirical studies conducted within Southeast Asia by scholars from outside the region.

In historical terms, 'plurality' characterizes Southeast Asia before the Europeans came and who, subsequently, divided the region into a community of 'plural societies.' Plurality here signifies a free-flowing, natural process not only articulated through the process of migration but also through cultural borrowings and adaptations. Politically speaking, polity was the society's political order of the day, a flexible nonbureaucratic style of management focusing on management and ceremony by a demonstrative ruler. States, governments, and nation–states, which constitute an elaborate system of bureaucratic institutions, did not really exist until Europeans came and dismantled the traditional polities of Southeast Asia and subsequently installed their systems of governance, using 'colonial knowledge,' which gave rise to the plural society complex.

Historically, therefore, plural society signifies both 'coercion' and 'difference.' It also signifies the introduction of knowledge, social constructs, vocabulary, idioms, and institutions hitherto unknown to the indigenous population (such as maps, census, museums, and ethnic categories), the introduction of market-oriented economy and systematized hegemonic politics. Modern nation–states or state–nations in Southeast Asia have emerged from this plural society context.

It is not difficult to show that the production of social scientific knowledge on Southeast Asia has moved along this plurality-plural society continuum. When scholars conduct research and write on pre-European Southeast Asia they are compelled to respond to the reality of Southeast Asian plurality during that period; a period which saw the region as the meeting place of world civilizations and cultures, where different winds and currents converged bringing together people from all over the world who were interested in 'God, gold and glory' and where groups

of indigenes moved in various circuits within the region to seek their fortunes. As a result, we have had, in Java, a Hindu king with an Arabic name entertaining European traders. In Champa, we had a Malay raja ruling a predominantly Buddhist populace trading with India, China, and the Malay archipelago. Whether we employ the orientalist approach or not, we cannot avoid writing about that period within a plurality framework, thus emphasizing the region's rich diversity and colorful traditions. In other words, the social reality of the region to a large extent dictates our analytical framework.

However, once colonial rule was established and the plural society was installed in the region, followed later by the formation of nation–states, the analytical frame, too, changed. Not only did analysts have to address the reality of the plural society but also the subsequent developments generated by the existence of a community of plural societies in the region. We began to narrow our analytical frame to nation–state, ethnic group, internation-state relations, intranation-state problems, nationalism, and so on. This gave rise to what could be called 'methodological nationalism,' a way of constructing and using knowledge based mainly on the 'territoriality' of the nation–state and not on the notion that social life is a universal and borderless phenomenon, hence the creation of 'Indonesian studies,' 'Malaysian Studies,' 'Thai Studies,' and so on.

With the advent of the Cold War and the modernization effort analysts became further narrowed in their frame of reference. They began to talk of poverty and basic needs in the rural areas of a particular nation, also focusing on resistance and warfare, slums in urban areas, and economic growth of smallholder farmers. The interests of particular disciplines, such as anthropology, became narrower still when it only focuses on particular communities in remote areas, a particular battle in a mountain area, a failed irrigation project in a delta, or gender identity of an ethnic minority in a market town.

In fact, in numerical terms, the number of studies produced on Southeast Asia in the plural society context supercedes many times those produced on Southeast Asia in the plurality context. Admittedly, social scientific studies about Southeast Asia developed much more rapidly after the Second World War. However, the focus became increasingly narrow and compartmentalized not only by academic disciplines but also in accordance to the boundaries of modern postcolonial nations. Hence, social scientific knowledge on Southeast Asia became, to borrow a Javanese term, *kraton*ized, or compartmentalized.

It is inevitable that a substantial amount of social scientific knowledge about Southeast Asia itself, paradigmatically, has been generated, produced, and contextualized within the plural society framework, because 'nation–state' as an analytical category matters more than, say, the plurality perception of the Penans of Central Borneo, who, like their ancestors centuries ago, move freely between Indonesia and Malaysia to eke out a living along with other tribal groups and outside traders, ignoring the existence of the political boundaries. In fact, anthropologists seem to have found it convenient, for analytical, scientific, and academic expedience, to separate the Indonesian Penans from those of Malaysia when, in reality, they are one and the same people.

Therefore, the plurality–plural society continuum is not only a 'knowledge baseline' but also a real-life social construct that was endowed with a set of ideas and vocabulary, within which people exist day-to-day in Southeast Asia.

3. *Constituting and Reproducing the Knowledge on Southeast Asia*

There are at least four major axes along which the construction, organization, and reproduction of social scientific knowledge about Southeast Asia and its societies have taken place.

The first axis is that of *discipline/area studies*. There is an ongoing debate between those who prefer to approach the study of Southeast Asia from a disciplinary perspective, on the one hand, and those who believe that it should be approached from an area studies dimension, employing an interdisciplinary approach, on the other.

The former prefer to start clearly on a disciplinary footing and treat Southeast Asia as a case study or the site for the application of particular set of theories that could also be applied elsewhere globally. The aim of such an approach is to understand social phenomena found in Southeast Asia and to make comparisons with similar phenomena elsewhere. Those preferring the latter approach see Southeast Asia as possessing particular characteristics and internal dynamics that have to be examined in detail using all available disciplinary approaches with the intention of unraveling and recognizing the indigenous knowledge without necessarily making any comparison with other regions of the world.

The bureaucratic implications of these two approaches can perhaps be discerned clearly in the way social scientific knowledge about Southeast Asia is reproduced through research and teaching. This brings us to the second axis, namely, the *undergraduate/graduate studies* axis.

Those who favor area studies often believe that Southeast Asian studies can be taught at the undergraduate level hence the establishment of Southeast Asian studies departments or programs, in a number of universities in Southeast Asia, combining basic skills of various disciplines to examine the internal dynamics of societies within the region. Acquiring proficiency in one or two languages from the region is

a must in this case. The problem with this bureaucratic strategy is that these departments have to be located in a particular faculty, say, in the arts, humanities, or social science faculty. This denies, for instance, those with a background in the natural sciences the opportunity to study in-depth about Southeast Asia.

Therefore, those discipline-inclined observers would argue that Southeast Asian studies should be taught at the graduate level to allow those grounded in the various disciplines, whether in the social or natural sciences or in other fields of study, to have an opportunity to specialize in Southeast Asian studies. Therefore, a geologist or an engineer who, for instance, is interested in the soil and irrigation systems of Southeast Asia could examine not only the physical make-up of Southeast Asia but also the human–environment relationship. This is particularly relevant at the present time since environmental and ecological issues have become global concerns.

This has made many individuals, institutions, and governments think carefully how they should invest their precious time and money when they are requested to support the setting up of, say, a program, center, or institute of Southeast Asian studies. They often ask whether universities should continue to have the prerogative on the teaching, research, and dissemination of knowledge about anything connected with Southeast Asia and its societies. Why not in nonuniversity institutions?

This takes us to the third axis, namely, the *university/nonuniversity* one. For many years, we imagined that only at the university we could acquire and reproduce knowledge about Southeast Asia, whether approached from the disciplinary or area studies perspective. However, many governments and international funding bodies felt that to obtain knowledge about Southeast Asia one need not go to university, but could acquire it through nonacademic but research-oriented institutions established outside the university structure to serve particular purposes. National research bodies such as LIPI (Indonesian Institute of the Sciences) in Jakarta and ISEAS (Institute of Southeast Asian Studies) in Singapore have been playing that role. 'Think-tanks,' such as the Center for Strategic Studies (CSIS), Jakarta, or the Institute of Strategic and International Studies (ISIS), Malaysia, have also played the role of the producer and reproducer of knowledge on societies in Southeast Asia outside the university framework. However, there seems to be a division of labor, based on differences in research orientation, in the task of producing and reproducing knowledge between the academic and nonacademic institutions.

The final axis is an *academic/policy-oriented research* one. While academic endeavors pursued within the context of Southeast Asian studies in the universities are motivated by interest in basic research, which is by definition scholarly, those pursued outside the universities are often perceived as not being scholarly enough because they are essentially applied or policy-oriented in nature and serving rather narrow, often political, interests of the powers that be in Southeast Asia.

It is argued that the critical difference between these two approaches is that the academic one is always open to stringent peer-group evaluation as a form of quality control, but that the applied one is not always assessed academically. In fact, the latter is often highly confidential and political in nature, thus denying it to be vetted by the peer group, hence its perceived inferior scholarly quality. The basic research-based academic endeavors are, therefore, seen as highly scholarly, whereas the nonacademic ones are perceived as highly suspect as scholarly works and not considered to contribute to the accumulation of knowledge on Southeast Asia societies.

However, research institutes like ISEAS in Singapore would argue that, even though it is essentially a policy-oriented research institute mainly serving the interests of the Singapore government, it still produces scholarly work of high quality and encourages basic research to be conducted by its research fellows either on an individual or a group basis. In other words, a nonuniversity research institute of Southeast Asian studies, such as ISEAS, could simultaneously conduct applied and basic research without sacrificing the academic and scholarly qualities of its final product; or put in another way, it is 'policy-oriented yet scholarly.'

The moot question is who are really the consumers of knowledge on Southeast Asians societies, hence Southeast Asian studies; the Southeast Asians or outsiders?

4. *Consuming the Knowledge on Southeast Asia*

It could be argued that social scientific knowledge about Southeast Asia and its societies is a commodity with a market value. Often the 'market rationale,' and not the 'intellectual rationale,' prevails in matters such as the setting-up of a Southeast Asian studies program, center, or institute, even in the government-funded academic institutions. However, the funding of research on Southeast Asian studies has often been dictated not by idealistic, philanthropic motives but by quite crass utilitarian desires, mainly political or economic ones. There are at least three important 'sectors' within which knowledge on Southeast Asia societies has been consumed: the public, the private, and the intellectual sectors.

Since the governments in Southeast Asia have been the biggest public sector investors in education, through public-funded educational institutions, they have been the largest employment provider. They have set their own preferences and priorities, in accordance to their general framework of manpower planning, in deciding what type of graduates and in which fields of

specialization they want to employ them. The pattern in Southeast Asian countries has been well established, that is, there is a higher demand for science graduates than the social sciences and the humanities. But amongst the latter there is no clear, expressed demand for Southeast Asian studies graduates. However, there seems to be a significant demand for the inclusion of the Southeast Asian studies content in all the non-natural science courses at the undergraduate level in most of the government-funded academic institutions in Southeast Asia. This is not unrelated to the fact that the awareness about ASEAN as a community has now become more generalized amongst the public hence the need for a more informed description on the different countries and societies within ASEAN (read Southeast Asia).

Outside Southeast Asia, such as in Japan and the United States of America, very rarely, specialization in Southeast Asian studies, or components of, has been considered highly desired in the job market of the public sector. Perhaps having a graduate-level qualification in Southeast Asian studies is more marketable in the public sector especially in government of semi-government bodies that deals with diplomatic relations or intelligence.

In the private sector, the demand for Southeast Asian studies as a form of knowledge and the demand for a potential employer who possesses that knowledge are both limited and rather specific. However, the number could increase depending on how large is the investment and production outfit a particular company has in Southeast Asia. Since some of the demand for the knowledge is rather short-term, often specific but detailed, therefore it has to be customized to the needs of a company, 'think-tanks,' or 'consultant companies' have often become the main supplier of such tailored knowledge. Many of such organizations are actually dependent on 'freelance' Southeast Asianists or academics who do such jobs on a part-time, unofficial basis.

It has been observed that the Japanese seem to be regular consumers of knowledge on Southeast Asia. This is hardly surprising because they have massive investments in Southeast Asia. There is, therefore, a constant need to know what is happening in the region. Research foundations from Japan, in particular the Toyota Foundation, have been very active, since the 1990s, in promoting 'Southeast Asian studies for Southeast Asians,' and supporting other research and exchange programs. Taiwan and Korea are the two other Asian countries having their own Southeast Asian studies research centers, along with the United States, United Kingdom, France, and The Netherlands, former colonial powers in Southeast Asia.

A more generalized demand for knowledge on Southeast Asian societies relates to marketing and this trend must not be underrated with the recent expansion of the middle class in the region. As the markets and clients in Southeast Asia become more sophisticated the need for in-depth knowledge on sectors of the Southeast Asia societies has increased. This in turn has increased the demand for graduates who have followed courses related to Southeast Asian studies.

In the intellectual sector, knowledge on Southeast Asia has been consumed generally by the NGOs, namely, those that are national-based ones as well as those that have regional networks. Because most of the NGOs are issue-specific based interest group, such as environmental protection, abused housewives, social justice, and the like, and often seeking funds for their activities from the governments and NGOs in developed countries, they find it more advantageous to operate on a regional basis because they get more attention and funding from the said source. The strength and success of their operation is very much dependent on the amount of knowledge they have about Southeast Asian societies in general as well as the specific issue that they are focusing on as a cause in their struggle.

With the popularity of the Internet and its increased usage around the world and within Southeast Asia, it has now become an important medium through which academic and popular knowledge on Southeast Asian societies has become available. The source of the knowledge could be located outside or within the region but are now much more accessible for commercial and noncommercial purposes.

As a conclusion, it could be said that Southeast Asian studies and what it constitutes is, first and foremost, a knowledge construct that represents only part of the region's social reality. In spite of this, it is the most important element, amongst the many, that gives Southeast Asia, the geo-physical region as well as its people and environment, its history, territory, and society. Because of the co-existence of different societal forms in the region, hence the unevenness of the tempo of social life in the region, the speed of social change thus also differs from one community to the other, from one area within the region to another. Only a multidisciplinary approach could capture these complexities embedded in the societies of Southeast Asia.

As the importance of the region increases in the globalizing world, both generalist and specialist knowledge about Southeast Asia become critical to the world and the region itself. In that sense, Southeast Asian studies as a knowledge construct transforms itself into a lived reality, especially for the Southeast Asians themselves. This knowledge, therefore, becomes indispensable both to those who study Southeast Asia and its society as well as to the Southeast Asian themselves.

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Modernization, Sociological Theories of; Plural Societies; Southeast Asian Studies: Culture; Southeast Asian Studies: Economics; Southeast Asian Studies: Politics

Bibliography

Bellwood P 1985 *Prehistory of Indo-Malayan Archipelago*. Academic Press, Sydney

Brown D 1994 *The State and Ethnic Politics in South-East Asia*. Routledge, London & New York

Collins J T 1996 *Malay, World Language: A Short History*. Dewan Bahasa & Pustaka, Kuala Lumpur

Evers H-D (ed.) 1980 *Sociology of South-East Asia: Reading on Social Change and Development*. Oxford University Press, Kuala Lumpur

Kemp H 1998 *Bibliographies on Southeast Asia*. KITLV Press, Leiden

Reid A 1988 *Southeast Asia in the Age of Commerce 1450–1680; Volume 1: The Lands Below the Winds*. Yale University Press, New Haven, CT

Reid A 1993 *Southeast Asia in the Age of Commerce 1450–1680; Volume 2: Expansion and Crisis*. Yale University Press, New Haven, CT

Steinberg J (ed.) 1987 *In Search of Southeast Asia: A Modern History*, rev. edn. Allen & Unwin, Sydney and Wellington

Tarling N (ed.) 1992 *The Cambridge History of Southeast Asia. Volume 1: From Early Times to c. 1800; Volume 2: The Nineteenth and Twentieth Centuries*. Cambridge University Press, Cambridge, UK

Wallace A R 1869 *The Malay Archipelago, the Land of the Orang-Utan and the Bird of Paradise: A Narrative of Travel with Studies of Man and Nature*. Macmillan, London

A. B. Shamsul

Southern Africa: Sociocultural Aspects

The southern African region comprises South Africa, Lesotho, Swaziland, Botswana, Namibia, Mozambique, Malawi, Zimbabwe, and Zambia. The recent history of the region, during the last half of the twentieth century, was dominated by the decolonization process and its aftermath. With a few exceptions, countries have struggled to establish viable political systems and bureaucratic states, to overcome ethnic, racial, and regional conflict, to diversify and stabilize economies often dominated by only one or two industries (such as copper mines in Zambia, tobacco in Zimbabwe), and to achieve economic development. The small countries of Lesotho, Botswana, and Swaziland are, for the most part, monoethnic states with little ethnic conflict, but have had significantly different internal political arrangements and economic fortunes. South Africa is the most economically and ethnically diverse country in the region, with the strongest state and institutions of civil society.

It dominates the region's economy. Lesotho is entirely surrounded by South Africa and almost completely dependent on it. All other countries, especially land-locked Zimbabwe, Zambia, and Malawi, depend to a large degree on South African transport services and harbors. South African investments and companies increasingly dominate the economies of its regional neighbors.

The southern African region shares similar and inter-related colonial histories, postcolonial interactions, and the broad outlines of the cultural histories of the majority of their populations. Close cooperation among governments of the region fostered by a half-century of political struggle has created tight political integration of the whole. The largest trading partners still remain outside of the region, primarily in Western Europe. At the beginning of the twenty-first century, all countries had begun to seek increasing intraregional cooperation in order to reverse the effects of conflict, isolation, economic decline, and political weakness of state structures. This was articulated by Thabo Mbeki, President of South Africa, as the search for an African Renaissance.

1. History

Archaeological evidence shows that southern Africa was 'the cradle of mankind.' The earliest prehumans and *Homo sapiens* evolved here. There is, however, no evidence that would link these early humans any more strongly to contemporary southern African populations than to the rest of humanity.

All languages in the region derive from only two of the six major language families of Africa (Greenberg 1966). The entire population of speakers of the Khoisan language group have lived in the region, and only here, from at least 20,000 years ago. Beginning around 2,000 years ago, Bantu-speaking peoples appear to have moved into the region quite rapidly from north of the Congo forest zone, in the region of today's Cameroun. They possessed technology to smelt iron and make durable fired pottery, and raised cattle, goats, and a mixture of crops from the Sudanic zone (millets and sorghum) and from the Indian Ocean region (sweet potatoes, bananas). At some time they acquired the New World crop, maize, which became central to the diet of the entire region well before European incursions.

Before the arrival of the agricultural and pastoral Bantu, most of southern Africa was sparsely populated by the Khoisan people. The 'Khoisan' language family includes two major branches, the 'Bushman' or 'San' languages and the KhoiKhoi. The primary distinction between these two groups, however, is social and economic. The Bushman were hunters and gatherers, living in small mobile bands of no more than 35–40 people. The Khoi were cattle and sheep-herders, living in larger tribal groupings. Early European settlers at the Cape of Good Hope began to trade with the Khoi peoples (then known as 'Hottentots') in the seventeenth century, but as trade relationships broke down and the European farmers ('Boers') began to move inland in the eighteenth century, the Khoi were

gradually absorbed as clients and laborers by the Boers. Bushman were killed in skirmishes, displaced to remote deserts, or incorporated as servants and wives by European people moving inland from the south, and by Bantu-speaking Africans moving southwards. By the nineteenth century Khoisan speakers remained only in the dry areas inland from the west coasts of South Africa, Namibia, and Angola, and in the Kgalagadi (Kalahari) desert of Botswana. There is still a relatively large Nama (Khoi) speaking group in Namibia. Indeed, the name 'Namibia' means 'land of the Nama.' In South Africa where 'Colored' people preserve much of the San and Khoi genetic heritage, Bantu languages such as Xhosa and Zulu preserve their linguistic heritage in the form of some vocabulary and the characteristic 'click' sounds found nowhere else among the world's languages. Their art, painted on rocks throughout the region, gives testament of a rich culture.

Although 'Bantu' refers strictly to a linguistic category, people who speak Bantu languages also share broadly similar cultures. With the exception of their spread to Indian Ocean islands such as the Comores and Zanzibar, they are restricted to tropical and subtropical continental southern, central, and eastern Africa. The culture and political economy of early Bantu speakers was characterized by residence in small dispersed agricultural and/or pastoral settlements led by a 'chief' who delegated authority to subchiefs or headmen, and to a council of senior or royal males in a form of informal democracy that usually excluded women. Precolonial religious beliefs focused on reverence for the spirits of (male) ancestors who were propitiated during regular ceremonies by individuals, families, clans, or larger tribal or village groupings. While there were generally no 'priests' in these traditional religious systems, all societies relied on diviners, healers, and prophets who acted either as mediums for spirits during healing rituals or occasionally interpreted the will of the spirits in a political context. The institution of religious prophecy continues to have significant political and social effects. The role of traditional prophets and spirit mediums in the *Chimurenga* or 'war of liberation' fought in Rhodesia/Zimbabwe was especially important, but resistance movements led by traditional prophets appeared in all southern African countries during colonial times as well. Strong beliefs in the existence of witches—individuals who caused evil or illness often unbeknown to themselves—and beliefs in the existence of animal familiars of witches or other kinds of spiritual powers inherent in plants and animals are also widespread. Today, majorities in all countries are at least nominally Christian, with small numbers of Muslims such as the Yao of Mozambique, and the Swahili-speakers in Mozambique, Malawi, and eastern Zambia. Many aspects of precolonial belief have been incorporated into indigenous Christian beliefs and practices, while minorities of people continue to practice traditional religion in preference to, and sometimes in combination with, Christianity or Islam.

The rapid establishment after 1816 of the Zulu kingdom under the warrior king Shaka, together with the British colonial advance in the latter half of the nineteenth century, provide the historical skeleton that unites the region's cultural diversity. Shaka was a member of the small Nguni-speaking 'Zulu' clan. He is credited with introducing a number of military innovations that eventually led to the establishment of his hegemony over a large conquest state in what is now northern KwaZulu-Natal Province of South Africa. Repercussions were felt as far north as Malawi and eastern Zambia, southern Tanzania and Mozambique, where groups of people fleeing Shaka's power established temporary hegemony. These people were known as the 'Ngoni' in the north, or Nguni in South Africa.

The rise of the Zulu Kingdom, together with movements inland of the Portuguese from the east coast, and the Boers from the Cape of Good Hope, led to general disruption of settled life in south central Africa. This, in turn, created opportunities for widespread incursions, first by African and Arab slave and ivory traders from the east coast, and later for European adventurers, hunters, traders, settlers, and missionaries. The general lack of strong centralized states or other large-scale political organizations in the region contributed to the advance of colonial rule and expropriation of land by Europeans for farms and mines. European Christian missionaries established mission stations in Nyasaland (now Malawi) in response to the devastation and social dislocation caused by Ngoni incursions and led the way for the British government to declare Nyasaland a Protectorate. Zulu offshoots established the Gaza kingdom in the southern central region of Portuguese Mozambique over the Tsonga-speaking people and settled in large numbers in southern Zimbabwe as the 'Ndebele' people. (The Ndebele of South Africa derive from an earlier Nguni expansion into Sotho-Tswana areas, and are historically and culturally distinct from the Zimbabwe Ndebele.) The Kololo people, a Tswana-speaking group originally deriving from the Hurutshe in what is now Botswana, fled north in response to Zulu incursions and established temporary hegemony over the Lozi people in western Zambia. They managed to impose much of their language and culture on the previously independent populations. In South Africa, the Zulu expansion and its repercussions had the effect of disrupting virtually all previous settlements of people during the *Mfecane* (or *Difeqane* in Sotho, 'the crushing'), a period of vast population movements and reorganization lasting through the middle of the nineteenth century. Out of this emerged several strongly integrated 'tribes' as followers of powerful chiefs: the Sotho of Basutoland (Lesotho), the Swazi of Swaziland, and the Tswana tribes of

Bechuanaland (Botswana). It also created an opportunity for white Afrikaans-speaking farmers, the Boers, to move north out of the Cape Province, a British Colony since 1802. After the Great Trek of 1835 they established several independent republics, including the Orange Free State and the Suid Afrikanse Republiek ('South African Republic,' known as 'The Transvaal').

2. *Colonialism and Independence Struggle*

In the second half of the century, Swaziland, Basutoland (now Lesotho), and Bechuanaland (Botswana) were declared British Protectorates after treaties were signed by their chiefs, partly in order to secure them as labor reserves to provide labor for European farmers and the mines, and partly to keep them from the hegemony of the Zulu on the one hand and the Afrikaner (Boer) states on the other. The Zulu, unwilling to enter into negotiations, were eventually defeated in 1879 by British forces. Their king was exiled briefly but eventually returned as king of the Zulu people under the authority of the Colony of Natal. The hyphenated name of the current province of South Africa, 'KwaZulu-Natal,' still reflects the regional power of the Zulu kingdom.

Mozambique was colonized by Portugal in the fifteenth century but only came under effective control in the late nineteenth century. Namibia was brought under German control as a result of the partition of Africa at the Berlin Conference in 1884. The remainder eventually came under the direct or indirect control of Britain. Zimbabwe and Zambia were given as concessions to the British South Africa Company under Cecil John Rhodes as Southern and Northern Rhodesia. In 1924, the British South Africa company was dissolved, and Northern Rhodesia and Southern Rhodesia became Crown Colonies under local white settler control. Together with the Protectorate of Nyasaland, they joined together in the Central African Federation in 1951 until independence was achieved by Northern Rhodesia and Nyasaland in 1964. The colony of Northern Rhodesia was renamed Zambia, while Nyasaland became Malawi.

Military conflicts in the later decades of the twentieth century severely limited the pace of economic and political development throughout the region. Zimbabwe/ Rhodesia, Namibia/South West Africa, and South Africa witnessed protracted guerrilla struggles between white settler populations and black African nationalist organizations that spanned most of the half-century and generated large cross-border refugee movements.

Southern Rhodesia continued to be ruled by its white minority settler population as Rhodesia, and declared its independence from Britain in 1965. Though unconstitutional and condemned by Britain and the United Nations, the largely white government continued to rule Rhodesia until 1980, despite massive trade and political sanctions. After the Zimbabwe African National Union (ZANU), with an ethnic base among the Shona, and the largely Ndebele Zimbabwe African Peoples Union (ZAPU) succeeded in defeating the white settler regime, Rhodesia became Zimbabwe under Robert Mugabe, the leader of ZANU. Independence was followed by a period of repression by ZANU against the Ndebele followers of ZAPU in the middle of the 1980s, though peace was eventually restored. Mugabe continued to rule Zimbabwe with an increasingly heavy hand. Twenty years after independence, he again declared 'war' on the country's white farmers who still retained about half of the arable land. At the turn of the millennium, Zimbabwe was on the brink of economic and political collapse.

The defeat of the Smith's Rhodesian government was made possible by close cooperation between ZANU and African nationalist forces in Mozambique. These had begun by the late 1960s to wage a war of liberation, and offered safe havens for ZANU forces. Operating from Tanzania, FRELIMO had succeeded in gaining control of northern Mozambique in the early 1970s. In 1974, the Portuguese government fell, due in large measure to the cost of its colonial wars in Angola and Mozambique. Mozambique was granted independence in the following year but suffered years of destructive civil war from shortly after the Portuguese withdrawal. To the west, the increasing success of the South West Africa Peoples Organisation (SWAPO) against the forces of the South African Apartheid government had also helped ZANU to achieve its aims in Zimbabwe. South West Africa had been wrested from German control by South African forces acting under British authority during World War I. It was then ceded to South Africa as a Trust Territory under the terms of the Armistice, and in accordance with a mandate of the League of Nations. After a protracted liberation struggle, it became independent Namibia.

Despite its Portuguese colonial heritage, Mozambique is clearly part of the Southern African region culturally and historically. The Tsonga-speaking people of southern Mozambique straddle the border with South Africa. Mozambicans have long been intensively involved in the South African economy, first as labor migrants to white farms from as early as the eighteenth century, and later as laborers on South African mines from the late nineteenth century until the present. South African capital has flowed strongly into Mozambique after the end of its protracted civil war. The African National Congress came to power in South Africa in 1994 after the release in 1990 of Nelson Mandela from 24 years of imprisonment. With the end of colonialism, in all countries except South Africa, large numbers of white people withdrew, either moving back to European countries or into South Africa. Today, well over 95 percent of the population of all countries except South Africa is black African.

3. States and Politics

State institutions in southern Africa are all relatively weak. The colonial states of Mozambique and North and South Rhodesia were never very strong, focusing mainly on controlling black labor migrations and political movements, and on delivering security and services to white farmers, miners, and business interests. The activity of liberation forces also removed large areas from effective government control. Parts of these countries remained effectively beyond the reach of the state, neither paying much in taxes nor receiving many services from the central authorities. At the end of the twentieth century, all except Swaziland had democratic governments elected under universal franchise. The benefits of the postcolonial transition to democratic politics have been limited, however, by long liberation struggles in some cases (Zimbabwe, South Africa, Namibia), postcolonial civil war (Mozambique), rampant corruption, mismanagement (Zambia, Malawi), and, for most, increasing poverty and environmental damage. Throughout the last half of the twentieth century the political process was dominated by relatively few persons and parties.

All countries have been ruled for long periods by one party or by only one or two presidents. Free political expression has often been suppressed. In South Africa the National Party, dominated by white Afrikaners, ruled from 1949 until 1994, one of the longest periods of rule by a single party anywhere in the postwar world. The National Party implemented a set of laws and practices known as Apartheid ('separateness') that attempted to maintain exclusive political and economic power by whites. Racism was institutionalized, and black opposition parties were banned. Their members were either jailed or forced into exile. Only whites were allowed to vote. Within these limits, however, the political process was orderly, with presidents stepping down after their terms of office had expired. Following this trend, Nelson Mandela, the first president of a free South Africa and leader of the African National Congress, stepped down after serving one five-year term to make way for Thabo Mbeki, his elected successor.

The problem of how to integrate 'traditional' African politics of the rural areas with the 'modern,' urban-centered political systems deriving from Europe (either liberal-democratic or centralist-Marxist) has persisted to the present. While Botswana, Lesotho, South Africa, and Zimbabwe maintained 'advisory' constitutional roles for 'traditional authorities' (that is, predominantly rural tribal 'chiefs'), others such as Mozambique abolished chiefship altogether. Tribal groups such as the Himba in Namibia and the Lozi of Zambia maintain chiefs or kings, but without recognition by the state's constitution. With the exception of Swaziland, all possessed more or less liberal democratic parliamentary political systems by the end of the twentieth century. Swaziland alone possessed an absolute monarch based on precolonial African traditions of chiefship. Botswana institutionalized traditional rulers who have a role in government through the House of Traditional Authorities.

In white-ruled Zimbabwe and South Africa, chiefs were brought under government control and seen as part of the system of oppression. In South Africa, during Apartheid, the internal 'homeland' republics also utilized the institution of 'modernized' (salaried and provided with minor bureaucratic support) chiefships that were based on traditional institutions. Nevertheless, the beginning of the twenty-first century saw a renaissance of chiefship in South Africa and Mozambique, with remnants of precolonial traditional political systems often playing public roles in most other countries of the region.

4. Livelihood

In the later decades of the twentieth century, all countries in the region in one period or other suffered increasing isolation from world markets as the result of political sanctions (Rhodesia/Zimbabwe and South Africa), distance from markets (Lesotho, Malawi, Zambia), warfare (Mozambique), or declining market prices for their primary products (Zambia). Worldwide political and economic sanctions were applied first to Rhodesia (Zimbabwe) throughout the 1970s during the period of white settler rule, and then to South Africa during the final decades of the struggle against Apartheid.

A polarity between the 'modern' urban centers and the surrounding 'traditional' rural regions has characterized the broad politico-economic structure of the region since colonial times. A long-term, occasionally rapid, migration of rural agriculturalists to urban areas has, in the past, skewed gender balances as males moved to urban areas in search of wage labor, while women tended to stay in the villages. Damage to family structures, unemployment, urban crowding, severely limited police capacity, and poor urban planning have often led to high levels of social insecurity and criminal violence. Rural–urban migration had stabilized by the end of the millennium, but high levels of mobility between city and country remain, with kin groups attempting to maintain bases in both.

In Zambia, urbanization proceeded rapidly during the colonial period, with miners and work-seekers clustering in the 'Copper belt' in the north-central Zambia. With collapsing copper prices during the 1980s and mismanagement of mines, many people returned to rural areas. By the beginning of the twenty-first century, over 95 percent of Whites and Indians in South Africa were living in urban areas. Black South Africans are nearly equally divided between urban and rural. Rural areas are characterized by dispersed residence, subsistence agriculture and pastoralism, producing no significant economic surplus for trade.

Urban areas throughout the region are weakly centralized and sprawling. All large urban areas derive from colonial patterns of segregated 'suburbs' loosely grouped around multiple trading and industrial centers. Single houses (or complexes of houses, outbuildings, and shacks) surrounded by walled 'gardens' or yards comprise the formal urban settlement pattern, while large informal shack settlements surround them.

The colonial economy based on mining, large-scale farming, and labor migration contributed to disruption of family life and systems of social control, especially in rural areas of Zambia, Zimbabwe, and South Africa. While mining and agriculture now account for less than 10 percent of the GDP in South Africa, these two sectors dominate the economies of other southern African countries as they did in South Africa in the early twentieth century. Lesotho and Malawi, lacking natural resources except water, have been largely labor-exporting countries, relying on remittances sent home by miners who work mainly in South Africa, but also Zambia and Zimbabwe. The bulk of Botswana's GDP is derived from diamond mines, while Zambia relies on exports of copper, cobalt, and other minerals. Namibia is mostly arid, and depends on mineral exports, especially uranium, for foreign exchange earnings. All countries, however, participate in sending labor to the mines located throughout the region but dominated by South African mining houses.

At the beginning of the twenty-first century, economic and social prospects are dimmed by the threat of HIV/AIDS pandemic. Southern Africa emerged as the hardest hit region in the world. As a consequence, the rate of population growth is predicted to fall below replacement rates in the first decade of the twenty-first century.

5. *Language and Communication*

The process of urbanization has resulted in the emergence of distinctive urban cultures with small manufacturing, trade, mining, and transport industries initially attracting large numbers of formerly rural people to urban areas where different languages and cultures came into contact. They blended to form new urban-based languages and cultures. Examples include *Isicamtu*, 'talking' or *tsotsitaal*, 'gangster language,' in South Africa, based on Zulu grammar but blending English, Afrikaans, and Sotho or Twana words; or *Fanagalo*, meaning 'Do-like-this,' a pidgin based on Zulu used on the mines and farms throughout southern Africa. New forms of music, dance, and religion blending African traditions and Western forms also arose.

No indigenous African language has achieved status or acceptance as a single national language or *lingua franca* anywhere in southern Africa. South Africa has 11 official languages, of which those with the largest number of speakers are Zulu (8.7 m speakers), Xhosa (6.8 m), Afrikaans (6.2 m), Pedi ('Northern Sotho'; 3.8 m), English (3.5 m), Sotho (Southern Sotho; 2.7 m), and Tswana (2.8 m) in addition to the smaller linguistic communities of Tsonga ('Shangaan'), Venda, Ndebele, and Swati ('Swazi'). Other languages spoken include Gujerati, Urdu, Hindi, Tamil, and Bhojpuri from India. Portuguese, Greek, and German is spoken by smaller communities from Europe. Chinese is spoken by recent immigrants from China and Taiwan, although the small community of Chinese who arrived in the nineteenth century speak only English. French is spoken primarily by central African refugees and migrants in South Africa.

The linguistic diversity is even greater in Zambia where 41 languages are spoken, including Bemba, Tonga, Nyanja (Chewa), and Lozi. Thirty-three languages are spoken in Mozambique and 28 in Namibia. Although over 70 percent of the population of Botswana speak Tswana, there are many small, nearly-extinct languages spoken by Khoisan people, and by refugees from surrounding countries.

There is some mutual intelligibility among some of these languages. For instance, Sotho, Pedi and Tswana are mutually intelligible, as are Zulu, Xhosa and Swazi in South Africa, Botswana, Swaziland and Lesotho. Several languages have become *lingua francas* within limited areas of these countries—for instance Zulu in South Africa, or Bemba and Nyanja/Chewa in Zambia and Malawi. Almost all people speak several regional languages. As a consequence of diversity, with the exception of Mozambique, English has increasingly become the dominant public language of government and the media. It is also the language of trade and diplomacy. While Mozambicans use Portuguese as the language of government and education, they are increasingly turning to English, too, as they develop deeper political and economic links with other nations. Nevertheless, the large majority of all people in the region continue to rely on local Bantu languages in their homes.

Until 1994, South Africa treated both English and Afrikaans as 'official,' but Afrikaans now figures as only one of the 11 official languages. Afrikaans is indigenous to South Africa where it evolved from dialects of Dutch in the Dutch colonial settlements at the Cape of Good Hope in South Africa. Today it is spoken throughout South Africa and Namibia, with small pockets in other southern African countries. In South Africa, Afrikaans is spoken by a larger number of people than English as a home language.

Within the region, South African society is the most complex with respect to color, 'race,' religion, language and dialect, region, and class. Ethnic, political, and other kinds of identities—such as preference for certain sports, or foods—are based on these differences, and also help to create and sustain them. None of these primary indicators of social difference, however, tend to define a consistent or bounded category

of people. In other words, people of many different colors or race speak the same language or set of languages, while people of the same region or class vary widely in what religion they follow, or in their political allegiances. People sharing the same religion and language—which is often a combination that defines a fairly closely-knit ethnic or national community in Europe, or Asia—may span widely different colors, classes, and regions in southern Africa. For instance, while the majority of South Africa's 'Colored' people speak Afrikaans and belong, like most White Afrikaners, to some denomination of the Reformed Church, until recently they considered each other to be eternally divided by their 'race.'

Surprisingly, while virtually all schools and public places were rigidly divided by 'race' in the 1980s, by the end of the century, schools, shopping centers, and other public places had quietly integrated with no more than a handful of violent or other racial incidents. On the other hand, as the business center of Johannesburg gradually became 'black,' the entire economic geography of this city of six million people shifted, with largely white owned business moving north into newer, dispersed business centers within the last few years of the 1990s. Despite the previously enforced divisions and centuries of conflict across many different cultures and language groups, South Africa exhibits a degree of cultural, economic, and political unity that is perhaps unique in Africa. Virtually all of its people manage to communicate with one another in one or several common languages, shop in the same stores, watch the same television channels, furnish their houses in similar ways, and—ironically—share the belief that they are much more different from one another than they actually are in terms of many objective measures of cultural or social difference.

See also: Apartheid; Central Africa: Sociocultural Aspects; Colonialism, Anthropology of; Colonization and Colonialism, History of; East Africa: Sociocultural Aspects; Historiography and Historical Thought: Sub-Saharan Africa; Imperialism, History of; Sub-Saharan Africa, Archaeology of; West Africa: Sociocultural Aspects

Bibliography

Barnard A 1992 *Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan People*. Cambridge University Press, Cambridge, UK

Colson E, Gluckman M (eds.) 1951 *Seven Tribes of British Central Africa*. Manchester University Press, Manchester, UK

Freund B 1984 *The Making of Contemporary Africa: The Development of African Society Since 1800*. Indiana University Press, Bloomington, IN

Greenberg J 1966 *The Languages of Africa*. Indiana University Press, Bloomington, IN

Grimes B F (ed.) 1996 *Ethnologue*, 13th edn. Summer Institute of Linguistics [http://www.sil.org/ethnologue/]

Hall M 1990 *Farmers, Kings, and Traders: The People of Southern Africa, 200–1860*. The University of Chicago Press, Chicago

Hammond-Tooke W D 1974 *The Bantu-speaking Peoples of Southern Africa*. Routledge, London

Hammond-Tooke W D 1993 *The Roots of Black South Africa*. Jonathan Ball Publishers, Johannesburg

R. J. Thornton

Sovereignty: Political

Sovereignty is a set of principles or rules that define appropriate governance structures. These rules were first explicated with the greatest clarity by Bodin and Hobbes in the sixteenth and seventeenth centuries, and elaborated by later thinkers including Jeremy Bentham and John Austin. The defining principle of these initial efforts to clarify the meaning of sovereignty was that there could be one, and only one, source of the law, and that this source, the sovereign, was either in practice or in theory not subject to any higher authority. A corollary is that each sovereign state is independent, subject to no external authority.

In the contemporary world, the meaning of sovereignty has become less certain. Sovereignty has become associated with several different rules and practices, rather than with some single notion of supreme authority within a given territory. The realization of any pure version of sovereignty of the sort that Bodin, and especially Hobbes, embraced is precluded by three factors. First, the relationship between order and a single hierarchy of authority has been belied by actual practice, especially the experience of modern democratic states in which authority is fragmented and divided. Second, power asymmetries and the absence of any final authority at the international level has meant that, for some states, juridical autonomy has been an empty or half empty shell because their domestic political structures have been infiltrated or organized by other, more powerful, states. Third, the sovereign right of states to enter freely into any agreement means that political leaders may in some cases choose to compromise the domestic autonomy of their own polities; one element of sovereignty, the right to make treaties, confounds another, the independence of domestic political structures.

1. Origins of Sovereignty

The concept of sovereignty was associated initially with the domestic ordering of authority and control. At its core was the notion that there can be only one supreme authority. Even if divine authority or natural law existed (which even Hobbes admitted) in practice, only the sovereign could make laws. There could not

be any earthly authority above the sovereign, because then the sovereign would not be sovereign. There could also be no external authority, no international or transnational actor, that could negate the will of the sovereign. A political order based on sovereignty or the sovereign state can be contrasted with the medieval world, in which there were overlapping authority claims within the same territory, and the traditional Sinocentric world in which China alone was supreme; other political entities, such as Korea, were subordinate or tributary states yet not part of the imperial core.

In Europe, a political world in which states with defined territories and some degree of autonomy from external, especially transnational authority (most notably the Catholic Church), evolved over several hundred years as a result of material factors, including long-distance trade and technological changes in warfare that favored states over other political entities such as city leagues, the Holy Roman Empire, or the Papacy, and ideational ones, most notably the Protestant Reformation (Tilly 1990, Spruyt 1994, Skinner 1978). It was, however, the European religious wars which provided the motivation to develop a doctrine that could legitimate a political order populated with one, and only one, final source of authority within a given territory. The antagonisms bred by sectarian beliefs precipitated bloody civil conflicts in the major polities of Western Europe. Bodin, who was almost killed in religious riots in France in 1572, wrote his *Six Books of the Commonwealth* in response to the Huguenot rebellion and the Huguenot writings that justified rebellion against tyranny. For Bodin, the fundamental objective of any government must be to insure order, and not to protect liberty. Sovereignty, which for Bodin was intimately associated with the right to give laws and not just to make judgments, had to be lodged somewhere, and he argued that monarchy was preferable to either an aristocracy in which sovereignty would rest with a few, or a democracy in which it would rest with the many. Bodin, like Machiavelli, dwelt on the frailty of political order. He acknowledged that a sovereign who became a tyrant might justly be killed by a foreigner, but he rejected any rationale for domestic resistance (Skinner 1978, pp. 284–8, Hinsley 1986, p. 120).

Bodin did not, however, regard the sovereign as unconstrained. The sovereign was bound by the laws of God and nature, and by custom. Because natural law dictated that promises ought to be kept, the sovereign was obligated to honor treaties made with other sovereigns, and agreements made with his own subjects. Taxation required the consent of the governed, because arbitrary confiscation of property would violate the laws of nature. In France, Bodin averred that the king could not change the Salic law, a customary practice, which limited succession to males (Skinner 1978, pp. 273, 289, Bodin 1992, Bk. I, Chap. 8).

Hobbes, writing 80 years later, was more emphatic in his defense of the power of the sovereign. Hobbes rejected even the modest constraints that Bodin had imposed, arguing that it is the sovereign who interprets the laws of God and nature. Once individuals had entered into a social contract they could not or ought not challenge the sovereign, although Hobbes acknowledged that individuals had a right to run away in the face of violent death. The alternative to the Leviathan, the sovereign, was the state of nature in which life was 'nasty, brutish and short.' The line of argument focusing on sovereignty as an attribute of domestic political structures was developed further by legal positivists, such as Austin, who asserted that the law properly understood emanated from one single source of authority untrammeled by any constraints. Bentham, for instance, argued that natural law was a nonsensical concept.

2. *Sovereignty in Practice*

While in theory sovereignty has often been associated with the concept of a single lawgiver untrammeled by any higher source of authority, except perhaps God, in practice the actual exercise of authority has been much more complicated. The kind of state envisioned by Bodin, and especially Hobbes, never came into existence. In Britain itself, the strife of the seventeenth century led to a political system in which authority was divided between king and Parliament. In the United States, the Founding Fathers established a constitutional order with a division of power among the executive, legislative, and judicial branches, as well as between the individual states and the federal government. In the twentieth century, centralized hierarchical systems such as Nazi Germany and the Soviet Union provided neither prosperity nor domestic tranquillity for their citizens. The classic view which associated political order with a single hierarchy of authority is irrelevant in the contemporary world.

2.1 *Sovereignty in the International System*

The notion of sovereignty however, continues to be invoked at the international level, where it is defined by two principles. First, the right to domestic autonomy or independence: regardless of their internal structure of authority (unitary, federal, democratic, autocratic, etc.), sovereign states exclude both *de jure* and *de facto* external sources of authority. Second, the right to make treaties: sovereign states can enter freely into contracts with other entities. In practice, the first of these principles, autonomy, has been contested persistently, and the second, the right to enter freely into contracts, can result in agreements that contradict the first. Although a few states have approximated the sovereign ideal of actual, and not just formal juridical autonomy, the United States being the most important

example, there has never been an historical moment during which the world could be described as divided among such states, each enjoying exclusive authority within its own territory. Even if states have had a monopoly over the legitimate use of force within their own borders, the purposes toward which such force might be used frequently have been influenced or determined by external actors.

2.2 Dilemmas of Sovereignty

There are two factors that preclude the full realization of any idealized form of sovereignty. First, because there is no authority above states, powerful states can violate the autonomy of their weaker neighbors. Second, because sovereignty implies that states have the right to make any agreements that they please, there is nothing to prevent them from entering into treaties that might subject them to external authority structures.

The most obvious violations of the principle of autonomy involve military force. Major powers have not been reluctant to dictate the domestic political structures of their weaker counterparts, especially within their own spheres of influence. The United States has intervened frequently in the smaller states of the Caribbean and Central America because of concerns about security, financial stability, human rights, and illegal drugs. After World War II the Soviet Union imposed communist regimes on its satellite states in central Europe. In 1999, the member states of NATO conducted an extensive bombing campaign over Yugoslavia designed to coerce the Yugoslav government into changing its practices toward Kosovo Albanians while at the same time continuing to recognize Kosovo as a part of Yugoslavia.

External influence on the domestic political structures of states has not, however, been the result of direct military action alone. All of the major peace treaties, beginning with the Peace of Westphalia of 1648, have legitimated to some extent external legitimation of domestic authority structures, sometimes as a result of fully voluntary and mutually beneficial accords, but at others as an outcome of asymmetrical bargaining power. The Peace of Westphalia, often described as a critical moment in the development of the modern state system, did delegitimate the Catholic Church as a transnational source of authority and initiate a long period in European history in which the secular balance of power rather than some notion of a unified Christendom defined the way in which leaders conceived of and practiced international politics. But at the same time, the specific provisions of the Westphalian Peace established a new constitution for the Holy Roman Empire including religious toleration within Germany, a position supported by the victors in the Thirty Years' War, France and Sweden, and accepted only reluctantly by the Habsburg monarch, Ferdinand III. The Peace did not create a world of autonomous political entities; the political units within the Empire (principalities, electorates, free cities) were still embedded within the larger institutional arrangements of the Empire which were, as a result of the Westphalian Peace, now subject to international guidance. The Vienna settlement following the Napoleonic Wars guaranteed the rights of Catholics in The Netherlands. During the nineteenth century the European powers, fearing that ethnic conflict in the Balkans would destabilize Europe (which, indeed, happened in 1914) insisted on religious toleration as a condition for recognizing all of the successor states of the Ottoman Empire, beginning with Greece in 1832 and ending with Albania in 1913. Perhaps most clearly, the peace settlements following World War I included extensive provisions for the protection of minority rights. Poland, for instance, as a condition of recognition and admission to the League of Nations, agreed to bilingual education in some primary schools, and to refrain from holding elections on Saturday because such balloting would have violated the Jewish Sabbath. While a few states, such Hungary, which had a small domestic minority population but many co-ethnics living in other states, welcomed these minority rights arrangements, most of the 30 or more states that were affected regarded them as a necessary evil, accepted only to secure their international status.

The interwar efforts at international constraints on domestic practices were a dismal, albeit internationally legitimated, failure, and after World War II, human rather than minority rights became the focus of attention. The United Nations Charter, for instance, endorsed both human rights and the classic sovereignty principle of nonintervention. Disheartening developments associated with the disintegration of Yugoslavia in the 1990s, however, revived earlier concerns with minority rights. Recognition of the Yugoslavian successor states was conditioned on their acceptance of constitutional provisions guaranteeing minority rights, and in Bosnia the Council of Europe established a set of external authority structures including a Human Rights Commission, a majority of whose members were appointed by the Council.

Political leaders may also use their sovereign right to enter freely into treaties to compromise voluntarily the autonomy of their own polity. In the contemporary environment, the clearest example of the tension between the right to contract and domestic autonomy is the European Union. The European Commission, the European Court of Justice and the European Parliament are supranational authority structures that can, in some issue areas, make decisions that are binding on member states, even states that have not agreed to them explicitly. The rulings of the Court, for instance, have a direct effect in national judicial systems and supremacy over domestic law, doctrines that emerged through judicial practice rather than any explicit treaty. The Single European Act adopted in 1987, and the Maastricht Treaty ratified in 1993

provided for majority or supramajority, but not unanimous, approval for decisions in many issue areas. Mutual recognition, which obligates member states to recognize each others' national regulatory measures, creates extraterritorial authority because legislation in one state governs the activities of enterprises in others.

The European Union is a particularly dramatic example of the tension between sovereignty understood as the right to enter freely into contracts, and sovereignty understood as autonomy or independence from external authority structures, but it is not unique. The member states of International Monetary Fund have accepted conditionality arrangements covering not only specific policy targets such as the level of the public debt, but also domestic authority structures such as the establishment of an independent central bank or the creation of a corruption commission.

3. *The Significance of Sovereignty*

The tensions between different attributes of sovereignty, the coercive and legitimated efforts by external actors to influence authority structures within specific states, and the abandonment of the notion that sovereignty requires some one single source of domestic authority, does not mean that the concept is inconsequential. Sovereignty remains a widely understood script, comprehensible not only to political leaders but to wider populations as well. Sovereignty is a rallying cry for dispossessed populations even in situations where the best possible outcome would be a state whose authority structures would not be of its own making as is, for instance, the situation in East Timor. The salience of sovereignty may make alternative arrangements more difficult, especially in situations where compromises reached by political leaders could be undermined by counter elites deploying the language of sovereignty. Other forms of political order require either coercion, as in the former Yugoslavia, or explicit agreement, as has occurred in Europe. Innovative structures, such as the relationship that China established with Hong Kong in the late 1990s which included the participation of a foreign judge on the Court of Final Appeal and a separate Hong Kong passport, have been explained in terms of the modern language of sovereignty even when they are more reminiscent of different ordering principles, such as the traditional Sinocentric tributary state system.

Sovereignty has become a set of principles more associated with the nature of political entities in the international environment (states as opposed to, for instance, protectorates or empires) and their relationship with each other (nonintervention), than with the organization of domestic political authority, which was the primary concern of most important early writers on sovereignty. But the absence of any ultimate source of authority in the international system, the conflicting norms embraced by political leaders and their followers, and clashes over issues of both security and material welfare mean that while sovereignty remains a consequential idea influencing political choices, it will also be persistently contested.

See also: Authoritarianism; Autonomy, Philosophy of; Bentham, Jeremy (1748–1832); Contractarianism; Democracy; Governments; Hobbes, Thomas (1588–1679); International Law and Treaties; Law: Overview; Machiavelli, Niccolo (1469–1527); Nations and Nation-states in History; Natural Law; Power: Political

Bibliography

Austin J 1954 [1832] *The Province of Jurisprudence Determined and the Uses of the Study of Jurisprudence*. Noonday Press, New York

Bentham J 1970 *Of Laws in General*. Hart H L A (ed.) Athlone Press, London

Bodin J 1992 *On Sovereignty: Four Chapters from The Six Books of the Commonwealth*. [Ed. and trans. J H Franklin]. Cambridge University Press, Cambridge, UK

Bull H 1977 *The Anarchical Society*. Columbia University Press, New York

Deudney D H 1995 The Philadelphian system: sovereignty, arms control, and balance of power in the American state-union, circa 1787–1861. *International Organization* **49**(2): 191–228

Fowler M R, Bunck J M 1995 *Law, Power and the Sovereign State: The Evolution and Application of the Concept of Sovereignty*. Pennsylvania State University Press, University Park, PA

Hinsley F H 1986 *Sovereignty*, 2nd edn. Cambridge University Press, Cambridge, UK

Hobbes T 1996 *Leviathan*. [Ed. R Tuck]. Cambridge University Press, Cambridge, UK

Jackson R H 1990 *Quasi-States: Sovereignty, International Relations and the Third World*. Cambridge University Press, Cambridge, UK

Krasner S D 1999 *Sovereignty: Organized Hypocrisy*. Princeton University Press, Princeton, NJ

Oppenheim L 1992 *Oppenheim's International Law*, 9th edn. [Eds. R Jennings, A Watts]. Longman, Harlow, UK

Ruggie J G 1998 *Constructing the World Polity: Essays on International Institutionalization*. Routledge, London

Skinner Q 1978 *The Foundations of Modern Political Thought:* Vol. 2. *The Age of Reformation*. Cambridge University Press, Cambridge, UK

Spruyt H 1994 *The Sovereign State and Its Competitors*. Princeton University Press, Princeton, NJ

Strang D 1991 Anomaly and commonplace in European political expansion: Realist and institutional accounts. *International Organization* **45**: 143–62

Tilly C 1990 *Coercion, Capital, and European States, AD 990–1990*. Basil Blackwell, Cambridge, MA

Wendt A 1999 *Social Theory of International Politics*. Cambridge University Press, Cambridge, UK

Vattel Emer de 1852 *The Law of Nations; or, Principles of the Law of Nature, applied to the Conduct and Affairs of Nations and Sovereigns*. [New edn. (trans. J. Chitty)]. T. and J. W. Johnson, Law Booksellers, Philadelphia

S. D. Krasner

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Soviet Studies: Culture

Soviet studies, like its other regional studies counterparts, is primarily a European and North American academic response to post-World War II geopolitics. It overlaps with *Slavic Studies* but also includes a research on non-Slavic cultures living on the territory of the Soviet Union, and makes a reference to the history and culture of the Russian Empire before 1917. Soviet Studies grew rapidly during the Cold War, and were generously supported by the US government in its effort to 'understand the enemy.' It integrated a number of disciplines including history, economics, literature, linguistics, sociology, and ethnology, but gave primacy to political science. From the outset, such study took the supranational unity of the Soviet Union as the master frame of reference for the study of languages, cultures, and social institutions, however diverse.

The status of *culture* in Soviet studies is complex and changing. Interest in the political origins of the Soviet system underwrote a long-term emphasis within Sovietology (the Cold War title for Soviet Studies) on *political culture*. Forms of governance, the nature of domination and subordination, the mechanisms of legitimation of the regime, and the emergence of a self-selected ruling elite were among the principal concerns of this type of research. At the other end of the spectrum was research on *artistic culture*, which was treated as expressive of national culture in general, and its producers, Soviet intelligentsia.

The 'cultural turn' of the mid-1980s in Western social science stimulated new development in Soviet studies. Increasingly, studies of Soviet society directed attention to social discourses, practices, and everyday life. This anthropologically informed idea of *culture* as a realm of socially shared meanings and a system of representations now proliferates in this regional research.

This article traces the development of Soviet and post-Soviet studies with an emphasis on how culture was used as an analytical framework. It also considers the diverse lines of influence that run between Western (post) Soviet studies and internal studies of culture in the Soviet Union and Russia. Throughout, the term 'Soviet culture' refers to cultural developments of the socialist period from 1917 to 1991. 'Post-Soviet culture' defines the contemporary stage of cultural development in the territory of the former Soviet Union.

1. *Culture in Russian Society and Russian Social Sciences*

The concept of *culture* has a complex provenance in the Soviet and Russian contexts. One sense of the term refers to education and the level of self-possession of an individual—*kul'turnost'*. The second meaning symbolizes people's unity based on language, tradition, and heritage, and reflects the profound role of culture in nationalist discourse. Finally, culture in its expressive forms, and literature in particular, constitute a relatively independent domain of meaning.

The term 'culture' did not enter widespread use in Russian society until the 1880s. Before then, occasional use of the word referred either to 'literary culture' or to the broader attainments of education, enlightenment, and civilization. As the process of national self-definition gathered speed in the late nineteenth century, a series of cultural dualisms emerged. Perhaps most significant was the opposition between East and West, which became central to discourse on national identity. As the twentieth century began, cultural dualism of this kind informed increasingly powerful notions of the 'Russian way of life' and 'Russian culture.' At the same time, non-Russian ethnic groups were ascribed particular cultures. 'Culture' became, in this context, the container of difference, and as such underwrote a variety of inquiries into the material culture and 'vanishing traditions' of 'traditional peoples,' including folklore studies. This interest became more pronounced in the Soviet era, reflecting a perceived need to study the effects of socialist modernization on the 'hinterlands' (Dunn and Dunn 1974).

Another formative distinction within Russian/Soviet culture is that between 'high' and 'low.' Originally the subject of nineteenth century debates between Slavophiles and Westernites, high–low distinctions became a major concern of communist rule. Communist party ideologues attempted to do away with the notion of a distinction between elitist and popular cultures by forging a 'culture for the masses.' New cultural initiatives were meant to speak to the peasants and the workers. Beginning in the 1920s, *kul'tura* was carefully planned and designed to eliminate cultural divides. In Stalin's time the word 'culture' acquired a distinctive suffix—*kul'turnost'* ('cultured-ness'). The suffix signified the inclusiveness of the new cultural order—embracing the new artistic cannon (Socialist realism) manners, and styles of life (Boym 1994). High art, for its part, would be democratized through museums, education, and cultural associations. Although greatly influenced by political constraints and censorship, this set of cultural policies produced a number of progressive trends in art and literary criticism that profoundly influenced Western cultural and linguistic theory.

The school of linguistics known as *Russian formalism* was one of these. It developed in the mid-1910s around Viktor Shklovsky and Yuryi Tynyanov, and in the 1920s merged with the Moscow Linguistic Circle, led by Roman Jackobson. Russian formalist theory tried to overcome the dualism of form and content by emphasizing the study of poetic language as such: the interplay of verbal forms, tropes, sounds, syntactic

constructions, etc. Their method of 'functional poetics' sought the underlining structure of works through the relationships between their internal elements (Erlich 1965).

Mikhail Bakhtin and Vladimir Voloshinov formed another group situated between the formalists and the official Marxist line on culture. Their work became an extensive study of what Bakhtin called the *dialogic* dimension of language—the construction of meaning in a linguistic field shaped by competing discourses such as official and popular cultural forms. Prohibited in the Soviet Union, these theorists found an audience among Western literary and cultural analysts, particularly in the areas of structuralism and semiotics (Holquist 1990).

Other movements were important within Soviet scholarship. During the 1960s, 1970s, and 1980s, the *Tartu School* associated with Yurii Lotman became prominent. These scholars developed a sophisticated approach to cultural history that emphasized cultural wholeness and totality in its search for the key to Russia's duality. They accorded culture a near metaphysical status—culture as an abstract force that reasserted itself onto national patterns of life (Lotman 1984).

As a result of political constraints and censorship, more analytical attention was paid to literary and art criticism than to other forms of cultural analysis. In response, literary and art criticism became a common means of articulating broader critiques of the Soviet system. This was the case with Boris Groys' *The Total Art of Stalinism* and Vladimir Papernyi's *Kul'tura 'Dva.'* Indigenous Soviet scholarship produced an extensive body of knowledge of Russian culture—though one that remained independent of Western theoretical influences until the 1990s.

2. *The Development of Soviet Studies in North America*

Although culture held a marginal position in Western Soviet studies until quite recently, the immediate postwar period between 1946 and the early 1950s provides a brief exception. For a time, linguists, sociologists, anthropologists, and even psychologists became interested in the study of Soviet society. They were encouraged in their work by the wave of war refugees who could provide first-hand knowledge about the life of the region. This effort collapsed shortly after it began. The comprehensive Soviet control of information thwarted Western inquiries until *Perestroika* in 1985. The inaccessibility of Soviet society to independent observers put severe methodological constraints on most forms of social science inquiry. Moreover, political considerations from the 1950s to mid-1980s played a major role in prioritizing certain kinds of research. High politics and economics were deemed crucially important and the Soviet political system became the focus of most attention. The 'ordinary man' nearly disappeared as a subject of inquiry. Political scientists and in some cases historians came to dominate this research agenda. Sociologists and anthropologists were more or less removed from the field.

One result of this reconfiguration was the so-called 'totalitarian model,' best represented by the line of political theory running from Hannah Arendt's *The Origins of Totalitarianism* (1953) to Martin Malia's *The Soviet Tragedy* (1994). The totalitarian model focused on the one-party political system and its methods of control: primarily terror and the destruction of meaningful social bonds outside the Party. This scholarship tended to suffer from several related problems. (a) It tended to take state ideologies at face value, reproducing Cold War prejudices about the diametrical opposition between capitalist and socialist systems, at the political as well as the economic level. (b) It took the indoctrination of the Soviet 'masses'—another self-proclaimed success of the Communist Party—as empirically proven.

Change came in the late 1970s and early 1980s, as debates arose between advocates of the totalitarian model and those who challenged totalitarian thinkers for overemphasizing the power of the Soviet political system. 'Revisionists' raised political questions about the nature of popular support and argued for evidence of more complex negotiation between the regime and the society. This is the moment in which 'culture' came to wider attention. Class identity, gender issues, social movements, and nationalism were introduced as necessary complements, and often as alternatives, to the political models. On their side, revisionists exposed themselves to the criticism of 'whitewashing' communist repression and purges.

The fall of the Soviet regime in 1991 transformed studies of culture on both sides. On the one hand, it provided access to Russia for Western scholars. Archives were opened and sociological studies of the general public became possible. On the other hand, it put Russian scholars, with their knowledge of the language and culture, in contact with the West. This new wealth of information and intellectual exchange was accompanied by subtler shifts in approach. In reaction to the revisionist effort to 'de-ideologize' the Soviet regime, a new generation of scholars returned to the question of how ideology created new and durable social structures. They did so, however, while retaining the focus on culture and organization of everyday life. The following section considers the major directions in research on culture in Soviet and post-Soviet studies that emerged around questions of social identities, everyday life, consumer, and popular cultures. It pays less attention to political science that continues to view culture as limited to 'political culture'—the expressions of economic and political processes.

3. Areas of and Research in Soviet and Post-Soviet Studies

Questions of 'modernity' and 'modernization' are common denominators for much of the Soviet and post-Soviet studies, from the official communist ideology of socioeconomic development to a broader trajectory of rationalization implicit in notions of 'modernity.' Against this backdrop, a number of cultural phenomena have attracted analytical attention.

3.1 The 'Soviet Individual'

The relationship between the regime and the society and the way it shapes individual identity is among the 'oldest' concerns in Soviet studies. From the 'totalitarian model' to the contemporary challenges of creating a civil society, many scholars have been preoccupied with what it meant to be a 'Soviet Individual.' Soviet modernization specified the need for an 'all-round personality' appropriate to a 'non-exploitative society,' and the state and the Party took responsibility for creating it. Facilitated by rapid urbanization, this 'socialization' proceeded through the imposition of cultural regulations and norms, altering the existing structures of identity (Volkov 2000). The state initiatives to reform individuals were not without limitations. Recent studies of socialist individuation show that a difference between public adherence and private life allowed for complex strategies of avoidance of Party control. Thus emerged the 'dissimulating subject' (Kharkhordin 1999). As the Soviet system calcified and as socialist prosperity became an increasingly remote fiction, endorsement of the official ideology became a means of securing a private existence out of reach of the intrusive power of the state (Yurchak 1997). Alexei Yurchak suggests that although the 'objective conditions of socialism' were seen as false, Soviet individuals continued to act as if they were true believers. This collective 'pretense of misrecognition' strengthened social solidarity and allowed for meaningful life in other areas.

3.2 Class and Stratification

Through the creation of the Soviet individual, the Soviet state aimed to remove distinctions among social groups. 'Class' became a focus of social policies and, as a result, of political studies. In Soviet studies, the Soviet Union provided a contrasting case for Western economic upward mobility. While Soviet officials insisted on 'giving power to a class,' Western Sovietologists viewed mobility primarily in individual terms, as resulting from individual opportunities to improve social standing. Other critiques treated the Soviet claim of the disappearance of a class society as a necessary fiction of the regime. Discussions of mobility and stratification were enriched in the 1980s by culturalist accounts of class consciousness as a formed and learned identity, drawing especially on E. P. Thompson's work on nineteenth century Britain. These directly challenged the dominant Soviet conception of class as an ascribed characteristic that had little connection to socioeconomic position (Fitzpatrick 1993). They were central to emerging work on a range of largely unexplored types of social behavior—especially the patterns of dissimulation, unmasking, and stigmatization that structured social identity and the relationship between individuals and the regime. Studies of class have been far less frequent in post-1991 scholarship, however, reflecting the rise of interest in other registers of identity.

3.3 National Identity

Socialist modernization also made claims to cultural progress in ethnic relations while preserving the national distinctiveness of Soviet society. The fate of non-Russian people under Soviet rule was the subject of numerous studies in the Soviet era, including historical studies of Central Asia (Eickelman 1993) and Siberia (Balzer 1997). Western Sovietologists also produced studies of non-Russian people—although these were, with rare exceptions, heavily influenced by cultural evolutionary paradigms. Modern anthropological approaches emerged only later (e.g., Caroline Humphrey's ethnography of collective farming in Buryatia (1983)).

Because nationalism was officially represented as an obstacle to communist internationalism and ostensibly suppressed, studies of post-Soviet ethnic relations reproduce a 'repressive hypothesis.' When, the argument goes, 'political opportunities' arose in the late 1980s, ethnic conflicts resulted from the discontent of the Russian 'core' and the intensifying local nationalisms of the non-Russian periphery (Kaiser 1994).

Yet, as Yuri Slezkin has argued, it may be more instructive to consider how socialism cleared the way for contemporary nationalism (Slezkin 1994). Despite opposition between socialism and nationalism on many levels, the principle of national difference was officially sustained in the federated states. The individual 'nation' was the only permitted form of social association other than the Communist party itself. This provided an opportunity for group mobilization along nationalist lines when the opportunity arose.

How this opportunity was translated into processes of cultural formation after the 1991 break-up is a complex story that is only beginning to be told: identity formation in the Ukraine (Wanner 1998) draws on different resources than that of 'traditional' peoples in Siberia (Grant and English 1999). The prominence of new concepts such as the *Russian diaspora* and the *near abroad*—designating Russians living in the republics—also reflects the new, post-Soviet realities of

national co-habitation. Research on population diasporas on the territory of the former Soviet Union offers new conceptual questions: what does it mean to become an ethnic and linguistic minority without having emigrated in any conventional sense (Shlapentokh et al. 1994)? David Laitin (1998) discusses new national identities in the post-Soviet space as outcomes of a 'tipping game'—a rational choice to learn new languages and cultural norms if one wishes to enjoy the benefits of citizenship. Laitin argues that under post-Soviet conditions, cultural assimilation depends neither on cultural nor linguistic similarity, nor on ethnic closeness, but rather on the degree to which the structural conditions of economic, political, and social life stimulate a desire to chose some and not other forms of identification (Laitin 1998).

3.4 History and Collective Memory

The interpretation of national history was also important for Soviet modernization. Soviet culture, although future-oriented, was deeply rooted in a particular historical vision. Peter the Great stood as a symbol of modernization, linked first to Bolsheviks and then to Mikhail Gorbachev; the French Revolution was a predecessor of the Russian Revolution, and the victory over Germany in World War II completed the genealogy of Russian military glory (Clark 1994). Stories of redemptive suffering in World War II, especially, sponsored belief in the unique position of Russia in world history. They provided a narrative of 'all-national struggle' and a precedent for the patriotic mobilization of a 'cynical and apathetic populace' (Tumarkin 1994, p. 277). Especially under Brezhnev, these narratives counterbalanced the numerous tensions that strained national unity while giving priority of place to ethnic Russians as the leaders of the Soviet people. After 1991, socialism itself became a form of narrative and came under scrutiny as a subject of historical interpretation.

3.5 Gender and Sexuality

One of the goals of socialist modernization was to solve the 'women question.' Western Sovietology tended either to accept communist claims that gender inequality had no place in the new society—Ella Winter's (1933) *Red Virtue* and William Mandel's (1975) *Soviet Women* took this line—or to severely criticize communists for indulging in mostly empty rhetoric on these issues, as Manya Gordon did in *Workers Before and After Lenin* (1941). The latter perspective reflected the tension between the official policy of women's emancipation and gender equality, and *de facto* sexism and inequality in everyday life (Lapidus 1978). The post-Soviet period has, by most accounts, seen further deterioration of this situation, as measured by discrepancies in male and female employment, sexual harassment in the workplace, and sexism in the media. Though some have argued that these phenomena are products of the rapid restructuring of society, most scholars of gender in Russia see connections between the formal promotion of women rights under socialism and the current regression. Many have argued that the Soviet constitutional definition of women as workers *and* mothers is partly responsible for discrimination directed against women in practice. In addition, the imposition of feminist ideology 'from above' did not reflect popular support 'from below' (Muller and Funk 1993). As the post-Soviet state fell into crisis, it has been unable to defend the full range of formal rights for women. One of the consequences, Mary Buckley (1992) notes, is that women have been increasingly driven away from participation in the public sphere.

Sexuality, as distinct from gender, was rarely examined in Soviet studies. Because studies of sexuality are methodologically demanding and complicated, they were severely restrained by Soviet censorship and rigid cultural norms. Peter Stanford's 1967 book, *Sexual Behavior in the Communist World*, presented an interesting but fragmented view of sexual practices under socialism. Most work on sexuality in Western Soviet studies was based on the analysis of available legal or literary texts and other cultural artifacts (Dunham 1961, Levitt and Toporkov 1999). Only in the 1990s has more theoretical, fieldwork-based research developed—partly in response to the dramatic reorganization of sexual norms in Russia. Laurie Essig's research on homosexuality in Russia is exemplary in this regard, offering not only rich empirical material but a critical comparative analysis of the distinctive formation of sexual subjectivity in Russia (Essig 1999).

3.6 Popular and Consumer Culture

During the Soviet years, the 'shortage economy' placed a premium on the hoarding, rather than consumption, of goods by individuals. Yet consumer goods were vitally important to socialist ideology. Images of communist abundance were crucial in securing the regime's legitimacy—especially in the context of the current sacrifices. In reality, the shortage of goods resulted in their informal distribution through a 'second economy.' Access to goods was a coveted and largely political privilege: those who possessed it were in some respects analogous to the Western moneyed and propertied class (Osokina 1998).

As new goods flow into the post-Soviet space and as the new consumer culture develops, new questions arise. Do the new relations of capitalism dissolve the fabric of pre-existing, socialist-era social relations? Does the international flow of goods homogenize cultures? How are we to understand the capitalist

'break': is there a link between the Soviet and post-Soviet consumer cultures (Humphrey 1995)?

Russian scholars and Western scholars of Russia alike have shown tremendous interest in the recent flowering of popular culture—a new feature of post-communism. Youth culture, night life, and new media are just a few areas of recent scholarly interest. As more cultural artifacts are imported to Russia from the West, the post-Soviet cultural landscape changes further. The study of popular culture's contemporary forms, however, requires careful consideration of the peculiar trajectory of the concept in the Soviet and Russian contexts (Barker 1999).

4. *New Social Realities And The Future of Post-Soviet Studies*

The demise of the Soviet Union in 1991 opened a new chapter in the studies of the region. It also brought considerable confusion to the field. The break up of the old territorial unit challenged the geographical coherence of the former Soviet Studies. It yielded new geographical groupings. Central Asia, for instance, could be plausibly integrated into Middle Eastern studies, while the Baltic republics, with their effort to join the European Union, could be approached in the context of Western Europe. Yet the shared socialist past underlies considerable similarity in the social, economic, and political development of these countries and precludes easy redefinition of the region.

The fall of the Soviet regime made it clear that certain perceptions of the Soviet Union were an integral part of the national fabric in Europe and the United States. Political scientists bore the most criticism for letting these perceptions cloud their understanding of fundamental social developments in the Soviet Union, leading to their collective failure to predict the end of communism. Despite this fact, revisions of totalitarian and modernization theories continue to be produced. New Russia centered accounts of capitalism's inevitability—part of a body of 'transition theory' modeled on European and North American capitalism—have come to the fore (Burawoy and Verdery 1999). Opposed to such visions of predetermined change are diverse efforts on the part of younger scholars to understand how family, work, and community refashion themselves in creative and resistive processes of change.

In spite of the accumulation of empirical research, few Cold War-era studies of Soviet culture were theoretically innovative. The opening of post-Soviet society and its ongoing transformation pose new challenges for Western cultural theory and studies of culture, from theories of capitalist development to nationalism and models of historical rupture. If taken critically, this challenge presents remarkable opportunities, both for the new Russian and regional studies, and for the study culture in general.

See also: National Traditions in the Social Sciences; Soviet Studies: Gender; Soviet Studies: Politics; Soviet Studies: Society; Totalitarianism

Bibliography

Balzer M M (ed.) 1997 *Shamanic Worlds: Rituals and Lore of Siberia and Central Asia*. North Castle Books, Armonk, NY

Barker A (ed.) 1999 *Consuming Russia*. Duke University Press, Durham, NC

Boym S 1994 *Common Places*. Harvard University Press, Cambridge, MA; London

Buckley M (ed.) 1992 *Perestroika and Soviet Women*. Cambridge University Press, Cambridge, UK; New York

Burawoy M, Verdery K (eds.) 1999 *Uncertain Transition: Ethnographies of Change in the Post-socialist World*. Rowman & Littlefield, Langam, MA

Clark K 1994 Changing historical paradigm in Soviet culture. In: Lahusen T (ed.) *Late Soviet Culture*. Duke University Press, Durham, NC

Dunham V 1961 Sex: from free love to puritanism. In: Inkeles A, Kent G (eds.) *Soviet Society: A Book of Readings*. Houghton Mifflin, Boston, MA

Dunn S, Dunn E (eds.) 1974 *Introduction to Soviet Ethnography*. Highgate Road Social Science Research Station, Berkeley, CA

Eickelman D (ed.) 1993 *Russia's Muslim Frontiers: New Directions in Cross-cultural Analysis*. Indiana University Press, Bloomington, IN

Erlich V 1965/1981 *Russian Formalism: History, Doctrine*, 3rd edn. Yale University Press, New Haven, CT

Essig L 1999 *Queer in Russia*. Duke University Press, Durham, NC

Fitzpatrick S 1993 Ascribing class: The construction of social identity. *Journal of Modern History* **75**(4): 745–70

Fitzpatrick S (ed.) 2000 *Stalinism: New Directions*. Routledge, New York

Gal S, Kligman G 2000 *The Politics of Gender After Socialism: A Comparative Historical Essay*. Princeton University Press, Princeton, NJ

Grant B, English (ed.) 1999 Neotraditionalism. In: Pika A (ed.) *The Russian North: Indigenous Peoples and the Legacy of Perestroika*. Canadian Circumpolar Institute, Edmonton and University of Washington Press, Seattle, WA

Holquist M 1990 *Dialogism: Bakhtin and his World*. Routledge, London; New York

Humphrey C 1983 *Karl Marx Collective: Economy, Society, and Religion in a Siberian Collective Farm*. Cambridge University Press, Cambridge, UK

Humphrey C 1995 Creating a culture of disillusionment. In: Miller D (ed.) *Worlds Apart*. Routledge, London and New York, pp. 43–68

Kaiser R 1994 *The Geography of Nationalism in Russia and the USSR*. Princeton University Press, Princeton, NJ

Kharkhordin O 1999 *The Collective and the Individual in Russia*. University of California Press, Berkeley, CA

Laitin D 1998 *Identity in Formation*. Cornell University Press, Ithaca, NY

Lapidus G 1978 *Women in Soviet Society*. University of California Press, Berkeley, CA

Levitt M, Toporkov A (eds.) 1999 *Eros and Pornography in Russian Culture*. Ladomir, Moscow

Lotman Iu U 1984 In: Shukman A (ed.) *The Semiotics of Russian Culture*. Dept. of Slavic Languages and Literature, University of Michigan, Ann Arbor, MI

Muller M, Funk N (eds.) 1993 *Gender Politics and Post-communism; Reflections from Eastern Europe and the Former Soviet Union*
Naiman E 1997 *Sex in Public*. Princeton University Press, Princeton, NJ
Osokina E 1998 *Za Fasadom 'Stalinskogo Izobiliia': raspredelenie i rynok v snabzhenii naseleniia v gody industrializatsii*. ROSSPEN, Moscow [Forthcoming translation and edition: *Our Daily Bread*. Edited and translated by Kate and Greta Bucher, trans. Sharpe ME]. Armonk, New York
Shlapentokh V, Sendich M, Payin E (eds.) 1994 *The New Russian Diaspora. Russian Minorities in the Former Soviet Republics*. M. E. Sharpe, Armonic, NY
Slezkin Y 1994 The Soviet Union as a communal apartment. *Slavic Review* **53**(2): 415–52
Tumarkin N 1994 *The Living & The Dead*. Basic Books, New York
Volkov V 2000 The concept of Cul'turnost'. In: Fitzpatrick S (ed.) *Stalininsm: New Directions*. Routledge, New York, pp. 210–30
Wanner C 1998 *Burden of Dreams: History and Identity in Post-soviet Ukraine*. Pennsylvania State University Press, University Park, PA
Yurchak A 1997 The cynical reason of the late socialism. *Public Culture* **9**(2): 161–88

O. Sezneva

Soviet Studies: Economics

With growing intensity during the years following World War II, Western economists studied the economy of the Soviet Union and similar East European variants bound loosely in the West by the paradigm of Soviet studies. Systematic analysis of the Soviet economy after World War II was motivated by a variety of factors. The dominant pragmatic motivation for studying the Soviet economy was the emergence of the Soviet Union, functioning in a Marxist–Leninist framework, as a major political and potential economic power, and the related Cold War typology of East–West relations as the dominant geopolitical framework in which the competing economic and social systems of East and West would interact. Beyond the obvious interests of national security in the West, analysis of the Soviet economy and similar systems of Eastern Europe was driven by a variety of other forces.

Economists understood that the economic system and related economic policies of the Soviet Union were fundamentally different from those found in Western market capitalist economies. These differences were arguably very important in determining how resources were allocated in these different economic systems, the core of the field of comparative economic systems. The Soviet Union became the dominant example of what would later be termed the administrative command economy. Moreover, economists realized that it would be a challenge to apply the research methods and traditional tools of neoclassical economic analysis to the peculiar organizational arrangements of the Soviet economic system, especially in a setting of Soviet secrecy.

In addition to issues of national interest and the inherent scholarly interest in the command economy, study of the Soviet Union was driven by pragmatic factors. The Soviet Union had, by most indicators, grown rapidly in the 1930s, a decade critical to understanding the subsequent Soviet economic experience. Moreover, the Soviet Union emerged from World War II to experience rapid economic growth in the 1950s, a decade when the subsequent pervasive growth slowdown was not evident. In an era when Soviet economic prospects looked bright, the issue of economic performance, and especially economic growth, was viewed as important. In addition, there was growing interest in pragmatic issues such as major sectors and regions, defense spending, foreign trade, and technology transfer for a largely closed economy functioning in an increasingly global economic environment. How would Western market-based firms interact with their Soviet counterparts, the Soviet firms operating in an isolated, secretive, and fundamentally different economic environment? Ironically, similar pragmatic interest would motivate later Western interest in other sectors of the Soviet economy, for example Soviet agriculture, the poor performance of which led to growing Soviet activity in foreign grain markets during the 1960s and thereafter. The growing world demand for energy has sustained interest in energy rich countries, the Soviet Union being a major example.

These factors led to a growth of interest in the Soviet economy and recognition of the need for a cadre of scholars whose research interests would focus on understanding resource allocation in the very different Soviet setting. This need to know led to the creation of area studies programs, most of which combined the usual postgraduate training in the discipline (economics) with associated area training in disciplines such as geography, political science, languages, and the like. Courses devoted to the Soviet economy and the more general issues of comparative economic systems became common in most colleges and universities. While economists understood the need to apply the tools of economic analysis to the Soviet case, they also understood the importance of a broader perspective. For such a broad scholarly objective, group identity in area studies programs and research centers was attractive, and yet there were costs associated with the pursuit of these objectives.

Study of the Soviet Union was a difficult task. Scholars were required to learn the Russian and possibly other languages. Moreover, access to all source materials was difficult, and for economists, much effort was required to assemble and to utilize

even minimal amounts of data, so critical for even basic statistical analysis. Soviet isolation from world agencies meant that data sources standard for any Western researcher were unavailable. Survey research was not generally possible, and Soviet official statistical publications (*The National Economy of the USSR*) contained limited information often of unknown origin and definition. Statistical and other information on the Soviet economy was limited, uneven, and often of questionable reliability. Archival research presented a major challenge notwithstanding the later availability of excellent guides to the information potentially available (Grimstead 1988). Field experience in the Soviet Union was essential, though the scholarly benefits of travel and work in the Soviet Union were difficult to realize in a setting of pervasive secrecy and limited access to information.

Finally, in the Western academic setting, it was not always easy to combine the professional demands of the discipline (economics) with the important but peripheral pursuits of area studies. Put differently, the postwar era witnessed major strides in the development of neoclassical economic analysis. In addition to debate over the appropriate paradigm within which to analyze the very different Soviet experience, major areas of development in neoclassical economics, for example, econometric theory and practice, and the development of computer technology for data analysis, were of only limited applicability in our study of the Soviet economy, given the limited availability of meaningful disaggregated data on the Soviet economy. In this context, there was often a pervasive gap between analysis of issues in Western economies and the analysis of comparable issues in Soviet-type economies. For example, as the availability of data and expanded computer access led to important new developments in the analysis of labor allocation in the West, analysis of labor allocation in the Soviet context remained difficult.

1. Soviet Studies: The Framework in Economics

The content of Soviet studies (economics) is very broad, probably best understood from a chronological perspective. Although it is arbitrary to identify specific time periods, the early postwar years were dominated by a major effort to assemble meaningful data series, and to begin an analysis of how the Soviet system actually worked (plan formulation and plan execution), an understanding essential to any meaningful assessment of performance. Thereafter, there was growing interest in sectoral analysis of the Soviet economy and issues of economic reform, especially following the reform attempts of the late 1950s and 1960s. Through the 1970s, economic analysis of the Soviet economy improved, though unevenly. For some sectors such as foreign trade, data availability facilitated more sophisticated analysis, while for other sectors such as labor, data limitations constrained the level of analysis that could be done. In general, data on money and price issues was almost nonexistent. However, through the 1980s, our analysis of the Soviet economy grew in both depth and breadth, with significant attention paid to important subjects such as consumption, investment, income distribution, and technological change. More information became available, new topics were investigated, and new methodological approaches applied, expanded by the new era of *perestroika*.

There is, however, an important sense in which a chronological approach is imperfect. During the transition era of the 1990s, the importance of the Soviet past for understanding the transition continued to be recognized. Likewise, throughout the postwar era, analysts of the Soviet economy recognized the fact that much of the Soviet system as we knew it emerged during the 1930s. Thus research interest in the early years of the Soviet economic experience sustains, and much research on these early years has helped analysts to understand later contemporary events (Davies 1980, Davies 1990, Nove 1982, Zaleski 1980).

In yet another dimension, the breadth of Soviet Studies is evident. While a chronological survey of the postwar era correctly identifies performance as a broad but key focus of research, scholars also developed and sustained an interest in issues well beyond performance narrowly defined. Thus in a country so large, naturally wealthy and diverse, a substantial literature emerged on a broad range of issues such as regional economic development, population, urbanization, and the like. As economists moved beyond the immediate frontiers of the discipline of economics, the richness of Soviet studies became evident. Moreover, there was often a close relationship among disciplines; geographers who studied regional issues often did so with methods familiar to economists; the work of political scientists was important for our understanding of bureaucracy.

2. The Content of Soviet Studies: Economics

Although considerable work on the Soviet economy was done prior to World War II, the dominant era of Soviet studies as an identifiable paradigm was the period 1950 through 1991. While intelligence analysis of the Soviet economy was done during World War II, the decade following the end of the war witnessed the growth of interest in the Soviet economy in the United States, Great Britain, France, and Germany in several major directions.

First, the Soviet economic system was dominated by state ownership with resource allocation conducted through a national economic plan, a document formulated by central planning authorities and implemented through ministries by industrial enterprises, agricultural enterprises, and other organizations. It was

evident from analysis of the early years of the command economy, that if this system was to be understood and its progress assessed as an alternative to market arrangements, it would be necessary to understand how it worked and how well it worked. To achieve these tasks, it would be necessary to assemble meaningful data on national income and its components. The focus on data collection and analysis dominated analysis of the Soviet economy and was complicated by Soviet methods of data collection and dissemination, official Soviet secrecy, and the use of a Marxian (conceptual) framework by the Soviet Union. The initial work on national income (Bergson 1961) became the basis for early assessment of Soviet economic growth and subsequently Soviet productivity, key elements for understanding the broader issues of Soviet economic performance and economic growth.

In addition to the development of meaningful data for analysis, it was necessary to understand both plan formulation and plan execution. Early efforts to understand plan formulation were cast primarily within the framework of input-output analysis, though subsequent work would also use mathematical modeling to understand the planning process. Plan implementation was the focus of early work (Granick 1954, Berliner 1957) combined with extensions into important sectors of the Soviet economy such as monetary and fiscal issues, capital accumulation, labor allocation, consumption, investment, and thereafter foreign trade.

From the 1960s through the 1970s, there was increasing attention paid to the Soviet economy, especially a growing number of in-depth sectoral studies, for example, in agriculture, and the growing use of econometric methods for statistical analysis and modeling (Green and Higgins 1977). Attention to the gathering of data took a major step forward in the late 1970s when the Soviet Interview Project (Millar 1987) resulted in a major database on the Soviet household. At the same time, interest in Soviet economic reform continued, albeit with a recognition that reform in the Soviet Union was having very limited impact on performance (Hewett 1988) even during *perestroika*, the last major reform of the Soviet era.

Western analysis of the Soviet economy continued to evolve through the 1980s with continuing emphasis on economic growth and productivity (Ofer 1986). In an era of declining performance, new themes emerged, while the analysis of basic issues sustained with important extensions. For example, while the theory of planning was in a sense the macroeconomics of the Soviet economy, the pervasive shortages and growing purchasing power of citizens led to interest in issues such as monetary overhang, arguably the beginnings of new thinking about the Soviet macro-economy.

Interest in sectoral allocation continued with growing interest in capital allocation and especially the issue of technological change. Moreover, analysis of the household permitted greater depth in the analysis of labor allocation. This depth was important, because it brought Western analysis of the Soviet economic system much closer to levels of sophistication present in the West and facilitated better understanding of the issues underlying Soviet economic growth and especially the continuing growth slowdown.

The latter years of the Soviet era witnessed the development and application of new theoretical and statistical methods. For example, the theory of the Soviet firm, so critical to understanding Soviet plan implementation, had been conducted by Western economists virtually in isolation from Western analysis of the firm. While variations of profit maximization dominated Western analysis of the firm in market settings, the analysis of the Soviet firm was thought to be more readily understood within the context of organization theory. This distinction was eroded as neoclassical theory placed new emphasis on property rights, agency theory, and information economics. Western analysis of the Soviet micro-economy assumed new dimensions, important for understanding the pervasive problems of the Soviet economy such as incentives, information problems, shortages, poor plan fulfillment, lack of technological change, and the like. During the 1980s, disequilibrium economics was applied with increasing frequency to analyze the peculiar allocation arrangements of the Soviet economy, and especially to understand the emergence of the second or informal economy. The development of the shortage economy (Kornai 1992) provided a broad theoretical underpinning to facilitate statistical investigation of the socialist economic system.

Finally, throughout the 1970s and 1980s, interest in the Soviet economy and similar examples elsewhere grew. Emphasis continued to be placed upon developing improved databases. While the Central Intelligence Agency continued to develop and disseminate data, other agencies provided important sectoral information. For example, the US Department of Agriculture was a major source of data and analysis on agriculture and the Soviet rural economy, while the Foreign Demographic Analysis Division of the Department of Commerce provided and analyzed population and labor force data. During these years, the publications of the Joint Economic Committee on China, Eastern Europe, and the Soviet Union became a basic resource for analysis of these economies. In addition to a growing volume of work on key economic aspects of the Soviet economy, important collections of scholarly work (Bergson and Kuznets 1963, Bergson and Levine 1983) became guideposts to the field of Soviet economics.

Throughout the postwar era, the Soviet economic experience was chronicled in a variety of additional sources. In addition to general guides (Konn 1992), attention to data problems (Marer et al. 1992, Treml and Hardt 1972) continued. There are a number of basic texts devoted to the Soviet economic experience

(Gregory and Stuart 2000, Nove 1961, Lavigne 1974, Campbell 1974, Sherman 1969).

3. Soviet Studies: The post-Soviet era

Understandably, the collapse of the Soviet empire fundamentally changed the dimensions of Western interest in Soviet-type command economies. Although there has been sustained interest in different economic systems from a scholarly perspective, interest in transition has assumed a central focus, with the demise of the Soviet Union as the primary example of the administrative command economy and the basic changes in the geopolitical framework of East–West relations. Our current interest in the Soviet economic experience is in part as a model for understanding the continuing issue of resource allocation in differing economic systems, interpreting the Soviet era and especially the impact of that era on the contemporary economies of Russia and the newly independent states of the region.

See also: Socialist Societies: Anthropological Aspects; Soviet Studies: Geography; Soviet Studies: Politics; Soviet Studies: Society

Bibliography

Bergson A 1961 *The Real National Income of Soviet Russia Since 1928*. Harvard University Press, Cambridge, MA

Bergson A, Kuznets S 1963 *Economic Trends in the Soviet Union*. Harvard University Press, Cambridge, MA

Bergson A, Levine H S 1983 *The Soviet Economy: Toward the Year 2000*. George Allen & Unwin, London

Berliner J 1957 *Factory and Manager in the USSR*. Harvard University Press, Cambridge, MA

Campbell R 1974 *The Soviet Type Economies: Performance and Evolution*, 3rd edn. Houghton Mifflin, Boston

Davies R W 1980 *The Industrialization of Soviet Russia*, 3 vols. Harvard University Press, Cambridge, MA

Granick D 1954 *Management of the Industrial Firm in the USSR*. Columbia University Press, New York

Green D W, Higgins C I 1977 *Sovmod I: A Macroeconometric Model of the Soviet Union*. Academic Press, New York

Gregory P R, Stuart R C 2000 *Russian and Soviet Economic Performance and Structure*, 7th edn. Addison Wesley Longman, Reading, MA

Grimsted P K 1988 *Archive and Manuscript Repositories in the USSR, Ukraine and Moldavia*. Princeton University Press, Princeton, NJ

Hewett E A 1988 *Restoring the Soviet Economy*. Brookings Institution, Washington, DC

Konn T (ed.) 1992 *Soviet Studies Guide*. Bowker-Saur, London

Kornai J 1992 *The Socialist Economy: The Political Economy of Communism*. Princeton University Press, Princeton, NJ

Lavigne M 1974 *The Socialist Economies of the Soviet Union and Europe*. Martin Robertson, London

Marer P, Arvay J, O'Connor J, Schrenk M, Swanson D 1992 *Historically Planned Economies: A Guide to the Data*. The World Bank, Washington, DC

Millar J R (ed.) 1987 *Politics, Work, and Daily Life in the USSR*. Cambridge University Press, New York

Nove A 1961 *The Soviet Economy: An Introduction*. Praeger, New York

Nove A 1982 *An Economic History of the USSR*. Penguin, New York

Ofer G 1987 Soviet economic growth, 1928–1985. *Journal of Economic Literature* **25**(4): 1767–833

Sherman H J 1969 *The Soviet Economy*. Little Brown, Boston

The National Economy of the USSR (Narodnoe khoziaistvo SSSR). Annual since 1956. Statistika, Moscow

Treml V G, Hardt J P 1972 *Soviet Economic Statistics*. University of North Carolina Press, Durham, NC

Zaleski E 1980 *Stalinist Planning for Economic Growth, 1933–1952*. University of North Carolina Press, Chapel Hill, NC

R. C. Stuart

Soviet Studies: Education

The vast majority of Soviet and Western studies of Soviet education since the 1930s have focused in positive ways on the curricular and instructional rigor of the educational system, and its contribution to Soviet social and economic development. The official line that dominated Stalinist and postwar Soviet research argued that the history of Soviet education was essentially unproblematic, with the Communist party–state leading the people upwards into enlightenment through literacy, science and technology, and militant materialism. Perhaps ironically, much of this was echoed in contemporary Western research and in modernization theory, which also considered the contribution of Soviet education to social and economic development to be direct and unproblematic. The only real point of contention was over the ultimate effects of rising literacy and the growth of an increasingly educated mass society. While Soviet scholars argued that these developments would lead to the realization of true Communism, Western scholars tended to argue that they would result in inevitable social differentiation and cultural liberalization on the Western model. Neither group seemed to consider the possibility that the Soviet educational system could be maladaptive or dysfunctional in its ability to contribute to sustainable social and economic development.

This unanimity in almost all studies of Soviet education since the 1930s also obscured the distinct historical phases and sharp policy conflicts that actually characterized the history of Soviet education. In fact, this history is much more contentious and significant for comparative international developments than the official Soviet line would suggest. As will be detailed below, the years after the Russian revolutions of 1917 were characterized by an official

Soviet policy of radical progressive education that was very much influenced by, and contributed to, international trends in educational policy. This was followed by increasingly conservative educational policies in the Stalinist 1930s and afterwards, when the enduring foundations of the Soviet educational system were created. In fact, while obscured by Soviet social and behavioral science, which always claimed that Soviet developments were utterly unique, this traditionalist restoration was actually also part of a broader international turn away from Deweyan progressive education in many nations during the middle decades of the twentieth century. Many of these historical shifts and policy changes, and the political and professional struggles that shaped them, have long been obscured by official Soviet historiography. The limits and failures of the Soviet educational system in terms of actual access and achievement have also long been obscured by opaque official Soviet statistics. More recent post-Soviet research, both Western and Russian, has begun to detail how, throughout its history, Soviet education was successful in many ways, and yet also heavily politicized, narrowly vocational, bureaucratically inflexible, and subordinated to the Soviet state and the Communist Party of the Soviet Union. Ultimately, it could be argued that the inflexibility of the system, and its difficulty in adapting to the needs of postwar social and economic development, caused Soviet education to become a serious obstacle to the viability of Soviet socialism. During the post-Stalinist period, from the 1950s through the 1980s, various reforms were attempted in an effort to 'humanize' and differentiate the highly bureaucratized and standardized system. The alleged failure of these reform efforts in the 1980s then led, in the post-Soviet years after 1991, to attempted radical reforms, questioning of the entire Soviet legacy, and deep systemic crisis in the 1990s.

1. The Soviet Educational Experiment and Radical Progressive Education

The models for what became Soviet education were sketched out in the decades before the Russian revolutions of 1917 in response to the need for rapid social and economic development, and especially in response to the poverty and social dislocations caused by World War One. The fundamental principles of Soviet progressive education included expanding the conception of education and culture to include a broad array of programs for adults, working people, and children of all ages. These policies also entailed the provision of extensive public health and welfare services through educational institutions, and a radical rethinking of the nature of curriculum and instruction to bring learning 'closer to life' and industrial work. Finally, educational access and promotion were to be radically expanded to make education open to all regardless of ethnic identity, class, or gender (Pinkevitch 1929).

These early Soviet educational policies were inspired by European and American progressive education, and were hailed as revolutionary by Western radicals and educators. However, Soviet education then went far beyond Western influences in its emphasis on including previously disadvantaged populations, as well as in its increasingly dogmatic emphasis on political indoctrination (Harper 1929). Soviet educational experiments in the 1920s were, in many ways, far ahead of their time (on this and what follows, see Kalashnikov and Epshtein 1927–1929). Soviet educators sought to synthesize general and vocational education for all students in a common curriculum that was known as polytechnicalism, although this rapidly proved impractical, and some vocational tracking was reintroduced. The educational experiments of the 1920s also included the use of interdisciplinary and thematic curricula, and sophisticated developmental instruction that was targeted to stages of students' psychological and intellectual development. Other innovations included student and faculty self-government, a focus on experiential learning and community service, and the curtailing of most formal assessments and criteria for promotion in the name of equality and universal access to education. Early Soviet education was also characterized by the extensive use of Western-style child study and psychological testing, which was known as pedology in Soviet practice. Resources were also invested in the mass provision of new media, libraries, and extracurricular activities for all ages. Finally, the 1920s also witnessed the development of written languages, new alphabets, primers, and comprehensive educational services for many indigenous peoples and small ethnic minorities.

Almost all accounts in the 1920s, both Soviet and Western, were extremely positive about these early progressive experiments, although some Western observers such as John Dewey were also critical of the Soviet educators' emphasis on ideological indoctrination and their philosophical dogmatism (Dewey 1929). For all of the early Soviet educators' idealism, however, there were clearly also many internal contradictions and serious problems with the progressive educational experiments of the 1920s. For example, while expanding access to previously disadvantaged populations, the regime also sought to deny educational opportunity to members of the prerevolutionary elites, or to imagined political or class enemies. This created serious social tensions, and drove out of the new system many of those who could have become leading teachers and students. Public education was often funded on a 'residual' basis, which meant that it was financed only after other budgetary obligations were met, even as the Soviet regime struggled to prevent private financing or tutoring from compensating for inadequate state provision. Nonetheless, early Soviet progressive edu-

cation, for all its utopianism and politicization, was undeniably an unprecedented and truly revolutionary experiment in the provision of mass education and culture for all.

2. *Stalinist Traditionalism and System Building*

However, as detailed in subsequent Soviet and Western research, the many problems with these early experiments effectively discredited them, and provoked a sharp reaction in the 1930s. Soviet accounts during the Stalinist period tended to blame the progressive educators for their allegedly 'anti-Leninist' errors, and often accused them of deliberate sabotage and counter-revolutionary 'wrecking' (Svadkovskii 1938). In fact, many of the progressive educators were purged in the 1930s, and their lives and work erased from the historical record, to be partially rediscovered only when the Soviet archives began to open up in the late 1980s (Fradkin 1990). Later Soviet accounts glossed over the distinctive policies of the 1920s, obscured the sharp conflicts of the time, and blamed the progressive educators for their impracticality and adherence to inappropriate 'bourgeois' theories (Kuzin et al. 1980). Most Western analyses simply argued that the progressive policies were not conducive to true social and economic modernization, and that it was virtually inevitable that more 'stable' and traditional curricula and instruction be restored to fit the Soviet educational system to the needs of modern industrial society (Anweiler 1964). Other Western analyses of the failure of the progressive educators focused on problems such as the acute lack of funding that crippled the attempted reforms in the 1920s, or the disruptions caused by conflicting party–state policies. Later Western authors focused on the popular and political backlash against the progressive experiments, and the broad social support for the upward mobility opened up by Stalinist educational policies (Fitzpatrick 1979). Almost all accounts since the 1930s agree that the progressive experiments became increasingly unpopular as they were accused of undermining the authority of teachers and parents, weakening curricular integrity and instructional rigor, and failing to produce the highly-trained technical specialists and skilled workers needed to 'build socialism.' From a broader and more critical perspective, with the coming of Stalinism in the 1930s, the task became less to mobilize and empower all citizens through progressive education, and more to discipline and train subjects for state-directed development.

Struggles over educational policy raged on in the 1930s, and resulted in the thorough repudiation of progressivism and the forging of a new and intensely conservative model of Soviet education. This also entailed the rejection of all child-centered or developmental schooling in the name of the struggle against 'pedological deviations,' and a return to a school-centered system and highly ritualistic formal 'didactics' (Kairov 1939). This Stalinist model was intensely traditional in curriculum and instruction; thoroughly politicized, especially in the social sciences; heavily dominated by the physical sciences and technology; and rigidly bureaucratized in its governance. From a broader perspective, it could be argued that the system of rigid party–state control over Soviet intellectual and professional life that resulted from these struggles became, over the following decades, debilitating for Soviet education as a whole. The fundamental principles of this Stalinist or traditionalist restoration are well known through the international attention they received in the postwar era, especially after the launch of the first satellite (Sputnik) in 1957. This system of Stalinist education entailed curricula organized along the lines of the traditional subject disciplines, and instruction that focused on the content of knowledge and its reproduction. The system was built upon a canon of standardized textbooks, lesson plans, and teaching guides; and emphasized lecturing, memorization, and rigorous oral examination. A central goal throughout was to reinforce the authority of the teacher or professor over their students, and of school directors and administrators over their staff. Formal assessments and rigid criteria for promotion were elaborated, with harsh sanctions for failure, lack of discipline, or dissidence. Many of the programs that sought to cultivate the languages and cultures of indigenous peoples and national minorities were curtailed, and many ethnic groups subjected to involuntary linguistic Russification and cultural Sovietization, often enforced through police terror. Overall, Soviet education in the 1930s and after became increasingly rigorous, hierarchical, and formal (Medynskii 1951). Stalinist education was permeated with a spirit of discipline and order, and was closely coordinated with the Soviet planned economy and national security state.

There were, undeniably, real successes for Soviet education in the Stalinist period, especially in raising literacy rates and building a comprehensive public education system. It also seems that the traditionalist restoration was quite popular with the new Soviet elites, many professional educators and teachers, and many culturally conservative parents. The system did produce unprecedented numbers of technical specialists and skilled and semiskilled workers, albeit in increasingly narrow vocational specializations. The system also proved successful at research and development, especially in mathematics, the physical sciences, and military technologies. As the Stalinist regime embraced conservative educational practices and cultural values, it also fostered significant compensatory legitimacy for Soviet socialism, as demonstrated by both Soviet and émigré memoirs. Yet for all its claims to universality and unanimity, unofficial cultural and educational traditions did survive, especially among religious and national minorities, as

well as among the Soviet intelligentsia in the form of unsanctioned teaching practices, and private and family tutoring (Kline 1957).

Official Soviet research in this period was, by definition, triumphal about the onward and upward march of literacy and enlightenment, and the indispensable leadership of the Soviet party–state (Konstantinov and Medynskii 1948). There was also a distinct tone of xenophobia, with Soviet developments depicted as both indigenous and superior to all comparable international trends. The emphasis on the role of state power tended to obscure the more complex relations between the party–state and admittedly dependent elements within the profession, as well as the role and interests of the Soviet public. While the disillusioned Western progressive George S. Counts was quite harsh in his judgement of the repressive and dogmatic aspects of Soviet education, he also acknowledged its rigor and utility for a forced-march version of modernization (Counts 1957). Almost all other Western and Soviet émigré accounts were positive in their judgment of the system's pedagogical principles and effectiveness, even if they sometimes acknowledged its rigidity and limitations.

However, alongside the successes and popularity of the conservative policies, there were severe problems as well. The human cost of Stalinist terror in education and culture was devastating, and resulted in the repression of many talented researchers, educational leaders, and teachers, especially among the national minorities. While adhering to the basic principles of modernization theory, some Western authors did draw attention to the more ambiguous effects of standardized Soviet education for the development of the non-Russian peoples of the Soviet Union (Bereday et al. 1971). Soviet education in this period became increasingly dogmatic and dominated by rigid official orthodoxy, with especially devastating effects in the social sciences and humanities. The system became top heavy and driven by administrative interests, with little space for professional autonomy or innovation. Research was often separated from teaching, and there was a clear assumption that theory and content should drive instructional practice. Furthermore, educational institutions and professional networks were isolated from one another through bureaucratic parallelism, as each economic sector and industry created its own educational and research hierarchies. Educational institutions and even individual teachers were closely supervised, even as many contrived elaborate and creative ways to subvert party–state controls. Most importantly, the entire system was structured into narrow vocational specializations that limited the development of the job skills and professional talents that were needed for sustainable postwar social and economic growth. These more critical perspectives only began to coalesce in the 1990s as some post-Soviet and Western authors began to focus on the more maladaptive or dysfunctional aspects of Stalinist educational policies. In conclusion, and taken as a whole, the Stalinist reforms in Soviet education successfully built a mass public education system, but at the price of bureaucratic rigidity, intellectual stagnation, and vocational obsolescence.

3. *The Potential and Limits of Postwar Reform*

Ironically, even as Western researchers were popularizing the curricular and instructional rigor of Soviet education in the post-Stalinist period after 1953, Soviet educators began to debate the inflexibility and limits of their own system. Attempted postwar reforms encompassed both more 'progressive' as well as quite 'traditional' ideas. For example, an attempted reform associated with Communist Party leader Nikita Khrushchev in 1958 sought to restore the polytechnical principles of the system, only this time as more rigorous and involuntary vocational tracking. This attempted reform sought to 'force down' the educational aspirations of the Soviet public through a requirement that all students entering higher education have at least two years of work or production experience (Goncharov and Korolev 1960). Significantly, this reform was essentially blocked by the combined resistance of enterprise managers, parents concerned about their childrens' upward mobility, and more traditionalist educators resistant to such experiments. The regime also attempted, throughout the post-Stalinist period, to strengthen proper 'civic–political' socialization in response to what was seen as growing juvenile delinquency and alienation from official ideological values.

Other reforms in the post-Stalinist period reflected the progressive innovations of the 1920s, especially the tradition of psychological research pioneered by Lev S. Vygotskii (1896–1934, see Vygotskii 1991). Vygotskean cultural psychology and developmental instruction, which also received significant attention from Western researchers in the 1980s and 1990s, may well prove to be the most internationally significant and enduring contribution of Soviet education. Other attempted reforms included greater curricular differentiation, the introduction of electives for students, and the development of programs for elite education in the arts, sciences, and sports (Matthews 1982). There was also a relaxation of linguistic Russification, with relatively more flexibility offered to educational institutions in the national minority regions. Yet it must be noted that, for all the potential of these reforms, they were often thwarted by the combined power of the party–state regime and the conservative educational establishment, especially in the various republic-level Ministries of Education and the Academy of Pedagogical Sciences.

Taken as a whole, the reform movements in Soviet education from the 1950s through the early 1980s did have the effect of consolidating the system, while also

diversifying and enriching curricular and instructional practices. Perhaps most importantly, the Soviet educational system by the 1970s had trained a vast new body of professional teachers, and essentially achieved universal literacy and nearly universal secondary education. Not surprisingly, official Soviet accounts focus on these achievements with little or no acknowledgement of any persistent problems or struggles over policy or practice (Kolmakova et al. 1986). Yet these achievements were, to a significant degree, gained at the expense of individual and professional freedom, creativity, and systemic flexibility. It must also be noted that, for all of its rhetoric of equality and universality, severe inequities persisted, especially between urban and rural areas, and that literacy and educational provision lagged behind in many national minority areas. The Soviet educational system was also slow to introduce new instructional technologies, and was especially weak in the provision of non-political continuing education and vocational retraining. It seems that little Western attention was paid to these systemic problems, perhaps in part because the significant Western research on Soviet education, which began in the 1950s, had narrowed considerably by the late 1970s. It also seems that international cooperation in this area was constrained by the conservatism of the neoStalinist educators and officials who continued to dominate the Soviet system into the early 1980s.

4. *The Crisis of Soviet Education and its Ambiguous Legacies*

The final phase of Soviet education in the 1980s and early 1990s witnessed both radical reforms and severe systemic crises. Several attempted reforms in the mid-1980s sought to tighten vocational tracking and increase party–state controls, but in a very real sense, Soviet educated society as a whole, as well as leading elements within the educational profession, were no longer willing to tolerate such reactionary policies. There was a growing sense that Soviet public education was in crisis, which led to the emergence of a growing 'innovation movement' among educators and teachers. This movement sought to humanize and democratize Soviet education; to loosen party–state controls over curriculum, instruction, and governance; to recover the spirit of the child-centered progressive experiments of the 1920s; and to expand contacts with international researchers and practitioners (Eklof and Dneprov 1993). After the collapse of the Soviet Union in 1991, the newly independent republics embraced an array of educational policies. Some tried, unsuccessfully, to simply preserve the old Soviet system, as in Belarus and parts of Central Asia. Others struggled to radically reform the old system through attempted privatization and decentralization, as in the Baltic states and the Russian Federation (on Russia, see Dneprov et al. 1991). The reforms of the 1990s have, undeniably, unleashed a veritable pedagogical revolution of innovation in curriculum and instruction, and finally joined together the long isolated Soviet educational community with its international counterparts. Much of this reintegration into the world community of educational research and practice has been facilitated by an unprecedented level of international assistance provided to post-Soviet educators and students in the form of research grants and exchanges. Yet, it must also be noted, all of the Soviet successor states have experienced severe if not catastrophic crises in the funding and provision of public education, and that both excellence and equity have suffered. In other words, now that post-Soviet educators have finally rejoined the world community from which they have been isolated since the 1920s, many lack the financial resources to remain in the profession or to take full advantage of the new opportunities for international or even domestic professional cooperation and development. Furthermore, the economic and demographic crises that have gripped the entire region threaten the health and readiness to learn of virtually an entire generation of children and young adults.

5. *Conclusions*

In conclusion, Soviet education both shaped and constrained the building of Soviet socialism. While Soviet and Western studies of Soviet education diverged sharply on some points of interpretation, almost all agreed that the Soviet educational system, as built in the 1930s and as endured into the 1980s, was fundamentally functional and conducive to modernization. While later Russian and Western authors have expressed some skepticism about the viability and functionality of the Soviet educational system, it did undeniably contribute, if in complex ways, to the development of a modern and literate mass society. Finally, the most rigid and repressive aspects of Soviet education were repudiated in the 1990s, and there have been some real successes in recent reform efforts. However, these years may ultimately be judged a historical tragedy of monumental proportions if the legacy of secular, free, and public education for all that was pioneered by Soviet educators is lost as well.

Bibliography

Anweiler O 1964 *Geschichte der Schule und Padagogik in Russland, vom Ende des Zarenreiches bis zum Beginn der Stalin-Ara*. Quelle & Meyer, Heidelberg, Germany

Bereday G Z F, Pennar J, Bakalo I 1971 *Modernization and Diversity in Soviet Education*. Praeger, New York

Counts G S 1957 *The Challenge of Soviet Education*. McGraw Hill, New York

Dewey J 1929 *Impressions of Soviet Russia and the Revolutionary World, Mexico–China–Turkey*. New Republic, New York
Dneprov E D, Lazarev V S, Sobkin, V S (eds.) 1991 *Rossiiskoe obrazovanie v perekhodnoi period: programma stabilizatsii i razvitiia* [*Russian Education in the Transition Period: Program of Stabilization and Development*]. Ministry of Education of the RSFSR, Moscow
Eklof A B, Dneprov E D 1993 *Democracy in the Russian School: The Reform Movement in Education since 1984*. Westview, Boulder, CO
Fitzpatrick S 1979 *Education and Social Mobility in the Soviet Union, 1921–1934*. Cambridge University Press, New York
Fradkin F A 1990 *A Search in Pedagogics: Discussions of the 1920s and 1930s*. Progress, Moscow
Goncharov N K, Korolev F F (eds.) 1960 *Novaia sistema narodnogo obrazovaniia v SSSR: Sbornik dokumentov i statei* [*The New System of Public Education in the USSR: Collection of Documents and Articles*]. Academy of Pedagogical Sciences of the RSFSR, Moscow
Harper S N 1929 *Civic Training in Soviet Russia*. University of Chicago Press, Chicago
Kairov I A 1939 *Pedagogika* [*Pedagogy*]. Uchpedgiz, Moscow
Kalashnikov A G, Epshtein M S (eds.) 1927–1929 *Pedagogicheskaia entsiklopediia* [*Pedagogical Encyclopedia*], 3 vols. Rabotnik prosveshcheniia, Moscow
Kline G (ed.) 1957 *Soviet Education*. Columbia University Press, New York
Kolmakova M N, Gellershtein L S, Ravkin Z I (eds.) 1986 *Ocherki istorii pedagogicheskoi nauki v SSSR (1917–1980 gg.)* [*Notes on the History of Pedagogical Science in the USSR, 1917–1980*]. Pedagogika, Moscow
Konstantinov N A, Medynskii E N 1948 *Ocherki po istorii sovetskoi shkoly RSFSR za 30 let* [*Notes on the History of the Soviet School in the Russian Soviet Federated Socialist Republic at 30 Years*]. Gosuchpedgiz, Moscow
Kuzin N P, Kolmakova M N, Ravkin Z I (eds.) 1980 *Ocherki po istorii shkoly i pedagogicheskoi mysli narodov SSSR, 1917–1941 gg*. [*Notes on the History of the School and Pedagogical Thought of the Peoples of the USSR, 1917–1941*]. Pedagogika, Moscow
Matthews M 1982 *Education in the Soviet Union: Policies and Institutions Since Stalin*. Allen and Unwin, London
Medynskii E N 1951 *Public Education in the USSR*, 2nd edn. Foreign Languages Publishing House, Moscow
Pinkevitch A P 1929 *The New Education in the Soviet Republic*. John Day Company, New York
Svadkovskii I F (ed.) 1938 *Protiv pedologicheskikh izvrashchenii. Sbornik statei* [*Against Pedological Deviations. Collection of Articles*]. Gosuchpedgiz, Moscow-Leningrad.
Vygotskii L V 1991 *Pedagogicheskaia psikhologiia* [*Pedagogical Psychology*]. (Davidov V V (ed.), Pedagogika, Moscow

M. S. Johnson

Soviet Studies: Environment

Environmental studies are by nature interdisciplinary, and this holds for Soviet studies of the environment. Yet approaches to the study of the Soviet environment were, like many topics, politicized within the USSR. Secrecy and ideological constraints made it difficult for Soviet biologists, geographers, and economists to discuss the full extent of pollution or the structural factors responsible for environmental waste and degradation. Political aspects of pollution were especially sensitive. Social scientists could at best only obliquely criticize the central planning, bureaucratic socialism, and militarized economy that were responsible for much of the country's environmental ills. Western experts had to rely on fragmented and incomplete data, yet they managed to construct reasonably accurate accounts of Soviet environmental practice.

The Soviet approach to the natural environment encapsulated, *in extremis*, all the elements that James C. Scott (1998) has identified as necessary for disastrous twentieth-century social engineering: the administrative ordering of nature and society; a high-modernist ideology relying on science and technology for problem solving; an authoritarian state fully willing to use its coercive powers for social experimentation; and a prostrate civil society unable to resist the authoritarian state's transformative plans. Seven decades of rapid, forced industrialization, ill-conceived agricultural experiments, botched industrial and irrigation projects, waste and mismanagement of natural resources, and disregard for the human consequences of development left a legacy of environmental destruction unrivaled in Europe or North America. With the collapse of communism the obstacles to serious environmental analysis have been removed; however, economic decline and political paralysis in the postcommunist era have severely constrained the former USSR's ability to address its environmental crisis.

1. Ideology and Attitudes

Soviet practices and policies toward the natural environment mirrored the political system's *raison d'être*. In ideology, the Soviets turned the environmental determinism of Marxism on its head, glorifying the radical transformation of nature through sheer political will. Soviet leaders formulated socialism into a doctrine that equated anthropogenic transformation of the environment with progress. As political scientist Joan DeBardeleben noted in *The Environment and Marxism–Leninism* (1985), ideology provided a conceptual framework that shaped and limited policy discourse.

The Soviet dictatorship showed respect neither for nature nor for humankind. Both were subordinated to the regime goal of constantly expanding material production, without regard for the wellbeing of the population, their quality of life, or for the natural environment. From an environmental perspective, the Soviet fixation with production focused the question: production to what end? However, Soviet scholars were not allowed to address such questions: while environmental economists in the West discussed

capitalist production in terms of the external costs associated with industry, Soviet economists had to echo the Party position that externalities were not possible in a socialist economy.

Soviet environmental studies, policies, and practices were molded and distorted by communist ideology. In the *Communist Manifesto*, Marx and Engels had praised capitalism's success in subjecting nature to man and machinery; Soviet communism's goal was to surpass this record of accomplishment. In addition, Marxism's labor theory of value denigrated natural resources to a position of secondary importance in the economy. This attitude was compounded by the belief that the USSR, by far the largest nation on earth, had illimitable stores of coal, land, timber, natural gas, and minerals. Only late in the history of the USSR did the regime acknowledge limits to exploiting nature. Even then, scholars focused on narrow technical issues of resource use and pollution, to avoid placing blame on government policies (Gerasimov et al. 1971).

2. *Industrial Development and Pollution*

Vladimir Lenin was favorably inclined toward conservation, including the system of nature preserves (*zapovedniki*) that predated the revolution. His primary emphasis, however, was on modernizing Russia in order to lay the foundations for communism. Joseph Stalin's comprehensive program of forced industrialization and agricultural collectivization embodied a strong anti-environment bias. Soviet developmental policies resulted in an all-out assault on nature by planners and workers. This determination to transform the natural world was comparable to the high modernist ideology found in late nineteenth- and early twentieth-century industrializing nations; layered on top, however, was the Communist Party's goal of subordinating all social and productive forces to its conscious guidance, and the insulation of the governing elites behind the regime's authoritarian political structure. The focus on industrial growth as the system's preeminent goal made continuous, linear progress a vital component of regime legitimacy, particularly after Marxist–Leninist ideology lost its appeal.

Maintaining conservation values was problematic in the highly-charged political atmosphere of Soviet politics. Some scientists urged caution in exploiting the natural environment, even during the most repressive period of Stalinism, at great personal risk, and small, relatively independent social movements dedicated to protecting the environment persisted into the Stalin era. Historian Douglas Weiner describes these subtle forms of resistance in his book *A Little Corner of Freedom* (1999). But all of Soviet science, including ecology and conservation, was adjusted to support the regime's political goals. Weiner's earlier *Models of Nature* (1988) details the impact of Stalinism on conservation in the Soviet Union. For example, acclimatization—the introduction of non-native species into certain environments in order to transform nature and enhance productivity—was advocated by the biologist Trofim Lysenko during the Stalin and Khrushchev eras. Constrained by a poor understanding of science and arrogance toward nature, the regime's acclimatization efforts proved disastrous for both the introduced species and the local ecology. Soviet biologist Zhores Medvedev chronicled how ideology, politics, and careerism affected Soviet genetics in *The Rise and Fall of T.D. Lysenko* (1969). More conventional genetics replaced Lysenkoism by 1966, after Party First Secretary Nikita Khrushchev had been ousted.

Geographers and economists wrote the earliest Western studies of Soviet environment. Philip Pryde's *Conservation in the Soviet Union* (1972) and economist Marshall Goldman's *The Spoils of Progress* (1972) laid the groundwork for later studies. These scholars found the major polluters of the environment to be heavy industry, particularly the energy, chemical, and metallurgical sectors favored by Soviet planners. Large urban areas suffered air and water pollution caused by dense concentrations of steel plants, oil refineries, ore smelters, and pulp and paper mills. Automobile pollution was generally less of a problem, since few people owned cars, but poorly maintained buses and trucks spewed tons of pollutants into the air. Contamination from radioactive waste, military installations, and nuclear power plants posed another serious hazard to public health. Agricultural runoff and inadequately treated industrial wastes polluted much of the Soviet Union's surface and subsurface water sources.

3. *The Late Soviet Era*

Public awareness of Soviet environmental issues emerged in the late 1960s, roughly parallel with the growth of environmentalism in the developed, industrialized democracies. The Soviet regime's focus on industrial development, official secrecy toward all systemic failings, and Communist Party assertions that environmental degradation characteristic of capitalist economies was not possible under socialism made it dangerous to criticize government policies openly. But the threatened pollution of Lake Baikal, the largest body of fresh water on the planet, led to open controversy over the building of two cellulose plants on its shores. Grigorii Galazii, Director of the Limnological Institute of the Siberian Academy of Sciences, led a crusade in the controlled press against the danger to the lake. Blunter criticism was expressed in an underground (*samizdat*) publication, the *Destruction of Nature*, published in the West by a pseudonomous Boris Komarov (Ze'ev Wolfson) (1978). Carefully managed public discussions on the environment did take place during the Brezhnev era, largely to highlight the regime's 'commitment to

environmental protection' and publicize ineffective factory managers, bureaucrats, or poachers.

Genuine discussion of environmental problems, and serious efforts to combat pollution and environmental degradation, however, were constrained by the authoritarian character of the system and the emphasis on covering up failings. A major theme in Komarov's manuscript was the pernicious effect of censorship and secrecy. The Soviet military frequently invoked national security to justify a range of environmentally destructive practices, such as locating cellulose plants to produce cord for aircraft tires on Lake Baikal, dumping nuclear wastes underground and in the oceans, and poaching in *zapovedniki*. But environmental destruction took its toll on the military, as it did on civilians. Recruits were often unfit for military duty because of poor health. Unhygienic living conditions, including contaminated drinking water and air pollution, led to high incidents of dysentery, skin disease, and respiratory illnesses within the military (Feshbach and Friendly 1992).

The Soviet experience forced environmentalists and economists to rethink many of their assumptions. Economic theory had held pollution to be an externalized cost to society resulting from production by private enterprise. Theoretically, social ownership internalized these costs, and a socialist economy therefore should pollute less than a capitalist economy. However, natural resources were not treated as scarce in the Soviet economy, which resulted in tremendous waste and inefficiency. A dominant theme in studies by Western economists (Goldman 1972), geographers (Pryde 1972, 1991), and political scientists (Ziegler 1987) was that concentrating power in the Communist Party and state bureaucracy inhibited the formation and implementation of responsible environmental policies. Socialist, centrally planned economies did not fully internalize external costs; rather, the absence of clearly identifiable property owners encouraged waste and abuse of the environment.

Soviet agriculture was possibly more environmentally destructive and more wasteful of natural resources than industry. Millions of hectares of land were impoverished or eroded by unscientific attempts to extract higher yields. Lysenko's ill-considered experiments in agrobiology caused significant damage to Soviet agriculture, as Medvedev (1969) documents. Khrushchev's Virgin Lands project and his determination to plant maize on unsuitable land constituted additional examples of gross agricultural mismanagement. Chemical fertilizers and pesticides were overused and applied haphazardly on the collective and state farms through ignorance, carelessness, and a skewed incentive structure. In Central Asia, diversion of the Amu Darya and Syr Darya rivers for cotton irrigation depleted the Aral Sea, saturated the land with salt and chemicals, and altered the region's microclimate. Structurally, socialized ownership of agricultural land generated a psychology in which farm workers and administration valued only short-term production and ignored long-term damage to the earth. Farmers were more careful with their private plots (and these were far more productive), but only 3 percent of land was set aside for private agriculture.

Western political scientists naturally argued that politics played an important role in Soviet environmental degradation, claiming authoritarian political systems were less responsive to environmental concerns than democratic, pluralist systems. Charles Ziegler, in *Environmental Policy in the USSR* (1987) found the lack of accountability to voters and the absence of environmental interest groups in the USSR ensured that decision makers could safely ignore public opinion. Until Gorbachev's *glasnost* eased controls, scholars within the USSR could publicly criticize neither the general policies nor the specific decisions of high Party and government officials.

Weak adherence to the rule of law also contributed to environmental degradation. The Soviet government adopted impressive environmental and conservation legislation, but these laws were largely symbolic and were consistently ignored by factories and ministries tasked with meeting production quotas. Enforcement of environmental regulations by government agencies was sporadic, and environmentalists could not utilize lawsuits, an effective tool of their Western counterparts.

Soviet scientific studies and government pronouncements consistently maintained that socialist economic planning, the national ownership of land and resources, and centralized regulation ensured a superior approach to environmental protection that could not be realized under capitalism. But in the Brezhnev period cautious scientists admitted that, in practice, natural resource use was inefficient and wasteful, and that erosion and water pollution were serious problems (Gerasimov et al. 1971). Studies by Western scholars suggested that Soviet environmental performance was no better than that of capitalist states, and often much worse, due to the distorting influence of bureaucratic state socialism. Using a comparative approach, political scientist Barbara Jancar found that political–economic factors in one-party systems played a greater role in defining environmental management than did regulation, regardless of whether the state was centralized, as was the USSR, or decentralized, as was Yugoslavia (Jancar 1987).

Perhaps unique to the Soviet Union was the environmental impact of the militarized Soviet economy. Military-related production accounted for approximately one-quarter of the total gross domestic product. Environmental concerns were routinely sacrificed in the interests of national security and, at least until *glasnost*, questioning military imperatives was forbidden. Issues like global warming might be treated skilfully, as they were in *The Night After … Climactic and Biological Consequences of a Nuclear War* (1985),

an edited collection of articles by prominent Soviet scientists. Nonetheless, these essays were published through an official government outlet and served a Soviet national security priority; namely, mobilizing opinion against the Reagan administration's Strategic Defense Initiative.

Given official secrecy and the politicizing of environmental issues, simple descriptive studies were an important contribution to the field, and a damning indictment of the Soviet system. Both Feshbach and Friendly (1992) and Komarov (1978) described how military bases contaminated towns and coastlines with fuel and lubricants, dumped garbage and scrap equipment indiscriminately, and polluted groundwater supplies. Troop exercises destroyed forests and ruined agricultural land. More pollution was generated by the production and storage of chemical weapons, while the production and testing of Soviet nuclear weapons contributed to radioactive pollution at the Semipalatinsk range in Kazakhstan and Novaya Zemlya in the Arctic. The Soviet navy dumped 135 reactors from decommissioned nuclear submarines in the waters around the Barents and Kara seas, and disposed of liquid nuclear waste from the Russian Far East by discharging it into the Sea of Japan. In postcommunist Russia the military continued to obfuscate and deny responsibility for environmental destruction, and officers were prosecuted for exposing the military's role in pollution.

Environmental legislation meant little in a nation not based on constitutional principles. Impressive sounding resolutions on air quality were adopted by the Council of Ministers as early as 1949; in the following decades additional laws were passed to protect water, forests, wildlife, and agricultural land. Scholars and officials could point to such progressive legislation as examples of the regime's commitment to the environment (see Galeeva and Kurok 1979, for the range of legislation). In reality, Soviet legislation consisted of grand pronouncements and largely unenforceable standards. Producers routinely violated environmental laws with impunity. Enforcement was entirely in the hands of government bureaucracy and, given the regime's emphasis on rapid and continuous economic growth, the enforcement of environmental laws that might interfere with production was actively or implicitly discouraged (Ziegler 1987).

4. *Reform and Collapse*

Environmental issues played a key role in the liberalization of the Soviet system during the Gorbachev era. By officially acknowledging the scope of environmental degradation, however grudgingly, the Communist Party admitted failing in its promise to provide a better life for its people, and undermining its claim to legitimacy. Guarded criticism had been voiced by scientists, journalists, and literary figures in the 1960s and 1970s of the issues of air and water contamination, the Aral Sea disaster, and planned diversion of Siberian rivers southward to Central Asia. With the easing of censorship under *glasnost*, and widespread fears of government deception and incompetence after the April 1986 explosion at the Chernobyl nuclear reactor in Ukraine, environmental activism mushroomed and scholars turned their attention to the new environmental politics (Ziegler 1991, Pryde 1991, pp. 246–65). Public opinion surveys discovered that an overwhelming majority of the population was concerned about the environmental problems and the threat to public health. Environmental groups dominated the new 'informals' (*neformaly*), nascent interest groups that emerged in the late 1980s. Women, concerned as they were with pollution-related birth defects and the health of their children and families, figured prominently in these movements.

Many of these spontaneously formed environmental groups linked environmental with national and separatist demands, especially in the more nationalistic Caucasian and Baltic republics. Under Soviet federalism Moscow treated the non-Russian republics like colonies, exploiting their natural resources, degrading the land, locating polluting industries on their territory, and in the process jeopardizing the health of the population. Soviet development strategies also led to an influx of ethnic Russians, who were concentrated in vital economic and political positions within the republics and who diluted the ethnic homogeneity of titular nationalities. Once political controls were relaxed, national discontent coalesced around the more obvious sources of Russian dominance—the Ignalina nuclear power plant in Lithuania, chemical plants in Armenia, nuclear testing in Kazakhstan, and oil pollution along Azerbaijan's coastline.

5. *Environmental Studies and the Transition*

Postcommunist Russia no longer promotes a transformationist ideology. Full representative democracy has proved elusive, but the government is no longer authoritarian. Environmental information is easily available, and there are few formal restrictions on environmental activism. The Transboundary Environmental Information Agency, for example, regularly publishes an electronic *Russian Environmental Digest* of newspaper and magazine articles. Drawing in part on studies of the Soviet environment, political scientists have theorized that democratic political systems should be more responsive to public demands for environmental protection (Payne 1995). But transitional Russians have become more concerned with acquiring material goods than with cleaning up the environment, notwithstanding the serious health implications of the legacy of Soviet pollution. These problems—which include higher morbidity and declining longevity rates, widespread birth defects, and outbreaks of such infectious diseases as diphtheria, cholera, and typhoid—have prompted a demo-

grapher–journalist team (Feshbach and Friendly 1992) to propose the Soviet Union's epitaph as death by ecocide.

Russia's natural environment may have improved in the 1990s as a result of the precipitous decline in industrial production, but the evidence is mixed. Viktor Danilov-Danilyan, head of Russia's State Committee for Environmental Protection, claimed air pollution from industrial sources at the beginning of 1999 was down 35 percent from 1991, while water pollution had decreased by 15–18 percent (Wren 1999) But Boris Porfiriev (1997), a systems analyst with the Russian Academy of Sciences, estimated that in 1995, 41 percent of Russia's population lived in cities where the maximum permissible concentrations of pollutants were exceeded, compared with only 18 percent in the USSR in 1989. While approximately 25 percent of Soviet waste water went untreated in 1989, 40 percent was untreated in Russia in 1995. Overall, investments for environmental protection measures decreased 20 percent between 1990 and 1995 (Porfiriev 1997, pp. 147, 152).

Environmental protection is guaranteed in the 1993 Constitution, but legal provisions remain inadequate. Resource-exploiting companies and ministries occupy advantageous, monopolistic positions relative to environmental agencies and lobby groups. Large budget deficits constrain Moscow's ability to address a range of social problems, of which the environment is only one. Local and national priorities often clash over development and environmental protection issues. Bureaucratic arrogance and ignorance, political neglect, entrepreneurial exploitation, and public indifference relegated environmental protection to a low priority on Russia's political agenda during the first postcommunist decade. And in a move disturbingly reminiscent of Soviet practices, in 2000 President Vladimir Putin dissolved Russia's State Environmental Committee, distributing its oversight functions to the Ministry of Natural Resources, an agency tasked with exploiting the environment.

With political restrictions largely gone, and institution-building on the agenda, a proliferation of environmental studies by social scientists might have been expected. This has not been the case. Western studies of Russia's transition have focused on privatization, the development of a market economy, ethnic questions, and democratization, with few according much attention to the natural environment. Studies by conservationists Michael Wells and Margaret Williams (1998) and Swedish economist Patrik Soderholm (1999) found the absence of legal, administrative, and regulatory institutions in transitional Russia complicated environmental protection measures. Russian natural scientists have continued to pursue environmental studies, but social scientists have not devoted much attention to environmental issues. Several articles in the 1998 and 1999 issues of the Russian journal *Military Thought* addressed various aspects of military-induced pollution, and Boris Porfirev (1997) has dealt with legal and organizational issues. However, there is still a great deal of work to be done on the Russian environment by researchers both inside and outside the country.

See also: Communism; Economic Transformation: From Central Planning to Market Economy; Environment Regulation: Legal Aspects; Environmental Challenges in Organizations; Environmental Change and State Response; Environmental Policy; Environmental Risk and Hazards; Environmental Vulnerability; Environmentalism, Politics of; Environmentalism: Preservation and Conservation; Soviet Studies: Culture; Soviet Studies: Economics; Soviet Studies: Geography; Soviet Studies: Politics; Soviet Studies: Society

Bibliography

DeBardeleben J 1985 *The Environment and Marxism–Leninism: The Soviet and East German Experience*. Westview Press, Boulder

Feshbach M 1995 *Ecological Disaster: Cleaning Up the Hidden Legacy of the Soviet Regime*. Twentieth Century Fund Press, New York

Feshbach M, Friendly A Jr 1992 *Ecocide in the USSR: Health and Nature Under Siege*. Basic Books, New York

Galeeva A M, Kurok M L (eds.) 1979 *Ob okhrane okruzhaiushchei sredy: Sbornik dokumentov partii i pravitel'stva (On The Natural Environment: A Collection of Party and Government Documents)*. Politizdat, Moscow

Gerasimov A P, Armand D L, Yefron K M (eds.) 1971 *Natural Resources of the Soviet Union: Their Use and Renewal*. WH Freeman, San Francisco

Goldman M I 1972 *The Spoils of Progress: Environmental Pollution in the Soviet Union*. MIT Press, Cambridge, MA

Jancar B 1987 *Environmental Management in the Soviet Union and Yugoslavia: Structure and Regulation in Federal Communist States*. Duke University Press, Durham, NC

Komarov B, (Wolfson Z) 1978 *Unichtozhenie prirody*. Possev, Frankfurt, Germany [1980 *The Destruction of Nature in the Soviet Union*. M E Sharpe, White Plains, NY]

Medvedev Z 1969 *The Rise and Fall of TD Lysenko*. Columbia University Press, New York

Payne R A 1995 Freedom and the environment. *Journal of Democracy* **6**: 41–55

Porfirev B 1997 Environmental policy in Russia: Economic, legal, and organizational issues. *Environmental Management* **21**: 147–57

Pryde P R 1972 *Conservation in the Soviet Union*. Cambridge University Press, Cambridge, UK

Pryde P R 1991 *Environmental Management in the Soviet Union*. Cambridge University Press, Cambridge, UK

Scott J C 1998 *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT

Soderholm P 1999 Pollution charges in a transition economy: The case of Russia. *Journal of Economic Issues* **33**: 403–10

Velikhov Y 1985 *The Night After … Climatic and Biological Consequences of a Nuclear War*. Mir, Moscow

Weiner D R 1988 *Models of Nature: Ecology, Conservation, and Cultural Revolution in Soviet Russia*. Indiana University Press, Bloomington, IN
Weiner D R 1999 *A Little Corner of Freedom: Nature Protection from Stalin to Gorbachev*. University of California Press, Berkeley, CA
Wells M P, Williams M D 1998 Russia's protected areas in transition: The impacts of perestroika, economic reform and the move towards democracy. *Ambio* **27**: 198–206
Wren C S 1999 World briefing. *New York Times* 19 January: 9
Ziegler C E 1987 *Environmental Policy in the USSR*. University of Massachusetts Press, Amherst, MA
Ziegler C E 1991 Environmental politics and policy under perestroika. In: Sedaitis J B, Butterfield J (eds.) *Perestroika from Below: Social Movements in the Soviet Union*. Westview Press, Boulder

C. E. Ziegler

Soviet Studies: Gender

This article will discuss the meaning of the term 'gender,' the changing pattern of gender roles throughout Soviet history, and the ways in which Western and Soviet writers have tackled gender issues in the Soviet Union from the time of the Revolution in 1917 up to the country's collapse in 1991.

1. *Clarification of Terms*

'Gender' refers to the personality traits and behavior associated with either men or women that are not rooted in biology but are determined by culture. Western feminists have been careful to distinguish between the terms 'gender' and 'sex,' the latter referring to the biological differences between men and women. However, Soviet writers and thinkers only acknowledged this distinction shortly before the disintegration of the Soviet Union, when, in 1990, the Centre for Gender Studies was founded in Moscow. The Soviet authorities always insisted, throughout the 70 years of the Soviet Union's existence, on their commitment to complete equality for men and women, but no attempt was made to analyze or understand the extent to which differences between men and women were biologically or culturally produced. Alexandra Kollontai, the country's principal theoretician on the 'woman question' at the time of the Revolution, did come close to addressing the concept of gender (though without actually using the term), but dismissed it as an irrelevancy, indeed, as a bourgeois preoccupation. What was important, she argued, was that all people, male and female, developed their personal potential. In her words:

> Leaving it to the bourgeois scholars to absorb themselves in discussion of the question of the superiority of one sex over the other, or the weighing of brains and the comparing of the psychological structure of men and women, the followers of historical materialism fully accept the natural specificities of each sex and demand only that each person, whether man or woman, has a real opportunity for the fullest and freest self-determination, and the widest scope for the development and application of all natural inclinations (Kollontai 1977, p. 581).

Yet despite their refusal to address the issue of gender openly, Soviet politicians were in reality very much concerned with it, and made determined attempts at different times either to eradicate or strengthen the cultural differences between men and women.

2. *Historical Overview*

In its first decade the Soviet Union concerned itself with challenging the traditional understandings of appropriate male and female behavior and social roles. At a time when the West was promoting and preserving separate male and female spheres, Soviet propagandists were urging women to leave that female bastion, the home, and enter the public world of work and politics, until then a largely male environment. Women were assured that State institutions would take over their domestic tasks. In any case, the family would become less important as people got used to living communally. Some theorists went so far as to argue that the family would serve no purpose at all in a communist society and in due course would simply 'wither away.'

However, there was no attempt to rethink both male and female roles. Women were simply to be incorporated into the hitherto male sphere of public service. Nor were women to be truly free of domestic responsibilities. Articles on the organization of communes made it clear that women would still be responsible for the cooking, laundry, child care, etc., but they would now take on this work collectively, for the commune as a whole, rather than individually, for their own families. There was also no question of them choosing not to have children. Rearing the next generation of workers was a patriotic duty, not a choice.

Widespread unemployment in the 1920s meant that few women were actually drawn into social production. This began to change in the late 1920s, when Stalin came to power. In 1929 he launched a huge industrialization drive. All workers were now needed, women as well as men, and by 1932 the number of women workers was almost double what it had been in 1928. At the same time, the pledge to free women from their traditional domestic duties was quietly abandoned. All resources were now required for industry, and if women were still willing to perform those traditional functions for free, the State would hardly fight for the right to fund them. Predictions about the demise of the family also ceased. On the contrary, the family was now held up as the 'basic cell' of Soviet society. There are a number of reasons for this change in approach. The first Five-Year Plan had produced an enormous social upheaval, one of the results of

which was an alarming drop in the birthrate, and it was hoped that rehabilitating the family might help reverse this trend. The familiarity of traditional family life might also provide some stability in people's lives in the face of such radical change. Furthermore, the family now took on a symbolic function; Soviet society as a whole was depicted as one large family, with Stalin the ultimate patriarch. Far from being an outdated institution that had no place in a communist society, the family assumed a new significance.

Women now found themselves laboring under a 'double burden.' They were required to go out to work: there were simply not enough men to bring about industrialization at the speed or to the extent required by Stalin; but they also had to perform their traditional domestic and nurturing functions. They were to do this with no assistance from men, nor, given the concentration on heavy rather than consumer industry, from any services or products that could relieve the domestic workload. Yet this combination of functions was ironically hailed as evidence of women's equality, and the basis of their self-fulfillment. The 'woman question' was declared solved.

When Khrushchev came to power and launched his 'destalinization' program, the 'woman question' was partially reopened once again. The press on the whole continued to applaud the country's successful establishment of women's equality, but some authors now acknowledged that the real situation was more complex, and that women fell behind men in both professional and political hierarchies. At the same time, sociologists began conducting 'time budget studies' which made it clear that women devoted an enormous amount of time to domestic work, and suggested that this was a significant factor in their inferior standing in public life. There were now some suggestions, albeit tentative, that husbands might be persuaded to ease their wives' burden by 'helping' them in the home.

From the 1970s, with Brezhnev now in charge, there was another shift in policy. The notion that women's sense of self-fulfillment rested on their combining the traditionally female roles in the family with the traditionally male roles in politics and social production began to be questioned, largely because of the seemingly intractable demographic implications: the birthrate, at least in the European republics of the Soviet Union, had remained stubbornly low. At the same time, women workers were essential in the Soviet Union's labor-intensive economy, and it was not possible simply to remove them from the workforce. Instead, there was an attempt to alter the balance between work and family in their lives, to persuade them to see the family as paramount and work as secondary. This meant that they would continue to work, but would devote less time and energy to trying to improve their qualifications and climb the professional ladder. Instead, they would see raising a family as their main source of satisfaction. Accordingly, a media campaign and a new school course were launched to raise the status of the family, persuade women to have more children, and encourage boys and girls to develop more traditional gender-specific traits. If having children had never been a choice for women, the attempted manipulation of their reproductive patterns now reached a new height.

Gender, though it was still not named as such, had become a major issue. Theorists now insisted that 'the study of psychological differences between members of the male and female sex is an important scientific and practical question, no less important than the continuing study of their biological and physical differences' (Khripkova and Kolesov 1981, p. 74). In the West there was also concern at this time with studying 'psychological differences between members of the male and female sex,' but in order to understand how they came about, the extent to which they challenged equality, and what could be done to overcome them. In the Soviet Union, the primary concern was to reinforce them.

When Gorbachev came to power in 1985, his program of economic reform introduced the prospect, for the first time since the 1920s, of widespread unemployment. The high level of female employment that had characterized the Soviet Union since the beginning of the Stalin era was no longer necessary. Women could now be withdrawn from the workforce, and this could result in two positive outcomes: an increase in the birthrate, and a reduction in the unemployment figure. This was presented to women themselves as a positive development: at last they had the 'luxury' of staying home with their children.

Yet the new openness (glasnost) of the Gorbachev era also meant that criticism of government policy was possible. Articles heatedly defending women's right to work began to appear in the press. This extended even to the Communist Party's own theoretical journal, *Kommunist*, in which three female scholars from the Institute of Socio-Economic Problems challenged what they saw as a clear trend to improve economic performance by shedding the less effective elements of the work force: 'and that means, in the first place, women.' This, they argued, lay behind the introduction of what seemed to be new benefits for women:

> In order to facilitate this female "exodus" from social production, it has been suggested that they be granted a series of privileges which will result in the reduction of their work load: the work day will be shortened, the period of paid leave to look after a new-born baby lengthened, and so on (Zakharova et al. 1989, p. 58).

Within two years of this article appearing its authors had founded the Centre for Gender Studies in Moscow and, as well as publishing material hitherto unimaginable in the Soviet Union, were mounting university courses and hosting conferences on gender issues in Soviet society.

3. Research on Gender Issues in the Soviet Union

Since the 1970s, Western research on Russia and the Soviet Union, in virtually all fields—literature, history and the social sciences—had paid considerable attention to the experience of women. Feminist scholars were initially attracted to the study of Russian and Soviet women, as the historian Linda Edmondson explains, because:

> it seemed (from a Western perspective) that a degree of equality and recognition had been won by women in the Soviet Union that had not been achieved in the West … they were impressed by Soviet claims not only to have legislated complete sexual equality, but also to have provided the conditions for its full realization; if not in the present, then at least in the foreseeable future (Edmondson 1992, p. 1).

They soon realized their mistake. However, it still seemed paradoxical that despite the Soviet Union's professed commitment to women's equality there was such apparent disinterest on the part of its scholars in discovering how it had supposedly been achieved, and that Western researchers working in the field 'were obliged to provide almost all the answers to their own questions' (Edmondson 1992, p. 2). Accordingly, the creation of the Centre for Gender Studies and the emergence of an indigenous feminist movement were greeted with considerable interest and enthusiasm.

Yet when Russian scholars did start providing their own answers, they made somewhat uncomfortable reading for their Western counterparts. The experience of living under state socialism, perhaps inevitably, led them to an understanding of equality and women's rights that did not accord with that of Western feminism. While feminist theory in the West had drawn attention to the distinction between the private and the public, Soviet women saw the main dichotomy as that between the family and the state. Western feminists had been largely critical of the patriarchal nature of the family, but for women in the Soviet Union it was seen as a refuge from the state-dominated bureaucracies that controlled the worlds of work and politics. While Western feminists tended to promote paid work as the essential route to women's financial independence and self-realization, women from the Soviet Union had experienced it as a form of exploitation which, in the words of Nanette Funk and Magda Mueller, reduced them 'to a cheap source of labor that could be plugged in anywhere without regard to specific capabilities, prerequisites, and ambitions' (Funk and Mueller 1993, p. 87). Western feminists had pointed to men's vested interest in keeping women in their place, but women in the Soviet Union were more inclined to see men as fellow sufferers under the yoke of state oppression. Indeed, some Soviet feminists now argued that men had had a harder time than women, since the latter were able to use their family responsibilities as an excuse to withdraw as far as possible from the public realm, where state control was at its greatest.

The supposed successes of the Soviet Union in securing women's equality—the high level of female involvement in paid work, the relatively strong representation of women at least in the lower echelons of politics, the extensive, if still inadequate, provision of pre-school childcare institutions—were also looked at with a critical eye by the new Soviet feminists. State socialism, it was argued, had been so obsessed with the idea of work that it had reduced women's equality to nothing more than their mass entry into social production. Being equal meant being treated like men, except that they also had the major responsibility for childrearing and domestic work. Many Soviet women began to define liberation in a surprising new way: as the right not to work, and to be treated as innately different to men. This essentialist position, which saw women as naturally oriented towards the home, was anathema to many Western feminists, but it was an understandable response to the conditions in which women had lived under what they now described as 'really existing socialism.' The differences which emerged in women's attitudes in East and West have not precluded a fruitful dialogue between them, but this has had to be conducted with considerable sensitivity and understanding.

Bibliography

Atkinson D, Dallin A, Lapidus G W (eds.) 1978 *Women in Russia*. Harvester, Hassocks, Sussex, UK

Attwood L 1990 *The New Soviet Man and Woman: Sex Role Socialization in the USSR*. Macmillan, London

Buckley M 1989 *Women and Ideology in the Soviet Union*. University of Michigan Press, Ann Arbor, MI

Edmondson L (ed.) 1992 *Women and Society in Russia and the Soviet Union*. Cambridge University Press, Cambridge, UK

Evans Clements B, Engel B, Worobec C D (eds.) 1991 *Russia's Women*. University of California Press, Berkeley, CA

Farnsworth B, Kollontai A 1980 *Socialism, Feminism, and the Bolshevik Revolution*. Stanford University Press, Stanford, CA

Funk N, Mueller M (eds.) 1993 *Gender Politics and Post-communism: Reflections from Eastern Europe and the Former Soviet Union*. Routledge, London

Gasiorowska X 1968 *Women in Soviet Fiction 1917–1964*. University of Wisconsin Press, Madison, WI

Goldman W Z 1993 *Women the State and Revolution: Soviet Family Policy and Social Life 1917–1936*. Cambridge University Press, Cambridge, UK

Ilic M 1999 *Women Workers in the Soviet Inter-war Economy*. Macmillan, London

Khripkova A G, Kolesov D V 1981 Devochka–podrostok–devuska [Girl–teenager–young woman] Prosveshchenic Moscow

Kollontai A 1977 *Selected Writing* (translated, and with commentary by Alix Holt). Allison & Busby, London

Kon I, Riordan J (eds.) 1993 *Sex and Russian Society*. Indiana University Press, Bloomington, IN

Lapidus G W 1979 *Women in Soviet Society: Equality, Development and Social Change*. University of California Press, Berkeley, CA

Liljestrom M, Muntysaari B, Rosenham A (eds.) 1993 *Gender Restructuring in Russian Studies*. Tampere University Press, Tampere, Finland
Stites R 1978 *The Women's Liberation Movement in Russia*. Princeton University Press, Princeton, NJ
Zakharova N, Posadskaya A, Rimashevskaya N 1989 Kak my reshaem zhenskii vopros [How we are resolving the Woman Question] *Kommunist* [Communist] No. 4, pp. 56–65

L. Attwood

Soviet Studies: Geography

The field of Soviet geography has undergone tremendous change. As a result of the significant social, economic and political changes of the post-1985 era, many practitioners of Soviet geography found the physical borders of their inquiry change. More often than not, the region of focus shifted from the Soviet Union to one or more of the former republics. In addition, research themes and theoretical approaches to scholarship expanded. Accordingly, the subject of Soviet geography has two distinct periods. Geographic research about the Soviet Union before *perestroyka* or the mid-1980s generally focused on amassing information and analyzing spatial patterns related to the natural environment, demographic patterns and economic activity (see Pryde et al. 1989 for an exhaustive review of research during this period). The post-Soviet period has continued previous traditions and added many new themes such as political change and the social and economic consequences of transition (see Engelmann et al. 2001) for a review of research themes and issues during this period).

1. *Soviet Geography*

During the Soviet period, the scope of geographical research often coincided with the strategic needs of policy makers. Because of Soviet secrecy regarding information about the natural environment and the location of natural resources, the location of economic activity and population dynamics, much of the work outside of the Soviet Union focused on amassing and interpreting information at a subnational scale in order to satisfy the demand for such knowledge among the policy-making community. As a result, geography was one of the few disciplines within Soviet studies to focus on regional and local issues throughout the Soviet period.

Despite limited interaction among Soviet colleagues and their counterparts elsewhere during the Cold War era, work by geographers in the Soviet Union became accessible to outsiders through various translation projects such as the journal *Soviet Geography and Translation*, founded by Theodore Shabad and renamed *Soviet Geography*, and a volume by Fuchs and Demko (1974)

Geographers in the Soviet Union and elsewhere published widely on the location of natural resources and the changing distribution of natural resource extraction, particularly on strategic resources like oil and natural gas. Several important volumes (e.g., Shabad 1969, Pryde 1972, Jensen et al. 1983) document the location of natural resources and discuss the wider implications such as economic development and conservation. Additional work by geographers brought to light important environmental issues related to renewable natural resources like the drying up of the Aral Sea and the resulting regional scale environmental degradation (e.g., Micklin 1988).

Numerous geographers traced the progress of various large-scale Soviet industrial projects related to natural resources and industrial development. Soviet officials perceived these projects as critical to national welfare, and other governments perceived understanding them as critical to their own national security. For example, geographers followed projects like the diversion of major Siberian rivers to provide water resources to Soviet Central Asia and the construction of a new railway line across Siberia, the Baykal–Amur main-line, to facilitate the transportation of natural resources. Many debated the strategic and economic consequences of constructing the new line. The journal *Soviet Geography* provided current information on these and many other projects related to natural resources and industrial development in its ten annual issues.

The location of economic activity and demographic dynamics comprised two other significant and interrelated research topics in Soviet geography. Geographers inside the Soviet Union and elsewhere analyzed urbanization patterns and the Soviet policy of city size limitations (e.g., Khorev 1971). In the United States, Harris (1970) initially promoted the study of Soviet urbanization using approaches and techniques that were standard at that time in Western urban geography, (e.g., central-place theory) and functionally categorized Soviet urban places. These works along with others (e.g., Lewis and Rowland 1979) evaluated aspects of Soviet central policy, like city size limits, and suggested that such policies did not appear to change overall urban growth rates.

Other works attempted to demystify the Soviet central planning system and show how regional and spatial planning occurred under this system and affected spatial development (e.g., Pallot and Shaw 1981). Shabad in his News Notes section of the journal *Soviet Geography* and elsewhere (Shabad 1984), provided data and analytical information about the location of economic activity and population. For example, he showed how the Soviet urban classification system according to population size and function could be used to reveal locations where central planners located economic activity or changed its nature. Liebowitz (1987) contributed to the ideologically generated debate over the effectiveness of

Soviet regional equalization policies. By using an oblast level scale of analysis, he concludes that the Soviet government had a strong commitment to reducing spatial inequality.

Investigative works such as those cited above distributed information to the scholarly and public policy communities about a mysterious place and system. They filled an important information gap about a significant part of the world. Salient questions in the field revolved around identifying what happens where and for what reasons. Other geographers often labeled the work as overly descriptive or devoid of theoretical or conceptual content. In reality, research during this period was implicitly theoretical and generally fell into two underlying and opposing frameworks: (a) Soviet exceptionalism due to the combination of communism, totalitarianism, and central planning; and (b) universalism due to the belief that human behavior is fundamentally the same irrespective of social, cultural, or economic context. Both perspectives mirrored broader trends in social sciences such as cold war dogma and modernization theory.

Even before the break up of the former Soviet Union, the field of Soviet geography began to change. New and expanded sources of information reduced the demand for amassing and publishing subnational data. Soviet geographers began to shift focus away from identifying spatial trends to interpreting and conceptualizing them. A new political imperative arose to document and interpret the spatial implications of 'democratization' and marketization. A modified set of supply and demand issues faced the scholarly community and produced a substantially increased set of research themes and theoretical perspectives.

2. *Post-Soviet Geography*

As a result of the break up, the geographical scope of research has generally narrowed to Russia as many Soviet specialists interested in other former Soviet states have aligned with regional specialists interested in Europe, the Middle East, and Asia. The research emphasis in the subdiscipline continues on regional and local issues in Russia with several changes. Among the most important changes is the increased collaboration with and access to colleagues in Russia. A review of articles in *Post-Soviet Geography and Economics* (formerly the journals *Soviet Geography* and *Post Soviet Geography*), shows much greater collaboration among international scholars and among non-regional specialists.

Also important are the increased amount of information available (usually through purchase) and the ability to conduct original field research (including ethnographic and survey studies). New ethical considerations have arisen. For example, we have begun to consider how non-Russian academics profit from interactions with Russian colleagues and informants. Many also consider the advantages to Russian colleagues who participate in joint research. In addition, the need to purchase information brings up issues related to the accessibility and dissemination of scholarly information as well as quality. The possibility of generating new forms of scholarly evidence leads to the ability to expand theoretical and methodological perspectives in Russian geography. Scholars within the field are not only in the business of amassing and interpreting information, but also of producing that information. While these changes have wrought considerable excitement among practitioners and new entrants into the field, the cohesiveness of the Soviet period seems to have disappeared.

Traditional research themes remain with a new emphasis on political issues. For example, while geographers still provide information on the location of natural resource extraction and the state of the natural environment, new themes related to the broader subdiscipline of nature–society relations have emerged. New work using ethnographic research methods includes a focus on the issue of the rights of indigenous people (e.g., Fondahl 1998). Research continues to document the nature of demographic change during this transition period. Several authors note important trends such as reversals of traditional demographic patterns, including population losses and deurbanization in the Russian East, which had traditionally been an urban migration destination (e.g., Heleniak 1997 and Rowland 1996). Other researchers place a new emphasis on the political dimensions of demographic change and ethnic identity politics (e.g., Kaiser 1994). These works continue the tradition of providing information about the region while explicitly showing a change in the intellectual content of the field by incorporating methods and topics that speak to the broader community of social scientists as well as to policy makers.

While documenting the location of economic processes still predominates, research during this period looks at the political aspects of economic relations and activity. Using theoretical approaches to study local governance in other countries, Mitchneck (1998) cites the example of the use of historical imagery for economic development to show that Russian local governments increased their participation in promoting local economic development. A focus on political relations also permeates the work of economic geography of Russia. Bradshaw and Lynn (1994) argue for the integration of new theoretical approaches into our study of economic relations in Russia. They use the example of world systems theory to show how we can enrich both the study of transition and social sciences at large by bringing analysis of Russia and the former Soviet Union into broader scholarly debates.

The direct examination of politics as a geographical process has also entered into the lexicon of Russian geography. Craumer and Clem produced a series of articles on electoral processes in Russia at various

scales of analysis (e.g., 1997) and O'Loughlin et al. (1996) also studied electoral processes. Also new to Russian geography is the analysis of the impact of broader social and economic change on the individual—especially differences across gender—and the introduction of social theory to the study of the transition period in Russia (see Bell 1997 and Pavlovskaya 1998).

In many ways, thematically, and theoretically, Russian geography appears more integrated into the overall field of geography than was Soviet geography. Research themes, methods, and theoretical approaches mirror the discipline at large more explicitly. The future of the discipline, however, is still quite uncertain. The productivity of scholars both in Russia and abroad seems somewhat sidetracked due to the need to restructure. In addition, the political imperatives that generated both scholarly interest and immense funding for research appear to have waned.

A promising area of future research for geographers studying Russia and other former Soviet countries would be contributing to transition theory and studying the impact of marketization on both the natural and built landscapes. Relatively few works to date have begun to deal directly with such issues (for exceptions see Bradshaw 1997, Ioffe and Nefedova 1998, Bradshaw and Hanson 1998, Smith 1999), in part, because of the limited time that has elapsed. But in this ever-evolving field, we have learned to expect the unexpected.

See also: Communism; Geography; Postsocialist Societies; Regional Geography; Socialist Societies: Anthropological Aspects; Soviet Studies: Economics; Soviet Studies: Environment; Soviet Studies: Politics; Soviet Studies: Society

Bibliography

Bell J E 1997 A place for community? Urban social movements and the struggle over the space of the public in Moscow. Dissertation, University of Washington.

Bradshaw M J (ed.) 1997 *Geography and Transition in the Post-Soviet Republics*. Wiley, Chichester, UK

Bradshaw M, Hanson P 1998 Understanding regional patterns of economic change in Russia: an introduction. *Communist Economics and Economic Transformation* **10**: 285–304

Bradshaw M J, Lynn N 1994 After the Soviet Union: the post-Soviet states in the world system. *The Professional Geographer* **46**: 439–49

Clem R S, Craumer P R 1997 Urban-rural voting differences in Russian elections, 1995–1996: a Rayon-level analysis. *Post-Soviet Geography and Economics* **38**: 379–96

Engelmann K, Bell J, Mitchneck B, Kaiser R 2001 Russia, Central Eurasia, East Europe. In: Gaile G L, Willmott C J (eds.) *Geography at the Dawn of the 21st Century*. Oxford University Press, Cambridge, MA

Fondahl G 1998 *Gaining Ground? Evenkis, Land and Reform in Southeastern Siberia*. Allyn and Bacon, Wilton, CT

Fuchs R, Demko G 1974 *Geographical Perspectives in the Soviet Union*. Ohio State University Press, Columbus, OH

Harris C 1970 *Cities of the Soviet Union: Studies in their Functions, Size, Density, and Growth*. Rand McNally, Chicago

Heleniak T 1997 Internal migration in Russia during the economic transition, (1989 to end-1996). *Post-Soviet Geography and Economics* **38**: 81–105

Ioffe G, Nefedova T 1998 Environs of Russian cities: A case study of Moscow. *Europe–Asia Studies* **50**: 43–64

Jensen R G, Shabad T, Wright A W (eds.) 1983 *Soviet Natural Resources in the World Economy*. University of Chicago Press, Chicago

Kaiser R J 1994 *The Geography of Nationalism in Russia and the USSR*. Princeton University Press, Princeton, NJ

Khorev B S 1971 *Problemy Gorodov*. Mysl', Moscow

Lewis R A, Rowland R H 1979 *Population Redistribution in the USSR, Impact on Society, 1897–1977*. Praeger, New York

Liebowitz R 1987 Soviet investment strategy: a further test of the 'equalization hypothesis'. *Annals of The Association of American Geographers* **77**: 396–407

Micklin P P 1988 Desiccation of the Aral Sea: A water management disaster in the Soviet Union. *Science* **241**: 1170–76

Mitchneck B 1998 The heritage industry Russian style: the case of Yaroslavl. *Urban Affairs Review* **34**: 28–51

O'Loughlin J, Shin M, Talbot P 1996 Political geographies and cleavages in the Russian parliamentary elections. *Post-Soviet Geography* and *Economics* **37**: 355–87

Pallot J, Shaw D J B 1981 *Planning in the Soviet Union*. Croom Helm, London

Pavlovskaya, M E 1998 Everyday life and social transition: gender, class, and change in the city of Moscow. Dissertation, Clark University.

Pryde P R 1972 *Conservation in the Soviet Union*. Cambridge University Press, London

Pryde P R, Berentsen W H, Stebelsky I 1989 Soviet and East Europe. In: Gaile G L, Willmott C J (eds.) *Geography in America*. Merrill, Columbus, OH

Rowland R H 1996 Russia's disappearing towns: new evidence of urban decline, 1979–1994. *Post-Soviet Geography and Economics* **37**: 63–89

Shabad T 1969 *Basic Industrial Resources of the USSR*. Columbia University Press, New York.

Shabad T 1984 The urban administrative system of the USSR and its relevance to geographic research. In: Demko G J, Fuchs R J (eds.) *Geographical Studies on the Soviet Union* (Department of Geography Research Paper No. 211). University of Chicago, Chicago, IL

Smith G 1999 *The Post-Soviet States: Mapping the Politics of Translation*. Arnold, London

B. Mitchneck

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Soviet Studies: Politics

The study of Soviet politics was a special area dealing with the analysis of political developments in the former Soviet Union: its political history and traditions, its political system and regime, the evo-

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

lution of its political leadership, its social structure, and the cultural roots of Communist totalitarianism.

With the fall of the Soviet Union and Communism in 1991, Soviet studies ceased to exist as an area of research. Today, a review of Soviet studies is primarily of historical interest. It is notable that, in its traditional form, Soviet studies was first and foremost exclusively the product of a Western community of researchers and politicians. Moreover, this brand of social science was largely an outcome of the Cold War between the West and the USSR, and served as the ideological and theoretical basis for that war. It was distinguished by a descriptive, narrative, and heavily ideological nature. The strong ideological undercurrent present in Soviet studies is understandable, given that its focus of research was the totalitarian Soviet regime, the very opposite of liberal democracy. Soviet studies was a closed area of research, removed from mainstream social science, and it effectively ignored modern research methods (i.e., comparative analysis). The field was a typical instance of area studies in that researchers felt little need to search for analogies and parallels beyond its confines. Indeed, Soviet studies (particularly those in the US) was dominated by émigrés who, in many cases, had left Communist countries as a result of persecution. This could not help but tinge their work with a certain amount of bitterness and strong emotion that were hardly conducive to objective analysis. Soviet studies, with their zeal and passionate condemnations, was not unlike Soviet social science and its attacks on the Western democratic system.

In any case, Soviet studies was unable to explain Gorbachev's *perestroika* or the collapse of the USSR that, incidentally, was not foreseen by any of the scholars in the field. This gives some indication of the standard of scientific analysis in the area. Still, a number of Sovietologists did provide the West with unbiased assessment of specific problems in Soviet development and of the processes that followed the collapse of Communism. Among these the most notable are Alex Dallin, Robert Legvold, Adam Ulam, George Breslauer, Arnold Horelick, Robert Tucker, Robert Daniels, Steven Cohen, Gail Lapidus, and Angela Stent.

After the fall of Communism, traditional Soviet studies petered out, not only in the US but also throughout the West. Effective research into the developments that shaped post-Communist Russia and other successor states of the USSR adopted new approaches developed by Western social inquiry. A complex analysis of post-Communism began to develop on the basis of transformation studies across all dimensions: politics, economics, culture, and the social sphere. The greatest progress in understanding post-Communism was achieved once researchers borrowed models from transition and comparative studies and started to explore both general and specific features of Russia's transformation from the perspective of other transition societies. At this stage, Juan Linz, Alfred Stepan, Terry Karl, Phillipp Schmitter, and Valery Bunce made significant contributions to the understanding of post-Communist Russia. Among the books that discuss the developments in the former USSR form a new perspective see Bunce (1999), Elster et al. (1998, Brown (1996).

Social science as practiced in the USSR prior to the fall of Communism was a discipline encumbered by ideological clichés. It was only in the *perestroika* years that research began to delve into the nature of the Soviet system, its evolution, and its relationship with historical tradition and national characteristics. With the breakup of the USSR and the fall of Communism, Russia's research community was faced with the need to reconsider its research methods completely and turn to international literature on social development. This has not been an easy process for Russian scholars and many have not been able to break free of Marxist philosophy and social science. Still, the 1990s have seen a gradual emergence of select social science disciplines that make use of internationally accepted research tools to study various social dimensions. Emerging areas of research include comparative political science, geopolitics, international relations, and transition studies. However, such research is largely concentrated in large university centers (such as Moscow, St. Petersburg, and Yekaterinburg) and around key magazines and journals (most importantly *Polis, Pro et Contra, Vlast*, and *Voprosy Philosophii*). Additionally, new communities of philosophers and political and social scientists have sprung up. In short, Russian social science has largely completed the initial phase of organizational and institutional development. It has done so with valuable support from the Western research community and financial assistance from certain Western funds and institutions, including the McArthur, Carnegie, Ford, Soros, and some European foundations.

Core works of Western social science have been translated into Russian and translations of contemporary research are readily available. Social science learning at Russian universities conforms with international standards to an ever-increasing extent. A new crop of researchers has done much to investigate various developments in Russian society. Among these are German Diligensky, Mikhail Ilyin, Andrey Melvil, Andranik Migranian, Igor Klyamkin, Vladimir Pastukhov, Oleg Bogomolov, Alexander Nekipelov, Nikolay Shmelev, and Yuri Levada.

Nevertheless, Russian social science is still in a transitional phase and its development remains difficult. There is a clear lack of stable avenues of interaction between various schools and perspectives, while research efforts are poorly differentiated and integrated. Inflows of new academic talent are limited. Sadly, some Russian observers consider Russian social science to be mired in crisis and apply this conclusion to all key social disciplines: political science, law,

history, social and cultural studies, etc. Russian research is relegated to the periphery of international research. Its investigation of major trends in social development is empirically weak and likewise flawed in a theoretical approach. Many research projects still rely on an oddly eclectic combination of outdated Marxist research methods and modern Western techniques, without due regard for the overall research context. This is a result of the parochial nature of Russian social sciences, their estrangement from international research, the belief in a unique Russian experience, and the reluctance of many scholars to study Russia in the context of other transformations and transitions from authoritarian rule. Knowledge of superior international works on the development of post-Communist Russia and other successor states of the USSR remains inadequate.

How can the weakness of Russian social science be explained? The first reason for it is the continuing crisis in Russian society and the indifference of authorities and the public to the development of science in general. Also part of the problem is the adverse tradition of excessively ideologizing research, the habit of some researchers to work for the authorities as well as the corresponding conformist attitudes that this engenders, and the lack of objective analyses, particularly when contemporary Russian developments are in question. Scientific prospects are blighted by an inadequate consolidation of the scholarly community, a lack of funding, the concentration of research effort in a handful of large cities, and the often-inadequate knowledge of foreign languages (particularly English) on the part of researchers, which prevents Russian science from breaking its parochial confines. The coming phase in the development of Russia's social science, i.e., its integration into the global framework, largely hinges on society's ability to overcome the current period of crisis and stagnation and acquire new perspectives and hopes. One of the major goals of the newly founded (1998) Russian Political Association which unites a major part of Russia's political scientists is the creation of a new framework for discussion and itellectual intercourse.

Thus, in 2001 Russian political studies have a complicated agenda. On the one hand, its goal is to integrate Russian studies into the general mainstream of world political science using its newest accomplishments, instruments, and approaches, on the other hand, its goal is to preserve the emphasis on area studies. The new step in the development of Russian political studies will demand more effort in analysis of the major problems of political history and traditions; of the political regimes; of the correlation between economic and political reform; of continuity and change; of stability and instability; and of the interaction between domestic and foreign policy. The success of this endeavor will depend on the ability of scientists to cooperate and share the results of their studies, and on the attempts to organize the collective effort of analysts from different fields in order to get a systemic and deeper understanding of the nature and puzzles of the post-Communist transformation.

See also: Communism; Communist Parties; Economic Transformation: From Central Planning to Market Economy; Socialist Societies: Anthropological Aspects; Soviet Studies: Economics; Soviet Studies: Society

Bibliography

Bonnel V, Breslauer G (eds.) 2000 *Russia in the New Century: Stability or Disorder?* Westview Press, Boulder, CO

Brown A 1996 *The Gorbachev Factor*. Oxford University Press, Oxford, UK

Bunce V 1999 *Subversive Institutions. The Design and Destruction of Socialism and the State*. Cambridge University Press, Cambridge, UK

Colton T, Legvold R (eds.) 1992 *After the Soviet Union*. W. W. Norton and Co, New York

Dallin A, Lapidus G (eds.) 1995 *The Soviet System: From Crisis to Collapse*. Westview Press, Boulder, CO

Elster J, Offe C, Preuss U K 1998 *Institutional Design in Post-communist Societies. Rebuilding the Ship at Sea*. Cambridge University Press, Cambridge, UK

Linz J, Stepan A 1996 *Problems of Democratic Transition: Southern Europe, South America, Post-communist Europe*. Johns Hopkins University Press, Baltimore, MD

L. Shevtsova

Soviet Studies: Society

'Soviet society' refers fundamentally to the society of the former Soviet Union (the Union of Soviet Socialist Republics). More broadly, scholars refer to a 'Soviet type of society,' i.e., a society structurally similar to that of the Soviet Union.

1. Main Perspectives

Social scientists have approached the study of Soviet society from several main perspectives. For some, Soviet society is of particular interest for what it can reveal about one of the world's superpowers. Those taking this view tend to observe and analyze all aspects of Soviet society through a political lens. They stress societal features potentially influencing or especially affected by its political system and policies. At another extreme, for those studying societies comparatively, Soviet society may be just another data point like American, British, Chinese, French, Indian, or Turkish society. Those adopting this perspective seek

to discern relationships among societal characteristics rather than the inner workings of Soviet society *per se*. For those with yet another perspective, Soviet society is a distinctive type of society; they think that studying its distinctive features can give a deeper understanding of how political, economic, and social structures are linked, and how history and culture impact the linkages among these structures.

These various perspectives have not received equal attention. The first was dominant during much of the twentieth century, especially in periods when the Soviet Union was regarded as a major military threat to the West. The second almost never leads to publications manifestly on Soviet society and has not predominated at any time; it is not covered in this article. The third, more sociological perspective has become prevalent since the Soviet Union collapsed.

2. *Soviet Society as a Distinctive Societal Type*

As a societal type, Soviet society was historically unique because its basic institutions were designed with the intent of implementing a particular theory of societal development and ideals, namely, that of Karl Marx. Although other societies were later constructed to realize Marx's vision, the Soviet Union's society was regarded as the prototype and the model to be followed.

According to Marx's theory of history, feudalism was succeeded by capitalism, and capitalism would be followed by communism. Communism would usher in a communist society, a superior stage of societal development. Although Marx wrote prolifically, his work mainly dealt with broad changes in economic and political systems and with his critical understanding of capitalism and capitalist society. Ironically, his writings furnish only a few general ideas about the nature of a communist society.

In the picture of a communist society painted by Marx, private capital and exploitation would be eliminated as the bases of society. There would be no social classes and no inherited privileges. People would be free to choose their activities; consequently, they would not be alienated from their work. Social relations would be based on cooperation rather than competition and exploitation; hence people would not be alienated from one another. Communist society would initiate a truly human and humane history. Communism would not only foster humane social relationships but also be economically more productive than other systems.

As evidenced in general by the actual societies designed to be communist and in particular by Soviet society, Marx's utopian vision of communism and of communist society proved to be wishful thinking. Its leaders pursued policies of rapid industrialization and urbanization, and it indeed achieved a relatively high level of economic output and became a largely urban society. Further, it succeeded in reducing (if not in eliminating) social inequalities and inheritance of social privileges. But productivity and economic growth were not as great as in capitalist countries, and the widespread use of force and repression to control people's lives gainsaid the claim that Soviet society was humane and had ended alienation.

Marx's contemporaries as well as succeeding generations of social thinkers challenged his theory and many of his claims, but few offered a different view of human history and human societies that was as broad and as encompassing as Marx's. In his monumental work, *A Study of History*, Arnold Toynbee (1947), was among the first to propose a vision of societal development that was not based on social classes. Like most scholars who have attempted to classify societal types, including Marx, he thought that at each stage of societal development, a certain mode of production was the driving force behind economic and social relations in a society. In Toynbee's scheme, pastoral societies were followed by agricultural societies, which were in turn succeeded by industrial societies. Later Daniel Bell (1973) argued that a new type of society came into being in the twentieth century: post-industrial society. Other scholars subsequently suggested particular kinds of postindustrial societies (e.g., information, service-based, postmodern).

Soviet society does not fit very well into the usual classifications of societal types, however, because it differed sharply from industrialized, urbanized societies that had market economies and democratic polities. Its structures for the organization of productive activities and for the allocation of economic outputs were so distinctive that the industrial basis of its economy provides comparatively little help in understanding Soviet society. The failure of the concept of 'industrial society' to identify fundamental differences between Soviet society and industrial societies in North America and West Europe has motivated scholars to study Soviet society as a distinctive type—to seek to identify its basic features that distinguish it from industrial societies with market economies and democratic polities.

3. *Origin and Key Elements of Soviet Society*

Communist ideology promised a better way of life to the many people living in industrializing societies who were relatively (and sometimes absolutely) deprived. In the nineteenth and early twentieth century, it was hard not to attribute much deprivation to the social upheavals that resulted when small numbers of entrepreneurs and other 'captains of industry' built large factories and other enterprises employing great numbers of men, women, and children who toiled very long hours at very low wages. Arguably even more deprived were the jobless and those without land to till. Not surprisingly, the communist vision was attractive to

many workers, peasants, intellectuals, and members of suppressed nationalities and ethnic groups. Some of them were mobilized by this vision and organized to change society.

According to Marx, social revolution is the only way that a working class exploited by the capitalist class can gain power in order to build a new communist society. In 1870 the Paris Commune tried to create the first communist society, but its efforts failed. Other attempted communist revolutions were sparked by the heavy losses of life and material destruction during World War I and by the hardships in core European countries in this war's aftermath.

Surprisingly, the first successful revolution of self-proclaimed communists occurred, not in a highly developed capitalist country, as Marx and Engels (1848/1962) had expected, but in Russia in 1917 where industrialization had then barely begun. A small group of intellectuals headed by Vladimir Lenin (née Ulyanov) was able to snatch power from Russia's ineffectual tsarist elite with comparative ease. They then managed to fight and win a three-year civil war (1918–1921) and to suppress dissenting factions in the communist movement. As a result, the Union of Soviet Socialist Republics (USSR) emerged as the world's first communist state.

During the early stages of the Soviet Union, its communist leaders—first Lenin and then his successor, Joseph Stalin—eliminated much of Russia's previous elite, not only its nobility but also its few industrialists. Numerous intellectuals with different ideas about societal goals were also removed. Well-to-do peasants, most of whom were suspicious of communism, were also got ten rid of. The Soviet Union's leaders justified the massive use of force to implement communist doctrine partly by the country's backwardedness: its largely peasant population was illiterate and had short time-horizons, and peasants were too disorganized, except on the village level, to create a new countrywide social system. In addition, the ideology itself excused the use of force by communist leaders. Social revolutionaries believe that their destiny is to force society to change in order to bring about a better society. The notorious ruthlessness of Russia's tsarist elite made this rationale more acceptable to many people.

Thus, it came about that Soviet society was built and maintained through the use of excessive force. Power and decision-making were concentrated at the very top. The real head of the Soviet Union was never elected, not even by the Central Committee of the Communist Party of the Soviet Union, which was officially the Party's leading body. The wishes and will of the general population were neglected.

In the Soviet Union's command economy, the state owned and controlled all productive resources. The Soviet system of extreme concentration of control over resources could be very effective. For example, it permitted this formerly backward country with its predominantly peasant population to industrialize with surprising speed, eliminate illiteracy and educate the general population, win World War II despite its initial unpreparedness, develop the first hydrogen bomb, launch the first satellite to orbit the earth, put the first cosmonaut in space, and become one of the world's two main superpowers during most of the twentieth century. It also achieved a considerable degree of social equality, though at a lower average level than in the United States and West Europe. While not overlooking the negative features of Soviet society, one should acknowledge its accomplishments.

Although the Soviet Union began as a dictatorship of the head of the Communist Party, the political system gradually evolved. By the 1960s, the apparatus of the Communist Party had been merged with the state apparatus, creating a party-state. During the final decades of the USSR, the *party-state* was the only actor in Soviet society that no one could legitimately contest; its authority was supreme.

The party-state's monopoly on public action greatly limited the circle of people able to influence how society developed. Individual initiative was condoned only in a person's private life. This depressed the populace's motivation to act in the interest of themselves personally or of the general public. It led people to shun responsibility in public life. Since individuals depended on the state for their basic sustenance, and since the economic system offered few incentives for hard work and achievements, ordinary people sensibly chose to focus on their own private, personal lives: their families, a few close friends, and their personal hobbies. Because material wellbeing was secure (though low), the average person living in the Soviet Union was able to concentrate on his or her personal interests and have a private life offering considerable personal satisfaction.

Soviet society did not, however, allow a civil society to come into being. It lacked a level of social organization above family and friends—a level of voluntary organizations and private clubs—that many scholars (e.g., Putnam 1993) regard as a major foundation of a democratic society. The party-state controlled not only all productive resources in the economy but also all social organizations linking individuals who were neither family members nor close personal friends (e.g., even choral groups and clubs of birdwatchers). Private individuals were not allowed to meet and form associations, even ones intended to foster private and personal goals.

The Soviet system successfully indoctrinated its citizens in Marxist ideology and established a public mentality oriented to the collectivity. In this ideology, a person's life had value only as a tool for the attainment of collective societal goals.

In the early decades of the Soviet Union, its leaders experimented even with ways to change family life. To emancipate women, marriage and divorce were made very easy. To dilute the influence of families on new generations, state organizations for tending young

children and for educating and training older children were instituted. Some of these experiments had unintended adverse consequences and were later abandoned. For example, it was discovered that in practice easy divorce meant a huge upsurge in unmarried mothers trying to raise children without the father's assistance. Because it was in the state's interest for fathers to support their children, the law was changed again to make divorce in the Soviet Union almost impossible. Nevertheless, despite various planned interventions of the party-state into family life, on the whole, family life and other personal aspects of individuals' lives in Soviet society seem to have been determined largely by the same kinds of factors as in noncommunist industrialized societies. Empirical research on this topic is, however, still scanty.

Two key institutions in modern capitalist societies, money and law, were not very important in Soviet society. The Soviet party-state did not need money to motivate people or to act as the medium of economic exchanges. Nor did it rely on laws to control people's behavior or to regulate economic activities. These ends were accomplished primarily through administrative orders and, when necessary, the threat of punishment. Indeed, both money and law were seen as potentially limiting the absolute power of the party-state and therefore as best avoided.

4. *One Society or Many?*

The Soviet Union was comprised of many different nations (ethnic groups) having different cultures, traditions, and histories. Within its borders, hundreds of different languages were spoken. While all national groups were predominantly rural before being incorporated into the Soviet Union, not all of them fit the prototypical image of Russian peasants living in villages. In Central Asia, there were nomadic herders; in Siberia, there were small tribal groups who survived by hunting. In some of the Baltic countries and in certain parts of Russia, there were independent farmers, some of whom produced food for world markets.

Forging its diverse national groups into a Soviet people and creating a 'homo Sovieticus' (Soviet man/woman) was an important objective of Soviet leaders. They used a wide variety of means, ranging from overt repression, to economic incentives, to formal indoctrination in schools and in compulsory lectures given at work sites, to more subtle efforts in songs, films, plays, and various literary works. After 70 years, these efforts could claim a moderate degree of success.

Even during the harshest years of repression under Stalin, however, people's ways of living were noticeably different in the various regions of the Soviet Union. Although the party-state institutional organization penetrated every part of the Soviet Union, every institution operated somewhat differently in the various major regions of the country. Not surprisingly, similarities were greatest in state institutions, where there were definite efforts to impose uniformity. For example, schools throughout the Soviet Union taught pupils of a given age the same subjects using essentially the same books. Remnants of traditional cultural patterns were most evident in family life and in patterns of interpersonal relations and friendships.

The last region conquered by the Russian Empire in the tsarist period was Central Asia. This enormous area has never been densely populated. Its native inhabitants are descended from many different ethnic groups that followed Islam and were tribally organized. However, it was not the Islamic religion as much as the traditional clan-based communities that made this part of Soviet society distinctive.

Uprisings in this region began during World War I. The Soviet state succeeded in restoring calm only after a series of mass slaughters in the 1930s. A colonial strategy of 'divide and conquer' was used to control it. Uzbeks, Kazakhs, Tajiks, Turkmens, and Kirgizes were given their own Soviet republics to establish the dominance of new nations in Central Asia.

Toward the end of the Soviet era, the Central Asian part of the Soviet Union was still predominantly rural and inhabited by nonSlavic peoples. But it also had capital cities and a few other urban centers with large industrial complexes inhabited and staffed by immigrant Slavic laborers. To a large extent, these urban areas were separate social oases whose residents had a way of life very different from that of those living in the other parts of these republics. The collapse of the Soviet Union left a large Russian diaspora in Central Asian cities. The titular nations who assumed the rule of their own countries have largely driven out the Russian diaspora. There remains a tiny, Soviet-educated native elite that has a say about who rules in this region, but the influence of this tiny native elite seems unlikely to endure. Leading groups, once again based on clans, are already being formed, and Central Asian societies formerly part of the Soviet Union are becoming more similar to other Central Asian societies, such as Afghanistan.

A second very distinctive part of the Soviet Empire was the Caucasus region in the south. The people of this region had distinctive religions and cultures. In the North Caucasus, some nations, such as the Avaris and Chechens, followed Islam; others were part of Orthodox Christianity; still others had their own distinctive and ancient national religions. The native peoples of this region were also organized in clans, but many had developed a more elaborate culture than was found in Central Asia.

To a considerable extent, the various peoples of the Caucasus region developed a national identity in the process of confrontations with Russians who sought to control their homelands. During the Soviet era, members of the national minorities in the Caucasus

were deeply involved in the Soviet power structure. Native elites sought to accommodate Moscow in some matters while at the same time acquiring more autonomy in handling affairs in their own traditional homelands. The top elite in the Caucasus in the Soviet era was politically very competent and has successfully protected its own turf. Evidence for their success is found in the fact that First Secretaries of the Communist Party in the former republics in this region have become the heads of some of the newly independent countries in this region (e.g., E. Shevardnadze in Georgia, G. Alijev in Azerbaijan).

A third tiny but clearly distinctive group lived in the western part of the Soviet Union, the Baltic peoples. The Baltic countries enjoyed independent statehood between World Wars I and II. In 1940 they were occupied by Soviet troops but were brought fully under Soviet control only after World War II ended. Soviet leaders used migration of nontitular peoples into the Baltic region as a method of sovietization. This policy was most successful in Latvia, where the 1989 Soviet Census revealed roughly equal numbers of Latvians and nonLatvians (most of whom were Russians or other Slavs). It was least successful in Lithuania. According to this same census, Lithuanians were still four-fifths of the population in 1989, and they stubbornly remembered the distant past when a Lithuanian Empire extended to the Black Sea.

Since the process of sovietization of the Baltic peoples and their societies was shorter than elsewhere in the Soviet Union, it never succeeded in erasing an antiSoviet mentality. Opposition to the Soviet regime in the Baltic region was silent but overwhelming. At the first opportunity presented by *perestroika*, mass movements organized in popular fronts along with native political leaders began to push for greater autonomy in the Baltic republics. Their aim was eventual restoration of independent Baltic countries, a goal achieved in 1991, the year when the USSR collapsed.

At the core of Soviet society were the major Slavic nations: Russians, Ukrainians, and Belarusians. In tsarist times, their languages were distinctive but mutually intelligible, and most of them were followers of Russian Orthodoxy. They also shared a largely similar heritage based on the life of peasants living in small villages loosely linked by towns functioning as trading posts. Stalin's program of collectivization destroyed village life and pushed many former peasants into new Soviet factories in cities. The destruction of the old way of life was as important as other direct methods in creating a new Soviet person.

The sovietization of Slavs reached its culmination during World War II, the Great Patriotic War. This war brought the greatest hardships in modern Russian history. Victory over Hitler's military forces was achieved only through millions of people sacrificing their lives and by millions of others toiling to support the Soviet war effort. This victory in the face of tremendous odds and losses bolstered the legitimacy of communist ideology and its collectivistic orientation. It made Soviet citizens proud and united them, especially those of Slavic origin. In the postwar period until the 1960s, the Soviet Union experienced rapid economic development and achieved important 'firsts' in the race to space. Soviet Slavic citizens' pride grew as their country acquired the status of a world superpower and leader in many arenas (e.g., science, sports).

Once the peoples of these regions were largely peasant Slavs, many of whom were descendants of former serfs. In the Soviet era, a majority of Slavs became city dwellers who lived in state-owned apartment complexes while working on factory assembly lines or doing various white collar jobs. They experienced great intergenerational mobility: geographical (rural to urban), educational (illiterate to a secondary education or higher), and social (peasant to manual or white collar worker). This great mobility reshaped people's personalities and mentalities and obliterated many memories of and allegiance to traditional ways of life.

When a period of high growth ends (which inevitably it does), educated people usually begin to wonder about their collective as well as personal future. After the Soviet Union crushed the Prague Spring in Czechoslovakia in the late 1960s, there came dreary years of ritualistic procedures of the Kremlin gerontocracy and little economic progress. Indeed, there were signs of economic retrenchment. Outside economic observers would label this as a shift from extensive to intensive economic development, but Soviet citizens of Slavic origin realized only that their lives were not getting better. The educated core of the Slavic people living in large cities (and not only the few public dissidents like Andrei Sakharov and Aleksandr Solzhenitsyn) began to speculate about the Soviet Union's future. Faith in the Soviet leadership, and even in a communist future, began to wither. A growing and widespread but diffuse view of Soviet society as collectively 'treading water' set the stage for the political and economic system to collapse from within.

5. *Studies of Soviet Society*

5.1 *Studies During the Soviet Era*

During most of the Soviet Union's existence, it was hard for observers to acquire objective information about Soviet society. In this situation, the study of Soviet society became very specialized and was not closely linked to other areas of social science research. *Sovietology* was the special field of social studies that addressed all aspects of Soviet society, though it focused especially on its political system and economy.

This special field arose as part of the political and ideological confrontation between communist and capitalist countries during the Cold War era. Specialists sought to reconstruct the reality of social life in Soviet society. Scholars were successful primarily in analyzing official Soviet documents and writings that gave evidence about the declared and actual policy goals of the Soviet Union.

Sovietological types of studies helped to manage the ideological and political confrontation between communist and Western democratic systems, but they were less valuable in providing a social scientific understanding of how Soviet society actually functioned. The basic bottleneck for Western social scientists was the dearth of high quality data on a wide array of topics.

During the Cold War era, social scientists conducted only a few sizable empirical projects that were intended to gather in-depth information about Soviet society *per se*. One of the best known was the so-called Soviet Émigré Project conducted by scholars at Harvard University in the late 1970s (cf. Millar 1987). In this project, émigrés from the Soviet Union were interviewed about their beliefs and their life experiences in the Soviet Union. These interviews were regarded as the best available substitute for direct surveys of random samples of individuals within the Soviet Union. This project indeed produced a richer and more complex picture of Soviet society than had been put together from scrutiny of the Soviet mass media and official public documents.

There were also careful and in-depth studies of the Soviet Censuses of Population and other official Soviet statistics (e.g., Clem 1986). A major conclusion was that the factual basis of official sources of Soviet data was highly suspect. At that time, it was widely believed that most errors in official data were intentional, designed to hide the true state of affairs in the Soviet Union. This was sometimes the case; however, mistakes often resulted from the same low level of quality control that was endemic in all nonmilitary production in the Soviet Union.

In a society designed to implement a particular view of history and societal development, there was no thirst for independent evidence that might reveal a picture different from the one offered by the party-state's leadership, or even worse, that might contradict the claims of the official view and expose actual flaws in the supposedly socialist society. Nevertheless, scholars within the Soviet Union made some useful studies of Soviet society.

In the 1960s, Yuri Levada initiated some theoretical analyses of the Soviet type of society. His analyses contained criticisms of how it worked, which led him to be barred from sociological work for the next two decades. In the 1970s, Boris Grushin began the Taganrog project that investigated the limits of the influence of mass media on society. In this same period, Aleksandr Yadov worked to develop middle-range theories justifying empirical sociological research outside an ideological framework, and Igor Kon applied Western sociological thought to various topics in Soviet society. Tatiana Zaslavskaya founded a program of sociological research that provided high-quality data on work behavior in Soviet industries and agriculture. Mikk Titma initiated longitudinal surveys designed to compare life careers of youths and young adults in various regions of the Soviet Union. The works of these and other Soviet scholars were published almost solely in Russian, and very few are available in English translation.

5.2 *Studies in the Post-Soviet Era*

With the demise of the Soviet Union, Soviet society is no longer of great political interest. As a major social experiment with important influences on the history of the twentieth century, Soviet society is now primarily of historical and social scientific interest.

Huge amounts of information about Soviet society and the Soviet system were released in the last years of the Soviet Union and during the first years of the new Russia Federation. Formerly secret archives, documents, and statistics became available to social researchers to study and analyze using the research methodologies and theoretical questions developed in the West. It also became much easier for social scientists to collect new data through interviews, social surveys, and various other kinds of field work. The first few years of this new openness allowed the quality of research on Soviet society and post-Soviet societies to reach the normal standards of the top journals in sociology, demography, economics, and political science in the United States and West Europe.

5.2.1 New insights into soviet society. Sovietological studies in the Cold War had usually depicted the Soviet system and society as not only highly centralized but monolithic and exceedingly stable. An important conclusion supported by new archival information is the remarkable evolution of the Soviet system during the seven decades of its existence. It developed from a crude personal dictatorship into a totalitarian society and then into a party-state with a complex system of institutions staffed by relatively educated officials.

The Soviet Union began as a relatively uncomplicated state with a top-down command system operated by the Communist Party apparatus, with terror used to guarantee that commands were obeyed. With a growing economy and a huge external empire in East Europe after World War II, the Soviet Union needed to expand the state and to develop a state apparatus having a legal basis to operate. A large new stratum, an intelligentsia with higher education, was

formed to serve in the state apparatus. Secondary education spread quickly and had become compulsory by the 1980s. University education expanded during the same postwar period. By the 1960s, the number of university graduates was already close to that in the United States. In the 1980s, white-collar workers had become a third of the labor force, an especially astonishing transformation when one considers that the Soviet Union was essentially a country of peasants in 1917.

It has become apparent that the Soviet Union's educated workforce transformed the state and Communist Party apparatus. The Constitution was changed to make the Communist Party the ruling party officially and as a matter of law. The Soviet Union developed a state bureaucracy that functioned relatively well. State institutions began to operate like typical bureaucracies pursuing their own organizational interests. Major Soviet institutions contested for power, largely in bureaucratic ways. Power became more widely dispersed among different state institutions and parts of the state apparatus. Even the KGB and other parts of the security apparatus were brought under the control of party-state in the postwar era. In the final years of the Soviet Union, most members of the Communist Party's Politburo were leading representatives of key economic ministries or of important republics within the Soviet Union. The Politburo, the Soviet Union's top executive organ, came to be comprised of the country's top executives; it was no longer just a group of Communist Party leaders.

Under Brezhnev's mediocre rule, internal dissatisfaction with a command system that limited initiative and restricted independent action even by loyal Soviet citizens soared. Cronyism and loyalty to a person's superior started to be the operative principle guiding behavior within the party-state apparatus. As repressiveness diminished and the threat of terror receded, criticism of the Soviet system became increasingly widespread, although initially not public. Gorbachev's policies of *perestroika* and *glasnost* brought these general feelings into the open. Though the top leaders of the Soviet state and Communist Party realized that change was needed, they had no clear idea what changes might have the desired effects of increasing efficiency while maintaining the basic power structure and its ideological foundation. Attempts to reform the command system failed. The finale came when the top ranks of party-state hesitated to use force to reestablish their complete control. They realized that they had no more positive alternative than the past and that they had lost the support of even the state bureaucracy, not to speak of the intelligentsia, which was actively opposing their leadership.

The collapse of the Soviet Empire and the party-state is a subject of special interest. There continue to be multiple explanations. The heavy burden of competition with the United States in the arms race certainly surpassed the capacity of the Soviet economy, but it cannot explain the regime's internal collapse. In 1991, the Soviet state could still marshal sufficient resources to have survived for some years. Challenges from the republics with strong nations were clearly a factor in the disintegration but could have been met and managed by applying sufficient force, as had been done in the past. Openness to the rest of the world during *perestroika* gave the Soviet elite a clear understanding that the Soviet Union offered a lower standard of living than the Western world. This realization eroded many people's faith in the communist ideology and system. A crucial factor was the paralysis of the party-state apparatus as a result of the loss of this faith in the long-term viability of the communist system, which thereby limited the capacity of the party-state's elite to govern.

5.2.2 Post-Soviet society and the transition process. In many instances, Soviet society and market societies are now seen as exemplifying very different ways in which an industrialized society can be organized. Rapid changes in postsocialist countries, but along different paths, permit rich comparisons between various postsocialist countries as well as between them and the well-developed societies in capitalist, democratic countries. It is often presupposed that all postsocialist societies are moving on a path toward a Western-type of society, that is, a democratic society with a well-developed market economy. Although some post-Soviet countries have unambiguously moved in this direction, it is too soon to be sure about some others.

The transformation of post-Soviet states from the previous system has raised a number of basic problems: the development of the state; strategies for developing democratic and market institutions; the basic principles of citizenship (ethnic or civic); the transformation of enterprises from the state sector to the private sector; factors promoting entrepreneurial activities; the emergence of greater and different kinds of social inequalities; the legacy of privileged positions in the Soviet era; the demographic crisis (a plummeting birth rate coupled with a rising mortality rate).

The Soviet Union's collapse has had an especially strong impact on studies of ethnicity and nationality. The emergence of fifteen post-Soviet states plus nearly a dozen other new nation states in the Soviet Union's 'outer empire' (i.e., the Soviet bloc in East Europe), most with sizable minority populations and many with unexpectedly high levels of ethnic hostilities, has stimulated social scientific studies of interethnic relations, nation building, and the role of nation states in the development of civil societies. Interethnic relations may be an issue in a multiethnic state, or between a majority and minority ethnic group within a single state, or between neighboring states each having an

ethnic diaspora in the other country. National self-determination is a different issue that develops when a cultural identity turns into a political identity. In a democratic regime, the outcome can be a peaceful divorce as occurred in the emergence of separate Czech and Slovak states. There is a clear difference between ethnic rights for cultural, social, economic, and political self-determination in the historic homeland of an ethnic group on the one hand and civic principles of interethnic relations in a modern state on the other hand. A civic society is only established with great difficulty when there are historical animosities between nationalities. Communist regimes suppressed ethnic divisions along with other features of a civil society. In contrast, a modern market-based society is premised upon civic principles. These civic principles allow ethnic tensions and hostilities to grow and offer no clear means to check them.

There has also been particular interest in the gains and losses from economic transformation for both individuals, groups, and society as a whole. In terms of the last, the main issue has been the advantages and disadvantages of economic shock therapy rather than a gradual transformation to a market economy. Available data on various postsocialist countries suggests that shock therapy does not always lead to quick improvements in economic and social life and that a gradual transition, as in China, tends to be more effective. Evidence also suggests that a macroeconomic, top-down approach to development excludes most people from participation in building a market economy. In contrast, a bottom-up approach motivates many people to seek and seize opportunities that markets provide. It allows a majority to participate, although naturally not everyone succeeds in the new economic competition. Indeed, a bottom-up process is how market societies emerged historically.

Both theory and comparative research have also debated the role of elites in the transformation of socialist societies. In post-Soviet and East European countries, the issue was initially politicized and often presented as the old (communist = bad) elite versus a new (noncommunist = good) elite. Sociological research has, however, focused mainly on whether the old elite retained their advantages in postsocialist societies, whether position in the old elite (e.g., political versus managerial elite) mattered, and whether members of the old elite tended to gain or lose relative to members of the non-elite who were otherwise comparable (e.g., in educational achievement). A related issue has been the value of social connections as contrasted with personal know-how. The old elite tended to be advantaged in both, and even in the level of formal education. To date, there have been no definitive conclusions about issues pertaining to the role of the former elites. Clearly, however, many members of the former elite have been able to retain an elite status although keeping an elite status has often involved moving into a different kind of social role (e.g., a former party functionary might become an entrepreneur or manager of an enterprise).

5.2.3 Major research projects. Information about Soviet society, and especially the societal transitions in post-Soviet states, has provided Western sociologists, demographers, political scientists, and economists with opportunities to gather new data and to test a wide array of hypotheses. In the post-Soviet era, a basic problem is the quality of data rather than its simple availability. Direct control over data collection is necessary to monitor and evaluate data quality, and Western researchers often lack direct control.

Major empirical research projects focusing on post-Soviet and East European countries include Hout and Wright's study of social stratification (see Hout et al. 1992, Gerber 1999); I. Szelenyi and Treiman's study of postcommunist elites, begun in 1994 (see Eyal et al. 1998); Kohn's surveys of class consciousness (see Kohn et al. 1997); Burawoy's case studies of the political economy of the transition from socialism (e.g., see Burawoy and Krotov 1992); and the longitudinal study of health and related issues by social scientists at the University of North Carolina (e.g., Zohoori et al. 1998, Lokshin and Popkin 1999). In the former Soviet Union *per se*, the longitudinal study of the life careers of secondary school graduates begun by Titma in 1983 has had two follow-ups in the post-Soviet period, one in 1993–94 and the other in 1997–98 (see Titma and Tuma 1995, Titma 1997). O'Brien and Patsiorkovski (1999) have collected data over time on households in Russian villages.

Many post-Soviet countries, especially Russia and the larger East European countries once in its sway, have been included in large cross-national studies. These include Inglehart's study of world values (see Inglehart and Baker 2000, World Value Survey Study Group 1994); the Eurobarometer surveys (see Reif et al. 1996); and numerous studies of elections (e.g., White et al. 1997) and public opinion (e.g., Clarke 1993). These various projects have enormously enlarged the amount of data relevant to many fields of sociology that come from postsocialist countries in general and post-Soviet countries more specifically.

A basic consequence of the greatly expanded base of data on Soviet and post-Soviet societies is that sociological research on the Soviet Union and post-socialist states has entered the mainstream of Western social science. It is no longer an area of specialization with strong connections to the foreign policies of Western States.

See also: Bureaucracy and Bureaucratization; Collectivism: Cultural Concerns; Communism; Communism, History of; Dictatorship; Elites: Sociological Aspects; Marx, Karl (1818–89); Marxism/Leninism;

Marxist Social Thought, History of; Peasants and Rural Societies in History (Agricultural History); Revolutions of 1989–90 in Eastern Central Europe; Soviet Studies: Politics; Totalitarianism: Impact on Social Thought; Utopias: Social

Bibliography

Bell D 1973 *The Coming of Post-industrial Society*. Basic Books, New York

Bobak M, Pikhart H, Hertzman C, Rose R, Marmot M 1998 Socioeconomic factors, perceived control and self-reported health in Russia. A cross-sectional survey. *Social Science and Medicine* **47**(2): 269

Burawoy M, Krotov P 1992 The Soviet transition from socialism to capitalism: worker control and economic bargaining in the wood industry. *American Sociological Review* **57**(1): 16–38

Clarke S 1993 Popular attitudes to the transition to a market economy in the Soviet Union on the eve of reform. *The Socological Review* **41** (November): 619–52

Clem R S (ed.) 1986 *Research Guide to the Russian and Soviet Censuses*. Cornell University Press, Ithaca, NY

Eyal G, Szelényi I, Townsley E 1998 *Making Capitalism Without Capitalists: Class Formation and Elite Struggles in Post-communist Central Europe*. Verso, London

Gerber T P 1999 Survey of employment, income, and attitudes in Russia (Seiar), January–March 1998 [Computer file]. ICPSR version. All-Russian Center for Public Opinion and Market Research (VTsIOM), Moscow [producer], 1998. Inter-university Consortium for Political and Social Research, Ann Arbor, MI [distributor]

Hout M, Wright E O, Sánchez-Jankowski M 1992 *Class Structure and Class Consciousness in Russia (MRDF)*. Survey Research Center, Berkeley, CA

Inglehart R, Baker W E 2000 Modernization, cultural change, and the persistence of traditional values. *American Sociological Review* **65** (February): 19–51

Kohn M L, Slomczynski K M, Janicka K, Khmelko V, Mach B W, Paniotto V, Zaborowski W, Gutierrez R, Heyman C 1997 Social structure and personality under conditions of radical social change: A comparative analysis of Poland and Ukraine. *American Sociological Review* **62** (August): 614–38

Lokshin M, Popkin B M 1999 The emerging underclass in the Russian Federation: Income dynamics 1992–96. *Economic Development and Cultural Change* **47**: 803–29

Marx K, Engels F 1962 (1848) *Manifesto of the Communist Party*. Foreign Languages Publishing House, Moscow, Russia

Millar J R (ed.) 1987 *Politics, Work, and Daily Life in the USSR: A Survey of Former Soviet Citizens*. Cambridge University Press, Cambridge, UK

O'Brien D J, Patsiorkovski V V 1999 Russian village household panel surveys, 1995–1997 (Computer file). D J O'Brien, University of Missouri, Department of Rural Sociology, Columbia, MO/V V Patsiorkovski, Russian Academy of Sciences, Institute for the Socio-Economic Studies of Population, Moscow (producers). Inter-university Consortium for Political and Social Research, Ann Arbor, MI (distributor)

Putnam R D 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ

Reif K, Cunningham G, Kuzma M 1998 Central and Eastern Eurobarometer 6: Economic and Political Trends, October–November 1995 (Computer file). ICPSR version. GfK EUROPE Ad hoc Research, Brussels, Belgium (producer), 1996 Zentralarchiv für Empirische Sozialforschung, Köln, Germany/Inter-university Consortium for Political and Social Research, Ann Arbor, MI (distributors.) 1998

Titma M (ed.) 1997 *Sotsial'noe rassloenie vozrastnoi kogorty [vypustkniki 80–x v postsovetskom prostranstve. Proiekt 'Puti pokoleniia'] (Social Stratification in an Age Cohort Graduates of the 1980s in Post-Soviet Countries. Project 'Paths of a Generation')*. Institute of Sociology of the Russian Academy of Sciences, Moscow, and the Center for Social Research in Eastern Europe, Tallinn, Estonia

Titma M, Tuma N B 1995 *Paths of a Generation: A Comparative Longitudinal Study of Young Adults in the Former Soviet Union*. Technical Report, Department of Sociology, Stanford University, Stanford, CA, and Center for Social Research in East Europe, Tallinn, Estonia

Toynbee A J 1947 *A Study of History*. Oxford University Press, New York

White S, Rose R, McAllister I 1997 *How Russia Votes*. Chatham House Publishers, Chatham, NJ

World Value Survey Study Group 1994 World Value Survey, 1981–1984 and 1990–1993 (Computer file). ICPSR version. Institute for Social Research, Ann Arbor, MI (producer), Inter-University Consortium for Political and Social Research, Ann Arbor, MI (distributor)

Zohoori N, Kozyreva P, Kosolapov M, Swafford M, Popkin B M, Glinskaya E, Lokshin M, Mancini D 1998 Monitoring the economic transition in the Russian federation and its implications for the demographic crisis: The Russian longitudinal monitoring survey. *World Development* **26**(11): 1977–93

M. Titma and N. B. Tuma

Space and Social Theory in Geography

Human geography's engagement with social theory has been enormously influential, leaving no part of the discipline untouched. 'Social theory' encompasses a heterogeneity of postpositivist philosophies, including, but not limited to, Marxist political economy, humanistic thought, structuration theory, feminism, postmodernism, and postcolonialism.

1. *Marxist Political Economy*

Starting with radical geography in the 1960s, Marxist geography stressed the analytical centrality of labor and the production process in the creation and transformation of economic landscapes (see *Marxist Geography*). Through labor, human beings enter into social relations, materialize ideas, and transform nature. Marxists emphasized the powerful role that class plays in the social and spatial division of labor, as

a vehicle through which social resources are distributed, as a central institution in shaping labor and housing markets, as a defining characteristic of everyday life, and as a fundamental dimension of political struggle. Underpinning this view is the labor theory of value and its implication that class relations rest upon the extraction of surplus value from the working classes, a process pregnant with politics. The Marxist perspective moved beyond simplistic dichotomies (e.g., base–superstructure) to incorporate the multiplicity of class relations over space and time, gender, the state, and ideology and culture. Marxism is thus not just an economic analysis, but a social, political, and moral one as well.

Marxist geography allowed landscapes to be understood as the creation of modes of production, most notably that of capitalism. Harvey (1982, 1989b) initiated a fascination with uneven development, in which regions are created, transformed, and annihilated by flows of capital over space in search of maximum rates of profit. Capital creates a geography of the built environment, a 'spatial fix' that must be recreated periodically over time. The uneven development of cores and peripheries was analyzed at several spatial scales, i.e., within the metropolitan region, within nation-states, and internationally. Competition forces innovation and induces an acceleration in the turnover rate of capital that rhythmically reconfigures the landscape of commodity production (Harvey 1982). Periodic crises emanate from the overaccumulation of surplus value, leading to broad restructurings of production and location. Viewed systemically, crises are necessary for capitalism's existence; commodity production is not only crisis-ridden, but crisis-dependent.

Marxist geography diversified in several ways. Marxist analyses of the state, or government, arose primarily from French structuralism. In contrast to neoclassical views of the 'public sector,' social theory highlighted social origins and the multiple ways in which the state shapes the territorial reproduction of social relations. Long dominated by a focus on capital, Marxists have examined the geography of labor, including, for example, the spatial dimensions of class struggle and unions.

In the 1980s, economic geography established linkages to French regulation theory, which theorized the multiple forms of capitalism over time and space (see *Regulation Theory in Geography*). The latest of many restructurings historically, late twentieth century capitalism, underwent a profound transformation from Fordism—characterized by large scale production, economies of scale, vertically integrated firms, oligopolies, homogeneous products, mass consumption, assembly lines, interchangeable parts, and Taylorism—to a post-Fordist regime, characterized by small-scale production, networks of cooperative firms, niche markets, and flexible (computerized) production methods clustered in large urban agglomerations. When local production complexes were linked to the global economy, Marxism offered a nuanced understanding of how regions are constructed around industries (see *Technology Districts*).

Geography's engagement with political economy led to an important retheorization of space. Improved transportation and communication systems create rounds of time–space compression, or in Marx's apocryphal words, the 'annihilation of space by time' (see *Time–Space in Geography*). Massey (1984) reincorporated the local uniqueness of places within wider divisions of labor, holding that the reproduction of uneven development occurs through the layering of different rounds of accumulation upon the physical and social landscapes of individual locales. The localities project recovered local places as fundamental to the understanding of how broad processes play out in specific contexts. The enormous time–space compression of late twentieth century capitalism, particularly through telecommunications, has embedded places within a worldwide 'space of flows' (Castells 1996). In the context of hypermobile, post-Fordist, postmodern capitalism, such processes are central to an increasingly service-dominated, globalized economy dominated by streams of money and information (see *Finance, Geography of*).

2. *World-systems Theory*

Political economy's engagement with geography extended to the origins and structure of the world-system, including colonialism and imperialism, the geopolitics of nation-states, multinational corporations, and the possibilities of socio-economic development. Widespread dissatisfaction with modernization theory led dependency theorists to reject neoclassical views of the international division of labor (see *Development Theory in Geography*). Third World poverty—indeed, the construction of the developing world itself—did not, in this view, simply 'happen,' but was the product of centuries of colonialism and continued neocolonial exploitation through transnational corporations, i.e., the development of underdevelopment. Dependency theory represented the first application of Marxist geographical thought on a global scale, locating the cause of poverty in the materialist, external workings of the world economy, not the ideological, internal order of societies, as did modernization theory. Dependency theory rejected the promise that imitation of the West would bring prosperity and liberty to developing nations. Rather, it argued, development and underdevelopment constitute two sides of global capital accumulation, i.e., the wealth of the developed world was, and is, derived from the extraction of surplus value from underdeveloped places. Every moment in the penetration of commodity relations into non-capitalist social formations, therefore, was viewed as inevitably an act of exploitation.

Wallerstein (1979) advanced understanding of the political economy of global geographies with world-systems theory. The capitalist world-system, in contrast to earlier world empires dominated by a single political structure, consists of numerous, independent nation-states, which collectively exert no effective control over the global market. The political geography of capitalism is the interstate system, not the nation-state *per se*. World-systems views divide nations into three broad categories, the core, i.e., the world's leading powers (typically dominated by a single hegemon); the periphery, which comprises large numbers of weak, impoverished states; and the semi-periphery, consisting of wealthier members of the developing world and less developed members of the West. Unlike dependency theory, which takes the nation-state as its unit of analysis, world-systems theory insists upon viewing the entire planet as an indivisible totality, overcoming the dualism between the global and the local. The possibilities of economic development are seen as more open-ended than in dependency theory. The success of the East Asian Newly Industrialized Countries (NICs) since World War II boosted the appeal of world-systems views considerably.

Political geography witnessed sophisticated interpretations of hegemony and territoriality at several spatial scales (see *Political Geography*). Agnew and Corbridge (1995), for example, dissect the 'power container' of the nation-state and its mounting 'leakages' to and from the world-system. O'Tuathail (1996) offers a poststructuralist critical geopolitics, that emphasizes the social origins of statecraft, foreign policy, and geographical knowledge.

3. *Humanistic Geography*

A parallel moment in geography's embrace of social theory occurred with the adoption of humanistic philosophies, particularly phenomenology. Humanists insisted emphatically on the central role of human consciousness, and sought to return purposeful human agents to the center of spatial analysis, as did Renaissance humanists. Philosophically, this line of thought drew upon Husserl, Heidegger, and, more recently, the semiotics of perception, cognition, and language (Buttimer 1976). Humanistic thought successfully refuted the assumption of naturalism, i.e., that there are no substantive differences between the human and natural sciences, revealing that social science is qualitatively different from the physical or natural sciences, so that metaphors drawn from those disciplines are inappropriate to the study of society (Barnes 1996). Because the human subject was poorly theorized in structural Marxism, humanists criticized its economic determinism and teleological reading of history (Duncan and Ley 1982).

Humanistic geography, with roots in the French Vidalian tradition, made numerous contributions to the analysis of place, including the roles of understanding and hermeneutics, the body, experience, and intentionality, and everyday life. Its form of understanding emphasized thought rather than behavior: 'truth,' in this view, could never be grounded simply in objective facts, but in the subjective interpretations that people give to them. This school gave cultural geography a badly needed platform to appropriate hitherto unavailable sources of data, such as literature, poetry, art, and mythology. Tuan's (1977) 'sense of place' studies elaborated upon places as abstract, intangible webs of symbolic meanings produced by and for people, emphasizing the notion of landscapes as authored by agents, for a purpose. Humanism also legitimized later geographical explorations of ethics, morality, and justice (e.g., Smith 1994). Finally, the emphasis on human consciousness was an important antidote to the teleology of structural Marxism; humanistic thought, emphasizing choice and contingency, revealed that history never rolls along a predetermined track.

However, humanistic geography also suffered from serious flaws. Most importantly, lacking any conception of social relations and the social construction of the individual subject, it lapsed into the atomistic ontologies of positivism and neoclassical economics. Unable to appropriate for itself the conceptual foundations necessary to comprehend production, class, the state, power, and struggle, humanism serves largely as a critique but not as an alternative to more structurally-oriented forms of social theory.

4. *Structuration and Realism*

Giddens's (1984) structuration theory effectively overcame the long-standing micro–macro division between social structures that act on their own accord on the one hand and ahistorical individuals on the other. Structuration theory began with the recognition that culture is what is taken-for-granted in daily life, i.e., common sense, the matrix of ideologies that people use to negotiate their way through their everyday worlds. Unlike the ahistorical subject of humanism, culture is acquired through a lifelong process of socialization: individuals never live in a social vacuum, but are socially produced from cradle to grave. Viewing culture as social practice upstaged the prevailing superorganic view in cultural geography, which reified culture as a force suspended above people's heads (Duncan 1980). By emphasizing that the deeply internalized rules and resources that people draw upon everyday simultaneously enable and constrain ways of thinking and acting, structuration theory rescued the living subject from the twin abysses of structural overdetermination and individualistic vol-

untarism. Social formations both produce, and in turn are produced by, their inhabitants (see *Time-geography*).

The counterpart to the unacknowledged preconditions of everyday life are the unintended consequences, the ways in which social relations continually escape the intentions of their creators. Drawing on Marx's famous aphorism that 'Men (sic) make history, but not by means of their own choosing,' structuration theory holds that social and spatial change occurs through intentional action, but not intentionally: in forming their biographies everyday, people unintentionally reproduce and transform their social worlds; individuals are simultaneously produced by, and producers of, history and geography. Thus, everyday thought and behavior do not simply mirror the world, they constitute it (Sayer 1984). Societies and regions are neither reducible to, nor independent of, the everyday lives of their inhabitants. This view linked the macrostructures of social relations to the microstructures of everyday life (Thrift 1983, Pred 1984).

Epistemologically, structuration theory was complemented by the philosophy of realism (Sayer 1984) (see *Critical Realism in Geography*). As with structuralism, realism holds that the interrelations among observations are ontologically real, even if they cannot be measured directly; social relations are thus 'real' even if they are not directly 'observed.' Explanation in realism centers on the distinction between necessary and contingent relations. Necessary relations, identified through theoretical abstraction, concern the essential mechanisms that produce change, not events or observations *per se*; contingent relations, in contrast, are specified empirically in concrete contexts. Whereas necessary conditions must exist in every social explanation, whether the causal properties they describe are activated or not empirically is contingent and indeterminate. The parallels here to structure and agency in structuration theory are clear. This distinction ruptures the unity of generalization and explanation: because different causal powers may generate similar empirical patterns, regularities are not sufficient for explanation; because necessary properties are contingently activated and may occur in only one instance, generalizations are not necessary for explanation either.

5. *Feminism*

Feminist scholarship arose from and contributed to dramatic changes in the status and lives of women, particularly since the women's movement of the 1960s. Feminism is not one unified school of thought, but a series of overlapping discourses, including liberal, Marxist, ecofeminist, radical, and postmodern variants (see *Gender and Feminist Studies in Geography*). Differentiating between the biological category of sex and the social category of gender, feminism stressed gender as a form of resource organization, power relation, and matrix of deeply sedimented presuppositions that underpin daily life. To ignore gender is to assume that mens' lives are 'the norm,' that there is no fundamental difference in the ways in which men and women experience and are constrained by social relations. Feminism served a critical role asserting gender as the first nonclass-based form of social determination.

Central to feminist critiques is the dissection of patriarchy, the domination and oppression of women by men, which is, if not a universal feature of social relations, extremely widespread historically and geographically. Are patriarchy and capitalism necessarily, and not contingently, related? As a political project, feminism is committed to the eradication of patriarchy, to the difference that gender makes in the life opportunities of men and women, to gender as a relation of power. It does not follow that feminism is about eradicating gender differences *per se*; as Hanson (1992) argued, the task is to theorize difference without power.

Feminist scholarship revealed how space is constructed around gender relations, including housing, work, and commuting patterns, access to public and private resources, and the meanings and behaviors that typically favor men and disadvantage women (Bowlby et al. 1990). While early feminist work focused simply on the 'geography of women,' later scholarship explored how gender is intertwined with relations of power, culture, ethnicity, and ideology.

Feminists played a major role in gendering economic geography, including spatial divisions of labor (Massey 1984) and gender differences in commuting patterns (Hanson and Pratt 1995). Feminists exposed as patriarchal the traditional equation of male roles with public spaces and female roles with private ones, which mirrors the broader schism between production and reproduction. Feminists also plunged into epistemology, showing how the construction of knowledge itself is pervasively gendered. Rose (1993) describes how geography's very maleness long colored its construction, arguing for an emancipatory geography that empowers women.

Gender, by definition, includes males. To equate 'gender' with women is to assume men have no gender, that it is not socially constructed or geographically instantiated. Like femininity, masculinity in its multiple forms is reproduced spatially through specific sites, practices, and discourses. Lastly, while gender and sexuality are far from synonymous, recent social theory has also explored the role of sexuality in the construction of personal spaces (see *Sexuality and Geography*).

As geography embraced political economy, it expanded the 'economy' to include the culture of the firm (Schoenberger 1997), the gendered nature of knowledge (Gibson-Graham 1996), and the role of

language and metaphor (Barnes 1996). These works emphasized the role of human actors rather than mechanistic institutions in the unfolding of economic processes, focusing on trust, knowledge, reputation, uncertainty, indeterminacy, and contingency rather than modes of production, class struggle, or technological change. The result has been a more thoroughly 'humanized' economic geography sensitive not only to time and space but to the complexities of social life.

6. *Postmodernism, Representations, Identity, and Postcolonialism*

Following the 'linguistic turn' of many social sciences, geographers exhibited a sustained interest in the politics of knowledge, ideology, discourse, language, symbols, and representations (Barnes and Duncan 1992) as well as the unity of knowledge and power. Far from constituting a web of selfpropelling ideas, all knowledge arises from and in turn contributes to power relations, either by naturalizing or challenging them. Every discourse and representation, whether self-consciously or not, is politically laden, serving some interests and not others; every theory, therefore, is not only an explanation but a legitimation of a particular interest.

Postmodernism contrasts sharply with the Enlightenment quest for universal metanarratives and the modernist assumption that reality is fundamentally ordered; postmodernism emphasizes complexity, randomness, disorder, and chaos. The postmodern picture of reality consists of a multitextured kaleidoscope that can never be adequately captured in a single theory: reality is more complex than any language can describe. Postmodernism celebrates heterogeneity, not commonalities; it accepts uncertainty and ambiguity; it emphasizes ephemerality rather than permanence. However, the simple dichotomy between modern and postmodern fails to do justice to the complexity of both. Marxists conceptualize postmodernism as a cultural response to the material and social transformation of post-Fordist capitalism (Harvey 1989a).

This view destabilized many received categories of geographic thought (Benko and Strohmeyer 1997) (see *Postmodernism in Geography*). Postmodernism's engagement with feminism problematized the unified metanarrative of gender that underpinned feminist work. Postmodernism and Marxism share an uneasy relationship (Harvey 1989a, Soja 1989). In the context of the localities debate, postmodernism holds that theory must be adapted to the temporal and spatial specifics of places.

Postmodernism also paved the way for the injection of semiotics into geographic thought and a wide-ranging debate about the nature of representations of space. Orthodox views, grounded in the correspondence theory of truth, portray representations as claims to truth, not attempts at persuasion; in this view, objectivity is both possible and necessary, and the act of interpretation is unproblematic. Under postmodernism, this perspective gave way to deconstructionist interpretations that emphasize every representation is a simplification, inevitably filled with silences (Barnes and Duncan 1992). Like landscapes, representations are always authored, situated in a context, inescapably partial and biased; thus, representations can only be understood intertextually, within the context of other representations. Further, the manner in which representations are interpreted/consumed is not necessarily how they are intended/produced: meanings of texts and landscape generally escape their authors. This line of thought is insistent that representations are social constructions: while discourses often become taken-for-granted as 'natural,' acquiring an asocial 'objective' status, there can be no value-free representations. Rather, representations of place are inescapably linked to power relations. Thus, they have social consequences (though not always intended ones), becoming part of the social reality that they describe: word-making is also world-making. Cartographers, for example, revealed maps to be discourses, laden with power relations, that constrain perceptions of space along some avenues and not others (Wood 1992, Harley 1989). Pickles (1995) argues that Geographic Information Systems have become implicated in the discursive rewriting of space, essentially underpinning a resurgence of positivist thought.

Human geography has also exhibited an interest in personal and social identity, the processes that shape human experience, establish the individual's boundaries of the self, and define their place in the world. Building on a long tradition of phenomenology, psychotherapy, and symbolic interactionism, this genre stressed the social construction of the individual self and sutured the formation and reproduction of individual and collective identities to place and space (Pile and Thrift 1995). Identities are power relations that reflect and contest relations of normality and marginality. In this light, identities are constructed by establishing difference, by the drawing of boundaries, by defining what they are not: there is always an Other, and 'othering' is a power relation. In contrast to the Enlightenment assumption of the unified human subject, identities can be contradictory bundles, constantly in flux, and inevitably interwoven with power relations, a view suited to the realities of late twentieth century, post-Fordist, hypermobile capitalism, in which the electronic mediation of experience is widespread, rapid time-space compression commonplace, and hyperreality the norm.

Social theory in geography only recently turned to race and ethnicity (Jackson and Penrose 1994), to the complex ways in which it is constructed and negotiated in everyday existence, how it interacts contingently with class and gender, and the complicated manner in

which it plays out over space. Racial and ethnic identity is fundamentally a power relation that typically reflects, reinforces, and naturalizes existing lines of inequality, although it may challenge hegemonic norms as well. The racialized geographies of urban poverty and the underclass have received attention as the products of restructuring, state policy, and local discrimination. To equate 'ethnicity' with nonwhites, however, presumes that whiteness is the norm, or 'default option;' for this reason, geographers have turned to the social construction of whiteness and its multiple political connotations.

The geographical understanding of ethnicity, discourse, and identity was greatly advanced by Said's (1978) *Orientalism*, which opened the door to an imaginative reconceptualization of the politics of spatial representation and identity formation at the worldwide level, revealing colonialism to be as much an ideological process as an economic or political one. Gregory (1994) laid bare geography's complicity in the power relations of global (see *Postcolonial Geography*) modernity: by framing space within 'occularcentric' worldviews of the Enlightenment—essentially, the assumption that the world is knowable, that it can be subjected to the dissection in terms of an all-seeing observer—modernity discursively constructed images, ideas, and identities that deepened the Western penetration of its Other, the non-West.

7. Concluding Comments

Paradoxically, contemporary geography illustrates both greater unity and greater divisions than perhaps at any time in its history. While enormous differences of opinion remain, a broad consensus has emerged that takes seriously political economy and social relations of knowledge and power as fundamental to the understanding of space. Social theory raised class, gender, ethnicity, and power as central foci of spatial analysis, including both their historically constituted material forms on the landscape and their ideological, taken-for-granted dimensions that inform daily life, discourse, and individual and social identities. Social reproduction has been put on a par with the dynamics of production as a key moment in the dissection of spatialities. In the process, geography became highly sensitized to issues of difference, to the politics of knowledge and the diverse ways in which places, and the people who inhabit them are represented to one another and to observers. The ways in which social and spatial relations are explained, justified, and described in politically-laden terms—in short, the manner in which discourse does not simply reflect the social world, but constitutes it—have also figured prominently. As a result of its intercourse with social theory, geography—long a discipline that lagged behind other fields in theoretical vigor—was catapulted from a naïve, atheoretical discipline into a theoretically informed and informative (i.e., generative) one.

Bibliography

Agnew J, Corbridge S 1995 *Mastering Space: Hegemony, Territory and International Political Economy*. Routledge, London

Barnes T J 1996 *Logics of Dislocation: Models, Metaphors, and Meanings of Economic Space*. Guilford, New York

Barnes T J, Duncan J S (eds.) 1992 *Writing Worlds: Discourse, Text and Metaphor in the Representation of Landscape*. Routledge, London

Benko G, Strohmayer U (eds.) 1997 *Space and Social Theory: Interpreting Modernity and Postmodernity*. Blackwell, Oxford, UK

Bowlby S, Lewis J, McDowell L, Foord J 1990 The geography of gender. In: Peet R, Thrift N (eds.) *New Models in Geography*. Unwin Hyman, London, Vol. 2

Buttimer A 1976 *Grasping the dynamism of life-world. Annals of the Association of American Geographers* **66**: 277–92

Castells M 1996 *The Rise of the Network Society*. Blackwell, Oxford, UK

Duncan J S 1980 The superorganic in American cultural geography. *Annals of the Association of American Geographers* **70**: 181–98

Duncan J, Ley D 1982 Structural Marxism and human geography. *Annals of the Association of American Geographers* **72**: 30–59

Gibson-Graham J K 1996 *The End of Capitalism* (*As We Knew It*). Blackwell, Oxford, UK

Giddens A 1984 *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, Berkeley, CA

Gregory D 1994 *Geographical Imaginations* Blackwell, Oxford, UK

Hanson S 1992 Geography and feminism: worlds in collision? *Annals of the Association of American Geographers* **82**: 569–86

Hanson S, Pratt G 1995 *Gender, Work and Space*. Routledge, London

Harley B 1989 Deconstructing the Map. *Cartographica* **26**: 1–20

Harvey D 1982 *The Limits to Capital*. University of Chicago Press, Chicago

Harvey D 1989a *The Condition of Postmodernity*. Blackwell, Oxford, UK

Harvey D 1989b *The Urban Experience*. Blackwell, Oxford, UK

Harvey D 1996 *Justice, Nature and the Geography of Difference*. Blackwell, Oxford, UK

Jackson P, Penrose J (eds.) 1994 *Constructions of Race, Place and Nation*. University of Minnesota Press, Minneapolis, MN

Massey D 1984 *Spatial Divisions of Labor*. Methuen, New York

O'Tuathail G 1996 *Critical Geopolitics: The Politics of Writing Global Space*. University of Minnesota Press, Minneapolis, MN

Pickles J (ed.) 1995 *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford Press, New York

Pile S, Thrift N (eds.) 1995 *Mapping the Subject: Geographies of Cultural Transformation*. Routledge, London

Pred A 1984 Place as historically contingent process: Structuration and the time-geography of becoming places. *Annals of the Association of American Geographers* **74**: 279–97

Rose G 1993 *Feminism and Geography: The Limits of Geographical Knowledge*. University of Minnesota Press, Minnesota, MN

Said E 1978 *Orientalism* 1st edn. Pantheon Books, New York
Sayer A 1984 *Method in Social Science: A Realist Approach*. Hutchinson, London
Schoenberger E 1997 *The Cultural Crisis of the Firm*. Blackwell, Oxford, UK
Smith D M 1994 *Geography and Social Justice*. Blackwell, Oxford, UK
Soja E W 1989 *Postmodern Geographies*. Verso, London
Thrift N-J 1983 On the determination of social action in space and time. *Environment and Planning D: Society and Space* **1**: 23–57
Tuan Y-F 1977 *Space and Place: The Perspective of Experience*. Edward Arnold, London
Wallerstein I 1979 *The Capitalist World Economy: Essays*. Cambridge University Press, Cambridge, UK
Wood D 1992 *The Power of Maps*. Guilford, New York

B. Warf

Space: Linguistic Expression

Spatial cognition is at the heart of much human thinking, as evidenced for example in the explanatory power of spatial metaphors and diagrams, and the explication of spatial concepts has thus played a key role in the development of Western mathematics and philosophy over two and a half millennia (see Jammer 1954). It is therefore easy to assume that spatial notions form a robust, universal core of human cognition. But naïve human spatial language turns out to vary significantly across languages, both in the way it is semantically organized and the way in which it is coded. This has major implications for how we should think about universals in human thought and language.

The Western intellectual tradition has provided an elaborate metalanguage for spatial concepts, but the conceptual underpinnings of naïve human spatial language utilize only a small subset of these concepts. Moreover, there are many aspects of our spatial perception and motor control which utilize much richer and more precise representations of space than are coded in language. The study of the representation of space in language, and its relation to cognition, must therefore be treated as an empirical enterprise, and this article reports on the still quite limited cross-linguistic information about how different languages structure the spatial domain (see Levinson and Wilkins in preparation).

1. *Space as a Semantic Domain*

Languages tend not to treat space as a single, coherent semantic field in the same way that they treat e.g., color (q.v.), kinship (q.v.), or ethnobotany—rather, space is structured as a large semantic field with many distinct subdomains, each of which is closely structured (and often varies considerably across languages). Nearly all languages have 'where'—questions which cover the entire field related (that is with the same or morphologically related forms for 'where-to,' 'where-from,' 'where-at') and describe the location of one object primarily with respect to another. But beyond these general properties, spatial expressions are structured in distinct subdomains, notably topology, frames of reference, and motion, with place names and spatial deixis as complementary areas. These subdomains cross-cut in certain ways (e.g., frames of reference may play a role in motion description), but because they have their own internal organization they need to be kept analytically distinct.

Although, the philosophical and scientific literature gives us many different ways to talk about space, natural languages tend to structure spatial description more in accord with the ideas of Leibniz than of Newton. That is, location is thought about as the place of one thing relative to another (rather than, as in Newtonian thinking, as specifications in the coordinates of an infinite, abstract three-dimensional 'box'). We will call the object to be located the 'figure,' and the object with respect to which the location is specified 'the ground' (other equivalent terms in the literature are 'theme' vs. 'relatum' or 'trajector' vs. 'landmark'). As Talmy (1983) noted, a good ground or landmark object is larger than and less mobile than the figure object to be located. The problems of location specification then largely boil down to the following functional considerations:

(a) Where figure F and ground G are contiguous or coincident, a static F may be said to be 'at G.' Where G is large, it can be subdivided, so that F can be said to be 'at the X-part of G.' This is, roughly, the subdomain of topology.

(b) Where F and G are separated in space, the problem is to specify an angle or direction from G in which F can be found (the search domain). This involves a coordinate system of some kind, a 'frame of reference,' of which there are three main types found in languages.

(c) Where F is in motion, two kinds of ground are especially relevant: the source or the goal of the motion. To indicate the progression of motion (i.e., the increase or decrease of distance from a ground), motion can be specified as 'F is moving towards goal G' or 'F is moving from source S.' In addition, motion can be thought of as bringing about, or destroying, a topological relation, as in 'F moves into G.' But to indicate a direction or vector of motion, both S and G must be specified, or a coordinate system of the kind mentioned in (b) utilized. (Of course, motion can also be specified as taking place within a location, in which case the systems in (a) and (b) can be utilized.

These functional distinctions give us a number of cleavages: stasis (or location) vs. kinesis (or motion), non-angular topological relationships vs. angular relations specified in frames of reference, with potentially cross-cutting domains as in Fig. 1.

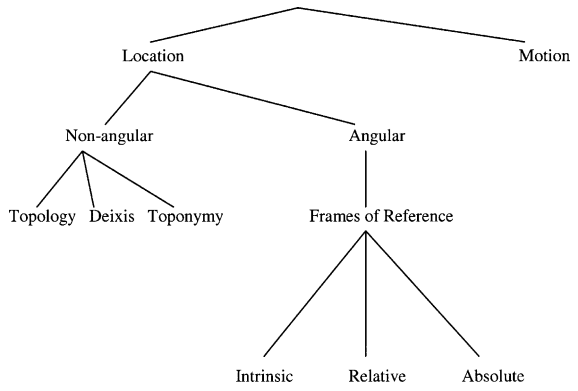

Figure 1
Cross-cutting subdomains of spatial language

Some further details in Fig. 1, like the three frames of reference, will be described below. But two subfields which will not receive further discussion need to be mentioned here. One is *deixis*, which concerns the way in which the location of the speech participants can constitute a special ground or landmark (as in *here*), which then makes interpretation dependent on determining that location (as in *He used to come here*). Languages universally seem to provide a structured set of deictic elements with distinctions that are often spatial (demonstratives like 'this' vs. 'that,' adverbs like 'here' vs. 'there'). Deictic anchoring of spatial expressions is, however, often implicit or pragmatic: *the local pub* implies a pub close to some reference point, which unless specified is taken to be the place of speaking.

A second subfield not treated further here is toponymy, or systems of place names. Place names have been much investigated from an historical point of view: since they are highly conservative they reveal much about the locations of ancient languages and peoples. However, the theory and cross-linguistic typology of place names—e.g., why some are descriptive, others not—is hardly developed at all (see e.g., Hunn 1993).

2. *Universals and Variation in the Semantic Parameters in Spatial Subdomains*

In this section, the semantic parameters involved in spatial description in two major subdomains are described to illustrate the different ways in which languages structure the spatial domain.

2.1 *Topology*

This subdomain, as mentioned, concerns contiguous relations between figure and ground. The term is potentially misleading—topology proper is a branch of geometry concerned with constancies under continuous distortions such as stretching, but is used in studies of spatial language following Piaget to describe the sort of spatial relations covered by the English prepositions *on*, *in*, *at* (Herskovits 1986). These spatial terms are amongst the earliest learnt by Western children, and it has therefore been supposed that they form a universal innate core of spatial notions. However, cross-linguistic studies show that universal semantic concepts are only to be found at a more abstract level. Take, for example, the spatial relations covered by English *on*, as in *The cup is on the table*, *The picture is on the wall*, or *The fresco is on the ceiling*—clearly the term covers a wide area, which is already subdivided in the adpositions (prepositions or postpositions) of many other European languages. Some languages (like Japanese) conflate 'on'-relations with 'over'-relations (as in *The light is over the table*) in their adpositions; others (like Yukatek) even include 'under'-relations as well (as in *The ball is under the table*). A detailed analysis of a dozen languages from different families suggests that universal semantic concepts in this domain lie at a more atomic level (Levinson and Wilkins, in preparation). If we think of the core prototype of English *on* as consisting of the notion of vertical superposition of figure over ground, with a horizontal ground supporting a non-attached figure, then the kind of notions that look universal are at the level of 'vertical superposition,' 'contact,' 'horizontal supporting surface,' (non) 'attachment.' Clearly these more abstract atomic concepts can be rearranged in many different ways to constitute the semantic relations expressible in an individual language.

Topological relations are coded in the European languages primarily in adpositions or case, but as we shall see these are not the only possibilities—some languages code topological relations entirely in nouns or verbs.

2.2 *Frames of Reference*

Linguists have presumed that there are two main frames of reference, which are often called 'deictic' and 'intrinsic,' and which can be illustrated by the ambiguity of *The boy is behind the bus*—either the bus is between the boy and us ('deictic') or he is at the rear of the bus ('intrinsic'). Psychologists have long known, however, that on the vertical dimension our perceptual systems use three frames of reference against which to measure uprightness: an object-centered frame, a viewer-centered frame, and an environmental frame. Cross-linguistic research shows that in fact these three frames of reference, which we will call the intrinsic, the relative, and the absolute frame respectively, may be used on the horizontal plane too (Levinson 1996). These three frames appear to be universally available, yet not all languages make systematic use of them all—some use the absolute frame, others the relative

frame, while most use the intrinsic frame, and some use all three (the main universal constraint appears to be that the relative frame is dependent on the possession of the intrinsic frame).

The intrinsic frame of reference, as its name implies, has to do with the intrinsic parts of ground objects. Recollect that frames of reference have to do with specifying angles or directions from a landmark. One way to achieve this is to designate a side or facet of an object, so that the figure can be said to be lying on an axis drawn from the center of the object through the designated side. Most languages provide a vocabulary for parts of objects, but the ways in which they partition objects can be very variable. In English, for example, the criteria for assigning the *front* of an object involve a complex mix of functional and orientational information: the front of a bus is the direction in which it canonically moves, the front of a building is the side in which it is normally entered, and the front of a television the side designed to be watched. In English the *top* of a bottle remains the top even when tipped over, but in other languages (like Zapotec) a more strictly orientational approach may assign the 'top' to whatever side is currently uppermost. Alternatively (as in Tzeltal), the assignment may rely entirely on the internal axial geometry of the object, so that e.g., the 'face' of an object may be a flat side on a secondary axis, even if normally this is undermost. What all these systems have in common is that they provide a system of designated sides which is independent of the position of the viewer, and (apart from the vertical axis) independent of an environmental orientation — how they achieve this differs not only in detail but in fundamentals.

Another way of obtaining angular distinctions on the horizontal plane is to use the body-axes of the viewer. We can then say that something is, for example, to the 'left' of something else in the visual field. Full versions of such systems use a left/right axis and an orthogonal front/back axis, but again the way in which they map bodily axes onto the ground object is very variable. Although these systems are often called 'deictic' in the literature, this is misleading—they typically have a deictic centre but they need not (as in *He kicked the ball to the left of the goal*), and the other two frames of reference can also have deictic centers. Hence these systems are here said to be 'relative' frames of reference, because they are relative to the body-axes of some observer. Some languages translate the bodily coordinates onto the ground object without rotation: in such languages (like Hausa) *The boy is to the left of the tree* means just what it does in English, but *The boy is in front of the tree* means what in English is described as *The boy is behind the tree*. In contrast, some languages rotate the axes, so that the front of the tree is towards us, but the left of the tree is what we would in English call its right (this is probably a minority option, but is found in some dialects of Tamil). Some languages reflect the axes, as in English, so that left and right remain in the same directions as the viewer's body, but front and back are reversed or 'flipped' over.

Notice that terms like 'front' and 'back' and 'left' and 'right' have intrinsic interpretations (as in *He's sitting at the chairman's left*), and for that reason ambiguities arise of the kind illustrated above with *The boy is behind the bus*. Relative systems are based on the intrinsic sides of humans, and probably arise as extensions of an intrinsic system to objects (like trees or balls) which resist the assignment of named sides. Hence the implicational relation: all relative systems are associated with intrinsic systems, but the possession of an intrinsic system does not necessarily imply the use of a relative one. Relative systems had been thought to be universal, but in fact it turns out that perhaps as many as a third of the world's languages do not employ them, instead making use of absolute systems.

Absolute systems rely on environmental gradients or 'geocentric' coordinates. Such systems use absolute directions, or fixed bearings, a bit like English North, but unlike in English, they are often used on all scales (as in *The cup to the north of yours is mine*, or even *There's a spider on your northern leg*), completely replacing the need for relative systems. There are many different varieties of such systems. Some of them (as in Australian Aboriginal languages) are fully abstract cardinal direction systems, with terms that may gloss 'North,' 'South,' etc., but which denote quadrants or arcs of specific angle, and are often based on bearings rotated from our cardinal directions. To use a system of this kind, speakers must always orient themselves precisely, using multiple environmental clues (wind directions, solar angles, sidereal ecliptics, landscape features, and the like). Other systems are based more directly on one major environmental cue, like prevailing winds (as in languages of the Torres Straits), river drainage (as in Alaska), or prevailing slopes in mountainous terrain (as in the Himalayas). For example, Tenejapan Tzeltal uses an 'uphill,' 'downhill,' 'across' system, based on the general lie of the land in the territory of Tenejapa (Chiapas, Mexico), but the system has been abstracted from the terrain to provide fixed orientation, so that the bearings remain constant outside the territory. Many systems use fixed orthogonal axes, but not all: for example, the Austronesian languages in the Pacific islands typically use the monsoon winds as the basis for one axis, but an inland-sea opposition as the basis for another axis—as one moves around the island, the angles between the axes constantly change.

3. *How Spatial Distinctions are Coded in Language Forms*

Much of the literature gives the impression that spatial distinctions are primarily coded in adpositions, as in

the prepositions of English. In fact, in most languages spatial information is distributed throughout the clause. Some languages have a score or more local cases (nominal suffixes) covering roughly the same distinctions as English prepositions (e.g., Finnish or the NE Caucasian language Avar). Other languages use a combination of case and adposition (as in Tamil), or case and spatial nominal (as in many Australian languages). Special classes of spatial nominals do a lot of the work in isolating languages like Chinese, and they play special roles in other languages (cf. *front* in the English complex preposition *in front of*, and *North* in the adverbial use *He went North*). Many spatial oppositions are made constructionally, as when German prepositions like *auf* take the dative to indicate location, and the accusative to indicate motion towards. Finally, verbs play a crucial role in spatial description. In the case of motion, there is often a special form-class of motion verbs, and languages tend to conflate motion with manner (as in English *crawl*) or direction or 'path' (as in Spanish *salir* 'to go out'), but not generally both (Talmy 1983). In many languages (e.g., Dutch), there is a small set of contrasting verbs (often derived from posture verbs) used to describe static location, encoding properties of the figure, or ground, or both. Some languages have a larger set (as in Tzeltal), and in this case most of the semantic load is taken by the verbs, which may encode such specific details as 'be located in a bowl-shaped container.'

In conclusion, it is clear that spatial concepts form a complex semantic field made up of structured linguistic subdomains. The semantic parameters involved may ultimately have a universal base, but the ways in which the semantic subdomains are organized displays considerable cross-linguistic variation. There is also considerable cross-linguistic variation in the way in which such distinctions are encoded in the form-classes of languages. This linguistic variation is especially interesting given the centrality of spatial concepts in human thinking—and indeed it can be shown that linguistic differences in frames of reference make a difference to how humans memorize and reason non-linguistically (Levinson 1996, in press, Pederson et al. 1998).

See also: Spatial Cognition

Bibliography

Bloom P, Peterson M, Nadel L, Garrett M (eds.) 1996 *Language and Space*. MIT Press, Cambridge, MA

Herskovits A 1986 *Language and Spatial Cognition*. Cambridge University Press, Cambridge, UK

Hunn E 1993 Columbia Plateau Indian place names: What can they teach us? *Journal of Linguistic Anthropology* **6**(1): 3–26

Jammer M 1954 *Concepts of Space: The History of Theories of Space in Physics*. Harvard University Press, Cambridge, MA

Landau B, Jackendoff R 1993 'What' and 'where' in spatial language and spatial cognition. *BBS* **16**(2): 217–65

Levinson S C 1996 Frames of reference and Molyneux's question: Crosslinguistic evidence. In: Bloom P, Peterson M, Nadel L, Garrett M (eds.) *Language and Space*. MIT Press, Cambridge, MA, pp. 109–70

Levinson S C in press *Space in Language and Cognition: Explorations in Linguistic Diversity*. Cambridge University Press, Cambridge, UK

Levinson S C Wilkins D (eds.) in preparation. *Grammars of Space: Towards a Semantic Typology*.

Lyons J 1977 *Semantics*. Cambridge University Press, Cambridge, UK

Miller G, Johnson-Laird P 1976 *Language and Perception*. Cambridge University Press, Cambridge, UK

Pederson E, Danziger E, Wilkins D, Levinson S C, Kita S, Senft G 1998 Semantic typology and spatial conceptualization. *Language* **74**: 557–89

Piaget J, Inhelder B 1956 *The Child's Conception of Space*. Routledge, London

Svorou S 1993 *The Grammar of Space*. Benjamins, Amsterdam

Talmy L 1983 How language structures space. In: Pick H, Acredolo L (eds.) *Spatial Orientation: Theory, Research and Application*. Plenum Press, New York

Talmy L 2000 *Toward a Cognitive Semantics*. MIT Press, Cambridge, MA, Vols. 1, 2

S. C. Levinson

Spatial Analysis in Geography

The proliferation and dissemination of digital spatial databases, coupled with the ever wider use of geographic information systems (GIS), is stimulating increasing interest in spatial analysis from outside the spatial sciences. The recognition of the spatial dimension in social science research sometimes yields different and more meaningful results than analysis which ignores it.

Spatial analysis is a research paradigm that provides a unique set of techniques and methods for analysing events—events in a very general sense—that are located in geographical space (see Table 1). Spatial analysis involves spatial modeling, which includes models of location-allocation, spatial interaction, spatial choice and search, spatial optimization, and space-time. Other entries in the encyclopedia take up these models (e.g., see *Location Theory*; *Spatial Interaction Models*; *Spatial Optimization Models*; *Spatial-temporal Modeling*); this article concentrates on spatial data analysis, the heart of spatial analysis.

1. Spatial Data and the Tyranny of Data

Spatial data analysis focuses on detecting patterns and exploring and modeling relationships between such patterns in order to understand processes responsible

Table 1
Popular techniques and methods in spatial data analysis

	Exploratory spatial data analysis	Model-driven spatial data analysis
Object data		
Point pattern	Quadrat methods Kernel density estimation Nearest neighbor methods *K*-function analysis	Homogeneous and heterogeneous Poisson process models and multivariate extensions
Area data	Global measures of spatial associations: Moran's *I*, Geary's *c*	Spatial regression models
	Local measures of spatial association: G_i and G_i* statistics, Moran's scatter plot	Regression models with spatially autocorrelated residuals
Field data	Variogram and covariogram Kernel density estimation Thiessen polygons	Trend surface models Spatial prediction and kriging Spatial general linear modeling
Spatial interaction data	Exploratory techniques for representing such data	Spatial interaction models
		Location-allocation models
	Techniques to uncover evidence of hierarchical structure in the data such as graph-theoretic and regionalization techniques	Spatial choice and search models
		Modeling paths and flows through a network

for observed patterns. In this way, spatial data analysis emphasizes the role of space as a potentially important explicator of socioeconomic systems, and attempts to enhance understanding of the working and representation of space, spatial patterns, and processes.

1.1 Spatial Data and Data Types

Empirical studies in the spatial sciences routinely employ data for which locational attributes are an important source of information. Such data characteristically consist of one or few cross-sections of observations for either micro-units such as individuals (households, firms) at specific points in space, or aggregate spatial entities such as census tracts, electoral districts, regions, provinces, or even countries. Observations such as these, for which the absolute location and/or relative positioning (spatial arrangement) is explicitly taken into account, are termed spatial data (e.g., see *Spatial Data*).

In the socioeconomic realm, points, lines, and areal units are the fundamental entities for representing spatial phenomena. This form of spatial referencing is also a salient feature of GIS (e.g., see *Geographic Information Systems*; *Spatial Data Infrastructure*). Three broad classes of spatial data can be distinguished:

(a) *object data* where the objects are either points (spatial point patterns or locational data, i.e., point locations at which events of interest have occurred) or areas (area or lattice data, defined as discrete variations of attributes over space),

(b) *field data* (also termed geostatistical or spatially continuous data), that is, observations associated with a continuous variation over space, given values at fixed sampling points, and

(c) *spatial interaction data* (sometimes called link or flow data) consisting of measurements each of which is associated with a link or pair of locations representing point or areas.

The analysis of spatial interaction data has a long and distinguished history in the study of a wide range of human activities, such as transportation movements, migration, and the transmission of information (see *Spatial Interaction*; *Spatial Interaction Models*). Field data play an important role in the environmental sciences, but are less important in the social sciences. This article, therefore, focuses on object data, the most appropriate perspective for spatial analysis applications in the social sciences. Object data include observations for micro-units at specific points in space, that is, spatial point patterns, and/or observations for aggregate spatial entities, that is, area data.

Of note is that point data can be converted to area data, and area data can be represented by point reference. Areas may be regularly shaped such as pixels in remote sensing or irregularly shaped such as statistical reporting units. When divorced from their spatial context such data lose value and meaning. They may be viewed as single realizations of a spatial stochastic process, similar to the approach taken in the analysis of time series.

1.2 The Tyranny of Data

Analyzing and modeling spatial data present a series of problems. Solutions to many of them are obvious, others require extraordinary effort for their solution. Data exercise a power that can lead to misinterpretation and meaningless results; therein lies the tyranny of data.

Quantitative analysis crucially depends on data quality. Good data are reliable, contain few or no mistakes, and can be used with confidence. Unfortunately, nearly all spatial data are flawed to some degree (e.g., see *Spatial Data*). Errors may arise in measuring both the location and attribute properties, but may also be associated with computerized processes responsible for storing, retrieving, and manipulating spatial data. The solution to the data quality problem is to take the necessary steps to avoid having faulty data determining research results.

The particular form (i.e., size, shape, and configuration) of the spatial aggregates can affect the results of the analysis to a varying, usually unknown, degree as evidenced in various types of analysis (see, e.g., Openshaw and Taylor 1979; Baumann et al. 1983). This problem generally has become recognized as the modifiable areal units problem (MAUP), the term stemming from the fact that areal units are not 'natural' but usually arbitrary constructs. For reasons of confidentiality, social science data (e.g., census data) are not often released for the primary units of observation (individuals), but only for a set of rather arbitrary areal aggregations (enumeration districts or census tracts). The problem arises whenever area data are analyzed or modeled and involves two effects: one derives from selecting different areal boundaries while holding the overall size and the number of areal units constant (the zoning effect). The other derives from reducing the number but increasing the size of the areal units (the scale effect). There is no analytical solution to the MAUP, but questions of the following kind have to be considered in constructing an areal system for analysis: What are the basic spatial entities for defining areas? What theory guides the choice of the spatial scale? Should the definition process follow strictly statistical criteria and merge basic spatial entities to form larger areas using some regionalization algorithms (see Wise et al. 1996)? These questions pose daunting challenges.

In addition, boundary and frame effects (i.e., the geometric structure of the study area) may affect spatial analysis and the interpretation of results. These problems are considerably more complex than in time series. Although several techniques, such as refined K-function analysis, take the effect of boundaries into account, there is a need to study boundary effects more systematically.

An issue that has been receiving increasing attention relates to the suitability of data. If the data, for example, are available only at the level of spatial aggregates, but the research question is at the individual respondent level, then the ecological fallacy (ecological bias) problem arises. Using area-based data to draw inferences about underlying individual-level processes and relationships poses considerable risks. This problem relates to the MAUP through the concept of spatial autocorrelation.

Spatial autocorrelation (also referred to as spatial dependence or spatial association) in the data can be a serious problem (e.g., see *Spatial Autocorrelation*), rendering conventional statistical analysis unsafe and requiring specialized spatial analytical tools. This problem refers to situations where the observations are nonindependent over space. That is, nearby spatial units are associated in some way. Sometimes, this association is due to a poor match between the spatial extent of the phenomenon of interest (e.g., a labor or housing market) and the administrative units for which data are available. Sometimes, it is due to a spatial spillover effect. The complications are similar to those found in time-series analysis, but are exacerbated by the multidirectional, two-dimensional nature of dependence in space rather than the unidirectional nature in time. Avoiding the pitfalls arising from spatially correlated data is crucial to good spatial data analysis, whether exploratory or confirmatory. Several scholars even argue that the notion of spatial autocorrelation is at the core of spatial analysis (see, e.g., Tobler 1979). No doubt, much of current interest in spatial analysis is directly derived from the monograph of Cliff and Ord (1973) on spatial autocorrelation that opened the door to modern spatial analysis.

2. Pattern Detection and Exploratory Analysis

Exploratory data analysis is concerned with the search for data characteristics such as trends, patterns, and outliers. This is especially important when the data are of poor quality or genuine *a priori* hypotheses are lacking. Many such techniques emphasize graphical views of the data that are designed to highlight particular features and allow the analyst to detect patterns, relationships, outliers, etc. Exploratory spatial data analysis (ESDA), an extension of exploratory data analysis (EDA) (Haining 1990, Cressie 1993), is geared especially to dealing with the spatial aspects of data.

2.1 Exploratory Techniques for Spatial Point Patterns

Point patterns arise when the important variables to be analysed is the location of events. At the most basic level, the data comprise only the spatial coordinates of events. They might represent a wide variety of spatial phenomena, such as cases of disease, crime incidents,

pollution sources, or locations of stores (e.g., see *Spatial Pattern, Analysis of*). Research typically concentrates on whether the proximity of particular point events, their location in relation to each other, represents a significant (i.e., nonrandom) pattern. Exploratory spatial point pattern analysis is concerned with exploring the first and second order properties of spatial point pattern processes. First order effects relate to variation in the mean value of the process (a large scale trend), while second order effects result from the spatial correlation structure or the spatial dependence in the process.

Three types of methods are important: Quadrat methods, kernel estimation of the intensity of a point pattern, and distance methods. Quadrat methods involve collecting counts of the number of events in subsets of the study region. Traditionally, these subsets are rectangular (thus, the name quadrat), although any shape is possible. The reduction of complex point patterns to counts of the number of events in quadrats and to one-dimensional indices is a considerable loss of information. There is no consideration of quadrat locations or of the relative positions of events within quadrats. Thus, most of the spatial information in the data is lost. Quadrat counts destroy spatial information, but they give a global idea of subregions with high or low numbers of events per area. For small quadrats more spatial information is retained, but the picture degenerates into a mosaic with many empty quadrats.

Estimating the intensity of a spatial point pattern is very like estimating a bivariate probability density, and bivariate kernel estimation can be adapted easily to give an estimate of intensity. Choice of the specific functional form of the kernel presents little practical difficulty. For most reasonable choices of possible probability distributions the kernel estimate will be very similar, for a given bandwidth. The bandwidth determines the amount of smoothing. There are techniques that attempt to optimize the bandwidth given the observed pattern of event location.

A risk underlying the use of quadrats is that any spatial pattern detected may be dependent upon the size of the quadrat. In contrast, distance methods make use of precise information on the locations of events and have the advantage of not depending on arbitrary choices of quadrat size or shape. Nearest neighbor methods reduce point patterns to one-dimensional nearest neighbor summary statistics (see Dacey 1960, Getis 1964). But only the smallest scales of patterns are considered. Information on larger scales of patterns is unavailable. These statistics indicate merely the direction of departure from complete spatial randomness (CSR). The empirical K function, a reduced second-moment measure of the observed process, provides a vast improvement over nearest neighbor indices (see Ripley 1977, Getis 1984). It uses the precise location of events and includes all event–event distances, not just nearest neighbor distances, in its estimation. Care must be taken to correct for edge effects. K-function analysis can be used not only to explore spatial dependence, but also to suggest specific models to represent it and to estimate the parameters of such models. The concept of K-functions can be extended to the multivariate case of a marked point process (i.e., locations of events and associated measurements or marks) and to the time-space case.

2.2 Exploratory Analysis of Area Data

Exploratory analysis of area data is concerned with identifying and describing different forms of spatial variation in the data. Special attention is given to measuring spatial association between observations for one or several variables. Spatial association can be identified in a number of ways, rigorously by using an appropriate spatial autocorrelation statistic (Cliff and Ord 1981), or more informally, for example, by using a scatter-plot and plotting each value against the mean of neighboring areas (Haining 1990).

In the rigorous approach to spatial autocorrelation the overall pattern of dependence in the data is summarized in a single indicator, such as Moran's I and Geary's c. While Moran's I is based on cross-products to measure value association, Geary's c employs squared differences. Both require the choice of a spatial weights or contiguity matrix that represents the topology or spatial arrangement of the data and represents our understanding of spatial association. Getis (1991) has shown that these indicators are special cases of a general formulation (called gamma) defined by a matrix representing possible spatial associations (the spatial weights matrix) among all areal units, multiplied by a matrix representing some specified nonspatial association among the areas. The nonspatial association may be a social, economic, or other relationship. When the elements of these matrices are similar, high positive autocorrelation arises. Spatial association specified in terms of covariances leads to Moran's I, specified in terms of differences, to Geary's c.

These global measures of spatial association can be used to assess spatial interaction in the data and can be visualized easily by means of a spatial variogram, a series of spatial autocorrelation measures for different orders of contiguity. A major drawback of global statistics of spatial autocorrelation is that they are based on the assumption of spatial stationarity, which implies *inter alia* a constant mean (no spatial drift) and constant variance (no outliers) across space. This was useful in the analysis of small data sets characteristic of pre-GIS times but is not very meaningful in the context of thousands or even millions of spatial units that characterize current, data-rich environments.

In view of increasingly data-rich environments, a focus on local patterns of association ('hot spots') and an allowance for local instabilities in overall spatial

association recently has been suggested as a more appropriate approach. Examples of techniques that reflect this perspective are the various geographical analysis machines developed by Openshaw and co-workers (see, e.g., Openshaw et al. 1990), the Moran scatter plot (Anselin 1996), and the distance-based G_i and G_i* statistics of Getis and Ord (1993). This last has gained wide acceptance. These G-indicators can be calculated for each location i in the data set as the ratio of the sum of values in neighboring locations (defined to be within a given distance or order of contiguity) to the sum over all the values. The two statistics differ with respect to the inclusion of the value observed at i in the calculation [included in G_i*, not included in G_i]. They can be mapped easily and used in an exploratory analysis to detect the existence of pockets of local nonstationarity, to identify distances beyond which no discernible association arises, and to find the appropriate spatial scale for further analysis (e.g., see *Spatial Association, Measures of*).

No doubt, ESDA provides useful means to generate insights into global and local patterns and associations in spatial data sets. The use of ESDA techniques, however, generally is restricted to expert users interacting with the data displays and statistical diagnostics to explore spatial information and to fairly simple low-dimensional data sets. In view of these limitations, there is a need for novel exploration tools sufficiently automated and powerful to cope with the data-richness-related complexity of exploratory analysis in spatial data environments (see, e.g., Openshaw and Fischer 1994).

3. *Model Driven Spatial Data Analysis*

ESDA is a preliminary step in spatial analysis to more formal modeling approaches. Model-driven analysis of spatial data relies on testing hypotheses about patterns and relationships, utilizing a range of techniques and methodologies for hypothesis testing, the determination of confidence intervals, estimation of spatial models, simulation, prediction, and assessment of model fit. Getis and Boots (1978), Cliff and Ord (1981), Upton and Fingleton (1985), Anselin (1988), Griffith (1988), Haining (1990), Cressie (1993), Bailey and Gatrell (1995) have helped to make model-driven spatial data analysis accessible to a wide audience in the spatial sciences.

3.1 *Modeling Spatial Point Patterns*

Spatial point pattern analysis grew out of a hypothesis testing and not out of the pattern recognition tradition. The spatial pattern analyst tests hypotheses about the spatial characteristics of point patterns. Typically, complete spatial random (CSR) represents the null hypothesis against which to assess whether observed point patterns are regular, clustered, or random. The standard model for CSR is that events follow a homogeneous Poisson process over the study region, that is, events are independently and uniformly distributed over space, equally likely to occur anywhere in the study region and not interacting with each other.

Various statistics for testing CSR are available. Nearest neighbor tests have their place in distinguishing CSR from spatially regular or clustered patterns. But little is known about their behavior when CSR does not hold. The K-function may suggest a way of fitting alternative models. Correcting for edge effects, however, might provide some difficulty. The distribution theory for complicated functions of the data can be intractable even under the null hypothesis of CSR. Monte Carlo tests is a way around this problem.

If the null hypothesis of CSR is rejected, the next obvious step in model-driven spatial pattern analysis is to fit some alternative (parametric) model to the data. Departure from CSR is typically toward regularity or clustering of events. Clustering can be modeled through a heterogeneous Poisson process, a doubly stochastic point process, or a Poisson cluster process arising from the explicit incorporation of a spatial clustering mechanism. Simple inhibition processes can be utilized to model regular point patterns. Markov point processes can incorporate both elements through large-scale clustering and small-scale regularity. After a model has been fitted (usually via maximum likelihood or least squares using the K-function), diagnostic tests have to be performed to assess its goodness-of-fit. Inference for the estimated parameters is often needed in response to a specific research question. The necessary distribution theory for the estimates can be difficult to obtain in which case approximations may be necessary. If, for example, clustering is found, one may be interested in the question of whether particular spatial aggregations, or clusters, are associated with proximity to particular sources of some other factor. This leads to multivariate point pattern analysis, a special case of marked spatial point process analysis. For further details see Cressie (1993).

3.2 *Modeling Area Data*

Linear regression models constitute the leading modeling approach for analyzing social and economic phenomena. But conventional regression analysis does not take into account problems associated with possible cross-sectional correlations among observational units caused by spatial dependence. Two forms of spatial dependence among observations may invalidate regression results: spatial error dependence and spatial lag dependence.

Spatial error dependence might follow from measurement errors such as a poor match between the spatial units of observation and the spatial scale of the

phenomenon of interest. Presence of this form of spatial dependence does not cause ordinary least squares estimates to be biased, but it affects their efficiency. The variance estimator is downwards biased, thus inflating the R^2. It also affects the t- and F-statistics for tests of significance and a number of standard misspecification tests, such as tests for heteroskedasticity and structural stability (Anselin and Griffith 1988). To protect against such difficulties, one should use diagnostic statistics to test for spatial dependence among error terms and, if necessary, take action to properly specify the spatially autocorrelated residuals. Typically, dependence in the error term is specified as a spatial autoregressive or as a spatial moving average process. Such regression models require nonlinear maximum likelihood estimation of the parameters (Cliff and Ord 1981, Anselin 1988).

In the second form spatial lag dependence, spatial autocorrelation is attributable to spatial interactions in data. This form may be caused, for example, by significant spatial externalities of a socioeconomic process under study. Spatial lag dependence yields biased and, also, inconsistent parameters. To specify a regression model involving spatial interaction, one must incorporate the spatial dependence into the covariance structure either explicitly or implicitly by means of an autoregressive and/or moving-average interaction structure. This constitutes the model identification problem that is usually carried out using the correlogram and partial correlogram. A number of spatial regression models, that is, regression models with spatially lagged dependent variables (spatial autoregressive models), have been developed that include one or more spatial weight matrices which describe the many spatial associations in the data. The models incorporate either a simple general stochastic autocorrelation parameter or a series of autocorrelation parameters, one for each order contiguity (see Cliff and Ord 1981, Anselin 1988).

Maximum likelihood procedures are fundamental to spatial regression model estimation, but data screening and filtering can simplify estimation. Tests and estimators are clearly sensitive not only to the MAUP, but also to the specification of the spatial interaction structure represented by the spatial weights matrix. Recent advances in computation-intensive approaches to estimation and inference in econometrics and statistical modeling may yield new ways to tackle this specification issue. In practice, it is often difficult to choose between regression model specifications with spatially autocorrelated errors and regression models with spatially lagged dependent variables, though the 'common factor' approach (Bivand 1984) can be applied if the spatial lags are neatly nested.

Unlike linear regression, for which a large set of techniques for model specification and estimation now exist, the incorporation of spatial effects into nonlinear models in general—and into models with limited dependent variables or count data (such as log-linear, logit and tobit models) in particular—is still in its infancy. The hybrid log-linear models of Aufhauser and Fischer (1985) are among the few exceptions. Similarly, this is true for the design of models that combine cross-sectional and time series data for areal units. See Hordijk and Nijkamp (1977) for dynamic spatial diffusion models.

4. *Toward Intelligent Spatial Analysis*

Spatial analysis is currently entering a period of rapid change leading to what is termed intelligent spatial analysis (sometimes referred to as geocomputation). The driving forces are a combination of huge amounts of digital spatial data from the GIS data revolution (with 100,000 to millions of observations), the availability of attractive softcomputing tools, the rapid growth in computational power, and the new emphasis on exploratory data analysis and modeling.

Intelligent spatial analysis has the following properties. It exhibits computational adaptivity (i.e., an ability to adjust local parameters and/or global configurations to accommodate in response to changes in the environment); computational fault tolerance in dealing with incomplete, inaccurate, distorted, missing, noisy and confusing data, information rules, and constraints; speed approaching human-like turnaround; and error rates that approximate human performance. The use of the term 'intelligent' is, therefore, closer to that in computational intelligence than in artificial intelligence. The distinction between artificial and computational intelligence is important because our semantic descriptions of models and techniques, their properties, and our expectations of their performance should be tempered by the kind of systems we want and the ones we can build (Bezdek 1994).

Much of the recent interest in intelligent spatial analysis stems from the growing realization of the limitations of conventional spatial analysis tools as vehicles for exploring patterns in data-rich GI (geographic information) and RS (remote sensing) environments and from the consequent hope that these limitations may be overcome by judicious use of computational intelligence technologies such as evolutionary computation (genetic algorithms, evolutionary programming, and evolutionary strategies) (see Openshaw and Fischer 1994) and neural network modeling (see Fischer 1998). Neural networks models may be viewed as nonlinear extensions of conventional statistical models that are applicable to two major domains: first, as universal approximators to areas such as spatial regression, spatial interaction, spatial choice, and space-time series analysis (see, e.g., Fischer and Gopal 1994); and second, as pattern recognizers and classifiers to allow the user to sift through the data intelligently, reduce dimensionality, and find patterns

of interest in data-rich environments (see, e.g., Fischer et al. 1997).

See also: Geographic Information Systems; Location Theory; Mathematical Models in Geography; Regional Science; Remote Sensing; Spatial Association, Measures of; Spatial Autocorrelation; Spatial Choice Models; Spatial Data; Spatial Data Infrastructure; Spatial Interaction; Spatial Interaction Models; Spatial Optimization Models; Spatial Pattern, Analysis of; Spatial Sampling; Spatial-temporal Modeling; Spectral Analysis; Urban Geography; Urban System in Geography

Bibliography

Anselin L 1988 *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht

Anselin L 1996 The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fischer M M, Scholten H J, Unwin D (eds.) *Spatial Analytical Perspectives on GIS*. Taylor & Francis, London

Anselin L, Florax R J G M (eds.) 1995 *New Directions in Spatial Econometrics*. Springer, Berlin

Anselin L, Griffith D A 1988 Do spatial effects really matter in regression analysis? *Papers, Regional Science Association* **65**: 11–34

Aufhauser E, Fischer M M 1985 Log-linear modelling and spatial analysis. *Environment and Planning A* **17**: 931–51

Bailey T C, Gatrell A C 1995 *Interactive Spatial Data Analysis*. Longman, Harlow, UK

Baumann J H, Fischer M M, Schubert U 1983 A multiregional labor supply model for Austria: The effects of different regionalizations in multiregional labour market modelling. *Papers, Regional Science Association* **52**: 53–83

Bezdek J C 1994 What's computational intelligence. In: Zurada J M, Marks II R J, Robinson C J (eds.) *Computational Intelligence: Imitating Life*. Institute of Electrical and Electronics Engineers, New York

Bivand R S 1984 Regression modeling with spatial dependence: An application of some class selection and estimation methods. *Geographical Analysis* **16**: 25–37

Cliff A D, Ord J K 1973 *Spatial Autocorrelation*. Pion, London

Cliff A D, Ord J K 1981 *Spatial Processes, Models & Applications*. Pion, London

Cressie N A C 1993 *Statistics for Spatial Data*, rev. edn. Wiley, New York

Dacey M F 1960 A note on the derivation of the nearest neighbour distances. *Journal of Regional Science* **2**: 81–7

Fischer M M 1998 Computational neural networks—a new paradigm for spatial analysis. *Environment and Planning A* **30**(10): 1873–91

Fischer M M, Getis A (eds.) 1997 *Recent Developments in Spatial Analysis—Spatial Statistics, Behavioural Modelling, and Computational Intelligence*. Springer, Berlin

Fischer M M, Gopal S 1994 Artificial neural networks. A new approach to modelling interregional telecommunication flows. *Journal of Regional Science* **34**: 503–27

Fischer M M, Gopal S, Staufer P, Steinnocher K 1997 Evaluation of neural pattern classifiers for a remote sensing application. *Geographical Systems* **4**: 195–223, 233–4

Fischer M M, Scholten H J, Unwin D (eds.) 1996 *Spatial Analytical Perspectives on GIS*. Taylor & Francis, London

Fotheringham S, Rogerson P (eds.) 1994 *Spatial Analysis and GIS*. Taylor & Francis, London

Getis A 1964 Temporal land-use pattern analysis with the use of nearest neighbor and quadrat methods. *Annals of the Association of American Geographers* **54**: 391–9

Getis A 1984 Interaction modelling using second-order analysis. *Environment and Planning A* **16**: 173–83

Getis A 1991 Spatial interaction and spatial autocorrelation: A cross-product approach. *Papers, Regional Science Association* **69**: 69–81

Getis A, Boots B 1978 *Models of Spatial Processes*. Cambridge University Press, Cambridge, UK

Getis A, Ord J K 1993 The analysis of spatial association by use of distance statistics. *Geographical Analysis* **25**(3): 276

Griffith D A 1988 *Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series*. Kluwer Academic, Dordrecht

Haining R 1990 *Spatial Data Analysis in the Social Sciences*. Cambridge University Press, Cambridge

Hordijk L, Nijkamp P 1977 Dynamic models of spatial autocorrelation. *Environment and Planning A* **9**: 505–19

Longley P, Batty M (eds.) 1996 *Spatial Analysis: Modelling in a GIS Environment*. GeoInformation International, Cambridge, UK

Openshaw S, Cross A, Charlton M 1990 Building a prototype geographical correlates exploration machine. *International Journal of Geographical Information Systems* **4**: 297–312

Openshaw S, Fischer M M 1994 A framework for research on spatial analysis relevant to geostatistical information systems in Europe. *Geographical Systems* **2**(4): 325–37

Openshaw S, Taylor P 1979 A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In: Wrigley N (eds.) *Statistical Applications in the Spatial Sciences*. Pion, London

Ripley B D 1977 Modelling spatial patterns. *Journal of the Royal Statistical Society* **39**: 172–212

Tobler W 1979 Cellular geography. In: Gale S, Olsson G (eds.) *Philosophy in Geography*. Reidel, Dordrecht

Upton J G, Fingleton B 1985 *Spatial Data Analysis by Example*. Wiley, New York

Wise S, Haining R, Ma J 1996 Regionalisation tools for the exploratory spatial analysis of health data. In: Fischer M M, Getis A (eds.) *Recent Developments in Spatial Analysis—Spatial Statistics, Behavioural Modelling and Computational Intelligence*. Springer, Berlin

M. M. Fischer

Spatial Association, Measures of

In the context of the following discussion, association means connectedness or relationship. If spatial association exists, the connectedness is over space (the earth's surface), and it is mappable. If variables are somehow held together tightly or loosely in space, as if a magnet were pulling them into a pattern reminiscent

of the well-known iron filings demonstration, one can say that the variables are spatially associated.

1. Introduction

Spatial association can characterize a single variable or many variables simultaneously. A single variable may be *spatially autocorrelated*; that is, values of the variable are somehow connected or related spatially (see *Spatial Autocorrelation*). Many variables—say, environmental variables such as air quality, water quality, certain species location, and human activities—may be associated one with another at one or more sites. In addition, spatial association and *spatial interaction* are related (see *Spatial Interaction*). If there is spatial interaction, for example, the movement of commuters to workplaces, there is also spatial association since home places are spatially connected to work places.

The scientist interested in the study of spatial association is motivated by the need to understand the spatial relationships between variables at specific sites or how variables of interest are spread across the earth's surface. Many disciplines are particularly concerned with the study of spatial association. Among them are geography, regional science, ecology, epidemiology, regional and urban economics, geology, hydrology and geomorphology. Since this is a broad and popular interest area, many methods and techniques of analysis have been created to confirm notions of spatial association between and among variables. Scientists test or theorize about variables to determine whether spatial association, either observed or expected, actually can be confirmed. As a result of this interest, statistical procedures have been developed for identifying and measuring the existence of spatial association. After a short discussion of maps depicting spatial association, the remainder of this entry will outline the characteristics of measurement procedures.

2. Maps Showing Spatial Association

Maps depicting spatial association fall into several categories: *in situ*, flows, time series, distance decay, and residuals. *In situ* maps reveal by means of overlays the possible association of two or more variables in a region (see Fig. 1). Flow maps show the interaction between sites (see Fig. 2). Time series usually are shown as a series of map views over time (days, years, etc). These represent spatial relationships as they evolve or change over time (see Fig. 3a, b, c). Distance decay maps emphasize Tobler's *First Law of Geography* that 'everything is related to everything else, [but] near things are more related than distant things'. A distance decay map might show that the effect of an earthquake declines with distance from its epicenter or

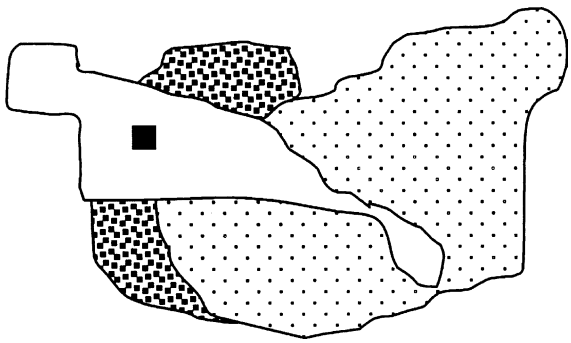

Figure 1
The study area represented by the black square indicates that there may be an *in situ* relationship among the three variables shown by different shading

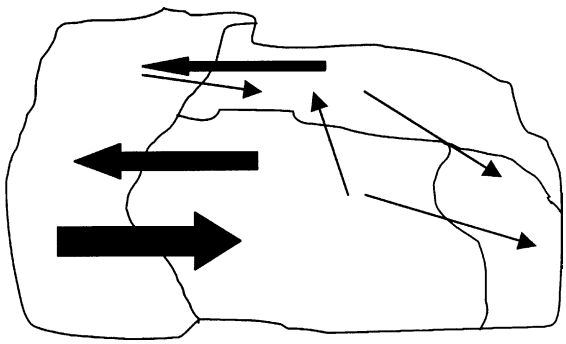

Figure 2
The width of the arrows indicates the magnitude of the interaction

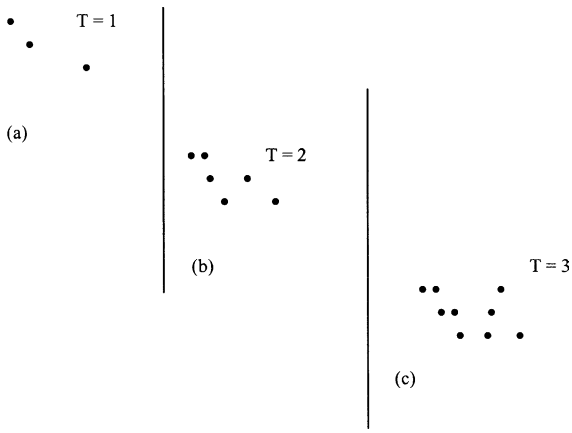

Figure 3
Over time {(a), (b), (c)}, there is an increase in the number of events at the same time as spacing between events is reduced

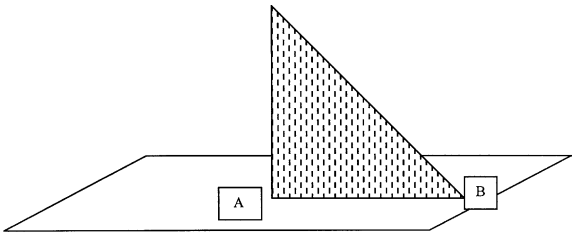

Figure 4
On the plane, as distance increases from A to B the effect on A of the variable indicated by the vertical triangle declines

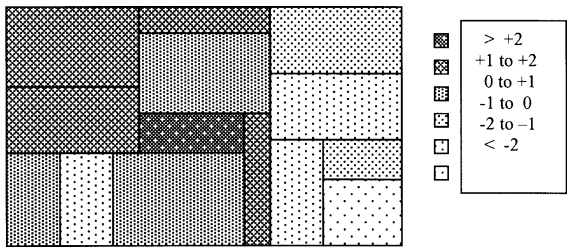

Figure 5
In standard deviational units, the map shows residuals from a regression model. It remains to be determined whether or not there is a spatially statistically significant association among the residuals

that the influence of a retail store declines with distance from its site (see Fig. 4). Residual maps show the remaining 'unexplained' variation after a model has been used to statistically explain a relationship between two or more variables (see Fig. 5). The unexplained spatial variation represents that part of whatever total variation exists in a mathematical model that is not accounted for by the variables included in the model. Residual maps help researchers identify new variables that might belong in a more complete model, such as a transportation cost, a regional effect, or a previously unidentified physical or human anomaly.

3. *Spatial Association in Matrix Terms*

In order to answer scientific questions about spatial association, one can reduce the complexities of what one sees on a map to a mathematically manipulable set of numbers (Hubert et al. 1981, Getis 1991). The following simple cross-product formula represents spatial association:

$$\Gamma = \sum_{i}^{n}\sum_{j}^{n} w_{ij}\, y_{ij} \qquad (1)$$

(a)

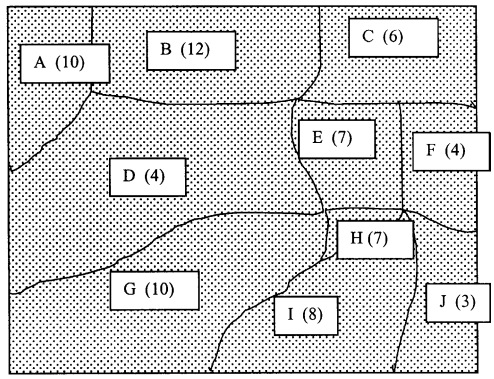

W =

	A	B	C	D	E	F	G	H	I	J
A	0	1	0	1	0	0	0	0	0	0
B	1	0	1	1	0	0	0	0	0	0
C	0	1	0	0	1	1	0	0	0	0
D	1	1	0	0	1	0	1	0	0	0
E	0	0	1	1	0	1	0	1	0	0
F	0	0	1	0	1	0	0	0	0	1
G	0	0	0	1	0	0	0	1	1	0
H	0	0	0	0	1	0	1	0	1	0
I	0	0	0	0	0	0	1	1	0	1
J	0	0	0	0	0	1	0	0	1	0

(b)

Y =

	A	B	C	D	E	F	G	H	I	J
A	0	2	4	1	3	6	0	3	2	7
B	2	0	6	3	5	8	2	5	4	9
C	4	6	0	3	1	2	4	1	2	3
D	1	3	3	0	2	5	1	2	1	6
E	3	5	1	2	0	3	3	0	1	4
F	6	8	2	5	3	0	6	3	4	1
G	0	2	4	1	3	6	0	3	2	7
H	3	5	1	2	0	3	3	0	1	4
I	2	4	2	1	1	4	2	1	0	5
J	7	9	3	6	4	1	7	4	5	0

(c)

Figure 6
The spatial pattern shown in (a) is reduced to a **W** matrix of spatial associations (b) and a **Y** matrix (c) made up of absolute differences between values associated with each region

Γ is the measure of spatial association, and the w_{ij} are contained in a matrix, **W**, of values that represent the relationship of each location of interest to all other sites (see Fig. 6a and b). Figure 6a is a hypothetical map of ten regions. The numbers in parentheses represent the value that a hypothetical variable, Y, takes in each of the regions. Suppose it is decided that regions that are contiguous to one another are associated. The **W** matrix (Fig. 6b) indicates the spatial relationship of each region to every other. For 6b, the convention is

	A	B	C	D	E	F	G	H	I	J
A	0	**2**	0	**1**	0	0	0	0	0	0
B	**2**	0	**6**	**3**	0	0	0	0	0	0
C	0	**6**	0	0	**1**	**2**	0	0	0	0
D	**1**	**3**	0	0	**2**	0	**1**	0	0	0
E	0	0	**1**	**2**	0	**3**	0	**0**	0	0
F	0	0	**2**	0	**3**	0	0	0	0	**1**
G	0	0	0	**1**	0	0	0	**3**	**2**	0
H	0	0	0	0	**0**	0	**3**	0	**1**	0
I	0	0	0	0	0	0	**2**	**1**	0	**5**
J	0	0	0	0	0	**1**	0	0	**5**	0

Gamma = 66

Figure 7
The value of Γ is made up of the sum of numbers shown in bold

adopted of placing a 1 in the matrix if a link (contiguity) exists and zero otherwise. The **Y** matrix (see Fig. 6c) contains numbers that are the absolute difference between values for all combinations of the regions. For example, the absolute difference between values in regions C and E is shown as an element of the matrix in row C and column E, that is, $|6–7| = 1$.

The two matrices, when summed as a cross-product as shown in Eqn. 1 (see Fig. 7), give the value of Γ. If the values in the **Y** matrix are small for those same rows and columns that correspond to the $w_{ij} = 1$ values in the **W** matrix, the value of Γ will be small. For the example shown, Γ equals 66, an average of 2.2 for each of the 30 links between contiguous regions. If the number 1 appeared in all the **W** cells, the total Γ would have been 300 or a mean of 3.3 for each of the hypothetical 90 links. The implication here is that the contiguous neighbors that characterize **W** gives a smaller gamma value than one might expect if the values y_{ij} were randomly placed in the 10 regions. Γ indicates that there may be a spatial association among the Y values. We use the term 'may be' to indicate that we have not yet used a statistical test to confirm that Y is in fact a spatially associated variable.

3.1 The W Matrix

In the preceding paragraph, we outlined a simple form of a **W** matrix, one that was constructed based on contiguous spatial relationships between regions. The **W** matrix embodies our preconceived or derived understanding of spatial relationships. If we believe, or if theory tells us, that a particular spatial relationship is distance dependent, then the **W** matrix should reflect that supposition. For example, if it is assumed that a spatial relationship declines in strength as distance increases from any given site, then the **W** matrix will show that nearby areas are weighted more highly than sites that are far from one another. Various distance-decay formulations theorized or derived for such phenomena as travel behavior, economic interaction, or disease transmission would require the elements within the **W** matrix to reflect these effects. Thus, a typical **W** matrix might contain elements of the form $w_{ij} = 1/d_{ij}^{\alpha}$ (α greater than 1). Thus, short distance relationships would be weighted more heavily than the long distances.

3.2 The Y Matrix

The nonspatial matrix, **Y**, gives us a view of how a variable(s) whose elements are georeferenced, that is, values of the variable are indexed at certain sites, *might* be associated with one another. The elements of the **Y** matrix, y_{ij}, represent the interaction of values of **Y**. Values can interact in a number of ways: addition, subtraction, multiplication, division, or combinations of these. An important type of interaction is that of covariance, that is, a value at one site may vary from the mean of all sites (the variation) and this difference is multiplied by the variation at another site. Taken over all sites, the *covariation* is

$$\sum_{i}^{n}\sum_{j}^{n} (y_i - \bar{y})(y_j - \bar{y})$$

and the *covariance* is this sum divided by the sample size, n.

3.3 Γ

When the matrix **W** and the matrix **Y** are similarly structured, one can conclude that **Y** values interact very much as the individual regions are spatially associated. Thus, Γ reveals the spatial association in the **Y** variable. In our previous example, we purposely created a **W** matrix that represented a relationship between contiguous regions. The fact that the values of the **Y** variable tend to be similar to one another when regions are contiguous allows us to conclude that the two matrices are similarly structured.

In the literature on spatial association, much attention has been given to the structure of **W** and **Y**. Each, in a sense, represents a related area of research. Those concerned with the **W** matrix have devised ever increasingly complex representations of spatial relations among spatial units. From the simple contiguity structure presented above, empiricists and theorists have constructed **W** matrices that represent supposed migration distances, disease transmission distances, assumed hierarchical spatial relations, political boundary effects, psychological distances, barriers to communication effects, economic effects of distance, social distances, and so on. In experimental work, the matrix **W** has also been used to represent concentric bands measured from various spatial foci or as representative of cumulative distances. These latter matrices help researchers to identify 'critical' distances, that is, distances at which association reaches some peak.

Research on the structure of the **Y** matrix usually is a response to a particular type of possible spatial

association. For example, the researcher may wonder whether the association is that of positive covariance. **Y** is structured in such a way that statistical tests can be carried out on the research hypothesis that statistically significant spatial association exists. Several of these formulations are given below. Each is an alternative method of identifying spatial association.

4. Spatial Association as Covariance

Typical Formula:

$$\mathrm{I} = n\sum_i^n\sum_j^n w_{ij}(y_i-\bar{y})(y_j-\bar{y})/\sum_i^n\sum_j^n w_{ij}\sum_i^n(y_i-\bar{y})^2, i \neq j \tag{2}$$

This is the well-known *Moran's I statistic*, explicated by Cliff and Ord (1981), and used widely as a measure of spatial association for spatial residuals from regression analysis. When the data used for testing purposes are georeferenced, the basic form of the linear regression model requires that the error term be spatially randomly distributed. Moran's *I* is used as a test for the randomness of residuals from regression. The elements of the **W** matrix are represented as w_{ij}. The **Y** matrix is the covariance and the remainder of the equation is a scaling device used to give the statistic useful testing properties.

5. Spatial Association as Difference

Formulas:

$$\gamma(h) = \sum_i^{n-h}\sum_{j-1+h}^{n} w_{ij}(y_i-y_j)^2/2n(h) \tag{3}$$

$$c = n-1\sum_i^n\sum_j^n w_{ij}(y_i-y_j)^2/2\sum_i^n\sum_j^n w_{ij}\sum_i^n(y_i-\bar{y})^2, i \neq j \tag{4}$$

The better known of these formulas is the first, called the *semivariance* or *variogram*. Widely used among geoscientists, this measure helps to identify the decline in spatial association as distance (h) increases away from the sites in question. It is the backbone for the large research area that yields continuous map surfaces common in studies of meteorology, geography, and geology (Cressie 1991). **W** is again represented by w_{ij}. **Y** is made up of $(y_i-y_j)^2$ and the remainder of the equation is used to scale the semivariance in terms of the variance. Equation 4 is known as Geary's c.

6. Spatial Association as Clusters

Formula:

$$K(d) = \left[A\sum_i^n\sum_j^n w_{ij}(d)/\pi n(n-1)\right]^{1/2} \tag{5}$$

This measure, called the *K-function*, helps to identify *spatial clustering* (Ripley 1981) within the region A. It is used primarily in ecological and epidemiological studies (Diggle 1983). The **W** matrix is made up of the w_{ij} elements, represented as ones, up to an arbitrary distance d. The **Y** matrix is invisible but is assumed to have each element equal to one. In practice, K values are found for a series of distances. Tests are usually carried out by simultating alternative patterns of points (see *Spatial Pattern, Analysis of*).

7. Focused Spatial Association

Formula:

$$G_i^*(d) = \sum_j^n w_{ij}(d)y_j/\sum_j^n y_j \tag{6}$$

For each type of measure shown above there is a counterpart (usually with some variation) that allows one to determine whether or not spatial association exists within the proximal region surrounding a georeferenced point. A popular measure of this type is the *Getis–Ord statistic*, G_i^*(d), which for the **Y** matrix each georeferenced point i is weighted (Getis and Ord, 1992). As distance increases from the site in question, values of G_i^* increase if clustering is in evidence, otherwise they decrease. Anselin's *Local Indicators of Spatial Association* or *LISA* statistics are designed for focusing Moran's and Geary's statistics (Anselin 1995).

8. Spatial Association as Interaction

The basic formula:

$$T_{ij} = y_i y_j / w_{ij}^{-\beta} \tag{7}$$

There are a number of interaction models; the one shown above is the simplest. These widely used models enable transportation and economic researchers to estimate the flow of automobiles, aircraft, migrants, goods and services, and so on, between places (Haynes and Fotheringham 1984). The **W** matrix usually represents a distance decay function while **Y** is usually a multiplicative interaction matrix.

All of these models have the Γ form. All can be used to test hypotheses about or to measure the extent of spatial association. Recent summaries on the measurement of spatial association can be found in Getis (1966, 1999) and Bailey and Gatrell (1995).

See also: Mathematical Models in Geography; Spatial Analysis in Geography; Spatial Interaction Models

Bibliography

Anselin L 1995 Local indicators of spatial association—LISA. *Geographical Analysis* **27**: 93–115

Bailey T C, Gatrell A C 1995 *Interactive Spatial Data Analysis*. Longman Scientific and Technical, Harlow, Essex, UK
Cliff A D, Ord J K 1981 *Spatial Processes: Models and Applications*. Pion Press, London
Cressie N A C 1991 *Statistics for Spatial Data*. Wiley, New York
Diggle P J 1983 *Statistical Analysis of Spatial Point Patterns*. Academic Press, London
Getis A 1991 Spatial interaction and spatial autocorrelation: a cross-product approach. *Environment and Planning A* **23**: 1269–77
Getis A 1966 Local spatial statistics. In: Longley P, Batty M (eds.) *Spatial Analysis: Modelling in a GIS Environment*, Geoinformation International, Cambridge, UK
Getis A 1999 Spatial statistics. In: Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds.) *Geographical Information Systems*, 2nd edn., Wiley, New York, Vol. 1
Getis A, Ord J K 1992 The analysis of spatial association by use of distance statistics. *Geographical Analysis* **24**: 189–206
Haynes K E, Fotheringham A S 1984 *Gravity and Spatial Interaction Models*. Sage, Beverly Hills, CA
Hubert L J, Golledge R G, Costanzo C M 1981 Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis* **13**: 224–33
Ripley B D 1981 *Spatial Statistics*. Wiley, New York

A. Getis

Spatial Autocorrelation

Spatial autocorrelation is the term used to describe the presence of systematic spatial variation in a variable. The variable can assume values either (a) at any point on a continuous surface (such as land use type or annual precipitation levels in a region); (b) at a set of fixed sites located within a region (such as prices at a set of retail outlets); or (c) across a set of areas that subdivide a region (such as the count or proportion of households with two or more cars in a set of Census tracts that divide an urban region). Positive spatial autocorrelation is the tendency for sites or areas that are close to one another to have similar values of the variable (i.e., both high or both low). Negative spatial autocorrelation is the tendency for adjacent values to be dissimilar (high values next to low values) which can arise in cases (b) and (c). The presence of spatial autocorrelation in map data has often been taken as indicative that there is something of interest in the map that calls for further investigation in order to understand the reasons behind the observed spatial variation (Cliff and Ord 1981). Whether identified formally through some statistic (see Sect. 3 below) or otherwise, if geographic phenomena display systematic spatial variability then this may provide the starting point for theorizing about the spatial outcomes of processes that may have their roots in, for example, economic, political, or sociological theory.

If spatial data display spatial autocorrelation, this has important implications for the application of statistical methods to such data. The term spatial autocorrelation arises in several different data analysis contexts. For example it arises in the fields of spatial statistics (Ripley 1981, Cressie 1991, Berliner), spatial econometrics (Anselin 1988) and geostatistics (Cressie 1991, Isaaks and Srivastava 1989, Berliner). If the values of a variable are known for a set of locations (or areas) then if there is spatial autocorrelation each value provides some information on the values of the same variable at nearby sites or for adjacent areas. One implication of this is that within the spatial data set there is information redundancy and for the purposes of hypothesis testing the size of the sample is not the same as the number of independent observations of the process. It is with the recognition of this property that statistical methods for analyzing spatial data part company with methods commonly employed in traditional statistics, sharing ground instead with the sorts of methods used to analyze time series data. The following sections will describe how spatial autocorrelation arises; how it is measured and its presence tested for; and the implications of spatial autocorrelation for spatial sampling, exploratory spatial data analysis, map interpolation, and confirmatory analysis including correlation and regression.

1. *How Does Spatial Autocorrelation Arise?*

Spatial autocorrelation can be a property of data that are conceptualized as either the outcome of a deterministic process or a stochastic (random) process. In the case of a deterministic process there is conceptually only one set of values (namely those that are observed) and it is the arrangement of values across the region that are defined as autocorrelated or not. The reference point for deciding whether such a distribution of values is autocorrelated is the geographic distribution of values that would have been obtained had those values been allocated at random across the region. In the case of a stochastic process, the observed values are but one set of realized values from an infinite set of possible realizations (a 'superpopulation') that can be generated by a spatial process or probability model. The variable may display spatial autocorrelation because it is a realization of a model where the random variables are spatially dependent. An example of such a model is the following, which is a particular case of the simultaneous autoregressive model (Whittle 1954). Let:

$$y(i) = \rho \sum_j w_{i,j} y(j) + e(i) \tag{1}$$

where $y(i)$ is the value of the variable for area i, ρ is a parameter, $w_{i,j}$ is non zero if area j ($j \neq i$) is a neighbor of area i (for example area j shares a common boundary with area i) and is zero otherwise, and $e(i)$ is independent and identically distributed and drawn

from a normal distribution with mean zero and variance σ^2. It can be seen from (1) that if $\rho > 0$ then the set of realized values of the variable Y will be positively spatially autocorrelated because the value at i will be a function of values in the neighboring areas. For other examples of these types of models see Besag 1974, 1975 and for reviews see Haining 1990, Cressie 1991, and Richardson 1992. There are other generators for spatially dependent variables and such models may derive from models that incorporate both space and time into their specification (see Berliner). In yet other contexts it is parameters (rather than the variables themselves) that are realizations of a spatially dependent probability model. This spatial dependency is transmitted to or inherited by the variables (Marshall 1991, Mollie 1996).

In the context of a stochastic process, the term spatial autocorrelation has a closely proscribed meaning and is associated with the second order moment (correlation) properties of the model. When thinking about actual problems or processes then spatial autocorrelation in a variable can arise in a number of different ways. Here are some examples:

(a) Many geographic phenomena display spatial continuity. If the spatial framework used to record the occurrence or the level of the phemonena is at a smaller scale than the scale at which the phenomena varies across the region, then values in adjacent areas of the spatial framework will tend to be similar. Remote sensing provides an illustration of this point. If land use remains the same over large tracts of land, and if a grid of small pixels records land use, then neighboring pixels will tend to record the same value of the variable. The pixel values are spatially autocorrelated.

(b) Even if the scale of the spatial framework used for recording data values is the same as the scale at which the phenomena varies across the landscape, measurement error can introduce spatial autocorrelation. This can occur with certain types of land use survey where recorders are given blocks of land to survey so that any operator-specific biases in the recording process will be similar within blocks and may therefore induce a form of spatial autocorrelation in the errors (Milne 1959). Spatial autocorrelation in errors also arises with remotely sensed data due to the nature of the recording device on the satellite. Due to light scattering, what is recorded by the sensor in one pixel, which supposedly refers to one parcel of land on the ground, is actually a function in part of terrain characteristics in neighboring parcels (Foster 1980).

(c) Variable values may show spatial autocorrelation because they depend on an underlying process in which spatial relationships affect the way the process unfolds. There are a number of examples. Diffusion processes are processes where some attribute, like an infectious disease, spreads through a fixed population and the probability of a new individual becoming infected depends on spatial proximity to an already infected individual (Cliff and Ord 1981). Spillover processes are processes where the attribute at some location is a function not only of conditions at that location but also of conditions in neighboring areas, for example, house prices as a function of local neighborhood conditions (Anas and Eum 1984), or city income levels as a function of income levels in neighboring cities and rural areas arising because of expenditure spillovers (Haining 1987). Interaction processes are where actions at one location influence actions at another location, such as price setting at one retail outlet influencing the prices set at other competitor sites (Haining 1983, 1984). Dispersal processes are where individuals themselves disperse (like seeds around a parent plant or new settlers colonizing a region) so that population counts are autocorrelated (Cliff and Ord 1981). In all these cases, at any point in time or when the process reaches some steady state then the variable may show evidence of spatial autocorrelation.

(d) The measured variable may not itself be the outcome of any of the foregoing situations but may be causally linked to another variable which is. In this case the spatial structure of the dependent variable inherits the spatial structure of the independent variable to which it is related. Spatial autocorrelation in a variable may also arise because of spatial autocorrelation in underlying parameters. The incidence of a non-infectious disease like a respiratory condition may be autocorrelated because the relative risk of the disease (which is a parameter and is itself a function of environmental characteristics such as air quality) is autocorrelated.

(e) Spatial autocorrelation can be a property of a derived set of values such as the residuals from a regression model. In this case spatial autocorrelation may have arisen because the regression model failed to include an important covariate in the model which had a spatially autocorrelated structure. The spatial structure of the omitted independent variable is then inherited by the regression residuals.

2. *Measuring and Testing for Spatial Autocorrelation*

2.1 *Autocorrelation Tests*

There are several statistics for measuring and testing for spatial autocorrelation. In the case of nominal data, the Join-Count test developed by Krishna Iyer (1949) is often used. The statistic is based on counting the number of joins of a particular color or colors on a map. Suppose each area is colored either black (B) or white (W). The Join-Count test is based on counting the number of BB, WW or BW adjacencies on the map. If a map is positively spatially autocorrelated then the number of BB and WW joins should be large and the number of BW joins small. If a map is

negatively spatially autocorrelated then the number of BW joins should be large and the number of BB and WW joins small (Haining 1990). The terms 'large' and 'small' are taken with reference to the number of such counts that would be expected if the black and white colors were allocated at random across the areas. This statistic can be generalized to handle maps where areas are classified into three or more colors (Dacey 1965).

In the case of ordinal and interval level data the two tests most commonly used are the Geary test and the Moran test (Geary 1954, Moran 1948). The Geary test is based on differences between the values of adjacent areas: $(y(i)-y(j))^2$ where $y(i)$ denotes the value observed in area i and $y(j)$ denotes the values observed in area j. If areas i and j are adjacent and if the map is positively autocorrelated then $(y(i)-y(j))^2$ should be small. The Geary statistic is based on summing these squared differences across the map:

$$\sum_i \sum_j \delta_{i,j}(y(i)-y(j))^2 \tag{2}$$

where $\delta_{i,j}$ is 1 if areas i and j are adjacent neighbors and zero otherwise ($\delta_{i,i}=0$). The measure (2) will be 'small' in the case of positive spatial autocorrelation relative to the case of a random distribution of values. On the other hand (2) will be 'large' in the case of negative spatial autocorrelation.

The Moran coefficient is based on cross products for adjacent areas:

$$\sum_i \sum_j \delta_{i,j}(y(i)-\bar{y})(y(j)-\bar{y}) \tag{3}$$

where $\bar{y}$ is the mean of the observations on Y. In the case of a positively autocorrelated map pattern, (3) will be positive since usually both terms in the cross product will be positive or negative. In the case of a negatively autocorrelated map pattern the terms will be negative, as if one of the terms is positive the other will usually be negative.

These statistics are appropriate to regular lattices but less appropriate for situations where the areal framework is irregular. Both (2) and (3) are invariant under certain alterations to the areal system but for any given area i some adjacencies are likely to be more important than others and should arguably therefore have greater weight in the calculation of any spatial autocorrelation statistic. In this case the binary weight ($\delta_{i,j}$) is replaced by a more general weight ($w_{i,j}$). The weights may be chosen to reflect some hypothesis about the degree of contact between areas (Cliff and Ord 1981). For example $w_{i,j}$ can be defined to reflect the length of the shared boundary between the two areas i and j, or the distance between the geographic centres of the two areas, or to reflect some measure of the intensity of interaction between the two areas in terms of, say, traffic flow.

The inference theories for the Geary and Moran statistics have been developed under two different conceptualizations of the null hypothesis of randomness or no spatial autocorrelation. In one case (the randomization hypothesis) the means and variances of the statistics are obtained assuming that the observed values are distributed at random across the region. In the other case (the normality assumption) the means and variances of the statistics are obtained assuming that the observed values are independent drawings from a normal probability distribution with the same mean and variance for each area. The assumption of equal variance is a strong assumption and often questioned in the case of irregular areal partitions where population counts vary substantially between areas and where the population count appears in the denominator of the variable (Waldhor 1996).

There are other techniques for exploring spatial autocorrelation. Perhaps he simplest technique is the scatterplot of paired data values $\{y(i), W^*y(i)\}$ where $y(i)$ is the value observed in area i and $W^*y(i)$ is the value in the ith position of the matrix product of W^* with y. W^* is the matrix of general weights $\{w_{i,j}\}$ where each row has been standardized to 1. So $W^*y(i)$ is a weighted average of the local values of y around i. If there is positive spatial autocorrelation present the scatterplot will have a distribution of points sloping upward from the bottom left of the graph. If there is negative spatial autocorrelation the scatterplot will have a distribution of points sloping downward from the top left of the graph. If the distribution of values is random the scatterplot will have no discernable trend. Cressie (1984) describes other graphical methods for inspecting for spatial structure.

2.2 Spatial Correlograms and Semi-variograms

The tests for spatial autocorrelation described above can be implemented at any number of spatial scales. That is, the neighbors of any area i need not be restricted to the adjacent neighbors but can be extended to include more remote neighbors. It may be informative to examine spatial structure at a range of distances simultaneously. Spatial correlograms (and semi-variograms) provide plots of spatial structure at a range of different distances. The correlogram at distance g has similarities with the Moran statistic. It is based on the expression:

$$\sum_i \sum_j w_{i,j}(g)(y(i)-\bar{y})(y(j)-\bar{y})$$

where $w_{i,j}(g)$ identifies all areas (j) that are a distance g from the area i. Distance g can be defined in terms of a metric or in terms of a shortest path on a graph that connects adjacent areas (Haining 1990). The spatial correlogram plot ($g = 1, 2, 3, \ldots, G$) thus provides a description of how spatial structure varies with distance.

The semi-variogram has similarities with the Geary statistic and is based on the expression:

$$\sum_i \sum_j w_{i,j}(g)(y(i)-y(j))^2$$

This, like the correlogram, provides a plot of spatial structure with distance thereby showing how spatial structure varies with distance (Isaaks and Srivastava 1989, Berliner). Both the correlogram and the semi-variogram may be used to suggest spatial models for the distribution of values across the map.

2.3 A Test for Local Autocorrelation

The statistics in 3.1 are often refered to as 'global' or 'whole map' tests of spatial autocorrelation because they compute an average measure for the entire map. It is recognized, however, that spatial autocorrelation might be a property of just a subset of the data. There might be a local tendency for similar values to cluster together but that this is not a general or 'global' property of the data.

A local test for spatial autocorrelation has been developed which computes the probability of getting the observed collection of values in a subset of the region if all the data values had been distributed at random over the area. The local Moran statistic is given by (Anselin 1995):

$$I(i) = (y(i)-\bar{y})\sum_j w_{i,j}(y(j)-\bar{y}) \quad i = 1, \ldots, n$$

where n is the number of areas so there are as many tests as there are cases. A high positive value of $I(i)$ signals a local spatial clustering of similar values, whilst a high negative value signals a local spatial clustering of dissimilar values.

3. Spatial Autocorrelation and Spatial Data Analysis

The presence of spatial autocorrelation in a variable has implications for a number of different areas of data collection and data analysis. These include spatial sampling, map interpolation, exploratory spatial data analysis, and confirmatory analysis including modeling. Each of these areas is briefly discussed in turn.

3.1 Spatial Sampling

Spatial sampling is the process of taking a subset of data drawn according to some specified rule and on the basis of this subset making inferences about the spatial population from which the data have been drawn. Consider the case where the aim is to estimate the proportion of an area covered by a particular type of land use. In this case we do not adopt a 'superpopulation' view of the problem but rather focus on the 'here and now' of the land use geography. A conventional estimator for the quantity is the sample mean:

$$\bar{y} = \frac{1}{n}\sum_i y(i)$$

where n is the size of the sample and $y(i)$ is 1 if sample site i is occupied by the particular land use, zero otherwise. The problem is to choose the most efficient sample design from random sampling, stratified random sampling, and systematic sampling. The spatial autocorrelation in the land use surface has an important influence on the relative efficiency of these three designs because, as noted above, the presence of spatial autocorrelation signals that data values recorded at one location carry information about data values at nearby locations. The stronger the spatial autocorrelation on the surface the more important it is to select sample sites that ensure that each sample point is not duplicating information contained in the rest of the sample. Random sampling is the least efficient sample design because there is no guarantee that sample points will not be clustered in particular areas of the map. Stratified random sampling overcomes the worst features of random sampling by ensuring coverage but sample sites can still be adjacent to one another on either side of a strata boundary. Systematic sampling is frequently the best because it ensures a minimum separation distance between sample sites (Dunn and Harrison 1993).

3.2 Map Interpolation

A range of estimators have been constructed to estimate values at specified locations where as yet data have not been collected. One of the more widely used methods, the kriging estimator, has the form:

$$\hat{y}(k) = \sum_j \lambda_{k,j} y(j)$$

where $\hat{y}(k)$ is the value to be estimated at location k where currently no data value is available (Berliner). The values $\{y(j)\}$ are known and $\{\lambda_{k,j}\}$ are coefficents that are a function of the spatial autocorrelation between locations k and j (Isaaks and Srivastava 1989). In this case the presence of spatial autocorrelation can be used to construct estimates of missing or unobserved values and in addition confidence intervals can be attached to the estimates. This confidence interval is influenced by how much information about the surface is available in the vicinity of the site. Confidence intervals will be large if the site k is far from sample points for which information is available, small if there are sample points with known values nearby.

3.3 Exploratory Spatial Data Analysis

Disease atlases are constructed to represent the geographic variability of a disease. Maps of standardized mortality rates (SMRs), for example, display the geography of relative risk for a particular disease and may be used to provide an indication of where the disease risk might be highest—perhaps as a preliminary to exploring for particular risk factors. Maps constructed using small geographic areas suffer from a number of problems including that the estimated rates tend to be unstable with the most extreme rates often associated with areas with the smallest populations. Bayes adjustment methods are often used to adjust such standardized rates. These methods incorporate information from the rest of the map in order to eliminate some of the random variation that is present in each area's SMR. A subset of these methods do not just use the SMRs from the rest of the map but place greatest emphasis on those estimates closest to the area to be adjusted. For this adjustment, spatial process models are used and incorporate the spatial autocorrelation in the relative risk rates into the adjustment process (Mollie 1996).

Other, simpler smoothing methods may also be employed to provide evidence on broad patterns in a spatial data set, for example, mean and median smoothers (Haining et al. 1998). These techniques all implicitly exploit some aspect of the concept of spatial autocorrelation—that places near to an area, i, provide information on the variable value at i. Similar methods have been used to 'clean' remotely sensed data where noise has blurred the image (Green 1996).

3.4 Information Loss in Statistical Tests

Spatial autocorrelation can undermine the use of classical statistical techniques. The correlation coefficient (r) is a common statistic for measuring the linear relationship between two variables (X and Y). The Pearson correlation coefficent varies between -1 and $+1$ with $+1$ signifying a perfect positive relationship between X and Y (as X increases, Y increases). The inference theory for the correlation coefficient is based on:

$$(n-2)^{1/2}|\hat{r}|(1-\hat{r}^2)^{-1/2} \tag{4}$$

where n is the sample size and $\hat{r}$ is the estimated correlation coefficient. Under the null hypothesis that the correlation in the population is zero (4) is t distributed with $(n-2)$ degrees of freedom. However the amount of statistical information carried by the n observations when the data are (positively) spatially autocorrelated is less than would be the case were the n observations to be independent. In the terminology of Clifford and Richardson (1985) the 'effective sample size' is less than the actual sample size n so the degrees of freedom for the test is less than n. The solution to the problem is to compute this effective sample size (N') which is obtained from the spatial correlogram (see Sect. 3.2). Then in (4) N' replaces n and it is t distributed with $(N'-2)$ degrees of freedom.

3.5 Modeling Spatial Variation

One of the statistical assumptions underlying the application of the classical regression model is that the errors are independent. If this assumption is violated tests of significance on the parameters of the independent variables are affected (the risk of a type I error is increased) and the coefficient of multiple determination measure of goodness of fit is inflated. This assumption has to be tested via the regression residuals (which are estimates of the errors). In the case of spatial data, an adaptation of the Moran test is often used although other tests are available (Anselin 1988, Cliff and Ord 1981). The inference theory of the Moran test has to be adapted in this context because the mean of the residuals is zero by construction. If the null hypothesis of no spatial autocorrelation is rejected then remedial action is needed to avoid mis-specification error. One approach is to seek out other independent variables that, because they were originally (and wrongly) excluded and because they were spatially autocorrelated, caused the residual autocorrelation. Other forms of model specification may also need to be considered depending on the analyst's conceptualization of the process underlying the behavior of the dependent variable. Thus there may be a need to include interaction terms, diffusion process terms or spillover terms (see Sect. 2). Thus a respecified regression model might include spatially lagged terms of the dependent variable, one or more of the independent variables or a spatial error model. (Examples of these and other models are described in Ord 1975, Anselin 1988, Haining 1990, Cressie 1991.)

Bibliography

Anas A, Eum S 1984 Hedonic analysis of a housing market in disequilibrium. *Journal of Urban Economics* **15**: 87–106

Anselin L 1988 *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Anselin L 1995 Local indicators of spatial association—LISA. *Geographical Analysis* **27**: 93–115

Besag J 1974 Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B* **36**: 192–236

Besag J 1975 Statistical analysis of non-lattice data. *The Statistician* **24**: 179–96

Cliff A, Ord J 1981 *Spatial Processes: Models and Applications*. Pion, London

Clifford P, Richardson S 1985 Testing the association between two spatial processes. *Statistics and Decisions* **2**(supplement): 155–60

Cressie N 1984 Towards resistant geostatistics. In: Verly G et al. (eds.) *Geostatistics for Natural Resources Characterization*. Reidel, Dordrecht, The Netherlands, pp. 21–44

Cressie N 1991 *Statistics for Spatial Data*. Wiley, New York

Dacey M 1965 A review of measures of contiguity for two and k-color maps. In: Berry B, Marble D (eds.) *Spatial Analysis: A Reader in Statistical Geography*. Englewood Cliffs, NJ, pp. 479–95
Dunn R, Harrison A 1993 Two dimensional systematic sampling of land use. *Applied Statistics* **42**: 585–601
Foster B 1980 Urban residential ground cover using Landsat digital data. *Photogrammetric Engineering and Remote Sensing* **46**: 547–58
Geary R 1954 The contiguity ratio and statistical mapping. *The Incorporated Statistician* **5**: 115–45
Green P 1996 MCMC in image analysis. In: Gilks G, Richardson S, Spiegelhalter D (eds.) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, pp. 381–99
Haining R 1983 Anatomy of a price war. *Nature* **304**: 679–80
Haining R 1984 Testing a spatial interacting-markets hypothesis. *Review of Economics and Statistics* **66**: 576–83
Haining R 1987 Small area aggregate income models: theory and methods with an application to urban and rural income data for Pennsylvania. *Regional Studies* **21**: 519–30
Haining R 1990 *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, UK
Haining R, Wise S, Ma J 1998 Exploratory spatial data analysis in a GIS environment. *The Statistician* **47**: 457–69
Isaaks E, Srivastava R 1989 *An Introduction to Applied Geostatistics*. Oxford University Press, Oxford, UK
Krishna Iyer P 1949 The first and second moments of some probability distributions arising from points on a lattice, and their application. *Biometrika* **36**: 135–41
Marshall R 1991 A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society, Series A* **154**: 421–41
Milne A 1959 The centric systematic area sample treated as a random sample. *Biometrics* **15**: 270–97
Mollie A 1996 Bayesian mapping of disease. In: Gilks G, Richardson S, Spiegelhalter D (eds.) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, pp. 359–79
Moran P 1948 The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B* **10**: 243–51
Ord J 1975 Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* **70**: 120–6
Richardson, S 1992 Statistical methods for geographical correlation studies. In: Elliott P, Cuzick J, English D, Stern R (eds.) *Geographical and Environmental Epidemiology: Methods for Small Area Studies*. Oxford University Press, Oxford, UK pp. 181–204
Ripley B 1981 *Spatial Statistics*. Wiley, New York
Whittle P 1954 On stationary processes in the plane. *Biometrika* **41**: 434–49
Waldhor T 1996 The spatial autocorrelation coefficient Moran's I under heteroscedasticity. *Statistics in Medicine* **15**: 887–92

R. P. Haining

Spatial Choice Models

The interrelationship between spatial structure and individual choice behavior has traditionally been at the forefront of geographic research. Geographers have realized that spatially varying consumer demand constitutes the basis for the evolution of spatial structures. In turn, the supply of facilities across space provides opportunities for individuals and households to conduct their activities, but at the same time also sets the space-time constraints that restrict behavior.

Over the years, many different attempts at modeling spatial choice behavior have been reported in the literature. These models predict the probability that an individual will choose a choice alternative as a function of its locational and non-locational attributes. In this article, the background, theoretical underpinnings, and basic methodology of these modeling approaches are examined. From the outset, it should be clear that limited available space prevents us from discussing modeling approaches in considerable detail.

1. Gravity Models

The modeling of spatial choice behavior gained considerable momentum when Wilson (1971) introduced a family of spatial interaction models, based on the principle of gravity, to describe and predict social phenomena. This principle states that the interaction between two bodies is proportional to their masses, and inversely proportional to the square of distance. Wilson elaborated this principle in a number of ways. First, he suggested that any distance decay function that reflects the notion that, *ceteris paribus*, the intensity of interaction decreases with increasing distance may be used to replace straight distance. Secondly, he argued that in spatial problems often the number of departures from particular origins and/or the number of arrivals at particular destinations is known, implying that so-called balancing factors have to be introduced in the model to satisfy these constraints.

Wilson distinguished an unconstrained spatial interaction model (departures and arrivals are not known), a production-constrained spatial interaction model (number of departures per zone is known), an attractions-constrained spatial interaction model (number of arrivals per zone is known), and the doubly constrained spatial interaction model (both number of departures and number of arrivals per zone are known). In the context of spatial choice behavior, the production constrained spatial interaction model is most interesting as it predicts which destination will be chosen. It can be expressed as:

$$p_{ij} = A_i O_i \mathbf{W}_j f(d_{ij}) \tag{1}$$

where, p_{ij} is the probability that an individual in zone i will choose destination j; A_i is a balancing factor, guaranteeing that the number of predicted interactions from zone i is equal to its number of observed interactions; O_i is the number of observed interactions from zone i; $\mathbf{W}_j$ is a vector of variables, representing

the attractiveness of destination j; $f(d_{ij})$ is some function of the distance between origin i and destination j.

Although much empirical work has been conducted on spatial interaction models since this seminal work, most of the relevant literature has been concerned with operational problems. Theoretical progress has been limited. For example, different measures of distance (straight-line distance, Euclidean distance, city-block metric, and travel time) and attractiveness have been suggested. Likewise, different functions for representing the distance decay effect have been proposed (power, exponential, and more complex functions). Finally, an exponent has been added to the specification of attractiveness term to allow for the fact that the larger destinations tend to have extra attraction beyond their size because of the benefits of economies of scale.

The estimation of spatial interaction models is typically based upon aggregate, zonal data of interaction flows (shopping, recreation, migration, traffic, etc). To that effect, the study area is delineated carefully to avoid large external flows, and divided into a number of smaller homogeneous zones. The spatial interaction model is calibrated using observations about interactions between these zones as input.

As a result, the parameters of the spatial interaction model are highly influenced by the geometry of the study area. This caused a search for improvements of the model, and for alternative modeling approaches. In this context, some authors suggested the development of models that estimate the effects of attraction variables endogenously. Another way of dealing with this problem was to incorporate some measure of spatial structure into a spatial interaction model (Fotheringham 1983). The simplest approach to the specification of this additional term is the use of a Hansen-type accessibility measure. This so-called competing destinations model may be expressed as:

$$p_{ij} = A_i O_i \mathbf{W}_j C_j^{\delta} d_{ij}^{-\beta} \tag{2}$$

where

$$A_i = \left[\sum_j \mathbf{W}_j C_j^{\delta} d_{ij}^{-\beta} \right]^{-1} \tag{3}$$

and

$$C_j = \sum_{k \neq j} \mathbf{W}_j d_{jk}^{-\gamma} \tag{4}$$

2. The Revealed Preference Model

The fact that the spatial interaction model is heavily influenced by the geometry of the study led Rushton (1969) to develop a preference-scaling model. The purpose of the model was to derive rules of spatial choice behavior.

To that effect, choice alternatives were classified into so-called locational types, which are a combination of an attractiveness category and a distance category. Spatial choice behavior is assumed to reflect a trade-off between attractiveness and distance separation. Pairwise choice data are then used to calculate the proportion of times a particular locational type is chosen, given that both are present. Based on this data, the locational types are positioned on a unidimensional preference scale using a nonmetric multidimensional scaling algorithm.

Over the years, this basic model has been elaborated in a number of important ways. First, Rushton (1974) showed how graphical methods, trend surface analysis, or conjoint analysis might be used to decompose the preference scale into the contributions of the two basic variables. Second, Girt (1976) suggested linking the preference function to overt behavior by relating distances on the preference scale to choice probabilities. Third, Timmermans (1979) suggested deriving the choice sets of individuals from data on information fields and consumer attitudes, and using these to scale the alternatives. Although Rushton's preference scaling model does have considerable appeal, it never received any major following. One of the reasons might be that the model is based on two explanatory variables only, while heterogeneity in preferences is not accounted for. The model has also been criticized because of its conceptual basis. It has been suggested that the model may not represent preferences at all, but rather inconsistencies in individual choice behavior.

3. Discrete Choice Models

The spatial interaction model has been criticized for its reliance on aggregate, zonal interaction flows. This criticism has led to the development of disaggregate discrete choice models that are based on individual choice behavior.

Discrete choice models may be derived from at least two formal theories: Luce's strict utility theory and Thurstone's random utility theory. In particular, Luce assumed that the probability of choosing a choice alternative is equal to the ratio of the utility associated with that alternative to the sum of the utilities for all the alternatives in the choice set. Luce thus assumed deterministic preference structures and postulated a constant-ratio decision rule.

In contrast, random utility theory is based on stochastic preferences in that an individual is assumed to draw a utility function at random on each choice occasion. An individual's utility for a choice alternative is assumed to consist of a deterministic component and a random utility component. Given the principle of utility maximizing behavior, the probability of choosing a choice alternative is then equal to the

probability that its utility exceeds that of all other choice alternatives in the choice set.

The specification of the model depends on the assumptions regarding the distributions of the random utility components. The best-known model is the multinomial logit (MNL) model. It can be derived by assuming that the random utility components are independently, identically Type I extreme value distributed. It can be expressed as:

$$p_{ik} = \frac{\exp[U(X_k, S_i)]}{\sum_j \exp[U(X_j, S_i)]} \tag{5}$$

where $U(X_k, S_i)$ is the deterministic part of the utility of choice alternative k of individual i with socioeconomic characteristics S_i.

One of the most important criticisms that have been expressed against the multinomial logit model concerned the fact that the utility of a choice alternative is independent of the existence and the attributes of other alternatives in the choice set (the Independence from Irrelevant Alternatives or IIA property). It predicts that the introduction of a new, similar choice alternative will reduce market shares in direct proportion to the utility of the existing alternatives, which is counterintuitive in the case of a high degree of similarity between particular alternatives.

Therefore, various alternative models have been developed to relax the IIA assumption (Timmermans and Golledge 1990). Similar work in this regard has been conducted by Williams (1977). The best-known model that circumvents the IIA-property is the nested logit model. Choice alternatives that are assumed to be correlated are grouped into the same nest. Each nest is represented by an aggregate alternative with a composite utility, consisting of the so-called inclusive value, and a parameter to be estimated. To be consistent with utility-maximizing behavior, the inclusive values should lie in the range between 0 and 1, and the values of the parameters should change consistently from lower levels to higher levels of the hierarchy.

4. *Conjoint Preference and Choice Models*

Unlike discrete choice models, the parameters of the conjoint preference and choice models are not derived from real-world data, but from experimental design data. Conjoint models involve the following steps when applied to spatial choice problems (e.g., Timmermans 1984).

First, each choice alternative is described by its position in a set of attributes, considered relevant for the problem at hand. Next, a series of levels for each attribute is chosen. Having defined the attributes and their levels, the next step involves creating an experimental design to generate a set of hypothetical choice alternatives. This design should be created such that the necessary and sufficient conditions to estimate the model of interest are satisfied. Respondents are then required to evaluate the resulting attribute profiles. Their responses are decomposed into the utility contribution of the attribute levels. Once part-worth utilities have been estimated, choice behavior can be predicted by assuming that an individual will choose the alternative with the highest overall evaluation. Alternatively, different probabilistic rules can be postulated.

To avoid such *ad hoc* rules, Louviere and Woodworth (1983) suggested estimating the preference function and choice model simultaneously. To that end, the attribute profiles should be placed into choice sets. Again, the composition of the choice sets depends on the properties of the choice model.

Over the years, the basic conjoint preference and choice models have been elaborated in many different ways. To avoid the problem of information overload and reduce respondent demand, Louviere (1984) suggested the Hierarchical Information Integration method. It assumed that individuals first group the attributes into higher-order decision constructs. Separate designs are constructed for each decision construct, and a bridging experiment is designed for measuring the tradeoff between the evaluations of the decision constructs. Oppewal et al. (1994) suggested an improved methodology by using multiple-choice experiments to test the implied hierarchical decision structure.

Other recent advances include the development of group preference models, the development of models of portfolio choice, the inclusion of contextual effects and constraints, and the combined use of experimental design and revealed preference data, to name a few.

5. *Complex Spatial Choice Behavior*

The previous models are all typically based on single-purpose, single-stop behavior. Over the years, however, it became clear that an increasing proportion of trips involved multistop behavior. Moreover, Hägerstrand's time geography had convincingly argued that behavior does not reflect preferences only, but also constraints. Thus, various attempts have been made to develop models of trip chaining and activity-travel patterns.

Originally, most models relied on semi-Markov process models or Monte Carlo simulations. More recently, utility-maximizing models have dominated the field. An important contribution in this regard was made by Kitamura (1984), who introduced the concept of prospective utility. It states that the utility of a destination is not only a function of its inherent attributes and the distance to that destination, but also

of the utility of continuing the trip from that destination. Dellaert et al. (1998) generalized Kitamura's approach to account for multipurpose aspects of the trip chain.

Trip chaining is only one aspect of multiday activity/travel patterns. Several models have recently been suggested to predict more comprehensive activity patterns. These models can be differentiated into the older constraints-based models (e.g., Lenntorp 1976), utility-maximizing models (e.g., Recker et al. 1986), rule-based or computational process models (e.g., Golledge et al. 1994, ALBATROSS – Arentze and Timmermans 2000).

6. *Conclusions*

This article has given a very brief overview of spatial choice modeling. Due to the limited available space, we have focused on some of the key modeling approaches, disregarding alternatives (e.g., dynamic models) altogether.

If we examine the accomplishments, one cannot escape the conclusion that the models developed over the years have been increasing in complexity. The theoretical underpinnings of the recent activity-based models are much richer than the Markov or gravity models of the 1960s. Moreover, the brief overview indicates that geographers have been using alternative data, such as experimental design data and activity diaries, in addition to the more conventional survey data.

See also: Spatial Analysis in Geography; Spatial Decision Support Systems; Spatial Interaction; Spatial Interaction Models; Spatial Optimization Models; Spatial Pattern, Analysis of; Spatial Statistical Methods; Spatial Thinking in the Social Sciences, History of

Bibliography

Arentze T A, Timmermans H J P 2000 *ALBATROSS: a Learning-based Transportation-oriented Simulation System*. EIRASS, Eindhoven University of Technology, The Netherlands

Dellaert B, Arentze T A, Borgers A W J, Timmermans H J P, Bierlaire M 1998 A conjoint based multi-purpose multi-stop model of consumer shopping centre choice. *Journal of Marketing Research* **35**: 177–88

Ettema D F, Timmermans H J P 1997 *Activity-based Approaches to Travel Analysis*. Elsevier Science, Oxford, UK

Fotheringham A S 1983 A new set of spatial interaction models: The theory of competing destinations. *Environment and Planning, A* **15**: 1121–32

Girt J L 1976 Some extensions to Rushton's spatial preference scaling model. *Geographical Analysis* **8**: 137–56

Golledge R G, Kwan M P, Gärling T 1994 Computational process modeling of travel decisions using a geographical information system. *Papers in Regional Science* **73**: 99–117

Kitamura R 1984 Incorporating trip chaining into analysis of destination choice. *Transportation Research* **18B**: 67–81

Lenntorp B 1976 Paths in space-time environments. *Lund Studies in Geography, Series B*, no. 44

Louviere J J 1984 Hierarchical information integration: a new method for the design and analysis of complex multiattribute judgment problems. In: Kinnear T C (ed.) *Advances in Consumer Research* **11**, ACRA, Provo, UT, pp. 148–55

Louviere J J, Woodworth G 1983 Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research* **20**: 350–67

Oppewal H, Louviere J J, Timmermans H J P 1994 Modeling hierarchical information integration processes with integrated conjoint choice experiments. *Journal of Marketing Research* **31**: 92–105

Recker W W, McNally M G, Root G S 1986 A model of complex travel behavior: part I – theoretical development. *Transportation Research, A* **20A**: 307–318

Rushton G 1969 Analysis of spatial behavior by revealed space preference. *Annals of the Association of American Geographers* **59**: 391–400

Rushton G 1974 Decomposition of space preference structures. In: Golledge R G, Rushton G (eds.) *Spatial Choice and Spatial Behavior*. Ohio State University Press, Columbus, OH, pp. 119–33

Timmermans H J P 1979 A spatial preference model of regional shopping behaviour. *Tijdschrift voor Economische en Sociale Geografie* **70**: 45–8

Timmermans H J P 1984 Decompositional multiattribute preference models in spatial choice analysis: a review of some recent developments. *Progress in Human Geography* **8**: 189–221

Timmermans H J P, Golledge R G 1990 Applications of behavioural research on spatial choice problems II: preference and choice. *Progress in Human Geography* **14**: 311–54

Williams H C W L 1977 On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning, A* **9**: 285–34

Wilson A G 1971 A family of spatial interaction models and associated developments. *Environment and Planning, A* **3**: 1–32

H. Timmermans

Spatial Cognition

Spatial cognition concerns the study of knowledge and beliefs about spatial properties of objects and events in the world. Cognition is about knowledge: its acquisition, storage and retrieval, manipulation, and use by humans, nonhuman animals, and intelligent machines. Broadly construed, cognitive systems include sensation and perception, thinking, imagery, memory, learning, language, reasoning, and problem-solving. In humans, cognitive structures and processes are part of the mind, which emerges from a brain and nervous system inside of a body that exists in a social and physical world. Spatial properties include location, size, distance, direction, separation and connection, shape, pattern, and movement.

1. The Geographic Study of Spatial Cognition

Geographers study spatial cognition for three reasons. First and foremost, spatial cognition provides questions of geographic interest in their own right, given that cognition about space and place is an expression of human-environment or human-earth relations. Second, geographers have hoped that spatial cognition would improve their understanding of traditional phenomena of interest; for example, where people choose to shop should depend in part on their beliefs about distances and road connections. Finally, there has been an interest in improving the use and design of maps and other geographic information products which depend in part on human understanding of depicted spatial relations.

In addition to geographers and cartographers, researchers from many different disciplines study spatial cognition, including psychologists, architects and planners, anthropologists, linguists, biologists, philosophers, and computer scientists. While overlapping and interacting with these disciplines, the geographer's interest in spatial cognition is distinct in some ways. Geographers are interested in the earth as the home of humanity. Their concern is with the cognition of humans, not the cognition of barn swallows or underseas robots. Geographers focus on spatial scales relevant to human activities on the earth's surface, such as temporary travel, migration, settlement, and economic activities; a geographer wants to understand human conceptualizations of the layout of a city or a region, not of an atom or the solar system. Finally, the geographer ultimately desires to describe, predict, and explain external human-environment relationships, and focuses on internal mental and neuropsychological structures and processes only insofar as it helps in the understanding of external relationships.

1.1 History of Spatial Cognition in Geography

The notion that a subjective understanding of space partially mediates human relationships with an objective physical world is undoubtedly quite old. Within the modern discipline of geography, the study of spatial cognition first appeared with work on geographic education early in the twentieth century. However, interest in spatial cognition flowered during the 1950s and 1960s. The subdiscipline of experimental cartography emerged with the insight that maps could be improved by understanding how humans perceive and think about information depicted in map symbols. In an attempt to explain human responses to floods and other natural hazards, the subdiscipline of environmental perception began to study people's subjective beliefs about the occurrence and consequences of hazard events. Most importantly for the study of spatial cognition, behavioral geographers began developing theories and models of the human reasoning and decision-making involved in spatial behavior e.g., migration, vacationing, and trip scheduling. This approach gained insight from the work of Kevin Lynch (1960), a planner who argued that 'images' of cities guide people's behavior and experiences of those cities. All of these historical developments constituted a new focus on disaggregate or microscale approaches within human geography, approaches that focused on the geographic behavior of individuals as an accompaniment to traditional approaches that focused on human settlements and larger aggregations of people.

2. Topics in the Geographic Study of Spatial Cognition

The study of spatial cognition in geography sheds light on questions such as how spatial knowledge and beliefs are acquired and develop over time; the nature of spatial knowledge structures and processes; how people navigate and stay oriented in space; how people use language to communicate with each other about space; and how aspects of spatial knowledge and reasoning are similar or different among individuals or groups.

2.1 Acquisition and Development

Humans acquire spatial knowledge and beliefs directly via sensorimotor systems that operate as they move about the world. People also acquire spatial knowledge indirectly via static and dynamic symbolic media such as maps and images, 3-D models, and language. Geographers are interested in how different media have consequences for the nature of acquired knowledge.

Spatial knowledge changes over time, through processes of learning and development. Both changes in the child's spatial knowledge and reasoning, and that of an adult visiting a new place for the first time, are of interest to geographers. Both physical maturation and experience influence development. A person's activity space; the set of paths, places, and regions traveled on a regular basis, is an important example of spatial experience that influences people's knowledge of space and place, no matter what their age. Most people know the areas around their homes or work places most thoroughly, for example (as in the anchor-point theory of Couclelis et al. 1987).

A widely accepted model of spatial learning suggests that spatial knowledge of places develops in a sequence of three stages or elements. First is landmark knowledge, unique features or views that identify a place. Second is route knowledge, based on travel routines that connect ordered sequences of landmarks. The final element is survey knowledge, knowledge of two-dimensional layouts that includes the simultaneous

interrelations of locations; survey knowledge would support detouring and shortcutting. Not everyone has detailed or comprehensive survey knowledge of their surrounds, and exposure to maps of places clearly advances the development of such knowledge.

2.2 Spatial Knowledge Structures and Processes

The metaphor of a cognitive map was coined by Tolman in a 1948 paper to refer to internally represented spatial models of the environment. The cognitive (or mental) map includes knowledge of landmarks, route connections, and distance and direction relations; nonspatial attributes and emotional associations are stored as well. However, in many ways, the cognitive map is not like a cartographic 'map in the head.' It is not a unitary integrated representation, but consists of stored discrete pieces including landmarks, route segments, and regions. The separate pieces are partially linked or associated frequently so as to represent hierarchies such as the location of a place inside of a larger region. Furthermore, spatial knowledge is not well modeled by metric geometries such as the Euclidean geometry of high school math. The ways in which spatial knowledge is internally represented in human minds lead to certain patterns of distortions in the way people answer spatial questions. For example, people often believe the distance from place A to B is different than from B to A. Turns are frequently adjusted in memory to be more like straight lines or right angles. At larger scales, most people have somewhat distorted ideas about the sizes and locations of continental land masses on the earth; for example, South America is thought to be due south of North America when it is actually rather far to the east as well.

2.3 Navigation and Orientation

Humans travel over the earth's surface to reach destinations. This typically requires planning and the ability to stay oriented while moving. Navigation is this coordinated and goal-directed travel through space. It consists of two components, locomotion and wayfinding. Locomotion refers to the guidance of oneself through space in response to local sensorimotor information in the immediate surrounds, and includes such tasks as identifying surfaces of support, avoiding obstacles, and moving toward visible landmarks. Locomotion generally occurs without the need for an internal model or cognitive map of the environment. Wayfinding refers to the planning and decision-making that allows one to reach a destination that is not in the immediate sensory field, and includes such tasks as choosing efficient routes, scheduling destination sequences, orientating to nonlocal features, and interpreting verbal route directions. Wayfinding tasks generally do involve a cognitive map of the environment. Various locomotion and wayfinding tasks vary greatly in their demand on attentional capacity. Following a familiar route is not very demanding on attention; finding one's way in a new city may be quite demanding.

Orientation is knowing 'where you are,' although the precision and comprehensiveness of this knowing varies greatly in different situations and for different people. Two fundamental processes are involved in orientation during navigation. The first process is the recognition of external features or landmarks–pilotage. In some cases, the recognized landmark is the goal, but more commonly the landmark aids orientation by serving as a key to knowledge of spatial relations stored in an internal cognitive or external cartographic map. The second process is dead reckoning, updating a sense of orientation by integrating information about movement speed, direction, and/or acceleration without reference to recognized features.

People also use symbolic media such as maps to navigate and stay oriented. Maps used for navigation show robust alignment effects, a common confusion in using maps when the top of the map is not the direction in the surrounds that one is facing when looking at the map. It results from the fact that spatial perception is 'orientation-dependent'; objects and pictures have a top and a bottom in the perceptual field. In the case of road or trail maps, many people turn the map to bring about alignment. Unfortunately, this option is not available with ubiquitous 'You-Are-Here' maps, which are surprisingly often placed in a misaligned way with respect to the viewer's perspective. The resulting alignment effect is reflected by the fact that misaligned maps are interpreted more slowly and with greater error by most people (Levine et al. 1984). While an annoyance in many situations, misaligned emergency maps placed in public buildings are potentially disastrous.

2.4 Spatial Communication via Language

Spatial information is often communicated verbally, in natural language such as English. People give and receive verbal route directions, read spatial descriptions contained in stories, and increasingly interact with computer systems via verbal queries. There are two notable characteristics of the way language expresses spatial information. One is that language expresses mostly nonquantitative or imprecise ('fuzzy') quantitative information about space. Statements about connections and approximate location are more important than precise statements. For example, people say 'turn left at the gas station' rather than 'turn 80° after you go 0.6 miles.' Great precision is typically unnecessary or even confusing. A second characteristic is that various aspects of the context of

communication are critical in interpreting spatial language. Context is provided by knowledge of who is speaking, where they are, physical features in the situation, the previous topic of conversation, and so on. 'The bicycle is near the church' is generally understood differently than 'Wisconsin is near Michigan.'

2.5 Individual and Group Similarities and Differences

No two individuals know exactly the same things or reason in exactly the same way about space and place. Some people are better at tasks such as wayfinding, learning spatial layouts, or reading maps. In some cases, there may be different ways to think about such problems, all of which may be effective. In such cases, one might speak of 'stylistic differences' in spatial cognition rather than 'ability differences.' Geographers are interested in measuring and explaining such individual differences. Traditional psychometric paper-and-pencil tests of spatial abilities tend not to measure these differences very well (Allen et al. 1996).

There are many factors that may contribute to variations in spatial cognition: body size, age, education, expertise, sex, social status, language, culture, and more. A first goal for geographic research is to measure and document ways these factors might covary with spatial cognition. Females and males appear to differ in some ways in their spatial abilities and styles, for example, but not in others (Montello et al. 1999). In addition to describing such covariations, geographers have the research goal of distinguishing their underlying causes. For example, the two sexes likely differ in their spatial cognition in part because of different travel experiences afforded them by their socialized roles. Explanations for such covariations are generally quite difficult to determine, however, as one cannot readily do randomized experiments on person characteristics such as age, sex, culture, and activity preferences.

3. Technologies and the Future of Spatial Cognition Research

A variety of technologies are poised to have a large impact on the questions and methods of spatial cognition research in geography. Global positioning system (GPS) receivers have recently become small and inexpensive, and are increasingly used by sailors, hikers, and other outdoor enthusiasts. A GPS receiver provides information about one's location by geometrically combining distance signals it picks up from several geosynchronous satellites constantly orbiting the earth. This system is also part of the technology of automated navigation systems built into automobiles to provide drivers up-to-date and locally relevant information about routes and destinations. Several research issues concern how locational information is best displayed or communicated, and how the availability of such information might change people's experiences and behaviors in space and place.

Other computer technologies will lead to new research and thinking in spatial cognition. There are a host of spatial cognition questions inspired by efforts to improve the effectiveness and efficiency of geographic information systems (GISs). How can complex geographical information be depicted to promote comprehension and effective decision-making, whether through maps, graphs, verbal descriptions, or animations? How does exposure to new geographic information technologies alter human ways of perceiving and thinking about the world? There are also research issues concerning the way that people do or do not spatialize their understanding of other types of information networks such as the World Wide Web; one often speaks metaphorically of 'navigating' the Web. However, undoubtedly the most dramatic technological development in spatial cognition will be the advent of computer-simulated worlds known as virtual environments or virtual reality.

See also: Cartography; Cognitive Maps; Geographic Education; Geographic Learning in Children; Knowledge (Explicit and Implicit): Philosophical Aspects; Space: Linguistic Expression; Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms; Spatial Thinking in the Social Sciences, History of

Bibliography

Allen G L, Kirasic K C, Dobson S H, Long R G, Beck S 1996 Predicting environmental learning from spatial abilities: An indirect route. *Intelligence* **22**: 327–55

Couclelis H, Golledge R G, Gale N, Tobler W 1987 Exploring the anchor-point hypothesis of spatial cognition. *Journal of Environmental Psychology* **7**: 99–122

Downs R M, Stea D (eds.) 1973 *Image and Environment*. Aldine, Chicago

Golledge R G 1987 Environmental cognition. In: Stokols D, Altman I (eds.) *Handbook of Environmental Psychology*, Wiley, New York

Levine M, Marchon I, Hanley G 1984 The placement and misplacement of you-are-here maps. *Environment & Behavior* **16**: 139–57

Liben L S, Patterson A H, Newcombe N 1981 *Spatial Representation and Behavior Across the Life Span*. Academic Press, New York

Lloyd R 1997 *Spatial Cognition: Geographic Environments*. Kluwer Academic, Dordrecht, The Netherlands

Lynch K 1960 *The Image of the City*. MIT, Cambridge, MA

Mark D M, Frank A U (eds.) 1991 *Cognitive and Linguistic Aspects of Geographic Space*. Kluwer Academic, Dordrecht, The Netherlands

Montello D R, Lovelace K L, Golledge R G, Self C M 1999 Sex-related differences and similarities in geographic and

environmental spatial abilities. Annals of the Association of American Geographers **89**(3): 515–34
Tolman E C 1948 Cognitive maps in rats and men. *Psychological Review* **55**: 189–208
Tversky B 1992 Distortions in cognitive maps. *Geoforum* **23**: 131–8

D. R. Montello

Spatial Data

Spatial data are the sum of our interpretations of geographic phenomena. Taken literally, spatial data could refer to any piece of information that associates an object with a location—from stars in the heavens to tumors in the human body. Typically, however, when used in a geographical context, spatial data is restricted to describing phenomena on or near the earth's surface. In digital form, the data are the primary information needed by geographic information systems (GIS), the software tools used for spatial-data analysis.

1. *What Are Spatial Data?*

'Spatial data' is used as an all-encompassing term that includes general-purpose maps, remotely sensed images, and census-tract descriptions, as well as more specialized data sets such as seismic profiles, distribution of relics in an archeological site, or migration statistics. Information about buildings, bridges, roads, streams, grassland, and counties are other examples of types of spatial data. Broadly, spatial data refers to any piece of information that has an absolute or relative location (see *Location: Absolute/Relative*). More commonly, spatial data are associated with information shown on maps and images of natural and man-made features that range in size from 1 m to 10 km. Such data are often referred to as geospatial data (see *Geographic Information Systems*; *Spatial Data Infrastructure*; *Remote Sensing*; *Landsat Imagery in Geography*; *Remote Sensing and Geographic Information Systems Analysis*; *Global Positioning System* for examples of the varying types and usages of spatial data). Spatial data sets are collections of locational and nonlocational information about selected geographic phenomena. This collection of information forms the heart of a spatial database. Once the information is in the database, computers provide the capabilities to handle geographic data in ways that were previously impossible.

2. *Data Models*

A digital spatial data set represents a certain model of geographic reality. These models describe both the locational and nonlocational aspects of the phenomena of interest. For users to make sense of the data they are receiving, they must understand the conceptual model underlying the data.

Table 1
Terminology for describing geographic features

Feature type	Example	Digital database term (and synonyms)
Point feature	Benchmark	Node (vertex, 0-cell)
Linear feature	Road centerline	Arc (edge, 1-cell)
Areal feature	Lake/pond	Polygon (face, 2-cell)

The locational aspect of a spatial data set is usually described by one of two types of models, either grid-cell (sometimes called raster) data models, or vector data models. Grid-cell data models use collections of fixed units of space, for example, cells of land 50 m on a side, to describe the location and extent of a geographic phenomenon. Digital remotely sensed images are collected in this form. Vector data models use graphic elements (points, lines, and polygons) to describe the location and extent of the phenomenon under study. Most digital map data are in this form.

The spatial data also must include all of the nonlocational information used to describe the features in the database. In the case of digital-image data, the amount of nonlocational data is relatively small, usually a set of values measuring characteristics such as the greenness of vegetation. With vector data the amount of nonlocational, or attribute, data can be quite large, even exceeding the amount of spatial data. This includes information about the various features included in the database (for example, the number of lanes of a highway) and their relationships to each other (for example, that the railroad tracks lie between the stream and the highway). In the context of GIS, features are the sum of our interpretations of geographic objects or phenomena. Buildings, bridges, roads, streams, grassland, and counties are examples of features.

2.1 *Feature Data Structures*

The real world of geographical variation is infinitely complex and often uncertain, but must be represented digitally in a discrete, deterministic manner. In one method of describing geographic reality, characterized by the information shown on maps, we think of the world as a space populated by features of various kinds—points, lines, and areas. This is commonly referred to as the vector or feature data model. Features have attributes that serve to distinguish them from each other. Any point in the space may be empty, or occupied by one or more feature. These features may exist in one combined database, or may be separated according to a theme or variable into a number of layers or classes. The terminology used to describe these features is given in Table 1.

The relationships between the nodes, arcs, and polygons are called topological relationships. In simple terms, the topological relationships are that arcs start and end only at nodes, and that polygons are enclosed by a series of arcs. Attention is given to neighboring geographical entities when coding arcs. For instance, the topological description of each arc includes not only the starting and ending bounding nodes but also the left and right designations of the co-bounding regions. In this way an arc is directed, that is, the first coordinate is that of the start node and the last coordinate that of the end node. Each polygon record might also include the identification of any interior islands or, if the polygon is itself an interior island, the identification of the single exterior polygon. This data structure not only accommodates the natural pattern of mapped features efficiently, but also provides the information needed to facilitate automated editing and manipulation of the database.

2.2 Raster Data Structures

Instead of describing discrete geographic objects, geographical variation can be represented by recording the locational pattern of one variable over the study area. When the spatial interval of this sampling is regular (e.g., rectangular), the resulting matrix of observations is called a raster data structure. If there are n variables, then n separate items of information are available for each and every point in the area. Digital terrain models data and digital images are typically stored in this structure. Other types of information such as categories of land-cover could also be encoded in a raster data structure. The cells of raster data (particularly raster data of images) are commonly referred to as pixels (derived from the term picture element).

2.3 Modeling Solids

So far our discussions have dealt with data on a surface. However, certain types of spatial data require a more complete modeling of the third dimension. Solid bodies, such as geologic structures (ore bodies, petroleum reservoirs, etc.), need to be encoded not only with x, y locations, but also with z or depth information. The surface models mentioned earlier can encode z values, but only one z value is permitted for each location. This results in what is termed a 2.5-D model. True solid or 3-D modeling requires that multiple z values can exist at a given x, y location.

One mechanism to model such data is to extend the raster data model into an additional dimension. Thus the 2-D squares become 3-D cubes. The cubes are often termed voxels (for volume elements). Each voxel is encoded with attribute data (such as rock type). The geolocation of any given voxel is implied by its location in the x,y,z array and the size of the cubes. The voxel representation allows us to model such things as hollow spheres, with a varying thickness of the shell, or solids with holes through them.

Another way to model solids is to use a technique called boundary representation. This is an extension of the vector data model to 3-D. A solid object is defined by its bounding surface. Just as one-dimensional lines bound two-dimensional areas, 2-D planar areas are located in 3-D space and are used to bound a 3-D solid. The various planar areas bound solids, much like the facets bound the shape of a jewel.

3. Describing Geographic Features

As noted earlier, spatial data models, in addition to providing locational data, must handle all of the nonlocational information used to describe the features in a database. For users to understand the nature of the data there must be a complete specification of the features being described. A comprehensive data specification includes the definition of the features as well as definitions of their attributes and attribute values. Other information required for a complete understanding includes delineation guidelines, data extraction rules, and representation rules. Understanding how the locational component of a feature is represented is of little use unless the user understands the nature of what is being described. What is really meant by terms such as 'forest,' 'marshlands,' 'residential land,' or 'bridge'? For example, one could declare a 'bridge' to be a feature. A bridge is one of a number of built-up structures on or near the surface of the earth. Its definition may be further refined to a structure erected to carry traffic over a depression or obstacle. Thus, the feature 'bridge' is an element of a set of phenomena (an 'erected structure'), with the common attributes of function ('to carry traffic') and location. It also has the common relationship of spanning another feature ('over a depression or obstacle'). The definitions and specifications of the data producers hopefully match to some degree the expectations of the user.

To help solve this data-definition problem a feature data specification can be used to provide a detailed definition of each feature and how it is modeled. This technique is used by national mapping agencies in the USA, France, and Germany. The feature data specifications define the domain of features that may be represented in a spatial database. For each feature in the feature data specification, a domain of attributes is defined. Each attribute is assigned a value, taken from the domain of values specified by the attribute definition authority. In addition to the definitions, guidance is given on how a given feature is delineated, when it is collected, and how it is represented in the spatial data model.

Specific relationships, termed semantic relationships, are used to express interactions that occur between features that are otherwise difficult to rep-

resent either with locational data, topological relationships, or feature descriptions. Such relationships can include the following: to flow, the relationship that occurs among the two-dimensional features composing a river system; to connect, the relationship that occurs (at junctions) among the two-dimensional transportation features composing a road network; to bound or to be bounded, the relationship that occurs between the perimeter of a city and its city limits; to be composed of, the relationship that occurs between a designated route of travel and the transportation features that compose it, or of the stream and pond features that comprise a watercourse; and to be above, the relationship that occurs between the features in an overpass or underpass of a freeway interchange. The intent of all of these relationships is to model the interactions between features sufficiently to support GIS applications and automated cartographic-product-generation requirements.

4. *Modeling Time*

Just as maps change over time and must be revised, so do spatial databases. There are two basic ways to handle the temporal dimension associated with spatial data. The simplest way is to maintain multiple versions of a database, akin to keeping a historical map file. When changes occur a new edition of the database is produced, released with the current data, and the previous edition is placed in a historical archive.

A more comprehensive, yet more complex, strategy is to maintain one database, where every database feature (and potentially its attributes) has time information associated with it. This is fairly straightforward for information in a raster form. For each cell, a variable-length list of attributes is maintained, noting any change in attribution and the time at which the change occurred. For feature-based vector data sets, the process is more complicated. Information is encoded into the database on when the feature was first incorporated into the database, if it presently exists, when it ceased to exist (e.g., a pond was filled in), and when it was recreated, perhaps in a different form (e.g., a road was rerouted). In this fashion the existence and modification of any given feature can be traced.

5. *Data Quality*

The utility of spatial data in GIS depends directly on the quality of the spatial data used in the process. Geospatial data quality is measured along at least three different axes: space (location, places), time, and theme (content, characteristics). The space axis characterizes positional accuracy; the time axis indicates temporal accuracy and completeness; and the theme axis reflects semantic accuracy, attribute accuracy, and completeness. These qualities are not independent, however. Indeed some data-quality aspects, such as logical consistency, directly refer to the interrelationships of these data-quality elements. Not all of these quality elements have well-defined quality metrics associated with them. Metrics of semantic accuracy, for example, are not well understood. And some of the measures used for other quality elements may be inadequate.

Also needed are linkages between the quality information present in the data and the GIS-processing algorithms, so that software can produce error estimates on the end results. The results of GIS analyses have often been presented as definitive to decision makers (e.g., this is the best site for the municipal water well), with little or no qualification. One would expect that decision makers would insist on qualifying statements on any results presented to them. The methods for doing this are not well understood. For example, due to the cumulative effects of error, it has been conjectured that as the number of data sets used in a GIS analysis increases, the accuracy of the result decreases. This runs counter to common practice, where multiple data sets are used with little regard for whether they are contributing more to the information content or to errors in the result.

6. *Metadata*

As technology provides faster and more efficient ways to transmit and process spatial data, data producers and users are realizing the potential value of exchanging and sharing spatial data. The ability to use existing spatial data is important to organizations that are trying to share data and to increase benefits from their data collections. This ability depends not only on being able to find data, but also on understanding the characteristics and quality of the data. Metadata, 'data about data,' provide information such as the characteristics of a data set, the history of a data set, and organizations to contact to obtain a copy of a data set. Standardized metadata elements provide a means to document data sets within an organization, to contribute to catalogs of data that help persons find and use existing data, and to help users understand the contents of data sets that they receive from others. Metadata support activities include determining existing data, determining the appropriate data set for a particular use, providing information to access data sets, and providing information to transfer data sets for processing and use.

While the benefits of metadata are manifest, several impediments hinder uniform adoption and use. These center on two topics: agreement on what metadata elements to collect, and the level of effort required to collect whatever metadata is required. Obviously the two items are linked. From the user perspective, detailed metadata is highly desirable. More information will help the user determine the suitability of a given data set. From the data producer's perspective,

metadata costs money; metadata elements should be constrained to a cost-effective set. Significantly more resources may be required to collect metadata about 'legacy' data holdings than data sets collected using techniques that capture the metadata as a part of the entire data-collection process.

Agreement on standardized terminology and definitions would not only provide benefits in the automation of the collection of metadata, but would also be exceedingly valuable in searching data catalogues. Several metadata 'standards' are in widespread use around the world. In general they have data elements that have been selected to provide information about four topics:

Availability—data needed to determine the sets of data that exist for a geographic location;

Fitness for use—data needed to determine if a set of data meets a specified need;

Access—data needed to acquire an identified set of data; and

Transfer—data needed to process and use a set of data.

These elements form a continuum where a user cascades through a pyramid of choices when determining what data are available, evaluating the fitness of data for use, accessing the data, and transferring and receiving the data. The exact nature of these data elements continues to be the subject of much debate and discussion.

A collection of metadata records combined with data management and search tools forms a data catalogue. The architecture of these data catalogues can vary from tightly held internal data indexes to widely accessible distributed catalogs spread across the Internet. Establishing a data catalog can be a major project. Among the items that need to be accomplished are to:

(a) determine which programs or individuals produce, manage, or disseminate data that are to be included in the catalog;

(b) identify what geospatial data an organization holds, where data are located, and their condition;

(c) identify priorities for documenting previously collected or produced data based on demand, geographic extent, uniqueness, or other criteria to be determined by the organization;

(d) create the metadata for new spatial data;

(e) ensure that the metadata comply with applicable standards, to the extent possible for the organization;

(f) provide training and technical assistance to metadata producers, quality-assurance personnel, and end users;

(g) update and maintain metadata as necessary;

(h) specify, acquire, and manage the hardware, software, and telecommunications services so that the metadata are electronically accessible;

(i) determine how to link or cross-reference spatial data and metadata to ensure consistency, maintainability, easy query or search, and easy access; and

(j) respond to questions about or orders for spatial data.

Clearly the catalog function can go far beyond the basic listing of content. Extensions that allow for direct ordering or download of data, user comment or feedback, as well as requests for new data collections are all possible. Metadata and data catalogs are the means by which data users can more economically find, share, and use spatial data.

Bibliography

Guptill S C, Morrison J L (eds.) 1997 *Elements of Spatial Data Quality*. Pergamon, Oxford, UK

Langran G 1992 *Time in Geographic Information Systems*. Taylor and Francis, Philadelphia, PA

Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds.) 1999 *Geographical Information Systems*, 2nd edn. John Wiley, New York

Robinson A H, Morrison J L, Muehrcke P C, Kimlerling A J, Guptill S C 1995 *Elements of Cartography*, 6th edn. John Wiley, New York

S. C. Guptill

Spatial Data Infrastructure

1. The Context

The nature of Geographical Information Systems (GIS) is that the basic principles are simple, sometimes apparently even trivial—but the rate of change in all aspects of the field is great and good practice is sometimes fiercely demanding intellectually and in other ways. This article reflects that situation; the continuously evolving characteristics of GIS and Spatial Data Infrastructure make this a snapshot of the situation in the early part of the twenty-first century (see *Geographic Information Systems*).

The history of GIS has been summarized by Coppock and Rhind 1991) and Foresman (1997). What began as a hugely expensive—and thus rare—mainframe computer tool for inventory and mapping purposes in the mid-1960s blossomed in the 1990s. The statistics suggest that there are now about two million users of GIS worldwide and this figure is growing at about 20 percent per annum. Many of the active users of GIS are carrying out tasks on behalf of large numbers of citizens, notably through the work of governments. In addition, there are many everyday users of services that include GIS elements, sometimes without this being obvious (e.g., in many Internet searches).

The annual commercial revenues generated from sales of GIS software are now over $1 billion worl-

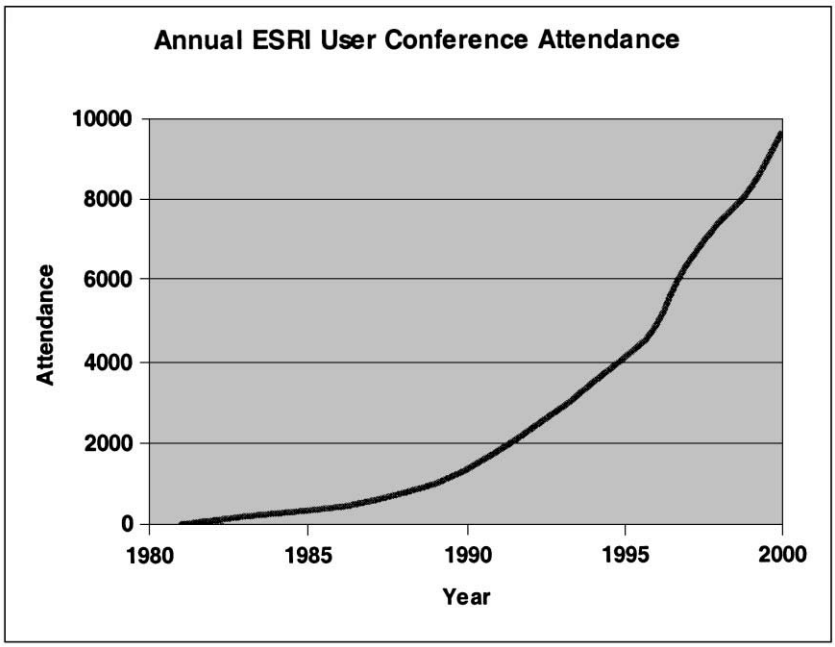

Figure 1
Attendance at the annual ESRI User conference since 1981—a surrogate for interest in GIS as a whole.
Source: Environmental Systems Research Institute

dwide. Typically this leverages a ratio of about 1:5 in terms of expenditures on hardware, staff training, data collection, and operations. Thus the total annual GIS expenditure is now likely—insofar as it can be estimated—to be of the order of $15 to $20 billion. This figure is supported by estimates of expenditure by known users e.g., cadastral and mapping agencies. The great bulk of the software revenues are earned by a small number of US-based firms but other GIS expenditures are typically more local. One surrogate for the growth of GIS is the number of people attending conferences: attendances at the most popular GIS software conference have grown from 23 in 1981 to over 10,000 in 2001 (Fig. 1). All this is not surprising for the range of applications of GIS is now truly extraordinarily huge (Longley et al. 2001). GIS has thus moved in 30 years from an esoteric side issue, based on the creation of software by academics and government, to a fully fledged global business and scientific endeavor.

2. *The Problem*

That said, views of what GIS is and why it is significant expanded greatly in the late 1990s. Issues such as the reliability of results from GIS-based analyses of multiple data sets, the impact of the technology on privacy, debates on public policies on access to information and its ownership, plus the international interactions of national policies, all involved and influenced thinking on GIS. The result was to re-engineer GIS as a science and as a political, managerial, and business issue as well as a technological one. This has a series of consequences, such as calling into question the existing, surprisingly homogeneous, GIS educational infrastructure existing in over 2,000 universities worldwide (Longley et al. 2001). Paradoxically, it has also led to a recognition that the 'island-like state' of GIS in respects other than education needs to be addressed: to date, most applications and uses of GIS have been project- based.

If GIS is to become totally ubiquitous—a reasonable aim given the ubiquity of use of geographical information in most decision making—then a number of major problems have to be overcome. These have been summarized by the US Federal Geographic Data Committee (1997) as follows:

> In the United States, geographic data collection is a multi-billion-dollar business. In many cases, however, data are duplicated. For a given piece of geography, such as a state or a watershed, there may be many organizations and individuals collecting the same data. Networked telecommunications technologies, in theory, permit data to be shared, but sharing data is difficult. Data created for one application may not be easily translated into another application. The problems are not just technical—institutions are not accustomed to working together. The best data may be collected

> on the local level, but they are unavailable to state and federal government planners. State governments and federal agencies may not be willing to share data with one another or with local governments. If sharing data among organizations were easier, millions could be saved annually, and governments and businesses could become more efficient and effective.
> Public access to data is also a concern. Many government agencies have public access mandates. Private companies and some state and local governments see public access as a way to generate a revenue stream or to recover the costs of data collection. While geographic data have been successfully provided to the public through the Internet, current approaches suffer from invisibility. In an ocean of unrelated and poorly organized digital flotsam, the occasional site offering valuable geographic data to the public cannot easily be found. Once found, digital data may be incomplete or incompatible, but the user may not know this because many data sets are poorly documented. The lack of metadata or information on the "who, what, when, where, why, and how" of databases inhibits one's ability to find and use data, and consequently, makes data sharing among organizations harder … If finding and sharing geographic data were easier and more widespread, the economic benefits to the nation could be enormous.

This statement actually understates the scale of the problems encountered in typical GIS projects or programs. These normally involve dealing with the assembly of data from multiple sources and coping with scarce staff skills, 'state of the art' technology, and frequently contested outcomes. Making effective and safe use of data encoded by different people to different levels of resolution and accuracy, collected at different times and without certification of quality, is self-evidently non-trivial. These are some of the problems that Spatial Data Infrastructures are designed to resolve.

3. *What is a Spatial Data Infrastructure?*

It seems (Table 1) that Spatial Data Infrastructures (SDIs) are already commonplace. Harris (1997) and Masser (1998) have described various policy aspects relating to Geographic Information and to the early stages of SDIs but the situation is constantly evolving. The use of terminology, however, conceals many different entities and states of development. In Australia the local manifestation is termed 'the Australian Spatial Data Structure;' in Canada 'the Canadian Geospatial Data Infrastructure;' in the Asia-Pacific group there is 'the Permanent Committee for GIS Infrastructure in Asia and the Pacific;' and in the UK it is the 'National Geospatial Data Infrastructure.' SDIs are not necessarily national: a number of US states have set one up and a global version is under discussion.

Table 1
Countries which have assembled or are assembling National Spatial Data Infrastructures as of mid-2000. Source: Henry Tom, Oracle Corporation

Argentina	Hungary	Northern Ireland
Australia	India	Norway
Canada	Indonesia	Pakistan
Colombia	Japan	Poland
Cyprus	Kiribati	Russian Federation
Finland	Macau	South Africa
France	Malaysia	Sweden
Germany	Mexico	United Kingdom
Greece	Netherlands	USA
	New Zealand	

The most commonly used terms in this arena are 'spatial data,' 'geospatial data,' and 'geographic(al) information.' The phraseologies can be regarded as synonymous in some uses but not in others (see, for instance, Longley et al.'s (2001) discussions of data, information, evidence, and knowledge). Broadly speaking, 'spatial data' is the most widely used term except in Europe where 'Geographic Information (GI)' is widely used.

Formal definitions of SDIs vary significantly around the world. Two definitions of the US National Spatial Data Infrastructure (NSDI) are shown below, together with that of the putative Global Spatial Data Infrastructure (GSDI):

> [The National Spatial Data Infrastructure is] 'the comprehensive and coordinated environment for the production, management, dissemination, and use of geospatial data. [It subsumes] the totality of the policies, technology, institutions, data and individuals that were producing and using geospatial data or GI within the United States.' Source: 1993 and 1994 reports of the Mapping Science Committee (MSC) of the United States National Research Council
> [NSDI is] 'the technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilization of geospatial data.' Source: Presidential Executive Order #12906: 1994
> The GSDI is the broad policy, organizational, technical and financial arrangements necessary to support global access to geographic information.' Source: Steering Committee of the Global Spatial Data Infrastructure (see http://www.gsdi.org)

The differences in these are subtle but important. For instance, the financial underpinnings of NSDIs were largely ignored in the early years, as were the commercial and private sector contributions. Many originally saw the entire NSDI enterprise being driven by new technologies, especially Geographical Information Systems; others saw the key issues as the information production process. The progressive acceptance of Geographical Information Science has now broadened the generally accepted concept of NSDIs to include a stronger research and education agenda. In the USA in particular, the NSDI was originally driven by federal government agencies and the Federal Geographic Data Committee in particular; this has led to much debate on how to create leadership in an enterprise that can or should not—according to Internet enthusiasts—be led.

4. The US Experience with an SDI

The President's Executive Order and the FGDC identified three primary areas to promote development of the NSDI. The first was the development of standards, the second improvement of access to and sharing of data by developing a National Geospatial Data Clearinghouse, and the third was the development of the National Digital Geospatial Data Framework. All of these efforts were to be carried out through partnerships among federal, state, and local agencies, the private and academic sectors, and non-profit organizations. Executive Orders in the USA are only enforceable in regard to federal agencies: such cross-sectoral adherence can only be achieved by influence and the leverage through federal funding.

The Federal Geographic Data Committee operates as a bureaucracy through a series of subcommittees based on different themes of geospatial data (e.g., soils, transportation, cadastral), each chaired by a different federal agency. Figure 2 illustrates the original subcommittees and which federal agency led each one. Several working groups have also been formed to address issues on which there is a desire among agencies to coordinate and which cross subcommittee interests (e.g., Clearinghouse, Standards, Natural Resource Inventories).

4.1 Standards

Many of these groups have developed standards for data collection and content, classifications, data presentation, and data management to facilitate data sharing. For example, the Standards Working Group developed the metadata standard, which was formally adopted by the FGDC in mid-1994 and has since been adapted in the light of international and other national developments. All of the FGDC-developed standards undergo an extensive public review process. This includes nationally advertised comment and testing phases plus solicitation of comments from state and

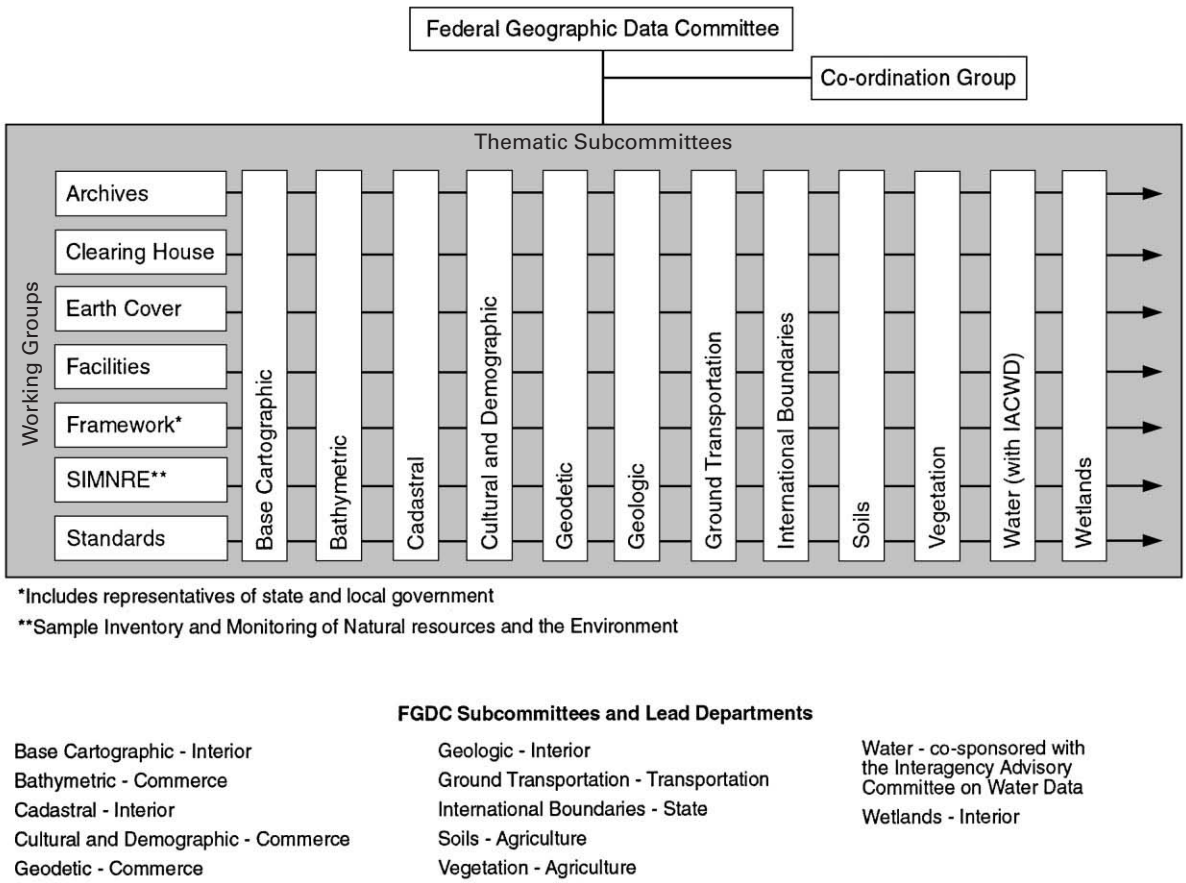

Figure 2
The original structure of the Federal Geographic Data Committee sub-committees and lead federal agency responsibilities for them.

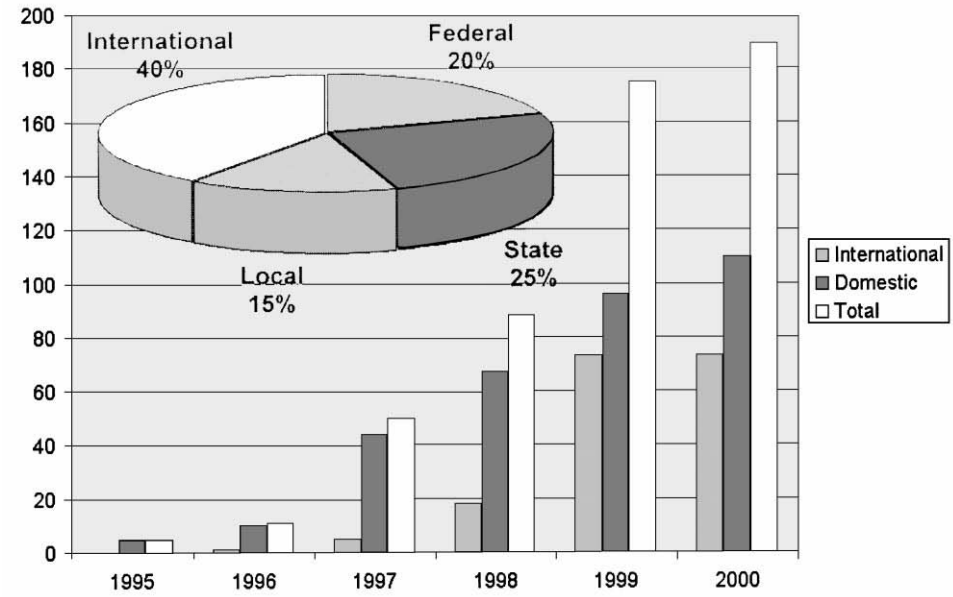

Figure 3
Growth in use of the US NSDI Clearinghouse. Source: Henry Tom, Oracle Corporation

local government agencies, private sector firms, and professional societies. The most recent NSDI standards at the time of writing (late 2000) are for utility data content, hydrography, geological maps, and a profile (i.e., a subset) of the shoreline metadata standard. The Executive Order mandated that all federal agencies must use all FGDC-adopted standards.

4.2 The Clearinghouse

The second activity area is intended to facilitate access to data, minimize duplication, and assist partnerships for data production where common needs exist. This has been done by helping to advertise the availability of data through development of a National Geospatial Data Clearinghouse. The strategy is that agencies producing Geographical Information describe its existence with metadata. They then serve those metadata on the Internet in such a way that they can be accessed by commonly used Internet search and query tools. The FGDC-adopted metadata standard describes the content and characteristics of geospatial data sets. President Clinton's Executive Order stipulated that agencies make these metadata accessible electronically. This has been rendered practicable by the development of metadata-creating software within GIS by the larger vendors. As a result of all this, nearly all federal agencies—as well as most States and numerous local jurisdictions—have become active users of the Internet for disseminating geospatial data. This model does not necessarily assume that GI will be distributed for free. Obtaining some of the data sets requires the payment of a fee; others are free. The Clearinghouse can also be used to help find partners for database development by advertising interest in or needs for GI. The growth in use of the Clearinghouse is shown in Fig. 3.

4.3 The Geographical Framework

The third NSDI activity area is the conceptualization and development of a US digital geographical (or geospatial) framework data set. Framework data are those which underpin all GIS operations, providing the geographical context and structure within which spatially manifested processes operate and on which other ('area coverage') data are assembled. There are two main reasons why a spatial framework is essential in almost all GIS operations. The framework provides both spatial context and spatial structure. In dealing with a satellite image, for example, the incorporation of place names facilitates linkage to the real world; moreover, the framework may be used to rectify distortions in, or subset and reorientate, the imagery. The second reason for the importance of the framework is that geographic information is typically collected by many different agencies, usually within the boundaries of national territories; and most of these agencies have historically collected data primarily to suit their own local purposes. Thus, if the benefits of added value are to be obtained from data

integration within a GIS, there must be a mechanism (or spatial 'key') to register the different data sets together in their correct relative positions and, beyond this, to ensure that the component objects within each data set are properly related. In general, only the geographical framework can provide this ubiquitous data linkage key through the use of coordinates or implicit geographical coding.

What constitutes the framework varies in different countries. In the USA, the Federal Geographic Data Committee (FGDC 1995a) has defined the information content of the US framework as consisting of:

- geodetic control
- digital orthoimagery
- elevation data
- transportation
- hydrography
- the geography of governmental units; and public land cadastral information.

Similar definitions—including place names—would in practice occur in many countries. Most typically, the framework is manifested in user terms through topographic mapping—the 'topographic template'-—which may contain all this information but partly in implicit form (e.g., geodetic control).

Hence the US framework was intended to form the foundation for the collection of other data, minimize data redundancy, and facilitate the integration and use of geospatial data. It should be pointed out that the USA is generally less well endowed in relation to a consistent geographical framework than are many other nations. The federal nature and physical extent of the country has ensured a multilevel and inconsistent hierarchy of geographical information is produced by the federal, state, and local governments, with local updating and tailoring by agencies and private sector bodies. This contrasts with Britain and many other countries where a single, consistent national framework has long been in existence and is consistently maintained up to date. The disadvantages of a multiplicity of frameworks are evident: the creation of multiple but different framework data sets will guarantee the creation of artifacts in any computer analysis using combinations of data sets related to each of the frameworks. Putting a value on such externalities is difficult, especially since the costs are visited on those at the end of supply chains and are often only discovered long after the data have led to mistaken conclusions.

Given the huge number of players involved, getting agreement on a multipurpose framework has been difficult in the USA. The Executive Order directed the FGDC to develop a plan for completing initial implementation of the framework by the year 2000 but this has proved impossible. A popular, partial solution has been to create one-meter resolution geometrically correct orthophotomaps in place of a conventional digital map framework. Attractions of this approach are that it could be automated to a great degree and could also be financed jointly by states and the federal government and implemented via the private sector.

The collaboration to create or refine the national NSDI has led to various tensions. These have occurred between federal agencies, between them and the states and local governments, and—to a more limited extent because of their more modest involvement until the late 1990s—with the private sector. These tensions arise partly because the NSDI permits the questioning of traditional, legacy roles for government bodies: for instance, the concept of assembling national databases by sewing together detailed ones held by states directly undermines the traditional role of some federal agencies. In addition, the NSDI has made public some latent demands which have ramifications for budget holders: in many cases the beneficiaries of spending are not those called on to divert resources from other activities.

Most recently, a US national report on NSDI pointed out the consequences of seeking a private/public sector partnership in developing the National Spatial Data Infrastructure. These consequences include an obvious requirement for the private sector to generate profits from trading in information and in financing infrastructure. These would necessitate, it was argued, a more cohesive and industry-friendly policy on information ownership and massive investment by government to convert its data into more coherent, consistent, and object-oriented form—which could then be marketed. Partnerships of this kind—as opposed to normal contractual arrangements with the private sector working for the government—are difficult to make work.

5. *The Global Spatial Data Infrastructure*

Most GI has been collected, assembled, and used hitherto in a national context. Yet there is a growing number of requirements for pan-national, even global data. For instance, Agenda 21 was designed to facilitate sustainable development. The manifesto was agreed in Rio in 1992 at the 'Earth Summit' and was supported by some 175 of the governments of the world. It is noteworthy that eight chapters of the Agenda 21 plan dealt with the need to provide geographic information. In particular, Chapter 40 aimed at decreasing the gap in availability, quality, standardization, and accessibility of data between nations. This was reinforced by the Special Session of the United Nations General Assembly on the Implementation of Agenda 21 held in June 1997. The report of this session includes specific mention of the need for global mapping, stressing the importance of public access to information and international cooperation in making it available.

The Global Spatial Data Infrastructure (GSDI) is an attempt to enhance the availability and accessibility

of global GI. It is the result of a voluntary coming together of national mapping agencies and a variety of other individuals and bodies. The organization has sought to define and facilitate creation of a GSDI by learning from national experiences. As part of this process, it funded a scoping study carried out by Australian consultants. They outlined in early 2000 their conceptual view of how the inputs to decision-making were linked to an SDI and the potential benefits that would arise from having one. They pointed out however that proof of the reality of the predicted benefits was difficult to find. The fundamental question for any GSDI business case is: what exactly is the product that governments and other agencies will be asked to fund? The consultants said bluntly that 'The broad definitions of GSDI are not sufficiently operational (or practical) to build a clear business case around.' They went on to raise some key questions which they recognized as being difficult to answer but had to be tackled to give credibility to a formal business case for GSDI. These included:

- Is GSDI a real global SDI or a federation of national or regional SDI?
- Is GSDI really about encouraging developing countries (or at least those that do not currently place emphasis on geospatial data to do so)? In this case, the product may be a suite of policy recommendations and technical assistance approaches that development agencies can apply in their client countries.
- Is GSDI a process, a general framework, or is it a particular product such as a world map or a comprehensive database?
- Who are the stakeholders?
- What about the demand side of GSDI? How do you isolate demand which arises from decision-making processes?

Given all this—and not surprisingly in view of the lack of any global government and the complications in building a national, let alone pan-national NSDI—progress on a GSDI has been slow to date. The primary reason is not inadequate technology but the requirement to clarify real need, obtain sources of finance, and secure political support.

6. *Spatial Data Infrastructures and the Social Sciences*

The diversity of multiple strands in the social sciences revealed by this encyclopedia complicate any brief attempt to link the book's subject matter to GIS and SDIs. This relationship is, however, manifestly in existence. Examples of factors which underpin the multiple and sometimes subtle relationships include:

- The increasingly 'taken for granted' nature of the role of GIS in supporting tactical (and sometimes strategic) decision making worldwide
- The growth of problem-oriented approaches within government as well as in business. These necessitate
  crossing traditional disciplinary boundaries and using information drawn from many sources e.g., relating to the natural environment as well as the social sciences—with all the ownership, epistemological, and other problems that entails
- The methodological implications of conflating information on people, places, institutions, and policies which have different conceptual bases and other characteristics
- Policy conflicts between different nations or other governed units such as US states (e.g., on the 'proper' role of the state as in the free distribution or charging for information)
- Institutional and interinstitutional politics and financing struggles within government (still the largest sources of Geographical Information). This can be encapsulated in the 'who leads' question which is difficult to answer given the permeation of SDI and GIS through many activities and responsibilities in different sectors.
- The interrelationship of the state and business. This is more complex than in many other cases for governments are both providers (in some case competitive with the private sector) and consumers of Geographic Information. The supply of some GI is a natural monopoly. There is a geography to this business/government relationship which is much more complex than national information policies would suggest (Longley et al. 2001).
- The legal issues involved in SDIs and GIS relate to liability, privacy (e.g. Curry 1999 and Longley et al. 2001), and 'fair dealing' as well as information ownership and protection. Again, there is a geography of the law related to SDIs.

In summary, SDIs are posited on worthy aims. But they raise fundamental theoretical and practical questions of the role of the state and of intersectoral and organizational partnerships on a scale hitherto unknown in many social sciences. Success in creating SDIs may both benefit and disadvantage the individual (e.g., in providing better and cheaper access to information but also in undermining privacy). It is also not clear how to assess success in an environment where competition and partnerships within governments and between them and the private sector are both inevitable. SDIs also raise a series of practical managerial questions. Ultimately, they impact on the lives and work of everyone. As a consequence of all of the above, SDIs provide a fertile laboratory for the social sciences.

See also: Geographic Information Systems: Critical Approaches; Geography; Remote Sensing; Remote Sensing and Geographic Information Systems Analysis; Spatial Analysis in Geography; Spatial Data

Bibliography

Coppock J T, Rhind D W 1991 The history of GIS. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems*. Longman Scientific and Technical, Harlow, UK pp. 21–43

Curry M R 1999 Rethinking privacy in a geocoded world. In: Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds.) *Geographical Information Systems: Principles, Techniques, Applications and Management*, 2nd edn. Wiley, New York pp. 757–66

Executive Order # 12906 1994 *Co-ordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure*. Washington DC

FGDC 1995a *Development of a National Digital Geospatial Data Framework*. Federal Geographic Data Committee, Washington, DC

FGDC 1995b *The National Geospatial Data Clearinghouse*. Federal Geographic Data Committee, Washington, DC

FGDC 1997 *A Strategy for the National Spatial Data Infrastructure*. Federal Geographic Data Committee, Washington, DC

Foresman T W 1997 (ed.) *The History of GIS*. Prentice-Hall, Upper Saddle River, NJ

Harris R 1997 *Earth Observation Data Policy*. Wiley, Chichester, UK

Longley P A, Goodchild M F, Maguire D J, Rhind D W 2001 *Geographical Information Systems and Science: A Student Companion*. Wiley, New York

Masser I 1998 *Governments and Geographic Information*. Taylor & Francis, London

National Research Council (US) Mapping Science Committee 1993 *Toward a Co-ordinated Spatial Data Infrastructure for the Nation*. Mapping Science Committee, National Research Council, National Academy Press, Washington DC

National Research Council (US) Mapping Science Committee 1994 *Promoting the National Spatial Data Infrastructure Through Partnerships*. Mapping Science Committee, National Research Council, National Academy Press, Washington DC

National Research Council (US) Mapping Science Committee 1997 *The Future of Spatial Data and Society*. Mapping Science Committee, National Research Council, National Academy Press, Washington DC

Rhind D W 1999 National and international geospatial data policies. In: Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds.) *Geographical Information Systems: Principles Techniques Management and Applications*. Wiley, New York pp. 767–87

D. Rhind

Spatial Decision Support Systems

People frequently make choices that involve evaluating geographic position or geographic context. Whether selecting routes, making service areas, or defining environmental hazardous areas, they use 'multiple spatial criteria' in making such choices and apply them to complex geographic data. In social science, such complex decision-making problems are often called 'ill-structured' and a set of theories, methods, and techniques has been developed for a wide variety of problem-solving contexts. It has become common to link these together in computer software that operates on spatial data. They are called spatial decision support systems and are increasingly being used in both business and public sectors today.

1. Spatial Decision Support for Ill-structured Problems

A number of problems may occur when making decisions for ill-structured problems. First, it may not be clear to the decision maker how much of the attainment of one spatial criterion should be sacrificed in order to attain a given amount of another. The decision maker often wants to explore alternative solutions and to reach a decision about the best solution only after extensive exploration of the spatial and the attribute solution space. That is, their preferences for a particular solution may be influenced by what can be attained in a particular context. Although they may have preferences for abstract entities, the implementation of these preferences is mediated by knowing what can be attained in the given context. Decision support systems are designed to facilitate the exploration of these spaces and thus to assist decision makers in their search for the best solutions to their problems.

A second problem is that the spatial databases containing relevant information may not be complete either with respect to all the variables that are relevant or to the accuracy with which they are measured. During the process of searching for a solution, decision-makers will sometimes want to add their unique spatial knowledge about an area or decide to search for additional facts about a location or area before making a decision.

A third problem is that though decision makers may want to view the results of models of processes, they may choose not to accept these results over concerns about their validity for the given situation. All three problems lead many decision makers to ask for computer-based, interactive, support systems with which to explore their problem and search for a solution.

1.1 Solution Spaces in SDSS

The solutions to all SDSS problems exist in both geographic and attribute space. As the geographic space of the solution to any problem changes, so its attribute space changes too. For example, in making districts from which voters will elect representatives, the geographic solution space is the map of the boundaries of the constituencies and the areas they contain. The corresponding attribute space describes the numbers of eligible voters in each district, their

party affiliations, their past voting record, and any attribute that a politician may view as relating to their likely voting behavior on a range of relevant issues in an upcoming election. Such attributes might be their ethnicity, income, age structure, religious preferences, etc. Thus, in an exercise to re-draw electoral districts following a decennial census, any participant in the process may want to see how any changes in the boundaries are accompanied by changes in any or all of these attributes. A system for simultaneously computing alternative regionalizations of territory would constitute a spatial decision support system for electoral redistricting. In this example, 'regionalizing' is the spatial function for which the support system is developed. Defined characteristics of the areas and the constraints for their definition specified by a particular political system would be the attributes for which the SDSS would be developed.

2. *Types of SDSS*

Different types of SDSS can be defined according to how geographic and attribute spaces are generated. In some SDSS, these spaces are generated interactively by the user. The capabilities of the SDSS then consist of the tools for manipulating the geographic information and reporting the characteristic attributes of the outcomes. Consequently, the simplest SDSS incorporate the capacity to convey information from the geographic to the attribute space or from the attribute space to the geographic space. These are key capabilities of geographic information systems (GIS).

GIS are simple spatial decision support systems. They permit users to query locations and learn about attributes at the locations. They also permit the definition of attributes—often following database queries such as sort or find—and then to see their geographic locations quickly. Good examples can be found in Birkin et al. 1996. Crossland et al. (1995) determined the efficacy of using GIS as SDSS for making location decisions. In an experiment, the students with access to GIS made better and quicker decisions with the SDSS than those without it.

The term SDSS is often reserved for systems that combine the functions of GIS with multicriteria decision analysis and optimizing models. These are developed for problem contexts that involve not only the manipulation of geographic information in GIS, but also the generation of preferences, the modeling of choices, and the search for locations or areas that best meet a user's defined needs. For such problems, GIS need to be integrated with spatial models. Integration is accomplished by visualization of related geographic and attribute spaces and by linking the analysis systems so that users can move easily between them using a unified user-interface (Densham 1994).

All SDSS link together a set of fundamental elements described below (see Fig. 1).

2.1 *Preference, Choice and Multicriteria Decision Models*

Many spatial decisions involve correct anticipation of how people will make selections from alternative locations when presented with any given set of alternatives. Subjective analysis of maps of existing or hypothetical future alternatives does not lead to accurate predictions of how many people will select one alternative over another. For this purpose spatial interaction or spatial choice models have been developed. When added to GIS with appropriate measures of the variables that affect choice, these models can provide invaluable assistance to people charged with making decisions about future facility locations, for example. These 'choice models' (Timmermans et al. 1984) examine the social and economic characteristics of small areas, and then, based on results from samples of the population, estimate how many of each population group will go to one facility and how many to another. When repeated for each small area, (Goodchild 1984) the sum of all the estimates is the aggregate estimate of the use of each facility.

In some cases, instead of modeling the choices of individuals, the user of the SDSS directly participates in the choice evaluation process. Jankowski and Ewart (1996), for example, used the spatial query capability of GIS to select plausible locations for rural health practitioners within an area. Their SDSS asked users to compare pairs of locations on given criteria, and then, using a multicriteria evaluation technique, to rank order the locations for this purpose. The GIS displayed the locations of the highest ranked alternatives.

2.2 *Location Optimizing Algorithms*

Many decision-makers want to select locations to optimize some defined outcome. A large literature describes models for finding optimal locations, routes, etc. SDSS often integrate such models with GIS so that, from a common user interface, the user can assemble the spatial data, solve the problem using a model selected from the interface, and then display the result. Ghosh and Rushton (1987) describe the evolution of location-allocation models from tools to find optimal locations for well-defined objectives to tools that can be incorporated in flexible analysis systems. In such systems, decision makers explore alternative location patterns and use optimizing algorithms to compute alternatives that they can evaluate against criteria of interest, some of which are highly qualitative in character.

Other SDSS are designed to select geographic routes for the delivery of goods or services (Keenan 1998). These are often used in a GIS environment supported by inputs from mobile units that are monitored by global positioning systems (GPS) that report their

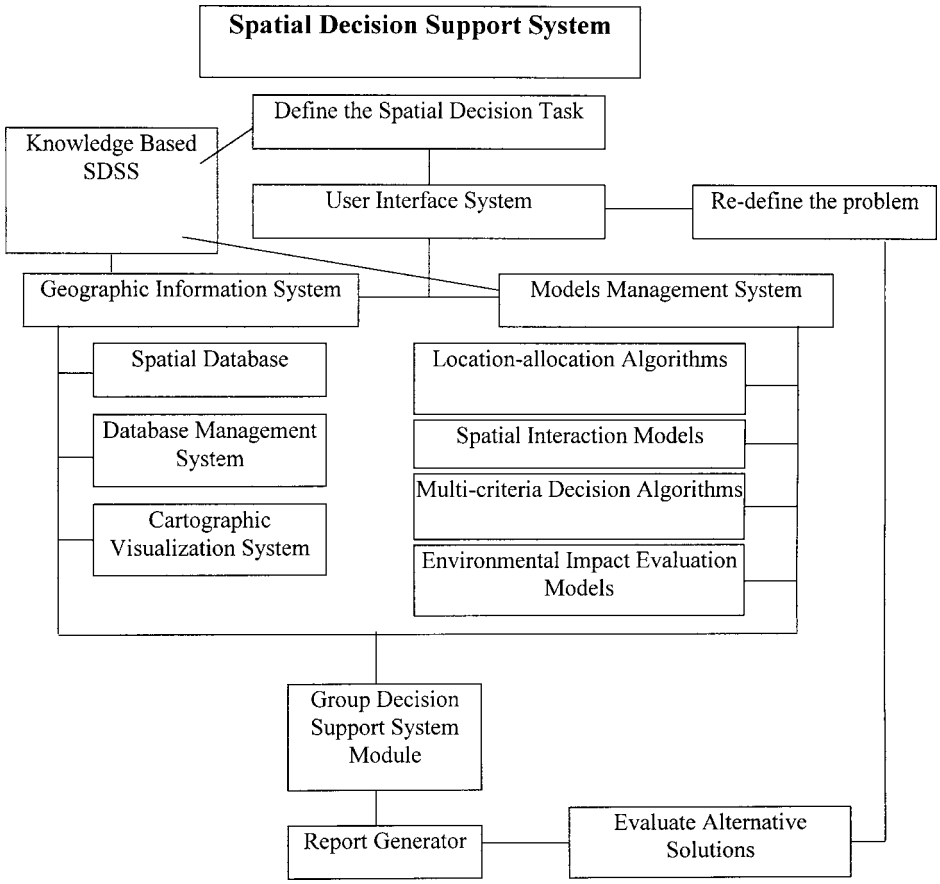

Figure 1
Fundamental elements of spatial decision support systems

locations at regular intervals to headquarters. This information is used in subsequent spatial assignment tasks.

The principal spatial task of some SDSS is to form regions (Densham and Rushton1996). Strapp and Keller (1996) developed a spatial decision support system to support the consolidation of agricultural land in the Czech Republic. Their link with GIS permitted the dynamic computation of spatial relationships between small pieces of territory and the farmsteads into which they are grouped at any point in the region formation process. Because of the ill-structured nature of this problem, they designed a spatial decision support system to permit users to work interactively with provisional regions to improve upon them.

2.3 Group Decision Support Systems

Spatial decisions are often made by groups of people, involving them in collaborative efforts. Nyerges and Jankowski (1997) discuss using adaptive structuration theory in a GIS environment to assist these efforts for situations in which groups are involved in public land planning controversies such as hazardous waste clean-up and habitat redevelopment.

Some SDSS incorporate methods for reconciling different views. Couclelis and Monmonier (1995) discuss this problem and offer a computer-supported system to help promote the resolution of NIMBY-type ('not in my back yard') planning controversies. Their system uses GIS capabilities to present spatial details of the planning situation but uses the concept of the graphic script as a device for encouraging dialogue and reconciliation of opposing views. Armstrong (1993) describes how computer supported cooperative work (CSCW) computing environments can enhance SDSS by enabling group members to observe, manipulate, and evaluate alternative plans collectively and then provide mechanisms for conflict resolution.

3. Knowledge-based SDSS

In many situations, the assistance required by users of an SDSS lies in the knowledge of experts who are not

available in person. Much of current research in SDSS entails capturing the knowledge of experts and incorporating it in the analysis system so that this knowledge becomes available at the appropriate time to the user. An early paper by Armstrong et al. (1990) outlined this approach. Examples of such systems are beginning to appear (Leung et al. 1997). In fully developed knowledge-based SDSS, many of the activities of the user-interface are bypassed (see Fig. 1) and are automated in the knowledge base. Some of the knowledge there is about the algorithms and their links to the decision task. Other knowledge is substantive knowledge about the processes that need to be modeled.

4. Implementations of SDSS

Many examples of prototype implementations of SDSS can be found in the academic literature. Commercial SDSS applications are usually domain-specific and are being developed to support decision making for specific problems that are complex, yet occur with sufficient frequency to support the overhead cost of developing and testing SDSS developed for the purpose. In many countries, following each census, there is a flourish of political redistricting activities. Increasingly, decision makers rely on SDSS for this task.

5. Conclusions

Currently, most SDSS are tailored for specific problem domains. Users often will not recognize that SDSS in quite different application domains have similar architectures and combine similar elements. The applications use the attribute language of the domain, even though the spatial data inputs, manipulations, and outcomes may be identical. SDSS represent the integration of ideas, methods, and techniques from many of the social, behavioral, and management sciences. Their usefulness in the policy and planning sciences is based on the degree to which they successfully incorporate the findings and approaches of these sciences. Their focus on the spatial and the geographic leads to a very important link with the evolving field of geographic information science. The rapid development of national spatial data infrastructures provides a rich and expanding database to which attribute information can be attached.

SDSS, at their inception, were a derivative science that built upon the successful development of models of behavioral, social, and political processes. However, with the expanding use of SDSS in a wide variety of decision situations, theories and models of these processes will have to account for the use of SDSS. When normal decision-making patterns involve SDSS use, they will become an important element in social science theories and models to explain decisions in these areas. This evolution from the use of social and behavioral models as exogenous units in SDSS to their incorporation as endogenous units is an important development in the social sciences. This development will mark the more explicit incorporation of geographic variables in social science theory and method.

See also: Geographic Information Systems; Locational Conflict (NIMBY); Spatial Choice Models; Spatial Equity; Spatial Optimization Models

Bibliography

Armstrong M P 1993 Perspectives on the development of group decision support systems for locational problem-solving. *Geographical Systems* **1**: 69–81

Armstrong M P, De S, Densham P J, Lolonis P, Rushton G, Tewari V K 1990 A knowledge-based approach for supporting locational decisionmaking. *Environment and Planning B: Planning and Design* **17**: 341–64

Birkin M, Clarke G, Clarke M, Wilson A 1996 *Intelligent GIS: Location Decisions and Strategic Planning*. Geoinformation International, Cambridge, UK

Couclelis H, Monmonier M 1995 Using SUSS to resolve NIMBY: How spatial understanding support systems can help with the 'Not in my backyard' syndrome. *Geographical Systems* **2**: 83–101

Crossland M D, Wynne B E, Perkins W C 1995 Spatial decision support systems: an overview of technology and a test of efficacy. *Decision Support Systems* **14**: 219–35

Densham P J 1994 Integrating GIS and spatial modeling: visual interactive modeling and location selection. *Geographical Systems* **1**: 203–21

Densham P J, Rushton G 1996 Providing spatial decision support for rural public service facilities that require a minimum work load. *Environment and Planning B: Planning and Design* **23**: 553–74

Ghosh A, Rushton G (eds.) 1987 *Spatial Analysis and Location-allocation Models*. Van Nostrand, New York

Goodchild M F 1984 ILACS: A location-allocation model for retail site selection. *Journal of Retailing* **60**: 84–100

Jankowski P, Ewart G 1996 Spatial decision support system for health practitioners: selecting a location for rural health practice. *Geographical Systems* **3**: 279–99

Keenan P B 1998 Spatial decision support systems for vehicle routing. *Decision Support Systems* **22**: 65–71

Leung Y, Leung K S, Shao Z, Lau C K 1997 A development shell for intelligent spatial decision support systems: 2. An application in flood simulation and damage assessment. *Geographical Systems* **4**: 39–57

Nyerges T L, Jankowski P 1997 Enhanced adaptive structuration theory: a theory of GIS-supported collaborative decision making. *Geographical Systems* **4**: 225–59

Strapp J D, Keller C P 1996 Towards an interactive spatial decision support system for agricultural land consolidation. *Geographical Systems* **3**: 15–31

Timmermans H J P, van der Heyden R, Westerveld H 1984 Decisionmaking between multi-attribute choice alternatives: a model of spatial shopping behaviour using conjoint measurements. *Environment and Planning A* **16**: 377–87

G. Rushton

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Spatial Equity

Although most social scientists may not know the term 'spatial equity,' almost everyone experiences it in everyday life. 'Spatial' implies that most of the experienced world, physical and social, takes up room; this has important social consequences. Elements of the natural and social world always have a location in space or a place, just as they do in time. Phenomena also have relationships of distance, spatial interaction and accessibility with one another, and these attributes shape the causes and consequences of things. 'Space' has structure, that is, institutions and relationships which constrain what and who can be where—in other words, everything and everyone's place on the landscape.

'Equity' is a social construct by individuals, or groups or societies, as to the degrees and kinds of inequalities that are acceptable or 'fair' or are unacceptable and 'unfair.' Thus, the measure of equity is inescapably bound up with the individual's or group's position with respect to the basic human dilemma of individual and collective rights and responsibilities. There is no measure of where 'fair' ends and 'unfair' begins; society redefines the line as needed according to custom and law. Classic works include Perin's *Everything in its Place* (1977), Rawls' *A Theory of Justice* (1971) and Harvey's *Social Justice and the City* (1973). The latter was recently commemorated in Merrifield and Swyngedouw's *The Urbanization of Injustice* (1997). On Rawl, see also Clark (1985).

1. *The Meaning of Spatial Equity*

Webster defines equity as 'a legal system which supplements statutes or laws to correct, if necessary, its failure to account for what is morally just and fair.'

Societies may deal with questions of equity and fairness in the justice system or via regulations and legislated standards. Thus a 13-year-old may not claim an equity issue in not being allowed to drive, but a healthy 80-year-old can.

The very fact of space implies that the notion of spatial equality (that all parts of space are of the same quality or value) is without meaning; every bit of space is different from every other. For this reason, equity is of heightened importance and difficult to obtain. In economic theory an efficient equilibrium should be equitable as well. Structurally, monopoly and inheritance, for example, lead to inefficient and inequitable outcomes. But space itself may make a structurally efficient pattern geographically inequitable. Consider a common situation of an uneven distribution of people, far denser on one side of a jurisdiction than the other. The most efficient, travel-minimizing location for a public facility will be strongly pulled to the densest core, leaving those in the low density edge in a highly unequal position (Morrill 1973).

The physical environment is incredibly diverse, and utterly unequal from place to place and over time in its utility or value to human production, residence, exchange or enjoyment. For physical, cultural, historic, and political reasons, individuals, groups, corporations, institutions, and societies are similarly diverse, and highly unequal in physical and mental capacity, experience, and mobilization of resources. They are profoundly unequal in their power to structure the landscape, to alter the environment, or to control more valuable locations.

Individuals, groups and societies are continually aware of their unequal positions in space, which often leads to perceptions of unfairness on the part of the less territorially advantaged (those on the 'wrong side of the tracks'), and on the part of moralists, ethicists, and liberals offended by or concerned with territorial inequality.

1.1 *Spatial Equity and Social Inequality*

Inequality of power and of rewards itself may lead to resentment and a profound sense of unfairness of the social system. At the societal level, the equity issue is the social system itself, and the battles are more structural than spatial, for examples, calls for socialism or communism as the solution for unequal power and wealth, or battles over minimum wages and tax rates. An exception is perhaps the demand for political self-determination of colonies—struggles that were more territorial and cultural than structural.

But within a nation and social system, as in the USA, although there may be many who want to change the system, the majority of US citizens, for example, accept the competitive and individual system and understand that unequal power and wealth are inherent to the system, that the rich and powerful can and will preempt the 'best places.' Thus a class-segregated society as such is not a spatial equity issue. After all, that poor person may win the lottery! Rather, issues of spatial equity arise when people feel unfairly treated, although they're playing within the rules of the game; when they feel that the rich and powerful manipulate the law itself to get more than the share their competitiveness and wealth warrant. An example is the conviction of poverty area residents that since they pay taxes, too, the potholes on their streets should have no lower priority than the potholes on the streets of the rich.

There are many examples of spatial inequality that raise equity issues. Starting at the global scale are historical colonialism and imperialism and contemporary dependency theory, hypothesizing that unfair power relations exploit and impoverish the Third World. Less technologically advanced cultures distrust

and fear the power of Western and especially of US culture. Smith (1977, 1994) provides extensive treatment of inequality from the global to the local scale. See also Knox (1975).

Within nations, prominent bases for raising issues of equity include (a) regional inequality, the sense of exploitation of the periphery by the core, or of poor regions by rich regions; (b) unequal access to public facilities and resources; (c) unequal or unfair treatment of areas because of their poverty, wealth or power; (d) spatial manifestations of the sports analogy that a competitive society should have a 'level playing field;' and (e) spatial aspects of unequal treatment of categories of people, because of race, ethnicity, language, religion, lifestyle, and gender.

1.2 Regional Inequality

Within the European Union, a dominant part of the budget and a large amount of research and rhetoric are devoted to the issue of lagging regions, built on the perception that it is somehow 'unfair' that some regions are prosperous and growing while others are poor and languishing; that it shows the failure of the system, rather than of the region. In US economic history, class struggles often seemed secondary to the resentment of the south and west against the hegemony of the northeast. In a generally private economy, the US has a long tradition of regional redistribution via public programs; for example, the Reclamation Act subsidizing irrigation development in the West, federal defense outlays and massive highway programs, TVA and the Bonneville Power Administration, and the Regional Development Commissions, were all designed to improve the competitive position of 'disadvantaged' parts of the country (Johnston 1980). Harrington's classic *The Other America* (1962) raised the nation's consciousness of poverty and helped lead to the War on Poverty.

Efficient market decisions frequently take the form of the closure of factories and consequent community decline, leading to forced out-migration. In these cases, 'pure market' solutions implying the 'shipment' of the surplus labor to growing regions confront the immense nonmarket value of places to their long-term residents, and a strong sense of unfairness arises, that the decline is not the fault of local people (Clark 1983). Societies often institute a variety of programs to reduce the social and economic costs, as through subsidies, community purchase of the closed plant, tax forgiveness, or retraining and relocation assistance.

The same principle is applied at the local level through federal or local programs such as slum clearance, public housing, and 'enterprise or empowerment zones,' reflecting a political consensus that the landscape had become unfairly unequal, presumably because of forces beyond local control.

1.3 Voting and Representation

In the USA, perceptions of unfairness in representation led to a series of Supreme Court decisions outlawing malapportionment (unequal population per district) and racial gerrymandering which disenfranchised minorities (initial decisions ruled against 'diluting' and 'packing' minority voters, but later rulings forbade gerrymandering to maximize minority representation). Ironically representation in the US Senate intentionally violated the equal population principle, since every state has two senators, but this inequality was defended on grounds of avoiding unfair domination of smaller by larger states.

1.4 Overcoming the 'Unfairness' of Nature

As communities develop and compete with other places, perceptions of being treated 'unfairly by nature' may arise. At the international scale, such perceptions may be used to justify invasions and wars. Within countries, governments at all levels spend huge sums to overcome natural limits, so they may 'compete on that more level playing field.' Examples are bridges, tunnels, and artificial harbors, built to improve the accessibility of a place to national and global markets.

1.5 Access to Public Facilities and Resources

This is the arena for which the word equity is most likely to be invoked, and for which legal remedies may be sought, on grounds of 'equal protection' under law (Pinch 1985). In the USA, until as recently as the 1960s it was the perception and probably the reality that weaker, poorer places among a set of jurisdictions, and weaker and poorer areas within a jurisdiction, were unequally and unfairly treated. Examples of inequitable access to public services and facilities entail the maintenance and quality of utilities and infrastructure, like roads; the location and quality of parks, playgrounds and pools, or of libraries; and the timeliness of police response calls, as if action were a function not of citizenship but of wealth—which it was. As a consequence of a series of court decisions over many years, cities have become far more attentive to more equal treatment across all parts of their jurisdictions, although huge historic inequities remain.

The construction and operation of public transportation systems have similarly been subject to scrutiny, as to the costs and benefits of subsidies to different parts of a region; poor, more transit-dependent districts no longer meekly accept inferior service and equipment. In most of the USA school districts are a distinct set of governments. District populations are astoundingly unequal in income, and districts in tax bases. In many but not all states, courts

have interpreted state constitutions to mean that all students should have 'equal access' to education, wherever they live. This has resulted in substantial redistribution of resources from richer to poorer districts, as well as to the subsidy of 'remote but necessary' small schools in very isolated communities.

1.6 The Location of Undesirable Facilities: Environmental Racism

A different issue of equity arises as to the differential ability of stronger and richer or weaker and poorer jurisdictions or parts of cities to ward off undesirable private and public activities. Less advantaged places and areas widely perceive that they have long been and remain dumping grounds for unattractive and even damaging facilities—prisons, halfway houses, incinerators and landfills, and services to the 'less desirable' folks (drug addicts, alcoholics, the homeless); adult shops and theaters, and polluting and noisy factories. If the obnoxious facilities or activities are in segregated racial minority areas, charges of 'environmental racism' may be made.

Poorer and minority areas perceive also that they are treated as tolerance zones for illicit activities and inadequate law enforcement. More generally, these areas are characterized by loss of jobs, retailing and services, and middle-class housing, leading to spatial mismatch between a declining inner city and more affluent suburbs. In more liberal cities and states, a consensus may arise for some degree of 'sharing' of the 'burden' of housing the less affluent, as through requirements for 'fair shares' of subsidized affordable housing, even in suburbs. In a very few metropolitan areas, where extreme political fragmentation results in highly inequitable resources for jurisdictions, a sense of injustice and unfairness has resulted in a degree of tax-base sharing, as for police and sewerage services.

1.7 Urban Inequality: Unequal treatment of categories of people

Many books focus on spatial inequality in US and European cities, e.g. Badcock (1984), Gottdiener (1985) and Katznelson (1981). An influential segment of the intellectual and planning community advocates fairly radical efforts to break down the high degree of racial and class segregation, as through stronger zoning control of settlement. The failed programs of mandatory bussing for school desegregation were similarly based on a sense of injustice and on the premise that mixing in classrooms would lead to mixing in neighborhoods. As with schools, current efforts toward class mixing seem unlikely to be a match for the stronger societal forces for segregation. Although men and women are usually residentially integrated, the history of gender inequality resulted in unequal access to opportunities and resources, and in insecure built environments. And because of this history the very subject of inequitably gendered landscapes has only recently been studied (Pratt and Hanson 1994).

2. Reactions to Spatial Inequity

As most of these examples show, the achievement of greater spatial equity seems limited to those aspects of everyday life in which public resources or facilities are involved. Social intervention in the form of redistribution of resources toward weaker places and areas seems to occur when a social consensus of 'market failure' arises—that the 'playing field' has become too uneven, within the prevailing social order. Within the social system, it proves easier also to address inequities in some arenas of life than in others. Not surprisingly it is less threatening to those on top, usually the holders of wealth, to accept equalizing measures with respect to political participation and voting and social dimensions (race, language and gender, although these remain highly contested), and even, if rarely, as to urban housing, than with respect to the truly key economic arena of income, taxes and inheritance.

In the USA, issues of equity and redistribution were fought mainly at the national level for most of the post-world war period. Since the mid-1980s and national shifts toward the right—that is, toward greater individual than collective responsibility—the arena of action has moved to local and state levels, where coalitions of activist groups have sometimes been able to gain enough power to enact redistributive plans unimaginable at the national level, for example, high minimum wages, rent controls, or mandates for affordable housing, in a process that might be called 'pluralist democracy' (see also *Postmodern Urbanism*). In other areas and arenas conservative coalitions stress faith in the market, privileging its efficiency over concerns for equity. Indeed, the concepts of efficiency and equity are the very symbols of differing ideological stances toward inequality. The treatment of equity (including this discussion) is usually based on a liberal assumption of people's similarity in needs and values—ideas going back at least to John Stuart Mill. The new politics of identity and difference may complicate the meaning and evaluation of equity (Keith and Pile 1993, Young 1990).

2.1 The Analysis of Spatial Equity

Maps, and now GIS (geographic information systems) have proven highly effective in revealing the structure of inequality in the landscape—patterns of segre-

gation, poverty, concentration of pollution, and unequal access, and have proven essential to convince legislative bodies and courts to address equity issues. A variety of both statistical and mathematical models have been used to measure inequality and its costs, and to assess the impacts of policy changes. Awareness of the human costs of spatial inequality may be increased through more ethnographic studies of the situation of individual households.

See also: Access: Geographical; Equity; Location: Absolute/Relative; Location Theory; Locational Conflict (NIMBY); Postmodern Urbanism; Racism, Sociology of; Residential Segregation: Geographic Aspects; Spatial Interaction Models; Spatial Pattern, Analysis of; Urban Poverty in Neighborhoods

Bibliography

Badcock B 1984 *Unfairly Structured Cities*. Basil Blackwell, Oxford, UK

Clark G 1983 *Interregional Migration, National Policy and Social Justice*. Rowan and Allenheld, Totowa, NJ

Clark G 1985 Making moral landscapes: John Rawl's original position. *Political Geography Quarterly* **5**: 147–62

Gottdiener M 1985 *The Social Production of Space*. University of Texas Press, Austin, TX

Harrington M 1962 *The Other America*. Macmillan, New York

Harvey D 1973 *Social Justice and the City*. Johns Hopkins University Press, Baltimore, MD

Jackson P, Smith S 1984 *Exploring Social Geography*. Allen and Unwin, London

Johnston R 1980 *Geography of Federal Spending in the United States*. Wiley, Chichester, UK

Katznelson I 1981 *City Trenches*. Pantheon, New York

Keith M, Pile S (eds.) 1993 *Place and the Politics of Identity*. Routledge, London

Knox P 1975 *Social Well-being: A Spatial Perspective*. Oxford University Press, London

Kodras J, Jones J P 1995 *Geographic Dimensions of US Social Policy*. Edward Arnold, New York

Merrifield A, Swyngedouw E 1997 *Urbanization of Injustice*. New York University Press, New York

Morrill R 1973 Efficiency and equity aspects of optimum location models. *Geographical Analysis* **9**: 215–26

Perin C 1977 *Everything in its Place*. Princeton University Press, Princeton, NJ

Pinch S 1985 *Cities and Services*. Routledge, London

Pratt G, Hanson S 1994 Geography and the construction of difference. *Gender, Place and Culture* **1**: 5–30

Rawls J 1971 *A Theory of Justice*. Harvard University Press, Cambridge, MA

Smith D 1977 *Human Geography—A Welfare Approach*. Edward Arnold, London

Smith D 1994 *Geography and Social Justice*. Blackwell, Cambridge, UK

Walzer M 1983 *Spheres of Justice*. Basic Books, New York

Young I 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

R. Morrill

Spatial Interaction

The study of spatial interaction has been carried on actively for many years in the field of geography, with its emphasis on the spatial aspects of cities and regions. Spatial interaction studies in geography may be divided into those dealing with spatial structure and those dealing with attempts to understand and model these structures. Spatial structure includes the study of: linkages the connections between centers; nodes, the centers themselves; hinterlands, the tributary area surrounding the nodes; and hierarchies, the systems composed of hinterlands, nodes, and linkages. Each of these will be examined briefly, with a particular emphasis on Chicago, as an example of a large, centrally located focus of spatial interaction. Then, some models of spatial interaction will be discussed as examples of attempts to describe and explain the structural systems of linkages, nodes, hinterlands, and hierarchies.

1. Structural Systems of Spatial Interaction

1.1 Linkages and Nodes

Linkages are a basic part of spatial interaction and may be studied at several different geographic scales. At the local scale, dense and complex networks of commuting linkages surround large metropolitan centers connecting city and suburban residential areas to employment areas. In the case of Chicago, certain corridors have been traced out by the flow of commuting traffic. These corridors developed initially in connection with commuter railroads and were later massively reinforced and ultimately replaced by expressways which both compete with and complement the rail linkages. Shopping and recreational flows form quite different sets of linkages with a more dispersed and multicentered pattern. In Chicago and other large metropolitan centers, beltways in the form of high-capacity peripheral expressways encircling the city have become important additional linkages connecting suburbs and connecting with the major radial corridors leading into the central city.

At the national scale, certain primary corridors have developed in the network of linkages. Chicago is at the western end of a major corridor extending through the Midwest to connect with the Megalopolis corridor extending from Washington to Boston. In air transportation, quite a different national pattern of linkages has developed, dominated by the traffic flows between large cities so that Chicago's strongest linkages are with New York and Los Angeles rather than with nearby cities.

Nodes within networks represent another structural aspect of spatial interaction study. The relative im-

portance of a node is determined by the size and nature of its linkages. At the local level important nodes are located along the major radial corridors and near the beltways. Nodes located in areas where beltways intersect radials have particularly high accessibility to the rest of the metropolitan area.

At the national level, the development of nodality has been closely linked to technological change. Chicago's national-scale nodality in the railroad era was unmatched by any other large US city. As the US highway system and the Interstate Highway System emerged, however, Chicago's nodality was no greater than that of a number of other large metropolitan areas. With the coming of high-speed, high-capacity jet airline service in the 1960s, Chicago again served as major focal point, ranking with New York and Los Angeles in number of high-density linkages. When national communication networks are considered, however, a different pattern of spatial interaction emerges. For television, entertainment, and publishing, there is a strong focus on the New York and Los Angeles nodes.

1.2 Hinterlands and Hierarchies

Spatial interaction has also been studied in terms of hinterlands and hierarchies. Systems of linkages and nodes give rise to groups of tributary areas, or hinterlands which, in turn, may be organized into hierarchical networks. Some of the earliest geographic studies of spatial interaction dealt with the delimitation of local hinterlands. These could be based on commuting, retail market areas, social trips, recreational trips, newspaper circulation, or shipments of goods. Chicago's large metropolitan area, which is mainly defined by its commuting linkages, also contains within it a number of small cities each with its own commuting and retail hinterlands. The result is a complex, fine-grained hierarchical structure of central places.

At the national scale, Chicago itself is embedded into several hierarchical systems. In air passenger traffic, for example, New York has become the single most dominant center. By the 1990s, it was the leading passenger generator not only for a group of adjacent cities but also for such large more distant cities as Chicago, Atlanta, and Los Angeles. In hierarchical fashion, each of these centers dominates its own surrounding hinterland. In terms of scheduled airline traffic as opposed to air-passenger origin and destination, a hierarchical hub-and-spoke system has developed. Short-haul connections for a particular airline focus on a single hub where passengers transfer to connecting flights to other hub centers.

Communications networks, such as those developed in the film and television industries, have a clearly bipolar structure focused on New York and Los Angeles. Other large metropolitan centers, such as Chicago, occupy far less central positions in the system. The rise of the Internet is in the process of creating still a different structure of spatial interaction. To date, the system has been a dispersed one in which it is difficult to detect major nodes such as those marking hierarchical or bipolar systems.

At the global scale, the early pattern of world trade routes focused on Britain and a number of strategically located ports has given way to an extensive system focused on major world cities such as London, New York, and Tokyo. There is also some evidence of a hub-and-spoke system, with the major world cities serving as regional centers, distributing international traffic to their surrounding areas. London serves European cities; New York, US cities; Miami, Latin American cities; and Tokyo, Hong Kong, and Singapore serve Asian cities.

2. Spatial Interaction Models

Models have ranged from the descriptive through the explanatory, predictive, and normative. Useful descriptions of the comparative accessibility of given networks, or of given nodes within a network, have been provided by graph-theoretic models. Hubs may be identified, and the differing accessibility of industrial cities may be measured and compared. Such measures, both for nodes and for entire networks, permit planners to evaluate the effects of proposed new linkages on nodal and network accessibility.

More explanatory models evolved from earlier descriptive efforts as attempts were made to identify some of the underlying forces shaping systems of spatial interaction. There is a large literature dealing with the observed tendency for interaction between any two cities to increase as the distance between them decreases. This relationship was first described as a gravity model: directly proportional to the product of the two cities' populations and inversely proportional to the distance between them. It soon became apparent, however, that spatial interaction influenced by many other factors. Distance had different effects for different kinds of interaction. Highway traffic, for example, was much more sensitive to distance than was air traffic. Other characteristics of cities such as per-capita income, economic complementarity, and community of interest had to be considered. The spatial structure of the regions surrounding the cities was also found to have effects in the form of intervening opportunities for spatial interaction and of the proximity of competing origins or destination cities. As the explanatory models continued to ramify, they evolved into a series of traffic demand equations incorporating a number of variables that had been shown empirically to affect spatial interactions.

Intra-urban spatial interaction, that is, the flow of traffic within large cities, has come in for special attention in recent years. Initially, aggregate explana-

tory models were used by planners and engineers as a basis for predicting and improving urban transportation. These models were codified and used in large traffic surveys and analyses. Cities were divided up into analysis zones and detailed surveys were carried out. Demand equations were used to estimate trip generation by analysis zones and the trips were then allocated to existing transport network facilities and used as the basis for future planning. Later, these aggregate models generally were replaced by models designed to facilitate an understanding of individual travel behavior. Data were collected and analyzed at an individual or household level rather than at an aggregate traffic-zone level. These behavioral studies addressed questions of travel choice at an individual level. How did such attributes as comfort, reliability, safety, waiting time, or privacy related to time and cost? How did the attributes interact with such travels characteristics as time and cost? What trade-offs were involved, as travelers would choose between such present or planned types of service as auto, bus, express bus, park-n-ride, car pools, high-occupancy traffic lanes, light rail, and others?

Spatial interaction was also modeled in a normative fashion. The most efficient patterns of linkages between a given set of supply nodes and a given set of demand nodes could be ascertained with programming models. Constraints could then be introduced to account for such things as bottlenecks, capacity limitations, or particular market characteristics. Also, if it were desirable to maximize profits rather than minimize transport, a more complex set of economic functions would be used at the various cities. In addition to optimal linkage patterns, programming models have been used to identify optimal hinterland or tributary area patterns. An optimal set of school districts, for example, can be obtained by minimizing the travel times of a given set of schools with varying capacities. Similar studies have been made for tributary areas of supermarkets, rural hospitals, and others.

Normative models have also been used for purposes other than the establishment of an optimal system of linkages or hinterlands. Sensitivity analysis permits the planner to view the consequences of a number of alternative proposed changes in the transport facilities. Each proposed change can be run through the model and the minimum costs identified. For example, in a school-district case, decisions as to the optimum location for adding or eliminating a school could be aided by running the model for each alternative. Trade-offs could then be evaluated. The cost in increased travel time of certain socially desirable alternatives, such as adding to the ethnic or racial diversity of each school, could be ascertained and evaluated in the light of other possible policy alternatives.

Thus, there is a gradation from descriptive models to explanatory, predictive, and normative models, all of which may provide improved understanding of spatial interaction as it is expressed in structural systems of linkages, nodes, hinterlands, and hierarchies.

See also: Access: Geographical; Communication: Geographic Aspects; Market Areas; Spatial Choice Models; Spatial Data Infrastructure; Spatial Interaction Models; Spatial Labor Markets; Spatial Optimization Models; Transportation Geography; Urban Activity Patterns

Bibliography

Fleming D Hayuth Y 1994 Spatial characteristics of transport hubs: Centrality and immediacy. *Journal of Transport Geography*, Vol. 2, pp. 3–18

Fotheringham A S, O'Kelly M E 1989 *Spatial Interaction Models: Formulations and Applications*. Kluwer Academic Publishers, Dordrecht

Golledge R G, Rushton G (eds.) 1976 *Spatial Choice and Spatial Behavior*. The Ohio State University Press, Columbus

Hagget P, Chorley R J 1969 *Network Analysis in Geography*. Edward Arnold, London

Hanson S (ed.) 1995 *The Geography of Urban Transport* (2nd edn.). Guilford Press, New York

Hoyle B S, Knowles R D (eds.) 1992 *Modern Transport Geography*. Belhaven Press, London

Janelle D 1969 Spatial reorganization: A model and a concept. *Annals of the Association of American Geographers* **5**: 348–54

Pred A R 1966 *The Spatial Dynamics of U.S. Urban-Industrial Growth, 1800–1914*. MIT Press, Cambridge

Taaffe E J, Gauthier H, O'Kelly M E 1996 *The Geography of Transportation*, 2nd edn. Prentice-Hall, Upper Saddle River, New Jersey

Ullman E 1956 The role of transportation and the bases for interaction. In: Thomas W L (ed.) *Man's Role in Changing the Face of the Earth*. University of Chicago Press, Chicago

Vance J E Jr 1980 *Capturing the Horizon: The Historical Geography of Transportation*. Harper and Row, New York

E. J. Taaffe

Spatial Interaction Models

1. Definitions and Uses

Spatial interaction is the general term for any movement of people, goods, or information over space that results from a decision-making process. Specific examples include movements such as migration, shopping trips, commuting, trips for recreational purposes, trips for educational purposes, freight flows, the spatial pattern of telephone calls, emails and world-wide web connections, of the use of healthcare facilities. Consequently, spatial interaction is a very broad topic, and spatial interaction models have a wide range of

applications (see also *Spatial Interaction*). Olsson (1970, p. 233) states:

> The concept of spatial interaction is central for everyone concerned with theoretical geography and regional science … Under the umbrella of spatial interaction and distance decay, it has been possible to accommodate most model work in transportation, migration, commuting, and diffusion, as well as significant parts of location theory.

Spatial interaction *models* are mathematical descriptions of spatial flows (see *Mathematical Models in Geography*). The models serve two purposes. One is to predict flows where they are unknown. For example, spatial interaction models have been used to predict the impact on traffic patterns of a new development (O'Kelly 1986), to predict the impact a new store will have on trade at existing stores (Fotheringham and O'Kelly 1989), and to predict the optimal spatial arrangement of healthcare facilities (Mayhew et al. 1986). The other is to yield information on the determinants of the flow system being analyzed. For example, spatial interaction models are commonly used to determine the influence of various store attributes on the choice of stores by consumers (Fotheringham and Trew 1993) or to examine the characteristics of places which affect in and out migration (Boyle et al. 1998, Flowerdew and Salt 1979).

Some questions that can be answered with the judicious use of spatial interaction models include:

(a) Is there sufficient demand between two cities to justify direct air service and what frequency of air service will be sufficient to meet the demand?

(b) Where is the optimal location for a new supermarket within a metropolitan area?

(c) What will the effect on traffic patterns be of building a new bridge or tunnel across a river?

(d) What can a store do to attract new customers—what features of the store are most likely to attract or deter customers?

(e) Are migrants becoming less constrained by distance over time?

(f) What influence does a political or administrative boundary have on students' choices of a university?

The remainder of this article examines some of the basic elements of spatial interaction modeling. It then describes four common types of spatial interaction models and outlines a typical application of each type of model. Brief discussions of model calibration and local forms of spatial interaction models are then given followed by a description of the various underlying theoretical structures that have been proposed to substantiate spatial interaction models.

2. *The Basic Elements of Spatial Interaction Models*

The starting point for spatial interaction modeling is a matrix of flows between a set of origins, each of which is usually labeled i, and a set of destinations, each of which is usually labeled j. The task is to represent the number of trips between any i and j, T_{ij}, as a function of origin characteristics, destination characteristics, and the separation between origins and destinations.

Hence spatial interaction models contain four basic elements:

(a) a measurement of the spatial interaction

(b) a measurement of the propulsiveness of each origin (that is, a measure of how many flows begin at each origin)

(c) a measurement of the attractiveness of each destination (that is, how many flows terminate at each destination)

(d) a measurement of the spatial separation between origins and destinations.

Each of these elements can be measured in different ways depending on the type of spatial interaction being modeled. Usually, distances between origins and destinations are used to measure spatial separation although cost and time can also be used. Regardless of how spatial separation is measured, it is often the most important determinant of spatial interaction patterns: individuals tend to be highly constrained by space for most types of spatial interactions. However, spatial separation is not the only determinant of spatial interaction patterns. For example, while an individual is likely to patronize a supermarket close to home, it may not always be the supermarket *closest* to home that is the most attractive: a supermarket farther away but offering a wider range of products or cheaper prices might be more attractive. This trade-off between proximity and other destination attributes is the essence of spatial interaction modeling.

Sometimes, the number of trips leaving each origin or the number of trips terminating at each destination is known from surveys or from some other exogenous source. In such cases, this information is used directly in the model. When such information is not known, the model has to contain surrogates of origin propulsiveness or destination attractiveness or both. Table 1 gives some ways in which origin propulsiveness and destination attractiveness might be measured for different types of spatial interaction. Different forms of spatial interaction model arise from having different amounts of information on the number of trips leaving each origin and terminating at each destination. These different types of spatial interaction model and how they arise are now described.

3. *A 'Family' of Spatial Interaction Models*

It was established above that in order to predict the volume of interaction between places, it is necessary to have three types of information about those places: a measure of origin propulsiveness, a measure of destination attractiveness, and a measure of spatial separation. The last is readily available as distances

Table 1
Possible measures of origin propulsiveness and destination attractiveness

Type of spatial interaction	Origin propulsiveness	Destination attractiveness
Shopping trips	Population Household income	Store size Number of parking spaces Variety of products Price levels
Commuting flows	Population Household income Car ownership rate	Number of jobs
Migration	Population Crime rate Unemployment rate Wage rate	Population Crime rate Unemployment rate Wage rate Climate

between locations can easily be calculated, especially if the locations are stored digitally. This produces four possible situations in which different levels of information are available on origin propulsiveness and destination attractiveness. In turn, these situations give rise to four different forms of spatial interaction model which are collectively referred to as a 'family of spatial interaction models.' (Fotheringham and Dignan 1984) demonstrate that they are actually four extreme points on a continuum of spatial interaction models.

3.1 The Unconstrained Model

Suppose a measure of spatial separation is available but there is no information on how many trips originate at each origin and how many trips terminate at each destination. These latter two values need to be estimated within the spatial interaction model by surrogate variables such as those described in Table 1. The appropriate spatial interaction model in such a situation is known as an unconstrained model with the following general form:

$$T_{ij} = k\, V_{i1}^{\alpha 1}\, V_{i2}^{\alpha 2} \ldots V_{ip}^{\alpha p}\, W_{j1}^{\lambda 1}\, W_{j2}^{\lambda 2} \ldots W_{jq}^{\lambda q}\, d_{ij}^{\beta} \quad (1)$$

where T_{ij} is the number of trips that begin at origin i and end at destination j, V_i represents an origin attribute that is a surrogate for the number of trips beginning at i, W_j represents a destination attribute that is a surrogate for the number of trips ending at j, and d_{ij} represents the spatial separation (assumed here to be measured by distance) between origin i and destination j.

The number of origin and destination variables included in the model depends on the type of spatial interaction being modeled and the information available. Here the number of origin variables is denoted by p and the number of destination variables is denoted by q. In some situations there might only be one origin attribute and one destination attribute in the model. The parameters α, λ, and β reflect the influence of each of the exogenous variables in the model on the volume of interaction and the values are obtained in the calibration of the model (see below). For example, if V_{i1} were population, it would be expected that $\alpha 1$ would be positive: as the population of an origin increased, a greater volume of interaction would emanate from it, everything else being equal.

Similarly, β would normally be negative: as the distance between any two places increases, the volume of interaction between them decreases, everything else being equal. The exact rate at which interaction decreases or 'decays' as distance increases, is a matter for empirical evaluation. One of the fascinating aspects about spatial interaction modeling is the ability to estimate these parameters in different systems and to compare spatial behavior in different systems and over time. For instance, the value of the distance-decay parameter, β, has been shown to vary across different economic systems, over time, by income group, and by gender (Southworth 1983). Generally, the deterrence of distance for interaction is less in more developed economies and there is a gradual reduction in distance-decay over time as transportation facilities improve and as knowledge about different locations increases. Finally, the parameter k in the model is a scale factor that simply balances the units on both sides of the equation and is of no great behavioral importance.

3.2 The Production-constrained Model

Suppose data on spatial separation *and* on the numbers of trips emanating from each origin are available but there are no data on the number of trips that terminate at each destination. The appropriate spatial

interaction model for this situation is known as a production-constrained model and it has the form:

$$T_{ij} = O_i \frac{W_{j1}^{\lambda 1} W_{j2}^{\lambda 2} \ldots W_{jq}^{\lambda q} d_{ij}^{\beta}}{\sum_j W_{j1}^{\lambda 1} W_{j2}^{\lambda 2} \ldots W_{jq}^{\lambda q} d_{ij}^{\beta}} \quad (2)$$

where O_i is the total number of trips beginning at origin i and it replaces the surrogate origin propulsiveness variables in the unconstrained version. The model essentially distributes the number of trips from an origin in proportion to the relative attractiveness of each destination as measured by the numerator of Eqn. (2). The denominator acts as a constraint on the predicted trips so that

$$\sum_j T_{ij}^* = O_i \text{ for every origin} \quad (3)$$

where T_{ij}^* is the predicted value of T_{ij} from the model. The model is sometimes represented as:

$$T_{ij} = A_i O_i W_{j1}^{\lambda 1} W_{j2}^{\lambda 2} \ldots W_{jq}^{\lambda q} d_{ij}^{\beta} \quad (4)$$

where

$$A_i = [\sum_j W_{j1}^{\lambda 1} W_{j2}^{\lambda 2} \ldots W_{jq}^{\lambda q} d_{ij}^{\beta}]^{-1} \quad (5)$$

Production-constrained forms of spatial interaction models are encountered very frequently, particularly in retailing where estimates of the number of trips leaving each origin can be obtained from census population data. The models are used to discover the characteristics of stores that attract consumers and to predict the proportion of people in each residential neighborhood shopping at each store. Once such a model is calibrated, it can be used to estimate a new store's potential market and its impact on the sales at existing stores. The model can be used to estimate sales at a variety of possible locations to aid in the optimal location of a store. It can also help retailers decide on the optimal strategy to attract more customers to a store. For example, would more customers be attracted to a store if the number of parking spaces were doubled or if price levels were reduced slightly?

3.3 The Attraction-constrained Model

Suppose data on spatial separation and on the numbers of trips terminating at each destination are available but there are no data on the number of trips beginning at each origin. In this case, an attraction-constrained spatial interaction model is the most appropriate form because it uses the known inflow totals into each destination and a set of surrogate variables to represent the total number of trips beginning in each origin. The model has the general form:

$$T_{ij} = D_j \frac{V_{i1}^{\alpha 1} V_{i2}^{\alpha 2} \ldots V_{ip}^{\alpha p} d_{ij}^{\beta}}{\sum_j V_{i1}^{\alpha 1} V_{i2}^{\alpha 2} \ldots V_{ip}^{\alpha p} d_{ij}^{\beta}} \quad (6)$$

where D_j represents the total number of trips terminating at destination j and this replaces the destination attractiveness surrogates in the unconstrained model.

The attraction-constrained model is the complement of the production-constrained model. In this case the model allocates the number of trips terminating at a destination across the various origins in proportion to the relative propulsiveness of each origin as measured by the numerator of Eqn. (6). The denominator again acts as a constraint on the predicted trips but in this case, the constraint is on the number of predicted trips into each destination so that

$$\sum_i T_{ij}^* = D_j \text{ for every destination} \quad (7)$$

where T_{ij}^* again represents the predicted value of T_{ij} from the model. The model can also be represented as:

$$T_{ij} = B_j D_j V_{i1}^{\alpha 1} V_{i2}^{\alpha 2} \ldots V_{ip}^{\alpha p} d_{ij}^{\beta} \quad (8)$$

where

$$B_j = [\sum_i V_{i1}^{\alpha 1} V_{i2}^{\alpha 2} \ldots V_{ip}^{\alpha p} d_{ij}^{\beta}]^{-1} \quad (9)$$

The terms A_i and B_j in the production-constrained and attraction-constrained models are sometimes referred to as 'balancing factors' since they constrain the row and column totals, respectively, of the predicted flow matrix.

Attraction-constrained spatial interaction models are encountered less frequently than production-constrained models but they can be used to estimate the demand for housing in different residential neighborhoods due to the location of a new enterprise. They can also be used to analyze geographic patterns in student enrollments where the total number of students attending a university is known. The attraction-constrained model can then be used to determine the characteristics of places that affect the number of students attending from different origins. The model can also be used in a normative way to derive 'expected' numbers of students enrolling from each origin (the predicted trips from the model) and to compare these expected values with the actual enrollments. This comparison can help identify areas from which the university is drawing more students than expected, given the characteristics of these areas, and areas where it is drawing fewer than expected. The latter could be targeted to increase enrollments. Obviously, this type of application could easily be extended to retailing or any other area where there is competition to attract individuals.

3.4 The Production-attraction Constrained Model (or Doubly Constrained Model)

Suppose data are available on spatial separation, the total number of trips beginning at each origin, *and* the total number of trips terminating at each destination. The appropriate spatial interaction model is then the production-attraction constrained, or doubly constrained, model. This has the form:

$$T_{ij} = A_i\, O_i\, B_j\, D_j\, d_{ij}^{\beta} \tag{10}$$

where

$$A_i = [\textstyle\sum_j B_j\, D_j\, d_{ij}^{\beta}]^{-1} \tag{11}$$

and

$$B_j = [\textstyle\sum_j A_i\, O_i\, d_{ij}^{\beta}]^{-1} \tag{12}$$

The A_i and B_j terms have to be estimated iteratively and they ensure that the constraints in Eqns. (3) and (7) are met.

The doubly constrained model allocates trips to the cells of the flow matrix (the origin-destination links) given row and column totals. It does this solely on the basis of the distance between origins and destinations. Its use has mainly been confined to transportation applications where the total numbers of commuters leaving each residential neighborhood are known and the total numbers of employment opportunities in each destination are also known. What is unknown, and can be predicted from the model, is the distribution of trips between origins and destinations.

4. A Note on Model Calibration

Although a detailed discussion of the nature and methods of calibrating spatial interaction models is beyond the scope of this article, it is worth mentioning very briefly how the parameters in a spatial interaction model might be estimated. For a fuller discussion of the calibration of spatial interaction models, the reader is referred to Fotheringham and O'Kelly (1989).

In order to use a spatial interaction model to answer questions such as those posed in Sect. 1, it is necessary to obtain values for the model's parameters. If the purpose of the model is to obtain information on the way either origin or destination characteristics affect spatial interaction patterns, the model must be calibrated with some known flow data. This is done by finding the values of the parameters that maximize some criterion, such as the accuracy with which the model replicates the known flow pattern. Parameter values that are far from their 'true' values will tend to produce very poor estimates of the flows, while parameters close to their 'true' values will tend to produce more accurate estimates. It is sometimes useful to determine how sensitive a model's predictions are to variations in the parameters used to derive the predictions.

If flow data are not available, model calibration as described above is impossible and 'educated guesses' have to be made for parameter values. This could be done, for example, by surveying the results of previous calibrations of similar models. However, in practice this is fraught with difficulties and should be avoided where possible as it is rare to find exactly the same spatial interaction model with the same combinations of explanatory variables being used in different applications. Also, it appears that the parameters of spatial interaction models are quite context dependent (Meyer and Eagle 1982).

5. Local Forms of Spatial Interaction Models

The forms of spatial interaction models described above are referred to as 'global' models because they provide a single set of parameter estimates for the whole flow system being analyzed. For instance, in each of the above models, a single estimate of distance-decay is obtained when the model is calibrated. In most circumstances, it would be useful to obtain more detailed information on the flow system and to obtain separate distance-decay parameter estimates for each origin in the system or for each destination. This is quite easy to achieve by calibrating what are known as origin-specific models and destination-specific models. An origin-specific form of the production-constrained model given in Eqn. (4) is:

$$T_{ij} = A_i\, O_i\, W_{j1}^{\lambda 1(i)}\, W_{j2}^{\lambda 2(i)} \ldots W_{jq}^{\lambda q(i)}\, d_{ij}^{\beta(i)} \tag{13}$$

where

$$A_i = [\textstyle\sum_j W_{j1}^{\lambda 1(i)}\, W_{j2}^{\lambda 2(i)} \ldots W_{jq}^{\lambda q(i)}\, d_{ij}^{\beta(i)}]^{-1} \tag{14}$$

In this model, each parameter is specific to origin i so that the spatial distribution of a parameter can be mapped to examine any spatial patterns in the behavior represented by that parameter. Interestingly, there is a long history of estimated origin-specific distance-decay parameters exhibiting spatial patterns in which centrally located origins have less negative parameter estimates than peripheral ones (Fotheringham 1981). The interpretation of such a pattern has provoked intense debate at times, but it appears that the distribution is due to a mis-specification bias which can easily be eliminated by the addition of a variable measuring the proximity of a destination to all the other destinations (Fotheringham et al. 2000, chap. 9). The models that result are termed 'competing destinations models' and have been shown to remove the previously observed spatial pattern of origin-specific, distance-decay parameters. Fotheringham and O'Kelly 1989). It is interesting to note that the mis-

specification of spatial interaction models became evident only when local forms of the models were calibrated.

For completeness, a destination-specific form of the attraction-constrained model in Eqn. (8) is:

$$T_{ij} = B_j D_j V_{i1}^{\alpha 1(j)} V_{i2}^{\alpha 2(j)} \ldots V_{ip}^{\alpha p(j)} d_{ij}^{\beta(j)} \qquad (15)$$

where

$$B_j = [\textstyle\sum_i V_{i1}^{\alpha 1(j)} V_{i2}^{\alpha 2(j)} \ldots V_{ip}^{\alpha p(j)} d_{ij}^{\beta(j)}]^{-1} \qquad (16)$$

In this model, parameters are obtained that are specific to each destination *j*.

6. *Theoretical Foundations of Spatial Interaction Models*

Although spatial interaction models generally perform well empirically and are generally assumed to be reasonable statements of spatial behavior, the models have been criticized for not having a sound theoretical base. The 'Holy Grail' of spatial interaction modeling has therefore been to provide a commonly acceptable theoretical derivation of spatial interaction models. Four main frameworks have been postulated for this purpose and in chronological order these are: (a) spatial interaction as *social physics*; (b) spatial interaction as *statistical mechanics*; (c) spatial interaction as *aspatial information processing*; and (d) spatial interaction as *spatial information processing*. It is not the place to describe each of these theoretical frameworks in detail (see Fotheringham et al. 2000, chap. 9, for such a description). However, a very brief discussion is given to demonstrate the progress that has been made in this area and how the modeling framework has moved far away from its original reliance on an analogy to laws of physical gravitation.

Attempts at understanding regularities in patterns of spatial flows began with the observation that the movement of people between cities was analogous to the gravitational attraction between solid bodies. That is, greater numbers of migrants were observed to move between larger cities than smaller ones, *ceteris paribus*, and between cities that were closer together than between cities that were farther apart, *ceteris paribus*. This led to the proposal of a simple mathematical model to predict migration flows between origins and destinations based on the Newtonian gravity model. However, there was no theoretical justification for this form of model until Wilson derived the family of spatial interaction models from entropy-maximizing principles (Wilson 1967). However, this framework, although a catalyst for a great volume of research in spatial interaction modeling, was behaviorally difficult to justify and was superseded by McFadden's random utility maximizing principle (1974). This linked spatial interaction modeling with the process of spatial choice (see *Spatial Choice Models*) although the spatial choice process it depended on was developed for aspatial choice processes and did not always fit well into a spatial context (Fotheringham and O'Kelly 1989). Consequently, the latest theoretical framework in which spatial interaction models have been embedded is that of *spatial* information processing (Fotheringham et al. 2000, chap. 9 1992). This has led to the development of the new forms of spatial interaction models mentioned in Sect. 4 termed 'competing destinations models'.

Spatial interaction models are the products of a spatial theoretical framework that incorporates concepts from economics, such as random utility maximization, and psychology, such as spatial cognition and information processing (see *Spatial Cognition*; *Behavioral Geography*). However, the models describe the constraints of space on important aspects of human behavior and hence are quintessentially geographic.

See also: Location Theory; Market Areas; Spatial Optimization Models; Transportation Geography

Bibliography

Boyle P J, Flowerdew R, Shen J F 1998 Modelling inter-ward migration in Hereford and Worcester: the importance of housing growth and tenure. *Regional Studies* **32**: 113–32

Flowerdew R, Salt J 1979 Migration between labour market areas in Great Britain, 1970–1971. *Regional Studies* **13**: 211–31

Fotheringham A S 1981 Spatial structure and distance-decay parameters. *Annals of the Association of American Geographers* **71**: 425–36

Fotheringham A S, Brunsdon C, Charlton M 2000 *Quantitative Geography: Perspectives on Spatial Data Analysis*. Sage, London

Fotheringham A S, Curtis A 1992 Encoding spatial information: the evidence for hierarchical processing. In: Frank A U, Campari I, Formentini U (eds.) *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, Lecture Notes in Computer Science. Springer-Verlag, Berlin, pp. 269–87

Fotheringham A S, Dignan T 1984 Further contributions to a general theory of movement. *Annals of the Association of American Georgraphers* **74**: 620–33

Fotheringham A S, O'Kelly M E 1989 *Spatial Interaction Models: Formulations and Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands

Fotheringham A S, Trew R 1993 Chain image and store choice modeling: the effects of income and race. *Environment and Planning A* **25**: 179–96

Mayhew L D, Gibberd R W, Hall H 1986 Predicting patient flows and hospital case-mix. *Environment and Planning A* **18**: 619–38

McFadden D 1974 Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed.) *Frontiers in Econometrics*. Academic Press, New York, pp. 105–142

Meyer R J, Eagle T C 1982 Context-induced parameter instability in a disaggregate-stochastic model of store choice. *Journal of Marketing Research* **19**: 62–71

O'Kelly M E 1986 Activity levels at hub facilities in interaction networks. *Geographical Analysis* **18**: 343–56

Olsson G 1970 Explanation, prediction and meaning variance: an assessment of distance interaction models. *Economic Geography* **46**: 223–33
Southworth F 1983 Temporal versus other impacts upon trip distribution model parameter values. *Regional Studies* **17**: 41–7
Wilson A G 1967 Statistical theory of spatial trip distribution models. *Transportation Research* **1**: 253–69

A. S. Fotheringham

Spatial Labor Markets

Everyone has seen headlines that report declining or increasing unemployment. If the articles that ensue do not report the unemployment data for distinct regions, the headlines are misleading. The idea that processes leading to unemployment are similar in all regions may be assumed, but it is not tenable. The labor market of a state cannot be thought of as a uniform entity; rather it comprises different spatial entities called spatial labor markets. The concept of a spatial labor market reflects the realities that mobility incurs costs, that information about a national labor market is not equally available everywhere and that people are not always willing to give up their home in order to take up a job where they are needed.

1. *Definition*

A spatial labor market is defined as an imaginary and sometimes also actual meeting place for the supply of and demand for labor. Thus, a spatial labor market is both a spatial entity and a mediation agency for paid work. Or, put another way: a spatial labor market is always both a territorial entity and a social institution that serves to distribute labor across the labor supply.

The dual nature of a spatial labor market can still be seen in many markets that emerge from time to time, as well as in more permanent markets. Markets are both actual meeting places for those offering and seeking goods and institutions that serve to set prices and distribute goods separated from the local and regional context.

2. *How a Spatial Labor Market Works*

From a functional point of view a spatial labor market thus corresponds to the general model of labor markets limited by the spatial range over which supply and demand can intersect and try to satisfy their requirements. It is the friction of distance that creates local and regional labor markets by limiting the spatial range over which supply and demand can intersect.

Nevertheless, the components and the mechanism of the spatial labor market corresponds to that of the nonspatial model. The labor supply offers its readiness to work and its qualifications while seeking to achieve high wages, job security, and career expectations. The labor demand, i.e., business people, entrepreneurs, public bodies, and representatives of all kinds of associations, seek this readiness to work and specific qualifications and endeavor on grounds of costs to pay low wages in return, while promising job security and career expectations only to certain employees. This searching process is accompanied on both 'sides' by organizations that have, through collective agreements, created the framework conditions which markedly shorten the individual negotiation process. For most workers the room for individual negotiations is narrowly restricted by wage agreements arrived at by collective bargaining, minimum wages, and labor laws.

The extent of labor supply and demand is not fixed. The number of people looking for work and in demand depends on many factors. On the supply side, the number of employable people of working age (usually between 15 and 60 years of age) is determined by the size of the population living in the region and its age structure, regardless of whether the people thus defined as 'employable' are in fact seeking work. The number of employable persons is increased directly by permanent, temporary, and pendular migration and natural population trends (fertility and mortality).

The proportion of employable people actually seeking work (workforce), and thus calculable as part of the labor supply, varies greatly. Factors determining whether people are seeking work include educational and professional qualifications, family status, gender, the regional infrastructure, and sociocultural norms. The proportion of people seeking work can be measured by the participation rate. This is obtained by dividing the number of persons employed (or unemployed) by the number of population. Finally, those seeking work offer their labor for a particular period of time, which is dependent on personal preferences and legal norms. In multiplying the labor supply by the period of work offered, it is possible to calculate the total volume of labor offered.

Set against this is the volume of labor demand. This can be calculated from the number of jobs available in a region. This is heavily dependent on the economic situation with an increase in the demand for goods and services tending to lead to an increase in the demand for labor, unless production is increased by new ways of organizing work and technical advances, which lead to a leveling-off in the volume of labor demand. The volume of labor demand is also dependent on the working hours, with a reduction and constant degrees of flexibility in the framework times clearly tending to an increase in demand.

The extent of the supply of and demand for labor is not fixed. On a spatial labor market the volume of labor supply and demand can fluctuate more strongly

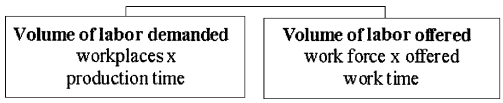

Figure 1
Supply and demand at the labor market (after Fassmann and Meusburger 1997)

than on a national labor market. As well as those mentioned above, other factors that come into play are the effects of the varied accessibility of the spatial labor market. Thus, the construction of roads and of rail-borne means of mass transport can affect the accessibility of a spatial labor market and, thus, the labor supply, just as the establishment or transfer of a firm can increase or decrease the demand (Fig. 1).

3. Size and Demarcation of a Spatial Labor Market

There is a consensus that a spatial labor market is based on distance which creates local and regional markets and leads to geographical subdivisions of a nonspatial labor market. To this extent the definition is clear. But it is still not clear how large a spatial labor market is and by what yardsticks and criteria it is demarcated. Two different demarcation approaches are possible.

The first approach involves using given administrative units (or aggregation of administrative units). This means that spatial labor markets can be limited by administrative boundaries. If one asks what is the labor market situation in Munich the answer will contain information on unemployment and employment in the administrative area of the city regardless of the fact that the functional area is much broader. The same is true if, for example, the labor market of old industrial areas, the rural labor market or the urban labor market is in the center of interest. The space is always a homogeneous and not a functional unit.

This contrasts with the prominent position of the functional integration that is found in the second approach to the demarcation of labor markets. The demarcation of a regional labor market depends on the commuter field (Eckey et al. 1990) and thus the accessibility and integration of a labor market center with its own hinterland. This is a common method for defining regional labor markets in the literature and considered the 'best' approach. Self-contained labor market areas, functionally demarcated on the basis of commuter data, would be an ideal model for taking stock of regional labour markets (Bröcker 1988).

Both approaches—the spatial labor market as an administrative (homogeneous) unit respective a functional unit—are overlooking important points. Theoretical and empirical research came to the conclusion that the idea of demarcation of a spatial labor market which is valid for all societal groups and for all times is wrong (Peck 1989, Hanson and Pratt 1995, Fassmann and Meusburger 1997). Spatial labor markets are always time-specific and constructed. Spatial labor markets do not exist as *a priori* territorial entities which require only a suitable analysis to discover, but are socially-constructed spatial entities depending on gender, occupation, and sociocultural norms. The functional labor markets for the male population is significantly different to that of the female, the societal acceptance of commuting to work on a daily or weekly basis over long distances varies from country to country and region to region. An administrative boundary can also change, just like the catchment area of a central location. The transport infrastructure is as significant in the formation of spatial labor markets as are company location decisions and the effect of institutions. The establishment of a large new company or the closure of an existing one can—depending on the analytical criteria— alter decisively the flow of commuters. Similarly, the opening up of previously closed borders and currency fluctuations can quickly redirect commuter flows in border regions. Labor market regions thus defined must therefore be re-examined every few years and possibly redemarcated.

The following is a further principal argument against the method of functional demarcation of a labor market region by the flow of commuters. Commuter data always underestimate the real extent of a regional labor market because commuter data are based on figures for the actual places of residence and work. The fact that someone lives in the commuter catchment area and looks for a job there does not mean, of course, that the person concerned has always lived there. It is equally possible that they migrated from a long distance in order to take up a job. The demarcation of a spatial labor market on the basis of commuter data is not, however, based on the actual place of residence and ignores previous migration. This is of major significance in assessing the extent of a regional labor market.

Another problem in defining a spatial labor market by using commuting data lies in the suggested precision that is, however, not guaranteed. Especially at the margins of a spatial labor market, it is not clear if a small spatial unit belongs to one or the other spatial labor market. For example, when 20 percent of the workforce commutes into one center and 20 percent into another, the classification ceases to be clear.

A continuing and as yet unsolved problem, however, is still the question of the size of a region integrated as a result of spatial interactions. Depending on the research question and the purpose chosen, there exist local, regional, and national labor market centers with corresponding catchment areas. A national labor market can be broken down into regional labor markets and the latter can be further split up into 'subregional' or local labor markets. Precise bound-

aries demarcating where a local labor market becomes a regional one and the yardsticks suitable for defining the regional labor market are lacking. Those that believe that universally-valid demarcation criteria exist are, however, mistaken.

4. Empirical Findings

The difficult question of demarcation should not obstruct the main finding here. Those that would like to distance themselves from the abstract model of the labor market of the economy and move towards a more realistic position are unavoidably approaching the model of a spatial labor market because labor markets are always spatial phenomena and research cannot therefore avoid taking the spatial dimension into consideration.

In any case, empirical surveys have drawn attention to the systematic differences between specific regional labor markets and different types of spatial labor markets. As an example of that one can ask how spatial labor markets are changing when comparing large metropolises, with the medium-sized centers and the rural areas. Some results will be presented (Fassmannn 1993, Fassmann and Meusburger 1997). Analogous to and very closely connected with the idea of central place theory is the distribution of highly qualified and leading positions in the secondary and tertiary sectors. Large towns offer the highest-ranking central services and connected with them a specific range of jobs. Public sector institutions such as ministries, the highest courts, museums, universities, and the head offices of major companies, together with all the service industries connected with them, are located in the respective 'primate cities' of the national network of cities. This phenomenon is being strengthened by the decline of manufacturing jobs. Industrial cities are being transformed into information and services centers with a consequent change in the demand for labor (Table 1).

The wide and attractive range of jobs offered in metropolises has led to a marked increase in employment. In metropolises the proportion of those seeking and taking up work is higher than on other labor markets. Women in particular are attracted by the demand from service industry companies. They benefit particularly from the expansion of the tertiary and quaternary sectors and the transfer and reduction of manufacturing jobs.

With the decline in manufacturing jobs and the dominance by now of highly qualified tertiary and quaternary jobs in central locations, the labor markets of large towns and cities experience an increase in well-paid and highly-qualified workers. Their purchasing power and consumer behavior stimulate the growth of less qualified service industry jobs at the other end of the labor market. Laundries, restaurants, and retail shops with long opening hours make new lifestyles possible and family structures that are growing in number. These phenomena are not found to the same extent in all towns and cities, and depend on the political regulation of the labor market and on the accessibility of cheap and mostly immigrant labor. The tendency appears to be in this direction, however, and was also dealt with thoroughly in the discussion of global cities (Sassen 1994).

The generally attractive range of jobs in large towns does not prevent high unemployment but, paradoxically, contributes to it. Unlike on the labor markets of small towns, old industrial areas, and rural areas the attractive supply of jobs leads to immi-

Table 1
Geographical features of labor markets in relation to the settlement system

Features	Rural areas	Central places	Metropoles
Size of labour	Severe job shortage negative commuter flow	Job surplus balanced commuter flow	Large job surplus positive commuter flow
Unemployment	Moderate to small and seasonal peaks	Tremendous spatial variation (old industrial areas vs. new growth centers)	Relatively high but more polarized
Structure of labour, supply and demand	Low demand, concentration on a few sectors	Varying demand, according to structure of central locations	Varying and polarized demand in manufacturing and services sectors
	Moderate labor supply; especially low female supply	Structure-specific labor supply	High labour supply; through women and immigrants
	Low labor qualifications (industrial labourers, agriculture)	Medium-level qualifications; skilled workers; middle-level white-collar staff	Polarises qualification (specialists and lowly-qualified service workers)

Source: Fassmann 1993

gration, commuter migration, and to high levels of employability among the population which persists even in times of temporary unemployment. Unemployment does not lead to employable people retreating into the reserve army. They continue to actively seek work, but are registered as unemployed.

Characteristic of the urban labor market, therefore, is the wide range of opportunities. For both the highly-qualified and the low-qualified immigrants and natives, men and women, old and young, the cities offer the best chance of finding work. In addition to this, moreover, the chances of changing jobs, of mobility, and social promotion are also greater. In short, the urban-labor market is a career-labor market. To achieve certain positions on the labor market it is almost essential to commute or to move to the urban centers. The 'reward' for this is a labor market open-ended at the top. If there were any truth in the myth that everybody could reach every social stratum, this would apply in the first instance to the urban labor market. As one goes further down the settlement structure to smaller and medium-sized towns, the structural features of the labor market change characteristically.

The number of high-ranking services declines, together with the proportion of highly-qualified and leadership jobs. The employment spectrum becomes narrower and is restricted to a middle-ranking sector. In particular, in those industrial towns not yet affected by internationalization and company transfers, traditional skilled jobs dominate the job structure at the middle level. Compared with the polarized job spectrum of the large towns, that of the small and medium-sized towns is markedly more homogeneous, with fewer highly-qualified jobs but also fewer jobs for the lowly-qualified in the industrial trade and service sectors.

It is women who are affected primarily by the increasing concentration in the number and quality of jobs. They are less often in work and more often drop out of paid work at certain stages of their lives (e.g., after marriage or after childbirth). One result is that they reduce the labor supply and prevent a marked rise in unemployment although there is tremendous variation in this phenomenon across countries. The tendency of leaving the labor market after marriage or after childbirth varies according to sociocultural norms and stages of modernization. A further step down the settlement hierarchy finally leads one to the rural area, which no longer offers any central services for the surrounding area. As a result, the job spectrum is further restricted to just a few sectors: agriculture, the construction industry, tourism, and industrial jobs in wood processing, the production of building materials, textiles, and food. On the whole these tend to be unskilled and poorly paid jobs. Only a small proportion of the population of working age has a job on the local labor market. The locations lack the conditions for higher-ranking and highly qualified jobs. Service industry companies and modern industrial companies that amount to more than just extended workbenches require transport links and the advantages of an agglomeration.

The lack of jobs does not, as might be expected in theory, lead to a constantly high level of unemployment but, on the contrary, to average and in some regions even markedly low levels of unemployment. Part of the labor supply withdraws from the labor market and at first does not even try to find paid work. Women, who cannot undertake long commuter journeys or any journeys at all, because of the lack of childcare facilities, give up looking for work. This decision is easier when wage levels are low and compare unfavorably with the costs of commuting and childcare. In addition women take on minor and part-time activities or withdraw into unpaid housework.

Those wishing to have a career and attend a higher level school or university and then look for a suitable job at the end of their educational training must commute or leave the area, because in rural areas qualified jobs are in short supply. The establishment of school centers in rural areas has changed this situation only to a limited extent and sometimes has even been counterproductive. Those graduates not finding a job teaching in their hometown have to move or be satisfied with a job below their qualifications. Establishing school centers without creating new jobs has therefore often led to an erosion of qualifications in rural areas and can thus to be seen as a failure of regional policy.

5. *Conclusion*

The demarcation and analyses of spatial labor markets leads to a differentiated and more realistic picture of the dynamic and structures of a national labor market. Trends observed in recent years on the labor market differ fundamentally from those on the labor markets in metropolises, old industrial areas, rural areas, peripheral border areas, and in the suburbs. In each case there are considerable differences regarding those population groups which were unemployed, and to what extent, and which groups found work or were forced out of active life into the silent reserve army. This draws attention to the need for a regional labor market policy that takes account of the specific causes of unemployment, underemployment and wage differentials among regions, while seeking solutions that are spatially acceptable.

See also: Employment and Labor, Regulation of; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Supply; Social Stratification; Unemployment: Structural; Uneven Development, Geography of; Urban Activity Patterns

Bibliography

Bröcker J 1988 Regionale Arbeitsmarktbilanzen 1978–1984. Methoden und Ergebnisse. *Raumforschung und Raumordnung* **3**: 87–97

Eckey H-F, Horn K, Klemmer P 1990 *Abgrenzung von regionalen Diagnoseeinheiten für die Zwecke der regionalen Wirtschaftspolitik. Gutachten im Auftrag des Untersuchungsausschusses der Gemeinschaftsaufgabe 'Verbesserung der Regionalen Wirtschaftsstruktur.'* N. Brockmeyer, Bochum, Germany

Fassmann H 1993 *Arbeitsmarktsegmentation und Berufslaufbahnen. Ein Beitrag zur Arbeitsmarktgeographie Österreichs. Beiträge zur Stadt- und Regionalforschung 11.* Verlag der Österreichischen Akademie der Wissenschaften, Vienna

Fassmann H, Meusburger P 1997 *Arbeitsmarktgeographie.* Teubner, Stuttgart, Germany

Fischer M M, Nijkamp P 1987 *Spatial labour market analysis: relevance and scope.* In: Fischer M M, Nijkamp P (eds.) *Regional Labour Markets. Contribution to Economic Analysis.* Elsevier Science, Amsterdam, New York, Oxford, Tokyo, pp. 1–26

Hanson S, Pratt G 1995 *Gender, Work and Space.* Routledge, London

Pearce D (ed.) 1986 *The MIT Dictionary of Modern Economics.* MIT Press, Cambridge, MA, pp. 233–34, 238

Peck J A 1989 Reconceptualizing the local labour market: space, segmentation and the state. *Progress in Human Geography* **3**: 42–61

Sassen S 1994 *Cities in a World Economy.* Pine Forge Press, Thousand Oaks, CA

H. Fassmann

Spatial Memory Loss of Normal Aging: Animal Models and Neural Mechanisms

This article will consider the effects of normal, age-related changes on certain forms of memory. The term 'normal aging' is used to refer to developmental processes that are expected to occur in organisms near the end of their lifespan. Alzheimer's disease and other neuropathological conditions in humans, for example, are not 'normal' or inevitable parts of the aging process, and thus are not considered here. The focus here will be on the identification of brain changes responsible for age-related learning and memory changes, which can be understood by examining any number of mammalian species. The motivation for this approach is ultimately the development of methods that can attenuate age-associated cognitive decline, thereby improving life quality of the elderly.

1. *The Hippocampus as a Model Brain Structure in which to Study Age-related Memory Deficits*

The way in which new events are learned and remembered changes over the lifespan. A great deal of research on learning and memory has focused on the hippocampus (Fig. 1), a structure implicated in these processes (see *Memory and Aging, Neural Basis of*). O'Keefe and Nadel (1978) proposed that the hippocampus functions as a cognitive map, and, in all mammals studied to date, optimal functioning of this structure is required for acquisition of the spatial knowledge required for effective navigation. The hippocampus is likely to participate in more than just spatial memory, but its role in spatial memory is emphasized here to constrain the topic of review (Squire 1992). It is certainly the case that in either humans or rats, extensive hippocampal damage prevents formation of new spatial maps that allow accurate large-scale navigation. It is these types of observations that suggest that the computation the hippocampal system performs is likely to be conserved across mammals.

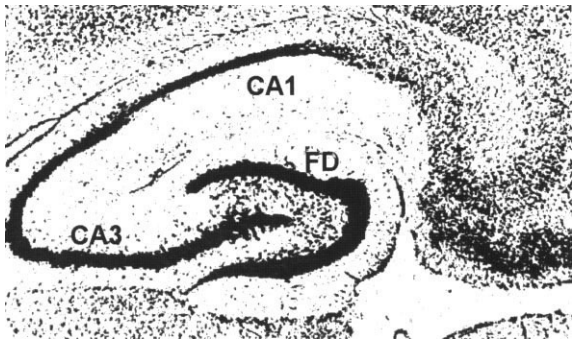

Figure 1
Coronal section through the rat hippocampus showing CA1 and CA3 pyramidal cell layers as well as the granule cell layer of the fascia dentata (FD)

1.1 *Changes in Navigational Behaviors in Normal Aging*

The acquisition of new spatial information is defective in old humans, old monkeys, and old rats. For example, Newman and Kaszniak (2000) tested healthy younger and older adults' ability to accurately locate a pole in relation to multiple cues present in a large tent-like environment. The older adults were not as accurate as were younger adults in placing the pole in the correct position, suggesting a difference in spatial knowledge between groups. Rapp et al. (1997) tested young and old monkeys in an eight-choice foraging task in which reward could be obtained only once at any location by the monkey. When the spatial memory demands of the task were made more difficult by imposing greater delays between choices, older monkeys became increasingly impaired. For rats, the circular platform task has been used to test spatial navigation ability. The object of the task is to navigate to the one place where it is possible to enter a dark chamber away from the brightly-lit open surface of the platform. The

consistent finding is that old rats do not remember this preferred dark chamber location from day to day as well as do younger rats (Barnes 1979). Taken together, the observations suggest that older mammals remember spatial information less accurately than do their young counterparts, and most often these spatial memory problems can be separated from other age-related sensory changes that might have a negative impact on performance.

2. *Changes in Hippocampal Connectivity and Functional Responsiveness*

Because normal aging results in hippocampal-dependent memory changes, changes in hippocampal structure and function over the lifespan might also be expected. In general, however, there appears to be rather little change in cell numbers, and only rather subtle and region-specific changes in cell function.

2.1 *Preservation of Hippocampal Cell Numbers and Function*

In contrast with some earlier anatomical studies, modern stereological cell counting methods have revealed no significant numerical differences in the major cell types within the hippocampus across the lifespan of humans (West 1993), monkeys, or rats. Similarly, studies of basic cellular neurophysiology suggest that the biophysical properties of hippocampal cells are largely preserved across the lifespan. This includes measurements such as membrane potential, cell input resistance, action potential heights, and membrane time constants measured *in vitro*, as well as average spontaneous cell firing rates measured in the awake, behaving state. Thus, normal aging does not result in widespread cell loss, or deterioration in cell viability and function (Barnes 1994).

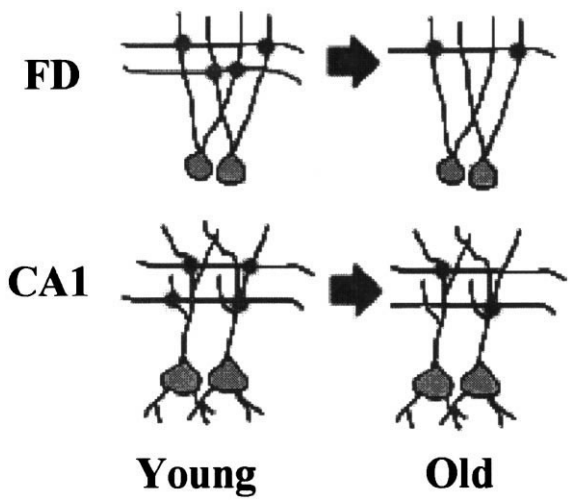

Figure 2
Schematic of age-related changes in granule cells of the fascia dentata (FD) and pyramidal cells in CA1 of rat hippocampus. Old rats show fewer axon collaterals and synapses, but those synapses that remain are more powerful in FD. In CA1 old rats show fewer synapses, but no change in synaptic strength

2.2 *Specific Age-related Hippocampal Structural and Functional Changes*

The age-related changes that do occur appear to be in connectivity and functional responsiveness, and the pattern of change is specific to the region examined. For example, from anatomy we know that there are fewer synaptic contacts from the main source of input to granule cells of the hippocampus (Fig. 1) from the entorhinal cortex (Geinisman et al. 1995, Fig. 2). From electrophysiology we know that there is pruning of these axons; however, the synaptic contacts that remain in old rats are more powerful (Barnes 1994, Fig. 2). This change in synaptic strength appears to help keep the probability of granule cell output relatively constant, in spite of the extensive deafferentation, and suggests a compensatory process at this synapse during aging. By contrast, in adjacent hippocampal regions to the granule cells (CA3 and CA1 in Fig. 1) there is no pruning of CA3 pyramidal cell axons that project to the pyramidal cells of CA1 (Fig. 2). The anatomy and physiology suggest that there are, however, fewer functional contacts per axonal branch in old rats. Interestingly, the efficacy of these individual synaptic contacts in CA1 does not change with age (Fig. 2), suggesting that functional compensation is not an obligatory process in response to age-related deteriorative change (Barnes 1994), but occurs selectively at only certain connections.

3. *How Does Aging Affect the Plasticity Characteristics of the Hippocampus?*

The changes in connectivity and cellular function during aging discussed above could impact the way in which information is stored in the hippocampal network. Hebb (1949) was among the earliest to propose that synaptic modification must take place during the acquisition and storage of associative memories. Modification of this type was described at the perforant path—granule cell synapse in the hippocampus by Bliss and Lomo (1973) and McNaughton et al. (1978) verified that it was, indeed, an associative process. Since that time an enormous body of empirical work has arisen describing the physiology, pharmacology, and molecular biology underlying this process known as long-term potentia-

tion (or LTP). Although there is no direct proof that LTP induced artificially at synapses in the brain reflects the process the nervous system normally uses to store information, there is a wide range of evidence consistent with this hypothesis. Thus, it has been important to examine the effect of normal aging on this process.

3.1 Hippocampal Plasticity Deficits in Aging

Although LTP can be robustly induced at synapses in the aging hippocampus (Landfield et al. 1977), such modification is more difficult to achieve in older rats. The factors contributing to these age-related changes include synapse sparsity, reduced glutamate receptor numbers, and reduced temporal summation of incoming input. Even if equivalent LTP induction levels are achieved, LTP can *decay* faster in old rats (Barnes 1979). Furthermore, processes that decrease synaptic efficacy, such as LTP reversal and long-term depression, are *more* effective in older animals (Norris et al. 1996) Importantly, the decay rate of LTP in hippocampus of rats correlates with spatial memory accuracy on the circular platform task, that is, old rats with the fastest decaying LTP also show the poorest retention of spatial information. These data suggest that the most promising candidate mechanism for information storage in the nervous system is compromised in old rats. Furthermore, the data suggest that if we understood the mechanism of faster decay of LTP, one might better understand why older animals show faster forgetting. In fact, a number of laboratories are investigating pharmacological interventions for extending LTP decay rates, in an effort to modify forgetting rates in aging.

4. How Do Changes in Plasticity Mechanisms Impact the Information Processing Characteristics of Hippocamal Ensembles?

It has recently become possible to record significant numbers of single hippocampal cells simultaneously when rats are awake, and free to behave (Wilson and McNaughton 1993). With this type of electrophysiological recording, the firing patterns of ensembles of cells are monitored without the use of artificial stimulation. Rather, cell firing occurs in response to the animal's behavior and sensory input, and thus provides a complementary window for viewing experience-dependent changes in brain function. Single cells in the hippocampus increase their firing rates when an animal is in a particular place within an environment, which is why O'Keefe first called these cells *place cells* (O'Keefe and Dostrovsky 1971). The area of space over which a given cell fires is called the cell's *place field*, and the firing pattern of many cells over a given environment is hypothesized to represent a cognitive map of that environment (O'Keefe and Nadel 1978). In fact, it is possible to reconstruct the animal's location, simply from knowing the pattern of hippocampal cell activity (Wilson and McNaughton 1993). These place field distributions are commonly referred to as *maps*. With repeated exposure to an environment, it is hypothesized that these maps can become stabilized and bound to features of the environment as a consequence of LTP at the synapses conveying external sensory information to the hippocampus. Such binding could help ensure accurate map retrieval. Two main findings from age comparisons using ensemble recording methods are discussed below, both suggesting ways in which faulty LTP mechanisms may impact the information processing characteristics of hippocampal ensembles.

4.1 Experience-dependent Place Field Expansion Deficits in Aging

There is a plastic change in the pattern of place cell discharge characteristics that occurs as a consequence of experience, strongly suggesting that synaptic modification is taking place in the system. This empirical observation was predicted by theoretical considerations of route learning, the fundamental idea for which goes back to Hebb's ideas on phase sequences, or linked cell assemblies. The LTP mechanism has an intrinsic asymmetry that predicts that an expansion of place fields should occur with repeated traversal of a given route. Such place field expansion, and a shift in the center of mass of place fields towards the origin of a route, has been verified experimentally (Mehta et al. 1997). Moreover, glutamate receptor blockade that blocks LTP induction also blocks this experience-dependent place field expansion effect. Because LTP mechanisms are disrupted in old rats, it was predicted that normal aging may alter place field expansion. This was confirmed by Shen et al. (1997) who showed that the expansion of place field size that occurs in young rats after a few traversals of a route is much less robust in old rats. The deficit in experience-dependent place field expansion in old rats is consistent with the likelihood of a failure of effective LTP mechanisms in these old animals.

4.2 Map Retrieval Deficits in Aging

Another example of how deficits in hippocampal synaptic plasticity may impact spatial cognition can be illustrated by a comparison of hippocampal map retrieval in young and old rats. When ensembles of hippocampal cells are recorded in young rats, different place cells fire selectively in the environment through which the rat moves. If the young rat is removed from a familiar environment, and then brought back later to

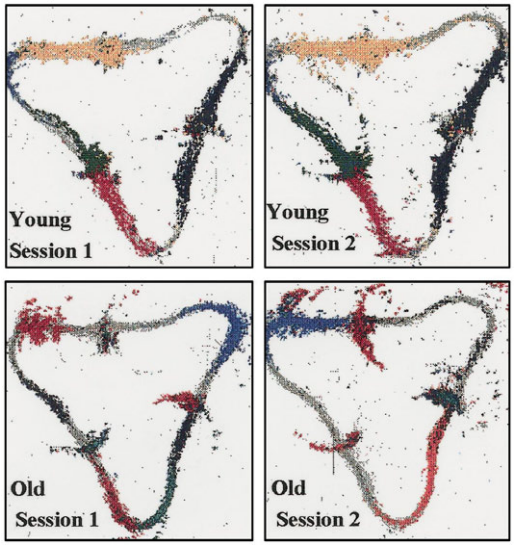

Figure 3
The firing of different hippocampal cells are shown in different colors, only a few of the simultaneously recorded cells are plotted for clarity. This illustrates hippocampal map retrieval accuracy on a triangular maze in a young rat on two different recording sessions, and a map retrieval error between sessions in an old rat

the same room, the same place field distribution, or same map, is retrieved at this later time (Fig. 3). On about two thirds of occasions when old rats are removed and returned to the same familiar room, they retrieve the same map, just like young rats. About one third of the time, however, a different place field distribution appears, as though the old rat had failed to retrieve the original map (Fig. 3). Thus, possibly because of a deficient binding to environmental cues or weak associative connections within the map itself, when old rats come back to the same situation a completely different pattern of cell activity can emerge (Barnes et al. 1997). This may well be associated with a place recognition failure. In addition to the map retrieval deficits described above, Tanila et al. (1997) have also shown that memory-impaired old rats do not create new maps as readily as do young rats when features are changed in an environment. Thus, it seems possible that altered LTP-like plasticity mechanisms may be partly responsible for the greater tendency of old organisms to become lost.

5. *Conclusions*

At the cellular level, normal aging leads to changes in hippocampal connectivity and plasticity mechanisms that are specific to different subregions, but leaves many biophysical parameters of neurons relatively unaffected. At the system level, aging also leads to changes in hippocampal network dynamics that result in nonlinear failures in information retrieval. These age-related changes in brain structure and function are likely to contribute to moderate, but clear changes in spatial learning and memory observed during normal aging.

See also: Aging Mind: Facets and Levels of Analysis; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Education in Old Age, Psychology of; Lifespan Theories of Cognitive Development; Memory and Aging, Cognitive Psychology of; Memory and Aging, Neural Basis of

Bibliography

Barnes C A 1979 *Journal of Comparative and Physiological Psychology* **93**: 74–104
Barnes C A 1994 *Trends in Neuroscience* **17**: 13–8
Barnes C A, Suster M S, Shen J, McNaughton B L 1997 *Nature* **388**: 272–5
Bliss T V P, Lomo T 1973 *Journal of Physiology* **232**: 331–56
Geinisman Y, de Toledo-Morrell L, Morrell F, Heller R E 1995 *Progress in Neurobiology* **45**: 223–52
Landfield P W, Rose G, Wohlstadter T C, Lynch G 1977 *Journal of Gerontology* **32**: 523–33
McNaughton B L, Douglas R M, Goddard G V 1978 *Brain Research* **157**: 277–93
Mehta M R, Barnes C A, McNaughton B L 1997 *Proceedings of the National Academy of Science USA* **94**: 8918–21
Newman M C, Kaszniak A W 2000 *Aging Neuropyschology and Cogition* **7**: 86–93
Norris C M, Korol D L, Foster T C 1996 *Journal of Neuroscience* **16**: 5382–92
O'Keefe J, Dostrovsky J 1971 *Brain Research* **34**: 171–5
O'Keefe J, Nadel L 1978 *The Hippocampus as a Cognitive Map*. Clarendon Press, Oxford, UK
Rapp P R, Kansky M T, Roberts J A 1997 *NeuroReport* **8**: 1923–8
Shen J, Barnes C A, McNaughton B L, Skaggs W E, Weaver K L 1997 *Journal of Neuroscience* **17**: 6769–82
Squire L R 1992 *Psychological Review* **99**: 195–231
Tanila H, Shapiro M, Gallagher M, Eichenbaum H 1997 *Journal of Neuroscience* **17**: 5155–66
West M J 1993 *Neurobiology of Aging* **14**: 287–93
Wilson M A, McNaughton B L 1993 *Science* **261**: 1055–8

C. A. Barnes

Spatial Mismatch

Spatial mismatch describes a broad set of geographical barriers to employment that stem from a disparity, or mismatch, between where people live and where appropriate job opportunities are located. A more specific form of spatial mismatch, the *spatial mismatch*

hypothesis (SMH), argues that the combined effects of residential segregation and economic restructuring limit geographical access to employment opportunities for black, inner-city residents. This article emphasizes geographical dimensions of the SMH, highlighting the changing spatial distributions of residences and job opportunities in cities, the measurement of geographical separation between the two, the sociospatial processes that underpin spatial mismatch, and policy implications.

Spatial mismatch is a historically specific concept that emerged in the 1960s amid widespread concern about racial discrimination and urban poverty. Although he never used the term 'spatial mismatch,' John Kain (1968) introduced the concept in analyzing the combined effects of residential segregation in housing markets and the postwar suburbanization of jobs on black unemployment. Poor geographical access to jobs, especially jobs in manufacturing, was said 'to reduce employment opportunities of Negroes who are already handicapped by employer discrimination and low levels of education' (p. 196). Although some subsequent researchers have defined spatial mismatch in narrow geographical terms, as long distances to jobs or small numbers of locally available jobs, Kain's research was notable in exploring the intersections between spatial and social barriers to employment. He considered how transportation access, employer discrimination, skills, and social networks affect workers' access to employment. These themes continue to the present in the SMH literature.

Kain's research stimulated discussion and debate during the early 1970s. After that, the field languished for a decade until William Julius Wilson's research on the urban underclass brought spatial mismatch back to the forefront (Wilson 1987). Since then the SMH has attracted research attention in a wide range of social science disciplines.

1. *Evidence of Spatial Mismatch*

The great appeal of the SMH is that it connects two prominent features of postwar American cities: the strong segregation of residential areas by race and ethnicity and the movement of jobs, especially jobs in manufacturing, to the suburbs and overseas. But do these processes necessarily result in spatial mismatch? Many central cities are losing employment, but they are also losing population (Stoll 1999). Thus, the balance of jobs to resident workers has remained relatively constant. Although the SMH emphasizes city–suburban differences, finer-grained patterns of residential and work location are much more relevant to job access. Cities and suburbs are not homogeneous; within each, African-Americans and other minorities are concentrated in a few areas.

Research using diverse indicators of geographical access to employment, in a variety of US cities, finds support for the SMH at the localized, intra-urban scale. Within central cities, black neighborhoods generally have a lower ratio of jobs to resident workers than do white neighborhoods, though the findings are highly sensitive to the geographical areas considered (Holzer 1991). A better approach is to examine the availability of jobs within a reasonable commuting radius. Research in cities like Chicago and San Francisco finds that blacks are less likely to work in their local neighborhoods than are comparable whites, and that the density of job opportunities is lower in and near predominately black neighborhoods (Ihlanfeldt and Sjoquist 1998).

In the 1990s researchers have started to combine the power of geographic information systems with detailed urban databases to compute new measures of employment accessibility (Ong and Blumenberg 1998). Transportation routes and physical and social barriers to movement can be layered together to describe the influences on travel and interaction patterns.

Spatial mismatch is also apparent in commuting times and flows. African-American workers have longer commute times on average than do comparable white workers, and the differences persist after controlling for wages, mode of transportation and personal and household characteristics (McLafferty and Preston 1996). Black workers are also more likely to have 'constrained' work trips that involve long travel times to low-wage jobs (Johnston-Anumonwo 1997). The propensity to work in one's local neighborhood—local working—is also low in black neighborhoods (Immergluck 1998). Spatial mismatch, then, is evident in long work trips to jobs that provide little economic return.

Although the time and cost of commuting are important in their own right, they have several limitations as measures of spatial mismatch. By definition, the measures are only based on a selected group, namely, employed workers. This 'selection bias' implies that commuting indicators may not accurately represent geographical barriers to employment for the unemployed (Cooke and Ross 1999). One way to address this problem is to analyse changes through time as people and jobs move away from central city locations. Case studies reveal that when firms move to the suburbs, black workers are more likely to quit and less likely to relocate in order to gain access to a new suburban work site than are comparable white workers (Ihlanfeldt and Sjoquist 1998). Black workers are more fixed in their residential locations, indicating the long-lasting effects of racial segregation and discrimination.

There are also problems in interpreting commuting patterns. Do long work trips reflect individual choices to travel further for high-wage, high-quality employment, or do geographical constraints compel long work trips? Many commuting decisions result from a complex mix of choice and constraints, as people juggle personal and family needs within the spatial and

social bounds of daily life (Gilbert 1998, Hanson and Pratt 1995). These choices and constraints vary between men and women, highlighting the importance of gender in the SMH.

Attention is turning to the role of place in spatial mismatch and the important variations in spatial mismatch within and among cities (Holloway 1998). In the USA, spatial mismatch is primarily a large-city phenomenon that is most evident in cities like New York, Chicago, Atlanta, and Los Angeles. It is also more apparent in the 'rustbelt' cities of the Northeast and Midwest where economic restructuring and suburbanization are in full force (Ihlanfeldt and Sjoquist 1998). Little has been written about spatial mismatch in racially and ethnically diverse cities outside the USA, although there are hints of mismatch in cities like London and Berlin where ethnic minority populations cluster in areas of limited job opportunity.

2. *Spatial Mismatch and Labor Market Processes*

Spatial mismatch is rooted in the social, economic, and spatial processes that shape urban labor markets. These influence where people and firms locate, the channeling of workers into certain types of occupations, workers' ability to search for and commute to job opportunities, and firms' willingness to hire workers. Connecting spatial mismatch to employment outcomes requires an understanding of labor market processes, as well as an understanding of the changing geographical patterns of firms and people.

Urban labor markets are strongly differentiated according to gender, race, ethnicity, and class. This labor market segmentation concentrates blacks and Latinos in a narrow set of occupations, and the particular occupations vary with gender. Historically black men worked as laborers, in transportation, manufacturing, and certain service occupations, whereas black women were concentrated in domestic service. Although these occupations have changed over time, there continues to be strong segmentation by race and gender. Segmentation implies that groups of workers face different sets of potential job opportunities that are unevenly distributed within metropolitan areas. Thus, the spatial mismatch may vary by gender, race, or class. Researchers have only begun to tease out the interactions between segmentation and spatial mismatch (Wyly 1996). For example, despite their different occupations, black men and women face similar spatial mismatch barriers in urban labor markets (McLafferty and Preston 1996).

Some labor market segmentation results from differences in education and skills that coincide, in some cases, with racial and ethnic divisions. The skills mismatch refers to the gap between the education and skill levels of inner-city residents and the requirements of available jobs. The transformation from a goods-producing to a service-producing economy is causing a loss of low-skill, manufacturing jobs from central cities, while the emerging 'information economy' demands a highly educated workforce. Poverty, poor quality schools, and social dislocations limit educational achievement in the inner city, resulting in the skills mismatch. The spatial and skills mismatches position workers and jobs in intersecting social and geographical spaces to limit employment opportunities for innercity residents.

Transportation is a crucial factor in spatial mismatch, influencing the ease and cost of overcoming distance in traveling to work. American cities are built around the automobile, the most flexible mode of transportation and the one that offers access to the widest geographical range of job opportunities. In contrast, travel by public transit is slow and limited to fixed routes and schedules. For inner-city residents, the hub and spoke design of most mass transit systems is poorly structured for travel to suburban job centers. Because blacks and Latinos have low rates of car ownership and licensing, they are much more dependent on public transit than are their white counterparts. Transit dependence affects all aspects of employment including job search and access to suburban jobs, and it is the single most important determinant of minority workers' long commute times. Some argue that this 'automobile mismatch' is a more serious problem than the spatial mismatch (Taylor and Ong 1995).

The process of spatial job search is also important to the SMH. People often search for jobs near their homes. As distance from the home increases, information about nearby job opportunities decreases and the costs of job search rise. Research indicates that blacks' and Latinos' job search patterns are centered upon their homes, and they tend to search in fewer areas and areas with weaker employment growth than do comparable whites (Stoll and Raphael 2000). Social networks play a key part in spatial job search. Racial and ethnic minorities rely heavily on family and friends for finding employment. For low-income, low-skill workers this may lead to job search in areas of low job growth and poor quality jobs, thus aggravating the spatial mismatch.

3. *Impacts of Spatial Mismatch*

Spatial mismatch affects peoples' lives in diverse ways. The SMH emphasizes the effects on poverty and unemployment and low incomes, but the empirical record is mixed. Perhaps the strongest support for the SMH comes from research on minority youth. For this group, unemployment is strongly correlated with

the average travel time to the types of jobs that youth normally hold and youths' proximity to areas of employment growth (Ihlanfeldt 1992). Minority youth who face long travel times or distances to areas of employment opportunity are much more likely to be unemployed. Looking beyond youth, research among vulnerable populations such as welfare recipients reveals that unemployment is highly sensitive to the concentration of low-wage jobs nearby (Ong and Blumenberg 1998). A 1998 review of the spatial mismatch literature concludes that the vast majority of recent studies support the SMH, and 'the lack of geographical access to employment is an important factor in explaining labor market outcomes' (Ihlanfeldt and Sjoquist 1998, p. 881).

Others follow Ellwood (1986) in arguing that 'race, not space' (p. 181) is the main determinant of poor labor market outcomes. Racial discrimination permeates urban labor markets in the United States, reducing minority employment irrespective of geographical access to jobs. Racial discrimination is often rooted in place, as firms avoid hiring workers from inner-city, minority communities, or use far-flung, racially-biased social networks in attracting employees (Kasinitz and Rosenberg 1996). Critics of the SMH point to examples where high unemployment and dense job opportunities exist side by side. The challenge in examining 'race versus space' is that the two are deeply intertwined: Race *is* space in the USA. Kain (1968) recognized this in the original formulation of the SMH, noting the impacts of racial discrimination by employers and discrimination in housing markets on geographical access to employment. Attempts at disentangling the effects of race and space find that both have large and significant effects on employment outcomes (Stoll 1999).

The implications of spatial mismatch beyond labor markets are discussed in Wilson's path-breaking work on the urban underclass (Wilson 1987, Wilson 1996). Wilson starts with spatial mismatch and economic restructuring and weaves a rich tapestry that includes the effects on male unemployment, marriage, household structure, and the implications for an array of social issues such as crime, migration, local institutions, and educational achievement. As jobs move away and incomes decline, the ties that bind communities, households, and institutions start to disintegrate. Spatial mismatch is not the only cause of these changes, but it is an important piece of the puzzle.

4. *Spatial Mismatch and Urban Policy*

Despite differences of opinion about the SMH, it has played an important role in urban policy formulation for several decades. Policies to address spatial mismatch involve efforts to bring people and jobs closer together or to improve the connections between them. Three general types of policy have been suggested. First, 'economic development' policies aim at creating jobs in inner-city areas. These place-based policies include public works programs and the provision of subsidies, training programs, and infrastructure to attract new industries to inner-city areas. The policies go by various names—the current incarnation is 'enterprise zones'—and they have had limited success. Many have failed to generate sustained, long-term employment because they did not alter the profitability of inner-city locations and benefits leaked out from inner-city neighborhoods to other areas.

Second, 'dispersal' policies aim at facilitating the movement of people to job-rich suburban areas. These strategies focus on the housing market and include the vigorous enforcement of fair housing laws, the development of public housing in suburban areas, and the provision of subsidies to help inner-city residents move to the suburbs. A good example of the latter is the Gautreaux program in Chicago, which offered housing subsidies to residents of public housing who wanted to move to the suburbs. A careful comparison of movers and stayers showed significant gains in employment for those who moved to the suburbs (Popkin et al. 1993).

Third, 'mobility' strategies emphasize improving inner-city residents' access to suburban job opportunities, rather than changing the locations of jobs or residences. Included here are improvements in public transit to facilitate reverse commuting, creation of car pools and van pools, employer-sponsored transportation from inner-city areas to job sites, and efforts to improve car ownership among the poor. Such policies have immediate and obvious benefits; however, they have never been implemented on a large scale. Also, the most effective mix of policies varies among cities depending on the density of urban development and the availability of transportation infrastructure.

After decades of experimentation, most analysts agree on the need for multipronged approaches that are sensitive to the particularities of place and context. Furthermore, there is general agreement that spatial mismatch cannot be isolated from the constellation of forces that affect poverty and unemployment. Breaking down the barriers of racial discrimination in job and housing markets, and improving health, housing, and education in disadvantaged communities could make spatial mismatch irrelevant in the long run.

Looking ahead, the historical specificity of the SMH highlights the fact that spatial mismatch is not a characteristic of particular locations but of particular places where people are disadvantaged based on a combination of race, ethnicity, class, and poor geographical access to work. Changing patterns of urban development, including the gentrification of inner-city neighborhoods, the growth of poverty in the suburbs, and evolving patterns of employment opportunity

mean that spatial mismatch is perpetually recast. Today's suburbs may become tomorrow's 'inner cities' as urbanization redefines spatial mismatch.

See also: Environmental Determinism; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Supply; Local Economic Development; Migration, Economics of; Rural Geography; Social Stratification; Spatial Interaction; Spatial Labor Markets; Transportation Geography; Transportation Planning; Urban Geography; Urban Growth Models; Urban Poverty in Neighborhoods

Bibliography

Cooke T J, Ross S L 1999 Sample selection bias in models of commuting time. *Urban Studies* **36**: 1599–613

Ellwood D 1986 The spatial mismatch hypothesis: Are there teenage jobs missing in the ghetto? In: Freeman R B, Holzer H T (eds.) *The Black Youth Employment Crisis*. University of Chicago Press, Chicago, pp. 147–85

Gilbert M R 1998 "Race", space and power: The survival strategies of working poor women. *Annals of the Association of American Geographers* **88**: 595–621

Hanson S, Pratt G 1995 *Gender, Work, and Space*. Routledge, New York

Holloway S R 1998 Male youth activities and metropolitan context. *Environment and Planning* **30**: 385–99

Holzer H J 1991 The spatial mismatch hypothesis: What has the evidence shown? *Urban Studies* **28**: 105–22

Ihlanfeldt K R 1992 *Job Accessibility and the Employment and School Enrollment of Teenagers*. W. E. Upjohn, Institute for Employment Research, Kalamazoo, MI

Ihlanfeldt K R, Sjoquist D V 1998 The spatial mismatch hypothesis: A review of recent studies and their implications for welfare reform. *Housing Policy Debate* **9**: 849–92

Immergluck D 1998 Job proximity and the urban employment problem: Do suitable nearby jobs improve neighborhood employment rates? *Urban Studies* **35**: 7–23

Johnston-Anumonwo I 1997 Gender, race and constrained work trips in Buffalo, 1990. *The Professional Geographer* **49**: 306–17

Kain J 1968 Housing segregation, negro employment, and metropolitan decentralization. *The Quarterly Journal of Economics* **82**: 175–97

Kasinitz P, Rosenberg J 1996 'Missing the connection': Social isolation and employment on the Brooklyn waterfront. *Social Problems* **43**: 180–96

McLafferty S, Preston V 1996 Spatial mismatch and employment in a decade of restructuring. *Professional Geographer* **48**: 420–31

Ong P, Blumenberg E 1998 Job access, commute and travel burden among welfare recipients. *Urban Studies* **35**: 77–93

Popkin S J, Rosenbaum J E, Meaden P M 1993 Labor market experiences of low-income, black women in middle-class suburbs. Evidence from a survey of Gautreaux program participants. *Journal of Policy Analysis and Management* **12**: 556–73

Stoll M A 1999 *Race, Space and Youth Labor Markets*. Garland Publishing, New York

Stoll M A, Raphael S 2000 Racial differences in spatial job search: Exploring the causes and consequences. *Economic Geography* **76**: 201–23

Taylor B D, Ong P M 1995 Spatial mismatch or automobile mismatch? An examination of race, residence and commuting in US metropolitan areas. *Urban Studies* **32**: 1453–73

Wilson W J 1987 *The Truly Disadvantaged: The Inner City, The Underclass, and Public Policy*. University of Chicago Press, Chicago

Wilson W J 1996 *When Work Disappears: The World of the New Urban Poor*. Knopf, New York

Wyly E K 1996 Race, gender, and spatial segmentation in the Twin Cities. *Professional Geographer* **48**: 431–45

S. McLafferty

Spatial Optimization Models

The concept of spatial optimization is quite simple. This topic involves identifying how land use and activities should be allocated and arranged spatially in order to optimize efficiency or some other measure of goodness. For example, how should fire protection resources be distributed across a city in order to maximize protection? How do we design a system of reserve lands that can ensure the survival of an endangered species? How should an area be divided into school attendance zones? These are but a few examples of spatial design problems. For each type of design problem there can be numerous if not an infinite number of alternative configurations or solutions. Optimization involves the attempt to identify the best of the alternatives. Simply put, spatial optimization is the science of optimal arrangement. The use of spatial optimization is growing due to the existence of better spatial data, the emergence of powerful geographic information systems, decreasing costs of computation, and better computer graphics and solution algorithms.

At first we might think that such a topic is futuristic, cutting edge. Actually, spatial optimization is a geographical problem that is as old as mankind. For example, hunters and gatherers had to decide how to allocate their efforts in collecting foodstuffs, moving encampments, etc. Such societies needed to be efficient in order to survive. Anthropologists and geographers have used optimization to analyze the efficiency of these early societies (see, for example, Van Reidhead (1979), Keene (1979), Church and Bell (1988)). Even the ancient Chinese art of Feng Shui deals with the arrangement of activities in order to be in a harmonious state. Just as we are interested in how spatial arrangements evolved in the past, we are equally interested in how future patterns of activities should be designed. For example, just where should that new store be located? Where should a dealer

provide service facilities for their products? How should a network of warehouses be designed and operated in order to supply a number of retail locations? Such questions are but the tip of the 'iceberg' of possible spatial location/allocation problems.

As with all geographic problems, the temporal domain is important. For example, between 1990 and 1998 Broward County, FL, located 15 new schools and operated up to 2,000 portable classrooms in order to cope with a growing school population. This means that school boundaries have been frequently adjusted and school capacities are often expanded with portable classrooms in order to cope with increasing enrollment. One of the major objectives in planning new schools and creating attendance areas is to plan an expansion that involves the least disruption of students and school reassignments (Lemberg and Church 2000). In a different example, to meet the fluctuations in daily emergency '911' call demand in San Diego County, CA, American Medical Response operates between two and 24 ambulances during each day. Posting ambulances at various locations and dispatching them over time is a complicated logistical/spatial optimization problem. Rest assured that the patterns of deployment vary over the day in order to reduce response times, but just where should such posts be located to get the most service out of a limited number of vehicles? Since it costs over $500,000 per year to keep an ambulance in operation, it is important to find an efficient service plan.

Geographic scale is an equally important issue in spatial optimization. There are typically three basic scales in spatial optimization problems: micro, meso, and macro. The macroscaled problem involves the highest level or area size (i.e., smallest map scale). At such a level, we might ask: 'in which cities should we have factories or warehouses?' We generally use no specific site information in making such a selection. Interactions between customers and facilities are modeled at an aggregated level. The objective is to identify those areas in which it makes the most sense to open a facility or activity. Such a decision might be considered 'strategic.' If we have a strategic plan to locate a regional warehouse in Reno, NV, the 'meso' level question is 'where in Reno should it be.' Such a problem would involve identifying the best available site, considering a wide range of criteria. This is a type of 'tactical' decision making and is often performed to support a strategic decision. Once we have a site or a place the microscaled problem involves carving up that space and allocating it into specific activities. For example, we may allocate store space to specific retail items or a shop floor to machine tools. Even malls are arranged spatially in order to take advantage of anchor stores and walking patterns (Bean et al. 1988). How people use such facilities and interact with their environment is an especially important aspect of this problem (Underhill 1999). In the industrial engineering literature, this type of problem is often referred to as a layout problem. Microscaled problems are akin to an 'operational level' problem.

1. *Historical Roots of Spatial Optimization*

Efficient land use activity was first discussed by Von Thunen (1842) when he described how agricultural activities including wood lots should be optimally arranged around an isolated village. Primary industries such as forestry now use spatial optimization to identify how each stand (tree type and age class) is to be managed over time. These management plans include the building of roads, the provision of environmental standards, and the return on investment (Church et al. 1998). In Chile, forest companies optimize nearly every feature of their industry, including the scheduling and routing of vehicles for log transport (so that logs arrive when just needed at loading facilities such as ports), the placement of harvest towers and cable ways for steep slope harvesting, and the minimization of land devoted to unproductive elements (e.g., roads). The US Forest Service has utilized a spatial optimization model called Forplan (and its successor SPECTRUM) for the last three decades in developing management plans (Kent et al. 1990) for virtually every national forest in the US.

Alfred Weber proposed one of the earliest industrial location models in 1908 (Wesolowsky 1993). As a simple example, he described the location of a manufacturing point between two localized raw materials and a market. He considered that the three points could be plotted on a Cartesian plane as shown in Fig. 1. Transport distances between the points were assumed to be Euclidean. Weber discussed how, under certain assumptions, the ideal location was the one that minimized transport costs of the raw materials to the manufacturing point and the product to market. Weber showed how this model could be solved geometrically for certain problems. The first optimal solution for the more generalized Weber model involving many markets and raw materials was developed in 1937 by Weisfeld and reinvented by Kuhn and Kuenne and Cooper in the 1960s (Wesolowsky 1993). Cooper subsequently proposed an extended model involving the location of multiple facilities. The multifacility Weber problem can now be solved optimally for only those problems involving less than 100 points.

In 1933 Walter Christaller (1966) wrote: 'Just as there are laws which determine the life of the economy, there are special economic-geographic laws determining the arrangement of towns.' Christaller attempted to find the laws that determined the number, size, and distribution of towns in a region of Southern Germany. Since that time geographers, archaeologists, and anthropologists have searched for Christaller's central place patterns in other regions and in other

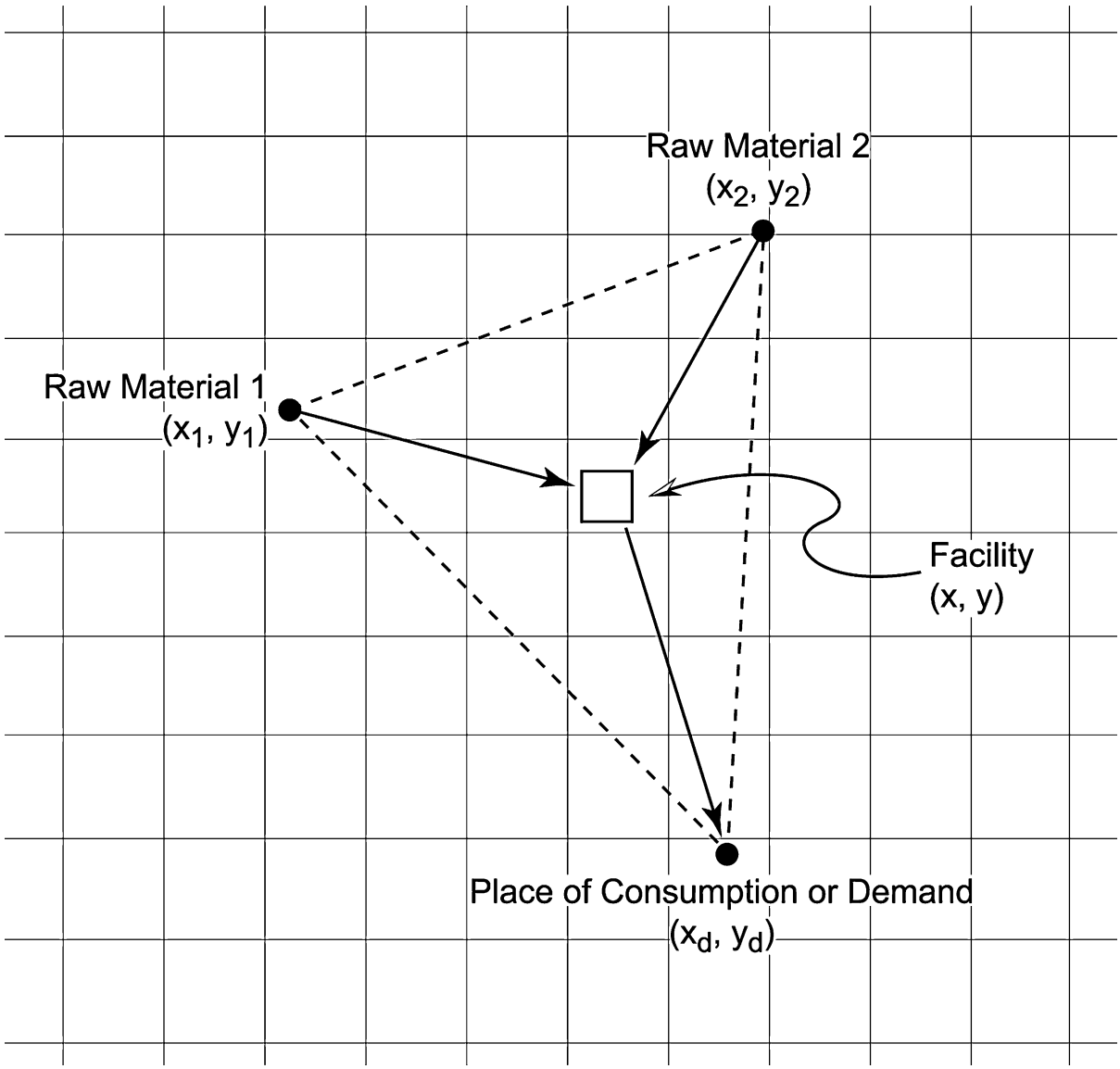

Figure 1
The classic Weber location triangle

applications, and have developed spatial optimization models that solve for the optimal spatial arrangement of towns and markets (Kuby 1989, Storbeck 1990). Virtually all early location theories have some type of spatial optimization counterpart today.

2. *Selected Examples of Spatial Optimization*

A large number of spatial design/arrangement/allocation problems have been approached using some form of optimization modeling. Techniques used to solve optimization models are typically found in the operations research (OR) literature. The field of OR traces its roots to World War Two and the need to solve large scale military logistics problems. One of the first OR models that was developed and can be solved optimally was the classical transportation problem. This problem involves a distributed set of supply points (called sources) and a distributed set of demand locations (called sinks). Each source has a finite capacity to supply a good and each destination has a specified amount of demand for that good that needs to be met. Between each unique pair of one source and one sink we can identify the most efficient route and

Table 1
Examples of spatial optimization problems

Problem name	Problem definition	General reference
Corridor location problem	Across a landscape, identify the optimal route for a right-of-way for a facility such as a pipeline, highway, or power transmission line that connects an origin and destination	Smith et al. 1989 Goodchild 1977 Lombard and Church 1993
Vehicle routing problem	Optimally route a vehicle from a starting point to visit a number of pre-specified locations and then return to the starting point.	Bodin 1990 Weigel and Cao 1999
Classical transportation problem	Allocate limited supplies at different locations to a number of spatially distributed demand points in such a manner as to minimize transportation costs of serving all demand	Yeates 1963 Camm et al. 1997
Simple plant location problem	Locate a set of facilities and allocate their supplies to demand points in such a manner as to minimize facility and transportation costs	Love et al. 1988
p-Median location problem	Locate p-facilities so that the total weighted distance of serving all demand is minimized	ReVelle and Swain 1970
Maximal covering location problem	Locate p-facilities so that the largest proportion of demand is covered or served within some maximal distance or time	Church and ReVelle 1974 Schilling et al. 1993
p-Hub location problem	Locate p-transportation hubs and allocate demand to specific hubs in such a manner as to minimize total transportation costs	O'Kelly and Miller 1994
Site configuration/design problem	Identify the best set of parcels or land units that optimizes costs and meets shape, contiguity, size, and suitability constraints	Wright et al. 1983 Cova and Church 2000
Districting problem	Partition spatial units into a set of districts that optimize one or more criteria; examples, school districts, political districts, etc.	Woodall et al. 1980 Lemberg and Church 2000
Spatial allocation problem	Allocate specific activities to defined spatial units	Hof and Bevers 1998

the cost to ship one unit of that good between the source-sink pair. The objective of the classical transportation problem is to allocate the supplies to the demands so that the total transport cost is minimized and the demand at each sink is met. Garrison (1959) was the first geographer to discuss the utility of the transportation problem as a model to analyze spatial structure. Yeates (1963) and others have since used this spatial optimization model to analyze efficiency of an existing arrangement or to develop an optimal transportation plan.

Table 1 gives a representative sample of spatial optimization problems that can be found in the literature, including the classical transportation problem. The list includes vehicle routing, districting, transportation, facility location, site search, and space allocation. For each of the 10 problems listed, the name, definition, and a few references are given. For each of these generic problems, a number of models have been developed. The references cite the original work, a review, or a recent work that contains an up-to-date list of references. All of the problems listed in

Table 1, with the exception of the corridor location problem and the classical transportation problem, have a problem complexity that is called NP (Non-deterministic Polynomial). The class NP refers to a set of problems for which there exists virtually no chance that procedures will be found which can solve all problems optimally within a reasonable computational time on a computer. One of the reasons for this is that these problems require the use of integer programming, an area of optimization that can be both computational intensive and limited in terms of the size of an individual application that can be solved optimally.

Because of this fact, it is common to see that different techniques are used for solving specific models and that heuristics are needed for most moderate to large sized problems. Heuristics are specialized solution procedures that are designed to approach problem solving using either a set of rules or some search strategy. An example of a heuristic is given in the next section.

3. *A Detailed View of a Spatial Optimization Problem*

Hakimi (1965) proposed a network location model called the p-median problem. Consider the network depicted in Fig. 2. Each node is connected to other nodes by arcs of given distances. Without any loss of generality, consider that each node represents a potential facility site as well as a point of demand. The measure of demand is represented as a demand weight (in units of people, tons, trips, or some other measure of need to be supplied by a facility). Assume that each demand will be supplied or served by their closest facility. The transportation cost is assumed to be directly proportional to the weight of a demand multiplied by the distance to the closest facility (note: this is the same assumption used in the Weber model). The p-median problem involves selecting the locations of p-facilities so that the total weighted-distance for all demand is minimized. Hakimi proved that at least one optimal solution to the p-median problem consists entirely of nodes of the network. Thus, the search for an optimal configuration can be limited to just the nodes of the network.

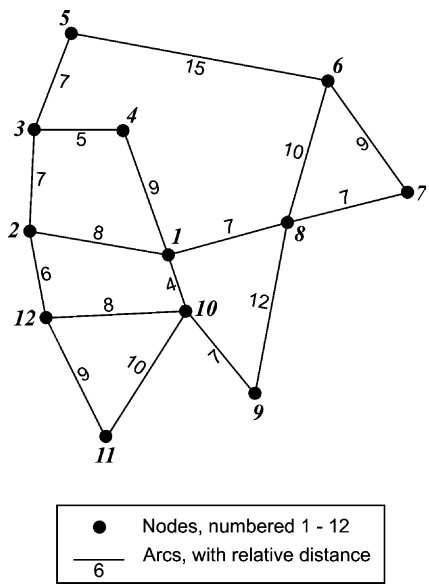

Figure 2
An example road network

Searching for the best nodal configuration, and hence an optimal configuration, seems simple. Why not enumerate all of the possible facility patterns, and pick the configuration with the lowest weighted-distance? Unfortunately, it isn't that simple. Consider, for example, that we wish to locate 10 facilities on a network of 100 nodes. Such an enumeration would involve generating and evaluating the following number of combinations

$$\left\{ \begin{matrix} 100 \\ 10 \end{matrix} \right\} = \frac{100!}{5!(100-5)!}$$

$$= \frac{100*99*98*97*96*95*94*\ldots..*1}{10*9*8*\ldots..*1*90*89*88*\ldots..*1}$$

or

$$\frac{100*99*98*97*96\ldots*91}{10*9*8*7*\ldots.*1} \cong 17{,}300{,}000{,}000{,}000$$

configurations
or alternatives

If a very fast computer could analyze 1,000 solutions each second, it would take 548 such computers working simultaneously to finish the task in a year. Thus, for many problems enumeration is impossible. In such cases, it is reasonable to consider optimization. ReVelle and Swain (1970) proposed the first optimal approach to solving the p-median problem. The basis of their approach is to formulate the p-median model as an integer-programming problem.

Consider a connected transportation network comprised of n nodes and at least $n-1$ arcs (note: it requires at least $n-1$ arcs to be connected). An example of a 12-node network is given in Fig. 2. In this review we will consider the most general form of the p-median problem where each node represents both a place of demand and a potential facility site. Given any pair of nodes on the network, a shortest path exists between the pair of nodes, and the distance of that path is easily calculated by a variety of efficient techniques. Consider the following notation

n = the number of nodes in the network

Table 2
Optimization model

Objective function:	$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i d_{ij} x_{ij}$	
Subject to:		
(1) Each demand must be served by either placing a facility at its node or by assigning to a facility elsewhere	$\sum_{j=1}^{n} x_{ij} = 1$	or each node i
(2) Assignment is possible to only open facilities	$x_{ij} \leqslant x_{jj}$	for each node i and j where $i \neq j$
(3) Exactly p-Facilities are located	$\sum_{j=1}^{n} x_{jj} = p$	
(4) Integer restrictions on variables	$x_{ij} = 0, 1$	for each node i and j

Table 3
Optimization model: modified second constraint

(2) Assignments are possible to a facility, only if it has been selected	$\sum_{i=1}^{n} x_{ij} \leqslant n x_{jj}$	For each node j

i, j = indices used to reference nodes,
where the nodes are numbered from 1 to n

a_i = the amount of demand at node i

d_{ij} = the distance from node i to node j

$$x_{ij} = \begin{cases} 1, \text{ if demand at node } i \text{ assigns to a facility at node } j \\ 0, \text{ otherwise} \end{cases}$$

$$x_{jj} = \begin{cases} 1, \text{ if node } j \text{ self assigns indicating that a facility is located at node } j \\ 0, \text{ otherwise} \end{cases}$$

Using the above notation, we can define the optimization model in Table 2 (ReVelle and Swain 1970).

The model in Table 2 can be solved by the use of integer programming techniques, most notably, linear programming with branch and bound (LP/BB). In a relatively high percentage of problems, the optimal linear programming solution is integer optimal as well and the branch and bound routine is not used. This property is called 'integer friendly.' Since branch and bound algorithms may take considerable computational time to resolve noninteger solutions, finding an optimal integer optima without BB usually means relatively fast computational times for a given problem size. Unfortunately, the above model is large in terms of the number or variables and constraints (n^2 and n^2+1 respectively). Consequently there is a limit as to how large a problem can be solved by this specific model formulation and LP/BB.

It is possible to formulate the p-median problem with a slightly different constraint set. Consider the following form of the second constraint (Table 3): the constraints that restrict demand assignments to only those nodes selected for a facility have been aggregated into one constraint for each facility. Consequently, this second model is more compact in that it contains only $2n+1$ constraints rather than n^2+1 constraints. As models are typically easier to solve when they have fewer constraints, this would at first seem to be an improvement. Unfortunately, this second model form is not 'integer friendly' and consequently the use of LP/BB typically relies on the BB algorithm and resulting solution times can be very large. The lesson here is that there are always alternate ways in which to formulate any given spatial optimization model, and the difficulty in solving a given problem can be in part due to the mathematical form of the problem.

There are practical limits, however, as to the size of a problem that can be solved optimally. For large problems it is simply not possible to find an optimal solution with certainty. Because of this there has been a great reliance on heuristic solution procedures. Heuristics are often designed to exploit some type of search strategy. For example one of the best known methods for solving the p-median problem is the 'swap' heuristic of Teitz and Bart (Church and Sorensen 1996). The 'swap' heuristic starts with a randomly generated solution. Given this starting solution, consider a candidate node as a possible replacement site for each of the currently selected nodes. If any swap involving this candidate node yields an improvement, the best swap with that candidate is made, creating a new and improved solution. Swaps are tested until no improvements can be made involving any node of the network. Repeating this procedure a few times has a good chance of finding a good, if not optimal solution. Other strategies are based upon 'survival-of-the-fittest' genetic algorithms, simulated annealing and statistical mechanics, Monte Carlo sampling, etc. All heuristics are rated for their capabilities in finding good solutions.

4. *Research Needs in Spatial Optimization*

Research in spatial optimization has expanded over the years. This expansion is due in part to increasing computational power and availability of computers,

better general purpose commercial products like GIS and optimization software, fourth generation computer language tools like visual studio and C++, and high resolution color monitors. Some spatial optimization models are now a part of general purpose GIS software. There are, however, a number of research directions that must be taken in order for such modeling tools to be commonly available. To make this a reality, research is needed in the following areas:

(a) develop better and more capable solution algorithms and heuristics. This is especially important in improving the capability to solve large problems of the simple plant location problem where the facilities have finite capacities;

(b) expand existing models and develop more comprehensive spatial optimization models, especially in ecological reserve design and land use planning;

(c) develop methods to better estimate the impact of data error and its propagation on model results. This is especially important since all geographic data sets of realistic size contain error;

(d) expand data model definitions to better link spatial optimization models with GIS. One application of considerable importance is the optimal location of cellular phone towers;

(e) develop better visualization methods along with graphical interface approaches that can support collaborative decision making processes;

(f) explore methods of data aggregation so that it may be possible to construct small problem representations that are easier to solve, but still retain all of the important problem characteristics.

Many industries and government agencies have been pushed to become more efficient. Increasing productivity has become a mantra of stockholder and citizen alike. In many settings new spatial optimization models are being proposed to help increase efficiency, promote equity, reduce environmental damages, and reduce energy needs. For example, new spatial/temporal optimization models that are being used for emergency ambulance operations can create plans that improve service response without additional investment or keep service levels constant and cut costs. The economic incentive to improve services and reduce costs will expand the application base that is already very large.

See also: Environmental and Resource Management; Industrial Ecology; Industrial Geography; Local Economic Development; Space and Social Theory in Geography; Spatial Equity; Spatial Pattern, Analysis of; Transportation Geography; Urban Places, Planning for the Use of: Design Guide

Bibliography

Bean J C, Noon C E, Ryan S M, Salton G J 1988 Selecting tenants in a shopping mall. *Interfaces* **18**: 1–9

Bodin L D 1990 Twenty years of routing and scheduling. *Operations Research* **38**: 571–9

Camm J D, Chorman T E, Dill F A, Evans J R 1997 Blending OR/MS, judgment, and GIS: Restructuring P&G's supply chain. *Interfaces* **27**: 128–42

Christaller W 1966 *Central Places in Southern Germany*. Prentice-Hall, Englewood Cliffs, NJ

Church R L, Bell T L 1988 An analysis of ancient Egyptian settlement patterns using location-allocation covering models. *Annals of the Association of American Geographers* **78**: 701–14

Church R L, Murray A T, Weintraub A 1998 Locational issues in forest management. *Location Science* **6**: 137–53

Church R L, ReVelle C S 1974 The maximal covering location problem. *Papers of the Regional Science Association* **32**: 101–18

Church R L, Sorensen P 1996 Integrating normative location models into GIS: Problems and prospects with the p-median model. In: Longely P, Batty L M (eds.) *Spatial Analysis–Modelling in a GIS Environment*

Cova T J, Church R L 2000 Contiguity constraints for single region site search problems. *Geographical Analysis*, in press

Garrison W L 1959 Spatial structure of the economy II. *Annals of the Association of American Geographers* **49**: 471–82

Goodchild M 1977 An evaluation of lattice solutions to the problem of corridor location. *Environment and Planning A* **9**: 727–38

Hakimi S L 1965 Optimum distribution of switching centers in a communication network and some related group theoretic problems. *Operations Research* **13**: 562–74

Hof J, Bevers M 1998 *Spatial Optimization for Managed Ecosystems*. Columbia University Press, New York

Keene A S 1979 Economic optimization models and the study of hunter-gatherer subsistence settlement systems. In: Renfrew C, Cooke K (eds.) *Transformations: Mathematical Approaches to Culture Change*. Academic Press, London

Kent B, Bare B B, Field R, Bradley G 1991 Natural resource land management planning using large-scale linear programming: The USDA Forest Service Experience with FORPLAN. *Operations Research* **39**: 13–27

Kuby M J 1989 A location-allocation model of Losch Central Place Theory-Testing on a uniform lattice network. *Geographical Analysis* **21**: 316–37

Lemberg D L, Church R L 2000 The school boundary stability problem over time. *Socio-Economic Planning Sciences* **34**: 159–76

Lombard K, Church R L 1993 The gateway shortest path problem: Generating alternative routes for the corridor location problem. *Geographical Systems* **1**: 22–45

Love R F, Morris J G, Wesolowsky G O 1988 *Facilities Location: Models and Methods*. North-Holland, New York

O'Kelly M E, Miller H J 1994 The hub network design problem: A review and synthesis. *Journal of Transport Geography* **2**: 31–40

ReVelle C S, Swain R W 1970 Central facilities location. *Geographical Analysis* **2**: 30–42

Schilling D A, Jayaraman V, Barkhi R 1993 A review of covering problems in facility location. *Location Science* **1**: 25–55

Smith T R, Peng G, Gahinet P 1989 Asynchronous, iterative, and parallel procedures for solving the weighted-region least cost path problem. *Geographical Analysis* **21**: 147–66

Storbeck J E 1990 Classical central places as protected thresholds. *Geographical Analysis* **22**: 4–21

Underhill P 1999 *Why We Buy: The Science of Shopping*. Simon and Schuster, New York

Van Reidhead V A 1979 Linear programming models in archaeology. *Annual Review of Anthropology* **8**: 543–78
Von Thunen J H 1842 *Der Isolierte Staat*. In: Hall P (ed.) *Von Thünens Isolated State*. Pergamon, London
Weigel D, Cao B 1999 Applying GIS and OR techniques to solve Sears technician-dispatching and home-delivery problems. *Interfaces* **29**: 112–30
Wesolowsky G O 1993 The Weber problem: History and procedures. *Location Science* **1**: 5–24
Woodall M, Cromley R G, Semple R K, Green M B 1980 The elimination of racially identifiable schools. *Professional Geographer* **32**: 412–20
Wright J, ReVelle C, Cohon J 1983 A multi-objective integer programming model for the land acquisition problem. *Regional Science and Urban Economics* **13**: 31–53
Yeates M 1963 Hinterland delimitation: A distance minimizing approach. *Professional Geographer* **15**: 7–10

R. L. Church

Spatial Pattern, Analysis of

Spatial data consist of measurements or observations (data values) taken at specific locations (data sites) in a geographic space. Typically, such data are presented visually in the form of a map, although they may also be represented in other analog forms such as aerial photographs and remotely sensed images. Since most processes operate over both space and time, such representations may be considered as cross-sectional snapshots of reality. The relative locations of the data values in the geographic space define a spatial pattern. Spatial pattern analysis involves describing and explaining such patterns. In general, this is done by comparing characteristics of an empirical pattern with those of null models representative of spatial randomness (i.e., no discernible spatial order). Once identified, the distinctive characteristics of the empirical pattern can be used as evidence to gain a better understanding of the processes which generated the pattern. A classic example of this is provided by the nineteenth-century physician John Snow who identified the cause of a cholera epidemic in London by observing that deaths were concentrated around a particular public water pump.

Methods of spatial pattern analysis can be classified in terms of the kinds of spatial data analyzed. Three types of spatial data are usually recognized: point data; geostatistical data; and lattice data. For each type, further classification can be achieved by considering if the data values are qualitative (categorical) or quantitative, and if they are univariate or multivariate. In addition, spatial pattern analysis originated in distinct academic communities. Point pattern techniques were developed in plant ecology, economic geography, and epidemiology; geostatistics in mining engineering and geology; and lattice data methods in econometrics. As a result, there are literally hundreds of different spatial pattern analysis techniques and almost as many texts describing them. Consequently, this essay will emphasize conceptual issues common to all techniques and illustrate them by outlining a widely used procedure for analyzing univariate, quantitative spatial data of each type. To simplify the presentation, it is assumed that the behavior of the process responsible for the pattern being analyzed is the same throughout the area under study. This condition is known as 'stationarity.' For more comprehensive treatments of spatial pattern analysis see Upton and Fingleton (1985), Cressie (1993), and Bailey and Gatrell (1995).

1. Point Data

Here the data sites consist of a finite number of point locations which themselves are the object of interest (e.g., the towns shown in Fig. 1). More formally, let $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$, where $\mathbf{s}_i = (s_{i1}, s_{i2})$, be the locations of n events in a bounded study region A of size $|A|$. Such patterns are typically analyzed by examining characteristics of either the number of events in subareas of A or interevent distances $\|\mathbf{s}_i - \mathbf{s}_j\|$. The most popular method of analysis, K function analysis, developed by the statistician Brian Ripley (Ripley 1981), combines a consideration of both characteristics. The procedure analyses the average number of events that lie within a distance h of a randomly chosen event. For a given empirical pattern, the K function, $K(h)$, may be estimated using

$$\hat{K}(h) = \hat{\lambda}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w(\mathbf{s}_i, \mathbf{s}_j)^{-1} I(\|\mathbf{s}_i - \mathbf{s}_j\| \leqslant h)/n$$
$$h > 0,\ i \neq j$$

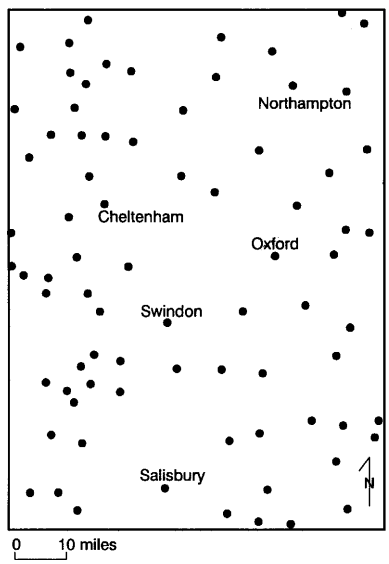

Figure 1
Bus service centers in south-central England, 1955

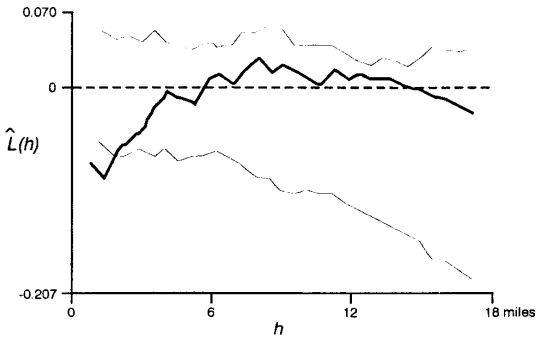

Figure 2
Observed and expected $\hat{L}(h)$ for bus service centers shown in Fig. 1.

where $\hat{\lambda} = n/|A|$ is an estimate of λ, the intensity (mean number of points per unit area) of the point process, $I(\cdot)$ is the number of events $\mathbf{s}_j$ within distance h of all events $\mathbf{s}_i$, and $w(\mathbf{s}_i, \mathbf{s}_j)$ is a correction for edge effects. The last is required since a disk, radius h, may extend outside the boundary of A. Usually, $w(\mathbf{s}_i, \mathbf{s}_j)$ is set equal to the proportion of the circumference of a circle centered at $\mathbf{s}_i$, passing through $\mathbf{s}_j$, that is inside A.

In order to evaluate $\hat{K}(h)$, it is compared to $K(h)$ which would result if the events were located independently in a completely homogeneous environment. Such a pattern is known as complete spatial randomness (CSR). More formally, CSR is the realization of a homogeneous planar Poisson process. This process has two conditions: every location in the study region has an equal chance of receiving an event (uniformity); and the selection of a location for an event in no way influences the selection of locations for any other events (independence). Under CSR,

$$K(h) = \pi h^2$$

To facilitate the comparison with CSR, the statistic $L(h)$ is often used in place of $K(h)$, where

$$L(h) = \sqrt{\frac{K(h)}{n}} - h$$

The expected value of $L(h)$ under CSR is 0. When the events in an empirical pattern are more dispersed over A than they would be in CSR (spatial segregation), the estimated value of $L(h)$, $\hat{L}(h)$, tends to be negative, whereas if the events are more clustered than expected under CSR (spatial clustering), $\hat{L}(h)$ tends to be positive. A test of the statistical significance of the difference between an empirical pattern and CSR can be made using a Monte Carlo simulation procedure. This involves simulating a set of x CSR patterns each with the same number of events as the empirical pattern in a study area identical to that of the empirical pattern. A confidence envelope is then defined by the extreme maximum and minimum values of $\hat{L}(h)$ for the x simulations. If $\hat{L}(h)$ for the empirical pattern extends outside of the confidence envelope, the null hypothesis of CSR is rejected (at a $1/(x+1)$ significance level) in favor of one of segregation or clustering, depending on the direction of the departure.

To illustrate the use of this technique, consider Fig. 1. This shows the locations of 74 central places in south-central England which provided bus services to surrounding rural areas in the 1950s. Figure 2 shows $\hat{L}(h)$ for this pattern together with a confidence envelope for CSR (based on 99 simulations). Note that for small values of h (approximately two miles or less), $\hat{L}(h)$ is negative and extends outside the CSR envelope. This indicates that the bus service centers have an inhibitive effect on each other since they are significantly more dispersed (at the 99 percent confidence level) than would be expected if they were independent.

2. *Geostatistical Data*

Geostatistical data consist of a set of n data sites, $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$, in a bounded study region A, at which the values, $(y_1, y_2, \ldots, y_n)$, of some spatially continuous variable Y have been measured, i.e., the data are assumed to represent sample observations from a surface defined by $(\mathbf{s}, y)$ (e.g., the snowfall data in Fig. 3). Geostatistical data can be analyzed by examining characteristics of the association between pairs of data values as some function of their relative location (e.g., the distance separating them). The most commonly used procedure, the semi-variogram, $\gamma(h)$, focuses on the variance of the data values. For empirical data, the semi-variogram is estimated for equidistant classes h using the sample semi-variogram

$$\hat{\gamma}(h) = \frac{1}{2n_h} \sum_{\|\mathbf{s}_i - \mathbf{s}_j\| = h} (y_i - y_{i+h})^2$$

where the summation is over all pairs of data sites separated by distance h, and n_h is the number of such pairs. In practice, if the data sites are irregularly spaced in A (as in Fig. 3), there may only be one pair of data sites for which $\|\mathbf{s}_i - \mathbf{s}_j\| = h$. In this case, a distance interval $h \pm \Delta$ can be used in place of h. The maximum value of $\hat{\gamma}(h)$ is often referred to as the 'sill' and the distance at which this occurs is known as the 'range.' The range defines the distance at which the data values cease to be related. In theory, $\gamma(h) = 0$ when $h = 0$, since the underlying surface is assumed to be continuous in A. However, in empirical data it is sometimes seen that as $h \to 0$, $\hat{\gamma}(h) \to c_0 > 0$. This is known as the 'nugget effect' and is considered to represent the inherent error induced by measurement errors and by micro-scale environmental variability. If there is no spatial dependence in the data, the sample variogram will consist of just a nugget effect, i.e., it is

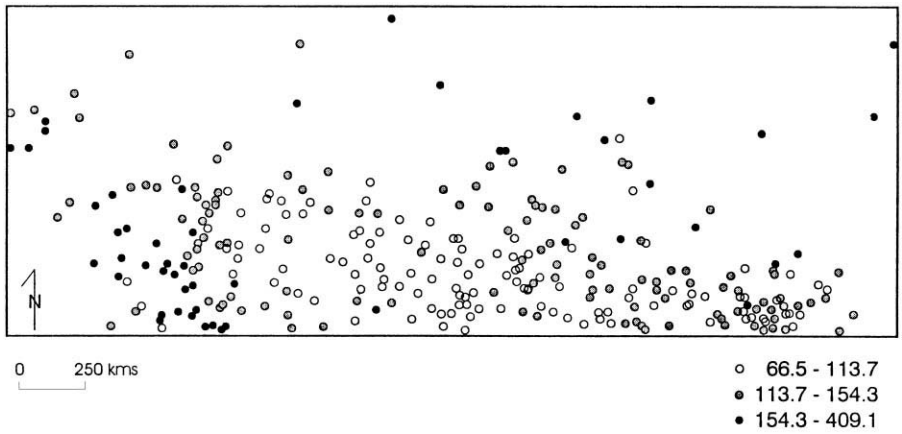

Figure 3
Thirty year averages (1961–90) of annual snowfall in centimeters at meteorological stations in western Canada, latitude 48.75–60 degrees north, longitude 94–119.5 degrees west. To simplify visualization of the data, the individual values are grouped into three classes.

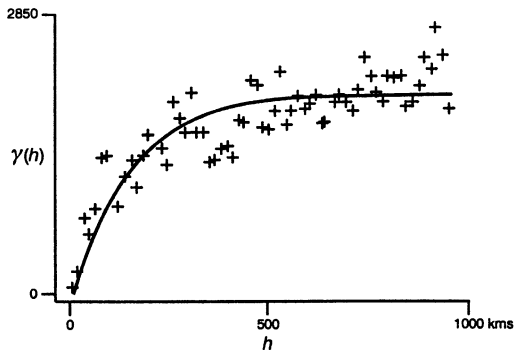

Figure 4
Empirical semi-variogram (crosses) and model semi-variogram (line) for snowfall data in Fig. 3.

horizontal. One of several parametric models can be fitted to the sample semi-variogram to provide a smooth, continuous description of the variance structure in the data. Such models can be used to predict the values of Y for non-sampled locations using a procedure called 'kriging' (Cressie 1993). Figure 4 shows the sample semi-variogram for the data mapped in Fig. 3. For this data the value of the sill is 2,100, reached at a range of 480 kilometers, and there is no nugget effect. An exponential model of the form

$$\gamma(h) = \sigma^2(1 - \exp^{-3h/r})$$

where σ^2 is the sill and r is the range has been fitted to the empirical data. The analysis indicates that while there is a clear spatial dependence in snowfall values at distances of up to about 500 kilometers, this is very marked for distances of up to 200 kilometers.

3. *Lattice Data*

In lattice data, the values, $(y_1, y_2, \ldots, y_n)$ of some variable Y are recorded for each member of a set of n areal units $(a_1, a_2, \ldots, a_n)$ which, taken collectively, cover the study region A, i.e., $a_1 \cup a_2 \cup \ldots \cup a_n = A$. The set of data values is considered to represent one possible realization of a spatial process operating over A. Lattice data are analyzed by examining characteristics of the association between pairs of data values as some function of their spatial association. This is referred to as 'spatial autocorrelation.' Since the data sites are areas, there are many different ways in which the spatial association w_{ij} between any two data sites a_i, a_j may be modeled. By far the most frequently used method is to set $w_{ij} = 1$ if a_i, a_j share a common boundary, and $w_{ij} = 0$, if they do not.

Moran's I, the most widely used measure of spatial autocorrelation, is given by

$$I = \left(\frac{n}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \right)_{i \neq j} \times \left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \right)$$

where: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The first term on the right-hand side of this equation is a normalizing factor, so that the range of I is approximately $-1 \leqslant I \leqslant 1$. The expected value of Moran's I under the assumption of inde-

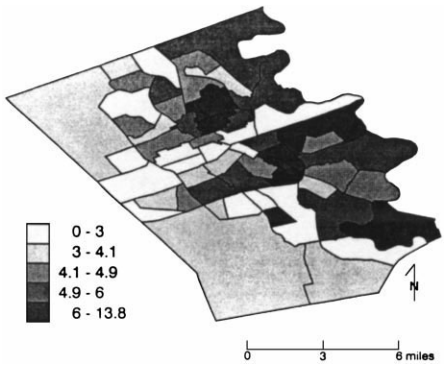

Figure 5
Percentage of census tract population with German mother tongue (Kitchener-Waterloo 1996)

pendence in the spatial pattern of values of Y (no spatial autocorrelation) is

$$\mathrm{E}(I) = -\frac{1}{n-1}$$

and the variance is

$$Var(I) = \left(\frac{n^2 S_1 - nS_2 + 3W^2}{W^2(n^2-1)} \right) - \frac{1}{(n-1)^2}$$

where

$$W = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}; \quad S_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij}+w_{ji})^2}{2};$$

$$S_2 = \sum_{i=1}^{n}(w_{i.}+w_{.j})^2; \quad w_{i.} = \sum_{j=1}^{n} w_{ij}; \quad w_{.j} = \sum_{i=1}^{n} w_{ij}.$$

An observed value of I which is larger than its expectation under the assumption of spatial independence indicates positive spatial autocorrelation (i.e., similar values of y, are found in spatial juxtaposition). In contrast, negative spatial autocorrelation occurs when neighboring values of y_i are mutually dissimilar, indicated by an observed value of Moran's I which is smaller than its expectation. For data sets of $n \geqslant 50$, I is approximately normally distributed so that the standard score

$$z = \frac{I - E(I)}{\sqrt{Var(I)}}$$

may be used to determine if the empirical pattern displays significant spatial autocorrelation.

As an example of this approach, consider Fig. 5 which shows the percentage of the 1996 population whose mother tongue was German for each census tract in the twin cities of Kitchener-Waterloo in Ontario, Canada. For this data set, $n = 56$, $\mathrm{E}(I) = -0.018$ and $\mathrm{Var}(I) = 0.0057$. While the observed value of $I = 0.115$ suggests that there is a tendency for census tracts with similar proportions of German speakers to cluster in space, the standard score of $z = 1.77$ indicates that the pattern of spatial association is not significantly different (at the 95 percent confidence level) from that which could arise by chance (i.e., no spatial autocorrelation).

The measurement of spatial autocorrelation can be extended to higher-order spatial neighbors where the spatially associated areas are $(k-1)$ intervening areas apart. A plot of the values of I for different spatial lags k, the spatial correlogram, is useful for detecting scale variations in the spatial pattern (Upton and Fingleton 1985).

4. *Changing Emphases and Recent Developments in Spatial Pattern Analysis*

Most recent developments in spatial pattern analysis have come in response to changes in two areas. The first involves characteristics of the typical spatial data set to be analyzed. Proliferation of automated monitoring devices, such as orbiting earth satellites, has meant that large (i.e., hundreds of thousands or more data sites) and often frequently updated data sets have become increasingly available. The second change is in computing; in particular, increases in computing power and memory capacity, coupled with cheaper hardware costs. One result of this is the appearance of more sophisticated geographical information systems capable of visualizing spatial data in a variety of ways.

The need to analyze large data sets has resulted in an increasing emphasis on spatial data exploration which, in turn, has led to a shift away from global methods of spatial pattern analysis to local ones (Getis and Ord 1996). While global methods attempt to summarize the entire spatial data set, local methods focus on searching for localized patterns within the data set by such means as identifying atypical locations (spatial outliers), clusters of abnormally high (hot spots) or low (cold spots) data values, and the presence of different patterns in sub-areas (spatial regimes). Since it is possible to generate a local statistic for each data site, the results of such analyses are best appreciated using some form of visualization (Hearnshaw and Unwin 1994).

For local statistics, the null models of spatial randomness, such as that of CSR described in Sect. 1, used to evaluate global statistics may not be appropriate. Instead, null hypotheses which take into account some form of pre-specified spatial structure

may be more appropriate. These may include, for example, allowing for some degree of significant spatial autocorrelation in the overall pattern or permitting the characteristics of the causal process to vary over subareas within the study area. Alternatively, a null hypothesis might be specified to evaluate if events cluster around one or more focal locations (e.g., disease incidences around toxic waste sites). While the characteristics of the simpler null models can be derived analytically, this is often not possible for the newer models, resulting in their properties being estimated from repeated simulations of the model processes.

Frequent updating of data sets has spurred an interest in real time surveillance of spatial patterns. This involves analyzing new data, as it becomes available, to detect any significant changes that have occurred in the characteristics of an existing spatial pattern (Rogerson 1997).

Since all of these innovations in spatial pattern analysis impose heavy computational demands, they have only become practical with the developments in computing previously noted. Indeed, in some situations it is possible that the amount of collected data may exceed human analytical ability. This has stimulated interest in the creation of automated procedures for spatial pattern analysis (Openshaw 1998). In particular, a new research field called spatial data mining, which integrates aspects of machine learning, database systems, data visualization, information theory and spatial statistics, has evolved with the aim of developing efficient methods for identifying the characteristics of very large spatial data sets.

See also: Spatial data; Spatial autocorrelation; Spatial association, measures of; GIS, critical approaches; Spatial analysis, in geography; Spatial sampling; Spatial statistical methods

Bibliography

Bailey T C, Gatrell A C 1995 *Interactive Spatial Data Analysis*. Longman Scientific and Technical, Harlow, UK

Cressie N 1993 *Statistics for Spatial Data*, rev. edn. John Wiley and Sons, Inc., New York

Getis A, Ord J K 1996 Local spatial statistics: an overview. In: Longley P, Batty M (eds.) *Spatial Analysis: Modelling in a GIS Environment*. GeoInformation International, Cambridge, UK

Hearnshaw H M, Unwin D J (eds.) 1994 *Visualization in Geographical Information Systems*. John Wiley and Sons, Inc., Chichester, UK

Openshaw S 1998 Building automated geographical analysis and explanation machines. In: Longley P A, Brooks S M, McDonnell R, Macmillan B (eds.) *Geocomputation: a Primer*. John Wiley and Sons, Inc., Chichester, UK

Ripley B D 1981 *Spatial Statistics*. John Wiley and Sons, Inc., New York

Rogerson P 1997 Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine* **16**: 2081–93

Upton G J G, Fingleton B 1985 *Spatial Data Analysis by Example*, Vol. 1: *Point Pattern and Quantitative Data*. John Wiley & Sons, Inc., Chichester, UK

B. Boots

Spatial Sampling

Sampling is undertaken in order to estimate an attribute of a variable of interest in a population. Sampling is used instead of undertaking a complete census for many reasons. For example undertaking a complete census may be impossible because the population is very, perhaps infinitely, large or it may be impractical because of expense. Some census data, for example household employment data in the UK, is only made available as a 10 percent sample from the 100 percent enumeration partly to help preserve confidentiality but also because the data requires manual coding which is expensive. In many situations a census is unnecessary because a value for the attribute is only required, or can only be measured, up to some predetermined level of precision and a sample size can be specified that will achieve that level of precision. Sampling is also undertaken to improve the accuracy of census data because some methods of recording are known to produce undercounts (e.g., in the case of crimes, not all events are reported to the police) and some methods of enumeration miss certain groups in the population (e.g., the homeless in the case of national censuses).

The term spatial sampling is used to describe the activity of sampling when the population is spatially or geographically distributed. All the individuals in the population have a georeference or geocode that identifies their location with respect to some local or global frame of reference. The population might be a surface of land-use values or ground-level air pollution values. In this case the population is the set of all points in the region and is infinitely large. The population might be the set of all households in a city in which case the population, whilst not infinite, is nonetheless very large. The population might refer to spatial aggregates. For example the population of households might be bundled into census units so that the population is the set of such units for which aggregate quantities are of interest.

Spatial sampling is usually undertaken by recording the data value at a point location such as a position on a line or a surface or recording some aggregate quantity for a polygon such as a pixel, a quadrat, or a census tract. In some spatial sampling a household is treated as a point (is the head of household unemployed?) in others as an area-based quantity (number of persons per room).

Designing a spatial sample to estimate an attribute involves selection of the required sample size (n) and

specification of the sample locations $(s_1,\ldots, s_n)$. The latter is called the sampling plan. The sampled data $\mathbf{z} = \{z(s_1),\ldots, z(s_n)\}$ is used to estimate the attribute which might be a 'nonspatial' attribute such as the mean value of the variable **Z,** or the proportion of the population with a given set of values on Z or the maximum value of Z. The attribute might be a 'spatial' attribute such as the autocorrelation function, semivariogram, or a map of the variable constructed using optimal interpolation. It is necessary to choose an appropriate estimator of the quantity and specify a measure of closeness (with respect to the attribute) that the sampling design seeks to minimize. A measure of closeness might for example be the sampling variance.

There may be other issues that affect spatial sampling designs. There may, for reasons associated with cost constraints, be a trade off that needs to be recognized between on the one hand the greater levels of precision obtainable by increasing the sample size versus the costs associated with collecting larger volumes of data. Samples designed to deliver acceptable levels of sampling error for a national survey won't achieve equivalent levels at smaller state or county-level scales without a larger sample and without stratification. Hence there needs to be some assessment of the benefits of having information at the smaller scale set against the costs of a larger sample. One solution to this problem, that can be used where some estimate is available at the local level, is to adapt methods based on 'borrowing strength.' National surveys yield precise estimates of national rates, but these are biased estimators of local area rates because local areas have different rates. Estimators that 'borrow strength' combine these two sources of information by using weights that reflect the relative precision of the two estimators (Ghosh and Rao 1994, Longford 1999). In disease mapping, 'borrowing strength' methods have been further adapted so that instead of borrowing strength from the national (or 'region wide') estimate, it is only borrowed from the geographically adjacent region (Elliott et. al. 1992).

There may be particular problems of accessibility to particular sample areas as well as the collection costs associated with adopting a dispersed distribution of sampling locations versus a more clustered distribution. Other issues may also need to be considered such as whether the sample design is required to undertake long term monitoring and whether there are to be several measured attributes recorded at each site so that some compromise is required between the different estimators (Haining 1990). The form of the sampling problem may be to decide how best to add to, or conversely rationalize an existing sampling network where there may be costs associated with any reconfiguration (Arbia and Lafratta 1997).

There are situations where some spatial object such as a polygon or a line is to be estimated, in which case the sample design is selected in order to estimate or represent characteristics of the object—such as the area of a polygon or the location of a line. In geographic information system databases the representation and resolution of boundary lines between polygons is the outcome of a sampling process. The precision with which any boundary or shape can be represented is a function of the spacing of the sample points (Griffith 1989).

1. Design and Model Based Approaches to Spatial Sampling

Suppose a region R is divided into a population of N units, where N might be very large. In the design-based approach it is assumed that $z(s_i)$ is the value of Z observed at the point s_i which is one of the N points in R. The value of Z at each of the N units is considered fixed. We let m(R) denote the mean of the population of these N values. The problem is to estimate the parameter m(R) and in the design-based approach Z is observed at n sample points where these sites have been selected according to some sampling plan where units have a known probability of being selected. Repeated sampling according to this scheme generates a distribution of values of the vector **z**. This distribution together with an estimator of m(R) defines a (design-based) sampling strategy. This estimator is used to make inferences about m(R) because, as it is based on **z**, it too will have a distribution of values under repeated sampling.

The model-based approach assumes an underlying random process with an assumed probability distribution that is generating the observed set of values in the N units of R. (See *Spatial Autocorrelation*; *Spatial Statistics*.) So the value $z(s_i)$ is the value of a random variable–it is only one of a distribution of values that could have been realized at location s_i. It is usual to specify properties of the model generating the N data values including the first order (mean, μ) and second order (variance-covariance, Σ) properties. Now m(R) is not a parameter but a random variable because it is a function of N random variables. To make clear the distinction this mean will be labeled as M(R) in the model-based case. It could also be of interest to estimate parameters such as the mean, μ, of the underlying model.

In the model-based approach the estimator of μ and the estimator (strictly speaking the predictor) of M(R) both incorporate the spatial model generating the data explicitly in order to produce good estimates (e.g., Haining 1988, Cressie 1991). For example, the best linear unbiased predictor of M(R) is of the form:

$$\sum_{i=1}^{n} a(i)\, z(s_i) \tag{1}$$

where the $\{a(i)\}$ depend on the model and are chosen to minimize the mean square prediction error. In this approach the sampling plan is of secondary interest.

By contrast in the design-based approach, the sample mean (a(i) = 1/n in Eqn. (1)) is used to estimate m(R) having been shown to have good properties (Bellhouse 1977). Attention focuses here on the sampling plan in order to identify which one will have the lowest sampling variance or sampling error. The advantage of this approach is that the estimator is easy to compute and it does not depend on specifying the underlying model. The model-based approach assumes a hypothetical super-population of possible (but unrealized) maps. The design-based approach is only concerned with the here and now of what has occurred. Perhaps more importantly from a conceptual perspective, the model-based approach can be used to estimate the parameters of the underlying signal (the parameters of the map generator, like μ) rather than the particular (noisy) map. For example, in relative risk estimation (also called disease mapping), it is a quantity like μ rather than M(R) that is usually of interest (Elliott et al. 1992).

2. *Sampling Plans*

There are a number of different ways to select the sites to be sampled. The main classes of sampling plan are random, stratified random, and systematic (Ripley 1981). We consider each of these in turn before turning to some other schemes.

In the case of random-sampling the n sample sites are selected so that each member of the population has an equal and independent chance of selection. Figure 1(a) shows a random sampling plan on a surface where each point on the surface had an equal chance of being selected. As has been commented before, a random sample often gives the appearance of some clustering in the distribution of selected sites. Figure 1(b) shows a stratified random sampling plan on the same surface that has been divided into nine strata. Within each stratum, one sample site has been selected according to the rule for a random sampling plan but more than one could be selected. Figure 1(c) shows one type of systematic sampling—the centric systematic sampling plan. The region has been divided into square strata and the central point in the strata selected. This method of sampling contains no randomization. However, randomization can be introduced by for example, randomizing on the location of the first sample point (the random start method). The other $(n-1)$ sample points occupy the same relative locations in the remaining $(n-1)$ strata. There are a number of variants of the systematic sampling scheme (Koop 1990). Stratified random and systematic sampling both raise questions about the size and shape of the strata to use. Shapes such as the hexagon or the triangle could be used for the strata. An understanding of the process generating the quantities of interest may help in designing these parameters of the sampling plan (Cressie 1991). A pilot survey may be important

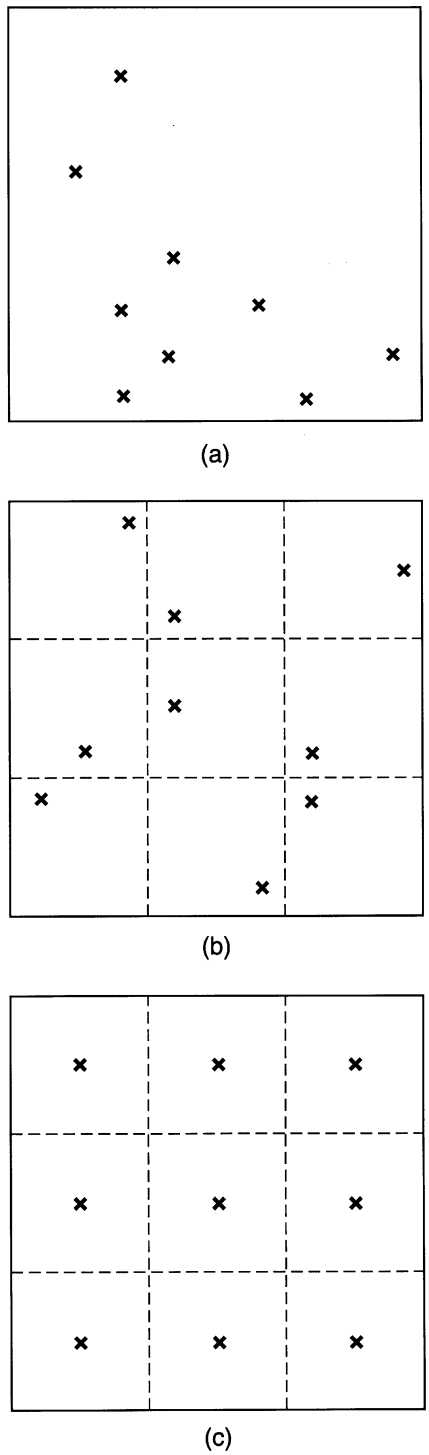

Figure 1

(a)

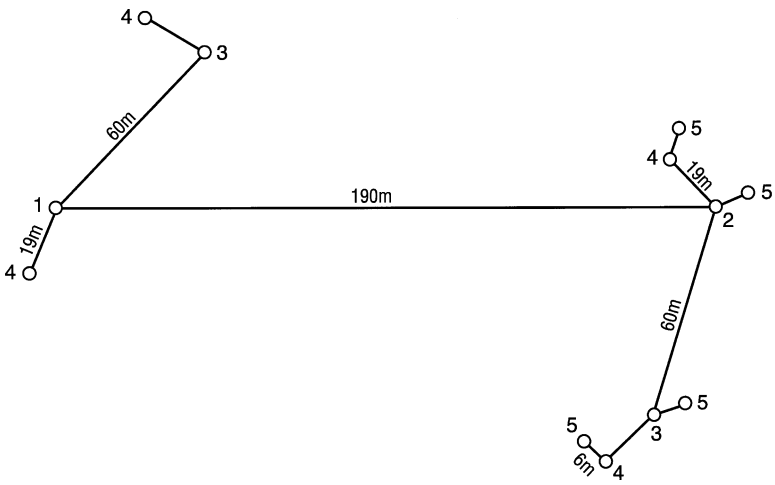

(b)

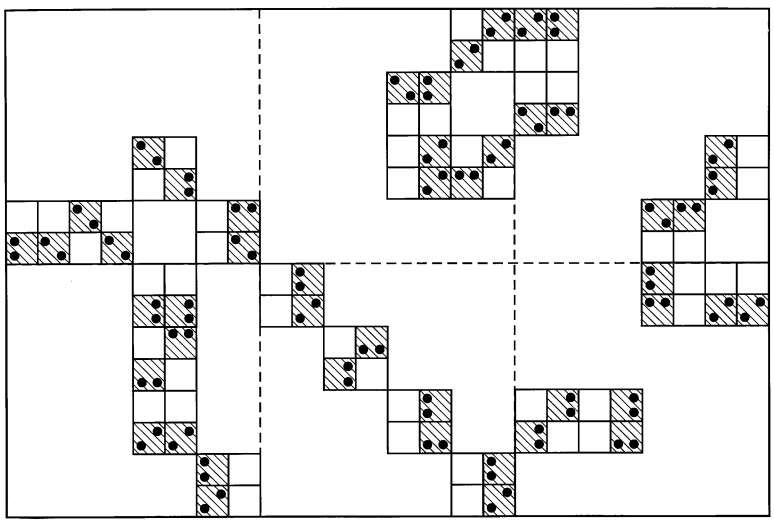

Figure 2

in helping to decide on the size of the strata and hence what density of points are needed.

In addition to these sampling plans there are others that may be appropriate in particular circumstances although there are sometimes special problems associated with computing sampling variances and hence interpreting results. Cluster sampling involves grouping the population into clusters (Lyberg and Cassell). These clusters are often defined in such ways that subsets of individuals comprising the population are spatially close together within any one cluster. These clusters are sampled and then the sample is drawn by selecting, usually at random, individuals from within the selected clusters. This sampling method is often cheaper and administratively easier to implement. Kahn and Sempos (1989) give the example of a health survey involving a large number of economically similar but geographically dispersed villages. Instead of drawing the sample of size n from all the villages (which could involve visiting a large number incurring considerable time and travel costs), a subsample of villages is drawn and the sample of size n is drawn from within this subset of villages.

Nested sampling may be used where variation in the quantity of interest is associated with different spatial scales–for example, different administrative tiers of a political system. Nested sampling may be employed, again to help keep down the time and cost of the

sampling exercise. The area is subdivided at the first (highest) tier of the system and units are chosen at random (similar to cluster sampling). Then these selected areas are subdivided at the second tier and these in turn are sampled—and so on until the lowest tier is reached. Within the selected areas at the lowest tier, individuals are sampled (Fig. 2(a)). As can be seen from Fig. 2(a) this method does not control for distance between the sample points so that if scales of variation vary continuously (rather than in terms of discrete hierarchical levels) as might be likely to arise with a surface of values, then this method of sampling may obscure the different scales of variation. Fixed interval point sampling as illustrated in Fig. 2(b) avoids this problem. The distance between sample points is fixed at those intervals the investigator wishes to study closely (Oliver and Webster 1986).

3. *Spatial Variation and Sampling Plans*

Spatial autocorrelation is the term used to describe the presence of systematic spatial variation in a variable and positive spatial autocorrelation, which is most often encountered in practical situations, is the tendency for areas or sites that are close together to have similar values. The presence of spatial autocorrelation in a population to be sampled has important implications for the sampling variance of the estimator. It is however necessary to distinguish between estimating surface properties in the case of the design-based approach (for example, estimating $m(R)$) and estimating underlying parameters in the case of the model based approach (for example estimating μ). We consider the design-based approach first.

If the purpose is to estimate $m(R)$ then the presence of positive spatial autocorrelation implies that each sample point carries information not only about its own sample site but also some information about neighboring site values. In the case of random sampling, the sampling variance of the sample means (as the estimator of $m(R)$) for any given sample size will be lower than that arising in the case of a population of independent individuals (Griffith et al. 1994). However, this does raise the question as to whether other sampling plans might do better.

Consider the three sampling plans described in Fig. 1. If the surface of values is positively spatially autocorrelated one of the problems with random sampling is that because adjacent values will be similar there will be information redundancy in the case of those sample points that happen to be close together. It is also evident from Fig. 1(a) that the random sampling plan leaves areas unsampled. Both these problems with the random sampling plan are reduced to some extent by the introduction of strata, but as is clear in Fig. 1(b) there are still some sample points that are close together resulting in information redundancy, and there are still some areas of the map not sampled—albeit smaller in extent. Systematic sampling, especially using the centric or the random start methods reduces both effects. Intuitively we therefore expect that for given sample size n, systematic sampling ought to produce estimates with lower sampling variances of a quantity such as m(R). This, by and large, proves to be the case because separating the sample sites improves the information content of the sample. However, while the gains of moving from random to either stratified random or systematic sampling are considerable, the gains of systematic sampling over stratified random sampling are generally much smaller and seem to depend quite critically on the specific form of the underlying spatial autocorrelation (Ripley 1981, Matern 1986, Dunn and Harrison 1993).

The sampling variance is estimated in the case of random sampling using:

$$\sum_{i=1}^{n} [z(s_i) - \bar{Z})]^2 / (n-1) \tag{2}$$

where

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} z(s_i)$$

Equation (2) provides an unbiased estimator in this case and Ripley (1981) discusses related methods for estimating the sampling variance under stratified random and systematic sampling.

The presence of spatial autocorrelation has implications for the other sampling plans. In the case of cluster sampling a critical factor is between-cluster and within-cluster variation. If the clusters are just random assemblages of individuals in the population then cluster sampling (with random selection of individuals from within each cluster) ought not to produce sampling variances any greater than say random sampling. If all the clusters differ greatly then selecting a sample of them will not lead to good population estimates. If clusters have internal positive spatial autocorrelation (so that they are more homogeneous than the population as a whole) then cluster-sampling variance is likely to be bigger than (simple) random sampling. In the case of (hierarchical) nested sampling, if the population shows spatial autocorrelation then, again there will be information redundancy because some sample sites are close together. This effect may be somewhat less of a problem in the case of fixed interval point sampling as a result of keeping sample sites apart.

In the model-based approach, if the purpose is to estimate μ (of a spatially autocorrelated model) using the sample mean, then this estimator's sampling variance is inflated relative to its sampling variance under the assumption of a population that is independent. If Eqn. (2) is used as the estimator of

the sampling variance it can severely underestimate the true sampling variance of the sample mean. Although the sample mean is an unbiased estimator of μ, the best linear unbiased estimator (BLUE) is usually preferred although it requires the correct specification of the underlying spatial model (see Eqn. (1)). As noted above the choice of sampling plan tends to be of secondary interest in this area of analysis but careful specification of the required sample size is needed if parameters are to be estimated to the desired level of precision (Haining 1988).

In addition to undertaking sampling to estimate non-spatial properties of an attribute, sampling may be needed in order to estimate spatial properties (the correlogram or semi-variogram) or to construct maps based on optimal interpolation. Evidence suggests that particularly for optimal interpolation and for constructing contour maps for a variable, centric systematic sampling grids are best with a density chosen to achieve required levels of precision and with a rectangular grid selected for practical reasons (Webster and Burgess 1981). Interpolation error will vary across the map as a function of the density of sample points and the effect of the survey boundary. If anisotropy is present (directional differences in the autocorrelation structure) then both alignment and grid specification for the systematic sampling will need to be adjusted (McBratney et al. 1981).

See also: Sample Surveys, History of; Sample Surveys: Methods; Sample Surveys: Model-based Approaches; Sample Surveys: Nonprobability Sampling; Spatial Analysis in Geography; Spatial Association, Measures of; Spatial Autocorrelation; Spatial Data; Spatial Data Infrastructure; Spatial Pattern, Analysis of; Spatial Search Models; Spatial Statistical Methods; Spatial-temporal Modeling

Bibliography

Arbia G, Lafratta G 1997 Evaluating and updating the sample design in repeated environmental surveys: Monitoring air quality in Padua. *Journal of Agricultural, Biological and Environmental Statistics* **2**: 451–66

Bellhouse D R 1977 Some optimal designs for sampling in two dimensions. *Biometrika* **64**: 605–11

Cressie N 1991 *Statistics for Spatial Data*. Wiley, New York

Dunn R, Harrison A R 1993 Two-dimensional systematic sampling of land use. *Applied Statistics* **42**: 585–601

Elliott P, Cuzick J, English D, Stern R 1992 *Geographical and Environmental Epidemiology: Methods for Small Area Studies*. Oxford University Press, Oxford, UK

Ghosh M, Rao J N K 1994 Small area estimation: An appraisal. *Statistical Science* **9**: 55–93

Griffith D A 1989 Distance calculations and errors in geographic databases. In: Goodchild M, Gopal S (eds.). *Accuracy of Spatial Databases*. Taylor and Francis, New York, pp. 81–90

Griffith D A, Haining R P, Arbia G 1994 Heterogeneity in attribute sampling error in spatial data sets. *Geographical Analysis* **26**: 300–20

Haining R P 1988 Estimating spatial means with an application to remotely sensed data. *Communications in Statistics: Theory and Methods* **17**: 573–97

Haining R P 1990 *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, UK

Kahn H A, Sempos C T 1989 *Statistical Methods in Epidemiology*. Oxford University Press, Oxford, UK

Koop J C 1990 Systematic sampling of two-dimensional surfaces and related problems. *Communications in Statistics: Theory and Methods* **19**: 1701–50

Longford N T 1999 Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society, A* **162**: 227–45

Matern B 1986 *Spatial Variation*, 2nd edn. Springer Verlag, Berlin, Germany

McBratney A B, Webster R, Burgess T M 1981 The design of optimal sampling schemes for local estimation and mapping of regionalized variables—I Theory and method. *Computers and Geosciences* **7**: 331–4

Oliver M A, Webster R 1986 Semi-variograms for modeling the spatial pattern of landform and soil properties. *Earth Surface Processes and Landforms* **11**: 491–504

Ripley B 1981 *Spatial Statistics*. Wiley, New York

Webster R, Burgess T M 1981 Optimal interpolation and isarithmic mapping of soil properties: III changing drift and universal kriging. *Journal of Soil Sciences* **32**: 505–24

R. P. Haining

Spatial Search Models

Spatial search is recognized as an important construct within behavioral geography because it is the preliminary, information acquisition phase of most locational decision processes. Spatial search is the task of identifying and investigating spatially distributed choice alternatives that serve as the basis for a locational decision. Spatial search processes range from purely mechanical, space covering procedures characteristic of remote sensing methods through complex, multi-stage searches drawing on indirect information sources as well as direct experience—the search for a new residence or a new place of employment, for example. No less important is the ongoing acquisition of geographic information which occurs as people move about their environment, consciously and unconsciously creating a practical knowledge base which they draw upon to inform the myriad locational decisions made during the normal course of a day—way finding and shopping decisions, for example (for a detailed review of the spatial search literature and for further references see Golledge and Stimson 1997).

Whether searching for a new place to live or shopping for a new pair of shoes, the set of locations visited during the search process describes a search pattern with predictable spatial characteristics and

important locational consequences. Four spatial regularities in the search patterns are particularly important: search patterns are limited in geographic scope and search effort is spatially persistent; the search patterns reflect the underlying spatial distribution of choice alternatives within the geographically limited search domain; the search pattern reflects the spatial biases in the information sources used to identify potential choice alternatives; and the search pattern is oriented around key locations or anchor points in the decision-maker's activity space. By implication, the spatial regularities and biases observed in residential mobility patterns, local labor mobility patterns, and in shopping patterns are reflections of systematic spatial regularities in the underlying search processes and in their associated search patterns.

Spatial search emerged as a major theoretical construct in behavioral geography in the 1960s as attention shifted from descriptive questions relating to the patterns of geographic activity to an interest in understanding the spatial choice and locational decision-making processes generating the geographic pattern. The early conceptualizations of the locational decision-making process by Wolpert (1965) and Brown and Moore (1970) emphasized the importance of understanding how search decisions affect both the probability of success and the geographic distribution of the locational alternatives included in the decision maker's choice set thereby affecting the location of the final choice. They argued that the first key decision in a locational decision process is the decision to begin active search for attractive alternatives. The decision to begin active search implies that: the decision maker is dissatisfied with the current situation; better alternatives are believed to exist within the decision maker's prospective search domain; and the expected returns to search are greater than the expected search costs. Subsequent work on the decision to begin the search process emphasizes the context dependent nature of the decision and the need to consider the prior knowledge and experience which the decision-maker brings to the spatial search decision.

Three search questions are triggered by the decision to begin a spatial search; the decision-maker must decide: which information sources should be used; where search effort should be focused; and when to terminate the search. Subsequent theoretical and empirical work on spatial search has been primarily concerned with the locational consequences of the decisions made in response to one or more of these three questions. Each of these questions has a geographic consequence and each decision is affected by the interaction between the locational preferences of the individual and the decision-making context.

The delineation of a geographically bounded search domain reflects the locational preferences of the decision-maker but is also closely tied to the choice of information source or sources used to identify search possibilities. The primary information source affecting the delineation of the search domain is the practical knowledge base that the individual brings to the search task. If an area is familiar to a person, search is more likely in that area; the search costs are likely to be lower as is the risk of making a bad choice. The search domain tends to be defined in terms of the individual's activity and the awareness space that varies in geographic scope depending upon the age, class, and ethnicity of the individual. Similar observations pertain for personalized information sources such as friends and relatives. Households relying on personalized sources of information are found to have much smaller search domains than individuals relying on impersonal sources such as newspapers and directories or professional sources such as real estate agents or employment agencies (Clark and Smith 1979).

In these instances, the search domain is dictated by the geographic reach of the information source. The role of information channels as sources of bias in the search process has received considerable attention in the racial segregation literature (Courant 1978). These investigations have focused specifically on racial steering and the potential role of real estate agents in creating racially homogeneous neighborhoods through selective filtering of housing information to black and white households. It is important to understand how geographic information is produced because whether systematic and mechanical or selective and subjective, spatial search processes produce representations of the external environment which are necessarily incomplete and potentially distorted both as a result of the geographically selective knowledge base and because information channels also selectively represent or filter the locational information available to the decision-maker. Most locational decisions and the search process leading up to the actual choice are bounded geographically in that the active investigation of choice alternatives is preceded by a decision to focus the search effort within a particular geographically delineated market. Once search begins within a particular market area, there is a strong tendency to continue searching within that same market area. Once the search for a new residence begins within a local housing market, for example, it is highly likely that search will continue within that same housing market. Similar patterns of spatial persistence are observed in job searches within local labor markets and in shopping decisions within local market areas. In other words, people continue to focus search effort within a targeted market area as a spatial satisfying strategy that discounts the potential acceptability of choice alternatives located outside the targeted market area. There is also evidence that area based search strategies continue to apply at more micro scales within the larger market area. In the case of residential search, for example, households tend to focus their search effort within specific community areas within

larger geographically defined housing markets (Huff 1986).

Decision makers decide where to focus their search effort based upon their locational preferences and upon what they know about the chances of being able to satisfy locational and nonlocational preferences by targeting search within a given area. Often there are pre-existing anchor points that have locational significance for the decision-maker and serve to focus or anchor the search effort (Golledge 1978). When households search for a new residence, for example, the job locations of the wage earners in the household and the location of the previous residence often serve as strong anchor points in the search process (Huff 1986). In the case of job search, the location of the residence often ties the search to a particular labor market and may result in a highly focused job search in the immediate proximity of the home (Hanson and Pratt 1988). In the case of search associated with shopping, home and work locations frequently are identified as key anchor points in the search process and at a more micro level, the strategic locations of anchor stores in shopping centers reflect an awareness of how these stores anchor the search behavior of shoppers within the mall.

How people decide when to conclude the search process has received a great deal of attention within the economics literature on job search (Lippman and McCall 1976). Conceptually, the answer to the stopping-rule question is quite simple: search will continue until the expected cost of further search exceeds the expected returns to further search. On the other hand, the empirical validation of stopping rule models has proven to be quite difficult. In the case of spatial search, both the expected costs and the expected returns to search are expected to vary from one geographic area to another (Smith et al. 1979) which implies that the length of search and the chance of successfully completing the search vary from one geographic area to another (Rogerson 1982).

Understanding the path dependent nature of the spatial search process is of vital importance with respect to the stopping rule question as well as the information use and the search location questions. In the stopping-rule case, the order in which choice alternatives are visited during the search process is found to affect the outcome of the search process. Different information sources with very different content and spatial coverage tend to be used during different phases of the search process. During the actual search process, the path dependencies often are spatial in nature. Spatial persistence in the decision to continue searching in the same area, the spatial clustering of choice alternatives visited during search highlight the existence of strong spatial dependencies in the search sequence. In the case of multi-purpose shopping behavior, spatial search often requires the resolution of both a routing or way finding problem and a stopping rule problem (Miller 1993).

Bibliography

Brown L A, Moore E A 1970 The intra-urban migration process: A perspective. *Geografiska Annaler* **52B**: 1–13

Clark W A V, Smith T R 1979 Modelling information use in a spatial context. *Annals of the Association of American Geographers* **69**: 575–88

Courant P 1978 Racial prejudice in a search model of the urban housing market. *Journal of Urban Economics* **5**: 329–45

Golledge R G, Stimson R J 1997 *Spatial Behavior: A Geographic Perspective*. Guilford Press, New York

Hanson S, Pratt G 1988 Reconceptualizing the links between home and work in urban geography. *Economic Geography* **64**: 299–321

Huff J O 1986 Geographic regularities in residential search behavior. *Annals of the Association of American Geographers* **76**: 208–27

Lippman S A, McCall J J 1976 The economics of job search: A survey. *Economic enquiry* **14**: 155–89

Miller H 1993 Modeling strategies for the spatial search problem. *Papers in Regional Science* **72**: 63–85

Rogerson P 1982 Spatial models of search. *Geographical Analysis* **14**: 217–28

Smith T R, Clark W A V, Huff J O, Shapiro P 1979 A decision making and search model for intraurban migration. *Geographical Analysis* **11**: 1–22

Wolpert J 1965 Behavioral aspects of the decision to migrate. *Papers and Proceedings of the Regional Science Association* **15**: 159–69

J. O. Huff

Spatial Statistical Methods

Spatial statistics is characterized by modeling and data analysis for processes displaying spatially-indexed dependence structures. Adjustment for spatial dependence has long been mandated in many disciplines. The very nature of geography is spatial. F. Stephan (1934) noted 'Data of geographic units are tied together … in dealing with social data, we know that by virtue of their very *social* character, persons, groups, and their characteristics are interrelated and not independent.' In the context of agricultural field experiments, Fisher (1935) described the basic intuition underlying much of spatial statistics: 'After choosing the area we usually have no guidance beyond the widely verifiable fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are far apart.' Though Fisher developed randomized experimental designs to avoid modeling spatial dependence explicitly, agricultural applications are an important motivation for spatial statistics (Besag 1991).

In the early 1950s the South African mining engineer D. G. Krige pursued methods for the spatial interpolation of data with the intent of predicting the distribution of ore (e.g., gold) in space. G. Matheron

(France) extended that work, formalizing mining geostatistics, and coined the term 'kriging' to refer to a class of spatial prediction methods. In parallel, L. S. Gandin (Soviet Union) developed essentially the same procedures for applications in meteorology where they are known as *optimal interpolation* and *objective analysis*.

While basic geostatistical methods remain very useful, the methodology was limited in terms of (a) applicability, since the analyses were historically confined to linear models and (b) volume of data analyzed. Technological advances regarding both data collection methods and computation have led to expansion of applications and methods of spatial statistics. Specifically, image analysis is and will continue to be a critical research area. Example problems include satellite remote sensing of the Earth's environment and ecology, including humans and their behavior and impacts; analysis of astronomical images from optical, radio-wave, infrared, etc., instruments; magnetic resonance imagery; and computer vision. Methods for image restoration (filtering out noise), classification, and object identification are sought. Similarly, the development of geographic information systems is a watershed of geographically referenced data, though the efficient extraction of information from such massive datasets offers new statistical challenges.

Spatial statistics also includes modeling and analysis of mapped point processes, in which the locations of observations are themselves random. For example, observations may be the locations and characteristics of trees in a forest (Stoyan and Penttinen 2000), or positions and types of galaxies in space (Feigelson and Babu 1992). For applications in archaelogy, see Hodder and Orton (1976); see also Stoyan (1988). Finally, an introduction to spatial modeling of random objects can be found in Cressie (1993); see also Grenander (1993).

1. Perspectives and Comparisons

Let D denote a collection of sites or locations denoted by $\mathbf{s}$, distributed in space. (Formally, D is taken to be a subset of d-dimensional Euclidean space. This permits mathematical consideration of the differences and distances between sites). In geostatistics, D is a collection of sites on the surface of the Earth, indexed in two dimensions either by $\mathbf{s} = (\text{latitude, longitude})$ or by Cartesian coordinates $\mathbf{s} = (x, y)$, or in three dimensions with the addition of altitude or depth. In image analysis, $\mathbf{s}$ typically represents a pixel located on a discrete lattice. For each $\mathbf{s}$ in D, there exists some random quantity, $Z(\mathbf{s})$, of interest. The stochastic process $\{Z(\mathbf{s}):\mathbf{s} \in D\}$ is a random field or spatial random process. Z could be a continuous variable (e.g., temperature); a discrete variable (e.g., a binary variate indicating the presence or absence of some characteristic, or a multivalued discrete variable of gray levels, especially in image analysis); a mixed variable (e.g., rainfall may be zero with positive probability at a location, yet continuous, if some rain occurs). Z may also be a multivariate variable.

Goals in spatial statistical analyses include distributional modeling of spatial random processes; estimation and inference for aspects of the spatial model based on observational data; prediction at sites where observations are not available; and filtering or noise reduction. Two approaches to spatial process modeling dominate current practice. In the first, jointly or simultaneously specified models for the process $\{Z(\mathbf{s}):\mathbf{s} \in D\}$ are developed. For example, in much of geostatistics, the variable $Z(\mathbf{s})$ is decomposed into 'mean' plus 'error' components. A critical challenge is the development of joint distributional information for the errors across space. A second strategy is to build conditionally specified models by modeling local dependence structures. This track builds directly on the intuition that variables in close spatial proximity tend to be related. This is quantified by formulating distributional features of $Z(\mathbf{s})$, conditional on the rest of the Z process, to depend only on the values of Z at locations in neighborhood of $\mathbf{s}$.

Practical modeling often focuses on specification of conditional means and variances. Two important issues arise. First, to enable spatial prediction accompanied by meaningful measures of uncertainty, a conditionally specified model should correspond to a proper joint distributional model. To that end, conditionally specified Markov random fields are often used since there is a well-developed theory regarding the existence of implied joint models (e.g., Cressie 1993, Chap. 6). The second issue involves dealing with variables located near the boundary of D, since this region typically dictates local dependence structures that are by nature different from those in the interior of D. Indeed, boundary problems are among the issues that generally distinguish spatial statistics from other applied statistics areas.

While geostatistics and imaging appear to have developed separately, they share some common formulations and goals. However, modeling and analysis methods are responsive to the typical types of data available. Geostatistical settings often feature small to moderately sized datasets, and data that are irregularly distributed in space, leading to difficulties in model estimation. In many imaging problems datasets are relatively large and are often regularly distributed in space. One impact of this difference is that treatment of measurement error is more accessible and common in imaging compared to its haphazard treatment in classical geostatistics.

The primary goal, particularly in geostatistics, is spatial prediction: Based on a vector of n observations, $Z \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$, we seek predictions of $Z(\mathbf{s}_0)$ where $\mathbf{s}_0$ is an arbitrary site in D. Depending upon the level of assumptions made, two essential principles guide the predictive analysis. First, in analyses based

on assumptions asserting only the trend and covariance structures, most predictions are derived as best linear unbiased predictions (BLUP). 'Best' refers to the property of having smallest prediction variance among all linear, unbiased predictors. In analyses which also include distributional assumptions, predictive analyses are based on the implied conditional distribution of $Z(\mathbf{s}_0)$ given the observed data Z. For example, assuming mean squared prediction error as the criterion for assessing the quality of prediction, the optimal predictor is $\hat{Z}(\mathbf{s}_0) = E(Z(\mathbf{s}_0) \mid Z)$; that is, the conditional mean of $Z(\mathbf{s}_0)$ given the data.

Both Bayesian and non-Bayesian viewpoints are applied in spatial statistics, though the use of these terms varies with the application areas. Since most strategies view the primary variable of interest Z as *random*, the principle that spatial prediction is based on the conditional distribution, or at least its moments, of the unobserved variables given the data is clearly tied to the Bayesian view in both interpretation and computational techniques. This is particularly evident in image analysis, where a potentially nonrandom, 'true image' is modeled as a realization of a stochastic spatial process (e.g., Geman and Geman 1984, Besag 1989). On the other hand, despite similar stochastic modeling of the primary variables of interest, the origins of geostatistics and spatial point process analysis appear to be primarily non-Bayesian in language and spirit.

In particular, early approaches to model fitting and parameter estimation were primarily non-Bayesian, though their interpretation as empirical Bayesian is quite natural. With the advent of powerful Bayesian computational strategies (i.e., Markov chain Monte Carlo), the Bayesian approach is undeniably on the rise in applications of all areas of spatial statistics. It offers opportunities for effective analyses in complex spatial problems, including multivariate and inhomogeneous datasets and data types, as well as highly nonlinear and non-Gaussian datasets, not readily amenable to competing viewpoints.

2. *Spatial Random Fields*

Assuming all quantities exist, let $\mu(\mathbf{s})$, $\mathbf{s} \in D$ be the mean (expected value) $\mu(\mathbf{s}) = E(Z(\mathbf{s}))$, and let $\mathrm{cov}(Z(\mathbf{s}), Z(\mathbf{s}'))$, defined for all pairs of sites $\mathbf{s}$ and $\mathbf{s}'$ in D, be the covariance function of the process. The process is (second-order or weakly) stationary if it has constant mean $\mu(\mathbf{s}) \equiv \mu$ and

$$\mathrm{cov}(Z(\mathbf{s}), Z(\mathbf{s}') = C(\mathbf{h}), \quad \text{for all } \mathbf{s} \in D,\ \mathbf{s}' \in D \quad (1)$$

where $\mathbf{h} = \mathbf{s} - \mathbf{s}'$. That is, the covariance between $Z(\mathbf{s})$ and $Z(\mathbf{s}')$ depends on the locations only through their difference. Note that this implies that the variance of $Z(\mathbf{s})$, $\sigma^2(\mathbf{s}) \equiv \sigma^2$, is constant. The function $C(\mathbf{h})$ is known as the *covariogram*. Strict or strong stationarity involves similar requirements applied to the joint distributions of $\{Z(\mathbf{s}):\mathbf{s} \in \mathscr{D}\}$ as opposed merely to first and second moments (Cressie 1993, Sect. 2.3).

Alternatively, again assuming that $\mu(\mathbf{s}) \equiv \mu$, the process is intrinsically stationary if

$$\mathrm{var}(Z(\mathbf{s}) - Z(\mathbf{s}')) = 2\gamma(\mathbf{h}), \quad \text{for all } \mathbf{s} \in D,\ \mathbf{s}' \in D \quad (2)$$

The function $2\gamma(\mathbf{h})$ is known as the *variogram*; $\gamma(\mathbf{h})$, is the *semivariogram*.

If the process is stationary, then calculation implies

$$0.5\ \mathrm{var}(Z(\mathbf{s}) - Z(\mathbf{s}')) = C(\mathbf{0}) - C(\mathbf{h}) = \gamma(\mathbf{h}) \quad (3)$$

Hence, stationarity implies intrinsic stationarity. However, variograms satisfying (2) may exist in some cases for which covariograms do not exist. For stationary processes, (3) allows one to view modeling covariograms and variograms as being equivalent, though some have argued that estimation based on data is enhanced for the variogram.

An intrinsically stationary process is isotropic if $\gamma(\mathbf{s} - \mathbf{s}') = \gamma(h)$ for all pairs of sites, where h is the distance between them. That is, the variogram depends only on the distance between sites (i.e., the length of the vector $\mathbf{h} = \mathbf{s} - \mathbf{s}'$), but not their relative directions. Anisotropy (the violation of isotropy) arises for processes that develop differentially in different directions of space (Cressie 1993, Sect. 2.3). For example, temperature values, or variables depending on temperature (e.g., type of agriculture, ecology, etc.), are moderated roughly by latitude. Hence, north–south covariance structures might well be different from their east–west counterparts. An example of an isotropic semivariogram is the exponential model:

$$\gamma(h) = \begin{cases} 0 & \text{if } h = 0 \\ \gamma_0 + \gamma_1(1 - e^{-h/a}) & \text{if } h > 0 \end{cases} \quad (4)$$

where $a > 0$. Note that as $h \to 0$, $\gamma(h) \to \gamma_0$, rather than 0. This discontinuous behavior, known as the *nugget effect*, occurs in a variety of commonly used semivariograms. Physically, γ_0 has been interpreted as arising from (a) very small-scale fluctuations in the Z process, and/or (b) measurement error. However, if the process modeled is a sufficiently smooth function of space, the only mathematically valid explanation for the nugget effect is measurement error.

The exponential covariogram corresponding to Eqn. (4) is

$$C(h) = \begin{cases} \gamma_0 + \gamma_1 & \text{if } h = 0 \\ \gamma_1 e^{-h/a}, & \text{if } h > 0 \end{cases} \quad (5)$$

Another covariogram sometimes used is the Gaussian model.

$$C(h) = \begin{cases} \gamma_0 + \gamma_1 & \text{if } h = 0 \\ \gamma_1 e^{-h^2/b}, & \text{if } h > 0 \end{cases} \tag{6}$$

where $b > 0$. (Use of the modifiers 'exponential' and 'Gaussian' accrues from similarities to the exponential and Gaussian probability density functions, rather than distributional assumptions for Z.) A very rich class of isotropic covariograms is the Matérn class, written here without a nugget for comparison:

$$C(h) = \frac{2\sigma^2}{\Gamma(\nu)}(h\sqrt{\nu}/a)^{\nu} K_{\nu}(2h\sqrt{\nu}/a) \tag{7}$$

where $\nu > 0$, σ^2 is the variance of $Z(\mathbf{s})$, $\Gamma(\cdot)$ is the usual gamma function, and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind.

3. Spatial Prediction

A standard spatial model takes the form

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \varepsilon(\mathbf{s}) \tag{8}$$

Where $\mu(\mathbf{s})$ is a spatial mean model and $\varepsilon(\mathbf{s})$ is a zero-mean spatial error process. Critical issues involve the choices of models regarding both the mean or trend (also known as drift) $\mu(\mathbf{s})$ and the error process $\{\varepsilon(\mathbf{s}): \mathbf{s} \in \mathscr{D}\}$. Such modeling is challenging in that both components explain aspects of the spatial structure present in Z. The conventional guideline is that the trend reflects large-scale spatial variation, while the error process explains residual small-scale spatial dependence. Implementation of this guideline and interpretation of results vary with the context as well as with the goals and knowledge of the analyst. As noted by Cressie (1993, p. 25), 'What is one person's (spatial) covariance structure may be another person's mean structure.' Also note that in modern statistical approaches, trend models based on both spatial location and covariates are recommended. Trend models can also be incorporated in conditionally specified models, either via conditionally specified mean models, or with globally specified mean models and conditionally specified error models.

3.1 The Basics of Kriging

Ordinary kriging asserts a constant trend model. Linear predictors take the form

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) \tag{9}$$

The unbiasedness condition, $E(\hat{Z}(\mathrm{s}_0)) = \mu$ adds the side condition that $\sum_{n=1}^{n} \lambda_i = 1$. Minimization of the predictive variance (or mean squared prediction error)

$$E(Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0))^2$$

subject to the side condition yields simple formulas for the optimal values of the λ_i as appropriate functions of the covariance function. Alternatively, the formula can be written in terms of the variogram, enabling spatial prediction for processes that are intrinsically stationary (Cressie 1993, pp. 121–3). The achieved optimal value of the predictive variance is then used to measure the precision of the prediction.

The kriging formulas are derived under the assumption that the variogram or covariance function is known. Typically, data-based estimates of these functions are simply plugged into the kriging formulas. A variety of methods for estimation of the variogram (or covariogram) have been suggested. A simple procedure is to apply a method-of-moments notion: an empirical version of the semivariogram is

$$\hat{\gamma}(\mathbf{h}) = \frac{\sum_{n(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_{i'}))^2}{\#n(\mathbf{h})} \tag{10}$$

where the sum is taken over those data points satisfying $\mathbf{s}_i - \mathbf{s}_{i'} = \mathbf{h}$ and $\#n(\mathbf{h})$ is the number of such pairs of data points. A simple estimate of the covariogram is

$$\hat{C}(\mathbf{h}) = \frac{\sum_{n(\mathbf{h})} (Z(\mathbf{s}_i) - \bar{Z})(Z(\mathbf{s}_{i'}) - \bar{Z})}{\#n(\mathbf{h})} \tag{11}$$

where $\bar{Z}$ is the simple mean of the observed values. However, since there are only n observations, these estimates cannot provide estimates for all $\mathbf{h}$. Practical estimates are typically constructed based on binning procedures. For example, under the additional assumption of isotropy, consider a collection of M non-overlapping intervals

$$I_m = [h_m - \delta_m, h_m + \delta_m]; \quad m = 1, \ldots, M$$

For each interval the above empirical sums and counts of pairs are taken over all pairs separated by distances in I_m. The resulting M values can then be smoothed to produce a function estimate.

The above estimates suffer from a variety of deficiencies, especially if the sample size n is small. A variety of alternatives have been suggested. One general approach is to parameterize the covariogram (or variogram), specifying it to be a member of some class (the Matérn class (7) is a popular choice). Methods for parameter estimation include least squares, weighted least squares, and maximum likelihood. A second avenue for estimation involves the use of nonparametric function estimation, especially

in the context of estimating the spectral representation of the covariogram. Also, robust estimates with respect to outlying observations have been developed. See Cressie (1993) and Stein (1999) for further discussion of classes of variograms and estimation procedures. Sampson and Guttorp (1992) consider estimation in the nonstationary case.

3.2 Trend Models and Their Estimation

Universal kriging refers to spatial prediction based on trend models of the form

$$\mu(\mathbf{s}) = \sum_{j=0}^{p} f_j(\mathbf{s})\beta_j \tag{12}$$

where the $f_j(\cdot)$ are $p+1$ specified functions of space and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ is a vector of $p+1$ unknown parameters. In vector notation, let $\mathbf{f}(\mathbf{s}) = (f_p(\mathbf{s}), \ldots, f_p(\mathbf{s}))'$, so the model is written compactly as

$$\mu(\mathbf{s}) = \mathbf{f}'(\mathbf{s})\boldsymbol{\beta}$$

Ordinary kriging is included by assuming $f_0(\mathbf{s}) = 1$ and $p = 0$.

A common example is a quadratic trend surface model: Writing locations in Cartesian coordinates $\mathbf{s} = (x, y)$, assume that

$$\mu(\mathbf{s}) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2$$

In addition to higher-order polynomial trends, nonpolynomial models, such as local polynomials have been considered (Venables and Ripley 1999). The use of splines is highly related; see Wahba (1990) and Cressie (1993, pp. 180–3).

Representation of the kriging model in matrix form provides a better understanding and enables parallels to familiar statistical linear regression modeling. The notation $\mathbf{X} \sim (\mathbf{m}, \mathbf{C})$ is read 'the random vector $\mathbf{X}$ has a multivariate distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$.' The notation $\mathbf{X} \sim N(\mathbf{m}, \mathbf{C})$ adds the assumption that the distribution is multivariate normal. The standard universal kriging model is then

$$\begin{pmatrix} Z(\mathbf{s}_0) \\ \mathscr{Z} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_0'\boldsymbol{\beta} \\ \mathbf{F}\boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \varepsilon(\mathbf{s}_0) \\ \boldsymbol{\varepsilon} \end{pmatrix} \tag{13}$$

where

$$\begin{pmatrix} \varepsilon(\mathbf{s}_0) \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^2(\mathbf{s}_0) & \boldsymbol{\sigma}' \\ \boldsymbol{\sigma} & \boldsymbol{\Sigma} \end{pmatrix} \right) \tag{14}$$

where $\boldsymbol{\sigma}$ is the $n \times 1$ vector of covariances between errors at site $\mathbf{s}_0$ and the observation sites; $\mathbf{F}$ is the $n \times (p+1)$ matrix dictated by the functions $f_p(\cdot)$ evaluated at the observation locations; and $\boldsymbol{\varepsilon}$ is the vector of errors at observation sites. These errors are assumed to have zero means and covariance matrix $\boldsymbol{\Sigma}$ whose entries are the covariances $\mathrm{cov}(\boldsymbol{\varepsilon}(\mathbf{s}_i), (\varepsilon(\mathbf{s}_{i'}))$ for pairs of data points.

Formulas for the optimal linear predictor or BLUP are readily derived (Cressie 1993, pp. 172–80):

$$\hat{Z}(\mathbf{s}_0) = \mathbf{f}'(\mathbf{s}_0)\boldsymbol{\beta} + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\mathscr{Z} - \mathbf{F}\boldsymbol{\beta}) \tag{15}$$

These formulas depend on appropriate values of the covariance function and $\boldsymbol{\beta}$, requiring estimation of both quantities. Note that use of (15) does not require spatial stationarity of the error process, though it is often assumed in estimation.

Estimation of $\boldsymbol{\beta}$ is typically accomplished via generalized least squares. Since the observations follow the model

$$\mathscr{Z} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{16}$$

The generalized least squares estimate of β is

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}\mathbf{F}\boldsymbol{\Sigma}^{-1}\mathscr{Z} \tag{17}$$

Estimation of the covariogram of the ε process proceeds as discussed earlier, though now the data used are the residuals from the fitted trend model: $\mathscr{Z} - \mathbf{F}\hat{\boldsymbol{\beta}}$. However, there is a clear difficulty. The very residuals needed to estimate the covariogram depend on the covariogram since $\hat{\boldsymbol{\beta}}$ is the function of $\boldsymbol{\Sigma}$. A variety of approaches to this problem have been suggested. One is to use some preliminary guess at $\boldsymbol{\Sigma}$ in (17), but base prediction formulas on the estimated covariogram of the resulting residuals. With the increase of readily available computation, the more modern approach iterates between generalized least squares estimation of $\boldsymbol{\beta}$ and covariogram estimation until some stability is achieved (Venables and Ripley 1999). Note that the extra uncertainty accruing from the use of data-based estimates of $\boldsymbol{\beta}$ and the covariance function is seldom accounted for in classical kriging.

Treatment of measurement error can be included in the modeling. Rather than observing $\mathscr{Z}$ directly, the observed data values, denoted by Y, may be modeled as $Y(\mathbf{s}_i) = Z(\mathbf{s}_i) + e(\mathbf{s}_i)$; $i = 1, \ldots, n$. Let $\mathscr{Y}$ denote the $n \times 1$ vector of observations. In matrix form we write

$$\mathscr{Y} \sim (\mathscr{Z}, \mathbf{D}_y) \tag{18}$$

where $\mathbf{D}_y$ is the covariance matrix of the measurement errors $\mathbf{e} = (e(\mathbf{s}_1), \ldots, e(\mathbf{s}_n))'$. A common, but not unequivocal, assumption is that $\mathbf{D}_y = \sigma_y^2\mathbf{I}$ (i.e., the measurement errors are mutually uncorrelated). Also, assume $\mathbf{e}$ is uncorrelated with $\mathscr{Z}$. If Z is a smooth function of space, σ_y^2 is a quantification of the nugget effect. To complete the formalization, interpret (18) as

a specification *conditional* on $\mathscr{Z}$. The model (16) for $\mathscr{Z}$ implies that the *unconditional* model for $\mathscr{Y}$ is

$$\mathscr{Y} \sim (\mathbf{F}\boldsymbol{\beta}, \boldsymbol{\Sigma} + \sigma_y^2 \mathbf{I}) \tag{19}$$

(Compare this covariance matrix to the piece-wise definitions of variograms in (5) and (6).) Without replicate observations at some of the observation locations, σ_y^2 is not directly estimable in classical kriging. Typically, estimates are obtained by extrapolation of covariogram estimates for small distances.

Note that the goals of the analysis now include not only prediction of $Z(\mathbf{s}_0)$, but also prediction of $\mathscr{Z}$. Predicting $\mathscr{Z}$ based on $\mathscr{Y}$ provides spatial smoothing, as opposed to spatial interpolation when σ_y^2 is assumed to be zero.

3.3 Multivariate Methods

In many applications, several variables may be of interest, suggesting the consideration of a random vector, $\mathbf{Z}(\mathbf{s}) = (Z_1(\mathbf{s}), \ldots, Z_K(\mathbf{s}))'$. In mining, the data might consist of measurements of amounts of K minerals at each observed site. In meteorology, data on temperature, pressure and precipitation are often collected. In satellite imaging, multichannel sensors collect intensities in various bandwidths.

The classical geostatistical treatment of multivariate spatial data is known as *cokriging*. The modeling strategy is a direct generalization of the univariate case. For each coordinate $Z_k(\mathbf{s})$ of $\mathbf{Z}(\mathbf{s})$, consider a mean or trend model (see (12))

$$\mu_k(\mathbf{s}) = \mathbf{f}_k'(\mathbf{s})\boldsymbol{\beta}_k, \quad k = 1, \ldots, K \tag{20}$$

Note that the spatial regression functions comprising the various $\mathbf{f}_k$ need not be the same for all k: indeed, the dimensions of the $\boldsymbol{\beta}_k$ may vary with k. The key second-order parameters are the covariance matrices,

$$\operatorname{cov}(\mathbf{Z}(\mathbf{s}), \mathbf{Z}(\mathbf{s}')) = \mathbf{C}(\mathbf{s}, \mathbf{s}'), \quad \mathbf{s} \in D, \mathbf{s}' \in D \tag{21}$$

The principles and basic formulas for best linear unbiased prediction remain essentially the same as for the univariate case. Since the notation is cumbersome, the reader is referred to Wackernagel (1995) for details.

Practical implementation of cokriging can be difficult. First, specification and subsequent estimation of $\mathbf{C}(\mathbf{s}, \mathbf{s}')$ is problematic, in that it involves specification not merely of K spatial covariance functions (one for each variable) but also $0.5K(K-1)$ cross-covariance functions, one for each pair Z_k, $Z_{k'}$, $k \neq k'$. The number of observations critically limits the ability to estimate the required covariance functions. Indeed, sample size typically mandates the need for extremely simple, often implausible, covariance models and/or a reduction in K. Further, the sampling pattern (e.g., observations of only some coordinates of $\mathbf{Z}$ at some sites) can moderate the ability to estimate parameters and predict.

An alternative modeling strategy is to build multivariate models hierarchically, namely as a sequence of conditional models. A precursor to this view from the geostatistics literature is kriging with external drift. For example, suppose that spatial prediction of near-surface ozone is of interest. It is well accepted that temperature is predictive for ozone. Hence, spatial prediction of ozone could be enhanced if appropriate temperature data were available and included in the model. An elementary approach is to view temperature as a (fixed) covariate and to include it in the formulation of a universal kriging model for ozone. Next, suppose temperature observations are unavailable at sites at which ozone predictions are sought. One could essentially form a kriging model for temperature, and rely on temperature predictions to produce ozone predictions. These notions can be made rigorous statistically. Specification of a complete spatial model for ozone conditional on temperature combined with a marginal spatial model for temperature does imply a joint model for both, as might be formed in a cokriging approach. However, the parameterizations of the models can be different. In particular, rather than direct specification and estimation of the cross-covariance function for ozone and temperature, one specifies a conditional mean relationship between the two. For further discussion and examples, see Royle and Berliner (1999).

4. Bayesian Spatial Statistics

4.1 Bayesian Geostatistics

Models for the Z process, such as the basic model (13) along with a measurement error stage as in (18) can readily be posed in the Bayesian framework. Process models for Z are viewed as priors; these priors may be specified simultaneously or conditionally. Specification of prior distributions on model parameters, such as the trend model coefficients $\boldsymbol{\beta}$, measurement error variance, and parameters of the modeled covariance function are also formulated. Predictive inference is then based on the posterior distribution of unknown quantities, given the data. (A general review of Bayesian analysis is given in Bernardo and Smith 1994.)

Direct analogs to conventional kriging often rely on multivariate normal distributions in both (13) and (18). This imposition of distributional assumptions has been criticized. However, neither standard kriging nor Bayesian analyses based on normal distribution assumptions are recommended for highly non-normal data and processes. The Bayesian approach readily offers valid measures of uncertainty (e.g., Le and Zidek 1992, Handcock and Stein 1993), unlike those arising from the 'plug-in' of estimates procedures dominating non-Bayesian kriging.

With the advent of Bayesian computing advances, non-normal models can be analyzed. For example,

settings in which the Z variable is discrete, such as in some imaging problems (Geman and Geman 1984) or is count data modeled as via Poisson or related distributions (Kaiser and Cressie 1997, Wolpert and Icktadt 1998), have been considered. See also Diggle et al. (1998).

4.2 Hierarchical Models

Hierarchical Bayesian models for collections of variables are combinations of component models or stages. This strategy is particularly useful in complex settings. Motivations for hierarchical modeling include:

(a) individual stages are often formulated more easily than is the full joint model directly;

(b) scientific understanding of the relationships among variables is often readily available for components of the model; and

(c) combining datasets consisting of very different types of data is often enhanced.

The first stage of a Bayesian hierarchical model is usually a measurement error model for the observations, analogous to (18). A variety of higher-level stages, interrelating spatial processes and model parameters, can then be formulated. Finally, a prior probability distribution for all model parameters is specified. See Royle et al. (1998) and Berliner (2000) for examples and discussion.

5. Spatial Point Processes

A spatial point process arises when the locations of events occur randomly in D. Basic data take the form $(\mathbf{s}_1, \ldots, \mathbf{s}_N)'$, where the number of locations, N is a random variable. The probability theory for point processes typically formulates a model for random points in $\mathscr{R}^2$. The model then implies a distribution for N, the number of events in the study region D, a bounded subset of $\mathscr{R}^2$, as well as the joint distribution of their locations. Generally, the treatment of edge effects then becomes a challenge (Ripley 1988, Chap. 2).

A standard point process model is the homogeneous Poisson spatial process with intensity λ: For any k and any selection of disjoint subsets $A_1, \ldots, A_k$ of $\mathscr{R}^2$, assume that the numbers of events, $N(A_i)$, $i = 1, \ldots, k$, are independent random variables each with probability

$$P(N(A_i) = n) = \frac{(\lambda|A_i|)^n}{n!} \exp(-\lambda|A_i|), \quad n = 0,1,\ldots \tag{22}$$

where $|A_i|$ is the area of A_i, $\int_{A_i} \mathbf{ds}$. Note that for any set A, the mean $E(N(A)) = \lambda|A|$. The Poisson process is viewed as completely spatially random, since counts in disjoint intervals are independent and the probabilities in (22) depend only on A_i through its area and not its shape or position. This model is often used as a null hypothesis. Rejection of the Poisson model suggests that the process has some spatially dependent structure, or dependence among separated regions. Numerous generalizations have been considered. The Poisson cluster process allows cluster locations to be distributed according to one Poisson spatial process, while events in clusters are distributed according to another Poisson model. Extensions to alternative measures on sets beyond simple area, including measures which vary in space leading inhomogeneous processes have been considered. A Cox process is generated as a Poisson process, but the measures associated with sets are themselves random.

As in geostatistics and related areas, selected moments of a spatial process play roles in both theory and data analysis. Define the intensity function of a point process as

$$\lambda(\mathbf{s}) = \lim_{|\mathbf{ds}| \to 0} \frac{P(N(\mathbf{ds}) > 0)}{|\mathbf{ds}|} \tag{23}$$

where $\mathbf{ds}$ is a small square centered at $\mathbf{s}$. If only one event may occur at a location, note that

$$\lambda(\mathbf{s}) = \lim_{|\mathbf{ds}| \to 0} \frac{E(N(\mathbf{ds}))}{|\mathbf{ds}|}$$

The second-order product density of a point process is

$$\lambda_2(\mathbf{s}, \mathbf{s}') = \lim_{|\mathbf{ds}| \to 0, |\mathbf{ds}'| \to 0} \frac{E(N(\mathbf{ds})N(\mathbf{ds}'))}{|\mathbf{ds}|\,|\mathbf{ds}'|} \tag{24}$$

If $\lambda_2(\mathbf{s}, \mathbf{s}')$ depends only on $\mathbf{h} = \mathbf{s} - \mathbf{s}'$, the process is stationary; if $\lambda_2(\mathbf{s}, \mathbf{s}')$ depends only on the distance h between the points, it is isotropic. In the isotropic case Ripley (1977) developed a theory for the K function,

$$K(h) = E(\text{number of other points within distance } h \text{ of an event})/\lambda \tag{25}$$

and its relationship to $\lambda_2(h)$. A Poisson process is isotropic with $K(h) = \pi h^2$.

The K function plays a critical role in theory, because it can be computed readily for a variety of important point processes, and in practice, because it can be estimated readily from data. Simple data-based estimates take the form

$$\hat{K}(h) = \frac{\sum_{i=1}^{N} \text{number of other points within distance } h \text{ of } \mathbf{s}_i}{\lambda N}$$

(λ is often estimated as $N/|D|$.) Note that this estimate is biased negatively, since events outside the data collection domain D are not included. The intuition behind the K function relates generally to statistical distance methods and specifically, nearest-neighbor methods. Distance methods involve statistics which are computed from the set of all distances between events, rather than the actual locations. Substantial theory regarding the distributions of distance-based statistics is available.

Extensions of point processes include marked point processes for which a (possibly multivariate) variable Z is associated with each event, leading to data of the form $Z \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_N))'$. A related multivariate class of problems involves the modeling and data analysis for multiple point processes.

6. Spatial–Temporal Statistics

Very complex model structures may arise when both spatial dependences and temporal dependences, and their interactions, are present. Early attempts in this area often viewed time as if it were another spatial dimension, leading to relatively simple treatments of spatial–temporal data via spatial methods. However, the special role of time (i.e., the direction of its flow) often renders such approaches ineffective. Alternatively, the Kalman filter adapted to space–time problems has had wide application in the sciences. The basic method is analogous to kriging in temporal settings. Spatial–temporal statistics is at the forefront of current statistical research. See Kyriakidis and Journel (1999) for a review in geostatistics. For recent Bayesian approaches and references, see Wikle et al. (1998) and Wikle (2000).

See also: Bayesian Statistics; Multivariate Analysis: Overview; Spatial Autocorrelation; Spatial Sampling

Bibliography

Berliner L M 2000 Hierarchical Bayesian modeling in the environmental sciences. *Allgemeines Statistisches Archiv* (*Journal of the German Statistical Society*) **84**: 141–53

Bernardo J M, Smith A F M 1994 *Bayesian Theory*. Wiley, New York

Besag J E 1989 Towards Bayesian image analysis. *Journal of Applied Statistics* **16**: 395–407

Besag J E 1991 Spatial statistics in the analysis of agricultural field experiments. In: *Spatial Statistics and Digital Image Analysis*. National Academy Press, Washington, DC

Cliff A, Ord J K 1981 *Spatial Processes: Models and Applications*. Pion, London

Cressie N A C 1993 *Statistics for Spatial Data*, rev. edn. Wiley, New York

Cressie N 1995 Bayesian smoothing of rates in small geographic areas. *Journal of Regional Science* **35**: 659–73

Diggle P J, Tawn J A, Moyeed R A 1998 Model based geostatistics (with discussion). *Applied Statistics* **47**: 299–350

Feigelson E D, Babu G J (eds.) 1992 *Statistical Challenges in Modern Astronomy*, Springer-Verlag, New York

Fisher R A 1935 *The Design of Experiments*. Oliver and Boyd, Edinburgh, UK

Geman S, Geman D 1984 Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern analysis and Machine Intelligence* ***PAMI-6***: 720–41

Grenander U 1993 *General Pattern Theory: A Mathematical Study of Regular Structures*. Clarendon Press, Oxford, UK

Handcock M S, Stein M L 1993 A Bayesian analysis of kriging. *Technometrics* **35**: 403–10

Hodder I, Orton C 1976 *Spatial Analysis in Archeaology*. Cambridge University Press, London

Kaiser M S, Cressie N 1997 Modeling Poisson variables with positive spatial dependence. *Statistics and Probability Letters* **35**: 423–32

Kyriakidis P C, Journel A G 1999 Geostatistical space–time models: A review. *Mathematical Geology* **31**: 651–84

Le N D, Zidek J V 1992 Interpolation with uncertain spatial covariance: A Bayesian alternative to kriging. *Journal of Multivariate Analysis* **43**: 351–74

Ripley B D 1977 Modelling spatial patterns. *Journal of the Royal Statistical Society*, *Series B* **39**: 172–92

Ripley B D 1988 *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge, UK

Royle J A, Berliner L M 1999 A hierarchical approach to multivariate spatial modeling and prediction. *Journal of Agricultural, Biological, and Environmental Statistics* **4**: 1–28

Royle J A, Berliner L M, Wikle C K, Milliff R 1998 A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador Sea. In: Gatsonis et al. (eds.) *Case Studies in Bayesian Statistics*. Springer-Verlag, New York, pp. 367–82

Sampson P D, Guttorp P 1992 Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association* **87**: 108–19

Stein M L 1999 *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York

Stephan F 1934 Sampling errors and interpretations of social data ordered in time and space. In: Ross F A (ed.) *Proceedings of the American Statistical Journal, New Series No*. **185A**. *Journal of the American Statistical Association*, **29** Suppl: 165–6

Stoyan D 1988 Thinnings of point processes and their use in statistical analysis of a settlement pattern with deserted villages. *Statistics* **19**: 45–56

Stoyan D, Penttinen A 2000 Recent applications of point process methods in forestry statistics. *Statistical Science* **15**: 61–78

Wackernagel H 1995 *Multivariate Geostatistics*. Springer-Verlag, New York

Wahba G 1990 *Spline Models for Observational Data*. SIAM, Philadelphia, PA

Wikle C K 2000 Hierarchical space–time dynamic models. In: Berliner L M et al. (eds.) *Studies in the Atmospheric Sciences*. Lecture Notes in Statistics, 144, Springer-Verlag, New York, 45–64

Wikle C K, Berliner L M, Cressie N 1998 Hierarchical Bayesian space–time models. *Journal of Environmental and Ecological Statistics* **5**: 117–54

Wolpert R L, Ickstadt K 1998 Poisson gamma random fields for spatial statistics. *Biometrika* **85**: 251–67

L. M. Berliner

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Spatial-temporal Modeling

'Spatial Temporal Models' (STMs) portray the variation of phenomena over time and across space. In the sections that follow, first the STMs are defined. Then the 'canonical' and 'compartment' types of STMs are described. The construction of STMs is discussed next. A number of trends that have influenced the evolution of spatial temporal modeling are touched upon in a 'concluding comments' section.

1. Introduction

A mathematical model qualifies as a STM if at least some of its variables are referenced spatially and temporally. Spatial temporal referencing means that the values of these variables are linked to points or intervals in time and simultaneously to entities in geographical space such as points, or areal units, or arcs or nodes of a network. The STMs are likely to include also purely spatial and purely temporal variables. Examples of spatial variables are distances, directions, spatial coordinates, locational quotients, measures of spatial inequality, measures of phenomena that vary over space or index geographic environments, and spatial dummies. Examples of purely temporal variables are time and functions of time, dummies associated with time intervals and points in time, and measures of phenomena that vary over time or index temporal environments.

The STMs primarily are concerned with three classes of phenomena: spatial flows, spatial diffusion, and spatial temporal change. The latter includes spatial variation of growth rates (for instance, of product per capita), qualitative spatial change (for instance, in land use), and change in spatial distributions (for instance, of crime events or of settlements). A systematic introduction to the spatial temporal modeling of several classes of substantive phenomena can be found in Haggett et al. (1977).

In the social sciences, much of the conceptual work that precedes or is intertwined with mathematical modeling is based on notions from system theory. Books dealing with the nexus between spatial temporal modeling and spatial systems include Bennett and Chorley (1978), Bertuglia et al. (1987), Huggett (1980), and Klassen et al. (1979). The STMs figure prominently in the edited collections by Anselin and Florax (1995), Fischer and Getis (1997), Griffith and Mackinnon (1981), Griffith and Lea (1983), Griffith and Haining (1986), and Wrigley (1979).

The STMs encompass a huge body of knowledge crisscrossed by many substantive and analytical literatures. In order to convey a feel for spatial temporal modeling the two most common types of STMs were singled out for a brief treatment in the next two sections. A subsequent section discusses an approach to the construction of STMs. In the concluding comments some trends that have influenced the evolution of the STMs are touched upon.

2. Canonical STMs

'Canonical single equation STMs' are models consisting of an equation with spatially referenced rates of change over time as the dependent variable, and in which at least some of the independent variables are spatial or spatially referenced. The 'canonical multiple equations STMs' instead are systems of equations, some or all of which qualify as canonical single equation STMs. In the social sciences, the most common STMs are the ones here referred to as canonical.

Examples of canonical single equation STMs can be found everywhere in the social sciences. For instance, they are the backbone of the empirical literature on the determinants of economic growth. Consider as a specific 'case' the regression model in Barro (1997). It has growth rates of GDP per capita as the dependent variable, and GDP per capita, male schooling, fertility rates, a democracy index, inflation rates, and others as independent variables. Barro estimated it using a sample of 114 countries, and growth rates calculated from several time intervals. In fact, the STMs in which the dependent variable is the rate of change of variables such as income, unemployment, population, welfare rolls, crime rates, for countries, counties, or regions, and in which the independent variables include indices of relevant characteristics of these geographical entities, are ubiquitous. Also very common are models in which rates of change are functions of distances or of spatial coordinates.

An example of canonical multi-equation STMs is Greenwood's 1981 model of metropolitan growth and migrations. The model consists of 14 equations with 14 endogenous and 16 exogenous variables. It focuses on the rates of change of income per capita, unemployment, and migrations into or out-of metropolitan and nonmetropolitan areas. A number of multi-equation canonical STMs are described in Putman (1979).

3. Spatial Compartments STMs

'Compartment models' describe systems made of subsystems interconnected by flows (De Palma and Lefevre 1987). In a number of instances the dynamics of a system and of its subsystems and the flows among these are influenced by controls or from policy instruments (Bennett and Chorley 1978). The term 'spatial compartment models' refers to spatial systems with subsystems corresponding to geographical entities such as regions, countries, or urban centers. For

instance, a world system may consist of spatial compartments made of aggregates of countries ('regions'). A system may be defined in terms of one or more 'elements.' For instance, population may be the only element in a system, and a four-elements system may be defined in terms of population, product, capital, and technology.

An example of spatial compartment STMs is the one employed in a 1977 UN study of the future of the world economy (Leontief 1977). The model is of the inter-regional input–output type. It has 48 producing and consuming sectors and 15 regions. The regions are aggregates of countries. A sector inputs from other sectors, and outputs to other sectors. The goods produced by sectors are classified as 'domestic' or 'internationally traded.' The internationally traded goods are exported to or drawn from international sectoral pools, and each region exports/imports a fixed share of the aggregate world exports/imports of any given internationally traded commodity.

The model is made of a total of 2625 equations grouped into 15 regional sets. Each regional set describes the inter-relations within a region of production, consumption, imports, and exports, of goods and natural resources. Each set involves 269 variables, of which 229 are region specific, and the remaining 40 represent the imports/exports pools of internationally traded goods and the balance/imbalance of the regions' international financial transactions. Thus, the model contains more variables than equations. To solve it, enough variables are set to fixed values to render the number of the remaining variables equal to the number of the equations.

This model was used to produce eight development paths of the world economy from the base year 1970 through 1980 and 1990 to the year 2000. These paths are based primarily on combinations of low/medium/high population growth scenarios, and low/high GDP per capita growth scenarios. In short, the questions asked focus upon what the world economy has to be like in 1980, 1990, 2000, if the population and product per capita follow specified scenarios that unfold within realistic technical and physical constraints.

Spatial compartment STMs can be constructed by combining input output models with econometric models. A review of pertinent issues and literatures is contained in Rey (1998). In the 'computable general equilibrium models' the demand and supply functions of products and of factors of production are specified and estimated. In the spatial compartment variants of these models the mechanisms affecting the spatial flows of products and factors are also brought forth and the equations describing these mechanisms are specified and estimated. Regional computable general equilibrium modeling has been surveyed by Partridge and Rickman (1998). Isard et al. (1998) review applied inter-regional equilibrium modeling. A discussion of the theory and estimation of multicountry trade models is contained in the book by Deardorff and Stern (1986). This 'Michigan' model represents a massive undertaking, and was developed for the purpose of evaluating the impacts of changes in tariffs on international trade flows. Linked macroeconomic models are discussed in Ball (1973) and in Fair (1984).

A spatial compartments follow up to the 'Limits to Growth' model (Meadows et al. 1972) is described in Mesarovic and Pestel (1974a) and is documented in the six volumes edited by the same authors (1974b). Their work builds upon the 'Limits to Growth' thesis that if the current population growth, economic growth, pollution increases, and natural resources depletion continue, the world will face a number of catastrophes. However, Mesarovic and Pestel maintained that before the middle of the twenty-first century we shall be confronted by regional catastrophes rather than by a collapse of the world system. These catastrophes will materialize in different regions for different reasons and at different times, and will be profoundly felt throughout the entire world (Mesarovic and Pestel 1974a).

Most of the spatial compartment models referred to in this section are constructed to produce forecasts, simulations, or scenarios. It should be noted, however, that there are extensive literatures dealing with spatial compartment STMs constructed for the purpose of investigating systemic optima. Intertemporal linear and non linear programming models based on spatial compartments are discussed in Takayama and Judge (1971) and in Judge and Takayama (1973). Optimal control spatial compartments STMs are reviewed in Tan and Bennett (1984).

4. *STMs Generated by Expansions*

In many cases, STMs in the social sciences have been imported from nonsocial science disciplines. For instance, some of the diffusion models in geography and economics originated from their counterparts in physics and engineering (Banks 1994). In general, the analogies between very different phenomena may trigger interdisciplinary transfers of methodologies and models. The notions of systems and of system dynamics and the schools of thought linked to them have played a major role in the construction of classes of STMs based on cross-disciplinary analogies.

However, in many circumstances the need may arise to construct STMs from purely spatial or purely temporal models, or to extend existing STMs to encompass relevant contexts that are outside their original scope. An approach to the incremental construction of mathematical models that can be easily applied to building STMs is the 'expansion method' (Casetti 1997). In a nutshell, the expansion method involves widening the scope of a simpler 'initial model' by redefining some or all of its parameters into

functions of expansion variables or variates that index substantively relevant contexts of the model. The 'terminal model' thus produced encompasses both the initial model and a specification of its contextual variation.

To demonstrate the operation of the method, let us discuss its application by adding a time dimension to a purely spatial model of urban population densities. This demonstration is based on a variant of the model presented in Casetti (1973). The urban studies literature has identified two spatial patterns of urban population densities that are often referred to as 'density cones' and 'density craters.' In the cones, the population densities decline exponentially with distance from city centers. Instead, a density crater type of trend is characterized by urban population densities that increase at first and then decline as the distance from the city center increases.

Let D, s, and ε denote, respectively, the natural logarithm of urban population density, distance from a city center, and an error term. Let

$$D = \alpha_0 + \alpha_1 s + \alpha_2 s^2 + \varepsilon \tag{1}$$

be an initial model subsuming both a density cone and a density crater as special cases. In fact, for $\alpha_1 < 0$ and $\alpha_2 \leqslant 0$ this model corresponds to a density cone, while for $\alpha_1 > 0$ and $\alpha_2 < 0$ it yields a density crater.

The literature has also identified distinctive patterns of temporal change of these spatial trends. They include the rising and flattening of the density cones, the widening of the density craters, and the gradual transformation of cones into craters. In order to extend the initial purely spatial model (1) so that it will encompass the spatial temporal trends recognized by the literature, the α parameters in it are expanded by the stochastic expansion equations

$$\alpha_i = \alpha_{i0} + \alpha_{i1} t + \theta_i \quad i = 0,1,2 \tag{2}$$

where the θs are error terms. The terminal models

$$D = \alpha_{00} + \alpha_{01} t + \alpha_{10} s + \alpha_{11} st + \alpha_{20} s^2 + \alpha_{21} s^2 t + u \tag{3}$$

$$u = \varepsilon + \theta_0 + \theta_1 s + \theta_2 s^2 \tag{4}$$

are obtained by replacing the right hand sides of the expansion equations for the αs in the initial model. Upon estimation, these terminal models can be used to test for the occurrence of specific spatial temporal trends, and to portray these trends if they occur.

The model appearing in this demonstration is intended for estimation and for the testing of statistical hypotheses. It should be noted, though, that the expansion method can be used to construct STMs of any kind. Any model that has parameters can be expanded, irrespectively of whether it is deterministic, stochastic, or mixed, and irrespectively of whether it is intended for solving, estimating, or optimizing.

5. *Concluding Comments*

Most STMs in the social sciences came into existence during or after the 1930s, and have been influenced or shaped by trends that kept altering the landscape of the quantitative analytical social sciences during this time horizon. The comments that follow highlight some such trends and influences.

Mathematical models in general, and STMs in particular, have been acquiring an important role in government and private sector activities as aids to decision making, to planning, to policy development and evaluation, and for forecasting. In the process they have been driven to address the complexities of the real world to an ever-increasing degree. The tendency toward more complex models is especially apparent in the canonical and compartment STMs, and in turn the proliferation of these models is a reflection of their practical usefulness. The movement toward the 'computational' counterparts of STMs that originated as purely mathematical formalisms is another aspect of these tendencies.

Increasingly complex STMs have also played an important role in the precise articulation of the theories and conceptual frameworks that constitute a substantial component of the social sciences. In the more theoretical applications this complexification has been intertwined with conceptual and technical advances imported from mathematics and the non social sciences. The STMs with nonlinear dynamics, and the ones centering on optimal control techniques are cases in point.

The entire evolution of STMs and of mathematical modeling at large has been permeated by the growth of stochastic modeling, and by the penetration of stochastic elements into models born deterministic. For instance, deterministic canonical STMs have been converted into 'mixed' or stochastic models by the addition of stochastic errors and/or of temporally or spatially autoregressive terms. An adequate stochastic specification and its attendant statistical econometric testing have increasingly become a must for empirical analyses to be credible.

See also: International Trade: Geographic Aspects; Location Theory; Spatial Decision Support Systems; Spatial Interaction Models; Spatial Optimization Models; Spatial Pattern, Analysis of; Spatial Search Models; Transportation Geography

Bibliography

Anselin L, Florax R J G M (eds.) 1995 *New Directions in Spatial Econometrics*. Springer, Berlin

Ball R J (ed.) 1973 *The International Linkage of National Economic Models*. North Holland, Amsterdam
Banks R B 1994 *Growth and Diffusion Phenomena: Mathematical Frameworks and Applications*. Springer, Berlin
Barro R J 1997 *Determinants of Economic Growth: A Cross-country Empirical Study*. MIT Press, Cambridge, MA
Bennett R J, Chorley R J 1978 *Environmental Systems: Philosophy, Analysis, and Control*. Princeton University Press, Princeton, NJ
Bertuglia C S et al. (ed.) 1987 *Urban Systems: Contemporary Approaches to Modelling*. Croom Helm, London
Casetti E 1973 Testing for spatial-temporal trends: An application to urban population density trends using the expansion method. *Canadian Geographer* **17**(2): 127–37
Casetti E 1997 The expansion method, mathematical modeling, and spatial econometrics. *International Regional Science Review* **20**(1/2): 9–33
Deardorff A V, Stern R M 1986 *The Michigan Model of World Production and Trade*. MIT Press, Cambridge, MA
De Palma A, Lefevre C 1987 The theory of deterministic and stochastic compartmental models and its applications. In: Bertuglia C S, Leonardi G, Occelli S, Rabino G A, Tadei R, Wilson A G (eds.) *Urban Systems: Contemporary Approaches to Modelling*. Croom Helm, London
Fair R C 1984 *Specification, Estimation, and Analysis of Macroeconometric Models*. Harvard University Press, Cambridge, MA
Fischer M M, Getis A (eds.) 1997 *Recent Developments in Spatial Analysis*. Springer, Berlin
Greenwood M J 1981 *Migration and Economic Growth in the United States: National, Regional, and Metropolitan Perspectives*. Academic Press, New York
Griffith D A, Haining R P (eds.) 1986 *Transformations Through Space and Time: An Analysis of Nonlinear Structures, Bifurcation Points, and Autoregressive Dependencies*. Mt Nijhoff, Dordrecht, The Netherlands
Griffith D A, Lea A C (eds.) 1983 *Evolving Geographical Structures: Mathematical Models and Theories for Space-time Processes*. M Nijhoff, The Hague, The Netherlands
Griffith D A, Mackinnon R (eds.) 1981 *Dynamic Spatial Models*. Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands
Haggett P, Cliff A D, Frey A 1977 *Locational Analysis in Human Geography*, 2nd edn. Wiley, New York
Hotelling H 1978 A mathematical theory of migration. *Environment and Planning*: *A* **10**: 1223–39
Huggett R J 1980 *Systems Analysis in Geography*. Clarendon Press, Oxford, UK
Isard W, Azis I J, Drennan M P, Miller R E, Saltzman S, Thorbecke E 1998 *Methods of Interregional and Regional Analysis*. Ashgate, Aldershot, UK
Judge G G, Takayama T (eds.) 1973 *Studies in Economic Planning Over Space and Time*. North Holland/American Elsevier, Amsterdam
Klassen L H, Paelinck J H P, Wagenaar S 1979 *Spatial Systems*. Saxon House, Farnborough, UK
Leontief W 1977 *The Future of the World Economy: A United Nations Study*. Oxford University Press, New York
Meadows D H, Meadows D L, Randers J, Behrens III W W 1972 *The Limits to Growth*. Universe Books, New York
Mesarovic M, Pestel E 1974a *Mankind at the Turning Point: The Second Report to The Club of Rome*, 1st edn. Dutton, New York
Mesarovic M, Pestel E (eds.) 1974b *Multilevel Computer Model of World Development System*. International Institute for Applied Systems Analysis, Laxenburg, Austria, vols. 1–6
Partridge M D, Rickman D S 1998 Regional computable general equilibrium modeling: a survey and critical appraisal. *International Regional Science Review* **21**(3): 205–48
Putman S H 1979 *Urban Residential Location Models*. Nijhoff, Boston
Rey S J 1998 The performance of alternative integration strategies for combining regional econometric and input-output models. *International Regional Science Review* **21**(1): 1–35
Takayama T, Judge G G 1971 *Spatial and Temporal Price and Allocation Models*. North Holland, Amsterdam
Tan K C, Bennett R J 1984 *Optimal Control of Spatial Systems*. Allen & Unwin, London
Wrigley N (ed.) 1979 *Statistical Applications in the Spatial Sciences*. Pion, London

E. Casetti

Spatial Thinking in the Social Sciences, History of

The objects upon which social sciences work, either people or artefacts, take up some room and are located in specific places. Cognition implies the building and mobilization of spatial skills, like orientation. Social scientists are generally more interested in particular spatial categories like direction, distance, place, area, landscape, village, town, city than in space: the distribution of things and people, the behaviors and practices of everyday life, or the plans developed for the future deal with concrete geographical realities rather than the abstract idea of space. Even geography was traditionally more oriented towards places, regions, or zones than towards space: the idea of geography as a spatial discipline did not develop before the 1950s, and was soon criticized. Among the other social sciences, economics was the only one with a specialty branch in the field: spatial economics, which began to flourish in the 1820s. Among the cognitive sciences, the psychology of orientation developed before Jean Piaget described the stages of space construction by children. As a result, spatial thinking in the social sciences often appears split into different fields, *urban* studies, *rural* studies, *landscape* studies, the significance of the *regional approach* and of *place*, the idea of *environment*, the problem of *orientation*, etc.

Spatial thinking refers to five types of intellectual operations: (a) using geographical observations as a starting point for social analysis, (b) explaining, generally within a functional perspective, the role of different aspects of space in the life of social groups, (c) understanding the way people invest symbolic meanings in their environment, (d) wondering about the significance given to space in the definition of concepts like society, economics, culture, language, polity, etc.,

or (e) deciphering the ways in which the human mind deals with space and builds and uses spatial categories. The first family was present from the earliest developments of the social sciences. The second family flourished mainly in the late nineteenth and most of the twentieth centuries. The third family is more characteristic of the rise of phenomenological approaches, since the 1930s or 1940s. The fourth family is mainly linked to the development of critical perspectives on the social sciences since the 1970s. The development of an interest for space by the cognitive sciences remained until recently independent from the other approaches.

1. Spatial Distributions as a Point of Departure for Social Thinking

We are living in social systems and so deeply embedded into them that generally we think they are an expression of the natural order. In order to build a science of social realities, the social scientist has to introduce a distance between the object. Travel narratives and geographical descriptions played a central role in this operation. Montaigne started to ponder about the relativity of social institutions when reading about the American cannibals. From Jean Bodin to Montesquieu, the diversity of political and social institutions nurtured the reflexion on social and political life and was interpreted through a simple model, the influence of climate.

The first steps of economics were linked with geographic evidence: the role of transport in the distribution of economic activities, the complementarities between cities and rural areas, the economic circuit served as bases for the development of a more conceptual vision, which started with the analysis of market mechanisms by Turgot and Adam Smith's reflexion on the division of labor.

With the development of travels and the improvement of their accounts, geographical comparisons became more sophisticated: they allowed for the development of new disciplines, like ethnography. The change of scenery could be provided by travel through time as well as space. Immanuel Kant emphasized the similarities between history and geography: spatial and temporal distributions were the first features to confront social scientists in their study of the World. He opposed spatio-temporal descriptions to systematic analyses. Partly under the influence of Kant, the idea that geography was essentially a description of spatial distributions developed and remained alive well into the twentieth century.

It was during the first decades of the nineteenth century that spatial comparisons played their most important role in the development of science. It was true for the natural sciences—Charles Darwin developed his evolutionary theory when traveling on the *Beagle*. In the social sciences, Alexis de Tocqueville built his analysis of democracy by comparing the US, where it permeated all forms of social life, and Old Europe, where aristocratics were still influencial. Karl Marx explored the opposition between medieval cities and rural areas, the Asiatic social and economic organization, and the industrial revolution in Britain before building the notion of mode of production.

By the middle of the nineteenth century, the cartographic model was increasingly replaced by models each with a logic dictated by the dynamic, immediate, interactive phenomena under study: the new specialized disciplines broke their earlier links with spatial thinking.

2. The Functional Approach to the Relations Between Space, Social Life, and Thought

When social and behavioral sciences severed their former links with distributional approaches, they generally considered space as a neutral stand and ceased to use it as a significant factor for explaining what they observed. Many scholars thought, however, that space mattered: they had the idea that social life differed according to its physical setting, the fact that people lived in empty or overpopulated areas, and were more or less distant from each other. For Emile Durkheim the foundation of sociology lay in social morphology, but he was unable to provide a clear picture of it. The only way to take it into account was to adopt a functional perspective: space could be conceived either as a collection of environments or a set of distances (see *Functionalism, History of*).

2.1 Space as Environment

In the environmental perspective, the emphasis was given to the local connexions between people, artefacts, and place. The first environmental approaches were developed by the sensualist psychology of the eighteenth century. Man was a product of whatever his senses impressed upon his mind. His aptitudes and personality reflected his environment. If it was disorderly or unsound, social life had many reasons to be stressed. By the end of the eighteenth century, social scientists were well aware of the social and political implications of such a conception. Jeremy Bentham conceived his *Panopticon* in order to provide criminals with an environment which would teach them safer forms of behavior. More generally, the idea that to reform society, you had to change its material setting was central to the nineteenth century utopian socialism and still present in the conception of urbanism developed by the International Movement of Modern Architecture in the 1930s.

Johann Gottfried Herder was more sensitive to the harmony which developed between peoples and the country they lived in than to the sensual influence of

these surroundings on their mind and behavior. The early development of geography in Germany was linked to the combined influence of Kant (evident in Alexander von Humboldt's conception of geographical description) and Herder (who was closer to Carl Ritter). Such a conception was important for all the social sciences in Germany, since it was conducive to a keener interest in peoples than in individuals or social classes. For many German geographers, the harmony between a people and the land they lived in was expressed through its landscapes (see *Romanticism: Impact on Social Thought*).

The Herderian environmentalism was a literary and poetic one. Natural evolutionism had a more direct impact on the social sciences, since it was a scientific theory. The early forms of evolutionism as developed in the eighteenth and early nineteenth centuries had no direct impact on the then emerging social sciences. The publication of *The Origin of Species* occurred at a time when history and the idea of progress had become popular in the public opinion, and original forms of evolutionism had developed in sociology: hence the impact of its theses in the social sciences. For sensualist psychology or Herderian philosophy, the influence of things upon the human mind was a direct one. For scientific evolutionism, the relations were mediated through biological mechanisms (see *Evolutionism, Including Social Darwinism*).

Did environments shape completely the destiny of humankind and of any of its groups? Hot discussions developed between environmentalists and the scientists who believed in human liberty. In order to depict the relations between organisms and their environment, a new discipline, ecology, appeared. Named by Haeckel in 1872, its first scientific achievements occurred at the end of the nineteenth century; it became mature in the 1930s and 1940s. Geographers did not wait for the development of ecology to propose a new conception of their discipline: *Anthropogeographie*, or human geography, was named and shaped by Friedrich Ratzel in the 1880s and 1890s. For him and for the first human geographers in France, a strict environmentalism was an unbearable position, but physical constraints did exist and shaped at least partly human behavior and history: geographers pleaded for possibilism, which gave room to physical and biological constraints, but acknowledged the role of other variables.

By 1920, it had become evident that environmentalism was not a very fruitful way of conceiving the role of space in social functioning.

2.2 Space as Distance

Economics had developed another approach to analyze the role of space in the functioning of society. In the early nineteenth century, some economists were unhappy with the elimination of space from mainstream theory. Hence, the development of spatial economics as a sideline. It was centered on two problems: the uneven distribution of natural factors and the role of distance. The name of David Ricardo is associated with the first orientation and the ensuing theory of international specialization and trade. Johann Heinrich von Thünen explained how distances to the market determined farming specialization even in a perfectly uniform setting (1826–52). The progress of spatial economics was a slow one: people had to wait until Alfred Weber for a coherent theory of industrial location (1909), and Walter Christaller and August Lösch for an understanding of the location of service activities and the nature of urban networks (early 1930s).

At the end of the nineteenth and the beginning of the twentieth centuries, some anthropologists also considered distance as a major factor of cultural differentiation. It soon appeared that diffusionism could not serve as an overall explanatory theory. Diffusion was, however, a significant cultural process (see *Anthropology, History of*).

When geographers, tired from the hopelessly descriptive character of the possibilist approach, tried in the late 1950s and 1960s to develop a more systematic view of their scientific enterprise, many of them decided that geography should become a discipline of distances: they incorporated what spatial economics and anthropology could offer and went further through the emphasis they gave to communication processes. They used what a new scientific field, the study of communication, was developing.

The functional approach to social problems through the analysis of distance was fruitful. Its role in planning theory and practice was essential during the 1960s and 1970s. It allowed for the design of well organized cities or suburbs. It did not prevent, however, the development of urban pathologies. In order to understand them, other approaches to the role of space in social life had to be developed.

3. Space as Experience

Functional approaches were specific to the social sciences. They did not offer many possibilities to the cognitive sciences. The situation is different for those who try to analyze space as experience.

3.1 Environmental Psychology

The curiosity of psychologists for spatial problems was initially focused on the problems of orientation and place recognition. Their interest changed with the genetic approach developed by Jean Piaget: his problem was to understand the way in which children explored their environment and built behavioral

schemes and mental categories out of their experience and inborn capabilities.

The relations between man and environment are mediated through the senses: what are the specific contributions of sight, hearing, smell, touch, taste, and cynesthesic sensation to the individual experience of space? Michel Foucault, for instance, was fascinated by the place given to sight by Western civilizations and its role in social control.

Because of its emphasis on experimental methods, mainstream psychology ignored the social dimension of spatial learning and experience. They were brought back by phenomenologists, Gaston Bachelard or Maurice Merleau-Ponty, for instance. Environmental psychology introduced the theme of proxemics.

3.2 The Psychological, Moral, and Religious Significance of Space

Through phenomenology and proxemics, space started to be conceived as a necessary component of communication systems: hence the contribution of linguistics and semiotics to spatial thinking. Landscapes ceased to be analyzed in a functional perspective. They were conceived as supports for messages. People experience different forms of connivance with them.

A difference exists between analytical and symbolic communication: in the first case, the problem is to transfer complex sets of information from individuals to individuals; in the second one, it is to stir simultaneously many persons through a unique and short message. Symbolism appears in this way as a way of mastering distance. Hence, the interest of geographers for this form of communication.

Space has other social meanings. Normative thinking relies on perspectives from the outside: people have to imagine Beyonds, either within beings or things (when the way of immanence is chosen) or outside them, generally in another World (when the way of transcendance is selected). These choices are important since they introduce a qualitative differenciation of space: places where the Beyonds touch the surface of our World are loaded with sacrality, in contrast to the surrounding profane areas. The building of Utopias, the belief in the Golden Age, or in a Land without Evil are ways to use history or terrestrial space to build other types of Beyonds. The dreams they nurture are important in the shaping of rural or urban areas.

The new approaches to geography developed since the 1970s incorporate this interest in the human experience of space.

3.3 Social Experience as a Localized Experience

People cannot develop an experience of society outside specific places. The *Time Geography* developed by Torstein Hägerstrand stresses the fundamental diversity of spatial and social experiences developed by people along their life trajectories: it means that social experiences differ according to persons. In the field of communication, for instance, people participate to different intersubjectivity circles depending on the places they live or had lived.

Such an observation is fundamental for all the social sciences: it reminds that societies, cultures, polities, languages do not exist in a spaceless social realm. They differ from place to place, because of the existence of *locales*, as expressed by Anthony Giddens. The social sciences of today can no more ignore spatial thinking: it is one of the main lessons of the postmodernist epistemological revolution.

Bibliography

Bachelard G 1957 *La Poétique de l'Espace*. Presses Universitaires de France, Paris

Dagognet F 1977 *Une Epistémologie de l'Espace Concret*. Vrin, Paris

Eliade M 1957 *Das Heilige und das Profane*. Rohwolt, Hamburg

Giddens A 1984 *The Constitution of Society*. Polity Press, Cambridge, UK

Gould P, White R 1974 *Mental Maps*. Penguin, Harmondsworth, UK

Lee T 1976 *Psychology and the Environment*. Methuen, London

Livingstone D N 1993 *The Geographical Tradition*. Blackwell, Oxford, UK

Merleau-Ponty M 1945 *Phénoménologie de la Perception*. Gallimard, Paris

Piaget J 1926 *La Représentation du Monde Chez l'Enfant*. Alcan, Paris

Ponsard C 1984 *History of Spatial Economic Theory*. Springer-Verlag, Berlin

Proshansky H M, Ittelson W H, Rivlin L G (eds.) 1970 *Environmental Psychology*. Holt, Rinehart and Wilson, New York

P. Claval

Spatially Explicit Analysis in Environmental Studies

Problems of the human–environment condition, including its change, are increasingly examined in terms of their spatial configuration and scale of analysis. It could be argued that much, if not all, of the field of geography studies these problems in this way, and there are numerous articles in this Encyclopedia dedicated to geographical problem solving and approaches. Spatial analysis of the human–environment condition (i.e., environmental and resources problems) has expanded beyond geography, however, becoming a multidisciplinary if not interdisciplinary

subfield of study. This multidisciplinary focus on spatially explicit problem solving, where the spatial aspect of the human behavior and/or the natural environment makes a crucial difference in the analysis and policy response to the problem, is treated here. Throughout, the emphasis is on geographical space (metric and relative distance), not on such social science concepts as 'social distance' (i.e., between social or economic classes), which are treated elsewhere.

1. Space in the Sciences

The natural sciences are further advanced than the social sciences in their theoretical developments of models of spatial processes. Landscape ecology and population biology are exemplary. In both of these fields, the spatial arrangement of the geophysical landscape, as well as flora and fauna on the landscape are the major focus of analysis. The advances in spatial analysis in the natural sciences, compared with the social sciences, are fairly straightforward to explain. There is a more developed physical theory of behavior over space, such as gravity and watershed impacts, or species interaction over space, than there are theories of spatial human behavior and individual preferences of spatial configuration. Also, until recently, there has been much more spatially explicit natural science data than social science data available for modeling and hypotheses testing, resulting in delayed response in the social sciences to spatially explicit problem solving (NRC 1997).

In geographical spatial models employed in the social sciences, space appears as a constraint on the system under consideration. For example, geographical space is reduced to distance in a transportation cost model of location. However, space cannot be considered only as a constraint or cost to be paid; it must also be considered in terms of spatial processes and must be modeled as such at the appropriate scale of analysis. In order to explain and predict these spatial processes, theoretical and empirical models must be developed to address where, when, and why these processes happen (Bockstael 1996).

The other social and behavioral sciences have lagged behind geography in developing theories and models of human behavior and choices over space. One of the main reasons for this lag is likely the paucity of disaggregate spatially explicit social science data. These data are necessary to test empirically the hypotheses derived from the theoretical constructs of individual human behavior. In addition to the increase in spatial data, the advances in geographical information systems (GIS) technology in availability and ease of use has increased the interest in spatial issues among the social sciences beyond geography. Also, as spatial data have become more available, so too have the statistical tools to use these data.

2. Topical Areas

Associated with the use of many renewable and nonrenewable natural resources are spatial effects, where the spatial location and configuration of the natural resource affects human behavior and consequences to the environment. Examples include stationary resources, such as energy resources, minerals, forestry; and nonstationary resources, such as fisheries and other wildlife, as well as the use of resources that can lead to spatial environmental problems. The specific examples that are focused on in this article include fisheries and other wildlife, forestry and urban land use, with an emphasis on the implications of where spatially explicit problem solving makes a critical difference in the analysis and policy response to the interaction between human systems and natural systems.

2.1 Fisheries and Wildlife Management

Historically, the primary focus of fisheries management models has been the common property feature of most fisheries, that is, not one person, industry or even country has control over the stock and location, and often, owing to mismanagement, this leads to open access problems resulting in overfishing and depletion of stocks. The policy responses to address these problems have included reduction in fishing season, and the development of tradable permits. Sanchirico and Wilen (1999) argue that these traditional models assume a spatially homogenous distribution of the resource, but if that is not case, then the policy implications concerning the associated human behavior will likely be incorrect. They have developed a model that includes the spatial configuration of the resource, focusing on patchy and heterogeneous distribution of the fish populations, especially the interconnections among and between patches. The main result is that by incorporating different forms of these dynamic spatial relationships, such important biological concepts as 'source' or 'sink' of the living resource can depend on both biological and socioeconomic parameters. Therefore, spatially differentiated management techniques, such as refuges or rotating harvesting zones that do not consider the spatial interaction of human behavior and the resource, could lead to suboptimal policies.

Similar to the spatial fishery models are spatial terrestrial wildlife models that analyze different wildlife management schemes. An example of the effects of the spatial interaction between beaver populations and forestry management is found in Bhat et al. (1996). As beaver populations move over space, management of the beavers on one property can have unintended spillovers on contiguous properties. By taking these interactions into account, the optimal beaver control strategies is derived for each location

and time period, leading to a more efficient management scheme than would be possible without taking into account the spatial dynamics of the system.

2.2 Forestry: Timber Management and Deforestation

The original emphasis of timber management forestry models was to maximize the value of the timber products deriving from a stand, and these stands were assumed to be spatially homogenous. These models assumed that the only benefit from the forest was the value of the timber, and that there were no off-site impacts from the growing forest or from timber harvesting. Swallow and colleagues (1997) argue that while these models are important, the lack of stand interactions in them can lead to faulty management and policy prescriptions, as the spatial relationships that lead to many ecosystem functions are ignored, such as the relationship between forest fragmentation and wildlife populations. By including these spatially explicit ecological interdependencies, Swallow and colleagues show that the optimal forestry management scheme can differ substantially from using a nonspatial modeling approach, including different harvesting periods for the different stands, and different amounts of the total benefits of recreation and forage availability to wildlife. Similarly, a spatially explicit cellular landscape ecology model that incorporates spatial ecological interdependencies derives the optimal spatial pattern of harvesting over time, a result that could not be obtained from a nonspatial approach (Hof et al. 1994). These results have important policy implications. For example, this approach can be used by public land managers who are managing the forest for multiple use and have to take into consideration nontimber outputs, such as wildlife, that are located on adjacent land that is not under their jurisdiction.

Tropical forest management issues have an important global policy context, owing to the linkages between tropical deforestation and the loss of biodiversity, and the decrease in carbon sequestration potential and those effects on global warming. The major uses of tropical forests include timber activities as in the temperate forest context, but to a greater degree than exists in temperate forests, they also include gathering of fuelwood, agroforestry and land clearing for cultivation. In a tropical forest management context, Albers (1996) develops a theoretical spatial model that incorporates the different potential land uses and their spatial interaction across plots. Tropical forest fragmentation and its spatial pattern have important effects on biodiversity and the ability of the forest to regenerate. The simulation model optimizes the pattern of land uses under different spatial and management scenarios, and investigates how these characteristics affect forest fragmentation and the associated ecological effects of the different scenarios. Albers concludes the interaction between land-use patterns and forest recovery generates flexibility in spatial planning for forest recovery that can generate income for land managers in the short run, while maintaining the option for large preserved areas in the long run.

For reasons not entirely clear, there has been much work on spatially explicit empirical modeling of tropical deforestation than in the temperate forestry (e.g., Mertens and Lambin 1997, Veldkamp and Fresco 1997). Included in these models are variables that can be derived from the satellite imagery and used with geographical information systems to create spatial variables, such as distance to roads and urban centers, as well as existing spatial geophysical data. As these models do not include any model of human behavior, the spatial pattern is not an outcome of a behavioral spatial process. The only possible simulations that can be performed are changes that are imposed on the explicit features of the landscape, therefore only simple policy analyses such as a moratorium on road-building or certain land uses can be performed. In contrast, models that develop an underlying structural model seeking to explain the human behavior that generates land-use patterns include Chomitz and Gray (1996) and Pfaff (1999). These models include advances by correctly including aggregate economic variables in a behavioral model. This makes these models policy-relevant, as behavioral responses to changes in policies can be predicted. However, spatially explicit tropical deforestation models have made advances in controlling for the spatial heterogeneity of the landscape, they all have a very simple spatial behavioral process, where distances to roads and markets are the main spatial elements linked with deforestation. More sophisticated spatial processes are needed for models of tropical deforestation where, in addition to distances to roads and markets, distances to neighboring land uses and the patterns of those neighboring land uses are considered in the probability of deforestation (e.g., Geoghegan et al. 2000).

2.3 Urban Land Use

The location, distribution and pattern of urban land uses and urban fringe development have major effects on the environment. It is the pattern of development that determines the amount and type of nonpoint source pollution into water bodies; loss of farmland and other open-space amenities; the amount of commuting with associated air pollution and congestion; as well as ecological effects, including hydrological disturbances and habitat fragmentation. There is a rich social science literature of spatial models of urban land use. The classic monocentric city model is an equilibrium model of urban spatial structure, where the distribution of land uses on a featureless plain

around a central business is a result of an equilibrium between the declining land price gradient and increasing transportation costs. In its simplest form, concentric rings of residential development around the urban center describe the resulting equilibrium pattern of land use and decreasing residential density as distance from the urban center increases. More sophisticated versions of the model have been developed, but nonetheless, the model's ability to explain observed land use patterns is weak. In comparing the model's predictions with actual land-use patterns, the model fails to explain the complexity of the spatial and temporal patterns of urban growth. This limitation is partly due to the treatment of space, which is assumed to be a 'featureless plain' and is reduced to a simple measure of distance from the urban center.

Some research has tried to overcome the 'featureless plain' assumption by including some of the heterogeneous landscape features that exist, for example, by focusing on how individuals value different landscape arrangements. In a study of individual preferences concerning agricultural land abandonment and subsequent spontaneous reforestation, Hunziker and Kienast (1999) use photographs of different stages of reforestation to elicit respondents' most preferred pattern of land uses and find that there are particular arrangements of the landscape that are preferred by individuals. In order to test the hypothesis that individuals value the pattern of land uses surrounding their homes, Geoghegan and colleagues (1997), have created spatial indices of land use fragmentation and diversity measured at different scales, and include these variables in a spatially explicit model of residential land values. Results demonstrate that individuals prefer to have a homogenous distribution of land use in the immediate area surrounding their homes, but a more heterogonous distribution of land uses on a larger scale.

The spatial interdependencies among land use, zoning regulations and on ecosystems are modeled in Bockstael and Bell (1997). A spatially explicit modeling approach is necessary in order to understand how these regulations affect development decisions across the landscape and in turn how the location of nutrients from point sources, such as sewer treatment plants, and nonpoint sources, for example, septic fields affect water quality. They find that differential zoning across regions deflects development from one region to another and that the amount of increased nitrogen loadings from a constant amount of new development varies from 4–12 percent, depending on the degree of difference across zoning regulations.

In the urban land-use models introduced above, the theoretical structure assumes a very simple concept of spatial human behavior—that individuals only respond to exogenously determined locations of landscape features. Other models allow for the potential spatial interactions among individuals. An approach that can be used to model this interaction explicitly is a cellular automata model, where behavior of a system is generated by a set of deterministic or probabilistic rules that determine the discrete state of a cell based on the states of neighboring cells. Despite the simplicity of the transition rules, these models, when simulated over many time periods, often yield complex and highly structured patterns. Researchers have used cellular automata models to analyze the process of urban growth (e.g. Clarke et al. 1997, White et al. 1997). In contrast to the standard urban economic models of urban structure introduced above, in which complex patterns are generated by imposing external conditions, these models demonstrate how complex structure arises internally from the interaction among individual cells. While these models fit the spatial process and the spatial evolution of urban land uses when compared with urban forms in the USA, because these models impose the land-use transition rules there is very little policy-relevant context to these models.

This limitation is overcome in a cellular automata approach used by Irwin and Bockstael (1999) to predict patterns of land use change in an urbanizing area, in which the transition probabilities are estimated as functions of a variety of exogenous variables and an interaction term that captures the effect of neighboring land use conversions. They then use a cellular automata model, with estimated parameters of land conversion, to demonstrate that a model in which there are spatial interactions between the land use decisions of individuals result in the evolution of a fragmented land-use pattern that is qualitatively much more similar to the observed pattern of development than the pattern that is predicted by models that ignore these spatial interactions. Models that include improved understanding of the spatial evolution of urban land use, when linked with a spatial ecosystem model, will be able to make much better predictions of the effects of government policies concerning land use on ecosystems.

3. *Conclusions*

In order to improve the understanding and modeling of spatially explicit human behavior in relationship to the environment, improved datasets, methods, and theory need to be developed (Irwin and Geoghegan 2000). All three of these issues are closely linked. As argued above, the historical paucity of spatially explicit social science data constrained spatial modeling of human behavior. As a result, interest in spatial issues in the social sciences, beyond the field of geography, has been sparse. But now that spatial social science data are becoming available, there has been a renewed interest in spatial issues. In a recent discussion of future research issues in natural resource and environmental economics, Deacon and colleagues (1998) comment, 'the spatial dimension of resource use may turn out to be as important as the exhaustively

studied temporal dimension in many contexts'—a conclusion that the regional and spatial science of the 1960s and 1970s argued previously. Therefore, having spatially explicit social science data should result in more sophisticated modeling of spatial human behavior and hypotheses testing. However, this necessitates improved empirical methods, such as those being developed within the field of spatial statistics and econometrics. A better understand of human behavior over space, coupled with better data and improved tools are necessary for advances in spatially explicit environmental problem solving.

See also: Ecological Economics; Economic Geography; Environment and Common Property Institutions; Environmental and Resource Management; Environmental Economics; Exhaustible Resources, Economics of; Land Use and Cover Change; Location Theory; Regional Science; Resource Geography; Spatial Analysis in Geography; Spatial Autocorrelation; Spatial Choice Models; Spatial Data; Spatial Pattern, Analysis of.

Bibliography

Albers H J 1996 Modeling ecological constraints on tropical forest management: spatial interdependence, irreversibility, and uncertainty. *Journal of Environmental Economics and Management* **30**: 73–94

Bhat M G, Huffaker R G, Lenhart S M 1996 Controlling transboundary wildlife damage: modeling under alternative management scenarios. *Ecological Modelling* **92**: 215–24

Bockstael N E 1996 Modeling economics and ecology: the importance of a spatial perspective. *American Journal of Agricultural Economics* **78**: 1168–80

Bockstael N E, Bell K P 1997 Land use patterns and water quality: the effect of differential land management controls. In: Just R E, Netanyahu S (eds.) *International Water and Resource Economics Consortium, Conflict and Cooperation on Trans-boundary Water Resources*. Kluwer Publishers, The Netherlands, pp. 169–91

Chomitz K M, Gray D A 1996 Roads, land use and deforestation: a spatial model applied to Belize. *World Bank Economic Review* **10**: 487–512

Clarke K C, Hoppen S, Gaydos L 1997 A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay Area. *Environment and Planning B: Planning and Design* **24**: 247–61

Deacon R T, Brookshire D S, Fisher A C, Kneese A, Kolstad C, Scrogin D, Smith V K, Ward M, Wilen J 1998 Research trends and opportunities in environmental and natural resource economics. *Environmental and Resource Economics* **11**: 383–97

Geoghegan J, Wainger L A, Bockstael N E 1997 Spatial landscape indices in a hedonic framework: an ecological economics analysis using GIS. *Ecological Economics* **23**: 251–64

Geoghegan J, Cortina Villar S, Klepeis P, Macario Mendoza P, Ogneva-Himmelberger Y, Roy Chowdhury R, Turner II B L, Vance C 2000 Modeling tropical deforestation in the southern Yucatan peninsular region: comparing survey and satellite data. In: *Agriculture, Ecosystems, and Environment* (under consideration)

Hof J, Bevers M, Joyce L, Kent B 1994 An integer programming approach for spatially and temporally optimizing wildlife populations. *Forest Science* **40**: 177–91

Hunziker M, Kienast F 1999 Potential impacts of changing agricultural activities on scenic beauty: A prototypical technique for automated rapid assessment. *Landscape Ecology* **14**: 161–76

Irwin E G, Bockstael N E 1999 Interacting agents, spatial externalities, and the endogenous evolution of residential land use pattern. Department of Agricultural and Resource Economics, University of Maryland, College Park, Working Paper 98–24

Irwin E G, Geoghegan J 2000 Theory, data, methods: developing spatially-explicit economic models of land use change. In: *Agriculture, Ecosystems, and Environment* (under review)

Mertens B, Lambin E F 1997 Spatial modeling of deforestation in southern Cameroon. *Applied Geography* **17**: 143–62

National Research Council 1997 *The Future of Spatial Data and Society: A Summary of a Workshop*. National Academy Press, Washington DC

Pfaff A S P 1999 What drives deforestation in the Brazilian Amazon? *Journal of Environmental Economics and Management* **37**: 26–43

Sanchirico J N, Wilen J E 1999 Bioeconomics of spatial exploitation in a patchy environment. *Journal of Environmental Economics and Management* **37**: 129–50

Swallow S K, Talukdar P, Wear D N 1997 Spatial and temporal specialization in forest ecosystem management under sole ownership. *American Journal of Agricultural Economics* **79**: 311–26

Veldkamp A, Fresco L 1997 Exploring land use scenarios: an alternative approach based on actual land use. *Agricultural Systems* **55**: 1–17

White R, Engelen G, Uljee I 1997 The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics. *Environment and Planning B: Planning and Design* **24**: 323–43

J. Geoghegan

Special Education in the United States: Legal History

People with disabilities in most parts of the world have often been excluded from many of the usual privileges and benefits of the societies in which they lived. This has been as true of schools as of any other social institution. Although the rhetoric of public education, in particular, has often emphasized equality of opportunity, the practice of education well into the twentieth century belied this rhetoric. A recent report on educational practices in China, for example, indicates that students with disabilities are routinely rejected by universities regardless of how intelligent or hard working they are. As a result, most of China's 20 million young people with physical disabilities stand

little chance of receiving a higher education (Mooney 2000).

Institutionalized inequalities in schools have been most evident in the ways that children of racial and ethnic minorities have been excluded from equitable learning opportunities, sometimes under the rationale that they could be provided a 'separate but equal' education. Less known is the history of exclusion of children with disabilities through legal and administrative practices. In 1893, a child was expelled from the Cambridge Public Schools in the USA because he was claimed to be 'weak in mind.' The expulsion was upheld by the Massachusetts Supreme Court (*Watson v. City of Cambridge* 1893).

The extent to which school decisions leading to the exclusion of children with disabilities were based on factors other than a child's academic ability is also illustrated by the wording of another court decision. In 1919, the Supreme Court of Wisconsin heard the case of a child with cerebral palsy who had the academic ability and physical capacity to be in school, but who drooled, had impaired speech, and had frequent facial muscle contractions. The court ruled that the child could be denied continuation in public school because he 'produces a depressing and nauseating effect upon the teachers and school children' (*Beattie v. State Board of Education*, *City of Antigo* 1919).

During the first half of the twentieth century, exclusion of children with disabilities from schooling was common and legal. A striking example of the justification given for segregated education for children with disabilities is the statement of the Superintendent of the Baltimore Public Schools in 1908. James Van Sickle said:

> If it were not for the fact that the presence of mentally defective children in a school room interfered with the proper training of the capable children, their (separate) education would appeal less powerfully … But the presence in a class of one or two mentally or morally defective children so absorbs the energies of the teacher and makes so imperative a claim upon her attention that she cannot under these circumstances properly instruct the number commonly enrolled in a class. School authorities must therefore, greatly reduce this number, employ many more teachers, and build many more school rooms to accommodate a given number of pupils, or else they must withdraw into small classes these unfortunates who impede the regular progress of normal children. The plan of segregation is now fairly well established in large cities, and superintendents and teachers are working on the problem of classification, so that they may make the best of this imperfect material (Van Sickle 1908, pp. 102-3).

The idea that children with disabilities have a right to an education, and that they should be educated in integrated settings, somewhat haltingly followed the racial integration implementation efforts that were put into motion following *Brown v. Board of Education* decision in the USA. Although there were increasing allocations of funds in the 1960s for training special education teachers and supporting special classes, these were usually incentive programs that school districts could choose to participate in, or not. The question of a right to an education for children with disabilities was not addressed for decades after the racial integration of schools was mandated.

In 1975, the interest and commitment in the US Congress to special education had evolved to the point that legislation was passed that provided the basic foundation for the right to education for students with disabilities that exists today. That legislation, entitled the Education of All Handicapped Children Act of 1975 (PL 94-142 1975), delineated that appropriate educational services be provided to all students with disabilities. It provided funding for the implementation of these services. It also mandated the manner in which these services should be planned and provided, and it included safeguards for students and their families.

The statement in the law that most directly addressed the issue of integrated education for students with disabilities is one that provides for what has come to be known as the least restrictive environment: 'Each public agency shall insure that special classes, separate schooling, or other removal of handicapped children from the regular educational environment occurs only when the nature or severity of the handicap is such that education in regular classes with the use of supplementary aids and services cannot be achieved satisfactorily' (*Public Law* 94-142 1975).

In 1986 the law was amended by PL 99-457. This legislation required a free appropriate education for preschool children aged three to five with disabilities. It also provided funding incentives to encourage the development of early intervention programs for infants and toddlers with disabilities, and those at risk of developing disabilities.

PL 94-142 was further amended in 1990. Among other modifications, this amendment changed the name of the existing law to the Individuals with Disabilities Education Act or, as it is now commonly called, IDEA. The IDEA included language that was very explicit in its call for including children with disabilities in regular school programs. It stated that schools must have a continuum of placements available in order to appropriately meet the needs of these students. It emphasized, however, that unless a child's individualized education program specifies some other necessary arrangement, the student must be educated in the school he or she would attend if he or she did not have a disability. Thus the IDEA reiterated the requirement that children with disabilities and those without disabilities be educated together to the maximum extent appropriate. The law also underscored the requirement that special classes, separate schooling, or other forms of segregation of students with disabilities occur only when the education of these students with the use of supplementary aids and

services cannot be achieved satisfactorily in regular classrooms (Individuals with Disabilities Education Act 1990).

1. The Institutionalizing of Special Education: A Conceptual History

It is ironic that PL 94-142 has been referred to as 'the mainstreaming law.' Mainstreaming, as both a concept and a term, has been used widely but defined very imprecisely. Like other educational terms, it has been used by professionals, parents, and politicians as if there were a common understanding of what was being communicated. In fact, that has often not been the case. Rather, the term has meant different things to different individuals and groups.

One early attempt to offer a model of mainstreaming outlined three elements that should characterize it: a continuum of services for students with disabilities, a reduction in the number of children 'pulled out' of regular classes, and increased provision of special services within regular classrooms rather than outside of those classrooms (Berry 1972).

The impetus for mainstreaming as it was defined by Berry, and as it was interpreted elsewhere under the term least restrictive environment, can perhaps be seen most clearly in Lloyd Dunn's 1968 article entitled 'Special Education for the Mildly Retarded: Is Much of It Justifiable?' In the article, Dunn called for special educators to consider carefully the efficacy research then available indicating that children with disabilities made greater academic progress in regular classrooms than in special classes. He urged special educators to resist 'being pressured into a continuing and expanding program (special classes) that we know now to be undesirable for many of the children we are dedicated to serve' (Dunn 1968, p. 5). He also argued that the labeling of children for placement in special classes led to a stigma that was very destructive to their self-concepts. Dunn asserted that 'removing a child from the regular grades for special education probably contributes significantly to his feelings of inferiority and problems of acceptance' (Dunn 1968, p. 9).

Following the publication of Dunn's article, there were others who called for the activating of Dunn's concerns in educational policy and practice. A 'zero reject model' was advocated that suggested that no child with mental retardation should be 'rejected' from a regular class and placed in a special class (Lilly 1970). The Council for Exceptional Children Policies Commission adopted a policy that specified that all students with disabilities 'should spend only as much time outside of regular classroom settings as is necessary to control learning variables' (Council for Exceptional Children Policies Commission 1973, p. 494).

Even though it was variously defined and understood, the concept of mainstreaming grew in importance and in its acceptance among professionals and the public during the 1970s and 1980s. However, the actual practice of mainstreaming was not, according to some critics, as common as the rhetoric that endorsed it. In 1986 a new call for the integration of children with disabilities into regular education programs was issued from the Assistant Secretary for Special Education and Rehabilitative Services of the US Department of Education. Secretary Madeline Will proposed what she called the Regular Education Initiative (REI). The REI proposed the restructuring of American education into a single system for the delivery of services to all children. Will argued that by merging special and regular education, a 'shared responsibility' would be created that would serve children without the stigma of diagnostic labels or segregated classrooms (Will 1986).

Will's proposal for the REI generated a great deal of controversy and concern in the special education profession. In 1987, for example, a group of leaders in the field met to assess the implications of the proposal. They praised the intent of Secretary Will's call for restructuring but cautioned that it 'is important to acknowledge that special education cannot seek institutional solutions to individual problems without changing the nature of the institution … school organization is what fundamentally must change' (Heller and Schilit 1987). This idea, that special education can change in fundamental ways only if the institution of the public school changes, is to be found repeatedly in the literature on special education reform that has been written in recent years (McLaughlin and Warren 1992).

The critics of the REI voiced strong concern over the implications of the proposal. Kauffman, for example, expressed fears that the special services needed by children with disabilities might be diluted or eliminated if the REI became a reality. He also argued that most regular classroom teachers do not have the training and/or the inclination to work with many of the students who are served by special education. In language reflective of his strong beliefs on this issue, Kauffman stated, 'The belief systems represented by the REI are a peculiar case in which both conservative ideology (for example, focus on excellence, federal disengagement) and liberal rhetoric (for example, nonlabeling, integration) are combined to support the diminution or dissolution of a support system for handicapped students' (Kauffman 1989, p. 273).

Supporters of the REI asserted that the response to Lloyd Dunn's challenge in 1968 to the efficacy of special classes had not been genuine or widespread. They questioned the essential validity of what had been tried and accomplished during the 1970s and 1980s in the name of mainstreaming. REI proponents maintained that only through a merger of special education into the regular education structure could all children be given the educational services they need without the stigma that had become associated with special education (Robinson 1990).

The interest in a merger of special and regular education as it was advocated by REI supporters proved, however, to be a one-dimensional interest. REI advocates were almost exclusively special educators. Ironically, regular education showed very little interest in the REI. This lack of interest in a merger of structures and efforts between regular and special education inspired an early critic to write: 'We have thrown a wedding and neglected to invite the bride. If this is an invitation to holy matrimony, it was clearly written by the groom (special education), for the groom, and the groom's family. In fact, the bride (general education) wasn't even asked. She was selected' (Lieberman 1985, p. 513).

2. *Inclusion and the Future of Special Education*

The most recent descriptor for the effort to create greater integration of children with disabilities into school programs is the term inclusion. For many educators, the term is viewed as a more positive description of efforts to include children with disabilities in genuine and comprehensive ways in the total life of schools. Whereas mainstreaming might have meant the same thing for many people, it could also be taken as having as its primary focus the physical presence of a child with disabilities in a regular classroom. Mainstreaming in this form could be only the appearance of integration. The REI was seen by some people, even some of its supporters, as meaning the dismantling of special education.

Inclusion, however, can mean that the goal of education for children with disabilities is the genuine involvement of each child in the total life of the school. Genuine inclusion means welcoming children with disabilities into the curriculum, environment, social interaction, and self-concept of the school.

Inclusion, of course, can (and already does) mean different things to different people. Some people interpret it to be just a new way of talking about mainstreaming. For others it may be seen as the REI with a new label. Some have even used the term inclusion as a banner for calling for 'full inclusion' or 'uncompromising inclusion,' meaning the abolishment of special education (Fuchs and Fuchs 1994).

Debates over the nuances and fine points of difference between mainstreaming, the REI, inclusion, full inclusion, and whatever the next term or concept may be are surely not reflective of the grassroots concerns of parents, teachers, and school leaders who are earnestly trying to do their best for children with disabilities. Coupled with the other vagaries and demands that are placed on parents and educators in our increasingly complex society, the confusion that exists over the rights and responsibilities of the various constituencies in the education arena must seem overwhelming.

The most sound basis for providing an inclusive education for students with disabilities in public schools must ultimately consist of the legislative foundations that create services for these students, and the interpretation that courts give to questions about this foundation.

The legislative foundation for inclusion in the USA, for example, is IDEA. It is clear that this statute stress two things: the policy of the least restrictive environment and the requirement that a continuum of services must be available so that an appropriate education may be provided to every child who has the need for special education. Court decisions about this foundation are a fitting topic for the close of this discussion of the institutional aspects of special education.

Although the cases and decisions of US courts concerning inclusion have varied, the principles derived from this litigation have been clear and compelling. Yell (1995) has shown that school districts must make every good-faith effort to maintain students with disabilities in general class placements with whatever appropriate aids and services might be needed to support a student in that environment. School districts that have not done so have consistently been found by courts to be in violation of the law. On the other hand, it is apparent from judicial review and decisions that some children may be so disruptive to general classrooms that these settings are not in their best interests or those of their nondisabled peers. The following suggestions to educators and others interested in the inclusion of children with disabilities in the total school environment are based on Yell's study (1995) of court decisions concerning the issue of the least restrictive environments for the education of students with disabilities:

(a) Determination of the LRE (least restrictive environment) must be based on the individual needs of the child.

(b) Good-faith efforts must be made to keep students in integrated settings. The beginning assumption should always be that a student belongs in a general classroom. Exceptions may be made to this assumption based on individual needs.

(c) A complete continuum of alternative placements and services must be available.

(d) In making LRE/inclusion decisions, the needs of the student's peers should be considered.

(e) When students are placed appropriately in more restrictive settings, they must still be integrated into general classes and other school activities to the maximum extent feasible and beneficial to the student.

(f) The LRE/inclusion decision-making process must be thoroughly documented.

The achievement of an individualized and appropriate educational program for any student is a dynamic process. Decisions regarding the most effective blend of special services and inclusive practice must be continually reviewed and changed in the best interests of the student.

See also: Disability, Demography of; Disability: Psychological and Social Aspects; Disability: Sociological Aspects; Law and People with Disabilities; Mental Retardation: Clinical Aspects; Mental Retardation: Cognitive Aspects

Bibliography

Beattie V 1919 State Board of Education, City of Antigo, 169 Wisc. 231, 172 N.W. 153

Berry K 1972 *Models for Mainstreaming*. Dimension, San Rafael, CA

1973 Proposed CEC Proposed CEC policy statement on the organization and administration of special education. *Council for Exceptional Children Policies Commission* **39**: 493–97

Dunn L M 1968 Special education for the mildly retarded: Is much of it justifiable? *Exceptional Children* **34**: 5–22

Fuchs D, Fuchs L 1994 Inclusive schools movement and the radicalization of special education reform. *Exceptional Children* **60**: 294–309

Heller H W, Schilit J 1987 The regular education initiative: A concerned response. *Focus on Exceptional Children* **20**: 1–7

1990 *Individuals with Disabilities Education Act* (PL 101-476) **20 USC:** Ch. 33, Sec. 1400–1485

Kauffman J M 1989 The regular education initiative as Reagan–Bush education policy: A trickle-down theory of education of the hard-to-teach. *Journal of Special Education* **23**: 256–78

Lieberman J M 1985 Special education and regular education: A merger made in heaven? *Exceptional Children* **51**: 513–16

Lilly M S 1970 Special education: A tempest in a tempest. *Exceptional Children* **37**: 43–48

McLaughlin M, Warren S 1992 *Issues and Options in Restructuring Schools and Special Education Programs*. Center for Policy Options in Special Education/University of Maryland, College Park, MD

Mooney P 2000 A college for disabled students meets indifference in Beijing. *The Chronicle of Higher Education* **46**: 1–6

PL94-142 1975 The Education of All Handicapped Children Act of 1975 *Public Law 94-142* 34 C.F.R. 300.550 (b) (2)

Robinson V F 1990 Regular education initiative: Debate on the current state and future promise of a new approach to educating children with disabilities. *Counterpoint* **5**: 3–5

Van Sickle J H 1908 Provision for exceptional children in the public schools. *Psychological Clinic* **2**: 102–11

Watson v. City of Cambridge 1893 157 Mass. 561, 32 N.E. 864

Will M C 1986 Educating children with learning problems: A shared responsibility. *Exceptional Children* **52**: 411–15

Yell M L 1995 Least restrictive environment, inclusion, and students with disabilities: A legal analysis. *Journal of Special Education* **28**: 389–404

J. D. Smith

Specialization and Recombination of Specialties in the Social Sciences

Specialization is an old phenomenon, observed as early as Auguste Comte. Recombination of specialties is also an old trend: Robert Merton noticed it in the early 1960s: 'the interstices between specialties become gradually filled with new cross-disciplinary specialties' (Merton 1963, p. 253). The multiplication of scientific activities and the growing complexity of knowledge have brought a division of labor, visible in increasing expertise in all domains. The history of scientific advancement is a history of concatenated specialization. In spite of this evidence, 'even since the social sciences began to develop, scholars have repeatedly expressed apprehension about the increasing fragmentation of knowledge through specialization' (Smelser 1967, p. 38).

Fragmentation and specialization are the two faces of the same coin. This double process is a condition of scientific advancement. The term 'discipline' is inherited from the vocabulary of the nineteenth century and is understood as a branch of instruction for the transmission of knowledge and as a convenient mapping of academic administration. In all universities, teaching, recruitment, promotion, peer review, and financing are still organized along disciplinary lines, but the great scientific laboratories and the most creative research centers are anchored on specialties, which in many cases cross the borders of formal disciplines, or are located at the interstices between disciplines. In the sociology of science the term 'discipline' is defined as a 'cluster of specialties' (Crane and Small 1992, p. 198), the specialty being the main source of academic recognition and of professional legitimation (Zuckerman 1988, p. 539).

In the natural sciences, specialization is celebrated as a mark of competence. Each contemporary natural science from astronomy to zoology is a conglomerate of dozens of specialties. Only in some social sciences, particularly in sociology, political science, psychology, and anthropology, does one hear complaints and lamentations about the fragmentation of the discipline into specialties. The word 'crisis' appears in the title of many books and articles about the identity and the contours of these disciplines (Horowitz 1993, S. Turner and J. Turner 1990). In reality, such 'crises' are symptoms of growth and expansion.

1. Specialization

Analysis means breaking things into parts. All sciences have made progress, from the sixteenth century on, by internal differentiation and cross-stimulation among emergent specialties. Each specialty developed a patrimony of knowledge. With the growth of these patrimonies specialization became a necessity. Increasing specialization has led to the creation of subdisciplines. A specialty requires a high qualification in a delimited domain. The process of specialization has tended to disjoin activities which had previously been united, and to separate scholars belonging to the same formal discipline, but who are interested in different fields. Specialties are considered as parts of a larger discipline but 'the modern scholar normally does not attempt to

master his discipline in its entirety. Instead he limits his attention to one field, or perhaps two' (Somit and Tanenhaus 1964, p. 49). Each formal discipline gradually becomes unknown in its entirety. No theory can any longer encompass the whole territory of a discipline. Talcott Parsons was the last sociologist to attempt such a theory. As a discipline grows, its practitioners become specialized and must neglect other areas of the discipline. It is for this reason that scissiparity, the amoeba-like division of a discipline into two, is a common process of fragmentation. Specialization provides researchers with methods so that they need not develop them anew. Refinements in methods are more easily transmitted among specialists. Specialization within disciplines is visible in national and international professional meetings. Everyone who has participated at a gathering of several thousand people has noticed the absence of coherence: 20–30 panels may run simultaneously, most of them mobilizing only a handful of people. The plenary sessions attract only a small minority, with most participants uninterested in issues encompassing the entire discipline.

Disciplines fragment along substantive, epistemological, methodological, theoretical, and ideological lines. To those in the field, the theoretical and ideological divisions are likely to seem more important than to others. Ralph Turner gives a description of this process in sociology, based on his experience as Editor of an important journal: 'In the 1930s and 1940s the aspiration to be a general sociologist was still realistic. There was a sufficiently common body of core concepts and a small enough body of accumulated research in most fields of sociology that a scholar might make significant contributions to several, and speak authoritatively about the field in general. It is difficult to imagine the genius necessary for such accomplishments today' (R. Turner 1991, p. 70).

The division of disciplines into specialties should be distinguished from their fragmentation into schools and sects. The term school refers to a group of scholars 'who stress a particular aspect'. The proponents of a school are emotionally committed to their approach. A school is simultaneously a subdivision of a field and a species of cult or sect (Smelser 1967, pp. 6–7). While the multiplication of specialties is related to the maturation of the discipline, 'the presence of numerous schools in a discipline generally betokens a relative scientific immaturity' (Smelser 1967, pp. 6–7). The same diagnosis is formulated by Gabriel Almond who, by crossing two dimensions, an ideological one and a methodological one, described four types of schools in political science (Almond 1990, p. 15). These types are not determined by the process of specialization.

Various authors stress the importance of the patrimony's expansion for subsequent fragmentation. As disciplines grow, they fragment and most parts become the patrimonies of individual subfields; a rare few are passed down in the patrimony of several formal disciplines and become the classics. For Randall Collins, the increasing specialization in sociology could be explained by the growth in scale in the profession during the late twentieth century: 'How does one make oneself visible when the sheer number of competitors increases? The field differentiates into various specialties. Instead of seeking recognition in the larger field, one opts for a smaller arena within which the discourse and the search for originality and for allies can go on' (Collins 1986, p. 1340). A process of differentiation without integration is visible across disciplinary borders. Specialists 'seem scarcely to recognize the names of the eminent practitioners in specialties other than their own' (Collins 1986, p. 1340).

Another scholar who emphasizes the importance of specialties in the organization of scientific communities is Harriet Zuckerman: 'Much historical and sociological evidence suggests that specialties have been prime audiences for many scientists: they are the explicit and tacit audiences—the reference groups—to which they address their work, just as they are the prime sources of wherewithal and rewards for that work' (Zuckerman 1988, p. 539). As in some cathedrals, there are more ceremonies in the chapels than under the principal nave.

2. *Specialties Versus Disciplines*

Analyzing the relationships between specialties *within* disciplines and between specialties *across* disciplines, two kinds of disciplines can be perceived: 'restricted disciplines, such as most physical sciences, which would be expected to exhibit a high degree of linkage between different research areas within the discipline, but less linkage to other disciplines,' and unrestricted sciences, such as most social sciences, which would be likely to exhibit relatively diffuse links among research areas both within and outside the disciplines (Crane and Small 1992, p. 200).

Depending on the definition we adopt, there are in the social sciences between 12 and 15 formal disciplines, but dozens of specialties, sectors, fields, and subfields. In sociology, for instance, there are some 50 specialties, as indicated by the list of research committees of the International Sociological Association. There are as many in the International Political Science Association. Most of these groups cooperate to a limited extent within their respective discipline.

There is more communication between specialties belonging to different disciplines than between specialties within the same discipline. Sociometric studies show that many specialists are more in touch with colleagues who belong officially to other disciplines than with colleagues in their own discipline. The 'invisible college' described by Robert Merton and other sociologists of science is an eminently interdisciplinary institution because it ensures communication

not only from one university to another and across all national borders, but also, and above all, between specialists attached administratively to different disciplines. The networks of cross-disciplinary influence are such that they are obliterating the old classification of the social sciences.

The division of disciplines into subfields tends to be institutionalized, as can be seen in the organization of large departments in many American and European universities. Another good indication of the fragmentation of the discipline is the increasing number of specialized journals. In the last two decades several hundred specialized journals in English relevant to social sciences have been launched. Most of these new journals cross the borders of two or three disciplines. Some other new journals have appeared in French and German. European unification has had an impact on the development of cross-national journals focusing on special fields. Specialization has consequences on the role of national professional associations and on the general journals. The size of specialties varies greatly: Egyptology and biosociology are small specialties; socialization and clientelism are very large. A segment which drifts away from the main body of the discipline has to reach a certain critical size before being recognized. Some of them enjoy a relative autonomy within the discipline.

A distinction has to be drawn between specialization within a formal discipline and specialization at the intersection of several monodisciplinary subfields. The latter can occur only after the former has become fully developed.

3. *Recombination of Specialties Across Disciplines*

Specialization within disciplines, accompanied by their fragmentation, is the first process. Recombination of specialties across disciplinary borders is the second. In the history of science a twofold process can be seen: on the one hand, a fragmentation of formal disciplines and, on the other, a cross-disciplinary recombination of the specialties resulting from fragmentation. The new field may become independent, like social psychology, or may continue to claim a dual allegiance, like historical sociology. In the latter case, librarians are not sure whether to place a work in the category of history or of sociology.

Such a recombination has been called hybridization (Dogan and Pahre 1990, p. 63). A hybrid is a combination of two branches of knowledge. Jean Piaget suggested a different analogy: genetic recombination (Piaget 1970, p. 524). The reasons for hybridization are clear enough: specialization leaves gaps between disciplines; specialties fill the gaps between two existing fields. Many of the new specialties are to a large degree hybrid specialties. The phenomenon is even more obvious in the natural sciences, e.g., in the awards of the Nobel prizes.

Some scholars recommend an 'interdisciplinary' approach. Such a recommendation is not realistic because it overlooks the process of specialization. Research enlisting several disciplines involves in fact a combination of segments of disciplines, of specialties, and not whole disciplines. The fruitful point of contact is established between sectors, and not along disciplinary boundaries. Considering the trends in the social sciences, the word 'interdisciplinarity' appears inadequate since it carries a hint of superficiality and dilettantism, and consequently should be avoided and replaced by hybridization of specialties. Different disciplines may proceed from different foci to examine the same phenomenon. This implies a division of territories between disciplines. In contrast, hybridization implies an overlapping of segments of disciplines, a recombination of knowledge in new specialized fields. Innovation in each discipline depends largely on exchanges with other fields belonging to other disciplines.

When old fields grow they accumulate such a mass of material in their patrimony that they split up. Each fragment of the discipline then confronts the fragments of other fields across disciplinary boundaries, losing contact with its siblings in the old discipline. A sociologist specialized in urbanization has less in common with a sociologist studying elite recruitment than with a geographer interested in the distribution of cities; the sociologist studying social stratification has more in common with his colleague in economics analyzing income inequality than he does with a fellow sociologist specialized in organizations; psychologists studying child development are much more likely to be interested in developmental physiology or in the literature on language acquisition than they are in other branches of psychology. A political scientist researching political socialization reads more sociological literature on the agents of socialization (family, church, school, street-corner society, cultural pluralism, etc.) than literature in political science on the Supreme Court, legislative processes, party leadership, or the recruitment of higher civil servants. Those working on the subfield of international relations in the nuclear age have little recourse to the classical literature in political science, but rather to economics, technology, military strategy, diplomatic history, game theory, nuclear physics, and engineering.

Most hybrid specialties and domains recognize their genealogical roots: political economy, political sociology, social geography, historical sociology, genetic demography, psycholinguistics, social anthropology, social ecology, biogeography, and many others. Some hybrid sciences do not indicate their filiation in their name: cognitive science, paleoarcheology. Eight social sciences also have roots in the natural sciences: demography, geography, psychology, anthropology, linguistics, archeology, criminology, and cognitive

science. The hybrid specialties branch out in turn giving rise, at the second generation, to an even larger number of hybrids.

Recombination of specialties appears in the borrowing of concepts, theories and methods (Dogan 1996). Concepts are vehicles for communication between specialties belonging formally to different disciplines. Concepts are not equally important in all specialties. For instance, alienation is more frequently used in social psychology than in political economy, segregation more in urban studies than in comparative politics, and charisma more in political sociology than in social ecology.

There are very few general theories accepted by the majority of scholars in any discipline; the best known theories are specialized theories: they are what Robert Merton calls middle-range theories, which do not encompass an entire discipline, but only a part of it. A specialized theory can cross several disciplines with a minimal effort of adaptation. For instance 'rational choice'—considered by some as a theory and by others as a school or sect—is, except for some stylistic features, basically the same in political science, in sociology, and in economics. The theory of methodological individualism does not change on moving from one discipline to another. Two to three dozen theories peregrinate freely among formal disciplines. The high interdisciplinary mobility of specialized theories and of specialized concepts, and also of research methods, gives some centripetal orientation to the wide range of social sciences. The cross-disciplinary diffusion of concepts and theories is in fact a diffusion among specialties belonging to different disciplines.

Methodology divides some disciplines as much as ideologies, and at the same time builds bridges between specialties across the disciplinary borders. The scholars who practice factor analysis understand each other well no matter what their official disciplinary affiliation, but most of them encounter some difficulties in communicating, *within* their discipline, with their colleagues who do not practice quantitative methods. In some departments methodological controversies generate factions, which may or may not coincide with divisions between specialties.

The citation patterns in academic publications permit us to measure empirically the degree of coherence of a discipline, the relations between specialties within disciplines, and the interferences of disciplines. If specialists in a given subdiscipline tend to cite mostly, or exclusively, specialists of the same subdiscipline, and if relatively few authors cite across the borders of their subdiscipline, then the formal discipline presents a low degree of internal coherence. In this case the real loci of research are the specialties. If, on the contrary, a significant proportion of authors cross the borders of specialties, then the discipline as a whole appears more or less as an integrated territory (Dogan 1997). Political scientists have borrowed in the 1930s from law, in the 1950s from sociology and law, and in the 1970s from sociology, philosophy, economics, history, and psychology. According to Craig Calhoun, 'Sociology is neither the most open nor the most insular of social sciences' (Calhoun 1992, p. 143). A study of patterns of citation during the period 1936–75 has found that sociologists cited 58 percent of the time articles in sociology journals, political scientists cited 41 percent of the time scholars from their own discipline, anthropologists referred 51 percent of the time to their colleagues, psychologists 73 percent to their own kin, and 79 percent of the economists did the same (Rigney and Barnes 1980, p. 117). In each social science a significant proportion of theoretical, methodological, and substantive communication has been with other disciplines, the most open being, in the late twentieth century, political science and the most autarchic economics.

In an analysis of journals identified as belonging to sociology and economics, a significant shift has been found from sociology to 'interdisciplinary sociology' and from economics to 'interdisciplinary economics,' between 1972 and 1987. The criteria for trade between disciplines was the proportion of citing references in journals of the respective discipline (Crane and Small 1992, pp. 204–5). An analysis by the same authors in terms of clusters of references shows a clear increase in interdisciplinary relationships.

Most social scientists are specialists contributing primarily to their own subdiscipline. For instance, the *Handbook of Sociology* edited by Neil Smelser (1988) is divided into major specialties, without an overarching structure. Smelser is aware of this heterogeneity. Noting that Talcott Parsons had 'exaggerated the internal unity of the discipline,' which 'if applied today would appear almost ludicrous' (p. 12), he writes: 'there appears to be no present evidence of an overarching effort at theoretical synthesis … and little reason to belief that such an effort is on the horizon,' adding that 'we have by now gone too far down the road of specialization and diversification' (p. 12). The ship of sociology is already built with watertight compartments. If the borders between specialties *within* disciplines are hermetic, in exchange, paradoxically enough, the frontiers between formal disciplines are open.

A distinction is needed between interdisciplinary amalgamation and hybridization by recombination of specialties belonging to different disciplines. A 'unified social science' was visible only in the early phase of the development of social sciences, when disciplines were not yet differentiated. Hybridization of specialties came much later, after the maturation of the process of internal fragmentation of disciplines. The word interdisciplinary is appropriate only for a certain phase of social sciences at the beginning of the twentieth century; it becomes misleading when it is used to describe contemporary trends, because at the end of the twentieth century only specialties are overlapping, not entire disciplines.

Some may confound recombination and synthesis. A synthesis brings a new interpretation, a personal or stylistic achievement. The difference is most clear in history. Arnold J. Toynbee's theory of history and Hobsbawn's *The Age of Empire* are syntheses. Fernand Braudel's *The Mediterranean World* is a recombination of segments of social sciences, largely history and geography. Perry Anderson's *Lineages of the Absolutist State* is largely synthesis, whereas Karl Wittfogel's *Oriental Despotism* is largely recombination.The recombinations are varied. Child development includes developmental physiology, language acquisition, and socialization. Indo-European studies encompass segments of historical linguistics, archeology, prehistory, and botany. Scholars in criminology come from subfields in law, sociology of deviance, social psychology, endocrinology, urban studies, social economics, and ethnopolitics. The study of artificial intelligence encompasses formal logic from philosophy, grammar and syntax from linguistics, and computer programming from computer science. Folklore studies includes sectors from historical linguistics, cultural anthropology, social history, and comparative literature.

In one generation the number of social scientists has increased 10-fold. Their status in the profession depends mostly on their contribution to knowledge, and much less on their performance as teachers. As teachers they are generalists, as researchers, specialists. A strong feeling of diminishing marginal returns is perceived among the mass of social scientists. This phenomenon of saturation and overcrowding has been called the 'paradox of density': the tendency of densely populated disciplines to produce less innovation per head in spite of the greater effort applied in these overcrowded domains (Dogan and Pahre 1990, p. 32). What is perceived by some scholars as 'dispersion' and 'disintegration' should be considered as manifestations of the growth of specialties, resulting in the weakening of the saturated disciplinary core, a process engendered by centrifugal and autonomous tendencies of the specialties across disciplines. The recombination of specialties across disciplines requires a remodeling of the intellectual map of the behavioral social sciences.

Bibliography

Almond G A 1990 *Discipline Divided, Schools and Sects in Political Science*. Sage, Newbury Park, CA

Calhoun C 1992 Sociology, other disciplines and the project of a general understanding of social life. In: Halliday T C, Janowitz M (eds.) *Sociology and its Publics: The Forms and Fates of Disciplinary Organization*. University of Chicago Press, Chicago, pp. 137–96

Collins R 1986 Is 1980s sociology in the doldrums? *American Journal of Sociology* **91**: 1336–55

Crane D, Small H 1992 American sociology since the seventies: the emerging identity crisis in the discipline. In: Halliday T C, Janowitz M (eds.) *Sociology and its Publics: The Forms and Fates of Disciplinary Organization*. University of Chicago Press, Chicago, pp. 197–234

Dogan M 1996 Political science and the other social sciences. In: Goodin R, Klingemann H D (eds.) *A New Handbook of Political Science*. Oxford University Press, Oxford, UK, pp. 97–130

Dogan M 1997 Les nouvelles sciences sociales: fractures des murailles disciplinaires. *Revue Internationale des Sciences Sociales* **153**: 467–82

Dogan M, Pahre R 1990 *Creative Marginality: Innovation at the Intersections of Social Sciences*. Westview Press, Boulder, CO

Horowitz I L 1993 *The Decomposition of Sociology*. Oxford University Press, New York

Merton R 1963 The mosaic of the behavioral sciences. In: Merton R K, Berelson B (eds.) *The Behavioral Sciences Today*. Basic Books, New York, pp. 247–72

Merton R 1968 On sociological theories of the middle range. In: *Social Theory and Social Structure*. Free Press, New York, pp. 39–53

Piaget J 1970 General problems of interdisciplinary research in UNESCO. In: Piaget J (ed.) *Main Trends of Research in the Social and Human Sciences*. Mouton, Paris, pp. 467–528

Rigney D, Barnes D 1980 Patterns of interdisciplinary citation in the social sciences. *Social Science Quarterly* **61**: 114–27

Smelser N J 1967 Sociology and the other social sciences. In: Lazarsfeld P, Sewell W, Wilensky H (eds.) *The Uses of Sociology*. Basic Books, New York, pp. 3–44

Smelser N J (ed.) 1988 *Handbook of Sociology*. Sage, Newbury Park, CA

Somit A, Tanenhaus J 1964 *American Political Science, a Profile of a Discipline*. Atherton, New York

Turner R 1991 The many faces of American sociology. A discipline in search of identity. In: Easton D, Schelling C S (eds.) *Divided Knowledge Across Disciplines, Across Cultures*. Sage, London, pp. 59–85

Turner S P, Turner J H 1990 *The Impossible Science: An Institutional Analysis of American Sociology*. Sage, Newbury Park, CA

Zuckerman H 1988 Sociology of science. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, pp. 511–74

M. Dogan

Species and Speciation

At first glance, 'species' should be a relatively easy term to define: a group of organisms sharing a common biological lineage. Yet biologists have argued over the details of the definition since around 1900. This is in large part because the definition inevitably reflects concepts about processes resulting in the formation of species (or speciation), which lie at the heart of Darwinian evolution. Those speciation processes also underlie observed patterns of biodiversity in both space and time.

1. Species Concepts

Species concepts proposed since around 1950 may be broadly grouped into three categories: typological (based on shared traits), biological (based on breeding habits), and evolutionary (based on lineage) concepts. In the first case, the goal is usually a definition that follows taxonomic practice in species descriptions. The second and third categories reflect ideas about the speciation process.

1.1 Typological Species Concepts

Typological species concepts are based on the premise that a group of individuals of one 'type,' sharing a list of traits, constitute a species. This set of species concepts includes the phenetic species concept (Sokal and Crovello 1970) and the genotypic cluster definition (Mallet 1995). In these cases, species are individuals sharing well-defined clusters of traits, with few to no individuals with intermediate traits. Some versions of what are usually referred to as phylogenetic species concepts (e.g., Cracraft 1983) also fall into the typological category, as species are defined by clusters of shared traits, although individuals are also grouped by population or lineage (for asexual organisms).

Traits may be genotypes or phenotypes; that is, the genes possessed by an organism or the characters resulting from the expression of those genes. However, as its name implies, the genotypic cluster definition focuses on genotypes. Phenotypic traits are commonly measured as length, width, etc., of some body part, or as proteins separated and visualized on a gel. Genotypic traits tend to be quantified as DNA or RNA base sequences. For example, bacterial species that cannot be grown in the laboratory are often described based on their DNA differences and similarities.

Taxonomists routinely use these typological species concepts when they identify individual organisms to species. The definitions have the advantage that they can be applied to almost any organism, dead or alive, fossil or recent. All one needs to be able to do is to make measurements of the phenotype and/or the genotype of the organism in question. However, this is also the disadvantage: one has to choose appropriate traits to measure, which will highlight the differences and similarities among individuals.

1.2 The Biological Species Concept and Its Relatives

The biological species concept states that species are populations of interbreeding organisms. This concept, strongly espoused by Mayr (1942), was dominant until the 1980s. It continues to be widely cited.

The initial formulation of the biological species concept focused on reproductive isolation and hence restriction of gene exchange between species. The inverse of inter-specific reproductive isolation, use of shared mate recognition and fertilization systems within a species, was later put forward by Paterson (1985). For example, two species of frogs could be discriminated upon based on their inability to produce viable offspring, or alternatively by shared mating calls and ability to produce viable offspring within each species.

Some relatively obvious difficulties arise in applying this species definition to individual cases. First, the requirement for interbreeding makes the application to asexual or clonal organisms impossible. Second, many populations are geographically separated and it is not known whether they could interbreed or not if they were brought together in the same place. Third, application to the fossil record is nearly impossible, as the probability of finding two fossil organisms interbreeding is exceedingly low, and even if that were to occur, the question of viable offspring remains. Nonetheless, the concept is useful as an embodiment of the necessity for restricted gene exchange between species.

1.3 The Evolutionary Species Concept and Its Relatives

Evolutionary species concepts place weight on defining species based on organisms' historical or evolutionary lineage. These concepts arose initially out of the recognition that species do occasionally interbreed. Interbreeding results in movement of genes from individuals of one species into the gene pool of another (introgression), but in some cases, species maintain their morphological and ecological distinctness even in the face of gene introgression. Definitions by Simpson (1951) and Wiley (1978) focused on species as lineages that have their own evolutionary histories and fates.

Most phylogenetic species concepts also fall under this heading. A species is defined as those organisms within a lineage between the initial splitting off of the species (cladogenesis) and either extinction of the lineage or further cladogenesis (e.g., Hennig 1966). The idea has been refined further to assert that such lineages must be monophyletic (derived from one common ancestor). However, the observation of gene introgression suggests that while individual traits in a species may be monophyletic, it is unlikely that all traits will be.

In practice, application of these species concepts requires knowledge of evolutionary history, which at present is obtained through comparisons of traits and constructions of cladograms (or genealogical trees) representing phylogenetic hypotheses. The data used are thus similar to that described in Sect. 1.1.

1.4 Pluralistic Views of the Species Concept

While broader conceptualizations of species may be unwieldy in terms of the data required to define a set of

individuals as a species, such concepts have been put forward as more realistic and defensible. Templeton (1989) offered a cohesion species concept, which defines species based on a combination of factors promoting genetic divergence between populations along with those promoting genetic similarity within populations, and includes factors that limit the spread of variation through genetic drift or natural selection. This concept combines biological species and recognition concepts outlined above with a more explicit incorporation of evolutionary processes. Identification of species starts with identification of an evolutionary lineage (as in evolutionary species concepts), followed by tests of reproductive isolation or mate recognition systems, and culminating with examination of the uniformity of selection regimes or ecological adaptations within groups.

1.5 Remaining Problems with Species Concepts

Most of these discussions of species assume that species are now formed as well-defined entities. This is clearly not the case for any of the definitions or concepts presented in Sects. 1.1–1.4. We can identify races and sub-species, which eventually may form species, using the typological approach; hybrid zones between species challenge the biological species approach; and gene introgression poses problems for monophyly. As a result, further refinement of the species concept is likely in the future.

2. Speciation Mechanisms

In practice, speciation occurs through formation of barriers to gene exchange between groups of organisms, concurrent with or followed by genetic differentiation of the groups.

2.1 Pre- and Post-zygotic Barriers to Gene Exchange

The barriers to gene flow are often lumped into two classes: pre-zygotic and post-zygotic mechanisms, named depending on whether they occur before or after the uniting of the gametes to form a zygote. Pre-zygotic mechanisms include all the factors associated with mate location, courtship, and mating, either behavioral or morphological. They also include the process of gamete union to form a zygote. Post-zygotic mechanisms include hybrid inviability or sterility. In general it is believed that even though post-zygotic mechanisms may initially isolate two species, selection should favor the development of pre-zygotic mechanisms whenever variation allowing that development occurs. This is because the investment by individuals in a hybrid offspring is higher if only post-zygotic mechanisms are in place.

2.2 Specific Barriers to Gene Exchange

Specific barriers to gene exchange include morphology, behavior, and genetic and cytoplasmic incompatibility. The first, morphology, has been well documented in a number of insect-pollinated plants. The location of the anthers (pollen bearing structures) relative to the nectaries combined with the body size and shape of the insects determines whether visiting insects actually brush against the anthers and hence transport pollen to other flowers.

Speciation may also involve behavioral divergence between two groups. Behavioral changes involved in speciation are usually associated with courtship and mating rituals. For example, species' differences in mating calls occur in frogs and crickets. Pheromones, or chemicals use to communicate between individuals, may also change. Moths, butterflies, and flies, for instance, use pheromones as integral parts of mate attraction or courtship. The chemical composition of the pheromone mix is generally species-specific. Species may also be isolated by differences in the timing of courtship and mating either within the year or within the day. Seventeen year and 13 year cicadas are an extreme example of such temporal isolation. These cicadas develop underground, emerging as adults to mate at the end of a 17 or 13 year lifespan, depending on the species. Each of three species of 17 year cicadas in the USA has a corresponding very similar 13 year cicada species. Yet, 17 and 13 year cicadas will only overlap as adults once every 221 years. Finally, species may be effectively isolated if they have different courtship or mating locations. The fruit fly, *Rhagoletis pomonella*, is a classic example. Courtship and mating occur on the fruit tree species where the female will later lay eggs. Initially in USA, hawthorn fruits were used; a new race of the fly moved onto apple in the nineteenth century, and then onto sour cherry in the mid-twentieth century. *Rhagoletis* races on each of these hosts respond to specific odors and color, shape and size cues, effectively isolating them (summarized in Bush 1998).

Genetic incompatibility between species arises in several ways. For example, egg and sperm, or pollen and stigma can possess surface proteins that prevent either fusion of the egg and sperm into a zygote, or pollen tube growth allowing fertilization of the plant ovum. A modification of these incompatibilities is seen in patterns of sperm or pollen precedence within a species, which limit gene exchange and could precede speciation (Howard 1999). Alternatively, once a hybrid zygote is formed, it may have low viability or be sterile. Mating between a horse and a donkey, for example, produces mules, which are sterile. Thus genes from the hybrid mating are not transferred beyond one generation. Empirically, the sex for which survival or fertility is more affected by hybridization tends to be the heterogametic sex (or the sex possessing dissimilar sex chromosomes, commonly called 'x' and

'y' chromosomes), in a phenomenon labeled 'Haldane's Rule' (Haldane 1922). This suggests that genes affecting hybrid sterility might be localized to the sex chromosomes. However, the genetics underlying hybrid sterility have been explored in *Drosophila*, which does show stronger male than female sterility in hybrid crosses. Genes affecting either male or female sterility are spread throughout the genome, and the number involved in male sterility alone sometimes exceeds 100 genes (summarized by Wu and Hollocher 1998).

Genetic barriers may also arise through changes in the number of chromosomes in new species (Ramsey and Schemske 1998). This is particularly common in plants. Gametes (in this case, pollen and ova) generally have a single copy of each chromosome (termed haploid), uniting to give a zygote with two copies of each chromosome (termed diploid). In the case of hybrids between two species, each of the two copies will come from different species, and gamete formation may be impeded because the cells cannot properly divide. Occasionally, however, errors in cell division occur, and diploid gametes are produced. If these form within hybrid individuals, then the resulting tetraploid zygote is completely fertile, as it now contains a complete set of the chromosomes from each of the original parent species. Subsequent gamete formation can proceed normally and the offspring are interfertile, but cannot mate successfully with the parental species. Such polyploidy has been demonstrated for example in the history of both agricultural tobacco and wheat. Other sorts of errors can produce individuals with other chromosome copy numbers; anything greater than diploid is generally called polyploid. No matter what their origin, if polyploid individuals are viable, the gametes they produce cannot usually produce viable offspring by uniting with gametes from their diploid parents.

Cytoplasmic incompatibility occurs when factors in the cytoplasm of the two gametes are not compatible, resulting in the death of the zygote. *Wolbachia* bacteria in insects, which are passed with the cytoplasm from mother to offspring, are one example. Currently, it is believed that males infected with *Wolbachia* produce sperm which must be 'rescued' by a factor in eggs produced by females that are also infected, before viable zygotes are produced. Whether this mechanism has been involved in speciation in the past, or is 'simply' a factor enhancing reproductive isolation at present is not yet established (see Werren 1998 for a review).

2.3 *Differentiation: Selection and Chance*

Both the development of reproductive isolation and the differentiation of other characters identifying organisms as members of a species could be influenced by selection or by genetic drift. In the former case, adaptation to a particular set of environmental conditions drives speciation. In the latter case, chance events, such as the genetic diversity present in individuals that become geographically isolated, or that happens to be included when a polyploid individual is formed, influence the results of speciation. The relative influence of these two forces has been the subject of debate; like many polarized debates in biology, most cases probably fall along the continuum between the two ends.

3. *Biogeography of the Speciation Process*

Evolutionary biologists have historically categorized speciation based on the geographic relationship of the newly evolved species. This classification reflects differences of opinion as to the relative role of geography in reducing gene exchange between putative species. Nonetheless, the actual evolutionary processes underlying speciation may be the same in each of these types of speciation.

3.1 *Allopatric Speciation*

'Allopatric' derives from 'allo' meaning 'separate' and 'patria' meaning 'country.' Allopatric speciation is thus speciation that occurs through geographic isolation of two populations or groups of populations. In this scenario, a population is split by some geographic barrier, such as a mountain range or river for terrestrial organisms, or a land mass for aquatic organisms. Over time, individuals in populations on each side of the barrier differentiate. This may be due to a founder effect, in which only part of the genetic variability originally present is represented in a new population; to chance loss of genetic variation through accidents of who happens to reproduce or not; or to adaptation to environmental differences on each side of the barrier. This differentiation between the divided populations is extensive enough to include the evolution of differences in reproductive biology, such that when the two populations are re-united by the removal of the barrier, they are no longer able to inter-breed and are viewed as separate species.

Support for the allopatric speciation model increased with the discovery of continental drift and plate tectonics, and the subsequent increased understanding of geologic history. Such continental rearrangement at a deep time level, and geographic changes associated with Pleistocene ice ages in more recent times, provide the geographic variation needed for allopatric speciation. Enhanced understanding of geologic events also provided the opportunity to test expected distributions of species under allopatric speciation through the study of vicariance biogeography, or the geographic distribution of species as a result of subdivision of ancestral species' ranges.

Allopatric speciation has been the dominant geographic model for speciation processes since around 1950, although other models are currently gaining support. This dominance derived from the argument

that development of genetic isolation would most likely occur if incipient species were physically isolated from each other for a time period, allowing the evolution of reproductive isolating mechanisms.

3.2 Parapatric Speciation

'Parapatric' derives from 'para' meaning 'near' and 'patria' meaning 'country.' Parapatric speciation thus occurs when a smaller population is isolated, usually at the periphery of a larger group, and becomes differentiated to the point of becoming a new species. Gene flow may remain possible between the two populations during the speciation process, and hybrid zones may be observed at the interface between the two populations as a result.

The best-known example of incipient parapatric speciation occurs in populations of the grass *Agrostis tenuis* which span mine tailings and normal soils. Individuals that are tolerant to heavy metals, a heritable trait, survive well on contaminated soil, but poorly on non-contaminated soil. The reverse occurs for intolerant populations. Gene flow occurs between sub-populations on and off mine tailings, but hybridization is inhibited by slight differences in flowering time between the two locations (McNeilly and Antonovics 1968, McNeilly and Bradshaw 1968).

3.3 Sympatric Speciation

'Sympatric' derives from 'sym' meaning 'together with' and 'patria' meaning 'country.' Sympatric speciation thus refers to speciation occurring within a single population. Because of the lack of geographic isolation, sympatric speciation was historically regarded by many biologists as being impossible. However, evidence is accumulating that sympatric speciation does indeed occur.

The most obvious mechanism that can result in sympatric speciation is speciation through polyploidy (Ramsey and Schemske 1998). A chance doubling of the chromosome number can lead to reproductive isolation through zygote inviability, without a preceding need for geographic isolation. Given that about half of all flowering plants are polyploid, this form of speciation may have played a large role in their formation.

Hybridization between two species, followed by rapid changes in the hybrid to enhance reproductive isolation from the parents has been held up as a form of sympatric speciation, particularly in plants. While the number of solid examples of this speciation mechanism are few, it does occur (Dowling and Secor 1997, Rieseberg 1997).

Genetic introgression through mechanisms other than hybridization, or genetic rearrangements ('jumping genes') are a potential source of variability within a population that could lead to speciation, but are essentially unexplored at present.

Small shifts in the ecological niche of individuals within a population are also argued to result in a form of sympatric speciation. The host races of *Rhagoletis* described above are a case in point. Another example is the radiation of species of cichlid fish within small crater lakes in Cameroon (Schliewen et al. 1994). Genetic analysis indicates that the species in each lake derive form a single ancestor. Although species differ slightly in habitat preference, the lakes are small enough that the different fish species are sympatric. In order for this mechanism to be effective, natural and sexual selection would need to act in the same direction. For example, if individuals diverge in resource use, and individuals with similar resource use prefer each other as mates, resource use races will evolve, which may eventually become full species. Note that this form of speciation has conceptual affinities both to the biological species concept (reproductive isolation) and to the recognition species concept (species cohesion).

4. Space: Patterns of Biodiversity Resulting From Speciation

The different modes of speciation outlined above each result in a different spatial pattern of biodiversity. Hybrid zones between newly formed species are the predicted outcome of allopatric speciation, if the geographically isolated populations have not developed full reproductive isolation prior to removal of the geographic barrier. Further, the establishment of a geographic barrier is expected to divide ancestral populations of many species within the flora and fauna of a region. Removal of that barrier and recontact of newly divergent flora and fauna can result in multiple individual hybrid zones that overlap, or suture zones. Such suture zones mark boundaries between regional floral and faunal assemblages. Remington (1968) outlined six major suture zones and a number of minor suture zones in the Nearctic region, including a compendium of known hybridizing pairs of species within each suture zone and an analysis of the likely geographic barriers whose removal resulted in the now-observable suture zone pattern.

Ring species also result from the interaction of speciation with a geographic barrier, but with a twist. As populations of a species expand into new areas, moving around a geographic barrier, they can form a ring around that barrier. While the immediately adjacent populations at each point around the ring interbreed, the terminal populations are reproductively isolated and cannot interbreed. Ring species thus can be thought of as a result of geographically extended parapatric speciation. Irwin et al. (2001) present an outstanding example of a ring species in the case of the greenish warbler (*Phylloscopus trochi-*

loides), whose six sub-species circle the Tibetan Plateau. The sub-species that are sympatric at the ends of the ring in central Siberia do not interbreed. The structure of songs involved in mate choice have diverged due to selection pressures favoring increased song complexity as the species expanded from the southwest around on the eastern and western sides of the plateau, eventually contacting again north of the plateau. Interestingly, there is now a gap in the ring in central China, which is probably due to habitat destruction by humans. The recentness of this gap has not allowed time for song divergence across the gap, but one could predict speciation across the gap in the future.

Sympatric speciation can generate a different pattern of biodiversity, hot spots of endemic species. Endemic species have very localized distributions, being found in one (usually small) area. Such species are not necessarily rare; the localized populations may be quite large. The cichlid fish species assemblages found in the crater lakes of Cameroon and cited above, along with the much richer cichlid diversity in the rift lakes of east Africa, are classic examples of hot spots of endemic species. In Lake Victoria, for example, over 300 species are found, of which very few are found anywhere else.

Overall biogeographic patterns of diversity within a region should therefore reflect the relative dominance of allopatric, parapatric, and sympatric speciation in shaping the flora and fauna. However, these biogeographic patterns are all too easily disturbed by human activities, which include habitat destruction resulting in species extinction and species transport resulting in new mixing of floras and faunas. This disturbance lends urgency to attempts to understand the causes of biogeographic patterns at the beginning of the twenty-first century, before the evidence is muddied by human activity.

5. *Time: Speciation Speed*

Based on the mechanisms of speciation outlined in Sects. 2 and 3, it is clear that speciation takes variable amounts of time to occur. Speciation via polyploidy, for example, can be nearly instantaneous. Speciation via changes in sexual signals involved in mate choice may be a much more gradual event.

From a lineage perspective, there are two models of speciation rates. First, speciation may occur through gradual change within the lineage, such that over time species 'a' is transformed into species 'b.' Second, lineages may split by either budding off a new lineage while the ancestral lineage remains unchanged, or by branching into two new lineages with the loss of the ancestral lineage. Thus, in the first model, change is slow; in the second, speciation events are rapid and interspersed among periods of no change or stasis. The first model has been dubbed 'gradualism' and the second, 'punctuated equilibrium' (Eldridge and Gould 1972). Large debates about the relative importance of gradualism and punctuated equilibrium raged among paleontologists throughout the 1970s and 1980s. The debate had a bearing on the role of natural selection and adaptation in speciation, as contrasted with genetic drift or chance founder effects. Supporters of the gradualist position usually argued that natural selection is of overwhelming importance, while supporters of punctuated equilibrium, initially at least, postulated a strong role for founder effects resulting in sub-sampling of the variation present in an ancestral species in a newly geographically isolated population, enhancing the speed of evolution of new species in those isolated populations. Support for both gradualism and punctuation exists in the fossil record, making resolution of the debate difficult. In addition, the time scale of the fossil record is very coarse, making it difficult to determine exactly how saltatorial a lineage might be.

See also: Adaptation, Fitness, and Evolution; Evolution, History of; Evolutionary Selection, Levels of: Group versus Individual; Evolutionary Theory, Structure of; Extinction of Species; Natural Selection; Primates, Evolution of

Bibliography

Bush G L 1998 The conceptual radicalization of an evolutionary biologist. In: Howard D J, Berlocher S H (eds.) *Endless Forms. Species and Speciation*. Oxford University Press, New York

Cracraft J 1983 Species concepts and speciation analysis. *Current Ornithology* **1**: 159–87

Dowling T E, Secor C L 1997 The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics* **28**: 593–619

Eldridge N, Gould S J 1972 Punctuated equilibrium: An alternative to phyletic gradualism. In: Schopf T J M (ed.) *Models in Paleontology*. Freeman, Cooper, San Francisco

Haldane J B S 1922 Sex ratio and unisexual sterility in animals. *Journal of Genetics* **12**: 101–9

Hennig W 1966 *Phylogenetic Systematics*. University of Illinois Press, Urbana, IL

Howard D J 1999 Conspecific sperm and pollen precedence and speciation. *Annual Review of Ecology and Systematics* **30**: 109–32

Howard D J, Berlocher S H (eds.) 1998 *Endless Forms. Species and Speciation*. Oxford University Press, New York

Irwin, D E, Bensch S, Price T D 2001 Speciation in a ring. *Nature* **409**: 333–7

Mallet J 1995 A species definition for the modern synthesis. *Trends in Ecology and Evolution* **10**: 294–9

Mayr E 1942 *Systematics and the Origin of Species*. Columbia University Press, New York

McNeilly T, Antonovics J 1968 Evolution in closely adjacent plant populations. IV. Barriers to gene flow. *Heredity* **23**: 205–18

McNeilly T, Bradshaw A D 1968 Evolutionary processes in populations of copper tolerant *Agrostis tenuis* Sibth. *Evolution* **22**: 108–18

Otte D, Endler J A (eds.) 1989 *Speciation and its Consequences*. Sinauer Associates, Sunderland, MA
Paterson H E H 1985 The recognition concept of species. In: Vrba E S (ed.) *Species and Speciation*. Transvaal Museum, Pretoria, South Africa
Ramsey J, Schemske D W 1998 Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics* **29**: 467–501
Remington C L 1968 Suture-zones of hybrid interaction between recently joined biotas. In: Dobzhansky T, Hecht M K, Steer W C (ed.) *Evolutionary Biology* **2**: 321–428
Rieseberg L H 1997 Hybrid origins of plant species. *Annual Review of Ecology and Systematics* **28**: 359–89
Schliewen U K, Tautz D, Paabo S 1994 Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature* **368**: 629–32
Simpson G G 1951 The species concept. *Evolution* **5**: 285–98
Sokal R R, Crovello T J 1970 The biological species concept: A critical evaluation. *American Naturalist* **104**: 127–53
Templeton A R 1989 The meaning of species and speciation. A genetic perspective. In: Otte D, Endler J A (eds.) *Speciation and its Consequences*. Sinauer Associates, Sunderland, MA
Werren J H 1998 *Wolbachia* and speciation. In: Howard D J, Berlocher S H (eds.) *Endless Forms. Species and Speciation*. Oxford University Press, New York
Wiley E O 1978 The evolutionary species concept reconsidered. *Systematic Zoology* **27**: 17–26
Wu C-I, Hollocher H 1998 Subtle is nature. The genetics of species differentiation and speciation. In: Howard D J, Berlocher S H (eds.) *Endless Forms. Species and Speciation*. Oxford University Press, New York

C. L. Boggs

Spectral Analysis

Spectral analysis is one of several statistical techniques necessary for characterizing and analyzing sequenced data. Sequenced data are observations that have been taken in one, two, or three dimensional space, and/or time. Examples might be observations of population density along a road, or of rainfall over an area, or of daily births at a hospital. One important limitation is that the observations be equally spaced in order that the analysis proceed efficiently. Spectral analysis refers to the decomposition of a sequence into oscillations of different lengths or scales. By this process, the observations in what is called the data domain are converted into the spectral domain. The reasons for doing this are that: (a) some forms of manipulation are easier in the spectral domain; and (b) the revealed scales are necessary statistical descriptors of the data and may suggest important factors that affect or produce such data. The following will provide brief descriptions of: (a) Fourier analysis and its use in manipulating data that are assumed to be periodic; (b) relevant statistics; and (c) one approach to spectral analysis of nonperiodic data including an example.

1. Fourier or Harmonic Analysis

Here data are represented by y where the position in the sequence is labeled by j, i.e., $y[j]$, where j varies from 0 to $(n-1)$. Because trigonometric functions require radians for their calculation the label j is converted to radians with $2\pi j/n$. What Fourier showed in 1822 was that almost any equally spaced sequence $y[j]$ could be fitted exactly by a set of cosine and sine curves. In addition, since no information was lost, these could be returned to the original data. There was one important assumption, that the data were periodic, i.e., that they repeated themselves both forwards from $(n-1)$ and backwards from 0 indefinitely.

A sinusoidal curve,

$$y[j] = A[k]\cos\{(2\pi kj)/n - \Phi[k]\},$$

as displayed by curve (a) in Fig. 1 is characterized by its frequency label, k, its amplitude, $A[k]$, and phase, $\Phi[k]$. The frequency, k, is the number of complete oscillations that occur over the total observations, n. For curve (a) this is 2. For n observations k may have integer values of 0 to $+n/2$ (n even). Frequency may also be expressed in observational units as $f = k/n\Delta t$ where Δt is the interval between observations. In the example, if y is observed daily for a year, the oscillation at $k = 2$ has a frequency of $f = 2/(365 \times 1 \text{ day}) = 1/182.5$th of a cycle per day $= 2$ cycles per year. It also has a wavelength (peak to peak distance) of $1/f = 182.5$ days $= 1/2$ year. The highest frequency that may be calculated is $f_{\max} = 1/(2\Delta t)$. The amplitude, $A[2]$, of curve (a) is 0.5. This also provides the variance of the curve, which is half the square of its amplitude, $A^2[2]/2 = 0.125$. The phase $\Phi[k]$, when it is divided by k to give the phase shift, gives the distance of the first peak from the origin. In this case $\Phi[2]/2 = 0.46$. Curve (a) can also be represented by the sum of a simple cosine curve ($y[j] = a[k]\cos(2\pi kj)/n$) as displayed as curve (b) and a simple sine curve ($y[j] = b[k]\sin(2\pi kj)/n$) given by curve (c). From trigonometry we can show that $A[k] = (a^2[k] + b^2[k])^{1/2}$, and $\Phi[k] =$

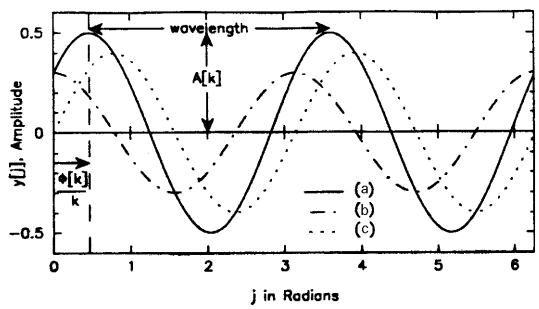

Figure 1
Examples of (a) sinusoidal, (b) simple cosine, and (c) simple sine curves for periodic data

arctan($b[k]/a[k]$). The coefficients $a[k]$ and $b[k]$ may be obtained from

$$a[k] = \frac{2}{n}\sum_{j=0}^{n-1} y[j]\cos\left(\frac{2\pi kj}{n}\right), \tag{1}$$

$$b[k] = \frac{2}{n}\sum_{j=0}^{n-1} y[j]\sin\left(\frac{2\pi kj}{n}\right), \tag{2}$$

where, for $k = 0$ and for $k = n/2$ when n is even, only Eqn. (1) exists and the factor 2 is replaced by 1. $a[0]$ is the mean of the series. A real set of data will be represented by the sum of the whole range of possible frequencies,

$$y[j] = \sum_{k=0}^{n/2}\left[a[k]\cos\left(\frac{2\pi kj}{n}\right) + b[k]\sin\left(\frac{2\pi kj}{n}\right)\right]. \tag{3}$$

Note that there were n $y[j]$ and a total of n $a[k]$ and $b[k]$. Thus the information present in the data domain, $y[j]$, is equivalent to the information in the spectral domain, $a[k]$ and $b[k]$. It is just arranged differently. A variable arranged according to k is said to be the spectrum of that variable. As an example, Fig. 4 represents the spectrum of the percentage variance.

All of these formulas refer to one dimension (along a road or through time). Similar equations may be written for spatial observations. It is nothing more than a curve- or surface-fitting procedure, like fitting a polynomial to observed data. One problem to be considered is how well the beginning of the data match the end of the data. If the beginning and ending points have very different magnitudes, the periodic assumption causes a big jump at those points. As a result, many frequencies will be required to account for it and these amplitudes may overwhelm the amplitudes within the data. Consequently, it is advisable to smooth off the ends of the data so that they meet at approximately the same magnitude. By this technique some information about the beginning and the end will be unavoidably lost.

What the Fourier series fitting does is to estimate any truly periodic elements present. In addition, it permits analytical procedures to be applied to the data even if they are not periodic. Often manipulations that are complex in the data domain are very simple in the spectral domain. For example, the degree of slope and azimuth of slope is a simple process of differentiation of Eqn. (3). The amplitudes are calculated by the process previously described, each amplitude is multiplied by its respective frequency, and the new amplitudes converted back to the data domain where they produce slope and orientation (Rayner 1972). As another example, pattern recognition is a simple multiplication process in the spectral domain (Rayner 1971). Yet another use is in the description or summarization of shape. Manipulation of digitized shape data such as national or other boundaries, by these Fourier techniques can reduce the number of original observations to less than one third without significant loss of information (Moellering and Rayner 1982).

2. Statistical Description

One use of statistics is to summarize large amounts of data. For example, Fig. 2 displays one year of daily birth data at one hospital. A first step is usually to calculate its probability distribution. Under ideal conditions that distribution will approach the 'normal probability distribution' or 'Gaussian probability distribution' form. Then the mean (μ_1) and variance (σ^2 or μ_2) are all that are needed to describe the data. Fig. 3 shows the probability distribution of the birth data which have a mean of 8.7 and a variance of 9.5. Superimposed is the normal curve. If the distribution is 'skewed' higher moments than the second moment (variance) will be required. Multimodal distributions will have to be subject to additional processing. In cases where more than one set of observations are being compared or related, 'covariance' and 'regression' techniques are applied. The true or population

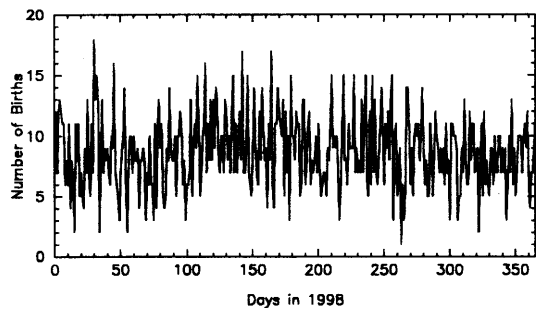

Figure 2
Daily birth data for one hospital during 1998

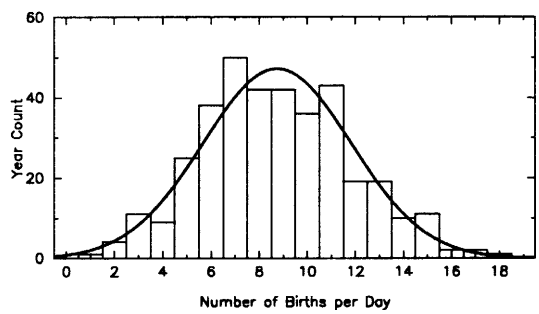

Figure 3
Probability distribution of birth data

characteristics of distributions are called 'parameters.' Since only samples of data are available, only estimates of these characteristics can be obtained. These estimates are called statistics and are usually denoted by 'hatted' symbols ($\hat{\sigma}^2$ or $\hat{\mu}_2$). The particular formula used to calculate a statistic is called an estimator so, for example, different variance estimators are produced by dividing the sums of squares of deviations from the mean either by n or by $(n-1)$, where n is the number of observations in the sample. Two important measures of the quality of an estimator are bias, which is the deviation of the mean magnitude of the estimate from the population magnitude, and the variance of the sample estimates. Unfortunately, estimators that have small bias typically have large variance, and vice versa. Another quality measure is consistency, in which both bias and variance should decrease as n increases. Spectral estimates are typically not consistent.

Statistical estimates are assumed to come from an ensemble of observations. That is, the sample is taken from all possible occurrences of the phenomenon. In fact, typically only one set of observations is available through time or in space. To be useful therefore, the data from the series must be assumed to be representative of the ensemble, the data are assumed to be ergodic data. This also assumes that the particular segment of time series is representative of the whole continuum. Then the statistics from the different segments should be equivalent. In other words, the data are said to be stationary data.

Now to return to the description of observations, if the data are independent and one observation is not affected by the previous or next one, the probability of observing a particular magnitude is dictated by the normal distribution and is not related to the sequence of magnitudes observed. Then the mean and variance are sufficient descriptors. When this assumption is not fulfilled (e.g., the population density one kilometer away from a high density within an urban area is also likely to be high and not given by the normal distribution), this influence may be specified by the autocovariance function (i.e., the covariance of each observation with its next neighbor, with its neighbor two observations away, with its neighbor three observations away, etc.),

$$yy[p] = \frac{1}{n-|p|} \sum_{j=0}^{n-1-|p|} y[j]y[j+p], \tag{4}$$

where p is the neighbor distance, which is often called the *lag*. The $y[j]$ are pre-processed to have a mean of zero. Because the possible number of summed pairs is reduced as the lag increases, there is a practical limit to p which is typically taken to be in the order of $n/10 = m$. Note that this is a symmetrical function around $p = 0$ having $2m$ terms, and it is not necessary to calculate for negative p. Eqn (4) produces a new set of numbers that typically have a maximum at $p = 0$, decline to zero, and then oscillate around zero. How may this function be interpreted? Because, like the original data, these numbers are not independent, that is not easy. However, if this sequence is transformed with an orthogonal function, the new sequence, which is independent, is interpretable. The function selected is the Fourier transform which is a limiting form of Fourier analysis. This produces a spectrum of the variance: an allocation of the variance into different scales, wavelengths of frequencies of fluctuations.

In Sect. 1 it was shown that harmonic analysis requires that the data be periodic. The autocovariance function, $yy[p]$, is truncated at $(m-1)$ and $-(m-1)$ and is certainly not periodic. The $yy[p]$ are really all that can be calculated from what may be seen through a blurred window of the original observations that theoretically extend from $-\infty$ to $+\infty$. How can this window be described, or what can be said about what is not seen? The oldest approach, also known as the classical approach (Priestley 1997), assumes that the ends of the series are smoothed off and the unseen $y[j]$ are zero. Alternative approaches have searched for other estimators which use knowledge of the data to obtain, for example, more realistic models of what is unseen. These use autoregression, moving average, and maximum entropy techniques, individually and in combination. They occupy a large amount of literature and all cannot be adequately summarized here where the classical technique will be used to illustrate the basic principles. For the other approaches reference should be made to books by Box and Jenkins (1976) and Kay (1988) which contains computer programs for empirical work.

3. *Classical Spectral Analysis*

As zeroes are added to the series, the number of frequencies that may be calculated is increased but their magnitudes are decreased. For example, if n is doubled the label for an equivalent frequency is doubled and its magnitude is halved. What this shows is that the individual amplitudes, which refer to integer k frequencies in harmonic analysis, become average amplitudes over a frequency band centered at k as $n \to \infty$. To distinguish them, the frequency counter is changed from k to f. The practical equations essentially stay the same but the interpretation changes. Also, now a and b are typically defined for two-sided functions with $a[f] = a[-f] = (1/2)a[k]$ and $b[f] = -b[-f] = (1/2)\ b[k]$, and Eqns. (1) and (2) become

$$a[f] = \frac{1}{n} \sum_{j=0}^{n-1} \left[y[j] \cos\left(\frac{2\pi kj}{n}\right) \right], \tag{5}$$

$$b[f] = \frac{1}{n} \sum_{j=0}^{n-1} \left[y[j] \sin\left(\frac{2\pi kj}{n}\right) \right]. \tag{6}$$

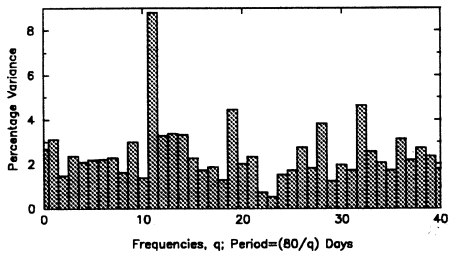

Figure 4
Spectrum of percentage variance of birth data

The calculation with the new interpretation of a and b is called a Fourier transform.

Because the autocovariance function is a symmetrical function only Eqn. (5) is required for its transformation,

$$YY[f] = \frac{1}{2m} \sum_{p=-(m-1)}^{m-1} \left[yy[p] \cos\left(\frac{2\pi kp}{m}\right)\right]. \qquad (7)$$

This still leaves the shape of the window to be taken into account. Various windows have been proposed. Their effect in the spectral domain is to perform a weighting average over adjacent frequencies, i.e.,

$$\text{ave}(YY[f_1]) = \sum_{-(m-1)}^{+(m-1)} YY[f]H[f_1 - f], \qquad (8)$$

where, for the 'hanning' window, $H[f]$ is equal to 1/4, 1/2, 1/4.

The above process is the same as a Fourier transform of the original $y[j]$ series that has had the mean removed, been windowed by the smoothing off of the ends, and had some zeroes added. Then, for the variance, the $(a^2[f] + b^2[f])$'s are summed over equally-sized groups of f values.

As an example of such an analysis the spectrum of the percentage variance of the birth data is given in Fig. 4 with $m = 40$ and $f = q/(2m\Delta t)$ where $0 \leqslant q \leqslant m$. One might expect the number of births to be random in time, in which case the spectrum should be flat (approximately constant variance over all frequencies). This is essentially what Fig. 4 reveals with one exception: a large peak at $q = 11$, at approximately one week wavelength. One may speculate on the cause.

The application of other techniques of spectral analysis, such as maximum likelihood, might tease out other interesting periods. The next step might be a regression analysis relating births to other factors. If the other factors are sequenced, then this calls for cross spectral analysis (Rayner 1971). It allocates the variance to other factors as a function of scale. It also provides an estimate at each frequency of any delay between births and the influencing factor. Of course, statistical techniques cannot assign a cause. They can only suggest interesting phenomena and apparent relationships for investigation by other means.

For more complex distributions, where higher statistical moments and products are required, higher order spectra, polyspectra, provide valuable insights (Rayner in press).

See also: Sequential Statistical Methods; Spatial Autocorrelation; Spatial Choice Models; Spatial Interaction Models; Spatial Pattern, Analysis of; Spatial Sampling; Spatial Search Models; Spatial Statistical Methods; Spatial-temporal Modeling; Statistical Pattern Recognition; Statistical Systems: Censuses of Population; Statistical Systems: Crime; Time Series: Advanced Methods; Time Series: Cycles; Time-use: Research Methods

Bibliography

Box G E P, Jenkins G M 1976 *Time Series Analysis, Forecasting and Control*, rev. edn. Holden-Day, San Francisco, CA

Kay S M 1988 *Modern Spectral Estimation: Theory and Application*. Prentice-Hall, Englewood Cliffs, NJ

Moellering H, Rayner J N 1982 The dual axis Fourier shape analysis of closed cartographic forms. *The Cartographic Journal* **19**: 53–89

Priestley M B 1997 A short history of time series. In: Rao T S, Priestley M B, Lessi O (eds.) *Applications of Time Series in Astronomy and Meteorology*, 1st edn. Chapman and Hall, London, pp. 3–23

Rayner J N 1971 *An Introduction to Spectral Analysis*. Pion, London

Rayner J N 1972 The application of harmonic and spectral analysis to the study of terrain. In: Chorley R J (ed.) *Spatial Analysis in Geomorphology*. Methuen, London, pp. 283–302

Rayner J N (in press) *The Basis of Dynamic Climatology*. Blackwell, Oxford, UK

J. N. Rayner

Speech Errors, Psychology of

The generation of spoken language is a complex task normally accomplished with high accuracy. Human beings are, on average, consummate 'talkers.' But, speakers do err occasionally, and in ways that reflect the mental processes underlying the act of speaking. That is the topic pursued here: What are the speech error patterns of normal language use, and what do they tell us about language processing?

Various departures from the ideal of perfectly articulated, fluent speech occur in ordinary conversation: hesitations, false starts, changes of mind in mid-sentence, and misarticulations. In this context, the term 'speech error' has come to be used in a semitechnical fashion to pick out particular errors

from the welter of everyday language. The focus is on fluently uttered forms in which some element or elements of the intended target utterance are missing, mis-selected or misordered. Errors of knowledge (i.e., those arising from ignorance of a word or of 'proper usage') and various nonfluencies are distinguished from speech error in this way of proceeding.

1. *A Methodological Note*

How is error data acquired? Several investigators have collected extensive records of errors heard during everyday language use. A classic such effort was that of Rudolf Merringer (Merringer and Meyer 1895) in the late nineteenth century. Cutler and Fay (1978) describe his methods in their introductory essay to a reissue of Merringer's work. The practices they describe are characteristic of later diary studies. Typically, a written record of an error is made soon after observation (normally, a few seconds to a few minutes), with annotations regarding pronunciation and setting. Such observation has the strong appeal of face validity for processes of language formulation. However, influences of perceptual filtering and memory distortion must be carefully weighed. It is clear that minor phonetic deviations may be missed, and only changes clearly discernable with respect to the speaker's target are well represented in such error corpora. See Cutler (1982) for detailed discussion of these and related issues.

Diary studies exist for several languages. It may be said with some assurance that whether the target language is Chinese, Dutch, English, German, Italian, Spanish, Turkish, or other, the same general types of phenomena arise. Specific language features may afford options for error —e.g., Chinese tone errors— not available in English or related languages. Nevertheless, the general behavior of errors involves similar mechanisms across languages. Discussion here focuses on English, with some commentary on representative language differences. Error examples are taken from papers in the bibliography.

2. *Types of Error*

In the popular literature, 'speech error' is synonomous with 'Spoonerisms', named after the Dean of New College, Oxford, Rev. W. A. Spooner (1844–1930), who was notorious for his misplacements of speech sounds, sometimes to extraordinary effect. For example, in lecturing a wayward student, he is alleged to have said: 'You have hissed all my mystery lectures. ... and tasted two worms' ('missed all my history lectures ... wasted two terms'). But Spoonerisms are special in several respects. They suggest that sound errors usually make words, and that they make a kind of wild and humorous sense. Neither condition is typical. Further, Spoonerisms primarily exploit sound error, and the relevant phenomena are much more general. No account of Rev. Spooner's performance is given here; for comment, see, e.g., Robbins (1966).

Notwithstanding the exclusion of Rev. Spooner's legacy, there is no denying the charm of suitably chosen errors as a way to display the hidden structure in language. The errors are intricate, and they illustrate essential features of language processing. Examples (a)–(n) are representative.

(a) '... sat down on a *s*ot *h*oddering iron' (... hot soldering ...)

(b) 'I won't hit the ba*rd* ha*ll*' (... ball hard)

(c) 'I have a case of l*ay* gr*o*de flu.' (... low grade ...)

(d) '... so I can start the s*tape*back up (... the tape ...)

(e) 'The curren*th* month is ... (The current ...)

(f) '..an anguage *l*acquisition device ... (... a language acquisition...)

(g) 'Is your *purr*y *kitt*ing' (... kitty purring)

(h) 'Make it so the *apple* has less *tree*s. (... so the tree has less apples.)

(i) 'She *sings* everything she *writes*. (... writes everything she sings)

(j) 'Bensinger takes to that left-field wall like *water* takes to *a duck*. (... like a duck takes to water)

(k) 'I've got whipped cream on my *mushroom*! (... on my mustache!)

(l) 'You put your *nose* ... your finger on it. (... put your finger on it)

(m) 'My! So that's the Porche *symblem*. (symbol/emblem)

(n) '*It doesn't care*. (It doesn't matter/I don't care)

A rough task analysis sets the stage for error interpretation. It is evident that systems for speaking include mechanisms that select from tens of thousands of available words just those appropriate to an intended message ('lexical retrieval'), mechanisms that integrate the words into sentence structures ('syntactic construction'), and mechanisms that orchestrate these decisions into control structures ('phonological and prosodic integration) that drive motor outputs. Psychological explanation must reconstruct this complex of mental events.

Descriptively, in examples (a)–(g), sounds move around in various ways—they are exchanged, anticipated, perseverated, and shifted. Important features of sound error include: Moved elements are *phonetically similar*, they correspond in *syllabic position*, and their immediate *phonetic environments are similar* (Boomer and Laver 1968, Dell 1984, Noteboom 1969, Caldognetti et al. 1987, del Viso et al. 1991, McKay 1970, Vousen et al. 2000). The fact that error likelihood is so strongly influenced by structural similarity suggests that errors arise, in part, out of mechanisms that associate independently specified processing elements (sounds, in this example) with a *planning framework* ('syllable or word frames', in this

instance). These errors and related phenomena indicate how sounds that make up words enter planning structures for sentences. In the task analysis, this reflects phonological and prosodic integration processes.

Examples (h)–(j) are 'word movement' errors. Some are of special interest because they reflect an analysis of words into stem and affix. In Example (h), 'apples' is decomposed into 'apple + s' —more abstractly ['apple' + plural marker]. The technical term in linguistic science for such a decomposition is *morphemic analysis*. Notice that the two types of morpheme (stem and affix) behave in very different ways. The stem moves, but the affix stays behind. One can also describe this error in terms of the notion of structural frame. By this hypothesis, the affix would be considered part of a phrasal frame that stem morphemes are assigned to as a sentence develops. This idea proves applicable to a wide range of error observations, and is strikingly manifest in languages with richly developed morphological structures. (Related facts are discussed in Sect. 3.4.) In the task analysis, these errors implicate syntactic integration.

Examples (k) and (l) provide another perspective on errors. Descriptively, these are *word substitutions*. Several points are relevant. First, (k) displays a phonetic similarity between target and intruding element; the phonetic similarities between 'mustache' and 'mushroom' are representative. Moreover, both words are nouns. Thus, the syntactic structure of the intended utterance is preserved. One may plausibly treat this as a failure—perhaps triggered by phonetic similarity—of mechanisms that retrieve elements from an inventory of word forms. But what about errors like (l)? Grammatical category is preserved, but phonetic similarity is not apparent. And, unlike (k), an obvious meaning relation holds between target and intrusion. Section 3.5 discusses this contrast. In the task analysis, these errors implicate lexical selection processes, and they encompass both form and meaning relations.

Examples 13 and 14 are *blend* errors. In morpheme, word, and phrase blends, two processing elements compete for a *single output slot*. Speakers are often aware of the competition, and can report experiencing it. The competing expressions are, with rare exception, of the same grammatical class and are semantically equivalent for the intended message. This phenomenon is a *prima facie* demonstration of some parallelism in sentence production. Such errors implicate lexical selection processes and aspects of phrasal and phonological composition.

3. *Interpretations of Error Phenomena*

Speech errors have many variants. The challenge is to explain their similarities and differences in ways that link them to accounts of production mechanisms at the necessary levels of representation (phrases, words, morphemes, syllables, sounds). A robust generalization across the error types is relevant: Like elements interact, i.e., sounds interact with sounds, words with words, and phrases with phrases. These interactions indicate vocabularies of 'units', and the error data shows they are linked in an architecture of processing stages discussed in Sects. 3.1–3.4.

3.1 *Accommodations in Error Processes*

Errors move elements of the intended utterance to 'new' positions. Because languages display many local dependencies of form, such displacements could violate these. But, typically they do not. Example (f) illustrates: '*a* language acquisition device' became '*an* anguage lacquisition device.' The indefinite article changed from 'a' to 'an', and so *accommodated* to the vowel onset ('anguage') created by the error. In Example (o), the bound morpheme deriving 'easily' from 'easy' shifts. But, the isolated stem 'easy' was correctly pronounced as /izi/ rather than with a reduced final vowel (as would have been appropriate had the affix remained in place).

(o) '... *easy enoughly* ...' (easily enough)

Many such variations triggered by changes in phonetic context appear in accommodations. By these observations, processes that fix the surface phonetic forms of words are ordered later than the error-generating structures: First, an abstract morphological and phonological representation is constructed, then regular processes flesh out its phonetic detail. This is a familiar notion in linguistics, and its expression in error phenomena is not greatly surprising (see Fromkin 1971). But, bear in mind that it could have turned out otherwise. The speaker of Example (f) might have said 'a anguage' but did not; the speaker of Example (o) might have said 'easuh enoughly', but did not; and so on. Error processes on this evidence are tightly bound to the structural systems that govern error-free speech.

3.2 *Other Evidence for Ordering of Processes: Sound Exchanges and Word Exchanges*

These error classes reveal syntactic and phonological planning structures. Sound interactions, and particularly exchanges as in (a)–(c), typically involve words *in the same phrase*. The relevant structure is prosodic (see Boomer and Laver 1968, Fromkin 1971, Garrett 1980). Prosodic grouping reflects syntax, but is also motivated by intonation and the phonological structure of the words chosen to express a given syntactic form. Treating the domain of sound exchange as prosodic may also explain some cross-language differences. For example, sound exchanges in English are predominantly between words, but in Spanish, within words (del Viso et al. 1992). The greater morphological complexity and length of Spanish words, as compared to English, would require exchanges be-

tween words to cross prosodic boundaries more often for the former than the latter.

Word and phrase exchanges sharply differ from sound exchanges. As Examples (i), (j), and (p)–(r) suggest, they arise for elements in distinct phrases that correspond in grammatical category.

(p) '... I left my *briefcase* in the *cigar*.' (... my cigar in the briefcase)

(q) 'What we want to do is train *its tongue* to move *the cat*.' (... the cat to move its tongue)

(r) 'As you *reap*, Roger, so shall you *sow*.' (As you sow ... so shall you reap)

Nouns exchange with nouns, verbs with verbs, adjectives with adjectives (Bierwisch 1970, Garrett 1975, Nooteboom 1967). On this evidence, operations at the point of error involve the grammatical roles of words (or minimal phrasal units). Sound exchanges differ in both respects. Recall also that sound exchanges implicate many detailed features of phonological structure; word exchanges are minimally so affected (see Dell and Reich 1981, del Viso et al. 1992). These differences in the role of phrasal domain, grammatical category, and phonological description reveal two processing systems. Syntactic constituents are built by a multiphrasal processor that does not represent the sound structure of its elements (word stems, morphemes, simple phrases); failures yield word and phrase exchanges. A subsequent process elaborates sound sequences for words within a phrase group; failures yield sound exchanges. Multiple phrasal units are present at the syntactic level, and a narrower planning window develops sound structures and pronunciation codes in prosodic groups.

3.3 Stranding Errors and Multiple Levels of Processing

The exchanges in Examples (g) and (h) decompose target words into their constituent morphemes and strand affixes. One finds stranding of a variety of productive structural markers, predominantly those of inflectional device (e.g., number, tense, and aspect markers, gerundives, comparatives). Evidence indicates that stranding arises at two points: within phrase (re prosodic groups for sound exchanges), and between phrase (re structures for word exchanges) (see Garrett 1993). These ideas extend the earlier discussion of accommodation. That discussion described it as a phonologically conditioned response to error movements. But, consider Examples (s) and (t).

(s) '··· that I would *hear* one if I *knew* it.' (...that I would know one if I heard it)

(t) '··· drei *Wochen* im *Tag* gearbeitet.' (drei Tage in der Woche gearbeitet).

Here there is accommodation, but, strikingly, it involves irregular forms of inflected words. In Example(s), the appropriate irregular verb forms for 'know' and 'hear' appear. The German example (Bierwisch 1970) shows several accommodations. Singular '*Woche*' takes plural marker 'en'; plural '*Tage*' has the singular form '*Tag*'. Further, German nouns bear grammatical gender: '*Woche*' is feminine, '*Tag*' is masculine. The articles required thus differ, and may interact with prepositions. Here, number and gender are appropriate, including interaction with the preposition ('*im*' is a contracted form of '*in dem*').

What accounts for this impressive range of accommodations? They are lexically dependent and so not derivable from phonological environments (as described in Sect. 3.2). In fact, lexically dependent accommodation is restricted to between phrase stranding exchanges for elements of corresponding grammatical category. In the account in Sect. 3.2, such errors occur at the level of abstract phrasal construction, prior to the phonological interpretation of content words. Development of the irregular forms follows from affiliation of the switched stems with stranded inflectional features.

3.4 Generalizing the Observations About Processing Frames

Stranding shows that inflectional elements are separately represented when stems are inserted into phrasal frames. This can be expressed by assuming that inflectional elements are parts of structural frameworks. If exchanges arise when lexical elements are misassigned to structural frames, it follows that parts of frames should not exchange. This expectation is borne out in error corpora; inflectional exchanges are rare (Garrett 1975, Stemberger 1984). Another class of observations (morpheme shifts), not discussed here, may also be connected to this picture (see Garrett 1980).

There is some basis for treating all functional elements both bound and free in this way. The claim is that functional elements (inflections, determiners, complementizers, etc.—the minor grammatical categories) are derivatives of structure and are thus recruited for incorporation into sentences by different mechanisms than those responsible for the selection of content items (nouns, verbs, prepositions, adjectives, adverbs—the major grammatical categories). This contrast has been discussed in the processing literature under the label of 'open class' and 'closed class' vocabularies. It might be construed as a distinction between those elements individually selected under conceptual control ('content elements') and those that are reflexes of the logic and syntax of the sentence (Levelt 1989, Bock and Levelt 1994). See Myers-Scotten and Jakes (2001), for a related account based on 'code-switching' (a bilingual language practice that fluently combines elements from two languages into a single utterance). Note, however, that some accounts of speech error link profiles for these vocabulary types to their relative frequency of occurrence in the language (Stemberger and McWhinney 1986, Dell 1991).

On this view, very high frequency rates, rather than structural roles, governs error behavior of functional elements.

3.5 *Lexical Selection Processes*: *Word Substitutions*

Word substitutions provide a data class distinct from movement error. These reflect failures in lexical retrieval; Examples (u)–(y) illustrate.

(u) 'Are you left-handed? No, I'm *amphibian*, no, *ambivalent*... arghh, ambidextrous!'

(v) 'If you can find a *garlic* around the house, I'll take it.' (gargle)

(w) 'All I want is something for my *shoulders*.' (elbows)

(x) 'He rode his bike to school *tomorrow*.' (yesterday)

(y) 'Yumm, are we having *lobsters* tonight?' (oysters)

A dissociation of errors with meaning similarities and those with form-based similarities is immediately apparent. Fay and Cutler (1977) in a seminal paper described these contrasts and proposed two selection processes, rooted respectively in content (semantic representations) and form (morphological and phonological representations). Depending on which process is compromised, errors should display quite different similarity metrics. Fay and Cutler's analysis, and subsequent studies, revealed multiple systematic form relations between target and error words that lacked a meaning relation—as Examples (k), (u), (v), while errors with meaning relations, as in Examples (m), (w), and (x), lack such phonological resemblances.

These substitutions also interact with sentence structure: Grammatical category correspondence between target and intrusion powerfully constrains word substitutions whether semantically or form driven. This links to the movement error analysis in Sects. 3.1–3.4. That analysis indicates staged processing in which phonologically uninterpreted elements are first associated with syntactic environments, followed by assignment of phonological representations to structural frameworks. Word substitution patterns fit into those processes. The distinction parallels the discussion of lemmas in *Speech Production, Psychology of*. Meaning-guided retrieval selects lexical representations (phonologically uninterpreted: lemmas) for input to the multiphrasal construction process that is the locus of word exchanges. Following this, word form retrieval (mapping from lemmas to word forms) provides inputs to the prosodic groups that are the locus of sound exchange errors.

The ordering is clear, but details of the sequencing less so. The differences between meaning and form-based substitutions have been a matter of sustained debate. A way to put this is to ask whether semantic selection is accomplished without feedback from form to meaning levels. Dell (1986) and others have argued that the process is to some degree interactive, and includes such feedback, yielding semantically related errors that also display a form relation, as illustrated in Example (y). Such errors are a minority, but occur with greater than chance likelihood. See Dell et al. (1997) for a recent review; and Levelt et al. (1999) for contrasting discussion.

4. *What Causes Error?*

Several causal factors have been identified. Broadly speaking, they are either 'noise' or 'competition' based. See Butterworth (1982) for a helpful exposition. He distinguishes causes of error in terms of 'sentence internal' and 'sentence external' processes, with emphasis on 'competing plans' as error sources.

Fluctuations in efficiency of the neural systems that underlie language processes could clearly cause error. But such accounts require explicit linkage to production mechanisms. This step is well represented in work by Dell (1986) and others using network models to simulate error patterns; see also Levelt et al. (1999). Error caused by competing sentence formulations is also readily identifiable in blends—as in Examples (m) and (n). These give strong evidence for multiple encoding paths within a sentence. It is easy to generalize such conditions to rationalize certain word substitutions and sound errors as well.

'Sentence external' processes provide additional sources of competing expressions. Such may be discourse-relevant variations on sentence form or discourse external representations for an independent line of thought. This latter may also include interaction of linguistic processing with attentional mechanisms that filter automatic responses to the environment. A 'filtering breakdown' may lead to incorporation of environmental material into speech (see e.g., Harley 1990).

Sigmund Freud (1901) is identified with accounts of errors motivated by thoughts related to unconscious wishes and desires. His account links those thoughts to the meaning of the words or phrases resulting from error. That requirement suggests that Freudian mechanisms are not a major factor in the generation of movement errors. A robust feature of virtually all such errors is their indifference to the meaning of error products (Bierwisch 1970). An alternative focus for Freud's claims may lie in editorial processes—i.e., accounts of self-monitoring and its relation to the mechanisms of sentence production (see Motley et al. 1982).

5. *The Broader Picture: Building Models of the Language Production Process*

Speech error analysis provides a rich source of information about the real-time processes that underlie human abilities to generate fluent, coherent, oral discourse. But, rich as it is, there are sharp limitations on its utility for certain questions about language

production. Issues of semantic and discourse structure are awkwardly represented or absent from error data, and multiple questions about the temporal relations of lexical and phrasal integration are undetermined by the patterns of naturally occurring speech error. At precisely this point one turns to experimental studies of production performance.

Bock (1987, 1990, 1995) and collaborators have used picture description tasks and related techniques to uncover detailed relations between sentence generation and the underlying conceptual structures that sentences encode. Models of lexical retrieval have been extensively investigated via picture naming and related paradigms (see *Speech Production, Psychology of*). A family of experiments that focuses on TOT (tip of the tongue) states provides complementary information on lexical staging for syntactic and phonological representations (see *Tip-of-the-tongue, Psychology of*). Studies using error induction techniques also provide points of contact with spontaneous error data. See Shattuck-Hufnagel (1987) for a comprehensive series of error induction studies relating sound error to lexical and phrasal structure, including stress factors. Work by Baars et al. (1975) led to a series of investigations showing lexical influences on the probability of sound exchanges. Some aspects of the experimentation point to monitoring processes as the source of such effects (see Levelt et al. 1999 for review). An extensive experimental literature on induced agreement errors has developed from original work by Bock and Miller (1991) on English number agreement, and has been applied to agreement issues in several additional languages (see, e.g., Vigliocco et al. 1999). See Bock and Levelt (1994) for a recent synthesis of several strands of production experimentation.

See also: Aphasia; Speech Production, Neural Basis of; Speech Production, Psychology of; Tip-of-the-tongue, Psychology of

Bibliography

Baars B, Motley M, McKay D 1975 Output editing for lexical status from artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior* **14**: 382–91

Bierwisch M 1970/1982 Linguistics and language error. In: Cutler A (ed.) *Slips of the Tongue*. Mouton, Amsterdam

Bock J K 1987 Coordinating words and syntax in speech plans. Ellis (ed.) *Progress in the Psychology of Language*, Vol. 3. Erlbaum, London, pp. 337–90

Bock J K 1990 Structure in language: Creating form in talk. *American Psychologist* **45**: 1221–36

Bock K, Levelt W J M 1994 Language production: Grammatical encoding. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, London, pp. 945–84

Bock J K, Loebell H, Morey R 1992 From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review* **99**: 150–71

Bock J K, Miller C A 1991 Broken agreement. *Cognitive Psychology* **23**: 45–93

Boomer D, Laver J 1968 Slips of the tongue. *British Journal of Disorders of Communication* **3**: 2–12

Butterworth B 1982 Speech errors: Old data in search of new thories. In: Cutler A (ed.) *Slips of the Tongue*. Mouton, Amsterdam

Caldognetto E, Tonelli L, Vagges K, Cosi P 1987 The organization of constraints on phonological speech errors. *Proceedings of the XIth International Congress of Phonetic Sciences*, Tallinn, U.S.S.R. Vol. 3, 50.4.1–50.4.4

Cutler A 1982 The reliability of speech error data. In: Cutler A (ed.) *Slips of the Tongue*. Mouton, Amsterdam

Cutler A, Fay D 1978 Introduction (ix–xxxv). In: New edn. *Versprechen und Verlesen: Eine Psycologisch-Linguistische Studie*. John Benjamins, Amsterdam

del Viso S, Igoa J, Garcia-Albea J 1991 On the autonomy of phonological encoding: Evidence from slips of the tongue in Spanish. *Journal of Psycholinguistic Research* **20**: 161–85

Dell G 1984 The representation of serial order in speech: Evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology, Learning Memory & Cognition* **10**: 222–33

Dell G 1986 A spreading activation theory of retrieval in sentence production. *Psychological Review* **93**: 283–21

Dell G 1991 Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes* **5**: 313–49

Dell G, Reich P 1981 Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior* **20**: 611–29

Dell G S, Schwartz M F, Martin N, Saffran E M, Gagnon D A 1997 Lexical access in aphasic and nonaphasic speakers. *Psychological Review* **104**(4): 801–38

Fay D, Cutler A 1977 Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry* **8**: 505–20

Freud S 1901 The psychopathology of everyday life. *Monatschrift fur Psychiatrie und Neurologie* **10**: 1–13

Fromkin V 1971 The non-anomalous nature of anomalous utterances. *Language* **47**: 27–52

Garrett M F 1975 The analysis of sentence production. In: Bower G (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory*. Academic Press, New York, Vol. 9

Garrett M 1980 Levels of processing in sentence production. In: Butterworth B (ed.) *Language Production: Vol 1. Speech and Talk*. Academic Press, London, pp. 177–220

Garrett M 1993 Errors and their relevance for models of language production. In: Dittman J, Blanken G (eds.) *Handbooks of Linguistics and Communication Science: Linguistic Disorders and Pathologies*. Walter De Gruyter, New York

Harley T 1990 Environmental contamination of normal speech. *Applied Psycholinguistics* **11**: 45–72

Levelt W J M 1989 *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA

Levelt W J M, Roelofs A, Meyer A 1999 A theory of lexical access in speech production. *Behavioral and Brain Sciences* **22**: 1–75

McKay D G 1970 Spoonerisms: The structure of error in the serial order of speech. *Neuropsychologia* **8**: 323–50

McKay D G 1972 The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology* **3**: 210–27

Meringer R, Meyer K 1895/1978 Versprechen und Verlesen: Eine Psychologisch Linguistische Studie. Stuttgart: Gos-

chensche Verlag. In: Cutler A, Fay D (eds.) John Benjamins, Amsterdam
Myers-Scotton C, Jake J 2001 Explaining aspects of code switching and their implications. In: Nicol J (ed.) *One Mind, Two Languages: Bilingual Language Processing*. Blackwell, Oxford, pp. 91–125
Motley M, Camden C, Baars B 1982 Covert formulation and editing of anomalies in speech production. *Journal of Verbal Learning and Verbal Behavior* **21**: 578–94
Shattuck-Hufnagel S 1987 The role of word onset consonants in speech production planning: New evidence from speech error patterns. In: Kellar E, Gopnik M (eds.) *Motor and Sensory Processing in Language*. Erlbaum Press, Hillsdale, NJ
Stemberger J 1985 An interactive activation model of language production. In: Ellis A (ed.) *Progress in the Psychology of Language*. Erlbaum, London, Vol. 1
Stemberger J, McWhinney B 1986 Form-oriented inflectional errors in language processing. *Cognitive Psychology* **18**: 329–54
Vousden J I, Brown G D A, Harley T A 2000 Serial control of phonology in speech production: a hierarchical model. *Cognitive Psychology* **41**: 101–75

M. Garrett

Speech Perception

It is a commonplace to have the impression that foreign languages are spoken much more rapidly than our own, and without silent periods between the words and sentences. Our own language, however, is perceived at a normal pace (or even too slowly at times) with clear periods of silence between the words and sentences. In fact, languages are spoken at approximately the same rate, and these experienced differences are solely due to the memory structures and psychological processes involved in speech perception. Thus we define speech perception as the process of imposing a meaningful perceptual experience on an otherwise meaningless speech input. The empirical and theoretical investigation of speech perception has blossomed into an active interdisciplinary endeavor, including the fields of psychophysics, neurophysiology, sensory perception, psycholinguistics, linguistics, artificial intelligence, and sociolinguistics.

1. *Psychophysics of Speech Perception*

In any domain of perception, one goal is to determine the stimulus properties responsible for perception and recognition of the objects in that domain (see *Psychophysics*). The study of speech perception promises to be even more challenging than other domains of perception because there appears to be a discrepancy between the stimulus and the perceiver's experience of it. For speech, we perceive mostly a discrete auditory message composed of words, phrases, and sentences. The stimulus input for this experience, however, is a continuous stream of sound (and facial and gestural movements in face-to-face communication) produced by the speech production process. Somehow, this continuous input is transformed into more or less a meaningful sequence of discrete events (see *Psychophysics*).

Although we have made great progress on this front, there is still active controversy over the ecological properties of the speech input that are actually functional in speech perception. One issue, revived by recent findings, is whether the functional properties in the signal are static or dynamic (changing with time). Traditionally, static cues (such as the location of formants (bands of energy in the acoustic signal related to vocal tract configuration), the distribution of spectral noise as in the onset of *saw* and *shawl*, and the mouth shape at the onset of a segment) have been shown to be effective in influencing speech perception. Dynamic cues such as the transition of energy between a consonant and the following vowel have also been shown to be important. For example, recent research has shown that the second formant (F2) transition defined as the change between the F2 value at the onset of a consonant–vowel (CV) transition and the F2 value in the middle of the following vowel is a reliable predictor of the place of articulation category (Sussman et al. 1998).

Controversy arises when research is carried out to argue for one type of cue rather than another. For example, investigators recently isolated short segments of the speech signal and reversed the order of the speech within each segment (Saberi and Perrott 1999). In this procedure, a sentence is divided into a sequence of successive segments of a fixed duration such as 50 ms. Each segment is time-reversed and these new segments are recombined in their original order, without smoothing the transition borders between the segments. In this fashion, the sentence could be described as between globally contiguous but locally time-reversed. The authors claimed that the speech was still intelligible when the reversed segments were relatively short (1/20th to 15th of a second). Their conclusion was that our perception of speech was demonstrated to be primarily dependent on higher-order dynamic properties rather than the short static cues normally assumed by most current theories. This type of study and logic follows a tradition of attempting to find a single explanation or influence of some psychological phenomenon. However, most successful research in psychology is better framed within the more general framework of *ceteris paribus* (all other aspects neutral). There is good evidence that perceivers exploit many different cues in speech perception, and attempting to isolate a single functionally sufficient cue is futile.

There is now a large body of evidence indicating that multiple sources of information are available to

support the perception, identification, and interpretation of spoken language. There is an ideal experimental paradigm that allows us to determine which of the many potentially functional cues are actually used by human observers, and how these cues are combined to achieve speech perception (Massaro 1998). The systematic variation of the properties of the speech signal and quantitative tests of models of speech perception allow the investigator to interpret the psychological validity of different cues. This paradigm has already proven to be effective in the study of audible, visible, and bimodal speech perception (Massaro 1998). Thus, this research strategy addresses how different sources of information are evaluated and integrated, and can identify the sources of information that are actually used.

2. *The Demise of Categorical Perception*

Too often, behavioral scientists allow their phenomenal impressions to spill over into their theoretical constructs. One experience in speech perception is that of categorical perception. Using synthetic speech, it is possible to make a continuum of different sounds varying in small steps between two alternatives. Listening to a synthetic speech continuum between /ba/ and /pa/ provides an impressive demonstration: students and colleagues usually agree that their percept changes qualitatively from one category to the other in a single step or two with very little fuzziness in-between. Consistent with this impression, there is no denying that we experience discrete categories in speech perception. This experience of discrete perception was developed as a description of the speech perception process and called categorical perception (CP), or the perceived equality of instances within a category. The CP of phonemes has been a central concept in the experimental and theoretical investigation of speech perception and has also spilled over into other domains such as face processing (Beale and Keil 1995, Etcoff and Mcgee 1992). CP was operationalized in terms of discrimination performance being limited by identification performance. Researchers at Haskins Laboratories (Liberman et al. 1957) used synthetic speech to generate a series of 14 consonant–vowel syllables going from /be/ to /de/ to /ge/ (/e/ as in gate). The onset frequency of the second formant transition of the initial consonant was changed in equal steps to produce the continuum. In the identification task, observers identified random presentations of the sounds as /b/, /d/, or /g/. The discrimination task used the ABX paradigm. Three stimuli were presented in the order ABX; A and B always differed and X was identical to either A or B. Observers were instructed to indicate whether X was equal to A or B. This judgment was supposedly based on auditory discrimination in that observers were instructed to use whatever auditory differences they could perceive.

The experiment was designed to test the hypothesis that listeners can discriminate the syllables only to the extent that they can recognize them as different phoneme categories. The CP hypothesis was quantified in order to predict discrimination performance from the identification judgments. The authors concluded that discrimination performance was fairly well predicted by identification. This rough correspondence between identification and discrimination has provided the major source of support for CP. There have been several notable limitations in these studies. Discrimination performance is consistently better than that predicted by identification, it is possible that participants are making their discrimination judgments on the basis of identification rather than on their auditory discrimination, and it is likely that any alternative theory would have described the results equally well (Massaro 1987).

In many areas of inquiry, a new experimental paradigm enlightens our understanding by helping to resolve theoretical controversies. For CP, rating experiments were used to determine if perceivers indeed have information about the degree of category membership. Rather than ask for categorical decisions, perceivers are asked to rate the stimulus along a continuum between two categories. A detailed quantitative analysis of the results indicated that perceivers have reliable information about the degree of category membership, contrary to the tenets of CP (Massaro and Cohen 1983). Although communication forces us to partition the inputs into discrete categories for understanding, this property in no way implies that speech perception is categorical. To retrieve a toy upon request, a child might have to decide between *ball* and *doll*; however, she can certainly have information about the degree to which each toy was requested.

Although CP has been discredited, it is often reinvented under new guises. Most recently, the perceptual-magnet effect (PME) has had a tremendous impact on the field, and has generated a great deal of research (Kuhl 1991). The critical idea is that the discriminability of a speech segment is inversely related to its category goodness. Ideal instances of a category are supposedly very difficult to distinguish from one another relative to poor instances of the category. If we understand that poor instances of one category will often tend to be at the boundary between two categories, then the PME is more or less a reformulation of prototypical CP. That is, discrimination is predicted to be more accurate between categories than within categories. In demonstrating its viability, the PME faces the same barriers that have been difficult to eliminate in CP research. In standard CP research, it is necessary to show how discrimination is directly predicted by identification performance. In the PME framework, it is also necessary to show how discrimination is directly predicted by a measure of category goodness. We can expect category goodness

to be related to identification performance. Good category instances will tend to be identified equivalently, whereas poor instances will likely be identified as instances of different categories. Lotto et al. (1998) found that discriminability was *not* poorer for vowels with high category goodness, in contrast to the predictions of the PME. The authors also observed that category goodness ratings were highly context sensitive, which is a problem for the PME. If category goodness is functional in discrimination, it should be relatively stable across different contexts.

Notwithstanding the three decades of misinterpreting the relationship between identification and discrimination of auditory speech, we must conclude that it is perceived continuously and not categorically. Recent research reveals conclusively that both visible and bimodal speech are perceived continuously (Massaro 1987, 1998). This observation pulls the carpet from under current views of language acquisition that attribute to the infant and child discrete speech categories (Eimas 1985, Gleitman and Wanner 1982). Most importantly, the case for the specialization of speech is weakened considerably because of the central role that the assumption of CP has played (Liberman and Mattingly 1985). Finally, several neural network theories such as single-layer perceptrons, recurrent network models, and interactive activation have been developed to predict CP (Damper 1994): its nonexistence poses great problems for these models.

Given the existence of multiple sources of information in speech perception, each perceived continuously, a new type of theory is needed. The theory must describe how each of the many sources of information is evaluated, how the many sources are combined or integrated, and how decisions are made. The development of a promising theory has evolved from sophisticated experimental designs and quantitative model testing to understand speech perception and pattern recognition more generally. A wide variety of results have been accurately described within the Fuzzy Logical Model of Perception (FLMP).

3. *The Fuzzy Logical Model of Perception (FLMP)*

The three processes involved in perceptual recognition are illustrated in Fig. 1 and include evaluation, integration, and decision. These processes make use of prototypes stored in long-term memory. The evaluation process transforms these sources of information into psychological values, which are then integrated to give an overall degree of support for each speech alternative. The decision operation maps the outputs of integration into some response alternative. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. The assumptions central to the model are: (a) each source of information is evaluated to determine the continuous degree to which that source specifies various alternatives; (b) the sources of information are evaluated independently of one another; (c) the sources are integrated to provide an overall continuous degree of support for each alternative; and (d) perceptual identification and interpretation follows the relative degree of support among the alternatives. In the course of our research, we have found the FLMP to be a universal principle of perceptual cognitive performance that accurately models human pattern recognition. People are influenced by multiple sources of information in a diverse set of situations. In many cases, these sources of information are ambiguous and any particular source alone does not usually specify completely the appropriate interpretation.

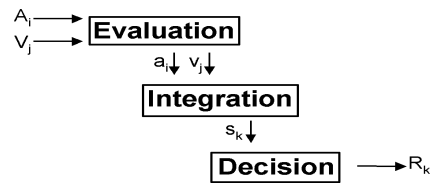

Figure 1
Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by upper-case letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j) These sources are then integrated to give an overall degree of support, s_k, for each speech alternative k. The decision operation maps the outputs of integration into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely

4. *Multimodal Speech Perception*

Speech perception has traditionally been viewed as a unimodal process, but in fact appears to be a prototypical case of multimodal perception. This is best seen in face-to-face communication. Experiments have revealed conclusively that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech (Massaro 1998). Consider a simple syllable identification task. Synthetic visible speech and natural audible speech were used to generate the consonant–vowel (CV) syllables /ba/, /va/, /a/, and /da/. Using an expanded factorial design, the four syllables

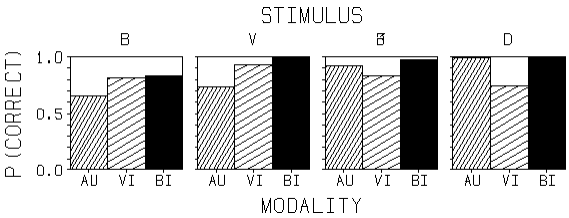

Figure 2
Probability correct identification of the unimodal (AU = Auditory, VI = Visual) and bimodal (BI) consistent trials for the four test syllables

were presented auditorily, visually, and bimodally. Each syllable was presented alone in each modality for $4 \times 2 = 8$ unimodal trials. For the bimodal presentation, each audible syllable was presented with each visible syllable for a total of 4×4 or 16 unique trials. Thus, there were 24 types of trials. Twelve of the bimodal syllables had inconsistent auditory and visual information. The 20 participants in the experiment were instructed to watch and listen to the talking head and to indicate the syllable that was spoken.

Accuracy is given in Fig. 2 for unimodal and bimodal trials when the two syllables are consistent with one another. Performance was more accurate given two consistent sources of information than given either one presented alone. Consistent auditory information improved visual performance about as much as consistent visual information improved auditory performance. Given inconsistent information from the two sources, performance was poorer than observed in the unimodal conditions. These results show a large influence of both modalities on performance with a larger influence from the auditory than the visual source of information.

Although the results demonstrate that perceivers use both auditory and visible speech in perception, they do not indicate how the two sources are used together. There are many possible ways the two sources might be used. We first consider the predictions of the FLMP.

In a two-alternative task with /ba/ and /da/ alternatives, the degree of auditory support for /da/ can be represented by a_i, and the support for /ba/ by $(1–a_i)$. Similarly, the degree of visual support for /da/ can be represented by v_j, and the support for /ba/ by $(1–v_j)$. The probability of a response to the unimodal stimulus is simply equal to the feature value. For bimodal trials, the predicted probability of a response, $P(/\mathrm{da}/)$ is equal to

$$P(/\mathrm{da}/) = \frac{a_i v_j}{a_i v_j + (1-a_i)(1-v_j)} \qquad (1)$$

In previous work, the FLMP has been contrasted against several alternative models such as a weighted averaging model (WTAV), which is an inefficient algorithm for combining the auditory and visual sources. For bimodal trials, the predicted probability of a response, $P(/\mathrm{da}/)$ is equal to

$$P(/\mathrm{da}/) = \frac{w_1 a_1 + w_2 v_j}{w_1 + w_2} = w a_i + (1-w) v_j \qquad (2)$$

The WTAV predicts that two sources can never be more informative than one. In direct contrasts, the FLMP has consistently and significantly outperformed the WTAV (Massaro 1998).

More generally, research has shown that the results are well-described by the FLMP, an optimal integration of the two sources of information (Massaro and Stork 1998). A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. For example, if the ambiguous auditory sentence, *My bab pop me poo brive*, is paired with the visible sentence, *My gag kok me koo grive*, the perceiver is likely to hear, *My dad taught me to drive*. Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro and Stork 1998).

Recent findings show that speech reading, or the ability to obtain speech information from the face, is not compromised by oblique views, partial obstruction or visual distance. Humans are fairly good at speech reading even if they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Massaro 1998).

5. *Higher-order or Top-down Influences*

Consistent with the framework being developed, there is now a substantial body of research illustrating that speech perception is influenced by a variety of contextual sources of information. Bottom-up sources correspond to those sources that have a direct mapping between the sensory input and the representational unit in question. Top-down sources or contextual information come from constraints that are not directly mapped onto the unit in question. As an example, a bottom-up source would be the stimulus presentation of a test word after the presentation of a top-down source, a sentence context. A critical question for both integration and autonomous (modularity, models is how bottom-up and top-down sources of information work together to achieve word recognition. For example, an important question is how early can contextual information be integrated with acoustic-phonetic information. A large body of research shows that several bottom-up sources are

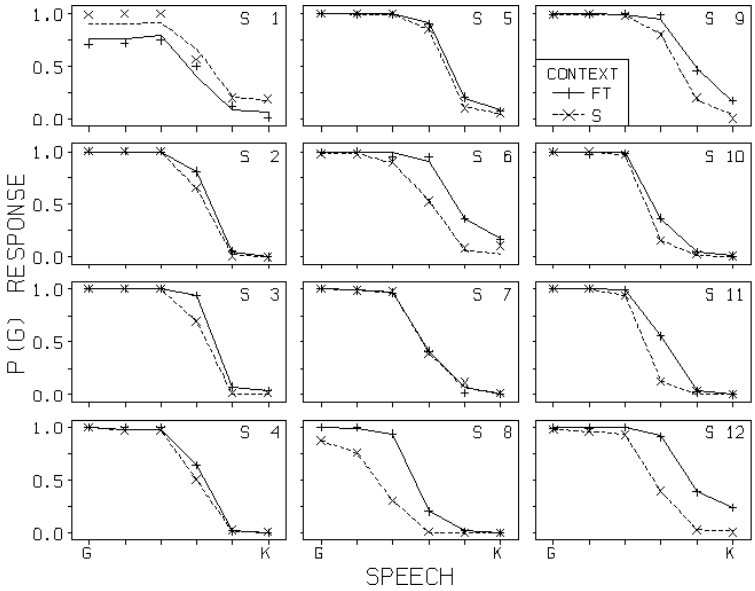

Figure 3
Observed (points) and FLMP's predictions (lines) of /g/ identifications for FT and S contexts as a function of the speech information of the initial consonant. Results from Pitt's (1995) Experiment 3a

evaluated in parallel and integrated to achieve recognition (Massaro 1987, 1994). An important question is whether top-down and bottom-up sources are processed in the same *manner*. A critical characteristic of autonomous models might be described as the language user's *inability* to integrate bottom-up and top-down information. An autonomous model must necessarily predict no perceptual integration of top-down with bottom-up information.

Pitt (1995) studied the joint influence of phonological information and lexical context in an experimental paradigm developed by Ganong (1980). A speech continuum is made between two alternatives, and the contextual information supports one alternative or the other. The initial consonant of the CVC syllable was varied in six steps between /g/ and /k/. The following context was either /Ift/ or /Is/. The context /Ift/ favors or supports initial /g/ because *gift* is a word whereas *kift* is not. Similarly, the context /Is/ favors or supports initial /k/ because *kiss* is a word whereas *giss* is not. Pitt improved on earlier studies by collecting enough observations to allow us to perform a subject-by-subject evaluation of the ability of specific models of language processing to account for the results. Previous tests of models using this task have been primarily dependent on group averages which may not be representative of the individuals that make the averages up.

The points in Fig. 3 give the observed results for each of the 12 subjects in the task. For most of the subjects, the individual results tend to resemble the average results reported by Pitt and earlier investigators. Ten of the 12 subjects were influenced by lexical context in the appropriate direction. Subject 1 gave an inverse context effect and Subject 7 was not influenced by context.

This FLMP was applied to the identification results of the 12 individual subjects in Pitt's Experiment 3a for which the greatest number of observations (104) were obtained for each data point for each subject. According to the FLMP, both the bottom-up information from the initial speech segment and the top-down context are evaluated and integrated. If s_i is the degree of support for the voiced alternative given by the initial segment and c_j is the support given by the following context, the total support for the voiced alternative i is given by

$$S(\text{voiced} \mid S_i C_j) = s_i \times c_j \tag{3}$$

The support for the voiceless alternative would be

$$S(\text{voiceless} \mid S_i C_j) = (1 - s_i) \times (1 - c_j) \tag{4}$$

The predicted probability of a voiced response is simply

$$P(\text{voiced} \mid S_i C_j) = \frac{S(\text{voiced} \mid S_i C_j)}{S(\text{voiced} \mid S_i C_j) + S(\text{voiceless} \mid S_i C_j)} \tag{5}$$

In producing predictions for the FLMP, it is necessary to estimate parameter values for each level of each

experimental factor. The initial consonant was varied along six steps between /g/ and /k/ and the following context was either /Ift/ or /Is/. Thus, there were six levels of bottom-up phonological information s_i and two contexts c_j. A free parameter is necessary for each level of bottom-up information but it is reasonable to assume that the contextual support given by /Is/ is one minus the lexical support given by /Ift/ so that only one value of c_j needs to be estimated. Thus, seven free parameters are used to predict the 12 independent points.

The lines in Fig. 3 also give the predictions of the FLMP. As can be seen in the figure, the model generally provides a good description of the results of this study. The root mean squared deviation (RMSD) between predicted and obtained is 0.017 on the average across all 12 independent fits. For the 10 subjects showing appropriate context effects, the RMSD ranges from 0.003 to 0.045 with a median of 0.007. Thus, for each of these individuals, the model captures the observed interaction between phonological information and lexical context: the effect of context was greater to the extent that the phonological information was ambiguous. This yields a pattern of curves in the shape of an American football, which is a trademark of the FLMP.

The model tests have established that perceivers integrate top-down and bottom-up information in language processing, as described by the FLMP. This result means that sensory information and context are integrated in the same manner as several sources of bottom-up information. These results pose problems for autonomous models of language processing.

See also: Audition and Hearing; Speech Production, Neural Basis of; Speech Production, Psychology of; Speech Recognition and Production by Machines

Bibliography

Beale J M, Keil F C 1995 Categorical effects in the perception of faces. *Cognition* **57**(3): 217–39

Damper R I 1994 Connectionist models of categorical perception of speech. *Proceedings of the IEEE International Symposium on Speech, Image Processing and Neural Networks,* Vol. 1, pp. 101–4

Etcoff N L, McGee J J 1992 Categorical perception of facial expressions. *Cognition* **44**: 227–40

Eimas P D 1985 The perception of speech in early infancy. *Scientific American* **252**: 46–52

Fitch H L, Halwes T, Erickson D M, Liberman A M 1980 Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics* **27**: 343–50

Ganong III W F 1980 Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance* **6**: 110–25

Gleitman L R, Wanner E 1982 Language acquisition: The state of the state of the art. In: Wanner E, Gleitman L R (eds.) *Language Acquisition: The State of the Art*. Cambridge University Press, Cambridge, UK

Kuhl P K 1991 Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics* **50**: 93–107

Liberman A M, Harris K S, Hoffman H S, Griffith B C 1957 The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* **54**: 358–68 , 753–71

Liberman A M, Mattingly I G 1985 The motor theory of speech perception revised. *Cognition* **21**: 1–36

Lotto A J, Kluender K R, Holt L L 1998 Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America* **103**: 3648–55

Massaro D W 1987 *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Erlbaum Associates, Hillsdale, NJ

Massaro D W 1994 Psychological aspects of speech perception: Implications for research and theory. In: Gemsbacher M (eds.) *Handbook of Psycholingristics*. Acadamie Press, New York pp. 219–63

Massaro D W 1998 *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press, Cambridge, MA

Massaro D W, Cohen M M 1983 Categorical or continuous speech perception: A new test. *Speech Communication* **2**: 15–35

Massaro D W, Stork D G 1998 Sensory integration and speechreading by humans and machines. *American Scientist* **86**: 236–44

Pitt M A 1995 The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition* **21**: 1037–52

Saberi K, Perrott D R 1999 Cognitive restoration of reversed speech. *Nature* **398**: 760

Sussman H M, Fruchter D, Hilbert J, Sirosh J 1998 Linear correlates in the speech signal: The orderly output constraint. *Behavioral & Brain Sciences* **21**(2): 241–99

D. W. Massaro

Speech Production, Neural Basis of

1. *Brain Regions Involved in Speech Production*

Over a century ago, the French neurologist Paul Broca demonstrated that speech mechanisms could be localized in the human brain. He did this by interviewing a patient with a severe speech production disorder with output limited to the recurring utterance, '*tan*.' Upon the patient's death, Broca examined the brain and concluded that the patient's inability to speak was due to a lesion in the inferior part of the frontal lobe (Broca 1861b). A second patient also had severely reduced speech and was subsequently found to have a similar cortical lesion (Broca 1861a). Since the time of Broca, scientists have found that lesions to Broca's area alone are not enough to produce lasting speech deficits (e.g., Alexander et al., 1989; Dronkers et al. 2000; Mohr 1976). Many have attempted to diagram

and map additional areas that might also subserve speech and language functions. This chapter will focus on evidence related to brain mechanisms of speech production. Specifically, evidence will be reviewed from lesion studies, electrocortical stimulation, and functional neuroimaging, all of which are helping to elucidate the neural basis of the planning and execution of articulatory movements.

The involvement of these neural structures can best be understood in the framework of a working model that follows the utterance from its conception to its ultimate articulation. Numerous studies have confirmed that the temporal lobe is important for the translation of concepts into linguistic representations. Once the utterance has been linguistically formulated, it must be transferred to the speech mechanisms that will carry out its production. Articulation itself requires planning, initiation, modification, and execution. In addition to Broca's area, it has become clear that many areas are involved in these aspects of speech production. New research has suggested a role for the anterior insula in articulatory planning, the supplementary motor area for the initiation of sequential speech movements, the basal ganglia and the cerebellum in the modification of pitch, loudness and rate, and the primary motor face cortex and pre-motor cortex in the execution of articulatory movements. Each of these areas will be discussed below in terms of the lesion studies that have helped to identify them and the functional neuroimaging research that has corroborated these findings. This chapter will begin by discussing the structures that are believed to convey linguistic information to frontal lobe speech mechanisms from posterior language cortex and end with a description of those mechanisms that execute the articulatory plan.

Superior longitudinal and arcuate fasciculi. The arcuate fasciculus is a bundle of fibers that connects the temporal and frontal lobes. As it travels through the parietal lobe, it becomes part of the superior longitudinal fasciculus and incorporates parietal fibers as well. The arcuate fasciculus was identified by Norman Geschwind as the tract that could potentially connect Wernicke's language area to Broca's speech area as described in Wernicke's original model (see Geschwind 1970; Wernicke 1874). The Wernicke-Geschwind model predicts that disruption of this tract results in isolated repetition deficits. However, recently a more comprehensive disorder has been reported. Dronkers and colleagues found that lesions that sever the arcuate/superior longitudinal fasciculus result in a severe production deficit with a complete loss of propositional speech (see Dronkers et al. 2000). Only automatized utterances remain. These patients have lost more than just repetition skills—they are unable to transfer any information from temporal lobe language areas to anterior speech mechanisms.

Broca's two original patients were also noted to have this same production disorder, which Broca termed 'aphemia' (Broca 1861a; 1861b). Dronkers et al. (submitted) recently had the opportunity to scan the postmortem brains of Broca's two original patients. They found that deep lesions also existed in these brains, in addition to the lesions Broca noted in the inferior frontal lobe. These deep lesions involved the superior longitudinal fasciculus in both cases, suggesting that this lesion also played a role in these historic patients' production deficits.

Insular cortex. The insula is a region of neocortex hidden beneath the intersection of the frontal, parietal, and temporal lobes. In the past, the insula has been implicated only vaguely in speech production (e.g., Ojemann and Whitaker 1978; Tognola and Vignolo 1980). More recently, Dronkers (1996) linked lesions in a specific region of the insula to an inability to plan and coordinate the appropriate movements necessary for articulation. She found that all patients studied with this disorder, known as 'apraxia of speech,' had lesions in the superior tip of the precentral gyrus of the insula while those without speech apraxia did not. Dronkers concluded that this small portion of the insula is critical for coordinating articulatory movements.

In keeping with this conclusion, recent neuroimaging studies have reported activation in the insula with tasks such as articulation of single words, word reading, picture naming, and word generation (see Indefrey and Levelt 2000). Interestingly, the insula appears to be activated only when tasks involve articulation of non-repeated and phonologically complex words (e.g., Wise et al. 1999), but not when only automatic or simple articulatory patterns are produced (e.g. Murphy et al. 1997).

Supplementary motor cortex. The supplementary motor area is located in the superior frontal gyrus and extends medially between the hemispheres. Penfield and others described the participation of the supplementary motor area in speech by documenting the effects of electrocortical stimulation on this region in patients undergoing surgical removal of epileptic tissue (see Penfield and Roberts 1959). Stimulation to the supplementary motor area either caused patients to make involuntary vocalizations or interrupted their ability to speak.

Infarctions of the supplementary motor area often result in a transcortical motor aphasia (See Rapcsak and Rubens 1994 for a review). Spontaneous speech is initially dysfluent, but comprehension, repetition, and reading remain relatively intact. Similarly, complete excision of the supplementary motor area causes a transient aphasia with deficits in articulation that resolve quickly (Penfield and Roberts 1959).

A review of the lesion literature suggests that the supplementary motor area is involved in the initiation of sequential, voluntary movements, including those for speech (e.g., Ziegler et al. 1997). This also seems to be true for the right supplementary motor area, as demonstrated by several case studies as well as

functional imaging data (e.g., Indefrey and Levelt 2000; Laplane et al. 1977). Other functional neuroimaging tasks that elicit supplementary motor cortex activation include the control of breathing for speech and vocalization (Murphy et al. 1997) and automatic speech (e.g., reciting months of the year; Ackermann et al. 1998).

Basal ganglia. The basal ganglia are subcortical nuclei comprised of the caudate, putamen, and globus pallidus. These structures interact with the cortex and with a number of other subcortical structures in a series of feedback loops that help to maintain motor activity (see Love and Webb 1996). Lesions to the basal ganglia can result in several types of movement disorders including Parkinson's and Huntington's disease, which each have their own speech characteristics (Duffy 1995). Patients with Parkinson's disease (which results from a lack of dopaminergic innervation of the caudate/putamen) show hypokinetic speech with decreased intensity and little modulation of pitch or loudness. In contrast, patients with Huntington's disease (which results from a loss of GABAergic, inhibitory neurons in the caudate) exhibit hyperkinetic speech with erratic control of pitch and loudness. These data suggest a role for the basal ganglia in the modulation of articulation and phonation.

Similar disorders can be seen in patients with focal lesions in the basal ganglia. For example, Pickett et al. (1998) reported a case study of a patient with bilateral lesions in the putamen and caudate. Her speech was initially characterized as dysarthric and was very difficult to understand, with poor articulatory and phonatory control. Similarly, Fabbro and colleagues reported at least two cases in which lesions in the basal ganglia resulted in hypophonia and a reduction of speech initiation and output (Fabbro et al. 1996).

In functional neuroimaging studies, caudate activation has been noted in only a small number of speech production studies (see Indefrey and Levelt 2000), but this may be due to the fact that it is not generally treated as a region of interest in imaging studies of speech.

Cerebellum. The cerebellum has traditionally been thought to be involved in fine coordination of motor acts including speech. Evidence for this comes from patients with focal cerebellar lesions. Ataxic dysarthria following cerebellar injury is a motor speech disorder that appears to be due to a discoordination in muscular control of the speech apparatus, as opposed to loss of initiation or execution (Duffy 1995). Mutism has also been associated with midline lesions in the cerebellum, in both children and adults (Coplin et al. 1997). However, this mutism is often transient and can resolve into ataxic dysarthric speech.

A broader theory of cerebellar function suggests that the cerebellum acts as a timing mechanism. That is, the cerebellum's role in motor coordination (for speech and other coordinated motor acts) is based on a more general mechanism that allows for the precise temporal control of such movements (Ivry and Keele 1989). Consistent with this hypothesis, Gandour and Dardarananda (1984) and others have shown that cerebellar patients have abnormal voice onset time distributions, indicating a timing disorder.

In the late 1990s, the cerebellum was implicated in speech production as a store for verbal short-term memory. Silveri et al. (1998) studied a patient following right cerebellar hemispherectomy and found that the patient had poor verbal short-term memory due to an impaired phonological output buffer. Functional neuroimaging data have also suggested that the cerebellum may play a role in verbal short-term memory (Ivry and Fiez 2000).

The role of the cerebellum in speech is probably multi-faceted, based on the functional heterogeneity of the cerebellum in motor control. Functional neuroimaging studies may help to tease out the functions of the various subregions of the cerebellum and their roles in speech and language.

Broca's Area. Traditionally, the literature has emphasized Broca's area as a major structure for speech production. However, the function of Broca's area has been re-evaluated. Data from stroke patients and from neurosurgical excision have revealed that removal/injury to Broca's area results in a transient mutism that resolves in 3–6 weeks (Dronkers et al. 2000; Penfield and Roberts 1959). Also, focal lesions to Broca's area do not result in a persisting Broca's aphasia, and Broca's aphasia may result from lesions outside of Broca's area (Dronkers et al. 2000).

With the increasing acknowledgment of the contribution of other brain areas to speech production (discussed in this chapter), Broca's area is no longer considered to be as crucial to speech as once thought. Though clearly involved in articulation, Broca's area functions within a network of brain regions that support speech production. Indeed, neuroimaging studies have reported networks of activation that include Broca's area in tasks involving phonological encoding and articulation, such as picture naming, word generation, and reading (see Indefrey and Levelt 2000).

Motor face and pre-motor cortex. The muscles of the vocal mechanism are innervated by cranial nerves V, VII, IX, X, XI, and XII, which receive impulses from the cortex via the corticobulbar tract. The motor face area of primary motor cortex, as well as pre-motor cortex, are believed to be the source of these projections (Duffy 1995). Lesions to these areas result in a pure motor speech disorder with slow, effortful speech and impaired articulation but intact language (Alexander et al. 1989). Such lesions are presumed to disrupt the corticobulbar fibers that activate the cranial nerves necessary for speech.

Penfield and colleagues originally showed that a number of regions in primary motor and pre-motor cortex were involved in speech output, as electro-

cortical stimulation of these sites resulted in speech arrest (see Penfield and Roberts 1959). Similarly, Ojemann and Mateer (1979) showed that, in four cases, speech arrest followed stimulation of inferior motor and premotor regions in the left frontal cortex.

Activation of primary motor cortex has been reported in fMRI studies of speech production. Wildgruber et al. (1996) found that tongue movements elicited comparable right and left activation in the inferior region of the motor strip. Monotone recitation of the months of the year activated more left than right motor cortex, while singing activated more right than left motor cortex. Ackermann et al. (1998) also found left motor cortex activation for silent repetition of months of the year.

2. Conclusions

Much has been learned about the neural processes involved in speech based on studies that have examined the specific deficits associated with lesions in particular brain regions. While language mechanisms cluster in posterior neocortical areas, speech production is generally supported by anterior cortical regions. Broca's area, motor face area, supplementary motor area, and the insula all play important roles in aspects of speech production, as evidenced by the effects of lesions in these areas. The superior longitudinal fasciculus provides the mechanism by which language information is transferred to these anterior speech regions. The basal ganglia and cerebellum also assist in speech production through modulation, coordination, and timing of speech movements. Functional neuroimaging studies have corroborated these findings and have allowed the visualization of the networks involved in speech production on-line.

See also: Language Development, Neural Basis of; Speech Perception; Speech Production, Psychology of; Syntactic Aspects of Language, Neural Basis of

Bibliography

Ackermann H, Wildgruber D, Daum I, Grodd W 1998 Does the cerebellum contribute to cognitive aspects of speech production? A functional magnetic resonance imaging (fMRI) study in humans. *Neurosci. Letters* **247**: 187–90

Alexander M P, Benson D F, Stuss D T 1989 Frontal lobes and language. *Brain and Language* **37**: 656–91

Broca P 1861a Nouvelle observation d'aphémie produite par une lésion de la troisième circonvolution frontale [New observations of aphemia produced by a lesion of the third frontal convolution]. *Bulletins de la Société d'Anatomie (Paris), 2e serie* **6**: 398–407

Broca P 1861b Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole) [Remarks on the location of the faculty of articulate language, followed by an observation of aphemia (loss of speech)]. *Bulletins de la Société d'Anatomie (Paris), 2e serie* **6**: 330–57

Coplin W M, Kim D K, Kliot M, Bird T D 1997 Mutism in an adult following hypertensive cerebellar hemorrhage: nosological discussion and illustrative case. *Brain and Language*. **59**: 473–93

Dronkers N F 1996 A new brain region for coordinating speech articulation. *Nature*. **384**: 159–61

Dronkers N F, Plaisant O, Iba-Zizen M T, Cabanis E A (submitted) Magnetic resonance imaging of the brains of Paul Broca's historic cases

Dronkers N F, Redfern B B, Knight R T 2000 The neural architecture of language disorders. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA

Duffy J R 1995 *Motor Speech Disorders: Substrates, Differential Diagnosis and Management*. C. V. Mosby-Year Book, St Louis, MO

Fabbro F, Clarici A, Bava A 1996 Effects of left basal ganglia lesions on language production. *Perceptual and Motor Skills* **82**: 1291–8

Gandour J, Dardarananda R 1984 Voice onset time in aphasia: Thai. 2. Production. *Brain and Language* **23**: 177–205

Geschwind N 1970 The organization of language and the brain. *Science* **170**: 940–4

Indefrey P, Levelt W J 2000 The neural correlates of language production. In: M S Gazzaniga (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA

Ivry R B, Fiez J A 2000 Cerebellar contributions to cognition and imagery. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press: Cambridge, MA

Ivry R B, Keele S W 1989 Timing functions of the cerebellum. *Journal of Cognitive Neuroscience* **1**: 136–52

Laplane D, Talairach J, Meininger V, Bancaud J, Orgogozo J M 1977 Clinical consequences of corticectomies involving the supplementary motor area in man. *Journal of the Neurological Sciences* **34**: 301–14

Love R J, Webb W G 1996 *Neurology for the Speech-language Pathologist*. (Third ed.). Butterworth-Heienemann, Boston, MA

Mohr J P 1976 Broca's area and Broca's aphasia. In: Whitaker H A, Whitaker H (eds.) *Studies in Neurolinguistics*, Vol. 1. Academic Press, New York

Murphy K, Corfield D R, Guz A, Fink G R, Wise R J, Harrison J, Adams L 1997 Cerebral areas associated with motor control of speech in humans. *Journal of Applied Physics* **83**: 1438–47

Ojemann G, Mateer C 1979 Human language cortex: Localization of memory, syntax, and sequential motor-phoneme identification systems. *Science* **205**: 1401–3

Ojemann G A, Whitaker H A 1978 Language localization and variability. *Brain and Language* **6**: 239–60

Penfield W, Roberts L 1959 *Speech and Brain Mechanisms*. Princeton University Press, Princeton, NJ

Pickett E R, Kuniholm E, Protopapas A, Friedman J, Lieberman P 1998 Selective speech motor, syntax and cognitive deficits associated with bilateral damage to the putamen and the head of the caudate nucleus: A case study. *Neuropsychologia* **36**: 173–88

Rapcsak S Z, Rubens A B 1994 Localization of lesions in transcortical aphasia. In: Kertesz A (ed.) *Localization and Neuroimaging in Neuropsychology*. Academic Press, San Diego, CA

Silveri M C, Di Betta A M, Filippini V, Leggio M G, Molinari M 1998 Verbal short-term store-rehearsal system and the

cerebellum: Evidence from a patient with a right cerebellar lesion. *Brain* **121**: 2175–7
Tognola G, Vignolo L A 1980 Brain lesions associated with oral apraxia in stroke patients: A clini-neuro-radiological investigation with the CT scan. *Neuropsychologia* **18**: 257–72
Wernicke C 1874 *Der Aphasische Symptomencomplex* [The aphasia symptom complex]. Cohn and Weigert, Breslau
Wildgruber D, Ackermann H, Klose U, Kardatzki B, Grodd W 1996 Functional lateralization of speech production at primary motor cortex: A fMRI study. *NeuroReport* **7**: 2791–5
Wise R J S, Greene J, Buchel C, Scott S K 1999 Brain regions involved in articulation. *Lancet* **353**: 1057–61
Ziegler W, Kilian B, Deger K 1997 The role of the left mesial frontal cortex in fluent speech: Evidence from a case of left supplementary motor area hemorrhage. *Neuropsychologia* **35**: 1197–208

N. F. Dronkers and J. V. Baldo

Speech Production, Psychology of

Speech production refers to the cognitive processes engaged in going from mind to mouth (Bock 1995), that is, the processes transforming a nonlinguistic conceptual structure representing a communicative intention into a linguistically well-formed utterance. Within cognitive psychology, research concerning speech production has taken various forms such as: research concerning the communicative aspect of speaking; research concerning the phonetics of the produced speech; and research concerning the details of the cognitive processing machinery that translates conceptual structures into well-formed linguistic utterances. We focus on the latter.

Native adult speakers produce on average two to three words per second. These words are retrieved from a lexicon of approximately 30,000 (productively used) words. This is no small feat: producing connected speech not only entails retrieving words from memory, but further entails combining this information into well-formed sentences. Considering the complexity of all the encoding processes involved, it is impressive that we produce speech at such a fast rate while at the same time remaining highly accurate in our production. Bock (1991) estimated that slips of the tongue occur in speech approximately every 1,000 words despite the ample opportunities for errors. In the following sections the cognitive processes subserving this ability will be discussed.

1. Methodological Issues

The prototypical empirical approach in cognitive psychology consists of systematically varying some properties of a stimulus (input), and to measure a corresponding behavior. Systematic relations between input and behavior are then used to infer the cognitive processes mediating between them. For example, researchers interested in language comprehension may systematically vary properties of a linguistic input such as the syntactic complexity of sentences or the frequency of words, and measure a corresponding behavior, such as reading times. Differences in the latter are taken as an indication of differences in the processes engaged in building an interpretation of a sentence (the output). Thus, the input is directly observable and completely accessible to systematic manipulation, while the output and the cognitive processes leading to it are inferred from the behavior.

The situation is different for research in speech production. The input (the conceptual structure to be expressed), is not directly observable and not readily accessible for experimental manipulation. By contrast, the output (i.e., the spoken utterance), is directly observable. Given this situation, research on speech production has started by focusing on properties of the behavior. Most prominent in this respect are analysis of speech errors, and of hesitations and pauses as they occur in spontaneous speech. These analyses motivated the general theoretical framework for speech production (Fig. 1) which is described in detail below. More recently researchers have started to make speech production accessible to standard experimental approaches from cognitive psychology (see Bock 1996).

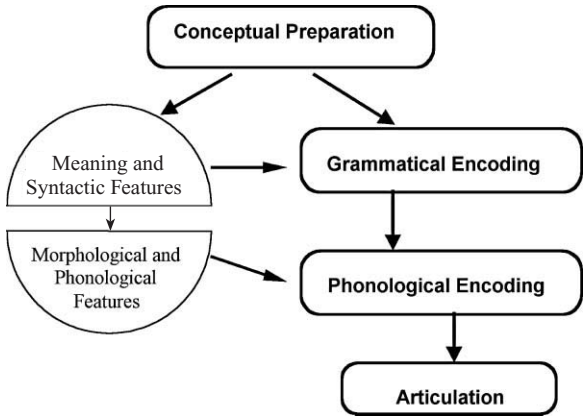

Figure 1
Levels of processing in speech production. Left part refers to stored representations, right part refers to processes integrating these representations into well-formed linguistic utterances

2. Levels and Processes

Figure 1 provides an outline of the levels of processing involved in speech production. These levels are shared by most current psycholinguistic models (e.g., Dell 1986, Garrett 1988, Levelt 1989). *Conceptual prep-*

aration refers to the processes that transform thoughts into verbalizable messages. *Grammatical encoding* refers to the processes that turn the messages into syntactically well-formed skeletons for sentences. *Phonological encoding* refers to the processes that provide flesh to the syntactic skeletons, realizing the phonological content of the sentences. At these latter two processing levels, stored lexical information is retrieved from a long-term memory store, the mental lexicon. In particular, during grammatical encoding it is assumed that abstract lexical representations, specifying the syntactic properties of the words, but not their phonological content, are retrieved. During phonological encoding, word form information is retrieved. Finally, *articulation* refers to the processes that implement the phonetic content and turn it into motor commands for speech.

These processing levels were originally inferred from properties of speech errors (see *Speech Errors, Psychology of*). Speech errors can be analyzed with respect to the type of linguistic unit involved (words, morphemes, phonemes), and with respect to constraints that hold for errors concerning a given type of unit. For example, word exchanges (e.g., 'on the room to my door' instead of 'on the door to my room') are constrained by syntactic factors (like the syntactic category of the exchanged words) whereas phoneme exchanges (e.g., 'heft lemisphere' instead of 'left hemisphere') are constrained by phonological factors (like their position as onset, nucleus, or coda of a syllable). These and related results suggest a distinction between a processing level that is primarily concerned with syntactic processing (grammatical encoding), and another level primarily concerned with phonological processing (phonological encoding). The fact that phonemes are relevant units in speech errors also shows that the phonological form of a word is not stored and retrieved as a whole, but built from smaller units, namely the phonemes.

Planning of a sentence is not completed on one processing level and then sent to the subsequent level. For example, phonological encoding starts as soon as the syntactic structure of some part of a sentence has been determined, while the grammatical encoding of the remainder of the sentence continues. This piece-by-piece generation of a sentence is referred to as incremental production. Although there is no consensus on the size of the relevant planning units at the different processing levels, it is generally agreed upon that the size of the planning units becomes smaller as one moves from conceptual preparation through grammatical encoding to phonological encoding and articulation. Incrementality provides a natural account of fluency in speech production. Furthermore, incremental production reduces the burden of storage of intermediate results of the production process at each level.

Although there is general agreement on the levels of processing, theories differ with respect to how the flow of information between levels is characterized. According to modular views, each level of processing first produces a complete output representation for a given planning unit. Only then is this output representation passed onto the next level (e.g., Levelt 1989). According to interactive views, every processing stage can pass partial results to the next processing level, before the output representation at the higher level has been completely specified, and lower levels of processing can provide feedback to higher levels of processing. As a consequence, processes at the phonological level can affect processes at the grammatical level (e.g., Dell 1986). In the following sections, we discuss the different processing steps in more detail.

2.1 Conceptual Preparation

A first important goal of conceptual preparation is to establish which parts of the conceptually available information are going to be encoded, and in what order (see Levelt 1989, for an overview). A second goal is to convert the conceptual information into a format that is a suitable for the linguistic formulation processes. One important open issue concerning conceptual preparation is how to characterize the mapping between conceptual and grammatical encoding. First is the question of whether conceptual preparation takes language specific properties into account (see *Language and Thought: The Modern Whorfian Hypothesis*). Languages differ in which conceptual or formal properties need to be realized as a detail of the sentential form. For example, in English the word 'friend' does not carry information concerning the sex of the friend. In Spanish the corresponding word is differentially inflected for a man ('*amigo*') or a woman ('*amiga*'). In English, adjectives used as predicates (e.g., 'tall' in 'The friend of Luis is tall') do not agree in gender with the noun; in Spanish they do (e.g., '*El amigo de Luis es alto*' or '*La amiga de Luis es alta*'). Thus, in these two languages, conceptual information concerning natural gender must (Spanish) or need not (English) be conveyed by a sentence. Second is the question of whether conceptual information permeates the processes occurring during grammatical encoding beyond providing its input (Vigliocco and Franck 1999).

2.2 Grammatical Encoding

Grammatical encoding refers to the processes involved in developing a syntactically well-formed sentence. It comprises first, those processes that map the relationships among the participants in a conceptual representation (e.g., agent, patient, etc.) onto functional syntactic relations between the words of a sentence (e.g., subject, direct object, etc.). Next, on the basis of the resulting hierarchically organized syntactic frame for the sentence, the words in the sentence are linearized in a manner allowed by the language being

spoken. The first step is also referred to as *functional level processing*, the second as *positional level processing* (Garrett 1988). Distinguishing between building hierarchical and linear frames provides a solution to an important problem that the language production system faces. Speech production has to be incremental to allow for fluent utterances, but at the same time the resulting utterance has to obey language specific constraints that force the use of only certain word orders. Assuming incremental conceptualization, however, the order in which parts of a conceptual message are processed do not necessarily correspond to a word order allowed in the speaker's language. This problem can be solved by separating the construction of hierarchical structures from the serial ordering of the words. In this way, hierarchical structures can be built in an incremental manner as soon as lexical elements are available, these can then be mapped to permissible linearly ordered positions (Vigliocco and Nicol 1998). Evidence compatible with such a separation comes from studies showing that the linear position of words can be primed by the previous presentation of the same linear order, even if the hierarchical structure differs (Hartsuiker et al. 1999).

An important open question with respect to grammatical encoding concerns whether the encoding at this level can be influenced by feedback from phonological encoding (Bock 1986). Research on the process of computing number agreement between a sentence subject and a verb suggests that this grammatical process is sensitive to whether number information in the subject noun phrase is overtly realized in the phonological form of the noun, suggesting the possibility of feedback from phonological encoding to grammatical encoding (e.g., Vigliocco et al. 1995; but see Sect. 2.5 for an alternative interpretation of the results).

2.3 Phonological Encoding

Phonological encoding refers to the processes that are responsible for determining the phonological word forms and prosodic content of the sentence (e.g., Levelt 1999). First, the phonemes of words are retrieved from the mental lexicon, together with a metrical frame which specifies stress pattern and number of syllables in the word. Following this retrieval process, the resulting sequence of phonemes is syllabified according to a language specific set of syllabification rules. The domain of syllabification is assumed to be the phonological word which can but need not coincide with a lexical word. It should be noted that in this view, the syllabic structure of an utterance is computed on-line. The reason for this assumption is that the actual syllabification of a word in running speech depends on the context in which it appears. For example, the word 'deceive' in isolation is syllabified as 'de-ceive,' but in the context of the utterance 'deceive us,' the syllable structure becomes 'de-cei-veus.' This observation also provides a functional reason why the phonological form of a word is not stored and retrieved as one entity: such a representation would have to be broken up into its constituent parts whenever the stored syllabification of a word does not agree with its syllabification in the context of running speech. The representation formed by phonological encoding, a syllabified phonological code, forms the input for the articulatory processes which realize this code as overt speech. It is an open issue whether the transition from phonological encoding to articulation involves accessing a 'syllabary,' i.e., memory representations specifying the motor tasks that have to be performed to generate each syllable (Levelt 1999).

2.4 Lexical Access

Grammatical and phonological encoding make use of stored representations of words. Parallel to the distinction between grammatical encoding and phonological encoding, two levels of stored representation are distinguished. On one level, words are represented as abstract syntactic (and semantic) entities (technically also referred to as lemmas), and on the other level their phonological form is specified. Evidence for two levels of representation in the mental lexicon comes from tip-of-the-tongue (TOT) states (see *Tip-of-the-tongue, Psychology of*). It has been demonstrated repeatedly that speakers in a TOT state can report grammatical properties of a word (e.g., the grammatical gender of a noun) above chance level even when they are unable to access the phonological form of the word.

It is unresolved whether, within the lexicon, the flow of information from meaning to form can be characterized as discrete, cascading, or fully interactive. Discrete serial models of lexical access (e.g., Levelt 1999) assume that, first, on the basis of some fragment of a conceptual structure, a set of lemmas becomes activated. This set includes the target lemma as well as semantically related lemmas. Phonological encoding of the word is initiated only after the target lemma has been selected. Hence, the phonological forms of words which are semantically related to the target word should not become activated. Cascading models (e.g., Peterson and Savoy 1998) assume that every activated lemma sends some activation to its corresponding word form, even before the actual target lemma has been selected. Thus, also the phonological forms of semantic competitors of the target word should receive some activation. Finally, interactive models (e.g., Dell 1986) allow for feedback from the phonological level to the lemma level. Thus not only will the phonological forms of semantic competitors receive activation, but the phonological forms will also send activation back to all lemmas with similar phonological forms.

The question of whether the phonological forms of semantic competitors do receive activation, as pre-

dicted by cascading models and feedback models, or not, as predicted by discrete serial models, is still an empirically unresolved issue. Phonological activation of semantic competitors has been demonstrated for the case of near-synonyms like 'couch' and 'sofa' (e.g., Jescheniak and Schriefers 1998, Peterson and Savoy 1998) and has been interpreted in favor of cascading of activation. However, as the evidence is presently restricted to this specific semantic relation, it has been suggested that near-synonyms might be a special case (e.g., Levelt 1999).

Concerning the question whether there is feedback from lower levels to higher levels in lexical processing, the lexical bias effect has been cited as evidence for feedback. The lexical bias effect concerns the observation that phoneme errors lead to existing words rather than nonwords more often than would be expected by chance. This finding can be explained by feedback from the level of phonological segments to some higher level representing words as one integral unit (e.g., Dell 1986). According to discrete serial models, by contrast, the chance of incorrect selection of a phoneme should be independent of whether the resulting string of phonemes forms an existing word or not (but see Sect. 2.5).

2.5 Self-monitoring and Repair

Speakers not only produce speech, but also monitor their own speech output and this monitoring may occur on overt speech as well as on 'internal speech.' Assuming an internal monitor implies that not all misfunctions of the speech production system will become visible in the eventual output. Rather, some will be discovered and repaired before they are actually articulated. Such covert repairs can play an important role in the interpretation of speech error results. For example, the lexical bias effect in phoneme errors has also been explained as the result of a pre-articulatory covert repair mechanism. This covert repair mechanism might be more likely to detect and correct phoneme errors leading to nonwords than those leading to words before the resulting error is actually produced. Hence, the lexical bias effect would not imply feedback of activation from lower to higher levels. It should be noted that related arguments have been put forward with respect to other empirical findings which appear to speak against discrete serial models, like the potential influence of phonology on grammatical encoding processes. It is a matter for future investigation to design empirical tests that allow us to differentiate between feedback and monitoring explanations of the relevant empirical phenomena.

See also: Aphasia; Dyslexia (Acquired) and Agraphia; Dyslexia, Developmental; Dyslexia, Genetics of; Speech Errors, Psychology of; Speech Production, Neural Basis of

Bibliography

Bock J K 1986 Meaning, sound and syntax: lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory and Cognition* **12**: 575–86

Bock J K 1991 A sketchbook of production problems. *Journal of Psycholinguistic Research* **20**: 141–60

Bock J K 1995 Sentence production. From mind to mouth. In: Miller J L, Eimas P D (eds.) *Handbook of Perception and Cognition. Speech, Language, and Communication*, 2nd edn. Academic Press, San Diego, CA, Vol. 11, pp. 181–216

Bock J K 1996 Language production: Methods and methodologies. *Psychonomic Bulletin and Review* **3**: 395–421

Dell G S 1986 A spreading-activation model of retrieval in sentence production. *Psychological Review* **91**: 283–321

Garrett M F 1988 Processes in language production. In: Nieumeyer F J (ed.) *Linguistics: The Cambridge Survey. Biological and Psychological Aspects of Language*. Cambridge University Press, Cambridge, UK, Vol. 3, pp. 69–96

Hartsuiker R J, Kolk H H J, Huiskamp P 1999 Priming word order in sentence production. *Quarterly Journal of Experimental Psychology* **52A**: 129–47

Jescheniak J D, Schriefers H 1998 Serial discrete versus cascaded processing in lexical access in speech production: Further evidence from the co-activation of near-synonyms. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**: 1256–74

Levelt W J M 1989 *Speaking. From intention to articulation*. MIT Press, Cambridge, MA

Levelt W J M 1999 Models of word production. *Trends in Cognitive Science* **3**: 223–32

Peterson R R, Savoy P 1998 Lexical selection and phonological encoding during language production: Evidence for cascaded processing. *Journal of Experimental Psychology: Language, Memory, and Cognition* **24**: 539–57

Vigliocco G, Butterworth B, Semenza C 1995 Constructing subject–verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language* **34**: 186–215

Vigliocco G, Franck J 1999 When sex and syntax go hand in hand: Gender agreement in language production. *Journal of Memory and Language* **40**: 455–78

Vigliocco G, Nicol J 1998 Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? *Cognition* **68**: B13–B29

H. Schriefers and G. Vigliocco

Speech Recognition and Production by Machines

The fields of speech recognition and speech production (text-to-speech or speech synthesis) have made great progress since the early 1990s. The technologies are now at the point of becoming commercially viable, and a number of products are currently available. Recent progress in both technologies has been charac-

terized by rapid incremental improvement rather than any theoretical breakthroughs. Both technologies are based on stochastic models and have benefited greatly from increased computing and data resources. This article will describe the techniques used for current state-of-the-art speech recognition and synthesis systems. It will give only a brief overview of the field. For further reading, see Jurafsky and Martin (2000) and Jelinek (1997).

1. Speech Recognition

This section will focus on large vocabulary continuous speech recognition (LVCSR), which is the most general form of the problem. There are other paradigms that are used to simplify the problem for some specific tasks. In isolated word recognition, as opposed to continuous recognition, the user is required to pause between words when speaking. This simplifies the problem by making it easier to find the word boundaries, so higher accuracies can be obtained. If the system only needs to detect a small set of words, and a small footprint is required, a technique called word spotting can be used. Word spotting attempts to find a small set of words in the input rather than trying to decode all of the speech. Both of these techniques can be viewed as special cases of LVCSR.

The general process of speech recognition is shown in Fig. 1. A signal processing component converts the analog speech signal into a sequence of discrete feature vectors. A decoding search process then matches internal acoustic models against the sequence of feature vectors to produce a sequence of words.

1.1 Signal Processing

While the exact signal processing configuration varies from system to system, the following description represents typical values. First, the analog signal is low-pass filtered with an 8 kHz cutoff and sampled at 16 kHz. A fast Fourier transform (FFT) is calculated using a 25 msec window, and the resulting amplitude spectrum is subjected to log compression. This log amplitude spectrum is then passed through another FFT resulting in a representation referred to as the cepstrum. Linear predictive coding (LPC) is used to calculate the coefficients for the cepstrum. In this scheme, each speech sample is represented as a linear combination of the N previous samples plus an additive term. The elements of the feature vector are the first N coefficients (for the linear combination) produced by the LPC calculation. A typical value for N is 12, and normally one feature vector is produced each 10 milliseconds.

1.2 The Matching Process

The most common form of speech recognition uses a maximum likelihood search. Given acoustic data A (as represented by the sequence of feature vectors), a maximum likelihood search attempts to find the word string W such that $P(W|A)$, the probability of the word sequence given the acoustic data, is maximized. This search evaluates hypotheses by applying Bayes Rule:

$$P(W|A) = \frac{P(A|W) \times P(W)}{P(A)}$$

where $P(A|W)$ is the probability of acoustic data A given word string W, $P(W)$ is the *a priori* probability of word string W, and $P(A)$ is the *a priori* probability of acoustic input A. $P(A|W)$ is estimated using acoustic models and $P(W)$ is estimated using language models. Typically, $P(A)$ is not estimated since it is difficult to estimate and it does not change the rank ordering of hypotheses since it is the same for all competing hypotheses.

1.3 Acoustic Models

Acoustic models are a stochastic mechanism for estimating $P(A|W)$, the probability of the observed acoustic feature sequence given a specified word sequence. The most common form of acoustic model

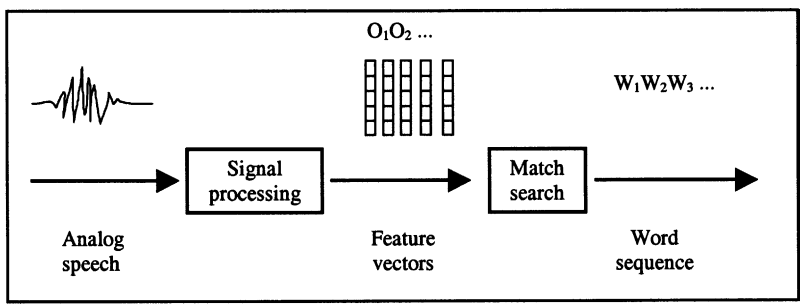

Figure 1
The speech recognition process

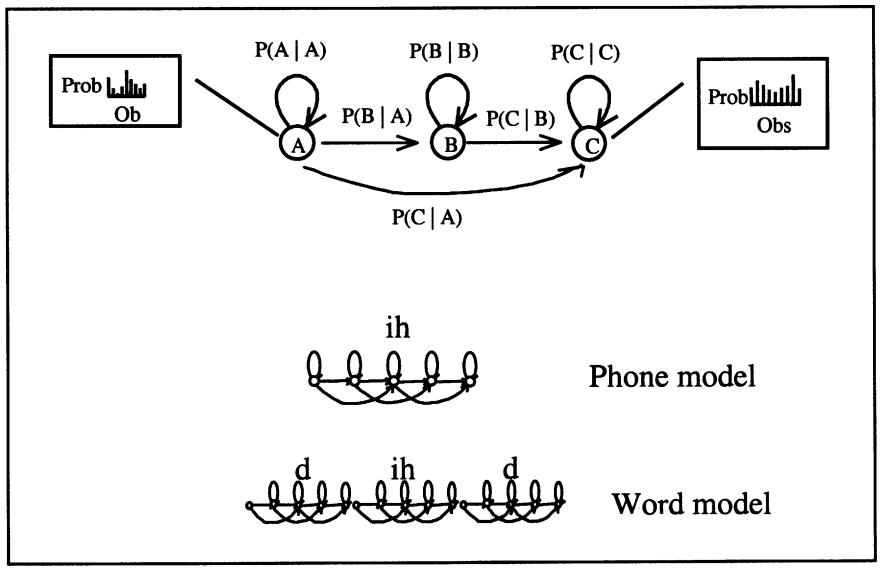

Figure 2
Hidden Markov models of speech units

used in speech recognition is the hidden Markov model (HMM). A hidden Markov model consists of a set of states S, a set of transition probabilities from each state to other states, and a set of observation probabilities for each state. This model is illustrated in Fig. 2. In order to apply the model to speech decoding, HMMs are associated with units of speech, such as phones, to model their acoustic realization. A word level model is generated by concatenating the HMMs for the phones comprising the word. In practice, more specific models are used which model phones in their left and right context. These models are referred to as triphones.

To create the models, the parameters of the HMMs, the transition and observation probabilities, must be estimated from training data. Recorded speech is transcribed, and the HMMs are aligned against the speech vectors using a procedure called the forward–backward algorithm. There is not a deterministic procedure for estimating the parameter values; rather an iterative procedure, Baum–Welch reestimation, is used to refine the parameters until a stopping criterion is achieved. While estimation of the transition probabilities is straightforward, estimation of the observation probabilities is less so since they are represented as a multidimensional space. For each state, the observation probability is the probability of observing the input feature vector given the state. There are two principal techniques for estimating the observation probabilities, Gaussian mixtures and neural networks.

In the Gaussian mixtures technique, the observation probability is estimated as a weighted sum of Gaussian probability density functions (PDFs). Every state has an associated set of multidimensional PDFs, each represented by a mean and covariance matrix. The feature vector is evaluated with respect to each of the distributions and the observation probability is a weighted sum of the values.

An alternate way of estimating the observation probabilities is to use neural networks to directly estimate the probability of the feature vectors. A recurrent neural network is associated with each Markov model state and gives the probability of the observation, given the state. These networks take a long time to train, but decode quickly.

1.4 Language Models

A language model is a stochastic mechanism for estimating $P(W)$, the *a priori* probability of a specified word sequence. $P(W)$ is normally decomposed to be applied in a left-to-right fashion. Using ! to represent the start of sentence, $P(W)$ is represented as:

$$P(W) = P(w_1|!) \times P(w_2|!w_1) \times P(w_3|!w_1w_2) \times \cdots \times P(w_n|!w_1w_2\ldots w_{n-1})$$

In this form, the probability of extending a sequence by the next word is conditioned on all the words in the sequence up to that point. In practice, it is not feasible to represent all possible histories, $!w_1w_2\ldots w_{n-1}$. The prevailing solution is to group all histories whose last $N-1$ words are the same, and thus to condition the probability of a word on the last $N-1$ words. These models are referred to as N-gram models. For $N = 3$,

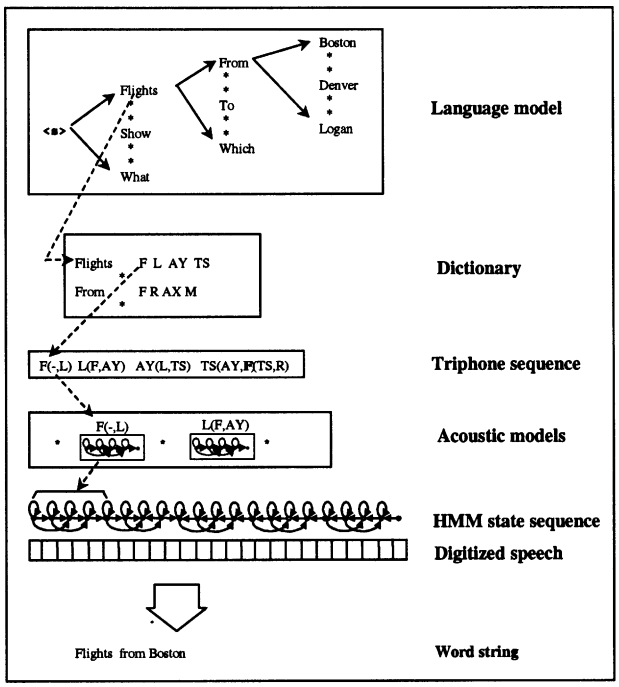

Figure 3
The speech decoding search

referred to as a trigram, $P(w_n|!w_1w_2\ldots w_{n-1})$ is estimated as $P(w_n|w_{n-2}w_{n-1})$. N-gram models are also trained from training data and estimated as:

$$P(w_n|w_{n-2}, w_{n-1}) = \frac{C(w_{n-2}, w_{n-1}, w_n)}{C(w_{n-2}, w_{n-1})}$$

where $C(w_{n-2}, w_{n-1}, w_n)$ is the count of the number of times that the sequence w_{n-2}, w_{n-1}, w_n was observed in the training data.

1.5 The Decoding Search

The general structure of the acoustic decoding process is shown in Fig. 3. It uses a language model, a dictionary, and an inventory of acoustic models to decode speech input. The recognizer processes input from left to right. It generates a language model probability for each word in the lexicon, given the start of utterance and matches these words against the start of the input. The system pursues hypotheses in parallel and extends them forward by assigning language model probabilities that each word will extend each hypothesis and matching that word against the corresponding portion of the speech input. In order to match a word against the digitized speech input, the system looks the word up in its pronunciation dictionary to find the sequence of phones describing the word pronunciation. The phones in the dictionary are context-independent, that is, the representation does not change for a phone dependent on its context, the phones to its left and right. Since the acoustic realization of a phone depends very much on its phonetic context, context-dependent models are actually used. The context-independent phones in the pronunciation are mapped to context-dependent phones; that is, notation is added to reflect the context of the phone to the left and to the right of the phone used. These-context dependent phone units are referred to as triphones. The system has a pretrained inventory of HMM acoustic models for the triphones. The triphone models are concatenated together to form a word model, which is matched against the speech input.

A dynamic programming search, the Viterbi algorithm, is used to align the speech feature vectors with the HMM triphone models. This search produces the sequence of Markov states that has the highest probability of generating the observed feature vectors. In reality the search does not generate real probabilities for hypotheses, but scores intended to approximate probabilities. The score for a hypothesis is a weighted combination of the acoustic match score $P(A|W)$ and the language model score $P(W)$. The match process is a large stochastic search to match models for word sequences against the speech input. In principle, all possible sequences of words in the

vocabulary are matched against the speech input. An exhaustive search clearly is not computationally viable, so pruning must be applied. The language model scores are very valuable in helping prune the search to look for the more likely word sequences. Most often, beam pruning is used. Hypotheses with scores more than a threshold value worse than the current best score are pruned.

1.6 Performance

The performance of a speech recognition system depends on many factors:

The time–accuracy tradeoff—since the recognition process is a pruned search, it has a time vs. accuracy tradeoff. Less pruning allows the system to explore more hypotheses and thus achieve higher accuracy, but also requires more run time. Research systems may run in many times real time, but systems interacting with users must run in very near real time. Increases in machine power can therefore translate into increased recognition accuracy by allowing less pruning while still running in real time.

Amount of training data—since the acoustic and language model parameters are estimated from training data, the amount and appropriateness of the training data critically affect the quality of the model.

Degree to which usage conditions are represented in training data—the weakest feature of current recognition technology is its lack of robustness. The systems generally perform well as long as the usage conditions are represented well by the training data. As the conditions under which the systems are used move away from the data on which they were trained, the performance degrades rapidly. If a system were trained only on male voices, it would not work well for most females. If it were trained on native English speakers, it would not work well for someone with a German accent (or almost any other accent). If trained with a close talking microphone, it would not work well over a telephone. So-called speaker-independent systems just use a training database containing as many different speakers as possible to get a reasonably broad coverage of speaker variability.

The commercial dictation systems that are currently available generally require a short enrollment process where the user is required to speak some prompted sentences into the system. This is to allow the system to adapt its general models to the specific user and insure at least a small amount of appropriate training data. Currently, commercial dictation systems use about 30,000 words and report average word error rates of approximately 5 percent.

2. Speech Production

Currently, the various approaches to speech recognition can be grouped into three basic categories: articulator models, source-filter models, and concatenative synthesis. An excellent brief discussion of the details of these approaches can be found in Pellom (1998). This section will present only a short overview.

2.1 Articulator Model Synthesis

Articulator model-based systems attempt to explicitly model the physical aspects of the human speech production system. They produce detailed models of the articulators and the airflow in the vocal tract. These models are generally based on detailed measurements of human speech production. Models of this type are not widely used because it is difficult to get the sensor data and because the models are computationally intensive.

2.2 Source-filter Based Synthesis

Source-filter-based systems use an abstract model of the speech production system (Fant 1960). Speech production is modeled as an excitation source that is passed through a linear digital filter. The excitation source represents either voiced or unvoiced speech, and the filter models the effect produced by the vocal tract on the signal. Within the general class of source-filter models, there are two basic synthesis techniques: formant synthesis and linear predictive synthesis.

Formant synthesis models the vocal tract as a digital filter with resonators and antiresonators (Klatt 1980). These systems use a low-pass filtered periodic pulse train as a source for voiced signals and a random noise generator as an unvoiced source. A mixture of the two sources can also be used for speech units that have both voiced and unvoiced properties. Rules are created to specify the time varying values for the control parameters for the filter and excitation. This is the technique used by DECTalk (Bruckert et al. 1983), probably the most used commercial synthesis system. Linear predictive synthesis uses linear predictive coding to represent the output signal. Each speech sample is represented as a linear combination of the N previous samples plus an additive excitation term. As in formant synthesis, the excitation term uses a pulse train for a voiced signal and noise for un-voiced.

While the source-filter-based systems are capable of producing quite intelligible speech, they have a distinct mechanical sound, and would not be mistaken for a human. This quality arises from the simplifying assumptions made by the model and therefore is not easily remedied.

2.3 Concatenative Synthesis

In recent years, much research has been focused on techniques for concatenative synthesis (Taylor et al. 1998). This technique uses an inventory of prestored units that are derived from real recorded speech. A database of speech is recorded from a single speaker

and is segmented to create an inventory of speech units. These units are concatenated together to create the synthesized output. The speech units differ between systems. One popular unit is diphones, which consist of parts of two phones, including the transitions between them (Peterson et al. 1958). The beginning and end of the diphone are positioned to be in stationary portions of the two phones. This reduces the problem of smoothing the junctions of concatenated units. There are also a manageable number of units, at most 1600 (40×40). The actual number is somewhat smaller, since not all possible pairs of phones occur in English. Context dependent phones, clustered by acoustic similarity, are also used. A search algorithm is used to find the optimal sequence of units (Hunt and Black 1996).

An extension of the small unit concatenative synthesis technique is to use variable sized units for concatenation (Yi and Glass 1998). In this method, task-specific sentences, phrases, and words are recorded from a single speaker. A search process is used to concatenate the largest units available in generating the output. In some cases, an entire prerecorded sentence or prompt can be used. Much of the time phrase level and word level units will be used. If a novel word is encountered, it is synthesized by concatenating phone sized units. This technique sounds very natural if word sized or larger units are being concatenated. The problem with this approach is that good coverage with large units is only possible in limited domains, and a new database must be recorded for each new domain. Work is in progress to unify the small and variable unit approaches to give a more general system.

See also: Speech Perception; Speech Production, Psychology of

Bibliography

Bruckert E, Minow M, Tetschner W 1983 Three-tiered software and VSLI aid developmental system to read text aloud. *Electronics* **56**: 133–8

Fant C G M 1960 *Acoustic Theory of Speech Production*. Mouton, The Hague, The Netherlands

Hunt A, Black A 1996 Unit selection in a concatenative speech synthesis system using a large speech database. *Proc. ICASSP-96* **1**: 373–6

Jelinek F 1997 *Statistical Methods For Speech Recognition*. MIT Press, Cambridge, MA

Jurafsky D, Martin J H 2000 *Speech and Language Processing*. Prentice-Hall, Englewood Cliffs, NJ

Klatt D H 1980 Software for a cascade/parallel formant synthesizer. *Journal of the Acoustics Society of America* **3**: 971–95

Pellom B 1998 Enhancement, segmentation, and synthesis of speech with application to robust speaker recognition. PhD thesis, Duke University, Durham, NC

Peterson G E, Wang W, Sivertsen E 1958 Segmentation techniques in speech synthesis. *Journal of the Acoustics Society of America* **8**: 739–42

Taylor P, Black A, Caley R 1998 The architecture of the festival speech synthesis system. *3rd ESCA Workshop on Speech Synthesis*. pp. 147–51

Yi J R W, Glass J R 1998 Natural-sounding speech synthesis using variable-length units. *Proc. ICSLP-98* **4**: 1167–70

W. Ward

Speed Reading

Modern society requires individuals to process an increasing amount of written information on a daily basis. The rising stacks of to-be-read books, articles, and documents often induce people to turn toward the commercially available courses that claim to radically increase reading speed. Most of these speed-reading courses claim that reading speed can be increased to 1,000 words per minute (wpm) with improved comprehension. This rate is over three times normal reading speeds (200–300 wpm), and some speed readers have claimed to read up to 100,000 wpm. Typical speed-reading training involves instructing readers to focus on fewer sections of the printed material, to refrain from rereading words and from lip movement or inner speech, to attempt to perceive meanings at a glance, and simply to concentrate on reading faster. While there is an abundance of anecdotal claims from individuals who report that they are able to read at amazing rates, the scientific evidence is not as convincing. The empirical studies that have supported speed-reading claims have been fraught with methodological problems. Those studies that are better designed have generally found that speed reading considerably inhibits comprehension and is comparable to scanning or skimming. This article reviews the history of speed-reading, the validity of speed-reading claims, evidence that refutes speed-reading claims, and some potentially positive contributions of speed-reading methods.

1. Brief History of Speed-reading

Increased fascination with anecdotes of speed reading emerged in the late nineteenth century popular press. This interest led to the first known speed-reading course, which was offered at Syracuse University in 1925. After a brief hiatus during World War II, the use of machines became popular to increase reading speed by modifying readers' eye movements. Normal reading involves rapid, jerky eye movements, called *saccades*. Words are only perceived during *fixations* between saccades. Ordinarily, readers make forward saccades; however, backward saccades called *regressions* are also made to fixate previous words. The number of letters or words that are processed during a single fixation is called the *perceptual span* (see *Eye*

Movements in Reading). Poor readers tend to make more regressions, and more frequent fixations (see *Reading Skills*). Hence, speed-reading advocates assumed that reducing regressions and fixations (with increased perceptual spans) would improve reading rate and consequently, reading ability. The focus after World War II on using mechanical devices to change readers' eye movements emerged from the success of using tachistoscopes (machines that flash images at controlled rates on a screen) to train pilots to more quickly and accurately identify aircraft. Similarly, tachistoscopes were used to train readers to reduce the number of fixations and widen the perceptual span. The results from these methods were initially promising. However, it eventually became evident that the effects of training were temporary and did not generalize to normal reading situations.

In the late 1950s, Evelyn Wood popularized speed-reading with her hand-pacing method, which involved training the eyes to follow a zigzag motion down each page. Combined with other reading and learning techniques, she introduced Reading Dynamics in 1959, which ultimately led to Evelyn Wood Reading Dynamics Institutes worldwide. The variety of speed-reading courses increased in the 1960s along with an abundance of anecdotal claims of their success.

2. *The Validity of Speed-reading Claims*

Are there individuals who can read at 1,000 wpm, or faster, and still comprehend at least 75 percent of the information? With its increasing popularity in the 1960s, many researchers began to challenge the astounding claims made by speed-reading advocates. Although many individuals have claimed to possess this ability, the empirical evidence was scant. Two principle areas of debate emerged in the literature. The first area of debate concerns the possibility of modifying readers' eye movements and perceptual span. The second issue concerns the methodology employed by speed-reading studies. The empirical research in support of speed-reading has tended to be fraught with methodological weaknesses, generally with respect to the measurement of comprehension during speed-reading tests.

2.1 *Eye Movements*

Many speed-reading courses attempt to modify readers' eye movements by increasing the perceptual span and reducing the number of fixations and regressions. However, researchers have argued that eye anatomy constrains the perceptual span to approximately three words from the fixation point; and more specifically, 3–4 letter spaces to the left and 14 letter spaces to the right of the center of an advanced reader's fixation point (Rayner 1986). The physiological boundaries of the perceptual span imply that reading rates exceeding 600–800 wpm are physiologically impossible if the reader is attempting to read all of the words (Taylor 1965). For example, a rate of 10,000 wpm would allow about 2.5 seconds per page of text with 7–8 fixations. Thus, each fixation would have to include a span of 55 words, which is impossible.

In response, speed-reading advocates have claimed that it is unnecessary to read every word of a passage, and that readers can learn to fixate only important words or sentences. In support of that claim, research has indicated that speed-readers make inferences between parts of the text by using prior knowledge (Carver 1983). That is, knowledge allows readers to develop an understanding of the text, despite not having read all of the words. This process is more successful when readers have more knowledge about the text, but virtually impossible without relevant knowledge. Indeed, most speed-reading courses acknowledge that speed-reading is only effective for familiar material.

Nevertheless, there is no guarantee that knowledge-based inferences will be correct, and evidence does not support the assumption that speed-readers learn to read only the most important words or sentences of a passage (Just and Carpenter 1987).

Many speed-reading courses also encourage readers to eliminate the tendency to make regressive eye movements. However, research has shown that regressions are virtually unavoidable with unfamiliar or complex material, and are adaptive when the segments that are reread are relevant to the reader's goals. Moreover, the relationship between eye movements and reading efficiency remains controversial and there is no available evidence that eye movement training improves reading ability. Although speed-reading training may alter eye movement patterns, these patterns revert to normal when comprehension is emphasized.

2.2 *Comprehension*

The second area of debate has centered on the comprehension of rapidly read material. There is little doubt that readers can turn pages quickly and scan printed material. The question is whether speed-readers can do so and maintain comprehension of the information. Speed-reading advocates have claimed that comprehension is not only maintained but improved. Others have claimed that speed-reading results in large comprehension decrements. The answer to this debate has been clouded by methodological problems (Carver 1990).

Much of the evidence in support of speed-reading has been anecdotal. In these cases, neither documentation nor comparisons to normal readers are generally provided. In addition, many empirical studies of speed-reading are flawed because comprehension was not assessed, or comprehension and rate were measured on separate passages. Without correspond-

ing evidence of comprehension and reading rate, a speed-reader is apt to 'read' rapidly when rate is measured (with low comprehension of the material) and read more slowly when comprehension is tested.

A second weakness of speed-reading research has stemmed from the use of comprehension questions that can be answered with reasonable accuracy without reading the passage. Generally, passages used to measure reading rate cover relatively familiar information. Consequently, passage-independent questions that can be answered on the basis of reasoning or prior knowledge inflate estimates of reading comprehension. For example, Carver (1972) demonstrated that previous estimates of speed-readers' comprehension (see Wood 1963) dropped from 68 percent to 17 percent for fictional material after correcting for passage-independent questions. Hence, tests showing that speed-readers maintain comprehension may not account for questions that can be answered without reading the material.

Similarly, the common use of multiple-choice questions to assess comprehension allows readers to guess the correct answers among the alternatives. For example, if two readers answer 20 of 30 questions correctly, but one reader provides answers for only 20 (missing 0, but skipping 10) and the other provides answers for all 30 questions (missing 10), the latter reader is showing greater evidence of guessing. Without correcting for guessing (see Carver 1990), the readers seem statistically equivalent in terms of their comprehension levels. This issue is particularly important because speed-readers often answer more questions, and guess more, than normal readers.

An additional weakness of speed-reading research concerns practice effects on the tests. When the same test or similar tests are used to assess comprehension before and after speed-reading training, the speed-reader may improve from having completed the test previously and not from training. Likewise, similar tests during and after training can lead to improvement on the post-training tests. These testing problems are most important for studies lacking a control group.

In summary, typical shortcomings of speed-reading research include: the lack of a comprehension test for the material used to measure reading rate, the use of passage-independent comprehension questions, not correcting comprehension performance for guessing, and the lack of a control group to measure practice effects on the tests.

3. *Evidence that Refutes Speed-reading Claims*

The bulk of empirical research has led to the conclusion that reading rapidly and maintaining comprehension is not possible. For example, Homa (1983) assessed two graduates of the American Speed-reading Academy who had reported reading rates above 100,000 wpm. He administered two visual identification tasks (letters and words) and a comprehension task. The speed-readers performed comparably to six undergraduates on the visual identification tasks but showed a greater tendency to guess the answers than the undergraduates. The speed-readers averaged 15,000 and 30,000 wpm on the comprehension task, but clearly understood little of what they had read, even after three attempts with the same material and the same tests. Homa concluded, 'the only noteworthy skill exhibited by the two speed-readers was a remarkable dexterity in page-turning.'

Carver (1985) examined the reading rate and comprehension of four groups of superior readers: speed-readers, professionals, high-skilled college students, and people who scored exceptionally high on tests such as the SAT. Carver found that speed-readers were similar in ability to other superior readers, except that they typically chose to skim and accept the lowered comprehension at rates higher than 1,000 wpm. The evidence strongly suggested that reading rates exceeding 600 wpm, with adequate comprehension, were not possible.

Just and Carpenter (1987) compared the reading abilities of 11 graduates of a Reading Dynamics course, 12 skimmers, and 13 normal readers. On two passages, the speed-readers, skimmers, and normal readers averaged 700 wpm, 600 wpm, and 240 wpm, respectively. The speed-readers' and skimmers' accuracy on comprehension questions was comparable, though speed-readers more accurately answered questions concerning the gist of an easier *Reader's Digest* passage. Both groups showed consistent comprehension decrements (7–15 percent) compared to the normal readers. Just and Carpenter concluded that reading speed can increase, but not without sacrificing comprehension.

4. *Potential Advantages of Speed-reading*

Although reading rates exceeding 600 wpm clearly impede comprehension, research has also indicated that many individuals read more slowly than necessary and that moderate increases in reading rate may improve comprehension. In addition, skilled readers show greater flexibility and efficiency by adjusting reading rate to their purpose and to the difficulty or familiarity of the material. However, it is debatable whether readers can be trained to be more flexible and to focus solely on pertinent information. Nevertheless, knowing when to read rapidly is considered an important skill in most reading programs.

5. *Conclusion*

The available research leads to the conclusion that speed-reading does not allow the reader to adequately understand the reading material. A distinction should be made between speed-reading, which assumes that normal comprehension is maintained, and scanning

and skimming, which accept a considerable decline in comprehension levels. Many researchers claim that speed-reading is essentially skimming. Nevertheless, proponents of speed-reading remain steadfast in their claims that speed-reading is effective. Research to support these claims would have to carefully control comprehension measurements and show equivalent or improved understanding of material read at rates exceeding 1,000 wpm. At this juncture, however, evidence suggests that *you get what you read*, and reading takes time.

See also: Eye Movements in Reading; Readability, Applied Psychology of; Reading, Neural Basis of; Reading Skills

Bibliography

Brozo W G, Johns J L 1986 A content and critical analysis of forty speed-reading books. *EDRS* (ED 270724)

Carver R P 1972 Comparisons among normal readers, speed-readers, and clairvoyant readers. *Yearbook of the National Reading Conference* **21**: 150–5

Carver R P 1983 Is reading rate constant or flexible? *Reading Research Quarterly* **18**: 190–215

Carver R P 1985 How good are some of the world's best readers? *Reading Research Quarterly* **20**: 398–419

Carver R P 1990 *Reading Rate*. Academic Press, San Diego, CA

Harris A J, Sipay E R 1985 *How to Increase Reading Ability*, 8th edn. Longman, New York

Homa D 1983 An assessment of two extraordinary speed-readers. *Bulletin of the Psychonomic Society* **21**: 123–6

Just M A, Carpenter P A 1987 *The Psychology of Reading and Language Comprehension*. Allyn & Bacon, Boston

Rayner K 1986 Eye movements and the perceptual span in beginning and skilled readers. *Journal of Experimental Child Psychology* **41**: 211–36

Taylor S E 1965 Eye movements in speed-reading: Facts and fallacies. *American Educational Research Journal* **2**: 187–202

Thompson M E 1985 Dimensions of speed reading: A review of research literature. *EDRS* (ED 263530)

Wood E N 1960 A breakthrough in reading. *Reading Teacher* **14**: 115–17

Wood E N 1963 Opinions differ on speed-reading. *NEA Journal* **46**: 44–6

D. S. McNamara

Spelling

In an alphabetic writing system, spelling refers to the norms or conventions by which the characters of the alphabet are arranged to form words. The term 'orthography' is also used to refer to a spelling system, especially within linguistics and literary studies.

The principle of alphabetic writing means that the characters used for writing correspond in some way, perhaps only roughly, to the individual phonemes of the language represented. In theory, it seems, spelling could be a simple matter of representing each phoneme by its corresponding alphabetic character. There are several reasons why this is not so, however.

(a) Spelling systems often, incorporate morphophonemic and/or grammatical information as well as phonological. They are rarely if ever simply representations of sounds.

(b) Spelling systems often are adaptations of spelling systems originally designed for other languages. This may result in some phonemic distinctions going unrepresented (or, conversely, subphonemic distinctions being indicated in spelling).

(c) The existence of a 'spelling system' implies a degree, often a high degree, of standardization. However, in every language, the written language represents an abstraction away from spoken varieties with potential variation along social, regional, and possibly other dimensions. There is always some mismatch between spoken words (even in a standard variety) and their written form.

(d) Over time, languages undergo grammatical and phonological change, but the spelling of the written language often remains conservative. Once established and conventionalized, a spelling system does not have to change at the same rate as the language it represents.

(e) Spelling systems, like other symbolic systems, are not purely a technical device but are subject to elaboration as part of the culture of their users. Spelling may communicate various kinds of social and ideological meaning as well as phonological information.

Thus, it is very rare for spelling systems to be wholly phonemic (or 'phonetic' as this term is often used outside linguistics); rather, they represent a complex of phonological, grammatical, historical and cultural information–only some of which may be apparent to their users on the occasion of use.

1. The Underlying Principles of Spelling Systems

Fundamental to spelling systems is an alphabetic principle, that is to say, that a finite, small number of symbols (typically less than 50) or 'letters of the alphabet' are used in combination to write linguistic units (morphemes or words). Alphabetic writing systems normally have discrete symbols to represent vowels and consonants (though some, like Arabic and Hebrew, do not represent vowels), and the alphabet typically has a fixed order (A, B, C, … etc.).

Closely associated with the alphabetic principle, at least historically, is the phonemic principle, which prescribes that there should be a correspondence between graphemes (letters or letter combinations) and phonemes of the language. While some words in some languages may be spelt more or less phonemically, it is rare for languages to be strictly phonemic in

their spelling systems. However, it is usual that, with a few exceptions, each letter in isolation has at least one phoneme of the language associated with it. In combination with other letters, it may keep its association with that phoneme or be associated with others. In English, for example, the ⟨c⟩ letter in isolation may be associated with /k/ or/s/ (although it cannot be pronounced without combining it with a vowel). In combination with ⟨h⟩ as ⟨ch⟩ it is associated with /k/ (as in 'ache') or /č/ as in cherry.

It is common for alphabetic spelling systems to observe a morphophonemic principle which preserves the spelling of specific morphemes even when they undergo conditioned phonological changes, Thus in English, the past tense ending is spelt ⟨ed⟩ even where, as in 'sailed,' it is pronounced [d], or, as in 'hoped,' it is pronounced [t]. Similarly, the unity of the German *Bild* [bIlt] 'picture' and *Bilder* [bildəR] 'pictures' is preserved by observing a morphonemic rather than a phonemic principle, as the latter would require the singular to be written as *Bilt*. As the change from /d/ to /t/ is predictable (because word-final plosives are always unvoiced) it is not reflected in the spelling.

An important cause of deviation from the phonemic principle in spelling systems is the operation of a historical principle, which in some cases might be called simply orthographic conservatism. The initial ⟨k⟩ in 'knee,' for example, has not been pronounced in most dialects of English for several centuries. Spelling systems are notoriously resistant to change, while phonological systems, by comparison, are not.

Another potential source of apparent inconsistencies is the etymological principle, typically applied to loan words from another language which shares the same alphabet. Thus in Dutch, the French loan-word *cadeau* ('gift') may be spelt ⟨*cadeau*⟩ (respecting the etymological principle) or ⟨*kado*⟩ which respects the phonemic principles inherent in the conventional spelling of native Dutch words. Likewise, considerations of etymology in English led to a class of words, coined on the basis of ancient Greek, in which /f/ is spelt as ⟨ph⟩: telephone, philosophy.

The historical principle and etymological principle together account for a phenomenon which is particularly apparent in English, though observable in many languages: the hybrid nature of the spelling system. English, though its inclusion of Anglo-Saxon, French, and Classical Greek and Latin words in the written language, with each set of lexis following the spelling tradition of the language in which it originated (the etymological principle) has become a hybrid system preserving (by the historical principle) a mixture of orthographic traditions, in particular the Anglo-Saxon and Norman French traditions (see Scragg 1974 for the history of English spelling). The longer the written tradition of a language, the more likely it is that the spelling system will be a hybrid showing the influence of different traditions. A further consequence of the hybrid nature of a spelling system like English is that it may become in part logographic, i.e., words which sound alike are distinguished by their 'shapes' or spellings (for example 'sole,' 'soul').

2. *Reading, Writing, and Spelling*

As pointed out by Augst (1986a, p. 26), 'no alphabetic script in the world is a representation of sounds and thus identical to a transcription.' Only the most naïve accounts of spelling see it simply as a matter of letters representing sounds. Many accounts, however, see spelling as a system of correspondences between phonemes (i.e., the potentially contrastive sound units of a language) and graphemes (written units, which could include single letters or letter combinations like ⟨sh⟩ in English representing the phoneme /š/). Such models assume that the processes of both reading and writing require access to the phonological level of representation of the word or morpheme. Augst suggests that a better model would be a trilateral sign, consisting of a representation of meaning, a phonemic schema and a graphemic schema. The reader could go directly from the graphemic schema (i.e., the written word on the page) to the meaning by recognizing the visual image, or vice versa in the case of writing. For words where the visual image was not familiar (or not familiar enough), the reader/writer would be able to access the phonemic schema—in lay terms, to 'sound out' the word by relying on the known correspondences between phonemes and graphemes. A common word like 'cat' might be read without accessing the phonemic schema, even though the phoneme-grapheme correspondence is exact; but an unusual word like 'tesla' might be read by matching graphemes to phonemes, and might be written 'phonetically' (and thus perhaps misspelt, e.g., 'tessler' or 'tezla' by a user who had heard the word but had not seen it in print.

The precise relationship between reading, writing, and spelling remains controversial. Which theoretical model is adopted clearly has consequences not just within linguistics, but for pedagogical theories as well, since in some countries pupils expend much time and effort to learn to spell (see Sect. 6).

3. *The Historical Origins of Spelling Systems*

The first written texts in a language are usually the work of writers who already know and use the written form of another language—one which already has a literate tradition. The conventions of writing are typically introduced from one culture to another through the mediation of a class of literate bilinguals. In terms of spelling, this means that the conventions associated with writing one language are transferred to a new language, for which they may or may not be well suited at the time. Spelling traditions therefore

tend to reflect social and historical movements and power relationships.

For example, the contemporary orthography of Manx (Sebba 1998a) reflects its origins in religious texts produced by a class of bilingual clergy who were the first writers of the language. More generally, as the vernaculars of Europe were transformed into written languages, the spelling conventions of Latin, in various forms, became part of the spelling traditions of the newly written languages.

During the era of imperial and colonial expansion, literacy was introduced to many languages using spelling systems based on the language of the colonial power. In many cases, the agents of literacy were missionaries or other clergy, or colonial administrators. Lepsius's (1863/1981) work set out to provide a 'standard alphabet' suitable for all languages, whereby the same symbol consistently represented the same sound (e.g., the symbol ⟨y⟩ for the [j] sound of English 'yes' and German *ja*). His work was the forerunner of phonemic orthographies introduced to many languages by trained linguists, both secular and missionary.

One issue debated by linguists providing 'new' orthographies for indigenous languages is whether such systems should reflect the spelling conventions of the local official language (e.g., Spanish in the case of language of Central America), enabling an easier transition to the dominant language of literacy—or more 'universal' principles, which may foster more independence for the language. See Barros (1995) for the history of this debate.

Languages with relatively short histories of literacy are often unfortunate enough to have several competing spelling systems, in spite of a dearth of printed literature. This may be the result of competing groups of missionaries (sometimes representing different linguistic philosophies or colonial powers) or competing political interests, or both (see Sect. 5).

4. *Spelling and Language Standardization*

The existence of a spelling system does not by itself imply standardization, but for languages which have gone down the road of grammatical and lexical standardization, the standardization of spelling seems to be an inevitable part of the process. In the histories of the main European languages of literacy, printers, publishers, and lexicographers have had a major part in this.

While English was successfully used as a written language for centuries even though the spelling system, was unstandardized, Stubbs (1980, pp. 44–5) points out how 'good spelling,' has become socially valued, while even the concept of a 'spelling mistake' in English is as recent as the last two centuries (Strang 1970, p. 107), following the publication of major dictionaries (*Johnson's*, in England, *Webster's*, in North America).

Publishers and printers have their own economic motives for wishing to standardize the forms of written words, and for English, had achieved this goal by about 1650, a century or so before spellings in private writing became standardised (Strang 1970, p. 107). It is not clear, however, whether the standardization of spelling produces gains in terms of reading efficiency. If reading is largely recognition of a visual image (see Sect. 2) then having just one image corresponding to each word might produce a gain in efficiency as fewer images will have to be stored and accessed. Standardization of spelling might then be expected to go hand in hand with mass literacy. This remains to be demonstrated however; and it does not fit with the fact that literacy in England was 'moderately widespread' in the two centuries before the industrial revolution (Strang 1970, p. 106), and then declined in percentage terms just as standardization approached completion.

Advocates of particular spelling systems (including the existing ones) rarely bother to justify the need for a standardized spelling. It is overwhelmingly taken for granted. This in itself is a good reason to ask why spelling standardization is necessary, and why such a high social value is put on being able to spell correctly.

5. *Spelling and Ideological Debate and Conflict*

Spelling systems, as a type of symbolic system, inevitably acquire social and cultural meanings which go far beyond their function as a means of representing a language. The circumstances of the introduction of a spelling system and the circumstances surrounding its use give rise to this cultural symbolism which may in turn lead to ideological contestation and conflict, with the spelling system as its focus. Some recent examples include the debate over an appropriate spelling system for Haitian Creole (Schieffelin and Doucet 1994), where a key issue was whether to use a 'gallicising' orthography, where certain letters had the same value as in French, or a more international one, where they did not. The choice between letters such as ⟨k⟩ and ⟨c⟩ was reinterpreted in the context of the debate by some protagonists as representing 'Anglo-Saxon imperialism' (⟨k⟩) vs. 'colonial subservience' (⟨c⟩).

A similar type of conflict exists in Galicia (Alvarez-Caccamo and Herrero Valeiro 1996), where the choice of orthography for Galician reflects a writer's view on whether Galician is a variety of Portuguese or an independent language. Those favoring 'reintegrationism' (with Portuguese) use mainly Portuguese spelling conventions; those favoring 'differentialism' use conventions closer to Spanish.

In less extreme cases, particular spelling systems may be unpopular because of their associations: for example, one orthography for Breton is disliked for its associations with Nazi collaborators during the World War II (McDonald 1989, p. 131ff).

Even where not codified, orthographic practices may represent ideological positions, especially a desire to differentiate. Thus Caribbean writers may write *kool*, *tuff* to distance Creole from English, although there is no significant pronunciation difference (Sebba 1998b). Within an Anarchist tradition in Spain, words such as *Okupación* ('Occupation (of buildings)') are spelt with ⟨k⟩ instead of ⟨c⟩ in graffiti, in defiance of the official spelling.

6. Spelling and Educational Policy

Most languages with a written tradition now have highly standardized spelling. Thus spelling is usually seen as one of the literacy skills which have to be acquired through formal schooling, with 'good spelling' taken as a sign that the writer is well educated. Education authorities differ, however, in the emphasis placed on the accomplishment of 'correct' spelling, as measured by the amount of school time devoted to it, and the penalties for failing to achieve the acquired standard. There are differences between Britain and the US, for example, in this respect, with Britain usually being seen as having a more tolerant attitude, though there have been recent policy changes. An insistence on 'correctness' in spelling and substantial penalties for underachieving are usually associated with political conservatism, with a more relaxed approach being seen as 'liberal,' in keeping with prescriptive approaches to other aspects of language.

In a few places, such as Norway, there are alternative spelling standards which take into account the dialectal preferences and competences of the local population (see Wiggen 1986). This is unusual; learners are much more likely to be subject to a national or supra-national spelling standard.

While there is almost universal acceptance of the notion that 'correct spelling' is a necessary educational outcome, for some languages there is a feeling that spelling is too difficult, and thus presents a barrier to educational achievement. This provides a major impetus for spelling reform, discussed in Sect. 7.

7. Spelling Reform

Spelling reform—usually meaning an adjustment to an existing system, rather than a root-and-branch rethink—is invariably controversial. Proposals for spelling reform are usually put forward when the current conventions are seen to be causing a 'problem'—typically an educational one. A commonly identified source of problems is language change, which results in the phonemic principle applying over a decreasing part of the lexicon; words where the spelling is out of step with the pronunciation are seen as causing difficulties for learners and leading to 'illiteracy.'

For English, the widespread perception of the spelling system as 'complex' and 'illogical' has led to many proposals for reform over the years, most of which have had no impact on actual spelling practices (see Carney 1994 for discussion). In England, the banner of spelling reform is carried by, among others, the *Simplified Spelling Society*. A few languages, such as Norwegian, have regular and fairly uncontentious reviews of spelling practices, carried out by a central authority. Other languages have implemented reforms, though seldom without resistance from the general public.

Recent examples of contentious reforms are those of Dutch (Jacobs 1997), Portuguese (Garcez 1995), and German, each of which required international negotiation. Reforms often founder as a result of conflicting interests: the general public (which nearly always seems to be opposed to changes), educationists (who are more likely to be in favor), publishers and lexicographers (who have a vested interest in certain types of change, which may produce increased demand from consumers), authors (who sometimes oppose change on the grounds that it will make earlier literary works unavailable to readers who do not know the pre-reform spelling), and national governments, who have to try to balance these opposing parties and promote their own interests. The outcome is often a set of reforms which satisfies no one, and the road to reform is often extremely rocky. The German reforms of the 1990s, for example, in spite of affecting only 0.05 percent of words in running text, have been subject to successful challenges in some regional courts, although these have been overturned by the Federal Constitutional Court (Johnson 2000).

See also: Dyslexia, Developmental; Dyslexia: Diagnosis and Training; Dyslexia, Genetics of; Reading Nonalphabetic Scripts; Reading Skills; Writing, Evolution of; Writing Instruction; Writing Systems; Writing Systems, Psychology of

Bibliography

Alvarez-Caccamo C, Herrero Valeiro M J 1996 O continuum das normas cscritas na Galiza: do espanhol ao português. *Agália: revista da Associaçom Galega da Lingua*, pp. 143–56

Augst G 1986a Descriptively and explanatorily adequate models for orthography. In: Augst G (ed.) *New Trends in Graphemics and Orthography*. Mouton De Gruyter, Berlin, pp. 25–42

Augst G (ed.) 1986b *New Trends in Graphemics and Orthography*. Mouton De Gruyter, Berlin

Barros C D M 1995 The missionary presence in literacy campaigns in the indigenous languages of Latin America. *International Journal of Educational Development* **15**(3): 277–87

Carney E 1994 *A Survey of English Spelling*. Routledge, London

Garcez P M 1995 The debatable 1990 Luso-Brazilian orthographic accord. *Language Problems and Language Planning* **19**: 181–78

Jacobs D 1997 Alliance and betrayal in the Dutch orthography debate. *Language Problems and Language Planning* **21**(2): 103–118

Johnson S 2000 The cultural politics of the 1998 reform of German orthography. *German Life and Letters* **53**(1): 106–25

Lepsius R 1981 *Standard Alphabet for Reducing Unwritten Languages and Foreign Graphic Systems to a Uniform Orthography in European Letters*. 2nd revised edn. (London 1863), Kemp J A (ed.) John Benjamins, Amsterdam

McDonald M 1989 *'We are not French!' Language, Culture and Identity in Brittany*. Routledge, London

Schieffelin B B, Doucet R C 1994 The "real" Haitian Creole: Ideology, metalingusitics and orthographic choice. *American Ethnologist* **21**: 176–200

Scragg D G 1974 *A History of English Spelling*. Manchester University Press, Manchester, UK

Sebba M 1998a Orthography as practice and ideology: The case of Manx. Lancaster University Department of Linguistics and Modern English Language, *Working Papers of the Centre for Language in Social Life*, No. 102

Sebba M 1998b Phonology meets ideology: The meaning of orthographic practices in British Creole. *Language Problems and Language Planning* **22**(1): 19–47

Strang B M H 1970 *A History of English*. Methuen, London

Stubbs M 1980 *Language and Literacy: The Sociolinguistics of Reading and Writing*. Routledge and Kegal Paul, London

Wiggen G 1986 The role of the affective filter on the level of orthography. In: Augst G (ed.) *New Trends in Graphemics and Orthography*. Mouton De Gruyter, Berlin, pp. 395–412

M. Sebba

Spencer, Herbert (1820–1903)

Herbert Spencer was born on April 27, 1820 in Derby, UK, and after a long and distinguished career as a major intellectual, he died on December 8, 1903 in Brighton, UK. During Spencer's childhood, he was privately tutored, first by his father and then, after the age of 13, by his uncle in Bath. Thus, Spencer received little formal education outside of his family, although his in-depth tutelage in mathematics and science was sufficient for him to contemplate as a relatively young man a grand Synthetic Philosophy in which a few general or 'first principles' (1862) could be used to explain the operative dynamics of physical science, biology (1864–67), psychology (1855), sociology (1874–96), and ethics (1892–98). He also was a moral crusader, writing important philosophical statements on such topics as education (1860) and social well-being (1850, 1892–98).

Spencer's sociology comes late in his career, beginning with a methodological treatise on sources of bias (Spencer 1873), continuing with an effort to systematically array cross-cultural and comparative data (Spencer 1873–34), and culminating with the use of data to illustrate the basic laws of sociology (Spencer 1898b [1874–96]). In this article, the goal is to see how Spencer's grand scheme developed and how it evolved into a profoundly important sociology [For a comprehensive listing of primary works by Spencer and of works on Spencer, see Perrin (1993)].

1. *Biographical Influences on Spencer's Sociology*

Several important aspects of Spencer's biography are important in examining his impact on social science. First, he never held an academic position, and hence, he did not have any students to carry forth his work after his death, although the wealth inherited from his uncle and from his book sales did allow him to commission other academics to perform research and, even after his death, to continue to completion the volumes of his *Descriptive Sociology* (Spencer 1873–34). Second, he lived an extremely spartan life, and yet at the same time, he interacted in London Clubs and various assemblies with the elite scientists and intellectuals of his time. Indeed, it is from these scholars—figures such as Thomas Huxley, Joseph Hooker, John Tyndall, and even Charles Darwin—that Spencer absorbed much of his knowledge about current discoveries in science. Third, Spencer was immensely popular in his time. His works generally came out in serial form in magazines and private subscriptions, later being packaged as whole books which sold more than 400,000 copies during Spencer's lifetime. Yet, despite his enormous popularity, Spencer was an enigma because he had no formal academic affiliation and because he lived a very private and, it appears, rather neurotic life punctuated by hypochondria. Yet, in all his works, several key influences are evident. Reading Thomas Malthus, Karl Erst von Baer, and Charles Darwin gave Spencer a view of the universe as evolving from simple to more complex through competition; extensive reading of the physical sciences gave Spencer a post-Newtonian conception of science, a view emphasizing that universal laws could explain all domains of the universe; and reading of Auguste Comte also gave Spencer a view of sociology as the search for general laws as well as a conception of social organisms as having structures devoted to meeting basic needs and of societal evolution as involving an increasing complexity of structures meeting these needs. Spencer (1968 [1864]) was, however, always rather defensive about perceptions of Comte's influence on him. Out of this mix of influences came Spencer's grand Synthetic Philosophy that was to unify physical science, psychology, biology, sociology, and ethics under first principles borrowed from the physics of Spencer's time.

2. *The Stigma of Spencer's Moral Philosophy*

Spencer's moral philosophy is one reason that his intellectual star began to fade in the early twentieth century. In his first major work, *Social Statics* (1850), almost a decade before Darwin published *On the Origin of Species*, Spencer coined the phrase 'the survival of the fittest' that was to haunt him for the rest of his career. The more general philosophy in this work simply argued that individuals should be free to pursue their happiness, as long as they do not abridge the rights of others to do the same, although there was a hard edge to this philosophy emphasizing that the less fit should not be supported by the state. Spencer (1904) complained bitterly in his autobiography about how he had become type cast as a result of this first book, but it is hard to feel sorry for him because he made the same 'libertarian' (in today's vocabulary) statements in his last work, *The Principles of Ethics*. For a discipline such as sociology, which is liberal in the twentieth century as opposed to the nineteenth century sense where freedom of the individual and lack of governmental activism are emphasized, this libertarian-sounding ideology has always worked to obscure the important contributions of Spencer to sociology.

3. *The 'First' or 'Cardinal' Principles*

Spencer's use of the physics of his time to formulate his 'cardinal principles' is now rather dated. Spencer postulated three basic principles: (a) the indestructibility of matter, (b) the continuity of motion in a given direction, and (c) the persistence of force behind movement of matter. From these general principles, three corollaries were derived: (d) the transferability of force from one type of matter and motion to another, (e) the tendency of motion to pass along the line of least resistance, and (f) the rhythmic nature of motion. From these principles, Spencer (1880[1862], p. 286) formulated his famous law of evolution as 'change from a less coherent form to a more coherent form, consequently on the dissipation of motion and integration of matter.' This law was presumed to apply to the evolution of celestial bodies, chemical compounds, organisms, and superorganic bodies (organization of organisms). Spencer was, therefore, the world's first general systems theorist.

The first or cardinal principles are less interesting than what Spencer did with them as he applied them to organic and superorganic evolution. In *Principles of Biology*, Spencer used the first principles to formulate the cornerstone of his biology and sociology: increasing size of an organic or superorganic mass requires an exponential differentiation of the structure necessary to support the larger mass. This principle is still used in biology, although its origins to Spencer are long forgotten; the principle is employed in sociology to study differentiation at all levels of social organization, with only a vague sense that this idea comes from Spencer. In many ways, this theoretical cornerstone of Spencer's Synthetic Philosophy actually comes more from his training and brief practice as an engineer than it does from his first principles; for within engineering it is well established that larger structures require more complex systems of support.

Spencer's star faded long ago in biology, and there is only a hint of his influence in education and philosophy today. Where Spencer still remains a figure deserving of at least mention, often in a somewhat derisive way, is in sociology. In order to appreciate what Spencer's legacy is today, we must review his contributions to sociology, where he still lives on as a marginal figure in sociology's pantheon of great masters that includes such giants as Karl Marx, Max Weber, Émile Durkheim, and George Herbert Mead.

4. *Spencer's Functionalism*

Spencer is famous for his 'organismic analogy' in which he lists in a more formal manner than Comte before him the parallel principles of organization between organic and superorganic bodies. But this analogy is often given too much emphasis in commentaries on Spencer because it hides a rather sophisticated functionalism. For Spencer, all superorganic systems composed of organic units must meet three basic requisites if they are to survive: (a) operation or (i) production of life-sustaining substances and (ii) reproduction of units; (b) regulation or coordination and control of units; and (c) distribution or movement of necessary information and substances among units. Populations which cannot meet these requisites will experience 'dissolution,' a notion first developed in *First Principles* (1880 [1862]). Thus, Spencer did not view evolution as an inevitable process of movement from simple to complex as populations sought to resolve problems of operation, regulation, and distribution; on the contrary, he recognized that societies often disintegrate. In this discussion of functional requisites, Spencer introduces two conceptions of 'selection.' On the one hand, Spencer argued that evolution of societies tends to increase complexity because, as societies come into conflict, the more complex and better organized generally will prevail, thereby ratcheting up the overall level of societal complexity. This argument is one of the first 'group selectionist' statements presented in either biology or the social sciences. On the other hand, there is another kind of selection which is at the core of all functionalist arguments. This selection occurs alongside Darwinian selection and revolves around problems and crises in production, reproduction, regulation, and distribution. When problems in any of these spheres emerge, they set into motion efforts to search for solutions by goal-seeking social units. Thus, social selection in superorganic systems is not blind because the very

units subject to selection can change their structure and thereby avoid being selected out. Spencer argues that most selection in sociocultural systems is of this second type, where problems of adaptation activate efforts to resolve crises with respect to production, reproduction, regulation, and distribution.

Since these requisites are the essential problems of survival in all superorganic systems, including both human and nonhuman societies, differentiation of structure during the course of evolution will revolve around these four axes. First will come differentiation of productive structures, followed by differentiation of reproductive units. This increased differentiation along these axes will generate selection pressures for their coordination and control, leading to the elaboration of political systems. As differentiation of operative and regulatory units occurs, new systems of distribution become necessary if dissolution is to be avoided. As these new systems of distribution emerge, e.g., markets, communication, and transportation infrastructures, they create new selection pressures for their regulation while at the same time providing positive feedback for further differentiation of production and, later, reproduction. Thus, movement of superorganic systems from simple to complex forms revolves around selection pressures along the axes of operation, regulation, and distribution as well as around the mutual feedforward and feedback effects of differentiation along one axis on the other axes. This kind of analysis is far more sophisticated than the typical commentary on functionalism acknowledges.

Spencer's substantive sociology also offers some very important models of system dynamics. One element of these models follows from the basic law: as a population grows, the increasing size generates escalated logistical loads for operation, regulation, and distribution which, in turn, set off selection pressures in the manner described above. Spencer's political sociology should be viewed in this context. For Spencer, once power is concentrated, this initial concentration will be used to gain even more regulatory control. As power becomes more concentrated, inequalities will escalate as those with power extract an ever-greater proportion of a population's resources. As inequalities increase, disintegrative pressures mount leading to the concentration of more power to suppress the potential conflict that always inheres in inequality; and as this process is iterated, system dissolution is likely. Geopolitics are also involved in these dynamics, in several senses. First, elites in the regulatory sector often engage in war to secure the resources of other populations which only increases the concentration of power and the scale and scope of inequalities. Second, such war-making efforts often distract subordinates in a system of inequality, at least for a time, from their true interests in changing the system of power generating inequality. Third, when power becomes too concentrated, resistance to this power inevitably increases, often leading to protest movements and other actions that lead to some deconcentration of power. Yet, the longer deconcentrated power lasts, problems of coordination and control escalate, creating selection pressures for reconsolidation of power. Societies thus cycle between concentrated profiles of power (what Spencer termed the 'militant') and deconcentrated power (what Spencer called 'industrial' profiles)—a line of analysis that predates Vilfredo Pareto's discussion of the circulation of elites.

5. *The Data Behind Spencer's Sociology*

These lines of argument have a very contemporary ring to them, and they can be found throughout Spencer's *Principles of Sociology*. Moreover, they are rather easily converted into laws or principles since this is how Spencer thought (see Turner 1985, for an effort to extract the principles from Spencer's work). One of the criticisms of Spencer's functionalism is that it is so speculative; while this may be so, Spencer also amassed enormous amounts of data to support his theoretical speculations. Indeed, the reason that Spencer's *Principles of Sociology* is such a long work is that it is filled with data. In fact, many scholars complained at its first publication in serial form and later as a three-volume book that the analytical ideas get lost in all the data. Thus, Spencer assembled data far in excess of almost any other sociologist in the history of the discipline. He hired professional academics to gather historical and ethnographic information, ordering it into categories that followed from his general functionalist scheme. There are many tragedies in the contemporary view of Spencer, but none is greater than the loss of what Spencer began in the 1860s and first published in 1873 as the inaugural volume of his *Descriptive Sociology* (the last volume was published in 1934).

Spencer was the first to execute what was later done by those assembling the Human Relations Area Files (HRAF) (e.g., Murdock 1953). Spencer recognized that in order to make generalizations about human societies, a large comparative database was necessary. But these data had to be arrayed so that they could be compared, leading Spencer to develop a category system for recording the data being assembled by professional academics. The result was the multivolume *Descriptive Sociology* (1873–34) which arrays data historically for those societies that have a written history and cross-sectionally for those less developed societies without a written history. The publication of these oversized volumes has been completely lost to modern social science, but if Spencer's lead had been followed by others in the nineteenth century, all social science would have a much firmer empirical base than is provided by HRAF. At the very time Spencer was assembling available information on 'primitive' societies, these societies were being changed or destroyed

by colonial forces, with the result that by the time anthropologists began to study them in detail in the 1930s, their pristine character had long been compromised. *Descriptive Sociology* could have been Spencer's greatest legacy, but like so much of his work, the contemporary world remains largely ignorant of his efforts to create a database for assessing the plausibility of his theories.

6. *What Might have Been, and What Can Still Be*

By the time of Spencer's death in 1903, his star had already begun to fade in intellectual circles. One reason for this decline can be attributed to Spencer as a lone scholar working outside of academia. Without students to carry on his work and without a school of thought institutionalized in academia, there was no one and no place to carry forward Spencerian sociology. Work still continued on *Descriptive Sociology* but this was much narrower in scope than Spencer's philosophy or even his sociology, and moreover, most of this work was contracted to scholars who were not committed to Spencer's grand scheme. So, when Talcott Parsons (1937) asked 'who now reads Spencer,' the answer was not very many scholars inside or outside of sociology. Spencer's star had fallen at the same time that evolutionary thinking came under attack in anthropology, and with no one to defend Spencer's sociology, he fell into obscurity. Yet, let us imagine for a moment what might have happened if Spencer had been defended by students or had established a school entrenched in academia like Durkheim's *année* school in France. In conducting this thought experiment, we can come to appreciate not only what might have been, but more fundamentally, what is still available in Spencer's legacy.

The first legacy would have been a Human Relations Area Files that came earlier, before colonialism completely destroyed indigenous populations. Moreover, sociology would have had a database that was comparative in several senses: the data would have been arrayed in common categories that are more sophisticated than George P. Murdock's (1953) HRAF; the data would have been both historical and ethnographic and would have allowed systematic comparisons of different types of societies at any point in time. Only in the last three decades has comparative-historical and ethnographic data been reintroduced back into sociology and, ironically, this reintroduction has been part of a movement to rekindle interest in the very evolutionary thinking that was used to discredit Spencer.

The second legacy is that the evolutionary paradigm would have survived and avoided the 50-year hiatus. Stage models have utility in sociological analysis, as has been demonstrated in the work of Gerhard Lenski (1966), Stephen K. Sanderson (1995), Jonathan Turner (1984, 1997), and others in recent years. Moreover, adopting Spencer's views on evolution would have avoided the view of evolution as inexorable, since dissolution would be a prominent point of analysis as would dialectical cycles within either evolution or dissolution.

A third legacy would have been a sophisticated functionalism that, unlike mid-twentieth century approaches, sought to articulate theoretical principles about the dynamics of social processes rather than simply cross-tabulate lists of functional requisites with lists of social processes. Functionalism fell into disuse not so much because it was dead wrong but because it was not done well. A reading of Spencer would have led the way to a functionalism that does not build conceptual schemes as much as highlight fundamental forces whose operative dynamics can be captured in theoretical principles.

A fourth legacy would have been a sociology that did not lose sight of its status as a science in which the goal is to generate theoretical principles that explain the operative dynamics of the social universe. Spencer's methodological treatise and his articulation of actual laws of human organization would have sustained the discipline better than Marx, Weber, Simmel, or Durkheim could, as various antiscience and excessively relativistic movements have come and gone for the whole of the twentieth century in sociology.

A fifth legacy would have been the actual substance of Spencerian sociology. Imagine if sociologists had mined *The Principles of Sociology* for insights into human organization with the same intensity as we have delved into Marx, Durkheim, Weber, or Mead. Spencer provided not only highly abstract theoretical principles but also many principles about the operation of stratification systems, polity, kinship, religion, economy, markets, ceremonies, professions, organizations, education, and many other key structures, and he documented these generalizations with copious amounts of historical, ethnographic, and for his time, contemporary data. The sociology of major institutions (e.g., family, education, religion,etc.) would be less parochial and descriptive than it is today because it would examine the dynamics of the institution with ethnographic as well as historical-comparative data organized by basic theoretical principles.

A sixth legacy would have been the continued connection between biology and sociology. This connection has only recently been revived along all the fronts found in Spencer's work. There is the ecological approach emphasizing Darwinian dynamics that Spencer saw before Darwin and that are now used to examine the relations between resources niches and a wide variety of social processes. Indeed, there is the revival of selection arguments in general (e.g., Runciman 1998), and moreover, an effort to move beyond Darwinian selection to considerations of group selection (e.g., Sober and Wilson 1998) and

Lamarckian processes so evident in Spencerian sociology (Turner 1995).

A seventh legacy is a political sociology that connects stratification processes within society to geopolitical dynamics in world systems. At the core of Spencer's sociology is a concern with power that is the equal of that to be found in either Marx or Weber. One of Spencer's main explanatory variables is how the concentration of power alters all other social processes, especially stratification and institutional systems, and his analysis of how war and commerce in the world system are implicated in power dynamics would not only have supplemented Marx and Weber but would have led to the emergence of world systems theory at the beginning rather than at the three-quarters mark of the twentieth century (e.g., Wallerstein 1974).

An eighth legacy is the emphasis on the fundamental relation between size, differentiation, and modes of integration. Much of this legacy was rediscovered in the mid-twentieth century in organizational studies on administrative intensity, but the basic dynamic operates at all levels of society. If sociologists had pursued Spencer's lead and focused in 1910 on this fundamental relationship in the social universe, not only would organizational studies been moved forward at an earlier point in history but also other social units might well have been analyzed for how these relations among size, differentiation, and integration play themselves out.

These eight lines of thinking that are part of Spencerian sociology were available to all sociologists by the early 1890s. In a sense we have squandered this incredible legacy, but it is not too late. The above thought experiment can be more than an intellectual exercise in the history of ideas. A reading of Spencer today can still inform sociology. There are gems of insight to be discovered and used to analyze societies in almost everything Spencer wrote. They simply sit there on library shelves, waiting to be used by scholars in the twenty-first century. There is probably little more to be gleaned from poring over the texts of others in sociology's early pantheon, but the neglect of Spencer for the whole of the twentieth century provides a rich resource for the twenty-first century.

See also: Conflict: Anthropological Aspects; Conflict: Organizational; Conflict Sociology; Cross-cultural Research Methods; Cultural Evolution: Theory and Models; Darwinism; Development: Organizational; Development: Social-anthropological Aspects; Elites, Anthropology of; Elites: Sociological Aspects; Ethics and Values; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolution: Optimization; Evolution: Self-organization Theory; Evolutionary Selection, Levels of: Group versus Individual; Functionalism, History of; Functionalism in Sociology; Groups, Sociology of; Hierarchies and Markets; Historical Change and Human Development; Individual/Society: History of the Concept; Motivational Development, Systems Theory of; Organizational Culture; Public Opinion: Political Aspects; Regulation, Economic Theory of; Religion: Definition and Explanation; Religion: Family and Kinship; Religion: Morality and Social Control; Social Evolution: Overview; Sociality, Evolution of; World Systems Theory

Bibliography

Lenski G E 1966 *Power and Privilege: A Theory of Social Stratification*. McGraw-Hill, New York

Murdock G P 1953 Social structure. In: Tax S, Eisely L, Rouse I, Voegelin C (eds.) *An Appraisal of Anthropology Today*. University of Chicago Press, Chicago

Parsons T 1937 *The Structure of Social Action*. McGraw-Hill, New York

Perrin R G 1993 *Herbert Spencer: A Primary and Secondary Bibliography*. Garland Publishers, New York

Runciman W G 1998 The selectionist paradigm and its implications for sociology. *Sociology*. **32**: 163–88

Sanderson S K 1995 *Social Transformations: A General History of Historical Development*. Blackwell, Cambridge, MA

Sober E D, Wilson S 1998 *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, Cambridge, MA

Spencer H 1873 *The Study of Sociology*. Kegan Paul, London

Spencer H 1873–1934 *Descriptive Sociology, or Groups of Sociological Facts*. D. Appleton, New York

Spencer H 1880 [1862] *First Principles*. A L Burt, New York

Spencer H 1887 *The Principles of Biology*, 2 Vols. D. Appleton, New York, originally published 1864–67

Spencer H 1888 *Education: Intellectual, Moral, and Physical*. D. Appleton, New York, originally published in 1861

Spencer H 1892 *Social Statics*. D. Appleton, New York, originally published in 1850

Spencer H 1892–98 *The Principles of Ethics*. D. Appleton, New York

Spencer H 1898a *The Principles of Psychology*. 2 Vols. D. Appleton, New York, originally published in 1855

Spencer H 1898b [1874–96] *The Principles of Sociology*. 3 vols. D. Appleton, New York, originally published between 1874–96

Spencer H 1904 *An Autobiography*. 2 Vols. Appleton, New York

Spencer H 1968 [1864] *Reasons for Dissenting from The Philosophy of M. Comte and Other Essays*. Glendessary Press, Berkeley, CA, originally published in 1864

Turner J H 1984 *Societal Stratification: A Theoretical Analysis*. Columbia University Press, New York

Turner J H 1985 *Herbert Spencer: A Renewed Appreciation*. Sage, Beverly Hills, CA

Turner J H 1995 *Macrodynamics: Toward a Theory on the Organization of Human Populations*. Rutgers University Press, New Brunswick, NJ

Turner J H 1997 *The Institutional Order*. Longman, New York

Wallerstein I 1974 *The Modern World System*. Academic Press, New York, Vol. 1

J. H. Turner

14898

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Sperry, Roger Walcott (1913–94)

Roger W. Sperry, scientist, teacher, philosopher, and humanitarian was born in Hartford, Connecticut, USA on August 20, 1913. He was married to Norma Dupree in 1949; they had 2 children. He died on April 17, 1994.

The diversity of Sperry's interests is not only reflected in his distinguished academic career, but also in his avocations. In addition to his major contributions to our understanding of brain structure and function, he was an unusually talented artist, sculptor, and paleontologist. His multiple talents across many areas attracted a constant stream of graduate students, postdoctoral fellows, and visiting scientists to his Psychobiology Unit at Caltech for four decades.

Sperry's scientific contributions centered on four major areas, the first of which began in the early 1940s, while he was still a graduate student in Paul Weiss' laboratory at the University of Chicago. Paul Weiss, whose views dominated the field of neural development at the time, held that the leading tips of growing neurons solely responded to local mechanical stimuli, and were in no way influenced by distant chemical cues. His numerous tissue culture studies had essentially closed the door on any possibility of chemical influences on neuronal growth. Weiss also held that the normal function of the nervous system was in no way dependent upon the precise patterns of neuronal connections. One example of this, according to him, was that when motor nerves were transplanted into any foreign muscle, normal movements would eventually occur after a period of training.

Sperry, as was to become his custom, challenged these ideas, and designed a typically simple, but ingenious experiment to test whether or not specific nerve connections are critical for normal function. Using rats, he transplanted motor nerves from leg muscles of one side to muscles of the opposite side, then tested their movements after the new nerve-muscle connections were established. Function was found to be severely impaired, and no amount of training could correct it (Sperry 1945). These results also had important implications for humans. Neurosurgeons at the time were transplanting nerves in wounded war veterans in attempts to help them regain digit and limb movements, which had been lost as a result of war injuries. In spite of extensive post-surgical training, these transplants met with little success. During his military service with the Office of Scientific Research and Development (1942–45), Sperry, along with Weiss, showed that surgical transplantation of nerves to antagonistic muscle groups followed by intensive training simply did not work because transplanted nerves were inherently different from the nerves which normally innervated those muscles. This finding resulted in significant modifications of treatment procedures, thus saving a great deal of time, frustration, and money.

During this same period it was reported that if the optic nerve were sectioned in amphibians, optic axons would regenerate into the brain, and normal function would be restored. Sperry immediately recognized that this finding offered the perfect opportunity to test his ideas regarding functional specificity and the possible role of chemical influences on neuronal growth patterns. He first confirmed the above results (Sperry 1950), then once again designed a brilliantly simple study—the now classic eye rotation experiment in the frog—to test his ideas. A hungry frog will accurately strike out with its tongue to obtain a fly held in front of it. Sperry sectioned the optic nerves of normal frogs, and rotated the eyeballs by 180 degrees. After regeneration had taken place and connections were reestablished in that part of the brain which is known to receive visual fiber input (an area called the optic tectum), he tested the frog's ability to capture a fly. The premise was that if the retinal axons had regrown into their normal sites of termination (in spite of the fact that the eye had been left/right reversed and turned upside-down), the frog should misdirect its strike by 180 degrees in any direction. If, however, the regenerating optic fibers disregarded their original target sites and established new central connections corresponding to the rotated eyes, the frog would be capable of striking accurately at the target fly. The answer was unequivocal. When the fly was held 90 degrees to the frog's right, it struck out 90 degrees to its left—180 degrees out of phase. The same result was obtained for all directions—left, right, up, down. The strike was always 180 degrees off, and no amount of retraining could correct it. The conclusion from both of the above lines of evidence was clear: peripheral nerves cannot be 'forced' to establish connections with foreign muscles, and centrally regenerating nerves will reconnect to their original sites, in spite of radical alterations of their starting positions. A number of other rerouting studies were carried out, each of them supporting the basic premise, which Sperry termed 'the chemoaffinity hypothesis.' Briefly stated, the growth and development of a pre-patterned neural network is determined by a genetically controlled chemical code, and the 'hard-wired' nature of this network results in a specific, highly adaptive range of behavior which is unalterable by training. Other studies over the next 20 years, using neuroanatomical and neurophysiological approaches, provided confirmation of this hypothesis. Attardi and Sperry (1963) provided further, histological confirmation by showing that regenerating retinal axons in the goldfish would always regrow to their original terminal sites in the brain.

Two major theoretical papers were published by Sperry in the mid-1960s (see Voneida 1997, Sperry 1965a), in which he precisely described his chemoaffinity theory. In spite of numerous attempts over the next three decades to disprove the theory, it has prevailed with only slight variations and elaboration

of the basic premise. It was primarily this work, begun as a predoctoral student in 1938 and pursued through the early 1960s that resulted in his election to the National Academy of Sciences.

A second major area of Sperry's work deals with cortical organization related to higher cognitive function in mammals. He joined the renowned psychobiologist Karl Lashley at Harvard for his postdoctoral work, where he began a new series of studies on the role of electrical fields in neocortical function. He was not comfortable with the idea that electrical fields or waves acting in a volume conductor (the brain) are critical for neocortical processing. His first study to challenge this was published in the late 1940s (see Sperry 1947). Here he demonstrated that motor coordination in monkeys remained virtually unaffected after multiple transections of sensorimotor cortex. He later confirmed this by demonstrating that neither cortical cuts, the insertion of numerous short circuiting wires or insulating plates into the cortex had any adverse effect on cortical function. These studies demonstrated that perception is dependent upon vertically oriented afferent and efferent cortical axons, predating the discovery of vertically oriented cortical columns. Sperry's premise was based on his keen understanding of neuroanatomy and neurophysiology, including the work of Santiago Ramon y Cajal and Lorente' de No, both of whom had demonstrated the importance of radial cortical connections to cortical function.

Thus, in a few carefully conducted experiments, Sperry had once again upset two major theories of brain function, Gestalt electric field theory of perception and the reduplicated interference pattern hypothesis. Indeed, when the renowned neuroembryologist Viktor Hamburger presented Sperry with the Ralph Gerard Award from the Society of Neuroscience in 1979, he proclaimed, 'I know of nobody else who has disposed of cherished ideas of both his doctoral and his postdoctoral sponsor, both at that time the acknowledged leaders in their fields.'

Sperry's third major contribution relates to the function of neuronal connections between the two brain halves (called commissures), interhemispheric communication, and lateralization of cognitive abilities in humans. Most of this work was done at Caltech, where he served as Hixon Professor of Psychobiology from 1954 to 1994. It was these studies, carried out between the mid-1950s through the 1970s (see Sperry 1981), which resulted in the Nobel Prize in Physiology and Medicine (shared with Hubel and Wiesel) in 1981. The early studies, carried out by Sperry and his students at Caltech from the mid-1950s through to the 1960s, elegantly elucidated some of the major functions of interhemispheric connections in memory transfer and eye-hand coordination. Restriction of sensory input to one brain hemisphere in animals having undergone surgical separation of the two brain halves was shown to limit the learning of various tasks to the 'input' hemisphere; the opposite side was capable of learning, but remained naive to those tasks until trained. Learning curves for each brain half were virtually identical; it was as if two separate brains were housed within a single cranial vault. It was during this period that he also began his work on the lateralization of cerebral function in humans. Neurosurgeons Philip Vogel and Joseph Bogen had performed a surgical section of the major pathway connecting the two brain halves on a World War II veteran who was suffering from progressively worsening epileptic seizures. Working in collaboration with Vogel and Bogen, Sperry and his students began a series of tests directed at understanding the effects of this surgical hemispheric separation on human perception, speech, and motor control. The work involved comparisons of cognitive abilities between the two separated brain halves, demonstrating differences theretofore unrecognized.

For the next 20 years, the work of Sperry and his collaborators revolutionized our understanding of brain function. Prior to this time, the right hemisphere had been virtually neglected by neuroscientists. At best, it was felt to play only a minor role in human cognition, with all major cognitive functions emanating from the left, or dominant half. As a result of Sperry's research, the right hemisphere gained its long overdue recognition as a full partner to the left, playing a major role in holistic, parallel, and spatial abilities. The left brain half was found to be superior to the right in tasks involving analytical, sequential, and linguistic processing. Perhaps most importantly, however, Sperry demonstrated that the combined effects of bihemispheric activity amounted to much more than the simple additive effects of the 2 separate hemispheres, and led him into the area of his fourth and final major contribution; namely, the relationship between brain and mind, science and values.

Sperry began thinking and writing about the subjective aspects of brain function as early as the late 1950s. In 1959, for example, as part of the discussion at the Josiah Macy Conference on the Central Nervous System and Behavior, he stated that he had never been entirely satisfied with the behaviorists' interpretation of brain function in purely objective terms, with no reference whatsoever to subjective experience. He pointed out that this does not mean that we should abandon the objective approach; it simply means that we should also include subjective experience in our working models of the brain. Again, in his 1965 paper, *Merging Mind, Brain, and Human Values* (see Sperry 1965b), he reinforces this thesis in his declaration that conscious forces must be included in any model or description of brain function, stating that in this model, '… conscious mental psychic forces are recognized to be the crowning achievement of some five hundred million years or more of evolution.'

Sperry continued to think and write about the mind, consciousness, and their relationship to human values right up to the end of his life. He postulated that the

mind is an emergent property of brain function, and like most emergent properties, bears little resemblance to the hundreds of millions of neurons in the central nervous system. He emphasized that while the mind is not the *same* as the brain, it is *dependent* upon it. This point is an important one, for it separates Sperry's views from the philosophical concept of dualism. Thus brain and mind are by no means separate entities, but function rather as coexistant partners through the processes of emergence and downward causation, or feedback. Taken together, these two properties result in a continuous, highly dynamic process of change from one moment to the next, drawing constantly on previous experiences by way of learning and memory, whether they occurred in the distant past or only moments ago. He strongly argued against reductionism, which dominated psychological thinking during the first half of the twentieth century. He believed that reducing consciousness to its separate components obliterates the emergent properties of mind, with '*... all its great power and uniqueness*' (see Sperry 1993).

Sperry then elevated his concept of mind as an emergent property of brain activity to a global level, stating that:

> The new paradigm affirms that the world we live in is driven not solely by mindless physical forces but more crucially, by subjective human values. Human values become the underlying key to world change (see Voneida 1997).

He urged his scientific colleagues to use the very unique and powerful forces of mind and consciousness toward improving and preserving the quality of life on our planet, rather than the reverse. His urgent appeal finally began to be heard when his long-term friend and colleague, Professor Rita Levi-Montalcini, herself a Nobel Laureate, convened the International Conference of Human Duties at the University of Trieste in November, 1992, to discuss these ideas in greater detail. Regrettably, Sperry's illness prevented him from attending the first meeting, which included 10 Nobel Laureates and numerous others, representing such widespread disciplines as neurobiology, chemistry, physics, economics, and theology. This conference led to the formation of The International Council of Human Duties, which formulated 'The Declaration of Human Duties' in 1994. The document has been published (Trieste Declaration 1997) and forwarded to The United Nations, for its consideration as a corollary to the 'Declaration of Human Rights.'

Sperry's thinking about subjective experience, consciousness, the mind and human values were crystallized in his paper 'The Impact and Promise of the Cognitive Revolution' (Sperry 1993), delivered at the American Psychological Association's centennial meeting in 1993. It expressed his great hope and sincere belief that if we humans could simply be persuaded to put our collective minds together and to use the enormous emergent power that they are capable of generating, we might very well begin the new millennium by not only improving the quality of life on the planet, but also by significantly increasing our chances for survival.

See also: Brain Asymmetry; Cognitive Neuroscience; Mind–Body Dualism; Neuromuscular System

Bibliography

Attardi D G, Sperry R W 1963 Preferential selection of central pathways by regenerating optic fibers. *Experimental Neurology* **7**: 46–64

Sperry R W 1945 The problem of central nervous reorganization after nerve regeneration and muscle transposition. *Quarterly Review of Biology* **20**: 311–69

Sperry R W 1947 Cerebral regulation of motor coordination in monkeys following multiple transection of sensorimotor cortex. *Journal of Neurophysiology* **10**: 275–94

Sperry R W 1950 Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology* **43**: 482–9

Sperry R W 1952 Neurology and the mind-brain problem. *American Scientist* **40**: 291–312

Sperry R W 1965a Embryogenesis and behavioral nerve nets. In: Dehaan R L, Ursprung H (eds.) *Organogenesis*. Holt, Rinehart and Winston, New York, pp. 161–85

Sperry R W 1965b Mind, brain and humanist values. In: Platt J R (ed.) *New Views of the Nature of Man*. University of Chicago Press, Chicago, pp. 71–92

Sperry R W 1981 Some effects of disconnecting the cerebral hemispheres. (Nobel lecture.) Almquist and Wiksell, Stockholm

Sperry R W 1983 *Science and Moral Priority: Merging Mind, Brain, and Human Values*. Columbia University Press, New York

Sperry R W 1993 The impact and promise of the cognitive revolution. *American Psychologist* **48**(3): 878–85

Trieste Declaration of Human Duties 1997 *A Code of Ethics of Shared Responsibilities*. Trieste University Press, Trieste, Italy

Voneida T J 1997 *Roger Walcott Sperry, 1913–1994*. The National Academy Press, Washington, DC

T. J. Voneida

Spirit Possession, Anthropology of

Since ancient times possession is a transcendental human experience that exists all over the world. Bourguignon (1973) found that in 90 percent of her sample of 488 societies, one or several forms of possession or, as she called it, altered state of conscience, existed. Even in Western societies, despite secularization, belief in spirits regains in importance due to new religious movements and the sweeping success of fundamentalist Christian churches. This

worldwide attractiveness asks for a rethinking of possession in relation to modernity and the disenchantment of the world.

1. Understanding Possession

The wide varieties of beliefs and cultural traditions make it difficult to define what possession is, and what it means for different people. A minimal classification is advanced by Boddy who defines spirit possession as 'the hold exerted over a human being by external forces or entities more powerful than she' (Boddy 1994, p. 407). In most possession cults these external forces are considered as spirits, god(s), or ancestor heroes. In Africa, possessing spirits are generically thought of as breath, air, or wind. Their individuality is produced in an interactive process, in which the medium, the afflicted person, and the public take part. Spirits may incarnate ancestors, cultural heroes, or former kings as well as specific ethnic or social groups, animals, material things, abstract principles, and even emotional states. They are refractions of social life and express the whole range of a society's cultural archive. Despite this objectification they are thought of as distinct personalities with special tastes for food, clothes, jewelry, colors, rhythms, and music.

The spirits' interaction with humans varies considerably. It is either conceived of as a form of communication through dreams, images, or voices and/or as a violent attack in which illness or misfortune is inflicted on an individual. During possession, humans are no longer themselves, but 'horses,' 'vessels,' or 'containers' of their spirits. They are not responsible for their acts or speeches, since they are dominated by forces alien to them. Lienhardt (1961) has called this mental stage passiones, a form of being acted upon, instead of being an actor. In order to determine the identity of the afflicting spirit, ritual performances are staged which serve as diagnosis and therapy at the same time. The process of initiation is more or less formal and in some cases dependent on public accountability. Complete unification with the spirit is achieved by way of trance or ecstasy which may be generated through music, dance, medicine, and drugs, or as Eliade (1951) has pointed out by way of archaic techniques, such as concentration, meditation, or rapture. Trance and ecstasy are used synonymously by many scholars. Others think they represent different bodily expressions of an altered state of consciousness. Trance, a sleep-like state and form of mental dissociation, may change into ecstasy: a form of frenzy, expressed in high pitched cries, hiccups, convulsive fits, speaking in tongues, or unknown languages, known as glossolalia. The identification with the spirit engenders a change in personhood and reverses former relations of subordination into relations of reciprocity. Spirits are accommodated and integrated into the medium's life and that of his/her family, since most rituals include wider kin. De Heusch (1962) called this process adorcism and differentiated it from exorcism that implies the removal of evil spirits or ghosts, which is typical of demonic possession.

2. Possession, Gender, and Empowerment

Possession is not a singular event but denotes a lifelong relationship whose meaning is culturally determined, socially transmitted, and individually learned and implemented. The great majority of the possessed are women. Their dominance in possession cults has been the object of many enquiries and the answers given have differed fundamentally. Lewis (1989) discussed possession as a means for subordinate women and (marginalized) men, to gain material benefits and improve their social status. Others have pointed out the close link between possession, femininity, and female identity that are playfully re-enacted, modified, or contested by transgressing moral boundaries. The heterosexual nature of most spirits rivals long established marriages and deconstructs male sexual hegemonies. It is this subversive nature of possession that is one reason for its ambivalent and ambiguous reputation.

Possession offers a wide range of opportunities for women and men to explore new identities, negotiate power relations between the sexes, and open up new spaces for social action. These are not limited to symbolic forms of behavior but refer to very mundane forms of praxis. Women and men competitively try to benefit from the professionalization and commercialization of healing which is more or less pronounced in different cults. They are engaged in ecological protection as 'guardians of the land' and exercise social and political power as cult leaders. Since colonial times spirit mediums are feared by colonial administrators as subversive, because they play(ed) eminent roles in wars of liberation and in the subsequent purification of war traumata. Due to the close association between possession, individual stress, and social change, possession cults are not limited to rural areas, but occur also in towns. In Malaysia female factory workers are struck by spirit possession (see Ong 1987) as a form of resistance against capitalist exploitation, while Kapferer describes exorcism in Sri Lanka as 'metacommunicative of the dynamics of class' (1991, p. 51). The transcending of class boundaries is also reported for Brazil and the Caribbean where middle-class people form a substantial number of cult members (Csordas 1987). In African towns a variety of possession cults compete for adherents belonging to all walks of life, including women as well as men. Although many cults are remarkably stable over time and extend over vast regions, a frequent redrawing of boundaries between town and village is quite typical.

3. Discourses on Possession

Since the beginning of the twentieth century the fascination with the subject on the part of scholars has generated a large body of literature. Whereas European case studies played an important role in earlier publications, examples from other continents began to dominate in the twentieth century. Due to the complexity of the subject all main theoretical debates are represented in the interpretations of spirit possession. Three of the most important approaches are:

3.1 Psychological Approaches

An important issue in early studies was the truth or falseness in the experience of possession that de Heusch (1962) classified as authentic or inauthentic possession. As a consequence, psychological interpretations loomed large from the very beginning, associating possession with multiple personalities and psychiatric symptoms. It was described as a form of pathology, due to personal stress or social change. Experiences of possession were predominantly discussed in metaphors of illness and healing which were dissociated from religious beliefs. Significant for this form of medicalization (Csordas 1987) was the neglect of its cultural and cosmological context. This reductionism was first challenged by more sophisticated case studies (Crapanzano and Garrison 1977, Obeyesekere 1981) which related psychological or psychoanalytical insights to the cultural construction of healing. A synthesis was achieved by Janzen (1992) who convincingly demonstrated that healing and religion are inseparable for the understanding of possession.

3.2 Sociological Approaches

Firth's (1964) distinction between spirit possession and spirit mediumship opened a sociological path for interpretation. He discussed spirit possession as a form of individual affliction whereas spirit mediumship was characterized 'as serving as an intermediary between spirit and men' (1964, p. 247). Beattie and Middleton (1969) elaborated this differentiation in their pioneering study on spirit mediums whose exercise of power they related to different forms of social organization. Arguing from a functionalist background they were concerned with the sociological and political functions of spirit mediumship and stressed its closeness to social change. Lewis (1989) took up their distinction and equated central possession cults with societies' main moral order being represented by men of social standing while peripheral possession cults were the domain of amoral spirits and powerless women. Being a weapon of the subordinate he described these cults as a 'thinly disguised protest movement directed against the dominant sex' (1989, p. 26). Another important discussion in his comparative sociology of ecstasy was his fervent pladoyer for the identification of spirit possession with shamanism. His arguments were directed against Eliade and de Heusch who had conceived the two as inverse processes. Whereas shamans rival with the gods and ascend to them, spirit mediums incarnate the spirits who descend and attack humans. Despite the impact of Lewis' book the instrumentalist portrayal of possession by most functionalists came increasingly under attack in the 1980s and led to another shift in methodology.

3.3 Hermeneutic Approaches

Due to the influence of Schütz and Geertz on anthropological theory during the 1980s, many authors favored a phenomenological or hermeneutic approach for a deeper understanding of the complexity of meanings acted out in these cults. Lambek who had earlier discussed curing ceremonies as performed 'texts' introduced a new topic by his ascription of possession as a means of knowledge (1993, p. 305f). He not only differentiated between objectified knowledge in Islam and embodied knowledge in possession, but also stressed the interdependence of objectification and embodiment for each type of knowledge. These dialectics explain, among other factors, as Boddy has pointed out, the counter-hegemonic praxis of possession without inhibiting the many forms of syncretism between the two. In her view possession beliefs constitute a metalanguage, 'a metaphoric construction on quotidian reality' (1989, p. 270). She describes how the world of humans is paralleled by the world of the spirits whose behavior, however, ridicules, satirizes or reinterprets human values. In an insightful analysis Leiris (1958) had already drawn attention to the playful and theatrical character of possession dances as arenas for negotiating identities. Fritz Kramer (1987) explored this venue further by interpreting 'performances of alien spirits as mimetic ethnographies' (Behrend and Luig 1999, p. 17). The distancing gaze on the other not only helps to recognize one's own self but also is a means of understanding past and present events. At the same time, this symbolic construction of the world is part of an ongoing appropriation of modernity that may be modified or even rejected through the intervention of the spirits. It is this creative aspect of the rituals of possession that allow for the regional or even international mediation of local practices that are to produce spirits of cosmopolitan nature. Spirit possession is thus in no way an archaic practice that runs counter to a disenchanted world, but rather acts as a metacommentary on its modernity.

See also: Mysticism; Myth in Religion; Ritual; Ritual and Symbolism, Archaeology of; Shamanism

Bibliography

Beattie J, Middleton J (eds.) 1969 *Spirit Mediumship and Society in Africa*. Routledge and Kegan Paul, London
Behrend H, Luig U (eds.) 1999 *Spirit Possession, Modernity and Power in Africa*. University of Wisconsin Press, Madison, WI
Boddy J P 1989 *Wombs and Alien Spirits. Women, Men and the Zar-Cult in Northern Sudan*. University of Wisconsin Press, Madison, WI
Boddy J P 1994 Spirit possession revisited: Beyond instrumentality. *Annual Review of Anthropology* **23**: 407–34
Bourguignon E 1973 *Religion, Altered States of Consciousness and Social Change*. Ohio State University, Columbus, OH
Crapanzano V, Garrison V (eds.) 1977 *Case Studies in Spirit Possession*. Wiley, New York
Csordas T J 1987 Health and the holy in African and Afro-American spirit possession. *Social Science and Medicine* **24**(1): 1–11
de Heusch L 1962, Cultes de possession et religions initiatiques de salut en Afrique. *Annales du Centre d' Etudes des Religions*
Eliade M 1951 *Le chamanisme et les techniques archaîques de l' extase*. Editions Payot, Paris
Firth R W 1964 *Essays on Social Organization and Values*. University of London, London
Janzen J M 1992 *Ngoma: Discourses of Healing in Central and Southern Africa*. University of California Press, Berkeley, CA
Kapferer B 1991 *A Celebration of Demons*. Berg Publishers, Oxford, UK
Kramer F 1987 *Der rote Fes. Über Besessenheit und Kunst in Afrika*. Athenäum, Frankfurt, Germany
Lambek M 1993 *Knowledge and Practice in Mayotte, Local Discourse of Islam, Sorcery, and Spirit Possession*. University of Toronto Press, Toronto, ON
Leiris M 1958 La possession et ses aspects théatraux chez les Ethiopiens de Gondar. In: *L'Homme: Cahiers d'Ethnologie, de Géographie et de Linguistique*. Plon, Paris
Lewis I M 1989 *Ecstatic Religion. An Anthropological Study of Shamanism and Spirit Possession*, 2nd edn. Routledge, London
Lienhardt R-G 1961 *Divinity and Experience. The Religion of the Dinka*. Clarendon Press, Oxford, UK
Obeyesekere G 1981 *Medusa's Hair. An Essay on Personal Symbols and Religious Experience*. Chicago University Press, Chicago
Ong A 1987 *Spirits of Resistance and Capitalist Discipline: Factory Women in Malaysia*. State University of New York Press, Albany, NY

U. Luig

Split Brain

Before the 1940s, the functions of the corpus callosum, estimated to have about 1,000 million axons, were unknown (Akelaitis 1944). Early in the 1950s, Roger Sperry at the University of Chicago, who was researching the anatomy of visual form perception in the cerebral cortex, proposed a surgical investigation that was to yield new insights into this tract and the brain systems of consciousness (Evarts 1990, pp. xix–xxi). Ronald Myers, a research student with Sperry, demonstrated by cutting the corpus callosum that a cat's visual experience was divided in two, and could be transferred by it between the hemispheres (Myers and Sperry 1953).

In the 1950s and 1960s Sperry's research group, now at the California Institute of Technology (Caltech), exploited what was now called the 'split brain' to study visual and tactile perceptions, memories and motor control in cats or monkeys (Sperry 1961). Studies with monkeys showed that, while ordinary pattern discrimination was completely split, less figurative visual experiences were not completely duplicated after split-brain surgery. This proved the importance of 'motor sets' or 'intentions to move' that 'anticipate' sensory information and that invoke undivided systems of the brain stem and cerebellum.

In 1962, the effects of commissurotomy, undertaken to arrest life-threatening epilepsy, were investigated in a human patient (Bogen 1985). Sperry directed psychological tests of consciousness, learning and various skills after the operation (Sperry 1967). Commissurotomy to treat multifocal epilepsy in other patients made it possible to test rational consciousness in the left and right hemispheres, their similarities and differences, their convergent influence over acts, ideas and feelings of the whole person, and the role of language (Sperry 1967, 1974, Sperry et al. 1969).

Commissurotomy research confirmed Sperry's theory of the causal potency of motor intentions, and of self-consistent conscious ideas and beliefs, and moral impulses (Sperry 1977, 1983). It stimulated the neuropsychological analysis of hemispheric differences unique to humans, and, by clarifying the special modes of processing in cortical territories, supported the 'cognitive revolution' of the 1960s (Sperry 1993). Sperry's research on the cognitive processes of the cerebral hemispheres of human beings gained him a Nobel prize in 1981 (Sperry 1982).

Since the classical period of split-brain research in the 1960s and 1970s, interest in hemispheric differences has flourished, and the cognitive processes, language processes and memory of commissurotomy patients have been studied in new, detailed ways. Nevertheless, the overall picture of cerebral functions has not significantly changed (Benson and Zaidel 1985, Nebes 1990, Trevarthen 1990a, Corballis 1991, Gazzaniga 1995). Though commissurotomy was effective in limiting seizures, the operation carries risks, and is now rarely performed.

1. *Split-brain Experiments with Animals*

Myers applied the stereo-microscope surgery developed by Sperry to cut the crossover of visual nerves in the optic chiasm under a cat's brain, so each eye would

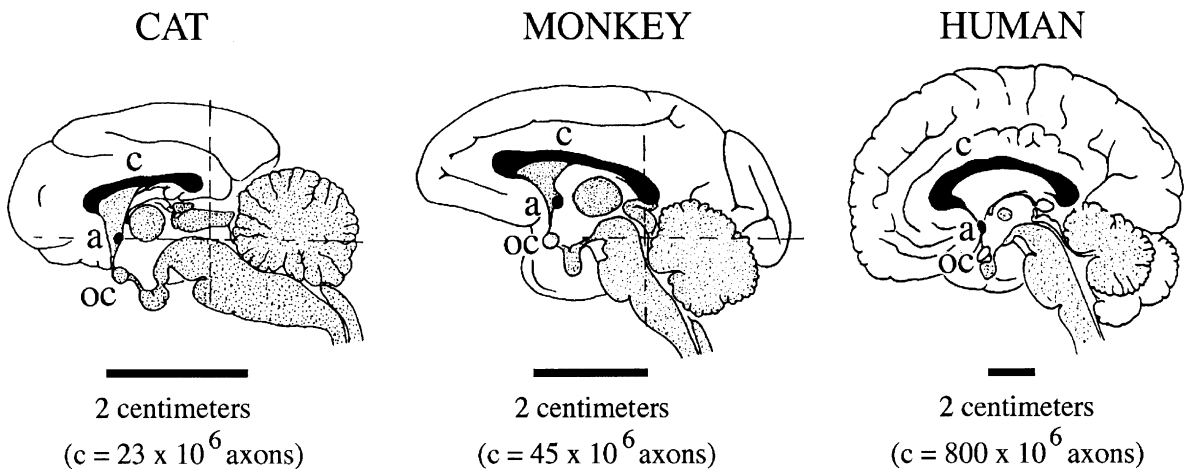

Figure 1
Anatomy of the cerebral commissures. Saggital sections of the brains of cat, monkey (rhesus macaque) and human adult, showing the corpus callosum (c), anterior commissure (a) and optic chiasm (oc). The estimated numbers of nerve axons in the corpus callosum of each species is indicated. This illustrates the enormous expansion of cortical functions in human beings, and their bilateral integration through the corpus callosum

lead to only one cerebral hemisphere (Figs. 1 and 2). In 1953, he and Sperry reported the transfer of visual memory between the hemispheres in cats with the chiasm divided, and the abolition of this transfer when the corpus callosum was cut (Myers and Sperry 1953). Outside carefully controlled test situations, the animals moved normally.

Monkeys also experience independent visual perception and memory in the two brain halves after the optic chiasm and forebrain commissures (corpus callosum and anterior commissure) are cut (Fig. 1). Again, the operated animals retained normal coordination of movements and attention when free to orient, correcting for loss of awareness due to the surgery. Split-brain monkeys learn two, mutually contradictory, visual discrimination habits in the right and left half-brains, showing that they possess double and separate realms of consciousness (Trevarthen 1962, Sperry 1977). In the cat, each disconnected hemisphere directed movements of the whole body, but with monkeys, a hand on the same side as one seeing hemisphere was, at least in the immediate postsurgical period, less willing to respond and, if forced into action, was clumsy, as if blind. Crossed pathways linking each cortex to the opposite hand were superior for guiding fine, purposeful movements of the fingers. Related split-brain projects demonstrated shifts of attention between the two separated halves of the cortex and the effects of preparatory sets to move in particular ways. The results revealed the global design of the mammalian brain for awareness, learning, and voluntary action (Sperry 1961, Berlucchi 1990).

What a chiasm-callosum sectioned (split-brain) monkey sees and remembers, i.e., in which hemisphere visual perception and learning occurs when both hemispheres are presented stimuli, depends on which hand the animal chooses for response. This is evidence that intention can instruct experience, that activities of the neocortex are directed intrinsically by purposeful impulses generated in the brain as a whole, and not just by processing the input of stimulus information at the cortical level.

When a monkey learns a complex two-handed skill, the memory of how to move is located more in the hemisphere opposite the dominant hand, i.e., the hand that the individual spontaneously prefers to use for fine, consciously regulated manipulations. A split-brain monkey's visual awareness can be turned on or off in one side of the brain by changing which hand the animal uses to manipulate the test apparatus to gain a reward. When using the right hand the monkey becomes conscious on the left side of its cerebrum and is unable to see with the right cortex. This causes the monkey to be inattentive to things in the left half of its experience, and makes it likely to imagine things that did not exist there. This clarifies 'perceptual neglect,' and 'illusory completion,' phenomena known from clinical neuropsychological research with consciousness of patients who have lost function in parts of their cerebral cortex (Trevarthen 1990b).

Tests of vision in the divided brain led to the theory that the sensory systems are linked to a hierarchy of motor coordinative systems, which either gear body movements to the layout of the surrounding world, or direct the subject to select objects that have life-

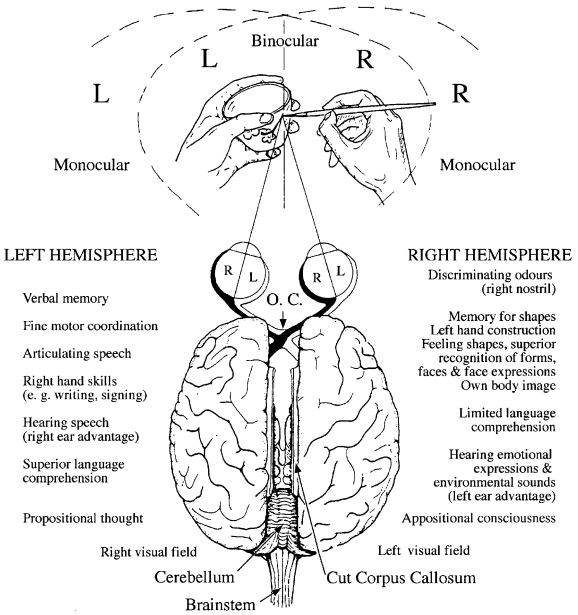

Figure 2
Human brain showing asymmetric functions in the cerebral hemispheres. The hemispheres are separated after commissurotomy (the 'split-brain' operation). Above, the two hands in the visual field, with the subject's attention fixated on the tip of the paintbrush. (L = left half of visual field; R = right half of visual field; o.c. = optic chiasm)

relevant qualities. Two active visual systems were distinguished: wide-field and part-conscious *ambient vision* informs whole-body locomotion and orientation; *focal vision*, employing the foveae to see within only a few degrees of visual angle and with precise intermittent fixation of gaze, inspects detail and identifies objects (Trevarthen 1968). The former, involving many subcortical centers, as well as parts of the parietal cerebral cortex, acts as motivator and map-reader for attention in the latter, more conscious vision of shapes, colors and patterns. The split-brain operation divides focal vision because its essential processes are confined to the cortex, but leaves ambient vision largely intact.

2. *Human Commissurotomy Patients*

In 1960, Joseph Bogen, a Los Angeles neurosurgeon, saw a patient who was having progressive, life-threatening seizures as a result of head injuries sustained as a paratrooper. In the light of the research with monkeys, which showed that commissurotomy could leave essentially normal control of whole-body action and the distribution of awareness, it was decided that cutting the corpus callosum and other interhemispheric commissures, 'total commissurotomy,' would reduce the epilepsy. This operation was carried out in 1962, and the epilepsy improved significantly. Sperry and Bogen, with Michael Gazzaniga, began systematic psychological tests at Caltech on the effects on perception, speech, and motor control. Instead of relying on standard neurological procedures, original tests were devised based on principles learned from split-brain animals (Sperry 1967, 1974, Sperry et al. 1969). The first report, with clinical observations made by Bogen, described the resultant left–right division of human consciousness (Gazzaniga et al. 1962). Norman Geschwind and Edith Kaplan, in Boston, described comparable 'deconnection' phenomena in patients with lesions that interrupted cortico-cortical links (Geschwind and Kaplan 1962).

Tests confirmed that awareness of objects seen by either half of the brain (in left and right halves of the visual field), or felt in the left or right hand, was essentially normal, but, as in the earlier animal studies, when free orientation was prevented, the left and right mental realms were as separate as if in separate heads (Fig. 2). Only when the subject shifted visual fixation or passed an object from one hand to the other did the one hemisphere know what the other was perceiving. As with the monkeys, human commissurotomy subjects experienced unified ambient visual functions, but split focal vision (Trevarthen and Sperry 1973, Trevarthen 1990b).

Though both hemispheres were conscious, only the left could speak or write. Surprisingly, however, given that left hemisphere lesions can produce total aphasia, the right hemisphere, though speechless, could nevertheless comprehend speech (Gazzaniga and Sperry 1967, Zaidel 1990b). It picked up information quickly from utterances made by the left half-brain, could read common words flashed to the left visual field, and could follow verbal instructions. It was, however, conspicuously weak at calculating and in comprehending logical propositions. Jerre Levy showed that in the nonspeaking right hemisphere, classically referred to as the 'minor' hemisphere, mental faculties that imagine spatial relations and transformations in space were more highly developed than on the left 'dominant' side of the brain. Two different, complementary kinds of cognitive strategy, two kinds of consciousness, were demonstrated (Levy-Agresti and Sperry 1968) (Fig. 2).

The surgically separated left hemisphere commands fluent speech and can generate and retain phonological images of speech sounds. It performs serial logical propositions such as are expressed in spoken, gestural or written language, and it is adept at the kind of combinatorial and systematic logic required for formal mathematics or score-based musical thought. The right hemisphere, in contrast, has superior mastery of the immediate 'appositional' experience (Bogen 1969)—of forms by sight, sound, or touch—and,

according to recent research, of odors. It can remember these experiences better, can distinguish and remember faces better, has richer reactions to colors, and to affective tones of the voice and its prosodic quality. It perceives more easily resemblances in topological but not Euclidean geometry. Commissurotomy patients have a color anomia, an inability to name colors, because the disconnected right hemisphere has a superior but inarticulate ability to perceive color. There is evidence that an undivided brain-stem color vision defines a dichromatic distinction between blue space and the reddish-brown solids, part of the natural layering of the conscious mind to serve different forms of action in the world (Trevarthen and Sperry 1973). Tests with commissurotomy patients support findings with brain-damaged and normal subjects showing the right half of the brain to be involved in emotional communication and the mediation of direct social interactions. The main hemispheric differences are summarized in Sperry's Nobel Address (Sperry 1982).

Jerre Levy and Colwyn Trevarthen used 'stimulus chimeras,' pictures showing different half objects either side of the midline, to observe biases in cognitive activities of the hemispheres induced by what a commissurotomy patient intended to do, or by what he or she had been told to think about (Levy and Trevarthen 1976). They called the motive sets that were capable of turning on or off one or other hemispheric cognitive mechanism, sometimes independently of the different cognitive aptitudes on the two sides, 'metacontrol.' Similar habitual arousal biases in normal subjects cause individual differences in cognitive aptitudes.

3. *Theoretical and Philosophical Implications*

Commissurotomy research stimulated both the neuropsychological investigation of effects of unilateral brain injury, and tests of normal perceptual, cognitive, and motoric asymmetries. The findings have been invoked to explain why children with similar opportunities for learning may be different in educational performance, and why adults show different intellectual aptitudes and job preferences. It is likely that hemispheric biases in cognition and intellectual differences between persons are products of inherent styles in purposes and emotions, as well as of different experiences and opportunities for practice and for communication with others.

In the normal brain the hemispheres work in tight partnership. The intuitive response to experience and the evocation of imagery linked to phenomenal reality by metaphor, on the one hand, and verbal analysis and prescription of aesthetic judgments, on the other, are two natural brain systems that develop as complements in the making and communicating of language, technology and art. There are large differences between individuals. Findings with commissurotomy patients have been used to support theories of 'modular intelligence' derived from neuropsychological and cognitive science fields and attempts have been made to apply these theories to evolutionary changes in hominid behavior and the different kinds of cognitive processing inferred. However, such models of the mind are incomplete if they do not incorporate explanations of the motivations that generate both action and awareness, and the communicable emotions that evaluate them.

Cognitive operations resist efforts to locate them in parts of the brain. The skills, concepts and memories of language are stored over wide areas. Parallel-distributed processes contribute to the comprehension of language, as for facial recognition, spatial awareness and categorical understanding of objects, or for executive planning and motor sequencing. The development of cortical units involved in all these functions depends on associative cortico-cortical links, both within each hemisphere and through the corpus callosum. These supplement equally massive connections to and from cell populations beneath the cortex.

A majority of humans are 'right-handed,' and have been since Neolithic times, preferring to take conscious control in the right hand, especially for communicating by gesture and for carefully directing implements. About 15 percent prefer the left hand. Most of us tend to move the right half of our mouths more, or earlier, in speaking, and look right as we speak. These elaborated learned skills originate in asymmetries implanted in subcortical systems before birth.

It has been clear for more than 100 years that language functions are more strongly represented in the left hemisphere. Right hemisphere lesions selectively impair object and event recognition by sight or hearing and may have devastating effect on the capacity to recognize who another person is and what their emotion may be. Right posterior lesions can also weaken a person's discrimination of their own expression of feeling, and lower self-awareness generally. The cognitive and performance functions of the right half of the brain are important to a person's cultural life in a community, and to the acquisition of many skills that language may find hard to describe, and they serve to support the language-mediating functions in the left hemisphere.

There has been much speculation about how recurring asymmetries in works of art might be related to the anatomy of the brain, and our common brain asymmetry does appear to have been unknowingly exploited by many painters to generate aesthetic effects. There is a school of art teachers who draw on split-brain research to guide pupils to make stronger use of the right hemisphere, claiming that this generates more dynamic, coherent and aesthetically pleasing forms, and some require right-handers to sketch unthinkingly with their left hands. We have to learn how to talk about what is beautiful in art, and it is

inherently difficult to recall what we say. Language offers a powerful tool for introspecting, for planning actions and for manipulating information in 'phenomenal absence' from its source, but it is limited in representational imagination. It is particularly difficult to translate back from word to picture.

The work with commissurotomy patients convinced Sperry that brain-generated awareness and evaluation have a determining role over the experience obtained by learning. He concluded that science has to give consciousness an interactive, causal role in brain function. Subjective mental states explain conscious behavior and its changes, and determine what a person is and does. This view gives the higher 'macro,' mental, vital and social forces equal importance, in scientific explanations of life, with physics and chemistry. Sperry's thinking about human consciousness and the importance of values and emotion is summarized in *Nobel Prize Conversations* (Sperry et al. 1985).

See also: Brain Asymmetry; Brain Damage: Neuropsychological Rehabilitation; Conscious and Unconscious Processes in Cognition; Consciousness and Sensation: Philosophical Aspects; Consciousness, Cognitive Psychology of; Consciousness, Neural Basis of; Epilepsy

Bibliography

Akelaitis A J 1944 A study of gnosis, praxis, and language following section of the corpus callosum and anterior commissure. *Journal of Neurosurgery* **1**: 94–102

Benson D F, Zaidel E 1985 *The Dual Brain*. Guilford Press, New York

Berlucchi G 1990 Commissurotomy studies in animals. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology,* Vol. 4. Elsevier, Amsterdam, New York, Oxford

Bogen J E 1969 The other side of the brain, II. An appositional mind. *Bull. Los Angeles Neurol. Soc.* **34**: 135–62

Bogen J E 1985 The callosal syndromes. In: Heilman K M, Valenstein R (eds.) *Clinical Neuropsychology*, 2nd edn. Oxford University Press, New York

Corballis M C 1991 *The Lopsided Ape: Evolution of the Generative Mind*. Oxford University Press, New York

Edwards B 1979 *Drawing on the Right Side of the Brain*. J-P. Tarcher, Los Angeles

Evarts E V 1990 Foreword: Coordination of movement as a key to higher brain function: Roger W. Sperry's contributions from 1939 to 1952. In: Trevarthen C (ed.) *Brain Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry*. Cambridge University Press, New York

Gardner H 1982 *Art, Mind and Brain*. Basic Books, New York

Gazzaniga M S 1970 *The Bisected Brain*. Appleton-Century-Crofts, New York

Gazzaniga M S 1989 Organization of the human brain. *Science* **245**: 947–52

Gazzaniga M S 1995 Principles of human brain organization derived from split-brain studies. *Neuron* **124**: 217–28

Gazzaniga M S, Bogen J E, Sperry R W 1962 Some functional effects of sectioning the cerebral commissures in man. *Proceedings of the National Academy of Sciences USA* **48**: 1765–9

Gazzaniga M S, Sperry R W 1967 Language after section of the cerebral commissures. *Brain* **90**: 131–48

Geschwind N, Kaplan E 1962 A human cerebral deconnection syndrome. *Neurology* **12**: 675–85

Innocenti G M 1986. General organization of the callosal connections in the cerebral cortex. In: Jones E G, Peters A (eds.) *Cerebral Cortex,* Vol. 5. Plenum, New York

Levy J 1988 Cerebral asymmetry and aesthetic eperience. In: Rentschler I, Herzberger B, Epstein D (eds.) *Beauty and the Brain. Biological Aspects of Aesthetics*. Birkhauser Verlag, Basel, Boston, Berlin

Levy J, Trevarthen C 1976 Metacontrol of hemispheric function in human split-brain patients. *Journal of Experimental Psychology: Human Perception and Performance*. **2**: 299–312.

Levi-Agresti J, Sperry R W 1968 Differential perceptual capacities in major and minor hemispheres. *Proceedings of the National Academy of Sciences USA* **61**: 1151

Myers R E, Sperry R W 1953 Interocular transfer of a visual form discrimination habit in cats after section of the optic chiasm and corpus callosum. *The Anatomical Record* **115**: 351–2

Nebes R D (ed.) 1990 The commissurotomised brain. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*, Vol. 4 Sect. 7. Elsevier, Amsterdam, New York, Oxford

Sperry R W 1961 Cerebral organization and behavior. *Science* **133**: 1749–57

Sperry R W 1967 Mental unity following surgical disconnection of the cerebral hemispheres. *The Harvey Lectures* **62**: 293–323

Sperry R W 1974 Lateral specialization in the surgically separated hemispheres. In: Schmitt F O, Worden F G (eds.) *The Neurosciences. Third Study Program*. MIT Press, Cambridge, MA

Sperry R W 1977 Forebrain commissurotomy and conscious awareness. *The Journal of Medicine and Philosophy* **2**(2): 101–26

Sperry R W 1982 Some effects of disconnecting the cerebral hemispheres. (Nobel lecture). *Science*, **217**: 1223–6

Sperry R W 1983 *Science and Moral Priority*. Columbia University Press, New York

Sperry R W 1993 The impact and promise of the cognitive revolution. *American Psychologist* **48**(3): 878–85

Sperry R W, Eccles J, Prigogine I, Josephson B, Cousins N 1985 *Nobel Prize Conversations*. Saybrook, Dallas, TX

Sperry R W, Gazzaniga M S, Bogen J E 1969 Interhemispheric relationships: The neocortical commissures; syndromes of hemisphere disconnection. In: Vinken P J, Bruyn G W (eds.) *Handbook of Clinical Neurology,* Vol. 4. North Holland, Amsterdam

Stephan M 1990 *A Transformational Theory of Aesthetics*. Routledge, London

Trevarthen C 1962 Double visual learning in split-brain monkeys. *Science* **136**: 258–9

Trevarthen C 1968 Two mechanisms of vision in primates. *Psychologische Forschung* **31**: 299–337

Trevarthen C 1987 Split brain and the mind. In: Gregory R L, Zangwill O L (eds.) *Oxford Companion to the Mind*. Oxford University Press, Oxford, New York

Trevarthen C 1990a Integrative functions of the cerebral commissures. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology,* Vol. 4. Elsevier (Biomedical Division), Amsterdam

Trevarthen C (ed.) 1990b *Brain Circuits and Functions of the Mind: Essays in Honour of Roger W. Sperry*. Cambridge University Press, New York

Trevarthen C 1999 Roger W. Sperry. In: Wilson R, Keil F (eds.) *The MIT Encyclopedia of Cognitive Sciences*. MIT Press, Cambridge, MA

Trevarthen C, Sperry R W 1973 Perceptual unity of the ambient visual field in human commissurotomy patients. *Brain* **96**: 547–70

Zaidel E 1990a Language functions in the two hemispheres following complete commissurotomy and hemispherectomy. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology,* Vol. 4. Elsevier, Amsterdam, New York, Oxford

Zaidel D 1990b Memory and spatial cognition following commissurotomy. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology,* Vol. 4. Elsevier, Amsterdam, New York, Oxford

C. Trevarthen

Sport and Exercise Psychology: Overview

Though the beginnings of sport psychological research can be traced back to the end of the nineteenth century it was not before the sixties that sport psychology emerged as an important and recognized academic and applied discipline (cf. Singer et al. 1993, ch. 1). The foundation of the International Society of Sport Psychology (ISSP) in 1965 and of several national organizations in Europe and in North America mark this period as the start of systematic scientific research and practice in sport psychology. Due to its increasing importance after World War II a systematic scientific research and support of elite sport gained significance. It became obvious that psychological processes were important aspects of athletic peak performance. Therefore, sport psychology was in this period mainly concerned with research on elite athletes and on performance enhancement in competitive sport. A distinction between the academic or research-oriented discipline of sport psychology and the discipline of applied sport psychology emerged in the 1980s, when not only the Association for the Advancement of Applied Sport Psychology (AAASP) was formed in 1985, but when also more and more books and journals were published worldwide that were dedicated to helping elite athletes reach their full potential. At the beginning of the twenty-first century it seems clear that academic and applied sport psychology are interdependent and profit from each other. Scientifically based methods of intervention are used in applied settings and research is stimulated by suggestions from practitioners.

During the 1980s another important development emerged. More and more researchers became involved in exercise psychology. In contrast to sport psychology that is focused on performance enhancement and competitive sports, exercise psychology is concerned with psychological aspects of primarily noncompetitive leisure, fitness, and health-oriented physical activity. This development can be nicely seen in the renaming of the *Journal of Sport Psychology*, first published in 1979, into the *Journal of Sport and Exercise Psychology* in 1988. Sport and exercise psychology, as it is called nowadays, deals with the psychological factors that influence sport performance and exercise behavior, including motor learning and motor development, as well as with the impact of sport and exercise on psychological growth and development, like personality, health, and well-being. The present overview will focus on individual, social, and motivational influences on sport and exercise behavior as well as on short- and long-term effects of sport and exercise. The topics of performance enhancement and of developmental sport psychology are dealt with in several other articles (see *Sport Psychology: Performance Enhancement*; *Developmental Sport Psychology*).

1. *Individual, Social, and Motivational Influences on Sport and Exercise Behavior*

1.1 *Individual Determinants of Sport and Exercise*

Research on personality differences in sport was greatly influenced by the paradigms of differential psychology. The trait approach was most popular until the 1970s. Numerous studies looked for differences in personality variables between athletes and nonathletes as well as between athletes of different performance levels. Talent detection via personality tests and performance enhancement due to personality training were the main impetus of this line of research. Taken together the results are rather inconclusive and show only correlational but no causal relationships. There were several methodological problems leading to these disappointing findings, like lacking adaptation to the sports context of the scales used and suboptimal research designs with respect to causal conclusions. Even the well-known finding from the 1960s showing a healthier personality profile of successful elite athletes than of nonathletes did not prove valid in the long run (Rowley et al. 1995). The interactional approach that followed is more promising and is still popular. Scales were developed within the sports context that differ between state and trait aspects of respective variables like, for instance, competitive anxiety, sport confidence, or cognitive strategies. Personality research in sport at the beginning of the twenty-first century is more concerned with either performance enhancement in sport or with effects of exercise on personality variables like, for instance, self-esteem (cf. Sect. 2.2).

1.2 *Social Processes in Sport and Exercise*

The beginnings of research on productivity in sport teams and about spectators' influence on performance can be traced back to the end of the nineteenth

century. Sport and exercise are often performed in groups or teams, and therefore social processes are deemed highly significant (Carron and Hausenblas 1998). Team cohesion, leadership, and aspects of group composition on individual and group outcomes are the most important topics. Whereas in the beginning cohesion was regarded as interpersonal attraction between team members, i.e., social cohesion, at the beginning of the twenty-first century cohesion in sport teams is conceptualized and assessed as social and task cohesion. The positive relationship of group cohesion and performance has been well-documented for sport teams and is mainly based on task cohesion. This relationship is quite robust, and exists irrespective of the group task. For example, highly cohesive handball, rowing, or golf teams alike are more successful than low cohesive teams. Less is known about the causal paths of the relationship. There is convincing evidence that success fosters cohesion more than vice versa (Mullen and Copper 1994). Looking at the relationship of cohesion and adherence to sport and exercise reliable differences can be found between dropouts perceiving a lower cohesion and adherers perceiving a higher coherence in exercise groups and in sport teams. Perceived high cohesion may lead to more satisfaction of the group members and their overall attraction to the group, whereas perceived low cohesion reduces the motivation to stay in the group. This result is of high importance for any setting, be it exercise, recreational, or competitive sport.

The predominant approaches to leadership behavior of coaches in sport are the mediational model of Smoll and Smith (1989) and the multidimensional model of Chelladurai (see Chelladurai and Riemer 1998). The mediational model was developed in youth sport and postulates that the effects of the coaches' behaviors are mediated by the meaning that players attribute to them. Smoll and Smith (1989) could demonstrate that coaches who give general technical instruction, who prefer positive feedback, and who give mistake-contingent instruction and encouragement are evaluated most highly by their athletes. Training in these skills led to more positive evaluation of youth sport coaches, to fewer dropouts from their groups, and to an increase in the athletes' self-esteem and in team cohesion compared to nontrained youth sport coaches. However, the coaches' behaviors had no direct impact on the athletes' performance.

The multidimensional model of leadership (MDL) by Packianathan Chelladurai postulates that individual and group outcomes (satisfaction and performance) are dependent on the congruence among required, actual, and preferred leader behavior. Preferred and perceived leader behavior is measured with a five-dimensional questionnaire called the Leadership Scale for Sports (LSS). This instrument has been tested and validated in quite a few studies with athletes and coaches in various sport contexts and countries, in North America, Asia, and Europe. Research was primarily focused on the perception of coach behavior by adult and high-level athletes. The relationship to outcome measures is inconclusive. Future research has to focus on this relationship between coaches' behavior and outcome measures such as performance and satisfaction in the group.

1.3 Motivation

Considering motivation as a term for explaining intensity and direction of an individual's effort, motivation research in the field of sport and exercise is on the one hand concerned with explaining why and how certain people are attracted to which kind of sport or exercise and why others are not. On the other hand strategies to enhance motivation, like mastery training, attributional retraining, or goal setting have been developed and tested (Roberts 1992). With regard to the first issue, achievement and competition have been a traditional focus of motivation research in sport. Three theoretical perspectives were most influential: need achievement theory, attribution theory and social-cognitive approaches, mainly goal orientation theory. Starting with need achievement theory in the 1950s and 1960s, high and low achievers in sport were found to differ in their dispositional orientations (motive to achieve success or to avoid failure), their task choice, and their emotional reactions to success and failure, as was predicted by the theory.

The attributional model of achievement motivation of Bernard Weiner led to a considerable number of studies since the 1970s (Singer et al. 1993, ch. 19). They showed that the initial attributional framework of ability, task, effort, and luck attributions needs an extension to more sport specific attributions like opponents' quality or audience support. Self-serving biases in attributions were also found in the sports context, whereas gender differences were rarely found. Reattribution training for low achievers was successfully applied in sport settings like physical education, elite sport, and exercise. Another important topic was the relationship of attributions and emotions in sport. Though an attributional model would predict that emotions after success and failure depend on attributions, emotions of athletes seem to depend more on outcome than on attributions (Willimczik and Rethorst 1995).

Since the 1980s social cognitive approaches have gained attention in sport psychology. For instance, the goal perspective approach of John Nicholls was adapted to sport settings by Joan Duda and tested in numerous studies (cf. Duda 1992). Goal perspective theory assumes that in achievement contexts like sport and exercise individuals are motivated because they want to demonstrate ability or competence. This can be done with two types of goal, which are conceptualized as dispositions. The ego or outcome orientation is easily aroused in competitive situations like

sport. Success or failure is dependent on the performance of others and perceptions of ability are referenced to the ability of others. In contrast, the task goal or mastery goal focuses on individual comparison. Perceptions of ability are self-referenced. Task orientation is typically aroused in situations where the focus is on learning or mastery. Results show that task orientation is correlated with intrinsic motivation, with more enjoyment in sport, with persistence in the face of failure, and with the selection of moderately difficult tasks. Task-oriented individuals are more like high achievers. In contrast, ego-oriented individuals, especially when low in perceived competence, are more at risk to drop out of sports, they are more prone to develop fear of failure, and they emphasize fairness in sport less than task-oriented individuals. Fostering task orientation more than ego orientation is therefore highly recommended. Goal perspective theory has also focused on socialization contexts and their impact on the development of one or the other kind of goal. It is assumed that the socialization context induces a motivational climate where either competition or mastery is emphasized. A competitive climate may enhance ego orientation, whereas a mastery climate enhances task orientation.

Though achievement motivation is clearly important it is not the sole reason for sport and exercise participation. For many youngsters and adults affiliation motivation, pleasure seeking, physical challenge, adventure, and risk motivation, as well as the wish to shape an attractive body and to keep one's 'fitness' by means of exercise play an equally or even more important role. Sport and exercise become integrated into a person's lifestyle, and the wish to express a certain lifestyle is an important motivation to practice exercise and sport. A lifestyle theoretical perspective in sport and exercise psychology (Dishman 1994, Abele et al. 1997) posits that sport is an important aspect of a person's general way of living. It is a means both to pursue different motives (affiliation, achievement, risk, fun, challenge, health), and to express certain attitudes (toward youth, health, physical attractiveness) at the same time.

With increasing knowledge about the physical and mental health benefits of exercise (see Sect. 2) psychology becomes increasingly concerned with exercise adherence (Dishman 1994, Abele et al. 1997). Knowing that more than half of the adults in Western countries do not exercise enough from a health-oriented point of view and knowing that more than half of the participants who have started an exercise program for health reasons quit it again after a short time points to the importance of research on the determinants of exercise adherence and on interventions aimed at enhancing adherence and preventing dropout. Whereas early models have stated that perceived vulnerability or health protection motivation are most important, research has shown that perceived vulnerability does not lead to long-term adherence. Rather positive factors like exercise-specific self-efficacy beliefs, social support in the person's immediate environment, perceived group cohesion (see Sect. 1.2), and a suitable program and program climate are important. Participation motivation has to be understood not as a state but rather as a process undergoing at least three phases (inactivity, intention to participate and begin participation, adherence) that need specific motivational support.

2. *Short- and Long-term Effects of Sport and Exercise*

2.1 *Exercise and Physical Health*

Because physical health and psychological well-being are interrelated, it is appropriate to briefly consider the effects of sport and exercise on physical health (cf. Abele et al. 1997). These may be differentiated into cardiovascular (reduction of heart rate, decrease of diastolic blood pressure, improvement of blood circulation); hemodynamic (improvement of the blood flow characteristics); metabolic (stronger and more efficient musculature, improvement of the bone substance); and endocrinological effects (increase of catecholamines and cortisol, etc.). Immunological effects also exist; they are, however, only understood with respect to elite sports, but not yet with respect to exercise. Another distinction relates to exercise-induced changes in endurance, vigor, and coordination skills. All these changes also have an impact on psychological well-being.

However, besides studying the beneficial effects of sport and exercise on physical and psychological health, sport and exercise psychology is also concerned with high risk sports, the determinants and their possibly dangerous effects.

2.2 *Exercise and Psychological Well-being*

Studies on exercise and psychological well-being have been conducted both with nonclinical and with clinical groups. Questions addressed were the role of exercise in the reduction of anxiety and depression, exercise and mood changes, positive and negative addiction to exercise, and the relation among exercise, personality changes, and cognitive functioning. Usually the impact of aerobic exercise like walking or running or fitness training, but sometimes also the impact of more competitive games like tennis, volleyball, or soccer, was analyzed. Several studies show that participants experience both short-term anxiety reduction and long-term depression reduction if they exercise regularly at least twice a week and for a time period of not less than six weeks. The longer they continue the more these negative psychological states are reduced. Run-

ning, for instance, can be an inexpensive, time-efficient adjunct to traditional psychotherapies (Weinberg and Gould 1999).

Nonclinical participants usually show the 'iceberg profile,' which means that compared to their preexercise mood their postexercise mood is characterized by elevated vigor, activity, and elation and reduced levels of anxiety, depression, confusion, fatigue, and anger. It has been argued that fitness activities versus sport games may serve different functions in the regulation of mood. Fitness activities may primarily lead to equilibration effects. A negatively experienced tension state in advance of the exercise is transformed into a positively experienced relaxation state after it. More competitive games, on the other hand, may lead to disequilibration effects. A positively experienced 'prestart' tension state is transformed into a positively experienced postgame relaxation state. Both short-term equilibration and disequilibration effects serve the participants' long-term well-being (Abele and Brehm 1993).

The causes of these mood changes are multifaceted and relate both to biophysiological processes (hormonal changes, increased blood circulation, muscle strengthening, and relaxation; see above) and to psychological processes. Important psychological processes may be distraction and time-out. During sport and exercise the participant is distracted from stress and urgent problems, and the time-out may lead to free-floating attention and to remote associations which, in turn, may help to find new and creative problem-solving strategies. Exercise can also become an addiction. As a positive addiction it is a regular activity integrated as one important and valuable aspect into the person's life. As a negative addiction it gradually becomes overwhelming and the person cannot stop doing it at the cost of neglecting other important life tasks. Negative exercise addiction needs psychological help and intervention.

Personality changes and changes in cognitive functioning have also been studied. Physical self-concept and self-esteem are positively influenced by regular exercise (Alfermann and Stoll 2000, Sonstroem 1997). It has also been posited that cognitive functioning may be enhanced by physical exercise. Though the idea that motor development is an important determinant of intellectual development in childhood is appealing, the research findings in adulthood show a correlational but not necessarily a causal relationship between exercise and cognitive functioning (Etnier et al. 1997).

3. Conclusions

Since the early 1960s sport and exercise psychology has undergone a rapid development both with respect to its theories and methods and with respect to its aims and topics. Regarding theoretical models, instead of directly taking over models from general psychology sport and exercise psychology now adapts and modifies these models contingent to the sport context and also builds new and more specific models. Such a context-specificity of theory construction leads to more specific hypotheses and may also help to refine more general theoretical models in psychology. Regarding methodological advances, similar developments can be observed. The measures used become more refined and more specific to the respective context. The general methodological approaches both in experimental and in field research become more refined.

Most important, sport and exercise psychology has broadened its topics and perspective from a predominant concern with performance of elite athletes to a concern with sport and exercise in everyday life, too. The psychological analysis of sport and exercise in schools, in recreational settings, in health settings, and in further sports related institutions has become as important as the analysis of psychological processes in elite sports. Exercise adherence as a topic, for instance, becomes more relevant the more people acquire health problems because of lack of exercise. Research on sports and lifestyle, as another example, must receive increased attention, not only because 'trend sports' are an important economic factor in Western societies, but also because people on average have more leisure time, and sport is one of the most important leisure activities. Research on psychological processes in physical education as a school subject, as a final example, may be directed at the impact of sport and exercise in schools on an individual's lifetime exercise behavior, as well as on its impact on tension reduction and aggressive behaviors in schools.

See also: Developmental Sport Psychology; Health Behavior: Psychosocial Theories; Health Behaviors; Health Behaviors, Assessment of; Health Psychology; Sport, Anthropology of; Sport Psychology: Training and Supervision; Sport, Sociology of; Sports as Expertise, Psychology of; Youth Sports, Psychology of

Bibliography

Abele A, Brehm W 1993 Mood effects of exercise versus sport games: Findings and implications for well-being and health. *International Review of Health Psychology* **2**: 53–80

Abele A E, Brehm W, Pahmeier I 1997 Sportliche Aktivität als gesundheitsbezogenes Handeln. In: Schwarzer R (ed.) *Gesundheitpsychologie*. Hogrefe, Göttingen, Germany, pp. 117–50

Alfermann D, Stoll O 2000 Effects of physical exercise on self-concept and well-being. *International Journal of Sport Psychology* **31**: 47–65

Carron A V, Hausenblas H A 1998 *Group Dynamics in Sport*, 2nd edn. Fitness Information Technology, Morgantown, WV

Chelladurai P, Riemer H A 1998 Measurement of leadership in sport. In: Duda J L (ed.) *Advances in Sport and Exercise Psychology Measurement*. Fitness Information Technology, Morgantown, WV, pp. 227–53

Dishman R K (ed.) 1994 *Advances in Exercise Adherence*. Human Kinetics, Champaign, IL
Duda J L 1992 Motivation in sport settings. In: Roberts G C (ed.) *Motivation in Sport and Exercise*. Human Kinetics, Champaign, IL, pp. 57–91
Etnier J L, Salazar W, Landers D M, Petruzzello S J, Han M, Nowell P 1997 The influence of physical fitness and exercise upon cognitive functioning. *Journal of Sport and Exercise Psychology* **19**: 249–77
Mullen B, Copper C 1994 The relation between group cohesiveness and performance: An integration. *Psychological Bulletin* **115**: 210–27
Roberts G C (ed.) 1992 *Motivation in Sport and Exercise*. Human Kinetics, Champaign, IL
Rowley A J, Landers D M, Kyllo L B, Etnier J L 1995 Does the iceberg profile discriminate between successful and less successful athletes? A meta-analysis. *Journal of Sport and Exercise Psychology* **17**: 185–99
Singer R N, Murphey M, Tennant L K (eds.) 1993 *Handbook of Research on Sport Psychology*. Macmillan, New York
Smoll F L, Smith R E 1989 Leadership behaviors in sport. *Journal of Applied Psychology* **18**: 1522–51
Sonstroem R 1997 The physical self-system: a mediator of exercise and self-esteem. In: Fox K R (ed.) *The Physical Self. From Motivation to Well-Being*. Human Kinetics, Champaign, IL, pp. 3–26
Weinberg R S, Gould D 1999 *Foundations of Sport and Exercise Psychology*, 2nd edn. Human Kinetics, Champaign, IL
Willimczik K, Rethorst S 1995 Cognitions and emotions in sport achievement situations. In: Biddle S H (ed.) *European Perspectives on Exercise and Sport Psychology*. Human Kinetics, Champaign, IL, pp. 218–44

A. E. Abele and D. Alfermann

Sport, Anthropology of

Theory production in anthropology has been characterized by the absence of sport as a field of analysis for understanding rituals, complex processes of identity construction, *rites de passage*, and bodily performances. For instance, Mauss' (1935) early piece on body techniques does not mention sport arenas and performances as privileged loci for the development of an anthropology of the body. Several reasons can be advanced as an explanation for this inattention to sport. The main explanation lies perhaps in the accepted idea that sports and games, because they produce asymmetry between winners and losers, flourish only in competitive industrial societies. In rituals, asymmetry is postulated in advance—initiated and uninitiated or sacred and profane—and the 'game' consists in making all the participants pass to the winning side by means of ritual participation (Lévi-Strauss 1966, p. 32). Therefore, almost by definition, anthropology, as the science of the nonmodern and 'primitive,' has in the past excluded the study of sports and competitive games because they were correctly perceived as central features of modernity.

The attention to the importance of sport in social theory is to a large extent related to the influence of the late Norbert Elias. In his historical sociological approach, sport was defined as a key area in the civilizing process of European societies. This process was characterized by continuous state control of the legitimate use of force, the development of social organizations to reduce open conflict among groups, and the elaboration of codes for social behavior oriented towards the exercise of individual self-control. Sport became a typical activity of leisure that, historically, fulfilled some of the required functions for the consolidation of the civilizing process. Clearly the exercise of self-control is central to sport. For Elias, one of the essential problems of many sports is how to reconcile two contradictory functions: 'the pleasurable de-controlling of human feelings and the full evocation of enjoyable excitement on the one hand, and on the other the maintenance of a set of checks to keep the pleasantly de-controlled emotions under control' (Elias 1986, p. 49). Modern sport is regulated by rules that assure equality among the contenders, make possible the maximization of inner pleasures and the relaxation of individual tensions, and contain the exercise of physical violence. Through Elias, sport becomes a privileged field for the analysis of individual and social tensions in modern societies.

In the critical tradition of social theory, best represented by Bourdieu (1984), the practice of sports has been seen as an efficient state, and bourgeois, mechanism to indoctrinate youth with values of sexism, nationalism, fanaticism, irrational violence, the cult of performance and competition, the cult of idols, and the uncritical acceptance of the values of capitalism. Bourdieu maintains that one of the perverse effects of sport is that it transforms the spectators—the consumers of the performances of the athletes—into a caricature of militancy. This is because their participation is imaginary, and their knowledge has been appropriated by the sports coaches, journalists, and bureaucrats (Bourdieu 1984, p. 185).

It is clear that the anthropological perspective of Lévi-Strauss, the critical tradition of social theory of Bourdieu, and the historical sociology of Elias have something in common: sport is perceived as an important field of analysis for achieving a better understanding of the appropriate or contradictory functioning of modern societies. In the arena of sport, individual self-fulfillment is related to games producing winners and losers, to exalted nationalism in an age of international competition, to passivity and ideological domination, and sometimes to an abnormal quest for excitement and violence. No one will deny that sport is associated with these practices and values, but recent anthropological research has shown that it is much more.

In the new type of research, the key assumption is that sport is a ritual and a game at the same time, and

as such a cultural construction that makes symbolic communication among its participants possible. The content of the communication may vary according to the degree of formality, rigidity, concentration of meaning, and redundancy. However, the ritual is also a performance in the sense that saying something is also doing something; hence ritual action makes possible a connection between the meanings and values mobilized by the participants. In every ritual, various types of participants can be distinguished: the experts in knowledge, the central participants or players, and the peripheral participants or audience.

This approximation permits the consideration of a range of discourses, practices, and identities produced by various kinds of participants. A historical analysis of the impact of sports situates these narratives and practices in specific periods and places, and enables us to follow their transformations. The empirical analysis of sport can contribute to general theories of the meaning of sport in identity construction or of the significance of bodily exploits for expanding the scope of theories of ritual and cultural performances. MacAloon's (1984) analysis of spectacle, ritual, festival, and games in Olympic competitions exemplifies a longstanding anthropological concern with ritual. In this direction, the anthropological analysis of sport is not a reflection of society, but a means of reflecting on society. Sport, understood as a central and not marginal activity, is a fruitful entrance for capturing important cultural, historical, and social processes. Sports represent a complex space for the display of identities as well as an arena for challenging dominant social and moral codes, contrary to what Bourdieu sustained.

1. *Gender and Nationhood: Sports and Politics*

It has been argued that a nation is an 'imagined political community': its members share a sovereign boundary and have a strong feeling of communion. Hence, the ideology of nationalism should be integrated in social practices that can create an image of 'the people' having 'something' in common. DaMatta (1982) has argued that Brazil is a society articulated by the sharp division between the 'home' and the 'street,' and between the family—a system of hierarchical social relations and persons—and the market and free individuals. The role of soccer is privileged because the personalized social world of the home and the impersonal universe of the street are combined in a public ritual. Impersonal rules regulating the game make possible the expression of individual qualities: 'soccer, in the Brazilian society, is a source of individualization … much more than an instrument of collectivization at the personal level' (DaMatta 1982, p. 27). The male players escape from the fate of class or race and construct their own successful biographies in an open arena. Soccer makes it possible to experience equality and freedom, important values of nationhood, in hierarchical contexts.

In order to triumph, soccer players (like *samba* dancers) must have *jogo de cintura*, the capacity to use the body to provoke confusion and fascination in the public and in their adversaries. The European identification of a Brazilian style of playing relating football and *samba*, manifested in the expression '*samba*-football,' is therefore not an arbitrary creation: it is rooted in Brazilian self-imagery (Leite Lopes 1997). This identification establishes important cultural differences—the existence and development of European styles of playing are not linked to music and dance.

The case of Brazil is not unique. In Cuba the ritual of playing baseball, the national sport, has been associated with dancing: 'each baseball match culminated in a magnificent dinner and dancing, for which orchestras were hired for playing *danzones*' (González Echeverría 1994, p. 74). As in the case of soccer, baseball was perceived as a democratic and modern game that made it possible for young male players of modest origins to experience social mobility. In Argentina, the arrival in the 1900s of millions of European immigrants in Buenos Aires and its hinterlands made possible the rapid expansion of football and tango. Leisure was transformed; the practice of different sports was accompanied by the crystallization of the choreography of tango and the formation of new orchestras. Tango was first exported to Europe in the 1910s, then to the rest of Latin America and North America, becoming a 'universal' dance to express eroticism and modernity. Argentinian soccer players were exported to Europe as early as in the 1920s. Argentinian performing bodies—dancers, musicians, and soccer players—became highly visible in the world arena of leisure (Archetti 1999).

It is interesting that both baseball and soccer, two sports originating outside the Latin American countries, were integrated into the construction of the 'national' masculinity as cricket in the West Indies, Pakistan, or India. Rules, esthetics, and style in cricket have been transformed into something that is no longer typically English because the hegemonic systems of values its practice evokes is questioned (Appadurai 1995). The rich ethnography of Alter (1992) shows that wrestling as a genuine Indian sport provides a context within which concerns with fitness, health, and strength articulate with various forms of nationalism. In the vision of many wrestlers, the various forms of self-discipline associated with the sport provide the moral coordinates around which a new India is imagined and embodied. Thus, cricket and wrestling enter into a complex relation between modernity, masculinity, decolonization, and nationalism.

The practice of sports is 'gendered' and should be understood as expressing and articulating gender

differences. MacClancy's (1996) analysis of female bullfighting in Spain is a case in point, showing how anthropologists have unproblematically accepted bullfighting as exclusively male, despite the fact that for over two centuries women have been bullfighters, taking the same risk and bearing the same scars as the male *matadores*. Brownell (1995) has demonstrated that the international success of Chinese sportswomen has allowed them to represent the nation in a way that is unlike Western women's representations of their nations. The ability to endure pain and sacrifice as feminine values is part of the national discourse transforming Chinese women into national models. Sánchez León (1993) has analyzed the gendered meaning of the two most popular sports in Peru. Football, a male practice, is related to freedom and improvisation, whereas volleyball, a female practice, displays the sense of responsibility and discipline of Peruvian women.

2. *The Production and Adoration of the Heroes*

The world of sport heroes is a world of creative enchantments because momentarily, like flashes of intense light, athletes become mythical icons representing mastery over mortality. The heroic sports figures must be seen in their cultural context in order to understand their communal impact (Holt et al. 1996). Oriard (1993) has shown that the athlete-hero in America embodies the 'land of opportunity' and is therefore the most attractive self-made man. The hero sustains the American Dream, personifies the democratic ideal of open accessibility to prestige and allows all citizens to share his glory. In relation to American football, Oriard has demonstrated that in the historical production of a restricted heroic masculinity certain competing qualities were at stake. On the one side, simple, unambiguous physical force was emphasized—what he calls an antimodern value. On the other side, the required 'manly' qualities were temperance, patience, self-denial, and self-control—which lie behind the modern and technocratic aspects of modernity (Oriard 1993, p. 192). Oriard asserts that football's cultural power derived in large part from this collision of the modern and the antimodern, one aspect of which was this dialectical embrace of competing notions of manliness.

In many cases, sporting heroes are a source of collective pride in both national and supranational settings. The case of Maradona, an Argentinian soccer player, is a clear illustration of 'transnational' heroism. He was like a God-like hero in Naples while he was playing for the Naples Football Club and in Argentina while playing on the national team. His heroism was related to mythological figures in the respective local traditions: in Naples he was identified with the mythical figure of the *scugnizzo* and in Argentina with the *pibe*. These prototypes have something in common: they are human beings endowed with grace, elegance, creativity, and freedom (Archetti 1997). The cult of sport heroes provides a social model to perceive paradoxes and dramas in society, and to recognize and question key values. When Garrincha, the extraordinary right winger of the Brazilian soccer team which won the World Cup in 1958 and 1962, died an alcoholic and in extreme poverty in 1983, the entire nation was shocked. His funeral became a part of national history. The crowds following his casket in silence in the streets of Rio de Janeiro were participants in a social drama, showing the most profound respect to a son of the working class who had suffered from injustice, maladjustment, and prejudice. The nation wondered how it had been possible to abandon so heartlessly the player defined as 'the joy of the people' (Leite Lopes and Maresca 1992).

3. *Engaged Audiences and Supporters*

Bourdieu's image of passive supporters and alienated audiences has been contradicted by ethnographic findings. Neves Flores (1982) has demonstrated that in the arena of soccer the supporters 'narrate' social stories and even produce what he calls 'ideologies of transformation.' The popular clubs, such as Flamengo in Rio de Janeiro and Corinthias in São Paulo, are seen as courageous, with lots of stamina and will to win. The 'elite' clubs, such as Fluminense in Rio de Janeiro, are *machista* and show discipline and fair play. The clubs where sons of immigrants are most represented are usually perceived as less national than the others. Neves Flores argues that the organized activity of the supporters displays the class and ethnic divisions in society and, in this sense, act against the dominant values in Brazilian society of national equality, populist integration, and political paternalism. Bromberger (1995) has shown that the narrative of the supporters of AC Torino, a soccer club from northern Italy, emphasizes local suffering and extreme loyalty, while the narrative of the supporters of Juventus, the 'other' club in the city, is based on continuous victory, international accomplishments, discipline, and efficiency. The history of Juventus is at the same time the history of Fiat, the Italian company owning it. The fans assert, in a creative way, the 'right' to identity and difference in modern societies. With his rich ethnography, Armstrong (1998) challenges the assumption that violence is wholly central to the match-day experience of soccer supporters in England. Rather, the creation of identity is at the roots of hooliganism, with all the cultural values and rituals, codes of honor and shame, and communal patterns of behavior and consumption that accompany it.

Klein (1991) has argued that baseball is a privileged field for the analysis of cultural hegemony and resistance in the Dominican Republic. Dominican supporters use the very form and symbol of American

domination—baseball—to promote resistance. Adopting the sport of baseball has also influenced cultural movements. 'Dominicans infused the game with their own raucous, melodramatic style, marked by a highly individualistic way of playing, and easy-going attitude toward the game both on the field and on the stands, music and dancing, and crowds by turn temperamental and tranquil' (Klein 1991, p. 152). Klein has shown something unusual in an era of global influences: working-class supporters in the Dominican Republic prefer wearing the shirts of national clubs rather than American ones.

Bibliography

Alter J S 1992 *The Wrestler's Body: Identity and Ideology in North India*. University of California Press, Berkeley, CA

Appadurai A 1995 Playing with modernity: The decolonization of Indian cricket. In: Breckenridge C A (ed.) *Consuming Modernity: Public Culture in a South Asian World*. University of Minnesota Press, Minneapolis, MN

Archetti E P 1997 'And give joy to my heart': ideology and emotions in the Argentinian Cult of Maradona. In: Armstrong G, Giulianotti R (eds.) *Entering the Field. New Perspectives on World Football*. Berg, Oxford, UK

Archetti E P 1999 *Masculinities. Football, Polo and the Tango in Argentina*. Berg, Oxford, UK

Armstrong G 1998 *Football Hooligans. Knowing the Score*. Berg, Oxford, UK

Bourdieu P 1984 *Questions de Sociologie*. Minuit, Paris

Bromberger C 1995 *Le Match du Football. Ethnologie d'une Passion Partisane à Marseille, Naples et Turin*. Editions de la Maison des Sciences de l'Homme, Paris

Brownell S 1995 *Training the Body for China: Sports in the Moral Order of the People's Republic*. University of Chicago Press, Chicago

DaMatta R 1982 Esporte na sociedade: um ensaio sobre o futebol Brasileiro. In: DaMatta R, et al (eds.) *Universo do Futebol: Esporte e Sociedade Brasileira*. Edicoes Pinakotheke, Rio de Janeiro, Brazil

Elias N 1986 Introduction. In: Elias N, Dunning E (eds.) *Quest for Excitement. Sport and Leisure in the Civilising Process B*. Blackwell, Oxford, UK

González Echeverría R 1994 Literatura, baile y beisból en el (último) fin de siglo Cubano. In: Ludmer J (ed.) *Las Culturas de Fin de Siglo en América Latina*. Beatriz Viterbo Editora, Rosario, Brazil

Holt R, Mangan J A, Lanfranchi P (eds.) 1996 *European Heroes. Myth, Identity, Sport*. F. Cass, London

Klein A M 1991 *Sugarball. The American Game, the Dominican Dream*. Yale University Press, New Haven, CT

Leite Lopes J S 1997 Successes and contradictions in 'multiracial' Brazilian football. In: Armstrong G, Giulianotti R (eds.) *Entering the Field. New Perspectives on World Football*. Berg, Oxford, UK

Leite Lopes J S, Maresca S 1989 La disparition de la 'joie du peuple'. *Actes de la Recherche en Sciences Sociales* **79**: 21–36

Lévi-Strauss C 1966 *The Savage Mind*. Weidenfeld & Nicolson, London

MacAloon J J 1984 Olympic Games and the theory of spectacle in modern societies. In: MacAloon J J (ed.) *Rite, Drama, Festival, Spectacle: Rehearsals Toward a Theory of Cultural Performance*. Institute for the Study of Human Issues, Philadelphia

MacClancy J 1996 Female bullfighting, gender stereotyping and the state. In: MacClancy J (ed.) *Sport, Identity and Ethnicity*. Berg, Oxford, UK

Neves Flores L F B 1982 Na zona de agriao. Algumas mensagens ideológicas do futebol. In: DaMatta R et al (eds.) *Universo do Futebol: Esporte e Sociedade Brasileira*. Edicoes Pinakotheke, Rio de Janeiro, Brazil

Oriard M 1993 *Reading Football. How The Popular Press Created an American Spectacle*. University of North Carolina Press, Chapel Hill, NC

Sánchez León A 1993 *La Balada del Gol Perdido*. Ediciones Noviembre Trece, Lima, Peru

E. P. Archetti

Sport, Geography of

The sports central to geography's concern are institutionalized contests, frequently professional in their highest form and attractive to spectators, involving the use of vigorous physical exertion between individuals or teams of human beings. The geography of sport is a recent development, largely cultural-historical in approach. It maps and explains the changing distribution patterns of individual sports and packages of national, major, and minor sports at levels ranging from the world to district (Bale 1982). It relates these changes to the character and form of stadiums and landscape ensembles. It also deals with the quality of theater—interactions between players and spectators—in different types of stadiums (Bowden 1995). New directions of research focus on sport in the formation of working-class culture, the effects of globalization and televised sports on the distribution of sports and the map of rapidly emerging women's sports.

1. Modern Sports: Origins and Dispersals

The distribution of modern sports in the twentieth century is explained in a model of (a) primary origins in southeast England; (b) dispersal to the British Isles, Anglo-America, and the British Empire; (c) secondary origins in northern Britain and in northeastern Anglo-America; and (d) dispersal from the latter mainly to the rest of North America and the Caribbean Basin, and from Britain to the rest of the world.

The birthplace of modern sports in Metropolitan England is attributed by some to the development of the mathematical-empirical world view after 1660. Marxists correlate sport with the birth of industrial

capitalism in the 1700s (Guttmann 1994, Holt 1989). Others point to Thomas Hughes's *Tom Brown's Schooldays* (1857) which glorified Thomas Arnold's education system at Rugby School, but maintained that moral development and character training inculcated by cricket and football outranked intellect in the curriculum.

Hughes's transformation of Arnold, abetted by the ruling classes, inspired the cult of athleticism in the elite schools and universities which built principally on the first modern game—cricket—codified by the aristocracy in the eighteenth century (see James 1963). It was joined in the public schools in the middle third of the nineteenth century by hockey and athletics and by the Rugby School and dribbling games. Neither of these football games were played after the 1870s, but seven modern forms of football developed from them and one of these forms of football has become the major team sport played by men in at least one season in practically every country in the world, and by women in most First World countries and in some Third World countries. Research in the 1990s goes far to explain the diffusion and distribution of the different types of football 1870–2001 as a classic case in the geography of sport.

2. *Football Exceptions to Soccer's World Dominance*

Modern association football originated in England. By the 1880s it was the national game in England and Scotland, and diffused thereafter to become the most popular team sport throughout the world. The exceptions are seven English-speaking countries: Wales and Ireland, and five countries settled substantially from Britain—USA, Canada, Australia, South Africa, and New Zealand. The explanation lies in (a) the preference of the majority of English public schools for the more manly rugby compared to the dribbling game played by the minority, and its consequence: rugby's greater popularity among the upper and middle classes in Britain and in a majority of regions in the British Isles 1848–78, and (b) a developing English upper and middle class prejudice against the dribbling game (association football after 1863), increasingly associated in the 1870s with working class industrial regions (Murray 1994, Russell 1997).

All of this ensured that the predominantly middle class colonists and their public-school and Oxbridge-educated leaders in New Zealand and Australia and in English-speaking Canada and South Africa in this midcentury period chose to carry with them Britain's most popular football game of the time—rugby (Holt 1989). That they did so to the exclusion of dribbling soccer is explained by Foster's law of simplification in which in the survival phase of early settlement, only one of many cultural forms in the diverse homeland is chosen by colonists, usually the form found in the region whence came most of the first settlers: rugby in southern England before 1878 (see Foster 1960).

The Rugby School game—the rugby current before the rules regularization banning the violent 'hacking' by the Rugby Union in 1871—provided the framework for the nationalist forms of football: Australian Rules played in Australia outside Queensland and New South Wales (Hibbins 1992) and Canadian and American football derived from the Harvard and McGill universities' encounters in 1874.

The peaks of British settlement occurred in eastern Australia, New Zealand, and South Africa late enough that the Rugby Union code became established there in the 1870s. This was challenged in the 1900s in eastern Australia and New Zealand by the new 13-a-side Rugby League professional code developed in northern England.

Before 1880 rugby was preferred over dribbling soccer in Ireland and Wales. The continued preference for rugby over the new passing soccer after 1880 results from cultural nationalism. In Ireland the Gaelic Athletic Association modernized Gaelic football to be the national game, and its ban on English games worked selectively against 'new' soccer relative to well-established rugby. In Wales the under-the-table payment of Welsh 'amateurs' pre-empted professional soccer's attractiveness to Welsh athletes and gave them the chance to beat the middle class English at their own game (Murray 1994, Holt 1989, Wagg 1995).

In the 1860s and 1870s in Argentina, Belgium, Sweden, and France both rugby and soccer were adopted. The attractiveness of the new soccer in the early 1880s relegated rugby thereafter to the status of a minor sport in Argentina, and replaced it in Sweden and Belgium. But in France rugby union parried the counterattack of new soccer after 1880, as did rugby league in southwest France after 1900, and both contributed to a long delay in soccer's elevation as the national team sport (Guttmann 1994, LaBlanc and Henshaw 1994).

3. *Diffusion of Soccer: The World's Sport*

The football adopted in every other country in the world after 1880 was soccer. The explanation for this lies in the revolutionary transformation of soccer from the dribbling game played by unspecialized players into a simple game of high skill emphasizing accurate long- and short-ball passing, wing-play, and the crossing of the ball to forwards specialized in shooting and heading, and positional specialization. The revolution began in northern Britain, produced two professional leagues (England 1888; Scotland 1891) and swept through Great Britain, leaving rugby union more popular only in southwest England and the

Scottish border, marginal rural regions with small towns and no large cities—and in south Wales (Russell 1997, Holt 1989, Murray 1994).

4. *Diffusion and Adoption*

During the first period of diffusion of soccer to independent countries, 1890–1940, the adoption of soccer in each country followed a standard pattern. A gradual transmission of the game to young members of the local elite from: (a) British resident businessmen and agents of trading houses in the large port(s); (b) members of the British colony in the capital; (c) expatriate British teachers in schools and colleges for British residents and the local elite; and (d) elite students returning with a soccer ball from educational sojourns in Britain. Often towards the end of this gradual phase of adoption and formation of teams by denizens, members of the middle class picked up the game from the elite and also in the large inland industrial and mining towns from British mining and railway engineers and textile mill owners and skilled workers. The initial 15–20 year phase of gradual adoption by the elite was followed by a filtering-down process and an explosive 10–15 years of rapid adoption by the middle and working classes in urban areas. The three phases of the S-curve pattern of the diffusion of innovations are completed with the 10–15 year slackening of the rate of adoption by the laggards: the market-townsmen and agricultural workers/peasants of the thinly populated countryside (Guttmann 1994, Russell 1997).

The critical signifier of the establishment of soccer as a country's major form of football (national team sport) is the date of creation by the national football association of the national club championship (data from LaBlanc and Henshaw 1994, Wagg 1995). The plotting of these dates of founding of national championships establishes three discrete models of the spread of soccer worldwide, variants of the classic patterns of spatial waves and S-curve time-series of diffusion (see Rogers 1983) in men's soccer to: (a) 35 independent nations 1890–1945, peaking in the 1920s; (b) 65 newly independent nations 1946–96, peaking in the 1960s; and, in women's soccer, to (c) 65 countries 1957–2000, peaking in the 1970s–1980s.

5. *Traditional Men's Soccer: Europe, the Middle East, and Latin America*

In the first diffusion-wave model, beachheads developed in Denmark, The Netherlands, Switzerland, Italy, and Argentina in the 1890s. In the 1900s the second wave pushed out to Norway–Sweden; Belgium–Germany; Austria–Hungary–Bohemia; Spain; Uruguay–Chile–Paraguay, and to a new beachhead in Mexico. The third wave of the 1910s, blocked in Europe by World War 1 and by continued Czarist and Ottoman imperial suppressions, broke only on the shores of Iceland, France, Portugal, Algeria–Morocco, Haiti, and Brazil. The fourth wave completed the conquest of the European outer margins (the three Baltic countries and Poland, Soviet Union, and three Balkan countries and Greece), swept into the Middle East and North Africa (Turkey, Israel/Palestine, Egypt, Tunisia), broke in northern Latin America (Bolivia–Peru–Ecuador–Colombia, Surinam, Costa Rica–Guatemala), and reached East Asia (Thailand, Japan/Korea). The last countries reached by the outer wave of the 1930s were Albania, El Salvador, and Honduras.

6. *Postcolonial Men's Soccer: Asia, Africa, and Islands*

Most of the countries that adopted modern soccer after 1945 were former colonies of European powers. Soccer was the team sport offered in missionary schools, in private schools mainly for Europeans, and to the local elite in the main cities. Mine-owners and railways used it as a means for social control of workers. Lack of investment by the colonial powers in the provision of social institutions for the vast majority ensured that the game did not spread widely beyond these two groups in the large towns. The filtering-down process often extended over 40 years before emergent nationalism and independence accelerated the process and united the masses behind national soccer teams.

Soccer was established as a national sport earliest in the former British, Dutch, and Belgian colonies because of soccer's earlier acceptance by the three colonial powers compared to France and Portugal and because Britain, Holland, and Belgium established the European national association/championship model in each colony's development. By contrast the imperial family model developed in the French and Portuguese colonies funneled the better colonial players into the French and Portuguese leagues thereby stunting the development of colonial ones.

Thus, in the post-World War II period the establishment of soccer as the national team sport took place first in the 1950s in the British colonies in the Mediterranean (Malta, Cyprus), North Africa (Ghana, Sudan), and East Asia (Burma, Malaysia), the Dutch colonies in the East and West Indies and in Congo–Kinshasa (formerly Belgian). Most of the 38 countries that formed national associations and club championships in an unprecedented decade of rapid establishment of soccer in the 1960s were the larger British Caribbean islands, former British colonies in eastern and southern Africa, and the coastal countries of Francophone Africa. In the long third phase of slackening adoption from 1970 onwards, the laggards were land-locked Francophone countries, war-torn

Angola and Mozambique, the smallest British Caribbean islands and island nations in the Atlantic, Indian, and Pacific oceans, their numbers boosted in the 1970s by the new Gulf States (Murray 1994, Wagg 1995).

7. *Postmodern Women's Soccer: The First World*

After decades of being rebuffed by national football associations and FIFA, women's soccer began its international phase of growth in the 1950s in its heartland—Germany and England—spreading slowly to seven neighboring European countries by 1969 when the English Football Association became the first to recognize women's soccer officially. The Women's Football Association founded there in 1971 was soon followed by other European national football associations, of which Germany's established the first national championship in 1974.

The European-only period of gradual adoption to 1969 gave way 1970–89 (with a prospective Women's World Cup in 1991) to an explosive adoption of women's soccer in another 40 countries, mainly in Europe and East Asia, but significantly in the new postmodern football cultures of Japan as well as USA, Canada, Australia, New Zealand, and South Africa where women are choosing soccer over the more physical and 'manly' forms of football favored by the majority of men. In the First World and in the northwest European heartland, until recently the unchallenged preserve of women's field hockey, it is the young middle class women of metropolitan suburbia with their greater freedom within the realms of lifestyle and leisure politics who dominate women's soccer. However, a small but significant minority of club and international players hail from working-class male strongholds (also a source of soccer-playing women in Brazil). Conspicuously under-represented on the world map of women's soccer are Latin American countries where working class men and macho values inhibit women's soccer. Absent are the countries of sub-Saharan Africa where female soccer is culturally proscribed, and those of Western Asia and North Africa where Muslims and Hindus demand sharp cultural divisions between male and female (Giulianotti 1999, LaBlanc and Henshaw 1994).

See also: Diffusion: Geographical Aspects; Diffusion, Sociology of; Entertainment; Play, Anthropology of; Sport, Anthropology of; Sport, History of; Sport, Sociology of; Sports, Economics of

Bibliography

Bale J 1982 *Sport and Place: A Geography of Sport in England, Scotland and Wales*. C Hurst & Co., London

Bowden M J 1995 Soccer. In: Raitz K (ed.) *The Theater of Sport*. The Johns Hopkins University Press, Baltimore, MD

Foster G M 1960 *Culture and Conquest*. Quadrangle Books, Chicago

Giulianotti R 1999 *Football: A Sociology of the Global Game*. Polity Press, Cambridge, UK

Guttmann A 1994 *Games and Empires: Modern Sports and Cultural Imperialism*. Columbia University Press, New York

Hibbins G M 1992 The Cambridge connection: The English origins of Australian Rules Football. In: Mangan J (ed.) *The Cultural Bond: Sport, Empire, Society*. Frank Cass, London

Holt R 1989 *Sport and the British: A Modern History*. Oxford University Press, London

James C L R 1963 *Beyond a Boundary*. Stanley Paul, London

LaBlanc M, Henshaw R 1994 *The World Encyclopedia of Soccer*. Visible Ink Press, Detroit, MI

Murray W J 1994 *Football: A History of the World Game*. Scolar Press, Aldershot, UK

Rogers E 1983 *Diffusion of Innovations*, 3rd edn. Free Press, New York

Russell D 1997 *Football and the English*. Carnegie Publishing, Preston, UK

Wagg S (ed.) 1995 *Giving the Game Away: Football, Politics, and Culture on Five Continents*. Leicester University Press, London

M. J. Bowden

Sport, History of

1. *Two Definitions of 'Sport'*

The English word 'sport' is etymologically derived from the late Latin 'deportare.' Originally it meant 'to divert yourself,' 'have fun,' 'amuse yourself'—in short everything which you would do 'in sport' or 'for love.' In the narrow sense of the word it meant the hunt for 'game,' a semantic connection derived from mediaeval aristocratic practices, a term which is still used in English. Horse-racing was also described as 'sport' in early-modern times. The word was first taken into everyday English usage, however, in the mid-nineteenth century when new games and pastimes became popular amongst the middle classes. In this period 'sport' became an umbrella term for ball games (cricket, football, croquet, and hockey), martial arts (boxing, fencing, and wrestling, etc.), heterogeneous outdoor pastimes (boating, riding, cycling, skating, swimming, and mountain-climbing) as well as for so-called 'athletics' (running, jumping, javelin throwing). This list of disciplines can be found in a general article on English sporting life published in 1899 (Aflalo 1899). One hundred years later a sociological survey for the city of Hamburg came to the conclusion that there were more than 240 different sporting disciplines, from BMX-cycling, to bungee jumping, and free climbing on artificial mountain walls (Heinemann and Schubert 1994).

The increasing variety of games and pastimes can have different forms of organization (neighborhood and circles of friends, clubs, schools and universities, the armed forces, political parties, or the State) and types of income (admission charges, private sponsoring, public grants). Some disciplines are popular,

others are not. Some are followed with the support of modern mass media, others have retained a more private character. Nonetheless, these games and pastimes all have one common feature: they are competitive. Therefore, they encompass only a narrow understanding of the term 'sport.' For, in the course of the development of competitive sports into a mass phenomenon and their growing significance in education, the armed forces, the economy, and culture, there has been an increasing tendency in many languages to use the word 'sport' (or its international usage as in 'deporte' (Spanish), 'desporto' (Portuguese), 'spor' (Turkish), 'сПорт' (Russian)) also for noncompetitive activities and institutions, for which there are many other—and mostly more precise—terms. Amongst these may be counted gymnastics and physical education, dancing, forms of active tourism such as rafting, and even rehabilitation. In this broad usage 'sport' today has become a collective term for many and diverse expressions of the general physical culture of mankind, including 'dead' cultures like the physical culture of the ancient Orient or the ἀγών (agon) of the ancient Greeks. Both usages of the word 'sport' have given rise to different directions in the study of the history of sporting activity.

2. *Sport History as the History of the Physical Culture of Mankind*

The direction based on the broader usage of the word—'sport history' as the history of physical culture—is the older of the two. Amongst the founders of the discipline in this sense can be counted the chronicler of the Greek ἀγών Pausanius (c. 115–180), the Enlightenment educationalist Johann Christoph Friedrich GutsMuths (1759–1839) and the 'father of gymnastics' Friedrich Ludwig Jahn (1778–1852). Their nineteenth and twentieth century descendants were physical education teachers, club and association officials, doctors, hygiene experts, military officers, journalists, and writers who have researched themes like mediaeval bathing cultures, physical bearing in the Renaissance, or the history of their local gymnastic clubs.

The great majority of these 'amateurs' were, and still are, not professional historians, but 'amateurs' and part-time writers. Thus, they have little experience in dealing with sources, do not know how to arrange their results in a more comprehensive context, and tend to indulge in off-the-cuff philosophizing about allegedly anthropological constants in human physical activities. Methodical studies are of little interest to these scholars; indeed, a good deal of them used their research into sport history to legitimize their own political standpoint. This is particularly conspicuous in publications of the 1920s and 1930s when continental Europe aired conflicts between so-called 'bourgeois' and working-class sport associations; and in later writings between 1950 and 1970 which were influenced by cold-war ideologies. They often yielded more information about the political and moral values of the respective authors than about the subject that they have researched.

The largest fund of knowledge in the field of sport history, however, is produced by academics who have made a name for themselves in university departments for physical education (*Institute für Leibesübungen*)—most particularly in German-speaking countries from the 1920s onwards and again since the 1960s. With the exception of some experts in ancient sport who are trained classical historians, a great many of these sport academics have been educated mainly in the natural sciences. Therefore, they do not make it their ambition to write elaborate historical studies. Nevertheless, they must be credited for their assiduous work in collecting and documenting sources and literature about sporting incidents, persons, institutions, and events in history.

All the more so because very early on they cast their gaze beyond national boundaries, as documented by representative volumes of collected essays like *Geschicte der Sport aller Völker und Zeiten* (Bogeng 1926) or *Geschichte der Leibesübungen* (Ueberhorst 1972/1989). Academic exchanges have occurred since the 1970s with the aid of societies like the *International Committee for the History of Physical Education and Sport* (ICOSH, founded in Prague 1967) and the *International Association for the History of Physical Education and Sport* (HISPA, founded in Zurich 1973); these two societies, representing the East and the West, amalgamated into the *International Society for the History of Physical Education and Sport* (ISHPES) in 1989. In addition, there are more specialized societies that promote exchanges of information on North America, Australia, and Europe.

3. *Sport History as the History of Competitive Games and Pastimes*

A second direction in sport history, which primarily regards modern competitive sport as its theme—i.e., is orientated towards 'sport' in the narrow sense of the word—has been adopted by professionals who have moved into sport history after a rigorous academic training in another academic discipline. Most of these scholars are historians, whilst the majority of the rest are sociologists, anthropologists, political scientists, and economists. A few are geographers, architects, or media specialists. Sport academics are only included in this group when they can be counted—as in France and the US—as trained social scientists.

This second category of sport historians is comparatively new. It sprang up in the 1970s in Great Britain, Australia, and the US, where sporting competition traditionally has been a part of the educational system. In the second half of the 1980s it also found

adherents on the European continent, and today it is in the process of spreading to South America, Africa, and Asia. Given the lack of local experts, however, studies on these regions of the earth have up to now mostly been conducted by European and American scholars.

International exchanges take place at conferences and in publications like *Sporting Traditions* (founded in 1977 in Australia) or the *British Journal of Sports History* which in 1987 became *The International Journal of the History of Sport*; further at research centers such as the *Center for Olympic Studies* at the University of Western Ontario, the University of Neufchâtel's *Centre International d'Étude du Sport*, or—in the UK—the *International Research Centre for Sport and Socialisation in Society* at the University of Strathclyde and the *International Centre for Sports History and Culture* at DeMontfort University, Leicester (this institute employs an international staff). In contrast to the first category of sport history, the international cooperation amongst these younger scholars is marked by a clear orientation towards problem-solving, and a certain pleasure in interdisciplinary experimentation, which in different ways enables themes to be looked at from completely heterogeneous perspectives. To give but one example: the phenomenon of spectators rioting in football stadiums which has been observed since the nineteenth century and which has been the object of comprehensive studies in the 1990s by historians not to speak of sociologists, social workers, architects, and media specialists (for an overview see Williams 1991).

A great many amateur historians are also active in the area of competitive sport, not least amongst them the oft-derided collectors of 'sportifacts,' who punctiliously generate lists of trophy winners, game scores and scorers, season and career averages, chronological events of team and club performances. Since these data, when put into context, have some value as source material, professional historians have a considerable interest in cooperating with the amateurs by supervising their work and giving them the chance of being published in special journals like *Sporting Heritage*.

The cooperation with academic sport scientists, however, is not as intense as it could be. One reason is that the two factions put the main stress on different subject matters, another is to be seen by the fact that scholars mainly belong to different generations. The leading sport historians in sport science were born before the World War II, whereas the social science faction, with few exceptions, were all born in the 1950s and after. Whereas the older generation grew up at a time when educationalists and politicians were interested in promoting virtues like self-discipline and strength in sport, the younger generation of research workers, which incidentally also includes women, has been imbued with a different understanding of sport. Massive state subventions during the cold-war period, expansion of educational opportunities and an increase in leisure since the 1960s, and above all the breakthrough of television and the concomitant commercialization tendencies, have not only transformed sport into a mass phenomenon but also helped to wipe out the old virtues and produce a mentality of hedonistic sporting consumption. While many of their older counterparts are irritated by, and even disgusted at this development, it is well known to the younger generation, and tracing back the traditions of this mentality is a matter of course for them.

4. *Sport History, Social History, and the Social Sciences*

The formation of this second category in sport history around the turn of the 1980s is remarkable because both professional historians and social scientists had tended to ignore sport up to this time. Therefore, alongside the generation-specific interest of young academics in sport, there needed to be a change in the disciplines of history and social sciences themselves in order to embrace sport history as a natural object of study.

The decisive change came with the rapid development of social history and historical sociology that began in the 1960s (see *Social History*). This boom was accompanied by an intensive reception of modernization theories, which made a fundamental contribution to the interest of scholars in the *social form* of modern sports, i.e., the rules, constraints, and organization of sporting competition (see *Modernization and Modernity in History*). For the features of this social form fitted very well into the categories of Max Weber and Talcott Parsons. 'Equality of opportunity to compete and in the conditions of competition,' 'specialization of roles,' 'rationalization,' 'bureaucratic organization,' 'quantification,' and 'quest for records': these were the crucial characteristics of modern sport as established by Guttmann (1978). Whereas historians and social scientists generally felt they had no responsibility for bodily movements and behavior in human beings—something which the first category of scholars from sport studies had no hesitation in writing about—the above definition seemed to imply that there might be a connection between the social formation of modern sport and the formation of modern Western society. Which political, economic, and social conditions were favorable to organized and thoroughly regulated competition? Who were the actors? What motives connected them with sport? What were the effects of their passion for sport? These and other questions were raised.

Significantly, the one modernization theorist still living at the time, the civilization theorist Elias, dedicated his work to sport research for a time. One of his hypotheses was that the emergence of organized sport as a form of relatively nonviolent physical combat was connected with a period in English history

when Parliament became a battleground for resolving and defusing interests and conflicts (Elias and Dunning 1986; see also Elias 1971).

As exciting as were scholars' attempts to emphasize the contexture of the development of sport, their theoretical ambitions began to evaporate as soon as they began exploring sources in more detail. During their researches they all discovered that modernization and other 'great' theories about society provide in many ways an unsuitable framework in which to interpret sport history.

One reason was the focus of 'great theories' on industrial society. This focus did not allow to take into account that competitive sport, at least in its 'mother country' England, had sprung up in pre-industrial times. There the most protracted impulse for development had been the early formation of a market society and the commercialization of everyday life: which explains why seemingly modern professional sport has a longer tradition in Britain than amateur sport (see *Consumption, History of*). On the other hand it was discovered that, precisely because of the rationality of its social form, sport could be effective as a transmitter of 'premodern' attitudes. Thus, in its land of origin, the notion of 'fairness,' e.g., evolved from informal aristocratic and popular cultures that can be traced back to the Middle Ages. There were also some other elements which, with the help of sport, bound 'chivalrous' gentlemen into modern civil society. Modernization theorists had not foreseen such paradoxes and even Elias, who was well acquainted with the sources, seemed at a loss to know what to do with them.

Further communication problems with interpretations based on modernization theories resulted from the increasing confirmation of research results which showed that, everywhere apart from Great Britain, modern sport was a cultural import, and that it was disseminated by international elites whose identities and loyalties were anchored not only in, but also across states, a factor which was partly responsible for disseminating a uniform, global image. Established historians and social scientists were unable to say much about this phenomenon because their studies were traditionally fixed on the modern nation state. Hence, the method of international comparison which they recommended was of some—if insufficient—help; and those sport historians who studied themes like the establishment of sporting culture in the British Empire or its diffusion in Europe as a side-effect of tourism, business contacts, and technology exchanges (Mangan 1988, Lanfranchi 1991, Guttmann 1994, Eisenberg 1997) became of necessity pioneers of cultural transfer studies in a global framework.

The significance of cultural transfer in the development of modern sport puts a different perspective on such research methods that, like oral history (see *Oral History*) or thick descriptions, were increasingly asserting themselves in social history in order to be able to reconstruct the development in small communities more accurately. True, such methods were used by sport historians, especially for Ph.D. theses (e.g., to research the social basis and the symbolic content of certain sports), and in individual cases (e.g., research work into how competitive sport was operated in totalitarian regimes) they brought to light highly surprising results. However, when it came to *explaining* social and sporting changes, this form of research contributed very little. Furthermore, completely different questions had to be addressed, such as the influence of single persons, governments, and military officers, especially in both world wars, the impact of rule changes by world sporting authorities, or media imposition of patterns of perception.

These findings gradually have led to a loosening of the relationship of sport historians to their 'parental' disciplines, historical science and historical sociology (see *Historical Sociology*), despite their fundamental ties. That this tension has not been expressed openly, nor led to an antitheoretical offensive, as happened some years ago with representatives of modern cultural history, can be attributed to the priority given to empirical research, and also to the fact that sport historians have been lucky enough to be able to pursue their interests elsewhere. The rejection of 'great theories' occurred at a period of intensive reception of 'smaller' problem-related 'theories of the middle range' (Robert Merton), which were being developed by colleagues in sociology and economics (Neale 1964, Loy and Kenyon 1969). Some historians such as Vamplew contributed actively to this discussion (Vamplew 1988).

This direction in research grew out of the awareness that social relations in sport present in many ways a different picture to those in 'real life.' A commonly given example is the nature of sporting competition. Although most businesses would regard a monopoly as the ideal market situation, in sport this would be an economic disaster: it is of little commercial value to be a world champion if you have no challengers. Another example is the financial actions of clubs and sports firms, where ambitions for success often override profit considerations. Furthermore, the borders of class, gender, and ethnicity often run differently in sport from those in society which, in addition, are considerably less fluid.

Such observations fascinated sport historians, not the least because they raised the value of the object under review. For they demonstrated that modern sport was in no way a 'mirror of society,' as was keenly asserted. It was much rather a relatively autonomous social subsystem, which functioned according to its own rules on the basis of its 'built-in' game character, and which in certain situations produced its own dynamic of development. Where this subsystem interacted with its social and economic environment, it would not only reproduce this environment, but also impose its own mark on this

environment. It could contribute to the erosion of outdated traditions like premodern value systems or social inequality; and conversely it could help to extend the life span of such anachronisms. Sport was an active, yet selective, impulse of social change.

5. The Future of Sport History

Reconstructing this capacity of sport from individual cases will be a central task for the future of sport history as a branch of social history. For this task to be fulfilled satisfactorily, sport historians, social historians, and social scientists must cooperate. Alas, the reception of the results of sport history research amongst the scientific community has left a lot to be desired. To write a synthesis of the scattered findings of the last few years is, therefore, not only a desideratum but an urgent necessity.

In addition, sport historians should target and close gaps in research. Given the fact that recent research work has almost exclusively covered the period from the eighteenth century to World War I this means, first of all, an extension of the chronological scope. A far broader look at the premodern phase of sport would be most fitting, especially a study of the effects in Great Britain of the Renaissance, which is a key epoch for questions of continuity and discontinuity between ancient and modern ideals in sport. This question, which can only be answered in cooperation with interested classical historians (cf. the journal *Nikephoros*), has been completely neglected to date.

It is however far more important for sport history to turn its attention to the later decades of the twentieth century. What is needed here is, first, the compilation of a comprehensive data collection covering the social basis of sport, the biographies of influential actors, the development of organizations on a national and international level, financing methods, etc., as has been done in the last few years for the nineteenth century. Second, the substantial entanglement of sporting associations and totalitarian regimes (e.g., Europe between the world wars, Eastern Europe, South America, and China after 1945) must be made the subject of research studies. In this respect it would be of great help if a special branch of political science could be evolved, analogous to those of sport economy and sport sociology in recent years. For there are some signs which show that relationships of dominance and power in sport take on different shapes from those in other areas of social life.

Furthermore, it is desirable for social science oriented sport historians to broaden their thematic range, i.e., extend their research activities to cover themes of the physical culture, which to date have mostly been left to amateurs and sport scientists: the movements of the human body, sexuality, education, gender-fixing, etc. Such studies should obviously avoid making naive statements about anthropological constants in humanity. They should rather reveal how regulated and organized modern competitive sport emanates into other areas of social life, what effects it has, and what consequences result from this. In other words, scholars should research subjects like the correlation between doping and the contemporary medical consumption (the first studies on the subject have already been conducted by Hoberman 1992), between the physical appearance of athletes and modern bodily ideals in the fashion industry and on the labor market in the present-day service economy, between certain easy-going manners and expressive forms of communication in sporting arenas, and in other areas of modern social life.

Because this is only possible in close cooperation with academic disciplines like media studies, which in many countries are still in the process of coming into being, there is a danger that sport history will lose its own profile and independence. But if, by doing so, it can become an established part of a methodically reflected project of writing social history in the twenty-first century, this would only be to its own advantage.

See also: Mass Media and Sports; Sport, Anthropology of; Sport, Geography of; Sport, Sociology of; Sports as Expertise, Psychology of; Sports, Economics of

Bibliography

Aflalo F G 1899 The sportsman's library. A note on the books of 1899. *Fortnightly Review* **72**: 968–76

Bogeng G A E (ed.) 1926 *Geschichte de Sports aller Völker und Zeiten*, 2 Vols. Seemann, Leipzig, Germany

Eisenberg C (ed.) 1997 *Fußball, soccer, calcio. Ein englischer Sport auf seinem Weg um die Welt*. Deutscher Taschenbuch Verlag, Munich, Germany

Elias N 1971 *The genesis of sport as a sociological problem*. In: Dunning E (ed.) *The Sociology of Sport*. Cass, London

Elias N, Dunning E 1986 *The quest for excitement in leisure*. In: Elias N, Dunning E (eds.) *Quest for Excitement: Sport and Leisure in the Civilizing Process*. Blackwell, Oxford, UK, pp. 63–90

Guttmann A 1978 *From Ritual to Record. The Nature of Modern Sports*. Columbia University Press, New York

Guttmann A 1994 *Games and Empires. Modern Sports and Cultural Imperialism*. Columbia University Press, New York

Heinemann K, Schubert M 1994 *Der Sportverein. Ergebnisse einer repräsentativen Untersuchung*. Hofmann, Schorndorf, Germany

Hoberman J M 1992 *Mortal Engines. The Science of Performance and the Dehumanization of Sport*. Free Press, New York

Holt R 1989 *Sport and the British. A Modern History*. Oxford University Press, Oxford, UK

Holt R 1992 Amateurism and its interpretations. The social origins of British sport. *Innovation* **5**: 19–31

Lanfranchi P 1991 Fußball in Europa 1920–1938. Die Entwicklung eines internationalen Netzwerkes. In: Horak R, Reiter W (eds.) *Die Kanten des runden Leders. Beiträge zur europäischen Fußballkultur*. Promedia, Vienna, pp. 163–72

Loy J W Jr., Kenyon G S (eds.) 1969 *Sport, Culture, and Society. A Reader on the Sociology of Sport*. Macmillan, London

Mangan J A ed. 1988 *Pleasure, Profit, Proselytism. British Culture and Sport at Home and Abroad 1700–1914*. Cass, London
Neale W C 1964 The peculiar economics of professional sport. *Quarterly Journal of Economics* **78**: 1–14
Nikephoros. Zeitschrift für Sport und Kultur im Altertum ff. Weidmannsche Verlagsbuchhandlung, Hildesheim, Germany
Poliakoff M B 1987 *Combat Sports in the Ancient World: Competition, Violence and Culture*. Yale University Press, New Haven, CT
Ueberhorst H (ed.) 1972/1989 *Geschichte der Leibesübungen*, 6 Vols. Bartels und Wernitz, Berlin, Germany
Vamplew W 1988 *Pay Up and Play the Game. Professional Sport in Britain 1875–1914*. Cambridge University Press, Cambridge, UK
Williams J 1991 Having an away day: English football spectators and the hooligan debate. In: Williams J, Wagg S (eds.) *British Football and Social Change. Getting into Europe*. Leicester University Press, Leicester, UK, pp. 160–84

C. Eisenberg

Sport Psychology: Performance Enhancement

Enhancing sport performance is part of the discipline of psychology known as sport psychology. Sport psychology is the scientific study of people in sport and exercise contexts. There are two main objectives of sport psychology: (a) learning how psychological factors affect an individual's physical performance, and (b) understanding how participation in sport affects a person's psychological development, health, and well-being (Weinberg and Gould 1995).

Although traditionally the emphasis has been on the first objective, both objectives will be considered. The rationale for this decision is that of the millions of individuals, especially children and adolescents, who play sports, only a tiny fraction will parlay those activities directly into a career in sport. For the rest, growing up means further defining their identity, discovering other skills and interests and, potentially, applying some of the principles learned during sport participation to their adult pursuits. Sport psychology, then, involves the use of sport to enhance competence and promote human development throughout the life span, and practitioners should be as concerned about 'life' development as they are about athletic development (Danish and Nellen 1997).

1. *Why Study Sports?*

The impact of sports on US society is pervasive. US culture places a tremendously high value on sport. It is a major source of entertainment for both young and old. It has been reported that between 20 and 35 million 5–18 year old youth in the USA participate in one or more nonschool sports and another 10 million 14–18 year old youth participate in school sports yearly. Only family, school, and television involve children's time more than sport. In sum, sport is a cultural phenomenon that permeates all of US society.

Sport is not only a relevant topic for study because of its importance in society; it is a significant factor in the development of a child's self-esteem and identity. However, whether it is a positive factor has been debated widely without a clear answer. Danish et al. (1990) concluded, following a review of the literature, that for sport to be a positive developmental experience, it must be designed specifically for the purpose of enhancing competence.

2. *A Brief History of Sport Psychology*

Sport psychology dates back to the turn of the twentieth century. Although Norman Triplett, a psychologist from Indiana University, is credited with conducting the first study on athletic performance in 1898, Coleman Griffith is known as the father of sport psychology. A psychologist at the University of Illinois who also worked in athletics he developed the first sport psychology laboratory, was involved in the first US coaching school, and wrote two groundbreaking books, *The Psychology of Coaching* (Griffith 1926) and *The Psychology and Athletics* (Griffith 1928). He also conducted a series of studies with the Chicago Cubs baseball team and the University of Illinois as well as consulting with a number of well-known athletes of his time.

As sport psychology evolved over the twentieth century, two different kinds of sport psychologists emerged. *Clinical sport psychologists* are trained primarily in applied areas of psychology such as abnormal, clinical, counseling, and personality psychology and are usually licensed psychologists. They tend to be less well trained in the sport sciences. *Educational sport psychologists* are usually not licensed psychologists. Their training is in exercise and sport science, physical education, kinesiology, and the psychology of human movement, especially as it is related to the context of sport. They often have additional training in counseling. They see themselves either as researchers or as 'mental coaches.'

For some time there has been conflict between these two groups, based on who is best trained to provide services that enhance sport performance. The target of the services are primarily elite and professional athletes, many of whom see little need for such services or who sometimes seek the help of individuals not trained in either area. More recently, a rapprochement between these two groups has taken place with the formation of the Association for the Advancement of Applied Sport Psychology, an organization of approximately 1,000 members, almost equally divided

between members trained in traditional areas of psychology and members trained in the sport sciences.

3. *A Framework for the Development of Performance Excellence*

Gould and Damarjian (1998) have proposed a psychological pyramid of peak performance (see Fig. 1). They consider three sets of factors that interact to produce peak performance. At the base of the pyramid is the personality, motivational, and philosophical foundation of the individual. This foundation consists of those enduring individual differences that influence one's motivational disposition to learn the psychological skills that enhance performance. Among these factors are goal orientation, self-confidence, anxiety, and attentional style.

On one side of the pyramid are the psychological strategies used to enhance peak performance. Among

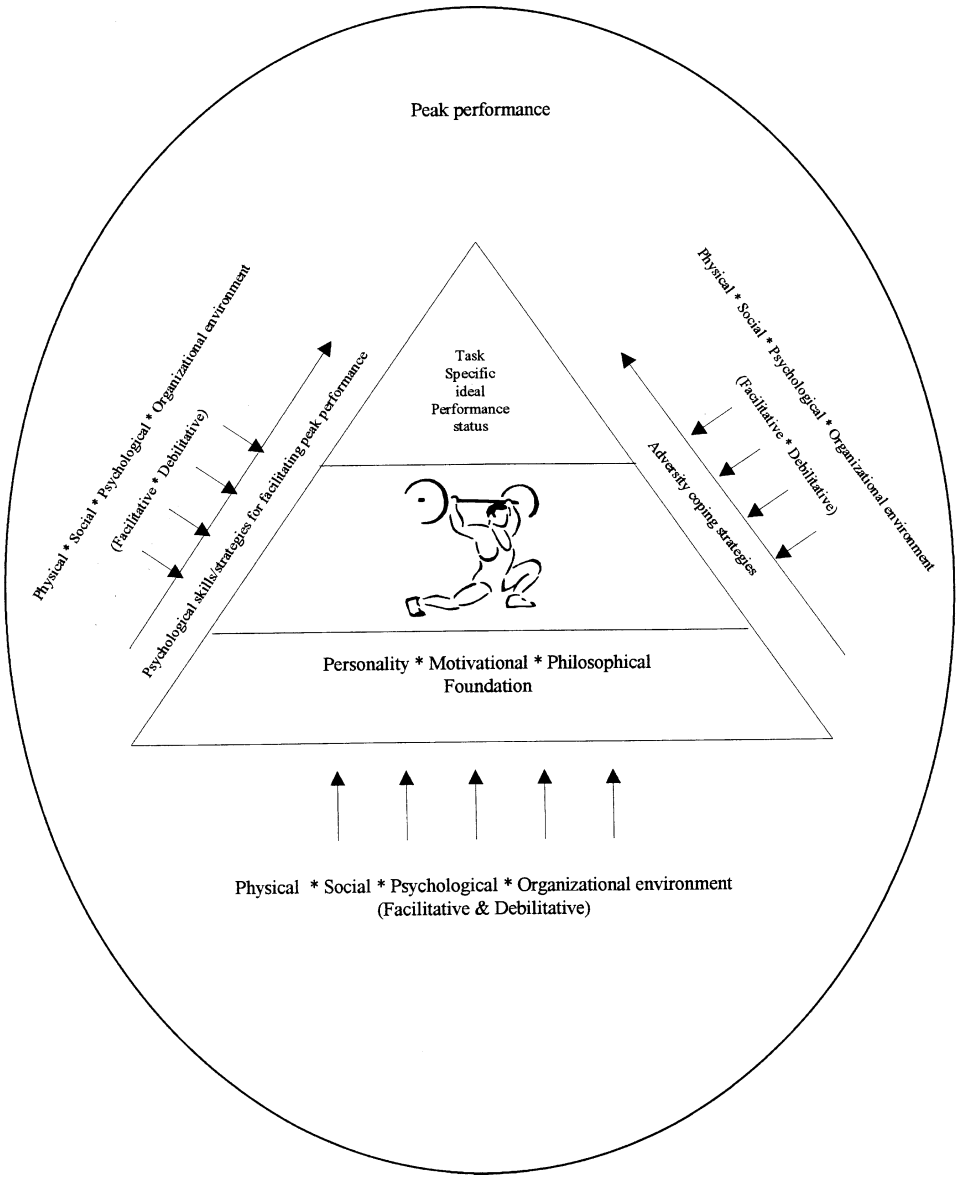

Figure 1
The Psychological Pyramid of Peak Performance (adapted from Gould and Damarjian 1998).

the skills taught are: goal setting, relaxation techniques, imagery and mental rehearsal, self-talk, the development of pre-performance and performance routines, and association/disassociation strategies. However, to achieve and maintain excellence, peak performers must learn how to deal with the adversity associated with excellence. Adversity coping strategies represent the other side of the pyramid. Stressors may come from an individual's own expectations, others' expectations, time demands, injuries, and general life concerns. As a result, learning effective coping strategies is essential. Although some of the strategies needed to cope may be the same as those used to enhance peak performance, the way they are used may be markedly different.

Athletes do not live, train, and compete in a vacuum. There are social, psychological, physical, and organizational factors in their environment that can have facilitative or debilitative effects on their performance. Gould et al. (1999) studied eight teams from the 1996 Atlanta Olympics. Four of these teams met or exceeded performance expectations and four failed to meet performance predictions. Through interviews and focus groups with the athletes and coaches, the authors found that teams that met or exceeded expectations participated in resident training programs, experienced family or friend support, employed mental preparation strategies, and were highly focused and committed. Teams that failed to meet expectations experienced planning and team cohesion problems, lacked experience, faced travel problems, encountered coaching problems, and had problems related to focus and commitment.

4. *Building the Foundation of a Peak Performer: Developing the Base*

B. S. Bloom (1985) and his colleagues followed 120 individuals who had achieved world-class success in sport (tennis and swimming), esthetics (music and the arts), and cognitive or intellectual (research mathematicians and research neurologists) endeavors. They found that the development of these individuals was divided into three phases: romance, precision, and integration. During the romance phase, development was multifaceted and characterized by play, exploration, and fun. It was the time when youth developed a love for learning and learned that hard work and fun were not incompatible. They learned the importance of having high expectations of themselves. The need to have high expectations was communicated through the expectations that parents, teachers, and the community in general had for the individual. Parents and teachers played a crucial role in the development of excellence. During this first phase motivation and effort counted more than any special gifts or qualities the child had. It was during this period that qualities such as character, emotional intelligence, and life skills were taught and acquired.

The romance phase was followed by a focus on precision. This was a period of systematic learning and skill development. The teachers/coaches of the individuals became more demanding and the children became more committed to the activity. Motivation became more internal than external. Skills were learned but not necessarily demonstrated very effectively.

It was not until the integration phase that individuals experienced the behavior as part of themselves. This phase was characterized by the perfection of their talent. The individuals became increasingly responsible for their own motivation. Much of their time was spent in their area of talent.

Bloom and his colleagues noted that by age 11 or 12 only about 10 percent of these individuals had progressed sufficiently to predict their later level of achievement. Among the special qualities that characterized these talented individuals were: a strong commitment to their particular talent field, a desire to reach a high level of attainment in their chosen field, an ability to learn quickly and well, and a willingness to invest great amounts of time and effort to reach high levels of achievement in their chosen field. For each field there were specific qualities that were important. Obviously, coordination, reflexes, endurance, and strength are needed for athletic excellence, but in contrast to what is believed by many, these conditions, while necessary, are clearly not sufficient.

5. *Teaching a Psychological Skills Training Program: Developing the Sides*

Athletes are active individuals. Their life experiences suggest they learn best by doing rather than by talking. They are accustomed to being 'coached' and expect concrete suggestions, feedback, and tasks. The process of teaching psychological skills should take advantage of their preferred learning style and expectations. As a result, although many of the specific strategies used to develop peak performance are based on traditional psychological approaches, they are best implemented using a skill-based, psychoeducational format (Danish et al. 1993). As such, these strategies may be quite similar to what Seligman and Csikszentmihalyi (2000) have referred to as 'positive psychology'.

A psychological or mental skills training program for developing all kinds of talent uses a psychoeducational approach as opposed to a traditional psychotherapeutic approach. The traditional psychotherapy model is problem- or illness-focused. The illness or problem is diagnosed, a treatment prescribed, therapy is undertaken and, if successful, a cure takes place. A psychoeducational model, on the other hand, begins with the individual's dissatisfaction or am-

bition. Goals are set and then the individual learns the skills to achieve the goal. The difference between the two approaches is both in the goals of the intervention and in the methods of delivery.

In terms of the goals, it is the difference between developmental and remedial help, between pathology and health promotion. Traditional psychotherapy focuses on identifying deficits and fixing them so the individual can return to 'psychological homeostasis'. The focus of psychoeducational approaches is on identifying assets and how to enhance them. Learning the skills serves as a catalyst for attaining personal and performance excellence.

One psychoeducational approach is life development intervention (LDI; Danish and D'Augelli 1983). The central strategy of LDI involves teaching individuals how to set a goal, develop a plan to reach the goal, and to overcome the roadblocks to goal attainment. Goals are actions undertaken to reach some desired end, not the end itself. Goals are different from results in that they are actions over which the participant has control. For example, for an athlete preparing for a competition, a goal might be to decide to learn how to develop a precompetition routine that will help focus on positive thoughts. To want to improve a time or score is not a goal, but a result. Individuals have control over their goals and only partial control over the results they want. In addition to having control over one's goals, three other elements are necessary for effective goal setting; (a) setting goals one is motivated to attain, (b) making sure that the goal is stated in positive terms, and (c) having a goal that is defined specifically and can be measured.

Being able to set a goal is one thing; being able to reach or attain the goal is another. There may be roadblocks that interfere with reaching a goal, regardless of whether it is of an athletic, career, or personal nature. These roadblocks are: a lack of knowledge, a lack of skill, being afraid to take risks in order to reach the goal, and/or a lack of adequate social support to work toward the goal. A lack of knowledge means that the individual is missing some information necessary for goal attainment. A lack of skills may refer to either physical or mental skills. A skill deficiency exists when the athlete does not know 'how to' attain a needed skill level. When individuals lack social support they need to know what support they need, who can provide them with the support (and who can not), and how to ask for the support. Support can either be of a tangible or emotional nature. The final roadblock relates to risk taking. Many individuals play and live their lives in a 'comfort zone' because they are afraid of failing, or possibly even succeeding. They may believe it is better not to try to extend themselves to reach a goal than to try and fail. Or they may feel that if they reach a goal, more will be expected of them on an ongoing basis. Risks are defined as the perceived benefits of an action minus the perceived costs. If the perceived costs outweigh the perceived benefits, efforts to reach a goal may be abandoned; if the perceived benefits outweigh the perceived costs, individuals may choose to work toward the goal. To optimize risk taking the benefits of an action must be increased and the costs decreased.

As part of the intervention, the roadblock(s) must be removed so that the individual may work toward the goal. Thus, teaching the necessary knowledge, skills, means of developing social support, and/or risk taking strategies to overcome the roadblocks to a goal is a major component of how the LDI approach is implemented.

The timing of when the intervention is implemented affects how it is delivered. For example, if goal setting were being taught, it could be taught to help an athlete prepare for the competition (enhancement), facilitate the athlete's ability to fine tune the goal during competition (support), or cope with an injury resulting from competition by helping develop a rehabilitation plan (counseling). Regardless of when the intervention is implemented, the building of rapport and the development of an alliance to work together toward the attainment of an individual's goals are essential.

6. *Developing Excellence in Sport and Life*

Not every individual can be a peak performer, but most can improve through performance enhancement strategies. Improving one's performance in sport or in other life domains begins with a decision to invest time and energy in deliberate practice. Bloom (1985) describes how a single-minded focus is an integral part of performance excellence. However, Marcia (1966) and Petitpas and Champagne (1988) point out that if individuals invest all their energies in sport, or in any single area, and achieve a level of success, they may decide to forego searching for a personal identity and commit to only that activity. By not exploring other areas, the individual may gain a sense of security at the expense of a search for identity. Baltes and Baltes (1980) refer to a single-minded focus as *selective optimization*. When individuals have a limited amount of time and energy, based on an assessment of environmental demands, their motivation, skills, and biological capacities, individuals select a pathway upon which to focus.

It should be pointed out that identity foreclosure and selective optimization are not in themselves harmful. These behaviors only become problematic when they prevent permanently the exploration of alternatives. It is through this process of exploration that individuals learn more about themselves. They receive feedback about their strengths and weaknesses by testing themselves in a variety of situations. They learn *life skills* (Danish 1995). For sport to serve as an effective model for learning life skills, the sport experience must be designed with this goal in mind.

Promoting personal and performance excellence simultaneously is not incompatible. However, it must be a planned outcome of sports participation. It occurs when athletes compete against themselves—more specifically, when that competition is focused on maximizing their own potential and achieving their own goals. It occurs when knowing oneself becomes as important as proving oneself.

De Coubertin ([1918] 1966), the founder of the Olympic movement, best described the dilemma faced by athletes who are so intent on proving themselves that they lack an awareness of what they have learned through sport:

> Sport plants in the body seeds of physio-psychological qualities such as coolness, confidence, decision, etc. These qualities may remain localized around the exercise which brought them into being; this often happens—it even happens most often. How many daredevil cyclists there are who once they leave their machines are hesitant at every crossroads of existence, how many swimmers who are brave in the water but frightened by the waves of human existence, how many fencers who cannot apply to life's battles the quick eye and nice timing which they show on the boards! The educator's task is to make the seed bear fruit throughout the organism, to transpose it from a particular circumstance to a whole array of circumstances, from a special category of activities to all the individual's actions (De Coubertin 1918).

Whether the purpose is to enhance sport performance or to improve other areas of one's life, the process is the same. What becomes essential is to ensure that what is learned can be generalized to other areas of the individual's life so that the game of life is a success.

7. *Concluding Comments*

The future of psychological interventions in sport lies in teaching mental skills to the masses, especially youth, and not just working with elite athletes. Sport participation among youth will likely continue to expand. However, it will not be taught in schools, where mandatory physical education and intramurals have all but disappeared. Instead, sport will be almost exclusively an after-school activity. As a result, the coaches will probably be youth workers and volunteers who may have little or no experience in teaching/coaching the different sports and will need special training.

Traditional programs will probably not be effective. The coaches at these youth-serving agencies tend to be underpaid and poorly trained, the volunteers are hard to recruit and retain, and the goals of these new coaches may not be the same as school-based coaches. Consequently, new programs must be developed that are modular, flexible, cost and time sensitive, multimedia in format, transportable, easily taught, and scalable.

There will be challenges and opportunities in developing the materials, training the new coaches, and expanding the sport experience to more youth. One of the biggest challenges is how to help youth develop both athletically and personally. It will require integrating the physical and mental aspects of the game. If we are successful, sport will become a vehicle for learning and applying some of life's most critical lessons.

See also: Sport and Exercise Psychology: Overview; Sport Psychology: Training and Supervision; Sports as Expertise, Psychology of; Sports Performance, Self-regulation of

Bibliography

Baltes P B, Baltes M M 1980 Plasticity and variability in psychological aging: Methodological and theoretical issues. In: Gurski G (ed.) *Determining the Effects of Aging on the Central Nervous System*. Shering, Berlin, Germany, pp. 41–60

Bloom B S (ed.) 1985 *Developing Talent in Young People*. Ballantine, New York

Danish S J 1995 Reflections on the status and future of community psychology. *Community Psychologist* **28**(3): 16–8

Danish S J, D'Augelli A R, Laquatin I 1983 *Helping Skills II: Life Development Intervention*. Human Sciences, New York

Danish S J, Nellen V C 1997 New roles for sport psychologists: Teaching life skills through sport to at-risk youth. *Quest* **49**(1): 100–13

Danish S J, Petitpas A J, Hale B D 1990 Sport as a context for developing competence. In: Gullotta T, Adams G, Monteymar R (eds.) *Developing Social Competency in Adolescence*. Sage, Newbury Park, CA, Vol. 3, pp. 169–94

Danish S J, Petitpas A J, Hale B D 1993 Life development intervention for athletes: Life skills through sports. *Counseling Psychologist* **21**(3): 352–85

De Coubertin P [1918] 1966 What we can now ask of sport. In: Dixon J [trans.] *The Olympic Idea: Discourses and Essays*. Carl Diem Institut, Köln, Germany

Gould D, Damarjian N 1998 Mental skills training in sport. In: Elliott B (ed.) *Training in Sport: Applying Sport Science*. Wiley, Chichester, UK, pp. 69–116

Gould D, Guinan D, Greenleaf C, Medbery R, Peterson K 1999 Factors affecting Olympic performance: Perceptions of athletes and coaches from more and less successful teams. *Sport Psychologist* **13**: 371–94

Griffith C R 1926 *Psychology of Coaching: A Study of Coaching Methods from the Point of View of Psychology*. Scribners, New York

Griffith C R 1928 *Psychology and Athletics, a general survey for athletes and coaches*. Scribners, New York

Marcia J E 1966 Development and validation of ego-identity status. *Journal of Personality and Social Psychology* **3**: 551–8

Petitpas A, Champagne D E 1988 Developmental programming for intercollegiate athletes. *Journal of College Student Development* **29**(5): 454–60

Seligman M E P, Csikszentmihalyi M 2000 Positive psychology – an introduction. *American Psychologist* **55**(1): 5–14

Weinberg R S, Gould D 1995 *Foundations of Sport and Exercise Psychology*. Human Kinetics, Champaign, IL

S. J. Danish

14928

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Sport Psychology: Training and Supervision

Sport psychology is a hybrid discipline more associated with physical education than with mainstream psychology. Debate over the training of sport psychologists is exemplified in the title of Brown's (1982) paper 'Are Sport Psychologists Really Psychologists?' Models of training and supervision will vary depending on how one answers that question. Sports psychologists and sport psychology organizations in different countries around the world have come up with different answers, and thus, different models of training and supervision have developed.

People pursue advanced degrees in sport psychology for reasons as varied as wanting to improve their coaching, preparing themselves for university academic positions, acquiring skills for clinical or counseling careers, and obtaining the training necessary to work in rehabilitation and sports medicine clinics (McCullagh and Noble 1996). Sport psychology as an academic discipline is almost always found in physical education/exercise science departments (other names include human movement studies and kinesiology) around the world (cf. Fujita 1987), but in many countries, those who practice come from psychology backgrounds. For example, in Flanders, virtually all of the members of the Flemish Society of Sport Psychology who practice have psychology degrees (Wylleman et al. 1999). In Australia, academic sport psychology is in human movement departments, but the current training of practitioners is in psychology departments, making graduates eligible for registration as psychologists (Morris 1995). This dual parenting of sport psychology has led to some lively debate about training.

The United States Olympic Committee (USOC 1983) stepped into the training debate when they began their Sport Psychology Registry (only those with doctoral degrees could apply). They divided their registry into three sections for three 'different' types of sport psychology service providers. 'Research sport psychologists' could apply for that category if they wished to conduct research with Olympic athletes on a variety of psychological topics. 'Educational sport psychologists' could deliver psychoeducational training (e.g., relaxation, goal setting, imagery, positive self-talk) to individuals and groups. 'Clinical sport psychologists' could deliver psychotherapeutic services for athletes with psychological problems (e.g., eating disorders, panic attacks).

This tripartite scheme was an attempt to accommodate individuals trained either in exercise science (many of the applications to the research and educational groups) or in psychology (most of the applications to the clinical group). The USOC was trying to be 'inclusive' of exercise science trained individuals who could not really use the protected title 'psychologist,' or be eligible for licensure, but who nevertheless often had extensive training in mental skills for performance enhancement. Also, one could apply for more than one designation. This classification scheme sparked some debate (e.g., Clarke 1984, Heyman 1984), and it also had a large influence on shaping how sport psychologists talked about themselves. Even today, many practitioners specifically call themselves 'educational sport psychologists' although there appears to be a trend away from using these, in many cases, somewhat restricting labels.

The question of training in sport psychology in the early to mid-1980s had become: training for what? Simplifying the USOC categories into to two broadly defined realms of research (academic posts and research center positions) and applied practice (performance enhancement, counseling, consultation, and clinical work) will help illustrate the basic paths for training that have evolved.

1. *Training*

The majority of people in academic research positions in sport psychology are still housed in exercise science departments. Sport psychology positions in psychology departments are relatively rare, and training for academic positions in sport psychology differ markedly around the world. In North America, a Ph.D. program would involve two to three years of extensive course work, involvement in a variety of research projects, and the completion of a substantial research dissertation, with many programs offering applied practica as part of the degree. In UK and other British Commonwealth nations (e.g., Australia, New Zealand), a Ph.D. in sport psychology is essentially a three-year long program of research (with little or no required course work), the product of which is often a series of three or four conceptually linked studies of a particular topic in the field. Thus, one system of training (e.g., the UK) has the doctoral student deeply immersed in the research process, and another system (e.g., the US) has extensive course work along with a major research project. Because the majority of academic research position training occurs in North America (nearly 100 universities offer graduate level programs), the majority of professional discussions on the training of sport psychologists has come from academics in those schools.

There is general agreement among sport psychologists that training should be interdisciplinary, covering both psychology and the exercise sciences (e.g., Dishman 1983). Feltz (1987), however, noted problems related to the integration of physical education/exercise science and psychology and suggested that expertise in psychology would be most appropriate to the applied aspects of the field. In the past decade many programs in North America, Australia,

and the UK have made opportunities for exercise science and psychology students to take classes in each others' departments. As sport psychology grows, it appears that training is becoming more and more interdisciplinary. Consensus on what training should be for sport psychologists, however, has not been reached. The field of sport psychology has no equivalent of the 'Boulder Model' in clinical psychology, but many programs describe themselves in their promotional material as espousing a scientist-practitioner model of training (see *Training in Clinical Psychology in the United States: Scientist–Practitioner Model*).

The organization that comes closest to setting training guidelines for those who wish to become practitioners is the Association for the Advancement of Applied Sport Psychology (AAASP) with their criteria for certification as a sport psychology consultant (see Sachs et al. 1998). Both the exercise sciences and scientific and applied psychology are well represented in the course work and training needed to become a certified consultant. The AAASP criteria also contain recommendations for supervised experience (at least 400 hours of supervised practicum working with athletes and coaches).

AAASP requires a doctoral degree for certification, but in many places around the world, a doctoral degree is not required to practice psychology. For example, in Australia sport psychologists are registered (equivalent of 'licensure' in the US) at the master degree level. Currently, Australia probably has the most integrated training and registration model of sport psychologists in the world. The Australian Psychological Society (APS) has several 'Colleges' (equivalent of Divisions in the American Psychological Association) within its structure, and one of them is the College of Sport Psychologists. The APS and the College of Sport Psychologists are the bodies that accredit master of applied psychology programs that emphasize working with sports people. Graduates of accredited programs are eligible for registration with State Psychologists Registration Boards. This integration of applied sport psychology training, program accreditation, and registration into mainstream applied psychology represents the latest level of professionalization in the field.

2. *Supervision*

The topic of supervision of sport psychology practica and internships, along with collegial supervision, arrived relatively recently on the professional practice of sport psychology debate scene. In 1993, the International Society of Sport Psychology published a short article on supervising trainees (Van Raalte and Andersen 1993), and supervision received a brief discussion in a book chapter on ethics (Sachs 1993). In 1994, however, two articles were published in the major journals of the field. The article by Andersen (1994) was a discussion piece on ethics and the supervision of applied sport psychology graduate students that offered recommendations for future training and supervision. The databased study on assessing the skills of sport psychology supervisors (Andersen et al. 1994) revealed the state of supervision in applied sport psychology training. Well over half the supervisors surveyed had never received any supervision of their own work with athletes. The median number of supervised practicum hours for students was just over 100. The opportunities for supervised experience in sport psychology graduate programs appeared seriously restricted, and the quality of the supervision could be questioned given the limited experience of many of the supervisors. Andersen et al. recommended that practitioners and trainers of future sport psychologists needed to pay more attention to supervision issues because it appeared that the training and supervision sport psychology graduates were receiving was, in many cases, inadequate.

In 1996, two more journal articles on supervision appeared. Andersen and Williams-Rice (1996) discussed, in a sport psychology service delivery context, the potential benefits of using psychodynamic, phenomenological, behavioral, cognitive-behavioral, and developmental models of supervision. They suggested that the cognitive-behavioral supervision model of Kurpius and Morran (1988) might be the most comfortable and appropriate for sport psychologists because it contains many features that sport psychologists use in practice (e.g., cognitive restructuring, mental rehearsal, cognitive self-instruction, covert modeling). Thus, a supervisor might have a supervisee mentally image and rehearse an up-coming encounter with an athlete, or supervisors might help supervisees overcome their own irrational beliefs about service delivery by using cognitive restructuring techniques. This supervision model has the potential to be the 'parallel process' model for sport psychology in that the interactions of the supervisor and the supervisee reflect similar processes that are going on with the supervisee and the athlete. Andersen and Williams-Rice also discussed transference and countertransference, along with peer and meta-supervision (supervision of supervision), misalliances, boundaries, and dual roles in supervision.

From the evidence in Andersen et al. (1994), there appeared to be very little training in how to be a sport psychology supervisor. Even in many clinical and counseling psychology programs, the main training one receives for becoming a supervisor is having been in supervision oneself. This state of affairs is similar to the wishful thinking that if one is well coached, one will become a good coach. Admittedly, current clinical and counseling programs are offering more training in supervision than they have in the past. In sport psychology education, however, training in how to supervise is quite rare.

Barney et al. (1996) presented a model for supervision training in sport psychology, borrowing and adapting Hart's (1982) three-stage developmental model of supervision. The Barney et al. model was actually a multitiered unified program of supervision and training. The model begins with background education in the sport sciences, physiology, physical education, and psychology. The training begins with a year-long course in sport psychology theory and research. Towards the end of that year students would begin their practica counseling athletes. Those beginning practicum students would be supervised by advanced students undergoing a 'supervision practicum' where they are being trained to become good supervisors. Barney et al. described three major phases for the supervisor-in-training that are interwoven. Working on the acquisition of supervisory skills (e.g., how to do a supervision intake), especially for a beginning supervisor, may take a considerable amount of time. Supervision may swing back and forth between skill acquisition and personal growth issues. The meta-supervisor helps supervisors-in-training to explore what they bring to the supervisor–supervisee relationship, their needs, their countertransferential thoughts and behaviors, and so forth. This self-exploration and personal growth are in the service of producing a more aware sport psychologist who will competently deliver service. The integration phase for the supervisor-in-training is combining new skills and personal growth for building dynamic helping relationships with athletes. Barney et al. placed great emphasis on the study of relationships in sport psychology (e.g., meta-supervisor–supervisor-in-training, supervisor–supervisee, psychologist–athlete) and how relationships are the key to successful sport psychology service delivery. This emphasis on relationships is something new in sport psychology, and this trend is exemplified in another recent article on training. Petitpas et al. (1999) have presented arguments and recommendations for making the examination of relationships one of the central features in the training of future applied sport psychologists. The recognition of the importance of relationships is nothing new in mainstream counseling and clinical psychology, but because sport psychology's original 'home' was in physical education, such counseling and clinical considerations did not arise until recently. The new emphasis on relationships represents the continuing 'psychologization' of sport psychology.

The debate on training in sport psychology has been with the field since the early 1980s, but the discussion of supervision is a relative newcomer. Supervision and training are currently becoming widely debated topics. For example, at the 1998 AAASP annual conference, there were over 20 presentations that were concerned with training and/or supervison. The study of sport psychology, even though it roots go back to Triplett (1898) and Griffith (1928), is a young, but maturing field. As applied sport psychology matures, interest in quality training and supervision will probably grow even more.

See also: Developmental Sport Psychology; Sport and Exercise Psychology: Overview; Sport Psychology: Performance Enhancement; Sports Performance, Age Differences in; Teamwork and Team Training; Youth Sports, Psychology of

Bibliography

Andersen M B 1994 Ethical considerations in the supervision of applied sport psychology graduate students. *Journal of Applied Sport Psychology* **6**: 152–67

Andersen M B, Van Raalte J L, Brewer B W 1994 Assessing the skills of sport psychology supervisors. *The Sport Psychologist* **8**: 238–47

Andersen M B, Williams-Rice B T 1996 Supervision in the education and training of sport psychology service providers. *The Sport Psychologist* **10**: 278–90

Barney S T, Andersen M B, Riggs C A 1996 Supervision in sport psychology: Some recommendations for practicum training. *Journal of Applied Sport Psychology* **8**: 200–17

Brown J M 1982 Are sport psychologists really psychologists? *Journal of Sport Psychology* **4**: 13–17

Clarke K S 1984 The USOC Sports Psychology Registry: A clarification. *Journal of Sport Psychology* **6**: 365–6

Dishman R K 1983 Identity crisis in North American sport psychology. *Journal of Sport Psychology* **5**: 123–34

Feltz D 1987 The future of graduate education in sport and exercise science: A sport psychology perspective. *Quest* **39**: 217–23

Fujita A H 1987 The development of sport psychology in Japan. *The Sport Psychologist* **1**: 69–73

Griffith C R 1928 *The Psychology of Athletics*. Scribner, New York

Hart G M 1982 *The Process of Clinical Supervision*. University Park Press, Baltimore, MD

Heyman S R 1984 The development of models for sport psychology: Examining the USOC guidelines. *Journal of Sport Psychology* **6**: 125–32

Kurpius D J, Morran D K 1988 Cognitive-behavioral techniques and interventions for application in counselor supervision. *Counselor Education and Supervision* **27**: 368–76

McCullagh P, Noble J M 1996 Education and training in sport and exercise psychology. In: Van Raalte J L, Brewer B W (eds.) *Exploring Sport and Exercise Psychology*, 1st edn. American Psychological Association, Washington, DC, pp. 377–94

Morris T 1995 Sport psychology in Australia: A profession established. *Australian Psychologist* **30**: 128–35

Petitpas A J, Giges B, Danish S J 1999 The sport psychologist-athlete relationship: Implications for training. *The Sport Psychologist* **13**: 344–57

Sachs M L 1993 Professional ethics in sport psychology. In: Singer R N, Murphey M, Tennant L K (eds.) *Handbook of Research on Sport Psychology*. MacMillan, New York, pp. 921–32

Sachs M L, Burke K L, Gomer S 1998 *The Directory of Graduate Programs in Applied Sport Psychology*, 5th edn. Fitness Information Technology, Morgantown, WV

Triplett N 1898 The dynamogenic factors in pacemaking and competition. *American Journal of Psychology* **9**: 507–33
— 1983 US Olympic committee establishes guidelines for sport psychology services. *Journal of Sport Psychology* **5**: 4–7
Van Raalte J L, Andersen M B 1993 Special problems in sport psychology: Supervising the trainee. In: Serpa S, Alves J, Ferreira V, Paulo-Brito A (eds.) *Proceedings: VIII World Congress of Sport Psychology*. International Society of Sport Psychology, Lisbon, pp. 773–6
Wylleman P, De Knop P, Delhoux J, Vanden Auweele Y 1999 Current status and future issues of sport psychology consultation in Flanders. *The Sport Psychologist* **13**: 99–106

M. B. Andersen

Sport, Sociology of

Sport, once considered a marginal area of interest (a subset of the sociology of leisure or sociology of culture at best) is now a distinct area of concern for sociologists. Since the 1960s, a significant body of sociological body of research has been produced constituting the area of sport sociology. The growing interest in sport as a site of research is the result of three factors: the increasing pervasiveness of competitive sport in everyday life, the growing acceptance of critical examination of sporting practices, and sport's distinctive and relatively autonomous structure. Sport sociologists are particularly interested in modern sport, a relatively new cultural form dating from the late 1800s. Modern sport shares attributes with leisure/recreation/play and with drama, but is distinct from both. It distinguishes itself from other forms of leisure/recreation/play in that it is physical, competitive and institutionalized. It can be distinguished from other forms of entertainment (e.g., drama, music) in that it is not in and of itself expressive (Gadamer 1975). The movement of a sport contest is governed by its rules. Athletes, if they are giving an 'honest effort,' are trying to win a game, not to communicate an idea or emotion. Thus, much of the social meaning of sport comes from outside of the actual playing of the game.

The relationship between spectators and sporting practices also distinguishes sport from other forms of entertainment. Sport spectators see themselves giving support to athletes. Where other forms of entertainment attempt to evoke an emotional response in the audience, the sport audience attempts to evoke an emotional response in athletes in an effort to impact the outcome of a contest. The fan/athlete bond, then, is rooted in the norms and obligations of reciprocity (Mauss 1954) that forms the foundation of the intense, sometimes fanatical, relationship between fans and sport.

Despite the rule-bound and relatively autonomous space of sport, the structure of a sport (i.e., its rules, amount of physical contact, who is allowed to compete, who governs the sport, etc.) and the meanings associated with a sport are a product of a particular culture. The social significance of a sport, argues Bourdieu (1988), cannot be apprehended in the absence of a much broader understanding of the structure of the society as a whole. For many sociologists, sport is a site through which questions related to social and political power, domination, and ideology are considered. Since the 1960s, sport sociology can be viewed as having two approaches: an institutional approach that emphasizes the unique and semi isolated nature of sporting practices and a cultural orientation which views sport as a cultural product or practice reflective of society as a whole.

1. *Sport as a Unique Institution*

The institutional approach to the study of sport has its roots in physical education. Physical education researchers in the 1960s and 1970s applied their empirical methods to sport (Kenyon and Loy 1965). For the most part, early researchers were looking for empirical proof for popular beliefs about sport's positive and character building attributes or as Coakley put it later 'socialization-as-internalization' research (1993). The results, by and large, were inconclusive (Fine 1987). At about the same time, sport and physical education became embroiled in the social movements of the 1960s and 1970s. In the United States, scholar and political activist Jack Scott published his manifesto, *Athletic Revolution* (1971) and Harry Edwards documented the protests and threatened boycott he led in 1968 in *Black Athlete Revolt* (1969). In France, Jean-Mari Brohm (1978), influenced by the student movement, argued that sport was alienated play and encouraged an anti-Olympic movement.

Beginning in the 1970s, empirical research on sport as an institution also took a critical turn, documenting structural inequalities within sport and challenging popular ideologies. One line of research following from these studies focuses on positional racial segregation on the field of play; a phenomenon dubbed 'stacking' (Curtis and Loy 1978). A related field of inquiry is the documentation of racial and gender inequalities in the hiring practices of teams and leagues. Accosta and Carpenter (1994), for example, document gender inequalities in hiring practices, salaries, and advancement of women in intercollegiate sport in the United States. More recently, sociologists have focused on disparities in media coverage. Socialization research also took a critical turn as sociologists

attempted to demonstrate a link between athletic participation and sexual assaults and other criminal or deviant behavior.

Another line of critical inquiry in sport has roots in a social problems perspective. The social problems perspective in sport tends to focus on contradictions within sport. Researchers frame sport as paradoxical, (i.e., sport promotes health and injury, both encourages and discourages drug use) or highlight positive deviance or over-conformity (i.e., a strong emphasis on winning encourages drug use or injury). Other research frames the problems of sport in the context of social-historical shifts. Debates about performance enhancing drug use, for example, are framed in light of changes in approaches to medical treatment. Commonly accepted medical interventions, such as cold remedies, therapeutic treatments for slowing down the aging process, or preventing depression, are marketed with the express intent to enhance performance in everyday life. As the standard of acceptable medicine changes, and the idea of what is a treatable ailment shifts, so will athletes' sense of fairness and prohibitions by sports organizations.

Independent of the burgeoning field of sport sociology, qualitative researchers in the 1960s, influenced by the Chicago School, began producing ethnographies of marginalized groups within sports world (Scott 1968, Polsky 1969). Ethnographers would find the relatively autonomous world of sport fertile ground for understanding common, everyday experiences. For example, Chambliss (1989) theorizes excellence through elite swimming, Adler and Adler (1991) detail the experience and consequences of role engulfment through college basketball, and Fine (1987) explores preadolescent male culture in his study of youth baseball.

2. *Cultural Orientation to Sport*

In the 1970s, Gruneau and other scholars associated with the Massachusetts School attempted to re-center the sociological study of sport within an overriding concern for political economy (Ingham and Donnelly 1997). The works of these scholars formed the basis of the cultural orientation in the United States and Canada. They found theoretical grounding in Gramsci's notion of hegemony. As a social practice, sports are produced by athletes and organizers, but within the context of historically-specific social constraints. This theoretical formulation provides the basis for Gruneau's analysis of class relations, reproductions and transformation of Canadian sport (1983). By extension, modern sport is viewed as a celebration of modern values of competition and winning and organized around principles of the division of labor, capitalism, and professionalism. Because sport offers so little social significance in and of itself (e.g., it has a veneer of neutrality), it becomes a powerful resource for constructing meaning—particularly for the benefit of the power elite. Central to the Massachusetts School is the Gramscian idea of counter hegemony or resistance. Sport, they note, could also be used by marginalized groups to challenge dominant ideologies and structures. From this perspective, sport is a site in which meaning is constructed, challenged, and reconstructed. Put slightly differently, sport is viewed as 'contested terrain.'

At about the same time in Europe, Eric Dunning, a student of Norbert Elias cultivated a somewhat different cultural orientation. Dunning incorporated Elias' grand theory of social development into a figurational approach to sport sociology (Elias and Dunning 1986, Dunning 1999). The figurational approach which focuses above all on social processes and interdependencies frame a vast range of works from hooliganism, athlete migrations and male bonding. The figurational approach distinguishes itself from the Massachusetts School in its focus on a total social structure—taking into account a wide variety of social processes that hold groups together such as emotions, gender, race, and communication patterns.

In the 1980s, feminist sport theorists incorporated aspects of the figurational approach and the Massachusetts School emphasis on hegemony and resistance and began to theorize about the changes in sport due, in large part, to the feminist movement. The influence of Dunning and Sheard's (1973) essay on rugby as a male preserve on the Massachusetts School is evident in the works of Theberge (1985) and Birrell (Birrell and Richter 1987). Arguing that sport is a male preserve that contributes to the oppression of women physically and sexually, they assert that sport nonetheless offers women the potential for subverting male dominance by enabling women to experience their bodies as strong and powerful or to restructure sport to fit feminist principles. Messner (1988), while discounting the simplistic notion that increasing participation by women signals ideological shifts, argues that images of athletic women are forged, interpreted, and contested in sport and may partially contradict masculine hegemony.

Feminist insights of contested terrain in sport were applied to a variety of concerns, including bodybuilding and masculinity (Klein 1993) and inequalities and Texas High school football (Foley 1990). These sport sociologists (often called critical sport sociologists) see sport as a site through which questions related to social and political power, ideology, and transformative possibilities are considered. The bulk of the research that follows from this conceptualization of sport is far less optimistic than the theory that inspired it. Bryson (1987), for example, found masculine hegemony is both resilient and more complex than had been previously theorized. Bryson documents the political tactics of men in Australia in response to women's athletic prowess that reframe their accomplishments to support masculine he-

gemony. In many ways, the more recent trends in sport follow from the discovery of sport's rather limited transformative potential.

3. Recent Trends, Future Challenges

Foucault, cultural studies, and postmodern theory have, to varying degrees, influenced critical sport sociology. One of the great difficulties for the sociology of sport is how to understand the physicality of sport. Most classical sociological theory ignores or diminishes the significance of body. Yet the empowerment and transformation of the body or body skills is at the heart of sport's transformative possibility for critical sport sociologists. Foucault and feminist theory enabled sport scholars to problemitize the body—to see it as a site of domination and through which power is exercised and enjoyed. Bodies are not simply empowered through sport, but they are constrained, shaped, and consumed and thus a central ideological resource (Cole 1994). Cultural studies encourages sport sociologists to make a critical analysis of the social meanings of sporting events. A growing literature offers critical readings of the discourse surrounding a sporting event or celebrity, both to expose the multiple meanings of sport and to analyze how these narratives do ideological work which reproduces and justifies race, class, and gender inequalities (Birrell and McDonald 2000). For example, images of black athletes, even those purporting to offer positive images of African–Americans in order to subvert racism, refer to and reinforce racial stereotypes associating African–Americans with crime, drugs, and poverty.

The postmodern perspectives suggest that we see sport as part of broad and fundamental social changes. One of the distinguishing features of contemporary post-industrial societies is the use of images and spectacles to enculturate individuals into a media and consumer culture. Social values, political debates, and everyday concerns are not garnered through active participation, but through the passive consumption of spectacles. Post-industrial media culture implodes sport into media spectacle. Celebrity athletes become images that are dynamic and complex—but, because they are disconnected from social life, they can be interpreted in a variety of ways. Postmodern perspectives encourage a shift in sociological concerns away from the active producers of sport and culture to an exploration of the passive consumption of sport spectacles (Kellner 1996).

Somewhat in response to the postmodern turn, another line of inquiry is the re-examination of the institutions of sport and their structural intersections with political power. These examinations of the organizational practices within the sport industry attempt to ground the discussion of power and domination in and through sport in a structural framework. The narratives surrounding sport become ancillary to the exercise of power—particularly global political and economic power—and how this shapes the institutions of sport.

One of the future challenges of sport sociology will be to find a synthesis between institutional and the cultural critiques. While the concerns of both approaches appear similar at a very basic level they are very different. The cultural critique relies essentially on the success and celebrity of people of color, women and other marginalized groups to demonstrate the resiliency of ideological domination while the institutional critique emphasizes barriers and limitations impeding the progress of these very same groups. Nonetheless the synthesis of these two perspectives seems to hold the promise for greater sociological understanding of sport.

See also: Critical Theory: Contemporary; Elias, Norbert (1897–1990); Entertainment; Hegemony: Cultural; Regulation: Empirical Analysis; Sport, Anthropology of; Sport, Geography of; Sports, Economics of

Bibliography

Acosta R V, Carpenter L J 1994 The status of women in intercollegiate athletics. In: Birrell S, Cole C L (eds.) *Women, Sport, and Culture*. Human Kinetics, Champaign, IL, pp. 111–18

Adler P A, Adler P 1991 *Backboards and Blackboards: College Athletics and Role Engulfment*. Columbia University Press, New York

Birrell S, McDonald M G (eds.) 2000 *Reading Sport: Critical Essays on Power and Representation*. Northeastern University Press, Boston

Birrell S, Richter D M 1987 Is a diamond forever? Feminist transformations of sport. *Women's Studies International Forum* **10**(4): pp. 395–409

Bourdieu P 1988 Program for sport sociology. *Sociology of Sport Journal* **5**(2): 153–61

Brohm J-M 1978 *Sport—A Prison of Measured Time*. Ink Links, London

Bryson L 1987 Sport and the maintenance of masculine hegemony. *Women's Studies International Forum* **6**(2): 135–53

Chambliss D 1989 The mundanity of excellence. *Sociological Theory* **7**(1): 70–86

Coakley J 1993 Sport and socialization. *Exercise and Sport Sciences Reviews* **21**: 169–200

Cole C L 1994 Resisting the canon: Feminist cultural studies, sport and technologies of the body. In: Birrell S, Cole C L (eds.) *Women, Sport and Culture*. Human Kinetics, Champaign, pp. 5–29

Curtis J, Loy J 1978 Positional segregation in professional baseball: Replications, trends, data and critical observations. *International Review of Sport Sociology* **4**: 5–21

Dunning E 1999 *Sport Matters: Sociological Studies of Sport, Violence and Civilization*. Routledge, London

Dunning E, Sheard K 1973 The rugby football club as a type of male preserve. *International Review of Sport Sociology* **8**: 5–24

Edwards H 1969 *The Revolt of the Black Athlete*. Free Press, New York

Elias N, Dunning E 1986 *Quest for Excitement*. Basil Blackwell, Oxford

Fine G A 1987 *With the Boys: Little League Baseball and Preadolescent Culture*. University of Chicago Press, Chicago
Foley D 1990 The great American football ritual: Reproducing race class and gender inequality. *Sociology of Sport Journal* **7**(2): 111–35
Gadamer H G 1975 *Truth and Method*. Sheed and Ward, London
Gruneau R 1983 *Class, Sport and Social Development*. University of Massachusetts, Amherst, MA
Ingham A G, Donnelly P 1997 A sociology of North American sociology of sport. *Sociology of Sport Journal* **14**(4): 362–418
Kellner D 1996 Sport media culture and race—Some reflections on Michael Jordon. *Sociology of Sport Journal* **13**(4): 458–67
Kenyon G S, Loy L W 1965 Toward a sociology of sport: A plea for the study of physical activity as a sociological and social psychological phenomenon. *Journal of Health, Physical Education and Recreation* **38**: 24, 68–69
Klein A 1993 *Little Big Men: Bodybuilding Subculture and Gender Construction*. State University of New York Press, Albany, NY
Mauss M 1954 *The Gift: Forms and Reason for Exchange in Archaic Societies*. W W Norton, New York
Messner M 1988 Sports and male domination: The female athlete as contested ideological terrain. *Sociology of Sport Journal* **5**(3): 197–211
Polsky N 1969 *Hustlers, Beats and Others*. Anchor Doubleday, New York
Scott J 1971 *The Athletic Revolution*. Free Press, New York
Scott M 1968 *The Racing Game*. Aldine, Chicago
Theberge N 1985 Toward a feminist alternative to sport as a male preserve. *Quest* **37**(2): 193–202.

T. W. Crosset

Sports as Expertise, Psychology of

Elite athletes represent a small and specific subset of the human population within whom the mastery of movement (motor) skill acquisition is at its most pronounced. The emerging field of sport expertise concerns itself with the detailed study of exceptional sports performers and the mechanisms underpinning the overt power, strength, grace, beauty, speed, accuracy, and apparent ease, efficiency, and effectiveness of their movement skills. Comparisons with less skilled individuals, along with detailed examinations of the developmental life histories of exceptional athletes, are used in an attempt to develop a principled understanding of the nature and limits of human perceptual–motor learning.

1. Theoretical and Applied Significance of Sport Expertise Research

Studying expert performance in sport has the potential to contribute both important theoretical and applied knowledge to the social and behavioral sciences. In a theoretical context, studying sport expertise is important because of its potential to provide valuable and unique insights into the nature of perceptual–motor skill learning. The attainment of expert performance in a whole range of domains (sport being no exception) appears to take at least 10 years, or approximately 10,000 hours, of daily, effortful practice, during which time literally millions of trials of practice are undertaken (Ericsson et al. 1993). Studying the expert therefore provides a window into learning, and skill optimization that conventional, laboratory-based learning studies of untrained participants simply cannot replicate. Studies of sports experts in particular are uniquely valuable in permitting broad assessment of the explanatory efficacy of theories and models of human skill acquisition and expertise (typically developed in domains other than sport) and, in conjunction with information from the field of kinesiology or human movement science, in helping isolate the limiting factors to performance in a range of different sport tasks.

In an applied context, sport expertise research offers promise as a means of providing information of practical significance to coaches and athletes regarding the selection of the most appropriate forms of practice and training, the assessment of skill progression, and the optimal design of sports systems and talent identification. For example, knowing what factors do, and equally importantly do not, discriminate expert and novice performers can provide a principled basis for guiding emphasis within practice. Knowing the specific information that must be learned in order to become an expert on a given task has direct implications for the structure of instruction and the provision of feedback. Knowing the early developmental experiences that adult experts in particular sports share in common can provide direct guidance with respect to the appropriate emphases to apply to the design of junior sport systems.

2. Historical Trends in the Study of Sport Expertise

While a strong case can be made for the potential importance of sport expertise research, historically it has not been a favored topic of study in either sport psychology or 'mainstream' cognitive psychology. Although a considerable number of studies comparing skilled athletes with less skilled people on a range of isolated 'abilities,' such as vision, reaction time, balance, movement speed, and agility, were undertaken throughout the 1930s to 1970s, this research work was, almost without exception, descriptive, atheoretical, and nonprogrammatic. Consequently, little progress was made during this period toward the development of a consolidated body of knowledge, nor toward the development of conceptual models or theories of sport expertise. In large part this was due to the concerted emphasis in the motor learning field at this same time

on research examining the limitations and capacities of putative information-processing stages—such work proceeded typically by studying untrained participants as they performed simple, but novel, laboratory tasks.

The foundations for a consolidated body of research on sport expertise were laid largely in the 1980s. During this period paradigms developed to study expertise in cognitive tasks (such as playing chess or solving mathematical problems) were applied progressively to the study of sport situations. Particularly prominent were the pattern-recognition paradigm developed by de Groot (1966), and Chase and Simon (1973), and the knowledge-based approaches to skill acquisition of Chi (1981), Anderson (1982) and others.

Studies using the pattern-recognition paradigm typically demonstrate superior pattern recognition and recall by experts for material that is domain-specific; however, this advantage disappears for unstructured material and for patterns derived from beyond the expert's domain. This observation holds true for sport tasks as well as more traditional cognitive tasks, although there is some controversy as to whether the expert's advantage on such tasks is a cause or merely a by-product of their expertise. The knowledge-based approaches have revealed systematically superior knowledge content and organizational structure for experts, again regardless of whether the experts are drawn from the cognitive or the motor domain (Thomas and Thomas 1994).

Studies derived from the pattern-recognition and knowledge-based approaches were invaluable in providing the impetus for the development of the sport expertise field, although on occasions the assumptions, strengths, and limitations in the use of these imported paradigms was probably not fully appreciated (Abernethy et al. 1993, 1994). For example, the verbal self-report method of data collection used within most of the knowledge-based approaches to cognitive expertise may have reduced veracity when used with sports experts. In sport settings, direct, conscious access to the mechanisms controlling highly automated, rapidly performed motor acts may simply not be possible, and any data derived from verbal self-report may need to be treated with caution.

The scholarly study of sports expertise has gained considerable momentum since the beginning of the 1980s. A number of factors have contributed to this. These include the growing interest in expert systems within cognitive science, the development of improved technology for the *in situ* experimental study of sport tasks, and a rejuvenated interest in learning as a pivotal issue within the broader motor-control field. Two additional factors are particularly noteworthy. The first is the emergence of a theoretical perspective (ecological psychology) which challenges many of the fundamental premises of cognitive psychology. Ecological psychology is influenced heavily by the work of James Gibson on perception (e.g., Gibson 1979), and by Nicholas Bernstein on action (e.g., Bernstein 1967). Its orientation toward functional couplings between perception and action places great emphasis on the study of natural actions as a central vehicle toward understanding skill learning, and such an orientation is compatible with a focus on sport tasks, and experts within such tasks (Abernethy et al. 1994). The second noteworthy factor is the renewed interest in the developmental bases and social support mechanisms for expert performance—a topic first presented in the work of Benjamin Bloom (Bloom 1985). The effective fusion of qualitative, social approaches to expertise development of the type developed by Bloom with more traditional, quantitative, empirical studies of experts represents one of the key challenges for sport expertise in the modern era.

3. *Current Emphases in Sport Expertise Research*

Many of the distinguishing characteristics of expert athletes surmised by both early skill-acquisition theorists and lay commentators have now been confirmed empirically (Abernethy et al. 1998). In comparison to less skilled individuals, exceptional athletes are demonstrably advantaged on a range of perceptual, cognitive, and motor parameters. Perceptually, they are faster and more accurate in recognizing patterns from their sport, superior in terms of both their speed and accuracy in anticipating the actions of an opponent, and better able to pick up essential biomechanical information from observation of the movement patterns of others.

Cognitively, they have superior knowledge of both factual (declarative) and procedural matters related to their sport, their knowledge is organized in a deeper, more richly structured form, they possess superior knowledge of situational probabilities for their sport, and they are better able to plan their own actions in advance. In terms of movement *per se*, expert athletes produce movements with greater consistency and adaptability, greater efficiency and effectiveness, and greater speed and accuracy than can less skilled performers. In addition, expert athletes have also been shown to be able to produce movement more automatically, with reduced attention demand and yet with superior ongoing monitoring, than can less skilled individuals.

An outstanding feature of expert performance is therefore the capacity to satisfy simultaneously competing task demands (such as speed versus accuracy, consistency versus adaptability, or attention versus monitoring) in a manner that optimizes task performance. Across each of the elements of expert sports performance, the challenge for researchers is to move beyond the existing descriptive knowledge about expertise toward the more fundamental goals of explaining and understanding how this expert ad-

vantage develops and is achieved. This challenge is being taken up at the time of writing in at least three major areas of focus for contemporary sports expertise research.

One major current research focus is the determination of the limiting factors to sports performance. While much is now known about factors that differentiate reliably the expert athlete from everyone else, an equally important body of knowledge is that accrued on factors which do *not* appear to be important discriminators. The accumulating evidence showing the inability of generalized perceptual, cognitive, and motor tests (such as simple visual tests, and generic measures of reaction time) to differentiate experts from others suggest that the attributes being measured by these tests are not the limiting factors to sports performance (Abernethy et al. 1998). Approaches which seek to improve sports skill by training these attributes are unlikely to be effective. The evidence on other aspects, such as whether or not experts and novices differ systematically in their sport-specific eye movement patterns, is less clear-cut, leaving issues as to the efficacy of training learners to model the scan patterns of experts unresolved at this time (Williams et al. 1998).

A second major current research focus is on the determination of possible differences in information usage and dependence between experts and novices. Determining the specific information (or cues) used by experts and novices in sport tasks is a challenging undertaking. To date, the main approaches that have been used involve selective temporal and/or spatial masking of potential information sources, manipulation of kinematic display information using computer digitization, and the recording of eye movement (or visual search) patterns. Such approaches indicate some systematic differences in the specific cues used by the experts and less skilled people—experts, in time-constrained sports at least, being able to pick up information from early-occurring cues that novices cannot utilize. More precise determination of the specific information sources guiding the selection and production of movement by experts in particular sport settings may, in the future, provide a more principled basis for training, and instructional and feedback augmentation than is currently possible.

A third major contemporary research focus is on the critical question of how learning to be an expert takes place. The approaches that describe sports expertise in terms of component skills and limiting factors, and in terms of differences in information usage, provide a valuable starting point for knowing the 'end product' differences between experts and novices; but they are silent with respect to the types and amount of practice, the critical role models, developmental experiences and social support structures that are necessary (if not sufficient) for expertise to be achieved. A perspective, advanced by Anders Ericsson and others (e.g. Ericsson et al. 1993), is that the sheer quantum of intense goal-directed (*deliberate*) practice is the single best predictor of the attainment of expertise. This perspective is stimulating significant research activity as others attempt to verify the validity and generality of this proposition across sports and attempt to determine the role played by the type of practice and by other contextual factors (such as parental support levels and relative age in first entering organized sport) in the development of expertise (e.g., see Singer and Janelle 1999, Starkes et al. 1996). An emerging view, though as yet supported by only a modest evidence base, is that relatively early exposure to competition with adults, opportunities to model and learn by watching from an early age, and the existence of broad diversity in practice conditions in the formative years, especially through settings that encourage creativity and experimentation without excess emphasis on outcomes, may all be important environmental ingredients for the nurturing of expert performers.

4. *Methodological Issues and Challenges*

Studying sports experts and understanding the basis of sports expertise poses some unique, and difficult, methodological challenges (Abernethy et al. 1993). Exceptional performers in any particular sport are, by definition, relatively few in number. A consequence of this is that many/most studies of sports expertise suffer from either limited statistical power (because of the difficulty in assembling a large sample size of genuine experts), or a dilution of the degree of 'expertise' within the expert group. With considerable variability evident in the stringency with which experts are defined, it is not an uncommon observation that the expert sample in one study may well qualify as the less skilled control sample in another. This absence of agreed, universal criteria or even convention as to what minimally constitutes an expert clearly works against the development of a consolidated knowledge base on sports expertise.

A further, vexing methodological problem relates to the dissociation of expertise effects from effects due not to expertise *per se* but rather due simply to experience. In the majority of the existing studies on sport expertise, key characteristics of the expert group are compared directly with those of a control group, usually consisting of task novices. As a consequence, the groups being compared differ not only in terms of their level of task expertise but also in their level of task experience. It is frequently difficult, therefore, to ascertain whether any observed differences are related genuinely to expertise or are simply a product of different levels of exposure to the task. Using additional groups differing in performance level but matched in terms of years of playing experience can go some way toward addressing this issue, although even with such controls the quality and quantity of experience are so tightly interrelated with skill level

that the confounding effect of experience cannot be completely disentangled from the issues of expertise.

Sport expertise is clearly not unidimensional, yet many of the prevailing methodological approaches used in an attempt to understand sport expertise have been of this type. As sport tasks provide the potential for minor deficiencies in some skill components to be offset by strengths in others, a raft of dependent measures tapping different facets of performance rather than unitary measures is obviously needed to capture fully the essence and individual variability that goes toward the making of an expert. More consistent use of multiple dependent measures is to be expected in future research endeavors in this field.

Unlike many other fields of study, where the key variable(s) of interest can be mapped effectively over relatively short time periods, the cumulative and uncertain nature of the development of expertise makes prospective studies—desirable as they may be—logistically very difficult. The accumulation of data on the lifespan development of experts therefore relies heavily on retrospective self-reports by athletes, yet such methods need considerable confirmation from other sources to avoid potential biases and inaccuracies. With the growing emphasis on mapping of the individual developmental histories of athletes, a key challenge for the future is to continue to modify and improve the reliability and validity of the methods used for acquiring and comparing lifespan data.

5. *Future Directions*

The major theoretical challenge for the sport expertise field is to understand, in a much more coherent way than is the case at the time of writing, the mechanisms responsible for the observable expert–novice differences in overall performance and in specific components of performance. Closer linkage of the research thrust in sports expertise with the present parallel theory development in motor control and learning, and in expert systems in cognitive science, appears imperative.

The major challenge from an applied psychology perspective is to translate the knowledge being acquired about sport expertise into useful practical directions for the guidance of sports training and instruction. Understanding more about the respective roles of implicit and explicit learning in expert performance, and utilizing fully the opportunities created through the new technologies (see *Simulation and Training in Work Settings*; *Virtual Reality*, *Psychology of*) will undoubtedly become increasingly critical in the future.

See also: Developmental Sport Psychology; Expert Systems in Cognitive Science; Expertise, Acquisition of; Sport and Exercise Psychology: Overview; Sport, Anthropology of; Sport, Geography of; Sport, History of; Sport Psychology: Training and Supervision

Bibliography

Abernethy B, Burgess-Limerick R J, Parks S 1994 Contrasting approaches to the study of motor expertise. *Quest* **46**: 186–98

Abernethy B, Thomas K T, Thomas J R 1993 Strategies for improving understanding of motor expertise. In: Starkes J L, Allard F (eds.) *Cognitive Issues in Motor Expertise*. Elsevier, Amsterdam, The Netherlands

Abernethy B, Wann J P, Parks S 1998 Training perceptual-motor skills for sport. In: Elliott B C (ed.) *Training in Sport: Applying Sport Science*. Wiley, Chichester, UK

Anderson J R 1982 Acquisition of cognitive skill. *Psychological Review* **89**: 369–406

Bernstein N 1967 *The Co-ordination and Regulation of Movements*. Pergamon Press, Oxford

Bloom B S 1985 *Developing Talent in Young People*. Ballantine, New York

Chase W G, Simon H A 1973 The mind's eye in chess. In: Chase W G (ed.) *Visual Information Processing*. Academic Press, New York

Chi M T H 1981 Knowledge development and memory performance. In: Friedman M P, Das J P, O'Connor N (eds.) *Intelligence and Learning*. Plenum Press, New York

de Groot A D 1966 Perception and memory versus thought. In: Kleinmuntz B (ed.) *Problem Solving Research, Methods and Theory*. Wiley, New York

Ericsson K A, Krampe R T, Tesch-Romer C 1993 The role of deliberate practice in the acquisition of expert performance. *Psychological Review* **100**: 363–406

Gibson J J 1979 *The Ecological Approach to Visual Perception*. Houghton-Mifflin, Boston

Singer R N, Janelle C M 1999 Determining sport expertise: From genes to supremes. *International Journal of Sport Psychology* **30**: 117–50

Starkes J L, Deakin J M, Allard F, Hodges N, Hayes A 1996 Deliberate practice in sports: What is it anyway? In: Ericsson K A (ed.) *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports and Games*. Erlbaum, Mahwah, NJ

Thomas K T, Thomas J R 1994 Developing expertise in sport: The relation of knowledge and performance. *International Journal of Sport Psychology* **25**: 295–312

Williams A M, Davids K, Williams J 1998 *Visual Perception and Action in Sport*. E & F N Spon, London

B. Abernethy

Sports, Economics of

1. *Introduction*

Sport is organized among individuals and among teams, and is professional or amateur. Most of the interesting questions on the economics of sport relate to professional team sport, which is the focus here. Over their histories, and until very recently, in some leagues, players have been restricted from moving between clubs within a league. Club owners have claimed that reservation of players or the imposition of transfer fees was necessary to maintain financial soundness of small clubs and to preserve competitive balance within leagues. In this article, pay and per-

formance under player reservation and under free agency are evaluated, along with issues relating to competitive balance. Under player reservation, player pay is less than the players' contribution to club incremental revenue. Under free agency, the players, ultimately, receive all of the incremental revenue. It is shown that the intrateam salary structure, with its wide pay differential between starting and backup players, is rational and wealth-maximizing. Competitive balance is value-maximizing, also. The effect of free agency, revenue-sharing among clubs, team salary caps, and syndicate ownership on competitive balance are evaluated.

2. Players' Market

2.1 Player Achievement: The Hot Hand and Sequences

In part, the appeal of professional team sport is the extraordinary playing ability of the athletes. This is demonstrated repeatedly on television, and commented on widely. Such feats are value-maximizing for the leagues, since they promote interest in the sport. A case can be made that baseball has been resuscitated by the record-breaking home run duel between Mark McGuire and Sammy Sosa, in the 1998 season.

While there are many ways of measuring player achievement, none excites more discussion than 'hot hands.' Mere mortals seek a causative explanation for the player who achieves a long sequence. But, hot hands is a myth, and the observed runs are no different from what one expects from a (possibly unbalanced) coin toss (successive heads in a row, when the probability of a head is some fixed number p). Even premier players in basketball do not make much more than half of their field goals. In the course of a season, such players would be expected to have one run of seven made shots in a row, and in the course of a career a sequence of 10 in a row. Naturally, players with higher shooting averages have longer sequences. Measured over the 50-odd year history of the National Basketball Association, the coin toss model predicts that at least one 0.5 shooter will make 16 shots in a row.

Research on sequences in sports (for basketball, see Gilovich et al. 1985, Camerer 1989) reveals only one exception to the coin toss model (Gould 1991). In 1941, Joe DiMaggio hit safely in 56 consecutive games. He hit 0.357 that year, and the coin toss model predicts that by chance one would see a sequence of that length once in 10 trillion trillion games. What an achievement!

2.2 Player Pay under Player Reservation and under Free Agency

Prior to the mid-1970s, baseball and basketball players could not negotiate freely with more than one team, and this restriction applied in football and in hockey, until more recent times. In European football (soccer), until 1995, player movement was restricted by the substantial transfer fee that had to be paid to the club owning the player's contract.

The result of player reservation in professional team sport was that players were exploited economically by the owners in the sense that they were paid less than the incremental revenue that their performance brought to the club (Scully 1974). It is not the case that the athletes in professional sports were poor by any means, since they were paid on average several times a working man's wage and the best of them received perhaps 10 times the pay of the less gifted performers. But, their pay was not equal to what they could have earned if they were free to offer their talent to the highest bidding club.

Veteran free agency has led to an enormous explosion in player salaries and in the share of club revenue captured by the players. Average player salaries are 50 or more times their level in the late 1960s, and player payroll costs are three or more times their share of club revenues (Scully 1989, 1995). While a superstar player then could expect to earn 10 times or so the pay of a rookie player, now a factor of 50 or more is not extraordinary. On the whole, then, the introduction of veteran free agency has eliminated monopsonistic exploitation of players in professional team sports. Nevertheless, pockets of exploitation exist, particularly among promising rookies, journeymen players, and first-round draft choices from the amateur ranks. This is so because there is not complete freedom in the players' market, such as would occur with teams competing for all players by offering one year or multiyear contracts for their services. New entrants into the professional ranks, if they survive the intense competition for their positions, must wait several years before they are free to negotiate with other clubs.

2.3 Pay Structure in Professional Team Sports

Payment for player performance is not piece rate, that is, a player who is twice as good as another, measured by some performance metric, does not receive twice the pay. Rather, relatively small differences in player performance often are associated with large differences in player compensation. The pay structure in professional team sport is hierarchical, as in a rank-order tournament (Lazear and Rosen 1981).

In professional sports, owners of clubs or sponsors of events, such as in tennis and golf, want all athletes to develop their skills to their maximum potential and for the athletes to expend maximum effort per contest. This desire is profit motivated: contests among highly skilled players who expend the maximum effort at winning are more valuable, and it is costly to monitor effort (prevent shirking). These facts imply that there will exist an optimal hierarchy of payments (an

optimal prize structure in golf and tennis, an optimal intraclub salary structure in professional team sport) that aligns player behavior (investments in developing playing skills and expending effort) with the owner's or sponsor's objective function. Individual sports, such as tennis and golf, are rank-order tournaments where the reward structure is predetermined and a competitor achieves the next higher payment by advancing one position in the rank of competitors. In professional team sports, there is a team rank-order tournament in which players compete for starting positions that are rewarded greatly and backup positions that pay a great deal less (Porter and Scully 1996). Starting players are always threatened by the performance of the lower-ranked backup players. This threat makes the performance of the starter like a serially repeated contest in the individual sports.

2.4 Pay and Performance Profiles

There are several features about player careers. Most enter the professional ranks as untested rookies, having spent a number of years in the minor leagues or in the amateur ranks. During that training period, pay is low. When a player gets a contract with a professional team, often they play infrequently and may be traded to another club. Only if their play is judged superior, will they play frequently. Player performance tends to improve over a certain portion of the playing career, reaches a peak, and then declines as age takes its toll on reflexes and physical ability. The rise in performance may be due in part to learning by doing, but mainly it is due to player and club investments in honing playing skills. The age at which the decline in performance sets in, and perhaps its rate, is related to player investments in skill and conditioning. Thus, the life cycle of performance and that of player pay is the result of a team rank-order tournament. Players invest endogenously in developing their skills, which determines their chance of securing a starting position. Under free agency, pay for starters is considerably more than for irregular players. The high rewards for starting players implies that every effort (training, conditioning, diet) will be made to maintain peak or near-peak performance for as long as possible.

Empirical evidence on the life cycle of performance and pay is measured most readily for baseball players (Scully 1995). During the monopsony era (sample period 1964–75), it was found that the peak in baseball player performance occurred at 1,596 games (about 10.6 years of play) and that the peak in salary occurred at 2,328 games (15.5 years). During the postfree agency era (sample period 1977–92), the peak in performance occurred much earlier (1,347 games or 9 years), and the peak in pay a great deal earlier (1,523 games or 10.2 years). The sharp increase in the hierarchy of earnings brought about by veteran free agency induced greater player investment in developing playing skills more rapidly. As such, baseball players reach their maximum performance potential about 1.5 years earlier under free agency.

Moreover, if one compares the profiles of performance and pay in the monopsony and the free agency periods one finds the following. Player productivity exceeds player pay up until 1,550 games (10.3 years) in the monopsony period, which is an indication of exploitation. In the free agency period, after 547 games (3.6 years) player pay exceeds player performance.

3. Competitive Balance

3.1 Competitive Balance within Professional Team Sports Leagues

An important dimension of playing quality is the relative strengths of the teams on the playing field (competitive balance). Fans want their favorite team to win under uncertainty. If the top club always beat the next ranked club, and so on, the games would be exhibitions, not contests, and they would have little market value. Some degree of uncertainty of outcome is necessary in sports, and this uncertainty is greater the more equal the playing strengths of the competitors. Greater equality of playing strengths among the clubs is wealth maximizing. Attendance at games and viewership on television is higher the closer the standings of the clubs are over a season.

Competitive balance depends mainly on the comparative financial strengths of the clubs. Teams earn revenue primarily from ticket sales and from broadcast fees. The main cost is player payroll. Attendance is determined chiefly by market (city) size and the club's record (Noll 1974). Since teams are located in cities of markedly different size, the prospect of fielding a competitive team depends in part on where the club is located, but also on how revenues are divided up among the clubs. In basketball and hockey, the home team receives all of the ticket revenue. In baseball, the gate split is about 85:15. In football, the gate division is 60:40. All national broadcast revenues are divided evenly among the clubs in all professional team sports. The difference in the formula for dividing up the gate receipts corresponds roughly to the dispersion in team financial strengths across the various leagues. The highest revenue-producing basketball team tends to have about 2.7 times the revenue of the lowest revenue-producing club, while it is about 2.9 times in hockey. The top club in baseball has about 2.5 times the revenue of the poorest drawing team. In football, the top team tends to have about 70 percent more revenue than the lowest revenue-producing team. With clubs distributed in unequal size markets and with a ban on unilateral relocation of clubs to the more lucrative markets, the more unequal the division of revenue among clubs the less will be the competitive balance within a league.

There are several ways of measuring competitive balance. The most obvious is to compare the actual number of championships with the theoretical number (one divided by the number of clubs times the number of years the club has existed). A second is to compare the historical average win records among the clubs. A third, and more precise, way is to compute the standard deviation among the win percentages of the clubs (Scully 1989). To make comparisons of competitive balance among the leagues, differences in the number of games played in a season needs to be taken into account, since the standard deviation of the win percent is inversely related to the number of games. Competitive balance is measured as the ratio of the actual standard deviation to the theoretical standard deviation under equality of playing strengths ($0.5/\sqrt{g}$, where g is the number of games).

Using this procedure, Quirk and Fort (1992) found for the period 1980–90 that football had the greatest competitive balance with a ratio ($\sigma/(0.5/\sqrt{g})$) of 1.53, while basketball had the least competitive balance with a ratio of 2.73. Baseball, with a ratio of 1.67 and hockey with 1.91 had relative playing strengths between football and basketball. Measured over the history of these leagues, the rankings of competitive balance remain unchanged. Football has always been the most competitively balanced, and basketball the least. Moreover, over time, competitive balance has improved significantly in baseball, somewhat in football, since the 1960s, but has remained substantially unchanged in basketball and hockey. Nevertheless, since the actual standard deviation of win percents greatly exceeds the theoretical standard deviation, one cannot conclude that competitive balance exists in any professional team sports league.

With the advent of veteran free agency, owners claimed that competitive balance would be destroyed. There is no difference in competitive balance in baseball in the period of veteran free agency compared to the player reservation period. While less clear cut, because freedom of player movement proceeded in steps in basketball, competitive balance has not changed in basketball in the era of freedom of player movement.

3.2 Freedom of Player Movement and Competitive Balance

Club owners, particularly in American baseball and European football, always have claimed that restrictions on player transfers are necessary to ensure competitive playing balance within a league and to protect the investments in the smaller city franchises. Economists, since Rottenberg (1956) have been very skeptical of this claim and there is no empirical evidence to support it. While it is true that there is a tendency for the best players to migrate to the larger city franchises, freedom of player movement or restrictions on player-initiated movement has no bearing on the distribution of playing talent within a league (El-Hodiri and Quirk 1971). Under free agency, players will migrate to the best paying clubs. The franchises that pay the most are those that expect the largest incremental revenue from those players' performances. Since an increment in the win percent raises revenue more in a large city than in a small city, the best players tend to migrate to the big city franchises. At issue is whether in a market where player-initiated movement is forbidden (player reservation, transfer fees), would a small city club keep a star player that it had acquired through a draft or an assignment? On the roster of that club, the player is worth, say, $1 million in incremental revenue. On a large city franchise, that player's performance is worth, say, $3 million. Without a scrupulously enforced league ban on cash sales of player contracts, the franchise in the small city has an incentive to sell the player's contract to the big city club, and capture some portion of that $2 million difference in incremental value. Therefore, players are allocated across clubs based on their highest incremental revenue. Whether the player or the club owns the contract (property right) is irrelevant to the allocation of playing talent within a league. That the efficient allocation of a resource is independent of who owns the property right is, of course, the famous Coase (1960) invariance theorem, but the essence of that theorem was discovered independently by Rottenberg (1956) in his earlier paper on the baseball players' market. While the distribution of playing talent is independent of the degree of freedom in the market for players, the distribution of the rents between owners and players is not. The players obtain a much higher share of the rents arising from their scarce talent under free agency than would be so under league restrictions on player-initiated transfers.

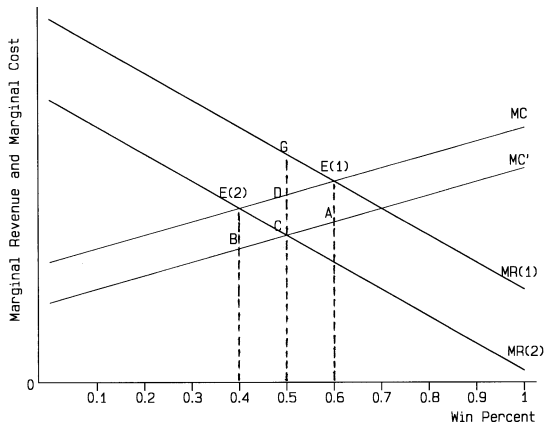

Figure 1
Freedom of player movement and competitive balance

The invariance of the allocation of playing talent to the degree of freedom in the players' market is illustrated in Fig. 1 for a two-club league. On the

vertical axis is club revenue and player cost per point in the win percent. On the assumption that revenue per point in the win percent declines as the win percent increases, marginal revenue (MR) declines with the win percent. MC is the cost per point in the win percent under free agency. Since top-playing talent is rarer than mediocre talent, its cost per unit is higher. Equilibrium, in the sense that both clubs maximize profits, occurs at $E(1)$ for the large city club (designated as team 1) with its higher marginal revenue, and where marginal revenue equals the marginal cost of playing talent ($MR(1) = MC$), and at $E(2)$ for the small city club (team 2), where $MR(2) = MC$. Thus, in equilibrium, $MR(1) = MR(2) = MC$. With $MR(1) > MR(2)$, the profit-maximizing distribution of playing talent is such that team 1 has a win record of 0.600 and team 2 has 0.400.

Now, assume that a reverse order of finish amateur draft, under which the club with the worst record gets to draft first, leaves each team with equal talent levels (each club will have a 0.500 record). Under this condition, the market is in disequilibrium. For club 1(cut point G), at a 0.500 win percent, marginal revenue exceeds marginal cost (the triangle GDE(1) represents potential profit), while, for club 2 (cut point D), marginal cost exceeds marginal revenue (the triangle CDE(2) represents a loss). Under free agency, club 1 will bid away playing talent from club 2 to the point at which a profit-maximizing equilibrium is re-established at $E(1)$ and $E(2)$, with club records of 0.600 and 0.400. Thus, under free agency, an initial equalizing redistribution of playing talent through a reverse order of finish drafting mechanism will be ineffective in maintaining equality of play with a league. The better players will migrate to the better paying, large city clubs, where their marginal revenue is higher.

Under player reservation, the player comes to terms with the club that owns their contract, or does not play. This condition leads to pay less than the incremental value associated with the player's contribution. The marginal cost per point in the win percent under monopsony is labeled MC′; in Fig. 1. Under monopsony, with an equal distribution of playing talent among the clubs (each club having a 0.500 record), team 2 has a level of playing talent such that marginal revenue equals marginal cost (point C in the figure), but club 1 has marginal revenue in excess of marginal cost (point G in the figure). There is more profit to be had on club 1, if it can purchase higher quality playing talent from club 2. It is in the interest of club 2 to sell that playing talent for cash, since by lowering its win percent profits increase. Equilibrium in the distribution of playing talent occurs when both clubs earn an equal profit from the distribution of playing talent (win percents). This occurs with a record of 0.400 for club 2, with $MR(2) > MC'$; by $BE(2)$, and 0.600 for club 1, with $MR(1) > MC'$; by $AE(1)$, and with $BE(2) = AE(1)$. Observe that under monopsony the profit-maximizing win percents remain as they

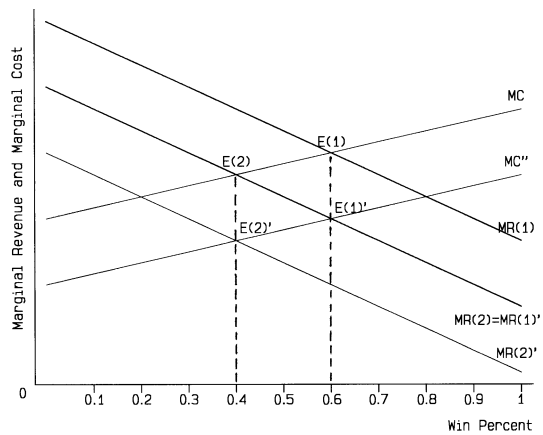

Figure 2
Revenue sharing and competitive balance

were under free agency. The gaps $AE(1)$ and $BE(2)$, in the figure, represent the monopsony rent transferred from the players to the owners under a player reservation rule. Thus, under monopsony, the distribution of playing talent remains the same as under free agency, but the big city club has cross-subsidized the small city team through a cash-sale purchase of player contracts. Equalization of playing talent is not possible under free agency or under player reservation with cash sales of player contracts. The only way in which reverse-order drafting will lead to equalization of playing talent among clubs that play in different size cities is under player reservation with a ban on cash sales of player contracts. But, note that under such a ban profit maximization is not achieved. Club 2 earns more profit with a win percent of 0.400 than with a record of 0.500, and club 1 earns more with a record of 0.600. Even if a formal ban on cash sales is imposed by a league, these profit opportunities are a powerful incentive to devise mechanisms for player reallocations (lopsided player trades, a player trade for compensation and a player named later) that will restore the distribution of win percents among the clubs toward their profit-maximizing distribution.

3.3 Revenue Sharing and Competitive Balance

If the degree of competiveness in the players' market does not affect competitive balance, will it be affected by the division of revenue among the clubs? Revenue division tends to moderate the distribution of revenue among teams within a league, but has no effect on competitive balance. The conclusion is illustrated in Fig. 2. Without revenue sharing, equilibrium is at $E(1)$ for team 1 and $E(2)$ for club 2. Profit maximization leads to 0.600 and 0.400 win records. Now, let revenue sharing occur such that each club has the same marginal revenue. The effect of revenue sharing is to

lower the marginal revenue for both clubs from $MR(1)$ to $MR(1)'$ for team 1 and from $MR(2)$ to $MR(2)'$ for club 2. Algebraic proof: under revenue sharing $R_i = sR_i + (1-s)R_j$, $i = 1, 2$, $j = 1, 2$, $i \neq j$, with s as the revenue share. In a two-team league, an increase in the number of games won (or, the win percent) for one club means a corresponding decrease in wins for the other club. Then, the change in revenue is $MR_i' = sMR_i - (1-s)MR_j$. The profit-maximizing equilibrium under revenue-sharing is at $E(1)'$ for club 1 and $E(2)'$ for club 2, where $MR(1)' = MR(2)' = MC''$. Note that the marginal cost of playing talent is proportionately lower under revenue division than in its absence ($MC'' < MC$). This is so because the worth of playing talent to the clubs is inversely proportional to the amount of revenue sharing. The distribution of playing strengths among the clubs is unaffected by the revenue division rule. While revenue sharing reduces the incentive of clubs to bid for scarce playing talent, it also reduces the financial incentive to win.

3.4 Team Salary Limits and Competitive Balance

Since the advent of veteran free agency in professional team sport, owners have sought mechanisms to constrain the competition for playing talent that they perceive as financially ruinous. One mechanism, invented by basketball, in 1980, is the team salary cap, a device that in theory imposes a maximum salary as a percent of gross revenue (and, a minimum percentage that applies to small city clubs). Football, since 1993, has a team salary cap, and baseball wants one, but has been blocked in its ambition by the players' union. Can a salary cap equalize playing strengths within a league? Suppose that a binding salary cap is imposed that fixes player salaries at a certain percentage of league revenues. The marginal cost of playing talent falls under the salary cap from its level under free agency. If scrupulously enforced, all clubs have the same team salary costs, and if there are no information asymmetries about expected playing talent and no differences in coaching quality, all clubs will have 0.500 records. Like the solution of player reservation with a ban on cash sales of player contracts, the distribution of win percents brought about by a hard salary cap is stable, but it is not profit maximizing, nor does it maximize league revenue. A large city team can earn more profit by acquiring better playing talent and a small market club can earn more by lowering its level of playing talent.

In basketball, the team salary constraint is relaxed significantly by the so-called Larry Bird exemption. Under the exemption, a team may re-sign a veteran free agent at whatever cost, and not have that player's salary count in the salary cap. This has led to most clubs having player payrolls well in excess of the cap, and this transforms the team salary limit concept into a soft or nonbinding constraint. The Larry Bird exemption was the main point of dispute between the basketball players and the owners, prior to the 1998–9 season. Salary caps have led, also, to creative financial arrangements such as large player signing bonuses relative to annual salary, back-end loaded long-term player salary contracts, deferred player compensation, and the shifting of revenue from team operations to other, nonteam entities. Sufficient erosion in the constraint of the salary cap, as with a player reservation system with player transfers for cash, can lead to the same distribution of playing talent as is attained with unrestricted free agency.

3.5 Syndicate Ownership and Competitive Balance

Major league soccer in the US is organized in an entirely different way than the main professional team sports leagues. The latter are composed of clubs that are privately owned and that are subject to various agreements with their respective leagues. In soccer, players contract with the league, not a club within the league, a fixed team salary is enforced, investors operate but do not own individual clubs, and all revenues are divided equally among the clubs. Thus, in theory, marginal revenue is equalized across all clubs at a win record of 0.500. Alternatively, there is no extra money to be made by a club from winning a championship. Nor is there any financial penalty for finishing in the cellar. While syndication will bring about equality of playing strengths, it raises very serious questions about club independence and competitive integrity. Baseball's early experiment with syndication failed. Interleague wars weakened many clubs and the aid from the stronger to the weaker clubs led to cross-ownership. In 1901, the surviving National League clubs formed a syndicate. Fans were not impressed, and believed that the outcomes of the games suspect. 'Trust' or cross-ownership in the National League was ended in 1902.

Bibliography

Camerer C 1989 Does the basketball market believe in the 'Hot Hand'? *American Economic Review* **79**: 1257–61

Coase R 1960 The problem of social cost. *Journal of Law and Economics* **3**: 1–44

El-Hodiri M, Quirk J 1971 An economic model of a professional sports league. *Journal of Political Economy* **79**: 1302–19

Gilovich T, Vallone R, Tversky A 1985 The Hot Hand in basketball: On the misperception of random sequences. *Cognitive Psychology* **17**: 295–314

Gould S J 1991 *The Streak of Streaks. Bully for Brontosaurus: Reflections on Natural History*. Norton, New York

Lazear E P, Rosen S 1981 Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* **89**: 841–64

Noll R 1974 Attendance and price setting. In: Noll R (ed.) *Government and the Sports Business*. Brookings, Washington, DC

Porter P K, Scully G W 1996 The distribution of earnings and the rules of the game. *Southern Economic Journal* **63**: 149–62
Quirk J, Fort R 1992 *Pay Dirt: The Business of Professional Team Sports*. Princeton University Press, Princeton, NJ
Rottenberg S 1956 The baseball players'-labor market. *Journal of Political Economy* **64**: 242–58
Scully G 1974 Pay and performance in major league baseball. *American Economic Review* **64**: 915–30
Scully G 1989 *The Business of Major League Baseball*. University of Chicago Press, Chicago
Scully G 1995 *The Market Structure of Sports*. University of Chicago Press, Chicago

G. W. Scully

Sports Performance, Age Differences in

'Sports,' wrote veteran journalist Frayne (1990) in his memoir, are just 'children's games that seem to enchant millions of grownups.' Frayne's irony is fitting. Although dictionaries define sports as pleasant pastimes, to the athletes, supporters, and wider society, sport is so much more. From classical times to the modern era, sports waxed and waned in their significance for religion, militarism, business, and occupation (McIntosh 1963). Innumerable people devoted their lives to sport, and far too many lost life because of sport. At least one war was blamed on sport, and at least one war halted because of sport (Morris 1981). Nowadays, television has made sport a truly international drama, enchanting, enraging, dismaying, or even boring audiences of tens of millions.

Social and behavioral scientists study sport for many reasons. Social scientists study societal influences on sports and of sports. Behavioral scientists study sports to gain knowledge about expertise in physical performance, including its developmental progression across the life span. This article examines the attributes of expert performance in different sports in relation to the aging of expert performers.

1. Expertise and Performance

1.1 Identification of Expertise

Who is an expert? A simple answer might be someone with acknowledged skill. This interpretation equates expertise with competence. Although most people are competent at skills practiced routinely in everyday life (e.g., driving), the appellation of expertise usually connotes a higher degree of exclusivity. Experts are not people who are simply competent, but those who exhibit exclusive competence consistently (Salthouse 1991).

There are two types of experts. First, experts who are normatively exclusive lie at the upper extreme of normal distributions for everyday skills (e.g., walking, running, cycling). However, normative expertise also depends upon the population providing the distribution. Running 1,500 meters in 6 minutes certainly denotes expertise within the overall adult population, falling within the top few percentile points. But compared to other runners, such a performance might be merely average. Unless that runner was a male aged 80 years or a female aged 65 years, in which case the performance would approximate a world record. Most sports accommodate to normative relativism by including age and gender distinctions within the competitive categories.

Second, other experts have exclusive skills not possessed by the majority of the population. Only seasoned race walkers can walk 1,500 meters in 6 minutes because none but race walkers have mastered that effective but peculiar style of walking. Most race walkers, before the 1950s, were of European ancestry. Similarly, equestrians might be considered experts because few people in relatively few countries ride horses competitively. These examples illustrate cultural and historical relativism in exclusive skill, with both pedestrian and equestrian pursuits being less popular today than in previous centuries.

Athletes at the upper levels of expertise are termed elite athletes. Elite athletes include those who break world records, win international championships, or compete at the highest levels within their country. Comparison of age differences in performances by elite athletes can clarify the limits to which life-span developmental processes impact on physical performance.

1.2 Paradigms for the Study of Expertise

Research on expertise gained impetus from studies of games like chess. Chase and Simon's (1973) research with chess players of different proficiencies elucidated cognitive processes contributing to superior performance. The same dominant paradigm of comparing people at different levels of proficiency found application in studies of occupation (e.g., surgical skill, typing, computer programming), recreation (e.g., bridge, video games), and sport (e.g., badminton, soccer). Spirduso (1980) pioneered a variant of this paradigm by comparing athletes and nonathletes at different ages to identify fast reaction time as an ability retained by athletes.

Other approaches to the study of expertise derived from ability, ecological, informational, and training models (Allard 1993). These paradigms respectively seek to identify abilities common to expert performers, environmental cues that trigger skilled performance, sequencing within performance, and the limits to which expertise can be enhanced by training. A recent development is the application of compensation theory to performances by aging athletes (Stones and Kozma 1995).

Statistical modeling of age trends in athletic performance has emerged as an influential paradigm in the study of sports expertise (Moore 1975, Stones and Kozma 1986). Age-class records in individual sports such as singles rowing, swimming, track and field, and weight lifting provide invaluable sources of data about age changes in physical performance. Not only is performance measured as meticulously as in any laboratory study, but the record holders (being the best in their country, or the world, at their respective age) can reasonably be assumed to be fit and healthy, with their performance relatively unaffected by chronic disease. Although sport for veterans became organized only during the 1970s, the numbers of competitors are now very large in some sports. The biannual championship of World Association of Veteran Athletes regularly attracts 5,000 or more track and field athletes from around the world, and the US Masters swimming organization has more than 37,000 members.

It is more difficult to evaluate performances by older athletes in team sports. Changes in team tactics, the composition of a team, the overall performance by a team, and transfer to another team are all factors that can affect the contributions of individual athletes. It is difficult to disentangle such extraneous effects from those of changing skill. However, two kinds of data enable insight into the sources of age differences in performance. First, aggregated longitudinal data show age trends within a sport. Examples include Stones and Kozma (1995), who analyzed aggregated data from ice hockey, suggesting that players compensate for late-career decreases in point scoring by aggressive play (i.e., penalty minutes). Also, Langford's (1987) research showed that home runs and batting averages in baseball peaked around the age of 27 years. Second, the ages of elite athletes can be compared across different sports. Limitations to this kind of comparison include changing historical trends, such that ages at peak performance and retirement in the 1900s may not mirror those of today.

1.3 Components of Expertise

The components of motor expertise include knowledge, skill, experience, and motivation. These components are hypothesized to change in relative importance over the course of athletes' careers (Abernethy et al. 1993). Knowledge becomes less important after an early acquisition phase; motor skills decrease in significance, particularly in later life; experience increases in importance over the life span; motivation may decrease or reflect changing priorities (see *Sports as Expertise*, *Psychology of*).

Accumulated experience contributes to performance in two ways. First, athletes become more adept in the use of strategy. Second, experience helps athletes to choose among different modes of compensation if performance falls below expectations. Examples of compensation include attempts improve skill, the substitution of an alternative skill, or modification of expectations (Stones and Kozma 1995).

Motivation can be differentiated into intrinsic and extrinsic components. Intrinsic motivations include desires to achieve, retain competence, and remain healthy. Extrinsic motivations include praise and payment. Most older athletes are long-time competitors who cite intrinsic benefits to health and competence as reasons for participation.

1.4 Categorization of Sports

Sports have been categorized on three dimensions other than individual vs. team sports. The first is between closed and open skills. Closed skills sports are performed in an unvarying environment, with the performance evaluated by the quality of movement patterns. Examples include individual sports such as figure skating, diving, and gymnastics, and team sports like synchronized swimming. Open skill sports are performed in a varying environment. The sources of variation include changes produced by an opponent (e.g., boxing, baseball, soccer), as well as in physical topography (golf). In contrast to closed skill sports, performance in open skill sports is evaluated by the outcome of movement rather than the motor pattern itself.

The second distinction refers to physical effort, the main components of which are power and endurance. Power events fall into two categories, both of which are 'fueled' by anaerobic sources that are depleted rapidly. Sprinting involves a continuous expenditure of high power over a short period. Peak power events involve an explosive use of power (e.g., jumping, throwing, pole vaulting, weight lifting). Events of longer duration require endurance rather than power (e.g., distance running, race walking). Such activity draws mainly on aerobic sources that provide for relatively low power expenditures over extended periods of time. Not all sports require much physical effort. Frayne (1990) wrote the following about baseball: 'Except for the pitcher and catcher, nobody does much of anything for long periods but stand around and wait.' Cricket is similar in this respect to baseball, and contrasts with sports like rugby and football that intermittently tax both power and endurance.

The final distinction refers to the use of strategy. Open skill sports such as motor racing, snooker, and tennis all are high strategy sports. Low strategy activities include most closed skill sports, but also open skill sports such as darts, swimming, and archery.

2. Age Differences in Different Sports

The preceding categorization suggests that sports can be grouped into four main categories. Closed skill sports are different from other sports because per-

formance is evaluated by the quality of movement rather than the outcome of movement. Although such sports vary in the physical effort required (e.g., divers expend less effort than gymnasts or skaters) practice is more important to performance than strategy. Open skill sports can be divided into three groups: sports requiring (a) effort rather than strategy, (b) strategy rather than effort, and (a) both strategy and effort. Differences in the ages of the elite performers are as follows.

2.1 Closed Skill Sports

Athletes in closed skill sports typically attain peak performance at a young age. Examples include Fu Mingxia, who had never practiced a complete set of dives before she became part of the Chinese juniors' diving team at age 11 years. Within 3 years she was both world champion and the youngest Olympic champion to that time. At the same 1992 Barcelona Olympics, another Chinese diver became a men's champion at the age of 13 years. Elite gymnasts are also youthful performers, with 14-year-olds regularly appearing in elite competition. In figure skating, Tara Lipinski of the US became the Olympic champion in 1998 at age 15 years, and the youngest ever champion in an individual event. Another figure skater, Sonja Henie, previously held that distinction.

Although approximately 15 percent of Olympic triple gold medal winners (i.e., winners of the same event in three successive games) competed in closed skill sports, few such athletes participate in elite competition beyond age 30 years. Many elite skaters become professional entertainers during retirement. Many gymnasts, including former Olympic champion Olga Korbut, become coaches. These examples illustrate compensation, wherein athletes substitute skills related to those formerly practiced.

2.2 Open Skill Sports Requiring Effort Rather than Strategy

Although swimmers may break world records in their teens, the recent history of track and field overturns traditional beliefs that athletes in power events both peak and fade at earlier ages than athletes in endurance events (Spirduso 1995). With the exception of the race walker Wan Yan, the youngest and oldest world record holders and Olympic champions all competed in power events.

Statistical modeling of age differences among veteran athletes suggests that performance decreases proportionate to the demands on power and endurance (Stones and Kozma 1986, 1996). Figure 1 illustrates these trends with swimming records.

Figure 1 is based on 1999 US Masters swimming records at different ages expressed as a ratio of 1999 world records. The ratios were combined for males and females in the 50 m and 200 m backstroke and butterfly. The butterfly requires more power than the backstroke (Astrand and Rodahl 1977), and the 200 m event taxes endurance more than the 50 m. The findings show performance times to increase exponentially with age, with greater age differences in events requiring higher power (butterfly) and more endurance (200 m). Comparable trends can be obtained from analysis of track and field records (Spirduso 1995, Stones and Kozma 1996).

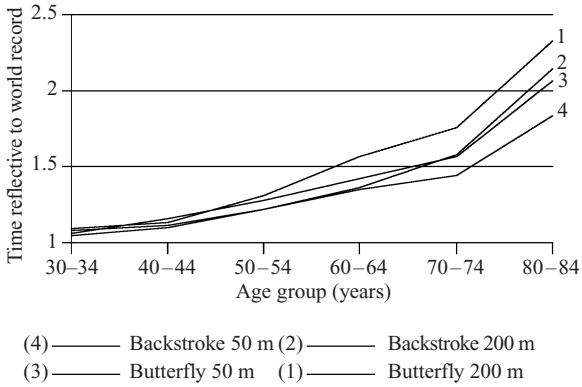

Figure 1
Age group records relative to world records in 50 m and 200 m backstroke and butterfly swimming events

2.3 Open Skill Sports Requiring Strategy Rather than Effort

Archery, baseball, cricket, equestrianism, golf, fencing, motor racing, and yachting all belong to this category. Although the athletes may earn elite status in their teens, their careers include the longest in any sport. Archer Galen Spencer won an Olympic gold medal at age 64 years. The youngest and oldest major league baseball players in the twentieth century were respectively aged 15 and 59 years. Seven Pakistani cricketers represented their country at 16 years or younger, and four cricketers from various countries competed internationally when aged over 50 years. In equestrian sport, the age range for the youngest and oldest qualifiers for the National Finals Rodeo in the US is the same as for baseball, and the career of Lester Piggott (England's most famous jockey) spanned 47 years (i.e., from his first win, at age 12 years, to his retirement at age 59 years). Ivan Osiier competed in Olympic fencing competitions over a 40-year span. The youngest and oldest golfers who won a national championship (Sri Lanka) were aged 12 and 54 years, respectively. Although the youngest Grand Prix motor racing winners were in their twenties, the oldest winners and competitors were aged in their fifties. Magnus Konow competed in Olympic yachting over a 40-year period. Experience undoubtedly contributes

to length of careers in these sports, with the practiced motor skills being resistant to age-related physical decline.

2.4 Open Skill Sports Requiring Strategy and Effort

Such sports include basketball, various forms of football (e.g., soccer), ice hockey, and tennis. The youngest players to compete at elite levels are in their mid-to-late teens, and the oldest generally in their forties. Examples include elite basketball players (youngest, 18 years; oldest, 43 years), Wimbledon tennis champions (youngest, 15 years; oldest, 43 years), and international soccer players (youngest, 17 years; oldest, 42 years). Within domestic soccer, Stones and Kozma (1995) reported that more than half of the oldest professional soccer players in the top European nations were goalkeepers, a position that requires the least physical effort. This finding suggests that careers may be longer if the demands on effort are lower. Very few athletes in this category compete at elite levels in their fifties, but with notable exceptions that include Sir Stanley Matthews in soccer and Gordie Howe in ice hockey.

3. Conclusions

Aging lowers the potential to perform at elite levels in sports that require power, endurance, endurance, and closed patterns of skill. Aging has less effect on performance in high strategy sports, and particularly those that require low effort.

See also: Aging, Theories of; Developmental Sport Psychology; Intelligence, Prior Knowledge, and Learning; Sport and Exercise Psychology: Overview; Sport Psychology: Performance Enhancement; Sports as Expertise, Psychology of

Bibliography

Abernethy B, Thomas K, Thomas J T 1993 Strategies for improving understanding of motor expertise (or mistakes we have made and things we have learned!). In: Starkes J L, Allard F (eds.) *Cognitive Issues in Motor Expertise*. Elsevier, North-Holland

Allard F 1993 Cognition, expertise, and motor performance. In: Starkes J L, Allard F (eds.) *Cognitive Issues in Motor Expertise*. Elsevier, North-Holland

Astrand P-O, Rodahl K 1977 *Textbook of Work Physiology: Physiological Bases of Exercise*, 2nd edn. McGraw-Hill, New York

Chase W G, Simon H A 1973 The mind's eye in chess. In: Chase W G (ed.) *Symposium on Cognition* (*8th*; 1972; *Carnegie-Mellon University*): *Visual Information Processing*. Academic Press, New York

Frayne T 1990 *The Tales of an Athletic Supporter*: *A Memoir*. McClelland & Stewart, Toronto

Langford W M 1987 *Legends of Baseball: An Oral History of the Game's Golden Age*. Diamond Communications, South Bend, IN

McIntosh P C 1963 *Sport in Society*. Watts, London

Moore D H 1975 A study of age group track and field records to relate age and running speed. *Nature* **253**: 264–5

Morris D 1981 *The Soccer Tribe*. Cape, London

Salthouse T A 1991 Expertise as the circumvention of human processing limitations. In: Ericsson K A, Smith J (eds.) *Toward a General Theory of Expertise*. Cambridge University Press, Cambridge, UK

Spirduso W W 1980 Physical fitness, aging, and psychomotor speed: A review. *Journal of Gerontology* **35**: 850–65

Spirduso W W 1995 *Physical Dimensions of Aging*. Human Kinetics, Champaign, IL

Stones M J, Kozma A 1986 Age trends in maximal physical performance: Comparison and evaluation of models. *Experimental Aging Research* **12**: 207–15

Stones M J, Kozma A 1995 Compensation in aging athletes. In: Dixon R, Backmän L (eds.) *Compensating for Psychological Deficits and Declines: Managing Losses and Promoting Gains*. Lawrence Erlbaum, Mahwah, NJ

Stones M J, Kozma A 1996 Activity, exercise and behavior. In: Birren J, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA

M. J. Stones J. Farrell, and J. Taylor

Sports Performance, Self-regulation of

Self-regulation refers to a regulation of the psychological state of an individual by the individual her- or himself. Self-regulation becomes necessary when internal or external barriers threaten the efficient realization of an intended action. Self-regulation consists of auxiliary processes that are based on meta-knowledge (metacognitive or metamotivational knowledge). These auxiliary processes support basic action processes like attention, motivation, etc., if those are not sufficient for optimum goal attainment.

Within the last decades it has been widely acknowledged that the so-called mental side is a decisive factor in athletic performance. As physical training has become more and more optimized for all athletes participating in major competitions, the winning edge is often furnished by an athlete's self-regulatory ability. Of two equally talented and equally well-prepared athletes, the one with better psychological skills is likely to win. Without such skills athletes may become their own worst enemy during competition, imposing psychological barriers upon themselves. Self-regulation helps to overcome these barriers.

1. A Brief History of Sport Psychology and Self-regulation

Self-regulation has a comparatively short history within the not very long history of sport psychology. Although it is sometimes claimed that sport psy-

chology has a history that dates back to ancient Greece, its scientific origins are not older than about 100 years. From the late nineteenth century on, the first decades of this discipline were mainly focused on the scientific study of athletic behavior. The first book with an applied perspective, *The Psychology of Coaching*, was published in 1926 by Griffith in the USA. During the 1960s and 1970s there was a considerable increase in the application of psychology in the field of athletics. However, self-regulation was mainly restricted to the use of relaxation techniques. At the same time self-regulation had become an evolving approach within behavior therapy. In the late 1960s, behavior therapists began to fill the black box, the organism variable, within the S-O-R model of behaviorism with cognitive processes individuals could actually use to control their own behavior (e.g., Kanfer 1970). In the late 1970s behavior therapists took the models and techniques of self-regulation developed in clinical psychology to the field of athletic behavior (e.g., Mahoney and Avener 1977). These self-regulation techniques became the core elements in 'mental training' or 'psychological skill training' for athletes which today is frequently seen as the major training content of applied sports psychology.

2. *Naive Self-regulation vs. Psychological Skill Techniques*

Individuals use self-regulation in everyday life, e.g., to resist the temptations of delicious food when they have committed themselves to a diet. Most of the self-regulation techniques used in daily situations are referred to as 'naive' because they are not derived from scientific knowledge and their effectiveness has not been systematically evaluated.

Usually, naive self-regulatory techniques are highly personalized coping strategies for situations constituting a challenge or a threat for the person. These techniques have developed through the person's experience with that situation and they are based on the person's subjective criteria. Some of these techniques are rites that have developed and some of them are simply instances of superstitious behaviors. These 'naive' self-regulation techniques can have positive effects. They may, for example, reduce athletes' fear of the uncontrollable, increase self-confidence, and induce a 'winning feeling.' However, because they do not systematically address disadvantageous mental performance conditions, they are often bound to fail.

In contrast to the 'naive' self-regulation techniques, those imparted by sport psychologists are 'a formal, structured application of psychological skill techniques to the enhancement of sport performances' (Morris and Bull 1991, p. 4). Sport psychologists train athletes in the use of these psychological skills. The objectives of such psychological skill training are manifold. A study by Mahoney et al. (1987) found sport psychologists and athletes to agree that the most basic skills to be acquired by athletes are anxiety management, concentration, motivation, self-confidence, and mental preparation.

2.1 *Self-regulation and Self-control*

In general, the approach to develop self-regulation skills in athletes is based on the premise that the right psychological climate helps mobilize physiological reactions that are essential to performing at one's best. A negative psychological climate such as feelings of frustration, fear, anger, and worry, or negative thinking in general, typically does the opposite. Loehr (1984) emphasized that neither negative nor positive psychological states just happen. Athletes can identify their own ideal performance state, and they can learn, intentionally or subconsciously, to create and maintain this state voluntarily so that their talents and physical skills thrive. This is the basic credo of self-regulation. The mental skills needed to trigger and sustain one's ideal performance state are learned through practice just as physical skills are learned. Some gifted athletes seem to be able to develop efficient 'naive self-regulation' techniques, but as mentioned above, these may have their flaws. Thus, most athletes need to be taught specific mental skill techniques. Psychology has systematically studied and developed such techniques, so that they are now available in sport psychology.

Self-regulation and self-control are key terms in motivation and action. They refer especially to volitional abilities that are needed to deal with adverse conditions during the execution of an intention. Kuhl and Beckmann (1994) subsume self-regulation and self-control processes under the more general conception of action control. Action control refers to the correspondence of the actual behavior with the intended action or, more specifically, the correspondence of the actual performance with the intended performance. Under adverse performance conditions, action control processes, i.e., self-regulation and self-control, are required to achieve a high degree of such correspondence.

In general, there are three objectives to be achieved through self-regulation:

(a) positive self-related control beliefs (e.g., self-efficacy)

(b) positive self-related emotions (high self-esteem)

(c) clear-cut distinctions between self and nonself (self-discrimination).

Most self-regulation techniques that are employed in sports address the first two objectives. The third objective is a more general developmental goal. It is based on the assumption that only individuals with access to their own needs and values (i.e., their integrated self) can perform up to their potential.

There are two different routes to master adverse conditions, i.e., to cope with threats to action control. One route is the intention-driven inhibition of the counterintentional processes. The other route is the promotion of conflict-reducing processes through supporting the facilitation of intention-relevant processes. Kuhl and Beckmann (1994) refer to the former as self-control and to the latter as self-regulation. Self-control is the less sophisticated action control strategy. It is based on metacognitive and metamotivational knowledge to a much lesser degree than self-regulation. Nevertheless, in some sport disciplines self-control can actually mediate top performance. Beckmann and Kazen (1994) found that athletes who primarily use self-control excel in sport disciplines requiring a short-lived, high-energy input like shot-putting, sprint and jump disciplines in track and field. But self-control failed in sport events demanding energy regulation, for example, long distance running. In these disciplines athletes with good self-regulatory skills showed the best performance.

Different tasks or sport disciplines require various kinds of self-regulatory abilities. In endurance sport, for example, it is important to stick to one's competition plan. In disciplines like sprint, shot-putting, or weight-lifting, maximum energy mobilization combined with high concentration is required, but only for a few seconds. In decathlon it is crucial to mobilize energy for a specific discipline but relax and regenerate between the heats.

3. *Self-regulation Techniques*

Several self-regulation techniques have been developed which an athlete can acquire through mental training. There is no unified terminology for psychological skills training. Seiler (1992, p. 31) speaks of 'a certain randomness of the classification, depending on the author's initial conception of what is important in sport and competition and the construction method itself.' As a result, different mental training programs stress different elements. Most programs share a focus on the following elements: activation regulation (relaxation and energization), attention regulation (concentration), self-talk (positive thinking, thought control), imagery (visualization, mental rehearsal). In the numerous mental training programs the different techniques usually blend in with one another. For example, a lot of techniques make use of images to support self-regulation. Only the major techniques of self-regulation will briefly be described in this article.

3.1 *Activation Regulation*

Self-regulation in sports was often equated with relaxation, especially in the early days of sport psychological interventions. The goal was to generate a state of psychovegetative functioning, which was considered optimal for the performance of a certain task. The guiding idea behind this intervention was the, so-called, Yerkes–Dodson law of the relationship between arousal and performance. In 1908, Yerkes and Dodson found an inverted U-shaped relationship between arousal and performance in a learning experiment. In this experiment, rats learned slowest when they were highly aroused, or when their arousal was low. Their learning was fastest when arousal was at a medium level. As a consequence of this finding generations of coaches and athletes were taught that medium arousal was optimal for performance. Because the majority of athletes show a high level of arousal before a competition, relaxation was promoted as a major technique of mental preparation for a competition.

But contradictory empirical findings show that the relationship between activation and performance is more complicated than postulated in the Yerkes–Dodson law. Neiss (1988) assumed that performance impairment found in highly aroused individuals is mediated through task-irrelevant cognitions (worry) rather than through the arousal component itself. Supporting this assumption, Dienstbier (1989) reported numerous studies that show a general positive relationship between physiological arousal and performance. Beckmann and Rolstad (1997) conclude that thorough relaxation may not be the optimal preparation immediately before starting a competition. Rather, athletes should aim at increasing their precompetition activation through so-called 'psyching up.' But this should be combined with the self-regulation technique of thought control (avoiding negative thoughts and worries).

Although relaxation techniques may not be adequate immediately prior to or during a competition, they still play an important role in psychological skills training. They are necessary as a first step in training other self-regulation techniques. On their own, they are also an important measure for personality development (developing an even temper, the ability to 'let go,' supporting recovery, higher tolerance for unexpected and uncontrollable events).

Relaxation methods taught in mental training are progressive muscle relaxation (PMR) and autogenous training (AT). These rather extensive relaxation programs intend to reduce bodily activation through focusing on relaxing muscles (PRM) or autosuggestive formulas like 'Heart beats calmly and steadily' (AT).

A very basic technique is 'controlled breathing.' The breathing is conducted using the diaphragm rather than the chest. Athletes concentrate on 'feeling the air enter the lungs and then slowly leave' (Smith 1998, p. 42). 'Controlled breathing' has proven to be a good means to 'let go,' which means to get rid of disturbing, negative thoughts. An advantage of 'controlled breathing' compared to other relaxation techniques is that it does not necessarily reduce athletes' readiness

for competition. Consequently, it is a relaxation technique that has its place in the preparation for competition.

3.2 Attention Control

For the majority of sport disciplines, attention is crucial for performance. Attention is threatened by various sources, especially during a competition. Through distractions athletes can lose the focus which allows them to execute their tasks effectively. Therefore, the ability to refocus in the face of distractions, to manage to concentrate on the task, is a critical self-regulatory skill.

Distractions come from a variety of sources, outside and inside athletes. The crowd, the competition conditions, a starting aircraft while an athlete tries to concentrate on her high jump, thoughts about the finishing place, or a specifically difficult and dangerous section of a downhill ski run while still gliding through an earlier section.

Nideffer (1980) made techniques of attention control popular among coaches and athletes. These techniques are used to focus or refocus attention more effectively during sport performance. One of these techniques is called 'centering': this is a technique used in a stressful situation to regain concentration. It involves taking a deep abdominal breath, with the individual focusing attention on the breath and then back out to a positive external cue.

Inducing positive affect (emotion control) is a frequently suggested aid in attention control. The advice is to focus on doing what will help keep the person positive and in control. The idea is that a strong positive focus protects against distractions. Usually a cognitive component is added: the athlete should develop specific, task-related goals that can capture attention; a focus that connects or reconnects to the task. This can be supported by images. Athletes can, for example, imagine being in an invisible force field. The slalom skier imagines that the slalom course is inside a tunnel of glass that keeps all possible distractions outside. Also, athletes can imagine a guiding line that carries them through their competition e.g., the downhill skier imagines a railroad track laid out along the ideal line of the downhill run.

Another, more general strategy involves relaxation training, in this case, as a form of personality development. Through relaxation training (PMR or AT), which is combined with other training routines on a regular basis, athletes can learn to 'let go.' They can learn not to carry mistakes with them and practice getting back on track quickly.

3.3 Self-talk and Thought Control

Another major route of self-regulatory skills for athletes became popular during the 1970s. These mental strategies are based on approaches in cognitive behavior modification which address the altering of 'private' cognitive processes. These techniques aim at stopping negative and destructive thoughts and promoting positive thoughts that support performance.

Gallwey (1976) has pointed out that tennis players lose their games in their heads first. When the 'inner dialogue' changes from a positive, confident focus to a negative one, the game is lost. Gallwey developed the so-called 'inner game' technique which became popular especially in tennis and golf. The inner game addresses the inner dialogue between what Gallwey (1976) calls Self 1, the analytical problem solver, and Self 2, the intuitive and emotional self. The inner dialogue can become critical, when Self 1 primarily articulates self-doubts or anxieties and thus threatens the maintenance of positive self-related control beliefs and positive self-related emotions. Strategies either have the task-related Self 2 completely occupy attentional capacity or alter the focus of Self 1 toward more positive, supportive thoughts.

The importance of having a positive inner dialogue is underlined by a study by Mahoney and Avener (1977). They found that the self-talk of US gymnasts who did not qualify for the Olympics was dominated by self-doubts and thoughts about failures. In contrast, the self-talk of those who qualified was characterized by self-confidence.

A simple technique that is often employed to stop negative thoughts is directly taken from behavior therapy: stop rules. Athletes are trained to shout 'stop' along with, for example, imagining a large stop sign to interrupt a negative thought sequence. But stopping the negative thoughts alone is not sufficient to get back into competition. After stopping the negative thoughts, it is important to refocus on an appropriate performance-related cue or positive thought. However, this technique involves an interruption of task performance or at least a momentary loss of concentration.

It is more efficient to prevent intrusive thoughts or negative thoughts from occurring beforehand. Beckmann (1993) used a thought-stopping technique to prevent negative thoughts, which has proven to effectively stop thoughts in numerous experiments in cognitive psychology. He had downhill skiers count numbers backwards beginning with 999 during a gliding section of a downhill run. It is impossible to think of anything else while counting the numbers. Thus, this technique prevented intrusive thoughts during that section. Because it is also quite easy to quit this task, it allows athletes to switch to task-relevant thoughts when approaching the next difficult section of the run.

3.4 Visualization

There are two applications of the visualization or mental imagery technique. One of these is used in

combination with relaxation techniques to induce a positive, relaxed state of mind. The athletes visualize themselves in a favorite relaxing scene. This can be a past successful performance, that will reinstall the positive affect, the feelings of competence, or in other words, the 'winning feeling' that was experienced then.

The second application of visualization is to mentally rehearse performance. In this regard, visualization can be used to assist the learning of a new technique, continue practice during a period of injury, or to prepare for a competition. Neurophysiological data shows that the visualization of a physical activity is not restricted to brain processes, but actually involves small movements of 'neuromuscular activity' of the muscles that would be used if the activity was actually performed (MacKay 1981). The mental rehearsal of performance yields the best results, however, when visualization is used in combination with actual performance.

Eberspächer (1995) points out that ultimately athletes should be able to imagine themselves perfectly performing a task from an inner perspective, the perspective of the actor him- or herself (this is referred to as ideomotor training in a specific sense). However, it seems necessary to step by step develop and practice this inner perspective. Otherwise, an athlete may not have sufficient control over the images that come up. In visualization that is not systematically developed, unconscious fears may intrude, for example, in their 'naive' visualizations slalom skiers may see themselves missing a gate that they considered critical during the inspection of the run. This would, of course, not constitute a good preparation for the competition. To avoid this, Eberspächer (1995) suggests three steps during the acquisition of visualization skills:

(a) subvocal training: write down the course of performance and then recite this course in an inner dialogue;

(b) covered perception training: see oneself perfectly perform the activity from an outside perspective, like seeing oneself on a video (observer perspective);

(c) ideomotor training: visualize the course of performance from an inner perspective (actor perspective).

To obtain the best results, and prevent negative thoughts and fears from intruding, it is important to execute a relaxation technique before beginning with the visualization.

3.5 Goal-setting

Numerous studies have shown that goal-setting is a self-regulation technique that works (Locke and Latham 1984). Specifically, if individuals commit themselves to specific, hard, but still realistic goals, they show better performance than when they only have a 'do your best' orientation. This technique supports task motivation, but also helps to focus attention on the task.

It is often important to help athletes to make their goals realistic. This means that they learn to define their goals in terms of what they can actually achieve given the state of their condition and their skills. This involves differentiating goals into short-term goals (upcoming competition), mid-term goals (standing at half-time of the season), and long-term goals (standing at the end of the season, or what one ultimately wants to achieve). In this perspective, it is helpful to introduce the image of a ladder that one intends to climb up. Take one step after another (short-term goals). You cannot expect to get to the top (long-term goal) directly from the bottom. Stay positive and confident while climbing up. You may slip down from a step from time to time (which means: don't expect to win every time), but stay oriented toward your goal, and keep climbing up steadily.

4. Conditions for the Application of Psychological Skills Training

The mental side plays an important role in athletic performance. Athletes can learn to control the mental side through the acquisition of self-regulation techniques. It should be noted that there are other psychological factors that influence performance and can be addressed through psychological interventions. Among these are, for example, the communication skills (between coach and athletes, and among team members) and team spirit.

The focus of the present chapter was on how athletes can regulate themselves in their preparation for a competition (including technique acquisition) as well as during a competition. Numerous self-regulation techniques have been developed by sport psychologists. Only a small sample of them could be described here.

These self-regulation techniques are acquired through psychological skills training. It should be pointed out that the techniques cannot be learned overnight, as some athletes, coaches, and officials seem to assume. Continued training for some period of time is necessary before these techniques will actually start to work. The acquisition of the basic techniques requires at least six months of regular practice.

Self-regulation techniques in sport are most effective when they do not have to be activated consciously. They work most effectively when they have become an automatically activated element in an athlete's behavioral repertoire. Ultimately, these techniques work tacitly and smoothly without interrupting performance. They help to achieve the optimum state of flow (Csikszentmihalyi 1975). For best effectiveness, psychological skills training should be integrated into the normal training routines.

5. Future Perspectives

During the last two decades quite a number of mental skills training programs have been developed. They provide sets of self-regulation techniques that athletes can learn in order to improve their performance. Although these techniques seem to work, it has also become obvious for practitioners that training the skills is not enough. The athletes the sport psychology consultant deals with show large individual differences. They have different intellectual capacities, different attitudes and needs, as well as different knowledge. Based on the existing set of self-regulation techniques, mental skills training has to be tailor-made for the individual athletes with their highly specific problem situation. Furthermore, short-term programs are not sufficient. Psychological training should accompany or better become integrated into physical training routines from adolescence on. Personality development is essential and should become a major focus of sport psychological training programs. The self is the core of self-regulation. Without a conflict-free self it becomes difficult to self-regulate. An integrated self may lead to the aspired maximizing of performance and achievement and at the same time to responsibility, improved decision-making, and happiness.

See also: Action Theory: Psychological; Activity Theory: Psychological; Health: Self-regulation; Self-regulation in Adulthood; Self-regulation in Childhood; Sport and Exercise Psychology: Overview; Sport Psychology: Performance Enhancement; Sport Psychology: Training and Supervision; Sports as Expertise, Psychology of; Virtual Reality, Psychology of

Bibliography

Beckmann J 1993 Aufgaben und Probleme bei der psychologischen Betreuung der deutschen alpinen Skimannschaft bei den Olympischen Winterspielen 1992. *Leistungssport* **23**(1): 23–5 [*Responsibilities and difficulties of psychological coaching of the German Alpine Ski Team at 1992 Winter Olympics*]

Beckmann J, Kazen M 1994 Action and state orientation and the performance of top athletes. In: Kuhl J, Beckmann J (eds.) *Volition and Personality. Action Versus State Orientation*. Hogrefe and Huber Publishers, Seattle, WA, pp. 439–51

Beckmann J, Rolstad K 1997 Aktivierung, Selbstregulation und Leistung. Gibt es so etwas wie Übermotivation? *Sportwissenschaft* **27**: 23–37 [*Activation, self-regulation, and performance. Is there something like overmotivation?*]

Csikszentmihalyi M 1975 *Beyond Boredom and Anxiety*, 1st edn. Jossey-Bass, San Francisco

Dienstbier R A 1989 Arousal and physiological toughness. Implications for mental and physical health. *Psychological Review* **96**: 84–100

Eberspächer H 1995 *Mentale Trainingsformen in der Praxis*. Sportinform, München [*Forms of Mental Training in Practice*]

Gallwey W T 1976 *Inner Tennis. Playing the Game*, 1st edn. Random House, New York

Griffith C R 1926 *Psychology of Coaching*. Scribner's, New York

Kanfer F H 1970 Self-regulation: Research, issues, and speculations. In: Neuringer C, Michael J L (eds.) *Behavior Modification in Clinical Psychology*. Appleton-Century-Crofts, New York, pp. 178–220

Kuhl J, Beckmann J (eds.) 1994 *Volition and Personality. Action and State Orientation*. Hogrefe and Huber Publishers, Seattle, Göttingen

Locke E A, Latham G P 1984 *Goal-setting: A Motivational Technique that Works!* Prentice Hall, Englewood Cliffs, NJ

Loehr J E 1984 How to overcome stress and play at your peak all the time. *Tennis* (**March**): 66–76

MacKay D 1981 The problem of rehearsal or mental practice. *Journal of Motorbehavior* **13**: 274–85

Mahoney M J, Avener M 1977 Psychology of the elite athletes: An exploration study. *Cognitive Therapy and Research* **1**: 135–41

Mahoney M J, Gabriel T J, Perkus T S 1987 Psychological skills and exceptional athletic performances. *The Sport Psychologist* **1**: 181–99

Morris T, Bull S J 1991 *Mental Training in Sport: An Overview* (BASS Monograph). British Association of Sports Sciences and National Coaching Foundation, Leeds, UK

Neiss R 1988 Reconceptualizing arousal: Psychobiological states in motorperformance. *Psychological Bulletin* **103**: 345–66

Nideffer R M 1980 The role of attention in optimal athletic performance. In: Klavova P, Daniel J (eds.) *Coach, Athlete and the Sport Psychologist*. University of Toronto Press, Toronto, ON, pp. 92–112

Seiler R 1992 Performance enhancement. A psychological approach. *Sport Science Review* **1**(2): 29–45

Smith M 1998 *Mental Skills for the Artistic Sports*. Johnson Gorman Publishers, Alberta

Yerkes R M, Dodson J D 1908 The relation of strength of stimuli to rapidity of habit-formation. *Journal of Comparative and Neurological Psychology* **18**: 459–82

J. Beckmann

Standpoint Theory, in Science

According to standpoint theorists, there are social positions from which people can obtain privileged perspectives on social relations, because those who are oppressed are in a good position to understand the social relations that constrain them. Recent work in Science and Technology Studies has shown that scientific knowledge is irreducibly social (e.g., see *Science, Sociology of*). Social relations are embedded in scientific knowledge and in different ways form topics and resources across the sciences, from physics and mathematics to primatology and psychology. Therefore standpoint theory could be relevant even in the natural sciences.

1. Standpoint Theory's Central Argument

With respect to science, standpoint theory has been most comprehensively worked out by feminist researchers. For this reason this article focuses primarily

on feminist contexts. But the arguments and conclusions can be extended to other types of standpoints: there are contexts in which the standpoints of non-Europeans or, returning to the roots of the theory, proletarians, might provide privileged epistemic positions within the sciences.

The central arguments for privileged perspectives stem from Hegel's 'master–slave' dialectic, Marx's discussions of ideology, and particularly from Georg Lukács' discussion of the privileged standpoint of the proletariat. In his 1922 essay 'Reification and the Consciousness of the Proletariat' (in Lukács 1971), Lukács argues that the proletariat potentially has a privileged standpoint from which to understand the rationalized structure of commodities and capital. Under capitalism, different classes perceive rationalization in different ways. The bourgeois, says Lukács, tend to accept their position as natural, just and even necessary, because it was created by them; they can retain an illusion of subjectivity, of themselves as actors. Through attempts to break out of the rationalized structure, the proletariat, on the other hand, can understand the constraining nature of social relations, and can also question the naturalness or necessity of those relations. As a result they can see that the system that removes much of their subjectivity is a specific historical product. Therefore particular social locations have the potential to provide insights when people in those locations run up against arbitrary structures. In essence, then, standpoint theory is an account of the blinding power of ideology, and how it can be overcome (see *Ideology, Sociology of*). If its arguments are right, it is relevant to science to the extent that scientific knowledge participates in such ideologies.

Within feminist discussions, standpoint theory has been developed in divergent directions, for example by Nancy Hartsock (1983), Patricia Hill Collins (1986), Dorothy Smith (1987), Sandra Harding (1991), and Rosemary Hennessy (1993). As do the rationalizations of capitalism, gender reifies social relations. Women have a distinctive experience, in the experience of discrimination, from which to see gender relations. And by working against discrimination they come to understand the unfairness and lack of necessity of many things normally taken for granted.

2. *Criticisms and Elaborations*

Because it has been taken up by many scholars and activists, the meaning of standpoint theory can be confusing. Its central and most controversial claim is that a particular social position potentially provides a *privileged* perspective, not merely *another* perspective. Claims that styles of thought are correlated with gender (e.g., Keller 1985) are sometimes grouped with feminist standpoint theories, but they do not contain the same sort of general argument for privilege. Even setting aside positions that do not recognize privilege, we can still differentiate between two versions of standpoint theory as it relates to scientific knowledge. A strong reading of the arguments of standpoint theorists entails that the social location of oppressed groups is a definite and positive resource for increasing the objectivity of science. A weaker reading entails only that objectivity tends to increase as more and more diverse groups are represented in science. The difference can be seen in Sandra Harding's notion of 'strong objectivity' (Harding 1991) and Helen Longino's more procedural notion of objectivity (Longino 1990). (See also *Objectivity: Philosophical Aspects*.)

The theory is often taken to imply that there is one correct feminist perspective, which will eventually reveal the true nature of gender relations. Setting aside good questions about whether the idea of a single complete theory is coherent, the claim can be challenged as overly essentialist. Essentialist assumptions of commonality in women's experiences or situations have been discredited by feminist researchers steadily finding diversity where once was seen uniformity. However, there is nothing necessarily essentialist about standpoint theory. Harding's formulation, based on 'thinking from women's lives' (Harding 1991), suggests that diversity can be built into the theory. It is not women's experiences alone that standpoint theory privileges, but experiences of working against discrimination, including in the production of knowledge. And there is no implied essentialism about women or their experiences when standpoint theory is framed, as it often is, in negative terms: interests and positions tend to blind people to features of their worlds. The call to think from women's lives does not exclude the possibility that men can acquire and develop feminist standpoints. By acting to oppose sexism, and by understanding obstacles faced by women, men can gain experience and knowledge of the social structures that maintain gender as oppressive.

Claims of privilege are limited by the fact that the oppressed typically have reduced resources with which to study their environments. Thus, even if the central arguments for standpoints are correct, they do not provide any cut-and-dried formula for correlating standpoints with more objective knowledge. Theorists representing oppressed groups may have better knowledge of the relevant social structures, but then again they may not. Thus standpoint theory's arguments are best given the weaker reading mentioned above.

The standpoint theory that emerges from these criticisms is modest and pragmatic. Without the determinism sometimes associated with standpoints—without the claim that, say, women necessarily understand some phenomena better than do men—standpoint theory is primarily an argument for the representation of diverse groups in the sciences. Those groups will be apt to take their research in

different directions, to discover different phenomena and see different strengths and weaknesses in arguments. Multiple standpoints can contribute to both the breadth and solidity of scientific thinking. But the theory does not claim that some people are simply to be believed because they command cognitively superior standpoints. The metaphor of 'privilege' is cashed out in terms of possibilities for understanding: genuinely inclusive scientific communities are likely to produce more balanced knowledge, being attuned to a broader array of social facts.

3. *The Domains of Standpoints in Science*

There are a number of contexts in the sciences for which even a cautious standpoint theory can be relevant. Three examples may be offered of broad contexts in which feminist and other standpoints can be applicable. The feminist examples easily could be modified for other standpoints.

3.1 *Gender and its Construction*

Social relations in themselves are topics of inquiry in the humanities, social, medical, and biological sciences. Research has shown ways in which law, literature, divisions of labor, and many other cultural products participate in the construction of important social categories. Many of these findings are due to the influx of feminist, Afro-American, post-colonial, and other critical theorists into academic social sciences and humanities departments, which has disrupted established patterns of thinking and added to various disciplines' understanding of their traditional as well as new subject matters.

Studies of sex differences, classifications of diseases, behavior patterns, and explanations of all of these in terms of evolutionary history also participate in the construction of the social world, for example in the construction of gender. Feminist biologists have found areas in which sexist ideology affects research, and have seen how scientists can reproduce, explain, and amplify social relations that they see in society (for a survey, see Schiebinger 1999). Standpoint theorists would claim that people who understand everyday social biases as the product of cultural formations are in a good position to see attempts to reify those formations and naturalize them, to see, for example, attempts to justify sexism by characterizing it as part of human biological heritage.

3.2 *Problem Choices*

Ideology affects the topics that are studied. For example, social scientists have looked to the structure of the household—usually taken for granted as a relatively fixed and uncontroversial unit—and have found conflicts, differences in interest, and social structures that had remained hidden when inquiry focused on the male world alone. Problem choices like these are common in the humanities, social sciences, and biological sciences, all of which self-consciously study social relations, though presumably such choices are not directly relevant in the physical sciences.

However, some areas of research may directly affect women, because the knowledge they produce can be used to reduce or exacerbate inequities (see *Gender and Technology*). Every area of inquiry has at least some instrumental uses, and many forms of applied knowledge have some relevance to the gendered structure of society. Topics within so-called 'pure' physics are connected to military uses, and are funded precisely because of those uses (see *Science, Technology, and the Military*). Women tend not to participate in benefits coming from the high status of military prowess, yet suffer many of the consequences of war. A feminist standpoint, then, is more likely to recognize and see as important the connections of certain areas of research to their consequences.

3.3 *Social Structures within Disciplines*

To varying degrees, sexism is a normal part of the working environment in all of the sciences. The allocation of funds, the structure of the tenure system, and old-boys networks can work against women. To recognize and study sexism within the sciences, however, is not just a specialized project of feminist sociology of science. To be a scientist means having knowledge (often tacit) of the norms and structures of one's discipline (see *Norms in Science*). For example, to convince colleagues to read an article, it is as important to know about form and rhetoric as it is to know something about 'subject matter.' Therefore, if feminist standpoint theory is right, feminist scientists will be in a better position than men to notice sexist aspects of their own discipline, aspects of their relations with other members of the field that perpetuate oppression.

4. *Conclusions*

Science and technology studies have pointed to the local and contextual origins of scientific knowledge (e.g., Knorr-Cetina 1981). A consequence of those origins is that scientific knowledge, even when it is deemed universal, may carry with it parochial assumptions about what is natural. As a result, science taken up outside its contexts of production needs to be corrected to take account of its own social assumptions and the knowledge of nonscientists (e.g., Wynne 1996). In this type of insight we might see the value of standpoint theory, showing how social location can be a resource for decontextualizing scientific knowledge.

Standpoint theory's arguments establish some grounds for privileged perspectives: people for whom social constraints are oppressive can more easily understand those constraints than can others. Recent studies of science and technology show that those grounds can be relevant in all branches of science: science is a thoroughly social process, and perspectives that tend to expose some of the effects of social structure are thus valuable throughout the sciences. In particular, feminist scholars have argued that because scientific communities have lacked diversity, they have typically lacked some of the resources to better enable them to see aspects of sexism and androcentrism in scientific work. Definite social locations, then, are relevant across the disciplines, though the range of particular situations in which they are relevant varies greatly in different contexts, sometimes affecting problem choices or norms of behavior, and sometimes directly affecting substantive claims. Therefore standpoint theory, together with the recognition of the social character of knowledge, shows that to increase objectivity, communities of research and inquiry should be diverse, representative, and democratic.

Bibliography

Collins P H 1986 Learning from the outsider within: The sociological significance of black feminist thought. *Social Problems* **33**: 14–32

Harding S 1991 *Whose Science? Whose Knowledge?: Thinking from Women's Lives*. Cornell University Press, Ithaca, NY

Hartsock N C M 1983 The feminist standpoint: Developing a ground for a specifically feminist historical materialism. In: Harding S, Hintikka M (eds.) *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology and Philosophy of Science*. Reidel, Dordrecht, The Netherlands

Hennessy R 1993 Women's lives/feminist knowledge: Feminist standpoint as ideology critique. *Hypatia* **8**: 14–34

Keller E F 1985 *Reflections on Gender and Science*. Yale University Press, New Haven, CT

Knorr-Cetina K 1981 *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Pergamon, Oxford, UK

Longino H 1990 *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, NJ

Lukács G 1971 *History and Class Consciousness; Studies in Marxist Dialectics*. MIT Press, Cambridge, MA

Schiebinger L 1999 *Has Feminism Changed Science?* Harvard University Press, Cambridge, MA

Smith D 1987 *The Everyday World as Problematic: A Feminist Sociology*. Northeastern University Press, Boston

Tuana N (ed.) 1989 *Feminism and Science*. Indiana University Press, Bloomington, IN

Wynne B 1996 May the sheep safely graze: A reflexive view of the expert–lay knowledge divide. In: Lash S, Szerszynski B, Wynne B (eds.) *Risk, Environment and Modernity*. Sage, London

S. Sismondo

State and Local Taxation

State and local taxation comprises those taxes that are collected at the subfederal government levels in order to finance state and local public services, assigning discretion in the determination of rates and bases of these taxes to subfederal governments. Consider a nation with different layers of government that are only serving administrative purposes. Each government level only distributes the public services to and collects the taxes from that nation's citizens upon which they agreed at the national level. In such a country, no state or local taxation is necessary since the level of public goods and services is decided upon uniformly for the whole country. This leaves citizens in some local jurisdictions, which prefer more public services than the national average, as much dissatisfied as citizens in local jurisdictions, which prefer less. It is possible to provide public services, the benefits of which are limited geographically and which are not national in scope, at different levels in different subfederal jurisdictions. Enabling citizens to consume publicly provided goods and services at different levels according to their preferences makes everyone better off without making someone worse off. In order to let people, who want to consume more, pay a higher tax price for those state and local services, taxes must be differentiated accordingly between states and local jurisdictions. In such a world with a differentiated supply of publicly provided goods, each individual can reside in a jurisdiction where a certain level of public services is provided to adequate tax prices. The art of state and local taxation is the assignment of different kinds of taxes to government levels such that an invisible hand properly guides a nation in an ideal world to an optimal multiunit fiscal system. This is not necessarily the case in the real world and some answers exist why fiscal systems are not optimal.

1. The Rationale of Tax Assignment to the State and Local Levels

The basic idea of a power to tax for states and local jurisdictions is that their public goods and services should be financed by those citizens benefiting from them and deciding on their quantity and quality. In this case, individuals choose their place of residence according to their preferences and their dislike for congestion. A set of fiscal jurisdictions consisting of equal taste communities then results from a process of 'voting by feet' (Tiebout 1956). Olson (1969) calls the basic condition for such an outcome the 'principle of fiscal equivalence.' It requires assignment to which public goods and services should be provided at which level of government. The guideline for this decision is the geographical incidence of those benefits and the costs borne by citizens in order to finance them. As some public goods and services reasonably cannot be

provided at the local level, some kinds of taxes may not be assigned to the local level for good reasons, too.

There are numerous reasons why the proper assignment of tasks and the power to tax to the subfederal levels are crucial to avoid suboptimal outcomes. For example, take the Scala in Milan, the famous Italian opera house. As it is partly financed by user charges and partly by subsidies paid from general taxes of the city, it will be congested because residents living in the suburban towns are able to visit it paying only the user charges and thus a lower price than the city's citizens. Nonresidents receive some of the benefits from the public service without paying an adequate tax price. This spillover is one of the externalities that Gordon (1983) summarized in his treatment of an optimal tax system for subfederal jurisdictions. Another externality exists if nonresidents pay some of the taxes of state or local jurisdictions. Taxes are exported to nonresidents and state and local governments thus have an incentive to provide public services at a higher level than their residents prefer. An example of tax exporting can be found in the taxation of tourist trade or of natural resources in some US states. For the year 1962, McLure (1967) estimated for the US states that overall tax export rates were between 19 and 28 percent in the short run. In the long run, only 10 states exported more than 24 percent of their taxes or less than 16 percent. The estimates are corroborated in a more recent study by Metcalf (1993). Mieszkowski (1983) estimated the 'total waste' in Alaska in the beginning of the 1980s that accrued from excessive public revenues and expenditures that were possible because of high natural resource taxes to about $500 million or 13 percent of estimated petroleum revenues. Tax exporting is not limited to commodity taxes or natural resource taxes. It may also occur if individuals, firms, or capital are mobile.

If factor mobility is allowed for, the problems of subfederal taxation are even enhanced. If a state or community levies a higher personal or corporate income tax than its neighboring jurisdictions, its mobile citizens or firms emigrate (or move their capital there) in order to enjoy a lower tax burden. Such a situation is characterized as tax competition. In doing so, mobile factors reduce the tax burden of residents and firms in the state or community they move to and increase the tax burden in those jurisdictions they leave. These changes in tax burdens usually are not considered by public authorities in these jurisdictions when deciding on the level of public goods and services such that these are provided at a lower than 'optimal' level. Since there is a local loss from taxation that does not correspond to a social loss, the cost of public services is overstated and jurisdictions will tend to underspend on state and local public services. Buchanan and Goetz (1972) call this effect a fiscal externality.

The effects of tax competition among states and local jurisdictions are similar to those of international tax competition and they vary according to different degrees of mobility of individuals and firms, to the number of mobile factors, to the number of jurisdictions involved in that competition, on the size of the jurisdictions involved, and to the number of tax instruments available to tax authorities. The assessment of the usefulness of these effects becomes rather complicated if competition using tax instruments is complemented by competition using public expenditures, like subsidies or specific infrastructural services, but also tax expenditures like tax holidays to attract mobile production factors.

The impact of tax competition on the possibilities of state and local governments to redistribute income by tax-transfer schemes is relatively clear-cut. Stigler (1957) already stated that 'redistribution is intrinsically a national policy.' Suppose that a region adopts a progressive income tax program designed to achieve a significantly more egalitarian distribution of income than exists in neighboring jurisdictions. If rich and poor households are mobile, such a program would create strong incentives for the wealthy to move out to neighboring jurisdictions. In this case, local redistribution induces sorting of the population, with the richest households residing in the communities that redistribute the least by income taxes. A more equal distribution of income would result, but it would be caused largely by an outflow of the rich and a consequent fall in the level of per capita income in the jurisdiction under consideration (Oates 1972). At the central level, this problem does not occur, because mobility is reduced, the larger a jurisdiction.

There are not many theoretical arguments against this line of reasoning. Most of them rely on imperfect mobility of individuals. But the more mobile people become, the better this mechanism should work, the less possible should state and local redistribution by a progressive income tax be. However, according to Buchanan (1975), all individuals have an incentive at the constitutional stage to agree to income redistribution because they are fundamentally uncertain about their future positions in income, health, and employment. At the postconstitutional level the rich might agree to income redistribution by the government because they, first, are interested in a public insurance scheme against fundamental privately uninsurable risks for themselves and their children, and, second, against exploitation by the majority of the poor residents and increasing crime rates. Particularly the second reason is important for the rich: they pay a premium for obtaining social peace. Janeba and Raff (1997) argue that individuals will agree to decentralized income redistribution because it is a credible commitment of the poor majority to the rich minority that the latter is not exploited at the postconstitutional stage. The rich allow for the agreed extent of redistribution because a constitutional competence for income redistribution at the federal level leads to a larger public sector with more income redistribution.

From this perspective, decentralized redistribution might be possible and even more efficient since redistribution is closer to citizens' preferences.

2. Guidelines for a Reasonable Tax Assignment

This reasoning leads to some guidelines for tax assignment of different taxes to the state and local levels. The bottom line is that state and local jurisdictions should finance the services they provide by charges or quasicharges to the beneficiaries (Musgrave 1983) or they should use those tax bases which have low interregional mobility (Wildasin 1980). Land and natural resource taxes and to a lesser degree also real estate taxes are thus especially suited for purpose of state or local taxation. A property tax, for example, is a good candidate for being assigned to the local level at least as far as residential property instead of business property is taxed. A property tax is basically a tax on land and construction built on it and it can thus be considered to be a tax on a relatively immobile tax base.

Consumption taxes are inappropriate at the local level because the interjurisdictional mobility leads to cross-border shopping or tax exporting, but they become suitable for the state level if the region is large enough. However, the technological development, in particular Internet shopping, reduces the possibility for state governments to levy a tax upon consumption. If individuals, firms, and financial capital are fully mobile a comprehensive income or profits tax should not be assigned to the state and local levels because it is necessary for such a taxation to be optimal so that a jurisdiction should be able to tax all income of its residents in the other jurisdictions. This is certainly not a realistic case. Beyond this, taxation of wage income tends to reduce this tax base less than the taxation of capital income and it could thus be more easily taxed at the state or even local level unless tax authorities cooperate and exchange information about taxpayers' incomes. And even profits taxes can be levied at the state or local levels if investment is specific to the locality such that a firm cannot easily relocate. The latter is particularly the case for not-in-my-back-yard (NIMBY) firms like waste incinerators or nuclear power plants. These firms have difficulties finding a location and assigning local jurisdictions the power to tax their profits gives these communities an incentive to provide such firms with a location.

Following from this reasoning is another rule for tax assignment: Progressive taxation, designed to secure redistributional objectives, should be assigned to the central level. A progressive income tax should thus be a federal tax. Moreover, as the case of Alaska illustrates, a large natural resource tax base permits local jurisdictions to provide public services at a low tax price leading to an attraction of individuals and firms at an inefficient scale. On these grounds, a case for central taxation of natural resources may be made whereby the central natural resource tax may apply to an excess base only, while leaving the average base for subfederal use.

All in all, these basic considerations led Musgrave (1983) to present the following tax assignment to government levels: The local level should rely on property taxes and payroll taxes, the state level on a resident income tax, a consumption tax, and a natural resource tax, and the federal level on an integrated progressive income tax and a natural resource tax for large natural resource bases. All government levels should additionally rely on fees and user charges as far as public services are attributable to particular users or certain groups of users. This possibility is probably reduced from the local to the federal level. According to Musgrave, corporate source income ideally would be included in the integrated base of the federal income tax. These propositions become substantially more complicated, if interjurisdictional tax agreements providing for tax credits or deductibility are considered. Federal tax deductibility of state and local taxes, for example, may mainly serve taxpayers with higher income and is thus regressive.

3. Tax Assignment in Some OECD Countries

A first look at real tax systems in OECD countries quickly reveals that tax assignment is not necessarily Musgravian (Pola et al. 1996; Pommerehne and Ress 1996). The US seem to come closest to those rules of tax assignment. The US states are prohibited by the US constitution from levying taxes on imports or exports but otherwise they may adopt any type of tax that does not egregiously discriminate against citizens of other states. US states have a lot of freedom in defining their tax bases, in setting tax rates, and in offering tax abatements to induce businesses to locate within their state. Finally, the states grant tax authority to differing degrees to local jurisdictions, too. While there are only minimal restrictions on state power to tax in the US, the federal level nevertheless relies mainly on the individual income and corporate income taxes, the states on sales taxes and natural resource taxes, and the most important source of revenue at the US local level is the property tax. Although state tax revenue from individual income and corporate income taxes is nonnegligible, it is only about one fifth of the respective federal revenue. Such a favorable judgment about the US tax system is, however, only justified at a first glance. According to Feldstein and Metcalf (1987), for example, the federal tax deductibility of personal income tax liabilities led many states to rely more heavily on deductible personal income taxes than on other type of revenue.

Apart from the US, for example, in Canada and Australia and particularly in Europe, the assignment

of taxes to the state and local levels differs considerably from the Musgravian model and varies widely across countries. In Canada, personal and corporate income taxes are the most important revenue sources of the federal level. The most visible of taxes in Canada, the personal income tax, is, however, also the most important revenue source of the provinces. On the one hand, there are fairly significant income tax rate differences between provinces. On the other hand, arrangements between the Canadian federal and provincial governments on sharing the bases for the personal and corporate income taxes have a long history going back at least to the Second World War. There is also a large variation in sales tax rates from Newfoundland to Alberta as the second important revenue source of the provinces. Canadian local jurisdictions, like those in the US, mainly rely on property taxes.

In Australia, the federal government exclusively levies taxes on personal and company income, sales taxes, and (nearly exclusively) excise taxes. The bulk of a state's tax revenues are derived from payroll taxes, stamp duties, and associated taxes on financial transactions, and so called franchise (i.e., business license) fees on alcohol, tobacco, and petroleum products that are not so much differing from excise taxes as well as lots of smaller revenue sources. Local governments, on the other hand, derive their revenues almost entirely from property taxes. Australian states also obtain revenue from the central government on the basis of a revenue sharing arrangement.

The Australian tax system is not as different from the German and Austrian system of state and local taxation. In particular Germany has established an extensive revenue sharing arrangement of personal and corporate income and value added taxes as the most important taxes accounting for about three quarters of total tax revenue. These are really shared taxes in the sense that German state governments decide on them together with the federal government. The federal government has at its own discretion only the gasoline tax and a surcharge on income taxes while the states have no power to tax. Local government relies on property taxes and local business taxes.

The less federalist European countries usually have taxing authority at the local level. In the UK, however, there is not only no regional or state level of government, but local authorities basically have not much power to tax, only some power in determining tax rates of the local property tax on real property. The formerly existing local business tax in the UK was abolished by the 'Thatcher' administration which cut the contributions of local taxes to local revenue sharply after the unsuccessful attempt to introduce the poll tax in the 1980s. There have been some changes in the UK at the end of the 1990's to establish regional authorities in Scotland, Wales, and Northern Ireland providing at least the Scottish government with some power to tax. The UK thus appears to be following the trend in Europe to a more decentralized government structure with a power to tax for subfederal governments as well.

Since the 1970s, systems of regional government have been given an enhanced role in taxation in France, Italy, and Spain. In France, there are three levels of subnational government: the local level with about 36000 cities and communities, about 100 *départements* and 22 regions. Since 1986, the importance of these subnational governments has increased. All three levels have the possibility to levy a property tax on real property (land and construction built on it), a local business tax ('*taxe professionelle*') and something similar to a residents' income tax ('*taxe d'habitation*'). All three taxes are subject to central regulation, but nevertheless there is a nonnegligible variation in tax rates.

In Belgium, there is a movement towards fiscal federalism since it is agreed that the sales and property taxes will become exclusive taxes for the regional and local levels. Apart from smaller taxes, the three Belgian regions additionally can levy a supplementary rate on the personal income tax which may be positive, a surcharge, or negative, a tax reduction. Although there is some regulation of the local and regional power to tax income, the variation of rates is considerable. In contrast to that, local governments in the Netherlands can only levy property taxes.

An interesting pattern of local taxation can be observed in the Nordic countries. Basically, Denmark, Finland, Norway, and Sweden are unitary countries with only two government levels. According to the studies in Rattsø (1998), local governments have considerable power to tax in Denmark and Finland, less so in Sweden, and only to a limited extent in Norway. In Denmark and Finland, local governments can rely on taxes on property and personal income (nearly) without restrictions for setting tax rates accompanied by a considerable variation. In Sweden, local taxation consists mainly of personal income taxes with less variation in rates due to the central regulation of local governments. In Norway, most tax revenue originates from income and wealth taxes shared with the central government giving local authorities taxing power only within a narrow band. Since 1979, all local governments apply the maximum rate. The property tax is not available to all local governments, but is restricted to urban areas and certain facilities (notably power plants). In addition, local property tax rates are also limited to a narrow band.

In Switzerland, states (cantons) have the basic power to tax personal and corporate income as well as capital. The local jurisdictions can levy a surcharge on cantonal direct taxes and raise their own property taxes. The central government relies mainly on indirect (proportional) taxes, the value added tax, and specific consumption taxes like the mineral oil tax. It can also rely on a source tax on income from interest. There is,

Table 1
State and local tax assignment in selected OECD countries

Country	States	Local authorities
USA	Income tax Sales tax Natural resource tax	Property tax Sales tax
Canada	Personal and corporate income tax Sales tax	Property tax
Australia	Payroll taxes Stamp duties Franchise fees	Property taxes
Germany	No taxing power	Property tax Business tax
UK	No state level	Property tax
France	Property tax Business tax Residents' income tax	Property tax Business tax Residents' income tax
Belgium	Personal income tax surcharge Property tax	Personal income tax surcharge Property tax
Switzerland	Personal and corporate income tax Inheritance tax Wealth tax	Personal and corporate income tax surcharge Property tax
Norway	No state level	Personal income tax Property tax
Denmark	No state level	Personal income tax Property tax
Musgravian 'optimal' assignment	Residents' income tax Consumption tax Natural resource tax	Property taxes Payroll taxes

finally, a small but highly progressive federal income tax, which, together with revenue from the source tax on interest income, amounted to 34 percent of total federal tax revenue in 1995, while the cantons and municipalities rely on income and property taxes to about 50 percent of their total revenue and 95 percent of their tax revenue. In addition, personal and corporate income taxes in Switzerland vary considerably between the cantons. Taking the value of the (weighted) average for Switzerland as 100, the index of the tax burden of personal income and property taxes varies from 55.6 in the canton of Zug to 135.8 in Fribourg in 1995, while the index of corporate income and capital taxes varied from 57.6 again in Zug to 145.6 in Nidwalden.

As the overview of these selected OECD countries in Table 1 indicates, there is no clear picture of which kinds of taxes are assigned to the state and local levels. Local jurisdictions regularly levy property taxes as well as fees and user charges. Some power to tax personal income, at least labor income, is attributed to state and local governments in the US, Canada, France, Belgium, the Nordic countries, and Switzerland. In some cases, for example, Germany, France and Switzerland, there exist even local business taxes in the discretion of subfederal jurisdictions. But on the whole, the assignment of taxes does not follow along Musgravian lines and there is a considerable variation in federal regulation of subfederal taxing power.

4. *The Political Economy of State and Local Taxation*

Given these differences in tax assignment to state and local authorities, the question arises which institutional and political factors explain the tax structure in different countries. An early contribution as to the choice of taxation in a political economy framework originates from Schneider and Pommerehne (1980). They develop a model of Australian tax and expenditure structure along with a model of selfish government behavior assuming that the re-election constraint is not always binding. In choosing among

tax instruments, a government in power is only urged to follow the preferences of a majority of voters if an election approaches and the re-election constraint is binding due to low government popularity. Between elections, the government can follow its ideological preferences. This modeling of government behavior proved to be clearly superior compared to the traditional models of government behavior in ex-post predictions of important variables (government spending, GDP, etc.). In contrast, Hettich and Winer (1999) assume that the governing party or coalition always maximizes the expected votes which by themselves depend on the consequences of policies for the economic welfare of voters. In making fiscal choices, political parties thus take into account economic, political, and administrative factors that determine the nature of electoral support and opposition associated with alternative ways of taxing. The authors present empirical evidence on the reliance on income taxation and income tax structure of the US states in 1985 and 1986 supporting their hypotheses.

Inman (1989) finds that redistributive concerns strongly influenced the decision of the 41 largest US cities between 1961 and 1986 to rely on property taxes, sales taxes, or user charges. The higher the demand of local voters for redistributive services, the less city governments rely on fees and user charges as well as sales taxes and the more on property taxes. Hunter and Nelson (1989) present evidence for the impact of interest groups on the local revenue structure. The higher the share of wealthy homeowners in local jurisdictions in Louisiana, the lower the share of property tax revenue and the higher the share of charges.

Besides vote-maximizing behavior of politicians and interest-group influence for redistribution in their favor, tax competition is a third factor determining the structure of taxation. However, the expected impact is ambiguous. On the one hand, tax competition for mobile capital and workers may lead to a convergence of tax rates to a low level. On the other hand, tax competition may lead to increases in tax rates if a state's geographic neighbors follow suit when it raises its tax rates. Such behavior might occur due to yardstick competition between states and local jurisdictions (Besley and Case 1995). Due to their incomplete political information, voters evaluate the fiscal performance of representatives in their jurisdiction by comparing it to the fiscal performance of neighboring jurisdictions. If a state's neighbor increases tax rates, this state has an increased possibility to raise its taxes as well. Arkansas State Senator Doug Brandon once said: 'We do everything everyone else does.'

Finally, the structure of state and local taxation, like federal taxation, is influenced by the degree of fiscal illusion among taxpayers. The more complex a tax system, the more incorrect voters' perceptions about their true tax burden and thus their tax resistance, the more easily representatives can increase taxes in order to finance new spending. Pommerehne and Schneider (1978) present evidence for Swiss local jurisdictions that the extent of fiscal illusion attributed to the complexity of the tax system is less pronounced, the more voters can determine taxes and spending directly in fiscal referenda. Moreover, Matsusaka (1995) finds that US states which allow for the use of initiatives in their constitutions rely more on fees and user charges than on broad based taxes increasing the visibility of contributions to the public sector and the equivalence between public services and tax prices.

5. *Concluding Remarks*

Although a comprehensive assessment is still too early, state and local taxation, appears to be the outcome of political decisions shaped by constitutional provisions and influenced by economic and administrative restrictions. Voters, representatives, bureaucrats, and interest groups deciding upon taxation under the decision making provisions of their constitutions and taking into account the mobility of tax bases and the accompanied tax competition and tax exporting determine to what extent state and local jurisdictions rely on property, income and sales taxes, or on user charges. The differences in these political, institutional, and economic factors explain why there are so many differences among countries in their reliance on tax sources at the state and local levels.

See also: Urban Government: Europe

Bibliography

Besley T, Case A 1995 Incumbent behavior: vote-seeking, tax-setting, and yardstick competition. *American Economic Review* **85**: 25–45

Buchanan J M 1975 *The Limits of Liberty*. University of Chicago Press, Chicago

Buchanan J M, Goetz C J 1972 Efficiency limits of fiscal mobility: an assessment of the Tiebout model. *Journal of Public Economics* **1**: 25–43

Feldstein M S, Metcalf G E 1987 The effect of federal tax deductibility on state and local taxes and spending. *Journal of Political Economy* **95**: 710–36

Gordon R H 1983 An optimal taxation approach to fiscal federalism. *Quarterly Journal of Economics* **98**: 567–86

Hettich W, Winer L 1999 *Democratic Choice and Taxation: A Theoretical and Empirical Analysis*. Cambridge University Press, Cambridge, UK

Hunter W J, Nelson M A 1989 Interest group demand for taxation. *Public Choice* **62**: 41–61

Inman R P 1989 The local decision to tax: evidence from large US cities. *Regional Science and Urban Economics* **19**: 455–91

Janeba E, Raff H 1997 Should the power to redistribute income be (de-) centralized? An example. *International Tax and Public Finance* **4**: 453–61

Matsusaka J G 1995 Fiscal effects of the voter initiative: Evidence from the last 30 years. *Journal of Political Economy* **103**: 587–623

McLure C E 1967 The interstate exporting of state and local taxes: estimates for 1962. *National Tax Journal* **20**: 49–77
Metcalf G E 1993 Tax exporting, federal deductibility, and state tax structure. *Journal of Policy Analysis and Management* **12**: 109–26
Mieszkowski P 1983 Energy policy, taxation of natural resources, and fiscal federalism. In: McLure C E (ed.) *Tax Assignment in Federal Countries*. Centre for Research on Federal Financial Relations, Canberra, Australia
Musgrave R A 1983 Who should tax, where and what? In: McLure C E (ed.) *Tax Assignment in Federal Countries*. Centre for Research on Federal Financial Relations, Canberra, Australia
Oates W E 1972 *Fiscal Federalism*. Harcourt/Brace/Jovanovich, New York
Olson M 1969 The principle of 'Fiscal Equivalence': the division of responsibilities among different levels of government. *American Economic Review*, *Papers and Proceedings* **59**: 479–87
Pola G, France G, Levaggi R (eds.) 1996 *Developments in Local Government Finance: Theory and Policy*. Edward Elgar, Cheltenham, UK
Pommerehne W W, Ress G (eds.) 1996 *Finanzverfassung im Spannungsfeld zwischen Zentralstaat und Gliedstaaten*. Nomos, Baden-Baden, Germany
Pommerehne W W, Schneider F 1978 Fiscal illusion, political institutions, and local public spending. *Kyklos* **31**: 381–408
Rattsø J (ed.) 1998 *Fiscal Federalism and State-Local Finance: The Scandinavian Perspective*. E Elgar, Cheltenham, UK
Schneider F, Pommerehne W W 1980 Politico-economic interactions in Australia: Some empirical evidence. *Economic Record* **56**: 113–31
Stigler G J 1957 The tenable range of functions of local government. In: US Congress, Joint Economic Committee (ed.) *Federal Expenditure Policy for Economic Growth and Stability*. US Government Printing Office, Washington, DC
Tiebout C M 1956 A pure theory of local expenditures. *Journal of Political Economy* **64**: 416–24
Wildasin D E 1980 Locational efficiency in a federal system. *Regional Science and Urban Economics* **10**: 453–71

L. P. Feld and F. Schneider

State and Society

The conceptual couple 'state/(civil) society' is a product of the protracted effort through which, starting with the Enlightment, a few generations of European intellectuals sought to characterize the advance of modernization. The juxtaposition of 'state' to '(civil) society' emphasizes the growing differentiation between, on the one hand, the structures and events directly connected with public authority, with the constitution and management of the polity, and on the other the structures and events which individuals generate as they pursue their private interests, express their emotions, acquire and exercise their capacity for judgment. Some institutional outcomes of this process can be still be recognized, according to many authors, in the shape and dynamics of contemporary societies.

1. *An Unsettled Distinction*

Speaking of a 'juxtaposition' between two concepts suggests that they lie on the same plane and are unproblematically complementary to one another. But this is not the only relation in which those concepts have stood to one another. In an earlier relation, the two concepts, instead, were closely overlapped. In a philosophical and juridical vocabulary that lasts between Aristotle and Kant, the expression 'civil society' designates a condition in which humans have left the state of nature by bringing about, and subjecting themselves to, political authority; one wrote of *societas civilis SIVE res publica*. Here, what made a society civil was the very fact of its being policed and governed.

Hegel (1953) breaks with this fusion of the two concepts in the *Philosophy of Right* of 1821. The civil society is understood as the set of individuals engaged in the competitive pursuit of their private interests (chiefly economic). The state is understood as an institutional site where identified collective interests—such as the increase in the state's security, or the maintenance of the public order—are pursued through expressly political activities.

But this juxtaposition is loaded with tensions, and Hegel presents conceptual entities—the family, the corporation, or the 'police' (which here means: the administration)—intended to mediate those tensions. Further, the juxtaposition may be perceived as a contrast between state and society, leading thinkers to take sides with the one or the other. Hegel himself opts for the state, but already Lorenz von Stein, who places the paired concepts state/society at the center of his main work, resolutely opts for society.

In a sense, so does Karl Marx, who like Hegel views civil society as the locus of the formation and confrontation of private interests of an economic nature. However, he endows it with a powerful, self-sustaining dynamic, grounded on means and relations of production, with respect to which the state itself loses the creative, autonomous power to drive historical development Hegel attributed to it.

Even in the course of the twentieth century, some authors prioritize one of two concepts and subsume under it the other. Oliver Williamson's (1975) theory of markets and hierarchies, for instance, posits economic exchange as the primordial social relation, and derives from 'market failures' the emergence of hierarchical, command-based relations, and thus of organizations. Implicitly, also in this view society, conceived in the first instance as the locus of market relations, precedes and in a sense justifies the state (if one sees in it a particularly significant organization). This priority of society, naturally, imposes tight

limitations on what the state can and should do (see Karlson 1993).

Such arguments imply that state/society is not a proper dichotomy, for one concept can be made to encompass and posit the other instead of accepting it as its peer. But there are also arguments to the effect not that one concept is sufficient, rather that more than two are needed.

Thus, one may object to the Hegelian/Marxian reduction of the civil society to the political economy or the market, and suggest for it an alternative referent. A model of a social realm focused on intellectual pursuits, and constituted and regulated without political mediation and power relations, can be found in the eighteenth century *Imperium Litterarum*, the international cosmopolis of *savants* and literati. Or one may seek to comprise within civil society the phenomena of culture and religion, or the institutions concerned with play, emotions, aesthetic expression, and other inter-individual relations not reducible to exchange. The market itself may either be made to adjust itself to the presence, under the conceptual umbrella of 'civil society,' of these new and rather different referents, or may be relocated to a third realm all of its own, as if to acknowledge its peculiarly powerful and demanding dynamic.

For instance, according to Cohen and Arato (1992), in advanced societies there are three fundamental institutional complexes—the state, the market, and civil society. The latter is conceived primarily as the site not of competition between private interests, rather of the formation of collective identities, of solidarity, of the formation and transformation of institutions by means of social movements. In this perspective, *tertium datur*. *Politics and Markets* (Lindblom 1977) do not exhaust the social space.

2. *Historical Roots of the Distinction*

One reason why, at the purely conceptual level, the relations between the two terms are so unsettled, may be that 'state' and '(civil) society' should not be conceived as universal categories. Whether we conceive them as juxtaposed or counterposed to one another, both expressions refer to historically unique ways of establishing and regulating both political and economic processes (and sociocultural ones, if a tripartition is accepted as more plausible).

The state/society distinction presupposes, and profoundly modifies, arrangements which had emerged in various European countries in the late Middle Ages, and which are often characterized by the German expression *Ständestaat* (literally, 'state of the estates'). But the expression is potentially misleading, for it may lead one to think of its referent as a 'state' in the modern sense of the term.

To avoid this misunderstanding the Austrian historian Otto Brunner (1968) has proposed replacing that expression with that of *ständische Gesellschaft*. By this is meant a context where constituted estates, through arrangements partly traditional partly agreed, take an active part in the government of a territory. However they do so chiefly in the exercise of their own privileges, not on a mandate from the central government; and in those privileges political aspects are tightly fused with both economic ones and with status ones, relating to honor and rank. Their activities of government are differentiated, but not as articulations of an overriding political function, rather on the basis of each estate's claims and responsibilities.

Furthermore, typically the component units of each estate are not individuals, but collective entities with a pronounced hierarchical structure. The dominant unit is the 'house,' understood as the seat and patrimony of an aristocratic dynasty. 'The lord of the house (*Hausherr*), at the same time that he performed the function, which today we consider private in nature, of head of the family, also exercised an authentic political power (jurisdictional, administrative, representative) over the members of the house subjected to him,' which tended to include the bulk of the local rural population (Schiera 1983). Furthermore, each 'house' tended to be also a fortified place, an assemblage of autonomous military resources, which the members of the family (again, including the rural dependants) would employ at the behest and under the command of the lord. In that capacity the house normally took part in the ruler's military undertakings; but equally could also protect the house's rights, arms in hand, from encroachments by other houses, including the ruler's own.

Finally, the typical house constituted an *oikos*, which is/was the seat of a complex plurality of activities intended to make it economically self-sufficient, as much as possible without recourse to exchange. The monitoring and organization of those activities (including, critically, the periodic delivery to the lord's household of unpaid produce and labor by the rural dependants) engaged the lord's faculties of jurisdiction and coercion.

3. *The Crisis of the* staendische Gesellschaft

The 'decline and fall' of these arrangements was caused by a critical aspect of state formation, the gathering at the center of a country of the faculties and facilities of organized violence, culminating, according to Max Weber's well known formula, in a monopoly of legitimate violence. Exactly because the aristocratic houses originally possessed an autonomous military capacity, that process, when it does not dissolve them outright, turns them into subordinate components of an unchallengeable public order, centered on the sovereign, their councils, standing army, and fiscal, administrative and judicial apparatus. The political autonomy of the houses, and thus of the estates they

collectively constitute, is lost. The sovereign's superiority, while it does not suppress the houses' internal hierarchical arrangements (including, of course, 'patriarchy') relativizes them, flattens them. Seen from the summit of sovereign power even the most exalted members of the realm begin to appear relatively similar to one another, relatively powerless.

At the same time that centralized political activities become the prerogative of a sovereign power that performs them in a more unitary and rational manner, economic activities become decentralized. An increasing number of subjects decide what goods and services to produce, in what ways, by interacting according to their own interests and capacities. The provision for the needs of the individuals depends increasingly on their capacity to enter into traffics with one another on the market.

In the countryside, commercialization reduces the part played in the totality of the economic process both by the aristocratic patrimonies and by the traditional rural units constituted by villages, which had shared the aristocrats' preference for self-sufficiency and their hostility to the market. It also diminishes the significance of city-based productive units regulated by guilds, which typically contrasted the market's tendency toward innovation and the free combination and recombination of resources.

To sum up: while state-building dissolves the *staendische Gesellschaft* by taking from it tasks, resources, and legitimacy pertaining to the political sphere, the advance of an economy based on traffics pushes into the market the units let loose by that dissolution. Against an expressly political organization with a strong vertical gradient, concerned with all relations grounded on *imperium*, there stands an open-ended set of increasingly numerous social units which, in the pursuit each of its own interest, establish with one another horizontal, or apparently horizontal relations.

4. *The Political Prerequisites of the Civil Society*

This can only take place, however, insofar as the process of state building reflects two peculiar principles: first, the political center takes on all expressly political concerns and resources but only those; second, it acknowledges the existence of private subjects who, while excluded from the political sphere, act autonomously in others. In both ways, political power expressly limits itself, through complex and delicate institutional arrangements: for instance, the progressive denial of political and juridical significance to the religious interests of individuals, or the acknowledgment and guarantee of private property and of contract.

Heinrich Popitz (1986) highlights the historical rarity of such arrangements: 'Only rarely has it been possible even just to pose the question of how to constrain institutionalized violence, in a systematic and effective manner. This has only taken place with the Greek *polis*, republican Rome, a few city-states, and the constitutional state of the modern era. The answers have been surprisingly similar: the principle of the primacy of the law and of the equality of all before the law (*isonomia*); the idea of setting boundaries to all legislation (basic rights), norms of competence (separation of powers, federalism), procedural norms (decisions to be taken by organs, their publicity, appeal to higher organs), norms concerning the distribution of political responsibilities (turn-taking, elections), public norms (freedom of opinion and of assembly).'

Within this small set of experiences of limited political power not all counterpose two realms we might call 'state' and '(civil) society.' For instance, in the Greek polis individuals exist chiefly as soldiers and citizens, and the private dimension of their life is little valued and expressed. Only vis-à-vis the modern constitutional state does (civil) society emerge, through the processes indicated above.

In many other civilizations the tendency of political power to aggrandize itself, to possess itself of all resources (beginning with land and labor power) suppresses all autonomy of the society, and may even reduce all individuals to subjects or slaves. In the modern West, by acknowledging and protecting private property in productive resources, the political center allows at least some subjects to stand against it as autonomous units. This, according to North and Thomas (1973), greatly favors the advance of the division of labor and the enlargement of the traffic economy. But it has also distinctive political effects, best perceived *a contrario* in the 'decline and fall' of empires which fail to acknowledge and protect property rights (Eisenstadt 1969).

The significance of private property has two implications. First, it justifies to an extent the Hegelian/Marxian understanding of civil society primarily as the site of traffics. Second, it entails that the civil society should not be envisaged as a 'flat' space, harboring no power relations; for private property, especially in the means of production, engenders a power difference between those who do and those who do not possess it. The state/society difference does not mean that power is institutionalized only in the former, but that the latter is structured (among other things) by power relations of its own, foremost among which are those established by property.

5. *State and Society in the Twentieth Century*

The complex interactions between the power forms that lie on each side of the divide shape, to a large extent, the dynamics of modern societies. In the twentieth century the prevailing trend could be called

'statism,' for it saw a considerable enlargement of the scope of state action. In totalitarian systems, this trend suppressed all institutional limitations upon that action, and threatened the existence itself of society as a separate realm. In other systems, it pushed back considerably the boundaries of that realm, but this in response also to demands originating from it. For instance, economically underprivileged strata sought to maximize the import of citizenship and to reduce that of their market position; many economically powerful groups sought the state's assistance in securing their profits and stabilizing or correcting the market itself, which for various reasons had lost its self-equilibrating capacity. A further component of the trend consisted in the strategies of groups (sections of the 'political class,' public employees, members of the so-called 'new professions') whose livelihood and whose opportunities for a greater sway on policy depended on an expansive definition of the state's role.

Whatever its protagonists, in non totalitarian countries that definition was much assisted by the existence of a 'public sphere' and by the institutionalization of adversary politics, which gave voice to broader social strata, and found its most recurrent theme in the respective scope of the state and the market. Political parties played the central role in this dynamic, sometimes opposing but more often supporting the 'statist' trend. This expressed itself, among other things, in the formation of bodies and agencies that sat astride the traditional divide between the public and the private realm—as with the Italian *parastato* or the British *quango*.

In the last decades of the twentieth century the trend was challenged by social forces expressing themselves politically through conservative and/or right-wing parties. These sought to at least slow down 'statism,' and proclaimed the superior rationality of the market, which in their discourse, but to an extent also in practice, reasserted itself as the chief, if not the only institutional site of the civil society. This late revisitation of the economy-centered, Hegelian/Marxian conception of the civil society acknowledges that in contemporary society also human potentialities and social processes—of a cultural, communicative, symbolic, apparent noneconomic nature—which other formulations treat as a *tertium* with respect to both state and market, are in fact targeted and controlled to a large extent by great concentrations of economic power. This is achieved through new material and social technologies, and subjected to the logic of accumulation, which is increasingly focused on informational and financial resources. Mostly those technologies are developed and operate transnationally, and thus increasingly escape the reach of the political competences and facilities vested in the states. This lends a new edge to the attempt to reduce, after decades of 'statism,' the significance of public and political action, and to enhance that of the market.

Also the end of the Cold War operated in that sense, for it reduced the necessity for strong political/military tutelage of the capitalist order. It also loosened the constraints of the Keynesian compact which the dominant strata had previously made with the lower orders and implemented largely through state action.

Both dominant trends—the explicit or implicit reduction of civil society to the market, and the slogan prevalent in the advanced countries at the end of the twentieth century: market, *si*, state, *no*—are potentially worrisome. The market itself remains the site of the formation and interaction between large economic forces, endowed with great power; and these are increasingly operating in a transnational space. In contrast the state—still wedded to the territory—can no longer perform even its most traditional functions, jurisdiction and police, let alone those functions of guidance and intervention in the economic realm it had increasingly claimed during most of the twentieth century.

These recent phenomena compromise the institutional tenability of the couple state/society nearly as much as it had been threatened, earlier, by the totalitarian state or by central economic planning. A seriously dis-empowered state cannot establish and monitor an authentic civil society complementary to itself. In this sense, the Aristotelian and medieval vision according to which only a politically organized society can truly call itself 'civil,' deserves to be revisited.

Bibliography

Brunner O 1968 *Neue Wege der Verfassungs- und Sozialgeschichte*. Vandenhoeck & Ruprecht, Göttingen, Germany

Cohen J L, Arato A 1992 *Civil Society and Political Theory*. MIT Press, Cambridge, MA

Eisenstadt S N 1969 *The Political Systems of Empires*. Free Press, New York

Hegel G W F 1953 *Philosophy of Right*. Clarendon Press, Oxford, UK

Karlson N 1993 *The State of State: An Inquiry Concerning the Role of Invisible Hands in Politics and Civil Society*. Almkvist & Wiksell, Stockholm, Sweden

Lindblom C E 1977 *Politics and Markets: The World's Political Economic Systems*. Basic Books, NY

Marx K 1904 *A Contribution to the Critique of Political Economy*. Kerr, Chicago, IL

North D C, Thomas R P 1973 *The Rise of the Western World: A New Economic History*. Cambridge University Press, London

Popitz H 1986 *Phänomene der Macht*. Mohr (Siebeck), Tübingen, Germany

Schiera P 1983 Società per ceti. In: Bobbio N, Matteucci N, Pasquino G (eds.) *Dizionario di Politica*. UTET, Turin, Italy

Williamson O E 1975 *Markets and Hierarchies. Analysis and Antitrust Implications: A Study in the Economics of Internal Organization*. Free Press, New York

G. Poggi

14964

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

State: Anthropological Aspects

The 'state' refers to a form of polity that links people and territory in 'societies' and thereby demarcates itself from other states. Although archeologists date states back to ancient Egypt and Mesopotamia (about 2500 BC), this article deals only with anthropological aspects of the modern state. It examines the implications of linking people to territory on subject formation, the administration of populations and their cultures, and on the practices of religiosity, kinship, and commerce as well as of anthropology itself.

1. *The Concept of Modern Statehood*

Modern statehood is a mode of political organization of 'society' in which the state is the primary agent charged with building the representation of society aware of itself and in harmony with itself. Modern law is the principal means by which the state attempts to create this harmony, an ideal state in which there is a coincidence between behavior and rules, and where the state can anticipate the conduct of its subjects (citizens) and develop normative frameworks (policies) to direct change (Nader 1990). States are only one of many rule-making institutions with administrative capacities. International and supranational bodies also increasingly make law, but states largely retain for themselves the power of legitimate enforcement. They remain the only institution in most parts of the contemporary world with an effective ideological and institutional apparatus that claims to represent the entire people within a particular territorial jurisdiction.

Statehood became the central principle asserted in international law in 1648, with the signing of the Treaty of Westphalia ending the Thirty-Year's War in North-Central Europe. Driven by industrial capitalism as a superior mode of production and accumulation, the state expanded its functions beyond the conduct of war, the administration of justice, and the regulation of commerce, to massive projects in the colonization of populations in the Americas, Asia, Africa, and the mid-East (Dirks 1993, Stoler 1995). Western European states, in particular, formulated projects of internal colonization and they began to usurp religious authority's control over the definitions of transcendence, the bodies of their subjects, and the institutionalization of their life courses.

Initially employed as a political program in the French Revolution and subsequently written into the United Nations Charter, the nation form grew out of transformed empires and tribes during the process of European state formation (Tilly 1990). At the same time, in a process of reciprocal copying, states in the Americas also emulated the nation form. This transformation of diverse peoples into unified nations appropriated local discourses of consanguinity and affinity and, in turn, reconstructed kinship. Unification was rarely accomplished without intermittent purges or other kinds of homogenizing processes. From the late fifteenth to the early seventeenth centuries, all Western European states engaged, with often-disastrous economic consequences, in variant forms of ethnoreligious 'cleansing' leading to population resettlement and the formation of victim groups.

In the twentieth century, national consolidations and the creation of new types of victim groups reached new heights, especially with the advent of fascist and racist ideologies originating in Europe but also employed elsewhere (Lefort 1986). The peace treaties concluding World War I introduced into Eastern and Southern Europe the two conditions of West European nation-states—homogeneity of population and rootedness in the soil. In Eastern Europe the implementation of these principles was largely elided (though tacitly encouraged) during the Cold War with the establishment of socialist states based on universal brotherhood and allied with the Soviet Union—a bloc that dissolved following the so-called 'velvet revolutions' of 1989. Western European states emerged weakened from World War II and unable to resist independence movements in their colonized states, primarily in Africa and Asia. These postcolonial states in turn nearly universally adopted the nation form to integrate diverse groups within territorial boundaries already drawn, arbitrarily, by European states. Nation-states have since become the mode by which all of the earth's polities are organized into formally comparable and sovereign governing units. Appropriation of the nation form effectively broadened the scope of the state to include the planned change of culture, variously defined, and the generation of forms of subjectivity in populations (Borneman 1992).

2. *Subject Formation through Linking Territory to People*

The 1933 Montevideo Convention defined the criteria for statehood as having: (a) a permanent population; (b) a defined territory; (c) government; and (d) the capacity to enter into relations with other states (Article 1, League of Nations Treaty Series No. 881)—with only the first two criteria consistently upheld. Between 1945 and 1999, the number of states recognized by international law as sovereign members of the United Nations increased from 51 to 189, with two nonmember states (the Holy See and Switzerland) also recognized. Defining statehood this way has numerous effects on subject formation. First, it produces pressures to unify ethnically mixed and territorially dispersed populations by centralizing legal and education systems and standardizing the life course. States uniformize by means of both bureau-

cratic technologies (the use of, e.g., interviews, surveys, and maps) and psychological appeals to the 'shared' symbolic universes they create (through the standardization of grammars and linguistic use, and the shaping of experience in ritual events and everyday routines). Second, it creates populations with mixed loyalties, settlers or migrants who live in a place among peoples separate from the place and peoples with whom they primarily identify. Third, it generates categories of automatic exclusion (e.g., resident stateless peoples, resident internal 'minorities,' nonresident external others), which are then used frequently to justify war against external or internal enemies. Fourth, the dependence of modern political representation on territorial attachment reinforces identifications with territory and the desire for statehood. These identifications exacerbate territorial conflicts and often lead to wars between potential antagonists who tend to receive international recognition only when their disagreements are articulated in terms of border violations.

Despite the ubiquity of the nation-state, different historical experiences of populations have produced a variety of ways of classifying states, some corresponding to intent of the ruler, others to the relation of ruler to ruled, and still others to their cultural form. In the twentieth century alone, classification of state forms include, for example, monarchy, oligarchy, theocracy, democracy, totalitarian, liberal, democratic, communist, fascist, socialist, authoritarian, welfare, theater, tributary, cephalic, bicephalic, romantic, satiric, colonial, postcolonial. In accounting for this variety of state forms and the historical conditions of their production, anthropologists have focused on the organization of public authority and forms of domestic control. The historian Foucault (1977, 1978) and the political scientist Anderson (1983) have inspired much recent scholarship.

The powers unleashed by science since the Enlightenment have generated what Foucault calls 'rituals of truth,' including disciplinary technologies that refigure space and the arts of governance charged with producing 'society' and 'culture' as objects of knowledge and identification. Moreover, the deployment of new procedures of power in the nineteenth century transformed state authority, from a sovereign's ability to 'take life or let live' (to order or require someone's death), to an investment in administering life by means of disciplining the body and regulating populations. States institutionalize particular cultural forms through projects of reform, education, development, incarceration, war, welfare, and family planning, along with the use of scientific techniques such as surveys, census, questionnaires, actuarial tables, economic forecasting, spatial mapping, case studies, and bureaucratic interviews. Anderson focuses on the historical conditions in which the 'imagined community' of the nation comes into being and achieves global currency, particularly in its relation to states following the collapse of empires. One of these conditions, print capitalism, enables advances in techniques of memory and collective narration through the proliferation of historicist myths and standardized accounts of experiences of national belonging across space and irrespective of local temporalities.

Because modern governing necessitates delineating a population and a territory, rulers deploy ethnographic assumptions about how to define a people. As well, through laws and policies states produce and circulate information about belonging (through families, schools, and the media), positively and negatively sanction prescriptive and proscriptive behaviors (about sexuality, gender, age, race, ethnicity, and work), and build libraries, monuments, and museums to display particular interpretations of the nation. Hence, modern states are distinctive in that they produce both laws and people; they not only administer and regulate already-existing cultural patterns but also produce new domains of objects and forms of subjectivity.

3. *The Discipline of Anthropology and the State*

Early anthropological work on the state, especially among archaeologists, was inspired above all by Friedrich Engels's study of the origin of the state and its relation to private property, Lewis Henry Morgan's evolutionary scheme, and Sir Henry Maine's theory of the progressive movement from status to contract in the development of state law (Service 1975). Without being fully conscious of it, fieldwork-based anthropologists, inspired by Bronislaw Malinowski and Franz Boas, initially participated in this expansion of the state by studying the generation, elimination, transformation, or preservation of indigenous peoples, tribes, peasants, and clans through the creation of encompassing, and simultaneously exclusionary, national forms of belonging. One of the major anthropological aspects of states has been, therefore, the deployment of anthropology: ways of defining and categorizing human populations and of systematizing knowledge about them (Ferguson 1990, Pemberton 1994). The discipline of anthropology has closely followed and defined its objects in dialogue with state projects, alternately supportive and oppositional: itemizing culture traits and social organizations of 'discrete' peoples, devising evolutionary schemes, tracing efforts at internal and external colonization, mapping the settling and resettling of territory and the transformation of property, and tracking diasporic populations, cultural diffusion, and global trade routes (Moore 1986).

By and large, however, anthropologists traditionally have focused on the production of the cultural or the social, or with forms of exchange among people

without states (Sahlins 1981). A preference for fieldwork in villages or small-scale communities also made it difficult for anthropologists to connect their subjects with states or the large-scale. Further, long-term observational fieldwork requires an intimate access that is difficult to obtain from rulers or those in the inner circles of power. Anthropologists have, therefore, often assumed an essential sociocultural opposition of the people to the state. This seemed credible, if not ethically imperative, when, as in the early twentieth century, the objects of study were so-called 'primitives' and 'indigenous peoples,' victimized and threatened with extinction, frequently through state-supported genocide. Or, alternatively, when the state was primarily a puppet of a colonial power. Nonetheless, the opposition is misleading, as society (*Gesellschaft*) vs. community (*Gemeinschaft*) originates as a project of the modern state.

Anthropology has in the meantime become a global discipline of alterity, no longer restricted to questions of origin or to particular peoples or places. Also, the relation of indigenous people to states has changed considerably, with many states now guaranteeing the 'preservation of indigenous cultures' and prioritizing their rights over those of more recent settlers. Moreover, the United Nations along with some nongovernmental organizations and a global tourist industry positively sanction states that protect and support what are called 'first peoples'—a movement that has reversed the trend toward extinction and strengthened claims of cultural authenticity or temporal priority of settlement. In any case, the notion of 'kin-based' or 'acephalic' polities as a pre-evolutionary stage to a 'state-based' polity, or of the replacement of status by contract, has become moot for ethnographic research as the entire planet is organized territorially by states unevenly integrated in global market exchanges, mutually regulated through legal contracts. The interaction of people with states is an unavoidable condition of modern life.

This interaction took three meta-narrative forms in the second half of the twentieth century: people living within the bipolar Cold War structure of socialist or communist states (planned economies, single Party rule, so-called 'Second World' of totalitarian states) vs. liberal states (free market economies, multiparty democracies, so-called 'First World') and a contested zone of people living in postcolonial states (mixed economies with diverse political forms liberated from empires, so-called 'Third World'). Until the late 1970s, anthropologists had little access to people within totalitarian states, and in other parts of the world they devoted much research to reconstructing the precolonial state. In the mid-1970s, a global trend toward democratization of authoritarian states began in southern Europe, was picked up in Latin America and parts of Asia in the 1980s, and spread to sub-Saharan Africa, Eastern Europe, and the Soviet Union in the late 1980s and 1990s, making the relation of these people to transforming states a subject of anthropological research. In effect, the space of the Second World has disappeared and the spaces of the Third and First Worlds are mixed, existing within each other, though in radically different proportions. The vast majority of states now call themselves democracies and are engaged in market reforms that open internal economic and cultural relations to global markets.

Of note at the beginning of the twenty-first century is the distinctive disjuncture between the global integration of financial markets and shifts in state strategies with respect to sovereignty over national economies, the proliferation of ethnic affirmation within states, and the denationalization of much military conflict (a process also driven by fear of the use of nuclear weapons). Whether states drive, resist, or submit to globalization, and hence thereby curtail or strengthen their sovereignty, varies greatly depending on size, wealth, location in the international system, and domestic legitimacy. Local cultural forms unable to resist pressures of cultural or economic globalization are nonetheless not homogenized, a fact apparent both in the local variations in 'society's' relation to the state, specifically to democratic form, and the continued diversity of everyday routines and rituals.

By the beginning of the twenty-first century, democracy has become the dominant form of internally organizing states, functioning as a mode of ruling, a device of power, and a strategy of self-representation. Furthermore, democracy requires systems of personal and public accountability, usually embodied in principles of 'the rule of law,' or it begins to degenerate into forms of despotism (Borneman 1997). As a mode of ruling, democracy means the introduction of regularized elections, the creation of ruling majorities and minorities, the institutionalization of an opposition, and the strengthening of principles of accountability that circumscribe the exercise of power. As a device of power, democracy is as dependent on the externalization or creation of negative others as on the internal dynamics of citizen formation. Practices of such externalization range from state-tolerated social discrimination to legal exclusion of refugees to war. As a strategy of self-representation, each democracy receives international recognition to the extent it is able to claim governance by 'rule of the people' and conformity to international norms and conventions.

Anthropologists charged with listening to others and often asked to give 'voice' to the people studied are involved in all three aspects of democracy. Hence, the discipline of anthropology unavoidably intervenes in local political struggles about defining 'the people.' Given the increasingly global nature of 'news' and accelerated information flows across state borders, local struggles about representation quickly enter inter- or supranational arenas, subjecting the state to pressure from authorities external to it and explicitly

drawing attention to the ways in which it exercises sovereignty. Such global flows also challenge state sovereignty by bringing to the world's attention local and subregional actors and events that might be in opposition to dominant groups within their own states.

The study of conflicts in democratizing states includes research on social and legal citizenship and on the relation of state policy to demands for representation and accountability. Along these lines, four basic contemporary conflicts about representation and states are observable: involving excluded people who want more social, economic, cultural, or legal inclusion into states in which they reside; diasporic people who want representation in both the place in which they reside and the place from which they are displaced; numerical minorities (usually represented as a 'culture' or a 'nation') who do not have their own state but want one; and citizens or social movements that demand more accountability from their rulers.

4. *Legitimacy and the Modern State*

Although the conception and functions of the modern state keep changing, its legitimacy is always based both on some notion of the 'consent' of its own citizens to rule and on the recognition of other states. This was not the case when rule was justified without recourse to 'the people,' in other words, rule based on coercion alone, principles of heredity or wealth, the ability to wage war and collect taxes, or the administration of justice. To obtain this consent and recognition, the state must usually go beyond Max Weber's famous dictum that its power rests on a monopoly on the legitimate use of violence.

The state does not, in fact, hold this monopoly in many parts of the world. And where the state has dissolved or weakened, the most vulnerable citizens tend to suffer the most from the lack of any protections. Indeed, today weak states are just as likely to lead to violent conflict as were strong imperial states in the last several centuries. In many places, the state's own abuse of this supposed monopoly—through genocide, strategic terrorism, the suppression of dissent, cynical uses of war—has traumatic effects on its citizens, leading to repeated and delayed suffering of past events. Irrespective of whether they initiated or were victims of the violence, states instrumentalize this trauma, not only by building monuments to the fallen soldiers but also though the construction of nationalist historiographies and support for a heritage industry that especially honors the idea of national sacrifice and organizes the process of mourning for losses after wars. So long as states continue to draw legitimacy from principles of territorial sovereignty and peoplehood, they will be motivated to generate, exacerbate, and institutionalize differences with neighbors or neighboring states. Anthropology is concerned less with the conduct of these conflicts and wars than with their effects on the organization of populations and communities (Tambiah 1992).

Also, the collapse of the Cold War's bipolar state system has resulted in the relatively unrestricted proliferation of arms, much of it driven by state-supported arms industries, and in a market for new and affordable technologies of death outside the control of states. Where the sources and agents of violence outside state control are increasing—a wide-ranging phenomena found, for example, in Colombia due to drug trafficking, in Sri Lanka and much of Africa due to civil war, and in the United States or Mexico due to poverty and class conflict—people employ private police forces.

Alternately, many states now lend their militaries for 'peace-keeping' efforts organized by the United Nations, and still others, such as states within the European Union, are relinquishing military sovereignty to continental or transcontinental military alliances. There is also a growing body of legal doctrine—and an international court established in 1999—holding states accountable to their own norms and to international covenants and agreements on issues such as human rights, environmental and maritime law, and commercial law. The state's sovereignty and legitimacy, then, is not granted as natural but conferred by its citizens and the recognition of other states and new international regimes.

The long history of linking the state to realization of the Idea of the people (Hegel), representation of the social (Durkheim), technical and bureaucratic rationality (Weber), or exploitation and class dominance (Marx), must now be supplemented by acknowledging that most states still resort to magical practices and cultivate religiosity in order to legitimate rule (Taussig 1997). State policies and edicts administered through 'law' often work through processes similar to magic: belief in the power of the utterance of the speaker, the performativity of speech, the linguistic re-articulation of group suffering and instrumentalization of memory, nationalistic narrativization of macro-political events, *ad hoc* rationalization of failures and successes. Even in the domain of criminal law in the most rational of states—Germany or Sweden, for example—most crimes are never solved and most criminals go free. Yet criminal law there still 'works'—that is, contributes to the legitimation of the state—largely through belief in the principles of justice that judicial inquiries, prosecutions, and trials efficaciously represent.

Likewise with respect to religion, states most successful in 'neutralizing' religion are those which recognize the significance of religiosity and incorporate religious traditions and symbolic forms into the secular rituals of political parties, elections, bureaucratic routines, and lifecourse *rites-de-passage* (Gauchet 1997, Herzfeld 1992). Such practices are

forms of group worship and they crystallize in patterns of national belonging when citizen and state narratives are brought into harmony. The struggle between the great world religions and the state to occupy and shape the symbolic forms of everyday life has not quieted, partly due to the development and ubiquity of new technologies of communication that make it possible for both instances to react independently and quickly to everyday events. Nonetheless, there has been a dispersion of the power over the control of religious expression, and the state is one of many benefactors from this process. In the last quarter of the twentieth century in most places of what is known as the First World, entertainment—especially involving sport and music—emerged as the field dominating both the management of experiences of transcendence and of people's affective attachments and identifications. State rituals—especially electoral campaigns, legislative debates, memorialization and mourning ceremonies, and political speeches—have, therefore, increasingly incorporated models and techniques of authority drawn from entertainment. This leads to a fusion of actors ('personalities') in the two fields and an incorporation of magic and infusion of 'charisma' in the rationalized arena of state activity.

With the end of the Cold War, First World states increasingly structure their relations to new and Third World states by promoting NGOs ('nongovernmental organizations'). Acting in a negative symbiosis with the state, such NGOs also incorporate magic into their missionary projects of rational reform and economic and political development. In much of the world they have replaced organized religion as competitors to the state in providing care (welfare) and promoting both local and global accountability in the exercise of power. Often, however, they serve primarily as instruments for the transfer of wealth or power between supranational elites. Notwithstanding liberal mistrust of the modern state, reaffirmed by the horrors of totalitarian experiments, the nation-state form still provides the possibility for local control of people's own destinies in a manner impossible for prenational communities. Stateless people are indeed the most defenseless actors in the contemporary world.

Hence, despite the wars, violent revolutions, and massive repressions of the nation-state, as the sociologist Lepsius (1988) argues, this form of polity has offered particular advantages to West European societies, including the institutionalization of peaceful conflict-solving through the rule of law, guarantees of individual freedom, the organization of interests through parliamentary democracies, and the integration of national economic development in the world economy. In each of these measures, autocratic and totalitarian states elsewhere have been at a permanent disadvantage, incapable of recognizing the subjects they seek to control for whom they are, and unable to organize power except through coercion. Such states destroy sociality and build false representations of the people, thereby creating a basis for the illusion of society against state.

See also: Nation: Sociological Aspects; Postcoloniality; Socialist Societies: Anthropological Aspects; State Formation; State, History of; State, Sociology of the; States and Civilizations, Archaeology of

Bibliography

Anderson B R O'G 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Borneman J 1992 *Belonging in the Two Berlins: Kin, State, Nation*. Cambridge University Press, Cambridge, UK

Borneman J 1997 *Settling Accounts: Violence, Justice, and Accountability in Postsocialist States*. Princeton University Press, Princeton, NJ

Dirks N B 1993 *The Hollow Crown: Ethnohistory of an Indian Kingdom*. University of Michigan Press, Ann Arbor, MI

Ferguson J 1990 *The Anti-Politics Machine: 'Development,' Depoliticization, and Bureaucratic Power in Lesotho*. Cambridge University Press, Cambridge, UK

Foucault M 1977 *Discipline and Punish: The Birth of the Prison* [trans. Sheridan A]. Vintage, New York

Foucault M 1978 *The History of Sexuality* [trans. Hurley R]. Penguin, London

Geertz C 1980 *Negara: The Theatre State in Nineteenth-century Bali*. Princeton University Press, Princeton, NJ

Gauchet M 1997 *The Disenchantment of the World: A Political History of Religion* [trans. Burge O]. Princeton University Press, Princeton, NJ

Herzfeld M 1992 *The Social Production of Indifference: The Symbolic Roots of Western Bureaucracy*. Berg, New York

Lefort C 1986 *The Political Forms of Modern Society: Bureaucracy, Democracy, Totalitarianism*. MIT Press, Cambridge, MA

Lepsius M R 1988 *Interessen, Ideen und Institutionen*. Westdeutscher Verlag, Opladen

Moore S F 1986 *Social Facts and Fabrications: 'Customary Law' on Kilimanjaro, 1880–1980*. Cambridge University Press, Cambridge, UK

Nader L 1990 *Harmony Ideology: Justice and Control in a Zapotec Mountain Village*. Stanford University Press, Stanford, CA

Pemberton J 1994 *On the Subject of Java*. Cornell University Press, Ithaca, NY

Sahlins M D 1981 *Historical Metaphors and Mythical Realities: Structure in the Early History of the Sandwich Islands Kingdom*. University of Michigan Press, Ann Arbor, MI

Service E R 1975 *Origins of the State and Civilization: The Process of Cultural Evolution*. Norton, New York

Stoler A L 1995 *Race and the Education of Desire: Foucault's History of Sexuality and the Colonial Order of Things*. Duke University Press, Durham, NC

Tambiah S J 1992 *Buddhism Betrayed? Religion, Politics, and Violence in Sri Lanka*. University of Chicago Press, Chicago

Taussig M 1997 *The Magic of the State*. Routledge, New York
Tilly C 1990 *Coercion, Capital and European States AD 990–1990*. Blackwell, Oxford, MA

J. Borneman

State Formation

A state consists of a set of institutions whose functions are social control and authoritative decision-making and implementation processes. In addition, modern states seek autonomy in their relations with other states. Beyond these very general criteria, the concept of the state quickly becomes multidimensional with respect to its historical development, its relevant legal considerations, and its theoretical debates. The institutions of a state are subject to conflict and disagreement over their structure and the role they play. Institutions change over time, and the changes both shape and reflect differing degrees of autonomy and strengths. In the twenty-first century there is considerable interest not only in the historical trajectory of state formation and the more proximate formation of new states—dozens of new states have been created in the second half of the twentieth century—but also in the concept of state capacity. Political scientists and political economists who measure and assess the strength and relative autonomy (sovereignty) of states in national and international arenas concern themselves with the proper role of states in the domestic and international political economy. States are in constant competition with other institutions such as intergovernmental institutions and for-profit and nonprofit institutions. Political scientists study and observe the role and function of each of these various national and international institutions. Not surprisingly, the proper role, the function, as well as the capacity of today's states has become a lively and contentious issue.

1. *The Problematic of the State*

Many prominent classical theorists studied state formation and the multifaceted issues surrounding the role and function of states. Theorists such as Machiavelli (*The Prince* 1988), Rousseau (*The Social Contract* 1987), Marx and Engels (*The Communist Manifesto* 1998), and Lenin (*State and Revolution* 1993) considered the formation and role of the state to be a theoretically interesting issue. More recent students of the formation of the modern state have focused their attention on such other topics as the absolutist states of Western Europe (Anderson 1979) and states and revolution (Skocpol 1979). Contemporary writers address issues such as the state and industrialization, nationalism, colonialism, and the formation of the state, in addition to many other topics concerned with communism and the postcommunist state. Other ongoing themes among students of state formation are the state and the economy, power and legitimacy, and the relationship of states to other states. Twenty-first century analysis of the forces of globalization and their impact on the state are also frequently discussed in the context of the historical evolution of state formation.

2. *History of the State*

The modern state is a Western European creation which emerged in the fifteenth and sixteenth centuries and became more firmly grounded in the seventeenth century. The modern state came into existence in the same time frame and in geographical locations that were the wellsprings of modern capitalism, modern science, and philosophy. Many scholars link the evolution of the modern state and Christianity in its Protestant form. It is generally agreed that the emergence of a new political form—the modern state—was intimately linked to the economic, intellectual, and religious developments of the time.

The first modern states to arise were Italy, the UK, France, and Spain. The principal technical tools of modern state building were the development of permanent bureaucracies and a standing army, but it was the availability of money that enabled rulers to acquire them. Once a ruler was able to control the professional bureaucracy and the army, the nobility of medieval times lost its political role and power. The principal producers of cash were the commercial and industrial middle class and the peasantry. By the late eighteenth and early nineteenth centuries the bourgeoisie was able to curtail and even abolish the privileged status of the aristocracy and the authority of the monarch. The bourgeoisie acquired increasing control over the state. Their victory, signaled by the French Revolution, transformed the dynastic, state into the modern state. However, to define the modern state as 'bourgeois,' meaning 'civic' signifies the civil and legal equality of all subjects regardless of status by birth or occupation. The abolition of the old feudal privileges did not mean that the propertied and educated classes ceased to have privileges in the modern state. Modern theorists of state formation like Charles Tilly (Tilly 1975) and many Marxist theorists like Miliband (1969), Poulantzes, (1969), and Offe (1975) view the state as 'organized coercion' orchestrated by an elite. Other social scientists view the state as the arena of legitimate authority embodied in the rules of politics and in governmental leadership and policies.

Some thinkers, especially Marxists, tend to treat the state as an analytic construct by which to explain the modes of production, class relations, and societal struggles in general. Others, including Theda Skocpol

(1979), argue that states are 'actual organizations' that control territories and peoples. Skocpol echoes the legal definition of a state much more closely than does Tilly for example, by pointing out that while the state is anchored in class-divided socio-economic structures it also needs to be seen as part of the international system.

As a practical matter, in order for a state to engage in international relations, other states need to recognize it as a fellow state. Modern sociologists call attention to the fact that a state is a human community that successfully claims the monopoly of legitimate use of physical force within a given territory. Whether a state is considered legitimate and on what basis its legitimacy is acquired and honored in the human community within the territorial boundaries of the state and in the large international community of fellow states, is to say the least, very historically and theoretically complex.

3. *The Modern State*

The modern state and the modern state system developed simultaneously. It required coercive capacity for the state to sustain itself both domestically and internationally. The state system of the nineteenth and twentieth centuries led to two world wars and the Cold War. These wars were significant in shaping contemporary states and the international system of states. However, contemporary concern and interest in the state is not confined to questions of the legitimacy of states or even their coercive capacity. At issue in the twenty-first century is the role of states as actors, competing with other actors, to determine what the ongoing proper role and function of a state ought to be with respect to its political, social, cultural, and economic obligations to its citizens.

State formation has never been simply a 'founding.' Quite the contrary, state formation is a process of formation and reformation based on the changing nature of societies within the state and the larger international system of states. The institutions and processes of a state and of states collectively are not static. To say that a state requires the capacity to carry out the proper functions of a state is a truism, but the capacities required at the turn of the twentieth century vary greatly from state capacities required at the turn of the twenty-first century. Furthermore, states compete with other national and international institutions. For example, states vie with interstate organizations, such as the European Union, the United Nations (UN), the World Trade Organization (WTO), and with nonprofit organizations such as Greenpeace and Oxfam, and for-profit corporations such as Microsoft and CNN. Financial capacity, knowledge capacity, monetary capacity, and even legal capacities of nonstate actors all challenge state capacity. This is not to suggest that states have become defunct but merely that what it means to be a state today and what capacities a state is expected to possess by its citizens and in the larger community of states continues to go through a process of formation and reformation.

4. *The Future of the State*

As the twenty-first century begins there are almost 200 so-called sovereign states in the world. It is useful to think of contemporary states in their subgroups of differing state types: first, the 'old states,' such as the UK, France, Spain, the US, China, and Japan, which have developed reasonably stable nation states over the centuries. Some of these states, like the US for example, are multinational and multiethnic. Others, such as Japan, are more homogeneous. What these states have in common is a citizenry that views the national government as legitimate, and these states on the whole, share stable political economies. Second, there are the 'new' states. Many of these states, for example, most of the sub-Saharan African states and many states in Asia and the Middle East, emerged out of colonialism during the twentieth century. These new states tend to meet the legal criteria for statehood, yet many lack legitimacy as states in the eyes of their citizens. Most often such states are also multiethnic and have yet to achieve a sense of nationhood—a shared belief on the part of citizens that they belong to the nation-state. Some of these states, for example, Somalia or the Congo, are, at the beginning of the twenty-first century, in a state of civil conflict. States which are not able to carry out the basic function of a state, are frequently referred to as 'failed' states. Such states have the legal trappings of a state but the nation-state formation has broken down and the state's capacity to govern its territory has failed. In the category of failed or failing states are all those for which the historical process of state formation has not yet succeeded.

One may observe that the historical timeframe during which state formation takes place will determine both the internal and external factors which have a bearing on the process, as well as the outcome, of state formation. In the post-Cold War era for example, state formation and reformation—in the case of former communist states—is taking place at a time when the dominant paradigm is a democratic form of governance within the context of free market capitalism. Accordingly, state formation in the new states and postcommunist states, takes place in an international setting which exerts a distinctive global leverage on the type of state formation for which political and economic support and acknowledgement is rendered.

However, today state formation and the reformation of all states, not just new or failing states, takes place in the context of globalization. Globalization brings with it active participation in the national and international life of states by nonstate actors and a new degree of interstate and inter-actor interdepen-

dence. Today global and regional institutions, along with states, participate in rule making, rule enforcing, knowledge creation, and in fields such as commerce, banking technologies, medicine, genetic engineering, and in defining the standards for human rights. The globalized world of the twenty-first century is reshaping the idea and the reality of the sovereign state. Yet states still are extremely important in the lives of most people and hence, state formation continues to be an important ongoing process, a process which is not merely a multifaceted historical process but a multidimensional work in progress.

See also: Regional Government; Regional Integration; State: Anthropological Aspects; State, History of; State, Sociology of the

Bibliography

Habermas J 1976 *Legitimation Crisis*. Heineman, London

Lenin V I 1993 *State and Revolution*. Penguin USA, New York

Machiavelli N 1988 *The Prince*. Cambridge University Press, Cambridge, UK

Marx K 1983 The eighteenth Brumaire. In: Held O (ed.) *States and Societies*. New York University Press, New York

Marx K, Engels F 1998 *The Communist Manifesto*. Signet Classics, New York

Miliband R 1969 *The State in Capitalist Society*. Weidenfeld and Nicolson, London

Nozick R 1974 Anarchy state and Utopia. Basic Books, New York

Offe C 1975 The theory of capitalist state and the problem of policy formation. In: Lindberg L N, Alford R, Crouch C, Offe C (eds.) *Stress and Contradiction in Modern Capitalism*. Lexington Books, Lexington, MA

Poulantzas N 1969 *Classes in Contemporary Capitalism*. Weidenfeld and Nicolson, London

Rousseau J 1987 *The Social Contract*. Penguin USA, New York

Skocpol T 1979 *States and Social Revolutions*. Cambridge University Press, Cambridge, UK

Strange 1996 *The Retreat of the State*. Cambridge University Press, Cambridge, UK

Tilly C (ed.) 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, UK

I. V. Gruhn

State, History of

1. Introduction: Problems of Definition

At present there are about 200 communities in the world that are considered sovereign nation states, according to standards developed during European history. Officially, alternative political models no longer exist. Therefore, it seems reasonable to adopt the narrow definition of modern statehood created by German professors of law (Jellinek 1959). According to them, a state is constituted by a homogenous population inhabiting a contiguous territory under one single government which is characterized by complete independence from any outside authority (sovereignty) and holding the monopoly of jurisdiction and the legitimate use of violence internally. History has shown that this kind of state need not be a nation and even less a constitutional democracy based on human rights. But today most states claim to be nations under the rule of law and a formal constitution. Nevertheless, these are recent achievements and rather exceptional in world history.

History abounds with different kinds of political organization that may also be called 'states.' This is not quite correct from the historian's point of view, because in most European languages the term 'state' was originally used to designate social status or real estate. But between the seventeenth and nineteenth centuries it became the name of a particular type of European political organization, achieving perfection during that same period. Before then, pre-state communities can be roughly classified into: (a) tribes without rulers, (b) chiefdoms, (c) city-states, and (d) empires (van Creveld 1999). Historically as well as typologically, humankind was first organized in face-to-face communities, either acephalous societies without rulers based upon kin groups, or chiefdoms based upon force, wealth, and consequent social stratification. In addition, many chiefs were distinguished by religious qualities or supported by a privileged group of priests.

All this is even truer of empires, which through religious ideology, a standing army, a kind of bureaucracy, and a tax system solved the problem of permanent control of large territories. This was a necessary precondition of future state-building, but not the sufficient one, which was to be the achievement of some city-states of Antiquity. Whereas everywhere else the distinction between ruler and owner of a polity did not exist or was at least blurred, in Greece and Rome for the first time citizens appointed certain persons to govern on behalf of the community and not for their own purposes, not as rulers, but as magistrates. These were the roots of the state as a transpersonal, abstract institution with an abstract legal personality (van Creveld 1999). Therefore, the history of the state is first of all the history of the transformation of the loosely structured 'chiefdoms' of European aristocrats and 'empires' of European kings into modern states between the high Middle Ages until the nineteenth century.

2. Theory

European state-formation was a multidimensional process, but most theories of state-building are still one-dimensional. Therefore, a multifactorial three-level theory of state-building which combines (a) the micro-level of individuals and groups, (b) the meso-level of the political system, and (c) the macro-level of

society is a more promising suggestion (Reinhard 1992).

State-building starts on the micro-level with the self-seeking lust for power of individuals, often with the competitive advantage of holding a kingship. Before the existence of the state as an abstract institution, the necessary transpersonal continuity was provided by the dynasty. Dynastic state-building consisted of the elimination or at least control of the rival holders of autonomous power dating from the prestate phase of history—the nobility, the church, urban, and rural communities—in order to establish a monopoly of power. To succeed dynasties needed the assistance of power elites, which in their own interest made the growth of state power their own cause. In the long run, lawyers of bourgeois origin proved better qualified for that role than members of the church or the nobility because, in contrast to the latter, lawyers owed all their status and power to the service of monarchs.

Far-reaching change on the meso-level of the political system resulted from successful exploitation of war, religion, and patriotism for the purpose of expanding dynastic power. The pre-existing rivalry of European monarchs inevitably grew with their power, because it became essential to outstrip their neighbors, to grow at their expense, and to protect themselves against the same objectives in turn. Therefore, they needed continuously growing armies and money in ever-increasing quantity to pay them. In its decisive phase of growth the modern state was a war state, which expanded its taxation, administration, and coercive apparatus mainly to wage war.

This resulted in a circular process, the coercion-extraction-cycle (Finer 1997) and finally in the internal and external monopoly of violence. Ultimately, only states wage war. Private wars, such as vendettas or feuds, noble or popular revolts were no longer legitimate under the powerful war and police state. 'Necessity' in the service of the common good served as key argument to legitimate this growth of state power. But when the competing 'confessional' churches lost most of their autonomy to the state after the Protestant Reformation—the price to be paid for political protection—religion became an instrument of emotional identification of the subjects with their country. 'Catholic' and 'Bavarian,' 'Polish,' or 'Spanish' became almost synonyms on the one hand, as well as 'Protestant' and 'English,' 'Prussian,' or 'Swedish' on the other hand.

Essential contributions originated from the societal and cultural environment on the macro-level. First, the geo-historical plurality of Europe was the incentive to the growth of state power through the coercion–extraction cycle. The result was a stable pluralism of internally strictly unitary states, an exceptional case worldwide. Universal empires never had a chance in Europe; the Holy Roman Empire of the Germans was at most a first among equals. But internal unity was not realized before the late eighteenth, the nineteenth, and in some cases even the twentieth century. For a long time most monarchies were composed of several parts of unequal status such as Castile and Aragon or Polonia and Lithuania.

Everywhere monarchs had to cope with a powerful system of autonomous local noble rule on the one hand, with a countrywide network of partly autonomous urban and rural communities on the other, once again European particularities. In addition, before the Reformation the Church considered itself an independent community, in some respect even a state before the state. This singular European dualism of spiritual and temporal, combined with the equally singular political pluralism turned out to be a precondition of political freedom, although neither Church nor state, neither noble landlords nor urban oligarchies were in favor of any liberty except their own. Finally, the strong position of the Church derived from its role as keeper of Latin culture. Roman law, to some extent transformed into the canon law of the Church, directly and indirectly proved fundamental not only to monarchic state-building, but also to individual liberty and property.

3. *The Age of Monarchy (Middle Ages to Eighteenth Century)*

European kings of the Middle Ages and the early modern period claimed to be emperors of their kingdoms, which is to say that like the pope in the church they held unrestricted power (*plenitudo potestatis*). For the first time in history the competence of a king was defined by the French constitution of 1791. Before that date royal power was limited only in the sense that there were some things a king could not do. The English king could not create statute law or raise taxes without consent of the realm, the French king could not change the succession to the throne or give away parts of the crown domain.

For the rest, every monarchy was 'absolute,' that is to say the king's prerogative included everything necessary to fulfill his fundamental duties as fountain of justice and keeper of peace. In that sense even Poland's 'republic of nobles' and England's parliamentary system were called 'absolute' by contemporaries. In Britain, where no written constitution defines powers, in theory the monarch today is still entitled to declare war and to conclude peace without consulting Parliament. In reality, however, since the later Middle Ages this was possible only if the necessary funds were raised with the help of representatives of the subjects. The monarchies of the European Ancien Régime were limited not so much by legal restraint as by small resources, traditional rules, and social conventions. Under these circumstances it was decisive for the process of dynastic state formation that the crowns of Castile, England, and France

became hereditary whereas the Empire and to some extent Denmark, Sweden, Poland, and Hungary remained elective monarchies. Compared with the universal competence of the modern state, including the right to extend this competence at discretion, the prerogatives of early modern monarchs were extremely modest. State activity expanded, when during the early modern period the so-called 'police' ordinances ('Policeyordnungen') started to cover new fields such as health and public welfare, moral order and communication, economic behavior, and labor relations.

This was no longer mere jurisdiction, reactive government by rescript as in the Middle Ages, but an expanding administration actively interfering with larger parts of the subjects' lives. But it was still not 'absolutism,' an anachronistic term of the nineteenth century which should be dropped (Henshall 1992), because the kind of monarchy without any legal restriction it designates never existed except in Denmark 1665–1834 and in Sweden for a few decades after 1693.

Similar to most monarchies of the world, European ones had a sacred character, practiced elaborate symbolic rituals, and were legitimized 'by the grace of God.' The coronation ceremony together with the coronation oath of the king and the corresponding oath of allegiance of the people were central in this respect. But there were varieties. In France political rituals and ideology constituted a true 'royal religion' ('religion royale'), and court life with its palaces, ceremonies, and festivities became an all-embracing total work of art. Some of its elements recur in England and elsewhere, especially the ritual of the royal touch, when French and English monarchs touched hundreds of people suffering from scrofula, the tubercular 'royal disease' ('morbus regius'), to cure them. Monarchs practiced this ritual with particular intensity when their monarchy needed additional emotional support: Henry IV and Louis XIII of France after the religious civil wars, Elizabeth I of England because of her succession problems, and Charles II after the restoration.

Other monarchies were more prosaic, including the Holy Roman Empire. And in spite of his politico-religious sanctuary, the Escorial, Philip II of Spain never wore a crown; there was not even a coronation ceremony in Castile. This was a way of demonstrating independence from the Church, but even in France political sacrality implied royal control of the Church and not the opposite.

Because of the Church's monopoly of culture, Europe's early political discourse was theological. The state was considered a consequence of the fall of man. With the reception of Aristotle natural man became a political being, a concept which was used to create an autonomous discourse of legitimation for the monarchy, erroneously, because 'political being' referred to the face-to-face community of the city and not to the state. But from Dante Alighieri down to John Locke, political theory remained essentially a theory of monarchical state-building, at most harmonizing limited monarchy with aristocracy and democracy into a mixed constitution, supposedly like the ancient Roman republic. Thomas Hobbes was the first to speak of the state instead of the monarchy, but this was not common practice before the eighteenth century, with Montesquieu and Jean-Jacques Rousseau.

This change coincided with the development of central institutions, which under the impact of enlightened rationality became more effective. Administrators were no longer the king's personal servants but transformed into professional officials of an abstract state, into modern civil servants ('Beamte') as described by Max Weber. The early medieval prince was assisted in his governmental functions by members of his court and council ('curia regis'), for written legal work by a chancellor, often an ecclesiastic. Most central institutions of European monarchies were created by segmentation of that curia. The court was reduced to a mere household, the chancery became a separate body.

Professional supreme courts and treasuries were established. To produce political consensus and financial aid an enlarged council of important subjects was created to become an assembly of estates or a parliament. What was left over of the ancient council was transformed from a multifunctional decision-making instrument such as the German aulic council ('Hofrat') into a differentiated system of governing bodies with a small group of confidants of the ruler at the core ('Privy Council,' 'Geheimer Rat,' 'Staatsrat,' etc.). From the sixteenth to the eighteenth century the conciliar principle of collegiality was common practice, to accumulate competence, to balance interests, to prevent disloyalty by mutual control, and to provide positions for as many members of the power elite as possible. Only a few very dedicated rulers such as Philip II of Spain, Louis XIV of France, and Frederic II of Prussia were still able to handle this complicated system of central government in person. Instead, many princes relied on a personal confidant as prime minister, such as Wolsey, Richelieu, Olivares, or used the new office of a secretary of state for coordination. In England, parliamentary management was a central task of royal confidants. The development of modern cabinet government was the consequence.

In France, Castile, Rome, and elsewhere many public offices were officially for sale, just another way to tap private accumulation in favor of public finance. The growth of state power was first of all a credit problem, because tax income had to be anticipated by credit to make the large sums of money which were necessary to wage war available on short term. Tax farming was another widespread technique used for that purpose. Direct taxes were based on the land, because movables and income were still not under administrative control. As usual, indirect taxes on

mass consumption proved more comfortable, whereas the rich were involved as creditors. But credit was still the personal credit of a prince, not of the state as institution. Therefore, repayment of loans was expected and interest rates were as high as risks. Under these circumstances, bankruptcy was the regular consequence of power policy until the Dutch and English invention of low interest long-term bonds with parliamentary guarantee, a decisive contribution to the further growth of the modern state.

4. *Premodern Alternatives and Their Political Heritage*

Traditionally, Europe respected man (and to some extent woman) as proprietor of his (her) person and goods. Therefore, in principle taxation was only possible with the consent of the taxed or at least their lords and guardians. For this purpose, most princes created structurally similar assemblies of important subjects or representatives of corporations, such as cities, called assemblies of estates, beginning 1188 in León. By the seventeenth century, however, most princes were strong enough to raise taxes without consent and to ignore their estates. This decline of estates has been considered the very essence of 'absolutism' by former research. In contrast, we know today that to some extent estates remained alive in most countries (with the exception of Denmark) and princes still had to bargain with them. But only in England seventeenth-century conflicts with the crown ended with victories of parliament and the definite establishment of financial control by that body. This system, imitated in most European countries after the French Revolution, finally led to parliamentary government and democracy.

In contrast to other parts of the world, in European political culture the relationship between ruler and subjects was considered a contractual one, based on legal mutuality. In case of a supposed breach of contract by the ruler, subjects felt entitled to armed resistance, because their legitimate lord had degenerated into an illegitimate tyrant. Therefore, early modern discourse and practice of the right of resistance through a series of European revolutions laid the foundations of modern liberty and democracy.

In most countries, the nobility dominated assemblies of estates, as holders of local power the true lords of the land. Not before the eighteenth century did the administrative institutions of the state start to penetrate this autonomous local power structure and not before the nineteenth century was noble rule to a large extent replaced by public administration. But beside or below noble rule existed innumerable more or less self-governing urban and rural communities, from independent republics like Venice, quasi-republics like German imperial towns, privileged English boroughs, down to villages of Polish serfs who were still entitled to regulate certain common affairs on their own. This is the 'Europe of thousand republics,' the Europe of 'communalism' (Blickle 2000), a world where members of face-to-face communities decided on their immediate concerns. These are the roots of European republicanism and of the tradition of European self-administration.

When assemblies of estates flourished, ecclesiastical power competing with princes was already declining. But the Roman church had been Europe's first state with the pope as the first absolute monarch and the priest as the first nonhereditary professional public servant. The Church practiced territoriality ('*Flächenstaat*') at a time when empires were still mainly based upon personal dependency ('*Personenverbandsstaat*'). And the church established the rules of political religion with the consequence that to the present day states remain transcendent entities in need of believers. In this sense the state is the last church because European political culture is a culture of believers.

In contrast, professional law courts such as the French 'parlements' or the English 'common law and equity courts' were not originally autonomous alternatives to state-building monarchies like aristocratic rule, communal self-government, or the Church but creations of the expanding state power. Nevertheless, because of the loose structure of early modern polities giving room for the development of specific group identity and interests, these judges under certain conditions could transform themselves into opponents of their king's policy and initiate resistance or revolution, as happened in England and France in the seventeenth and eighteenth centuries. But in this way the principles of rule of law, of an independent judiciary, and even of a judicial review of legislation were introduced into political culture.

In the end, however, the fully developed modern state replaced noble rule by his officials, reduced local self-administration to a commission of the central government, and established complete control of the churches and the legal apparatus. The principle of justice existing beyond, above, before the state was largely replaced by the arbitrary use of the state's monopoly of legislation. But nevertheless the heritage of political alternatives remained alive: liberty and democracy, rule of law and independent judges, the claim to self-government, and to have a state one could believe in.

5. *The Age of Equality (Nineteenth and Twentieth Centuries)*

The modern European state, so far developed by state-building monarchies, reached maturity with the French Revolution and the political modernization of Europe which followed. Modernity means unity, unity of territory, people, government, sovereignty, and violence. The revolutionary process took the last step

to modernity and replaced the hierarchy of orders and the plurality of privilege which had characterized the political society of the Ancien Régime with a society where stratification became a private distinction because politically all subjects were now equal before the state and its law. From now on, not families or corporations were the basic components of the state, but individuals. But mobilizing individuals emotionally in favor of the state into some kind of participation—not necessarily a democratic one—resulted in an ultimate increase of state power with incredible dimensions: the total state.

With the quasi-natural order of estates and corporations gone, the modern state of equal individuals needed a new organization. With the exception of Britain and Sweden (until 1975) this was provided from now on by a constitution. But in spite of the American and French models of 1787 and 1791 nineteenth-century European constitutions established rule of law, but neither democracy or fundamental rights. Their first purpose was equal integration and not equal participation. Most countries remained monarchies, but deprived of sacrality since the period of enlightenment, monarchy lost unquestionable legitimacy. Step by step monarchies were transformed into democratic republics or quasi-republics. But in most countries during the nineteenth century democracy was still far away. Not only informally as usual, but even legally, in their right to vote, all subjects were equal, but some were more equal than others (Orwell 1945), mostly because of their wealth. In 1830 the following percentage of the total population was entitled to vote: in Austria, Denmark, Prussia, and Spain 0 percent, France 0.5 percent, Belgium 1.1 percent, Britain 2.1 percent. In 1910 the numbers were: in Denmark 17 percent, Britain 18 percent, Austria, Belgium, Germany and Spain 21–23 percent, and France 29 percent (Reinhard 1999).

Percentages in the twenties indicate universal suffrage of adult males. Until the twentieth century, state-building and politics remained almost exclusively a business for men. There were some outstanding queens and empresses in European history—Isabella of Castile, Elizabeth I of England, Christina of Sweden, Maria Theresa of Austria, Catherine II of Russia, and others. But some of these great women had their difficulties with male political culture. And there were no female ministers and officials. Even the enlightened philosopher Immanuel Kant considered women political non entities because they lacked economic independence. After Wyoming 1869 and New Zealand 1893 female suffrage was introduced in Finland in 1906, Austria, Germany and Poland in 1918, the USA in 1920, Britain in 1928, Spain in 1931, France in 1944, and Switzerland in 1971. But in the late twentieth century female heads of state or government, ministers, parliamentarians, and even female civil servants were still a minority.

Originally, humans identify themselves mainly with their face-to-face community. 'Patria,' the land of the fathers, and 'Natio,' the community of birth, even during the early modern period quite often designated rather small territorial entities. However, through state-building people learned to apply them to their country in the large. Finally, after the French Revolution self-defined linguistic or cultural communities, peoples in search of a state considered themselves nationalities, that is, nations to be. In contrast to ideology, nations are not natural, but invented entities. This does not imply complete artificiality and arbitrariness. There is always some kind of basis in reality, such as common descent, language, religion, or territory. Most important is a common history and the consequent option for its continuity. Nationalism as a system of symbols constitutes a community as a nation, integrates its members, and defines the difference from other communities.

For radical nationalists the nation is the highest value and the ultimate source of meaning; nationalism becomes a religion and as such a way to tap the material and emotional resources of subjects, especially in war. Therefore, nation-building became the final phase of European state-building, whereas new states in Europe and worldwide tried to constitute themselves as nation-states right from their foundation.

Expanding political participation and emotional mobilization of subjects correspond to a further extension of government, of state power. A common denominator is a high degree of social discipline in the subject. Church and State had been working on this since the late Middle Ages, but achieved this goal only in the nineteenth and twentieth centuries with the assistance of armies, schools, and factories (Weber 1964). The citizens of a fully developed modern state enjoy the legal protection of their person and property and some political participation. In return, a citizen is subject to three fundamental duties, compulsory school attendance, compulsory military service, and general liability to taxation. Based upon the fiction of the sovereignty of the people the state apparatus is able to extend its competence at its own discretion and to transform itself into a totalitarian system as happened under the impact of nationalist, racist, and communist ideologies in the twentieth century. Some of them proclaimed the end of the state, but that proved impossible. In contrast, the totalitarian state turned out to be the final stage of development of the modern state. And at a closer look, even the modern welfare state proves to be a soft variety of totalitarianism. Beginning with limited social security programs before World War I (1883–89 creation of an insurance system for workers in Germany) the welfare state exploded after World War II. Between 1950 and 1983 the West-European average of social security expenses rose from 9 percent of the GNP to 25 percent. But the total care for the citizens means also total

regulation of their lives. Even the neo-liberal reaction is based upon large-scale government action in the service of a general economic and social policy; goals, not methods differ from former socialist welfare policy.

Six hundred years of European expansion made the modern European state the world's standard model of political organization. The colonial regimes of Spanish and British America established in the sixteenth and seventeenth centuries adopted premodern forms of European statehood and political culture. Therefore, most Latin American states remain authoritarian but weak states even today, whereas traditional self-government in a comparatively egalitarian society made the USA a successful nation, although it lacked certain traits of the modern state such as belief in the state as a value in itself, in bureaucratic centralism, in a strong professional army, and in an established church. From 1839 to 1935 Britain's white settler colonies became sovereign states, mostly combining the British and US models, whereas in the nineteenth and early twentieth centuries East Asian and Muslim empires tried to transform themselves into modern states to cope with the weakness of their traditional regimes when confronted with the West, in the case of Japan with tremendous success. Finally, during several waves of decolonization after World War II, the lot of quite artificially created colonies became equally artificial nation states. For the first time a global community of international law and moderately effective international organizations such as the United Nations (UN) became reality. Quantitatively and qualitatively the modern state reached its culminating point during the twentieth century.

6. *The Disintegration of States and Possible New Forms of Political Life*

In the last third of the twentieth century, culmination of the modern state turned into decline. To begin with, most 'new' states worldwide suffer from serious shortcomings, such as lack of democracy and efficiency, general corruption, and political disintegration. The export of European state models was a success only in three specific constellations: (a) where it was accompanied with the export of European civil society and political culture as in the cases of the USA and the British dominions; (b) where it met a polity and society compatible with its own structure as in the case of Japan; and (c) where colonial rule lasted long enough to create a Western-style civil society as in the case of India. For the rest, the new states of the former colonial and the former communist worlds tend to loose their sovereignty internally to the para-statehood of old or new chiefdoms or ethnic communities, externally to international organizations like the UN or the International Monetary Fund (IMF). They become mere 'quasi-states' (Jackson 1990).

At the same time European states were also, to some extent, disintegrating into 'tribes,' were loosing sovereignty to 'chiefs,' pressure groups, and international organizations, and were no longer able to provide complete social security. As a consequence of totalitarianism and mass participation, the modern state turned self-destructive. The excesses of nationalism, racism, and communism destroyed the necessary faith in the state. Social discipline was questioned by the common desire of self-realization. New ethnic, regional, social movements started to compete with the state for the loyalty of the subjects, some on the basis of democratic self-determination, some with the use of violence. Withdrawal of loyalty led to an underground economy expanding at the expense of the tax and welfare systems. The necessity to win elections by bargaining with interest groups and by providing additional welfare benefits for the electorate ended in a debt-trap with almost the whole budget allocated in advance. In addition, economic globalization allows important tax-payers to move to less restrictive countries with lower wages and taxes.

However, for economic and security reasons most states loose more of their sovereignty to international organization. The European Union has not yet become a federal state and suffers from deficient democracy and social policy, but nevertheless common economic interest has released remarkable binding forces. The right to declare and to wage war, once the very essence of sovereignty, has become obsolete. Instead of interstate wars and peace treaties we have hundreds of 'low intensity conflicts' (van Creveld 1991) which tend to respect neither non-combatant status or national borders.

Nevertheless, the national state is still there. But it has definitely lost the monopoly of power and policy. In power processes, it tends to become one competitor among others. Unity, the constituent principle of modernity, the monocratic polity, is being replaced by a postmodern plurality of multiple political units coordinated by the necessity to cooperate and not by state power. Certainly, this pluralism might end in chaos. But it might also prove a new way for humankind to seek the political pursuit of happiness.

See also: Absolutism, History of; Aristocracy/Nobility/Gentry, History of; Balance of Power, History of; Bureaucratization and Bureaucracy, History of; Citizenship, Historical Development of; Citizenship: Political; Citizenship: Sociological Aspects; Communism, History of; Conservatism: Historical Aspects; Contemporary History; Democracy; Ethnic Groups/Ethnicity: Historical Aspects; Frontiers in History; Global History: Universal and World; Historiography and Historical Thought: Current Trends; History and Memory; Human Rights, History of; Interest Groups, History of; International Relations, History of; Liberalism; Military History; Modern-

ization and Modernity in History; Nation: Sociological Aspects; National Liberation Movements; National Socialism and Fascism; Nationalism: General; Nationalism, Historical Aspects of: The West; Nations and Nation-states in History; Parliaments, History of; Peacemaking in History; Play, Anthropology of; Political History (History of Politics); Political Parties, History of; Psychology: Historical and Cultural Perspectives; Racism, History of; Social Movements, History of: General; Socialism; State and Society; State: Anthropological Aspects; State Formation; State, Sociology of the; Urban Anthropology; Violence, History of; Warfare in History; Welfare State

Bibliography

Anderson M S 1993 *The Rise of Modern Diplomacy, 1450–1919*. Longman, London

Bayart J-F 1989 *L'état en Afrique. La Politique du Ventre*. Fayard, Paris [English translation 1996 *The State in Africa. The Politics of the Belly*. Longman, London]

Berman H J 1983 *Law and Revolution. The Formation of the Western Legal Tradition*. Harvard University Press, Cambridge, MA [1991 *Recht und Revolution. Die Bildung der westlichen Rechtstradition*. Suhrkamp, Frankfurt, Germany]

Blickle P (ed.) 1997 *Resistance, Representation, and Community*. Clarendon Press, Oxford, UK [1998 *Résistance, Représentation et Communauté*. Presses Universitaires de France, Paris]

Blickle P 2000 *Kommunalismus. Geschichte einer gesellschaftlichen Organisationsform*. 2 Vols., Oldenbourg, Munich, Germany

Bonney R (ed.) 1995 *Economic Systems and State Finance*. Clarendon Press, Oxford, UK [1996 *Systèmes Économiques et Finances Publiques*. Presses Universitaires de France, Paris]

Coleman J (ed.) 1996 *The Individual in Political Theory and Practice*. Clarendon Press, Oxford, UK [1996 *L'individu Dans La Théorie Politique et Dans la Pratique*. Presses Universitaires de France, Paris]

Contamine P (ed.) 2000 *War and Competition Between States*. Clarendon Press, Oxford, UK

Creveld M v 1991 *The Transformation of War*. Free Press, New York [1998 *Die Zukunft des Krieges*. Gerling, Munich, Germany]

Creveld M v 1999 *The Rise and Decline of the State*. University Press, Cambridge, UK [1999 *Aufstieg und Untergang des Staates*. Gerling, Munich, Germany]

Edelman M 1972 *Politics as Symbolic Action*. University of Illinois Press, Urbana, IL [1976 *Politik als Ritual. Die symbolische Funktion staatlicher Institutionen und politischen Handelns*. Campus, Frankfurt, Germany]

Edelman M 1974 *The Symbolic Uses of Politics*. Markham, Chicago, IL [1976 *Politik als Ritual. Die symbolische Funktion staatlicher Institutionen und politischen Handelns*. Campus. Frankfurt, Germany]

Ellenius A (ed.) 1998 *Iconography, Propaganda, and Legitimation*. Clarendon Press, Oxford, UK

Finer S 1997 *History of Government From the Earliest Times*, 3 vols. University Press, Oxford, UK

Hamilton K, Langhorne R 1995 *The Practice of Diplomacy. Its Evolution, Theory and Administration*. Routledge, London

Henshall N 1992 *The Myth of Absolutism. Change and Continuity in Early Modern European Monarchy*. Longman, London

Hobsbawm E J 1990 *Nations and Nationalism since 1780: Programme, Myth, Reality*. University Press, Cambridge, UK [1991 *Nationen und Nationalismus. Mythos und Realität seit 1780*. Campus, Frankfurt, Germany; 1997 *Nations et Nationalisme Depuis 1780: Programme, Mythe, Réalité*. Gallimard, Paris]

Hoffman J 1995 *Beyond the State*. Polity Press, Oxford, UK

Jackson R H 1990 *Quasi-states. Sovereignty, International Relations, and the Third World*. University Press, Cambridge, UK

Jellinek G 1959 *Allgemeine Staatslehre*, 3rd edn. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany [1st edn. 1900]

Jouvenel B D 1945 *Du Pouvoir. Histoire Naturelle de sa Croissance*. Bourquin, Geneva, Switzerland [1972 *Über die Staatsgewalt. Naturgeschichte ihres Wachstums*. Rombach, Freiburg, Germany; 1993 *On power. The Natural History of its Growth*. Liberty Fund, Letchworth, UK]

Koenigsberger H G (ed.) 1988 *Republiken und Republikanismus im Europa der Frühen Neuzeit*. Oldenbourg, Munich, Germany

Mann M 1986–1993 *The Sources of Social Power*, 2 Vols. Cambridge University Press, Cambridge, UK

1995–2000 *The Origins of the Modern State in Europe, 13th–18th Centuries*. 7 vols. Clarendon Press, Oxford, UK [1996–1998. *Les Origines de l' État Moderne en Europe, XIIIe–XVIIIe Siècle*. 7 vols. Presses Universitaires de France, Paris; 1998 sq. *Los Orígenes del Estado Moderno en Europa, Siglos XIII a XVIII*. 7 vols. Fondo de Cultura Económica, Madrid, Spain]

Orwell G 1945 *Animal Farm*. Secker & Warburg, London

Padoa-Schioppa A 1997 *Legislation and Justice*. Clarendon Press, Oxford, UK

Reinhard W 1992 Das Wachstum der Staatsgewalt. Historische Reflexionen. *Der Staat* **31**: 59–75

Reinhard W 1996 *Power Elites and State Building*. Clarendon Press, Oxford, UK [1996 *Les Élites du Pouvoir et la Construction de l'État en Europe*. Presses Universitaires de France, Paris; 1997 *Las Élites del Poder y la Construcción del Estado*. Fondo de Cultura Económica, Madrid, Spain]

Reinhard W 1999 *Geschichte der Staatsgewalt. Eine Vergleichende Verfassungsgeschichte Europas von den Anfängen bis zur Gegenwart*. Beck, Munich, Germany

Reinhard W (ed.) 1999 *Verstaatlichung der Welt? Europäische Staatsmodelle und außereuropäische Machtprozesse*. Oldenbourg, Munich, Germany

Ritter G A 1991 *Der Sozialstaat. Entstehung und Entwicklung im internationalen Vergleich*, 2nd edn. Oldenbourg, München, Germany

Weber M 1964 *Wirtschaft und Gesellschaft. Studienausgabe*. Kiepenheuer und Witsch, Köln, Berlin, Germany [1st edn. 1921, 1978 *Economy and Society*. University of California Press, Berkeley, CA]

W. Reinhard

State, Sociology of the

Few terms are as ubiquitous in the social sciences as that of the 'strong state,' yet there is no agreement on how to define its scope. The strong state crops up in

the writings of economists, political scientists, and sociologists. It has been labeled both a destroyer and creator of wealth. It has been connected both to chronic underdevelopment and, just as frequently, to high levels of economic performance. If everyone knows what a strong state is, why do they disagree over its role?

1. Strong and Weak States

In the conventional, 'neoclassical' view of the state and economic development, two types of government exist—strong and weak—with the latter usually given better grades for promoting economic performance (Klitgaard 1991). In this view, strong states typically choose interventionist strategies for development. This notion is largely based on the strong states of Latin America, where state interventions in economic affairs include populism, import substitution, and regulatory manipulation of markets, often designed to obtain substantial and, typically, unsustainable redistributive goals (see Dornbusch and Edwards 1991).

The usual justification for developing countries to adapt strong activist states has been that the private sector would not provide the massive public investments needed for economic growth. In reality, however, strong states often ended up pursuing power for the sake of their own leaders; elitism, corruption, and misallocation of resources resulted. Disillusionment with the performance of state-led growth strategies caused critics to champion weak states, which foster growth by leaving markets alone (Friedman and Friedman 1980). The minimalist state is touted for doing no more than providing a stable, credible currency and a strong legal system designed to enforce private property rights and contract law. On the other hand, such weak states have historically been unable to prevent the capture and distortion of economic policy by oligopolies and rent-seeking elites.

The conventional language of *strong* vs. *weak* thus fails to explain a government's ability to foster or hinder economic performance. East Asia's high-performing economies—South Korea, Malaysia, Taiwan, Hong Kong, and Singapore—dramatically demonstrated the inadequacy of the traditional 'neoclassical' categories. In this region strong, not weak, states enabled the private sector to make a major contribution to growth. Contrary to expectation, the weakest states in East Asia—Indonesia and the Philippines—did not nurture strong private sectors; in fact, they were much less successful than the others in terms of generating home demand and accumulating capital and skills. In response to the economic successes of these activist East Asian governments, a new generation of statist scholars has emerged to advocate the role of a strong state in promoting economic development (see Campos et al. 1994, Johnson 1982, Wade 1990a, 1990b, Evans 1992). The result is a strong polarity in the literature. In one camp, a strong state is viewed as an obstacle to economic development; in the other, state activism is touted as an essential component in the most celebrated cases of economic growth since the end of World War II. Both of these views contain important elements of truth, but both are too limited to account for the varied performance of developing nations.

2. Weak States

A weak government is handicapped by the inability to enforce property rights, or even many of its own laws. Because of the legal inadequacy of a weak state, its interventions are based not on rules, but on administrative or executive discretion; this results in corruption and opportunism rather than supervision. In fact, insufficient legal clarity is often intentional, thus facilitating random interventions by political authorities. For example, to compensate for rampant tax evasion, governments in developing countries devise tax regimes that allow considerable official discretion in assigning rates. But the capricious application of expansive tax powers drives private firms into the informal sector. Deprived of these tax revenues, the government cannot provide basic services such as law enforcement, and few benefits remain for those firms that do disclose profits. When transparency becomes a risk, firms will have difficulty attracting capital, and the government will become weaker still.

When states are too weak to control their own officials, agents of the state may act independently of one another. Thus, a policy of reform may be announced by one part of the government, while another part blocks its implementation. The inability to enforce rules reduces private business activity unless some important actor of the state, such as the president's family or inner circle, intervenes (for example, Indonesia under Suharto, the Philippines under Marcos, or Russia under Yeltsin). Political power is needed to defend property rights when the impartiality of the legal system is compromised or when the rules are themselves inadequate. When countervailing institutions such as an independent legislature or judiciary are weak, the leader can provide market access to chosen favorites in return for their loyalty, disabling the discipline of market forces.

Bureaucrats in a weak state learn to use their power and access to information to extract value from proposed economic activity rather than to promote that activity. With corruption flourishing at all levels of government (Shleifer and Vishny 1993), the deal-making authority of leadership becomes essential. In this regard, a weak state acquires the same market-destroying tendencies of strong discretionary states, in which property rights enforcement is at the discretion of a handful of individuals.

Although weak states respond to interest groups, they often cannot enforce bargains made with important constituents. Interventionist rules typically cannot be consistently enforced. As a result, reform measures, property rights, and contracts are not durable because the state lacks the means to implement and maintain them. Unable to enforce these minimal conditions, weak states can rarely sustain competitive markets. Instead, networks of individuals with special connections to the political regime are needed to sustain trade. Inevitably, officials of weak states are likely to collude with select, but powerful, private groups to extract the nation's resources. A weak state easily becomes a predatory state.

Many of the postcolonial African states fall into this category. Publicly disinterested bureaucracies in which pay and promotion were linked to performance did not emerge, compromised by governments that disbursed employment in exchange for loyalty. Dependency on foreign capital, including aid, discouraged accountability to citizens while eroding national sovereignty. Ethnic, religious, and regional fragmentation gave rise to political parties that enjoyed only narrow social support. Excessive centralization was often pursued to overcome these forces of national disintegration. However, highly centralized governments allowed bureaucracy to be replaced by a single-party or no-party system with national planning bodies controlling public goods and services. Local governments that were allowed neither independent judicial authority nor the resources to develop managerial and technical capacity failed to deliver basic services. The legislature and judiciary were dominated by the executive branch, eliminating independent oversight of budgetary allocations. The result of this excessive centralization was often a weak state dominated by the unbridled discretion of the executive, degenerating into personal rule and kleptocracy. Support for government collapsed, and many African states became incapable of providing basic public order.

3. *Strong and Unlimited States*

Strong and unlimited states, too, can act in a discretionary fashion. Lacking constitutionally defined limits, they are likely to be both interventionist and confiscatory. Able to alter rights and markets at will, they are incapable of making a credible commitment to private sector development. (Contemporary examples include Vietnam, Syria, Ethiopia, Iraq, and the USSR.) To induce participation in markets, strong and unlimited states must offer protection in the form of monopoly rights, preferential tariffs, or guild or trade union privileges. The exercise of unlimited political discretion allows them to promise excessive benefits to win the support of particular groups. This, in turn, creates political risk, reducing economic investments and undermining the possible outcomes of reform. The irony is that due to their unlimited power, no contract with the government or of interest to the government is secure. A strong and unlimited state has much in common with a weak state; both are likely to become predators of private sector wealth.

4. *Strong but Limited States*

A third type of state, which arose in Western Europe, is not only strong enough to establish and maintain property rights, but is constitutionally prevented from violating these rights. This state requires an administrative structure that keeps the economic and political activities of regime officials separate. Such a state must be strong enough not only to adopt rules that can sustain a competitive private sector, but also to prevent itself from responding to the political forces that inevitably arise to monopolize access to the marketplace. Only this type of state will ensure competitive access to open markets.

Strong but limited states are often mistaken for minimal states, since in comparison with interventionist states, they directly provide a relatively small number of goods and services. But the term *minimal* gives the impression of weakness, which is misleading. These states must instead be strong enough to withstand the political pressure on government to intervene in markets. Hong Kong under British rule was perhaps a perfect example of such a state in which 'positive noninterventionism' protected the market from pressure groups, rent seekers, and unscrupulous officials. Like the governments of Western Europe and North America, East Asia's other high-performing economies—Japan, Malaysia, South Korea, Singapore, and Taiwan—all accepted limits on their discretion over private sector profits. All adopted mechanisms to separate the political and economic functions of government and to protect property and contract rights.

5. *Implications*

In distinguishing strong but limited government from the conventional perspective that identifies only weak and strong government, the advantages of each and the contradictions between them can be understood. In the conventional view, the two types of strong states are not distinguished. A state that carries out and maintains highly interventionist policies, such as import substitution or subsidizing a highly unionized urban workforce, is confused with a strong but limited state that plays a neutral role in the allocation of rents and subsidies. Federalism in the USA, for example, offers an institutional foundation for a strong but limited state which is able to foster equitable growth (see Weingast 1995, 1997).

In many emerging states, technical and practical knowledge is unevenly distributed. Often a small group of capitalists dominates local industry, agriculture, and access to the external market. The superior technical knowledge of these small groups forces the government to depend on them to build the market, which allows representatives from that group to gain control over the government's regulatory apparatus and ultimately over the state. When economic power is concentrated in a few governmental ministries, it is more vulnerable to capture than when decision-making powers are dispersed among institutions that have veto power over each other.

Strong but unlimited states often promote corruption and rent-seeking, exacerbating social and regional inequality. As noted earlier, both strong unlimited and weak states are prone to plunder the markets. What 'neoclassical' wisdom calls a weak state can often more aptly be described as a strong but limited state that can maintain the market, providing the necessary public goods while resisting capture by interest groups.

While the champions of state activism raise reasonable doubts about the earlier neoclassical view, they do not explain why some strong states promote economic development while others pursue policies that hinder it. After all, the majority of countries that pursued activist policies experienced negative growth along with exacerbated social inequality. Even if the arguments advocated by statists—that these states have made a direct and significant contribution to economic growth—are taken at face value, they have done no more than show that this role for the state is feasible, something long known.

But why have some strong states promoted growth rather than other political goals? The conventional view offers little help. By contrast, the example of East Asia's successful regimes suggests that states are strong when critical and durable limits channel government behavior into activities compatible with economic development.

6. *Achieving Strong but Limited Government*

The development of strong but limited government depends on the existence of institutions that make it in the interests of political officials to abide by limits on their power. In many reform environments, the desire to create strong but limited government often drives leadership to increase the executive's discretionary power. Both Mikhail Gorbachev in the former USSR and Boris Yeltsin in Russia succumbed to this approach. Many Latin American leaders have also been attracted to it. Reform by mechanisms such as presidential decree is a two-edged sword, however. On the benefit side, the reform-minded leader gains the power to design and implement a reform package by eliminating many of the veto points blocking reform. On the cost side, all such reform programs can also be undone by decree and hence are only as good as the paper on which they are written.

For economic reform to be durable, it must be costly for political authorities to renege on promises once made. As the US federalists realized, durability cannot depend upon altruism. No government should expect public service to attract only godlike figures who always place the needs of the public before those of their family or friends. Institutions must be designed so that political actors have incentives to maintain and preserve the system, while carefully delineating where their power to intervene in private economic activity begins and ends.

A government's commitment to market-led growth requires the design of appropriate government institutions that ensure clear boundaries between private and public, and that protect both property rights and agreements with the business sector. The creation of such boundaries will shift the investment horizons of investors toward the long-term so that growth can be sustained.

Are boundaries that protect private wealth from government plunder more likely to be erected in weak or strong states? In their effort to establish competitive market economies, weak states often falter because narrow interest groups obtain control over the market, over the judiciary, and ultimately over the polity. Oligopolies often arise to overcome systemic market risk caused by the weak enforcement of laws and contracts. In countries as diverse as Indonesia, the Philippines, and Russia, elites have enough resources to block necessary reforms and to prevent transparent governmental practices from being implemented. Strong and unlimited states will falter because their leaders may refuse to implement reforms that dilute centralized power and allow a decentralized private sector to grow.

7. *The Asian Miracle and the Sociology of the State*

The successful economies of East Asia offer examples of strong states that encouraged economic growth by institutionalizing the following set of limits on executive discretion: (a) channels of representation for businesses to induce them to make long-term investments, (b) wealth-sharing schemes to elicit the support of the wider population for development; (c) a competent bureaucracy encouraging credible government commitment to long-term, growth-enhancing policies.

The significant improvements in income distribution that accompanied sustained economic growth in the high-performing East Asian economies resulted from effective use of public and private sector institutions and organizations. Common throughout East Asia were deliberation councils, which, although limited to particular sectors, offered input into the

policy process and had veto power over the outcome. The councils provided greater predictability, improved transparency, and procedural guidelines and precedents for the exchange of views between leadership and the private sector. The leadership of East Asia generally avoided undertaking major policy directives without first consulting with the major players targeted by the regulation. The councils made information about who gets what from government a matter of record, thereby limiting the opportunities for preferred firms to engage in behind-the-scenes favor seeking or private appropriation of gains from insider information. Ultimately, deliberation councils served as commitment devices, helping to reduce the tendency of government to renege on contracts with private holders of wealth.

A range of institutional innovations promoting broad-based health, education, and housing were designed to make credible the regime's promise of shared wealth and growth. These efforts enabled leadership to gain the cooperation of the citizenry. Businesses invested in fixed capital because the potential risks of social unrest were mitigated. Encouraged by anti-inflationary macroeconomic policies, the wider population saved a considerable portion of their income and invested in education, confident of being able to enjoy the future gains. The introduction of institutions that encouraged equitable growth was an explicit policy choice by government that contributed directly to an improved investment environment.

Sharp contrasts between the successful and less successful bureaucracies in Asia are instructive. The bureaucracies of Japan, Korea, Singapore, and Taiwan, and some bureaus in Malaysia and Thailand, enjoyed merit-based recruitment and promotion, well-defined career paths, competitive salaries relative to the private sector, pension systems, and the possibility of retirement to a board seat in a large private or public enterprise. All but Thailand and Indonesia upheld harsh penalties for corruption that generated (relative) competence and integrity among bureaucrats. Even in the less competent bureaucracies in Indonesia and Thailand, explicit spending limits and a binding commitment to monetary stability limited discretion.

Improving the productivity of the civil service so that it could enforce rules and ensure standards was a first step in the encouragement of private sector confidence among all of East Asia's high performers. The civil service was rewarded for its conformity with agreed-upon rules that circumscribed arbitrary and capricious interventions in its oversight of the economy. Merit supplanted personal patronage so that government would be more responsive to general interests.

Although direct political participation was weak, nonelites in all of the high performers shared the fruits of growth, as evidenced by declining poverty and a reduction in income disparity. In fact, the countries that grew fastest had the most income equality. Wealth-sharing mechanisms created opportunities for the nonelites to pursue productive rather than politically destabilizing activities. Moreover, growth-promoting policies contributed to regime legitimacy, helping to encourage broad social support for the government. While each country determined its own specific pattern of institutional mechanisms and the purposes to which they were devoted, their experience shows that progress toward sustained, self-reliant, and more equitable development depends, in part, on the strength and quality of a country's institutional and organizational capacity. The creation and utilization of predictable procedures distinguishes effective development leadership. The difference between development based on political will and that based on institution building is essentially that the latter is sustainable, while the former is not.

As indicated by the success of East Asia, Western Europe, and North America, strong but limited states have historically had the best record for economic performance. Having defined the character of institutions that promote growth, there remains the conundrum of why only a small minority of world leaders have based their political survival on the competent selection of economic policies.

East Asia's success firmly supports the generalization that governments will realize the goals of economic efficiency only when leadership depends on broad social foundations. The erection of such foundations continues to be the most elusive component of economic growth around the world.

See also: Citizenship, Historical Development of; Civil Society/Public Sphere, History of the Concept; Development and the State; Nation: Sociological Aspects; Nationalism: General; Nationalism, Historical Aspects of: The West; Nations and Nation-states in History; Policy History: State and Economy; Power in Society; Power: Political; State and Society; State: Anthropological Aspects; State Formation; State, History of; Women's Suffrage

Bibliography

Campos E, Levi M, Sherman R 1994 Rationalized bureaucracy and rational compliance, Working Paper 105, Institutional Reform and the Informal Sector. University of Maryland, College Park, MD

Dornbusch R, Edwards S 1991 *The Macroeconomics of Populism in Latin America*. University of Chicago Press, Chicago

Evans P 1992 The state as problem and solution: Predation, embedded autonomy, and structural change. In: Haggard S, Kaufman R (eds.) *The Politics of Economic Adjustment*. Princeton University Press, Princeton, NJ, pp. 139–81

Friedman M, Friedman R 1980 *Free to Choose: A Personal Statement*. Harcourt Brace Jovanovich, New York

Johnson C 1982 *MITI and the Japanese Miracle*. Stanford University Press, Stanford, CA

Klitgaard R 1991 *Adjusting to Reality: Beyond 'State versus Market.'* ICS Press, San Francisco
Shleifer A, Vishny R W 1993 Corruption. *Quarterly Journal of Economics* **108**: 559–617
Wade R 1990a *Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization*. Princeton University Press, Princeton, NJ
Wade R 1990b Industrial policy in East Asia: Does it lead or follow the market? In: Gereffi G, Wyman D (eds.) *Manufacturing Miracles: Paths of Industrialization in Latin America and East Asia*. Princeton University Press, Princeton, NJ, pp. 231–66
Weingast B R 1995 The economic role of political institutions: Market-preserving federalism and economic development. *Journal of Law, Economics and Organization* **1**(April): 1–31
Weingast B R 1997 The political foundations of democracy and the rule of law. *American Political Science Review* **91**(June): 245–63

H. L. Root

States and Civilizations, Archaeology of

1. Civilizations

'Civilization' has had two meanings in reference to ancient societies both of which require critical review. First, civilization has designated complex prehistoric cultures with productive economies, class stratification, and striking intellectual and artistic achievements. Civilization in this sense has contrasted with 'primitive' or 'tribal' societies i.e., societies with modest subsistence economies, egalitarian social relations, and less extravagant art and science. Civilizations have been equated with progress and cultural superiority, a view promoted by European elites of the early modern era. Using ideas of civilization and progress, these elites assumed the role of civilizing agents to legitimate their domination of the working class at home and colonial peoples abroad. Modern anthropologists reject narratives of civilization and progress, preferring less ethnocentric accounts of non-Western cultures. Anthropologists now use more neutral terms to refer to the complex cultures of the ancient world, calling them complex societies or archaic states.

Civilization has also referred to the enduring cultural traditions of the specific regions of the world. Distinctive cultural principles characterized large geographical areas of the world for long periods of time, outliving the expansion and collapse of individual polities. For example, the idea of cyclical time was a fundamental principle among the Olmec, Maya, and Aztec people of ancient Mesoamerica, and it is an idea that persists today among indigenous peoples in Mexico and Guatemala. While contemporary anthropologists recognize the persistence of such orienting concepts, they reject the possible inference that these long-lived ideas imply cultural stasis. In the past as in the present, old concepts were reinterpreted in the face of new situations. Nor do these enduring principles imply uncontested social harmony. People of different social standing might share a set of cultural principles, but still arrive at different conclusions regarding appropriate behavior.

2. States

State has been a more useful term for analyzing complex societies. States are powerful, specialized institutions for public decision-making and control. State government was the most important element of complex ancient societies. The state created a surplus economy that supported many of the other institutions characteristic of complex societies: infra-structural investments, administrative bureaucracies, full-time craft specialization, long distance trade, market exchange, law courts, armies, and priesthoods. Every archaic state had among its key institutions: (a) some form of surplus extraction e.g., corvée, tribute, or taxation; (b) a bureaucracy that supervised the upward flow of tribute payments and the downward flow of tribute demands, and (c) coercive institutions, including armies, law courts, and police, to ensure tribute payment.

Archaeologists once characterized archaic states as having relatively simple systems of class stratification: nobles and commoners, or tribute-receivers and tribute-payers. But more recently, archaeologists have emphasized the diversity of groups within the state, their complex interactions, and the state's chronic instability. Archaic states were constantly threatened by both internal and external forces. High-ranking nobles led factions that competed for positions of rule. Provincial leaders stood ready to declare their independence from state control. Lineages and ethnic groups resisted the state's appropriation of their collective assets. Households tried to evade the state's demands on their goods and labor. The history of states and civilizations is the history of strategy and counter-strategy deployed by oppositional groups, leading to the development of complex societies and their dissolution.

3. Key Characteristics of Archaic States—Economics

Economic activity in archaic states was guided by the principles of efficiency, security, and control. All economic actors took into account the efficiency (e.g., the cost:return ratio) of economic alternatives. But peasants and household-based craft specialists often

sacrificed efficiency for security; that is, they avoided economic alternatives that placed their annual subsistence at risk. States, on the other hand, often sacrificed efficiency for control; that is, they organized some forms of production so as to monopolize output rather than to maximize returns. Peasants, too, were sometimes guided by their desire for control; they were inclined to produce goods whose attributes (e.g., perishability) made state appropriation less likely.

3.1 Agricultural and Pastoral Systems

Archaic states were characterized by highly productive agricultural systems. The vast majority of these states were dependent upon agriculture augmented by some form of water-control, terracing, and/or fertilizer. Mexico is famous for its chinampas (raised fields), Peru for its terrace systems, Mesopotamia for its irrigation canals, and China for its rice paddies. Such systems enhanced both agricultural productivity and reliability.

At one time, archaeologists believed that states came into existence in order to construct and manage intensified agricultural systems as a response to the problem of population pressure (this is the 'hydraulic hypothesis' of state formation). But periods of state formation do not coincide with periods of population pressure, and the study of contemporary irrigation systems in tribal societies suggests that such systems can be built and managed without centralized forms of government.

States did have good reason to construct agricultural facilities. First, such systems provided income to the state. Agricultural facilities constructed with corvée labor represented prime agricultural land that the state could claim as its own without alienating existing property owners. Second, agricultural facilities attracted tax-paying settlers. Peasants were drawn to newly available, high quality land, and they were willing to pay for access to it. Third, agricultural facilities expanded food production within the immediate hinterlands of political capitals, making it possible to feed the urban populace. The desire for more land, greater crop security, and larger quantities of marketable food also led peasant households to invest in agricultural facilities on their own. State-sponsored agricultural investment can be distinguished from peasant investment through careful study: uniformity in the timing and techniques used in facility construction are the best indicators of state sponsorship.

The long-term environmental consequences of agricultural development was a potentially destabilizing factor in all ancient states. In Mesopotamia, irrigation led to the salinization of the soil, first causing agriculturalists to switch from wheat to more salt-tolerant barley and then to abandon salinized areas all together.

3.2 Tribute, Taxes, and Private Estates

All archaic states needed wealth to fund their centralized power. Taxes could be paid in labor, goods, or currency. A tribute in labor was least likely to provoke peasant resistance, since labor could be 'reciprocated' with food and drink, presented to workers by the state in a party-like atmosphere. This form of taxation may characterize the earliest states (i.e., states that still had to contend with kinship groups and kinship norms) or states that legitimated their rule with ideologies of collective wellbeing. A tribute in goods may have its origins in the spoils of war, tribute representing spoils paid on a sustained-yield basis. Agricultural staples and woven cloth were the most common forms of tribute goods, but metals, precious stones, and other exotic valuables were also collected. Tribute foodstuffs were used to feed state personnel. Tribute luxury goods enabled rulers to reward loyal servants and to control the formation of social identities.

Because of their bulk, tribute goods were often stored in provincial warehouses, potentially funding revolts by provincial leaders. Substituting a monetary tax for tribute goods reduced transport costs and provided more centralized control over wealth. But monetary taxation occurred only in conjunction with market systems, where money could be converted into needed goods and services.

Tribute demands often transformed rural landscapes and households. Greek farm families under Roman rule abandoned their rural farmsteads and moved to towns where economic diversification could supplement their incomes and provide the necessary cash for taxes. In the Valley of Oaxaca, the state's demand for maize led to the cultivation of more distant fields which, in turn, altered routines of food preparation. More dry foods (i.e., tortillas) were made so that meals could be carried and eaten away from home.

Peasant households sometimes tried to escape tribute payments. In Mexico under the Aztecs, maguey cultivation expanded, perhaps because its perishable sap was not easily appropriated by the state. In Mesopotamia, peasant families wishing to avoid taxation sometimes fled to the mountains, becoming nomadic pastoralists.

3.3 Urbanism

Archaic states provided the setting for the earliest cities i.e., large, socially-heterogeneous, permanent settlements. These cities were religious and political centers; their economic functions were much less important than is the case for modern cities. Cities housed the administrative machinery of the state, usually in a multi-roomed palace complex. High-ranking nobles, bureaucrats, and military personnel resided in the capital so that their activities could be

directed and monitored by the ruler. Most state capitals also contained shrines, temples, and the residences of high-ranking priests. The concentration of political elites in the capital provided a clientele for enterprising craftspeople, merchants, and other service providers, who swelled the urban populace. Early cities did not supply many of the tools or meager furnishings used by rural peasants; these were usually produced in hinterland towns and villages, by part-time craft specialists.

In several cases, the growth of ancient cities was accompanied by the emptying out of the countryside. Both Uruk (Mesopotamia) and Teotihuacan (Mexico) grew explosively at the expense of rural settlements. In these cities, peasant farmers and farm laborers made up as much as 75 percent of the urban population. Both push and pull factors may have moved farmers from the country to the city. The security, sanctity, and commercial opportunities of the city provided powerful draws; at the same time, rural people may have been forcibly removed to the cities in order to facilitate supervision by the state and to loosen their claims on the land.

Cities developed only minimally in early Egypt. In Egypt, the concentration of population along the banks of the Nile, and the ease of water-bourne communication up and down the river, may have made it unnecessary to concentrate elites and rituals in urban centers.

3.4 Markets

Markets often appear in conjunction with cities. Anthropologists are divided on the issue of whether markets developed from the 'top down' or from the 'bottom up.' 'Top down' theorists postulate that markets developed from the demands of urban elites for food and status goods. 'Bottom up' theorists view markets as developing from the efforts of peasants to enhance the efficiency and/or economic security of their households by participating in specialized production and exchange.

Regardless of their origins, markets attracted the attention of political leaders. Market taxes were an important source of revenue to rulers, and markets provided an efficient means of provisioning cities. States provided both direct and indirect stimuli to market growth. Rulers promoted markets by offering merchants protection against thieves and bandits and maintaining market order. Indirectly, states stimulated market participation by peasant households. Tribute demands forced peasants to become more efficient producers, engaging in specialized production and market exchange. In addition, the social hierarchies that accompanied state formation sometimes prompted peasants to consume specialist-produced goods to enhance their social standing.

In the long run, markets weakened the state's control over the economy. Merchant guilds, formed during periods of state collapse, resisted state control of the economy. The wide array of goods offered for market sale made it difficult to enforce the state's sumptuary rules limiting certain forms of food and dress to state-designated groups. Buying and selling conferred economic power upon individuals who were not state functionaries.

3.5 Full-time Craft Specialists

Many forms of craft specialization antedate the origins of states. In tribal societies, it is not unusual for households or entire villages to specialize in ceramic production, stone tool production, or even metallurgy. This craft activity is carried out on a part-time basis in conjunction with basic subsistence activities. Craft products are distributed in reciprocal exchanges to strengthen social ties and/or to acquire goods from other specialists. Rural, part-time specialization of this sort was also present in archaic states, often enhanced by the opportunities for exchange provided by urban markets.

In situations of incipient social inequality, some craft specialists produced very elaborate goods that required heavy expenditures of labor. These 'hypertrophic' goods (such as embroidered textiles, jade earspools, stone sculptures, heavy bronze vessels, and gold jewelry) were commissioned by aspiring leaders to legitimate their claims to social worth. The craftsmen filling these commissions were part-time specialists either living in their own households or as members of the leader's high-ranking family.

The earliest full-time craft specialists in archaic states were the dependents of state rulers. Rulers supported craft specialists to maintain exclusive control over their output. Controlling sumptuary goods enabled rulers to consolidate their power by rewarding loyal servants and controlling social identities. Thus, for example, the Inka ruler supported a certain number of yanakuna and aqllakuna, full-time specialists in metallurgy and elaborate textile production. Gold and silver jewelry and embroidered textiles were used by the ruler to denote statuses within the state hierarchy. Maya rulers sponsored elite kinsmen who specialized in the production of elaborate painted ceramic vessels. Used in royal funerary rituals, these vessels confirmed the noble status of families. Egypt's pharaohs supported entire towns of craftsmen dedicated to the preparation of royal tombs. Royal patronage accounts for the high levels of skill and artistry that characterize the finest works of archaic states.

Full-time craft specialization also emerged in more commercial contexts when a high demand for craft goods (usually for export trade) combined with developed transportation networks and a steady supply of marketed food. These conditions were met most

fully in the Mediterranean world where the waterbourne transport of food staples from one broad region to another underwrote really large-scale production and exchange. In late precolonial India and China, reliable agricultural systems combined with riverine and oxcart transport yielded similarly high levels of commercial production and exchange.

4. Key Characteristics of Archaic States—Social Organization

Social heterogeneity is the defining characteristic of complex societies. In addition to the occupational diversity described above, the intersection of class, gender, and ethnicity created an array of social statuses and life histories in the ancient world.

4.1 Class Stratification

In the earliest states, the division between classes was based on political rather than economic status. A key division within society was between those who paid tribute and those who received it. Tribute was paid to the ruler who distributed it to servants of the state (usually close relatives). Key personnel were endowed with their own streams of tribute flow. Thus, the tribute-collecting class in archaic states comprised the ruler and his close relatives, high-ranking military personnel, government administrators, priests, and surviving elements of the pre-state nobility. For the governing class, wealth, power, and prestige coincided and were mutually reinforcing. Conversely, tribute-payers suffered a lower standard of living, a lack of formal power, and low social status.

Archaeological and historical research reveals significant variability within the commoner class. For example, Aztec households varied by size, composition, wealth, and ethnic affiliation. Rural households chose different combinations of agricultural production and craft activity. This variability complicates the traditional view that commoners were a homogeneous class. It suggests that commoners were divided by demographic, economic, and social factors.

Slaves constituted a class below commoners. Slaves were captured in warfare, purchased through the slave trade, or sentenced to servitude as punishment for crimes or failed debts. Among Aztecs and Mayas, slaves were used as household servants, but they were never very numerous. In Mesopotamia, large numbers of female war captives worked at textile production, and this may have set a precedent for the use of slave labor in commodity production. The use of slave labor in commercial enterprises was extremely important in highly commercialized economies of ancient Greece and Rome.

Private property sometimes developed out of warfare-based economic systems. In China and Aztec Mexico, lands that were seized in warfare and redistributed to loyal soldiers constituted an early form of private property. In Mesopotamia, private property developed long after states arose, through the gradual sale of lands belonging to kinship units to secular or religious elites.

4.2 Gender

In *The Origin of the Family, Private Property, and the State* (1972 [1884]), Frederick Engels argued that the development of private property brought about the state (to maintain wealth differences) and the patriarchal family (to insure the legitimacy of inheriting offspring). In the process, women's status declined from a position of approximate equality with men to an inferior position. This was the 'world historic defeat of the female sex.' Modern historic and archaeological scholarship suggests that institutions such as private property and the patriarchal family are not universal characteristics of ancient states, nor is the subordination of women. Nevertheless, gender relations and ideologies were frequently renegotiated during the process of state formation, with results that were detrimental to women.

The status of women in states was undermined by several factors. First, states were commonly created and maintained through warfare, and war was usually men's work. Through warfare, men gained access to spoils, a form of wealth that was not household-dependent. In addition, rulers often glorified male warriors and rewarded them with administrative positions in order to maintain their loyalty. Thus, warfare provided avenues of wealth, prestige and power for men that were not available to women. Furthermore, rulers dismantled kinship groups, which were nodes of resistance to state rule. These kinship groups often provided the primary rationale and defense of women's rights in ancient societies, and when they were destroyed, women were left without social recourse. It is not surprising, therefore, that many archaic states favored men. Ancient Rome and ancient China both awarded great authority to male household heads. The Classic Maya and the Aztecs conferred public prestige upon male warriors, and men dominated these state governments.

However, the subordination of women to men was not universal. States that legitimated their rule with ideologies of collective wellbeing did not emphasize warfare and did not stress gender difference. These states might have included Teotihuacan (Mexico), the Harappan states of the Indus Valley, and Egypt under the pharaohs. In the Andes, a long tradition of parallel descent favored gender equality. Even in states where patriarchy was the norm, high-ranking women were wealthier and more powerful than lower class men (i.e., class trumped gender). In peasant families, women's contribution to the household economy often translated into approximate gender equality. In ad-

dition, women found forums for resistance outside of the state. Female oracles provided social critique in Oaxaca Mexico, ancient Greece, and the Andes.

4.3 Ethnicity

Ethnicity is an identity based upon a presumption of shared history and common cultural inheritance. Ethnic identity is shaped by both ethnic affiliation and ethnic attribution. Ethnic affiliation refers to individuals' own sense of group membership and the characteristics of the group as defined by its members. Ethnic attribution concerns the characteristics of the group as defined by outsiders.

States acted opportunistically and inconsistently in dealing with ethnicity. Sometimes they suppressed ethnic affiliation to weaken the resistance of subject groups to the state. At other times they encouraged ethnic affiliation to accentuate division within the commoner class. States also used derogatory ethnic stereotypes to legitimize their exploitation of subject peoples. For example, Aztecs depicted their Otomí subjects as lazy, improvident, untrained blockheads and gaudy dressers. Stereotyping and ethnic prejudice generally heightened ethnic consciousness and perpetuated ethnic heterogeneity within ancient states.

5. Key Characteristics of Archaic States—Religion and World Views

5.1 State Religion

All archaic states had an official religion with standardized temples and full-time priests. Rulers' patronage of religion may have sprung from either piety or cynical efforts to manipulate the subordinate class. Either way, religion provided a powerful political resource that was actively exploited by ancient rulers.

Many archaeologists argue that state religion was used to forge a unified polity out of distinctive subject peoples and to legitimize the ruler's authority. For example, Chinese rulers tried to force cultural uniformity upon all the groups that they conquered, and Confucian philosophers declared that just rulers held the 'Mandate of Heaven.' The Incas' effort to impose cultural uniformity on their subjects is archaeologically visible in provincial capitals where large plazas were constructed using Inca architectural forms. Local people gathered in these plazas to attend state-sponsored rituals.

In contrast, Aztec state religion was not geared to produce a uniform world view within the empire, nor to sanctify Aztec rulers in the eyes of their subjects. Rather, Aztec religion was intended to win the loyalty of a relatively small target group, the young men who formed the core of the Aztec army, by imbuing their military exploits with cosmological significance. In the Aztec hinterlands, distinctive forms of ceramic figurines suggest that commoners did not accept the state's ideology; on the contrary, commoners successfully maintained their own world view. Distinctive commoner ideologies may have been present in all archaic states, challenging the official point of view.

5.2 Monumental Architecture

The earliest monuments were megalithic tombs and earthworks built by groups that were still organized using kinship principles. These monuments were probably constructed during communal rituals, when aspiring leaders sought social prominence by sponsoring group festivities. Under state rule, monumental construction increased in scale, creating some of the most striking icons of the ancient world: Egypt's pyramids, Mesopotamia's ziggurats, the colossus of Zeus at Rhodes. It is estimated that the great Pyramid of Khafu (Cheops) required 13,440,000 person-days of labor.

Some monuments may have been constructed by a willing populace, motivated by civic pride and religious devotion. This would have been most likely in states that legitimated their rule with ideologies of collective wellbeing. At Teotihuacan, for example, the Temple of the Sun and the Temple of the Moon may have been constructed in a surge of popular enthusiasm. In states relying upon coercive rule, monument construction expressed submission to the state. Aztec rulers invited subject communities to participate in the construction of the Great Temple in Tenochtitlan, and refusals to participate were considered acts of rebellion.

Through their size, monumental structures symbolized the ruler's exclusive link to supernatural forces. Monumental structures also expressed overwhelming power of the state and the fact of commoners' subordination. Large buildings and massive sculptures imposed themselves on social space and human psyches. In the words of Antonio Gilman (1996), they showed 'who was boss.'

5.3 Writing

Writing developed in archaic states for many different reasons. The earliest texts, from Mesopotamia at 3300 BCE, recorded the economic transactions of large temple estates. In ancient Egypt and Mesoamerica, writing was first used to record the exploits of rulers. The earliest Chinese inscriptions were questions carved on turtle carapaces and the shoulder blades of cattle for the purposes of divination. In all cases, writing was an elite activity. Writing can be regarded the product of elite efforts to establish and to preserve a single, authoritative record of events and to make its voice stronger than the spoken words of its subjects.

6. The Rise and Fall of Archaic States

In the 1970s, archaeologists believed that states arose to manage environmental problems. Since states are powerful administrative institutions, state officials can function as highly effective problem solvers. They can construct large-scale irrigation networks, insure populations against crop failure, administer markets, organize warfare, and sponsor long-distance trade. Archaeological evidence confirms that archaic states did oversee some of these activities. However, the evidence for these activities usually dates to a few centuries after state formation; thus, the problem-solving capacity of states does not appear to account for their origins.

More recently, archaeologists have begun to account for state origins by examining how individuals acquire social power. In one model, dynamic, ambitious leaders assemble coalitions of followers by sponsoring feasts and distributing gifts. Each leader with his followers then competes against other leaders and followers to establish dominance, first on the local, then on the regional, level. In these systems, personal prestige is emphasized, and dominance may be maintained through naked force. In a second model, leaders attract a broader following, by appealing to ideologies of collective wellbeing. State formation is more 'volunteeristic,' resembling a religious movement, and there is broad-based participation in collective rituals and public monument-building. Leaders are more anonymous, and when force is used, it is used on behalf of society as a whole. It has been suggested that the flamboyant Maya and the Shang rulers might exemplify the outcome of the first (individualizing) strategy, whereas the Teotihuacan (Mexico) and Harappan states might be products of the second (corporate) strategy.

Archaic states were complex, often fragile, social structures. Social dominance required balancing the needs and power of diverse social segments. Sudden shocks and long-term problems threatened the delicate balance. Over-population, excessive tax burdens, climatic deterioration or invading barbarians, all threatened to topple archaic states at one time or another. Egypt's Old Kingdom collapsed when provincial officials became increasingly independent of the pharaoh's rule; the Middle Kingdom fell because bureaucrats at the center absorbed too much of the state's wealth. Competition among the independent Classic Maya kingdoms led the Maya to over-exploit their resources and, finally, to abandon their homes. The surprise is not that archaic states eventually collapsed, but that ruling elites could devise institutional structures that enabled states to endure, sometimes for centuries at a time.

See also: Agricultural Change Theory; Civilization, Concept and History of; Civilizational Analysis, History of; Civilizations; Intensification and Specialization, Archaeology of; Pastoralism in Anthropology; State, History of; Trade and Exchange, Archaeology of

Bibliography

Alcock S E, D'Altroy T N, Morrison K D, Sinopoli C M (eds.) 2001 *Empires: Perspectives from Archaeology and History*. Cambridge University Press, Cambridge, UK

Blanton R E, Feinman G M, Kowalewski S A, Peregrine P N 1996 A dual-processual theory for the evolution of Mesoamerican civilization. *Current Anthropology* **37**: 1–14

Childe V G 1950 The urban revolution. *Town Planning Review* **21**: 3–17

Classen H J M, Skalník P (eds.) 1978 *The Early State*. Mouton, The Hague

DeMarrais E, Castillo L J, Earle T 1996 Ideology, materialization, and power strategies. *Current Anthropology* **37**: 15–31

Engels F 1972 [1884] *Origin of the Family, Private Property, and the State*. Leacock E (ed.). International Publishers, New York

Feinman G M, Marcus J (eds.) 1998 *Archaic States*. School of American Research, Santa Fe, NM

Gilman A 1996 Comments on 'Ideology, materialization, and power strategies.' *Current Anthropology* **37**: 56–7

Gledhill J, Bender B, Larsen M T (eds.) 1988 *State and Society: The Emergence and Development of Social Hierarchy and Political Centralization*. Unwin Hyman, London

Patterson T C 1997 *Inventing Western Civilization*. Monthly Review Press, New York

Richards J, Van Buren M (eds.) 2000 *Order, Legitimacy and Wealth in Ancient States*. Cambridge University Press, Cambridge, UK

Sanders W T, Webster D 1988 The Mesoamerican urban tradition. *American Anthropologist* **90**: 521–46

Silverblatt I 1988 Women in States. *Annual Review of Anthropology* **17**: 427–60

Stein G J 1998 Heterogeneity, power, and political economy: Some current research issues in the archaeology of Old World complex societies. *Journal of Archaeological Research* **6**: 1–44

Yoffee N, Cowgill G L (eds.) 1988 *The Collapse of Ancient States and Civilizations*. University of Arizona Press, Tucson, AZ

E. M. Brumfiel

Statistical Analysis: Multilevel Methods

Multilevel, or hierarchical, data are blocked or clustered. Children within families; individuals within organizational units; longitudinal data on individuals; repeated measures on individuals; organizational units within larger organizations; multistage sample surveys; and time series of cross-sections exemplify multilevel or hierarchical data structures. When data are clustered, each observation counts for less than it would if the data points were independent. If clustering is ignored, statistical inference can be faulty. With

clustering, regression methods that assume independent observations provide inefficient and possibly biased coefficient estimates, and associated confidence intervals that are too narrow. In the context of regression modeling, any method that attempts to take account of clustering is multilevel.

Several generalized regression methods are commonly used in the analysis of hierarchical data. Hierarchical or random coefficient models are multilevel. So too are fixed effects models, marginal models, classical single-level estimation followed by regression coefficient covariance matrix adjustment, and hybrid approaches. The choice of an approach for a particular research problem must depend on the goals of the research and assessment of the design and measurement deficiencies presented by a particular body of data, and is often consequential. Multilevel modeling approaches emphasize different aspects of the data, vary in the nature and stringency of their assumptions, and support different kinds of descriptions and interpretations. Poor choice of method can lead to the application of an incorrectly specified model, and thereby to good estimates of the wrong quantities.

1. *Random Coefficient Models*

The random coefficient model offers a compelling formulation that is consistent with the social scientific goal of understanding how units at one level affect, and are affected by, units at other levels. Consider a design with two levels, and suppose that there is a single level-1 regressor (X), a single level-2 regressor (G) that is by definition *constant* over all level-1 observations within a context, and that the dependent variable (Y) is Gaussian distributed conditional on X and G. The random coefficient model can be expressed as

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij} \tag{1}$$

with

$$\begin{aligned} \beta_{0j} &= \eta_{00} + \eta_{01} G_j + \alpha_{0j} \\ \beta_{1j} &= \eta_{10} + \eta_{11} G_j + \alpha_{1j} \end{aligned} \tag{2}$$

where $j = 1, 2, \ldots, J$ denotes contexts (level-2 units), and $i = 1, 2, \ldots, n_j$ denotes individuals within contexts (level-1 units). The level-2 equations (Eqn. (2)) assert that the level-1 intercepts and slopes vary over contexts as linear functions of G. There is often justification for supposing that the within-context coefficients depend on contextual characteristics. It is appealing that the dependence includes random error components (α_{0j} and α_{1j}).

Assumptions about the error structure of this two-level model are that, for each context, the level-1 errors follow the Gaussian distribution, are centered on zero, have constant variance, and are independent. The error variances may vary over contexts. The level-2 errors are also Gaussian distributed, centered on zero, independent within equations, and have constant variances. The level-2 errors may be correlated across equations. It is further assumed that the level-1 errors are orthogonal to the level-2 errors, and that all error terms are orthogonal to X and G. The orthogonality assumptions are needed to secure consistent estimates of the η's.

Many generalizations of this model are possible—coefficients can be constrained to be nonrandom; the dependent variable may be discrete, counted, truncated, ordinal, or time to event; there may be more regressors at any level as well as more than two levels; the nesting of levels may be cross-classified; different assumptions about the functional forms of the distributions of the errors can be made; and the assumption of independence of errors within level-2 (or higher) units can be relaxed. Nevertheless, the assumptions of orthogonality between regressors and errors, and of cross-level orthogonality between error components, continue to remain important. Until that is no longer so, the credibility of an application of a random coefficient model rests on the plausibility of these assumptions, which should be thought of as inextricably linked to the systematic components of the model, and not as the fine print of a contract that is never read, much less enforced.

However compelling it may be on substantive grounds to specify the random coefficient multilevel model as Eqns. (1) and (2), substitution of Eqn. (2) into Eqn. (1) results in the equivalent mixed-model representation

$$\begin{aligned} Y_{ij} = \eta_{00} + \eta_{01} G_j + \eta_{10} X_{ij} + \eta_{11} X_{ij} G_j \\ + (\alpha_{0j} + \alpha_{1j} X_{ij} + \varepsilon_{ij}) \end{aligned} \tag{3}$$

which demonstrates that treating intercepts and slopes as dependent on G is the same as supposing that Y depends on X, G, and their cross-level (product) interaction.

Random coefficient models with Gaussian errors are commonly estimated by maximum likelihood or restricted maximum likelihood. For binary response the prevalent methods are marginal and penalized quasi-likelihood, although Bayesian methods based on Markov Chain Monte Carlo (MCMC) estimation are also beginning to be used. A consensus on preferred methods of estimation of multilevel binary response and other nonlinear (non-Gaussian) models has yet to emerge. Estimation of multilevel models is iterative and computationally intensive, more so for non-Gaussian cases. Especially for these cases, researchers are well advised to check numerical estimates across algorithms, methods, and software. Other problems, such as substantial design imbalance in combination with context size, also need to be eval-

uated for their potential impact on numerical estimates (see *Estimation: Point and Interval*; *Hierarchical Models: Random and Fixed Effects*; *Meta-analysis: Overview*; *Markov Chain Monte Carlo Methods*).

2. Fixed Effects Models

The fixed effects model does not require the assumption of orthogonality between X and α_0 inherent to the random coefficient model. It can be written as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ij} G_j + \delta_j + u_{ij} \quad (4)$$

where the δ_j are between-context contrasts, and the u_{ij} are assumed to be classical Gaussian errors. Relative to the random coefficient model, the additive term in G has been dropped, and the intercept and slope are no longer allowed to vary randomly over contexts, although each context is allowed its own intercept and the cross-level interaction is retained. The δ_j absorb the additive component of *all possible* G-variables, not just the finite number that can be included in Eqn. (3). Additive components for any level-1 regressors that do not vary within contexts are also absorbed by the δ_j.

In the fixed effects approach, 'fixing' the between-context contrasts means that the contexts are themselves assumed to be fixed, in the sense that statistical inference is conditional on the particular selection of contexts present in the data. Because the contrasts can be defined as the coefficients of a set of indicator variables that assign individuals to contexts, there is no need to assume that context membership is orthogonal to X. This is a critical difference between the fixed effect and random coefficient models. If the assumption of orthogonality between X and α_0 is implausible in the random coefficient model, and it can be when the list of G-variables is incomplete, then the fixed effects model provides an alternative that solves the problem. In the Gaussian case, the fixed effects model is a conventional regression model. When J is large—and often when it is not—there may be little interest in describing estimated values of the δ_j, in which case estimation of the other covariate coefficients can be performed after within-context centering of the variables. For a binary response variable, conditional logistic regression can be used to estimate fixed effect models (Chamberlain 1980). In this case, the context contrasts are not estimated, although additive context differences are controlled. Software for fixed effects estimation is widely available.

If it is desired to obtain estimates of the additive component of the contextual variables, then the fixed effects approach is not the method of choice. Also, in the binary response case, conditional logistic regression does not necessarily use all of the data, and in this sense may be inefficient (see *Economic Panel Data*; *Linear Hypothesis: Regression (Basics)*; *Multivariate Analysis: Discrete Variables (Logistic Regression)*).

3. Coefficient Variance Estimates Adjusted for Clustering

The adjusted coefficient variance estimator in common use for multilevel analysis is a robust estimator. In actual applications, care must be taken to ensure that the robust estimator selected includes an extension of the algorithm for the general robust estimator that accommodates clustering. With the robust estimator, it is possible to specify a model whose systematic component is identical to that of Eqn. (3), and whose error component need not be independently and identically distributed, Gaussian distributed, or orthogonal to the regressors. The robust equivalent of Eqn. (3) can be expressed as:

$$Y_{ij} = b_0 + b_1 G_j + b_2 X_{ij} + b_3 X_{ij} G_j + e_{ij} \quad (5)$$

Least-squares computation can be used to produce coefficients whose estimated variances and covariances are subsequently adjusted for context membership. The coefficients will be unbiased estimates of the corresponding terms in Eqn. (3) if the orthogonality assumption is satisfied. The standard errors will be adjusted for clustering and other sources of heteroscedasticity regardless of whether this is so. When Eqn. (3) is true, confidence intervals for its coefficients obtained by prevailing methods will tend to be smaller than confidence intervals based on the robust clustering adjustment associated with Eqn. (5).

The robust approach is also an alternative to the fixed effects approach. The latter achieves orthogonality between the regressors and the error component by including context membership in the systematic component of the model, but is unable to include main effects of contextual variables even though it controls for all possible contextual variables. In the robust approach coefficients of contextual regressors are estimated, but all possible contextual variables are not controlled. Coefficient standard errors are adjusted for clustering, but the coefficient estimates are not efficient.

The robust, variance adjustment approach to multilevel analysis has been extended to models of discrete and counted outcomes, and to survival problems, and can be used where a fixed effects approach is not available. Nevertheless, where it can be justified, explicit incorporation of clustering into model specification is preferable to *post hoc* adjustments of estimators that ignore clustering (see *Sample Surveys: The Field*; *Sample Surveys: Survey Design Issues and Strategies*; *Sample Surveys: Model-based Approaches*).

4. Marginal Models

When Y is a discrete or count response variable, it is possible to specify a multilevel model in which the within-context correlatedness of observations is in-

corporated as a side condition on the model, rather than as a consequence of a random intercept. For Y binary and ω_{ij}, the logit of 'success' for the ijth observation, the multilevel dependence of Y on regressors, can be written as

$$\omega_{ij} = \beta_0 + \beta_1 G_j + \beta_2 X_{ij} + \beta_3 X_{ij} G_j \quad (6)$$

with an additional specification indicating the pattern of correlation between observations within a context (corr(Y_{ij}, $Y_{i'j}$), $i \neq i'$). Specifying that all pairs of observations within a context are equally correlated corresponds to the assumption of correlatedness induced by a random coefficient model, but virtually any pattern can be specified.

Models that characterize the marginal distribution of Y as a function of regressors are said to be marginal models. The model of Eqn. (6) and its side condition on correlated observations is a member of this class. The generalized linear model (McCullagh and Nelder 1989), a unifying theoretical framework for the linear model and discrete and count response variables, has been extended to allow for the analysis of multilevel data (Liang and Zeger 1986, Zeger and Liang 1986). This extension, known as the generalized estimating equations approach (GEE), provides a marginal model (population-average) alternative to random coefficient models when Y is discrete or a count. Equation (6) can be estimated using the GEE method.

Gaussian random effects models sustain a marginal model interpretation of regression coefficients. Random effects models for discrete and count response variables do not. Marginal models are appropriate when the analytic goal is to infer how Y depends on covariates for a population of level-1 units averaged over contexts. Random effects models are appropriate when within-context estimates are desired. For the logistic response case, the estimated coefficients of marginal models tend to be smaller in absolute value than the corresponding coefficients in random effects models (Diggle et al. 1994, p. 141).

Marginal models for Y discrete or a count can be estimated for a broad range of structures of potential within-context correlation; random coefficient models for discrete and count response variables currently cannot. Computation is fast (iterative generalized least squares), and estimation is relatively straightforward compared to binary response random coefficient models, although issues surrounding optimal characterization of within-context correlation and estimation of coefficient standard errors continue to attract attention.

5. *Omitted Variables*

Omitted variables in random coefficient models are of particular concern when they are endogenous, that is, correlated with Y and the included regressors, because they induce coefficient bias. The fixed effect approach provides one solution to this problem when the omitted variables are contextual. There are others: measurement of omitted variables, random assignment to contexts, and instrumental variables estimation.

Measuring the omitted variables is not an option in secondary analysis or in the midst of primary analysis—by then the data have been collected and returning to the field is generally infeasible. This should not, however, preclude future attempts to extend the range of measurement.

Random assignment to contexts presumes that an experimental design is feasible in the study of a particular phenomenon. Often it is not. Even when randomization is possible, it is not necessarily well suited to answering certain kinds of questions (see *Experimental Design: Randomization and Social Experiments*; *Quasi-Experimental Designs*; *Experimental Design: Large-scale Social Experimentation*).

Instrumental variables estimation is often used in an effort to solve an endogeneity problem. When an instrumental variables approach is used, aspects of the clustering problem may be sacrificed to solve what is perceived as a much more important one—selection into contexts. In Eqn. (3) replace X by T, suppose that the underlying data structure is quasi-experimental, and let $T_{ij} = 1$ if an individual is treated (0 otherwise). If Y is a health outcome and contexts are hospitals, and if assignment to treatment is not random, then ε potentially contains unmeasured variables correlated with T and also with Y. In quasi- and nonexperimental data, whether a patient receives a particular kind of treatment can be nonrandom with some causes known, others not. An instrumental variables approach to this problem attempts to find one or more covariates W correlated with T and Y, where Y and W controlling T are uncorrelated, and W and ε are uncorrelated. It is often difficult to find plausible instruments (W), although contextual variables are sometimes used. Thus, instrumental variables estimation cannot be employed with the automaticity of any of the other approaches to clustered data.

6. *Examples*

An example of a Gaussian application, and another of a binary response application, illustrate the kinds of differences in findings that can occur using alternative multilevel approaches. The data are from a one percent clustered sample of the 1990 Census of China, and are specific to Hainan province. Clustering is at the enumeration district level, and within clusters enumeration is 100 percent.

6.1 *Gaussian Example*

In the Gaussian example, there are 17,662 men in the data, distributed over 332 clusters, the smallest having 114 men and the largest having 3,735. The occupational socioeconomic status (SEI) of men ages 20–65

Table 1
Alternative multilevel estimates, occupational socioeconomic status of men

	Regressions		
Covariates	Ordinary least squares Unadjusted/Robust/Cluster (1)/(2)/(3)	Fixed effect (4)	Random coefficient (5)
Age	0.09	−0.01	0.02
	(7.78)/(7.92)/(1.32)	(−0.91)	(0.50)
Education	0.34	0.36	0.43
	(7.29)/(6.00)/(1.00)	(6.96)	(2.36)
Rural	−4.92		−0.72
	(−18.47)/(−13.50)/(−1.01)		(−0.35)
Age·CED²	0.31	0.66	0.55
	(13.16)/(13.03)/(1.72)	(17.83)	(5.75)
Ed·CED²	2.67	2.75	2.38
	(25.51)/(21.78)/(3.06)	(20.81)	(4.72)
Intercept	15.50	11.48	12.89
	(33.08)/(32.15)/(4.63)	(29.42)	(6.71)
R^2	0.45	0.57	

Source: One percent sample, males ages 20–65, Hainan Province, 1990 Census of China.
$N = 17{,}662$. Quantities in parentheses are ratios of coefficients to standard errors. The coefficients of Regressions (1)–(3), estimated by ordinary least squares, are identical. The standard errors of (1) are unadjusted; those of (2) are adjusted, but without taking explicit account of clustering; those of (3) are adjusted for cluster membership. The context contrasts in the fixed effects regression (4) are not shown. The level-2 error variances in the random coefficient regression (5), not shown, are significant.

is a function of age and years of schooling at the individual level, type of place of residence (rural vs. urban), and the square of a contextualized measure of education. Type of place of residence is a characterization of place, and constant within contexts. Substantive theory and exploratory data analysis lead to this random effects model:

$$\begin{aligned} SEI &= \beta_{0j} + \beta_{1j} Age + \beta_{2j} Ed + \varepsilon \\ \beta_{0j} &= \eta_{00} + \eta_{01} Rural + \alpha_{0j} \\ \beta_{1j} &= \eta_{10} + \eta_{11} CED^2 + \alpha_{1j} \\ \beta_{2j} &= \eta_{20} + \eta_{21} CED^2 + \alpha_{2j} \end{aligned} \tag{7}$$

Table 1 reports coefficient estimates and ratios of coefficents to standard errors ('t-ratios') for several approaches. Regressions 1–3 are the same ordinary least squares (OLS) regression, differentiated to keep track of the type of standard error. When clustering is ignored, the t-ratios are far larger (regressions 1 and 2). Unstructured adjustment of standard errors makes little difference. Adjusting OLS standard errors for clustering radically reduces the t-ratios (regression 3). The fixed effects coefficients (regression 4) are roughly the same as the OLS coefficients, with larger t-ratios than obtained with the clustering adjustment. Only one term is 'significant' when the clustering adjustment is used, whereas three of the four estimable coefficients are highly 'significant' in the fixed effects model. The coefficients of the random effects model (regression 5) are comparable with those for the fixed effects model (where present), but the t-ratios are smaller. For those terms present in both the fixed effects and random coefficient models, coefficients are either significant in both models, or not significant in both. In computations not displayed here, the random coefficient specification was computed twice within two different programs, once using maximum likelihood and once using restricted maximum likelihood. There was agreement across programs and methods. Although the basic conclusion is the same for the fixed effects and random coefficient models and differs somewhat from that provided by the clustering adjustment, the variation in coefficient magnitudes and t-ratios across the models is cautionary. Even in the Gaussian case, it can be misleading to estimate a model using only one approach.

6.2 *Binary Response Example*

In the binary response example, data from the 1990 Hainan sample are again used. Infants and toddlers born in the 18 months prior to the starting date of the Census enumeration are the level-1 units of observation, and enumeration districts are level-2 observations. The available data do not record child mortality, which has been inferred from maternal information elicited in the Census questionnaire. The response variable is whether an imputed death to a child born after 1988 occurred prior to July 1, 1990. There are 43 imputed deaths out of 2,572 mother–child pairs. A majority of enumeration districts have no imputed deaths. For this reason, only the within-context intercept is treated as random. For ω, the logit of a death within the first 18 months of life, the specifi-

Table 2
Binary response example, imputed infant mortality

	Regression				
				Random coefficients	
Variables	Classical cluster adj.	GEE	Fixed effects	PQL	MCMC
	(1)	(2)	(3)	(4)	(5)
Maternal Ed	−0.25 (−2.62)	−0.284 (−2.93)	−0.27 (−2.77)	−0.29 (−3.19)	−0.32 (−3.56)
Male	−0.96 (−1.26)	−1.18 (−1.95)	−1.25 (−2.13)	−1.20 (−2.07)	−1.43 (−2.55)
Ed · Male	0.24 (1.73)	0.29 (2.48)	0.30 (2.63)	0.29 (2.60)	0.32 (2.84)
Rural	2.07 (2.17)	1.99 (1.48)		1.81 (2.03)	1.38 (1.38)
Intercept	−4.74 (−4.71)	−4.74 (−3.50)		−4.75 (−5.07)	−4.74 (−4.24)
Level-2 Var.				0.85 (1.78)	2.05 (1.53)

Source: One percent sample, infants to age 18 months paired with mothers, Hainan Province, 1990 Census of China.
$N = 2{,}572$. Quantities in parentheses are ratios of coefficients to standard errors.

cation treats the probability of death as a function of maternal years of schooling, the sex of the child, and the interaction between maternal education and sex of child. The intercept is treated as a function of type of place of residence (the same for all residents of an enumeration district) and a random error:

$$\omega_{ij} = \beta_{0j} + \beta_1 Ed + \beta_2 Male + \beta_3 Ed \cdot Male$$
$$\beta_{0j} = \eta_{00} + \eta_{01} Rural + \alpha_{0j} \qquad (8)$$

Table 2 presents a selection of regressions based on alternative multilevel approaches and algorithms. Regression 1 is the result of classical maximum likelihood estimation followed by an adjustment for clustering. Using this approach, neither the sex of the child nor the sex–education interaction is 'significant.' Using the GEE approach, the coefficient for sex is at the cusp of conventional significance and the sex–education interaction is significant (regression 2). The fixed effects results, using conditional logistic regression, indicate significance for all included terms (regression 3). Penalized quasi-likelihood estimates are provided by regression 4. These are consistent with the fixed effects results for those terms present across the fixed and random effects specifications. Regression 5 presents Bayesian results based on MCMC. The type of place of residence coefficient is smallest and not significant in the Bayesian regression. Application of a clustering adjustment to the GEE results (not shown) also substantially affects conclusions about statistical significance.

It is not unusual to encounter sparse data. This binary response example is perhaps even more cautionary than the Gaussian example. Researchers are not yet routinely able to determine whether convergence is to a local or a global maximum, whether sufficient iterations have been calculated using MCMC, whether all of the assumptions required for a particular method are valid in a given instance, or whether software is performing as intended.

7. *Conclusions*

Analysis of multilevel data structures using methods that ignore the clustering of observations in the data is unsatisfactory because it can lead to confidence intervals that are too narrow, and to coefficient bias. Random coefficient models—and their specializations and generalizations—are attractive tools for the analysis of multilevel data. When there is selection into contexts, the random coefficient approach is subject to coefficient bias, as it is more generally when there are omitted regressors at any level. The fixed effects approach provides one kind of solution to the omitted variables problem. Instrumental variable estimation offers another. A posterior adjustment to accommodate clustering, applied to the estimated variances of coefficients, is robust but inefficient when the random effects model is correct. Although this adjustment does not require orthogonality between errors and covariates, robust standard errors for biased coefficients are far from universally desirable. For discrete and count responses, marginal models estimated using GEE provide an alternative to random effects models. Marginal models are attractive when there is little interest in within-context estimates, or when the pattern of within-context correlation is not simple. The choice of a multilevel modeling approach is ideally governed by research goals, the study design, and the nature of the data. There is also value in cross-checking results across methods and software.

8. *Further Reading*

For an introduction to random coefficient models (terminology often used interchangeably with random effects models and hierarchical models), see Snijders and Bosker (1999). The level of exposition in Bryk and Raudenbush (1992) ranges from introductory to advanced, as is also true for Goldstein (1995). The

exposition in Longford (1993) is technically demanding. Littell et al. (1996) survey mixed models in general, and include material on hierarchical models. Gilks et al. (1998) and Gelman et al. (1995) are useful for Bayesian random coefficient models. Draper (2002) provides a detailed introduction to random coefficient models from the Bayesian perspective.

Fixed effects modeling is treated in econometrics texts (e.g., Greene 2000), and is often discussed as one approach to the analysis of panel data (Baltagi 1995). For the binary response case, see Pendergast et al. (1996), who provide a review with broad scope. Robust adjustments of standard errors for clustering (e.g., Rogers 1993) are an extension of the methodology of Huber (1967) and White (1980, 1982). The GEE approach is well exposited in Diggle et al. (1994), who provide detailed comparisons with random coefficient and other models. In this same vein, Allison's (1999) introduction should not be missed. Instrumental variables estimation is a standard subject in econometrics texts (e.g., Greene 2000).

For an illuminating use of random coefficient models in the study of neighborhoods see Sampson et al. (1997). Geronimus and Korenman (1992) use a fixed effects approach in the study of teen pregnancy. Hamilton and Hamilton (1997) use a fixed effects approach in the analysis of surgery outcomes. For the same kind of problem, but using a different data structure, McClellan et al. (1994) employ an instrumental variables approach. Applications of marginal models are found largely in the epidemiological and biomedical literature. Multilevel analysis has long been mainstream in economics, and will soon become so in the rest of the social sciences.

9. *Software*

Dedicated random coefficient packages include HLM5 (Raudenbush et al. 2000) and MLwiN (Goldstein et al. 1998). HLM5 additionally offers GEE estimation; MLwiN includes Bayesian MCMC estimation. Most major statistical packages include facilities for all of the approaches discussed here, as integral components or contributed macros. Software for multilevel analysis is in continuous development. The multilevel models mail list is an essential portal for current information (multilevel@jiscmail.ac.uk), as is the website hosted by the Multilevel Models Project (http://www.ioe.ac.uk/multilevel/).

Bibliography

Allison P D 1999 *Logistic Regression Using the SAS System*. SAS Institute, Cary, NC

Baltagi B H 1995 *Economic Analysis of Panel Data*. Wiley, New York

Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models*. Sage, Newbury Park, CA

Chamberlain G 1980 Analysis of covariance with qualitative data. *Review of Economic Studies* **47**: 225–038

Diggle P J, Liang K-Y, Zeger S L 1994 *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK

Draper D 2002 *Bayesian Hierarchical Modeling*. Springer-Verlag, New York

Gelman A, Carlin J B, Stern H S, Rubin D B 1995 *Bayesian Data Analysis*. CRC Press, Boca Raton, FL

Geronimus A T, Korenman S 1992 The socioeconomic consequences of teen childbearing reconsidered. *Quarterly Journal of Economics* **107**: 1187–214

Gilks W R, Richardson S, Spiegelhalter D J 1998 *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London

Goldstein H 1995 *Multilevel Statistical Models*, 2nd edn. Halstead, New York

Goldstein H, Rasbash J, Plewis I, Draper D, Browne W, Yang M, Woodhouse G, Healy M 1998 *A User's Guide to MLwiN*. Multilevel Models Project, Institute of Education, University of London, London

Greene W H 2000 *Econometric Analysis*, 4th edn. Prentice-Hall, Upper Saddle River, NJ

Hamilton B H, Hamilton V H 1997 Estimating surgical volume—outcome relationships applying survival models: Accounting for frailty and hospital fixed effects. *Health Economics* **6**: 383–95

Huber P J 1967 The behavior of maximum likelihood estimators under nonstandard conditions. In: LeCam L M, Neyman J (eds.) *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA, Vol. 1, pp. 221–33

Liang, K-Y, Zeger S L 1986 Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22

Littell R C, Milliken G A, Stroup W W, Wolfinger R D 1996 *SAS System for Mixed Models*. SAS Institute, Cary, NC

Longford N T 1993 *Random Coefficient Models*. Oxford University Press, New York

McClellan M, McNeil B J, Newhouse J P 1994 Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? *Journal of the American Medical Association* **272**: 859–66

McCullagh P, Nelder J A 1989 *Generalized Linear Models*. Chapman and Hall, New York

Pendergast J F, Gange S J, Newton M A, Lindstrom M J, Palta M, Fisher M R 1996 A survey of methods for analyzing clustered binary response data. *International Statistical Review* **64**: 89–118

Raudenbush S, Bryk A, Cheong Y F, Congdon R 2000 *HLM 5: Hierarchical Linear and Nonlinear Modeling*. Scientific Software International, Lincolnwood, IL

Rogers W H 1993 Regression standard errors in clustered samples. *Stata Technical Bulletin* **13**: 19–23

Sampson, R J, Raudenbush, S W, Earls F 1997 Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* **277**: 918–24

Snijders T A B, Bosker R 1999 *Multilevel Analysis*. Sage Publications, London

White H 1980 A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**: 817–38

White H 1982 Maximum likelihood estimation of misspecified models. *Econometrica* **50**: 1–25

Zeger S L, Liang K-Y 1986 Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**: 121–30

W. M. Mason

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Statistical Analysis, Special Problems of: Grouped Observations

Often data are available to social and behavioral scientists only in grouped form, i.e., data presented in tables or using display techniques such as frequency polygons or histograms. Reasons for having the data only in grouped form might be because of the confidentiality of data or because access to the individual observations might be too costly (e.g., data suitable for an international comparison might be found in tabular form in the United Nations Demographic Yearbook, but to collect the original data even if it is available will be costly and time consuming) or because the data sets are vast and it is desirable to group it in order to economize on data handling and computation. In the last two cases the researchers must judge the cost of obtaining the original data against the loss of precision in estimating from grouped, relative to ungrouped, data.

The literature on grouped observations can be divided into three broad categories. The first is inference made from a sample of grouped observations about the ungrouped population. The second involves the relationship between moments calculated from grouped observations, before and after grouping, and the third is that of optimal grouping. Since the latter two were quite intensively discussed in Haitovsky (1983), most of the present entry is devoted to a discussion of the cases which are of major interest to the social and behavioral scientist. They are:

(a) Estimation of linear models (or regression models) from observations cross-classified by all the explanatory variables in the model in which only the cell means and frequencies are reported; or when the observations are differently grouped by subsets of the explanatory variables and only the corresponding marginal means and their frequencies are available to the researcher.

(b) The estimation of distribution curves when only mean values and frequencies are available. A case in point is the estimation of income or wealth distributions, when the individual data points are not available either because of confidentiality or its abundance. This has been historically the typical case of estimation of income or wealth distributions. A related curve is the so called Lorenz Curve or Relative Concentration Curve, used for the representation of the size distribution of income or wealth and inequality measures based on the latter, such as the Gini coefficient.

It is interesting to note that in spite of a history of over hundred years of analyzing the relationships between various statistics estimated from grouped observations and their ungrouped counterparts and of making statistical inference from ungrouped observations about the parent, ungrouped, distribution, a rigorous theory about the legitimacy of the use of grouped observations for statistical inference has not been developed until very recently. A unified theory about data coarsening, of which grouping is a special case, with the emphasis on coarsening at random (CAR) has just started to evolve in the last five years.

1. *Definition and Notation*

Data grouping is the process by which any variable W with a given distribution function $F(w)$, continuous or discrete, is condensed into a discrete distribution function, i.e.,

$$p_i = \int_{c_{i-1}}^{c_i} dF(w), \quad i = 1,\ldots, m \tag{1}$$

where $W \in [c_0, c_m]$ is partitioned by $c_0 < c_1 < \cdots < c_m$ into m disjoint and exhaustive groups, where the c_i's are set in advance. The c_i's are called interval limits or boundaries and (c_{i-1}, c_i) is the ith interval or class. Thus, p_i is the probability of an observation falling in the ith class. Hence, grouping is a transformation of one distribution function into a multinomial distribution function. The number of cases within the ith class is denoted by n_i, $\sum_i^m n_i = n$, and is called the ith group frequency. The condensations are usually into the interval mid-point $k_i = (c_{i-1} + c_i)/2$ or into the interval mean or centroid $\bar{w}_i$. The latter is the conditional mean of W, given $c_{i-1} \leqslant W < c_i$, i.e., $\bar{w}_i = \int_{c_{i-1}}^{c_i} w dF(w)/p_i$. If all the intervals are equally spaced, i.e., are of the same width $h = c_i - c_{i-1}$ $\forall i$, we will refer to them as equidistant.

It is useful to represent the grouping by an $m \times n$ grouping matrix $G = N^{-1}A$ where A is an $m \times n$ aggregation matrix with elements 0 or 1, the 1's in the ith row indicate the unit to be aggregated into the ith class and N is an $m \times m$ diagonal matrix with the class frequencies along its main diagonal. Hence, N^{-1} is an averaging matrix. Thus, the ith row of A sums to $n_i (\sum n_i = n)$ due to the exhaustiveness of the classification and each column sums to 1, due to the disjointedness of the classification.

Note that irrespective of how the observations are arranged.

$$GG' = \begin{bmatrix} \frac{1}{n_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{n_2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \frac{1}{n_m} \end{bmatrix} \tag{2}$$

The transformation of ungrouped to grouped data is obtained by $\bar{W} = GW$. Cross-classifications can be presented in a similar manner: the two (or more) different groupings can be performed by different grouping matrices, say, G_i $(i = 1.2)$ and the two matrices, which have necessarily the same number of columns, but not necessarily of the same number of rows, are stacked one upon the other to produce the piled matrix G_p to yield

$$G_p G_p' = \begin{bmatrix} N_1^{-1} & N_1^{-1} Q^T N_2^{-1} \\ N_2^{-1} Q N_1^{-1} & N_2^{-1} \end{bmatrix} \tag{3}$$

where N_i $i = 1, 2$ are diagonal matrices with group frequencies of the ith classification along their main diagonal and Q is the matrix with the joint frequencies as entries. For a simple numerical example the reader is referred to Haitovsky (1973, pp. 3–5). A k-way classification will be represented by stacking k matrices of type G.

2. *The Linear (Regression) Model*

Consider the linear (regression) model

$$Y = X\beta + \varepsilon \tag{4}$$

where Y is an $n \times 1$ vector of the original (ungrouped) observations on the dependent variable, X is an $n \times p$ matrix of n original observations on p explanatory variables, ε is an $n \times 1$ vector of disturbances, and β is the $p \times 1$ vector of unknown parameters to be estimated. If instead observations grouped into only m classes are available, the model can be transformed into:

$$\bar{Y} = \bar{X}\beta + \bar{\varepsilon}, \tag{4a}$$

where

$$\bar{Y} = GY$$
$$\bar{X} = GX \quad \text{and}$$
$$\bar{\varepsilon} = G\varepsilon,$$

where G is a grouping matrix of either type G or G_p from above.

Furthermore, if it is assumed that

$$E\{\varepsilon_i\} = 0 \quad \text{and} \quad E\{\varepsilon_j \varepsilon_j'\} = \sigma^2 I_n \tag{5}$$

by grouping the observations, we group the corresponding disturbances, resulting in:

$$\left.\begin{aligned} E\{\bar{\varepsilon}\} &= GE\{\varepsilon\} = 0 \quad \text{and} \\ E\{\bar{\varepsilon}\bar{\varepsilon}'\} &= GE\{\varepsilon\varepsilon'\}G' = \sigma^2 GG'; \end{aligned}\right\} \tag{5a}$$

that is, unless the grouping scheme is such that the same number of observations appear in each class of each classification, the transformation matrix destroys the homoscedastic (equal-variance) properties of the regression disturbances. In order to restore homoscedasticity, we use generalized least-squares estimation, resulting in

$$\bar{\beta} = \{\bar{X}'(GG')^{-1}\bar{X}\}^{-1}\bar{X}'(GG')^{-1}\bar{Y} \tag{6}$$

$$V(\bar{\beta}) = \sigma^2\{\bar{X}'(GG')^{-1}\bar{X}\}^{-1} \tag{7}$$

$$(\bar{Y} - \overline{X\beta})'(GG')^{-1}(\bar{Y} - \overline{X\beta})/(m-p)$$
$$\text{is an unbiased estimate of } \sigma^2, \tag{8}$$

and the coefficient of determination R^2, which serves as a measure of the goodness with which the model fits the data, is computed as

$$R^2 = \frac{\bar{\beta}'\bar{X}'(GG')^{-1}\bar{Y}}{\bar{Y}'(GG')^{-1}\bar{Y}} \tag{9}$$

Finally, under the assumption of the normality of the disturbances ε the analysis-of-variance table can be constructed as shown in Table 1.

For testing the null hypothesis $\beta = 0$ one uses the central F-statistics, where

$$F_{(p,\,m-p)} = \frac{\bar{\beta}'\bar{X}'(GG')^{-1}\bar{Y}}{\bar{Y}'(GG')^{-1}\bar{Y} - \bar{\beta}'\bar{X}'(GG')^{-1}\bar{Y}} \frac{m-p}{p}, \tag{10}$$

where the subscripts of F denote the numbers of degrees of freedom of the F-distribution.

The intuitive explanation of the proposed procedure is that, in the absence of direct information on all individuals in the sample, the best one can do is to assign the class centroids to all individuals falling within that class. The weighting has the effect of repeating the values recorded in the tables as many times as the number of individuals comprising the category, without of course attributing more degrees of freedom in the calculation of the variance. That is, each sum, sum of squares and sum of cross-products, is weighted by the corresponding frequencies of the

Table 1

Source of variation	Degrees of freedom	Sums of squares	E {sum of squares}
Total ungrouped	n	$Y'Y$	
Due to grouping	$n-m$	$Y'(I - G'(GG')^{-1}G)Y$	
Total after grouping	m	$\bar{Y}'(GG')^{-1}\bar{Y}$	
Due to grouped regression	p	$\bar{\beta}'\bar{X}'(GG')^{-1}\bar{Y}$	$\beta'\bar{X}'(GG')^{-1}\bar{Y} + p\sigma^2$
Residual after grouped regression	$m-p$	$\bar{Y}'(GG')^{-1}\bar{Y} - \bar{\beta}'\bar{X}'(GG')^{-1}\bar{Y}$	$(m-p)\sigma^2$

ungrouped data in each category. The typical elements in the cross-product matrices obtained by so doing will be $\sum_j n_j x_{ij} x_{sj}$ and $\sum_j n_j x_{ij} y_j (j = 1, ,\ldots, m_i; i, s = 1, \ldots, p)$.

2.1 How to Group

There still remains the question: how should the original observations be grouped to assure estimates as close as possible to those obtained from the original ungrouped observations? The question was addressed by Cramer (1964) for the case of one explanatory variable model and generalized by Haitovsky (1973, pp. 10–11) to the multivariate case, who showed that the observations should be cross-classified by all explanatory variables which are included in the model and each cell in the cross-classified table should include the sums (or centroids) of all the variables in the model—explanatory and dependent—falling within each cell and the number of observations making each cell, namely, the joint frequencies. We refer to such grouping as 'full cross-classification.' Fully crossed-classified data maximize the correlation between grouped and the corresponding ungrouped estimates, or alternatively it maximizes their respective relative efficiencies. In this sense it assures the closeness of the respective estimates. The reader is referred to Haitovsky (1973) for numerical examples that show how bad estimates not based on fully cross classified data can get, even though they are unbiased.

2.1.1 Estimation with differently grouped observations. The problem is that tables with data cross-classified by more than two variables are rarely available, while models with more than two explanatory variables are very prevalent in the social and behavioral sciences. What is often available are one-way-classification tables classified by each one of the model explanatory variables, where all the variables in the model and their frequencies are reported in each column of the table. In other words, the observations are differently grouped by all explanatory variables in the model, constituting the marginal means and frequencies of the fully cross classified tables. For this data set up Houthakker (reported by Haitovsky 1966) devised a Generalized-Least-Squares-type estimators of the form (6)–(10) except that here $\bar{X}$ and $\bar{Y}$ are the stacked $\bar{X}$ matrix and $\bar{Y}$ vector from the different one-way-classification tables, the number of degrees of freedom for the estimators of the model variance is the number of independent rows of the stacked $\bar{X}$ minus the number of estimated parameters required for the estimation of the model variance, and GG' is replaced by $G_p G_p'$. (Note that care must be taken before inverting $G_p G_p'$ since it is a singular matrix: either the appropriate rows and the corresponding columns must be deleted (Haitovsky 1973, p. 15) or generalized inverses must be applied (see Farebrother 1979 and the correction by Baksalary 1982). An apparent shortcoming of Houthakker's method is that there is no reason to believe that, if the full cross-classification of the necessary variables is not available, the corresponding joint frequencies will be available. However, there are quite a few good and relatively easy ways to construct the joint frequencies from the marginal frequencies.

The way to estimate the joint frequencies from the marginal frequencies varies according to the researcher's assumption about the nature of the cross classification. Farebrother (1979) considers the case where the grouping by the explanatory variables is independent and therefore the joint frequencies can be approximated by $n_{ij} = n_{i\cdot} n_{\cdot j} / \sum_i^{m_1} \sum_j^{m_2} n_{ij}$ where $n_{i\cdot}$ $(i = 1,\ldots, m_1)$ are the row sums and $n_{\cdot j}$ $(j = 1_j, \ldots, m_2)$ are the column sums of the joint frequency matrix N. Bewley (1989) showed however that, under the independence of grouping assumption, Farebrother's method can be very closely mimicked by stacking the marginal $\bar{X}$ matrices and $\bar{Y}$ vectors from the different tables and then applying to them the method of weighted least squares, where the weights are the corresponding marginal frequencies (see also Haitovsky 1973, footnote p. 59). Thus there is no need to construct the joint frequencies. Fukushige and Hatanaka (1991) suggested an improvement upon Farebrother's method by using Goodman's (1985) RC correlation method for estimating the joint frequencies from the marginals, which entails the calculation of correction factors based on the standardized scores of the rows and columns of the marginal frequencies and the correlation between them, applied to $n_{ij} = n_{i\cdot} n_{\cdot j} / \sum\sum n_{ij}$. Fukushige and Hatanaka (1991) also suggested a Bayesian estimation of the joint frequencies considering them as being jointly generated by a multinomial density, subject to $\sum_i \sum_j n_{ij} = n$, with a Dirichlet density as a natural conjugate prior. Fukushige and Hatanaka (1991) then went further to estimate the model parameters by the ordinary Bayes method as applied to the linear model, given the joint frequencies estimated either by the Bayesian or by Goodman's methods.

2.1.2 Haitovsky's methods. Estimations not based on Houthakker's method are Haitovsky's (1966, 1973, Chap. 4) unbiased estimators and extentions (1973, pp. 35–36), shown by Farebrother (1979) to be instrumental variables estimators (he also corrected an error in Haitovsky's calculation of the variance of the estimator), and Haitovsky's 'Efficient Estimator' (1968, 1973, Chap. 5), which also requires iterations.

Haitovsky's (1973) unbiased estimators utilize the property that only estimates from observations group-

Table 2
Effect of different grouping and estimation methods on the estimates of a model with two explanatory variables using Houthakker's data

Model	Intercept	Coefficient of X_1	Coefficient of X_2	$\hat{\sigma}^2$
Ungrouped data ($n = 1218$)	17.10 (7.30)[a]	0.7578 (0.1398)	−0.1778 (0.0367)	5250
Fully cross classified data ($n = 56$)	16.47 (10.32)	0.7473 (0.1203)	−0.1624 (0.0323)	3914
Classified by X_1 ($n = 7$)	10.86 (5.78)	0.5505 (1.6139)	0.0382 (1.8752)	9027
Classified by X_2 ($n = 8$)	73.74 (11.26)	−0.6532 (2.5391)	−0.0931 (0.1572)	1345
Haitovsky's 'unbiased method' ($n = 15$)	18.0335 (6.6065)[b]	0.7271 (0.1033)	−0.1718 (0.0282)	4337
Haitovsky's 'unbiased method' explanatory variables are differently grouped[c] ($n = 15$)	17.8659	0.7319	−0.1723	
Haitovsky's '2 stage efficient method' ($n = 15$)	16.7025 (8.9025)	0.7560 (0.2006)	−0.1725 (0.0187)	
Weighted regression applied to stacked data[d] ($n = 14$)	18.5065 (5.2765)	0.7133 (0.1320)	−0.1698 (0.0355)	4331
Estimation based on Houtakker's method				
Joint frequencies are available ($n = 14$)	18.08 (5.87)	0.7263 (0.1259)	−0.1719 (0.0338)	4285
Joint frequencies are available, using generalized inverse ($n = 15$)	18.0745 (5.8691)	0.7263 (0.1259)	−0.1719 (0.0338)	4286
Joint frequencies are estimated by RC correlation method ($n = 15$)	18.491	0.7147	−0.1704	4322
Bayes estimates, joint frequencies are available[e]	18.542	0.7174	−0.1728	5419
Bayes estimates joint frequencies are estimated by RC correlation method[e]	18.581	0.7120	−0.1700	5238

Source: Haitovsky (1973), lines 1–5, 7–9, Farebrother (1979), line 10, Fukushige & Hatanaka (1991), lines 11–13.
[a] Figures in parentheses are the estimated standard errors of the coefficients above them. [b] The standard error is a correction suggested by Farebrother (1979). [c] Here the tables of Y and X_1 classified by X_1 and Y and X_2 classified by X_2 have 7 and 8 classes as in the previous entry in Table 1, but the X's are cross classified in four classes in each X_1 and X_2. [d] The estimates are identical to those obtained by Houthakker's method where the joint frequencies are constructed from the marginal frequencies under the independence of the cross classification assumption. See Bewley (1989). [e] The estimates reported here are based on 10000 replicates and on the highest level of precision imposed on the hyper parameter in the Bayesian estimates. The differences between the extreme values for the different level of precision are 4%, 1%, and 4% for the estimates for the intercept, coefficient of X_1, and coefficient of X_2.

ed by all explanatory variables featuring in the model, either cross-classified or differently classified, assures the credibility of the estimates in the following way. Suppose that the researcher estimates only $\hat{\beta}_j$ ($j = 1, \ldots, p$) from the table grouped by the jth variable, using the least squares formula for the estimation of a simple, one explanatory variable, model, namely, $\sum_i^{(j)} y_i x_{ji} / \sum_i^{(j)} x_{ji}^2$, where x_{ji} and y_i are measured as deviation from their respective weighted means, the superscripts designate the table used for the estimation, and the weighting by the marginal frequencies is subsumed under the summation sign. But since the appropriate model is given by eqn. (4a), upon substituting it in the last formula and replacing the β_j by $\hat{\beta}_j$ one obtains

$$\frac{\sum_i^{(j)} y_i x_{ji}}{\sum_i^{(j)} x_{ij}^2} = \hat{\beta}_j + \sum_{l \neq j}^{p} \hat{\beta}_l \frac{\sum_i^{(j)} x_{ji} x_{li}}{\sum_i^{(j)} x_{ji}^2} \quad (j = 1, \ldots, p) \qquad (11)$$

which is a system of p equations linear in the p unknown $\hat{\beta}_j$'s. Multiplying both sides by $\sum_i^{(j)} x_{ji}^2$ (assuming they are all strictly positive), one obtains a normal-equation-like system, in which the jth equation is calculated from the table classified by the jth variable. If there are as many one-way-classification tables as there are explanatory variables, where each one of the tables is classified by each of the explanatory variables, and in each of these tables all the explanatory variables appear as data (centroids), the $\hat{\beta}_j$'s can be derived yielding unbiased, or conditionally unbiased, estimates depending on whether the x variables are random or fixed. This method is the 'unbiased method.'

However, if the dependent variable and only one explanatory variable are classified by the latter in a one-way-classification table and in addition each one of the explanatory variables is classified by each of the others, and both the classified and classifying variables appear in all the tables, eqn (11) can be solved to yield unbiased (or conditionally unbiased) estimates. The advantage of the latter option is that the different tables need not be grouped by the same number of classes, nor with the same interval limits.

As for Haitovsky's 'efficient method' the reader is referred to Haitovsky (1968).

2.1.3 A numerical example. To complete this section, it will be instructive to present some numerical results. The data set originally used by Houthakker (see Haitovsky 1973, Appendix 1) has been extensively used to examine the performance of various methods by comparing to the estimates obtained from the ungrouped or fully cross classified data. As can be seen from Table 2, all the regression coefficients based on cross-classified data or on properly weighted data, classified in a one way classification table, are remarkably close to the ungrouped estimates, providing that there are tables classified by all explanatory variables. It is true even in cases where the information about the ungrouped observations is substantially reduced by grouping.

3. Estimation of Distribution Functions

Another important application of grouped observations is in the estimation of distribution functions. A case in point is the estimation of income or wealth distributions. Typically, confidentiality disallows the reporting of the original observations by, say, the US Internal Revenue Service. Instead only the grouped observations including the class boundaries, the class frequencies and often the centroids are available. Another case is data measured on a discrete, rather than continuous scale, and grouped into classes. The latter often happens in the estimation of survival curves where a subject is inspected for failure only at prespecified time intervals. Note that in this case the midpoints rather than the centroids are available, except for open-ended intervals; the latter are to be treated differently.

3.1 Estimation of Income or Wealth Distributions

Income and wealth distributions are often assumed to follow a parametric distribution, such as Pareto or Log-Normal distributions, and the grouped data are applied in the estimation of the parameters that characterize these distributions. Alternatively, a nonparametric approach is taken, namely, the underlying data generation process is ignored and a distribution free curve is fitted to the available income points. For the latter, we quote Mazzarino (1986) for an interesting experiment investigating the effect of the number of groups on the estimated curve. A natural starting point for the two approaches is to consider the group frequencies as a random sample from a multinomial distribution and write the likelihood function $L(\theta)$ of the vector of parameters θ as proportional to $L \propto \prod_{i=1}^{m} p_i^{n_i}$, p_i being the probability of an observation falling in the ith group, as defined in Sect. 2 (see Kulldorff 1961).

In addition to the maximum likelihood estimate of Pareto's scale parameter, Aigner and Goldberger (1970) apply a regression framework to the estimation of the scale parameter, making use of the mean, variance and covariance of the multinomial distribution and applying the least squares principle to the differences between the observed relative group frequencies n_i/n and their corresponding probabilities p_i. Alternatively, least squares can be applied to the differences between the observed cumulative 'greater than' relative frequencies and the corresponding probabilities or to the log cumulative frequencies relative to the corresponding log probabilities. Aigner and Goldberger (1970) found that, although the latter is easiest to compute, its relative asymptotic efficiency deteriorates fast when compared to the maximum likelihood and generalized least squares estimates, the two most (asymptotically) efficient estimators considered by them, as the scale parameter grows beyond 1.5.

3.2 Estimation of Lorenz Curve

The Lorenz curve (also called relative concentration curve) is widely used for the representation and analysis of the size distribution of income (or wealth). It relates the cumulative income (wealth) fraction earned by various proportions of the population to the proportion receiving this income, when the units are arranged in increasing order of their income (wealth). Thus, if income is given by the probability density function $f(u)$, the proportion of units having income less than w is

$$F(w) = \int_0^w f(u)\,du$$

while their mean income is

$$\bar{w} = \int_0^w u\,f(u)\,du,$$

the latter is often called the incomplete first moment. Given that μ, the average of all the units receiving income, exists, then $L(w) = \bar{w}/\mu$ is the cumulative relative income earned by those whose income is less than w. The Lorenz curve is obtained by relating $L(w)$ to $F(w)$ and plotting its values in a unit square with $L(w)$ as the ordinate and $F(w)$ as the abscissa, or as was defined by Lorenz:

> plot along one axis the cumulated percent of the population from the poorest to the richest, and along the other axis the percent of the total wealth held by these percents of the populations. (1905)

For grouped data we define here

$$p_i = \sum_{j=1}^{i} n_j \Big/ \sum_{j=1}^{m} n_j,$$

the relative population frequencies of the first i groups, and $\bar{w}_j$ the average income within group i, $L(w)$ is estimated by

$$\hat{L}(p_i) = \sum_{j=1}^{i} n_j \bar{w}_j \Big/ \sum_{j=1}^{m} n_j \bar{w}_j$$

and the fraction of the population earning less than c_i is given by $\hat{F}^{-1}(p_i)$. For estimating the Lorenz curve, Kakwani and Podder (1973, 1976) assume income to follow Pareto's law, from which the Lorenz curve is easily derived, and then follow Aigner and Goldberger's (1970) framework to estimate the Lorenz curve. The formula derived from Pareto's law suggests two extensions (Kakwani and Podder 1973) which seems to fit the data better. Another extension was proposed by Ortega et al. (1991), who also list a few more functional forms previously suggested for the estimation of Lorenz curves. Another interesting formulation that takes account of some mathematical restrictions on the Lorenz curve and hence improves its estimation was proposed by Kakwani and Podder (1976). A non-parametric approach is suggested by Gastwirth and Glauberman (1976), who use Hermite interpolation for smoothing the curve. An alternative approach to distribution-free methods was suggested by Brown and Mazzanino (1984). Still another approach was taken by Gastwirth (1975), Gastwirth and Krieger (1975), and Krieger (1979). They put lower and upper bounds on the Lorenz curve and on its moments.

A bivariate version of the Lorenz curve has been applied to measure the progressivity of taxation (Suits 1977) and to the design of welfare dominating tax policies (Yitzhaki and Slemrod 1991). Another application is in the field of finance where the second degree of stochastic dominance criterion (Levy 1992) is applied to ordering risky prospects in accordance with maximization of expected utility, when the shape of the utility function is not known (Yitzhaki 1982) and for the evaluation of rising prospects (Shalit and Yitzhaki 1994).

3.3 Estimation of Inequality Measures Based on Lorenz Curves

Perfect equality is represented by the 45° line in the Lorenz curve unit box, connecting the origin with the opposite corner, hence called the egalitarian line. The larger the area between that line and the Lorenz curve the less egalitarian is the society represented by the latter. There are numerous ways to calculate the Gini coefficient, but most of them are not suitable for its calculation from grouped data because they ignore the intra-class inequality and thus bias the measure downward (see Lerman and Yitzhaki 1989). A simple way to estimate the Gini coefficient non-parametrically is by

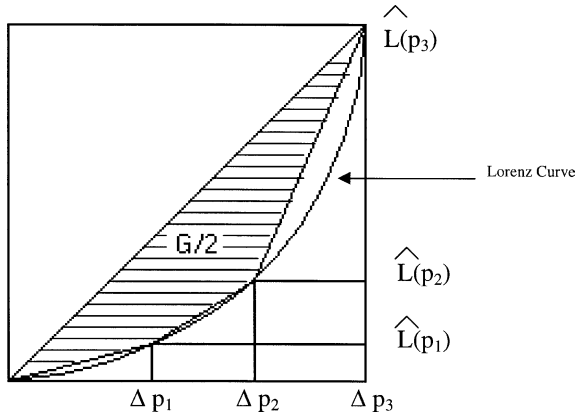

Figure 1
Lorenz curve and the estimation of Gini coefficient

$$G = 1 - 2\sum_{i=1}^{m} \Delta p_i Q_i \tag{12}$$

where

$$Q_1 = \hat{L}(p_1)/2,\ Q_i = \hat{L}(p_{i-1}) + n_i \bar{w}_i / 2 \sum_{j=1}^{m} n_j \bar{w}_j,$$

$$i = 2\ldots, m,$$

$$\Delta p_1 = p_1, \quad \Delta p_i = p_i - p_{i-1} \qquad i = 2,\ldots m,$$

and $\sum_{i=1}^{m} \Delta p_i Q_i$ is the area of the rectangles and triangles in Fig. 1.

Note that $\sum_{i=1}^{m} \Delta p_i Q_i$ overestimates the area to the right of the Lorenz curve in the typical strictly convex Lorenz curve and hence G underestimates the Gini coefficient, but this bias decreases rapidly as the number of groups increases. The Gini coefficient, as estimated by G, lends itself to various types of decompositions, such as decomposition by income source or by population class, e.g., urban/rural division, where both can be further decomposed by, say, regions, resulting in a two-tier decomposition. Another possible decomposition is by intra-class, inter-class, and overlapped components. The latter arises from the fact that poor people in a high-income class may be worse off than rich people in a low income-class (Yao 1999). Bounds on the Gini coefficient were proposed by Gastwirth (1975) and by Gastwirth and Krieger (1975). Parametric estimation can be derived from the parametric estimation of

Lorenz curves by the method of moments or maximum likelihood (Slottje 1990). Other parametric estimation were proposed by Kakwani and Podder (1976) and by Kakwani (1976) to overcome the downward bias inherent in most derivations of the Gini coefficients from grouped observations.

4. Coarsening at Random

A question that only recently has been investigated concerns the effect of coarsening mechanism (grouping in the context of our article) on statistical inference using grouped observations about the ungrouped distribution. Alternatively, the question is: when can the grouping mechanism be ignored if the above inference is desired? Gill et al. (1997) and Gill and Robins (1997) investigated this question in terms of the conditional distribution of the grouped observations given the corresponding ungrouped observations. This question is of particular interest in cases of statistical inference when the above-mentioned conditional distribution is being used, e.g., in the calculation of score functions, in using the EM algorithm (the E-step), or in going through the Gibbs' sampler. The intuition in the grouped observations case is straightforward: if we know, or can safely assume, that the grouping mechanism is independent of the ungrouped observations, the grouped observations can be legitimately used for statistical inference without the explicit introduction of the grouping mechanism into the analysis.

To put it more formally, does the marginal distribution of the grouped observations, indexed by the parameters that characterize the ungrouped distribution and the particular mechanism applied in the grouping process respectively (i.e., the joint likelihood for these two parameters) factor into the conditional distribution of the grouped observations, given the ungrouped observations, and the distribution of the ungrouped observations, indexed by its parameter? If such factorization is possible, the grouping mechanism can be ignored in making the inference from the grouped to the ungrouped distribution. In the grouped observations case, if the grouping is done by intervals set in advance and independent of the ungrouped observations and the statistics used in the inference are calculated from the observations grouped this way, the above factorization is possible and the statistics calculated from the grouped observations can be legitimately used for the desired inference. The good news is that indeed most grouping mechanisms obey this rule.

5. Conclusion

Although it is evident from the discussion above that methods are available for drawing inference from grouped observations with little loss of precision, the cost of storing, handling and analyzing large data sets has gone down with the fast growing speed and capacity of computers and the parallel development of the necessary software for its utilization. It seems that inference from grouped observations will be mainly needed for cases in which access to the individual data is too costly or impossible, e.g., because of data confidentiality, and therefore the use of grouped observations will be cost efficient.

Bibliography

Aigner D J, Goldberger A C 1970 Estimation of Pareto's law from grouped observations. *Journal of the American Statistical Association* **65**: 712–23

Baksalary J K 1982 An invariance property of Farebrother's procedure for estimation with aggregated data. *Journal of Econometrics* **22**: 317–22

Bewley R 1989 Weighted least squares estimation from grouped observations. *Review of Economics and Statistics*. **71**: 187–8

Brown J A C, Mazzarino G 1984 Drawing the Lorenz curve and calculating the Gini concentration index from grouped data by computer. *Oxford Bulletin of Economics and Statistics* **46**: 273–78

Cramer J S 1964 Efficient grouping, regression and correlation in Engel curve analysis. *Journal of the American Statistical Association* **59**: 233–50

Farebrother R W 1979 Estimation with aggregated data. *Journal of Econometrics* **10**: 43–55

Fukushige M, Hatanaka M 1991 Estimation of a regression model on two or more sets of differently grouped data. *Journal of Econometrics* **47**: 207–26

Gastwirth J L 1975 The estimation of a family of measures of economic inequality. *Journal of Econometrics* **3**: 61–70

Gastwirth J L, Glauberman M 1976 The interpolation of the Lorenz curve and Gini index from grouped data. *Econometrica* **40**: 479–83

Gastwirth J L, Krieger A M 1975 On bounding moments from grouped data. *Journal of the American Statistical Association* **70**: 468–71

Gill R D, van der Laan M J, Robins J M 1997 Coarsening at random: Characterizations, conjectures, counter-examples. In: Lin D Y, Fleming T R, (ed.) *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*. Springer-Verlag, pp. 255–94

Gill R D, Robins J M 1997 Sequential models for coarsening and missingness. In: Lin D Y, Fleming T R (ed.) *Proceedings of the First Seattle Symposium on Biostatistics: Survival Analysis*. Springer-Verlag, pp. 295–305

Goodman L A 1985 The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and symmetry models for contingency tables with or without missing entries. *Annals of Statistics* **13**: 10–69

Haitovsky Y 1966 Unbiased multiple regression coefficients estimated from one-way-classification tables when the cross classifications are unknown. *Journal of the American Statistical Association* **61**: 720–8

Haitovsky Y 1968 Regression analysis of grouped observations when the cross classifications are unknown. *Review of Economic Studies* **35**: 77–89

Haitovsky Y 1973 *Regression Estimation from Grouped Observations*. Griffin, London

Haitovsky Y 1983 Grouped data. In: Johnson N L, Kotz S, Read C B (eds.) *Encyclopedia of Statistical Sciences*. Wiley, New York. Vol. 3, pp. 527–537
Heitjan D F, Rubin D B 1991 Ignorability and coarse data. *Annals of Statistics* **19**: 2244–53
Kakwani N 1976 On the estimation of income inequality measures from grouped observations. *The Review of Economic Studies* **43**: 483–93
Kakwani N C, Podder N 1973 On the estimation of Lorenz curves from grouped observations. *International Economic Review* **14**: 278–91
Kakwani N C, Podder N 1976 Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations. *Econometrica* **44**: 137–48
Krieger A M 1979 Bounding moments, the Gini index and Lorenz curve from grouped data for unimodal density functions. *Journal of the American Statistical Association* **74**: 375–8
Kulldorff G 1961 *Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples*. Almquist and Wiksell, Stockholm, Sweden
Lerman R, Yitzhaki S 1989 Improving the accuracy of estimates of Gini coefficient. *Journal of Econometrics* **42**: 43–7
Levy H 1992 Stochastic dominance and expected utility: Survey and analysis. *Management Science* **83**: 555–93
Lorenz M O 1905 Method for measuring of wealth. *Journal of the American Statistical Association* **9**, New series No. 70, 209–19
Mazzarino G 1986 Fitting of distribution curves to grouped data. *Oxford Bulletin of Economics and Statistics* **48**: 189–200
Ortega P, Martin G, Fernandez A, Ladoux M, Garcia A 1991 A new functional form for estimating Lorenz curves. *Review of Income and Wealth* **37**: 447–52
Shalit H, Yitzhaki S 1994 Marginal conditional stochastic dominance. *Management Science* **40**: 670–84
Slottje D J 1990 Using grouped data for constructing inequality indices. *Economics Letters* **32**: 193–7
Suits D B 1977 Measurement of tax progressivity. *American Economic Review* **67**: 747–52
Yao S 1999 On the decomposition of Gini coefficients by population class and income source: A spreadsheet approach and application. *Applied Economics* **31**: 1249–64
Yitzhaki S 1982 Stochastic dominance, mean variance, and Gini's mean difference. *American Economic Review* **72**: 178–85
Yitzhaki S, Slemrod J 1991 Welfare dominance: An application to commodity taxation. *American Economic Review* **81**: 480–96

Y. Haitovsky

Statistical Analysis, Special Problems of: Outliers

Anyone analyzing data may expect sometimes to encounter one or more observed values which appear to be inconsistent or out of line with the rest of the data. Such a value is called an outlying observation or outlier. Judged subjectively, it looks discrepant or surprising. There are two main ways of dealing with outliers, discordancy testing (Sects. 2, 3) and accommodation (Sect. 4). In simpler types of data set, the presence of an outlier is evident by inspection. In more complex data situations such as designed experiments, an outlying observation may disrupt the expected pattern of results, though its presence is not obvious by inspection and a statistical procedure is needed to detect it (Sect. 5).

1. Two Simple Examples

1.1 A Disputed Paternity Case

In 1921 Mrs Kathleen Gaskill was sued for divorce in the British courts by her husband, a serving soldier, because she had given birth to a child 331 days after he had left the country to serve abroad. He claimed that the child could not be his, since 331 days was not credible as a duration of pregnancy. She denied adultery and claimed, with medical support, that the length of pregnancy, though unusual, was genuine. Viewed as a duration of pregnancy, 331 days is an outlier. In the legal case (which the court decided in favor of the wife), the only evidence of the alleged adultery was the abnormal length of this duration. This exemplifies the situation where the judgment of an outlier depends purely on its position in relation to a main data set. Table 1 gives the durations of pregnancy for 1669 women who gave birth in a maternity hospital in the north of England in 1954 (summarized from Newell 1964). Clearly 331 days, exceeded certainly by 1 case and possibly by 2 out of 1669 in this data set, cannot be regarded as an incredible duration.

1.2 Predicting College Grades

Figure 1 shows a scatter plot of points (x, y) for a small sample of 22 college students; x is the student's score in an aptitude test, used as a predictor of college performance, and y is the student's freshman year grade-point average. There are two outlying points $A(x_A, y_A)$ and $B(x_B, y_B)$; the desired predictive relationship of y from x could be modeled satisfactorily by a linear regression omitting these two points. Should they be included in the relationship or not? This is discussed in Sect. 5. See *Linear Hypothesis: Regression (Basics)*.

2. Outliers in Univariate Samples

The nature of outliers and the range of issues in dealing with them are first discussed in the simple situation of a univariate sample of n observations $x_1, x_2, \ldots, x_n$ of a variable X. Denote the observations arranged in ascending order by $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$,

Table 1
Durations of pregnancy for 1669 women

Duration (weeks)	11–20	21–25	26–30	31–35	36–38	39–41	
Numbers	4	10	23	113	333	962	
Duration (weeks)	42	43	44	45	46	47	56
Numbers	144	53	12	10	3	1	1

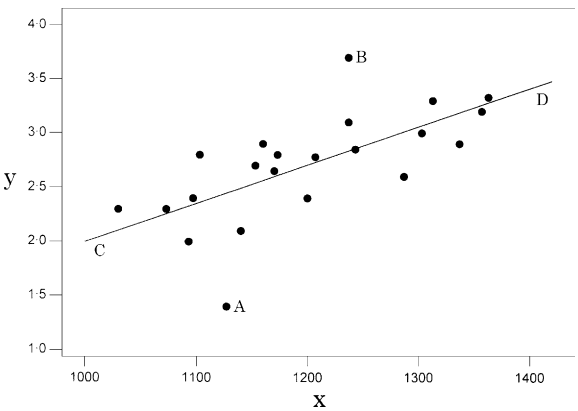

Figure 1
Scatter plot of grade-point average y vs. predictive test score x for 22 college students

$x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. If, say, the upper extreme observation $x_{(n)}$ appears (to an analyst or investigator examining the data) to be, not only higher than the rest of the sample but surprisingly high, it is declared an outlier. This is a subjective judgment. Is the outlier credible as a value of X? This depends on the form the investigator is assuming for the distribution of X, e.g., whether it is, say, normal or on the other hand a skew distribution such as exponential in which a high value $x_{(n)}$ is less surprising. Call this assumed model for the distribution of X the basic model and denote it by F; $x_{(n)}$ is an outlier relative to model F (e.g., normal), but may not be an outlier relative to a different model (e.g., exponential).

Table 2, with dotplot Fig. 2, gives the scores by 30 children in a reading accuracy test. There is an upper outlier $x_{(30)} = 62$ and possibly a second one $x_{(29)} = 53$. The variable X is here the reading accuracy test score obtained by a child in a population of children of the relevant age group and educational characteristics. A reasonable working assumption is that its distribution is normal; this is supported by a plot of ordered values against normal order scores, which is roughly linear.

2.1 Causes of Outliers

No statistical issue arises if there is a tangible explanation for the presence of an outlier, e.g., if in the above example the score was misrecorded, or component subscores were wrongly totaled, or if the child tested was known to be unrepresentative of the population under study. The outlying value would simply be removed from the data set, or corrected where appropriate, e.g., in the case of wrong totaling.

Otherwise, a statistical judgment is required for interpreting the outlier and deciding how to deal with it, based on the apparent discrepancy between it and the pattern the investigator has in mind for the data—the basic model F. This judgment requires a test of the null hypothesis or working hypothesis H that all n observations belong to F. On this hypothesis, the upper outlier $x_{(n)}$ is a value from the distribution of a variable $X_{(n)}$, the highest observation in a random sample of size n from F. If H is rejected by a statistical test at some level (e.g., 5 percent), $x_{(n)}$ is judged to be not credible (at the level of the test) as the highest value in a sample of size n from F. It is then said to be a discordant outlier (relative to the basic model F); the test of H is a discordancy test. (See *Hypothesis Testing in Statistics*; *Significance, Tests of*. See also Sect. 2.3.)

In a Bayesian approach, there is no direct analogue of a test of discordancy. Its place is taken by procedures designed to obtain a posterior probability assessment of the 'degree of contamination' of the data. This means the degree to which the data contain observations ('contaminants') generated not by F but by some other model G; a basic model is used which contains additional parameters relating to a possible G (see *Bayesian Statistics*).

2.2 Treatment of Outliers: Retention, Identification, Rejection, Incorporation

If the result of the test is to accept H and find $x_{(n)}$ nondiscordant, all the data can be taken to come from F, including the outlier. Although it has been thought to

Table 2
Reading accuracy test scores (in ascending order) for 30 children

2	4	5	8	9	12	12	12	12	14	16	16	18	19	20	21	23	23
23	26	27	27	28	30	34	38	40	42	53	62						

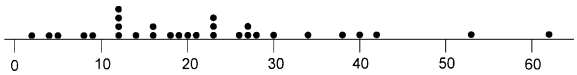

Figure 2
Dotplot of the 30 test scores in Table 2

look somewhat unusual, it is consistent with having arisen purely by chance under the basic model and can be retained in the data set.

When on the other hand the outlier is adjudged discordant, there are three possibilities.

If the focus of interest is the outlier itself (as in the disputed paternity case), it is identified as a positive item of information, inviting interpretation or further examination; it might reveal, for instance, a novel teaching method causing an exceptional test score.

The other two possibilities, rejection and incorporation, arise when the investigator's concern is with the main data set, and the aim of the research is to study characteristics of the underlying population, e.g., to estimate the mean of a distribution of examination scores. The outlier is only of concern insofar as it affects these inferences about the underlying population. It is not of interest in itself; for example, the students giving rise to points A and B in Fig. 1 are effectively unknown!

Either the model F is firmly adopted without challenge, and estimates or tests of the population parameters are to be based on F; or there is the option of changing the model.

If the model is not under question and inferences are to be based on it, the outlier, being adjudged inconsistent with F, is a nuisance item and should be rejected. Historically, outlying values have been studied since the mid eighteenth century, often in the context of astronomical observations. Until the 1870s rejection was often the only procedure envisaged for dealing with outliers, as typified by such publications as Peirce (1852) 'Criterion for the rejection of doubtful observations' and Stone (1868) 'On the rejection of discordant observations.'

The fact that an outlier is omitted for purposes of analyzing a data set does not necessarily mean that it is erroneous or undesirable. Consider, for example, Fig. 3. The ordinates are annual figures for the population of the UK, in millions, published in each year's *Annual Abstract of Statistics*. Those for 1985 through 1990 are estimates based on the 1981 census, while the outlier at 1991, the true value from the 1991 census, is the only genuine data point!

The alternative possibility, when the outlier is adjudged discordant relative to F, is that all the observations, including the outlier, belong to the same distribution, but this distribution is not satisfactorily represented by the initial basic model F. The analysis then needs to start afresh with a more appropriate choice of model F* (e.g., a skew distribution instead of a symmetrical F), in relation to which $x_{(n)}$ is not

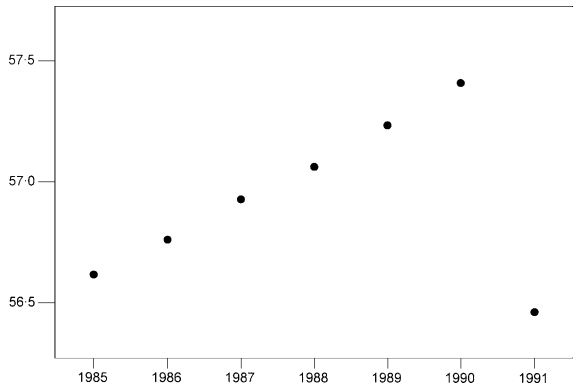

Figure 3
Some annual figures for the UK population

discordant. It was an outlier relative to F, but is not an outlier relative to F*, which satisfactorily models the entire data set. It has been incorporated in the data set in a non-discordant fashion, and the outlier problem disappears.

2.3 Other Meanings of the Word 'Outlier'

In the entry, an observation which appears inconsistent with the main data set is called an outlier; this is a subjective judgment. If a statistical test provides evidence that it cannot reasonably be regarded as consistent with the main data set, it is a discordant outlier. It is important to bear in mind, however, that some writers use the word 'outlier' for a discordant outlier, and refer to an observation which looks anomalous but may or may not be discordant as a 'suspicious' or 'suspect' value.

Note also the different meaning given to the word 'outlier' in graphical displays in exploratory data analysis (EDA), such as boxplots. With Q_1 and Q_3 denoting (essentially) the lower and upper quartiles in the sample, observations greater than $Q_3+k(Q_3-Q_1)$ or less than $Q_1-k(Q_3-Q_1)$ are flagged as outliers. These values are sometimes outliers and sometimes not. With the typical value of 1.5 for k, a normal sample of size 100 has more than 50 percent chance of containing one or more of these 'outliers'! (See *Exploratory Data Analysis: Univariate Methods*.)

3. Multiple Outliers: Block and Consecutive Tests

Any particular discordancy test (taking univariate samples as context) is specified by three items: the particular outlier configuration (e.g., a lower outlier; an upper and lower outlier pair), the distribution of the data on the working hypothesis (e.g., normal with unknown mean and variance), and the test statistic. Taking Table 2 for illustration, there is a discordancy

test for $x_{(30)} = 62$, in a normal sample, based on its difference from the sample mean $\bar{x} = 22.53$ in relation to the sample standard deviation $s = 14.13$, with test statistic $T = (x_{(n)} - \bar{x})/s = (62 - 22.53)/14.13 = 2.79$.

For $n = 30$ the tabulated 5 percent point of T on hypothesis H is 2.74; 62 is declared discordant (just) at the 5 percent level. The observation $x_{(29)} = 53$, which is also high, can now be tested for discordancy in relation to the other 28 values; this is a consecutive test procedure for 2 outliers. The value of T is $(53 - 21.17)/12.22 = 2.61$, $P \simeq 8$ percent, the evidence of discordancy is weak.

Alternatively, 62 and 53 could be considered together as an outlying pair in relation to the other 28 observations. This can be tested, in a block test procedure for 2 outliers, by the statistic $(x_{(n-1)} + x_{(n)} - 2\bar{x})/s$, which here $= (53 + 62 - 2 \times 22.53)/14.13 = 4.95$, to be compared with the tabulated 1 percent point 4.92. So, on this block test, there is quite strong evidence that (62, 53) are a discordant pair.

3.1 Masking and Swamping

If $x_{(29)}$ were nearer to $x_{(28)} = 42$, the T-value for $x_{(30)}$ would be higher than 2.79, and the evidence for the discordancy of $x_{(30)}$ would be stronger. This is called the masking effect. A discordancy test of the most extreme observation is rendered insensitive by the nearness of the next most extreme observation; the second observation has masked the first.

Conversely, a block test could be insensitive if the second observation $x_{(n-1)}$ is close to $x_{(n-2)}$ and not outlying. The pair $x_{(n)}$, $x_{(n-1)}$ might not reach discordancy level when tested jointly, even though $x_{(n)}$ on its own is discordant in relation to the other $n-1$ observations. This is called the swamping effect of $x_{(n-1)}$. If $x_{(n-1)}$ is unduly high, there is the risk of masking; if it is unduly low, there is the risk of swamping.

4. Accommodation of Outliers

When the aim is to make inferences about a population of values of a variable X by analyzing a sample of observations of X, there is an alternative to detecting outliers and testing them for discordancy. This is to use procedures (testing or estimation) which are relatively unaffected by possible outliers, by giving extreme values less weight than the rest of the observations, whether they are discordant or not. Such a procedure is called robust against the presence of outliers (see *Robustness in Statistics*).

There is a great range of procedures available in the literature. The approach is well illustrated by the following simple methods for robust estimation of the mean and variance of a univariate sample $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$.

4.1 Trimming

The mean is calculated after one or more extreme values have been omitted from the sample. Omission of the r (lowest) and the s (highest) values gives the (r, s)-fold trimmed mean

$$(x_{(r+1)} + x_{(r+2)} + \cdots + x_{(n-s)})/(n - r - s).$$

4.2 Winsorizing

Instead of omitting the r (lowest) values $x_{(1)}, \ldots, x_{(r)}$, they are each replaced by the nearest value which is retained unchanged, i.e., $x_{(r+1)}$; similarly for the s (highest) values. The mean of the resulting modified sample, still of size n, is the (r, s)-fold Winsorized mean

$$((r+1)x_{(r+1)} + x_{(r+2)} + \cdots + x_{(n-s-1)} + (s+1)x_{(n-s)})/n.$$

The choice of r and s depends on various factors such as the form of the distribution of X (e.g., symmetrical or asymmetrical) and the level of robustness required. The greatest possible amount of trimming or Winsorizing, with $r = s$, gives the sample median.

Robust estimates of variance can be based in various ways on the deviations $x - m$, where m is a trimmed or Winsorized mean, instead of the usual $x - \bar{x}$.

For an extensive account of accommodation issues and procedures, see Barnett and Lewis (1994).

5. Outliers in More Complex Data Situations

In a regression of a variable y on a single regressor variable x, an outlying point is apparent by inspection, as instanced by Fig. 1. For a regression on 2 or more regressor variables $x_1, x_2, \ldots$, this is no longer the case. The data may nonetheless contain an outlying point (or more than one), i.e., a point (x, y) which departs to a surprising degree from the overall pattern of the data. Its observed response value y deviates 'surprisingly' from the fitted value from the regression equation; in other words, it has an outlying residual. Detection of an outlying point requires essentially a statistical examination of residuals.

To illustrate this, consider the regression data in Fig. 1, with two outlying points A and B. Denote the fitted regression line CD by $y = a + bx$ ($a = -1.528$, $b = 0.003528$), with residual mean square $s^2 = 0.148$. The residual for the outlying point A is $y_A - (a + bx_A) = 1.4 - 2.447 = -1.047$, and its estimated standard deviation is

$$s\sqrt{\left\{1 - \frac{1}{n} - \frac{(x_A - \bar{x})^2}{\sum(x - \bar{x})^2}\right\}} = 0.371.$$

A discordancy test statistic for the point A is its studentized residual (the residual divided by its esti-

mated standard deviation), viz. $-1.047/0.371 = -2.82$; this is just significant at the 5 percent level. On a similar test the outlying point B is not discordant at all. The most outlying point in the regression (in this case, A) is the one with the largest studentized residual, ignoring sign. See *Linear Hypothesis: Regression (Basics)*.

An equivalent criterion to the studentized residual for testing discordancy is the proportional reduction in the residual sum of squares when the outlier is omitted, or, equivalently again, the increase in the maximized log-likelihood. The test then reflects directly how much the fit of the regression model to the data is improved on omission of the outlier. This approach—the deletion method—obviously extends to the detection and testing of multiple outliers, with the choice between block and consecutive deletion (see *Likelihood in Statistics*). It is also the appropriate approach to outliers in multivariate data; see *Multivariate Analysis: Overview*.

Similar considerations apply to the detection and testing of outliers in designed experiments and outliers in contingency tables, with use of residual-based or deletion procedures. See *Experimental Design: Overview*. Accommodation procedures are also available in the literature. Outliers also occur in time series (see *Time Series: General*). For details of methods for dealing with outliers in the various data situations mentioned above, see Barnett and Lewis 1994.

Some work has been done on outliers in sample surveys, but the methodology for outliers in this vital field urgently needs developing. See *Sample Surveys: Methods*; *Sample Surveys: Model-based Approaches*; *Sample Surveys: The Field*.

In multilevel, or hierarchical, data the concept of an outlier becomes more complex, and much work needs to be done. See *Hierarchical Models: Random and Fixed Effects*; *Statistical Analysis: Multilevel Methods*; Lewis and Langford (2001). See also Langford and Lewis (1998), from which the following passage is quoted:

In a simple educational example, we may have data on examination results in a two-level structure with students nested within schools, and either students or schools may be considered as being outliers at their respective levels in the model. Suppose … that … a particular school is found under test to be a discordant outlier; we … need to ascertain whether it is discordant owing to a systematic difference affecting all the students measured within that school or because one or two students are responsible for the discrepancy. At student level, an individual may be outlying with respect to the overall relationships found across all schools or be unusual only in the context of his or her particular school. … [Again,] masking [see Sect. 3.1] can apply across the levels of a model. The outlying nature of a school may be masked because of another similarly outlying school or by the presence of a small number of students within the school whose influence brings the overall relationships within that school closer to the average for all schools.

6. Conclusion

'The concern over outliers is old and undoubtedly dates back to the first attempt to base conclusions on a set of statistical data. … The literature on outliers is vast … ' (Beckman and Cook 1983). As this article has shown, outliers occur commonly in data and the problem of how to deal with them in any particular case is an unavoidable one. A massive armory of principle, model, and method has built up over the years. For a review of the relevant literature, see the extensive coverage of outlier methodology by Barnett and Lewis (1994), whose book, which includes some 970 references, still remains the only comprehensive treatment of the subject. A valuable survey of many outlier problems is also given in Beckman and Cook (1983).

Bibliography

Barnett V, Lewis T 1994 *Outliers in Statistical Data*, 3rd edn. Wiley, Chichester, UK

Beckman R J, Cook R D 1983 Outlier …… s (with discussion). *Technometrics* **25**: 119–63

Chatterjee S, Hadi A S 1988 *Sensitivity Analysis in Linear Regression*. Wiley, New York

Cook R D, Weisberg S 1982 *Residuals and Influence in Regression*. Chapman and Hall, London

Hawkins D M 1980 *Identification of Outliers*. Chapman and Hall, London

Hoaglin D C, Mosteller F, Tukey J W 1983 *Understanding Robust and Exploratory Data Analysis*. Wiley, New York

Langford I H, Lewis T 1998 Outliers in multilevel data (with discussion). *Journal of the Royal Statistical Society A* **161**: 121–60

Lewis T, Langford I H 2001 Outliers, robustness and the detection of discrepant data. In: Leyland A H, Goldstein H (eds.) *Multilevel Modelling of Health Statistics*. Wiley, Chichester, UK

McKee J M 1994 *Multiple Regression and Causal Analysis*. F E Peacock, Itasca, IL

Newell D J 1964 Statistical aspects of the demand for maternity beds. *Journal of the Royal Statistical Society A* **127**: 1–33

Peirce B 1852 Criterion for the rejection of doubtful observations. *Astronomical Journal* **2**: 161–3

Rousseeuw P J, Leroy A M 1987 *Robust Regression and Outlier Detection*. Wiley, New York

Rousseeuw P J, van Zomeren B C 1990 Unmasking multivariate outliers and leverage points (with discussion). *Journal of the American Statistical Association* **85**: 633–51

Stone E J 1868 On the rejection of discordant observations. *Monthly Notices of the Royal Astronomical Society* **28**: 165–8

von Eye A, Schuster C 1998 *Regression Analysis for Social Sciences*. Academic Press, San Diego, CA

T. Lewis

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Statistical Analysis, Special Problems of: Transformations of Data

Data transformations such as replacing a variable by its logarithm or by its square-root are often used to simplify the structure of the data so that they follow a convenient statistical model or are more suitable for graphing. Transformations have one or more objectives, including:

(a) inducing a simple systematic relationship between a response and predictor variables in regression;

(b) stabilizing a variance, that is, inducing a constant variance in a group of populations or in the residuals after a regression analysis; and

(c) inducing a particular type of distribution, e.g., a normal or symmetric distribution, or removing extreme skewness so that the data are more appropriate for plotting.

The second and third goals are concerned with simplifying the 'error structure' or random component of the data.

The need for transformation can be seen in Fig. 1, showing assets and sales for a sample of 79 'Forbes 500' companies. The data were taken from the Data and Story Library at Carnegie Mellon University's StatLib (http://lib.stat.emu.edu/DASL/). Panel (a) shows that both assets and sales are highly right skewed and that the scatter of sales, conditional on the value of assets, is greater for larger values of assets. Because of this extreme right skewness, most of the data are crowded into the lower left-hand corner of the plot and it is difficult to ascertain what type of regression model is appropriate for these data.

The normal probability plot in panel (b) also shows the right skewness of sales. Panel (c) is a normal plot of the residuals from a linear regression of sales on assets. Notice that the extreme residuals are much larger in absolute value than the corresponding normal quantiles. These outliers are due to the nonconstant conditional variance of sales, given assets; the outliers are from the observations with large values of assets. The skewness and heteroscedasticity in this example is quite typical of economic data.

Figure 2 uses the same raw data as Fig. 1, but in Fig. 2 both assets and sales are log transformed. Notice in panel (a) that log(sales) appear linearly related to

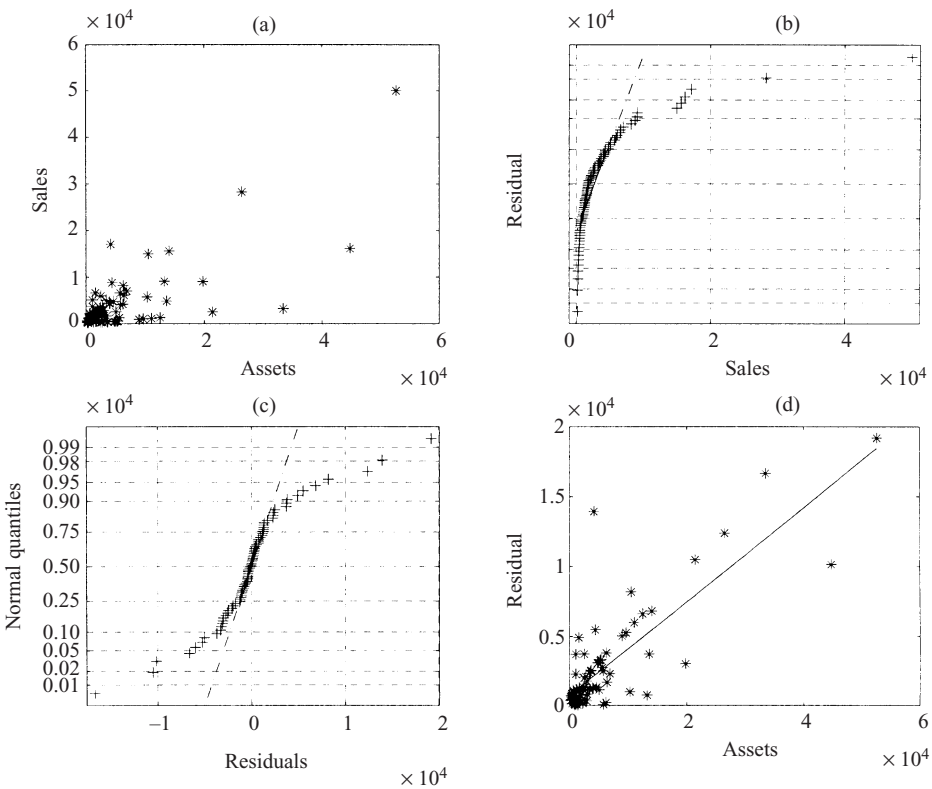

Figure 1
Assets and sales of a sample of Forbes 500 companies: (a) plot of sales and assets; (b) normal probability plot of sales; (c) normal probability plot of residuals from a linear fit of assets to sales; (d) plot of absolute residuals vs. assets with a scatterplot smooth

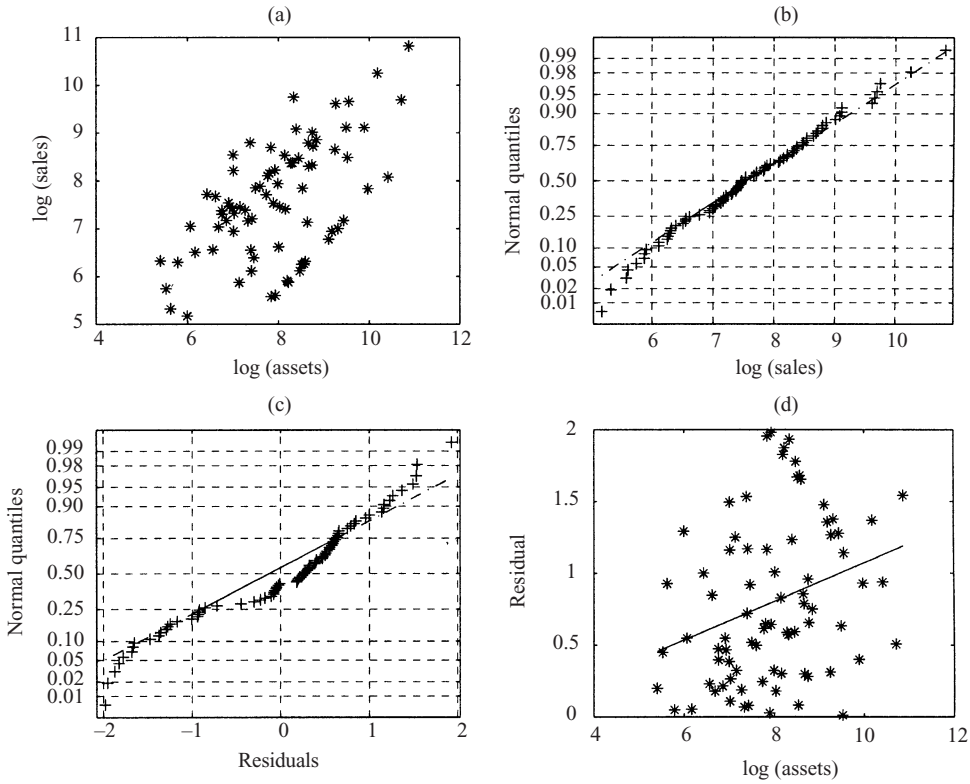

Figure 2
Log(assets) and log(sales) of a sample of Forbes 500 companies. (a) plot of log(sales) and log(assets); (b) normal probability plot of log(sales); (c) normal probability plot of residuals from a linear fit of log(assets) to log(sales); (d) plot of absolute residuals versus log(assets) with a scatterplot smooth

log(assets). The normal probability plots in panels (b) and (c) suggest that both log(sales) and the residuals from the regression of log(sales) against log(assets) are approximately normally distributed. Panel (d) shows that these residuals are far less heteroscedastic than those from the untransformed data.

Data transformations were developed at a time when there were far fewer tools available to the applied statistician than now. Many problems that in the past would have been tackled by data transformation plus a linear model are now solved using more sophisticated techniques; for example, generalized linear models, nonlinear models, smoothing, and variance function models. Nonetheless, model simplification is still important, and data transformation is still an effective and widely used tool.

1. *Transformations to Simplify the Systematic Model*

In regression, if the expected value of the response y is related to predictors $x_1, \ldots, x_J$ in a complex way, often a transformation of y will simplify this relationship by inducing linearities or removing interactions. For example, suppose that

$$y = \beta_0 x_1^{\beta_1} \ldots x_J^{\beta_J} + \varepsilon \tag{1}$$

where ε is a small random error and y and $x_1, \ldots, x_J$ are all positive. The Taylor series approximation $\log(\mu+\varepsilon) \approx \log(\mu) + \varepsilon/\mu$ is appropriate for small values of ε. Using this approximation with $\mu = \beta_0 x_1^{\beta_1} \ldots x_J^{\beta_J}$, one finds that $\log y$ follows approximately the linear model

$$\log y = \beta_0^* + \beta_1 x_1^* + \ldots + \beta_J x_J^* + (\beta_0 x_1^{\beta_1} \ldots x_J^{\beta_J})^{-1}\varepsilon \tag{2}$$

where $\beta_0^* = \log(\beta_0)$ and $x_j^* = \log(x_j)$. Model (1) is nonlinear in each x_j and has large and complex interactions.

In contrast, model (2) is linear and additive in the transformed predictors. However, transformations should not be used blindly to linearize a systematic relationship. If the errors in Eqn. (1) are *homo-*

scedastic, then the errors in Eqn. (2) will be *heteroscedastic*. Also, the log transformation can induce skewness and transform the data nearest 0 to outliers. On the other hand, if the original data exhibit heteroscedasticity and skewness, then these problems might be reduced by transformation. Any intelligent use of transformations requires that the effects of the transformations on the error structure be understood; see Ruppert et al. (1989), where the use of transformations to linearize the Michaelis–Menten equation is discussed.

In our example of Forbes 500 companies, log(sales) seems linearly related to log(assets), which suggest that the untransformed data satisfy

$$\text{sales} \approx \beta_0 \text{assets}^{\beta_1}. \tag{3}$$

Moreover, the log transformed data is less skewed and heteroscedastic than the original data, so transformation brings the data more closely in line with usual assumptions of linear regression.

2. *Transformations to Simplify Error Structure*

The effects of a transformation upon the distribution of data can be understood by examining the transformation's derivative. A nonlinear transformation tends to stretch out data where the transformation's derivative is large in absolute value, and to squeeze together data where the absolute value of that derivative is small. These effects are illustrated in Fig. 3(a) where the logarithm transformation is applied to normalize right-skewed lognormally distributed data. Along the x-axis one sees the probability density function of the lognormal data. The log transformation is shown as a solid curve. On the y-axis is the normal density of the transformed data. The mapping of the 5th percentile of the original data to the same percentile of the transformed data is shown with dashed lines, and similar mappings for the median and 95th percentiles are shown with *solid and dashed lines*.

The log is a concave transformation; that is, its derivative is decreasing. Therefore the long right tail

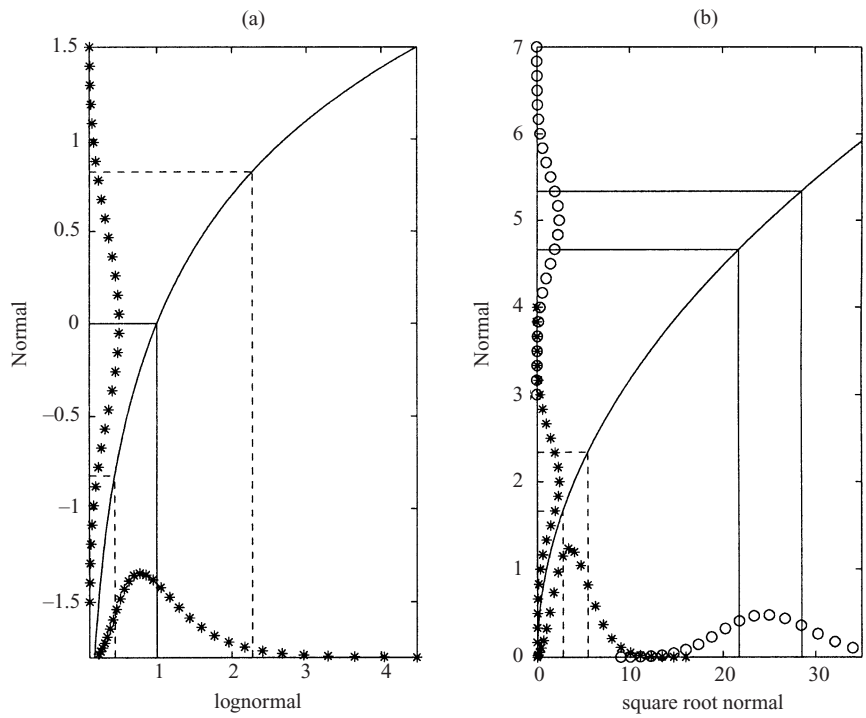

Figure 3
(a) A transformation to remove right skewness. Here the concave log transformation converts right-skewed lognormal data to normally distributed data. The densities of the original and transformed data are shown on the x and y axes, respectively. The lines show the mappings of the 5th, 50th, and 95th percentiles; (b) A variance-stabilizing transformation. Here the square-root transformation converts heteroscedastic data consisting of two populations to data with a constant variance; '*' and 'o' denote the two populations. The lines show the mappings of the first and third quartiles of the two populations

and short left tail of the lognormal distribution are shrunk and stretched, respectively, by the mapping to produce a symmetric distribution of the transformed data. The density on the y-axis is evaluated at equally spaced points. Their inverse images on the x-axis are spread out increasingly as one moves from left to right. The same effect can be seen in the percentiles.

In general, concave transformations are needed to symmetrize right-skewed distributions. Analogously, convex transformations which have increasing first derivatives are used to symmetrize left-skewed distributions.

Data often exhibit the property that populations with larger means also have larger variances. In regression if $E(Y|\mathbf{X}) = \mu(\mathbf{X})$ is the mean of the conditional distribution of the response Y given a vector $\mathbf{X}$ of covariates, then we might find that the conditional variance $\mathrm{var}(Y|\mathbf{X})$ is a function of $\mu(\mathbf{X})$, i.e., $\mathrm{var}(Y|\mathbf{X}) = g\{\mu(\mathbf{X})\}$ for some function g. If Y is transformed to $h(Y)$, then a delta-method linear approximation (Bartlett 1947) shows that

$$\mathrm{var}(h(Y)|\mathbf{X}) \approx \{h'(\mu(\mathbf{X})\}^2\, g\{\mu(\mathbf{X})\}. \qquad (4)$$

The transformation h stabilizes the variance if $\{h'(y)\}^2\, g(y)$ is constant. For example, if $g(y) \propto y^{\alpha}$, then $h(y) \propto y^{1-\alpha/2}$ is variance-stabilizing for $\alpha \neq 2$. For Poisson data, $\alpha = 1$ and the square-root transformation is variance-stabilizing. For data with a constant coefficient of variation, $\alpha = 2$ and the log transformation is variance-stabilizing. Exponentially distributed data, or, more generally, gamma-distributed data with a constant shape parameter, are examples with a constant coefficient of variation. For binomial data with a constant sample size, n, $g(y) = ny(1-y)$ and the 'angular' transformation, $\sin^{-1}(\sqrt{y})$, is variance-stabilizing.

When g is an increasing function, then the variance-stabilizing transformation will be concave, as illustrated in Fig. 3(b). The original data on the x-axis are such that their square roots are normally distributed data, except that they are truncated in the extreme tails to be strictly positive. There are two original populations with different means and variances. After a square-root transformation, the two populations have different means but equal variances. The dashed lines show the mapping of the first and third quartiles of the first population of the original data to their transformed values. The solid lines show the same mapping of the second population.

When $\mathrm{var}(Y|\mathbf{X})$ is not a function of $\mu(\mathbf{X})$, then there is no variance-stabilizing transformation, and variance function modeling (Carroll and Ruppert 1988) should be used as an alternative, or at least an adjunct, to transformation of Y.

3. *Power Transformations*

There are several parametric families of data transformation in general use. The most important are the power or Box and Cox (1964) family and the power-plus-shift family. The Box–Cox power transformations, that are used for positive y, are:

$$\begin{aligned} h(y;\lambda) &= (y^{\lambda}-1)/\lambda \quad \lambda \neq 0 \\ &= \log(y) \qquad \lambda = 0. \end{aligned}$$

For $\lambda \neq 0$ the ordinary power transformation, y^{λ}, is shifted and rescaled so that $h(y;\lambda)$ converges to $\log(y)$ as λ converges to 0. This embeds the log transformation into the power transformation family in a continuous fashion. The shift and rescaling are linear transformations and have no effect on whether the transformation achieves the three objectives enumerated in Sect. 1. Power transformations with $\lambda = -1/2, 0, 1/2, 1, 1.5, 2$, and 2.5 are shown in Fig. 4. Notice that the transformations are upward bending (convex) when $\lambda > 1$ and downward bending (concave) when $\lambda < 1$. Transformations which are either convex or concave are called 'one-bend' by Kruskal (1968).

As mentioned before, power transformations are variance-stabilizing when the variance is a power of the mean, a model that often fits well in practice. The cube-root transformation, called the Wilson–Hilferty (1931) transformation, approximately normalizes, and therefore symmetrizes, gamma-distributed data.

The effect of $h(y;\lambda)$ on data depends on its derivative $h'(y;\lambda) = (\partial/\partial y)h(y;\lambda) = y^{\lambda-1}$. Consider two points $0 < y_1 < y_2$. The 'strength' of the transformation can be measured by the ratio.

$$\frac{h'(y_2;\lambda)}{h'(y_1;\lambda)} = \left(\frac{y_2}{y_1}\right)^{\lambda-1}$$

which shows how far data are spread out at y_2 relative to y_1. This ratio is one for $\lambda = 1$ (linear transformation), is less than one for $\lambda < 1$ (concave transformations), and is greater than one for $\lambda > 1$ (convex transformation). For fixed y_1 and y_2, the ratio is increasing in λ, so that its concavity or convexity becomes stronger as λ moves away from 1. If a certain value of λ transforms a right-skewed distribution to symmetry, then a larger value of λ will not remove right-skewness completely, while a smaller value of λ will induce some left-skewness. Analogously, if the variance is an increasing function of the mean, and a certain value of λ is variance-stabilizing, then for a larger value of λ the variance of the transformed data is still increasing in the mean, while for a smaller value of λ, the variance is decreasing in the mean. Understanding this behavior is important when a symmetrizing or variance-stabilizing power transformation is sought by trial and error.

The shifted-power transformation is

$$h(y;\lambda,\theta) = h(y+\theta;\lambda)$$

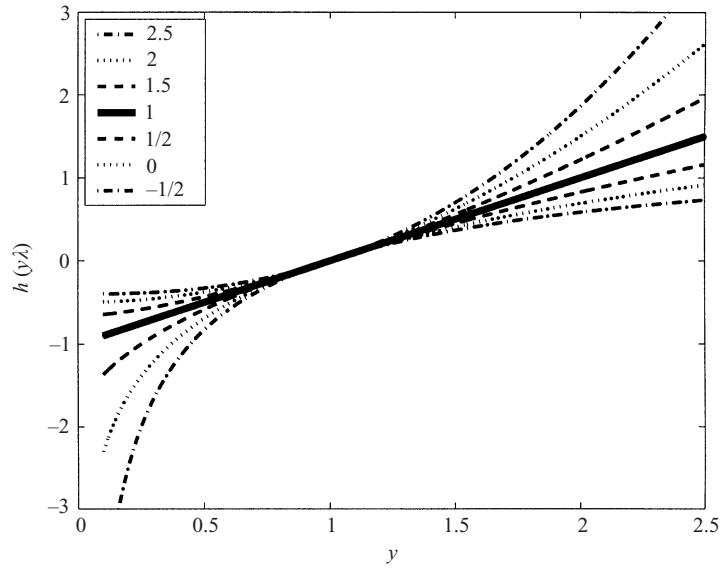

Figure 4
Box–Cox power transformations. The linear transformation with $\lambda = 1$ is shown as a heavy solid line. The convex transformations with $\lambda > 1$ are above this line and the concave transformations with $\lambda < 1$ are below

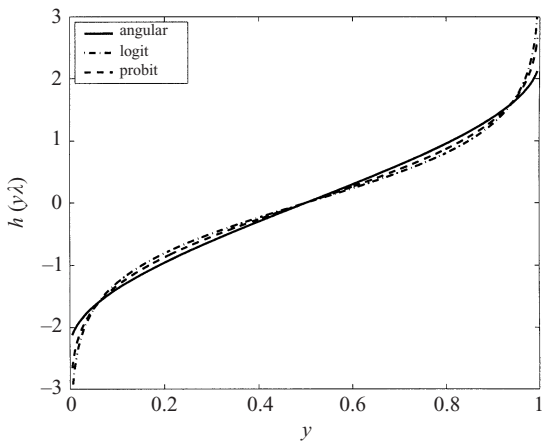

Figure 5
Angular, logit, and probit transformations. To facilitate comparisons of shapes, each of the transformations was shifted and scaled so that the mean and variance of the transformed values were 0 and 1, respectively

which is a location shift of y to $y+\theta$ followed by a Box–Cox power transformation. The ordinary power transformation can be applied only to nonnegative data, but the shifted power transformation can be used for any data set provided that $\theta > -\min(Y)$, where $\min(Y)$ is the smallest value of the data. Thus, shifted power transformations are particularly well suited to variables that have a lower bound, but no upper bound.

Kruskal (1968) and Atkinson (1985) discuss families of transformations that are appropriate for percentages, proportions, and other variables that are constrained to a bounded interval. The most common transformations for variables constrained to lie between 0 and 1 are the angular transformation, $\sin^{-1}(\sqrt{y})$, the logit transformation, $\log\{y/(1-y)\}$, and the probit transformation, $\Phi^{-1}(y)$, where Φ is the normal (0, 1) cumulative distribution function. These transformations, shown in Fig. 5, are concave for $y < 1/2$ and convex for $y > 1/2$; Kruskal (1968) calls them 'two-bend' transformations.

For data that are bounded neither above nor below, Manly (1976) proposes the exponential data transformations.

$$\begin{aligned} h(y;\lambda) &= \{\exp(\lambda y)-1\}/\lambda, \quad \lambda \neq 0 \\ &= y, \qquad\qquad\qquad\quad \lambda = 0 \end{aligned}$$

which is the transformation from y to $\exp(y)$, a positive variate, followed by the Box–Cox power transformation.

4. *Estimation of Transformation Parameters*

There are situations where a transformation may be selected based on *a priori* information. For example, if we know that model (1) holds, then we know that the log transformation will linearize and remove inter-

actions. In other situations, transformations are often based on the data, though perhaps selected from a predetermined parametric family, e.g., shifted power transformations. Trial and error can be an effective method of parameter selection. For example, one might try power transformations with λ equal to $-1, -1/2, 0, 1/2$, and 1, and see which value of λ produces the best residual plots.

When all aspects of the data are modeled parametrically, then the transformation parameter, as well as other parameters, can be estimated simultaneously by maximum likelihood. Semiparametric transformations are used when the model is not fully parametric, e.g., it is assumed that there is a parametric transformation to symmetry but not to a specific symmetric parametric family.

4.1 Maximum Likelihood Estimation

Maximum likelihood estimation is possible once a parametric model is specified. However, there are at least three possible parametric transformation models in common use: the Box and Cox (1964) model; the transform-both-sides (TBS) model of Carroll and Ruppert (1981); and the Box and Tidwell (1964) model. In practice it is essential to consider which, if any, of these models is most appropriate.

Box and Cox (1964) assumed that after a parametric transformation of the response, the data follow a simple linear model with homoscedastic, normally distributed errors. Therefore, the Box–Cox model is that

$$h(Y;\lambda) = \beta_0 + \beta_1 x_p + \ldots \beta_p x_p + \varepsilon \tag{5}$$

where ε is normal $(0, \sigma^2)$ for some unknown value of λ. They find the likelihood for this model as a function of $(\lambda, \sigma^2, \beta_0, \ldots, \beta_p)$ and show how to maximize likelihood to estimate the transformation parameter, the regression parameters, and the variance of the errors. The Box–Cox model attempts to achieve each of the three objectives in Sect. 1 with a single transformation. Although it may be rare for all three goals to be achieved exactly, often in practice both the systematic model and the error structure can be simplified by the same transformation, and because of this the Box–Cox technique has seen widespread usage. Simultaneous estimation of a power and shift by maximum likelihood is not recommended, for technical reasons (Atkinson 1985), but λ can be estimated by maximum likelihood for a fixed value of the shift, which can be chosen by trial and error.

Model (5) with only an intercept, i.e., $h(Y;\lambda) = \beta_0 + \varepsilon$, is commonly used to choose a normalizing transformation for a univariate sample. When maximum likelihood estimation was applied to this model using the Forbes 500 data, the maximum likelihood estimations of λ were -0.07 and -0.04 for sales and assets, respectively. These values are quite close to the log transformation, $\lambda = 0$, which partially justifies the use of the log transformation for this example.

Carroll and Ruppert (1981, 1988) apply maximum likelihood estimation to a somewhat different parametric model than that of Box and Cox, called TBS. They assume that there is a known theoretical model relating the response to the predictors, i.e., that

$$Y \approx f(x_1, \ldots, x_p; \beta_1, \ldots, \beta_q)$$

where $f(x_1, \ldots, x_p; \beta_1, \ldots, \beta_q)$ is some nonlinear regression model. Transforming only the response will destroy this relationship, so they transform both the response and the theoretical model in the same way. It is assumed that after transforming by the correct λ, the errors are homoscedastic and normally distributed. Therefore, the Carroll–Ruppert model is

$$h(Y;\lambda) = h\{f(x_1, \ldots, x_p; \beta_1, \ldots, \beta_q); \lambda\} + \varepsilon.$$

TBS simplifies only the error structure while leaving the systematic model unchanged.

In the Forbes 500 example, if one starts with the power model (3), then the model used in Fig. 2, i.e.,

$$\log(\text{assets}) = \beta_0^* + \beta_1 \log(\text{assets}) + \varepsilon, \quad \beta_0^* = \log(\beta_0) \tag{6}$$

is the TBS model with $\lambda = 0$.

The model of Box and Tidwell (1962) assumes that the untransformed Y is linear in the transformed predictors, for example, that

$$Y = \beta_0 + \beta_1 h(x_1;\lambda_1) + \ldots + \beta_p h(x_p;\lambda_p) + \varepsilon.$$

This model is a special case of nonlinear regression and can be fit by ordinary nonlinear least-squares. Atkinson (1975) discusses a combination of the Box–Tidwell model and the Box–Cox model, where

$$h(Y;\lambda) = \beta_0 + \beta_1 h(x_1;\lambda_1) + \ldots + \beta_p h(x_p;\lambda_p) + \varepsilon.$$

A special case of Atkinson's model restricts the transformation parameters so that $\lambda = \lambda_1 = \ldots = \lambda_p$. Atkinson's model with $\lambda = \lambda_1 = 0$ is equivalent to the TBS model (6).

4.2 Semiparametric Estimation

Hinkley (1975) estimates a transformation to symmetry by setting a skewness measure of the transformed data to zero and solving for the transformation parameter.

Ruppert and Aldershof (1989) estimate a variance-stabilizing transformation by minimizing the correlation between the squared residuals and the fitted values.

4.3 Effects of Outliers and Robustness

An outlier is an observation that is poorly fit by the model and parameter values that fit the bulk of the data. When the model includes a transformation, outliers can be difficult to identify. For example, if the original data are lognormally distributed, apparent outliers in the long right tail may be seen to be conforming after the normalizing transformation. Conversely, the smallest observations may be quite close to the bulk of the original data but outlying after a log transformation. It is essential that the transformation parameter not be chosen to fit a few outliers but rather to fit the main data set. Diagnostics and robust transformations can be useful for this purpose. The semiparametric estimators discussed above are often chosen to be insensitive to outliers, e.g., in Hinkley (1975), by choosing λ to minimize a robust measure of skewness. Because of the potentially serious effects of not accounting for the effects of outliers properly, users of transformations are advised to consult the monographs by Atkinson (1985) and Carroll and Ruppert (1988) for practical advice.

4.4 Inference after a Transformation

There has been some debate on whether a data-based transformation can be treated as fixed when one makes inferences about the regression parameters (Box and Cox 1982). This is a serious practical question. In the Box–Cox model, the values of the regression parameters are highly dependent on the choice of transformation, and there is a general consensus that inference for regression parameters should treat the transformation parameters as fixed; see Box and Cox 1982 for discussion. If one is estimating the expected response or predicting responses at new values of the predictors, then the predictions are more variable with an estimated transformation than when the correct transformation is known (Carroll and Ruppert 1981). Prediction intervals can be adjusted for this effect to obtain coverage probabilities closer to nominal values (Carroll and Ruppert 1991).

5. Multivariate Transformations

Most of the classical techniques of multivariate analysis assume that the population has a multivariate normal distribution, an assumption that is stronger than that the individual components are univariate normal. Andrews et al. (1971) generalize the Box–Cox model to multivariate samples. They transform each coordinate of the observations using the same transformation family, e.g., power transformations for all coordinates, but with coordinate-specific transformation parameters. It is assumed that within this family of multivariate transformations, there exists a transformation to multivariate normality. All parameters in this model are estimated by maximum likelihood. See Gnanadesikan (1997) and Ruppert (2001) for further discussion.

6. Nonparametric Transformations

For interpretability, a data transformation should be continuous and monotonic, but parametric assumptions are made only for simplicity. A nonparametric estimate of a transformation allows one to select a suitable parametric family of transformation, to check if an assumed family is appropriate, or to forgo parametric assumptions when necessary. Tibshirani's (1988) transformation to additivity and variance stabilization (AVAS) method uses nonparametric variance function estimation and Eqn. (2) to estimate the variance-stabilizing transformation. Nychka and Ruppert (1995) use nonparametric spline-based transformations within the TBS model. Nonparametric transformations should not be confused with semiparametric transformation, as discussed above; in the latter a transformation is selected from within a *parametric* family but the data are not assumed to have a density within a parametric family.

7. Further Information

Atkinson (1985), Carroll and Ruppert (1988), and Draper and Smith (1998) are good places to learn more about the practical aspects of data transformation.

See also: Distributions, Statistical: Approximations; Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics)

Bibliography

Andrews D F, Gnanadesikan R, Warner J L 1971 Transformations of multivariate data. *Biometrics* **27**: 825–40

Atkinson A C 1985 *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press, New York

Bartlett M S 1947 The uses of transformations. *Biometrics* **3**: 39–52

Box G E P, Cox D R 1964 An analysis of transformations (with discussion). *Journal of the Royal Statistical Society*, *Series B* **26**: 211–46

Box G E P, Cox D R 1982 An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association* **77**: 209–10

Box G E P, Tidwell P W 1964 Transformations of the independent variables. *Technometrics* **4**: 531–50

Carroll R J, Ruppert D 1981 On prediction and the power transformation family. *Biometrika* **68**: 606–19

Carroll R J, Ruppert D 1984 Power transformation when fitting theoretical models to data. *Journal of the American Statistical Association* **79**: 321–8

Carroll R J, Ruppert D 1985 Transformations in regression: A robust analysis. *Technometrics* **27**: 1–12

Carroll R J, Ruppert D 1988 *Transformation and Weighting in Regression*. Chapman and Hall, New York
Carroll R J, Ruppert D 1991 Prediction and tolerance intervals with transformation and/or weighting. *Technometrics* **33**: 197–210
Draper N R, Smith H 1998 *Applied Regression Analysis*, 3rd edn. Wiley, New York
Gnanadesikan R 1997 *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York
Hinkley D V 1975 On power transformations to symmetry. *Biometrika* **62**: 101–11
Kruskal J B 1968 Transformations of data. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York
Nychka D, Ruppert D 1995 Nonparametric transformations for both sides of a regression model. *Journal of the Royal Statistical Society, Series B* **57**: 519–32
Manly B F J 1976 Exponential data transformations. *Statistician* **25**: 37–42
Ruppert D 2001 Multivariate transformations. In: El-Shaarawi A H, Piegorech W W (eds.) *Encyclopedia of Environmentrics*. Wiley, New York
Ruppert D, Aldershof B 1989 Transformations to symmetry and homoscedasticity. *Journal of the American Statistical Association* **84**: 437–46
Ruppert D, Carroll R J, Cressie N 1989 A transformation/weighting model for estimating Michaelis–Menten parameters. *Biometrics* **45**: 637–56
Tibshirani R 1988 Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association* **83**: 394–405
Wilson E B, Hilferty M M 1931 The distribution of chi-square. *Proceedings of the National Academy of Sciences The United States of America* **17**: 684–88

D. Ruppert

Statistical Clustering

Every formal inquiry begins with a definition of the objects of the inquiry and of the terms to be used in the inquiry, a *classification* of objects and operations. Since about 1950, a variety of simple explicit classification methods, called clustering methods, have been proposed for different standard types of data. Most of these methods have no formal probabilistic underpinnings. And indeed, classification is not a subset of statistics; data collection, probability models, and data analysis all require informal subjective prior classifications. For example, in data analysis, some kind of binning or selection operation is performed prior to a formal statistical analysis.

However, statistical methods can assist in classification in four ways:

(a) in devising probability models for data and classes so that probable classifications for a given set of data can be identified;

(b) in developing tests of validity of particular classes produced by a classification scheme;

(c) in comparing different classification schemes for effectiveness; and

(d) in expanding the search for optimal classifications by probabilistic search strategies.

These statistical methods are referred to as *cluster analysis* or *statistical classification*. The second term is preferred, because a cluster, from its astronomical origin, tends to connote a crowd of points close together in space, but a class is a set of objects assembled together for any kind of reason, and so the more general term *classification* is to be preferred. Part of the reason for the narrower term *cluster analysis* is that the term *classification* was once used in Statistics to refer to *discriminant analysis* (see *Multivariate Analysis: Classification and Discrimination*). In discriminant analysis, a classification rule is developed for placing a multivariate observation into one of a number of established classes, using the joint distribution of the observations and classes. These problems are regression problems: the variable to be predicted is the class of the observations. For example, logistic regression is a method for discriminant analysis with two classes (see *Multivariate Analysis: Discrete Variables (Logistic Regression)*). More recently, classification in statistics has come to refer to the more general problem of constructing families of classes, a usage which has always been accepted outside of statistics.

1. Algorithms

The first book on modern clustering techniques is Sokal and Sneath (1963). The most comprehensive recent review of a huge diverse literature is Arabie et al. (1996) which includes in particular an excellent survey of algorithms by Arabie and Hubert and of probabilistic methods by Bock.

A classification consists of a set of classes; formal clustering techniques have been mainly devoted to either *partitions* or *hierarchical clustering* (see, however, the nonhierarchic structures in Jardine and Sibson 1968 and the block clusters in Hartigan 1975).

1.1 Partitions

A partition is a family of *classes*, sets of objects with the property that each object lies in exactly one class of the partition. *Ideal-object* partitions are constructed as follows. Begin with a set of given objects, a class of ideal objects, and a numerical dissimilarity between any given object and any ideal object. A fitted ideal object is associated with each class having minimum sum of dissimilarities to all given objects in the class.

The algorithm aims to determine k ideal objects, and a partition of the objects into k classes, so that the dissimilarity between an object and the fitted ideal object of its class, summed over all objects, is minimized. The usefulness of such a clustering is that the

original set of objects is reduced with small loss of information to the smaller set of fitted ideal objects. Except in special circumstances, it is impracticable to search over all possible partitions to minimize the sum of dissimilarities; rather some kind of local optimization is used instead.

For example, begin with some starting partition, perhaps chosen randomly. First fit the ideal objects for a given partition. Second reallocate each object to the class of the fitted ideal object with which it has smallest dissimilarity. Repeat these two steps until no further reallocation takes place. The *k-means* algorithm, MacQueen (1967), is the best known algorithm of this type; each object is represented by a point in p-dimensional space, the dissimilarity between any two points is squared euclidean distance, and the ideal object for each class is the centroid of the class. Asymptotic behaviour of the k-means algorithm is described in Pollard (1981).

1.2 Hierarchical Clustering

Hierarchical clustering constructs *trees* of clusters of objects, in which any two clusters are disjoint, or one includes the other. The cluster of all objects is the root of the tree.

Agglomerative algorithms, Lance and Williams (1967), require a definition of dissimilarity between clusters; the most common ones are *maximum or complete linkage*, in which the dissimilarity between two clusters is the maximum of all pairs of dissimilarities between pairs of points in the different clusters; *minimum or single linkage or nearest neighbor*, Florek et al. (1951), in which the dissimilarity between two clusters is the minimum over all those pairs of dissimilarities; and *average linkage* in which the dissimilarity between the two clusters is the average, or suitably weighted average, over all those pairs of dissimilarities.

Agglomerative algorithms begin with an initial set of singleton clusters consisting of all the objects; proceed by *agglomerating* the pair of clusters of minimum dissimilarity to obtain a new cluster, removing the two clusters combined from further consideration; and repeat this agglomeration step until a single cluster containing all the observations is obtained. The set of clusters obtained along the way forms a hierarchical clustering.

1.3 Object Trees

A different kind of tree structure for classification is a graph on the objects in which pairs of objects are linked in such a way that there is a unique sequence of links connecting any pair of objects. The *minimum spanning tree*, Kruskal (1956), is such a graph, constructed so that the total length of links in the tree is minimized. There may be more than one such optimal graph. The minimum spanning tree is related to single linkage clusters: begin by removing the largest link, and note the two sets of objects that are connected by the remaining links, continue removing links in order of size, noting at each stage the two new sets of objects connected by the remaining links. The family of connected sets obtained throughout the removal process constitutes the single linkage clusters.

In a Wagner tree, the given objects are end nodes of the tree, and some set of *ideal* objects are the possible internal nodes of the tree. A Wagner tree for the given objects is a tree of minimum total length in which the given objects are end nodes and the internal objects are arbitrary. A Wagner tree is a minimum spanning tree on the set of all objects and the optimally placed internal nodes. This tree has application in evolutionary biology as the *most parsimonious* tree reaching the given objects (Fitch and Margoliash 1967).

A third kind of tree structure arises from the CART algorithm of Breiman et al. (1984). This algorithm applies to regression problems in which a given target variable is predicted from a recursively constructed partition of the observation space; the prediction is constant within each member of the partition. If the predictors are numerical variables, the first choice in the recursion looks over all variables V and all cutpoints a for each variable, generating the binary variables identifying observations where $V < a$. The binary predictor most correlated with the target variable divides the data set into two sets, and the same choice of division continues within each of the two sets, generating a *regression tree*. It is a difficult problem to know when to stop dividing. These regression trees constitute a classification of the data adapted to a particular purpose: the accurate prediction of a particular variable.

Usually, only the final partition is of interest in evaluating the regressive prediction, although the regression tree obtained along the way is used to decide which variables are most important in prediction, for example, the variables appearing in the higher levels of the tree are usually the most important. However, if there are several highly correlated variables, only one might appear in the regression tree, although each would be equally useful in prediction.

2. Probabilistic Classification

The aim is to select an appropriate classification C from some family Γ based on given data D.

2.1 Mixture Models

It is assumed that there are a number of objects, denoted $x_1, \ldots, x_n$, which are to be classified. A probability model specifies the probability of an

observation x when x is known to come from a class with parameter value θ, say $p(x|\theta)$. The parameter value θ corresponds to an ideal point in the partition algorithm [Sect. 1.2]. The *mixture* model assumes that there is a fixed number of classes k, that an object x_i comes from class C_j with mixing probability p_j, so that the probability of observation x is $p(x) = \Sigma_j p_j p(x|\theta_j)$, and collectively the observations have probability $\Pi_i p(x_i)$. The parameters p_j and θ_j usually are estimated by maximum likelihood in a non-Bayesian approach. No single partition is produced by the algorithm, but rather the conditional probability of each partition at the maximum likelihood values of the parameters. How do these statistical partitionings compare with the dissimilarity based partitions of Sec. 1.2? Suppose the partition itself is a parameter to be estimated in the statistical model; estimate the optimal partition $C = \{C_1, \ldots, C_k\}$, and the parameters p_j, θ_j to maximize the probability of the data given the partition $\Pi_j \Pi_{x_i \in C_j} p_j p(x_i|\theta_j)$. Thus, the optimal partition in the statistical model is the optimal partition in a dissimilarity model with dissimilarities between the object x_i and the ideal object θ_j

$$d(x_i, \theta_j) = \log(p_j) + \log p(x_i|\theta_j) \tag{1}$$

It is, thus, possible to translate back and forth between statistical models and dissimilarity models by taking dissimilarities to be log probabilities.

However, it is wrong to estimate the optimal partition by maximizing the probability of the data given the partition and parameters over all choices of partitions and parameters; as the amount of data increases, the limiting optimal partition does not correspond to the *correct* partition obtained when the parameters p_j, θ_j are known. (The correct partition assigns x_i to the class that maximizes $p_j p(x_i|\theta_j)$. In contrast, the estimate of p_j, θ_j that maximizes the data probability averaged over all possible partitions, is consistent for the true values under mild regularity conditions, and the partition determined from these maximum likelihood estimates is consistent for the correct partition.

In a Bayesian approach, the probability model is a joint probability distribution over all possible data D and classifications C. The answer to the classification problem is the conditional probability distribution over all possible classifications C given the present data D. Add to the previous assumptions a prior distribution on the unknown parameters p_j, θ_j (Bock 1972). Again, consistent estimates of the correct classification are not obtained by finding the classification C and parameters p, θ of maximum posterior probability. Use the marginal posterior distribution of p, θ, averaging over all classifications, to give consistent maximum posterior density estimates $\hat{p}$, $\hat{\theta}$ say. Then find the classification C of optimal posterior probability C given the data and given $p = \hat{p}$, $\theta = \hat{\theta}$. This classification will converge to the optimal classification for the mixture distribution with known p and θ.

If in addition a prior distribution is specified for the number of clusters, a complete final posterior distribution is available for all possible partitions into any number of components.

2.2 Hierarchical Models

Many of the hierarchical methods assume a dissimilarity function specifying a numerical dissimilarity $d(i, j)$ for each pair of objects i, j. If d satisfies the *ultrametric inequality* for every triple i, j, k

$$d(i, j) \leqslant \max[d(i, k), d(j, k)] \tag{2}$$

then the various hierarchical algorithms discussed in Sect. 1.2 all produce the same clusters. The dissimilarity between a given object and another object in a given cluster C is less than the dissimilarity between that object and another object not in C. Thus hierarchical clustering may be viewed as approximating the given dissimilarity matrix by an ultrametric; methods have been proposed that search for a good approximation based on a measure of closeness between different dissimilarities.

With slightly more generality, *evolutionary distances* between a set of objects conform to a tree in which the objects are end nodes and the internal nodes are arbitrary. A nonnegative link distance is associated with each link of the tree, and the evolutionary distance between two objects is the sum of link distances on the path connecting them. Evolutionary distances are characterised by the following property: any four objects i, j, k, l may be partitioned into two pairs of objects in three ways; for some pairing (i, j), (k, l) say, the *evolutionary equality* is satisfied

$$d(i, k) + d(j, k) = d(i, l) + d(j, k) \tag{3}$$

The pairing (i, j), (k, l) means that the tree path from i to j has no links in common with the tree path from k to l. We can construct an evolutionary tree on a set of objects by approximating a given dissimilarity by an evolutionary distance.

One case where evolutionary distances arise probabilistically is in the factor analysis model in which the objects to be classified form a set of v variables on which a number of independent v-dimensional observations have been made (Edwards and Cavalli-Sforza 1964). The variables are taken to be end-nodes in an evolutionary tree whose links and internal objects remain to be determined. Associated with each link of the tree is an independent normal variable with mean zero and variance depending on the link; this variance is the link distance. Each observation for the v variables is generated by assigning the random link values; then the variable $V_i - V_j$ is a sum of independent normal variables along the path of directed

links connecting i and j. The random normal associated with each link is of opposite sign when the link is traversed in the opposite direction. The quantities $d(i, j) = E(V_i - V_j)^2$ form a set of evolutionary distances, which are estimated by $\hat{d}(i, j) =$ sample have $(V_i - V_j)^2$.

For each observation $\{V_1 - V_i, i = 2, \ldots p\}$ is multivariate normal with a covariance matrix determined by the link variances, and the log likelihood may be written, for a certain weight function w

$$\sum_{i,j} (\hat{d}(i, j) - d(i, j)) w_{i,j}(\sigma) \tag{4}$$

This expression is maximized to estimate the unknown σ values. In this case the sample distances form a sufficient statistic for the parameters, and no information is lost in using distances rather than the complete data set. Similar discrete Markovian models due to Felsenstein (1973) reconstruct evolutionary trees from DNA sequence data.

A nonparametric model for hierarchical clustering assumes that the objects $x_1, \ldots, x_n$ to be clustered are sampled randomly from a space with values $x \in S$ and density f with respect to some underlying measure, usually counting measure or lebesgue measure. It is necessary to have a concept of connectedness in the space S. The family of connected sets of points having density $f(x) > c$, computed for each nonnegative c, together form a hierarchical clustering called *high density* clusters (Bock 1974, Hartigan 1975). The statistical approach estimates this hierarchical clustering on the density f from the given sample $x_1, \ldots x_n$ by first estimating the density f by $\hat{f}$ say, then forming the estimated clusters as the high density clusters in $\hat{f}$. There are numerous parametric and nonparametric estimates of density available.

How do the standard hierarchical clustering methods do in estimating these high density clusters? Badly. The single linkage clusters may be re-expressed as the high density clusters obtained from the nearest-neighbor density estimator; the density estimate at point x is a decreasing function of $d(x, x_i)$, where x_i is the closest sample point to x. This estimator is inconsistent as an estimate of the density, and the high density clusters are correspondingly poorly estimated. However, it may be shown that the single linkage method is able to consistently identify distinct modes in the density. On the other hand, complete linkage corresponds to no kind of density estimate at all, and the clusters obtained from complete linkage are not consistent for anything at all, varying randomly as the amount of data increases. In evaluating and comparing clustering methods, a statistical approach determines how well the various clustering methods recover some known clustering when data is generated according to some model based on the known clustering (Milligan 1980). The most frequently used model is the multivariate normal mixture model.

3. *Validity of Clusters*

A typical agglomerative clustering produces $n-1$ clusters for n objects, and different choices of dissimilarities and cluster combination rules will produce many different clusterings. It is necessary to decide 'which dissimilarity, which agglomeration method, and how many clusters?'.

In partition problems, if a parametric mixture model is used, standard statistical procedures described in Sect. 2.1 are available to help decide how many components are necessary. These models require specific assumptions about the distribution within components that can be critical to the results. The nonparametric *high density* cluster model associates clusters with modes in the density. The question about number of clusters translates into a question about number of modes.

The *threshold or kernel width* test, (Silverman 1981), for the number of modes is based on a kernel density estimator, for example, given a measure of dissimilarity, and a threshold d, use the density estimate at each point x consisting of the sum of $1 - d(x, x_i)/d$ over observed objects x_i within dissimilarity d of x. The number of modes in the population is estimated by the number of modes in the density estimate. To decide whether or not two modes are indicated, the quantity d is increased until at $d = \hat{d}_1$, the estimated density $\hat{f}_1$ first has a single mode. If the value $\hat{d}_1$ required to render the estimate unimodal is large enough, then more than one mode is necessary. To decide when $\hat{d}_1$ is large, resample from the unimodal distribution $\hat{f}_1$, obtaining a distribution of resampled $\hat{d}_1$ values for those samples, the values of d in each of those resamples required to make the density estimate unimodal. The tail probability associated with the given $\hat{d}_1$ is the proportion of resampled $\hat{d}_1$ that exceed the given $\hat{d}_1$. A similar procedure works for deciding whether or not a larger number of modes than 2 is justified. An alternative test that applies in 1-dimension for two modes vs. one uses the DIP statistic (Hartigan and Hartigan 1985), the maximum distance between the empirical distribution function and the unimodal distribution function that is closest to the empirical distribution function. The null distribution is based on sampling from the uniform.

4. *Probabilistic Search*

Consider searching for a classification C in a class Γ that will maximize $l(C)$ the log probability of C given the data. Γ is usually too large for a complete search, and a local optimization method is used. Define a neighborhood of classifications $N(C)$ for each classification C. Beginning with some initial classification C_0, steepest ascent selects C_i to be the neighbor of C_{i-1} that maximizes $l(C)$, and the algorithm terminates at C_m when no neighbor has greater probability. Of course, the local optimum depends on the initial classification

C_0, and the local optimum will not in general be the global optimum.

The search may be widened probabilistically by *randomized steepest ascent or simulated annealing* (Klein and Dubes 1989), as follows: rather than choosing C_i to maximize $l(C)$ in the neighborhood of C_{i-1}, choose C_i with probability proportional to $\exp(\mathrm{l}(\mathrm{C})/\lambda)$ from the neighborhood of C_{i-1}. In running the algorithm, we keep track of the highest probability classification obtained to date, and may return to that classification to restart the search. The degree of randomness λ may be varied during the course of the search. To do randomized ascent at randomness λ add to the log probability of each classification the independent random quantity $-\lambda \log E$, where E is exponential; then do deterministic steepest ascent on the randomized criterion.

To be specific, consider the probabilification of the *k-means* algorithm, that constructs say k clusters of n objects in v dimensions to minimize the total within cluster sum of squares S. Use a probability model in which within cluster deviations are independent normal with variance σ^2, the cluster means are uniformly distributed, and σ^2 has density $\sigma^{-(2vk+4)}$ (these prior assumptions match maximum likelihood), the log probability of a particular classification into two clusters given the data is

$$-[\tfrac{1}{2}(n+1)vk+1]\log S \tag{5}$$

Steepest ascent k-means switches between clusters the observation that most decreases S; randomized steepest ascent switches the observation that most increases the randomized criterion

$$-[\tfrac{1}{2}(n+1)vk+1]\log S-\lambda \log E \tag{6}$$

Finally, to choose the number of clusters, allowing k to vary by permitting an observation to 'switch' to a new singleton cluster. Penalize against many clusters by the BIC penalty; the number of parameters (cluster means and the variance) specifying k clusters in v dimensions is $vk+1$, so the final randomized criterion is

$$-\tfrac{1}{2}nvk\log S-\lambda \log E-\tfrac{1}{2}(vk+1)\log n \tag{7}$$

This formula adjusts only for the increase in log probability due to selecting optimal parameter values, and not for the increase in the log probability caused by selecting the highest probability partitions. A number of alternative choices of information criterion are proposed by Bozdogan (1988).

5. Conclusion

Although probabilistic and statistical methods have much to offer classification, in providing formal methods of constructing classes, in validating those classes, in evaluating algorithms, and in spiffing up algorithms, it is important to realise that classifications have something to contribute to statistics and probability as well. The early extensive development of taxonomies of animals and plants in Biology proceeded quite satisfactorily without any formal statistics or probability through 1950; the big methodological change occurred in the late-nineteenth century following Darwin, who demanded that biological classification should follow genealogy.

Classification is a primitive elementary basic form of representing knowledge that is prior to statistics and probability. Many of the standard algorithms were developed without allowing for uncertainty. Many of them, such as complete linkage, say, do not seem supportable by any statistical reasoning. Others will benefit by statistical evaluations and by probabilistic enhancements, as suggested in Hartigan (2000) and Gordon (2000).

Classification is both fundamental and subjective. If an observer classifies a set of objects to be in a class, for whatever reason, that class is valid in itself. Thus, there is no correct natural classification that our formal techniques can aim for, or that our statistical techniques can test for. Since classification is fundamental, it should be used to express formal knowledge, and probabilistic inferences and predictions should be based on these classificatory expressions of knowledge, Hartigan (2000). More formal and refined classifications may be assisted and facilitated by probability models reflecting and summarizing the prior fundamental classifications.

See also: Algorithmic Complexity; Algorithms; Hierarchical Models: Random and Fixed Effects; Mixture Models in Statistics

Bibliography

Arabie P, Hubert L J, De Soete G (eds.) 1996 *Clustering and Classification*. World Scientific, Singapore

Bock H H 1972 Statistiche Modelle und Bayes'sche Verfahren zur Bestimmung einer unbekannten Klassifikation normalveteilter zufälliger Vektoren. *Metrika* **18**: 120–32

Bock H H 1974 *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung and Strukturierung von Daten* (*Clusteranalyse*). Vandenhoek & Ruprecht, Göttingen, Germany

Bozdogan H 1988 A new model selection criterion. In: Bock H H (ed.) *Classification and Related Methods of Data Analysis*. North Holland, Amsterdam, pp. 599–608

Breiman L, Friedman J, Olshen R, Stone C 1984 *Classification and Regression Trees*. Chapman and Hall, New York

Edwards A W F, Cavalli-Sforza L L 1964 Reconstruction of evolutionary trees. In: Heywood V N, McNeill J (eds.) *Phenetic and Phylogenetic Classification*. The Systematics Association, London, pp. 67–76

Felsenstein J 1973 Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22**: 240–49

Fitch W M, Margoliash E 1967 Construction of phylogenetic trees. *Science* **155**: 279–84

Florek K, Lukaszewicz J, Perkal J, Steinhaus H, Zubrzycki S 1951 Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum* **2**: 282–5
Gordon A D 2000 *Classification*. Chapman and Hall, New York
Hartigan J A 1975 *Clustering Algorithms*. Wiley, New York
Hartigan J A 2000 Classifier probabilities. In: Kiers H A L, Rasson J-P, Groenen P J F, Schader M (eds.) *Data Analysis, Classification, and Related Methods*. Springer-Verlag, Berlin, pp. 3–15
Hartigan J A, Hartigan P M 1985 The dip test of unimodality. *Annals of Statistics* **13**: 70–84
Jardine N, Sibson R 1968 The construction of hierarchic and non-hierarchic classifications. *Computer Journal* **11**: 77–84
Klein R W, Dubes R C 1989 Experiments in projection and clustering by simulated annealing. *Pattern Recognition* **22**: 213–20
Kruskal J B 1956 On the shortest spanning subtree of a graph and the travelling salesman problem. *Proceedings of the American Mathematical Society* **7**: 48–50
Lance G N, Williams W T 1967 A general theory of classificatory sorting strategies. I: Hierarchical systems. *Computer Journal* **9**: 373–80
MacQueen J 1967 Some methods for classification and analysis of multivariate observations. In: LeCam L M, Neyman J (eds.) *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA, pp. 281–97
Milligan G W 1980 An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika* **45**: 325–42
Pearson K 1984 Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London, Series A* **185**: 71–110
Pollard D 1981 Strong consistency of *k*-means clustering. *Annals of Statistics* **9**: 135–40
Silverman B W 1981 Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society B* **43**: 97–9
Sokal R R, Sneath P H A 1963 *Principles of Numerical Taxonomy*. Freeman, San Francisco

J. A. Hartigan

Statistical Data, Missing

1. Introduction

Missing values arise in social science data for many reasons. For example:

(a) In longitudinal studies, data are missing because of *attrition*, i.e., subjects drop out prior to the end of the study.

(b) In most surveys, some individuals provide no information because of noncontact or refusal to respond (*unit* nonresponse). Other individuals are contacted and provide some information, but fail to answer some of the questions (*item* nonresponse). See *Sample Surveys: The Field*.

(c) Information about a variable is partially recorded. A common example in social science is *right censoring*, where times to an event (death, finding a job) are recorded, and for some individuals the event has still not taken place when the study is terminated. The times for these subjects are known to be greater that corresponding to the latest time of observation, but the actual time is unknown. Another example of partial information is *interval censoring*, where it is known that the time to an event lies in an interval. For example in a longitudinal study of employment, it may be established that the event of finding a job took place some time between two interviews.

(d) Often indices are constructed by summing values of particular items. For example, in economic studies, total net worth is a combination of values of individual assets or liabilities, some of which may be missing. If any of the items that form the index are missing, some procedure is needed to deal with the missing data.

(e) Missing data can arise by design. For example, suppose one objective in a study of obesity is to estimate the distribution of a measure Y_1 of body fat in the population, and correlate it with other factors. Suppose Y_1 is expensive to measure but a proxy measure Y_2, such as body mass index, which is a function of height and weight, is inexpensive to measure. Then it may be useful to measure Y_2 and covariates, X, for a large sample but Y_1, Y_2 and X for a smaller subsample. The subsample allows predictions of the missing values of Y_1 to be generated for the larger sample, yielding more efficient estimates than are possible from the subsample alone.

Unless missing data are deliberately incorporated by design (such as in (e)), the most important step in dealing with missing data is to try to avoid it during the data-collection stage. Because data are likely to be missing in any case, however, it is important to try to collect covariates that are predictive of the missing values, so that an adequate adjustment can be made. In addition, the process that leads to missing values should be determined during the collection of data if possible, since this information helps to model the missing-data mechanism when an adjustment for the missing values is performed (Little 1995).

Three major approaches to the analysis of missing data can be distinguished:

(a) discard incomplete cases and analyze the remainder (complete-case analysis), as discussed in Sect. 3;

(b) impute or fill in the missing values and then analyze the filled-in data, as discussed in Sect. 4; and

(c) analyze the incomplete data by a method that does not require a complete (that is, a rectangular) data set.

With regard to (c), we focus in Sects. 6, 7, and 8 on powerful likelihood-based methods, specifically maximum likelihood (ML) and Bayesian simulation. See *Bayesian Statistics*. The latter is closely related to multiple imputation (Rubin 1987, 1996), an extension of single imputation that allows uncertainty in the

imputations to be reflected appropriately in the analysis, as discussed in Sect. 5.

A basic assumption in all our methods is that missingness of a particular value hides a true underlying value that is meaningful for analysis. This may seem obvious but is not always the case. For example, consider a longitudinal analysis of employment; for subjects who leave the study because they move to a different location, it makes sense to consider employment history after relocation as missing, whereas for subjects who die during the course of the study, it is clearly not reasonable to consider employment history after time of death as missing; rather it is preferable to restrict the analysis of employment to individuals who are alive. A more complex missing data problem arises when individuals leave a study for unknown reasons, which may include relocation or death.

Because the problem of missing data is so pervasive in the social and behavioral sciences, it is very important that researchers be aware of the availability of principled approaches that address the problem. Here we overview common approaches, which have quite limited success, as well as principled approaches, which generally address the problem with substantial success.

2. Pattern and Mechanism of Missing Data

The *pattern* of the missing data simply indicates which values in the data set are observed and which are missing. Specifically, let $Y = (y_{ij})$ denote an $(n \times p)$ rectangular dataset without missing values, with ith row $y_i = (y_{i1}, \ldots, y_{ip})$ where y_{ij} is the value of variable Y_j for subject i. With missing values, the pattern of missing data is defined by the *missing-data indicator matrix* $M = (m_{ij})$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present.

Some methods for handling missing data apply to any pattern of missing data, whereas other methods assume a special pattern. An important example of a special pattern is *univariate* nonresponse, where missingness is confined to a single variable. Another is *monotone* missing data, where the variables can be arranged so that $Y_{j+1}, \ldots, Y_p$ are missing for all cases where Y_j is missing, for all $j = 1, \ldots, p-1$. The result is a 'staircase' pattern, where all variables and units to the left of the broken line forming the staircase are observed, and all to the right are missing. This pattern arises commonly in longitudinal data subject to attrition, where once a subject drops out, no more subsequent data are observed.

The missing-data mechanism addresses the reasons why values are missing, and in particular whether these reasons relate to values in the data set. For example, subjects in a longitudinal intervention may be more likely to drop out of a study because they feel the treatment was ineffective, which might be related to a poor value of an outcome measure. Rubin (1976) treated M as a random matrix, and characterized the missing-data mechanism by the conditional distribution of M given Y, say $f(M \mid Y, \phi)$, where ϕ denotes unknown parameters. When missingness does not depend on the values of the data Y, missing or observed, that is, $f(M \mid Y, \phi) = f(M \mid \phi)$ for all Y, ϕ, the data are called missing completely at random (MCAR). MCAR does not mean that the pattern itself is random, but rather that missingness does not depend on the data values. An MCAR mechanism is plausible in planned missing-data designs as in example (e) above, but is a strong assumption when missing data do not occur by design, because missingness often does depend on recorded variables.

Let Y_{obs} denote the observed values of Y and Y_{mis} the missing values. A less restrictive assumption is that missingness depends only on values Y_{obs} that are observed, and not on values Y_{mis} that are missing. That is:

$$f(M \mid Y, \phi) = f(M \mid Y_{obs}, \phi) \quad \text{for all } Y_{mis}, \phi$$

The missing data mechanism is then called missing at random (MAR). Many methods for handling missing data assume the mechanism is MCAR, and yield biased estimates when the data are not MCAR. Better methods rely only on the MAR assumption.

3. Complete-case, Available-case, and Weighting Analysis

A common and simple method is complete-case (CC) analysis, also known as *listwise* deletion, where incomplete cases are discarded and standard analysis methods applied to the complete cases. In many statistical packages this is the default analysis. Valid (but often suboptimal) inferences are obtained when the missing data are MCAR, since then the complete cases are a random subsample of the original sample with respect to all variables. However, complete-case analysis can result in the loss of a substantial amount of information available in the incomplete cases, particularly if the number of variables is large.

A serious problem with dropping incomplete cases is that the complete cases are often easily seen to be a biased sample, that is, the missing data are not MCAR. The size of the resulting bias depends on the degree of deviation from MCAR, the amount of missing data, and the specifics of the analysis. In sample surveys this motivates strenuous attempts to limit unit nonresponse though multiple follow-ups, and surveys with high rates of unit nonresponse (say 30 percent or more) are often considered unreliable for making inferences to the whole population.

A modification of CC analysis, commonly used to handle unit nonresponse in surveys, is to weight respondents by the inverse of an estimate of the probability of response. A simple approach is to form adjustment cells (or subclasses) based on background

variables measured for respondents and nonrespondents; for unit nonresponse adjustment, these are often based on geographical areas or groupings of similar areas based on aggregate socioeconomic data. All nonrespondents are given zero weight and the nonresponse weight for all respondents in an adjustment cell is then the inverse of the response rate in that cell. This method removes the component of nonresponse bias attributable to differential nonresponse rates across the adjustment cells, and eliminates bias if within each adjustment cell respondents can be regarded as a random subsample of the original sample within that cell (i.e., the data are MAR given indicators for the adjustment cells).

With more extensive background information, a useful alternative approach is *response propensity* stratification, where (a) the indicator for unit nonresponse is regressed on the background variables, using the combined data for respondents and nonrespondents and a method such as logistic regression appropriate for a binary outcome; (b) a predicted response probability is computed for each respondent based on the regression in (a); and (c) adjustment cells are formed based on a categorized version of the predicted response probability. Theory (Rosenbaum and Rubin 1983) suggests that this is an effective method for removing nonresponse bias attributable to the background variables when unit nonresponse is MAR.

Although weighting methods can be useful for reducing nonresponse bias, they do have serious limitations. First, information in the incomplete cases is still discarded, so the method is inefficient. Weighted estimates can have unacceptably high variance, as when outlying values of a variable are given large weights. Second, variance estimation for weighted estimates with estimated weights is problematic. See *Estimation: Point and Interval*. Explicit formulas are available for simple estimators such as means under simple random sampling (Oh and Scheuren 1983), but methods are not well developed for more complex problems, and often ignore the component of variability arising from estimating the weight from the data.

Available-case (AC) analysis (Little and Rubin 1987 Sect. 3.3) is a straightforward attempt to exploit the incomplete information by using all the cases available to estimate each individual parameter. For example, suppose the objective is to estimate the correlation matrix of a set of continuous variables $Y_1, \ldots, Y_p$. See *Multivariate Analysis: Overview*. Complete-case analysis uses the set of complete cases to estimate all the correlations; AC analysis uses all the cases with both Y_j and Y_k observed to estimate the correlation of Y_j and Y_k, $1 \leqslant j, k \leqslant p$. Since the sample base of available cases for measuring each correlation includes at least the set of complete cases, the AC method appears to make better use of available information. The sample base changes from correlation to correlation, however, creating potential problems when the missing data are not MCAR or variables are highly correlated. In the presence of high correlations, there is no guarantee that the AC correlation matrix is even positive definite. Haitovsky's (1968) simulations concerning regression with highly correlated continuous data found AC markedly inferior to CC. On the other hand, Kim and Curry (1977) found AC superior to CC in simulations based on weakly correlated data. Simulation studies comparing AC regression estimates with maximum likelihood (ML) under normality (Sect. 6) suggest that ML is superior even when underlying normality assumptions are moderately violated (Little 1988a). Although AC estimates are easy to compute, standard errors are more complex. The method cannot be generally recommended, even under the restrictive MCAR.

4. *Single Imputation*

Methods that impute or fill in the missing values have the advantage that, unlike CC analysis, observed values in the incomplete cases are retained. A simple version of the method imputes missing values by their unconditional sample means based on the observed data. This method can yield satisfactory point estimates of means and totals, but it yields inconsistent estimates of other parameters such as variances, covariance matrices or regression coefficients, and it overstates precision (Little and Rubin 1987, Chap. 3). Thus it cannot be recommended in any generality.

An improvement is *conditional mean* imputation, in which each missing value is replaced by an estimate of its conditional mean given the values of observed values. For example, in the case of univariate nonresponse with $Y_1, \ldots, Y_{p-1}$ fully observed and Y_p sometimes missing, one approach is to classify cases into cells based on similar values of observed variables, and then to impute missing values of Y_p by the within-cell mean from the complete cases in that cell. A more general approach is regression imputation, in which the regression of Y_p on $Y_1, \ldots, Y_{p-1}$ is estimated from the complete cases, and the resulting prediction equation is used to impute the estimated conditional mean for each missing value of Y_p. For a general pattern of missing data, the missing values for each case can be imputed from the regression of the missing variables on the observed variables, computed using the set of complete cases. Iterative versions of this method lead (with some important adjustments) to ML estimates under multivariate normality (Little and Rubin 1987, Chap. 8).

Although conditional mean imputation incorporates information from the observed variables and yields best predictions of the missing values in the sense of mean squared error, it leads to distorted estimates of quantities that are not linear in the data, such as percentiles, correlations and other measures of

association, as well as variances and other measures of variability. A solution to this problem of point estimation is to use random draws rather than best predictions to preserve the distribution of variables in the filled-in data set. An example is *stochastic regression* imputation, in which each missing value is replaced by its regression prediction plus a random error with variance equal to the estimated residual variance.

Hot-deck methods impute actual values observed in the dataset. An example is the Current Population Survey (CPS) hot deck for imputing income items in that survey. For each nonrespondent on one or more income items, the CPS hot-deck finds a matching respondent based on variables that are observed for both; the missing income items for the nonrespondent are then replaced by the respondent's values. When no match can be found for a nonrespondent based on all of the variables, the algorithm searches for a match at a lower level of detail, obtained by omitting some variables and collapsing the categories of others. A more general approach to hot-deck imputation matches nonrespondents and respondents using a distance function based on the observed variables. For example, *predictive mean matching* (Little 1988b) that have similar predictive means from a regression of the missing variable on observed variables. This method is somewhat robust to misspecification of the regression model used to create the matching metric.

The imputation methods discussed so far assume the missing data are MAR. In contrast, models that are not missing at random (NMAR) assert that even if a respondent and nonrespondent to Y_p appear identical with respect to observed variables $Y_1, \ldots, Y_{p-1}$, their Y_p-values differ systematically. Lillard et al. (1986), for example, discussed NMAR imputation models for missing CPS data. It is also possible to create a NMAR hot-deck procedure; for example, respondents' values that are to be imputed to nonrespondents could be multiplied by an inflation or deflation factor that depends on the variables that are observed. A crucial point about the use of NMAR models is that there is never direct evidence in the data to address the validity of their underlying assumptions. Thus, whenever NMAR models are being considered, it is prudent to consider several NMAR models and explore the sensitivity of analyses to the choice of model (Rubin 1977, Little and Wang 1996).

5. *Multiple Imputation*

A serious defect with imputation is that it invents data. A single imputed value cannot represent the uncertainty about which value to impute, so analyses that treat imputed values just like observed values generally underestimate uncertainty, even if nonresponse is modeled correctly and randomly drawn imputations are created. Large-sample results show that for simple situations with 30 percent of the information missing, single imputation under the correct model results in nominal 90 percent confidence intervals having actual coverages below 80 percent. The inaccuracy of nominal levels is even more extreme in multiparameter testing problems (Rubin 1987, Chap. 4).

A modification of imputation that fixes this problem is multiple imputation (MI) (Rubin 1987). Instead of imputing a single set of draws for the missing values, a set of M (say $M = 5$) datasets are created, each containing different sets of draws of the missing values from their joint predictive distribution. We then apply the standard complete-data analysis to each of the M datasets and combine the results in a simple way. In particular for scalar estimands, the MI estimate is the average of the estimates from the M datasets, and the variance of the estimate is the average of the variances from the five datasets plus $1+1/M$ times the sample variance of the estimates over the M datasets (The factor $1+1/M$ is a small-M correction). The last quantity here estimates the contribution to the variance from imputation uncertainty, missed (i.e., set to zero) by single imputation methods. Another benefit of multiple imputation is that the averaging over datasets results in more efficient point estimates than does single random imputation. Often MI is not much more difficult than doing a single imputation—the additional computing from repeating an analysis M times is not a major burden, and methods for combining inferences are straightforward. Most of the work involves generating good predictive distributions for the missing values, which is needed for consistent point estimation in any case. A more detailed discussion of multiple imputation appears elsewhere in this encyclopedia.

6. *Maximum Likelihood for Ignorable Models*

There are statistical methods that allows the analysis a nonrectangular dataset without having to impute the missing values. In particular, maximum likelihood (ML) avoids imputation by formulating a statistical model and basing inference on the likelihood function of the incomplete data. Define Y and M as above, and let X denote an $(n \times q)$ matrix of fixed covariates, assumed fully observed, with ith row $x_i = (x_{i1}, \ldots, x_{iq})$ where x_{ij} is the value of covariate X_j for subject i. Covariates that are not fully observed should be treated as random variables and modeled with the set of Y_j's (Little and Rubin 1987 Chap. 5). The data and missing-data mechanism are modeled in terms of a joint distribution for Y and M given X. *Selection models* specify this distribution as:

$$f(Y, M \mid X, \theta, \psi) = f(Y \mid X, \theta) f(M \mid Y, X, \psi) \quad (1)$$

where $f(Y \mid X, \theta)$ is the model in the absence of missing values, $f(M \mid Y, X, \psi)$ is the model for the missing-

data mechanism, and θ and ψ are unknown parameters. The likelihood of θ, ψ given the data Y_{obs}, M, X is then proportional to the density of Y_{obs}, M given X regarded as a function of the parameters θ, ψ, and this is obtained by integrating out the missing data Y_{mis} from Eqn. (1), that is:

$$L(\theta, \psi \mid Y_{obs}, M, X) = \text{const} \times \int f(Y, M \mid X, \theta, \psi)\, dY_{mis} \quad (2)$$

The likelihood of θ *ignoring the missing-data mechanism* is obtained by integrating the missing data from the marginal distribution of Y given X, that is:

$$L(\theta \mid Y_{obs}, X) = \text{const} \times \int f(Y \mid X, \theta)\, dY_{mis} \quad (3)$$

The likelihood (3) is easier to work with than (2) since it is easier to handle computationally; more important, avoids the need to specify a model for the missing-data mechanism, about which little is known in many situations. Hence it is important to determine when valid likelihood inferences are obtained from (3) instead of the full likelihood (2). Rubin (1976) showed that valid inferences about θ are obtained from (3) when the data are MAR, that is:

$$p(M \mid X, Y, \psi) = p(M \mid X, Y_{obs}, \psi) \quad \text{for all } Y_{mis}, \psi$$

If in addition θ and ψ are distinct in the sense that they have disjoint sample spaces, then likelihood inferences about θ based on (3) are equivalent to inferences based on (2); the missing-data mechanism is then called *ignorable* for likelihood inferences. Large-sample inferences about θ based on an ignorable model are based on ML theory, which states that under regularity conditions

$$\theta - \hat{\theta} \sim N_k(0, C) \quad (4)$$

where $\hat{\theta}$ is the value of θ that maximizes (3), and $N_k(0, C)$ is the k-variate normal distribution with mean zero and covariance matrix C, given by the inverse of an information matrix; for example $C = I^{-1}(\hat{\theta})$ where I is the observed information matrix $I(\theta) = -\partial^2 \log L(\theta \mid Y_{obs}, X)/\partial\theta\partial\theta^T$, or $C = J^{-1}(\hat{\theta})$ where $J(\theta)$ is the expected value of $I(\theta)$. As in (Little and Rubin 1987), Eqn. (4) is written to be open to a frequentist interpretation if $\hat{\theta}$ is regarded as random and θ fixed, or a Bayesian interpretation if θ is regarded as random and $\hat{\theta}$ fixed. Thus if the data are MAR, the likelihood approach reduces to developing a suitable model for the data and computing $\hat{\theta}$ and C.

In many problems maximization of the likelihood to compute $\hat{\theta}$ requires numerical methods. Standard optimization methods such as Newton-Raphson or Scoring can be applied. Alternatively, the *Expectation-Maximization* (EM) algorithm (Dempster et al. 1977) can be applied, a general algorithm for incomplete data problems that provides an interesting link with imputation methods. The history of EM, which dates back to the beginning of this century for particular problems, is sketched in Little and Rubin (1987) and McLachlan and Krishnan (1997).

7. *The EM Algorithm for Ignorable Models*

For ignorable models, let $L(\theta \mid Y_{obs}, Y_{mis}, X)$ denotes the likelihood of θ based on the hypothetical complete data $Y = (Y_{obs}, Y_{mis})$ and covariates X. Let $\theta^{(t)}$ denote an estimate of θ at iteration t of EM. Iteration $t+1$ consists of an E-step and M-step. The E-step computes the expectation of $\log L(\theta \mid Y_{obs}, Y_{mis}, X)$ over the conditional distribution of Y_{mis} given Y_{obs} and X, evaluated at $\theta = \theta^{(t)}$, that is, $Q(\theta \mid \theta^{(t)}) = \int \log L(\theta \mid Y_{obs}, Y_{mis}, X) f(Y_{mis} \mid Y_{obs}, X, \theta^{(t)})\, dY_{mis}$. When the complete data belong to an exponential family with complete-data sufficient statistics S, the E-step simplifies to computing expected values of these statistics given the observed data and $\theta = \theta^{(t)}$, thus in a sense 'imputing' the sufficient statistics. The M-step determines $\theta^{(t+1)}$ to maximize $Q(\theta \mid \theta^{(t)})$ with respect to θ. In exponential family cases, this step is the same as for complete data, except that the complete-data sufficient statistics S are replaced by their estimates from the E-step. Thus the M-step is often easy or available with existing software, and the programming work is mainly confined to E-step computations. Under very general conditions, each iteration of EM strictly increases the loglikelihood, and under more restrictive but still general conditions, EM converges to a maximum of the likelihood function. If a unique finite ML estimate of θ exists, EM will find it.

Little and Rubin (1987) provided many applications of EM to particular models, including: multivariate normal data with a general pattern of missing values and the related problem of multivariate linear regression with missing data; robust inference based on multivariate t models; loglinear models for multiway contingency tables with missing data; and the general location model for mixtures of continuous and categorical variables, which yields ML algorithms for logistic regression with missing covariates. An extensive bibliography of the myriad of EM applications is given in Meng and Pedlow (1992).

The EM algorithm is reliable but has a linear convergence rate determined by the fraction of missing information. Extensions of EM (McLachlan and Krishnan 1997) can have much faster convergence: ECM, ECME, AECM, and PX-EM (Liu et al. 1998). EM-type algorithms are particularly attractive in problems where the number of parameters is large. Asymptotic standard errors based on bootstrap methods (Efron 1994) are also available, at least in simpler situations. An attractive alternative is to switch to a Bayesian simulation method that simulates the posterior distribution of θ (see Sect. 9).

8. Maximum Likelihood for Nonignorable Models

Nonignorable, non-MAR models apply when missingness depends on the missing values (see Sect. 2). A correct likelihood analysis must be based on the full likelihood from a model for the joint distribution of Y and M. The standard likelihood asymptotics apply to nonignorable models providing the parameters are identified, and computational tools such as EM, also apply to this more general class of models. However, often information to estimate both the parameters of the missing-data mechanism and the parameters of the complete-data model is very limited, and estimates are sensitive to misspecification of the model. Often a sensitivity analysis is needed to see how much the answers change for various assumptions about the missing-data mechanism.

There are two broad classes of models for the joint distribution of Y and M. *Selection* models model the joint distribution as in Eqn. (1). *Pattern-mixture* models specify:

$$f(Y, M \mid X, \pi, \phi) = f(Y \mid X, M, \phi) f(M \mid X, \pi) \quad (5)$$

where ϕ and π are unknown parameters, and now the distribution of Y is conditioned on the missing-data pattern M (Little 1993). Equations (1) and (5) are simply two different ways of factoring the joint distribution of Y and M. When M is independent of Y, the two specifications are equivalent with $\theta = \phi$ and $\psi = \pi$. Otherwise (1) and (5) generally yield different models. Rubin (1977) and Little (1993) also consider *pattern-set mixture* models, a more general class of models that includes pattern-mixture and selection models as special cases.

Although most of the literature on missing data has concerned selection models of the form (1) for univariate nonresponse, pattern-mixture models seem more natural when missingness defines a distinct stratum of the population of intrinsic interest, such as individuals reporting 'don't know' in an opinion survey. However, pattern-mixture models can also provide inferences for parameters θ of the complete-data distribution, by expressing the parameters of interest as functions of the pattern-mixture model parameters ϕ and π. An advantage of the pattern-mixture modeling approach over selection models is that assumptions about the form of the missing-data mechanism are sometimes less specific in their parametric form, since they are incorporated in the model via parameter restrictions. This idea is explained in specific normal models in Little and Wang (1996), for example.

Heitjan and Rubin (1991) extended the formulation of missing-data problems via the joint distribution of Y and M to more general incomplete data problems involving coarsened data. The idea is replace the binary missing-data indicators by coarsening variables that map the y_{ij} values to coarsened versions $z_{ij}(y_{ij})$. Full and ignorable likelihoods can be defined for this more general setting. This theory provides a bridge between missing-data theory and theories of censoring in the survival analysis literature. The EM algorithm can be applied to ignorable models by including the missing data (or coarsening) indicator in the complete-data likelihood.

9. Bayesian Simulation Methods

Maximum likelihood is most useful when sample sizes are large, since then the log-likelihood is nearly quadratic and can be summarized well using the ML estimate of θ and its large sample variance-covariance matrix. When sample sizes are small, a useful alternative approach is to add a prior distribution for the parameters and compute the posterior distribution of the parameters of interest. For ignorable models this is

$$\begin{aligned} p(\theta \mid Y_{obs}, M, X) &\equiv p(\theta \mid Y_{obs}, X) \\ &= \text{const} \times p(\theta \mid X) \times f(Y_{obs} \mid X, \theta) \end{aligned}$$

where $p(\theta \mid X)$ is the prior distribution and $f(Y_{obs} \mid X, \theta)$ is the density of the observed data. Since the posterior distribution rarely has a simple analytic form for incomplete-data problems, simulation methods are often used to generate draws of θ from the posterior distribution $p(\theta \mid Y_{obs}, M, X)$. Data augmentation (Tanner and Wong 1987) is an iterative method of simulating the posterior distribution of θ that combines features of the EM algorithm and multiple imputation, with M imputations of each missing value at each iteration. It can be thought of as a small-sample refinement of the EM algorithm using simulation, with the imputation step corresponding to the E-step (random draws replacing expectation) and the posterior step corresponding to the M-step (random draws replacing MLE).

An important special case of data augmentation arises when M is set equal to one, yielding the following special case of the 'Gibbs' sampler.' Start with an initial draw $\theta^{(0)}$ from an approximation to the posterior distribution of θ. Given a value $\theta^{(t)}$ of θ drawn at iteration t:

(a) Draw $Y_{mis}^{(t+1)}$ with density $p(Y_{mis} \mid Y_{obs}, X, \theta^{(t)})$;

(b) Draw $\theta^{(t+1)}$ with density $p(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}, X)$.

The procedure is motivated by the fact that the distributions in (a) and (b) are often much easier to draw from than the correct posterior distributions $p(Y_{mis} \mid Y_{obs}, X)$ and $p(\theta \mid Y_{obs}, X)$. The iterative procedure can be shown in the limit to yield a draw from the joint posterior distribution of Y_{mis}, θ given Y_{obs}, X. The algorithm was termed chained data augmentation in Tanner (1991). This algorithm can be run independently K times to generate K iid draws from the approximate joint posterior distribution of θ and Y_{mis}.

Techniques for monitoring the convergence of such algorithms are discussed in a variety of places (Gelman et al. 1995, Sect. 11.4). Schafer (1996) developed data

augmentation algorithms that use iterative Bayesian simulation to multiply-impute rectangular data sets with arbitrary patterns of missing values when the missing-data mechanism is ignorable. The methods are applicable when the rows of the complete-data matrix can be modeled as iid observations from the multivariate normal, multinomial loglinear or general location models.

10. Overview

Our survey of methods for dealing with missing data is necessarily brief. In general, the advent of modern computing has made more principled methods of dealing with missing data practical (e.g., multiple imputation, Rubin 1996). Such methods are finding their way into current commercial and free software. Schafer (1996) and the forthcoming new edition of Little and Rubin (1987) are excellent places to learn more about missing data methodology.

Bibliography

Dempster A P, Laird N M, Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B* **39**: 1–38

Efron B 1994 Missing data, imputation and the bootstrap. *Journal of the American Statistical Association* **89**: 463–79

Gelman A, Carlin J, Stern H, Rubin D B 1995 *Bayesian Data Analysis*. Chapman & Hall, New York

Haitovsky Y 1968 Missing data in regression analysis. *Journal of the Royal Statistical Society B* **30**: 67–81

Heitjan D, Rubin D B 1991 Ignorability and coarse data. *Annals of Statistics* **19**: 2244–53

Kim J O, Curry J 1977 The treatment of missing data in multivariate analysis. *Sociological Methods and Research* **6**: 215–40

Lillard L, Smith J P, Welch F 1986 What do we really know about wages: the importance of nonreporting and census imputation. *Journal of Political Economy* **94**: 489–506

Little R J A 1988a Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics* **37**: 23–38

Little R J A 1988b Missing data adjustments in large surveys. *Journal of Business and Economic Statistics* **6**: 287–301

Little R J A 1993 Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**: 125–34

Little R J A 1995 Modeling the drop-out mechanism in longitudinal studies. *Journal of the American Statistical Association* **90**: 1112–21

Little R J A, Rubin D B 1987 *Statistical Analysis with Missing Data*. Wiley, New York

Little R J A, Wang Y-X 1996 Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* **52**: 98–111

Liu C H, Rubin D B, Wu Y N 1998 Parameter expansion for EM acceleration: the PX-EM algorithm. *Biometrika* **85**: 755–70

McLachlan G J, Krishnan T 1997 *The EM Algorithm and Extensions*. Wiley, New York

Meng X L, Pedlow S 1992 EM: a bibliographic review with missing articles. *Proceedings of the Statistical Computing Section, American Statistical Association 1992*, 24–27

Oh H L, Scheuren F S 1983 Weighting adjustments for unit nonresponse. In: Madow W G, Olkin I, Rubin D B (eds.) *Incomplete Data in Sample Surveys, Vol. 2, Theory and Bibliographies*. Academic Press, New York, pp. 143–84

Rosenbaum P R, Rubin D B 1983 The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55

Rubin D B 1976 Inference and missing data. *Biometrika* **63**: 581–92

Rubin D B 1977 Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* **72**: 538–43

Rubin D B 1987 *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York

Rubin D B 1996 Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**: 473–89 [with discussion, 507–15; and rejoinder, 515–7]

Schafer J L 1996 *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London

Tanner M A 1991 *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. Springer-Verlag, New York

Tanner M A, Wong W H 1987 The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**: 528–50

R. J. Little and D. B. Rubin

Statistical Identification and Estimability

A statistical identification problem arises when a sample of infinite size, assumed to be generated by an element of a class of statistical models indexed by a parameter, does not lead to certain knowledge of one or more components of that parameter. Roughly, a model is identified if enough observed data could allow one to distinguish elements of the parameter set from one another.

Statistical identification problems are ubiquitous, occurring in fields as diverse as reliability, optimal control theory, experimental design, and epidemiology. A rich array of statistical identification problems appear in the social sciences: latent structure analysis, factor analysis, and analysis of covariance structures are based on multivariate statistical models that possess, by construction, this problem. Nonresponse and data censoring in the conduct of sample surveys often lead to an identification problem and the problem is a central feature of the analysis of econometric structural equation models and of errors-in-variables regression models. Experimental designs in which treatments are confounded lead to an identification problem. A nonparametric identification problem occurs in the theory of competing risks.

Understanding if a statistical model is identified, non-identified, or unidentified is a necessary prelude to the construction of a sensible method for making inferences about model parameters from observed data.

1. Background

The problem of statistical identifiability appears in economics as early as Pigou (1910). Thurstone discusses identification in his applications of factor analysis to intelligence (1935, 1947) and, in his latent structure analysis of opinion research and attitudes, Lazarsfeld deals with identification issues (1950).

Principal contributions to a rigorous mathematical treatment of the problem began in the 1950s with the work of Haavelmo (1944), Hurwicz (1950), Koopmans et al. (1950) and Koopmans and Reiersøl (1950) and spawned a huge literature on the subject. Fisher (1966) provides a thorough rigorous treatment of identification for simultaneous equation systems. He laid to rest the issue of whether or not identifiability is a discrete or continuous property of parameter estimation by demonstrating that estimators which are consistent given a correct model specification converge in probability to true parameter values as specification error approaches zero. Thus identifiability may be regarded as a continuous property of a set of parameters and small specification errors lead to small inconsistencies in parameter estimation. Rothenberg (1971) and Bowden (1973) extend identification theory to embrace nonlinear models, distinguishing local from global identifiability. Hannan (1969, 1971) generalizes the theory of identification for linear structural equation systems with serially uncorrelated residual errors in another direction: he establishes conditions for identifiability for the important case of auto-regressive moving average error terms (see *Time Series: General*).

Probabilistic models employed in linear system optimal control theory have much in common with econometric structural (simultaneous) equation systems in which observations are measured with random error. In the psychometric literature, the later class of models goes under the acronym LISREL (Linear Structural Relationships). Jöreskog and Sörbom (1984) have developed methods for specifying LISREL model structure, algorithms for calculating various types of parameter estimators and a framework for interpretation of results. As with econometric structural equation systems, applications of LISREL span a wide domain of social science applications, marketing, psychology, sociology, and organizational behavior among them.

A Bayesian treatment of identification began with Dawid (1979) and Kadane's (1974) studies of identification in Bayesian theory and Drèze's (1972), (1974) analysis of econometric simultaneous equation systems from a Bayesian perspective. Identification through a Bayesian lens is currently a vigorous area of research yielding new insights into how to construct meaningful prior distributions for parameters.

2. Example 1

A sample survey is conducted to determine attitudes of individuals in a population toward a particular subject. The sample frame consists of N individuals, each of whose response is classified as either 'Favorable' or 'Unfavorable.' This sample yields $N_{R\cdot}$ respondents and $N_{S\cdot} = N - N_{R\cdot}$ non-respondents. Of the $N_{R\cdot}$ respondents, N_{FR} are classified as 'Favorable' and $N_{UR} = N - N_{FR}$ are classified as 'Unfavorable.' In the absence of information bearing directly on the fraction of non-respondents who would be classified as 'Favorable,' statistics available for inference are N_{FR}, N_{UR}, $N_{S\cdot}$ subject to $N_{FS} + N_{US} = N_{S\cdot}$ and to $N_{FR} + N_{UR} + N_{S\cdot} = N$. Suppose that the process generating data about both respondents and non-respondents is modeled as a double dichotomy with probabilities p_{FR}, p_{UR}, p_{FS} and $p_{US} > 0$ and $p_{FR} + p_{UR} + p_{FS} + p_{US} = 1$ as shown in Table 1. Table 2 shows the corresponding tables for the model of observed data.

The probability distribution for observations N_{FR}, N_{UR} and $N_{S\cdot}$ in Table 1 is proportional to

$$(p_{FS} + p_{US})^{N_{S\cdot}} p_{FR}^{N_{FR}} p_{UR}^{N_{UR}}$$

Observations N_{FR}, N_{UR} and $N_{S\cdot}$ are sufficient for joint inference about $p_{S\cdot} = p_{FS} + p_{US}$, p_{FS} and p_{UR} but even an infinite sample will not lead to knowledge of p_{FS} and p_{US} with certainty: p_{FS} and p_{US} are not identifiable in the presence of non-response as an infinite sample yields knowledge only that p_{FS} and p_{US} satisfy $p_{FS} + p_{US} = p_{S\cdot}$

3. Definitions

A statistical 'model' is a family $\{P_\theta, \theta \in \Theta\}$ of probability distributions indexed by a parameter $\theta \in \Theta$. A statistical 'structure' is P_θ for a particular $\theta = \theta_0 \in \Theta$. Parameters θ and $\theta' \in \Theta$ are 'observationally equivalent' if and only if $P_{\theta'}(A) = P_{\theta'}(A)$ for all A in a measurable space $\mathscr{S}$.

Definition: Θ is 'identifiable' if and only if $\theta = \theta'$ for all observationally equivalent pairs $\theta, \theta' \in \Theta$.

In practice, it is useful to couch these definitions in terms of distribution functions: let X be an observable random variable with range $\mathscr{S}_X$ and distribution function $F_\theta(x) \equiv P(X \leqslant x \mid \theta)$ belonging to a model $\{F_\theta, \theta \in \Theta\}$. If there exists at least one pair (θ, θ') for which $\theta, \theta' \in \Theta$ and $\theta \neq \theta'$ such that $F_\theta(x) = F_{\theta'}(x)$ for

Table 1

	Parameters			Observations		
Favorable (F)	p_{FR}	p_{FS}	$p_{\cdot F}$	N_{FR}	N_{FS}	$N_{\cdot F}$
Unfavorable (U)	p_{UR}	p_{US}	$p_{\cdot U}$	N_{UR}	N_{US}	$N_{\cdot U}$
	$p_{R\cdot}$	$p_{S\cdot}$	1.0	$N_{R\cdot}$	$N_{S\cdot}$	N

Table 2

	Parameters		Observations		
Favorable (F)	p_{FR}		N_{FR}		
Unfavorable (U)	p_{UR}		N_{UR}		
	$p_{R\cdot}$	$p_{S\cdot}$	$N_{R\cdot}$	$N_{S\cdot}$	N

all $x \in \mathscr{S}_x$ then we say that $\boldsymbol{\Theta}$ is 'non-identifiable' by X. If $\boldsymbol{\Theta}$ is not non-identifiable by X it is identifiable. These definitions of identifiability can be recast to embrace nonparametric models as well (Basu 1981, Berman 1963).

4. *Example 2*

In general, unobservables in an 'econometric simultaneous equation model' are restricted to residuals (errors in model equations). A 'factor analysis model,' on the other hand, possesses both unobservable latent and unobservable explanatory variables. Each is a special case of a more general 'structural equation model' in which a 'latent variable model' displaying interactions among latent variables and a 'measurement model' specifying how observable variables are related to latent variables are conjoined. Properties of the covariance matrix of observable variables are at center stage and structural equation test procedures are designed to measure the quality of fit between an observed (sample) covariance matrix $\boldsymbol{\Sigma}$ and a model based covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ indexed by a vector $\boldsymbol{\theta}$ of model parameters.

Even when the structure of relations between observable variables and latent variables and the structure of relations among latent variables are both linear, the model covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ may be a complicated, nonlinear function of elements of the parameter vector $\boldsymbol{\theta}$. In this setting, in order for an element of the parameter vector $\boldsymbol{\theta}$ to be identified it must be an explicit, known function of one or more elements of the sample covariance matrix $\boldsymbol{\Sigma}$ under the hypothesis that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. (See Bollen (1989) for a comprehensive treatment. Also Hoyle (1995) for applications.) If PDS $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is $(k \times k)$ it has at most $\frac{1}{2}k(k+1)$ functionally independent components so for $\boldsymbol{\theta}$ to be identifiable, it can possess at most $\frac{1}{2}k(k+1)$ functionally independent components. This is a generic, easy-to-apply, necessary but not sufficient condition for identifiability of $\boldsymbol{\theta}$.

In the notation of Bollen 1989 we have a

Latent Variable Model:

$$\boldsymbol{\eta} = \mathbf{B}^*\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

The $(m \times m)$ coefficient matrix $\mathbf{B}^*$ induces interactions among latent endogenous variables $(m \times 1)$ $\boldsymbol{\eta}$, the $(m \times r)$ coefficient matrix $\boldsymbol{\Gamma}$ establishes the effect of exogenous variables $(r \times 1)$ $\boldsymbol{\xi}$ on $\boldsymbol{\eta}$ and $(m \times 1)$ $\boldsymbol{\zeta}$ represents a random residual error term assumed to be uncorrelated with $\boldsymbol{\xi}$ and to possess mean zero and finite (PDS) variance matrix $\boldsymbol{\Psi}$. In addition $\mathbf{B} \equiv \mathbf{I} - \mathbf{B}^*$ is assumed to be nonsingular.

Measurement Model:

$$\mathbf{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$
$$\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}$$

The elements of the $(p \times m)$ coefficient matrix $\boldsymbol{\Lambda}_y$ determine the magnitude of the effect of levels and changes in levels of latent variables $\boldsymbol{\eta}$ on observable variables $\mathbf{y}$; $(q \times r)$ $\boldsymbol{\Lambda}_x$ relates exogenous variables $\boldsymbol{\xi}$ to observable variables in $\mathbf{x}$ in a similar fashion. Residual errors $(p \times 1)$ $\boldsymbol{\varepsilon}$ and $(q \times 1)$ $\boldsymbol{\delta}$ are random variables, generally assumed to be uncorrelated with one another, with $\boldsymbol{\eta}$ and with $\boldsymbol{\xi}$ and to possess common mean zero, but possibly different (PDS) covariance matrices Θ_ε and Θ_δ respectively.

To be specific, focus on the composition of the variance matrix Var($\mathbf{y}$) of $\mathbf{y}$ as a function of latent variable model parameters $\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ and measurement model parameters $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Theta}_\varepsilon$:

$$\mathrm{Var}(\mathbf{y}) = \boldsymbol{\Lambda}_y\mathbf{B}^{-1}[\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^t + \boldsymbol{\Psi}](\mathbf{B}^{-1})^t\boldsymbol{\Lambda}_y^t + \boldsymbol{\Theta}_\varepsilon$$

The $(p \times p)$ PDS matrix Var($\mathbf{y}$) possesses $\frac{1}{2}p(p+1)$ functionally independent components. The variance matrix $\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}_\varepsilon$ alone also possesses $\frac{1}{2}p(p+1)$ functionally independent components and, in the absence of restrictions, other than just $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are PDS, $\boldsymbol{\Lambda}_y$, $\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ possess $\frac{1}{2}m(m+1) + \frac{1}{2}r(r+1) + m(m+p+r)$ additional functionally independent components. As the above equation suggests, it

can be a formidable task to verify by direct algebraic analysis whether or not a particular model parameter is, or is not identifiable.

Clearly, parameter matrices $\mathbf{\Lambda}_y$, $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Phi}$, and $\mathbf{\Theta}_\varepsilon$ cannot be expressed as unique functions of elements of $\mathrm{Var}(\mathbf{y})$ unless restrictions are placed upon them. When one or more components of $\mathbf{B}$ cannot be recovered uniquely as a function of elements of $\mathrm{Var}(\mathbf{y})$, such components are called 'under-identified' or 'non-identified.' A central task then, is to select constraints on latent variable model parameters and measurement model parameters that renders desired parameters 'identifiable.'

If a priori restrictions on parameters of latent variable and measurement model parameters are specified in a way that leads to two or more distinct representations of some components of $\mathbf{B}$ as a function of components of $\mathrm{Var}(\mathbf{y})$, those components of $\mathbf{B}$ are called 'over-identified.' It can happen, of course, that some elements of $\mathbf{B}$ are identified, some are under-identified and others are over-identified. Bartels (1985) provides an easily accessible treatment of these concepts.

A common form of a priori constraint is to 'specify' that some parameter values are equal to zero. For example, if the modeler assumes that $\mathbf{y}$ and $\mathbf{x}$ are measured without error and that $\mathbf{\Lambda}_y = (m \times m)\,\mathbf{I}$, $\mathbf{\Lambda}_x = (r \times r)\,\mathbf{I}$, $\mathrm{Var}(\boldsymbol{\eta}) = \mathrm{Var}(\boldsymbol{\xi}) = \mathbf{\Theta}_\varepsilon = \mathbf{\Theta}_\delta = \mathbf{0}$ and the variance of $\mathbf{y}$ becomes

$$\mathrm{Var}(\mathbf{y}) = \mathbf{B}^{-1}[\mathbf{\Gamma\Phi\Gamma}^t + \mathbf{\Psi}](\mathbf{B}^{-1})^t.$$

If, in turn, as is often the case in econometric applications, exogenous variables $\mathbf{x}$ are assumed to be predetermined and known with certainty, $\mathbf{\Phi} = \mathbf{0}$. With all of these restrictions in force $\boldsymbol{\eta} = \mathbf{y}$, $\boldsymbol{\xi} = \mathbf{x}$ and the latent variable model becomes $\mathbf{B}\,\mathbf{y} = \mathbf{\Gamma z} + \boldsymbol{\delta}$. Then observable $\mathbf{y} = \mathbf{\Pi z} + \mathbf{u}$, with $\mathbf{\Pi} \equiv -\mathbf{B}^{-1}\mathbf{\Gamma}$, $\mathbf{u} = \mathbf{B}^{-1}\boldsymbol{\delta}$ and $\mathrm{Var}(\mathbf{y}) = \mathbf{B}^{-1}\mathbf{\Psi}(\mathbf{B}^{-1})^t \equiv \mathbf{\Omega}$, is called the 'reduced form' of this model.

Provided that predetermined variables $\mathbf{z}$ are uncorrelated with variables $\mathbf{u}$ (weaker assumptions are possible) and that the asymptotic variance matrix of the $\mathbf{z}$'s exists, both $\mathbf{\Pi}$ and $\mathbf{\Omega}$ can be consistently estimated by ordinary least squares. However, the distribution of $\mathbf{y}$ given $\mathbf{z}$ now depends exclusively on parameters $\mathbf{\Pi}$ and $\mathbf{\Omega}$. Pre-multiplication of both sides of $\mathbf{B}\ \mathbf{y} = \mathbf{\Gamma z} + \boldsymbol{\delta}$ by any nonsingular $(m \times m)$ matrix leads to the same reduced form, so all such structures are observationally equivalent and without further restrictions the system is underidentified.

Fisher (1966) provides an elegant and thorough exposition of conditions for identifiability of linear simultaneous equation system model parameters. Among the most important are 'order' and 'rank' conditions established by Koopmans et al. (1950). Suppose interest focuses on a single equation and that exclusion restrictions alone on $\mathbf{y}$ and $\mathbf{z}$ are in force. A (necessary but not sufficient) 'order condition' for parameters of this equation to be identifiable is that the number of variables excluded be greater than or equal to $m-1$—one less than the number of components in (endogenous) $\mathbf{y}$. Alternatively, the total number of excluded predetermined variables must be greater than or equal to the number of variables not excluded from this equation, less one (Fisher 1966, pp. 40–1).

The following argument due to Fisher is a transparent explanation of the origin of the 'rank condition' for linear simultaneous equation systems, first proved by Koopmans et al. (1950). Exclusion of the lth element of the first row $\mathbf{a}_1$ of $m \times (m+r)\mathbf{A} \equiv [\mathbf{B}, \mathbf{\Gamma}]$ is representable as $\mathbf{a}_1\mathbf{c}^{(l)} = 0$, $\mathbf{c}^{(l)}$ an $(m+r) \times 1$ column vector with a 1 at element l and zeros elsewhere. Thus k such exclusions of distinct variables may be written in terms of $(m+r) \times k$ $\mathbf{C}$ as $\mathbf{a}_1\mathbf{C} = \mathbf{0}$. Clearly if $\mathbf{a}_1$ is some linear combination of other rows of $\mathbf{A}$, it is not identifiable. Thus, a sufficient condition for $\mathbf{a}_1$ to be identified is that $\mathbf{xAC} = \mathbf{0}$ holds only for $(1 \times m)$ $\mathbf{x}$ of the form $= (\alpha, 0, \ldots, 0)$, α a scalar. In the absence of other a priori restrictions on elements of $\mathbf{a}_1$, this condition is also necessary. When it obtains, the rank of $\mathbf{AC}$ must be $m-1$. The order condition follows from the observation that the rank of $\mathbf{AC}$ must be less than or equal to that of $\mathbf{C}$, so the number of independent restrictions imposed by $\mathbf{C}$ must be at least $m-1$.

5. *Estimation*

When a linear statistical model is just identified, both consistent and unbiased estimators of its parameters exist. A more general notion of estimability for linear statistical models, first proposed by Bose (1944, 1947), says that a parameter $\boldsymbol{\theta}$ of such a model is (linearly) estimable if there exists a linear function of observations with expectation equal to $\boldsymbol{\theta}$ for each $\boldsymbol{\theta}$ in a prespecified set $\mathbf{\Theta}$. This special case of identifiability, apparently developed independently of the more general theory of identification outlined in this review, has generated a large literature of its own (see *Factor Analysis and Latent Structure: Overview*).

Deistler and Seifert (1978) address the interplay between identification and estimability for a quite general (possibly nonlinear) version of an econometric structural equation model without measurement error. They answer the question 'When, for such systems, does identifiability imply existence of consistent estimators?'

Jöreskog (1973) and Wiley (1973) were among the first to specify practical procedures for estimation of parameters by analysis of the covariance structure of observations generated by a general structural equation model of the form described in Example 2. For a detailed description of parameter estimation via maximum likelihood, instrumental variables and least squares see, for example, Jöreskog and Sörbom (1984).

Simultaneous Equation Estimates (Exact and Approximate), Distribution of is a crisp summary of methods designed to recover consistent estimates of structural equation parameters, such as Full Information Maximum Likelihood (FIML), Limited Information Maximum Likelihood (LIML), R-class estimators, two- and three-stage least squares (2 SLS, 3 SLS). A more leisurely presentation of econometric simultaneous equation system identification issues, couched in terms of a two-equation supply and demand model, and the nature of estimation problems they pose, appears in *Simultaneous Equation Estimation: Overview*.

6. *Bayesian Identification*

Non-Bayesians typically assign exact numerical values to a subset of parameters of a statistical model in order to achieve identification of one or more other parameters. Fisher (1961) suggested broadening this special type of a priori specification by replacing exact numerical values with probabilistic constraints. Drèze (1962) took up the challenge in his treatment of Bayesian identification for simultaneous equation estimation models. As before, let $\mathbf{By}+\mathbf{\Gamma z}=\mathbf{u}$ be a 'structural equation system' indexed by matrix parameters consisting of nonsingular $\mathbf{B}$, $\mathbf{\Gamma}$, and PDS $\mathrm{Var}(\mathbf{u})=\mathbf{\Sigma}$ with reduced form $\mathbf{y}=\mathbf{\Pi z}+\mathbf{v}$, $\mathbf{\Pi}=-\mathbf{B}^{-1}\mathbf{\Gamma}$ and disturbance variance $\mathrm{Var}(\mathbf{v})=\mathbf{B}^{-1}\mathbf{\Sigma}(\mathbf{B}^{-1})^{t}\equiv\mathbf{\Omega}$. Because the likelihood function for observations $\mathbf{y}$ can be written as a function of reduced form parameters $\mathbf{\Pi}$ and $\mathbf{\Omega}$ alone, full knowledge of the exact numerical values of elements of $\mathbf{\Pi}$ and $\mathbf{\Omega}$ does not permit recovery of unique values for elements of $\mathbf{B}$, $\mathbf{\Gamma}$, and $\mathbf{\Sigma}$. The map from $\mathbf{\Pi}$ and $\mathbf{\Omega}$ to $\mathbf{B}$, $\mathbf{\Gamma}$, and $\mathbf{\Sigma}$ is one to many. In effect, all nonsingular matrices $\mathbf{B}$ are observationally equivalent and $\mathbf{B}$ is only identifiable in the special case when $f(\mathbf{B}\,|\,\mathbf{\Pi},\mathbf{\Omega})$ is concentrated on a single point in the space of nonsingular matrices $\mathbf{B}$. Nevertheless, one can assign a prior density $g(\mathbf{B},\mathbf{\Gamma},\Sigma)$ to structural parameters, observe data $(\mathbf{Y},\mathbf{Z})$ and compute a posterior density $f(\mathbf{B},\mathbf{\Gamma},\mathbf{\Sigma}\,|\,\mathbf{Y},\mathbf{Z})$ for $\mathbf{B}$, $\mathbf{\Gamma}$, and $\mathbf{\Sigma}$ whether or not the model is identified in a classical sense. Because identifiability is a property of the likelihood function alone, if a proper prior distribution is assigned to a complete set of parameters ($\boldsymbol{B}$, $\mathbf{\Pi}$, and $\mathbf{\Omega}$ here) the posterior for these parameters is proper whether or not the likelihood function is identified. Such priors must be chosen judiciously: too tight a prior will anchor the posterior on the prior until sample size becomes very large; a very loose proper prior may fail to reflect relevant prior knowledge. In order to streamline computation of posteriors, Raiffa and Schlaifer (1961) introduced the concept of 'natural conjugate' priors. The functional form of a natural conjugate prior is derived by interpreting parameter as sufficient statistic and sufficient statistic as parameter in the likelihood function. A consequence is that both prior and posterior possess the same functional form, a form dictated by the shape of the likelihood function. For simultaneous equation systems, multivariate-, matric- and poly-t posterior distributions for parameters follow and figure prominently in work by Drèze (1976), Drèze and Morales (1976), Drèze and Richard (1983), and Mouchard (1976). (See *Bayesian Statistics*.)

To minimize the impact of a priori evidence or, in some cases, to exploit invariance with respect to mapping from one representation of a parameter to another, a diffuse (vague, noninformative) prior can be the prior of choice. However, in some settings, this choice comes at a cost as a diffuse prior may lead to an improper posterior. Lindley and El Sayeed (1968) foreshadow much recent work on the character of diffuse priors. They were perhaps the first to recognize several important features of prior to posterior analysis. First, a proper prior always yields a proper posterior. Second, '…in the functional relationship problem [improper priors] can cause considerable difficulties….' Third, even when the likelihood function is not identifiable, observed data can lead to a posterior that is informative; i.e., different from the prior. Fourth, 'Uncritical use of improper distributions can lead to inconsistent estimates.'

A simple example often used to introduce the idea of nonidentifiability highlights issues raised by diffuse priors.

7. *Example 3*

Suppose observations to be independent, identically normal with mean $\theta=\mu_1+\mu_2$ and variance 1. Then μ_1 and μ_2 are non-identifiable. Nevertheless, assignment of a proper prior to μ_1 and μ_2 leads to a proper posterior. As the number of observations approaches infinity, the posterior for θ converges in distribution to a true value θ_0 of θ and the (proper) posterior for μ_1 and μ_2 concentrates on the line $\theta_0=\mu_1+\mu_2$.

If, however, a diffuse prior $d\mu_1$, $-\infty<\mu_1<\infty$, is assigned to μ_1 and proper prior $g(\theta)$ to θ, the posterior for μ_1 remains diffuse irrespective of sample size. Replace the proper prior $g(\theta)$ for θ with a proper prior $h(\mu_2)$ for μ_2. Then, provided the prior for μ_1 is diffuse, the posterior for μ_2 remains the same as the prior $h(\mu_2)$, no matter how many observations are made.

These features of Example 3 suggest that uncritical choice of a diffuse prior can lead to a posterior with unsatisfactory or even unacceptable properties, in particular when the parameter space is of large dimension. For example, Kleibergen and Van Dijk (1997) point out that for simultaneous equation systems '… the order condition reflects overall identification while the rank condition reflects local (non) identification.' They show that local non-identification coupled with a diffuse prior leads to posterior pathology when one attempts Limited Information analy-

sis of a single equation and suggest an alternative approach. (See Gelfand and Sahu (1999) for other examples.)

Dawid (1979) and Kadane (1974) use conditional independence as a device for a definition of non-identifiability which embraces both Bayesian and non-Bayesian perspectives. Suppose that an observable X possesses a distribution F_θ determined by a parameter $\boldsymbol{\theta}$ which we partition into two pieces $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. If the distribution of $\boldsymbol{X}$ is fully determined by $\boldsymbol{\theta}_1$ alone, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is non-identifiable.

When a Bayesian assigns a prior distribution $g(\boldsymbol{\theta})$ to $\boldsymbol{\theta}$, Bayes theorem tells her that the posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ for a parameter $\boldsymbol{\theta}$ given data $\mathbf{y}$ is proportional to the product of the likelihood function $\mathbf{L}(\boldsymbol{\theta}|\mathbf{y})$ for $\boldsymbol{\theta}$ given $\mathbf{y}$ and the prior $g(\boldsymbol{\theta})$ assigned to $\boldsymbol{\theta}$. The (posterior) distribution for $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$ and data $\mathbf{y}$ given prior $g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)h(\boldsymbol{\theta}_1)$ is

$$f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y}) \propto \mathbf{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)h(\boldsymbol{\theta}_1)$$

If $\boldsymbol{\theta}_2$ is absent from the likelihood function,

$$f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y}) \propto g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)h(\boldsymbol{\theta}_1|\mathbf{y}) \propto g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$$

so that while data will modify her prior judgements about $\boldsymbol{\theta}_1$, no amount of observable data $\mathbf{y}$ will change her a priori probability assessment $g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ for $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$. Data $\mathbf{y}$ is, however, unconditionally informative about $\boldsymbol{\theta}_2$; as data accrues and we learn more about $\boldsymbol{\theta}_1$, prior judgements about $\boldsymbol{\theta}_2$ are modified in accord with

$$f(\boldsymbol{\theta}_2|\mathbf{y}) \propto \int g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)h(\boldsymbol{\theta}_1|\mathbf{y})\,\mathrm{d}\boldsymbol{\theta}_1$$

provided that both $g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ and $h(\boldsymbol{\theta}_1|\mathbf{y})$ are proper distributions.

Return to Example 1. As Bayesians, (noting that $\mathrm{p}_{FR}+\mathrm{p}_{UR}+\mathrm{p}_{FS}+\mathrm{p}_{US} = 1$), we assign a prior distribution $g(\mathrm{p}_{FR}, \mathrm{p}_{UR}, \mathrm{p}_{FS})$ to p_{FR}, p_{UR}, p_{FS}. If observed data $\mathrm{y} = (\mathrm{N}_{S\cdot}, \mathrm{N}_{FR}, \mathrm{N}_{UR})$ is generated in accord with Table2, the likelihood function for p_{FR}, p_{UR}, p_{FS},

$$[1-\mathrm{p}_{FR}-\mathrm{p}_{UR}]^{N_{S\cdot}}\,\mathrm{p}_{FR}^{\mathrm{N}_{FR}}\mathrm{p}_{UR}^{\mathrm{N}_{UR}}$$

does not depend on p_{FS}. Consequently,

$$f(\mathrm{p}_{FS}|\mathrm{p}_{FR}, \mathrm{p}_{UR}, \mathbf{y}) \propto g(\mathrm{p}_{FS}|\mathrm{p}_{FR}, \mathrm{p}_{UR})$$

This matches our intuition: given $\mathrm{p}_{FR}+\mathrm{p}_{UR} = \alpha$ we then know with certainty that $\mathrm{p}_{FS}+\mathrm{p}_{US} = 1-\alpha$, p_{FS}, $\mathrm{p}_{US} \geqslant 0$. Data of type $(\mathrm{N}_{S\cdot}, \mathrm{N}_{FR}, N_{\mathrm{UR}})$ refines our knowledge about the value of α, but even an infinite sample of this type will not lead to knowledge certain of p_{FS}. The conditional distribution of p_{FS} given $\mathrm{p}_{FR}+\mathrm{p}_{UR} = \alpha$ prior to observing $\mathbf{y}$ is the same as the conditional distribution of p_{FS} given $\mathrm{p}_{FR}+\mathrm{p}_{UR} = \alpha$ posterior to observing $\mathbf{y}$.

While the theory of identification in a classical statistical context is now well understood, the story of identification from a Bayesian viewpoint is still unfolding. Markov Chain Monte Carlo [MCMC] methods breathe new life into Bayesian inference, allowing efficient computation of normalizing constants and parameter estimates for complex models. An explosion of Bayesian application of MCMC methods surfaces new identification issues. In their study of Bayesian identifiability for 'Generalized Linear Models' (GLMs), Gelfand and Sahu (1999) present conditions under which assignment of an improper prior distribution to some components of a parameter set leads to a proper posterior distribution. More particularly, if a posterior distribution for parameters is improper, under what conditions does there exist a lower-dimensional parameter set that possesses a proper posterior distribution? Ghosh et al. (2000) prove that when data $\mathbf{y}$ is uninformative about $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$ so that $\boldsymbol{\theta}_2$ is not identifiable, the posterior density $f(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1|\mathbf{y})$ is proper if and only if both $h(\boldsymbol{\theta}_1|\mathbf{y})$ and $g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ are proper, a fact that Gelfand and Sahu use to establish that for GLMs at least, even when the posterior distribution for all parameters is improper, there are embedded lower dimensional models that possess a proper posterior distribution, although in general, that distribution is not uniquely determined. (See Gelfand and Sahu (1999) for conditions that guarantee uniqueness.) They warn, however, that model fitting using MCMC methods is '…somewhat of an art form, requiring suitable trickery and turning to obtain results in which we can have confidence.'

Bibliography

Bartels R 1985 Identification in Econometrics. *The American Statistician* **39**(2): 102–4

Basu A P 1981 *Statistical Distributions in Scientific Work* **5**. In: Taillie C, Patil, Baldessari (eds.) D Reidel, Dordrecht, The Netherlands, pp. 335–48

Berman S M 1963 *Annals of Mathematical Statistics* **34**: 1104–6

Bollen K A 1989 *Structural Equations with Latent Variables*. Wiley, New York

Bose R C 1944 Proceedings of the 31st Indian Sci. Congr., Delhi, Vol. 3, pp. 5–6

Bose R C 1947 Proceedings of the 34th Indian Sci. Congr., Delhi, Part II, pp. 1–25

Bowden R 1973 The theory of parametric identification. *Econometrica* **41**(6): 1069–74

Dawid A P 1979 Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society* Ser. B **41**: 1–31

Deistler M, Seifert H-G 1978 Identifiability and consistent estimability in econometric models. *Econometrica* **46**(4): 969–80

Dreze J H 1962 The Bayesian approach to simultaneous equation estimation. *ONR Research Memorandum No. 67*. Northwestern University

Drèze J H 1972 Econometrics and Decision Theory. *Econometrica* **40**: 1–17

Drèze J H 1974 Bayesian theory of identification in simultaneous equations models. *Studies in Bayesian Econometrics and Statistics*. North-Holland, Amsterdam, pp. 159–174 (presented to Third NBER-NSF Seminar on Bayesian Inference Econometrics, Harvard Univ., October 1971)
Drèze J H 1976 Bayesian limited information analysis of the simultaneous equations model. *Econometrica* **44**: 1045–75
Drèze J H, Morales J A 1976 Bayesian full information analysis of simultaneous equations. *Journal of the American Statistical Association* **71**(376): 919–23
Drèze J H, Richard J F 1983 Bayesian analysis of simultaneous equation systems. *Handbook of Econometrics* **1**
Fisher F M 1961 On the cost of approximate specification in simultaneous equation estimation. *Econometrica* **29**(2)
Fisher F M 1966 *The Identification Problem in Econometrics*. McGraw-Hill, New York
Gelfand A, Sahu S K 1999 Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association* **94**(445): 247–53
Ghosh M, Chen M H, Ghosh M, Agresti A 2000 Non-informative priors for one parameter item response models. *Journal of Statistical Planning and Inference* **88**: 99–115
Hannan E J 1969 The identification of vector mixed auto-regressive moving average systems. *Biometrika* **57**: 223–5
Hannan E J 1971 The identification problem for multiple equation systems with moving average errors. *Econometrica* **39**: 751–65
Haavelmo T 1944 The probability approach in econometrics. *Econometrica* **12**, Supplement
Hoyle 1995 *Structural Equation Modeling: Concepts, Issues and Applications*. Sage, Thousand Oaks, CA
Hurwicz L 1950 Generalization of the concept of identification. *Statistical Inference in Dynamic Economic Models. Cowles Commission Monograph 10*. Wiley, New York
Johnson W O, Gastwirth J L 1991 Bayesian inference for medical screening tests: Approximations useful for the analysis of AIDS. *Journal of the Royal Statistical Society Ser. B* **53**: 427–39
Jöreskog K G 1973 In: Krishnaiah P R (ed.) *Analysis of Covariance Structures. Multivariate Analysis III*. Academic Press, New York
Jöreskog K G, Sörbom D 1984 LISREL VI *Analysis of Linear Structural Relations by Maximum Likelihood, Instrumental Variables, and Least Square Methods. User's Guide*. Department of Statistics, University of Uppsala, Uppsala, Sweden
Kadane J B 1974 The role of identification in Bayesian theory. *Studies in Bayesian Econometrics and Statistics*. North-Holland, Amsterdam, pp. 175–91 (presented to Winter Meeting of Econometric Society, Toronto, December 1972).
Kleibergen F, Van Dijk H K 1997 *Bayesian Simultaneous Equations Analysis using Reduced Rank Structures*. Erasmus University Technical Report 9714/A
Koopmans T C, Reiersøl O 1950 The identification of structural characteristics. *Annals of Mathematical Statistics* **21**: 165–81
Koopmans T C, Rubin H, Leipnik R B 1950 Measuring the equation systems of dynamic economics. *Statistical Inference in Dynamic Economic Models. Cowles Commission Monograph 10*. Wiley, New York
Lazarsfeld P F 1950 The logical and mathematical foundation of latent structure analysis. The interpretation of some latent structures. *Measurement and Prediction* **4**: Studies in Psychology of World War II
Leamer E E 1978 *Specification Searches*. Wiley, New York
Mouchard J-F 1976 *Posterior and Predictive Densities for Simultaneous Equation Models*. Springer-Verlag, New York
Neath A A, Samaniego F J August 1997 On the efficacy of Bayesian inference for nonidentifiable models. *The American Statistician* **51**(3): 225–32
Pigou C 1910 A method of determining the numerical values of elasticities of demand. *Economic Journal* **20**: 636–40
Poirier D J 1996 *Revising Beliefs in Non-Identified Models*. Technical Report, University of Toronto, Canada
Raiffa H, Schlaifer R O 1961 *Applied Statistical Decision Theory*. Harvard Business School, Boston
Rothenberg T J May 1971 Identification in parametric models. *Econometrica* **39**: 577–91
Thurstone L L 1935 *The Vectors of Mind*. University of Chicago Press, Chicago
Thurstone L L 1947 *Multiple-Factor Analysis*. University of Chicago Press, Chicago
Van der Genugten B B 1997 *Statistica Neerlandica* **31**: 69–89
Wald A 1950 Note on the identification of economic relations. *Statistical Inference in Dynamic Economic Models. Cowles Commission Monograph 10*. Wiley, New York
Wiley D E 1973 The identification problem for structural equation models with unmeasured variables. In: Goldberger A S, Duncan O D (eds.) *Structural Equation Models in the Social Sciences*. Seminar Press, New York

G. M. Kaufman

Statistical Methods, History of: Post-1900

1. Introduction

In *Statistical Methods for Research Workers*, Fisher (1925) characterizes statistics as '(i) the study of populations, (ii) the study of variation, (iii) the study of methods of the reduction of data.' This is still the field of statistics in a nutshell. The three forms of studies are interwoven, and one thread cannot fully be understood without the other in the fabric of statistics that they weave.

During the early decades of the twentieth century, the theory of probability and of statistical methodology of the least-squares tradition from Laplace and Gauss continued to be developed and the two fields married permanently. Methods of regression, correlation, and analysis of variance found their final form. As did survey sampling, which was given a probabilistic footing. The discussion of Bayesian analysis and inverse probability continued and led to a split between the frequentist and the Bayesian schools of thought.

This period also saw the emergence of fundamentally new ideas and methodologies. Statistical inference, e.g., in the forms of hypothesis testing and confidence interval estimation, was identified as distinct from data description. Parametric models, allow-

ing exact inference, even for small samples, were found, and led to constructive concepts of inductive logic and optimality for estimation and testing.

This impressive development in the concepts and methods peaked with Fisher's 1922 paper. It was a paradigmatic revolution of statistics, along those of Laplace and Gauss around 1800 (Hald 1998). It took about 20 years and many papers for each of these authors to work out their basic ideas in detail and it took half a century for the rest of the statistical community to understand and develop the new methods and their applications.

The impact of statistics on society and the social sciences also changed and expanded fundamentally. Statistical thinking spread and was integrated in several of the social sciences as part of their basic thinking. The development of psychology, sociology and economics was influenced in essential ways by statistics and statistics itself benefited from ideas and methods developed in these fields, as well as in biology and other sciences.

2. *Impacts of Statistics on the Social Sciences*

The term 'statistics' has its origin in the social sciences. Statistics meant the assemblage of numerical data describing the state, its population, institutions etc., with focus on the particulars. With Adolphe Quetelet this changed. Statistical mass phenomena attracted more and more interest from social scientists throughout the nineteenth century. It was asked whether distributions in society followed laws of chance, such as the binomial, the Poisson, and the normal distribution. Most phenomena were found to show 'supernormal' variation, and the question arose whether the population could be stratified to obtain 'normal' variation within strata.

The emphasis in statistical work changed around 1900 from the general study of variability to more specific issues. Statistical data were requested and brought forward into discussions of social policy and social science. In 1910, Karl Pearson and Ethel M. Elderton, for example, gathered data to see whether children of heavy drinkers were more disabled than other children. Their conclusion (Elderton and Pearson 1910) was in the negative:

> the alcoholic propensities of parents seemed utterly unrelated to any measurable characteristics of the health or intelligence of offsprings, and what relations were found were as likely to be the reverse of what the temperance movement expected as not.

Stigler (1999) summarizes this paper and the ensuing discussion with the leading economist Alfred Marshall and others. Pearson (letter to the *Times*, 12 July 1910) scolded the economists for only arguing theoretically and not bringing forward empirical support:

> to refute these data the only method is to test further equally unselected material … Professor Marshall suggested that the weaklings among the sober had grandparents who were profligate. This may be so but, as a statistician, I reply: 'Statistics—and unselected statistics—on the table, please.' I am too familiar with the manner in which actual data is met with the suggestion that other data, if collected, might show another conclusion to believe it to have any value as an argument. 'Statistics on the table, please,' can be my sole reply.

The call for focused unbiased statistics spread, and the growth in statistical data was impressive both in official statistics and in social and other sciences. The toolbox of statistical methods also grew and the concepts and methods were steadily used more widely.

Even though statistics and probability originated in the social sphere with gambling, insurance and social statistics, statistical thinking in the modern sense came late to the core social sciences. It was, in a sense, easier for Gauss to think statistically about errors in his measurement of the deterministic speed of a planet and for the experimental psychologist to randomize their design and to use corresponding statistical models in the analysis, than it was for Pearson to study how children were affected by their parents' drinking and than it was for the economists to study business cycles. The difficulty is, according to Stigler (1999) that statistics has a much more direct role in defining the object of study in the social sciences. This relative unity of statistical and subject matter concepts was a taxing challenge that delayed progress. When statistical thinking took hold, however, the fields themselves changed fundamentally and received new energy.

This delayed but fundamental impact of statistics can be traced in various disciplines. Consider the study of business cycles. Good times follow bad times, but are these cycles of business regular in some way? While the nineteenth century economists mostly were looking for particular events that caused 'crises' in the economy, Karl Marx and others started to look for regular cyclic patterns. Moving averages and other linear filtering techniques were used to remove noise and trend from the series to allow cycles of medium length to be visible. Statisticians worried whether a cycle of wavelength 20 years were due to the data or to the filtering methods. George Udny Yule found in 1926 that 'nonsense correlations' between time series of finite length are prone to show up and stand conventional tests of significance simply because of internal auto-correlation. He had earlier coined the term 'spurious correlation,' but found that nonsense-correlation are flukes of the data that 'cannot be explained by some indirect catena of causation' (Yule 1926). Eugene Slutsky's Russian paper of 1927 made things even worse for descriptive business cycle analysis. By a simulation experiment using numbers drawn in the People's Commissariat of Finance lottery, he demonstrated that linear filters can produce seemingly

regular cycles out of pure noise. It has later been found that the filter used to find the popular 20-year cycles tends to produce 20-year cycles even from random data.

Slutsky's experiment and Yule's analysis highlighted the limitations of model-free descriptive statistical analysis. It gradually became clear that economic data had to be interpreted in light of economic theory, and the field of econometrics emerged. This revolution peaked with Trygve Haavelmo's *Probability Approach in Econometrics* (1944). Mary S Morgan (1990) holds further that the failure of descriptive business cycle analysis was instrumental in the development of dynamic models for whole economies, starting with the work of Ragnar Frisch and Jan Tinbergen.

Other disciplines saw similar developments of delayed but fundamental impact of statistical methodology, but perhaps not to the same extent as economics. The study of intelligence and personality was, for example, transformed and intensified by the advent of statistical factor analysis stemming from Spearman (1904) and Thurstone (1935).

3. *Statistics Transformed*

3.1 *The Study of Populations*

Populations can be real, such as the population of Norway in 1894 with respect to the number of sick and disabled. Theoretical populations, on the other hand, are the hypothetical possible outcomes of an experiment, an observational study, or a process in society or nature.

For finite realized populations, Kiær proposed in 1895 to the International Statistical Institute that his representative method should be considered as an alternative to the method of the day, which was to take a complete census. The idea of survey sampling was extensively debated and was gradually accepted internationally. Results from Kiær's 1894 survey of sickness and disability were, however, questioned in Norway by Jens Hjorth. He had no problem with sampling *per se* but he criticized the particulars of the survey and the questionnaire. From actuarial reasoning he claimed that Kiær had underestimated the number of disabled. At issue was a proposed universal insurance system for disability and Hjorth argued that a snapshot picture of the population obtained from a survey would be insufficient. Estimates of transition probabilities to disability were also needed. In the background of Hjorth's reasoning was, in effect, a theoretical population, e.g., a demographic model, with disability probabilities by age, occupation group and other determinants as primary parameters. Equipped with the statistical competence of the day, Hjorth presented error bounds based on the binomial distribution for the fraction of disabled and he discussed the sample size necessary to obtain a given precision for various values of the true fraction. Kiær apparently did not understand probability. In 1906, he had to concede that the number of disabled was grossly underestimated in his 1894 survey, and he became silent on his representative method for the rest of his life.

This debate over survey sampling continued over the ensuing decades until Jerzy Neyman (1934) described in clear and convincing form the modern frequentist probability model for survey sampling including stratification, cluster sampling, and other schemes. Neyman's model for a survey of a given finite population can be seen as a theoretical population consisting of the collection of all the possible samples that could be drawn, with a probability associated with each.

The distinction between sample quantities and theoretical parameters were not always observed. For example, instead of asking how data could be used to draw inferences about theoretical quantities, Karl Pearson and others proposed to assess the degree of association in two-by-two tables by means of various coefficients, e.g., the coefficients of mean square contingency and tetrachoric correlation, by emphasizing the importance of substantive theory, and the distinction between theoretical and empirical quantities.

R A Fisher cleared up the confusion caused by the wilderness of coefficients. By 1920, he had introduced the notion of 'parameter' for characteristics of the theoretical population and he used the word 'statistics' for estimates, test variables, or other quantities of interest calculated from the observed data. Fisher's interest was not focused on describing data by statistics, but on using them to draw inferences about the parameters of the underlying theoretical population. In the mid-1920s, Fisher pursued these notions of inference and he incorporated his evolving principles of statistical estimation into a new chapter in the second 1928 edition of his pioneering 1925 book, *Statistical Methods for Research Workers*. Here he presented his new measure of the statistical information in the data with respect to a particular parameter and he stressed the importance of using efficient estimators. Fisher's ideas were based on the concept of a parametric statistical model that embodies both the substative theory, e.g., relevant aspects of Mendel's theory of genetics, and a probabilistic description of the data generating process, i.e., the variational part of the model. This probabilistic model led Fisher to formulate the likelihood function, which in turn led to a revolutionary new methodology for statistics.

Fisher also picked up on William S. Gosset's 1908 development of the t-distribution for exact inference on the mean of a normal distribution with unknown variance. From a simulation experiment and guided

by distributional forms developed by Karl Pearson, Gosset guessed the mathematical form of the *t*-distribution. Later, Fisher was able to prove formally the correctness of Gosset's guess and then go on to show how the *t*-distribution also provided an 'exact solution of the sampling errors of the enormously wide class of statistics known as regression coefficients.' (Fisher 1925)

3.2 The Study of Variation

There are two related aspects of the term variation: uncertainty in knowledge, and sampling variability in data, estimates or test statistics. In the tradition of inverse probability stemming from Thomas Bayes, no distinction was drawn between uncertainty and variation. This inverse-sampling probability tradition dominated well into the twentieth century, although unease with some of the implicit features was expressed repeatedly by some critics as early as the mid-nineteenth century.

Inverse probability is calculated via Bayes' theorem, which turns a prior distribution of a parameter coupled with a conditional distribution of the data given the parameter into a posterior distribution of the parameter. The prior distribution represents the *ex ante* belief/uncertainty in the true state of affairs and the posterior distribution represents the *ex post* belief. The prior distribution is updated in the light of new data.

Before Fisher, statistical methods were based primarily on inverse probability with flat prior distributions to avoid subjectivity, if they were grounded at all in statistical theory. Fisher was opposed to the Bayesian approach to statistics. His competing methodology avoided prior distributions and allowed inference to be based on observed quantities only. When discussing Neyman's paper on survey sampling Fisher (1934), stated:

> All realized that problems of mathematical logic underlay all inference from observational material. They were widely conscious, too, that more than 150 years of disputation between the *pros* and the *cons* of inverse probability had left the subject only more befogged by doubt and frustration.

The likelihood function was the centerpiece of Fisher's methodology. It is the probability density evaluated at the observed data, but regarded as a function of the unknown parameter. It is proportional to the posterior density when the prior distribution is flat. The maximum-likelihood estimate is the value of the parameter maximizing the likelihood. Fisher objected, however, to the interpretation of the normed likelihood function as a probability density. Instead, he regarded the likelihood function and hence the maximum likelihood estimator as a random quantity, since it is a function of the data which are themselves considered random until observed.

To replace the posterior distribution, he introduced the fiducial distribution as a means of summing up the information contained in the data on the parameter of interest. In nice models, the fiducial distribution spans all possible confidence intervals by its stochastic quantiles. A confidence-interval estimator for a parameter is a stochastic interval obtained from the data, which under repeated sampling bounds the true value of the parameter with prescribed probability, whatever this true value is. Neyman was the great proponent of this precise concept. He even claimed to have introduced the term 'degree of confidence,' overlooking Laplace, who had introduced the large sample error bounds and who spoke of '*le degré de confiance*' in 1816 (Hald 1998). Fisher's fiducial argument helped to clarify the interpretation of error bounds as confidence intervals, but Fisher and Neyman could not agree.

Neyman thought initially that confidence intervals, which he introduced in his (1934) paper on sampling, were identical to Fisher's fiducial intervals and that the difference lay only in the terminology. He asked whether Fisher had committed a '*lapsus lingua*' since Fisher spoke of fiducial probability instead of degree of confidence (Neyman 1941). Neyman's point was that no frequentist interpretation could be given for the fiducial probability distribution since it is based on the observed data *ex post* and is a distribution over the range of the parameter. He shared with Fisher the notion of the parameter having a fixed but unknown value and to characterize it by a probability distribution made no sense to him. Fisher stressed, however, that his fiducial distribution is based on the logic of inductive reasoning and provides a precise representation of what is learned from the observed data. He could not see that the frequentist property of the confidence interval method allowed one to have 'confidence' in the actually computed interval. That intervals computed by the method in hypothetical repetitions of the experiment would cover the truth with prescribed frequency was of some interest. It was, however, the logic and not this fact that made Fisher say that the computed interval contained the parameter with given fiducial probability.

Harrold Hotelling (1931) staked out an intermediate position. He generalized Gosset's *t*-distribution to obtain confidence regions for the mean vector of multivariate normal data with unknown variance-covariance matrix. He concluded

> confidence corresponding to the adopted probability P may then be placed in the proposition that the set of true values [the mean vector] is represented by a point within this boundary [the confidence ellipsoid].

Although fierce opponents on matters of inference, Fisher and Neyman agreed that inverse probability had left the subject only more 'befogged by doubt and frustration' (Fisher 1934) and they united in their fight against Bayesianism. The tradition from Bayes survived, however. One solution to the flawed indifference

principle is to go subjective. By interpreting the prior distribution as a carrier of the subjective belief of the investigator before he sees the data, the posterior distribution is interpreted correspondingly, as representing the subjective believes *ex post* data. This was the approach of Bruno de Finetti. Harold Jeffreys proposed another solution. He argued that the prior should be based on the statistical model at hand. For location parameters, the flat prior is appropriate since it is invariant with respect to translation. Scale parameters, like the standard deviation, should have flat priors on the logarithmically transformed parameter. These and other priors were canonized from invariance considerations.

The basic idea of statistical hypotheses testing had been around for some 200 years when Karl Pearson developed his chi-square test in 1900 (Hald 1998). The chi-square test and Gosset's t-test initiated a renewed interest in the testing of statistical hypotheses. A statistical test is a probabilistic twist of the mathematical proof by contradiction: if the observed event is sufficiently improbable under the hypothesis, the hypothesis can reasonably be rejected. Fisher formulated this logic in his p-value for tests of significance.

A given statistical hypothesis can usually be tested by many different test statistics calculated from the data. Fisher argued that the likelihood ratio was a logical choice. The likelihood ratio is the ratio of the likelihood at the unrestricted maximum to the likelihood maximized under the hypothesis. Neyman and Egon S. Pearson, son of Karl Pearson, were not satisfied with this reasoning. They looked for a stronger foundation on which to chose the test. From Gosset, they took the idea to formulate the alternative to the hypothesis under test. To them, testing could result either in no conclusion (there is not sufficient support in the data to reject the null hypothesis), or in a claim that the alternative hypothesis is true rather than the null hypothesis.

The alternative hypothesis, supposed to be specified *ex ante*, reflects the actual research interest. This led Neyman and Pearson to ask which test statistic would maximize the chance of rejecting the null hypothesis if indeed this hypothesis and not the null hypothesis is true, when the probability of falsely rejecting the null hypothesis is controlled by a pre-specified level of significance. In nice problems, they found the likelihood ratio statistic to produce the optimal test in 1933. This result is the jewel of the Neyman–Pearson version of frequentist inference. As important as their optimality result is, perhaps, their emphasis on the alternative hypothesis, the power of the test, and the two types of error (false positive and false negative conclusion) that might occur when testing hypotheses.

Gauss asked in 1809 whether there exists a probability distribution describing the statistical variability in the data that makes the arithmetic mean the estimator of choice for the location parameter. He showed that the normal distribution has this property and that his method of least squares for linear parameters is indeed optimal when data are normally distributed. This line of research was continued. The distribution that makes the geometric mean the estimator of choice was found and John M. Keynes discussed estimating equations and showed that the double exponential distribution makes the median an optimal estimator. In 1934 Fisher followed up with the more general exponential class of distributions. When the statistical model is within this class, his logic of inductive inference gives unambiguous guidance.

Probability remained as the underpinning of formal statistical methods and the probability theory saw many advances in the early twentieth century, particularly in continental Europe. The Central Limit Theorem, which is at the heart of Laplace's approximate error bounds, was strengthened. Harald Cramér (1946) summed up this development and he exposed its statistical utility since many estimators and other statistics are approximately linear in the observations and thus are approximately normally distributed. Related to the Central Limit Theorem is the use of series expansions to find an approximation to the distribution of a statistic. Thorvald N. Thiele completed his theory of distributional approximation by way of half invariants, which Fisher later rediscovered under the name of cumulants. Francis Y Edgeworth proposed another series expansion in 1905 (see Hald 1998). Probability theory is purely deductive and should have an axiomatic basis. Andrei N Kolmogoroff gave probability theory a lasting axiomatic basis in 1933, see Stigler (1999) who calls 1933 a magic year for statistics.

3.3 The Study of Methods of the Reduction of Data

Raw statistical data need to be summarized. Fisher (1925) says

> no human mind is capable of grasping in its entirety the meaning of any considerable quantity of numerical data. We want to be able to express all *relevant* information contained in the mass by means of comparatively few numerical values … It is the object of the statistical process employed in the reduction of data to exclude the irrelevant information, and to isolate the whole of the relevant information contained in the data.

It is theory that dictates whether a piece of data is relevant or not. Data must be understood in the light of subject matter theory. This was gradually understood in economics, as well as in other fields. Fisher (1925) showed by examples from genetics how theory could be expressed in the statistical model through its structural or parametric part. The likelihood function is, for Fisher, the bridge between theory and data. Fisher introduced the concept of sufficiency in 1920: a statistic is sufficient if the conditional distribution of the data given the value of the statistic is the same for

all distributions within the model. Since statistics having a distribution independent of the value of the parameter of the model is void of information on that parameter, Fisher was able to separate the data into its sufficient (informative) statistics, and ancillary (non-informative) statistics.

The likelihood function is sufficient and thus carries all the information contained in the data. In nice cases, notably when the model is of the exponential class, the likelihood function is represented by a sufficient statistic of the same dimensionality as the parameter. It might then be possible to identify which individual statistic is informative for which individual parameter. If, say, the degree of dependency expressed by the theoretical odds ratio in a two by two contingency table is what is of theoretical interest, the marginal counts are ancillary for the parameter of interest. Fisher suggested therefore conditional inference given the data in the margins of the table. In the testing situation, this he argued leads to his exact test for independence. Conditional inference is also often optimal in the Neyman–Pearson sense.

4. Conclusion

Statistical theory and methods developed immensely in the first third of the twentieth century with Fisher as the driving force. In 1938, Neyman moved from London to Berkeley. The two men were continents apart and kept attacking each other. The question is, however, whether this was more grounded in personal animosity than real scientific disagreement (Lehmann 1993). In the US, where Neyman's influence was strong, there has been a late-twentieth-century resurgence in Fisherian statistics (Efron 1998) and even more so in Bayesian methods. Fisher and Neyman both emphasized the role of the statistical model. Both worked with the likelihood function, and suggested basically the same reduction of the data. Fisher gave more weight to the logic of inductive reasoning in the particular case of application, while Neyman stressed the frequentistic properties of the method. To the applied research worker, this boils down to whether the level of significance or the confidence coefficient should be chosen *ex ante* (Neyman), or whether data should be summarized in fiducial distributions or *p*-values. These latter *ex post* benchmarks are then left to the reader of the research report to judge.

Over the period, probabilistic ideas and statistical methods were slowly dispersed and were applied in new fields. Statistical methods gained foothold in a number of sciences, often through specialized societies like Econometric Society and through journals such as *Econometrica* from 1933 and *Psychometrica* from 1936.

Statistical inference was established as a new field of science in our period, but it did not penetrate empirical sciences before World War II to the extent that research workers in the various fields were required to know its elements (Hotelling 1940). Statistical ideas caught the minds of eminent people in the various sciences, however, and new standards were set for research. Fisher's *Statistical Methods for Research Workers* has appeared in 14 editions over a period of 47 years. His ideas and those of Neyman and others spread so far and so profoundly that C Radhakrishna Rao (1965) declared, perhaps prematurely, statistics to be the technology of the twentieth century.

See also: Bayesian Statistics; Fiducial and Structural Statistical Inference; Fisher, Ronald A (1890–1962); Statistical Methods, History of: Pre-1900; Statistics, History of

Bibliography

Cramér H 1946 *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ

Efron B 1998 R. A. Fisher in the 21st Century. *Statistical Science* **13**: 95–114

Elderton E M, Pearson K 1910 *A First Study of the Influence of Parental Alcoholism on Physique and Ability of the Offspring*. The Galton Laboratory, London

Fisher R A 1934 Discussion of Neyman. *Journal of the Royal Statistical Society* **97**: 614–19

Fisher R A 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, Series A, **222**: 309–68

Fisher R A 1925 *Statistical Methods for Research Workers*. Reprinted in *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press, Oxford, 1990

Hald A 1998 *A History of Mathematical Statistics From 1750 to 1930*. Wiley, New York

Haavelmo T 1944 The probability approach in econometrics. *Econometrica* **12**(Supplement): 1–118

Hotelling H 1931 The generalization of Student's ratio. *Annals of Mathematical Statistics* **2**: 360–378

Hotelling H 1940 The teaching of statistics. *Annals of Mathematical Statistics* **11**: 457–70. Reprinted (1988) as 'Golden Oldy,' *Statistical Science* **3**: 57–71

Lehmann E L 1993 The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? *Journal of American Statistical Association* **88**: 1242–9

Morgan M S 1990 *The History of Econometric Ideas*. Cambridge University Press, Cambridge, UK

Neyman J 1934 On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**: 558–625

Neyman J 1941 Fiducial argument and the theory of confidence intervals. *Biometrika* **32**: 128–50

Rao C R 1965 *Linear Statistical Inference and its Applications*. Wiley

Spearman C 1904 General intelligence, objectively determined and measured. *American Journal of Psychology* **15**: 201–93

Stigler S M 1999 *Statistics on the Table. The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA

Thurstone L L 1935 *The Vector of Mind*. University of Chicago Press, Chicago, IL

Yule G U 1926 Why do we sometimes get non-sense correlation between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society* **18**: 1–63

T. Schweder

Statistical Methods, History of: Pre-1900

Statistics first assumed its recognizably modern form, as a set of tools and rationales for analyzing quantitative data, during the decade of the 1890s. Even then, it remained a far-flung and heterogeneous endeavor. Before this time, statistical methods were developed and put to use in the context of a variety of scientific and bureaucratic activities. These were sometimes recognized as analogous, but there were no common textbooks, nor even a shared name. The history of statistical methods up to 1900 means an examination of the development of statistical knowledge in various social and intellectual settings, and of the migration of techniques and understandings among them. The fields involved include mathematical probability, observational astronomy, quantitative natural history, physics, and an array of social science projects.

1. From Political Arithmetic to Statistics, 1660–1830

The name 'statistics' was coined in eighteenth-century Germany for a descriptive science of states. Whether or not statistical descriptions should involve numbers was left open at first. After 1800, in the polarizing era of the French Revolution, this question became more ideological, and the mere purveying of numerical tables was denounced for its slavishness, and for viewing the state as lifeless and mechanical. It thus entailed a change of mentality, suggestive also of new structures of government and administration, when this 'state science' began to be defined in terms of the collection and deployment of numbers. That happened first in Britain and France, where it was largely settled by 1840. The new science of statistics was closely related to censuses and other official information agencies, which expanded enormously in this same period (Hacking 1990).

The new statistics, commonly called a science, was practiced mainly by reformers and administrators. Its methods involved organization and classification rather than mathematics. Procuring reliable numbers, then as now, was no small achievement. Among those who hoped to move this social science in the direction of finding laws and abstract relationships, the highest priority was to standardize categories. In this way, they hoped, the 'statist' could compare rates of crime, disease, or illegitimacy across boundaries, and try to discover how they varied according to nationality, language, religion, race, climate, and legal arrangements. Such figures might also enable them to predict the potential for improvement if, for example, laws were changed or educational systems were put in place. A series of statistical congresses beginning in 1853 undertook to negotiate statistical uniformity, with very little success.

The comparative neglect of mathematics among the social statisticians was not simply a consequence of a lack of appropriate tools. Social numbers had been collected since the late seventeenth century under the rubric of 'political arithmetic,' a bolder and more freewheeling enterprise than statistics. John Graunt, a tradesman by origin, was elected to the Royal Society of London for his 'Observations upon the Bills of Mortality' (1662), a compilation and analysis of mortality records. William Petty, rather more daring, used what figures he could procure to calculate such quantities as the total wealth of Britain, including the monetary value of its human lives. On this basis he even proposed modestly to increase the wealth of the nation by transporting the Irish to England. In the eighteenth century, probability theory was deployed by the most illustrious of mathematicians to investigate social numbers. Jakob Bernoulli's *Ars conjectandi* (1713) addressed the convergence of observed ratios of events to the underlying probabilities. A discussion was provoked by the discovery that male births consistently outnumber female ones, in a ratio of about 14 to 13. In 1712, John Arbuthnot offered a natural theology of these birth numbers. The ratio was nicely supportive of belief in a benevolent God, and indeed of a faith without priestly celibacy, since the excess of male births was balanced by their greater mortality to leave the sexes in balance at the age of marriage. The odds that male births should exceed female ones for 82 consecutive years by 'chance' alone could be calculated, and the vanishingly small probability value stood as compelling evidence of the existence of God. Against this claim, Nicolas Bernoulli of the famous Swiss mathematical family showed that the steadiness of these odds required no providential intercession, but was about what one would anticipate from repeated tosses of dice with—he revised slightly the ratios—18 sides favoring M and 17 F (Hacking 1975).

Pierre Simon Laplace investigated birth ratios in the 1780s, and also discussed how to estimate a population from a partial count. He was party to an alliance in France between mathematicians and administrators, which also included the mathematician and philosopher Condorcet (Brian 1994). Laplace's population estimate was based on a registry of births for all of France, and on the possibility of estimating the ratio of total population to births through more local surveys. Proceeding as if the individuals surveyed were independent, and the birth numbers accurate, Laplace

used inverse probabilities to specify how many should be counted (771,469) in order to achieve odds of 1,000 to 1 against an error greater than half a million. The estimated population, which had to be fed into the previous calculation, was 25,299,417, attained by multiplying the average of births in 1781–2 by a presumed ratio of 26 (Gillispie 1997).

Laplace and Condorcet wrote also about judicial probabilities. Supposing a jury of 12, requiring at least 8 votes to convict, and that each juror had a specified, uniform probability of voting correctly, these mathematicians could calculate the risk of unjust convictions and undeserved acquittals. They also determined how the probabilities would change if the number of jurors were increased, or the required majority adjusted. The study of judicial probabilities reached its apex in the 1830s when Siméon-Denis Poisson made ingenious use of published judicial records to inform the calculations with empirical content. By this time, however, the whole topic had become a dinosaur. The use of mathematics to model human judgment now seemed implausible, if not immoral, and inverse probabilities (later called Bayesian) came under challenge (Daston 1988). Political arithmetic had come to seem weirdly speculative, and probabilistic population estimates both unreliable and unnecessary. In the new statistical age, sound practice meant a complete census. On this basis, some of the more mathematical methods of statistical reasoning were rejected around 1840 in favor of empirical modesty and bureaucratic thoroughness. Not until the twentieth century was sampling restored to favor.

One important application of probability to data grew up and flourished in place of these old strategies and techniques. That was the method of least squares. The problem of combining observations attracted a certain amount of attention in the eighteenth century. The first formal method to gain wide acceptance was worked out between about 1805 and 1810. It was initially published in 1805 by the French mathematician Adrien-Marie Legendre, who was then at work on the problem of measuring a line of meridian in order to fix the meter. Carl Friedrich Gauss asserted soon afterwards that he had already been employing least squares for several years, though there are reasons not to take his claim at face value (Stigler 1999, Chaps. 17 and 20). From 1809 to 1811, Gauss and then Laplace published probabilistic derivations of the method, and provided interpretations that shaped much of its subsequent history (see *Gauss, Carl Friedrich (1777–1855)*). Particularly crucial was its connection to what became known as the astronomer's error law, and later as the Gaussian or normal distribution. Gauss derived least squares by assuming that errors of observations conform to this law. Laplace showed how the error curve would be produced if each observation was subject to a host of small disturbances, which could equally well be positive or negative. The mathematics of error, and of least squares, formed the topic for a line of mathematical publications throughout the century. It provided a means of analyzing quantitative data in astronomy, meteorology, surveying, and other related observational sciences. Later, through a paradoxical transformation, the same mathematics was used to describe and analyze variation in nature and society.

2. *Social Science and Statistical Thinking*

The mathematical tools of statistics worked out by Laplace and Gauss were extended but not decisively changed in the period from 1830 to 1890. They engaged the attention of some excellent mathematicians, including Auguste Bravais, I.-J. Bienaymé, and A.-L. Cauchy (Hald 1998). The work of the Russian school of probability theory, initiated by P. L. Chebyshev, had consequences for statistical methods. But the really decisive development in this period involved statistics as a form of reasoning, worked out mainly in relation to social science, and involving, for the most part, comparatively rudimentary mathematics. This *statistical thinking* arose first of all in relation to the social science that bore the name 'statistics.' Statistics as a method was conceived by analogy to this substantive science. It was, to rely here on anachronistic terms, less a mode of *inference* than a family of *models*. They involved, at the most basic level, the production of collective order from the seemingly irregular or even chaotic behavior of individuals.

Like most discoveries, this one had ostensible precedents. Since, however, the Belgian Adolphe Quetelet had in the 1820s learned probability from Laplace, Poisson, and Joseph Fourier in Paris, and published several memoirs on vital statistics, his astonishment in 1829 should count for something. In that year, inspired by newly published French statistics of criminal justice, he attached a preface on crime to a statistical memoir on the Low Countries. He declared himself shocked at the 'frightening regularity with which the same crimes are reproduced,' year after year. He introduced a language of 'statistical laws.' At first he worried that such laws of moral actions might conflict with traditional doctrines of human free will, though he eventually rejected the idea of a 'strange fatalism' in favor of interpreting these laws as properties of a collective, of 'society.' His preoccupation with statistical regularities reflected the new emphasis in moral and political discussion on society as an autonomous entity, no longer to be regarded as subordinated to the state. But there was also something more abstract and, in the broadest sense, mathematical, in Quetelet's insight. It implied the possibility of a quantitative study of mass phenomena, requiring no knowledge at the level of individuals (Porter 1986).

Quetelet was an effective publicist for his discovery, which indeed was not uniquely his. Statistical laws

were widely discussed for half a century in journalism, literature, and social theory by authors like Dickens, Dostoevsky, and Marx. Moralists and philosophers worried about their implications for human freedom and responsibility. Natural scientists invoked social-statistical analogies to justify the application of this form of reasoning to physical and biological questions. The creation of a statistical physics by James Clerk Maxwell, Ludwig Boltzmann, and Josiah Willard Gibbs, and of statistical theories of heredity and evolution by Francis Galton and Karl Pearson, attests to the increasing range of these statistical theories or models. Precisely this expansion made it increasingly credible to define statistics as a method rather than a subject matter. At the same time, a new body of analytical tools began to be formulated from within this discourse, drawing from probability and error theory, but applying and adapting them to issues in social statistics.

The most pressing of these issues was precisely the stability of statistical series. A particularly forceful and uncompromising restatement of Quetelet's arguments in Henry Thomas Buckle's 1857 *History of Civilization in England* provoked an international debate about statistics and free will. One strategy of response was to explore this stability mathematically. The presentation and analysis of time series had long been associated with economic and population statistics, and with meteorology (Klein 1997). These debates provided a new focus. Was there really anything remarkable in the much-bruited regularities of crime, marriage, and suicide? Robert Campbell, schoolboy friend of the physicist Maxwell, published a paper in 1859 comparing Buckle's examples with expectations based on games of chance. He found the social phenomena to be much less regular than were series of random events. In Germany, where statistics was by this time a university-based social science, this question of the stability of social averages formed the core of a new mathematical approach to statistics. Wilhelm Lexis set out to develop tools that could simultaneously raise the scientific standing of statistics and free it from the false dogmas of 'French' social physics.

The normal departure of measured from true ratios in the case of repeated independent random events was expressed by Poisson as $\sqrt{(2pq/n)}$, where p and q are complementary probabilities and n the number of repetitions. Lexis began his investigation of the stability of statistical series in 1876 with a paper on the classic problem of male to female birth ratios. In this case, the tables of annual numbers were reasonably consistent with Poisson's formula. He called this a 'normal dispersion,' though it appeared to be unique. Nothing else—be it rainfall or temperature, or time series of births, deaths, crimes, or suicides, whether aggregated or broken down by sex or age—showed this much stability. He understood Buckle's deterministic claims as implying 'subnormal dispersion,' that is, an even narrower range of variation than the Poisson formula predicted. This could easily be rejected. Lexis's limits of error, defined by three times the Poisson number, corresponded with a probability of about 0.999978. Virtually every real statistical series showed a 'supernormal' dispersion beyond this range.

Laplace and Poisson seemed often to use legal and moral questions as the occasion to show off their mathematics. Lexis was serious in his pursuit of a quantitative social science, one that would reveal the social and moral structure of society. Against Quetelet, who had exalted the 'average man' as the type of the nation, he undertook to show that human populations were highly differentiated—not only by age and sex, but also by religion, language, and region. His program for social science was to begin by disaggregating, seeking out the variables that differentiate a population, and using statistical mathematics to determine which ones really mattered. Society, it could be shown, was not a mere collection of individuals with identical propensities, but a more organic community of heterogeneous groups. The wider than chance dispersions revealed that society was not composed of independent atomic individuals.

3. *Origins of the Biometric School*

Institutionally, the crucial development in the rise of modern methods of statistics was the formation of the biometric school under Karl Pearson at University College, London. The biometric program began as a statistical theory or model of natural processes, not a method of inference or estimation. Francis Galton, who pioneered the statistical study of heredity and evolution, derived his most basic conceptions from the tradition initiated by Quetelet. He was impressed particularly by the error curve, which Quetelet had applied to distributions of human measurements such as height and circumference of chest. Since he regarded the applicability of the error law as demonstrating the reality of his 'average man,' he minimized the conceptual novelty of this move. The mathematics, for him, showed that human variation really was akin to error. Galton strongly rejected this interpretation. The 'normal law of variation,' as he sometimes called it, was important for statistics precisely because it provided a way to go beyond mean values and to study the characteristics of exceptional individuals. The biometricians used distribution formulas also to study evolution, and especially eugenics, to which Galton and Pearson were both deeply committed (MacKenzie 1981).

Galton attached statistics to nature by way of a particulate theory of inheritance, Darwin's 'Pangenesis.' His statistical concepts were also, at least initially, biological ones: 'reversion' meant a partial return of ancestral traits; 'regression' a tendency to return to a

biological type; and 'correlation' referred to the tendency for measures of different bodily parts of one individual to vary in the same direction. By 1890, however, he recognized that his statistical methods were of general applicability, and not limited to biology. Pearson drew his statistical inspiration from Galton's investigations, though he may well have required the mediation of Francis Edgeworth to appreciate their potential (Stigler 1986). Edgeworth, an economist and statistician, was, like Pearson, a skillful and even original mathematician of statistics, but no institution builder.

Thus it was Pearson, above all, who created and put his stamp on this new field. Interpreting it in terms of his positivistic philosophy of science, he strongly emphasized the significance of correlation as more meaningful to science than causation. While he was committed above all to biometry and eugenics, he also gave examples from astronomy, meteorology, and education. Beginning in the mid-1890s, he attracted a continuing stream of postgraduate students, many but not all intending to apply statistics to biology. Among the first was George Udny Yule, an important statistician in his own right, who worked mainly on social and economic questions such as poverty. Pearson fought with many of his students, and also with other statisticians, most notably R. A. Fisher, and his reputation among modern-day statisticians has suffered on this account. Nevertheless, the 1890s were decisive for the history of statistical methods. Their transformation was not exclusively the work of the biometricians. This is also when the Norwegian A. N. Kiaer drew the attention of statisticians once again to the possibilities of sampling, which Arthur Bowley began to address mathematically. These trends in social statistics, along with the new mathematics of correlation and the chi-square test, were part of something larger. This was the modern statistical project, institutionalized above all by Pearson: the development of general mathematical methods for managing error and drawing conclusions from data.

See also: Bernoulli, Jacob I (1654–1705); Galton, Sir Francis (1822–1911); Laplace, Pierre-Simon, Marquis de (1749–1827); Pearson, Karl (1857–1936); Quetelet, Adolphe (1796–1874); Statistics, History of; Statistics: The Field

Bibliography

Brian E 1994 *La mesure de l'Etat: Administrateurs et géomètres au XVIIIe siècle*. Albin Michel, Paris

Daston L 1988 *Classical Probability in the Enlightenment*. Princeton University Press, Princeton, NJ

Desrosières A 1998 *The Politics of Large Numbers: A History of Statistical Reasoning*. Harvard University Press, Cambridge, MA

Gillispie C C 1997 *Pierre-Simon Laplace, 1749–1827: A Life in Exact Science*. Princeton University Press, Princeton, NJ

Hacking I 1975 *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical inference*. Cambridge University Press, London, UK

Hacking I 1990 *The Taming of Chance*. Cambridge University Press, London, UK

Hald A 1998 *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York

Klein J 1997 *Statistical Visions in Time: A History of Time-Series Analysis, 1662–1938*. Cambridge University Press, Cambridge, UK

Krüger L, Daston L, Heidelberger M (eds.) 1987 *The Probabilistic Revolution, Vol. 1: Ideas in History*. MIT Press, Cambridge, MA

MacKenzie D 1981 *Statistics in Britain, 1865–1930: The Social Construction of Scientific Knowledge*. Edinburgh University Press, Edinburgh, UK

Matthews J R 1992 *Mathematics and the Quest for Medical Certainty*. Princeton University Press, Princeton, UK

Porter T M 1986 *The Rise of Statistical Thinking, 1820–1900*. Princeton University Press, Princeton, NJ

Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press of Harvard University Press, Cambridge, MA

Stigler S 1999 *Statistics on the Table: The History of Statistical Concepts and Methods*. Belknap Press of Harvard University Press, Cambridge, MA

T. M. Porter

Statistical Pattern Recognition

Statistical pattern recognition is concerned with the problem of using statistical techniques to design machines that can classify patterns of measurement data in useful ways. For example, such techniques have been successfully exploited in areas such as object recognition, speech recognition, fingerprint recognition, data analysis, signal processing, cognitive modeling, and artificial intelligence. After identifying key 'features' of information in the measurement data, a statistical pattern recognition machine uses its expectations about the 'likelihood' of those features and the 'losses' of making situation-specific decisions to select a classification action which minimizes an appropriate expected risk function. In many cases, the statistical pattern recognition machine must 'learn' through experience the relative likelihood of occurrence of patterns of features in its environment. Although typically viewed as a branch of artificial intelligence (see *Decision Support Systems*) statistical pattern recognition methods have been extensively and successfully applied to many behavioral modeling and data analysis problems (see *Multidimensional Scaling in Psychology*; *Multivariate Analysis: Overview*). The present entry discusses some of the essential features of the generic statistical pattern recognition problem (see Andrews 1972, Duda and Hart 1973, Golden 1996, Patrick 1972).

1. Generic Statistical Pattern Recognition Problem

As shown in Fig. 1, the generic statistical pattern recognition problem may be described in terms of four major components: feature selection and feature extraction, probabilistic knowledge representation, Bayes decision rule for classification, and a mechanism for learning.

1.1 Feature Selection and Extraction

The pattern of measurement data to be classified is represented as a measurement vector (i.e., a point in 'measurement space'). The choice of the measurement space is primarily determined by problem-imposed considerations such as the types of transducers which are available and the types of patterns which must be processed. The particular pattern classification methodology embodied in the pattern recognition machine usually has little or no influence over the choice of the measurement space. Each measurement vector is then mapped into a feature vector (i.e., a 'point' in feature space) whose dimensions are primarily chosen to facilitate pattern classification by the pattern recognition machine. This process which maps a measurement vector **s** into a feature vector **x** is called the feature selection and extraction stage of the pattern recognition process.

1.2 Probabilistic Knowledge Representation

Knowledge is implicitly represented in a statistical pattern recognition machine in several ways. The particular process for mapping points in measurement space into feature space (as well as the dimensions of the feature space) is one important knowledge source. However, the expected likelihood of observing a particular feature vector and the associated losses for making different situation-specific decisions are also critical knowledge sources (see *Utility and Subjective Probability: Contemporary Theories*).

Specifically, let the symbol ω denote the pattern class. Thus, ω_1 refers to the first pattern class (e.g., a set of photographs of 'dogs') that might have generated the observed feature vector **x**, while the symbol ω_2 refers to a second competing pattern class (e.g., a set of photographs of 'cats') that might have generated **x**. The likelihood of observing a feature vector **x**, corresponding to a particular pattern class is expressed in terms of two types of probability distributions. The *a priori* class probability distribution $p(\omega_i)$ indicates the likelihood that a feature vector will be generated from the pattern class ω_i before the feature vector **x** has been observed. The class conditional probability distribution $p(\mathbf{x}|\omega_i)$ indicates the conditional likelihood of observing the feature vector **x** given that the feature vector **x** had in fact been generated from pattern class ω_i.

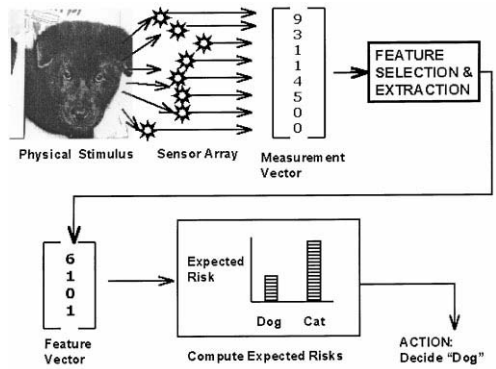

Figure 1
Flow chart diagram of a generic statistical pattern recognition machine. In this example, the sensor array (represented as a measurement vector) detects a pattern of light intensities produced from light reflecting off the head of a dog. A feature selection and extraction stage transforms the measurement vector into a feature vector. Using the feature vector and its probabilistic knowledge representation the pattern recognition machine selects the minimum expected risk decision (i.e., 'Decide Dog')

Knowledge regarding the expected consequences of taking a particular classification action α_j, given the feature vector came from pattern class ω_i is also exploited. Typically, the classification action takes the form of a decision indicating which of a finite number of pattern classes generated the feature vector. Formally, it is assumed that a loss $\lambda(\alpha_j|\omega_i)$ is incurred when classification action α_j is taken when the feature vector was in fact generated from pattern class ω_i.

1.3 Bayes Decision Rule for Classification

The optimal decision rule for selecting a classification action α_j if the pattern class ω_i generated the feature vector simply requires that the classification action α^* which minimizes the loss for the pattern class ω_i be chosen (i.e., $\lambda(\alpha|\omega_i) \geqslant \lambda(\alpha^*|\omega_i)$ for all α). In practice, however, the pattern class ω_i which actually generated the feature vector is unknown. The statistical pattern recognition machine only has access to the observed feature vector **x** and its probabilistic knowledge base.

One possible rational inductive decision making strategy is called the Bayes decision rule for classification; it forms the basis for classification decisions in most statistical pattern recognition algorithms. The Bayes decision rule for classification dictates that the pattern recognition machine should select the classification action α^* which minimizes the expected loss or

conditional risk given feature vector $\mathbf{x}$ is observed. Formally, the risk $R(\alpha|\mathbf{x})$ is defined by the formula:

$$R(\alpha|\mathbf{x}) = \sum_{i=1}^{M} \lambda(\alpha|\omega_i) p(\omega_i|\mathbf{x}) \tag{1}$$

The important minimum probability of error case of Eqn. (1) arises when $\lambda(\alpha_k|\omega_i) = 1$ when $i \neq k$ and $\lambda(\alpha_k|\omega_i) = 0$ for $i = k$. In this case, Eqn. (1) becomes:

$$R(\alpha_k|\mathbf{x}) = \sum_{i \neq k}^{M} p(\omega_i|\mathbf{x}) = 1 - p(\omega_k|\mathbf{x}) \tag{2}$$

The action α_k which minimizes the probability of error case of the risk function in Eqn. (1) is usually called a maximum *a posteriori* (MAP) estimate since α_k maximizes the *a posteriori* distribution $p(\omega_j|\mathbf{x})$ (i.e., the distribution obtained after observing the feature vector $\mathbf{x}$).

Now note that the risk $R(\alpha|\mathbf{x})$ in Eqn. (1) can be expressed in terms of the *a priori* class likelihoods and class conditional likelihoods using the formula:

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{p(\mathbf{x})} \tag{3}$$

Substituting this expression into Eqn. (1), it follows that the risk, $R(\alpha|\mathbf{x})$, for selecting action α given feature vector $\mathbf{x}$ is observed can be expressed as:

$$R(\alpha|\mathbf{x}) = \left(\frac{1}{p(\mathbf{x})}\right) \sum_{i=1}^{M} \lambda(\alpha|\omega_i) p(\mathbf{x}|\omega_i) p(\omega_i) \tag{4}$$

For computational reasons, some monotonically decreasing function of $R(\alpha_i|\mathbf{x})$ is usually exploited and referred to as a discriminant function $g_i(\mathbf{x})$. For example, a useful discriminant function for Eqn. (4) is:

$$g_i(\mathbf{x}) = -\sum_{i=1}^{M} \lambda(\alpha_i|\omega_i) p(\mathbf{x}|\omega_i) p(\omega_i) \tag{5}$$

which does not require computation of $p(\mathbf{x})$.

Such discriminant functions define 'decision boundaries' in feature space. For example, suppose that there are only two pattern classes with corresponding discriminant functions g_1 and g_2. In this situation, the feature vector $\mathbf{x}$ is classified as a member of pattern class ω_1 if $g_1(\mathbf{x}) - g_2(\mathbf{x}) < 0$ and a member of pattern class ω_2 if $g_1(\mathbf{x}) - g_2(\mathbf{x}) > 0$. Thus, the 'decision boundary' between the two pattern classes in feature space is defined by the set of feature vector which satisfy the equation: $g_1(\mathbf{x}) - g_2(\mathbf{x}) = 0$.

1.4 Learning as Bayes Estimation

Just as the classification problem involves selecting that classification action which minimizes an appropriate expected classification loss function, the learning problem can also be formulated within a minimum expected loss framework as well. Typically, the measurement vector, feature vector mapping from measurement space into feature space, *a priori* class distributions, and loss functions are known. Also the family of class conditional distributions or the 'class conditional probability model' is typically indexed by the values of some model parameter β. Given this latter assumption, a model is a set such that each element is a class conditional probability distribution $p(\mathbf{x}|\omega;\beta)$. The goal of the learning problem is then reduced to the problem of selecting a particular value of the parameter β.

1.4.1 A Bayes risk function for learning. Assume the measurement vector to feature vector mapping, classification decision loss function λ, and *a priori* class probabilities $\{p(\omega_1), \ldots, p(\omega_M)\}$ have been specified. In addition, assume a class conditional probability model defined by a set of class conditional distributions of the form $p(\mathbf{x}|\omega;\beta)$. The learning problem then reduces to the problem of estimating a value of the parameter β given a set of training data.

Let $\mathbf{X}_n = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}]$ denote a set of n feature vectors which define a set of training data. For expository reasons, simply assume $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are observed values (realizations) of independent and identically distributed (i.i.d.) random vectors. Let $p(\beta)$ denote the *a priori* probability distribution of β (i.e., the distribution of the parameter vector β before the training data $\mathbf{X}_n$ has been observed). Let $l(\beta, \beta^*)$ denote a loss function associated with the action for selecting a particular value of β when the correct choice of the parameter β is in fact β^*. A popular and general risk function for selecting parameter β given the training data $\mathbf{X}_n$ involves finding a β that minimizes the Bayes risk function for learning (e.g., Patrick 1972 pp. 71–2) given by the formula:

$$E(\beta|\mathbf{X}_n) = \int l(\beta, \beta^*) p(\beta^*|\mathbf{X}_n) d\beta^* \tag{6}$$

The quantity $p(\beta|\mathbf{X}_n)$ (known as the *a posteriori* density or *a posteriori* mass function for parameter vector β) is not explicitly provided since it can be expressed in terms of $p(\mathbf{x}|\omega;\beta)$ (the class conditional likelihood function), $p(\beta)$ (the *a priori* likelihood of β which is not dependent upon the pattern recognition machine's experiences), and the 'training data' $\mathbf{X}_n$. It then follows from the i.i.d. assumption that:

$$p(\beta|\mathbf{X}_n) = \frac{p(\mathbf{X}_n|\beta)}{p(\beta)} = \frac{\prod_{j=1}^{n} p(\mathbf{x}^{(i)}|\beta)}{p(\beta)} \tag{7}$$

For expository reasons, assume β is quantized so that it can take on only a finite number of values. Then minimizing Eqn. (6) using the loss function $l(\beta, \beta^*) = 1$ if $\beta \neq \beta^*$ and $l(\beta, \beta^*) = 0$ if $\beta = \beta^*$ corresponds to

maximizing Eqn. (7) (i.e., computing the MAP estimate or most probable value of β). It can be shown that for many popular *a priori* and class conditional distribution functions, $p(\beta|\mathbf{X}_n)$ can be well approximated by the 'likelihood of the observed data,' $p(\mathbf{X}_n|\beta)$ for sufficiently large sample size n (e.g., Golden 1996 pp. 299–301). Accordingly, a popular approach to parameter estimation known as maximum likelihood estimation involves maximizing the likelihood function $p(\mathbf{X}_n|\beta)$ by selecting a parameter vector value, β_n, such that $p(\mathbf{X}_n|\beta_n) \geqslant p(\mathbf{X}_n|\beta)$ for all β.

1.4.2 Nonparametric and parametric learning algorithms. Learning algorithms that make minimal assumptions regarding how the data was generated are referred to as nonparametric learning algorithms. For such algorithms, the parametric form of the class conditional distribution is often quite flexible. Nonparametric Parzen-window-type learning algorithms (e.g., Duda and Hart 1973, Patrick 1972) essentially count the observed occurrence of different values of the feature vector $\mathbf{x}$. Learning algorithms that incorporate considerable prior knowledge about the data generating process are referred to as parametric learning algorithms. The classical Bayes classifier which assumes that the class conditional distributions have a Gaussian distribution is a good example of a parametric learning algorithm (e.g., Duda and Hart 1973, Patrick 1972).

Nonparametric learning algorithms have the advantage of being relatively 'unbiased' in their analysis of the data but may require relatively large amounts of the right type of training data for effective learning. Parametric learning algorithms are 'biased' algorithms but if the right type of prior knowledge is 'built-in' to the learning algorithm then a parametric learning algorithm can exhibit superior learning from even poor quality training data. Statistical pattern recognition algorithms which are typically viewed as 'nonparametric' can be transformed into parametric learning algorithms if a sufficient number of constraints upon the class conditional distributions are imposed. For example, polynomial discriminant function (e.g., Duda and Hart 1973, Patrick 1972) and the closely related multilayer backpropagation neural network and perceptron type learning algorithms (e.g., Golden 1996; also see *Perceptrons*) can learn to approximate almost any desired form of the class conditional distribution function (given sufficiently many free parameters). Thus, such learning algorithms are best viewed as nonparametric statistical pattern recognition algorithms. On the other hand, prior knowledge in the form of constraints on the parametric form of the class conditional distributions can usually be introduced resulting in parametric variations of such classical nonparametric learning algorithms (e.g., Golden 1996).

1.4.3 Supervised, unsupervised, and reinforcement learning algorithms. Supervised, unsupervised, and reinforcement learning algorithms have the common goal of estimating the class conditional distribution from the training data. The essential difference among these different learning algorithms is the availability of information in the feature vector indicating which pattern class generated the feature vector. In supervised learning, each feature vector in the training data set contains explicit information regarding which pattern class generated that feature vector. In unsupervised learning, information regarding which pattern class generated a particular feature vector is not available in either the training data set or the test data set. In reinforcement learning, 'hints' about which pattern class generated a particular feature vector are provided either periodically or aperiodically throughout the learning process (Sutton and Barto 1998).

2. Applications

2.1 Artificial Intelligence

Statistical pattern recognition methods have been extensively applied in the field of artificial intelligence. Successful applications of these methods in the field of computer vision include extraction of low-level visual information from visual images, edge detection, extracting shape information from shading information, object segmentation, and object labeling (e.g., Chellapa and Jain 1993, Duda and Hart 1973). Statistical pattern recognition methods such as Hidden Markov models play an important role in speech recognition algorithms and natural language understanding (Charniak 1993). Bavesian networks defined on directed acyclic graphs and the closely related Markov random field methods are being applied to problems in inductive inference (Chellapa and Jain 1993, Golden 1996, Jordan 1999). In the fields of robot control and neurocomputing, the general stochastic dynamical system control problem (see *Optimal Control Theory*) is naturally viewed within a statistical pattern recognition framework where the control signal corresponds to the pattern recognition machine's 'action' (Golden 1996).

2.2 Behavioral Modeling

In an attempt to develop and refine theories of human behavior, some psychologists have compared the behavior of a class of statistical pattern recognition machines known as 'connectionist' algorithms with human behavior (see Golden 1996, and Ellis and Humphreys 1999, for some introductory reviews also see *Connectionist Models of Concept Learning*; *Connectionist Models of Development*; *Connectionist Models of Language Processing*). Although such algorithms

are typically not presented within a statistical pattern recognition context, they can usually be naturally interpreted as statistical pattern recognition machines (Golden 1996).

2.3 Data Analysis

Increasingly sophisticated methods for discovering complex structural regularities in large data sets are required in the social and behavioral sciences. Many classical statistical pattern recognition techniques such as factor analysis, principal component analysis, clustering analysis, and multidimensional scaling techniques have been successfully used (see *Multidimensional Scaling in Psychology*; *Multivariate Analysis: Overview*). More sophisticated statistical pattern recognition methods such as artificial neural networks (see *Artificial Neural Networks: Neurocomputation*; *Connectionist Approaches*) and graphical statistical models will form the basis of increasingly more important tools for detecting structural regularities in data collected by social and behavioral scientists.

See also: Categorization and Similarity Models; Classification: Conceptions in the Social Sciences; Configurational Analysis; Feature Representations in Cognitive Psychology; Letter and Character Recognition, Cognitive Psychology of; Monte Carlo Methods and Bayesian Computation: Overview; Multivariate Analysis: Classification and Discrimination; Object Recognition: Theories; Signal Detection Theory: Multidimensional; Spatial Pattern, Analysis of

Bibliography

Andrews H C 1972 *Introduction to Mathematical Techniques in Pattern Recognition*. Wiley, New York

Charniak E 1993 *Statistical Language Learning*. MIT Press, Cambridge, MA

Chellappa R, Jain A 1993 *Markov Random Fields: Theory and Application*. Academic Press, New York

Chen C H, Pau L F, Wang P S P 1993 *Handbook of Pattern Recognition & Computer Vision*. World Scientific, River Edge, NJ

Duda R O, Hart P E 1973 *Pattern Classification and Scene Analysis*. Wiley, New York

Ellis R, Humphreys G 1999 *Connectionist Psychology: A Text with Readings*. Psychology Press, East Sussex, UK

Golden R M 1996 *Mathematical Methods for Neural Network Analysis and Design*. MIT Press, Cambridge, MA

Jordan M I 1999 *Learning in Graphical Models*. MIT Press, Cambridge, MA

Nilsson N J 1965 *Learning Machines*. McGraw-Hill, New York

Patrick E A 1972 *Fundamentals of Pattern Recognition*. Prentice-Hall, Englewood Cliffs, NJ

Schervish M J 1995 *Theory of Statistics*. Springer-Verlag, New York

Sutton R S, Barto A G 1998 *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA

R. M. Golden

Statistical Sufficiency

In the language of statistical theory, a statistic is a function of a set of data, and a sufficient statistic contains as much information about the statistical model as the original set of data. Statistical sufficiency has served as a powerful concept in the theory of inference in helping to clarify the role of models and data, and in providing a framework for data reduction.

1. Introduction

Statistical sufficiency is a concept in the theory of statistical inference that is meant to capture an intuitive notion of summarizing a large and possibly complex set of data by relatively few summary numbers that carry the relevant information in the larger data set. For example, a public opinion poll might sample one or two thousand people, but will typically report simply the number of respondents and the percentage of respondents in each of a small number of categories. A study to investigate the effectiveness of a new treatment for weight loss may well report simply the average weight loss of a group of patients receiving the new treatment. This study would usually report as well some measure of the range of weight loss observed; variability in this setting would be a crucial piece of information for evaluating the effectiveness of the new treatment.

In order to make precise the vague notion of an informative data summary, it is necessary to have a means of defining what we mean by relevant information. This in turn is highly dependent on the use we wish to make of the data. The formal mechanism for doing this is to use a *statistical model* to represent a (usually idealized) version of the problem we are studying. Statistical sufficiency and information are defined relative to this model. A statistical model is a family of probability distributions, the central problem of statistical inference being to identify which member of the family generated the data currently of interest. The basic concepts of statistical models and sufficiency were set out in Fisher (1922, 1925) and continue to play a major role in the theory of inference.

For example, we might in an idealized version of the weight-loss problem assume that the potential weight loss for any particular patient can be described by a model in the family of *normal distributions*. This specifies a mathematical formula for the probability that the weight loss of an individual patient falls in any possible interval, and the mathematical formula depends on two unknown parameters, the mean and standard deviation. If we are willing to further assume that the same statistical model applies to all the potential recipients of the weight-loss treatment, then the data collected in a careful study will enable us to

make an inference on the two parameters that depends only on two data summaries: the average weight loss in the data and the standard deviation of the weight loss in the data.

The normal distribution is also referred to as the 'bell curve,' but there is in fact a family of bell curves, each differing from the others by a simple change of mean, which fixes the location of the bell curve, and standard deviation, which fixes the scale of the bell curve. Standardized tests for measuring IQ are designed so that the scores follow a bell-curve model with parameter values for the mean and standard deviation of 100 and 15, respectively. The parameters for the potential weight loss under our hypothetical new treatment are unknown, and the role of the data collected on the treatment is to shed some light on their likely values. Should the weight loss bell curve turn out to be centered at or near zero, one would be wise not to invest too heavily in the new company's stock just yet. As long as we are willing to accept the bell-curve model for the weight loss data, *all* the information about that model is contained in two numbers: the average and the standard deviation of the weight loss recorded in our data set.

An idealized model that lies behind most media reporting of polls is a *Bernoulli* or *binomial* model, in which we assume that potential respondents will choose one of two categories with a fixed but unknown probability, that respondents choose categories independently of each other, and with the same, constant probability. In such an idealized version of the problem, we have complete information about this unknown probability from the knowledge of the number of responses and the proportion of responses in each of the two categories.

These two examples are by construction highly idealized, and it is easy to think quickly of a number of complexities that have been overlooked. An educated reader of report on the weight-loss study would want to know how the patients were selected to receive the new treatment, whether the patients were in any way representative of a population of interest to that reader, whether any comparison has been made with alternative, possibly conventional, treatments, the length of time over which the study was conducted and so on. Public opinion polls typically offer more than two categories of response, even to a 'yes/no' question, including something equivalent to 'don't know,' or 'won't say,' and also often break down the results by gender, geographic area, or some other classification. Randomized selection of respondents by computerized telephone polls is very different to a 'reader response' survey often conducted by magazines and media outlets. These and a wide variety of other issues are studied in the fields of experimental design and sample survey design, by statisticians and researchers in a variety of subject matter areas, from economics to geophysics. See also *Observational Studies: Overview* and *Experimental Design: Overview*.

Statistical sufficiency is concerned with a much narrower issue, that of formalizing the notion that for a given family of models, a complex set of data can be summarized by a smaller collection of numbers, and lead to the same statistical inference. The following sections describe statistical sufficiency with reference to families of models and related concepts in the theory of statistics. Section 2 provides the main definitions and examples, Sect. 3 some related concepts, and Sect. 4 some more advanced topics.

2. *Statistical Models and Sufficiency*

2.1 *Basic Notation*

We describe a statistical model by a *random variable* Y and a family of *probability distributions* for Y, $\mathscr{F}$. A random variable is a variable that takes values on a *sample space*, and a probability distribution describes the probability of observing a particular value or range of values in the sample space. The sample space is in a great many applications either *discrete*, taking values on a countable set such as (some subset of) the nonnegative integers, or *continuous* taking values on (some subset of) the real line $\mathscr{R}$ or p-dimensional Euclidean space $\mathscr{R}^p$. A typical data set will be a number of realizations of Y, and the goal is to use these observations to deduce which member of $\mathscr{F}$ generated these realizations. The members of $\mathscr{F}$ are often conveniently indexed by the probability mass function or probability density function $f(y)$, where

$$f(y) = \text{Prob}(Y = y) \tag{1}$$

if the sample space is discrete, and

$$f(y) = \lim_{h \to 0} h^{-1}\{\text{Prob}(Y \leqslant y + h) - \text{Prob}(Y \leqslant y)\} \tag{2}$$

if the sample space is continuous. If Y is a vector then Prob $(Y \leqslant \text{y})$ is defined component-wise, i.e. Prob $(Y_1 \leqslant y_1, \ldots, Y_p \leqslant y_p)$.

Very often the mathematical form of the probability distributions in $\mathscr{F}$ is determined up to a finite number of unknown constants called parameters, so we write $\mathscr{F}_\theta = \{f(\cdot;\theta)\}$, or more precisely $\mathscr{F}_\theta = \{f(\cdot;\theta);\ \theta \in \Theta\}$, where Θ is a set of possible values for θ, such as the real numbers $\mathscr{R}$ or real-valued vectors of fixed length p, $\mathscr{R}^p$, or [0, 1], and so on. The problem of using data to make an inference about the probability distribution is now the problem of making inference about the parameter θ, and is referred to as *parametric inference*. Sufficiency also plays a role in *nonparametric inference* as well, and we illustrate this at the end of this section.

We assume in what follows that the random variable $Y = (Y_1, \ldots, Y_n)$ is a vector of length n with observed

valued $y = (y_1,\ldots, y_n)$. A *statistic* is a function $T = t(Y)$ whose distribution can be computed from the distribution of Y. Examples of statistics include the mean, $T_1 = n^{-1}\Sigma Y_i$, the variance $T_2 = (n-1)^{-1}\Sigma (Y_i - T_1)^2$, the standard deviation $T_3 = T_2^{1/2}$, the range $T_4 = \max Y_i - \min Y_i$, and so on. The probability distribution of any statistic is obtained from the probability distribution for Y by a mathematical calculation.

2.2 Sufficiency and Minimal Sufficiency

A statistic $T = t(Y)$ is sufficient for θ in the family of models $\mathscr{F}_\theta = \{f(y; \theta); \theta \in \Theta\}$ if and only if its conditional distribution

$$f_{Y|T}(y \mid t) \tag{3}$$

does not depend on θ.

Example 1. Suppose $Y_1,\ldots, Y_n$ are independent and identically distributed with probability mass function

$$f(y_i; \theta) = \theta^{y_i}(1-\theta)^{1-y_i}, \quad y_i = 0, 1; \quad \theta \in [0, 1] \tag{4}$$

which can be used to model the number of successes in n independent Bernoulli trials, when the probability of success if θ. The joint density for $Y = (Y_1,\ldots, Y_n)$ is

$$f(y; \theta) = \prod \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i} \tag{5}$$

and the marginal density for $T = \Sigma Y_i$ is

$$f(t; \theta) = \sum_{y:\Sigma y_i = t} \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i} = \theta^t(1-\theta)^{n-t}c(t, n) \tag{6}$$

where $c(t,n) = n!/t!(n-t)!$ is the number of vectors y of length n with t ones and $n-t$ zeroes. The conditional density of Y, given T, is the ratio of (5) and (6) and is thus free of θ.

The intuitive notion that the definition is meant to encapsulate is that once the sufficient statistic T is observed, i.e. once we condition on T, no further information about the θ is available from the original data Y.

The definition of sufficiency in terms of the conditional distribution is not very convenient to work with, especially in the case of continuous sample spaces where some care is needed in the definition of a conditional distribution. The following result is more helpful.

Factorization theorem. A statistic $T = t(Y)$ is sufficient for θ in the family of models $\mathscr{F}_\theta = \{f(y; \theta); \theta \in \Theta\}$ if and only if there exist functions $g(t; \theta)$ and $h(y)$ such that for all $\theta \in \Theta$

$$f(y; \theta) = g(t; \theta)h(y) \tag{7}$$

Example 2. Let $Y_1,\ldots, Y_n$ be independent, identically distributed from the normal distribution with mean μ and standard deviation σ:

$$f(y_i; \mu, \sigma) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \tag{8}$$

The joint density of $Y = (Y_1,\ldots, Y_n)$ is

$$\begin{aligned} f(y; \mu, \sigma) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i - \mu)^2\right\} \\ &= \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i - \bar{y})^2 + \frac{n}{2\sigma^2}(\bar{y} - \mu)^2\right\} \\ &\quad \times \left(\frac{1}{\sqrt{2\pi}}\right)^n \end{aligned} \tag{9}$$

where $\bar{y} = n^{-1}\Sigma y_i$ is the saple mean, $\theta = (\mu, \sigma)$, and $\Theta = \mathscr{R} \times \mathscr{R}^+$. From (9) we see that $T = \{\bar{Y}, \Sigma(Y_i - \bar{Y})^2\}$ is sufficient for θ, with $g(t; \theta)$ identified with all but the final factor in (9).

Strictly speaking, there are a number of other sufficient statistics in both the examples above. In Example 2, $(\Sigma_1^4 Y_i, \Sigma_5^n Y_i, \Sigma_1^7(Y_i - \bar{Y})^2, \Sigma_8^n(Y_i - \bar{Y}))$ is also sufficient, as is, trivially, the original vector Y itself. Clearly $(\bar{Y}, \Sigma_1^n(Y_i - \bar{Y})^2)$ is in some sense 'smaller' than these other sufficient statistics, and to be preferred for that reason. A *minimal sufficient statistic* is defined to be a function of every other sufficient statistic. One-to-one functions of the minimal sufficient statistic are also minimal sufficient and are not distinguished, so in Example 2 $(\Sigma_1^n Y_i, \Sigma_1^n Y_i^2)$ is also minimal sufficient. Any statistic defines a partition of the sample space in which all sample points leading to the same value of the statistic fall in the same partition, and the minimal sufficient statistic defines the coarsest possible partition.

Since the central problem of parametric statistical inference is to reason from the data back to the parameter θ, extensive use is made of the *likelihood function*, which is (proportional to) the joint density of the data, regarded as a function of θ:

$$L(\theta; y) = c(y)f y; \theta) \tag{10}$$

The factorization theorem suggests, and it can be proved, that the minimal sufficient statistic in the family of models $\mathscr{F}_\theta$ is the likelihood statistic $L(\cdot; Y)$. This result is easily derived from the factorization theorem if Θ is a finite set, but extending it to cases of more practical interest involves some mathematical intricacies (Barndorff-Nielsen, 1978, Chap. 4).

Example 3. To illustrate the concept of sufficiency in a nonparametric setting, suppose that $Y_1,\ldots, Y_n$ are independent, identically distributed on $\mathscr{R}$ with probability density function f; i.e. $\mathscr{F}$ consists of all

probability distributions which have densities with respect to Lebesgue measure. Denote by $Y_{(\cdot)}$ the vector of ordered values $Y_{(\cdot)} = (Y_{(1)} \leqslant Y_{(2)} \leqslant \ldots \leqslant Y_{(n)})$ (see *Order Statistics*). Then $Y_{(\cdot)}$ is sufficient for $\mathscr{F}$, as

$$f_{Y|Y_{(\cdot)}}(y|y_{(\cdot)}) = \begin{cases} \dfrac{1}{n!} & \text{if } y \text{ is a permutation of } y(\cdot) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

is the same for all members of $\mathscr{F}$

Example 4. Let $Y_1, \ldots, Y_n$ be independent, identically distributed according to the uniform distribution on $(\theta-1, \theta+1)$, where $\theta \in \mathscr{R}$. The order statistic $Y_{(\cdot)}$ is sufficient, as this is a special case of Example 3 above, but is not minimal sufficient. The minimal sufficient statistic is determined from the likelihood function:

$$L(\theta; y) = \begin{cases} \prod_{i=1}^{n} (\tfrac{1}{2}), & y_{(n)} - 1 \leqslant \theta \leqslant y_{(1)} + 1 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

from which we see that the minimal sufficient statistic is $T = (Y_{(1)}, Y_{(n)})$, as this is a one-to-one function of the likelihood statistic.

2.3 Exponential Families

A random vector $Y \in \mathscr{R}^p$ is said to be distributed according to an exponential family of dimension k if the probability mass or density function can be expressed in the form

$$f(y; \theta) = \exp\{\varphi(\theta)^T t(y) - c(\theta) - d(y)\} \tag{13}$$

for some functions $\varphi(\theta) = \{\varphi_1(\theta), \ldots, \varphi_k(\theta)\}$, $t(y) = \{t_1(y), \ldots, t_k(y)\}$, $c(\theta)$ and $d(y)$.

Many common distributions can be expressed in exponential family form, including Example 1 with $\varphi(\theta) = \log\{\theta/(1-\theta)\}$, $t(y) = \Sigma y_i$ and Example 2 with

$$\varphi(\theta) = \left(\frac{\mu}{\sigma}, -\frac{1}{2\sigma^2} \right)$$
$$t(y) = (\textstyle\sum y_i, \sum y_i^2)$$

Other examples of exponential families include the Poisson, geometric, negative binomial, multinomial, exponential, gamma, and inverse Gaussian distributions. Generalized linear models are regression models built on exponential families that have found wide practical application. See Dobson (1990) or McCullagh and Nelder (1989) and *Analysis of Variance and Generalized Linear Models*.

From (13) we can see that exponential families are closed under independent, indentically distributed sampling, and that the minimal sufficient statistic for θ is the k-dimensional vector $T = \Sigma t(Y_i)$. Exponential families are said to permit a *sufficiency reduction* under sampling; we can replace the original n observations by a vector of length k. Possibly for this reason exponential families are used widely as models in applied work.

A partial converse of this is also true, that if a finite dimensional minimal sufficient statistic exists in independent sampling from the same density, and the range of values over which the density is positive does not depend on θ, then that density must be of exponential family form.

Versions of the material in this appear in most mathematical statistics textbooks. Suggested references, roughly in order of increasing difficulty, are Azzalini (1996), Casella and Berger (1990), Lehmann and Casella (1998), Cox and Hinkley (1974), Pace and Salvan (1997), and Barndorff-Nielsen and Cox (1994).

3. Some Related Concepts

3.1 Ancillary Statistics

A companion concept for the reduction of data to a number of smaller quantities is that of *ancillarity*. A statistic $A = a(Y)$ is said to be ancillary for θ in the family of models $\mathscr{F}_\theta$ if the marginal distribution of A,

$$f_A(a) = \int_{\{y:\, a(y) = a\}} f(y; \theta)\, \mathrm{d}y \tag{14}$$

is free of θ. If every other ancillary statistic is a function of a given ancillary statistic, the given statistic is a maximal ancillary statistic.

Example 4 (continued). In independent sampling from the $U(\theta, \theta+1)$ distribution, $T = (Y_{(1)}, Y_{(n)})$ is a minimal sufficient statistic, and $A = Y_{(n)} - Y_{(1)}$ is a maximal ancillary statistic.

If an ancillary statistic exists, all the information about θ is in the conditional distribution of Y, given the ancillary statistic. If a minimal sufficient statistic exists, all the information about θ is in the marginal distribution of T. These two data reductions, to A or to T, are complementary, and can sometimes be combined.

Example 5. Let $Y_1, \ldots, Y_n$ be independent, identically distributed from the one parameter *location* family $f(y_i; \theta) = f_0(y_i - \theta)$, $\theta \in \mathscr{R}$, where $f_0(\cdot)$ is known. Then the vector of residuals $A = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})$ is ancillary. An extension of this to the location scale family is possible: let $f(y_i; \theta) = \sigma^{-1} f_0\{\sigma^{-1}(y_i - \mu)\}$, then

$$A = \{(Y_1 - \bar{Y})/S, \ldots, (Y_n - \bar{Y})/S\}, \tag{15}$$

where $S^2 = \Sigma(Y_i - \bar{Y})^2$, is a maximal ancillary. The normal distribution of Example 4 is a location-scale

model, with minimal sufficient statistic $T = (\bar{Y}, S^2)$, which in this case is independent of A.

The importance for the theory of statistics in concepts of dimension reduction such as ancillarity and sufficiency is that inference for a k-dimensional parameter is encapsulated by a probability distribution on a k-dimensional space. This immediately provides tests of significance, confidence intervals, and point estimates for the unknown parameter. From this point of view it now appears that ancillarity is possibly a more central notion to the theory of statistics, although sufficiency is regarded widely as more natural and intuitive.

It should be noted that inference from a Bayesian point of view automatically provides a k-dimensional distribution for inference for a k-dimensional parameter, which is the posterior distribution given the observed data y. This is achieved at the expense of constructing a probability model for the unknown parameter θ. See *Bayesian Statistics*.

3.2 Asymptotic Sufficiency

The normal distribution is used widely for modelling data from a variety of sources, but is also very important in the theory of inference in providing approximate inferences, when exact inferences may be unavailable. While the likelihood function is equivalent to the minimal sufficient statistic, a special role in approximate inference is played by some functions computed from the likelihood function.

The *maximum likelihood estimator* of θ in the model $f(y; \theta)$ is defined by

$$\sup_{\theta} L(\theta; y) = L(\hat{\theta}; y) \tag{16}$$

and the *observed information function* for θ is

$$j(\theta) = -\partial^2[\log\{L(\theta; y)\}]/\partial\theta\partial\theta' \tag{17}$$

Under regularity conditions on the family of models for $Y = (Y_1, \ldots, Y_n)$ we have the following asymptotic result for a scalar parameter θ as $n \to \infty$:

$$(\hat{\theta} - \theta_0)\,\{j(\theta_0)\}^{1/2} \xrightarrow{d} \theta_0\, N(0, 1) \tag{18}$$

which means that for sufficiently large n the distribution of the maximum likelihood estimator is arbitrarily close to a normal distribution with mean zero and variance $j^{-1}(\theta_0)$, under sampling from $f(y; \theta_0)$. A similar result is available for vector θ.

The maximum likelihood estimator has by construction the same dimension as the unknown parameter θ, so (18) provides a means of constructing an inference for θ using the normal approximation. This suggests that a notion of *asymptotic sufficiency* could be formalized, the maximum likelihood estimator having this property. This formalization relies on showing that the likelihood function is asymptotically of normal form, with minimal sufficient statistic $\hat{\theta}$; see Fraser (1976, Chap. 8). The argument is very similar to that showing that the posterior distribution for θ is asymptotically normal. More generally, in models with smooth likelihood functions, a Taylor series expansion of the likelihood function around $\hat{\theta}$ suggests a refinement of asymptotic sufficiency: the set of statistics $\hat{\theta}$ and second and higher order derivatives of the likelihood function evaluated at $\hat{\theta}$ form a set of *approximate sufficient statistics* in a certain sense (Barndorff-Nielsen and Cox 1994, Chap. 7).

In general models the maximum likelihood estimator is not the minimal sufficient statistic, although it is always a (many-to-one) function of the minimal sufficient statistic. Much of the recent development of the parametric theory of inference has been concerned with combining sufficiency or approximate sufficiency with ancillarity or an appropriate notion of approximate ancillarity, to effect a dimension reduction with minimal loss of information. In general models if the maximum likelihood estimator is used in place of the minimal sufficient statistic, information is lost. The idea of recovering this information, via conditioning on an ancillary statistic, was put forth in Fisher (1934). With some exceptions this idea was not exploited fully in the statistical literature until about 1980, when it was discovered that an asymptotic version of Fisher's argument holds in very general situations. The subsequent 20 years saw continued development and refinement of Fisher's original idea. A brief overview is given in Reid (2000), and the books Barndorff-Nielsen and Cox (1994) and Pace and Salvan (1997) give a detailed account.

3.3 Inference in the Presence of Nuisance Parameters

Many statistical models have vector valued parameters θ, but often some components or scalar valued functions of θ are of particular interest. For example, in evaluating the hypothetical new weight-loss treatment under a normal model, the mean parameter μ is in the first instance of more interest than the standard deviation σ. In studying results from a poll we may wish to introduce parameters describing various aspects of the population, such as gender, age group, income bracket, and so on, but the main purpose of the poll is to construct an inference about the proportion likely to vote for our candidate, for example.

The idealized modelling of these types of situations is to describe θ as consisting of two components $\theta = (\psi, \lambda)$, of *parameters of interest*, ψ and *nuisance parameters*, λ. An extension of this would be to define $\psi = g(\theta)$ as determined by a set of constraints on θ,

but only the component version will be considered here.

There are in the literature several different definitions of sufficiency for a parameter of interest ψ in the presence of a nuisance parameter λ. To illustrate some of the issues we consider again the normal distribution with $\theta = (\mu, \sigma)$, $T = (\bar{Y}, s^2)$. We have

$$f_T(t; \mu, \sigma) = f_{\bar{Y}}(\bar{y}; \mu, \sigma) f_{S^2|Y}(s^2; \sigma^2) = f_{\bar{Y}}(\bar{y}; \mu, \sigma) f_{S^{12}}(s^2; \sigma^2) \quad (19)$$

where $f_{\bar{Y}}(\cdot)$ is the density of a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, and $f_{S^2}(\cdot)$ is proportional to that of a chi-squared distribution on $n-1$ degrees of freedom. As the conditional distribution of S^2, given $\bar{Y}$ is free of μ, we could define $\bar{Y}$ as sufficient for μ, by analogy to (3). However it is not sufficient for μ in the sense of providing everything we need for inference about μ, since its distribution depends also on σ. In this latter sense S^2 is sufficient for σ.

There have been a number of attempts to formalize the notions of partial sufficiency and partial ancillarity, but none has been particularly successful. However, the notion of dimension reduction through exact or approximate sufficiency and ancillarity has played a very important role in the theory of parametric inference.

Sufficiency in the presence of nuisance parameters is considered in Barndorff-Nielsen (1978, Chap. 4), Jorgensen (1997) and Bhapkar (1991). A discussion of partitions like (19) and the role of conditional and marginal inference is given in Reid (1995).

4. Conclusion

A sufficient statistic is defined relative to a statistical model, usually a parametric model, and provides all the information in the data about that model or the parameters of that model. In many cases the sufficient statistic provides a substantial reduction of complexity or dimension from the original data, and as a result the concept of sufficiency makes the task of statistical inference simpler. The related concept of ancillarity has a similar function. Recent developments in likelihood inference have emphasized the constructions of approximately sufficient and approximately ancillary statistics in cases where further dimension reduction is needed after invoking sufficiency and ancillarity. Excellent textbook references for the definitions and examples include Casella and Berger (1990) and Azzalini (1996). More advanced accounts of recent work are given in Barndorff-Nielsen and Cox (1994) and Pace and Salvan (1997).

See also: Analysis of Variance and Generalized Linear Models; Bayesian Statistics; Distributions, Statistical: Approximations; Experimental Design: Overview; Fiducial and Structural Statistical Inference; Frequentist Inference; Observational Studies: Overview; Order Statistics

Bibliography

Azzalini A 1996 *Statistical Inference*. Chapman and Hall, London

Barndorff-Nielsen O E 1978 *Information and Exponential Families in Statistical Theory*. Wiley, New York

Barndorff-Nielsen O E, Cox D R 1994 *Inference and Asymptotics*. Chapman and Hall, London

Bhapkar V P 1991 Sufficiency, ancillarity and information in estimating functions. In: Godambe V P (ed.) *Estimating Functions*. Oxford University Press, Oxford, UK

Casella G, Berger R L 1990 *Statistical Inference*. Brooks-Cole, Pacific Grove, CA

Cox D R, Hinkley D V 1974 *Theoretical Statistics*. Chapman and Hall, London

Dobson A J 1990 *An Introduction to Generalized Linear Models*. Chapman and Hall, London

Fisher R A 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of The Royal Society London, Series A* **222**: 309–68. Reprinted. In: Kotz S, Johnson N L 1992 (eds.) *Breakthroughs in Statistics*. Springer, New York

Fisher R A 1925 Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**: 700–25

Fisher R A 1934 Two new properties of mathematical likelihood. *Proceedings of the Royal Society Series A* **144**: 285–307

Fraser D A S 1976 *Probability and Statistics: Theory and Applications*. Duxbury Press, North Scituate, MA

Jorgensen B 1997 The rules of conditional inference: Is there a universal definition of non-formation? *Journal of the Italian Statistical Society* **3**: 355–84

Lehmann E L, Casella G 1998 *Theory of Point Estimation*, 2nd edn. Springer, New York

McCullagh P, Nelder J A 1989 *Generalized Linear Models*, 2nd edn. Chapman and Hall, London

Pace L, Salvan A 1997 *Principles of Statistical Inference From a Neo-Fisherian Perspective*. World Scientific, Singapore

Reid N 1995 The roles conditioning in inference (with discussion). *Statistical Science* **10**: 138–57

Reid N 2000 Likelihood. *Journal of the American Statistical Association* **95**: 1335–40

N. Reid

Statistical Systems: Censuses of Population

1. Early Sporadic Censuses

'The Census at Bethlehem,' painted by Pieter Bruegel (the Younger, around 1600) shows the Virgin Mary on a donkey, going to Nazareth to be counted at home by the Roman Census. There are countries even at the

beginning of the twenty-first century where the patriotic 'ceremony' of the census requires people to be at their usual (*de jure*) home for easier counting. A census was also conducted in AD 2 in the relatively newly unified China, where they counted 60 million persons, about a quarter of the world's population then. (At 1.3 out of 6 billions today, China is still one-fifth of the world's population and would be one-quarter without her rigorous population controls.) A millennium later in 1086, a sort of census was conducted in England for William the Conqueror, known as the Domesday Book. These are three prominent examples from several known to have been undertaken in different countries at different times, and they illustrate the importance placed on population counts. However, they were only sporadic and highly variable attempts, for which we have no complete records, especially because some of them were kept as 'state secrets.' There were also many censuses for only single cities, districts, or provinces. Furthermore, they were not usually counts of households or of all persons. Some counted only men of military age for allocating levies of troops; some counted only chimneys, used as an easy proxy for households. The chief aims of censuses were to apportion to administrative units legislative seats, the burden of taxation, army levies, or of work assignments. These aims survive today, but others have been added in modern times. All of these aims are supported by the details for small domains that complete censuses yield. (See Goyer 1992, Taeuber 1986).

2. Rise of Modern Censuses

The first modern, national census of households and persons was established in the Constitution of the new USA in 1790 and has been continued decennially since then. Its first major purpose was allocating seats in the new national Congress, but it also served other purposes and needs (Anderson 1988). Those purposes and needs increased gradually and the variables collected by the census increased likewise. The idea spread and decennial censuses began in the UK in 1801 and in Belgium in 1846. The International Statistical Institute's Congresses of 1853, 1872, and 1897 passed recommendations in favor of decennial censuses for all countries, with some basic requirements, aiming at standardization for better international comparability. It also urged open publication and dissemination of results. In 1900 some 68 different countries already conducted and made available population censuses, which included 43 percent of the world's population. The quality and content of the censuses improved gradually. For example, it is only since 1850 that the data for USA censuses covered all individuals rather than just families.

The greatest international impetus came after 1946 from the Population Division and the Statistical Division of the new United Nations. The stimulation, guidance, and coordination of the UN Statistical Division were crucial for spreading the establishment of censuses and their comparability (UN 1980). The efforts of the UN's sister agencies (FAO, WHO, and UNESCO) were also important. Beyond the growing coverage of most countries and improved coverage within countries, population censuses have advanced in five other dimensions. First, the quality of the census data has been improved. Second, that improvement has been subjected to measurement by quality checks, often with duplicate observations. Third, both the contents and the quality have been increasingly standardized across countries in order to facilitate international comparability of data. Fourth, the number, diversity, and depth of census variables have been greatly increased. Fifth, the analyses of census variables have been vastly deepened and improved, with the help of modern computing machines and methods.

In addition to censuses of population and housing, many countries also conduct censuses of agriculture, business, manufacturing and mineral industries, etc. These are usually done separately, often by different and specialized staffs and enumerators. The censuses of agriculture may still be linked to population censuses in less developed countries, where most households still also operate small farms. However, in industrialized countries most agricultural products come from large firms of 'agro-business.'

2.1 Contents of Population Censuses

Beyond mere head counts, modern population censuses obtain several classes of variables, as shown below, condensed from the UN's recommended list:

Residences—usual, on the census date, duration, previous, at birth; place of work

Sex, age, marital status, relation to head of family and household

Citizenship, religion, language, national/ethnic origin

Children: born alive, living. Births last year: live and dead

Marriage: age, duration, order

Education: attained, literacy, qualifications, attendance

Occupation: status, industry sector, time worked

Income, socioeconomic status, economic dependency

From this recommended list, most countries omit some topics or include them only in some years. On the other hand, some countries add other topics, of course, as desired. With all those variables, and with increasingly complex and detailed analyses, the published outputs of censuses grew into census monographs of many volumes, and then took years to finish (Farley

1995). Therefore statistical offices are continuing to develop cheaper and faster methods of dissemination, with the aid of computers and electronic publishing.

2.2 Timing of Censuses

Censuses of the population must assign specific locations on specific dates (the census day or reference period) to people. The collection period for enumerating may be one week or a few weeks. However, the reference period is a single census day for items like residence and age; for items like income (or crops) it may be the preceding year. However, humans are highly mobile and they are becoming even more so: with longer and more vacations, longer travel to work, more second homes, more split families, and so on. *De facto* is the census term for the actual location of persons on census day and *de jure* for their 'usual place of residence' (home). Without this distinction city workers would make the central cities larger in the daytime when their 'usual place of residence' locates them in the suburbs. But location problems are more difficult for students living away from home in university cities; for death rates in cancer hospitals far from home; large institutional populations in prisons or in the military; for second vacation homes, etc. These are all examples that census definitions and operations must cover, in the field and in the office. Nomadic populations pose a great variety of special problems in timing and locations, of too great a variety for us to attempt to cover here.

For the census work in most countries, tens or hundreds of thousands of enumerators must be hired and trained for just one or two weeks of fieldwork and a few months in the offices. In China and India they need millions of enumerators. Some experts have doubts about how well they can be trained. However, others contend that the 'ceremony,' the patriotic campaign that can be mounted every 10 years, with educational campaigns, and exhortations in the media, increases awareness, participation, and response rates. Still others fear that the campaign arouses the latent and dormant opposition to fact-finding.

Decennial censuses may be regarded as systematic samples of a continuous time dimension: samples of a single census day in the 3,652 days of 10 years, or one year every 10 years for variables like yearly income. Estimates for intercensal and postcensal periods must depend on models and methods, and these are called intercensal and postcensal estimates. Other methods are also available for national and large domain estimates; but what are lacking above all are estimates for small areas and small domains that are also more detailed in time than decennial censuses. These needs have given impetus to methods of small area estimates and small domain estimates (Platek et al. 1987, Purcell and Kish 1986). Their accuracy depends on relations to auxiliary data from registers and sample surveys; also on the stability of variables like total population. But they are not accurate for unstable and fluctuating variables, like unemployment, infectious diseases, and financial variables. Sample surveys have been developed for some of these (see Sect. 3).

2.3 Coverage by Censuses

All censuses have problems of coverage; and even the best suffer a few percentages of noncoverage (or missing units). Furthermore, typically the units (households, persons) are not missed evenly, or at random; they are typically missed more often among the lower socioeconomic classes, some ethnic groups, and areas. These uneven coverages persist despite efforts to find and to measure them. However, the size and bias of this noncoverage may often be smaller than the even greater distortions due to the uses of obsolete decennial censuses for 10 years. All national census results also consist of combinations of regions or provinces or states, also of districts or counties or communes, and so on down to small domains. The populations of the provinces may differ greatly in culture and standards and the national results represent averages of the provincial levels. For some provinces different languages and questionnaires may be needed. Further, deliberate omissions may exclude some areas because: (a) they are vast, 'uninhabited' mountains, deserts, jungles, or small islands; or (b) they may be areas under hostile populations or rebellions. Changes of boundaries that have occurred, either losses or acquisitions, after wars and treaties cause problems for publishing and reading census monographs, tables, and graphs.

2.4 Costs of Censuses

The fine spatial detail of complete censuses is bought with high total costs. Because the total costs of censuses are high, they are usually only held every 10 years; and only a few countries had five-year censuses. The growing number of topics covered by censuses have also increased their total high costs and these prevent more frequent census taking, despite demands for annual data with spatial detail. At the same time the large size of the census is also an obstacle to obtaining greater richness of data (with more variables) with spatial details. This high cost is one cause of the sporadic political controversies about censuses in many countries. But even more serious are the conflicts due to the policy implications of census results, especially those that are used for apportionment between provinces and between ethnic groups. Apportionment of seats in Congress (Parliament), of taxation, of armies, of welfare funds—the very reasons for collecting census data—can also lead to conflicts, and even (rarely) to suppression of data. The collection

and open publication of reliable and trusted statistical data is (like liberty) 'the fruit of constant vigilance.'

The costs of censuses must vary with the methods used to collect and process them. Door-to-door personal interviews have been the standard throughout history. However, now they have been modified in places with telephoned and mailed-in questionnaires, and also with mailed-out-mailed-in postal methods. Processing and editing have been speeded with mechanical and electronic methods; but the savings seem to be swallowed by increased complexities of analyses. The comparison of total costs across times, and between countries are complicated by population sizes, exchange rates, and living standards; but guesses are made that the total cost of censuses per capita population equal about 1/2 to 1 hour of the countries' median hourly wage for every 10 years.

3. Sampling and Censuses

Two other methods for collecting population data have emerged as both complementary to and competitive with complete census taking: sampling methods and administrative registers. Sampling methods have come to be used in several distinct ways in different countries.

3.1 Census Samples

The addition of great numbers of desirable topics increased the costs of censuses, and since 1950 a solution to this problem was increasingly found in the use of census samples. In addition to the complete count using the 'short form' containing a few basic questions, a sample of the population completes a 'long form' with many variables. This sampling reduces both the total cost of the census and the burden on the majority of respondents in 2000. The rates and methods of sampling differ between countries. The sampling units may be enumeration districts (areas—EDs or EAs), or households, or persons. Selection of persons has lower sampling variance, but selecting EDs may have better enumerative controls and greater flexibility. The sample population may only get the long form or they may get duplicates of both the short and the long forms. Census samples generate large samples, whether 5 or 10 or 20 percent of the complete count, and thus they can provide detailed spatial data, but with reduced costs. They are still much larger than ordinary survey samples, which have selection rates of 1/100, 1/1,000 or 1/10,000 of complete counts.

3.2 Postenumerative Surveys, Quality Checks, Coverage Checks

Samples may be taken for quality checks with measurements that may use more costly and (hopefully) improved methods. These measurements may be in addition to, or instead of, the regular census data collection and the sampled units may be persons, households, or enumeration districts. Coverage checks are attempts to find the sort of persons missed by the regular census count. The surveys are often called postenumerative because they are usually done after the census collection, but they may be done simultaneously.

3.3 Sample Censuses

These have been conducted by a few countries, and sometimes are called 'minicensuses,' perhaps based on one-percent samples and with fewer topics than the sample surveys described below. The greatest expressed need is for annual data to supplement the decennial censuses, which are increasingly judged to be inadequate because of obsolescence. Typically, census monographs containing the richer data of the long form become available five or more years after the census year, although the simple counts may be released quickly (Farley 1995).

3.4 Sample Surveys

Generally, sample surveys differ greatly from complete censuses; they are much smaller; they are much cheaper, but they also lack detail in spatial (geographic) and other domains. But periodic samples yield much needed temporal detail, and also have other advantages. They can yield much richer data, especially because the content and the methods can be focused on specific contexts. Also because that greater concentration can be achieved for the questionnaires, for the interviewers and for the respondents as well.

'Representative sampling' was first proposed by F. N. Kiaer in 1895 in Norway, for replacing censuses. The fundamental concept requires a careful, scientific method for selecting a sample from the population and then estimating the population variables from the sample statistics. Many statisticians, academic and practical, in many journals and in the International Statistical Institute for about 50 years, opposed sampling as a substitute for complete enumeration. Meanwhile probability samples came into use for social surveys, notably for unemployment rates and then for the famous quarterly surveys of the labor force, which became the Current Population Surveys of the US Bureau of the Census (1978). All countries in the European Union and many others now have quarterly or monthly surveys of the national labor force. These surveys, mostly of 5,000 to 100,000 households, are financed because governments need current monthly or quarterly data on (un)employment and its periodic fluctuations. National statistical offices generally conduct many other samples every year, some of them only once, some repeatedly, and

some periodically. Some of the censuses are also produced by sampling methods and the Food and Agriculture Agency (FAO) has been in the forefront of promoting agricultural censuses by means of sample surveys (see *Social Survey, History of*).

4. Administrative Registers

Administrative registers related to population counts, sometimes called civil or population registers, may be of different kinds, purpose, and quality in different countries and at various times. Vital registers of births and deaths may be the most common examples, but others exist for voter registration, or recently auto licenses, telephone numbers, etc. The most relevant registers date from 1749 in Sweden and the Nordic countries, where the official church started keeping records of the parish populations, which continues today. Some statisticians propose replacing decennial censuses with data from administrative registers of different kinds, or with the combined social resources available in specific countries (see *Databases, Core: Demography and Registers*). The data are inexpensive, almost without cost, because some other agency, with other aims and resources, pays for them. Furthermore, they are often detailed spatially, and also detailed temporally, timely, and they are kept up-to-date. They have been used in several Scandinavian countries to replace their population counts. However for two chief reasons, most countries cannot use administrative registers to replace censuses. First, for lack of the conditions that facilitate the successes of Scandinavian registers: a highly educated and cooperative society, which is highly motivated to register promptly their own changes of residence. Second, no registers, not even the Scandinavian registers, have ever obtained the richer data (beyond a few lines) that modern censuses need. In addition, some statisticians fear the misuse of registers by careless, or malevolent, governments and societies.

5. Future Developments

Modern censuses face some gravely conflicting demands: higher quality and better coverage of the population, but with reduction of mounting costs. We need spatial detail, but also temporal detail and timeliness. Censuses yield more spatial detail, but sample surveys are superior in all of the three other aspects. The greatest need is for more timeliness, because decennial data do not satisfy modern needs, especially for annual data with spatial details. At the other extreme the monthly surveys provide adequate national statistics and satisfy the other requirements, but they yield no spatial detail for local needs. This conflict gave rise to designs of 'cumulative representative samples' and then to 'rolling samples,' in the last few years. The National Household Health Survey in the USA yields separate samples of about 1,000 households each week, and they are cumulated into 52,000 households yearly, with health data on about 150,000 persons. Thus the yearly cumulations yield data for even moderately rare diseases and infirmities. Furthermore, the American Community Survey of the US Census Bureau is conducting a large pilot study for a 'rolling sample' by AD 2003, which may replace the long form of the US Census of 2010 (Alexander 1996, Kish 1998). Many statisticians in AD 2000 felt that the greatest need generally is for regular, standardized annual surveys (sample or census) that are harmonized internationally, and also large enough to yield reliable data for small domains. It is likely that the censuses of the twenty-first century will be quite different from those of the twentieth century.

Bibliography

Alexander C 1996 A rolling example survey for yearly and decennial uses. *Bulletin of the International Statistical Institute*. Helsinki, 52nd session

Alterman H 1969 *Counting People, The Census in History*. Harcourt, Brace and World, New York

Anderson M 1988 *The American Census: A Social History*. Yale University Press, New Haven, CT

Farley R 1995 *State of the Union, Vols I and II*. Russell Sage, New York

Goyer D S 1992 *Handbook of National Population Censuses: Europe*. Greenwood Press, NY

Kish L 1998 Space/time variations and rolling samples. *Journal of Official Statistics* **13**: 31–46

Kish L, Verma F 1986 Complete censuses and samples. *Journal of Official Statistics* **2**: 381–95

Platek R, Rao J N K, Sarndal C E, Singh M P 1987 *Small Area Statistics*. Wiley, New York

Purcell N J, Kish L 1987 Estimation for small domains. *Biometrics* **35**: 365–84

Shryock H S, Siegel J S 1973 *The Methods and Materials of Demography*. US Bureau of the Census, Washington, DC

Taeuber C 1986 Census. In: *International Encyclopedia of the Social Sciences*. Macmillan, New York

United Nations 1980 Principals and recommendations for population and housing censuses. *Series M No. 67* (Sales E.80.XV11.8)

US Bureau of the Census 1978 The current population survey: Design and methodology. *USCB, Technical Paper 40*, Washington

L. Kish

Statistical Systems: Crime

1. Crime Statistics

Crime statistics are largely generated from three avenues—administrative data, surveys, and ethnographic studies. Administrative data are often referred to as the 'official' statistics as they are drawn most often from the records kept by state institutions. The

first official statistics were published in France in 1827 (Coleman and Moynihan 1996). Self-report surveys are the other major source of crime data or statistics and involve interviewing perpetrators and victims of crime. In general ethnographic studies do not result in 'statistics' as such. They are based on small samples that are usually unrepresentative of the population. However, such rich materials play an important role in understanding what lies behind the numbers used by crime analysts.

The collection and monitoring of official crime statistics has largely been taken over by government and government sponsored agencies such as the US Bureau of Justice Statistics or the UK Home Office (Coleman and Moynihan 1996). These agencies draw on the major criminal justice institutions such as the police, prisons, and the courts for these official statistics. On the surface the collection of such data might appear straightforward and relatively uncontentious yet in reality it remains one of the most controversial areas in the analysis of crime. Official crime statistics have been subject to critique since the mid-nineteenth century through to the present, the beginning of the twenty-first century (Coleman and Moynihan 1996).

There are a variety of factors that contribute toward this controversy (Kempf 1990, Jackson 1990). Some of the more important reasons include:

(a) The study of crime itself is often controversial.
(b) What is defined as a 'crime' is a social construction and not an empirical fact.
(c) There is not always consensus on how to collect crime statistics.
(d) There is no one data source that can address all the issues raised in the study of crime.

2. *Official Crime Statistics*

Administrative databases are used to generate official crime statistics yet they are first and foremost a management tool. As a result, using data generated from this system as the sole source of crime statistics is problematic. Unlike many other economic and social activities, the act of committing a crime can result in the loss of liberty and in some cases the loss of life. Understandably, most people do not admit to committing crime. As a consequence detection of crime is dependent on victims reporting such activities to the police, and the police successfully apprehending the suspects. Police make decisions about whether to proceed with charges, then prosecutors make decisions about whether to continue with the charge and if so which charges they will pursue. Finally, the courts then make a determination as to guilt or innocence. There are many factors that intervene at different points in this process ultimately affecting the 'numbers' that appear in official statistical collections. Because of the potential for error at various stages through the recording process official homicide statistics are largely regarded as having the most credibility.

3. *Police Data*

Counting rules used by the collection agencies complicates interpretation of police statistics. Data can be confusing to the novice as it can on occasion refer to the offences/incidents and at other times to the offenders and victims. For example, one offender may be responsible for 50 offences. Similarly, five people may be involved in one offence (e.g., a bank robbery by a group of people). Where an individual has committed a range of offences often only the most serious offence is recorded for statistical purposes. For example, prison statistics in Australia are based on the most serious offence of the prisoner. Thus if an individual is convicted on a range of offences on a particular day only the most serious offence is recorded. As a result official statistics can under-report the extent of crime in the community. Police data are limited by:

(a) Definitions of what constitutes a crime.
(b) Variability in reporting by citizens.
(c) Law enforcement organizational structures.
(d) Variations in counting rules across jurisdictions.
(e) Low rates of detection and clearance.

4. *Court and Corrections Data*

These particular administrative databases represent a subset of police data for they record crimes that have been processed further down the criminal justice chain. For example only one in 20 people who appear before the criminal courts in Australia serve time in prison. Data from these systems are used to produce crime statistics on matters relating to sentencing, prison composition, recidivism, probation, and parole. However, as with police databases they are highly restrictive in terms of the limited nature of the data recorded. Usually only the most basic demographic and arrest characteristics are collected. Counting rules also vary across the courts. In Australia, for example, the unit of count is the charges or appearances in the lower court while in the higher court it is distinct persons. Crime statistics from these databases will be biased as they are based on those who represent the most serious offenders and offences.

5. *Improvements in Official Statistics*

Despite these limitations there have been significant advances in the recording of crime within administrative systems. Many countries have moved or are moving toward collating national data on crime that is comparable across areas by standardizing definitions and counting rules (UNAFEI and AIC 1990). In addition jurisdictions are seeking to link data across

criminal justice institutions so that a more complete picture of crime, offending, and offenders can be determined. Improved and cheaper technologies, and the development of statistical packages, have meant that analysts and researchers are demanding unit record files (URF) rather than aggregated statistics. As a result new administrative databases are being established to provide greater detail in response to research and policy demands. A major development has been the National Incident-Based Reporting System by the United States Federal Bureau of Investigation (Maxfield 1999).

There are three major factors that are driving improvements in crime statistics. The first is the demand for accountability by governments for the way in which scarce public money is spent. To justify their activities criminal justice institutions are increasingly relying on statistics to support their assertions and resource allocations. This results in both a greater awareness of statistics and a greater scrutiny of these same statistics. The second major factor has been the critical scrutiny of these data particularly by independent academic researchers. The third has been the development of technology that enables the cost-effective storage and retrieval of data.

6. *Crime Victim Surveys*

Researchers have long been concerned about a 'hidden figure' of crime that represents a large proportion of unrecorded crimes and criminals. Crime victims' surveys from a range of countries have shown that the majority of people do not, for a variety of reasons, report crimes to the police. There are other biases in the official statistics. White-collar crime is rarely recorded or dealt with under criminal law (Sutherland 1949). Some researchers argue that the important feature of crime data is the long-term trend in crime and not the overall amount of crime.

Others have argued that even the long-term trends are difficult to interpret—is crime really increasing or is it a reflection of other things? For example, changes in police practices and priorities, or changes in the interpretation and classification of particular offences can help to explain changes in crime statistics. There have also been changes in the police management/processing of some forms of crime, such as rape and domestic violence, that have impacted on the number of recorded crimes.

Given the sustained criticism of official statistics governments and academics have sought ways to improve upon their collections. To address the issue of under-reporting of crime to police, crime victim surveys are routinely undertaken in many countries. Individuals within households are randomly selected and asked about their experience of crime and victimization. From this source of information the nature and extent of crime has been shown to consistently be more prevalent than appears from 'official' statistics. However, such surveys have also been criticized on a number of grounds including (Koffman 1996, Coleman and Moynihan 1996, Skogan 1981):

(a) The definition of what constitutes a 'crime' in the survey instrument does not always match the legal definition in the official statistics making comparisons difficult.

(b) Particular subgroups are either more or less likely to self-report particular crimes affecting the reliability of the estimates. For example, those with higher levels of education are more likely to report assault, while women are less likely to report rape/domestic violence.

(c) Memory failure is a problem with respondents being more likely to recall incidents where the offender was not known to them than where the offender was known or related. There are also potential problems in terms of ignorance of events, deception, and 'forward and backward telescoping.'

(d) Methodological problems include who should be interviewed in the household and the phrasing and order of questions.

7. *Self-report Surveys*

In addition to victimization surveys there has been increasing use of self-report data from offenders. Dissatisfaction with official crime statistics has been a major factor in the development of offender surveys. This source of information has tended to be dominated by academics rather than government statistical collection agencies. One of the reasons is because these data need to be more carefully informed by theoretical models that can only come from an in-depth understanding of the substantive material. Self-reports are now regarded as the most common methodology for collecting data on crime statistics and large scale self-reported studies have taken place in a range of countries (see Junger-Tas and Marshall 1999).

Self-report data have generally confirmed that official data underestimates the level of crime in the community as described in official statistics. This particularly applies to victimless crime such as prostitution and illegal drug use. It is in the area of deviant behavior that self-reported crime statistics are superior to official statistics and these behaviors, if defined as illegal, are rarely reported to the police. Hence, self-report data are widely used to provide more accurate statistics on the prevalence and incidence of engagement in criminal activities by offenders. In addition self-report data have been used to:

(a) examine the correlates of offending;

(b) test theoretical models of criminal behavior; and

(c) evaluate intervention programs (Junger-Tas and Marshall 1999).

However, biases also apply to many self-report studies. Unless extraordinarily large sample sizes are obtained in population surveys the capacity to examine the correlates of offending or test theoretical models is limited by the highly skewed distribution of criminal behavior; this form of deviance is a rare event. For example, less than half a percent of the general population self-report having tried heroin in the past 12 months in household surveys in the USA and Australia.

To generate 100 people who have used heroin in the past 12 months would require a sample of 20,000 people. In addition most household surveys tend to under-represent those engaged in criminal activity; this is partly due to the sampling frame and partly due to the differential participation amongst particular groups (Weis 1986).

One solution to overcoming the problem of large sampling error is to use disproportionate stratified samples or to focus on specific groups such as incarcerated offenders. The latter will ensure that the sample has the desired characteristics but generalizability is severely limited.

Given the different methodologies employed to obtain crime statistics it is not surprising that the absolute rates vary between official and self-report data from offenders and victims. Consequently, researchers have advocated that relative rates should be used in making comparisons (see Junger-Tas and Marshall 1999). When this is done there is considerable consistency between the statistical sources from a range of difference countries (Junger-Tas et al. 1994).

8. Impediments to the Improvement of Crime Statistics

There are three major impediments to the improved development of crime statistics:

(a) the application of ethics/privacy/informed consent provisions;

(b) access to primary or unit record data; and

(c) the rising costs of undertaking major research studies.

The movement toward increased ethics (or human) oversight of projects can be problematic particularly in the area of crime. In practice the self-report methodology asks people to incriminate themselves. This poses real dilemmas for ethics committees, many of who are not trained in the discipline. As a result committees may be reluctant to give ethics approval. It also poses problems for researchers who in many countries and jurisdictions do not have protection under the law; they may be legally required to release information that was gained under assurances of confidentiality to the informant. Instigating informed consent forms, particularly when it requires a signature may conflict with assurances of confidentiality or anonymity thus affecting response rates. Junger-Tas and Marshall (1999) allude to the problem of informed consent on participation rates for delinquency studies. These issues militate against the study of crime and could well reduce the empirical study of crime and the collection of crime data outside of the official statistics.

The major improvements to crime statistics have largely occurred because of academic scrutiny of published statistics and secondary analysis of unit record data. The USA has had a long tradition of sharing and releasing statistical data for secondary analysis; this is not the norm. Access to official data collections for secondary analysis is often difficult and highly controlled by state institutions in many countries, like Australia. Until state institutions release full unit record data for secondary analysis this situation will continue to impede the improvement of national crime statistics.

Finally, funding constraints balanced alongside the increasing costs of data collection and research will affect both the government and nongovernment sectors in their ability to improve on their statistical collections. This is compounded by improvements in statistical methods that allow social scientists to undertake more sophisticated analysis. But this entails more sophisticated statistical collections and these are usually much more expensive.

Bibliography

Coleman C, Moynihan J 1996 *Understanding Crime Data*. Open University Press, Buckingham, UK

Jackson P 1990 Sources of data. In: Kempf K L (ed.) *Measurement Issues in Criminology*. Springer-Verlag, New York

Junger-Tas J, Marshall I H 1999 The self-report methodology in crime research. In: Tonry M (ed.) *Crime and Justice, A Review of Research*. University of Chicago Press, Chicago, Vol. 25, pp. 291–367

Junger-Tas J, Terlouw G-J, Klein M W (eds.) 1994 *Delinquent Behaviour Among Young People in the Western World*. Kugler, New York

Kempf K L (ed.) 1990 *Measurement Issues in Criminology*. Springer-Verlag, New York

Koffman L 1996 *Crime Surveys and Victims of Crime*. University of Wales Press, Cardiff, UK

Maxfield M G 1999 The national incident-based reporting system: research and policy applications. *Journal of Quantitative Criminology* **15**: 119–49

Skogan W G 1981 *Issues in the Management of Victimization*. US Department of Justice: Bureau of Justice Statistics

Sutherland E H 1949 *White Collar Crime*. Dryden, New York

UNAFEI and AIC 1990 *Crime and Justice in Asia and the Pacific*. Australian Institute of Criminology, Canberra, Australia

Weis J 1986 Methodological issues in criminal career research, Vol. 1. In: Blumstein A, Cohen J, Roth J A, Visher C (eds.) *Criminal Careers and 'Career Criminals.'* National Academy Press, Washington, DC

T. Makkai

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Statistical Systems: Economic

'Economic statistics' in the widest sense of the term refers to those statistics the purpose of which is to describe the behavior of households, businesses, public sector enterprises, and government in the face of changes in the economic environment. Accordingly, economic statistics are about incomes and prices, production and distribution, industry and commerce, finance and international trade, capital assets and labor, technology, and research and development. Nowadays it is customary to include under economic statistics estimates of the value and quantity of stocks of natural assets and measurements of the state of the physical environment.

1. *Economic Statistics and Official Statistics*

In its strict sense the term economic statistics has a narrower scope as a distinction is drawn between economic statistics and economic accounts (as in the national or international accounts). Some statistics occupy a gray area—between social and economic statistics. They comprise measurements that play a role in describing social conditions and are used intensively in the making of social policies but at the same time relate intimately to the economic fabric of society. Examples are the statistics on income distribution commonly used for the measurement of poverty, the occupational make-up of the labor force, the expenditure patterns of households, and the structure of their income. For purposes of the present article the term 'economic statistics' excludes the national accounts but includes the statistics that are simultaneously used for economic and social purposes. Most of the references in this article apply to official statistics, i.e., those statistics that are either produced by government statistical agencies or by other bodies but are legitimized by being used by government. The usage adopted corresponds to the way in which government statistical offices are commonly organized.

2. *Directly Measured Statistics and Derived Statistics*

The term economic statistics is usually applied to statistics derived from direct measurement, i.e., either from surveys addressed to economic agents or from official administrative procedures that yield numerical data capable of being transformed into statistics. For example, the statistic on the national output of cement is usually derived by directly asking all (or a sample of) cement producers how much cement they produced over a defined period. The statistic on how many jobs were filled during a given period may be derived from the social security forms that employers are requested to send to the social security administration from time to time. This is different from the way in which the economic accounts of the nation are derived.

These are formed by combining all the existing economic (or basic) statistics through an interrelated system of accounting identities and from those inferring identity values that may not be available directly. They are also a means of assessing the strengths and weaknesses of economic statistics by noting how close they come into balance when the system of identities is applied to them. In other words, economic statistics are the raw material on the basis of which the economic or national accounts are compiled. The compilation of the national accounts in addition to summarizing and systematizing economic statistics plays the important role of showing up gaps in the range of economic statistics required to describe the functioning of the economic system (Table 1).

2.1 *Economic Statistics and Business Statistics*

The kernel of economic statistics is formed by statistics related to business and derived from questionnaires addressed to business, from questionnaires addressed to business employees, or from official records which businesses create in their dealings with government. Such statistics include descriptions of the business activity (what does the business produce, how does it produce it, for how much does it sell, what is the surplus generated by the sale once materials and labor are paid for, etc.) as well as of the nature of the factors engaged in production (structures and equipment used in production by type; labor employed by type of skill and occupation), and in addition statistics on the type of customers who purchase goods and services from businesses.

3. *How Economic Statistics Are Measured*

In spite of the increasing popularity of using administrative records as sources to generate economic statistics, on the whole, economic statistics are compiled mostly from surveys addressed to business. Generally speaking, the surveys consist of a list of respondents, a formal questionnaire—which can be used as basis for a personal interview (or an interview by telephone or Internet), a procedure followed when dealing with exceptional cases (not finding the right person to whom the questions should be addressed, finding that the business is not in the expected activity, or that the answers are incomprehensible or contradictory)—and a set of steps designed to aggregate individual answers so as to produce statistical totals, averages, or measures of distribution.

3.1 *Problems of Defining What Is a Business*

Part of the process of conducting a business survey consists in locating the addressees—either all the

Table 1
Types of economic statistics and sources (in italics) from which they are derived

Economic agent	Production	Distribution	Resources employed	Prices
Households	Household income/*surveys of household income*	Wages and salaries/*surveys of household income*, *labor force surveys*	Current population estimates, participation in labor force, employment/ *labor force surveys*	Consumer prices and consumer price index/ *surveys of prices*
Business	Value and amount of production; production by type of activity and by goods and services produced, orders and deliveries/*censuses and surveys of business activity*	Value and amount of sales; stocks of finished goods, materials and goods in process; sales by activity of seller and by product sold; sales by type of customer (exports and imports)/ *censuses and surveys of business activity; customs records*	Numbers employed by occupation; remuneration of labor employed; capital assets used; depreciation; profits; loans and equity borrowing/ *censuses and specialized business surveys; company balance sheets*	Producer prices, unit values in manufacturing, producer price index/*surveys of industrial prices, surveys of service activities, current agricultural surveys, customs records*
Government (central and local, public sector corporations, etc.)	Government expenditures and revenues by function and by source/*public sector accounts*	Public sector sales of goods and services/ *public sector accounts*	Public sector employment; public sector wages and salaries/*public sector reports; budget and supporting accounts*	Prices subject to regulation/ *special price surveys*
Other (nonprofit organizations, labor unions, employers confederations, etc.)	Income and expenditure of nonprofit sector/ *special surveys*		Employment in nonprofit sector; wages in nonprofit sector/ *special surveys*	

addressees in the survey universe if the survey is of a census type or some of the addressees if it is a sample survey. In either case two issues must be settled beforehand: (a) that all the addressees conduct the activity they are supposed to and (b) that there is no one else who conducts the targeted activity and has been left out. A related question has to do with what is a business. The 'business' can be defined as the entity that is liable under the law to pay taxes; that has a management capable of deciding its scale of operations and how those operations should be financed; and that *inter alia* conducts the activity that is germane to the survey. An alternative definition is one that seeks that part of the business (and only that part) that engages in the target activity directly and excludes all other parts of the business. In either case it is important that the respondents to the survey understand the scope of the questions and are capable of answering them out of records available to the business's management.

3.2 Business Records

Businesses tend to keep all the records they require to ensure informed management in addition to what they have to keep in order to comply with tax authorities and any other regulatory bodies that affect them. Usually those records fall into five categories: tax related; social insurance; engineering records on the performance of physical production; personnel and

safety records; and environmental records if the productive activity is of sufficient importance and of a kind that may detrimentally affect the physical environment. The efficiency of statistical surveys is measured partly by the ability of the statistical agency to request that information, but mostly by that information which is readily available from the records the business tends to keep anyway.

3.3 Household Surveys

Usually the major surveys addressed to households include a survey of households and how their members relate to the labor force and a survey of family income and expenditure. The former is the basis upon which statistics on the percentage of the labor force that is unemployed are calculated. The latter usually produces two key statistics: income distribution and structure of household expenditure. The latter is used to determine the various proportions of expenditure, which in turn are used in the compilation of the consumer price index. The surveys about household behavior—variation in the structure of expenditure and savings according to income, location of residence, family size, and age distribution—also provide key social statistics.

3.4 Public Accounts and Administrative Records

There are at least two categories of records that make explicit survey taking redundant: reports on public accounts and administrative registers created as a result of regular and comprehensive administrative procedures. The two have different properties. The accounts of the public sector (and of para-public enterprises such as nationalized industries or public corporations) are usually made available to the public as a means of displaying transparency and accountability. If the statistical requirement is limited to the displaying of aggregate information, what is published is usually sufficient as a description of the income, expenditure, and balance sheet of all public sector entities. In addition, all operations relative to the collection of taxes generate comprehensive records, susceptible to replace statistics to the extent the latter target the same variables. For example, the collection of income tax generates statistics on the income structure of all households (or individuals) liable to pay tax. The collection of VAT-type taxes generates records on the amount of sales, the nature of the seller, and in many cases the broad economic sector to which the buyer belongs. Social security records tend to generate information on the nature of the worker from whom amounts for social insurance are withheld, their place of work, the hourly, daily, or weekly remuneration that serves as a basis for the estimation of the amount withheld, etc. And of course, the oldest and most universal administrative operation of all—that connected with the activity of customs administrations—generates comprehensive information on the value, composition, conveyance used, origin, and destination of merchandise that either enters or leaves the national territory. This information, also known as merchandise exports and import statistics, is to be found amongst the earliest collections of numerical data of relevance to the administration of a country. To the extent that these administrative records are accessible to the compilers of statistics they tend to constitute acceptable replacements for estimates derived from surveys or from censuses.

4. Economic Classifications and Conceptual Frameworks

Economic statistics—like any other kind of descriptive statistics—are usually presented within the framework of standard classifications. In addition to geographic classifications, the most popular of these classifications are the Standard Industrial Classification (the SIC, or in its international version, the ISIC) and the Standard Product Classification (the SPC or, in its international version, the Standard International Trade Classification or currently the Central Product Classification). The SIC refers to the activity (or production technique) used to produce a specified list of goods and services. Similar (or homogeneous) production techniques are grouped into industries. The overall result of these groupings is a classification that views economic activity from the supply side. The SPC groups goods and services by their intrinsic characteristics rather than by the way in which they were produced. It is a classification that views the results of economic activity from the demand side. The use of international classifications is vital to ensure intercountry comparability of economic statistics.

The history of international cooperation in economic statistics is inextricably bound with the history of the two major economic classifications. Thus, the United Nations Statistics Commission, at its first meeting immediately after the end of World War II, commissioned the Secretariat to produce a classification of economic activities and shortly afterwards requested the development of a classification of the goods that entered into international exchanges (imports and exports). The Committee on Statistics of the League of Nations had worked on an initial version of such a classification. Almost immediately after the war the Statistics Commission approved the International Standard Industrial Classification of Economic Activities (ISIC) and the Standard International Trade Classification (SITC). Today, both classifica-

tions are into their fourth version (ISIC Rev. 3 and SITC Rev. 3, respectively).

Currently, the economic statistics of the members of the European Union and of many central and eastern European countries are classified by the NACE (Nomenclature des Activités économiques dans les Communautés Européennes). The three North American countries have their own system of economic classifications (North American Industrial Classification System—NAICS). Most other countries use one or another of the various versions of the ISIC.

The classification of goods was taken over by the body created to co-ordinate the activities of customs administrations worldwide (the Customs Co-ordination Council) and ended up as the Harmonized Commodity Description and Coding System (HS). As such it has been adopted for international trade statistics by the overwhelming majority of countries. Indeed, a number of them use the same classification for purposes of presenting their commodity output in the context of their industrial statistics. The requirement to develop an international classification of services prompted—again at the request of the United Nations Statistical Commission—the development of the Central Product Classification (CPC) which is being increasingly used to compare statistics on international trade in services with the internal production of and commerce in services.

4.1 Other Classifications

These apply to a variety of economic statistics but they (with the possible exception of the International Standard Classification of Occupations (ISCO)) are neither as well developed as the ones on activity and product nor are they at this stage the subject of international agreement. There are several examples of such classifications. In the case of capital assets there is still the question of whether a classification of assets ought to be embedded in the standard classifications of goods or whether it should be *sui generis*. There is no suitable classification of intangible assets. In the realm of labor inputs, recent innovations in the contractual relationship between employers and employees have produced a variety of forms of employment, none of which has so far found its way into an accepted classification.

4.2 Conceptual Frameworks

The conceptual framework for economic statistics is a hybrid, somewhere between business accounting and the system of national accounts. But, because economic statistics are the prime ingredient for the compilation of the national accounts, an explicit reconciliation between the former and the latter is required. Current trends suggest that because of pressures to keep government generated paperwork down, economic statistics tend to follow the patterns of business records. Where required, supplementary information is collected so that the basic statistics may be reconciled to the requirements of the national accounts compilers. But typically the necessary adjustments are conducted within the statistical agency. This is usually the case for key balances such as profits.

5. The Evolution of Economic Statistics as a System

The traditional manner in which economic statistics were derived was by the periodic taking of an economic census. Indeed in earlier times the census of population was taken together with the economic census. More recently, the economic census is addressed at businesses and retrieves information on their activity, location, legal form of organization, and on how many and what kind of people they employ. Supplementary information is usually obtained through lengthier questionnaires addressed not to all businesses but rather to a selection, usually above a certain size but sometimes to a randomly selected fraction of all those in the scope of the census.

5.1 Censuses and Registers

Economic censuses have become steadily less popular, largely due to their cost, the time it takes to process them, and the fact that, because of rapid changes in the size, make up, and even activity of businesses, census information loses much of its value shortly after it is compiled (see *Statistical Systems: Censuses of Population*). Instead, there has been an equally steady increase in the popularity of registers of businesses, usually derived from administrative data and used for much the same purposes as censuses but updated far more frequently. Indeed, in a number of cases, the introduction of new administrative procedures connected with tax gathering has been managed bearing in mind possible requirements of importance to the government's statistical agency.

The lists of businesses derived from administrative or tax-gathering procedures, commonly referred to as 'business registers,' are used mostly in two modes: as sampling frames and as a basis for indicators of economic performance such as 'net business formation' (the difference between the numbers of businesses created and the number of businesses closed down in a given period of time). As sampling frames, business registers are invaluable to support *ad hoc* inquiries into the state of the economy, usually conducted by contacting a randomly chosen group of businesses and gauging the opinion of their chief executive officers.

There are several requirements that must be satisfied for the modern approach to the compilation of business statistics to work. In particular, the source of information for the register must be comprehensive; it must be kept up to date as a result of the administrative procedures on which it is based; and the statistical agency must have unrestricted access to the source. The benefits that accrue to a statistical office from a business register are that: (a) it can dispense with the task of conducting a periodic census; (b) it can conduct a coherent set of sample surveys and link the resulting information so as to form a composite picture of the economy and its performance; and (c) it can control the paper burden that it imposes on individual business respondents through its multiple surveys.

5.2 *Applications of Business Registers*

Business registers are maintained at two distinct levels. There is a list of all businesses through their registration with some government department (usually the tax authority). This list conveys no more than there is a possible business behind a tax account. The physical counterpart of the tax account is usually determined by direct contact once it is established that the business is more than just a legal entity. The relationships between businesses as legal entities and the establishments, mines, mills, warehouses, rail depots, etc. that comprise them are a source of interesting information, but they are also the most difficult part of the management of a business register.

6. *Current Economic Indicators*

Under the impulse of macroeconomic stabilization policies and the belief in policy makers' ability to fine tune effective demand, the early 1960s saw a worldwide push to develop current economic indicators of demand by sector. A great deal of the empirical work had been done in the US, spearheaded by the National Bureau of Economic Research and resulting in a classification of short term economic statistics according to whether they lead, coincided, or lagged behind the 'reference' economic cycle. Most attention was placed on the performance of manufacturing industry, particularly on the output of equipment for/of fixed investment as well as for export; on the construction of plant and of other nonresidential structures; and in general on leading indicators of capital formation. At the same time considerable effort was made to improve the reliability and the timeliness of the key variables for purposes of economic stabilization: the rate of unemployment and the change in the consumer price index. Lastly, the balances in the main markets of the economy were looked at through quarterly national accounts, which in turn demanded major developments in the quality and the quantity of basic economic statistics.

7. *Today's Array of Economic Statistics*

In today's array of economic statistics the following would be the expected components in the more advanced statistical agencies: (a) statistics on the structure of the economy derived from either an economic census or from administrative data complemented by a survey on structural data (type of activity, breakdown of production costs, capital stock etc.); (b) exports and imports of goods and services, with the goods part broken down by commodity, country of origin and destination; (c) producer and consumer prices and indexes thereof; (d) short term indicators of economic change such as order books, deliveries of manufacturing goods and stocks thereof; (e) monthly physical output of key commodities and an index of industrial production; (f) business balance sheets (income, profit and loss, financial) and employment and unemployment as well as household income and expenditure. In addition, a statistical version of the accounts of the public sector would be an expected component.

In order to support these statistics, statistical agencies would maintain an up-to-date version of a classification by economic activity, usually an internationally agreed classification with a national supplement, and a classification of goods and services. They would also have their own or an international version of the classification of occupational employment.

See also: Consumer Economics; Databases, Core: Demography and Registers; Databases, Core: Sociology; Economic Growth: Measurement; Economic History, Quantitative: United States; Economic Panel Data; Economics: Overview; Microdatabases: Economic; Quantification in the History of the Social Sciences; Statistical Systems: Education; Statistical Systems: Labor; Statistics, History of

Bibliography

Cox B et al. 1995 *Business Survey Methods*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York

ISI, Eurostat, BEA 1996 *Economic Statistics, Accuracy, Timeliness and Relevance*. US Department of Commerce, Washington, DC

Statistics Canada 1997 *North American Industry Classification System*. Statistics Canada, Ottawa

Turvey R 1989 *Consumer Price Indices, an ILO Manual*. International Labor Office, Geneva

United Nations 1990 International Standard Industrial Classification of All Economic Activities, *Statistical Papers Series M No. 4 Rev. 3*. United Nations, New York

United Nations 1991 A Model Survey of Computer Services, *Statistical Papers Series M No. 81*. United Nations, New York

United Nations 1993 System of National Accounts. United Nations, New York

United Nations 1997 Guidelines on Principles of a System of Price and Quantity Statistics, *Statistical papers Series M No. 59*. United Nations, New York
Voorburg Group 1997 *Papers and Proceedings*. United Nations Statistical Commission, New York

J. Ryten

Statistical Systems: Education

Educational activities of some kind occur in every human system such as the family, workplace, and governmental organizations such as schools and colleges. The statistical systems available to provide measurements of educational activities, such as individual learning and institutional support, are found in the institutions that administer those activities: schools and local, state, and federal government agencies. This article will describe the status of statistical information about the US educational system and educational attainment of individuals at the year 2000. It will discuss types of information collected by state and federal agencies, international agencies, and education associations. These agencies provide the basic statistical resources for many forms of social science analysis of educational practices and outcomes.

The educational statistical system was originally created to provide measurement of administrative capacity, such as the distribution of financial resources, rather than the capacity of individuals, such as what is learned by individuals. In recent years, national policies about education have shifted from providing access to resources to encouraging greater accountability by the school organizations themselves. Thus, the statistical system concentrated more attention between 1965 and 1999 on measures of student achievement and its correlates. The complete system of education statistics includes information about students, teachers, administrators, the general population, and finances.

1. *Organization of National and International Education Statistics*

For the United States, the constitutional power and responsibility for providing schooling for children resides in the laws of each state. Local school systems and colleges and universities collect some forms of statistical information about themselves to administer their own programs. Each school system and state education agency or college and university assembles basic facts about itself such as numbers of students enrolled, number of teaching staff, number of graduates, and amount of financial resources available. Since the nineteenth century, the federal government has been given authority to assemble aspects of school system statistics into national summaries at a central agency, such as the Department of Education, to provide for national trends in finance, enrollment, and completions of schooling and for monitoring the administration of programs in education. Standard definitions are assembled by the National Center for Education Statistics (NCES) of the US Department of Education to assist local agencies, states, and colleges and universities develop common statistical measures (NCES 1999c).

International organizations that provide central functions for the reporting of educational statistics include the Organization for Economic Co-operation and Development (OECD) and UNESCO. International standards were created in the 1970s by UNESCO to enable worldwide assembly of statistics of these basic enrollments, graduates, and financial accounts into a single report (UNESCO 1998). Statistics collected under these procedures were published annually in the UNESCO *Statistical Yearbook* until 1998 but may be discontinued after that year (UNESCO 1998). The ISCED definitions are used also by the OECD for the collection of educational statistics for its member countries which include Europe, Japan, Korea, Australia, New Zealand, Mexico, Turkey, and the United States (OECD 1999). OECD publishes an annual report, *Education at a Glance*, that provides comparable statistics across countries on demographic, financial resources, school participation, school environment, graduate output, student achievement, and labor market outcomes (OECD 1997).

The Office of Management and Budget (OMB) provides general oversight of the federal statistical agencies in the US government by evaluating their programs and establishing funding priorities. It also oversees and coordinates the administration's information and regulatory policies to develop better performance measures and to reduce burdens on the public. Survey instruments of the population proposed by federal agencies are evaluated by the OMB offices to insure that survey questions are not asked of the general public unnecessarily.

2. *US Department of Education*

In recognition that local schools and states needed to be able to monitor their own performance in enrollments in schools by comparing themselves with other states, a US Department of Education was originally created in 1867 for the purpose of collecting information on schools and teaching that would help the states establish efficient school systems. The most recent US Department of Education was created in 1979 by the Department of Education Organization Act (93 Stat. 668; 20 USC 3401). The responsibility for education statistics belongs to an agency within the Department of Education: The National Center for Education Statistics (NCES). The General Education

Provision act says, 'The purpose of the Center shall be to collect, and analyze, and disseminate statistics and other data related to education in the United States and in other nations' (20 U.S.C. 1221e-1, as amended). The programs of statistical data collection and reporting by NCES are described in their report *Programs and Plans* (NCES 1999c).

Congress allocates funds to NCES for the purpose of collecting, analyzing, reporting, and training persons to use educational statistics. These funds are then used in nationally competed contracts with companies that carry out survey collection and analysis functions for the government. The quality of the statistics are monitored by a staff of NCES to insure that unbiased collection procedures are followed and reported to all those who would use these statistical services. The agency conducts a few research studies of its own to improve, expand and modernize its techniques of statistical collection and reporting. Oversight of the agency is provided by a board of researchers from university settings and by other governmental agencies such as the Office of Management and Budget, which seeks to assure the public that data collection requests are not too heavily burdensome on the organizations that are required to provide the funds.

3. *The Census Bureau*

The first statistical information on schooling for the Nation was collected in the Census of 1840. The 1840 census required that 'the marshals and assistants should collect and return in statistical tables … all such information in relation to mines, agriculture, commerce, manufactures, and schools, as will exhibit a full view of the pursuits, industry, education, and resources of the country' (Wright 1900). This first report resulted in national statistics from school systems rather than from individuals and it resulted in a number of errors such that in some states more students were reported than children counted in the population. Beginning in 1850 individuals were asked whether they had 'attended school at any time during the year preceding the census' (Folger and Nam 1967 p. 2). Censuses during the twentieth century continued to ask individuals whether they were enrolled in school, but did not ask about grade of enrollment until 1940. The censuses since 1940 all contained grade of enrollment by demographic characteristics such as residence, sex, and race.

The census of 1840 also initiated a series of questions for adults on the ability to read and write. The reliability of this item was said to be weak, and to be reported for whites only. Approximately 1 in 5 persons in 1840 answered the census takers that they were illiterate (Folger and Nam 1967). A question on ability to read and write was continued in the censuses until 1930. But, by 1940 illiteracy rates had dropped to about 3 percent and adults were asked how many years of school they had completed (Folger and Nam 1967 p. 131). This question has continued until the census of 2000 which has modifications to expand the categories of persons with more than a high school education. A complete description of the actual census items used in measuring educational levels is contained in *200 Years of U.S. Census Taking* (US Bureau of the Census 1989).

The Census Bureau conducts monthly surveys of employment for the Bureau of Labor Statistics and includes items on educational attainment and school enrollment regularly so that analysis of changes in employment patterns, income, marriage, voting, and mobility might be examined for persons of different educational levels. The March Current Population Survey which collects information on income for all respondents contains questions about years of school completed. Statistics on these factors are reported in the annual series of reports on the Current Population Survey and are included in the public use tapes available to researchers. Occasionally, at the request of other agencies, the Census Bureau asks its respondents expanded questions about their education, such as their field of study (US Bureau of the Census 1999). The measure of educational levels was changed to type of degree rather than years of school completed in 1993.

4. *National Science Foundation and National Academy of Sciences*

The National Science Foundation monitors the production of scientists and engineers through the collection and analysis of surveys on doctorates and of financial support for basic research by government agencies as well as by private businesses. A summary of these statistical studies are reported every two years in a report titled, *Science & Engineering Indicators* (National Science Board 2000). NSF also provides survey data to researchers intended to study issues in academic science and engineering resources (such as the production, employment, and funding of scientists) in publicly available databases. Or by accessing the website (http://www.nsf.gov/sbe/srs/stats.htm) some of the statistical data sources can be accessed directly.

The National Academy of Sciences is a non-governmental organization established by President Lincoln to provide the Federal Government with the best advice of scholars throughout the country. It receives funding from federal agencies to conduct studies on specific issues, including the quality of statistical data collection activities of the government and the interpretation of statistical reports on student achievement and international comparisons. Thus, discussions and reports by the National Academy of Sciences indirectly influence the policies of the government in procedures for their statistical system. Reports

of the National Academy are publicly available through their website (http://www.nas.edu).

5. School Participation and Graduation

The number and percentage of the population attending schools of different kinds was a significant indicator of changes in institutional support for education during the nineteenth and twentieth centuries when the education system was expanding. In more recent years, a more significant indicator of educational levels is the completion of secondary school, of college, and field of study of graduates. Information about enrollments and graduates has been collected in household surveys, the census of population, and from reports of schools and colleges to the Commissioner of Education Statistics since 1840 and are summarized annually in the Digest of Education Statistics (NCES 1999a). The federal statistical agencies provide aggregate summaries of local, state, and household reports into a national summary.

Such descriptive statistics are of more interest to historian's analysis of changes in American institutions than to social science analysis of individual differences in progress through the education system. The study of personal choice within a school setting is conducted in a series of studies that follow a sample of students through school with paper and pencil questionnaires. Many such longitudinal studies are carried out by federal agencies in education, labor, and health for the purpose of analyzing consequences of student performance and choices made at an early age on later life chances (West et al. 1998). Many of these studies are conducted by joint participation of researchers and government agencies for the design and support of the analysis of significant issues in education and educational policy. Beginning in the1990s, most of this information is made available through agency websites.

6. Student Achievement

Every classroom in the United States evaluates student performance directly or indirectly. A national system of testing companies measuring student performance has grown since World War II to provide school teachers and administrators with professionally developed assessment instruments to evaluate the performance of whole schools compared with national averages or for individual student scores used for screening for college entry. Because of concerns that student performance as measured by college entrance exams appeared to be declining during the 1960s, but that the national estimates were not reliable, the Federal government created the National Assessment of Education Progress (NAEP) in 1966 as an independent measure of changes in student performance of the country. The National Assessment surveys and reporting of statistics has been supported and managed by NCES and results of the assessments for mathematics, science, reading, writing, and civics can be found in their annual publications. Establishing national standards for education performance and developing a methodology for conducting national assessments was made the responsibility of the National Assessment Governing Board by Congress in 1988 (see http://www.NAGBE.org). Planning and research to develop international comparisons of student performance was first initiated in 1958 and first conducted among 14 year-old students in mathematics in 1965 (Husen 1967). The countries that participate in these international comparisons formed an informal network of research institutes called the International Organization for the Evaluation of Educational Achievement (IEA). Several international studies of student performance in science, civics, writing, science, and mathematics were conducted between 1970 and 1999 (Medrich and Griffith 1992, OECD 2000). The US government, through the NCES and NSF, has provided support and monitoring of these studies because the results of the studies have received significant national publicity and have been effective in mobilizing local school districts to pay greater attention to student performance levels. The growth of the American form of testing in many other countries can be seen in the large number of countries that participated in the Third International Study of Mathematics and Science in 1995 when 50 countries participated in some aspects of the study and new countries were added to the assessment carried out in 1999. New international studies will be carried out after the year 2000 that use European definitions of the end of secondary schooling and measure student performance on skills and knowledge of mathematics, science, and reading literacy (OECD 2000).

Federal agencies conduct a number of studies that are intended to inform the general public about knowledge and attitudes held by adults. For example, the International Adult Literacy Survey was conducted in member countries of the OECD in 1994 to describe the literacy and numeracy of the adult populations in those countries (http://nces.ed.gov/pubs98/98053.pdf). A new survey of adults called the 'Life Skills Survey' is being planned for the year 2002 that will include national and international comparisons to identify and measure adult skills linked to social and economic success. The survey items in this survey will be based on a framework derived from American theories of intelligence and intends to create paper and pencil tests that would demonstrate the distribution of these skills in the population (NCES). This study and another study conducted for the National Science Foundation seeks to identify the distribution of knowledge and attitudes about science (science literacy) among adults in the United States and other countries (National Science Board 2000). Each of these surveys serves to

provide policymakers with information about the level of information held by adults.

Studies and reports about methods for assessing student performance are conducted by the National Academy of Sciences in its Board on Testing and Assessment and the Board on International Comparative Studies in Education. These committee members are selected from academic researchers who have written about student assessment procedures. Their discussions and reports raise questions in their meetings about the statistical quality of the intended and unintended affects of reporting of student assessment by statistical agencies and thus provide helpful guidance to federal policies about data collection.

7. *Teachers, Teaching, and Curriculum and Instruction*

A complete system of statistics for measuring changes in the US educational system would have to include aspects of the educational processes itself: teaching, curriculum, and instruction (Shavelson et al. 1989). Jeannie Oakes and Neil Carey say,

> Curriculum is the what of science and mathematics education. Curriculum in content—the topics, concepts, processes and skills that students learn in science and mathematics classes—but it includes more than that. It includes the depth to which students explore content; the way teachers organize, sequence, and present it; and the textbooks and materials schools use (Shavelson et al. 1989, p. 96).

International studies have helped introduce the methods of measuring the content of education systems. However, the statistical system has yet to put together all of the elements of this system in one place so that the operating events in schools can be compared and contrasted. Surveys are separately conducted about student achievement, teacher characteristics, curriculum, and teaching practices (National Science Board 2000).

Little consensus has been reached among education researchers about the most significant events in schools. Studies are conducted of pedagogy, curriculum content, and home environment (The Coleman Report 1966). The statistical system produces statistical studies that permit the analysis of social characteristics of students and parents as well as some of the classroom. But no single survey itself can summarize all of the elements essential to describing the processes and outcomes of the educational system.

8. *School Finance*

Differences in community capacity to support school systems results in large differences across the United States in revenues for public elementary and secondary schools. Since the differences may result from either differences in cost of living or in unequal spending, the collection and analysis of statistics on school finance of public elementary and secondary school systems is a major activity for NCES (NCES 1999c). In higher education, the methods that students and families use for payment of their tuition and living expenses while in college are a national policy concern and a subject of intense data collection and analysis (NCES 1999c). Surveys are conducted to determine what the effect of federal policy on the subsidies for tuition is to individuals and institutions.

The connection between educational level of individuals and personal income is a subject of data collection in surveys by the US Census Bureau where individuals are asked to report both their educational levels and income. In only a few instances has personal income been associated with measurement of specific educational experiences, such as field of study in college in Census Bureau surveys.

9. *Statistics and the Internet*

The results of education statistical studies are being disseminated more rapidly and thoroughly through the Internet than in previous years. All federal and many state education agencies are making available on their Websites statistical tables and analytical reports. Also, the databases themselves can be addressed over the Internet. In some instances, it is possible to request a specific table by electronic mail from a particular survey that generates the specific summaries. These changes in production and distribution of statistical materials have revolutionized the availability of statistics and makes their use much more likely to occur in many reports.

10. *Statistics and Public Policy*

School systems of all kinds have been undertaking reform of their practices of teaching and instruction for a number of years. Statistical measurement of the outcomes of schooling is important for the public discussion of the purposes and directions of these reforms. For example, the results of the Third International Mathematics and Science Study which reported that US students compared well with other countries in the 4th grade, less well at the 8th grade, and were much lower than other countries in the 12th grade attracted the attention of the President and many members of Congress. The production of statistical studies have affected deliberations about the importance of legislation regarding the improvement of education in the United States.

Bibliography

The Coleman Report 1966 *Equality of Educational Opportunity*. US Department of Health, Education, and Welfare, US Government Printing Office, Washington, DC

Folger J K, Nam C 1967 *Education of the American Population*. US Government Printing Office, Washington, DC
Husen T (ed.) 1967 *International Study of Achievement in Mathematics*. John Wiley, New York
Medrich E, Griffith J 1992 *International Mathematics and Science Assessments: What Have We Learned?* US Department of Education, NCES 92-011
National Science Board 2000 *Science and Engineering Indicators: 2000*. National Science Foundation, Arlington, VA, NSB 00-01
Office of Management and Budget 1999 *Statistical Programs of the US Government*: FY 2000
Organization for Economic Co-operation and Development 1997 *Education at a Glance*. OECD 2, rue Andre-Pascal, Paris
Organization for Economic Co-operation and Development 1999 *The OECD Data Base*. OECD 2, rue Andre-Pascal, Paris http://www.oecd.org/statlist.htm
Organization for Economic Co-operation and Development 2000 *Program for International Student Assessment*. http://www.pisa.oecd.org
Shavelson R J, McDonnell L, Oakes J (eds.) 1989 *Indicators for Monitoring Mathematics and Science Education: A Sourcebook*. RAND, Santa Monica, CA
UNESCO 1998 *UNESCO Statistical Yearbook 1998*. Unesco, 7 place de Fontenoy, 75352 Paris 07 SP
US Bureau of the Census 1989 *200 Years of U.S. Census Taking, Population and Housing Questions, 1790–1990*. US Government Printing Office, Washington, DC
US Bureau of the Census 1990 *1990 Census of Population, Social and Economic Characteristics, United States*. (1990 CP-2-1) Appendix B or http://www.census.gov/td/stf3/append_b.html
US Bureau of the Census 1999 *Education Statistics*. http://-www.census.gov/population/
US Department of Education, National Center for Education Statistics 1999a *The Digest of Education: 1999*. May 1999. NCES 1999-032
US Department of Education, National Center for Education Statistics 1999b *Classification Evaluation of the 1994–95 Common Core of Data*. January 8
US Department of Education, National Center for Education Statistics 1999c *Programs and Plans of the National Center for Education Statistics for 1999*. NCES, no. 1999-027
West K, Hauser R, Scanlan T (eds.) 1998 *Longitudinal Surveys of Children*. National Academy Press, Washington, DC
Wright C 1900 *History and Growth of the United States Census*. US Government Printing Office, Washington, DC

L. Suter

Statistical Systems: Health

Statistics is the science of collecting, analyzing, presenting, and interpreting data. A statistic is an item of quantitative data and so the term statistics can also mean a collection of quantitative data i.e., the results of collecting and analyzing data.

The term health statistics system can be used to describe any organized or structured mechanism that involves health statistics. It can therefore include systems that are used for collecting, analyzing, interpreting, and making available quantitative data in relation to health status or health services activity, as well as referring to the actual collections of quantitative health related data. Increasingly these statistics systems are computerized, and the increased availability and ease of access of information technology means that this trend towards computerization of health statistics systems is likely to continue in the future. However in many countries, paper based systems are still used for the collection of health statistics and, in some cases, the aggregation and dissemination of these health statistics are also not computerized. Therefore focusing solely on computerized health statistics systems, when looking for health information can mean that much valuable information is missed.

Decisions about health and health care, whether made by policy makers, by health professionals or by the public, are influenced by the information that is available. Much of this information is derived from official and other health statistics. Effective health care is therefore dependent on the process of gathering, analyzing, and disseminating timely and appropriate information. The quality and accessibility of the health statistics systems, from which this information is derived, can have a major influence on how health and health care systems are run and used.

When discussing statistical systems the terms data and information are often used interchangeably, but it is important to recognize the differences. Data are specific facts recorded in some way for possible future use, for example, the annual expenditure for a specific department in a hospital. Data becomes information when it is has meaning, usually by putting it in some relationship with other data. For example, relating expenditure for a hospital department to the services provided by that department and comparing this with the data for similar departments in other hospitals, provides information on how expensive it is to deliver the service in the hospital concerned compared to other hospitals. Health statistics are derived from data, and should provide information.

Another term that is increasingly used in relation to health information is knowledge. By putting information in the context of other information, including both quantitative data and textual information, deductions can be made that increase knowledge. Knowledge provides the basis for making decisions in a rational way. Knowledge, based on information, allows appropriate actions to be selected in order to achieve specific objectives. This is often referred to as evidence based decision-making. This is one of the main purposes of gathering data and producing health statistics.

1. Scope of Health Statistics

In some respects, the terms health information system and health statistics system can be used inter-

changeably. Where the focus is on the collection of the data, or on the statistical calculations that need to be carried out on the data to produce useful information, then the term health statistics may be used more often. However, if the focus is on the results of this data collection and analysis, and on making this information available, then the term health information is used more often. The use of telemedicine is expanding the scope of health care information, and so the definition of health information is growing to include not just health statistics but also multimedia audio, video, and high-resolution still images (Balch and Tichenor 1997).

The scope of what health statistics systems cover can be defined in several ways:

(a) by the type of data and information included;

(b) by where the data are produced;

(c) by where and how the information is held and made available; and

(d) by the purposes for which the information is used.

A narrow definition of health statistics could conclude that the purpose for collecting statistical information should be primarily to satisfy the needs of government. This would mean that health statistics systems would just cover those systems designed to produce official health statistics for government use. However, the current view in many countries, and of many of the producers of official statistics, is that the statistics are produced to serve not only government but also the wider community. The members of this wider community of health statistics users come from different organizational backgrounds, including legislative bodies, government departments, health service organizations, researchers, voluntary organizations, and patients, as well as the public in general. The health service organizations that need to use health statistics can also be wide ranging, depending on the health system in operation in the country concerned. They are likely to include both the public and private sector, and both providers of health care services and purchasing organizations.

This broader view of health statistics users is both driven by, and enabled by, the rapid spread of the use of the Internet and the World Wide Web. Familiarity with accessing information via the web has raised the expectations of both health professionals and the public in relation to the availability of health information. Increasingly producers of health statistics have started to respond to this by making the information they produce available in different formats and through different mechanisms to suit the differing needs of the user community.

The widespread use of the Internet has meant that this is now used as a common method of disseminating information about statistics, as well as providing access to the actual data in some cases. Access to statistics held on data archives, many of which include health statistics, is also possible via the web. Web-based approaches can also allow information from different systems to be brought together, either by making these available via a web site, or by using web based technologies to access information held on different systems.

The design and coverage of health statistics systems, described later in this article, therefore needs to be seen as being currently very much in a transitional stage. For many systems, developments are ongoing which will take advantage of the opportunities, and also meet the challenges, of the technology that is now available for bringing together information from different sources and for providing wider access to this information (Sandiford et al. 1992).

2. *Use of Health Statistics*

There are many different individuals and organizations involved in planning, managing, running, investigating, and using the health care system. Health statistics systems should ideally be designed to meet the information requirements of this health care community. Some of the main uses of health statistics are summarized in Sect. 2.1 to 2.9.

2.1 Information for Policy Makers and for the Management of Health Care Delivery

Policy makers at both national and local level need information to support decisions on planning and funding of services. This applies both to policy makers responsible for publicly funded services, and to organizations such as health insurance companies involved in the funding and provision of private health care services (Roos 1999, Roos et al. 1999). Health information systems can provide decision-makers with the capability to:

(a) Make critical comparisons across regions and subregions of residents' health status, socioeconomic risk characteristics, use of hospitals, nursing homes, and physicians.

(b) Carry out analyses, in relation to the population served, of demographic changes, expenditure patterns, and hospital performance.

(c) Carry out outcome research across hospitals and countries to support recommendations on effective treatments,

(d) Make comparisons of expenditure relative to outcomes to support purchasing decisions

(e) Carry out longitudinal research on the impact of health policy changes.

This information helps to answer questions such as what health care services do particular population groups need and, by comparing this need to the current service provision, which populations need more services and which need less? (See Slater 1999, Azubuike and Ehiri 1999, De Kadt 1989.)

In addition, health information systems are important support tools in the management and im-

provement of the delivery of health care services. Experience in several countries suggests that comparison of hospital-specific data influence hospital performance. If statistics on provider performance are made available this ensures not only prudent and cost-effective health care purchasing, but also it gives health care providers comparable information to enable them to improve the care they deliver.

2.2 Health Surveillance Systems

In health terms, surveillance is the close monitoring of the occurrence of selected health events in the population. This implies that the events to be monitored can be specified in advance. This may include setting up a notification system for the health event, to supplement data available from routine health data collection systems, for example in the case of infectious diseases such as tuberculosis. However, the events monitored can also include certain signs and symptoms which can be seen as 'indicator events', to give warning of unforeseen problems.

Surveillance methods have been developed as part of efforts to prevent or control specific health events. Surveillance systems are therefore dynamic, in contrast to archival health information, which can be used for epidemiological analysis. The events that can be controlled in surveillance systems include infectious diseases, maternal and child health, chronic non-infectious diseases, injuries, and the effects of exposure to occupational or environmental risk factors.

2.3 Types of Health Information

The ultimate aim of any statistical work in the health sector should be to assist in the maintenance and improvement of the health of the population and in the delivery of health care services. Health information can therefore be defined as any information that is needed to support these activities.

It is important to be able to access both health service and non-health service data when looking at health needs, since the determinants of health often lie outside the health service. Work in relation to prevention of illness and the promotion of good health therefore require access to data from a wide range of sources. In response to this requirement health statistics systems need to cover, or provide access to, information not only on all aspects of the health care systems, but also on factors that influence health and the availability and use of health care.

There are different models of health systems used in different countries, and these have an impact on the health statistics systems, both in terms of the data collected, and hence the information potentially available, and in terms of where the information is held and who has responsibility for it. There may also be variations in the organization of health care, hospitals and hospital information systems within a given national environment (Saltman and Figueras 1997). However, the differences in health system models usually relate mainly to funding mechanisms, rather than to the basic activities that need to be carried within the health sector. This means that the types of information held are usually similar, even if the organizations involved are different.

The types of information covered usually include:

(a) Health status
—Health of the population
—Vital statistics (births, deaths)
—Health surveillance—for both infectious and non infectious diseases

(b) Registers
—Patient registers e.g., persons insured, persons registered with a family doctor
—Disease registers e.g., cancer

(c) Health service resources
—Health care facilities
—Health care staff
—Other resources e.g., medical equipment, drugs

(d) Health service activity
—Hospital and other secondary care activity
—Primary care activity
—Financial information

(e) Explanatory factors
—Economic information
—Sociodemographic information
—Environmental information

(f) Other related data (often government or public sector statistics)
—Social care sector activity
—Police—crime figures, road traffic accidents
—Education

Some of the main types of information listed are discussed in more detail in later sections. In addition to the quantitative data listed in Sect. 2.3, the interpretation of information from health statistics systems often requires reference to textual information in relation to current policies and guidance, and also to research findings.

Some of the information is obtained using administrative data, which are routinely collected as part of the process of running and monitoring the health system. In addition information is available from censuses, either using data from national population censuses or from censuses of specific groups, e.g., census of patients in hospital at a particular time. Health statistics are also available from a wide range of survey data. These can be international, national, regional, or local surveys, and can include surveys that are part of a regular survey program as well as 'one off' surveys carried out for a specific research or policy analysis study.

The increased use of computerization in health statistics systems has already been mentioned. Computers can store large volumes of data and can process them quickly. This means that if data are entered on to

a computer at source it may not be necessary to summarize data before onward transmission. For example, every patient's record can be stored and sent rather than just counts of patients with particular diseases. Consequently converting data into information can be postponed until a specific issue raises the need for a particular piece of information. Much more flexibility is therefore available in how the data are aggregated to create information.

Also data held in electronic form is no longer confined to being in one place at a time. Several people can simultaneously be reading the same record in quite different locations. This creates possibilities for sharing information, which were never feasible with paper systems. However it also increases the risk of unauthorized access to data, and data confidentiality issues in relation to health statistics systems are discussed in a later section.

2.4 Health of the Population

Health monitoring requires measures of incidence, prevalence, and severity of illnesses, and the consequences of illness, classified by age, sex, socio-economic, and geographical characteristics of the population. There are no ideal measures, and several sources need to be used in combination, for example using both health surveys and information about patients in hospital. The varied sources of population health information provide different pieces of the overall health picture.

Ultimately the objective of health care expenditure is to improve the health of the population not for its own sake but so that people can live happier more prosperous lives. Gains in health status might be regarded as the ultimate measure of a health program's success. Composite measures of health status such as the DALY (Disability Adjusted Life Year) and the QALY (Quality Adjusted Life Year), although developed from different disciplinary backgrounds, both attempt to summarize the potential gains from medical interventions (Williams 1999, Jenkinson 1994). Whereas health economists developed QALYs, DALYs have their roots in epidemiology. As a consequence, the perspective of the first is based on patient assessment of pain and suffering while the second is disease driven with a medical assessment of the amount of suffering.

2.5 Vital Statistics—Mortality Data

The most absolute measure of ill health is death, and mortality statistics are one of the principal sources of information about the changing patterns of disease and the differences in these patterns between countries. Since mortality statistics are widely available death rates have traditionally been used as crude, but often effective, surrogates for more comprehensive measures of disease. This is often justified by the fact that death is unambiguous and the data collection is complete. However use of mortality data alone can provide misleading impressions of the burdens of disease in society. For example, in certain parts of the world, musculoskeletal disorders (including rheumatism and arthritis) are the most important current causes of limiting long-standing illness but are one of the least prevalent causes of death.

Death rates vary by age and by sex and so, if comparisons are to be made between populations with different age and sex composition then some adjustments are needed. A common approach is to use age standardized rates and ratios (often referred to as Standardized Mortality Ratios or SMRs). These are calculated by applying the rates found for a particular population to an agreed standard population. Standardization is also used when comparing other measures, such as hospitalization rates, which are likely to be affected by the age and sex structure of the population. Another measure commonly used when considering premature mortality or avoidable deaths is Potential Years of Life Lost (PYLL). This highlights the loss to society as a result of early deaths (for example, below age 65 or 75). The figure is the sum, over all persons dying from that cause, of the years that those persons would have lived had they survived to the stated age (e.g., 65).

2.6 Vital Statistics—Births Data

Although systems vary between countries, the systems from which birth and maternity data can be derived are likely to include civil registration of births, the notification of births, and hospital-based data collection systems. Birth registration is the oldest source of data about births. In most countries the primary reason for setting up systems for registering births, marriages, and deaths was to provide civil documents to record the events for legal purposes. To obtain a full picture of birth and maternity care, data also needs to be collected about facilities and staffing of the maternity services and about the socio-economic context in which childbirth occurs.

2.7 Patient Registers

Where health insurance systems are in operation, there will be registers for the population covered by the schemes. These registers are likely to include information such the individual's insurance status, name, date of birth, address, and identification number, plus possibly other socio-economic or medical status data, e.g., whether the person is a diabetic. Other data related to individuals would be the health insurance contributions made on their behalf from employers or through other mechanisms. Information on dependents covered by an employed person's contributions may also be included.

In many health systems patients will be registered with a family doctor or primary care center. These registers are likely to include some of the same basic patient information as for the health insurance registers, as well as some health status information. Records of preventive services such as immunizations carried out by family doctors may also be included.

2.8 Disease Registers

Separate registers are often kept, usually at local or regional level, for patients with specific diseases, and cancer registration is an important example of this. Since, in many parts of the world, cancer-related services are consuming ever-increasing health resources; the information from cancer registries is important in helping health care planners, researchers, and policy-makers formulate strategies to meet this challenge.

Cancer registration is the collection and classification of data on all cases of cancer occurring within a defined territory. The primary source of data on which cancer registration is based is the patient's clinical case notes. This holds true whether the information is derived from hospital departments, primary or community care, or the private sector.

Cancer registration is a worldwide process. The longest established cancer registries are in Europe and North America, but increasingly they are being developed in the countries of Asia, Africa, and South America. Attempts to record comprehensive information on the incidence and distribution of cancer have been made as part of individual censuses and studies since the last century, but it was not until the late 1940s that cancer registration really got underway. In 1959 the World Health Organization (WHO) produced recommendations for the establishment of cancer registries, and, in 1965 the International Agency for Research on Cancer (IARC) was established as a specialist research center of WHO.

2.9 Hospital Data

A wide variety of data is available from hospitals for the purpose of health service analysis and performance monitoring. Where computerized systems are not available, much of this is obtained as aggregate counts of patients, and of resources used, falling into predefined categories. However, where computerized health information systems are in operation, individual records of care can be accessed for each patient which contain wide ranging details of the diagnoses, the treatments delivered, and the resources used.

2.10 Non-health Service Data

As has already been mentioned, it is important to be able to access non-health service data when looking at health needs, and when comparing statistics on health service activity. In many countries there is also an increasing emphasis within the health sector on partnership working with other agencies, and the need to share information on common and related issues. Health and health care activities can be affected by activities in other sectors, and monitoring and planning of health services activity need to take this into account.

Non-health service data sources include population estimates, vital statistics, and census data. Vital statistics covering data on births and deaths have already been mentioned. Population estimates and census data are discussed elsewhere in this publication. Various deprivation indices, some of which are derived from Census data, are also used when analyzing and interpreting statistics relating to the health sector.

The links between housing and health have been well established. Some data on housing can usually be obtained from population censuses, and information about housing stock may also be available from local government statistical systems. Poor air quality is also considered to be detrimental to health. Once again some information may be available from government statistical systems supplemented by special studies. Other statistics relevant to health that may be available from government sources include information on education, income, unemployment, transport, and crime.

3. Standard Coding/Classification Systems

If health statistics are to be compared across different organizations, regions, and countries, it is important that common coding and classification systems are used throughout the health sector and that they are used consistently. The WHO International Classification of Diseases is a classification of specific medical conditions and groups of conditions which is determined by an internationally representative group of experts who advise the World Health Organization (WHO), which publishes periodic revisions. It was first introduced a hundred years ago and the version in current use worldwide is the tenth revision (ICD 10).

In addition to the use of international classifications such as ICD10, many countries and international communities have undertaken initiatives to standardize the health data they collect in order to facilitate comparisons across countries. An example of this is the European Union's program 'Community Action on Health Monitoring' that was adopted for a five-year period with effect from 1 January 1997 (for EU health information projects, see CORDIS website). Its stated objective was to contribute to the establishment of a Community health monitoring system that makes it possible to:

(a) Measure health status, trends and determinants throughout the Community.

(b) Facilitate the planning, monitoring and evaluation of Community programs and actions.

(c) Provide member states with appropriate health information to make comparisons and to support their national health policies.

(d) Develop methods and tools necessary for analysis and reporting on health status, the effect of policies on health and on trends and determinants in the field of health.

The program is structured in three pillars as follows:

(a) Pillar A deals with the establishment of Community health indicators and with conceptual and methodological work related to the process of making the data comparable and for identifying and developing suitable indicators;

(b) Pillar B deals with the development of a Community wide network for the sharing and transferring of health data between member states, the European Commission and international organizations;

(c) Pillar C deals with the development of methods and tools necessary for analysis and reporting and the support of analyses and reporting on health status, trends and determinants and on the effect of policies on health.

These activities are being carried out in close cooperation with the member states and also with institutions and organizations which are active in the field of health monitoring, in particular: the World Health Organization, the Organization for Economic Cooperation and Development, the International Labor Organization (OECD Health Data, WHO Health for All Database, WHO World Health Report 2000, WHO Health in Transition profiles).

4. *Managing Confidentiality*

One important aspect of health information, and therefore of the systems that relate to health information, is confidentiality. This is because the information held may relate to identifiable individuals. Many countries, as well as international communities such as the European Union, have strict laws requiring organizations, including government agencies, to register the types of personal information they hold, the purposes they use it for and who they share it with. The law gives people the rights to inspect the data which organizations hold about them and to challenge inaccuracy and mis-use. Information systems and data flows have to be designed to respect privacy with personal identifier data removed when it is not needed and security precautions in place when it is needed.

Some principles that commonly apply to collecting, holding, and using health data are that:

(a) Health authorities should justify the collection of personally identifiable information,

(b) Patients should be given basic information about data practices,

(c) Data should be held and used in accordance with fair information practices,

(d) Legally binding privacy and security assurances should attach to identifiable health information with significant penalties for breach of these assurances,

(e) Disclosure of data should be made only for purposes consistent with the original collection,

(f) Secondary uses beyond those originally intended by the data collector should be permitted only with informed consent.

5. *Recent Developments*

Some of the trends in technology affecting health information systems have already been described. Another feature of planned developments in systems is a move towards integrated, patient-centered computing. Currently information is often collected according to the activity carried out e.g., by hospital department, which can make it difficult to obtain a complete set of information for an individual patient, or group of patients. The patient centered approach means that the computer systems and files are designed around capturing information along the whole continuum of care for a patient (Nold 1997).

It is also expected that health statistics systems will move towards higher-speed data transmission, and greater network capacity, following national and international computing standards. This will allow more rapid access to health statistics, as well as facilitating the bringing together of information from different sources—leading to improved information and greater knowledge about population health and health care (Mendelson and Salinsky 1997).

See also: Health Behaviors; Health Care Delivery Services; Health Economics; Health Education and Health Promotion; Health Interventions: Community-based; Health Policy; Health Surveys

Bibliography

Azubuike M C, Ehiri J E 1999 Health information systems in developing countries: Benefits, problems, and prospects. *Journal of the Royal Society for the Promotion of Health* **119**(3): 180–4

Balch D C, Tichenor J M 1997 Telemedicine expanding the scope of health care information. *Journal of the American Medical Informatics Association* **4**(1): 1–5

CORDIS—website http://dbs.cordis.lu/

De Kadt E 1989 Making health policy management intersectoral: Issues of information analysis and use in less developed countries. *Social Science & Medicine* **29**(4): 503–14

ICD10 http://www.who.int/whosis/icd10

Jenkinson C (ed.) 1994 *Measuring Health and Medical Outcomes*. UCL Press, London

Mendelson D N, Salinsky E M 1997 Health information systems and the role of state government. *Health Affairs* **16**(3): 106–19

Nold E G 1997 Trends in health information systems technology. *American Journal of Health-System Pharmacy* **54**(3): 269–74
OECD Health Data—website http://www.oecd.org/els/health/
Roos N P 1999 Establishing a population data-based policy unit. *Medical Care* **37**(6 Suppl): JS15–26
Roos N P, Black C, Roos L L, Frohlich N, De Coster C, Mustard C, Brownell M D, Shanahan M, Fergusson P, Toll F, Carriere K C, Burchill C, Fransoo R, MacWilliam L, Bogdanovic B, Friesen D 1999 Managing health services: How the Population Health Information System (POPULIS) works for policy-makers. *Medical Care* **37**(6 Suppl): JS27–41
Saltman R B, Figueras, J 1997 *European Health Care Reform—Analysis of Current Strategies*. WHO Regional Publications, European Series, No 72
Sandiford P, Annett H, Cibulskis R 1992 What can information systems do for primary health care? An international perspective. *Social Science & Medicine* **34**(10): 1077–87
Slater C H 1999 Outcomes research and community health information systems. *Journal of Medical Systems* **23**(4): 335–647
Williams A 1999 Calculating the global burden of disease: Time for a strategic reappraisal? *Health Economics* **8**(1): 1–8
WHO Health for All Database—website for Health for All in the 21st Century, http://www.who.int/archives/hfa/
WHO Regional Office for Europe, European Health Observatory Health in Transition profiles, http://www.observatory.dk/frame_index_main.htm
WHO World Health Report 2000 http:/www.who.int/
World Bank Data, http://www.worldbank.org/

D. Leadbeter

Statistical Systems: Labor

Labor statistics represent a large body of information relevant to many of the social and economic issues of our time. They are used as indicators of how well the general economy is functioning and also as a basis for designing, implementing, and monitoring employment and social programs and policies. Labor statistics are generally part of a national statistical program and are interrelated with statistics in other fields, for example, with statistics on education, health, and production. International comparisons of labor statistics can help in assessing a country's economic performance relative to other countries (for example, by comparing unemployment rates and employment growth) and in evaluating the competitive position of a country in international trade (for example, by comparing productivity and labor cost trends).

1. *Topics and Classifications*

Article 1 of Labor Statistics Convention Number 160 of the International Labor Office (ILO) established in 1985 a minimum set of topics to be covered in official national labor statistics programs (ILO 2000). The list is as follows.

(a) Economically active population, employment, where relevant unemployment, and, where possible, visible underemployment.

(b) Structure and distribution of the economically active population, for detailed analysis and to serve as benchmark data.

(c) Average earnings and hours of work (hours actually worked or hours paid for) and, where appropriate, time rates of wages and normal hours of work.

(d) Wage structure and distribution.

(e) Labor cost.

(f) Consumer price indices.

(g) Household expenditure or, where appropriate, family expenditure and, where possible, household income or, where appropriate, family income.

(h) Occupational injuries and, as far as possible, occupational diseases.

(i) Industrial disputes.

ILO convention 160 recommends that each member state (almost all countries are members of the ILO) 'collect, compile, and publish basic labor statistics, which shall be progressively expanded in accordance with its resources to cover [these] subjects.' The USA and most other developed nations collect and publish data for all subjects listed in Article 1. On the other hand, developing nations often do not yet have the full array of labor statistics. The publication *Key Indicators of the Labour Market* (Johnson et al. 1999) provides information about the availability of many of the basic labor statistics in almost all countries of the world. Deficiencies in the availability and timeliness of labor statistics are particularly apparent in Africa and parts of Asia.

1.1 *National Classifications*

The presentation of labor statistics requires many distinctions or categories such as occupations and branch of economic activity. In order to facilitate data collection and analysis, these categories are organized into classifications. These are instruments that group together units of a similar kind, often in hierarchies, thus allowing for different degrees of detail to describe the characteristics in a systematic and simplified way. For example, in the USA, the Standard Occupational Classification and the Standard Industrial Classification have been the structures that define occupations and industries for many years. Such systems need periodic revision to keep up with the changing economy and to accommodate new industries and occupations. *Revising the Standard Occupational Classification System* (1999) provides an account of the 1998 revision to the US occupational classification as well as a brief history of the industrial classification system.

The North American Industry Classification System (NAICS) represents a complete restructuring of the

US industrial classification (1997). This system was prepared jointly because the North American Free Trade Agreement of 1994 increased the need for comparable statistics from the USA, Canada, and Mexico. Each country can individualize the new system to meet its own needs, as long as data can be aggregated to the standard NAICS categories.

1.2 Other International Classifications

International standard classifications serve to enhance cross-national comparisons. The International Standard Industrial Classification of all Economic Activities was established under the auspices of the United Nations, and the latest International Standard Classification of Occupations was established by the ILO's 14th International Conference of Labor Statisticians in 1987. Summaries of these classifications appear in the appendices to the ILO's *Yearbook of Labour Statistics* (1999). Countries often prepare concordances or 'crosswalks' that allow their national classifications to mesh with the international classifications.

1.3 International Concepts

The ILO's periodic Conferences of Labor Statisticians have established guidelines for the measurement of many aspects of labor statistics (ILO 2000). These guidelines serve as a model for countries trying to develop national statistics on the topic concerned and they also provide, to some extent, a benchmark for international comparisons. For example, ILO standards for measuring labor force, employment, and unemployment are followed by many countries. Varying interpretations of these guidelines, however, particularly with regard to different aspects of unemployment, create differences across countries that require adjustments to enhance comparability of unemployment rates (Sorrentino 2000).

2. Institutional Framework

Labor statistics are primarily collected by governments rather than the private sector, although in some countries there has been a movement toward privatizing some data collections. In the USA, which has a decentralized statistical system, over 70 Federal agencies are engaged in the collection of statistics. The Bureau of Labor Statistics (BLS) in the US Department of Labor is the principal data-gathering agency of the US Federal Government in the broad field of labor economics. Most of the Bureau's data come from voluntary responses to surveys of businesses or households conducted by BLS staff, by the Bureau of the Census (under contract to BLS), or in conjunction with cooperating State and Federal agencies.

Unlike the decentralized system of the USA, other countries tend to have central statistical offices, and the collection of labor statistics is accomplished within these national agencies. However, labor ministries often also play a role in the collection of labor statistics abroad.

Several international organizations work in the realm of labor statistics, obtaining and publishing labor information from countries, providing technical assistance to member countries, and establishing conceptual guidelines. National agencies cooperate with the international organizations by transmitting the data requested on a regular basis and also for *ad hoc* studies. Major international organizations involved in labor statistics are discussed in Sect. 6.

3. Sources of Labor Statistics

Labor statistics are generally collected in censuses or sample surveys of households or business establishments, or processed on the basis of data resulting from administrative procedures of public agencies. Many of the major programs of the US BLS are based on sample surveys. For instance, the Current Population Survey (CPS) is a monthly survey based on a sample of about 50,000 households; data are collected by computer-assisted personal and telephone interviews. The survey is conducted by the Bureau of the Census for BLS and is the source of national statistics on the labor force, employment, and unemployment. About 70 countries conduct regular labor force surveys, on an annual, quarterly, or monthly basis.

Another major survey of BLS is the monthly Current Employment Statistics Survey. This survey is a sample of almost 400,000 establishments conducted by State employment security agencies in cooperation with BLS, and it provides employment, hours, and earnings data collected from payroll records of business establishments. Similar surveys are conducted by many other countries.

Labor statistics may also be based on administrative records. Although the primary purpose of administrative records is not the production of statistics, these systems provide a rich information source that has been exploited to varying degrees in different countries. Among the uses in the USA, administrative data from the Unemployment Insurance system are used in combination with CPS data to produce unemployment data for States, labor market areas, counties, and cities. In several European countries, employment office registrations are the primary monthly indicator of the labor market. This is likely to be the case in countries that do not have a monthly household survey as a source for unemployment statistics.

Some labor statistics are constructed by combining data from different sources. For example, industry

labor productivity data (output per employee or per hour) require data on both labor input and business output. The labor input measures are generally constructed from a combination of hours data from labor force and establishment surveys, while the output components typically are based on national income and product accounts which are derived from business establishment surveys.

4. Dissemination of Labor Statistics

Disseminating statistics means presenting and distributing the data to users in various forms and preparing analyses of the data. The BLS makes available the information it produces through a broad publication program that includes news releases, periodicals, reports, and bulletins. Articles in the *Monthly Labor Review* present analyses of labor data. Some BLS material is available on magnetic tape, computer diskette, CD-ROM, and microfiche. Current BLS data, publications, and articles are also made available through the Internet. Other countries use similar methods of data dissemination.

The publications *Major Programs of the Bureau of Labor Statistics* (1998) and the *Handbook of Methods* (1997), as well as the BLS web site noted in the bibliography, provide further information on the Bureau's programs and how to obtain BLS data. Links to the statistical offices of other countries as well as to international organizations are included on the BLS web site.

In 1996, the International Monetary Fund established a Dissemination Standards Bulletin Board on a web site (IMF DSBB) to assist countries in following the best practices of dissemination of economic and financial data to the public. One of the aspects of this, the General Data Dissemination System (GDDS), guides countries in the provision of comprehensive, timely, accessible, and reliable data, and the site provides information on data produced and disseminated by participating member countries. The GDDS system is built around four dimensions—data characteristics, quality, access, and integrity. Information on the formats and titles of national statistical publications is included. The GDDS is a source of metadata (that is, descriptions of current practices on data concepts, sources, and methods) on labor statistics as well as on other types of economic statistics. The detailed metadata provide users with a tool to better assess the usefulness of the data for their own particular purposes.

5. International Comparisons

An international perspective can often enrich analysis of national labor markets, but such comparisons must be made with caution because statistical concepts and methods vary from country to country. Simply comparing national data across countries could lead to erroneous conclusions. The BLS and several international organizations derive meaningful comparisons by selecting a conceptual framework for comparative purposes; analyzing foreign statistical series and selecting those that most nearly match the desired concepts; and adjusting statistical series, where necessary and feasible, for greater inter-country comparability.

The BLS comparisons generally are based on US concepts so that they will be on the conceptual basis most familiar to US data users. Comparisons made by international organizations frequently are based on the international standards recommended by the ILO, discussed in section 2.3. Brief descriptions of the comparative programs of the BLS, three international organizations, and a unique international micro-database follow. Further information is available from their respective web sites, cited in the bibliography.

5.1 Bureau of Labor Statistics

From its inception, BLS has published comparative information on labor conditions and developments abroad. Foreign labor research and statistical analyses have been undertaken because: (a) comparisons between US and foreign labor conditions shed light on US economic performance relative to other nations; (b) comparisons provide information on the competitive position of the USA in foreign trade, which has an important influence on the US economy and employment; (c) information on labor conditions published by a majority of foreign countries is not readily available to US labor representatives, employers, Government officials, and others, and frequently is not available in English; and (d) often, only an expert can judge the quality and comparability of foreign statistical data.

The BLS comparisons program currently covers three key areas: (a) labor force, employment, and unemployment; (b) manufacturing productivity and labor costs; and (c) manufacturing hourly compensation costs. Only major developed countries are covered in the labor force and productivity comparisons, whereas the hourly compensation comparisons cover 28 countries at various stages of development. Outside the USA, national statistical agencies generally do not produce their own international comparisons. They depend on the comparative work of the BLS as well as that of international bodies such as those described below.

5.2 International Labor Office

The ILO in Geneva plays a major role in the arena of international labor statistics. This agency's role has already been mentioned in terms of promoting the

production of labor statistics and setting international standards for statistics. The worldwide scope of the ILO's membership ensures that this organization provides the most comprehensive geographic coverage among the international organizations involved in labor statistics.

The ILO has several publications or programs that provide varying degrees of comparable labor statistics. The annual *Yearbook of Labour Statistics* (ILO 1999) is the most comprehensive source for worldwide labor statistics. The *Yearbook* compiles national labor data according to international classifications of industry and occupation, with annotations concerning deviations from these classifications, coverage, and sources. However, no adjustments are made for conceptual differences in labor force measures. All data are supplied by the national statistical authorities. The *Yearbook* contains data on labor force, employment, unemployment, hours of work, wages, labor cost, consumer prices, occupational injuries, and strikes and lockouts. Some statistics are presented as time series, others are provided for only the latest year.

The ILO's program of ILO-Comparable labor force statistics takes the process a step further by making adjustments of the labor force and unemployment data to ILO concepts for over 30 countries. Data from the ILO-Comparable program are available online (Lawrence 1999) and are also used in the publication *Key Indicators of the Labour Market* (Johnson et al 1999). The latter publication is well annotated concerning comparability issues for all of the 18 indicators provided.

5.3 Organization for Economic Cooperation and Development (OECD)

The Paris-based OECD comprises 29 member countries, most of which are the advanced countries in North America and Western Europe, as well as Australia, New Zealand, and Japan. In addition, some of the transition countries of Eastern and Central Europe, Mexico, and Korea became members of the OECD in the 1990s. Most OECD labor statistics cover the member countries only, but some series are being expanded to cover non-member Asian and Latin American countries.

The main labor-related publications of the OECD are the annual and quarterly editions of *Labour Force Statistics* (1999). The annual publication compiles two decades of statistics, while the quarterly editions show selected data for the past five years. These publications provide statistics on key elements of the labor force, rather than on the broader range of labor statistics contained in the ILO *Yearbook*. Through the OECD's provision of extended time series, it is possible to identify structural changes that have taken place in the labor force of the countries with respect to gender, age, and economic activities. Data on part-time employment and duration of unemployment are included. The detailed statistics on employment are generally in conformity with the International Standard Industrial Classification.

Another program of the OECD, the Standardized Unemployment Rates, provides unemployment rates adjusted to ILO concepts for 24 member countries. The Standardized Rates are published in an annex to the aforementioned quarterly editions and they also are available from the OECD web site.

Additional labor statistics are published in the annual *OECD Employment Outlook*, expanding the range considerably to include special international analyses of training statistics, indicators of the degree of employment protection and regulation, trends in self employment, income distribution, minimum wages and poverty, the transition from education to the labor market, trends in working hours, and workforce aging, to name some of the special studies. The OECD analyses pay careful attention to comparability and often contain annexes that explain the data comparison issues in detail.

5.4 Statistical Office of the European Communities (Eurostat)

Eurostat's main role is to process and publish comparable statistical information for the 15 member countries of the European Union. Eurostat tries to arrive at a common statistical language that embraces concepts, methods, and statistical and technical standards. A significant portion of Eurostat's statistics relates to labor. The *Labor Force Survey* (1998) is a joint effort by member states to coordinate their national surveys. The result is a detailed set of comparative data on employment, working time, and unemployment. Eurostat also publishes monthly harmonized unemployment rates for all member countries. Comparable data are critically important for the European Union because the Community's social funds are distributed according to a formula that includes unemployment as a significant component.

Eurostat's Harmonized Indices of Consumer Prices (HICPs) are designed for international comparisons of consumer price inflation. The HICPs are the result of the collaboration between Eurostat and national statistical offices of member states on harmonizing the different methods used to compile price indices. The harmonization process has resulted in an improved quality of the indices. HICPs are not intended to replace national consumer price indices (CPIs).

Besides the labor force survey and CPI statistics, Eurostat also publishes results from its harmonized labor cost surveys and conducts a European household panel that allows for analysis of income, employment, social exclusion, health, education, and many other topics.

5.5 Luxembourg Income Study (LIS)

The LIS, begun in 1983, is a unique cooperative international research project with a membership that includes 25 countries on four continents. The project is funded mainly by the national science and social science research foundations of its member countries. LIS comprises a collection of household income survey sets of microdata (files of data on individual households) for cross-country research. A companion database established in 1993 is the Luxembourg Employment Study, or LES. The data sets of LIS and LES have been harmonized in order to facilitate comparative studies. LIS and LES have become a rich source for comparative research in the areas of poverty and income distribution as well as labor market behavior. Moreover, these databases foster interdisciplinary research in the fields of economics, sociology, and political science. Researchers in member countries may use the LIS and LES databases at no charge, after signing a confidentiality agreement. Copies of the studies based on these databases are available from the LIS web site.

Disclaimer

All views expressed are those of the author and do not necessarily reflect the views of the Bureau of Labor Statistics.

See also: Employment and Labor, Regulation of; International Research: Programs and Databases; International Trade: Geographic Aspects; Labor Organizations: International; Labor Supply; Statistical Systems: Censuses of Population; Statistical Systems: Economic; Statistical Systems: Education; Statistical Systems: Health

Bibliography

Eurostat (Statistical Office of the European Communities) web site: http://europa.eu.int/comm/eurostat/

1998 *Labour Force Survey*. Eurostat, Luxembourg. International Labor Office web site: http://www.ilo.org

Johnson L Sorrentino C et al. (eds.) 1999 *Key Indicators of the Labour Market*. ILO, Geneva, Switzerland. Web site: http://www.ilo.org/public/english/employment/strat/KILM/index.htm

ILO 1999 *Yearbook of Labour Statistics*. ILO, Geneva, Switzerland. The statistics of the *Yearbook* are available online at http://laborsta.ilo.org

ILO 2000 *Current International Recommendations on Labour Statistics*. ILO, Geneva, Switzerland. The ILO's conventions and recommendations are also available at the following web site: http://www.ilo.org/public/english/bureau/stat/standards/index/htm

International Monetary Fund Dissemination Standards Bulletin Board web site: http://dsbb.imf.org/

Lawrence S 1999 ILO-comparable annual employment and unemployment estimates. *ILO Bulletin of Labour Statistics* **3**: XI–XLIII. The ILO-comparable data are available at the following website: http://laborsta.ilo.org

Luxembourg Income Study (and Employment Study) web site: http://www.lis.ceps.lu/index.htm

Office of Management and Budget 1997 *North American Industry Classification System*. Bernan Press, Lanham, MD

Organization for Economic Cooperation and Development web site: http://www.oecd.org. For the Standardized Unemployment Rates: http//www.oecd.org/media/new-numbers

2000 *Employment Outlook*. OECD, Paris, France

1999 *Labour Force Statistics 1978–1998*. Also *Quarterly Labour Force Statistics*. OECD, Paris, France

US Department of Labor Bureau of Labor Statistics (BLS) web site: http://stats.bls.gov. For the BLS international comparisons series: http://stats/bls.gov/flshome.htm

1997 *BLS Handbook of Methods*. BLS Bulletin 2490, BLS, Washington, DC

1998 *Major Programs of the Bureau of Labor Statistics*. BLS Report 919. BLS, Washington, DC

1999 *Revising the Standard Occupational Classification System*. BLS Report 929, BLS, Washington, DC. This publication is available at the following web site: http://stats.bls.gov/pdf/socrpt929.pdf

Sorrentino C 2000 International unemployment rates: How comparable are they? *Monthly Labor Review* **6**: 3–20. Web site: http://www.stats.bls.gov/opub/MLR/2000/06/contents.htm

C. Sorrentino

Statistics as Legal Evidence

1. Introduction

The subject of statistics has sometimes been defined as decision-making in the face of uncertainty; from this perspective, the legal use of statistics is a natural rather than a surprising development. Indeed, probabilistic and statistical modes of reasoning and inference can be found in documents as old as the Talmud and the use of pseudo-quantitative degrees of proof is one element of medieval canon law. But all such attempts to introduce quantitative elements into the law remained necessarily informal until the rise of probability and statistics as formal mathematical disciplines.

Shortly after the birth of mathematical probability in the seventeenth century, mathematicians such as Leibniz and James Bernoulli examined the legal implications of the subject (the latter in his masterpiece, the *Ars conjectandi*) and, in the eighteenth and nineteenth centuries, analyses of the reliability of testimonial evidence became commonplace in the treatises of the day on mathematical probability. But it was only in the hands of the French mathematicians Condorcet, Laplace, and Poisson that models of jury deliberation were first proposed, analyzed, and fitted to data. The models of Condorcet and his successors became of renewed interest in the 1970s

when some United States jurisdictions considered permitting juries to either have fewer than 12 members, or non-unanimous verdicts. Subsequently, critics such as Comte and Bertrand in France and Mill and Venn in England questioned whether the subjective, qualitative, and individual could be analyzed by methods that were supposed to be objective, quantitative, and based on the collective; and in the end such academic discussions had little practical legal impact.

But in the twentieth century, and especially in the years following World War II, when the uses and applications of statistics found ubiquitous applications throughout science, commerce, government, and society, the law began to turn increasingly to the methods of statistical inference for assistance in a wide variety of fields. Such applications now range from sampling, proof of discrimination, and the determination of damages, to the establishment of paternity, human identification, and the prediction of future violent behavior. The applications that occur fall into two natural classes: in some cases, statistical methods are used to determine specific facts relevant to a particular case being decided (did this employer discriminate against that employee? did the defendant commit the offense charged?); in other cases, statistical methods are used are answer more general questions thought to be relevant to underlying law (does capital punishment deter homicides?).

2. *Sampling*

The use of sampling in civil and administrative cases represents a natural and early application of statistical methods. A company seeking a refund for overpaid taxes but wishing to avoid the time and expense of examining each and every sales receipt at issue, might instead draw a sample from the population of all receipts; a government regulatory agency, suspecting inappropriate behavior on the part of a company trading commodity futures might conduct an audit of a random selection of accounts, rather than examine and analyze potentially tens of thousands of such accounts; an insurance company engaged in an arbitrated dispute might agree to accept the results of a sample of policies drawn by a neutral third party. Such applications are natural and represent some of the earliest serious uses of statistics in the law precisely because the benefits were obvious and the methodology (if properly conducted) difficult to challenge.

3. *Random Selection*

A properly drawn scientific sample is accepted as representative because the randomness in question has been imposed by the statistician (using either a computer, random number table, pseudo-random number generator, or physical randomization). Sampling practitioners know however that true randomness is often difficult to achieve and that samples not drawn in accord with appropriate protocols may turn out to exhibit serious biases. Thus in some legal applications of statistics, whether or not a sample has been randomly drawn is itself the question of interest. Casinos reporting unusually low cash receipts during a year are sometimes challenged on the basis of a statistical analysis of possible variation in yearly revenue. Governments also resort to the use of random selection or lotteries, for example, in jury selection, the military draft, random audits of tax returns, and spot checks of baggage or individuals at airports, in ways that may be subjected to statistical scrutiny and legal challenge. One famous example of this is the 1968 United States draft lottery (which determined the order in which individuals were drafted into the army): individuals born later in the year were more likely to be drafted before individuals born earlier in the year. Sometimes the purpose of a government lottery is not just to achieve randomization, but also to ensure a popular perception of fairness. Aristotle, in his *Athenian Constitution*, describes a complex multistage procedure for juror selection that presumably satisfied both of these functions.

Even if the mechanism for selection of a jury or grand jury does not employ randomization, there may be a legal requirement that the results of whatever method of selection has been used should at least appear to be random. In any discussion of random selection, it is important to distinguish between the process used to generate an outcome, and the absence of pattern in the outcome. The process may employ a true element of physical randomness, yet the outcome might by chance display some form of pattern; while the output of some deterministic processes, such as a pseudo-random number generator, may appear to be pattern free. Thus in the United States, for example, juror pools are often based on samples of administrative convenience but may, as a matter of law, be required, however selected, to be a representative cross-section of the population. Jury verdicts are therefore sometimes challenged on the ground of possible discrimination in the selection of the jury, it being alleged that either minorities or other groups had been excluded or underrepresented. Thus, for example, in the trial of Dr. Benjamin Spock for advocating resistance to the draft, his attorneys (believing that women would be more favorably disposed towards their client) appealed his conviction in part on the ground that women had been disproportionately excluded during one stage in the jury selection process and based this contention on a statistical analysis.

4. *Employment Discrimination*

In scientific sampling, the model of random selection arises from the actions of the statistician and, in jury selection, the randomness of the outcome may be a

legal requirement. But in some instances the model of random selection is a benchmark imposed by the analyst. In employment discrimination cases, for example, it may be claimed that racial, ethnic, sexual, or age discrimination took place in the hiring, promotion, or termination of individuals. In the workplace, such discrimination, if it does take place, is seldom explicit today and resort is often made to statistical methods of proof to document the underrepresentation of a protected class.

In the United States, the law distinguishes two basic classes of employment discrimination: disparate treatment and disparate impact. In disparate treatment, intentional discrimination on the part of the employer is charged and statistical proof of underrepresentation may be introduced to support such a claim. In disparate impact cases, in contrast, the intent of the employer is not at issue: it is charged that, for whatever reason, the hiring, promotion, or termination practices of the employer impact to a greater extent on one or more protected classes, for example, Hispanics and African-Americans, than others. Rulings of the US Supreme Court in the 1970s found that, in certain circumstances, disparate impact was enough to sustain a finding of discrimination, unless the employer can defend the practice at issue on the grounds of 'business necessity'; for example, a promotional examination may have a disparate impact but test for a necessary job skill. Such rulings were of great importance: by creating a purely statistical category of employment discrimination, they resulted in a rapid assimilation of statistical methods into the law, including many areas other than employment discrimination, due in large measure to the imprimatur given them by the Supreme Court.

After statistical methods of proof of employment discrimination became routinely accepted in US courts, increasingly sophisticated methods began to be introduced. In cases of salary discrimination, for example, complex regression models are sometimes used in an attempt to argue that a minority group or other protected class is being underpaid even after the salary of the individual is adjusted for all variables (such as experience, skill, and credentials). Because both the hiring and promotion of individuals and the salaries they are paid are not the outcome of a lottery but the result of human choice and agency, the appropriateness of using as a yardstick the random selection model underlying all such statistical models has been questioned. One serious concern is that as the methods used become more complex and in consequence more removed from the intuition of the lay trier of fact, the potential for misunderstanding and misuse increases. One illustration of this is the not uncommon interpretation of a mere finding of statistical significance as being a direct proxy for one of substantive significance.

It is not always obvious at what level an analysis should take place in a discrimination case. Simpson's paradox refers to situations where a disparity between two groups is found at one level of analysis, but disappears at another. For example, in the 1960s, the Graduate School of the University of California at Berkeley discovered that a substantially higher percentage of men than women were being admitted for graduate studies. In an attempt to identify those departments responsible for the disparity, the Graduate School examined the records of individual departments and found that in each department women were doing at least as well as, and sometimes better than, their male counterparts. There is a simple explanation for the phenomenon in this case: women were applying preferentially to departments that had a lower admission rate, men were applying preferentially to departments that had a higher admission rate.

5. *Forensic Applications*

Much of forensic science is devoted to the identification of substances or individuals, and in both cases important statistical issues arise. Classical gunpowder residue, blood alcohol level, and drug tests, three examples of tests for the suspected presence of substances, must all take into account the statistical variability inherent in the testing process. In the case of human identification, additional issues arise. The determination of human identity has classically exhibited two extremes. In fingerprint analysis, it is usually accepted that an adequate fingerprint exhibiting enough characteristics suffices to uniquely identify its source. In classical serological testing, in contrast, determining the ABO type of a biological specimen (usually blood, saliva, or semen) merely suffices to narrow down the class of possible donors to perhaps 25 percent to 1 percent of the population. Such evidence, although probative, obviously cannot by itself establish identity. It is only recently, with the advent of sophisticated methods of DNA testing, that the potential for human identification has dramatically increased.

In a DNA profile, one or more locations (loci) on the human chromosome are examined and typed. Because human chromosomes come in pairs, two alleles are present for each locus and the list of allele pairs for each locus constitutes the DNA profile for the individual. Because the loci used in forensic applications are typically hypervariable (many possible alleles can occur at each locus), the resulting DNA profiles can be highly discriminating (few individuals in the population will have the same profile) and the typical profile may have a frequency in the population ranging from 1 in several thousand, to 1 in several million, billion, or more (depending on the system used). Such impressive frequency estimates necessarily make certain statistical or population genetic assumptions (independence of alleles within and between loci,

sometimes termed either 'gametic' and 'zygotic,' or 'Hardy-Weinberg' and 'linkage' equilibrium), assumptions whose validity must be tested and may vary substantially from one population to another. More sophisticated approaches estimate parameters that attempt to correct for the presence of possible subpopulations.

Initially the introduction of such frequency estimates was attended by considerable controversy, but the introduction of polymerase chain reaction (PCR) methods in recent years has resulted in the ability to type increasing numbers of loci. At first, the FBI laboratory only typed four restriction fragment length polymorphism (RFLP) loci but in the not too distant future, the resulting DNA profiles may well come to be viewed as essentially establishing identity.

Nevertheless, even in situations where the computation of a profile frequency is not regarded as being by itself controversial, a number of statistical complexities still arise:

(a) the calculation of profile frequencies requires knowledge of the underlying allele frequencies, and knowledge of these requires either a true random sample, or at least, if a sample of convenience is used, one thought to be representative of the population,

(b) profile frequencies can vary considerably from one racial or ethnic group to another and information for some population subgroups may not be readily available (for example, some American Indian populations),

(c) in some cases, especially sexual assaults, a mixture of two or more sources of DNA may be present and such complex mixture cases can present formidable difficulties in both interpretation and appropriate statistical analysis,

(d) if a DNA profile is identified by searching through a large databank of such profiles (for example, profiles of convicted felons), the statistical weight afforded to such a match needs to take this into account. (From a frequentist perspective, it is sometimes argued that this should be done by adjusting for the multiple comparisons being performed; from a Bayesian perspective, the adjustment would reflect the differing priors in a situation where a suspect is in hand and a suspectless crime resulting in a databank search.)

(e) matching DNA profiles are much more common among relatives than unrelated individuals in a population, the computation of profile frequencies for relatives leads naturally to another important application of DNA testing.

6. *Paternity Testing*

Closely related to its use in criminal forensics, DNA profiles are also used in establishing paternity. Suppose the profiles of a child, mother, and putative father have been determined. If the putative father is not excluded as a possible father, then two statistics are commonly encountered to measure how unusual such a failure to exclude is. One statistic, termed the 'random man not excluded' (RMNE), computes what fraction of the male population would similarly not be excluded if their DNA profiles were also tested. The second statistic, termed the 'paternity index,' computes the likelihood ratio for two hypotheses conditional on the observed evidence (the three profiles): the first hypothesis, that the putative father is indeed the father; the second, that the true father is a random member of the population. Such calculations were initially performed using knowledge of the ABO and other blood groups, but in the current era of DNA testing often suffice to effectively establish paternity.

There are a number of variant calculations that can also be performed. One can compute paternity indices if the profile of the mother is unknown or in more complex situations where the profiles of relatives other than the mother are known. Or one can attempt to identify the source of biological material when the profile of the individual is unavailable but the profiles of relatives are available. This is often the case in body identification, in particular when mass disasters are involved.

7. *Violence Prediction*

Society often has a stake in attempting to predict future violent behavior: are there adequate grounds for confining a mentally ill person on the grounds that they pose a significant risk if left at liberty; if a prison inmate is paroled, is it likely that they will commit another crime after release; if a sexual predator is released after their prison sentence is completed, how likely is it that they will commit another sexually violent offense after their release? Such predictions, especially in the last category, have traditionally suffered from the same difficulties as testing for a rare disease: even if a test for a disease has both high sensitivity and specificity, when the disease is rare enough, most of the positives are false positives. Despite such difficulties, in recent years some psychologists have attempted to construct predictive 'instruments' (tests) using complex logistic regression models, and have analyzed their performance using survival analysis techniques.

8. *Interpretation*

Statistical analyses may be performed by highly competent individuals but their results must be interpreted and used by persons lacking formal training or quantitative understanding. This presents challenges both in successful presentation and in meeting the important responsibilities placed on the statistician to

ensure that his results are not overvalued. Terms such as 'statistical significance' are easily and frequently misunderstood to imply a finding of practical significance; levels of significance are all too often interpreted as posterior probabilities, for example, in the guise that, if a DNA profile occurs only 1 in 10,000 times, then the chances are 9,999 to 1 that an individual having a matching profile is the source of the profile. As a result, individuals may tend to focus on the quantitative elements of a case, thereby overlooking qualitative elements that may in fact be more germane or relevant.

9. Other Applications

The last decades of the twentieth century saw a remarkable variety of legal applications of statistics: attempting to determine the possible deterrent effects of capital punishment, estimating damages in antitrust litigation, epidemiological questions of the type raised in the book and movie *A Civil Action*. Some of the books listed in the bibliography discuss a number of these.

See also: Juries; Legal Reasoning and Argumentation; Legal Reasoning Models; Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox); Sample Surveys: The Field

Bibliography

Bickel P, Hammel E, O'Connell J W 1975 Is there a sex bias in graduate admissions? Data from Berkeley. *Science* **187**: 398–404. Reprinted In: Fairley W B, Mosteller F (eds.) *Statistics and Public Policy*. Addison-Wesley, New York, pp. 113–30

Chaikan J, Chaikan M, Rhodes W 1994 Predicting Violent Behavior and Classifying Violent Offenders. In: *Understanding and Preventing Violence, Volume 4: Consequences and Control*. National Academy Press, Washington, DC, pp. 217–95

DeGroot M H, Fienberg S E, Kadane J B 1987 *Statistics and the Law*. Wiley, New York

Evett I W, Weir B S 1998 *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*

Federal Judicial Center 1994 *Reference Manual on Scientific Evidence*. McGraw Hill

Fienberg S E 1971 Randomization and social affairs: The 1970 draft lottery. *Science* **171**: 255–61

Fienberg S E (ed.) 1989 *The Evolving Role of Statistical Assessments as Evidence in the Courts*. Springer-Verlag, New York

Finkelstein M O, Levin B 1990 *Statistics for Lawyers*. Springer-Verlag, New York

Gastwirth J L 1988 *Statistical Reasoning in Law and Public Policy*, Vol. 1: *Statistical Modeling and Decision Science*; Vol. 2: *Tort Law, Evidence and Health*. Academic Press, Boston

Gastwirth J L 2000 *Statistical Science in the Courtroom*. Springer-Verlag, New York

Harr J 1996 *A Civil Action*. Vintage Books, New York

Lagakos S W, Wessen B J, Zelen M 1986 An analysis of contaminated well water and health effects in Woburn, Massachusetts. *Journal of the American Statistical Association* **81**: 583–96

Meier P, Zabell S 1980 Benjamin Peirce and the Howland Will. *Journal of the American Statistical Association* **75**: 497–506

Tribe L H 1971 Trial by mathematics. *Harvard Law Review* **84**: 1329–48

Zeisel H, Kaye D H 1997 *Prove It With Figures: Empirical Methods in Law and Litigation*. Springer-Verlag, New York

S. L. Zabell

Statistics, History of

The word 'statistics' today has several different meanings. For the public, and even for many people specializing in social studies, it designates numbers and measurements relating to the social world: population, gross national product, and unemployment, for instance. For academics in 'Statistics Departments,' however, it designates a branch of applied mathematics making it possible to build models in any area featuring large numbers, not necessarily dealing with society. History alone can explain this dual meaning. Statistics appeared at the beginning of the nineteenth century as meaning a 'quantified description of human–community characteristics.' It brought together two previous traditions: that of German '*Statistik*' and that of English *political arithmetic* (Lazarsfeld 1977).

When it originated in Germany in the seventeenth and eighteenth centuries, the *Statistik* of Hermann Conring (1606–81) and Gottfried Achenwall (1719–79) was a means for classifying the knowledge needed by kings. This 'science of the state' included history, law, political science, economics, and geography, that is, a major part of what later became the subjects of 'social studies,' but presented from the point of view of their utility for the state. These various forms of knowledge did not necessarily entail quantitative measurements: they were basically associated with the description of specific territories in all their aspects. This territorial connotation of the word 'statistics' would subsist for a long time in the nineteenth century.

Independently of the German *Statistik*, the English tradition of political arithmetic had developed methods for analyzing numbers and calculations, on the basis of parochial records of baptisms, marriages, and deaths. Such methods, originally developed by John Graunt (1620–74) in his work on 'bills of mortality,' were then systematized by William Petty (1627–87). They were used, among others, to assess the population of a

kingdom and to draw up the first forms of life insurance. They constitute the origin of modern demography.

Nineteenth-century 'statistics' were therefore a fairly informal combination of these two traditions: taxonomy and numbering. At the beginning of the twentieth century, it further became a mathematical method for the analysis of facts (social or not) involving large numbers and for inference on the basis of such collections of facts. This branch of mathematics is generally associated with the *probability theory*, developed in the seventeenth and eighteenth centuries, which, in the nineteenth century, influenced many branches of both the social and natural sciences (Gigerenzer et al. 1989). The word 'statistics' is still used in both ways today and both of these uses are still related, of course, insofar as quantitative social studies use, in varied proportions, inference tools provided by mathematical statistics. The probability calculus, on its part, grounds the credibility of statistical measurements resulting from surveys, together with random sampling and confidence intervals.

The diversity of meanings of the word 'statistics,' maintained until today, has been heightened by the massive development, beginning in the 1830s, of *bureaus of statistics*, administrative institutions distinct from universities, in charge of collecting, processing, and transmitting quantified information on the population, the economy, employment, living conditions, etc. Starting in the 1940s, these bureaus of statistics became important 'data' suppliers for empirical social studies, then in full growth. Their history is therefore an integral part of these sciences, especially in the second half of the twentieth century, during which the mathematical methods developed by university statisticians were increasingly used by the so-called 'official' statisticians. Consequently the 'history of statistics' is a crossroads history connecting very different fields, which are covered in other articles of this encyclopedia: general problems of the 'quantification' of social studies (Theodore Porter), mathematical 'sampling' methods (Stephen E. Fienberg and J. M. Tanur), 'survey research' (Martin Bulmer), 'demography' (Simon Szreter), 'econometrics,' etc. The history of these different fields was the object of much research in the 1980s and 1990s, some examples of which are indicated in the Bibliography. Its main interest is to underscore the increasingly closer connections between the so-called 'internal' dimensions (history of the technical tools), the 'external' ones (history of the institutions), and those related to the construction of social studies 'objects,' born from the interaction between the three foci constituted by university research, administrations in charge of 'social problems,' and bureaus of statistics. This 'co-construction' of objects makes it possible to join historiographies that not long ago were distinct. Three key moments of this history will be mentioned here: Adolphe Quetelet and the average man (1830s), Karl Pearson and correlation (1890s), and the establishment of large systems for the production and processing of statistics.

1. *Quetelet, the Average Man, and Moral Statistics*

The cognitive history of statistics can be presented as that of the tension and sliding between two foci: the *measurement of uncertainty* (Stigler 1986), resulting from the work of eighteenth-century astronomers and physicists, and the *reduction of diversity*, which will be taken up by social studies. Statistics is a way of taming chance (Hacking 1990) in two different ways: chance and uncertainty related to protocols of observation, chance and dispersion related to the diversity and the indetermination of the world itself. The Belgian astronomer and statistician Adolphe Quetelet (1796–1874) is the essential character in the transition between the world of 'uncertain measurement' of the probability proponents (Carl Friedrich Gauss, Pierre Simone de Laplace) and that of the 'regularities' resulting from diversity, thanks to his having transferred, around 1830, the concept of average from the natural sciences to the human sciences, through the construction of a new being, the average man.

As early as the eighteenth century, specificities appeared from observations in large numbers: drawing balls out of urns, gambling, successive measurements of the position of a star, sex ratios (male and female births), or the mortality resulting from preventive smallpox inoculation, for instance. The radical innovation of this century was to connect these very different phenomena thanks to the common perspective provided by the 'law of large numbers' formulated by Jacques Bernoulli in 1713. If draws from a constant urn containing white and black balls are repeated a large number of times, the observed share of white balls 'converges' toward that actually contained by the urn. Considered by some as a mathematical theorem and by others as an experimental result, this 'law' was at the crossroads of the two currents in epistemological science: one 'hypothetical deductive,' the other 'empirical inductive.'

Beginning in the 1830s, the statistical description of 'observations in large numbers' became a regular activity of the state. Previously reserved for princes, this information henceforth available to 'enlightened men' was related to the population, to births, marriages, and deaths, to suicides and crimes, to epidemics, to foreign trade, schools, jails, and hospitals. It was generally a by-product of administration activity, not the result of special surveys. Only the population census, the showcase product of nineteenth-century statistics, was the object of regular surveys. These statistics were published in volumes with heterogeneous contents, but their very existence

suggests that the characteristics of society were henceforth a matter of scientific law and no longer of judicial law, that is, of observed regularity and not of the normative decisions of political power. Quetelet was the man who orchestrated this new way of thinking the social world. In the 1830s and 1840s, he set up administrative and social networks for the production of statistics and established—until the beginning of the twentieth century—how statistics were to be interpreted.

This interpretation is the result of the combination of two ideas developed from the law of large numbers: the generality of 'normal distribution' (or, in Quetelet's vocabulary: the 'law of possibilities') and the regularity of certain yearly statistics. As early as 1738, Abraham de Moivre, seeking to determine the convergence conditions for the law of large numbers, had formulated the mathematical expression of the future 'Gaussian law' as the limit of a binomial distribution. Then Laplace (1749–1827) had shown that this law constituted a good representation of the distribution of measurement errors in astronomy, hence the name that Quetelet and his contemporaries also used to designate it: the law of errors (the expression 'normal law,' under which it is known today, would not be introduced until the late nineteenth century by Karl Pearson).

Quetelet's daring intellectual masterstroke was to bring together two forms: on the one hand the law of errors in observation, and on the other, the law of distribution of certain body measurements of individuals in a population, such as the height of conscripts in a regiment. The likeness of the 'Gaussian' looks of these two forms of distribution justified the invention of a new being with a promise of notable posterity in social studies: the average man. Thus, Quetelet restricted the calculation and the legitimate use of averages to cases where the distribution of the observations had a Gaussian shape, analogous to that of the distribution of the astronomical observations of a star. Reasoning on that basis, just as previous to this distribution there was a real star (the 'cause' of the Gaussian-shaped distribution), previous to the equally Gaussian distribution of the height of conscripts there was a being of a reality comparable to the existence of the star. Quetelet's average man is also the 'constant cause,' previous to the observed controlled variability. He is a sort of model, of which specific individuals are imperfect copies.

The second part of this cognitive construction, which is so important in the ulterior uses of statistics in social studies, is the attention drawn by the 'remarkable regularity' of series of statistics, such as those of marriages, suicides, or crimes. Just as series of draws from an urn reveal a regularity in the observed frequency of white balls, the regularity in the rates of suicide or crime can be interpreted as resulting from series of draws from a population, some of the members of which are affected with a 'propensity' to suicide or crime. The average man is therefore endowed not only with physical attributes but also 'moral' ones, such as these propensities. Here again, just as the average heights of conscripts are stable, whereas individual heights are dispersed, crime or suicide rates are just as stable, whereas these acts are eminently individual and unpredictable. This form of statistics, then called 'moral,' signaled the beginning of sociology, a science of society radically distinct from a science of the individual, such as psychology (Porter 1986). Quetelet's reasoning would ground the one developed by Durkheim in *Suicide: A Study in Sociology* (1897).

This way of using statistical regularity to back the idea of the existence of a society ruled by specific 'laws,' distinct from those governing individual behavior, dominated nineteenth-century, and, in part, twentieth-century social studies. Around 1900, however, another approach appeared, this one centered on two ideas: the distribution (no longer just the average) of observations, and the correlation between two or several variables, observed in individuals (no longer just in groups, such as territories).

2. *Distribution, Correlation, and Causality*

This shift of interest from the average individual to the distributions and hierarchies among individuals, was connected to the rise, in late-century Victorian England, of a eugenicist and hereditarian current of thought inspired from Darwin (MacKenzie 1981). Its two advocates were Francis Galton (1822–1911), a cousin of Darwin, and Karl Pearson (1857–1936). In their attempt to measure biological heredity, which was central to their political construction, they created a radically new statistics tool that made it possible to conceive partial causality. Such causality had been absent from all previous forms of thought, for which A either is or is not the cause of B, but cannot be so somewhat or incompletely. Yet Galton's research on heredity led to such a formulation: the parents' height 'explains' the children's, but does not entirely 'determine' it. The taller fathers are, the taller are their sons *on average*, but, for a father's given height, the sons' height dispersion is great. This formalization of heredity led to the two related ideas of regression and correlation, later to be extensively used in social studies as symptoms of causality.

Pearson, however, greatly influenced by the antirealist philosophy of the physicist Ernst Mach, challenged the idea of 'causality,' which according to him was 'metaphysical,' and stuck to the one of 'correlation,' which he described with the help of 'contingency tables' (Pearson 1911, Chap. 5). For him, scientific laws are only summaries, brief descriptions in mental stenography, abridged formulas, a condensation of perception routines for future use and forecasting. Such formulas are the limits of observa-

tions that never perfectly respect the strict functional laws. The correlation coefficient makes it possible to measure the strength of the connection, between zero (independence) and one (strict dependence). Thus, in this conception of science associated by Pearson with the budding field of mathematical statistics, the reality of things can only be invoked for pragmatic ends and provided that the 'perception routines' are maintained. Similarly, 'causality' can only be insofar as it is a proven correlation, therefore predictable with a fairly high probability. Pearson's pointed formulations would constitute, in the early twentieth century, one of the foci of the epistemology of statistics applied to social studies. Others, in contrast, would seek to give new meaning to the concepts of reality and causality by defining them differently. These discussions were strongly related to the aims of the statistical work, strained between scientific knowledge and decisions.

Current mathematical statistics proceed from the works of Karl Pearson and his successors: his son Egon Pearson (1895–1980), the Polish mathematician Jerzy Neyman (1894–1981), the statistician pioneering in agricultural experimentation Ronald Fisher (1890–1962), and finally the engineer and beer brewer William Gosset, alias Student (1876–1937). These developments were the result of an increasingly thorough integration of so-called 'inferential' statistics into probabilistic models. The interpretation of these constructions is always stretched between two perspectives: the one of science, which aims to prove or test hypotheses, with truth as its goal, and the one of action, which aims to make the best decision, with efficiency as its goal. This tension explains a number of controversies that opposed the founders of inferential statistics in the 1930s. In effect, the essential innovations were often directly generated within the framework of research as applied to economic issues, for instance in the cases of Gosset and Fisher.

Gosset was employed in a brewery. He developed product quality-control techniques based on a small number of samples. He needed to appraise the variances and laws of distribution of parameters calculated on observations not complying with the supposed 'law of large numbers.' Fisher, who worked in an agricultural research center, could only carry out a limited number of controlled tests. He mitigated this limitation by artificially creating a randomness, itself controlled, for variables other than those for which he was trying to measure the effect. This 'randomization' technique thus introduced probabilistic chance into the very heart of the experimental process. Unlike Karl Pearson, Gosset and Fisher used distinct notations to designate, on the one hand, the theoretical parameter of a probability distribution (a mean, a variance, a correlation) and on the other, the estimate of this parameter, calculated on the basis of observations so insufficient in number that it was possible to disregard the gap between these two values, theoretical and estimated.

This new system of notation marked a decisive turning point: it enabled an inferential statistics based on probabilistic models. This form of statistics was developed in two directions. The estimation of parameters, which took into account a set of recorded data, presupposed that the model was true. The information produced by the model was combined with the data, but nothing indicated whether the model and the data were in agreement. In contrast, the hypothesis tests allowed this agreement to be tested and if necessary to modify the model: this was the inventive part of inferential statistics. In wondering whether a set of events could plausibly have occurred if a model were true, one compared these events—explicitly or otherwise—to those that would have occurred if the model were true, and made a judgment about the gap between these two sets of events.

This judgment could itself be made according to two different perspectives, which were the object of vivid controversy between Fisher on the one hand, and Neyman and Egon Pearson on the other. Fisher's test was placed in a perspective of truth and science: a theoretical hypothesis was judged plausible or was rejected, after consideration of the observed data. Neyman and Pearson's test, in contrast, was aimed at decision making and action. One evaluated the respective costs of accepting a false hypothesis and of rejecting a true one, described as errors of Type I and II. These two different aims—truth and economy—although supported by close probabilistic formalism, led to practically incommensurable argumentative worlds, as was shown by the dialogue of the deaf between Fisher on one side, and Neyman and Pearson on the other (Gigerenzer et al. 1989, pp. 90–109).

3. *Official Statistics and the Construction of the State*

At the same time as mathematical statistics were developing, so-called 'official' statistics were also being developed in 'bureaus of statistics,' for a long time on an independent course. These latter did not use the new mathematical tools until the 1930s in the United States and the 1950s in Europe, in particular when the random sample-survey method was used to study employment or household budgets. Yet in the 1840s, Quetelet had already actively pushed for such bureaus to be set up in different countries, and for their 'scientification' with the tools of the time. In 1853, he had begun organizing meetings of the 'International Congress of Statistics,' which led to the establishment in 1885 of the 'International Statistical Institute' (which still exists and includes mathematicians and official statisticians). One could write the history of these bureaus as an aspect of the more general history of the construction of the state, insofar as they developed and legitimized a common language specifically combining the authority of science and that of

the state (Anderson 1988, Desrosières 1998, Patriarca 1996).

More precisely, every period of the history of a state could be characterized by the list of 'socially judged social' questions that were consequently put on the agenda of official statistics. So were co-constructed three interdependent foci: representation, action, and statistics—a way of describing and interpreting social phenomena (to which social studies would increasingly contribute), a method for determining state intervention and public action, and finally, a list of statistical 'variables' and procedures aimed at measuring them.

Thus for example in England in the second third of the nineteenth century, poverty, mortality, and epidemic morbidity were followed closely in terms of a detailed geographical distribution (counties) by the General Register Office (GRO), set up by William Farr in 1837. England's economic liberalism and the Poor Law Amendment Act of 1834 (which led to the creation of workhouses) were consistent with this form of statistics. In the 1880s and 1890s, Galton and Pearson's hereditarian eugenics would compete with this 'environmentalism,' which explained poverty in terms of the social and territorial contexts. This new 'social philosophy' was reflected in news forms of political action, and of statistics. Thus the 'social classification' in five differentiated groups used by British statisticians throughout the twentieth century is marked by the political and cognitive configuration of the beginning of the century (Szreter 1996).

In all the important countries (including Great Britain) of the 1890s and 1900s, however, the work of the bureaus of statistics was guided by labor-related issues: employment, wages, workers' budgets, subsistence costs, etc. The modern idea of unemployment emerged, but its definition and its measurement were not standardized yet. This interest in labor statistics was linked to the fairly general development of a specific 'labor law' and the first so-called 'social welfare' legislation, such as Bismarck's in Germany, or that developed in the Nordic countries in the 1890s. It is significant that the application of the sample survey method (then called 'representative' survey) was first tested in Norway in 1895, precisely in view of preparing a new law enacting general pension funds and invalidity insurance: this suggests the consistency of the political, technical, and cognitive dimensions of this co-construction.

These forms of consistency are found in the statistics systems that were extensively developed, at a different scale, after 1945. At that time, public policies were governed by a number of issues: the regulation of the macroeconomic balance as seen through the Keynesian model, the reduction of social inequalities and the struggle against unemployment thanks to social-welfare systems, the democratization of school, etc. Some people then spoke of 'revolution' in government statistics (Duncan and Shelton 1978), and underscored its four components, which have largely shaped the present statistics systems. National accounting, a vast construction integrating a large number of statistics from different sources, was the instrument on which the macroeconomic models resulting from the Keynesian analysis were based. Sample surveys made it possible to study a much broader range of issues and to accumulate quantitative descriptions of the social world, which were unthinkable at a time when the observation techniques were limited to censuses and monographs. Statistical coordination, an apparently strictly administrative affair, was indispensable to make consistent the observations resulting from different fields. Finally, beginning in 1960, the generalization of computer data processing radically transformed the activity of bureaux of statistics.

So, 'official statistics,' placed at the junction between social studies, mathematics, and information on public policies, has become an important research component in the social studies. Given, however, that from the institutional standpoint it is generally placed outside, it is often hardly perceived by those who seek to draw up a panorama of these sciences. In fact, the way bureaus of statistics operate and are integrated into administrative and scientific contexts varies a lot from one country to another, so a history and a sociology of social studies cannot omit examining these institutions, which are often perceived as mere suppliers of data assumed to 'reflect reality,' when they are actually places where this 'reality' is instituted through co-constructed operations of social representation, public action, and statistical measurement.

See also: Estimation: Point and Interval; Galton, Sir Francis (1822–1911); Neyman, Jerzy (1894–1981); Pearson, Karl (1857–1936); Probability and Chance: Philosophical Aspects; Quantification in History; Quantification in the History of the Social Sciences; Quetelet, Adolphe (1796–1874); Statistical Methods, History of: Post-1900; Statistical Methods, History of: Pre-1900; Statistics: The Field

Bibliography

Anderson M J 1988 *The American Census. A Social History*. Yale University Press, New Haven, CT

Desrosières A 1998 *The Politics of Large Numbers. A History of Statistical Reasoning*. Harvard University Press, Cambridge, MA

Duncan J W, Shelton W C 1978 *Revolution in United States Government Statistics, 1926–1976*. US Department of Commerce, Washington, DC

Gigerenzer G et al. 1989 *The Empire of Chance. How Probability Changed Science and Everyday Life*. Cambridge University Press, Cambridge, UK

Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK

Klein J L 1997 *Statistics Visions in Time. A History of Time Series Analysis, 1662–1938*. Cambridge University Press, Cambridge, UK

Lazarsfeld P 1977 Notes in the history of quantification in sociology: Trends, sources and problems. In: Kendall M, Plackett R L (eds.) *Studies in the History of Statistics and Probability*. Griffin, London, Vol. 2, pp. 213–69
MacKenzie D 1981 *Statistics in Britain, 1865–1930. The Social Construction of Scientific Knowledge*. Edinburgh University Press, Edinburgh, UK
Patriarca S 1996 *Numbers and Nationhood: Writing Statistics in Nineteenth-century Italy*. Cambridge University Press, Cambridge, UK
Pearson K 1911 *The Grammar of Science*, 3rd edn rev. and enl.. A. and C. Black, London
Porter T 1986 *The Rise of Statistical Thinking, 1820–1900*. Princeton University Press, Princeton, NJ
Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press of Harvard University Press, Cambridge, MA
Szreter S 1996 *Fertility, Class and Gender in Britain, 1860–1940*. Cambridge University Press, Cambridge, UK

A. Desrosières

Statistics: The Field

Statistics is a term used to refer to both a field of scientific inquiry and a body of quantitative methods. The field of statistics has a 350-year intellectual history rooted in the origins of probability and the rudimentary tools of political arithmetic of the seventeenth century. Statistics came of age as a separate discipline with the development of formal inferential theories in the twentieth century. This article briefly traces some of this historical development and discusses current methodological and inferential approaches as well as some cross-cutting themes in the development of new statistical methods.

1. Introduction

Statistics is a body of quantitative methods associated with empirical observation. A primary goal of these methods is coping with uncertainty. Most formal statistical methods rely on probability theory to express this uncertainty and to provide a formal mathematical basis for data description and for analysis. The notion of variability associated with data, expressed through probability, plays a fundamental role in this theory. As a consequence, much statistical effort is focused on how to control and measure variability and or how to assign it to its sources.

Almost all characterizations of statistics as a field include the following elements:

(a) Designing experiments, surveys, and other systematic forms of empirical study.

(b) Summarizing and extracting information from data.

(c) Drawing formal inferences from empirical data through the use of probability.

(d) Communicating the results of statistical investigations to others, including scientists, policy makers, and the public.

This article describes a number of these elements, and the historical context out of which they grew. It provides a broad overview of the field, that can serve as a starting point to many of the other statistical entries in this encyclopedia.

2. The Origins of the Field

The word 'statistics' is related to the word 'state' and the original activity that was labeled as statistics was social in nature and related to elements of society through the organization of economic, demographic, and political facts. Paralleling this work to some extent was the development of the probability calculus and the theory of errors, typically associated with the physical sciences (see *Statistical Methods, History of: Pre-1900*). These traditions came together in the nineteenth century and led to the notion of statistics as a collection of methods for the analysis of scientific data and the drawing of inferences therefrom.

As Hacking (1990) has noted: 'By the end of the century chance had attained the respectability of a Victorian valet, ready to be the logical servant of the natural, biological and social sciences' (p. 2). At the beginning of the twentieth century, we see the emergence of statistics as a field under the leadership of Karl Pearson, George Udny Yule, Francis Y. Edgeworth, and others of the 'English' statistical school. As Stigler (1986) suggests:

> Before 1900 we see many scientists of different fields developing and using techniques we now recognize as belonging to modern statistics. After 1900 we begin to see identifiable statisticians developing such techniques into a unified logic of empirical science that goes far beyond its component parts. There was no sharp moment of birth; but with Pearson and Yule and the growing number of students in Pearson's laboratory, the infant discipline may be said to have arrived. (p. 361)

Pearson's laboratory at University College, London quickly became the first statistics department in the world and it was to influence subsequent developments in a profound fashion for the next three decades. Pearson and his colleagues founded the first methodologically-oriented statistics journal, *Biometrika*, and they stimulated the development of new approaches to statistical methods. What remained before statistics could legitimately take on the mantle of a field of inquiry, separate from mathematics or the use of statistical approaches in other fields, was the development of the formal foundations of theories of inference from observations, rooted in an axiomatic theory of probability.

Beginning at least with the Rev. Thomas Bayes and Pierre Simon Laplace in the eighteenth century, most early efforts at statistical inference used what was known as the method of inverse probability to update a prior probability using the observed data in what we now refer to as Bayes' Theorem. (For a discussion of who really invented Bayes' Theorem, see Stigler 1999, Chap. 15). Inverse probability came under challenge in the nineteenth century (see *Statistical Methods, History of: Pre-1900*), but viable alternative approaches gained little currency. It was only with the work of R. A. Fisher on statistical models, estimation, and significance tests, and Jerzy Neyman and Egon Pearson, in the 1920s and 1930s, on tests of hypotheses, that alternative approaches were fully articulated and given a formal foundation. Neyman's advocacy of the role of probability in the structuring of a frequency-based approach to sample surveys in 1934 and his development of confidence intervals further consolidated this effort at the development of a foundation for inference (cf. *Statistical Methods, History of: Post-1900* and the discussion of 'The inference experts' in Gigerenzer et al. 1989).

At about the same time Kolmogorov presented his famous axiomatic treatment of probability, and thus by the end of the 1930s, all of the requisite elements were finally in place for the identification of statistics as a field. Not coincidentally, the first statistical society devoted to the mathematical underpinnings of the field, The Institute of Mathematical Statistics, was created in the United States in the mid-1930s. It was during this same period that departments of statistics and statistical laboratories and groups were first formed in universities in the United States.

3. Emergence of Statistics as a Field

3.1 The Role of World War II

Perhaps the greatest catalysts to the emergence of statistics as a field were two major social events: the Great Depression of the 1930s and World War II. In the United States, one of the responses to the depression was the development of large-scale probability-based surveys to measure employment and unemployment. This was followed by the institutionalization of sampling as part of the 1940 US decennial census. But with World War II raging in Europe and in Asia, mathematicians and statisticians were drawn into the war effort, and as a consequence they turned their attention to a broad array of new problems. In particular, multiple statistical groups were established in both England and the US specifically to develop new methods and to provide consulting. (See Wallis 1980, on statistical groups in the US; Barnard and Plackett 1985, for related efforts in the United Kingdom; and Fienberg 1985). These groups not only created imaginative new techniques such as sequential analysis and statistical decision theory, but they also developed a shared research agenda. That agenda led to a blossoming of statistics after the war, and in the 1950s and 1960s to the creation of departments of statistics at universities—from coast to coast in the US, and to a lesser extent in England and elsewhere.

3.2 The Neo-Bayesian Revival

Although inverse probability came under challenge in the 1920s and 1930s, it was not totally abandoned. John Maynard Keynes (1921) wrote *A Treatise on Probability* that was rooted in this tradition, and Frank Ramsey (1926) provided an early effort at justifying the subjective nature of prior distributions and suggested the importance of utility functions as an adjunct to statistical inference. Bruno de Finetti provided further development of these ideas in the 1930s, while Harold Jeffreys (1938) created a separate 'objective' development of these and other statistical ideas on inverse probability.

Yet as statistics flourished in the post-World War II era, it was largely based on the developments of Fisher, Neyman and Pearson, as well as the decision theory methods of Abraham Wald (1950). L. J. Savage revived interest in the inverse probability approach with *The Foundations of Statistics* (1954) in which he attempted to provide the axiomatic foundation from the subjective perspective. In an essentially independent effort, Raiffa and Schlaifer (1961) attempted to provide inverse probability counterparts to many of the then existing frequentist tools, referring to these alternatives as 'Bayesian.' By 1960, the term 'Bayesian inference' had become standard usage in the statistical literature, the theoretical interest in the development of Bayesian approaches began to take hold, and the neo-Bayesian revival was underway. But the movement from Bayesian theory to statistical practice was slow, in large part because the computations associated with posterior distributions were an overwhelming stumbling block for those who were interested in the methods. Only in the 1980s and 1990s did new computational approaches revolutionize both Bayesian methods, and the interest in them, in a broad array of areas of application.

3.3 The Role of Computation in Statistics

From the days of Pearson and Fisher, computation played a crucial role in the development and application of statistics. Pearson's laboratory employed dozens of women who used mechanical devices to carry out the careful and painstaking calculations required to tabulate values from various probability distributions. This effort ultimately led to the creation of the *Biometrika Tables for Statisticians* that were so widely used by others applying tools such as chi-square tests and the like. Similarly, Fisher also

developed his own set of statistical tables with Frank Yates when he worked at Rothamsted Experiment Station in the 1920s and 1930s. One of the most famous pictures of Fisher shows him seated at Whittingehame Lodge, working at his desk calculator (see Box 1978).

The development of the modern computer revolutionized statistical calculation and practice, beginning with the creation of the first statistical packages in the 1960s—such as the BMDP package for biological and medical applications, and Datatext for statistical work in the social sciences. Other packages soon followed—such as SAS and SPSS for both data management and production-like statistical analyses, and MINITAB for the teaching of statistics. In 2001, in the era of the desktop personal computer, almost everyone has easy access to interactive statistical programs that can implement complex statistical procedures and produce publication-quality graphics. And there is a new generation of statistical tools that rely upon statistical simulation such as the bootstrap and Markov Chain Monte Carlo methods. Complementing the traditional production-like packages for statistical analysis are more methodologically oriented languages such as S and S-PLUS, and symbolic and algebraic calculation packages. Statistical journals and those in various fields of application devote considerable space to descriptions of such tools.

4. *Statistics at the End of the Twentieth Century*

It is widely recognized that any statistical analysis can only be as good as the underlying data. Consequently, statisticians take great care in the the design of methods for data collection and in their actual implementation. Some of the most important modes of statistical data collection include *censuses*, *experiments*, *observational studies*, and *sample surveys*, all of which are discussed elsewhere in this encyclopedia. Statistical experiments gain their strength and validity both through the random assignment of treatments to units and through the control of nontreatment variables. Similarly sample surveys gain their validity for generalization through the careful design of survey questionnaires and probability methods used for the selection of the sample units. Approaches to cope with the failure to fully implement randomization in experiments or random selection in sample surveys are discussed in *Experimental Design: Compliance* and *Nonsampling Errors*.

Data in some statistical studies are collected essentially at a single point in time (cross-sectional studies), while in others they are collected repeatedly at several time points or even continuously, while in yet others observations are collected sequentially, until sufficient information is available for inferential purposes. Different entries discuss these options and their strengths and weaknesses.

After a century of formal development, statistics as a field has developed a number of different approaches that rely on probability theory as a mathematical basis for description, analysis, and statistical inference. We provide an overview of some of these in the remainder of this section and provide some links to other entries in this encyclopedia.

4.1 *Data Analysis*

The least formal approach to inference is often the first employed. Its name stems from a famous article by John Tukey (1962), but it is rooted in the more traditional forms of descriptive statistical methods used for centuries.

Today, data analysis relies heavily on graphical methods and there are different traditions, such as those associated with

(a) The 'exploratory data analysis' methods suggested by Tukey and others (see *Exploratory Data Analysis: Univariate Methods*).

(b) The more stylized correspondence analysis techniques of Benzecri and the French school (see *Scaling: Correspondence Analysis*).

(c) The alphabet soup of computer-based multivariate methods that have emerged over the past decade such as ACE, MARS, CART, etc. (see *Exploratory Data Analysis: Univariate Methods*).

No matter which 'school' of data analysis someone adheres to, the spirit of the methods is typically to encourage the data to 'speak for themselves.' While no theory of data analysis has emerged, and perhaps none is to be expected, the flexibility of thought and method embodied in the data analytic ideas have influenced all of the other approaches.

4.2 *Frequentism*

The name of this group of methods refers to a hypothetical infinite sequence of data sets generated as was the data set in question. Inferences are to be made with respect to this hypothetical infinite sequence. (For details, see *Frequentist Inference*).

One of the leading frequentist methods is significance testing, formalized initially by R. A. Fisher (1925) and subsequently elaborated upon and extended by Neyman and Pearson and others (see below). Here a null hypothesis is chosen, for example, that the mean, μ, of a normally distributed set of observations is 0. Fisher suggested the choice of a test statistic, e.g., based on the sample mean, $\bar{x}$, and the calculation of the likelihood of observing an outcome as or more extreme as $\bar{x}$ is from $\mu = 0$, a quantity usually labeled as the p-value. When p is small (e.g., less than 5 percent), either a rare event has occurred or the null hypothesis is false. Within this theory, no probability can be given for which of these two conclusions is the case.

A related set of methods is testing hypotheses, as proposed by Neyman and Pearson (1928, 1932). In this approach, procedures are sought having the property that, for an infinite sequence of such sets, in only (say) 5 percent for would the null hypothesis be rejected if the null hypothesis were true. Often the infinite sequence is restricted to sets having the same sample size, but this is unnecessary. Here, in addition to the null hypothesis, an alternative hypothesis is specified. This permits the definition of a power curve, reflecting the frequency of rejecting the null hypothesis when the specified alternative is the case. But, as with the Fisherian approach, no probability can be given to either the null or the alternative hypotheses (see *Hypothesis Testing in Statistics*).

The construction of confidence intervals, following the proposal of Neyman (1934), is intimately related to testing hypotheses; indeed a 95 percent confidence interval may be regarded as the set of null hypotheses which, had they been tested at the 5 percent level of significance, would not have been rejected. A confidence interval is a random interval, having the property that the specified proportion (say 95 percent) of the infinite sequence, of random intervals would have covered the true value. For example, an interval that 95 percent of the time (by auxiliary randomization) is the whole real line, and 5 percent of the time is the empty set, is a valid 95 percent confidence interval.

Estimation of parameters—i.e., choosing a single value of the parameters that is in some sense best—is also an important frequentist method. Many methods have been proposed, both for particular models and as general approaches regardless of model, and their frequentist properties explored. These methods usually extended to intervals of values through inversion of test statistics or via other related devices (see e.g., *Fiducial and Structural Statistical Inference*). The resulting confidence intervals share many of the frequentist theoretical properties of the corresponding test procedures (see *Estimation: Point and Interval*).

Frequentist statisticians have explored a number of general properties thought to be desirable in a procedure, such as invariance, unbiasedness, sufficiency, conditioning on ancillary statistics, etc. (see *Invariance in Statistics*, *Estimation*: *Point and Interval*, *Statistical Sufficiency*). While each of these properties has examples in which it appears to produce satisfactory recommendations, there are others in which it does not. Additionally, these properties can conflict with each other. No general frequentist theory has emerged that proposes a hierarchy of desirable properties, leaving a frequentist without guidance in facing a new problem.

4.3 Likelihood Methods

The likelihood function (first studied systematically by R. A. Fisher) is the probability density of the data, viewed as a function of the parameters. It occupies an interesting middle ground in the philosophical debate, as it is used both by frequentists (as in maximum likelihood estimation) and by Bayesians in the transition from prior distributions to posterior distributions. A small group of scholars (among them G. A. Barnard, A. W. F. Edwards, R. Royall, D. Sprott) have proposed the likelihood function as an independent basis for inference (see *Likelihood in Statistics*). The issue of nuisance parameters has perplexed this group, since maximization, as would be consistent with maximum likelihood estimation, leads to different results in general than does integration, which would be consistent with Bayesian ideas.

4.4 Bayesian Methods

Both frequentists and Bayesians accept Bayes' Theorem as correct, but Bayesians use it far more heavily (see *Bayesian Statistics*). Bayesian analysis proceeds from the idea that probability is personal (see *Probability: Interpretations*) or subjective, reflecting the views of a particular person at a particular point in time. These views are summarized in the prior distribution over the parameter space. Together the prior distribution and the likelihood function define the joint distribution of the parameters and the data. This joint distribution can alternatively be factored as the product of the posterior distribution of the parameter given the data times the predictive distribution of the data.

In the past, Bayesian methods were deemed to be controversial because of the avowedly subjective nature of the prior distribution. But the controversy surrounding their use has lessened as recognition of the subjective nature of the likelihood has spread. Unlike frequentist methods, Bayesian methods are, in principle, free of the paradoxes and counterexamples that make classical statistics so perplexing. The development of hierarchical modeling and Markov Chain Monte Carlo (MCMC) methods have further added to the current popularity of the Bayesian approach, as they allow analyses of models that would otherwise be intractable (see *Monte Carlo Methods and Bayesian Computation: Overview* and *Markov Chain Monte Carlo Methods*).

Bayesian decision theory, which interacts closely with Bayesian statistical methods, is a useful way of modeling and addressing decision problems of experimental designs and data analysis and inference (see *Decision Theory: Bayesian*). It introduces the notion of utilities and the optimum decision combines probabilities of events with utilities by the calculation of expected utility and maximizing the latter (e.g., see the discussion in Lindley 2000).

Current research is attempting to use the Bayesian approach to hypothesis testing to provide tests and p-values with good frequentist properties (see Bayarri and Berger 2000).

4.5 Broad Models: Nonparametrics and Semiparametrics

These models include parameter spaces of infinite dimensions, whether addressed in a frequentist or Bayesian manner. In a sense, these models put more inferential weight on the assumption of conditional independence than does an ordinary parametric model. Methods and properties of these models are still being developed and some of these are described in various entries of this encyclopedia (see *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)*, *Nonparametric Statistics: Rank-based Methods*, *Semiparametric Models*).

4.6 Some Cross-cutting Themes

Often different fields of application of statistics need to address similar issues. For example, dimensionality of the parameter space is often a problem. As more parameters are added, the model will in general fit better (at least no worse). Is the apparent gain in accuracy worth the reduction in parsimony? There are many different ways to address this question in the various applied areas of statistics.

Another common theme, in some sense the obverse of the previous one, is the question of model selection and goodness of fit. In what sense can one say that a set of observations is well-approximated by a particular distribution? (cf. *Goodness of Fit: Overview*). All statistical theory relies at some level on the use of formal models, and the appropriateness of those models and their detailed specification are of concern to users of statistical methods, no matter which school of statistical inference they choose to work within.

5. Statistics in the Twenty-first Century

5.1 Adapting and Generalizing Methodology

Statistics as a field provides scientists with the basis for dealing with uncertainty, and, among other things, for generalizing from a sample to a population. There is a parallel sense in which statistics provides a basis for generalization: when similar tools are developed within specific substantive fields, such as experimental design methodology in agriculture and medicine, and sample surveys in economics and sociology. Statisticians have long recognized the common elements of such methodologies and have sought to develop generalized tools and theories to deal with these separate approaches (see e.g., Fienberg and Tanur 1989).

One hallmark of modern statistical science is the development of general frameworks that unify methodology. Thus the tools of Generalized Linear Models draw together methods for linear regression and analysis of various models with normal errors and those log-linear and logistic models for categorical data, in a broader and richer framework. Similarly, graphical models developed in the 1970s and 1980s use concepts of independence to integrate work in covariance section, decomposable log-linear models, and Markov random field models, and produce new methodology as a consequence. And the latent variable approaches from psychometrics and sociology have been tied with simultaneous equation and measurement error models from econometrics into a broader theory of covariance analysis and structural equations models.

Another hallmark of modern statistical science is the borrowing of methods in one field for application in another. One example is provided by Markov Chain Monte Carlo methods, now used widely in Bayesian statistics, which were first used in physics. Survival analysis, used in biostatistics to model the disease-free time or time-to-mortality of medical patients, and analyzed as reliability in quality control studies, are now used in econometrics to measure the time until an unemployed person gets a job. We anticipate that this trend of methodological borrowing will continue across fields of application.

5.2 Where Will New Statistical Developments be Focused?

In the issues of its year 2000 volume, the *Journal of the American Statistical Association* explored both the state of the art of statistics in diverse areas of application, and that of theory and methods, through a series of vignettes or short articles. These essays provide an excellent supplement to the entries of this encyclopedia on a wide range of topics, not only presenting a snapshot of the current state of play in selected areas of the field but also affecting some speculation on the next generation of developments. In an afterword to the last set of these vignettes, Casella (2000) summarizes five overarching themes that he observed in reading through the entire collection:

(a) Large datasets.
(b) High-dimensional/nonparametric models.
(c) Accessible computing.
(d) Bayes/frequentist/who cares?
(e) Theory/applied/why differentiate?

Not surprisingly, these themes fit well those that one can read into the statistical entries in this encyclopedia. The coming together of Bayesian and frequentist methods, for example, is illustrated by the movement of frequentists towards the use of hierarchical models (see *Hierarchical Models: Random and Fixed Effects*) and the regular consideration of frequentist properties of Bayesian procedures (e.g., Bayarri and Berger 2000). Similarly, MCMC methods are being widely used in non-Bayesian settings and, because they focus

on long-run sequences of dependent draws from multivariate probability distributions, there are frequentist elements that are brought to bear in the study of the convergence of MCMC procedures. Thus the oft-made distinction between the different schools of statistical inference (suggested in the preceding section) is not always clear in the context of real applications.

5.3 The Growing Importance of Statistics Across the Social and Behavioral Sciences

Statistics touches on an increasing number of fields of application, in the social sciences as in other areas of scholarship. Historically, the closest links have been with economics; together these fields share parentage of econometrics. There are now vigorous interactions with political science, law, sociology, psychology, anthropology, archeology, history, and many others.

In some fields, the development of statistical methods has not been universally welcomed. Using these methods well and knowledgeably requires an understanding both of the substantive field and of statistical methods. Sometimes this combination of skills has been difficult to develop.

Statistical methods are having increasing success in addressing questions throughout the social and behavioral sciences. Data are being collected and analyzed on an increasing variety of subjects, and the analyses are becoming increasingly sharply focused on the issues of interest.

We do not anticipate, nor would we find desirable, a future in which only statistical evidence was accepted in the social and behavioral sciences. There is room for, and need for, many different approaches. Nonetheless, we expect the excellent progress made in statistical methods in the social and behavioral sciences in recent decades to continue and intensify.

Bibliography

Barnard G A, Plackett R L 1985 Statistics in the United Kingdom, 1939–1945. In: Atkinson A C, Fienberg S E (eds.) *A Celebration of Statistics: The ISI Centennial Volume*. Springer-Verlag, New York, pp. 31–55

Bayarri M J, Berger J O 2000 P values for composite null models (with discussion). *Journal of the American Statistical Association* **95**: 1127–72

Box J 1978 *R. A. Fisher, The Life of a Scientist*. Wiley, New York

Casella G 2000 Afterword. *Journal of the American Statistical Association* **95**: 1388

Fienberg S E 1985 Statistical developments in World War II: An international perspective. In: Anthony C, Atkinson A C, Fienberg S E (eds.) *A Celebration of Statistics: The ISI Centennial Volume*. Springer-Verlag, New York, pp. 25–30

Fienberg S E, Tanur J M 1989 Combining cognitive and statistical approaches to survey design. *Science* **243**: 1017–22

Fisher R A 1925 *Statistical Methods for Research Workers*. Oliver and Boyd, London

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L 1989 *The Empire of Chance*. Cambridge University Press, Cambridge, UK

Hacking I 1990 *The Taming of Chance*. Cambridge University Press, Cambridge, UK

Jeffreys H 1938 *Theory of Probability*, 2nd edn. Clarendon Press, Oxford, UK

Keynes J 1921 *A Treatise on Probability*. Macmillan, London

Lindley D V 2000 1932 The philosophy of statistics (with discussion). *The Statistician* **49**: 293–337

Neyman J 1934 On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society* **97**: 558–625

Neyman J, Pearson E S 1928 On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika* **20A**: 175–240

Neyman J, Pearson E S 1932 On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series. A* **231**: 289–337

Raiffa H, Schlaifer R 1961 *Applied Statistical Decision Theory*. Harvard Business School, Boston

Ramsey F P 1926 Truth and probability. In: *The Foundations of Mathematics and Other Logical Essays*. Kegan Paul, London, pp.

Savage L J 1954 *The Foundations of Statistics*. Wiley, New York

Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, Cambridge, MA

Stigler S M 1999 *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA

Tukey John W 1962 The future of data analysis. *Annals of Mathematical Statistics* **33**: 1–67

Wald A 1950 *Statistical Decision Functions*. Wiley, New York

Wallis W 1980 The Statistical Research Group, 1942–1945 (with discussion). *Journal of the American Statistical Association* **75**: 320–35

S. E. Fienberg and J. B. Kadane

Status and Role, Social Psychology of

The social psychology of status and role refers to the study of the social interactive processes that involve people in displaying, creating, and changing statuses and roles. Status and role link individuals to society, to institutional structures, social groups, etc.

1. History of Status and Role

Status is a term originally referring to legally enforceable rights related to societal positions, such as the rights implied in the expression 'the status of citizens.' In the social sciences it has been used to refer to societal positions, such as ascribed (race, gender, age) and achieved (occupational) statuses. Historically different societal positions have been attributed vary-

ing degrees of legal rights (slaves, women, peasants); thus the term status carries with it an invidiousness. This connotation has been conceptualized as amounts of societal honor, esteem, and resources attributed to individuals or groups. The social psychological conception of role is derived from the theater applied to everyday life; individuals are conceived as playing roles in society similar to actors upon a stage. Role also has more than one meaning. Role is conceptualized as cultural norms, prescriptions, and expectations associated with statuses such as those of fulfilling the achieved status of physician or the ascribed status of woman. Role also has been formulated as the real practices associated with a status.

Both concepts have long histories. What is now termed symbolic interaction, a form of social psychology associated with the University of Chicago early in the twentieth century employed the concept of role as developed by George Herbert Mead and Robert E. Park (Stryker and Statham 1985). Mead's (1934) conceptions of 'role-playing' or 'taking the role of the other' are normative formulations used to theorize social processes such as the socialization of children, the development and evolution of self in interactions with others, and the acquisition of meanings in conversations. Park formulated roles as ubiquitous public performances in which people engage to construct the social organizations of society (Turner 1968). It was the anthropologist Ralph Linton who explicitly linked status to role. Linton theorized statuses as ascribed and achieved societal positions. Attached to each status is a role described as normative prescriptions (i.e., rights and duties, Linton 1936, p. 113) expressing expected behaviors associated with the status that the role incumbent is obliged and committed to fulfill. For Linton status is a structural unit of society separate from the individual who fills it and role is social process ('A *role* represents the dynamic aspect of a status' Linton 1936, p.114). Linton illustrates these by describing age and gender (he uses the word sex) as ascribed statuses associating these with culture-specific behaviors that constitute these roles in different societies.

2. *Elaborating the Social Psychology of Status and Role: 1940s to 1970s*

In the period following World War II Theodore Sarbin, a social psychologist, along with two sociologists, Erving Goffman and Ralph H. Turner, shaped the social psychological discourse of status and role as social processes. Sarbin moved Linton's focus from cultural prescriptions embedded in status shaping role incumbents' behavior to factors for effectively accomplishing roles. Goffman elaborated the metaphor of social life as theater, focusing on the processes of role playing in creating selves and definitions of the situations. Turner conceptualized indeterminacy and individuals' agency in status-role playing. Despite their different focuses the authors share intellectual heritages in Park and G. H. Mead.

In 1943, 1954, and 1968 Theodore Sarbin provided extended statements on role theory. The 1954 and 1968 essays were published in the influential *Handbook of Social Psychology*. In the 1954 essay Sarbin used, but collapsed, the terminology of status and role, conceiving of roles as simultaneously normative obligations ('expectations,' 'rights and obligations,' Sarbin 1954, p. 226) and as social positions ('We define position as a system of role expectations, p. 223). This formulation framed an evolving trend for these three interactively inclined authors; the concept of status became muted as the dynamic processes of role were moved center stage in their analyses.

In the late 1960s Sarbin and Allen formulated role enactment (1968, p. 497) as a dependent variable, analyzing the influences of independent variables that enhance or inhibit the effectiveness of a role's display. By effectiveness the authors mean motor and linguistic actions expressed in roles that are congruent with roles' prescriptions and expectations (1968, p. 550). Sarbin and Allen provide six independent variables that impact upon the effectiveness of role enactment. The variables (1968, pp. 546–7) are: (a) role expectations (validity of role expectations held by the actor); (b) role location (accuracy locating others and the self in the proper role system); (c) role demands (sensitivity to situationally generated role demands); (d) role skills (available general and specific skills for role performances); (e) self-role congruence (congruence of self-conception with the role); (f) audience effects (role reinforcement properties of the audience). Detailed analyses of the six variables indicate that role incumbents possess differing linguistic, motor, and emotional capacities furnishing them with the wherewithal to perform roles more or less effectively (i.e., convincingly, appropriately, authentically—1968, p. 490). Although Sarbin emphasizes individuals' capacities for effective role enactments, his role theorizing is interactive (e.g., 'role enactment always occurs in the social context of complementary roles,' p. 545) as evidenced by the concern with the causes, personal and social, for variation in role performances in social settings. In a closing section of the 1968 essay the authors turn their analysis around asking how role enactments, treated as independent variables, affect the creation, evolution, and change of role incumbents' social identities; this analysis of social identities parallels Goffman's conception of the construction of self in interaction.

Sarbin insists that his portrayal of status and role is within the context of a dramaturgical (social life as theater) metaphor (Sarbin and Allen 1968, pp. 488–9, Sarbin 1968, p. 546). That metaphor is even more appropriately applied to Erving Goffman's theorizing. In the preface to his *The Presentation of Self in Everyday Life* Goffman (1959, p. xi) so anoints his approach, and in the remainder of the book he

elaborates the metaphor. However, the focal concept of the book and of Goffman's corpus is the self. Goffman has devoted much less space to theorizing status and role than he has to the self. However, it is individual role players' actual public accomplishments in interactive performances that construct selves for others and self (Goffman 1959, pp. 252–3). Goffman defines role similarly to Linton as cultural prescriptions situated in statuses but also as typical real activities in which role players engage (i.e., role enactments, Goffman 1961, pp. 85, 93); he conceives of status as social position related to other positions and as honors and resources attached to the status.

Within the context of the dramaturgical metaphor Goffman theorizes a conception of interactive performances in which every rule and practice used to construct selves and define situations are sociological features of role playing. Goffman describes two settings in which role playing occurs: focused and unfocused interactions. The latter refers to the myriad trivial interchanges of daily life such as those between shoppers and supermarket cashiers, or among a group of unacquainted passengers riding an elevator (Goffman 1963). Sustained interactions are focused and particularly relevant; among these are organized work settings. Institutionalized work teams interest Goffman because role playing in these are staged in bounded structural and interactive circumstances that are specifiable precisely for analysis.

It is from the work setting of doctors and nurses as a hospital surgical team that Goffman formulates a theory of the processes of status and role (Goffman 1961). Goffman frames the essay by distinguishing between role embracing and distancing. 'To embrace a role is to disappear completely into the virtual self available in the situation' …, (Goffman 1961, p.106). It is the role incumbent engaged as Linton might describe '… to act solely in terms of the normative demands upon someone in his position' (p. 85). In contrast role distancing suggests that the role incumbent, while engaging in the normative performances of the role will also perform activities that are not part of, or may even be inconsistent with, the role in order to perpetuate the organization's work functions (pp. 108, 114–15).

Nowhere should role incumbents more totally embrace their roles than among doctors and nurses engaged in the serious work of surgery. Yet in typical ways doctors and nurses exhibit role-distancing activities in the operating theater. In addition to performing surgery a chief surgeon may find that he or she must act as 'team manager' expending status resources (e.g., their dignity) by playing the role of 'nice guy or gal' in order to reduce neophyte team members' anxieties, increasing their poise and communication during the surgery (Goffman 1961, p.122). Surgeons may mimic, jest, joke, sing, or discuss recent personal and news events to dispel tension. Surgeons may perform tasks not associated with their role and beneath their status (again expending status honor) to maintain the team's solidarity and commitments to the tasks. These role-distancing activities, Goffman suggests, are systemic and typical because they are derived from the sociological role-set in which surgeons are embedded. Some distancing activities evolve from the surgeon's ties to the hospital administration, others from their status as instructor and mentor to interns, still others from connections with patients and their family, and even the surgeon's own family may be the source for role-distancing performances. No matter where they originate from, these distancings are endemic to the surgical theater; they are performed to preserve and perpetuate the ordered functions of the team and its effective accomplishment of the surgical task.

Goffman argues that Linton's conception of the structural determination of roles (an embracing of role norms) is anomalous. He insists that adequate theorizing requires integrating normative role obligations with situationally required role distancing. Only such a complementary formulation proffers an empirically sound theory for the effective accomplishment of institutional tasks.

Turner does not emphasize the dramaturgical metaphor; instead he methodically and systematically theorizes and investigates the processes of status and role. Turner centers interactional, in contrast to structural, determination of status and role. Linton's structural conceptualization of role behavior determined by status did not sit well with Turner's epistemic roots in the Meadian tradition. Early in his career Turner published 'The Navy Disbursing Officer as Bureaucrat' (1947) addressed to this issue. In that essay Turner suggests that some status incumbents performed the role consistent with its bureaucratic normative role prescriptions, but others were capable of creating alternate solutions for the Navy Disbursing Officer role. These officers constructed novel, better adapted, more effective courses of action, innovating upon the cultural, structural, and personal constraints placed upon them as role incumbents.

This early essay indicated Turner's inclinations toward interactive determination. His subsequent writings became more ardently so inclined. Thus, he ultimately theorized that even role performances that publically conform to normatively expected behaviors may not be taken for granted as structurally determined. Instead, Turner conceives of status occupants as interactively engaged in constructing their roles and positions. For example, Turner argues that Sarbin's and Goffman's conceptions of 'appearances and props,' such as a military uniform or the spacial arrangement in classrooms such as the teacher's position at the head of class, do not assure incumbents' statuses nor their role performances. Turner advocates that in every instance it is the role performances, interactively achieved, that create an incumbent's right to claim a status and to play a role (Turner 1985).

Turner's conception of status and role is that they are volitionally constructed within the constraints of social settings. Paraphrasing his language, incumbents engage in processes of interactively creating their statuses and roles; they engage in role making (an intractable aspect of role playing) as well as role taking (Turner 1962).

While Turner assumes normatively prescribed roles exist, their prescriptions are general propositions (*gestalts* regarding such as values, goals, sentiments, styles, etc., Turner 1985), not specific dictates to behaviors, and therefore they are indeterminate. Individuals use these gestalts as schema to organize and interpret their own and others' behaviors. It is individuals' interpretations of these gestalts, within specific social contexts, that guide their performances. This proposition involved Turner in an epistemic coherent 'inductive' approach to the study of a variety of related topics on statuses and roles. However, after three decades of such publications Turner moved to unify his social psychology of status and role.

As the 1970s ended, a theoretical language to study the processes of status and role had evolved. The language assumed that while status and role exist as structural and normative features of society their enactment is determined by individual incumbents' qualities and accomplishments in situated interactions. The language formulated the possibilities not only for individuals' recapitulation of culturally prescribed status and role but also for actors' fashioning and performing differentiated variations of them. Status and role are linked to the development and character of selves and identities, as well as to definitions of situations, social organizations, and institutions. The language proffered an empirically grounded interactive theory of status and role social processes. While the theoretical language was bold, it was also preliminary. Nevertheless, as a series of loosely connected propositions and hypotheses it stimulated an array of empirical and theoretical topics, from status and role enactment, to status and role strain and conflict, differentiation and change, etc. The social psychology of status and role had progressed a long way from its initial formulations delivered by Mead, Park, and Linton.

3. *Consolidations in the 1980s*

The 1980s began a period of critical review and theoretical consolidation. Of the three influential social scientists, only Turner remained involved with the topic. Turner provides a unique effort at consolidating the social psychology of status and role. He, and his colleague Paul Colomy, introduced axioms to explain role processes (Turner and Colomy 1988, Turner 1990). The essays focus on the processes of role allocation, differentiation, and change. Three axioms determine these processes. They are: (a) functional efficiency of roles, (b) costs and benefits for role incumbency, and (c) positive and negative cultural representation of roles. Each of these three axioms separately and in combination explain the processes of role allocation and differentiation. Turner provides a similar but more complex analysis for role change (Turner 1990). The intent of these essays is to integrate role theory using a small, limited set of explanatory axiomatic factors. At the same time it is an effort to coherently integrate macro structural features to role processes (Turner 1990, pp. 106–8). The authors provide remarkably incisive models. However, their commitment to indeterminacy in the construction of statues and roles deters them from achieving theoretical closure. The models also still require empirical substantiation.

A second form of consolidation is the attempt to synthesize interactive and structural approaches (earlier attempts are in Coser 1966 and Merton 1975.) In the 1980s 'impartial' scholars tried to resolve the differences between the approaches in an effort to provide a unified theory. An underlying theme in these endeavors is that both approaches had advanced theoretically to where they were no longer in opposition with one another.

Warren Handel, in a cautiously reasoned essay (1979), suggests that interactive and structural approaches are 'compatible and complementary.' Structural approaches' increasing theoretical emphasis on role conflict undermines a previously held core conception of role incumbents' conformity to internalized norms permitting stability and variation in status and role performances. Handel suggests that interactive approaches' conception of negotiated meanings do not undermine structural conceptions of patterned conduct. Negotiated meanings may coexist within a working consensus among interacting individuals in a particular situation. Handel concludes by legitimating his synthetic formulation, suggesting it has been designed in accord with Mead's interactive theorizing. Jerold Heiss provides a similar synthesis between the two approaches. His reasoning is more ad hoc than is Handel's but it comes to a similar conclusion, 'The *general* role theory that I foresee for the future will combine the macroanalyses of the structuralists with the microanalyses of the interactionists' (Heiss 1981, p. 101, emphasis in the original).

In a third form of consolidation Sheldon Stryker attempts to develop a unified theory of roles by integrating interaction and structure. The framework is entitled a 'structural symbolic interactionism' (Stryker 1983, Stryker and Statham 1985), for some an oxymoron. The theory assumes that macro and micro social structures may constrain role improvisation. Among the micro and macro structures that do so are: persons' selves and identities, expectations embedded in roles, others' potential sanctions, structural closure and openness, cultural symbols' impositions. Each is a variable, thus its constraint of improvisation depends

on the relative capacity to impose itself upon role incumbents. Stryker generalizes that in interactions when role incumbents' selves and identities and their goals and interests are satisfied they will conform to the role expectations. In contrast, disruptive social circumstances will encourage role improvisation. Stryker's commitment to Mead's symbolic interaction inspires him also to suggest that the same structural conditions (e.g., high employee turnover) may provoke antithetical interpretations simultaneously suggesting opportunities for role innovation and demands for conformity. Stryker suggests that his theory is a work in progress which 'to this point has produced only a few reasonably rigorous theories. It largely consists of a set of assumptions and concepts ...' (Stryker and Statham 1985, p. 368).

4. Conclusion: To the Millennium and Beyond

The theoretical link between status and role was severed soon after World War II. The structural conception of status is now confined to such as 'status attainment studies,' i.e., investigations of individuals' and groups' relative position in stratification systems that are the result of acquired resources such as income, education, occupation. Status in social psychology is conceived as individuals' or groups' relative societal evaluation, thus referring to degrees of honor, esteem, or resources. Status in this form is used as an independent variable, analyzing its effects upon other social processes.

A second fault line, already discussed, occurred between structural and interactive approaches. Despite their controversy, both camps reduced the use of the term status; a situation particularly accented among interactionists who came to use role as both societal position and its cultural norms. The social psychology of status and role is now tantamount to the social psychology of roles. Additionally, recent studies of status and roles have relied upon and elaborated established theory rather than created theory anew.

Baker and Faulkner's (1991) conception of roles as employed by persons to acquire resources and combine into concrete positions is exemplary of the terminological and theoretical uses of contemporary role theory. Their study of Hollywood film producers, directors, and screen writers refers to these as roles and not as social positions or statuses. Their analysis then focuses upon combinational relations among the three roles before and after the rise of the movie blockbuster. It relates films' financial successes to innovative role combinations. In the production of films the stereotypical organization is functional specificity with three different persons playing the three roles. However, Baker and Faulkner find that functional specificity of producer and functional dedifferentiation of director-screen writer was the most effective arrangement for achieving financial successes during the blockbuster era. The authors present quotes from secondary sources to illustrate that role participants were aware of the advantages of this organizational arrangement. Although their study is primarily focused at the structural level the authors assume that organization is the result of persons' volitional activities, and they attribute the structural reorganization to movie industry leaders' personal agency. Neither in theory nor in practice is there a division between structural and interactive determination of roles and organization for the authors. This position permits the authors to conclude that role innovations may be used as organizational resources.

Each of the first three editions of the *Handbook of Social Psychology* (1954, 1968, 1985) devoted a chapter to role theory. The fourth edition of the *Handbook*, published in 1998, contains no chapter exclusive to roles. Instead role is introduced throughout the fourth edition in relation to its functions and displays in groups, gender, race, stereotyping, distributive justice and so forth. Although the language is still not unified, role and status are now part of the social science lexicon and are employed conceptually when empirical issues appropriately call for their use.

See also: Ethnomethodology: General; Goffman, Erving (1921–82); Group Processes, Social Psychology of; Groups, Sociology of; Interactionism: Symbolic; Mead, George Herbert (1863–1931); Personality and Conceptions of the Self; Self: History of the Concept; Social Psychology; Social Psychology: Sociological; Social Psychology, Theories of; Status and Role: Structural Aspects

Bibliography

Baker W E, Faulkner R R 1991 Role as resource in the Hollywood film industry. *American Journal of Sociology* **97**: 279–309

Coser R L 1966 Role distance, sociological ambivalence, and transitional status systems. *American Journal of Sociology* **72**: 173–87

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday Anchor Books, New York

Goffman E 1961 *Encounters: Two Studies in the Sociology of Interaction*. Bobbs-Merrill, Indianapolis, IN

Goffman E 1963 *Behavior in Public Places: Notes on the Social Organization of Gatherings*. Free Press, New York

Handel W 1979 Normative expectations and the emergence of meaning as solutions to problems: Convergence of structural and interactionist views. *American Journal of Sociology* **84**: 855–81

Heiss J 1981 Social roles. In: Rosenberg M, Turner R H (eds.) *Social Psychology: Sociological Perspectives*. Basic Books, New York, pp. 94–129

Linton R 1936 *The Study of Man: An Introduction*. Appleton-Century-Crofts, New York

Mead G H 1934 *Mind, Self and Society: From the Standpoint of a Social Behaviorist*. University of Chicago Press, Chicago

Merton R K 1975 Structural analysis in sociology. In: Blau P (ed.) *Approaches to the Study of Social Structure*. Free Press, New York, pp. 21–52
Sarbin T R 1954 Role theory. In: Lindzey G (ed.) *The Handbook of Social Psychology, Vol. 1: Theory and Method*. Addison-Wesley, Cambridge, MA, pp. 223–58
Sarbin T R 1968 Role: Psychological aspects. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan and Free Press, New York, Vol. 13, pp. 546–52
Sarbin T R, Allen V L 1968 Role theory. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*, 2nd edn. Addison-Wesley, Reading, MA, Vol. 1, pp. 488–567
Stryker S 1983 Social psychology from the standpoint of a structural symbolic interactionism: Toward an interdisciplinary social psychology. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 16, pp. 181–218
Stryker S, Statham A 1985 Symbolic interaction and role theory. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*, 3rd edn. Random House, New York, Vol. 1, pp. 311–78
Turner R H 1947 The Navy disbursing officer as a bureaucrat. *American Sociological Review* **12**: 342–48
Turner R H 1962 Role-taking: Process vs. conformity. In: Rose A (ed.) *Human Behavior and Social Process*. Houghton-Mifflin, Boston, pp. 20–40
Turner R H 1968 Role: Sociological aspects. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan and Free Press New York, Vol. 13, pp. 552–7
Turner R H 1985 Unanswered questions in the convergence between structuralist and interactionist role theories. In: Halle H J, Eisenstadt S N (eds.) *Micro-sociological Theory: Perspectives on Sociological Theory*. Sage, Beverly Hills, CA, Vol. 2, pp. 22–36
Turner R H 1990 Role change. In: Scott W R, Blake J (eds.) *Annual Review of Sociology*. Annual Reviews, Palo Alto, CA, Vol. 16, pp. 87–110
Turner R H, Colomy P 1988 Role differentiation: Orienting principles. In: Lawler E J, Markovsky B (eds.) *Advances in Group Processes A Research Annual*. JAI Press, Greenwich, CT, Vol. 5, pp. 1–27

G. M. Platt

Status and Role: Structural Aspects

The closely connected concepts of 'status' and 'role' have a long and complex history in a number of different traditions of social thought, but their contemporary meanings developed during the first half of the twentieth century. Despite it having been Shakespeare who claimed that 'all the world's a stage,' it was the social psychologists of the early twentieth century who systematically employed the concept of role and so paved the way for later 'dramaturgical' approaches to social life. This work also prepared the way for those who would develop the idea of status as the structural position corresponding to a role. The term status, however, has been used in two quite different ways, only one of which is directly relevant to this discussion. One usage has referred to hierarchical differences of prestige and social honor in systems of social stratification (see *Social Stratification*); the other has referred, more generally, to any structural position or location in a social system. It is the second sense that will be discussed in this article.

1. Status as Structural Position

The hierarchical concept of status was popularized by Max Weber, who contrasted stratification by status with stratification by class (Scott 1996). In the 1930s, however, cultural anthropologists introduced the idea of status as a position or location in a social structure (Linton 1936). This usage was rapidly adopted by structural functionalist sociologists. While this initially led to some confusion about whether or not particular statuses were hierarchically ranked, the new usage soon became established (Parsons 1942b). Some resolved this confusion by using 'position' or 'location' in preference to status (Warner 1952, p. 46), and this practice will be followed in this article.

The idea of status as structural position originated in the idea that social structure consisted of culturally defined institutions. This had its origins in Durkheim, but it was the development of a clear understanding of culture by American anthropologists (Kroeber 1917, Kluckhohn 1954) that gave the idea its contemporary form. It became a central concept in sociological analysis (Parsons 1951).

It was held that the culture of a society establishes a set of social institutions. These are standardized normative patterns that people are expected to follow in their social actions. They regulate individual and group behavior by defining the nature of their social relations. Institutions define 'proper, legitimate or expected modes of action or of social relationship' (Parsons 1940, p. 53), and so guide and channel behavior. Parsons defined social structure as 'a patterned system of social relationships of actors' (Parsons 1945, p. 230).

The institutionalized expectations that people share define their social positions relative to other positions, and to the performances of the individuals who occupy these (Parsons 1942b, p. 143). This knowledge of social positions provides people with the conceptual maps that they use to organize their social actions. Like all ideas, they are contained in individual minds. However, because they are *shared*, there is a reciprocity of expectations that solidifies them into distinctively collective representations (Davis 1948, p. 87). Social positions are Durkheimian social facts with the power to constrain the behavior of individuals. We do not simply 'know' the social positions that make up our society, we experience social pressure to conform to the expectations they involve.

These pressures are felt through role expectations. Roles define the specific rights and obligations that are

entailed in a social position (Dahrendorf 1958, p. 36). They tell us what to do and what to expect others to do (Parsons 1940, p. 54). They provide us with a definition of the situation that sets the limits within which we may legitimately act (Williams 1960).

Some writers used the concept of role to describe the *actual* behavior associated with a position, but most have recognized that a normative focus requires a concern for *expected* behavior (Levy 1952, pp. 158–60). The relationship between expected behavior and actual behavior is very complex, and there is no one-to-one correspondence between norms and behavior. Roles are definitions of what people are expected to do in different situations. They define the goals, the means, and the attitudes that are culturally recognized as legitimate for those who occupy particular social positions.

2. *Role Theory*

These ideas were taken up during the 1950s as the basis of a 'role theory' that tried to set out a comprehensive paradigm for role analysis. The most influential formulation of this was that of Gross et al. (1958), who used it to explore the behavior of school superintendents. They did this using Merton's (1957) argument that each social position is associated with an array of role-specific forms of behavior that together comprise a 'role set.' A medical student, for example, must act as a student not only in relation to other students, but also in relation to teachers, doctors, nurses, social workers, patients, medical technicians, and so on. In each of these relations, the student is likely to encounter different expectations about her or his behavior (Merton 1957, p. 112).

Gross et al. showed that the school superintendent had to negotiate the conflicting expectations held by teachers, parents, children, governors, and ancillary staff. They recognized that this produces varying degrees of strain or conflict in role expectations. The concept of 'role conflict' has been central to many applications of role theory.

While this work concentrated on the structural aspects of position and role, it was complemented by those who worked on a social psychology of roles (Newcomb 1950). These writers explored the learning processes through which roles are acquired and the social pressures to conformity that exist in face-to-face situations. Socialization and conformity were central principles for role theorists, and their views have been subject to persistent criticism.

3. *Socialization and Conformity*

Role theorists see the norms that define social positions as learned through a process of socialization. The expectations of others are *internalized* in the personality as it is formed. They become objects of moral commitment, and individuals are likely to be strongly attached to them. They tend to regard conformity as a duty (Parsons 1940).

This is associated with the argument that there is consensus over social norms. Members of a society share their moral commitments and so violations are met with guilt or shame in the violator and indignation or condemnation in those whom the deviance affects. Deviance from these shared norms is likely to result in 'corrective forces' or sanctions that may vary from mild disapproval to overt coercion. For many, this led to the conclusion that deviance was both infrequent and automatically corrected. Societies were seen as self-regulating systems in which actors had little choice but to conform.

The degree of consensus in a society rarely reaches this level, however, and individuals are never perfectly socialized (Wrong 1961). These points were especially forcefully raised against Parsons, but even he recognized that the 'completely institutionalized' social relationship was a rarity. Although broad descriptive contrasts could be drawn between societies in terms of their common values, 'when dynamic problems of directions and processes of change are at issue, it is essential to give specific attention to the elements of malintegration, tension and strain in the social structure' (Parsons 1942a, p. 117). Nevertheless, the mainstream of role theory overemphasized both consensus and internalization. It was conflict theorists such as Dahrendorf (1958) who showed that role expectations have to be seen in relation to the distribution of power and that this can give them a solidity that does not depend on a consensus of opinion.

What passes for consensus may be the ideas of a dominant or powerful group (Rex 1961, Shils 1961). A minority may be able to use its resources to institutionalize its own norms for the population as a whole. This may be especially effective where the majority of the population forms a relatively unorganized mass without strong counterinstitutions of its own.

4. *Role-taking and Role-making*

Structural functional and conflict theories of position and role, supported by social psychological theories of socialization and conformity, lay at the heart of mainstream sociology from the 1940s to the 1960s. During the 1960s, however, radical and forceful objections were raised against this (see *Action Theory: Psychological*). The immediate origin of the sociological concept of role had been Mead's (1927) account of taking the role of the other, and critics of the mainstream turned to Mead for support. They questioned cultural determinism and its oversocialized image of the actor that saw people as the mere puppets of structural forces. Blumer (1962, p. 189) held that a

structure of positions is simply a 'framework inside of which action takes place.' It is a conceptual guide for action, not an external, substantial entity or an independent coercive power.

Turner (1962) took this argument to the heart of role theory. He held that the emphasis on conformity gave only a partial view of role behavior. It stressed imitation and ignored innovation, failing to recognize that role-takers were also 'role-makers.' Individuals do not simply take over roles as templates for conformist behavior. Turner held that role-taking always required improvisory behavior. The role knowledge that individuals learn during their primary and secondary socialization does not give them precise, programmed instructions for behavior in all the many unique and unpredictable circumstances in which they are likely to find themselves. Individuals tentatively interpret and reinterpret each other's actions in the situations that they encounter, recreating their roles from the raw materials provided to them during socialization (Turner 1962, p. 23). Shared knowledge of social positions sets the general conditions for action, but it does not fully determine it (Blumer 1962, p. 190). The normative functionalist approach to role-taking has to be broadened into a more comprehensive account.

The fullest development of this view was that of Erving Goffman (1959, 1963), who explored the ways in which individuals actively and creatively construct the images that they present in their role behavior. Goffman systematically employed the theatrical metaphor of role-playing, holding that individuals are never merely actors following a script, but are authors as well. Using such ideas as props, scenery, front stage, and back stage, Goffman showed how public performances depend on the more private situations to which individuals can withdraw. He introduced a number of novel concepts, most notably 'role distance.'

5. *A Sense of Social Structure*

The criticisms made by the symbolic interactionists added a whole new dimension to the analysis of position and role. While some of the critics of mainstream views saw this role as a complete alternative to the orthodoxy, others saw it as complementing the structural account. These differences persist, and have been compounded by a newer and more radical line of argument. Influenced by phenomenology and ethnomethodology, Aaron Cicourel has produced the most systematic statement of this critique.

Cicourel asked the fundamental question: how is role-taking possible? His answer was that the taking and making of roles rests on a set of cognitive processes through which actors give meaning to the world and so sustain a 'sense of social structure' (Cicourel 1972, p. 11). People do carry role information in their heads, but they must also be able to recognize when one particular position or role is relevant, and they must be able to infer what expectations others have of their behavior. They cannot make sense of their social world simply by drawing on the role and positional knowledge that they have learned during their socialization. Before they can apply norms in particular situations, they must arrive at some kind of understanding of what kind of situation it is. This means that 'members of a society must acquire the competence to assign meaning to their environment so that surface rules and their articulation with particular cases can be made' (Cicourel 1968, p. 52). This ability to infer and to impute meaning to situations is a practical skill that is an essential condition for any social life at all. Cicourel sees this skill as an interactional competence, making explicit parallels with Chomsky's concept of linguistic competence. Though he does not adopt Chomsky's own rationalist theory of the mind, Cicourel does take over his stress on the generative capacities that are provided by human competences. It is their practical, meaning-making skills that allow people to use their knowledge of social norms to generate appropriate role behavior. The structural aspects of positions and roles, therefore, are seen by Cicourel as resting on the possession of a complex set of cognitive procedures (also termed inductive, interpretive, or inference procedures) that operate in the same way as the deep structure grammatical rules of a language.

Cicourel illustrated these cognitive procedures by drawing on Alfred Schutz' discussion of the assumptions that people must make for social interaction to be possible. Schutz held that individuals must assume a reciprocity of perspectives between themselves and their potential partners, they must fill in the gaps in their knowledge through the *et cetera* principle, and they must assume that things occur as 'normal form.' These and similar cognitive procedures constitute the mental module that makes it possible for actors to generate appropriate but innovative responses in changing circumstances, despite the fact that they have only fragmentary and uncertain evidence available to them. They allow people to assign meaning and relevance to the objects in their environment and to construct definitions of the situation that allow them to infer which of the norms stored in their memories are relevant. People build a sense of social structure that allows them to orient themselves appropriately in the various situations that they encounter. Once the meaning of a situation has been decided, norms can be invoked on the assumption that there is a consensus among those with whom they interact and that these are, indeed, the appropriate norms. Normative order and role behavior, therefore, are negotiated and constructed on the basis of the underlying sense of social structure that interactional competence makes possible.

6. Conclusion

Status—or position—and role remain central concepts in the normative account of social structure. This normative account does not tell us everything that we need to know about social structure (Lockwood 1956, 1964), but it is an essential aspect of any structural account. The use of these normative concepts, however, must incorporate an awareness of the creative and innovative role-making activities that are integral to all role behavior. It is important to recognize that improvisation and negotiation are essential features of the construction of social action. It is also important to recognize, however, that these processes are possible only because individuals possess certain cognitive skills that allow them to infer the meaning of social situations and to build a sense of the social structures in which they must act. Structural positions and roles are human accomplishments, but they have real consequences for the people who occupy them.

See also: Action Theory: Psychological; Dramaturgical Analysis: Sociological; Functionalism in Sociology; Institutions; Norms; Phenomenology in Sociology; Social Stratification; Status and Role, Social Psychology of

Bibliography

Blumer H 1962 Society as symbolic interaction. In: Rose A (ed.) *Human Behavior and Social Processes*. Routledge and Kegan Paul, London

Cicourel A V 1968 The acquisition of social structure: Towards a developmental sociology of language and meaning. In: Cicourel A V (ed.) 1973 *Cognitive Sociology*. Penguin, Harmondsworth, UK

Cicourel A V 1972 Interpretive procedures and normative rules in the negotiation of status and role. In: Cicourel A V (ed.) 1973 *Cognitive Sociology*. Penguin, Harmondsworth, UK

Dahrendorf R 1958 Homo Sociologicus: On the history, significance, and limits of the category of social role. In: Dahrendorf R (ed.) 1968 *Essays in the Theory of Society*. Routledge and Kegan Paul, London

Davis K 1948 *Human Society*. Macmillan, New York

Goffman E 1959 *The Presentation of Self in Everyday Life*. Penguin, Harmondsworth

Goffman E 1963 *Relations in Public Places*. Free Press, New York

Gross N, Mason W S, Mc Eachern A W 1958 *Explorations in Role Analysis: Studies of the School Superintendency Role*. John Wiley and Sons, New York

Kluckhohn C 1954 Culture and behavior. In: Lindzey G (ed.) *Handbook of Social Psychology*. Addison-Wesley, Cambridge, MA, Vol. 2

Kroeber A 1917 The superorganic. *American Anthropologist* **19**: 162–213

Levy M J 1952 *The Structure of Society*. Princeton University Press, Princeton, NJ

Linton R 1936 *The Study of Man*. D. Appleton Century, New York

Lockwood D 1956 Some remarks on the social system. *British Journal of Sociology* **7**: 134–46

Lockwood D 1964 Social integration and system integration. In: Zollschan G, Hirsch W (eds.) 1976 *System, Change and Conflict*. Houghton-Mifflin, New York

Mead G H 1927/1934 *Mind, Self and Society From the Standpoint of Social Behaviorism*. University of Chicago Press, Chicago

Merton R K 1957 The role set: Problems in sociological theory. *British Journal of Sociology* **8**: 106–20

Newcomb T M 1950 *Social Psychology*. Dryden Press, New York

Parsons T 1940 The motivation of economic activity. In: Parsons T (ed.) 1954 *Essays in Sociological Theory*, 2nd edn. Free Press, New York

Parsons T 1942a Democracy and social structure in pre-Nazi Germany. In: Parsons T (ed.) 1954 *Essays in Sociological Theory*, 2nd edn. Free Press, New York

Parsons T 1942b Propaganda and social control. In: Parsons T (ed.) 1954 *Essays in Sociological Theory*, 2nd edn. Free Press, New York

Parsons T 1945 The present position and prospects of systematic theory in sociology. In: Parsons T (ed.) 1954 *Essays in Sociological Theory*, 2nd edn. Free Press, New York

Parsons T 1951 *The Social System*. Free Press, New York

Rex J A 1961 *Key Problems of Sociological Theory*. Routledge and Kegan Paul, London

Scott J 1996 *Stratification and Power: Structures of Class, Status and Command*. Polity Press, Cambridge, UK

Shils E A 1961 Centre and periphery. In: Shils E A (ed.) 1975 *Centre and Periphery: Essays in Macrosociology*. University of Chicago Press, Chicago

Turner R 1962 Role-taking: Process versus conformity. In: Rose A (ed.) *Human Behavior and Social Processes*. Routledge and Kegan Paul, London

Warner W L 1952 *The Structure of American Life*. University of Chicago Press, Chicago

Williams R 1960 *American Society: A Sociological Investigation*. 2nd edn, Alfred A Knopf, New York

Wrong D 1961 The oversocialized concept of man in modern sociology. *American Sociological Review* **26**: 183–93

J. Scott

Statutes of Limitations

1. Definition

Statutes of limitations are procedural rules that limit legal actions on the basis of time. They derive from both legislative and judicial sources and exist in many continental and non-continental legal systems. Specifically, they regulate the amount of time that a potential plaintiff has to initiate a formal legal claim. In general, once a statute of limitation expires, a plaintiff cannot press a particular legal claim, regardless of its underlying substantive merits.

It is important to distinguish statutes of limitations from the closely related yet distinct common law doctrine of 'laches.' The doctrine of laches vests courts with discretionary authority to determine whether a plaintiff, while commencing a legal action within the

time requirements set forth by the relevant statute of limitations, nonetheless has unreasonably delayed bringing a lawsuit to the detriment of a defendant (Heriot 1992). Both statutes of limitations and the doctrine of laches limit a plaintiff's access to court.

2. *Purposes*

Statutes of limitations seek to reconcile numerous individual and social interests. These interests, broadly construed, include a desire for an orderly legal system that resolves claims promptly, efficiently, and equitably and also promotes legal stability.

Within the context of litigation, these general interests become more individualized, acute, and particularized. Plaintiffs benefit from having a reasonable amount of time within which to press a legal claim. Some amount of time is necessary so that a potential plaintiff can contemplate possible litigation and gather and assess evidence that will inform a decision about whether to pursue formal legal remedies. Statutes of limitations induce plaintiffs to pursue legal claims diligently and promptly by punishing those who do not by foreclosing the opportunity to pursue a legal claim. Also, defendants have a legitimate interest in repose as well as not having to defend against 'stale' claims. Moreover, defendants should not be disadvantaged in court due to lost evidence, faded memories, and the disappearance of witnesses—all possibilities that increase with the passage of time.

Numerous social interests cut in different directions. It is, of course, well known that societies benefit when their judicial systems vindicate their members' legal rights. Other social interests include a desire for a secure and stable legal system. Moreover, a desire to minimize court congestion is shared by many. Finally, the probability of legal error increases when plaintiffs wait too long before commencing legal action and degrade components critical to a fair, just, and thorough trial.

3. *History and Development*

Statutes of limitations have existed for centuries; scholars trace the origins of statutes of limitations to the early days of Roman Law (Lowenthal 1950). By 1236 in England, statutes of limitations governing the transfer of property were enacted (Pollock and Maitland 1898). During the reign of Henry VIII, Parliament passed legislation establishing some time limits that affected real property transfers (Heriot 1992). The Limitation Act of 1623 marks the advent of the first systematic and reasonably effective rules governing time limitations for commencing legal actions. Notably, the Act expressly tolled limits for, among others, infants as well as the aged, imprisoned, and infirm. In addition, the Act codified the common law doctrine of adverse possession which generally provides an individual with legal title to property that the individual has occupied openly, notoriously, and continuously for a prescribed period of time. The Act also marks the beginning of the modern law of limitations governing personal actions in common law (Lowenthal 1950) and served as the model adopted by many American legislatures (Heriot 1992).

Statutes of limitations are found in many common law and non-common law settings. In common law settings, such as the United States, a recent survey of California law uncovered more than 32,000 statutes of limitations (or time-limiting bars) in that state alone (Montoya 1996). These enactments provide insights into a legal system's moral and ethical orientation, and in particular, their preference for human life issues over legal claims relating to other issues. In California, for example, statutes of limitations for slander or libel claims typically prevent any such claim being initiated more than one year after the action. On the other hand, there is no statutory limit whatsoever for murder.

Statutes of limitations are found in non-common law settings as well. For example, general 30-year statutes of limitation for civil actions operate in France and Germany (Martindale-Hubbell 2000). Of course, deviations from the general rule peculiar to each country also exist. In Germany, a general four-year statute of limitation governs certain business claims; three years for tort actions (Martindale-Hubbell 2000). In France, a ten-year limit governs most tort actions; a two-year limitation bounds claims on most insurance policies (Martindale-Hubbell 2000).

4. *Rationale*

Reflecting statutes of limitations' effort to reconcile conflicting principles, rationales advanced by scholars justifying such statutes vary. Both efficiency and equity (or 'fairness') emerge in the literature as common rationales for statutes of limitations (see, for example, Lowenthal 1950). Epstein (1986) advances a utilitarian justification that, in a world where everyone has an equal probability of being a plaintiff or defendant, the application of statutes of limitations and the resulting reduction in legal error, administrative and transaction costs bring about a net gain for the benefit of all.

5. *Criticisms*

Most critics of statutes of limitations focus on three points. One critique dwells on the particular time limits imposed on particular legal claims (see, for example, Bibas 1994, Thomas 1991). A second critique focuses on the operation of statutes of limitations,

especially on when time limits begin to run against a possible plaintiff (see, for example, Glimcher 1982). A third, broader critique flows from the courts' application of time limits in a 'rule-like' manner. Some critics of the rule-like application of statutes of limitations advance as one alternative administration by judges on a case-by-case basis (see, for example, Richardson 1997).

However, even most critics recognize that a rule-like application of statutes of limitations yields certain advantages. Such an application provides benefits from the fact that it is comparatively easy to administer, since decisions depend on a limited number of ascertainable facts and results can often be predicted *ex ante*. Moreover, well-crafted statutes convey in advance clear requirements for those contemplating pressing legal claims. Finally, a rule-like application of statutes of limitations reduces the potential for judicial abuse of discretionary powers.

Problems with statutes of limitations arise when the application of rules generates difficult or inequitable results. Critics point to such results as evidence for a more flexible application of statutes of limitations and for a close examination, case-by-case, of the way in which a particular application advances or impedes the underlying rationale for the time limits.

A case-by-case or individualized (i.e., non-rule-based) approach to applying statutes of limitations possesses certain advantages. Principally, this discretionary approach permits a closer examination of whether a specific claim falls within the range of difficulties that statutes of limitations were designed to address. Judicial discretion can reduce the possibility of injustice generated by a mechanical application of a statute of limitation. Ironically, greater judicial discretion in the application of such rules as statutes of limitations entails costs as well as benefits, because they generate greater uncertainty and thus increase the danger of judicial abuse.

Over one hundred years ago, Justice Oliver Wendell Holmes (1897) asked, 'What is the justification for depriving a man of his rights, a pure evil as far as it goes, in consequence of the lapse of time?' Justice Holmes' rhetorical question highlights a general ambivalence surrounding statutes of limitations and their applications (Ochoa and Winstrich 1997).

Benefits flowing from the application of statutes of limitations are frequently long-term and sometimes difficult to discern. In contrast, the costs imposed by statutes of limitations are short-term and more immediate and visceral. By seeking to reconcile critical principles that sometimes collide in litigation, statutes of limitations generate ambivalence and unease which weakens—but by no means dislodges—their secure and long-standing position in many—if not most—legal systems.

See also: Common Law; Legal Systems, Classification of; Rule of Law; Rules in the Legal Process

Bibliography

Bibas S A 1994 The case against statutes of limitations for stolen art. *Yale Law Journal* **103**: 2437–69

Epstein R A 1986 Past and future: The temporal dimension in the law of property. *Washington University Law Quarterly* **64**: 667–722

Glimcher S D 1982 Statutes of limitations and the discovery rule in latent injury claims: An exception of the law? *University of Pittsburgh Law Review* **43**: 501–23

Heriot G L 1992 A study in the choice of forum: Statutes of limitation and the doctrine of Laches. *Brigham Young University Law Review* **1992**: 917–70

Holmes O W 1897 The path of the law. *Harvard Law Review* **10**: 457–78

Lowenthal M A 1950 Developments in the Law: Statutes of Limitations. *Harvard Law Review* **63**: 1177–270

Martindale-Hubbell International Law Digest 2000

Montoya J M 1996 California statutes of limitations. *Southwestern University School of Law* **25**: 745–1373

Ochoa T T, Winstrich A J 1997 The puzzling purposes of statutes of limitations. *Pacific Law Journal* **28**: 453–514

Pollock F, Maitland F W 1898 *The History of English Law Before the Time of Edward I*, 2nd edn. Cambridge University Press, Cambridge, UK (Ledlie's translation)

Richardson E J 1997 Eliminating the limitations of limitations law. *Arizona State Law Journal* **29**: 1015–74

Thomas R L 1991 Adult survivors of childhood sexual abuse and statutes of limitations: A call for legislative action. *Wake Forest Law Review* **26**: 1245–95

M. Heise

Stereotypes, Social Psychology of

1. Origin and History

Introduced to the social sciences by Lippmann in his book *Public Opinion* (1922), the concept of stereotype refers to beliefs, knowledge, and expectations of social groups. To capture the idea of a stereotype, Lippmann made famous the phrase 'pictures in our heads' to refer to an internal, mental representation of social groups in contrast to their external reality. At a time when mental constructs of any type were undeveloped and soon to be frowned upon in American psychological discourse, it is especially remarkable that Lippman and other early theorists detected and described the psychological importance of stereotypes. Not only was such a possibility theorized, the results of empirical tests of stereotypes were made available by Katz and Braly (1933). Their checklist method of asking respondents to assign trait adjectives (e.g., intelligent, dishonest) to ethnic and national groups (e.g., Jews, Germans) dominated the field, and in modified form, remains in use even today. With the exception of a hiatus in research on stereotypes at mid-century, the concept of stereotype has occupied a dominant position in psychology since the 1920s.

Reviews of historical and contemporary research on stereotypes are liberally available (see Fiske 1998).

The concept of stereotype, in many ways, remains largely in keeping with its original formulation, although some shifts in definition and emphasis are worth noting. In the earliest proposals, stereotypes were regarded to be inaccurate assessments of groups; in fact, according to Lippmann, a stereotype was a 'pseudoenvironment' or 'fiction' and according to Katz and Braly (1933) it was an unjustified and contradictory reaction to an outgroup. In contemporary psychology, the accuracy with which a stereotype captures the essence of a social group is not of central theoretical interest. The belief that 'Many X's are Y' may well be accurate (e.g., many neurosurgeons are men) but if such a belief is applied in judging an individual member of a group (e.g., Female X is not, or cannot be, a neurosurgeon), a stereotype is seen to be in play. This process, by which an individual is given or denied an attribute because of membership in a group, is regarded to be of psychological and social interest. Such a shift is the result of a change in emphasis from examining beliefs about social groups per se (stereotype content), to an interest in the mental mechanics by which they influence interpersonal and intergroup perception and interaction (stereotype process). It is obvious that both aspects of stereotypes, their content and process, are critical to an understanding of their nature and function in social interaction. In fact, attention to implicit or automatic stereotypes in recent years has rekindled an interest in stereotype content and strength and in their relationship to related constructs such as explicit stereotypes, implicit and explicit prejudice, and group identity.

From the earliest use of the term in psychology, stereotypes have been regarded as the cognitive (thought) as opposed to the affective (feeling) component of mental representations of social groups. As such, the construct is tied to but differentiated from the concept of attitude, preference, or liking as well as the concept of discrimination. 'I do not like group X' is a verbal statement of an attitude or prejudice, while an act denying friendship or intimacy with a member of group X may be regarded as a behavioral indicator of prejudice. To say 'I believe group X to be incompetent' is to express a stereotype, while an act of denying employment to a member of group X based on such a belief is seen as the behavioral indicator of that stereotype.Of course, such arbitrary distinctions between mental and behavioral measures disappear when measures of brain activation are introduced. Verbal self-reports, bodily gestures, decisions about hiring are all viewed as behavioral, as opposed to brain indicators of stereotypes and prejudice. It is likely that just as measures of activation in particular sub-cortical structures like the amygdala have been shown to be associated with behavioral measures of prejudice (Phelps et al. 2000), the future will yield similar studies of stereotypes and their neural correlates. Such research has the potential to unify social, cognitive, and neural levels of analysis by demonstrating that the learning of stereotypes that culture and social environments make possible can be synchronously detected in observable behavior as well as in brain activation.

The lure of a cognitive analysis in psychology led to a preponderance of attention to stereotypes, to the exclusion of its sister concept, prejudice. Fiske notes a 5:1 ratio in published work on the two concepts between 1974 and 1995. That trend appears to be shifting with attention to both stereotype and prejudice in recent years and especially to their relationships as observed at both conscious and unconscious levels. In addition, research has focused on the role of mood states in dictating stereotype expression with the counterintuitive proposal that positive mood increases rather than decreases reliance on stereotypes (Bodenhausen et al. 1994). The relationship between stereotype (cognition) and prejudice (attitude) has been a complex one with early work linking the two concepts closely enough so as to regard them as synonymous. Today, the relationship between the two is an uneasy one, without a clear sense of the exact nature of the relationship between stereotypic thoughts and prejudicial feelings. It is assumed that stereotypes and prejudice need not be evaluatively compatible. Attitude (liking) can be dissociated from stereotypes (competence, respect, etc.).

When searching for large-scale theories of stereotypes or theories in which the role of stereotypes is central, one comes up relatively empty-handed. The focus of research appears not to have been on the development of broad theories, although two exceptions must be noted, in each of which stereotypes play a role in the larger framework: a psychodynamic, individual difference approach advancing the notion of an authoritarian personality to understand ethnocentrism (Adorno et al. 1950), and a social-cognitive approach based on social identity and self-categorization (Tajfel and Turner 1986) to understand individual–group relationships and their psychological and social consequences. The paucity of large-scale theories of stereotyping and prejudice is compensated by creative experiments that form the groundwork of what may in the future yield unified theories of stereotypes and related constructs of prejudice and discrimination. The last three decades are likely to be remembered for noteworthy experimental discoveries and diversity in the manner in which stereotypes have been measured, with special emphasis on the role of consciousness in thought and feeling about social groups.

2. *Categorization and Beyond*

To perceive is to differentiate, and social perception inherently involves the ability to see differentiation among groups, for example, to see women as dif-

ferentiated from men, elderly as differentiated from young, European as differentiated from Asian. Individual humans each belong to multiple social groups, with gender, age, race/ethnicity, class, religion, and nationality being the most obvious such categories and numerous other differentiations also easily perceived along dimensions such as political orientation, physical attractiveness, etc. A fundamental feature of social perception is the classification of individuals into groups, and the process by which such classification is achieved is assumed to be automatic (Fiske 1998). Once categorized, attributes believed to be associated with the group are generalized to individuals who qualify for group membership. Thus, for example, women may come to be seen as nurturant and nice, men as strong and competent whether they as individuals, deserve those ascriptions or not.

Stereotypes are fundamental to the ability to perceive, remember, plan, and act. Functionally, they may be regarded as mental helpers that operate in the form of heuristics or short-cuts. Historically, this view of stereotypes as devices that allow a sensible reading of a complex world marked a breakthrough in research on stereotypes. Advanced by Allport (1954) and Tajfel (1969), the idea that stereotypes were inherent to the act of social categorization now forms the basis of the modern view. As such, stereotypes are regarded to be ordinary in nature, in the sense that they are the byproducts of basic processes of perception and categorization, learning, and memory. This cognitive view of stereotypes has dominated the field since the early 1980s, and from this standpoint several noteworthy discoveries about the nature of stereotypes have emerged.

3. *Discoveries*

Early research in this cognitive tradition showed that within-group differences are minimized and between-group differences are exaggerated, with greater confusion, for example, in memory for within-race and within-gender comparisons than between group comparisons. Stereotypes can be generated through a misperception about the frequency with which an attribute is associated with a group. For example, perceivers may come to see an illusory correlation based on the shared distinctiveness of particular attributes. Specifically, when a doubly distinctive event occurs (e.g., when a statistical minority performs a low frequency behavior) such an event is overemphasized in memory and produces a bias in perception such that their occurrence is overestimated.

Hamilton and Gifford (1976) have shown how such a process can lead to the assessment that members of minority groups perform more evaluatively negative behaviors than is actually the case. Research in this tradition has also provided a good understanding of how stereotypes, in the form of expectancies of what social groups are capable of, can lead to a bias to confirm rather than disconfirm expectancies. Moreover, the targets who are stereotyped may themselves fulfill the expectancies the perceiver holds about them to complete the biased cycle of intergroup interaction. Stereotypes are the vehicles of essentialist thinking about social groups. Dispositional group attributions, or the belief that groups are inherently the way they are, can lead to the assessment that attributes associated with groups are stable and unchanging.

Stereotypes play their role from the earliest moments in the information processing sequence, giving preference to stereotype-consistent options. To date the evidence overwhelmingly suggests that stereotypes influence the manner in which information is sought, perceived, remembered, and judged. Stereotypes limit the amount of information that is required to make a judgment by giving meaning to partial and even degraded information, and they allow decisions to be reached even when time is short. Stereotypes facilitate initial identification of congruent information and disallow attention to incongruent information. Although research has focused on the role of stereotypes in thinking, they are viewed as serving a socially pragmatic function—as helping social interaction proceed smoothly and with ease (Fiske 1998).

From such research it may appear that the only judgments of individuals that ensue are those that rely entirely on knowledge of and beliefs about social groups. Yet, that is not the case. Depending on the context and the motivation of the perceiver, category-based stereotypic judgments may well yield to more individualized, person-based judgments (Brewer 1988, Fiske and Nuberg 1990). Indeed, experiments on the mechanisms of stereotype expression have regularly shown that stereotypes may or may not play a role depending on the extent to which such knowledge is seen as meaningful or justified (Banaji et al. 1993).

The routine simplification function of stereotypes can have far-reaching effects. In the work of Tajfel and co-workers, categorization of individuals into social groups was shown to produce a heightening of perceived differences, and these differences (even when the distinction between groups was arbitrary and minimal) created intergroup discrimination in the form of greater resource allocation to members of one's own than another group. Thus, emanating from ordinary categorization, stereotypes can play a crucial role in interpersonal and intergroup relations. Besides routine simplification, stereotypes are also regarded as serving the function of protecting a stable and psychologically justified view of the world and the place of humans, as members of social groups, in it. Jost and Banaji (1994) argued that holding negative stereotypes of another's group may serve not only an ego-protective function ('I am better than you') and a group-protective function ('My group is better than yours') but a system justifying function as well. The counterintuitive hypothesis from such a perspective is

that when status hierarchies relegate groups to relative positions of inferiority and superiority, members of disadvantaged groups may themselves come to hold negative beliefs about their groups in the service of a larger system in which social groups are arranged.

3.1 Implicit Stereotypes

With its roots in the ideas of Allport and Tajfel, the notion that stereotypes may operate without conscious awareness, conscious intention, and conscious control is hardly surprising. In fact, throughout the twentieth century, experiments have shown that in one form or another stereotypes emerge spontaneously from initial categorization and continue to have a life of their own independent of conscious will. Yet, it would be fair to say that a direct interest in implicit or unconscious social cognition is relatively recent, with theoretical input from theories of unconscious mental life and methodological input from the development of new measurement tools and techniques.

Contrast the following two measures of stereotypes. A respondent is asked to indicate, using a traditional verbal self-report scale, the extent to which African Americans are scholarly and athletic. Or, a respondent is asked to rapidly pair words like 'scholar' and 'athlete' with faces of African–Americans, and the time to do so is measured. The first measure assumes the ability to respond without self-presentational concerns, and more importantly, it assumes the ability to be able to adequately reflect on the content of one's thoughts and provide an accurate indication of the complex association between race and psychological attributes. The second measure, although not within the traditional view of stereotype assessment, provides a measure of the strength of association between the group and the attributes. Such a measure has been taken to be an indicator of the stereotype and its strength. To investigate the implicit or automatic manner in which stereotypes of social groups may express themselves, investigators have used a variety of techniques from measuring response latencies (i.e., the time to make a response), to examining errors in memory and biases in linguistic reports. The largest single body of work has used response latencies as indicators of automatic stereotypes and prejudice and the data from such measures have yielded several new results and debates about them (see Banaji 2001).

Stereotypes can be activated by the mere presentation of symbols of social group or group-related attributes. It appears that although conscious prejudice and stereotypes have changed, their less conscious, automatic expressions are strikingly strong. As measured by the Implicit Association Test (Greenwald et al. 1998) automatic stereotypes appear to exist in robust form; large effect sizes are the hallmark of automatic stereotypes (see Nosek et al. in press). A priming measure has also been widely utilized in which prime-target pairs are presented in close succession and response latency to the target serving as the measure of automatic stereotypes. For example, responses are reliably faster to female first names ('Jane') when the immediately preceding word is stereotypically consistent ('nurse') than inconsistent ('doctor'). Such effects are obtained with words and pictures and they generalize to a variety of social groups.

Given the socially significant consequences of stereotype use, investigations of the variability and malleability of automatic stereotypes have been examined. Research has focused on the relationship between conscious and unconscious expressions of stereotypes and prejudice. As Devine (1989) showed, evidence of automatic race stereotypes is present irrespective of the degree of conscious prejudice toward Black Americans. Additionally, Banaji and Hardin (1996) showed that automatic gender stereotypes were manifested irrespective of endorsement of conscious attitudes and beliefs about gender egalitarianism. Such results point to the dissociation between conscious and unconscious social stereotypes, but it is clear that a simple dissociation may not adequately or accurately capture this relationship. Rather, results are now available that indicate that those with higher levels of conscious prejudice may also show higher levels of automatic or implicit prejudice. It appears that studies using multiple measures of each stereotype and statistical tools to uncover latent factors will yield evidence in favor of a relationship between conscious and unconscious stereotypes, while also revealing their unique and non-overlapping nature.

Questions concerning the controllability of automatic stereotypes are hotly debated (Fiske 1998). It appears that a desire to believe that stereotypes can be controlled, perhaps because of their pernicious social consequences, can result in the wishful assessment that they are indeed controllable. Automatic stereotypes do not appear to be controllable by ordinary acts of conscious will. However, habitual patterns of thought, feeling, and behavior toward social groups that cohere with broader value systems and ideology appear to predict automatic responses. In addition, Greenwald et al. (in press) have shown that automatic identity with one's group can predict stereotypes held about the group and attitudes toward it and have put forth a unified theory of self, group stereotypes, and attitudes. In support, they have found that attitudes toward mathematics and science can be predicted by the strength of the automatic stereotype that math is male or masculine. Women who hold a stronger math = male stereotype also show more negative attitudes toward mathematics.

The effects of minor interventions to activate stereotype-incongruent associations (e.g., female–strong) can be detected in weaker automatic stereotypes (Blair et al. in press). Such findings point to the flexibility of the representations of social stereotypes. Although the category 'strong women' may be

counter-stereotypic, interventions that highlight this association can produce a lowering of the default stereotype of female = weak. The possibility of such strategies for inducing a shift in automatic stereotypes and the potential to track stereotypes through both behavioral and brain activation measures has the potential, in the future, to inform about stereotype representation, process, content, and mechanisms for social change.

See also: Mental Representation of Persons, Psychology of; Prejudice in Society; Schemas, Frames, and Scripts in Cognitive Psychology; Schemas, Social Psychology of; Small-group Interaction and Gender; Stigma, Social Psychology of

Bibliography

Adorno T W, Frenkel-Brunswick E, Levinson D J, Sanford R N 1950 *The Authoritarian Personality*. Harper, New York

Allport G 1954 *The Nature of Prejudice*. Doubleday, New York

Banaji M R 2001 Implicit attitudes can be measured. In: Roediger H L III, Nairne J S, Neath I, Surprenant A (eds.) *The Nature of Remembering: Essays in Honor of Robert G. Crowder*. American Psychological Association, Washington, DC

Banaji M R, Hardin C D 1996 Automatic stereotyping. *Psychological Science* **7**: 136–41

Banaji M R, Hardin C, Rothman A J 1993 Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology* **65**: 272–81

Blair I V, Ma J E, Lenton A P in press Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*

Bodenhausen G V, Kramer G P, Susser K 1994 Happiness and stereotypic thinking in social judgment. *Journal of Personality and Social Psychology* **66**: 621–32

Brewer M 1988 A dual-process model of impression formation. In: Srull T K, Wyer R S Jr (eds.) *Advances in Social Cognition*. Erlbaum, Hillsdale, NJ, pp. 1–36

Devine P G 1989 Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* **56**: 5–18

Fiske S T 1998 Stereotyping, prejudice, and discrimination. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology* 4th edn. McGraw Hill, New York, pp. 357–411

Fiske S T, Neuberg S L 1990 A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In: Zanna M P (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, pp. 1–74

Greenwald A G, Banaji M R, Rudman L, Farnham S, Nosek B A, Mellott D in press A unified theory of social cognition. *Psychological Review*

Greenwald A G, Mellott D, Schwartz J 1998 Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* **74**: 1464–80

Hamilton D L, Gifford R K 1976 Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology* **12**: 392–407

Jost J T, Banaji M R 1994 The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology* **33**: 1–27

Katz D, Braly K 1933 Racial stereotypes in one hundred college students. *Journal of Abnormal and Social Psychology* **28**: 280–90

Lippmann W 1922 *Public Opinion*. Harcourt, Brace, Jovanovich, New York

Nosek B, Banaji M R, Greenwald A G in press Harvesting intergroup attitudes and stereotypes from a demonstration website. *Group Dynamics*

Phelps E A, O'Connor K J, Cunningham W A, Funayama S, Gatenby J C, Gore J C, Banaji M R 2000 Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience* **12**: 729–38

Tajfel H 1969 Cognitive aspects of prejudice. *Journal of Social Issues* **25**: 79–97

Tajfel H, Turner J C 1986 The social identity theory of intergroup behavior. In: Worchel S, Austin W A (eds.) *Psychology of Intergroup Relations*. Nelson-Hall, Chicago, pp. 7–24

M. R. Banaji

Stevens, Stanley Smith (1906–73)

S. Smith Stevens (1906–73) was a leading figure in the field of experimental psychology during the middle part of the twentieth century. He is best known for his work in psychophysics and especially for the development of the psychophysical power law, but he also made important contributions to the study of hearing, and to an understanding of measurement theory and its role in psychology. Aside from his many research reports, his most influential publications include the classic *Hearing: its Psychology and Physiology* (1938) co-authored with Hallowell Davis, the *Handbook of Experimental Psychology* (1951), and *Psychophysics: Introduction to its Perceptual, Neural and Social Prospects* (1975) edited by his wife, Geraldine Stevens.

1. Beginnings

To understand fully the significance of any scientist's contribution, it is important to know the context in which it occurred. It is particularly useful to be acquainted both with the scientists who paved the way and with those who followed. Historians of experimental psychology have traced the beginnings of the field to the German scientist Wilhelm Wundt (1832–1920), who established the first laboratory for psychological research in 1879. Although other scientists had made important contributions to the emerging field, it was Wundt who first identified an entire range of questions in psychology amenable to laboratory inves-

tigation, and it was to his laboratory in Leipzig that students from many countries came for training in experimental psychology. Stevens, whose distinguished career culminated in his position as Professor of Psychophysics at Harvard University, stands in a direct line of scholarly descent from Wundt as a member of the fourth generation. He was trained by E. G. Boring (1886–1968) at Harvard, who in turn had studied at Cornell with E. B. Titchener (1867–1927), an Englishman who had received his degree in Leipzig from Wundt in 1892. And whereas Wundt may be considered the first to devote himself to Experimental Psychology, it may well be that Stevens was the last. His work marked and contributed to a transition from one era to another in the evolution of scientific psychology and an understanding of that transition makes an interesting story.

Stevens was born on November 4, 1906 in Ogden, Utah, home to the Mormon Church, and, following the customs of that institution, he departed at the age of 17 to spend 3 years as a missionary in Belgium and later in France despite a complete ignorance of the language. It seems likely that this experience contributed to the development of persistence and determination, traits that were to serve him well in his career as a scientist. On his return he entered the University of Utah and, after two years, transferred to Stanford, where he graduated in 1931 having taken such a wide variety of courses that the identification of a major was problematic. He was accepted by the Harvard Medical School but demurred on the grounds that the required payment of a $50 fee and the need to spend the summer taking a course in organic chemistry seemed equally unattractive. But he did decide to study at Harvard, and duly enrolled in the School of Education as the easiest (and most economical) path to Harvard's resources. The transforming experiences there were E. G. Boring's course on perception and service as Boring's unpaid research assistant. By the end of his second year, Stevens had earned a PhD in the Department of Philosophy (from which the Psychology Department emerged a year later). A year studying physiology under Hallowell Davis and another as a research fellow in physics led to appointment in 1936 as an instructor in the Psychology Department, where he remained until his death on January 18, 1973 in Vail, Colorado. By 1946 he achieved the rank of full professor, and in 1962 the University acceded to his request that he be named Professor of Psychophysics in honor of the field founded a century earlier by G. T. Fechner, whose theories, ironically, Stevens worked hard to refute.

2. *Stevens's Psychophysical Power Law*

Scholars familiar with his work quickly could agree in making a list of Stevens's major achievements, but could and do disagree about their relative importance and their lasting value. For some the single most important accomplishment was his discovery of the psychophysical power law that sometimes does and always should carry his name (Stevens 1957). He not only revived a view broached by some in the nineteenth century, but also created a body of evidence that firmly established what is sometimes called the Psychophysical Power Law and sometimes Stevens's Law. It concerns the relation between the strength of some form of energy, such as the sound pressure level of a tone, and the magnitude of the corresponding sensory experience, loudness in this example. It is easily discovered that sensation strength is nonlinearly related to stimulus intensity (two candles do not make a room seem twice as bright as one), but it is harder to say what that relation is. For over a century before Stevens approached the problem, the prevailing view, Fechner's Law, asserted that, as stimulus strength grew geometrically (by constant ratios), sensation strength increased arithmetically. This logarithmic relation had little supporting evidence but remained in many texts for lack of a good alternative. Stevens came to a different conclusion on the basis of data obtained by asking observers who were presented with stimuli varying in intensity to make numerical judgments of their subjective experience. In the case of loudness he found that judged loudness grew as a power function of sound pressure (with an exponent of about 0.6), not as the logarithmic relation predicted by Fechner (Stevens 1961). A similar relation was found to hold between judged brightness and luminance, and eventually, such a relation was found for dozens of other continua. In each case, as the amount or intensity of a stimulus grew by ratio steps, so too did the observer's estimate of his subjective experience of it. Stevens's Law seemed applicable to all intensive perceptual continua.

Equally interesting was the discovery that for each perceptual continuum, there was a distinctive value of the exponent relating numerical judgment to stimulus intensity. In hearing, for example, a huge range of sound pressures was compressed into a much smaller range of loudness, whereas for electric shock applied to the fingertips, a quite small range of currents between detectability and pain was expanded into a larger range of perceived intensities.

Although Stevens originally argued that the method of magnitude estimation provided a 'direct measure' of sensation (see *Sensation and Perception: Direct Scaling*) and that the power law described how energy in the environment was transformed through the relevant sensory apparatus into sensation, critics soon showed this assumption to be untenable. What remained was an empirical principle relating judgment to stimulus intensity.

Stevens went on to show that numerical magnitude estimation was a special case of a more general paradigm, cross-modal matching. If the observer is asked to match loudness, for example, by manipu-

lating the luminance of a light source until its brightness matches the level of the target loudness, the power relation is again the result. And, neatly enough, the exponent of that function is exactly predictable by the exponents identified through magnitude estimation of brightness and loudness alone (Stevens 1969). It is this ability of observers to match perceived magnitudes across many different perceptual attributes, and the discovery that in all cases the result is described by a power function, that are the two pillars on which Stevens's Law rests. Despite the many controversies over its meaning, this simple empirical principle stands securely on a mountain of evidence and must be accommodated by any theory of psychophysics.

3. *Scales of Measurement*

Others attach greater importance to Stevens's ideas about scales of measurement. In a seminal paper (1946) he created a taxonomy of measurement, which identified four classes of scales—nominal, ordinal, interval, and ratio—distinguishing among them on the basis of the operations that leave the form of the scale invariant. Nominal scales assigned numbers that serve only to distinguish one event or class of events from another (the classic example is the set of numbers used to identify each member of a team); any one-to-one substitution is permissible because player identity is preserved. Ordinal scales of measurement permit a rank ordering of measured events (relative hardness of materials, for example) and will remain invariant under monotonic increasing transformations. Interval scales entail a constant unit of measurement and so permit the calculation of differences between any two values (as for the common scales of temperature) and remain invariant under linear transformations. Ratio scales (the bread and butter of physics) also feature constant units of measurement, but in addition allow the expression of two values as a ratio, since a true zero exists; weight is a good example. Such a scale is invariant under a simple multiplicative transformation.

It is harder to assess the impact of this conceptualization of measurement than that of the work on cross-modal matching judgments. The latter was instrumental in creating new bodies of knowledge, especially in the field of sensory psychology, and continues to play an influential role in psychophysical theory and research. But Stevens's four-step hierarchy of measurement scales encountered an assortment of reactions that leaves its present status in some doubt.

Contemporary theorists find the nominal scale of little interest because it imposes no ordering on the measured entities. And several authorities have noted the omission of a log interval scale from what Stevens had proposed as an exhaustive listing. The relevance of this classification to the selection of statistical procedures was hotly debated (e.g., see Binder 1984). Stevens had argued that the type of measurement placed constraints on which statistics can be used, but statisticians were quick to cite seeming exceptions, and the usefulness of Stevens's measurement taxonomy as a guide to the selection of statistical techniques became uncertain.

But, for psychologists, his ideas about measurement were, and in some ways continue to be, of great importance. At a time when Stevens was trying to provide support for his newly-created scale of loudness, he felt it important to show that it met the same criteria of legitimacy as did the familiar scales of the more developed sciences. It was a time when the status of psychology as a science was a topic of lively debate, and Stevens was much involved in efforts to identify the critical features of science and the ways in which psychology shared them. During the 1940s and 1950s questions about the nature of science and psychology's place in it were discussed at length, and Stevens's analysis of measurement in general and the ways in which psychologists could use scales of measurement was very influential. The idea that a measurement scale—or, for that matter, any lawful relation—could be expressed in terms of the operations that left it invariant, played an important role in the thinking of his students, and no doubt for scientists in other disciplines as well. The work that Stevens began has now evolved into a mathematical specialty in its own right, yielding results that confirm some of what Stevens intuited, and disconfirming other aspects of his work. (It must be remembered that the ideas he was advancing lent themselves to mathematical formalization, though he himself did not provide it.) That his ideas about measurement have continued to be developed and refined by other scientists today may be the ultimate accolade for his pioneering work in this field (e.g., see Luce and Narens 1999).

4. *Hearing*

If Stevens had never established the psychophysical power law and had left the study of measurement untouched, he would still be known and remembered as a pioneer in the field of psychoacoustics, the scientific study of the perception of sound (e.g., see Scharf 1999). His collaboration with Hallowell Davis in the late 1930s resulted in *Hearing*, a book that for many years was the fundamental source on the subject and has been ranked by some scholars as second only to the work of Helmholtz in its originality and breadth. In addition to his measurements of loudness, he did important work on auditory localization, and on the relation of pitch to intensity. At the behest of the US Army Air Corps he established the Psycho-Acoustic Laboratory at Harvard in 1940 to study the effect of noise on pilot psychomotor efficiency. As its director

he was able to attract a group of students and colleagues of extraordinary ability, among them W. R. Garner, K. D. Kryter, J. C. R. Licklider, and G. A. Miller, all of whom made important contributions to psychoacoustics and beyond. It was an interesting aspect of his personality that, despite a rather gruff manner, he functioned as an intellectual magnet, attracting to his laboratory some of the most talented researchers in the world. Not the least of these was G. von Békésy, who spent almost 20 years in Stevens' basement lab and won a Nobel Prize during that period for his early work (in Budapest and Stockholm) on the ear.

Throughout his career there were several periods of intense activity as a laboratory scientist. But in the period before World War II he was also active in the discussion of philosophical questions about the nature of psychology. He was impressed by the work of the physicist P. W. Bridgman on the central role of operationism in science, and argued for its use in defining the concepts of psychology. And for a period he was associated closely with a group of logical and methodological positivists whose ideas about the nature of science seemed especially hospitable to a behaviorist psychology (Stevens 1939). Perhaps one manifestation of this outlook was a reluctance to extend his thinking very far beyond his data. In a period when many experimental psychologists were developing ever more complex theoretical systems, Stevens's own forays into the realm of theory were few and limited in scope. He seemed happiest examining his data (as well as those of others) to see what useful information could be extracted. Much of his time in the laboratory was spent plotting and replotting data in the pursuit of clarity and simplicity, and his research reports were models of both qualities.

5. *Impact on Experimental Psychology*

Two additional characteristics of his work deserve mention. While the majority of his contemporaries were engaged in a search for relevant variables (e.g., does 'meaningfulness' affect ease of remembering?), Stevens's goal was the discovery of functional relations, mathematical descriptions of the association of one factor with another, as epitomized by his psychophysical power law. A corollary of this belief was an attitude toward sampling theory and statistical testing that ranged from indifference to disdain. The object for him was not to see whether 80 dB tones received an average loudness judgment greater than did 60 dB tones (at the 5 percent level of significance), the equivalent of what most of his contemporaries were doing. What was important to Stevens was whether a lawful relation could be seen when data were plotted in a thoughtful way. Large numbers of experimental subjects were not needed to reveal the kind of robust relations he sought; a few colleagues or students recruited in the hallways were enough for most of his purposes.

A landmark publication for research psychologists in the 1950s was the *Handbook of Experimental Psychology* edited by Stevens (1951). It was, as he described it in the preface, 'a technical survey that would systematize, digest, and appraise the mid-century state of experimental psychology.' It was not the first of its kind, but it was the most recent, and thus defined for teachers and students alike the set of shared interests that constituted experimental psychology. It was a time still when one might be expert in one aspect of the field and fairly knowledgeable about many others. Since then the amount of knowledge in psychology has grown at an explosive rate, and now the achievement of competence in one area increasingly restricts knowledge of any other. At the beginning of the twenty-first century, the tent of 'experimental psychology' has few remaining occupants; there are instead cognitive scientists, brain researchers, vision scientists, and a host of others in specialized niches. Just as Stevens's mentor E. G. Boring may have been among the last to teach an all-inclusive account of psychology in a single course, so Stevens may have been among the last to view the phrase 'experimental psychology' as a useful rubric.

6. *Conclusion*

Stevens can be seen as a transitional figure in the evolution of psychological science, yet much of what he did continues to influence the way it is practiced today. His lasting contributions to the fields of psychoacoustics and psychophysics have been discussed here. But, in addition, he created a roadmap for much of modern sensory neurophysiology and for those who theorize in that area. Many of the phenomena for which today's brain researchers seek explanations were established by Stevens or by others using the methods he developed. Were he alive today, it seems likely that he would embrace the new technology and would be developing ingenious techniques for exploring the neural aspects of sensory events. Although an account of his influence remains an unfinished story, at the least he will be remembered as one who brought new and powerful ideas to the study of psychophysics, leaving it a field reinvigorated and evolving as a subject of great intrinsic interest and many useful applications (e.g., see Baird 1997, Norwich 1993, Stevens 1966).

See also: Audition and Hearing; Experimentation in Psychology, History of; Galton, Sir Francis (1822–1911); Helmholtz, Hermann Ludwig Ferdinand von (1821–94); Lashley, Karl Spencer (1890–1958); Psychology: Historical and Cultural Perspectives; Psychology: Overview; Psychophysics; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Baird J C 1997 *Sensation and Judgment: Complementarity Theory of Psychophysics*. Erlbaum, Mahwah, NJ
Binder A 1984 Restrictions on statistics imposed by method of measurement: Some reality, much mythology. *Journal of Criminal Justice* **12**: 467–81
Luce R D, Narens L 1999 S. S. Stevens, measurement, and psychophysics. In: Killeen P R, Uttal W R (eds.) *Fechner Day 99: The End of Twentieth Century Psychophysics*. The International Society for Psychophysics, Tempe, AZ
Miller G A 1974 Stanley Smith Stevens: 1906–1973. *American Journal of Psychology* **87**: 279–88
Norwich K H 1993 *Information, Sensation, and Perception*. Academic Press, San Diego, CA
Scharf B 1999 S. S. Stevens's contributions to an understanding of hearing. In: Killeen P R, Uttal W R (eds.) *Fechner Day 99: The End of Twentieth Century Psychophysics*. The International Society for Psychophysics, Tempe, AZ
Stevens S S 1939 Psychology and the science of science. *Psychological Bulletin* **36**: 221–63
Stevens S S 1946 On the theory of scales of measurement. *Science* **103**
Stevens S S (ed.) 1951 *Handbook of Experimental Psychology*. Wiley, New York
Stevens S S 1957 On the psychophysical law. *Psychological Review* **64**: 153–81
Stevens S S 1961 To honor Fechner and repeal his law. *Science* **133**: 80–6
Stevens S S 1966 A metric for the social consensus. *Science* **151**: 530–41
Stevens S S 1969 On predicting exponents for cross-modality matches. *Perception & Psychophysics* **6**: 251–6
Stevens S S 1975 *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. Wiley, New York
Stevens S S, Davis H 1938 *Hearing: its Psychology and Physiology*. Wiley, New York

R. Teghtsoonian

Stigler, George Joseph (1911–91)

1. Personal History

George Stigler, one of the great economists of all time, with a gift for writing matched among modern economists only by John Maynard Keynes, was born January 17, 1911 in Renton, Washington, a suburb of Seattle, the only child of Joseph and Elisabeth Stigler. His parents had separately migrated to the United States at the end of the nineteenth century, his father from Bavaria, his mother from what was then Austria-Hungary. Stigler went to public schools in Seattle, and then to the University of Washington, receiving a BA in 1931. Although he says that he had 'no thought of an academic career' when he graduated from college, he applied for and was awarded a fellowship at Northwestern University for graduate study in the business school, receiving an MBA in 1932 (Stigler 1988, p.15). More important, at Northwestern he developed an interest in economics as an object of study in its own right and decided on an academic career. After one further year of graduate study at the University of Washington, he moved to the University of Chicago where he found the intense intellectual atmosphere he sought. Chicago became his intellectual home for the rest of his life, as a student from 1933 to 1936, a faculty member from 1958 to his death in 1991, and a leading member of, and contributor to, the 'Chicago School' throughout. He received his PhD in 1938, with a thesis, now a classic, on the history of neoclassical theories of production and distribution (Stigler 1941).

At Chicago, Stigler was particularly influenced by Frank H. Knight, under whom he wrote his dissertation, Jacob Viner, who taught economic theory and international economics, John U. Nef, economic historian, and their younger colleague, Henry Simons, who became a close personal friend and whose *A Positive Program for Laissez Faire* greatly influenced many of Stigler's contemporaries.

'At least as important to me,' Stigler writes (1988, pp. 23–5), 'as the faculty were the remarkable students I met at Chicago,' and goes on to list W. Allen Wallis, Milton Friedman, Kenneth Boulding, and Robert Shone from Britain, Sune Carlson from Sweden, Paul Samuelson, and Albert G. Hart—all of whom subsequently had distinguished careers.

In 1936, Stigler was appointed an assistant professor at Iowa State College (now University), and shortly thereafter was married to Margaret Mack. The Stiglers had three sons: Stephen, a professor of statistics at the University of Chicago, David, a lawyer in private practice, and Joseph, a businessman. Mrs. Stigler died in 1970, tragically without any advance warning. George never remarried.

In 1938, Stigler accepted an appointment at the University of Minnesota, going on leave in 1942 to work first at the National Bureau of Economic Research and then at the Statistical Research Group of Columbia University, where he joined a group directed by W. Allen Wallis that was engaged in war research on behalf of the armed services. When the war ended in 1945, George returned to the University of Minnesota, but remained only one year, leaving in 1946 to accept a professorship at Brown University, where he also remained only one year before moving to Columbia University in 1947 and then, in 1958, to the University of Chicago where he remained the rest of his life. At Chicago, he became an editor of the *Journal of Political Economy*, established the Industrial Organization Workshop, which achieved recognition as the key testing ground for contributions to the field of industrial organization, and in 1977 founded the Center for the Study of Economy and the State, serving as its director until his death in Chicago on December 1, 1991.

Stigler served as President of the American Economic Association in 1964, and of the History of

Economics Society in 1977. He was elected to the National Academy of Sciences in 1975. He received the Alfred Nobel Memorial Prize in Economic Science in 1982 and the National Medal of Science in 1987.

Stigler's governmental activities included: serving as a member of the Attorney General's National Committee to Study the Antitrust Laws, 1954–5; Chairman, Federal Price Statistics Review Committee, 1960–1; member, Blue Ribbon Panel of the Department of Defense, 1969–70; Vice Chairman, Securities Investor Protection Corporation, 1970–3; Co-Chairman, Blue Ribbon Telecommunications Task Force, Illinois Commerce Commission. 1990–1.

2. *Scientific Contributions*

Stigler made important contributions to a wide range of economic areas: (a) history of thought, (b) price theory, (c) economics of information, (d) organization of industry, and (e) theory of economic regulation. In addition, (f) he wrote many influential essays and reviews, and was a long-time editor.

2.1 *Intellectual History*

Intellectual history was Stigler's first field of specialization and remained a continuing interest throughout his life. His doctoral dissertation, published as *Production and Distribution Theories*, was followed by a steady flow of perceptive, thoughtful, and beautifully written articles and books interpreting the contributions of his predecessors. Almost single-handedly, he kept the field alive, despite widespread lack of interest within the profession. More recently, he succeeded in bringing about a revival of interest in the field, which is currently flourishing. In the words of one of his students, 'The versatility of his work in [the history of economic thought] has been matched by its depth and insight—and unmatched by any other historian of economic thought' (Sowell 1987, p. 498; the rest of this paragraph draws from Sowell's excellent summary of 'Stigler as an Historian of Economic Thought'). In addition to analyzing the theories of particular economists, he examined particular concepts, such as perfect competition and utility, traced trends in the economics profession, considered the influence of economic theories on events and vice-versa, and the meaning of originality. Many of his important articles in this area are collected in a 1965 volume, *Essays in the History of Economics*.

The history of thought provided a rich seedbed for Stigler's wide-ranging scientific work. Stigler regarded economic theory, in the words of Alfred Marshall, as 'an engine for the discovery of concrete truth,' not as a subject of interest in its own right, a branch of mathematics. Few economists have so consistently and successfully combined economic theory with empirical analysis, or ranged so widely.

2.2 *Price Theory*

Stigler's first important publication after his doctoral thesis was a textbook, *The Theory of Competitive Price*, published in 1942, followed by four revised versions under the title *The Theory of Price* (1946, 1952, 1966, 1987). His textbook is unique among intermediate textbooks in price theory in its systematic linking of highly abstract theory to observable phenomena. It is unique also in its concise yet rigorous exposition.

One of his earliest papers (1939) broadened the concept of a short-run cost curve to accommodate the observation that firms faced with fluctuating demand could build flexibility into their fixed equipment at a cost of not achieving the lowest possible cost for a particular output.

Another early paper (1945) presented a formal analysis of determining the minimum cost of an adequate diet, supplemented by an empirical calculation of the minimum cost—one of the earliest applications of linear programming. 'The procedure,' he explained, 'is experimental because there does not appear to be any direct method of finding the minimum of a linear function subject to linear conditions' (p. 310). Two years later George Dantzig provided such a direct method, the simplex method, now widely used in many economic and industrial applications (1987).

Like these two examples, Stigler's many later contributions to price theory were a by-product of seeking to understand puzzling real world phenomena, and are best discussed in that context.

2.3 *Economics of Information*

'The Economics of Information' is the title of a seminal article, published in 1961. In his intellectual autobiography, Stigler termed it, 'My most important contribution to economic theory' (1988, pp. 79–80). The article begins, 'one should hardly have to tell academicians that information is a valuable resource: knowledge is power. And yet it occupies a slum dwelling in the town of economics. Mostly it is ignored' (Stigler 1961, p. 213).

This brief 12-page article is a splendid illustration of several of Stigler's unique virtues: creativity, which he defined as consisting 'of looking at familiar things or ideas in a new way'; the capacity to extract new insights about those seemingly familiar things; and the ability to state his main points in a provocative and eminently readable way.

Stigler illustrates the importance of subjecting information to economic analysis with two familiar observations—that there is generally no single 'market price' but a variety of prices that 'various sellers (or buyers) quote at any time' and that 'advertising … is an enormously powerful instrument for the elimination of ignorance … and is normally "paid" for by the seller.' In each case, he presents a sophisticated analysis that explains what determines the observed phenomena, an analysis that has empirical implications capable of being contradicted by observation. He concludes that 'The identification of sellers and the discovery of their prices are only one sample of the vast role of the search for information in economic life' (quotes from Stigler 1988, pp. 213, 220, 221, 224).

As occurred throughout his career in every field Stigler touched, this article was less important for the particular familiar observations that it analyzed than for giving birth to an essentially new area of study for economists. There followed an outpouring of further studies by him but even more by his students and others that have radically altered the accepted body of economic theory.

This aspect of his influence was stressed in the citation for his Nobel Prize, which noted that Stigler's 'studies of the forces which give rise to regulatory legislation have opened up a completely new area of economic research … Stigler is also recognized as the founder of "economics of information" and "economics of regulation," and one of the pioneers of research in the intersection of economics and law.'

2.4 *The Organization of Industry*

The Organization of Industry is the title of a book published in 1968. Its 'main content,' Stigler says in the preface, 'is a reprinting of 17 articles I have written over the past two decades [including the "Economics of Information"] in the area of industrial organization. Although the main topics in industrial organization are touched upon, the touch is often light. The ratio of hypotheses to reasonably persuasive confirmation is distressingly high in all of economic literature, and it must be my chief and meager defense that I am not the worst sinner in the congregation.' Far from being the worst sinner, Stigler's main contribution to the field, both in this book and later writing, was the use of empirical evidence to test hypotheses designed to explain features of industrial organization. Article after article combines subtle theoretical analysis with substantial nuggets of empirical evidence, presented so casually as to conceal the care with which the data were compiled and the effort that was expended to determine what data were both relevant and accessible. These articles record the shift in Stigler's views on antitrust—from initial support of an activist antitrust policy to skepticism about even a minimalist policy.

One of the most influential of the articles in the book was on 'The Economies of Scale' in which he introduced 'the survivor principle' as a way to use market data to determine the efficient size of an enterprise in a particular industry. The approach has since become widely accepted.

In 1962, he published an article, in collaboration with Claire Friedland, his long-time associate, entitled 'What Can Regulators Regulate? The Case of Electricity,' which concluded that regulation of electric utilities had produced no significant effect on rates charged. Schmalensee (1987, p. 499) notes that 'This essay stimulated a great deal of important empirical work on public regulation in the USA. It also posed a basic problem: if regulation does not generally achieve its stated objectives, why have so many agencies been established and kept in existence?'

2.5 *Economic Regulation*

'The Theory of Economic Regulation' (1971b) attempts to answer that question. It is one of a series of contributions by Stigler to what has since come to be called 'public choice' economics: the shift from viewing the political market as not susceptible to economic analysis, as one in which disinterested politicians and bureaucrats pursue the 'public interest,' to viewing it as one in which the participants are seeking, as in the economic market, to pursue their own interests, and hence subject to analysis with the usual tools of economics. The seminal work that deserves much of the credit for launching 'public choice,' *The Calculus of Consent*, by James Buchanan and Gordon Tullock (1962), appeared in the same year as the Stigler–Friedland article.

The 'central thesis of the article,' Stigler wrote, 'is that, as a rule, regulation is acquired by the industry and is designed and operated primarily for its benefit.' He notes that two 'alternative views of the regulation of industry are widely held. The first is that regulation is instituted primarily for the protection and benefit of the public at large or some large subdivision of the public. The second view is essentially that the political process defies rational explanation.' He then gives example after example to support his own thesis, which is, by now, the orthodox view in the profession, concluding: 'The idealistic view of public regulation is deeply imbedded in professional economic thought. The fundamental vice of such a [view] is that it misdirects attention'—to preaching to the regulators rather than changing their incentives (quotes from Stigler 1971b, p. 3).

'Smith's Travels on the Ship of State' (1971a), published in the same year as 'The Theory of Economic Regulation,' raises the same question on a broader scale. Smith gives self-interest pride of place in analyzing the economic market, but does not give it the same role in analyzing the political market. Smith's

failure to do so constitutes Stigler's main—indeed, nearly only—criticism of the *Wealth of Nations*, that 'stupendous palace erected upon the granite of self interest' (Rosenberg 1993, p. 835). The same theme pervades many of Stigler's later publications.

2.6 *Essays, Editorship, and Reviews*

In addition to his prolific academic publications, Stigler published a great many essays for a broader audience, many given as talks to a wide variety of audiences. The most important are collected in three volumes, *The Intellectuals and the Marketplace* (1963), *The Citizen and the State* (1975), and *The Economist as Preacher* (1982). 'There he (the intelligent layman) will find a potpourri of wit and seriousness blended with a high writing style' (Demsetz 1982, p. 656). 'For 19 years Stigler was a very successful editor of the *Journal of Political Economy*. Under his leadership this journal solidified its high reputation among economists' (Becker 1993, p. 765).

Stigler was also a prolific reviewer of books. His complete bibliography lists 73 reviews in 24 publications ranging from strictly professional, such as the *Journal of Political Economy* (22) and the *American Economic Review* (10) to the popular, such as the *Wall Street Journal* (5), and the *New York Times* (3), and dating from 1939 to 1989.

Stigler's last book, his intellectual autobiography, published in 1988, *Memoirs of an Unregulated Economist*, is a delight to read. As I described it at the time: 'Stigler's memoirs are a gem: in style, in wit, and above all, in substance, they reflect accurately his own engaging personality and his extraordinarily diverse contributions to our science.'

3. *Concluding Comments*

An evaluation of Stigler's scientific contribution would be incomplete without recognizing his role as a teacher and colleague. As teacher, he inspired his students by his own high standards, and instilled a respect for economics as a serious subject concerned with real problems. As one of his students, Thomas Sowell, wrote (1993, p. 788): 'What Stigler really taught, whether the course was industrial organization or the history of economic thought, was intellectual integrity, analytical rigor, respect for evidence—and skepticism toward the fashions and enthusiasms that come and go.'

Stigler supervised many doctoral dissertations at both Columbia and the University of Chicago. According to Claire Friedland, Stigler's associate for many years, he served on more than 40 thesis committees at Chicago, and perhaps 40 more at Columbia, and chaired a considerable fraction of those committees. His students come close to dominating the field of industrial organization.

Stigler was an extremely valuable colleague. He provided much of the energy and drive to the interaction among members of the Chicago economics department, business school, and law school that came to be known as 'The Chicago School.' His Workshop in Industrial Organization was an outgrowth of a law school seminar started by Aaron Director that Stigler cooperated in running when he came to Chicago. His relations were especially close with Aaron Director, Gary Becker, Richard Posner, Harold Demsetz, and myself, enhancing significantly the scientific productivity of all of us.

See also: Economics, History of; Employment and Labor, Regulation of; Firm Behavior; Information, Economics of; Institutional Economic Thought; Law and Economics; Market Areas; Market Structure and Performance; Markets and the Law; Political Economy, History of; Regulation, Economic Theory of; Regulation Theory in Geography; Utilitarian Social Thought, History of

Bibliography

Becker G S 1993 George Joseph Stigler, January 17, 1911–December 1, 1991. *Journal of Political Economy* **101**: 761–7

Buchanan J M, Tullock G 1962 *The Calculus of Consent*. University of Michigan Press, Ann Arbor, MI

Dantzig G B 1987 Linear programming. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics*. Stockton Press, New York, Vol. 3, pp. 203–5

Demsetz H 1982 The 1982 Nobel Prize in economics. *Science* **218**: 655–7

Rosenberg N 1993 George Stigler: Adam Smith's best friend. *Journal of Political Economy* **101**: 833–48

Schmalensee R 1987 Stigler's contributions to microeconomics and industrial organization. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics*. Stockton Press, New York, Vol. 4, pp. 499–500

Sowell T 1987 Stigler as an historian of economic thought. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics*. Stockton Press, New York, Vol. 4, pp. 498–9

Sowell T 1993 A student's eye view of George Stigler. *Journal of Political Economy* **101**: 784–92

Stigler G J 1939 Production and distribution in the short run. *Journal of Political Economy* **47**: 305–27

Stigler G J 1941 *Production and Distribution Theories, 1870 to 1895*. Macmillan, New York

Stigler G J 1942 *The Theory of Competitive Price*. Macmillan, New York

Stigler G J 1945 The cost of subsistence. *Journal of Farming Economy* **27**: 303–14

Stigler G J 1946 *The Theory of Price*, rev. edns. 1952, 1966, 1987. Macmillan, New York

Stigler G J 1961 The economics of information. *Journal of Political Economy* **69**: 213–25

Stigler G J 1963 *The Intellectual and the Market Place, and Other Essays*. Free Press of Glencoe, New York

Stigler G J 1965 *Essays in the History of Economics*. University of Chicago Press, Chicago, IL

Stigler G J 1968 *The Organization of Industry*. Irwin, Homewood, IL
Stigler G J 1971a Smith's travels on the ship of state. *History of Political Economy* **3**: 265–77
Stigler G J 1971b The theory of economic regulation. *Bell Journal of Economics and Management Strategy. Sci.* **2**: 3–21
Stigler G J 1975 *The Citizen and the State: Essays on Regulation*. University of Chicago Press, Chicago, IL
Stigler G J 1982 *The Economist as Preacher, and Other Essays*. University of Chicago Press, Chicago, IL
Stigler G J 1988 *Memoirs of an Unregulated Economist*. Basic Books, New York
Stigler G J, Friedland C 1962 What can regulators regulate? The case of electricity. *Journal of Law and Economics* **5**: 1–16

M. Friedman

Stigma, Social Psychology of

A person who possesses an attribute that risks their full acceptance from others is said to possess a stigma. As a consequence, the individual is stigmatized, reduced in our minds from a whole and usual person to a tainted, discounted one (Goffman 1963). A person with a stigma may be described as possessing a failing, a shortcoming, or a handicap, be viewed as deviant or abnormal, and thus be subject to marginalization or oppression. Thus, stigmatized individuals are the targets of prejudice, stereotyping, and discrimination (see *Intergroup Relations, Social Psychology of*; *Stereotypes, Social Psychology of*). In order to cope with their situation, some stigmatized individuals may make use of a variety of psychological and behavioral coping strategies (see *Coping across the Lifespan*).

There are several categories of attributes in our culture that have been the basis for pervasive stigmatization (e.g., race, ethnicity, religion, physical ability, physical appearance, gender, and sexual orientation). Goffman (1963) described three general categories/types of stigma: abominations of the body (e.g., physical deformities), blemishes of individual character (e.g., criminal behavior, addiction, or sexual orientation), or tribal stigma (e.g., race, nation, religion, etc.). Most characterizations of stigma emphasize its relational nature: an attribute cannot be stigmatizing without another attribute being viewed as more usual or more acceptable. For example, an unusually tall woman may not feel stigmatized in the company of other very tall people, however, she might feel stigmatized in the company of average height people. Depending upon the context, we all have the potential to be stigmatized.

The nature of stigmas may vary across several dimensions. For instance, Jones et al. (1984) described six dimensions: concealability of the stigmatizing attribute, course (duration and permanence) of the stigmatizing attribute, disruptiveness of the stigmatizing attribute in social interactions, esthetic qualities of the stigmatizing trait, origin or cause of the stigmatizing trait, and perilousness of the stigmatizing trait to others. Crocker et al. (1998) argued that two of these traits are the most critical to understanding the experience of a stigmatized person: the visibility (or concealability) of the attribute, and the controllability of the attribute (or the origin, reason for the person's having that attribute; for example, personal choice vs. genetic inheritance).

1. *History of Research on Stigma*

Early research on stigma was very descriptive in nature, bringing the circumstances and concerns of stigmatized individuals to the attention of the social sciences. This early research typically focused on one particular stigmatizing trait (e.g., physical disability, mental illness, or criminal status) in single-study pieces, often conducted by sociologists or clinicians using qualitative methods such as interviews and intensive case studies. It was not until 1963 that this varied literature was integrated into a single theoretical review in Goffman's (1963) book *Stigma: Notes on the Management of Spoiled Identities*. His book provided a unifying theory of stigma, and insights into issues stigmatized people face in terms of their social and personal identity and their management of interactions with nonstigmatized people.

Early theorizing and research can be characterized as attempts to promote sympathy and understanding by emphasizing the perceived pitiable situation of stigmatized individuals. This is illustrated by Lewin's (1941/1948) discussion of self-hatred among Jews and Allport's (1954/1985) discussion of traits due to victimization. As Allport noted 'One's reputation, whether false or true, cannot be hammered, hammered, hammered, into one's head without doing something to one's character.' It was during this phase that the famous 'doll studies' were conducted in an attempt to show the damage to the self-concept caused by prejudice and segregation (e.g., Clark and Clark 1939). The conclusions often drawn from this research were that black children showed self-hatred by preferring white to black dolls whereas no such self-hatred was shown by white children. Although later refuted (see below), these doll studies were used as evidence in the landmark case of Brown vs. Board of education (1954) to illustrate the damage caused by segregation.

The second phase of research can be considered a transitional phase, in that it represents a merger of research on stigma and research on prejudice, which had been largely independently conducted. This merger is illustrated by the second major integrative work on stigma, Jones et al.'s (1984) book *Social Stigma: The Psychology of Marked Relationships*. Jones and his colleagues illustrate clearly the connection between prejudice and stigmatization pro-

viding further analysis of the implication of stigma for social interactions. This phase can be considered in part a result of the civil rights movement in that it was during this time that a new cultural norm emerged as to how to perceive stigmatized group members, providing new rights, higher social standing, and group pride to members of a number of stigmatized groups. During this time more affirmative theories of self-concept and group identity emerged (see *Ethnic Identity, Psychology of*; *Social Identity, Psychology of*) as well as theories about processes by which the stigmatized can protect their self-esteem. It was also during this time of upward social mobility of oppressed group members that the large body of empirical work on intergroup contact and tokenism began.

The latest development in the research on stigma can be termed 'The Target's Perspective' that has extended the research on stigma in two major ways. First, it broadens our view of the stigmatized. It highlights the importance of understanding how stigmatized individuals perceive prejudice (Swim et al. 1998) and the reasons for their responses to prejudice, which might be misinterpreted if not viewed within the larger context of their life experiences (Miller and Myers 1998). It also highlights understanding variations within stigmatized groups such as variations in extent and type of group identity (Cross 1991). Moreover, the target's perspective research argues for a more empowering view of stigmatized individuals, not as passive recipients of prejudice but as active copers (Swim et al. 1998). Second, the target's perspective research goes beyond simply merging the literatures on stigma and prejudice, instead arguing that the study of the target's experience with prejudice should be used to enlighten the larger theory concerning prejudice and its perpetrators. The target's perspective researchers have asserted that the research on prejudice primarily has studied perpetrators, failing to give a voice to the stigmatized.

2. *Current Understanding of Stigma*

Research on stigma can be conceived of as representing three primary topic areas: (a) perceptions of prejudice and discrimination; (b) effects of prejudice and discrimination on the stigmatized; and (c) coping with prejudice and discrimination. Below we highlight illustrative and consequential work in these areas.

2.1 *Perceptions of Prejudice and Discrimination*

Perceptions of prejudice include general perceptions of the prevalence of prejudice against specific groups or oneself and whether people label specific incidents that occur to themselves or others as discriminatory. An intriguing finding is that people tend to perceive that they personally are less likely to experience prejudice than are members of their group (Taylor et al. 1994). This finding spurred research examining motivational factors (e.g., a desire to not accuse others or to protect sense of personal control) and cognitive factors (e.g., difficulties in aggregating information) influencing stigmatized members to deny their personal experiences with prejudice. Consistent with a denial explanation for the personal/group discrepancy, research has revealed that members of stigmatized group are more likely than members of nonstigmatized groups to underestimate the likelihood that specific ambiguous negative incidents are a function of discrimination against them (Rugerrio and Major 1998).

Several factors influence the likelihood that people will perceive prejudice and discrimination. First, individual differences in sensitivity to prejudice and discrimination influence perceptions. For example, the more people endorse sexist beliefs, the less likely they are to label certain incidents as sexual harassment (Swim and Cohen 1997) and African-Americans who have a general cultural distrust may more readily perceive prejudice than those who are more trusting (Feldman Barrett and Swim 1998). Second, several features of particular incidents influence perceptions. Some beliefs and actions are more readily defined as prejudicial. For example, women are more likely to favorably rate a man who endorses benevolent sexist beliefs than hostile sexist beliefs perhaps in part because they do not recognize the link between these two types of sexist beliefs (Kilianski and Rudman 1998). Benevolent sexist beliefs include the endorsement of complementary gender differentiation, heterosexual intimacy, and paternalism and men who tend to endorse these beliefs tend to also endorse hostile sexist beliefs. Also, some people are more likely to be perceived to be prejudiced than others are. That is, people hold stereotypes about stereotypers such as believing that whites are more likely to discriminate against blacks than blacks are (Inman and Baron 1996). Third, there are personal costs and benefits to perceiving prejudice (Feldman Barrett and Swim 1998). Perceiving prejudice when none existed can lead to interpersonal disruption and distrust in outgroup members as well as to confirmatory biases (see *Self-fulfilling Prophecies*) and not perceiving prejudice when it does exist can endanger an individual's self-esteem and can lead to internalization of prejudicial or stereotypical beliefs.

2.2 *Effects of Prejudice and Discrimination on the Stigmatized*

As noted above, there has been much concern that prejudice psychologically impacts the stigmatized. Perhaps the best-known early set of empirical work is the previously mentioned doll studies originating in

work by Clark and Clark (1939). The conclusion that these studies illustrate self-hatred among African-Americans has been disputed (e.g., Cross 1991) and countered by findings illustrating, for example, that African-Americans have higher self-esteem than European Americans as measured on standardized measures of self-esteem (Gray-Little and Hafdahl 2000). Despite these findings, it is important to recognize that some stigmatized groups (e.g., heavy women) may be more likely to have lower self-esteem than nonstigmatized people, which can be a function of internalizing prejudice (Quinn and Crocker 1998). There also is variation within stigmatized groups with some group members, such as those that do not strongly identify with a group, being more effected by prejudice than others (Gray-Little and Hafdahl 2000). Further, experiencing discrimination can lead to other psychological effects such as increased negative daily moods (Swim et al. 2000) and long-term psychiatric symptoms such as obsessive-compulsiveness, depression, and anxiety (Klonoff et al. 2000).

Decreases in academic performance are also important consequences of being in a stigmatized group. Compared with the nonstigmatized, stigmatized individuals may more often find themselves solo members of groups which can lead to distraction due to impression management concerns that can interfere with performance (Saenz and Lord 1989). Further, situations can induce stigmatized individuals to fear that others will stereotype them and this 'stereotype threat' can impair test performance (Crocker et al. 1998). For instance, telling women that there are gender differences on a difficult exam in a stereotype relevant domain of importance to the individual is sufficient to result in decrements in exam scores. Similar findings have occured for other groups, such as African–Americans.

2.3 Coping with Prejudice and Discrimination

Coping strategies used by stigmatized individuals can be divided into psychological and behavioral responses. In many cases, coping strategies protect people from the negative effects of prejudice and discrimination and have potential costs. Disengagement or disidentification is a possible psychological response to discrimination that can allow one to not attend to potentially demeaning feedback from others (Crocker et al. 1998). However, this can result in disidentification with domains that might be useful for promoting one's academic success. One type of behavioral response is preventative. Stigmatized individuals may choose to avoid situations where they feel that they may encounter prejudice (Swim et al. 1998) or to engage in behaviors that might compensate for the negative expectations that others might have of them (Miller and Myers 1998). Although these prevention strategies may protect the individual from future encounters with prejudice, they too have their costs. Avoidance can cause people to miss opportunities, prevent disconfirmation of beliefs, and prevent the development of meaningful relationships. People can overcompensate for prejudice, which may appear out of place or fake, or they may undercompensate particularly if they are in situations where they do not feel threatened. Other behavioral responses occur after an event and these responses can be both nonassertive and assertive (Swim et al. 1998). People may choose to politely confront a person, for instance, by asking the perpetrator to explain themselves, by attempting to educate the perpetrator, or by engaging in collective action against the perpetrator. While these responses may result in change and may leave the target feeling more empowered, there are potential costs, such as retaliation on the part of the perpetrator especially with the more assertive responses.

3. Future Directions

As noted above, current research on stigma has broadened our understanding of the stigmatized and we anticipate that this process will continue to occur. For instance, there are still relatively few studies that illustrate the strengths of the stigmatized. However, we anticipate that this will change with more people focusing, for instance, on how the stigmatized support each other, confront perpetrators, and set and seek their life goals. Further, we anticipate that there will be more research on variations in experiences across stigmatized groups, such as differences in experiences of those from different Asian countries or differences in identity and economic standing within any stigmatized group. Moreover, there will be more complexity in these analyses as mixed relationships develop where children are born to parents of different ethnic heritages or adopted to parents who do not share, for instance, the same ethnic heritage as their parents.

We also anticipate that research on the stigmatized will continue to inform and be informed by other research areas. For instance, there is a strong connection between the literature on stress and stigma in that experiences with prejudice can be considered forms of stressors. Further research on stigma likely to lead to major changes in how people study prejudice and intergroup relationships. Given greater emphasis on similarities and differences between the stigmatized, there may be greater attention paid to similarities and differences between racism, sexism, heterosexism, and antifatism. There may also be greater recognition in the subjectivity in defining what is and is not prejudice in that members of different groups will likely disagree as to what constitutes prejudicial behavior. Further, there will likely be more attention paid to the contributions that the stigmatized have to intergroup relationships, such as understanding their attempts to

maintain smooth intergroup interactions as well as the origins and characteristics of their prejudices against both dominant groups and other stigmatized groups.

See also: Discrimination; Discrimination: Racial; State, Sociology of the; Stereotypes, Social Psychology of

Bibliography

Allport G W 1985 [1954] *The Nature of Prejudice*. Addison-Wesley, Reading, MA

Brown vs. Board of Education of Topeka 1954 347 US 483

Clark K P, Clark M P 1939 Segregation as a factor in the racial identification of Negro pre-school children: A preliminary report. *The Journal of Experimental Education* **8**: 161–3

Crocker J, Major B, Steele C 1998 Social stigma. In: Gilbert D, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, Vol. 2, pp. 504–53

Cross W E 1991 *Shades of Black: Diversity in African–American Identity*. Temple University Press, Philadelphia

Feldman Barrett L, Swim J K 1998 Appaisals of prejudice and discrimination. In: Swim J K, Stangor C (eds.) *Prejudice: The Target's Perspective*. Academic Press, San Diego pp. 11–36

Goffman E 1963 *Stigma: Notes on the Management of Spoiled Identities*. Prentice-Hall, Englewood Cliffs, NJ

Gray-Little B, Hafdahl A R 2000 Factors influencing racial comparisons of self-esteem: A quantitative review. *Psychological Bulletin* **126**: 26–54

Inman M L, Baron R S 1996 The influence of prototypes on perceptions of prejudice. *Journal of Personality and Social Psychology* **70**: 727–39

Jones E E, Farina A, Hastorf A H, Markus H, Miller D, Scott R A 1984 *Social Stigma: The Psychology of Marked Relationships*. Freeman, New York

Kilianski S E, Rudman L A 1998 Wanting it both ways: Do women approve of benevolent sexism? *Sex Roles* **39**: 333–52

Klonoff E A, Landrine H, Compbell R 2000 Sexist discrimination may account for well-known gender differences in psychiatric symptoms. *Psychology of Women Quarterly* **24**: 93–9

Lewin K 1948 [1941] Self-hatred in Jews. In: Lewin G W (ed.) *Resolving Social Conflicts—Selected Papers on Group Dynamics*. Harper & Brothers, New York, pp. 186–200

Miller C T, Myers A M 1998 Compensating for prejudice: How heaveyweight people (and others) control outcomes despite prejudice. In: Swim J K, Stangor C (eds.) *Prejudice: The Target's Perspective*. Academic Press, New York, pp. 191–219

Quinn D, Crocker J 1998 Vulnerablitiy to the affective consequences of the stigma of overweight. Experiencing everyday prejudice and discrimination. In: Swim J K, Stangor C (eds.) *Prejudice: The Target's Perspective*. Academic Press, San Diego, pp. 127–48

Ruggiero K M, Major B N 1998 Group status and attributions to discrimination: Are low- or high-status group members more likely to blame their failure on discrimination? *Personality & Social Psychology Bulletin* **24**: 821–37

Saenz D S, Lord C G 1989 Reversing roles: A cognitive strategy for undoing memory deficits associated with token status. *Journal of Personality & Social Psychology* **56**: 698–708

Swim J K, Cohen L L 1997 Overt, covert, and subtle sexism: A comparison between the attitudes toward women and modern sexism scales. *Psychology of Women Quarterly* **21**: 103–18

Swim J K, Cohen L L, Hyers L L 1998 Experiencing everyday prejudice and discrimination. In: Swim J K, Stangor C (eds.) *Prejudice: The Target's Perspective*. Academic Press, San Diego, pp. 38–61

Swim J K, Hyers L L, Cohen L L, Ferguson M J 2000 Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social Issues* **57**

Taylor D M, Wright S C, Porter L E 1994 Dimensions of perceived discrimination: The personal/group discrimination discrepancy. In: Zanna M P, Olson J M (eds.) *The Psychology of Prejudice: The Ontario Symposium*. Erlbaum, Hillsdale, NJ, Vol. 7, pp. 233–56

J. K. Swim and L. L. Hyers

Stochastic Dynamic Models (Choice, Response, and Time)

The term 'stochastic dynamic model' is used in human experimental psychology to describe models of choice and reaction time (RT) in which noise-perturbed information about stimulus identity accrues in continuous time. It has gained currency since it was used by Luce (1986) to refer to a class of models first investigated in a pioneering series of articles by Pacut (e.g., 1976), in which information accrual was described by linear stochastic differential equations (SDEs). Because of the relatively rich dynamics that can be represented by equations of this kind, the term is also used in a more general sense to describe models in which informational accrual is not constant, but varies with time or state. This entry describes the mathematical methods that are used to analyze models of this kind.

Like the classical sequential-sampling models (see *Auditory Models*, *Sequential Decision Making*), the main application of stochastic dynamic models is to performance in simple decision or simple information-processing tasks, typically, those that require a speeded choice between two alternatives. Along with Signal Detection Theory and its variants (see *Signal Detection Theory: Multidimensional*), these models are members of a larger class of statistical decision models, which assume that the representation of stimuli in the central nervous system is inherently statistical, by virtue of the mass action properties of neural coding. This leads to models in which the stimulus representation, viewed as a sequence of events unfolding in time, is treated formally as a stochastic process.

Because of the assumed noisiness of stimulus representations, statistical decision models typically assume that decisions are made by an averaging or integration device, whose function is to smooth out transitory fluctuations in neural activation, thereby rendering decisions less susceptible to random perturbations in

the stimulus process than would otherwise be the case. In Signal Detection Theory, the period of integration is assumed to be fixed, or to be determined by factors external to the stimulus process, whereas in sequential-sampling models integration continues until a criterion quantity of evidence for a particular response is obtained.

From a theoretical standpoint, sequential-sampling models have greater explanatory power than signal-detection models, because they can account for choice probabilities and RT within a unified theoretical framework, whereas signal-detection models provide an account of choice probabilities only. However, sequential-sampling models are more complex to analyze mathematically because they require solution of a first passage time problem. This solution gives the distribution of times needed for the accrued evidence to attain a criterion level, which is identified in such models with the decision component of RT. In contrast, the analysis of signal detection models requires only the mean and variance of the accrued evidence as a function of time, as these suffice to determine the choice probabilities.

In the sequential-sampling literature, two main classes of model have predominated: random-walk models, in which evidence accrues as a single, signed total and counter, or accumulator models, in which evidence for competing response alternatives accrues in parallel. When formulated in continuous time, these models become diffusion process models and Poisson counter models, respectively. Although a number of researchers have contributed to the development of these models, the standard versions are those described, respectively, by Ratcliff (1978) and Townsend and Ashby (1983).

Recently, dynamic generalizations of these 'classical' models have begun to appear, in which the assumption of a constant rate of information accrual is replaced by time-dependent accrual, state-dependent accrual, or some combination of the two. These assumptions are motivated by the need to capture the complex dynamics that seem to underlie performance in some domains. State dependence can arise when the accrual process is imperfect or 'leaky' (Diederich 1995, Rudd and Brown 1997, Smith 1995), or when it is governed by resistive, or approach-avoidance, dynamics (Busemeyer and Townsend 1993).

Time dependence arises when, for any of a number of reasons, the physical mechanism that generates the stimulus process is itself time varying (Heath 1992). Examples of the latter include models in which the accrual process is driven by time-varying encoding mechanisms in the visual cortex or by a noisy, continuous-time neural network, and models in which information accrual is gated in time by selective attention. In all such cases, the theoretical principles that underlie the model lead naturally to the assumption of time-dependent or state-dependent information accrual.

1. Diffusion Process Models

Figure 1 shows a two-barrier, dynamic, diffusion process model of RT. The main elements of this model are an encoding function, a source of noise, and a decision stage. The encoding function, $\mu(t)$, is a deterministic function of time, which represents the average information available to the decision stage at time t. ('Information' is used loosely here, to refer to the output of relevant neural systems, rather than in its technical sense of entropy.) In general, the value of $\mu(t)$ is assumed to reflect the results of an ongoing matching operation between perceptual and memorial representations of a stimulus. In Fig. 1, the encoding function arises as the difference between a time-varying stimulus representation (the continuous curve) and a sensory referent or 'mental standard,' c (the dashed line), which results in a function that takes on both positive and negative values. Positive values are evidence for one response, r_A; negative values are evidence for the alternative response, r_B. In classical sequential-sampling models, the encoding function is assumed to be constant for a given experimental trial ($\mu(t) = \mu$), but in general dynamic models it may vary with time.

At each instant, t, the encoding function is additively perturbed by noise, $W(t)$. In diffusion process models, the noise is assumed to be white noise, that is, broad-spectrum (delta correlated) Gaussian noise. To make a decision, successive values of the resulting noisy comparison process are integrated, as shown in the box on the right of Fig. 1. The value of the integral as a function of time is a stochastic process, $X(t)$, which represents the accrued stimulus information at time t. The process of accrual continues until $X(t)$ exits from the region bounded by two absorbing barriers $a_A(t)$ and $a_B(t)$.

If, at the time of the first boundary crossing, $X(t) \geqslant a_A(t)$, then response r_A is made; if $X(t) \leqslant a_B(t)$, then response r_B is made. Psychologically, the time of the first boundary crossing is identified with the

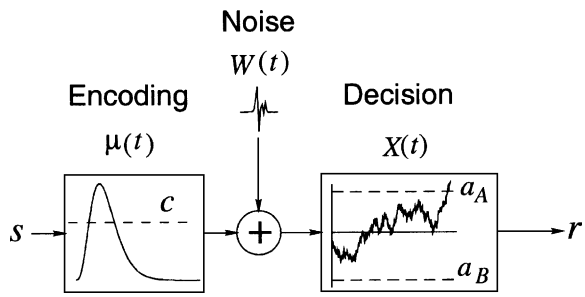

Figure 1
Dynamic diffusion process model showing cognitive operations intervening between presentation of an external stimulus (s) and production of a binary identification response (r)

decision time component of RT, and the probability that the first boundary crossed is $a_A(t)$ is identified with the proportion of r_A responses. Although the figure shows $a_A(t)$ and $a_B(t)$ as constant, in general dynamic models the absorbing boundaries and the encoding function may all depend on time. Such variations may be useful when modeling the response to external time constraints, such as RT deadlines.

1.1 Stochastic Differential Equations for Information Accrual

In general, the process of information accrual in stochastic dynamic models can be described by a linear SDE of the form

$$dX(t) = [\mu(t) + b(t)X(t)]\,dt + [\sigma(t) + c(t)X(t)]\,dB(t) \quad (1)$$

This Eqn. defines $dX(t)$, the random change in the accrual process $X(t)$ in a small time interval dt. The change consists of two parts, one deterministic, the other stochastic. The deterministic change, or drift, is expressed by the quantity $\mu(t) + b(t)X(t)$. The first term $\mu(t)$, which is purely time dependent, is determined by the temporal properties of the encoding process. The second term $b(t)X(t)$ expresses a joint dependency on time and on the current (random) state. Pure state dependency arises in the special case in which $\mu(t) = 0$ and $b(t)$ is constant.

The stochastic part of the change, the so-called diffusion coefficient, is represented by the second term on the right of Eqn. 1. Stochastic perturbation is introduced via the white noise process, $dB(t)$, which is the formal derivative of the Brownian motion or Wiener diffusion process. The stochastic perturbation at any instant can also be represented as the sum of two terms, one purely time dependent, $\sigma(t)$, and another, $c(t)X(t)$, that is jointly dependent on time and state. Under appropriate regularity conditions, the process $X(t)$ that solves Eqn. 1 is a diffusion process, that is, a continuous sample path Markov process. Solutions to linear stochastic differential equations of the form of Eqn. 1 are obtained by using the Itô calculus (e.g., Karatzas and Shreve 1991).

In applications, a number of special cases of Eqn. 1 are of interest. These are instances of the case $c(t) = 0$, in which there is no state dependence in the diffusion coefficient. Under these circumstances, the solution to Eqn. 1 is a Gaussian process, $X(t)$, whose distribution at time t can be characterized uniquely by its first two moments. By means of the Itô calculus it may be shown that, for a process with initial condition $X(\tau) = y$, the value of $X(t)$ is normally distributed with mean

$$E[X(t)\,|\,X(\tau) = y] = \int_\tau^t \exp\left[\int_s^t b(u)\,\mathrm{d}u\right]\mu(s)\,\mathrm{d}s + y\exp\left[\int_\tau^t b(u)\,\mathrm{d}u\right] \quad (2)$$

and variance

$$\mathrm{var}[X(t)|X(\tau) = y] = \int_\tau^t \exp\left[2\int_s^t b(u)\,\mathrm{d}u\right]\sigma^2(s)\,\mathrm{d}s. \quad (3)$$

These eqns. provide the basis for a dynamic signal-detection model that generalizes the classical model of Signal Detection Theory. It extends in a straightforward way to multiple correlated dimensions, to give a dynamic formulation of the General Recognition Theory of Ashby and Towsend (1986).

To date, the most important subcases of the Gaussian process in Eqns. 2 and 3 are those in which $\sigma(t) = \sigma$ (constant) and $b(t) = 0$ or $b(t) = -\gamma$. These subcases represent, respectively, dynamic versions of the Wiener and Ornstein-Uhlenbeck (OU) diffusion process. From a psychological perspective, the Wiener and OU processes may be thought of as describing the dynamics of stochastic accrual processes in which information integration is perfect, or 'leaky,' respectively.

1.2 Solution of First Passage Time Problems for Time-inhomogeneous Diffusions by Numerical Integral Equations

In models of RT such as the one shown in Fig. 1, the quantities of interest are the times required for $X(t)$ to first cross one of the two absorbing boundaries, which represent decision criteria for the responses r_A and r_B. The distribution of these times are obtained by solving the first passage time problem for $X(t)$. Classical methods for solving first passage time problems treat the problem as a boundary value problem in the theory of partial differential equations (e.g., Ratcliff 1978). However, these methods are usually not tractable when the drift term in Eqn. 1 is not time-homogeneous. Instead, such problems can often be solved using numerical integral equation techniques that were developed in mathematical biology to model the properties of integrate-and-fire neurons. These methods yield integral equation representations for $g_A[a_A(t), t\,|\,x_0, t_0]$ and $g_B[a_B(t), t\,|\,x_0, t_0]$, the first passage time densities of the process $X(t)$ through the absorbing boundaries $a_A(t)$ and $a_B(t)$, given the initial condition $X(t_0) = x_0$.

To give a formal characterization of these methods, we define the transition density of $X(t)$, the unconstrained diffusion process, by

$$f(x, t\,|\,y, \tau) = \frac{\mathrm{d}}{\mathrm{d}x}P[X(t) \leqslant x\,|\,X(\tau) = y].$$

That is, the transition density is the partial derivative with respect to the spatial coordinate of the transition distribution $P[X(t) \leqslant x\,|\,X(\tau) = y]$. Classically, the transition distribution is obtained by solving the

partial differential equation that is variously known as the Kolmogorov forward equation or Fokker-Planck equation, with specified initial condition. For the special Gaussian case $c(t) = 0$ of Eqn. 1, it may be obtained directly from the solution given by the Itô calculus, namely,

$$P[X(t) \leqslant x \mid X(\tau) = y] = \Phi \left\{ \frac{x - E[X(t) \mid X(\tau) = y]}{\sqrt{\mathrm{var}[X(t) \mid X(\tau) = y]}} \right\}$$

where Φ is the normal distribution function and where the moments are given by Eqns. 2 and 3.

For an arbitrary diffusion process, which may be both temporally and spatially inhomogeneous, the first passage time densities $g_A[a_A(t), t \mid x_0, t_0]$ and $g_B[a_B(t), t \mid x_0, t_0]$ satisfy the Fortet eqns.

$$f[a_A(t), t \mid x_0, t_0] = \int_{t_0}^{t} g_A[a_A(\tau), \tau \mid x_0, t_0] f[a_A(t), t \mid a_A(\tau), \tau] \mathrm{d}\tau + \int_{t_0}^{t} g_B[a_B(\tau), \tau \mid x_0, t_0] f[a_A(t), t \mid a_B(\tau), \tau] \mathrm{d}\tau \quad (4a)$$

and

$$f[a_B(t), t \mid a_0, t_0] = \int_{t_0}^{t} g_B[a_B(\tau), \tau \mid x_0, t_0] f[a_B(t), t \mid a_B(\tau), \tau] \mathrm{d}\tau + \int_{t_0}^{t} g_A[a_A(\tau), \tau \mid x_0, t_0] f[a_B(t), t \mid a_A(\tau), \tau] \mathrm{d}\tau. \quad (4b)$$

These eqns. express the unknown first passage time densities as functions of the free transition density $f(x, t \mid y, \tau)$ of $X(t)$. They are obtained by considering the sample paths of $X(t)$ that pass through one of two small open intervals $a_A^{\varepsilon}(t) \equiv (a_A(t), a_A(t) + \varepsilon)$ and $a_B^{\varepsilon}(t) \equiv (a_B(t), a_B(t) - \varepsilon)$, on the upper and lower absorbing barriers, respectively, at time t. These intervals are both in the absorbing region for $X(t)$, so any sample path of this kind must have made at least one boundary crossing at some time $\tau, \tau \leqslant t$. It may therefore be decomposed into an initial segment representing a transition from the starting point x_0 to the point of the first boundary crossing at $a_A(\tau)$ or $a_B(\tau)$, followed by a second segment representing the subsequent transition into $a_A^{\varepsilon}(t)$ or $a_B^{\varepsilon}(t)$. Because $X(t)$ is a Markov process, the probability density for such sample paths, obtained by shrinking ε to zero, may be decomposed into a product of two density functions. One describes the first passage through an absorbing boundary at τ; the other describes the subsequent free transition (irrespective of the number of intervening boundary crossings) to $a_A(t)$ or $a_B(t)$. Equations 4a and 4b provide a mutually-exclusive and exhaustive decomposition of the probability densities of sample paths of this kind.

Formally, Eqns. 4a and 4b are Volterra integral equations of the first kind, which can be solved analytically only in special cases. While the general case can in principle be handled numerically, by approximating the integrals with sums, any simple method of approximation will be numerically unstable, because as the interval of approximation, h, becomes small, the transition density $f(x, t \mid y, t-h)$ approaches a Dirac delta function. In the terminology of integral equation theory, the kernel of the equation is said to be singular. Consequently, recent theoretical work has sought to find ways to transform the eqn. to remove the singularity.

It was shown by Buonocore et al. (1990) and subsequently, in somewhat greater generality, by Gutiérrez Jáimez et al. (1995), that for a large class of diffusion processes, Eqns. 4a and 4b can be transformed into Volterra integral equations of the second kind of the form

$$g_A[a_A(t), t \mid x_0, t_0] = -2\Psi[a_A(t), t \mid x_0, t_0] + 2\int_{t_0}^{t} g_A[a_A(\tau), \tau \mid x_0, t_0] \Psi[a_A(t), t \mid a_A(\tau), \tau] \mathrm{d}\tau + 2\int_{t_0}^{t} g_B[a_B(\tau), \tau \mid x_0, t_0] \Psi[a_A(t), t \mid a_B(\tau), \tau] \mathrm{d}\tau \quad (5a)$$

and

$$g_B[a_B(t), t \mid x_0, t_0] = 2\Psi[a_B(t), t \mid x_0, t_0] - 2\int_{t_0}^{t} g_A[a_A(\tau), \tau \mid x_0, t_0] \Psi[a_B(t), t \mid a_A(\tau), \tau] \mathrm{d}\tau - 2\int_{t_0}^{t} g_B[a_B(\tau), \tau \mid x_0, t_0] \Psi[a_B(t), t \mid a_B(\tau), \tau] \mathrm{d}\tau. \quad (5b)$$

These eqns. express the first passage time densities at time t as a function jointly of their values at all preceding times, $\tau < t$ and of a kernel function $\Psi(x, t \mid y, \tau)$. The latter may be chosen in such a way that its value goes to zero as τ approaches t, which means that, unlike Eqns. 4a and 4b, the right-hand sides of Eqns. 5a and 5b can be approximated numerically, with finite sums. A full account of this method, including details of how to obtain the kernel functions for diffusions such as the Wiener and OU processes, may be found in Smith (2000).

1.3 Finite-state Markov Chain Approximations

Another approach to modeling decision-making using diffusion processes has been used by Busemeyer and

Townsend (1992) and collaborators, in which the process is approximated numerically as the limit of a finite-state Markov chain. Computationally efficient solutions to the first passage time problem for such chains can be obtained using spectral methods when the process is time-homogeneous, as in the case of the OU process with a constant comparison function, $\mu(t) = \mu$. More recently, Diederich (1997) has extended this method to the case in which the encoding function is only piecewise time-homogeneous, that is, $\mu(t) = \mu_i = 1, 2, \ldots$, where the index i denotes one of a sequence of consecutive, nonoverlapping observation intervals. This describes a situation in which the encoding function undergoes a sequence of instantaneous shifts during the course of an experimental trial. As such, it resembles models previously considered for piecewise random walks and Wiener processes by Ratcliff (1980) and Heath (1981), respectively. Its main application has been to model shifts in attention between stimulus attributes in cognitive decision tasks, in settings in which the time scale of the decision is long relative to that of the attention shifts. In tasks in which decisions are made more rapidly, as is the case in simple information-processing tasks, it may be more relevant to represent shifts in attention as progressive rather than instantaneous events. The dynamics of such shifts can then be modeled in continuous time using the techniques of the preceding paragraphs.

2. Poisson Counter Models

Figure 2 shows an example of the second class of stochastic, dynamic models, the Poisson parallel-counter. This model consists of three main stages: an encoding stage, a point-process generation stage, and a stochastic accrual stage. The output of the encoding stage is a function $\mu(t)$ that describes the time course of the stimulus representation. This is compared simultaneously to the mental representations of the stimuli that correspond to the response alternatives r_A and r_B. The result of this comparison is a pair of independent point processes, $X_A(t)$ and $X_B(t)$, whose intensities reflect the degree of the match between the stimulus and the associated mental representations. In the classical model, these stochastic comparison processes are assumed to be homogeneous Poisson processes. More generally, however, the intensities of these processes may be arbitrary functions of time, denoted $a(t)$ and $b(t)$. Under these circumstances, $X_A(t)$ and $X_B(t)$ become time-inhomogeneous.

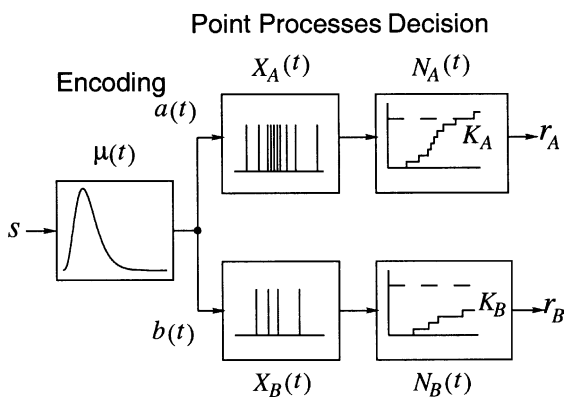

Figure 2
Poisson parallel-counter model with time-inhomogeneous intensities

These processes drive a pair of counting processes, $N_A(t)$ and $N_B(t)$, which count the number of point events in $X_A(t)$ and $X_B(t)$, respectively. The accrual of counts continues in this way until the total in one or another counter exceeds an associated critical count, specifically, until $N_A(t) \geqslant K_A$ or $N_B(t) \geqslant K_B$. At this point, accrual ceases and the response associated with the counter in question is given. The RT density functions for this model are determined by the integrated event rate functions $A(t)$ and $B(t)$, which are defined as

$$A(t) = \int_0^t a(s)\,\mathrm{d}s$$

$$B(t) = \int_0^t b(s)\,\mathrm{d}s.$$

The first passage time densities $g_A(t)$ and $g_B(t)$ may then be written:

$$g_A(t) = \frac{[A(t)]^{K_A-1} a(t) e^{-A(t)}}{(K_A-1)!} \left\{ \sum_{j=0}^{K_B-1} \frac{[B(t)]^j}{j!} e^{-B(t)} \right\} \tag{6a}$$

and

$$g_B(t) = \frac{[B(t)]^{K_B-1} b(t) e^{-B(t)}}{(K_B-1)!} \left\{ \sum_{j=0}^{K_A-1} \frac{[A(t)]^j}{j!} e^{-A(t)} \right\} \tag{6b}$$

respectively. Like the time-inhomogeneous diffusion process, closed-form expressions for choice probabilities and mean RT cannot in general be obtained for this model, although these quantities may always be computed from the RT density functions by numerical integration. However, for some special cases that are of interest in applications, closed-form expressions for these statistics are readily obtainable. Details may be found in Smith and Van Zandt (2000).

3. Models with Multivariate Diffusion Processes

The preceding sections described the simplest versions of the main classes of stochastic dynamic models. However, other alternatives are possible, for example, models in which several diffusion processes race one

another in parallel, as first proposed by Ratcliff (1978). Subsequent authors have considered general dynamic models in a similar vein, such as the model of simple RT described by Smith (1995), which incorporates both time inhomogeneity and state dependence. Providing the constituent processes themselves are independent, such models may be analyzed straightforwardly using the methods of Sect. 1. Usually, however, considerable complexity accompanies any attempt to relax the independence assumption in a useful way.

When the processes are not independent, their combined behavior must be modeled as a multivariate diffusion process with specified covariance properties. Distributional properties for a fairly wide variety of correlated diffusion processes may be obtained by solving the multivariate analogue of Eqn. 1 using the Itô calculus. This suffices to give the time-dependent mean and covariance function of the process, unconstrained by absorbing barriers, which provides the basis for a dynamic signal-detection model. For models of RT, however, it is usually necessary to solve the associated first passage time problem in multivariate form.

One exception to this prescription is the stochastic version of General Recognition Theory developed by Ashby (2000), which seeks to characterize how people make binary classification judgments on stimuli defined in a multivariate perceptual space. In this model, the multivariate problem can be reduced to a univariate problem because, geometrically, all the information relevant to the decision is contained in the projection of the diffusion process onto the normal to the hyperplane dividing the response categories. However, in models not possessing this structure it is necessary to solve the multivariate first passage time problem explicitly. Unfortunately, there are no general methods for solving such problems and the fragmentary results that have been obtained in special cases are usually not applicable to behavioral data. From a modeling perspective, this represents a significant limitation, because in some domains, models in which there is crosstalk between channels are of considerable theoretical interest. In these cases, because of the difficulties in obtaining relevant results in other ways, the properties of the associated models have been investigated using Monte Carlo simulation. A recent example of the use of simulation methods to investigate stochastic dynamic models is the work of Wenger and Townsend (2000).

4. Conclusions and Future Directions

This article has surveyed the mathematical methods used to obtain RT and accuracy predictions for the two main classes of continuous-time, dynamic models used to model decision-making in cognitive tasks: diffusion process models and Poisson parallel-counter models. The attraction of such models is that they allow richer and more complex ideas about the neural representation of stimuli to be embodied in mathematical models. The disadvantage of this additional freedom is that the resulting models may be more difficult to test empirically than their classical counterparts. For this reason, future developments in this area are likely to be in domains in which the dynamic characteristics of a model are strongly constrained by existing theory. Such constraints are needed to ensure models that are expressions of theoretical principles rather than simply vehicles for fitting data. Areas with a body of theory that is sufficiently well articulated to support models of this kind include early vision, mathematical studies of memory, and selective attention.

See also: Majorization and Stochastic Orders; Neural Systems and Behavior: Dynamical Systems Approaches; Self-organizing Dynamical Systems; Stochastic Models

Bibliography

Ashby F G 2000 A stochastic version of general recognition theory. *Journal of Mathematical Psychology* **44**: 310–29

Ashby F G, Townsend J T 1986 Varieties of perceptual independence. *Psychological Review* **93**: 154–79

Buonocore A, Giorno V, Nobile A G, Ricciardi L 1990 On the two-boundary first-crossing-time problem for diffusion processes. *Journal of Applied Probability* **27**: 102–14

Busemeyer J, Townsend J T 1992 Fundamental derivations from decision field theory. *Mathematical Social Sciences* **23**: 255–82

Busemeyer J, Townsend J T 1993 Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review* **100**: 432–59

Diederich A 1995 Intersensory facilitation of reaction time: Evaluation of counter and diffusion coactivation models. *Journal of Mathematical Psychology* **39**: 197–215

Diederich A 1997 Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology* **41**: 260–74

Gutiérrez Jáimez R, Román Román P, Torres Ruiz F 1995 A note on the Volterra integral equation for the first-passage-time probability density. *Journal of Applied Probability* **32**: 635–48

Heath R A 1981 A tandem random walk for psychological discrimination. *British Journal of Mathematical and Statistical Psychology* **34**: 76–92

Heath R A 1992 A general nonstationary diffusion model for two choice decision making. *Mathematical Social Sciences* **23**: 283–309

Karatzas I, Shreve S E 1991 *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York

Luce R D 1986 *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press, New York

Pacut A 1976 Some properties of threshold models of reaction latency. *Biological Cybernetics* **28**: 63–72

Ratcliff R 1978 A theory of memory retrieval. *Psychological Review* **85**: 59–108

Ratcliff R 1980 A note on modeling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology* **21**: 178–84

Rudd M E, Brown L G 1997 A model of Weber and noise gain control in the retina of the toad *bufo marinus*. *Vision Research* **37**: 2433–53

Smith P L 1995 Psychophysically-principled models of visual simple reaction time. *Psychological Review* **102**: 567–93

Smith P L 2000 Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology* **44**: 408–63

Smith P L, Van Zandt T 2000 Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology* **53**: 293–315

Townsend J T, Ashby F G 1983 *The Stochastic Modeling of Elementary Psychological Processes*. Cambridge University Press, Cambridge, UK

Wenger M J, Townsend J T 2000 Basic response time tools for studying general processing capacity in attention, perception and cognition. *Journal of General Psychology* **127**: 67–99

P. L. Smith

Stochastic Models

The simplest stochastic experiment is coin-toss. Each toss of a fair coin has two possible results and each of these results has probability of one half. This experiment is mathematically modeled with a random variable. A random variable is characterized by a state space and a probability distribution; in coin-toss the state space is {head, tail} and the distribution is 'probability of head = 1/2,' 'probability of tail = 1/2'.

This paper deals essentially with three issues. The first one is a description of families of independent random variables and their properties. The law of large numbers describes the behavior of averages of a large number of independent and identically distributed variables; for instance, it is 'almost sure' that in a million coin-tosses the number of heads will fall between 490 thousands and 510 thousands. The central limit theorem describes the fluctuations of the averages around their expected value. The Poisson approximation describes the behavior of a large number of random variables assuming values in {0, 1}, each one with small probability of being 1.

The second issue is phase transition and is discussed in the setup of percolation and random graph theory. For example, consider that each i belonging to a bidimensional integer lattice (that is, i has two coordinates, $i = (i_1, i_2)$, each coordinate belonging to the set of integer numbers $\{\ldots, -1, 0, 1, \ldots\}$) is painted black with probability p and white otherwise, independently of everything else. Is there an infinite black path starting at the origin of the lattice? When the answer is yes it is said that there is percolation. A small variation on the probability p may produce the transition from absence to presence of percolation.

The third question is related to dynamics. Markov chains describe the behavior of a family of random variables indexed by time in {0, 1, 2,...} for which the probabilistic law of the next variable depends only on the result of the current one. Problems for these chains include the long time behavior, law of large numbers and central limit theorems. Two examples of Markov chains are discussed. In the birth-and-death process at each unit of time a new individual is born or a present individual dies, the probabilities of such events being dependent on the current number of individuals in the population. In a branching process each new individual generates a family that grows and dies independently of the other families.

Special attention is given to interacting particle systems; the name refers to the time evolution of families of processes for which the updating of each member of the family depends on the values of the other members. Two examples are given. The simple exclusion process describes the evolution of particles jumping in a lattice subjected to an exclusion rule (at most one particle is allowed in each site of the lattice). The study of these systems at long times in large regions is called the hydrodynamic limit. In particular, when the jumps are symmetric the hydrodynamic limit relates the exclusion process with a partial differential equation called the heat equation, while when the jumps are not symmetric, the limiting equation is a one-conservation law called the Burgers equation. The voter model describes the evolution of a population of voters in which each voter updates his opinion according to the opinion of the neighboring voters. The updating rules and the dimension of the space give rise to either unanimity or coexistence of different opinions.

The stochastic models discussed in this article focus on two commonly observed phenomena: (a) a large number of random variables may have deterministic behavior and (b) local interactions may have global effects.

1. *Independent Random Variables*

The possible outcomes of the toss of a coin and its respective probabilities can be represented by a random variable. A random variable X is characterized by a set S of possible outcomes and the probability distribution or law of these outcomes: for each x in S, $\mathbb{P}(X = x)$ denotes the probability that the outcome results x. In coin-toss the set S is either {head, tail} or {0, 1} and the distribution is given by $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = 1/2$. The mean or expected value $\mathbb{E}(X)$ of a random variable X is the sum of the possible outcomes multiplied by the probability of the outcome: $\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x)$. The variance $\mathbb{V}(X)$ describes the square dispersion of the variable around the mean:

$$\mathbb{V}(X) = \sum_x (x - \mathbb{E}X)^2 \mathbb{P}(X = x). \tag{1}$$

A process is a family of random variables labeled by some set. Independent random variables are characterized by the fact that the outcome of one variable does not change the law of the others. For example, the random variables describing the outcomes of different coins are independent. The random variables describing the parity and the value of the face of a dice are not independent: if one knows that the parity is even, then the face can only be 2, 4, or 6.

1.1 Law of Large Numbers and Central Limit Theorem

A sequence of independent random variables is the most unpredictable probabilistic object. However, questions involving a large number of variables get almost deterministic answers. Consider a sequence of independent and identically distributed random variables X_1, X_2,... with common mean μ and finite variance σ^2. The partial sums of the first n variables is defined by $S_n = X_1 + \cdots + X_n$ and the empiric mean of the first n outcomes by S_n/n. The expectation and the variance of S_n are given by $\mathbb{E}(S_n) = n\mu$ and $\mathbb{V}(S_n) = n\sigma^2$. The law of large numbers says that as n increases, the random variable S_n/n is very close to μ. More rigorously, for any value $\varepsilon > 0$, the probability that S_n/n differs from its expectation by more than ε can be made arbitrarily close to zero by making n sufficiently large.

The Central Limit Theorem says that the distance between S_n and $n\mu$ is of the order $\sqrt{n}$. More precisely it gives the asymptotic law of the difference when n goes to ∞:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leqslant a \right) = \Phi(a) \tag{2}$$

where Φ is the normal cumulative distribution function defined by

$$\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} \mathrm{d}x \tag{3}$$

1.2 Law of Small Numbers and Poisson Processes

Consider a positive real number λ and a double labeled sequence $X(i, n)$ of variables with law $\mathbb{P}[X(i, n) = 1] = \lambda/n$, $\mathbb{P}[X(i, n) = 0] = 1 - \lambda/n$; n is a natural number, that is, in the set $\{0, 1,\ldots\}$ and i is in the 'shrinked' lattice $\{1/n, 2/n,\ldots\}$ (the nomenclature comes from the fact that as n grows the distance between points diminishes). In this way, for fixed n there are around n random variables sitting in each interval of length one but each variable $X(i, n)$ has a small probability of being 1 and a large probability of being 0 in such a way that the mean number of ones in the interval is around λ for all n. When n goes to infinity the above process converges to a Poisson process. This process is characterized by the fact that the number $N(\Lambda)$ of ones in any interval Λ is a Poisson random variable with mean $\lambda m(\Lambda)$, where $m(\Lambda)$ = length of Λ: $\mathbb{P}[N(\Lambda) = k] = e^{-\lambda m(\Lambda)}[\lambda m(\Lambda)]^k/k!$; furthermore, the numbers of ones in disjoint regions are independent. A bi-dimensional lattice consists of points with two coordinates $i = (i_1, i_2)$ with both i_1 and i_2 natural numbers and analogously the points in a d-dimensional lattice have d coordinates. When i runs in a multidimensional shrinked lattice, the limit is a multidimensional Poisson process. This is called the law of small numbers.

The one-dimensional Poisson process has the property that the distances between successive ones are independent random variables with exponential distribution (it is said that T has exponential distribution with parameter λ if $\mathbb{P}(T > t) = e^{-\lambda t}$). In this paper, one-dimensional Poisson process will be called a Poisson clock and the times of occurrence of ones will be called rings.

The classic book for coin-tossing and introductory probability models is Feller (1971). Modern approaches include Fristedt and Gray (1997), Brémaud (1999) and Durrett (1999a). Aldous (1989) describes how the Poisson process appears as the limit of many distinct phenomena.

2. Static Models

2.1 Percolation

Suppose that each site of the two-dimensional lattice is painted black with probability p and white with probability $1-p$ independent of the other sites. There is a black path between two sites if it is possible to go from one site to the other jumping one unity at each step and always jumping on black sites. The cluster of a given site x is the set of (black) sites that can be attained following black paths started at x. Of course, when p is zero the cluster of the origin is empty and when p is one, it is the whole lattice. The basic question is about the size of the cluster containing the origin. When this size is infinity, it is said that there is percolation. It has been proven that there exists a critical value p_c such that when p is smaller than or equal to p_c there is no percolation, and when p is bigger than p_c there is percolation with positive probability. This is the typical situation of phase transition: a family of models depending on a continuous parameter which show a qualitative change of behavior as a consequence of a small quantitative change of the parameter. Oriented percolation occurs when the paths can go from one point to the next following only two directions; for example South–North and West–East. The standard reference for percolation is Grimmett (1999).

2.2 Random Graphs

Consider a set of n points and a positive number c. Then, with probability c/n, connect independently each pair of points in the set. The result is a random graph with some number of connected components. Let $L_1(n)$ be the number of points in the largest connected component. It has been proven that as the number of points n increases to infinity the proportion $L_1(n)/n$ of points in the largest component converges to some value γ and that if $c > 1$ then $\gamma > 0$ and that if $c \leqslant 1$, $\gamma = 0$. This means that if the probability of connection is sufficiently large, a fraction of points is in the greatest component. In this case this is called the giant component. The number of points in any other component, when divided by n, always converges to zero; hence there is at most one giant component. The fluctuations of the giant component around γn, the behavior of smaller components, and many other problems of random graph theory can be found in Bollobás (1998).

3. Markov Chains

A Markov chain is a process defined iteratively. The ingredients are a finite or countable state space S, a sequence of independent random variables U_n uniformly distributed in the interval [0, 1], and a transition function F which to each state x and number u in the unit interval assigns a state $y = F(x, u)$. The initial state X_0 is either fixed or random and the successive states of the chain are defined iteratively by $X_n = F(X_{n-1}, U_n)$. That is, the value of the process X_n at each step n is a function of the value X_{n-1} at the preceding step and of the value of a uniform random variable U_n independent of everything. In this sense a Markov chain has no memory. The future behavior of the chain given the present position does not depend on how this position was attained. Let $Q(x, y) = \mathbb{P}(F(x, U_n) = y)$, the probability that the chain at step n be at state y given that it is at state x at time $n-1$. Since this probability does not depend on n the chain is said to be homogeneous in time. The transition matrix Q is the matrix with entries $Q(x, y)$.

When it is possible to go from any state to any other the chain is called irreducible, otherwise reducible. The state space of reducible chains can be partitioned in subspaces in such a way that the chain is irreducible in each part. Irreducible chains in finite state spaces stabilize in the sense that the probability of finding the chain at each state converges to a probability distribution π that satisfies the balance equations: $\pi(y) = \sum_x \pi(x)Q(x, y)$. The measure π is called invariant for Q because if the initial state X_0 has law π then the law of X_n will be π for all future time n. When there is a unique invariant measure π the proportion of time the chain spends in any given state x converges to $\pi(x)$ when time goes to infinity no matter how the initial state is chosen. This is a law of large numbers for Markov chains. The expected time the chain needs to come back to x when starting at x at time zero is $1/\pi(x)$. When the state space has a countable (infinite) number of points it is possible that there is no invariant measure; in this case the probability of finding the chain at any given state goes to zero; that is, the chain 'drives away' from any fixed finite region. It is possible also to define Markov chains in noncountable state spaces, but this will not be discussed here.

3.1 Coupling

Two or more Markov chains can be constructed with the same uniform random variables U_n. This is called a coupling and it is one of the major tools in stochastic processes. Applications of coupling include proofs of convergence of a Markov chain to its invariant measure, comparisons between chains to obtain properties of one of them in function of the other, and simulation of measures that are invariant for Markov chains.

Recent books on Markov chains include Chen (1992), Bianc and Durrett (1995), Fristedt and Gray (1997), Brémaud (1999), Durrett (1999a), Schinazi (1999), Thorisson (2000), Häggström (2000), and Ferrari and Galves (2000).

3.2 Examples: Birth–Death and Branching Processes

The state space of the birth–death process is the set of natural numbers $\{0, 1, \ldots\}$, the probability to jump from x to $x+1$ is p_x, a number in [0, 1] and the probability to jump from x to $x-1$ is $1-p_x$. It is assumed that $p_0 = 1$, that is, the process is reflected at the origin. The transition function is given by $F(x, u) = x + H(x, u)$, with $H(x, u) = 1$ if $u < p_x$ and $H(x, u) = -1$ if $u \geqslant p_x$. It is possible to find explicit conditions on p_x that guarantee the existence of an invariant measure. For instance, if $p_x < a$ for all x for some $a < 1/2$, then there exists an invariant measure π. The reason is that the chain has a drift towards the origin where it is reflected. On the other hand, if the drift towards the origin is not sufficiently strong, the chain may have no invariant measure.

The $(n+1)$th generation of a branching process consists of the daughters of the individuals alive in the nth generation:

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_{n,i} \tag{4}$$

where Z_n is the total number of individuals in the nth generation and $X_{n,i}$ is the number of daughters of ith

individual of the nth generation. $Z_0 = 1$ and $X_{n,i}$ are assumed independent and identically distributed random variables with values in $\{0, 1, 2, \ldots\}$. When $X_{n,i} = 0$ the corresponding individual died without a progeny. The process so defined is a Markov chain in $\{0, 1, 2, \ldots\}$; the corresponding transition function F can be easily computed from the distribution of $X_{n,i}$. The mean number of daughters of the typical individual $\mu = \mathbb{E}X_{n,i}$ (a value that does not depend on n, i) is the crucial parameter to establish the behavior of the process. When $\mu > 1$ the probability that at all times there are alive individuals is positive; in this case the process is called supercritical and in case of survival the number of individuals increases geometrically with n. When $\mu \leqslant 1$ the process dies out with probability one (subcritical case). Guttorp (1991) discusses branching processes and their statistics.

4. *Interacting Particle Systems*

Processes with many interacting components are called interacting particle systems. Each component has state space S and $X_t(i)$ denotes the state of component i at time t; t may be discrete or continuous, generally i belongs to a d-dimensional lattice. The update law of each component typically depends on the state of the other components; this dependence is in general 'local,' that is, determined only by the configuration of a few other neighboring components. The references for interacting particle systems are Liggett (1985) and (1999), Durrett (1988) and (1999b), Chen (1992), Schinazi (1999), Fristedt and Gray (1997).

4.1 *The Exclusion Process*

In this process i belongs to the d-dimensional lattice and $X_t(i)$ assumes only the value 0 and 1. If $X_t(i) = 0$, site i is empty at time t and if $X_t(i) = 1$, site i is occupied by a particle at time t. The process evolves as follows. Independent Poisson clocks are attached to the sites; when the clock rings at site i (at most one clock rings at any given time), if there is a particle at i, then a site j is chosen at random and the particle tries to jump to j. If j is empty then the particle jumps; if j is already occupied the jump is suppressed. The interaction between particles is embedded in the law of the jump. The simplest case is when the distribution of the jump does not depend on the configuration. In this case the interaction is just reduced to the suppression of jumps attempted to occupied sites and the motion is called the simple exclusion process.

The measure obtained when a particle is put at each site with probability ρ independently of the other sites is called a product measure. A surprising feature of the simple exclusion process with spatially homogeneous jump laws is that if the initial configuration is chosen according to a product measure, then the configuration at any future time will also be distributed with this measure. In other words, product measures are time invariant for the process. Due to the conservative character of the dynamics, i.e., the initial density is conserved, the process accepts a family of invariant measures; there is one invariant product measure for each density ρ in $[0, 1]$.

The main questions are related to the asymptotic behavior of the process when time runs. In many cases it has been proven that the process starting with an initial configuration having an asymptotic density converges as time goes to infinity to the product measure with that density. More detailed questions are related to hydrodynamic limits obtained when large space regions are observed at long times. A particularly interesting case is the symmetric simple exclusion process, for which jumps of sizes y and $-y$ have the same probability. The large space, long time behavior of the symmetric simple exclusion process is governed by the heat equation. For example, in the one-dimensional case it has been proven that the number of particles in the region $[aN, bN]$ at time tN^2 divided by N converges to $\int_a^b u(r, t)\,\mathrm{d}r$, where $u(r, t)$ is the solution of the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial r^2} \tag{5}$$

with initial condition $u(r, 0)$ (the initial particle distribution must be related with this condition in an appropriate sense). This is called a law of large numbers or hydrodynamic limit. The limiting partial differential equation represents the macroscopic behavior while the particle system catches the microscopic behavior of the system. The law of the configuration of the process in sites around rN at time tN^2 converges as N goes to infinity to the product measure with density given by the macroscopic density $u(r, t)$. This is the most disordered state (or measure). This is called propagation of chaos.

When the motion is not symmetric the behavior changes dramatically. The simplest case is the totally asymmetric one-dimensional nearest neighbor simple exclusion process. The particles jump in the one-dimensional integer lattice $\{\ldots, -1, 0, 1, \ldots\}$. When the Poisson clock rings at site i, if there is a particle at that site and no particle at $i+1$, then the particle jumps to $i+1$; otherwise nothing changes. To perform the hydrodynamic limit time and space are rescaled in the same way: the number of particles in the region $[aN, bN]$ divided by N at time tN converges to $\int_a^b u(r, t)\,\mathrm{d}t$, the solution of the Burgers equation

$$\frac{\partial u}{\partial t} = -\frac{\partial u(1-u)}{\partial r} \tag{6}$$

with initial condition $u(r, 0)$. Similar equations can be obtained in the d-dimensional case. Propagation of chaos is also present in this case. Contrary to what happens in the symmetric case, the Burgers equation develops noncontinuous solutions. In particular, consider $\rho < \lambda$ and the initial condition $u(r, 0)$ equal to ρ if $r < 0$ and to λ otherwise. Then the solution $u(r, t)$ is ρ if $r < (1-\rho-\lambda)t$ and λ otherwise. The well-known phenomenon of highways occur here: the space is divided in two regions sharply separated; one region of low density and high velocity and the other region of high density and low velocity. This phenomenon that was first observed for the Burgers equation (Evans (1998)) was then proven to hold for the microscopic system for the nearest neighbor case. The location of the shock is identified by a perturbation of the system: start with a configuration η coming from a product shock measure, that is, the probability of having a particle at site i is ρ if i is negative and λ if i is positive, and put a particle at the origin. Consider another configuration η' that coincides with η at all sites but at the origin. η' is a perturbation of η. Realizing two processes starting with η and η' respectively, using the same Poisson clocks, there will be only one difference between the two configurations at any given future time. The configuration as seen from the position of this perturbation is distributed approximately as η, that is, with a shock measure. The perturbation behaves as a second class particle; the sites that are occupied by particles of both η and η' are said to be 'occupied by a first class particle,' while the site where η differs from η' is said to be 'occupied by a second class particle.' Following the Poisson clocks, the second class particle jumps backwards when a first class particle jumps over its site and jumps forwards if the destination site is empty. See Liggett (1999).

Special interest has been devoted to the process when the initial condition is 'there are particles in all negative sites and no particles in the other sites.' The solution $u(r, t)$ of the Burgers equation with this initial condition is a rarefaction front: $u(r, t)$ is one up to $-t$, zero from t on and interpolates linearly between 0 and 1 in the interval $[-t, t]$. This particular case has been recently related to models coming from other areas of applied probability. In particular with increasing subsequences in random permutations, random matrices, last passage percolation, random domino tilings and systems of queues.

Hydrodynamics of particle systems is discussed by Spohn (1991), De Masi and Presutti (1991), and Kipnis and Landim (1999).

4.2 Voter Model

In the voter model each site of a d-dimensional integer lattice (this is the set of d-uples of values in the integer lattice $\{\ldots, -1, 0, 1, \ldots\}$) is occupied by a voter having two possible opinions: 0 and 1. At the rings of a Poisson clock each voter chooses a neighbor at random and assumes its opinion. The voter model has two trivial invariant measures: 'all zeros' and 'all ones.' Of course, if at time zero all voters have opinion zero, then at all future times they will be zero; the same occurs if all voters have opinion one at time zero. Are there other invariant measures? The answer is no for dimensions $d = 1$ and $d = 2$ and yes for dimension $d \geqslant 3$. The proof of this result is based on a dual process. This process arises when looking at the opinion of a voter i at some time t. It suffices to go back in time to find the voter at time zero that originates the opinion of i at time t. The space–time backwards paths for two voters may coalesce. In this case the two voters will have the same opinion at that time. The backwards paths are simple random walks. In one or two dimensions, nearest-neighbors random walks always coalesce, while in dimensions greater than or equal to three there is a positive probability that the walks do not coalesce. Hence there are invariant states in three or more dimensions for which different opinions may coexist.

A generalization of the voter model is the random average process. Each voter may have opinions corresponding to real numbers and, instead of looking at only one neighbor, at Poisson rings voters update their opinion with a randomly weighted average of the opinions of the neighbors and adopt it. As in the two-opinions voter model, flat configurations are invariant, but if the system starts with a tilted configuration (for example, voter i has opinion λi for some parameter $\lambda \neq 0$), then the opinions evolve randomly with time. It is possible to relate the variance of the opinion of a given voter to the total amount of time spent at the origin by a symmetric random walk. This shows that the variance of the opinion of a given voter at time t is of the order of t in dimension one, $\log t$ in dimension 2, and constant in dimensions greater than or equal to 3. The random average process is a particular case of the linear processes studied in Liggett (1985).

5. Future Directions

At the start of the twenty-first century, additional developments for the exclusion process include the behavior of a tagged particle, the convergence to the invariant measure for the process in finite regions, and more refined relations between the microscopic and macroscopic counterpart. The voter model described here is 'linear' in the sense that the probability of changing opinion is proportional to the number of neighbors with the opposite opinion. Other rules lead to 'nonlinear' voter models that show more complicated behavior. See Liggett (1999). The rescaling of the random average process may give rise to nontrivial limits; this is a possible line of future development.

The contact process, a continuous-time version of oriented percolation, has not been discussed here. The contact process in a graph instead of a lattice may

involve three or more phases (in the examples discussed here there were only two phases). The study of particle systems in graphs is a promising area of research.

An important issue not discussed here is related to Gibbs measures; indeed, phase transition was mathematically discussed first in the context of Gibbs measures. One can think of a Gibbs measure as a perturbation of a product measure: configurations have a weight according to a function called energy. Gibbs measures may describe the state of individuals that have a tendency to agree (or disagree) with their neighbors. Stochastic Ising models are interacting particle systems that have as invariant measures the Gibbs measure related to a ferromagnet. Mixtures of Ising models and exclusion processes give rise to reaction diffusion processes, related via hydrodynamic limit with reaction diffusion equations.

An area with recent important developments is Random tilings, which describes the behavior of the tiles used to cover an area in a random fashion. The main issue is how the shape of the boundary of the region determines the tiling. Dynamics on tilings would show how the boundary information transmits. The rescaling of some tilings results in a continuous process called the mass-less Gaussian field; loosely speaking, a multidimensional generalization of the central limit theorem.

An important area is generally called spatial processes. These processes describe the random location of particles in the d-dimensional real space instead of a lattice. The Poisson process is an example of a spatial process. Analogously to Gibbs measures, other examples can be constructed by giving different weight to configurations according to an energy function. The study of interacting birth and death processes having as invariant measure spatial process is recent and worth to develop. This is related also with the so-called perfect simulation; roughly speaking this means to give a sample of a configuration of a process as a function of the uniform random numbers given by a computer. See, for instance, Häggström (2000).

See also: Diffusion and Random Walk Processes; Dynamic Decision Making; Markov Decision Processes; Markov Processes for Knowledge Spaces; Probability: Formal; Probability: Interpretations; Sequential Decision Making; Stochastic Dynamic Models (Choice, Response, and Time)

Bibliography

Aldous D 1989 *Probability Approximations via the Poisson Clumping Heuristic*. Applied Mathematical Sciences, 77. Springer-Verlag, New York

Bianc P, Durrett R 1995 Lectures on probability theory. In: Bernard P (ed.) *Lectures from the 23rd Saint-Flour Summer School held August 18–September 4, 1993*. Lecture Notes in Mathematics, 1608. Springer-Verlag, Berlin

Bollobás B 1998 *Modern Graph Theory*. Graduate Texts in Mathematics, 184. Springer-Verlag, New York

Brémaud P 1999 *Markov Chains. Gibbs Fields, Monte Carlo Simulation, and Queues*. Texts in Applied Mathematics, 31. Springer-Verlag, New York

Chen M F 1992 *From Markov Chains to Nonequilibrium Particle Systems*. World Scientific Publishing, River Edge, NJ

De Masi A, Presutti E 1991 *Mathematical Methods for Hydrodynamic Limits*. Lecture Notes in Mathematics, 1501. Springer-Verlag, Berlin

Durrett R 1988 *Lecture Notes on Particle Systems and Percolation*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA

Durrett R 1999a *Essentials of Stochastic Processes. Springer Texts in Statistics*. Springer-Verlag, New York

Durrett R 1999b *Stochastic Spatial Models*. Probability Theory and Applications (Princeton, NJ, 1996), 5–47, IAS/Park City Mathematics Series 6, *American Mathematical Society*, Providence, RI

Evans L C 1998 *Partial Differential Equations*. American Mathematical Society, Providence, RI

Feller W 1971 *An Introduction to Probability Theory and its Applications*, Vol. II, 2nd edn. Wiley, New York

Ferrari P A Galves A 2000 *Coupling and Regeneration for Stochastic Processes*. Associatión Venezolana de Matemáticas, Mérida

Fristedt B, Gray L 1997 *A Modern Approach to Probability Theory: Probability and its Applications*. Birkhauser Boston, Boston, MA

Grimmett G 1999 *Percolation*, 2nd edn. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 321. Springer-Verlag, Berlin

Guttorp P 1991 *Statistical Inference for Branching Processes*. Wiley, New York

Häggström 2000 *Finite Markov Chains and Algorithmic Applications*. To appear.

Kipnis C, Landim C 1999 *Scaling Limits of Interacting Particle Systems*. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 320. Springer-Verlag, Berlin

Liggett T M 1985 *Interacting Particle Systems*. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Science], 276. Springer-Verlag, New York

Liggett T M 1999 *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 324. Springer-Verlag, Berlin

Schinazi R B 1999 *Classical and Spatial Stochastic Processes*. Birkhäuser Boston, Boston, MA

Spohn H 1991 *Large Scale Dynamics of Interacting Particles*. Springer-Verlag, Berlin

Thorisson H 2000 *Coupling, Stationarity, and Regeneration: Probability and its Applications*. Springer-Verlag, New York

P. A. Ferrari

Stock Market Predictability

One of the most enduring cornerstones of modern financial theory has been that stock markets are reasonably efficient in reflecting new information and

that price movements closely fit a random walk model where future prices cannot be predicted from an examination of past prices. Toward the end of the twentieth century, however, financial economists and statisticians began to question received doctrine, and many came to believe that stock prices were at least partially predictable. Moreover, although far more controversial, many have argued that these predictable patterns enable investors to earn excess risk-adjusted rates of returns and provide evidence that the efficient market theory should be rejected. This survey reviews the findings of the empirical work concerning stock market predictability. It reviews the work done on stock prices that appears to reject the random walk hypothesis and looks at the body of studies suggesting that future returns can be predicted on the basis of such fundamental variables as dividend yields, price earnings, and price-to-book value ratios, market capitalization (size), etc. The survey also examines the relationship between predictability and market efficiency.

1. *The Efficient Market Theory and its Critics*

The basic idea behind the efficient market theory is that securities markets are extremely efficient in digesting information about individual stocks or about the stock market in general. When information arises about a stock (or the market as a whole), the news spreads very quickly and is immediately incorporated into the prices of securities. Thus, neither technical analysis (an analysis of past price patterns to determine the future) nor fundamental analysis (an analysis of a company's earnings, dividends, future prospects, etc. to determine a stock's proper value) will help investors to achieve returns greater than would be obtained by buying and holding one of the broad stock market indices. The efficient market theory is associated with the idea of a 'random walk,' which is a term loosely used in the finance literature to characterize a price series where all subsequent price changes represent random departures from previous prices. The logic of the random walk idea is that if the flow of information is unimpeded and information is immediately reflected in stock prices, then tomorrow's price change will reflect only tomorrow's news and will be independent of the price changes today. But news is by definition unpredictable and, thus, resulting price changes must be unpredictable and random. As a result, prices fully reflect all known information, and even uninformed investors buying a diversified portfolio at the tableau of prices given by the market will obtain a rate of return as good as that achieved by market professionals. Discussions of the random walk hypothesis can be found in Samuelson (1965) and Fama (1970).

Academic work published mainly during the 1980s and 1990s has tended to cast doubts on many of the tenets of the efficient market theory. This work has suggested that the random walk model is not supported and that future stock prices and returns can be predicted, at least in part, by an examination of past stock prices and certain 'fundamental' valuation methods such as the ratio of stock prices to earnings and book values. Moreover, some of this work has been interpreted as implying that markets are inefficient in the sense that arbitrage opportunities exist that enable investors to earn excess risk-adjusted returns.

2. *The Major Predictable Patterns that Have Been Discovered*

This article will now focus on the findings of empirical work about stock market predictability done during the 1980s and 1990s. The general thrust of much of the published academic work has been along the following two lines. First, it appears that there are many predictable patterns in the stock market. The stock market is *not* random. Indeed, a book published by Lo and MacKinlay (1999) has the arresting title, *A Nonrandom Walk Down Wall Street*. Second, although this is far more controversial, many have argued that these predictable patterns enable an investor to earn big risk-adjusted rates of return and are proof that the efficient market theory is wrong.

2.1 *There is Some Evidence of Short-run Momentum in the Market*

When stock returns for individual stocks and market averages are calculated over days, or weeks, or even months, there is positive serial correlation. In other words, a positive rate of return one week is more likely than not to be followed by a positive return the next week. The market is not a perfect random walk. Further evidence provided by Lo and MacKinlay uses a simple specification test based on variance estimators. Their test exploits the fact that the variance of the increments of a random walk is linear in the sampling interval. For example, if stock prices are generated by a random walk then the variance of monthly sampled log-price relatives must be four times as large as the variance of a weekly sample. Using data over a 23-year time frame from 1962 through 1985, Lo and MacKinlay reject the random walk hypothesis. The rejection cannot be explained completely by infrequent trading (where news affects smaller company stocks with a lag because they trade less frequently) or by time-varying volatilities.

This work suggests that quantitative managers who use momentum to help guide purchases have some empirical support on their side. Whether this is a true inefficiency or not, however, is open to more question. Certainly any investor who pays transactions costs is not able to outperform a buy-and-hold strategy because the kind of momentum that has been found is

small relative to the transactions costs involved, and it is far from clear how dependable any patterns that can discovered will be. The market is not statistically random, but it is nearly so.

2.2 But there is Mean Reversion

Returns may be positively correlated for short periods, such as days or weeks, but when returns are measured over longer periods, such as three or four years, there tends to be negative correlation. Just as in the Bible seven lean years followed seven fat years, so in the stock market extraordinary returns are somewhat more likely to be followed by lackluster returns. For example, Fama and French (1988) demonstrated that long holding period returns are negatively serially correlated. The serial correlation is significant and implies that 25–40 percent of the variation in long holding period returns can be predicted on the basis of past returns. Similar results were found by Poterba and Summers (1988), who conclude that there is substantial mean reversion in stock market returns at longer horizons. Some studies have attributed this forecastability to the tendency of stock market prices to 'overreact.' DeBondt and Thaler (1985), for example, argue that investors are subject to waves of optimism and pessimism that cause prices to deviate systematically from their fundamental values and later to exhibit patterns of reversal. These findings give some support to investment techniques that rest on a 'contrarian' strategy, that is buying the stocks, or groups of stocks, that have been out of favor for long periods of time, and avoiding those stocks that have had large run-ups over the last several years.

While there is considerable support for this view, it should be pointed out that such mean reversion is quite a bit weaker in some decades than it is in other periods. Indeed, the strongest empirical results come from periods including the Great Depression. Moreover, such return reversals for the market as a whole may be quite consistent with the efficient functioning of the market and could result, in part, from the volatility of interest rates. There is a tendency when interest rates go up for stocks to go down, and as interest rates go down for stocks to go up. If, in fact, interest rates fluctuate over time, one will tend to get return reversals, or mean reversion, and this is quite consistent with the efficient functioning of markets where stock returns will need to go up or down to be competitive with bonds.

Moreover, it may not be possible to profit from the tendency for individual stocks to exhibit patterns of return reversals. For example, Fluck et al. (1997) simulated over the period a strategy of buying stocks that had particularly poor returns over the past three to five years. They found very strong statistical evidence of return reversals, but it was truly reversion to the mean, *not* an opportunity to make extraordinary returns. They found that stocks with very low returns over the past three to five years had higher returns in the next period. Stocks with very high returns over the past three to five years had lower returns in the next period, but the returns in the next period were similar for both groups. While they found strong evidence of mean reversion, they could not confirm that a contrarian approach would yield higher than average returns. There was a statistically strong pattern of return reversal, but not one that implied an inefficiency in the market that would enable investors to make excess returns.

2.3 There do Seem to be Some Seasonal Patterns in the Stock Market

A number of researchers have found that January has been a very unusual month for stock market returns. Returns from an equally weighted stock index have tended to be unusually high during the first two weeks of the year. The return premium has been particularly evident for small stocks i.e., those with relatively small total capitalizations, as has been demonstrated by Keim (1983). Haugen and Lakonishok (1988) detail the high January returns in a book entitled *The Incredible January Effect*.

There also appears to be a number of seasonal effects. For example, French (1980) documents significant Monday returns. There appear to be significant differences in average daily returns in countries other than the USA, as summarized by Hawawini and Keim (1995). There also appear to be some patterns in returns around the turn of the month, as shown by Lakonishok and Smidt (1988), as well as around holidays as shown by Ariel (1990).

The general problem with these anomalies, however, is that they are *not* dependable from period to period. The January effect has not been useful to investors through much of the 1990s. Moreover, these nonrandom effects (even if they were dependable) are small relative to the transactions costs involved in trying to exploit them. They do not appear to offer investors any arbitrage opportunities that would enable them to make excess returns.

2.4 The Size Effect

Probably one of the strongest effects that investigators have found is the so-called size effect: the tendency over long periods of time for small stocks to generate larger returns than those of large stocks. Since 1926, small-company stocks in the USA have produced a rate of return about 1.5 percentage points larger than the returns from large stocks.

Fama and French (1992) looked at data from 1963 to 1990 and divided all stocks into deciles according to their size as measured by total capitalization. Decile 1

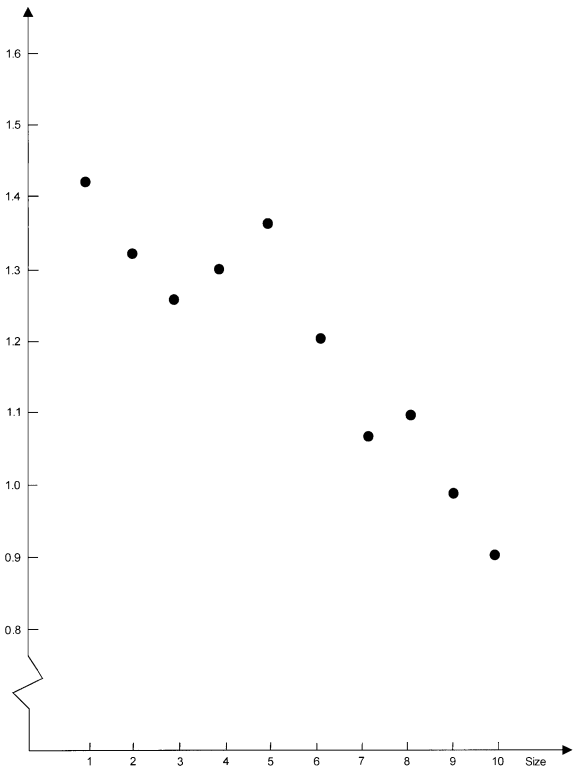

Figure 1
Average monthly returns for portfolios, formed on the basis of size, 1963–90 (source: Fama and French 1992)

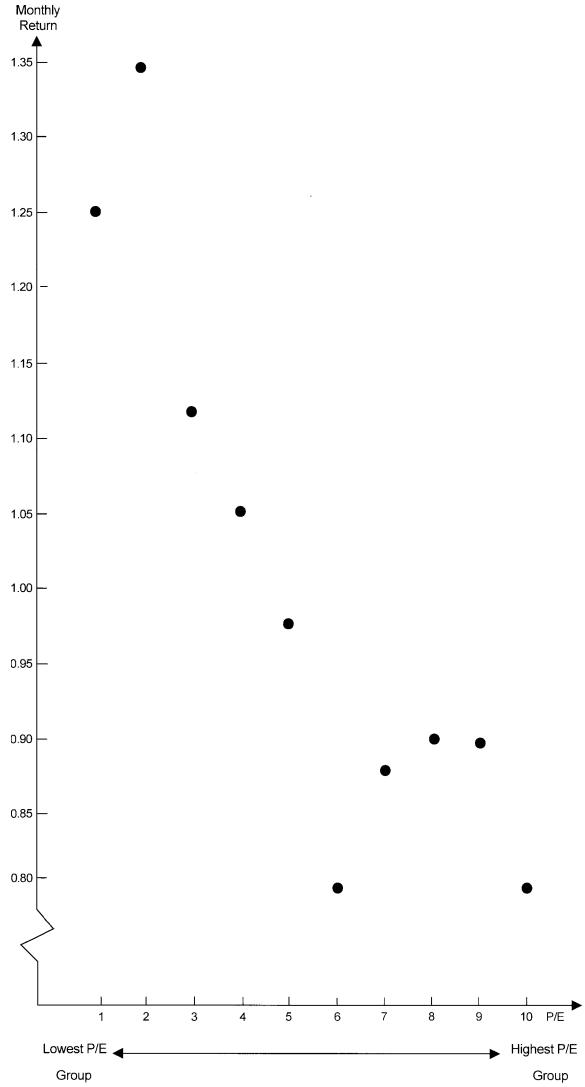

Figure 2
Average monthly returns for portfolios, formed on the basis of price-to-earnings (P/E) ratios, 1962–90 (source: Hawawini and Keim 1995)

contained the smallest 10 percent of all stocks while decile 10 contained the largest stocks. The results plotted in Fig. 1 show a clear tendency for the deciles made up of portfolios of smaller stocks to generate higher average monthly returns than deciles made up of larger stocks.

If the 'Beta' measure of systematic risk from the capital asset pricing model is accepted as the correct risk measurement statistic, the size effect can be interpreted as indicating a market inefficiency. This is so because portfolios consisting of smaller stocks have excess risk-adjusted returns. But any conclusion that an inefficiency exists involves accepting the joint hypothesis that a size effect is present *and* that the capital asset pricing model (CAPM) is correct. Fama and French point out, however, that the average relationship between 'Beta' and return is flat—not upward sloping as the CAPM predicts. Moreover, within Beta deciles, ten portfolios constructed by size display the same kind of positive relationship shown in Fig. 1. On the other hand, within size deciles, the relationship between Beta and return continues to be flat. Fama and French conclude that size may well be a far better proxy for risk than Beta, and therefore that their findings should not be interpreted as indicating that markets are inefficient.

The dependability of the size phenomenon is also open to question. Certainly during the decade of the 1990s there has been little to gain from holding smaller stocks. Indeed, in most world markets it has been the large stocks that have done especially well. Finally, it is also possible that the small-firm effect is simply a result of survivorship bias in currently available computer tapes of past returns. Today's lists of companies include only small firms that have survived, not the ones that later went bankrupt. Thus, a

researcher who examined the 10-year performance of today's small companies would be measuring the performance of those companies that survived—not of the ones that failed.

2.5 *There is Evidence that Stocks with Low Price Earnings Multiples Generate Higher Returns than those with High Price Earnings Multiples*

As Fig. 2 shows, evidence first published by Nicholson (1960) and later confirmed by Ball (1978) and *Basu* (1983) shows that low price earnings (P/E) stocks provide higher rates of return than high P/E stocks. This finding is consistent with the views of behavioralists (see, for example, Kahneman and Riepe 1998) that investors tend to be overconfident of their ability to project high earnings growth and thus overpay for 'growth' stocks. The finding is also consistent with the views of Benjamin Graham and David Dodd (1934), first expounded in their classic book on security analysis and later championed by the legendary US investor, Warren Buffett. Similar results have been shown for price/cash-flow multiples, where cash flow is defined as earnings plus depreciation and amortization. See Hawawini and Keim (1995).

2.6 *Stocks with Low Price-to-book Value (P/BV) Ratios Generate Larger Returns than Stocks with High P/BV Ratios*

The ratio of stock price to book value (the value of a firm's assets minus its liabilities divided by the number of shares outstanding) has also been found to be a useful predictor of future security returns. Low price-to-book (along with low P/E) is considered to be another hallmark of so-called 'value' in equity securities, and is also consistent with the view of behavioralists that investors tend to overpay for 'growth' stocks that subsequently fail to live up to expectations. Fama and French (1992) concluded that size and P/BV together provide considerable explanatory power for future returns and once they are accounted for, little additional influence can be attributed to P/E multiples. Fama and French (1997) also conclude that the P/BV effect is important in many world stock markets other than that of the USA.

Lakonishok et al. (1994) argue that such results raise questions about the efficiency of the market. But these findings do not necessarily imply inefficiency. P/BV may be a more accurate proxy for the systematic risk priced in the stock market than traditional estimates of the CAPM Beta because of measurement errors in the traditional estimates. Indeed, this is the explanation Fama and French give for this predictable pattern. Moreover, especially with so much merger and restructuring activity, book value is increasingly hard to interpret. Thus, it is far from clear that the pattern will continue. In fact, stocks with low P/BV ratios underperformed the general market during the 1990s. Indeed, if one looks at the behavior of growth stocks and value stocks (those with low P/BV and P/E) in the USA over a 75-year period to 2000, it appears that the Fama–French period from the early 1960s through 1990 may have been a unique period in which value stocks rather consistently produced higher rates of return.

2.7 *Predicting Market Returns from Initial Dividend Yields*

Another apparently predictable relationship concerns the returns realized from stocks and the initial dividend yields at which they were purchased. The data since 1926 for the USA are presented in Fig. 3. The figure was produced by measuring the dividend yield of the broad stock market (in this case, the Standard and Poor's 500 Stock Index) each quarter between 1926 and 2000, and then calculating the market's subsequent 10-year total return. The observations were then divided into deciles depending upon the level of the initial dividend yield. In general, the figure shows that investors have earned a higher rate of return from the stock market when they purchased stocks with an initial dividend yield that was relatively high, and a below average rate of return when they purchased them at initial dividend yields that were low.

Formal statistical tests of the ability of dividend yields (dividend-price ratios) to forecast future returns have been conducted by Fama and French (1988) and Campbell and Shiller (1988). Depending on the forecast horizon involved, as much as 40 percent of the variance of future returns can be predicted on the basis of initial dividend yields.

These findings are not necessarily inconsistent with efficiency. Dividend yields of stocks tend to be high when interest rates are high, and they tend to be low when interest rates are low. Consequently, the ability of initial yields to predict returns may simply reflect the adjustment of the stock market to general economic conditions. Moreover, the dividend behavior of US corporations may have changed over time (see Fama and French 1999). Companies in the twenty-first century may be more likely to institute a share repurchase program rather than increase their dividends. Thus, dividend yield may not be as meaningful as in the past. Moreover, it is worth pointing out that dividend yields were unusually low at the start of 1995. An investor who avoided US stocks then would have missed out on one of the steepest five-year increases in stock prices in history.

Finally, it is worth noting that this phenomenon does not work consistently with individual stocks. Investors who simply purchase a portfolio of individual stocks with the highest dividend yields in the market will not earn a particularly high rate of return.

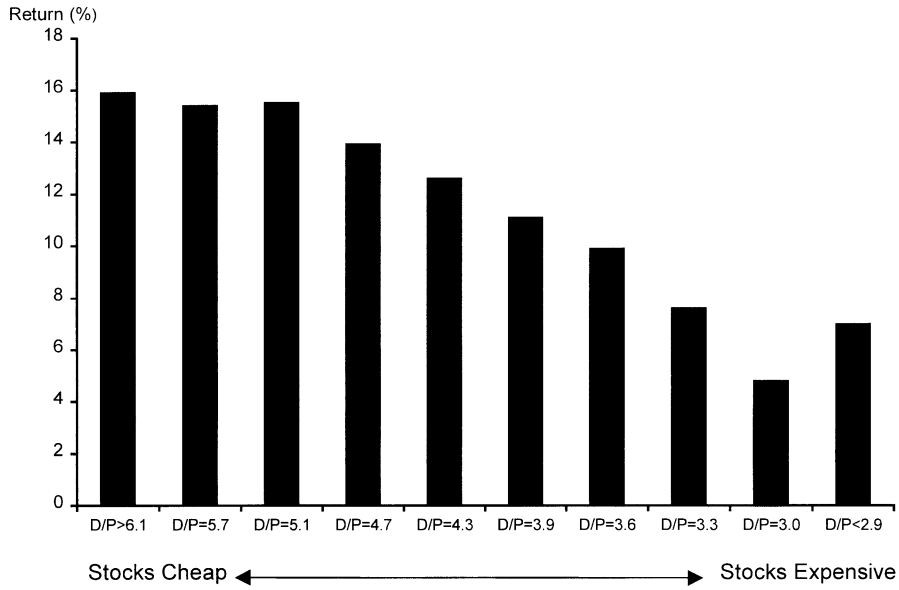

Figure 3
The future 10-year rates of return when stocks are purchased at alternative initial dividend yields (D/P) (source: The Leuthold Group)

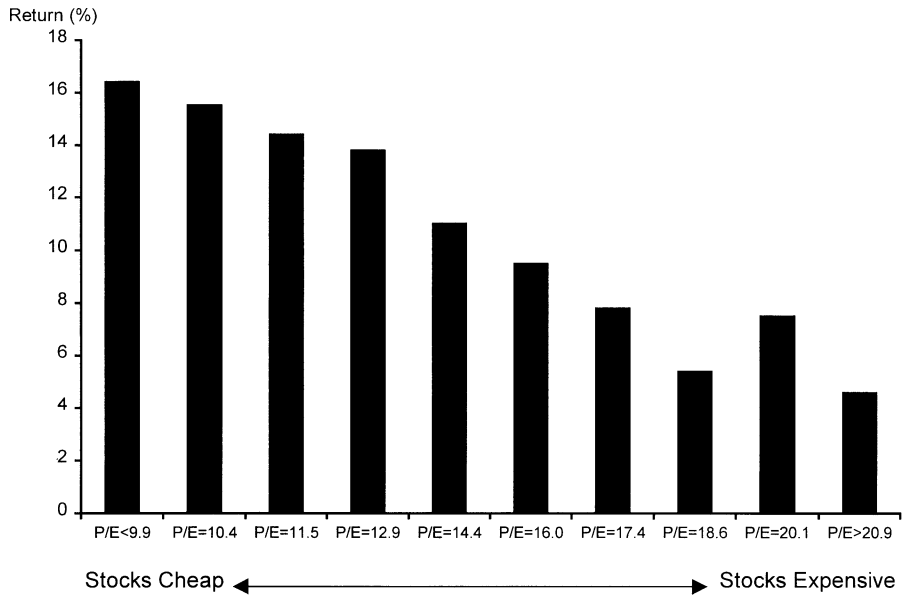

Figure 4
The future 10-year rates of return when stocks are purchased at alternative initial price-to-earnings (P/E) multiples (source: The Leuthold Group)

One popular implementation of such a 'high dividend' strategy is the 'Dogs of the Dow Strategy,' which involves buying the ten stocks in the Dow Jones Industrial Average with the highest dividend yields. For some periods this strategy handily outpaced the overall average, and so several 'Dogs of the Dow' funds were brought to market and aggressively sold to individual investors. Such funds have generally underperformed the market averages during the 1995–9 period.

2.8 Predicting Market Returns from Initial Price Earnings Multiples

The same kind of predictability for the market as a whole has been shown for price earnings ratios. The data are shown in Fig. 4. The figure presents a decile analysis similar to that described for dividend yields above. Investors have tended to earn larger long-horizon returns when purchasing stocks at relatively low price earnings multiples. Campbell and Shiller (1988) report R^2s (a measure of goodness of fit that takes on a maximum value of one) of over 40 percent and conclude that equity returns have been predictable in the past to a considerable extent. Whether such historical relations are necessarily applicable to the future is another matter. Such an analysis conducted by Shiller in January 1996 predicted a long-horizon (nominal) return for the US stock market close to zero.

3. Concluding Comments

In summary, there are many statistically significant predictable patterns in the stock market. It is important to emphasize, however, that these patterns are not dependable in each and every period, and that some of the patterns based on fundamental valuation measures of individual stocks may simply reflect a better proxy for measuring risk. Moreover, many of these patterns could self-destruct in the future, as many of them have already done. Indeed, this is the logical reason why one should be cautious not to overemphasize these anomalies and predictable patterns.

Suppose, for example, that one of the anomalies is really true. Suppose that there is a truly dependable January effect, and that the stock market—especially stocks of small companies—will generate extraordinary returns during the first five days of January. What will investors do? They will buy late in December, on the last day of December, and sell on January 5. But then investors find that the market rallied on the last day of December and so they will begin to buy on the next-to-last day; and because there is so much 'profit taking' on January 5, investors will have to sell on January 4 to take advantage of this effect. Thus, to beat the gun, investors will have to be buying earlier and earlier in December, and selling earlier and earlier in January so that eventually the pattern will self-destruct. Any truly repetitive pattern that can be discovered in the stock market and can be arbitraged away will self-destruct. Indeed, the January effect became undependable after it received considerable publicity.

Similarly, suppose there is a general tendency for stock prices to underreact to certain new events, leading to abnormal returns to investors who exploit the lack of full immediate adjustment (see DeBondt and Thaler 1995 and Campbell et al. 1997 for a discussion of event studies and possible under reaction of market participants). 'Quantitative' investment managers will then develop strategies in an attempt to exploit the pattern. Indeed, the more potentially profitable a discoverable pattern is, the less likely it is to survive.

Moreover, many of the predictable patterns that have been discovered may simply be the result of data mining. The ease of experimenting with financial data banks of almost every conceivable dimension makes it quite likely that investigators will find some seemingly significant but wholly spurious correlation between financial variables or among financial and non-financial data sets. One amusing one that has been uncannily accurate in the past for US markets is the Super Bowl Indicator, which predicts the sign of market returns for the year by the winner of January's football Super Bowl. Given enough time and mass-aging of data series it is possible to tease almost any pattern out of every data set. Moreover, the published literature is likely to be biased in favor of reporting such results. Significant effects are likely to be published in professional journals while negative results, or boring confirmations of previous findings, are relegated to the file drawer or discarded. Data-mining problems are unique to nonexperimental sciences, such as economics, which rely on statistical analysis for their insights and cannot test hypotheses by running repeated controlled experiments.

An exchange at an academic conference (1992) between Robert Shiller (1981), an economist who is sympathetic to the argument that stock prices are partially predictable and skeptical about market efficiency, and Richard Roll, an academic financial economist who also is a businessman managing billions of dollars of investment funds, is quite revealing. After Shiller stressed the importance of inefficiencies in the pricing of stocks, Roll responded as follows:

> I have personally tried to invest money, my client's money and my own, in every single anomaly and predictive device that academics have dreamed up … I have attempted to exploit the so-called year-end anomalies and a whole variety of strategies supposedly documented by academic research. And I have yet to make a nickel on any of these supposed market inefficiencies … a true market inefficiency ought to be an exploitable opportunity. If there's nothing investors can exploit in a systematic way, time in and time out, then it's very hard to say that information is not being properly incorporated into stock prices. (*Journal of Applied Corporate Finance*, Spring, 1992)

Finally, there is a remarkably large body of evidence suggesting that professional investment managers are not able to outperform index funds that simply buy and hold the broad stock market portfolio (see, for example, Malkiel 1995). Three-quarters of professionally managed funds are regularly outperformed by a broad index fund with equivalent risk, and those that do appear to produce excess returns in one period are not likely to do so in the next. The record of

professionals does not suggest that sufficient predictability exists in the stock market to produce exploitable arbitrage opportunities.

Pricing irregularities and predictable patterns in stock returns do appear over time and even persist for periods. Undoubtedly, with the passage of time and with the increasing sophistication of our databases and empirical techniques, we will document further apparent departures from efficiency, and further patterns in the development of stock returns. But the end result will not be an abandonment of the belief of many in the profession that the stock market is remarkably efficient in its utilization of information and that whatever patterns do exist are unlikely to provide investors with a method to obtain extraordinary returns.

See also: Decision Research: Behavioral; Markets and the Law; Risk: Theories of Decision and Choice

Bibliography

Ariel R A 1990 High stock returns before holidays: Existence and evidence on possible causes. *Journal of Finance* **45**: 1611–26

Ball R 1978 Anomalies in relationships between securities' yields and yield-surrogates. *Journal of Financial Economics* **6**: 103–26

Basu S 1983 The relationship between earning's yield, market value and the returns for NYSE common stocks: Further evidence. *Journal of Financial Economics* **12**: 129–56

Campbell J Y, Lo A W, MacKinlay A C 1997 *The Econometrics of Financial Markets*. Princeton University Press, Princeton, NJ

Campbell J Y, Shiller R J 1988 Stock prices, earnings, and expected dividends. *Journal of Finance* **43**: 661–76

DeBondt W F M, Thaler R 1985 Does the stock market overreact? *Journal of Finance* **40**: 793–805

DeBondt W F M, Thaler R 1995 Financial decision-making in markets and firms: A behavioral perspective. In: Jarrow R, Maksimovic V, Ziemba W (eds.) *Handbook in Operations Research & MS*. Elsevier Science, Vol. 9

Fama E F 1970 Efficient capital markets: A review of theory and empirical work. *Journal of Finance* **25**: 383–417

Fama E F, French K R 1988 Permanent and temporary components of stock prices. *Journal of Political Economy* **96**: 246–73

Fama E F, French K R 1997 Value vs. growth: The international evidence. *Journal of Finance* **53**: 1975–99

Fama E F, French K (forthcoming) Disappearing dividends: Changing firm characteristics or increased reluctance to pay? *Journal of Financial Economics*

Fluck Z, Malkiel B G, Quandt R E 1997 The predictability of stock returns: A cross-sectional simulation. *Review of Economics and Statistics* **79**: 176–83

French K R 1980 Stock returns and the weekend effect. *Journal of Financial Economics* **8**: 55–69

Graham B, Dodd D L 1934 *Security Analysis: Principles and Techniques*. McGraw-Hill, New York

Haugen R A, Lakonishok J 1988 *The Incredible January Effect*. Dow Jones-Irwin, Homewood, AL

Hawawini G, Keim D B 1995 On the predictability of common stock returns: Worldwide evidence. In: Jarrow R, Maksimovic V, Ziemba W (eds.) *Handbooks in Operations Research & MS*. Elsevier Science, Amsterdam New York, vol. 9, pp. 497–544

Kahneman D, Riepe M W 1998 Aspects of investor psychology. *Journal of Portfolio Management* **24**: 52–65

Keim D B 1983 Size-related anomalies and stock return seasonality: Further empirical evidence. *Journal of Financial Economics* **12**: 13–32

Lakonishok J, Schleifer A, Vishny R W 1994 Contrarian investment, extrapolation, and risk. *Journal of Finance* **49**: 1541–78

Lakonishok J, Smidt S 1988 Are seasonal anomalies real? A ninety-year perspective. *Review of Financial Studies* **1**: 403–25

Lo A W, MacKinlay A C 1999 *A Non-random Walk Down Wall Street*. Princeton University Press, Princeton, NJ

Malkiel B G 1995 Returns from investing in equity mutual funds 1971 to 1991. *Journal of Finance* **50**: 549–72

Nicholson S F 1960 Price-earnings ratios. *Financial Analysts Journal*. **July/August**: 43–50

Poterba J M, Summers L H 1988 Mean reversion in stock returns: Evidence and implications. *Journal of Financial Economics* **22**: 27–59

Roll R, Shiller R J 1992 Comments: Symposium on volatility in US and Japanese stock markets. *Journal of Applied Corporate Finance* **1**: 25–9

Samuelson P 1965 Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* **6**: 41–9

Shiller R J 1996 Price-earnings ratios as forecasters of returns: The stock market outlook in 1996. Unpublished manuscript, Yale University

Shiller R J 1981 Do stock prices move too much to be justified by subsequent changes in dividends?*American Economic Review* **71**: 421–36

B. G. Malkiel

Stockholders' Ownership and Control

Theory and research on the relations among top managers, company directors, stockholders, and external contenders for corporate control experienced a remarkable flowering during the 1990s. Early work addressed the central puzzle raised by the widespread separation of ownership and control among large American corporations, namely, why would any sensible person—much less thousands of them—invest their savings in businesses run by unaccountable professional managers? As Berle and Means (1932) framed the problem, those who ran such corporations would pursue 'prestige, power, or the gratification of professional zeal' in lieu of maximizing profits. Shareholders weakened by their fractionation could do little to stop them. Yet generations of individuals and financial institutions continued to invest in these firms. Why?

1. Resolving the Separation of Stock Ownership from Company Control

Answering this question led to the creation of a new theory of the firm that portrayed the public corporation as a 'nexus of contracts.' In this model, the mana-

gers of the corporation were disciplined to optimize shareholder value by an array of mechanisms, from incentive compensation and vigilant directors to an unforgiving 'market for corporate control.' Taken together, these devices worked to vouchsafe shareholder interests even when ownership was widely dispersed. Research in this tradition flourished in the 1980s, as takeovers of under-performing firms in the US became common and restive institutional investors made their influence known. Studies focused on assessing the effectiveness of such contrivances as governing boards dominated by outsiders, executive pay tied to stock performance, and susceptibility to hostile acquisition (Walsh and Seward 1990).

Three developments, however, expanded research on ownership and control beyond these mechanisms and beyond American business. The first was for investigators to view the governance structure of the firm—the set of devices that evolve within the firm to monitor managerial decision making—as an *ensemble*. Rather than regarding any particular aspect of the firm's structure as essential, researchers came to study them as complements or substitutes. Remuneration highly leveraged around share price, for instance, could act as a substitute for a vigilant board, and together they could really tighten the leash.

The second development was a growth of comparative and historical research that highlighted the idiosyncrasy of the American obsession with ownership and control. American-style corporate governance, aimed at 'solving' the problems created by the separation of ownership and control, came to be seen as only one of several feasible systems. Indeed, among some advanced economies with well-developed business enterprise, the problem needed no solving at all since the separation of ownership from control remained much the exception (Roe 1994).

The third development was the articulation of a reflexive stance on the theory of ownership and control. While agency theory can be viewed as an empirical theory of the corporation, it also came to be considered a prescriptive theory, an explanation of not only what *is* but also a vision of what *could* or *should* be. Its influence on public policy debates during the 1980s and 1990s is evident in the widespread rise of the rhetoric of shareholder value in the business press and executive suite (Useem 1996).

2. *Shareholder Influence on Top Management and the Corporation*

When ownership is dispersed widely, shareholders possess few direct devices to ensure their control—or at least their influence—on the company's direction and performance. This is the essence of the 'agency problem' identified by Berle and Means (1932), and it remains potentially potent among many US firms. Three-quarters of the 100 largest US corporations in 1999 lacked even a single ownership block of 10 percent or more and, of the top 25, the biggest single holder averaged a scant four percent (Brancato 1999).

Scattered shareholders in effect relinquish their control to directors whom they elect to serve as their agents in selecting and supervising top management. Dispersed holders can influence the firm only through trading shares and voting directors. They have no real voice in the selection of executives, or even who serves on the board. By one argument, shareholder inattention is in fact the only rational course for small shareholders since the cost of learning about the governance of the firm and mobilizing change is sure to be far overshadowed by the marginal and nonselective benefits of any resulting performance improvement. This problem of collective action could result in management unaccountability in some capital markets, but investors nonetheless rest easy in the US because of a well-developed set of mechanisms that have arisen to ensure share price maximization without overt investor control (Easterbrook and Fischel 1991).

The most forceful instrument for riveting attention on share price is the 'market for corporate control.' By hypothesis, a poorly-run corporation suffers a depressed market valuation, and this opens an opportunity for outsiders with better management skills to acquire the firm at an implicit discount, oust top management, and rehabilitate the firm, thus restoring value to its long-suffering stockholders. Unwanted takeovers and their threats provide a potent mechanism to shock top managers who under-serve shareholder interests; they know that executives generally find themselves on the street after an unfriendly takeover. The market for corporate control is thus the most visible hand of Darwinian discipline, weeding out poorly run firms and protecting shareholders from bad management (Manne 1965).

Hostile takeovers are a blunt instrument for corporate control, and as a result, some US institutional investors have moved toward more fine-grained methods of influence. Large pension funds, for example, often field shareholder proxy proposals for better governance. Though little evidence yet confirms that this mode of shareholder activism has immediate impact on the share price of target firms, it does push targeted firms to improve their governance and thereby further open them further to investor influence (Karpoff 1998).

If hostile takeovers are often too blunt an instrument and shareholder activism too diffuse, stock analysts provide an intermediate form of shareholder influence on company behavior. Analysts investigate companies in order to render judgments on their prospects for future earnings and thus to estimate their appropriate valuation. Analysts are often allowed privileged access to corporate executives and are in a position to give strategic and governance advice

directly to senior managers, something rarely afforded to the firm's small shareholders. In principle, the rewards go to the most accurate analysts, giving them incentives to act as corporate watchdogs. In practice, however, some analysts are not entirely dispassionate observers: those who work for firms doing business with a corporation give systematically more favorable evaluations than do others. While analysts may serve as a more refined channel for shareholder influence, this avenue of control remains problematic as well (Useem 1996).

3. *Between Ownership and Control: Top Management*

A handful of company executives make the defining decisions, whether to launch a new product, enter a promising region, or resist a tender offer. It is they who most directly control the fate of the firm. It is they who stand between corporate owners and company results.

Firms often define their 'top' as no more than the seven or eight most senior officers. To much of the world beyond the company walls, however, top management is personified almost solely by the chief executive. Academic researchers had long been drawn to the same pinnacle of the pyramid, partly on the conceptual premise that the chief executive is the manager who really matters, and partly on the pragmatic ground that little is publicly available on anybody except the CEO. We have, thus, benefited from a long accumulation of work on their family histories, educational pedigrees, and political identities, and we know as a result that their origins are diverse, credentials splendid, and instincts conservative. More contemporarily, we have learned much about their tangled relations with directors and investors as well. To know the CEO's personal rise and board ties is to anticipate much of the firm's strategic intent and performance promise. They are also a good predictor of the CEO's own fortunes. An elite MBA degree accelerates movement to the top, a wealthy background attracts outside directorships, and a hand-picked board enhances pay and perquisites (Useem and Karabel 1986).

The conceptual and pragmatic underpinnings for shining the light solely on the CEO as the fulcrum for ownership and control, however, have eroded in recent years, as companies redefined their operations and researchers reconceived their methods. A central thrust of company restructuring during the 1980s and 1990s has been to transform the chief executive's office into the office of the executive. At many companies, CEOs have fashioned teams among their top officers that meet frequently and resolve jointly, their vertical control giving way to lateral leadership. As a result, researchers expanded their field of view from the chief executive to the entire upper echelon. They applied fresh strategies, ranging from personal interviews to direct observation and internet surveys, to acquire data on top managers other than the chief executive who now matter far more to company performance and stock analysts. A host of studies subsequently confirmed what many executives already appreciated. A better predictor of a company's performance than the chief executive's capability is the quality of the team that now runs the show (Finkelstein and Hambrick 1996).

Yet even this widening of research attention beyond the chief executive has not broadened enough. Many companies have expanded the concept of top management to include one, two, or even three hundred senior managers whose decisions have decisive bearing on firm performance. The concept of top management, then, has expanded in the minds of company managers and, to a lesser extent, academic researchers, from just the boss to the boss's team and now to the boss's court.

If shareholders are increasingly insistent that company directors require top managers to deliver steadily rising returns on their investments, their action is premised on a critical assumption that these top management teams do make the difference. While the premise may seem intuitively obvious to those inhabiting this world, it is far from self-evident to those who study it. Observers diverge from the assumption of influential executives in two opposing directions, one viewing top managers as all powerful, others seeing them as all but powerless.

Vilfredo Pareto and kindred 'elite' theorists early articulated the all-powerful view, but C. Wright Mills captured it best for the US in his classic work, *The Power Elite* (1956). Company executives, in his acid account, had joined with government officials and military commanders in an unholy alliance to exercise a commanding control over companies and the country. To the extent that top management is indeed all powerful, shareholders possess a silver bullet for righting whatever has gone wrong in a company: install new management.

For the community of management scholars, Pfeffer (1981) well articulated the opposite, all-*powerless* view with the contention that market and organizational constraints so tied top management's hands that it was much the captive, not a maker, of its own history. Production technologies and competitive frays in this view are far more determining of company results than the faceless executives who sit in their suites. It counts little who serves in top management, and, by extension, on the board. If top management is as powerless as the imagery suggests, investors seeking new management in response to poor performance are surely wasting their time.

Company executives do complain of a seeming helplessness at times, but almost anybody in working, consulting, or research contact with top managers reports just the opposite. Direct witnesses typically describe instead a commanding presence of top management, and they characterize its capabilities as

making the difference between a company's success or failure. Investors themselves think otherwise as well. When companies announce executive succession, money managers and stock analysts are quick to place a price on the head of the newly arrived, and, depending on the personality, billions can be added to or subtracted from the company's capitalization.

The importance that investors and directors attribute to top management for growing their fortunes can also be seen in the ultimate punishment for failure to do so: dismissal. Whether the US, Japan, or Germany, the likelihood of a CEO's exit in the wake of a stock free fall is increased by as much as half. And investors are often the engine behind the turnover. Study of Japanese companies stunned by a sharp reversal of fortune, for example, reveals that those whose top ten shareholders control a major fraction of the firm's stock are more often than others to dismiss the president and replace directors (Kaplan 1997, Kang and Shivdasani 1997).

Similarly, company boards are observed to reward successful executives far more generously than in the past. For each additional $1,000 added to a company's value in 1980, directors on average provided their chief executive an extra $2.51. By 1990, the difference in their payments had risen to $3.64, and by 1994 to $5.29. Put differently, in 1980 the boards of companies ranked at the 90th percentile in performance gave their CEOs $1.4 million more than did boards whose firms ranked at the 10th percentile (in 1994 dollars). But by 1990, the boards had expanded that gap to $5.3 million and by 1994 to $9.2 million (Hall and Liebman 1998).

4. *Between Stockholders and Top Management: The Governing Board*

If companies make semiconductors and concrete much the same way the world around, they organize their directors in almost as many ways as there are national economies. American boards tend to include only two or three insiders, while Japanese boards rarely include even two outsiders. German and Dutch governance is built around a two-tier governance structure, employees holding half of the upper tier seats; British and Swiss governance is designed around a single-tier, management-dominated structure. Some systems give formal voice to labor, others none: German law requires that labor representatives serve on the board, while French law places labor observers in the boardroom. American law mandates nothing, but some boards have added a labor leader on their own (Charkham 1994).

Yet such varieties in national practices may well diminish for two reasons. First, as equity markets internationalize, with companies seeking capital from all corners of the globe, investors predictably prefer relatively consistent director models that they believe will optimize shareholder value and performance transparency regardless of country. Their mental models of what's right may not always be right, but that is beside the point. Consider the preference expressed by some institutional investors in the US for a board chair who is not a sitting executive, believing that separation improves monitoring. Whatever the separatist penchant and however well it may have served some companies, research reveals that such a division creates no clear advantage for company performance (Baliga et al. 1996).

The diversity in directorship practices is likely to diminish, however, for a second, more factual reason. Certain practices do engender better performance, regardless of the setting, and as companies increasingly compete worldwide for customers and investors, they are likely to adopt what indeed does prove best. A case in point here is the number of directors on the board. Research on teams' success suggests that bigger is not better when boards exceed 15 or 20 members. When the number of directors climbs far above middle range, the engagement of each diminishes, and so too does their capacity to work in unison. Studies of American and European firms reveals that smaller boards displayed stronger incentives for their chief executive, greater likelihood of dismissing an under-performing CEO, and larger market share and superior financial performance (Conyon and Peck 1998).

It is not surprising, then, to see companies migrating worldwide toward smaller boards in search of improved performance. Directors in principle protect owner interests, direct company strategy, and select top management. In practice, they had concentrated more on strategy and selection and less on owner interests. But the rising power of investors has made for greater director focus on creating value and less coziness with top management. American and British directors, following the 'Anglo-Saxon' model, are already more focused on value than most. But directors in other economies can be expected to gravitate slowly toward the mantra of shareholder supremacy as well.

5. *Shareholders, Directors, and Executives at the Table*

With the rise of professional investors and their subjugation of national boundaries, those who occupy the executive suite and those who put them there are drawing far more research and policy attention, and justifiably so. In a by-gone era when markets were more steady and predictable, when airline and telephone executives confidently knew what to expect next year, the identity of top management mattered little. When shareholders were far smaller and decidedly quieter, when airline and telephone directors comfortably enjoyed inconsequential board meetings, the composition of the governing body mattered little

either. During the 1990s, however, all of this has changed in the US and UK, with other economies close behind. As shareholders have sought to reclaim authority over what they owned, they have brought top management and company directors back in. Company strategy and financial results can no longer be understood without understanding the capabilities and organization of those most responsible for delivering them.

The market is no longer viewed as so impersonal, the company no longer so isolated. Ownership and financial decisions are embedded in a complex network of working relations among money managers, stock analysts, company directors, senior executives, and state regulators. Given the right combination of features, that lattice can yield high investment returns and robust national growth. Given the wrong amalgam, it can lead instead to self-dealing, frozen form, and economic stagnation. The road ahead thus depends on how well researchers understand what control mechanisms and leadership styles work well both within national settings and across cultural divides, and on the extent that top managers, company directors, and stockholders learn and apply what is best.

See also: Business Law; Capitalism; Capitalism: Global; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; Institutional Investors; International Law and Treaties; International Organization; International Trade: Commercial Policy and Trade Negotiations; Internet: Psychological Perspectives; Law and Economics; Law and Economics: Empirical Dimensions; Markets and the Law; Monetary Policy; Regulation, Economic Theory of; Regulation: Working Conditions; Stock Market Predictability; Technological Innovation; Venture Capital

Bibliography

Baliga B R, Moyer R C, Rao R S 1996 CEO duality and firm performance: What's the fuss? *Strategic Management Journal* **17**: 41–53

Berle Jr A A, Means G C 1932 *The Modern Corporation and Private Property*. Commerce Clearing House, New York

Brancato C 1999 *International Patterns of Institutional Investment*. Conference Board, New York

Charkham J 1994 *Keeping Good Company: A Study of Corporate Governance in Five Countries*. Oxford University Press, New York

Conyon M J, Peck S I 1998 Board size and corporate performance: Evidence from European companies. *European Journal of Finance* **4**: 291–304

Davis G F, Thompson T A 1994 A social movement perspective on corporate control. *Administrative Science Quarterly* **39**: 141–73

Easterbrook F H, Fischel D R 1991 *The Economic Structure of Corporate Law*. Harvard University Press, Cambridge, MA

Finkelstein S, Hambrick D C 1996 *Strategic Leadership: Top Executives and Their Effects on Organizations*. West Publishing, Minneapolis-St. Paul, MN

Hall B J, Liebman J B 1998 Are CEOs really paid like bureaucrats? *Quarterly Journal of Economics* **113**: 653–91

Kang J-K, Shivdasani A 1997 Corporate restructuring during performance declines in Japan. *Journal of Financial Economics* **46**: 29–65

Kaplan S N 1997 Corporate governance and corporate performance: A comparison of Germany, Japan, and the U.S. *Journal of Applied Corporate Finance* **9**: 86–93

Karpoff J M 1998 The impact of shareholder activism on target companies: A survey of empirical findings. Unpublished, University of Washington

Manne H G 1965 Mergers and the market for corporate control. *Journal of Political Economy* **73**: 110–20

Pfeffer J 1981 Management as symbolic action: The creation and maintenance of organizational paradigms. In: Cummings L L, Staw B M (eds.) *Research in Organizational Behavior*. JAI, Greenwich, CT, Vol. 3

Roe M J 1994 *Strong Managers, Weak Owners: The Political Roots of American Corporate Finance*. Princeton University Press, Princeton, NJ

Useem M 1996 *Investor Capitalism: How Money Managers Are Changing the Face of Corporate America*. Basic Books, New York

Useem M, Karabel J 1986 Pathways to top corporate management. *American Sociological Review* **51**: 184–200

Walsh J P, Seward J K 1990 On the efficiency of internal and external corporate control mechanisms. *Academy of Management Review* **15**: 421–58

M. Useem and G. Davis

Strategic Human Resources Management

Defining 'strategic human resource management' and placing it in context involves several related questions. In the first place, what do we mean by 'strategy'? Then, what is 'human resource management' and does it differ from '*strategic* human resource management'? What are the characteristics and issues concerning different approaches to strategic human resource management? What is the relationship between strategic human resource management and organizational performance?

1. What is Strategy?

First, a simple answer: Strategy may be defined as a set of fundamental choices about the ends and means of an organization. More recently it might be defined as policies and procedures that are aimed at securing competitive advantage. That said, there are several different approaches in exploring the question 'what is strategy?' Some focus on the *content* of strategy and its positioning in relation to factors internal or external to

the organization that might be seen as salient to its securing competitive advantage. An example of the internal, firm-focused approach would be the 'resource-based' view of strategy of the firm (e.g., Barney 1991). This perspective defines firms as unique bundles of resources and suggests that the route to developing sustained competitive advantage is to develop resources internally that meet the criteria of value, rarity, being costly to imitate, and nonsubstitutability. Examples of the latter, market-focused approach might be Porter's (1980) generic strategies (low cost, differentiation, segment, or niche) in relation to his five-forces model or Miles and Snow's (1978) categorization of firms as prospectors, analyzers, defenders, and reactors, both models emphasizing market-oriented choices. Theories that bring together internal and external perspectives are life-cycle theories of the firm or product (for example, the Boston Consultancy Group's portfolio planning growth-share matrix).

A second approach, which can subsume the content approaches, focuses on the *processes* of strategy making. A famous distinction is that between 'planned' and 'emergent' strategy (Mintzberg and Walters 1985). Building on this is a useful typology of Whittington (1993), based on two axes, relating to continua of outcomes (profit maximizing/pluralistic) and of processes (deliberate/emergent). Four perspectives on strategy are identified: the 'classical' (profit maximizing/deliberate), 'evolutionary' (profit maximizing/emergent), 'processual' (pluralistic/emergent) and the 'systemic' (pluralistic/deliberate). The processes of strategy making that are 'deliberate' comprise of rational, planned, top-down decision making ('classical' model) and decision making shaped by the social systems in which organizations are embedded ('systemic' model). 'Emergent' processes are those that reflect the very bounded, incremental, and political nature of decision making ('processual') or, in contrast, the 'population ecology' view of the competitive processes of natural selection, where the market picks the winners, rendering ineffective any deliberate strategizing ('evolutionary'). Writers typical of each of these perspectives would be Ansoff (1965) ('classical'), Hannan and Freeman (1988) ('evolutionary'), Mintzberg and Walters (1985) ('processual'), and Granovetter (1985) ('systemic').

2. *What is Strategic HRM?*

Even leaving aside the work on HRM by European writers, such as Anthony, Keenoy, Legge, Townley, and Willmott, that lies within a critical discourse tradition and which, as such, will not be considered here, there has been much debate on this question since the mid-1980s (see, e.g., Beer and Spector 1985, Fombrun et al. 1984, Guest 1987, Walton 1985). For many years, especially in the USA, human resource management was just another term for personnel management. Both terms referred to the management of the employment relationship and of the indeterminacies in the employment contract. In the early 1980s, however, a combination of factors—increased globalization of markets and perceived intensification of competition, the rise of the 'enterprise culture,' the Japanese 'Janus' and models of excellence, the decline in trade union pressure and disillusionment with 'traditional' personnel management, changes in the nature of work and in the workforce—awoke senior managers to the importance of human resources in the achievement of competitive advantage, and to a recognition that HRM/personnel management was too important to be left to personnel specialists.

Indeed, whether HRM was considered to be different from personnel management largely depended on the point of comparison. Sharp distinctions and contrasts emerged if the normative aspirations of HRM were compared to the descriptive practices of personnel management, but otherwise faded into several different emphases, all of which, though, pointed to HRM, in theory at least, being an essentially more central strategic management task than personnel management. As one commentator put it, the real difference between HRM and personnel management was 'not what it is, but who is saying it. In a nutshell HRM represents the discovery of personnel management by chief executives' (Fowler 1987, p. 3). Hence the idea of *strategic human resource management*.

However, different commentators espouse different models and approaches to strategic HRM. A major distinction may be found between the 'hard' model of the Michigan school (Fombrun et al. 1984), reflecting what has been termed a *utilitarian instrumentalism* and the 'soft,' 'mutuality' model of the Harvard school (Beer and Spector 1985, Walton 1985), more reminiscent of *developmental humanism* (Hendry and Pettigrew 1990). The 'hard' model stresses HRM's focus on the crucial importance of the close integration of human resources policies, systems, and activities with business strategy, requiring that they are not only logically consistent with and supportive of business objectives, but achieve this effect by their own coherence. From this perspective, employees are regarded as a headcount resource to be managed in exactly the same rational, impersonal way as any other resource, i.e., to be exploited for maximal economic return. Thus its focus is on human *resource management* and the implied perspective on strategy seems to have much in common with Whittington's 'classical' profit-maximizing/deliberate model of strategy.

In contrast, the 'soft,' 'developmental humanism' model, while still emphasizing the importance of integrating HR policies with business objectives, sees this as involving treating employees as valued assets, a source of competitive advantage through their commitment, adaptability, and high quality (of skills, performance, etc.). Employees are proactive and

resourceful rather than passive inputs into the productive process. Rather than exploitation and cost minimization, the watchwords in this model are investment and value added, with the focus on *human resource* management. The strategic perspective here, in Whittington's terms, is, therefore, pluralistic and echoes some of the ideas of the resource-based view of strategy.

Throughout the 1990s this model of HRM became increasingly seen, in both the USA and UK, as the exemplar of what HRM, as a distinctive philosophy of managing the employment relationship, was all about (see, e.g., Becker and Gerhart 1996, Guest 1987, Huselid 1995, MacDuffie 1995). Termed variously, the 'high commitment' (HCM), 'high performance,' or 'mutuality' model of HRM, at an operational level it was seen to comprise such policies as:

(a) Careful recruitment and selection with an emphasis on traits and competencies;

(b) Extensive use of systems of communication;

(c) Team working with flexible job design;

(d) Emphasis on training and learning;

(e) Involvement in decision making with responsibility (empowerment);

(f) Performance appraisal linked to contingent reward systems.

Increasingly, throughout the 1990s, particularly among US commentators, this model came to be equated with strategic HRM, reflecting the popularity of resource-based views of strategy in its belief in acquiring and developing employees as assets (see, e.g., Becker and Gerhart 1996, Boxall 1996). In some ways this reflects the interests of HR specialists and academics alike. Fitting HR policy and practice to business strategy always leaves HR as the 'third order' strategic decision, designed to reinforce prior 'first order' decisions about portfolio planning and 'second order' business decisions about organizational design and operating procedures, still reactive if no longer as powerless as traditional personnel management. The resource-based view of strategy accords a more proactive role to HRM as it can be valued not just for reinforcing predetermined business strategies, but for developing strategic capability, improving the long-term resilience of the firm.

However, it would be a mistake to see these two contrasting models of HRM as *necessarily* incompatible. If the stress is on integrating HRM policies with business strategy, in certain situations, choosing HCM policies may be appropriate to driving the business strategy. HCM appears the logical choice where employee discretion is crucial to the delivery of added value, and reciprocal high-trust relations are essential to organizational performance. Such a situation is often equated with organizations pursuing a strategy of producing high value-added goods and services, in a knowledge-based industry, where it adopts a policy of value-added goods and services rather than asset management. But if the organization's chosen business strategy is to compete in a labor-intensive, high-volume, low-cost industry, generating profits through increasing market share by cost leadership, such a policy of treating employees as assets may be deemed an expensive luxury. For such an organization the HR policies that may be most appropriate to driving strategic objectives are likely to involve treating employees as a variable input and cost to be minimized. Further, *human resource* management, in line with the 'soft' HCM model, is only viable if the organization is seen to be economically successful or 'hard,' if it manages effective human *resource management*.

At the heart of the distinction between 'hard' and soft' models of HRM lies a fundamental contradiction that personnel management/HRM—whatever the nomenclature—is called upon to mediate: needs it to achieve both the control and the consent of employees. Integration with business strategy may require the design and interpretation of an employment contract that veers toward control (with assumptions of labor as a commodity, embedded structural conflict, low trust, and transactional contracts) for some groups of employees; but consent (with assumptions about the employee as a valued asset, a unitary frame of reference, high trust, and relational contracts) for others.

This raises a major question about what precisely we mean by 'integrating HRM with business strategy,' the cornerstone of the 'hard' model of HRM. 'Integration' appears to have two meanings: integration or 'fit' with business strategy, and the integration or complementarity and consistency of 'mutuality' employment policies aimed at generating employee commitment, flexibility, quality, and the like. This double meaning has been referred to also as the 'external' and 'internal' fit of the HRM policies. The problem is that while 'fit' with strategy, as suggested above, would argue a *contingent* design of HRM policy, internal consistency—at least with the 'soft' human resource values associated with 'mutuality'—would argue a *universalistic* approach to the design of employment policy. Can this contradiction be reconciled without stretching to the limit the meaning of HRM as a distinct approach to managing people at work? Indeed, should we focus on HRM as a 'special variant' of personnel management, reflecting a particular ideology about how employees should be treated? Or should we regard it as a variety of very different policies and practices designed to achieve the desired employee contribution, judged solely 'against criteria of coherence and appropriateness (a less rigid term than "fit")?' (Hendry and Pettigrew 1990, pp. 8–9).

This distinction is critical as the contingent and universalistic approaches rest on very different and contradictory theoretical perspectives about organizational competitveness. The universalistic approach is consistent with institutional theory and arguments about organizational isomorphism (DiMaggio and

Powell 1983). In other words, the assumption here is that organizations that survive do so because they identify and implement the most effective, 'best' policies and practices. As a result, successful organizations get to look more and more like each other through practices like benchmarking and so on. (Here we have echoes of Whittington's 'evolutionary' perspective on strategy.) In HR terms this equates with the belief that treating employees as assets via the HCM model will always pay off, irrespective of circumstances. Contingency approaches, on the other hand, are consistent with the resource-based theories, referred to earlier, that argue that sustained competitive advantage rests not on imitating so-called 'best practice,' but on developing unique, nonimitable competencies (Barney 1991). This approach rests on a recognition of the importance of idiosyncratic contingencies that result from causal ambiguity and path dependency. From this perspective organizational performance is not enhanced merely by following 'best' HR practice (the HCM model), but from knowledge about how to combine, implement, and refine the whole potential range of HR policies and practices to suit the organization's idiosyncratic contingencies (Boxall 1996).

This brings us to what has become the key research question about strategic HRM, in both the USA and UK: what is its relationship to organizational performance?

3. *Strategic HRM and Organizational Performance*

Is strategic HRM linked to organizational performance on a universalistic (additive), configurational (patterned), or contingency (idiosyncratic) basis (Delery and Doty 1996)? At first sight the greatest support appears to be for the universalistic model: that the greater the extent to which the characteristics of the 'high commitment'/'high performance' model of HRM are adopted the better the organizational performance (see, e.g., empirical studies in both the UK and USA by Delery and Doty 1996, Guest 1997, 1999, Huselid 1995, MacDuffie 1995, Patterson et al. 1997). (For support of the contingency position, see Arthur 1992.) However, on examination, the empirical evidence for such universalism is more equivocal than it might appear at first sight.

The problems stem from methodological concerns with research design (see Legge 2000) and empirical observation (see Cully et al. 1999). Leaving aside disagreements about the specification and measurement of HCM HR practices, there are several major difficulties in drawing clear inferences from this research. Much of the work has been cross-sectional and survey-based, relying on the responses of one senior executive. The level of analysis is generally at corporate level, given the reliance on financial measures of performance, as it is at this level that much of the publicly available financial data exist. This leads to a range of difficulties. Where the research design is cross-sectional (as opposed to longitudinal), while causality may be inferred from correlation, technically it is not tested. This results in three possibilities.

A causal relationship may exist in the direction inferred, i.e., HCM HR policies and practices may give rise to positive outcomes, although even if this did exist, it might reflect no more than a temporary 'Hawthorne' effect in response to a change program (Purcell, 1999). (Of course it is possible, with the same direction of causality, that strategic HRM, particularly contingently chosen 'hard' HRM, may give rise to negative outcomes, or, positive outcomes on organizational performance measures but negative outcomes on employee measures, such as turnover and absenteeism.) Or, reverse causality may exist. Thus, as a firm becomes more profitable, it may invest in HCM HR practices, such as expenditure on training, or profit sharing, in the belief that such practices will further enhance performance, or that they will reduce the risk of performance declines, or from a belief in the justice and efficacy of wealth distribution. However, it is the profits that generate HCM practices rather than vice versa.

A further possibility is that the observed relationship between HR practices and performance may be an artefact of the implicit performance theories of survey respondents. In other words, if respondents have little detailed knowledge of the HR practices in their firm (highly likely if the firms are large, diverse and multisite and the research is being conducted at corporate level) but know that the firm is performing well in terms of productivity and profitability, they may infer that 'high performance' HR practices *must* exist, given this level of performance, based on the implicit theory that such practices are related to high performance (Wright and Gardener 2000). Similarly, the groups of employees that are likely to be most visible to such a respondent are those core employees to whom HCM is most likely to be applied, *not* the contingent workers to whom more transactional contracts are likely to be applied. The workers on the periphery of an organization may simply not be visible to the senior management single respondent to a corporate-level survey (Purcell 1999). Issues of workforce segmentation appear to be consistently neglected in such research (Boxall 1996).

Second, there is a theoretical problem with the heavy reliance on financial measures of performance (return on assets, Tobin's q) or organizational measures (productivity, quality, customer satisfaction) without equal concern to measure employee outcomes. (This is particularly marked in the US studies cited. Some recent UK studies, such as Guest 1999 and Cully et al. 1999 have attempted to measure employees' responses, such as satisfaction and employee relations climate.) The HCM model of HR practices rests on

the assumption that they will result in a positive psychological contract between employees and the organization that will give rise to employees' high commitment, quality, and flexibility, that in turn will have a positive effect on organizational performance and, ultimately, financial results. Reliance solely on financial and performance measures seems misguided as the theoretical rationales of how HR effects organizational performance rest on the assumption that it is *through* employee behaviors and outcomes. However, while the corporate level of analysis is consistent with the collection of financial data, it is not conducive to assessing employee attitudes and behaviors. Failure to open the 'black box' of employee behavior is a major limitation of many of these studies. However, to do so within the confines of quantitative positivistic designs (deemed necessary to establish generalizable causal relationships) raises immense problems about how many and what variables should be in the 'black box' and how these variables should be specified.

Then there is the issue of distinguishing between espoused and actual HR practices and employee skills and behaviors. The greater the number of intervening variables identified and the greater the level of specificity, the greater the multiplicative effect in determining the processes of a model, as the model building requires specification of the relationships between each of the specifications of the major intervening variables, giving rise to potentially unmanageable complexity (Wright and Gardner 2000).

Turning to empirical observation, we are confronted with a puzzling finding. If we are to accept at face value the research studies that suggest that there is a universalistic relationship between 'high commitment,' 'soft' model HRM and organizational performance, why do we find such relatively little implementation of this model of HRM? In the UK, for example, the latest Workplace Employee Relations Survey (WERS) in 1998 (Cully et al. 1999) found that only 14 percent of the responding workplaces have 'soft' HRM fully in place (defined as eight plus out of 15 'high commitment' management practices) as opposed to 29 percent which had three or fewer, 22 percent of which, with three or fewer HCM practices and no unions, may be defined in Guest's own memorable phrase as constituting a 'black hole.' As several commentators have observed, the implementation of any management best practice is likely to fall foul of vested interests, politicking, organizational history, and culture—precisely the factors that undermine the exercise of the 'classical' model of strategizing, but which are central to the 'processual' model. Indeed, it has been suggested that in the Anglo-American world, since about the mid-1980s, the dominant emphasis in practice has been on short-term survival rather than the development of long-term resource-based advantage, with the widespread, ad hoc adoption of the 'hard' policies of delayering, downsizing, and increasingly contingent forms of employment (Boxall 1996). Such initiatives seem consistent with Whittington's 'processual' model of strategizing, emerging as a stream of incremental, politically feasible actions in response to pressing problems.

4. *Future Directions of Research on Strategic HRM*

Whether from a practitioner or academic viewpoint, interest in the HRM/performance relationship appears to be here to stay. Most of the commentators already cited have suggestions for the improvement of research designs (e.g., more within-industry, business-level and plant-level designs; more consistent and valid measures of HR practices; more longitudinal studies; recognition of a broader range of organizational outcomes; opening up the 'black box', and so on). Guest (1997) and Purcell (1999) present two contrasting approaches, one from a universalistic and one from a contingent perspective.

Guest (1997, 1999) suggests that expectancy theory might provide a theory of process to link HRM practices and performance, as it links motivation and performance. Specifically, expectancy theory proposes that, at the individual level, high performance depends on high motivation, coupled with the necessary skills and abilities and appropriate role design and perception, equating skills and abilities with quality, motivation with commitment and role structure, and perception with flexibility. HR practices designed to foster these HR outcomes (e.g., selection and training for skills and abilities/quality, contingent pay and internal promotion for motivation/commitment, and team-working design for appropriate role design and perception/flexibility) should facilitate high individual performance, which in turn is a contributory factor in high performance outcomes (e.g., high productivity, low absence and labor turnover) that should contribute (other things being equal) to financial outcomes. In an empirical study of employees' reactions to HCM HR policies. Guest (1999) suggests that the psychological contract may be a key intervening variable in explaining the link between HR practices and employee outcomes such as job satisfaction, perceived job security, and motivation. Furthermore, although Guest's research is cross-sectional and, hence, raises the usual caveats about causality, the inferred direction of his empirical findings is supported by a similar longitudinal study (Patterson et al. 1997).

Nevertheless, Guest acknowledges the limitations of these models in terms of explaining organizational outcomes. While we may be able to measure the impact of HRM practices on HRM outcomes (quality, commitment, and flexibility), the measurable impact on organizational and financial outcomes is likely to become progressively weaker because of the range of potentially intervening variables. Guest also suggests

that we need a theory about the circumstances when human resources matter more and a theory about how much of the variance between HR practices and performance can be explained by the human factor.

Purcell (1999, pp. 36–8), building on a resource-based view of strategy has some interesting observations here. He recognizes that, on the one hand, the claim that the bundle of 'best practice' HCM HR is universally applicable 'leads us into a utopian cul-de-sac,' ignoring dual labor markets, contingent workers, and business strategies that logically do not require 'high commitment' HR practices to achieve financial success. On the other hand, the search for a contingency model of HRM is a 'chimera,' 'limited by the impossibility of modelling all the contingent variables, the difficulty of showing their interconnection, and the way in which changes in one variable have an impact on others, let alone the need to model idiosyncratic and path-dependent contingencies.' The way forward, Purcell argues, is the analysis of how and when HR factors come into play in the management of strategic change. Purcell suggests that we should explore how organizations develop successful transition management, build unique sets of competencies and distinctive organizational routines, and, in situations of 'leanness,' with greater dependency on all core workers, develop inclusiveness and the trust of such workers. The focus of research on strategic HRM (indeed, the focus of strategic HRM itself) should be on 'appropriate HR architecture and the processes that contribute to organizational performance in the short and medium term, and which positively contribute to the achievement of organizational flexibility or longevity.'

See also: Decision Making, Psychology of; Human Resource Management, Psychology of; Industrial Relations and Collective Bargaining; Industrial Sociology; Industrial–Organizational Psychology: Science and Practice; Organizational Behavior, Psychology of; Organizational Change; Organizational Decision Making; Organizations, Sociology of; Rational Choice and Organization Theory

Bibliography

Ansoff H 1965 *Corporate Strategy*. Penguin, Harmondsworth, UK

Arthur J B 1992 The link between business strategy and industrial relations systems in American steel minimills. *Industrial and Labor Relations Review* **45**: 488–506

Barney J 1991 Firm resources and sustained competitive advantage. *Journal of Management* **17**(1): 99–120

Becker B, Gerhart B 1996 The impact of human resource management on organizational performance: Progress and prospects. *Academy of Management Journal* **39**(4): 779–801

Beer M, Spector B 1985 Corporate-wide transformations in human resource management. In: Walton R E, Lawrence P R (eds.) *Human Resource Management, Trends and Challenge*. Harvard Business School Press, Boston, pp. 219–53

Boxall P 1996 The strategic HRM debate and the resource-based view of the firm. *Human Resource Management Journal* **6**(3): 59–75

Cully M, Woodland S, O'Reilly A, Dix G 1999 *Britain at Work*. Routledge, London

Delery J, Doty H 1996 Models of theorizing in strategic human resource management: Tests of universalistic, contingency, and configurational performance predictions. *Academy of Management Journal* **39**(4): 802–35

DiMaggio P, Powell W 1983 The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review* **48**(2): 147–60

Fombrun C, Tichy N M, Devanna M A (eds.) 1984 *Strategic Human Resource Management*. Wiley, New York

Fowler A 1987 When chief executives discover HRM. *Personnel Management* **19**(1): 3

Granovetter M 1985 Economic action and social structures: The problem of embeddedness. *American Journal of Sociology* **91**(3): 481–510

Guest D E 1987 Human resource management and industrial relations. *Journal of Management Studies* **24**(5): 503–21

Guest D E 1997 Human resource management and performance: A review and research agenda. *International Journal of Human Resource Management* **8**(3): 263–90

Guest D E 1999 Human resource management—the workers' verdict. *Human Resource Management Journal* **9**(3): 5–25

Hannan M T, Freeman J 1988 *Organizational Ecology*. Cambridge University Press, Cambridge, MA

Hendry C, Pettigrew A M 1990 Human resource management: An agenda for the 1990s. *International Journal of Human Resource Management* **1**(1): 17–43

Huselid M 1995 The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal* **38**: 635–72

Legge K 2000 Silver bullet or spent round? Assessing the meaning of the 'high commitment management'/performance relationship. In: Storey J (ed.) *Human Resource Management: A Critical Text*. Thomson Learning, London, pp. 20–36

MacDuffie J P 1995 Human resource bundles and manufacturing performance: Organizational logic and flexible production systems in the world auto industry. *Industrial and Labor Relations Review* **48**: 197–221

Miles R E, Snow C C 1978 *Organizational Strategy, Structure and Process*. McGraw-Hill, New York

Mintzberg H, Walters J A 1985 Of strategies, deliberate and emergent. *Strategic Management Journal* **6**: 257–72

Patterson M, West M, Lawthom R, Nickell S 1997 *Impact of People Management Practices on Business Performance*. IPD, London

Porter M 1980 *Competitive Strategy*. Free Press, New York

Purcell J 1999 Best practice and best fit: Chimera or cul-de-sac? *Human Resource Management Journal* **9**(3): 26–41

Walton R E 1985 From control to commitment in the workplace. *Harvard Business Review* **63**(2): 77–84

Whittington R 1993 *What is Strategy and Does it Matter?* Routledge, London

Wright P M, Gardener T M 2000 Theoretical and empirical challenges in studying the HR practice-performance relationship. Paper presented at the special workshop on Strategic Human Resource Management, EIASM, INSEAD, Fontainebleau, France, 30 March–1 April

K. Legge

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Strategic Intelligence

Intelligence is the provision of information about targets of concern, mostly foreign, for the use of decision makers, mostly governmental. The information normally prioritizes secrets and/or forecasts, and the degrees of animosity toward the targets (ranging from hostility all the way over to friendliness) are likely to shape the various efforts. The intelligence that is called 'strategic' is especially alert to the more competitive ranges of this spectrum, and is also loosely distinguished from what is called 'tactical' intelligence—the strategic seen as having wider scope, longer time horizons, and organizationally higher standing. Such strategic intelligence is several thousand years old, yet it lacked widely known, sharply focused manuals and treatises until the twentieth century (except in the narrow domain of cryptology).

The long delay was no doubt related to the traditional secretiveness of most of the intelligence activities—but also to slowness in differentiating them from broad statecraft and diplomacy and generalship, and from each other, and eventually institutionalizing various intelligence functions: distinct modes of information collection and analysis and interface with decision makers, and systematic counterintelligence shielding of one's own efforts from those of others. With these developments in the twentieth century have come incentives for elaborate doctrine and theorizing inside the intelligence organizations—and outsiders are increasingly being allowed to play a part, often driven also by their own concerns for social science applications and/or for greater democratic accountability. A significant literature is emerging.

1. *History of the Emergence of Intelligence Institutions and a Literature*

Some of the most ancient Mesopotanian tablets are cryptograms, and Western memory has highlighted espionage activities since the sieges of Troy and Jericho. Systematic literature on generalship, statecraft, and diplomacy became abundant, and commonly included passages on the need to have spies and to do assessments of information, but little, however, about exactly how to do the assessments, and even less about the subtleties of exactly how to direct one's spies, and almost nothing about how to be a spy and perform espionage. For the latter, from among published works that ever achieved wide circulation, one would have had to rely on some fiction and a few memoirs.

Sheer secretiveness does not suffice to explain the slowness of emergence of a systematic intelligence literature: such literature began to appear with the emergence of modern science—and appropriately was first about cryptology (ca. 1550), the aspect of intelligence that could most readily benefit from the application of the mathematics, etc., with 'how to' orderliness. Traditional espionage had required secretiveness too—but about methods that were thought to be sufficiently inculcated through age-old conventional wisdom and craftsmanship/apprenticeship not to require formal treatises for transmission of new specialties to novice practitioners or for continued scholarly updating, as cryptology soon did.

Many advances in science and technology were of course also contributing importantly to a general, lurching modernization of warfare and statecraft in the seventeenth and eighteenth centuries—and to a precipitation (which finally became rapid in the late nineteenth century) of new institutionalized specialties, crystalized out of the age-old fluid mixes of generalship, statecraft, and diplomacy. Again cryptology helped lead the way, from 'black chambers' of the seventeenth century to 'signals intelligence' (SIGINT), salient with the advent of telegraphy.

But until after Napoleon, the crystalization of intelligence functions in Europe was mostly discontinuous, ad hoc, and personalistic. Generals and admirals, rulers and their top advisers constructed intelligence outfits (usually rather small) as they saw fit and employed them idiosyncratically. There was a general presumption that diplomats would engage in espionage, discreetly, as part of their role as reporters to their home governments; likewise postal services dutifully would intercept accessible messages; merchant captains would be available to be debriefed, similarly traders and merchant bankers; clerics would commonly collaborate; and so forth, ad hoc.

Yet eventually the long century of relative peace after 1815 enabled the career diplomats to consolidate for themselves, as reporters, a professional status and ethic of personal abstention from conducting espionage—though their embassies and consulates might continue to harbor other functionaries in roles that did provide cover for agent running. Also, military attaches (likewise naval and later air) acquired a formally recognized 'licensed spy' status that was mutually convenient to the governments concerned. And the spread of democracy led to emphasis on differentiating one's counterintelligence security arrangements, domestic from international.

In the twentieth century, technical collection (TECHINT) methods proliferated, utilizing new technologies in SIGINT and beyond it: notably overhead reconnaissance (imagery intelligence, IMINT). Such new or reconstituted intelligence specialties, specifically identified as such, were incompatible with continuation of ad hocery; they required distinct stable career bureaucracies, at least within the armed services, and probably also outside lest political, economic, and other civilian priorities be absorbed subordinately into military juggernauts, even in peacetime. Instead, civilian intelligence services could do some of the espionage and could highlight open

sources of information, contributing to a more complete ('all-source') system of collection. To be sure, that would complicate coordination in all-source assessment/analysis of the collected information and in dissemination of it to the policy makers. So additional new institutions would be needed to reintegrate the recently proliferating institutions. And each of them, old and new, would come to need doctrines, for its own stable identity and its missions.

All of this was in fact the swiftly evolving setting for twentieth-century discourse about strategic intelligence.

2. *Components of the Twentieth Century Intelligence Literature*

Informed discourse about intelligence requires access to evidence. Yet successful conduct of intelligence requires much secretiveness and deception. The conflict between these two principles is the central dilemma retarding scholarship about the subject—in this era when the complexity and importance of the functions and institutions has finally achieved general recognition, while the prevailing standards of respectability in social science do not comport easily with the dubious quality of evidence available for open discourse about intelligence. The field is nearly unique in that evidence is not only hard to find because it is deliberately hidden but additionally is hard to trust because often it has been deliberately planted to deceive. At least some governmental relaxation of long enforced secrecy restrictions is virtually prerequisite for building an evidentiary base usable by outsiders. Yet various possible governmental motives for relaxation rouse suspicions that even the releases are being directly or indirectly rigged. Governments do hold the keys. Investigative journalists, especially those with current events preoccupations, can be stymied; defectors and adversaries can be denigrated and largely discredited; veterans can be held to oaths of silence; archives can be kept closed and/or purged. Nevertheless, whenever stonewalling erodes and instead validation gradually proceeds, any and all of these potential sources can come cumulatively to constitute the evidentiary base for informed open discourse.

That is what has in fact been slowly occurring—especially with momentum from the United States—in a widening circle of democratic and ex-communist countries during the decades since the USA's traumatic encounter with Vietnam. Exposes have led step by step to treatises. And much that in the past had indeed been factual but had been presented as intelligence fiction, so as to elude restrictions, can now be accorded less ambiguous credibility, even become part of the accepted data base. Historiography and some social science can take hold.

Among outside historians, about twentieth-century intelligence the most outstanding achievement thus far is the UK's Christopher Andrew's successive masterworks (1985, 1995, Andrew and Gordievsky 1990) on the services of each of the major intelligence powers. For coverage of earlier centuries also, the general masterwork is German: (Piekalkiewicz 1988); about cryptology it is American (Kahn 1996). Now, even for intelligence history as recent as the Cold War, multi-archival international research collaborations are sometimes successful, e.g. Russians and Americans (Fursenko and Natfali 1997, Murphy et al. 1997). For sheer massive contemporary updated data accumulation by an outsider no one anywhere rivals the USA's Jeffrey Richelson (1999).

But most important for durably connecting strategic intelligence with social science are the works that do not just add to the base of openly available evidence but use it to rethink and generalize intelligence problems. In this genre, usually the writers who have been best positioned to take early advantage of the new permissiveness come from among former intelligence practitioners (or supervisory associates) who happen to be of uncommonly reflective—even scholarly—disposition, have familiarized themselves with the open literature currently available in their respective periods, and soberly resolve to contribute their own perspectives to its cumulative advancement. Milestone works of this kind in the United States have been Kent (1966), Johnson (1989), Shulsky (1991), and in the UK, Herman (1996), all of them distinguished for comprehensiveness as well as for insider/outsider complementarity of strengths.

What kinds of contributions can they make, in the widening discourse based on increasingly available data and personal contacts? And how about authors who remain essentially outsiders yet want to study the arcane world of ongoing intelligence—not just its history? What does the existing literature already exhibit and foreshadow about the feasibility and value of various research directions?

Scholars can act like investigative journalists and 'speak the unspeakable,' in reports that are also additional evidence for generalization by others.

Scholars can act as systematic ethicists, judging different forms or instances of intelligence activity. Much of such effort has been devoted to evaluating covert action (intrigue, etc.), but not much yet about intelligence in the stricter sense, where problems like loyalties and means–ends relationships also abound, worth normative exploration.

Epistemologists and cognitive psychologists have wide room to apply their disciplinary skills studying problems of perception/misperception/deception that arise in intelligence collection and assessment, and in counterintelligence where these complexities are loosely tagged a 'wilderness of (interfacing) mirrors.' The pioneering studies here were done as academic political psychology (Jervis 1970, 1976).

There is room for management-school faculties to contribute insights from their fields of management

psychology and of operations research. Here quantitative methods are central whereas most other outsiders' work on intelligence, thus far, has been entirely qualitative or only rudimentally quantitative. However, rational-choice/game-theory modelers' increasing emphasis on their players' information levels may come to alter this imbalance, if the applicability to intelligence studies gets recognized.

Political scientists as well as psychologists can conceptualize and taxonomize in their studies of intelligence. This is a function that insiders are very familiar with on their own; they eagerly construct, debate, and reconstruct their secret lexicons, knowing full well the value of verbal precision (and deliberate imprecision). But such efforts on the inside are prone to express conflicting parochialisms.

Taxonomies by outsiders may be more objective, a moderating influence, and sometimes their conceptualization can even introduce or clarify ideas that have not already become familiar to insiders blinkered by their own secret terminology.

Similarly, outside social scientists can sometimes venture into speculative scenario writing, which implicitly can be a useful cross-check upon the range and content of comparable exercises done inside, even when the outsiders are not privy enough to the latter for the cross-checking to be made explicit.

Political scientists as well as like-minded historians can make cross-national comparisons leading to some mid-level theorizing–for example about how much the interface between intelligence analysts and top policymakers is affected by characteristics of different countries' governmental systems, be they pluralistic or relatively centralized.

All these modes of widening discourse about strategic intelligence, besides having benefits for the scholars and some practitioners, contribute to making civil society in general more knowledgeable about this important dimension of world affairs. Thus the discourse can tend to enhance broad democratic accountability, with long-term advantages that probably outweigh the occasional damage from excessive leaks and disclosures. But wide acquiescence by insiders comes only very slowly.

See also: Cold War, The; Information and Knowledge: Organizational; Information Technology; Intelligence: Organizational; Knowledge Societies; Military Sociology; Organizations, Sociology of; Technology and Social Control; War, Sociology of

Bibliography

Andrew C 1985 *Her Majesty's Secret Service: The Making of the British Intelligence Community*. Viking, New York

Andrew C 1995 *For the President's Eyes Only: Secret Intelligence and the American Presidency from Washington to Bush*. HarperCollins, New York

Andrew C, Gordievsky O 1990 *KGB: The Inside Story*. 1st edn. HarperCollins, New York

Fursenko A V, Naftali T 1997 *One Hell of a Gamble: Khrushchev, Castro, and Kennedy, 1958-1964*. 1st edn. Norton, New York

Herman M 1996 *Intelligence Power in Peace and War*. Cambridge University Press, Cambridge, UK

Jervis R 1970 *The Logic of Images in International Relations*. Princeton University Press, Princeton, NJ

Jervis R 1976 *Perception and Misperception in International Politics*. Princeton University Press, Princeton, NJ

Johnson L K 1989 *America's Secret Power: The CIA in a Democratic Society*. Oxford University Press, Oxford, UK

Kahn D 1996 *The Codebreakers: The Story of Secret Writing*, rev. edn. Scribner, New York

Kent S 1966 *Strategic Intelligence for American World Policy*, rev. edn. Princeton University Press, Princeton, NJ

Murphy D E, Kondrashev S A, Bailey G 1997 *Battleground Berlin: CIA vs. KGB in the Cold War*. Yale University Press, New Haven, CT

Piekalkiewicz J 1988 *Weltgeschichte der Spionage: Agenten, Systeme, Aktionen* [World History of Espionage: Agents, Systems, Operations]. Sudwest Verlag, Munich, Gemany

Richelson J T 1999 *The U.S. Intelligence Community*, 4th edn. Westview, Boulder, CO

Shulsky A N 1991 *Silent Warfare: Understanding the World of Intelligence*. Brassey's (US), Washington, DC

H. B. Westerfield

Strategic Planning

Strategic planning may be defined as 'a disciplined effort to produce fundamental decisions and actions that shape and guide what an organization (or other entity) is, what it does, and why it does it' (Bryson 1995, pp. 4–5). This definition implies that strategic planning is not a single thing, but is instead a set of concepts, procedures, and tools that must be carefully tailored to situations if desirable outcomes are to be achieved. Many authors have suggested generic strategic planning processes. Poister and Streib (1999) in a review of the field of strategic planning and management assert that strategic planning:

(a) Is concerned with identifying and responding to the most fundamental issues facing an organization;

(b) Addresses the subjective question of purpose and the often-competing values that influence mission and strategies;

(c) Emphasizes the importance of external trends and forces, as they are likely to affect the organization and its mission;

(d) Attempts to be politically realistic by taking into account the concerns and preferences of internal, and especially external, stakeholders;

(e) Relies heavily on the active involvement of senior level managers, and sometimes elected officials, assisted by staff support where needed;

(f) Requires the candid confrontation of critical issues by key participants in order to build commitment to plans;

(g) Is action oriented and stresses the importance of developing plans for implementing strategies; and

(h) Focuses on implementing decisions now in order to position the organization favorably for the future.

It is important to highlight what strategic planning is not. Strategic planning is not a substitute for strategic thinking and acting. It may help people do that, but used unwisely it may hinder strategic thinking and acting. Strategic planning is not a substitute for leadership. At least some key actors must be committed to the process or it is bound to fail. Strategic planning is also not a substitute for an organizational or community strategy. Strategies have numerous sources, both planned and unplanned. Strategic planning is likely to result in a statement of organizational or community intentions, but what is realized in practice will be some combination of what is intended with what emerges along the way. Finally, strategic planning is not synonymous with what is called 'comprehensive planning' for communities in the USA, or has been called 'structure planning' in Europe. There may be little difference if the agency doing the comprehensive or structure planning is tied directly to governmental decision-makers. However, in practice, there may be three significant differences.

First, the plans are often prepared to meet legal requirements and must be formulated according to a legally prescribed process with legally prescribed content. As legal instruments, these plans have an important influence. On the other hand, the plans' typical rigidity can conflict with the political process with which public officials must cope. Therefore strategic plans can provide a bridge from legally required and relatively rigid policy statements to actual decisions and operations.

Second, comprehensive or structure plans are usually limited to a shorter list of topics than a government's full agenda of roles and responsibilities. For that reason, they are of less use to key decision makers than strategic planning, which can embrace all of a government's actual and potential roles before deciding why, where, when, and how to act.

Third, as Kaufman and Jacobs (1987) argue, strategic planning on behalf of a community is more action-oriented, more broadly participatory, more emphatic about the need to understand the community's strengths and weaknesses as well as the opportunities and threats it faces, and more attentive to inter-community competitive behavior. Strategic planning is thus typically more comprehensive than comprehensive planning or structure planning, while at the same time producing a more limited action focus.

Strategic planning also is typically distinguished from strategic management. Strategic management is a far more encompassing process 'concerned with managing an organization in a strategic manner on a continuing basis' (Poister and Streib 1999, p. 310). Strategic management links strategic planning and implementation by adding ongoing attention to budgeting; performance measurement, evaluation, and management; and feedback relationships among these elements.

1. *Brief History of Development and Use of Strategic Planning*

Strategic planning for organizations has primarily developed in the private sector. This history has been amply documented by others (Eden and Ackermann 1998). However, since around 1980, public and nonprofit use of strategic planning has skyrocketed. This experience, and a growing body of literature, have indicated that strategic planning approaches either developed in the private sector, or else strongly influenced by them, can help public and nonprofit organizations, as well as communities or other entities, deal in effective ways with their dramatically changing environments.

In the USA—where strategic planning has proceeded the furthest—a substantial majority of municipal and state governments, and an overwhelming majority of federal agencies and nonprofit organizations, make use of strategic planning. Strategic planning has been principally applied to public and nonprofit organizations, but applications to communities have also increased substantially (Berman and West 1998). Other nations are also making increased use of strategic planning concepts, procedures, and tools for public and nonprofit organizations and communities (e.g., Osborne and Plastrik 1997, Salets and Faludi 2000).

2. *Benefits*

When done well, strategic planning offers a number of benefits. Advocates usually point to four main potential benefits. The first is the 'promotion of strategic thought and action.' The second is 'improved decision making,' while the third is 'enhanced organizational responsiveness and improved performance.' Finally, strategic planning can 'directly benefit the organization's people' by helping them better perform their roles, meet their responsibilities, and enhance teamwork and expertise. There is no guarantee, however, that these benefits will be achieved. For one thing, strategic planning is simply a set of concepts, procedures, and tools that must be applied wisely to specific situations. Also, even when applied wisely there is no guarantee of success.

3. Approaches to Strategic Planning

Strategic planning may be divided into 'process' and 'content' approaches. We start first with broadly encompassing process approaches, then consider more narrowly conceived process approaches, and finally discuss two content approaches.

3.1 The Harvard Policy Model

The Harvard policy model was developed as part of the business policy courses taught at the Harvard Business School since the 1920s (Bower et al. 1993). The approach provides the principal (though often implicit) inspiration behind the most widely cited recent models of public and nonprofit sector strategic planning. The main purpose of the Harvard model is to help an organization develop the best 'fit' between itself and its environment; that is, to develop the best strategy for the organization. One discerns the best strategy by analyzing the internal strengths and weaknesses of the company and the values of senior management, and by identifying the external threats and opportunities in the environment and the social obligations of the firm. Then one designs the appropriate organizational structure, processes, relationships, and behaviors necessary to implement the strategy, and focuses on providing the leadership necessary to implement the strategy.

Effective use of the model presumes that senior management can agree on the organization's situation and the appropriate strategic response, and has enough authority to enforce its decisions. A final important assumption of the model, common to all approaches to strategic planning, is that if the appropriate strategy is identified and implemented, the organization will be more effective. For the model to be useful for public and nonprofit purposes, it typically has to be supplemented with other approaches, such as the portfolio and strategic issues management approaches discussed below. A portfolio approach is needed because a principal strategic concern at the organizational level is oversight of a portfolio of agencies, departments, or programs. Strategic issues management is needed because much public and nonprofit work is typically quite political, and articulating and addressing issues is the heart of political decision making.

The systematic assessment of strengths, weaknesses, opportunities and threats—known as a SWOT analysis—is a primary strength of the Harvard model. This element of the model appears to be applicable in the public sector to organizations, functions, and communities. Another strength is its emphasis on the need to link strategy formulation and implementation in effective ways. The main weaknesses of the Harvard model are: (a) it does not draw attention to strategic issues; (b) it does not consider the full range of stakeholders; and (c) it does not offer specific advice on how to develop strategies, except to note that effective strategies will build on strengths, take advantage of opportunities, and overcome or minimize weaknesses and threats.

3.2 Strategic Planning Systems

Strategic planning is often viewed as a system whereby leaders and managers go about making, implementing, and controlling important decisions across functions and levels in the organization. When viewed this way, strategic planning is synonymous with strategic management. Strategic planning systems vary along several dimensions: the comprehensiveness of decision areas included, the formal rationality of the decision process, and the tightness of control exercised over implementation of the decisions, as well as how the strategy process itself will be tailored to the organization and managed. The strength of these systems is their attempt to coordinate the various elements of an organization's strategy across levels and functions. Their weakness is that excessive comprehensiveness, prescription, and control can drive out attention to mission, strategy, and innovation, and can exceed the ability of participants to comprehend the system and the information it produces (Mintzberg 1994, Roberts and Wargo 1994).

Strategic planning systems are applicable to public organizations (and to a lesser extent communities), for, regardless of the nature of the particular organization, it makes sense to coordinate decision making across levels and functions and to concentrate on whether the organization is implementing its strategies and accomplishing its mission (Boschken 1988, p. 1994). However, it is important to remember that a strategic planning system characterized by substantial comprehensiveness, formal rationality in decision making, and tight control will work only in an organization that has a clear mission, clear goals and objectives, relatively simple tasks to perform, centralized authority, clear performance indicators, and information about actual performance available at reasonable cost. While some public and nonprofit organizations—such as hospitals and police and fire departments—operate under such conditions, most do not. As a result, most public and nonprofit sector strategic planning systems typically focus on a few issues, rely on a decision process in which politics plays a major role, and control something other than program outcomes (e.g., budget expenditures) (Osborne and Plastrik 1997).

3.3 Stakeholder Management Approaches

Freeman (1984) states that strategy can be understood as an organization's mode of relating to or building

bridges to its stakeholders. For Freeman, a stakeholder may be defined as any individual, group, or organization that is affected by, or that can affect, the future of the organization. He argues, as do others who emphasize the importance of attending to stakeholders, that a strategy will only be effective if it satisfies the needs of multiple groups. Because it integrates economic, political, and social concerns, the stakeholder model is one of the approaches that is most applicable to the public sector. Many interest groups have stakes in public organizations, functions, and communities. If the model is to be used successfully, however, there must be the possibility that key decision-makers can achieve reasonable agreement about who the key stakeholders are and what the response to their claims should be.

The strengths of the stakeholder model are its recognition of the many claims—both complementary and competing—placed on organizations by insiders and outsiders and its awareness of the need to satisfy at least the key stakeholders if the organization is to survive. Because of its attention to stakeholders, the approach may be particularly useful in working collaboratively with other organizations, since each party to a collaboration must be viewed as a stakeholder (Huxham 1993). The weaknesses of the model are the absence of criteria with which to judge competing claims and the need for more advice on developing strategies to deal with divergent stakeholder interests. A number of other encompassing process approaches to strategic planning have been developed primarily in the UK in the field of operations research. These include Journey Making (Eden and Ackermann 1998) and Strategic Choice (Friend and Hickling 1997). They are now beginning to be used more widely around the world.

3.4 Strategic Issues Management Approaches

These approaches are process components, pieces of a larger strategic planning process. The concept of strategic issues first emerged when practitioners of corporate strategic planning realized a step was missing between the SWOT analysis of the Harvard model and the development of strategies. That step was the identification of strategic issues, or crucial challenges facing the organization. Many organizations now include a strategic issue identification step as part of full-blown strategy revision exercises, and also as part of less comprehensive annual strategic reviews between major strategy revisions. In addition, many organizations have developed strategic issue management processes actually separated from annual strategic planning processes. Many important issues emerge too quickly to be handled as part of an annual process.

Strategic issue management is clearly applicable to public organizations, since the agendas of these organizations consist of issues that should be managed strategically (Eden and Ackermann 1998). The strength of the approach is its ability to recognize and analyze key issues quickly. The approach also applies to functions or communities, as long as some group, organization, or coalition is able to engage in the process and to manage the issue. The main weakness is that in general the approach offers no specific advice on exactly how to frame the issues other than to precede their identification with a situational analysis of some sort. Nutt and Backoff (1995) have gone the furthest in remedying this defect. The final two process approaches to be discussed are 'process strategies.' They are logical incrementalism and strategic planning as a framework for innovation. Process strategies are approaches to implementing a strategy that already has been developed in very broad outline and is subject to revision based on experience with its implementation. Other important process approaches not discussed in this article, due to space limitations, include total quality management (Cohen and Brand 1993), strategic negotiations (Susskind and Cruikshank 1987), collaboration (Gray 1989), and the management of culture (Schein 1992).

3.5 Logical Incrementalism

In incremental approaches, strategy is a loosely linked group of decisions that are handled incrementally. Decisions are handled individually below the organizational level because such decentralization is politically expedient—organizational leaders should reserve their political clout for crucial decisions. Decentralization also is necessary since often only those closest to decisions have enough information to make good ones.

The incremental approach is identified principally with Quinn (1980), although the influence of Lindblom (1959) is apparent. Quinn developed the concept of 'logical incrementalism'—or incrementalism in the service of overall organizational purposes—and as a result transformed incrementalism into a strategic approach. Logical incrementalism is a process approach that, in effect, fuses strategy formulation and implementation. The strengths of the approach are its ability to handle complexity and change, its emphasis on minor as well as major decisions, its attention to informal as well as formal processes, and its political realism. A related strength is that incremental changes in degree can add up over time into changes in kind (Mintzberg 1987). The major weakness of the approach is that it does not guarantee that the various loosely linked decisions will add up to fulfillment of corporate purposes.

Logical incrementalism would appear to be very applicable to public and nonprofit organizations, as long as it is possible to establish some overarching set of strategic objectives to be served by the approach.

When applied at the community level, there is a close relationship between logical incrementalism and collaboration. Indeed, collaborative purposes and arrangements typically emerge in an incremental fashion as organizations individually and collectively explore their self-interests and possible collaborative advantages, establish collaborative relationships, and manage changes incrementally within a collaborative framework (Stone 2000).

3.6 Strategic Planning as a Framework for Innovation

As noted in Sect. 3.5, strategic planning systems can drive out attention to mission, strategy, and innovation. The systems, in other words, can become ends in themselves and drive out creativity, innovation, and new ventures without which the organization might become irrelevant or die. Many organizations, therefore, have found it necessary to emphasize innovative strategies as a counterbalance to the excessive control orientation of many strategic planning systems (Light 1998). The framework-for-innovation approach relies on many elements of the approaches discussed above, such as SWOT analyses and portfolio methods. This approach differs from earlier ones in four emphases: (a) innovation as a strategy; (b) specific management practices to support innovation; (c) development of a mission or 'vision of success' that provides the decentralized and entrepreneurial parts of the organization with a common set of super-ordinate goals toward which to work; and (d) nurture of an entrepreneurial organizational culture (Light 1998, Osborne and Plastrik 1997).

The main strength of the approach is that it allows for innovation and entrepreneurship while maintaining central control. The weaknesses of the approach are that typically—and perhaps necessarily—a great many, often costly, mistakes are made as part of the innovation process and that there is a certain loss of accountability in very decentralized systems (Mintzberg 1994). Those weaknesses reduce the applicability to the public sector, in particular, in which mistakes are less acceptable and the pressures to be accountable for details (as opposed to results) are often greater (Light 1998). Many nonprofit organizations will also have trouble pursuing this approach because a shortage of important resources will magnify the risks of failure. 'Process' approaches do not prescribe answers, though good answers are presumed to emerge from appropriate application. In contrast, 'content' approaches do yield answers. In fact, the models are antithetical to process when process concerns get in the way of developing the 'right' answers. In the next two sections, we discuss two content approaches: portfolio models and competitive analysis. Other important content approaches not covered in this article, due to space limitations, include 'reinventing government' (Osborne and Gaebler 1992), systems analysis (Senge 1990), and reengineering (Linden 1994).

3.7 Portfolio Models

The idea of strategic planning as managing a portfolio of 'businesses' is based on an analogy with investment practice in which an investor manages a portfolio of stocks to balance risk and return. Although the applications of portfolio theory to the public sector may be less obvious than those of the approaches described above, they are nonetheless just as powerful (Nutt and Backoff 1992). Many public and nonprofit organizations consist of 'multiple businesses' that are only marginally related. Often resources from various sources are committed to these unrelated businesses. That means public and nonprofit managers must make complex portfolio decisions, though usually without the help of portfolio models that frame those decisions strategically. The strength of portfolio approaches is that they provide a method of measuring entities of some sort (e.g., programs, proposals, or problems) against dimensions that are deemed to be of strategic importance (usually attractiveness of the option versus capability to implement it). Weaknesses include the difficulty of knowing what the appropriate strategic dimensions are, difficulties of classifying entities against dimensions, and the lack of clarity about how to use the tool as part of a larger strategic planning process.

Portfolio approaches can be used in conjunction with an overall strategic planning process to provide useful information on an organization, function, or community in relation to its environment. However, unlike the process models portfolio approaches provide an 'answer'; that is, once the dimensions for comparison and the entities to be compared are specified, the portfolio models prescribe how the organization or community should relate to its environment. Such models will work only if a dominant coalition is convinced that the answers they produce are correct.

3.8 Competitive Analysis

Another important content approach that assists strategy selection has been developed by Michael Porter (1998a) and his associates. Porter (1998a) hypothesizes that five key competitive forces shape an industry: relative power of customers, relative power of suppliers, threat of substitute products, threat of new entrants, and the amount of rivalry activity among the players in the industry. For many public and nonprofit organizations, there are equivalents to the forces that affect private industry. An effective organization in the public and nonprofit sector,

therefore, must understand the forces at work in its 'industry' in order to compete effectively, and must offer value to its customers that exceeds the cost of producing it. On another level, planning for a specific public function (health care, transportation, or recreation) can benefit from competitive analysis if the function can be considered an industry.

In addition, Porter (1998b) points out that for the foreseeable future self-reinforcing agglomerations of firms and networks are crucial aspects of successful international economic competition. In effect, not just firms or nations, but metropolitan regions (Shanghai, the Silicon Valley, New York, London, and Paris) are key economic actors. Regions interested in competing on the world stage should therefore try to develop the infrastructure necessary for virtuous (rather than vicious) cycles of economic growth to unfold. In other words, wise investments in education, transportation and transit systems, water and sewer systems, parks and recreation, housing, and so on, can help firms reduce their costs—particularly the costs of acquiring an educated labor force—and thus improve the firms' ability to compete internationally.

The strength of competitive analysis is that public and nonprofit organizations can use competitive analysis to discover ways to help the private firms in their regions. When applied directly to public and nonprofit organizations, however, competitive analysis has two weaknesses: it is often difficult to know what the 'industry' is and what forces affect it, and the key to organizational success in the public and nonprofit world is often collaboration instead of competition. Competitive analysis for the public organizations, therefore, must be coupled with a consideration of social and political forces and the possibilities for collaboration (Huxham 1993, Osborne and Plastrik 1997).

4. Conclusions

It should be noted that careful studies of corporate-style strategic planning in the public and nonprofit sectors are few in number. Nevertheless, it is possible to reach some tentative conclusions. First, it should be clear that strategic planning is not a single concept, procedure, or tool. In fact, it embraces a range of approaches that vary in their applicability to the public and nonprofit sectors and in the conditions that govern their successful use. Second, while any generic strategic planning process may be a useful guide to thought and action, it will have to be applied with care in a given situation, as is true of any planning process (Christensen 1999). Because every planning process should be tailored to fit specific situations, every process in practice will be a hybrid. Third, public and nonprofit strategic planning is well on its way to becoming part of the standard repertoire of public and nonprofit leaders, managers, and planners. Because of the often dramatic changes these people and their organizations confront, we can hypothesize that the most effective leaders, managers, and planners at the beginning of the twenty-first century are, and will be increasingly in the future, the ones who are best at 'strategic' planning.

Finally, research must explore a number of theoretical and empirical issues in order to advance the knowledge and practice of public-sector strategic. In particular, strategic planning processes that are responsive to different situations must be developed and tested. These processes should specify key situational factors governing their use; provide specific advice on how to formulate and implement strategies in different situations; be explicitly political; indicate how to deal with plural, ambiguous, or conflicting goals or objectives; link content and process; indicate how collaboration as well as competition is to be handled; and specify roles for those involved in the process. Other topics in need of attention include: the nature of strategic leadership; ways to promote and institutionalize strategic planning across organizational levels, functions that bridge organizational boundaries, and intra- and interorganizational networks; and the ways in which information technologies can help or hinder the process. Progress has been made on all of these fronts, but work is clearly necessary if we are to understand better when and how to use strategic planning.

See also: Nonprofit Sector, Geography of; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Public Goods: International; Strategy: Organizational

Bibliography

Berman E M, West J P 1998 Productivity enhancement efforts in public and nonprofit organizations. *Public Productivity and Management Review* **22**: 207–19

Boschken H L 1988 *Strategic Design and Organizational Change*. The University of Alabama Press, Tuscaloosa, AL

Boschken H L 1994 Organizational performance and multiple constituencies. *Public Administration Review* **54**: 308–12

Bower J, Bartlett C, Chistensen C, Pearson J 1991 *Business Policy: Text and Cases*, 7th edn. Irwin, Homewood, IL

Bryson J M 1995 *Strategic Planning for Public and Nonprofit Organizations*, rev. edn. Jossey-Bass, San Francisco, CA

Christensen K S 1999 *Cities and Complexity: Making Intergovernmental Decisions*. Sage, Los Angeles, CA

Cohen S, Brand R 1993 *Total Quality Management in Government*. Jossey-Bass, San Francisco, CA

Eden C, Ackermann F 1998 *Making Strategy: The Journey of Strategic Management*. Sage, Los Angeles, CA

Freeman R E 1984 *Strategic Management: A Stakeholder Approach*. Pitman, Boston

Friend J, Hickling A 1997 *Planning Under Pressure: The Strategic Choice Approach*. Butterworth-Heinemann, Oxford, UK

Gray B 1989 *Collaborating: Finding Common Ground for Multiparty Problems*. Jossey-Bass, San Francisco, CA

Huxham C 1993 Pursuing collaborative advantage. *Journal of the Operational Research Society* **44**: 599–611
Kaufman J, Jacobs H M 1987 A public planning perspective on strategic planning. *Journal of the American Planning Association* **53**(1): 23–33
Light P C 1998 *Sustaining Innovation: Creating Nonprofit and Government Organizations that Innovate Naturally*. Jossey-Bass, San Francisco, CA
Lindblom C E 1959 The science of muddling through. *Public Administration Review* **19**: 79–88
Linden R M 1994 *Seamless Government: A Practical Guide to Re-Engineering in the Public Sector*. Jossey-Bass, San Francisco, CA
Mintzberg H 1987 Crafting strategy. *Harvard Business Review* **87**(4): 66–75
Mintzberg H 1994 *The Rise And Fall Of Strategic Planning*. Free Press, New York
Nutt P C, Backoff R W 1992 *Strategic Management for Public and Third Sector Organizations: A Handbook for Leaders*. Jossey-Bass, San Francisco, CA
Nutt P C, Backoff R W 1995 Strategy for public and third sector organizations. *Journal of Pubic Administration Research and Theory* **5**(2): 189–211
Osborne D, Gaebler T 1992 *Reinventing Government*. Addison Wesley, Reading, MA
Osborne D, Plastrick R 1997 *Banishing Bureaucracy: The Five Strategies for Reinventing Government*. Addision-Wesley, Reading, MA
Poister T H, Streib G 1999 Strategic management in the public sector: Concepts, models, and processes. *Public Management and Productivity Review* **22**(3): 308–25
Porter M 1998a *Competitive Advantage: Creating and Sustaining Superior Performance*. Free Press, New York
Porter M 1998b *The Competitive Advantage of Nations: With New Introduction*. Free Press, New York
Quinn J B 1980 *Strategies for Change: Logical Incrementalism*. R D Irwin, Homewood, Ill
Roberts N, Wargo L 1994 The dilemma of planning in large-scale public organizations: The case of the United States Navy. *Journal of Public Administration Research and Theory* **4**: 469–91
Salets W, Faludi A (eds.) 2000 *The Revival of Spatial Strategic Planning*. Royal Netherlands Academy of Arts and Sciences, Amsterdam
Schein E 1992 *Organizational Culture and Leadership*, 2nd edn. Jossey-Bass, San Francisco, CA
Senge P M 1990 *The Fifth Discipline: The Art and Practice of the Learning Organization*. Doubleday, New York
Stone M M 2000 Exploring the effects of collaborations on member organizations: Washington county's welfare-to-work partnership. *Nonprofit and Voluntary Sector Quarterly* **29**(1) Supplement: 98–119
Susskind L E, Cruikshank J 1987 *Breaking the Impasse: Consensual Approaches to Resolving Public Disputes*. Basic Books, New York

J. M. Bryson

Strategy: Organizational

The subject of strategy in organizations has become one of the dominant features of the contemporary world of academics and practitioners. The use of the word is pervasive. The genesis and development of the concept has been relatively swift yet has produced a rich variety of interpretations and uses. Its core meaning was taken by Chandler, to center on a logical sequence of goal-setting and resource allocation within firms. In his widely-adopted formula (Chandler 1962, p. 13): 'Strategy is the determination of the basic long-term goals of an enterprise, and the adoption of courses of action and the allocation of resources necessary for carrying out these goals.' Subsequent work has emphasized the competitive element so that the purpose of a strategy is to maintain or improve performance. Strategies are judged by their ability to neutralize threats and exploit opportunities for the organization while capitalizing on strengths and avoiding weaknesses (Barney 1997, p. 27). Strategy also refers to the direction of an organization, which matches its resources to its environment in order to meet stakeholders expectations (Johnson and Scholes 1993, p. 10). Strategic management is concerned with the process by which strategies are adopted and put into practice.

1. Origins

The practical origin of the term strategy is found in military history. In the ancient world *strategos* was applied to the generalship of an army commander, 'the art of the general.' By 330 BC it also included notions of administration and the use of armed force in order not only to defeat enemies but also establish systems of governance (Quinn et al. 1988, p. 1). The word strategy was first used in English in the seventeenth century. By 1870, military dictionaries distinguished between strategy (planning conducted away from the enemy) and tactics (the immediate engagement of an adversary) (Whipp 1996, p. 13).

The idea of organizational strategy in a business or commercial sense was a product of the second half of the twentieth century. Yet its emergence as a separate subject of teaching and research in the 1960s owed much to its military roots. The pioneering book of Ansoff (1965) on *Corporate Strategy*, for example, distinguished between corporate strategy and business strategy, paralleling the military ideas of grand strategy and battle strategy. Corporate strategy was seen as selecting what business to be in; business strategy dealt with how to operate in a given business area. A second landmark text (Andrews 1971) also relied on these military distinctions. He emphasized the aim of strategy as establishing a clear dominance in a business domain while simultaneously planning for the movements of hostile adversaries (see also Quinn et al. 1988, p. 42).

The second main influence on the development of the subject of strategy was the discipline of economics. In the early decades of the twentieth century, research on industry form, competition and innovation had pointed to the role of individual firms in economic

relations. Such writers were firmly set in the traditions of microeconomics. The approach was a rational one with a reliance on profit maximizing behavior by all economic agents, accessible information and the mobility of resources. Andrews' view of strategy was therefore clear: 'rivalry amongst peers for prizes in a defined and shared game' (Pettigrew and Whipp 1991, p. 29).

The third force to shape the emergence of strategy as a subject was the geographical and corporate context. The flourishing of large, multi-product corporations, such as General Motors and Du Pont, took place in the 1920s and 1930s in the USA. As they grew commercially and organizationally more complex, so did their need for more sophisticated means of assessing markets and competitors and matching them with the resources of the firm. The early experts on strategy worked closely with these corporations in response to their needs. Ansoff was an executive at Lockheed Aerospace while Chandler was related to the Du Pont family. As the demand for technical experts to produce strategic analysis rose, so business schools and consulting firms grew in number and importance during the 1960s and 1970s. Outside of the US, corporate forms were much slower to evolve and in Europe the large-scale growth in business schools and the study of strategy did not occur until 20 years later.

2. *Uneven Development*

Given the late advent of strategy as a subject, its development since the 1960s has often been more colorful than more mature areas of the social sciences. New approaches have arisen, offering challenges to the main orthodoxies. Insights and techniques from outside of economics have been applied to enrich and extend the subject. As the study of strategy has progressed, so its full complexity has become apparent. The growth of the area is best understood by reference to three waves of activity: the classical era of the 1960s, the challenges set by organizational analysts in the 1970s and early 80s, and the production of a more composite approach to meet the circumstances of the 1980s and early 1990s.

During the 1960s, strategy experts placed a heavy reliance on planning techniques. Financial planning tools were created to link capital investments with production and to derive profit forecasts. Forecast-based planning was constructed in the light of unstable market behavior. Business portfolio matrices and product lifecycle tools were examples of devices produced to make forecasts (Bourgeois 1996, p. 5). Early computing technology raised the expectations of strategic planners regarding the potential of this planning-led approach. The conviction that the environment of the firm was eminently capable of being captured by such techniques was strong.

In the following decades, this planning-led approach to strategy continued as the mainstream, particularly in North America. Alternative voices appeared on both sides of the Atlantic with their attention directed not to planning strategy but the process of implementation. The process perspective was built upon decision-making theory (e.g., see *Organizational Decision Making*). First Lindblom (1959), and later March and Olsen (1976), demonstrated how the decisions of senior management were often less than rational, notwithstanding their use of planning apparatus. Trade-offs, compromises, and coalition formation were shown to be routine. These writers drew attention to the haphazard and highly subjective qualities of strategic decisions and their mode of implementation. Writers, such as Mintzberg (1978), revised the profile of the strategic decision making, and implementation activities in an organization, suggesting that these constituted important change processes in their own right and should be analyzed and managed accordingly. Concepts of power and culture were applied to the problem for the first time (e.g., *Organizations: Authority and Power*). The result was a more realistic appreciation of the problems strategic management involved, including the issues of political influences, aborted ideas and unintended outcomes.

Whilst this process-based approach has continued as a separate tradition in the study of strategy (often labelled strategic change), it also had a major influence on the third wave of activity in the 1980s. The driving force behind these new developments was a dramatic shift among international competitors. Japanese firms in the manufacturing and consumer goods sectors enjoyed spectacular gains in overseas markets, notably the USA and Western Europe. Their market success was based on the ability to combine product innovations with new standards in quality, underpinned by (in the eyes of western managers) novel forms of organization. These events provoked two responses from academics.

First, was a renewed interest in competitive analysis, drawing on the microeconomic perspective (outlined in Sect. 1). Porter played the leading role by highlighting 'generic strategies' (1980)—such as cost leadership vs. differentiation—and the need for organizations to clarify their choice of strategy. He went on to offer a model of extended competition (Porter 1985) which forced firms to treat seriously new sources of competition from outside a sector (e.g., see *Competitive Strategies: Organizational*). This fresh emphasis on competitor analysis was augmented by a rejuvenation during the 1980s of the resource-based theory of the firm. Here particular attention was paid to the identification and maintenance of the unique collection of capabilities (technological, organizational, financial, for example) found in successful firms but which others would find difficult to imitate. A preoccupation grew with 'strategic intent': the

ability to create extremely long-term aspirations for an organization and link them to immediate operations (Hamel and Prahalad 1989). The insights of the strategic change analysts (e.g., see *Organizational Change*) became especially relevant to this new conception of strategy and competition (Pettigrew and Whipp 1991).

3. Current Themes

The emphasis in current strategy research, in part, arises from ideas and incomplete projects, which began in earlier eras; this is particularly true for the interest in strategic capabilities and transformation. In the empirical sphere the dominant characteristics of strategic studies has been their entry into hitherto unresearched settings. At the same time, it has been noticeable how other specialisms in the business studies area, such as human resource management have attempted to incorporate a strategic perspective (e.g., see *Strategic Human Resources Management*).

The harsh competitive conditions of the 1980s and 1990s, produced a general interest amongst business people in organizations, which could prosper in such circumstances. Companies such as Honda, Chrysler or ABB received intense scrutiny from practitioners and academics alike because of their superior strategic management. The result was that academics became exceptionally concerned with explaining their success. To this end, research has been targeted at the unusual ways in which such firms analyze their relationship with their environments, create ways of handling those relationships and do so in a consistent pattern of business behavior (Kay 1993, p. 9). The core of their success is attributed to their delivery of innovations in products or services. Research has shown that producing these innovations is explained by the existence of a set of supporting capabilities. In the case of Honda, its unique record of product innovation worldwide rests on a collection of distinctive capabilities involving: problem solving, project management and working with suppliers (Pascale 1990).

The circumstances, which prompted these exceptional feats of strategic management, have, in turn, become the subject of academic attention. The exploitation of information technology has transformed entire industries. New forms of communication have created totally novel means of acquiring information and purchasing goods or services. Information technology has also led to completely new types of products. The character of sectors, such as retail banking and insurance, has been transformed. The ability of organizations to manage the wholesale reworking of their strategic assumptions and operational methods has become a source of fascination to practitioners and academics alike. Whereas in the 1980s, the emphasis had been on the adjustment of company strategies to new rules of competition, today, the contemporary agenda is dominated by the prospect of companies facing the growth of new products, markets and sectors on a scale not seen since the twentieth century in Europe and the early twentieth century in the USA.

In parallel to these high profile projects on strategic capabilities and transformations, strategy scholars have moved away from their preoccupation with corporations. The result has been an engagement with strategic management within small businesses, voluntary and not-for-profit sectors and professional organizations. These alternative contexts have provided useful tests for established theories. Sometimes the characteristics of these under-researched sites have forced analysts and managers in the corporate sector to review their models. The ability of voluntary organizations (e.g., see *Voluntary Organizations*) to succeed with few resources and yet motivate staff is a case in point (Johnson and Scholes 1993, pp. 23–30).

4. Methods

The strategy field has witnessed the use of an extensive battery of research methods during its comparatively brief existence to-date. In the early phase, the quantitative methods associated with economics became well established. These techniques have continued to play a major role, especially in the measurement of performance and competition. From the 1970s, as the insights from psychology, politics and sociology were used to open up the strategic management process (see Sect. 2), so more qualitative approaches (including ethnography in the study of organizational culture, for example) were employed. The qualitative and quantitative currently co-exist productively (Kaplan and Norton 1996). More contentious is the common reference by strategy writers to individual company examples. Outsiders may confuse teaching cases (usually of a single organization) with full-scale research-based case studies. Nevertheless, popular accounts of corporate success stories and the personal reflections of leading executives which are light on formal research methodologies, have created an impression of superficiality which academics often have to live down.

5. Future Research

Given the youthful nature of the subject, the study of strategy in the future will continue to unravel the problems associated with understanding competitive environments, how they may be matched to the resources of an organization and how the resulting decisions can be implemented. Fresh opportunities for research will appear in the international sphere. This will include not only the continued relevance of global markets and organizations but also the further exploration of national contexts. The differences of the financial and business systems between the USA,

Europe and Japan are well known, detailed comparisons with other parts of Asia or South America await. In addition, the fragmentation of social and economic relations—around, for example, network structures or self-organized commercial ventures—will pose a challenge to a subject which was founded on the problems of large corporations. Other tests will be supplied by links between strategy writers and specialists from other subject areas. Examples include, complexity theory and knowledge management experts from organization studies (e.g., see *Learning: Organizational*). There are also signs that those who are skeptical of the precepts of strategic management will increase their attention. Their assertions that strategy is merely
a specialized form of language which enhances the power in society of management is the opening salvo in what could be a long exchange of intellectual fire power. The response by the strategy experts will demonstrate how far the subject has come of age into its mature form.

See also: Management: General

Bibliography

Andrews K R 1971 *The Concept of Corporate Strategy*. Irwin, Homewood, IL

Ansoff H I 1965 *Corporate Strategy*. McGraw-Hill, New York

Barney J B 1997 *Gaining and Sustaining Competitive Advantage*. Addison-Wesley, Reading, MA

Bourgeois L J 1996 *Strategic Management from Concept to Implementation*. Dryden, Fort Worth, TX

Chandler A 1962 *Strategy and Structure*. MIT Press, Boston, MA

Hamel C, Prahalad C K 1989 Strategic intent. *Harvard Business Review*. **May-June**: 63–76

Johnson G, Scholes K 1993 *Exploring Corporate Strategy*, 3rd edn. Prentice-Hall, Hemel Hempstead

Kaplain R S, Norton D P 1996 *The Balanced Scorecard*. Harvard Business School Press, Boston

Kay J 1993 *Foundations of Corporate Success*. Oxford University Press, Oxford, UK

Lindblom L 1959 The science of muddling through. *Public Administration Review* **19**: 78–88

March J, Olsen J 1976 *Ambiguity and Choice in Organizations*. Universitetsforlaget, Bergen, Norway

Mintzberg H 1978 Patterns in strategy formation. *Management Science* **24**: 934–48

Pascale R T 1990 *Managing on the Edge*. Viking, London

Pettigrew A, Whipp R 1991 *Managing Change for Competitive Success*. Blackwell Oxford, UK

Porter M 1980 *Competitive Strategy*. Free Press, New York

Porter M 1985 *Competitive Advantage: Creating and Sustaining Superior Performance*. Free Press, New York

Quinn J B, Mintzberg H, James R M 1988 *The Strategy Process*. Prentice Hall, Englewood Cliffs, NJ

Whipp R 1999 Strategy and organisations. In: Clegg S, Hardy C, Nord W R (eds.) *Handbook of Organization Studies*. Sage, London

R. Whipp

Street Children: Cultural Concerns

Since the early 1980s street children have become a focus for attention in the media, an embarrassment to many bureaucrats, and a matter of priority for international agencies and numerous non-governmental organizations. At the same time, challenging academic and practical questions have been raised regarding who 'street children' are and the best strategies for helping them.

1. Definitions and Problematic Categorizations

The term 'street children' came into widespread use following the United Nations *Year of the Child* (1979) when it was adopted by international agencies and promptly translated into many different languages (as *meninos de rua* in Portuguese, *niños de la calle* in Spanish, *enfants de la rue* in French). It arose from attempts to develop a term free of negative connotations about children popularly known as street urchins, vagrants, gaminos, rag-pickers, or glue-sniffers, and historically as street Arabs or vagrants. Utilized in 1951 by UNESCO to refer to vagrant children following World War II, it was applied in 1986 to the developing world by UNICEF, the agency primarily concerned with children's wellbeing, to designate minors who worked or slept on the streets (Williams 1993).

At a most basic level, a street child is 'a homeless or neglected child who lives chiefly in the streets' (*Oxford Dictionary*). Welfare agencies, however, have reworked their definitions many times, encountering enormous problems in devising meaningful statements about street children and making appropriate distinctions between categories of children.

Problems ensued even after the consensual definition formulated in 1983 by the Inter-NGO Programme for Street Children and Street Youth: 'Street children are those for whom the street (in the widest sense of the word that is unoccupied dwellings, wasteland, etc …) more than their family has become their real home, a situation in which there is no protection, supervision or direction from responsible adults' (see Ennew 1994, p. 15). Several terms could lead to confusion: what is meant by 'family,' 'protection,' or 'responsible' adults (Hecht 1998)? Even what is meant by 'home' is conceptualized differently across cultures: being homeless is rendered as *desamparado* (without protection or comfort of other people) in Latin America, *furosha* (floating) in Japan, and *khate* (rag-picker) in Nepal, evoking concepts of disaffiliation, transience, and marginal work rather than notions of residential access or type of abode (Desjarlais 1996).

Several attempts to categorize street children, developed to acknowledge their heterogeneous circumstances and lifestyle, have also proved problematic. In

one typology following UNICEF, 'children *of* the street' have a family accessible to them but have made the streets their home; 'children *on* the street' are child workers who return at night to their families; 'abandoned' children literally have no home to go to, while children 'at high risk' are those likely to be drawn into street-life. Distinctions have also been made between 'abandoned' and 'abandoning' street children (Felsman 1984), and among children according to their degree of family involvement and psychological characteristics. These terminologies have proved difficult to uphold in practice, given the fluidity of children's movements on and off the streets and a lack of correspondance with the ways children themselves relate their experiences. Many labels such as children 'without family contact' and 'abandoned' lacked precision and were used ad hoc rather than analytically. The focus on categories of children, which led to the popular distinction between children *of* and *on* the streets (translated as *de la calle/en la calle* in Spanish, or *de la rue/dans la rue* in French) is falling into disuse.

The term 'street children' is still widely used to highlight a set of working and living conditions rather than to refer to the personal or social characteristics of the children in question. But while street children are identified as those who occupy the public spaces of urban centres and whose activities are largely unsupervised by adults, it has proved important to move beyond definitions based solely on physical, social, and economic criteria.

2. *Statistics and the Creation of a Social Issue*

A great deal of controversy also surrounds the process of enumerating street children. Different agencies issue often widely different estimates, partly because they find street children difficult to count, partly because they are not always talking about the same children. In Brazil alone, the figure of 7 million street children is most often cited, typically with reference to the homeless, although careful surveys suggest a national estimate less than 1 percent of that figure (39,000 in Hecht 1998). The 1991 video 'About the UN Rights of the Child' stated that an estimated 200 million children around the world live on the streets, while another global figure of 100 million street children has also gained credibility. But how are the children counted and who exactly is included? Huge numbers, produced to draw attention to children's plight, rest upon largely elastic and nebulous definitions of homeless and working children; they are usually the result of guesswork and have 'no validity or basis in fact' (Ennew 1994, p. 32).

The same problems of enumeration and group definition beset studies of the homeless in the developed world. The definitions of 'homeless' may shift to include those who sleep rough, live in shelters, squat, or double up with other families, encompassing the literally homeless and the precariously housed (Glasser and Bridgman 1999). Estimates of homeless youth—whether 'runaways,' in care, or with a parent—are often inflated by welfare agencies to legitimate their role, yet minimized by bureaucratic institutions with legal or financial responsibilities; thus statistics reflect the various agendas of organizations which collect them (Hutson and Liddiard 1994). The process of identifying and counting street youths is problematic in itself and it is also a part of the construction and management of homelessness as a social issue.

Notwithstanding their imprecision, estimates can help to identify significant groups and appropriate interventions in different cultural contexts. It is important to know that the great majority of street children in developing countries are not 'homeless' but return at night to their families (they are children *on* the streets). The fact that they are predominantly boys has alerted attention on why local circumstances should differ according to gender. Care is taken to reveal the age distribution between teenage adolescents, those in late childhood, and very young children with families, who may experience different reactions from the public, as well as ethnic distribution. Estimates also help to appreciate changing trends. In the US, the fastest growing subgroup among the homeless are presently young children.

3. *Representations of Street Children*

Several different explanations have been offered to explain the phenomenon of homelessness and the dramatic rise in the numbers of street children. There are, broadly speaking, two major schools of thought: one links homelessness to personal pathology, the other to external structural factors (Glasser and Bridgman 1999). The former focuses on problems such as family dysfunction and substance abuse or mental illness, which increase an individual's risk of taking to the streets. The latter emphasizes broad economic and sociopolitical factors such as poverty and marginalization under conditions of rapid urbanization, rural-to-urban migration, or conflict, which affect the stability of households. Whether linked to biographic or structural factors, street children are often portrayed as victims of deprived environments and helpless with respect to changing circumstances, or else as social deviants (Veale et al. in Panter-Brick and Smith 2000).

The 'characteristics' of street-children—their tenuous links with families and independence from adults—do not conform to commonly held views about what constitutes a 'proper' childhood in contemporary thought. Popular images of street-children—as victims or villains—rest upon particular representations of children and twentieth century discourses about child-

hood that arose among the wealthier classes of the western industrial world (Boyden 1990). The apposition of 'street' and 'children' suggests that street location is a peculiar and significant mark of identity outside the normal frame: 'after all, children can also be found using fields, lofts, and gardens without there being any apparent need to coin terms such as 'field children,' 'loft children,' or 'garden children' (Glauser 1990). Street children upset broad expectations both about the street—a place for commerce, men, and prostitutes—and about childhood—a time for carefree and protected innocence (Hecht 1998). From the modern Western viewpoint, children who are not 'at home' and nurtured by responsible adults are 'forsaken' or 'deviant,' and street children in particular are 'out of place' in society and 'outside childhood' (Connolly and Ennew 1996).

It has been argued that such portrayals are actually unhelpful to the children themselves (Ennew 1994). Many interventions which emphasise a vulnerable and 'lost' childhood have focused on 'rescuing' children from the streets by placing them in institutions or back with a family; this tends to ignore the social networks and coping strategies developed by the children, and generally has failed to provide lasting solutions. The focus of children's activities 'on the street' has also tended to promote unidimensional accounts of their lives; for instance, it is not often realized that children who live on the street commonly attend school and retain contacts with their families (Ennew and Milne 1997). Furthermore, public condemnation of street murders in Latin America has unwittingly reduced children's status to that of hunted prey, which conveys little understanding of the ways in which violence is actually experienced: in Brazil, while harassment at the hands of police is ubiquitous and extermination or torture not uncommon, street children are far more likely to die at the hand of their peers than as victims of death squads (Hecht 1998).

4. *Child-centered Perspectives and Research Issues*

In the late 1990s fundamental assumptions about children and appropriate methods of research began to be re-evaluated, in the context of international legislation of children's 'best interests' (see *Children, Rights of: Cultural Concerns*) and the development of participatory methodology (Ennew and Milne 1997). The realization was made in social science research that children are not helpless or passive, but capable of making informed decisions about the future, and that they express views and aspirations that may well differ from those of adults. Child-centered perspectives and participatory research methods have emphasized the need to recognize the agency and social competence of street children and to enhance their participation. The key to research and project planning has changed to working 'with' children rather than 'for' them: 'they are not 'objects of concern' but people. They are vulnerable, but not incapable. They need respect, not pity' (Ennew 1994, p. 35). Remedial action in the best interests of children—as stressed by the United Nations Convention of the Rights of the Child—is not a matter of rescuing children, but of listening to them, and of fostering child participation.

A range of participatory approaches have been attempted, such as having street youths to conduct actual research, encouraging children to represent their world through drawings or plays, and staging workshops or group discussions on topics selected by the children themselves. These methods, more appropriate than the ubiquitous questionnaire survey, aim to involve the children in the research process rather than simply gather information or anecdotal testimonies. They have the potential to generate more valid data on children's lives but raise their own set of ethical questions.

Two significant research problems have been the lack of comparative information on peer groups within the same country and a failure to evaluate longitudinally the 'career' outcomes of street children (see Baker and Panter-Brick in Panter-Brick and Smith 2000). The tendency has been to focus on snap shot descriptions of street children with little time-depth or contextual information. Street children live complex and varied lifestyles—they may move home and back to the street, go to school, or work at times for an employer. Few street children remain on the street into adulthood, although the range of outcomes is enormously variable and shaped by cultural context. While prison, insanity, or death are the common expectations of Brazilian street children, stable employment, marriage, and children are achievable career paths for Nepali street children.

Comparative and longitudinal research is critical, but not easily achieved. For instance, it is easy to contend that homelessness endangers children's physical and emotional health, and indeed most studies of the homeless in the West and of street children in developing countries have emphasized the debilitating and deprived aspects of street-life. Recently, however, a number of comparative studies have questioned the expectation that homeless children should be the most vulnerable in deprived environments, arguing that poverty, not homelessness *per se*, may carry the most significant risks for health and behavior outcomes (Ziesemer et al. 1994, Bassuk et al. 1997, Veale et al. in Panter-Brick and Smith 2000). It is also recognized that the public discourse on street children has detracted from the issue of widespread poverty and violence affecting children who live at home.

The issues of violence, perpetrated by both adults and other street children, of substance abuse, of sexual exploitation, and sexually transmitted diseases, are often raised but among those most difficult to research.

The nature of appropriate interventions designed to enhance children's own coping strategies, such as social networks with street peers and resourcefulness in obtaining money, have also been discussed extensively. A compilation of studies across different continents and topics of research are found in Mermet (1997), Ennew and Milne (1997), and dedicated web sites.

5. *Conclusion*

Studies of street children and research on homelessness were first concerned with describing lifestyle situations in terms of the uses of public spaces and the links with family or institutions, hinging various categories of street life upon the constructs of home, family, and a proper childhood. Twenty years on, such research is concerned with identifying the subjective and cultural interpretations of homelessness (Desjarlais 1996), with considering explicitly the context of poverty and social exclusion (Mingione 1999), and with seeking to follow career paths and identify what transitions are possible from one situation to another. At the international level, advocacy for children is less focused on the 'street' as it is on abusive work situations, sexual exploitation, or criminality, and with promoting responses that are consistent with children's rights (Ennew 1995, Bartlett et al. 1999).

See also: Child Abuse; Childhood Sexual Abuse and Risk for Adult Psychopathology; Children and the Law; Children, Rights of: Cultural Concerns; Children, Value of; Early Childhood: Socioemotional Risks; Infant and Child Mortality in the Less Developed World; Poverty and Child Development; Street Children: Psychological Perspectives; War, Political Violence and their Psychological Effects on Children: Cultural Concerns

Bibliography

Bartlett S, Hart R et al. 1999 *Cities for Children: Children's Rights, Poverty and Urban Management*. Earthscan, London

Bassuk E L, Weinreb L F et al. 1997 Determinants of behavior in homeless and low-income housed preschool children. *Pediatrics* **100**: 92–100

Boyden J 1990 Childhood and the policy makers: A comparative perspective on the globalization of childhood. In: James A, Prout A (eds.) *Constructing and Reconstucting Childhood: Contemporary Issues in the Sociological Study of Childhood*. Falmer Press, London, pp. 184–215

Connolly M, Ennew J 1996 Introduction—Children out of place. *Childhood—A Global Journal of Child Research* **3**(2): 131–45

Desjarlais R 1996 Some causes and cultures of homelessness. *American Anthropologist* **98**(2): 420–5

Ennew J 1994 *Street and Working Children: A Guide to Planning*. Save the Children, London

Ennew J 1995 Outside childhood: Street children's rights. In: Franklin B (ed.) *The Handbook of Children's Rights: Comparative Policy and Practice*. Routledge, London and New York

Ennew J, Milne B 1997 *Methods of Research with Street and Working Children: An Annotated Bibliography*. Radda Barnen, Stockholm and Weden

Felsman J K 1984 Abandoned children: A reconsideration. *Children Today* **13**: 13–8

Glasser I, Bridgman R 1999 *Braving the Street: The Anthropology of Homelessness*. Berghahn, Oxford, UK

Glauser B 1990 Street children: Deconstructing a construct. In: James A, Prout A (eds.) *Constructing and Reconstructing Childhood: Contemporary Issues in the Sociological Study of Childhood*. Falmer Press, London, pp. 138–56

Hecht T 1998 *At Home in the Street: Street Children of Northeast Brazil*. Cambridge University Press, Cambridge, UK

Hutson S, Liddiard M 1994 *Youth Homelessness—The Construction of a Social Issue*. Macmillan, London

Mermet J 1997 *Bibliography on Street Children*. Henry Durant Institute, Geneva

Mingione E 1999 The excluded and the homeless: The social construction of the fight against poverty in Europe. In: Mingione E (ed.) *Urban Poverty and the Underclass: A Reader*. Blackwell, Oxford, UK, pp. 83–104

Panter-Brick C, Smith M 2000 *Abandoned Children*. Cambridge University Press, Cambridge, UK

Williams C 1993 Who are 'street children?' A hierarchy of street use and appropriate responses. *Child Abuse & Neglect* **17**: 831–41

Ziesemer C, Marcoux L et al. 1994 Homeless children: Are they different from other low-income children? *Social Work* **39**(6): 658–68

C. Panter-Brick

Street Children: Psychological Perspectives

The United Nations defined street children as 'any boy or girl … for whom the street (in the widest sense of the word, including unoccupied dwellings, wasteland, etc.) has become his or her habitual abode and/or source of livelihood; and who is inadequately protected, supervised, or directed by responsible adults' (quoted in Lusk 1992 p. 294). The definition of street children plays a pivotal role in research and may be a source of disagreement about the results of studies (Koller and Hutz 1996). Children and adolescents who look like drifters (wear shabby, dirty clothing, beg for food or money, sell small objects, work, or wander without a purpose on the streets) can be found in large cities all over the world. The appearance of abandonment singles them out as belonging to the same group. However, their life histories, family character-

istics, street life experiences, and prognoses are very different (Dallape 1996). Some researchers define street children based on characteristics such as sleeping location, family ties, school attendance, leisure, survival activities, occupation on the street environment, etc. Such definitions can lead to broad categorizations as: *children of the streets or children in the streets* (e.g., Barker and Knaul 1991, Campos et al. 1994, Forster et al. 1992).

Children *of* the streets would be those who actually live on the street, all day and at night, who do not attend school, and do not have stable family ties. They fulfill their needs and are socialized on the streets. In contrast, children *in* the streets would be those who live with their families, may attend school, but spend all or part of their days on the streets, trying to earn money for themselves or their families (Hutz and Koller 1999). The relationship with the family has been considered as a key feature of the definition of street children. Felsman (1985) identified three groups of street children in Colombia: (a) orphaned or abandoned children, (b) runaways, and (c) children with family ties. Leisure and occupation on the streets were added to family ties in Martins' work (1996) to identify three different groups of street children in Brazil. He found a group of children with stable family ties who worked on the streets and went home every night. These children played in their neighborhood or on the streets where they worked and many attended school. A second group had unstable family ties. Although they lived on the streets, these children knew their families and, occasionally, went home to visit or even to stay for a while. Finally, there was a group of children who were on their own on the streets and who had lost all contact with their family.

Nevertheless, it is difficult and it may even be misleading, to define a child as belonging to a specific category. Hutz and Koller (1999) claimed that in their research they rarely found children who had completely lost contact with their family. They also identified many children who lived at home and worked on the streets, but occasionally slept on the street, and children who periodically left home and lived on the streets for weeks or months, and then went back home. The variability within these groups regarding the frequency of family contact, sleeping location, occupation on the streets, the destination for the money they earn, school attendance, and several other variables (including physical and sexual abuse, sexual activity, etc.) may be so large that the distinction between the *of the street group* and the *in the street group* may be meaningless or even misleading for research or intervention purposes. These authors suggested that it would be more appropriate to categorize street children as a function of the risks to which they are exposed (e.g., contact with gangs, use of drugs, dropping out of school, lack of proper parental guidance, prostitution, etc.) and the protective factors available to them (e.g., school attendance, supportive social networks, contact with caring adults, etc.). Researchers could then determine how vulnerable children are to developmental risk and what appropriate actions could be taken in each specific case.

1. *Developmental Implications*

Children living on the streets are still children undergoing development, despite their life conditions. They experience risks and challenges that, at the same time, may jeopardize their development and promote the acquisition of strategies for dealing with life on the streets. There is some evidence that economical pressures and emotional disturbances in the family expose children to larger risks than do the conditions of the street (Hecht 1998, Hutz and Koller 1997, Matchinda 1999). Street children often face larger risks than children in general because they are exposed to negative physical, social, and emotional factors at home and still have to deal with the challenges of life on the streets. On the other hand, there is evidence that the conditions of life on the streets lead to the development of coping strategies that are adaptive and that may help to strengthen their cognitive and social skills.

2. *Social Development*

Street children are usually targets of social rejection and discrimination. They have to develop their social identity and sense of belonging to a society that views them either as victims, who deserve pitying, or as criminals who must be taken off the streets and locked in jail. On the one hand, they are seen as victims, because they do not have shelter, clothes, food, or adult protection, have to work on the streets instead of going to school, are sexually exploited, and so on. On the other hand, they are perceived also as transgressors because they often use drugs, commit robbery, make noise, and are grouped in threatening gangs. The adult environment is usually very hostile to street children. The police aggress them often, causing physical harm and humiliation. They are also harassed by street adults and gangs, which fight for space and better survival conditions.

An effective strategy to survive and develop in such a hostile environment is to belong to a group on the streets. Therefore, street children will often join gangs and develop different kinds of peer relationships that leads to the development of *emotional groups* (appropriate to spend the night and to have fun together) and *business groups* (organized to dodge street life risks and fulfill their survival needs). Another strategy consists of going to social institutions for food and shelter. However, such institutions often fail to help

them effectively, because they aim at taking the children off the streets, whereas the children seek them, because they perceive the institutions as part of street life, and not as a way out of it (Hecht 1998).

3. Emotional Development

Aptekar (1989, 1996) stated that street children are mentally healthier than their siblings who stayed at home. Koller and Hutz (1996) observed that these children have the ability to reorganize their lives on the streets, in spite of their risky conditions and their life histories. Most of them left home because their parents failed to provide a safe, nurturing, and affective environment. Many children also report sexual or physical abuse, drug use at home, and economic exploitation as reasons for leaving home (Raffaelli et al. 1995).

Some evidence presented by DeSouza and collaborators (DeSouza et al. 1995) indicated that street children were not in greater psychological distress than children of a low social economic status who lived with their families. Koller et al. (1996) investigated subjective wellbeing of street children and of children who lived in deep poverty and their findings also did not show significant differences between these groups.

4. Physical Development

Life on the streets represents a constant source of risk to children and adolescents. Their safety and survival demand energy and coping strategies to confront the daily risks. Conflicts between gangs, police harassment, and adult street dwellers physical abuse are some examples of daily violence that street children have to deal with successfully to stay safe. Cold weather, lack of food and shelter, traffic accidents, untreated injuries and illnesses, exposure to drugs and unprotected sexual activities are also important risks to their health and physical integrity (Donald and Swart-Kruger 1994, Hecht 1998).

Street children must develop adaptive strategies to survive and stay safe in spite of those risks. As previously mentioned, they form groups (emotional and business groups) that protect them from street violence and help them to survive. Also, often they find shelter and food in institutions that have rules that must be obeyed. They learn to cope with such rules, even when they do not agree with them, but their behavior becomes opportunistic and often ingenious (Donald and Swart-Kruger 1994).

5. Cognitive Development

Most street children, even those who go to school, are illiterate and have negative school experiences. Their attention span, memory, and cognitive development in general may be affected by malnutrition, drug use and intoxication, untreated illnesses, and accidental injuries. Often, they have difficulty adapting to the formal school system because it requires discipline, attention to specific tasks and schedules, planning ahead, and other routines with which they cannot deal effectively. Language, critical thinking, and intelligence also develop more slowly and may present significant deficits because street children interact mostly with their peers and have very little contact with adults. In fact, some researchers have noted that street children find it rewarding to talk to adults who will listen and speak with them in a friendly manner (Hutz and Koller 1999). Middle-class children, as a rule, are exposed daily to caring adults who talk to them, tell them stories, listen to their tales, and spend time interacting with them. Street children do not have this experience.

Spatial skills and very well developed visual and auditory discrimination, for example, are required to detect, avoid, and escape street risks. Carraher et al. (1985) noted that street children who worked at the market were very capable of dealing with money and doing sophisticated calculations to figure out the price or value of products (although they failed when presented with standard school math problems). Aptekar (1989) also referred to what he calls *street knowledge* as an important skill for social interaction.

6. Conclusion

Children living on the streets are a social problem in many countries, a problem that has to be fought by all means available. To fight this social ill requires that individuals and groups in society take the social and political responsibility to develop effective prevention and intervention projects. Children on the streets are vulnerable to risks but they manage to develop 'coping strategies' that often make them resilient (Hutz et al. 1996). They behave as children when they play or interact with peers on the streets. But, they must also act as adults when they have to provide for their subsistence and safety. In spite of their circumstances, street children are still developing persons that require appropriate health care, education, a nurturing home, safety, and human rights in order to grow with dignity and to become adjusted and productive citizens.

See also: Childhood: Anthropological Aspects; Children, Rights of: Cultural Concerns; Infant and Child Mortality: Central and Eastern Europe; Infant and Child Mortality in Industrialized Countries; Infant and Child Mortality in the Less Developed World; Mental Health Programs: Children and Adolescents; Poverty and Child Development; Socialization in Infancy and Childhood; Street Children: Cultural Concerns; Violence and Effects on Children; War,

Political Violence and their Psychological Effects on Children: Cultural Concerns

Bibliography

Aptekar L 1989 Colombian street children: Gamines or chupagruesos. *Adolescence* **24**: 783–94

Aptekar L 1996 Crianças de rua nos países em desenvolvimento: Uma revisão de suas condições [Street children of developing countries: A review of their conditions]. *Psicologia Reflexão e Crítica* **9**: 153–85

Barker G, Knaul F 1991 *Exploited Entrepreuners: Street and Working Children in Developing Countries*. Working Paper 1, Childhope, New York

Campos R, Raffaelli M, Ude W, Greco M, Ruff A, Rolf J, Antunes C M, Halsey N, Greco D 1994 Social networks and daily activities of street youth in Belo Horizonte, Brazil. *Child Development* **65**: 319–30

Carraher T N, Carraher D, Schliemann A 1985 Mathematics in the streets and in the schools. *British Journal of Developmental Psychology* **3**: 21–9

Dallape F 1996 Urban children: A challenge and an opportunity. *Childhood* **3**: 283–94

DeSouza E, Koller S H, Hutz C S, Forster L M 1995 Preventing depression among Brazilian street children. *InterAmerican Journal of Psychology* **29**: 261–5

Donald D, Swart-Kruger J 1994 The South-African street child: Developmental implications. *South-African Journal of Psychology* **24**: 169–74

Felsman J K 1985 Abandoned children reconsidered: Prevention, social policy, and the trouble with sympathy. *ERIC Document* ED268457

Forster L M, Barros H T, Tannhauser S L, Tannhauser M 1992 Meninos de rua: Relação entre abuso de drogas e atividades ilícitas [Street children: The relationship between drug use and illicit activities]. *Revista ABP-APAL* **14**: 115–20

Hecht T 1998 *At Home in the Street*. Cambridge University Press, Cambridge, UK

Hutz C S, Koller S H 1997 Questões sobre o desenvolvimento de crianças em situação de rua [Issues regarding the development of street children]. *Estudos de Psicologia* **2**: 175–97

Hutz C S Koller S H 1999 Methodological issues in the study of street children. In: Raffaelli M, Larson R (Volume eds.) Damon W (Series ed.) *Homeless and Working Street Youth Around the World: Exploring Developmental Issues. New Directions in Child Development*. Jossey Bass, San Francisco, Vol 85, pp. 59–70

Hutz C S, Koller S H, Bandeira D R 1996 Resiliência e vulnerabilidade em crianças em situação de risco [Resilience and vulnerability in children at risk]. *Coletâneas da ANPEPP* **1**(12): 79–86

Koller S H, Hutz C S 1996 Meninos e meninas em situação de rua: Dinâmica, diversidade e definição [Boys and girls on the streets: Dynamics, diversity, and definition]. *Coletâneas da ANPEPP* **1**(12): 11–34

Koller S H, Hutz C S, Silva M 1996 June. Subjective well-being of Brazilian street children. Paper presented at XXVI International Congress of Psychology, Montreal, Canada

Lusk M 1992 Street children of Rio de Janeiro. *International Social Work* **35**: 293–305

Martins R A 1996 Censo de crianças e adolescentes em situação de rua em São José do Rio Preto [A census of street children and adolescents in São José do Rio Preto, Brazil]. *Psicologia Reflexão e Crítica* **9**(1): 101–22

Matchinda B 1999 The impact of home background on the decision of children to run away: The case of Yaounde city street children in Cameroon. *Child Abuse and Neglect* **23**: 245–55

Raffaelli M, Siqueira E, Payne-Merrit A, Campos R, Ude W, Greco M, Greco D, Ruff A, Halsey N (The Street Youth Study Group) 1995 HIV-related knowledge and risk behaviors of street youth in Belo Horizonte, Brazil. *AIDS Education and Prevention* **7**(4): 287–97

S. H. Koller and C. S. Hutz

Stress and Cardiac Response

The origin of the concept of stress in health-related disciplines is closely linked to the idea that organisms react physiologically to the presence of danger or threat. This reactivity has a protective function since it provides the logistic and instrumental basis for adaptive behaviors such as fight or flight. However, if maintained for long periods, this protective or defensive reactivity may become a health risk, compromising the normal functioning of the organs involved. For many scientists the excessive physiological reactivity associated with stress is the main mechanism linking stress to illness.

1. *Cardiac Response*

The heart is a vital organ that reacts rapidly to physical and psychological demands. This reactivity is controlled by the central nervous system through neural and humoral pathways.

1.1 *Neural Pathways*

The neural pathway to the heart involves the two branches of the autonomic nervous system: the sympathetic and the parasympathetic. When the sympathetic nerves are activated, the heart rate accelerates and the heart contraction becomes stronger, resulting in a greater blood volume discharge and greater blood pressure in the vessels. When the parasympathetic nerves are activated, the heart rate decelerates, reducing the blood volume discharge and the blood pressure. The two branches can act reciprocally (one activated and the other inhibited), working in the same direction (increasing or decreasing the heart activity), or nonreciprocally (both activated or inhibited simultaneously), competing against each other for the final cardiac outcome.

1.2 *Humoral Pathways*

The humoral pathway involves the neural activation of the endocrine system, resulting in the secretion by

the adrenal glands of various hormones into the bloodstream. One of these hormones, adrenaline, when it reaches the heart, has the same effect as the sympathetic nerves: it increases the heart rate and the stroke volume. Other adrenal hormones, such as cortisol, facilitate the transfer of lipids and glucose into the bloodstream and also have an important indirect effect on the heart and on the arteries that supply blood to the heart muscle: the coronaries. The humoral effects on the heart are slower than the neural effects but they last longer, contributing to maintaining cardiac reactivity during prolonged periods.

1.3 Cardiac–Respiratory Interactions

The cardiovascular and the respiratory systems are both responsible for maintaining the continuous exchange of gases, oxygen and carbon dioxide, between the atmosphere and the tissues. The cardiovascular system reacts to changes in the concentration of these gases in the blood. Therefore, changes in the respiratory pattern which induce changes in gas concentration produce changes in the blood flow and cardiac activity in order to restore the appropriate gas concentration. Two examples of cardiac reactivity due to changes in respiration are the *hyperventilation* and *respiratory sinus arrhythmia* phenomena.

1.3.1 Hyperventilation. Hyperventilation is caused by rapid respiratory patterns provoking reduced carbon dioxide concentration in the blood. The cardiovascular response is a reduced blood flow to the brain and heart arteries, and an increase in cardiac activity due to parasympathetic inhibition.

1.3.2 Respiratory sinus arrhythmia. In contrast, respiratory sinus arrhythmia is caused by slow and relaxed respiratory patterns. The consequence is a rhythmic deceleration of the heart rate, due to parasympathetic activation, coinciding with each expiratory cycle.

2. Stress and Cardiac Defense

Cardiac defense can be defined as the cardiac response to perceived danger or threat. The concept of defense has a long tradition in physiology and psychology and it is in this context that the concept of stress was first introduced (see *Stress and Coping Theories*).

2.1 Historical Antecedents

At the beginning of the twentieth century, Ivan Pavlov and other Russian reflexologists used the term defense reflex to refer to protective physiological responses elicited by noxious stimulation, such as hand withdrawal from an electric shock, eyeblink to a puff of air, or vomit to bad food. A few years later, Walter Cannon used the term defense to refer to the 'fight or flight' response, a sympathetically mediated cardiovascular response to emergency situations aimed at providing energy supply to the body to facilitate adaptive behaviors such as attack (fight) or escape (flight). By the middle of the century and following Cannon's ideas, Hans Selye had introduced the concept of stress and used the term 'alarm' response to refer to the first stage of the physiological response to stressful situations.

2.2 Classical Approaches to Cardiac Defense

There have been two major approaches to the study of cardiac defense: (a) the cognitive or sensory processing approach built on Pavlov's distinction between orienting and defense reflexes, and (b) the motivational or emotional approach built on Cannon's ideas on the fight or flight response.

2.2.1 The cognitive approach. The cognitive approach assumes that cardiac changes to environmental stimuli reflect attentional and perceptual mechanisms aimed at facilitating or inhibiting the processing of the stimuli (Sokolov 1963, Graham 1979). The orienting response (a deceleration of the heart rate to moderate or novel stimulation) facilitates attention and perception of the stimuli, whereas the defense response (an acceleration of the heart rate to intense or aversive stimulation) reduces attention and perception as a form of protection against the stimuli. To study cardiac defense, this approach has developed tasks that present discrete sensory stimuli, such as noise, light, pictures, or words, manipulating their physical features—intensity, duration, and rise time—or their affective content—pleasant vs. unpleasant pictures or words. According to this approach, cardiac defense is characterized by a heart-rate accelerative response, peaking between 3 and 6 s after stimulus onset, which habituates slowly with stimulus repetition.

2.2.2 The motivational approach. The motivational approach assumes that cardiac changes in response to environmental demands reflect metabolic mechanisms aimed at providing the necessary energy to the body to support behavioral adjustments (Obrist 1981). If the appropriate behavior is to be quiet and relaxed, then the cardiac response will be a heart-rate deceleration. If the appropriate behavior is to be active, either physically or psychologically, then the cardiac response will be a heart-rate acceleration.

To study cardiac defense, this approach has developed tasks that involve emotionally or cognitively challenging situations maintained throughout several minutes, such as keeping the hand in cold water, pressing a key as fast as possible to avoid shocks, or doing complex arithmetic tasks. The term 'mental stress' was introduced in this context to refer to tasks that require 'mental effort' instead of 'physical effort.' Under such conditions cardiac defense is characterized by a heart-rate acceleration maintained during the task duration, accompanied by reduced cardiac variability linked to hyperventilation and loss of respiratory sinus arrythmia.

3. *A New Approach to Cardiac Defense*

The cognitive and motivational approaches to cardiac defense have been difficult to reconcile in the past. However, new data have been accumulated on defense responses, both in animals and humans, and on the neurophysiological brain circuits controlling such responses, which not only facilitate integration of the classic approaches but also have potential implications for the concept of stress and its relation to cardiovascular illness.

3.1 *Natural Defense*

In natural settings, such as the imminence of a predator, the defense reaction is not a unique response that could be conceptualized as cognitive vs. motivational, or vice versa. Rather, the defense reaction follows a dynamic sequential process with initial phases in which attentional factors predominate, aimed at the detection and analysis of the potential danger, and if the process continues, later phases in which behavioral factors predominate, aimed at active defense such as fight or flight. This sequential process can be divided into three stages: pre-encounter, post-encounter, and circa strike. Depending on the spatial and temporal proximity of the predator, different components of the defense reaction may take place successively.

3.2 *Protective Reflexes*

Motor startle and cardiac defense are two protective reflexes that have been extensively studied in recent years in relation to both cognitive and emotional modulatory factors. Startle is a whole-body motor flexor response, including an eyeblink that is elicited by abrupt stimulation such as a brief, intense noise or light. Startle is augmented when the organism is in an aversive emotional state and inhibited when the organism is in a positive emotional state. Similarly, startle is potentiated in the first stages of the natural defense reaction when attentional factors predominate, and inhibited during the later phases when behavioral factors predominate (Lang 1995).

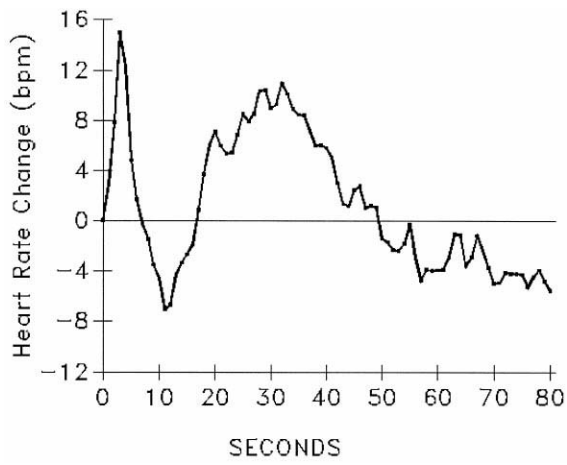

Figure 1
Cardiac defense response to an intense noise of half a second duration. The horizontal axis represents 80 s after noise onset. The vertical axis represents the cardiac changes in beats per minute (bpm) from the baseline

On the other hand, studies on the cardiac components of the defense response to unexpected noises or shocks have questioned the classical models on response characteristics and their functional significance (Vila 1995, Vila et al. 1997). The cardiac response to an unexpected noise is not only a heart-rate acceleration. The response is a complex pattern of heart-rate changes, observed throughout the 80 s after stimulus onset, with two large accelerative and two smaller decelerative components in alternating order (see Fig. 1). The evocation of the response has both parasympathetic (first acceleration and deceleration) and sympathetic (second acceleration and deceleration) influences. As regards its psychological significance, contrary to the classical cognitive model, cardiac defense is accompanied by an increase in sensory processing (sensory intake). On the other hand, emotional factors are also involved, producing similar modulatory effects as in motor startle. For example, phobic subjects presented with aversive noise while viewing pictures of their phobic object show a clear augmentation of the two accelerative components of the response.

3.3 *The Brain's Defense System*

The above data support a description and interpretation of cardiac defense as a sequence of heart-rate changes with both accelerative and decelerative components, with both parasympathetic and sympathetic

influences, and with both cognitive (attentional) and emotional (energetic) significance. The structural foundation of this response is in the brain's aversive motive system. What is known about this system comes from animal research, particularly from studies on defense reactions and fear conditioning (see LeDoux 1996 and Lang et al. 1997 for a comprehensive review). Neuroscientists have traced the chain of neural activation in the brain, starting from the sensory input, proceeding through the subcortical and cortical connecting structures, and finishing in the autonomic and motor effectors. A critical structure in this circuit is the amygdala, a subcortical center located within the temporal lobes that receives inputs from the sensory thalamus and various cortical structures. Downstream from the amygdala, the defense circuit branches into different structures, controlling separate forms of defense. This hierarchical organization, together with the learned plasticity in the anatomy and functioning of the brain's structures, provides considerably variety in the pattern of defensive output. Depending on the situational context and on the weight of the local pathways within the circuit, the aversively motivated responses can be escape or avoidance, aggressive behavior, hormonal and autonomic responses, or the augmentation of motor startle.

4. *Conclusions: Implications for Stress and Cardiovascular Disease*

The classical models of defense did not emphasize the dynamic character of the defense response—obvious in natural settings—and the simultaneous involvement of different psychological processes—cognitive and motivational. The new model matches better the prevalent model of stress in health-related disciplines: the cognitive-transactional model (see *Stress and Coping Theories*). The continuous dynamic interaction between the situation and the person, leading to activation or deactivation of the stress response, is the main feature of the transactional model. In addition, the new approach to defense can help to understand the specific nature of stress, differentiating it from other related concepts such as anxiety or arousal. Based on the new defense model, stress can be defined as the sustained activation of the brain's defense motivational system. A state of maintained activation of this system implies a variety of specific defense responses being continuously elicited, including neural and humoral cardiovascular responses that correlate with poor cardiovascular health: maintained increases in heart rate, stroke volume and blood pressure, and maintained decreases in respiratory sinus arrhythmia and cardiac variability. These sustained responses are instrumental for the development of cardiovascular diseases such as essential hypertension and coronary heart disease.

See also: Cardiovascular Conditioning: Neural Substrates; Coronary Heart Disease (CHD), Coping with; Coronary Heart Disease (CHD): Psychosocial Aspects; Hypertension: Psychosocial Aspects; Stress and Health Research; Stress, Neural Basis of; Stress: Psychological Perspectives

Bibliography

Graham F K 1979 Distinguishing among orienting, defense and startle reflexes. In: Kimmel H D, van Olst E H, Orlebeke F J (eds.) *The Orienting Reflex in Humans*. Erlbaum, Hillsdale, NJ, pp. 137–67

Lang P J 1995 The emotion probe: Studies of motivation and attention. *American Psychologist* **50**: 372–85

Lang P J, Bradley M M, Cuthbert B N 1997 Motivated attention: Affect, activation, and action. In: Lang P J, Simons R F, Balaban M T (eds.) *Attention and Orienting*. Erlbaum, Mahwah, NJ, pp. 97–135

LeDoux J 1996 *The Emotional Brain*. Pergamon, New York

Obrist P A 1981 *Cardiovascular Psychophysiology: A Perspective*. Simon & Schuster, New York

Sokolov E N 1963 *Perception and the Conditioned Reflex*. Pergamon, Oxford, UK

Vila J 1995 Cardiac psychophysiology and health. In: Rodríguez-Marín J (ed.) *Health Psychology and Quality of Life Research*. University of Alicante Press, Alicante, Spain, pp. 520–40

Vila J, Pérez M N, Fernández M D, Pegalajar J, Sánchez M 1997 Attentional modulation of the cardiac defense response in humans. *Psychophysiology* **34**: 482–7

J. Vila

Stress and Coping Theories

For the last five decades the term stress has enjoyed increasing popularity in the behavioral and health sciences. It first was used in physics in order to analyze the problem of how man-made structures must be designed to carry heavy loads and resist deformation by external focus. In this analysis, stress referred to external pressure or force applied to a structure, while strain denoted the resulting internal distortion of the object (for the term's history, cf. Hinkle 1974, Mason 1975a, 1975c). In the transition from physics to the behavioral sciences, the usage of the term stress changed. In most approaches it now designates bodily processes created by circumstances that place physical or psychological demands on an individual (Selye 1976). The external forces that impinge on the body are called stressors (McGrath 1982).

1. *Theories of Stress*

Theories that focus on the specific relationship between external demands (stressors) and bodily processes (stress) can be grouped in two different categories:

approaches to 'systemic stress' based in physiology and psychobiology (among others, Selye 1976) and approaches to 'psychological stress' developed within the field of cognitive psychology (Lazarus 1966, 1991, Lazarus and Folkman 1984, McGrath 1982).

1.1 Systemic Stress: Selye's Theory

The popularity of the stress concept in science and mass media stems largely from the work of the endocrinologist Hans Selye. In a series of animal studies he observed that a variety of stimulus events (e.g., heat, cold, toxic agents) applied intensely and long enough are capable of producing common effects, meaning not specific to either stimulus event. (Besides these nonspecific changes in the body, each stimulus produces, of course, its specific effect, heat, for example, produces vasodilatation, and cold vasoconstriction.) According to Selye, these *nonspecifically caused changes* constitute the stereotypical, i.e., *specific*, response pattern of systemic stress. Selye (1976, p. 64) defines this stress as 'a state manifested by a syndrome which consists of all the nonspecifically induced changes in a biologic system.'

This stereotypical response pattern, called the 'General Adaptation Syndrome' (GAS), proceeds in three stages. (a) The *alarm reaction* comprises an initial shock phase and a subsequent countershock phase. The shock phase exhibits autonomic excitability, an increased adrenaline discharge, and gastro-intestinal ulcerations. The countershock phase marks the initial operation of defensive processes and is characterized by increased adrenocortical activity. (b) If noxious stimulation continues, the organism enters the *stage of resistance*. In this stage, the symptoms of the alarm reaction disappear, which seemingly indicates the organism's adaptation to the stressor. However, while resistance to the noxious stimulation increases, resistance to other kinds of stressors decreases at the same time. (c) If the aversive stimulation persists, resistance gives way to the *stage of exhaustion*. The organism's capability of adapting to the stressor is exhausted, the symptoms of stage (a) reappear, but resistance is no longer possible. Irreversible tissue damages appear, and, if the stimulation persists, the organism dies.

Although Selye's work influenced a whole generation of stress researchers, marked weaknesses in his theory soon became obvious. First of all, Selye's conception of stress as a reaction to a multitude of different events had the fatal consequence that the stress concept became the melting pot for all kinds of approaches. Thus, by becoming a synonym for diverse terms such as, for example, anxiety, threat, conflict, or emotional arousal, the concept of stress was in danger of losing its scientific value (cf. Engel 1985). Besides this general reservation, specific critical issues have been raised. One criticism was directed at the theory's core assumption of a nonspecific causation of the GAS. Mason (1971, 1975b) pointed out that the stressors observed as effective by Selye carried a common emotional meaning: they were novel, strange, and unfamiliar to the animal. Thus, the animal's state could be described in terms of helplessness, uncertainty, and lack of control. Consequently, the hormonal GAS responses followed the (specific) emotional impact of such influences rather than the influences as such. In accordance with this assumption, Mason (1975b) demonstrated that in experiments where uncertainty had been eliminated no GAS was observed. This criticism lead to a second, more profound argument: unlike the physiological stress investigated by Selye, the stress experienced by humans is almost always the result of a cognitive mediation (cf. Arnold 1960, Janis 1958, Lazarus 1966, 1974). Selye, however, fails to specify those mechanisms that may explain the *cognitive transformation* of 'objective' noxious events into the subjective experience of being distressed. In addition, Selye does not take into account *coping mechanisms* as important mediators of the stress–outcome relationship. Both topics are central to psychological stress theories as, for example, elaborated by the Lazarus group.

A derivative of the systemic approach is the research on *critical life events*. An example is the influential hypothesis of Holmes and Rahe (1967), based on Selye's work, that changes in habits, rather than the threat or meaning of critical events, is involved in the genesis of disease. The authors assumed that critical life events, regardless of their specific (e.g., positive or negative) quality, stimulate change that produces challenge to the organism. Most of this research, however, has not been theoretically driven and exhibited little empirical support for this hypothesis (for a critical evaluation, see Thoits 1983).

1.2 Psychological Stress: The Lazarus Theory

Two concepts are central to any psychological stress theory: *appraisal*, i.e., individuals' evaluation of the significance of what is happening for their well-being, and *coping*, i.e., individuals' efforts in thought and action to manage specific demands (cf. Lazarus 1993).

Since its first presentation as a comprehensive theory (Lazarus 1966), the Lazarus stress theory has undergone several essential revisions (cf. Lazarus 1991, Lazarus and Folkman 1984, Lazarus and Launier 1978). In the latest version (see Lazarus 1991), stress is regarded as a *relational* concept, i.e., stress is not defined as a specific kind of external stimulation nor a specific pattern of physiological, behavioral, or subjective reactions. Instead, stress is viewed as a relationship ('transaction') between individuals and their environment. 'Psychological stress refers to a relationship with the environment that the person appraises as significant for his or her well being

and in which the demands tax or exceed available coping resources' (Lazarus and Folkman 1986, p. 63). This definition points to two processes as central mediators within the person–environment transaction: *cognitive appraisal* and *coping*.

The concept of *appraisal*, introduced into emotion research by Arnold (1960) and elaborated with respect to stress processes by Lazarus (1966, Lazarus and Launier 1978), is a key factor for understanding stress-relevant transactions. This concept is based on the idea that emotional processes (including stress) are dependent on actual expectancies that persons manifest with regard to the significance and outcome of a specific encounter. This concept is necessary to explain individual differences in quality, intensity, and duration of an elicited emotion in environments that are objectively equal for different individuals. It is generally assumed that the resulting state is generated, maintained, and eventually altered by a specific pattern of appraisals. These appraisals, in turn, are determined by a number of personal and situational factors. The most important factors on the personal side are motivational dispositions, goals, values, and generalized expectancies. Relevant situational parameters are predictability, controllability, and imminence of a potentially stressful event.

In his monograph on emotion and adaptation, Lazarus (1991) developed a comprehensive emotion theory that also includes a stress theory (cf. Lazarus 1993). This theory distinguishes two basic forms of appraisal, primary and secondary appraisal (see also Lazarus 1966). These forms rely on different sources of information. Primary appraisal concerns whether something of relevance to the individual's well being occurs, whereas secondary appraisal concerns coping options.

Within *primary appraisal*, three components are distinguished: *goal relevance* describes the extent to which an encounter refers to issues about which the person cares. *Goal congruence* defines the extent to which an episode proceeds in accordance with personal goals. *Type of ego-involvement* designates aspects of personal commitment such as self-esteem, moral values, ego-ideal, or ego-identity. Likewise, three *secondary appraisal* components are distinguished: *blame or credit* results from an individual's appraisal of who is responsible for a certain event. By *coping potential* Lazarus means a person's evaluation of the prospects for generating certain behavioral or cognitive operations that will positively influence a personally relevant encounter. *Future expectations* refer to the appraisal of the further course of an encounter with respect to goal congruence or incongruence.

Specific patterns of primary and secondary appraisal lead to different kinds of stress. Three types are distinguished: harm, threat, and challenge (Lazarus and Folkman 1984). *Harm* refers to the (psychological) damage or loss that has already happened. *Threat* is the anticipation of harm that may be imminent. *Challenge* results from demands that a person feels confident about mastering. These different kinds of psychological stress are embedded in specific types of emotional reactions, thus illustrating the close conjunction of the fields of stress and emotions.

Lazarus (1991) distinguishes 15 basic emotions. Nine of these are negative (anger, fright, anxiety, guilt, shame, sadness, envy, jealousy, and disgust), whereas four are positive (happiness, pride, relief, and love). (Two more emotions, hope and compassion, have a mixed valence.) At a molecular level of analysis, the anxiety reaction, for example, is based on the following pattern of primary and secondary appraisals: there must be some goal relevance to the encounter. Furthermore, goal incongruence is high, i.e., personal goals are thwarted. Finally, ego-involvement concentrates on the protection of personal meaning or ego-identity against existential threats. At a more molar level, specific appraisal patterns related to stress or distinct emotional reactions are described as *core relational themes*. The theme of anxiety, for example, is the confrontation with uncertainty and existential threat. The core relational theme of relief, however, is 'a distressing goal-incongruent condition that has changed for the better or gone away' (Lazarus 1991).

Coping is intimately related to the concept of cognitive appraisal and, hence, to the stress-relevant person-environment transactions. Most approaches in coping research follow Folkman and Lazarus (1980, p. 223), who define coping as 'the cognitive and behavioral efforts made to master, tolerate, or reduce external and internal demands and conflicts among them.'

This definition contains the following implications. (a) Coping actions are not classified according to their effects (e.g., as reality-distorting), but according to certain characteristics of the coping process. (b) This process encompasses behavioral as well as cognitive reactions in the individual. (c) In most cases, coping consists of different single acts and is organized sequentially, forming a coping *episode*. In this sense, coping is often characterized by the simultaneous occurrence of different action sequences and, hence, an interconnection of coping episodes. (d) Coping actions can be distinguished by their focus on different elements of a stressful encounter (cf. Lazarus and Folkman 1984). They can attempt to change the person–environment realities behind negative emotions or stress (*problem-focused coping*). They can also relate to internal elements and try to reduce a negative emotional state, or change the appraisal of the demanding situation (*emotion-focused coping*).

1.3 Resource Theories of Stress: A Bridge between Systemic and Cognitive Viewpoints

Unlike approaches discussed so far, resource theories of stress are not primarily concerned with factors that

create stress, but with resources that preserve well being in the face of stressful encounters. Several social and personal constructs have been proposed, such as *social support* (Schwarzer and Leppin 1991), *sense of coherence* (Antonovsky 1979), *hardiness* (Kobasa 1979), *self-efficacy* (Bandura 1977), or *optimism* (Scheier and Carver 1992). Whereas self-efficacy and optimism are single protective factors, hardiness and sense of coherence represent tripartite approaches. Hardiness is an amalgam of three components: internal control, commitment, and a sense of challenge as opposed to threat. Similarly, sense of coherence consists of believing that the world is meaningful, predictable, and basically benevolent. Within the social support field, several types have been investigated, such as instrumental, informational, appraisal, and emotional support.

The recently offered conservation of resources (COR) theory (Hobfoll 1989, Hobfoll et al. 1996) assumes that stress occurs in any of three contexts: when people experience loss of resources, when resources are threatened, or when people invest their resources without subsequent gain. Four categories of resources are proposed: object resources (i.e., physical objects such as home, clothing, or access to transportation), condition resources (e.g., employment, personal relationships), personal resources (e.g., skills or self-efficacy), and energy resources (means that facilitate the attainment of other resources, for example, money, credit, or knowledge).

Hobfoll and co-workers outlined a number of testable hypotheses (called principles) derived from basic assumptions of COR (cf. Hobfoll et al. 1996).

(a) Loss of resources is the primary source of stress. This principle contradicts the fundamental assumption of approaches on critical life events (cf. Holmes and Rahe 1967) that stress occurs whenever individuals are forced to readjust themselves to situational circumstances, may these circumstances be positive (e.g., marriage) or negative (e.g., loss of a beloved person). In an empirical test of this basic principle, Hobfoll and Lilly (1993) found that only loss of resources was related to distress.

(b) Resources act to preserve and protect other resources. Self-esteem is an important resource that may be beneficial for other resources. Hobfoll and Leiberman (1987), for example, observed that women who were high in self-esteem made good use of social support when confronted with stress, whereas those who lacked self-esteem interpreted social support as an indication of personal inadequacy and, consequently, misused support.

(c) Following stressful circumstances, individuals have an increasingly depleted resource pool to combat further stress. This depletion impairs individuals' capability of coping with further stress, thus resulting in a loss spiral. This process view of resource investment requires to focus on how the interplay between resources and situational demands changes over time as stressor sequences unfold. In addition, this principle shows that it is important to investigate not only the effect of resources on outcome, but also of outcome on resources.

2. Coping Theories

2.1 Classification of Approaches

The Lazarus model outlined above represents a specific type of coping theory. These theories may be classified according to two independent parameters: (a) trait-oriented versus state-oriented, and (b) microanalytic versus macroanalytic approaches (cf. Krohne 1996). Trait-oriented and state-oriented research strategies have different objectives: The *trait-oriented* (or dispositional) strategy aims at early identification of individuals whose coping resources and tendencies are inadequate for the demands of a specific stressful encounter. An early identification of these persons will offer the opportunity for establishing a selection (or placement) procedure or a successful primary prevention program. Research that is *state-oriented*, i.e., which centers around actual coping, has a more general objective. This research investigates the relationships between coping strategies employed by an individual and outcome variables such as self-reported or objectively registered coping efficiency, emotional reactions accompanying and following certain coping efforts, or variables of adaptational outcome (e.g., health status or test performance). This research strategy intends to lay the foundation for a general modificatory program to improve coping efficacy. *Microanalytic* approaches focus on a large number of specific coping strategies, whereas *macroanalytic* analysis operates at a higher level of abstraction, thus concentrating on more fundamental constructs.

S. Freud's (1926) 'classic' defense mechanisms conception is an example of a *state-oriented*, *macroanalytic* approach. Although Freud distinguished a multitude of defense mechanisms, in the end, he related these mechanisms to two basic forms: repression and intellectualization (see also A. Freud 1936). The *trait-oriented* correspondence of these basic defenses is the personality dimension repression–sensitization (Byrne 1964, Eriksen 1966). The distinction of the two basic functions of emotion-focused and problem-focused coping proposed by Lazarus and Folkman (1984) represents another macroanalytic state approach. In its actual research strategy, however, the Lazarus group extended this macroanalytic approach to a *microanalytic* strategy. In their 'Ways of Coping Questionnaire' (WOCQ; cf. Folkman and Lazarus 1988, Lazarus 1991), Lazarus and co-workers distinguish eight groups of coping strategies: confrontative coping, distancing, self-controlling, seeking social support, accepting responsibility, escape-avoidance,

planful problem-solving, and positive reappraisal. The problem with this conception and, as a consequence, the measurement of coping is that these categories are only loosely related to the two basic coping functions.

Unlike the *macroanalytic*, *trait-oriented* approach that generated a multitude of theoretical conceptions, the *microanalytic*, *trait-oriented* strategy is mostly concerned with constructing multidimensional inventories (overviews in Schwarzer and Schwarzer 1996). Almost all of these measurement approaches, however, lack a solid theoretical foundation (cf. Krohne 1996).

2.2 Macroanalytic, Trait-Oriented Coping Theories

Research on the processes by which individuals cope with stressful situations has grown substantially over the past three decades (cf. Lazarus 1991, Zeidner and Endler 1996). Many trait-oriented approaches in this field have established two constructs central to an understanding of cognitive responses to stress: *vigilance*, that is, the orientation toward stressful aspects of an encounter, and *cognitive avoidance*, that is, averting attention from stress-related information (cf. Janis 1983, Krohne 1978, 1993, Roth and Cohen 1986). Approaches corresponding to these conceptions are repression–sensitization (Byrne 1964), monitoring-blunting (Miller 1980, 1987), or attention-rejection (Mullen and Suls 1982). With regard to the relationship between these two constructs, Byrne's approach specifies a unidimensional, bipolar structure, while Miller as well as Mullen and Suls leave this question open. Krohne, however, explicitly postulates an independent functioning of the dimensions vigilance and cognitive avoidance.

2.2.1 Repression–sensitization. The repression–sensitization construct (cf. Byrne 1964, Eriksen 1966) relates different forms of dispositional coping to one bipolar dimension. When confronted with a stressful encounter, persons located at one pole of this dimension (repressers) tend to deny or minimize the existence of stress, fail to verbalize feelings of distress, and avoid thinking about possible negative consequences of this encounter. Persons at the opposite pole (sensitizers) react to stress-related cues by way of enhanced information search, rumination, and obsessive worrying. The concept of repression–sensitization is theoretically founded in research on perceptual defense (Bruner and Postman 1947), an approach that combined psychodynamic ideas with the functionalistic behavior analysis of Brunswik (1947).

2.2.2 Monitoring and blunting. The conception of monitoring and blunting (Miller 1980, 1987) originated from the same basic assumptions formulated earlier by Eriksen (1966) for the repression–sensitization construct. Miller conceived both constructs as cognitive informational styles and proposed that individuals who encounter a stressful situation react with arousal according to the amount of attention they direct to the stressor. Conversely, the arousal level can be lowered, if the person succeeds in reducing the impact of aversive cues by employing avoidant cognitive strategies such as distraction, denial, or reinterpretation. However, these coping strategies, called blunting, should only be adaptive if the aversive event is uncontrollable. Examples of uncontrollable events are impending surgery or an aversive medical examination (Miller and Mangan 1983). If control is available, strategies called monitoring, i.e., seeking information about the stressor, are the more adaptive forms of coping. Although initially these strategies are associated with increased stress reactions, they enable the individual to gain control over the stressor in the long run, thus reducing the impact of the stressful situation. An example of a more controllable stressor is preparing for an academic exam.

The general relationship between a stressor's degree of controllability and the employment of monitoring or blunting strategies can be moderated by situative and personal influences. With regard to situation, the noxious stimulation may be so intense that blunting strategies, such as attentional diversion, are ineffective with respect to reducing stress-related arousal. Concerning personality, there are relatively stable individual differences in the inclination to employ blunting or monitoring coping when encountering a stressor.

2.2.3 The model of coping modes. Similar to Miller's monitoring-blunting conception, the model of coping modes (MCM) deals with individual differences in attention orientation and emotional-behavioral regulation under stressful conditions (Krohne 1993). The MCM extends the (largely descriptive) monitoring-blunting conception (as well as the repression–sensitization approach) in that it relates the dimensions vigilance and cognitive avoidance to an explicative cognitive-motivational basis. It assumes that most stressful, especially anxiety evoking, situations are characterized by two central features: the presence of *aversive stimulation* and a high degree of *ambiguity*. The experiential counterparts of these situational features are *emotional arousal* (as being primarily related to aversive stimulation) and *uncertainty* (related to ambiguity). Arousal, in turn, should stimulate the tendency to cognitively avoid (or inhibit) the further processing of cues related to the aversive encounter, whereas uncertainty activates vigilant tendencies.

These two coping processes are conceptually linked to personality by the hypothesis that the habitual preference for avoidant or vigilant coping strategies reflects individual differences in the susceptibility to emotional arousal or uncertainty. Individuals who are

especially susceptible to states of stress-induced emotional arousal are supposed to habitually employ cognitive avoidance. The employment of avoidant strategies primarily aims at shielding the person from an increase in arousal (*arousal-motivated coping behavior*). Individuals who are especially affected by the uncertainty experienced in most stressful situations are supposed to habitually employ vigilant coping. Thus, the employment of vigilant strategies follows a plan that is aimed at minimizing the probability of unanticipated occurrence of aversive events (*uncertainty-motivated coping behavior*).

The MCM conceives the habitual coping tendencies of vigilance and cognitive avoidance as independent personality dimensions. That means, *aggregated* across a multitude of stressful encounters, the employment of vigilant strategies and of avoidant ones does not preclude each other. Thus, four *coping modes* can be defined. (a) Persons who score high on vigilance and low on cognitive avoidance are called *sensitizers*. These persons are primarily concerned with reducing uncertainty by directing their attention towards stress-relevant information. (b) Individuals with the opposite pattern are designated as *repressers*. These persons minimize the experience of arousal by avoiding aversive information. (c) *Nondefensives* have low scores on both dimensions. These persons are supposed to flexibly adapt to the demands of a stressful encounter. Instead of frequently employing vigilant or avoidant coping strategies, they prefer to act instrumentally in most situations. (d) Individuals who exhibit high scores on both dimensions are called *high anxious*. In employing vigilant as well as avoidant coping strategies, these persons try to reduce both the subjective uncertainty and the emotional arousal induced by stressful encounters. Because the two goals are incompatible in most situations, high-anxious persons are assumed to show fluctuating and therefore less-efficient coping behavior. Approaches to assess individual differences in vigilance and cognitive avoidance are described in Krohne et al. (2000). Empirical results related to predictions derived from the MCM are presented in Krohne (1993, 1996), and Krohne et al. (1992).

3. Future Perspectives

Although the fields of stress and coping research represent largely explored territory, there are still fertile perspectives to be pursued in future research. Among the promising lines of research, two perspectives will be mentioned here.

(a) Compared to the simplistic stimulus-response conception of stress inherent in early approaches on stress, the 'psychological' (i.e., cognitive transformation) approach of the Lazarus group clearly represents progress. However, in advocating a completely 'subjective' orientation in conceptualizing stress, Lazarus overstated the 'cognitive turn' in stress research. In stating that 'we might do better by describing relevant environments and their psychological meanings through the lenses of individuals' (Lazarus 1990, p. 8) he took a stand that is at variance with the multivariate, systems-theory perspective proposed in his recent publications on stress and emotions (Lazarus 1990, 1991).

First, the stress process contains variables to be assessed both subjectively and objectively, such as constraints, temporal aspects, or social support networks, as well as responses to be measured at different levels (cf. Lazarus 1990, Table 1). Second, the fact that most objective features relevant to stress-related outcomes exert their influence via a process of cognitive transformation (Mischel and Shoda 1995) does not mean that objective features can be neglected. It is of great practical and theoretical importance to know which aspects of the 'objective' environment an individual selects for transformation, and how these characteristics are subjectively represented. Third, as far as response levels are concerned, it is obvious that stressors do not only create subjective (cognitive) responses but also reactions at the somatic and the behavioral-expressive level. In fact, many individuals (especially those high in cognitive avoidance) are characterized by a *dissociation* of subjective and objective stress responses (cf. Kohlmann 1997; for an early discussion of the psychological meaning of this dissociation see Lazarus 1966). These individuals may manifest, for example, relatively low levels of subjective distress but at the same time considerable elevations in autonomic arousal. In recent years, the concept of subjective-autonomic response dissociation has become increasingly important in clarifying the origin and course of physical diseases and affective disorders.

(b) It is important to define central person-specific *goals* (or reference values) in coping, such as reducing uncertainty, inhibiting emotional arousal, or trying to change the causes of a stressful encounter. These goals are not only central to understanding the stress and coping process, they are, in fact, 'the core of personality' (Karoly 1999). Goals define the trans-situational and trans-temporal relevance of certain stressors, serve as links to other constructs such as self-concept or expectancies, influence regulatory processes such as coping, and define the efficiency of these processes (cf. Karoly 1999, Lazarus 1991, Mischel and Shoda 1995). Instead of applying global and relatively content-free trait concepts in stress and coping research such as anxiety, depression, or optimism, a more fertile perspective would be to study personality in this field by paying attention to what people are trying to do instead of only observing how they actually respond to stressful events.

See also: Coping across the Lifespan; Coping Assessment; Health: Self-regulation; Self-regulation

in Adulthood; Self-regulation in Childhood; Stress: Measurement by Self-report and Interview; Stress, Neural Basis of; Stress: Psychological Perspectives

Bibliography

Antonovsky A 1979 *Health, Stress, and Coping*. Jossey-Bass, San Francisco, CA

Arnold M B 1960 *Emotion and Personality* (2 Vols.). Columbia University Press, New York

Bandura A 1977 Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* **84**: 191–215

Bruner J S, Postman L 1947 Emotional selectivity in perception and reaction. *Journal of Personality* **16**: 69–77

Brunswik E 1947 *Systematic and Representative Design of Psychological Experiments: With Results in Physical and Social Perception*. University of California Press, Berkeley, CA

Byrne D 1964 Repression–sensitization as a dimension of personality. In: Maher B A (ed.) *Progress in Experimental Personality Research*. Academic Press, New York, Vol. 1, pp. 169–220

Engel B T 1985 Stress is a noun! No, a verb! No, an adjective. In: Field T M, McCabe P M, Schneiderman N (eds.) *Stress and Coping*. Erlbaum, Hillsdale, NJ, pp. 3–12

Eriksen C W 1966 Cognitive responses to internally cued anxiety. In: Spielberger C D (ed.) *Anxiety and Behavior*. Academic Press, New York, pp. 327–60

Folkman S, Lazarus R S 1980 An analysis of coping in a middle-aged community sample. *Journal of Health and Social Behavior* **21**: 219–39

Folkman S, Lazarus R S 1988 *Ways of Coping Questionnaire Research edition*. Consulting Psychologists Press, Palo Alto, CA

Freud A 1936 *Das Ich und die Abwehrmechanismen* [The ego and the mechanisms of defense]. Internationaler Psychoanalytischer Verlag, Vienna

Freud S 1926 *Hemmung, Symptom und Angst* [Inhibitions, symptoms and anxiety]. Internationaler Psychoanalytischer Verlag, Vienna

Hinkle L E 1974 The concept of 'stress' in the biological and social sciences. *International Journal of Psychiatry in Medicine* **5**: 335–57

Hobfoll S E 1989 Conservation of resources: A new attempt at conceptualizing stress. *American Psychologist* **44**: 513–24

Hobfoll S E, Freedy J R, Green B L, Solomon S D 1996 Coping reactions to extreme stress: The roles of resource loss and resource availability. In: Zeidner M, Endler N S (eds.) *Handbook of Coping: Theory, Research, Applications*. Wiley, New York, pp. 322–49

Hobfoll S E, Leiberman J R 1987 Personality and social resources in immediate and continued stress-resistance among women. *Journal of Personality and Social Psychology* **52**: 18–26

Hobfoll S E, Lilly R S 1993 Resource conservation as a strategy for community psychology. *Journal of Community Psychology* **21**: 128–48

Holmes T H, Rahe R H 1967 The social readjustment rating scale. *Journal of Psychosomatic Research* **11**: 213–8

Janis I L 1958 *Psychological Stress: Psychoanalytic and Behavioral Studies of Surgical Patients*. Wiley, New York

Janis I L 1983 Stress inoculation in health care: Theory and research. In: Meichenbaum D, Jaremko M (eds.) *Stress Reduction and Prevention*. Plenum, New York, pp. 67–99

Karoly P 1999 A goal systems self-regulatory perspective on personality, psychopathology, and change. *Review of General Psychology* **3**: 264–91

Kobasa S C 1979 Stressful life events, personality, and health: An inquiry into hardiness. *Journal of Personality and Social Psychology* **37**: 1–11

Kohlmann C-W 1997 *Persönlichkeit und Emotionsregulation: Defensive Bewältigung von Angst und Streß* [Personality and the regulation of emotion]. Huber, Bern, Switzerland

Krohne H W 1978 Individual differences in coping with stress and anxiety. In: Spielberger C D, Sarason I G (eds.) *Stress and Anxiety*. Hemisphere, Washington, DC, Vol. 5, pp. 233–60

Krohne H W 1993 Vigilance and cognitive avoidance as concepts in coping research. In: Krohne H W (ed.) *Attention and Avoidance. Strategies in Coping with Aversiveness*. Hogrefe & Huber, Seattle, WA, pp. 19–50

Krohne H W 1996 Individual differences in coping. In: Zeidner M, Endler N S (eds.) *Handbook of Coping: Theory, Research, Applications*. Wiley, New York, pp. 381–409

Krohne H W, Egloff B, Varner L J, Burns L R, Weidner G, Ellis H C 2000 The assessment of dispositional vigilance and cognitive avoidance: Factorial structure, psychometric properties, and validity of the Mainz Coping Inventory. *Cognitive Therapy and Research* **24**: 297–311

Krohne H W, Hock M, Kohlmann C-W 1992 Coping dispositions, uncertainty, and emotional arousal. In: Strongman K T (ed.) *International Review of Studies on Emotion*. Wiley, Chichester, UK, Vol. 2, pp. 73–95

Lazarus R S 1966 *Psychological Stress and the Coping Process*. McGraw-Hill, New York

Lazarus R S 1974 Psychological stress and coping in adaptation and illness. *International Journal of Psychiatry in Medicine* **5**: 321–33

Lazarus R S 1990 Theory-based stress measurement. *Psychological Inquiry* **1**: 3–13

Lazarus R S 1991 *Emotion and Adaptation*. Oxford University Press, New York

Lazarus R S 1993 Coping theory and research: Past, present, and future. *Psychosomatic Medicine* **55**: 234–47

Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York

Lazarus R S, Folkman S 1986 Cognitive theories of stress and the issue of circularity. In: Appley M H, Trumbull R (eds.) *Dynamics of Stress. Physiological, Psychologcal, and Social Perspectives*. Plenum, New York, pp. 63–80

Lazarus R S, Launier R 1978 Stress-related transactions between person and environment. In: Pervin L A, Lewis M (eds.) *Perspectives in Interactional Psychology*. Plenum, New York, pp. 287–327

Mason J W 1971 A re-evaluation of the concept of 'non-specifity' in stress theory. *Journal of Psychiatric Research* **8**: 323–33

Mason J W 1975a A historical view of the stress field. Part I. *Journal of Human Stress* **1**: 6–12

Mason J W 1975b A historical view of the stress field. Part II. *Journal of Human Stress* **1**: 22–36

Mason J W 1975c Emotion as reflected in patterns of endocrine integration. In: Levi L (ed.) *Emotions: Their Parameters and Measurement*. Raven, New York, pp. 143–81

McGrath J E 1982 Methodological problems in research on stress. In: Krohne H W, Laux L (eds.) *Achievement, Stress, and Anxiety*. Hemisphere, Washington, DC, pp. 19–48

Miller S M 1980 When is little information a dangerous thing? Coping with stressful events by monitoring versus blunting.

In: Levine S, Ursin H (eds.) *Health and Coping*. Plenum, New York, pp. 145–69
Miller S M 1987 Monitoring and blunting: Validation of a questionnaire to assess styles of information seeking under threat. *Journal of Personality and Social Psychology* **52**: 345–53
Miller S M, Mangan C E 1983 Interacting effects of information and coping style in adapting to gynecologic stress: Should the doctor tell all? *Journal of Personality and Social Psychology* **45**: 223–36
Mischel W, Shoda Y 1995 A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review* **102**: 248–68
Mullen B, Suls J 1982 The effectiveness of attention and rejection as coping styles: A meta-analysis of temporal differences. *Journal of Psychosomatic Research* **26**: 43–9
Roth S, Cohen L J 1986 Approach, avoidance, and coping with stress. *American Psychologist* **41**: 813–9
Scheier M F, Carver C S 1992 Effects of optimism on psychological and physical well-being: Theoretical overview and empirical update. *Cognitive Therapy and Research* **16**: 201–28
Schwarzer R, Leppin A 1991 Social support and health: A theoretical and empirical overview. *Journal of Social and Personal Relationships* **8**: 99–127
Schwarzer R, Schwarzer C 1996 A critical survey of coping instruments. In: Zeidner M, Endler N S (eds.) *Handbook of Coping: Theory, Research, Applications*. Wiley, New York, pp. 107–32
Selye H 1976 *The Stress of Life* (rev. edn.). McGraw-Hill, New York
Thoits P A 1983 Dimensions of life events that influence psychological distress: An evaluation and synthesis of the literature. In: Kaplan H B (ed.) *Psychosocial Stress: Trends in Theory and Research*. Academic Press, New York, pp. 33–103
Zeidner M, Endler N S (eds.) 1996 *Handbook of Coping: Theory, Research, Applications*. Wiley, New York

H. W. Krohne

Stress and Health Research

1. Introduction

1.1 The Origin of Stress Research

Stress is a term that is often related to a psychological or physical state of health (cf. Sapolsky 1994) and is used by many psychological, sociological, and medical scientists as well as laymen. A first definition of the term stress was proposed by Cannon (1935). He postulated that passing a critical stress level, triggered by physical and emotional stimuli, endangers the homeostatic control of an organism. However, the name that is most closely associated with stress is Hans Selye, who defined stress as a nonspecific response of the body to any demand characterized by the secretion of glucocorticoids. Selye developed the concept of the *general adaptation syndrome* (*GAS*), which encompasses the alarm reaction, the stage of resistance, and finally the stage of exhaustion. The bodily responses to massive or ongoing stress are summarized in the *stress triad*, including (a) the enlargement and hyperactivity of the adrenal cortex, (b) the shrinking or atrophy of the thymus, spleen, lymph nodes, and the lymphatic system, and (c) the appearance of gastrointestinal and bowel ulcers (Selye 1936, 1937). Regarding health consequences, Selye argued that stress plays a role in the development of every disease. According to Selye, there is always a particularly weak organ or system (due to heredity or external conditions) that is thus likely to break down under stress, which is why he concluded that individuals can develop different types of diseases under the influence of the same kind of stressor.

In the concept of *four basic variations of stress*, Selye (1983) emphasized that stress can be based on overstress (hyperstress) as well as understress (hypostress). He also contrasts harmful, damaging stress (distress) from good stress (eustress).

Selye's idea of an unspecific stress response to all kinds of stimuli was challenged by Mason (1968, 1975), who underlines the importance of specific emotional reactions that determine a specific endocrine stress response. Mason showed that specific situational characteristics, such as novelty, uncontrollability, unpredictability, ambiguity, anticipation of negative consequences, and high ego-involvement, lead to specific hormonal stress responses.

1.2 Definitions of Stress

To date, a large number of definitions of stress are available (see *Stress and Coping Theories*). Some researchers have even proposed to drop the term stress. Levine and Ursin (1991) have offered a comprehensive definition of stress, distinguishing between (a) input (stress stimuli), (b) (individual) processing, and (c) outcome (stress reaction). In this concept, those stimuli that require processing are defined as loads. An individual appraisal process determines whether a load becomes a stressor. Differences in processing are based for example on genetic, ontogenetic, and social factors, and early life as well as lifelong experiences. A related idea has been introduced earlier by Lazarus and Folkman (1984). According to the cognitive-transactional model by Lazarus and Folkman, stress is experienced as a process that is initially triggered by situational demands, but then mainly by the cognitive appraisal of these demands. The characteristics of a situation (primary appraisal) are evaluated simultaneously in line with the available coping capacities or resources (secondary appraisal). This results in a cognitive appraisal of challenge, threat, or harm/loss,

and, subsequently, in emotions, coping attempts, and adaptational outcomes. This process can be cyclical, and it may lead to a number of reappraisals that change the nature of the stress episode.

At the level of stress responses, physiological, behavioral, and subjective/verbal reactions could be distinguished. Physiological responses primarily include the *sympathetic–adrenal–medullary* (*SAM*) *axis*, *hypothalamic–pituitary–adrenal* (*HPA*) *axis*, and *immune system*. Behavioral responses, among others, cover attention, arousal, and vigilance; subjective/verbal reactions are interpretations, cognitions, and emotions. The stress reactions to physical and psychological stimuli are primarily determined by the individual interpretation, but also by the social context, the social status, genetic factors, gender, developmental stage, and individual lifelong experiences.

Finally, Levine and Ursin (1991) contrast an unspecific stress response (general alarm) and a specific individual stress reaction. With this idea they offer an integrational view of Selye's (1983) and Mason's (1968) approaches to a definition of stress.

2. *The Model of Allostatic Load*

2.1 *Introduction*

While homeostatic regulation reflects the stability of constants, allostatic regulation means stability through change. Homeostatic systems (e.g., body temperature, blood oxygen, blood pH, glucose levels) must be maintained within a narrow range. That means that deviations in a homeostatic system trigger a restorative response to correct the changes. In contrast, the regulation of *allostatic* (adaptive) systems can operate in a relatively broad range of regulation. Following Sterling and Eyer (1988), allostasis is the regulation of the internal milieu through dynamic change in hormonal and physical parameters. Allostasis is defined as the ability of the body to increase or decrease vital functions to a new steady state on challenge.

McEwen and Stellar (1993) concluded that the concept of homeostasis is not helpful for explaining the hidden costs of chronic stress on the organism. Therefore, they extended the concept of allostasis over the dimension of time and introduced the idea of *allostatic load*. Allostatic load is defined as the cost of chronic exposure to elevated or fluctuating endocrine or neural responses resulting from chronic or repeated challenges that the individual experiences as stressful. Stress, for instance psychological demands, physical threat/danger, or adverse life experiences, activates various adaptive (allostatic) systems, initiating adaptation and coping processes. The SAM and the HPA axes are the main endocrine stress systems of the organism involved in an allostatic response.

In the face of an internal or external challenge, these systems are activated. (a) Catecholamines are released from the sympathetic nerves and the adrenal medulla. (b) Corticotropin-releasing factor (CRF) from the nucleus paraventriculares (PVN) of the hypothalamus triggers the discharge of adrenocorticotrophic hormone (ACTH) from the anterior pituitary gland. ACTH stimulates the secretion of glucocorticoids (e.g., cortisol) from the adrenal cortex. These hormones serve as mediators of adaptational processes. Besides the sympathetic nervous system and the HPA axis, the cardiovascular, metabolic, and immune system play a crucial role in allostatic processes.

In sum, allostatic systems enable the body to respond adequately to changes in social and physical conditions and therefore protect the body in the short run. In the long run, however, the bodily responses to stress can cause damage and finally promote several diseases. The organism's 'hidden toll' of short-term adaptation to challenges is described as *allostatic load*, which constitutes a wear and tear on the body from chronic overactivity or underactivity of allostatic systems.

2.2 *Types of Allostatic Load*

Four different scenarios can cause allostatic load: (1) frequent exposure to stress, (2) inability to habituate to repeated challenges, and (3) inability to terminate a stress response. In these three types, allostatic load is promoted by an organism's increased exposure to stress hormones and other allostatic mediators. Finally, (4) an inadequate allostatic response in one allostatic system could be related to an increased activation of another allostatic system. An inadequate response of one allostatic system could trigger an inadequate (compensatory) response in another allostatic system.

The model emphasizes that allostatic load reflects not only the impact of lifelong experiences (wear and tear), but also covers early life experiences, genetic predispositions, environmental factors, as well as psychological and behavioral parameters. The cascading relationships between these factors could explain individual differences in the susceptibility to stress and, in some cases, to disease.

For example, Meaney and co-workers (1994) showed in animal studies that early life experiences can calibrate the lifelong pattern of physiological stress responses. Whereas unpredictable stress in newborn rodents led to an overactivity of HPA and SAM axis responsiveness, neonatal handling resulted in a reduced reactivity of the stress systems for the entire lifespan. Concerning the impact of genes on HPA axis functioning, Wüst et al. (2000) report on a medium-sized yet distinct genetic influence on cortisol levels after awakening. The role of genes in the regulation of stress systems is discussed in detail by Koch and Stratakis (2000). Psychological factors, such as an-

ticipating negative consequences, pessimism, anxiety, or worry also contribute to allostatic load. While the origin of allostatic load can be based on an individual psychological appraisal process, psychological factors can prolong, intensify, expand, or aggravate the amount of existing allostatic load. Finally, there seems to be a bidirectional influence of stress and lifestyle factors, for example nutrition, physical exercise, sleeping habits, alcohol, caffeine, and nicotine consumption, with significant effects on the extent of allostatic load (cf. Steptoe 2000).

3. *Stress and Health*

3.1 *Types of Allostatic Load and Health*

The four different types of allostatic load may possibly be related to increased risks of the development of different health problems. Repeated ups and downs of physiological responses or a heightened activity of physiological systems due to internal or external challenges result in allostatic load, which could predispose the organism to disease. There is evidence that acute and chronic stress are significant risk factors for cardiovascular, immunological, and psychological health problems as well as impairments in the central nervous system (CNS). However, it has to be taken into consideration that evidence for stress effects on health are mainly based on correlational analyses.

Animal and human data point to the idea that allostatic load of type 1 and type 3 may possibly be related to the risk of myocardial infarction. Repeated blood pressure elevation (allostatic load type 1), for instance, could promote atherosclerosis. The failure to recover from blood pressure increase following an acute stressor (allostatic load type 3) may lead to hypertension, which also increases the risk of atherosclerosis. Furthermore, heightened cortisol levels promote an increase in insulin secretion, resulting in an endocrine condition that may also accelerate atherosclerosis. Atherosclerosis itself appears to increase the risk of myocardial infarction.

Allostatic load type 2 covers the failure to habituate or the extent of habituation to repeated stress exposure. After repeated speaking and mentally solving arithmetic problems in front of an audience (Trier Social Stress Test), most participants showed habituation in their cortisol response, but about one-third of them still had high cortisol increases every time they were exposed to the stressor. The relationship between the inability of the HPA axis to habituate to repeated exposure to this psychosocial laboratory stress protocol and various health parameters is currently being investigated.

Chronically elevated SAM and HPA axis activity can lead to decreased bone mineral density, body weight loss, amenorrhea and/or anorexia nervosa, as observed in depressed patients or high-performance athletes (allostatic load type 3). Concerning bone mineral density, Michelson et al. (1996) showed that moderately elevated glucocorticoid levels over a prolonged period were associated with reduced levels of osteocalcin in depressed women, therefore inhibiting bone formation (allostatic load type 3).

Allostatic load type 4 is illustrated by observing that a hyporeactive HPA axis (decreased release of glucocorticoids after stress) could lead to an increased response of the immune system. In case one system does not show an adequate stress response, another system increases its activity because the underactive system does not provide the normal counter-regulatory control. For instance, lewis rats, which are known to have an HPA axis hyporeactivity, show increased susceptibility for autoimmune and inflammatory disturbances, and those rats which developed an HPA axis hyporeactivity due to a subordinate position in their social network are believed to die earlier than their counterparts (allostatic load type 4) (cf. McEwen 2000).

3.2 *Stress and the Cardiovascular System*

Animal and human studies indicate that a relationship exists between stress and cardiovascular disturbances. Stress plays a crucial role in cardiovascular diseases as a risk factor in the progressive development of hypertension and atherosclerosis and as a trigger of acute incidences, such as myocardial infarction, stroke, or even death. While stress provokes cardiovascular and endocrine changes that could be beneficial even in healthy human beings, the same stimuli can be detrimental to patients with coronary artery disease (see *Stress and Cardiac Response*).

Manuck et al. (1988) observed that socially subordinate female primates and dominant male primates in unstable social hierarchies have an increased risk of developing atherosclerosis. In humans, occupational strain, for example, jobs with high psychological demands and low control, leads to elevated blood pressure, increased progression of atherosclerosis, and increased risk of coronary heart disease (CHD) (cf. McEwen 2000). Allen and colleagues (1993) concluded that women are more likely to be cardiac reactors, whereas men are more prone to be vascular reactors, suggesting a gender-specific allostatic load within the cardiovascular system.

3.3 *Stress and the Immune System*

A large amount of literature shows the impact of stress on changes in immune system functioning (cf. Biondi 2001). Chronic psychological stress or increased allostatic load is generally shown to be associated with decreased immune responses as well as enhanced severity of infections. Consequently, a stress-related suppression of cellular immune system functioning can, for instance, lead to increased susceptibility to or

severity of the common cold (Cohen et al. 1991). While chronic stress seems to precede blunted immune response, acute stress may be potentially beneficial: experiencing acute stress can enhance the trafficking of immune cells to the site of acute challenge, resulting in an immune-enhancing effect over several days (cf. McEwen 2000). Acute stress may call immune cells 'to their battle stations,' thus enhancing short-term immunity (cf. Dhabar 2000, Dhabar and McEwen 1997).

Possible mechanisms involved in the relationship between psychological stress and immune function changes, covering the stress–disease connection, are discussed in Biondi (2001). There seems to be a complex connection between emotional stress and the development of cancer. In sum, human data on psychological factors and cancer are controversial, but apparently there is support that psychological stress may contribute to disease manifestation and progression. Concerning infectious diseases, some studies have reported negative findings. However, human data point to a possible role that stress plays in increasing the risk of tuberculosis, influenza, upper respiratory infections in children and adults, as well as recurrences of herpes labialis virus and genital herpes virus, or HIV progression. Biondi (2001) summarizes that psychological stress could act as a cofactor in the pathogenesis, course, and outcome of infectious diseases in animals and humans. The exact role of psychological factors in the pathogenesis and course of specific autoimmune diseases (e.g., rheumatoid arthritis, systemic lupus erythematosis, multiple sclerosis, ulcerative colitis, Crohn's disease, joint inflammation, psoriasis) is still poorly understood. In view of the clinical complexity of these illnesses, a simple connection between psychological stress and the status of an autoimmune disease cannot be presumed. For asthma bronchiale, there appears to be some support for the assumption that this illness could be exacerbated by psychological stress via immune–inflammatory mechanisms. Finally, the process of wound healing appears to be impaired under chronic as well as minor psychological stress, a process which may be mediated by proinflammatory cytokines (Glaser et al. 1999).

3.4 Stress and the Central Nervous System

Besides numerous peripheral effects of stress, stress-induced secretion of adrenal steroids can also affect the central nervous system (CNS). For example, elevated glucocorticoid levels can impair hippocampal and temporal lobe functioning, brain areas which are involved in declarative, episodic, and short-term memory. While acute stress can profoundly impair memory (effects reversible and relatively short-lived; Kirschbaum et al. 1996), repeated stress can end up in atrophy of the dendrites of pyramidal neurons in the CA3 region of the hippocampus, causing long-term cognitive impairments. Besides its role in memory and cognition, the hippocampus is most important for the negative feedback regulation of the HPA axis. Allostatic load during a lifetime may cause loss of resilience of HPA axis function, resulting in an attenuation of the negative feedback signal that may lead to hyperactivity of this stress system. The connection between the wear and tear on this region of the brain and the dysregulation of the HPA axis is elaborated in detail in the *glucocorticoid cascade hypothesis* (Sapolsky et al. 1986).

Animal data show that in aged rats, impairments of declarative, episodic, and spatial memory are associated with HPA axis hyperactivity. In elderly humans, Lupien et al. (1998) observed that hippocampal atrophy and cognitive impairment were related to progressive elevations of cortisol levels over a four-year interval.

In sum, animal as well as human data support the idea that elevated glucocorticoid levels could lead to a reduction in hippocampal functioning. Consequences may include the dysregulation of the HPA axis as well as cognitive impairments.

3.5 Stress and Psychological Diseases

Stress is said to be one important parameter, among others, in the manifestation or aggravation of several psychological disturbances, psychosomatic illnesses, and psychiatric disorders, including burnout, anorexia nervosa, chronic fatigue syndrome (CFS), asthma, neurodermatitis, major depression, schizophrenia, or post-traumatic stress disorder (PTSD), which are sometimes referred to as stress-related disorders. The notion that stress contributes somehow to onset or progression of these disorders does not indicate how strong the influence of stress is compared to other important contributing factors, such as genetic predisposition or organic diseases.

However, changes in the regulation of the HPA axis can be related to several stress-related disorders. For example, chronic stress, major depression, schizophrenia, or anorexia nervosa seem to be associated with a *hyper*active HPA axis, whereas a *hypo*reactive HPA axis is found more often in CFS, autoimmune processes, or PTSD. Since there are only few longitudinal data available to date on the relationship between HPA activity and stress-related disorders, the question of cause and effect remains to be answered in future research.

4. Gender Differences in Stress Responses and Health

Gender-specific prevalence rates reveal that men and women are at differential risk for a number of illnesses (see *Gender and Physical Health*). Whereas women suffer

more often from autoimmune diseases (e.g., multiple sclerosis or systemic lupus erythematosus), men are more prone to develop coronary heart diseases and appear to be more susceptible to infectious diseases. Also, the incidence of psychiatric disorders is not equally distributed between men and women. Diagnoses such as major depression, anxiety, phobia, panic disorder, and obsessive compulsive disorder are more common in women, whereas men more often develop antisocial behavior, abuse substances, or commit suicide. Life expectancy is about seven years higher for women compared to men, although women consistently appear to report more physical as well as somatoform symptoms than men.

Besides this clear *gender–disease dimorphism*, gender differences also exist in subjective and biological aspects of the human stress response (cf. Kudielka et al. 2000). Concerning subjective/verbal reactions to stress, several studies (ranging from everyday hassles to posttraumatic stress reactions) show that women experience and/or report more stress than men. As to HPA axis responses to psychological stress, men showed repeatedly higher ACTH and free, *bioavailable*, cortisol responses compared to women, but similar total cortisol increases. Only in the luteal phase of the menstrual cycle did women show similar free cortisol responses to those of men (Kirschbaum et al. 1999). There is evidence that these differences are, at least partly, explained by gender differences in the availability of gonadal steroids, especially estrogens.

The finding that men have higher blood pressure changes and higher elevations in epinephrine concentrations under psychological stress parallels the observation that some women seem to be protected from CHD (see *Gender and Cardiovascular Health*). Estrogen availability is discussed as being a preventive agent for the onset or progression of CHD, but the use of hormone replacement therapy (HRT) for prevention of heart disease is more complex than initially believed. The heart and estrogen/progestin replacement study (HERS) was a first large randomized, blind, placebo-controlled secondary prevention trial to evaluate the efficacy and safety of estrogen and progesterone replacement therapy in reducing CHD risk (cf. e.g., Hulley et al. 1998). It was conducted from January 1993 through July 1998, with a mean follow-up of 4.1 years at 20 centers. However, in sum, the results of the HERS study could not support the notion that HRT is a protective measure against CHD.

5. *Summary and Conclusion*

This article has dealt with the origin of stress research, briefly describing the basic aspects of the stress concepts of Cannon, Selye, and Mason. Then, a comprehensive definition of stress by Levine and Ursin (1991) was offered, distinguishing between stress stimuli, individual processing, and stress reactions on three different levels (physiological, behavioral, subjective/verbal).

The relationship between stress and health was elucidated using the allostatic load model introduced by McEwen and Stellar (1993): the body's cost of short-term adaptation to internal and external challenges is allostatic load, which constitutes a wear and tear on the organism. Four different types of allostatic load were presented and related to different health problems, including the cardiovascular system, the immune system, and the CNS. Finally, gender differences in stress responses and health outcome, including psychosomatic and psychiatric diseases, were briefly discussed, with the conclusion that the gender–disease dimorphism could be associated with gender-specific stress responses.

See also: Emotional Inhibition and Health; Environmental Stress and Health; Gender and Cardiovascular Health; Gender and Physical Health; Gender Role Stress and Health; Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of; Job Stress, Coping with; Posttraumatic Stress Disorder; Social Support and Stress; Stress and Cardiac Response; Stress and Coping Theories; Stress at Work

Bibliography

Allen M T, Stoney C M, Owens J F, Matthews K A 1993 Hemodynamic adjustments to laboratory stress: The influence of gender and personality. *Psychosomatic Medicine* **55**: 505–17

Biondi M 2001 Effects of stress on immune functions: An overview. In: Ader R, Felten D L, Cohen N (eds.) *Psychoneuroimmunology*, 3rd edn. Academic Press, San Diego, CA, Vol. 2, pp. 189–226

Cannon W B 1935 Stresses and strains of homeostasis. *American Journal of the Medical Sciences* **189**(1): 1–14

Cohen S, Tyrrell D A, Smith A P 1991 Psychological stress and susceptibility to the common cold. *New England Journal of Medicine* **325**: 606–12

Dhabar F S 2000 Stress-induced enhancement of immune function. In: Fink G (ed.) *Encyclopedia of Stress*. Academic Press, San Diego, CA, Vol. 2, pp. 515–23

Dhabar F S, McEwen B S 1997 Acute stress enhances while chronic stress suppresses immune function in vivo: A potential role for leukocyte trafficking. *Brain Behavior and Immunity* **11**: 286–306

Glaser R, Kiecolt-Glaser J K, Marucha P T, MacCullum R C, Laskowski B F, Malarkey W B 1999 Stress-related changes in proinflammatory cytokine production in wounds. *Archives of General Psychiatry* **56**: 450–6

Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, Vittinghoff E 1998 Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin replacement study/HERS) research group. *Journal of the American Medical Association* **280**(7): 605–13

Kirschbaum C, Kudielka B M, Gaab J, Schommer N C, Rohleder R, Hellhammer D H 1999 Impact of gender,

menstrual cycle phase and oral contraceptives on the activity of the hypothalamic–pituitary–adrenal axis. *Psychosomatic Medicine* **61**(2): 154–62
Kirschbaum C, Wolf O T, May M, Wippich W, Hellhammer D H 1996 Stress- and treatment-induced elevations of cortisol levels associated with impaired declarative memory in healthy adults. *Life Sciences* **58**(17): 1475–83
Koch C A, Stratakis C A 2000 Genetic factors and stress. In: Fink G (ed.) *Encyclopedia of Stress*. Academic Press, San Diego, CA, Vol. 2, pp. 205–12
Kudielka B M, Hellhammer D H, Kirschbaum C 2000 Sex differences in human stress response. In: Fink G (ed.) *Encyclopedia of Stress*. Academic Press, San Diego, CA, Vol. 3, pp. 424–9
Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York
Levine S, Ursin H 1991 What is stress? In: Brown M R, Koob G F, Rivier C (eds.) *Stress—Neurobiology and Neuroendocrinology*. Marcel Dekker, New York, Chap. 1
Lupien S J, de Leon M, de Santi S, Convit A, Tarshish C, Nair N P, McEwen B S, Hauger R L, Meaney M J 1998 Cortisol levels during human aging predict hippocampal atrophy and memory deficits [published erratum appears in *Nature Neuroscience* 1998, **1**(4): 329]. *Nature Neuroscience* **1**: 69–73
Manuck S B, Kaplan J R, Adams M R, Clarkson T B 1988 Studies of psychosocial influences on coronary artery atherogenesis in cynomolgus monkeys. *Health Psychology* **7**: 113–24
Mason J W 1968 A review of psychoendocrine research on the sympathetic–adrenal medullary system. *Psychosomatic Medicine* **30**: 631–53
Mason J W 1975 A historical view of the stress field. *Journal of Human Stress* **1**: 6–12
McEwen B S 2000 Allostasis and allostatic load. In: Fink G (ed.) *Encyclopedia of Stress*. Academic Press, San Diego, CA, Vol. 1, pp. 145–50
McEwen B S, Stellar E 1993 Stress and the individual. *Archives of Internal Medicine* **153**: 2093–101
Meaney M J, Tannenbaum B, Francis D, Bhatnagar S, Shanks N, Viau V, O'Donnell D, Plotsky P M 1994 Early environmental programming hypothalamic pituitary–adrenal responses to stress. *Seminars in Neuroscience* **6**: 247–59
Michelson D, Stratakis S, Hill L, Reynolds J, Galliven E, Chrousos G, Gold P 1996 Bone mineral density in women with depression. *New England Journal of Medicine* **335**: 1176–81
Sapolsky R M 1994 *Why Zebras Don't Get Ulcers*. W H Freeman, New York
Sapolsky R M, Krey L C, McEwen B S 1986 The neuroendocrinology of stress and aging: The glucocorticoid cascade hypothesis. *Endocrine Reviews* **7**: 284–301
Selye H 1936 A syndrome produced by diverse nocuous agents. *Nature* 32
Selye H 1937 Studies on adaptation. *Endocrinology* **21**(2): 169–88
Selye H 1983 The stress concept: past, present, and future. In: Cooper C L (ed.) *Stress Research*. Wiley, Chichester, UK, pp. 1–20
Steptoe A 2000 Health behavior and stress. In: Fink G (ed.) *Encyclopedia of Stress*. Academic Press, San Diego, CA, Vol. 2, pp. 322–6
Sterling P, Eyer J 1988 Allostasis: A new paradigm to explain arousal pathology. In: Fisher S, Reason H S (eds.) *Handbook of Life Stress, Cognition and Health*. Wiley, New York, pp. 629–49
Wüst S, Federenko I, Hellhammer D H, Kirschbaum C 2000 Genetic factors, perceived chronic stress, and the free cortisol response of awakening. *Psychoneuroendocrinology* **25**: 707–20

B. M. Kudielka and C. Kirschbaum

Stress at Work

1. The Changing Nature of Work and the Role of Stress

The nature of work has changed considerably over the past several decades in economically advanced societies. Industrial mass production no longer dominates the labor market. This is due, in part, to technological progress, in part to a growing number of jobs available in the service sector. Many jobs are confined to information processing, controlling, and coordination. Sedentary rather than physically strenuous work is becoming more and more dominant. New management techniques, including quality management, are introduced, and economic constraints produce work pressure, rationalization, and cut-down in personnel. These changes go along with changes in the structure of the labor market. More employees are likely to work on temporary contracts, on fixed term, or in flexible job arrangements. The workforce is getting older, and an increasing proportion of women enters the labor market, with an increase of double exposure among women with children or, more generally, in dual career families. Most importantly, overemployment in some segments of the workforce is paralleled by underemployment, job instability, or structural unemployment in other segments. This latter trend currently hits the economically most advanced societies as well as economies that lag behind. Overall, a substantial part of the economically active population is confined to insecure jobs, to premature retirement, or job loss.

Why is work so important for human well-being, and how does work contribute to the burden of stress and its adverse effects on health? In all advanced societies work and occupation in adult life are accorded primacy for the following reasons. First, having a job is a principal prerequisite for continuous income and, thus, for independence from traditional support systems (family, community welfare, etc.). Increasingly, level of income determines a wide range of life chances. Second, training for a job and achievement of occupational status are the most important goals of socialization. It is through education, job training, and status acquisition that personal growth and development are realized, that a core social identity outside the family is acquired, and that

goal-directed activity in human life is shaped. Third, occupation defines an important criterion of social stratification. Amount of esteem in interpersonal life largely depends on type of job and level of occupational achievement. Furthermore, type and quality of occupation, and especially the degree of self-direction at work, strongly influence personal attitudes and behavioral patterns in areas that are not directly related to work, such as leisure or family life (Kohn and Schooler 1983). Finally, occupational settings produce the most pervasive continuous demands during one's lifetime, and they absorb the largest amount of active time in adult life, thus providing a source of recurrent negative or positive emotions. It is for these reasons that stress research in organizations where paid work takes place is of particular relevance.

It is important to recognize that traditional occupational hazards, such as exposure to toxic substances, heat, cold, or noise, are no longer the dominant challenges of health at work. Rather, distinct psychological and emotional demands and threats are becoming highly prevalent in modern working life. There has been a recognition that the importance of work goes beyond traditional occupational diseases and, indeed, it is likely that work makes a greater contribution to diseases not thought of as 'occupational' in conventional terms. At a descriptive level, recent reports find that, for instance, almost half of the workforce are exposed to monotonous tasks or lack of task rotation; 50 percent work at a very high speed or to tight deadlines. Thus, over- and underload at work are highly prevalent. Every third employee has no influence on work rhythm, and every fifth is exposed to shiftwork (Paoli 1997). There is now growing awareness among all parties of the labor market that psychosocial stress at work produces considerable costs, most importantly a high level of absenteeism, reduced productivity, compensation claims, health insurance, and direct medical expenses. Permanent disability and loss of productive life years due to premature death add to this burden. In the European Union, the costs of stress to organizations and countries are estimated to be between 5 and 10 percent of GNP per annum (Cooper 1998). At the same time scientific evidence on associations between psychosocial stress at work and health is growing rapidly.

2. *Concepts of Work-related Stress and Selected Findings*

Research on psychosocial work-related stress differs from traditional biomedical occupational health research by the fact that stressors cannot be identified by direct physical or chemical measurements. Rather, theoretical models are needed to analyze the particular nature of the psychosocial work environment. A theoretical model is best understood as a heuristic device that selectively reduces complex reality to meaningful components. Components are meaningful to the extent that they provide the material from which the researcher can deduce explanations and, thus, produce new knowledge. Ideally, a theoretical model of psychosocial stress at work with relevance to health should encompass a wide variety of different occupations and should not be restricted to a specific time or space.

Before describing some prominent theoretical models of occupational stress, the basic terminology needs to be clarified. Much critique has been raised against the ambiguity of the term 'stress.' To avoid this, the following terms were suggested: 'stressor' is defined as an environmental demand or threat that taxes or exceeds a person's ability to meet the challenge; 'strain' is defined as the person's response to such a situation in psychological and physiological terms. Psychological responses include specific cognitions and negative emotions (e.g., anger, frustration, anxiety, helplessness), whereas physiological responses concern the activation of the autonomic nervous system and related neuro-hormonal and immune reactions. Stressors, in particular novel or dangerous ones, are appraised and evaluated by the person, and as long as there is some perception of agency on the part of the exposed person efforts are mobilized to reverse the threat or to meet the demands. Such efforts are termed 'coping,' and they occur at the behavioral (even interpersonal), cognitive, affective, and motivational level. Clearly, when judging strain, the quality and intensity of a stressor as well as the duration of exposure have to be taken into account, as well as individual differences in coping and in vulnerability to strain reactions. Recent research indicates that only part of human strain reactions are subject to conscious information processing, whereas a large amount bypasses awareness. The term 'stressful experience' is introduced to delineate that part of affective processing that reaches consciousness. Stressful experiences at work are often attributed to adverse working conditions by exposed people themselves. While they usually refer to some common sense notions of stress, it is crucial to note that these attributions differ from the explanatory constructs of stressful experience at work that have been identified by science.

2.1 *Person–Environment Fit*

The first significant contribution to modern occupational stress research dates back to studies conducted more than 30 years ago at the Institute for Social Research at the University of Michigan. They were guided by a theoretical concept termed 'person–environment fit' (Caplan et al. 1980). Stressful experience at work, in this model, is conceptualized in two ways: first, as an experience where the work

environment does not provide adequate supplies to meet the person's needs; second, as an experience where the abilities of the person fall short of demands that are prerequisite to receiving supplies. In both conditions stressful experience results from a misfit between needs or abilities of the working person and demands or opportunities of the work environment. Further job opportunities may fail to fulfill the person's needs as a consequence of unmet demands. As an example, a worker with limited skills is excluded from any promotion prospects, although, to meet his level of living, he badly needs better pay.

This approach represents a type of theorizing that is termed process theory. As such, it is applicable to a wide range of work but does not specify the particular content dimensions on which person and environment should be examined (Edwards et al. 1998). As its basic terms are rather broad, it may be difficult to reach a consensus on how to measure core constructs. Nevertheless, person–environment fit offers an elaborate set of propositions of how critical constellations of work-related environmental and personal characteristics contribute to the development of stressful experience. At the empirical level, most up-to-date studies have tested adverse effects on subjective health. Therefore, it is not quite evident to what extent these types of misfit are capable of eliciting sustained physiological strain and thus contribute to the development of bodily disease.

2.2 Demand, Support, and Control

The argument of how strain reactions affect physiological regulation has received special attention in a second influential contribution to current occupational stress research that focuses on the theoretical construct of control. Animal stress research and psycho-physiological experiments in humans confirmed the critical impact on physiological and psychological strain reactions produced by lack of control in a challenging situation (Frankenhaeuser 1979, Henry and Stephens 1977). Control is defined as the ability of a person to choose their own actions to cope with a challenge. In the workplace it ranges from task autonomy to participation in decision making, and strain evolves from challenging situations at work that leave the person with limited choice on how to react. This idea was elaborated in terms of the demand–control model of occupational stress by Karasek (Karasek and Theorell 1990, see *Occupational Health*). This model posits that strain is contingent on the combined effects of high demand and low control; though both high demands and low control may also have independent additive effects on strain and health.

Stress research based on the concept of control resulted in several important discoveries. First, it was shown that physiological strain reactions are not uniform, but may vary according to the degree of perceived control. For instance, stressful experience in terms of striving without success was found to elicit strain reactions that exert particularly harmful effects on the cardiovascular system (Kaplan et al. 1983). Second, epidemiologic studies on work-related control and health, most importantly in the framework of the demand–control model, help explaining the social distribution of some major chronic diseases, especially coronary heart disease (CHD). An inverse social gradient of CHD has been documented. Exposure to low control jobs is more frequent among lower status groups as is the prevalence of CHD. In fact, recent studies found evidence that a considerable part of the social variation in CHD is attributed to adverse psychosocial work conditions (Marmot et al. 1999). Finally, research on control at work has put special emphasis on coping resources, in particular the buffering effects produced by social support. Strain-reducing effects of social support are due to positive experience resulting from social relationships at the instrumental, emotional, and cognitive level (House 1981). This evidence has implications both for a broader conceptualization of occupational stress (e.g., the revised demand–support–control model (Karasek and Theorell 1990)) and for the design of workplace intervention measures.

2.3 Effort and Reward

In physiological terms, theories of control at work deal with the experience of power and influence which, at the psychological level, is paralleled by feelings of autonomy and self-efficacy. An alternative model of stressful experience at work, the model of effort–reward imbalance, is concerned with distributive justice, that is, with deviations from a basic 'grammar' of social exchange rooted in the notions of reciprocity and fairness (Siegrist 1996). This model assumes that effort at work is spent as part of a socially organized exchange process to which society at large contributes in terms of rewards. Rewards are distributed by three transmitter systems: money, esteem, and career opportunities including job security. The model of effort–reward imbalance claims that lack of reciprocity between costs and gains (i.e., high 'cost'/low 'gain' conditions) elicits sustained strain reactions. For instance, having a demanding, but unstable job, achieving at a high level without being offered any promotion prospects, are examples of high cost/low gain conditions at work. In terms of current developments of the labor market in a global economy, the emphasis on occupational rewards including job security reflects the growing importance of fragmented job careers, of job instability, underemployment, redundancy, and forced occupational mobility including their financial consequences.

According to this model, strain reactions are most intense and long-lasting under the following conditions: (a) lack of alternative choice in the labor market may prevent people from giving up even unfavorable jobs, as the anticipated costs of disengagement (e.g., the risk of being laid-off) outweigh the costs of accepting inadequate benefits; (b) unfair job arrangements may be accepted for a certain period of one's occupational trajectory for strategic reasons; by doing so employees tend to improve their chances for career promotion at a later stage; (c) a specific personal pattern of coping with demands and of eliciting rewards characterized by overcommitment may prevent people from accurately assessing cost–gain relations. 'Overcommitment' defines a set of attitudes, behaviors, and emotions reflecting excessive striving in combination with a strong desire for being approved and esteemed (Siegrist 1996). At the psychological level, experience of effort–reward imbalance is often paralleled by feelings of impaired self-worth or self-esteem.

2.4 Empirical Evidence

Evidence of impaired health due to exposure to stressors that are defined by the concepts described is growing rapidly. In particular, a number of well-designed prospective social-epidemiological studies found moderately elevated relative risks (two- to threefold) of incident CHD in economically active middle-aged men and women who were suffering from high demand/low control or effort–reward imbalance. Other health effects observed in prospective and cross-sectional studies concern distinct cardiovascular risk factors, gastrointestinal symptoms, mild psychiatric disorders, poor subjective health, and musculo-skeletal disorders. In addition, rates of absenteeism are elevated (Marmot et al. 1999). There is also evidence of changes in physiological patterns, release of stress hormones, and altered immune competence following stressful experience at work (see *Occupational Health*). Taken together, the burden of illness that can be attributed, at least in part, to exposure to an adverse psychosocial work environment turns out to be impressive. While further research is needed to address important issues, practical implications of this new evidence deserve special attention.

3. Practical Implications and Future Directions

Overall, worksite health promotion activities are realized with little concern about scientific advances obtained in the field of occupational stress research. Eventually, this may change in the future as all three models described in the last paragraph (as well as additional models not mentioned in detail, for summary see Cooper 1998) provide a rationale for evidence-based preventive activity. Person–environment fit theory concludes that work demands need to be tailored according to the worker's abilities. This involves both specific measures of organizational development and person development. Moreover, organizational interventions must suit the needs of affected individuals that are closely linked to the work role (e.g., motivation, learning, satisfaction; Edwards et al. 1998). In the demand–support–control model, special emphasis is put on autonomy, skill discretion, and personal growth. Measures of work reorganization include job enlargement, job enrichment, skill training, enhanced participation, and teamwork (Karasek and Theorell 1990). Practical implications of the effort–reward imbalance model concern the development of compensatory wage systems, the provision of models of gain sharing, and the strengthening of nonmonetary gratifications. Moreover, ways of improving promotional opportunities and job security need to be explored. Supplementary measures are interpersonal training and skill development, in particular leadership behavior.

Clearly, the power of economic life and the constraints of globalization limit the options and range of worksite health promotion measures. Moreover, such measures are difficult to apply to medium-sized and small enterprises where the majority of the workforce in advanced societies is employed. Finally, it must be stated that many instrumental activities that qualify as work are no longer confined to the conventional type of workplace located within an enterprise or organization, particularly with the advent of advanced communications technology (Cooper 1998).

3.1 Future Challenges for Research

There are at least three obvious future directions for occupational stress research. First, the theoretical concepts of stressful experience at work discussed need to be examined in a comparative perspective, with emphasis on their potential overlap or cumulative effects. For instance, recent studies found independent statistical prediction of the demand–control and the effort–reward imbalance model when tested simultaneously (Marmot et al. 1999). Moreover, the estimation of health risk was considerably improved when based on combined information derived from both models.

A second direction of research addresses the link or interface between work and nonwork settings, in particular work and family. In advanced societies work and family have been identified as the two most central domains, especially so in midlife. Therefore, stress-enhancing and stress-buffering experiences in these two domains significantly contribute to personal health and well-being. This interface is conceptualized in terms of spillover, compensation, or cumulation. Spillover refers to the transfer of experience and behavior from one domain to the other. For instance,

work overload or shiftwork schedules may have negative effects on family life. Alternatively, experience of competence and success at work can have beneficial effects on marital life and relationships between parents and children (Kohn and Schooler 1983). Compensation represents efforts to reduce stressful experience in one domain by improving satisfaction in another domain. For instance, leisure and family activities are enhanced to the extent that decreasing involvement in a stressful work environment is feasible. Compensation often occurs in terms of resource management in a stressful situation. In conceptual terms, there is some overlap between the buffering model of social support and compensation processes (House 1981). Cumulation refers to the fact that the critical components of stressful psychosocial experience, such as lack of control or effort–reward imbalance, generalize across core life domains. For instance, disease vulnerability may be highest among people with insecure, badly paid, effortful jobs who also live in a family arrangement that offers little emotional and material satisfaction. It is desirable to extend the theoretical models discussed, both conceptually and operationally, to life domains other than work.

A third direction of research concerns intervention studies. The implementation and evaluation of theory-based intervention is a promising area for creating new knowledge although it is admitted that established methodological principles of scientific trials cannot be fully met. Recent theory-based intervention studies have documented favorable health effects (Cooper 1998, Karasek and Theorell 1990, Siegrist 1996, see *Occupational Health*). In view of the importance of work for health, it is well justified to advance the field of occupational stress both at the science and policy level.

See also: Environmental Stress and Health; Job Stress, Coping with; Occupational Health; Performance Evaluation in Work Settings; Psychological Climate in the Work Setting; Social Support and Stress; Stress and Cardiac Response; Stress and Coping Theories; Stress and Health Research; Stress in Organizations, Psychology of; Stress Management Programs; Stress: Psychological Perspectives; Well-being and Burnout in the Workplace, Psychology of; Worksite Health Promotion and Wellness Programs

Bibliography

Caplan R D, Cobb S, French J R P, Harrison R V, Pinneau S R 1980 *Job Demand and Worker Health: Main Effects and Occupational Differences*. Institute for Social Research, Ann Arbor, MI

Cooper C L (ed.) 1998 *Theories of Organizational Stress*. Oxford University Press, Oxford, UK

Edwards J R, Caplan R D, van Harrison R 1998 Person–environment fit. In: Cooper C (ed.) *Theories of Organizational Stress*. Oxford University Press, Oxford, UK, pp. 28–67

Frankenhaeuser M 1979 Psycho-neuroendocrine approaches to the study of emotions as related to stress and coping. In: Howe H E, Diensbier R A (eds.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, pp. 123–61

Henry J P, Stephens P M 1977 *Stress, Health, and the Social Environment*. Springer, Berlin

House J 1981 *Work, Stress and Social Support*. Addison-Wesley, Reading, MA

Kaplan J R, Manuck S B, Clarkson T B 1983 Social stress and atherosclerosis in normocholesterolemic monkeys. *Science* **220**: 733–5

Karasek R A, Theorell T 1990 *Healthy Work. Stress, Productivity, and the Reconstruction of Working Life*. Basic Books, New York

Kohn M, Schooler C 1983 *Work and Personality: An Inquiry into the Impact of Social Stratification*. Ablex, Norwood, NJ

Marmot M, Siegrist J, Theorell T, Feeney A 1999 Health and the psychosocial environment at work. In: Marmot M G, Wilkinson R (eds.) *Social Determinants of Health*. Oxford University Press, Oxford, UK, pp. 105–31

Paoli P 1997 *Working Conditions in Europe. The Second European Survey on Working Conditions*. European Foundation, Dublin

Siegrist J 1996 Adverse health effects of high effort/low reward conditions. *Journal of Occupational Health Psychology* **1**: 27–41

J. Siegrist

Stress in Organizations, Psychology of

Stress is a term widely used in the social sciences, but not well defined by them; it may refer to any noxious or unwanted stimulus, or to the effects of such stimuli. The concept of stress itself seems to have been borrowed from the fields of physics and engineering, where it is defined more precisely as any force applied to a body, and the resulting alteration or distortion is referred to as strain. This article, following those definitions, will refer to the relevant stimuli as stressors, their presence as stress, and their effects on the individual as strain.

Despite their differences in terminology and the variety of theoretical models with which they work, researchers on organizational stress agree on some or all of the following causal sequence: an objective stressor, its perception by individuals exposed to it, their appraisal of it, their immediate reactions, and its longer term effects. Models or frameworks for research along these lines have been proposed by numerous authors since the early 1960s. The major relationships specified in these models, along with some of the variables most frequently studied, are included in the

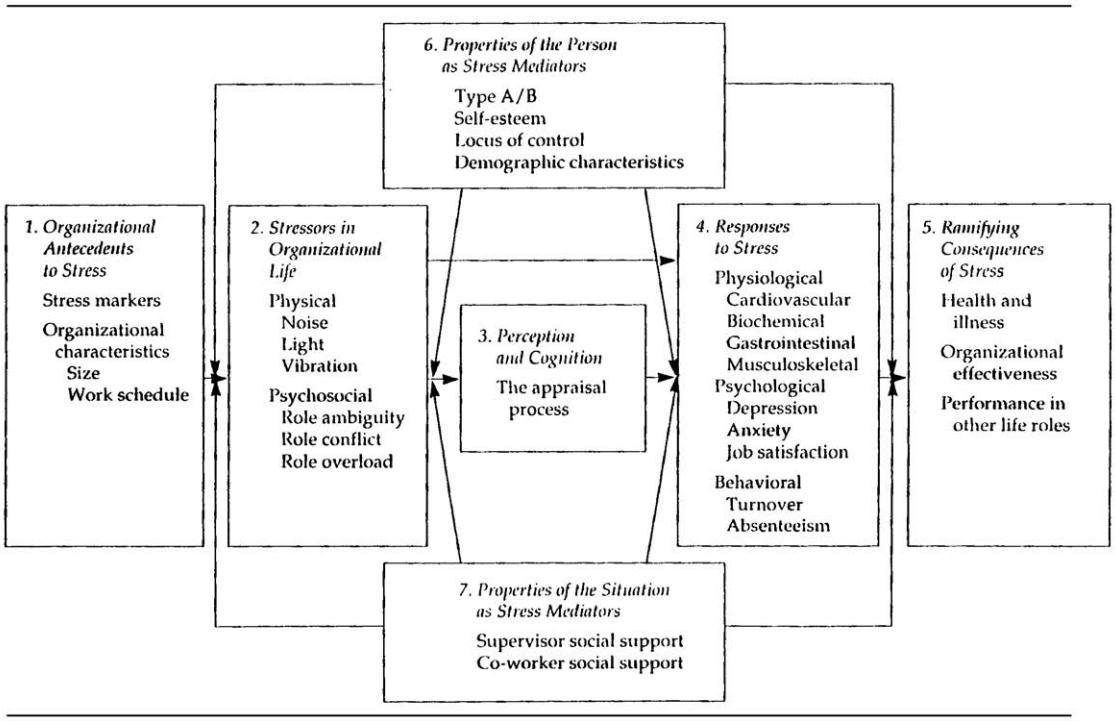

Figure 1
Theoretical framework for the study of stress in organizations (Kahn and Byosiere 1992)

'ISR model' developed by French and Kahn (1962) and elaborated by Kahn and Byosiere (1992) (for a review of stress research in organizations see Fig. 1).

1. *Early Empirical Research*

Empirical organizational studies, especially in the early years, were less comprehensive than these models suggest; most research on stress was limited to undesirable job characteristics and their effects on individual employees. Some studies, however, examined the extent to which this causal sequence is moderated by enduring characteristics of the individuals and by properties of the organization or its parts.

Still earlier, research and clinical observations of stress effects were carried out by physicians rather than psychologists. For example, medical doctors recognized angina pectoris and other diseases as the apparent result of 'stress and strain' in their patients' lives, although they did not identify working conditions as hypothetical causes. Selye (1982), who pioneered research on the physiological effects of stress, worked almost exclusively in the laboratory, with rats as experimental subjects. The stressors in his experiments, however, are suggestive of some occupational conditions. These included alarm, sustained vigilance, and physical constraint, all of which led to consequences that Selye called the *stress response* or *general adaptation syndrome* (*GAS*). The GAS included damage to the adrenal glands and the lymphatic system, development of stomach ulcers, and other indications of physiological dysfunction. Selye insisted that the GAS was evoked by any of a wide range of stressful stimuli or 'noxious agents.'

Research that was more specific to role demands in real-life settings, especially military service, was initiated by psychiatrists and psychologists during World War II (Grinker and Spiegel 1945, Janis 1951). Criterion variables in this research were mainly psychiatric symptoms. In the 1950s and 1960s, interest in occupational stress increased substantially, and experimental and field research on the subject was conducted in a number of countries. This research was stimulated by the 'human relations movement,' an unorganized but widespread concern about working conditions and their effects on employees. The emphasis on human relations, as the term implies, included psychological variables, both as causes and effects, in stress research. Centers of this early work were located at the University of Stockholm and the Karolinska Institute in Sweden, the Tavistock Institute in England, the Work Research Institute in Norway, the Institute for Social Research at the

University of Michigan, and in a number of other academic settings.

One of the early Michigan studies (Kahn et al. 1964) investigated the effects of two occupational variables, role conflict and role ambiguity. The underlying assumption was that for each person in an organization, there is a set of other people—supervisors, subordinates, and peers—whose expectations that person must meet. Role conflict for an individual occurs when the members of his or her set disagree about what that individual should do, and role ambiguity exists when the expectations themselves are unclear. In this study the effects of these two stressors, based entirely on self-report, included increased tension, internal conflict about appropriate action, and reduced satisfaction with the job, with supervisors, and with the organization as a whole.

The Swedish research, both in laboratory and organizational settings, included physiological as well as psychosocial effects of stress. Effects of stress were measured as changes in catecholamine (adrenaline and noradrenaline) excretion, creatinine clearance, and elevations in triglyceride and cholesterol levels (Levi 1972, Frankenhaeuser 1979).

2. *Recent Research Findings*

The summary of more recent research findings, with recent defined as the years since 1980, follows the sequence shown in Fig. 1, from organizational and societal characteristics that are stress-generating, to the stressors themselves, their perception and appraisal by individuals, the physical and emotional and behavioral responses they evoke, and the longer term effects on those exposed to the stressors and on others associated with them.

2.1 *Organizational Antecedents to Stress*

Job loss has been perhaps the most intensively investigated work-related stressor. Much of the research has been conducted around plant closings and the consequences for the affected workers and their families. A program to mitigate these negative effects was developed by Price and Vinokur (1995) and experimental designs to measure its effectiveness have been carried out, first in the USA and more recently in many other countries.

The psychological and physiological strains of job loss have been demonstrated beyond question at the individual level. It is plausible therefore that high levels of general unemployment would have negative effects on health. Sociologists and economists who work with variables at institutional and societal levels, however, have seldom been concerned with the measurement of stress and its effects on individuals. Brenner and his colleagues are exceptions; they have demonstrated that the health of populations is affected by changes in economic conditions at regional and national levels (Brenner and Mooney 1983). Economic growth or decline and economic instability are substantially correlated with rates of institutionalization, suicide, and all-cause mortality. These correlations, appropriately time lagged, have been replicated in nine industrialized countries. The intervening processes of stress and strain, however, are neither specified in the theory nor included in the research.

Stress-related findings are similarly sparse at the organizational level. An early correlation from a national population sample suggested that organizational size was a source of stress, presumably because it led to increased formality and rigidity. But research in The Netherlands found a curvilinear relationship between organizational size and stress, with employees in medium-sized companies reporting higher levels of both stressor presence and strain experienced. In short, research evidence indicates that organizations differ significantly in their levels of stress and strain, but does not say what structural properties explain these differences.

Somewhat more research has been done on the stress patterns associated with different positions within organizations. These studies, almost none of them replicated, suggest that span of control and location at an organizational boundary expose employees to increased stress. Data on the effects of hierarchical position are inconsistent, with some studies showing maximum stress at highest positions and other studies showing curvilinear patterns.

Occupation, which perhaps can be considered an organizational factor once removed, has been much studied for its health consequences. Most occupational studies by social scientists, however, do not deal explicitly with stress; they report outcome differences in health or other indicators of well-being and leave the intervening processes to the imagination of the reader.

2.2 *Organizational Stressors*

Job characteristics that have been identified as stressors are a miscellaneous set, some of them technological, some policy-determined, and still others interpersonal in origin. Most studied are such physical conditions as monotony and repetitiveness, shift work, the need for sustained vigilance, exposure to extremes of temperature, and ambient presence of noise and vibration. A few other job characteristics, more social psychological than physical, have also been identified as stressors; these include role conflict, in which a person is subjected to incompatible demands; role ambiguity, in which a person is not given adequate information about the requirements of the job; role overload, in which the quantitative or qualitative demands of the job exceed the person's resources or

ability; and lack of autonomy, perhaps the most frequently heard complaint of people in nonsupervisory positions.

2.3 Perception and Cognition

Some stressors, as toxicological research has demonstrated, can have effects without having been perceived by people exposed to them. Most organizational sources of stress, however, are recognized by the people they affect; their seriousness is assessed, and their physical and psychological impact is mediated by that cognitive assessment. Most research on organizational stress does not distinguish between the objective stressor and its subjective perception and evaluation by people exposed to it. Lazarus and his colleagues, however, have concentrated on this sequence of perception and cognition, which they call the appraisal process (Lazarus and Folkman 1984). There is strong experimental evidence that an individual's appraisal of a stressor and the degree of its threat to his or her well-being has effects that are independent of its objective properties. Lazarus's work has been almost exclusively experimental and not directly concerned with organizational sources of stress. Beehr and Bhagat (1985), however, have brought together those organizational studies that utilized appraisal as a predictor of strain.

2.4 Responses to Stress

Responses to organizational stress are usually classified as physiological, psychological, and behavioral. Ideally, the effects of a stressor would be measured in all three of these domains. In practice, however, the disciplinary background of researchers seems to determine the choice of criterion variables. Physicians and epidemiologists concentrate on physiological responses; psychologists are more likely to emphasize self-reports of strain and behavioral change.

2.4.1 Physiological responses. Job-related stress has been repeatedly shown to affect three kinds of physiological responses—cardiovascular, biochemical, and gastrointestinal. Cardiovascular responses to job stress include elevations in blood pressure and heart rate. Biochemical responses include elevations in cholesterol level and in catecholamines (epinephrine and norepinephrine). Gastric symptoms have been assessed primarily by self-report, sometimes with the inference of peptic ulcer. Among the stressors that evoke these symptoms are role conflict and ambiguity, unpredictability of work load and other demands, lack of control over pace and method of work, and ambient distracting noise. Repetitive motions, on jobs that require heavy lifting and on those that are limited to keyboard work, have been interpreted as causing a variety of musculoskeletal complaints, from back pain to carpal tunnel syndrome. Swedish research has been a major source of information on physiological responses to work-related stress. For summaries that include this research, see Kahn and Byosiere 1992 and Lundberg 1999.

Finally, job insecurity and actual job loss must be counted among the more serious sources of work-related stress. Studies of unemployment and job loss have been numerous for many decades, especially at times when unemployment rates were high. More recently, studies of job loss and of job insecurity have been stimulated by widespread reductions in force, usually referred to as corporate downsizing. Most such research has not included physiological measures. The early study of plant closings by Cobb and Kasl (1977) is an exception; it showed elevations in blood pressure, heart rate, cholesterol level, and catecholamine output. These symptoms occurred during anticipation of job loss and continued through the time of actual dismissal. A gradual return to prestressed levels then began and was completed as workers found new employment and adjusted to their new jobs. More recent research on the impact of job loss has included findings on the mitigating effect of an experimental program of counseling and peer support (Price and Vinokur 1995).

2.4.2 Psychological responses. Many studies of job stress include psychological responses, and they have been summarized in a number of review articles (Holt 1982, Cooper and Payne 1988, Kahn and Byosiere 1992, Lundberg 1999). As these reviews indicate, most of the psychological responses to job stress in this research are self-reported and unlinked to clinical validation or physiological measures. An exception is the work of colleagues at Stockholm University, both at the Karolinska Institute and the Department of Psychology (Stockholm University 1999).

By far the most frequently cited response to job stress is dissatisfaction with the job. More intense affective responses include anger, frustration, irritation, and hostility toward supervisors and the organization as a whole. Other negative responses correlated with job stress are more passive; among them are boredom, burnout, fatigue, feelings of helplessness and hopelessness, and depressed mood. Many of these words are virtual synonyms, a fact that reflects the tendency of researchers to design and label their own measures rather than replicate previous work. Replication in diverse organizations would be very useful, however, as would the combination of physiological and behavioral criteria with psychological responses.

2.4.3 Behavioral responses. Two behavioral correlates of job stress are well replicated: reductions in

job performance and increased absence. Several studies also reported increased turnover and increased use of tobacco. A review of these (Kahn and Byosiere 1992), along with a larger number of unreplicated findings, subsumed the behavioral responses to job stress into five broad categories:

(a) Degradation/disruption of the work role itself (job performance, accidents, errors, etc.);

(b) Aggressive behavior at work (stealing, purposeful damage, spreading rumors, etc.);

(c) Flight from the job (absenteeism, turnover, early retirement, etc.);

(d) Degradation/disruption of other life roles (marriage, friendships, citizenship);

(e) Self-damaging behaviors (drug use, alcohol use, smoking, etc.).

2.5 Mediators of Stress

The model used in this brief review shows two sets of variables that might act to mediate or buffer the effects of job stress: enduring properties of the person and properties of the work situation itself.

That people differ in their sensitivity of job stress is beyond argument, and many demographic and personality characteristics have been proposed as stress mediators. No single characteristic mediates the effects of stress in all cases, but evidence for a stress-buffering effect has been found for at least three personality traits: Type A, locus of control, and self-esteem. Type A individuals (competitive, impatient, aggressive) showed greater strain than Type B individuals working under comparable conditions. People high in internal control (belief that their own choices and decisions make a difference) show less strain than people who score high in external control (belief that their circumstances are determined by 'external' forces beyond their control). High self-esteem also seems to offer some protection from the effects of job stressors.

Few situational factors have been studied for their ability to buffer the effects of job stress. Social support is the exception; both its main effects and its buffering or interaction effects have been frequently tested. Findings are more consistent for main effects than for interactions; social support from various sources—supervisors, co-workers, family, and friends—predicts lower overall strain. Buffering effects have been frequently found as well, but the reasons for its failure in some cases are not clear. An earlier review (LaRocco et al. 1980) proposed that social support buffered (i.e., reduced) the relationship between various job stressors and indicators of mental and physical health but not between job stressors and more specific job strains (boredom, job dissatisfaction, dissatisfaction with workload).

Autonomy on the job is a second factor in the work situation that has been studied for its buffering as well as its main effects on strain. Karasek and his colleagues (1981), analyzing national survey data from the USA and Sweden, found that, among workers with heavy job demands, those who had low decision latitude (little control over pace and method) reported symptoms of physiological and psychological strain; for those who had high decision latitude, however, the relationship between job demands and strain did not hold.

3. Unresolved Issues in the Study of Stress

Research on organizational sources of stress continues and it is appropriate, therefore, to conclude by mentioning some issues that require the attention of stress researchers. They can be summarized in three categories: unresolved questions of concept and theory and of method and design.

3.1 Questions of Concept and Theory

A longstanding issue in stress theories is the distinction between stimuli that are stressful and those that are not. Stress theories must not imply that the optimum human state is absence of stimuli; to the contrary, research on sensory deprivation has demonstrated the negative effects of that state. Moreover, some stimuli are pleasant or exciting at some levels or for brief periods, but unpleasant or damaging at higher levels or longer duration. Furthermore, physical or mental demands that are experienced as stressful or challenging may have training or strengthening effects rather than effects that are weakening or damaging in other ways.

Selye (1982) attempted to distinguish between two kinds of task-related stress, eustress and distress, on the basis of the prospect for successful task completion or achievement. He did not argue that eustress avoids the long-term physiological effects of stress, but rather that to some extent these are unavoidable in life. Eustress and the psychological rewards of achievement, however, make the effort worthwhile. Levi (1972) had earlier dealt with this problem in more quantitative terms. He proposed that stress be thought of in terms of a continuum of stimulation, in which conditions of both understimulation (deprivation) and overstimulation (excess) were stressors. Somewhere between these extremes were the domains of pleasant stimulation, challenge, and training effect.

Another issue in stress research, which is as important for methodology as for concepts and theory, is the distinction between crisis events and chronic conditions. Stressful life events—job loss, bereavement, illness, and the like—have been prominent in stress research, at least since Holmes and Rahe (1967) published a scale that gave researchers a basis for rating their intensity and their combined demand as stressors. Lazarus and his colleagues (Lazarus and Folkman 1984) have emphasized the stress effects of chronic conditions, *daily hassles*, rather than the

stressful life events (*SLEs*) that characterize the work of Holmes and Rahe (1967) and of Dohrenwend and her colleagues (Dohrenwend and Dohrenwend 1984). More recent work by McEwen and his colleagues (McEwen 1998) has introduced the concept of *allostatic load*, which has something in common with the daily hassles of Lazarus. Allostatic load refers to the stress effects of sustained physical or psychological demands, even when individuals are coping with them successfully. Theories of psychological stress have yet to deal fully with these issues.

3.2 Issues of Method and Design

The main weaknesses of design in organizational research on stress are three, all of them long-standing: overreliance on self-report, on cross-sectional designs, and on single organizations. Exclusive dependence on self-reported data means that all respondents in a study describe their perceptions of some working condition and the symptoms of strain that they experience. At best, this design limitation generates data of uncertain validity as descriptors of the work situation; at worst, it is a near tautology.

The reliance on cross-sectional rather than longitudinal designs is by no means a special defect of research on job stress; it is a common weakness in social research. The criticism is similarly familiar; causal inferences from cross-sectional data are notoriously weak. The usual reliance on single-organization designs makes it impossible to bring organizational variables into the analysis of job stress.

If one could write prescriptions for research on job stress, therefore, there would be more designs that were longitudinal, that included objective measures of job characteristics, that included physiological as well as self-reported measures of strain, and that traced the effects of job stress on the performance of both work and nonwork roles. All these improvements are within reach and each of them is occasionally exemplified in current research. Progress in stress research requires that they become more general in future research.

See also: Stress: Psychological Perspectives

Bibliography

Beehr T A, Bhagat R S (eds.) 1985 *Human Stress and Cognition in Organizations*. Wiley, New York

Brenner M H, Mooney A 1983 Unemployment and health in the context of economic change. *Social Science and Medicine* **17**(16): 1125–38

Cobb S, Kasl S V 1977 *Termination: The Consequences of Job Loss*. US Department of Health, Education and Welfare. US Government Printing Office, Washington, DC

Cooper C L, Payne R 1988 *Causes, Coping and Consequences of Stress at Work*. Wiley, Chichester, UK

Dohrenwend B S, Dohrenwend B P 1984 *Stressful Life Events and their Contexts*. Rutgers University Press, New Brunswick, NJ

Frankenhaeuser M 1979 Psychoneuroendocrine approaches to the study of emotion as related to stress and coping. In: Howard H E, Dienstbier R A (eds.) *Nebraska Symposium on Motivation, 1978*. University of Nebraska Press, Lincoln, NE

French J R P Jr., Kahn R L 1962 A programmatic approach to studying the industrial environment and mental health. *Journal of Social Issues* **18**(3): 1–47

Grinker R R, Spiegel J 1945 *Men Under Stress*. Blakiston, Philadelphia

Holmes T H, Rahe R H 1967 The social readjustment rating scale. *Journal of Psychosomatic Research* **11**: 213–18

Holt R R 1982 Occupational stress. In: Goldberger L, Breznitz S (eds.) *Handbook of Stress: Theoretical and Clinical Aspects*. Free Press, New York, pp. 419–44

Janis I L 1951 *Air War and Emotional Stress*, 1st edn, McGraw-Hill, New York

Kahn R L, Byosiere P 1992 Stress in organizations. In: Dunnette M D, Hough L M (eds.) *Handbook of Industrial and Organizational Psychology*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA, Vol. 3, pp. 571–650

Kahn R L, Wolfe D M, Quinn R P, Snoek J D 1964 *Organizational Stress: Studies in Role Conflict and Ambiguity*. Wiley, New York

Karasek R A, Baker D, Marxer F, Ahlbom A, Theorell T 1981 Job decision latitude, job demands, and cardiovascular disease: A study of Swedish men. *American Journal of Health* **71**: 694–705

LaRocco J M, House J S, French J R P Jr 1980 Social support, occupational stress and health. *Journal of Health and Social Behavior* **21**: 202–18

Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York

Levi L 1972 Stress and distress in response to psychosocial stimuli. Laboratory and real life studies on sympathoadrenomedullary and related reactions. *Acta Medica Scandinavica*, 191, Supplement No. 528, p. 14. Stockholm, Sweden

Lundberg U 1999 Workplace stress. In: McGuigan F J (ed.) *Encyclopedia of Stress*. Allyn and Bacon, Boston

McEwen B S 1998 Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York Academy of Sciences* **B40**: 33–44

Price R H, Vinokur A D 1995 Supporting career transitions in a time of organizational downsizing: The Michigan JOBS Program. In: London M (ed.) *Employees, Careers and Job Creation: Developing Growth-oriented Human Resource Strategies and Programs*, 1st edn., Jossey-Bass, San Francisco, pp. 191–209

Selye H 1982 History and present status of the work stress concept. In: Goldberger L, Breznitz S (eds.) *Handbook of Stress: Theoretical and Clinical Aspects*, 1st edn. Free Press, New York, pp. 7–17

Stockholm University 1999 *Annual Report XXXIX*. Department of Psychology, Stockholm University, Sweden

R. L. Kahn

Stress Management Programs

Rationale and contents of intervention programs developed for improving stress management skills and for cultivating a personal sense of control are described. Focus is on the cognitive-behavioral programs

which cover the most frequently used components of intervention. The range of application of stress management programs across different groups and settings, such as workplace or chronic illness, is illustrated. Results on the evaluation of the effectiveness of interventions with regard to the different outcome measures used are reported. Implications for the implementation of stress management programs are discussed.

1. Introduction

Stress is generally considered a major impediment to individual health and well-being and to social and societal functioning. Stress is associated with psychological symptoms, diminished well-being, acute and chronic health problems, and impaired performance in work and social settings. To alleviate such stress-related problems, theory-based stress management programs have been developed whose aim is to improve personal skills in dealing more effectively with stressful conditions and to cultivate a personal sense of self-efficacy and control in problem situations. In the next section of this article, the main theories and key concepts that underlie stress management interventions will be delineated. Subsequently, program designs, intervention techniques, and the range of application of stress interventions will be discussed. The fourth section presents results of the evaluation of stress management programs across different outcome measures. Moreover, methodological aspects of intervention studies will be examined. Finally, prospects for future research and program development will be considered.

2. Rationale and Theoretical Background

The majority of stress management programs are based on the following theories: (a) transactional theories of stress and coping, and (b) theories for cognitive-behavioral psychotherapy developed in clinical psychology. In addition to these psychological theories, knowledge of the physiology of stress and the physiological systems activated by stress has resulted in the view that relaxation must be a major objective in healthy stress management. Relaxation techniques have since been advocated that are aimed at reducing heightened arousal and at preventing its short- and long-term deleterious effects.

2.1 Transactional Theories of Stress and Coping

The most influential theory of stress and coping was developed by Lazarus and Folkman (1984) who defined stress as resulting from an imbalance between perceived external or internal demands and the perceived personal and social resources to deal with them. According to Lazarus and Folkman, two cognitive appraisal processes can be distinguished. The initial appraisal, defined as *primary appraisal*, involves the analysis of whether an event is personally relevant. Events perceived as personally relevant can be appraised as either positive or stressful (the latter including possible harm, threat, or challenge). If individuals perceive events as stressful, they evaluate their own resources to deal with the demands. This constitutes the process of *secondary appraisal*. Stress occurs when the demands are perceived as either exceeding or taxing the resources and coping responses become activated. Lazarus and Folkman (1984) defined *coping* as cognitive and behavioral efforts to deal with situations appraised as stressful. Generally, cognitive appraisal and coping processes are influenced by personality factors, personal and social resources, characteristics of the situation, and other variables (see *Stress and Coping Theories*).

A very similar stress model has been developed by Palmer (1997) to guide health professionals in the selection of appropriate stress interventions. It incorporates Arnold Lazarus' seven interacting response modalities comprising behavior, affect, sensory, imaginal, cognitive, interpersonal, and drugs/biology. According to this multimodal–transactional model, stress and coping proceed in five stages. In *Stage 1*, an external or internal stressor is perceived by the individual to be emerging either from an external source or from internal bodily sensations. In *Stage 2*, the individual appraises his or her capacity to deal with the stressor. The person then decides whether he or she has the resources to cope. If the individual perceives that he or she cannot cope, stress may be experienced, and stress responses are likely to become activated. Appraisal of coping capacity is influenced by social and cultural beliefs and attitudes that determine the importance of an event. An example is the belief that one must meet every deadline. If individuals perceive that they cannot cope, they progress to *Stage 3*. In this stage, stress responses occur that include behavioral, affective, sensory, imaginal, cognitive, interpersonal, and physiological changes. In *Stage 4*, the individuals appraise the effectiveness of the coping strategies they have used. If they perceive themselves to have been successful, they return to a state of equilibrium. If they believe, however, that they have failed to cope with the situation, this perceived failure exacerbates existing problems and turns into an additional strain. *Stage 5* relates to the continuing process and long-term consequences. If an individual uses strategies that alter the situation and reduce stress, he or she will return to a state of equilibrium. If the problems persist, however, eventually the person's health may be negatively affected.

In sum, transactional theories of stress and coping focus on the cognitive and affective aspects of an

individual's interactions with the environment and the coping strategies he or she uses or lacks.

2.2 Cognitive Approaches to Psychotherapy

Many of the techniques adopted in stress management programs have their origin in clinical psychology. In addition to the stress and coping theories, stress management is an explicit goal of intervention in cognitive and cognitive-behavioral approaches to psychotherapy. The basis of this form of clinical intervention is the belief that negative thoughts and cognitive distortions (e.g., 'catastrophizing' and overestimating negative outcomes) play a causal role in many clinical disorders. Cognitive interventions seek to modify biased and maladaptive beliefs by cognitive reeducation. The most influential cognitive therapies are Beck's cognitive therapy (Beck 1976) and Ellis' rational emotive behavior therapy (Ellis 1962, Ellis et al. 1997). Within the framework of his *rational emotive behavior therapy*, Ellis has developed the ABC model, which posits that an emotional, behavioral, or physiological reaction (C, for consequence) to an activating event (A) is influenced mainly by the beliefs (B) pertaining to an event and not so much by the activating event *per se*. According to Ellis, irrational beliefs such as 'awfulizing' or dogmatic and absolutist beliefs will lead to self-defeating behavioral and emotional consequences. In psychotherapy, such beliefs are challenged by asking logical, empirical, and pragmatic questions. Generally, cognitive interventions are employed in combination with behavioral techniques and the acquisition of problem solving and social skills.

3. Program Designs

3.1 Intervention Techniques

The importance of transactional theories of stress (Lazarus and Folkman 1984, Palmer 1997) and cognitive therapy approaches (Beck 1976, Ellis 1962, Meichenbaum 1985) lies in their recognition that the subjective interpretation of events plays a critical role in the experience of stress. Moreover, lack of appropriate skills in dealing with problem situations or inappropriate or rigid use of coping strategies may exacerbate stress. Thus, stress intervention now focuses on two aspects: dysfunctional cognitive appraisal such as cognitive distortions and irrational beliefs, and inefficient coping attempts and deficits in a wide range of personal and social skills. Furthermore, in accord with the physiology of stress, reducing the level of physiological arousal is acknowledged as a major objective in stress management. Based on this theoretical framework, the main intervention techniques employed in stress intervention include: (a) *relaxation techniques* that can be used to achieve short-term, *in situ* relaxation, as well as prolonged relaxation in the private setting (e.g., progressive muscle relaxation, meditation, breathing exercises, biofeedback, and autogenic training); (b) *cognitive re-education and cognitive restructuring*, mainly by challenging maladaptive cognitions and negatively biased beliefs; and (c) *skills training* in areas such as problem solving, time management, negotiation, assertive behavior, and the activation of social support (Bunce 1997, Murphy 1996, Palmer 1996, 1997).

According to reviews of stress intervention studies, combinations of individual techniques, especially the combination of relaxation and cognitive-behavioral skills training, are the most frequently used forms of intervention (Bunce 1997, Kaluza 1997, Murphy 1996). A very influential composite stress management program that represents the multicomponent approach is '*stress inoculation training*' (Meichenbaum 1985). The stress inoculation procedure typically includes three stages. The first stage, *education*, focuses on the assessment of an individual's problems, the identification of strategies usually adopted to deal with stressful situations, and reconceptualization of the individual's responses to stressors. Thus, the client learns to become aware of how he or she habitually responds to stressful situations. In the second stage, *rehearsal*, the client learns and practices various coping-skill techniques such as relaxation, problem solving, and cognitive restructuring. During the third stage, *application*, the newly acquired coping skills are practiced under simulated conditions and transferred into everyday life.

In sum, stress management programs seek to reduce physiological arousal, and to provide the individual with a positive perception of stressful situations that is more conducive to well-being. Moreover, the programs aim to teach a wide range of coping skills to be used in a flexible, situation-appropriate way. Above all, the goal of stress interventions lies in promoting a sense of control and self-efficacy in coping with problem situations (see *Self-efficacy and Health*).

3.2 Range of Application

Stress management programs have been implemented in a variety of settings. They are usually *prevention oriented* and focus on psycho-education and cognitive-behavioral skills training. Accordingly, most programs are offered to all who seek such forms of intervention for themselves (Murphy 1996). In contrast to psychotherapeutic approaches, stress management programs are primarily developed for subclinical populations. However, these programs may also address high-risk groups such as hypertension patients, persons suffering from coronary heart disease (Bennett and Carroll 1990, Dusseldorp et al. 1999), or individuals with specific psychological problems such as high levels of stress, anxiety (Murphy 1996), or anger (Novaco 1975). For these groups, there may be

no clear distinction between preventive stress management programs and psychotherapy. Moreover, stress management programs are offered to persons coping with specific stressors or chronically difficult living conditions; this may include caregiving or chronic financial problems (Kaluza 1997). Modified stress interventions are also offered to children and adolescents (Wolchik and Sandler 1997).

Stress management programs are particularly prevalent within *work settings* (Bunce 1997, Murphy 1996). The high frequency of stress interventions that target the workplace and occupational stress is a result of the recognition that stress is a major cause of problems such as absenteeism, accidents, poor work performance, and burnout (see *Stress at Work*). Coping with *chronic illness* is another important domain for which stress management programs have been developed and applied (Devins and Binik 1996, Dusseldorp et al. 1999). The most common forms of stress interventions for patients are education, behavioral training, relaxation, cognitive restructuring, and support groups. Usually, these interventions are part of composite intervention or self-management programs. Such interventions are generally designed to assist the patients to engage in active coping, to change risk behavior, and to feel less hopeless and helpless (see *Chronic Illness, Psychosocial Coping with*).

3.3 Formats

Though stress management programs can be administered individually, most programs are implemented in *small groups*. Typically, a number of weekly training sessions which may last one hour or longer are either conducted over a number of weeks or condensed into block sessions. Learning theories see advantages in both formats (Brown et al. 1998). Interest in working with *large groups* which meet for 'massed learning' format workshops may increase in the future, however, as preventive psycho-education and skills training become part of general health promotion. Brown et al. (1998) have shown that large-scale, day-long stress management workshops are just as effective as the traditional programs consisting of one weekly session. The value of large-scale stress management programs lies in their potential to make preventive training accessible to the general public and to induce more persons to avail themselves of psychological intervention measures.

4. Program Evaluation

4.1 Outcome Measures

In the studies that evaluate the effects of stress-management interventions, a wide range of health outcome measures were used to assess program effectiveness. Frequently employed measures include indices of *subjective well-being* (which, for the most part, are negatively defined as distress reduction, e.g., anxiety, anger/hostility, depression), *perceived strain* associated with different life domains, self-reported *somatic symptoms* and complaints, *control and self-efficacy beliefs*, and *physiological measures*. Moreover, some studies rely on *biochemical measures* (Kaluza 1997, Murphy 1996). According to a meta-analysis of 36 intervention studies, only five studies assessed *coping strategies* (Kaluza 1997). This is especially noteworthy in light of the fact that the acquisition of adaptive coping strategies is the primary focus of stress intervention. Obviously, the existing program evaluations have major shortcomings. In workplace intervention studies, organizational outcome measures such as job satisfaction, work performance, absenteeism, and turnover were additionally assessed (Murphy 1996). Evaluations of programs aimed at coping with chronic illness frequently assess illness-related knowledge, illness-related behavior and cognitions, and disease-related outcomes in addition to measures of subjective well-being (Devins and Binik 1996, Dusseldorp et al. 1999).

4.2 Results

There are several systematic and comprehensive reviews of studies which evaluate the success of stress management programs (Devins and Binik 1996, Murphy 1996). While there exist only a few meta-analyses of preventive stress management interventions, especially of *primary preventive* programs, they clearly attest to positive treatment effects (Kaluza 1997, Lipsey and Wilson 1993; for secondary preventive programs see Dusseldorp et al. 1999). In their review of meta-analyses of psychological treatment effects, Lipsey and Wilson (1993) reported two studies of stress management programs, with the two mean effect sizes indicating positive treatment results. Similarly, a more recent meta-analysis of 36 studies (31 of them conducted in the US) reported small to medium effect sizes found in randomized control studies. Positive effects were found for different outcome measures such as subjective well-being, anger/Type A behavior, coping strategies, physiological measures, and cognitions (Kaluza 1997). Notably, mean effect sizes were greater for follow-up measures assessed between one and six months after the intervention. Particularly remarkable is the increase in mean effect size found in five studies that assessed subjective well-being more than six months after post-test. These results suggest that the effectiveness of stress management programs may be stable or may even increase over time, at least in regard to subjective well-being.

According to a systematic review of 64 studies on *workplace* management programs (Murphy 1996), stress interventions were generally found to be effective, but the effectiveness varied with the outcome measure used. Cognitive-behavioral techniques were

more effective with respect to psychological and cognitive outcomes, mainly anxiety, whereas relaxation was more effective in regard to physiological outcomes, predominantly assessed by blood pressure measurements. Meditation was used in only six studies, but produced the most consistent results across outcome measures. Generally, composite programs which used a combination of techniques proved to be more effective than single-technique programs. Job and organizational outcome measures, mainly job satisfaction, were employed in only 40 percent of the studies. According to Murphy, the low frequency of organizational outcomes may reflect the fact that stress management programs target the individual worker, not the organization. Of the studies reviewed, two-thirds reported no increase in job satisfaction. It appears that the interventions had a greater effect on subjective well-being and physiological arousal, which are subject to personal control, than on outcomes such as job satisfaction and somatic complaints, which are more influenced by external working conditions. The conclusion may be that workplace stress management needs to be more comprehensive and should address personal coping skills as well as the work environment (Murphy 1996).

In sum, meta-analyses and systematic reviews attest to the general effectiveness of stress management programs, especially with regard to psychological and physiological well-being. The programs had less of an effect on outcomes that may be heavily influenced by external factors. Inconclusive are the results for coping skills, which were seldom assessed. However, in interpreting the results one has to bear in mind that studies examining the success of stress management programs face the problem that it is more difficult to demonstrate a significant effect in participants who enter the training with subclinical levels of complaints than it is in participants who enter with high levels of stress. For example, in the review of Murphy (1996), 79 percent of the prevention-oriented studies yielded positive effects, compared with 94 percent of the treatment-oriented studies that involved persons complaining of high stress or anxiety.

4.3 Methodological Aspects

The variety of different intervention techniques raises the question of whether some techniques are more effective than others. Results for contrasting different intervention techniques are equivocal. Murphy (1996) reported technique-specific effects, whereas other reviewers came to the conclusion that seemingly different techniques appear to have similar outcomes, which is generally known as the '*equivalence paradox*' (Bunce 1997, Palmer 1996). According to Bunce (1997), outcome equivalence may be due to two factors. First, methodological and design weaknesses may underlie outcome equivalence. Second, all the different intervention techniques or combinations of techniques may have common features which are responsible for change. The methodological shortcomings that could add to outcome equivalence include: (a) combining techniques even if only one technique is used so that it is not possible to identify which specific components are responsible for the changes that are observed; (b) small sample sizes that may be associated with insufficient statistical power to detect technique-specific effects even if they exist; (c) inappropriate or unspecific outcome measures; and (d) participant heterogeneity.

The second interpretation suggested by Bunce to explain outcome equivalence is that *nonspecific* factors common to distinct contrasting techniques produce similar effects. Such nonspecific factors may be the experience of being in a group and meeting people with similar problems, expectancy effects, or insights gained by participation in respect to the self, others, and one's environment. Indeed, studies that controlled for placebo effects, e.g., with a control group directed to sit quietly or to passively receive information, have demonstrated nonspecific effects (Bunce 1997, Murphy 1996). One may argue, however, that nonspecific effects are no less valuable. In fact, nonspecific effects may be constituent parts of psychological treatment, involving mechanisms of social interaction, expectations, and changes in beliefs and attitudes (Lipsey and Wilson 1993).

The quality of the studies that evaluate intervention effects has been a point of severe methodological critique in the past. Recent reviews, however, reveal increasing methodological rigor (Ivancevich et al. 1990, Kaluza 1997, Murphy 1996). According to a recent review of 64 intervention studies in the workplace, just over half of the studies were true experiments with random assignment to a training or control group (Murphy 1996). However, about 25 percent of the studies did not use a control or comparison group, and 25 percent did not randomly assign participants to training and control groups. Generally, the studies involving randomized groups yielded lower effects, which underscores the importance of randomized control trials. Of the 36 studies that were included in the meta-analysis of Kaluza (1997), 61 percent were true experimental studies with random assignment to training and control groups. In the remaining studies, participants were not randomized into an intervention or control condition. In addition to post-treatment assessment, 53 percent of the studies reported results for a follow-up between one and over six months after post-test.

5. Prospects for Future Program Development

Stress management programs are designed for a wide range of possible stressors and chronically difficult living conditions. Most of them are preventive in nature and open to all persons who wish to improve their skills in dealing with stressful situations and

enhance their feelings of control. Unspecific and stressor-specific stress management programs provide an invaluable service for health promotion by offering psycho-education, skills training, and adaptive perceptions of problem situations. Generally, systematic reviews and meta-analyses of intervention studies attest to the effectiveness of preventive stress management programs. Clearly, however, further research is needed, especially with nonclinical populations attending preventive stress management programs (Kaluza 1997, Palmer 1996). Future research should focus on the following four problems.

(a) Studies on stress intervention should seek to identify the factors that are responsible for change and the *mechanisms of change*. Such mediators by which changes may be engendered could be psychological variables such as personal sense of control and self-efficacy, the use of adaptive coping strategies, and insights into self, others, and work (Bunce 1997).

(b) Outcome measures should be used that reflect the specific goals of stress management intervention such as realistic and adaptive stress-related cognitions, increased skills for dealing with everyday stress, and improved social competencies (Kaluza 1997). Currently, most outcome measures pertain to subjective well-being, which may indicate improvement in coping skills, but does not clearly reflect actual changes in stress-related interactions between the individual and his or her environment. Moreover, the effectiveness of stress intervention should be evaluated by long-term follow-up assessments in order to ascertain whether the interventions guard against future stress experiences.

(c) Biographical and psychological individual differences such as age, gender, self-esteem, or locus of control may systematically influence changes effected by stress interventions (Bunce 1997). According to Bunce, future research must consider *individual differences* and should examine how these factors might act as moderators and mediators of outcome variance. Matthews and Wells (1996) assert that intervention is likely to be most successful when it is based on systematic analyses of individual styles of attention and coping. Miller has shown that interventions may be more effective when matched with the dispositional information-processing style (e.g., Miller and Mangan 1983).

(d) In the development of stress intervention programs, *cultural aspects* should be taken into consideration in addition to biographical and psychological individual differences. For instance, minority groups in work or social settings may experience stress due to racial prejudices that could become the target of special stress intervention measures (Palmer 1996). Generally, cultural differences may play an important role in successful health promotion. Matching stress interventions with cultural beliefs and models of health and illness is a challenging task for multicultural societies (Devins and Binik 1996).

There is every reason to believe that stress will continue to be a salient problem in the near future. Increasingly, individuals and organizations will seek effective interventions to improve coping skills and reduce stress-related problems. Providing methodologically sound and effective unspecific and stressor-specific stress management programs will remain a predominant task for all who are involved in health promotion.

See also: Chronic Illness, Psychosocial Coping with; Social Support and Stress; Stress and Coping Theories; Stress and Health Research; Stress at Work

Bibliography

Beck A T 1976 *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York

Bennett P, Carroll D 1990 Stress management approaches to the prevention of coronary heart disease. *British Journal of Clinical Psychology* **29**: 1–12

Brown J S, Cochrane R, Mack C F, Leung N, Hancox T 1998 Comparison of effectiveness of large scale stress workshops with small stress/anxiety management training groups. *Behavioural and Cognitive Psychotherapy* **26**: 219–35

Bunce D 1997 What factors are associated with the outcome of individual-focused worksite stress management interventions? *Journal of Occupational and Organizational Psychology* **70**: 1–17

Devins G M, Binik Y M 1996 Facilitating coping with chronic physical illness. In: Zeidner M, Endler N S (eds.) *Handbook of Coping*. Wiley, New York

Dusseldorp E, van Elderen T, Maes S, Meulman J, Kraaij V 1999 A meta-analysis of psychoeducational programs for coronary heart disease patients. *Health Psychology* **18**: 506–19

Ellis A 1962 *Reason and Emotion in Psychotherapy*. Lyle Stuart, New York

Ellis A, Gordon J, Neenan M, Palmer S 1997 *Stress Counselling: A Rational Emotion Behaviour Approach*. Cassell, London

Ivancevich J M, Matteson M T, Freedman S M, Philipps J S 1990 Worksite stress management interventions. *American Psychologist* **45**: 252–61

Kaluza G 1997 Evaluation von Streßbewältigungstrainings in der primären Prävention—eine Meta-Analyse (quasi-)experimenteller Feldstudien [Evaluation of stress management interventions in primary prevention—a meta-analysis of (quasi-)experimental studies]. *Zeitschrift für Gesundheitspsychologie* **5**: 149–69

Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York

Lipsey M W, Wilson D B 1993 The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist* **48**: 1181–209

Matthews G, Wells A 1996 Attentional processes, dysfunctional coping, and clinical intervention. In: Zeidner M, Endler N S (eds.) *Handbook of Coping*. Wiley, New York

Meichenbaum D 1985 *Stress Inoculation Training*. Pergamon, New York

Miller S M, Mangan C E 1983 Interacting effects of information and coping style in adapting to gynecologic stress: Should the doctor tell all? *Journal of Personality and Social Psychology* **45**: 223–36

Murphy L R 1996 Stress management in work settings: A critical review of the health effects. *American Journal of Health Promotion* **11**: 112–35
Novaco R 1975 *Anger Control: The Development and Evaluation of an Experimental Treatment*. Heath, Lexington, MA
Palmer S 1996 Developing stress management programs. In: Woolfe R, Dryden W (eds.) *Handbook of Counseling Psychology*. Sage, London
Palmer S 1997 Stress, counselling, and management: Past, present and future. In: Palmer S, Varma V (eds.) *The Future of Counselling and Psychotherapy*. Sage, London
Wolchik S A, Sandler I N 1997 *Handbook of Children's Coping*. Plenum, New York

H. Weber

Stress: Measurement by Self-report and Interview

The term 'stress' is used in a variety of different ways. Sometimes it is employed to refer to an experience of physical tension or distress stimulated by the environment and sometimes to the occurrence of events or experiences that are assumed to involve significant adaptive demands. In this latter case, the focus is on stressors, measured in terms of the occurrence of events and circumstances of stress-evoking potential. Even though it is useful to distinguish stressors from stress, the two terms are often used interchangeably. Despite varying characterizations of stress, Cohen et al. (1995, p. 3) have argued that all perspectives share a concern with the process in which 'environmental demands tax or exceed the adaptive capacity of an organism, resulting in psychological and biological changes that may place persons at risk for disease.'

This article is focused on stress measurement in social as opposed to laboratory contexts. For that reason, strategies for assessing the environmental activation of physiological systems are not covered. Initial consideration will be given to measures of eventful stressors, which fall into two categories—those using a checklist approach and those that rely on data obtained in detailed interviews. This will be followed by brief discussions on the measurement of chronic stress and of exposure to major, potentially traumatic events.

1. *Life Event Checklists*

By far the largest amount of available evidence linking stress exposure to physical and mental health has come from studies employing checklist measures of social stress. Stimulated by Adolf Meyer's life chart procedure, Hawkins et al. (1957) developed the Schedule of Recent Experiences (SRE). This 42-item questionnaire covered events in areas such as marital problems, relationship changes, work and residential changes, and family relationships. A decade later Holmes and Rahe (1967), elaborating on the SRE, developed the Social Readjustment Rating Scale (SRRS) by employing a panel of judges to rate the amount of adjustment each event would typically require. Although research has not supported the expectation that the weights derived from these ratings would improve estimates of stress exposure, their contribution to the face validity of event checklists was evidently substantial. Following the publication of the SRRS there was a huge outpouring of research on stressful life events and it was, for a considerable period, the best-known and most widely used instrument.

The conceptualization that initially informed the checklist approach was that an individual's level of experienced stress was embodied in the cumulative amount of change or readjustment brought about by personal events occurring in the recent past (Holmes and Rahe 1967). This idea that the amount of change is the event property responsible for the stressful impact of life events seems to have been based on Selye's (1956) contention that stress is comprised of nonspecific biological change elicited in response to environmental events and that even relatively small changes occurring in close succession can influence susceptibility and thus disease.

Over the years since the publication of the Social Readjustment Rating Scale (SRRS) there has been an active debate over its adequacy and about alternative approaches to measuring life events and to evaluating their stress-engendering properties. Indeed, at least 20 critical reviews of stressful life event methodology have been published since 1975 (Turner and Wheaton 1995). However, the checklist method has evolved in response to criticism, and a substantial array of different life event scales have been developed. Turner and Wheaton (1995) provide a nonexhaustive list of 47 different life event inventories—many targeted to specific age groups.

Despite several clearly justified criticisms and associated questions about the reliability and validity of resulting scores, the various published and unpublished life event inventories collectively represent the traditional, and still dominant, research procedure for estimating variations in stress exposure. A substantial preponderance of the now massive literature assessing the linkage between social stress and health is based on checklist measures of life stress. This literature has indicated a reliable association between life stress and the occurrence of both psychological distress and risk for physical health problems (e.g., Thoits 1983, Jemmot and Locke 1984).

A conclusion that seems to follow from the consistency with which these relationships have been observed is that, whatever the shortcomings of the method, event inventories yield meaningful estimates

of stress exposure. However, there are two major challenges to this conclusion. One is the possibility that observed associations are an artifact of operational confounding arising from the fact the many checklists, perhaps the majority historically, have included items that are arguably symptoms of physical or mental health problems, or the consequences of such problems. Although evidence on this possibility has not been entirely consistent, the majority of studies have found that the relationships at issue persist, and are little attenuated, when potentially confounding items are removed from consideration (e.g., Tausig 1982).

The second challenge involves the hypothesis of reverse causality. Concern that illness or distress also produces either real or artifactually higher numbers of reported events has been a prominent theme within the life events literature. A crucial point in this regard is that the possibility that illness also causes stress exposure does not question the causal status of life event scores with respect to illness; it simply complicates the proper estimation of this causal impact (Turner and Wheaton 1995). In this connection, longitudinal studies consistently suggest that if one controls for prior or baseline distress or mental health, life events have a net impact on change in distress or mental health over time. Thus, at least with respect to mental health, where the causal direction debate has been most intense, the literature supports two essential criteria of causality: (a) statistical association, and (b) temporal precedence.

The primary remaining issue is the possibility of spuriousness in the apparent effect of life events on health outcomes. Although this explanatory hypothesis cannot be entirely set aside, the literature is filled with examples of studies that have attempted to control for confounding factors, and almost none that have claimed to explain away the life events—illness relationship (e.g., Kessler et al. 1985, Turner and Lloyd 1999).

1.1 Change Events Versus Negative Change Events

As previously noted, the conceptualization that guided developers of earlier event checklists was that change *per se*, whether positive or negative in nature, required adaptation and can have a stressful impact. Although this perspective continues to be represented among a number of prominent investigators, the majority of life event researchers have come to focus upon undesirable change assessed with lists containing only putatively negative events. The debate about whether it is change or undesirability that is relevant to the experience of stress is difficult to resolve because there is no wholly satisfactory criterion measure—at least not one that can be applied in field settings. However, a number of researchers have argued that power to predict distress or illness represents a reasonable criterion (e.g., Turner and Wheaton 1995). From this perspective, considerable evidence on the change versus undesirability issue has accumulated. Zautra and Reich (1983) published a review of 17 studies that examined the relationships of desirable and undesirable events to measures of psychological distress. They observed a consistent pattern of clear positive relationships with negative events in contrast to weak and contradictory findings with respect to positive events. Results suggesting the dominant significance of negative compared to positive events as risk factors for distress and illness have been quite consistent across a range of study populations, dependent variables, and life event inventories (Turner and Wheaton 1995). However, the conclusion that ostensibly positive life changes play no role in either elevating or conditioning stress experiences cannot be made with confidence. The mechanisms involved may simply be more complex than research has so far been able to reveal. The data do support the recommendation that, in the context of the limited time or space that characterizes most questionnaires, event checklists be restricted to events that are presumptively undesirable in nature.

1.2 Event Timing and Duration

Most checklist studies on the health effects of social stress have employed a one-year time frame for inquiring about experienced events. This time frame appears to be favored because it is regarded as long enough to obtain a reasonable estimate of variations in exposure to recent life events, and short enough to avoid the substantial decline in the ability of respondents to recall events that appears to occur beyond one year. In contrast, studies employing two- to five-year time frames have sometimes reported higher correlations than those that have considered shorter periods. In general, the choice of an appropriate time frame must involve consideration of the nature of the outcome of interest and the estimated or assumed causal lag of event effects for that outcome. When relatively long time periods are selected, group decline in event reporting can be investigated as a function of months prior to interview and cut-points for the time frame established based on the results.

In this connection, it should be recognized that the problem of reliability of reporting or of remembering, and the contention that stressors may be time limited in their effects, cannot be applied equally to all events that may be experienced. For example, it would rarely be the case that a subject would not remember, or fail to report in response to specific questions, that his or her parent had died, or child had died, or parents had divorced that he or she had served in combat, or had been physically or sexually assaulted. Evidence that a number of major events or circumstances are relevant to mental health and that the duration of their effects is closer to a lifetime than to one year suggests the utility of considering such events within efforts to index a person's burden of stressful experience. The

measurement of major, potentially traumatic events is discussed briefly below.

It is widely recognized that both discrete life events and relatively enduring problems are relevant to distress and illness. The experience of eventful stress may alter the meanings of existing strains, generate new strains, or magnify existing strains (Pearlin 1983). Alternatively, the impact of recent events may be amplified in the presence of chronic stress (Brown and Harris 1978). To test the idea that eventful and chronic stressors act synergistically requires the capacity to distinguish stressors within the two categories. Although life event checklists aim to assess discrete events, it is clear that some occurrences of a given event are time-limited in nature, whereas other occurrences of the same event may represent the beginning of a long-term difficulty.

In a large-scale community study ($n = 1561$), Avison and Turner (1988) obtained information for each reported event on when it ended as well as when it had occurred. Analyses confirmed that event checklists routinely reflect differences in exposure to relatively enduring as well as discrete stressors. Moreover, none of the 31 negative events assembled by these investigators from established lists were found to be uniformly experienced as either discrete or chronic stressors. This research thus indicates that enduring or chronic stress is part of what is reflected in associations of checklist scores with health outcomes and that it is possible to distinguish its contribution. However, the enduring stressors that might be indexed within an event checklist are likely to be too limited and narrow to yield adequate estimates of variations in exposure to chronic stress. This issue is discussed further below.

1.3 Event Weighting

Considerable attention has been given to differential weighting of life events, presumably because the idea that there are no differences in impact potential across events is counterintuitive. There are several types of weighting, but the two most common are 'objective' ratings of change, or importance, or seriousness, of events by 'judges,' selected in random samples of defined populations, in purposive samples of individuals who have experienced the event being weighted, or by expert panels, and 'subjective' ratings assigned by individuals to their own events and considered only at the individual level. A primary difference between these two is the level of aggregation of the weight: objective weights reflect an average assigned by a group of raters; subjective weights involve self-reporting by the individual on how serious or stressful the event was for them.

Thus, 'objective' weights are attempts to capture differences in event characteristics that imply differences in impact potential that are applicable across individuals, whereas 'subjective' weights attempt to capture differences in impact potential associated with variations in the meaning of an event for individuals themselves. As noted above, despite numerous attempts to demonstrate otherwise, research has failed to yield evidence for the effectiveness of the 'objective' weighting of events. Nevertheless, some researchers have called for additional efforts, arguing that the issue is not yet settled (Turner and Wheaton 1995).

Concern about subjective weights most often revolves around the issue of confounding estimates of stress exposure with measures of outcome. Reports of illness or distress and of the seriousness or stressfulness of reported events may reflect the same domain of experience. This problem is arguably less serious when the outcome of interest is physical health status and, with longitudinal data, bias can be substantially reduced by partialling out the effects of events on later health. However, subjective weights are problematic in another respect: they tend to confound differential stress exposure with differential vulnerability to stress. The weight attached by a respondent to an event will be substantially a function of his or her capacity, real or perceived, to handle that event emotionally and/or practically. This capacity is largely determined by the individual's coping skills and the availability of social and personal resources. Thus, subjective ratings confound the effects of stress mediators/moderators with those of stress exposure and should therefore be avoided even in prospective studies and studies on physical health status.

2. Interview Measures of Life Events

Interview measures are distinguished from the checklist approaches by their use of qualitative probes. These probes are employed to more precisely specify the characteristics of life events that might be stressful. The approach avoids the violence to logic implicit in assuming that widely differing events have equivalent stress-evoking potential by specifically attempting to establish the degree of severity of each event exposure (Wethington et al. 1995). Other major strengths of intensive interview measures include achievement of greater comprehensiveness of event coverage through flexible domain-specific elicitation and self-nomination of events not evoked by direct questioning, and the ability to establish the time ordering of stress exposure and health outcomes with greater confidence. The major disadvantage of interview measures is that they are extremely time-consuming and expensive relative to checklist methods.

2.1 Life Events and Difficulties Schedule (LEDS)

The Life Events and Difficulties Schedule (LEDS) is the most widely used personal interview method and has produced perhaps the most compelling evidence available for the role of stress in psychiatric disorder (Brown and Harris 1978). This semistructured in-

strument yields a narrative story of each experienced event that is used by investigators to rate the importance of events. These investigator-based ratings attempt to estimate what the impact would be in a specific context for the average person. It is the experience of a 'severely' threatening event or of a severe chronic difficulty that is hypothesized to increase risk for the occurrence of physical or psychiatric disorder. Severity is established through ratings of 'long-term contextual threat,' aided by the use of 'dictionaries' of examples of rated events and difficulties, the use of which requires a formal training course (see Wethington et al. 1995).

The considerable strengths of this approach are offset by the very high cost noted earlier. This fact, along with the extensive amount of time required to complete the rating process, may make the method unsuitable for large-scale community studies. From certain points of view, a second disadvantage is that the method, in evaluating social context within the severity rating process, ends up aggregating stress exposure with factors that may amplify or moderate its impact. That is, what has been described as the stress process is substantially represented in the ratings of contextual threat. Although this results in powerful predictions of health outcomes, information that is crucial to interventionists is obscured when the effects of stress exposure and of differential vulnerability are confounded. As Wethington et al. (1995) have noted, Brown and his colleagues continue to refine the measure to address this criticism, at least to some extent.

2.2 *The Standardized Event Rating System*

The Standardized Event Rating System (SEPRATE) utilizes a structured interview format to develop descriptions of each event or difficulty reported (Dohrenwend et al. 1993). In addition to the narrative description, the probes allow a standardized assessment of separate aspects of events and situations thought to condition level of stress experienced. A major difference from the LEDS interview is that material that might infer differences in stress vulnerability is stripped from narratives to avoid the inclusion of vulnerability estimates within event ratings.

This interview measure is relatively new and has not been as widely tested as the LEDS. Like the LEDS it is relatively expensive to use, is lengthy to administer, and requires specialized training from the authors.

3. *Chronic Stress*

Despite growing evidence that chronic stressors affect physical and psychological health, there have been few attempts toward comprehensive measurement of such stressors. The relative lack of attention in this area may arise from concern that reports of enduring role-related or other stressors may be more a consequence or reflection, than a cause of disorder. A promising 51-item inventory has recently been offered by Wheaton (1997) who has presented evidence contrary to this concern about causal confounding. The inclusion of this measure, along with a life events checklist, has been shown to account for about 2.5 times as much variance in depressive symptomatology as typically reported in the literature (Turner et al. 1995). This measure assesses a wide range of role-related and role-unrelated chronic stressors and the observation that it is unique association with depression suggests the need in future research for including a separate sampling of items addressed to the general domain of chronic stress. There is also a variety of less comprehensive questionnaires that deal with specific domains such as work or relationship strains. These have been enumerated and critiqued by Lapore (1995) and Herbert and Cohen (1995).

4. *Lifetime Exposure to Major/Traumatic Events*

There is evidence that traumatic events or circumstances, individually or cumulatively, are relevant to mental health (Rutter 1989, Kessler and Magee 1993, Turner and Lloyd 1995). There appears to be a range of events that, presumably, can be reliably measured and that can have significant mental health consequences despite their occurrence years, or even decades, earlier. This argues strongly for the systematic inclusion of lifetime experience of such events as part of the effort to assess difference in stress exposure. Analyses should consider both the direct individual and cumulative effects of lifetime traumas and the extent to which they may condition the impact of more contemporaneous stressors. Brief inventories of major/traumatic events are presented in Kessler and Magee (1993) and Turner and Lloyd (1995). However, because the nature of traumatic experience varies across populations and age cohorts, event selection and coverage should be specific to each study population.

It is important to understand that the health significance of social stress has never been effectively tested because we have yet to adequately measure differences in stress exposure. The conceptual consequences of this inadequate measurement can be profound. This is so because unmeasured group differences in stress exposure masquerade within research findings as group differences in vulnerability. Failure to take account of differences in traumatic stress or in the stress of discrimination can inappropriately suggest group differences in coping skills or other personal attributes relevant to adaptation. It seems clear that efforts to measure stress exposure should be as comprehensive as possible and take

account of eventful stressors, more enduring strains, lifetime occurrence of traumatic events, and stress associated with discrimination and acculturation.

See also: Coping across the Lifespan; Coping Assessment; Stress and Coping Theories; Stress Management Programs; Stress, Neural Basis of; Stress: Psychological Perspectives

Bibliography

Avison W R, Turner R J 1988 Stressful life events and depressive symptoms: Disaggregating the effects of acute stressors and chronic strains. *Journal of Health and Social Behavior* **29**: 253–64

Brown G W, Harris T O 1978 *Social Origins of Depression*. Free Press, New York

Cohen S, Kessler R C, Gordon L U 1995 *Measuring Stress: A Guide for Health and Social Scientists*. Oxford University Press, New York

Dohrenwend B P, Raphael K G, Schartz S, Stueve A, Skodol A 1993 The structured event probe and narrative rating method for measuring stressful life events. In: Goldberg L, Breznitz S (eds.) *Handbook of Stress*. Free Press, New York, pp. 174–99

Hawkins N G, Davies R, Holmes T H 1957 Evidence of psychosocial factors in the development of pulmonary tuberculosis. *American Review of Tuberculosis and Pulmonary Diseases* **75**: 768–80

Herbert T B, Cohen S 1995 Measurement issues in research of psychosocial stress. In: Kaplan H (ed.) *Psychosocial Stress: Perspectives on Structure, Theory, Life-course and Methods*. Academic Press, New York, pp. 295–332

Holmes T H, Rahe R H 1967 The social adjustment rating scale. *Journal of Psychosomatic Research* **11**: 213–8

Jemmot J B, Locke S E 1984 Psychosocial factors, immunologic mediation and human susceptibility to infectious diseases: how much do we know? *Psychological Bulletin* **95**: 78–108

Kessler R C, Magee W J 1993 Childhood adversities and adult depression: Basic patterns of association in a US national survey. *Psychological Medicine* **23**: 679–90

Kessler R C, Price R H, Wortman C B 1985 Social factors in psychopathology: Stress, social support, and coping processes. *Annual Review of Psychology* **36**: 531–72

Lapore S J 1995 Measurement of chronic stressors. In: Cohen S, Kessler R C, Gorden L V (eds.) *Meassuring Stress: A Guide for Health and Social Scientists*. Oxford University Press, New York

Pearlin L I 1983 Role strains and personal stress. In: Kaplan H B (ed.) *Psychosocial Stress: Trends in Theory and Research*. Academic Press, New York, pp. 3–31

Ross C E, Mirowsky J 1979 A comparison of life event weighting schemes: change, undesirability, and effect proportional indices. *Journal of Health and Social Behavior* **20**: 166–77

Rutter M 1989 Psychosocial risk trajectories and beneficial turning points. In: Doxiadis S, Stewart S (eds.) Early influences in shaping the individual. *NATO Advanced Science Institutes Series, Series A: Life Sciences*. Plenum Press, New York

Selye H 1956 *The Stress of Life*. McGraw-Hill, New York

Tausig M 1982 Measuring life events. *Journal of Health and Social Behavior* **23**: 52–64

Thoits P A 1983 Dimensions of life events that influence psychological distress: An evaluation and synthesis of the literature. In: Kaplan H B (ed.) *Psychosocial Stress: Trends in Theory and Research*. Academic Press, New York, pp. 33–101

Turner R J, Lloyd D A 1995 Lifetime traumas and mental health: The significance of cumulative adversity. *Journal of Health and Social Behavior* **36**: 360–76

Turner R J, Lloyd D A 1999 The stress process and the social distribution of depression. *Journal of Health and Social Behavior* **40**: 374–404

Turner R J, Wheaton B 1995 Checklist measurement of stressful life events. In: Cohen S, Kessler R C, Gordon L U (eds.) *Measuring Stress: A Guide for Health and Social Scientists*. Oxford University Press, New York

Turner R J, Wheaton B, Lloyd D A 1995 The epidemiology of social stress. *American Sociological Review* **60**: 104–24

Wheaton B 1997 The nature of chronic stress. In: Gottlieb B (ed.) *Coping With Chronic Stress*. Plenum, New York

Wethington E, Brown G W, Kessler R C 1995 Interview measurement of stressful life events. In: Cohen S, Kessler R C, Gordon L U (eds.) *Measuring Stress: A Guide for Health and Social Scientists*. Oxford University Press, New York

Zautra A J, Reich J W 1983 Life events; and perceptions of life quality: Developments in a two-factor approach. *Journal of Community Psychology* **1**: 121–32

R. J. Turner

Stress, Neural Basis of

The term stress is semantically confusing and difficult to define. The confusion of what is meant by stress is compounded by its use both as a noun, i.e., a stressful experience or life event, and as a verb, i.e., a stress response. In addition, stress is usually seen as a negative state particularly as it is often used to describe the state of having too much to do in too little time. Despite the negative connotations, the dictionary defines stress, as something that weight is placed upon, something important. Synonyms for stress include emphasis, accent, and significance; not exactly definitions that should elect such negative responses. Another definition is raised from the field of physics, where the typical synonym is strain. That strain is typically measured through changes in the physical properties of matter. This definition has some utility when considering the neural basis of stress. That is, a given amount of stress can be useful to material, testing its strength. But when the amount of stress is increased precipitously, those physical properties can change, often irrevocably. Combining these two perspectives, then, it is tempting to assert that stress is the physical response to environmental change beyond the expected level of force. This would necessarily exclude changes in sensory processing, i.e., perception, which may or may not be stressful. However, it also excludes two of the more salient features of the stress response: subjectivity and emotionality.

With regard to subjectivity, what is stressful to one person is not necessarily stressful to another. And with respect to emotionality, few would deny the emotional

content of a stressful experience. The true definition lies somewhere between the mechanistic/physical view and a more emotional/subjective one. For the purposes of this submission, we will consider stress to be the subjective response to environmental changes beyond an expected level of force. Experience of stress and the neural processes that mediate it will be discussed here.

Possibly the most accepted definition of stress was proposed by the eminent psychologist Hans Seyle who defined it as the nonspecific response of the body to any demand placed on it. With an emphasis on non-specificity, Seyle was referring to the most well described stress response systems in the body: the hypothalamic-pituitary-adrenal axis (HPA). In response to the stressor, the paraventricular nucleus of the hypothalamus releases the corticotrophin-releasing factor (CRF), which induces the release of adrenocorticotrophic hormone (ACTH) from the anterior lobe of the pituitary. ACTH exerts its effects on the adrenal cortex promoting the synthesis and secretion of glucocorticoids, cortisol in humans, into the bloodstream. Glucocorticoids are the final effectors of the HPA axis and exert homeostatic regulatory effects including energy mobilization and inhibition of anabolic processes. The HPA axis is based on negative feedback resulting from the release of glucocorticoids into the blood that suppress the release of CRF from the hypothalamus and ACTH secretion from the pituitary, and results in termination of the stress response. Concomitantly with the release of ACTH from the pituitary, the opioid peptide beta-endorphin is also released having wide-ranging modulatory effects on neurotransmission and hormone regulation in the brain. Seemingly paradoxical, beta-endorphin plays a role in stress-induced analgesia and prevents the stress response from overwhelming the capacity of the organism to contend with the stressor.

Glucocorticoids are released in response to every type of stressor known to humans, from novelty to chronic pain, to death. They are also released in response to more positive emotions and experiences such as pleasure and learning. Because glucocorticoids are released into the bloodstream and thus have access to nearly every known organ in the body, the emphasis on non-specificity is not unfounded. Activation of the HPA is known to have widespread consequences and under supra-threshold conditions can be detrimental to neuronal functioning. However, Seyle probably did not mean to infer that the stress response in its entirety was nonspecific and certainly did not mean to infer that it was out of control. In fact, there is a recent appreciation for the role that glucocorticoids play as anti-inflammatory agents and in bringing the system back to homeostasis after a stressful experience.

The many different neural and neuroendocrine pathways activated as a result of stressful experience further exemplify the oversimplification of Seyle's emphasis on non-specifity. In addition to the secretion of glucocorticoids by the adrenal cortex, the adrenal medulla and autonomic nervous system also have prominent actions in mediating the stress response. The autonomic nervous system consists of two branches, the sympathetic and parasympathetic, which work in opposition to each other. Activation of the sympathetic division of the autonomic nervous system stimulates the chromaffin cells of the adrenal medulla to secrete the catecholamines epinephrine and norepinephrine into the bloodstream. Epinephrine and norepinephrine induce a variety of physiological changes that provide the organism with the necessary energy resources to carry out its actions under stressful experiences in preparation for the 'fight or flight' response, a term used by Walter Cannon. As a result, respiration increases to facilitate the exchange of oxygen in the lungs, while cardiac output is increased to divert blood from digestive areas to muscles. Sympathetic activation also results in an increase in the supply of glucose via the inhibition of insulin and increased glucagon production that converts stored glycogen into glucose. Although counterintuitive, the inhibition of insulin is very adaptive since it increases the availability of glucose to the brain thereby enhancing neuronal function. In contrast to the sympathetic, the parasympathetic division is promptly inhibited by stress and thereby postpones the costly non-essential anabolic processes of vegetative function such as digestion.

In addition to these peripheral responses to stress, there are numerous central responses. Among other purposes, they are important in establishing the 'what' and 'where' of the stressful experience, as well as the formation of memories about the experience. For example, CRF neurons form part of the critical HPA response to stress and are innervated by limbic areas including the amygdala, bed nucleus of the stria terminalis, and lateral septum. It is suggested that they provide emotional tone to or perception of the stressful stimulus. Brainstem afferents from catecholaminergic neurons have a stimulatory action on HPA activation, as does serotonergic input from the dorsal raphe nucleus. Inhibitory input is provided by local GABA neurons that decrease the release of CRF. CRF neurons also receive input from other hypothalamic nuclei including the superchiasmatic nucleus, which serves to synchronize circadian rhythmicity and control the daily oscillations of hormone secretion. The stress response in turn activates several brain regions and pathways. For example, exposure to stressful experience stimulates the mesolimbic/mesocortical dopamine systems, which in turn innervate the prefrontal cortex, an area involved in anticipatory phenomenon and cognitive function. Furthermore, the mesolimbic/mesocortical systems are linked to the nucleus accumbens, which are implicated in reinforcement/motivational and reward processes.

The amygdala and hippocampal components of the limbic system are also stimulated during stress. The

amygdala is important for the emotional analysis of the stressor and has direct connections with the hypothalamus, cortex, and regions of the brainstem involved in autonomic control. Thus, the amygdala may serve as an integrator linking regions responsible for detection of stressors involving cognitive processing (e.g., the cortex and hippocampus) and those that serve as effectors of the stress response (e.g., the hypothalamus and brainstem controls). The hippocampus, on the other hand, appears to be involved in inhibition of the HPA axis and thus is part of the feedback control of ACTH release. The inhibitory role of the hippocampus depends on activation of hippocampal glucocorticoid receptors, which stimulate the fimbria-fornix pathway from the lateral septum and bed nucleus of the stria terminalis inducing the release of the neuropeptide, somatostatin, to inhibit the PVN. Thus, the interaction between the stress system and neural elements function to regulate the autonomic, endocrine, and behavioral responses to stress allowing for perception of the stressor, emotional coding, and initiation of action.

The HPA axis of the stress system exerts a profound influence on the regulation of growth, reproduction, and the immune function. There is a stress related suppression of growth hormone that inhibits growth in the young organism and the repair of existing tissue in adults. In both the male and female reproductive systems, stress induced CRF and beta-endorphin secretion suppress the gonadtropin-releasing hormone (GNRH) at the level of the hypothalamus. Furthermore, glucocorticoids decrease pituitary responsiveness resulting in suppression of the leutenizing hormone (LH) and the release of the follicle-stimulating hormone (FSH). Testicular and ovarian sensitivity to LH is also decreased by glucocorticoids. Stress can thereby disrupt ovulation and implantation in females and reduce sperm production and testosterone secretion in males. Stressors also exert multiple effects on the inflammatory/immune reaction. Glucocorticoids mediate stress-induced immunosuppression via a negative feedback loop. Glucocorticoid secretion resulting from HPA axis stimulation by cytokines (i.e., interleukin-1, tumor necrosis factor-alpha) suppresses the stimulatory effects of cytokines on the HPA axis. It is obvious that the impact of stressor exposure is widespread and extends to many physiological and neural systems.

Defining stress and its neural consequences is made more difficult by the parameters of stress. For example, the duration of stressful stimulus can be either chronic or acute and the neural indices differ dramatically depending in these temporal characteristics. Stressful experience also differs in its intensity. Minor stressful experiences can induce a range of responses that are much different from more severe forms of stress. This is contradictory to Seyle's view of a common stress response regardless of the nature of the stressor. The stress system produces adaptive responses to acute stressors. However, the same stressor, when chronic or activated without sufficient physiological cause, becomes pathogenic resulting in the emergence of stress related disease. Prolonged activation of the catabolic processes of the stress response can result in cardiovascular (hypertension), metabolic (myopathy, fatigue), gastrointestinal (gastric ulcers), and reproductive (impotence and amenorrhea) consequences, as well as complications associated with immunosuppression (impaired disease resistance). Although there is consistency of responses activated during stress, responses can be specific to the particular characteristics of the stressor.

Stress physiology is not only modulated by physical stressors, but also by psychological variables. In fact, psychological factors can initiate a stress response even in the absence of physical insult and homeostatic disruption. Animals and humans have been shown to exhibit the classic stress response during circumstances such as bereavement, difficult cognitive tasks, and conditioned fear. The stress response initially evolved as a way of contending with environmental challenges but over time came to be activated by cognitive phenomenon. The magnitude of a stressful event is modulated by not only the nature of the stressor and the resulting physiological perturbations, but also individual perception and interpretation of the stressor. Lack of predictability and control are the two major elements underlying the cognitive appraisal of stress. Uncertainty about an aversive event reduces feelings of control over the situation thereby inducing higher levels of stress. Related to this is the well-documented phenomenon of learned helplessness in which prolonged exposure to uncontrollable and inescapable aversive stimuli can produce emotional disruption and some deficits in performance, while exposure to the same amount of controllable stress has no effect. Furthermore, the psychological component of stressors underscores the element of individuality of the stress response. What is stressful to one individual may not be to another. Factors such as gender, prenatal and early life psychological and physical stress have profound and long-lasting effects on the basal activity of the stress system. Just as chronic activation of the stress response can lead to a manifestation of physical pathologies, so it can lead to psychiatric syndromes including melancholic depression, anorexia nervosa, obsessive compulsive disorder, and panic anxiety.

It has been demonstrated thus far that stressors have extensive impact on various neural and peripheral systems. But, the aforementioned neuromodulatoy systems also contribute to a range of behaviors allowing for maximal ability of the organism to contend with the stressful stimulus. This reciprocal relationship between behavior and physiology is essential to overcoming a stressor and re-establishing homeostasis. Reproductive and feeding behaviors are suppressed allowing for appropriate provision of

energy to overcome the stressor. The degree of arousal and cognition is also altered. As already demonstrated, acute stress can produce adaptive behavioral responses. Such a response would include increased arousal and enhanced learning and memory resulting from blocked glucose transport to cells benefiting neuronal function by providing more glucose to neurons. As a result, the animal is able to contend with the stressful experience. On the other hand, chronic stress can be detrimental to neural function. Elevated glucocorticoids damage hippocampal neurons resulting in the inability of the hippocampus to inhibit CRF release. Consequently, there is an increased secretion of glucocorticoids leading to additional hippocampal degeneration. Such damage can produce deficits in learning and memory. An inverted U function could thus be used to represent the relationship between stress and performance. Optimal performance occurs with moderate levels of demand but is compromised at the extremes. Low levels of stress, such as boredom or drowsiness, are detrimental to performance. As stress increases, performance is facilitated up to an optimal level beyond which there is deterioration. Thus, physiological and behavioral processes are involved in the adaptive response to stress.

Although stress is difficult to define, the responses produced by stress are not. Stressors are pervasive and challenge the physiologic and behavioral well being of the organism. All stress-inducing stimuli, both physical and psychological, exert their initial effects on the central nervous system (CNS). Thus, the stress response system is predominantly under neural control. The CNS regulates the reactions of the organism through close interactions with other neuroanatomical, peripheral and functional elements to produce adaptive responses necessary for maintaining homeostasis. The stress system is designed for quick activation and disengagement. Chronic stimulation results in maladaptive responses and a multitude of physical, emotional, and behavioral pathologies. Thus, a contemporary perspective of stress has evolved from Seyle's original concept of a single stereotypical response resulting from any stress put on the body to one that reflects different patterns of biochemical, behavioral, and physiological responses of varying degrees to particular stressors.

To this point, we have focused primarily on the nonspecific or general responses to stress. However, the past few decades of research on the neural basis of stress would suggest that many responses to stressful experience are not nonspecific and are indeed specific at many levels: from their temporal characteristics, to regional location, to molecular identity. To illustrate this point, we will discuss the responses to one specific stressor, that of acute exposure to brief, low-intensity, inescapable tail shocks in the rat. Exposure to this one experience induces a number of responses that are both regionally and temporally specific. At the transmitter level, exposure to the event enhances the release of acetylcholine in the hippocampus and prefrontal cortex but not in the nucleus accumbens. At the receptor level, exposure to the stressor activates n-methyl-d-aspartate (NMDA) receptors in the basolateral amygdala, while enhancing the affinity of the AMPA type of glutamate receptor in area CA1 but not area CA3 of the hippocampus. At the biochemical level, exposure enhances the binding of phorbol esters, a marker for protein kinase C activity. This effect occurs specifically in the basolateral nucleus but does not occur in either the central nucleus of the amygdala or the hippocampus. At the physiological level, exposure to the acute stressful event reduces spontaneous unit activity in the basolateral amygdala, while enhancing theta rhythms and reducing long-term potentiation in hippocampal subregions. At the molecular level, activation of immediate early gene expression occurs in the several layers of the cortex and area CA1 of the hippocampus but not in the dentate gyrus. At the cognitive level, exposure to the stressor has been reported to impair instrumental responding while enhancing classical conditioned responses.

These responses to stress are also dependent on the sex of the animal. For example, exposure to the stressor, while enhancing classical conditioned responses in the male rat, dramatically impairs conditioning in the female rat. Finally, many of these responses to stress are immediately induced while others are persistently expressed. Moreover, a subset of these responses can be reactivated long after the stressful experience has ceased by simply placing the animal in the context where the stressful event occurred. Thus, there is a psychological component (or memory) of the experience that is sufficient to activate many indices of stress in the brain. From this one example, it is clear that the responses to stress can be both regionally and temporally specific as well as sex specific. Understanding the neural basis of stress will require an appreciation and documentation of these specifics.

See also: Anxiety and Fear, Neural Basis of; Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of

Bibliography

Cannon W B 1929 The wisdom of the body. *Physiological Review* **9**: 399–431

Hebb D 1955 Drives and the CNS (conceptual nervous system). *Psychological Review* **62**: 243–54

Hennessy J W, Levine S 1979 Stress, arousal, and the pituitary-adrenal system: A psychoendocrine hypothesis. *Progress in Psychobiology and Physiological Psychology* **8**: 133–78

McEwen B S, Sapolsky R M 1995 Stress and cognitive function. *Current Opinions in Neurobiology* 205–16

Sapolsky R M 1992 *Stress, the Aging Brain, and the Mechanisms of Neuron Death*. MIT Press, Cambridge, MA

Seyle H 1956 *The Stress of Life*. McGraw Hill, New York
Shors T J 1998 Stress and sex effects on associative learning: For better or for worse. *The Neuroscientist* **4**: 353–64

T. J. Shors and B. Horvath

Stress: Psychological Perspectives

Psychological stress, like the general term 'stress,' is a common term that defies a universally agreed-upon definition. These difficulties are not due to a lack of interest in the concept; rather, they are due to a widespread fascination with psychological phenomena related to organism–environment interactions. It is the inherent complexity of this enterprise that attracts and vexes investigators. In the place of a standard definition of psychological stress, three different perspectives have evolved with their own definitions, as well as their own conceptual and methodological traditions.

1. *Historical Considerations and Concerns*

Interest in the implications of emotions for health is age-old, dating back to the origins of holistic approaches to medicine exemplified by the writings of Hippocrates (400 BC). Interest in the influence of psychological factors on emotional functioning has consequently represented a parallel theme of interest over the centuries (see *Environmental Stress and Health*). The basic premise is that stress gives rise to distressing emotions, which in turn influence disease. Yet the connections between these words and ideas are quite rich, and more complex than current viewpoints suggest. Indeed, 'stress' is derived as a shortened form of 'distress'; disease, in turn, originally referred to 'disease,' or distress (Rees 1976). It is in this lexicographic context that varied connections between stress, distress, and disease can be appreciated, only some of which are useful for scientific purposes.

As rich in explanatory power as psychological stress might be for understanding the causes of distress and disease, the linguistic overlap of these terms also signifies problems. The intuitive appeal of the ideas and their relations may foster false theories and findings. There is a tendency to 'explain away' disorders of unknown etiology based on prevailing beliefs about the unwanted or adverse social or psychological circumstances (Monroe and McQuaid 1994). For example, before the physical causes of particular infectious diseases were discovered, the origins of these illnesses were commonly ascribed to psychological stress and distressing emotions (Sontag 1978). The intuitive appeal of explanations based on psychological stress and related ideas requires firm thinking and sound science to differentiate valid associations from plausible and convenient fictions.

2. *The Major Perspectives and Traditions*

Two traditions may be discerned which, together, have given rise to three perspectives on psychological stress. The first tradition derives from clinical observations and field studies of humans exposed to threatening circumstances. The second tradition arises from the animal laboratory, where systematic control over stimulus conditions and response alternatives of the organism has been exerted. The common denominator of the three perspectives on psychological stress emerging from these traditions is an interest in the psychological concomitants and sequelae associated with challenging or threatening environmental conditions. The differences between these perspectives, as explained next, is the relative emphasis placed on the environment, organism, or environment–organism interactions over time as a defining feature of psychological stress.

2.1 *The Stimulus or Environmental Perspective*

The stimulus or environment perspective holds that psychological stress is anchored in the environment and its characteristics. The underlying assumption is that objective external circumstances can be classified as psychologically demanding, threatening, benign, and so forth. In human field or laboratory research, the stimulus parameters were recorded in naturalistic contexts or manipulated in laboratory settings. Animal laboratory research, on the other hand, could readily control the environment, and thereby more easily operationalize and manipulate the stimulus conditions.

Other factors, too, were influential in promoting the environmental perspective. Attention to the environment for creating stress and illness became especially evident toward the mid-twentieth century; the experiences of World War II reshaped the ways in which a breakdown in human functioning was viewed. Prior to this time, it was thought that individual vulnerability (i.e., biological) largely accounted for who succumbed to stress. Normal people maintained health and survived the ravages of stress; the biologically predisposed or tainted broke down. Given that one-third of American casualties in World War II sustained no physical injury (Weiner 1992), the idea that severe stress could precipitate psychological and physical breakdown in previously normal individuals was incorporated into medical and psychiatric thinking (Dohrenwend 1998). This brought to the foreground

the importance of extreme environments for defining psychological stress.

Finally, a surge of interest and activity in psychological stress for an environmental perspective came from another avenue of inquiry: life events questionnaires. Thomas Holmes and Richard Rahe developed the idea that life events could be assessed in an objective manner. These researchers developed a list of 43 events observed to occur prior to the onset of disease, including such items as marriage, trouble with the boss, change in sleeping habits, death in the family, and vacation. The events were collated and weighted, producing the Schedule of Recent Experiences (SRE; Holmes and Rahe 1967). Literally thousands of studies were conducted with the SRE and the many derivative life event checklists it spawned (Monroe and McQuaid 1994).

2.2 The Response Perspective

The second perspective on psychological stress can be traced to the writings of Darwin and the theory of natural selection. It is in these ideas that the attention began to focus on the implications of challenges and dangers of the environment, forcing the organism's *adaptation* to ever-altering external circumstances. The response features of stress take on special importance in this framework, as it is recognized that evolution also fostered skills in *anticipation* of dangers and adversity (Weiner 1992). Being able to prepare for and prevent problems in the future drew upon mental, emotional, and behavioral capabilities that, over time, enhanced survival and reproduction. From this perspective, psychological stress is viewed as the *response* of the organism to the complex adaptive demands of the environment.

The twentieth century brought more sustained and intense interest in probing the intricacies of psychological stress, with an emphasis on the response perspective. The work on physiology by Walter Cannon posed questions about how the body maintained its regulatory functions in the face of changing environmental demands. Hans Selyé introduced the term 'stress' into more formal scientific arenas, and investigated the physiology of organisms in response to acute and extreme environmental conditions.

2.3 The Environment–Organism Interaction Perspective

The stimulus and response perspectives on psychological stress each possess certain limitations. For example, knowledge of the environment does not always predict the responses; individual differences in response to similar stimuli are a common finding. From a response perspective, there are numerous indicators to draw upon (different emotions, psychophysiologic, or hormonal indices), many of which displayed different relationships with distinct environmental challenges (Weiner 1992).

To address some of these limitations, psychological stress has been proposed to represent a particular *relationship* between organism and environment. This third perspective holds that psychological stress is the product of the individual's ongoing transactions with ambient circumstances, particularly circumstances that are perceived as taxing or exceeding adaptive resources. Several possibilities are entertained within this perspective, ranging from an emphasis on stressful *experiences* (Weiner 1992) through emphases on the ongoing process of event occurrence, appraisal, coping, reappraisal, and so on (Lazarus and Folkman 1984). The core feature of this perspective is that of denoting an *active process* between a dynamic environment and an adapting organism over time.

Broadening the definition of psychological stress to encompass the interactions between organism and environment over time also enlarged the conceptual scope of research. This shift raises additional considerations, to which we now turn.

3. Conceptual Challenges and Problems

After the 1950s, stress became a common term, cutting across sociology, psychology, medicine, physiology, and biology. A broad array of constructs were united under this common rubric. At least two issues arise from this explosive growth in research on stress: (a) individual differences, and (b) generality vs. specificity of psychological stress.

3.1 Individual Differences

Despite the promising leads of early research on psychological stress, it became apparent that there was individual variability in response to stressful circumstances. Not all people succumbed to even the most extreme of situations, while some people broke down in the face of apparently minimal demands. Theory and research needed to account for such variation, thereby introducing factors hypothesized to moderate stress effects.

Many factors have been proposed as stress moderators. Attention to this matter invoked such individual difference variables as appraisal of threat and coping (Lazarus and Folkman 1984, Monroe and Kelley 1995) and more remotely a variety of other considerations believed to moderate the effects of stress (e.g., personality factors, social support, etc.; Brown and Harris 1989, Goldberger and Breznitz 1993). The net effect was to broaden the conceptual arena of stress research considerably and to introduce multifactorial models, taking into account a host of factors that could influence stressors and responses

over time (see *Coping Across the Lifespan*; *Control Behavior: Psychological Perspectives*; *Social Support and Stress*; *Stress and Coping Theories*).

3.2 Generality vs. Specificity of Psychological Stress

A second concern was related to the omnibus nature of the stress concept. On the one hand, the stress united a variety of cognate constructs and emphasized the general nature of stress (e.g., Selyé 1976). On the other hand, the unique implications of specific environmental challenges and responses could be obscured within such a general viewpoint (e.g., Weiner 1992). Recent research from both the animal laboratory and human field studies suggests that responses to stressors can be very specific with respect to the particular demands imposed, with specificity being key for understanding many of the adaptive implications (Dohrenwend 1998, Weiner 1992) (see *Stress, Neural Basis of*).

Overall, the challenge for research on psychological stress was to remain true to its roots in extracting the 'essence' of psychological phenomena related to adversity, while at the same time not becoming so overly inclusive as to become essentially meaningless or obscure other underlying associations. Such an agenda raises important methodological issues, as discussed next.

4. Methodological Considerations

Measuring psychological stress in a reliable and valid manner has been a continuing challenge for stress researchers. For example, animal laboratory research has proven useful for more precise specification of both stimulus parameters and response characteristics associated with psychological stress. Yet the methods adopted often constrained the array of possible responses, creating unnatural stressor conditions which, in turn, yielded misleading views about how adaptation proceeds under more ecologically valid conditions (Weiner 1992) (see *Stress Management Programs*).

In terms of human field research, there has been considerable debate about measurement procedures for assessing stressful life events. Research on self-report life event checklists such as the SRE revealed serious psychometric limitations (Brown and Harris 1989, Monroe and McQuaid 1994). Naïve research designs also led to spurious findings. For example, being depressed could cause life events such as 'trouble at work,' and 'difficulties with spouse,' rather than the life events causing subsequent depression. Additionally, many of the life events on the SRE and similar checklists were direct indicators of disorder or illness (i.e., 'major change in eating habits,' 'major change in sleeping habits'). Although more refined and comprehensive measures were developed to address such concerns, fundamental problems with self-report checklists remain (Brown and Harris 1989, Monroe and McQuaid 1994).

In response to these methodological concerns, other investigators designed semistructured interviews and developed explicit guidelines, decision rules, and operational criteria for defining and rating life events (Brown and Harris 1989, Dohrenwend 1998). In general, these procedures enhance the reliability of life event assessments and provide better prediction of outcomes. Although these approaches are more time- and labor-intensive, they represent among the best available tools for measuring psychological stressors. Importantly, too, these approaches provide flexibility in measuring life stress, allowing for ratings of specific dimensions of stress and for taking into account the influences of moderators of stress impact. Thus, in theory, such approaches can help to bridge the gap between a strictly stimulus perspective, and to incorporate useful features of the response and organism–environment interaction perspectives.

5. Future Directions

The three perspectives on psychological stress are converging on a viewpoint of psychological stress as a comprehensive *process* of complex transactions over time between organism and environment. Yet each approach continues to be of value for emphasizing particular features of these complex models that do not receive adequate attention within the other perspectives.

The stimulus perspective is essential to provide valid measures of the stressor characteristics of environments. Without standardized indices reflecting the input characteristics of the circumstances animals or people face, little can be said about 'downstream' elements of the models. For example, individual differences in appraisal or coping have little meaning without considering what is being appraised or coped with. It is in relation to well-specified environmental conditions that the implications of individual differences in appraisal and coping may be most readily understood in terms of the impact on health and well-being (Monroe and Kelley 1995). Future work may benefit from further specification and standardization of the stressor characteristics associated with adverse environments and their relations to varied outcomes (Monroe and McQuaid 1994).

The response perspective suggests a number of avenues to enrich future inquiry. With respect to animal laboratory research, simple singular response indices are being replaced by more complex patterns of behavioral and biological effects. These may range from the immediate 'exquisitely specific' neuroendocrine response profiles through long-term structural changes in the central nervous system (Weiner 1992). With

respect to human field studies, retrospective recall of conditions and emotions may be enhanced through such developments as ecological momentary assessments. These procedures target multiple, immediate reports from people in their usual environments. These methods avoid many of the pitfalls of procedures based on the retrospective reconstruction of events and experiences (Stone et al. 1999).

Perhaps most promising but challenging is the interactional perspective. Formidable challenges in operationalizing key constructs exist (e.g., appraisal and coping; Monroe and Kelley 1995, Cohen et al. 1995). Yet interesting progress has been made recently, employing sophisticated longitudinal research designs with in-depth, microanalytic techniques evaluating day-to-day transactions between the person and environment (Lazarus 2000). Further work along these lines is essential for differentiating the facts from the fictions in delineating the implications of psychological stress for health and well-being.

See also: Coping Across the Lifespan; Job Stress, Coping with; Social Support and Stress; Stress and Coping Theories; Stress and Health Research; Stress at Work; Stress in Organizations, Psychology of; Stress Management Programs; Stress, Neural Basis of

Bibliography

Brown G W, Harris T O (eds.) 1989 *Life Events and Illness*. Guilford Press, London

Cohen S, Kessler R C, Gordon L U (eds.) 1995 *Measuring Stress: A Guide for Health and Social Scientists*. Oxford University Press, New York

Dohrenwend B P 1998 Overview of evidence for the importance of adverse environmental conditions in causing psychiatric disorders. In: Dohrenwend B P (ed.) *Adversity, Stress, and Psychopathology*. Oxford University Press, New York, pp. 523–38

Goldberger L, Breznitz S (eds.) 1993 *Handbook of Stress: Theoretical and Clinical Aspects*, 2nd edn. The Free Press, New York

Holmes T H, Rahe R H 1967 The social readjustment rating scale. *Journal of Psychosomatic Research* **11**: 213–18

Lazarus R S 2000 Toward better research on stress and coping. *American Psychologist* **55**: 665–73

Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York

Monroe S M, Kelley J M 1995 Measurement of stress appraisal. In: Cohen S, Kessler R C, Gordon L U (eds.) *Measuring Stress: A Guide for Health and Social Scientists*. Oxford University Press, New York, pp. 122–47

Monroe S M, McQuaid J R 1994 Measuring life stress and assessing its impact on mental health. In: Avison W R, Gotlib I H (eds.) *Stress and Mental Health: Contemporary Issues and Prospects for the Future*. Plenum, New York, pp. 43–73

Rees W L 1976 Stress, distress, disease. *British Journal of Psychiatry* **128**: 3–18

Selyé H 1976 *The Stress of Life*, rev. edn. McGraw-Hill, New York

Sontag S 1978 *Illness as Metaphor*. Farrar, Straus and Giroux, New York

Stone A A, Shiffman S S, DeVries M W 1999 Ecological momentary assessment. In: Kahneman D, Diener E, Schwarz N (eds.) *Well-being: The Foundations of Hedonic Psychology*. Russell Sage Foundation, New York

Weiner H 1992 *Perturbing the Organism: The Biology of Stressful Experience*. The University of Chicago Press, Chicago

S. M. Monroe

Stressful Medical Procedures, Coping with

Many moderately invasive, potentially life saving procedures, e.g., colonoscopy to examine the colon and remove polyps, can be performed on an outpatient or 'same-day' basis. Many patients endure these procedures with minimal expressions of discomfort, while others experience and manifest moderate to high levels of pain and distress. Laboratory and clinical studies designed to assess the effects of psychological preparation in reducing distress have increased our understanding of the processes underlying these individual differences in three ways. First, the pain experience is an integration of a sensory experience and an emotional, distress response; pain is not just a sensation that is proportional to the extent of injury. Second, the type and intensity of the emotional response, e.g., fear-distress, is dependent on the patient's representation of the experience, i.e., how it is perceived and interpreted, and the perceived availability of a means of controlling it. Finally, the patient's representation of a procedure is based on information from doctors and nurses and by the ideas, beliefs, memories, and perceptions based on their own experience and observations of others.

1. Pain is Sensation and Emotion

Pain has often been conceptualized as a sensory process (see *Pain, Health Psychology of*). Thus, an injury was thought to activate receptors that sent pain messages to the brain, and pain and distress were expected to be directly proportional to the extent of injury (Leventhal and Everhart 1979). Patients who reported or showed greater discomfort than expected in response to an injury or to a treatment were assumed to have personalities that predisposed them to be excessively sensitive to pain signals or to be malingerers. Observations of wounded servicemen during World War II (Beecher 1946) contradicted the sensory model in two ways. The first was the absence of pain and distress when survival was enhanced by active coping. The second was the failure to alleviate pain and distress with analgesics such as morphine supposed to block conduction of sensory nerves. Administration of barbiturates to reduce emotional

reactions of fear and anxiety reduced pain-distress, suggesting that emotional reactions were a component of the pain response that could amplify pain reactions.

Anatomical and physiological studies as well as behavioral data show that the pain-distress experience is a product of parallel and interacting sensory and emotional systems (Melzack and Wall 1970). Situational factors that increase emotional reactions and amplify distress include the unexpectedness of the noxious sensations, the perception that more serious consequences are yet to follow, and the perception that the sensations are not controllable. Inhibition of the emotion system and minimization of pain-distress occurs when the sensations are expected, perceived as benign (imply no serious consequences), and can be controlled. Two mechanisms appear to be involved in this inhibitory process: the body's opiates, the endorphins, and direct, neural inhibition of pain centers (Strausbaugh and Levine 2000).

2. *Distress Reduction in Laboratory Settings*

A classic study by Egbert and colleagues (1964) demonstrated that postsurgical distress could be reduced by preparatory information; it triggered both laboratory and clinical studies of preparation. The laboratory studies showed how subtle differences in preparatory information had a substantial impact on the level of pain-distress. For example, subjects about to undergo a cold pressor test kept their hands submerged in the ice water for less time if they were told they were participating in a 'Vasoconstrictive Pain Study' (painful label) than if the study was described as a 'Discomfort Study' (less painful label; Hirsch and Liebert 1998). The discomfort group also gave lower ratings to their distress. In our earlier studies, subjects told to focus on the benign, sensory experience of the ice water during cold pressor immersion (pins and needles, color of the hand, etc.) experienced less distress and showed an increase in skin temperature in comparison to subjects told the experience would be painful. Distraction often proved ineffective in blocking these noxious experiences unless it was in the same modality as the pain stimulus (auditory; skin sensation).

3. *Distress in Response to Invasive Medical Procedures*

The study by Egbert and colleagues (1964) and the laboratory studies indicated that a patient's response to an invasive medical procedure can be shaped by the way their doctors and nurses frame the somatic sensations evoked by the procedure. As Janis and Leventhal pointed out years ago (Janis and Leventhal 1968), the florid symptoms often observed among surgical patients and previously attributed to underlying psychopathology were often a product of the meaning of the situation and the availability of coping reactions. For example, although the painful sensations experienced during procedures such as endoscopy or bronchoscopy are readily attributed to the procedure, their meaning is ambiguous. They could imply that the instrument has caused damage or contacted a cancer, or suggest to that patient that he will become increasingly frightened and will be unable to retain emotional control if the pain lasts much longer. Ambiguity also can extend to post-treatment symptoms, e.g., the pain experienced after coronary bypass surgery can be identified as wound pain from a healing incision or damage to the heart. The implication of the latter is far more threatening and fear arousing than the first. In the absence of information from medical practitioners, the meaning of somatic experience will reflect the patient's limited knowledge, memories of similar experiences, anecdotes from others, and their worst fears.

Providers can provide patients with information so that they will experience procedures in a nonthreatening way and minimize pain-distress. Two types of information were given to patients prior to an endoscopic examination and each had beneficial effects: sensory information regarding what they would feel during the examination (e.g., having air pumped into your stomach will make it feel that you have just eaten a large meal) and behavioral plans (e.g., techniques for mouth breathing and swallowing to reduce gagging while swallowing the endoscopic tube). Patients given both types of information were least likely to gag when swallowing the endoscope (36.4 percent) and had lower heart rates in contrast to 90 percent of those gagging who received neither type of information. Patients given only one type of information fell in between (Johnson and Leventhal 1974). Thus, providing patients with information that minimizes the threat value of an experience reduces their anxiety and helps patients to coordinate behavioral procedures with their somatic sensations. The efficacy of this combination is well illustrated in studies of preparation for childbirth. The simple act of monitoring the sensory features of labor contractions lowered postdelivery pain reports in women who had participated in LaMaze classes and were trained in the behavioral responses needed to facilitate child birth (E. Leventhal et al. 1989). Women who were offered various ways of distracting themselves did not benefit with reduced pain from their childbirth classes. Thus, monitoring increased familiarity with body sensations, '… reduced emotional behaviors and lowered the levels of pain, by facilitating the coordination of breathing and pushing with contractions' (p. 369).

Practitioners seeking to help patients cope with stressful procedures need to be aware of the influence of current levels of pain-distress on retrospective pain reports. For example, chronic pain patients underestimated the pain they had reported on electronic

diaries during the prior week if their retrospective reports were given after physical therapy when their current pain was low (Smith and Safer 1993); patients who reported their prior pain before physical therapy overestimated their diary reports. Redelmeier and Kahneman (1996) found that retrospective ratings of the overall painfulness of colonoscopy and lithotripsy was related to the peak level of pain during the procedure and the level of pain experienced at the end of the procedure. The duration of the procedure was unrelated to overall judgments of pain. Thus, memories of pain-distress during medical procedure are biased upward if patients are experiencing much pain when making their judgments, and can be biased downward if the procedure is designed to end at a low level of pain and distress. This bias may be of clinical importance as the recall of pain may affect willingness to undergo future, protective examinations.

There is evidence that interventions to reduce pain-distress are particularly helpful for patients who are most vulnerable to upset. Baron and Logan (1993) proposed that patients who have a high desire for control during a dental, i.e., root canal, procedure (perhaps because they interpret the procedure as threatening), yet perceive themselves as having little control (because they feel incapable of coping with the threat), are highly vulnerable to upset and experience the greatest anticipatory anxiety before treatment and the greatest discomfort during it. When given sensory and procedural information that reframed the procedure in a less toxic way, these patients with high desire for control but low perceived control were no more distressed by their root-canal procedure than other, nonvulnerable patients. Vulnerable patients given only 'placebo,' filler information were highly distressed.

An important caveat is whether the factors effective for the reduction of pain-distress in response to brief procedures such as colonoscopy will be effective in reducing pain and distress for prolonged stressors such as the pain and distress experienced following surgery. Informing people about somatic sensations may be ineffective in reducing pain postsurgery (Johnston and Vogele 1993) unless the information provides clear and believable nonthreatening interpretations of these somatic events. Recent work on recovery from myocardial infarction suggests that this is critical. Behavioral techniques for distress control seem to be effective in long-term stress settings if the control is effective. For example, hospitalized patients allowed to self-administer analgesics use less analgesic and report less pain.

4. *Summary*

The pain and distress experienced during invasive procedures can be modified by effective preparation. The key constituents of such preparation appear to be nonthreatening interpretations of the sensory cues evoked during the procedure and strategies for effective situational control linked to these cues. This combination appears to be effective for regulating the emotional component of the pain-distress response. Individual differences, however, can modify the effectiveness of behavioral preparation; patients lacking a sense of efficacy in controlling either the situation or their own affective reactions may benefit more when they see trusted others in control of the procedure. Preparation may also have different effects on different outcomes, some reducing pain experience, others physiological reactions, and others the time for completing the procedure. Though studies are needed to identify the mechanisms underlying the control of pain-distress during noxious medical procedures (Strausbaugh and Levine 2000), it is clear that patients can benefit from information. A humane system will provide information with regard to what patients will experience and help them to interpret these experiences as an expected part of a healing process. It will also provide behavioral procedures to facilitate the control of these noxious sensations and their associated affects.

See also: Control Beliefs: Health Perspectives; Pain, Management of; Pain, Neural Basis of; Social Support and Recovery from Disease and Medical Procedures; Stress Management Programs

Bibliography

Baron R S, Logan H 1993 Desired control, felt control, and dental pain: Recent findings and remaining issues. *Motivation and Emotion* **17**(3): 181–204

Beecher H K 1946 Pain in men wounded in battle. *Annals of Surgery* **123**: 96–105

Egbert L D, Battit G E, Welch C E, Bartlett M K 1964 Reduction of post operative pain by encouragement and instruction of patients. *New England Journal of Medicine* **270**: 825–7

Hirsch M S, Liebert R M 1998 The physical and psychological experience of pain: The effects of labeling and cold pressor temperature on three pain measures in college women. *Pain* **77**(1): 41–8

Janis I L, Leventhal H 1968 Human reactions to stress. In: Borgatta E, Lambert W (eds.) *Handbook of Personality Theory and Research*. Rand McNally, Chicago, pp. 1041–85

Johnson J E, Leventhal H 1974 Effects of accurate expectations and behavioral instructions on reactions during a noxious medical examination. *Journal of Personality and Social Psychology* **29**(5): 710–8

Johnston M, Vogele C 1993 Benefits of psychological preparation for surgery: A meta-analysis. *Annals of Behavioral Medicine* **15**(4): 245–56

Leventhal E A, Leventhal H, Shacham S, Easterling D V 1989 Active coping reduces reports of pain from childbirth. *Journal of Consulting and Clinical Psychology* **57**(3): 365–71

Leventhal H, Everhart D 1979 Emotion, pain, and physical illness. In: Izard C E (ed.) *Emotions in Personality and Psychopathology*. Plenum, New York, pp. 263–99

Melzack R, Wall P D 1970 Psychophysiology of pain. *The International Anesthesiology Clinics* **8**: 3–34

Redelmeier D A, Kahneman D 1996 Patients memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain* **66**: 3–8
Smith W B, Safer M A 1993 Effects of present pain level on recall of chronic pain and medication use. *Pain* **55**: 355–61
Strausbaugh H J, Levine J D 2000 Pain. In: Fink G (ed.) *Encyclopedia of Stress*. Academic Press, San Diego, CA, Vol. 3, pp. 115–8

H. Leventhal

Strikes: Sociological Aspects

The withholding of labor is known to have occurred during antiquity, the Middle Ages, and the Renaissance (Crouzel 1887). It was not, however, until the spread of wage labor under industrial capitalism that strikes became a routine expression of workers' claims. The very words by which we know the phenomenon in many European languages ('strike' in English, with its derivatives: German *streik*, Dutch *strijk*, and Swedish *strejk*, the French *grèves*, the Spanish *huelga*, the Italian *sciopero*) reveal a mid- to late nineteenth century origin. Only in English, expressions such as *strike*, *strike tools*, or *strike of work* became increasingly common starting during the mid-eighteenth century, keeping pace with the industrial revolution. 'This day the hatters *struck* and refused to work till their wages are raised' we read in the *Annual Register* on 9 May 1768. By 1793, the expression must have been common enough in England for G. Dyer to write: 'The poor … seldom strike, *as it is called*, without good reasons … The colliers had struck for more wages' (*The Oxford English Dictionary*, entry for 'strike'). In French, we have to wait until 1805 to come across the first use of the word *grève* in its meaning of withholding of labor by extension of a much earlier meaning of the word as 'unemployed' or 'without work.' But it is not until 1845–1848 that the new meaning became common (Le Grand Robert de la Langue Francaise, *grève*). The Italian *sciopero* or the Spanish *huelga* did not become part of the daily vocabulary until the later half of the nineteenth century, again by adapting a word used for centuries to refer to unemployment and lack of work (*Deli-Dizionario Etimologico della Lingua Italiana*; *Diccionario Crítico etimológico castellano e hispánico*).

1. Strike Statistics

Systematic collection and publication of strike data started in most countries between 1870 and 1900. Until then, strike data were kept secret in the hands of Ministries of the Interior. After all, strikes were viewed as crimes under Combination Acts (Perrot 1974). Only when the conception of strikes shifted from one of crime to one of 'social illness' did strikes become the concern of specialized Labor Ministries. It is these Ministries that initiated systematic recording and publishing of strike data and that standardized definitions and collecting procedures. Early statistics provided rich information on each individual strike: the immediate causes of strikes, their outcomes, their organizational bases, the occurrences of violent incidents and the number of establishments affected by any given strike. There is none of that richness in today's strike statistics. Furthermore, strike data are highly aggregated (Fisher 1973). Typically, they do not go much beyond three aggregate measures of strike activity: number of strikes (S), number of workers involved (W), and number of hours lost (H).

On the basis of these three indicators, several different measures have been constructed in the literature (Forcheimer 1948, Knowles 1952, Shorter and Tilly 1974, Hibbs 1976): *frequency* or number of strikes per thousand employed workers; *size* or average number of strikers per strike (W/S); *duration* or average number of hours lost per strikers (H/W); and *volume*, defined as frequency × size × duration = (S/thousand employed workers) × (W/S) × (H/W) = (H/thousand employed workers); the Galambos' and Evans' composite strike index based on an unweighted average of S, W, and H (Galambos and Evans 1966). Graphically, these different components of strike activity can be represented in a three-dimensional space as a parallelepiped—the 'shape' of strikes—whose sides are given by the size, duration, and frequency of strikes and whose volume corresponds to the product of the three sides (Shorter and Tilly 1974, Hibbs 1976, Franzosi 1995).

2. Data Reliability

The reliability of strike statistics varies from country to country, but it is generally low (Shalev 1978). Some countries carry out data collection through specialized offices within Departments of Labor; others rely on police departments (e.g., Italy). Some use newspapers as sources of data; others rely on self-reporting, and still others base their statistics on police records. Different countries adopt different definitions of strikes, leading to difficulties in cross-national comparisons. It is generally difficult to gauge the bias of strike statistics under any data collection arrangement. Strike size can be just as easily overestimated as underestimated. If employers report strike size, they may be inclined to play down the strike to embarrass the unions or to portray a public image of peaceful labor relations (particularly vis-à-vis clients), but they may also wish to magnify the strike in order to justify future retaliatory actions or police/government in-

tervention, or to contest the imposition of penalties by clients for late delivery. Union reported figures are more likely to be overestimated in a show of organizational strength. If police offices themselves estimate the size of the strike it may be no more than pure guesswork. Strike frequency, on the other hand, is likely to be always greatly underestimated regardless of the efficiency of collection: small, short, shop-level spontaneous work stoppages, that represent the bulk of all strike occurrences, probably escape recording, as scores of field investigations have shown (Turner et al. 1967, Korpi 1978, 1981, Batstone et al. 1978).

Bias in the data is not without consequences. When single strikes across a city or an industry are recorded separately, the number of strikes is inflated but the average strike size is reduced. To the extent that strike size is related to the workers' organizational capacities, a bias in this measure would misrepresent the organizational basis of strikes (Batstone et al. 1978, p. 22). Under-reporting of small, plant-level strikes obscures the role of informal leadership and the existence of diffused, grass-root discontent, often directed against official unions as much as against employers (Korpi 1981). The systematic lack of information on strike tactics (e.g., walk-out, sit-down, checker-board, with demonstrations through the shop, marches to city hall), and on the broader spectrum of working class actions (such as slowdowns, work to the rule, overtime ban, rallies, demonstrations), fundamentally biases the nature and extent of working-class mobilization. In any case, exclusive emphasis on what workers do and on just one type of action (strike) distorts the broader social relations in which strikes occur. 'It takes two to tango,' Franzosi argues, 'it takes at least two to fight' (Franzosi 1995, pp. 15–18). Strikes emerged in the nineteenth century out of the incessant interaction among three main actors—workers, employers, and the state; it is this interaction that ultimately defined the rules and the acceptable forms of protest (Tilly 1978, p. 161). It is not possible to understand what workers do in isolation of what employers and the state do.

Data aggregation adds problems of its own. There is no way of telling which combination of strike frequency and duration produces any given figure of number of hours lost (Batstone et al. 1978, p. 23). A yearly total of 100,000 strikers could refer to 100,000 different workers or to the same 10,000 going on strike 10 different times (Batstone et al. 1978, p. 21). 'Average' measures, such as size and duration, may mask wide differences in the distribution of actual strike sizes and durations, in particular when used for international comparisons. The same 'average' size, in fact, could be the result of a few, large and official strikes for the renewal of industry-wide, collective agreements and a large number of small, plant-level strikes (such as the Swedish or the Italian case), or of a great deal of intermediate size strikes, as in the United States, where plant-level bargaining is the rule but workforce concentration is higher (Edwards 1981, p. 231). The same 'volume' can mask marked differences in each of its constituent components. Standardized indicators, such as frequency, have been computed using different deflating procedures: total civilian labor force, non-agricultural labor force, unionized labor force, employment figures, etc., leading to an unnecessary proliferation of strike indicators that stand in the way of comparability of findings (Stern 1978).

3. *Patterns of Strikes: What we Know*

The ready availability of strike data for over a century has resulted in a rich scholarly tradition of (mostly quantitative) research on strikes, focusing on the interindustry propensity to strikes (e.g., Kerr and Siegel 1954), but notably on the temporal dynamics of strikes. From as early as the beginning of the 20th century there has been a concern with the relationship between strikes and the short- and medium-term fluctuations in the level of economic activity. Several statistical studies have shown that strike frequency follows the business cycle and the movement of unemployment in particular—the higher the level of unemployment, the lower the number of strikes (see the seminal work by Ashenfelter and Johnson 1969). As American historian David Montgomery wrote: 'Large scale strikes erupted [in the early 1900 in America] whenever the level of unemployment fell off sufficiently to give the strikers a ghost of a chance of success' (Montgomery 1980, p. 94).

Indeed, the dependence of strike frequency on the business cycle is one of the clearest findings of econometric strike research. Yet, economic theories of strikes have not gone unchallenged. For one thing, economic determinants of strikes can be statistically significant without being the main or exclusive influence (Vanderkamp 1970). Furthermore, while there is strong statistical evidence in favor of the relationship between strike frequency and the business cycle, it is not so for the other strike measures: size, duration, and volume. While strike frequency is negatively related to the level of unemployment, there is evidence supporting a positive relationship for duration (e.g., Turner et al. 1967). If economic determinants seem to play a significant role only on frequency, and each strike component measures a different aspect of strike activity, what are the determinants of strike size, duration, and volume?

No doubt, it is workers' organizational capacities that have most deeply affected the overall shape of strikes, size and duration, in particular. As early as 1911–13, French economist March showed the close statistical connection between strike size and unionization. But it is Shorter and Tilly (1974) who most systematically investigated the effect of organization on strike shape in what to date remains a landmark in strike research. For one thing, national, large-scale

strikes are only possible in the presence of national, large-scale labor organizations. But even at the local level, the organizational basis of trade unions helps to bring out on strike a larger portion of the workforce, leading to larger average strike sizes. Labor organizations have also contributed to reduce the average strike duration. The rise of labor organizations has typically shifted the timing of strikes from economically 'bad' times to 'good' times. Through organization, as Hobsbawm put it, industrial workers learned 'the common sense of demanding concessions when conditions are favorable, not when hunger suggests' (Hobsbawm1964, pp. 126, 144). By timing strikes during periods of prosperity unions also contributed to shortening the average duration of strikes. In fact, strikes are more likely to be long and drawn-out during periods of recession when employers are under no pressure to seek a quick settlement.

Changes in the timing of strikes and the consequent shortening of their average duration have also brought about changes in the likelihood of successful settlements and of the occurrence of violence during a strike. Long disputes are likely to be unsuccessful for two reasons: (a) because they tend to occur in economically unfavorable times and (b) because they have a greater likelihood of being violent (and violent strikes decrease the chances of success) (Snyder and Kelly 1976).

Yet the statistical significance of a unionization variable in econometric models of strike size hardly offers a causal explanation. After all, the reverse causal order has produced statistically significant results in union growth models, where unionization becomes the 'dependent' variable and number of strikers the 'independent' one (Ashenfelter and Pencavel 1969, Bain and Elsheikh 1976). In the end, we may have done no more than observe a strong correlation (Goetz-Girey 1965, pp. 146–7, Edwards 1981, p. 9). The exclusive reliance on highly aggregated strike and unionization figures in econometric work has obscured the causal mechanisms in the relationship between these different aspects of mobilization processes; it has also exaggerated the role of formal organization at the expense of informal shop-floor leadership (Batstone et al. 1978). Indeed, the use of temporally disaggregated firm-level data shows how rises in unionization follow (rather than precede) surges in the level of strikes, successful strikes in particular (Franzosi 1995, pp. 124–34).

The political position of labor in national power structures has also deeply affected the shape of strikes. Both strike volumes (Hibbs 1978) and strike sizes (Korpi and Shalev, 1980) have gone down in countries where labor-oriented social-democratic parties acquired stable and lasting control over governments, such as in the Scandinavian social democracies and in Austria (Korpi and Shalev, 1980 p. 320). Direct access to political power offers labor alternative and less costly means to achieve a more favorable distribution of resources: the government machinery itself, rather than the strike (Shorter and Tilly 1974, Hibbs 1976, Korpi and Shalev, 1980). When the working class has control of the government, the locus of conflict over the distribution of resources, national income in particular, shifts from the labor market and the private sector, where strike activity is the typical means of pressure, to the public sector, where political exchange prevails. Short-lived labor governments, on the other hand, or outright exclusion from the government of communist parties, are associated with higher, rather than lower, levels of strike activity (Hibbs 1976, Paldam and Pedersen 1982). Franzosi's work on postwar Italy shows that the 'historic compromise' strategy of the Italian Communist Party drastically changed the shape of strikes between 1975 and 1978 during the Governments of National Solidarity (Franzosi 1995).

4. *The Withering Away of Strikes and Cycles of Strikes*

In the 1950s arguments about the 'embourgeoisement' of the working class and the 'end of ideology' were popular. Postwar prosperity and the institutional mechanisms for conflict resolution and collective bargaining, the arguments went, have resulted in permanent industrial peace, 'the withering away of the strike, the virtual disappearance of industrial conflict in numerous countries' (Ross and Hartman 1960, p. 6, Dahrendorf 1959, pp. 257–79, Kornhauser 1954, pp. 524–5). History was to prove these arguments wrong, with the resurgence of conflict in the 1960s to unprecedented levels in those same numerous countries (Crouch and Pizzorno 1978). But once again the 'virtual disappearance' of strikes in the 1980s and 1990s rekindled the debate on the withering away of strikes, with a novel, and perhaps more powerful trump to the argument: the very working class has now virtually disappeared, with the drastic shrinking of manufacturing and the soaring of services. Since service workers are largely white collar, with a high percentage of women—both actors being traditionally little strike prone—changes in the occupational structure should result in the 'withering away of strikes' once and for all.

These arguments are compelling. No doubt the change in the location of conflict from industry to services has deeply changed the nature of labor conflicts. For one thing, it has changed the very nature of the two traditional actors involved in labor conflicts: workers and employers. The typical worker is no longer male and blue-collar, and the typical employer is public rather than private: the state has become the largest service employer in most Western countries (for comparative evidence, see Tiziano 1987). Historically, the state intervened in labor disputes in-

directly, regulating conflict via legislative means, only rarely taking a more direct role (either to repress or to arbitrate). As employer the state cannot escape being drawn directly into service disputes as a principal bargaining actor. Second, the 'essential' nature of many services (transportation, health care, etc.) and the monopoly conditions under which most services are provided has dramatically increased strikers' disruptive power (i.e., their ability to inflict severe losses to the counterpart at minimal cost to strikers). The more centralized and monopolistic a service is, and the larger the number of users affected by a service strike. Disruption is always intrinsic to service disputes, regardless of the intentions of the actors. The law-and-order potential of highly disruptive strikes further draws the state into service disputes, even when these disputes are located in the private sector.

But from here to argue in favor of the 'withering away of strikes' is to ignore fundamental characteristics of strikes. First, strikes, as we have seen, are closely linked to wage labor. The nature of employment (wage labor) does not fundamentally change with changes in the occupational structure. Technicians, teachers and college professors, nurses and doctors, pilots and air controllers have shown to be as willing to go on strike as blue-collar workers. Second, strikes are a cyclical phenomenon. Strike research has made that quite clear (Franzosi 1995, pp. 345–8). Short-term seasonal fluctuations weave in and out of medium-term fluctuations of some years duration, mostly related to the business cycle, which in turn ziz-zag around long-term fluctuations of some 40–50 years in length. Research on these long-term cycles of strikes suggests that these 'strike waves' are linked to the long-term movements of capitalist economies (the so-called Kondratieff's cycles; Mandel 1980, Cronin 1980). During wave years the characteristics of strikes are different than in surrounding years: strikes tend to be longer; they involve a higher proportion of the workforce, both at the plant and at the industry level, and a higher number of establishments (Hobsbawm 1964, p. 127; Shorter and Tilly 1974, pp. 142–6; Franzosi 1995). Strike tactics become more radical and innovative during strike waves, being mainly plant-based and in the hands of informal leaders, outside the control of the union and the institutionalized rules of the game. Demands become radicalized as well. Strike waves also often bring about legislative reforms (Shorter and Tilly 1974, p. 145; Franzosi 1995, pp. 302–6). Finally, the resurgence of conflict after long periods of peace is typically sudden and overwhelming. It takes employers, the state, and even the unions by surprise (French May of 1968 *docet* or the Italian *autunno caldo* of 1969).

If these theories are correct, the decline in strike activity in recent decades is only the result of a long-term cycle rather than a permanent downward trend leading to the withering away of strikes. If the past is anything to go by, the mobilization process will catch us by surprise with its sudden, sharp, and sweeping bouts of conflict: 1840s, 1880s, 1920s, 1960s … we may not have to wait long to be surprised.

See also: Conflict Sociology; Industrial Relations and Collective Bargaining; Industrial Relations, History of; Industrial Sociology; Labor History; Labor Movements, History of; Labor Unions; Work and Labor: History of the Concept; Work, Sociology of; Working Classes, History of

Bibliography

Ashenfelter O, Johnson G E 1969 Bargaining theory, trade unions, and industrial strike activity. *American Economic Review* **59**(40): 35–49

Ashenfelter O, Pencavel J H 1969 American trade union growth: 1900–1960. *Quarterly Journal of Economics* **83**: 434–48

Bain G S, Elsheikh F 1976 *Union Growth and the Business Cycle*. Blackwell, Oxford, UK

Batstone E, Boraston I, Frenkel S 1978 *The Social Organization of Strikes*. Blackwell, Oxford, UK

Cronin J E 1980 Stages, cycles and insurgencies: The economics of unrest. In: Hopkins T K, Wallerstein I (eds.) *Processes of the World-system*. Sage, Beverly Hills, CA, pp. 101–18

Crouch C, Pizzorno A (eds.) 1978 *The Resurgence of Class Conflict in Western Europe Since 1968*. Holmes and Meier, New York

Crouzel A 1887 *Etude historique, économique, juridique sur les coalitions et les grèves dans l'industrie*. A. Rousseau, Paris

Dahrendorf R 1959 *Class and Class Conflict in Industrial Society*. Stanford University Press, Stanford, CA

Edwards P K 1981 *Strikes in the United States 1881–1974*. Blackwell, Oxford, UK

Fisher M 1973 *Measurement of Labour Disputes and Their Economic Effects*. OECD, Paris

Forcheimer K 1948 Some international aspects of the strike movement. *Bulletin of the Oxford Institute of Statistics* **10**: 9–24

Franzosi R 1995 *The Puzzle of Strikes: Class and State Strategies in Postwar Italy*. Cambridge University Press, Cambridge, UK

Galambos P, Evans E W 1966 Work-stoppages in the United Kingdom, 1951–1964: A quantitative study. *Bulletin of the Oxford University Institute of Economics and Statistics* **28**: 33–55

Goetz-Girey R 1965 *Le mouvement des grèves en France*. Editions Sirey, Paris

Hibbs D Jr 1976 Industrial conflict in advanced industrial societies. *American Political Science Review* **70**(4): 1033–58

Hibbs D Jr 1978 On the political economy of long run trends in strike activity. *British Journal of Political Science* **8**(2): 153–75

Hobsbawn E 1964 *Labouring Men*. Basic Books, New York

Kerr C, Siegel A 1954 The interindustry propensity to strike. An international comparison. In: Kornhauser A, Dubin R, Ross A (eds.) *Industrial Conflict*. McGraw-Hill, New York, pp. 189–212

Knowles K G J C 1952 *Strikes, BA Study in Industrial Conflict*. Blackwell, Oxford, UK

Kornhauser A 1954 The undetermined future of industrial conflict. In: Kornhauser A, Dubin R, Ross A (eds.) *Industrial Conflict*. McGraw-Hill, New York, pp. 519–26

Korpi W 1978 *The Working Class in Welfare Capitalism*. Routledge and Kegan Paul, London
Korpi W 1981 Unofficial strikes in Sweden. *British Journal of Industrial Relations* **19**(1): 66–86
Korpi W, Shalev M 1980 Strikes, power, and politics in the Western nations, 1900–1976. In: Zeitlin M (ed.) *Political Power and Social Theory*. JAI Press, Greenwich, CT, Vol. 1, pp. 301–34
Mandel E. 1980 *Long Waves in Capitalist Development*. Cambridge University Press, Cambridge, UK
Montgomery D 1980 Strikes in nineteenth-century America. *Social Science History* **4**(1): 81–104
Paldam M, Pedersen P 1982 The macroeconomic strike model: A study of seventeen countries, 1948–1975. *Industrial and Labor Relations Review* **35**(4): 504–21
Perrot M 1974 *Les ouvriers en grève* (*France 1871–1890*). Mouton, Paris
Ross A H, Hartman P T 1960 *Changing Patterns of Industrial Conflict*. Wiley, New York
Shalev M 1978 Lies, damned lies, and strike statistics: The measurement of trends in industrial conflicts. In: Crouch C, Pizzorno A (eds.) *The Resurgence of Class Conflict in Western Europe Since 1968. Volume 1. National Studies*. Holmes and Meier, New York, pp. 1–20
Shorter E, Tilly C 1974 *Strikes in France, 1830–1968*. Cambridge University Press, Cambridge, UK
Snyder D, Kelly W R 1976 Industrial violence in Italy, 1878–1902. *American Journal of Sociology* **82**(1): 131–62
Stern R N 1978 Methodological issues in quantitative strike analysis. *Industrial Relations* **17**(1): 32–42
Tilly C 1978 *From Mobilization to Revolution*. Addison-Wesley, Reading, MA
Tiziano T (ed.) 1987 *Public Service Labour Relations: Recent Trends and Future Prospects*. International Labour Organisation, Geneva
Turner H A, Clack G, Roberts G 1967 *Labour Relations in the Motor Industry*. Allen & Unwin, London
Vanderkamp J 1970 Economic activity and strike activity. *Industrial Relations* (*University of California*, *Berkeley*) **9**(2): 215–30

R. Franzosi

Strong Program, in Sociology of Scientific Knowledge

The 'Strong Program' is a set of methodological claims about the proper way to conduct sociological enquiry into the nature of knowledge, including scientific knowledge. The program enjoins the investigator to adopt four requirements, which may be labeled, respectively, (a) causality, (b) impartiality, (c) symmetry, and (d) reflexivity. 'Causality' means that the aim is not merely descriptive or interpretive but also explanatory. The concern is with what brings about and sustains a body of knowledge. 'Impartiality' means that the investigator disregards his or her sense of whether the beliefs under scrutiny are true or false, or rational or irrational. Equal curiosity should attend both sides of the evaluate divide. 'Symmetry' means that this equally distributed curiosity should issue in the same general kinds of sociological explanation. All beliefs confront the same general problem of credibility and depend on the same contingencies to solve the problem. True beliefs have no more intrinsic credibility than false ones. Finally, 'reflexivity' means that sociologists of knowledge must not claim access to any transcendent standpoint, or engage in special pleading, to justify their own claims. Put simply: no sociological theory of knowledge is acceptable unless it applies to itself. These requirements all stem from the desire that the approach should itself be 'scientific' in spirit, although it is widely perceived by critics as antiscientific. To offset such misunderstandings each of the points calls for amplification.

1. *Causality*

The commitment to causality does not mean that only sociological causes are at work. It is taken for granted that belief systems are the achievements of biological entities that interact with one another and their material environment. People have their eyes open, and sensory inputs influence individual and collective belief. Unless they had effective sense organs and some innate cognitive propensities, they could neither create nor transmit a shared culture. Knowledge cannot be 'merely' or 'purely' social. Sociologists can therefore only contribute partial explanations of the overall phenomenon of cognition. They must work alongside, say, sensory psychologists—but no adequate account of human knowledge can ever proceed without due acknowledgment of the role of the social. The 'strength' of the program lies in the necessity, not the sufficiency, of the social dimension.

The idea that material reality plays little or no role in science is a form of philosophical 'idealism.' Its association with the 'Strong Program' has been solely through the imputations of critics. For example, Latour (1993) says the program is one of a family of theories based on the idea that knowledge has two ingredients ('nature' and 'society') which trade-off against one another—the more of one meaning the less of the other. Within this framework the 'Strong Program' is said to be an extreme case where one of the ingredients, nature, actually makes a nil contribution. The strength of the program is thus wrongly said to reside in the claim that knowledge (and indeed reality) is purely social.

This false view of the 'Strong Program' is widespread. The sociologist Cole (1992) who himself accepts the zero-sum picture, seeks to distance himself from the program by insisting that nature plays some role rather than no role. He suggests that it is an

empirical question to decide just how much influence is exerted by nature by comparison to that exerted by society. A variant of the error also underlies Gottfried and Wilson's (1997) article in *Nature*. They claim that the program is incompatible with science having a good record of predictive success. The allegation appears to rest on the idea that if scientists were only attending to what is going on in society, and were not attending to what is happening in nature, then their ability to make predictions would be unintelligible. That, however, is not the 'Strong Program's' claim.

The real claim of the 'Strong Program' is that scientists understand nature through and with society, not in spite of it. Society is the vehicle of understanding, not a screen or distraction. Society does not trade-off against nature, rather it enables scientific access to it. Society no more interferes with knowledge than the eye interferes with vision. They both operate as the relevant organ of cognition and as a necessary constituent of it. No one asks 'how much' of our visual experience is contributed by the eye, and 'how much' by the object seen. (We might ask 'what' is contributed, but not 'how much.') The question makes no sense, and in this framework neither does Cole's allegedly 'empirical' research program to find out how much of knowledge depends on nature and how much on society.

How do social processes enable us to have scientific access to the world? What manner of 'vehicle' or 'organ' does society provide? The answer given by Kuhn (1962) illustrates the general case. Cognition within a group is coordinated by reference to a shared exemplar or paradigm, that is, a concrete achievement on which activity is modeled. By virtue of their orientation towards that achievement, and their knowledge of its standing in the eyes of others, members of a group endow it with a normative status. Something is constituted as a paradigm in the way that money is constituted as money by being regarded as money, and a leader by being seen as and known as a leader. In this way society gets woven into the very fabric of knowledge and into the struggle to make sense of intransigent, real-world processes. Establishing a paradigm produces a qualitative change in the mode of cognition from mere individual belief to a collective relation to the world that we can recognize as scientific. Without a paradigm there may be scientists, but there is something less than science.

2. *Impartiality and Symmetry*

The reference to Kuhn puts us in a better position to understand impartiality and symmetry. He did not argue that the idea of a paradigm merely helps us make sense of old theories about natural processes, but no longer applies to current or correct theories. On the contrary, he applied the idea to all cases, that is, impartially and symmetrically, because he was searching for invariant principles. Every shared approach to nature must overcome the problem that our experience can be given more than one interpretation. The shared interpretation has to be sustained in the face of continuing difficulty, anomaly, and the threat of fragmentation into subjective opinion.

The requirements of impartiality and symmetry face a major obstacle because they run counter to some aspects of common sense. In general, our curiosity has a markedly partial and asymmetric structure. We typically confine the request for causal explanation to the breakdowns in our routine expectations. When things go smoothly we do not ask why, because nothing more needs to be said. If nothing needs to be said, perhaps nothing can be said, and we may be tempted to think that causality does not apply, or that qualitatively different kinds of causes are in operation. Thus, some philosophers have argued that while illusory perception calls for causal explanation, veridical perception does not. No professional psychologist would speak like this—just as no medical scientist would say there are causes for disease but no causes for health, and no engineer would say there are causes for why bridges collapse but none for why they stay up. If psychologists and other scientists can restructure their professional curiosity, and give it a locally symmetrical character, then this is also an appropriate posture for sociologists.

A deeper problem for symmetry comes from our capacity to follow rules. Here is the argument. Everyone accepts that the rules of a game, such as chess, are conventional and hence social. But once the rules have been agreed the question of whether or not a move conforms to the rule seems to be nonconventional. It is a matter of conceptual match or logical consistency and inconsistency. More generally, the meaning of any concept may be a matter of convention, but the question of whether an object falls under that concept seems wholly different. Meaning is conventional; truth is not. A sociologist may legitimately try to explain the choice of rules and definitions, but not, surely, matters of consistency and truth. The first is a question of causation, the latter belongs to the realm of logic and rationality. In as far as we follow rules and respond to considerations of meaning and truth our behavior must be explained rationally not sociologically. So, once again, asymmetry asserts itself.

This argument is widely accepted, but it may be challenged on the grounds that, in reality, every single act of concept application is problematic and negotiable. Not every single act is negotiated, in practice they are often routine, but as the history of science, logic and mathematics shows us, they are negotiable. Past applications of a concept do not serve to fix meanings and thereby determine in advance the norms of proper concept application or rule following. The move to the next case is always problematic. This position, known as 'meaning finitism,' follows from the determined effort to adhere to the causal and

scientific perspective. It marks an important point of contact between the 'Strong Program' and the later philosophy of Wittgenstein which centered on the analysis of rule following and the problematic character of 'sameness.'

3. Reflexivity

This requirement immediately raises the issue of relativism and self-refutation. If sociologists, own claims are socially determined and relative' why believe what they say? That the program is relativist is beyond question, but this can be seen as a strength not a weakness. Relativism (which should not be confused with idealism) is the opposite of absolutism and implies that there are no absolute justifications for any knowledge claim. All justifications end in something unjustified and merely taken for granted. But this does not mean that there is never any reason to accept what anyone says, as if it were all an ideological smokescreen. It just means that in the end reasons are local and contingent. Such an admission is not self-defeating. To draw that conclusion would be to argue in a circle from the premise that the only real reasons are absolute ones, thus begging the question on behalf of absolutism. Rather, it is the attempt to produce absolute justifications that should be viewed with suspicion.

4. Conclusion

The 'Strong Program' was formulated by an interdisciplinary group in Edinburgh in the early 1970s (see Barnes 1974, Bloor 1976). The immediate aim was to codify and clarify an emerging body of case-studies, particularly by historians of science (see Shapin 1982). The real life-blood of the sociology of scientific knowledge lies with such empirical work, for example, to name but a small sample: Mackenzie 1981) on the history of statistics and its connections with eugenics, Shapin and Schaffer (1985) on the disputes surrounding Boyle's experiments with the air-pump, Harwood (1993) on styles of research in German genetics before 1933, and Kusch (1999) on the different theories of mind generated in different psychological laboratories in the early years of the century. The program continues to serve its function, but its rich contributions to specific historical and sociological projects have been obscured to some degree by criticism of its philosophical foundations.

See also: Campbell, Donald Thomas (1916–96); Empiricism, History of; Experiment, in Science and Technology Studies; Experimentation in Psychology, History of; Kuhn, Thomas S (1922–96); Popper, Karl Raimund (1902–94); Reflexivity, in Science and Technology Studies; Reflexivity: Method and Evidence; Scientific Knowledge, Sociology of; Truth and Credibility: Science and the Social Study of Science; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Barnes B 1974 *Scientific Knowledge and Sociological Theory*. Routledge, London

Bloor D 1976 *Knowledge and Social Imagery*. Routledge, London

Cole S 1992 *Making Science. Between Nature and Society*. Harvard University Press, Cambridge, MA

Gottfried K, Wilson K G 1997 Science as a cultural construct. *Nature* **386**: 545–7

Harwood J 1993 *Styles of Scientific Thought. The German Genetics Community 1900–1933*. University of Chicago Press, Chicago

Kuhn T S 1962 *The Structure of Scientific Revolutions*. Chicago University Press, Chicago

Kusch M 1999 *Psychological Knowledge. A Social History and Philosophy*. Routledge, London

Latour B 1993 *We Have Never Been Modern*. Harvard University Press, Cambridge, MA

Mackenzie D A 1981 *Statistics in Britain 1865–1930*. Edinburgh University Press, Edinburgh, UK

Shapin S 1982 History of science and its social reconstructions. *History of Science* **20**: 157–211

Shapin S, Schaffer S 1985 *Leviathan and the Air Pump. Hobbes, Boyle and the Experimental Life*. Princeton University Press, Princeton, NJ

D. Bloor

Structural Contingency Theory

An effective organizational structure provides the coordination necessary for people to work together so that an organization can perform at a high level. There is no 'one best way,' that is, one structural type that makes all organizations perform highly. On the contrary, the effect of an organizational structure on performance is contingent upon whether it fits certain factors called the contingency factors. An effective structure for an organization is one that fits the level of its contingency factors. Research has shown that there are three main contingency factors: size, task uncertainty, and diversification. In overview, increasing size requires a more bureaucratic structure, task uncertainty requires reducing the degree of bureaucratization, and diversification requires a divisional or a matrix structure.

1. Bureaucratic Structure and its Contingencies

The structure of an organization includes how the work of the organization is divided up and allocated among people, in terms of job specialization, sections,

and departments. It also includes the number of levels in hierarchy and how many subordinates report directly to each manager, that is, his or her span of control. Another aspect of organizational structure is the degree of formalization: how far rules, standard operating procedures, and manuals govern the way people work. A further aspect of structure is the degree of decentralization, the degree to which decision-taking authority is distributed down the hierarchy (Jones 1998).

Many of these structural variables tend to be correlated so that they compose a single structural dimension of the degree of bureaucratic structure (Donaldson 2001). An organization that is low on bureaucratization has low specialization, few departments and sections, narrow spans of control, low formalization, few hierarchical levels, and high centralization. Conversely, an organization that is high on bureaucratization has high specialization, many departments and sections, wide spans of control, high formalization, many hierarchical levels, and low centralization. Organizations are distributed widely along this continuum of bureaucratic structure.

2. *Bureaucratic Structure and the Size Contingency*

Size is the number of people who must be organized, that is the number of employees or members of the organization. For high performance, the organization needs to have the degree of bureaucratic structure that fits its size (Child 1984).

An unbureaucratic structure fits an organization of small size, while a highly bureaucratic structure fits a large size. For small size, the required, unbureaucratic, structure is a short hierarchy, with most of the decisions being taken by the top manager. The top manager controls the organization directly by giving instructions (Child 1984). There is little formalization because the top manager is making decisions on each case, rather than bothering to identify abstract rules to anticipate future cases. There is low specialization and few departments and sections, because there are few people among whom the work could be divided. The span of control is narrow, because each manager must supervise subordinates who are doing a broad range of tasks.

Conversely, large size entails a bureaucratic structure that has a tall hierarchy because an additional level of management has to be introduced each time the number of employees exceeds the feasible span of control of their supervisor. Size also brings increased complexity and this combines with the tall hierarchy to force top management to delegate decisions down the hierarchy and so decentralization increases. With so many people to be organized in the large organization, the work is finely divided up among them, producing an elaborate division of labor. There is high specialization and many departments and sections. The span of control is wide, because each manager can supervise more subordinates because the subordinates are doing a limited range of tasks. There is high formalization because, with larger size, many decisions recur about similar cases, leading to codification of best practice into rules and standard operating procedures (e.g., a hiring procedure). Such formalization relieves managers from having to attend to each decision, so that administration can be delegated to clerks who follow rules, which speeds decision-making, cuts costs, and is another reason why managerial spans increase with size. Overall, top management has relinquished much direct control and instead controls the organization indirectly and impersonally, through formalization (Child 1984). As the organization grows in size over time, so it needs to increase progressively the degree of bureaucratization of its structure in order to maintain the fit that sustains high performance.

Information technologies, whether used in operations or administration, tend to reinforce the effects of size (Donaldson 2001). Information technologies increase bureaucratization, such as by increasing formalization, because they reduces uncertainty, though their effects are weaker than size. Moreover, computerization itself is really a form of bureaucratization. Work is executed according to computer programs that are rules (e.g., 'If the checking account balance is less than $500, then apply a $10 fee to every transaction'). Knowledge workers and other employees working with computers are constrained in their actions by the programs in their computers. Thus computerization is electronic formalization that allows top management more indirect, impersonal control over lower-level employees, so that top management can further increase decentralization.

Bureaucratic structure is sometimes criticized for having too many managers and administrators relative to operating personnel (e.g., workers). However, the ratio of managers and administrators to operating personnel tends to *fall* as size increases. While larger organizations have more levels in their hierarchy, their spans of control increase so as to more than compensate (Jones 1998). Thus bureaucratic structure brings economies of scale in administration (Donaldson 1996).

3. *Bureaucratic Structure and the Task Uncertainty Contingency*

While a bureaucratic structure fits low task uncertainty, a less bureaucratic structure is required to fit high task uncertainty.

The creation of effective rules relies on all the cases of the same occurrence being similar. In routine tasks this holds, allowing rules to be formulated. However, some tasks are uncertain, so that one case varies

markedly from another, and future cases differ from past ones. Some uncertainty in organizations comes from the environment, for example, competitor pricing is hard to predict. Other task uncertainty comes from innovation inside the organization, such as designing a new product or developing a new technology to distribute a service.

Whatever the source of the uncertainty, tasks high in uncertainty cannot be handled effectively by rules. Tasks of medium uncertainty can be handled effectively by the hierarchy. However, the greater the uncertainty, the more information that has to be processed by the organization. Because of the pyramidal shape of the hierarchy, it soon becomes inadequate for dealing with more than moderate uncertainty. High task uncertainty, such as involved in radically new products or services, or making many innovations at the same time, requires reliance upon experts. The high uncertainty in their tasks prevents the organization from governing their work effectively by rules. Some decision-making authority needs to be delegated to these experts, even if they are at low hierarchical levels. Similarly, managers of innovative programs need to be able to act autonomously. Thus task uncertainty is fitted by a structure that is lower in formalization and centralization (Jones 1998).

The effect of the task uncertainty on bureaucratic structure is secondary to that of size. High task uncertainty partially offsets the high formalization that large size causes by lowering formalization. Conversely, high task uncertainty reinforces the high decentralization that large size causes by further increasing decentralization. Thus a large organization with highly uncertain tasks is less formalized and more decentralized than a large organization with highly certain tasks. The reduction in formalization decreases bureaucratic structure, while the increase in decentralization increases that aspect of bureaucratic structure.

Moreover, these effects of task uncertainty are localized to specific parts of the organization. For a firm that is developing new products, the uncertainty of its tasks is greatest in some parts, such as the research and development department that is inventing the new product. In contrast, the manufacturing department of the firm focuses mainly on the routine production of existing products, so that the uncertainty of its tasks is lower. Thus the required structure of the research and development department is low in formalization, while the manufacturing department is high. Similarly, the research and development department employs qualified scientists and technologists and devolves some decision-making to them so that they exercise autonomy and creativity. Conversely, the manufacturing department employs mainly less educated personnel who lack professional qualifications and trains them to conform to the operating procedures of the firm, so the department is run in a more formalized and centrally planned way. Nevertheless, innovation entails solving some novel problems that involve all the functional departments. For example, research needs to know whether manufacturing can make their new design on its existing machines. The integration required between the functions necessitates the adoption of cross-functional project teams. These contain representatives from each of the functional departments who are brought together for collaboration (Jones 1998). The teams provide lateral linkages that run across the hierarchy and rely upon *ad hoc* internal coordination rather than planning or formalization. Therefore they are a further way task uncertainty reduces bureaucratic structure.

4. *Divisional and Matrix Structure and the Diversification Contingency*

4.1 *Functional and Divisional Structures*

Another aspect of structure is how activities and responsibilities are grouped in the organization. In a functional structure, the managers reporting to the top manager (i.e., the CEO; chief executive officer) are specialized by function, including operational functions (e.g., manufacturing). In contrast, in the divisional structure, the senior managers are specialized by their products or services, customers, or geographical areas, and each has a broad range of operational functions within their division. A divisional structure is more specialized, formalized, and decentralized than a functional structure and so divisionalization is an extension of bureaucratization (Donaldson 2001). However, the contingency causing divisionalization is primarily diversification, with size playing a minor role (Donaldson 2001) so that the contingency determination of divisionalization differs from that of bureaucratization.

Organizations vary in their levels of diversification of their products, services, customers, or geographical areas. For brevity, product diversification will be discussed here but not service and customer diversification because similar considerations apply. Increasing diversification reduces the interdependence between the tasks involved in one product and the others. This requires less coordination between products, so each can be organized as a separate subunit acting autonomously from the other subunits and from the corporate center. Consequently, the corporate center needs to contain fewer operational functions. The greater the diversification, the greater the decentralization needs to be and thus the smaller the head office. As diversification varies from low to high, the fitting structure goes from being a pure function to a pure divisional structure. At intermediary levels of diversification, the role of the central functions decreases with each level and the role of the product heads increases. Where a firm is undiversified, such as

producing only one product, then specialization by function fosters economies of scale and avoids duplication. Each function is highly interdependent on the next, so that operational coordination is needed from the CEO, producing centralization. Also, each function is a cost center, with profitability only being assessed for the firm overall.

The next level of product diversification is where a firm offers products that are vertically integrated with each other. For example, an aluminum-refining firm might vertically integrate backward by acquiring the bauxite mine that supplies its feed material. It might also vertically integrate forward by creating a plant that fabricates its refined aluminum into window frames. Each stage along the value-added chain, that is, mining, refining, and fabricating, becomes an organizational subunit, which may be termed a division. Some decisions will be taken autonomously by each division, such as the types of production equipment to purchase, so that the structure is more decentralized than a functional structure. However, a large flow of materials must be coordinated between divisions. This necessitates some restrictions on divisional autonomy and the presence of corporate-level staff with expertise in transportation and the planing of production and inventory. Thus the corporate center contains many staff additional to the administrative functions such as accounting. Whereas the fabrication division is a profit center, the mining and the refining divisions should be cost centers because their sales are to other divisions making their profitability a contentious issue. Divisional general managers need to have their bonuses tied to corporate profitability to encourage them to cooperate.

A medium level of diversification is where a firm has two or more *related* products. The relatedness between the products lies in similarities among their materials, production technologies, skills, brands, sales force, or distribution systems. Each product is organized as a division and is run quite autonomously, except where the divisions are related and so synergies may be extracted through central coordination. For example, an electronics firm has three product divisions: avionics, computers, and domestic appliances, with each focusing on their product market. But there are commonalties in the underlying technologies so that there is a large central research and development laboratory that pursues major innovations beyond those being researched in the divisions. Therefore, the corporate center contains functions beyond the merely administrative and is quite large in size. Each division is a profit center, but the transfer pricing of shared components moderates its accountability. This is also moderated by large corporate overhead charges for the central laboratories and by corporate decisions about which division receives the profitable, new products developed in the central laboratory. Divisional general managers need to have their bonuses partly tied to that of their own division and partly to corporate profitability to encourage them to cooperate with the other divisions and the corporate staff. The divisional structure boosts innovation by having functions for each product which facilitates functional interaction. However, this is at the cost of some duplication of assets across divisions. Thus, where cost pressures are paramount, the organization instead needs to use a functional structure.

A high level of diversification is where a firm has two or more *unrelated* products, e.g., a conglomerate. There are no operating synergies to be extracted, so little central coordination of operation is required and the divisions are highly autonomous. Therefore the corporate level needs no functions beyond the merely administrative and so can be small. Each division can be straightforwardly held profit accountable. Divisional general managers need to have their bonuses tied solely to the profitability of their own division.

Area diversification has similar effects. Area diversification is how far the organization contains an area that operates separately from other areas, that is, it designs, manufactures, and sells the product all within its own area. For example, area divisions are frequently found in the food industry, where low price relative to weight makes intercontinental transportation too costly. Low area diversification fits with the functional structure, while high area diversification fits with the area divisional structure.

4.2 Matrix Structures

In a matrix structure, a subordinate reports to more than one boss. Despite the inherent complexity and resulting conflict potential of matrices, they fit some contingencies. A related product organization attains both some product innovation and some cost reduction by using a project-functional matrix structure. Each project manager heads a temporary team, which fosters close interaction between different functional specialists, to facilitate product innovation. Additionally, the heads of the functions coordinate the deployment of resources across projects to reduce costs. The relatedness of the products (e.g., at Lockheed-Gelac all the products are transport airplanes) allows the same resources to be used by the projects. This resource sharing is facilitated by the projects each being at different stages so that the same resources are not required simultaneously. The project-functional structure fits medium product diversity and low area diversity so that operations are geographically proximate and there is no area diversity that needs managing.

In contrast, an area type of matrix is appropriate for an organization that requires coordination along the area dimension in addition to coordination by either function or product. For example, a multinational corporation has offices and plants scattered around the world, each reporting up to functional or product

heads. However, the corporation also needs to have a local manager coordinating each geographical area to deal with local issues, e.g., the host government. The increase in environmental diversity, caused by cultural and political heterogeneity, that is, medium area diversification, requires that the organization mirrors and manages this diversity through having a manager for each area. Thus, area type matrices fit medium area diversification. There interdependencies exist between the areas, requiring coordination through high-level functional or product heads, at the same level as the area managers, thereby creating the matrix. Where area diversification is medium and product diversification is low, then core operations need to be structured functionally and so the fitting structure is the area-functional matrix. However, where area diversification is medium and product diversification is also medium, then core operations need to be structured divisionally and so the fitting structure is the area-product matrix (e.g., Asea Brown Boveri). Medium product diversification obtains not only for related, but also for unrelated, product diversification. This is because the medium area diversification introduces links across the separate products due to their geographical co-location (e.g., being in France and so having to deal jointly with the government there). Thus, the level of product diversification has to be determined jointly with the level of area diversification, that is, interdependence is established by examining together products and areas.

The heart of structural contingency theory is the concept of fit of organizational structure to the contingency. Many of the more promising developments in structural contingency theory delve into the concept of fit, using it to illuminate organizational performance and organizational change. A major conceptual statement about fit that has influenced subsequent research is by Van de Ven and Drazin (1985). Some researchers have extended the analysis by considering the fit of an organizational structural variable to multiple contingencies (e.g., Keller 1994). Other researchers have analyzed the fit of multiple organizational structural variables to a contingency (e.g., Drazin and Van de Ven 1985). Still other researchers have analyzed the fit of multiple organizational structural variables to multiple contingencies simultaneously (e.g., Pennings 1987). Other researchers have analyzed the differential effects of fit on different aspects of organizational performance such as financial performance vs. social costs (e.g., Kraft et al. 1995). Structural contingency theory has sometimes been criticized for being static, however, on the contrary, it contains a theory of change dynamics in that fit and misfit play crucial roles in organizational structural change (Donaldson 2001). An organization in fit tends thereby to have high performance, which leads it to increase its contingencies, such as size and diversification, so that it moves into misfit and so has lower performance. An ensuing crisis of low performance causes the organization to adopt a new structure that fits the new levels of its contingencies, thereby restoring fit and performance (Donaldson 2001). Whether the performance level of a firm in misfit becomes low enough to trigger structural change depends on other causes of performance that interact with misfit. These other causes include debt and divestment and can reinforce misfit or offset its depressive effect on performance. The set of causes of organizational performance and their effects on organizational growth and adaptation have been formalized into a new organizational theory, organizational portfolio theory (Donaldson 1999).

Fit is traditionally conceptualized as a line of iso-performance, meaning that a fit of organizational structure to a low level of the contingency (e.g., small size) produces the same performance as a fit of organizational structure to a high level of the contingency (e.g., large size). Such a model fails to provide an incentive for an organization to increase its level of the contingency (e.g., size), so that there is no explanation of why an organization changes its level of the contingency, such as by growing larger. A reconceptualization of fit has been proposed in which fit is a line of hetero-performance, in which fits to higher values of the contingency variable produce greater performance than fits to lower levels (Donaldson 2001). Therefore there is an incentive for the organization to change its level of contingency such as by growing larger over time. In this way, reconsidering the effect of fit on performance allows the explanation of changes in the contingency (e.g., size) of an organization in fit, thereby rendering structural contingency theory more comprehensive and dynamic. Thus, deeper analyses of fit, both empirically and conceptually, provide a major route for the ongoing development of structural contingency theory (for developments synthesizing structural contingency theory with other organizational theories see Donaldson 1995).

See also: Authority: Delegation; Bureaucracy and Bureaucratization; Diversification and Economies of Scale; Labor, Division of; Organizational Decision Making; Organizational Size; Organizations: Authority and Power; Organizations, Sociology of; Risk: Empirical Studies on Decision and Choice; Risk: Theories of Decision and Choice

Bibliography

Child J 1984 *Organization: A Guide to Problems and Practice*, 2nd edn. Harper and Row, London; Hagerstown, MD

Donaldson L 1995 *American Anti-Management Theories of Organization: A Critique of Paradigm Proliferation*. Cambridge University Press, Cambridge, UK; New York

Donaldson L 1996 *For Positivist Organization Theory: Proving the Hard Core*. Sage, London; Thousand Oaks, CA

Donaldson L 1999 *Performance-Driven Organizational Change: The Organizational Portfolio*. Sage, Thousand Oaks, CA
Donaldson L 2001 *The Contingency Theory of Organizations*. Sage, Thousand Oaks, CA
Drazin R, Van de Ven A H 1985 Alternative forms of fit in contingency theory. *Administrative Science Quarterly* **30**: 514–39
Jones G R 1998 *Organizational Theory: Text and Cases*, 2nd edn. Addison-Wesley Publishing Company, Reading, MA
Keller R T 1994 Technology–information processing fit and the performance of R&D project groups: a test of contingency theory. *Academy of Management Journal* **37**: 167–79
Kraft K L, Puia G M, Hage J 1995 Structural contingency theory revisited: main effects versus interactions in Child's National Study of Manufacturing and Service Sectors. *Revue Canadienne des Sciences de l'Administration—Canadian Journal of Administrative Sciences* **12**: 182–94
Pennings J M 1987 Structural contingency theory: a multivariate test. *Organization Studies* **8**: 223–40
Van de Ven A H, Drazin R 1985 The concept of fit in contingency theory. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behaviour*. JAI Press, Greenwich, CT, Vol. 7, pp. 333–65

L. Donaldson

Structural Equation Modeling

In 1980, Peter Bentler (1980, p. 420) stated that structural equation modeling held 'the greatest promise for furthering psychological science.' Since then, there have been many important theoretical and practical advances in the field. So much so, in fact, that Muthén (2001) announced a 'second generation' of structural equation modeling. The purpose of this article is to provide a description of the 'first generation' of structural equation modeling and to outline briefly the most recent developments that constitute the 'second generation' of structural equation modeling. This article will not devote space to the comparison of specific software packages. For the most part, existing software packages differ in terms of the user interface. The competitive market in structural equation modeling software means that new developments available in one software package will soon be available in others.

1. *The First Generation: Definition and Brief History of Structural Equation Modeling*

Structural equation modeling can be defined as a class of methodologies that seeks to represent hypotheses about the means, variances, and covariances of observed data in terms of a smaller number of 'structural' parameters defined by a hypothesized underlying conceptual or theoretical model. Historically, structural equation modeling derived from the hybrid of two separate statistical traditions. The first tradition is factor analysis developed in the disciplines of psychology and psychometrics. The second tradition is simultaneous equation modeling developed mainly in econometrics, but having an early history in the field of genetics and introduced to the field of sociology under the name *path analysis*.

The subject of this article concerns the combination of the latent variable/factor analytic approach with simultaneous equation modeling. The combination of these methodologies into a coherent analytic framework was based on the work of Jöreskog (1973), Keesling (1972) and Wiley (1973).

2. *The General Model*

The general structural equation model as outlined by Jöreskog (1973) consists of two parts: (a) the structural part linking latent variables to each other via systems of simultaneous equations, and (b) the measurement part which links latent variables to observed variables via a restricted (confirmatory) factor model. The structural part of the model can be written as

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{1}$$

where $\boldsymbol{\eta}$ is a vector of endogenous (criterion) latent variables, $\boldsymbol{\xi}$ is a vector of exogenous (predictor) latent variables, $\mathbf{B}$ is a matrix of regression coefficients relating the latent endogenous variables to each other, Γ is a matrix of regression coefficients relating endogenous variables to exogenous variables, and $\boldsymbol{\zeta}$ is a vector of disturbance terms.

The latent variables are linked to observable variables via measurement equations for the endogenous variables and exogenous variables. These equations are defined as

$$\mathbf{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2}$$

and

$$\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \tag{3}$$

where $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$ are matrices of factor loadings, respectively, and $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ are vectors of uniqueness, respectively. In addition, the general model specifies variances and covariances for $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{\delta}$, denoted Φ, Ψ, Θ_ε, and Θ_δ, respectively.

Structural equation modeling is a methodology designed primarily to test substantive theories. As such, a theory might be sufficiently developed to suggest that certain constructs do not affect other constructs, that certain variables do not load on certain factors, and that certain disturbances and measurement errors do not covary. This implies that some of the elements of B, Γ, Λ_y, Λ_x, Φ, Ψ, Θ_ε, and Θ_δ

are *fixed* to zero by hypothesis. The remaining parameters are *free* to be estimated. The pattern of fixed and free parameters implies a specific structure for the covariance matrix of the observed variables. In this way, structural equation modeling can be seen as a special case of a more general *covariance structure model* defined as $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\Omega})$, where $\boldsymbol{\Sigma}$ is the population covariance matrix and $\boldsymbol{\Sigma}(\boldsymbol{\Omega})$ is a matrix valued function of the *parameter vector* $\boldsymbol{\Omega}$ containing all of the parameters of the model.

3. Identification and Estimation of Structural Equation Models

A prerequisite for the estimation of the parameters of the model is to establish whether the parameters are *identified*. We begin with a definition of *identification* by considering the problem from the covariance structure modeling perspective. The population covariance matrix $\boldsymbol{\Sigma}$ contains population variances and covariances that are assumed to follow a model characterized by a set of parameters contained in Ω. We know that the variances and covariances in $\boldsymbol{\Sigma}$ can be estimated by their sample counterparts in the sample covariance matrix $\mathbf{S}$ using straightforward equations for the calculation of sample variances and covariances. Thus, the parameters in Σ are identified. What needs to be established prior to estimation is whether the unknown parameters in Ω are identified from the elements of Σ.

We say that the elements in Ω are *identified* if they can be expressed uniquely in terms of the elements of the covariance matrix $\boldsymbol{\Sigma}$. If all elements in Ω are identified then the model is identified. A variety of necessary and sufficient conditions for establishing the identification of the model exist and are described in Kaplan (2000).

Major advances in structural equation modeling have, arguably, been made in the area of parameter estimation. The problem of parameter estimation concerns obtaining estimates of the free parameters of the model, subject to the constraints imposed by the fixed parameters of the model. In other words, the goal is to apply a model to the sample covariance matrix $\mathbf{S}$, which is an estimate of the population covariance matrix $\boldsymbol{\Sigma}$ that is assumed to follow a structural equation model. Estimates are obtained such that the discrepancy between the estimated covariance matrix $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\Omega}})$ and the sample covariance matrix $\mathbf{S}$, is as small as possible. Therefore, parameter estimation requires choosing among a variety of *fitting functions*. Fitting functions have the properties that (a) the value of the function is greater than or equal to zero, (b) the value of the function equals zero if, and only if, the model fits the data perfectly, and (c) the function is continuous. We first consider estimators that assume multivariate normality of the data. We will discuss estimators for non-normal data in Sect. 5.

3.1 Normal Theory-based Estimation

The most common estimation method that rests on the assumption of multivariate normality is the method of *maximum likelihood*. If the assumption of multivariate normality of the observed data holds, the estimates obtained from maximum likelihood estimation possess certain desirable properties. Specifically, maximum likelihood estimation will yield unbiased and efficient estimates as well as a large sample goodness-of-fit test.

Another fitting function that enjoys the same properties under the same assumption is the *generalized least squares* (GLS) estimator developed for the structural equation modeling context by Jöreskog and Goldberger (1972).

4. Testing

Once the parameters of the model have been estimated, the model can be tested for global fit to the data. In addition, hypotheses concerning individual parameters can also be evaluated. We consider model testing in this section.

4.1 Testing Global Model Fit

A feature of maximum likelihood estimation is that one can explicitly test the hypothesis that the model fits data. The null hypothesis states that the specified model holds in the population. The corresponding alternative hypothesis is that the population covariance matrix is any symmetric (and positive definite) matrix. Under maximum likelihood estimation, the statistic for testing the null hypothesis that the model fits in the population is referred to as the *likelihood ratio* (*LR*) *test*. The large sample distribution of the likelihood ratio test is chi-squared with degrees of freedom given by the difference in the number of nonredundant elements in the population covariance matrix and the number of free parameters in the model.

4.2 Testing Hypotheses Regarding Individual Parameters

In addition to a global test of whether the model holds exactly in the population, one can also test hypotheses regarding the individual fixed and freed parameters in the model. Specifically, we can consider three alternative ways to evaluate the fixed and freed elements of the model vis-à-vis overall fit.

The first method rests on the difference between the likelihood ratio chi-squared statistics comparing a given model against a less restrictive model. Recall that the initial hypothesis, say H_{01}, is tested against the alternative hypothesis H_a that Σ is a symmetric positive

definite matrix. Consider a second hypothesis, say H_{02} that differs from H_{01} in that a single parameter that was restricted to zero is now relaxed. Note that the alternative hypothesis is the same in both cases. Therefore, the change in the chi-squared value between the two models can be used to test the improvement in fit due to the relaxation of the restriction. The distribution of the difference between the two chi-squared tests is distributed as chi-squared with degrees of freedom equaling the difference in degrees-of-freedom between the model under H_{01} and the less restrictive model under H_{02}. In the case of a single restriction described here, the chi-squared difference test is evaluated with one degree of freedom.

There are other ways of assessing the change in the model without estimating a second nested model. The *Lagrange multiplier* (*LM*) test can be used to assess whether freeing a fixed parameter would result in a significant improvement in the overall fit of the model. The LM test is asymptotically distributed as chi-squared with degrees of freedom equaling the difference between the degrees of freedom of the more restrictive model and the less restrictive model. Again, if one restriction is being evaluated, then the LM test will be evaluated with one degree of freedom. The LM test is also referred to as the *modification index*.

Finally, we can consider evaluating whether fixing a free parameter would result in a significant decrement in the fit of the model. This is referred to as the *Wald test*.

The Wald test is asymptotically distributed as chi-squared with degrees of freedom equaling the number imposed restrictions. When interest is in evaluating one restriction the Wald test is distributed as chi-squared with one degree of freedom. The square root of the Wald statistic in this case is a z statistic. The reason why the LM and Wald tests can be used as alternative approaches for testing model parameters is that they are asymptotically equivalent to each other and to the chi-squared difference test.

4.3 Alternative Methods of Model Fit

Since the early 1980s, attention has focused on the development of alternative indices that provide relatively different perspectives on the fit of structural equation models. The development of these indices has been motivated, in part, by the known sensitivity of the likelihood ratio chi-squared statistic to large sample sizes. Other classes of indices have been motivated by a need to rethink the notion of testing exact fit in the population—an idea that is deemed by some to be unrealistic in most practical situations. Finally, another class of alternative indices has been developed that focuses on the cross-validation adequacy of the model.

In this section we will divide our discussion of alternative fit indices into three categories: (a) measures based on comparative fit to a baseline model, including those that add a penalty function for model complexity, (b) measures based on population errors of approximation, and (c) cross-validation measures.

4.3.1 Measures based on comparative fit to a baseline model. Arguably, the most active work in the area of alternative fit indices has been the development of what can be broadly referred to as measures of *comparative fit*. The basic idea behind these indices is that the fit of the model is compared to the fit of some baseline model that usually specifies complete independence among the observed variables. The fit of the baseline model will usually be fairly large and thus the issue is whether one's model of interest is an improvement relative to the baseline model. The index is typically scaled to lie between zero and one, with one representing perfect fit relative to this baseline model. The sheer number of comparative fit indices precludes a detailed discussion of each one. We will consider a subset of indices here that serve to illustrate the basic ideas.

The quintessential example of a comparative fit index is the normed-fit index (NFI) proposed by Bentler and Bonett (1980). This index can be written as

$$\mathrm{NFI} = \frac{\chi_b^2 - \chi_t^2}{\chi_b^2} \tag{4}$$

where χ_b^t is the chi-squared for the model of complete independence (the so-called baseline model) and χ_t^2 is the chi-squared for the target model of interest. This index, which ranges from zero to one, provides a measure of the extent to which the target model is an improvement over the baseline model.

The NFI assumes a true null hypothesis and therefore a central chi-squared distribution of the test statistic. However, an argument could be made that the null hypothesis is never exactly true and that the distribution of the test statistic can be better approximated by a non-central chi-squared with noncentrality parameter λ. An estimate of the noncentrality parameter can be obtained as the difference between the statistic and its associated degrees of freedom. Thus, for models that are not extremely misspecified, an index developed by McDonald and Marsh (1990) and referred to as the *relative noncentrality index* (RNI) can be defined as

$$\mathrm{RNI} = \frac{[(\chi_b^2 - df_b) - (\chi_t^2 - df_t)]}{\chi_b^2 - df_b} \tag{5}$$

The RNI can lie outside the 0–1 range. To remedy this, Bentler (1990) adjusted the RNI so that it would lie in the range of 0–1. This adjusted version is referred to as the *comparative fit index* (CFI).

Finally, there are classes of comparative fit indices that adjust existing fit indexes for the number of parameters that are estimated. These are so called *parsimony-based* comparative fit indices. The rationale behind these indices is that a model can be made to fit the data by simply estimating more and more parameters. Indeed, a model that is just identified fits the data perfectly. Therefore, an appeal to parsimony would require that these indices be adjusted for the number of parameters that are estimated.

It is important to note that use of comparative indices has not been without controversy. In particular, Sobel and Borhnstedt (1985) argued early on that these indices are designed to compare one's hypothesized model against a scientifically questionable baseline hypothesis that the observed variables are completely uncorrelated with each other. Yet, as Sobel and Borhnstedt (1985) argue, one would never seriously entertain such a hypothesis, and that perhaps these indices should be compared with a different baseline hypothesis. Unfortunately, the conventional practice of structural equation modeling as represented in scholarly journals suggests that these indices have always been compared with the baseline model of complete independence.

4.3.2 Measures based on errors of approximation. It was noted earlier that the likelihood ratio chi-squared test assesses an exact null hypothesis that the model fits perfectly in the population. In reality, it is fairly unlikely that the model will fit the data perfectly. Not only is it unreasonable to expect our models to hold perfectly, but trivial misspecifications can have detrimental effects on the likelihood ratio test when applied to large samples. Therefore, a more sensible approach is to assess whether the model fits approximately well in the population. The difficulty arises when trying to quantify what is meant by 'approximately.'

To motivate this work it is useful to differentiate among different types of discrepancies. First, there is the *discrepancy due to approximation*, which measures the lack of fit of the model to the population covariance matrix. Second, there is the *discrepancy due to estimation*, which measures the difference between the model fit to the sample covariance matrix and the model fit to the population covariance matrix. Finally, there is the discrepancy due to overall error that measures the difference between the elements of the population covariance matrix and the model fit to the sample covariance matrix.

Measures of approximate fit are concerned with the discrepancy due to approximation. Based on the work of Steiger and Lind (1980) it is possible to assess approximate fit of a model in the population. The method of Steiger and Lind (1980) utilizes the root mean square error of approximation (RMSEA) for measuring approximate fit. In utilizing the RMSEA for assessing approximate fit, a formal hypothesis testing framework is employed. Typically an RMSEA value $\leqslant 0.05$ is considered a 'close fit.' This value serves to define a formal null hypothesis of close fit. In addition, a 90% confidence interval around the RMSEA can be formed, enabling an assessment of the precision of the estimate (Steiger and Lind 1980). Practical guidelines recommended suggest that values of the RMSEA between 0.05 and 0.08 are indicative of fair fit, while values between 0.08 and 0.10 are indicative of mediocre fit.

4.3.3 Measures that assess cross-validation adequacy. Another important consideration in evaluating a structural model is whether the model is capable of cross-validating well in a future sample of the same size, from the same population, and sampled in the same fashion. In some cases, an investigator may have a sufficiently large sample to allow it to be randomly split in half with the model estimated and modified on the *calibration* sample and then cross-validated on the *validation* sample. When this is possible, then the final fitted model from the calibration sample is applied to the validation sample covariance matrix with parameter estimates fixed to the estimated values obtained from the calibration sample.

In many instances, investigators may not have samples large enough to allow separation into calibration and validation samples. However, the cross-validation adequacy of the model remains a desirable piece of information in model evaluation. Two comparable methods for assessing cross-validation adequacy are the Akaike information criterion (AIC) and the expected cross-validation index (ECVI). In both cases, it is assumed that the investigator is in a position to choose among a set of competing models. The model with the lowest AIC or ECVI value will be the model that will have the best cross-validation capability.

5. *Statistical Assumptions Underlying Structural Equation Modeling*

As with all statistical methodologies, structural equation modeling requires that certain underlying assumptions be satisfied in order to ensure accurate inferences. The major assumptions associated with structural equation modeling include: multivariate normality, no systematic missing data, sufficiently large sample size, and correct model specification.

5.1 *Sampling Assumptions*

For the purposes of proper estimation and inference, a very important question concerns the sampling mechanism. In the absence of explicit information to the

contrary, estimation methods such as maximum likelihood assume that data are generated according to simple random sampling. Perhaps more often than not, however, structural equation models are applied to data that have been obtained through something other than simple random sampling. For example, when data are obtained through multistage sampling, the usual assumption of independent observations is violated resulting in predictable biases. Structural equation modeling expanded to consider multistage sampling will be discussed in Sect. 6.

5.2 Non-normality

A basic assumption underlying the standard use of structural equation modeling is that observations are drawn from a continuous and multivariate normal population. This assumption is particularly important for maximum likelihood estimation because the maximum likelihood estimator is derived directly from the expression for the multivariate normal distribution. If the data follow a continuous and multivariate normal distribution, then maximum likelihood attains optimal asymptotic properties, viz., that the estimates are normal, unbiased, and efficient.

The effects of non-normality on estimates, standard errors, and tests of model fit are well known. The extant literature suggests that non-normality does not affect parameter estimates. In contrast, standard errors appear to be underestimated relative to the empirical standard deviation of the estimates. With regard to goodness-of-fit, research indicates that non-normality can lead to substantial overestimation of likelihood ratio chi-squared statistic, and this overestimation appears to be related to the number of degrees of freedom of the model.

5.2.1 Estimators for non-normal data. The estimation methods discussed in Sect. 3 assumed multivariate normal data. In the case where the data are not multivariate normal, two general estimators exist that differ depending on whether the observed variables are continuous or categorical.

One of the major breakthroughs in structural equation modeling has been the development of estimation methods that are not based on the assumption of continuous multivariate normal data. Browne (1984) developed an asymptotic distribution free (ADF) estimator based on weighted least-squares theory, in which the weight matrix takes on a special form. Specifically, when the data are not multivariate normal, the weight matrix can be expanded to incorporate information about the skewness and kurtosis of the data. Simulation studies have shown reasonably good performance of the ADF estimation compared with ML and GLS.

A problem with the ADF estimator is the computational difficulties encountered for models of moderate size. Specifically, with p variables there are $u = 1/2p(p+1)$ elements in the sample covariance matrix $\mathbf{S}$. The weight matrix is of order $u \times u$. It can be seen that the size of the weight matrix grows rapidly with the number of variables. So, if a model was estimated with 20 variables, the weight matrix would contain 22,155 distinct elements. Moreover, ADF estimation required that the sample size (for each group if relevant) exceed $p + 1/2p(p+1)$ to ensure that the weight matrix is non-singular. These constraints have limited the utility of the ADF estimator in applied settings.

Another difficulty with the ADF estimator is that social and behavioral science data are rarely continuous. Rather, we tend to encounter categorical data and often, we find mixtures of scale types within a specific analysis. When this is the case, the assumptions of continuity as well as multivariate normality are violated. Muthén (1984) made a major contribution to the estimation of structural equation models under these more realistic conditions by providing a unified approach for estimating models containing mixtures of measurement scales.

The basic idea of Muthén's (1984) method is that underlying each of the categorical variables is a latent continuous and normally distributed variable. The statistics for estimating the model are correlations among these latent variables. For example, the Pearson product–moment correlations between observed dichotomous variables are called *phi coefficients*. Phi coefficients can underestimate the 'true' or latent correlations between the variables. Thus, when phi coefficients are used in structural equation models, the estimates may be attenuated. In contrast, the latent correlations between dichotomous variables are referred to as *tetrachoric correlations*, and their use would result in disattenuated correlation coefficients. Other underlying correlations can also be defined.

The procedure developed by Muthén (1984) handles mixtures of scale types by calculating the appropriate correlations and the appropriate weight matrix corresponding to those correlations. These are entered into a weighted least-squares analysis yielding correct goodness-of-fit tests, estimates, and standard errors. Unfortunately, as with the ADF estimator, the method of Muthén is computationally intensive. However, research has shown how such estimation methods can be developed that are not as computationally intensive.

5.3 Missing Data

Generally, statistical procedures such as structural equation modeling assume that each unit of analysis has complete data. However, for many reasons, units may be missing values on one or more of the variables

under investigation. Numerous *ad hoc* approaches exist for handling missing data, including listwise and pairwise deletion. However, these approaches assume that the missing data are unrelated to the variables that have missing data, or any other variable in the model, i.e., missing completely at random (MCAR)—a heroic assumption at best. More sophisticated approaches for dealing with missing data derive from the seminal ideas of Little and Rubin (1987).

Utilizing the Rubin and Little framework, a major breakthrough in model-based approaches to missing data in the structural equation modeling context was made simultaneously by Allison (1987) and Muthén et al. (1987). They showed how structural equation modeling could be used when missing data are missing-at-random (MAR), an assumption that is more relaxed than missing-completely-at-random.

The approach advocated by Muthén et al. (1987) was compared with standard listwise deletion and pairwise deletion approaches (which assume MCAR) in an extensive simulation study. The general findings were that Muthén et al.'s approach was superior to listwise and pairwise deletion approaches for handling missing data even under conditions where it was not appropriate to ignore the mechanism that generated the missing data.

The major problem associated with Muthén et al.'s approach is that it is restricted to modeling under a relatively small number of distinct missing data patterns. Small numbers for distinct missing data patterns will be rare in social and behavioral science applications. However, it has proved possible to apply MAR-based approaches (Little and Rubin 1987) to modeling missing data within standard structural equation modeling software in a way that mitigates the problems with Muthén et al.'s approach. Specifically, imputation approaches have become available for maximum likelihood estimation of structural model parameters under incomplete data assuming MAR. Simulation studies have demonstrated the effectiveness of these methods in comparison to listwise deletion and pairwise deletion under MCAR and MAR.

5.4 Specification Error

Structural equation modeling assumes that the model is properly specified. Specification error is defined as the omission of relevant variables in any equation of the system of equations defined by the structural equation model. This includes the measurement model equations as well as the structural model equations. As with the problems of non-normality and missing data, the question of concern is the extent to which omitted variables affect inferences.

The mid-1980s saw a proliferation of studies on the problem of specification error in structural equation models. On the basis of work by Kaplan and others (summarized in Kaplan 2000) the general finding is that specification errors in the form of omitted variables can result in substantial parameter estimate bias. In the context of sampling variability, specification errors have been found to be relatively robust to small specification errors. However, tests of free parameters in the model are affected in such a way that specification error in one parameter can propagate to affect the power of the test in another parameter in the model. Sample size also interacts with the size and type of the specification error.

6. Modern Developments Constituting the 'Second Generation'

Structural equation modeling is, arguably, one of the most popular statistical methodologies available to quantitative social scientists. The popularity of structural equation modeling has led to the creation of a scholarly journal devoted specifically to structural equation modeling as well as the existence of SEMNET, a very popular and active electronic discussion list that focuses on structural equation modeling and related issues. In addition, there are a large number of software programs that allow for sophisticated and highly flexible modeling.

In addition to 'first-generation' applications of structural equation modeling, later developments have allowed traditionally different approaches to statistical modeling to be specified as special cases of structural equation modeling. Space constraints make it impossible to highlight all of the modern developments in structural equation modeling. However, the more important modern developments have resulted from specifying the general model in a way that allows a 'structural modeling' approach to other types of analytic strategies. The most recent example of this development is the use of structural equation modeling to estimate multilevel data as well as specifying structural equation models to estimate growth parameters from longitudinal data.

6.1 Multilevel Structural Equation Modeling

Attempts have been made to integrate multilevel modeling with structural equation modeling so as to provide a general methodology that can account for issues of measurement error and simultaneity as well as multistage sampling.

Multilevel structural equation modeling assumes that the levels of the within-group endogenous and exogenous variables vary over between-group units. Moreover, it is possible to specify a model which is assumed to hold at the between-group level and that explains between-group variation of the within-group variables. An overview and application of multilevel structural equation modeling can be found in Kaplan (2000).

6.2 Growth Curve Modeling from the Structural Equation Modeling Perspective

Growth curve modeling has been advocated for many years by numerous researchers for the study of interindividual differences in change. The specification of growth models can be viewed as falling within the class of multilevel linear models (Bryk and Raudenbush 1992), where level 1 represents intraindividual differences in initial status and growth and level 2 models individual initial status and growth parameters as a function of interindividual differences.

Research has also shown how growth curve models could also be incorporated into the structural equation modeling framework utilizing existing structural equation modeling software. The advantages of the structural equation modeling framework to growth curve modeling is its tremendous flexibility in specifying models of substantive interest. Developments in this area have allowed the merging of growth curve models and latent class models for the purpose of isolating unique populations defined by unique patterns of growth.

7. Pros and Cons of Structural Equation Modeling

Structural equation modeling is, without question, one of the most popular methodologies in the quantitative social sciences. Its popularity can be attributed to the sophistication of the underlying statistical theory, the potential for addressing important substantive questions, and the availability and simplicity of software dedicated to structural equation modeling. However, despite the popularity of the method, it can be argued that 'first-generation' use of the method is embedded in a conventional *practice* that precludes further statistical as well as substantive advances.

The primary problem with the 'first-generation' practice of structural equation modeling lies in attempting to attain a 'well-fitting' model. That is, in conventional practice, if a model does not fit from the standpoint of one statistical criterion (e.g., the likelihood ratio chi-squared test), then other conceptually contradictory measures are usually reported (e.g., the NNFI). In addition, it is not uncommon to find numerous modifications made to an ill-fitting model to bring it in line with the data, usually supplemented by *post hoc* justification for how the modification fit into the original theoretical framework.

Perhaps the problem lies in an obsession with null hypothesis testing—certainly an issue that has received considerable attention. Or perhaps the practice of 'first-generation' structural equation modeling is embedded in the view that only a well-fitting model is worthy of being interpreted. Regardless, it has been argued by Kaplan (2000) that this conventional practice precludes learning valuable information about the phenomena under study that could otherwise be attained if the focus was on the predictive ability of a model. Such a focus on the predictive ability of the model combined with a change of view toward strict hypothesis testing might lead to further substantive and statistical developments. Opportunities for substantive development and model improvement emerge when the model does not yield accurate or admissible predictions derived from theory. Opportunities for statistical developments emerge when new methods are developed for engaging in prediction studies and evaluating predictive performance.

That is not to say that there are no bright spots in the field of structural equation modeling. Indeed, developments in multilevel structural equation modeling, growth curve modeling, and latent class applications suggest a promising future with respect to statistical and substantive developments. However, these 'second-generation' methodologies will have to be combined with a 'second-generation' epistemology so as to realize the true potential of structural equation modeling in the array of quantitative social sciences.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Factor Analysis and Latent Structure, Confirmatory; Factor Analysis and Latent Structure: Overview; Growth Curve Analysis; Indicator: Methodology; Latent Structure and Causal Variables; Linear Hypothesis: Regression (Basics); Statistical Analysis: Multilevel Methods

Bibliography

Allison P D 1987 Estimation of linear models with incomplete data. In: Clogg C C (ed.) *Sociological Methodology*. Jossey-Bass, San Francisco, pp. 71–103

Bentler P M 1980 Multivariate analysis with latent variables: causal modeling. *Annual Review of Psychology* **31**: 419–56

Bentler P M 1990 Comparative fit indices in structural models. *Psychological Bulletin* **107**: 238–46

Bentler P M, Bonett D G 1980 Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin* **88**: 588–606

Browne M W 1984 Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* **37**: 62–83

Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park, CA

Jöreskog K G 1972 Factor analysis by generalized least squares. *Psychometrika* **37**: 243–60

Jöreskog K G 1973 A general method for estimating a linear structural equation system. In: Goldberger A S, Duncan O D (eds.) *Structural Equation Models in the Social Sciences*. Academic Press, New York, pp. 85–112

Kaplan D 2000 *Structural Equation Modeling: Foundations and Extensions*. Sage, Newbury Park, CA

Keesling J W 1972 Maximum likelihood approaches to causal analysis. Unpublished Ph.D. thesis, University of Chicago

Little R J A, Rubin D B 1987 *Statistical Analysis with Missing Data*. Wiley, New York

McDonald R P, Marsh H W 1990 Choosing a multivariate model: noncentrality and goodness-of-fit. *Psychological Bulletin* **107**: 247–55

Muthén B 1984 A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**: 115–32

Muthén B 2001 Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In: Sayer A, Collins L (eds.) *New Methods for the Analysis of Change*. American Psychological Association, Washington, DC, pp. 291–322

Muthén B, Kaplan D, Hollis M 1987 On structural equation modeling with data that are not missing completely at random. *Psychometrika* **51**: 431–62

Sobel M, Bohrnstedt G W 1985 Use of null models in evaluating the fit of covariance structure models. In: Tuma N B (ed.) *Sociological Methodology*. Jossey-Bass, San Francisco, pp. 152–78

Steiger J H, Lind J M 1980 Statistically based tests for the number of common factors. Paper presented at the Psychometric Society, Iowa City, IA

Wiley D E 1973 The identification problem in structural equation models with unmeasured variables. In: Goldberger A S, Duncan O D (eds.) *Structural Equation Models in the Social Sciences*. Academic Press, New York, pp. 69–83

D. Kaplan

Structuralism

For expository reasons, a specific sense of structuralism will be distinguished from a loose sense. The specific sense is the program which Claude Lévi-Strauss, borrowing from structural linguistics, tried to introduce into anthropology. It is defined by the various, not always compatible, statements by Lévi-Strauss. The loose sense is reserved for those authors who deserve to be discussed when speaking about a style of approach which has manifested itself in different ways in different times and places, but which in the history of ideas can generally be traced back to certain writings of Emile Durkheim and Marcel Mauss. Included in this groups are figures who wrote well before Lévi-Strauss, such as Dutch authors who made many of the analytic discoveries found in Lévi-Strauss's *The Elementary Structures of Kinship* (1969a), but who did not speak of structuralism, as well as contemporaries of Lévi-Strauss who though knowledgeable of Lévi-Strauss's writings did not regard themselves necessarily as structuralists in his sense, or did so only part of the time.

The specific sense of structuralism poses its own problems. Which of Lévi-Strauss's many doctrines and which of the many intellectual influences on Lévi-Strauss are to be regarded as structuralist? Is it not arbitrary to regard some as structuralist and to disregard the others? Yet taken as a whole, Lévi-Strauss's style of thought is so personal as to be impossible to imitate in its entirety. Given the elusive nature of his language and his many self-contradictions, many would not want to. Lévi-Strauss's thought and prose, for example, have been said to be influenced by the French symbolist poets and the surrealists (e.g., Leach 1970, p. 16); and according to his own testimony he has been influenced by such composers as Debussy, Stravinsky, and Wagner (Lévi-Strauss 1969b, p. 15). Elsewhere (Lévi-Strauss 1955, pp. 48–50) he describes the influence on his thought of geology, psychoanalysis, and Marxism. 'All three showed that understanding consists in the reduction of one type of reality to another; that true reality is never the most obvious of realities … .' In response Leach (1970, p. 13) has written, 'the relevance of Marxist ideology for an understanding of Lévi-Strauss is difficult to determine. Lévi-Strauss' use of dialectic, with the formal sequence of thesis–antithesis–synthesis, is Hegelian rather than Marxist and his attitude to history seems to be quite contrary to Marxist dogma.' Lévi-Strauss dedicated *The Savage Mind* (1966) to the phenomenologist Merleau-Ponty. How are we precisely to relate the influence on him of American anthropologists such as Boas, Kroeber, and Lowie to his structuralism?

Perhaps we are on surer ground when it comes to the debts which he admits to owing Marcel Granet, Georges Dumézil, and Henri Grégoire (Lévi-Strauss 1969b, p. 15, n. 5). Also there is less difficulty in seeing the similarity between his structuralism and Gestalt psychology with its emphasis or the perception of patterns (Lévi-Strauss 1963a, pp. 324–41). The relevance of Gestalt was recognized by another of his influences, namely Wiener (1948, p. 18, ff.), one of the founders of communications theory. Indeed games theory, the mathematical theory of communication and cybernetics are obvious and acknowledged sources of some of his more unquestionably structuralist ideas (Lévi-Strauss 1963a, p. 283; for a discussion see Ardener 1971, liv ff.). The most direct source of course is the linguistic methods of the structuralist linguists Troubetzkoy, Jakobson, and Halle (cf. Benveniste 1962). In an anticipatory response to a statement of Wiener (1948, p. 24) that sociology and anthropology lacked the long runs of statistics under constant conditions necessary for the applications of the mathematics of communications theory, Lévi-Strauss wrote in a paper of 1945 on *Structural Analysis in Linguistics and in Anthropology* that, 'It [linguistics] is probably the only one [social science] which can truly claim to be a science.' Lévi-Strauss's scientism (a characteristic of his thought), like that of A. R. Radcliffe-Brown, is thus programmatic rather than realized. However, it is the methods which have allegedly turned linguistics into a scientific social science that Lévi-Strauss wants to use to achieve the same transformation in social anthropology.

Referring to Troubetzkoy (1933), Lévi-Strauss reduced the structural method to four basic operations. First, structural linguistics shifts from the study of conscious linguistic phenomena to study of their unconscious infrastructure; second, it does not treat terms as independent entities, taking instead as its basis of analysis the relations between terms; third, it introduces the concept of system. 'Modern phonemics does not merely proclaim that phonemes are always part of a system; it shows concrete phonemic systems and elucidates their structure'; finally, structural linguistics aims at discovering general laws, either by induction 'or … by logical deduction, which would give them an absolute character.' Thus, for the first time, a social science is able to formulate necessary relationships (Lévi-Strauss 1963a, p. 33, 1962).

The crucial influence here was the conception of the phoneme. This idea was based on a distinction between the significant and the insignificant sounds in a given language. Only a small set of sounds in any given language are capable of determining differences of meanings. These individual phonemes are defined by contrastive features which are the distinctive features of a given languages, and these distinctive features are defined by a system of oppositions—the sort of thing that makes the difference between pat and bat.

In the early 1945 paper he drew an analogy between the phoneme and the terms of a kinship terminology: 'Like phonemes, kinship terms are elements of meaning; like phonemes, they acquire meaning only if they are integrated into systems. Kinship systems, like "phonemic systems," are built by the mind on the level of unconscious thought … in the case of kinship as well as linguistics, the observable phenomena result from the action of laws which are general but implicit' (Lévi-Strauss 1963a, p. 34).

Lévi-Strauss also took over from Ferdinand de Saussure, from whom structuralist linguistics derives, the notion of the arbitrary nature of the linguistic sign and the realization that each linguistic term derives its value from its opposition to all the other terms (de Saussure 1922, p. 126). As Ardener (1971, p. 36) comments, 'Here we have clearly stated the now fashionable anthropological view that elements in the system define themselves in opposition to all other elements in the system.'

In contrasting his views on social structure to those of Radcliffe-Brown, Lévi-Strauss held that social structure is a model made up after empirical reality. It is, therefore, a method with the following features. It is a system made of elements, none of which can change without changing other elements. As a model, it should take its place in a series of transformations resulting in a group of models of the same type. It should be possible to predict how the model will react if its elements are modified. Importantly and finally, the model should make intelligible all of the facts. The best model is the simplest model which accounts for all of the facts (Lévi-Strauss 1963a, pp. 279–81). Lévi-Strauss can be said to have brought the intellect into anthropology. As he said, 'each culture has its own theoreticians whose contributions deserve the same attention as that which the anthropologist gives to colleagues.' 'Many "primitive" cultures have built models of their marriage regulations which are much more to the point than models built by professional anthropologists' (Lévi-Strauss 1963a, p. 282). The anthropologist should take these models into consideration and after comparing them to what people actually do, form his own models which for Lévi-Strauss are the structure.

These ideas are already present in his early classic *The Elementary Structures of Kinship* (1969a) in which he treated certain positive marriage rules as transformations across a range of logically possible models, of which only some had empirical realization. They directed his study of systems of classification in his *Totemism* (1963b) and *The Savage Mind* (1966). A characteristic argument is that 'so-called totemic institutions' turn on 'an homology between two systems of differences,' that is, the differences between species is homologous to the differences between groups. These ideas are also carried into his several papers and four books on myth analysis which were intended to reveal the constraints on the human mind (Lévi Strauss 1969b, p. 10). 'Mythical thought always progresses from the awareness of oppositions toward their resolution' (Lévi-Strauss 1963a, p. 224). Lévi-Strauss (1963a, p. 228) early on proposed that every myth corresponds to a formula of the following type

$$Fx(a):Fy(b) \sim Fx(b):Fa-1(y).$$

Later he has he returned to it (Lévi-Strauss 1988, p. 57, 156, 163, 170). This formula he describes as representing 'an inversion of terms and relations, under two conditions: (1) that one term be replaced by its opposite (in the above formula, a and $a-1$); (2) that an inversion be made between the function value and the term value of two elements (above, y and a). Mosko (1991) has modified the formula and said that it works. For some, however, the formula may be incomprehensible. However, elsewhere Lévi-Strauss warns his readers against taking his mathematical formulas and symbols too seriously, as their resemblances to those of the mathematicians are only superficial (Lévi-Strauss 1969b, p. 30). Nevertheless, apart from the above formula, Lévi-Strauss never shows us how the fundamental structure of myth differs from his initial ideas of opposition, mediation and transformation. When he does set forth the structure of a given myth, the demonstration is so heavily freighted with content (the concrete elements that are in opposition) that it is hard to see the difference between the formal structure and the concrete myth.

Dumont (1975, p. 333), following Pocock (1961, p. 76), has written of a convergence between Evans-Pritchard and Lévi-Strauss in the pages of *The Nuer* (1940a), in a passing from the functionalism of

Radcliffe-Brown to meaning. Radcliffe-Brown used structure as a synonym for social structure, the true subject of social anthropology. In the closing pages of *The Nuer*, Evans-Pritchard makes several attempts to characterize his achievement and state what he means by structure, finally stating that, 'We have defined structure by what amounts to the presence of group segmentation,' that is, the segmentary principle (Evans-Pritchard 1940a, p. 265). If this is a convergence, it is a convergence before the fact; for Lévi-Strauss wrote his essay on social structure more than a decade after Evans-Pritchard wrote *The Nuer*. It is striking that in that essay Lévi-Strauss mentions *The Nuer* only in a footnote in the midst of a long list of books which have had something to say about economic formations (Lévi-Strauss 1963a, p. 319, n. 49). At the risk of reading the present into the past and taking the lead of Pocock and Dumont, we could say that in 1939 or thereabouts, Evans-Pritchard did not have the vocabulary available to talk about his achievement which would have been available a little over a decade later. Taking the better parts of Lévi-Strauss's essay, he might later have been able to write that what he had done in *The Nuer* is present a series of models, starting with the models of the Nuer themselves, that is, their meanings, and that the most salient and most abstract model he had used was the segmentary principle, which was essentially his final definition of structure.

Dumont noted too that in Evans-Pritchard's treatment of factual oppositions there were underlying conceptual oppositions and that, 'the author has well and truly discovered structuralism on his own account without direct borrowing—as was the case in other quarters—from contemporary developments in neighboring disciplines (phonology, Gestalt theory)' (Dumont 1975, p. 335). Dumont also noted that 'what is properly and strictly the structural aspect of *The Nuer* has never really taken root in England' (Dumont 1975, p. 334). For example, 'immediately afterwards, the opposite tendency to reify or solidify, in the manner of Radcliffe-Brown, was vigorously asserted by Meyer Fortes' (Dumont 1975, p. 340). Thus, Fortes stated that 'The most important feature of unilineal descent groups in Africa … is their corporate organization' (Fortes 1953, p. 25). More recently Holy stated that 'Evans-Pritchard saw the lineage system in terms of the paradigm of hierarchically structural segments, each of which was a corporation' and lists as one of his errors 'ascribing corporateness to what are empirically … only categories of people' (Holy 1996, pp. 88–9). As evidence of Evans-Pritchard's views he cites Fortes. In fact, both Fortes and Holy overlook the fact that Evans-Pritchard stated quite plainly in *The Nuer* (Evans-Pritchard 1940a, p. 89) and in his contribution to *African Political Systems* (1940b, pp. 286–7) that, 'Nuer lineages are not corporate, localized, communities' and 'clans and their lineages are not distinct corporate groups.'

In *The Nuer* structure remains close to social structure, but by 1956 in chapter IV, 'Spirit and the Social Order,' in *Nuer Religion*, Evans-Pritchard was applying similar ideas to conceptions of divinity. 'Given the segmentary political and lineage structure of the Nuer it is understandable that the same complementary tendencies towards fission and fusion and the same relativity that we find in the structure are found also in the action of Spirit in the social life' (Evans-Pritchard 1956 p. 115). This chapter understandably bears a close similarity in inspiration to Chap. 4 in his student Needham's *Structure and Sentiment* (1962) and to that mine of structuralist inspiration *Primitive Classification* (Durkheim and Mauss 1963) with its concern about the relation between the structure of society and the structure of classification (for a discussion of its influence see Needham 1963, pp. 31–2). Evans-Pritchard was one of the few Africanists who had the slightest understanding of what Asianists such as Dumont, Leach, and Needham were trying to do in bringing structuralist approaches into British anthropology, particularly, but not exclusively, in relation to the topic of marriage alliance.

A convergence between Lévi-Strauss's structuralism and early work by Dutch anthropologists was recognized and celebrated by Josselin de Jong (1952). P. E. Josselin de Jong (1977) provided a service by bringing out a selection of this early Dutch structuralism (to give it a retrospective designation) in English translation in his *Structural Anthropology in the Netherlands*, which includes Louis Onvlee's essay on the construction of the Mangili dam, a crystal clear example of pre-Lévi-Straussian structuralism that achieves more than Lévi-Strauss ever did at a similar level, but which remains entirely free of the characteristic blemishes of Lévi-Strauss's thought.

The terms structuralism and structuralist were unknown to some of the authors alluded to here, accepted retrospectively by some, and avoided by others. Leach wrote (1970, p. 9), 'I myself was once a pupil of Malinowski and I am, at heart, still a 'functionalist' even though I recognize the limitations of Malinowski's own brand of theory. Although I have occasionally used the 'structuralist' methods of Lévi-Strauss to illuminate particular features of particular cultural systems the gap between my general position and that of Lévi-Strauss is very wide.' In *Political Systems of Highland Burma* (1954), which Leach later characterized as rationalist, that is, structuralist, unlike his later book *Pul Eliya* (1961) which he characterized a empiricist, that is, functionalist (Leach 1976, p. 6), Leach writes that as opposed to social structure, culture provides the form and 'dress' of the social situation (Leach 1954, p. 16). These views did not prevent him from publishing structuralist treatments of the Bible (Leach 1962) and of Michelangelo's paintings on the ceilings of the Sistine Chapel (Leach 1977). Leach's attitude appears to be that structural-

ism is a technique to be used when it is advantageous to do so, rather than a movement within which the anthropologists must always travel. In all probability Dumont would have agreed with him. When considering north Indian kinship in relation to south Indian kinship, Dumont (1966, p. 103) found that his structuralist methods, which were so productive in the south, were unproductive, or at least not productive to the same extent, in the north. It can be said that some peoples are structuralist and others are not. Perhaps it could be said that there are structuralist topics and non-structuralist topics about a given people. The sensible course is to use whatever techniques suggest themselves and produce a result.

See also: Anthropology, History of; Durkheim, Emile (1858–1917); Kinship in Anthropology; Linguistics: Overview; Saussure, Ferdinand de (1857–1913); Structuralism, History of; Structuralism, Theories of

Bibliography

Ardener E 1971 Introductory essay: Social anthropology and language. In: Ardener E (ed.) *Social Anthropology and Language* (ASA Monographs 10). Tavistock, London

Benveniste E 1962 'Structure' en linguistique. In: Bastide R (ed.) *Sens et usages du terme structure*. Mouton, The Hague

Dumont L 1966 North India in relation to South India. *Contributions to Indian Sociology* **9**: 90–103

Dumont L 1975 Preface by Louis Dumont to the French edition of *The Nuer*. In: Beattie J H M, Lienhardt R G (eds.) *Studies in Social Anthropology*. Clarendon Press, Oxford, UK

Durkheim E, Mauss M 1963 *Primitive Classification* (trans. Needham R). University of Chicago Press, Chicago

Evans-Pritchard E E 1940a *The Nuer*. Clarendon Press, Oxford, UK

Evans-Pritchard E E 1940b The Nuer of the Southern Sudan. In: Fortes M, Evans-Pritchard E E (eds.) *African Political Systems*. Oxford University Press, London

Evans-Pritchard E E 1956 *Nuer Religion*. Clarendon Press, Oxford, UK

Fortes M 1953 The structure of unilineal descent groups. *American Anthropologist* **55**(1): 17–41

Holy L 1996 *Anthropological Perspectives on Kinship*. Pluto Press, London

Josselin de Jong J P B 1952 *Lévi-Strauss's Theory on Kinship and Marriage*. Brill, Leiden

Josselin de Jong P E (ed.) 1977 *Structural Anthropology in the Netherlands* (Koninklijk Instituut voor Taal-, Landen Volkenkunde Translation Series 17). Nijhoff, The Hague

Leach E 1954 *Political Systems of Highland Burma*. Bell, London

Leach E 1961 *Pul Eliya*. Cambridge University Press, Cambridge, UK

Leach E 1962 Genesis as myth. *Discovery* May: 30–5

Leach E 1970 *Lévi-Strauss*. Collins, London

Leach E 1976 *Culture and Communication*. Cambridge University Press, Cambridge, UK

Leach E 1977 Michelangelo's genesis: Structuralist comments on the paintings on the Sistine Chapel ceiling. *Times Literary Supplement* March **18**: 311–3

Lévi-Strauss C 1955 *Tristes Tropiques*. Plon, Paris

Lévi-Strauss C 1962a Les limites de la notion de structure en ethnologie. In: Bastide R (ed.) *Sens et usages du terme structure*. Mouton, The Hague

Lévi-Strauss C 1963a *Structural Anthropology*. Basic Books, New York

Lévi-Strauss C 1963b *Totemism* (trans. Needham R). Beacon, Boston

Lévi-Strauss C 1966 *The Savage Mind*. Weidenfeld and Nicolson, London

Lévi-Strauss C 1969a *The Elementary Structures of Kinship* (trans. Harle Bell J, von Sturmer J R, Needham R). Eyre & Spottiswoode, London

Lévi-Strauss C 1969b *The Raw and the Cooked: Introduction to a Science of Mythology*. Harper and Row, New York, Vol. 1

Lévi-Strauss C 1988 *The Jealous Potter* (trans. Chorier B). University of Chicago Press, Chicago

Mosko M S 1991 The canonic formula of myth and nonmyth. *American Ethnologist* **18**(1): 126–51

Needham R 1962 *Structure and Sentiment*. University of Chicago Press, Chicago

Needham R 1963 *Introduction: Primitive Classification*. University of Chicago Press, Chicago

Pocock D 1961 *Social Anthropology*. Sheed and Ward, London

Radcliffe-Brown A R 1952 *Structure and Function in Primitive Society*. Macmillan, London

de Saussure F 1922 *Cours de linguistique générale*, 2nd edn. Payot, Geneva

Troubetzkoy N 1933 *La phonologie actuelle. Psychologie du langage*. Paris

Wiener N 1948 *Cybernetics, or Control and Communication in the Animal and the Machine*. Wiley, New York

R. H. Barnes

Structuralism, History of

The triumph of structuralism during the 1950s and 1960s was of such a remarkable nature that it became identified with the ensemble of French intellectual history since 1945. During these years, little attention was paid to anything outside of this trend which appeared as a new perspective on the world and human culture. Emphasising critical thinking and expressing the emancipatory will of nascent social sciences in search of scholarly and institutional legitimation, structuralism generated a real sense of collective enthusiasm for the whole of French intelligentsia over at least two decades.

Then suddenly, at the beginning of the 1980s, everything was overturned. Most of the French heroes of this intellectual adventure disappeared and with them their work, anxious to be buried by a new epoch, accentuating thus the impression of the end of an era. In this way, the working through of mourning, necessary in order to do justice to what had been one of the most fruitful periods of our intellectual history, had been avoided.

Taken in its most general sense, the word structure served as a password for a large part of the human sciences (*sciences humaines*), enabling a unitary pro-

gram. Where does the concept of structuralism, one that has led both to such infatuation and to such opprobrium, originate? Derived from the word structure, it starts out as an architectural term. Structure was first and foremost 'the way in which a building is constructed' (*Dictionnaire de Trévoux*, 1771 edition).

During the seventeenth and eighteenth centuries, the meaning of the term structure is modified and enlarged by the analogy with living beings. Thus, the human body is perceived as a construction in Fontenelle as is language in Vaugelas and Bernot. The term, therefore, is taken to refer to a description of how the components of a concrete object make up the whole. It can have various applications (anatomical, psychological, geological, mathematical structures, etc.). The structural approach is not really applied to the field of the human sciences until the more recent nineteenth century, with Spencer, Morgan and Marx.

It is then taken to refer to a lasting phenomenon that connects, in complex terms, the components of an aggregate according to a more abstract schema. The term structure, rarely found in Marx, with the exception of the preface to *A Contribution to the Critique of Political Economy* (1859), is consecrated by Durkheim at the end of the nineteenth century in *The Rules of Sociological Method* (1895). Structure thus gives rise to that which, in André Lalande's *Vocabulaire*, is qualified as a neologism—structuralism between 1900 and 1926. Structuralism emerges among psychologists in opposition to the functional psychology of the turn of the century. However, structuralism's real starting point, in terms of its modern meaning and standpoint of domination among all the human sciences, originates in the evolution of linguistics.

Although Saussure uses the term structure only three times in *The Course in General Linguistics*, it is the Prague School (Troubetzkoy and Jakobson) that is to make widespread the use of the terms structure and structuralism. The Danish linguist, Hjemslev, who founded the journal Acta Linguistica in 1939, asserts the reference to structure as a charter program. The first article in this journal treats the subject of 'structural linguistics.'

1. *The Structural Years*

Departing from this linguistic nucleus, the term triggers a veritable revolution throughout the human sciences at the heart of the twentieth century. To the extent that Michel Foucault observed that structuralism 'is not a new method, but the awakened and anxious conscience of modern knowledge.' Jacques Derrida defined this approach as an 'adventure in perception' and Roland Barthes saw structuralism as the passage from symbolic to paradigmatic conscience, or as the advent of the awareness of paradox. It therefore suggests a movement of thought and a new relationship to the world, much wider than a simple binary methodology confined to one or another field of specific study. This interpretation, seeing itself as unitary, privileges the sign at the expense of meaning, space at the expense of time, the object at the expense of the subject, relation at the expense of content, and culture at the expense of nature.

Primarily, structuralism functions as a paradigm for a philosophy of suspicion, or an act of exposure on behalf of intellectuals that aimed at demystifying the *doxa* and denaturalise meaning, to destabilise it while searching behind the scenes for an expression of bad faith. It is postulated that the class position of the speaker or his libidinal position enables the denouncement of the illusion under which all types of discourse are condemned to exist. This situation of exposure is inscribed in the ancestry of French epistemological tradition that proposes a position of overhang or rupture between scientific competence and common sense, the latter seen as being tied to illusion. Behind this paradigm of suspicion, a profound pessimism envelops intellectuals as well as a critique of western modernity the antithesis of which they attempt to grasp by reference to the figures of the child, the insane or the savage. Thus, beneath the liberatory discourse of Enlightenment, the discipline of the body, the confinement and imprisonment of the social body within an infernal logic of knowledge and power is exposed. Roland Barthes, then, says: 'I profoundly refuse my civilisation, to the point of nausea' and the 'Finale' of Claude Lévi-Strauss's *The Naked Man* ends with the word 'NOTHING,' written in capitals, evoking a requiem or the twilight of mankind. This structuralist paradigm proposes the prevailing of a suspension of meaning, as a means of opposing both eurocentrism and various forms of westernised teleology, in favour of a differentialist mode of thought.

The second dimension of the structuralist paradigm involves philosophy's take-over of three of the social sciences undergoing emancipation and sharing a common valorisation of the unconscious as the locus of truth. These disciplines are general linguistics when, with Saussure, they separate language (*langue*) as a legitimate object from speech (*parole*) forced back to the non-scientific domain; anthropology, with its concern with the encoding of messages rather than with messages themselves and psychoanalysis's view upon the unconscious as an effect of language. This relates to a quest for both legitimation and standing by nascent social sciences confronted with the established force of the classical humanities and by the tradition and conservatism of the old Sorbonne. Structuralism, at this stage, offers itself as a third discourse, bridging literature and the hard sciences, seeking institutionalisation through socialisation, by-passing the academic center of the Sorbonne by way of the peripheral universities. Publishing and the press through the respected institution of the *Collège De France*, serves

as a refuge for research on the cutting edge. This involves a veritable battle between the ancients and the moderns in which the themes of breaking and rupture come into play at several levels. The social sciences then attempt to free themselves by cutting the umbilical cord that tied them to philosophy by erecting the efficiency of a scientific method. On their part, some philosophers, taking into account the importance of these endeavors, capture their program and redefine the function of philosophy as a philosophy of the concept. This was to become what is known as 'effectology,' or the use of research in the human sciences for a philosophical project that simultaneously deconstructs the classifications employed from within these practices. Philosophy, therefore, retains its central position while declaring its own demise. The major role played by philosophers in this matter also relies on their ability, in the middle of the 1960s, to impose a unitary program, an endeavor in which the structuralist etiquette serves as a rallying call.

The second explanatory idea that accounts for this history is that of generation. Thus, a number of landmark events make a lasting impression upon a generation indelibly marked by the Second World War. The difficulty of thinking after Auschwitz, expressed by Adorno, meant the difficulty of thinking optimistically, as was possible at the turn of the century, of the history of a west that had turned towards atrocity, to crimes against humanity. This history was to become the very locus of doubt, questioning, and severe criticism. Hell is no longer the other, but oneself. A questioning of the self and of its illusory capacities for mastery then aroused by a society that had reduced the individual to nothing more than a number. This despair that engenders the critical distance of suspicion becomes accentuated during the 1950s and 1960s with the progressive liberation of the colonised peoples who had broken free of the colonial yoke and achieved their independence. Many intellectuals were to see this rejection of western influence as a confirmation of their critical position and held the figure of the 'other,' or the epitome of absolute alterity from the west, to be the very locus of the expression of truth and of a certain purity. This release from western history, that enabled the Bororos or the Nambikwaras to be perceived as expressing the purified cradle of humanity, is reinforced by the successive revelations about what some had taken to be the incarnation of their hopes for the real Marxism of Communist countries.1956 is to become a vital date for a generation of Marxist intellectuals that takes refuge in a structuralised Marxism, a 'vacuum-packed Marxism,' as it is dubbed by Jean-Marie Domenach, upon which the disasters of real Communism would have no bearing. Many among these intellectuals are to break with Communist culture, stealing away on tiptoes from history and taking refuge in the enclosure of text, science, and in the evacuation of the subject and the signified.

2. *Founding the Scientificity of a Third Discourse*

The hope for the scientific renovation of the social sciences found its method, a common language capable of effecting change, within structural linguistics. Linguistics then became the model for a whole range of social sciences aspiring to formalism. The discipline was progressively diffused from anthropology, to literary criticism, and to psychoanalysis and it led to the profound renewal of the mode of philosophical questioning.

The areas most affected by the linguistic contagion were those disciplines that found themselves in an institutionally precarious position. Alternatively, disciplines were in search of an identity tainted by the internal contradiction of their pretensions to scientific positivity and their relationship to the political, as in the case of sociology and those such as literary studies and philosophy, fully immersed in the dispute between tradition and modernity. This connection served to blur the boundaries between disciplines. Structuralism offered itself here as a unifying project. This temptation was most clearly expressed by Roland Barthes who advocated a general semiology capable of grouping all the human sciences around the study of the sign.

Modernization becomes articulated at this point with the interdisciplinary, as it is necessary to defy the sacrosanct boundaries in order to allow the linguistic model to infiltrate throughout the field of the human sciences. As everything is linguistic, everyone is linguistic and the world is language. This interdisciplinariness, that infringes upon the Humboltdian model of the university in which each discipline has its strictly delineated place and is related to the delegitimation of meta-narratives, creates a veritable enthusiasm for all variants of formalism, for knowledge immanent to itself. The key word of the times is 'communication' that, even beyond the journal bearing the same title is evocative of the multi-disciplinary euphoria of this period. Lévi-Strauss was the first to formulate this unifying program for the human sciences in the post-war period. To be sure, the constellation he elaborated gravitated around the social anthropology that he represented and that alone was capable of leading this totalising enterprise. The basis for anthropology's particular vocation, according to Lévi-Strauss, is its ability to bridge the natural and human sciences and, therefore, it does not reject being classed among the natural sciences at the final call of judgement. Strengthened by his fruitful encounter with Jakobson during the war in the USA, Lévi-Strauss privileges the linguistic model within his anthropological project. His search for invariants and his paradigmatic and syntactic deconstructions incorporate the teaching of Jakobson's phonology, for example binary opposites and differential variations. His contribution ensured that linguistics engendered a particularly fecund area of scholarship during the post-war period. Although Lévi-Strauss guides anthropology in a cultural di-

rection, due to the emphasis placed on language and the decoding of signs, he does not, however, neglect his ambition for unity. His quest for mental perimeters also seeks out the field of biology. The wholeness, to which Lévi-Strauss aspires, incorporating also Marcel Mauss's ambition to construct the 'total social fact,' thus aims to embrace the entire scientific domain. It also aims to finally make an anthropology of the Science of Man; a federation of auxiliary sciences strengthened by logical-mathematical models, the contribution of phonology and by a limitless field of study, able at the same time to encompass societies devoid of history or writing across the globe.

The anthropologist is therefore able to access the unconscious of social practices and can reinstate the complex combinatories of the rules in force in all human societies. Such an ambition represented a major challenge for all sciences involved in the study of human beings. It also encouraged competitive reactions to this program from within other fields of knowledge or, on the other hand, a dependency upon a process of conquest with the aim of attaining a standing for those still marginalised disciplines in search of legitimation. The excessive nature of this ambition matches the difficulty faced by anthropology at its outset to position itself within the institution. Although anthropology alone did not succeed in emancipating the human sciences, structuralism, taking up the challenge, in fact became the common paradigm if not a common school for a whole range of disciplines while converging to construct a whole and unified science.

In 1953, against a backdrop of open crisis, Lacan gives his Rome Report . He is charged with paving an attractive and specifically French way towards the unconscious. In order to win this gamble, he searches for foundations, for institutional and theoretical guarantees. Lacan heads on the quest for lifelines. Therefore he is responsible for breathing new life into psychoanalysis, for curbing its crisis by using a strategy of offence and a dynamic of alliance. While Lacan employs all the means at his disposal he nonetheless uses any available intellectual material to turn things to his advantage, here proving rather more successful.

Lacan's re-reading of Freud is inscribed in the Saussurian tradition, prioritising the synchronic dimension. With this, he shares in the structuralist paradigm and prompts a new reading of Freud that no longer privileges the importance of stage theory but submits this to a basic oedipal structure characterized by its universality, independent of the relationship to temporal or spatial contingencies and preceding all history.

Contrary to Saussure, for whom the primary object was language, Lacan privileges speech, a displacement necessary due to the therapeutic practice. But this word does not represent the utterings of a conscious subject in control of his speech, but rather the opposite. Speech is forever cut off from all access to reality. They merely convey the signifiers that rebound between them. Man, therefore, only exists according to his symbolic function that, in turn, should apprehend him. Lacan thus presents a radical reversal of the notion of the subject as a product of language and its effect, as implied by the notorious proposition that 'the unconscious is structured like a language.' There is, therefore, no point in searching for the human essence anywhere else but in language. This is what Lacan meant when he says, 'language is an organ;' 'the human being is characterized by the fact that his organs are external to himself.' This symbolic function, upon which man's identity is founded, is opposed by Lacan in his Rome lecture to the language of bees that is entirely based on the stability of its relation to the reality it signifies. Lacan finds in the Saussurian sign, cut off from the referent, the semi-ontological kernel of the human condition.

This new vision of a decentered, split subject is compatible with the notion of the subject being developed in other structuralist fields within the human sciences. This subject is in some ways fictive, not existing outside of its symbolic dimension and of the signifier. Although the signifier is prioritised over the signified, there is nonetheless no question of the latter's abandonment. These two different levels therefore interact, a process that Lacan relates to Freud's discovery of the unconscious thus making of him, in Lacan's eyes, the first structuralist. The signifier even makes the signified undergo a sort of passion. As becomes clear here, Lacan submits Saussure to a certain twisting of his concepts and in as much as the notion of the sliding of the signified under the signifier would not have made sense to Saussure, so too the notion of the unconscious escaped him. Lacan takes up the two big rhetorical figures of metaphor and metonomy used by Jakobson to highlight the deployment of discourse and assimilates the two to the functionary mechanism of the unconscious that, structured like language, situates itself in total isology from the rules of the latter.

3. *A Global Semiological Project: 1964–66*

In 1964 there were several breakthroughs in the classical schema. Structuralism now has the possibility of appearing to be a program shared by several disciplines. It is, in fact, in 1964 that the new university at Nanterre was opened, the very center of literary modernity. It is in the same year that a special edition of *Communications* appears, edited by Roland Barthes and dedicated to semiology. Barthes here defines 'structuralist activity' as the conscience of paradox and the denaturalisation of meaning. At the same time, Lacan, definitively excommunicated by the International Freudian Association, establishes the ECF (*l'École de la cause freudienne*). His seminar moves from Sainte-Anne to the *École Normale*

Supérieure at the *rue d'Ulm*, the epicenter of humanities studies that had fallen prey to the veritable fever of scientism and epistemology favoured by Louis Althusser. Lacan was not only welcomed by Althusser there but he was also welcomed to the bosom of Marxism in a text entitled 'Lacan and Freud' in which he subscribes to the new direction taken by psychoanalysis as defined by Lacan—seeing in the latter's return to Freud the mirroring of his own return to Marx. It is in the same year that Lévi-Strauss embarks upon his signature work, *Mythologiques*, a tetralogy on myths beginning with the publication of *The Raw and the Cooked* in which mythical thought is described to be structured-like language.

It is around philosophy that the formulation of a global program going beyond the positivity of any specific discipline in the social sciences can be found. At the *École Normale Supérieure d'Ulm*, the epicenter of the circumvention of the strictly classical Sorbonne, the intellectual home of structuralism and that of Lacan and Derrida, philosophy around Louis Althusser is to become the 'Theory of Theoretical Practice,' the appropriation of the contribution of the new social sciences and their internal subversion. It is due to the return to Marx, evident in the seminars beginning at the outset of the 1960s, that Althusser proposes re-*Reading Marx* by establishing his work as the advent of Science following the epistemological break discernible in it around 1845. Two versions of Marx are, therefore proposed of which Althusser refers only to the second, that of science, structural causality and over-determination. Althusser reads Marx in a way that suggests a theoretical anti-humanism and a 'process devoid of either subject or object' as many of the themes being elaborated within the structuralist project of the 1960s in which man is spoken rather than speaking.

The other attempt led by the structuralist program in the effort to globalise itself is led by Michel Foucault during structuralism's peak year—1966—the year in which *The Order of Things* was published. Foucault, thus, announces the death of philosophy and its substitution by active thought, brought about due to logic and linguistics. Foucault prioritises two (social) sciences in particular, psychoanalysis and ethnology in that he sees them as 'counter-sciences' that enable the destabilising of both history and of the subject. Foucault sees humanism as our 'middle-ages' and the figure of man as a transitory one destined to disappear. It is once again within the figure of the unconscious of the human sciences that their truth may be uncovered behind their illusory pretension to reinstate man in a command made forever impossible by the three narcissistic wounds provoked by Galileo, Darwin and Freud.

The structuralist program is, however, deconstructed from within itself in that it departs from a principled double negation that could not but cause it to fall into an aporia of the subject and of historicity. The ontologisation of structure that resulted, that which Paul Ricoeur called 'Kantianism without the transcendental subject,' could not but lead it into a dead end. I support Paul Ricoeur's warning against the untimely dissociation of the absolute complementarity of the explanatory level, represented by semiological analysis and the interpretative stage or the reappropriation by the subject of the meaning of text, allowing him to endow it not only with meaning but also with signification. The joining of the semiological and hermeneutic approach enables a reappropriation of the discourse's various components. Paul Ricoeur presents them as a 'quadrilateral of discourse:' (a) the speaker as a singular event of speech; (b) the interlocutor's bringing out of the dialogical character of the discourse; (c) meaning or the very theme of the discourse which is the unique and exclusive dimension of structuralism to the exclusion of all three other dimensions; and (d) reference or that being spoken about. This kind of reappropriation allows for the emergence, not of a transparent subject or master of meaning, but of an ontology of action based upon a triangular relation systemised by Ricoeur in *Oneself as Another*, the relationship drawn between the self (ipse), the other (alterity) and the institutional foundations (community). This perspective thus seeks to 'regroup' the metaphysics of the act, interpersonal ethics and the problem of social ties, a task made all the more pressing by the discovery in the twentieth century of the 'fallibility' of democracy. Paul Ricoeur thus enables us to develop action open to historicity, creativity and to narrativity in order to understand better the constitution of both our personal and collective narrative identities.

See also: Anthropology; Deconstruction: Cultural Concerns; Postmodernism: Philosophical Aspects; Structuralism; Structuralism, Theories of

Bibliography

Barthes R 1953 *Le Degré Zéro de l'Écriture*. Editions du Seuil, Paris
Barthes R 1957 *Mythologies*. Editions du Seuil, Paris
Barthes R 1964 *Essais Critiques*. Editions du Seuil, Paris
Brandel F 1969 *Ecrits sur l'Histoire*. Flammarion, Paris
Dosse F 1991/1992 *Histoire du Structuralisme*. Editions la Découverte, Paris, 2 Vols.
Ducrot O, Todorov T 1972 *Dictionnaire Encyclopédique des Sciences du Language*. Editions la Découverte, Paris
Foucault M [1961] 1972 *Histoire de la Folie à l'Âge Classique*. Gallimard, Paris
Jakobson R 1963 *Essais de Linguistique Générale*. Minuit, Paris
Lacon J 1966 *Ecrits*. Editions du Seuil, Paris
Lévi-Strauss C 1949 *Structures Élémentaires de la Parenté*. Plon, Paris
Lévi-Strauss C 1952 [1973] Race et histoire. In: *Anthropologie Structurale Deux*. Plon, Paris
Lévi-Strauss C 1958 *Antropologie Structurale*. Plon, Paris

Lévi-Strauss C 1962 *La Pensée Sauvage*. Plon, Paris
Todorov T, Ducrot O, Sperber D, Safouan M, Wahl F 1968 *Qu'est-ce que Le Structuralisme?* Editions du Seuil, Paris
Vernant J P 1965 *Mythe et Pensée chez les Grecs*, 2nd edn. Maspero, Paris

F. Dosse

Structuralism, Theories of

Structuralism is an intellectual tendency that seeks to understand and explain social reality in terms of social structures. Influenced by the rising prestige of mid-nineteenth century science, theories of structuralism focused on structural form as an organizing principle underlying whole cultures and societies. Structural form is analytically distinct from cultural content such as meanings and norms, although culture and structure can be defined in terms of each other as, e.g., in cultural structuralism or in the idea of structure as a cultural pattern or model. In contrast to both culture and the reductionist view of society as an aggregate of individuals, social structure is defined as the forms of social relations (e.g., difference, inequality) among a set of constituent social elements such as positions, units, levels, regions and locations, and social formations (see *Structure: Social*).

Structuralism proceeds on two analytic levels: (a) as a method of analysis or procedure of knowing (epistemology), and (b) as an ontology or metaphysical design of social reality. It approaches its subject matter in terms of two meta-theoretical perspectives on social reality: (a) social structure as an empirical or historical reality, and (b) social structure as a model or representation of reality (see *Sociology, Epistemology of*; *Theory: Sociological*). The conceptual property space created by these analytical dimensions generates a fourfold typology that accommodates the major theories of structuralism existing today: (a1) sociological structuralism, (a2) symbolic structuralism, (b1) historical structuralism, and (b2) orthodox structuralism. Since the major symbolic and orthodox structuralist theories privilege simultaneity and synchrony over history and diachrony, b1 will only be discussed briefly in contrast to genetic structuralism in Sect. 2.

1. Sociological Structuralism

Marx and Durkheim are the two most important theorists to introduce structural concepts into the study of society. Both conceived of social structure in holistic and nonreductionist terms, i.e., not as an aggregation of individual actors, but as social facts *sui generis*, explainable only by other social facts. For Marx, the central social fact was the political economy; for Durkheim it was social ties and norms. Both were committed to the use of positivist scientific procedures in the study of society and believed in the necessity of empirical evidence and the importance of establishing objective causal connections between the phenomena to be explained and the structures explaining them. Both offered an analysis of the structural features of modern societies and were centrally concerned with the transition from premodern to modern forms of social organization. However, they developed different structural concepts and theories of historical and structural transformation.

1.1 Marx's Materialist Conception of History

Marx formulated the transition from feudal to capitalist society as a historical and structural shift in the mode of production, a high-level structural concept comprising the productive forces as the active source of continuous change and the social relations of production as the historical outcome of conscious human productive activity or self-transforming praxis. The developing contradictions between ever-changing productive forces and increasingly constraining social relations generate the possibility of the transformation of a given mode of production and the emergence of a new mode of production. Structural features such as the division of labor, the ownership and control of capital, the class structure of society, and the ideological representation of class conflict constitute a sociohistorical totality which tends to change with the progressive transformation of the mode of production. Resistance to change by social classes invested and interested in the status quo is the most proximate cause of class conflict and the possibility of social revolution. Marx's ([1859] 1904) reliance on the Hegelian causal mechanism of dialectical contradiction and on the internal dynamics of the mode of production led him to overestimate the likelihood of revolutionary change and to underestimate the relatively autonomous role of the state, legal institutions, and social movements in mediating between pressures for change and the collective accommodation to the status quo. The use of a self-transformative principle of dialectical causality intrinsic to the economic sphere to explain historical social change led to the charge that Marx's theory is one-dimensional, self-confirming, and not independently falsifiable or testable, hence worthless as a scientific theory. It can be argued, however, that the rise of the regulatory state in the twentieth century as well as specific forms of radical state intervention under state socialist or national socialist regimes interrupted global capitalist expansion, thus providing the negative evidence necessary to test and refute Marxian economic predictions. By the same token, the collapse of state socialism and the worldwide decline of regulation and intervention under neoliberal auspices have spurred the resumption of globalization. These historical processes, in turn, are likely to confirm certain Marxian expectations and to contribute to the theoretical rehabilitation of

Marxian theory (see *Marxism in Contemporary Sociology*).

1.2 *Durkheim's Sociological Realism* (*Holism*)

Durkheim ([1893] 1964) undertook to explain the transition to modern industrial society by invoking a different set of structural forces: population growth due to the agricultural revolution, the resulting rise of population density, the increase in the rate of social interaction, competition, and division of labor among differentiated and specialized social functions, and the emergence of social relations and associations among social groups and occupations. This transition from the mechanical solidarity of segmented but undifferentiated societies to the 'moral density' and organic solidarity of highly differentiated ones implied a structural theory of social change which was perfected in his work on the social causes of suicide. The resulting theory of social integration continues to be tested by empirical research. Durkheim formulated a set of methodological rules designed to put sociology on a scientific footing. His conception of social structure as a nonreducible 'social fact' was paralleled by culturalist notions of collective representations, collective conscience, as well as cognitive and normative constraints which he referred to as 'ways of acting, thinking, and feeling.' This formulation implied a theory of institutions combining structural and normative elements as 'social facts,' a theoretical framework transcending the distinction between structure and culture, just as Marx acknowledged the constant interplay of collective action, consciousness, ideology, and structure. Elements of Durkheim's and Marx's structuralism reappear in contemporary approaches, e.g. Bourdieu's (1989) constructivist structuralism and Giddens's (1984) theory of structuration.

1.3 *Modern Network Structuralism*

An important development in contemporary sociological structuralism is social network analysis (Scott 1991). Social networks, defined as the patterns of social relations among a set of actors or nodes, are generic social structures. Networks differ from organizations and institutions in that they are informal, private, self-organizing, noncontractual, unregulated, unaccountable, nontransparent, and typically of limited focus, size, and duration. Networks are the least structured social forms that can be said to possess any structure at all, yet whose structural configurations affect the behavior of their members. While networks are typically seen as real, empirical structures, they can be represented by formal models. As such, they are closely related to concepts like 'intersecting social circles' and the 'web of group affiliations' in Simmel's (1971, p. 23) formal sociology and Blau's (1998) theory of structural sociology (see *Networks: Social*).

2. *Symbolic Structuralism*

Sociological empiricism and positivism gave rise to a countervailing epistemological tendency claiming that the holistic structures identified by the new social sciences were nothing but cognitive models of reality since they necessarily had to rely on categories that represented, constituted, or constructed reality. This resurgence of neo-Kantian philosophy influenced Weber, Simmel, and Lévi-Strauss, among others, and served to justify the rejection of positivist causality and historical materialism.

Symbolic structuralism is a method of analysis that conceives of structure as a model or representation of reality. Typical is Saussure ([1915] 1966), who introduced certain methodological innovations into the study of language that would move linguistics from a descriptive historical discipline to an analytical, scientific one. The linguistic sign was seen as a holistic combination of two structural elements: a form that signifies (signifier) and a concept to which the form refers (signified). The relationship between signifier and signified is not natural or fixed, but social and arbitrary, hence infinitely variable. Linguistic symbols are meaningful only insofar as they produce identity-in-contrast, binary oppositions (difference), and formal distinctions. This method shifts the analytic focus from such metaphysical concepts as 'object,' 'reality,' or 'thing-in-itself' to the formal properties and the internal structural relations of sign systems. The content and use-value of symbols and objects are subordinated to their relational identity, their position, function, and exchange value within a given system.

A little-noted aspect of Saussure's linguistic structuralism is its origin in the categories of the capitalist economy, especially the 'radical duality' of economic history and the political economy as a system (1966, p. 79). Both economics and linguistics are 'confronted with the notion of *value*'; both are '*concerned with a system for equating things of different orders*—labor and wages in one and a signified and signifier in the other' (Saussure 1966, p.79, italics in original).

Saussure elevates the duality of system and history to central principles of linguistic analysis: the distinction between language and speech, and the priority of synchrony over diachrony. This methodological imperative entails the epistemological precedence of the code over its application, of the structure of the system over its origins and transformation, of exogenous over endogenous change. 'The linguist who wishes to understand a state must discard all knowledge of everything that produced it and ignore diachrony' (Saussure 1966, p. 81). Historical change involves merely the displacement of one form or structure by another. History becomes eventless history.

Piaget (1970) showed that these structuralist methods spread from linguistics to other disciplines,

e.g., gestalt psychology, psychoanalysis (e.g., the structure of the unconscious in Freud and Lacan), and structural anthropology (Lévi-Strauss). Piaget's (1970) own genetic structuralism replaces the historical study of social change with the notion of structure as a system of transformations. Three key elements here are the idea of wholeness, of structured wholes as determined by internal rules of transformation, and of the relative autonomy and self-regulation of structures. Piaget tacitly acknowledges the transition of structuralism from a method to an ontology by asserting that even though 'structures consist in their coming to be; that is, their being "under construction,"' they nevertheless depend on higher-level structures; i.e., 'on a prior formation of the instruments of transformation—transformation rules or laws' (1970, pp. 140–1). Distinct from genetic structuralism, historical structuralism (Marc Bloch, Lucien Febvre, Fernand Braudel and the Annales School) assumes the determinative influence of historical deep structures such as the *longue durée*.

The epistemological ambiguity of structures as models of reality or as pregiven structures appears in other sociological approaches, for example Georg Simmel's (1971) social forms and social types, Max Weber's 'ideal types' of religion and rationality, and Alfred Schutz's 'typifications,' the 'first-order constructs' (the primary, common-sense construction of social reality by participants) and 'second-order constructs' (the secondary, derivative social construction by the scientific observer) of the structures of the life-world (see Dosse 1997, p. 42 and *Phenomenology in Sociology*).

3. *Orthodox Structuralism*

The convergence of structure as a model or representation of reality and the ontological view of structure as substance constitutes orthodox structuralism. The analytic differences between structure and system are blurred. The idealist version is best represented by the work of Claude Lévi-Strauss, the materialist version by Louis Althusser.

3.1 *Lévi-Strauss and Anthropological Structuralism*

Lévi-Strauss's structural anthropology was influenced by Saussure's structuralist linguistics by way of Roman Jacobson and the phonological school of Prague as well as by Durkheim and Marcel Mauss (Dosse 1997, pp. 21–2, 29–30). In his initial attempt to formulate a method for producing a model of social reality, Lévi-Strauss develops a systemic concept of social structure implying the internal interdependence of elements or 'elementary structures' as, for example, in kinship systems. This conceptualization set the stage for the idea of a 'formal structural homology,' the reproducibility of structures on different levels such that a series of transformations can produce a group of models of the same general type, ascending from the 'order of elements' in the social structure to the 'order of orders,' and ultimately to some universal order (Lévi-Strauss 1967, pp. 302–9). For Lévi-Strauss, the advantage of structural models is that they make it possible to arrive at general laws and to specify the internal logic of social systems such that all the observed facts are immediately visible, an assumption shared with gestalt psychology and other approaches based on a coherence concept of truth rather than on a positivist correspondence concept of truth.

The search for general laws leads Lévi-Strauss to shift his attention from the study of conscious phenomena (speech, opinions, ideology) to that of the unconscious infrastructure of cognitive forms, basic codes, and myths. Here, then, the reality of social structure is but the surface manifestation of the deep structure of the mind. Ultimately, Lévi-Strauss can be said to take the structuralist model of reality from the agnostic dualism of Kant's *a priori* categories back to the pre-Socratic belief that reality can be found only in timeless, changeless 'Being.' Thus, it would be logically absurd to see history and social change as anything other than the recurrent rearrangements of internal relations and structural building blocks.

3.2 *Althusser's Marxist Structuralism*

The immediate textual source for the adaptation of Marxism to structuralism is Frederick Engels' formulation of historical materialism in terms of the complex interaction between the economic base and the political and ideological superstructure of society. In this tripartite distinction between relatively autonomous regions of a given social formation, each representing its own form of practice, 'the economic element finally asserts itself as necessary' (Engels, in Heydebrand 1981, p. 88). Althusser adapts Engels' dialectical concept of interaction among levels of social structure by developing a deterministic (rather than probabilistic) scientific model of 'contradiction and overdetermination.' This model rejects empiricism, construes dialectics as a dualism of opposites, separates 'science' from ideology, and asserts that the superstructure 'is relatively autonomous but the economy is determinant in the last instance' (Althusser 1970, p. 177). Since the structuralist concept of causality is deterministic rather than stochastic, a 'structure is always the co-presence of all its elements and their relations of dominance and subordination—it is an ever pre-given structure' (Althusser 1970, p. 255). For Althusser, theory is the work of transforming ideology into science. This 'theoretical practice' is subject only to its own determinations and is prior to, and independent of, the base–superstructure problematic. This formulation of materialist structuralism occurs at such a high level of abstraction that it becomes practically indistinguishable from idealist struc-

turalism. Althusser later recognized this problem and rejected his notion of theoretical practice as a 'theoreticist error, i.e., as an epistemological quest for categorical constants and orthodox theoretical principles that falsely impose a fixed, ontological order on the 'real object' by means of a reified 'object of knowledge.'

The controversies surrounding orthodox structuralism as an intellectual and ideological tendency have contributed to the clarification of structural concepts in the social sciences, but failed to formulate testable propositions. Sociological, symbolic, and historical structuralism, however, continue to exercise considerable theoretical influence in the social sciences and in the post-structuralism of Derrida and Foucault.

See also: Durkheim, Emile (1858–1917); Macrosociology–Microsociology; Marx, Karl (1818–89); Networks: Social; Piaget, Jean (1896–1980); Saussure, Ferdinand de (1857–1913); Simmel, Georg (1858–1918); Social Networks and Gender; Sociology: Overview; Structuralism; Structuralism, History of

Bibliography

Althusser L 1970 *For Marx*. Vintage, New York

Blau P 1998 *Culture and Social Structure*. In: Sica A (ed.) *What is Social Theory*. Blackwell, Oxford, UK, pp. 265–75

Bourdieu P 1989 Social space and symbolic power. *Sociological Theory* **7**(1): 14–25

Dosse F 1997 *History of Structuralism,* 2 vols. University of Minnesota Press, Minneapolis, MN

Durkheim E [1893] 1964 *The Division of Labor in Society*. Free Press, New York

Giddens A 1984 *The Constitution of Society: Outline of a Theory of Structuration*. Polity Press, Oxford, UK

Heydebrand W 1981 Marxist structuralism. In: Blau P, Merton R K (eds.) *Continuities in Structural Inquiry*. Sage, London, pp. 81–119

Lévi-Strauss C 1967 *Structural Anthropology*. Doubleday Anchor, New York

Marx K [1859] 1904 *A Contribution to the Critique of Political Economy*. Charles Kerr, New York

Piaget J 1970 *Structuralism*. Harper Torch, New York

Saussure F de [1915] 1966 *Course in General Linguistics*. McGraw-Hill, New York

Scott J 1991 *Social Network Analysis*. Sage, London

Simmel G 1971 *Georg Simmel on Individuality and Social Forms* [Levine D (ed.)]. University of Chicago Press, Chicago

W. V. Heydebrand

Structure: Social

The term structure (Latin *structura* from *struere*, to construct) has been first applied to construction and only later on, during the classical period, to the scientific field of biology. It began to be used, in the anatomic language, to refer to the interdependence of parts and to the corresponding mode of organization in living bodies. What is worth noticing is that the metaphor of construction on the one hand, the analogy with the organism on the other, are plainly involved in the two first main uses of the term with reference to human societies, in the mid-nineteenth century.

The metaphor is clearly used by Karl Marx, when he writes in The Preface to *A Contribution to Critique of Political Economy* (1859), the well-known sentence according to which 'the sum total of [the] relations of production constitute the economic structure (*Struktur*) of society, the real foundation, on which rises a legal and political superstructure (*Überbau*).' It was a characteristic of Herbert Spencer's thought to dwell on the analogy between organism and society, even if he used it mainly 'as a scaffolding': from the 'Prospectus of a System of Philosophy' (1858) on, he worked with a mental scheme associating structure and function, respectively, linked to anatomical phenomena and physiological processes.

These two original ideas have deeply influenced the development of the field and it should be noted that from the beginning, social structure has been conceptualized in quite different ways. Perhaps it may be added that the Marxian conception in terms of a 'real foundation' already implies the assumption of a sort of 'deep structure.'

The pronounced growth of (social) anthropology and sociology made social structure a key concept of both sciences but there never was an uncontested agreement about its definition; on the contrary, the term has been used in so many different senses that it will be quite difficult to draw up their complete list. Therefore, Merton was quite right in beginning his 'fourteen stipulations for structural analysis' by remarking 'that the … notion of ''social structure'' is polyphyletic and polymorphous' (Merton 1975).

We shall then limit our analysis to some landmarks in the history of the concept and in the more recent developments. Rather than trying at first to give a general definition, we shall emphasize the main dimensions which are, so to speak, the substratum of the most elaborated conceptions of social structure.

Our last section will be devoted on the one hand to the formulation of some theoretical principles which nowadays are (or can be) widely accepted and, on the other hand, to the major open questions scholars in the field are still faced with.

1. Some Landmarks in the History of the Concept and in Contemporary Advances

In this area of research as in many others, Emile Durkheim's work is a significant link between the

pioneers and later developments. *The Division of Labour in Society* (Durkheim 1893) can be read as based on a structural analysis. In this respect the most typical argument is to be found in Book II, in which Durkheim brings out the causes of division of labor in terms of increase of the population in the society, of its 'material density' and, mainly, of its 'dynamic or moral density.' However, it is not to be forgotten that in Book I he lays great stress on the function of a developing division of labor, which is to generate ('organic') solidarity. Thus, Durkheim gave an impulse to the morphological view on social structure, emphasizing the material and spatial distribution of population, while paying attention to the moral dimension attuned to the complexity of social relations in heterogeneous societies.

In keeping with some ideas derived from Durkheim (but not his attempt at synthesis), Radcliffe-Brown maintained the organic analogy, with the crucial distinction between structure and function and the attention given to the description of the interdependence of the component parts in human society. As he put it, 'in the study of social structure the concrete reality with which we are concerned is the set of actually existing relations … which link together … human beings' (Radcliffe-Brown 1940). What is required by 'scientific purposes is an account of the form of the structure, … abstracted from the variations of particular instances.' The British structural school of anthropology was imbued with Radcliffe-Brown's views. There is yet a book which stands out as a theoretical advance, namely Nadel's *The Theory of Social Structure* (Nadel 1957), in which he works out a conception of social structure in terms both of an elaborate idea of network and of a theory of roles.

In his attempt at building up such a theory, Nadel was bound to discuss some parts of Parsons' work, in which role is conceived as a major structural category. In Parsons' words, 'the structure of social systems … consists in institutionalized patterns of normative culture' (Parsons 1961). But norms vary according to the positions of the individual actors and also to the type of activity. On the one hand, they define roles, with the corresponding rules of behavior, on the other they constitute institutions. Social structure is thus to be studied in its essentially regulative properties and this approach strongly constrasts with the conception emphasizing the distributive dimension of social structure; we hit here upon an unmistakable dividing line.

There is another big divide which is worth noticing from an epistemological point of view, even if it conspicuously took place in the field of anthropology, bringing about dispute between different schools. It would be pointless in this article to outline an overview of such a composite movement as structuralism. However, for the social sciences it was significant in two main respects: first it meant the parting of the suitable ways for sociology and anthropology in so far as cultural anthropologists were requested to follow the lead of structural linguistics; and secondly it claimed a novel conception of structure. Levi-Strauss specificably states that 'the term "social structure" has nothing to do with empirical reality but with models which are built up after it' and consequently asserts that 'social structure can, by no means, be reduced to the ensemble of the social relations to be described in a given society' (Lévi-Strauss 1953). Therefore Levi-Strauss' followers are very keen on contrasting this '*transformational* structuralism,' emphasizing the transformability of a single relational set, with the still prevailing empirical views of social structure, grounded in the positivistic notion of 'correspondence rules.' It may perhaps be added that the rules of transformation connecting specific arrangements of the assumed underlying structure do not prevent it from being arbitrary; the argument is still needing some kind of empirical substantiation. Accordingly, Levi-Strauss is the most convincing, when he could draw upon a whole research tradition. His kinship models cannot be cut off from the large and momentous set of earlier studies in the field.

The term 'model' has often been used loosely in the social sciences; but, with regard to social structure, it generally has formal connotations. The third dividing line we want to emphasize is mainly methodological and deals with the relevant approach to social structure: must it be formal, as Peter Blau (Blau 1977) would defend it to be, or not? We have here to go back to a classical 'master of sociological thought,' namely Georg Simmel, for finding the father of the formal approach. He dwelt on 'The Number of Members as Determining the Sociological Form of the Group,' translated by Albion Small as early as 1902–3, stressed the importance of triads (vs. dyads) and analyzed the main conditions of group persistence, one of the most essential of which is organization. Peter Blau resumed this line of thought and enlarged it in a decidedly macrosociological direction. In his words, 'a social structure can be defined as a multidimensional space of different social positions among which a population is distributed.' His primary interest is 'in the two generic forms [social] differentiation assumes,' that is inequality and heterogeneity in 'the degree to which [they] intersect' and in their effect upon the integration of various segments of society; and he endeavors 'to construct a deductive theory of social structure,' by stating theorems deduced from primitive propositions. We lack space to enumerate all of them but at least two are worth noticing: the more heterogeneous social structure, the more likely intergroup relations; and the less correlated social differences, the more likely intergroup relations. These theoretical statements were later on put to an empirical test by Blau and Schwartz (Blau and Schwartz 1984).

Simmel likewise is held up as a pioneer by scholars belonging to a somewhat different and developing research subtradition—social network analysis. Its approach is positively structural, but its substantive

focus is generally microsociological; therefore in most cases it does not convey an overall view of social structure. Yet it can contribute a lot, by clarifying specific levels of structure and carrying on with other intellectual tools Merton's study of status—and role—sets and Homans' work on 'internal systems.' All the more as some synthesis has been brought about by Fararo and Skvoretz (Fararo and Skvoretz 1987) between social network analysis and the macro-structural approach, represented by Granovetter's 'strength-of-weak ties' theory and Blau's formal deductive one, respectively.

Harrison White's work has also been wholly dedicated to the investigation of structural problems. His last book, *Identity and Control* (White 1992) opens up new vistas, even if, as a commentator put it, it often requires as much attention as *Finnegans Wake*.

2. *Guiding Principles and Open Questions for Sociological Research*

First of all, to avoid any misunderstanding, we do not ingenuously intend to prescribe relevant rules for research. We have the only aim of stating theoretical propositions about which there is (or could be) general agreement. Some of them have been explicitly formulated by scholars. On other points, they have not and our statements will accordingly be more tentative. We shall succinctly comment on all of them.

To begin with, a first major principle has been tersely conveyed by Nadel: 'It seems impossible to speak of social structure in the singular.' No matter whether structure is conceived as a network of roles (Nadel) or whether it is conceptualized in terms of a multiplicity of social positions (Blau), the statement holds true. The ordering always is partial and relative both to the specific property being analyzed and to the systemic frame of reference chosen to describe it.

A second central principle has been laid down by Thomas Fararo (Fararo 1989): 'Structure is not an ultimate given but a conditional attainment.' Persistence (which is not to be mistaken for permanence) and endurance of social structures are problematic. In any case their stability has to be accounted for; the idea of equilibrium may be used as a heuristic device, provided attention is paid to moving equilibria as well as to static ones. Structures are maintained, when and if they do so, by underlying processes. These two principles are basic but some others are easily derived from them and there still are some points which are worth noticing.

Although it sounds like a cliché, it is to be remembered that structural characteristics are always constituted by the social activities of individuals, however contrary to their purposes the outcomes may be. The idea to be stressed upon here is that structures are generated by men's actions and interactions. It entails a decisive consequence for sociological research: structural analysis cannot be divorced from action theory. In other words, when we are directing our efforts to problems of social structure, we had better, as Fararo put it, 'start with … principles and models of action in social situations' (Fararo 1989).

Social structure is no more conceived as the determining factor of individuals' behavior and this view has been exploded. The Mertonian idea of a 'core process of choice between socially structured alternatives' in Stinchcombe's words (Stinchcombe 1975), is more to the point. To go further in the matter, we could say that social structure indeed is constraining but it is also enabling.

The notion of 'structural effects' is not questionable at all, although it would be better to conceptualize the corresponding phenomena in terms of emerging structural properties. But the idea of structural causality is at least dubious. There has been an impressive demise of what Raymond Boudon aptly called 'magic structuralism' (1968), even if it never was as widespread in the social sciences as in the cultural and literary studies.

It certainly is a moot point, whether and how social structure and culture are linked. But whatever importance is to be attached to cultural dimensions, social systems have to be, on an analytical level, clearly distinguished from cultural systems: Parsons has here given the lead. We can thus point out a major flaw in symbolic or transformational structuralism: it fails to specify 'the ways through which' the symbolic rules, the so-called universal rules of the human mind, 'impinge on the actual workings of societies' and 'to identify the social mechanisms through which … the principles of deeper structure are activated within institutional frameworks' (Eisenstadt in Blau and Merton 1981).

In the first part of this article, we dwelt on two contrasting epistemological positions: on one side, there are scholars linking social structure to empirical reality, on the others, scholars linking it to models. Now, this dividing line can be tackled from another point of view and perhaps understood in a less radical way. As Nadel and Blau both make it clear, the main question is whether scholars intend social structure for description or explanation. These two conceptions obviously are quite different; still you cannot describe without some selection, i.e., abstraction process: description always is the first step in explanation. In this respect, Mario Bunge's distinction between a model object, which defines a representation principle for a class of empirical phenomena, and a theoretical model, constructed with the intention of explaining these phenomena, perhaps is not irrelevant.

The stress put on the selection and abstraction process may be useful for guarding against too naturalistic a view of social structure. Structural properties are what we are interested in, not an 'entity'

called social structure. In this regard, Giddens' statement, 'structure only exists as "structural properties"' (Giddens 1979), makes good sense. But this conception, implying a principle of 'order,' a 'real' virtuality,' in Gudmund Hernes' words (in Hedström and Swedberg 1998) is still compatible with a realist epistemology.

A quite different point is worth clearing up: we began by quoting Marx but later on we did not deal with class structure and the reader may wonder why. It is indeed a significant field to study and it would deserve a paper by itself. But class structure is no more to be understood as the rock-bottom foundation of society: the continental-European 'structural marxism' is now obsolete, as the whole of our theoretical and epistemological argument, we hope, makes it clear.

Finally, we shall borrow from Merton his last and wise stipulation: 'like any other theoretical orientation in sociology, structural analysis can lay no claim to being capable of accounting exhaustively for social and cultural phenomena' (Merton 1975).

Such are the guiding lines along which scholars' work can proceed; but there are still major questions on the research agenda. Two of them especially are, in our opinion, worth noticing. The first one requires a kind of theoretical synthesis. We have previously in this paper distinguished two main dimensions of social structure: the distributive and the regulative. Yet they are strongly connected and we need an approach accounting for both of them. The durkheimian lead seems to have been forgotten and it would theoretically be a strategical move to follow it again: it is somewhat ironic that Durkheim could be read both ways (and brilliantly, by Parsons and by Blau). As regards the latter one, the scholarly investigation must fully take account of the conception of social structure as a multilevel phenomenon. Therefore due attention must be paid to the micro-macro link and to the social mechanisms and processes through which it is operating; and the on-going interaction of the micro- and macro-levels would need a more general model (from macro to macro through micro) and the identification, at each step, of the appropriate types of mechanisms.

Of course social scientists will address problems of social structure with diverse theoretical views and differing 'domain assumptions.' But today all of us know that theoretical pluralism is not a sign of the immaturity of sociology as a science.

See also: Biology's Influence on Sociology: Human Sociobiology; Class: Social; Functionalism in Sociology; Institutions; Integration: Social; Labor, Division of; Macrosociology–Microsociology; Marx, Karl (1818–89); Networks: Social; Parsons, Talcott (1902–79); Social Change: Types; Social Constructivism; Social Geography; Sociology, History of; Spencer, Herbert (1820–1903); Status and Role: Structural Aspects; Structuralism, Theories of; System: Social; Theory: Sociological

Bibliography

Blau P M 1977 *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. Free Press, New York

Blau P M, Merton R K (eds.) 1981 *Continuities in Structural Inquiry*. Sage Publications, London and Beverly Hills

Blau P M, Schwartz J 1984 *Cross-cutting Social Circles*. Academic Press, Orlando, CA

Boudon R 1998 *A quoi sert la notion de "structure"?* Gallimard, Paris. [trans. Vaugham M 1971 *The Uses of Structuralism*. Heinemann, London]

Durkheim E 1893 *De la division du travail social*. Alcan, Paris. [trans. Halls W D 1984 *The Division of Labor in Society*, Free Press, New York]

Fararo T J 1989 *The Meaning of General Sociological Theory*. Cambridge University Press, Cambridge, UK

Fararo T, Skvoretz J 1987 Unification research programs: Integrating two structural theories. *American Journal of Sociology* **92**: 1183–209

Giddens A 1979 *Central Problems of Social Theory: Action, Structure and Contradiction in Social Analysis*. Macmillan, London

Hedström P, Swedberg R (eds.) 1998 *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge University Press, Cambridge, UK

Lévi-Strauss C 1949 *Les structures élémentaires de la parenté*, 1st edn. Gallimard, Paris [trans. Bell J H, von Sturmer J R, Needham R (ed.) 1969 *The Elementary Structures of Kinship*. Beacon Press, Boston]

Lévi-Strauss C 1953 Social structure. In: Kroeber A L (ed.) *Anthropology Today: An Encyclopedic Inventory*. The University of Chicago Press, Chicago

Marx K 1913 *A Contribution to Critique of Political Economy*. Kerr, Chicago

Merton R K 1975 Structural analysis in sociology. In: Blau P M (ed.) *Approaches to the Study of Social Structure*. Free Press, New York

Nadel S F 1957 *The Theory of Social Structure*. Cohen and West, London

Murdock G P 1949 *Social Structure*. Macmillan, New York

Parsons T 1961 An outline of the social system. In: Parsons T, Shils E, Naegele K D, Pitts J R (eds.) *Theories of Society: Foundations of Modern Sociological Theory*. Free Press of Glancoe, New York, pp. 30–79

Radcliffe-Brown A R 1940 Social structure. In: Evans-Pritchard E E, Eggan F (eds.) *Structure and Function in Primitive Society*. Cohen & West, London, pp. 188–204

Spencer H 1858 Prospectus of a system of philosophy. In: Rumney J (eds.) 1966 *Herbert Spencer's Sociology: A Study in the History of Social Theory*. Atherton, New York, pp. 297–303

Simmel G 1902–3 The number of members as determining the sociological form of the group. *American Journal of Sociology* **8**: 1–46, 158–96

Simmel G 1908 *Soziologie: Untersuchungen über die Formen der Vergesellschaftung*. Dunker & Humblot Leipzig, Germany [trans. Wolff K H, Bendix R 1955 *Conflict and the Web of Group Affiliations*. Free Press, Glencoe, IL]

Spencer H 1876–96 *Principles of Sociology*. Appleton, London
Stinchcombe A L 1975 Merton's theory of social structure. In: Coser L (ed.) *The Idea of Social Structure: Papers in Honor of Robert K. Merton*. Harcourt, Brace, Jovanovich, New York
White H C 1992 *Identity and Control: A Structural Theory of Social Action*. Princeton University Press, Princeton, NJ

F. Chazel

Subaltern History

1. Gramsci in India

In the form in which it is currently known, subaltern historiography is derived from the writings of a group of historians of modern South Asia whose work first appeared in 1982 in a series entitled *Subaltern Studies* (Amin and Chakrabarty 1996, Arnold and Hardiman 1994, Bhadra et al. 1999, Chatterjee and Jeganathan 2000, Chatterjee and Pandey 1992, Guha 1982–9). The term 'subaltern' was borrowed by these historians from Antonio Gramsci, the Italian Marxist, whose writings in prison in the period 1929–35 had sketched a methodological outline for a 'history of the subaltern classes' (Gramsci 1971). In these writings, Gramsci used the word 'subaltern' (*subalterno* in the Italian) in at least two senses. In one, he used it as a code for the industrial proletariat. But against the thrust of orthodox Marxist thinking, he emphasized that in its rise to power, the bourgeoisie did not simply impose a domination through the coercive apparatus of the state but transformed the cultural and ideological institutions of civil society to construct a hegemony over society as a whole, even eliciting in the process the acquiescence of the subaltern classes. In Gramsci's analysis of capitalist society, the central place is occupied by questions such as the relation of state and civil society, the connections between the nation, the people, the bourgeoisie and other ruling classes, the role of intellectuals in creating the social hegemony of the bourgeoisie, strategies for building a counter-hegemonic alliance, etc.

In the second sense, Gramsci talked of the subaltern classes in precapitalist social formations. Here he was referring to the more general relationship of domination and subordination in class-divided societies. But specifically in the context of southern Italy, he wrote about the subordination of the peasantry. Gramsci was very critical of the negative and dismissive attitude of European Marxists towards the culture, beliefs, practices and political potential of the peasantry. Positioning himself against this attitude, he wrote of the distinct characteristics of the religious beliefs and practices, the language and cultural products, the everyday lives and struggles of peasants, and of the need for revolutionary intellectuals to study and understand them. But he also highlighted the limits of peasant consciousness which, in comparison with the comprehensiveness, originality and active historical dynamism of the ruling classes, was fragmented, passive and dependent. Even at moments of resistance, peasant consciousness remained enveloped by the dominant ideologies of the ruling classes. These discussions by Gramsci were turned to productive use by South Asian historians writing in the 1980s.

2. Against Elitist Historiography

Ranajit Guha, who edited the first six volumes of *Subaltern Studies*, wrote in the introductory statement of the project: 'The historiography of Indian nationalism has for a long time been dominated by elitism—colonialist elitism and bourgeois-nationalist elitism' (Guha 1982). The objective of subaltern historiography was to oppose the two elitisms. The field of modern South Asian history was dominated in the 1970s by a debate between a group of historians principally located in Cambridge, UK, and another based mainly in Delhi, India. The former argued that Indian nationalism was a bid for power by a handful of Indian elites who used the traditional bonds of caste and communal ties to mobilize the masses against British rule. The latter spoke of how the material conditions of colonial exploitation created the ground for an alliance of the different classes in Indian society and how a nationalist leadership inspired and organized the masses to join the struggle for national freedom. Guha argued that both these views were elitist—the former representing a colonial elitism and the latter a nationalist elitism. Both assumed that nationalism was wholly a product of elite action. Neither history had any place for the independent political actions of the subaltern classes.

In setting their agenda against the two elitisms, the historians of *Subaltern Studies* focused on two main issues. One was the difference between the political objectives and methods of colonial and nationalist elites on the one hand and those of the subaltern classes on the other. The second was the autonomy of subaltern consciousness. Pursuing the first question, the historians of *Subaltern Studies* showed that the claim of colonialist historians that the Indian masses had been, so to speak, duped into joining the anti-colonial movement by Indian elites using primordial ties of kinship and clientelism was false. They also showed that it was untrue to say, as nationalist historians did, that the political consciousness of the subaltern classes was only awakened by the ideals and inspiration provided by nationalist leaders. It was indeed a fact that the subaltern classes had often entered the arena of nationalist politics. But it was also a fact that in many instances they had refused to join,

despite the efforts of nationalist leaders, or had withdrawn after they had joined. In every case, the goals, strategies and methods of subaltern politics were different from those of the elites. In other words, even within the domain of nationalist politics, the nationalism of the elites was different from the nationalism of the subaltern classes.

The second question followed from the first. If subaltern politics was indeed different from that of elite politics, what was the source of its autonomy? What were the principles of that politics? The answer that was suggested was that subaltern politics was shaped by the distinct structure of subaltern consciousness. That consciousness had evolved out of the experiences of subordination—out of the struggle, despite the daily routine of servitude, exploitation and deprivation, to preserve the collective identity of subaltern groups. Where was one to look for the evidence of this autonomous consciousness? It could not be found in the bulk of the archival material that historians conventionally use, because that material had been prepared and preserved by and for the dominant groups. For the most part, those documents only show the subaltern as subservient. It is only at moments of rebellion that the subaltern appears as the bearer of an independent personality. When the subaltern rebels, the masters realize that the servant too has a consciousness, has interests and objectives, methods and organization. If one had to look for evidence of an autonomous subaltern consciousness in the historical archives, then it would be found in the documents of revolt and counterinsurgency.

The first phase of the work of *Subaltern Studies* was dominated by the theme of peasant revolt. Ranajit Guha's *Elementary Aspects of Peasant Insurgency in Colonial India* (1983) is the key text in this area. But most other scholars associated with the project wrote on the history of peasant revolt from different regions and periods of South Asian history. They were able to discover a few sources where the subaltern could be heard telling his or her own story. But it was always clear that there would be few such sources. What became far more productive were new strategies of reading the conventional documents on peasant revolts. The historians of subaltern politics produced several examples of ways in which reports of peasant rebellion prepared by official functionaries could be read from the opposite standpoint of the rebel peasant and thus used to shed light on the consciousness of the rebel. They also showed that when elite historians, even those with progressive views and sympathetic to the cause of the rebels, sought to ignore or rationally explain away what appeared as mythical, illusory, millenarian or utopian in rebel actions, they were actually missing the most powerful and significant elements of subaltern consciousness. The consequence, often unintended, of this historiographical practice was to somehow fit the unruly facts of subaltern politics into the rationalist grid of elite consciousness and to make them understandable in terms of the latter. The autonomous history of the subaltern classes, or to put it differently, the distinctive traces of subaltern action in history, were completely lost in this historiography.

3. *The Imbrication of Elite and Subaltern Politics*

The analysis of peasant insurgency in colonial India and of subaltern participation in nationalist politics by the historians of *Subaltern Studies* amounted to a strong critique of bourgeois–nationalist politics and of the postcolonial state. Writing about peasant revolts in British India, Ranajit Guha (1983) and Gautam Bhadra (1994) showed how this powerful strand of anticolonial politics, launched independently of bourgeois-nationalist leaders, had been denied its place in established historiography. Gyanendra Pandey (1984, 1990), David Hardiman (1984), Sumit Sarkar (1984) and Shahid Amin (1995) wrote about the two domains of elite and subaltern politics as they came together in the nationalist movement led by the Congress. Dipesh Chakrabarty (1989) wrote about a similar split between elite and subaltern politics in the world of the urban working class. Partha Chatterjee (1986, 1993) traced the development of nationalist thought in India in terms of the separation of elite and subaltern politics and the attempts by the former to appropriate the latter. The postcolonial nation-state had, it was argued, included the subaltern classes within the imagined space of the nation but had distanced them from the actual political space of the state. Although the political emphasis was not the same in each writer, there was nevertheless a strong flavor in *Subaltern Studies* of the Maoist politics that had hit many parts of India in the 1970s. Many critics thought they could detect in it a romantic nostalgia for a peasant armed struggle that never quite took place. Others alleged that by denying the unifying force of the nationalist movement and stressing the autonomous role of the subaltern classes, the historians of *Subaltern Studies* were legitimizing a divisive and possibly subversive politics.

Another connection that was often made in the early phase of *Subaltern Studies* was with the 'history from below' approach popularized by British Marxist historians. Clearly, the work of Christopher Hill, E. P. Thompson, Eric Hobsbawm or the *History Workshop* writers was eagerly mined by subalternist historians for methodological clues for doing popular history. But there was a crucial difference because of which subaltern history could never be a 'history from below.' The forgotten histories of the people, pulled out from under the edifice of modern capitalist civilization in the West, had undoubtedly made the story of Western modernity more detailed and complete. But there did not seem to be any narrative of

'history from below' that could persuasively challenge the existence, stability or indeed the historical legitimacy of capitalist modernity itself. Not surprisingly, 'history from below' was invariably written as tragedy. But in countries such as India, 'history from below' could not be confined within any such given narrative limits. The subalternist historians refused to subscribe to the historicist orthodoxy that what had happened in the West, or in other parts of the world, had to be repeated in India. They rejected the framework of modernization as the necessary emplotment of history in the formerly colonial countries. They were skeptical about the established orthodoxies of both liberal–nationalist and Marxist historiographies. As a result, they resisted in their writings the tendency to construct the story of modernity in India as the actualization of modernity as imagined by the great theorists of the Western world. This resistance, apparent even in the early phase of *Subaltern Studies*, was to be expressed later in arguments about 'other modernities.'

4. *Reorientations*

One line of argument of the early *Subaltern Studies* project that ran into serious problems concerned the existence of an autonomous subaltern consciousness. Much of the study of the insurgent peasant was a search for a characteristic structure of peasant consciousness, shaped by its experience of subordination but struggling ceaselessly to retain its autonomy. One problematic question that arose was about the historicity of this structure. If subaltern consciousness was formed within specific historical relations of domination and subordination, then could that consciousness change? If so, why could it not be said that the Indian peasantry was transformed, modernized, turned into citizens by its experience of nationalist politics? Why the resistance to a progressive narrative of history? Or was there another narrative of a changing subaltern consciousness? It was a classic structuralist impasse to which there was no easy historical answer.

A related problem was about the notion of the subaltern as an active historical agent. Research into subaltern history had shown that the subaltern was both outside and inside the domains of colonial governance and nationalist politics. To the extent that it was outside, it had retained its autonomy. But it had also entered those domains, participated in their processes and institutions and thereby transformed itself. Every bit of historical evidence was pointing to the fact that the subaltern was 'a deviation from the ideal'. Why then the search for a 'pure structure' of subaltern consciousness? Moreover, argued Gayatri Spivak (1987, 1988) in two influential articles, subaltern history had successfully shown that the 'man' or 'citizen' who was the sovereign subject of bourgeois history writing was in truth only the elite. Why was it necessary now to clothe the subaltern in the costume of the sovereign subject and put him on stage as the maker of history? Subaltern historiography had in fact challenged the very idea that there had to be a sovereign subject of history possessing an integral consciousness. Why bring back the same idea into subaltern history? It was only a myth that the subaltern could directly speak through the writings of the historian. In fact, the historian was only representing the subaltern on the pages of history. The subaltern, announced Spivak, cannot speak.

The new turn in *Subaltern Studies* began more or less from the fifth and sixth volumes published in 1987–89. It was now acknowledged with much greater seriousness than before that subaltern histories were fragmentary, disconnected, incomplete, that subaltern consciousness was split within itself, that it was constituted by elements drawn from the experiences of both dominant and subordinate classes. Alongside the evidence of autonomy displayed by subalterns at moments of rebellion, the forms of subaltern consciousness undergoing the everyday experience of subordination now became the subject of inquiry. Once these questions entered the agenda, subaltern history could no longer be restricted to the study of peasant revolts. Now the question was not 'What is the true form of the subaltern?' The question had become 'How is the subaltern represented?' 'Represent' here meant both 'present again' and 'stand in place of'. Both the subjects and the methods of research underwent a change.

One direction in which the new research proceeded was the critical analysis of texts. Once the question of the 'representation of the subaltern' came to the fore, the entire field of the spread of the modern knowledges in colonial India was opened up for subaltern history. Much studied subjects such as the expansion of colonial governance, English education, movements of religious and social reform, the rise of nationalism—all of these were opened to new lines of questioning by the historians of *Subaltern Studies*. The other direction of research concentrated on the modern state and public institutions through which modern ideas of rationality and science and the modern regime of power were disseminated in colonial and postcolonial India. In other words, institutions such as schools and universities, newspapers and publishing houses, hospitals, doctors, medical systems, censuses, registration bureaux, the industrial labor process, scientific institutions—all of these became subjects of subaltern history-writing.

5. *Other Modernities*

A major argument that has been developed in the more recent phase of writings in and around *Subaltern Studies* is that of alternative or hybrid modernities.

The focus here is on the dissemination of the ideas, practices and institutions of Western modernity under colonial conditions. The framework of modernization theory invariably produces the history of modernity in non-Western colonial countries as a narrative of lag or catching up. As Dipesh Chakrabarty has put it, these societies seem to have been consigned for ever to 'the waiting room of history'. The universality of Western modernity erases the fact that like all histories, it too is a product of its local conditions. When transported to other places and times, it cannot remain unaffected by other local conditions. What happens when the products of Western modernity are domesticated in other places? Do they take on new and different shapes—shapes that do not belong to the original? If they do, are we to treat the changes as corruptions? As deviations from an ideal? Or are they valid as examples of a different modernity?

To argue the latter is both to provincialize Europe and to assert the identity of other cultures even as they participate in the presumed universality of modernity. Chakrabarty (2000), Arnold (1993), Prakash (1999), Spivak (1999) and Chatterjee (1995), for example, have explored various aspects of this process of the 'translation' of modern knowledges, technologies and institutions. They have tried to show that the encounter between Western forms of modernity and colonized non-Western cultures was not a simple imposition of the one on the other, nor did it lead to corrupt or failed forms of modernity. Rather, it produced different forms of modernity whose marks of difference still remain subject to unresolved contestations of power. Here, the work of South Asian subalternist historians has often overlapped with and contributed to what has become known as post-colonial studies. The historical study of modern discourses and institutions of power in colonial India has fed into a growing literature on the production of hybrid cultural forms in many different regions of the formerly colonial world. More significantly, post-colonial studies has extended the argument about hybrid cultural forms to the understanding of contemporary cultures in the Western metropolitan countries themselves, instanced most immediately in the diasporic cultures of immigrants but also, less obviously, in the role of the colonial experience in the formation of Western modernity even in its purportedly 'original' form. Historical and social science disciplines have tended to merge here with the concerns and methods of the literary and cultural disciplines to break new, and still mostly uncharted, theoretical grounds.

6. *Rethinking the Political*

To return to the initial political questions raised by subaltern history writing: where have they led? The early emphasis on peasant rebellion and consciousness had widened, even in the first phase of the project, to include the resistance of other dominated and marginalized groups in colonial society. Once the idea of a paradigmatic structure of subaltern consciousness became less persuasive, the subject of resistance began to be approached from a variety of angles and without a fixed design either of the reproduction of a traditional structure or of the necessary transition to new structures. Recent subaltern historiography has vigorously participated in three sets of South Asian debates—over religious minorities, caste and gender. These debates have opened up the way to rethinking the political formation of the nation as well as the political process of democracy.

In India, the broader political debate over religious minorities has been carried out between two opposed groups—the Hindu chauvinists on the one hand and the secularists on the other. What the researches of subalternist historians have shown is that the debate between secularism and communalism is in no way a struggle between modernity and backwardness. Both the rival political positions are firmly planted in the soil of modern government and politics. Second, the two groups are pursuing two different strategies of consolidating the regime of the modern nation-state. Both strategies are elitist, but they involve two different modes of representation and appropriation of the subaltern. Third, faced with these rival elitist strategies, subaltern groups in India are devising, in their own ways, independent strategies of coping with communal as well as secularist politics. The recent experience of ethnic violence and authoritarian politics in Sri Lanka and Pakistan have raised even more fundamental questions about the adequacy of the nation-form as a field where subaltern politics might be negotiated (Chatterjee and Jeganathan 2000).

The second question on which there has been significant recent discussion is caste. There has been a transformation in the politics of caste in India following the agitations in 1990–1 over the Mandal Commission report. By looking at the politics of caste in its discursive aspect, it has become clear that the supposedly religious basis of caste divisions has now completely disappeared from public debate. The conflicts now are almost exclusively centered on the relative positions of different caste groups in relation to the state. Second, the debate over whether or not to recognize caste as a criterion for affirmative action by the state once again reflects two different elitist strategies of representing and appropriating the subaltern. And subaltern groups too, in their efforts to establish social justice and self-respect, are devising various strategies of both resisting the state as well as of using the opportunities offered by its electoral and developmental functions. Strategies of alliance between castes at the middle and bottom rungs of the ritual hierarchy with other oppressed groups such as tribals and religious minorities have produced significant electoral successes. But with the creation of new

political elites out of subaltern groups, the questions of 'who represents?' and 'to what end?' are being asked with a new urgency.

The third debate concerns the social position of women. In one sense, all women living in patriarchal societies are subalterns. Yet, it is not true that women are not identified by class, race, caste, and community. Hence, just as it is valid to analyse the subordination of women in a society ruled by men, so also is it necessary to identify how the social construction of gender is made more complex by the intervention of class, caste, and communal identities. Recent discussions on this question have focused on the Indian social reform movements of the nineteenth century, especially the various legal reforms to protect the rights of women, in the context of the colonial state and nationalist politics. Subaltern feminist writings have raised questions about the adequacy of a modernizing agenda of legal reform from the top without facing up to the challenge of reforming the actual structures of patriarchal power within the local communities which continue to flourish outside the reach of the law.

The recent debates raise new questions about conceptualizing old modernist ideas such as nation, citizenship and democracy. By virtue of these implications, the recent subaltern history writings from South Asia have been productively used in writings on the history of modernity in other parts of the formerly colonized world, such as, for instance, on nationalism and gender in the Middle East or on the politics of peasant and indigenous groups in Latin America. Having traveled from Italy to India, the idea of subaltern history has now produced a generally available methodological and stylistic approach to modern historiography that could be used anywhere.

See also: Caste; Colonialism, Anthropology of; Colonization and Colonialism, History of; Gramsci, Antonio (1891–1937); Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Historiography and Historical Thought: South Asia; Minorities; Peasants and Rural Societies in History (Agricultural History); Peasants in Anthropology

Bibliography

Amin S 1995 *Event, Metaphor, Memory: Chauri Chaura 1922–1992*. Oxford University Press, Delhi, India

Amin S, Chakrabarty D (eds.) 1996 *Subaltern Studies IX*. Oxford University Press, Delhi, India

Arnold D 1993 *Colonizing the Body: State Medicine and Epidemic Disease in Nineteenth-century India*. University of California Press, Berkeley, CA

Arnold D, Hardiman D (eds.) 1994 *Subaltern Studies VIII*. Oxford University Press, Delhi, India

Bhadra G 1994 *Iman o nishan: banglar krishak chaitanyer ek adhyay* [*Faith and the Flag: An Aspect of Peasant Consciousness in Bengal*]. Subarnarekha, Calcutta, India

Bhadra G, Prakash G, Tharu S (eds.) 1999 *Subaltern Studies X*. Oxford University Press, Delhi, India

Chakrabarty D 1989 *Rethinking Working-class History: Bengal 1890–1940*. Princeton University Press, Princeton, NJ

Chakrabarty D 2000 *Provincializing Europe: Postcolonial Thought and Historical Difference*. Princeton University Press, Princeton, NJ

Chatterjee P 1986 *Nationalist Thought and the Colonial World: A Derivative Discourse*? Zed Books, London

Chatterjee P 1993 *The Nation and its Fragments: Colonial and Postcolonial Histories*. Princeton University Press, Princeton, NJ

Chatterjee P (ed.) 1995 *Texts of Power: The Disciplines in Colonial Bengal*. University of Minnesota Press, Minneapolis, MN

Chatterjee P, Jeganathan P (eds.) 2000 *Subaltern Studies XI*. Permanent Black, Delhi, India

Chatterjee P, Pandey G (eds.) 1992 *Subaltern Studies VII*. Oxford University Press, Delhi, India

Gramsci A 1971 *Selections from the Prison Notebooks* [trans. Hoare Q, Nowell Smith G]. International Publishers, New York

Guha R (ed.) 1982–9 *Subaltern Studies I–VI*. Oxford University Press, Delhi, India

Guha R 1982 On the historiography of Indian nationalism. In: Guha R (ed.) *Subaltern Studies I*. Oxford University Press, Delhi, India, pp. 1–8

Guha R 1983 *Elementary Aspects of Peasant Insurgency in Colonial India*. Oxford University Press, Delhi, India

Hardiman D 1984 *Peasant Nationalists of Gujarat*. Oxford University Press, Delhi, India

Pandey G 1984 *The Ascendancy of the Congress in Uttar Pradesh, 1926–1934*. Oxford University Press, Delhi, India

Pandey G 1990 *The Construction of Communalism in Colonial North India*. Oxford University Press, Delhi, India

Prakash G 1999 *Another Reason*. Princeton University Press, Princeton, NJ

Sarkar S 1984 *Popular Movements and Middle-class Leadership in Late Colonial India*. K P Bagchi, Calcutta, India

Spivak G C 1987 Subaltern studies: Deconstructing historiography. In: Guha R (ed.) *Subaltern Studies IV*. Oxford University Press, Delhi, India, pp. 338–63

Spivak G C 1988 Can the subaltern speak? In: Nelson C, Grossberg L (eds.) *Marxism and the Interpretation of Culture*. University of Illinois Press, Urbana, IL

Spivak G C 1999 *A Critique of Postcolonial Reason*. Harvard University Press, Cambridge, MA

P. Chatterjee

Subaltern Studies: Cultural Concerns

Subaltern Studies is an Indian school of historiography whose inspiration lay in the Maoist movement of the 1970s, and whose *raison d'être* has been the critique of the perceived elitist bias of Indian nationalist discourse in history writing. Since 1983, when the first

volume appeared, Subaltern Studies has produced 10 volumes of collected research articles, which comprise the main corpus. After the appearance of the sixth volume, a collective has managed editorial work. Individual members of the collective have also written texts that exemplify the 'subaltern' viewpoint.

1. The Origins

The school was founded by Ranajit Guha, a Marxist intellectual from Bengal. Once a member of the Communist Party of India, Guha was influenced by the radicalism of the Chinese Cultural Revolution in the late 1960s to align himself with Indian Maoism, which characterized independent India as a semifeudal and semicolonial state. Arguably, the project's impetus derived from an effort to establish the truth of this proposition: 'the price of blindness about the structure of the colonial regime as a dominance without hegemony has been, for us, a total want of insight into the character of the successor regime too as a dominance without hegemony' (Guha 1989, p. 307). However, Subaltern Studies has changed a great deal since then. Ranajit Guha is acknowledged by the collective to be its intellectual driving force and edited the first six volumes. He is also the author of *Elementary Aspects of Peasant Insurgency in Colonial India* (Guha 1983), recognized as a seminal work in the genre of radical Indian historiography.

Subaltern Studies began with an attempt to apply the approach known as 'history from below' in the Indian context. The term *subaltern* was inspired by the Italian Marxist Antonio Gramsci, and is used to indicate powerlessness in a society wherein class differentiation, urbanization, and industrialization had proceeded very slowly. Gramsci is the source of the subalternist respect for 'culture' and the 'fragment' in history writing. Subaltern Studies was also influenced by the critical Marxism of the English historian E. P. Thompson, who attempted to move beyond economistic definitions of class interest, as well as by Foucauldian critiques of power/knowledge systems. Its (earlier) Maoist orientation and polemical concern with the nationalism of Indian colonial elites made for a slant toward the investigation of peasant rebellion under colonial rule as well as peasant recalcitrance vis-à-vis Gandhian nationalism and the Indian National Congress.

2. The Significance of 'Culture'

Initially overtly political in its stance, Subaltern Studies was popular among young Indian historians and scholars of modern India abroad as a radical alternative to an uncritical academic celebration of Independence. The Indian national movement was seen as a failed hegemonic project (Sen 1987, pp. 203–35, Guha 1983, 1989, pp. 210–309). Following Guha's investigation of elementary forms of insurgent peasant consciousness, the school attracted the hostility of the Indian Marxist establishment for being 'idealist' (see Chakrabarty 1986, p. 364)—a sign that it had departed from economic reductionism. In a society where cultural symbols play an important role in everyday life as well as in political mobilization, this was a fruitful departure, necessary for comprehending phenomena such as charismatic leadership and communal conflict (see, e.g., Amin 1984, pp. 1–55, 1995, Pandey 1982, pp. 143–97, 1983, pp. 60–129, 1984, pp. 231–70, 1990, Hardiman 1982, pp. 198–231, 1984, pp.196–230, 1987). It has since expanded 'beyond the discipline of history' to engage 'with more contemporary problems and theoretical formations' (*Subaltern Studies* 1998, Vol. 9, Preface). These include the politics of identity and literary deconstruction. The shift in the 1990s was marked by the interest shown in the project by literature scholars and critics such as Edward Said and Gayatri Chakravorty Spivak. The analysis of discourse has since become a major preoccupation for subalternist research.

It would be wrong to adduce a uniformity of vision in the output of *Subaltern Studies* over the years. The explicit, if unorthodox Marxian bent of its 'history from below' phase has been superseded by a recent preoccupation with community, rural innocence, and cultural authenticity. The epistemological approval of the Leninist 'outsider' as the bearer of 'higher' revolutionary consciousness (Chaudhury 1987, pp. 236–51) sits in unresolved tension with the oft-expressed critique of elitism and statism in historiography (Pandey 1994, pp. 188–221), and the belief in the immanence of culturally mediated forms of universal community (Chatterjee 1989, pp. 169–209). A thematic that persists, however, is the opposition of two 'domains'—that of the elite, meaning the colonial state and its allies, and their forms of politics, knowledge, and power, vs. the subaltern. The latter has been variously interpreted as the peasantry, community, locality, and traditional domesticity and distinguished by its resistance to colonization. The difficulty caused by the problem of mass complicity has been dealt with by valorizing the recalcitrance of 'fragments.'

The subalternist stance of giving voice to the repressed elements of South Asian history has engendered valuable research. A prominent example is Shahid Amin's meticulous and thought-provoking investigation of the prolonged aftermath of the (in)famous Chauri Chaura riot of 1922, which resulted in the death of 22 policemen, the suspension of the first noncooperation movement, and the subsequent punishment by hanging of 19 accused rioters (Amin 1987, pp. 166–202, 1995). This unravelling of 'an event which all Indians, when commemorating the nation, are obliged to remember—only in order to forget,' (Amin 1995, p. 1) relentlessly juxtaposes event to

nationalist metaphor and existential reality to ideological representation. It will remain an outstanding text in the subalternist corpus.

Pandey's intricate account of cow-protection movements in eastern India in the late nineteenth century exposes the interplay of symbolism, class interest, and public space. This path-breaking essay in the prehistory of communal politics (Pandey 1983, pp. 60–129), along with his writings on the 'construction' of communalism in colonial India (Pandey 1989, pp. 132–68, 1990) has contributed significantly to a raging historical debate. Guha's (1987, pp. 135–65) 'Chandra's Death' skilfully uses a legal narrative from mid-nineteenth century Bengal to analyze the workings of patriarchal culture and indigenous justice with great sensitivity to the existential predicament of 'low-caste' women. In a brilliant passage, Guha qualifies a description of conventional systems of asylum thus:

> this other dominance did not rely on the ideology of Brahmanical Hinduism or the caste system for its articulation. It knew how to bend the relatively liberal ideas of Vaishnavism and its loose institutional structure for its own ends, demonstrating thereby that for each element in a religion which responds to the sigh of the oppressed there is another to act as an opiate. (Guha 1987, p. 159)

3. *Theoretical Tensions*

Subaltern Studies' concern with issues of ideological hegemony elicits a questioning of the school's own theoretical tensions. The juxtaposition of statist versus subalternist history; or the tyrannical march of Western-inspired universals versus the resistance/occlusion of heroic subalterns expresses a view of a society divided into discrete social zones—with a concomitant oversight regarding the osmosis between these 'domains.' The idea that contemporary history encompasses a grand struggle between the narratives of Capital and Community, that the latter is the truly subversive element in modern society ('community, which ideally should have been banished from the kingdom of capital, continues to lead a subterranean, potentially subversive life within it because it refuses to go away'—P. Chatterjee 1997, p. 236), raises the question of why 'class' has been demoted from the estate of subalternity, even though it too refuses to go away. In an era wherein the assertion of community is rapidly transiting from the realm of peasant insurgency to that of mass-produced identity, might not 'community' actually function as the necessary metaphysic of Capital rather than its unassimilable Other?

The tensions extend beyond theory to that of discursive choice and indeed, silence. The political success of the movement for Pakistan, for example, the transformation in this case, of 'communalism' into 'nationalism,' has not been investigated despite Guha's pointers 1989, pp. 210–309, Guha 1992, pp. 69–120), and despite the urgent need for reflection on the Indian communists' transitory but significant support in the 1940s for the two-nation theory and Partition. Nor has the subsequent history of Pakistan and the emergence of Bangladesh been addressed—a rich field for those interested in the ever-shifting paradigms of nationhood and identity in South Asia. Has subalternist thinking confined itself ideologically within the fragmentary remainder of 1947? Similarly, the category of labor and the history of the working class is absent from the main corpus of research after Chakrabarty's (1983, pp. 259–310, 1984, pp. 116–52, 1989) publications on the jute-mill workers of Calcutta.

Nor would it appear from this fleeting passage of workers through the subalternist corpus, that the national movement and nationalism had any impact on them. Despite insightful commentaries by P. Chatterjee (1984, pp. 153–95, 1986) on Gandhian ideology and Amin's work on Chauri Chaura, the widespread popular appeal of Gandhi and his *ahimsa* remains an underexamined theme. The scholar who tires of negativity and is looking for answers on the role of charisma might find emotional sustenance as well as food for thought in an essay by Dennis Dalton (1970). Dalton is neither a historian nor a subalternist.

This lacuna coexists with a reluctance to tackle the history of the communist movement, within India or internationally. Given its founder's abiding interest in the failure of radical historiography to produce a 'principled and comprehensive (as against eclectic and fragmentary) critique of the indigenous bourgeoisie's universalist pretensions' (Guha 1989, p 307), it would have been intellectually appropriate for him to address the history and historiographical practice of the movement to which he owed theoretical inspiration. The subalternist antipathy towards what is perceived as the representational pretension of the Gandhian Congress, its habit of translating a constricted, bourgeois aspiration into a nationalist universal (see Guha 1992, pp. 69–120), elicits a query about the political practice of the 'true representatives' of the workers and peasants.

Guha can hardly be faulted for polemical shyness—and this makes Subaltern Studies' sustained avoidance of 'principled and comprehensive' research on the fractious and tragic meanderings of Indian communism quite remarkable. The observations on Guha's own political trajectory (Amin and Bhadra 1994, pp. 222–5) which refer to his disillusionment with the Soviet invasion of Hungary in 1956, and to his association with Maoist students in Delhi University in 1970–1, throw no light on the nature and content of his theoretical transformation. These are not mere matters of biographical detail. They are linked to vital historical questions on Bolshevism and its impact on anticolonial struggles in the twentieth century, and not just in India. The fact that *Subaltern Studies* has

carried the occasional essay on non-Indian societies, implies that the project of exposing elitist bias and ideological camouflage has been (notionally at least) thrown open to cross-national debate. Yet its sole addressal of Leninism in 17 years takes the form of a theoretical apology (Chaudhury 1987, pp. 236–51) without raising the matter of political hegemonism and subalternity in the USSR, initiatives 'from below' in the Russian Revolution, or the impact of Stalinism on the international communist movement.

4. A Radical Historiography?

As a discursive field Subaltern Studies has produced provocative research on the history of colonial India and, of late, into more recent developments. It has been a forum for fresh scholarship on a variety of themes, ranging from 'low' caste and 'tribal' peasant insurgency, middle-class ideologies of nationalism, prison life, and disciplinary structures under colonialism, to the politics of liquor, the significance of myth, and interpretations of 'bondage.' It has also contributed important theoretical reflections on questions of nationalism, colonial science, caste, gender, and identity. This includes an evaluation of the historiographical antecedents of *Hinduttva* or majoritarian nationalism (Chatterjee 1994, pp. 1–49), a critique of colonial penology (Arnold 1994, pp. 148–87), a commentary on recent developments in Indian feminism (Tharu and Niranjana 1996, pp. 232–60), research on concubinage and female domestic slavery (Chatterjee 1999, pp. 49–97) and an epistemological analysis of colonial ethnography (Ghosh 1999, pp. 8–48).

The inquisitive scholar will also find it worthwhile to read a critique of the school written by an erstwhile member of the collective. Sarkar's (1998) essay on 'The Decline of the Subaltern in Subaltern Studies' challenges what he considers to be its valorization of the indigenous, its 'enshrinement of sentimentality', and the shift in its polemical target from capitalist and colonial exploitation to Enlightenment rationality. A caveat might also be entered on the status of any scholarly claim to 'represent' the voice, interest or agency of a preferred subject—the historian's discipline may indeed never be free of bias, but surely it must be as committed to the ideal of truth-as-the-whole, and balance, as to polemic. Be that as it may, Subaltern Studies has raised the level of debate in Indian historiography—the corpus may be critiqued, but certainly not ignored. It has had an impact on the orientation of many scholars, within and outside the discipline of history, and beyond the frontiers of India. Whether it will retain its original radical impetus by engaging boldly with questions posed by its own practice and the rapidly changing social and political environment in the post-Soviet global order remains to be seen.

Bibliography

Amin S 1984 Gandhi as Mahatma: Gorakhpur District, Eastern UP, 1921–2. In: Guha R (ed.) *Subaltern Studies 3*. Oxford University Press, Delhi, India

Amin S 1987 Approver's testimony, judicial discourse: The case of Chauri-Chaura. In: Guha R (ed.) *Subaltern Studies 5*. Oxford University Press, Delhi, India

Amin S, Bhadra G 1994 Ranajit Guha: A bibliographical sketch. In: Hardiman D, Arnold D (eds.) *Subaltern Studies 8, Essays in Honour of Ranajit Guha*. Oxford University Press, Delhi, India

Amin S 1995 *Event, Metaphor, Memory: Chauri Chaura, 1922–1992*. Oxford University Press, Oxford, UK

Arnold D 1994 The colonial prison: Power, knowledge and penology in nineteenth century India. In: Hardman D, Arnold D (eds.) *Subaltern Studies 8, Essays in Honour of Ranajit Guha*. Oxford University Press, Delhi, India

Chakrabarty D 1983 Conditions for knowledge of working class conditions: Employers, government and the jute workers of Calcutta, 1890–1940. In: Guha R (ed.) *Subaltern Studies 2*. Oxford University Press, Delhi, India

Chakrabarty D 1984 Trade unions in a hierarchical culture: The jute workers of Calcutta, 1920–1950. In: Guha R (ed.) *Subaltern Studies 3*. Oxford University Press, Delhi, India

Chakrabarty D 1986 Discussion: Invitation to a dialogue. In: Guha R (ed.) *Subaltern Studies 4*. Oxford University Press, Delhi, India

Chakrabarty D 1989 *Rethinking Working Class History: Bengal, 1890–1940*. Oxford University Press, Oxford, UK

Chatterjee P 1984 Gandhi and the critique of civil society. In: Guha R (ed.) *Subaltern Studies 3*. Oxford University Press, Delhi, India

Chatterjee P 1986 *Nationalist Thought and the Colonial World—A Derivative Discourse*. Zed/Oxford University Press, Oxford, UK

Chatterjee P 1989 Caste and subaltern consciousness. In: Guha R (ed.) *Subaltern Studies 6*. Oxford University Press, Delhi, India

Chatterjee P 1994 Claims on the past: The genealogy of modern historiography in Bengal. In: Hardman D, Arnold D (eds.) *Subaltern Studies 8, Essays in Honour of Ranajit Guha*. Oxford University Press, Delhi, India

Chatterjee P 1997 *The Nation and Its Fragments: Colonial and Post Colonial Histories*. Oxford University Press, Oxford, UK

Chatterjee I 1999 Colonising subalternity: Slaves, concubines and social orphans in early colonial India. In: Bhadra G, Prakash G, Tharu S (eds.) *Subaltern Studies 10*. Oxford University Press, Delhi, India

Chaudhury A K 1987 In search of a subaltern Lenin. In: Guha R (ed.) *Subaltern Studies 5*. Oxford University Press, Delhi, India

Dalton D 1970 Gandhi during partition. In: Philips C H, Wainright M D (eds.) *The Partition of India: Policies and Perspectives 1935–1947*. Allen and Unwin, London, pp. 222–44

Ghosh K 1999 A market for aboriginality: Primativism and race classification in the indentured labour market of colonial India. In: Bhadra G, Prakash G, Tharu S (eds.) *Subaltern Studies 10*. Oxford University Press, Delhi, India

Guha R 1983 *Elementary Aspects of Peasant Insurgency in Colonial India*. Oxford University Press, Oxford, UK

Guha R 1987 Chandra's death. In: Guha R (ed.) *Subaltern Studies 5*. Oxford University Press, Delhi, India

Guha R 1989 Dominance without hegemony and its historiography. In: Guha R (ed.) *Subaltern Studies 6*. Oxford University Press, Delhi, India
Guha R 1992 Discipline and mobilise. In: Chatterjee P, Pandey G (eds.) *Subaltern Studies 7*. Oxford University Press, Delhi, India
Hardiman D 1982 The Indian faction: A political theory examined. In: Guha R (ed.) *Subaltern Studies 1*. Oxford University Press, Delhi, India
Hardiman D 1984 Adivasi assertion in South Gujarat: The Devi movement of 1922–3. In: Guha R (ed.) *Subaltern Studies 3*. Oxford University Press, Delhi, India
Hardiman D 1987 *The Coming of the Devi: Adivasi Assertion in Western India*. Oxford University Press, Oxford, UK
Pandey G 1982 Peasant revolt and Indian nationalism: The peasant movement in Awadh 1919–22. In: Guha R (ed.) *Subaltern Studies 1*. Oxford University Press, Delhi, India
Pandey G 1983 Rallying round the cow: Sectarian strife in the Bhojpuri region, 1888–1917. In: Guha R (ed.) *Subaltern Studies 2*. Oxford University Press, Delhi, India
Pandey G 1984 Encounters and calamities: The history of a North India Qasbah in the nineteenth century. In: Guha R (ed.) *Subaltern Studies 3*. Oxford University Press, Delhi, India
Pandey G 1989 The colonial construction of 'communalism': British writings on Benares in the nineteenth century. In: Guha R (ed.) *Subaltern Studies 6*. Oxford University Press, Delhi, India
Pandey G 1990 *The Construction of Communalism in Colonial North India*. Oxford University Press, Oxford, UK
Pandey G 1994 The prose of otherness. In: Hardiman D, Arnold D (eds.) *Subaltern Studies 8, Essays in Honour of Ranajit Guha*. Oxford University Press, Delhi, India
Prakash G 1990 *Genealogies of Labour Servitude in Colonial India*. Cambridge University Press, Cambridge, UK
Sarkar S 1998 *Writing Social History*. Oxford University Press, Oxford, UK
Sen A 1987 Capital, class and community. In: Guha R (ed.) *Subaltern Studies 5*. Oxford University Press, Delhi, India
Tharu S, Niranjana T 1996 Problems for a contemporary theory of gender. In: Amin S, Chakrabarty D (eds.) *Subaltern Studies 9*. Oxford University Press, Delhi, India
Subaltern Studies 1982–99 10 Vols, Oxford University Press, Delhi, India

D. Simeon

Subculture, Sociology of

More than the notion of culture, that of subculture poses more problems for sociologic analysis than it is able to solve. Just as specialists have enumerated several dozens of definitions of the word culture (Kroeber and Kluckhohn 1952, Shils 1961), it would be easy to do likewise for the notion of subculture.

1. *The Concept of Subculture*

When the concept of culture is taken in a normative manner, as being the set of works of the mind, art, philosophy, religion, and science, thus entailing a high level of work and development, the notion of subculture would seem to subtend the degraded or 'vulgar' forms of that culture. Within the hierarchical framework of culture, the subcultures may be understood as the popular cultures, often marginal, and especially the mass culture engendered by the cultural industries. On the one hand, a whole conservative tradition (Bloom 1987) and on the other, a critical one, the Frankfurt School (Adorno 1980), has analyzed this mass subculture as being a mixture of propaganda, commerce and alienation, and as being a degradation of both the aristocratic and rationalist ideal of a great culture, as widespread as civilization itself. In short, this view holds subculture to be an inferior culture.

While much thinking on mass culture has considered subculture as being centered on leisure activities, entertainment and cultural consumption, the limitations of such a view were soon revealed. First (Morin 1962), it was demonstrated that mass culture could play an integrating and communicative role at the same time as the traditional social cultures and affiliations were declining, such as those involving classes and religions. Moreover, this form of culture does not remain superficial when it induces strong cultural transformations, and when it is partially associated with social movements such as youth culture and the production and consumption of music. For many years thought of as a soft form of propaganda, the mass subculture conveyed by the media plays a more ambiguous and probably more democratic role than one might believe (Lazarfeld et al. 1968). Finally, the notion of mass cultures comprises a set of specific forms of cultural consumption, ranging from art to leisure activities, including more private activities which have become the object of public cultural policies playing a leading role in the definition of ways of life and social identities. This mass culture might even be a degradation of 'real' culture; indeed, while some of its productions are authentic cultural works, the question arises as to whether cinema, sport and television are an integral part of culture, or whether they are only subcultures. In fact, the issue is hardly of scientific interest and has perhaps come about only through the ethnocentricity of culturally dominant and legitimate groups. At best, the notion of subculture in this view is only a polemic device, the danger being that we may become blind to the diversity of mass culture and to the value of studying its modes of production and forms of consumption.

2. *The Hierarchy of Cultures*

When culture is considered from an anthropologic viewpoint as a set of values, representations, and ways of being and acting, the problem of the hierarchy of cultures is resolved, since each group is logically defined by a culture (Lévi-Strauss 1952). In this case, however, it is only the nature of the difficulty which changes, because the major issue is that concerning the

demarcation between cultures held to be coherent and relatively closed entities and consequently defining where the frontier with subcultures lies. In fact, there is a risk of associating a subculture to just about every identifiable activity/practice, and therefore of diluting the concept into an ongoing continuum.

It is clear that any group, social category or set of activities/practices may be more or less clearly defined by a subculture. It is valid to talk of youth subcultures, peasant subcultures, class subcultures, the subculture of chic or delinquent city quarters, and even professional subcultures. Numerous authors have investigated these themes and there is now a whole anthropology of modernity. However, the notion of subculture took on a particular meaning with studies on the culture of poverty. R. Hoggart (1957) highlighted two essential dimensions to working-class culture: first, the way a community culture can be based on social distance and how the working class is locked into its own values; second, the manner in which a culture serves to ratify the social domination that the working class undergoes, with the result that one's class destiny is widely accepted. In this view, the subculture is seen as a way of maintaining social order. In describing the culture of poverty, O. Lewis (1963) treads similar ground, but with more clarity. For him, poverty engenders a specific culture which, more than just a form of adaptation to economic constraints, is rather a normative interiorization of poverty, thus explaining the way it is reproduced and why its actors are unable to escape from their situation in life. Thus, poverty may be explained by its own subculture. This idea has been heavily criticized owing to its conservative undertones, suggesting that the poor are victims, more of their culture, than of the social conditions that are imposed on them.

Beyond this criticism, the notion of subculture is problematic, since it is difficult to know where the frontiers between cultures lie in a porous society where there is interpenetration, imitation and opposition between groups and cultures (Bourdieu 1979). Moreover, this view tends to reify cultures and subcultures, to give them more homogeneity than they really have, and to explain the conduct of actors only by their cultural identification.

3. *The Links Between Culture and Action*

To these first two difficulties is to be added a third, i.e., the links between culture and action. While it is clear that social action is normatively and culturally orientated and that it proceeds in 'patterns' of action, it is problematic to explain action as the simple application of a cultural program, as postulated by the most radical thinkers on culturalism (Benedict 1935, Kardiner 1939, Linton 1968). More precisely, by explaining social conduct in terms of cultures and subcultures, there is a great risk of getting into a tautologic loop. For example, deviant behavior could be explained by the existence of a deviant subculture, authoritarian behavior by the existence of authoritarian subcultures, and people might play soccer because it likewise has its own subculture. Even if such an explanation were to be admitted, the main scientific issue remaining would be the ability to explain the development and logic of subcultures. Regarding deviance, for example, there are several competing theories of the subculture, because they explain the development of the subcultures differently.

It is clear that delinquent behavior supposes specific norms, a particular form of socialization, a conception of social life and its values, if only to neutralize the interiorization of conformist norms (Matza 1964). Three schools of thought may be distinguished here. The first (Thrasher 1927, Jankowski 1991) considers the subculture of gangs of youngsters as being a response to a phenomenon of social disorganization and anomie. Enhancement of the bonds of solidarity in the gang, the logic of confronting other gangs and the police, and the obsession with honor may all be seen as ways of constructing a subculture, guaranteeing limited social integration for individuals who cannot play a full role in social life. The second school of thought regarding delinquent subculture derives from the Mertonian model of anomie (Merton 1965). Here, subcultures are seen as a way to reduce the structural tension brought about by the juxtaposition of profound social inequality with the democratic culture of achievement and its concomitant realization of the individual in economic success. The various deviant cultures may therefore be considered to enhance either delinquency as a response to frustrated conformism (Cohen 1955), or social retreat, withdrawal from the game and holing up in a primary group (Cloward and Olhin 1960). Finally, other more marginal theories hold deviant subculture to be the product of confrontation between the working class and the conformism of the middle classes. Consequently, the actors of social control are seen as engendering deviance and the labeling of individuals by way of their own professional subcultures. Hence, social conflicts may be considered to be conflicts between subcultures.

To be clear and not fall foul of tautology, subculture may be explained as the product of social relations and of the strategies that the actors implement. In general, subcultures result from the meeting of wider cultures and particular social situations (Dubet 1987). Therefore, subcultures could be considered as the way in which actors interpret whole cultural settings in the light of the situations and contexts in which they find themselves. In these terms, subcultures are 'constructed' by actors and are not imposed on them as principal determinants of their action. Moreover, it is difficult to free the notion of subculture from the value judgments that define it. Indeed, it is very often used in a naive ideological way by researchers oscillating

between populist enchantment and aristocratic loftiness.

See also: Counterculture; Critical Theory: Frankfurt School; Culture, Sociology of; Popular Culture; Values, Sociology of; Youth Gangs

Bibliography

Adorno T W 1980 *Minima Moralia. Reflexions sur la Vie Mutilée*. Payot, Paris
Benedict R 1935 *Patterns of Culture*. Routledge, London
Bloom A 1987 *L'âme Désarmée, Essai sur la Culture Générale*. Julliard, Paris
Bourdieu P 1979 *La Distinction*. de Minuit, Paris
Cloward R A, Ohlin L E 1960 *Delinquency and Opportunity*. The Free Press, New York
Cohen A K 1955 *Delinquent Boys. The Culture of the Gang*. The Free Press, New York
Dubet F 1987 *La Galère. Jeunes en Survie*. Fayard, Paris
Hoggart R 1957 *The Uses of Literacy*. Chatto and Windus, London
Jankowski M S 1991 *Islands in the Streets: Gangs and American Urban Society*. University of California Press, Berkeley, CA
Kardiner A 1939 *The Individual and his Society*. Columbia University Press, New York
Kroeber A L, Kluckhohn C 1952 *Culture: A Critical Review of Concepts and Definitions*. Vintage Books, New York
Lazarsfeld P, Berelson B, Gaudet H 1968 *The People's Choice*. Columbia University Press, New York
Lévi-Strauss C 1952 *Race et Histoire*. UNESCO, Paris
Lewis O 1963 *Les Enfants des Sanchez*. Gallimard, Paris
Linton R 1936 *The Study of Man*. Appleton Century, New York
Matza D 1964 *Delinquency and Drift*. Wiley, New York
Merton R K 1965 *'Structure Sociale, Anomie et Déviance', Eléments de Theorie et de Méthode Sociologique*. Plon, Paris
Morin E 1962 *L'esprit du temps*. Grasset, Paris
Shils E 1961 Mass society and its culture. In: *Culture for the Millions*. D. Van Nostrand, Princeton
Thrasher F 1927 *The Gang*. University of Chicago Press, Chicago, 1963

F. Dubet

Subjective Probability Judgments

1. Introduction

Life is rife with uncertainty. Do our enemies have biological weapons? Will it rain next Sunday? How much oil can be found by drilling in this spot? Important decisions in areas as diverse as medicine, business, law, public policy, and one's private life, must be made under uncertainty, hence hinge on odds and probabilities. 'Probability,' in the famous words of Bishop Butler (1736), indeed 'is the very guide to life.' Yet the theories of mathematical probability or of statistical inference cannot deliver the probabilities sought. So we have to make do with subjective probabilities—the probabilities that people generate in their own minds to express their uncertainty about the possibility of the occurrence of various events or outcomes. How do people generate these subjective probabilities? Since the 1970s, this question has attracted much research attention from cognitive psychologists. The present article will review this research.

Although statistics and probability theory can seldom replace intuitive judgments, a favored paradigm in the study of subjective probabilities has been to use the norms and dictates these theories provide as yardsticks against which people's intuitions can be compared, thereby illuminating the cognitive heuristics people use, and the biases to which they are subject. Appropriately, the dominant school in this line of research is called *the heuristics and biases approach* (Tversky and Kahneman 1974, Kahneman et al. 1982). At its core are two heuristics for judging probabilities—the representativeness heuristic, and the availability heuristic.

2. Representativeness

This is a heuristic for judging the probability that a certain target case, presented by some individuating information, belongs to a certain category (or comes from a certain source, or possesses a certain feature). This is a common type of probability judgment: a doctor judges the probability that a particular patient has pneumonia, a juror judges the probability that a particular defendant committed murder, or an employer judges the probability that a particular job candidate would be a good employee. Its essence is to base the probability judgment on the extent to which the target case's relevant particulars (e.g., the patient's symptoms, the defendant's motives, or the job candidate's background) match, resemble, or represent the attributes which the judge considers typical of, or central to, the category (Kahneman and Tversky 1972).

Similarity judgments, however, obey different rules than probability. For example, a person could well be more *like* a typical New Yorker than *like* a typical American—but a person can never be more *likely* to be a New Yorker than an American. This difference between resemblance and probability can bias judgments of the latter when they are based on the former. Consider, for instance, the following example.

Example 1: After graduating from college, Sarah took a trip around the world, writing frequent letters home. One contained the following description: 'It is so clean here, you could eat off the streets. Snow-capped mountains frame picturesque log cabins, with geraniums in every window.' Where do you think this letter was written?

Respondents given this question sensibly thought it was more likely that the letter was written from Switzerland than from Brazil. But they also thought it was more likely that the letter was written from Switzerland than from ... Europe! When asked to select the most likely of several possibilities (which included Switzerland, Brazil, and Europe, among others), they ranked Switzerland—not Europe—first (Bar-Hillel and Neter 1993).

Probability theory does not provide an answer to the question: 'What is the probability that this letter was written from Switzerland?' But once this probability is assessed, it sets constraints on certain other probabilities. Thus, whatever the probability that the letter was written from Switzerland, it cannot possibly be higher than the probability that it was written from Europe. This is the *extension principle*, and people realize its validity. Those that rely on it when performing the experimental task do not rank Switzerland before Europe. Most, however, just compare the description in the letter with their idea of the places under consideration. The match between the description and Switzerland is the best—better even than the match between it and Europe. So, in accordance with the representativeness heuristic, Switzerland is ranked most likely.

Results such as this point out both what people often do (namely, base probability judgments on judgments of representativeness), and what they often fail to do (namely, rely on formal rules of probability). One factor, which plays an important role in formal probability calculations but is relatively neglected in subjective probability reasoning, is size—either category size, or sample size. This is shown in the following example.

Example 2: A certain town is served by two cab companies, the Blue Cab and the Green Cab, named after the color of cab they operate. The Blue Cab Company has many more cabs than the Green Cab Company. One dark and foggy night, a pedestrian was knocked down by a hit-and-run cab. She later testified that it was green. Her accuracy rate in distinguishing green from blue cabs under the conditions that existed that night was found to about 85 percent. What is the probability that the errant cab was a Green Cab?

This problem, and many variants thereof, has been studied in scores of studies (e.g., Tversky and Kahneman 1982a, Bar-Hillel 1980, 1983). Formatted like a high school math problem, it invites explicit numerical calculation. The modal response is to assign the witness' testimony a probability that matches her accuracy, 85 percent. The same happens when the witness is said to have identified the cab as belonging to the larger company, Blue Cab. Indeed, the relevance of the relative size of the cab companies to the question eludes most people. After all, they argue, the pedestrian was run over by a single cab, not by the entire fleet. Nonetheless, when the witness is unable to identify the color at all, most people properly realize that the errant cab is probably from the larger company, simply because *it is* the larger company. Normatively, the witness' testimony should be integrated with this base-rate consideration, rather than replacing it.

The inclination to neglect category size is known as *the base-rate fallacy*. Since Example 2. does not involve judgment by representativeness, it shows that base-rate neglect is not merely a side effect of this heuristic, as it was in Example 1. Example 3. shows that size considerations are neglected not only when considering the source population, but also when considering the sample.

Example 3.: A game of squash can be played either to 9 or to 15 points. If you think you are a better player than your opponent, which game gives you a higher chance of winning, the short game, or the longer game? Suppose, instead, that you are the weaker player, which game gives you a higher chance now?

If you are inclined to favor the same length in either case, consider this theorem from probability theory: The larger the sample of rounds (i.e., 15 rather than 9), the likelier it is to yield the expected outcome, namely victory to the stronger player. So if you believe you are the stronger player, you should prefer the longer game, but if you believe you are weaker, your chances are better with a shorter game. Intuitively, though, winning a game seems equally representative of player strength whether it is a shorter or longer game (Kahneman and Tversky 1982b).

This is not to deny that people can—spontaneously or by mild prompting—appreciate that larger samples are better for estimating population parameters than smaller samples. But they erroneously think that the advantage lies in being *relatively* larger (in terms of the ratio of sample size to population size) rather than *absolutely* larger. Many, for example, indicate equal confidence in the accuracy of a survey of 1 in every 1,000 people in predicting election outcomes in a city of a million voters as in a survey of 1 in every 1,000 people in a town of 50,000—even though the former survey samples 20 times as many people (Bar-Hillel 1979).

Example 4., which follows, demonstrates another judgmental bias, which is related to size neglect and is known as *failure to regress*.

Example 4.: When Julie was 4, she was rather tall for her age—in the 80th percentile (i.e., taller than about 80 percent of her peers). She is now 10 years old. What would you guess is her present height percentile?

If you think that Julie may have moved up, or she may have moved down, but she probably is still taller than most of her peers, you are quite right. However, if you also think that Julie is as likely to be in a higher percentile as in a lower percentile, and that by and large she is probably still most likely to be approximately in the 80th percentile, then you are sharing a common, but erroneous, disregard for the phenomenon known as regression to the mean. This is the fact that the mean of a second sample from a population

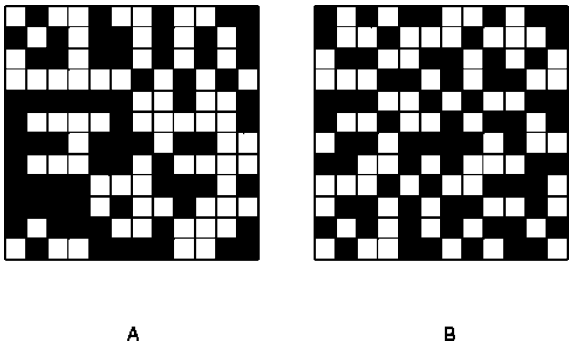

Figure 1
Which matrix is more 'randomly' mixed?

can always be expected to be closer to the population mean than was the first sample's (unless the first sample's mean happens to have coincided with the population mean to begin with). In the present example, this follows from the fact that a larger proportion of the population of children were shorter than Julie (80 percent) than were taller than her (20 percent).

Failure to recognize regression to the mean can result in inventing causal explanations for what may be mere statistical artifacts (Kahneman and Tversky 1973), such as for the observation that very bright women tend to have somewhat less bright husbands.

The following example demonstrates some difficulties that people have with the concept of randomness.

Example 5: Consider the two 12×12 matrices in Fig. 1. Each consists of 50 percent black cells and 50 percent white cells. In one, the colors are mixed in a manner typical of truly random mixing. In the other, the cells are not mixed randomly at all. Which do you think is the 'random' matrix?

Most people perceive B as more 'random' (Falk 1975). A seems to them too 'clumpy.' But in fact, about half the pairs of adjacent cells in B are of the same color—exactly as much clumpiness as expected in a randomly mixed matrix. On the other hand, in A, almost two-thirds of neighboring cells are of different colors, which is more alternation than could be expected statistically. But people's intuitions of randomness (see, e.g., Bar-Hillel and Wagenaar 1991) are characterized by 'local representativeness'—the erroneous expectation that the short run should exhibit the properties which *the law of large numbers* promises only in the long run. Thus, local representativeness is like a *law of small numbers* (Tversky and Kahneman 1971).

3. *Availability*

If the representativeness heuristic tends to overlook category size, the availability heuristic is used primarily for judging category size—or rather, relative size. Category size relates to probability inasmuch as probability is sometimes taken to be the proportion, in the population as a whole, of instances that belong to some specific category, or possess some specific feature. A statistician who wants to estimate relative frequencies draws a random sample. The cognitive counterpart of the statistician's procedure is to 'sample in the mind'—namely, to bring to mind, either by active search or spontaneously, examples of the target population. Sampling in the mind, however, is hardly random. It is subject to many effects that determine the ease with which examples come to mind (Tversky and Kahneman 1973). Examples of such effects are salience, recency, imaginability, and—fortunately—even actual frequency.

Effects that determine the ease of calling to mind are known as *availability effects*. The availability heuristic is the judgmental procedure of reliance on mental sampling, and is demonstrated in the following example.

Example 6: Subjects were read a list of 39 names of celebrities. In List 1, the 19 women were rather more famous than the 20 men, and in List 2 it was the 19 men who were more famous than the 20 women. Each person heard a single list. They were then asked whether they had heard more female names or more male names. As predicted, when the men [women] were more famous, male [female] names were more frequently recalled. They were also judged, by independent respondents, to be more frequent.

It turns out to be unnecessary actually to bring examples to mind, as it is enough just to estimate how many could be brought to mind. People seem to have a pretty good notion of availability (Tversky and Kahneman 1973), and use it to estimate probabilities.

In Example 7., the availability heuristic leads to a violation of the aforementioned extension principle, mentioned before.

Example 7.: Imagine the following scenarios, and estimate their probability. (a) A massive flood somewhere in America, in which more than 1,000 people die. (b) An earthquake in California, causing massive flooding in which more than 1,000 people die.

Respondents estimated the first event to be less likely than the second (Tversky and Kahneman 1983). As discussed earlier, the extension principle logically precludes this possibility: a flood with unspecified causes in a relatively unspecified location is necessarily more likely than one with a more specific cause and location. An earthquake in California, however, is a readily imaginable event which greatly increases the availability—hence the subjective probability—of the flood scenario.

A particularly prominent result from the availability heuristic is *the unpacking effect*, shown in Example 8.

Example 8.: A well-known 22-year-old Hollywood actress was admitted to a hospital emergency room with pains in the lower right abdomen, which had

lasted over 12 hours. What could she be suffering from?

A group of general practitioners from the Stanford area were given this question. Half were asked to assign a probability to the following possibilities: (a) Indigestion (b) Extra-uterine pregnancy (c) Other. The list given to the other half was: (a) Indigestion (b) Extra-uterine pregnancy (c) Acute appendicitis. (d) Inflamed kidney. (e) Cervical inflammation (f) Other. The first group assigned a probability of 50 percent to the 'Other' category. The second group assigned a probability of 69 percent to categories (c) to (f), although of course these are but a more detailed breakdown ('unpacking') of 'Other' (Redelmeier et al. 1995). A similar effect is found even when it is hard to argue that the unpacking itself provides additional information. When the category of 'homicide' is broken down into 'homicide by a stranger' and 'homicide by an acquaintance,' the latter two are given a higher total probability as causes of death than the unbroken category (Tversky and Koehler 1994). Here, apparently, the different descriptions of the categories merely call attention to different aspects of the categories. Either by affecting memory or by affecting salience, enhancing the availability of subcategories through more detailed descriptions enhances their perceived probability.

4. *Subjective Uncertainty*

When judging probability, people can locate the source of the uncertainty either in the external world or in their own imperfect knowledge (Kahneman and Tversky 1982). The latter is what Bayesian theory means by the term *subjective* probability (whereas in this article, recall, the adjective refers to the mode of assessment, rather than to the locus of the uncertainty). When assessing their own uncertainty, people tend to underestimate it. The two major manifestations of this tendency are called *overconfidence*, and *hindsight bias*.

Overconfidence concerns the fact that people overestimate how much they actually know: when they are p percent sure that they have answered a question correctly or predicted correctly, they are in fact right on average less than p percent of the time (e.g., Lichtenstein et al. 1982, Keren 1991). Hindsight bias concerns the fact that people overestimate how much they would have known had they not possessed the correct answer or prediction: events which are given an average probability of p percent before they are known to have occurred, are given, in hindsight, probabilities higher than p percent (Fischhoff 1975)—a phenomenon sometimes known as 'I knew it all along.'

Both these biases can result from biased availability. After all, what we know is almost by definition more available to us than what we do not know. Regarding hindsight, it is enough to note that once something (such as the final outcome) has been brought to mind, it is nearly impossible to ignore (try to 'not think of a white elephant').

Regarding overconfidence, it has been shown that it can be dampened by instructing people to 'think of reasons you might be wrong' prior to asking them to assess their own confidence (Koriat et al. 1980)—a direct manipulation to increase the availiability of the complementary event, the one competing for a share of the total 100 percent probability. In addition, overconfidence turns to underconfidence in very hard questions (e.g., May 1986). Perhaps in answering easy questions, one's best guess is mentally prefaced by: 'I'm not sure I know, but I think that …' which focuses on the available. In answering hard questions, on the other hand, one's best guess might be mentally prefaced by: 'I don't think I know, but I guess that …' which focuses on the unavailable.

5. *Conclusion*

The survey of heuristics and biases involved in people's judgments under uncertainty shows that subjective probabilities (in the sense of judged probabilities) are not just imperfect or inaccurate versions of objective probabilities, but rather are governed by cognitive principles of their own. These may often lead to good judgments—but in many circumstances entail violations of normative dictates, and hence systematic departures from more rational judgments. This view stands in contrast to earlier views, which saw 'Man as an intuitive statistician,' (Peterson and Beach 1967). It also contrasts with recent views (e.g., Cosmides and Tooby 1996, Gigerenzer 1998), which challenge the very appropriateness of a research program that asks people to make probability judgments. This school of thought insists that we evolved to judge frequencies, not probabilities, so we are not to be faulted for the biases in our subjective probabilities. But as long as people are called upon, as in modern life they so often are, to make probability judgments, awareness of the phenomena this article described would benefit them greatly.

See also: Bayesian Theory: History of Applications; Bounded and Costly Rationality; Decision Making: Nonrational Theories; Decision Theory: Bayesian; Heuristics for Decision and Choice

Bibliography

Bar-Hillel M 1979 The role of sample size in sample evaluation. *Organizational Behavior and Human Performance* **24**: 245–57

Bar-Hillel M 1980 The base-rate fallacy in probability judgments. *Acta Psychologica* **44**: 211–33

Bar-Hillel M, Neter E 1993 How alike is it versus how likely is it: A disjunction fallacy in stereotype judgments. *Journal of Personality and Social Psychology* **65**(6): 1119–32
Bar-Hillel M, Wagenaar W A 1991 The perception of randomness. *Advances in Applied Mathematics* **12**: 428–54 [also in Keren G, Lewis C (eds) 1993 *A Handbook for Data Analysis in the Behavioral Sciences* Lawrence Erlbaum, Hillsdale, NJ, 369–93]
Butler J 1736 *The Analogy of Religion* ['Introduction' cited in *The Oxford Book of Quotations*, 3rd edn. (1979), Oxford University Press, Oxford, UK, p. 117 : 9]
Cosmides L, Tooby J 1996 Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* **58**: 1–73
Falk R 1975 The perception of randomness. Ph D dissertation [in Hebrew], The Hebrew University, Jerusalem, Israel
Fischhoff B 1975 Hindsight = foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology*: *Human Perception and Performance* **1**: 288–99
Gigerenzer G 1998 Ecological intelligence: An adaptation for frequencies. In: Dellarosa Cummins D, Allen C (eds.) *The Evolution of Mind*. Oxford University Press, New York
Kahneman D, Tvesky A 1982a Evidential impact of base rates. In: Kahneman D, Slovic P, Tvesky A (eds.) 1982 *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK
Kahneman D, Slovic P, Tversky A (eds.) 1982b *Judgment under Uncertainty*: *Heuristics and Biases*. Cambridge University Press, Cambridge, UK
Kahneman D, Tversky A 1972 Subjective probability: A judgment of representativeness. *Cognitive Psychology* **3**: 430–54
Kahneman D, Tversky A 1973 On the psychology of prediction. *Psychological Review* **80**: 237–51
Kahneman D, Tversky A 1982 On the study of statistical intuitions. *Cognition* **11**: 123–41
Keren G 1991 Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica* **77**: 217–73
Koriat A, Lichtenstein S, Fischhoff B 1980 Reasons for confidence. *Journal of Experimental Psychology*: *Human Learning and Memory* **6**: 107–18
Lichtenstein S, Fischhoff B, Phillips L 1982 Calibration of probabilities: The state of the art to 1980. In: Kahneman D, Slovic P, Tversky A (eds.) *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK, pp. 306–34
May R 1986 Inferences, subjective probability and frequency of correct answers: A cognitive approach to the overconfidence phenomenon. In: Brehmer B, Jungermann H, Lourens P, Sevoaan A (eds.) *New Directions in Research on Decision Making*. North Holland, Amsterdam, pp. 174–89
Peterson C R, Beach L R 1967 Man as an intuitive statistician. *Psychological Bulletin* **68**(1): 29–46
Redelmeier D, Koehler D J, Liberman V, Tversky A 1995 Probability judgment in medicine: Discounting unspecified alternatives. *Medical Decision Making* **15**: 227–30
Tversky A, Kahneman D 1971 The belief in the 'law of small numbers.' *Psychological Bulletin* **76**: 105–10
Tversky A, Kahneman D 1973 Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* **5**: 207–32
Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* **185**: 1124–31
Tversky A, Kahneman D 1983 Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* **90**: 293–315
Tversky A, Koehler D J 1994 Support theory: A nonextensional representation of subjective probability. *Psychological Review* **101**: 547–67

M. Bar-Hillel

Sub-Saharan Africa, Archaeology of

1. *The African Stone Age: A Hard Act to Follow*

People evolved in sub-Saharan Africa from a common ancestor with the other African apes about 5 or 7 million years ago. The earliest recognizable artefacts, which are stone tools associated with animal bones, come from a number of sites in east Africa about 2.5 million years ago. The earliest members of our own genus, *Homo*, date from a little after that time. The earliest claimed evidence for the controlled use of fire and the earliest persuasive bone artefacts come from sites in eastern and southern Africa dated to just before or just after 2 million years ago. Some time before 1 million years ago, a first 'out of Africa' expansion took archaic humans across the land bridge into Eurasia. The earliest handaxes, beautiful as well as functional, spread from Africa some time after 1 million years ago. Current opinion is that between a half and a quarter of a million years ago African early *Homo* populations evolved into our own species, *Homo sapiens*, whilst relatives in southeastern Asia and in Europe took other directions. Eventually, perhaps 100,000 years ago, a second 'out of Africa' movement, or perhaps a series of waves, arguably caused the replacement of archaic humans by modern ones, as *Homo sapiens* completed the process of global colonization. These early modern people in southern and eastern Africa left behind the earliest evidence for systematic marine food use and substantial early evidence for the manufacture of pigment, presumably for ritual use. If this sounds somewhat 'earlier than thou,' it nevertheless underlines the massive contribution of sub-Saharan Africa to the story of human evolution and the emergence of modern behavior.

John Goodwin effectively began the systematization of sub-Saharan stone age prehistory when, in 1929, along with van Riet Lowe, he published *The Stone Age Cultures of South Africa*, in which he established the terminological framework still used throughout the continent. He recognized similarities between the stone artefact assemblages in South Africa and Europe,

particularly in the earlier phases, but elected to coin a new set of terms, the Earlier, Middle, and Later Stone Ages. His vision was that these major units, broadly equivalent to the Lower, Middle, and Upper Palaeolithic of Europe, could be subdivided into geographically bounded assemblage sets that he called cultures or variants. Whereas the Earlier and Middle Stone Age assemblages closely resembled their European counterparts, the Later Stone Age (LSA) was quite different from the Upper Palaeolithic of Europe. At least one of his cultural constructs, the Wilton culture, became popular as a designation of microlithic Later Stone Age assemblages from the Cape of Good Hope to Cape Horn. After the foundation of the Pan African Congress meetings in the 1940s, which created a structure for French and English speaking archaeologists working in sub-Saharan Africa to better coordinate their efforts, the notion of 'intermediates' between Earlier and Middle and between Middle and Later Stone Ages was briefly entertained. These were to contain assemblages that illustrated the technological and typological transitions, but for reasons explained later were unsuccessful.

The most serious attempt to standardize what had become a terminological nightmare was the brave but ineffective set of recommendations of the Burg Wartenstein conference, held in Austria in 1964. This template envisaged a set of hierarchically organized terms (occurrences, phases, industries, and industrial complexes), replacing the older tripartite system of Goodwin's. It never found much favor and archaeologists seem simply to have moved on from terminological angst, choosing to use older or newer terms informally and to investigate behavior rather than debate the structure of the archaeological record. Significantly, this early obsession with stones as determining materials has been diffused by attention to other materials and other issues.

2. *Human Origins*

A question of considerable interest is that of whether the archaeological record is an accurate reflection of the distributions of early people. Fossil fragments of the genus *Australopithecus* (and *Ardipithecus*) were known only from the sedimentary lake basins of East Africa and the dolomite caverns of South Africa until the discovery of australopithecines from the deserts of Chad. The classic sites of eastern and southern Africa are depositional environments, which might suggest that australopithecines were more widespread but not preserved in more erosional landscapes without such catchments. Chad is certainly a hint of wider distributions. Stone tools, however, are less affected by poor preservational contexts and establish the significance of the lake basins of the African Rift Valley as the focal places in early human technological history.

Mary Leakey's pioneering archaeology of living floors at Olduvai Gorge in East Africa established that the hand axe containing Acheulean assemblages of the African Earlier Stone Age, which Goodwin had wanted to call the the Stellenbosch culture in the Cape, were preceded by earlier assemblages without such bifaces. With the first potassium argon dates of about 1.8 million years from the base of Bed 1 at Olduvai came the realization of the huge antiquity of stone tool making in Africa. The earliest assemblages, naturally referred to as Oldowan, had a variety of cores, flakes, and rather crudely retouched pieces. Pre-hand axe assemblages are now known to date from as early as 2.5 million years ago, at a number of sites in the stratified lake basin deposits of the East African Rift Valley. The assemblages vary, at least in part, with the variable nature of the stone raw materials available. Very little has yet emerged as to the likely precedents of these Oldowan assemblages, but presumably a long period of more informal tool use and modification precedes the localized concentrations now claimed as the earliest sites. By comparison with recently recorded chimp tool uses, we might expect that the changes in behavior which led to the localization of tool making debris in specific places is what is actually reflected in the archaeological record. Initial experiments with stone tool making must substantially predate 2.5 million years, but may never be recognized.

The earliest assemblages with recognizable bifaces, characteristically both roughly teardrop shaped hand axes and blade ended cleavers, date to about 1.6 million years or so, soon after the earliest appearance of the fossil remains of what used to be African *Homo erectus*, more recently termed *Homo rudolfensis*. These are now universally referred to as Acheulean assemblages, named after the much later, though discovered earlier, French sites, in contradiction to Goodwin's wishes. Three remarkable aspects of these occurrences have continued to puzzle archaeologists. First is the observation that hand axes are extremely widespread, from the river gravels of the Somme and Thames in England, through the beach terraces of the Mediterranean to the lakeside camps of the East African Rift and on to the caves and coastal dunes of the Cape. A spatial distribution of several million square kilometers and a time span of more than a million years implies a broad functionality and a lasting usefulness, a veritable Palaeolithic Swiss army knife. Suggestions as to the uses of hand axes have included digging up underground plant foods, butchering animal carcasses, implementation as missiles, and suspension as parts of animal snares. This issue is not independent of the other two enduring characteristics of these often beautifully made tools.

Many archaeologists have noted that in many, but not all, cases hand axes appear to have been flaked beyond the apparent functional requirements. Indeed the pursuit of symmetry and regularity of outline is

undeniable in many examples, raising the question of whether the objects are functional or works of art, or both. Associated with this observation is the remarkable number of such hand axes at many, but again not all, sites. At Olorgesailie and other East African sites, and particularly at the hand axe localities along the Orange River in South Africa, there are thousands, and in some cases millions of bifacially worked hand axes. Even allowing for the fact that such sites are obvious palimpsests, there seems to be a case for questioning the production strategy of artefacts that often appear to the casual observer to be still relatively usable. Why did people put so much energy into the finishing of such pieces, and why did they make so many? A recent suggestion has been that both of these characteristics imply that the making of the tool has to be seen as just as important as its use. Like the tail of the peacock, the ability to lavish attention onto the making of a hand axe, and then discard it after little use, must imply a set of evolutionary capacities to impress the opposite sex. We cannot say who was making and who was impressed, but the assumption of many archaeologists is that men made and women gasped.

There is little doubt that all developments in human genetic evolution and technological history prior to about 1.8 million years ago happened in Africa. The earliest non-African sites at Dmanisi in Georgia and Ubeidiya in Israel are not Acheulean assemblages. Although not strictly a matter of African prehistory, the absence of the hand axe and cleaver, and hence the designation Acheulean, from southeast Asian stone tool assemblages, has caused much comment. If it is confirmed that the first 'out of Africa' movement preceded the invention of these bifacial forms, and if it is further confirmed that genetic exchange between southeast Asia and Africa has been minimal after this movement, then a long independent trajectory for both hominid evolution and artefactual tradition in southeast Asia becomes plausible. By contrast, the artefactual records of Europe, southwestern Asia, and Africa remain rather tightly linked throughout most of the Pleistocene.

In an intriguing cycle of argument and counter argument, it now appears that Raymond Dart's belief that australopithecines or early members of the *Homo* lineage were both firemakers and bone tool makers may have some support. Dart's original ideas were based on his work at Makapansgat, South Africa, which he initially believed to be a place where australopithecines lived. The hearths, however, turned out to be manganese staining. C. K. (Bob) Brain, in an inspiring programme of taphonomic, experimental, and ethno-archaeological research, showed that the complex stratigraphies of the (then) Transvaal dolomite caves were places where early hominids died and were deposited by predators. Patterns in associated faunal remains were produced by the actions of predators and other forces on bones of variable density, not by tool makers' choices. It is ironic, then, that working at the stratigraphically better understood site of Swartkrans, Brain later began to record the existence there of evidence for both fire and bone tools. It should be recognized that the Swartkrans evidence is much later than that claimed by Dart from Makapansgat, and that the argument is much more robust, if still somewhat vulnerable. The current claims from Swartkrans are that over a hundred small pointed bones were used for excavating underground plant foods and that charred bones among the faunal remains reflect the controlled use of fire, both claims for behaviors predating 1 million years.

3. *Regionality and the Appearance of Modern Humans*

The transition from Earlier to Middle Stone Age assemblages is poorly understood, in part because it happened much before radiocarbon dating became useful and after the time period best suited to potassium argon dating. Many Middle Stone Age (MSA) sites are open sites difficult, in any case, to date. Most archaeologists presume that the assemblages without hand axes, characterized by some facetted platform flakes and radial and Levallois cores, first appear about 250,000 years ago, but there is much room for debate. The decision no longer to make hand axes or cleavers from about this time may signal the invention and dissemination of methods of mounting and hafting stone points onto wooden shafts. Assuming this required smaller stone components, we might interpret the many kinds of points, unretouched, unifacial, bifacial, leaf-shaped, hollow-based, tanged and serrated, as versions of the spear point. As Desmond Clark pointed out some time ago, regionalization of point 'style' characterizes the MSA and distinguishes it from the preceding Earlier Stone Age (ESA), where change was slow and dissemination of form more encompassing. For some archaeologists this regionality within the MSA, along with the more rapid replacement of one form by another, is a mark of symbolic storage, because it reflects decisions by presumably identity-conscious groups, choices that reflect not function but style, a first indication of 'them' and 'us.'

The most dramatic development in thinking about the African Middle Stone Age has been its remarkable rise, in Desmond Clark's words, 'from peripheral to paramount.' In the late 1960s, bolstered by a rather naive belief in the accuracy of some of the earliest radiocarbon results obtained, MSA assemblages were thought to be contemporary with Upper Palaeolithic assemblages in Europe. Because of the obvious similarities between MSA and Middle, rather than Upper, Palaeolithic stone tool technologies and types, this was taken as a reflection of Africa's backwardness in

progressing along the (presumed standard) tool-making trajectory. It was also wrongly believed at that time that the Kabwe (formerly Broken Hill) cranium and the Saldanha calvarium were 'African neanderthalers.' After Vogel and Beaumont began to doubt the radiocarbon 'dates' for MSA assemblages, new, more experimental dating techniques, such as luminescence and ESR dating, became available (see *Chronology*, *Stratigraphy*, *and Dating Methods in Archaeology*). These have shown that almost all of the MSA is beyond the range of radiocarbon dating and confirmed the rough contemporaneity of Middle Palaeolithic and Middle Stone Age. We should also not underestimate the impact of the genetic work on mitochondrial DNA of the mid 1980s on archaeologists' thinking about the age of the MSA and the significance of the morphology of associated hominid remains. The 'out of Africa' genetic model, a re-evaluation of the modernity of MSA hominids and new dating assessments for African MSA sites have transformed our understanding of the origins of anatomically and behaviorally modern people. In this process, Africa has emerged as paramount, the likely source area of modernity in both senses.

Central to this change of thinking has been the set of assemblages known to John Goodwin and referred to by him as Howiesons Poort. These formed part of the short-lived notion of a 'second intermediate' between the Middle and Later Stone Ages, but have been shown repeatedly to lie stratigraphically within the suite of MSA assemblages, most effectively at Klasies River (South Africa) main site. Howiesons Poort assemblages are MSA in that they contain radial cores and sub-triangular flakes with facetted platforms, but they have an interesting set of novel characters that make them, in some estimations, precocious manifestations of later stone tool making patterns. Unlike assemblages above and below them, for example, there is a tendency for tool makers to prefer finer grained rocks and to manufacture substantial numbers of blades, some of them punch-struck, from them. There are also many backed pieces, including segments, some of them microlithic, trapezes and truncated blades, and in some assemblages, burins. These assemblages are increasingly well dated to beyond 60,000 years, which has interesting implications for these 'Upper Palaeolithic-like' characteristics, which seem to appear in Europe, along with unquestionably modern people, some 20,000 or more years later.

The African Middle Stone Age, but not specifically the Howiesons Poort, is increasingly viewed as the product of anatomically modern people. From Dar es Soltan in the north to Klasies River main site in the south, MSA hominids are anatomically modern, as indeed, or nearly so, are their contemporaries at Middle Palaeolithic sites in southwest Asia. Klasies River main site is, in fact, the most persuasive example of a set of associations that constitute anatomical and behavioral modernity. A deep depositional pile that is unquestionably a shell midden, associations of food-waste and stone tools with hearths, well dated by radiometric dates and faunal observations to Marine Isotope Stages 5 and 4, and with excellent stratigraphic associations of MSA artefacts, including a Howiesons Poort component, with enough hominid fragments to be sure of a modern morphology, sets the standard for claims about modernity. Elsewhere in southern Africa, MSA assemblages are associated with abundant, clearly utilized ochre, bone tools, and decorated pieces of ostrich eggshell. So far it is not clear whether the relative paucity of these kinds of associations elsewhere in Africa is due to incomplete survey, poor preservation or the geography of ancient innovations. Claims for MSA manufacture of bone harpoons at Katanda would, if confirmed, support the innovative character of African MSA people.

Why, we may wonder, has Africa become so clearly paramount, so definitively not peripheral in the evolutionary history of modern people. One possibility is the accumulating evidence for the necessary consumption of marine and freshwater foods as the nutritional substrate for intellectual advances. Rather than being a consequence of increased intelligence, something that modern people did, the regular gathering of shellfish and other marine and freshwater organisms, appears to be essential if people were to acquire the structural lipids necessary to build a bigger brain. Coastal, stream, and lakeside site locations have been the pattern in Africa since the time of early hominids: it may be as much the available foods as the accessibility of water that had significance. Because the earliest shell middens are located along the shores of the Mediterranean-type terrestrial ecosystems at the northern and southern extremities of Africa, it is from here that the surge to modernity may have been fueled.

4. *The Survival of Hunting and Gathering*

Our understanding of the transition from Middle to Later Stone Age (LSA) is complicated by poor dating resolution, highly episodic site occupations, and by the fact that it apparently happened when much of the central parts of northern and southern Africa were even more arid than they are today. Additionally, at this time, approximately 20,000 to 50,000 years ago, the mean sea level was lower and some of the most attractive parts of the subcontinent were the exposed continental shelves, now lost to study by the rise in sea level. One interesting consequence of a lengthy expansion during Marine Isotope Stages 2, 3, and 4 of the Kalahari and Sahara arid zones, is the probable reduction in genetic exchanges between subtropical African populations and their contemporaries in the two Mediterranean-type ecosystems to the north and south. It was perhaps during this time, after the emergence and expansion out of Africa of modern

people, that the current geographic patterns in African human phenotypes became established. Apparently supporting this projection is the increasingly regionally differentiated and localized nature of the African LSA.

Goodwin was correct in recognizing little similarity between the African LSA as a whole and the European Upper Palaeolithic. The transition from the MSA to the LSA technologically involved the abandonment of radial and Levallois core techniques and the focusing on small bipolar and unipolar cores that had already appeared in the MSA (see *Lithics and Archaeology*). It may have happened at different times in different parts of the sub-continent, something that now appears likely in the European transition from Middle to Upper Palaeolithic as well. Microlithic assemblages appear much earlier in Kenya than in southern Africa, but robust regional chronological patterns are yet to emerge.

Certainly by 20,000 years ago the sub-continent is widely but patchily occupied by hunter gatherers making a variety of (mostly) microlithic tools on fine grained raw materials and, where preservation is good, leaving behind tools of ostrich eggshell, bone and other organic raw materials that loosely resemble those of the recent ethnographic past. There are, however, many important technological, social, and demographic changes within the last 20,000 years, especially as pastoralism and farming began to spread slowly and irregularly through the ecosystems of sub-Saharan Africa.

An important component of the heritage of LSA hunter–gatherers is the large corpus of paintings and engravings, commonly referred to as rock art (see *Prehistoric Art*). Throughout southern and eastern Africa there are images made by farming people, usually in white paint, often presenting geometric forms or images of colonial times and almost always made with the finger. Easily distinguished from these is a tradition that extends south from Zambia, Zimbabwe, northern Botswana, and Namibia to the Cape, characterized by relatively naturalistic fine line images of people and animals, in places overlain by handprints. These paintings and engravings may have begun to appear as early as 25,000 years ago, were certainly widespread by 10,000 years ago and were still being made in the century before last. From the research of Partricia Vinnicombe and David Lewis-Williams, we can now relate the images to the world view of surviving San hunter–gatherers and their ancestors. A particularly important set of reference material has been the texts recorded by Wilhelm Bleek and his sister in law Lucy Lloyd in the 1870s and related to them by a group of Xam San informants from the South African interior. Recent Kalahari hunter–gatherers do not themselves paint or engrave, but clearly subscribed to the same cosmology as the now extinct Xam. A point of some interest is the corpus of Tanzanian fine line images that resemble those from much further south. There is here an artistic parallel to the enigma of 'click' languages in more or less the same part of eastern Africa, also isolated from a host of such languages in the south. No doubt there is still much to learn about the history of the interface between Khoesan people and their traditions with other Africans to the north.

There is a significant irony in the discovery of Africa's paramount role in the story of human evolution. Because European science gave rise to the intellectual discipline of archaeology, European scientists have always been searching for origins in Europe. Gradually and serially these have failed, to be replaced by a growing realization of the African contribution. But it is tragic that with few exceptions this legacy of African prehistory has been revealed by people of European ancestry, albeit ones with a strong sense of belonging to Africa. But then, in the longer view, we are all Africans.

See also: African Studies: History; Hunter–Gatherer Societies, Archaeology of; Hunting and Gathering Societies in Anthropology

Bibliography

Goodwin A J H, van Riet Lowe C 1929 The Stone Age cultures of South Africa. *Annals of the South African Museum*. **27**: 1–289

Klein R G 1999 *The Human Career*, 2nd edn. Chicago University Press, Chicago

J. Parkington

Substance Abuse in Adolescents, Prevention of

Substance abuse is a problem that countries throughout the world have struggled with for many decades. In the USA where substance abuse is the highest in the industrialized world, the annual prevalence of illicit drug use among adolescents peaked in 1979, declined for more than a decade, and then rose sharply throughout most of the 1990s (Johnston et al. 1995/96; see Fig. 1). Concern for substance abuse as a problem derives from the fact that it can lead to a host of adverse consequences including long-term disability and premature death. Broadly conceived, substance abuse involves the inappropriate use or misuse of one or more psychoactive substances including tobacco,

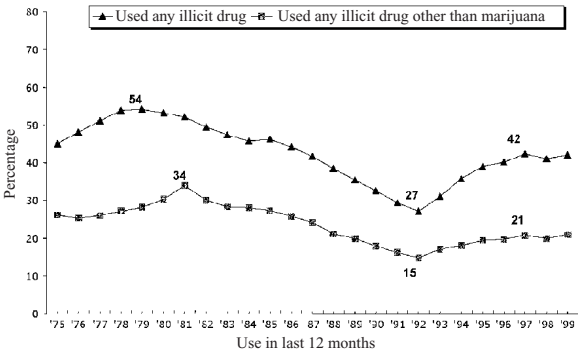

Figure 1
Trends in annual prevalence of an illicit drug use index for 12th graders (Johnston et al. 1995/96)

alcohol, and illicit drugs such as marijuana, cocaine, and heroin.

For most individuals, substance abuse has its roots in childhood and adolescence (Hartel and Glantz 1999). It begins with a brief period of 'experimental' use of one or more substance followed by escalation in both the frequency and amount of the substance(s) used. As use increases, so too does tolerance for the pharmacological effects of most substances, leading to further increases in the amount used per occasion and the frequency of use as well, as the eventual development of psychological and/or physical dependence. Thus, the developmental course of substance abuse progresses from onset and occasional use to frequent and/or heavy addictive use. It also typically progresses from the use of widely available legal substances (alcohol and tobacco) to the use of illicit substances.

Efforts to combat the problem of substance abuse have included actions by law enforcement agencies to interdict the flow of drugs, legislative initiatives, treatment interventions, mass media campaigns, and prevention programs. In view of the obvious difficulty of solving the problem of substance abuse through enforcement or treatment approaches, many experts have long argued that prevention offered the greatest potential. Since the early 1980s, there has been considerable progress in understanding both the causes of substance abuse and how to prevent it. This article summarizes what is currently known about both effective and ineffective prevention approaches.

1. *Informational and Affective Approaches*

Efforts to prevent substance abuse have largely emphasized the dissemination of factual information about psychoactive substances, their pharmacological effects, and the adverse consequences of use (Botvin et al. 1995b). These approaches assume that individuals begin using both licit and illicit substances after logically considering the potential risks and expected benefits. It is further assumed that individuals who elect to use drugs do so because they are unaware of the adverse consequences of using drugs. Given this, prevention approaches were initially designed to provide information about the risks associated with substance use so that people would be more likely to choose not to use drugs. Closely related to information dissemination were fear arousal approaches to prevention that attempted to deter use by dramatizing the dangers of substance abuse. Despite the intuitive appeal of these approaches, evaluation studies have consistently found that they do not decrease substance use (Schaps et al. 1981). Some studies suggest that these approaches may even increase substance use by stimulating curiosity.

Another approach to substance abuse prevention that was widely used during the 1960s and 1970s attempted to prevent substance abuse by promoting affective development. These 'affective education' approaches, as they were called, were designed to increase self-understanding and acceptance through activities such as values clarification and responsible decision making, improving interpersonal relations by promoting effective communication and assertiveness, and providing students with the opportunity to fulfill their basic needs through existing social institutions. Once again, evaluation studies have found little evidence to support the effectiveness of these approaches (Schaps et al. 1981).

2. *Social Influence Approaches*

In contrast to the cognitive and affective approaches, prevention approaches targeting the social factors that promote and sustain adolescent substance abuse have shown considerably more promise. These approaches were based on studies showing the important role that social factors play in the initiation and development of substance abuse. They are also based on social psychological theories of human behavior, particularly social learning theory, persuasive communications theory, and problem behavior theory.

2.1 *Psychological Inoculation*

Derived from persuasive communications theory, psychological inoculation involves first exposing adolescents to weak social influences promoting substance use, followed later by progressively stronger social influences. Through this process, it was hypothesized that the exposed adolescents would build up resistance to more powerful prodrug messages, similar to what they would be likely to encounter as they entered junior high school. Despite the prominence of psychological inoculation in the initial formulation of the social influence prevention approach, it has not been supported by the empirical evidence and consequently has received considerably less emphasis in

more recent variations of the social influence model. Other components of the social influence model have assumed greater emphasis, particularly normative education and resistance skills training.

2.2 Normative Education

The prevalence of smoking, drinking, and illicit drug use is usually overestimated by adolescents (Fishbein 1977). This overestimate leads to the formation of inaccurate normative expectations—the forming of a set of normative beliefs that support substance use. In order to reduce pressure to use drugs, social influence approaches to substance abuse prevention generally include a component designed to correct the misperception that many adults and most adolescents use drugs. Two different strategies have been used to modify or correct normative expectations. The first is straightforward and simply involves providing adolescents with information from national or local survey data concerning the prevalence of substance use among adolescence and/or adults. A second method involves having adolescents conduct their own local surveys of substance use within their school or community. This information provides adolescents with a more realistic view of the prevalence of substance use and, because substance use in most communities is far lower than adolescents believe, it corrects the misperception that most people use drugs and helps establish antisubstance use norms.

2.3 Drug Resistance Skills

Drug resistance skills are the hallmark of most contemporary approaches to substance abuse prevention (Ellickson and Bell 1990). The primary focus of these prevention approaches is on teaching adolescents a set of skills or tactics for resisting influences to smoke, drink, or use illicit drugs—particularly influences coming from the media and peers. Adolescents are taught how to identify and respond to prodrug messages in advertisements, movies, rock videos, etc. They are also taught to identify the persuasive appeals used by advertisers to promote the sale of tobacco products and alcoholic beverages, to analyze advertisements and their messages, and to formulate counterarguments to common advertising appeals (Flay et al. 1989). Teaching these skills is designed to provide adolescents with the tools for actively resisting advertisements promoting cigarette smoking and alcoholic beverages, particularly given the growing recognition that advertisers have developed rather sophisticated ad campaigns explicitly designed to target the youth market.

Perhaps the best known aspect of resistance skills training involves teaching adolescents a set of skills that they can use to resist offers from their peers to smoke, drink, or use illicit drugs. Prevention programs that include resistance skills training emphasize verbal and nonverbal skills for resisting these offers. These programs teach adolescents what to say when they are offered or pressured to engage in some form of substance use as well as teaching them how to say it in the most effective way possible. Adolescents are taught to identify 'high-risk' situations where they are likely to experience peer pressure to use drugs, are taught to avoid those situations whenever possible, and are taught to develop 'action plans' so that they are prepared to handle situations involving peer pressure to use drugs.

In order to be optimally effective, resistance skills training must provide adolescents with a general repertoire of verbal and nonverbal skills that they can use when confronted by peer pressure to use drugs. Adolescents must not only develop these skills, but also the confidence that they can apply them when necessary. To accomplish this, prevention programs teach students these resistance skills and also provide opportunities to practice the use of these skills.

3. Personal and Social Skills Training

Since the early 1980s, prevention approaches have also been developed and tested that are broader in their focus than social influence approaches. These approaches are designed to teach information and skills that are specifically related to the problem of substance abuse within the context of enhancing the development of general personal competence, hence they are also referred to as 'competence enhancement' approaches. Although these approaches include elements of the social influence model, a distinguishing feature is an emphasis on teaching a variety of personal self-management skills and general social skills. An underlying assumption of this approach is that merely teaching drug refusal skills is not likely to be effective with adolescents who are motivated to use drugs. Put differently, teaching adolescents how to say no to drugs will not succeed among those adolescents who want to say yes. Moreover, advocates of this approach argue that prevention approaches that target a more comprehensive set of risk and protective factors are more likely to be effective than those intended to target a limited number of these factors.

Competence enhancement approaches teach generic skills such as general problem-solving and decision-making skills, cognitive skills for resisting interpersonal and media influences, skills for increasing self-control and self-esteem, skills for coping effectively with anxiety or stress, general social skills, and assertive skills. One prominent approach, called *life skills training* (LST), teaches several types of self-management skills such as self-appraisal, goal setting and self-directed behavior change, decision making and independent thinking, and anxiety management skills (Botvin et al. 1990). Adolescents are also taught

a set of general social skills including communications skills, conversational skills, skills for asking someone out for a date, complimenting skills, and both verbal and nonverbal assertive skills. All of these skills are taught using a combination of instruction and demonstration, behavioral rehearsal (practice), feedback, social reinforcement (praise), and extended practice through behavioral homework assignments.

4. Evidence of Effectiveness

A growing literature shows that several approaches to substance abuse prevention can substantially reduce adolescent substance use. However, it is now clear that not all prevention approaches are effective. In fact, the only prevention approaches that have consistently reduced substance use are those that teach resistance skills and normative education, either alone or in combination with components designed to enhance general personal and social competence. Studies testing these approaches provide strong evidence for the short-term effectiveness of prevention, but somewhat weaker evidence of long-term effectiveness.

4.1 Short-term Effects

Studies testing the effectiveness of social influence approaches to cigarette smoking have generally shown that they are able to reduce the rate of smoking by 30 to 50 percent after the initial period of intervention (e.g., Luepker et al. 1983, Ellickson and Bell 1990, Sussman et al. 1993). One study has also demonstrated that the social influence approach can reduce the use of smokeless tobacco (Sussman et al. 1993). Studies testing interventions that contain social influence components in combination with personal and social skills training components show that they can cut cigarette smoking by 40 to 75 percent (Botvin et al. 1995a, Schinke 1984).

Social influences approaches have also been shown to reduce alcohol and marijuana use by roughly the same magnitude as for tobacco use (Ellickson and Bell 1990). Similarly, studies show that the personal and social skills training approaches can also reduce the use of alcoholic beverages and marijuana (Botvin et al. 1990). When comparing the relative impact of these prevention approaches on tobacco, alcohol, and marijuana use, the strongest effects are for cigarette smoking and marijuana use, while there are somewhat weaker and less consistent effects on the use of alcoholic beverages.

It is important to recognize that the social influence approaches and the competence enhancement approaches discussed here both include program components designed to increase drug resistance skills and promote antidrug norms. The results from one study (Caplan et al. 1992) suggests that generic competence enhancement approaches to substance abuse prevention may not be effective unless they also teach normative education and/or resistance skills training. Although approaches that combine the key elements of the social influences approach and the broader competence enhancement approach appear to be the most effective based on current evidence, additional research is necessary to better understand how and why effective prevention programs work.

4.2 Long-term Effects

The results of longer-term follow-up studies present a somewhat mixed picture. Evidence from several studies indicates that current prevention approaches have a reasonable degree of durability. Studies show prevention effects lasting anywhere from one to four years (Luepker et al. 1983, Flay et al. 1989).

Several factors appear to play a role in determining how durable these prevention approaches are. These include (a) the length of the initial intervention, (b) the inclusion of ongoing (multiyear) intervention or booster sessions, (c) the quality and completeness of implementation, and/or (d) the number of risk and protective factors targeted. Prevention approaches that produced longer-lasting effects generally have a stronger initial dosage (eight class sessions or more), include booster sessions, are implemented with high fidelity, and utilize a more comprehensive prevention approach.

Results from a large-scale, randomized field trial involving nearly 6,000 seventh graders from 56 public schools in the USA found long-term reductions in smoking, alcohol, and marijuana use at the end of the twelfth grade (Botvin et al. 1995a). In this study schools were randomly assigned to prevention and control conditions. Participating students in the prevention condition received the LST program during the seventh grade (15 sessions) with booster sessions in the eighth grade (10 sessions) and ninth grade (five sessions). No intervention was provided during grades 10 through 12. According to the results, the prevalence of cigarette smoking, alcohol use, and marijuana use for the students in the prevention condition was as much as 44 percent lower than for controls. Significant reductions were also found with respect to the percentage of students who were using multiple substances one or more times per week. These results suggest that to be effective, prevention programs need to be more comprehensive and have a stronger initial dosage than most studies using the social influence approach. Another implication is that prevention programs also need to include at least two additional years of interventions (booster sessions) and must be implemented in a manner that is consistent with the underlying prevention model.

4.3 Effectiveness with Ethnic Minority Youth

Several studies have examined the impact of prevention approaches on minority youth. Two strategies

have been followed. One strategy, common to many local prevention efforts, assumes that the etiology of substance abuse is different for different populations. Thus, many prevention providers have developed interventions designed to be population-specific (e.g., projects for African-American youth with an Afrocentric cultural focus). A second strategy assumes that the etiology of substance abuse is more similar than different across populations. This approach involves the development of preventive interventions designed to be generalizable to a broad range of individuals from different populations. As information concerning the etiology of substance use among minority youth has accumulated, it has become clear that there is substantial overlap in the factors promoting and maintaining substance use among different populations. The results of these studies suggest that prevention approaches may have application to a broad range of individuals from different populations. Indeed, studies show that both the social influence approach and the combined social influence/competence enhancement approach are effective with minority youth (Botvin et al. 1995b). Research also shows, however, that in environments with homogeneous populations, prevention effects can be further enhanced by tailoring the intervention to the target populations.

5. *Concluding Comments*

Substantial progress has been made since the early 1980s in the area of substance abuse prevention. Much has been learned concerning the etiology of substance abuse and how to prevent it. Despite the advances in prevention, additional research is needed to determine the long-term effectiveness of current approaches; further refine existing approaches, and identify new ones that may be more powerful; identify and test the mechanisms through which effective programs work; develop and test more comprehensive prevention strategies that combine the best features of existing prevention modalities; and extend the findings from substance abuse prevention to other related health and social problems such as HIV/AIDS, teen pregnancy, violence, and delinquency. Finally, it is vitally important that a concerted effort be made to translate research into practice in order to capitalize on the progress that has been made in prevention research.

See also: Crime and Delinquency, Prevention of

Bibliography

Botvin G J, Baker E, Dusenbury L, Botvin E M, Diaz T 1995a Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population. *Journal of the American Medical Association* **273**(14): 1106–12

Botvin G J, Baker E, Dusenbury L, Tortu S, Botvin E M 1990 Preventing adolescent drug abuse through a multimodal cognitive-behavioral approach: Results of a three-year study. *Journal of Consulting and Clinical Psychology* **58**: 437–46

Botvin G J, Schinke S, Orlandi M 1995b *Drug Abuse Prevention with Multi-ethnic Youth*. Sage, Newbury Park, CA

Caplan M, Weissberg R P, Grober J S, Sivo P, Grady K, Jacoby C 1992 Social competence promotion with inner-city and suburban young adolescents: Effects of social adjustment and alcohol use. *Journal of Consulting and Clinical Psychology* **60**: 56–63

Ellickson P L, Bell R M 1990 Drug prevention in junior high: A multi-site longitudinal test. *Science* **247**: 1299–1305

Fishbein M 1977 Consumer beliefs and behavior with respect to cigarette smoking: A critical analysis of the public literature. In: Anonymous (ed.) *Federal Trade Commission Report to Congress Pursuant to the Public Health Cigarette Smoking Act of 1976*. US Government Printing Office, Washington, DC

Flay B R, Keopke D, Thomson S J, Santi S, Best J A, Brown K S 1989 Long-term follow-up of the first waterloo smoking prevention trial. *American Journal of Public Health* **79**(10): 1371–6

Hartel C R, Glantz M D 1999 *Drug Abuse: Origins and Interventions*. APA Books, Washington, DC

Johnston L D, O'Malley P M, Bachman J G 1995/96 *National Survey Results on Drug Use From the Monitoring the Future Study, 1975–1998. Vol. II: College Students and Young Adults*. National Institute on Drug Abuse, Rockville, MD

Luepker R V, Johnson C A, Murray D M, Pechacek T F 1983 Prevention of cigarette smoking: Three year follow-up of educational programs for youth. *Journal of Behavioral Medicine* **6**: 53–61

Schaps E, Bartolo R D, Moskowitz J, Palley C S, Churgin S 1981 A review of 127 drug abuse prevention program evaluations. *Journal of Drug Issues* **11**: 17–43

Schinke S P 1984 Preventing teenage pregnancy. *Progress in Behavior Modification* **16**: 31–64

Sussman S, Dent C W, Stacy A W, Sun P 1993 Project towards no tobacco use: 1-year behavior outcomes. *American Journal of Public Health* **83**: 1245–50

G. J. Botvin

Suicide

1. *Definition*

The term suicide was mentioned for the first time in medieval times in 1177 (van Hooff 1990) according to the Latin *sui cidium* (self-killing) and *sui caedere* (to kill himself/herself). Now, this term is used in the international scientific literature. It avoids judgment, as do terms like 'self murder' and 'voluntary death.'

In the realm of the term suicide, there are other terms that have to be defined. Suicide ideas include: thinking about death either of one's own or someone else's or death in general; the wish to die; suicide ideas in a more narrow sense. It is not easy to define the term suicide attempt. The definition most commonly ac-

cepted is the following: 'An act with nonfatal outcome, in which an individual deliberately indicates a non-habitual behavior that, without the intervention of others, will cause self-harm, or deliberately ingests a substance in excess of the prescribed or generally recognized therapeutic dosage, and which is aimed at realizing changes which the subject desired via the actual or expected physical consequences' (WHO 1989, Platt et al. 1992). This definition includes an active intention for self-harm but not necessarily self-killing. Suicide and suicide attempt are also considered as suicidal behavior.

2. *Importance of Suicide*

Suicide belongs to the 10 most frequent causes of death among adults in industrialized European and North American countries. In these countries, suicide is considered the second most frequent cause of death among adolescents and young adults (15–35 years). Suicide in young male adults has increased constantly during the last 20 years. Table 1 shows the frequencies of causes of death worldwide (WHO 1990) disregarding the most frequent causes of death such as cardiovascular diseases, stroke, cancer, infectious diseases, and malnutrition. These figures have to be taken *cum grano salis* because of a dearth of or inaccurate statistics on death in developing countries. As shown in Table 1, only fatal car accidents had a similar frequency of death as compared with suicide rates. War activities, crime, or even AIDS at the beginning of the 1990s had a considerably lower death rate.

3. *Epidemiology*

The most recent available suicide rates from all over the world are depicted in Fig. 1. This shows that a high variability of suicide rates exists across cultures and countries. Looking at the rank order of the frequency of suicide rates, a relatively stable pattern can be described—at least for European countries. There exist a North–South down gradient and a East–West down gradient of suicide rates: eastern and northern countries have higher suicide rates than western and southern countries. Hereby, sociological factors seem to play an important role in terms of secularization and industrialization with the collapse of the family group as a consequence as well as climate factors, which cannot be disentangled sufficiently.

Table 1
WHO statistics: causes of death worldwide (1999)

856,000	Death because of car accidents
818,000	Death because of suicide
322,000	Death because of war
292,000	Death because of crime
291,000	Death because of Aids

There exist distinct sex differences with a preponderance of males in suicide and preponderance of females in suicide attempt. This distribution is stable across all countries and cultures. Only China reports a prevalance of suicide among women in selected areas.

The highest rate of suicide is found in people of older age and that of suicide attempt in younger people. Suicidal behavior has been found very rarely in children but, without doubt, there is clinical evidence that children from 4 years of age have suicide ideas and make a suicide attempt or kill themselves. However, suicide occurring in children before prepuberty is very rare. Furthermore, it has to be taken into consideration, that until prepuberty (8–10 years) children normally do not experience death as something irreversible (Pfeffer 1986). Cultural differences exist in terms of using different suicide methods. So-called weak suicide methods (ingestion of drugs, wrist cutting) and hard suicide methods (shooting, hanging, jumping, drowning, etc.) will be distinguished. Weak suicide methods prevail in suicide attempts, hard methods in successful suicide. In general, females more often prefer weak methods than males. The most frequent method for suicide in females and males in North America is shooting, in Europe hanging.

Suicide rates in several European countries such as Germany, England and Wales, France, Hungary, and the Nordic countries show an astonishing stability over decades and even a century. The same holds true for counties such as Saxonia, Brandenburg, and Thuringia, where suicide rates have been stable over a century. Figure 2 depicts suicide rates over one century in Germany (Deutsches Reich and Federal Republic of Germany) (Wedler 1992). Such stability can be traced also (in other countries) back to the nineteenth century such as in Prussia, France, and the Nordic countries. On the other hand, there have been dramatic changes in suicide rates as during the nineteenth century, where secularization, industrialization, and individualization have taken place in western and northern European societies. During World War I and II further dramatic changes occurred not only affecting both males and females but also countries which have not been involved in these wars, such as Switzerland. Finally, the last changes have occurred after the fall of the Iron Curtain in Eastern Europe between 1985 and 1995 with a dramatic decrease in suicide rates by up to 40 percent (Diekstra 1996). (See *Mental Illness, Epidemiology of*.)

4. *Differing and Common Aspects between Suicide and Suicide Attempts*

There are considerable differences between suicide attempters and suicide victims (Stengel 1964). The group of suicide victims is rather small in comparison

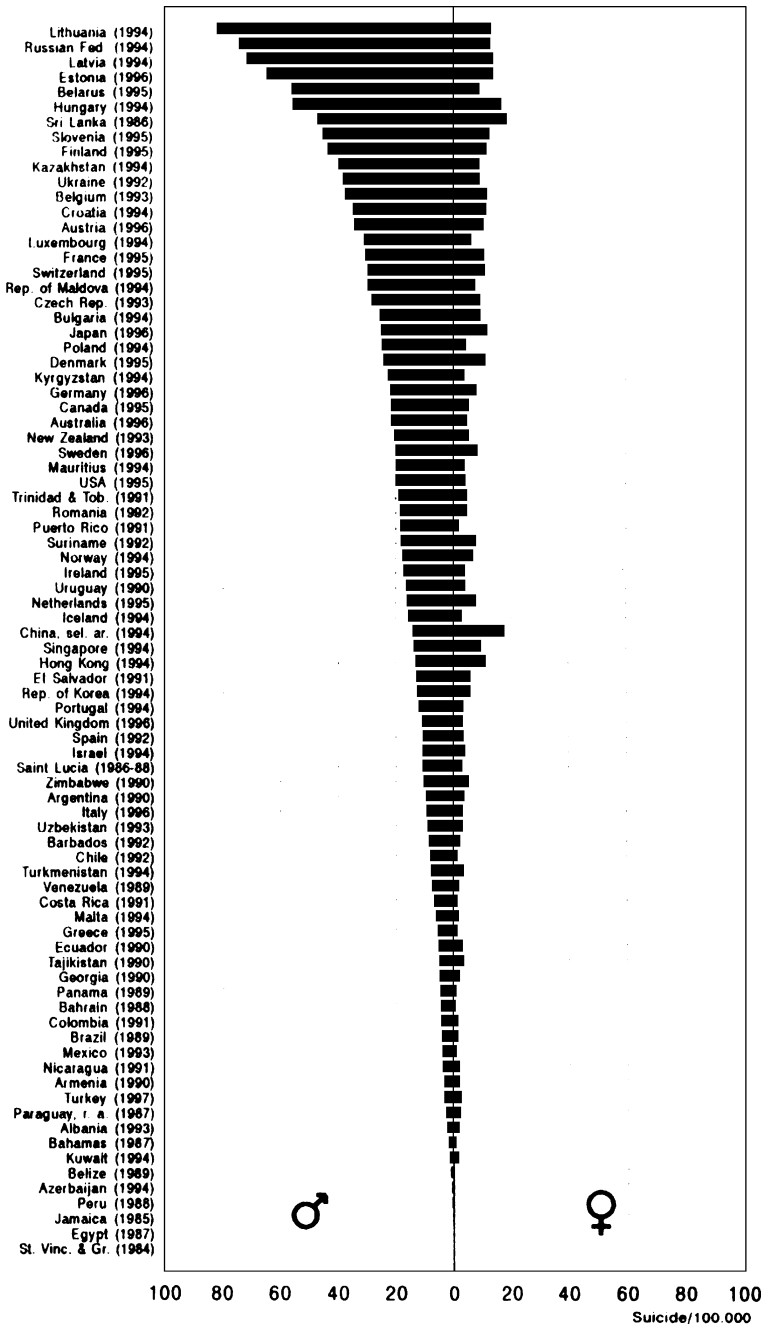

Figure 1
WHO statistics: suicide rates of males and females (1994–96)

with the group of suicide attempters. The sociodemographic characteristics are quite different: suicide victims are older and predominantly male, suicide attempters are younger age and predominantly female. The tendency of a transition from the group of suicide attempters to the group of suicide victims is rather small with a 10 percent chance. Whereas the life of the victim is terminated by suicide, a nonfatal suicide

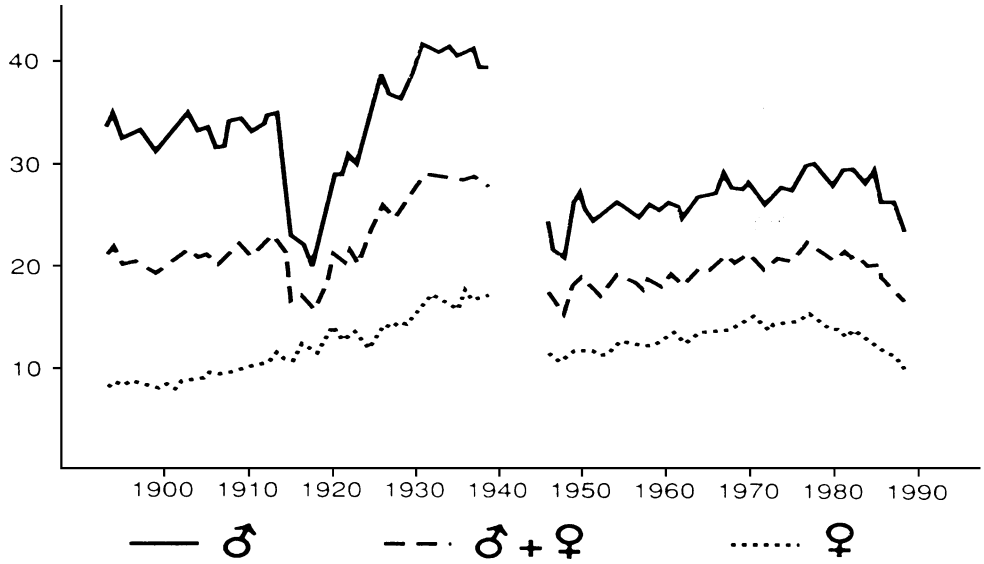

Figure 2
Suicide rates per 100,000 inhabitants in the German Reich from 1900 to 1940 and in West Germany from 1945 to 1990 (Wedler 1992)

attempt becomes a very important event in the life of this person, often with a change in the further life trajectory.

Nevertheless, there are also some common aspects between these two groups: the probability of suicide is higher, the more often suicide attempts happen. Despite the fact that more males are committing suicide and more females make a suicide attempt, there is a considerable overlap between both groups. The same holds true for age distribution: a considerable proportion of young males and females commit suicide, while older males and females make suicide attempts. Most of the risk factors are equally valid for both suicide victims and suicide attempters (see below).

5. *Risk Factors for Suicide and Suicide Attempts*

Risk factors for suicide and suicide attempts are as follows: increased age; marital status like being separated, divorced, widowed; single; unemployment; gender and economical status (suicide occurs more often in males, suicide attempts occur more often in females and the lowest social class). The most important risk factors for suicide and suicide attempts, however, are a previous suicide attempt and a psychiatric disorder (Kreitman 1977).

Psychiatric hospitals bear a great risk of losing a patient to suicide. It has been estimated that 15 percent of depressed patients, 15 percent of patients with a drug or alcohol addiction, and 10 percent of patients with schizophrenia, who have been treated in a psychiatric hospital, finally die by committing suicide. Patients with anxiety disorders or personality disorders also have a considerable risk of committing suicide. From an epidemiological point of view, subjects with a depression or an addiction have a 15-fold to 60-fold higher risk of committing suicide as compared with the general population (Bronisch 1999).

There are differences between urban and rural areas with respect to suicide rates. The highest suicide rates are seen in urban areas, whereas the lower rates are seen in rural areas. An exception are the states of the former USSR, where a preponderance of suicide rates in rural areas is found.

Climate seems also to play an important role in the distribution of suicide rates. The highest suicide rates are found during late spring and summer time. This distribution can be found across the northern as well as the southern hemispheres. Countries in southern parts of Europe show a lower suicide rate than countries in the northern parts of Europe.

Religious affiliation also plays in important role in the frequency of suicides. Especially in those countries with a predominance of Catholics, suicide rates are lower, whereas in countries with a predominance of Protestants, suicide rates are higher. However, if the degree of urbanization and the distribution of social classes are taken into consideration, the differences between catholic and protestant countries are fading (Kreitman 1977).

6. Origins of Suicidal Behavior

6.1 Suicide as a Basic Human Existential Problem

Ever since the written history of mankind, suicide has been considered a basic human existential phenomenon that has been reflected in many philosophical and religious contributions as well as in all kinds of arts. The attitude towards suicide did not only depend on the *Zeitgeist* of the respective historical period, but also on the individual stance of different philosophical schools such as of Stoa and Epikur in antiquity and different religions such as Christianity and Buddhism. The sociological theories are a consequence of the enlightment of the seventeenth and eighteenth centuries. During the second half of the nineteenth century, psychological theories developed upon the influence of psychological thinking and emerging psychoanalysis. During the second half of the twentieth century, biological theories were finally stimulated by the natural sciences focusing on the brain.

Suicide requires self-reflecting properties of an individual, i.e., an individual who can differentiate between an observing and experiencing ego. Suicide does not exist in animals. Even nonhuman primates do not show suicidal behavior, neither in natural settings nor in preservations. However, suicide has been found in Neolithic societies.

6.2 Sociological Aspects

6.2.1 Durkheim's theory. Sociological theories are influenced by the theory of Durkheim (1897) based upon the evaluation of suicide statistics of western European countries (France and Prussia) during the second half of the nineteenth century. Durkheim proposed four different types of suicide depending on the degree of successful adaptation of the individual to different types of societies. A prerequisite to adaptation of the individual to society is that individualization is neither too weak nor too strong. If the individualization is too strong, the individual will be isolated from society and the danger of an egoistic suicide increases. If the individualization is too weak, the danger of an altruistic suicide increases.

On the other hand, a prerequisite to the equilibrium between the individual and society is that social norms are neither too strong nor too unspecific. Norms that are too strong facilitate fatalistic suicide such as an expiatory death. Norms that are too unspecific favor anomic suicide. With the help of the existing suicide statistics, Durkheim tried to verify three of the four types (he did not discuss the fatalistic suicide). He tried to prove that Protestant countries have a higher suicide rate than Catholic countries, since Protestantism allows the individual a higher degree of freedom than does Catholicism: it is called egoistic suicide. He explained the increase in suicide rates during political crises (Dreyfuss affair in France) in terms of an egoistic suicide. During times of flourishing economic growth, implying a derestriction of social norms, suicide rates should lead to an anomic suicide, whereas during times of war, stronger norms lead to a stronger coherence of society resulting in a decrease in suicide rates.

6.2.2 General sociological trends. During the nineteenth century, only parts of Durkheim's theories could be confirmed such as the differences between Catholic and Protestant countries in western Europe and the decrease of suicide rates during both world wars. Generally speaking, Durkheim's data and statistical data from the twentieth centuries could demonstrate an increase in suicide in Europe during the nineteenth and twentieth centuries depending on the development of industrialization, secularization, and individualization of the respective countries, i.e., the more pronounced these developments were the more frequent suicide and suicide attempts occurred (Retterstol 1992). However, through the fall of the Iron Curtain in Eastern Europe between 1985 and 1995, the theory of anomic suicide has been opposed. Also, a remarkably stable sex difference with male preponderance in suicide and female preponderance in suicide attempts has been shown in nearly all cultures and societies. Another quite remarkable trend of the last decades has been the increase in suicide rate of 15–35-year-old adolescents and young adults. In most western societies, this development has been correlated with an increasing rate of depression and addiction potential. An increasing urbanization, loss of social structures and familial bonds, as well as changes in sex roles may play an important role in this development (Klerman 1987).

6.3 Psychological Aspects

Psychological models include psychodynamic, attachment, and cognitive-behavioral theories that are only partially substantiated by empirical research. The main theories are concerning aggression directed inward instead of outward, the loss of a symbiotic relationship, and a severe personal insult resulting in severe loss of self-esteem. Finally, classical conditioning, operant conditioning, and model learning are, in terms of the behavioral point of view, also of great importance.

6.3.1 Aggression directed against the ego. Freud (1917) conceptualized one of the first psychological theories about the development of suicidal behavior in 'mourning and melancholia.' To him, the psychodynamics of suicidal behavior and depression were

indistinguishable, i.e., suicidal behavior and depression were seen as a turn of aggression against the ego and an ambivalent attitude towards other people. Aggression directed against other people will be followed by loss of a desperately needed person. Turn of the aggression against the ego is expressed by feelings of guilt, self-devaluation, and finally self-killing.

6.3.2 Distorted cognitive schemes. Beck (1967) also emphasized a strong connection between suicidal behavior and depression. The depressive sees himself, the world, and the future (so called cognitive triade) negatively because of their distorted cognitive schemes. Catastrophization, minimizing, maximizing, and overgeneralization belong to typically distorted cognitive schemes. As a consequence of such a cognitive triade, helpnessless, hopelessness, and powerlessness result. In this case, the only solution for an individual may be to end their miserable life by committing suicide.

6.3.3 Narcissistic crisis. Henseler (1974/1984) formulated the first narcissistic theory of suicidal behavior. Individuals with narcissistic personality traits are people who can be hurt easily by significant others through criticism or lack of special attention. They are lacking a stable self-esteem and their self-perception is distorted: they oscillate between overestimation and underestimation of their own abilities. This is not restricted to one's own character but also extends onto a corresponding distorted view of other people. Insults through criticism or lack of special attention lead to aggressive outbursts that can be directed either towards others or their own character in terms of a suicide or suicide attempt.

6.3.4 Appeal to human bonding. Attachment of the newborn to the mother is a characteristic common to all mammals. In all mammals, the newborn is very dependent and the reproductivity rate at least in the nonhuman and human primates is rather small. Therefore, the strong attachment is a life-preserving measure of the evolution to protect the dependent newborn until it is grown up. The emotional bonding in primates is very intense, a separation of the mother will be experienced as very painful (Bowlby 1977a, 1977b).

As has been known from many empirical studies, most suicides and suicide attempts are consequences of imminent separation, separation, or divorce of males and females. The intention to die is ambivalent in most suicide attempts, i.e., besides a true tendency to end life in nearly all suicide attempts there are tendencies to survive. That can be seen easily in the choice of suicide methods, which often are not so dangerous, i.e., ingestion of small amounts of pills, or in the arrangement of suicide attempt that leaves a rescue open, i.e., by attempting suicide in the presence of a significant other.

The painfully expected or experienced separation from a significant other and the appeal of the significant others not to be left behind both seem to play an important role in the meaning of a suicide attempt (Bronisch 1995).

6.3.5 Classical and operant conditioning. Basic principles of behavioral theory and therapy can be described using the Stimulus Organism Reaction Kontingency Consequence Scheme (SORKC scheme) of Kanfer (Schmidtke 1988) (see Fig. 3). The SORKC scheme tries to describe each behavior in a broader context of conditions, which can be placed linearly in a chain or circle. The constituents of the SORKC scheme for suicidal behavior are the following:

(a) life events and social situations as Stimuli variables (S), which may trigger suicidal behavior;

(b) behavioral repertoire based upon organic conditions of the individual such as psychiatric disorders and the personality (e.g., cognitive styles and stances) as Organism variable (O), which may predispose to suicidal behavior;

(c) suicidal behavior as a reaction (R) to the S and O variables;

(d) Kontingency (K), i.e., the reinforcement conditions, which support or suppress suicidal behavior; and behavior according to a regular or irregular order of the consequences;

(e) Consequences (C) in the sense of negative reinforcement (e.g., death) or positive reinforcement (attention of the environment) of suicidal behavior.

The process hereby is rather circular, i.e., consequences can again influence further suicidal behavior. This might be possible by changing the individual (O variable) and the surroundings (S variable).

6.3.6 Model learning. Imitation of suicidal behavior is a new area in suicide research. The observation that suicidal behavior is running in certain families, cultures, and countries could be explained by genetic transmission, psychological conflicts, or by model learning. In regard to model learning, three hypotheses have been proposed: the imitation hypothesis, the contagion hypothesis, and the suggestion hypothesis. All three hypotheses have in common that suicidal behavior of a model will be imitated, whereby predisposition of the individual (so-called 'presuicidal personality'), lack of social support in coping with life crises and enduring social burden are the contributing factors. Furthermore, a high degree

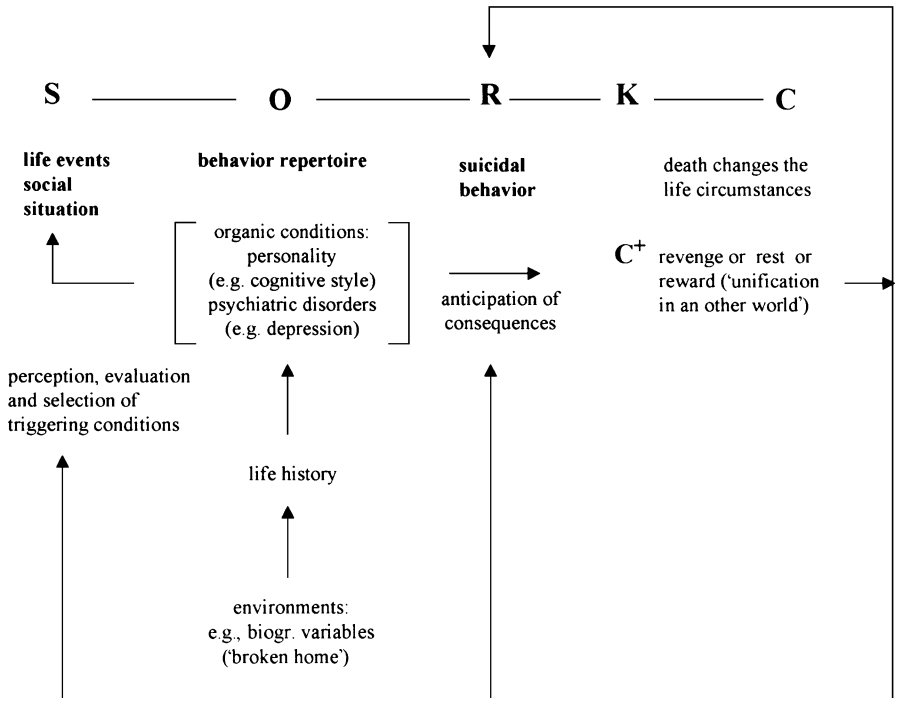

Figure 3
The cognitive-behavioral scheme for the development of suicidal behavior

of social suggestibility of predisposed individuals plays an important role.

During the 1980s and 1990s the influence of different media, such as books, newspapers, journals, television, and movies, on increasing suicide rates can be demonstrated (Phillips and Lesyna 1995). Suicide victims have resembled the person portrayed by the respective media: there have been similarities between this person and the suicide victim in terms of age, gender, motives for the suicide, and the methods used for suicide. The increase of suicide rate was in temporal relationship to the reported suicide cases and was not compensated by a decrease in later time, i.e., these suicides were committed earlier than planned. Also very impressive are studies that demonstrate a decrease in suicide rates not only by changing the information towards being more informative and being less sensational, but also by reducing information about suicide (Phillips and Lesyna 1995). One example is a study of railway suicides reported in newspapers, which could be reduced by cooperation with local newspapers in reducing and confining to nonsensational reports on this matter (Sonneck et al. 1994).

In a broader sense, traditions in different cultures and countries concerning either permissive or restrictive attitudes towards suicide can also be regarded as a form of model learning. These attitudes may be transferred over generations such as in Hungary where an extremely high suicide rate can be observed throughout an entire century.

6.4 *Biological Aspects*

During the nineteenth century, Esquirol (1838) was the first to observe that suicidal behavior ran in families and, therefore, postulated that suicidal behavior was a kind of mental illness. Family, twin, and adoption studies gave hints about suicidal behavior having a genetic component.

There is a familial burden due to suicide and suicide attempts (of suicide victims or suicide attempters as compared with subjects without suicide or suicide attempts). Separation of nature (heritage) and nurture (rearing) is not possible within family studies, but is within twin and adoption studies. In twin studies, identical (identical genes) and fraternal twins (half of the genes are nonidentical as in sisters and brothers) are compared with regards to suicide or suicide attempts. The respective studies demonstrate a significantly higher frequency of suicide and suicide attempts in identical as compared with fraternal twins (Roy et al. 1991). In adoption studies, identical twins have been compared with each other, whereby one twin has

been raised with his biological parents, while the other twin has been adopted right away after birth. Significantly fewer twins raised by their biological parents developed suicidal behavior or committed suicide than twins who had been adopted following birth. Furthermore, within biological families, a significantly higher rate of suicides or suicide attempts was shown than within adoptive families of the respective identical twin. All these studies also demonstrate that suicidal behavior has to be seen apart from a specific psychiatric diagnosis, especially apart from a depressive illness. On the other hand, the distinctive feature that might be inherited may not be suicidal behavior as such but rather personality traits such as lack of impulse control or aggressiveness (Wender et al. 1986). Postmortem studies of suicide victims and in vivo studies (investigation of cerebrospinal fluid and the platelets in the blood of suicidal persons concerning molecular genetic, biochemical and physiological parameters) support the so-called serotonin hypotheses of suicidal behavior: impulsiveness, aggression directed inwards or outwards, and violent suicidal behavior are correlated with a dysfunction of the serotonergic transmitter system in the human brain (Mann 1998).

7. *Prevention*

Prevention can be divided into primary, secondary, and tertiary prevention. With regard to primary prevention the main focus implies a reduction of tendencies within the society that support suicidal tendencies. A verdict of suicidal behavior by means of social ostracizing or punishment has not been successful in reducing suicide rates in medieval and modern times (Minois 1995). However, primary prevention also concerns the prevention of suicidal behavior by ameliorating life conditions in general and socioeconomical conditions in particular as well as effective treatment of psychiatric disorders. It also refers to a support of tolerance for individual life styles. Furthermore, the restricted approach to frequently used methods, such as guns for shooting, can be helpful. Primary prevention implies also an amelioration of power of resistance against suicidal tendencies in the individual. Education of lay people and professionals in detecting and treating suicidal persons could be the most effective strategies. Secondary and tertiary prevention provided by fellow citizens (e.g., Samaritans) or medical or psychological professional personnel mean prevention of relapses in suicidal behavior or rehabilitation of patients with suicidal behavior. Empirical studies proving the effect of the above mentioned preventive strategies are still lacking. There is now an emerging trend to develop national suicide prevention programs in several western and northern European countries (Silverman and Maris 1995).

8. *Therapy*

Therapy has to be divided into crisis intervention, with actual psychiatric-psychotherapeutic intervention, into pharmacotherapy and psychotherapy. Crisis intervention primarily means prevention of an imminent danger of suicide. Crisis intervention should provide care and protection to the individual, i.e., protection of life, since suicide attempts often require medical care such as admission to an internal intensive care unit for detoxification or a surgical intervention. Crisis intervention also includes immediate evaluation of still existing suicidal tendencies of the patient in order to make decisions for further preventive measures: admission to a closed psychiatric ward for relapse prevention or crisis intervention on an inpatient or outpatient basis. Crisis intervention also includes acute psychotherapy and pharmacotherapy of the patient. The most important determinant of the psychotherapy is hereby the development of a supportive relationship between the suicidal person and the therapist. The greater the time lapse between the suicide attempt and help offered, the more often suicidal people are inclined to disregard this help. A supportive relationship means that the therapist is accepting the patient's 'cry for help' and is making a commitment trying to understand the patient, however, is not willing to accept the patient's decision to end their life as a form of reasonable problem solving. They clarify the triggering situation that has led to the suicide attempt. Together with the patient, the therapist tries to develop other strategies to resolve seemingly insoluble conflicts in the patient's life. Also, an important determinant of a successful intervention is the involvement of significant others of the patient as well as the social environment in general. The acute pharmacological intervention means primarily sedation of the patient in order to keep him or her from further suicidal behavior.

Psychotherapy means treatment of psychopathological states such as depression, dysfunctional behaviors like impulsiveness, or pathological personality traits such as low self-esteem (Bronisch 1999). It includes the classical types of psychotherapy, i.e., psychodynamic approaches and cognitive-behavioral approaches. So far, empirical studies could only demonstrate efficacy for some groups of patients, such as patients with Borderline Personality Disorders by cognitive-behavioral therapies (Linehan 1997).

Pharmacotherapy of suicidal behavior primarily is focused on the underlying psychiatric illness of the patient such as schizophrenia (antipsychotics), depressive disorders (antidepressants), anxiety disorders (anxiolytics), and addiction (drugs against withdrawal symptoms). However, during the 1990s drugs specifically targeting the typical characteristics of suicide attempters such as impulsiveness and aggressiveness seem to have been successful in reducing suicide and suicide attempts: lithium (Ahrens 1997) and selective

serotonin reuptake inhibitors (Montgomery 1997). (See *Antidepressant Drugs*.)

9. Concluding Remarks

Suicide and suicide attempt are a behavior confined to human beings. They are rooted fundamentally in the human existence and have as a prerequisite self-reflecting properties of the individual, i.e., conscious acting with the consequence or the attempt to extinguish the own existence.

The attempts to describe suicide and suicide attempt based on a single cause in only one dimension has been tried by Esquirol (1838) in biology, Durkheim (1897) in sociology, or Freud (1917) in psychology and seems to be inappropriate. Suicide and suicide attempt are multidetermined acts (Roy 2000). However, the weighting of different aspects of suicide and suicide attempt seems to be rather difficult to explain, so that in the following, only a synopsis of the empirical findings will be given.

Suicide is lacking in animals and even in nonhuman primates. However, it has already been found in Neolithic societies as well as later on in the old Greek and Latin societies. Precise empirical data about suicide in different countries and cultures has been lacking up to the nineteenth century so that a description of the development of suicide from the antiquity via the medieval to modern times is impossible. However, there is a considerable increase of suicides in societies with a high degree of secularization, industrialization, and individualization starting in the middle of the nineteenth century. For individual countries, there is often a stability of suicide rates on a high or low level depending at least in Europe on the degree of industrialization, religious affiliation and climate: northern countries with a high degree of industrialization, Protestant religious affiliation and rather few sunny days have considerably higher suicide rates than southern countries with a low degree of industrialization, Catholic religious affiliation, and many sunny days have lower suicide rates. The differences could best be explained by human bonding as a protective factor, whereas climate could influence biological rhythms, hence, triggering suicidal behavior.

In regard to sex differences and suicidal behavior, very stable differences exist between males and females with a male preponderance in suicide and female preponderance in suicide attempts. The reasons for these sex differences, which seem to be stable across countries, cultures, and times, have not been well investigated and need further empirical research. There is no doubt that psychological motives play not only an important role as triggers for suicidal behavior but also are important determinants for the development of suicidal behavior. The most important motive for suicidal behavior is certainly the impending loss of a human relationship, even if the wish to die outweighs the wish to survive in the suicidal act, i.e., in the seriousness of the suicide attempt. Besides that, the turning of aggressive tendencies towards the own character and narcissistic vulnerability are personality traits predisposing to suicidal behavior. From the cognitive-behavioral point of view, classical conditioning, operant conditioning, and model learning are basic learning principles, which also can be transferred to behaviors such as suicide and suicide attempt. This way, model learning is not only important for perpetuating suicidal behavior in families and peer groups but also in societies with traditional high (Hungary) suicide rates (so called imitation hypothesis).

In certain families, a genetic burden of suicidal behavior could be confirmed by twin and adoption studies. This genetic burden is often connected to psychiatric illnesses such as depressive disorder, addiction, anxiety disorder, schizophrenia, and personality disorder. Biological studies (postmortem brain and peripheral biochemical studies) could demonstrate a dysfunction of the serotonergic system in suicide victims and suicide attempters, which in psychopathological terms is correlated to a loss of impulse control, aggressiveness, affective instability, and violent suicide attempt.

As empirical studies could confirm, suicidal behavior is not a 'normal' behavior, but is rooted in psychological and biological abnormalities of the respondent with the corollary of prevention and therapy. However, empirical data about successful prevention of suicidal behavior is still lacking and successful therapeutic interventions only exist for subgroups of suicide attempters such as depressives with a successful treatment with lithium and serotonin reuptake inhibitors as well as cognitive-behavioral therapy for patients with a borderline personality disorder. Life as a precious good, desperation, and suffering of the respondents as well as suffering of the relatives should bring the focus on suicide more into the scientific scope of view as other life threatening illnesses already have been for a long time.

See also: Bowlby, John (1907–90); Depression; Depression, Clinical Psychology of; Depression, Hopelessness, Optimism, and Health; Durkheim, Emile (1858–1917); Freud, Sigmund (1856–1939); Suicide: Ethical Aspects; Suicide, Sociology of

Bibliography

Ahrens B 1997 Mortality studies and the effectiveness of drugs in long-term treatment. *Pharmacopsychiatry* **30**(suppl.): 57–61

Beck A T 1967 *Depression: Clinical, Experimental and Theoretical Aspects*. Harper and Row, New York, Evanston, IL, London

Bowlby J 1977a The making and breaking of affectional bonds I. *British Journal of Psychiatry* **130**: 201–210
Bowlby J 1977b The making and breaking of affectional bonds II. *British Journal of Psychiatry* **130**: 421–31
Bronisch T 1995 *Der Suizid. Ursachen—Warnsignale—Prävention*. Beck, Munich, Germany
Bronisch T 1999 Suizidalität. In: Möller H J, Laux G, Kapfhammer H P (eds.) *Psychiatrie und Psychotherapie*. Springer, Berlin, Heidelberg, New York, pp. 1673–91
Diekstra R F W 1996 The epidemiology of suicide and parasuicide. *Archives of Suicide Research* **2**: 1–29
Durkheim E 1897 *Le Suicide*: *Etude de Sociologie* (trans. Spaulding J A, Simpson G 1951). Free Press, Glencoe, IL
Esquirol J E D 1838 *Des maladies mentales: Considerées sous les rapports médicale, hygienique et medicolegale; Accompagneé des 27 Planches gravées*. Bruxelles. TI & II, Paris
Freud S 1917 Trauer und Melancholie. In: Johnes E (ed.) *Collected Papers*. Hogarth Press, London, Vol. 4, pp. 137–56
Henseler H 1974/1984 *Narzißtische Krisen. Zur Psychodynamik des Selbstmords*. Westdeutscher Verlag, Obladen, Germany (1984), Reinbek bei Hamburg: Rowohlt (1974)
van Hooff A J L 1990 *From Autothanasia to Suicide: Self-killing in Classical Antiquity*. Routledge, London
Klerman G L 1987 Clinical epidemiology of suicide. *Journal of Clinical Psychiatry* **48**(Suppl.): 33–8
Kreitman N 1977 *Parasuicide*. Wiley, New York
Mann J J 1998 The neurobiology of suicide. *Nature Medicine* **4**: 25–30
Minois G 1995 *Histoire du suicide*. Fayard, Paris
Montgomery S A 1997 Suicide and antidepressants. In: Stoff D M, Mann J J (eds.) *The Neurobiology of Suicide from the Bench to the Clinic*. Annals of the New York Academy of Sciences, New York, pp. 329–38
Pfeffer C R 1986 *The Suicidal Child*. Guilford Press, New York, London
Phillips D P, Lesyna K 1995 Suicide and the media research and policy implications. In: Diekstra R F W, Gulbinat W, Kienhorst I, De Leo D (eds.) *Preventive Strategies on Suicide*. Brill, Leiden, Germany, pp. 231–61
Platt S, Bille-Brahe U, Kerkhof A, Schmidtke A, Bjerke T, Crepet P, De Leo D, Haring C, Lonnqvist J, Michel K, Philippe A, Pommereau X, Querejeta I, Salander-Renberg E, Temesvary B, Wasserman D, Sampaio Faria J 1992 Parasuicide in Europe: The WHO/EURO Multicentre Study on Parasuicide I. Introduction and preliminary analysis for 1989. *Acta Psychiatrica Scandinavica* **85**: 97–104
Retterstøl N 1992 Suicide in the Nordic countries. *Psychopathology* **25**: 254–65
Roy A 2000 Suicide. In: Sadock B J, Sadock V A (eds.) *Kaplan & Sadock's Comprehensive Textbook of Psychiatry II*. Lippincott Williams & Wilkins, Philadelphia, Baltimore, New York, London, Buenos Aires, Hong Kong, Sydney, Tokyo, pp. 2031–40
Roy A, Segal N L, Centerwall B S, Robinette C D 1991 Suicide in twins. *Archives of General Psychiatry* **48**: 29–32
Schmidtke A 1988 *Verhaltenstheoretisches Erklarungsmodell suizidalen Verhaltens*. S. Roderer, Regensburg, Germany
Silverman M M, Maris R W 1995 *Suicide Prevention Toward the Year* 2000. Guilford Press, New York, London
Stengel E 1964 *Suicide and Attempted Suicide*. Penguin, UK
Wedler H 1992 Some remarks on the frequency of suicides in Germany. In: Crepet P, Ferrari G, Platt S, Bellini M (eds.) *Suicidal Behavior in Europe. Recent Research Findings*. Libbey, Rome, pp. 79–83
Wender P H, Kety S S, Rosenthal D, Schulsinger F, Ortmann J, Lunde I 1986 Psychiatric disorders in the biological and adoptive families of adopted individuals with affective disorders. *Archives of General Psychiatry* **43**: 923–29
World Health Organization 1990 *Consultation on Strategies for Reducing Suicidal Behaviors in European Region: Summary Report*. World Health Organization, Geneva

T. Bronisch

Suicide: Ethical Aspects

In 1917, perhaps still troubled by the suicides of two of his brothers and of a close friend, Ludwig Wittgenstein wrote:

> If suicide is allowed then everything is allowed.
> If anything is not allowed, then suicide is not allowed.
> This throws a light on the nature of ethics, for suicide is, so to speak, the elementary sin…

In contrast, in 1961 the Nobel physicist Percy Bridgman, then almost 80 years old and suffering from terminal cancer, shot himself, leaving a final note:

> It isn't decent for society to make a man do this thing himself.
> Probably this is the last day I will be able to do it myself.

These views frame the scope of the contemporary debates over the ethics of suicide. At one extreme, some hold the view that suicide is, as Wittgenstein put it (though he questioned even this view), 'the elementary sin,' the act most clearly 'not allowed.' At the opposite extreme, others hold that control over one's own life is a matter of right, and ending that life something for which, at least in circumstances like painful terminal illness, a person ought to receive assistance and social support.

These debates are both ancient and new. In the West, prior to the beginning of the twentieth century, suicide had often been viewed as a moral question; at the beginning of the twentieth century, however, 'scientific' views of suicide as the product of social forces or of individual psychopathology worked to remove suicide from ethical scrutiny. Suicidality came to be seen largely as a clinical matter for prevention and treatment. Not until the emergence of the right-to-die movement in the latter part of the twentieth century did ethical issues about physician-assisted suicide and hence about suicide itself begin to re-emerge as a matter for philosophical reflection and for public debate.

1. Conceptual Issues

In popular discourse, the term 'suicide' carries extremely negative connotations. However, there is little

agreement on a formal definition. Some authors count all cases of voluntary, intentional self-killing as suicide. Others include only cases in which the individual's primary intention is to end his or her life. Still others recognize that much of what is usually termed 'suicide' is neither wholly voluntary nor involves a genuine intention to die, including many of the cases usually labeled 'suicide'—especially those associated with depression or other mental illness. Many writers exclude cases of self-inflicted death which, while voluntary and intentional, appear aimed to benefit others or to serve some purpose or principle: for instance, Socrates' drinking the hemlock, Captain Oates' walking out into the Antarctic blizzard to allow his fellow explorers to continue without him, or the self-immolation of war protesters. These cases are usually not called 'suicide' but 'self-sacrifice,' 'martyrdom,' 'heroism,' or other terms with strongly positive connotations. However, attempts to differentiate these positive cases often seem to reflect moral judgments, not genuine conceptual differences.

Nonwestern traditional cultures offer many additional conceptual puzzles. Is the label 'suicide' appropriate for wives and royal retainers who allow themselves to be buried alive at the death of the king? For widows who, responding to strong social expectation, commit *sati* by throwing themselves on their husband's funeral pyre? For *jihad* or religiously motivated holy-war acts certain to cause one's own death? For kamikaze missions undertaken by volunteers who know they will not return but who are seen as making a supremely heroic sacrifice for the good of the nation? In all cultures, cases of death from self-caused accident, self-neglect, chronic self-destructive behavior, victim-precipitated homicide, high-risk adventure, and self-administered euthanasia—which all share many features with suicide but are not usually termed such—cause still further conceptual difficulty. Consequently, some authors claim that it is not possible to reach a rigorous formal definition of suicide, and that only a criterial or operational approach to characterizing uses of the term can be successful. Nevertheless, conceptual issues surrounding the definition of suicide are of considerable practical importance in policy formation, as for instance in coroners' practices in identifying causes of death, psychiatric protocols, insurance disclaimers, religious prohibitions, and laws either prohibiting or permitting aiding suicide in situations such as terminal illness.

2. *Traditional and Contemporary Ethical Arguments Concerning Suicide*

Because recent scientific views of suicide have tended to presuppose that suicide is not a voluntary act, much of the most significant argumentation concerning the moral status of suicide is to be found in the historical tradition. Traditional discussions of suicide employ a wide range of types of arguments, including principle-based, consequentialist, and virtue-based arguments. In each of these categories, some arguments are directed toward the conclusion that suicide is immoral and/or should be prohibited; others lead to the conclusion that suicide is morally permissible and/or should be permitted; a few lead to the conclusion that suicide is morally obligatory in certain sorts of circumstances.

2.1 *Principle-based Arguments*

Many thinkers in Western philosophical and religious traditions which employ principle-based, duty-based, or other deontological forms of moral reasoning have argued that suicide is intrinsically wrong, and, hence, that it is prohibited. Conversely, others argue that because suicide is prohibited (by God), it is known to be intrinsically wrong. For example, traditional Judaism cites the Biblical passage at *Genesis* 9:5, 'For your lifeblood I will surely require a reckoning,' as the basis of the principle that suicide is wrong; the only exceptions involve self-killing to avoid spiritual defilement, a practice known as *Kiddush Hashem* (as in the mass suicide at Masada), and temporary insanity. In classical Greece, Socrates, as portrayed in Plato's *Phaedo*, cites the secret, 'whispered' Orphic principle that man is a 'prisoner' or 'possession' of the gods and ought not run away; thus suicide would violate the wishes of the gods. Early Christian thinkers implied or asserted that suicide is wrong *per se*, since, as Augustine argued, it is to reject God's gift of life and runs counter to the will of God. In exceptional cases like those of Samson pulling down the temple on the Philistines—and himself—, and Saul falling on his sword, Augustine insists, suicide must have been commanded by God, but such cases are very, very rare and do not excuse ordinary persons.

In the high Middle Ages, Thomas Aquinas definitively established Christianity's opposition to suicide with a set of five arguments. Central among these is the principled claim that suicide is unnatural: 'everything naturally seeks to keep itself in being.' From this principle, Thomas drew the conclusion that '… suicide is, therefore, always a mortal sin in so far as it stultifies the law of nature …' Still later, philosopher Immanuel Kant argued that suicide can be shown to be intrinsically wrong by application of the Categorical Imperative: one cannot will without contradiction that it be universal law. Many other moralists have held that, as a matter of basic ethical principle, suicide is intrinsically wrong.

At the same time, however, other thinkers employing deontological or principle-based reasoning have argued that suicide can be morally permissible or even

praiseworthy. Typically, these thinkers assert that the principle of autonomy is fundamental, claiming that unless the person is not acting voluntarily or does not have adequate information, or unless serious harm to others will follow, a person's choice of suicide is to be respected. Defenses of suicide by Seneca, by Rousseau's character St. Preux in the novel *Heloise*, by Schopenhauer, and by Mme. de Stael (though she later argued against suicide) are of this form: they assert that as a matter of principle, suicide is not intrinsically wrong and that suicide can be a morally acceptable, sometimes praiseworthy thing to do.

Although the language of moral rights is not available to the earlier thinkers, some have elevated the principle of autonomy into the basis of a fundamental right to suicide. For example, Schopenhauer and Nietzsche see suicide as a matter of the basic right of the individual to choose whether to continue to exist in the world. As Nietzsche put it, 'suicide is man's right and privilege.'

2.2 *Consequentialist Arguments*

Other thinkers have argued that suicide is morally permissible in certain situations, though not in others. Typically, these arguments are consequentialist or utilitarian in character—outcome-oriented, so to speak—assessing the permissibility of suicide in terms of its impact on the suicidal person, on family members, on friends and associates, and on the larger society. Aristotle, for example, says that suicide is wrong because it 'treats the state unjustly'; he is concerned about its impact on society. Christian and Islamic thinkers often argue not only that suicide is wrong in itself but that it will mean damnation for the victim. Blackstone argued that the suicide is an 'offense against the king,' who has an interest in the preservation of all his subjects. Samuel Pepys, recounting in his diary his efforts to intervene on behalf of the widow of a man who has just drowned himself, points to the disastrous impact suicide could have on family members, inasmuch as the English legal system made a suicide's property forfeit to the Crown.

Using similarly consequentialist arguments, many contemporary authors also point to the consequences suicide can have for family members (especially children) and other close associates. They particularly emphasize the sense of guilt and shame that may be aggravated in survivors. Suicide, on these consequentialist arguments, damages its victim, damages family members, and can damage the society of which they are part.

However, not all consequentialist arguments speak against the morality of suicide. Greek and Roman Stoics, defending suicide in principle, also point to the beneficial consequences suicides like that of Cato might have in protecting the freedoms of society. Both traditional and contemporary authors acknowledge that the suicide of an aggressive or abusive person may have positive consequences for others by removing a threat, that the suicide of a person suffering an extended, expensive illness may relieve burdens on an overwrought family, and that suicides of self-sacrifice and principle can have substantial benefits for society. Because these arguments appear to favor or even encourage suicide in many sorts of circumstances, they are currently viewed as extremely controversial. They are rarely openly discussed, though they play an influential role in the background of concerns about 'slippery slope' pressures on vulnerable persons.

2.3 *Virtue-based Arguments*

Still other thinkers treat suicide as an index of character. For example, Aristotle asserts that suicide is 'cowardly'—that is, that it is evidence of a defect of character, a failure to act virtuously by exhibiting insufficient courage in facing a difficult situation. Many later thinkers, including Augustine and Thomas Aquinas, have adopted the view that suicide is 'cowardly.' These and other Christian writers also celebrate the willing acceptance of suffering, particularly in emulating the sacrifice made by Christ. As Mme. de Stael put it in her second essay on suicide opposing the libertarian stance of her first, 'Suffering is a blessing…it is a privilege to be able to suffer.'

In contrast, Greek and Roman Stoics favoring suicide, as well as many later authors, also employ virtue criteria, but argue to the opposite conclusion. The Stoics held that real cowardice held in fearing death, and that committing suicide itself requires courage. Examples cited in this tradition include the heroic suicides of Cato, Lucretia, and Seneca, respectively involving commitment to an ideal of political liberty, sexual virtue, and personal courage in the face of an untrustworthy emperor. For some virtue theorists, suicide can involve other virtues as well: fortitude, loyalty, civic responsibility, and many others. Thus, on a virtue-based analysis, too, arguments may be made both against and for the ethical permissibility of suicide.

3. *Social Correlates, Biochemical Findings, and Markers of Suicide Risk*

The question of whether suicide is wrong—whether in principle, or because of its consequences, or as a fault of character—seems to presuppose that the act is at least to some degree under the agent's voluntary control. Traditional accounts of the morality of suicide, both those which allow it in some or all circumstances—including Plato, the Stoics, Hume, Nietzsche, and others—and those which oppose it in all or virtually all circumstances—Aristotle, Augustine, Thomas Aquinas, Kant, and others—agree in

assuming that the act is sufficiently voluntary that the person whose death it will be can appropriately be blamed (or praised) for so acting—after all, he or she could have refrained from suicide and kept on living.

Scientific views of suicide, introduced by Durkheim and Freud at the beginning of the twentieth century, have challenged this assumption. Durkheim's anomie theory held that suicide is a function of the degree of integration of the social body: suicide can be altruistic, where suicide practices are highly institutionalized; it can be egoistic, where individuals do not respond to social expectations and regulations; or it can be anomic, where society does not provide adequate regulation of its members, as, Durkheim believed, in modern industrial society. In contast, Freud insisted that suicide was the product of mental illness. For much of the twentieth century, whether suicide was understood as the product of social forces or as the consequence of underlying depression or other mental disorder, in either case it has been seen as largely beyond the voluntary control of its victim.

Suicide prevention, pioneered by Edwin Shneidman in the mid-twentieth century and continued by many later suicidologists, works in part to identify risk factors associated with elevated rates of suicide: risk is greater for males (especially older white males), for alcohol abusers, for those with low religiosity, for those with poor or rigid coping skills, for those less willing to seek professional help and more likely to ignore the warning signs of suicide, for those with poorer support systems, for those with access to more lethal technology such as guns, and for those for whom failure is more obvious in the primary adult role (defined for males as economic success; for women, as success in relationships). Also at heightened risk are teens who listen to heavy-metal music and white males who listen to country music, though these factors are not causes but proxies for other variables, like alcohol or drug use or divorce. Copycat effects following celebrity suicides, holiday dips and post-holiday peaks, and seasonal variations in suicide have all been explored in the search for explanations, as has the concept of suicidal 'careers' as a form of coping mechanism.

At the same time, new findings in psychiatry, biochemistry, epidemiology, and genetics are beginning to erode assumptions about voluntariness as they reveal biological correlates of suicidal behavior. Depression, it has long been known, is strongly associated with suicide (some studies suggest that it is present in as many as 80–90 percent of cases, though in other studies the rate is lower). Levels of 5-HT, or serotonin, a neurotransmitter, and its metabolite 5-HIAA are decreased, as measured in CNS spinal fluid. Fenfluramine used as a challenge agent evokes an impaired response in people who have made serious suicide attempts. Prolactin release is stimulated by serotonin; it too is decreased in suicidal persons. There is some evidence of heritable genetic patterns in families with multiple suicides: adoption studies show that adoptees with suicide in their biological families are at greater risk for suicide than other adoptees, and there is some evidence of significant association between genetic variation in the gene for tryptophan hydroxylase (TPH) and violent suicide attempts. It may even be the case that depression and suicidality are associated with second-trimester exposure during pregnancy to winter flu.

Such findings suggest that suicide, as well as suicidal ideation and suicidal and parasuicidal behavior, along with aggressive behavior generally, may be to some degree biologically determined, either directly or in conjunction with social and environmental factors. If so, suicidality would thus be less a focus for moral blame or praise than for clinical treatment and concern. It may even be possible to develop a biological test—for example, a blood test, a saliva assay, or a urine analysis—that would screen for CNS neurotransmitter abnormalities associated with suicidal behavior. While work to date establishes correlations rather than causes, it nevertheless points to psychological and situational factors that may play a causal role in suicide.

Some theorists have assumed that such research will eclipse ethical issues in suicide altogether. However, this research faces two challenges: first, the definitional issue of who counts as a 'suicide' (as distinct from martyr, hero, self-sacrificer, or terminally ill patient seeking an easier death), and second, the prudential issue of whether suicide can ever serve the interests of the person who commits it. Definitional challenges to research programs are numerous: should research on completed 'suicides' include among its cases, for example, self-sacrificing parents who put themselves in harm's way to protect their children, or social protesters who immolate themselves for political causes, or the sea captain who elects, as tradition has required him, to go down with the ship? The prudential challenge asks whether there can be 'good reason' for suicide in some cases, such as painful terminal illness or extreme old age, and hence whether research is misguided if it focuses primarily on negatively valued cases; that is, those called 'suicides' in ordinary parlance.

Some theorists suggest that there are two basic groups of suicides: on the one hand, people with neurotransmitter abnormalities or other biological defects who suffer from mental illness, depression, or extreme emotional stress, people who cannot consider their best alternatives, and on the other hand biologically normal people who can make such considerations but who find themselves caught in situations in which suicide appears to be the most reasonable choice, a way out of their difficulties. These groups are sometimes distinguished as 'irrational' or mentally ill, and 'rational' suicides. A clear differentiation is not adequately supported by current findings, primarily because research on 'suicide' typically includes just

those attempts and completed cases conventionally labeled suicide, overlooking cases not usually identified as suicides where the motivation is viewed as understandable or heroic, as in, for example, cases of political protest or of hastened dying in terminal illness. By some estimates, 95 percent of suicide cases are 'irrational' and five percent 'rational,' though this estimate is itself challenged by conceptual concerns about what counts as 'suicide' in the first place.

Nevertheless, this distinction influences much of contemporary discussion about ethical issues in suicide. 'Irrational' suicides are assumed not to pose moral issues concerning the agent, since they are assumed not to be voluntary in a robust sense, though of course second parties like family members and physicians may face moral questions about whether and how to intervene. 'Rational' suicides, on the other hand, are assumed to be appropriately considered in ethical terms: these are people who make a choice to end their own lives, and the moral issue is whether it can ever be right to choose as they do.

4. *The Contemporary Debate over Physician-assisted Suicide*

Beginning in the latter part of the twentieth century, the so-called 'right-to-die' movement began to challenge life-prolongation practices in the treatment of the terminally ill. Patients, it argued, should have the right to control the care given them as they are dying, whether by refusing treatment like respirators or antibiotics or by discontinuing treatment already in progress—even if this means they will die. By giving a central role to patients' autonomous decisions about medical care at the end of life, the right-to-die movement also raised the issue of whether the terminally ill patient may deliberately and directly end his or her own life—and receive help from his or her physician in doing so. As of the year 2001, physician-assisted suicide had become legal only in the Netherlands and the US state of Oregon, and had briefly been so in the Northwest Territory of Australia, but the issue was being widely debated in many countries, especially in the USA, Canada, the UK, Belgium, Australia, Colombia, and several Scandinavian countries.

Four principal ethical arguments, all to be found in earlier, historical discussions of the ethics of suicide, contribute to this broadly international public debate over whether physician-assisted suicide (and, in some countries, voluntary active euthanasia) should be regarded as morally acceptable and legally permitted. The two principal arguments on the pro side are (a) the argument from autonomy, that the autonomous choice of a competent person is to be respected, and (b) the argument from mercy, an argument for freedom from suffering, that the dying patient has the right to try to avoid pain and suffering as much as possible, including by means of a deliberately undertaken earlier death if that is the only effective way to avoid them. At least on some views, both conditions are necessary and neither sufficient in itself: a person must both choose autonomously and have no realistic alternative in seeking freedom from suffering in the process of dying to underwrite an adequate case for physician-assisted suicide. Neither voluntary choice alone in the absence of terminal suffering nor terminal suffering without genuine choice would be sufficient to make such a case.

On the opposing side, two principal arguments against moral recognition or legal acceptance of physician-assisted suicide also inform the current debates, (a) the argument from the intrinsic wrongness of killing—that killing, and hence self-killing, is in itself wrong, and (b), that permitting assisted suicide would risk the 'slippery slope': it would invite pressures on patients by overwrought or greedy families, unscrupulous physicians, or cost-conscious institutions, all of which might force a vulnerable patient to 'choose' suicide when they did not really wish to die.

Each of these arguments is open to rebuttal. For instance, the claim that killing is not morally permissible can be countered by the objection that killing in self-defense, in war, and in capital punishment are regarded as permissible. The claim that death may be the only way to avoid pain and suffering is countered with the possibility of deep sedation. Such counter-arguments are themselves open to rebuttal: for example, that it is only the killing of innocents that is prohibited, or that deep, irreversible sedation is tantamount to causing death, in that it permanently ends the possibility of human experience. These counter-counter-arguments are in turn open to still further reply. In practice, the disputes over physician-assisted suicide are hugely extended and extremely complex, simmering in many disciplinary areas—philosophy, law, medicine, psychology, sociology, political theory, theology, social anthropology, economics, and public policy—in many countries around the world.

These disputes have had a focusing effect, redirecting the larger normative debate about suicide, quiescent since Durkheim, Freud, and the beginning of the twentieth century, toward a specific type of case, terminal illness. Contemporary disputes about the ethical issues in suicide no longer address the larger range of cases in which suicide had been considered permissible or impermissible by various cultures at various times: in shame or dishonor, in chronic illness, in old age, in extreme poverty, for blindness or other disability, on being widowed, for social protest or the defense of principle or religious belief, to avoid degrading labor or coercion in unacceptable activities, and many other reasons. The debate over right-to-die issues and physician-assisted dying that has emerged in the latter part of the twentieth century is indeed a debate about the ethics of suicide, but it is a highly

specific, highly focused debate, raised virtually exclusively with regard to physical suffering in the context of terminal illness; it does not yet address whether and when suicide is morally impermissible, permissible, or obligatory in any of a much wider range of situations.

While the right-to-die debates over the ethics of physician-assisted suicide in terminal illness continue to ferment, primarily in the developed nations, work is re-intensifying in suicide prevention, focused especially on teens and on some groups of persons at higher identified risk, and on improvement in the care of dying patients, particularly on more effective pain control. Some believe that these measures will be able to reduce both the incidence of suicide and the call for physician-assisted suicide in the future.

5. *The Future*

In 1975, Henry Van Dusen, former president of Union Theological Seminary, and his wife Elizabeth—both leaders in American theological life—committed suicide together by taking overdoses. Mr. Van Dusen was 77 and had had a disabling stroke five years earlier; Mrs. Van Dusen was 80, blind, and suffering from serious arthritis. They left a letter saying that 'We still feel this is the best way and the right way to go,' predicting that it would become more frequent and more publicly accepted. It is not yet possible to say whether they were right; some observers argue that the ethical issues in suicide will recede again out of public view as modern medicine becomes more and more capable of alleviating pain and suffering now feared by the dying, while others think the ethical issues will become more and more broadly explored, as patients seek greater and greater autonomy in the matter of their own dying. And some think that these ethical issues will be central in the moral life of many societies for the foreseeable future, as life expectancies increase and the likelihood of dying from late-life degenerative disease grows. If so, ethical scrutiny in the matter of suicide will turn to two sorts of issues: the morality of approaching terminal illness and unavoidable death by 'designing' one's own dying, and the risks of social encouragement and pressures that could develop as it becomes a normal, socially tolerated, expected thing to do. Whether such trends will in fact develop is impossible to predict, but it is possible to predict what ethical issues such trends would raise.

See also: Consequentialism Including Utilitarianism; Suicide; Suicide, Sociology of; Virtue Ethics

Bibliography

Améry J 1976 1999 *On Suicide: A Discourse on Voluntary Death*. Indiana University Press, Bloomington and Indianapolis, IN

Battin M P 1982, 1995 *Ethical Issues in Suicide*. Prentice-Hall, New York

Battin M P, Rhodes R, Silvers A (eds.) 1998 *Physician-Assisted Suicide: Expanding the Debate*. Routledge, New York

Beauchamp T L (ed.) 1996 *Intending Death. The Ethics of Assisted Suicide and Euthanasia*. Prentice-Hall, Upper Saddle River, NJ

Bridgman P 1961 letter in *Bulletin of the Atomic Scientists*, quoted by Max Delbrück, education for suicide, interview in *Prism*, a publication of the American Medical Association 2 (1974), p. 20

Brody B A (ed.) 1989 *Suicide and Euthanasia: Historical and Contemporary Themes*. Kluwer, Dordrecht/Boston/London

Durkheim E 1897 1951 *Suicide: A Study in Sociology*. Free Press, New York

Emanuel L L (ed.) 1998 *Regulating How We Die. The Ethical, Medical and Legal Issues Surrounding Physician-Assisted Suicide*. Harvard University Press, Cambridge, MA

Ethics **109**(3) (April 1999) Symposium on Physician-Assisted Suicide

Kaplan K J, Schwartz M B (eds.) 1998 *Jewish Approaches to Suicide, Martyrdom, and Euthanasia*. Jason Aronson, Northvale, NJ and Jerusalem

Larson E J, Amundsen D W 1998 *A Different Death: Euthanasia and the Christian Tradition*. Inter Varsity Press, Downers Grove, IL

Maris R 1981 *Pathways to Suicide*. Johns Hopkins University Press, Baltimore

Maris R W, Berman A L, Silverman M M 2000 *Comprehensive Textbook of Suicidology*. Guilford Press, New York

Minois G 1995 *Histoire du suicide: La société occidentale face a la mort volontaire*. Libraire Artheme Fayard [1999 *History of Suicide. Voluntary Death in Western Culture*. Johns Hopkins University Press, Baltimore]

Murray A 1998 *Suicide in the Middle Ages*. Vol. I: *The Violent against Themselves*. Oxford University Press, Oxford

Stack S 2000 Suicide: A 15-year review of the sociological literature: Part I: Cultural and economic factors. *Suicide and Life-Threatening Behavior* **30**(2): 145–76

Weir R F (ed.) 1997 *Physician-Assisted Suicide*. Indiana University Press, Bloomington and Indianapolis, IN

Werth J L Jr 1996 *Rational Suicide? Implications for Mental Health Professionals*. Taylor and Francis, Washington, DC

Wittgenstein L *Notebooks 1914–16*. von Wright G H, Anscombe G E M (eds.) trans. Anscombe G E M, Blackwell, Oxford, 1961, p. 91e. Entry dated January 10, 1917

Zucker M B (ed.) 1999 *The Right to Die Debate: A Documentary History*. Greenwood Press, Westport, CT

M. P. Battin

Suicide, Sociology of

Although the word itself has been used only since the seventeenth century, the reality that is suicide has been part of human history for a long time. Many suicides occurred in Greece in the fifth century BC and in Rome during the first century AD. The first writings about suicide present it as a religious, moral, or pathological phenomenon, frequently attributed to supernatural causes or to mental illness. Empirical studies appear only at the beginning of the nineteenth century, with still moral overtones. Statistics are analyzed according to an individual or social patho-

logical perspective and the suicide rates are correlated with such variables as the regional economic situation and the homicide or mental illness rates.

1. *Durkheim*

The first truly sociological writing on the subject is *Le Suicide. Étude de Sociologie* by Durkheim (1967) (see *Durkheim, Emile (1858–1917)*). Making use of official statistics, he points out the variations in suicide rates throughout time and space. For him, the rate of suicide constitutes a social fact. Each society has a specific tendency toward suicide. The structure of a society decides, so to speak, its own contingent of voluntary deaths.

Durkheim refutes the theories that explain suicides by extrasocial individual factors. He not only dismisses psychopathic or normal psychological factors, but also cosmic factors and the imitation factor which includes contagion. He constantly distinguishes between the cause of suicide, which is always social, and the motives, which are merely pretexts. Applying the concepts of integration and regulation to religious, domestic, and political societies, Durkheim constructs a typology of suicides. Egotistic suicide results from insufficient social integration and differs from altruistic suicide that is seen in societies too strongly integrated. Anomic suicide occurs in societies where rapid changes are not, or insufficiently, regulated; on the contrary, fatalistic suicide, scarcely touched upon by Durkheim, follows from an excess of social regulation.

Several criticisms have been directed against Durkheim's work. In particular, Baechler (1975) compares his lack of interest in the victim of suicide with his great concern for the social rate of suicide. Baechler observes that, for Durkheim, it is not individuals who take their own life, but society itself that commits suicide through some of its members. Durkheim's analysis of the low suicide rate among women is also severely criticized, along with his lack of development of fatalistic suicide and the confusion he perpetuates between anomy and egotism.

2. *In Durkheim's Wake*

Controversy was not the only reaction to *Le Suicide*. The book awakened a keen interest in the study of suicide, although that interest showed itself only in the 1950s, especially in the USA. Durkheim's research generated a number of studies seeking to replicate, extend, clarify, and even to rectify not only his conceptual analysis but also his empirical results. Still widely taught today in many countries, Durkheim continues to exert a powerful influence on the sociology of suicide.

Using the notions of social status and social roles, Gibbs and Martin (1964) developed a theory around the Durkheimian concept of social integration. When a society engenders a status incompatibility and therefore role conflicts, it has a greater risk of suicide. For Gibbs and Martin, a relationship exists between the suicide rate and the integration of status and roles. It gives rise to their theorem: the suicide rate of a given society varies inversely with the degree of integration of the status and the roles which make up that society.

Lester (1989, 1997) is one of the sociologists who have contributed the most—and he continues to do so—to the operationalization of Durkheim's concepts, especially social integration, and to the widest possible verifying of his empirical conclusions. Although he remains open to psychological and individual perspectives, Lester is part of the Durkheimian movement, as witnessed by his numerous statistical and comparative studies of national and regional suicide rates throughout the world and his insistence on deepening the direct relationship between suicide and the modernization of society. In this regard, Lester recommends greater precision in the criteria for modernization, beyond the vague notions of industrialization, urbanization and secularization. Declining birth and death rates and the increases in the divorce rate and in the gross national product are specific indices that should be taken into account.

3. *Psychosocial Approach*

Notwithstanding the studies that throw light on the social factors of suicide, they do not satisfy several researchers who are interested in studying how such factors influence the persons who commit suicide and what moves them, as individuals, to take their life. The study of suicide thus moves more and more inward, refusing to see the causes of suicide as only social and external. Without neglecting social factors, the study of suicide now considers psychological and individual factors, in both their affective and cognitive dimensions.

3.1 *Affective Dimension*

Halbwachs (1930) gives a different interpretation to the Durkheimian notion of anomic suicide. For him, the increase in the number of suicides should be more correctly attributed to the loneliness felt by those individuals who are faced with an increasingly complex society that makes it difficult for them to establish satisfactory social contacts. The key concept is to be found in the lifestyle of people, since it determines the opportunities for a relationship with others. One therefore finds fewer suicides in rural areas because of a more harmonious lifestyle. Loneliness with its accompanying social isolation, brought on by a complex and conflicting urban society, thus becomes the cause of suicide, which is consequently explained by psychosocial factors.

Henry and Short (1954) suggest another psychosocial explanation of suicide by incorporating the

Freudian concepts of frustration and aggression. An integrated society, in which external constraints increase, has a lower rate of suicide. This is because external constraints have more effect on persons of low status; the subsequent frustration causes their aggressiveness to be turned outwards, with the potential for homicide. Conversely, persons of high status, who impose their demands on others, have no one but themselves to blame if they feel frustrated; their aggressiveness is directed inwards upon themselves, with the potential for suicide. Far from being unanimously accepted, Henry and Short's theory still captured the attention of some researchers. Maris (1981) endorses the idea of social constraints. However, while they are fewer for the person socially isolated, they nevertheless increase the likelihood of suicide.

Giddens (1966) also puts forward a psychosocial approach to suicide. He presents a typology of suicide, using the Durkheimian notions of egotism and anomy with which he associates the Freudian theory of depression. Different kinds of societies spawn various depressed and suicidal personalities. On the one hand, an egotistic society produces introverts who are incapable of relating to others; their suicidal behavior seeks expiation or social reintegration. On the other hand, an anomic society creates personalities dissatisfied with their performance; their suicide aims at surpassing themselves. In any event, for Giddens, all cases of suicide presuppose social isolation, marked by the absence of significant relationships.

3.2 Cognitive Dimension

Sociology also progressively integrates the cognitive dimension that is linked to a concrete suicide. The possible meanings of suicidal behavior gradually come to revolve around the idea that suicide is the solution to one's problems.

Douglas (1967) has left his mark on this approach to the study of suicide by his work on meanings. Influenced by Weber, Douglas states that suicide is a social act that includes subjective meanings constructed from interactions between the victim and his cultural context. He thus advances a social constructionist theory of meanings. Suicide is seen as 'a means of transforming the soul from this world to the other world,' 'a transformation of the substantial self in this world or in the other world,' 'a means of achieving fellow-feeling,' or 'of getting revenge' (Douglas 1967, pp. 284, 300, 304, 310).

Also inspired by Weber, Baechler (1975) attempts to understand what makes human beings put an end to their lives. He defines suicide as any behavior that seeks and finds the solution to an existential problem in the act of putting an end to one's life. His strategic theory includes several types of suicidal behavior: the escapists, who see a way out, the aggressives, who wish to attack someone, the oblates, who sacrifice themselves, and the gamblers, who take a chance with death.

Taylor (1978) also studies the significance of the suicide act for the actor himself. Even as they commit suicide, individuals are rarely seeking death. Aligning himself with Baechler, Taylor suggests two meanings to this behavior. The first is the 'ordeal' character of suicidal action that represents the risk, the game, the betting with death. The second is the aspect of 'rebirth' that the suicide act may take. In risking his life, the individual hopes to escape his problems and reinvest in a new life that will make more sense to him. He is seeking a renewal of life.

4. Psychological and Biological Approaches

Since the 1980s, the increasing influence of the psychological approach on the sociology of suicide has become evident. For Shneidman (1985), suicide is a multidimensional distress of which the main component is always psychological. Suicidal people always suffer from an unbearable psychological pain and from a cognitive constriction, a 'tunnel vision.' Suicide becomes the only solution to their problems. Following Shneidman, Kral (1994) also insists on the psychache felt by the suicidal person. However, for Kral, the direct cause of the fatal act is the idea itself of suicide. Suggested by the victim's culture, it acquires more and more credibility in his eyes and is gradually transformed into an acceptable alternative for getting rid of his psychache. The cultural schema of suicide is progressively internalized into a personal schema.

Some sociologists underline the influence of biology on suicidal behavior. De Catanzaro (1981) adopts a sociobiological standpoint. In his literature review on suicide, Maris (1981) takes note of a biopsychosocial model of suicide. For certain individuals, their life has been so strewn with multiple problems that their capacity for adaptation is severely impaired, possibly owing to their biological condition, e.g., a failure of serotonin or an imbalance of neurotransmitters triggered by social and psychological factors.

5. Differentiation of Suicidal Behaviors

The sociology of suicide progressively differentiates. First interested in *completed suicides*, it examines their many particular forms; multiple studies grouped around gender, ethnic origin, and age group proliferate rapidly. Then, sociologists gradually widen their field to include studies of *noncompleted suicides*.

5.1 Various Forms of Completed Suicides

Instead of a general analysis of the completed suicides, several sociologists isolate particular groups of the

fatal acts. Some of them concentrate on suicides by imitation or contagion, studying the fluctuation of suicides following publicity about certain voluntary deaths. When there is widespread coverage by the press, there follows an increase in the number of suicides. This is known as the 'Werther effect,' referring to suicides committed in 1774 after the publication of a novel by Goethe. Tinged with controversy and a lack of unanimity in the results, a good number of studies have been carried out around that point since the early 1980s (Stack 1987).

Other particular forms of completed suicides are examined according to sex, ethnic origin and age group. Analyses of the sex of persons who committed suicide become more detailed and more discriminating. Invoking his concepts of integration and regulation, Durkheim himself had accounted for a higher suicide rate among men than among women. In most societies, researchers still observe far more male suicides than female, with the particular socialization of each gender often offered as an explanation. Researchers are also looking more and more at links between suicide and homosexual orientation, because of the stigmatization and social exclusion that result from the latter.

Sociologists continue to be deeply interested in the ethic origin of persons who died by their own hand. The suicide rate of immigrants to the USA is analyzed and found similar to the rate in their country of origin. Even the methods for suicide used by Asian Americans differ from those chosen by the majority of Americans. The immigrants' culture of origin is thus a determining factor in their suicide. Intrigued by the low suicide rate of African American men and women in spite of the several risk factors associated with them, some sociologists point out the protective influence of their cultural patterns, particularly of their religious family roles. Black women are protected by their network of social support. As for Chinese women, almost the only ones who take their life more often than their male counterpart, it is because of their traditional culture making them responsible for family problems.

Suicide rates according to age group are carefully and more painstakingly examined. The peaking of suicide in old age is confirmed by Lester (1997) on an international scale. On the other hand, several researchers observe an increasing rate of suicide among the youth in contemporary Western societies. Suicides by imitation and suicide pacts are studied among teenagers who, it is well known, are vulnerable to this phenomenon.

5.2 Other Suicidal Behaviors

It is only in the 1960s that sociologists began to explore the subject of noncompleted suicides. For some researchers, attempted suicide and completed suicide do not show the same characteristics and therefore constitute two distinct social facts. The attempted suicide is presented either as a typical behavior of egotistic societies or as a form of blackmail or a call for help. Often individuals who attempt suicide are hoping more for an intervention that will reconcile them with their social environment than they are seeking death.

Other sociologists consider noncompleted suicides as part of a process of which the final step is a completed suicide. Maris (1981) underlines the importance of attempted suicides in the suicidal process. Identifying 15 predictors of completed suicides during the lifetime of a suicidal person, he puts forward a model of suicide careers. One of the predictors is the greater number of nonfatal suicide attempts, especially among young women. In fact, a higher frequency of attempted suicides and of 'suicidal ideations'— a concept only timidly discussed in sociology—is generally found among girls.

In their review of the literature published between 1963 and 1993, Arensman and Kerkhof (1996) make a distinction between those whose suicide attempts have been rather weak, those whose attempts have been serious, and those who have ended up killing themselves. For these authors, the link between mild attempts and completed suicides is not shown; they refer to those attempts as parasuicides and reserve the expression 'attempted suicide' for severe attempts.

6. Methods Used

In the wake of Durkheim's work, the analysis of the social factors of suicide was mostly done by means of quantitative and statistical studies of the macrosociological type. Along with its focusing on the suicidal persons themselves, sociology gradually opened itself to qualitative studies of a more microsociological type that favored sociological individualism (see *Macrosociology–Microsociology*).

6.1 Quantitative Studies

Durkheim's use of statistics has been severely criticized, particularly by Douglas (1967), who reproaches Durkheim for choosing statistics that confirmed his theory, while giving scant attention to those that did not. However, the school of Columbia University, especially via Selvin, praised the work of Durkheim and greatly contributed to reviving the interest for his study of suicide in the USA and in France. In particular, Selvin (1958) pointed out the cleverness and the innovative character of Durkheim's statistical analysis, not merely contented with simple correlations but using multivariate analysis to validate the complex links between suicide and many social factors.

Macrosociological studies of suicide involving official statistics continue to abound, indeed are in the majority. The relative inaccuracy of such statistics has been argued by some authors, but the hesitation about using them is less and less relevant in contemporary advanced societies where they enjoy greater credibility (Lester 1997). More and more sophisticated statistical tests allow for rigorous longitudinal analyses, as well as precise verification of the correlations between suicide and numbers of variables: gender, marital status, hereditary and biological characteristics, family history, profession, religious affiliation, ethnic origin, immigration, use of drugs or alcohol, emotional problems, previous suicide attempts, economic, geographical and seasonal context of suicides, and means used to commit suicide.

6.2 *Qualitative Studies*

Correlational studies identify the associations between suicide and different social and psychological factors (see *Positivism: Sociological*), without delving too deeply into the concrete influence such variables actually exert on individuals. With time, sociology, turning more to the person who took his own life, wished to go to the heart of this human and social drama and capture the dynamics and the diversity of its significance. Thus, qualitative studies of suicide were undertaken.

In this type of research, the sociology of suicide often prefers the biographical approach in either of its two forms. It is a life story, if it is reconstructed from accounts and documents furnished by the persons themselves who had suicidal ideas or had attempted suicide. The expression life history refers to accounts given by others as well as to various documents allowing the reconstruction of the person's history in situations of completed suicides.

Baechler (1975) retraces, through accounts written by others, short life histories of individuals who have attempted or completed suicide. Adopting the Weberian idealtype, he ends up with several types of meanings in suicidal behavior, which have been mentioned above. Gratton (1996) writes on her own thorough life stories of some Quebec youngsters who committed suicide, using their personal material left behind as well as 29 in-depth interviews with their kin. On that basis, she constructs four Weberian idealtypes involving, for young people, a lack of connectedness between their values and resources. Two of those idealtypes are centered around the values of the youth, either excessive or inadequate, and two around their resources, either turned powerless or so weak that they depend on others.

Published biographies are used to analyze the suicide of 30 famous people in order to check which theories best clarify the meaning of their suicide. Each biography is analyzed according to situations corresponding to each of the theories. Some theories, such as those of Maris and Shneidman, explain better than others those suicides. The least applicable was Freud's theory.

Suicide notes, in particular, offer a privileged instrument with which to capture the significance that the actors themselves attribute to their suicide. Researchers are more and more sensitive to the analysis of such notes. For example, the collective suicide of Jim Jones and his sect members, in 1978, was investigated through that medium. Seven hundred audiotapes were found at the scene of the drama. The words exchanged between Jones and his followers are deciphered in the light of the Durkheimian concepts of altruistic and fatalistic suicides. The leaders who influenced the collective suicide appear to have committed an altruistic suicide. On the other hand, the climate of despair, hostility, and degradation which reigned in that community seems to have provoked a fatalistic suicide for the majority of Jones' disciples.

Life histories of suicide committers also provide the means to map out their psychological autopsies. At first, these autopsies were meant to retrace the social and psychological events that preceded the suicide, in order to understand their subjective significance. However, even if this instrument is still mentioned by Maris (1981) to establish the life trajectory of suicide victims, it is often used by psychologists and psychiatrists merely in the form of standardized questionnaires, aimed at determining the presence or absence of mental illness in those people. Other researchers, interested in intergenerational suicide behavior, will use the 'génogramme' in the reconstruction of life histories; it is a graphic representation of the main events that marked the history of a family for at least three generations.

In summary, when sociology looks at suicide as a social phenomenon of a group or of a society, it tends to favor quantitative and generalizing approaches coming from the natural sciences. However, as soon as sociology becomes concerned with the social actors themselves who adopt suicidal behavior and with the diversity of their experiences, it turns to cultural sciences and uses qualitative approaches focused on the individuals.

7. *Conclusion*

It seems that, ever since Durkheim, the sociology of suicide has been developing by accretion, advancing in three major stages. First of all, Durkheim's *Le Suicide* was a pioneer text, identifying suicide as a social phenomenon whose causes had to be sought in the society itself. The sociology of suicide thus made its appearance on the scene of theoretical and empirical knowledge. However, in the following 50 years, it

made almost no progress. Durkheim's study reigned practically alone. It did not, during that period, awaken a strong current of research on suicide.

Then, in the 1950s, the sociology of suicide began to flourish, especially in the USA. Sociologists were busy verifying, extending, refining, and revising Durkheim's research. Interest in suicide and its social factors broadened progressively, to include the study of the victims of suicide themselves and of their different suicidal behaviors, and to admit psychosociological and psychological approaches. In this regard, social and clinical psychology, Freud's psychoanalysis and Weber's comprehensive sociology have been determining influences in the development of the sociology of suicide.

In the third stage, the prevailing one at the start of the twenty-first century, the sociology of suicide is diversifying greatly, as much in its research subjects as in its methodological tools. It has become a field of study, called suicidology, that is increasingly complex and autonomous, with stretchable boundaries and where multidisciplinarity takes on more and more importance (Kral 1994). In that regard, it is emphasized that the various disciplines need not be in competition with each other, since their common ground is suicide. What they ought to do is to confront their different bodies of knowledge and articulate them in a common reflection. The recent publication of Maris et al. (1999) bears witness to this necessary sharing of knowledge about suicide.

Simultaneously, interest in the sociology of suicide is rapidly becoming universal, embracing researchers in almost all corners of the world. Studies on the subject abound. A search of *Sociological Abstracts* alone, from 1986 to 1999, reveals more than 150 articles from authors in the UK, France, Germany, Spain, Italy, Switzerland, The Netherlands, Norway, Finland, Iceland, Brazil, New Zealand, Japan, Canada, and notably the USA. Periodicals devoted exclusively to suicide have been published: *Suicide and Life-threatening Behavior, Archives of Suicide Research, Giornale Italiano di Suicidologia*.

Finally, given the gravity of the suicidal phenomenon in several societies or groups, particularly among the youth, suicidologists from all scientific disciplines are no longer content merely to study suicide. More and more they demonstrate the desire and the need to act socially. With the goal of intervening to prevent suicide, they join national and international associations and take part in seminars and colloquia around the world. The sociology of suicide has contributed, as have other disciplines, to the fact that scientific research is nowadays trying to apply itself to improving society.

See also: Anomie; Durkheim, Emile (1858–1917); Integration: Social; Quantification in the History of the Social Sciences; Social Integration, Social Networks, and Health; Solidarity: History of the Concept; Solidarity, Sociology of; Status and Role: Structural Aspects; Suicide; Suicide: Ethical Aspects; Violence, History of

Bibliography

Arensman E, Kerkhof A F M 1996 Classification of attempted suicide: review of empirical studies, 1963–1993. *Suicide and Life-threatening Behavior* **26**: 46–67

Baechler J 1975 *Les suicides*. Calmann-Levy, Paris

De Catanzaro D 1981 *Suicide and Self-damaging Behavior. A Sociobiological Perspective*. Academic Press, New York

Douglas J D 1967 *The Social Meanings of Suicide*. Princeton University Press, Princeton, NJ

Durkheim É 1967 *Le suicide. Étude de sociologie*, 2nd edn. Presses Universitaires de France, Paris (1st edn., 1897)

Gibbs J P, Martin W T 1964 *Status Integration and Suicide*. University of Oregon Press, Eugene, OR

Giddens A 1966 A typology of suicide. *Archives of European Sociology* **7**: 276–95

Gratton F 1996 *Les suicides d'être de jeunes québécois*. Presses de l'Université du Quebec Sainte-Foy, Canada

Halbwachs M 1930 *Les causes du suicide*. Arnopress, Glencoe, IL

Henry A F, Short J F Jr. 1954 *Suicide and Homicide*. Free Press, Glencoe, IL

Kral M J 1994 Suicide as social logic. *Suicide and Life-threatening Behavior* **24**: 245–55

Lester D 1989 *Suicide from a Sociological Perspective*. Thomas, Springfield, IL

Lester D 1997 Part III: International perspectives. Suicide in an international perspective. *Suicide and Life-threatening Behavior* **27**: 104–11

Maris R W 1981 *Pathways to Suicide. A Survey of Self-destructive Behaviors*. Johns Hopkins University Press, Baltimore, MD

Maris R W, Canetto S S, McIntosh J L, Silverman M M 1999 *Review of Suicidology, 2000*. Guilford Press, New York

Selvin H C 1958 *Durkheim's Suicide and Problems of Empirical Research. American Journal of Sociology* **63**: 607–19

Shneidman E 1985 *Definition of Suicide*. Wiley, New York

Stack S 1987 The sociological study of suicide: methodological issues. *Suicide and Life-threatening Behavior* **17**: 133–50

Taylor S 1978 Suicide and the renewal of life. *Sociological Review* **26**: 373–90

F. Gratton

Sun Exposure and Skin Cancer Prevention

Sun exposure can be defined as the exposure of human beings to natural and artificial ultraviolet radiation (UVR). UVR is a component of natural sunlight and is produced artificially (e.g., by sunbeds) for cosmetic (e.g., to get tanned skin) or medical reasons (e.g., therapy of psoriasis). Sun exposure is the most

important environmental risk factor for all forms of skin cancer (Marks 1995). Sun protection behavior, on the other hand, is the most effective form of skin cancer prevention. This article provides an overview of the incidence and causation of skin cancer, of the psychological conditions for sun exposure and sun protection behavior, and of skin cancer prevention programs.

1. *Skin Cancer: Incidence and Causation*

The two major classes of skin cancer are malignant melanoma (MM) and nonmelanoma skin cancers (NSC; subtypes: basal cell carcinoma, squamous cell carcinoma). In many countries, skin cancer is the cancer with the most rapidly increasing incidence rate. In the United States, the incidence rates of skin cancer are 12.0 per 100,000 Caucasians for MM (mortality rates: men 3.3 per 100,000 people/year, women 1.7) and 236.6 per 100,000 Caucasians for NSC (mortality rates: men 0.7, women 0.3) (Marks 1995). Incidence rates for people with darker skin are much lower (see *Cancer Screening*).

The development of skin cancer depends on constitutional and environmental factors (for an overview, see Marks 1995). The major constitutional factors are skin color, skin reaction to strong sunlight, and, additionally for MM, large numbers of moles, freckles, and family history of MM. The major environmental risk factor for skin cancer is sun exposure. Cumulative sun exposure is related to the development of NSC. For MM, the role of sun exposure is less clear. Increased risk of MM seems to be related to severe sunburns, in particular in childhood, and intermittent recreational sun exposure.

2. *Predictors of Sun Exposure*

Sun exposure not only accompanies many working and recreational outdoor activities, it is also seeked intentionally (for an overview of studies on psychosocial factors of sun exposure, see Arthey and Clarke 1995, Eid and Schwenkmezger 1997). The major motivational predictor for unprotected sun exposure is the desire for a suntan. Level of suntan is related to the perception of other people as being healthy and attractive in a curvilinear manner, with a medium tan considered the healthiest and most attractive skin color. Therefore, getting a suntan helps people to make a good impression on others. The belief that a suntan improves one's physical appearance is the strongest predictor of suntanning. Voluntary sun exposure is related to thinking more about one's physical appearance, to a higher fear of negative evaluations by others, and to a higher body self-consciousness and body self-esteem. In addition to these variables of personal appearance, suntanning is related to a positive attitude toward risk taking in general and to having friends who sunbathe and use sunbeds (see *Health Behavior: Psychosocial Theories*).

3. *Predictors of Sun Protection Behavior*

The harmful effects of UVR on the skin can be prevented by sun protection behavior. The four most important protection behaviors are: (a) Avoiding direct sunlight, particularly during the peak hours of daylight (10 am to 4 pm), (b) seeking shade, (c) wearing a hat, (d) wearing skin-covering clothing, and (e) using sunscreen with a sun protection factor of at least 15. Even in high-risk countries such as Australia, the prevalence rates of these protective behaviors are rather low indicating that about half of the population are inadequately protected on sunny days (e.g., Lower et al. 1998). Voluntary sun exposure and the use of sun protection are two largely independent and only weakly correlated behaviors. Moreover, the predictors for sun protection behavior differ from the predictors for sun exposure. In contrast to sun exposure, appearance-oriented variables are less important for predicting sun protection behavior. The major predictors for sun protection behavior are perceived threat of skin cancer, the benefits and barriers of different sun protection behaviors, social factors, and knowledge about skin cancer (Eid and Schwenkmezger 1997, Hill et al. 1984).

Perceived threat of skin cancer depends on the perceived severity of and the vulnerability to skin cancer. People who rate skin cancer as a severe disease and who think that they are more vulnerable to skin cancer are more likely to use sun protection. Furthermore, individuals with a sensitive and fair skin, both which are more prone to skin cancer, use sun protection more often. This risk perception is strongly related to gender. Women have stronger vulnerability and severity beliefs. Moreover, women use sunscreen more often and apply sunscreen with a higher sun protection factor than men (see *Gender and Cancer* and *Gender and Health Care*).

The actual display of specific sun protection behaviors depends on (a) the belief that these behaviors are effective in preventing skin damage, sunburns, and skin cancer, and (b) the specific barriers of each behavior (Hill et al. 1984). Specific barriers to the application of sunscreen is the belief that sunscreen is greasy and sticky and the fact that sunscreen has to be applied repeatedly. An additional barrier for men is the belief that sunscreen makes men appear seemingly sissy and looks unattractive. Barriers against wearing hats are the beliefs that hats are inconvenient to wear as well as causes problems with hairstyle. Additionally, for men, wearing a hat is associated negatively with the beliefs that wearing a hat causes baldness and a sweaty head, is inconvenient on windy days, and gets in the way when playing sports. The major barriers against

wearing shirts are the discomfort from heat and the feeling of being overdressed. Therefore, a goal of skin cancer prevention programs must be the removal of these specific barriers against sun protection behavior. Because these barriers partly depend on temporary fashion influences, the present-day barriers of specific sun protection behaviors have to be examined before a prevention campaign is planned.

Social factors also have an important influence on sun protection. Use of sun protection is correlated with peers' sun protection behavior, parental influence, and parental sun protection behavior. Furthermore, people who know a person suffering from skin cancer use sun protection more often. Finally, knowledge about the risks of sun exposure has been proven a significant predictor for sun protection in many studies. In particular, people with more knowledge about skin cancer use sunscreen with a higher sun protection factor.

4. *Skin Cancer Prevention*

With respect to the target groups considered, sun exposure and sun protection modification programs can be grouped into three classes: (a) programs using mass media, (b) community intervention programs, and (c) educational programs (for an overview of these programs, see Eid and Schwenkmezger 1997, Loescher et al. 1995, Morris and Elwood 1996, Rossi et al. 1995).

4.1 *Mass Media Programs*

Typical mass media programs use informational pamphlets, comic books, newspaper reports, TV and radio advertisements, or videotapes to carry their messages. These mass media campaigns are either often not evaluated, or the evaluations are published in reports that are not accessible to the public, or the evaluations suffer from methodological problems (e.g., missing control group). However, there are some studies evaluating single videotapes or informational pamphlets by an experimental or control-group design (for an overview see Eid and Schwenkmezger 1997, Morris and Elwood 1996). According to these studies, videotapes and informational pamphlets can increase knowledge about, perceived severity of, and vulnerability to skin cancer, as well as increase individual sun protection intentions. Studies focusing on long-term effects on behavior, however, are missing. Furthermore, these studies show that an emotional prevention video imparting knowledge by employing a person with skin cancer might be more appropriate than a more unemotional presentation of facts. Moreover, educational texts focusing on the negative effects of sun exposure for physical appearance (wrinkles, skin aging) might be more appropriate than texts on the negative health-related consequences. This, however, might only be true for people with a low appearance orientation, whereas appearance-based messages can result in boomerang effects for high appearance-oriented people. Finally, the manner in which a health message is framed is important. Detweiler et al. (1999) demonstrated that messages on sun protection behavior are more effective when they highlight potential positive effects of the displayed sun protection behavior (gain-framed messages) than when they focus on the negative effects of the omitted behavior (loss-framed messages).

4.2 *Community Intervention Programs*

Mass media are often a component of community intervention programs, but community intervention programs can implement several other methods as well (see *Gender and Cancer*). Community-based programs can make use of local peculiarities (e.g., the daily UVR rate in a community) and can focus on local places of risk behavior, for example, swimming pools and beaches. Based on principles of learning theory, Lombard et al. (1991) designed an intervention program with pool lifeguards as models for sun protection behaviors and daily feedback posters of the sun protection behavior displayed the day before. This program was effective with respect to two sun protection behaviors (staying in the shade, wearing shirts) but not other ones. Rossi et al. (1995) describe several new techniques such as special mirrors (sun scanner) and photographs (sun damage instant photography) that have been used in intervention programs on beaches to show the effects of sun damage and photoaging in a more dramatic fashion.

4.3 *Educational Programs*

Because of the cumulative risk of sun exposure and the risks of sunburn in childhood, intervention programs are particularly important for children and adolescents. Educational programs have been developed for students of different ages from preschool to high school (see *Health Promotion in Schools*). In general, these programs are effective in enhancing knowledge about skin cancer and the awareness of the risks of unprotected sun exposure. Changing attitudes toward sun protection, however, have only been found after participation in programs consisting of more than one session. With respect to changes in behavioral intentions the results are inconsistent. Significant changes in intentions have been found in only a few studies and not consistently across the intentions to different behaviors. Changes in observed behavior, however, have not been analyzed. The failure of educational programs in changing behavioral intentions might be

due to the fact that these programs have focused more on motivational than on behavioral factors. Multimethod educational programs including components on the implementation of behavioral routines that are very important for developing health behavior in childhood are desiderata for future research.

5. Future Directions

Sun exposure and skin cancer prevention are relatively new fields of research. Future research will profit from the use of multimethod assessment strategies of behavior that does not only assess self-reported behavioral intentions but also uses observational and physical measures of behavior (e.g., spectrophotometers for measuring the melanin skin content). Furthermore, research on new components of intervention programs is needed. For example, future programs could consist of behavior change methods in addition to attitude change methods as well as strategies for weakening the association between suntan and attractivity.

See also: Cancer-prone Personality, Type C; Cancer: Psychosocial Aspects; Cancer Screening; Diabetes: Psychosocial Aspects; Health Behaviors; Health Education and Health Promotion

Bibliography

Arthey S, Clarke V A 1995 Suntanning and sun protection: A review of the psychological literature. *Social Science Medicine* **40**: 265–74

Detweiler J B, Bedell B T, Salovey P, Pronin E, Rothman A 1999 Message framing and sunscreen use: Gain-framed messages motivate beach-goers. *Health Psychology* **18**: 189–96

Eid M, Schwenkmezger P 1997 Sonnenschutzverhalten (Sun protection behavior). In: Schwarzer R (ed.) *Gesundheitspsychologie* (*Health Psychology*). Hogrefe, Göttingen

Hill D, Rassaby J, Gardner G 1984 Determinants of intentions to take precautions against skin cancer. *Community Health Studies* **8**: 33–44

Loescher L J, Buller M K, Buller D B, Emerson J, Taylor A M 1995 Public education projects in skin cancer. *Cancer Supplement* **75**: 651–56

Lombard D, Neubauer T E, Canfield D, Winett R 1991 Behavioral community intervention to reduce the risk of skin cancer. *Journal Applied Behavioral Analysis* **4**: 677–86

Lower T, Grigis A, Sanson-Fisher R 1998 The prevalence and predictors of solar protection use among adolescents. *Preventative Medicine* **27**: 391–9

Marks R 1995 An overview of skin cancers: Incidence and causation. *Cancer Supplement* **75**: 607–12

Morris J, Elwood M 1996 Sun exposure modification programmes and their evaluation: A review of the literature. *Health Promotion International* **11**: 321–32

Rossi J S, Blais L M, Redding C A, Weinstock M A 1995 Preventing skin cancer through behavior change. Implications for interventions. *Dermatol. Clin.* **13**: 613–22

M. Eid

Supply Chain Management

1. Introduction

Supply chain management is one of the most essential aspects of conducting business. Many people outside of the direct community (in research and industry) do not realize this because an ordinary consumer often experiences only its effects. Recall the times when the item that you wanted was not available in your favorite garments or grocery store, recall how many times you got a great 'deal' at the end of the season, recall the sudden increases in gas prices due to shortages, recall the times when your e-commerce site promised availability but later could not send the required product or sent you the wrong product, or recall the times when your customized product (like a personal computer or kitchen cabinet) was delayed to a great extent. All the above and several other experiences that consumers have on a routine basis are direct consequences of supply chain practices followed by firms. As opposed to business-to-consumer transactions, supply chain practices have immediate impact on business-to-business transactions. A few years back Toyota had to shut down its manufacturing facility in Japan due to supply shortages for its brake pedals, Boeing took a charge for several million dollars in the late 1990s due to insufficient capacity and part shortage resulting from an inability of the supply base to ramp up production, while firms such as Dell Computers, Wal-Mart, and 7-Eleven Japan have consistently outperformed competition due to their great strengths in supply chain management.

2. Definition

Supply chain management is such a vast topic that as a result people often give it a different definition based on their own personal experience. To some, supply chain management is all about managing the supplier base, determining what to outsource and to whom, and managing relationships with the various suppliers. To some others it is efficient ways of transferring goods from one place to another taking into account the distribution and transportation costs. To another set of people it is all about how the different firms in the distribution channel or value chain are integrated in terms of information systems and inventory management practices. To yet another group it is effective management of fixed and variable assets required for running the business. In a sense all these definitions are like the blind men defining the elephant based on its different organs. A comprehensive definition of supply chain management can be given as follows.

A supply chain is the set of entities that are involved in the design of new products and services, procuring

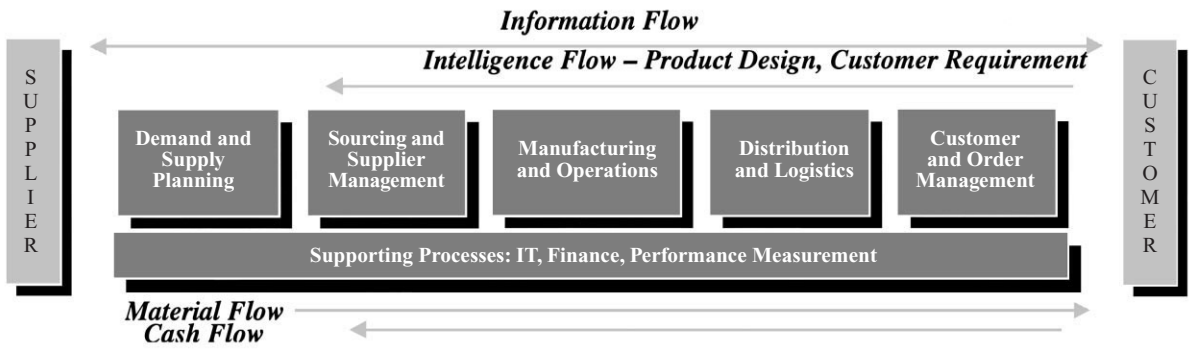

Figure 1
Supply chain process

raw materials, transforming them into semifinished and finished products and delivering them to the end customer. Supply chain management is efficient management of the end-to-end process starting from the design of the product or service to the time when it has been sold, consumed, and finally disposed of by the consumer. This complete process includes product design, procurement, planning and forecasting, production, distribution, fulfillment, and after-sales support (see Fig. 1).

Supply chain management issues can be classified into two broad categories—configuration and coordination. Configuration-level issues relate to the basic infrastructure on which the supply chain executes and coordination-level issues relate to the actual execution of the supply chain.

2.1 Configuration-level Issues

2.1.1 Supply base decisions. How many and what kinds of suppliers to have? Which parts to outsource and which to keep in house? How to standardize and streamline procurement practices? Should one use vertical marketplaces for auctions or should one invest in developing highly integrated supply partnerships? How long or short contracts with suppliers should be?

2.1.2 Plant location decisions. Where and how many manufacturing, distribution, or retail outlets to have in a global production distribution network? How much capacity should be installed at each of these sites? What kind of distribution channel should a firm utilize—traditional brick and mortar, direct to consumer via Internet or phone, or a combination?

2.1.3 Product portfolio decisions. What kinds of products and services are going to be supported through the supply chain? How much variety to provide to customers? What degree of commonality to have across the product portfolio?

2.1.4 Information support decisions. Should enterprise resource planning software be standardized across functional units of a firm? Should the supply chain work on standard protocols such as XML (extended markup language) or on proprietary standards?

2.2 Coordination-level Issues

2.2.1 Material flow decisions. How much inventory of different types of products should be stored? Should inventory be carried in finished form or semifinished form? How often should inventory be replenished? Should a firm make all of its inventory decisions or is it better to have the vendor manage the inventory? Should suppliers be required to deliver goods just in time?

2.2.2 Information flow decisions. In what form is information shared between different entities in the supply chain—paper, voice via telephone, EDI (electronic data interchange), XML? To what degree does collaborative forecasting take place in the supply chain? What kind of visibility is provided to other entities in the supply chain during execution? How much collaboration takes place during new product or service development among the supply chain partners?

2.2.3 Cash flow decisions. When do suppliers get paid for their deliveries? What kinds of cost reduction

efforts are taken across the supply chain (or expected of suppliers)? In a global firm, in which currency will a supplier be paid?

2.2.4 Capacity decisions. How to optimally utilize the existing capacity in terms of manpower and machines? How to schedule on a manufacturing line to complete jobs on time? How much buffer capacity to have for abnormal situations with excess demand?

As is evident, configuration and coordination issues are interdependent. Configuration issues can be viewed as strategic long-term decisions whereas coordination issues are medium-to-short term decisions. Generally, firms develop a strategy for the configuration-level decisions and then constrain the coordination decisions based on those.

3. Complexities Associated with Supply Chains

As evident from discussions in the earlier section, supply chain management spans several functional and geographical areas. This introduces complexities both in terms of design and execution of supply chains. Some of the pertinent factors that complicate supply chain management decisions are as follows.

3.1 Multiple Agents

Supply chain issues need to be decided by different entities sometimes having different interests. For example, a retailer may want that the distributor provide very high availability for the products but at the same time not charge anything additional from the retailer. The distributor may sometimes agree to that but in turn may want information about actual customer sales which the retailer may not want to share. Even when decisions have to be made within the same firm there could be incentive issues. For example, the marketing or sales department, typically a revenue center, presents the future demand forecast to the manufacturing department that is a cost center. Clearly, there is incentive for the former to over-forecast and the latter to under-produce (as compared to the forecast). This creates several difficulties while deciding on the amount of inventory to be stocked. Another related issue is encountered where the marketing department may push for huge amount of variety in the product/service offerings; the manufacturing department may not want to embrace that because additional complexities are created during execution.

3.2 Uncertainty

Accurately matching supply and demand is the ultimate goal of effective supply chain management but that is complicated by uncertainty at various levels of the process. There is uncertainty in product and technology development, in predicting customer demand, in day-to-day operations and manufacturing, and supply. Typically, uncertainty creates more inefficiency in the system. For example, if the final demand for a sweater at a store cannot be predicted accurately then the firm either stocks too little (in which case it suffers from stock-outs) or produces too much (in which case it has to salvage the inventory through a huge sale at the end of the season). Similarly, the uncertainty in supply may necessitate additional buffer inventory.

3.3 Information Asymmetry

Since supply chain processes extend across multiple functional units within a firm and often across different firms there is a high degree of asymmetry in terms of information. This is caused primarily by two main reasons—one relates to lack of adoption of information technology and the other relates to reluctance to share information with other supply chain partners. The lack of information causes several problems during actual fulfillment. For example, when a consumer goes to an e-commerce site and buys an item off the electronic catalog, the consumer expects to receive the product on time. The consumer is not aware that the inventory status on the product may be updated only once a week and that the information on the site may be outdated. As a result, the consumer is disappointed when the product does not arrive on time.

3.4 Lead Time

Each and every task in the supply chain process needs time to be completed and the resources (labor, machines, or computers) have limited processing capacity. As a result, not all tasks can be completed after the actual demand is known and some of the tasks need to be done up front (which may or may not get utilized based on the actual demand realized). Further, the limited capacity associated with the resources creates variability in the actual realized lead-time, which in turn necessitates greater resource requirements at the next stage in the supply chain. The above complexities lead to several types of inefficiencies in the supply chain that are often perceived as the 'bad effects' of inefficient supply chain management. Some of the major inefficiencies can be classified into the following categories.

4. Inefficiencies of Supply Chain Management

4.1 Poor Utilization of Inventory Assets

One common effect of poor supply chain management is having excess inventory at various stages in the

supply chain, at the same time having shortages at other parts of the supply chain. Since inventory forms a substantial part of working assets of a firm, poor management could lead to huge inefficiencies. Lee and Billington (1992) provide an excellent overview of pitfalls and opportunities associated with inventory management in supply chains.

4.2 Distortion of Information

Another effect relates to lack of visibility of demand and supply information across the supply chain which causes the bullwhip effect. This effect describes how a small blip in customer demands may get amplified down the supply chain because the different entities in the supply chain generate and revise their individual forecasts and do not collaborate and share actual demand information. Lee et al. (1997) describe the causes and controls for this effect.

4.3 Stock-outs

Poor supply chain management also results in late deliveries and large stock-outs. Fundamentally, these effects are caused due to an inability of the firm to predict the requirement for raw material and equipment capacity together with the uncertainty associated with obtaining deliveries of products on time from its suppliers. Fisher et al. (1994) describe how accurate forecasts in the apparel industry could potentially reduce this inefficiency.

4.4 Customization Challenges

As the degree of customization has increased in the marketplace, one of the immediate effects of poor supply chain management relates to late deliveries of customized products. Firms are developing several strategies in order to provide variety while keeping costs under control. These include delaying differentiation of the product and introducing more commonality and modularity in product lines (see Swaminathan and Tayur 1998).

5. Supply Chain Models: Past, Present, and Future

The science related to supply chain management traces its history back to the early 1950s when several researchers were interested in understanding the optimal policies related to inventory management. One of the first pieces of work in this stream relates to the models developed by Clark and Scarf (1958) for managing inventories at multiple echelons. Several hundreds of researchers have studied related inventory problems under stochastic and deterministic environments since the 1950s.

This research is captured concisely in the research handbook edited by Graves et al. (1993). There is a large amount of literature in the area of transportation and distribution as well as plant location models in the context of supply chain management. Traditional researchers focussed on developing optimal policies and rules for specific supply chain issues assuming a centralized control of the supply chain. In the 1990s researchers have started to study problems which take a decentralized multi-agent approach to analyzing supply chain problems, integrate information availability across the supply chain with logistics decisions, develop new models for supply contracts, and demand forecasting and integrate product design with supply chain management. A collection of prominent pieces of research in this area is contained in Tayur et al. (1999). In addition to academic research, several firms in the 1990s have developed successfully and employed large analytical and simulation models for supply chain optimization and execution. Arntzen et al. (1995) describe one such system developed for Digital Equipment and recently, IBM was awarded the Franz Edelman Award by INFORMS (The Institute for Operations Research and Management Sciences) for a supply chain optimization project which led to $750 million in inventory savings.

In the twenty-first century, firms face severe challenges in terms of global competition and customer requirement for greater variety, shorter and reliable delivery times, and lower prices. The advent of e-commerce has created immense opportunities but at the same time has made firms more vulnerable to logistics pitfalls. Today customers do not just buy products but they buy delivered products. As a result fulfillment is as important as making the sale. As opposed to traditional channels where inventory could be stored to hide other inefficiencies in terms of lead time and poor forecasts, in the fast-paced electronic business environment such arrangements are not as useful. As a result, firms are beginning to pay more attention towards supply chain management. Both business-to-consumer and business-to-business e-commerce environments have introduced several issues related to supply chain management which are likely to be studied by researchers in the near future.

The prevalence of the Internet has led to the development of vertical market places that promise to reduce the inefficiencies in the buying process in several industries. On one hand, these market places are likely to reduce the cost of goods for the manufacturer due to more competition leading to better prices. This line of thought indicates that in future supply chains may be more agile and supplier relationships may be short-term oriented. On the other hand, several firms realize that greater benefits can be attained if some of these market places can in fact be used for process integration and collaboration across the supply chain.

In such an environment, firms need to develop greater trust so that they would be willing to share information with their supply chain partners. Researchers today are trying to identify under what conditions one or the other scenario may play out and what kinds of new models and analysis need to be developed. A related effect of the Internet is the expansion of global supply chains. Today it is much easier for any supplier located in a remote part of the world to bid on contracts from large firms in developed nations with whom they may not have done business in the past. Issues related to coordination of global supply chain management are likely to be an integral part of supply chain management research in the future. Another important research topic is the reverse logistics issues related to supply chain management. Traditionally, researchers have only concerned themselves with efficient movement of goods from supplier to the customer. Now a greater number of researchers are studying problems related to disposal of used products, refurbishing old products, making packaging more environmentally friendly, and basing supplier selection on environmental criteria in addition to traditional criteria related to cost, quality, and reliability. Another new stream of research is the study of supply chains in the service industry. As opposed to traditional manufacturing-oriented supply chains, service supply chains are more complicated due to the inability to store inventory. Uncertainty is handled in those cases using additional buffer capacity. Finally, researchers are beginning to focus more on the integration of product design and supply chains and developing models to understand the concepts related to design for supply chain management.

See also: Commodity Chains; Location Theory; Market Areas; Market Structure and Performance; Marketing Strategies; Retail Trade

Bibliography

Arntzen B C, Brown G G, Harrison T P, Trafton L L 1995 Global supply chain management at Digital Equipment Corporation. *Interfaces* **25**(1): 69–93

Clark A J, Scarf H 1958 Optimal policies for a multi-echelon inventory problem. *Management Science* **6**: 475–90

Graves S C, Rinnooy Kan A H G, Zipkin P H 1993 *Logistics of Production and Inventory, Handbook in OR and MS*. North Holland, Amsterdam, Vol. IV

Fisher M L, Hammond J H, Obermeyer W R, Raman A 1994 Making supply meet in an uncertain world. *Harvard Business Review* **72**(3): 83–9

Lee H L, Billington C 1992 Managing supply chain inventory: Pitfalls and opportunities. *Sloan Management Review* **33**(3): 65–73

Lee H L, Padmanabhan V, Whang J 1997 Information distortion in a supply chain: The bullwhip effect. *Management Science* **43**(4): 546–8

Swaminathan J M, Tayur S 1998 Managing broader product lines through delayed differentiation using vanilla boxes. *Management Science* **44**(12): S161–72

Tayur S, Ganeshan R, Magazine M J 1998 *Quantitative Models for Supply Chain Management*. Kluwer, MA

J. M. Swaminathan

Support and Self-help Groups and Health

1. Nature and Use of Self-help and Support Groups

As psychosocial interventions, support groups and mutual aid, self-help (MASH) groups have a great deal in common. Their members share a similar health concern, life event, role transition, or chronic stressor, and they exchange a variety of types of support through face-to-face interaction, bringing to bear their experiential knowledge to gain an understanding of their personal and collective predicament and of ways of coping with it. The differences between these two small group structures are that support groups are organized, governed, and led by professionals who provide supplementary expert knowledge or skill training, whereas MASH groups are self-governing, and led by individuals who are peers by virtue of sharing the group's problem or concern. In addition, support groups tend to have a closed membership whereas MASH groups are open, and support groups tend to have a fixed duration whereas MASH groups are indefinite in duration. Finally, some MASH organizations engage in advocacy efforts on behalf of marginalized or vulnerable segments of the population (Reissman and Carroll 1995), whereas support groups tend to look inward rather than outward. Both support and MASH groups differ quite dramatically from therapy groups; the leader does not engage in any clinical activities, such as psychological assessment, and the members are not assigned to a group on the basis of their symptomatology, diagnostic label, or prognosis.

A recent national US survey revealed that approximately 7 percent of adults had participated in a MASH group in the prior 12 months, and 18 percent had done so at some point in their lifetime (Kessler et al. 1997). A similar study in Canada, also based on a national probability sample, revealed a much lower prevalence of 2 percent of adults who had participated in a MASH group in the prior 12 months (Gottlieb and Peters 1991). Perhaps the self-help ethos is stronger in the US than in Canada, where a national healthcare system may also serve as a disincentive to MASH participation. Humphreys and Klaw (1998) report

that, in the US, there are MASH organizations for almost every major chronic condition and leading cause of mortality. Examples of the former include groups for arthritis sufferers, people diagnosed with ischemic heart disease, cancer patients, diabetics, and people with hearing and visual impairments. MASH organizations are heavily utilized by people with substance abuse and emotional problems, Alcoholics Anonymous (AA) being the most widely used of all MASH organizations and probably the most widely accepted by professional referral agents such as family physicians and addictions counselors. AA also has been subjected to more evaluation research than other MASH groups, one recent study finding that, over three years, participants decreased their daily alcohol intake by 75 percent while also effecting substantial cost savings relative to those receiving professional outpatient treatment (Humphreys and Moos 1996).

Although no corresponding census has been taken of public participation in support groups, prior reviews of the literature have spotlighted the abundance of groups for cancer patients (Fawzy et al. 1995, Helgeson and Cohen 1996), family caregivers of older adults, particularly those suffering from dementia (Bourgeois et al. 1996, Lavoie 1995, Toseland and Rossiter 1989), and people coping with loss through death or divorce (Hughes 1988). The fact that support groups have been subjected to far more research than MASH groups testifies to the greater ease of access they provide, due in part to their professional sponsorship and more casual attitude toward maintaining anonymity.

Both support and MASH groups function as specialized personal communities that supplement or compensate for deficiencies in the participants' natural networks (Helgeson and Gottlieb 2000). These deficiencies may stem from the natural network's inability to comprehend or tolerate the thoughts, feelings, and behaviors that arise from relatively novel stressors, the network's impatience with or exhaustion from lengthy periods of rendering support and aid, or from other interpersonal dynamics that compromise the network's prosocial behaviors, such as conflicts about appropriate ways of coping (Gottlieb and Wagner 1991) and emotional overinvolvement (Coyne et al. 1988). In some circumstances, coparticipants substitute for a natural network that has proved destructive, such as when the fellowship offered by AA supercedes former drinking buddies (Humphreys and Noke 1997). In some contexts coparticipants serve functions complementary to the natural network's, such as when a weight loss group helps to maintain the gains effected by dietary changes at home.

2. *Group Design and Dynamics*

From the perspective of an observer, meetings of MASH groups and support groups would be virtually indistinguishable except for two differences. First, the support group would devote a portion of its time to education or training provided by an expert, whereas the MASH group would devote the entire meeting to the swapping of experiences and the exchange of support. Second, many, but not all MASH groups would operate on a 12-step model or a variation thereof, which consists of quite elaborate ideological formulations about the cause and source of difficulty and about the ways members need to think about their situations in order to come to terms or resolve them (Antze 1976). Underlying the 12-step belief system are the ideas of spiritual surrender and personal struggle, along with the need for moral cleansing and mutual aid. In contrast, support groups rarely import an ideology but develop their own norms, with whatever cognitive restructuring occurring as a result of the resources introduced by the professional leader and the influences arising from the members' interactions.

Aside from these two differences, the same set of structural elements need to be considered in designing both MASH and support groups. These include: (a) the setting and scheduling of the group sessions, including their number, length, duration, and intervals; (b) the leadership, which may be consistent or rotating, and involve one or two (co-led) facilitators; (c) the group composition, taking into account demographic (e.g., age, gender, socioeconomic status), stressor-related (e.g., stage, severity), and coping-related (e.g., extent of distress and resource mobilization) criteria; and (d) the format, including the extent of meeting structure, extra-group contacts among members, and behavioral training. To date, there has been relatively little research comparing the effects on both the process and outcomes, as well as attrition, resulting from systematic manipulation of these structural elements. For example, little is known about ways of composing the groups to optimize the social comparisons that members make, and whether aspects of the social mix or of the problem they share more strongly affect members' attraction, participation, social learning and supportive exchanges in the group (Gottlieb 1998).

The beneficial psychosocial dynamics that underlie the mutual aid occurring in both support and MASH groups include the emotional venting along with the validation and normalization of emotions that are gained from contact with similarly afflicted peers, the opportunity to enhance one's sense of worth and control by providing support to others, a process that has been called helper-therapy (Riessman 1965), and the opportunities for self-enhancement and self-improvement that derive from the social comparisons that members make. These latter comparisons may take an upward, downward, or lateral direction, ideally respectively serving self-improvement, self-enhancement, and normalizing functions. For example, in a widely cited study, Taylor (1983) found that the vast majority of women with breast cancer

found a real or imagined dimension on which they compared more favorably to other women with the same diagnosis.

These psychosocial dynamics have yet to be integrated within a comprehensive model of the process through which support and MASH groups exert an impact on health outcomes, whether psychiatric or medical. That is, to understand the mechanisms of action through which these psychosocial dynamics influence health and morale, competing theoretical models should be specified and tested. Cohen (1988), Stroebe et al. (1996), and Wills and Filer (1999) have described several potential direct and indirect pathways through which social support may affect health, including physiological mechanisms, appraisal and reactivity mechanisms, and behavioral mechanisms of action. For example, it is presently unknown whether the increased survival time documented by Spiegel et al. (1989) and Fawzy et al. (1993) for women with metastatic breast cancer and malignant melanoma respectively resulted from mediating improvements in their immune and endocrine function, the enhancement of positive health behaviors such as diet, exercise, and compliance with their therapy regime, or a purely psychological process involving hope and a sense of being needed by interdependent group members. Naturally, there may also be a complex interaction among these mechanisms. The important point is that support and MASH groups that are informed by these models should be designed to impact the hypothesized mediating processes, and measures that are sensitive to change in these processes must also be included.

3. *Group Impacts*

Evaluation of the effects of MASH groups has consisted largely of feedback regarding consumer satisfaction, not randomized controlled trials that administer standardized outcome measures at several intervals following the intervention. Indeed, since most MASH groups are open-ended in nature, allowing members to flexibly come and go as their needs dictate, it is virtually impossible to assess anything that resembles a uniform intervention. Moreover, many MASH groups are not oriented to personal change, recovery, or improved adjustment; instead, they offer acceptance, support, and fellowship to people who feel marginalized (e.g., people with developmental disabilities) or who have been stigmatized (e.g., people with mental illnesses), or people who live with and care for someone who has one of those statuses. In short, many MASH groups view support as an end in itself rather than as a means to other ends, rendering conventional outcome evaluation irrelevant.

Those MASH groups that have been rigorously evaluated consist of both *ad hoc* groups designed along self-help principles and groups that are part of a MASH organization, such as AA. Examples of *ad hoc* groups are parents of premature infants (Minde et al. 1980), bereaved wives (Marmar et al. 1988), and older adult men who have diabetes (Gilden et al. 1992). All three of these studies report significant desirable health effects that lasted for several months after the group ended, compared to less effective or equally effective but more costly comparison interventions such as psychotherapy. Though limited in number, evaluations of groups that are part of a larger MASH organization show similar beneficial effects, including Humphreys and Moos' (1996) study of participants in AA, and Lieberman and Videka-Sherman's (1986) study of bereaved participants in THEOS (They Help Each Other Spiritually). One provocative finding of the latter study is that participants who developed friendships through their group experience achieved significantly better adjustment than those who did not, suggesting that MASH group effects are magnified for those who have the social skills, personality, or motivation to convert a tie based on shared organizational membership and stressful circumstances into a strong personal tie.

As previously noted, there exist far more evaluations of the health impact of support groups than of MASH groups, including review papers on groups for cancer patients, the family caregivers of elderly persons with dementia, and the bereaved. Both within and across these three stressor contexts there is much variability in the characteristics of the participants, the structure, content, and format of the meetings, and the measures used to tap the intended effects. The only constant is the 'dosage,' with the vast majority of groups meeting between 6 and 10 times for 1–2 h on each occasion. As for the impacts, there seems to be far more consensus about the limited impact of support groups for family caregivers than for the other two stressor contexts.

Specifically, the evidence shows that the mental health of family caregivers does not improve much from such a brief intervention, but that longer-term groups and groups that involve skill training that is tailored to situational demands and needs effect larger gains, whether measured in terms of general (e.g., depressive affect and anxiety) or domain specific (e.g., role performance and hostility toward the recipient of care) indicators (Gottlieb 1998, Lavoie 1995, Toseland and Rossiter 1989, Ostwald et al. 1999). It appears that a short-term intervention is not matched appropriately to the chronic character of the caregiving experience, and it is also hard to see how such a limited group experience can mitigate the distress that is a natural biproduct of helplessly witnessing the deterioration of a family member.

Fawzy et al. (1995) also concluded their review of the effects of support groups for cancer patients with the recommendation that short-term groups were best suited to the needs of newly diagnosed patients who are early in their treatment, whereas long-term groups were better suited to those with advanced metastatic disease. An example of the latter is a year-long support

group intervention that involved weekly 90-min meetings plus abundant extra-group contact among the members, women who had metastatic breast cancer (see *Gender and Cancer*). Co-led by a professional and a counselor whose breast cancer was in remission, the group achieved better adjustment (e.g., mood disturbance, tension-anxiety, vigor, confusion) than a control group by the 12-month measurement interval, but not by the two preceding 4-month and 8-month intervals. Moreover, by the time of the 10-year follow-up, the support group participants had increased their survival time by 18 months relative to the control group. Although this study gives reason for optimism about the health benefits of longer-term support groups, the findings must be replicated before any definitive conclusions can be drawn. Moreover, future research should devote more attention to identifying the mechanisms underlying the group's ameliorative effects and determining which subgroups of participants and which disease diagnoses benefit most from this type of psychosocial intervention. Clearly, people who are more extraverted, inclined to cope by seeking information and venting their feelings, and have the skills and motivation to form attachments with co-participants, are more likely to be attracted to this medium in the first place and perhaps more prone to benefit from the provisions it supplies. Conceivably, the group offers a sense of belonging and being needed by others, and these perceptions may stimulate more active immune system functioning, which may account for the superior outcomes that have been observed.

Helgeson and Cohen (1996) limited their review of psychosocial interventions for cancer patients to a comparison of peer support vs. education vs. combined support and education groups. The former concentrate on the provision of emotional support, whereas the latter mainly deliver informational support. Although their conclusions are hedged due to the numerous methodological weaknesses of the studies and differences in the samples and disease sites and stages, they offer important insights into the mechanisms that may be implicated in psychological and physical adjustment to cancer (see *Cancer: Psychosocial Aspects*). These mechanisms include enhanced self-esteem, gains in perceived control, optimism regarding the future, improved processing of emotions, and opportunities to derive meaning from their adversity. These and other pathways, including biobehavioral processes, deserve greater scrutiny in future intervention trials.

Support groups for people who have experienced losses through death or divorce have proved to speed the adjustment process, compared to people who have not used any professional or mutual aid resource (Bloom et al. 1985, Hughes 1988, Vachon et al. 1980, Silverman 1986). Groups for children whose parents have divorced have also been found to foster improved behavioral, academic, psychological and social adjustment (see Grych and Fincham, 1992, for a review). Support groups have also been offered to adolescents with eating disorders and to children of varying ages who have a parent with a mood disorder (Beardslee et al. 1993) or with an alcohol or substance abuse problem (Emshoff 1990).

Evaluation of support groups is complicated by the fact that these interventions usually consist of multiple active ingredients, not just the support that is exchanged among the members. In fact, in a rare comparison of support groups with and without a skill-training component, Stolberg and Mahler (1994) found that the group that lacked skills training was no more effective than the control group for children whose parents had divorced. More studies that adopt such a dismantling design (West et al. 1993) need to be conducted in order to demonstrate that changes in social support are responsible at least in part for improved outcomes. This means that measures of support need to be taken and then related to the distal outcomes of the intervention as a strategy of confirming the mediating role of support.

Finally, recognizing that many people who might be attracted to a support group live in geographically remote communities, have disabilities that limit their mobility, or prefer not to divulge their identity, a number of investigators have implemented computer-mediated groups, using the technology that has recently become available at relatively low cost. For example, Dunham et al. (1998) trained 50 poor, young, single mothers to use a bulletin board system to post both private and network-wide messages to one another and to hold text-based teleconferences with up to eight mothers at a time. With the mothers' permission, Dunham et al. (1998) analyzed the extent and nature of the participants' communications, and found a significant pre- to post-test decrease in parenting stress for those mothers who participated consistently over the 6-month study period. Telephone support groups have also been implemented for family caregivers of persons with dementia (Goodman and Pynoos 1990), a vulnerable population that is largely home-bound. A distinct psychological advantage of support groups that use computer and telephone media is that they moderate the emotional intensity of the group experience, which in turn may decrease its threat. On the other hand, without direct exposure to the other group members, the social comparison process (see *Social Comparison, Psychology of*) may be more limited and the possibility of friendship formation, which has proved to be so vital in the bereavement and cancer studies reviewed earlier, is diminished.

4. *Conclusion*

MASH and support groups are relatively low-cost and highly accessible helping resources that promote health and well being through many psychological

processes. In the health field, these groups have been employed both as an alternative and a complement to professional services. Moreover, as the actual and potential health benefits of these groups become more widely recognized, mutual aid strategies are more likely to become integrated within health care and health promotion programs (Humphreys and Klaw 1998). More systematic investigation of different ways of structuring the groups and tailoring them to the needs and coping styles of prospective members, promises to augment their attractiveness and health impacts.

See also: Social Support and Health; Social Support and Recovery from Disease and Medical Procedures; Social Support and Stress; Social Support, Psychology of

Bibliography

Antze P 1976 The role of ideologies in peer psychotherapy organizations: Some theoretical considerations and 3 case studies. *Journal of Applied Behavioral Science* **12**: 323–46

Beardslee W R, Salt P, Porterfield K, Rothberg P C, van de Velde P, Swatling S, Hoke L, Moilanen D L, Wheelock I 1993 Comparison of preventive interventions for families with parental affective disorder. *Journal of the American Academy of Child Psychology* **32**: 254–63

Bloom B L, Hodges W, Kern M B, McFadden S 1985 A preventive intervention program for the newly separated: final evaluations. *American Journal of Orthopsychiatry* **55**: 9–26

Bourgeois M S, Schulz R, Burgio L 1996 Interventions for caregivers of patients with Alzheimer's Disease: A review and analysis of content, process, and outcomes. *International Journal of Aging and Human Development* **43**: 35–92

Cohen S 1988 Psychosocial models of the role of social support in the etiology of physical disease. *Health Psychology* **7**(3): 269–97

Coyne J C, Wortman C B, Lehman D 1988 The other side of support: emotional overinvolvement and miscarried helping. In: Gottlieb B H (ed.) *Marshaling Social Support: Formats, Processes, and Effects*. Sage, Newbury Park, CA, pp. 305–30

Dunham P J, Hurshman A, Litwin E, Gusella J, Ellsworth C, Todd P W 1998 Computer-mediated social support: Single young mothers as a model system. *American Journal of Community Psychology* **26**: 281–306

Emshoff J G 1990 A preventive intervention with children of alcoholics. *Prevention in Human Services* **7**: 225–54

Fawzy F I, Fawzy N W, Arndt L A, Pasnau R O 1995 Critical review of psychosocial interventions in cancer care. *Archives of General Psychiatry* **52**(2): 100–13

Fawzy F I, Fawzy N H, Hyun C S, Elashoff R, Guthrie D, Fahey J L, Morton D 1993 Malignant melanoma: Effects of an early structured psychiatric intervention, coping, and affective state on recurrence and survival 6 years later. *Archives of General Psychiatry* **50**(9): 681–89

Gilden J L, Hendryx M S, Clar S, Casia C, Singh S P 1992 Diabetes support groups improve health care of older diabetic patients. *Journal of the American Geriatric Society* **40**(2): 147–50

Goodman G, Pynoos J 1990 A model telephone information and support program for caregivers of Alzheimer's patients. *Gerontologist* **30**(3): 399–404

Gottlieb B H 1998 Support groups. In: Friedman H (ed.) *Encyclopedia of Mental Health*. Academic Press, San Diego, CA, pp. 635–48

Gottlieb B H, Peters L A 1991 National demographic portrait of mutual aid group participants in Canada. *American Journal of Community Psychology* **19**(5): 651–66

Gottlieb B H, Wagner F 1991 Stress and support processes in close relationships. In: Eckenrode J (ed.) *The Social Context of Coping*. Plenum, New York, pp. 165–88

Grych J H, Fincham F D 1992 Interventions for children of divorce: Toward greater integration of research and action. *Psychological Bulletin* **111**(3): 434–54

Helgeson V S, Cohen S 1996 Social support and adjustment to cancer: Reconciling descriptive, correlational, and intervention research. *Health Psychology* **15**(2): 135–48

Helgeson V S, Gottlieb B H 2000 Support groups. In: Cohen S, Underwood L, Gottlieb B H (eds.) *Social Support Measurement and Intervention: A Guide for Health and Social Scientists*. Oxford University Press, New York, pp. 221–45

Hughes R J 1988 Divorce and social support: A review. *Journal of Divorce* **11**(3–4): 123–45

Humphreys K, Klaw E 1998 *Expanding the Self-Help Group Movement to Improve Community Health and Well-being*. Center for Health Care Evaluation, VA Palo Alto Health Care System, Stanford, CA

Humphreys K, Moos R H 1996 Reduced substance abuse-related health care costs among voluntary participants in Alcoholics Anonymous. *Psychiatric Services* **47**(7): 709–13

Humphreys K, Noke J M 1997 The influence of post-treatment mutual help group participation on the friendship networks of substance abuse patients. *American Journal of Community Psychology* **25**(1): 1–16

Kessler R C, Mickelson K D, Zhao S 1997 Patterns and correlates of self-help group membership in the United States. *Social Policy* **27**: 27–46

Lavoie J P 1995 Support groups for informal caregivers don't work! Refocus the groups or the evaluations? *Canadian Journal of Aging* **14**(3): 580–603

Lieberman M A, Videka-Sherman L 1986 The impact of self-help groups on the mental health of widows and widowers. *American Journal of Orthopsychiatry* **56**(3): 435–49

Marmar C R, Horowitz M J, Weiss D S et al. 1988 A controlled trial of brief psychotherapy and mutual-help group treatment of conjugal bereavement. *American Journal of Psychiatry* **145**(2): 203–9

Minde K, Shosenberg N, Marton P, Thompson J, Ripley J, Burns S 1980 Self-help groups in a premature nursery: a controlled evaluation. *Journal of Pediatrics* **96**(5): 933–40

Ostwald S K, Hepburn K W, Caron W, Burns T, Mantell R 1999 Reducing caregiver burden: A randomized psychoeducational intervention for caregivers of persons with dementia. *Gerontologist* **39**(3): 299–309

Riessman F 1965 The helper therapy principle. *Social Work* **10**: 29–38

Reissman F, Carroll D 1995 *Redefining Self-help*. Jossey-Bass, San Francisco

Silverman P R 1986 *Widow-to-widow*. Springer, New York

Spiegel D, Kraemer H C, Bloom J, Gottheil E 1989 Effect of psychosocial treatment on survival of patients with metastatic breast cancer. *Lancet* **2**(8668): 888–91

Stolberg A L, Mahler J 1994 Enhancing treatment gains in a school-based intervention for children of divorce through skill training, parental involvement, and transfer procedures. *Journal of Consulting Clinical Psychology* **62**(1): 147–56

Stroebe W, Stroebe M, Abakoumkin G, Schut H 1996 The role of loneliness and social support in adjustment to loss: A test of attachment versus stress theory. *Journal of Personal Social Psychology* **70**(6): 1241–9

Taylor S E 1983 Adjustment to threatening events: A theory of cognitive adaptation. *Am. Psychology* **38**(11): 1161–73

Toseland R W, Rossiter C 1989 Group interventions to support family caregivers: A review and analysis. *Gerontologist* **29**(4): 438–48

Vachon M L, Lyall W, Rogers J, Freedman-Letofsky K, Freeman S 1980 A controlled study of a self-help intervention for widows. *American Journal of Psychiatry* **137**(11): 1380–84

West S G, Aiken L S, Todd M 1993 Probing the effects of individual components in multiple component prevention programs. *American Journal of Community Psychology* **21**(5): 571–605

Wills T A, Filer M 1999 Social networks and social support. In: Baum A, Revenson T (eds.) *Handbook of Health Psychology*. Erlbaum, Mahwah, NJ, pp. 209–34

B. H. Gottlieb

Suprachiasmatic Nucleus

The suprachiasmatic nucleus (SCN) of the hypothalamus is the principal circadian pacemaker in the mammalian brain and, as such, it generates circadian rhythms in rest and activity, core body temperature, neuroendocrine function, autonomic function, memory and psychomotor performance, and a host of other behavioral and physiological processes. The SCN is the central player in an important neural system, the circadian timing system (CTS).

1. Circadian Rhythms are Adaptations to the Solar Cycle

The physical environment of living organisms presents extraordinarily diverse circumstances requiring complex adaptations. The most prominent, regularly recurring stimulus in the physical environment is the solar cycle of light and dark and, not surprisingly, nearly all species, plant and animal, have evolved daily rhythms in organismal functions and behavior in response to the solar cycle which enhance adaptation and survival. Such rhythms, when they have special properties, to be outlined below, are designated circadian rhythms. Circadian rhythms are found in nearly all living organisms, from prokaryotes through eukaryotes, from simple to higher plants, and simple invertebrates to the human. The nature of these rhythms is diverse but in mammals, the major subject of this article, the most evident expression of circadian function is the sleep–wake cycle. This rhythm, and associated rhythms in physiological and endocrine functions, are expressions of the activity of a specialized neural system, the CTS.

Mammals present two patterns of behavioral adaptation to the solar cycle. In those species in which olfaction and audition dominate the animals' perception of their environment, the rest-activity (sleep–wake) rhythm is 'nocturnal'; the animals are awake at night and sleep during the day. In animals in which vision is the major sensory modality, the rest-activity rhythm is diurnal with activity during the day and rest at night. The nocturnal pattern is typical of rodents and carnivores, for example, whereas the diurnal pattern is exemplified by birds and primates.

2. Circadian Rhythms Have Two Principal Properties

Under normal conditions of the light-dark cycle, circadian rhythms have a period of exactly 24 hours. If photic cues are removed, however, by placing animals in constant darkness or constant light, the rhythms are maintained but with a period that differs from 24 hours. In the first circumstance, in which the period of the rhythm is the same as that of the solar cycle, the rhythm is said to be 'entrained.' When the period differs from 24 hours in constant conditions, the rhythm is termed 'free-running' (Fig. 1). These two properties, entrainment and endogenous maintenance of free-running rhythmicity, provide us with insight into the necessary organization of the CTS. That is, the phenomenon of a free-run in the absence of environmental cues indicates that circadian rhythmicity is generated by a clock-pacemaker within the animal. Entrainment implies that photoreceptors and a visual pathway to the clock-pacemaker are necessary components of the system and the diverse functions under circadian control imply the presence of pathways from the clock-pacemaker to sites expressing effector rhythms (Fig. 1). The effect of circadian regulation is not to alter a function but to provide a temporal organization for its expression.

3. The SCN is the Center of the CTS

3.1 The SCN is an Endogenous Circadian Pacemaker

The SCN is a small, paired hypothalamic nucleus that lies immediately about the optic chiasm and just lateral to the third ventricle in all mammals. Typically it is composed of small neurons that are compactly organized. There are four lines of evidence supporting the conclusion that the SCN is a circadian pacemaker.

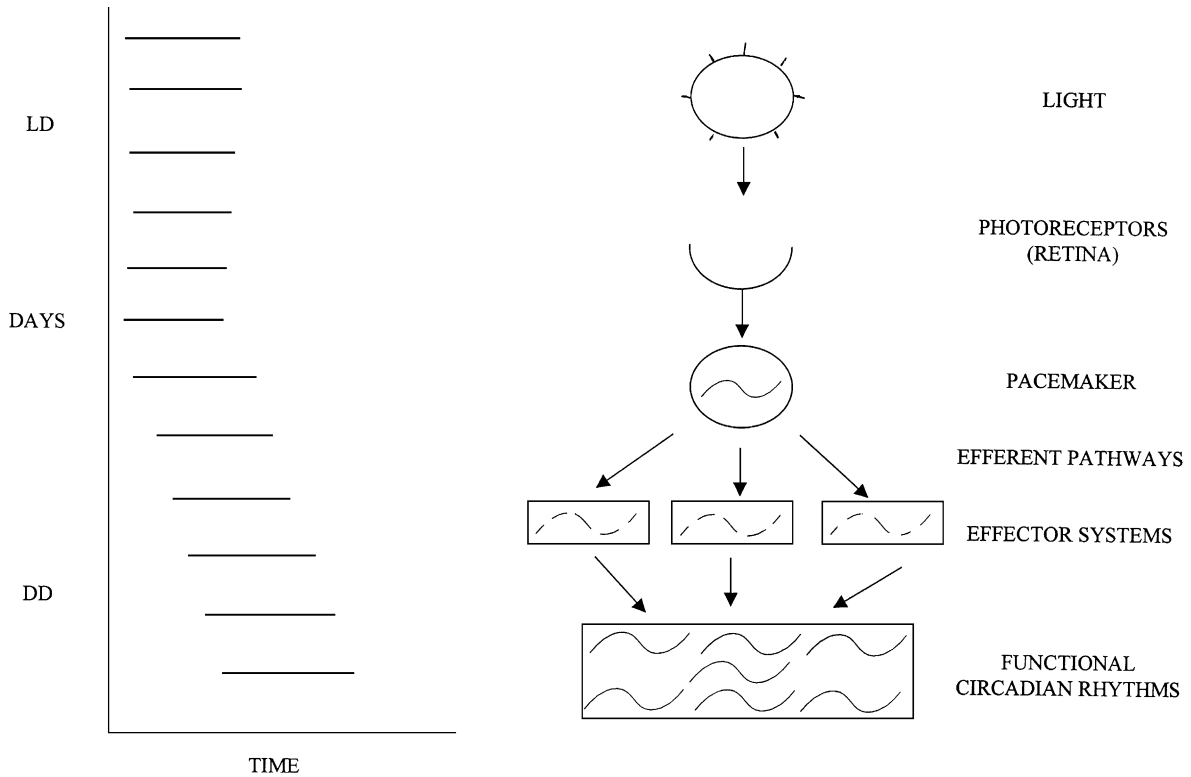

Figure 1
The left panel shows an entrained and free-running rhythm in diagram. Each line represents a day and the length of the line would be the duration of the walking period in a wake–sleep cycle. LD refers to a light–dark cycle and DD to constant darkness. The first six days are entrained in LD and the last six free-running in DD. The right panel is a representation of the circadian timing system with light acting upon retinal photoreceptors to be transduced to neural information to be carried through an entrainment pathway to the pacemaker. The pacemaker generates a circadian signal which is transmitted to effector systems which express circadian regulation of function

3.1.1 The SCN is the site of termination of the retinohypothalamic tract—an essential entrainment pathway. For an explanation of this, see below.

3.1.2 Ablation of the SCN abolishes circadian rhythms. The initial studies of SCN ablation were performed in the early 1970s. Following these lesions, there is a complete loss of temporal organization of functions under circadian regulation. For example, SCN-lesioned animals show an arrhythmic pattern of sleep-wake behavior yet the amount of time spent awake is normal as is the amount of time spent in REM and non-REM sleep. Similarly, the rhythm in adrenal steroid secretion is lost but the total amount of steroid secreted is normal and responses to stress are maintained. Thus, in these and many other functions, the effect of SCN lesions is only to alter the temporal organization of functions, not the functions themselves.

3.1.3 Circadian rhythmicity is maintained in the SCN in isolation in vivo and in vitro. In elegant experiments, Inouye and Kawamura were able to isolate the SCN from the rest of the brain in intact animals and record from SCN neurons in these animals. Although the animals were arrhythmic in behavior, and other brain areas lost rhythms in neuronal firing, SCN neurons maintained a stable rhythm of firing with high levels during the day and low levels at night, and the firing rate rhythm free-runs in constant conditions. Subsequent studies by a number of investigators confirmed these observations in SCNs maintained *in vitro* in the absence of the rest of the brain. This has not been found for other brain areas. Thus, the SCN is a unique brain area in that it main-

tains rhythmicity in isolation both from environmental cues and input from other brain areas.

3.1.4 Circadian rhythmicity is restored in SCN-lesioned animals that receive transplants of fetal SCNs. In arrhythmic animals with SCN lesions, placement of small transplants of fetal hypothalamus containing the SCN into the third ventricle restores the circadian sleep–wake rhythm. The transplants must contain SCN tissue and be placed in the relative vicinity of the lesioned area to be effective. It was unclear until recently how the transplants functioned but work by Silver and Lehman and co-workers, who were early leaders in the transplant studies, showed conclusively that the transplants can communicate with the rest of the brain through a humoral signal released into the cerebrospinal fluid to diffuse into the adjacent brain, hypothalamus and thalamus. These data on transplantation of SCN provide the final, conclusive evidence that the SCN is the principal circadian pacemaker.

3.1.5 SCN Neurons are circadian oscillators. There are two potential mechanisms by which the SCN might generate circadian rhythms. The first is that individual SCN neurons are born as circadian oscillators that are then coupled to make a pacemaker. The second is that circadian function arises as an emergent property of coupled SCN neurons which are not themselves oscillators. The SCN is formed embryologically from precursor cells located in the diencephalic germinal epithelium. As these precursor cells divide, the newly formed neurons migrate from the germinal epithelium to form the SCN. In the rat, the most extensively studied species, SCN neurons are formed in late prenatal life with the process completed by three days before birth. At 2 days before birth, the SCN is very primitive and contains no synapses, indicating that its neurons do not participate in a neural network. Nonetheless, a circadian rhythm in glucose utilization appears in the SCN on that day, and this is not generated by maternal influences, or other fetal influences, indicating that SCN neurons have become rhythmic and, hence, at least a subset must be individual circadian oscillators. Subsequently, the neurons are coupled through neural connections to make a pacemaker. The validity of this interpretation of SCN organization is shown by other studies in which individual SCN neurons, isolated in tissue culture, exhibit circadian rhythms in firing rate. This type of pacemaker organization is also consonant with observations on invertebrate pacemakers.

3.1.6 Circadian function in the SCN is genetically determined and generated through protein interactions. Data presented over many years have established that circadian rhythmicity is a genetically-determined function. Among the studies contributing to this understanding are experiments in Neurospora and Drosophila that have indicated the existence of clock genes. Beginning in the mid-1980s, clock genes were cloned and analyzed, largely by Dunlap and his co-workers for Neurospora and by Hall, Rosbash, and Young for Drosophila. This work shows that circadian rhythmicity, at the molecular-cellular level, is the expression of clock genes producing proteins that, in turn, regulate expression of the clock genes, that is, the clock mechanism can be viewed as a simple feedback loop. This field is evolving very rapidly at present and recent findings indicate that there is a substantial conservation of function from invertebrates to mammals. Homologues of the Drosophila clock genes are present in the mammalian SCN and cycle in a pattern similar to that in Drosophila. This work will continue to expand rapidly and we should expect a reasonably complete understanding of the molecular-cellular events underlying circadian rhythmicity in the mammalian SCN to be elucidated in the near future.

3.2 The SCN is Coupled to Effector Systems by Neural Pathways

The major output of the SCN is to hypothalamus and thalamus with very limited projections to basal forebrain. Within this context, the densest projections are to anterior hypothalamus for the control of body temperature and autonomic function and to the paraventricular and arcuate nuclei for neuroendocrine regulation and to the posterior hypothalamus and thalamus for behavioral state regulation. Thus, these projections form the basis circadian control of effector systems. A scheme for this is shown diagrammatically in Fig. 2.

4. The SCN in Nonmammalian Vertebrates

There appears to be an SCN, or an SCN homologue, in the brains of all vertebrates. The function of the SCN, however, differs in nonmammalian vertebrates in comparison to mammals. As described above, the SCN is the circadian pacemaker responsible for the temporal organization of behavior into 24-hour sleep-wake cycles in virtually all mammals. Relatively little work has been done on fish, amphibia, and reptiles but birds have been studied extensively. In avian species there are three patterns of organization of the circadian timing system. The first, represented by the domestic chicken, is somewhat similar to mammals. The SCN is the pacemaker responsible for the temporal organization of behavior but the pineal is

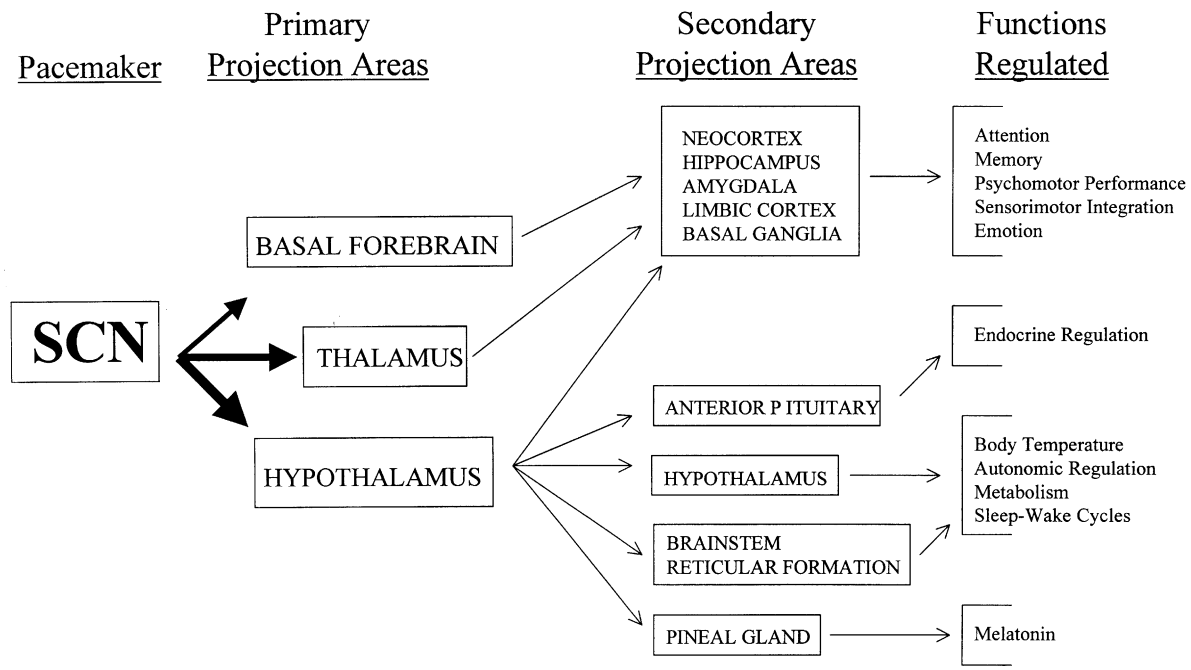

Figure 2
Diagrammatic representation of the functional organization of SCN control of effector system

capable of independent circadian activity. The second, represented by passerine species, differs in that the pineal is the principal pacemaker but the pineal acts through effects on the SCN, that is, if the pineal is removed from animals in constant conditions, the animals become arrhythmic and transplantation of the pineal to the anterior chamber of the eye restores rhythmicity. If the SCN is ablated in animals with intact pineals, behavioral rhythmicity is lost even though the pineal continues to put out a rhythmic signal. The third variation, seen in some quail species, is characterized by three structures, the pineal, SCN, and retina being capable of generating rhythmicity. Thus, in avian species, the SCN, the pineal or the retina are potential circadian pacemakers. This is of interest in that recent data indicate that the mammalian retina is capable of generating intrinsic circadian rhythms and there is some evidence that mammals can exhibit a very limited amount of circadian function in the absence of the SCN. This is only evident in specialized environmental situations and, under usual circumstances, SCN-lesioned animals do not recover circadian function and remain arrhythmic throughout their lives.

5. *Conclusions*

The adaptation of animals to their environment is characterized by a temporal organization of behavior into cycles of rest (sleep) and activity. In mammals, this temporal organization of behavior, facilitated by a precise temporal organization of physiological, neural, and endocrine processes is established by a small, paired hypothalamic nucleus, the suprachiasmatic nucleus. The suprachiasmatic nucleus is comprised of individual neuronal oscillators coupled to constitute a network that functions as a circadian pacemaker. The precise output of the pacemaker is modulated by photic and nonphotic neural afferents to provide maximal adaptation. This output acts predominantly through connections to hypothalamus and, to a lesser extent, to thalamus and basal forebrain. Although limited, these connections orchestrate the precise pattern of behavioral and physiological processes that represent the circadian contribution to adaptation.

See also: Circadian Rhythms; Hypothalamus; Sleep and Health; Sleep Disorders: Psychiatric Aspects; Sleep Disorders: Psychological Aspects; Sleep States and Somatomotor Activity; Sleep: Neural Systems

Bibliography

Hastings M H 1997 Central clocking. *Trends in Neuroscience* **20**: 459–64

Klein D C, Moore R Y, Reppert S M 1991 *Suprachiasmatic Nucleus: The Mind's Clock*. Oxford University Press, New York

Meijer J H, Rietveld W J 1989 Neurophysiology of the suprachiasmatic circadian pacemaker in rodents. *Physiological Reviews* **69**: 671–707
Moore R Y 1996 Entrainment pathways and the functional organization of the circadian system. *Progress in Brain Research* **111**: 103–19
Moore R Y 1997 Circadian rhythms: Basic neurobiology and clinical applications. *Annual Review of Medicine* **48**: 253–66
Moore R Y, Leak R K 1999 Suprachiasmatic nucleus. In: Takahashi J S, Moore R Y, Turek F W (eds.) *Biological Rhythms*. Plenum Press, New York

R. Y. Moore

Suprasegmentals

Suprasegmental features of speech are associated with stretches that are larger than the segment (whether vowel or consonant), in particular *pitch*, *stress*, and *duration* (Lehiste 1970). Respectively, these terms refer to the sensation of higher and lower tone, to the prominence patterns of words, and to durational differences between segments, whether between different segments or between pronunciations of the same segment in different contexts. In general, speech can be analyzed from two points of view: the phonetic reality, most readily accessible as the acoustic speech wave form, but equally present in the articulatory behavior of the speaker and the perceptual analysis of the auditory signal by the listener, and the discrete mental representation which serves as the set of instructions from which the phonetic signal is built by the speaker. One way of structuring a discussion of this topic is to take these three suprasegmental features as a point of departure, consider their somewhat varied manifestations at the different stages of the speech chain, that is, their articulation, acoustics, and perception, and tabulate the ways in which they function in languages. This approach would reveal that the terms pitch, stress, and duration do not refer to concepts at a single stage in the speech chain. Stress is largely a phonological notion (cf. section on the foot). Pitch is a perceptual concept, the articulatory correlate of pitch being rate of vocal cord vibration, and the acoustic correlate, in the usual case, being periodicity in the speech signal, the repetition of nearly the same pattern of vibration. Another approach takes the phonological representation as the starting point, and refer to the phonetic features that are involved in their realization, which is what will be done here.

Recently, the phonological object of study that corresponds to the traditional suprasegmental features has come to be known as *prosody*. Prosodic structure has two components. First, there is a hierarchical compartmentalization of the segment string into phonological constituents, like the syllable and the intonational phrase, called the *prosodic hierarchy* (Selkirk 1978, Nespor and Vogel 1986, Hayes 1989). Second, overlaid on the string of segments, there is a *tonal structure*, which is synchronized in specific ways with the segmental and prosodic structure (Pierrehumbert and Beckman 1988, Ladd 1996). All languages have both of these structural components.

1. The Prosodic Hierarchy

The phonological structure of an English sentence like *Too many cooks spoil the broth* does not just consist of a linear segment string [tuː meni kʊks spɔɪl ðə brɔθ]. The segments are grouped in a hierarchical set of constituents, where constituents at each rank include those of the rank below. There are seven syllables and five feet (*Too*, *many*, *cooks*, *spoil*, and *broth*), each of which is also a phonological word; next, there are three phonological phrases (*Too many cooks*, *spoil*, and *the broth*), two intonational phrases (*Too many cooks* and *spoil the broth*), and one utterance (*Too many cooks spoil the broth*).

The *foot* captures the notion stress. It is a constituent which in the unmarked case consists of a stressed (or strong: S) and an unstressed (or weak: W) syllable, with languages choosing between *trochees* (S–W) and *iambs* (W–S). Often, there are restrictions on segmental contrasts in weak syllables, as in Russian, which has five vowels in stressed syllables but only three in unstressed syllables, and in English, which, depending on the variety, has between 13 and 17 monophthongs and diphthongs in stressed syllables, but only three in unstressed syllables (cf. the final syllables in *villa*, *folly*, and *fellow*; Bolinger 1986, p. 347ff.). Commonly, feet are disyllabic. English, a trochaic language, also has marked feet: a monosyllabic foot occurs in *cooks* and *manifest*, where *-fest* occurs after the disyllabic foot *mani-*; a trisyllabic foot occurs in *cavity* and *university*, where the foot *-versity* occurs after the disyllabic foot *uni-*. One of the feet of a word is selected to bear *primary stress* or *main stress* ['], the others having *secondary stress* [ˌ]. In English, the primary stress is on the rightmost branching foot, as shown by *'mani ˌfest*, *uni'versity*, and *ˌApaˌlachi'cola*, but there are many exceptions, like *'salaˌmander*, *ˌperso'nnel*. Other languages will make other choices. For instance, Bengali, Czech, Finnish, Icelandic, and Pintupi (Australia) always have the primary stress on the first trochee, that is, all words have initial stress, and secondary stresses may be found on alternate syllables, as in Pintupi ['jumaˌɽiŋkaˌmaraˌtʲuɽaka] 'because of mother-in-law.' Foot structure assignment may take place *left-to-right*, as in Pintupi, where odd-syllabled polysyllabic words end in trisyllabic feet, or *right-to-left*, and instead of a trochee, the foot may be an iamb. In

quantity-sensitive languages, long vowels, or long vowels and closed syllables, do not tolerate being in a weak position, and will thus interrupt any regular count through the word. Hawaiian is a quantity-sensitive trochaic language which assigns feet right-to-left, so that [ko'hola] 'reef' has penultimate stress, but [ˌkoho'laː] 'whale' has final stress. When a language has *extrametricality*, the regularities obtain minus an edgemost (usually final) syllable. Finally, within limits imposed by the typology of the language, some words may have exceptional stress, in which case the foot structure is marked in the underlying representation. English is a quantity-sensitive, trochaeic, right-to-left language with extrametricality and exceptions (cf. Hayes 1985).

Stressed positions in phonological representations not only attract more phonological contrasts, as the English and Russian vowel systems show, but also lead to less casual articulation than unstressed positions. The intensity at the higher end of the frequency spectrum tends to fall off less steeply (even *spectral balance*), spectral targets of vowels are approached more closely (less *vowel reduction*), and their duration is longer; overall intensity measures like RMS intensity appear to be unreliable indicators of stress (e.g., Beckman 1986, Sluijter 1995). Stressed syllables also have different pitch characteristics from unstressed syllables. These are caused by the fact that stressed syllables attract *intonational pitch accents*, tones, or tone complexes, which (together with boundary tones, see below) make up the tonal structure of the language. The presence of a pitch accent on a stressed syllable makes that syllable *accented*, as indicated by capitalization. When said in isolation, an English word like *theory* is pronounced *THEory*, while there may also be a pitch accent on a secondary stress before the primary stress, so that *Japanese* will be *JapaNESE* or *JAPANESE*. English has a number of different pitch accents, like H*, L*, L*H, etc., where the starred tone is time-aligned with the stressed syllable. While in isolation, the pronunciation will be *THEory*, in *SET theory* the syllable *the-* is unaccented, because in compounds only the first constituent is accented. Other reasons for *deaccentuation* are stress shift and lack of focus (see below).

A frequently applied diagnostic for the *phonological word* is syllabification, more specifically the extent to which consonants syllabify as onsets with vowels to their right. The phonological word may be smaller or larger than the morphological word. In many languages, it is smaller in compounds, where each constituent is a separate syllabification domain (e.g., English *cat's eye* syllabifies as ['kæts.ˌaɪ], not as *['kæt.ˌsaɪ]); it is larger in cases of cliticization (e.g., English *What is he* ['wɔt.si] *saying?*). The *phonological phrase* may determine the distribution of accents, favoring first and last accents. Although *TOO MANY*! will have two accents, *Too many COOKS* is likely to lose the pitch accent on *many*.

The term *stress shift* is applied to cases where such rhythmic deaccentuation leads to the sole presence of a pitch accent on a foot with secondary stress, as in *JAPANESE*, but *JAPANESE FURniture* (Gussenhoven 1991). The *intonational phrase* typically comes with *boundary tones* at its edges. For instance, the intonational phrase *Too many cooks* might end with a H-tone, pronounced after the pitch accent on *cooks*. Boundaries of this type are often felt to correspond with a comma in writing. The utterance, finally, roughly corresponds with the intuitive notion of spoken sentence. Instead of, or in addition to, the intonational phrase, the phonological phrase and the utterance may also be provided with boundary tones.

In (1), a full prosodic representation is given, together with frequently used symbols; the subscript *I* in the tonal structure indicates a boundary tone.

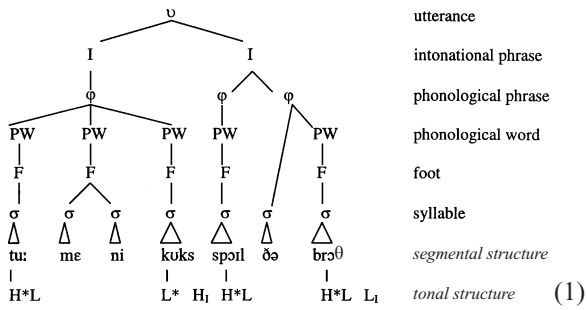

(1)

The choice of tones is one of many possible combinations, each of which will produce a different *intonation contour*. At the higher ranks, some optionality also exists in the phrasing, depending on style, constituent length, and sometimes language variety. Thus, *spoil the broth* may form a single phonological phrase, as shown by varieties like American English which allow stress shift in verb-object combinations, like *MAINtain ORder* (cf. *ORder to maintain*). The sentence can also be pronounced as a single intonational phrase.

The constituents in the prosodic hierarchy often define the context of segmental restrictions. For instance, syllable-final obstruents are voiceless in Turkish; voicing of Dutch [s] occurs between a sonorant segment and a vowel if [s] occurs at the end of a prosodic word; in Kimatuumbi, all vowels are shortened except any long vowel that ends up in the last word in the phonological phrase. Higher rank prosodic constituents in particular tend to undergo (phonetic) *preboundary lengthening*, which is usually confined to the final syllable.

There is no consensus on how many prosodic constituents there are, in part due to apparent differences between languages. For instance, Japanese has an *accentual phrase*, which seems to lie halfway

between PW and ϕ. A *clitic group* has been claimed for Italian and English, for instance, providing a level between PW and ϕ. A constituent below the syllable is the *mora*, which captures quantity contrasts. A syllable with a short vowel has one mora, one with a long vowel or diphthong two. A geminate consonant adds one mora, a single consonant none. Disyllabic words in Finnish, which has quantity contrasts in vowels and consonants in stressed syllables, may have two moras (CVCV), three (CVVCV or CVCCV), or four (CVV-CCV). *Exhaustive parsing* refers to the ideal situation in which each constituent includes all and only the segments included in the constituents of the rank below. For instance, the inclusion of [m] in the syllable [mɛ] implies that it is also included in the foot [mɛni] and the phonological word [mɛni] (which makes the structure distinct from that of *Tomb any*). A violation occurs in (1) for the syllable *the*, which lacks the normal properties of F and PW, and is directly included in the ϕ.

2. Tonal Structure

The tonal structure can be thought of as a string of tones arranged in parallel to the segmental structure provided by the string of consonant and vowels, that is, as an *autosegmental* representational tier (Goldsmith 1976). It varies greatly across languages. There are two ways in which a typology can be set up, a morphosyntactic one ('How do tones signal what words, phrases, or clause types are used?') and a phonological one ('Where are tones located and how many contrasting tones are there?). A frequently applied morphosyntactic criterion is whether the tone(s) enter into the specification of (also segmentally specified) words or morphemes. If they do, the language is a *tone language*, examples of which can readily be found in Asia (Sino-Tibetan languages, Japanese), Africa (Bantu languages, Khoi-San languages, Nilo-Saharan languages), and many languages in the Americas. Examples of tonal distinctions ([ˋ] = low, [ˊ] = high, [¯] = mid) of this kind are Japanese [hásì] 'chopsticks,' [hàsí] 'bridge,' [hàsī] 'end,' Mandarin (Sino-Tibetan) [má:] 'mother,' [mǎ:] 'hemp,' [mà:] 'horse,' [mâ:] 'vituperate,' Mamvu [máàkà] 'type of seasoning' vs. [màákà] 'cat.' Frequently, tone is involved in the morphology in particular in languages spoken in Africa, as shown by Makonde [nìndì tálèékà] 'I cooked' vs. [nìndà tàlèékà] 'I will cook' and Mamvu [áfù] 'man,' [àfú] 'men' (all examples from Kaji 1999). About half of the languages in the world fall in this class (Maddieson 1978). If they do not use tone to determine the identity of morphemes, they are not tone languages; sometimes non-tone languages are referred to as *intonation languages*, but this term incorrectly suggests that tone languages do not use tone intonationally.

In nontone languages, tones are used to indicate discourse meanings, like 'interrogativity,' 'finality,' 'continuation,' etc. This is illustrated by English *There's no cutlery*: when *cutlery* is spoken with a low-to-high rise, the sentence is meant as a question; with falling-rising pitch it could either be a reminder or again a question, while a high-to-low fall signals a statement of fact. Many tone languages use tone in this intonational way, like Japanese, which distinguishes a L-tone in statements from a H-tone in questions utterance-finally. By contrast, some languages, like the Western Aleut dialect of Unangan, would appear to restrict their tonal structure to the insertion of specific initial and a final boundary tones at the beginning and end of every intonational phrase, and perhaps one more at the end of the utterance. Here, the only function of tone is to mark the phrasing. There are thus widely varying functional claims made by the phonologies of different languages on pitch variation. Not all acoustic variation, however, is due to differences in phonological representation. Speakers of all languages will also use pitch variation to signal universal meanings. Variations in overall pitch range, for example, may signal such general meanings as surprise, interest, emphasis, etc., many of which have an ethological origin (Ohala 1994).

A phonological typology of tone might be based on tonal density: how many locations are specified for tone, that is, have tonal *associations*? In other words, languages differ in the selection of the *tone bearing unit*: in the 'densest' case they specify every mora for tone, and in the sparsest case they just mark the phrasing. When more than one tone is associated with a syllable, like HL or LH, the combination is known as a *contour tone*, while a single tone is a *level tone*.

After edges of prosodic constituents, stressed syllables rank as a popular location for tone. In (1), all tones are intonational. The Venlo Dutch sentence (2a) also has intonational boundary tones and an intonational pitch accent (H*), but the second H-tone on the word for 'play' is a lexical tone, whose presence makes this word distinct from (the segmentally identical) word for 'rinse,' given in (2b). The difference comes out in that the syllable [spϕ:] has high pitch in (2a) but falling pitch in (2b), where the boundary tone spreads to the second mora of the stressed syllable. Thus, although (2a) and (2b) differ in pitch, they have the same intonation.

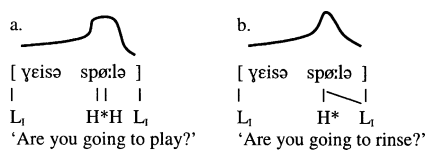

(2)

Languages that, like the Dutch dialect of Venlo, restrict the lexical tonal contrasts to one per word are

called *pitch accent languages*. (Observe the different meaning of 'pitch accent' from that given earlier.) Most of these use the syllable with word stress for this location (*metrical pitch accent languages*), but the location may also be arbitrary, as is in Japanese (*lexical pitch accent languages*), where the HL-tone complex is located on one of the syllables of accented words. (Unaccented words are unaccented because they—again arbitrarily—do not have this tone complex on any of their syllables.) In Japanese, the accented syllable has none of the phonetic characteristics of stressed syllables (Beckman 1986). Metrical pitch accent languages can also be found in Europe (Norwegian, Swedish, Central Franconian dialect of German, Limburgian dialect of Dutch, Serbocroatian dialects, Basque dialects, Lithuanian; cf. van der Hulst 1999).

The number of tonally specified locations is usually greater than the number of tones. For instances, moras may be specified by a *spreading* tone: when a tone spreads (left or right), it specifies more than one tone bearing unit. Also, suffixes may have a tone with the opposite value from the tone of the stem (*polar rule*), etc. Moras, syllables, or words that are unspecified for tone are pronounced with a pitch that results from an *interpolation* between the two pitches specified at the nearest locations to the right and to the left. If a language just has an initial L-tone and a final H-tone in the intonational phrase, for instance, the pitch in the phrase may begin low, and rise towards the end. In fact, the pitches indicated in the unaccented Japanese word for 'end' are due to the interpolation between an initial LH tone complex of the accentual phrase and a final L-tone of the utterance, and both pitches are thus due to intonational tones, the mid pitch on the second syllable being a curtailed interpolation between H and L.

A second typological question concerns the number of *contrasting tones*. While most languages, and most probably all nontonal languages, contrast H with L, some languages have three or even four contrasting level tones.

Both the prosodic hierarchy and the autosegmental tonal structure are *nonlinear* forms of representations, differing radically from the linear segment string given at the beginning of Sect. 1.

3. The Expression of Focus

Languages frequently exploit prosodic structure for the expression of *focus*, the marking of information status in sentences. Languages appear to vary in the kind of information status which is marked (counter-presupposition, counter-assertiveness, topichood), but a common operationalization of focus is the answer to a question. Thus, if *Sue doesn't like cheese* is an answer to *Who doesn't like cheese?*, the focus is *Sue*, since it expresses the information the question seeks to elicit, while the rest of the sentence is outside the focus. However, if it is an answer to *Why are you changing the menu?*, the whole sentence is focus. The latter type is called *broad focus*, sentences with unfocused constituents having *narrow focus* (Ladd 1996). Prosodic encoding of focus distribution may take several forms. One is phrasing, in which case the beginning or end of a medium-rank prosodic constituent coincides with, respectively, the beginning or end of the focus constituent (e.g. Japanese, Bengali). Some languages use different types of intonational pitch accent for focused and non-focused parts of sentences (e.g., Italian), while others deaccent non-focused parts (Germanic), or de-accent *and* use different pitch accents for broad and narrow focus (Bengali) (Hayes and Lahiri 1991). Lastly, accentuation may signal other meanings than focus. If sentence (1) were to lose its accent on *spoil*, it would lose its implicit conditional status ('If there are too many cooks …') and imply that the number of cooks that actually end up spoiling the broth is too large.

See also: Linguistics: Overview; Phonology; Speech Perception; Speech Production, Psychology of; Speech Recognition and Production by Machines

Bibliography

Beckman M E 1986 *Stress and Non-Stress Accent*. Foris, Dordrecht, The Netherlands

Bolinger D 1986 *Intonation and Its Parts: Melody in Spoken Language*. Stanford University Press, Stanford, CA

Goldsmith J 1976 Autosegmental phonology. Doctoral dissertation, MIT

Gussenhoven C 1991 The English Rhythm Rule as an accent deletion rule. *Phonology* **8**: 1–35

Hayes B 1989 The prosodic hierarchy in meter. In: Kiparsky P, Youmas G (eds.) *Rhythm and Meter*. Academic Press, Orlando, FL, pp. 201–60

Hayes B 1995 *Metrical Stress Theory*. University of Chicago Press, Chicago

Hayes B, Lahiri A 1991 Bengali intonational phonology. *Natural Language and Linguistic Theory* **9**: 47–96

Hulst H van der 1999 *Word Prosodic Systems in the Languages of Europe*. Mouton de Gruyter, Berlin

Hyman L M in press Tone systems. In: Haspelmath M, König E, Oesterreicher W, Raible W (eds) *Language Typology and Language Universals: An International Handbook*. Mouton de Gruyter, Berlin

Kaji S 1999 (ed.) *Proceedings of the Symposium on Cross-linguistic Studies of Tonal Phenomena: Tonogenesis, Typology, and Related Topics*. Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA), Tokyo University of Foreign Studies

Ladd D R 1996 *Intonational Phonology*. Cambridge University Press, Cambridge, UK

Lehiste I 1970 *Suprasegmentals*. MIT Press, Cambridge, MA

Maddieson I 1978 Universals of tone. In: Greenberg J (ed.) *Universals of Human Language, Vol. 2: Phonology*. Stanford University Press, Stanford, CA, pp. 335–65

Nespor M, Vogel I 1986 *Prosodic Phonology*. Foris, Dordrecht, The Netherlands
Ohala J J 1994 The frequency code underlies the sound-symbolic use of voice pitch. In: Hinton L, Nichols J, Ohala J J (eds.) *Sound Symbolism*. Cambridge University Press, Cambridge, UK
Pierrehumbert J, Beckman M E 1988 *Japanese Tone Structure*. MIT Press, Cambridge, MA
Selkirk E O 1978 On prosodic structure in relation to syntactic structure. In: Fretheim T (ed.) *Nordic Prosody*, Vol. 2. TAPIR, Trondheim, Norway, pp. 111–40
Sluijter A 1995 *Phonetic Correlates of Stress and Accent*. Foris, Dordrecht, The Netherlands

C. Gussenhoven

Supreme Courts

Supreme courts are courts of final appeal within national judicial systems. There is a degree of ambiguity about which institutions properly may be called both a 'court' and 'supreme.' Civil law countries have, effectively, three (or more) separate court systems covering different areas of the law, with separate institutions at the head of each, of which only the supreme civil court is called a 'court.' Federal systems commonly have, along with a national supreme court, separate supreme courts for their states or provinces. The development of supranational courts (e.g., the European Court of Justice), further complicates the matter, as these have become the courts of final appeal on some matters. Among these many institutions that arguably function, in one way or another, as 'supreme courts,' this article focuses on national, not supranational or sub-national, courts, and on courts whose cases arrive principally by way of appeal from lower courts.

Over the latter part of the twentieth century scholarly interest in these courts has grown significantly due to several developments. The growing capacity of rights-advocacy and other organizations to pursue sustained litigation, growing complaints about bureaucratic abuses, and a weakening of centralized governmental power have encouraged many supreme courts, in a wide array of countries, to intervene increasingly frequently and actively in the policy process (Epp 1998, Tate and Vallinder 1995). At the same time, many observers have contended that strong and independent supreme courts play a key role in fostering the conditions for constitutional democracy, the rule of law, and the protection of individual and minority rights (Henkin and Rosenthal 1990, Stotzky 1993), as well as property rights and economic development. These are heavy burdens to bear, and the capacity of supreme courts for achieving a wide range of policy goals is often exaggerated. Although supreme courts are not without significant influence, their power is constituted and thus also limited by their countries' particular political, social, and economic structures. Nonetheless, the conditions favoring active policy intervention by supreme courts show no signs of abating, and scholarly attention to them is likely to increase.

1. Institutional Structure

1.1 The Main Types

Supreme courts take a wide variety of forms, and are related to other political institutions in a wide variety of ways, but nonetheless fall within two main types reflecting the broad division between the common law and civil law traditions (see *Legal Systems, Classification of*; *Common Law*; *Civil Law*). In common-law countries, a single national supreme court has general jurisdiction over a wide range of cases in virtually all areas of the law arising in lower courts. These courts are frankly recognized as law-making bodies (through their development of caselaw), and traditionally have been valued for their roles in checking executive discretion and lending coherence and consistency to their country's caselaw. They exercise these roles almost exclusively in cases arising out of concrete disputes between two or more parties; the losing party in a lower court may appeal the decision to a superior court, and a small proportion of these appeals ultimately reaches the country's supreme court where review, in theory, is confined to disputes over law (and not facts) (see *Appeals: Legal*). This model is found today in countries throughout the former British empire, particularly in the United States, Canada, Australia, India, and in Africa, particularly Ghana, Kenya, Nigeria, as well as South Africa.

In some common-law countries, particularly the USA and countries following its model, the supreme court possesses the power of judicial review to strike down legislation as unconstitutional, but typically only incidentally to resolving disputes between two or more parties (a system known as 'concrete review') (see *Judicial Review in Law*). In other common-law countries—those with parliamentary sovereignty, particularly Britain—the supreme court has no such power but nonetheless exercises relatively broad authority to check administrative discretion. Following the enormous symbolic influence of the US Supreme Court's attack on racial segregation and its creation of new individual rights in the 1950s and 1960s, some supreme courts in common-law countries, particularly Canada and India, shifted markedly from the British to the American model by increasingly exercising the power of judicial review, particularly in the area of individual rights (Epp 1998; Knopff and Morton 1992). Under the growing influence of European law, even the Appellate Committee of the House of Lords,

Britain's supreme court, has made decisions that appear to be edging toward a frank exercise of judicial review, a trend that is likely to be reinforced as Britain's new bill of rights gains authority.

Civil law systems, by contrast, have been heavily influenced by the French Revolution's efforts to limit the power of courts and to foreclose the development of an independent, powerful, and active judiciary headed by a single supreme court. Consequently, the supreme judicial power in many civil law countries is divided into three (or more) distinct institutions, each with a specialized jurisdiction (Jacob et al., 1996, Merryman 1985). Thus, in France, the ordinary judiciary, which is headed up by the Court of Cassation, traditionally has been strictly forbidden from exercising judicial review or in any other way challenging legislative or administrative acts. The value of some mechanism for checking administrative acts was quickly recognized, however, resulting in the creation of a separate administrative court, the Contentious Section of the Council of State (hereafter, the Council of State), which now functions as the supreme court for the administrative court system. With the growing interest in constitutionalism after World War II, many civil law countries added a separate constitutional court with the power of judicial review over legislation, reflecting the belief that the ordinary supreme court was insufficiently statesmanlike, too conservative, or too tied to a prior regime, to give meaningful effect to constitutional provisions (Cappelletti 1989, p. 145, Favoreu 1990, Merryman and Vigoriti 1967) (see *Constitutional Courts*). In countries in which the constitutional court decides challenges to existing laws, that court and the ordinary supreme court have at times developed sharply conflicting interpretations of the validity of particular laws, which has led to significant tensions between these courts (see e.g., Krug 1997, Merryman and Vigoriti 1967). The civil law model—with multiple courts of last resort—has been adopted throughout much of Europe, many parts of Africa, the Middle East, and Latin America.

The strength of the pressures favoring active supreme court participation in the policy process are evident in the history of the French system. The French Court of Cassation originated as a legislative tribunal with power only to quash incorrect judicial interpretations of law, but over time it adopted a forthright appellate function and eventually was formally converted into the ordinary court system's supreme court; it has since engaged in significant judicial policy-making (Merryman 1985); (for parallels in Britain, see Stevens 1978). Similarly, although the Council of State is not a part of the regular judiciary, it has evolved to a position of great independence within the administrative branch, adjudicates the legal claims of individuals against the state bureaucracy, exercises judicial review over administrative regulations (which constitute a substantial proportion of French law), and has served as a model for supreme courts elsewhere (Brown and Bell 1993, Hill 1993, Merryman 1985). Although formally there is no doctrine of *stare decisis* in civil law systems, in practice the decisions of the Court of Cassation and the Council of State gain wide adherence in lower courts and constitute a body of 'jurisprudence' that is similar in many ways to caselaw (Merryman 1996).

The institutional structures of supreme courts in Latin America mix elements of the US and French models (see e.g., Barker 1986). Thus, most Latin-American countries have a single supreme court with the power of judicial review and with general jurisdiction over appeals from a variety of ordinary courts, but many deny to the inferior courts the power of judicial review, and even the supreme court's power of judicial review in most of these countries is limited to suspending the application of a law in a particular case, not invalidating the law itself. Some countries in the region follow the French model by denying the supreme court jurisdiction over the administrative courts; some countries' supreme courts hear constitutional issues that are referred by the legislature, and Columbia has both a regular supreme court and a separate constitutional court.

1.2 Agenda-setting, Workload, and Internal Structure

The workload of many supreme courts has increased tremendously in recent years, providing significant opportunities for creative judicial policy-making (McWhinney 1986). To better manage rising caseloads, some countries have created intermediate appellate courts and have granted their supreme court significant discretion over which cases to decide. Such reforms acknowledge and enhance the policy-making role of the supreme court, and courts that have this discretion typically use it to expand their policy-making activities by building a coherent agenda that commonly has focused on constitutional issues (Kagan et al. 1977, Pacelle 1991). Granting the supreme court discretion over its judicial agenda is more typical of common law than civil law systems, as appellate review traditionally has been viewed in the former as an extraordinary resort, whereas in the latter it is viewed as a regular aspect of the judicial process (Damaska 1986, Shapiro 1981). Countries that have not granted discretion over case selection to their supreme courts—for instance, India and France—are virtually forced by rising caseloads to expand the number of judges on the court; this is often accompanied by a division of the court into several panels (or 'chambers' or 'benches') specializing in different areas of the law.

Supreme courts differ, too, in how decisions are issued. The decisions of common-law supreme courts typically report significant disagreements among the

deciding justices. Some, like the supreme courts of Canada and the United States, issue a single majority opinion along with dissenting opinions; others report separate ('*seriatim*') opinions written by each of the justices. Civil law courts, by contrast, present a public face of consensus that conceals any politically motivated disputes among the judges. Modes of publication differ as well: generally, the larger the court's caseload, the smaller the percentage of decisions that is published. Although the most important decisions typically are published, a body of research shows that in lower courts decision-making patterns in published and unpublished decisions differ significantly.

These various internal structures provide senior supreme court justices with different opportunities and constraints for manipulation of the decision agenda. Thus, in courts that are divided into separate panels, the chief justice may influence outcomes in particular cases through exercise of authority to assign justices and cases to particular panels. In courts that meet as a whole, as in the USA, the chief justice or the senior justice in the majority similarly may influence case outcomes by exercising authority to assign the responsibility for writing the majority opinion (see e.g., Epstein and Knight 1998, pp. 125–35).

2. *Supreme Courts in the Political Process*

2.1 *Political Influences on Supreme Courts and their Legitimacy*

Supreme courts are political institutions and are influenced by the political process; but the ways in which they are political, and the nature and paths of political influence vary considerably. Although countries have adopted a wide array of constitutional and statutory provisions aimed at striking a balance between judicial independence from undue political manipulation and judicial responsiveness to widespread popular values, the effects of these measures are strongly shaped by several key factors related to the broader structure and distribution of political and organizational power in each society.

First, the more centralized is governmental power, particularly in the executive, the greater is the degree of direct manipulation of the supreme court by officials in these other institutions. At one extreme, military coups or the exercise of emergency powers by governing regimes are commonly associated with sharp limitations on the authority of a country's supreme court and direct manipulation of the court by the regime (Tate 1993). At the other extreme, in countries in which governing power is greatly divided either by a federal structure or an American-style separation of powers, or both, the paths of influence on the supreme court are more attenuated, affording the supreme court greater room for maneuver. Between these ends of the continuum lie many intermediate instances. The supreme courts of Latin-American republics and Japan, where political power is highly centralized, rarely challenge governmental policies or intervene in the policy process (Itoh 1989, Stotzky 1993, Verner 1984). Nonetheless, even supreme courts in pluralistic systems—e.g., the United States and Canada—rarely challenge directly the key policies favored by the national governing coalition.

Second, political influences on supreme courts are conditioned by the degree of pluralistic conflict over selection of justices and the degree of diversity in the pool from which justices are drawn (see *Judicial Selection*). At one extreme, in a pattern more typical of common law than civil law countries, are courts whose appointment processes are subject to open political conflict between competing factions and whose justices are drawn from a relatively diverse population that includes politicians, government officials, and law professors. Thus, in the United States, the constitutional requirement for Senate confirmation of nominations made by the President has afforded the opportunity for open, pluralistic conflict over appointments to the Supreme Court (Maltese 1995). At the other extreme are systems that insulate appointment processes from popular pressure and which draw judges from an insular elite, a pattern typical of civil law countries. In some civil law countries, the judiciary is largely self-regulating; judges are selected by competitive examination after receiving specialized legal training, and they advance through the judicial hierarchy on the basis of seniority and the approval of superior judges. In other civil law countries, the executive dominates judicial recruitment and promotion to the supreme courts. Often the two factors are found together, resulting in highly-insular supreme courts.

For instance, in Chile the Supreme Court annually evaluates and ranks all of the country's judges, and appointments to the Court are made by the executive from a short list of nominees selected by the Court itself from among the lower judiciary; the system discourages the development of diversity of opinion and creativity within the judicial ranks, and, as a consequence, the Court has remained very quiescent over most of its history (Stotzky 1993, Vaughn 1992–93). Although the differences between civil law and common law patterns of appointment to supreme courts thus loom large, the dichotomy does not account for all variations. In the common-law world, the process of appointment to the US Supreme Court is notoriously open and conflictual, while for the British higher judiciary it is highly insular. In France, judges on the Court of Cassation are drawn exclusively from inferior courts; by contrast, members of the Council of State are drawn not only from inferior administrative tribunals but also from the admin-

istrative bureaucracy and academia (Brown and Bell 1993, pp. 78–80). In general, the insularity of the recruitment process in continental Europe appears to be gradually eroding, with potentially significant implications for the judicial role.

Third, the degree of popular legitimacy of supreme courts and the nature and diversity of external coalitions of support for them place constraints on direct manipulation of these courts by other governmental officials. Some supreme courts have gained significant popular legitimacy, while others remain relatively unsupported; judicial legitimacy, ironically, may be cultivated by courts themselves through active intervention in the policy process that attracts external coalitions of support (Gibson et al. 1998). The ability of courts to cultivate supporters, however, is constrained by the extent of litigational resources and organizational pluralism in civil society; strong societies composed of numerous litigation-advocacy organizations are a key condition for the development of active judicial agendas (Epp 1998). Where coalitions of support are diverse and politically strong, they have played key roles in supporting supreme courts against measures aimed at curbing their power. This is clear, for instance, in comparing the limited success of attacks on the US Supreme Court prior to 1937 (Ross 1994) with the Gandhi government's relatively greater influence over the Supreme Court of India in the 1970s (Baxi 1980, Epp 1998). In general, judges in common-law countries tend to be accorded relatively high status and respect, while the civil law tradition has cultivated a lasting and deeply-entrenched disrespect for the judicial role and for judges (Merryman 1996).

Although the various factors just summarized are conceptually distinct, they are mutually reinforcing in practice. Governing systems in which political power is decentralized have tended to accord greater symbolic status to their judges, particularly to their supreme court justices; have tended to develop pluralistic and even conflictual appointment processes that select justices from a diverse population; have tended to cultivate broad coalitions of support around their supreme courts; and their supreme courts have tended to be relatively actively engaged in the policy process. By contrast, governing systems in which political power is centralized have tended to accord their judges, and even their supreme court justices, the lowly status of legal bureaucrats; have tended to select these justices through an executive-dominated process from highly insular pools of legal professionals; and, as a consequence, these courts have developed highly formalistic decision processes that yield politically-acquiescent decisions and which rarely cultivate support in the broader population. The latter extreme is exemplified by the insular judicial bureaucracies of Latin America that have few sources of popular support—a pattern that is surely compounded by the effects of economic inequality on the degree of organizational pluralism in civil society.

2.2 The Growth in Policy Intervention by Supreme Courts, and the Limits of their Power

Supreme courts, particularly in governing systems in which political power is not highly centralized, have increasingly intervened in the policy process in recent decades, in areas ranging from individual rights to environmental policy. Judges in active policy-making supreme courts commonly have developed relatively high degrees of individuality, and, consequently, these courts are subject to relatively open internal conflict that has made possible the quantitative study of decision-making patterns. Several partially conflicting strains have developed in these studies. One, called the 'attitudinal model,' posits that justices' votes directly reflect their political attitudes toward judicial policies; this theory, although originally based on studies of the US Supreme Court (Segal and Spaeth 1993), has now been applied elsewhere as well (Morton et al. 1995). A second strain of research, sometimes labelled 'new institutionalism,' rejects the claim that justices' decisions directly reflect their policy preferences (Clayton and Gillman 1999). One variant, influenced by rational choice theory, posits that in attempting to achieve their policy goals, supreme court justices temper the raw exercise of their preferences through strategic action that takes into account the preferences of other actors (and thus also of certain institutional norms) (Epstein and Knight 1998). A second variant of new institutionalism emphasizes the influence of legal norms: these studies show that many decisions, particularly in the agenda-setting process, are constrained by legal factors and are reached unanimously (Perry 1991), and that legal assumptions establish the foundations and boundaries of judicial decision-making and thereby constrain the choices recognized by the justices (Gillman 1993).

The extent to which supreme courts are capable of effecting substantial changes in governmental policy, too, remains a matter of dispute. It is clear that supreme courts have no enforcement authority of their own and thus depend for implementation of their decisions on the actions of public officials, private individuals and organizations. On this basis, Gerald Rosenberg (1991) has argued that policy reform decisions of the US Supreme Court have led to actual reforms only in very limited circumstances, particularly when they are aided by sympathetic officials or a sympathetic population, or are subject to implementation by market forces. Michael McCann (1994), however, has shown that judicial decisions provide strategic resources that may be used by social movements, organized groups, and public officials to substantial effect (albeit often in ways unanticipated by the judiciary) (see *Law, Mobilization of*). In some countries, indeed, particular supreme court decisions undoubtedly have led to substantial policy changes. In the USA, time-series analyses have revealed that major decisions by the US Supreme Court have significantly

influenced news media attention to particular issues (Flemming et al. 1997). Taken as a whole, the increasing policy intervention of supreme courts, and of courts in general, has contributed to the 'judicialization' or 'legalization' of policy and administrative processes (see e.g., Tate and Vallinder 1995).

3. The State of Social Scientific Research on Supreme Courts

Research on supreme courts has advanced remarkably in the late twentieth century. Scholars have developed sophisticated quantitative databases of decision-making patterns and of relations between supreme courts and some aspects of governmental systems and societies at large. Additionally, research agendas are increasingly based on social scientific theories and thus increasingly attempt to explain rather than merely to describe. Yet important gaps remain. The cognitive revolution in psychology has as yet been largely ignored in studies of supreme court decision-making. Scholarly attention, moreover, is often directed exclusively to courts' internal decision-making processes even though it is widely recognized that judicial policies are commonly transformed in the implementation process. Most of the existing social scientific research on supreme courts has focused on a handful of highly active courts, particularly those of the United States and Canada, while many other courts, particularly the supreme courts of cassation in the civil-law world, have been largely ignored. Finally, although research on supreme courts in many parts of the world has expanded in the 1990s, only a few truly comparative studies have been done.

Nonetheless, our understanding of supreme courts has grown. The body of social scientific research suggests both that supreme courts may contribute to the broader goals that many observers in recent years have pinned on them—the protection of individual and minority rights and the cultivation of constitutional normal more generally—but also that the extent to which they support these goals is conditioned by a range of factors, particularly the degree to which political power in the governing system is centralized, and the degree to which these courts enjoy popular legitimacy and are supported by coalitions of organizations in civil society.

See also: Civil Law; Constitutional Courts; Courts and Adjudication; Judicial Review in Law

Bibliography

Barker R S 1986 Constitutional adjudication in Costa Rica: a Latin American model. *Inter-American Law Review* **17**: 249–74

Baxi U 1980 *The Indian Supreme Court and Politics*. Eastern Book Company, Lucknow, India

Brown L N, Bell J S 1993 *French Administrative Law*. 4th edn. Clarendon, Oxford, UK

Cappelletti M 1989 *The Judicial Process in Comparative Perspective*. Clarendon, Oxford, UK

Clayton C W, Gillman H 1999 *Supreme Court Decision-making: New Institutionalist Approaches*. University of Chicago Press, Chicago

Damaska M R 1986 *The Faces of Justice and State Authority: A Comparative Approach to the Legal Process*. Yale University Press, New Haven, CT

Epstein L, Knight J 1998 *The Choices Justices Make*. CQ Press, Washington, DC

Epp C R 1998 *The Rights Revolution: Lawyers, Activists, and Supreme Courts in Comparative Perspective*. University of Chicago Press, Chicago

Favoreu L 1990 American and European models of constitutional justice. In: Clark D S (ed.) *Comparative and Private International Law: Essays in Honor of John Henry Merryman on His Seventieth Birthday*. Duncker and Humblot, Berlin, pp. 105–36

Flemming R B, Bohte J, Wood B D 1997 One voice among many: the supreme court's influence on attentiveness to issues in the United States, 1947–1992. *American Journal of Political Science* **41**: 1224–50

Garcia F S 1998 The nature of judicial reform in Latin America and some strategic considerations. *American University International Law Review* **13**: 1267–325

Gibson J L, Caldeira G A, Baird V A 1998 On the legitimacy of national high courts. *American Political Science Review* **92**: 343–58

Gillman H 1993 *The Constitution Besieged: The Rise and Demise of Lochner Era Police Powers Jurisprudence*. Duke University Press, Durham, NC

Henkin L, Rosenthal A 1990 *Constitutionalism and Rights: The Influence of the United States. Constitution Abroad*. Columbia University Press, New York

Hill E 1993 Majlis al-Dawla: The administrative courts of Egypt and administrative law. In: Mallat C (ed.) *Islam and Public Law: Classical and Contemporary Studies*. Graham and Trotman, London, pp. 207–28

Itoh H 1989 *The Japanese Supreme Court: Constitutional Policies*. Wiener, New York

Jacob H, Blankenburg E, Kritzer H M, Provine D M, Sanders J 1996 *Courts, Law and Politics in Comparative Perspective*. Yale University Press, New Haven, CT

Kagan R A, Cartwright B, Friedman L M, Wheeler S 1977 The business of state supreme courts, 1870–1970. *Stanford Law Review* **30**: 121–56

Knopff R, Morton F L 1992 *Charter Politics*. Nelson Canada, Scarborough, Canada

Krug P 1997 Departure from centralized model: The Russian supreme court and constitutional control of legislation. *Virginia Journal of International Law* **37**: 725–87

Maltese J A 1995 *The Selling of Supreme Court Nominees*. Johns Hopkins University Press, Baltimore, MD

McCann M 1994 *Rights At Work: Pay Equity Reform and the Politics of Legal Mobilization*. University of Chicago Press, Chicago

McWhinney E 1986 *Supreme Courts and Judicial Law-making: Constitutional Tribunals and Constitutional Review*. Martinus Nijhoff, Dordrecht, The Netherlands

Merryman J H 1985 *The Civil Law Tradition: An Introduction to the Legal Systems of Western Europe and Latin America*, 2nd edn. Stanford University Press, Stanford, CA

Merryman J H 1996 The French deviation. *American Journal of Comparative Law* **44**: 109–19
Merryman J H, Vigoriti V 1967 When courts collide: Constitution and cassation in Italy. *American Journal of Comparative Law* **15**: 665–86
Morton F L, Russell P H, Riddell T 1995 The Canadian charter of rights and freedoms: A descriptive analysis of the first decade, 1982–1992. *National Journal of Constitutional Law* **5**: 1–60
Pacelle R L Jr 1991 *The Transformation of the Supreme Court's Agenda: From the New Deal to the Reagan Administration*. Westview Press, Boulder, CO
Perry H W 1991 *Deciding to Decide: Agenda Setting in the United States Supreme Court*. Harvard University Press, Cambridge, MA
Rosenberg G N 1991 *The Hollow Hope: Can Courts Bring About Social Change?* University of Chicago Press, Chicago
Ross W G 1994 *A Muted Fury: Populists, Progressives, and Labor Unions Confront the Courts, 1890–1937*. Princeton University Press, Princeton, NJ
Segal J A, Spaeth H J 1993 *The Supreme Court and the Attitudinal Model*. Cambridge University Press, New York
Shapiro M M 1981 *Courts: A Comparative and Political Analysis*. University of Chicago Press, Chicago
Solomon D H 1992 *The Political Impact of the High Court*. Allen and Unwin, St Leonards, Australia
Stevens R B 1978 *Law and Politics: The House of Lords as a Judicial Body, 1800–1976*. University of North Carolina Press, Chapel Hill, NC
Stotzky I P (ed.) 1993 *Transition to Democracy in Latin America: The Role of the Judiciary*. Westview Press, Boulder, CO
Tate C N 1993 Courts and crisis regimes: A theory sketch with Asian case examples. *Political Research Quarterly* **46**: 311–38
Tate C N, Vallinder T (eds.) 1995 *The Global Expansion of Judicial Power*. New York University Press, New York
Vaughn R G 1992–93 Proposals for judicial reform in Chile. *Fordham International Law Journal* **16**: 577–607
Verner J G 1984 The independence of supreme courts in Latin America. *Journal of Latin American Studies* **16**: 463–506

C. R. Epp

Surprise Examination

Examinations, and tax audits, are not always announced in advance. There are 'surprise' examinations. What is more, it seems that there might be surprise ones consistent with it being announced that they will occur in some specified period of time. There is, however, a line of argument to the conclusion that it is impossible to have a surprise examination when it has been announced ahead of time that it will occur in some specified period of time. This is very hard to believe, which is why the line of argument constitutes a paradox.

This article starts by stating the line of argument and various attempts to identify the fallacy. It concludes with a short discussion of some related issues concerning cooperative social arrangements, a matter often raised under the heading of the backwards induction paradox. This article does not discuss certain technical issues to do with self-reference sometimes raised in the context of the surprise examination paradox (for these see, e.g., Shaw 1958).

1. *The Paradox*

Suppose that on Monday morning the head of a school makes two announcements to the students.

(a) An examination will be held at 14:00 one afternoon this week.

(b) You will not know until shortly before 14:00 on the day which afternoon it is.

The examination will, that is, be a surprise one.

The argument to paradox runs as follows.

Suppose that the examination is held on Friday. In this case, it will be known in advance when the examination is to be held. For it will be known on Thursday after 14:00 that as the only day left to hold the examination is Friday, the examination has to be on Friday. It follows that the examination cannot be held on Friday consistent with the announced conditions (a) and (b). Now suppose that the examination is held on Thursday afternoon. After 14:00 on Wednesday, it will be known that the only afternoons left to hold the examination are Thursday and Friday. But Friday has just been ruled out as a possible day to hold a surprise examination, so that leaves Thursday as the only possible day. But then the students will know after 14:00 on Wednesday that the examination has to be on Thursday. But this is to say that Thursday as well as Friday can be ruled out as a possible day for a surprise examination. Suppose next that the examination is held on Wednesday. On Tuesday, after 14:00, it will be known that the only afternoons left for the examination are Wednesday, Thursday, and Friday. Thursday and Friday are not possible days for a surprise examination, by the reasoning just given, which leaves Wednesday as the only possible day. But then, if the examination is held on Wednesday, it will be known in advance that it is on Wednesday: Wednesday can, that is, be ruled out as a possible day for a surprise examination. Repeating the same reasoning rules out Monday and Tuesday. It follows that it is impossible to hold an examination under the announced conditions, that announcing that the examination will be held in a given period, and that it will be a surprise one makes it impossible for it to be a surprise one.

There is clearly a fallacy somewhere in this reasoning but it has proved hard to say exactly where it goes wrong. This is part of the reason the paradox has generated a large literature (see Kvanvig 1998).

2. *Replies that Draw on the Possibility of Doubt About (a) or (b)*

2.1 *Uncertainty About (a)—the Examination Being Held at All*

In most discussions of the paradox, it is granted that the examination cannot be held on the last day and still be a surprise. The focus of the discussion is on where the argument goes wrong *after* the step where Friday is excluded. However, as Quine (1953) points out, the matter is not that simple. If there is some doubt about whether or not the conditions set by the head will be satisfied, the argument that rules out the last day can fail. In particular, if there has been no examination by late Thursday afternoon and the students find themselves, perhaps as a result of this fact or perhaps for some other reason, in doubt about whether condition (a) will be satisfied, about whether, that is, there will be an examination at all in the week in question, they do not know that there will be an examination on Friday. There maybe, for all they know, no examination at all. In this case, if the head proceeds to hold the examination on the Friday at 14:00, both conditions set down will be satisfied. The examination will have been held at 14:00 on one day in the week in question (which satisfies condition (a)), and the students will not have known in advance which day it was to be (which satisfies condition (b)).

However, as many commentators point out, the students might be in a position to be certain that the examination will be held in the week in question. Perhaps they know that holding the examination some time during that week is an unshakable prerequisite for government registration of the school. However, it seems that it would still be possible that the examination should be a surprise. There must, therefore, be something more to say about where the argument to the conclusion that a surprise examination is impossible goes wrong. The mistake cannot simply be the possibility of doubt about the examination being held in the relevant period, because the paradox survives the removal of that doubt.

2.2 *Uncertainty About (b)—the Examination Being a Surprise*

Suppose that it is certain that the examination will be held in the week in question, in which case the students can rule out Friday as a possible day for a surprise examination. This is consistent with it being a matter of doubt whether or not the examination will be a surprise one. Now suppose that there is doubt about this condition—(b)—being satisfied. Wright and Sudbury (1977) point out that this undermines the students' reasoning on Thursday morning in the case where Monday through Wednesday prove to be exam-free. The students cannot argue that Friday is excluded as a possible day for the examination to be held on the ground that then they would know in advance when the examination was to be held. For, if there is doubt about (b), they must allow that maybe the examination will not be a surprise. They must, therefore, allow that there are two possible days for the examination to be held: Friday, in which case (b) will be violated, or Thursday. In this case, the head can hold the examination on Thursday consistent with (a) and (b). This is because the students cannot, on Thursday morning, rule out Friday owing to doubt about the examination being a surprise. In effect, the head can exploit doubt about the examination being a surprise—which leaves Friday as a possible day for the examination in the students' minds—to ensure that it is a surprise by holding it on Thursday.

In sum, doubt about (a) would make it possible for the examination to be held on Friday consistent with (a) and (b). Doubt about (b), even in the absence of doubt about (a), would make it possible to hold the examination on Thursday consistent with (a) and (b).

The obvious question to ask at this point is what happens when there is no doubt about either (a) or (b). Could it not be the case that it is certain that the examination will be held on one afternoon of Monday through Friday, and certain that the students will not know in advance which afternoon it is? Isn't there still a paradox to be addressed? In order to address this question, it is necessary to draw a distinction between whether there is doubt on some matter and whether there *would* be doubt on some matter were certain things to happen.

3. *What is Open to Doubt vs. What Would be Open to Doubt*

It is not open to doubt given current evidence that Hitler died in the bunker, but were certain discoveries to be made, it would become open to doubt. Similarly, for most people who are here and now awake, it is not open to doubt that they are here and now awake, but were they to have the experience as of waking up from an extraordinarily life-like dream, it would be open to doubt. What is here and now open to doubt is distinct from what would be open to doubt were certain things to happen.

The view that the points made in the previous section do not dispose of the surprise examination paradox arises from the conviction that a surprise examination is possible when it is beyond doubt both that the examination will occur during the week and that the students will not know in advance. How then, runs the challenge, can we dispose of the paradox by pointing out that when there is doubt about (a) or (b), the student's reasoning fails at some key point or other? However, on careful inspection, the failure of the students' reasoning does not depend on there actually being doubt about (a) or (b). It depends on it

being the case that there would be doubt were certain things to happen.

Provided it is the case that were Friday morning to come without the examination having so far taken place, there would then be doubt about the examination happening in the week in question, that is, about (a), the students' cannot rule out the examination being held on Friday. If there would, *should Friday morning arrive before the examination makes its appearance*, be doubt on that Friday morning about (a), the head can hold the examination on Friday without violating either (a) or (b)—as was noted above. This does not require that there be, at any stage, actually any doubt. It requires simply that were Friday morning to arrive without an examination having been held, there would be doubt about (a).

A similar point applies to the role of doubt about the examination being a surprise, that is, of doubt about (b). There need not actually be any doubt that the examination will be a surprise. The students' reasoning would be undermined simply by the fact that there would be doubt should enough days pass without the examination taking place. As was noted above, should Thursday morning arrive without there having been an examination and were this to mean that there would then be doubt about (b), the students cannot rule out Friday even if they are certain of (a). It is not required that there ever actually be any doubt about (b). All that is required is that should certain things come to pass, there would be doubt about (b).

It might be suggested that it could be built into the paradox that not only is there no doubt at any stage about (a) and (b), there would be no doubt about them no matter how many days pass without an examination. But this would be to build something absurd into the statement of the paradox. There would *have* to be doubts about one or both of (a) and (b) should enough days pass without the examination having taken place. For example, should Friday morning arrive without the examination having taken place, the students would *know* that one or both of (a) and (b) are open to doubt. It would be inconsistent of them to hold in that eventuality both that it is beyond doubt that the examination will take place on one of the designated days and that they cannot know in advance which it is. For example, if it would still be certain that the examination will take place on one of the designated days should Friday morning arrive without the examination having been held, they would know that should this happen, the examination would not be a surprise (for a longer development, see Jackson 1987).

4. The Backward Induction Paradox

4.1 The Rationality of Cooperation

There are parallels between the style of argument in the surprise examination paradox and certain arguments concerning the rationality or otherwise of cooperative social arrangements. Many cooperative arrangements that have both costs and benefits to the involved parties rest on the basis of expected returns in the future to all parties to the arrangement. The Robinsons work with the Smiths on something that advantages the Smiths in the expectation that in the future the Smiths will work with the Robinsons on something that advantages the Robinsons. A classic example is helping with building houses. It is much easier to build a house when there are a number of workers to lift heavy beams and the like. So the Robinsons give of their time to help the Smiths build the Smiths' house, but do so in the expectation that the Smiths will later help the Robinsons build the Robinsons' house. But why, in terms of self-interest, should the Smiths later help the Robinsons; they already have their house? The answer is that there will be other opportunities for cooperation, and the Smiths can expect to benefit from some of them.

This intuitively appealing answer can be challenged by a 'backwards induction.' It will not be rational (in the sense of being advantageous) for the 'giving' party to cooperate on the *last* occasion for possible cooperation, because there is no prospect of their being the 'receiving' party in the future. But then, it seems, there is no reason for the giving party to cooperate on the second last occasion either, assuming rationality on the part of both parties. The only reason for the giving party to cooperate on the second last occasion is in the hope of being the receiving party on the last occasion, and that could only happen if the giving party on the last occasion is irrational. Similar reasoning applies to the third last, fourth last, etc. occasions, so yielding the conclusion that it is never rational to be a giving party in a cooperative arrangement.

It would be a mistake to reply to this argument by challenging the self-interest conception of rationality it employs. The puzzle is that it clearly is in parties' self-interest sometimes to cooperate when they are the givers, in the expectation of being receivers in the future. The error in the argument is that it neglects the kind of epistemological points we saw to be crucial in diagnosing the error in the surprise examination paradox. Although it would be irrational for the giving party to cooperate on the *known* last occasion for cooperation, normally it is not known when the last occasion for cooperation is. Similarly, there may be doubt about all parties being rational.

4.2 Sequences of Prisoner's Dilemmas

A special version of the backwards induction paradox is sometimes deployed to argue that it is rational to defect throughout a sequence of prisoner's dilemmas as well as in a one-off prisoner's dilemma. A prisoner's dilemma is defined as follows. There are two parties, X

and Y. If they both cooperate, they each do better than if they both defect. However, X and Y each do better if they defect, independently of whether or not the other cooperates or defects—defecting dominates cooperating, as it is put (see Resnick 1987).

Most game theorists accept that this means it is rational for each to defect despite the fact that if they both defect, they each end up worse off than if they both do the irrational thing of cooperating. What has proved harder to accept is the claim that, in a sequence of prisoner's dilemmas of known length involving rational players, it would be rational to defect throughout. Surely it would be rational, sooner or later, for X and Y to start cooperating in order to set up an expectation of cooperation which means that, at least sometimes, they end up both cooperating, which is better for each than their both defecting?

A backwards induction argument has been used to resist this thought. Clearly, runs the resistance, it is not rational to cooperate at the last prisoner's dilemma—engendering expectations of future cooperation has no pay off in that case. But cooperation at the second last prisoner's dilemma is only rational if it leads to cooperation at the last prisoner's dilemma, so this means that cooperation at the second last cannot be rational *unless* it leads to irrationality at the last prisoner's dilemma. But that is to say that if X and Y are steadfastly rational, defection is rational at the second last prisoner's dilemma as well as at the last. Repeating the reasoning delivers the result that it is rational to defect throughout the sequence of prisoner's dilemmas.

A full discussion of this last backwards induction argument is beyond the scope of this article (but see Pettit and Sugden 1989 and Bicchieri 1998). But it should be noted, first, that belief in rationality, or lack of it, is crucial to the argument, and, second, that if cooperating at early stages in a sequence of prisoner's dilemmas sets up certain expectations, these early stages will not strictly speaking be prisoner's dilemmas, for taking into account the longer term benefits means that they will not be cases where defecting dominates cooperation.

See also: Prisoner's Dilemma, One-shot and Iterated

Bibliography

Bicchieri C 1998 Decision and game theory. In: Craig E (ed.) *Routledge Encyclopedia of Philosophy*. Routledge, London, Vol. 2, pp. 823–35

Kvanvig J L 1998 Paradoxes, epistemic. In: Craig E (ed.) *Routledge Encyclopedia of Philosophy*. Routledge, London, Vol. 7, pp. 211–14

Jackson F 1987 *Conditionals*. Basil Blackwell, Oxford, UK

Pettit P, Sugden R 1989 The backward induction paradox. *Journal of Philosophy* **86**: 169–82

Quine W V 1953 On a so-called paradox. *Mind* **62**: 65–7

Resnick M D 1987 *Choices*. University of Minnesota Press, Minneapolis, MN

Shaw R 1958 The paradox of the unexpected examination. *Mind* **67**: 382–4

Wright C, Sudbury A 1977 The paradox of the unexpected examination. *Australasian Journal of Philosophy* **55**: 41–58

F. Jackson

Survey and Excavation (Field Methods) in Archaeology

1. The Term 'Archaeology'

The term 'archaeology' comes from the Greek word, *arkhaiologia*, meaning the 'discussion or discourse about ancient things.' By definition, archaeology is the science by which the remains of the ancients can be methodically and systematically studied to obtain as complete a picture as possible of ancient culture and society, and thereby to reconstruct past ways of life. An antiquarian doctor of Lyon, Jacques Spoon, first used the term 'archaeology,' in the seventeenth century. Archaeology is a discipline that involves the scientific method of study, observation, recording, and experiment. Archaeologists study in detail the complete life of cultures and their patterns of change in an attempt to delineate the causes and effects of the cultural process. The fascination of archaeology lies in the solving of long-standing problems and in the discovery of the meaning of things. The excitement of archaeology lies in the new problems which recent findings produce.

2. The History of Archaeology

The earliest 'archaeologist' was the sixth century BCE King Nabonidus of Babylon who excavated a temple down to its foundations which had been laid thousands of years before. Even the Roman emperor Augustus was a collector of 'giants' bones and ancient weaponry. With time, Europeans began to collect burial urns, chipped flint stone and ground stone tools which some believed were amulets. In 1719 the Austrian Prince d'Elboef led the first excavation to Herculaneum, which had been destroyed along with Pompeii in AD 79 when Mount Vesuvius erupted. Thereafter when pieces of classical sculpture began to turn up, artists began to study their shapes and decoration and collectors began to collect, proudly display, and discuss their forms. With time, more people began to describe what they recognized as ancient monuments

and excavations were undertaken to retrieve more and more objects. In the mid eighteenth and early nineteenth centuries digging for ancient objects became the pursuit of gentlemen of the leisured class and professionals: they became known as antiquarians and formed clubs to discuss the merits of ancient artifacts.

Archaeology as a pursuit using scientific systems of recovery began in the nineteenth century, along with the growing realization that humankind had an ancient past that could be investigated in a methodical way. The names of most important archaeologists from the early days include Sir William Matthew Flinders Petrie in Egypt, Robert Koldewey in Babylon, the duplicitous Heinrich Schliemann in the Aegean at Troy, Mycenae and Tiryns, and General Augustus Pitt-Rivers in Britain. Later pioneers were Sir R. E. Mortimer Wheeler in Britain and India, George Andrew Reisner and Sir Leonard Woolley in the Near East, Max Uhle and Alfred Vincent Kidder in America, and François Bordes and André Parrot in France.

In the present day, archaeology has grown into a multidisciplinary, methodical, problem-oriented search for the past. It now involves above-ground research as well as excavation, and a host of specialists is required, depending on the site. Background research as well as excavation has become meticulous. The object study of pottery, for example, involves not only typologies but sophisticated scientific testing by neutron activation analysis to understand the chemical components and to identify the sources and, cultural patterns as well as to assess trade relationships.

Archaeological questions extend far beyond the excavation and interpretation of any single cultural group, geographic region, or chronological period of time. In theory, all archaeological study is aimed toward an explanation that will provide a conceptual basis for the reconstruction of the past, and the where, how, and why of the changes that took place.

3. *The Importance of Archaeology*

Archaeology is unique in that it encompasses the whole world over the past 2.5 million years in order to reconstruct the past. It includes the research of all kinds of sites—underwater wrecks or port installations, caves, jungles, deserts, and mountain-top sites; wherever past activity has taken place. Archaeology wants to discover where and when humankind developed and began to practice art, agriculture, and technology and became urbanized. It is our route to our understanding of ourselves, for in order to know where we are going we have to reconstruct and understand the past and where we have come from.

Archaeology continues to give us a broad and long-term view about the past and the events that conditioned us. There is an implosion of new answers to research on how, where, and when humankind began, our development of technology, the arts, the expression of our ideas in writing, how we harnessed the environment brought about the domestication of plants and animals, and developed agricultural systems. Archaeology helps us to understand the processes of becoming urbanized how we developed complex societies, and to understand our basic needs.

4. *Archaeological Sites*

Archaeology covers a wide diversity in time, space, subject matter, and approach. An archaeological site can be any place that has been a scene of past human activity, including a briefly settled campsite, a shipwreck, a monumental city, or the evidence of any number of different combinations of human remains. Any place where human beings have established themselves, even momentarily, is considered a site.

As each archaeological site is a unique time capsule, each has its own distinct character and problems. Types of sites are often labeled by their prehistoric or historic context, their function, their general topographic nature, or according to whether they are composed of one or many levels of occupation. Sites represent a body of data relevant to their setting and their cultural patterning and must be interpreted both in relation to this local setting and as a link between cultures. Sites may be small or large. Small sites are far more numerous throughout the world and may range from hunting or fishing stations to small village communities. There are prehistoric plains or riverside terraces such as Olduvai Gorge in Tanzania, Africa; there are caves such as Chou Kou Tien near Peking, China where the Peking Man was found; or those containing Upper Paleolithic art, like Lascaux in France.

Large sites are generally cities, temples, large tells—artificially built-up mounds in the Middle East with one city built on top of the other (known as *höyük* in Turkish or *tepe* in Persian). The ancient mound of Jericho, Israel is particularly important, revealing evidence dating back to 8000 BCE. A more recent tell of the Mississippi Basin was built by the Mound Builders of North America in about 1300 CE. There are monumental cities such as classical Rome, Meso-American Tikal in Guatemala, or Angkor Vat in Cambodia. There are sacred areas reserved for ritualistic functions such as Stonehenge in England, or the statues of Easter Island. There are cemeteries and tombs such as those of Mycenae in Greece, the Pyramids in Egypt, or the mounds of Ohio, Illinois or Mississippi. Famous sites also include underwater shipwrecks and harbors discovered worldwide off the shores of the Mediterranean, or near Oslo, Norway, where early Viking ships have been recovered.

The selection of a site for excavation therefore depends upon what the excavator–principal investigator wants to find, and what problems will be answered by excavation. What is the nature of the evidence? It usually involves a problem to be solved—how has the site been built up over time? What is its environment and how has this changed? Using a research design, a review process is undertaken which includes the preliminary analysis of the site, the evaluation of the probable results, the examination of the costs, and the evaluation of organizational abilities and needs to carry out the tasks involved.

5. *Archaeological Studies*

Many types of archaeology have evolved over the years. There are Old World (Eastern Hemisphere) and New World (Western Hemisphere) prehistoric specialties, based on the stone tools humankind produced. The Paleolithic, Mesolithic, Neolithic, Bronze Age, and Iron Age are specialties. These are broad artificial classifications with subdivisions. As far as the discipline in which it has received attention, Old World archaeology is a separate subject in some schools, or is associated with departments of classics or Near Eastern studies, history, or art history. In the New World prehistoric specialists may concentrate on the Paleolithic, Archaic, and Woodland periods. Then both Old and New World archaeology moves into historical archaeology. In the Old World there are experts who are devoted to classical archaeology (the worlds of Greece and Rome) or Biblical archaeology (the areas encompassed by the Holy Land).

Similarly, many kinds of specialties have developed in archaeology, for example, industrial archaeology; the study of manufacturing and salvage archaeology; the rescue work done before a modern construction takes place such as land development, or urban renewal, or when a pipeline, road, or dam is built. There are also technical specialties of people who concentrate on specific techniques such as ceramics, lithics, paleobotany, palynologists (specialists in pollen), and dendrochronologists (specialists of tree ring growth) among many others, including computer specialists who lend support to all these areas.

Since the 1960s in America, archaeology as a subdiscipline also has been associated with anthropology. As anthropology is the study of humanity, anthropologists generally observe the behavior of present-day people in order to explain and reconstruct their behavior patterns. In so doing, they have to make assumptions about past behavioral patterns as set against the environment of the present.

Anthropology is divided into several subfields among which are social and/or cultural anthropology, physical or biological anthropology, linguistic anthropology, which includes the interpretation of oral history, demographic anthropology, ethnic anthropology, cognitive and symbolic anthropology, enthnoarchaeology, and archaeo-astronomy (the study of ancient astronomical events).

6. *Dating Systems in Archaeology*

The term 'culture' is used when assuming that a group of people exhibit shared behavior. Archaeologists want to be able to establish accurate space and time frameworks or dates for the cultures they are researching. In short archaeologists want to establish the site context, culture, and chronology. Laboratory methods provide what are known as absolute dating methods, and relative dates are ascribed to archaeological strata based on the site deposition or stratigraphy and the discovery of cultural artifacts similar to those found at other sites.

Absolute dating is a date given in specific years, in terms of chronometrical dating or the calendar. A date given as BP is a date before 1950. BC means 'before Christ' (although nowadays BCE is used, meaning 'before the Common Era'). CE (Common Era) is used instead of the old term AD, meaning *Anno Domini*, or after the birth of Christ. Absolute dates can be established by laboratory analyses like radiocarbon dates (C-14), potassium-argon, thermoluminesence (TL), archaeomagnetism, and palaeomagnetism, among others. These are geochronologic methods used for determining the absolute date of a deposit.

Relative dates are based on the stratigraphic deposition—the principles of superposition, association, and correlation. Superposition assumes that the uppermost stratum (soil level) was deposited last and the lowest stratum was deposited first; thus the reasoning assumes that lower strata are generally older than the layers above them. Association assumes that the artifacts found in the same closed and sealed archaeological context are contemporary and are therefore presumed to be culturally and temporally associated. Correlation assumes that, once cultural associations have been determined, artifacts can be cross-culturally correlated with artifacts from other site cultures and they can be expected to have been manufactured at the same time, and therefore to possess the same relative date.

Archaeologists use seriation as a method of relative dating. Seriation involves the ordering of specific artifact types from an earlier to a later time sequence by the use of a typology. In creating a type series, the archaeologist assumes that artifacts have formal attributes that can be measured like raw material, length, width, thickness, color, weight, and so on, and that these nonrandom attribute factors were used as a standard by the maker. Therefore, a typology is comprised of types which are a cluster of attributes in a class of objects and these can be given relative dates that can be used to date the stratification.

Once associations have been defined cultural levels can be inferred. A cultural level represents the choices of a particular social group, such as the classical Greeks. The data collected in such a cultural level depict a known and recognizable set of ideas that are then represented as the cultural products of that society.

7. *Archaeological Survey*

Survey includes exhaustive historical research, map-reading, geomorphologic and geographic analysis, aerial photography, and ground-level surveys, including electronic and magnetic surveying techniques. Historical research includes all previous research of the area—excavation reports, artifact studies, and the conclusions drawn by previous researchers. In-depth bibliographic research provides essential background information.

All archaeologists are concerned with surveying as a means of planning and laying out the area to be studied or excavated. Surveying is generally the first step in all archaeological projects, and is often the last step before the completion of the excavation. Using maps, aerial photos, and whatever background information is available, ground level surveys can be undertaken on foot to understand the physical features of the site as well as the artifacts—in short, all the material cultural remains of the site or the culture. The site is then located according to Global Positioning Systems on a topographic government map with a scale of 1:50,000, and its location is established and officially reported.

As a branch of engineering, surveying is the science that accurately determines the shape, area (size), and positions of a site's surface by the measurement of certain points. It includes the measurements of horizontal distances and elevations, and determines the direction and relationships of angles by the use of optical instruments. There are several principal classes of survey: *geodetic* surveys, *photogrammetric* (picture-measured) aerial surveys, and *plane* surveys. Geodetic surveys are used for the study of large areas like countries or continents, and take into account the curvature of the earth's surface. Satellite photographs have become important to the archaeological analyses of large landforms. Photogrammetric surveys are the science of taking detailed vertical or oblique photographs from an aircraft for the reliable measurement of the earth's features. Plane surveying is used for small areas of archaeological focus in which the curvature of the earth's surface is not measured. The preliminary survey also consists of a systematic walking over the site recording standing structures and features as well as artifact materials and scatters. Depending upon what type of survey is to be conducted, artifact concentrations may be collected, processed and studied to indicate the type of site and its artifact range. The survey should establish the site size, what kind of site it may be, its time span, and, if possible, its depth of deposit.

Depending on the site, three principal types of survey are usually conducted before excavation: (a) the land survey for the proper planning and laying out of the site which includes the site grid; (b) The topographic survey which is the measurement of elevations representing the surface of the area; and (c) the ongoing excavation survey, which is made during excavation to control the location of archaeological structures, features, and objects and providing these with elevations. Anomalies and physical features are then thoroughly mapped and described as well as factors like vegetation and soils.

Decisions then may be taken as to what additional surveying equipment will be used to look at the subsurface features. Electrical resistivity meters and magnetic surveying by using any number of different magnetometers are instruments that are linked to computer-processing programs; they help to determine what disturbances or structures exist in the area. Only when the comprehensive survey has been completed can the researcher decide if excavation is justified, and if it is, where it is to begin, and what course it is to take.

8. *Excavation and Ethical Considerations*

Even before a site is excavated, preservation and consolidation measures must be seriously considered and implemented. How will the site be left after the excavation is completed? As excavation is destruction, one of the most important considerations is the site's stratigraphy, its depth of deposit, and its characteristics. We need to conserve and protect the excavated sites and monuments from all periods and in all parts of the world. Furthermore we need to protect sites against the threats from human intervention, mass tourism, and building development, as well as from the elements of natural deterioration and neglect. Looting is an age-old problem. There are those who clandestinely excavate to sell artifacts at an enormous profit. Although artifact-rich countries have the legislation to prevent the illegal excavation and export of their country's patrimony, they lack consistent regulatory measures to police looters, dealers, and the exporters of their antiquities. If we are to preserve our common human heritage, the protection of cultural patrimony has to become everyone's priority.

There have been changes in the ethical standards observed by archaeologists today. Ethical standards in archaeology may be defined as the obligations of a professional or an amateur to the excavation, to the country, State or area in which it is located, to the public at large, and to fellow archaeologists. Today's archaeologists strive to work and live according to a

high standard of behavior, applying their knowledge to enhance the reputation and achievements of archaeology as a whole and for the benefit of human knowledge. Many archaeological associations have drawn up ethical codes that have been accepted by international archaeological societies. The codes of the Archaeological Institute of America and the Society of Professional Archaeologists are among these.

Who will be responsible for the care and upkeep of the site over time? And what should be done to care for site artifact collections? Where are they going to be stored, who is going to study them and on what time schedule? What is to be their final disposal—a museum, reburial, and, if the latter, where will they be buried? Today the archaeologist has the responsibility of addressing such issues and these should be best addressed before excavation takes place.

9. *Excavation*

Excavation is the deliberate recovery of buried objects in relation to the layer or stratum and other associated objects in its original deposit. The principle of excavation is that artifacts, nonartifactual materials, and architectural features are not randomly distributed, but are deposited by people. Their position in the earth reflects the cultural, economic, and social behavior of humankind that can be studied by means of excavation. This is the basis for the functional reconstruction of the site. It is only through excavation that the archaeologist can collect and extract a wide range of physical information or obtain a sense of the people who lived, worked, and perhaps worshipped at a site. The archaeologist does this by describing the evidence the people left behind—the architectural features and objects. It is only through excavation that we can align these physical features in time and space so that the functional reconstruction of the processes involved can be determined. In this way we can better understand and interpret the causes, changes, and processes which the ancients used to bring about change.

Excavation is the method of removing objects and uncovering stationary features that have been concealed by later deposits. It involves the removal of layers or strata in reverse order to the way in which they were laid down, gradually revealing each successive stage in the history of the people who used the site. Because excavation is the means by which information is unearthed, it must be conducted methodically so that whatever is found can be seen and studied within its own context. The main objectives of excavation are to uncover the position of objects as well as floral and faunal data, and to plot their location on a horizontal plan. It also involves the meticulous recording and sequencing of successive layers on the vertical section by position and depth measurements. Finally, the goal is to relate such data to the total environment of the site and to assess its relationship to other sites in the area.

10. *Modern Developments in Archaeology*

Major developments in archaeology took place in the late 1990s with the advent of computer technology—remote sensing, web pages, virtual reality, computer simulations, digital cameras, and the greater use of the hard sciences including DNA analysis.

In the 1960s American archaeologists began to challenge the traditional archaeological academy and argued that the discipline had become simplistic. As a result of this critique the 'new archaeology' was developed, emphasizing the importance of the processes within the society being studied. This was known as 'processual archaeology,' and it emphasized the value of objective, logical explanation and the descriptive processes that can produce testable assumptions. They believed that culture should be studied as sets of systems and subsystems as they relate to the environment, the economy, and patterns of subsistence. They also emphasized that the interrelationships between different sectors of the society should be explainable so that specific developments could be established throughout time.

They failed to stimulate the traditionalist majority who practiced archaeology as they had before and who believed that there were no universal laws governing human behavior.

The merit to the then-new approach to archaeological data opened the way to a vast array of new theories including Marxist, neo-Marxist positivist, structuralist, post-structuralist, post-processualist, deconstructionist, and probabilist views along with other theoretical constructs. The post-processual school of archaeologists still adheres to a so-called interpretative archaeology, which insists on emphasizing history, philosophy and, whenever possible, literary studies. Interpretative archaeologists place importance on the fundamental uniqueness and diversity of culture, believing there is no one governing law that can apply either to archaeological research or to the interpretation of the past. These theories hold some truth, provide stimulating hypotheses for our explanation of the past and are used by many who thoughtfully interpret the probabilities of the evidence. Now archaeologists are wrestling with the ideas of multivariate explanations—and many continue to build idealized models for their interpretations.

11. *Conclusion*

By and large, archaeologists are dedicated researchers and teachers. As for future directions in theory and research, evaluation can be said to drive archaeological standards. By clearly specifying the outcomes

in measured ways, archaeologists will change the way they excavate and publish. In the training of students, field directors hope they will train people who possess the qualities and attributes desired in a competent archaeologist. As an archaeological project is expected to foster a shared sense of a shared mission between the field researcher and the principal investigator, both will strive to reach a common goal. Archaeology is best taught through active learning processes, and principal investigators help students by gauging their progress and identifying and overcoming barriers to the excavation as a whole. The developmental process of archaeology is most effective when the goals are known so that the whole team can work together toward the shared goals, collaboratively devising strategies to overcome any obstacles. Our understanding of the past is in a constant state of change due to new discoveries and advances in our knowledge. It will never be finished.

Today archaeology has developed into a complex discipline with a vast literature and a range of specialties that are well beyond the competence of a single researcher. More and more data are being produced, museums are overwhelmed with objects and, even with the advent with the computer, many artifacts have remained unanalyzed or unregistered. There is a crisis in the discipline due to the fact so much data has been produced, and as excavation involves the destruction of the site, archaeologists must be aware that the discipline may self-destruct unless it puts its house in order. It faces considerable challenges and risks ahead and it must develop an international dialogue among its leadership that is committed to a strategy which will ensure the saving of the past for the future.

Bibliography

Deetz J 1996 *In Small Things Forgotten: An Archaeology of Early American Life*. Anchor Books/Doubleday, New York

Greene K 1995 *Archaeology: An Introduction: The History, Principles and Methods of Modern Archaeology*. University of Pennsylvania Press, Philadelphia, PA

Hodder I 1999 *The Archaeological Process: An Introduction*. Blackwell, Oxford, UK

Hodder I 1991 *Reading the Past: Current Approaches to Interpretation in Archaeology*. Cambridge University Press, Cambridge, UK

Renfrew C, Bahn P G 2000 *Archaeology: Theories, Methods and Practice*, 3rd edn. Thames & Hudson, New York

Sharer R J, Ashmore W 1992 *Archaeology: Discovering Our Past*. Mayfield, Mountain View, CA

Schiffer M B 1987 *Formation Processes of the Archaeological Record*. University of Utah, New Mexico Press, Albuquerque, NM

Trigger B 1989 *A History of Archaeological Thought*. Cambridge University Press, Cambridge, UK

Willey G R, Sabloff J A 1995 *A History of American Archaeology*. W H Freeman, San Francisco, CA

M. S. Joukowsky

Survey Research: National Centers

National centers for survey research are a rare species among the masses of survey research centers, companies and offices. However, the impact of the few centers on the development of systematic survey research and its methodology cannot be overestimated. Four centers in three different countries will be examined as blueprints or models. These are, in chronological order, the National Center for Opinion Research (NORC, Chicago), the Institute for Social Research (ISR, Ann Arbor), the National Center for Social Research (NATCEN, London), and the German Society of Social Science Infrastructure Institutes (GESIS, Mannheim). The focus is on survey research; thus surveys must be defined clearly and distinguished from other instruments for social research, like tests and census. Next, national centers for survey research will be distinguished from survey research centers, institutes, companies, and census offices at large. The four model centers will then be portrayed with some emphasis on their history and their impact on survey research.

1. Social Surveys

Surveys are defined in this article as *observing societal facts using interviews as the source of information and systematic sampling as the base for selecting information sources* (cf. de Heer et al. 1999), and they are quantitative instruments to identify distributions of societal characteristics. Questions or items in a survey are understood as indicators pointing to a latent construct or dimension. Psychological tests, including many marketing studies, differ in two ways from surveys: they focus on characteristics of individuals or small groups but not societal characteristics. In most cases the selection of the respondent is not based on systematic sampling. Survey research is, on the contrary, neither interested in information about a single individual nor in small, identifiable groups. Information gathered on the individual level in a survey will be aggregated across groups and even societies. A population census may share items or question wording with survey research studies. However, census studies are not designed to identify latent constructs. By its name, a census is an investigation involving all members of a given population not a sample.

Survey research, hence, requires among others specific methods and techniques for designing sampling frames, dealing with nonresponse, and constructing and validating measurement instruments which must be in turn suitable for complex statistical modeling across all strata of a society. Survey research is a complex, often high-tech, production process. Consider as an example a National General Social Survey like the American GSS or the German ALLBUS (Davis and Smith 1992, Davis et al. 1994).

The target is to collect data from about 3,000 respondents. To do so, one has to contact up to 6,000 households and to identify whether they are eligible and willing to participate. Up to 600 interviewers in 630 sampling points will collect the data. About 50–100 other staff will back them up, supervise the process, or edit the incoming data. The whole process will be directed and supervised by 5–10 senior social researchers. The goal is to collect from each and every respondent all information as accurately as possible. The result will be about 1.2 million different information units (answers to questions, etc. by all respondents).

This complexity of the survey production process requires adequate organizational and professional production structures. Big survey research companies like Gallup, Harris, Infratest-Burke, Inra, or Mori to name but a few and famous, provide such organizational and professional structures for polls, big market surveys, etc. In the public, nonprofit domain, independent, academic survey research centers serve the academic communities, governments, and even compete with private companies on many sectors.

It is the integration of academic methodological standards and perspectives into the applied survey research process, which makes academic survey research centers different from private companies' research units. Publishing rigorous methodological research, going beyond the 'it works' to find out 'how and why it works,' is thus a must for an academic survey research institute. Hence it is no coincidence that the two most prominent survey research centers today, ISR and NORC, are also centers of methodological excellence. However, methodological research, important as it may be for survey research, is a very small area in the sea of modern social research. It is almost always interdisciplinary and cooperative. In short, methodological survey research is not suited for a career at a university. National research centers by their very size and their mission to accomplish high quality surveys provide a natural base for long-term high quality methodological research also.

National centers for survey research cover national and international issues, which are quite distinct from regional centers. In addition such centers have a broad focus on methods, target audience, and topics. Funding and organization differ from center to center and from nation to nation.

The term 'national center' as it is used here does not exclude several centers within one nation, nor does it imply that the word 'national' in the name immediately qualifies an institute to be a national center. They also serve as hosts for important national surveys. Studies like the General Social Survey (NORC), the Panel Study of Income Dynamics (ISR), British Election Studies (NATCEN), or the German General Social Survey (ZUMA/GESIS) are cornerstones of their countries' social science information systems. Finally, they have an enormous impact on the development of multinational survey research: NATCEN, NORC, ZUMA are founding members of the International Social Survey Programme (ISSP, www.issp.org), ISR hosts the World Value Survey.

Rigorous methods, the public as an audience, free distribution of major nationwide data sets, intensive, published methodological research, and an international perspective are what mark national centers for survey research.

Today, 88 years after the first survey (de Heer et al. 1999, p. 28) centers which truly qualify are still very small in number. This is despite the enormous growth in interviews conducted around the globe (in 1978 the gross revenue of the survey industry was estimated at $4 billion; de Heer et al. 1999, p. 32); the annual turnover reported by ESOMAR in 1998 was about $13 billion with an increase of 10 percent since 1997 (www.esomar.nl).

There are actually only a handful institutions worldwide which can be counted as a national center for survey research: in the US the National Opinion Research Center (NORC, Chicago, founded 1941), the Survey Research Center in the Institute for Social Research (ISR, Ann Arbor, 1946), in Great Britain the National Center for Survey Research (NATCEN (formerly SCPR), London, 1969), the Norwegian Social Science Data Services (NSD, Bergen, 1971), the German Social Science Infrastructure Institutes (GESIS, Mannheim, an umbrella organization for ZUMA in Mannheim, Zentralarchiv in Cologne and the Social Science Information Center in Bonn, 1986).

Other institutes like social science data archives, for instance the Roper Center, ICPSR, or the Essex Archive (cf. CESSDA, www.nsd.uib.no/cessda/ or www.ifdo.org), or the numerous social research institutes listed in clearing houses (www.abacon.com/sociology/soclinks with references to other lists) either do not carry or contract out major national surveys or are not heavily involved in methodological survey research apart from secondary data analysis. Hence they are not covered in this section.

In the following, short bios of the four centers will be presented. These bios will not go into the history of these institutes; rather their current standing will be emphasized. In concluding, the impact to these centers on survey research will be outlined at the beginning of the twenty-first century.

2. *NORC: National Opinion Research Center, Chicago*

NORC was established in 1941 as a nonprofit corporation through a grant from the University of Denver and the Field Foundation. In 1947, NORC moved to the University of Chicago campus (www.norc.uchicago.edu). The center conducts survey research in the public domain as well as for private

organizations. It is an independent nonprofit organization with strong ties to the University of Chicago.

NORC has a reputation for both complex surveys and innovations in survey research methodology. It is the birthplace of surveys conducted for the benefit of the sociological profession at large, the General Social Survey (GSS) (Davis and Smith 1992). The first GSS was conducted as early as 1972. A GSS is a generalized tool which repeatedly measures important characteristics. It is released immediately after completion to the social science community. Thus researchers, students, and other interested parties have up-to-date survey data to hand which they can analyze on their own terms. Such a complex, long-term enterprise requires a strong and well-equipped infrastructure, like the one provided by NORC as a national survey research center. Other strongholds of NORC are methodological studies concentrating on measurement error. Norman Bradburn and Roger Tourangeau are but two among the number of researchers who contributed to this field from the 1980s.

3. *ISR: Institute for Social Research, Ann Arbor*

Shortly after NORC moved to Chicago, in 1948, researchers from the Survey Research Center and the Research Center for Group Dynamics founded ISR in Ann Arbor (www.isr.umich.edu). From its very beginning ISR advocated interdisciplinary research. Its continuous studies range from National Election Studies to the Panel Study of Income Dynamics. It provides major databases as well as substantial contributions in group dynamics and political science to the scientific community. ISR is also a world leader in survey research methodology and statistics. Researchers like Kish, Schuman, Converse, Presser, Groves, Couper, Singer, and Schwarz pave the way for modern survey research processes in areas such as sampling statistics, response error, household-non-response, and survey cognition processes.

Another impact area of ISR is training and data publishing. Since 1962 ISR has hosted the Inter-university Consortium for Political and Social Research (ICPSR) which was founded as an archive for quantitative social science data. From a safe haven for key punched cards ICPSR developed the world's largest survey data publisher/distributor where researchers have access to thousands of data files.

ISR offers three training programs open to researchers worldwide. These are the Summer Institute in Survey Research Techniques, the ICPSR Summer Program in Quantitative Methods and, more recently, the Research Center on Group Dynamics Program in Experimental Methodology and Statistics. The institute puts much emphasis on international cooperation and knowledge dissemination in all its activities.

Apart from a healthy competitive attitude, NORC and ISR are complementary institutes with strong ties to their respective universities (NORC—University of Chicago, ISR—University of Michigan at Ann Arbor). NORC puts more emphasis on fielding many complex studies, while ISR concentrates on specific studies. ISR is very strong in survey methodology, while NORC is stronger in survey process management and sampling design. Both provide a base for international cooperation.

4. *NATCEN: National Centre for Social Research, London*

NATCEN, formerly SCPR, was founded in 1969. Today, it is Britain's largest nonprofit survey research organization (www.natcen.ac.uk). It is the home of British Social Attitudes as Britain's general social survey, of the British Elections studies, and major health surveys, to name but a few. Within the tradition of British social research, NATCEN conducts both quantitative and qualitative studies. Among its divisions are a Survey Methods Centre (research, consulting, and teaching) and joint centers with universities like CASS (Centre for Applied Social Surveys) which hosts a survey item bank. Researchers at NATCEN are leading in applied survey research. Among their innovations is the International Social Survey Programme (ISSP) initiated by Roger Jowell and the first ever deliberative poll. It will be the 'command and engine room' of the European Social Survey to come in 2002.

NATCEN is the only center among the survey research centers discussed here which names itself a 'national center.' However, it is a national center not by Her Majesty's appointment, but *de facto* due to the specific structure of the British science funding system. As an independent charitable trust it forms the base for collaborative mid-range and long-range projects like CREST, the Centre for Research into Elections & Social Trends. Funded as a research center of the Economic and Social Research Council (ESRC), it is a joint venture of Nuffield College in Oxford and NATCEN in London. The above-mentioned CASS, on the other hand, is an ESRC research center joining NATCEN with the University of Southampton.

5. *GESIS: German Society of Social Science Infrastructure Institutes, Mannheim, Cologne, and Bonn*

Founded in 1986 by act of the federal government and all states, GESIS comprises three major centers: ZUMA, Center for Survey Research and Methodology in Mannheim, the Zentralarchiv in Cologne, and the Information Center (www.gesis.org). Together

these three centers form a comprehensive national survey research center as a permanent infrastructure for the social sciences.

ZUMA consults academics in all areas of quantitative survey methods; it runs continuous research streams in areas like standardization of demographics, intercultural survey research, and, recently in web surveys. It is, among others, the home of ALLBUS, the German General Social Survey, the German ISSP, and a participant in the German Social Welfare Studies. Together with researchers from five other European institutes it will be part of the European Social Survey 'engine room' (see above). It is also the birthplace of the German Standard Demography, now a major project of the German Census Office, ZUMA, and other professional groups. The center bridges the gap between German Official Statistics and the social science profession by providing and/or enabling access to official microdata.

The Zentralarchiv, founded as early as 1960, is the German national archive for quantitative survey data. It played a major role in establishing the international archive network. The archive also provides training courses in survey methods.

The Information Center provides in-depth bibliographies of German social science publications, keeps record of ongoing research projects, and serves as a platform for information interchange between Eastern and Western Europe.

Taken together the three centers form a unique infrastructure for their national scientific community. Access to all services is free, products are nominally charged.

The three institutes were independent and differently funded units until the National Science Board (*Wissenschaftsrat*) voted to form a unified national survey research center, namely GESIS. As a social science infrastructure it has a strict supportive function for academic research. Substantive research is different than in America or Britain, and is not a core issue of GESIS. However, as a dedicated social science infrastructure GESIS is fully funded by the federal government and all German states.

6. *Conclusion*

National centers for survey research are defined as institutions that provide databases, methodological research, and training for their scientific community. In addition, national centers focus on the national rather than the regional level and bridge national research with international. There may be more institutions around the globe than the four introduced here. Many others may fulfill one or other of the functions. However, the core of a national center for survey research is dedicated methodological studies embedded into long-term methodological and substantial research programs.

In looking back to get a vision of the future, one could arguably say that the fast progress one could observe in survey research technology and methodology since the 1950s was caused by the very existence of the two American centers. Others countries still lag behind the United States both in advancement of scientific knowledge and in survey engineering. Germany, for instance, where ZUMA was founded as late as 1974 with limited capacities (no fieldwork staff, no own surveys), had to bridge a huge methodological and technological gap since the 1980s. The founding of GESIS as a national center was a late but not too late remedy. It is thus no coincidence that the lack of national centers or their substitutes correlates with the standards of survey research in a country. Because all modern societies need sound information about their own state, efforts should be made to either found more national centers or to transform existing ones into real or virtual transnational centers. Finally, in view of the new survey technologies, like net-based multimedia surveys, and their likely abuse and misuse, national centers become even more indispensable as a means to sort good methods and data from information rubbish.

See also: Data Archives: International; Databases, Core: Political Science and Political Behavior; Databases, Core: Sociology; Economics, History of

Bibliography

Davis J A, Mohler P P, Smith T W 1994 Nationwide general social surveys. In: Borg I, Mohler P P (eds.) *Trends and Perspectives in Empirical Social Research*. Walter de Gruyter, Berlin

Davis J A, Smith T W 1992 *The NORC General Social Survey—A User's Guide*. Sage, Newbury Park, CA

De Heer W, De Leeuw E, van der Zouwen J 1999 Methodological issues in survey research: A historical review. *Bulletin de Methodologie Sociologique* **64**: 25–48

P. P. Mohler

Surveys and Polling: Ethical Aspects

Polls and surveys are used to investigate opinions and attitudes, and to obtain factual information. Two broad aspects of ethical practice with respect to polls and surveys can be distinguished. The first has to do with the extent to which research designs, instruments, data collection procedures, and analysis permit valid and reliable conclusions. The second pertains to the researcher's obligations toward others, especially respondents, but clients and the larger public as well. Because polls and surveys are really indistinguishable

except in terms of location, with polls typically carried out by commercial organizations and surveys by academic institutions (Schuman 1997), the present discussion of ethical practices is intended to apply to both.

Although the discussion is framed in terms of the code of ethics of one organization in a particular country—the Code of Ethics of the American Association for Public Opinion Research (AAPOR), a professional organization including academic, commercial, and government survey researchers among its members—the principles presented here apply much more generally. But because the legal requirements for protecting respondent confidentiality, as well as the ability to do so in practice, vary from one country to another, such variations are not discussed here. Also excluded from the discussion are the fundamentally different ethical issues raised by administrative uses of data gathered for statistical purposes (see, for example, Seltzer 1998).

1. *Standards for the Conduct of Research*

AAPOR's Code of Ethics, which obligates members to adhere to the Code and is one of the few such instruments to provide a mechanism for punishing violations, is necessarily couched in general terms, simply asserting that members will 'exercise due care' in carrying out research and 'take all reasonable steps' to assure the validity and reliability of results. In May 1996, AAPOR adopted two more specific statements, one defining a set of performance standards that embody current 'best practices' in the field of survey research, the other a set of practices that are universally condemned. Members are urged to adhere to the first and refrain from the second. Among the 'best practices' are scientific methods of sampling, which ensure that each member of the population has a measurable chance of selection; follow-up of those selected to achieve an adequate response rate; careful development and pretesting of the questionnaire, with attention to known effects of question wording and order on response tendencies; and adequate training and supervision of interviewers. Among practices defined as 'unacceptable' are fund raising, selling, or canvassing under the guise of research; representing the results of a poll of self-selected or volunteer respondents (for example, readers of a magazine who happen to return a questionnaire, callers to an 800- or 900-number poll, or haphazard respondents to a web-based survey) as if they were the outcome of legitimate research; and 'push polls,' which feed false or misleading information to voters under the pretense of taking a poll. For example, a push-poll might ask if the respondent would object to learning that candidate X has been accused of child abuse, has defaulted on a loan, etc. Revealing the identity of individual respondents to anyone without their permission, and misleading the public by selecting or manipulating methods of research or analysis to reach a desired conclusion, are included among 'unacceptable' practices but they are also mentioned in the AAPOR Code.

2. *Standards for Dealing with Clients, the Public, and Respondents*

2.1 *Obligations to Clients*

Ethical behavior toward clients requires, in the first place, undertaking only those research assignments that can reasonably be accomplished, given the limitations of survey research techniques themselves or limited resources on the part of the researcher or client. Second, it means holding information about the research and the client confidential, except as expressly authorized by the client or as required by other provisions of the Code. Chief among the latter is the requirement that researchers who become aware of serious distortions of their findings in public must disclose whatever may be needed to correct these distortions, including a statement to the group to which the distorted findings were presented. In practice, this provision of the Code may obligate researchers to issue a corrective statement to a legislative body, a regulatory agency, the media, or some other appropriate group if they become aware of the release of such distorted results.

2.2 *Obligations to the Public*

The AAPOR Code of Ethics attempts to satisfy the profession's obligations to the larger public by mandating the disclosure of minimal information about a poll or survey whose findings are publicly released. The elements that must, at a minimum, be disclosed are the following: (a) who sponsored the survey, and who conducted it; (b) the exact wording of questions asked, including the text of any preceding instruction or explanation to the interviewer or respondent that might reasonably be expected to affect the response; (c) a definition of the population under study and a description of the sampling frame; (d) a description of the sample selection procedure, making clear whether respondents were selected by the researcher or were self-selected; (e) size of sample and, if applicable, completion rates and information on eligibility criteria and screening procedures; (f) a discussion of the precision of the findings, including, if appropriate, estimates of sampling error and a description of any weighting or estimating procedure used; (g) which results, if any, are based on parts of the sample rather than the entire sample; (h) method, location, and dates of data collection. Together, these items of information communicate some idea of the error likely to

be associated with estimates from a given survey, or, conversely, the degree of confidence warranted by the results. The Code requires this information to be included in any report of research results, or, alternatively, made available at the time the report is released. In other words, researchers are not required by the Code of Ethics to follow the best professional practices, but they are required to disclose the practices they used. Thus, the Code relies on the 'marketplace of ideas' to drive out bad practices over time.

2.3 Obligations to Respondents

Few, if any, surveys put respondents at physical or psychological risk in ways that psychological experiments or, particularly, medical research may do. Ordinarily, the biggest risk to survey respondents comes from the possibility that their identity will be disclosed, either inadvertently, because adequate precautions were not taken to maintain the confidentiality of the data, or deliberately, as a result of special efforts by an 'intruder' to obtain the identity of one or more respondents. Thus, survey organizations have a special responsibility both to protect the confidentiality of the data they collect and to inform respondents of the risks that remain. These are two of the ethical issues that will be discussed in this section; the third arises from declining survey response rates and attempts to compensate for these by special persuasion strategies, including the offer of refusal conversion payments.

2.3.1 Informed consent in surveys and polls. One important purpose of informed consent is to protect respondents from harm. Another purpose, less often cited, is to give them meaningful control over information about themselves, even if the question of harm does not arise. For example, informing potential respondents of the voluntary nature of their participation in the interview is essential to ethical survey practice, even if this leads some potential respondents to refuse to participate. Similarly, obtaining the respondents' permission to record an interview should be standard practice, even though no harm is likely to come to them as a result. Currently, many states permit such recording with only the interviewer's consent.

At present, obtaining informed consent, which may be defined as the 'knowing consent of an individual or his legally authorized representative … without undue inducement or any element of force, fraud, deceit, duress, or any other form of constraint or coercion,' is not legally mandated for participation in most surveys in the United States. Even those surveys that are subject to federal rules for the protection of human subjects are exempt unless they collect respondent identification and disclosure of responses could reasonably be damaging to respondents' good reputation, employability, or financial standing or could put them at risk of civil or criminal liability (*Federal Register* 1991, 28:012). (Surveys of populations that are seen as especially vulnerable, such as minors or incapacitated or institutionalized adults, are required to institute special procedures for obtaining informed consent, however.) But surveys conducted by federal agencies *are* subject to the Privacy Act of 1974, which requires the disclosure of some information to respondents—for example, the principal purpose for which the survey is to be used (Duncan et al. 1993).

Although informed consent is not *legally* required for most surveys in the United States, some major professional associations (for example, the American Statistical Association and the American Sociological Association) do make mention of it in their ethical codes. AAPOR, however, does not, noting only that 'We shall strive to avoid the use of practices or methods that may harm, humiliate, or seriously mislead survey respondents.'

It is not difficult to understand this omission. Although full disclosure of sponsorship and purpose, for example, may be highly desirable from an ethical point of view, it may seriously bias responses to a survey. It may do so not only because some people will choose not to participate in surveys sponsored by certain organizations or with a particular purpose, but also because those who do participate may distort their responses in order to thwart or advance the purpose as they understand it. From another standpoint, the issue is how much detail should be communicated in order to 'inform' respondents— should they, for example, be told in advance that they are going to be asked their income near the end of the interview? Although Presser (1994) argues persuasively that the uses to which a survey will be put are often impossible to predict and even harder to communicate to respondents with any accuracy, he acknowledges that in the end, there 'may be no escaping the conflict of ethical obligations with research ends' (Presser 1994, p. 454).

2.3.2 Protection of respondents' confidentiality. As has already been noted, probably the most important source of danger to survey respondents comes from identification of their responses, that is, from deliberate or inadvertent breaches of confidentiality. The more sensitive the information requested—for example, HIV status or whether the respondent has engaged in some illegal activity—the greater the harm that a breach of confidentiality might entail. Not unexpectedly, respondents require strong assurances of confidentiality when they are asked to provide such data about themselves (Singer et al. 1995).

A variety of threats to the confidentiality of survey data exist. Most common is simple carelessness—not

removing identifiers from questionnaires or electronic data files, leaving cabinets unlocked, not encrypting files containing identifiers. Although there is no evidence of respondents having come to harm as a result of such negligence, it is important for data collection agencies to raise the consciousness of their employees with respect to these issues, provide guidelines for appropriate behavior, and ensure that these are observed.

Less common, but potentially more serious, threats to confidentiality are legal demands for identified data, either in the form of a subpoena or as a result of a Freedom of Information Act (FOIA) request. Surveys that collect data about illegal behaviors such as drug use, for example, are potentially subject to subpoena by law-enforcement agencies. To protect against this, researchers studying mental health, alcohol and drug use, and other sensitive topics, whether federally funded or not, may apply for Certificates of Confidentiality from the Department of Health and Human Services. Such certificates, which remain in effect for the duration of the study, protect researchers from being compelled in most circumstances to disclose names or other identifying characteristics of survey respondents in federal, state, or local proceedings (42 C.F.R. 2a.7, 'Effect of Confidentiality Certificate'). The Certificate protects the identity of respondents; it does not protect the data.

However, when surveys are done expressly for litigation, demands for data from one party or the other are not uncommon, and this protection may not be available. Several such cases are reviewed by Presser (1994). In one, the case was settled before the data were to be produced; in two others, the courts ordered the researchers to turn over raw data, including respondent identifiers, to the defendants. Presser (1994) discusses several actions that might be taken by professional survey organizations to enhance protections of confidentiality in such cases—for example, mandating the destruction of identifiers in surveys designed for adversarial proceedings. If this became standard practice, individual researchers would be better able to defend their own use of such a procedure.

A final threat to data confidentiality that should be mentioned here is what is known as 'statistical disclosure.' So far, the discussion has been in terms of explicit respondent identifiers—name, address, and social security number, for example. But there is increasing awareness that even without explicit identifiers, re-identification of respondents may be possible given the existence of high-speed computers, external data files to which the survey data can be matched, and sophisticated matching software. Longitudinal studies may be especially vulnerable to such re-identification.

There are essentially two ways to reduce the risk of statistical disclosure: restricting access to the data, for example by enclave or bonding arrangements, or restricting the data that may be released, for example, by top-coding age and income and releasing detailed data only for large geographic areas. Although guidelines for reducing the risk of statistical disclosure exist (Federal Committee on Statistical Methodology 1994), some researchers believe that the best way to safeguard confidentiality is to create synthetic data sets derived from the actual information provided by respondents (Rubin 1987).

In view of the diverse threats to confidentiality reviewed above, there is a growing consensus among researchers that absolute assurances of confidentiality should seldom, if ever, be given to respondents (Duncan et al. 1993). The 'Ethical Guidelines for Statistical Practice' of the American Statistical Association (ASA), for example, caution statisticians to 'anticipate secondary and indirect uses of the data when obtaining approvals from research subjects.' At a minimum, the Guidelines urge statisticians to provide for later independent replication of an analysis by an outside party. They also warn researchers not to imply protection of confidentiality from legal processes of discovery unless they have been explicitly authorized to do so, for example, by a Certificate of Confidentiality. Research by Singer (1978) and Singer et al. (1995) suggests that for surveys on nonsensitive topics, neither the response rate nor data quality is likely to suffer as a result of such qualified assurances of confidentiality. For those collecting sensitive data, researchers should make every effort to obtain a Certificate of Confidentiality, which would permit them to offer stronger assurances to respondents. Even such Certificates, however, cannot guard against the risk of statistical disclosure.

2.3.3 Efforts at persuasion. An issue that merits brief discussion here is the dilemma created by the need for high response rates from populations that spend less time at home, have less discretionary time, and are less inclined to cooperate in surveys. This dilemma leads survey organizations to attempt more refusal conversions and to make greater use of refusal conversion payments. Respondents see both practices as undesirable (Groves et al. 1999). They also regard the paying of refusal conversion payments as unfair, but awareness of such payments does not appear to reduce either the willingness to respond to a future survey or actual participation in a new survey (Singer et al. 1999).

2.3.4 Polling on the Net. Polls and surveys conducted on the Internet have shown explosive growth in the last several years, and such surveys bring an entirely new set of ethical problems to the fore, or bring new dimensions to old problems. For example, maintenance of privacy and confidentiality in such surveys requires technological innovations (for example, 'secure' websites, encryption of responses) beyond those

required by older modes of data collection. Other ethical issues simply do not arise with older modes of conducting surveys—for example, the possibility that a respondent might submit multiple answers, thus deliberately biasing the results—and these, too, call for new technical solutions requiring sophisticated knowledge of computer hardware and software. These problems, and proposed solutions, are too new to permit more than brief mention of them here.

3. Concluding Observations

A number of developments in recent years have increased ethical dilemmas for survey researchers. There is growing demand for information about individuals on the part of both business and government. At the same time, there is growing concern about personal privacy on the part of the public. Technological developments have made safeguarding privacy and confidentiality increasingly difficult, and for this as well as other reasons, there is increasing reluctance on the part of the public to participate in polls and surveys, including government censuses. How these dilemmas are resolved in the next few years will have profound implications for the future of survey research.

See also: Confidentiality and Statistical Disclosure Limitation; Ethical Dilemmas: Research and Treatment Priorities; Personal Privacy: Cultural Concerns; Privacy: Legal Aspects; Privacy of Individuals in Social Research: Confidentiality; Psychiatry: Informed Consent and Care; Research Conduct: Ethical Codes; Research Ethics: Research; Research Subjects, Informed and Implied Consent of; Social Survey, History of

Bibliography

Duncan G T, Jabine T B, DeWolf V A (eds.) 1993 *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academy Press, Washington, DC

Federal Committee on Statistical Methodology 1994 *Working Paper 22: Report on Statistical Disclosure Limitation Methods*. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC

Federal Register 1991 Federal policy for the protection of human subjects. *Federal Register* **June 18**: 28,002–31

Groves R M, Singer E, Corning A D, Bowers A 1999 A laboratory approach to measuring the joint effects of interview length, incentives, differential incentives, and refusal conversion on survey participation. *Journal of Official Statistics* **15**: 251–68

Presser S 1994 Informed consent and confidentiality in survey research. *Public Opinion Quarterly* **58**: 446–59

Rubin D B 1987 *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York

Schuman H 1997 Polls, surveys, and the English language. *The Public Perspective* **April/May**: 6–7

Seltzer W 1998 Population statistics, the Holocaust, and the Nuremberg trials. *Population and Development Review* **24**: 511–52

Singer E 1978 Informed consent: Consequences for response rate and response quality in social surveys. *American Sociological Review* **43**: 144–62

Singer E, Groves R M, Corning A D 1999 Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation. *Public Opinion Quarterly* **63**: 251–60

Singer E, Von Thurn D R, Miller R E 1995 Confidentiality assurances and response: A quantitative review of the experimental literature. *Public Opinion Quarterly* **59**: 266–77

E. Singer

Survival Analysis: Overview

1. Failure-time Data and Representation

The term survival analysis has come to embody the set of statistical techniques used to analyze and interpret time-to-event, or failure-time data. The events under study may be disease occurrences or recurrences in a biomedical setting, machine breakdowns in an industrial context, or the occurrence of major life events in a longitudinal-cohort setting. It is important in applications that the event of interest be defined clearly and unambiguously. A failure-time variable T is defined for an individual as the time from a meaningful origin or starting point, such as date of birth, or date of randomization into a clinical trial, to the occurrence of an event of interest. For convenience, we refer to the event under study as a failure.

A cohort study involves selecting a representative set of individuals from some target population and following those individuals forward in time in order to record the occurrence and timing of the failures. Often, some study subjects will still be without failure at the time of data analysis, in which case their failure times T will be known only to exceed their corresponding censoring times. In this case, the censoring time is defined as the time elapsed from the origin to the latest follow-up time, and the failure time is said to be right censored. Right censorship may also arise because subjects drop out the study, or are lost to follow-up. A key assumption underlying most survival analysis methods is that the probability of censoring an individual at a given follow-up time does not depend on subsequent failure times. This independence assumption implies that individuals who are under study at time t are representative of the cohort and precludes, for example, the censoring of individuals at time t because, among individuals under study at that time, they appear to be at unusually high (or low) risk of failure. (This assumption can be relaxed to independence between censoring and failure times conditional on the preceding histories of covariates.) The

reasonableness of the independent censoring assumption, which usually cannot be tested directly, should be considered carefully in specific applications.

The distribution of a failure time variate $T > 0$ is readily represented by its survivor function $F(t) = P(T > t)$. If T is absolutely continuous, its distribution can also be represented by the probability density function $f(t) = -dF(t)/dt$ or by the hazard, or instantaneous failure rate, function defined by

$$\lambda(t) = \lim_{\Delta t \downarrow 0} P(t \leqslant T < t + \Delta t \mid T \geqslant t)/\Delta t = f(t)/F(t^-). \tag{1}$$

Note that

$$F(t) = \exp\{-\int_o^t \lambda(u)\,\mathrm{d}u\}, \quad \text{and}$$

$$f(t) = \lambda(t)\exp\{-\int_o^t \lambda(u)\,\mathrm{d}u\}$$

for $0 < t < \infty$

In a similar way, the distribution of a discrete failure time variate T, taking values on $x_1 < x_2 < \cdots$, can be represented by its survivor function, its probability function $f(x_i) = P(T = x_i)$, $i = 1, 2 \ldots$, or its hazard function $\lambda(x_i) = P(T = x_i | T \geqslant x_i) = f(x_i) | F(x_i^-)$. Note that

$$F(t) = P(T > t) = \prod_{i \mid x_i \leqslant t} [1 - \lambda(x_i)] \tag{2}$$

as arises through thinking about the failure mechanism unfolding sequentially in time. To survive past time t where $x_j \leqslant t < x_{j+1}$, an individual must survive in sequence each of the points $x_1, \ldots, x_j$ with conditional probabilities $1 - \lambda(x_i)$, $i = 1, \ldots, j$.

It is often natural and insightful to focus on the hazard function. For example, human mortality rates are known to be elevated in the first few years of life, to be low and relatively flat for the next three decades or so, and then to increase rapidly in the later decades of life. Similarly, comparison of hazard rates as a function of time from randomization for a new treatment compared with control can provide valuable insights into the nature of treatment effects. Hazard function modeling also integrates nicely with a number of key strategies for selecting study subjects from a target population. For example, if one is studying disease occurrence as it depends upon the age of study subjects, hazard function models readily accommodate varying ages at entry into the study cohort since the conditioning event in (1) specifies that the individual is surviving and event-free at age t. Again one needs to assume that selected subjects have hazard rates that are representative (conditional on covariates) of those at risk for the target population.

2. Estimation and Comparison of Survival Curves

Suppose that a cohort of individuals has been identified at a time origin $t = 0$ and followed forward in time. Suppose that failures are observed at times $t_1 < t_2 < \cdots < t_k$. Just prior to the failures observed at time t_j, there will be some number, r_j, of cohort members that are under active follow-up. That is, the failure and censoring times for these r_j members occur at time t_j or later. Of these r_j individuals, some number, d_j, are observed to fail at t_j. The empirical failure rate or estimated hazard at t_j is then d_j/r_j, and the empirical survival rate is $(1 - d_j/r_j)$. The empirical failure rate at all times other than $t_1, \ldots, t_k$ in the follow-up period is zero. The Kaplan and Meier (1958) estimator of the survivor function F at follow-up time t is

$$\hat{F}(t) = \prod_{\{t_j \leqslant t\}} (1 - d_j/r_j) \tag{3}$$

the product of the empirical survivor rates across the observed failure times less than or equal to t. Note that $\hat{F}$ estimates the underlying survivor function F at those times where there are cohort members at risk. For (3) to be valid, the cohort members at risk at any follow-up time t must have the same failure rate operating as does the target population at that follow-up time. Thus, the aforementioned independent censoring assumptions are needed.

The hazard rate estimators d_j/r_j, $j = 1, \ldots, k$ can be shown to be uncorrelated, giving rise to the Greenwood variance estimator

$$\mathrm{var}\,\{\hat{F}(t)\} = \hat{F}(t)^2 \sum_{\{t_i \leqslant t\}} d_j/\{r_j(r_j - d_j)\} \tag{4}$$

for $\hat{F}(t)$. In order to show asymptotic results for the estimator $\hat{F}(t)$, for example that $\hat{F}(t)$ is a consistent estimator of $F(t)$, it is necessary to restrict the estimation to a follow-up interval $[0, \tau)$ within which the size of the risk set (i.e., number of individuals under active follow-up) becomes large as the sample size (n) increases. Under these conditions, $n^{1/2}\{\hat{F}(t) - F(\mathrm{t})\}$; $0 \leqslant t < \tau$ can be shown to converge weakly to a mean-zero Gaussian process, permitting the calculation of approximate confidence intervals and bands for F on $[0, \tau)$. Applying the asymptotic results to the estimation of $\log\{-\log F(t)\}$, which has no range restrictions, is suggested as a way to improve the coverage rates of asymptotic confidence intervals and bands (Kalbfleisch and Prentice 1980, pp. 14, 15).

Consider a test of equality of m survivor functions $F_1, \ldots, F_m$ on the basis of failure-time data from each of m populations. Let $t_1 < t_2 < \cdots < t_k$ denote the ordered failure times from the pooled data, and again let d_j failures occur out of the r_j individuals at risk just prior to time t_j, $j = 1, \ldots, k$. Also, for each sample $i = 1, \ldots, m$, let d_{ij} denote the number of failures at t_j, and r_{ij} denote the number at risk at t_j^-. Under the null

hypothesis $F_1 = F_2 = \cdots = F_m$ the conditional distribution of $d_{1j},\ldots, d_{mj}$ given $(d_j, r_{1j},\ldots, r_{mj})$ is hypergeometric from which the conditional mean and variance of d_{ij} are respectively $u_{ij} = r_{ij}(d_j/r_j)$ and $(V_j)_{ii} = r_{ij}(r_i - r_{ij})d_j(r_j - d_j)r_j^{-2}(r_j - 1)^{-1}$. The conditional covariance of d_{ij} and $d_{\ell j}$ is

$$(V_j)_{i\ell} = r_{ij}r_{\ell j}d_j(r_j - d_j)r_j^{-2}(r_j - 1)^{-1}. \tag{5}$$

Hence the statistic $v_j' = (d_{1j} - u_{1j},\ldots, d_{mj} - u_{mj})$ has conditional mean zero and variance V_j under the null hypothesis. The summation of these differences across failure times,

$$v = \sum_1^k v_j \tag{6}$$

is known as the logrank statistic (Mantel 1966, Peto and Peto 1972). The contributions to this statistic across the various failure times are uncorrelated, and $v'V^{-1}v$ generally has an asymptotic chi-square distribution with $m-1$ degrees of freedom where $V = \sum_1^k V_j$ and a 'prime' denotes vector transpose.

This construction also shows that a broad class of 'censored data rank tests' of the form $\sum_1^k w_j v_j$ also provide valid tests of the equality of the m survival curves. In these tests, the 'weight' w_j at time t_j can be chosen to depend on the above conditioning event at t_j, or on data collected before t_j, more generally. The null hypothesis variance of these test statistics is readily estimated by $\sum_1^k w_j^2 V_j$, again leading to an asymptotic chi-square $(m-1)$ test statistic. The special case $w_j = \hat{F}(t_j^-)$ generalizes the Wilcoxon rank sum test to censored data in a natural fashion (e.g., Peto and Peto 1972, Prentice 1978) and provides an alternative to the logrank test that assigns greater weight to earlier failure times. More adaptive tests can also be entertained (e.g., Harrington and Fleming 1982).

3. *Failure-time Regression: The Cox Model*

3.1 *Model Development*

The same type of conditional probability argument generalizes to produce a flexible and powerful regression method for the analysis of failure-time data. This method was proposed in the fundamental paper by Cox (1972) and is referred to as Cox regression or as relative risk regression.

Suppose that each subject has an associated vector of covariates $z' = (z_1, z_2,\ldots)$. For example, the elements of z may include primary exposure variables and confounding factor measurements in an epidemiologic cohort study, or treatment indicators and patient characteristics in a randomized clinical trial. Let $\lambda(t; z)$ denote the hazard rate at time t for an absolutely continuous failure time T on an individual with covariate vector z. Now one requires the censoring patterns, and the cohort entry patterns, to be such that the hazard rate $\lambda(t; z)$ applies to the cohort members at risk at follow-up time t and covariate value z for each t and z. The Cox regression model specifies a parametric form for the hazard rate ratio $\lambda(t; z)/\lambda(t; z_0)$, where z_0 is a reference value (e.g., $z_0 = 0$). Because this ratio is nonnegative, it is common to use an exponential parametric model of the form $\exp\{x(t)\beta\}$. The vector $x(t) = \{x_1(t),\ldots, x_p(t)\}$ includes derived covariates that are functions of the elements of z (e.g., main effects and interactions) and possibly product (interaction) terms between z and t. The vector $\beta' = (\beta_1,\ldots, \beta_p)$ is a p-vector of regression parameters to be estimated. Denoting $\lambda_0(t) = \lambda(t; z_0)$ allows the model to be written in the standard form

$$\lambda(t; z) = \lambda_0(t)\exp\{x(t)\beta\}. \tag{7}$$

As in previous section let $t_1 < t_2 < \cdots < t_k$ denote the ordered distinct failure times observed in follow-up of the cohort, and let d_j denote the number of failures at t_j. Also let $R(t_j)$ denote the set of labels of individuals at risk for failure at t_j^-. We suppose first that there are no ties in the data so that $d_j = 1$. The conditional probability that the ith individual in the risk set with covariate z_i fails at t_j, given that $d_j = 1$ and the failure, censoring (and covariate) information prior to t_j, is simply

$$\frac{\lambda(t_j; z_i)}{\sum_{\ell \in R(t_j)}\lambda(t_j; z_\ell)} = \frac{\exp\{x_i(t_j)\beta\}}{\sum_{\ell \in R(t_j)}\exp\{x_l(t_j)\beta\}}. \tag{8}$$

The product of such terms across the distinct failure times

$$L(\beta) = \prod_{j=1}^k \frac{\exp\{x_j(t_j)\beta\}}{\sum_{\ell \in R(t_j)}\exp\{x_\ell(t_j)\beta\}} \tag{9}$$

is known as the partial likelihood function for the relative risk parameter β.

If one denotes $L(\beta) = \prod_{j=1}^k L_j(\beta)$, then the fact that the scores $\partial \log L_j(\beta)/\partial\beta$, $j = 1,\ldots, k$ have mean zero and are uncorrelated (Cox 1975) suggests that, under regularity conditions, the partial likelihood (9) has the same asymptotic properties as does the usual likelihood function. Let $\hat{\beta}$ denote the maximum partial likelihood estimate that solves $\partial \log L(\beta)/\partial\beta = 0$. Then, $n^{1/2}(\hat{\beta} - \beta)$ will generally converge, as the sample size n becomes large, to a p-variate normal distribution having mean zero and covariance matrix consistently estimated by $-n^{-1}\partial^2 \log L(\hat{\beta})/\partial\hat{\beta}\partial\hat{\beta}'$. This result provides the usual basis for tests and confidence intervals on the elements of β.

Even though we are working with continuous models, there will often be some tied failure times. If the ties are numerous, it may be better to model the data using a discrete failure-time model. Often, however, a simple modification to (9) is appropriate.

When the d_j are small compared with the r_j, the approximate partial likelihood (Peto and Peto 1972, Breslow 1974)

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp\{s_j(t_j)\beta\}}{[\sum_{\ell \in R(t_j)} \exp\{x_l(t_j)\beta\}]^{d_j}} \tag{10}$$

is quite adequate for inference, where $s_j(t_j) = \sum_{i \in D_j} x_i(t_j)$ and D_j represents the set of labels of individuals failing at t_j. Efron (1977) suggested an alternate approximation that is only slightly more complicated and is somewhat better. Both the Breslow and Efron approximations are available in software packages.

3.2 Special Cases and Generalizations

Some special cases of the modeled regression vector $x(t)$ illustrate the flexibility of the relative risk model (7). If one restricts $x(t) = x$, to be independent of t, then the hazard ratios at any pair of z-values are constant. This gives the important proportional hazards special case of (7). For example, if the elements of z include indicator variables for $m-1$ of m samples and $x(t) = (z_1,\ldots, z_{m-1})$ then each e^{β_j}, $j = 1,\ldots, m-1$ denotes the hazard ratio for the jth sample compared with the last (mth) sample. In this circumstance the score test for $\beta = 0$ arising from (9) reduces to the logrank test of the preceding section.

The constant hazard ratio assumption, in this example, can be relaxed by defining $x(t) = \{g(t)z_1,\ldots, g(t)z_{m-1}\}$ for some specified function g of follow-up time. For example $g(t) = \log t$ would allow hazard ratios of the form t^{β_j} for the jth sample compared with the mth. This gives a model in which hazard ratio for the jth vs. the mth population increase ($\beta_j > 0$) or decrease ($\beta_j < 0$) monotonely from unity as t increases. The corresponding score test would be a weighted logrank test for the global null hypothesis. More generally one could define $x(t) = \{z_1,\ldots, z_{m-1}, g(t)z_1,\ldots, g(t)z_{m-1}\}$. This model is hierarchical with main effects for the sample indicators and interactions with $g(t)$. A test that the interaction coefficients are zero provides an approach to checking for proportionality of the hazard functions. Departure from proportionality would imply the need to retain some product terms in the model, or to relax the proportionality assumption in some other way.

As a second special case consider a single quantitative predictor variable z. Under a proportional hazards assumption one can include various functions of z in the modeled (time independent) regression vector $x(t) = x$ for model building and checking, just as in ordinary linear regression. The proportionality assumption can be relaxed in various ways including setting $x(t) = \{xI_1(t),\ldots, xI_q(t)\}$ where $I_1(t),\ldots, I_q(t)$ are indicator variables for a partition of the study follow-up period. The resulting model allows the relative risk parameters corresponding to the elements of $x = x(z)$ to be separately estimated in each follow-up interval. This provides a means for testing the proportionality assumption, and relaxing it as necessary to provide an adequate representation of the data.

The above description of the Cox model (7) assumed that a fixed regression vector z was measured on each study subject. It is important to note that (7) can be relaxed to include measurements of stochastic covariates that evolve over time. This does not complicate the development of (9) or (10). Specifically, if $z(t)$ denotes a study subject's covariates at follow-up time t, and $Z(t) = \{z(u); u < t\}$ denotes the corresponding covariate history up to time t, then the hazard rate $\lambda(t; Z(t)\}$ can be modeled according to the right side of (7) with $x(t)$, comprising functions of $Z(t)$ and product terms between such functions and time.

Careful thought must be given to relative risk parameter interpretation in the presence of stochastic covariates. For example, the interpretation of treatment effect parameters in a randomized controlled trial may be affected substantially by the inclusion of postrandomization covariates when those covariates are affected by the treatment. However, careful use of such covariates can serve to identify mechanisms by which the treatment works.

3.3 Survivor Function Estimation

The cumulative 'baseline' hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)\mathrm{d}u$ (7) can be simply estimated (Breslow 1974) by

$$\hat{\Lambda}_0(t; \hat{\beta}) = \sum_{t_j \leqslant t} \left[d_j \Big/ \sum_{\ell \in R(t_j)} \exp\{x_\ell(t_j)\hat{\beta}\} \right]. \tag{11}$$

If $x(t) = x$ is time-independent, the survivor function $F(t; z) = p\{T > t|z\} = \exp\{-\Lambda_0(t)e^{x\beta}\}$ is estimated by

$$\hat{F}(t; z) = \exp\{-\hat{\Lambda}_0(t; \hat{\beta})\mathrm{e}^{x\hat{\beta}}\} \tag{12}$$

A slightly more complicated estimator $\hat{F}(t; z)$ is available if x is time-dependent.

A practically important generalization of (7) allows the baseline hazard to depend on strata $s = s(z)$, or more generally $s = s(z, t)$ so that

$$\lambda(t; z) = \lambda_{0s}(t) \exp\{x(t)\beta\} \tag{13}$$

where s takes integer values $1, 2,\ldots, m$ indexing the strata. This provides another avenue for relaxing a proportional hazards assumption. The stratified partial likelihood is simply a product of terms (9) over the various strata.

3.4 Asymptotic Distribution Theory

The score function $U(\beta) = \partial \log L(\beta)/\partial\beta$ from (9) involves a sum of terms that are uncorrelated but not

independent. The theory of counting processes and martingales provides a framework in which this uncorrelated structure can be described, and a formal development of asymptotic distribution theory. Specifically, the counting process N for a study subject is defined by $N(t) = 0$ prior to the time of an observed failure, and $N(t) = 1$ at or beyond the subject's observed failure time. The counting process can be decomposed as

$$N(t) = \Lambda(t) + M(t) \tag{14}$$

where, under independent censorship,

$$\Lambda(t) = Y(t)\int_0^t \lambda\{u; Z(u)\}\mathrm{d}u \tag{15}$$

and $M(t)$ is a locally square integrable martingale. Note that in the 'compensator' Λ the 'at risk' process Y is defined by $Y(t) = 1$ up to the time of failure or censoring for the subject and zero thereafter. The decomposition (14) allows the logarithm of (9), and its derivatives, and the cumulative hazard estimator (11), to be expressed as stochastic integrals with respect to martingales, to which powerful convergence theorems apply. Anderson and Gill (1982) provide a thorough development of asymptotic distribution theory for the Cox model using these counting process techniques, and provide weak conditions for the joint asymptotic convergence of $n^{1/2}(\hat{\beta}-\beta)$ and to $\hat{\Lambda}_0(\cdot, \hat{\beta})$ to Gaussian processes. The books by Fleming and Harrington (1991) and Andersen et al. (1993) use the counting process formulation for the theory of each of the statistics mentioned above, and for various other tests and estimators.

4. Other Failure-time Regression Models

4.1 Parametric Models

Parametric failure-time regression models may be considered as an alternative to the semiparametric Cox model (7). For example, requiring $\lambda_0(t)$ in (7) to be a constant or a power function of time gives exponential and Weibull regression models respectively. The maximum partial likelihood estimate $\hat{\beta}$ of the preceding section is known to be fully efficient within the semiparametric model with $\lambda_0(t)$ in (7) completely unspecified. While some efficiency gain in β estimation is possible if an assumed parametric model is true, such improvement is typically small and may be more than offset by a lack of robustness.

Of course there are many parametric regression models that are not contained in the Cox model (7). One such class specifies a linear model for the logarithm of time to failure,

$$\log T = \alpha + x\beta + \sigma u \tag{16}$$

where $x = x(z)$ is comprised of functions of a fixed regression vector z. The parameters α and σ are location and scale parameters, β is a vector of regression parameters and u is an error variate. Because regression effects are additive on log T, or multiplicative on T, this class of models is often referred to as the accelerated failure-time model. If the error u has an extreme value distribution with density $f(u) = \exp(u - \mathrm{e}^u)$, then the resulting model for T is the Weibull regression model previously mentioned. More generally, any specified density for u induces a hazard-function model $\lambda(t; z)$ for T to which standard likelihood estimation procedures apply, and asymptotic distribution theory can be developed based again on counting process theory.

4.2 A Competing Semiparametric Model

The model (16) has also been studied as a semiparametric class in which the distribution of the error u is unspecified. Prentice (1978) considered the likelihood for the censored data rank vector based on the 'residuals' $v(\beta_0) = \log T - x\beta_0$ to generate a linear rank statistic

$$\sum_{j=1}^{k}(x_j - \bar{x}_j)w_j v_j(\beta_0) \tag{17}$$

for testing $\beta = \beta_0$. In (17), $v_j(\beta_0)$ is the jth largest of the uncensored $v(\beta_0)$ values, x_j is the corresponding modeled covariate value, $\bar{x}_j$ is the average of the x-values for the individuals having 'residual' equal or greater than $v_j(\beta_0)$, and w_j is a weight, usually taken to be identically equal to one (logrank scores). An estimator of $\hat{\beta}$ of β in (17) can be defined as a value of β_0 for which the length of the vector (17) is minimized. The calculation of $\hat{\beta}$ is complicated by the fact that (17) is a step function that requires a reranking of the residuals at each trial value of β_0. However, simulated annealing methods have been developed (Lin and Geyer 1992) for this calculation, and resampling methods have been proposed for the estimation of confidence intervals for the elements of β (Parzen et al, 1994). See Ying (1993) for an account of the asymptotic theory for $\hat{\beta}$.

The model (16) implies a hazard function of the form

$$\lambda(t; z) = \lambda_0(te^{-x\beta})\mathrm{e}^{-x\beta}. \tag{18}$$

This form suggests a time-dependent covariate version of the accelerated failure-time model (e.g., Cox and Oakes 1984, Robins and Tsiatis 1992) given by

$$\lambda\{t, Z(t)\} = \lambda_0\left\{\int_0^t \mathrm{e}^{-x(u)\beta}\mathrm{d}u\right\}\mathrm{e}^{-x(t)\beta} \tag{19}$$

which retains the notion of the preceding covariate

history implying a rescaling or acceleration, by a factor $e^{-x(t)\beta}$, of the rate at which a study subject traverses the failure-time axis.

5. Multiple Failure Times and Types

5.1 Multiple Failure Types and Competing Risks

Suppose that individuals are followed for the occurrence of a univariate failure time, but that the failure may be one of several types. The type i specific hazard function at follow-up time t, for an individual having preceding covariate $Z(t)$ can, for example, be specified as

$$\lambda_i\{t; Z(t)\} = \lambda_{0i}(t) \exp\{x(t)\beta_i\} \tag{20}$$

where λ_{0i} is a baseline hazard for the ith failure type and β_i is a corresponding relative risk parameter. Estimation of $\beta_1, \beta_2, \ldots$, under independent censorship is readily based on a partial likelihood function that is simply a product of terms (9) over distinct failure types (e.g., Kalbfleisch and Prentice 1980, Chap. 7). In fact, inference on a particular regression vector β_i and corresponding cumulative hazard function Λ_{0i} is based on (9) upon censoring the failures of types other than (i) at their observed failure times. Accelerated failure time regression methods similarly generalize for the estimation of type-specific regression parameters.

Very often questions are posed that relate to the relationship between the various types of failure. It should be stressed that data on failure time and failure type alone are insufficient for studying the interrelation among the failure mechanisms for the different types, rather one generally requires the ability to observe the times of multiple failures on specific study subjects.

5.2 Recurrent Events

Suppose that there is a single type of failure, but that individual study subjects may experience multiple failures over time. Examples include multiple disease recurrences among cancer patients, multiple job losses in an employment history study, or multiple breakdowns of a manufactured item. The counting process N at follow-up time t is defined as the number of failures observed to occur on the interval $(0, t]$ for the specific individual. The asymptotic theory of Andersen and Gill (1982) embraced such counting processes in which the failure rate at time t was of Cox model form

$$\lambda\{t; H(t)\} = \lambda_0(t) \exp\{x(t)\beta\}. \tag{21}$$

In this expression, the conditioning event $H(t)$ includes the individual's counting process history, in addition to the corresponding covariate history prior to follow-up time t. Hence the modeled covariate $x(t)$ must specify the dependence of the failure rate at time t both on the covariates and on the effects of the individual's preceding failure history. Various generalizations of (21) may be considered to acknowledge these conditioning histories in a flexible fashion. For example, λ_0 can be stratified on the number of preceding failures on the study subject; or the argument in λ_0 may be replaced by the gap time from the preceding failure (Prentice et al 1981).

A useful alternative to (21) involves modeling a marginalized failure rate in which one 'forgets' some aspects of $H(t)$. For example, one could replace $H(t)$ in (21) by the covariate history $Z(t)$ and the number $s = s(t)$ of failures for the individual prior to time t. A Cox-type model $\lambda_{0s}(t) \exp\{x(t)\ \beta_s\}$ for the resulting failure rate function would allow one to focus on covariate effects without having to model the timing of the s preceding failures. While the parameters β_s can again be estimated using the score equations from a stratified version of (9) a corresponding sandwich variance formula is needed to acknowledge the variability introduced by the nonincreasing conditioning event (Wei et al. 1989, Pepe and Cai 1993, Lawless and Nadeau 1995).

5.3 Correlated Failure Times

In some settings, such as group randomized intervention trials, the failure times on individuals within groups may be correlated, a feature that needs to be properly accommodated in data analysis. In other contexts, such as family studies of disease occurrence in genetic epidemiology, study of the nature of the dependency between failure times may be a key research goal.

In the former situation, hazard-function models of the types described above can be specified for the marginal distributions of each of the correlated failure times, and marginal regression parameters may be estimated using the techniques of the preceding sections. A sandwich-type variance estimator is needed to accommodate the dependencies between the estimating function contributions for individuals in the same group (Wei et al 1989, Prentice and Hsu 1997).

If the nature of the dependence between failure times is of substantive interest it is natural to consider nonparametric estimation of the bivariate survivor function $F(t_1, t_2) = p(T_1 > t_1, T_2 > t_2)$. Strongly consistent and asymptotically Gaussian nonparametric estimators are available (Dabrowska 1988, Prentice and Cai 1992), though these estimators are nonparametric efficient only if T_1 and T_2 are independent (Gill et al. 1995). Nonparametric efficient estimators away from this independence special case can be developed through a grouped data nonparametric maximum likelihood approach (Van der Laan 1996), though this estimation problem remains an active area of investigation.

5.4 Frailty Models

There is also substantial literature on the use of latent variables for the modeling of multivariate failure times. In these models, failure times are assumed to be independent given the value of a random effect, or frailty, which is usually assumed to affect the hazard function in a multiplicative (time-independent) fashion.

One of the simplest such models arises if a frailty variate has a gamma distribution. This generates the Clayton (1978) bivariate model with joint survivor function

$$F(t_1, t_2) = \{F_1(t_1)^{-\theta} + F_2(t_2)^{-\theta} - 1\}^{-1/\theta} \qquad (22)$$

where θ controls the extent of dependence between T_1 and T_2 with maximal positive dependence as $\theta \to \infty$, maximal negative dependence as $\theta \to -1$, and independence at $\theta = 0$ (e.g., Oakes 1989).

Nielsen et al. (1992) generalize this model to include regression variables and propose an expectation-maximization algorithm with profiling on θ for model fitting. Andersen et al. (1993) and Hougaard (2000) provide fuller accounts of this approach to multivariate failure-time analysis.

6. Additional Survival Analysis Topics

Space permits only brief mention of a few additional survival analysis topics: Considerable effort has been devoted to model building and testing aspects of the use of failure-time regression models. The monograph by Therneau and Grambsch (2000) summarizes much of this work in the context of the Cox model (7).

The cost of assembling covariate histories has led to a number of proposals for sampling within cohorts, whereby the denominator terms in (9) are replaced by suitable estimates, based on a subset of the subjects at risk. These include nested case-control sampling in which the denominator at t_j is replaced by a sum that involves the individual failing at t_j and controls randomly selected from $R(t_j)$, independently at each distinct t_j, and case-cohort sampling in which the denominator summation at t_j consists of the failing individual along with the at risk members of a fixed random subcohort (see Samuelson (1997) for a recent contribution).

Considerable effort has also been directed to allowing for measurement error in regression variables, particularly in the context of the Cox model (7) with time-independent covariates. Huang and Wang (2000) provide a method for consistently estimating β in (7) under a classical measurement error model, and provide a summary of the literature on this important topic.

Testing and estimation on regression parameters has also been considered in the context of clinical trials under sequential monitoring. The recent book by Jennison and Turnbull (2000) summarizes this work.

Additional perspective on the survival analysis literature can be obtained from the overviews of Anderson and Keiding (1999) and of Oakes (2000). Goldstein and Harrell (1999) provide a description of the capabilities of various statistical software packages for carrying out survival analyses.

Bibliography

Andersen P K, Gill R D 1982 Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10**: 1100–20

Andersen P K, Borgan O, Gill R D, Keiding N 1993 *Statistical Models Based on Counting Processes*. Springer, New York

Andersen P K, Keiding N 1999 Survival analysis, overview. *Encyclopedia of Biostatistics*. Armitage P, Colton T (eds.). Wiley, New York, Vol. 6, pp. 4452–61

Breslow N E 1974 Covariance analysis of censored survival data. *Biometrics* **30**: 89–99

Clayton D 1978 A model for association in bivariate life tables and its application in studies of familial tendency in chronic disease incidence. *Biometrika* **65**: 141–51

Cox D R 1972 Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B,* **34**: 187–220

Cox D R 1975 Partial likelihood. *Biometrika* **62**: 269–276

Cox D R, Oakes D 1984 *Analysis of Survival Data*. Chapman & Hall, London

Dabrowska D 1988 Kaplan-Meier estimate on the plane. *Annals of Statistics* **16**: 1475–89

Efron B 1977 Efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**: 557–65

Fleming T R, Harrington D D 1991 *Counting Processes and Survival Analysis*. Wiley, New York

Gill R D, Van der Laan M, Wellner J A 1995 Inefficient estimators of the bivariate survival function for three models. *Annals of the Institute Henri Poincaré, Probability and Statistics* **31**: 547–97

Goldstein R, Harrell F 1999 Survival analysis, software. *Encyclopedia of Biostatistics*. Armitage P, Colton T (eds.). Wiley, New York, Vol. 6, pp. 4461–6

Harrington D P, Fleming T R 1982 A class of rank test procedures for censored survival data. *Biometrika* **69**: 133–43

Hougaard P 2000 *Analysis of Multivariate Survival Data*. Springer, New York

Huang Y, Wang C Y 2000 Cox regression with accurate covariates unascertainable: A nonparametric correction approach. *Journal of the American Statistical Association* **95**: 1209–19

Jennison C, Turnbull B W 2000 *Group Sequential Methods with Application to Clinical Trials*. Chapman & Hall, New York

Kalbfleisch J D, Prentice R L 1980 *The Statistical Analysis of Failure Time Data*. Wiley, New York

Kaplan E L, Meier P 1958 Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**: 457–81

Lawless J F, Nadeau C 1995 Some simple and robust methods for the analysis of recurrent events. *Technometrics* **37**: 158–68

Lin D Y, Geyer C J 1992 Computational methods for semiparametric linear regression with censored data. *Journal of Computational and Graphical Statistics* **1**: 77–90

Lin D Y, Wei J L, Ying Z 1998 Accelerated failure time models for counting processes. *Biometrika* **85**: 608–18
Mantel N 1966 Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer and Chemotherapy Reports* **50**: 163–70
Nielsen G G, Gill R D, Andersen P K, Sorensen T I A 1992 A counting process approach to maximum likelihood estimation in frailty models. *Scand Journal of Statistics* **19**: 25–43
Oakes D 1989 Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**: 487–93
Oakes D 2000 Survival analysis. *Journal of the American Statistical Association* **95**: 282–5
Parzen M I, Wei L J, Ying Z 1994 A resampling method based on pivotal estimating functions. *Biometrika* **68**: 373–9
Pepe M S, Cai J 1993 Some graphical displays and marginal regression analysis for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**: 811–20
Peto R, Peto J 1972 Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society Series A*, **135**: 185–206
Prentice R L 1978 Linear rank tests with right censored data. *Biometrika* **65**: 167–79
Prentice R L, Cai J 1992 Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79**: 495–512
Prentice R L, Hsu L 1997 Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika* **84**: 349–63
Prentice R L, Williams B J, Peterson A V 1981 On the regression analysis of multivariate failure time data. *Biometrika* **68**: 373–9
Robins J M, Tsiatis A A 1992 Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* **79**: 311–9
Samuelsen S O 1997 A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84**: 379–84
Therneau T, Grambsch P 2000 *Modeling Survival Data: Extending the Cox Model*. Springer, New York
Van der Laan M J 1996 Efficient estimation in the bivariate censoring model and repairing NPMLE. *Annals of Statistics* **24**: 596–627
Wei I J, Lin D Y, Weissfeld L 1989 Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of American Statistical Association* **84**: 1065–73
Ying Z 1993 A large sample study of rank estimation for censored regression data. *Annals of Statistics* **21**: 76–99

R. L. Prentice and J. D. Kalbfleisch

Sustainability Transition: Human–Environment Relationship

The reconciliation of society's human developmental goals with the planet's environmental limits over the long term is the foundation of an idea known as sustainable development. This idea emerged in the early 1980s from scientific perspectives on the interdependence of society and environment, but even as it garnered increasing political attention and acceptance around the world, its scientific base weakened. The decade of the 1990s has seen an effort to re-engage the scientific community around the requirements for a sustainability transition. Beyond its commonplace meaning as a transition towards a state of sustainable development, a sustainability transition was studied as a series of interlinked transitions, as a process of adaptive management and social learning, and as a set of indicators and future scenarios.

1. *Sustainable Development*

The origins of sustainable development can be traced back through the 1980 World Conservation Strategy and the 1972 Stockholm Conference on the Human Environment to the early days of the international conservation movement (O'Riordan 1988, Adams 1990). But the contemporary linking of environment and development is little more than a decade old, stemming from *Our Common Future*, the 1987 report of the World Commission on Environment and Development (also known as the Brundtland Commission). The Commission defined sustainable development as the ability of humanity 'to ensure that it meets the needs of the present without compromising the ability of future generations to meet their own needs' (WCED 1987). An extraordinarily diverse set of groups and institutions have taken this widely accepted statement of sustainable development and projected upon it their own hopes and aspirations. While sharing a common concern for the fate of the earth, proponents of sustainable development differ in what they think is to be sustained, what is to be developed, how each is related to the other, and for how long a time.

The most common view as to what is to be sustained are planetary life-support systems, particularly those supporting human life (Kates 1994). These systems include natural resources as well as environments that provide aesthetics, recreation, pollution absorption and cleansing, and other ecosystem services (Daily 1997). Contrasting this anthropocentric view is one of sustaining nature itself in all its biodiversity and assemblages of life forms. These ought to be sustained, not only for their utilitarian service to humans, but because of the moral obligations of humanity arising either as 'stewardship'—still acknowledging the primacy of humans—or as a form of 'natural rights' in which earth and its other living things have equal claims for existence and sustenance. Also for some, cultural diversity is seen as the counterpoint to biological diversity, and various communities that include distinctive cultures, particular groups of people, and specific places need to be sustained.

The most common view of what is to be developed is the economy that provides employment, needed

income, desired consumption, and the means for investment in physical and human capital, as well as environmental maintenance and restoration. An alternative emphasis is on people-centered human development which focuses on the quantity of life, as seen in the survival of children and their increased life expectancy, and the quality of life through education and equal opportunity. Finally, some discussions of what is to be developed emphasizes society itself, addressing the well-being and security of national states, regions and institutions.

The relationship between what is to be sustained and what is to be developed also differ. Some conceptual statements, provide equal emphasis for both; others, while paying homage to sustainable development, seem to be saying 'sustain only,' 'develop mostly,' or 'sustain or develop subject to some minimal constraint on the other.' Similarly the implied time horizons for sustainable development differ. There is universal acceptance that such time horizons are intergenerational, but these then range between a single generation of 25 years, several generations, or as an unstated, but implicit, forever. Each of these time periods present very different prospects and challenges for sustainable development. Over the space of a single generation, almost any development appears sustainable; over an infinite forever none does.

At the century's end, the difficulties of actually delivering on these diverse hopes that people around the world have attached to the idea of sustainable development had become increasingly evident. In part, these difficulties reflected political problems, grounded in questions of financial resources, equity and the competition of other issues for the attention of decision-makers. In part, they reflect the differing views about what should be developed, what should be sustained and over what period. Additionally, however, the political impetus that carried the idea of sustainable development so far and so quickly in public forums has also increasingly distanced it from its scientific and technological base.

2. *Interlinked Transitions*

Thus during the 1990s several groups of scientists sought to reinvigorate their historic connection with the concept of sustainable development by focusing not on the larger concept, with all its actual and implied diverse meanings, but on a sustainability transition. The interest in transitions arose in part from the widespread perception that the continuation of current trends in both environment and development would not provide for the desired state of sustainable development. To the extent that transitions represent breaks in such trends, there was growing interest in identifying the needed or desired transitions in the relationships between society and environment.

There was also widespread interest in the post-Cold War world in the nature of transitions themselves. Attention was focused on the economic transition from state to market control and on the civil society transition from single-party, military or state-run institutions to multiparty politics and a rich mix of nongovernmental institutions (Mathews 1997). Earlier transitions had also been identified, such as that in settlement patterns from rural to urban; in agricultural productivity from increases in production derived from additional land to increases derived from greater yields; and in health from early death by infectious diseases to late death by cancer, heart attack, and stroke. For the environment, significant transitions were seen for biogeochemical cycles in a shift from dominance by natural processes to dominance by human releases; for specific pollutants from increasing to decreasing rates of emissions; and for land cover in temperate zones a change from deforestation to reforestation (Turner et al. 1990).

But the greatest interest lay in the successful confirmation of the 'demographic transition'—the change in population regimes from ones of high birth and death rates, to ones of low birth and death rates. The demographic transition appeared credible because it met scientific criteria: it was partly supported by theory, matched well the data and had predictive power (Kates 1996). It was also seen as both needed and desired because of the widespread consensus that rapidly growing population made more difficult human development, while further stressing the natural systems that needed to be sustained.

On a global scale, at the century's end, the transition in birth rates was more than halfway towards birth stability, the 2.1 births required to achieve eventual zero population growth, with an average of less than three children for each woman of reproductive age, compared with five at the post World War II peak of population growth. The death transition was even more advanced, with life expectancy at birth having increased from 40 years to 66 years, about three-quarters of the way to the stabilizing stage of the transition, when life expectancy is likely to reach 75 years. Mid-range projections (in 1998) foresaw further slowing of population growth rates with 8.9 billion people in the world by 2050 and stabilizing to 9.5 billion by the century's end (United Nations 1999).

Thus, in the first major conceptualization of the 'sustainability transition' Speth (1992) set forth a series of five inter-linked transitions (demographic, technological, economic, social and institutional) as collective requirements for a sustainability transition. In addition to the desired demographic transition, Speth saw the need for a transition towards technologies that were environmentally benign and reduced sharply the consumption of natural resources, and the generation of waste and pollutants. For the economy, the desired transition was towards one in which prices reflected their full environmental costs. A social transition

would move towards a fairer sharing of economic and environmental benefits both within and between countries. All of these transitions required a fifth: a transition in the institutional arrangements between governments, businesses and people that would be less regulation-driven and more incentive-led, less confrontational and more collaborative. To these five, Gell-Mann (1994) further suggested an ideological transition towards a sense of solidarity that can encompass both the whole of humanity and the other organisms of the biosphere and an informational transition that integrates disciplinary knowledge and disseminates it across the society.

Together, Speth and Gell-Mann present these transitions as requirements for a more sustainable world: if each individual transition is completed successfully the result would constitute a sustainability transition. And all of these transitions may be underway as there are some trends in each of the desired directions. To investigate this concept they organized a major, albeit uncompleted, research effort known as the '2050' project (World Resources Institute, Brookings Institutions, Santa Fe Institute) that sought to track these transitions and understand their interactions.

3. *Adaptive Management and Social Learning*

A different approach was adopted by the National Academy of Sciences (USA) Board on Sustainable Development (1999). It too adopted a two generation, to the year 2050, time horizon. But in *Our Common Journey: A Transition Toward Sustainability*, the Board doubted that any specific set of trends or transitions constituted necessary or sufficient conditions for sustainability. It argued that the pathway to a sustainability transition could not be mapped in advance. Instead, it would have to be navigated adaptively through trial and error, and conscious policy experimentation (Holling 1978, Lee 1993) and a process of social learning (Social Learning Group 2001) as well other devices such as indicators and scenarios (described below).

What could be selected were goals for a sustainability transition. In the Board's judgment, the primary goals of a transition toward sustainability over the next two generations should be to meet the needs of a much larger but stabilizing human population, to sustain the life-support systems of the planet, and to reduce hunger and poverty substantially. Specific targets for these goals for human well-being and environmental preservation have been defined over the past few decades through extensive processes of international political debate and intergovernmental action.

In the area of human needs, internationally agreed-on targets exist for providing food and nutrition, nurturing children, finding shelter and providing an education, although not for finding employment. Thus, there is an implicit hierarchy of needs that favors children and people in disasters, and favors feeding and nurturing first, followed by education, housing and employment.

Compared with targets for meeting human needs, quantitative targets for preserving life-support systems are fewer, more modest and more contested. Global targets now exist for ozone-depleting substances and greenhouse gases, and regional targets exist for some air pollutants. Absolute prohibitions (zero targets) exist for ocean dumping of radioactive wastes and some toxics, for the taking and/or sale of a few large mammals (whales, elephants and seals), migratory birds when breeding or endangered, and certain regional fishing stocks. Water, land resources and ecosystems such as arid lands and forests have, at best, qualitative targets to achieve sustainable management or restoration. International standards exist for many toxic materials, organic pollutants and heavy metals that threaten human health, but not for ecosystem health.

A major action and research agenda was developed in *Our Common Journey* based on both current scientific understanding and the requirements of an emerging sustainability science. For the core sectoral areas of sustainable development identified more than a decade ago by the Brundtland Commission—human population and well-being, cities, agriculture, energy and materials, and living resources—the Board identified concrete goals and appropriate next steps to accelerate major transitions underway or needed in each sector. To create a research agenda of what might be called 'sustainability science,' the Board would promote the creation of usable knowledge from scientific understanding; would seek to integrate global and local perspectives to shape a 'place-based' understanding of the interactions between environment and society; and initiate focused research programs on a small set of understudied questions that are central to sustainability transition.

4. *Indicators and Scenarios*

The largest scientific effort to date has been the creation of indicators of sustainable development and to use them to track a transition toward sustainability. It sought to make more realistic the diverse meanings of sustainable development by attempting to develop indicators of the different values so implied. A major effort was launched by the Scientific Committee on Problems of the Environment (Moldan et al. 1997) and an important alternative effort by the so-called Balaton Group (Meadows 1998). These scientific efforts complemented similar efforts by various local, national, and international groups to create indicators as ways of expressing desired values for sustainable

Table 1
Sustainable development in the USA: an experimental set of indicators

Economic	Environmental	Social
Capital assets	Surface-water quality	Population of USA
Labor productivity	Ratio of renewable water supply to withdrawals	Life expectancy at birth
Domestic product	Fisheries utilization	Births to single mothers
Income distribution	Acres of major terrestrial ecosystems	Children living in families with only one parent present
Consumption expenditures per capita	Invasive alien species	Educational achievement rates
Unemployment	Conversion of cropland to other uses	Educational attainment by level
Inflation	Soil-erosion rates	Teacher training level and application of qualifications
Investment in R&D as a percentage of GDP	Timber growth to removals balance	People in census tracts with 40 percent or greater poverty
Federal debt to GDP ratio	Outdoor recreational activities	Crime rate
Energy consumption per capita and per $ of GDP	Contaminants in biota	Participation in the arts and recreation
Materials consumption per capita and per $ of GDP	Quantity of spent nuclear fuel	Contributing time and money to charities
Homeownership rates	Identification and management of superfund sites	
Percentage of households in problem housing	Metropolitan air quality nonattainment	
	Status of stratospheric ozone	
	Greenhouse-gas emissions	
	Greenhouse climate response index	

futures (http:}}iisd1.iisd.ca/measure}compendium.htm and United Nations Commission for Sustainable Development 1996). A set of 40 economic, environmental and social indicators for the USA that use existing data sets are shown in Table 1 (http://www.sdi.gov/reports.htm).

In contrast to indicators such as these that can inform society to what extent progress is being made in a transition toward sustainability, alternative pathways for a transition were explored through long-range development scenarios that reflect the uncertainty about how the global system might unfold. While they are rigorous, reflecting the insights of science and modeling, these scenarios are told in the language of words as well as numbers because assumptions about culture, values, lifestyles, and social institutions require qualitative description. Sets of scenarios were created that reflect current trends and reform proposals, various forms of social and environmental breakdown, and more fundamental transitions or transformations (Raspin et al. 1997, Hammond 1998). Based on these and other data, the National Academy of Science's Board (USA) on Sustainable Development (1999) concluded that:

> Although the future is unknowable … a successful transition toward sustainability is possible over the next two generations. This transition could be achieved without miraculous technologies or drastic transformations of human societies. What will be required, however, are significant advances in basic knowledge, in the social capacity and technological capabilities to utilize it, and in the political will to turn this knowledge and know-how into action.

Bibliography

Adams W M 1990 *Green Development: Environment and Sustainability in the Third World.* Routledge, London

Daily G C 1997 *Nature's Services: Societal Dependence on Natural Ecosystems.* Island Press, Washington DC

Gell-Mann M 1994 *The Quark and The Jaguar: Adventures in the Simple and Complex*. W. H. Freeman, New York
Hammond A 1998 *Which World? Scenarios for the 21st Century*. Island Press, Washington DC
Holling C S (ed.) 1978 *Adaptive Environmental Assessment and Management*. Wiley, Chichester, UK
Kates R W 1994 Sustaining life on earth. *Scientific American* **271**: 114–22
Kates R W 1996 Population, technology, and the human environment: A thread through time. *Daedalus* **125**(3): 43–71
Lee K N 1993 *Compass and Gyroscope: Integrating Science and Politics for the Environment*. Island Press, Washington DC
Mathews J T 1997 Power shift. *Foreign Affairs* **76**: 50–66
Meadows D 1998 *Indicators and Information Systems for Sustainable Development. A report to the Balaton Group*. The Sustainability Institute, Hartland Four Corners, VT
Moldan B, Billharz S, Matravers R (eds.) 1997 *Sustainability Indicators: A Report on the Project on Indicators of Sustainable Development*. Wiley, New York and Chichester, UK
Board on Sustainable Development, National Research Council 1999 *Our Common Journey: A Transition Toward Sustainability*. National Academy Press, Washington, DC
O'Riordan T 1988 The politics of sustainability. In: Turner R K (ed.) *Sustainable Environmental Management: Principles and Practice Boulder*. Westview Press, Boulder, CO
Raskin P, Gallopin G, Gutman P, Hammond A, Swart R 1998 *Bending the Curve: Toward Global Sustainability. A Report of the Global Scenario Group*. Stockholm Environment Institute, Boston
Speth J 1992 The transition to a sustainable society. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 870–72
The Social Learning Group 2001 *Learning to Manage Global Environmental Risks: Vol 1. A Comparative History of Social Responses to Climate Change, Ozone Depletion, and Acid Rain. Vol 2. A Functional Analysis of Social Responses to Climate Change, Depletion, and Acid Rain*. MIT Press, Cambridge, MA
Turner B II, Clark W, Kates R, Richards J, Mathews J, Meyer W 1990 *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere Over the Past 300 Years*. Cambridge University Press, Cambridge, UK
United Nations 1999 *World Population Prospects: The 1998 Revision*. United Nations Population Division, New York
United Nations Commission for Sustainable Development 1996 *Indicators of Sustainable Development: Framework and Methodologies*. United Nations, New York
World Commission on Environment and Development 1987 *Our Common Future*. Oxford University Press, Oxford, UK

R. W. Kates

Sustainable Development

In the late 1980s, the World Commission on Environment and Development (WCED), otherwise known as the Brundtland Commission, defined sustainable development as 'development that meets the needs of the present without compromising the ability of future generations to meet their own needs' (WCED 1987). Since then, sustainable development has received significant attention from the global community at the local, national, and international level amid concerns about climate change, biodiversity loss, tropical deforestation, and other environmental depletion and degradation problems.

This article discusses briefly the emergence of the concept of sustainable development and provides a broad definition of sustainable development. Two distinct interpretations of sustainable development—*weak* and *strong* sustainability—are compared and contrasted. Recent advances in three areas of the literature that are relevant to understanding sustainable development and to designing appropriate policy responses are examined. These are: ecological functioning and resilience; environmental Kuznets' curve (EKC) and the environment–growth debate; and endogenous growth, technological innovation, and resource dependency. For each area, the basic concept is explained and the implications for sustainable development are discussed. Finally, the policy issues and options for achieving sustainable development are discussed. Four key options are identified: measuring the ecological and economic impact of declining natural resources; improving our estimates of the economic costs of depleting natural capital; establishing appropriate incentives, institutions, and investments for sustainable management of natural capital; and encouraging interdisciplinary collaboration between the relevant social and behavioral sciences undertaking research on sustainable development.

1. Concept of Sustainable Development

1.1 Definition of Sustainable Development

The broad concept of sustainable development (Pearce and Barbier 2000) encompasses considerations of equity across and within generations, taking a longer-term perspective and accounting for the value of the environment in decision-making (Pearce et al. 1989). As noted by Turner (1997), the concept of sustainable development has been challenged on several grounds: as an oxymoron, as a means of imposing a particular political position, as an approach to divert attention from more pressing socioeconomic problems, and given the history of the human–environment relationship (see *Human–Environment Relationships*). However, sustainable development continues to receive increasing international recognition and it has become a key guiding principle for the global society at the start of the new millennium (National Research Council 1999).

There have been a substantial number of diverse and wide-ranging interpretations of sustainable development, both across and within scientific disciplines. Pearce et al. (1989) provide a collection of

such definitions, and for summaries of recent developments see Goldin and Winters (1995) and Pearce and Barbier (2000). In general, sustainable development implies that economic development today must ensure that future generations are left no worse off than present generations. This means that future generations should be entitled to at least the same level of economic opportunities—and thus at least the same level of economic welfare—as is currently available to present generations. As Pezzey (1989) succinctly states, the welfare of society should not be declining over time.

1.2 Total Capital Stock

The total stock of capital employed by the economic system, which includes physical, human, and natural capital, determines the full range of economic opportunities and welfare available to both present and future generations. Natural capital comprises of natural resources and environments and provides various goods and services to society. For example, tropical forests provide *direct use benefits* in the form of timber, fruits, and nuts, *indirect use benefits* such as recreation, watershed protection, biodiversity protection, carbon storage, and climate maintenance, and *nonuse benefits* which include existence and bequest values (Barbier et al. 1994). Sustainable development requires managing and enhancing the total capital stock so that its aggregate value does not decline over time. Therefore, society must decide how best to use its total capital stock today to increase current economic activities and welfare, and how much it needs to save or even accumulate for the welfare of future generations.

However, it is not simply the aggregate stock of capital in the economy that matters but also its composition. In particular, whether present generations are 'using up' one form of capital to meet the needs of today. For example, much of the recent interest in sustainable development has risen out of concern that current economic development may be leading to the rapid accumulation of physical and human capital, but at the expense of excessive depletion and degradation of natural capital. The major concern has been that the irreversible depletion of the world's stock of natural wealth will have detrimental implications for the welfare of future generations (see *Consumption, Environmentally Significant*).

While generally it is accepted that economic development around the world is leading to the irreversible depletion of natural capital, there is widespread disagreement as to whether this necessarily implies that such development is inherently unsustainable. From an economic standpoint, the critical issue of debate is not whether natural capital is being irreversibly depleted. Instead, as argued by Mäler (1995), the more crucial issue for sustainable development is whether we can compensate future generations for the current loss of natural capital, and if that is the case, how much is required to compensate future generations for this loss? This depends on whether natural capital is considered to have an unique or essential role in sustaining human welfare and whether special 'compensation rules' are required to ensure that future generations are not made worse off by natural capital depletion today.

1.3 Weak and Strong Sustainability

Two distinct interpretations of sustainable development are referred to commonly in the literature as 'weak' and 'strong' sustainability. According to the *weak sustainability* view, there is essentially no inherent difference between natural and other forms of capital, and hence the same optimal depletion rules ought to apply to both. That is, as long as the natural capital that is being depleted is replaced with even more valuable physical and human capital, then the value of the aggregate stock—comprising human, physical, and the remaining natural capital—is increasing over time. Maintaining and enhancing the total stock of all capital alone is sufficient to attain sustainable development.

In contrast, proponents of the *strong sustainability* view argue that physical or human capital cannot substitute for all the environmental resources comprising the natural capital stock, or all of the ecological services performed by nature. This view questions whether human, physical, and natural capital comprises a single homogeneous total capital stock. *Uncertainty* over many environmental values, in particular the value that future generations may place on increasingly scarce natural resources and ecological services, further limits our ability to determine whether we can compensate adequately future generations for irreversible losses in *essential* natural capital today. Thus, the strong sustainability view suggests that environmental resources and ecological services that are essential for human welfare and cannot be substituted easily by human and physical capital should be protected and not depleted. Maintaining or increasing the value of the total capital stock over time in turn requires keeping the nonsubstitutable and essential components of natural capital at least constant over time.

2. Aspects of Sustainable Development

2.1 Ecological Functioning and Resilience

The ecological economics literature has called attention to the important role of *ecological functioning* and *resilience* in sustaining human livelihoods and

welfare (Perrings et al. 1992) (see *Ecological Economics*). By ecological functioning, ecologists usually mean those basic processes of ecosystems, such as nutrient cycling, biological productivity, hydrology, and sedimentation, as well as the overall ability of ecosystems to support life. Ecologists consider the collective range of *life-support functions* to be the key characteristic that defines an ecosystem, as well as the source of the many key ecological resources and services that are important to human welfare (e.g., drinking water, food, and soils).

Although there is some dispute in the ecological literature over precise interpretations of the term *ecological resilience*, ecologists generally use this term to mean the capacity of an ecosystem to recover from external shocks and stresses (e.g., drought, fire, earthquakes, pollution, biomass removal) (Holling 1986, Walker 1992). As the health of an ecosystem usually is determined by its capacity to deliver the life-support functions appropriate to its stage of ecological succession, then the ecological resilience of the system is linked inherently to its ecological functioning. That is, if an ecosystem is resilient, then it should recover sufficiently from any human-induced or natural stresses and shock and function normally. However, if ecosystems are insufficiently resilient to recover from persistent problems of environmental degradation then the ability of ecosystems to function normally and deliver important biological resources and ecological services will be affected. At some threshold level of ecological functioning, ecosystems that are subject to persistent environmental degradation and disruption begin to malfunction, breakdown, and ultimately, collapse (see *Human–Environment Relationship: Carrying Capacity*).

The loss of ecosystem functioning and resilience will lead to a decline in the availability of some economically important biological resources and ecological services (see *Extinction of Species*). For example, the loss of fish stocks as the result of pollution impacting on marine systems, the inability of rangeland pastoral systems to recover from drought, and the decline of the watershed protection function of degraded forests. At first these ecological and welfare impacts may be localized, but if they occur on a sufficiently great enough scale, they may generate wider national, regional, or even global effects, such as climate change (see *Climate Change, Economics of*). Although the scale of such impacts and their ultimate effects on human welfare are often uncertain and difficult to determine, the irreversible loss of key biological resources and ecological 'services' through the disruption and breakdown of ecosystems could constitute a growing problem of *ecological scarcity* (Barbier 1989). The rising 'scarcity' of ecological resources and services will mean that they will increase in value relative to human-made goods and services produced in economic system. Moreover, persistent and rising ecological scarcity may be an indicator of more widespread and frequent disruptions to the functioning and resilience of ecosystem, and increase the likelihood of ecological collapse and catastrophe on ever widening scales (Daily and Ehrlich 1992, Meadows et al. 1992, UNPF 1993).

2.2 Implications of Ecological Functioning and Resilience for Sustainable Development

Given the rate at which ecological disruption is occurring globally as a result of the cumulative impacts of pollution, habitat modification, and conversion and over-exploitation of some biological resources, the ecological functioning and resilience upon which many important ecological resources and services depend may be undermined. Thus, not only is there a danger of depleting essential ecological components of natural capital, but the widespread disruptions to the functioning and resilience of ecosystems may mean that these essential components are being lost irreversibly. Therefore, the ecological economics literature generally endorses the strong sustainability view.

Ensuring sustainable development first will require identifying those essential ecological resources and services that are most at risk from current patterns of economic activity and development. Second, it is necessary to take appropriate policy measures to protect the irreplaceable components of natural capital. Finally, it is important to maintain the functioning and resilience of the key ecosystems on which they depend. This implies a fundamental rethinking in the way in which countries currently manage their environments, as well as the relative importance of policies for environmental protection in the process of economic development. The additional costs that are required to adjust current patterns of economic activity and development could be significant, especially in the short and medium term, and may constrain such a comprehensive shift in policy.

2.3 Environmental Kuznets' Curves and the Environment–Growth Debate

There has been a long-standing debate about the relationship between economic growth and the environment. An early view was expressed in the publication of the Club of Rome's *Limits to Growth* in 1972 that there exists a *trade-off* between economic growth (measured by rising real per capita incomes) and the environment, whereby an improvement in one implies a reduction in the other (Meadows et al. 1972, Daly 1987). The modern sustainable development debate has focused more on the potential *complementarity* between growth and the environment, where growth is considered compatible with improved environmental quality (see *Cassandra/Cornucopian Debate*).

A recent extension to this literature looks at the analysis of *environmental Kuznets' curves* (EKC). This is the hypothesis that there exists an 'inverted U' shaped relationship between a variety of indicators of environmental pollution or resource depletion and the level of per capita income (Kuznets 1955, Grossman and Kreuger 1995, Shafik 1994). The implication of this hypothesis is that environmental degradation should be observed to increase initially, but to eventually decline, as per capita income increases. To some extent, this combines the two previous views on the relationship between economic growth and the environment. That is, at low incomes there exists a trade-off between economic growth and environmental quality, while at higher incomes economic growth is complementary to improved environmental quality.

Pearce and Barbier (2000) and Stern et al. (1996) provide a review of the relevant empirical studies. These studies suggest that EKC relationships are more likely to hold for only a few pollutants with short-term and local impacts, such as sulphur dioxide (SO_2) and to a lesser extent solid particulate matter (SPM) (see *Air Pollution*). In addition, even when an EKC relationship is estimated, the 'turning point' on the curve where environmental degradation starts to decline with per capita income is very high relative to the current per capita GDP levels of most countries. This is a particular problem for less developed countries with low per capita income levels (see *Environment and Development*). However, several recent studies have also demonstrated that EKCs are highly susceptible to structural economic shifts and technological changes, which are in turn influenced by policy initiatives such as the Montreal Protocol and the Second Sulphur Protocol.

2.4 Implications of EKCs for Sustainable Development

Recent EKC studies have revived the wider 'growth vs. the environment' debate. Some commentators have argued that the empirical evidence of EKCs supports the general proposition that the solution to combating environmental damage is economic growth itself (Beckerman 1992). For example, the EKC literature offers limited evidence that for certain environmental problems, particularly air pollutants with localized or short-term effects, there is an eventual reduction in emissions associated with higher per capital income levels, which may be attributable to the 'abatement effect' that arises as countries become richer (Panayotou 1997). Also, both the willingness and the ability of political jurisdictions to engage in and enforce improved environmental regulations, to increase public spending on environmental research and development, or even to engage in multilateral agreements to reduce emissions may also increase with per capita income levels (see *Greening of Technology and Ecotechnology*).

Other commentators have been more cautious, noting that conclusive evidence of an EKC relationship applies only to a few pollutants, which makes it difficult to support more general growth-environment linkages. It has also been noted that, even for those pollutants displaying EKC characteristics, aggregate global emissions are projected to rise over time, demonstrating that the existence of an EKC relationship for such pollutants does not imply necessarily that, at the global level, any associated environmental damage is likely to disappear with economic growth. Therefore, the recent EKC literature should not be taken to imply that economic growth on its own will foster environmental improvement automatically. As Panayotou (1997) concluded, 'when all effects are considered, the relationship between growth and the environment turns out to be much more complex with wide scope for active policy intervention to bring about more desirable, and in the presence of market failures, more efficient economic and environmental outcomes.' Or, as Arrow et al. (1995) have succinctly put it: 'economic growth is not a panacea for environmental quality; indeed it is not even the main issue.'

2.5 Endogenous Growth, Technological Innovation, and Resource Dependency

Another recent advance in the literature with important implications for sustainable development looks at *endogenous growth*, *technical innovation*, *and resource dependency*. A key feature of recent endogenous growth models is that technological innovation, that is, the development of new technological ideas or designs, is *endogenously* determined by private and public sector choices within the economic system (hence the term 'endogenous' growth) rather than determined outside (i.e., 'exogenous') to the system (Romer 1990). This implies that the level of technology in an 'endogenous growth' economy can be advanced perpetually through public and private investments, such as research and development expenditures and increases in human capital skills. Since *technical innovation* adds to physical capital already used in production, the economy can potentially 'sustain' growth rates indefinitely. Furthermore, as the economy develops, its growth rate should eventually diminish until it reaches its long-run growth rate (Barro and Sala-I-Martin 1995).

The recent debate over the role of innovation in economic growth has led to empirical studies across countries and regions in the factors underlying long-term economic growth. Empirical evidence seems to support the notion that rich countries have higher rates of savings or more 'effective' institutions, and

implies that persistently bad government policies and ineffective institutions are inhibiting the long-run growth prospects of low-income countries. In addition, poor countries may fail to achieve higher rates of growth because they fail to generate or use new technological ideas to reap greater economic opportunities.

Institutional and policy failures in poor economies are important determinants of their inability to innovate sufficiently to achieve higher long-term growth rates. Another important factor may be the structural economic dependence of these economies on their natural resource endowments (see *Environmental Vulnerability*). Recent explanations as to why *resource dependence* may be a factor in influencing economic growth point to a number of possible fundamental linkages between environment, innovation, and long-term growth relevant to poor economies. Previous studies of an economy's dependence on natural resources and its long-run growth (Dasgupta and Heal 1979, Stiglitz 1974) have been extended by Barbier (1997) to include an endogenous growth economy in order to determine whether an economy's dependence on an exhaustible natural resource is necessarily a binding constraint on long-term growth. The results of the analysis show that sufficient allocation of human capital to innovation will ensure that in the long run resource exhaustion can be postponed indefinitely, and the possibility exists for sustained long-term per capita consumption. However, when negative feedback effects between the rate of resource utilization and innovation exist, the conditions under which it is possible to sustain per capita consumption over the long run are much more stringent and again assume that the economy is able to build up sufficient human capital and to innovate.

2.6 Implications of Endogenous Growth Theory, Technical Innovation, and Resource Dependency for Sustainable Development

Endogenous growth theory offers some support for the weak sustainability view. That is, provided that obstacles such as persistent policy distortions, political instability, and institutional failures can be overcome, any economy—even a low-income and resource dependent developing country—should be able to foster endogenous innovation to substitute human and physical capital for a declining natural capital base in order to sustain economic opportunities and welfare indefinitely. The key appears to be developing effective policies and institutions, including the necessary public and private investments to enhance research and development and human capital skills.

Unfortunately, in most low-income countries current economic policies and investments in agriculture, forestry, and other resource-based sectors have led to rapid changes—frequently with adverse economic consequences—in resource stocks and patterns of use. Demographic trends have often worsened the relationship between population and resource carrying capacity in many regions, and chronic environmental degradation may itself be contributing to social conflict and political instability, thus undermining the social and economic conditions necessary to foster long-term innovation and development (Homer-Dixon 1995). In addition, the type of 'resource-saving' innovations envisioned in most endogenous growth models are likely to be technologies to abate pollution and other waste products, and to conserve the use of raw material and energy inputs. However, many ecological resources and services, such as biodiversity, amenity services, and ecological functions, are less amenable to substitution by the conventional resource-saving innovations developed in the economic process. Thus, innovations are less likely to lead to opportunities for the substitution of the more unique, and possibly essential, range of ecological resources and services, including the general life-support functioning and resilience of ecosystems.

3. Policy Issues and Options for Sustainable Development

This section discusses the policy issues and options for sustainable development. There are some important areas of common ground for policy makers that make good progress towards sustainable development, regardless of the viewpoint on strong and weak sustainability. For example, as there exists considerable scope for substituting human and physical capital for natural capital through conservation and substitution of resource inputs, reducing pollution, and improving environmental protection, then designing policies for sustainable development must begin with these basic objectives in mind. To do this requires progress in four areas.

First, it is important to improve existing efforts to measure the economic and ecological consequences of natural capital depletion and degradation. That is, in order to consider how well an economy is substituting physical and human capital for natural capital, it is necessary to measure how much the value of the natural capital stock has been 'depreciated' through current efforts to use environmental resources and services to promote greater levels of consumption and investment in the economy. Determining the 'sustainable' level of income and consumption, that is, accounting for the increases in the total available goods and services in the economy *net* of the depreciation in the value of the natural capital stock, is an important economic approach to measuring sustainability. Equally, determining the ecological 'limits' to exploitation of the natural capital stock, including ecological thresholds, carrying capacities, and key ecological goods, services and natural environments

that might be considered 'essential,' are important ecological contributions to measuring sustainability.

Second, many of the resources and services of our natural capital endowment are either 'unpriced,' meaning that unlike other goods and services in our economy there are no markets for them, or they are 'underpriced,' meaning that where markets do exist for some environmental goods and services the prices in these markets do not necessarily reflect their true contribution to human welfare. As long as environmental goods and services continue to be 'unpriced' and 'underpriced,' then their contribution to human welfare relative to other goods and services in the economy is effectively 'undervalued.' Thus, in order to promote better policies for sustainable development it is necessary to determine the correct economic values for environmental goods and services and to develop a variety of economic tools for assessing these values (see *Resources and Environment: Contingent Valuation*).

Third, determining an appropriate 'mix' of human, physical, and natural capital to ensure sustainable development is ultimately about designing appropriate incentives, institutions, and investments for efficient and sustainable management of the natural environment. In order to promote sustainable development it is important to determine the causes of environmental degradation, particularly where the failure of institutions, markets, and government policy are the main contributing factors, and to demonstrate how correcting these failures can lead to improved incentives and investments for efficient and sustainable management of natural capital (see *Environmental Policy: Protection and Regulation*).

Finally, in order to make genuine progress towards sustainable development, it is necessary for the conventional social and behavioral sciences, such as economics and ecology, to work together towards a common goal of sustainable development.

See also: Development: Rural Development Strategies; Development: Sustainable Agriculture

Bibliography

Arrow K, Bolin B, Costanza R, Dasgupta P, Folke C, Holling C S, Jansson B-O, Levin S, Mäler K-G, Perrings C A, Pimentel D 1995 Economic growth, carrying capacity, and the environment. *Science* **268**: 520–21

Barbier E B 1989 *Economics, Natural Resource Scarcity and Development: Conventional and Alternative Views*. Earthscan Publications, London

Barbier E B 1997 Endogenous growth and natural resource scarcity. *Environmental and Resource Economics* **14**(1): 51–74

Barbier E B, Burgess J C, Folke C 1994 *Paradise Lost? The Ecological Economics of Biodiversity*. Earthscan, Publications, London

Barro R J, Sala-I-Martin X 1995 *Economic Growth*. McGraw Hill, New York

Beckerman W 1992 Economic growth and the environment: Whose growth? *World Development* **20**: 481–96

Daily G C, Ehrlich P R 1992 Population, sustainability and Earth's carrying capacity. *BioScience* **42**: 761–71

Daly H E 1987 The economic growth debate: What some economists have learned but many have not. *Journal of Environmental Economics and Management* **14**(4): 323–36

Dasgupta P S, Heal G E 1979 *The Economics of Exhaustible Resources*. Cambridge University Press, Cambridge, UK

Goldin I, Winters L A (eds.) 1995 *The Economics of Sustainable Development*. Cambridge University Press, Cambridge, UK

Grossman G M, Kreuger A B 1995 Economic growth and the environment. *Quarterly Journal of Economics* **110**(2): 353–77

Holling C S 1986 Resilience of ecosystem: Local surprise and global change. In: Clark E C, Munn R E (eds.) *Sustainable Development of the Biosphere*. Cambridge University Press, Cambridge, UK

Homer-Dixon T 1995 The ingenuity gap: Can poor countries adapt to resource scarcity. *Population and Development Review* **21**: 587–8

Kuznets S 1955 Economic growth and income inequality. *American Economic Review* **49**: 1–28

Mäler K-G 1995 Economic growth and the environment. In: Perrings C, Mäler K-G, Folke C, Holling C S, Jansson B-O (eds.) *Biodiversity Loss: Economic and Ecological Issues*. Cambridge University Press, Cambridge, UK.

Meadows D H, Meadows D L, Randers J, Behrens W 1972 *The Limits to Growth*. Universe Books, New York

National Research Council Board on Sustainable Development, Policy Division 1999 *Our Common Journey: A Transition Towards Sustainability*. National Academy Press, Washington, DC

Panayotou T 1997 Demystifying the environmental Kuznets curve: Turning a black box into a policy tool. *Special Issue on Environmental Kuznets Curves, Environment and Development Economics* **2**(4)

Pearce D W, Barbier E 2000 *Blueprint for a Sustainable Economy*. Earthscan Publications, London

Pearce D W, Markandya A, Barbier E B 1989 *Blueprint for a Green Economy*. Earthscan Publications, London

Perrings C, Folke C, Maler K-G 1992 The ecology and economics of biodiversity loss: The research agenda. *AMBIO* **21**(3): 201–11

Pezzey J 1989 Economic Analysis of Sustainable Growth and Sustainable Development. Environment Department Working Paper No. 15, World Bank, Washington, DC

Romer P 1990 Endogenous technological change. *Journal of Political Economy* **98**: S71–S102

Shafik N 1994 Economic development and environmental quality: An econometric analysis. *Oxford Economic Papers* **46**: 757–73 (Suppl. S.)

Stern D I, Common M S, Barbier E B 1996 Economic growth and environmental degradation: The environmental Kuznets curve and sustainable development. *World Development* **24**(7): 1151–60

Stiglitz J E 1974 Growth with exhaustible natural resources: Efficient and optimal growth paths. *Review of Economic Studies. Symposium on the Economics of Exhaustible Resources*, pp. 123–38

Turner B L II 1997 The sustainability principle in global agendas: Implications for understanding land use/cover change. *The Geographic Journal* **163**(2): 133–40

UNPF (United Nations Population Fund) 1993 *Population, Growth and Economic Development*. Report on the Consultative Meeting of Economists convened by the United

Nations Population Fund, 28–29 September 1992, UNPF, New York
Walker B H 1992 Biodiversity and ecological redundancy. *Conservation Biology* **6**: 18–23
WCED (World Commission on Environment and Development) 1987 *Our Common Future*. Oxford University Press, Oxford

J. C. Burgess and E. B. Barbier

Sustainable Transportation

The concept of sustainable transportation applies the idea of societal sustainability to a particular area of human activity: the movement of people and goods. Sustainability, or sustainable development, has been adopted by the United Nations agencies as an overarching goal of economic and social development (UNCED 1992). Although there are several alternative definitions of sustainability, the most widely cited definition and the one generally recognized as initiating the global consideration of the subject, is that of the World Commission on Environment and Development (also known as the Brundtland Commission).

> Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs. (WCED 1987, p. 43)

This and most other definitions constitute a normative concept of intergenerational responsibility or fairness. It is also significant that the context of the Brundtland definition was the issue of economic development and resulting conflicts with the environment. The concept of sustainability balances the right of less advantaged nations to economic development, with the imperative to preserve the resources, especially the environmental resources, that future generations will need for their own well-being. Thus, a concept of equity, both intergenerational and international, has been inherent in the idea of sustainability since its inception.

Sustainability pertains to the responsibility of an entire generation of society to future generations; whether it can meaningfully be applied to a single area of human activity such as transportation has been a subject of debate. That is, sustainability must be satisfied by the integral activities of a society and so, in this sense, it is not possible to judge whether one sector of society is sustainable on its own. On the other hand, it would not be possible for society to achieve sustainability, at least not in the long run, if even one sector were not sustainable. Thus, sustainable transportation may be considered a necessary but not sufficient condition for the sustainability of society.

Economic development has been central to the issue of sustainability since its inception. Economists have argued that the Bruntland definition implies non-decreasing well-being, and some have suggested that measures such as gross domestic product or real consumption per capita may be appropriate measures of sustainable growth. Others argue that well-being must be defined more broadly to include other areas of human development, such as education, health, environmental quality, and social justice (e.g., Pearce et al. 1994). In this broader view, the human institutions passed from one generation to another are just as important a resource as any other institution.

The question of how to translate the Brundtland imperative into a measurable and verifiable concept is still being worked out. In the economic literature, efforts to resolve this question have revolved around the concepts of weak and strong sustainability. The doctrine of weak sustainability holds that manufactured capital (machinery, buildings, infrastructure, etc.) is perfectly substitutable for natural capital (natural and environmental resources). This view holds not only that manufactured capital can be substituted for resources such as coal or iron ore, but that it can be substituted for environmental functions such as the ability of the air to absorb pollution or even the atmospheric and ocean processes that produce the global climate. Strong sustainability, on the other hand, maintains that there are limits to the substitutability of manufactured and natural capital. In particular, it maintains that there are certain environmental processes and functions that must be protected and preserved because they cannot be replaced by created capital. Ecosystems and biodiversity are often-cited examples of irreplaceable environmental processes.

How economic growth and expanding human welfare could continue indefinitely in the face of intensifying environmental impacts and finite natural resources is a central question for theorists and researchers. Human activity, and especially economic activity, are inextricably linked to the environment by the ever-expanding quantities of natural resources extracted, processed and consumed (Pearce and Warford 1993). Because mass and energy must be conserved due to the laws of nature, all the resources and energy used must end up somewhere in the form of waste (unless recycled). A useful way to conceive of the relationship between human activity and the environment is to imagine a coefficient of environmental intensity that relates total environmental damage to the total level of activity, perhaps measured, albeit incompletely by gross domestic product. Similarly, one can imagine a coefficient of natural resource requirements for input to the world economy. Because emissions controls, recycling, and greater energy efficiency can change the environmental coefficient, technological change and behavioral change clearly must play a key role in achieving sustainability. If the environmental coefficient can be reduced as fast or faster than the growth of human activity, and if increasing efficiency and

recycling can similarly reduce the requirements for energy and other natural resource inputs, then it should be possible for human well-being to expand in a sustainable way.

The above line of reasoning identifies sustainability not as an equilibrium or final state of being able to continue indefinitely, but as a steady state in which the rate of improvement exceeds the rate of exhaustion of manufactured capital, natural resources, environmental services, and human and institutional capital. In this view, sustainability requires reducing demands on the environment per unit of activity as fast as or faster than the growth in activity. It requires increasing the efficiency of energy and resource use, or expanding the scope of useful resources at a rate at least as fast as the growth of human activity. It may even be interpreted more broadly as a matter of improving human institutions faster than challenges emerge to threaten them. While these concepts can and have been applied to transportation, a comprehensive theory of sustainable transportation that is measurable and verifiable is still yet to emerge.

1. *Sustainable Transportation and Economic Development*

Throughout history, the expansion and improvement of transportations systems has been an engine of economic development (Greene and Wegener 1997). The fast and reliable movement of persons and commodities contributes more than simply a reduction in the costs of production. It facilitates the reorganization of production and consumption to permit specialization, scale economies, inventory reduction, and more. Today, transportation activity continues to increase at roughly the rate of economic growth in both the developed and developing economies.

Motorized transport is the source of essentially all the demands transportation imposes on natural and environmental resources. The motorization of transport, fully accomplished in the developed economies during the twentieth century, is now proceeding rapidly in the developing world. In 1950, 90 percent of the world's 50 million motor vehicles could be found in Europe or the United States. Today, half of the world's roughly 700 million vehicles are to be found outside of those countries. The developing economies own more than one-third of the world's motor vehicles, and vehicle ownership outside of the developed economies is growing at twice the rates seen in the developed economies. The global potential for motorization is enormous. In developed economies, motor vehicle ownership rates average 500 per 1,000 persons. In the US, there is more than one motor vehicle per person of driving age. Ownership rates in the developing economies are about one-twentieth as high.

Expansion of modern, motorized transportation in the developing economies appears to be essential to their sustainable development. The World Bank has identified three broad types of sustainability in considering transport investments in developing countries: (a) economic and financial sustainability; (b) environmental sustainability; and, (c) social sustainability (World Bank 1996). These categories reflect the view that to be sustainable, transportation systems must have stable financing in the context of market economies, must address the many adverse impacts of transportation on health and ecosystems, and must serve the mobility needs of the poor as well as those with the ability to pay.

2. *Environmentally Sustainable Transport*

Transportation systems have diverse and widespread impacts on the environment, affecting climate, air and water quality, noise, habitats, and biodiversity. Worldwide, transportation accounts for one-fourth of global carbon dioxide emissions. Scientists have concluded that the accumulation of carbon dioxide and other less common anthropogenic greenhouse gases in the world's atmosphere since the onset of the industrial revolution are now having a perceptible impact on the earth's climate. While it is not yet clear what the impacts on climate will be, they are likely to include increases in storm activity, changes in the patterns of rainfall, a general warming leading to higher sea levels, and other potentially harmful or even catastrophic effects. The quantity of carbon dioxide emission produced by transportation depends on the carbon content of the transportation fuels and the energy efficiency of transport systems. Today, motorized transport is entirely dependent on fossil fuels. The combustion of fossil fuels is the predominant source of greenhouse gas emissions from human activities. A typical passenger car adds 1 to 2 tonnes of carbon to the atmosphere each year. Despite 50–100 percent increases in the fuel economy of automobiles over the past 30 years, there appears to be no prospect of reducing global greenhouse gas emissions from transportation for at least 20 to 30 years. Clearly, continuing to force alterations in the earth's climate systems in unpredictable ways does not satisfy the sustainability doctrine.

Transportation is a major source of local and regional air pollution. In Europe and the United States, transportation vehicles emit roughly 45 percent of nitrogen oxides and 40 percent of volatile organic compounds that mix in the atmosphere to form ozone, a pollutant that exacerbates a variety of respiratory problems and causes damage to vegetation as well. Air pollutants from transport have an immediate impact on human health and on vegetation. Ozone pollution has been observed to change the distribution of tree species in forests, retard the growth of vegetation,

and damage crops. Prolonged exposure to ozone is strongly believed to reduce biodiversity of plant communities and diminish variability in their gene pools (NRC 1997). Emissions of oxides of nitrogen and sulfur from motor vehicles contribute to acid rain with widespread negative impacts on terrestrial and aquatic ecosystems. Volatile organic compounds also include a number of toxic pollutants, including known and suspected carcinogens, such as benzene, formaldehyde, cyanide, and dioxin.

Transportation vehicles produce significant quantities of fine particulate pollution via several mechanisms. Vehicle tailpipe emissions contain particles produced by incomplete combustion. Particles and aerosols are formed when certain motor vehicle emissions react in the air. Finally, dust from roads is a major source of air-borne particulate matter. Fine particulate matter can be inhaled deeply into the lungs, contributing to a variety of respiratory problems. Epidemiological studies have correlated elevated morbidity and mortality rates with long-term, low-level exposure to fine particulates, as well as to episodes of acute pollution, suggesting that particulate pollution may be the most serious form of motor vehicle pollution from the perspective of human health (NRC 1997).

Since the 1950s, enormous technological progress has been made in reducing the pollutant emissions from new transportation vehicles. A properly operating motor vehicle manufactured in 2000 produces 1 percent to 10 percent of the pollution (depending on the pollutant) of a car made 35 years earlier. Still, even this rate of progress has been insufficient to achieve air quality standards in many urban areas because the growth of travel has outpaced the rate of technological improvement. Further significant improvements in conventional vehicles and fuels will be made from 2000 to 2010. Yet achieving sustainability will probably require entirely new energy sources and propulsion technologies ultimately.

3. *Solid Waste*

Transportation infrastructure and vehicles generate enormous amounts of solid waste. Some of this waste, such as iron and steel, has a long history of recycling. Other components like plastics most often end up in landfills or incinerators. Waste oil and fluids, batteries, and other components often contain toxic materials that pose serious health and environmental hazards if not properly disposed of. The maintenance and reconstruction of transportation facilities also generate substantial quantities of mostly asphalt and concrete waste. Not only increased recycling, but life-cycle management of materials for transportation systems has been recognized as essential to achieving sustainable transportation systems (OECD 1997).

4. *Noise*

It may seem odd to include noise among the issues affecting transportation's sustainability. Noise, after all, is inherently transitory and its impacts on society and the environment would seem to be also. On the other hand, transportation systems operate more or less continuously, generating noise around the clock. Studies carried out in Europe have shown that about 20 percent of the population are exposed to daytime transport noise exceeding acceptable levels (65 dB(A)) (European Commission 1995). Another 170 million are exposed to noise levels that cause serious annoyance as defined by the World Health Organization. Road noise is the dominant source, followed by rail and air transport. Furthermore, data for the 1980–95 period showed no significant improvement in exposure to traffic noise for the population of Europe. The principal impacts of transportation noise are annoyance, interference with work and sleep and, less commonly, hearing loss. Estimates of the costs of noise pollution in Europe range between 0.1 percent and 2 percent of GDP.

5. *Natural Resources*

Transportation is a major consumer of two critical, limited resources: land and petroleum. Worldwide, transportation accounts for 56 percent of global petroleum consumption and between 1973 and 1997 was responsible for all of the increase in world oil consumption (IEA 2000). Recent assessments of the world's supplies of conventional petroleum indicate that sometime between 2005 and 2025, the world will have consumed half of all the conventional petroleum that is known or believed to exist (EIA 2000 and USGS 2000). Once this point is passed, it is virtually certain that production of conventional petroleum will begin to decline, and unconventional energy sources will have to be substituted at an increasing rate. The world's transportation system, which is 97 percent dependent on petroleum for energy, will not be sustainable unless it can make a transition to alternative energy sources. While there are more than adequate resources of fossil fuels that can be converted to transport fuels (e.g., coal, oil shale, and tar sands), the greater environmental impacts associated with using these fuels raises serious questions about the sustainability of a motorized world transport system based on fossil fuels.

Perhaps even more serious are the questions raised by transportation systems' expanding direct and indirect uses of land. Not only does transportation infrastructure such as road, airports, and railroad tracks, occupy millions of hectares of land, but the infrastructure and vehicles that use it have significant impacts on surrounding ecosystems. Linear transportation structures serve as deadly barriers to the

migration and movements of species, and alter water flow and groundwater. But perhaps more importantly, transportation infrastructure indirectly alters the patterns and density of human settlements, with even more pervasive impacts on ecosystems and biodiversity (NRC 1997).

6. *Sustainable Communities*

Transportation is recognized as a key element of sustainable communities (PCSD 1997). Human development patterns and the transportation systems that serve them are strongly interdependent. Transportation choices made by present generations will constrain the opportunities of future generations because built environments can be changed only slowly and at great cost. The high level of mobility provided by automotive transport has engendered dispersed low-density development, sometimes referred to as sprawl. The area of built-up urban land in the United States increased by an estimated 14 million acres between 1982 and 1992, driven in large part by the expansion of highway infrastructure (BTS 1996, p. 162).

As demand for mobility has continued to grow, while the land available for expanding transportation infrastructure in urban areas has not, chronic congestion has become a major problem in cities around the world. Increasing levels of traffic congestion are not only a source of stress and waste the time of present generations, but can restrict the opportunities for mobility of future generations. If the capacities of transportation systems cannot be expanded to accommodate future growth without incurring unacceptable trade-offs, the options available to future generations will be diminished.

7. *Safety*

Safety must also be included in assessing the sustainability of transportation systems (e.g., Black 2000). While absolutely compelling as a human tragedy, the sense in which transportation safety is a sustainability issue may be less obvious. Globally, there are about 700,000 motor vehicle fatalities annually, and many millions of serious injuries. On the other hand, fatality rates (e.g., per 100,000 vehicle miles) are generally declining. In developed economies such as the United States, recent years have seen declines in total fatalities even while motor vehicle travel and population size continue to grow. In countries undergoing rapid motorization, however, total fatalities are increasing at the same time rates are declining. When both fatalities and fatality rates are increasing, the sustainability of transportation systems is in question. The resource at issue is human life itself, and the needs of future generations would clearly appear to be threatened by increasingly unsafe transportation systems. Where both total fatalities and fatality rates are declining, an argument can be made for sustainability. Without a doubt, safety is the single most important problem facing transportation. Yet, to date there is little agreement on how safety should be interpreted in the context of the sustainability doctrine.

8. *Equity*

Equity among generations is clearly integral to the concept of sustainability. For sustainability to exclude equity considerations among individuals within a generation would be inherently inconsistent. The Charter of European Cities and Towns Towards Sustainability states that the goals of sustainable development are '…to achieve social justice, sustainable economies, and environmental sustainability' (Greene and Wegener 1997). Furthermore, the context of the Brundtland sustainability definition was clearly the need to preserve the global environment while permitting developing countries the economic development that is their right. Clearly social justice and particularly equity have been integral to the concept of sustainability. It has been argued that the concept of intergenerational fairness embodied in sustainability, in fact arises from a societal commitment to equality of opportunity among contemporaries (Howarth 1997). Exactly how equity concerns are to be incorporated in an operating definition of sustainability has yet to be resolved, as have the implications for transportation. At a minimum, equity considerations include the interests of developing economies to pursue higher levels of mobility, as well as the needs of disadvantaged citizens for access to mobility in all societies.

9. *Measuring Sustainability*

If policies are to be formulated and implemented to achieve sustainable transportation, indicators must be found to measure it (OECD 1994). If the steady-state concept of sustainability is accepted, these indicators should show:

whether the world's environmental resources are being reduced by transportation's impacts on air or water pollution;

whether the global climate is being altered by greenhouse gas emissions;

whether habitats are being destroyed through direct or indirect land-use impacts; and

whether the generation of solid waste is increasing or being restrained by recycling.

Indicators must show whether the fundamental natural resources required for transportation, such as energy and land, are being expanded as fast or faster than they are being consumed, and whether trans-

portation systems are being developed and managed so as to enhance social equity by improving access to mobility for all. Until the consequences of transportation for sustainability are made visible measuring them, progress toward sustainable transportation will be limited.

See also: Air Pollution; Desertification; Ecotourism; Environmental Economics; Resource Geography; Safety, Economics of; Tourism, Geography of; Transportation Geography; Transportation Planning; Transportation: Supply and Congestion

Bibliography

Black W R 2000 Socio-economic barriers to sustainable transport. *Journal of Transport Geography* **8**: 141–7

Bureau of Transportation Statistics (BTS) 1996 *Transportation Statistics Annual Report 1996*. US Department of Transportation, Washington, DC

Energy Information Administration (EIA) 2000 *Long-term World Oil Supply*. US Department of Energy, Washington, DC. Available on the World Wide Web at http://www.eia.doe.gov

European Commission 1995 *Towards Fair and Efficient Pricing in Transport*. COM(95)691, Directorate-General for Transport, DG-VII, Brussels, Belgium

Greene D L, Wegener M 1997 Sustainable transport. *Journal of Transport Geography* **5**(3): 177–90

Howarth R B 1997 Sustainability as opportunity. *Land Economics* **73**(4): 569–79

International Energy Agency (IEA) 2000 *Key Energy Statistics*. OECD, Paris. Available on the World Wide Web at http://www.iea.org

National Research Council (NRC), Committee for a study on transportation and a sustainable environment 1997 *Toward a Sustainable Future*. Special Report 251, Transportation Research Board, National Academy Press, Washington, DC

Organization for Economic Co-operation and Development (OECD) 1994 *Environmental Indicators*. OECD, Paris

OECD 1997 *Recycling strategies for road works. Road Transport Research*. Organization for Economic Co-operation and Development, Paris

Pearce D W, Atkinson G D, Dubourg W R 1994 The economics of sustainable development. *Annual Review of Energy and the Environment* **19**: 457–74

Pearce D W, Warford J J 1993 *World Without End*. International Bank for Reconstruction and Development, Washington, DC

President's Council on Sustainable Development (PCSD) 1997 *Sustainable Communities Task Force Report*. Washington, DC

United Nations Conference on Environment and Development (UNCED) 1992 *Agenda 21*. United Nations, New York

United States Geological Survey (USGS) 2000 *US Geological Survey World Petroleum Assessment 2000—Description and Results*. Department of the Interior, Denver, CO

World Bank 1996 *Sustainable Transport*. International Bank for Reconstruction and Development, Washington, DC

World Commission on Environment and Development (WCED) 1987 *Our Common Future*. Oxford University Press, Oxford, UK

D. L. Greene

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Sy

Symbolic Boundaries: Overview

1. Definition and Intellectual Context

'Symbolic boundaries' are the lines that include and define some people, groups and things while excluding others (Epstein 1992, p. 232). These distinctions can be expressed through normative interdictions (taboos), cultural attitudes and practices, and more generally through patterns of likes and dislikes. They play an important role in the creation of inequality and the exercise of power. The term 'symbolic boundaries' is also used to refer to the internal distinctions of classification systems and to temporal, spatial, and visual cognitive distinctions in particular (see *Expression and Interaction: Microfoundations of Expressive Forms*). This article focuses on boundaries within and between groups. It discusses the history, current research, and future challenges of work on this topic.

The literature on symbolic boundaries has gained importance since the 1960s due to a convergence between research on symbolic systems and indirect forms of power. Writings by Pierre Bourdieu, Mary Douglas, Norbert Elias, Erving Goffman, and Michel Foucault on these and related topics have been influential internationally across several disciplines, but particularly in anthropology, history, literary studies, and sociology. In North America, a renewed cultural sociology has produced wide-ranging empirical research agendas on symbolic boundaries and inequality. In other fields including community, cognition, deviance, gender, immigration, knowledge and science, nationalism, professions, race and ethnic studies, and social movements, issues of boundaries have gained analytical prominence, although some authors analyze boundary work without using the language of symbolic boundaries.

2. History

Two of the founding fathers of sociology played central roles in shaping the literature on symbolic boundaries: Emile Durkheim and Max Weber. Their contributions are reviewed before turning to the 'neo-classical' writings of Mary Douglas, Norbert Elias, and Thornstein Veblen, which illustrate the lasting influence of Durkheim and Weber on this literature up to the 1960s. While Durkheim brings attention to classification systems and their relationship with the moral order, Weber is more concerned with their impact on the production and reproduction of inequality.

One of the most widely used examples of symbolic boundaries is taken from Durkheim's later work, *Les formes élementaires de la vie religieuse* (*The Elementary Forms of Religious Life* 1965 [1911]). In this book, Durkheim argues that the distinctiveness of the religious experience from other types of experiences rests in the fact that it involves a symbolic distinction between the realms of the sacred and the profane (pp. 234, 250). The meanings of these realms are mutually exclusive and are defined relationally, through interdictions and rituals that isolate and protect the former from the latter (e.g., a Roman Catholic sinner cannot receive communion until he is purified through confession) (p. 271).

The distinction between the sacred and the profane extends to the whole universe of objects and people in which it takes place. For instance, the status of members of a community is defined by the types of relationship they have with sacred objects (e.g., Roman Catholic women cannot celebrate mass). In this sense, religious systems provide a cosmology, i.e., a general interpretation of how the world is organized and how its elements relate to one another and to the sacred. This cosmology acts as a system of classification and its elements are organized according to a hierarchy (counterpoising for instance the pure with the impure). The belief invested in this 'order of things' structures people's lives to the extent that it limits and facilitates their action. In Durkheim's words, 'the power attached to sacred things conducts men with the same degree of necessity as physical force' (p. 260).

Moving beyond the religious realm, Durkheim points to the existence of a moral order, i.e., a common public system of perception of reality that regulates, structures, and organizes relations in a community. This system operates less through coercion than through intersubjectivity (p. 238). In fact, Durkheim defines society by its symbolic boundaries: it is the sharing of a common definition of the sacred and the profane, of similar rules of conducts and a common compliance to rituals and interdictions that defines the internal bonds within a community. Hence, he posits that the boundaries of the group coincide with those delimitating the sacred from the profane.

Unlike Durkheim, Max Weber is more concerned with the role played by symbolic boundaries (honor) in the creation of social inequality than in the creation of social solidarity. In *Wirtschaft und Gesellschaft*

(1922; *Economy and Society* 1978), he describes human beings as engaged in a continuous struggle over scarce resources. In order to curb competition, they discriminate toward various groups on the basis of their cultural characteristics, such as lifestyle, language, education, race, or religion (Chap. 2). In the process, they form status groups whose superiority is defined in relation to other groups. They cultivate a sense of honor, privilege relationships with group members, and define specific qualifications for gaining entry to the group and for interacting with lower status outsiders (e.g., opposing miscegenation). They invoke their higher status and shared rules of life to justify their monopolization of resources. Hence, cultural understandings about status boundaries have a strong impact on people's social position and access to resources.

Thornstein Veblen's *The Theory of the Leisure Class* (1979 [1899]) parallels Weber's writings on status groups. Veblen is also concerned with the mechanisms that produce boundaries between status groups. This American economist suggests that habits of thought ('classifying and demarcating') are central to these mechanisms and are often organized around notions of superiority and inferiority concerning employment, consumption, and leisure. In his words, 'the concept of dignity, worth, or honor, as applied either to persons or conduct, is of first-rate consequence in the development of class and class distinctions' (p. 15). For instance, idleness symbolizes status because it signifies pecuniary status. It is a way for 'successful men to put their prowess in evidence.' Evidence of nonproductive consumption of time includes 'quasi-scholarly or quasi-artistic accomplishments.' Also, 'refined tastes, manners, and habits of life are useful evidence of gentility because good breeding requires time, application, and expenses' (p. 49) and are therefore not available to those 'whose time and energies are taken up with work.'

Veblen also developed the concept of 'conspicuous consumption' in the context of an acerbic critique of the excesses of the business class. He argues that the possession and display of wealth confers honor: as an invidious distinction, it symbolizes ranking within a group. This way of manifesting superiority is more common when predatory aggression or war are less frequent. Veblen's analysis assumes that there is a usual tendency to change standards of sufficiency as one's pecuniary situation improves, so that one becomes restless with creating 'wider and ever-widening distance' between herself and the average standard.

Also paralleling Weber's work is the work of German sociologist Norbert Elias, *Über Den Prozess der Zivilisation* (*The Civilizing Process* 1982 [1939]). Elias analyzes the historical emergence of a boundary between civilized and barbarian habits by using evidence from Western manner manuals written between the late middle ages and the Victorian period. His attention centers on 'natural' bodily functions such as spitting, defecating, eating, and blowing one's nose to show an 'advance in the threshold of self-control' over time. He demonstrates the growing centrality of shame and embarrassment in instituting norms of behavior in public and private. At a more general level, he shows transformations in standards of behavior and feelings and in personality structures (what he calls 'habitus' or habits emerging from social experience). He argues that these vary across hierarchical groups in society and that these variations are key to pacification and the exercise of power.

To turn now to the lasting influence of Durkheim's work, in *Purity and Danger* (1966), Mary Douglas is concerned with the order-producing, meaning-making, and form-giving functions of classification systems and the role of rituals in creating boundaries grounded in fears and beliefs. In *Natural Symbols* (1970), she takes on the idea of a correspondence between classification systems and social organization advanced by Durkheim and Mauss (1963 [1903]). She describes the structure of binary symbolic systems as 'reflecting' that of group structures. Like Elias, she is also concerned with the moral order and centers her attention on the system of social control as expressed through the body and through the observable artifacts of everyday life (food, dirt, and material possessions). She argues that the very basis of order in social life is the presence of symbols that demarcate boundaries or lines of division.

One of Douglas' main concerns is how communities differentiate themselves from one another and how they are internally differentiated. She distinguishes groups on the basis of their degree of social control and of the rigidity of their grid (by which she means the scope and coherent articulation of their system of classification or the extent to which it is competing with other systems). In societies with high social control and great cultural rigidity (i.e., what she calls high grid and group), there is a concern to preserve social boundaries; the role structure is clearly defined and formal behavior is highly valued and well-defined in publicly insulated roles. Through 'the purity rule,' formality screens out irrelevant organic processes, 'matters out of place.' Douglas suggests that the more complex the system of classification and the stronger the pressure to maintain it, the more social intercourse pretends to take place between disembodied spirits, i.e., the more the purity rule applies.

3. *Current Theory and Research*

In the contemporary literature on symbolic boundaries, both the neo-Weberian and neo-Durkheimian heritage remain strong. The question of how boundaries intersect with the production of inequality has attracted great interest in recent years, following the publication of Pierre Bourdieu's impressive corpus. In

the United States in particular, cultural sociologists have been working to assess some of Bourdieu's theoretical claims and to use his work as a stepping stone for improving our understanding of the cultural aspects of class, gender, and racial inequality. Other important developments concern the study of identity through boundary work, and research on moral order, community, and symbolic politics. As is argued in the final section of this article, social scientists will soon face the challenge of integrating work from a wide range of fields that have used the concept of boundaries to various ends.

3.1 Culture and Inequality

In the last 20 years, a large neo-Weberian literature emerged around the study of processes of closure, as illustrated most notably by the work of Frank Parkin and Randall Collins. Parkin (1979) drew on Weber to propose an analysis of class relationship that focuses on the distributive struggle for monopolizing or usurping resources within and across classes—with an emphasis on the right of ownership and credentialism, i.e., the use of educational certificates to monopolize positions in the labor market. Equally inspired by Durkheim, Collins (1998) extended his earlier work on credentialism and interaction rituals to analyze how intellectuals compete to maximize their access to key network positions, cultural capital, and emotional energy, which generates intellectual creativity. These authors' contributions intersect with those of Pierre Bourdieu and his collaborators, although their ideas followed an independent path of development.

In *Reproduction* (1977 [1970]), Pierre Bourdieu and Jean-Claude Passeron proposed that the lower academic performance of working class children cannot be accounted for by their lower ability but by institutional biases against them. They suggest that schools evaluate all children on the basis of their familiarity with the culture of the dominant class (or cultural capital), thus penalizing lower-class students. Extensive vocabulary, wide-ranging cultural references, and command of high culture are valued by the school system and students from higher social backgrounds are exposed to this class culture at home. Hence, children of the dominated classes are overselected by the educational system. They are not aware of it, as they remain under the spell of the culture of the dominant class. They blame themselves for their failure, which leads them to drop out or to sort themselves into lower prestige educational tracks.

This work can be read as a direct extension of Karl Marx and Friedrich Engels' 'dominant ideology thesis,' which centers on the role of ideology in cementing relations of domination by camouflaging exploitation and differences in class interests. However, Bourdieu and Passeron are more concerned with classification systems than with representations of the social world itself, i.e., with how representations of social relationships, the state, religion, and capitalism contribute to the reproduction of domination. Implicitly building on the work of Antonio Gramsci, they made inroads in analyzing the subjective process of consolidation of class domination, focusing on the shaping of cultural categories. The control of subjectivity in everyday life through the shaping of common sense and the naturalization of social relations is the focus of their attention. They broaden Marx and Engels by suggesting that crucial power relations are structured in the symbolic realm proper, and are mediated by meaning. They *de facto* provide a more encompassing understanding of the exercise of hegemony by pointing to the incorporation of class-differentiated cultural dispositions mediated by both the educational system and family socialization.

In *Distinction* (1984 [1979]), Bourdieu applies this analysis to the world of taste and cultural practice at large. He shows how the logic of class struggle extends to the realm of taste and lifestyle, and that symbolic classification is key to the reproduction of class privileges: dominant groups define their own culture and ways of being as superior. Thereby they exercise 'symbolic violence,' i.e., impose a specific meaning as legitimate while concealing the power relations that are the basis of its force (Bourdieu and Passeron 1977 [1970], p. 4). They define legitimate and 'dominated' cultures in opposition: the value of cultural preferences and behaviors are defined relationally around binary oppositions (or boundaries) such as high/low, pure/impure, distinguished/vulgar, and aesthetic/practical (p. 245). The legitimate culture they thereby define is used by dominant groups to mark cultural distance and proximity, monopolize privileges, and exclude and recruit new occupants to high status positions (p. 31). Through the incorporation of '*habitus*' or cultural dispositions, cultural practices have inescapable and unconscious classificatory effects that shape social positions.

A large American literature applying, extending, assessing, and critiquing the contributions of Bourdieu and his collaborators developed in the wake of their translation in English (for a review, see Lamont and Lareau 1988). For instance, DiMaggio (1987) suggests that boundaries between cultural genres are created by status groups to signal their superior status. Lamont (1992, Chap.7) critiqued Bourdieu (1979) for exaggerating the importance of cultural capital in upper-middle class culture and for defining salient boundaries *a priori* instead of inductively. By drawing on interviews with professionals and managers, she showed that morality, cultural capital, and material success are defined differently and that their relative importance vary across national contexts and by subgroups. Lamont also showed variations in the extent to which professionals and managers are tolerant of the lifestyles and tastes of other classes, and argued that cultural *laissez-faire* is a more important

feature of American society than French society. High social and geographic mobility, strong cultural regionalism, ethnic and racial diversity, political decentralization, and relatively weak high culture traditions translate into less highly differentiated class cultures in the United States than France.

Other sociologists also argue that cultural boundaries are more fluid and complex than cultural capital theory suggests. In particular, in his study of group variation in home decoration, Halle (1993) suggests that art consumption does not necessarily generate social boundaries. He finds that the meaning attached to living room art by dwellers is somewhat autonomous from professional evaluations, and is patterned and influenced by a wide range of factors beyond class—including neighborhood composition. He also finds that cultural consumption is less differentiated than cultural capital theory suggests—with landscape art being appreciated by all social groups for instance. Finally, he suggests that 'the link between involvement in high culture and access to dominant class circles … is undemonstrated' (p. 198). For his part, Hall (1992) emphasized the existence of heterogeneous markets and of multiple kinds of cultural capital. He proposes a 'cultural structuralism' that addresses the multiplicity of status situation in a critique of an overarching market of cultural capital.

Bryson (1996), Erickson (1996), and Peterson and Kern (1996) suggest that cultural breadth is a highly valued resource in the upper and upper-middle classes, hence contradicting Bourdieu's understanding of the dominant class which emphasizes exclusively the boundaries they draw toward lower class culture. Bryson (1996) finds that musical exclusiveness decreases with education. She proposes that cultural tolerance constitutes a multicultural capital more strongly concentrated in the middle and upper classes than in the lower classes. Erikson (1996) suggests that although familiarity with high status culture correlates with class, it is useless in coordinating class relations in the workplace. She writes that the 'culture useful for coordination is uncorrelated … with class, popular in every class' (p. 248) and that 'the most useful overall cultural resource is variety plus a well-honed understanding of which [culture] genre to use in which setting' (p. 249). For their part, Peterson and Kern (1996) document a shift in high status persons from snobbish exclusion to 'omnivorous appropriation.' These studies all call for a more multidimensional understanding of cultural capital as a basis for drawing boundaries, and counter Bourdieu's postulate that the value of tastes is defined relationally through a binary or oppositional logic.

3.2 Identity and Boundary Work

The growing literature on identity is another arena where the concept of symbolic boundaries has become more central. In particular, sociologists and psychologists have become interested in studying boundary work, a process central to the constitution of the self.

Social psychologists working on group categorization have been studying the segmentation between 'us' and 'them.' Brewer's (1986) social identity theory suggests that 'pressures to evaluate ones' own group positively through in-group/out-group comparison lead social groups to attempt to differentiate themselves from each other.' This process of differentiation aims 'to maintain and achieve superiority over an out-group on some dimension' (Tajfel and Turner 1985, pp. 16–17). While these authors understand the relational process as a universal tendency, sociologists are concerned with analyzing precisely how boundary work is accomplished, i.e. with what kinds of typification systems, or inferences concerning similarities and differences, groups mobilize to define who they are.

The concept of boundary work was proposed originally by Gieryn in the early 1980s to designate 'the discursive attribution of selected qualities to scientists, scientific methods, and scientific claims for the purpose of drawing a rhetorical boundary between science and some less authoritative residual non-science' (1999, pp. 4–5). In recent years, sociologists have become interested in analyzing this process by looking at self-definitions of ordinary people, while paying particular attention to the salience of various racial and class groups in boundary work. For instance, Newman (1999) analyzes how poor fast-food workers define themselves in opposition to the unemployed poor. Lamont (1992) studies the boundary work of professionals and managers, while Lamont (2000) examines how workers in the United States and France define worthy people in opposition to the poor, 'people above,' blacks, and immigrants, drawing moral boundaries toward different groups across the two national context. Lichterman (1999) explores how volunteers define their bonds and boundaries of solidarity by examining how they articulate their identity around various groups. He stresses that these mappings translate into different kinds of group responsibility, in 'constraining and enabling what members can say and do together.' Binder (1999) analyzes boundaries that proponents of Afrocentrism and multiculturalism build in relation to one another in conflicts within the educational system. Finally, Gamson (1992) analyzes how the injustice frames used in social movements are organized around 'us' and 'them' oppositions.

Jenkins's (1996) study of social identity provides useful tools for the study of boundary work. He describes collective identity as constituted by a dialectic interplay of processes of internal and external definition. On the one hand, individuals must be able to differentiate themselves from others by drawing on criteria of community and a sense of shared belonging within their subgroup. On the other hand, this internal

identification process must be recognized by outsiders for an objectified collective identity to emerge. Future research on the process of collective identity formation may benefit from focusing on the dynamic between self-identification and social categorization.

3.3 Moral Order, Community, and Symbolic Politics

A third strand of work on symbolic boundaries presents more palpable neo-Durkheimian influences. Several empirical studies have centered on moral order and on communities. Wuthnow (1987, p. 69) writes 'Order has somehow to do with boundaries. That is, order consists mainly of being able to make distinctions—of having symbolic demarcations—so that we know the place of things and how they relate to one another.' A recent example of this neo-Durkheimian line of work is Alexander's (1992) semiotic analysis of the symbolic codes of civic society. The author describes these codes as 'critically important in constituting the very sense of society for those who are within and without it.' He also suggests that the democratic code involves clear distinctions between the pure and the impure in defining the appropriate citizen. His analysis locates those distinctions at the levels of people's motives and relationships, and of the institutions that individuals inhabit (with 'honorable' being valued over 'self-interested' or 'truthful' over 'deceitful' in the case of the democratic code).

It should be noted that the last decades have produced several studies of status politics that documented precisely how groups sharing a lifestyle made such distinctions, engaged in the maintenance of the moral order, and simultaneously bolstered their own prestige. Particularly notable is Gusfield (1963) who analyzed the nineteenth century American temperance movement in favor of the prohibition and the Eighteenth Amendment to the constitution. Gusfield understands this movement as a strategy used by small-town Protestants to bolster their social position in relation to urban Catholic immigrants. Along similar lines, Luker (1984) describes the worldviews of anti-abortion and pro-choice activists. She shows that they have incompatible beliefs about women's careers, family, sexuality, and reproduction, and that they talk past one another and define themselves in opposition to one another. The literature on social movements includes numerous additional studies that focus on the process by which categories of people are turned into categories of enemies (Jasper 1997, Chap. 16).

4. Challenges and Future Directions

Two main challenges concerning the study of boundaries are pointing at the horizon of sociological scholarship. They concern (a) a necessary synthesis of the various strands of work that speak to boundary issues across substantive areas; and (b) the study of the connection between objective and subjective boundaries.

The concept of boundaries is playing an increasingly important role in a wide range of literatures beyond those discussed above. For instance, in the study of nationalism, citizenship, and immigration, scholars have implicitly or explicitly used the boundary concept to discuss criteria of membership and group closure within imagined communities (e.g., Anderson 1983, Brubaker 1992). Some of these authors are concerned with the established rules of membership and boundary work. Yet others study national boundary patterns, i.e., the ways in which nations define their identity in opposition to one another (Saguy 2000).

The concept of boundary is also central in the study of ethnicity and race. The relational process involved in the definition of collective identity ('us' vs. 'them') has often been emphasized in the literature on these topics. The work of Barth (1969) for instance concerns objective group boundaries and self-ascription, and how feelings of communality are defined in opposition to the perceived identity of other racial and ethnic groups. Along similar lines, Bobo and Hutchings (1996) understand racism as resulting from threats to group positioning. They follow Blumer (1958) who advocates 'shift[ing] study and analysis from a preoccupation with feelings as lodged in individuals to a concern with the relationships of racial groups…[and with] the collective process by which a racial group comes to define and redefine another racial group' (p. 3).

Finally, gender and sexual boundaries are also coming under more intense scrutiny. For instance, Epstein (1992) points out that dichotomous categories play an important part in the definition of women as 'other' and that much is at stake in the labeling of behaviors and attitudes as feminine or masculine (also see Gerson and Peiss 1985). Those who violate gender boundaries in illegitimate ways often experience punishment in the workplace. More recently, Tilly (1997) argues that dichotomous categories such as 'male' and 'female' (but also 'white' and 'black') are used by dominant groups to marginalize other groups and block their access to resources. He extends the Weberian scheme by pointing to various mechanisms by which this is accomplished, such as exploitation and opportunity hoarding. He asserts that durable inequality most often results from cumulative, individual, and often unnoticed organizational processes.

Because these various literature all deal with the same social process, boundary work, it may be appropriate at this point to begin moving toward a general theory of boundaries by, for instance, identifying similarities and differences between boundaries drawn in various realms—moral, cultural, class, racial, ethnic, gender, and national boundaries. This could be

accomplished by focusing on a number of formal features and characteristics of boundaries, such as their visibility, permeability, boundedness, fluidity, and rigidity. We may also want to compare embedded and transportable boundaries; explicit and taken-for-granted boundaries; positive and negative boundaries; and the relationship between representations of boundaries and context. Social scientists should also think more seriously about how different types of boundaries can combine with one another across local and national contexts (e.g., how cultural or moral boundaries combine with race, gender, or class boundaries).

A second challenge will be to understand the connection between objective boundaries and symbolic boundaries. Students of objective boundaries have focused on topics such as the relative importance of educational endogamy vs. racial endogamy among the college educated; racial hiring and firing; the extent of residential racial segregation; the relative permeability of class boundaries; and the process of creation of professional boundaries. Lamont (1992) has argued that symbolic boundaries are a necessary but insufficient condition for the creation of objective boundaries. More empirical work is needed on the process by which the former transmutes into the latter.

See also: Collective Identity and Expressive Forms; Discourse and Identity; Expressive Forms as Generators, Transmitters, and Transformers of Social Power; Leisure and Cultural Consumption; Networks and Linkages: Cultural Aspects

Bibliography

Alexander J 1992 Citizens and enemy as symbolic classification: on the polarizing discourse of civil society. In: Lamont M, Fournier M (eds.) *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago

Anderson B 1983/1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Barth F 1969 Introduction. In: Barth F (ed.) *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*. George Allen and Unwin, London

Binder A 1999 Friend and foe: boundary work and collective identity in the Afrocentric and multicultural curriculum movements in American public education. In: Lamont M (ed.) *The Cultural Territories of Race: Black and White Boun*daries. University of Chicago Press, Chicago and Russell Sage Foundation, New York

Blumer H 1958 Race prejudice as a sense of group position. *Pacific Sociological Review* **1**: 3–7

Bobo L, Hutchings V L 1996 Perceptions of racial group competition: extending Blumer's theory of group position to a multiracial social context. *American Sociological Review* **61**: 951–72

Bourdieu P 1979 *Distinction, Critique sociale du jugement*. Editions de Minuit, Paris [1984 *Distinction: A Social Critique of the Judgment of Taste*. Harvard University Press, Cambridge, MA]

Bourdieu P, Passeron J-C 1970 *Reproduction. Elements pour une théorie du système d'enseignment*, Editions de Minuit, Paris [1977 *Reproduction in Education, Society, and Culture*. Sage, Beverly Hills, CA]

Brewer M B 1986 The role of ethnocentrism in intergroup conflict. In: Worchel S, Austin W G (eds.) *Psychology of Intergroup Relations*. Nelson-Hall Publishers, Chicago

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Bryson B 1996 'Anything but Heavy Metal': symbolic exclusion and musical dislikes. *American Sociological Review* **61**(5): 884–99

Collins R 1998 *The Sociology of Philosophies: A Global Theory of Intellectual Changes*. Harvard University Press, Cambridge, MA

DiMaggio P 1987 Classification in art. *American Sociological Review* **52**: 440–55

Douglas M 1970 *Natural Symbols*. Vintage, New York

Douglas M 1966 *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. Routledge & Kegan Paul, London

Durkheim E 1911 *Les formes élémentaires de la vie religieuse*. Alcan, Paris [1965 *The Elementary Forms of Religious Life*. Free Press, New York]

Durkheim E, Mauss M 1903 De quelques formes primitives de classification: contribution à l'étude des représentations collectives. *Année Sociologique* **6**: 1–72 [1963 *Primitive Classification*, University of Chicago Press, Chicago]

Elias N 1939 *Über Den Prozess der Zivilisation. Soziogenetische und Psychogenetische Untersuchungen*, 2 Vols. Haus Zum Flakin, Basel [1982 *The Civilizing Process*. Blackwell, New York]

Epstein C F 1992 Tinker-bells and pinups: the construction and reconstruction of gender boundaries at work. In: Lamont M, Fournier M (eds.) *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago

Erickson B 1996 Culture, class, and connections. *American Journal of Sociology* **102**: 217–51

Gamson W 1992 *Talking Politics*. Cambridge University Press, New York

Gerson J M, Peiss K 1985 Boundaries, negotiation, consciousness: reconceptualizing gender relations. *Social Problems* **32**: 317–31

Gieryn T F 1999 *Cultural Boundaries of Science: Credibility on the Lines*. University of Chicago Press, Chicago

Gusfield J 1963 *Symbolic Crusade*. University of Illinois Press, Urbana, IL

Hall J R 1992 The capital(s) of cultures: a non-holistic approach to status situations, class, gender, and ethnicity. In: Lamont M, Fournier M (eds.) *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago

Halle D 1993 *Inside Culture. Art and Class in the American Home*. University of Chicago Press, Chicago

Jasper J M 1997 *The Art of Moral Protest: Culture, Biography and Creativity in Social Movements*. University of Chicago Press, Chicago

Jenkins R 1996 *Social Identity*. Routledge, London

Lamont M 1992 *Money, Morals, and Manners: The Culture of the French and American Upper-Middle Class*. University of Chicago Press, Chicago

Lamont M 2000 *The Dignity of Working Men: Morality and the Boundaries of Race, Class, and Immigration*. Harvard Uni-

versity Press, Cambridge MA and Russell Sage Foundation, New York
Lamont M, Lareau A 1988 Cultural capital: allusions, gaps, and glissandos in recent theoretical developments. *Sociological Theory* **6**: 153–68
Lichterman P 1999 *Civic Culture Meets Welfare Reform: Religious Volunteers Reaching Out*. Paper presented at the Annual Meeting of the American Sociological Association, Chicago
Luker K 1984 *Abortion and the Politics of Motherhood*. University of California Press, Berkeley, CA
Newman K 1999 *No Shame in My Game: The Working Poor in the Inner City*. Knopf and Russell Sage Foundation, New York
Parkin F 1979 *Marxism and Class Theory: A Bourgeois Critique*. Columbia University Press, New York
Peterson R A, Kern R 1996 Changing highbrow taste: from snob to omnivore. *American Sociological Review* **61**: 900–7
Saguy A C 2000 Sexual harassment in France and the United States: activists and public figures defend their positions. In: Lamont M, Thévenot L (eds.) *Rethinking Comparative Cultural Sociology: Polities and Repertoires of Evaluation in France and the United States*. Cambridge University Press, Cambridge, UK
Tajfel H, Turner J C 1985 The social identity theory of intergroup behavior. In: Worchel S, Austin W G (eds.) *Psychology of Intergroup Relations*. Nelson-Hall, Chicago
Tilly C 1997 *Durable Inequality*. Harvard University Press, Cambridge, MA
Veblen T 1899/1979 *The Theory of the Leisure Class*. Penguin, New York
Weber M 1922/1956 *Wirtschaft und Gesellschaft. Grundisse der Verstehenden Soziologie*, Hohrn, Tubingen [1978 *Economy and Society*. University of California Press, Berkeley, CA, Vol. 1]
Wuthnow R 1987 *Meaning and Moral Order: Explorations in Cultural Analysis*. University of California Press, Berkeley, CA

M. Lamont

Symbolic Interaction: Methodology

The term 'symbolic interactionism' (SI) was coined in an article by Blumer (1937). SI presented a radical break from the traditional, macrostructural theories of sociology. SI rejected the scientific method and its reliance on predetermined hypotheses. SI's goal was not predicting about society, as structuralist methodologies did, but understanding the nature of the social world. In basing its research in the real world of everyday life, SI relied heavily on the theoretical work of the pragmatists, especially that of Mead (1934).

1. *Blumer and Naturalistic Research*

In his book containing reprints of his earlier articles, Blumer (1969) applies a methodology, naturalistic research, to the theoretical work of Mead. In symbolic interactionism, theory and methodology are intertwined very closely. Mead's social pragmatism advocated a study of society that takes place in small group interaction; namely, the relation of self and others. This reciprocal interaction is derived, for Mead, through the use of symbols, primarily language but others as well. Thus, society, for Mead, is a consensual, intersubjective world where the sharedness of meanings among its members allows some stability in an ever-changing flux (cf. Adler and Adler 1980).

Blumer, who was first a student and then a colleague of Mead at the University of Chicago, followed the latter's theoretical ideas and stated that society exists in action, in the reciprocal exchange of symbols, and the meaning these have for the members of society. If society is to be found in the interactional processes of its members we should study it in 'the real world,' i.e., the world created by societal members in their interaction, rather than in artificial settings or hypothetical situations. For Blumer, it follows that rather than constrain the study of social processes with controlled variables we should proceed with minimal intervention by the researcher. Also, in trying to maintain the integrity of the phenomenon under investigation we should not fit the 'real world' to our method, as typical in survey research, but should adapt out methodology to the world and thus use flexible data collection strategies. Finally, the model of the study should not be predetermined and then tested, but should be everchanging to fit the changes and events in the real world.

For Blumer the naturalistic research has two phases: exploration and inspection. Exploration allows the researcher to gain a clearer understanding of the study, what methods to use, and to conceptualize the area of study in order to be able to talk from facts rather than speculations (Blumer 1969). Exploration should be flexible and change as necessary to understand the topic at hand and thus should rely on a multiplicity of methods including observation, interviewing, life histories, and personal documents. Exploration is the first step of naturalistic inquiry for Blumer but must be complemented by another, more analytic step, inspection. This latter step aids the researcher in developing an analytic understanding of the elements of inquiry and their relations. This understanding is sensitizing rather than definitive, since the world studied is a process, i.e., ever changing. Furthermore, inspection allows for the testing of theoretical propositions. Thus, naturalistic research uses exploration to define the parameters of the study, then turns to inspection to develop sensitizing concepts about the study, finally developing theoretical concepts about society and its members (Hammersley 1989).

1.1 *The Chicago School*

Blumer mentioned some works of the Chicago School and particularly *The Polish Peasant in Europe and*

America (Thomas and Znaniecki 1918–20) as exemplars of the naturalistic method. Indeed, many works of the Chicago School are considered the antecedents of SI, since they tended to rely on some facet of the naturalistic method. Both Mead and Blumer were at the University of Chicago in the 1920s and 1930s when the school had great influence on American sociology. Among many other important contributors to the naturalistic studies of the Chicagoans, besides W. I. Thomas, who was one of the theoretical leaders (but left as early as 1918), Robert Park must be mentioned. He was an ex-journalist trained in philosophy and took some courses from the pragmatists, John Dewey and William James. From his previous profession Park brought an understanding of studying the real world 'out there,' thus, being the inspirator of many studies of naturalistic nature. It must be noted that the Chicago School has been criticized for saying very little about their methods in their writings (Hammersley 1989) and for not really being a school of thought at all (Becker 1999). Some of the most notable works from the Chicago naturalistic tradition are *The Ghetto* (Wirth 1928), *The Delinquent Gang* (Thrasher 1927), *The Gold Coast and the Slum* (Zorbaugh 1929), *The Hobo: The Sociology of the Homeless Man* (Anderson 1923). For a detailed study of the Chicago School see Faris (1967) and Bulmer (1984).

2. *Other Methods Based on Mead*

While Blumer's adaptation of Mead's theories is the methodological mainstay of SI, there are other methodologies based on SI, and these will be mentioned next. Analytic induction was first discussed by Znaniecki (1928). It was later used, with minor variations by Lindesmith (1937, 1968) (he was a graduate student of Blumer), Cressey (1950) (a student of Lindesmith), Becker (1963) (see Hammersley 1989), and others. Analytic induction, according to Znaniecki, recognizes the fact that objects in the world are open to an infinite number of description and, thus, our account of them must be selective; this selectivity will be based on the interest at hand, which for sociologists is primarily social and cultural systems; commonly used sociological methods relying on pre-identification (deductive) or superficial description (inductive) will not work, only analytic induction will accomplish the task. The researcher will select a small number of cases (10–12, usually) and study them in depth, continually defining and redefining the event and formulating and reformulating theoretical propositions until they will fit all cases. Negative cases must also be examined (this was Lindesmith's idea).

Another student of Blumer, Strauss, together with Glaser, developed another SI method, grounded theory (Glaser and Strauss 1967). Closely related both to Blumer's methodology and to analytic induction, grounded theory placed more emphasis on the generation and development of theory. Relying on the inductive method, grounded theory is akin to Blumer's inspection, only much more elaborate. Rather than relying on a priori population, in analytic theorizing one continues to study new cases until the point of saturation, generating theoretical categories.

Not all SI methods followed the constructionist approaches outlined above. A notable exception came from the Iowa School of Sociology. Kuhn (1964) adopted a much more deterministic approach to Mead's discussion of the self and the nature of the 'me,' the various roles and images we have of ourselves. The methodology he adopted to discover the nature of the self was called the Twenty Statements Test (TST), a series of open-ended questions about the self. Kuhn felt that rather than use the oblique method of observing people one ought to ask them directly about the nature of their inner feelings and they would honestly disclose them to the researcher. The results of TST would be used, by Kuhn, to outline generic laws that would apply to human beings in different situations. Other positivistic oriented symbolic interactionists are Sheldon Stryker, described as a 'structural role theorist,' who influenced numerous students at the University of Indiana and Carl Couch, who was a stalwart of the discipline, with his 'Behavioral Sociology' at the University of Iowa (cf. Reynolds 1993).

In the 1960s and 1970s a plethora of theoretical approaches, largely based on the naturalistic method, appeared. Some were based on basic Meadian tenets, such as dramaturgy (Goffman 1959), and labeling (Becker 1963). Others based their constructionist approach not only on the ideas of Mead but on those of the phenomenologists (Husserl, Schutz,Heidegger, Dilthey) and the existentialists (Merleau-Ponty, Sartre), and ordinary language philosophers (Wittgenstein). They are phenomenological sociology, existential sociology, ethnomethodology, and the sociology of emotions (see Douglas et al. (1980) for a survey of these sociologies and a list of references to them; also, see Adler et al. (1987)).

3. *Postmodern-informed Symbolic Interactionism*

More recently, with the influence of postmodernism on the social sciences, the methodology of SI has undergone some dramatic changes (Marcus and Fischer 1986, Dickens and Fontana 1994). Again, the closeness of theory and methodology in SI affects methodological approaches to SI. Postmodernism (Lyotard 1984) advocates a disenchantment with metatheories, claiming all we can possibly know about society is much smaller in scope, and it is relative due to its historical, temporal, regional nature, as well as other elements. Thus, society becomes fragmented and

we should question and deconstruct all overarching metatheories and reduce the scope and intent of our studies to our immediate surroundings, to the details of everyday life.

These beliefs impact the naturalistic method of SI in a number of ways. First, we can no longer claim to be aiming at universal laws of human behavior since that would be a metatheory. Second, metatheories come under scrutiny and are deconstructed, i.e., their basic assumptions are rendered problematic and questioned. This leads to the questioning of the role of the researcher and author in terms of authority of selectively choosing data to record and report, thus biasing the study. Postmodern-informed interactionism attempts to remedy this authorial bias by the use of polyphony, by this meaning to allow the many voices of the respondents to speak for themselves. Additionally, researchers disclose as much information as possible about themselves—role, interest, possible bias, and more. This is only a partial solution to the authorial influence, which is ultimately, irremediable, as we cannot help but to perceive and report selectively.

Other methodological changes take aim at the relation between the researcher and the researched. Traditionally, interactionists have attempted to remain as neutral and invisible as possible in order not to influence the study. Postmodern-informed interactionists make their presence felt in an attempt to establish a one-to-one partnership with the respondents. In so doing researchers are no longer aloof observers collecting data but become a part of the interaction and influence the research findings. Now data stems from a negotiated process between researcher and respondent (Gubrium and Holstein 1997) (notice that this goal contradicts the previous goal of reducing authorial bias).

Another change involves the notion of 'value free' research. Some postmodern-informed interactionists wish to overthrow traditional sociology (Clough 1998, Denzin 1997, 1999a) which they consider patriarchal, instead they envision a new researcher who is a 'feminist, communitarian reseacher … who builds collaborative, reciprocal, trusting, and friendly relations with those studied' (Denzin 1997).

4. *Current Trends*

Postmodern-informed interactionists not only have made dramatic methodological changes from Blumerian interactionism but are experimenting with radically different reporting styles, which are closely intertwined with the new methodological ethos. Some of these styles include poetry and performances, as well as the writing of short stories about one's personal past (autoethnographies). While some of these reports are highly engaging (Richardson 1997, Denzin 1999b), they raise a question about standards—should these works be judged and evaluated by the standards of sociology or poetry and literature? Some interactionists have been highly critical of postmodern-informed interactionism (Seidman 1991, Shalin 1993, Best 1995, Prus 1996). As a result the field of interactionism is currently divided in many groups with different methodological approaches.

Adler and Adler (1999), in describing the varied groups of interactionists at the turn of the millenium, use the metaphor of 'the ethnographers' ball' and sit the factions at various tables. We will rely upon the Adlers' description in the following section. The first table sits the 'postmodern-informed' ethnographers, next to them is the 'autoethnography group,' a bit further down sit the 'interpretive discourse analysts' who attempt to integrate interactionism, ethnomethodology, and postmodernism. At another table the Adlers have the 'classic interactionists,' in the wake of the old Chicago School. Next to them sit the 'no mo pomo' (Adler and Adler 1999), the Blumerian group critical of the new postmodern influence. Near by are the followers of Erving Goffman, the 'dramaturgists,' studying everyday life by using the metaphor of the theater. They are followed by the 'grounded theorists,' next to the 'conversational analysts' who are continuing the work of ethnomethodologists such as the late Harvey Sacks and others. Next, the Adlers have a table for those who practice 'the extended case study,' using ethnography to do applied research. Finally we see the 'visual interactionists' who capture symbolic interaction through the eye of the camera lens. As the Adlers themselves say, 'the list can go on' (Adler and Adler 1999).

We would add to the Adler's list three more tables, before we run out of room in the ballroom. One would be the 'phenomenological/existential' table sitting Jack Douglas and his former students at the San Diego campus of the University of California. Another table, right next to the postmodernists, would be the 'feminist interactionists' who are 'united in their criticisms of "Eurocentric masculinist approaches"' (Collins, quoted in Denzin 1997). Finally, there would be a table of 'theoretical interactionists' who have written extensively in the area without actually engaging in ethnographic work (cf. Lyman and Scott 1989).

The methodology of Symbolic Interactionism has become the methodologies of symbolic interactionism, and can no longer be grouped under Blumer's guiding ideas, yet their diversity is exaggerated and exacerbated by rivalry among various subgroups. By and large, while differing in some methodological points, the great majority of interactionists still wish to study the real world and the meaning that its members make of it and how they manage to maintain and achieve social order as a negotiated interactional accomplishment.

See also: Ethnomethodology: General; Grounded Theory: Methodology and Theory Construction; In-

teractionism: Symbolic; Postmodernism: Methodology; Postmodernism: Philosophical Aspects

Bibliography

Adler P, Adler P 1980 Symbolic interactionism. In: Douglas J D, Adler P A, Adler P, Fontana A, Freeman C R, Kotarba J A (eds.) *Introduction to the Sociologies of Everyday Life*. Allyn and Bacon, Boston

Adler P, Adler P 1999 The ethnographers' ball revisited. *Journal of Contemporary Ethnography* **28**(5): 442–50

Adler P, Adler P, Fontana A 1987 Everyday life sociology. *Annual Review of Sociology* **13**: 217–35

Anderson N 1923 *The Hobo: The Sociology of the Homeless Man*. University of Chicago Press, Chicago

Becker H S 1963 *Outsiders: Studies in the Sociology of Deviance*. Free Press of Glencoe, London

Becker H 1999 The Chicago school, so-called. *Qualitative Sociology* **22**(1): 3–12

Best J 1995 Lost in the ozone again. *Studies in Symbolic Interaction* **17**: 120–30

Blumer H 1937 Social psychology. In: Schmidt E P (ed.) *Man and Society*. Prentice-Hall, New York

Blumer H 1969 *Symbolic Interactionism: Perspective and Method*. Prentice-Hall, Englewood Cliffs, NJ

Bulmer M 1984 *The Chicago School of Sociology*. University of Chicago Press, Chicago

Clough P T 1998 *The End(s) of Ethnography: From Realism to Social Criticism*, 2nd edn. Peter Lang, New York

Cressey D 1950 The criminal violation of financial trust. *American Sociological Review* **15**: 738–43

Denzin N K 1989 *Interpretive Interactionism*. Sage, Newbury Park, CA

Denzin N 1997 *Interpretive Ethnography: Ethnographic Practices for the 21st Century*. Sage, Thousand Oaks, CA

Denzin N 1999a Interpretive ethnography for the next century. *Journal of Contemporary Ethnography* **28**(5): 510–19

Denzin N 1999b Performing Montana. In: Glassner B, Hertz R (eds.) *Qualitative Sociology as Everyday Life*. Sage, Thousand Oaks, CA

Dickens D R, Fontana A 1994 *Postmodernism and Social Inquiry*. The Guilford Press, New York

Douglas J D, Adler P A, Adler P, Fontana A, Freeman C, Kotarba J A (eds.) 1980 *Introduction to the Sociologies of Everyday Life*. Allyn and Bacon, Boston

Faris R E L 1967 *Chicago Sociology 1920–1932*. Chandler Pub. Co., San Francisco

Glaser B G, Strauss A L 1967 *The Discovery of Grounded Theory*. Aldine, Chicago

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, Garden City, New York

Gubrium F, Holstein J A 1997 *The New Language of Qualitative Methods*. Oxford University Press, New York

Hammersley M H 1989 *The Dilemma of Qualitative Methods: Herbert Blumer and the Chicago Tradition*. Routledge, London

Kuhn M 1964a Major trends in symbolic–interaction theory in the past twenty-five years. *Sociological Quarterly*. **5**: 61–84

Lindesmith A R 1937 *The Nature of Opiate Addiction*. The University of Chicago Libraries, Chicago

Lindesmith A 1968 *Addiction and Opiates*. Aldine, Chicago

Lyman S, Scott M 1989 *Sociology of the Absurd*, 2nd edn. Harcourt Brace, New York

Lyotard J-F 1984 *The Postmodern Condition: A Report on Knowledge*. [Trans. Bennington G, Massumi B] University of Minnesota Press, Minneapolis, PA

Marcus G E, Fischer M M J 1986 *Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences*. University of Chicago Press, Chicago

Mead H G 1934 *Mind, Self and Society*. University of Chicago Press, Chicago

Prus R 1996 *Symbolic Interaction and Ethnographic Research*. State University of New York Press, New York

Reynolds L 1993 *Interactionism: Exposition and Critique*. General Hall, Dix Hills, New York

Richardson L 1997 *Fields of Play: Constructing an Academic Life*. Rutgers University Press, New Brunswick, NJ

Sanders C 1995 Stranger than Fiction: Insights and pitfalls in post-modern ethnography. *Studies in Symbolic Interaction* **17**: 89–104

Schmidt E P (ed.) 1937 *Man and Society*. Prentice-Hall, New York

Seidman I E 1991 *Interviewing as Qualitative Research*. Teachers College Press, New York

Shalin D 1993 Modernity, postmodernism and pragmatist inquiry: An introduction. *Symbolic Interaction*. **16**: 303–32

Thomas W I, Znaniecki F 1918–20 *The Polish Peasant in Europe and America*. University of Chicago Press, Chicago, Vols.1–5

Thrasher F M 1927 *The Gang: A Study of 1,313 Gangs in Chicago*. University of Chicago Press, Chicago

Wirth L 1928 *The Ghetto*. University of Chicago Press, Chicago

Znaniecki F 1928 Social research in criminology. *Sociology and Social Research* **12**: 307–22

Zorbaugh H W 1929 *The Gold Coast and the Slum*. University of Chicago Press, Chicago

A. Fontana

Symbolism in Anthropology

Symbolism, as it is known and practised in anthropology, is the positive side of a negative effect called 'the subject,' or 'subjectivity.' The symbol and its symbolism live a life of their own, even in the efforts to determine what they may be or mean, so that the whole meaningful life of our species is compounded within their parameters. It is because it is difficult or well-nigh impossible to separate the life of symbols from our own that the social sciences have come into existence. What we identify as 'symbolic' is a recollective fantasy or educated guess as to what may have been going on as it was being formulated, and the lived or thought continuity it presupposes is a cumulative reprojection, or collective daydream, compounded of such educated guesses. It is the scholar's or the writer's way of dealing with 'the emotions.'

The secret of this has very much to do with 'yes' and 'no.' For it was Gregory Bateson who demonstrated that analogy is not exclusive to human beings, that the dog that wishes to be friendly will deliver 'the bite that is not a bite,' affirming positive intention by not

closing its teeth on one's hand. We, it seems, invented 'no,' or at least 'delivered' it. The simplest version is a joke: 'Hurt me!' said the masochist. 'No,' said the sadist, smiling in his teeth. Ever since, we have had the subject to contend with.

'Hurt me!' said the phenomenologist; 'No,' said the anthropologist. Symbolism gives us a deniability of the spoken word in the action of writing or reading it, and a deniability of the written word in the action of speaking from it. We know from ourselves it is that way, and we are a very quick study.

In its most radical determination, anthropological symbolism is concerned with the human ability to reflect ideas, emotions, and perceptions in representational form, hence to reflect upon itself. As an analyzable artifact of cultural or psychological experience, symbolism renders conscious what might otherwise be thought to be unconscious or subconscious. As a totalizing perspective, integrating or at least classifying the whole of a personal, interpersonal, cultural, and social conceptualization of things, symbolism essays the whole compass of the concept. It is world-in-the-person and person-in-the-world. 'DENY ME, CRUCIFY ME' said God. 'NO,' said the being, smiling in its teeth, and just now suddenly HUMAN.

The problems of psychology pertain to the individual person or personality, those of sociology involve a collective or consensual entity, and anthropology defines itself through the differences and similarities that these two polarities share in common. As a conjoint therapy, a medium through which each of these functions is determined by means of the other, symbolism takes on a life of its own, becomes its own mythology about itself.

That might help to explain why anthropologists like Edward Sapir, Alfred Kroeber, Gregory Bateson, Margaret Mead, and especially Ruth Benedict, took such a deep interest in psychological problems. And also why psychologists beginning with Freud, Jung, and Adler, tried to capture anthropology within the net of their all-too-specific visions. It is as though the medium of symbolisation, a self-reflexive agency come into its own, simultaneously divided the two approaches and assimilated each to the other. Or as though the very medium of interpretation stole its message back from those who sought to interpret or dissect it.

In the fine phrasing of the Urapamin people of Papua New Guinea, according to Joel Robbins, 'the eye steals it and eats it.' Interpretation is a thief, and not necessarily a visual one. Another anthropologist, F. E. Williams, was told in Papua that 'The sky has no mouth, the sea has no mouth; only old men have mouths.' Just how much of hermeneutical theory was made in envy of the spoken word? The favorite form of Melanesian immortality is what they call 'the talk that never dies.'

Treated formalistically, subjected to its own definition of itself, it becomes, instead, the symbolism that never lives. The structuralism and ethnosemantics of the 1960s shifted the interpretative message into the symbols, signs, and markers themselves, grounded in schemes for the classification of things, and ultimately the classification of classification itself. The economy or self-condensing advantage of the structuralist shift in emphasis is easily misunderstood in that structuralism and semiotics do not deny the reflexive (e.g., 'psychological') ramifications of symbolism so much as they incorporate them within their analytic designs. The so-called 'poststructuralist' or 'postmodern' approaches have basically nothing to add to this, for in 'unpacking' the 'text,' pretending lived social reality as an imaginary diagram or document, one is subversively packing it.

An alternative approach was suggested in the writings of Sir Edmund Leach and Mary Douglas in which the significative and self-classifying properties of symbolization became secondary to the liminal categorization of properties developed from their usage. It was the anomalies, ambivalences, and ambiguities encountered in classification, and thus only indirectly classification itself, that motivated, and so defined, symbolism. (Thus the consumption of pork would have been proscribed in the ancient Middle East because the pig was an anomalous creature, one that 'divideth the hoof but cheweth not the cud.') The result was a theory of symbolism that was forced to deny its subject matter as a means of affirming it. Victor W. Turner noted that the possibility of an 'anti-structure,' as he called it, is a necessary concomitant of anything that could be called structural. ('Deny the liminal, the reality between the words,' challenged Turner. 'No,' said the ardent symbolist, smiling in his teeth.)

'What is the representational or reflectional status of a symbolism that compromises its own ability to mean or represent, and so exists through the riddle it makes of its own existence?' Essentially all subsequent approaches to symbolism in anthropology devolved from the point of that impasse, offering alternative solutions to the problem by denying or qualifying certain of its drawbacks. One might, for instance, argue that there is no definable symbolism, but only an evolutionary or historical process, thereby postponing the subject of the enquiry indefinitely. In that case the symbolist would have to be what Claude Levi-Strauss called a *bricoleur*, a crafty opportunist anxiously fitting together what the inevitable process was casually taking apart. *Bricolage* would be a form of the Hegelian or Marxian dialectic, but, alas, the dialectic would also be a form of *bricolage*.

Another alternative was posed in the American conception of a 'cultural system,' which has its only unity in a sort of either/or proposition. Either symbolism is metaphorical to its own usage, an open invitation to what Clifford Geertz has called its 'understanding,' or it is not. In his classic study of American kinship, David M. Schneider defined the

symbol as that which stands for what it is not, and it was but a step from this to his later conclusion that 'kinship' simply does not exist. ('Deny *relationship*,' said the anthropologist, 'it is only a function of the *person*.' 'No,' said the informant, smiling in his teeth.)

Is misunderstanding simply a more arcane version of what Geertz and the phenomenologists have called 'understanding?' It would be impossible to form substantive conclusions about anthropological symbolism and what it means without in some way participating in the very effect one was trying to cast into relief. It would likewise be very difficult to live within or experience a kind of 'construction of reality' without in some way penetrating or disqualifying the facade it is said to impose. Virtually any example of symbolism one could know or imagine would have to correspond with some version of the anti-symbolic insight, and thus give the lie to Max Weber's famous dictum about human beings 'suspended in webs of meaning that they themselves have created.' Should Weber have added '… or uncreated,' and smiled in his teeth?

The thing that stands for what it is not is also not what it stands for. The prescient quality of the symbol, and of the symbolism on which it is founded, has just that little to do with content and form, with the kinds or varieties of its usage and recognizable categorization. The lesson in this, which has whole semiotic literatures at its beck and call, is not simply moronic, but actually oxymoronic, like a breath of fresh air. It is more clearly analogous to the lyrics of an American 'country' song: 'Just what part of "no" don't you understand?' Is the metaphor really a metonym in disguise, or is it the other way around; is the 'syntagmatic chain,' a code phrase for the ways convention strings symbols together, a sort of learner's permit for the paradigmatic, the illusion of all the meanings coming together at once? 'There are no easy horses,' says another country song, 'but you've got to learn to ride.' 'Americans,' said the cowboy–humorist Will Rogers, 'are lucky enough not to get all the government they pay for.' Are symbolists even luckier?

The symbolism, or theory of symbolism, that does not partake of the emotional quality it engenders is just another sign, a part of 'no' that, properly misunderstood, might well become a 'yes.' We affirm our denials symbolically, as they say. If we can have no emotions that are not part of the process of determining what they themselves might be, then symbol and emotion are like siamese twins, joined at the hip. They are in a physical relationship; no kind of therapy or 'talking cure' will be able to cut them apart, but only, if lucky, make them feel better about their plight. Denial is also emotional, very strongly and suggestively so, but writing a theory of denials would be like photocopying the air instead of breathing it.

A large part of the 'symbolic' problem seems to be that the predominant mental ethos in what we like to pretend to ourselves as a 'modern world civilization' has excused itself from the insights of the medieval nominalist philosophers. These were the gadflies who insisted, against the worship of essences, the Platonic Realism of their own version of civilization, that there is nothing in what we know save in the names we give to it. We know the names of the stars and their constituent elements and processes. We have given very fancy names to the molecules that make us up, and to their processes and transformations, and are lucky enough not to get all the ones we pay for. But, according to Peter Abelard, we would know very little else.

'The *named* is the mother of the myriad creatures' said Lao Tsu. Phenomenalism, epistemology, and, of course, symbolism, have a very ancient reading in the texts of Aristotle, the Arabic and Jewish, Indian, and Chinese philosophers, and even Heraclitus. But it would seem that the medieval problem of symbolism began with the assertion, by Berengar of Tours, that the Holy Christian Sacrament is but a symbol. He may have smiled in his teeth but was very nearly killed for this. He meant that the accidental properties of things (including the names we give them), and not their essences or substances (whatever those might be) make up the whole reality by which we know them. Thanks to his protector, Fulk the Black of Anjou, Berengar lived to a ripe old age. But he committed symbolic suicide for the medieval church.

Dying by design may be the first and the last thought of the would-be suicide, and of the architect of gothic cathedrals. It was no accident that Berengar's insight happened in medieval times, and provoked the theologies of sacramental doctrine into being. It was no accident either that the subsequent scientific revolution built virtual gothic cathedrals of the accidental properties of things, the matter of 'substance' having gone to subatomic particles, energies, uncertainty-principles or worse, committing real homicide instead of symbolic suicide. 'Understand me,' said Berengar. 'No,' said modern world civilization, smiling in its nuclear fallout. Lucky enough not to get all the Berengar it has paid for.

Not surprisingly, it was a twentieth-century philosopher, Ludwig Wittgenstein, who completed Berengar's 'sentence.' The sacrament, or anything else, is a symbol because 'We picture facts to ourselves.' In other words, important as understanding or even interpretation may be, the essential fact is what Wittgenstein called 'making *sense* of things.' This means that sense is a total phenomenon, and that reason, logic, and even the *overstanding* of a problem that we call an 'explanation,' are its derivative qualities. The whole perspective is in the picture, and no abstraction or ideal quality, however profound, is ever-more useful or effective than the sense that models its meanings for us.

More to the point of this definition, sense gives the gift of deniability to logic and rational ordering. This

would have to mean that all of Einstein's conclusions about general and special relativity were beholden to the 'thought experiments,' through which he imagined them to himself, the 'picture facts.' It would mean that the chemistry of DNA is the micromanagement of Watson and Crick's 'double helix' model, and that all subsequent determinations of subatomic structure are footnotes to the initial mistake of the Bohr atom.

It means that the fossil itself is a cartoon of the reconstructions that we make of it and call a 'dinosaur,' and that everything that we may remember about the past lies in the future of the present moment, that anything we may know of culture or through our own version of it, is a by-product of the initial invention of the concept. We make a solid object of its sense, we 'make book on it.' Imaginal deniability suggests that 'culture' and symbolization enjoy a positive (e.g., 'double negative') relation to each other, that each lives a life of its own within the shadow of the other. 'Deny me, dissect me, disqualify me,' said the culture. 'No,' said the symbol, smiling in its teeth.

The English reformer John Wycliffe suggested that neither substance nor accident have very much to do with Berengar's symbolic sacrament, that what the partaker is after is an 'indwelling of Christ in the soul,' a better sense of Christ's sacrifice, or denial of his life, in the body. The self-denial in Christ's sacrifice is essential to its meaning, or self-affirmation. Likewise, metaphor, the pictorial essence of sense in the symbol, is language's way of falling on its sword. And of liking it.

If what we call 'consciousness' is the mind's way of surprising itself, then symbolism is the reaction it has when it will not take 'no' for an answer. No surprise/no consciousness; no consciousness/no symbol. Everything in the world of letter and number, and even the picture, has the quality of marks, usually in black and white. Everything else has the quality of surprise. Then consciousness itself has no other birth, and no death, and no other phenomenology either that would not come as a surprise. If consciousness begins in pain, as in the joke quoted, then, never mind the change of subject, it would have to finish in satisfaction.

Put more bluntly, this means that if pain is the consequence of taking the human subject too seriously, then symbolism would be the consequence of passing it off as a joke. Then morphine, *pace* Christ, would be the sufferer's best friend, for with morphine 'all the pain is there, I can feel it in my bones, but, insofar as I cannot sense it, I quite honestly don't give a damn,' said the human subject, smiling in its teeth.

We have plenty of subjects in anthropology, and the surprising thing is that we have no valid ethnography about how or why the subject should take itself seriously, care or not care about its culture. Not even in Colin Turnbull's work on the *Ik*. And this makes it difficult to either care or not care about anthropology itself. We have all too many anthropological studies of the ideological substitute for care or concern, the backlash of symbolization. The target of pity becomes a pathetic (e.g., phenomenological) subject when forced to turn that pity back on itself, and an object or willing tool when it lashes out against its plight. The giver of pain or of pity becomes an empathetic subjector. 'Hurt me,' implores the pathetically cultural masochist. 'I am hurting,' says the ideological sadist, smiling in his teeth. Symbolism does not come from inspired self-rejection; it goes there.

Symbolism does not merely encourage personal solipsism; it positively enjoins it, and forces the solipsist to make the necessary compromises or adjustments in attitude. 'Every man his own priest,' was the motto of the Reformation; 'every person their own guru, world-renouncer, and inventor of the world,' sings the modern-day choir in response, 'we are all hurting.' The quality of mercy is strained through the sieve of cognition. Symbolism may or may not be the live wire, nature's way of short-circuiting the human race, but it is definitely culture's way of short-circuiting anthropology.

Symbolism is the edged weapon, the most ancient, *pace* language, part of our technology; it is the distinction that cuts the person as the person cuts the distinction. Without the cutting tool, linguistic or whatever, we could not establish figure against ground, or ground against figure; we would not be grounded at all for want of knowing the difference between the two.

Without the essential tension between affirmation and denial, what the symbolism 'stands for' and how it does not do so, the yes-and-no that Peter Abelard called *sic et non*, its significance would indeed be a mere construction. The so-called 'interpretation' merely dramatizes, and the so-called 'construction' deconstructs, an effect that is in fact causal to their very possibilities, like the trick that 'emotion' plays upon our feelings and sensibilities. ('How do you *feel* about yourself? Tell me about it.') Symbolism is both affirmation and denial at the same time, and the tension between its alternative propositions is the time-space, what Mikhail Bakhtin called the 'chronotope,' of our lives. It is the conscience, the judgment, the question and the answer, the argument and the satisfaction.

In effect, symbolism makes the person necessary in a way that biology or biological evolution could never do by itself. It sustains an adversarial counterpoint without which the fictions of individual and collective, conviction and apathy or their moral and ethical surrogates, would be the empty idealities or meaningless categorizations that marry contemporary political ideologies to social science thinking. How, then, might one propose? Smiling, in one's teeth?

See also: Genealogy in Anthropology; Liminality; Mental Imagery, Psychology of; Metaphor and its Role in Social Thought: History of the Concept; Myth

in Religion; Ritual; Ritual and Symbolism, Archaeology of; Symbolic Boundaries: Overview; Symbolism (Religious) and Icon

Bibliography

Bateson G 1972 *Steps To an Ecology of Mind*. Chandler, San Francisco
Douglas M 1966 *Purity and Danger*. Routledge and Kegan Paul, London
Geertz C 1973 *The Interpretation of Cultures*. Basic Books, New York
Leach E 1976 *Culture and Communication*. Cambridge University Press, Cambridge, UK
Levi-Strauss C 1966 *The Savage Mind*. Weidenfeld and Nicolson, London
Schneider D M 1980 *American Kinship*. University of Chicago Press, Chicago
Turnbull C M 1972 *The Mountain People*. Simon and Schuster, New York
Turner V W 1969 *The Ritual Process*. Aldine, Chicago
Wagner R 1986 *Symbols that Stand for Themselves*. University of Chicago Press, Chicago

R. Wagner

Symbolism (Religious) and Icon

1. Religious Symbols: Their Nature and Functions

1.1 Symbols, Transcendence, and Salvation

Religious symbols arise out of the tension which exists between consciousness and the world. Indeed, the *raison d'être* of religious symbolism is to transcend that tension and move beyond, into 'the other,' into the world of transcendence. Religious symbols in other words are bridges between consciousness, the world, and the sacred. They are not just boundaries between the sacred and the profane, they are also means and processes of their inversion and transformation. Through symbols the material becomes spiritual and the spiritual becomes empirical and is communicated in visible form. As Eliade (1959, p. 12) put it: 'By manifesting the sacred any object becomes *something* else, yet it continues to remain *itself*, for it continues to participate in the surrounding cosmic milieu.' To transcend the world is to give it meaning at another level and attempt to transform it through symbols. A symbol, as Geertz (1975, p. 91) defines it, may be 'any object, act, event, quality, or relation which serves as a vehicle for a conception.' For Geertz, as for Sperber (1975), to symbolize is to conceptualize. Religious symbols, thus, serve as vehicles of conceptualization of this world in terms of another world. Along with conceptualizing the sacred they also serve as means to separate it and conceal it from the profane. In mediating transcendence, symbols function at the same time as vehicles of transformation and stabilization of the world, but the conservative or revolutionary political function of symbols is a matter of degree and depends on political and socioeconomic conditions.

There is an intentional and existential element in religious symbolism which links ultimate reality, social reality, and the quest for salvation which is innate in the human condition and, in one sense or another, underpins all religions. The limitations posed by the human biological, psychological, and social life and above all death, become embodied in religious images which transcend these limitations and give meaning to an otherwise chaotic reality. From this premise Marx, following Feuerbach, and without any theoretical elaboration of the symbolic, produced a reductionist, compensatory theory of religion and understood all religious symbols to be projections and epiphenomena of unfulfilled material needs and limitations imposed by socioeconomic conditions. Freud also produced a reductionist theory of religious symbols in the framework of his psychoanalytic theory. Religious symbolism underwent sophisticated analysis, especially in anthropology during the twentieth century, but its religious meaning *per se* and the problem of salvation were rarely taken on board. Reductionist interpretations of religious symbols have declined in the last decades of the twentieth century but the postmodernists have been criticized for their extreme relativism (Gellner 1992). Weber (1948, p. 280) did not get involved in the analysis of symbols and rituals but he argued that 'the world images that have been created by ideas have, like switchmen, determined the track along which action has been pushed by the dynamic of human interest.' He made these images central to his sociology of religion and explored them as an integral part of human action in the pursuit of salvation. The symbolic representation of salvation then, as pursued within the various religious images of the world in various social and historical contexts, form the general framework of the functions and meanings of religious symbols.

Durkheim (1915 p. 47) described 'sacred things' as 'set apart and forbidden.' Yet in the context of salvation the world of the sacred and that of the secular must be, and are, brought into interaction. Salvation means that chaos, limitation, and disunity are surmountable and God meets humankind. Within the Abrahamic faiths especially, salvation means to bring everything under the unity of God. The forces of evil and death must, and can, be overcome. A bridge between God and humans must be built with the initiative of God and the synergy of humankind. Because of this symbols, as tools of transcendence and salvation, are essentially ambiguous. Their ultimate function is to mediate, reveal, express, and com-

municate that which in essence cannot be fully grasped, spoken of, or communicated. God is both: known and unknown, present and absent, hidden and revealed, God and humans. As Martin (1980, p. 7) points out, 'To transcend is to go beyond. So we must name it and refuse to name it.' What Rudolf Otto (1953) has called *mysterium tremendum et fascionum* cannot be contained in ordinary human knowledge and logic. So the believer's attitude to the sacred is basically one of faith, awe, fascination, and submission.

For the same reason the sacred is not subject to the ordinary rules of formal logic. The nonlogical, the miraculous, and the extraordinary are basic properties of the sacred. Leach (1976, pp. 69–71) calls the logic involved in religious discourse 'mytho-logic' and argues that 'One characteristic feature of such nonlogic (mytho-logic) is that metaphor is treated as metonymy.' Sacred agents can be here, there, and everywhere simultaneously and according to the believer this is not just metaphorical language. It expresses (describes) true properties of divine agents which human agents cannot possess because they are limited, because they are human. Unlike humans, gods can do anything they please, whenever they please. If their doings are to make sense at all they must be in terms of some general scheme of salvation. They are not subject to biological and physical constraints (Christ walking on the water or moving through closed doors). The anthropomorphic, physical features of divine agents are usually represented in exaggerated and distorted form. The Hindu pantheon is characteristic for its multisemic, multiform, and weird symbolic representation. God Siva's commonest representation, for instance, is the 'linga' (the phallus curved on stones) and the sacred bull. Victor Turner (1968, p. 580) describes trickster gods as agents of liminality and antinomian character with ambiguous sexual characteristics. Thus the Greek god Hermes was represented as hermaphrodite and was symbolized by ithyphallic statues and the Yoruba god Eshu is represented with a long headdress curved as a phallus.

Christian symbolism is full of ambiguity and paradox which makes sense primarily in terms of salvation and the Kingdom. God the Son 'begotten' of God the Father, becoming man in Christ, being born of a virgin, dying on the cross, rising from the dead and ascending into heaven, and coming again in glory to judge the quick and the dead: all this is part of the *Nicene Creed* which is the symbol of the Church. The body and blood of Christ, 'eaten' and 'drunk' by Christians, are symbols of reconciliation (redemption) between humans and God. Adam is restored through Christ, the new Adam, who is God-Man (indivisible and without confusion according to the christological doctrine). A dominant, condensed symbol (Turner 1967, pp. 28–30, Douglas 1970, p. 10, Firth 1973, pp. 87–93), which reenacts and represents all that and more, is the Christian liturgy or Eucharist. The point to be stressed here is that religious symbols as units do not make sense in isolation and are systematically related in terms of corresponding schemes of salvation which at the same time correspond to world images and social experience. That may, of course, be above all a theological interpretation of religious symbolism which may be rejected by many sociologists, anthropologists, and psychologists. David Martin (1980), however, has demonstrated that a deep and very sophisticated sociological analysis and interpretation of Christian symbolism can take place by encompassing transcendence in the context of the general soteriological scheme of Christianity. The point to be stressed, in any case, is that without taking seriously the implications for salvation no sociological, anthropological, or psychological interpretation of religious symbolism will be adequate. Indeed, reductionist interpretations may distort the essential anthropological meaning of religious symbols.

1.2 Symbols, Boundaries, Inversion, and Power

Religious symbolism acquired specific cultural significance and potency within what have been called axial civilizations (Eisenstadt 1986). In the monotheistic religions of Judaism, Christianity, and Islam transcendence and unity, and the eschatology implicit in them, became central cultural forces with a cosmic historical dialectic. Transcendence and the pursuit of salvation necessitate boundaries. To go beyond is to move from one boundary to the next, from one set of conditions to another or from one status to another. To be redeemed means to be reborn. The Kingdom of God makes good sense in the context of imperfect, unjust, and corrupt human kingdoms. The heavenly city, where 'there shall be tears no more, no sorrow, no death' (Rev. 21:4), is the opposite of earthly cities where sin, death, suffering, injustice, misery, and tears abide. So the polar oppose of the city of God is the city of Babylon. Within the Kingdom there is unity and divine bliss. Divisions of any kind, injustice, and biological and social boundaries do not apply. Blood bonds and other social bonds as well as divisions of all kinds are characteristic of this fallen world. There are no churches or temples in heaven but the sacred and the profane realms, though separate, are interdependent because the Kingdom, where God shall be all in all, has not come yet. The two kingdoms are separated by ambiguous symbolic boundaries of which paradox and inversion are basic characteristics. Sacred symbols function as sacred markers which challenge or reinforce the powers of this world with a power of their own.

Entering a church, a synagogue, a mosque, or any sacred place one crosses a boundary: the boundary between the world of everyday life and that of another world of sacred, cosmic power and significance. Within the sacred places there are other boundaries

designating a hierarchy of meaning reaching up to the highest, the most sacred and the ineffable. In those specifically sacred parts, the holy of holies, only persons of special sacred status may enter in order to perform specific ritual acts. These acts guarantee that the holiness, the perfection, and the autonomy of the sacred are protected and maintained, while symbolically mediated to the secular world. In the Christian Eucharist the priest carries out the sacrifice of Christ on behalf of Christ, the congregation, and humankind. So hierarchy and status become intertwined from the highest to the lowest levels and transition from one status to another require *rites de passage* (Van Gennep 1960). Such rites symbolize losing the old status and entering a new one and involve what Turner (1969, p. 80) has called a state of liminality, a state 'betwixt and between,' rich in ambiguous symbolism linked 'to death, to being in the womb, to invisibility, to darkness, to bisexuality, to the wilderness … .' In the early church, to be baptized meant to enter the community of the saints, to pass from death to life. Immersion in, and rising out, of the water stood for purification, death, and resurrection. This is symbolized by stripping off the old clothes and putting on new white robes. The color white has transcultural symbolic significance linked to life and to light, as black signifies darkness and death.

Martin (1980), Leach (1976), Douglas (1970), Turner (1969), and others have shown that inversion and status reversal are central mechanisms in religious symbolism. Christ, the Lord and Master of all, washes the feet of the disciples. Humility and hierarchy go together, as the irreconcilable tension between hierarchy and its total abolition are endemic in Christianity. Biological and social reproduction and family and social bonds are counteracted by the concept of virginity and the forming of celibate brotherhoods and sisterhoods. Being born again means entering spiritually and physically the community of the elect. 'Those who will lose their life will gain it.' God defeats death through death. Such inversions and logical contradictions form the substance of religious symbolism.

Symbolic boundaries link the spiritual power that 'moves mountains' and the powers of this world which permeate the human condition. Again, in most religions, this is through paradox and inversion. For Christianity, 'the infinite power of the Almighty' is demonstrated by Christ's suffering on the cross and by icons like that of the Pantocrator in Daphni outside Athens. A thin and weak man like Ghandi could challenge the British Empire by advocating nonviolence. Mary Douglas (1970, pp. 37–53) mentions the 'bog Irish' in London and the Apocryphal characters of old Eleazar in ancient Israel (2 Macc. 6:18) as exercising symbolic religious power against secular power, one by abstaining from meat on Friday and the other by refusing to eat pork on pain of death. The ultimate clash between the powers of this world and the power of heaven and glorious and total victory of the latter over the former is depicted most colourfully in the book of Revelation. Yet for Christianity, as for all religions, the tension between the two kingdoms is always present because the Kingdom of God is yet to come. There is a specific kind of formidable power which is generated through martyrdom which the powers and the rulers of this world seem to fear most. Millenarists down the centuries have celebrated the imminence of the Apocalypse by burning their homes and possessions, by weird behavior, by noise and ecstasy or by total silence, or, paradoxically, by adopting rigid absolute rules in their effort to reject all rules. But the Apocalypse remains in the future and nonviolent religious movements like early Christianity became in time powerful Churches embodying both sacred and secular power. However, as long as the Kingdom is not fully and finally here, religious symbols will continue to reveal and conceal the sacred, and their eschatological pull will continue to move the world.

2. *Icon*

Icons represent the sacred in condensed pictorial form. In iconography the holy is grasped not by reasoning but by feeling. As Langer (1980, p. 15) argues 'the thing we do with images is to envisage a story.' Icons speak to the heart and according to St John of Damascus 'they are to the eye what speech is to the ear' (Uspenski 1976, p. 9). Religious images engage the emotions and express and communicate the sacred directly. Iconography, in most religions, is anthropomorphic and this is true of 'primitive,' Indian, Egyptian, Hindu, Buddhist, Greco-Roman as it is of Christian religion. Central feature of icons is the human body although in certain religions it is presented in a distorted form. In Hinduism, for instance, God *Visnu* is invariably represented with four hands which symbolize his various attributes. In Judaism, the making of images of God or idols are strictly prohibited (Ex. 20:4, Dt. 4:15–18, Dt. 27:15) so iconography did not develop. In Islam, although the Qur'an does not prohibit divine images, pictorial representations were forbidden from the beginning so abstract, geometric, decorative designs and calligraphy developed instead (Schimmel 1987, p. 64).

In a restricted sense, icons refer to the specific form of iconography which developed within Eastern Christianity during the Byzantine period. They constitute the almost exclusive type of religious art in the Orthodox Church to the present day as frescoes, mosaics but mostly painted on wood (Sendler 1988, Kokosalakis 1995). Icons developed out of the early Christian iconography of the catacombs which, in large part, conveyed coded messages of basic Christian themes. Christ, for instance, was represented as fish, which in Greek is ΙΧΘΥΣ, which is an acronym for:

Ιησούς, Χριστός, Θεού, Υιός, Σωτήρ (Jesus Christ Son of God Savior). By the fourth century icons could be found in most Christian communities and constituted a central part of Christian teaching parallel to biblical texts. The formalization of the Christian doctrine during the fourth and fifth centuries consolidated with it the basic themes of Christian iconography: Christ, the Virgin, the saints, and various biblical themes. The representation of Christ in particular went through a long evolution and for certain icons, especially of the Virgin, it was believed that they were not made by human hand (αχειροποίητοι).

By the end of the sixth century icons occupied a central place in collective and private worship in Byzantium but their theological meaning became explicit during the iconoclastic period (AD 726–843) when Byzantine society was shaken to its foundations. Devotion to icons had led to excesses and Emperor Leo III, supported by some bishops and certain heretical groups who rejected external worship, issued the first decree against icons in AD 726. There followed a long period of theological controversy and social unrest which, apart from religious, had also serious political and economic etiology. Icons were finally and decisively reinstated in AD 843 by Empress Theodora and since then have occupied a central place in Orthodox theology and worship.

Icons in the Orthodox Church are religious symbols *par excellence* with central cultural significance. They are related to neoplatonic ideas about the expression of spiritual reality in matter with a substantive religious meaning around the concept of the person/image which refers directly to both the creation of humans in the image of God (κατ' εἰκόνα) and their salvation through the incarnation of the word of God. Icons embody the universal and the particular simultaneously. Some icons are attributed thaumaturgical powers by the faithful. In Eliade's (1959, pp. 24–29) terms such icons constitute 'theophanies.' Orthodox pictorial art adopted the inverted, two-dimensional perspective so that the image enters the viewer instead of the viewer entering the image. For the Orthodox Church, icons are windows of revelation and vehicles of direct communication with divine agents.

See also: Durkheim, Emile (1858–1917); Eliade, Mircea (1907–86); Myth in Religion; Religion, Sociology of; Ritual; Symbolic Boundaries: Overview; Symbolism in Anthropology

Bibliography

Douglas M 1970 *Natural Symbols: Explorations in Cosmology*. Barrie and Rockliff, The Cresset Press, London

Durkheim E 1915 *The Elementary Forms of the Religious Life*. Allen & Unwin, London

Eliade M 1959 *The Sacred and the Profane*. Harcourt, Brace, New York

Eisenstadt S N (ed.) 1986 *The Origins and Diversity of Axial Age Civilizations*. SUNY Press, Albany, NY

Firth R 1973 *Symbols: Public and Private*. Allen & Unwin, London

Geertz C 1975 *The Interpretation of Cultures*. Hutchinson, London

Gellner E 1992 *Postmodernism, Reason and Religion*. Routledge, London

Kokosalakis N 1995 Icons and non-verbal religion in the Orthodox tradition. *Social Compass* **42**(4): 433–49

Langer S 1980 *Philosophy in a New Key: A Study in the Symbolism of Reason, Rite and Art*, Harvard University Press, Cambridge, MA

Leach E 1976 *Culture+Communication: The Logic by which Symbols are Connected*. Cambridge University Press, Cambridge, UK

Martin D 1980 *The Breaking of the Image: A Sociology of Christian Theory and Practice*. Blackwell, Oxford, UK

Otto R 1953 *Das Helige Breslau* [*The Idea of the Holy*. Harvey J (trans.)]. Oxford University Press, London

Schimmel A 1987 Iconography: Iconography Islamic. In: Eliade M (ed.) *Encyclopedia of Religion*. Macmillan, London, pp. 64–71

Sendler E 1988 *The Icon, Image of the Invisible: Elements of Theology, Aesthetics and Technique*. Oakwood Publications, Rotondo Beach, CA

Sperber D 1975 *Rethinking Symbolism*. Cambridge University Press, Cambridge, UK

Turner V 1967 *The Forest of Symbols: Aspects of Ndembu Ritual*. Cornell University Press, Ithaca, NY

Turner V 1968 Myth and symbol. In: Shils D (ed.) *The Encyclopedia of the Social Sciences*. Macmillan, London, Vol. 9, pp. 576–82

Turner V W 1969 *The Ritual Process*. Routledge, London

Uspenski B 1976 *The Semiotics of the Russian Icon*. The Peter Reider Press, Lisse, The Netherlands

Van Gennep A 1960 *The Rites of Passage* [Vizedom M and Caffee G (trans.)]. Routledge, London

Weber M 1946 The social psychology of the world religions. In: Gerth H H, Wright Mills C (eds.) *From Max Weber: Essays In Sociology*. Routledge, London, pp. 267–301

N. Kokosalakis

Symptom Awareness and Interpretation

Physical symptoms often are ambiguous and open for subjective interpretation. Even though people certainly have some awareness of their internal states, that awareness and the accuracy of interpreting internal sensations may be fairly limited (Pennebaker 1982). Many authors state that people are poor at tracking physiological processes. Nevertheless, they hold the belief that they are able to do so. In general, individuals are motivated to maintain an understanding of their bodily condition. Both vague internal states and information about possible diseases lead to interpretive activities. Understanding the processes that guide symptom identification and interpretation

at the initial stage of the occurrence and also during the course of disease has been found to be critical for understanding how people act upon symptoms and disease.

Traditionally, mainly fields such as psychophysiology and the medical sciences dealt with questions related to symptom awareness and interpretation. Pennebaker (1982), in his groundbreaking book, brought the information processing perspective to the area. This approach has proven useful in understanding phenomena of symptom awareness and interpretation. Specifically, Pennebaker proposed that symptoms or sensations are best understood as representing information about internal states. Taking the information processing perspective, this article will emphasize the role of the following aspects: (a) information organization and mental representation of symptoms, (b) knowledge activation and use, and (c) an explanation and attribution of symptoms. Throughout, the behavioral implications of the reviewed research for disease-management, help-seeking behavior, and compliance and adherence to medical regimens will be illustrated taking selected diseases as examples (hypertension and diabetes). Finally, open questions and probable directions of future theory and research will be addressed.

1. *Information Organization: The Mental Representation of Symptoms*

Given the quantity of information that is available to individuals at any given time, it is functional for the organism to organize and reduce the incoming data. This is achieved by cognitively (re)structuring, synthesizing, and organizing the incoming information. Cognitive psychology has introduced several theoretical constructs in this context, namely schemata, cognitive sets, prototypes, expectancies, and scripts. Studies on the structure of illness schemata consistently found five dimensions according to which symptom and illness experiences are organized (Lau and Hartman 1983): illness schemata include information about the identity or label of the disease, its consequences (i.e., the symptoms), its causes, its duration, appropriate treatments, and expectations about its curability. Cognitive structures such as illness schemata or prototypes help to organize information from internal sensations (symptoms) and disease-related information gathered from the external environment.

Certainly, the most relevant model of symptom perception within the tradition of information processing has been proposed by Leventhal et al. (1984). Their approach is rooted in cognitive psychology and builds on the work on cognitive schemata and prototypes (see *Health and Illness: Mental Representations in Different Cultures*). In specific, Leventhal et al. (1984) propose that individuals tend to construct their own individual representation of symptoms or illness and that this idiosyncratic representation will, in turn, influence their behavior (e.g., help-seeking, adherence, and compliance). Specifically, they proposed that illness representations (or schemata) are a function of an individual's semantic knowledge about symptoms and disease and specific contextual factors such as the nature of somatic changes and the situations in which these occur. This semantic knowledge accumulates across the life span and is acquired through the media, through personal experience, and from family and friends who have had experience with the disease. People's common sense models of illness strongly influence which symptoms a person will search for and will ultimately perceive. Work by Meyer et al. (1985) on hypertension illustrates this point. Hypertensives hold one or a combination of disease models about high blood pressure. As Meyer et al. (1985) demonstrated, some patients hold the belief that hypertension is a disease of the heart, others believe that it is an arterial disorder, and a third group might associate hypertension with emotional upset. Importantly, the kind of belief patients hold affects the way they monitor their body: Symptoms that are consistent with their specific illness belief are more likely to get noticed (see *Hypertension: Psychosocial Aspects*). The common sense model of illness approach has also been used successfully to explain such phenomena as placebo effectiveness and mass psychogenic illness (MPI). For example, Pennebaker (1982) stated that placebos are effective because they change the way in which individuals process sensory information. Individuals who are worried about their condition—after taking a (placebo) drug—change their focus from sensations that may signal illness to sensations that signify improved health.

In addition to general concepts of illness, people also hold organized conceptions—termed disease prototypes—for particular diseases (Bishop 1991, Bishop and Converse 1986). For example, a person may have a disease prototype of heart disease. Similar to general illness conceptions, prototypes of specific diseases help people organize and evaluate information about bodily sensations that might otherwise not be interpretable. Thus, a person who holds the belief that he or she is vulnerable to heart disease is more likely to interpret chest pain in accord to his or her prototype of heart disease than a person who does not hold this belief (Bishop and Converse 1986). This latter person might instead regard the chest pain as signaling a gall bladder problem.

Altogether, organized representations of illness-related knowledge and experiences strongly determine symptom awareness and interpretation, and the expectations individuals generate respecting their consequences and causes. This has important implications for health behavior as the kind of beliefs patients hold about symptoms and disease may lead them to alter or fail to adhere to their treatment regimens (e.g., see *Patient Adherence to Health Care Regimens*).

2. Knowledge Activation and Use

2.1 The Role of Knowledge Activation in Symptom Awareness

What directs individuals to become aware of a symptom and how to interpret it? As noted above, individuals use hypotheses and subjective theories (their prototypes) about diseases to interpret initially ambiguous symptoms. Once a plausible hypothesis is identified, information to verify the hypothesis is sought. This is performed by selectively scanning stored knowledge for matching evidence. Given the rather vague nature of many symptoms, confirming evidence often can easily be found.

The accessibility of knowledge stored in memory has been demonstrated to increase when it has been recently primed. Mechanic (1972), for example, documented a common phenomenon in medical schools, the medical students' disease. He reported that about 70 percent of first-year medical students perceive symptoms of those illnesses studied in their courses. Obviously, studying the symptoms leads students to focus on their own internal states and to compare what they hear with their concrete personal experience at the moment. Vague, ambiguous physical sensations are then interpreted in accordance with the disease that was made salient in their course.

Importantly, Leventhal et al. (1980) showed that this process can work both ways. They proposed the symmetry rule, which holds that individuals will (a) seek symptoms when given a diagnostic label (such as in the medical students disease) and (b) seek a diagnostic label when experiencing symptoms, that is, the perception of a change in internal sensation creates a pressure for finding a label. Similarly, being provided with a label of a specific disease generates pressure for a perceptual referent. Schachter and Singer's (1962) classic study can be regarded as evidence for the symptom-to-label part of the symmetry rule since it shows that people will come up with a specific interpretation or label for their somatic sensations that can explain their condition (see below, attribution of symptoms). Bishop's (1991) and Bishop and Converse's (1986) studies on differences in symptom perception as a function of differential beliefs about hypertension mentioned earlier can also be interpreted as reflecting the symptom-to-label part of the symmetry rule.

The second part of the symmetry rule (that people will seek symptoms when given a label) is consistent with findings by Meyer et al. (1985) on hypertension. Hypertension is largely a symptomless disease. However, many patients erroneously believe that they can tell when their blood pressure is high (e.g., they believe their blood pressure to be high when their face gets warm, their heart beat increases, or they are feeling tense) (Baumann and Leventhal 1985). In fact, though, the correlation between beliefs about level of blood pressure and actual blood pressure is low (Meyer et al. 1985). In accord with the label-to-symptom part of the symmetry rule, the authors found that patients in hypertension treatment became more likely to perceive their blood pressure the longer they were in treatment (even though high blood pressure, in fact, is unsymptomatic).

Importantly, illness representations strongly influence the way patients will act upon (subjectively perceived) symptoms. Taking hypertension as an example, patients have been found to treat their condition based on their belief that they are able to detect when their blood pressure is high and when it is not. When they 'feel fine,' they may discontinue taking medication, and when they believe their blood pressure is high they might take it. A similar phenomenon has been documented for the self-management of diabetes. Both hyperglycemia and hypoglycemia produce unpleasant bodily sensations that can be vividly described by patients (Hampson 1997). Diabetes patients often fail to self-monitor their blood glucose level and instead, like hypertension patients, rely on what their blood glucose level 'feels like.' However, the relation between symptom experience and acute blood glucose levels is also often imprecise, and patients are misled by their illness belief. Hence, diabetes patients need to be advised that their illness management should not rely on subjective perceptions of their blood glucose level. Instead, in deciding when to take measures to increase or decrease their blood glucose level, they should rely exclusively on testing their blood glucose (Hampson 1997, see *Diabetes: Psychosocial Aspects*).

2.2 The Role of Knowledge Activation in Symptom Interpretation

What determines which knowledge will be activated and used in order to *interpret* perceived symptoms? The evaluation of symptoms is very subjective and has been found to be influenced by the nature of their attributes. Symptom attributes can include such aspects as how salient (noticeable) and how unexpected they are, or which consequences they may have for one's life. Not all features of a physical sensation receive equal attention. Most likely, symptoms that are visible or cause pain draw most attention. Similarly, symptoms that affect highly valued parts of the body (e.g., the eyes, face, or heart) usually are interpreted as being more serious and as more likely to require attention than symptoms that affect less valued organs. In sum, symptoms that hurt, symptoms that change quickly, and symptoms that are incapacitating are more likely than their opposites to be interpreted as signs of a serious illness and therefore more likely to induce a person to seek medical treatment (Safer et al. 1979).

This selective attention to symptoms can have its costs, however, when symptoms that are not salient

but nevertheless serious remain unnoticed. For example, people with heart attacks often wait for several hours until they seek help (Matthews 1982). As a consequence, 70 percent of all deaths due to heart attacks occur in the first few hours. Matthews noticed that a person will wait longer before seeking help the less salient the first symptoms are, the more depressed and exhausted the person is, and the more effortful the activity is that they are engaged in at the time of symptoms' onset. The latter effect can also be explained by Pennebaker's (1982) *Competition of Cues Model*. In his model, Pennebaker states that external and internal sensory information compete for people's attention. The likelihood with which internal sensory perceptions are perceived then is a function of the proportion of external to internal information. When people are focused externally, for example, in a highly distracting environment or during a vigorous physical activity, they are slower at noticing symptoms. In contrast, when the external environment lacks stimuli, attention is more likely directed toward the body. This increases the likelihood that cues that suggest illness are noticed (Pennebaker 1982). In one study, for example, Pennebaker demonstrated that people are more likely to cough during boring parts of a movie than during more interesting parts. Apparently people focus their attention more on their internal states when there are few external stimuli, and are therefore more likely to notice itching or tickling in their throats to which they then respond with coughing.

Altogether, the awareness and interpretation of internal sensations is a function of the relation between external and internal stimuli. And it is also a function of the beliefs that individuals hold. The beliefs cause them to selectively attend to specific internal information or contextual information, i.e., not all knowledge is used in evaluating internal states.

3. *Explanation and Attribution of Symptoms*

It has been suggested that people are 'common-sense' medical scientists who search for the meaning of somatic events and search to attribute it to a specific cause. Research has been presented which demonstrates that sometimes not the mere presence of symptoms impact on behavioral implications, but rather the attributions made for these symptoms. For example, almost two-thirds of the people who seek medical attention for low-level symptoms have been found to suffer from anxiety or depression instead. This may occur because symptoms of these emotional disorders are mistakenly thought to be indicative of a physical disease. Attributions of symptoms have also been found to affect adherence to medical regimens. For example, in their study on HIV-infected women, Siegel and Gorey (1997) found decreased adherence to medication and even discontinuation of medication intake when symptoms were attributed to side effects of the medicine taken. According to the authors, adherence was reduced because these symptoms were interpreted as evidence that the medication was making them sicker or had more risks than benefits. Similarly, failure of medications to relieve symptoms also negatively influenced adherence because they are interpreted as not having any effects.

3.1 *What Guides Symptom Attribution?*

Valins (1966) was the first to suggest that the mere belief that a change in physiological 'arousal' had occurred would lead individuals to seek for information to label this change. In a recent demonstration, McCaul et al. (1992) found that students who were informed falsely that they were at risk for gum disease reported higher rates of gum bleeding (a symptom of gum disease) 2 days later than students in the control group who were informed that they had no gum disease. Mechanic's notions on the medical students' disease and Schachter and Singer's (1962) study on the attribution of physical arousal mentioned earlier also fit into this conception: When persons have access to a certain sickness label (e.g., because they just read about it or met a sick person) they will be particularly likely to attribute their physical sensations to this disease.

The explanation and attribution of symptoms is, however, also influenced by motivational factors. Specifically, it has been shown to be governed by optimistic tendencies that support a positive view of oneself and one's physiological condition (Weinstein 1980, see *Health Risk Appraisal and Optimistic Bias*). This means that an individual will be less likely to interpret a novel symptom (e.g., bleeding of a mole) as a sign of a serious, life-threatening disease (e.g., skin cancer) and will be more declined to neglect it or to attribute it to a nonserious cause (e.g., a scratch). In general, the optimistic bias results in selectively seeking information that confirms the less threatening hypothesis. Obviously, this benign hypothesizing results in delay to seek help (e.g., see *Illness Behavior and Care Seeking*).

3.2 *The Influence of Mood, Emotions, and Stress*

Mood, emotions, and life stressors may be a further important determinant of symptom attribution and in consequence of care seeking and adherence. People who are in a positive mood report fewer symptoms (Salovey and Birnbaum 1989). People in a bad mood, however, report more symptoms, are more pessimistic that any actions they might take would relieve their symptoms, and perceive themselves as more vulnerable to future illness. In the presence of life stress, ambiguous physical sensations (e.g., fatigue, headaches, or stomachaches) have been found to be

interpreted as signs of stress instead of illness (Cameron et al. 1995). Hereby, stress attributions are more likely for sets of symptoms that are unfamiliar, novel, and ambiguous (Pennebaker 1982).

3.3 Individual Differences in Symptom Attribution

The attribution of perceived symptoms is further determined by both individual and cultural differences. Age and gender as well as ethnic and racial differences in symptom attribution have been documented (e.g., see *Aging and Health in Old Age*; *Culture as a Determinant of Mental Health*; *Health and Illness: Mental Representations in Different Cultures*; *Gender and Physical Health*). For example, identifying initially ambiguous symptoms as signs of a beginning disease is particularly difficult in the elderly. First, older individuals experience more physical complaints than younger persons. They therefore have to evaluate the significance of new symptoms against an increasingly complex background of age related somatic changes (Leventhal and Crouch 1997). Second, distinguishing symptoms that can be attributed to normal aging from symptoms that signal serious disease can be a complex task in old age. Importantly, the 'wait and see' strategy that is widely used at all ages in an attempt to rule out benign causes of ambiguous bodily signals may be more dangerous in old age because of a less vigorous immune protection. Discounting and ignoring treatable symptoms and waiting to seek help can delay diagnosis and treatment of potentially life-threatening disorders and can create unnecessary risk for morbidity and mortality. In sum, older individuals may be at greater risk because they are more likely to make aging attributions and because it has more serious consequences when they delay seeking treatment for a serious disease (Leventhal and Crouch 1997).

4. Remaining Questions

Much progress has been made in the theoretical understanding of the mental representation of symptom sets and diseases and the behavioral implications of these beliefs. Most research in this area has so far focused on just a few diseases: asthma, diabetes, and hypertension. As a consequence, the mental representation of these diseases and their consequences for self-management is, by now, quite well understood. Much more research is needed, however, with regard to the understanding of diseases that until now have received less research attention (e.g., allergies).

Clearly, social and cultural factors shape both the awareness and interpretation of physical symptoms and the behavioral responses chosen for controlling them. A key area for future work is to understand the self-regulatory processes of symptom perception and interpretation within the social and cultural context in which they occur. With a marked increase in the proportion of older people in many countries, research in this area has to move more strongly into the direction of taking a life-span perspective. A better understanding of age-related changes in the awareness and interpretation of symptoms will be vital in the process of optimizing the self-management of chronic diseases. At the same time, with many chronic diseases on the rise among children (e.g., asthma and allergies), more knowledge is needed to understand how they perceive and interpret symptoms and how contextual factors of their life situation can be worked into a better adherence to problematic medical regimens.

See also: Culture as a Determinant of Mental Health; Diabetes: Psychosocial Aspects; Gender and Physical Health; Health and Illness: Mental Representations in Different Cultures; Health Risk Appraisal and Optimistic Bias; Hypertension: Psychosocial Aspects; Illness Behavior and Care Seeking; Patient Adherence to Health Care Regimens; Schemas, Frames, and Scripts in Cognitive Psychology; Stress Management Programs

Bibliography

Baumann L J, Leventhal H 1985 I can tell when my blood pressure is up, can't I? *Health Psychology* **4**: 203–18

Bishop G D 1991 Understanding the understanding of illness: Lay disease representations. In: Skelton J A, Croyle R T (eds.) *Mental Representations in Health and Illness*. Springer-Verlag, New York, pp. 32–59

Bishop G D, Converse S A 1986 Illness representations: A prototype approach. *Health Psychology* **5**: 95–114

Cameron L, Leventhal E A, Leventhal H 1995 Seeking medical care in response to symptoms and life stress. *Psychosomatic Medicine* **57**: 1–11

Hampson S E 1997 Illness representations and the self-management of diabetes. In: Petrie K J, Weinman J A (eds.) *Perceptions of Health and Illness*. Harwood Academic, Amsterdam, pp. 323–47

Lau R R, Hartman K A 1983 Commonsense representations of common illnesses. *Health Psychology* **2**: 167–85

Leventhal E, Crouch M 1997 Are there differences in perceptions of illness across the lifespan?. In: Petrie K J, Weinman J A (eds.) *Perceptions of Health and Illness*. Harwood Academic, Amsterdam, pp. 77–102

Leventhal H, Meyer D, Nerenz D 1980 The common sense representation of illness danger. In: Rachman S (ed.) *Contributions to Medical Psychology*, Vol. 2. Pergamon, New York

Leventhal H, Nerenz D R, Steele D J 1984 Illness representations and coping with health threats. In: Baum A, Singer J (eds.) *A Handbook of Psychology and Health*, Vol. 4. Erlbaum, Hillsdale, NJ, pp. 219–52

Matthews K A 1982 Psychological perspectives on the Type A behavior pattern. *Psychological Bulletin* **91**: 293–323

McCaul K D, Thiesse-Duffy E, Wilson P 1992 Coping with medical diagnosis: The effects of at-risk versus disease labels over time. *Journal of Applied Social Psychology* **22**: 1340–55

Mechanic D 1972 Social psychological factors affecting the presentation of bodily complaints. *New England Journal of Medicine* **286**: 1132–39

Meyer D, Leventhal H, Gutmann M 1985 Common-sense models of illness: The example of hypertension. *Health Psychology* **4**: 115–35
Pennebaker J W 1982 *The Psychology of Physical Symptoms*. Springer-Verlag, New York
Safer M A, Tharps Q J, Jackson T C, Leventhal H 1979 Determinants of three stages of delay in seeking care at a medical care clinic. *Medical Care* **17**: 11–29
Salovey P, Birnbaum D 1989 Influence on mood on health-relevant cognitions. *Journal of Personality and Social Psychology* **57**: 539–51
Schachter S, Singer J E 1962 Cognitive, social and physiological determinants of emotional state. *Psychological Review* **69**: 379–99
Siegel K, Gorey E 1997 HIV infected women: Barriers to AZT use. *Social Science and Medicine* **45**: 15–22
Valins S 1966 Cognitive effects of false heart rate feedback. *Journal of Personality and Social Psychology* **4**: 400–8
Weinstein N D 1980 Unrealistic optimism about future life events. *Journal of Personality and Social Psychology* **39**: 806–20

B. Knäuper

Synapse Formation

Formation of specialized contacts, the so-called synapses, between presynaptic axonal terminals of neurons and their postsynaptic targets, such as neurons and muscle cells, is a decisive step in building the nervous system. It enables rapid formation transfer between contacting cells via spatially and temporally restricted activation of postsynaptic receptors by neurotransmitters secreted from the nerve terminals. Considerable progress in the understanding of this complex phenomenon came from studies of synaptogenesis of the vertebrate neuromuscular junction (Fig. 1A). During the last decade investigations of glutamatergic neuron-neuron contacts in mammalian brains (Fig. 1B) and the neuromuscular junction in *Drosophila* highlighted the diversity of molecular interactions and universality of certain principles underlying formation of different types of synapses. These two issues are addressed in the present article.

1. *Approaching the Target*

Growth cones are guided to their targets by attractive and repulsive cues that can be diffusible or integrated into the cell surface or the extracellular matrix (Mueller 1999). Semaphorins, ephrins and their receptors constitute large groups of guidance molecules that evoke repulsive and adhesive responses, often bidirectionally in the interacting cells. Cell adhesion molecules of the immunoglobulin superfamily promote neurite outgrowth via homophilic and heterophilic interactions with a variety of extracellular matrix and membrane-bound molecules. Recently, the L1 immunoglobulin superfamily molecule has also been found to be involved in ephrin receptor mediated repellent signaling. Netrins compose a family of proteins that can attract or repel axons, acting via distinct netrin receptors. Binding of recognition molecules to receptors activate signal transduction mechanisms, mostly involving receptor and non-receptor tyrosine kinases. These signals have been shown to converge in the regulation of activities of several small guanosine triphosphate (GTP) binding proteins of the Rho subfamily. Two of these proteins, Cdc42 and Rac1, are under the control of attractive guidance cues, whereas Rho is under control of repulsive guidance cues (Hall 1998). Recent studies show that one and same molecule can evoke attractive or repulsive responses depending on the internal state of the cell. For instance, three different extracellular cue molecules—recognition molecule netrin-1 brain-derived neurotrophic factor (BDNF) and neurotransmitter acetylcholine—induce attractive or repellent turning of the growth cones depending on the intracellular level of cyclic adenosine monophosphate (AMP) (Ming et al. 1997).

2. *Selecting the Target*

Selection of the appropriate postsynaptic target and search strategies of axonal growth cones are interrelated processes. Removal of the repulsive cues or their axonal receptors often results in a decrease in axon bundling, so-called defasciculation, indicating that axon–axon interactions are no longer favored over axon–substrate interactions. This enables the fibers to grow into the regions that normally are not invaded and there form ectopic synapses. At the *Drosophila* neuromuscular junction, the immunoglobulin superfamily molecule fasciclin II promotes and semaphorin II inhibits promiscuous synaptogenesis. These molecules are expressed by all muscles in wild-type flies. A more restricted expression is seen for netrin B which is synthesized by a subset of muscles where it attracts some axons and repels others. Underexpression of repellent molecules or overexpression of adhesion-promoting fasciclin II upregulate formation of promiscuous and ectopic synapses. These experiments argue against the view that pre- and postsynaptic partners have unique complementary labels, and show that the relative balance of adhesive and repellent molecules controls the number of the synapses formed (Winberg et al. 1998). This is also demonstrated by genetic deletion of semaphorin IIIa and its receptor, neuropilin, which resulted in defasciculation and abnormal spreading of cranial nerves in mice. Also, more synapses were formed on wild-type hippocampal neurons expressing the neural cell adhesion molecule (NCAM), the closest mammalian homolog of fasciclin II, when compared to

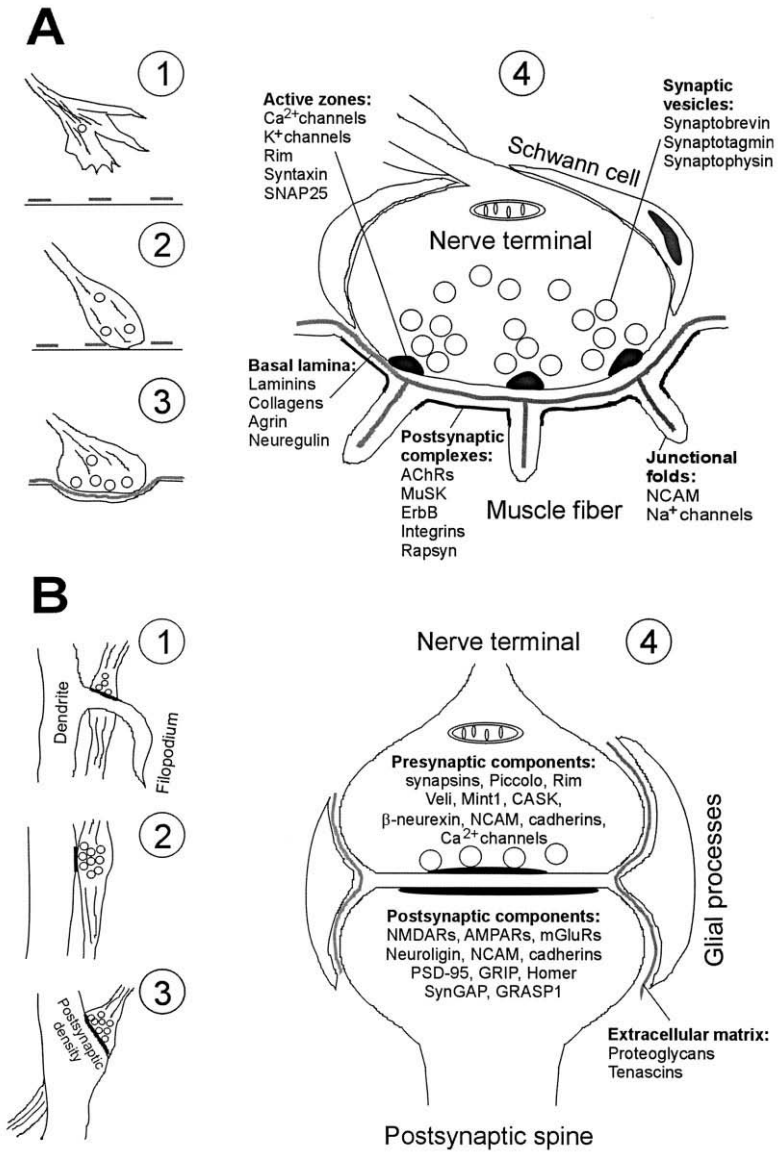

Figure 1
Formation and differentiation of the vertebrate neuromuscular junction (A) and excitatory neuron–neuron connection (B). The numbers 1–4 refer to four consecutive stages of maturation, during which accumulation of synaptic vesicles in presynaptic active zones and clustering of postsynaptic receptors occur. For the neuromuscular junction (A), after the growth cone (stage 1) establishes contact with a muscle fiber, it is a simple bulbous enlargement (stage 2). During further maturation, extracellular material is accumulated in the synaptic cleft forming the basal lamina, multiple junctional folds are formed, and nerve terminals are coated by glial processes (stages 3–4). For the neuron–neuron synapse (B), induction of a synapse by dendritic filopodium (stage 1) is followed by retraction of the filopodium and formation of the shaft synapse (stage 2) that can be further transformed into stubby (stage 3) and to spine synapse (stage 4). For mature synapses (stage 4) the localization of some molecules is shown. Indicated in (A) synaptic vesicle and active zone proteins are common for these types of synapses. A more detailed description of protein functions and involved molecular interactions is given in the text

neighboring neurons from NCAM-deficient mice maintained in co-cultures. However, NCAM-deficient neurons maintained in a no-choice situation— separately from NCAM-expressing cells—receive normal numbers of contacts. This implies that NCAM mediates one of several signals which are integrated by the developing axon to confer preference for contact formation at sites where the integrated signal is optimal (Dityatev et al. 2000).

Cooperative actions of some N-acetyl-galactosamine-terminated glycoconjugates, N-cadherin and BDNF are known to regulate formation of synapses of retinal axons in the chick optic tectum. Lamina-specific connections in the hippocampus illustrate that synaptogenic cues can be expressed by 'guidepost' cells that are not definitive synaptic targets but transient targets for growing axons. Such 'guidepost' cells, for instance, are Cajal–Retzius cells which occupy the *stratum lacunosum-moleculare* and calbindin-positive interneurons occupying the *stratum radiatum* of the hippocampus. During early development, entorhinal axons form synapses onto Cajal–Retzius cells before the dendrites of pyramidal cells reach the molecular layer. Calbindin-positive interneurons are preferred targets for commissural axons. Later, most of guidepost cells die and synapses are transferred to the pyramidal dendrites (Sanes and Yamagata 1999).

3. *Exchanging Signals*

When an axon approaches its postsynaptic target, it releases anterograde signals and receives retrograde influences from the postsynaptic side. As discussed below, anterograde signals can induce clustering of postsynaptic receptors at synaptic sites and down-regulation of receptor expression at nonsynaptic sites. Also, when neonatal muscles, destined to become fast- or slow-twitch, are denervated and cross-reinnervated with each other's nerves, they acquire contractile characteristics appropriate for their new innervation. Such 'conversion' of postsynaptic cells is mediated by acetylcholine release and electrical activity inducing changes in expression of proteins encoding fiber type-specific isoforms of contractile proteins.

Retrograde signaling plays an important role in elimination of incorrect or excessive synaptic inputs. Most vertebrate muscle fibers are initially innervated by more than one motor axon, retaining only a single input during early postnatal life. Similarly, projections from both eyes initially overlap in the lateral geniculate nucleus, with eye-specific laminae forming from the initially diffuse projections late in embryogenesis. These processes appear to depend on postsynaptic electrical activity. First, competition and selection of afferents could then arise due to limited amounts of maintenance factors released from postsynaptic cells and taken up by axons in an activity-dependent manner. Nitric oxide and neurotrophic factors, like BDNF, nerve growth factor (NGF) and fibroblast growth factor 2 (FGF2), are among the favorite candidate molecules mediating retrograde strengthening of synapses. Second, electrical activity can evoke release of proteases such as thrombin or plasminogen activator, destabilizing cell interactions at the nerve terminal unless they are protected by endogenous inhibitors. Interestingly, the heparan sulfate proteoglycan agrin, implicated in clustering of nictotinic acetylcholine receptors (AChRs) at the neuromuscular junction, may function also as a protease inhibitor. Third, postsynaptic activity can redistribute the localization of adhesion molecules leading to retraction of axons and loss of synapses via downregulation of adhesive interactions in inappropriate sites, whereas synaptic strengthening should involve accumulation of adhesion molecules, such as NCAM, or changes in their molecular configuration as has been shown for cadherins, which dimerize upon depolarization. Such dimers are protected from proteolytic degradation and can form transsynaptic associations (Tanaka et al. 2000).

4. *Postsynaptic Differentiation*

Maturation of synaptic contacts leads to the accumulation of transmitter receptors at postsynaptic sites. If AChRs are uniformly distributed in immature muscles at a density of $1000\,\mu m^{-2}$, their expression in the mature muscles exceeds $10000\,\mu m^{-2}$ at the synapse, whereas the density drops to $10\,\mu m^{-2}$ extrasynaptically. Many molecules have been reported to promote clustering of AChRs *in vitro*. Of these, only one has so far been implicated in synaptogenesis *in vivo*. This is the nerve-derived alternatively spliced isoform of agrin that interacts with the transmembrane protein tyrosine kinase MuSK. A crucial effector of receptor clustering downstream of MuSK is the cytoplasmic protein rapsyn that bears distinct domains responsible for association with the membrane, multimerization, and interaction with receptors. Functionally similar synapse-associated proteins were discovered in neuronal synapses (Garner et al. 2000). They bear protein motifs called PDZ domains that can bind ligand- or voltage-gated ion channels and adhesion molecules. In *Drosophila*, discs-large protein interacts with and determines synaptic localization of fasciclin II and Shaker type of voltage-gated K^+ channels. In vertebrates, the PDZ domain-containing postsynaptic density protein (PSD-95), glutamate receptor interacting protein (GRIP), and Homer interact with N-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-isoxazole-4-propionic acid (AMPA) and metabotropic glutamate receptors, respectively. PSD-95 can also interact with voltage-gated K^+ channels, the synaptic cell adhesion molecule neuroligin, and kainate receptors, whereas GRIP interacts with ephrin B ligands. These complexes

are linked to the postsynaptic cytoskeletal matrix and signaling cascades. For example, PSD-95 and GRIP are likely to couple glutamate receptors to the Ras GTPase-activating protein SynGAP and GDP-GTP exchange factor GRASP1, respectively.

Differential expression of transmitter receptors in postsynaptic sites at the vertebrate neuromuscular junction is achieved not only by clustering of diffusely distributed receptors, but also by transcriptional activation of AChR genes in the muscle in subsynaptic nuclei, and transcriptional repression of AChR expression in nonsynaptic nuclei. Repression of synthesis is regulated by postsynaptic activity, inducing entry of Ca^{2+} and activation of protein kinase C. Critical targets of this kinase are transcriptional factors (myoD, myf5, MRF4, and myogenin) that down-regulate transcription of AChRs by binding to the so-called E-box containing sequences in several AChR subunit genes. These repressive signals can be counteracted at synaptic sites by the nerve-derived neuregulin that is incorporated into synaptic basal lamina and activates synthesis of AChRs in myocytes. Neuregulin acts via the epidermal growth-factor-related receptor tyrosine kinase and downstream cascades of effector molecules, including Ras, Raf, and Erk kinases. These effectors regulate the activity of transcription factors of the ets family, which can bind to N-box elements in some AChR subunit genes. Neuregulins also trigger synthesis of some receptors in the central nervous system and some proteins are likely to be synthesized locally at synaptic sites in dendrites.

5. *Presynaptic Differentiation*

Spontaneous and evoked neuromuscular transmission has been seen to begin within minutes after an axon contacts the postsynaptic cell. However, the synaptic activity is initially low, not only because of the small number of postsynaptic receptors, but also because only few vesicles are released. During maturation of a synapse, synaptic vesicles increase in number and become clustered at active zones, whereas some cytoskeletal elements characteristic of the growing axon are lost. These changes are accompanied by an increase in synaptic volume that leads to a prominent increase in synaptic activity. The factors that induce presynaptic differentiation are unknown, but several molecules, which can be retrogradely released or are present in extracellular matrix, have been implicated in stimulating clustering and release of synaptic vesicles. They include neurotrophic factors, BDNF and FGF2, and extracellular matrix molecules such as agrin and synaptic isoforms of laminin. As for the postsynaptic site, presynaptic differentiation involves redistribution of synaptic machinery, and changes in expression of different isoforms of synaptic proteins. Accumulation and mobilization of vesicles is mediated by synapsins and the PDZ domain-containing proteins Piccolo and Rim. Possibly, assembly of pre- and postsynaptic components in neuron–neuron connections depends on formation of transsynaptic molecular complexes. Thus, postsynaptically localized neuroligin, associated with PSD-95, can bind to presynaptically localized β-neurexin associated via CASK with two PDZ domain-containing proteins, Veli and Mint1. The latter molecule interacts with N-type Ca^{2+} channels that are responsible for the entry of Ca^{2+} and transmitter release at the nerve terminals.

6. *Synaptogenesis as a 'Never Ending Story'*

Formation and loss of synapses continue during the whole lifespan of animals. It is likely that high cognitive functions, such as learning and memory, utilize the mechanisms evolved to wire the brain during early development. Indeed, many molecules involved in early synaptogenesis, such as cell adhesion molecules of the immunoglobulin superfamily, integrins, cadherins, and ephrin receptors, have proved to play important roles in long-term potentiation and depression of synaptic connections, underlying learning and memory in adults (see *Long-term Depression (Cerebellum)*; *Long-term Depression (Hippocampus)*; *Long-term Potentiation (Hippocampus)*; *Memory: Synaptic Mechanisms*). However, synaptogenesis in the adult brain occurs in the context of a very different milieu than that of early development. Mature synapses are subject to more severe constraints for further modifications, due to reduced NMDA receptor-mediated currents, and well-developed inhibitory circuitry. Additional restrictions are imposed by the established adhesive interactions within and outside of synapses, and a firm extracellular matrix enriched in repellent signals, thus possibly reducing structural plasticity. It remains to be seen which mechanisms active during development contribute to the formation of synapses in the adult brain and which mechanisms were additionally 'invented' to provide stability, but also flexibility in synaptic activity underlying learning and memory.

See also: Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Recovery of Function: Dependency on Age; Synapse Ultrastructure; Synaptic Efficacy, Regulation of; Synaptic Transmission

Bibliography

Broadie K 1998 Forward and reverse genetic approaches to synaptogenesis. *Current Opinion on Neurobiology* **8**: 128–8

Dityatev A, Dityateva G, Schachner M 2000 Synaptic strength as a function of post- versus presynaptic expression of the neural cell adhesion molecule NCAM. *Neuron* **26**: 207–17

Fitzsimonds R M, Poo M M 1998 Retrograde signaling in the development and modification of synapses. *Physiological Review* **78**: 143–70

Flanagan J G, Vanderhaeghen P 1998 The ephrins and Eph receptors in neural development. *Annual Review of Neuroscience* **21**: 309–45
Hall A 1998 Rho GTPases and the actin cytoskeleton. *Science* **279**: 509–14
Garner C C, Nash J, Huganir R L 2000 PDZ domains in synapse assembly and signalling. *Trends in Cell Biology* **10**: 274–80
Lee S H, Sheng M 2000 Development of neuron–neuron synapses. *Current Opinion in Neurobiology* **10**: 25–31
Ming G L, Song H J, Berninger B, Holt C E, Tessier-Lavigne M, Poo M M 1997 cAMP-dependent growth cone guidance by netrin-1. *Neuron* **19**: 1225–35
Mueller B K 1999 Growth cone guidance: first steps towards a deeper understanding. *Annual Review of Neuroscience* **22**: 351–88
Sanes J R, Lichtman J W 1999 Development of the vertebrate neuromuscular junction. *Annual Review of Neuroscience* **22**: 389–442
Sanes J R, Yamagata M 1999 Formation of lamina-specific synaptic connections. *Current Opinion in Neurobiology* **9**: 79–87
Schachner M 1997 Neural recognition molecules and synaptic plasticity. *Current Opinion in Cellular Biology* **9**: 627–34
Tanaka H, Shan W, Phillips G R, Arndt K, Bozdagi O, Shapiro L, Huntley G W, Benson D L, Colman D R 2000 Molecular modification of N-cadherin in response to synaptic activity. *Neuron* **25**: 93–107
Walsh F S, Doherty P 1997 Neural cell adhesion molecules of the immunoglobulin superfamily: role in axon growth and guidance. *Annual Review of Cellular and Developmental Biology* **3**: 425–56
Winberg M L, Mitchell K J, Goodman C S 1998 Genetic analysis of the mechanisms controlling target selection: complementary and combinatorial functions of netrins, semaphorins, and IgCAMs. *Cell* **93**: 581–91

A. Dityatev and M. Schachner

Synapse Ultrastructure

The chemical synapse, which is the primary location for communication between neurons, is characterized by the apposition of two neuronal membranes with unequal membrane densities and a cluster of synaptic vesicles close to the synaptic site. Synaptic ultrastructure refers to the study of the physical components that make up the chemical synapse. This article examines some of the important findings in this area and considers the potential functional relevance of changes in synaptic ultrastructure that have been observed following neuronal activation, neural lesions, learning, and memory.

1. *Why is Synaptic Ultrastructure Important?*

Ramon y Cajal (1893; cited in Bliss and Collingridge 1993) theorized that learning and memory formation resulted from the formation of new synapses. Others speculated that learning might result from changes in the relative strength or efficacy of existing synapses. Donald Hebb (1949) postulated that one neuron's involvement in firing a second neuron would strengthen the connection between the two neurons. The most likely location for this type of change in connective strength is the synapse. At the time of Hebb's proposal there was little evidence to suggest that synapses were capable of change, either anatomically or physiologically. Based on advances in research methodology, however, it is clear that synapses are indeed capable of these types of changes.

As a result of this early research (Cajal 1893, Hebb 1949), the structure and functioning of the synapse has become an area of intense study. While neuronal, axonal, and dendritic changes are also undoubtedly important, the synapse probably represents the primary location for activity dependent neural plasticity. Research on synapses includes examining synaptic chemical transmission (see *Synaptic Transmission*), how synapses form (see *Synapse Formation*), and the differences in synaptic connections throughout the brain. While describing specific synaptic components has also been an important pursuit, quantifying the number and dimensions of synapses is the primary goal of research on synapse ultrastructure.

Quantifying synapses may be especially important for functional reasons when synapses change in number or dimension following neural activation, neural lesions, learning, and memory formation. While a comprehensive review of the literature in this area is beyond the scope of this article, the basic components of the synapse are described, some of the research examining the structural plasticity of the synapse is discussed, and a series of ultrastructural changes that may support the anatomical storage of information is proposed.

2. *Basic Ultrastructural Components of the Synapse*

The following is a brief description of the basic physical components of the synapse. The synapse can be divided into presynaptic components, the synaptic cleft, and postsynaptic components. The presynaptic component usually arises from the end of an axonal branch and the postsynaptic component is usually a dendrite or a dendritic spine. Although many other synaptic contact types exist (e.g., axo-axonic (axon to axon), axosomatic (axon to neuron cell body), etc.), the majority of synapses in the more plastic commonly studied regions of the brain (e.g., the hippocampus and cerebral cortex) consist of axon terminals contacting dendritic spines.

The synaptic junction contains a specialized area referred to as the active zone where synaptic transmission is known to take place. This area is composed of two plasma membranes separated by a cleft, with

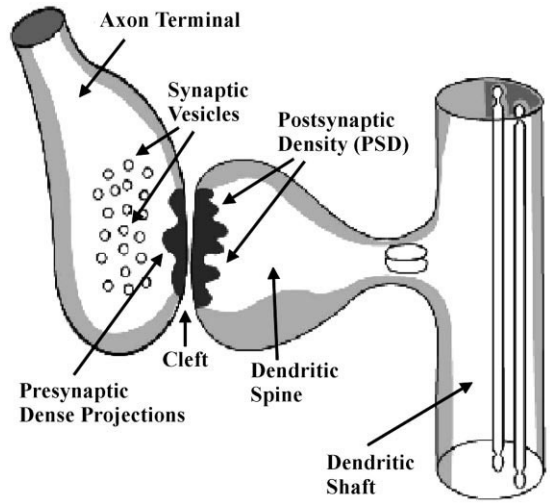

Figure 1
Basic ultrastructural components of the synapse

dense structural specializations (referred to as densities, dense projections, or thickenings; see Fig. 1). The presynaptic axon terminal contains synaptic vesicles situated within the presynaptic dense projections. These dense projections are tufts of electron dense material assumed to be essential for vesicular release. The postsynaptic element contains the postsynaptic density (PSD), which is comprised of various proteins (Bloomberg et al. 1977) and is associated with actin, a calcium calmodulin dependent protein kinase (Cohen et al. 1985). While PSDs are also associated with other kinases etc., actin may play a particularly important role in synaptic ultrastructural change as it is involved in maintaining the structure of the postsynaptic element (see Sect. 3.2). The PSD is also the site of postsynaptic neurotransmitter receptors and is considered integral to modulation of synaptic events (Seikevitz 1985).

Gray (1959) proposed that chemical synapses come in two basic morphological types (Type I and Type II). Type I synapses are thought to be excitatory, house round vesicles, and have a wider PSD region (asymmetric synapses). Type II synapses are thought to be inhibitory, house oblong or oval vesicles and have a smaller PSD (symmetric synapses).

3. *Ultrastructural Plasticity of the Synapse*

It has become clear that neuronal growth, changes in connectivity, and changes in activity are associated with modulation of the structure of synapses (Bailey and Kandel 1993). Changes have been observed in synaptic number, synaptic curvature, synaptic size, and synaptic configuration (e.g., perforated synapses, see Sect. 3.4). Please note that this article does not include a consideration of the ultrastructure of dendritic spines (see Harris and Kater 1994 for review).

3.1 *Quantifying Synaptic Number*

When examining the ultrastructure of synapses in a given brain region the initial analysis usually involves some estimate of the total number of synapses in that region. In early research synapses were often counted on a single plane within the region of interest. That is, researchers used an electron microscope to photograph one large two-dimensional area and simply counted how many synapses were present. Gundersen (1977) and others began to realize that this form of counting involved a serious bias. Coggeshall and Lekan (1996) point out that considering only one plane biases the synaptic counts because the number of two-dimensional synaptic profiles observed depends not only on the actual number of whole synapses but also on the size, shape, orientation, etc. of the synapses. As a result, the method of using stereological dissectors to estimate synaptic number was developed. Put simply this involves examining a series of micrographs (photographs) that move through the tissue in three dimensions. Synapses are only counted once when they first appear which rules out the bias identified above and includes all synapses present (see Geinisman et al. 1996b for more details). Synaptic number in a brain region is assumed to important functionally as more synapses would yield stronger connections between neurons (Petit 1995).

3.2 *Changes in Synaptic Curvature*

Changes in synaptic curvature are also known to occur following various forms of synaptic activation (see Petit 1995 for review). Markus and Petit (1989) evaluated synaptic curvature using four possible states: concave (presynaptic element protrudes into the postsynaptic element), flat (no curvature), convex (postsynaptic element protrudes into the presynaptic element), and irregular (w-shaped or more than one curvature). One possible explanation for curvature changes is that plastic internal actin and myosin cytoskeletal networks may create changes in response to activity dependent calcium influx. Fifkova (1985) found that cytoskeletal networks that contain only highly plastic actin structures are found exclusively in developing neurons and mature dendritic spines (see Deller et al. 2000 for an update on this research).

Functionally, curvature is thought to effect synaptic efficacy by increasing contact area and the probability of the transmitters reaching their target (see Markus and Petit 1989 for review). Computer modeling experiments have also suggested that concave shaped syn-

apses, seem to confine the diffusion of presynaptic intracellular calcium (Ghaffari et al. 1997). This would lead to higher calcium concentrations in the presynaptic terminal and an increased probability of vesicular (transmitter) release (see *Synaptic Transmission*).

3.3 Changes in Synaptic Size

Most researchers have theorized that larger synapses are stronger synapses. Conceivably, more neurotransmitter could be released (presynaptically), more postsynaptic receptors could be present, and more conductance through to the postsynaptic cell could occur (Petit 1995). A recent study used cultured cortical neurons to study the effect of synaptic size on the quantal capacity of the synapse (Mackenzie et al. 1999). By imaging Ca^{2+} and then conducting morphological analyses on the same synapse, Mackenzie et al. concluded that synaptic size correlates positively with the amplitude of the postsynaptic response. This suggests that larger synapses create larger synaptic signals.

3.4 Changes in Synaptic Configuration

One of the main changes in synaptic ultrastructure is the transition of some synapses from a continuous synapse to a perforated synapse. A perforated synapse can be defined as any synapse with a discontinuous or segmented PSD (Geinisman et al. 1991). Carlin and Siekevitz (1983) proposed that at some optimal size the postsynaptic material would perforate, eventually segment, and potentially new simple synapses would form. There is, however, evidence against the notion that synapses split to form new simple synapses. Jones et al. (1991) found few double-headed spines, axo-spinous non-perforated synapses lying adjacent to one another, and spinules completely traversing the presynaptic terminals of perforated synapses. Functionally, perforations may indicate synapses ready to divide and this may lead to more synapses and a larger neural signal (Toni et al. 1998). Alternatively, perforations may allow calcium channels to be located closer to the vesicular release apparatus, which could result in greater neurotransmitter release (Jones et al. 1991).

4. Changes in Synaptic Ultrastructure following Neural Events

As mentioned in Sect. 1, a comprehensive review of the literature on changes in synaptic ultrastructure is beyond the scope of this article. As a result synaptic changes following the onset of disease and other models are not discussed. The following is a selection of findings associated with models that involve activation of synapses primarily in the hippocampal formation due to its involvement in learning and memory (see *Hippocampus and Related Structures*).

4.1 Long-term Potentiation

Bliss and Lomo (1973) are credited with the discovery of long-term potentiation (LTP), which has proven to be a popular model system for learning and memory (see *Long-term Potentiation (Hippocampus)*). Put simply LTP is a sustained increase (hours, days, or weeks) in the amplitude of the response evoked in a cell, or population of cells, following tetanic (repeated) stimulation of those cells. The ultrastructural characteristics of synapses have been shown to change following the induction of LTP (Geinisman et al. 1991, Desmond and Levy 1983, Weeks et al. 1999). It is not clear, however, what role, if any, these changes play in the enhanced neural signal observed during the maintenance of LTP (see Sect. 3.2–3.4 for some possibilities).

While LTP can be induced in several brain regions, the majority of the research on changes in synaptic ultrastructure following LTP has been conducted in the hippocampus. Geinisman et al. (1996a) found more numerous axo-dendritic synapses (synapses directly on the dendritic shaft) 13 days following the induction of LTP in the rat dentate gyrus (a sub-region of the hippocampus). Functionally, these axo-dendritic synapses may have a more direct effect on the soma and could therefore combine to create a potentiated signal (see Harris and Kater 1994 for discussion). Desmond and Levy (1983) analyzed synaptic curvature in the rat dentate gyrus and found that the number of concave shaped synapses increased as early as 2 minutes after the induction of LTP. General increases in synaptic size following LTP have also been reported (Chang and Greenough 1984, Desmond and Levy 1988). These studies found an enlargement of the presynaptic and postsynaptic elements, and an increase in length of the presynaptic dense projections and postsynaptic density (PSD).

Geinisman et al. (1991) found an increase in the number of perforated synapses per neuron following the induction of LTP in the rat dentate gyrus. Importantly, the significant increase in perforated synapses observed at 1 hour post-induction was no longer evident at 13 days post-induction (Geinisman et al. 1996a). Buchs and Muller (1996) utilized calcium accumulation markers to identify activated synapses and found that these activated synapses were more perforated following LTP induction. Toni et al. (1999) reported an increase in synaptic perforations in activated synapses during the first 30 minutes following *in vitro* LTP induction in the rat CA1 area (another sub-region of the hippocampus). At 60 minutes post-induction, they found an increase in the proportion of double synapses (two spines connected to the same

presynaptic terminal) suggesting that perforated synapses may eventually split to form new simple synapses. Contradictory evidence has come from Sorra and Harris (1998) who did not find any ultrastructural changes 2 hours post-LTP induction in area CA1 of the rat.

In a series of experiments, Weeks et al. (1999, 2000, 2001) considered the pattern of synaptic ultrastructure at 1 hour, 24 hours, and 5 days post-LTP induction. LTP was associated with an increase in concave shaped synapses at 24 hours. Synapses were larger overall at 1 hour and 5 days but not different at 24 hours. These differences in length were particularly evident in concave-shaped synapses, which were longer at 1 hour, shorter at 24 hours, and longer at 5 days. The proportion of perforated synapses was increased at 1 hour post-LTP induction but did not differ from controls at later time periods. Taken together and added to the other research in this area these results form a model of ultrastructural remodeling that occurs in activated synapses following LTP. Synapses appear to initially grow in size, become concave in shape and perforate by 1 hour post-LTP induction, either split or form new smaller concave synapses at 24 hours and then grow again in size by 5 days.

4.2 Learning and Memory

Synaptic ultrastructure has also been shown to change following learning. Reempts et al. (1992) found an increased number of concave perforated synapses in the rat dentate gyrus following the acquisition of a one-way active avoidance task. Concave synapses were also found to be increased in size relative to controls following training. This result is similar to the findings at 1 hour post-LTP induction (Weeks et al. 2000).

Synapse ultrastructure was also analyzed following the hippocampus-dependent trace eye-blink conditioning in the rabbit (see *Classical Conditioning, Neural Basis of*; *Eyelid Classical Conditioning*). Geinisman et al. (2000) found that while the number of synapses did not change following conditioning the PSDs were significantly larger in the trained animals. Geinisman et al. speculated that this increase in size might represent the addition of signal transduction proteins (thus increasing synaptic strength).

Learning that involves the cerebellum (see *Cerebellum: Associative Learning*) has also been associated with changes in synaptic ultrastructure. Kleim et al. (1996) found that motor skill learning caused synaptogenesis (formation of new synapses) in the rat cerebellum (see *Motor Control Models: Learning and Performance*). The number of synapses per purkinje cell increased in rats after they learned to navigate an aerial maze compared to motor controls. Further, the number of multiple varicosities (clusters of synapses) increased following learning as did the amount of branching in the purkinje cell processes. Kleim et al. also found that as the number of synapses increased, their average size decreased. This result is similar to the changes observed in the rat dentate gyrus 24 hours post-LTP induction (Weeks et al. 1999).

4.3 Synaptic Ultrastructure Following Neural Lesions

Reactive synaptogenesis (formation of new synapses after a neural lesion) in the hippocampus was examined by Anthes et al. (1993). Following ipsilateral entorhinal cortical lesions (see *Hippocampus and Related Structures*), synapses were quantified in the rat dentate gyrus at 3, 6, 10, 15, and 30 days post-lesion. Results showed that the lesions caused an initial 88 percent synaptic loss at day 3, which was followed by rapid synaptogenesis from day 6 through to day 15. Anthes et al. speculated that, following the loss of entorhinal input, previously dormant or suppressed fibers and their synapses became active in the absence of the primary innervation. Interestingly, synaptic size was found to decrease during the phase of rapid synaptogenesis. As synaptogenesis returned to baseline levels (approximately day 15), synaptic size also returned to control or pre-lesion levels. Another important finding was that the number of perforated synapses was greatest at the peak of synaptogenesis (days 10–15) and returned to control levels by day 30 post-lesion. Despite the differences in the length of time post-lesion, the sequence of synaptic structural changes observed by Anthes et al. is very similar to that observed by Weeks et al. (2001) following LTP (see Sect. 4.1).

5. Proposed Sequence of Synaptic Ultrastructural Change

There are limitations to the scope of the sequence of synaptic ultrastructural changes proposed. This sequence is based primarily on changes observed at excitatory, axo-spinous (axon to dendritic spine) synapses. Therefore, these modifications may not take place in inhibitory, axo-dendritic, or axo-somatic synapses, etc.

The data required to make a detailed comparison between the synaptic structural changes associated with LTP, reactive synaptogenesis, learning, and memory is far from complete but it is tempting to speculate about a common mechanism. One intriguing possibility, which was suggested by the growth of clusters of synapses observed by Kleim et al. (1996), is that a subset of synapses become involved or recruited in the various forms of neural plasticity. This subset of synapses then follows a sequence of structural changes (an initial increase in size, concavity, and perforations, then a decrease in size but an increased number of synapses, followed by a return to baseline levels) to

produce more numerous and stronger synapses within the activated groups of afferent fibers. This finding directly supports Hebb's (1949) postulate (see Sect. 1) and allows for lasting ultrastructural change in synapses that form activated networks within the nervous system. These activated networks may then form a basis for learning, memory, and other psychological processes.

See also: Long-term Depression (Cerebellum); Long-term Depression (Hippocampus); Long-term Potentiation and Depression (Cortex); Long-term Potentiation (Hippocampus); Memory: Synaptic Mechanisms; Neurotransmitters; Synapse Formation; Synaptic Efficacy, Regulation of; Synaptic Transmission

Bibliography

Anthes D, LeBoutillier J, Petit T 1993 Structure and plasticity of newly formed adult synapses: A morphometric study in rat hippocampus. *Brain Research* **626**: 50–62

Bailey C, Kandel E 1993 Structural changes accompany memory storage. *Annual Review of Physiology* **55**: 397–426

Bliss T, Collingridge G 1993 A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature* **361**: 31–9

Bliss T, Lomo T 1973 Long lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* **232**: 331–56

Bloomberg F, Cohen R, Siekevitz P 1977 The structure of postsynaptic densities isolated from dog cerebral cortex. II. Characterization and arrangement of some of the major proteins within the structure. *Journal of Cell Biology* **78**: 204–25

Buchs P, Muller D 1996 Induction of long-term potentiation is associated with major ultrastructural changes of activated synapses. *Proceedings of the National Academy of Science USA* **93**: 8040–5

Carlin R, Siekevitz P 1983 Plasticity in the central nervous system; do sysapses divide? *Proceedings of the National Academy of Science USA* **80**: 3517–21

Chang F, Greenough W 1984 Transient and enduring morphological correlates of synaptic activity and efficacy change in the rat hippocampal slice. *Brain Research* **309**: 35–46

Coggeshall R, Lekan H 1996 Methods for determining numbers of cells and synapses: A case for more uniform standards of review. *Journal of Comparative Neurology* **364**: 6–15

Cohen R S, Chung S K, Pfaff D W 1985 Immunocytochemical localization of actin in dendritic spines of the cerebral cortex using colloidal gold as a probe. *Cellular and Molecular Neurobiology* **5**: 271–84

Deller T, Merten T, Roth S, Mundel P, Frotscher M 2000 Actin-associated protein synaptopodin in the rat hippocampal formation: Localization in the spine neck and close association with the spine apparatus of principal neurons. *Journal of Comparative Neurology* **418**: 164–81

Desmond N, Levy W 1983 Synaptic correlates of associative potentiation/depression: An ultrastructural study in the hippocampus. *Brain Research* **265**: 21–30

Desmond N, Levy W 1988 Synaptic interface surface area increases with long-term potentiation in the hippocampal dentate gyrus. *Brain Research* **453**: 308–14

Fifkova E 1985 Actin in the nervous system. *Brain Research Reviews* **9**: 187–215

Geinisman Y, de Toledo-Morrell L, Morrell F 1991 Induction of long-term potentiation is associated with an increase in the number of axospinous synapses with segmented postsynaptic densities. *Brain Research* **566**: 77–88

Geinisman Y, de Toledo-Morrell L, Morrell F, Persina I, Beatty M 1996a Synapse restructuring associated with the maintenance phase of hippocampal long-term potentiation. *Journal of Comparative Neurology* **367**: 413–23

Geinisman Y, Disterhoft J, Gundersen H, Mcechron M, Persina I, Power J, Van Der Zee E, West M 2000 Remodeling of hippocampal synapses after hippocampus-dependent associative learning. *Journal of Comparative Neurology* **417**: 49–59

Geinisman Y, Gundersen H, Van Der Zee E, West M 1996b Unbiased stereological estimation of the total number of synapses in a brain region. *Journal of Neurocytology* **25**: 805–19

Ghaffari T, Liaw J, Berger T 1997 Impact of synaptic morphology on presynaptic calcium dynamics and synaptic transmission. *Society for Neuroscience Abstracts* **23**: 2105

Gray E G 1959 Axo-somatic and axo-dendritic synapses of the cerebral cortex: An electron microscope study. *Journal of Anatomy* **93**: 420–33

Gundersen H J G 1977 Notes on the estimation of the numerical density of arbitrary profiles: The edge effect. *Journal of Microscopy* **111**: 219–23

Harris K, Kater S 1994 Dendritic spines: cellular specializations imparting both stability and flexibility to synaptic function. *Annual Review of Neuroscience* **17**: 341–71

Hebb D 1949 *The Organization of Behaviour*. Wiley, New York

Jones D G, Itarat W, Calverly R K S 1991 Perforated synapses and plasticity: A developmental overview. *Molecular Neurobiology* **5**: 217–28

Kleim J A, Lussing E, Schwarz E, Comery T, Greenough W 1996 Synaptogenesis and FOS expression in the motor cortex of the adult rat after motor skill learning. *Journal of Neuroscience* **16**: 4529–35

Mackenzie P, Kenner G, Prange O, Shayan H, Umemiya M, Murphy T 1999 Ultrastructure correlates of quantal synaptic function at single CNS synapses. *Journal of Neuroscience* **19**: RC13 (1–7)

Markus E, Petit T 1989 Synaptic structural plasticity: Role of synaptic shape. *Synapse* **3**: 1–11

Petit T 1995 Structure and plasticity of the Hebbian synapse: The cascading events for memory storage. In: Spear L P, Spear N E, Woodruff M (eds.) *Neurobehavioral Plasticity: Learning, Development and Response to Brain Insults*. Erlbaum, Hillsdale, NJ, pp. 185–205

Reempts J, Dikova M, Werbrouck L, Clincke G, Borgers M 1992 Synaptic plasticity in rat hippocampus associated with learning. *Behavioral Brain Research* **51**: 179–83

Siekevitz P 1985 The postsynaptic density: A possible role in long-lasting effects in the central nervous system? *Proceedings of the National Academy of Science, USA* **82**: 3494–8

Sorra K, Harris K 1998 Stability of synapse number and size at 2 hr after long-term potentiation in hippocampal area CA1. *Journal of Neuroscience* **18**: 658–71

Toni N, Buchs P, Bron C, Muller D 1999 LTP promotes formation of multiple spine synapses between a single axon terminal and a dendrite. *Nature* **402**: 421–5

Weeks A, Ivanco T, LeBoutillier J, Racine R, Petit T 1999 Sequential changes in the synaptic structural profile following long-term potentiation in the rat dentate gyrus: I. The intermediate maintenance phase. *Synapse* **31**: 97–107
Weeks A, Ivanco T, LeBoutillier J, Racine R, Petit T 2000 Sequential changes in the synaptic structural profile following long-term potentiation in the rat dentate gyrus: II. The induction/early maintenance phase. *Synapse* **36**: 286–96
Weeks A C W, Ivanco T L, LeBoutillier J, Racine R, Petit T 2001 Sequential changes in the synaptic structural profile following long-term potentiation in the rat dentate gyrus: III. Long-term maintenance phase. *Synapse* **40**: 74–84

A. C. W. Weeks

Synaptic Efficacy, Regulation of

Multiple biological signals and mechanisms regulate the efficacy of synaptic transmission, thus providing rich combinatorial possibilities for modifying neural communication. These processes have been studied particularly well in invertebrate preparations, whose large neurons and relatively simple neural circuits permit detailed mechanistic and functional analyses.

1. *Multiple Processes Regulate Synaptic Efficacy*

The processes regulating synaptic efficacy vary along several dimensions, including (a) the nature of the change itself (e.g., an increase or decrease in efficacy); (b) mechanisms of induction —that is, the nature of the neural signals that initiate the synaptic change; (c) mechanisms of expression, the proximal causes of synaptic changes; (d) mechanisms of maintenance, the steps involved in their persistence; and (e) time course, ranging from transient forms of regulation lasting only a few seconds, to relatively permanent long-term changes persisting a day or more.

2. *Simple Forms of Synaptic Regulation*

2.1 *Decreases and Increases of Synaptic Efficacy*

Simple forms of synaptic regulation include decreases and increases in synaptic efficacy. Synaptic depression is typically expressed as a decrease in the amplitude or other parameters of a postsynaptic potential (PSP) at a given synaptic connection (Fig. 1). Other things being equal, the larger the synaptic potential, the greater the influence a presynaptic neuron exerts on its postsynaptic target neuron. Hence, in synaptic depression, the presynaptic neuron exerts less control over its postsynaptic target. Facilitation, enhancement, augmentation, and potentiation all refer to an increase in synaptic strength, often expressed as an increase in PSP amplitude (Fig. 1).

Synaptic regulation has been studied most extensively at chemical synapses. However, changes in the efficacy of electrical synapses can also occur.

2.2 *Mechanisms of Induction*

A separate but related issue is the type and locus of cellular activity necessary to induce, or trigger, the change in synaptic efficacy.

Homosynaptic changes arise from activity that occurs at the same ('homo') synapse that is modified. Thus, homosynaptic depression is a decrease in synaptic strength ('depression') that arises from activity at that synapse (hence, 'homosynaptic'). In contrast, heterosynaptic changes arise from activity in neurons other than those being modified. Both are instances of non-associative induction mechanisms, in that only one type or locus of neural activity is involved in triggering the synaptic change.

In contrast, conjunctive or associative mechanisms of synaptic regulation depend on activity in more than one cell. Associate mechanisms are of particular interest from a theoretical perspective, because they involve a more advanced regulatory logic: changes in synaptic strength occur if and only if two conditions are met (see *Associative Modifications of Individual Neurons*).

2.3 *Regulation of Synaptic Efficacy in the* Aplysia *Siphon-withdrawal Circuit*

The following four examples illustrate these different forms of induction as well as certain other aspects of synaptic regulation in the context of the gill- and siphon-withdrawal circuit of the marine mollusk *Aplysia* (Fig. 2).

As detailed by Eric Kandel and colleagues (Kandel and Schwartz 1982, Kandel and Pittenger 1999), the *Aplysia* gill- and siphon-withdrawal reflex to a tactile siphon stimulus is mediated in part by the LE siphon sensory neurons, which connect both directly and indirectly to LF_S and other siphon and gill motor neurons. The strength of LE sensorimotor connections can be modified in several ways (see *Memory: Synaptic Mechanisms*). Similar changes have also been found at connections of *Aplysia* tail mechanosensory neurons.

2.3.1 Homosynaptic depression and homosynaptic facilitation. Low-frequency activation of LE siphon sensory neurons depresses transmitter release at sensorimotor connections and hence also depresses the postsynaptic response (Fig. 1; Castellucci and Kandel 1976). Functionally, homosynaptic depression contributes to behavioral habituation of the gill-and siphon-withdrawal reflex. The siphon is a small fleshy spout that expels seawater from the mantle cavity. A

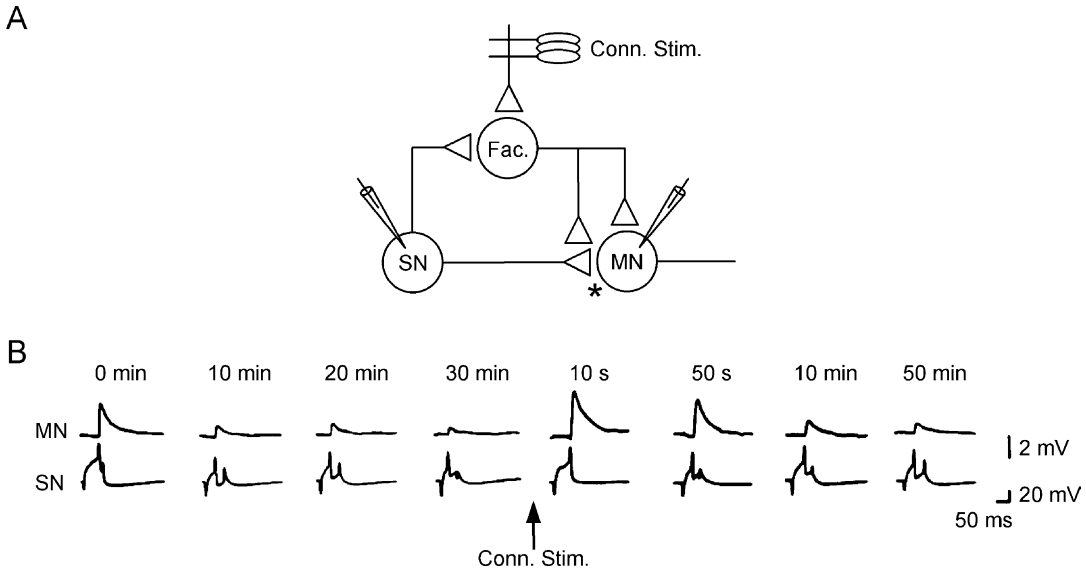

Figure 1
Homosynaptic depression and heterosynaptic facilitation at *Aplysia* sensorimotor synapses **A.** Simplified circuit and experimental arrangement. Siphon sensory neurons (SN) project monosynaptically and polysynaptically to gill and siphon motor neurons (MN). Extracellular stimulation of the connective, which carries input from tail and head to the abdominal ganglion, activates facilitatory interneurons (Fac.) that produce heterosynaptic facilitation at sensorimotor connections (asterisk) by enhancing transmitter release. **B.** Example of homosynaptic depression (left) and heterosynaptic facilitation (right). Intracellular stimulation evokes an action potential in the sensory neuron that elicits a monosynaptic excitatory synaptic potential (EPSP) in the motor neuron. With repeated stimulation, the amplitude of the EPSP declines (synaptic depression). Connective stimulation (Conn. Stim.) increases the EPSP amplitude (heterosynaptic facilitation). Source: Castellucci and Kandel 1976. © American Association for the Advancement of Science. Adapted with permission

tactile stimulus to the siphon causes both the gill and the siphon to withdraw. After repeated taps of the siphon, however, the amplitude and duration of the reflex diminish. Habituation is an important and ubiquitous form of learning that enables an organism to screen out stimuli of little consequence.

In addition to homosynaptic depression, *Aplysia* sensorimotor connections exhibit post-tetanic potentiation (PTP) in response to high-frequency activation, which is due at least in part to homosynaptic facilitation (Clark and Kandel 1984).

2.3.2 Heterosynaptic facilitation and heterosynaptic inhibition. In *Aplysia* siphon-withdrawal circuits, strong noxious stimuli (e.g., tail shock) activate a group of facilitatory interneurons (e.g., L29 cells) that synapse onto LE siphon sensory cells. Activation of facilitatory interneurons modifies sensory cells and increases their transmitter release via multiple mechanisms. Because this increase in synaptic efficacy is regulated by an extrinsic input, it is an example of heterosynaptic facilitation (Fig. 1; Castellucci and Kandel 1976). Behaviorally, presynaptic facilitation contributes to sensitization of the siphon-withdrawal reflex, increasing its amplitude and duration. With very strong stimuli, heterosynaptic inhibition can also be induced.

2.3.3 Activity-dependent enhancement of presynaptic facilitation. In this associative mechanism (also termed activity-dependent neuromodulation and activity-dependent extrinsic modulation), activity in sensory neurons enhances the heterosynaptic presynaptic facilitation normally evoked by facilitatory interneurons (Hawkins et al. 1983). The facilitation is greater than can be accounted for by a simple combination of a homosynaptic increase (PTP arising from sensory neuron activity) and heterosynaptic facilitation (arising from activation of facilitatory interneurons) and thus is associative. This mechanism is believed to contribute to classical conditioning of the siphon-withdrawal reflex, in which a siphon tap conditioned stimulus (CS) that activates siphon sensory neurons is paired with a tail shock unconditioned stimulus (US) that activates neuromodulatory facilitatory interneurons. Such pairings result in greater synaptic

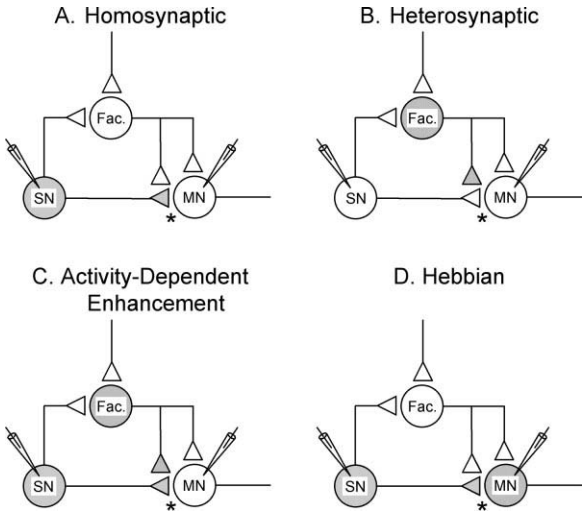

Figure 2
Four induction mechanisms for synaptic facilitation at *Aplysia* sensorimotor connections. In each case, the synapse being strengthened is the sensorimotor connection from the siphon sensory neuron (SN) to the siphon motor neuron (MN), as indicated by the asterisk. The locus of cellular activity critical for the induction of facilitation is indicated by shading of the cell or synapse. **A.** High-frequency activity of the SN alone (shaded) produces homosynaptic facilitation (e.g., PTP). **B.** Activity in facilitatory interneurons enhances transmitter release from the SN, resulting in heterosynaptic facilitation. **C.** Conjoint activation of the sensory neuron and the facilitatory interneuron produces even greater facilitation, indicating activity-dependent enhancement of heterosynaptic facilitation. **D.** Conjoint activity in the presynaptic SN and postsynaptic MN produces Hebbian facilitation

facilitation and a greater increase in the siphon-withdrawal response than that produced by unpaired CS and US presentations.

2.3.4 Hebbian facilitation. Hebbian facilitation (Hebb 1949) also contributes to increases in synaptic efficacy at LE-LF_s connections (Glanzman 1995). The critical inductive signal for Hebbian facilitation is conjoint activity of the presynaptic sensory neuron and activity or large depolarization of the postsynaptic motor neuron. The associative properties presumably arise from a glutamate N-methyl-D-aspartate (NMDA)-type receptor, which acts as a molecular 'AND-gate' that requires both neurotransmitter binding *and* large postsynaptic depolarization (to remove Mg^{2+} block of the channel) in order for Ca^{2+} influx to occur through the channel and trigger synaptic change (see *Associative Modifications of Individual Neurons*; *Long-term Potentiation (Hippocampus)*). Behaviorally, Hebbian facilitation also contributes to classical conditioning of the reflex. The tactile siphon CS activates siphon sensory neurons, and the tail shock US activates and depolarizes siphon motor neurons, thereby providing the key conjunctive events.

2.4 Expression, Maintenance, and Time Course of Synaptic Regulation

In principle, increases and decreases in synaptic efficacy can result from either a presynaptic expression mechanism (e.g., a decrease in transmitter release from the presynaptic cell), or a postsynaptic expression mechanism (e.g., a decrease in the sensitivity or number of transmitter receptors on the postsynaptic target cell), or both.

Although the mechanisms for expression and for maintenance and duration of synaptic regulation are logically and biologically separable, they are often functionally related and will be considered together here. Four broad categories of mechanisms allow the regulation of synaptic efficacy to span a wide range of temporal domains: (a) transient regulation, which is voltage- and Ca^{2+}-dependent, lasting milliseconds to seconds; (b) short-term regulation, involving second-messenger mediated covalent modifications of pre-existing proteins, typically lasting minutes to an hour; (c) intermediate-term regulation, dependent on protein synthesis, persisting several hours; and (d) long-term regulation, dependent on both RNA synthesis and protein synthesis, persisting for hours and days (Fig. 3).

2.4.1 Transient, voltage- and Ca^{2+}-dependent presynaptic regulation. As demonstrated by studies at the squid giant synapse, which permits intracellular stimulation and recording from both presynaptic and postsynaptic elements, transmitter release is regulated by the voltage of the presynaptic terminal, which in turn influences the basal concentration of intracellular Ca^{2+}, the ion responsible for transmitter release. Depolarization of the presynaptic terminal elevates intracellular Ca^{2+} concentrations and increases release, whereas hyperpolarization decreases intracellular Ca^{2+} concentrations and decreases release. Axo-axonic synapses (connections from one neuron onto presynaptic terminals of another) can depolarize or hyperpolarize presynaptic terminals, thereby regulating release and producing presynaptic facilitation or inhibition, respectively. This form of synaptic regulation is typically transient, as it is dependent on the continuation of the afferent axo-axonic input, so it allows a moment-to-moment gating of information flow. Homosynaptic spike activity can also influence residual Ca^{2+} levels and thus modify transmitter release relatively directly.

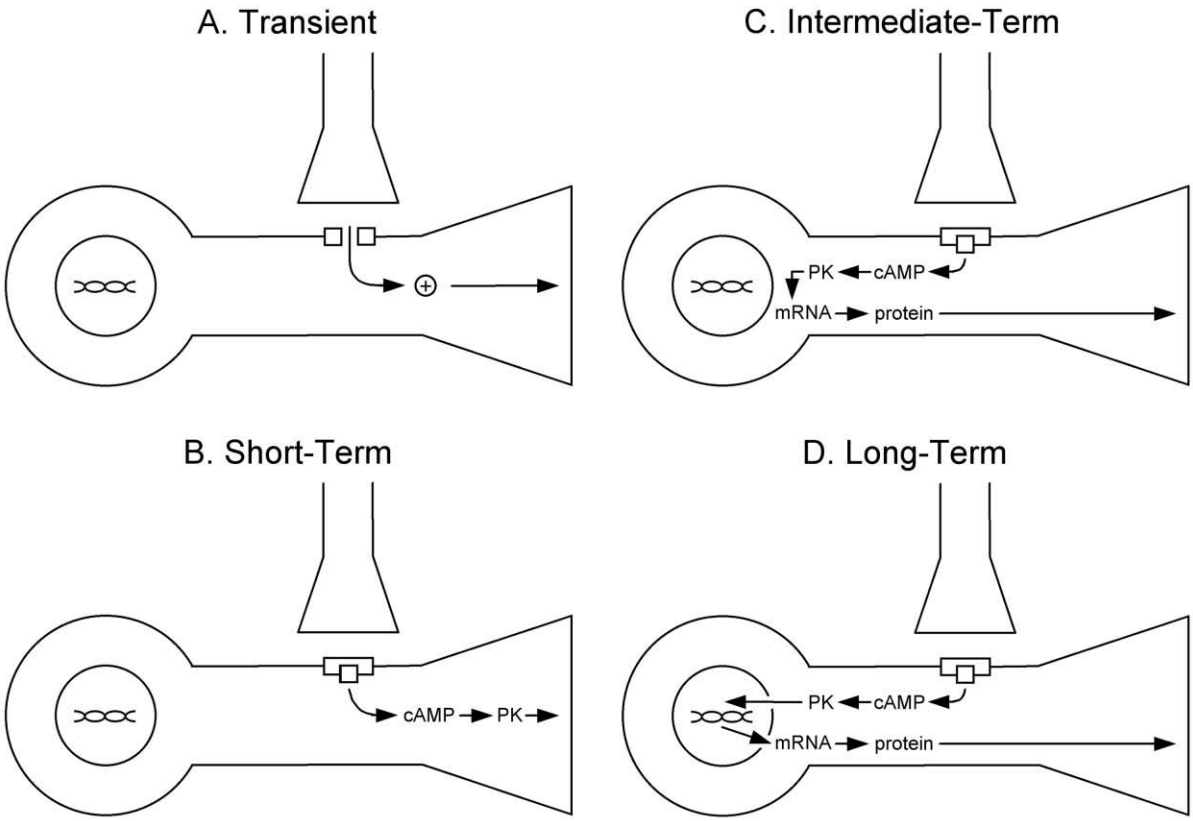

Figure 3
Maintenance of Enhanced Synaptic Efficacy. Four classes of mechanisms allow the regulation of synaptic efficacy to span a wide range of temporal domains, illustrated here by four examples of presynaptic facilitation. **A.** Transient voltage- and Ca^{2+}-dependent facilitation. Activation of presynaptic ionotropic receptors depolarizes the presynaptic terminal, increasing basal Ca^{2+} levels and enhancing transmitter release. The facilitation terminates quickly (within milliseconds to seconds) after afferent activity subsides and the terminal repolarizes. **B.** Short-term covalent modifications, as occur at *Aplysia* sensorimotor connections. Facilitatory inputs elevate concentrations of second messengers (e.g., cAMP), activate protein kinases, and phosphorylate substrate proteins, enhancing transmitter release. Short-term facilitation persists beyond the termination of afferent facilitatory input, but subsides on a time scale of minutes as substrate proteins are dephosphorylated. **C.** Intermediate-term facilitation at *Aplysia* sensorimotor connections lasting a few hours requires translation of mRNA to protein but not transcription of DNA to RNA. **D.** Long-term facilitation at *Aplysia* sensorimotor connections. Here the second messenger cascade regulates gene expression as well as protein synthesis, leading to relatively permanent synaptic facilitation lasting a day or more

2.4.2 Short-term regulation via second-messenger mediated covalent modifications. Short-term regulation of synaptic efficacy, lasting minutes to approximately an hour, typically involves second-messenger cascades and the subsequent covalent modification of pre-existing proteins, often via kinase-dependent phosphorylation. For example, in short-term facilitation at *Aplysia* sensorimotor connections, facilitatory interneurons activate both the cAMP-dependent protein kinase (PKA) and protein kinase C (PKC) in sensory cells. In the cAMP-dependent cascade, the facilitatory transmitter (e.g., serotonin), the 'first messenger,' binds to a receptor coupled to G-protein that activates the enzyme adenylyl cyclase. Adenylyl cyclase activation converts adenosine triphosphate (ATP) to cyclic adenosine monophosphate (cAMP), the 'second messenger.' cAMP in turn activates PKA by binding to the regulatory subunits of the kinase, causing them to dissociate from the catalytic subunits. The freed catalytic subunits then phosphorylate the substrate proteins, invoking conformational and functional changes. Similarly, in the

PKC cascade, serotonin activates phospholipase C via a separate G-protein coupled receptor, which in turn elevates the levels of the second messenger diacylglycerol and activates PKC. Together, these two second-messenger cascades produce a broadening of sensory neuron action potentials, enhanced Ca^{2+} influx, and an enhancement of spike-release coupling, thereby increasing transmitter release.

Second-messenger cascades modify a variety of substrate proteins, allowing for coordinated regulation of multiple steps involved in synaptic transmission. Because substrate phosphorylation persists even after the brief input from facilitatory interneurons terminates, synaptic facilitation is maintained beyond the presence of sensitizing stimulus itself. However, biochemical mechanisms for substrate dephosphorylation also exist. Hence, covalent modifications of substrate proteins by themselves are not well suited for more permanent maintenance of synaptic changes.

2.4.3 Intermediate-term and long-term regulation of synaptic efficacy involve translation and transcription. Intermediate-term (lasting hours) and long-term forms of synaptic regulation (lasting a day or more) typically require the recruitment of additional steps to those implicated in short-term regulation. The relationship among these different forms of cellular memory has been particularly well studied at *Aplysia* sensorimotor connections. Unlike short-term facilitation, certain forms of intermediate-term facilitation require the synthesis of new proteins (translation of mRNA to protein). Long-term facilitation involves not only protein synthesis, but also the modification of gene expression (transcription of DNA to RNA) (Montarolo et al. 1986). These forms of synaptic regulation are thus distinct on a mechanistic as well as phenomenological level.

With repeated sensitizing stimuli or their analogs (e.g., serotonin application) used to induce long-term facilitation, PKA phosphorylates not only substrate proteins involved in short-term facilitation at the synapse, but also the transcriptional activator cAMP-response element binding protein (CREB) in the sensory neuron nucleus. CREB then binds to and activates the promoter element CRE (the cAMP response element), inducing gene expression. The new gene products involve not only 'effector' molecules directly involved in the expression of long-term facilitation, but also additional transcription factors, such as C/EBP, that regulate the expression of other genes, thereby producing a cascade of gene regulation. Interruption of this gene cascade at early time points by inhibitors of RNA or protein synthesis blocks subsequent long-term facilitation, but leaves short-term facilitation intact. In contrast, inhibition of RNA and protein synthesis at later time points has little effect. Thus the dynamics of macromolecular synthesis in long-term facilitation mimic the time-sensitivity of long-term memory itself. cAMP and CREB have been implicated in neural substrates of memory in other systems, such as *Drosophila* (see *Memory in the Fly, Genetics of*).

Just as short-term facilitation involves the coordinated facilitation of several processes involved in transmitter release, long-term facilitation involves multiple effector genes. Enhanced expression of the enzyme ubiquitin carboxyterminal hydrolase promotes the degradation of the inhibitory regulatory subunits of PKA, thereby making the kinase persistently active, even after cAMP levels return to baseline. Other gene products lead to the growth of new sensory neuron synaptic terminals (which does not occurn short-term facilitation), thus providing a stable, enduring substrate for long-term memory. Structural changes in sensory neurons are accompanied by corresponding growth in postsynaptic targets, providing further evidence for the coordinated regulation of synaptic efficacy on multiple levels.

3. *Functional Consequences of Simple Synaptic Regulation*

3.1 *Modulation of Input–Output Pathways*

The functional consequences of simple forms of synaptic regulation can be relatively straightforward in circuits where there is a relatively uniform direction of information flow from input to output. For example, as described above, decreases in synaptic strength at *Aplysia* sensorimotor connections contribute to decreases in the siphon-withdrawal response (e.g., habituation), and increases in synaptic strength contribute to increases in the siphon-withdrawal response (e.g., sensitization and classical conditioning).

3.2 *Reconfiguration of Multifunctional Networks*

In other instances, however, even relatively simple forms of synaptic change can have more complex functional consequences. Many neural networks (or their components) are multifunctional networks, participating in more than one behavior. Because the operation of such networks depends on interactions among multiple non-linear processes at the cellular, synaptic, and network levels, regulation of synaptic efficacy can profoundly alter network operation, reconfiguring the network into functionally distinct modes.

3.2.1 Escape swimming in Tritonia. One interesting example is the reconfiguration of a premotor network from a resting reflex withdrawal circuit into a rhythmic central pattern generator (CPG) swim circuit in the marine mollusk *Tritonia* (Katz and Frost

1996). None of the three major interneuron types in this circuit have endogenous bursting properties; the oscillation of the network instead arises from, and hence depends upon, their synaptic interactions. The dorsal swim interneurons (DSIs) are essential elements of the CPG circuit for escape swimming. DSI activity contributes directly to the generation of the swim motor program; in addition, it rapidly enhances the synaptic efficacy and excitability of a second interneuron, C2. C2 has been implicated in the reconfiguration mechanism; its activity disinhibits the DSI neurons, enabling rhythmic dorsal flexions. Thus, the rapid enhancement of C2 contributions may functionally rewire the network, allowing the reconfiguration from simple reflex withdrawal to rhythmic swimming to occur. Because the regulation of C2 synaptic efficacy develops and decays dynamically with DSI activity, it may also play a role in the cycle-by-cycle generation of the swim rhythm itself.

3.2.2 CPGs in lobster stomatogastric ganglion. The gastric mill CPG and pyloric CPG in the lobster stomatogastric ganglion are also polymorphic networks that can be functionally reconfigured to produce fundamental variations in output under different circumstances (Harris-Warwick and Marder 1991). Such alterations arise both from regulation of synaptic efficacy in the circuit and from changes in the intrinsic properties of the component neurons. For example, stimulation of the anterior pyloric modulator (APM) neuron activates the pyloric motor pattern by enhancing the plateau potential capabilities of pyloric CPG neurons. The platean maintains neurons in a depolarized state and so allows them to fire repetitively. As a consequence, a brief excitatory synaptic input that was previously ineffective can trigger a transition of the neuron to the tonically active plateau condition. In the plateau state, neurons also actively resist hyperpolarization, thus rendering inhibitory inputs less effective. The combined enhancement of synaptic excitation and diminution of inhibition result in a more active network.

Modulatory transmitters produce bursting pacemaker potentials in the anterior burster (AB) neurons, the most important pacemaker of the pyloric CPG. Synaptic interactions within the stomatogastric ganglion are mediated partly by non-spiking release, which varies continuously with membrane potential. Hence, larger bursting potentials enhance transmitter release from AB terminals, increasing AB inhibition of certain postsynaptic CPG targets and removing them from the functional circuit, thereby altering the pyloric network and rhythm. Because different modulatory amines exert different effects on AB bursts, the anatomical circuit can be sculpted in a variety of ways to yield a number of different functional networks.

4. Qualitative Regulation of Synaptic Efficacy

The decreases and increases in synaptic efficacy described above involve primarily quantitative modifications of synaptic strength, rendering synapses either more or less potent. However, regulation of synaptic efficacy can also involve a number of qualitative changes in synaptic transmission, thus extending the range of adaptive mechanisms that the nervous system utilizes.

4.1 Regulation of Transmitter Type

An important additional form of synaptic regulation involves changes in the type, rather than the amount, of transmitter that a neuron releases. Such changes can arise either from shifts in the relative release of different co-transmitters resulting from different rates and patterns of neural firing, or from changes in transmitter synthesis (Simerly 1990). Release of one co-transmitter or another can produce qualitative changes in the overall functional effect of synaptic transmission on a particular postsynaptic target (e.g., from excitation to inhibition), rather than simple quantitative changes in a pre-existing synaptic response. Regulation of the transmitter type can also effectively add or remove a particular postsynaptic pathway from operation, thus allowing functional rewiring of the neural circuit without the need for anatomical reorganization. In principle, similar molecular switching might be accomplished by regulation of the relative number or sensitivities of different postsynaptic receptors (see *Long-term Potentiation (Hippocampus)*).

4.1.1 Lobster stomatogastric ganglion. Changes in co-transmitter release affect the pyloric CPG in the lobster stomatogastric ganglion. The inferior ventricular nerve (IVN) cells evoke both excitatory and inhibitory responses in the AB neurons, the most important pacemaker of the pyloric CPG, as described above. At low frequencies of IVN activation, one (unknown) excitatory transmitter is released, and the frequency of the pyloric rhythm increases. In contrast, at higher frequencies, the inhibitory transmitter histamine is preferentially released, and the pyloric rhythm is disrupted. In a variation on this theme, the neuromodulatory transmitter dopamine changes the sign of the synaptic connection between certain identified neurons in the CPG from inhibitory to excitatory by reducing the efficacy of the inhibitory synapse and uncovering a weak electrical connection.

4.1.2 Feeding in Aplysia. Differential release of co-transmitters may also provide for intrinsic self-regulation of a network. For example, the motor neurons in the feeding circuit of *Aplysia* use a variety of

peptide co-transmitters that modulate neuromuscular transmission, extending the dynamic range over which the muscles can operate effectively. High frequency activation of motor neuron B15 enhances the co-release of small cardioactive peptide B. This neuromodulatory peptide increases the amplitude and quickens relaxation time of the muscle, thereby enabling larger muscle contractions to be driven at a higher rate. Similarly, in the leech, the myogenic heart muscle is innervated by heart excitatory motor neurons that co-release acetylcholine and FMRFamide, a modulatory peptide that regulates amplitude and duration of heart beat. Amines also modulate the amplitude of potentials at many crustacean neuromuscular junctions.

4.2 Regulation of Synaptic Plasticity: Metaplasticity

An additional form of synaptic regulation, recently termed 'metaplasticity' (Abraham and Bear 1996), is manifested as a modification of the ability of the synapse to undergo changes in synaptic efficacy, rather than as a change in synaptic efficacy *per se*. For example, long-term potentiation (LTP) not only increases the population spike evoked in dentate granule cells of mammalian hippocampus by afferent stimulation, it also reduces non-linear response properties so that synaptic efficacy is less affected by the recent history of activity at that synapse. LTP thus increases the fidelity as well as the gain of the system (Berger and Sclabassi 1985). Regulation of more long-lasting forms of synaptic plasticity also occurs at hippocampal synapses, involving, for example, shifts in the tendency of synapses to exhibit LTP or long-term depression (LTD).

Metaplasticity has recently been implicated at identified inhibitory synapses of the siphon-withdrawal circuit of *Aplysia* (Fischer et al. 1997). The inhibitory interneuron neuron L30 exhibits multiple forms of synaptic enhancement that control the gain of the siphon-withdrawal reflex, including frequency facilitation during initial phases of high-frequency activation and synaptic augmentation and potentiation that persist after activation. Tail shock reduces the baseline strength of L30 connections onto L29 cells; it also attenuates augmentation/potentiation while leaving frequency facilitation relatively intact. Behaviorally, tail shock decreases the inhibition normally produced by cutaneous stimuli, perhaps by reducing augmentation of L30 inhibition, thus suggesting a functional role for metaplasticity within a defined neural circuit.

5. Regulation of Synaptic Efficacy and of Other Neural Properties

Although not the explicit subject of this article, changes in neural excitability and other intrinsic neural properties can occur in addition to modifications of synaptic efficacy. Synaptic and non-synaptic changes can co-occur in the same neurons, and can be functionally and even mechanistically related. For example, both *Aplysia* mechanosensory cells and *Hermissenda* type B photoreceptors exhibit learning-related increases in excitability (increased firing to a given stimulus), as well as enhanced synaptic strength (Schuman and Clark 1994). These two modifications work in concert to increase both the number and the potency of the synaptic signals conveyed to their postsynaptic targets. Similarly, as described above, reconfiguration of the *Tritonia* escape swim CPG and the pyloric CPG in lobster stomatogastric ganglion involve both synaptic and non-synaptic forms of neural regulation.

One important distinction between regulation of synaptic efficacy and of other neural properties is the greater specificity that synaptic changes may exhibit. Because any given neuron may form pre- or postsynaptic contacts with thousands of other cells, the ability to modify only selected connections of a neuron confers a high degree of precision in the modification of neural pathways. In contrast, intrinsic changes such as enhanced excitability are often expressed cell-wide or nearly cell-wide, enhancing responses to virtually all inputs and enhancing virtually all outputs. In *Aplysia* siphon sensory cells, several short-term forms of synaptic plasticity, including heterosynaptic facilitation, heterosynaptic inhibition, and Hebbian facilitation, exhibit synapse specificity (Clark and Kandel 1984, Clark 2001). Synapse specificity is perhaps not surprising in short-term regulation, which involves covalent modifications of pre-existing proteins. However, long-term heterosynaptic facilitation, which depends on regulation of transcriptional and translational events, is also in part synapse-specific (Clark 2001, Martin et al. 1997). Synapse-specific long-term facilitation raises the intriguing question of how new gene products are preferentially utilized at facilitated synapses, compared with other synapses of the same neuron. An emerging answer, consistent with comparable findings in hippocampal LTP, is that synapse-specific long-term facilitation involves synaptic tagging and depends in part on local protein synthesis. According to this hypothesis, relatively short-lived, local regulatory events at treated synapses allow the subsequent capture of new gene products that are shipped cell-wide, thus conferring the specificity. Irrespective of the particular mechanistic details, long-term synapse-specific facilitation provides an additional illustrative example of the rich interplay and elegance of the multiple biological processes involved in synaptic regulation.

See also: Associative Modifications of Individual Neurons; Ion Channels and Molecular Events in Neuronal Activity; Memory: Synaptic Mechanisms; Synaptic Transmission

Bibliography

Abraham W C, Bear M F 1996 Metaplasticity: The plasticity of synaptic plasticity. *Trends in Neurosciences* **19**: 126–30

Berger T W, Sclabassi R J 1985 Nonlinear systems analysis and its application to study of the functional properties of neural systems. In: Weinberger N M, McGaugh J L, Lynch G (eds.) *Memory Systems of the Brain*. Guilford Press, New York

Castellucci V, Kandel E R 1976 Presynaptic facilitation as a mechanism for behavioral sensitization in *Aplysia*. *Science* **194**: 1176–8

Clark G A 2001 The ins and outs of classical conditioning: Maxims from a mollusk. In: Steinmetz J E, Gluck M, Solomon P (eds.) *Model Systems and the Neurobiology of Associative Learning*. Erlbaum, Hillsdale, NJ

Clark G A, Kandel E R 1984 Branch-specific facilitation in *Aplysia* siphon sensory cells. *Proceedings of the National Academy of Sciences USA* **81**: 2577–81

Fischer T M, Blazis D E J, Priver N A, Carew T J 1997 Metaplasticity at identified inhibitory synapses in *Aplysia*. *Nature* **389**: 860–5

Glanzman D L 1995 The cellular basis of classical conditioning in *Aplysia* californica—It's less simple than you think. *Trends in Neurosciences* **18**: 30–6

Harris-Warrick R M, Marder E 1991 Modulation of neural networks for behavior. *Annual Review of Neuroscience* **14**: 39–57

Hawkins R D, Abrams T W, Carew T J, Kandel E R 1983 A cellular mechanism of classical conditioning in *Aplysia*: Activity-dependent amplification of presynaptic facilitation. *Science* **219**: 400–5

Hebb D O 1949 *The Organization of Behavior*. Wiley, New York

Kandel, E R, Pittenger C 1999 The past, the future and the biology of memory storage. *Philosophical Transactions of the Royal Society London B* **354**: 2027–52

Kandel E R, Schwartz J H 1982 Molecular biology of learning: Modulation of transmitter release. *Science* **218**: 433–43

Katz P S, Frost W N 1996 Intrinsic neuromodulation: Altering neuronal circuits from within. *Trends in Neurosciences* **19**: 54–61

Martin K C, Casadio A, Zhu H E-Y, Rose J C, Chen M, Bailey C H, Kandel E R 1997 Synapse-specific, long-term facilitation of *Aplysia* sensory to motor synapses: A function for local protein synthesis in memory storage. *Cell* **91**: 927–38

Montarolo P G, Goelet P, Castellucci V F, Morgan J, Kandel E R, Schacher S 1986 A critical period for macromolecular synthesis in long-term heterosynaptic facilitation in *Aplysia*. *Science* **234**: 1249–54

Schuman E M, Clark G A 1994 Synaptic facilitation at connections of *Hermissenda* type B photoreceptors. *Journal of Neuroscience* **14**: 1613–22

Simerly R B 1990 Hormonal control of neuropeptide gene expression in sexually dimorphic olfactory pathways. *Trends in Neurosciences* **13**: 104–10

G. A. Clark

Synaptic Transmission

As explained in detail by J. C. Eccles in 1964, nerve cells communicate via specialized regions of close juxtaposition that were first named 'synapses' by C. S. Sherrington, who used this term to denote the anatomical region where the axon terminal of one neuron contacts another neuron. Sherrington always used synapse in its functional sense (for example, to account for the one-way conduction in the reflex arc). However, it must now be considered from an anatomical perspective as well since, despite variations, there are common structural features consistent with some of the physiological properties of synapses. For example, given that there is a high resistance gap of 200 Å width separating neuronal membranes at most synapses, as first claimed by Rámon y Cajal, current generated by one cell cannot spread significantly to the next one. Thus, it is advantageous for postsynaptic terminals to utilize a chemical mediator to secure transmission. This concept was only established after decades of controversy and efforts to distinguish between electrical and chemical models, for both excitation and inhibition. Yet, although it is beyond the scope of this overview, electrotonic interactions between nerve cells, in the periphery and in the brain, can indeed occur, due to either (a) areas of continuity between them via gap junctions (Bennett 1977); or (b) the channeling of currents in adjacent cells by an unusually high external resistance. The second case is called ephaptic or, better stated, field effect transmission (Faber and Korn 1989). These two electrical modes, which should not be confused, allow signaling to occur instantaneously, and are well suited for the mediation of rapid behaviors, such as escape. In contrast, the hallmark of chemical synapses is the irreducible delay (-1 msec) between the incoming stimulus and the postsynaptic response, which reflects activation of the presynaptic machinery involved with transmitter release (Katz 1969, Llinas 1982). Chemical synapses, however, have remarkable adaptive advantages, such as: (a) variable gain; (b) a large repertoire of polarity and kinetics, resulting in excitations and inhibitions of different durations, depending upon their specific chemistry; and (c) the ability to store information on a long-term basis, due to plastic properties which have become a paradigm for studies of learning and memory.

1. Integrative Properties of Neurons and Synaptic Correlates

Since any behavior is ultimately produced by combined activity in neuronal assemblies, it is of critical interest to understand how each individual neuron reacts to incoming signals initiated in cells connected to it through synaptic junctions, as in a simple reflex chain (Fig. 1A). In fact, information is distributed widely in the nervous system, such that a single neuron may both have many targets and receive inputs from numerous presynaptic axonal terminals, according to the principles of divergence and convergence, respectively. Recording electrodes inserted in the various

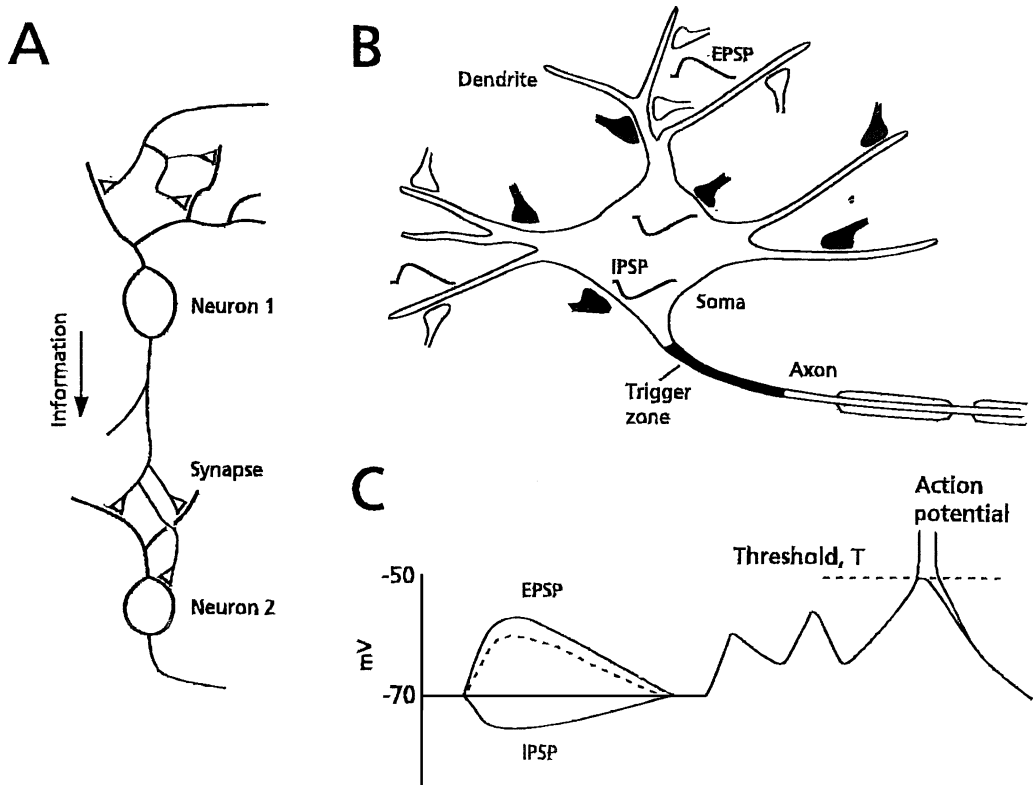

Figure 1
Integrative properties of neurons and mode of generation of action potentials

parts of a neuron (soma, dendrite, axon), in combination with anatomical studies of the synapses contacting these loci, have revealed electrical signals whose polarity reflects their function. The signals are named excitatory postsynaptic potentials (EPSPs), and inhibitory postsynaptic potentials (ISPSs). Due to nerve cell cable properties, which are determined by the membrane resistivity and capacitance, signals generated at a distance from the axon are attenuated there, but nevertheless contribute to the summation of responses at the so-called trigger zone (Fig. 1B), which decides whether or not some information will be passed on to the next neuronal relay. This essential integrative property of the cell can be quite complex, with there being multiple trigger zones (Llinas et al. 1969), and summation not always being algebraic (non-linear). Yet, whatever the rules are in a given neuronal subtype, the tonic and phasic bombardment of a cell by incoming inputs can only bring it to such a decision when excitation dominates inhibition and reaches a critical value, the threshold for generating an action potential (Fig. 1C) that will propagate along the axon towards the next synaptic relay. It can already be noted that reaching threshold is the result of the collective activity of large populations of inputs, each of which may, however, act randomly, as will be described below. This cooperative nature of the nervous system pertains whether it is viewed from a molecular, synaptic, network, or behavioral perspective.

2. *What Synapses Have in Common*

The meaning of the term 'synapse' has evolved over time, in parallel with the development of the morphological and physiological tools available for studies of the brain. Today, its strict definition requires electron microscopic demonstration of ultrastructural features associated with chemical transmission. A comparison of central synapses with the neuromuscular junction (nmj), which was the Holy Grail of pioneering studies, allows one to recognize their basic unity.

At the frog nmj, the terminals of the motor axon occupy the end plate, a specialized depression in the muscle membrane, which forms folds. Opposite the crest of each fold, where transmitter receptors are the

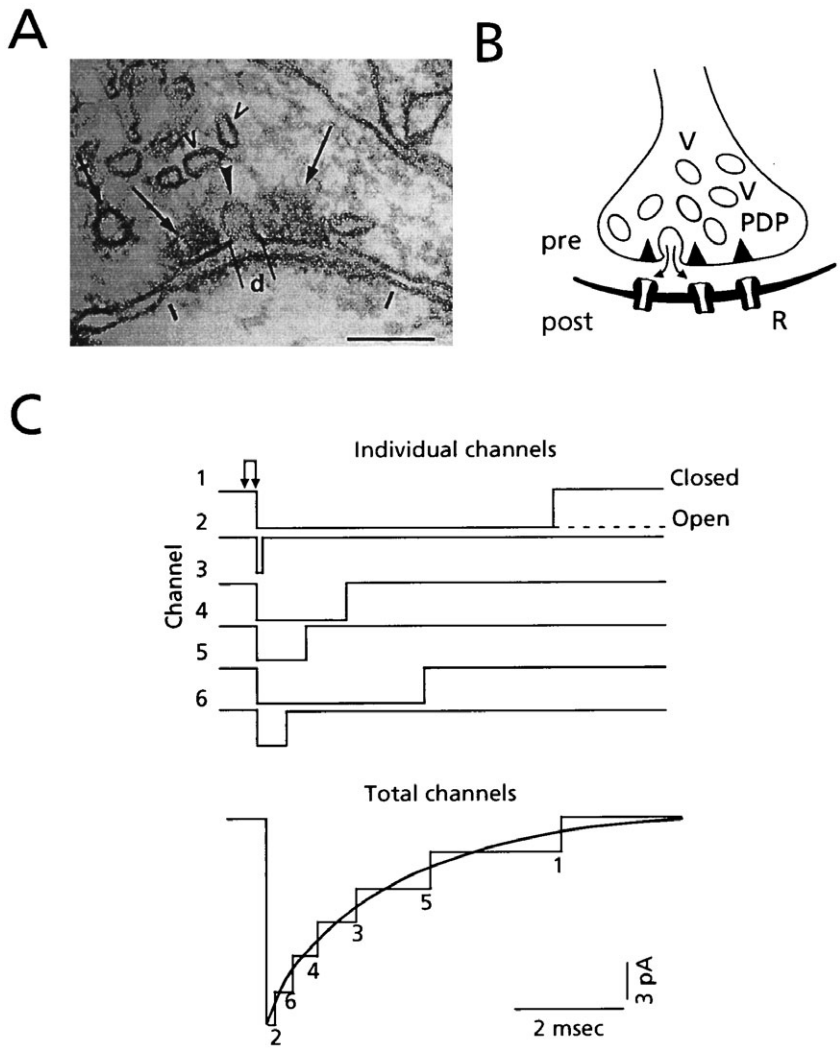

Figure 2
Presynaptic and postsynaptic components of quantal transmission

most abundant, the axon terminal contains dense bands and spherical organelles that package 5,000–10,000 molecules of the transmitter acetylcholine (Kuffler and Yoshikami 1975) and are called synaptic vesicles. It was later demonstrated that the opening of synaptic vesicles (exocytosis) takes place at the borders of the bands, which therefore were designated active zones (Couteaux and Pecot-Dechavassine 1970). An equivalent term for physiologists is 'release site,' as vesicle openings were observed there (Heuser et al. 1974). The motor end plate can thus be considered as a number of synaptic units, each represented by a dense band with its associated vesicles and a postsynaptic density (for review, see Peters et al. 1991).

Synapses in the central nervous system (CNS) have more restricted apposition zones. They most often consist of synaptic boutons (Fig. 2A, B) that contain the vesicles and are separated from the postsynaptic element by the 10–20 nm wide cleft. The plasma membranes on both sides exhibit marked densities which, together with the assembly of vesicles, are referred to as a synaptic complex (Palay 1958). Here the analog of the active zone is the presynaptic grid, formed by electron-opaque densities called presynaptic dense projections, which are arranged in triangular arrays and, in *en face* view, take the form of an 'egg box.' This structure is postulated to 'guide' the vesicles towards the plasma membrane (Pfenninger et

al. 1972). As at the nmj, the postsynaptic membrane facing this release apparatus contains the receptors which, however, are restricted to localized clusters (Triller et al. 1985). Thus, as further shown by physiological studies, the basic functional unit of synaptic transmission is a complex containing the presynaptic active zone and its neighboring synaptic vesicles, together with the aggregate of ligand-gated receptors in the opposing postsynaptic membrane. It thus corresponds to the synaptic complex.

At either type of chemical synapse, the depolarization of the terminals associated with a presynaptic action potential opens voltage sensitive Ca^{++} channels, with the consequent Ca^{++} influx initiating a cascade of molecular events that can culminate in the opening of a vesicle fused to the plasma membrane at the release site and the rapid discharge of its contents into the cleft (Fig. 2B). The postsynaptic receptors are thus exposed to a high concentration of transmitter, on the order of mM, albeit briefly (Fig. 2C, double arrow). Consequently, the ligand-gated channels open nearly synchronously, allowing specific ions to flow down their electrochemical gradient. However, as each individual opened channel in the postsynaptic receptor matrix closes randomly, the postsynaptic current, which is the sum of all individual channel activity, decays exponentially. This 'population' phenomenon is illustrated in Fig. 2C, where the summed current rises rapidly and relaxes slowly, a time course typical of those at the so-called fast synapses used by the brain for information processing. At the nmj, the contents of one vesicle open about 1000 receptor-channels, while at central synapses the number can be much less.

One should note that in the CNS the current flow through synaptic channels can be either inward (i.e., directed towards the cell interior) and excitatory, or outward and inhibitory, depending upon the selectivity of the channel, i.e., which ion(s) it allows to pass. Furthermore, in most neuronal systems, glycine and γ-aminobutyric acid (GABA) are inhibitory, and glutamate and acetylcholine are excitatory, although a given transmitter can have either function, depending upon the type of channel associated with that transmitter's receptor (Hille 1984).

3. *Why Synaptic Potentials Fluctuate in Integer Steps*

With a recording electrode placed at the nmj, Katz and colleagues made two related discoveries. First, at rest, spontaneous potentials of similar amplitudes occur randomly. Second, when the nerve is stimulated repeatedly, the resulting evoked endplate potentials fluctuate in amplitude in a steplike manner from trial to trial. Furthermore, amplitude histograms show that the apparently discrete size classes are integer (1, 2, 3, … n) multiples of a basic unit which corresponds in time course and average size to the spontaneous event (del Castillo and Katz 1954). These elementary 'miniature' endplate potentials were thus called 'quanta,' since they are representative of the basic unit of the evoked response. Remarkably, the underlying statistics of transmitter release, which account for the variable number of quanta issued after each impulse (based on experiments conducted in a low concentration of extracellular calcium ($[Ca^{2+}]_0$), satisfied the Poisson process. Release occurred with statistical independence at any one of a large number, n, of 'quantal units available for release' at the junction, with a low release probability (p) that was identical for each unit. Finally, it was postulated that each quantum corresponded to the action of the transmitter packaged in one synaptic vesicle, and since n is large in the Poisson equation, it was assumed to represent the total number of vesicles ready to be released (Katz 1969).

Despite difficulties inherent to the analysis of signals recorded from brain cells, such as the inability to relate any miniature response to one of the numerous junctions impinging on the neuron being investigated, and the low signal to noise ratio, the conclusion illustrated in Fig. 3A, namely that responses evoked by activating a single presynaptic neuron are quantal, was also reached for central synapses (Korn and Faber 1991, Redman 1990, see also Jonas et al. 1993). However, a statistical formulation other than the Poisson most often provides a better description of the amplitude fluctuations. Here the statistics are either a simple binomial model, where n is finite and p is relatively large, or a compound binomial, where p may be different at individual releasing units (see also below). Regardless of the mathematical details (Martin 1977 and see below), n is smaller than originally thought. Also, in experiments combining physiological recordings with anatomical reconstructions of the cells studied, the value derived for this parameter turned out to be equivalent to the number of active zones or release sites in the activated presynaptic terminals. The result lead to the 'one vesicle' hypothesis, according to which a presynaptic action potential causes the exocytosis of the contents of at most one vesicle at each active zone (reviewed in Korn et al. 1990). Despite controversies about multivesicular release, for example during development, this concept is still commonly used for interpreting experimental data (see Stevens 1993).

4. *Consequences of Quantal Transmission*

At the skeletal nmj, neurotransmission in normal Ca^{++} has a high safety factor (large n and p), such that a presynaptic impulse always leads to muscle contraction. On the other hand, the full implications and richness of the quantal nature, and hence uncertainty of transmitter release, are better exploited, albeit still puzzling, at central synapses. From a functional point

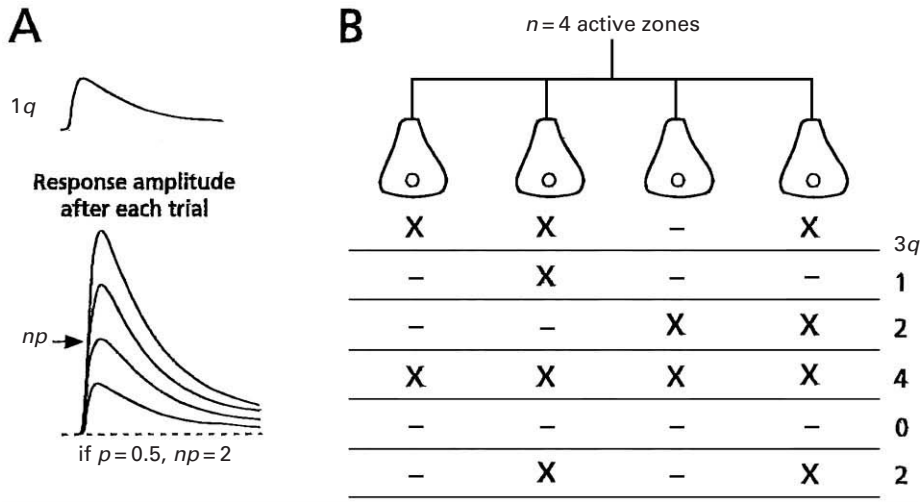

Figure 3
Binomial model of transmitter release

of view, the most novel feature is that communication across central synapses is not guaranteed systematically even when p is relatively high: after each impulse, a synapse flips a coin. This probabilistic aspect is exemplified in Fig. 3A, 3B, which stress that each synaptic unit operates in an 'all or none' manner. It also shows that over a large number of trials the average response amplitude equals the $n \cdot p \cdot q$, product, where n and p are presynaptic parameters, and q, the size of a quantum, is proportional to the number of channels opened by a single exocytotic event. This product defines the so-called synaptic weight or strength of a connection between two neurons, which is a dynamic term, as it can be modulated pre- or postsynaptically. Indeed, persistent increases in synaptic weights are considered to underlie higher-order processes such as learning and memory, from invertebrates to primates. For example, activity-dependent long-term potentiation (LTP; Bliss and Collingridge 1993) in mammalian hippocampus, which is considered a well-suited candidate for a memory mechanism, can result from an increase in any one of the three quantal parameters (Stevens 1993). The reverse form of plasticity is long-term depression (LTD; Bear and Abraham 1996) (first described in the cerebellum), which may be a cellular correlate of erasing memory traces. It presumably represents a second mechanism for allowing neurons to respond preferentially to selected subsets of their inputs.

Since a neuron may receive information from thousands of junctions, the binary mode of transmission illustrated in Fig. 2A may be better suited for computational purposes than having a graded process at each site, which would lead to excess or redundant information. Along this line, a synapse can be viewed metaphorically as a symbolic communication system with, according to the scheme of Shannon and Weaver, the encoding process localized to the release site. Knowledge of p and n then allows calculation with the classical entropy equation of the amount of information, in bits, at a connection between neurons. For example, with a given value of p, the increased reliability of neuronal connections with larger ns is counterbalanced by a decreased number of bits per quantum (Korn and Faber 1987).

5. *Diversity of Synapses*

Central synapses exhibit a great deal of structural and functional heterogeneity, which is built upon the scaffolding described above. For example, the transmitter at adjacent contacts on to a given postsynaptic could be an amino acid such as glycine or acetylcholine, a biogenic amine such as serotonin or dopamine, or a peptide such as somatostatin, substance P, or Vasoactive Intestinal Peptide (VIP), and, furthermore, combinations of these substances may be packaged in one terminal ending (Hokfelt et al. 1984). The consequences of this colocalization and of the cotransmission that is often associated with it are not yet resolved, particularly in puzzling cases where two classical transmitters with the same function can be released together (Jonas et al. 1998). In addition, there are multiple receptor subtypes for most transmitters, and more than one may be colocalized to the same postsynaptic locus (Somogyi 1998). Some are ionotropic, with binding causing channel opening or closing, while others are metabotropic, which means that the ligand-bound receptor activates an intracellular metabolic cascade which can modulate kinases or channel activity. In fact, listing the differences

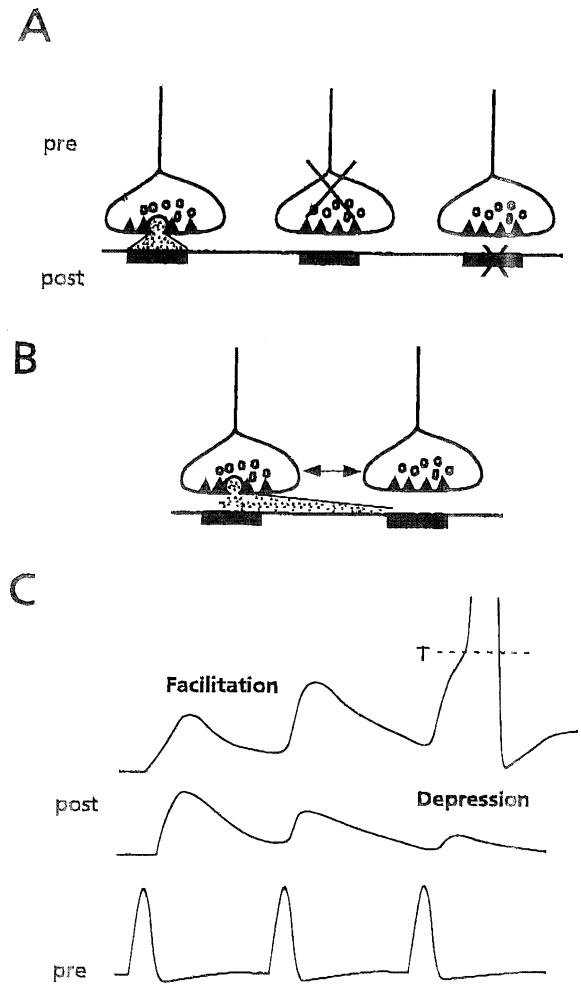

Figure 4
'Latent' synapses and mechanisms of short-term modifications of synaptic efficacy

between synapses is almost an insurmountable task. One would first have to consider constitutive variations in pre- or postsynaptic parameters, namely the probability of release, the kinetics and conductances of single channels. The mechanisms for limiting the duration of action of a transmitter such as its diffusion in the synaptic cleft, its uptake by activated terminals or glia, its chemical inactivation, and receptor desensitization once the channels have opened, add another layer of complexity (Johnston and Wu 1995). One then would have to factor in the synergistic or antagonistic effects of the growing inventory of modulators (Kaczmarek and Levitan 1987), which can influence these various functions in normal and pathological conditions, as well as during development. This diversity may confer a higher degree of functional uniqueness to individual synapses and neurons than is commonly recognized. In any case, it influences the richness of animal behaviors and higher functions which result from activity in complex neuronal assemblies that are synchronized via their synaptic connectivity.

Some elements of synaptic diversity and their functional implications deserve mention. First, although most junctions are fully operational in the normal brain, some of them may remain silent, either because a presynaptic action potential does not evoke transmitter release, or because the postsynaptic receptors are unresponsive (Fig. 4A). These may be viewed as 'latent' rather than 'silent' synapses, because they may be turned on in specific circumstances, such as during learning (Malenka and Nicoll 1997). Second, by virtue of the close packing of synapses, the transmitter released at one of them can diffuse laterally and interact with the postsynaptic receptors of its neighbors (Fig. 4B). This spillover may violate the principle of independence of individual junctions and result in synergistic interactions (Korn et al. 1990), and it may contribute to phenomena such as LTP. Finally, synaptic strength can vary on a millisecond timescale when a presynaptic neuron discharges repetitively (Fig. 4C), and this plasticity can be expressed as either a facilitation or depression (Johnston and Wu 1995, Markram and Tsodyks 1996). Both processes are presynaptic in origin, due to a corresponding change in the release probability, and typically meet the functional requirements of the related networks. For example, facilitation may contribute to recruit neuronal targets that are otherwise weakly excited.

6. *Quantal Analysis and its Limitations*

A statistical dissection of postsynaptic electrical recordings, based on the quantal hypothesis, remains an unsurpassed tool for understanding how different classes of nerve cells communicate with each other. The models used, and their corresponding mathematical formulations, are the same as used in game theory and many branches of the social, behavioral and political sciences; that is, the laws of chance. Stated otherwise, the release of a packet of transmitter after an impulse is equivalent to rolling dice. Despite their apparent complexity, these laws are quite straightforward. For all models, the average number of units released over a large series of trials, which is called *mean quantal content* or m,

$$p_1 + p_2 + \ldots p_n = \sum_{i=1}^{n} p_i$$

where p_1 is the release probability at each site. This is the essence of the *compound binomial release model*, which encompasses all others that assume independent probabilities of release.

If there were no ambiguities about *quantal size*, q, i.e., the magnitude of the postsynaptic response to a quantum, amplitude distribution histograms of evoked responses would have discrete peaks corresponding to transmission *failures*, single quantum responses, doublets and so on, and these distributions could be described with one of the quantal models.

For a compound binomial:

$$\begin{aligned} p_x = {} & p_1 p_2 \ldots p_x (1-p_{x+1}) \ldots (1-p_n) \\ & + p_1 p_3 \ldots p_{x+1} (1-p_2)(1-p_{x+2}) \ldots (1-p_n) \\ & + (1-p_1) \ldots (1-p_{n-x})(p_{n-x+1}) \ldots p_n, \end{aligned}$$

where each line represents one way to get x successes and $(n-x)$ failures. For example, in the first line, units 1 to x release quanta, whereas the other units ($x+1$ to n) do not. The probability of failure for one quantal unit is $(1-p_1)$.

For a simple bionomial:

$$p_x = \left(\frac{n}{x} \right) p^x (1-p)^{(n-x)}$$

where

$$\left(\frac{n}{x} \right) = \frac{n!}{(n-x)!x!}$$

gives the number of combinations by which x quanta can be released out of an available pool of size n.

For a Poisson:

$$p_x = \frac{e^{-m} m^x}{x!}$$

with $m = np$

As developed elsewhere (Martin 1977), most investigators have relied on the two latter models and on indirect methods derived from them rather than trying to fit the distributions. These methods involve either counting the number of failures, n_0, in N trials or measuring the coefficient of variation, CV, of the amplitude distribution, since according to a Poisson:

$$\begin{aligned} m &= -\ln n_0/N \text{ (method of failures)} \\ &= \mathrm{CV}^2 \text{ (coefficient of variation method).} \end{aligned}$$

In rare cases, m can be determined directly, by counting quanta or by averaging miniature events arising from the same synapses. Corresponding expressions allow the last two methods to be applied to a binomial process if p is finite (see Martin 1977). Also, since the binomial $\mathrm{CV} = ((1-p)/np)^{1/2}$ and is independent of q, it has been used as a tool for identifying the pre- vs. postsynaptic locus of a change in synaptic strength, but this approach is liable to misinterpretations (Faber and Korn 1991).

Correct application of these tools requires the accurate identification of the quantal response, which is often difficult at central connections due to quantal variance, distortion by cable properties, and instrumental noise. As a consequence, discrete peaks in amplitude histograms are blurred and failures, single quantal responses, doublets, etc., cannot be distinguished. Various fitting algorithms have been developed, with some success (reviewed in Korn and Faber 1987; see also Stricker et al. 1994), their limitations, however, leaving some questions unresolved. One is the size of the quantum, which is related to the number of channels opened by an exocytotic event, and to their unitary conductance, and can be quite small at central contacts (Finkel and Redman 1983, Edwards et al. 1990). Another is that at some junctions there may be a mismatch between the large number of transmitter molecules released and the relatively few receptors available postsynaptically, which led to the suggestion of a saturated response (Jack et al. 1981). If so, the principles of quantal analysis would no longer hold, as quantitization would be imposed post- rather than presynaptically.

A related issue is the difficulty in determining the pre- or postsynaptic locus of the increased synaptic weight that occurs during LTP, using quantal analysis alone. Raging controversies have not yet been entirely resolved, and mechanisms on both sides of the synapse have been invoked (see Stevens 1993, Liao et al. 1995, Isaac et al. 1996). One of the most intriguing proposals is that although the induction process is postsynaptic, its maintenance can be expressed as an increased release probability, which requires signaling in the reverse direction across the synapse, by a retrograde messenger (reviewed in Bliss and Collingridge 1993). This debate could be obscuring the essential notion of the unity of the synapse where pre- and postsynaptic regulatory factors could be implicated separately or in unison, depending upon functional requirements.

7. *Other Avenues for Synaptic Studies*

Like all scientific fields, synaptic studies cannot remain confined to their historical roots, such as statistics, and need to be enriched by other approaches. The most direct is imaging of dynamic changes at the synapse, associated with the ability to tag pre- and postsynaptic structural components of activated junctions with fluorophores (Malgaroli et al. 1995). In addition, computer-aided modeling of single synapses, or of populations, is necessary to understand biophysical properties of vesicular exocytosis, the fate of the released molecules in the cleft, and their interactions with the postsynaptic receptors, all of which govern the size and kinetics of the postsynaptic response (Bartol et al. 1991, Kruk et al. 1997). Together with

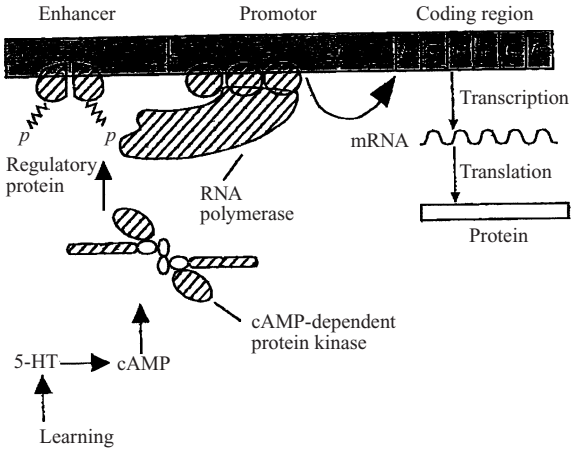

Figure 5
Modulation of gene expression during learning

anatomical correlates, these techniques should not only solve issues such as those related to single vs. multivesicular release and receptor saturation, but also deepen our understanding of brain development, and learning and memory, and how synapses are affected in pathological conditions. Along this line, it can already be envisioned how synapses contribute to genetic or acquired illnesses of behavior and cognition, including psychiatric disorders (Kandel 1998). Figure 5 illustrates the manner in which expression of a gene, leading to the production of a specific protein, could be altered following synaptic activation associated with learning. Conversely, genetic alterations can lead to significant psychiatric disorders, and it can be postulated that they do not leave synapses untouched. This example shows once more that synaptic studies represent a bridge between the neural sciences, cognitive psychology, and behavioral sciences.

See also: Ion Channels and Molecular Events in Neuronal Activity; Long-term Potentiation (Hippocampus); Memory: Synaptic Mechanisms; Neural Plasticity; Neurons and Dendrites: Integration of Information; Neurotransmitters; Synapse Formation; Synapse Ultrastructure; Synaptic Efficacy, Regulation of

Bibliography

Bartol T M Jr., Land B R, Salpeter E E, Salpeter M M 1991 Monte Carlo simulation of miniature endplate current generation in the vertebrate neuromuscular junction. *Biophysicical Journal* **59**: 1290–307

Bear M F, Abraham W C 1996 Long-term depression in hippocampus. *Annual Review of Neuroscience* **19**: 437–62

Bennett M V L 1977 Electrical transmission: A functional analysis and comparison to chemical transmission. In: Brookhart J M, Mountcastle V B, Kandel E R, Geiger S (eds.) *Handbook of Physiology: The Nervous System*. Williams and Wilkins, Baltimore, MD, pp. 357–416

Bliss T V, Collingridge G L 1993 A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature* **361**: 31–9

Couteaux R, Pecot-Dechavassine M 1970 Vesicules synaptiques et poches au niveau des zones actives de la jonction neuromusculaire. *Comfetes Rertus le l'Academic des Sciences* **271**: 2346–9

Del Castillo J, Katz B 1954 Quantal components of the end-plate potential. *Journal of Physiology* **124**: 560–73

Eccles J C 1964 *The Physiology of Synapses*. Springer-Verlag, Berlin

Edwards F A, Konnerth A, Sakmann B 1990 Quantal analysis of inhibitory synaptic transmission in the dentate gyrus of rat hippocampal slices: A patch-clamp study. *Journal of Physiology (London)* **430**: 213–49

Faber D S, Korn H 1989 Electrical field effects: Their relevance in central neural networks. *Physiological Review* **69**: 821–63

Finkel A S, Redman S J 1983 The synaptic current evoked in cat spinal motoneurones by impulses in single group la axons. *Journal of Physiology* **342**: 615–32

Heuser J E, Reese T S, Landis D M D 1974 Functional changes in frog neuromuscular junctions studied with freeze-fracture. *Journal of Neurocytology* **3**: 109–31

Hille B (ed.) 1984 *Ionic Channels of Excitable Membranes*. Sinauer Associates, Sunderland, MA

Hokfelt T, Johansson O, Goldstein M 1984 Chemical anatomy of the brain. *Science* **225**: 1326–34

Isaac J T, Hjelmstad G O, Nicoll R A, Malenka R C 1996 Long-term potentiation at single fiber inputs to hippocampal CA1 pyramidal cells. *Proceedings of the National Academy of Science USA* **93**: 8710–15

Jack J J, Redman S J, Wong K 1981 The components of synaptic potentials evoked in cat spinal motoneurones by impulses in single group Ia afferents. *Journal of Physiology (London)* **321**: 65–96

Johnston D, Miao-Sin Wu S (eds.) 1995 *Foundations of Cellular Neurophysiology*. MIT Press, Cambridge, MA

Jonas P, Bischofberger J, Sandkuhler J 1998 Corelease of two fast neurotransmitters at a central synapse. *Science* **281**: 419–24

Jonas P, Major G, Sakmann B 1993 Quantal components of unitary EPSCs at the mossy fibre synapse on CA3 pyramidal cells of rat hippocampus. *Journal of Physiology (London)* **472**: 615–63

Kaczmarek L K, Levitan I B (eds.) 1987 *Neuromodulation: The Biochemical Control of Neuronal Excitability*. Oxford University Press, Oxford, UK

Kandel E R 1998 A new intellectual framework for psychiatry. *American Journal of Psychiatry* **155**: 457–69

Katz B (ed.) 1969 *The Release of Neural Transmitter Substances*. Charles C. Thomas, Springfield, IL

Korn H, Faber D S 1987 Regulation and significance of probabilistic release mechanisms at central synapses. In: Edelman G M, Gall W E, Cowan W M (eds.) *Synaptic Function*. A Neuroscience Institute Publication. Wiley, NY, pp. 57–108

Korn H, Faber D S 1991 Quantal analysis and synaptic efficacy in the CNS. *Trends in Neuroscience* **14**: 439–45

Korn H, Faber D S, Triller A 1990 Convergence of morphological, physiological, and immunocytochemical techniques for the study of single Mauthner cells. In: Bjorklund A, Hokfelt T, Wouterlood F G, van den Pol A N (eds.) *Handbook of Chemical Neuroanatomy, Vol. 8: Analysis of Neuronal*

Microcircuits and Synaptic Interactions. Elsevier Science Publishers, Amsterdam, pp. 403–80
Kruk P J, Korn H, Faber D S 1997 The effects of geometrical parameters on synaptic transmission: A Monte Carlo simulation study. *Biophysics Journal* **73**: 2874–90
Kuffler S W, Yoshikami D 1975 The number of transmitter molecules in a quantum: An estimate from iontophoretic application of acetylcholine at the neuromuscular synapse. *Journal of Physiology (London)* **251**: 465–82
Liao D, Hessler N A, Malinow R 1995 Activation of postsynaptically silent synapses during pairing-induced LTP in CA1 region of hippocampal slice. *Nature* **375**: 400–4
Llinas R R 1982 Calcium in synaptic transmission. *Scientific American* **4**: 56–65
Llinas R, Nicholson C, Precht W 1969 Preferred centripetal conduction of dendritic spikes in alligator Purkinje cells. *Science* **163**: 184–7
Malenka R C, Nicoll R A 1997 Silent synapses speak up. *Neuron* **19**: 473–6
Malgaroli A, Ting A E, Wendland B, Bergamaschi A, Villa A, Tsien R W, Scheller R H 1995 Presynaptic component of long-term potentiation visualized at individual hippocampal synapses. *Science* **268**: 1624–8
Markram H, Tsodyks M 1996 Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature* **382**: 807–10
Martin A R 1977 Junctional transmission. II: Presynaptic mechanisms. In: Brookhart J M, Mountcastle V B, Kandel E R, Geiger S (eds.) *Handbook of Physiology: The Nervous System*. Williams and Wilkins, Bethesda, MD, pp. 357–416
Palay S L 1958 The morphology of synapses in the central nervous system. *Experimental Cell Research* **5**: 275–93
Peters A, Palay S L, Webster H DeF (eds.) 1991 *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*. Oxford University Press, Oxford, UK
Pfenninger K H, Akert K, Moor H, Sandri C 1972 The fine structure of freeze-fractured presynaptic membranes. *Journal of Neurocytology* **1**: 129–49
Redman S J 1990 Quantal analysis of synaptic potentials in neurons of the central nervous system. *Physiological Reviews* **70**: 165–98
Somogyi P 1998 Precision and variability in the placement of pre- and postsynaptic receptors in relation to transmitter release sites. In: Faber D S, Korn H, Redman S J, Thompson S M, Altman J S (eds.) *Central Synapses: Quantal Mechanisms and Plasticity*. pp. 82–95
Stevens C F 1993 Quantal release of neurotransmitter and long-term potentiation. *Neuron* **10**: 55–63
Stricker C, Redman S, Daley D 1994 Statistical analysis of synaptic transmission: Model discrimination and confidence limits. *Biophysics Journal* **67**: 532–47
Triller A, Cluzeaud F, Pfeiffer F, Betz H, Korn H 1985 Distribution of glycine receptors at central synapses: An electron microscopy study. *Journal of Cell Biology* **101**: 683–8

H. Korn and D. S. Faber

Syncretism

The meaning of the term 'syncretism' has changed drastically over the course of time. Currently, the term refers to the mixing together of elements from different religions. This mixing is sometimes viewed with contempt, especially by the clergy of the religions concerned. Scholarly use of the term is more objective and does not entail this kind of value judgement. In this article we shall describe the history of the term. Special attention will be given to current debates on power mechanisms in syncretism. Lastly, syncretism will be compared with fundamentalism.

1. *History of the Term 'Syncretism'*

1.1 *Earliest Uses*

The first known use of the term 'syncretism' is found in Plutarch's (46–125) *Moralia* (Morals), a treatise on moral principles. When suggesting that one should consider one's brother's enemies and friends as one's own enemies and friends, he uses the example of the inhabitants of Crete. Plutarch uses the word syncretism to refer to their habit of forgetting internal conflicts and differences when confronting a common enemy. It may be that the use of the word was inspired not only by the Cretans' example but also by a word play, since the Greek *syngkrasis* means a 'mixing together' (Stewart and Shaw 1994).

The second author known to have used the term was the humanist Erasmus of Rotterdam (1496–1536). He quoted Plutarch and, like him, recommended syncretism, although here as a way of seeking a synthesis of seemingly disparate philosophical and theological opinions and as an attempt to understand the other completely, despite differences of view.

After the Reformation, with its increase in divergent theological views, the term acquired a negative connotation for the first time. Opponents of the German Lutheran theologian George Calixtus (1586–1656), an early defendant of ecumenism and theological reconciliation, used the term pejoratively as a mixing of points of view that, in their opinion, was to be avoided. The debate became known as the 'syncretistic controversy.'

1.2 *Views from Religious Studies and Missionary Practice*

In the nineteenth century, when research on world religions began to flourish and colonialism brought cultures and religions into contact, syncretism was understood more objectively as the mixing of elements from different religions, although purity continued to be viewed implicitly as the norm (Rudolph 1979).

Christianity also reached the height of its expansion in the nineteenth century. Missionaries objected to the tendency of new converts to continue practices from their prior religion. Syncretism then became the concept used for this—condemned and combated—tendency. This meant that Christian theological use of

the term was no longer restricted to internal differences between theologians but referred to the mixing of religious elements, especially Christian and non-Christian. Unlike its usage in religious studies, syncretism was considered censurable and illegitimate. The term still retains this meaning, even though ecumenical missionary circles have become more understanding of the need to translate the Christian message into local cultural terms (so-called inculturation; see Schreiter 1993). It has also been suggested that Christianity itself is the result of a syncretistic process (Pannenberg 1970). Yet, despite this development, the theological criterion continues to be how intact the core message of Christianity remains.

For a long time scholars in religious studies have worked with the term and defined its meanings (Rudolph (1979) offers a useful summary), whereas, as will be shown in the next section, the term has a much shorter history in anthropology. Scholars in the field of religious studies, such as historians and phenomenologists of religion, view syncretism objectively and point to its general presence at the founding of new religions (even in religions that now oppose syncretism) and in their mutual contact. Scholars distinguish between symmetrical and equal relationships between religions on the one hand and asymmetrical and unequal ones on the other. Typologies of the degree to which elements can be mixed have been proposed, using terms such as identification, amalgamation, and symbiosis. It has been acknowledged that syncretism may involve elements other than religious ones, since religions are linked to cultural contexts.

The basic ambiguity of meanings in syncretistic contexts has been emphasized (Pye 1971). Sometimes syncretism is presented as the result but more often as the process itself. Syncretism is presented as part of popular religion, while the representatives of official religion generally condemn it. Correspondingly, syncretism has often been depicted as unconscious, implicit and spontaneous, as typical of what Redfield called the little tradition of the unreflective many (Rudolph 1979).

There have also been proposals to excise the term from scholarly vocabulary (Baird 1971, Schineller 1992, Van der Veer (1994) in Stewart and Shaw 1994) for being too vague, subjective, and too broad. If syncretism occurs in all religions and if borrowing and blending are universals in any event, the term seems to be insufficiently specific and therefore superfluous. Moreover, virtually no religious person sees himself or herself as syncretistic. However, terms are not easily eliminated from scholarly vocabularies. Precise definition or even redefinition may resolve this dilemma. In view of the increase in contact between believers of different religions, some term is needed to refer to the inevitable mixing of religious elements. In addition, the long history of the term has shown that its meaning may change.

1.3 Views from Cultural Anthropology

While students of religion from outside cultural anthropology focused on world religions, the first anthropological texts on syncretism referred to Latin American and Caribbean examples, especially the Afro-Christian syncretism that was followed from the presence of slaves and their descendants in the region. Of course, syncretism was present in other areas, but where the negative missionary use of the term was the norm, as in Africa, anthropologists avoided using the term, even with respect to obvious cases of syncretism, as occurred in the so-called African independent churches founded by Africans and not by missionaries (Stewart and Shaw 1994).

When the American anthropologist Melville Herskovits began using the term in anthropology, following his Brazilian student Arthur Ramos, he took his examples from Afro-American religions (Greenfield and Droogers 2001). His interest in the topic was partly ideological since he understood syncretism as a tool for promoting a unified culture in the United States, to be formed from the many cultures that immigrants, including those who had involuntarily come from Africa, had brought with them. This melting-pot idea was, of course, a way of rehabilitating the Afro-American strand in US culture over against white racism. Moreover, Herskovits saw syncretism as a half-way stage in the process of acculturation, the intermingling of cultures in contact. He was very much interested in the way people gave meanings to selected old and new elements, mixing cultural standards, and thus initiating change towards a new synthesis.

The alternative course is an emphasis on authenticity and a return to tradition and to the past, as in the case of the black consciousness movement. Nourished by the political ideal of integration the concept of acculturation may have suggested a one-sided linear evolution.

Another scholar who used the term syncretism in the Latin American context was the French ethnologist Roger Bastide (Bastide 1978). On the basis of fieldwork in Afro-Brazilian religions, he emphasized the systematic way in which analogous elements from different religious sources are mixed. He also pointed to the role of power mechanisms, especially in the contact between slaves and their masters. The structural similarity between African, Catholic, and Amerindian worldviews facilitated syncretism. Thus African gods could be identified with Catholic saints and Amerindian spirits. In the oppression that the slaves suffered, supernatural help was most welcome. This redressed the balance of power somewhat, such as when masters were the object of slave rituals intended to harm them. The redress of this balance also became clear when Catholic elements were selectively adopted through the application of African criteria, and not vice versa, as the masters wished when they baptized their slaves. In practice, as a strategic

device, identification with and differentiation from the masters' religion were both used in a double game. Thus the apparent adoption of a Catholic ritual attitude in Mass could serve as an alibi for the continuation of African ritual practices.

2. Current Issues

2.1 Power

The political use of the concept of syncretism has received more attention in the 1990s (Droogers 1989, Stewart and Shaw 1994). The negative sense of the term is an indication of a conflict between clergy and laity on the right to produce legitimate religion. Leaders of a religion whose adepts are involved in a syncretistic process will, from a purist point of view, use their power to condemn any effort to mix elements from different religions. Stewart and Shaw (1994) accordingly proposed the term antisyncretism.

As suggested above, the subjective meaning of the term could be a reason for abandoning the term in scholarly use. Yet it has the advantage of revealing the power dimension of the syncretism process. In that sense, syncretism is one of a number of terms used in anthropology despite their pejorative meanings, such as magic or popular religion.

An attitude of antisyncretism is not only held by leaders of affected religions but can also be found in representatives of religions that are the result of a process of syncretism. Thus some leaders of the very syncretistic Afro-Brazilian religions adopt an anti-syncretistic attitude, favoring a return to the African roots and the removal of all Catholic influence.

2.2 Syncretism and Fundamentalism

In various senses syncretism and fundamentalism are opposites. Whereas syncretism uses more than one religious source, fundamentalism defends the uniqueness of the one source, often codified in a sacred text. Syncretism implies that more versions of the experience with the sacred are acceptable and can be combined, whereas fundamentalists defend one exclusive version. Syncretists experiment with religious symbols, as many as seem to be applicable to their situation, whereas fundamentalists jealously guard the one central symbol of their religion. Syncretism is usually criticized by religious specialists who feel that they are losing control and who may react by defending a fundamentalist view. Syncretists interpret the opportunity offered by the growing contact between cultures and religions as a chance to combine religious repertoires for thought and action, whereas fundamentalists react to the moral side effects of this contact by retreating to the traditional repertoire, even though they may make use of modern means of communication to spread their message. Both syncretism and fundamentalism seem to flourish as a consequence of modernization and globalization but for different reasons: one embracing the new situation, the other being critical of it because it is felt to threaten cherished traditions and morals. Syncretists behave as global citizens, whereas fundamentalists divide the world into 'us' and 'them.' Syncretists are not organized and rarely present themselves as such, whereas fundamentalists tend to organize themselves and to occupy power positions.

See also: Creolization: Sociocultural Aspects; Cultural Assimilation; Cultural Contact: Archaeological Approaches; Hybridity; Mission: Proselytization and Conversion

Bibliography

Baird R D 1971 *Category Formation and the History of Religion*. Mouton, The Hague, The Netherlands

Bastide R 1978 *The African Religions of Brazil, Toward a Sociology of the Interpenetration of Civilizations*. Johns Hopkins University Press, Baltimore, MD

Droogers A 1989 Syncretism: The problem of definition, the definition of the problem. In: Gort J D, Vroom H M, Fernhout R, Wessels A (eds.) *Dialogue and Syncretism, an Interdisciplinary Approach*. Eerdmans and Rodopi, Grand Rapids, MI, pp. 7–25

Greenfield S, Droogers A (eds.) 2001 *Reinventing Religions: Syncretism and Transformation in Africa and the Americas*. Rowman & Littlefield, Lanham, MD

Pannenberg W 1970 *Basic Questions in Theology*. SCM, London

Pye M 1971 Syncretism and ambiguity. *Numen* **18**: 83–93

Rudolph K 1979 Synkretismus vom theologische Scheltwort zum religionswissenschaftlichen Begriff. *Humanitas Religiosa: Festschrift für Haralds Biezais zu seinem 70. Geburtstag*. Almqvist & Wiksell, Stockholm, pp. 193–212

Schineller P 1992 Inculturation and syncretism: What is the real issue? *International Bulletin of Missionary Research* **16**: 50–3

Schreiter R J 1993 Defining syncretism: An interim report. *International Bulletin of Missionary Research* **17**: 50–3

Stewart S, Shaw R (eds.) 1994 *Syncretism/Anti-syncretism, the Politics of Religious Synthesis*. Routledge, London

Van der Veer P 1994 Syncretism, multiculturalism and the discourse of tolerance. In: Stewart S, Shaw R (eds.) *Syncretism/Anti-syncretism, the Politics of Religious Synthesis*. Routledge, London, pp. 196–211

A. Droogers

Syndromal Diagnosis versus Dimensional Assessment, Clinical Psychology of

A syndrome is a pattern of symptoms occurring in a disease; a diagnosis implies history and prognosis as well. Dimensional assessment refers to the severity of the associated symptoms or characteristics. In 1895

the countries of Europe met to define the important diseases. They could not agree, but in 1905 when they next met, they just had to agree to define diseases important to public health, if only to be able to share information about prevention and treatment. Progress is possible only when all share a common language, and definition of disorders has been as much product of practical need as the result of informed research. In clinical psychology, the dominant classifications are now the American Psychiatric Association's Diagnostic and Statistical Manual, 4th revision (DSM-IV) (1994) and the World Health Organization's International Classification of Disease, 10th revision (ICD-10) (1993). Both are lists of characteristics (symptoms and durations, and other diagnoses that preempt) but neither routinely includes history or cause. A dimensional assessment is particular to an individual, a fine grained description of symptoms and other attributes currently present. Both should inform about clinical picture and beneficial treatments: the syndromal diagnosis in respect to what is known about others with the same complaint; the dimensional assessment in respect to this individual.

Historically there was a clear distinction—when doctors made diagnoses they said this person belongs to a category, and from that category one could infer etiology, natural history, and response to treatment; whereas when psychologists completed a dimensional assessment (see Minnesota Multiphasic Personality Inventory) they could profile the person's strengths and weaknesses in respect to the normative values for the general population. A diagnosis is of value because it pulls together the ways that symptoms reflect a class phenomenon. A dimensional assessment is of value because it highlights the particular issues in an individual case. To some extent the dichotomy posed in the title of this article is an artifact, for good clinicians who make a syndromal diagnosis go on to make a formulation in which the individual characteristics are taken into account, and good clinicians who make a dimensional assessment acknowledge the syndromal information that might inform prognosis and treatment. For most, the syndromal diagnosis informs the broad choice of treatment, while the dimensional assessment is the key to measuring progress.

Syndromal diagnosis is not straight forward. It varied until the American Psychiatric Association in 1980 and the World Health Organization in 1990 published detailed diagnostic criteria for each mental disorder (see *Mental and Behavioral Disorders, Diagnosis and Classification of*). Prior to this there were significant variations between countries in the way that serious mental disorders like schizophrenia or manic depressive illness were diagnosed, and probably even more intercountry variations in the way that the nonpsychotic anxiety, depressive, substance use, and somatoform disorders were identified, although these variations were rarely studied. The current DSM (APA 1994) and ICD (WHO 1993) criteria have resulted in clinicians being able to agree much of the time. But as Meehl had originally remarked and Kendell (1989) has affirmed, two clinicians will vary in the way they ascertain the symptoms, and in the way they interpret and weight them toward the criteria, so that disagreement, even given an agreed classification, will still occur. There are no validated laboratory or x ray tests for any mental disorder and no gold standard method of determining diagnosis. Diagnosis depends solely on the pattern of illness complained of by the patient or elicited by the clinician. It is becoming increasingly clear that the present diagnostic criteria, largely based on the work of Kraepelin in Europe at the beginning of the twentieth century, are but surface markers for hypothetical underlying concepts about which clinicians agree. In Jablensky's (1999) view, we have a *nomenclature*, not a logical classification of mental disorders. The present classifications, even though sanctified by tradition, and now reliable and stable, may not even be correct.

The ICD and DSM classifications, while differing in minor ways, seem as though they should function equivalently. The advent of agreed diagnostic criteria has meant that fully structured diagnostic interviews could be developed (e.g., the Composite International Diagnostic Interview or CIDI) and as these interviews address both DSM and ICD classifications, the correspondence between the two classifications could be examined. The results are interesting (Andrews et al. 1999). Only two-thirds of people who met criteria for a diagnosis on either system met criteria on both DSM and ICD classifications, the agreement varying from a high of 85 percent for depression and dysthymia to a low of 35 percent for substance abuse and post-traumatic stress disorder. Most of the disagreements appeared to come from the way the criteria were defined and operationalized, and not from any basic disagreement about the underlying concept. Epidemiological studies show that the demographic correlates among people meeting criteria for a diagnosis of a mental disorder—age under 45, being female, currently unmarried, unemployed, and of lower socioeconomic status—were the same in both classifications, as though the cases identified by either classification shared similar underlying characteristics.

Nevertheless, such demographic correlates have lead some to question the validity of the concept of mental disorders—'Isn't it understandable that the disenfranchised have symptoms of anxiety, depression and somatic dysfunction?' Is mental illness therefore a social construct? Thomas Szasz (1961), appalled by the ease with which a supposition of psychosis could lead to involuntary hospitalization, argued that it was a social construct. The surprising feature is constancy of the syndromic nature of the disorders; psychosis has similar symptoms and natural history across cultures, in the developing and in the developed world, in people with and without the listed social correlates.

Major depressive disorder, agoraphobia, or neurasthenia are equally invariant in their symptoms and natural history across countries and across peoples. This invariant pattern is not evidence of a stereotyped response to stress, for contrary to popular belief, stress plays a very small part in mental disorders. It looks as though these are the patterns of symptoms to which we are heir.

How important are the mental disorders? The World Bank began a project to estimate the important consequences of all diseases by adding the years of life lost due to a disease to the years lived with a disability, weighted for severity. Worldwide say Murray and Lopez (1996), mental disorders, as currently defined, account for 9 percent of the total burden of disease. In established market economies like North America, Europe, and Australasia, they account for 22 percent of the total burden, not because there is an epidemic of mental disorders in those countries, but because their relative importance has increased as the burden due to infection and chronic physical disease has declined. Mortality does not contribute a great amount, the burden of mental disorders largely comes from the morbidity—the years lived with a disability. Among the mental disorders, the anxiety and depressive disorders account for half this morbidity, not so much because they are so disabling, but because they are common, whereas the psychoses, which are more disabling, are relatively rare, and so account for less of the total morbidity. In total, mental disorders are an important cause of human capital loss.

Syndromal diagnoses are important for political reasons. They are essential if epidemiological studies are to answer the usual questions asked by health planners—how many people have which disorders, how disabled are they by these disorders, what does it cost to treat them, and what services should we provide? Reliable case identification is also critical if therapeutics is to advance. Randomized controlled trials of a new remedy against placebo or established treatment are the only way that the efficacy of new treatments can be assessed. Being able to specify the diagnostic class to which the patients belong is critical, both to the scientific question, and to the marketing of the remedy once it has been shown to be efficacious. Last, funders often no longer pay for the treatment of a person, but pay on the basis of the syndromic diagnosis, i.e., a set casemix fee for a hospital to treat a new case of schizophrenia, or a set number of visits or set fee for a clinician to provide ambulatory treatment for a case of depression. Classification by syndromal diagnosis is essential to all these activities—defining the unmet need, identifying and marketing the treatments that work, and reimbursing the clinicians who provide the service. The whole edifice of mental health care would collapse were we to remove syndromal diagnosis even though we have no proof that these diagnoses are fundamentally valid.

There are a number of problems that stem from accepting the present definitions. The first problem is the seemingly arbitrary nature of the syndrome definitions themselves. In part, the diagnosis of a major depressive episode requires the presence of five of nine symptoms known to occur in depressive illness. But people with four symptoms—depression, loss of interest, inability to concentrate, feelings of worthlessness—would not meet criteria even though they could share the same demographic correlates, response to treatment, and level of disability as did some people who did meet criteria. The diagnosis of schizophrenia in part depends on a disturbance persisting for at least six months characterized by the presence of two of five symptoms of psychosis during a one month period with concomitant social/occupational dysfunction. If the disorder has been present for less than a month it is classified as Brief Psychotic Disorder, if present for one month to five months it is classified as Schizophreniform Disorder, and when it has continued for six months it becomes schizophrenia. While there are practical reasons that relate to cause and outcome that led to this division based on chronicity, it is obvious that it is an arbitrary subdivision of a spectrum of psychotic disorders. Both the ICD and the DSM classifications of mental disorders contain many more examples, of a point on a dimension of symptoms marking a criterion for a disorder.

Many presume that the existence of a specific treatment might validate a disorder, but there is a lack of specificity among current treatments for mental disorders. Given the millions of dollars spent, one would have thought that drug remedies for depression would cure depression specifically. Recent evidence suggests that three-quarters of the benefit from Prozac and other drugs sold as antidepressants is due to spontaneous remission or response to placebo and only one quarter of the observed improvement is due to the specific effects of the drug (Enserink 1999). In most jurisdictions the same drugs have won approval as being efficacious for panic disorder, social phobia, or for obsessive compulsive disorder. Clearly the drugs have little specific action and, at best, may act on some underlying brain mechanism common to depressive and anxiety disorders.

There are a number of additional reasons why the skeptic might doubt the current diagnostic systems. The first problem is that multiple diagnoses are more common than one would expect if the individual syndromes were independent, four times more common in fact (Andrews et al. 1990). At the clinical level one is surprised when a person who was hospitalized with an episode of mania—days of unnatural elation and overactivity coupled with being overtalkative and grandiose, sexually unrestrained, and careless of money and reputation—recovers and is rehospitalized two years later with a diagnosis of schizophrenia—deluded and hallucinated, with disorganized thinking and speech, coupled with a steady occupational and social role deterioration over the past months. No

one should be so unfortunate as to suffer two major mental disorders, yet clinical conferences to review the management plans of difficult to treat patients not infrequently learn that, at various times, a number of different diagnoses have been used to justify treatment. Comorbidity is to be expected if an underlying causal factor is related to both the disorders experienced, or if one disorder is a response to the other, but two or more apparently independent disorders occurring in the same person should be less common than it is.

Comorbidity should refer to the concurrence of two or more disorders (see *Comorbidity*). Structured diagnostic interviews, like the CIDI, ask whether a person has ever had the symptoms that satisfy the criterion for a disorder. Comorbidity has come to refer to people who have satisfied the criteria for more than one disorder, sometime in their lifetime. Comorbidity is common and in a number of surveys it has been shown that people who meet criteria for one disorder, will on average meet criteria for an additional disorder, which is four times the expected risk. This finding is not just due to respondents being prone to say 'yes' to any symptom question. In the anxiety and depressive disorders comorbidity has been related to the presence of a predisposing vulnerability factor of high trait anxiety and poor coping. Sequential effects are also evident, with the stress of one illness acting as a precipitant of additional disorders in vulnerable persons. For example people with agoraphobia are likely to become depressed, and people with social phobia can become dependent on the alcohol that they initially used to manage their disorder. However the comorbidity comes about, people identified as having more than two illnesses in their lifetimes to date are more likely to meet criteria for a current disorder, to be more disabled, and to be and have been very high utilizers of health services. Comorbidity is not just an artifact of measurement, it is also an indicator of severity and unmet need.

Kendell (1989) has revised the original Robins and Guze (1970) criteria for determining whether a clinical syndrome was valid as follows:

(a) identification and description of the syndrome, either by clinical intuition or by statistical analysis;

(b) demonstration of boundaries or 'points of rarity' between related syndromes by discriminant function analysis or latent class analysis;

(c) follow-up studies establishing a distinctive course or outcome;

(d) therapeutic trials establishing a distinctive treatment response;

(e) family studies establishing that the syndrome 'breeds true'; and

(f) association with some more fundamental abnormality—histological, psychological, biochemical, or molecular.

These six criteria have not been achieved in any group of mental disorders—the affective, anxiety, substance use, personality, or psychotic disorders. While identification and description of each syndrome was well advanced by the ICD and DSM classifications, clear boundaries have been difficult to establish, distinctive outcomes are rare, a unique response to one treatment is uncommon, and multiple diagnoses over time are common. The diagnostic classifications have attempted to deal with the last problem by establishing hierarchical rules as to which diagnosis precludes others, but the strategy is neither systematic nor wholly convincing. Information from studies of the human genome might go someway to resolve this impasse by showing discontinuities between genotypes and the accepted phenotypes, and so force a revision of the current classifications (Smoller and Tsuang 1998). Twenty years ago, it was believed that neurochemistry and neuroimaging studies would do this and they haven't. It is not that there are no data, it is just that reliable data are so patchy that it is not yet possible to say that disease X is distinguished by specific information addressing three or more of the Kendell criteria. Until there is, the validity of any syndrome is more a belief than a fact.

Health planners and researchers, clinicians and their patients, and funding agencies, all act as though the diagnostic categories are an established fact. Like the representatives at the International Health Conference in 1905, they have to. However, even good clinicians turn a blind eye when a depressed person has less than the required five symptoms—if, in the light of their history, the person probably will respond to the usual therapy for depression, the diagnosis will be made. Similarly, even if classic symptoms of schizophrenia have lasted for less than six months most clinicians will make the diagnosis and begin antipsychotic drugs. Most health planners will be pleased and so will most funding agencies. Good clinicians already use syndrome specific questionnaires or rating scales of thoughts, feelings and behaviors, and of global measures of disability. They know that the task is to get the person well, that is, no longer troubled by symptoms, and no longer disabled. This dimensional assessment is the key to the measurement of outcome and as outcome assessment is an increasing requirement, it is likely such dimensional assessment will soon rank in importance with the establishment of the syndromal diagnosis. Only with the personality disorders is there a strong move to discard the categories altogether and use a dimensional assessment based on measurement of personality traits as being more accurate and parsimonious.

In summary, the categories of the common mental disorders are well established and accepted and appear to represent the domains of symptoms complained of by people in both developed and developing countries. While health planners, clinicians, and funding agencies pay lip service to these categories, most realize that they are dealing with dimensions of illness on which the categories impose an artificial dichotomy and so will advise or accept the appropriate treatment even if

the person does not strictly meet the criteria. Hopes that statistical analysis of symptom patterns, or distinctions evident in treatment response, or variations in natural history would inform categorical classification have proved illusory. Some think that the advances in molecular genetics might assist in rewriting the topography of mental disorders but even if this is so, most accept that the issue will still be dimensional, 'just how much of disorder X should one cope with before seeking help,' not 'should all people with the gene for disease X be treated.' Moreover the likely response will also be dimensional and range from public education, through self-help strategies and assistance at the primary care level, to specialist mental health care (Andrews 2000). Diagnostic criteria and dimensional assessment are but two sides of the same coin. Good clinicians use both.

See also: Behavioral Assessment; Classification: Conceptions in the Social Sciences; Clinical Assessment: Interview Methods; Clinical Psychology: Validity of Judgment; Comorbidity; Differential Diagnosis in Psychiatry; Mental and Behavioral Disorders, Diagnosis and Classification of

Bibliography

American Psychiatric Association (APA) 1994 *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. (DSM-IV). American Psychiatric Association, Washington, DC

Andrews G 2000 Meeting the unmet need through disease management. In: Andrews G, Henderson S (eds.) *The Unmet Need in Psychiatry*. Cambridge University Press, London

Andrews G, Slade T, Peters L 1999 Classification in psychiatry: ICD-10 versus DSM-IV. *British Journal of Psychiatry* **174**: 3–5

Andrews G, Stewart G, Morris-Yates A, Holt P, Henderson S 1990 Evidence for a general neurotic syndrome. *British Journal of Psychiatry* **157**: 6–12

Enserink M 1999 Can the placebo be the cure? *Science* **284**: 238–40

Jablensky A 1999 The nature of psychiatric classification: Issues beyond ICD-10 and DSM-IV. *Australian and New Zealand Journal of Psychiatry* **33**: 137–44

Kendell R E 1989 Clinical validity. *Psychological Medicine* **19**: 45–55

Murray C J L, Lopez A D 1996 Evidence-based health policy: Lessons from the global burden of disease study. *Science* **274**: 740–3

Robins F, Guze S B 1970 Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry* **126**: 983–7

Smoller J W, Tsuang M T 1998 Panic and phobic anxiety: Defining phenotypes for genetic studies. *American Journal of Psychiatry* **155**: 1152–62

Szasz T 1961 *The Myth of Mental Illness*. Harper and Row, New York

World Health Organization 1993 *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization, Geneva

G. Andrews

Syntactic Aspects of Language, Neural Basis of

Natural languages are built on three different knowledge components: the sound of words, called phonology; the meaning of words, called semantics; and the grammatical rules according to which words are put together, called syntax. This article concerns the neural mechanism underlying the knowledge and use of syntax. Three methods have been used to study the neural basis of syntax: (a) the observation of the correlation between brain lesion and behavior; (b) functional imaging approaches; and (c) event-related brain potential methods. Evidence from these three approaches with respect to syntactic processing will be reviewed and an explanatory model covering major aspects of the data will be suggested.

1. Lesion–Behavior Approaches

Lesion–behavior approaches to investigating the neural basis of language date back to the middle of the nineteenth century (Broca 1861, Wernicke 1874). Using such an approach it was observed that patients with lesions in the temporal region of the left hemisphere (the so-called Wernicke's area, Brodmann's area, BA 22/42) are characterized by poor comprehension, whereas their speech is fluent and often syntactically well structured although semantically empty. Patients with lesions in the left inferior frontal gyrus (the so-called Broca's area, BA 44), in contrast, demonstrate a selective deficit for syntactic aspects in language production and in language comprehension. Patients with lesions in Broca's area produce agrammatic speech, that is utterances containing content words, but lacking most of the grammatical morphology (function words and inflections). They fail to comprehend sentences whenever the correct interpretation depends on the full and adequate processing of the grammatical morphology. In the 1980s a central topic of discussion centered around three possible ways of understanding this behavior. A first view took this behavior to be due to a lack of grammatical knowledge as such (Caramazza and Zurif 1976, Berndt and Caramazza 1980). A second approach interpreted it as being caused by computational deficit either in accessing lexical information or in accessing the syntactic information encoded in function words and inflectional morphology (Bradley et al. 1980). A third view assumed the computational deficit to be based either on reduced memory resources (Kolk and Van Grunsven 1985) or on reduced processing resources leading to a trade-off between syntactic and semantic interpretative processes (Linebarger et al. 1983).

It has been shown that Broca's patients are able to judge a sentence's grammaticality (Linebarger et al.

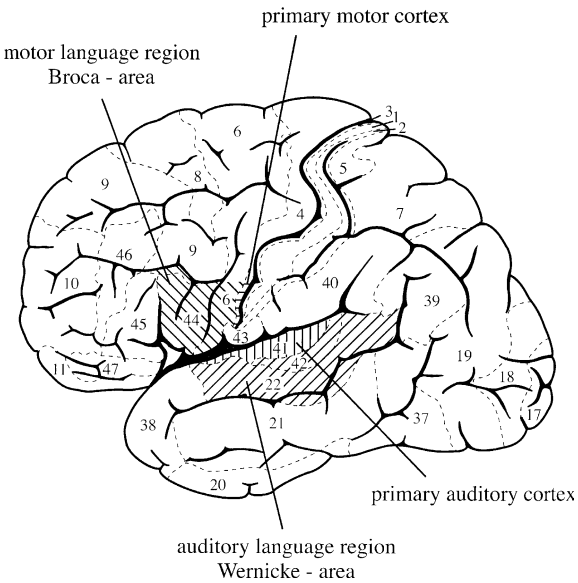

Figure 1
Left hemisphere (numbers refer to Brodmann's cytoarchitectonically defined areas)

1983) and even show grammatical priming (i.e., grammatically correct context sentences enhanced the recognition of a target word compared to ungrammatical context sentences; see Haarmann and Kolk 1991). These findings lead to the assumption that Broca's area may not be the brain area where syntactic knowledge is represented, but rather that this brain tissue supports syntactic computation. While some view this brain region to be the home of very specific syntactic computations necessary to process dislocated elements (e.g., moved constituents that leave behind traces, as Grodzinsky 2000), as others consider this brain area to support on-line processes of syntactic phrase-structure building (Friederici 1995).

There is, however, evidence that might challenge these views. It has been reported that patients with lesions outside Broca's area show syntactic comprehension deficits in off-line tasks (Caplan et al. 1985, 1996). Moreover, both Broca's and Wernicke's patients display a certain similarity in their perfomance in off-line comprehension tasks (Bates et al. 1987). It was demonstrated that patients speaking a language which marks syntax by inflectional morphology are generally more deficient, irrespective of their classification, than those speaking a language marking syntax by strict word order. However, the finding that in a highly inflected language Broca's patients are more affected than Wernicke's patients suggests a particular deficit of Broca's aphasics for syntactic morphology. Thus it appears that similarities between different patient groups can be observed when using off-line comprehension tasks. As these tasks only reflect the end point of the comprehension process they may imply certain on-line operations possibly supported by Broca's area or adjacent tissue, as well as other operations possibly supported by other areas (see also Zurif et al. 1993) (see Fig. 1).

2. *Functional Neuroimaging*

New imaging techniques during the 1990s have provided additional insight into the neural basis of syntactic parsing. Positron emission tomography (PET) studies (see Sect. 4.3.3) have investigated the reading of syntactically simple and complex sentences. It was found that the regional blood flow increases in the left Broca's area as a function of syntactic complexity (Stromswold et al. 1996). A recent study using functional magnetic resonance imaging (fMRI) to register the brain's activity during the reading of sentences of different complexity found a significant increase of the hemodynamic response in the left Broca's and Wernicke's area as a function of the sentence's complexity (Just et al. 1996).

Other studies investigated the brain–syntax relationship by comparing the brain activity during the processing of word lists with that during sentence processing. In an event-related fMRI study Friederici et al. (2000) presented four different auditory stimulus types: a list of real words (nouns); a list of pseudowords; normal sentences containing function words and real content words; and sentences containing function words and pseudowords instead of real content words. Both sentence types differed from the two word lists, showing an additional activation in the anterior portion of the left superior temporal gyrus. The sentences that contained only grammatical morphemes and pseudowords instead of content words displayed an additional activation in the left frontal operculum in the vicinity of the Broca's area (i.e., medial to BA 44).

The combined imaging data suggest that parts of the left superior temporal gyrus and the left frontal operculum are the brain systems that support syntactic processes. When speculating about the specific contribution of the two brain regions one may again consult the patient data. Given the observation that Broca's patients are able to perform grammaticality judgments (Linebarger et al. 1983), but are deficient in computing syntactic information on-line (Friederici 1995), and the finding that Wernicke's patients are deficient in performing judgments (Blumstein et al. 1982), but perform similar to normals in some on-line comprehension tasks (Zurif et al. 1993), the following account may be formulated. While the left temporal region is necessary for the representation of syntactic knowledge, the left inferior frontal region supports syntactic procedures in real time.

3. Neurophysiological Approaches

The temporal course of the neural activity underlying syntax processing has been evaluated by means of electroencephalography (EEG) (see Sects. 4.3.5 and 4.3.6) and more recently also by means of magnetoencephalography (MEG) (see Sect. 4.3.7). With their seminal paper Kutas and Hillyard (1983) introduced the event-related brain potential (ERP) method to investigate language processes. They observed a particular ERP component correlated with lexical-semantic processes: a negative potential peaking around 400 msec after the onset of the critical event, called the N400 component. The critical event in their study was a sentence final word which was highly unexpected and semantically incoherent (for example: 'He spread his bread with butter and socks.'). It was shown that the amplitude of the N400 varied as a function of a word's semantic expectancy and also as a function of a word's lexical frequency. Importantly, the N400 component has been reported in a number of different languages including English, German, Dutch, French, Hebrew, and even sign language in the visual as well as in the auditory domain (for a review, see Van Petten 1995). Note that unexpected events in a non-semantic domain elicit a positivity (P300) rather than an N400 (Donchin 1981).

3.1 Syntax-related Electrophysiological Correlates

The picture is somewhat more complex for the ERP pattern correlated with syntactic processing. There are two syntax-related components that have been identified: an early left anterior negativity, and a late centroparietal positivity. While the anterior negativity is only present when processing outright violations, the late positivity is observed when processing outright violation *and* when processing violations of structural preferences.

3.1.1 The syntax-related late positivity. The late positivity is a centroparietally distributed positivity usually at around 600 msec and beyond, called P600. P600 components are observed both in correlation with outright syntactic violation and with violations of syntactic preferences. Violations of the latter kind are present in so-called 'garden path' sentences, that is sentences for which it becomes clear at a particular word that the perceiver was misguided structurally by following a simple instead of a more complex syntactic structure (for example: 'The broker hoped to sell the stock was sent to jail' vs. 'The broker persuaded to sell the stock was sent to jail,' Osterhout and Holcomb 1992). At critical points in the sentence (at the word 'to' and the word 'was') the reader/listener has to revise and reanalyze the structure initially followed. The P600 is also present in sentences containing an outright syntactic violation, be it a phrase-structure violation realized by a word category error (Neville et al. 1991, Friederici et al. 1993) or be it a morphosyntactic violation realized by an incorrect inflection (Gunter et al. 1997, Münte et al. 1997). Functionally, the P600 is taken to reflect costs of reprocessing (Osterhout et al. 1994), syntactically guided processes of reanalysis and repair (Friederici 1995), or syntactic integration difficulties in general (Kaan et al. 2000).

3.1.2 The syntax-related anterior negativity. For outright syntactic violations an anterior negativity, that is, a frontally distributed negative waveform usually with a maximum over the left, is observed. It is called ELAN (early left anterior negativity) or LAN (left anterior negativity), depending on its latency of occurrence. Before we discuss the particular appearance of this syntax-related negativity in more detail we shall recapitulate briefly that syntactic information is multiply encoded in a given language.

3.2 Domains of Syntax and their Neurophysiological Correlates

There are at least four syntactic domains that have to be considered: (a) word category information necessary to build up local syntactic structures; (b) morphosyntactic information such as verb inflections necessary to establish agreement between different elements within and across phrases or case markings allowing the assignment of thematic roles (who is doing what to whom); (c) verb argument structure information necessary to establish subcategorization relations; and (d) long-distance dependencies necessary to establish the relation between a moved element (filler: 'who') and its original position in the sentence (gap: __) in a sentence such as 'Who did Peter see __?' Some, but not all, aspects of these different domains have been investigated using neurophysiological recordings. We shall take up the different domains in turn.

3.2.1 Word category information. The anterior negativity was found in correlation with phrase-structure violations realized as word category errors and with the processing of function words. It was observed between 125 and 200 msec after the critical word's onset (italicized) in a number of studies using auditory presentation (for example: 'Der Freund wurde im *besucht*' 'The friend was in the visited' vs. for 'Der Freund wurde *besucht*'/ 'The friend was visited,' Friederici et al. 1993, Hahne and Friederici 1999). (But see Neville et al. 1991, who found an early and late negativity in the visual domain.) The distribution of the word category error related negativity is often

frontal with a maximum to the left, but has also been found to be less lateralized (Friederici et al. 1999). So far it is not yet clear what determines the distributional differences. Current research is trying to solve this issue. With respect to the possible neural generators for the anterior negativity it is interesting that left anterior cortical lesions eliminate the anterior word category related negativity whereas left anterior lesions restricted to the basal ganglia do not (Friederici et al. 1999). Functionally, the word category related negativity was interpreted as reflecting initial structure building processes (Friederici 1995), which according to syntax first models (Frazier 1987), are based on word category information only.

A comparable left anterior negativity also related to word category information was observed for the processing of function words in syntactically correct sentences (Neville et al. 1992, see also Pulvermüller et al. 1995). It was present around 280 msec post onset when reading sentences. Similar to the above discussed negativity, the function word related negativity was eliminated with left anterior lesions (Ter Keurs et al. 1999). In line with the proposal by Friederici (1995), the authors interpret the left anterior negativity to reflect processes of word category identification.

3.2.2 Morphosyntactic information. Morphosyntactic violations (for example: 'The plane took *we* to paradise' vs. 'The plane took us to paradise') usually elicit a left anterior negativity between 300 and 500 msec after the critical word's onset (italicized). It was found in the visual domain in a number of different languages (Münte et al. 1993, Gunter et al. 1997, Coulson et al. 1998), as well as in the auditory domain (although with a somewhat larger distribution, Friederici et al. 1993). Note that the types of morphosyntactic violations investigated in these studies do not alter the critical word's syntactic category, but influence the relation between the verb and its arguments (e.g., subject–verb agreement). Under the view that initial phrase-structure building is concerned primarily with word category information, and that the lexical information encoded in an inflection only becomes relevant when thematic roles are to be assigned, it is conceivable that the violation of the former information type elicits an earlier brain response than the violation of the latter information type.

3.2.3 Verb argument structure information. The processing of verb argument structure information has been investigated in a number of different studies using different types of violation. In an early study, Rösler et al. (1993) registered ERP for sentences which contained subcategorization violations (for example: 'Der Lehrer wurde gehustet'/'The teacher was coughed'). A left anterior negativity between 300 and 500 msec was found in correlation with the violation. In this study, the sentence final verb (one-place verb) does not match the thematic grid opened up by the passive construction (two-place verb). When the mismatch involved subject vs. object arguments, again a left anterior negativity was reported (Coulson et al. 1998). Thus it appears that violations concerning the relation between external (subject) and internal arguments (objects) also evoke a left anterior negativity.

Interestingly, the argument structure related negativity is present between 300 and 500 msec, that is, the time domain in which the lexical-semantic ERP component is found. This is not surprising, given that the availability of the verb argument structure information requires full access to the verb's lexical entry.

3.2.4 Long-distance dependencies. Long-distance dependencies that establish the relation between a moved constituent (filler) and its original position (gap) have been investigated by Kluender and Kutas (1993) in a number of constructions. They found a LAN at the gap position at which the previously perceived filler has to be reactivated. As it is assumed that the filler has to be kept in memory until the gap is met, they claim that the LAN is a reflection of verbal memory. The two functional interpretations of the LAN as a syntactic violation related LAN and a filler-gap related LAN has recently been solved by showing that the former is present locally (at the violation) and that the latter spans from the filler to the gap position over the sentence (Kluender et al. 1998).

In conclusion, it appears that outright syntactic violations elicit a local anterior negativity with a maximum over the left hemisphere, in addition to a P600. Moreover, the combined data suggest that word category related anterior negativities are present earlier than anterior negativities found in correlation with lexically encoded morphosyntactic information and lexically encoded verb argument information.

4. *Modeling the Neurocognition of Syntactic Processes*

The data provided by recent ERP and functional imaging studies have advanced our knowledge on the neural basis of language, and of syntax in particular. The network identified to support syntactic processes clearly includes superior temporal and inferior frontal areas dominantly in the left hemisphere. Their specific function still has to be determined in detail. It appears, however, that the left temporal region is relevant for the representation of syntactic knowledge, whereas the

inferior frontal region supports syntactic procedures. The timing of the brain's activity correlated with syntactic processes can be described to consist of different stages: an early stage during which first-pass parsing processes (based on word category information) take place, and a next stage during which lexically encoded syntactic information is processed. Both are reflected in a left anterior negativity in the ERP. A late stage of syntactic processes is reflected by a late positivity in the ERP. This component is observed whenever syntactic integration is difficult. Future research will have to determine whether different aspects of syntactic integration difficulties are correlated with distinct topographies of the late positivity which would indicate the involvement of different neural structures.

See also: Connectionist Models of Language Processing; Semantic Knowledge: Neural Basis of; Sign Language: Psychological and Neural Aspects; Syntax

Bibliography

Bates E, Friederici A D, Wulfeck B 1987 Comprehension in aphasia: A cross-linguistic study. *Brain and Language* **32**: 19–67

Berndt R S, Caramazza A 1980 A redefinition of the syndrome of Broca's aphasia: Implications for a neuropsychological model of language. *Applied Psycholinguistics* **1**: 225–78

Blumstein S E, Milberg W, Schrier R 1982 Semantic processing in aphasia: Evidence from an auditory lexical decision task. *Brain and Language* **17**: 301–15

Bradley D C, Garrett M F, Zurif E B 1980 Syntactic deficits in Broca's aphasia. In: Caplan D (ed.) *Biological Studies on Mental Processes*. MIT Press, Cambridge, MA, pp. 269–86

Broca P 1861 Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletins de la Société Anatomique de Paris* **6**: 330–57

Caplan D, Baker C, Dehaut F 1985 Syntactic determinants of sentence comprehension in aphasia. *Cognition* **21**: 117–75

Caplan D, Hildebrandt N, Makris N 1996 Location of lesions in stroke patients with deficits in syntactic processing in sentence comprehension. *Brain* **119**: 933–49

Caramazza A, Zurif E B 1976 Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language* **3**: 572–82

Coulson S, King J, Kutas M 1998 Expect the unexpected: Event-related brain responses of morphosyntactic violations. *Language and Cognitive Processes* **13**: 21–58

Donchin E 1981 Surprise!…Surprise? *Psychophysiology* **18**: 493–513

Frazier L 1987 Theories of sentence processing. In: Garfield J (ed.) *Modularity in Knowledge Representation and Natural-language Processing*. MIT Press, Cambridge, MA

Friederici A D 1995 The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and Language* **50**: 259–81

Friederici A D, Meyer M, Von Cramon D Y 2000 Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. *Brain and Language* **74**: 289–300

Friederici A D, Pfeifer E, Hahne A 1993 Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research* **1**: 183–92

Friederici A D, Von Cramon D Y, Kotz S A 1999 Language related brain potentials in patients with cortical and subcortical left hemisphere lesions. *Brain* **122**: 1033–47

Grodzinsky Y 2000 The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences* **23**: 1–71

Gunter T C, Stowe L A, Mulder G 1997 When syntax meets semantics. *Psychophysiology* **34**: 660–76

Haarmann H J, Kolk H H J 1991 Syntactic priming in Broca's aphasics: Evidence for slow activation. *Aphasiology* **5**: 247–63

Hahne A, Friederici A D 1999 Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience* **11**: 193–204

Just M A, Carpenter P A, Keller T A, Eddy W F, Thulborn K R 1996 Brain activation modulated by sentence comprehension. *Science* **274**: 114–16

Kaan E, Harris A, Gibson G, Holcomb P J 2000 The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes* **15**: 159–201

Kluender R, Kutas M 1993 Subjacency as a processing phenomenon. *Language and Cognitive Processes* **8**: 573–633

Kluender R, Münte T, Wind-Cowles H, Szentkuti A, Walenski M, Wieringa B 1998 Brain potentials to English and German questions. *Journal of Cognitive Neuroscience Supplement* **24**

Kolk H H J, Van Grunsven M J F 1985 Agrammatism as a variable phenomen. *Cognitive Neuropsychology* **2**: 347–84

Kutas M, Hillyard S A 1983 Event-related potentials to grammatical errors and semantic anomalies. *Memory & Cognition* **11**: 539–50

Linebarger M C, Schwartz M F, Saffran E M 1983 Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition* **13**: 361–92

Münte T F, Heinze H J, Mangun G R 1993 Dissociation of brain activity related to syntactic and semantic aspects of language. *Journal of Cognitive Neuroscience* **5**: 335–44

Münte T F, Matzke M, Johannes S 1997 Brain activity associated with syntactic incongruencies in words and pseudowords. *Journal of Cognitive Neuroscience* **9**: 318–29

Neville H J, Mills D L, Lawson D S 1992 Fractionating language: Different neural subsystems with different sensitive periods. *Cerebral Cortex* **2**: 244–58

Neville H J, Nicol J L, Barss A, Forster K I, Garrett M F 1991 Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience* **3**: 151–65

Osterhout L, Holcomb P J 1992 Event-related potentials and syntactic anomaly. *Journal of Memory and Language* **31**: 785–804

Osterhout L, Holcomb P J, Swinney D 1994 Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology-learning Memory and Cognition* **20**: 786–803

Pulvermüller F, Lutzenberger W, Birbaumer N 1995 Electrocortical distinction of vocabulary types. *Electroencephalography and Clinical Neurophysiology* **94**: 357–70

Rösler F, Friederici A D, Pütz P, Hahne A 1993 Event-related brain potentials while encountering semantic and syntactic constraint violations. *Journal of Cognitive Neuroscience* **5**: 345–62
Stromswold K, Caplan D, Alpert N, Rauch S 1996 Localization of syntactic comprehension by positron emission tomography. *Brain and Language* **52**: 452–73
Ter Keurs M, Brown C M, Hagoort P, Stegemann D F 1999 Electrophysiological manifestations of open- and closed-class words in patients with Broca's aphasia with agrammatic comprehension: An event-related brain potential study. *Brain* **122**: 839–854
Van Petten C 1995 Words and sentences: Event-related brain potential measures. *Psychophysiology* **32**: 511–25
Wernicke C 1874 *Der aphasische Symptomenkomplex*. Cohn and Weigert, Breslau, Germany
Zurif E B, Swinney D, Prather P, Solomon J, Bushell C 1993 An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. *Brain and Language* **45**: 448–64

A. D. Friederici

Syntactic Control

The term 'control' is usually attributed to the generative linguist Paul Postal (1970) (see *Generative Syntax*). The term expresses metaphorically the observation that the reference of an understood grammatical subject in a subordinate clause, the controllee, may depend on the reference of another constituent, the controller, in the main clause. An example illustrating the control relation is given in (1), with its relevant syntactic structure represented in (1a). In (1a) constituents with the same reference are marked by identical subscripts; the bracketed part of the sentence contains a nonfinite subordinate clause with an implicit subject symbolized as Ø (for 'empty subject'). These conventions will be followed throughout the article.

Mary convinced Billy to do her homework. (1)

Mary_i convinced Billy_j [$Ø_j$ to do her_i homework]. (1a)

In (1) the noun phrase *Billy* controls the reference of the empty subject of the infinitival complement clause. Every native speaker of English will understand (1) as meaning that Billy—and not Mary herself—will do her homework. The challenging linguistic problem is to find a principled account for why in (1) the main clause object is the controller, why the main clause subject cannot have this function, and why the empty subject cannot just refer to a third person not mentioned in the sentence at all.

Sentences such as (1) exhibit the notion of obligatory control, i.e., a case in which the empty subject must be coreferential with a main clause constituent. Obligatory control contrasts with free (also called arbitrary or nonobligatory) control as in:

[Ø to wait in the crowded airport] would be a mistake. (2)

In the subject clause in (2), the reference of the understood subject is not determined by any other noun phrase in the sentence but depends on the discourse context.

1. Some Control Phenomena

In this article, the term 'control' will be understood as meaning 'obligatory control.' The following nonexhaustive list gives some sentences that exemplify obligatory control.

(a) Infinitival Purpose Clauses

John_i bought a book_j [$Ø_i$ to amuse himself with]. (adapted from Nishigauchi 1984, p. 229) (3)

Mary_i bought the little boy_j a toy [$Ø_j$ to amuse himself with]. (adapted from Nishigauchi 1984, p. 229) (4)

(b) Infinitival Complement Clauses

John_i begged Mary_j [$Ø_j$ to leave]. (5)

Mary_i begged [$Ø_i$ to leave]. (6)

(c) Gerundial Complement Clauses

Paul_i thanked Mary_j for [$Ø_j$ writing the report]. (7)

Paul_i thanked Mary_j for [$Ø_i$ having been admitted to the tennis club]. (8)

Most work on control has been done on nonfinite complement clauses, i.e., on clauses exemplified by sentences (5) through (8). The remainder of the article will therefore be concerned with control in nonfinite complement clauses.

2. A Short History of Control

In the early days of generative grammar (see *Linguistics: Incorporation*) in the 1960s and 1970s, it was generally believed that control is a syntactic phenomenon, i.e., that the controller of the empty subject in the subordinate clause can be determined on a purely structural basis. More recently, it has become increasingly evident that semantic and pragmatic factors play an equally, if not more, important role in control.

2.1 Syntactic Approaches

The syntactic approach to control assumes that control relations can be determined on a purely structural basis. The most influential proposal along these lines was made by Peter Rosenbaum (1967, 1970). He postulated a Minimal Distance Principle (MDP), which predicts that the reference of the (empty) subject of a nonfinite complement clause depends on the closest available controller in the main clause. In sentence (1) above, the main clause noun phrase closest to the empty subject Ø is *Billy*.

Indeed, this noun phrase controls the reference of the empty subject, and the MDP thus makes the right prediction.

There are, however, many sentences for which the MDP is falsified. Difficulties emerge with constructions that display more than one potential controller in the main clause, as in (9) and (10):

Mary$_i$ persuaded Bill$_j$ [Ø$_j$ to clear her$_i$ desk]. (9)
Mary$_i$ promised Bill$_j$ [Ø$_i$ to clear her$_i$ desk]. (10)

For sentence (9) the MDP correctly selects the closest noun phrase in the main clause, *Bill*, as controller. *Persuade* belongs to the class of control verbs that normally select the main clause object as the controller. However, for sentence (10) the MDP wrongly selects *Bill* as controller; in fact, the main clause subject *Mary* is the controller. *Promise* belongs to the class of control verbs that usually select their subject as controller. Subject control verbs of the *promise* class constitute a problem for the MDP and have to be marked as exceptional in the lexicon.

Another problem that the MDP faces derives from the fact that one and the same verb may have either subject or object control:

Bill$_i$ thanked the coach$_j$ for [Ø$_j$ selecting him$_i$]. (11)
Bill$_i$ thanked the coach$_j$ for [Ø$_i$ having been admitted to the tennis club]. (12)

In (11), which can be regarded as the default case, the controller is *the coach* (object control), whereas in (12) a control switch or shift takes place: the most likely referent of the implicit subject Ø is *Bill* (subject control). The MDP wrongly predicts that in both sentences *the coach* should be the controller.

Other syntactic control theories have been proposed by Joan Bresnan (1982) (see *Lexical Functional Grammar*) and Emmon Bach (1979) in a logical framework (see *Montague Grammar*). What these theories have in common is that they regard syntactic structure as the decisive factor for determining the control relationship. However, there is a substantial body of evidence that suggests strongly the importance of semantic and pragmatic factors in the determination of control relations.

2.2 Semantic Approaches

A comparison of sentences (9) and (10), which look alike structurally, invites the conclusion that their differing control properties are caused by the meanings of the verbs *persuade* and *promise*, respectively. To persuade somebody to do something means that the addressee of an act of persuasion is led to believe that he or she should do something. This prospective action is expressed in the infinitival complement clause. Since in an active sentence such as (9) the addressee is the grammatical object of the main clause, it is not surprising that this constituent functions as controller of the empty subject of the subordinate clause. In contrast, a promise involves a commitment of the speaker to a future course of action. In an active sentence such as (10), the 'promiser' is the grammatical subject and one would therefore expect the subject to be controller of the understood subject of the complement clause.

The most complete account that acknowledges the importance of the meaning of control verbs is probably Sag and Pollard's (1991) control theory. These authors use semantic criteria to distinguish three classes of control verbs: (a) the *order/permit* class, which typically induces object control; (b) the *promise* class, which typically induces subject control; and (c) the *want/expect* class, which again induces subject control. The *order/permit* and the *promise* classes conceptually imply actions, whereas verbs of the *want/expect* class express what Sag and Pollard call an 'orientation' in their complement clause.

2.3 Pragmatic Approaches

Since the mid-1980s, the insight has grown that, in addition to the meaning of control verbs, the context in which a sentence is uttered and world knowledge about the participants' status play a significant role in assigning the coreferential links between controller and controllee (e.g., Comrie 1984, 1985, Růžička 1999; see *Pragmatics: Linguistic*).

These additional pragmatic factors are illustrated by the following German sentences (adapted from Wegener 1989):

Der Anwalt$_i$ versprach dem Häftling$_j$ [Ø$_j$ bald aus dem Gefängnis entlassen zu werden]. (13)
'The lawyer promised the prisoner that he (= the prisoner) would be released from jail soon.'
Der Häftling$_i$ versprach seiner Frau$_j$ [Ø$_i$ bald aus dem Gefängnis entlassen zu werden]. (14)
'The prisoner promised his wife that he (= the prisoner) would be released from jail soon.'

The most plausible interpretation for both (13) and (14) is that the prisoner will be released from jail. These interpretations follow from the speech act verb *versprechen*, 'promise,' the status of the referents of the noun phrases *der Anwalt, dem Häftling*, and *seiner Frau*, and the meaning of the passive complement clause. A promise implies a promiser who is committed to a future action, and a promisee who benefits from the action of the promiser. This semantic–pragmatic scenario can be applied to the interpretation of sentences (13) and (14) and the most plausible controller can be inferred (cf. Panther and Köpcke 1993, Panther 1994).

In (13) the lawyer is the promiser; hence the lawyer is committed to some future action. However, the complement clause does not express an action explicitly. A step toward a reasonable interpretation is to assume that the complement clause expresses the result of an action performed by the promiser. The action itself has to be inferred. Furthermore, the result of the

action is beneficial to the prisoner. Therefore, the most plausible controller is the noun phrase *dem Häftling*. On this account, the assignment of the controller is not a matter of finding the noun phrase that is syntactically closest to Ø in the matrix clause, but a matter of common-sense knowledge about the roles of lawyers and prisoners.

For (14), a coherent interpretation results if it is assumed that the prisoner can behave in such a way that he causes his release from jail. Again, the complement clause is interpreted as the result of an action—an action carried out by the prisoner to the effect that he will be released from prison. His wife can be regarded as the beneficiary of this action.

The interpretations sketched above are not the only ones possible. On the assumption that the lawyer is in jail, (13) could be interpreted as conveying that the lawyer will be released from jail; and (14) could have the interpretation, 'The prisoner promised his wife that she would be released from jail soon' given the background knowledge that the prisoner's wife is in prison. It is in the nature of the pragmatic approach to control that different background assumptions have an impact on what is the most plausible controller of Ø.

Similar observations have been made for Chinese by Huang (1994, p. 153), who contrasts the following sentences (referential subscripts and bracketing have been added):

Xuesheng$_i$ daying laoshi$_j$ [Ø$_i$ mingtian jiao zuowen]. (15)
'The pupil promises the teacher that he (= the pupil) will hand in the essay tomorrow'

Laoshi$_i$ daying xuesheng$_j$ [Ø$_j$ mingtian fang yi tian jia]. (16)
The teacher promises the pupil that he (= the pupil) will have a day off tomorrow'

The control verb *daying*, 'promise,' can be used either in the 'unmarked,' i.e., normal, sense 'commitment-to-action, as in (15), or in the 'marked' sense 'commitment-to-permission,' as in (16). These senses are derived pragmatically; they constitute 'preferred' interpretations that can be inferred on the basis of the meaning of the speech act verb *daying* and world knowledge about teacher–pupil interactions in the institutional framework of a school.

To summarize, the interpretation of control sentences relies not only on syntactic and semantic information but also, to a substantial degree, on world knowledge and contextual clues about the utterance situation.

3. *Language-specific Coding Differences of Control Relations*

One might surmise that the mechanisms of control are very similar across languages, but on closer inspection it turns out that there are language-specific differences in the way that control structures are coded. A selective list of these coding differences is given below (for a detailed account, see Růžička 1999).

3.1 *Modal vs. Non-modal Coding*

For many speakers of English the following sentence is ambiguous:

Johnny$_i$ asked his mother$_j$ [Ø$_{i/j}$ to go to the movies]. (17)

Sentence (17) can mean either that Johnny wants his mother to go to the movies or that Johnny asks for permission to go to the movies himself. This latter interpretation seems to be the most likely one, assuming that Johnny's mother has authority over Johnny (see Sect. 2.3).

As Růžička (1999, p. 45) points out, the 'ask-for-permission' (subject control) reading is also available alongside the 'ask-for-action' interpretation (object control) in Spanish, although the concept of permission is not coded explicitly.

María$_i$ le$_j$ pidió a Juan$_j$ [Ø$_i$ hablar con los muchachos]. (18)
'María asked Juan that she be allowed to talk to the boys'

The 'ask-for-permission' interpretation is not possible, however, if the same sentence is translated literally into German:

Hänschen$_i$ bat seine Mutter$_j$ [Ø$_j$ ins Kino zu gehen]. (19)
'Johnny asked his mother to go to the movies'

Sentence (19) can only mean that Johnny's mother is supposed to go to the movies. In order to express the 'ask-for-permission' sense, with Johnny as controller, German speakers have to resort to an utterance that codes permission by means of the modal auxiliary dürfen:

Hänschen$_i$ bat seine Mutter$_j$ [Ø$_i$ ins Kino gehen zu dürfen]. (20)
'Johnny asked his mother to be allowed to go to the movies'

3.2 *Causative vs. Non-causative Coding*

Another important coding difference between English and languages like German, Russian, and Czech is exemplified by

Peter$_i$ persuaded his father$_j$ [Ø$_j$ to be examined by a doctor]. (21)

Sentence (21) means that Peter's father should undergo a medical examination. This sense is, however, not explicitly coded in sentence (21), but is nevertheless conventionally conveyed by it. What is coded in the subordinate clause is the result of the action, not the action itself. If sentence (21) is translated literally into German, for many speakers ungrammaticality results and, for those who accept

the sentence, an interpretation with Peter as controller is more plausible:
?Peter$_i$ überredete seinen Vater$_j$ [Ø$_i$ von einem Arzt untersucht zu werden]. (22)
'Peter persuaded his father that he (= Peter) should be examined by a doctor'

The only way of rendering the intended message of (21) in German or Russian is to code *his father* as an active participant who engages in an action which results in a medical examination by a doctor:
Peter$_i$ überredete seinen Vater$_j$ [Ø$_j$ sich von einem Arzt untersuchen zu lassen]. (23)
'Peter persuaded his father to have himself (= his father) examined by a doctor'
Petr$_i$ ugovoril otca$_j$ [Ø$_j$ podvergnut' sebja osmotru vrač]. (Růžička 1999, p. 58) (24)
'Peter persuaded father that he (= father) should undergo a medical examination'

Both German and Russian have to code the agent role of father explicitly. This is done by the causative verb *lassen* in German and the action verb *podvergnut' sebja* 'submit oneself' in Russian.

3.3 Implicit Control vs. Explicit Control

Another important difference between English and languages such as German and Russian concerns the ability for a conceptually present controller to remain unexpressed. In German it is possible to say:
Der Professor$_i$ bat Ø$_j$ [Ø$_j$ den Zeitschriftenartikel zu kopieren]. (25)
'The professor requested (of some person) that this person photocopy the journal article'

In (25) the controller of the understood subject remains unexpressed. The person or group of persons supposed to comply with the request is indeterminate. This phenomenon may be called implicit control (cf. Panther 1997).

However, a word-for-word translation of (25) would not have the same effect in English:
The professor$_i$ asked Ø$_j$ [Ø$_i$ to photocopy the journal article] (26)

As the referential subscripts indicate, the only possible interpretation of (26) is that the professor himself wants to photocopy the journal article and asks for permission to do so. In German, however, the translation of (26) requires modal coding with *dürfen* 'be allowed to' (see Sect. 3.1).

Another example of implicit control is (27):
Der Arzt$_i$ riet Ø$_j$ [Ø$_j$ den Fettanteil in meinem Essen zu verringern]. (27)
'The doctor advised (someone) that this person should reduce the amount of fat in my diet'

The literal translation of (27) in English leads to ungrammaticality (marked by an asterisk):
*The doctor$_i$ advised Ø$_j$ [Ø$_j$ to reduce the amount of fat in my diet]. (28)

These examples illustrate a general tendency in English: in obligatory infinitival control constructions the controller has to be made explicit, whereas in languages such as German and Russian it can remain unexpressed. However, sometimes there is a way to circumvent the principle in English by resorting to a gerund construction. A possible translation of (27) is:
The doctor$_i$ advised Ø$_j$ [Ø$_j$ reducing the amount of fat in my diet]. (29)

4. Conclusion

Human languages employ various means to keep track of referents on the sentence and the discourse level. Control relations can be seen in this larger framework of coreference phenomena in general (see *Anaphora*). As with other coreference relations, it turns out that the interpretation of control structures involves a complex interplay of syntactic, semantic, and pragmatic parameters.

See also: Linguistics: Incorporation; Syntax; Syntax–Semantics Interface; Valency and Argument Structure in Syntax

Bibliography

Bach E 1979 Control in Montague Grammar. *Linguistic Inquiry* **10**: 515–31

Bresnan J 1982 Control and complementation. *Linguistic Inquiry* **13**: 343–434

Comrie B 1984 Subject and object control: Syntax, semantics, pragmatics. *Berkeley Linguistics Society: Proceedings of the Tenth Annual Meeting*. Berkeley Linguistics Society, Berkeley, CA

Comrie B 1985 Reflections on subject and object control. *Journal of Semantics* **4**: 47–65

Huang Y 1994 *The Syntax and Pragmatics of Anaphora: A Study with Special Reference to Chinese*. Cambridge University Press, Cambridge, UK

Nishigauchi T 1984 Control and the thematic domain. *Language* **60**: 215–50

Panther K-U 1994 *Kontrollphänomene im Englischen und Deutschen aus semantisch-pragmatischer Perspektive*. Narr, Tübingen

Panther K-U 1997 An account of implicit complement control in English and German. In: Verspoor M, Lee K D, Sweetser E (eds.) *Lexical and Syntactical Constructions and the Construction of Meaning*. Benjamins, Amsterdam

Panther K-U, Köpcke K-M 1993 A cognitive approach to obligatory control phenomena in English and German. *Folia Linguistica* **27**: 57–105

Postal P M 1970 On coreferential complement subject deletion. *Linguistic Inquiry* **1**: 439–500

Rosenbaum P S 1967 *The Grammar of English Predicate Complement Constructions*. MIT Press, Cambridge, MA

Rosenbaum P S 1970 A principle governing deletion in English sentential complementation. In: Jacobs R, Rosenbaum P S (eds.) *Readings in English Transformational Grammar*. Ginn, Waltham, MA
Růžička R 1999 *Control in Grammar and Pragmatics*. Benjamins, Amsterdam
Sag I A, Pollard C 1991 An integrated theory of complement control. *Language* **67**: 63–113
Wegener H 1989 'Kontrolle'—semantisch gesehen: Zur Interpretation von Infinitivkomplementen im Deutschen. *Deutsche Sprache* **17**: 206–28

K.-U. Panther

Syntax

The linear arrangement of the words in a sentence is rarely arbitrary, and the morphological shape which words take often depends on other words they co-occur with. The part of grammar that deals with the pertinent regularities is called syntax.

1. *Functions of Syntax*

Loosely speaking, the variations in form and position that constitute syntax fulfill a number of functions: The words and phrases of a sentence stand in various meaning relations to each other, and these relations are encoded by syntactic differences. A basic function is the expression of so-called 'thematic relations' or 'argument linking': a verb such as *kill* expresses a two-place predicate KILL (x, y), a relation between an *agent* x and a *patient* y in an event of killing. Such argument roles of a predicate must be linked to the argument expressions in a sentence, and at least in the ideal case, this linking is carried out systematically in natural language, for example by word order, as in English (1), or by case, as in German (2):

(1) the tiger killed the buffalo.
agent predicate patient

(2) dass der Tiger den Büffel tötete
that the.nom tiger the.acc buffalo killed.

Argument linking is not the only semantic-pragmatic function figuring in syntax. Argument places of verbs can, for example, be questioned (*who did he see?*), and languages may require that question phrases appear in specific positions (e.g., clause initially, as in English). Information structure ('new vs. old information,' 'focus vs. topic') is a relevant dimension for sentence construction too, and may influence word order, for example. Unlike (2), German (3) seems to be fully felicitous only in a situation in which the crucial new information is the identification of the tiger as the agent of killing.

(3) dass den Büffel der Tiger tötete.

A characterization of the functions that syntax encodes must be complemented by an identification of the means of encoding. Word order and shape variation are two options. The latter comes in two varieties. The fact that a noun is the agent of a verb may be encoded by shape variation of the noun (= case), or by shape variation of the verb (agreement markers)—or a combination of both. 'Word order,' on the other hand, seems to be the result of mapping the two-dimensional hierarchical structure of a sentence onto the unidimensional sound (or letter) stream (see Sect. 4 below). Intonational differences often serve to express pragmatic and semantic functions, but they seem to be employed only rarely in argument linking.

2. *Argument Roles*

In a sentence such as (4), we can distinguish words and phrases that are linked to argument places (in italics) of the predicate (boldface), and words and phrases that are not; which have an adjunct function (underlined).

(4) *Jack* **saw** *the girl* <u>in</u> <u>the</u> <u>park</u> <u>yesterday.</u>

Let us confine our attention to argument roles. They are linked to (the meaning of) the predicate *hierarchically*. For a verb such as *kill*, the agent is higher in this hierarchy than the patient. Grammatical differences reflect these hierarchies systematically. When two arguments of a predicate refer to the same entity, the higher argument will, for example, be realized in its usual shape, whereas the lower argument takes a special form (it is, e.g., realized as a reflexive pronoun) but not vice versa.

(5) (I expect) Bill to wash himself.
(I expect) *himself to wash Bill.
[Ungrammatical strings are preceded by a*]

For prototypical instances of 'transitive' relations, that is, for verbs expressing instantaneous causation of an event affecting a patient by an agent (*kill*, *beat*), the hierarchy seems to be the same in all languages, with agents always being more prominent than patients. Potential exceptions to this (Dyirbal, Arctic Eskimo; see e.g., Marantz 1984) can presumably be explained away. For less prototypical relations—e.g., for psychological predicates—grammatical hierarchies are linked to semantics less systematically, as the difference between *like* and *please* suggests: the 'experiencer' of the psychological state is encoded as the subject (as the higher role) with *like*, but as an object (the lower role) with *please*. Languages may differ in this respect—in some, nearly all psychological predicates follow the model of *please* (Basque), in others, all relevant verbs pattern with *like* (Lezghian), and in still others (like English), both construction types occur (see Bossong 1998).

In addition to identifying their place in the hierarchy, argument roles have often been classified in terms of notions that we have already used informally, namely/thematic ('theta') roles (Gruber 1965, Dowty

1991, Jackendoff 1990; but also Hjelmslev 1972, 1935–7, Jakobson 1971, 1936) such as 'agent,' 'instrument,' 'goal,' 'experience,' 'source,' 'patient,' etc. One needs to refer to these notions in the context of selecting the proper preposition or semantic case, but for capturing core syntactic laws, the necessity of integrating theta roles (in addition to hierarchy statements) into grammatical descriptions has yet to be established (apart, perhaps, from the special role played by the agent function). At least, most if not all current grammatical models can do without them easily (Fillmore 1968 and approaches following him are notable exceptions).

3. *Grammatical Functions*

One issue of syntactic theory is whether argument roles can be linked directly to the formal means of encoding, or whether the rules of encoding crucially employ ancillary notions, namely, the grammatical functions 'subject,' 'direct object,' 'indirect object,' etc. Some grammatical models (Relational Grammar, Perlmutter (1983); Lexical Functional Grammar, (Bresnan 1982)) assume that such grammatical relations are indispensable primitive notions of syntax. Other approaches, such as the Standard Theory (Chomsky 1965), or the Government and Binding (GB) framework (Chomsky 1981) take grammatical functions (at best) to be mere classificatory concepts of little explanatory value, which can be defined in terms of structure and hierarchy: a 'direct object' is a noun phrase 'governed' by the verb, or it is a noun phrase which is a structural sister of the verb.

Thus, Lexical Functional Grammar assumes that sentences are linked to a *functional structure*, in which, e.g., the verb *kill* is linked to two abstract grammatical functions, subject and object. How these grammatical relations are spelled out is a function of language-particular rules. In English, grammatical functions are encoded by word order (more precisely, hierarchically)—the direct object has to follow the verb immediately (1). In German (2, 3), the grammatical functions seem encoded by case marking (direct objects bear accusative case). The mapping from argument structure to formal means of expression would thus proceed as in (6):

(6) kill (role 1, role 2)
kill (subject, object)
kill (nominative, accusative)

Lexical Decomposition Grammar (Wunderlich 1997) translates argument hierarchies directly into morphological markings or order relations; that is, the second step in (6) is skipped. The Standard Theory and the GB framework assume that sentences are linked to an abstract 'deep structure,' in which the words and phrases combine hierarchically, with argument role hierarchies being reflected by argument expression hierarchies in deep structure, as in (7). Such deep structures are mapped on to audible representations in various ways, e.g., by case marking, reordering, etc.

(7) $[_{\text{Infl-Phrase}}$ Infl $[_{\text{VP}}$ [the tiger] [killed [the buffalo]]]]

In other words, there is no consensus concerning the status of grammatical functions. It seems fair to say however, that no universally applicable definitional criteria for, e.g., subjecthood, have been found so far. This suggests that grammatical functions must either be considered fuzzy concepts, or that they reflect a grammatical perspective that may be adequate in some languages only.

The nature of grammatical functions has often been discussed in the context of the claim that one type of linking argument expressions to argument places, namely the hierarchical-positional ('configurational') one that we find in English, is more fundamental than the others (such as case marking) in the sense that it underlies the grammatical system of *all* languages: at deep structure level, all languages follow roughly the constructional logic of English. This is the position held by the mainstream version of the GB framework, and has been taken over in the Minimalist program (Chomsky 1995). The contrary view that there are no privileged systems has been a minority position in GB (e.g., Hale 1983), and is the position held by Lexical Functional Grammar and Relational Grammar. Debates in the 1980s primarily contrasted languages with rigid word order and little case marking (English, Chinese) with languages with a relatively elaborate case system and free constituent order (German, Japanese). One outcome is the insight that a strong hierarchy-based configurational component can be identified in the grammar of languages with a fairly rich case marking too (e.g., Haider 1993).

On the other hand, it may also be uncontroversial that polysynthetic languages (such as Mohawk) with a very rich system of verbal agreement have a grammatical system that differs radically from, say, English and Japanese. It appears as if nominal expressions in the polysynthetic languages do not show a grammatical behavior that is even remotely reminiscent of subjects and objects—rather, their grammar is comparable to what holds for adjuncts (like adverbs) in English. Argument linking and other important grammatical factors seem to be dealt with in the morphology only. Why this is so, and whether this means that there are languages without a syntax proper (a position strongly argued against by Baker 1996) is an open issue.

4. *Movement*

If more than one semantic–pragmatic relation has to be expressed, and if the same encoding mechanism is used for the different functions, the encoding rules

may impose incompatible requirements on a phrase in a sentence. In English, the positional system of identifying arguments requires that objects *follow* the verb, whereas question phrases have to appear at the *beginning* of a clause. These requirements cannot be met simultaneously:

(8) (a) *(I do not care) you like who
(b) (I do not care) who you like

One way of dealing with this problem has been the postulation of various *levels of representation*, such that different requirements have to be met at different levels. In the Chomskyian tradition (see e.g., Chomsky 1957, 1965, 1981) a level called deep structure/d-structure, representing the thematic relations among arguments and predicates has played a prominent role until quite recently. This d-structure (corresponding to (9a)) would then be *mapped* on to representations coming close to the phonetic 'surface' by a set of structure-changing operations ('transformations'), of which the movement (more precisely: fronting) of phrases has been the most important one (relative to (9a), *who* has moved to clause initial position in (9b)). The law that requires that complement questions begin with a question word applies at 'surface structure' (s-structure) only, not at deep structure.

(9) (a) you like who?
(b) who do you like?

Details changed in the 1990s, but the distinction between structure building operations encoding argument linking and structure-changing operations (movement) is still a key property of all grammatical models inspired by Chomsky. Most alternative approaches (e.g., Generalized Phrase Structure Grammar; Gazdar et al. 1985) differ at a purely technical level only. Optimality Theory borrowed from phonology (Prince and Smolensky 1993; see also Grimshaw 1997, Kager 1999, Müller 2000) allows the handling of conflicts between incompatible grammatical requirements in a quite different way: the grammatical principles are ordered, and lower principles (such as: the object appears to the right of the verb) may be violated in the interest of higher ones (such as: question phrases appear clause-initially). Whether such options will eventually reduce the importance of movement metaphors in grammatical descriptions is an open question at the time of writing.

Beginning with the seminal work of Ross (1967), investigations on syntactic movement have brought to light a new type of syntactic law. There is a set of 'locality conditions' on movement, in the sense that words and phrases cannot be moved out of certain kinds of construction. Thus, in English, simple complement clauses are transparent for movement, while complement questions or relative clauses are not. (The place where a phrase has moved from is indicated by.)

(10) (a) you think that she likes someone
who do you think that she likes _
(b) you wonder when she kissed someone
***who** do you wonder *when* she kissed _
(c) you know a man who kissed someone
***who** do you know a man *who* kissed _

In (10b) and (10c), we observe an 'intervention' effect. The phrase *who* should be placed in the position indicated by _ at d-structure (because of the laws of argument linking), and in sentence initial position at s-structure because of the laws of question formation. The 'preposing' of *who* yields ungrammaticality in (10b) and (10c) because phrases cannot cross similar phrases (like *when* in (10b)) when they move to their surface position.

A similar intervention effect can be observed in (11) when a choice must be made between different phrases that could be placed into sentence-initial position: again, a phrase of type X must not cross a phrase of the same type in this context. This is the so-called superiority effect (see Chomsky 1973, also Chomsky 1995, Rizzi 1990).

(11) (a) **who** do you expect _ to buy *what*
(b) ***what** do you expect *who* to buy _

Furthermore, only complement clauses with an object function are transparent for movement, but not subject clauses (12a) and adjunct clauses (12b):

(12) (a) *who is to kiss _ fun?
(b) *who do you weep because you met _ ?

The Chomskyan approach (in the narrower sense) tries to identify a set of purely formal and universally valid syntactic laws that characterize such 'island' effects (see Chomsky 1981, 1986). The empirical challenge is that the constraints proposed so far are not being respected in all languages in the same way.

5. *Logical Forms*

Not all languages require that constituent questions be formed by the fronting of a question word or phrase (wh-phrase). In Chinese, wh-phrases are realized in the position where they should appear according to the laws of argument linking.

(13) ni xihuan shei
you like who
'who do you like?'

Interestingly, the scope options of Chinese wh-phrases are constrained, and in a way reminiscent of what we saw in the preceding section for English. Thus, a wh-phrase which takes scope over the whole matrix clause may appear as part of an object clause (14), but cannot show up in a subject clause (15). This is identical to what we have just observed for English (10a), (12).

(14) Zhangsan xiangxin shei mai-le shu
Zhangsan believe who bought books
'who does Zhangsan believe bought books?'

(15) *Zhangsan tao-le shei zhen kexi
Zhangsan marry who real pity
'for which *x*: that Zhangsan married *x* is a real pity'

Such observations gave rise to the idea (Huang 1982), quite popular in the GB framework, that *shei* undergoes movement to the clause initial position corresponding to its semantic scope as well (so that it is clear why its scope behavior is constrained by restrictions on movement), but 'invisibly' so. There are two ways in which the relevant observations may be interpreted, one that gives priority to syntax, and one that does not.

The GB framework postulated a grammatical architecture as in (16): by putting together words from the lexicon, a deep structure arises. This d-structure undergoes a number of transformations, mainly movement, in order to reach the interface levels Phonetic Form (PF) (interface with the articulatory–perceptory system) and Logical Form (LF) (interface with the cognitive system). The first steps, so-called 'overt' syntax, are common to these two derivations targeting PF and LF; they yield surface structure (s-structure). 'After' s-structure, the two paths split. The syntactic derivation of LF proceeds in the standard way, but no longer affecting phonetic material (covert syntax).

(16)

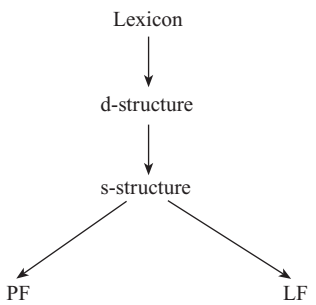

In this model, the Logical Form of (14) is (17), i.e., question words are preposed in Chinese too—the only difference from English being that the movement of *shei* has applied 'after' s-structure, so that it lacks a phonetic effect. All languages, it is claimed, have the same Logical Form, but they may make (among other things) different decisions as to whether a certain movement operation (like the fronting of wh-phrases) applies in overt or covert syntax.

(17) shei Zhangsan xiangxin _ mai-le shu
who Zhangsan believe bought books

Note that the major argument for such a view is the empirical observation that conditions that constrain c visible movement in some languages (see (12)) restrict the scope-taking behavior of certain phrases that have not been moved visibly (see (15)) in other languages. If one presupposes or shows that the relevant constraints c are formal-syntactic in nature, the conclusion that there is a syntactic component with invisible movement is quite compelling—but one must keep in mind that the overt movement operations constrained by c serve the purpose of scope-taking too.

Thus, one can also take quite a different perspective on facts as represented by (12) and (15): language could be governed by an elaborate system of *semantic constraints* on scope taking, and apparent syntactic conditions on movement might turn out to be superficial consequences of these constraints. Such semantic constraints affect the distribution of, for example, question words quite independently of whether they are moved syntactically or not. Proposals that can be related to this idea can be found in the literature (see, e.g., Erteshik-Shir 1977, Szabolsci and Zwarts 1997), but a full semantic account of constraints on movement has yet to be developed.

6. *Structural Hierarchies*

The previous section illustrated the difficulty of disentangling syntax and semantics in the domain of constraints on movement. This difficulty reappears when one considers the role which the geometry of abstract syntactic structures plays for quite a variety of relations and processes. (18)–(20) illustrate probably universal asymmetries that are cases in point: the pronoun *he* can be bound semantically by *no one* in (18a) only, but not in (18b): (18b) cannot have the interpretation sketched in (18c). *Ever* in (19) is a so-called 'negative polarity item.' It can appear in a sentence only if there is a negating expression present in that sentence. *Nobody* manages to license the polarity element *ever* in (19a), but not in (19b). (20) illustrates the fact that question (wh-) phrases are almost always moved to the left (if at all), but not the right.

(18) (a) No one thinks that he is incompetent.
(b) *He thinks that no one is incompetent.
(c) $\neg \mathrm{V}x$ (x thinks: x is incompetent)

(19) (a) Nobody thinks he will ever have enough money.
(b) *Bill ever thinks nobody will have enough money.

(20) (a) you tell me that she likes **who**
who do you tell me that she likes?
(b) you tell **who** that she likes Mary
*you tell that she likes Mary **who**

Such asymmetries were noted very early in modern syntax theory. After several unsuccessful attempts involving crucial reference to precedence, Culicover (1976) and Reinhart (1976) discovered the relevant purely structural factor at more or less the same time. The term *c-command* goes back to Reinhart (1976). We have reformulated it here as in (21), in a way that fits easily into present-day assumptions.

(21) a *c-commands* b if b is part of the structural sister of a

Recall that argument hierarchies are mapped, e.g., on to hierarchical abstract deep-structures, which are then transformed to surface structures by further operations. For (22), one might postulate a surface

structure such as (23) in the bracket notation, and (24) in the tree notation.

(22) the dangerous tiger has killed the old buffalo

(23) [the [[dangerous] [tiger]]] [has [[killed] [the [[old] [buffalo]]]]]]

(24)

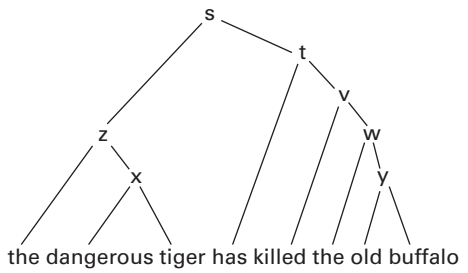

the dangerous tiger has killed the old buffalo

We have labeled the nodes in the tree (24) for ease of reference. We say that A *dominates* B iff there is a uniquely downwards directed path in the tree from A to B. Thus, *s* dominates *x*, but neither *y* nor *z* dominate *w*. A immediately dominates B iff A dominates B and there is no C that dominates B but not A. A is a sister of B if there is a C such that C immediately dominates A and B. Therefore, (21) implies that *z* c-commands *w* but not vice versa.

The following generalizations then hold for natural languages:

(25) A noun phrase A can bind a pronoun B only if A c-commands B.

(26) A negating expression A licenses a polarity element B only if A c-commands B.

(27) An element can move from position A to position B only if A c-commands B.

Let us exemplify the impact of these statements with (25) only. The examples in (18) translate into the simplified partial structural representations (28). It is easy to figure out that *no one* c-commands *he* in (28a), but not in (28b), so that (25) correctly predicts that *no one* can bind the pronoun in (18a) only.

(28) (a) [no one [thinks [that [he [is [incompetent]]]]]]

(b) [he [thinks [that [no one [is [incompetent]]]]]]

Reference to the structural notion c-command in (25) cannot be replaced by precedence, because such an account would fail to explain why *nobody* can neither bind the pronoun nor license the polarity element in (29).

(29) *a letter to nobody would ever please him

Presumably, the establishment of syntactic relations between A and B always presupposes that A c-commands B, or vice versa.

Kayne (1994) has proposed that elements α that asymmetrically c-command β precede β in all natural languages. If his approach is correct, we would understand why the basic orders subject–verb–object (SVO, English), SOV (Japanese), and VSO (Irish) predominate among the world's languages; subjects occupy a hierarchically higher position in sentence structure than objects, and should therefore precede them if Kayne is right.

7. *Operations Affecting Grammatical Relations*

That the formation of questions, relative clauses, topicalization, etc. involves a preposing operation in certain languages such as English seems to be accepted quite generally, There may even be psycholinguistic evidence suggesting that such movement dependencies are processed in a special way by the human parser (see, e.g., Frazier and Clifton 1993).

In addition to such dependencies, in which phrases are displaced in order to encode semantic or pragmatic functions (so-called A-bar-movement), there are grammatical operations such as 'passive' (as in (30)) or 'raising to subject' (as in (31)) which primarily affect the grammatical functions of arguments: in a thematic sense, *a moose* in (30c) has properties of an object (it expresses the patient role); in other respects, it behaves like a subject (in terms of position and agreement with the verb).

(30) (a) one shot a moose

(b) there was a moose shot _

(c) a moose was shot _

(31) (a) it seems that he likes Mary

(b) he seems _ to like Mary

That such operations affecting grammatical functions should be expressed in terms of movement (see (32)) is less obvious (at least for languages in which grammatical functions are not straightforwardly encoded positionally) and less widely accepted.

(32) *d-structure*

_ was shot a moose

s-structure

a moose was shot _

Although the proper analysis of these processes is still subject to some debate, a number of locality restrictions on the process have, however, been uncovered, which are illustrated in (33).

(33) (a) I shoot a moose

a moose was shot

(b) I laugh at him

he was laughed at

(c) I expect her to win

she was expected to win

(d) I expect she will win

*she was expected will win

(e) I prefer very much for you to kiss Mary

*Mary was preferred very much for you to kiss

The passive operation may map a direct object (33a), an object in a prepositional phrase (33b), and the subject of a nonfinite complement (33c) to the subject position, but fails to be able to establish such a relation for the subject of an embedded finite clause (33d), or any object in an embedded clause (33e).

The restrictions may be slightly different in other languages, as the contrast between English and German in (34) shows, but the crucial observation is that similar locality restrictions can be observed with the binding of reflexive pronouns, at least in the most restrictive subcase of this domain, as it is realized in English; see (35).

(34) (a) he forgot to repair the car
*the car was forgotten to repair

(b)

er	vergass	den Wagen
he.nom	forgot	the.acc car
zu.repaieren		
to repair		
der Wagen	wurde	zu reparieren
the.nom car	was	to repair
vergessen		
forgotten		

(35) he shot himself
he laughs at himself
he expects himself to win
*he expects himself will win
*he prefers very much for her to kiss himself

The set of constructions to which these locality restrictions common to passivization and the grammar of reflexives apply is larger: it includes quantifier scope. A quantifier s can always take scope over a quantifier t if s c-commands t at s-structure. Thus, since subjects are higher in the structural hierarchy than objects, sentence (36a) allows a reading in which the existential quantifier takes scope over the universal quantifier

(36) (a) some man loves every woman
(b) there is a man such that this man loves every woman
(c) for every woman there is a man (possibly a different one for each woman) such that that man loves the woman

In English, however, a structurally lower quantifier s in position S may additionally take scope over t in position T if S and T could interact in terms of, say, passive formation, or the binding of reflexive pronouns. Thus, (36a) allows the further interpretation indicated in (36c), in which the object quantifier takes scope over the subject quantifier. A comparison of (33), (35), and(37) exemplifies the parallelism between wide scope options for a hierarchically lower quantifier, and the possibility of passive formation/binding of reflexive pronouns. See also Hornstein (1995).

(37) some woman laughs at every man
(wide scope for *every man* **possible**)
some woman expects every man to be a loser
(wide scope for *every man* **possible**)
some woman expects every man will be a loser
(wide scope for *every man* **impossible**)
some woman prefers very much for me to kiss every man
(wide scope for *every man* **impossible**)

If formal syntax is responsible for the restrictions exemplified in (33), the binding rules for reflexives and the scope taking behavior of quantifiers must be reducible to syntactic operations too. However, it is hard to exclude the possibility that some extra-syntactic constraints governing quantifier scope or reflexive binding have side effects on passivization or raising, so that (33) would find a nonsyntactic account. Note, furthermore, that the impression that the same factor constrains the three different types of process discussed here may be spurious. Thus, in many languages (e.g., Swedish) the domain for the binding of reflexive pronouns is much larger than the domain for passive formation.

8. *A Note of Autonomy*

As a reaction to so-called Generative Semantics, the hypothesis of the autonomy of syntax governed much of the history of current syntactic theorizing (see Newmeyer 1980): syntax was considered to be the only 'generative' component, the results of which are merely *interpreted* by semantics and by phonology. The preceding sections have revealed that it is often difficult to draw a sharp boundary between syntax and semantics, and strong adherence to the autonomy hypothesis may nowadays reflect a rather parochial perspective.

The same may be true for the interaction of phonology and syntax. The laws of the placement of clitics (e.g., unstressed pronouns) as we find them in Romance and Slavic languages often make crucial reference to phonological conditions (no clitic element must appear in the initial position of an intonation phrase) and syntactic conditions (elements bearing a given case etc. must appear in the leftmost position of, say, the verbal projection) at the same time. Phonological conditions may even be responsible for the application of syntactic operations such as verb movement.

The idea of a completely autonomous and encapsulated syntactic component, in which all processes can be understood in purely formal syntax-based terms only, in which syntactic processes are driven by syntactic conditions only, which was most popular in the GB tradition, therefore has to be modified, at least.

9. *Approaches to Universals*

The discovery of syntactic universals and their explanation has always been the major goal of syntactic theory. The typological tradition tries to attain this goal by comparing a large and representative sample of languages with respect to specific grammatical properties. Often this presupposes that the constructions can only be looked at in less detail than in the generative tradition. In the generative approach, the idea is to uncover abstract universal laws by an in-depth analysis of a few languages only.

The GB approach (Chomsky 1981) tried to identify a small set of simple yet universally valid syntactic principles, defining a Universal Grammar. It was assumed, especially in later stages of the theory, that formal differences between languages reduce to differences in the lexical inventory. The quest for such simple yet universally valid syntactic generalization was not successful, however, so the GB framework was replaced by, essentially, two new models. The Minimalist Program (Chomsky 1995) tries to understand syntactic properties in terms of so-called interface conditions (phonology, semantics). It may have explanatory value, but its empirical coverage is sometimes limited. Optimality Theory (OT) (Prince and Smolensky 1993, Grimshaw 1997) assumes that there are universally valid and simple conditions. Phonology, syntax, and semantics are not necessarily isolated subcomponents. OT acknowledges that simple principles will be in conflict with each. Such conflicts are resolved by ordering the principles hierarchically. Hierarchies are different in various languages, the only source of grammatical differences.

In particular, the major principles assumed in OT may correspond to 'ideal types' that languages may come close to in different degrees (as a function of the ranking of the relevant principles), so OT may have moved the generative enterprise closer to typology. This may be true in yet another sense; in contrast to GB, OT declared explicitly itself as a theory of variation and its explanation.

See also: Constituent Structure: Syntactic; Linguistics: Incorporation; Syntactic Aspects of Language, Neural Basis of; Syntactic Control; Syntax–Phonology Interface; Syntax–Semantics Interface

Bibliography

Baker M 1996 *The Polysynthesis Parameter*. Oxford University Press, New York

Bossong G 1998 Le marquage de l'expérient dans les langues d'Europe. In: Feuillet J (ed.) *Actance et Valence dans les Langues de l'europe*. Mouton de Gruyter, Berlin, pp. 259–94

Bresnan J (ed.) 1982 *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA

Chomsky N 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands

Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA

Chomsky N 1973 Conditions on transformations. In: Anderson S, Kiparsky P (eds.) *A Festschrift for Morris Halle*. Holt, Rinehart and Winston, New York, pp. 232–86

Chomsky N 1981 *Lectures on Government and Binding*. Foris, Dordrecht, The Netherlands

Chomsky N 1986 *Barriers*. MIT Press, Cambridge, MA

Chomsky N 1995 *The Minimalist Program*. MIT Press, Cambridge, MA

Culicover P 1976 A constraint on coreferentiality. *Foundations of Language* **14**: 109–18

Dowty D 1991 Thematic protoroles and argument selection. *Language* **67**: 574–619

Erteshik-Shir N 1977 *On the Nature of Island Constraints*. IULC, Bloomington, IN

Fillmore C 1968 The case for case. In: Bach E, Harms R (eds.) *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York, pp. 1–88

Frazier K, Clifton C 1993 *Construal*. MIT Press, Cambridge, MA

Gazdar G, Klein E, Pullum G, Sag I 1985 *Generalized Phrase Structure Grammar*. Blackwell, London

Grimshaw J 1997 Projections, heads and optimality. *Linguistic Inquiry* **28**: 373–422

Gruber J 1965 *Studies in lexical relations*. Diss., MIT, Cambridge

Haider H 1993 *Deutsche Syntax—generativ*. Narr, Tübingen, Germany

Hale K 1983 Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory* **1**: 5–47

Hjelmslev L 1972 [1935–7] *La catégorie des cas. Étude de grammaire générale*. W Fink, Munich, Germany

Hornstein N 1995 *Logical Form*. Blackwell, Oxford, UK

Huang C T J 1982 Move WH in a language without Wh movement. *The Linguistic Review* **1**(4): 369–416

Jackendoff R 1990 *Semantic Structures*. MIT Press, Cambridge, MA

Jakobson R 1971, [1936] Beitrag zur allgemeinen Kasuslehre. Gesamtbedeutungen der russischen Kasus. *Selected Writings II*. Mouton, The Hague, pp. 23–71

Kager R 1999 *Optimality Theory*. Cambridge University Press, Cambridge, UK

Kayne R 1994 *The Antisymmetry of Syntax*. MIT Press, Cambridge, MA

Marantz A 1984 *On the Nature of Grammatical Relations*. MIT Press, Cambridge, MA

Müller G 2000 *Elemente der optimalitätstheoretischen Syntax*. Stauffenburg, Tübingen, Germany

Newmeyer F 1980 *Linguistic Theory in America*. Academic Press, New York

Perlmutter D (ed.) 1983 *Studies in Relational Grammar 1*. University of Chicago Press, Chicago, IL

Prince A, Smolensky P 1993 Optimality theory Ms., Rutgers University, NJ

Reinhart T 1976 The syntactic domain of anaphora. Ph.D. diss., MIT, Cambridge, MA

Ross J R 1967 Constraints on variables in syntax. Ph.D. diss., MIT, Cambridge, MA

Rizzi L 1990 *Relativized Minimality*. MIT Press, Cambridge, MA

Szabolcsi A, Zwarts F 1997 Weak islands and an algebraic semantics for scope taking. In: Szabolcsi A (ed.) *Ways of Scope Taking*. Kluwer, Dordrecht, The Netherlands

Wunderlich D 1997 Cause and the structure of verbs. *Linguistic Inquiry* **28**(1): 27–68

G. Fanselow

Syntax–Phonology Interface

An individual word, pronounced in isolation, will have a characteristic pronunciation. Yet the pronunciation of a sentence, made up of a sequence of words, is not merely a stringing together of these individual pronunciations. The pronunciation of a sentence, i.e.,

its phonetic realization, is a function of the surface phonological representation of the sentence. Standard models of the organization of a grammar hold that this phonological representation interfaces with the surface syntactic representation of the sentence, the latter being made up of the sequence of the underlying phonological representations of the words and morphemes that are actually pronounced, and their organization into a syntactic word and phrase structure. This organization of the grammar permits in principle that the syntactic representation of a sentence may influence its phonological representation, and therefore allows an explanation for why sentence phonology is not simply the phonology of the individual words of the sentence strung together. Consider, for example, the sentence pairs *No people will go* vs. *No, people will go* and *People will go happily* vs. *People will go, happily*. In both cases, commas indicate a difference in syntactic structure within the pairs. That syntactic structure difference is reflected in a difference in pronunciation, specifically here in a difference in intonation. Effects like these of syntax on phonology have been studied most commonly, in part because earlier models of generative grammar sanctioned only effects in this direction. It is also possible that effects might go in the opposite direction, with phonological principles constraining the range of acceptable syntactic representations at the interface; some recent research explores this question.

Two major lines of thinking on the influence of syntax on phonology can be distinguished. The *prosodic structure hypothesis* holds that the phonological representation of a sentence is organized into a prosodic constituent structure which is independent of, but related to, the surface syntactic structure of a sentence (Selkirk 1986, 1995; Nespor and Vogel 1986; Truckenbrodt 1999). It claims, moreover, that the syntax of a sentence can impinge directly only on this prosodic structure, namely on the organization into phonological words or phrases, and on the distribution of prosodic heads (or stresses). Other apparent effects of syntax on phonology, e.g., the choice of segmental variants or the placement of certain phrasal tones, are claimed to be indirect, mediated by the prosodic structure organization. And since the phonetic interpretation of a sentence is based on surface phonological representation, it follows that there are no direct effects of syntax on phonetic form. If the prosodic structure hypothesis embodies the correct theory of the domain structure for sentence phonology, then the theory of grammar must distinguish between the surface phonological representation of a sentence (PR), which would include this prosodic structure, and its surface syntactic representation (PF). The *direct access hypothesis* (Kaisse 1985, Odden 1995), by contrast, holds that phenomena of sentence phonology and phonetics may be defined directly from the surface syntactic word and phrase structure. It does not require a theory of grammar that distinguishes the surface syntactic representation of a sentence crucially from its surface phonological representation, at least in its hierarchical structure. Evidence appears to favor the prosodic structure hypothesis.

1. *The Prosodic Structure Hypothesis*

A prosodic structure is a well-formed labeled bracketing or tree. The constituents of prosodic structure belong to distinct prosodic categories, arranged in a *prosodic hierarchy*.

(1) The Prosodic Hierarchy:
Utterance (Utt)
Intonational Phrase (IP)
Major Phonological Phrase (MaP)
Minor Phonological Phrase (MiP)
Prosodic Word (PWd)
Foot
Syllable

In the unmarked case, it is claimed, prosodic structure is *strictly layered*, in the sense that a constituent of a higher level in the hierarchy immediately dominates only constituents of the next level down in the hierarchy. In addition, within a prosodic constituent, in the unmarked case, one of the daughter constituents constitutes the *prosodic head*, the locus of prominence or stress. Compelling support for the claim that the domain structure of sentence phonology and phonetics has the formal properties of such a structure comes from evidence of *domain convergence* and *domain layering* within individual languages. Yet a direct access approach could potentially model these effects. Crucial support for the prosodic structure hypothesis comes from the fact that the constraints which define the hierarchical phonological domain structure are heterogeneous in type, including *prosodic structure markedness constraints* which are properly phonological in character, appealing only to properties of surface phonological representation, and *syntax-prosodic structure interface constraints*, which call for features of the syntactic representation to be reflected in phonological representation. Effects of prosodic markedness constraints show that the domain structure of surface phonology is not strictly determined by the syntax, only influenced by it.

1.1 *Domain Convergence*

In English declarative sentences, the right edge of a major phrase (MaP) in prosodic structure is marked by the presence of a low tone. In (1a,b) a L- marks the low target tone found on the final syllable of a MaP.

(2) (a) (No^{L-}) (animals are allowed^{L-}) =
No, animals are allowed.
[iænɨmlz]
(b) (No animals are allowed^{L-}) =
No animals are allowed.
[ænɨmlz]

We also see a segmental reflex of that same MaP organization: a glottal stop appears as the onset of the first syllable of *animals*, [ɨænɨmlz], in (2a), where it is phrase-initial, but not in (2b). This glottalization effect is arguably phonetic in character (Dilley and Shattuck-Hufnagel 1996). In (3a,b) we see an additional MaP effect: a contrast in the pronunciation of the function word *to*.

(3) (a) (They're allowed to graze there^{L-}) (by law^{L-}).
[tə]
(b) (They're allowed to^{L-}) (by law^{L-})
[tuw]

To appears in its stressless, vowel-reduced weak from [tə] in (3a), where it is medial within the MaP, but in its stressed, full vowel strong form [tuw] when it appears at the right edge of MaP in (3b) due to the ellipsis of the following verb phrase. In English, monosyllabic function words like *to* generally appear in strong form only at the right edge of MaP (or when they bear contrastive stress) (Selkirk 1995). These three different phenomena which converge on major phrase in English together show that both edges of these constituents are simultaneously relevant to defining phonetic and phonological phenomena.

In Bengali sentence phonology (Hayes and Lahiri 1991), MaP is marked at its right edge by a H- tone. The left edge of MaP is the locus of further phenomena: the word that is leftmost in a MaP is the head of that phrase, and a Low pitch accent, L*, is located on the initial syllable of this stressed word:

(4) (ʃæ^{L*}moli^{H-}) (ra^{L*} m-er bari^{H-}) (dʰu^{L*}ketʃʰil^{H-})$^{L°}$
Shamoli Ram's house entered
Shamoli entered Ram's house.

In addition, there are segmental assimilation phenomena which apply optionally both within and between words. These assimilations operate across the span of the MaP, but are blocked if the sequence of segments belongs to different phrases:

(5) (ɔ^{L*}mor^{H-}) (tʃa^{L*}dor^{H-}) (ta^{L*}ra-ke^{H-}) (díe^{L*}tʃʰe^{H-})$^{L°}$
[r] [tʃ] [r] [t]
Amor scarf Tara–to gave
Amor gave the scarf to Tara.

(6) (ɔ^{L*}mor tʃador tara-ke^{H-}) (díe^{L*}tʃʰe^{H-})$^{L°}$
[tʃtʃ] [tt]
Amor gave the scarf to Tara.

The faster pronunciation in (6) is organized into fewer MaP than in the more deliberate pronunciation of the same sentence in (5). This difference in domain organization is indicated both by the patterns of tone distribution and by the assimilation of /r/. This Bengali phrasal domain, then, shows right and left domain-edge phenomena, as well as phenomena defined on the prominent head of the domain and across the span of the domain.

1.2 Domain Layering

The edge of a clause in the syntax typically coincides with the edge of what prosodic structure theory calls the intonational phrase (IP), the prosodic constituent immediately superordinate to the MaP in the prosodic hierarchy. In English, a final rising tonal contour on the last syllable of word (notated$^{L-H°}$), sometimes referred to as the continuation rise, is taken to indicate the presence of the right edge of IP (Beckman and Pierrehumbert 1986), indicated with curly braces.

(7) {(Since herds of grazing cows^{L-})
[hrdz]
(have been allowed in her meadow^{L-})$^{H\%}$}
[əv] [ɾ]
{(her flowers have all disappeared^{L-})}
[hɾ] [əv]

The right edge of an IP typically is accompanied by a short pause or a significant lengthening of the final syllable. In English (and other languages) the left edge of IP is, moreover, the site of phonetic strengthening effects that are potentially different in kind or in degree from those seen with the major phrase (Keating et al. 1998). For example, an underlying /h/ in English will typically fail to be pronounced if it is in an unstressed syllable (compare stressed *herds* to stressless *her* in the first clause), but at the beginning of an IP /h/ must necessarily be pronounced, even in a stressless syllable (compare stressless *her* in the second clause to stressless *her* in the first). By contrast, there is the possibility of absence of /h/ in the stressless *have* which is MaP-initial in the first clause. This difference in appearance of [h] provides further testimony of the difference in the two levels of domain structure.

In Japanese, a rise (LH) in pitch which is referred to as Initial Lowering is found at the left edge of a minor phrase (MiP) (Poser 1984, Kubozono 1993), indicated with angle brackets in (8b).

(8) (a) Yamaai-no yamagoya-no uraniwa-no umagoya-ni kabi-ga hae^{H*}-ta
[[[[mountain village-GEN] hut-GEN] backyard-GEN] barn-LOC] [mold-NOM] [grow-PAST]
The barn in the backyard of a hut in a mountain village grew moldy.
(b) $^{L}<^{H}$ Yamaai-no yamagoya-no $>^{L}$ $<^{H}$ uraniwa-no umagoya-ni $>\uparrow^{L}<^{H}$ kabi-ga hae$^{H*}\downarrow$-ta $>)^{L}$

Sentence (8b) shows an organization into binary MiP. In (9b), by contrast, where all the words have a lexical accent, we see that there are as many instances of MiP, and hence of Initial Lowering, as there are of accents:

(9) (a) Yama^{H*} gata-no yama^{H*} dera-no one:^{H*} san-ga mayone:^{H*} zu-o ho:ba^{H*} tteiru-wa
[[[Yamagata-GEN] mountain temple-GEN] young-lady-NOM] [[mayonnaise-ACC] filling-her-mouth]
The young lady from a mountain temple in Yamagata was filling her mouth with mayonnaise.
(b) ($^{L}<^{H}$yama$^{H*}\downarrow$gata $>^{L}<^{H}$yama$^{H*}\downarrow$dera $>^{L}$ $<^{H}$one:$^{H*}\downarrow$san-ga $>\uparrow^{L}<^{H}$ mayone:$^{H*}\downarrow$zu-o $>^{L}<^{H}$ho:ba$^{H*}\downarrow$tteiruwa $>)^{L}$

The difference in patterns of minor phrasing seen in (8) and (9) are the result of prosodic markedness constraints, to be discussed below.

At the level of MaP, the level above MiP in the prosodic hierarchy, two tonal phenomena, are defined: catathesis (↓), and upward pitch reset (↑) (Poser 1984, Pierrehumbert and Beckman 1988). Catathesis is an accent-induced downstepping of the pitch range, while upward pitch reset returns the pitch range to a higher level. Only material within the same MaP undergoes catathesis. Upwards reset, seen in both (13) and (14), appears at the left edge of MaP. These examples, then, illustrate the basic strict layering of unmarked hierarchical prosodic structure: here a MaP is parsed into a sequence of MiP.

1.3 Nonsyntactic Determinants of Phonological and Phonetic Domain Structure

A direct access theory could model patterns of edge-sensitive phonetic and phonological phenomena as long as the edges were definable solely in syntactic terms. But demonstrably nonsyntactic factors in phonological domain organization, such as prosodic markedness constraints and speech rate, show that the influence of syntax is indirect.

1.3.1 The interaction of prosodic markedness constraints and interface constraints. In optimality theory (Prince and Smolensky 1993), phonological markedness constraints call for some ideal phonological target output shape. Of particular importance to our topic, markedness constraints compete with a class of constraints called faithfulness constraints, which call for properties of the surface phonological representation to reflect, or be identical to, properties of a related grammatical representation. Syntax–phonology interface constraints can be understood as a variety of faithfulness constraint, one of the input–output variety, if surface syntactic representation (PF) and surface phonological representation (PR) are in an input–output relation, or one of an output–output variety, if the PF and PR relation are evaluated in parallel, not serially.

In the examples of Japanese minor phrasing cited in (8) and (9), prosodic markedness constraints complement the effects of syntax–phonology interface constraints, producing phrasing where it would not be called for by the syntax. A prosodic constraint calling for a MiP to be binary in its PWd composition, call it Binary MiP, results in the binary phrasing seen in (8). Another prosodic constraint, call it MiP Accent, calls for a MiP to contain at most one accent; it is responsible for the appearance of the nonbinary MiP in (14). In Japanese, MiP Accent takes precedence over Binary MiP, as in (9), a relation that can be expressed in the grammar of Japanese with an optimality-theoretic constraint ranking: MiP Accent ≫ Binary MiP.

The ideal binary shape for MiP, which appears in (8), fails to appear in sentences with somewhat different syntactic structures, showing that certain syntax–phonology interface constraints also outrank the Binary MiP markedness constraint in the grammar of Japanese. The simple three word subject–object verb sentence (10) divides into two major phrases (MaP) at the break between subject and object, with the result that the subject consists of a nonbinary minor phrase (MiP), in violation of Binary MiP:

(10) (a) Inayama-ga yuujin-o yonda
[[Inayama-NOM] [[friend-ACC] call-PST]]
Mr Inayama called his friend.
(b) $({}^{L}<{}^{H}$Inayama-ga$>\uparrow({}^{L}<{}^{H}$yuujin-o yonda$>)^{L}$

The MaP break here is imposed by a syntax–phonology interface constraint expressible as the alignment constraint (Selkirk and Tateishi 1988, 1991):

(11) Align-L (XP, MaP)
The left edge of a maximal projection in syntactic representation (PF) corresponds to the left edge of a major phrase in surface phonological representation (PR).

The noun phrase (12), which consists of an accentless Adjective–Adjective–Noun sequence, consists of a single MaP (as shown by lack of internal upward pitch range reset), but nonetheless shows a MiP break between the two adjectives, leaving the first MiP nonbinary.

(12) (a) Amai akai ame-ga
[sweet [red [candies]]]
(b) $({}^{L}<{}^{H}$amai$>{}^{L}<{}^{H}$akai ame-ga$>)^{L}$
NOT: $*({}^{L}<{}^{H}$amai akai ame$>)^{L}$

At play here is an interface constraint sensitive to syntactic branching (Kubozono 1993):

(13) Align- L (X^{br}, MiP).
The left edge of a branching constituent in syntactic structure corresponds to the left edge of a minor phrase in prosodic structure.

These two syntax–phonology alignment constraints outrank Binary MiP in the grammar of Japanese, causing the violations of Binary MiP seen in (10) and (12):

(14) Align-L (XP, MaP), Align- L (X^{br}, MiP) ≫ Binary MiP.

It is because neither of these interface constraints is applicable in the left-branching structure of the subject noun phrase in (8) that we see the effects of the lower-ranked Binary MiP emerge there. In other words, where the lower-ranked Binary MiP is not forced to be violated by the higher ranked interface constraints, it produces effects that are *complementary* to the effects of the interface constraints.

We also find cases where an expected effect of a syntax–phonology interface constraint *fails to appear* because of pressure from a competing prosodic markedness constraint. This can be modeled by

ranking the prosodic constraint higher than the relevant interface constraint. (15) shows the expected MaP break at the left edge of VP in a neutral Focus sentence in Japanese.

(15) A neutral Focus sentence in Japanese

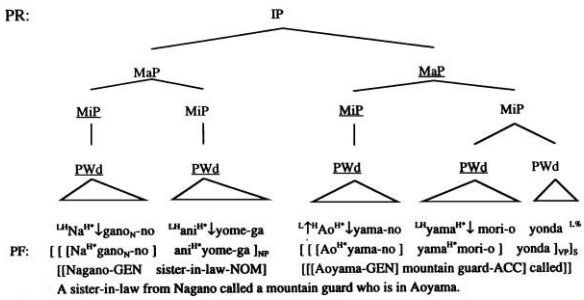

Sentence (16) contains a text which differs only in having a contrastive Focus on the last word of the subject NP, *ani'yome*. The phonological domain structure of this sentence shows an absence of the syntactically expected MaP break at the left edge of the VP, as well as the unexpected presence of a MaP break at the left edge of the Focus constituent (Poser 1984, Pierrehumbert and Beckman 1988, Nagahara 1994).

(16) The same sentence, except with *aniyome* as FOCUS

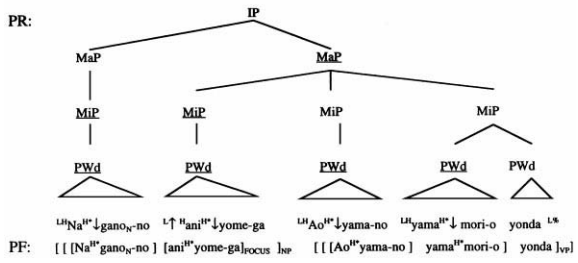

Truckenbrodt (1995) argues that the source of these phrasing effects in Focus sentences lies in a syntax–phonology interface constraint calling for a Focus to be more prosodically prominent than any other element within the Focus domain.

(17) Focus Prominence

The string in PR which corresponds to a string in PF dominated by Focus must contain a prosodic prominence that is greater than any other prominence contained in the Focus domain.

In the case where the domain of the Focus is the entire sentence, as in (16), respect of this constraint means that the Focus element and all the prosodic phrases that dominate it will be the head(s) of the intonational phrase corresponding to that sentence in PR (underlined). A prominent constituent is subject to a prosodic markedness constraint that calls for it to be aligned with an edge of the superordinate constituent (McCarthy and Prince 1993). The violation in (17) of the interface constraint (11) calling for a MaP edge at the left edge of VP is claimed to be driven by the higher-ranked prosodic constraint Align-R (MaP_{head}, IP).

This constraint calls for the prominent MaP in the IP that dominates the Focus in (17) to align with the right edge of the intonational phrase, and so prevents the existence of any other MaP following the Focus. This analysis is captured in the ranking in (18).

(18) Focus Prominence, Align-R (MaP_{head}, IP) $\gg$ Align-L (XP, MaP).

So, here, a syntactic effect is overridden by a prosodic effect.

1.4 Speech Rate

One additional sort of nonsyntactic factor on phonological domain structure that has been recognized for some time in the literature is speech rate. In Bengali, a faster speech rate gives rise to larger phonological phrases (see Sect. 1.1). Speech rate plays a role as well in MiP organization in Japanese; in slower speech, single words may easily constitute a MiP (Selkirk and Tateishi 1988). Since syntactic structure does not vary with speech rate, these rate-based differences in phrasing give additional evidence for an autonomous, nonsyntactic, prosodic structural representation of phonological domains.

2. The Syntax–Prosodic Structure Interface: Serialist or Parallelist?

The classic generative model of the relation between syntax and phonology is an input–output model, in which surface syntactic representation, the output of the syntactic component, is input to the phonological component, whose output is the surface phonological representation. Most approaches to modelling the syntax–prosodic structure interface have assumed this serialist, unidirectional, input–output model, with the prediction that syntax may influence phonology, but not vice versa. They have posited constraints, algorithms or rules relating the two representations which have the general form 'If the surface syntactic representation of the sentence has property S, then the surface phonological representation will have property P.'

In the Align/Wrap theory of the interface (Selkirk 1986, 1995, Truckenbrodt 1999), the demarcational Align class of constraints calls for the edge of any syntactic constituent of type C_S in the surface syntax to align with the edge of a prosodic constituent of type C_P in the surface phonology, while the cohesional Wrap class of constraints calls for a syntactic constituent of type C_S in the surface syntax to be contained within a

prosodic constituent of type C_P in the surface phonology. There have been other, relation-based, proposals concerning the nature of syntax–prosodic structure interface constraints as well, including Nespor and Vogel 1986, McHugh 1990, Kim 1997, all understood to be consistent with the prosodic structure hypothesis, and with the input–output architecture.

An input–output model of the syntax–phonology interface makes the essential claim that constraints on phonological output representation do not interact with constraints on the surface syntactic representation. Two sorts of apparent challenges for the input–output model have recurred in recent literature. One sort concerns clitics whose distribution is governed by both syntactic and prosodic constraints, such as the second position clitics of Serbo-Croatian (Inkelas and Zec 1990). Yet because the syntax itself affords options in the positioning of second position clitics, it is not necessary to construe the prosodic subcategorization of a clitic as outranking a constraint on syntactic word order. Another sort of challenge concerns the alleged dependence of word order on the positioning of focus-related prosodic prominences in the sentence, as suggested by Vallduvi (1991) for Catalan, or Zubizarreta (1998) for Spanish. However, since the latter sort of phenomenon potentially can receive a purely syntactic treatment in which the Focus properties of syntactic representation are crucial in determining word order (see, e.g., Grimshaw and Samek-Lodovici 1998), this sort of case is not yet compelling. It remains to be seen whether the syntax–phonology interface is an input–output relation or whether it is instead a two-way street.

See also Generative Grammar; Phonology; Syntax

Bibliography

Beckman M, Pierrehumbert J 1986 Intonational structure in English and Japanese. *Phonology* **3**: 255–309

Dilley L, Shattuck-Hufnagel S 1996 Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* **24**: 424–44

Grimshaw J, Samek-Lodovici V 1998 Optimal subjects and subject universals. In: Barbosa P, Fox D, Hagstrom P, McGinnis M, Pesetsky D (eds.) *Is the Best Good Enough? Optimality and Competition in Syntax*. MIT Press, Cambridge, MA, pp. 193–220

Hayes B, Lahiri A 1991 Bengali intonational phonology. *Natural Language and Linguistic Theory* **9**: 47–96

Inkelas S, Zec D 1990 Prosodically constrained syntax. In: Inkelas S, Zec D (eds.) *The Phonology–Syntax Connection*. CSLI, Chicago, IL, pp. 365–78

Kaisse E M 1985 *Connected Speech: The Interaction of Syntax and Phonology*. Academic Press, San Diego, CA

Keating P, Fougeron C, Cho T, Hsu C-S 1998 Domain-initial articulatory strengthening. *Papers in Laboratory Phonology 6*. Cambridge University Press, Cambridge, UK

Kim N-J 1997 *Tones, Segments, and their Interaction in North Kyungsang Korea*. Ph.D. diss. Ohio State University, Columbus, OH

Kubozono H 1993 *The Organization of Japanese Prosody*. Kurosio, Tokyo

McCarthy J J, Prince A 1993 Generalized alignment. In: Booij G, Marle J v (eds.) *Yearbook of Morphology*. Kluwer, Dordrecht, The Netherlands, pp. 79–153

McHugh B 1990 The phrasal cycle in Kirunjo Chaga tonology. In: Inkelas S, Zec D (eds.) *The Phonology–Syntax Connection*. University of Chicago Press, Chicago, IL, pp. 217–42

Nagahara H 1994 *Phonological Phrasing in Japanese*, Ph.D. diss. UCLA, Los Angeles

Nespor M, Vogel I 1986 *Prosodic Phonology*. Foris, Dordrecht, The Netherlands

Odden D 1995 Phonology at the phrasal level in Bantu. In: Katamba F (ed.) *Bantu Phonology and Morphology*. LINCOM EUROPA, Munich, Germany, pp. 40–68

Pierrehumbert J, Beckman M 1988 *Japanese Tone Structure*. MIT Press, Cambridge, MA

Poser W J 1984 *The Phonetics and Phonology of Tone and Intonation in Japanese*. Ph.D. diss. Massachusetts Institute of Technology, Cambridge, MA

Prince A, Smolensky P 1993 Optimality theory: Constraint interaction in generative grammar. Ms, Rutgers University Center for Cognitive Science, New Brunswick, NJ

Selkirk E 1986 On derived domains in sentence phonology. *Phonology* **3**: 371–405

Selkirk E 1995 The prosodic structure of function words. In: Beckman J, Walsh Dickey L, Urbanczyk S (eds.) *Papers in Optimality Theory*. GLSA, Amherst, MA, pp. 439–70

Selkirk E, Tateishi K 1988 Constraints on minor phrase formation in Japanese. In: *Proceedings of the 24th Annual Meeting of the Chicago Linguistics Society*, Chicago Linguistic Society, Chicago, IL

Selkirk E, Tateishi K 1991 Syntax and downstep in Japanese. In: Georgopoulos C, Ishihara R (eds.) *Interdisciplinary Ap proaches to Language: Essays in Honor of S.-Y. Kuroda*. Kluwer, Dordrecht, The Netherlands, pp. 519–44

Truckenbrodt H 1995 *Phonological Phrases: Their Relation to Syntax, Focus and Prominence*, Ph.D. diss. Massachusetts Institute of Technology, Cambridge, MA

Truckenbrodt H 1999 On the relation between syntactic phrases and phonological phrases. *Linguistic Inquiry* **30**: 219–56

Vallduvi E 1991 The role of plasticity in the association of focus and prominence. In: *Proceedings of the Eastern States Conference on Linguistics* (*ESCOL*) *7*, pp. 295–306

Zubizarreta M L 1998 *Prosody, Focus and Word Order*. MIT Press, Cambridge, MA

E. Selkirk

Syntax–Semantics Interface

Anybody who speaks English knows that (1) is a sentence of English, and also has quite a precise idea about what (1) means.

(1) Every boy is holding a block.

English speakers know for infinitely many sentences that they are sentences of English, and they also know the meaning of these infinitely many sentences. A

central idea of generative syntax (see *Linguistics: Incorporation*) is that the structure of a sentence is given by a recursive procedure, as this provides the means to derive our knowledge about the infinite number of sentences with finite means. For the same reason, semanticists have developed recursive procedures that assign a meaning to sentences based on the meaning of their parts.

The syntax–semantics interface establishes a relationship between these two recursive procedures. An interface between syntax and semantics becomes necessary only if they indeed constitute two autonomous systems. This is widely assumed to be the case, though not entirely uncontroversial as some approaches (see *Grammar: Functional Approaches* and *Montague Grammar*) do not subscribe to this hypothesis.

Consider two arguments in favor of the assumption that syntax is autonomous: one is that there are apparently purely formal requirements for the well-formedness of sentences. For example, lack of agreement as in (2) renders (1) ill-formed (customarily marked by prefixing the sentence with an asterisk), though subject–verb agreement does not seem to make any contribution to the meaning of (1).

(2) *Every boy hold a block.

The special role of uninterpretable features for syntax comes out most sharply in recent work by Chomsky (1995), who regards it as one of the main purposes of syntax to eliminate such uninterpretable features before a sentence is interpreted.

On the other hand, there are also sentences that are syntactically well-formed, but do not make any sense (often marked by prefixing the sentence with a hatch mark). Chomsky's (1957, p. 15) famous example in (a) makes this point, and so does (b).

(3) (a) #Colorless green ideas sleep furiously.
(b) #She arrived for an hour.

Independent of the value of these arguments, the separation of syntax and semantics has led to tremendous progress in the field. So, at least it has been a successful methodological principle.

1. Basic Assumptions

Work on the syntax–semantics interface by necessity proceeds from certain assumptions about syntax and semantics. We are trying to keep to a few basic assumptions here. In the specialized articles on generative syntax (*Linguistics: Incorporation*), and *quantifiers* (*Quantifiers, in Linguistics*) many of these assumptions are discussed and justified in more depth.

For the syntax, we assume that sentences have a hierarchical constituent structure that groups words and subconstituents into constituents. Also, we assume that constituents can be moved from one part of the tree structure to another, subject to constraints of the kind Ross (1967) first described. The constituent structure of a sentence is captured by the kind of phrase structure tree illustrated in (4)(a). (4)(b) shows a phrase structure tree with a movement relation.

(4(a))

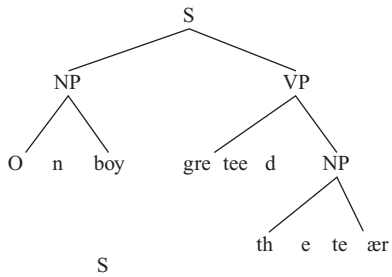

(4(b))

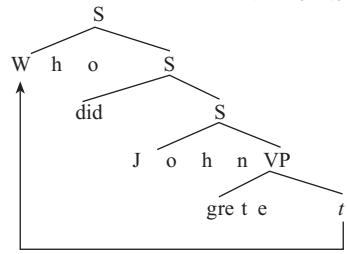

The task of semantics is to capture the meaning of a sentence. Consider first the term 'meaning.' In colloquial use, 'meaning' includes vague associations speakers may have with a sentence that do not exist for all speakers of a language: e.g., the sentence 'I have a dream' may have a special meaning in this colloquial sense to people familiar with Martin Luther King.

Semantics, however, is at present concerned only with reproducible aspects of sentence meaning. In particular, semanticists have focused on the question of whether a sentence is judged true or false in a certain situation. Part of what any speaker of English knows about the meaning of (1) is that it is true in the situation shown in picture A, but false in the situation shown in picture B.

(5) Every boy is holding a block.

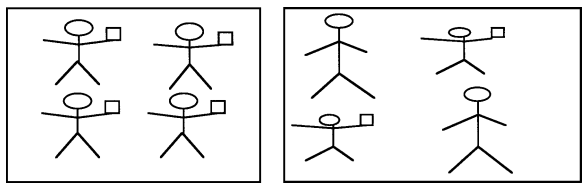

Picture A Picture B

All speakers of English are equipped with certain mental mechanisms that enable them to make this truth value judgment for (1) and similar judgments for infinitely many other sentences. An explicit theory can be given using techniques similar to those used in mathematical logic. In this approach, the meaning of a constituent is modeled by a mathematical object (an individual, a function, a set, or a more complicated

object). Complete declarative sentences such as (1) correspond to functions that assign to possible situations one of the truth values True or False. The basic intuition of entailment between sentences as in (6) is captured if for every situation to which the meanings of the premises (a) and (b) assign True, the meaning of the conclusion (c) also assigns to this situation True.

(6) (a) Every boy is holding a block.
(b) John is a boy.
(c) Therefore, John is holding a block.

2. Syntax–Semantics Correspondences

Though syntax and semantics are two autonomous recursive procedures, most researchers assume that there is a relationship between the two to be captured by the theory of the syntax–semantics interface. In particular, it seems to be the case that the steps of the recursion are largely the same. In other words, two phrases that form a syntactic constituent usually form a semantic constituent as well (Partee (1975) and others).

Consider (7) as an illustration of this.

(7)	(a) A smart girl	bought	a thin book.
	(b) A thin girl	bought	a smart book.
	Subject	**Verb**	**Object**

Syntacticians have argued that the subject and the object in (6) form constituents, which we call noun phrases (abbreviated as NPs). We see in (6) that the adjective that occurs in a NP also makes a semantic contribution to that NP. This is not just the case in English: as far as we know, there is no language where adjectives occurring with the subject modify the object, and vice versa.

Further evidence for the close relation of syntactic and semantic constituency comes from so many phenomena that we cannot discuss them all. Briefly consider the case of idioms. On the one hand, an idiom is a semantically opaque unit whose meaning does not derive from the interpretation of its parts in a transparent way. On the other hand, an idiom is syntactically complex. Consider (8).

(8) (a) Paul kicked the bucket. ('Paul died.')
(b) The shit hit the fan. ('Things went wrong.')

The examples in (8) show a verb–object idiom and a subject–verb–object idiom. What about a subject–verb idiom that is then combined transparently with the object? Marantz (1984, pp. 24–28) claims that there are no examples of this type in English. Since syntacticians have argued that the verb and the object form a constituent, the verb phrase, that does not include the subject (the VP in (4)), Marantz's generalization corroborates the claim that idioms are always syntactic constituents, which follows from the close relationship between syntax and semantics.

If syntactic and semantic recursion are as closely related as we claim, an important question is the semantic equivalent of syntactic constituent formation. In other words, what processes can derive the interpretation of a syntactically complex phrase from the interpretation of its parts? Specifically, the most elementary case is that of a constituent that has two parts. This is the question to which the next two sections provide an answer.

2.1 Constituency

2.1.1 Predication as functional application. An old intuition about sentences is that the verb has a special relationship with the subject and the objects. Among the terms that have been used for this phenomenon are 'predication', which we adopt, 'theta marking,' and 'assigns a thematic role.'

(9) John gave Mary *Brothers Karamazov*.

One basic property of predication is a one-to-one relation of potential argument positions of a predicate and actually filled argument positions. For example, the subject position of a predicate can only contain one nominal: (a) shows that two nominals are two many; and (b) shows that none is not enough.

(10) (a) *John Bill gave Mary *Brothers Karamazov*.
(b) *gave Mary *Brothers Karamazov*.

Chomsky (1981) describes this one-to-one requirement between predication position and noun phrases filling this position as the theta-criterion.

The relationship between 'gave' and the three NPs in (9) is also a basic semantic question. The observed one-to-one correspondence has motivated an analysis of verbs as mathematical functions. A mathematical function maps an argument to a result. Crucially, a function must take exactly one argument to yield a result, and therefore the one-to-one property of predication is explained. So, far example, the meaning of 'gave Mary *Brothers K.*' in (10) would be captured as a function that takes one argument and yields either True or False as a result, depending on whether the sentence is true or false.

The phrase 'gave Mary *Brothers K.*' (the verb phrase) in (9) is itself semantically complex, and involves two further predication relations. The meaning of the verb phrase, however, is a function. Therefore the semantic analysis of the verb phrase requires us to adopt higher order functions of the kind explored in mathematical work by Schönfinkel (1924) and Curry (1930). Namely, we make use of functions the result of which is itself a function. Then, a transitive verb like 'greet' is modeled as a function that, after applying to one argument (the object), yields as its result a function that can be combined with the subject. The ditransitive verb 'give' is captured as a function that yields a function of the same type as 'greet' after combining with one argument.

Functional application is one way to interpret a branching constituent. It applies more generally than just in NP–verb combinations. Another case of predi-

cation is that of an adjective and a noun joined by a copula, as in (11). We assume here that the copula is actually semantically empty (i.e., a purely formal element), and the adjective is a function from individuals to truth values, mapping exactly the red-haired individuals on to True (see Rothstein 1983).

(11) Tina is red-haired.

Similarly, (12) can be seen as a case of predication of 'girl' on the noun 'Tina.' For simplicity, we assume that in fact not only the copula but also the indefinite articles in (12) are required only for formal reasons. Then, 'girl' can be interpreted as the function mapping an individual who is a girl on to True, and all other individuals on to False.

(12) Tina is a girl.

2.1.2 Predicate modification as intersection. Consider now example (13). In keeping with what we said about the interpretation of (11) and (12) above, (13) should be understood as a predication with the predicate 'red-haired girl.'

(13) Tina is a red-haired girl.

The predicate 'red-haired girl' would be true of all individuals who are girls and are red-haired. But clearly, this meaning is derived systematically from the meanings of 'red-haired' and 'girl.' This is generally viewed as a second way to interpret a branching constituent: by intersecting two predicates, and essentially goes back to Quine (1960).

2.1.3 Predicate abstraction. A third basic interpretation rule can be motivated by considering relative clauses. The meaning of (14a) is very similar to that of (13). For this reason, relative clauses are usually considered to be interpreted as predicates (Quine 1960, p. 110 f.).

(14) (a) Tina is a girl who has red hair.
(b) Tina is a girl who Tom likes.

Syntacticians have argued that the relative pronouns in (14) are related in some way to an argument position of the verb. We adopt the assumption that this relationship is established by syntactic movement of the relative pronoun to an initial position of the relative clause.

(15) who_x x has red hair
who_x Tom likes x

Looking at the representation in (15), the semantic contribution of the relative pronoun is to make a predicate out of a complete clause which denotes a truth value. An appropriate mathematical model for this process is lambda-abstraction (Church 1941). The definition in (16) captures the intuition that (14b) entails the sentence 'Tom likes Tina,' where the argument of the predicate is inserted in the appropriate position in the relative clause (see, e.g., Cresswell (1973) or Partee et al. (1990) for more precise treatments of lambda abstraction).

(16) $[\lambda_x$ XP] is interpreted as the function that maps an individual with name A to a result r where r is the interpretation of XP after replacing all occurrences of x with A.

Functional application, predicate modification and lambda abstraction are probably the minimal inventory that is needed for the interpretation of all hypothesized syntactic structures.

2.2 Scope

2.2.1 Quantifier scope. A second area where a number of correspondences between syntax and semantics have been found are semantic interactions between quantificational expressions. Consider the examples in (17) (adapted from Rodman (1976, p. 168):

(17) (a) There's a ball that every boy is playing with.
(b) Every boy is playing with a ball.

The difference between (17a) and (17b) is that only (17b) is true when every boy is playing with a different ball, as in picture B; (17a) is only judged to be true in the situation depicted in picture A, where every boy is playing with the same ball.

(18)

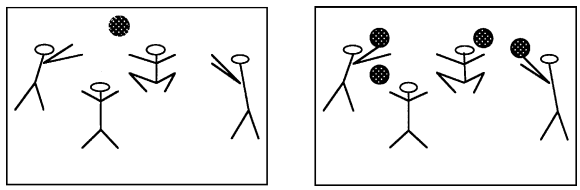

Picture A Picture B

The semantic difference between (17a) and (17b) correlates with a difference in syntactic structure. Namely, in (17b) 'a ball' is part of the sister constituent of 'every boy,' while it is not so in (17a).

The special role the sister constituent plays for semantic interpretation was first made explicit by Reinhart (1976) building on the work of Klima (1964) and others. She introduced the term c-command for the relationship where a phrase enters with any other phrase inside its sister constituent, and proposed that the domain of c-commanded phrases (i.e., the sister constituent) is the domain of semantic rule application. The c-command domain is then equated with what logicians call the scope.

On the recursive interpretation procedure developed in the previous section, Reinhart's generalization is captured in a straightforward manner. This can be seen without being precise about the semantics of quantifiers (see Sect. 3.1, and *Quantifiers, in Linguistics*). Consider the interpretation of the relative clause constituent in (a), which is shown with the lambda operator forming a predicate in (19). This

predicate will only be true of an object *a* if every boy is playing with *a*. Such an individual is found in picture B, but not in picture A.

(19) Every boy is playing with *x*.

For sentence (17b), however, we want to interpret the verb phrase 'is playing with a ball' as a predicate that is true of an individual *a*, if there is some ball that *a* is playing with. (Below, in Sect. 3.2, we sketch a way to derive this meaning of VP from the meanings of its parts.) Crucially, this predicate is fulfilled by every boy in both picture A and picture B.

2.2.2 Binding. A second case where c-command is important is the binding of pronouns by a quantificational antecedent. Consider the examples in (20).

(20) (a) Every girl is riding her bike.
(b) Every girl is riding John's bike.

Sentence (a) is judged true in a situation where the girls are riding different bicycles—specifically, their own bicycles. Example (b), however, is only true if John's bike is carrying all the girls. The relationship between the subject 'every girl' and the pronoun 'her' in (a) can only be made more precise by using the concept of variable binding that has its origin in mathematical logic. (Note that, for example, the relevant interpretation of (a) is not expressible by replacing the pronominal with its antecedent 'every girl,' since this would result in a different meaning.)

We assume that the interpretation of the VP in (20a) can be captured as in (21) following Partee (1975) and others.

(21) λx is riding x's bike.

The relevance of c-command can be seen by comparing (20a) with (22). Though (22) contains the NP 'every girl' and the pronoun 'her,' (22) cannot be true unless all the boys are riding together on a single bike. The interpretation that might be expected in (22) in analogy to (20a) but which is missing can be paraphrased as 'For every girl, the boys that talked with her are riding her bike.'

(22) The boys who talked with every girl are riding her bike.

The relevance of c-command corroborates the recursive interpretation mechanism that ties syntax and semantics together. Because binding involves the interpretation mechanism (16) operating on the sister constituent of the λ-operator, only expressions c-commanded by a λ-operator can be bound by it. In other words, an antecedent can bind a pronoun only if the pronoun is in the scope of its λ-operator.

3. Syntax–Semantics Mismatches

3.1 Subject Quantifiers

When we consider the semantics of quantifiers in more detail, it turns out that the view that predication in syntax and functional application in semantics have a one-to-one correspondence, as expressed above, is too simple. Consider example (23) which is repeated from (1).

(23) Every boy is holding a block.

Assume that the verb phrase 'is holding a block' is a one-place predicate that is true of any individual who is holding a block. Our expectation is then that the interpretation of (23) is achieved by applying this predicate to an individual who represents the interpretation of the subject 'every boy.' But some reflection shows that this is impossible to accomplish—the only worthwhile suggestion to capture the contribution of 'every boy' to sentence meaning by means of one individual, is that 'every boy' is interpreted as the group of all boys. But the examples in (24) show that 'every boy' cannot be interpreted in this way.

(24) (a) Every boy (*together) weighs 50 kilos.
(b) All the boys (together) weigh 50 kilos.

The interpretation of (24) cannot be achieved by applying the predicate that represents the VP meaning to any individual. It can also be seen that predicate modification cannot be used to assign the right interpretation to examples like (24a). The solution that essentially goes back to Frege (1879) and was made explicit by Ajdukiewicz (1935) is that actually the quantificational subject is a higher order function that takes the VP predicate as its argument, rather than the other way round. This interpretation is given in (25):

(25) [every boy] is a function mapping a predicate P to a truth value, namely it maps P to True if P is true of every boy.

Note that this analysis still accounts for the exactly one-argument requirement, which was observed above, because the result of combining the subject quantifier and the VP is a truth value, and therefore cannot be combined with another subject quantifier.

The Frege/Ajdukiewicz account of (23) does, however, depart from the intuition that the verb is predicated of the subject. Note that the semantic difference between quantifiers and nonquantificational NPs does not affect the syntax of English verb–subject relationships. Still in (1) the verb agrees with the subject, and the subject must precede the verb.

3.2 Object Quantifiers

Consider now (26) with a quantificational noun phrase in the object position. We argued in Sect. 2.1.1 that the interpretation of a transitive verb is given by a higher order function that takes an individual as an argument and results in a function that must take another individual as its argument before resulting in a truth value. But then (26) is not interpretable if we assume the Frege/Ajdukiewicz semantics of quantifiers.

(26) John greeted every boy.

From the partial account developed up to now, it follows that 'greet' and 'every boy' should be combined semantically to yield a predicate representing the meaning of the VP. However, the meaning of 'every boy' is a higher order predicate that takes a predicate from individuals to truth values as its argument. But the meaning of 'greet' is not such a one-place predicate.

Any solution to the interpretability problem we know of posits some kind of readjustment process. One successful solution is to assume that before the structure of (26) is interpreted, the object quantifier is moved out of the verbal argument position (see, e.g., Heim and Kratzer 1998). For example the structure shown in (27) is interpretable.

(27) John λx [every boy] λy [x greeted y]

The movement process deriving (27) from (26) is called 'quantifier movement.' Quantifier movement in English is not reflected by the order in which words are pronounced, but this has been claimed to be the case in Hungarian (Kiss 1991).

3.3 Inverse Scope

The original motivation for quantifier movement was actually not the uninterpretability of quantifiers in object position, but the observation that the scope relations among quantifiers are not fixed in all cases. In example (17a) in Sect. 2.2.1, the scopal relation among the quantificational noun phrases was fixed. But consider the following sentence:

(28) A boy held every block.

The example is judged true in the situation in picture A, though there is no single boy such that every block was held by him.

(29)

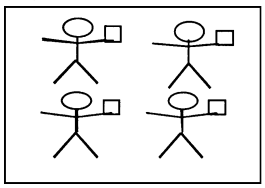

Picture A

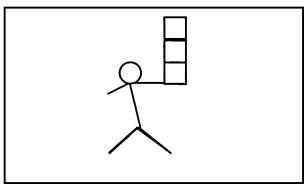

Picture B

The truth of (28) in this situation is not predicted by the representation in (30) that in some way maintains the syntactic constituency of the verb and the object to the exclusion of the subject.

(30) [a boy] λx [every block] λy [x held y]

For this reason, May (1977) and Chomsky (1975) propose to allow also quantifier movement that creates a constituent which includes the subject and the verb but not the object (similar ideas go back to Montague (1970) and Lewis (1972)). This structure is shown in (31).

(31) [every block] λy [a boy] λx [x held y]

Hence, the hypothesis that movement of quantificational noun phrases prior to interpretation is possible (and indeed obligatory in some cases) provides a solution not only for the problem with interpreting object quantifiers, but also for the availability of inverse scope in some examples.

4. Concluding Remarks

Many of the questions reviewed above are still topics of current research. We have based our discussion on the view that syntax and semantics are autonomous, and that there is a mapping from syntactic structures to interpretation. However, recent work by Fox (2000) argues that in some cases properties of interpretation must be visible to the syntax of quantifier movement. As the basic inventory of interpretation principles, we have assumed functional application, predicate modification, and lambda abstraction over variables. Other approaches, however, make use of function composition and more complex mathematical processes to eliminate lambda abstraction (Jacobson and others). In a separate debate, Sauerland (1998) challenges the assumption that movement in relative clauses should be interpreted as involving binding of a plain variable in the base position of movement. While the properties of scope and binding reviewed in Sect. 2.2 are to our knowledge uncontroversial, current work extends this analysis to similar phenomena such as modal verbs, tense morphemes, comparatives and many other topics (Stechow and Wunderlich (1991) give a comprehensive survey).

Research activity on the syntax–semantics interface is currently expanding greatly, as an increasing number of researchers become proficient in the basic assumptions and the formal models of the fields of both syntax and semantics. Almost every new issue of journals such as *Linguistic Inquiry* or *Natural Language Semantics* brings with it some new insight on the questions raised here.

See also: Linguistics: Incorporation; Movement Theory and Constraints in Syntax; Semantics; Syntax; Valency and Argument Structure in Syntax

Bibliography

Ajdukiewicz K 1935 Die syntaktische Konnexität. *Studia Philosophica* **1**: 1–27

Chomsky N 1957 *Syntactic Structures*. Mouton, Den Haag

Chomsky N 1975 Questions of form and interpretation. *Linguistic Analysis* **1**: 75–109

Chomsky N 1981 *Lectures on Government and Binding*. Foris, Dordrecht

Chomsky N 1995 *The Minimalist Program*. MIT Press, Cambridge, MA

Church A 1941 *The Calculi of Lambda-Conversion*. Princeton University Press, Princeton, Vol. 6

Cresswell M J 1973 *Logic and Languages*. Methuen, London

Curry H B 1930 Grundlagen der kombinatorischen Logik. *American Journal of Mathematics* **50**: 509–36
Fox D 2000 *Economy and Semantic Interpretation*. MIT Press, Cambridge, MA
Frege G 1879 *Begriffsschrift. Eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Neubert, Halle
Heim I, Kratzer A 1998 *Semantics in Generative Grammar*. Blackwell, Oxford
Jacobson P 1999 Towards a variable-free semantics. *Linguistics and Philosophy* **22**: 117–84
Kiss K 1991 Logical structure in linguistic structure. In: Huang J, May R (eds.) *Logical Structure and Linguistic Structure*. Kluwer, Dordrecht, The Netherlands, pp. 111–47
Klima E 1964 Negation in English. In: Fodor J A, Katz J J (eds.) *The Structure of Language*. Prentice-Hall, Englewood Cliffs, NJ, pp. 246–323
Lewis D 1972 General semantics. *Synthese* **22**: 18–67
Marantz A 1984 *On the Nature of Grammatical Relations*. MIT Press, Cambridge, MA
May R 1977 The grammar of quantification. Ph.D. dissertation, MIT
Montague R 1970 Universal grammar. *Theoria* **36**: 373–98
Partee B H 1975 Montague grammar and transformational grammar. *Linguistic Inquiry* **6**: 203–300
Partee B H, ter Meulen A, Wall R E 1990 *Mathematical Methods in Linguistics*. Kluwer, Dordrecht
Quine W 1960 *Word and Object: Studies in Communication*. MIT Press, Cambridge, MA
Reinhart T 1976 The syntactic domain of anaphora. Ph.D. dissertation, MIT
Rodman R 1976 Scope phenomena, movement transformations, and relative clauses. In: Partee B H (ed.) *Montague Grammar*. Academic Press, New York
Ross J R 1967 Constraints on variables in syntax. Ph.D. dissertation, MIT
Rothstein S D 1983 The syntactic form of predication. Ph.D. thesis, MIT
Sauerland U 1998 The meaning of chains. Ph.D. thesis, MIT
Schönfinkel M 1924 Über die Bausteine der mathematischen Logik. *Mathematischen Annalen* **92**: 305–16
Stechow A von, Wunderlich D 1991 *Semantik. Ein internationales Handbuch zeitgenössischer Forschung*. Walter de Gruyter, Berlin, New York

U. Sauerland and A. von Stechow

System: Social

A system is a set of differentiated and interdependent components. With such an abstract definition, the notion of system explicitly or implicitly underlies all or most of the theories of the social sciences. It is therefore useful to distinguish three ways of conceptualizing social systems, each one corresponding to a major paradigm. First, the *historicist* paradigm, according to which society constitutes a global system subject to a necessary law of historical evolution; this paradigm was mainly represented by Comte and Marx in the nineteenth century. Second, the *structural-functional* paradigm, according to which a society is a system whose components are the social institutions (the family, religion, etc.); this approach was developed in anthropology by Radcliffe-Brown and Malinowski in the 1930s, and then imported into sociology by Merton and Parsons. Third, the *actionist* paradigm, in which the social systems are constituted by actors and their interactions; one of the main sources of this paradigm is the subjectivist and marginalist revolution that happened in the 1870s in economics.

1. *The Historicist Conception of Social Systems*

According to the historicist conception, society is destined to go through a necessary succession of great historical stages, each of which is a social system. The aim of sociology, then, is to determine the nature of these systems, and especially the historical 'law' governing their succession. Thanks to this historical 'law,' it becomes possible to know the nature of the social system that will appear in the future, the nature of the next great historical stage. What is the use of knowing in advance an inescapable future? Comte and Marx have the same answer: it gives us the power to facilitate and accelerate the advent of the ideal future society, since according to them the future social system will be the ultimate stage in the progress of humanity.

1.1 *Comte and Marx*

In his *positivist* theory, Comte (1824) analyzes the past and future evolution of society as the succession of three great social systems. The *theological* system, first, was represented by the European absolute monarchies: social phenomena were then interpreted in the frame of the Christian religion, the spiritual power was held by the priests and the temporal power by the aristocracy (the sword). Then came the *metaphysical* system ushered in by the French Revolution: social phenomena were interpreted in the frame of the philosophy of the Enlightenment, the spiritual power gave way to freedom of conscience, and the temporal power came into the hands of the people (democracy). According to Comte, this metaphysical system is only transitional. It has a destructive nature and is thus unable to constitute a lasting social order. It will be replaced in the future by the *positive* system, in which sociology will become a 'positive' science (founded, just like modern physics and chemistry, upon observation and invariable natural laws), and in which the spiritual power will be held by the scientists and the temporal power by the producers (the manufacturers). Each social system is thus defined by its 'civilization' (the state of our understanding of social phenomena) and by its 'social organization' (as defined by who holds the spiritual and of the temporal powers). Social organization is here supposed to be determined by the state of civilization, and the latter is determined by scientific progress. Comte believed that this progress obeyed the *law of the three stages*: he thought that

sociology, like the other positive sciences, would appear at the end of a three-stage process ('theological,' 'metaphysical,' and 'positive') through which knowledge is progressively liberated from its animistic and anthropomorphic bias. In this way, the components of each social system and social evolution itself are determined by the progress of knowledge, and this progress is in turn determined by the law of the three stages.

Marx's historicism (1859), or *dialectical materialism*, describes a succession of four 'modes of production' (or social systems), antique, feudal, capitalistic, and socialistic. Each one of these systems is made up of two major components: the 'material productive forces' (techniques, tools, machinery) and the 'production relations' (property laws). The productive forces are supposed to determine the production relations, and also all the other ideological features of society (politics, religion, esthetics, philosophy). Marx illustrated this idea with a famous sentence: 'the hand mill gives you feudal society, the steam mill industrial capitalism' (*Misère de la philosophie* 1847). In a watered-down version of his historicism, Marx (1859, Preface) explains that the succession of these systems is a historical necessity arising from the development of the productive forces (i.e., of the technical capacities of production). Technical and economic progress end up being incompatible with existing production relations. When this contradiction reaches its climax, the production relations are unshackled, and a great social revolution takes place, marking the transition between two modes of production. In *Das Kapital* (1867, Vol. 1), Marx attempts to study in a much more detailed way, with an actionist methodology, the necessary future succession between capitalism and socialism (see Sect. 3 for actionism).

The historicist view was widely held in the nineteenth century. It is to be found in Tocqueville's analysis of the secular transition between aristocratic and democratic society (*De la Démocratie en Amérique*, Vol. 1, 1835). Durkheim, greatly influenced by Comte, attempts to explain in *De la Division du Travail Social* (1893) the necessary transition between the primitive societies, where social cohesion is 'mechanical' (cohesion comes from the similarity between individuals), and modern society, where social cohesion is 'organic' (cohesion arises from the differences between and the complementarity of individuals). In Spencer's historicism (*The Principles of Sociology*, 1891), the historical law of succession between the primitive and modern societies is a law of evolution from a simple to a complex organization, based in part on the biological principles of Darwinism.

1.2 Limits and Errors of Historicism

It is easy today to perceive the limits of historicist theories. From a static point of view, social systems are represented in an oversimplified way, and their internal causal relations are debatable (the alleged primacy of 'civilization' over 'social organization' in Comte, and of 'productive forces' over 'production relations' in Marx). But these theories are especially questionable from a dynamic point of view: regardless of the obvious errors of prediction of a Comte or a Marx, the historicist project may be conclusively refuted with the help of an epistemological argument. Historicist 'laws' are supposed to be the equivalent in the field of social phenomena of physical laws in the field of natural phenomena. But physical laws (for instance, the law of falling bodies) explain *reproducible* phenomena, i.e., phenomena that may be repeatedly observed because they recur a great number of times. Historicist 'laws,' on the other hand, try to explain a *unique* phenomenon, human history. This phenomenon, which happens only once and is definitely not reproducible, cannot be explained by a scientific law in the sense of the natural sciences (Popper 1957). In other words, the claim of historicism to possess a scientific character is completely unfounded. Historicist authors are unable to establish, in a convincing way, the causal relations that are supposed to determine the necessary succession between the great social systems or historical stages (for an analysis of other deep contradictions in the positivism of Comte see Hayek 1952, and in the dialectical materialism of Marx see Mises 1957).

2. The Structural-functional Conception of Social Systems

In the structural-functional paradigm, social institutions are the components of society considered as a system. On the one hand, these institutions play a role in the functioning of society (*functionalism*) and on the other hand are held together by relationships of complementarity that constitute a structure (*structuralism*). The historicist sociologists of the nineteenth century were immersed in a vast process of sociological, economic, and political change. This process seemed to be explainable in terms of a necessary succession of great social systems. The structural-functional paradigm was born in a totally different intellectual context. It originated from the anthropological study of primitive societies, whose long-term history was unknown. As a consequence, this paradigm was completely severed from the most questionable aspect of historicism, namely the search for a global historical 'law' of succession between great social systems.

2.1 Society as a Functional System

As formulated by anthropologist Radcliffe-Brown (1935), the basic idea of functionalism is that in a given society, every institution, every social process, every

cultural form possesses a 'function,' i.e., it plays a definite and even indispensable role in the survival and perpetuation of this system. This reasoning is borrowed from biology, where each organ plays a necessary role in the survival of the organism. Faced with a social institution, whatever it may be, the social scientist is supposed to establish its function from the standpoint of the global system constituted by society. For instance, Radcliffe-Brown explains the existence of ancestor worship in traditional societies by the function of strengthening solidarity: thanks to these rites, individuals learn that they can count on the magical help of their forebears, and also that they must submit to the traditional rules. A narrower conception of functionalism was offered by Malinowski (1944). According to him, the economic, educational and political institutions of a society have the function of satisfying—directly or indirectly—the 'primary' needs related to the biological survival of individuals (food, medical care, reproduction). These primary needs give birth to 'derivative' needs. Satisfying the need for food, for instance, implies a cooperation between the members of society. Since this cooperation requires rules of conduct, these rules themselves become a need, but an 'indirect' one in the sense that it is derived from a primary need.

Functionalism was then imported into the sociology of complex societies by Merton (1949). His *functional analysis* is more qualified than the basic or 'absolute' functionalism developed by the anthropologists of traditional societies. Merton begins by distinguishing the functions that are 'manifest' (i.e., deliberately sought for or designed by the actors taking part in the system) and the functions that are 'latent' (i.e., those that play a role unknown to the participants in the system). The function of the cult of the ancestors according to Radcliffe-Brown is an example of a latent function. Merton then observes that in modern societies some institutions can be functional for a given group, and at the same time dysfunctional for another group (a technical advance, for instance, can increase the average productivity of labor, but also create a redundancy in some branches of production and compel some workers to search for a new job). A modern society, considered as a system, shares none of the 'functional unity' of biological organisms. Furthermore, some institutions or customs may not have any function. Finally, Merton criticizes the idea that there exist a certain number of functions that are necessary to the survival of any society, and that the same specific institutions would be indispensable to perform these functions across all societies.

2.2 The Structure of Social Systems

The conception of society as a functional system of institutions leads quite naturally to an analysis of *social structure*, that is, to an analysis of the complementarity of these institutions. A famous illustration is given by Parsons (1951). He observes that a highly industrialized society requires actors to be socially and geographically mobile. But this mobility is possible only if the actors are not too closely connected to their parents: there must be a clear-cut separation between an actor's 'orientation' family (where he or she was born) and the 'procreation' family (the one created by his or her marriage). If we follow this reasoning to its logical conclusion, industrialization is seen to be incompatible with powerful and extensive traditional family ties. Those two components—industrialization on the one hand and the separation between orientation and procreation families on the other—constitute a consistent social structure typical of modern societies. Conversely, the absence of industrialization and extensive and powerful family ties are the components of another stable social structure, typical of traditional societies. This example shows how the existence of one institution can imply the existence of another institution, their combination representing a stable and coherent social structure.

In anthropology, with Lévi-Strauss (1958), the study of social structure became a *structuralism* based largely upon linguistics. Lévi-Strauss realized that in traditional societies kinship relations (brother/sister), (husband/wife), (uncle/nephew), (father/son) constitute a system characterized by a law of opposition. If, in a given society, the relations between a brother and his sister are generally 'good' (friendly), then (husband/wife) relations are 'bad' (reserved or hostile); in the same way, if (uncle/nephew) relations are 'good,' then (father/son) relations are 'bad,' or vice versa. The system of kinship (brother/sister)(husband/wife)/(uncle/nephew)(father/son) can thus exist in only four structures in the different societies: '+ −/+ −,' '+ −/− +,' '− +/+ −,' or '− +/− +.' The first of these structures, '+ −/+ −,' means that (brother/sister) relations are good, as symbolized by the '+' sign, (husband/wife) relations are bad, '−' sign, (uncle/nephew) relations are good, and (father/son) relations are bad.

This law of opposition, in which relations take precedence over the components they relate, is perfectly similar to the law that had been previously discovered by the linguists N. S. Troubetzkoy and R. Jakobson in the area of structural phonology (these authors showed that language as a communication system rests on a law of opposition between the phonemes).

2.3 The Explanatory Limits of Structural-functional

This paradigm has been criticized mainly for its lack of explanatory power. In principle, the 'absolute' functionalism of a Radcliffe-Brown or a Malinowski permits us to explain the existence of an institution by the role it plays in the working and thus in the

existence of a given society. But the more qualified conception of functionalism to be found in Merton does not enable us to explain the existence of an institution by its latent function, since this institution or function is not considered as a functional requisite of the society anymore. The existence of an institution, then, can only be explained with the help of a 'genetic' analysis consisting of the study of the historical origin of this institution. Two important criticisms have been aimed at structural-functional from an actionist viewpoint by Homans (1987): first, this paradigm takes for granted the actors' obedience to social rules, and does not explain why the actors choose to follow these rules; second, the existence of social structures is also taken for granted, and no explanation is offered as to how these structures emerge from social interaction.

3. *The Actionist Conception of Social Systems*

The actionist paradigm aims at discovering the causes of social phenomena in the *actions* of individuals, and more precisely in the combination of these actions (e.g., see *Action, Theories of Social*; *Methodological Individualism in Sociology*). Social systems are defined as systems of actions or of interactions, from the smallest (two actors) to the largest ones (for instance, the international division of labor). In this paradigm, individual actions are generally considered as rational. This rationality may be utilitarian (appropriateness of means to ends), cognitive (rationality of positive beliefs, i.e., of factual propositions), or axiological (rationality of normative beliefs). The focus of attention, here, is neither on the global evolution of society (historicism), nor on social institutions seen from the standpoint of their functional role and of their mutual relationships (structural-functional), but on human action: social phenomena are considered as emergent and possibly unintentional effects whose causes are actions constituting a system. The actionist paradigm is widely used in economics as well as in sociology. It is quite old since it implicitly underlies, for instance, the works of a classical economist like A. Smith (*An Inquiry into the Nature and Causes of the Wealth of Nations*, 1776). Its importance was reinforced in the 1870s by the marginalist and subjectivist revolution that took place in economics. This revolution consisted of a much more rigorous analysis of human choices and actions in an environment characterized by the scarcity of goods, and made possible the development of really satisfying theories of prices and interest rates.

3.1 *The Systems of Economic Exchange*

The most simple system of exchange is composed of two actors, either of which possesses a good and decides or refuses to exchange it against the other good, according to his or her subjective preferences. For an exchange to take place, the good supplied must have less subjective value than the good demanded (and this is true of each of the two actors). This system can be generalized to a market comprising many sellers and buyers trading one good for another. It may then be generalized to a whole economic system made up of a great numbers of markets. In the tradition originating in L. Walras (*Éléments d'Économie Politique Pure*, 1874), the representation of a global economic system rests on the concept of *general equilibrium*. In a situation of general equilibrium, all the actions of all the actors on all the markets are compatible with one another: they can all be carried out without any actor feeling any regret or disappointment. Thanks to a mathematical formalization of the subjective preferences of the consumers and the possibilities of production of the producers, the existence of a general equilibrium can be logically demonstrated: under some conditions, there exists a set of prices of consumer goods and factors of production such that all the actions of buying and selling are compatible (Arrow and Hahn 1971). This theory is rather complex from a mathematical point of view, and it suffers from two main defects: it does not demonstrate that the system converges towards a general equilibrium, and it does not take into account the radical uncertainty of the future.

In another school of thought, the Austrian school (Mises 1949), the focus is on *market processes*. Economic analysis here focuses on the consequences in the economic system of a change in one of its elements. If, for instance, the consumers change their preferences and decide to spend more than before on good *A* and less on good *B*, the demand for *A* rises and its price tends to rise. The rate of profit also tends to rise in the branch of production of good *A* (since the price of this product rises when its costs of production remain the same). If this rate of profit becomes higher than the current rate of profit, more resources will be invested in the production of good *A*. Two consequences follow: first, the demand for the factors of production of *A* will rise and thus the prices of these factors will tend to rise; and second, the supply of *A* will increase and thus its price will tend to diminish. In this way, the rate of profit in the branch of production of *A* is brought down to the current level (since the price diminishes and the costs rise). Conversely, as the demand for *B* and its price are lowered, the rate of profit falls below the current level, and a disinvestment occurs. Demand for the factors of production of *B* decreases, resulting in lower prices. As a consequence, the supply of *B* tends to drop and its price to rise. As a consequence, the rate of profit in the production of *B* is brought back to the current level (the price rises and the costs diminish). A first adaptation of the economic system to the change of demand is thus achieved: the branch of production of *A* expands and the branch of *B* contracts. But the changes are not restricted to these

two branches of production. To the extent that the demand for the factors of production of *A* or *B* (and for the factors of production of these factors) is also changed, the changes diffuse throughout the system and a new balance between the various branches of industry is established. It is possible to study in the same way the diffusion of the effects of a technical advance, or of a variation in the supply of a natural resource, in the quantity of money, in savings, and so on.

3.2 The Actionist Systems in Sociology

The actionist tradition in sociology was born in Germany at the end of the nineteenth century. Weber (1956), for instance, developed a 'comprehensive' method, which consists of trying to understand actions from the point of view of the actors themselves. This method constitutes a generalization of the analysis of human choices derived from the subjectivist revolution of the 1870s in economics. The German sociological tradition had a major influence on sociology in the twentieth century, for instance through the 'interactionist' approach (e.g., see *Interactionism, Symbolic*). The use of actionist systems enables us to explain in a convincing way a number of important and puzzling phenomena. Boudon (1973) has thus explained 'Anderson's paradox' in the field of sociology of education. From the 1960s on, an important democratization of schooling took place, as the proportion of children of low socioeconomic status increased in the colleges and universities. Since future social status is strongly related to academic achievement, vertical social mobility was expected to increase in developed countries. But, and this is the paradox, it didn't. In spite of a reduction of this inequality, social mobility was not greater than before. Boudon showed that this paradox could be explained due to an actionist system based on the motivations and choices of individuals. Many other examples could be taken from the various fields of sociology (see Boudon 1979 for a survey).

4. The General System Theory

In the 1940s, Bertalanffy (1968) began to develop a 'general system theory,' aiming at transcending the frontiers between a wide range of disciplines—physics, chemistry, biology, psychology, and the social sciences. His basic idea was that the systems analyzed in these different branches of learning share a number of features that can and should be the subject of a science of systems as such. But if it is admitted that the aim of science is to find satisfactory explanations of important phenomena, then the analogies that can be discovered between the systems originating from different branches of science do not matter very much. These analogies can constitute genuine scientific contributions only if they suggest new ways of explaining phenomena, or if they improve the existing explanations. Structural anthropology (Lévi-Strauss 1958) is a good example of a set of theories that originated from an abstract theory of the structure of systems (see Sect. 2 above), but it is an exception. Most of the time, the study of social systems can dispense with a general science of systems (obviously, the opposite is not true).

See also: Action, Theories of Social; Comte, Auguste (1798–1857); Durkheim, Emile (1858–1917); Exchange: Social; Functional Explanation: Philosophical Aspects; Functionalism, History of; Functionalism in Sociology; Interactionism: Symbolic; Marx, Karl (1818–89); Methodological Individualism in Sociology; Organizations, Sociology of; Parsons, Talcott (1902–79); Positivism: Sociological; Rational Choice Theory in Sociology; Rationality in Society; Social Change: Types; Spencer, Herbert (1820–1903); Structuralism, Theories of; Structure: Social

Bibliography

Arrow K J, Hahn F H 1971 *General Competitive Analysis*. North-Holland, Amsterdam

von Bertalanffy L 1968 *General System Theory: Foundation, Development, Applications*. George Braziller, New York

Boudon R 1973 *L'Inégalité des Chances. La Mobilité Sociale dans les Sociétés Industrielles*. Armand Colin, Paris [1974 *Education, Opportunity, and Social Inequality*. Wiley, New York]

Boudon R 1979 *La Logique du Social. Introduction à l'Analyse Sociologique*. Hachette, Paris [1981 *The Logic of Social Action: An Introduction to Sociological Analysis*. Routledge and Kegan Paul, Boston]

Comte A 1824 Plan des travaux nécessaires pour réorganiser la société européenne. In: Arnaud P (ed.) *Du pouvoir spirituel*. Le Livre de Poche, Paris [1978]

von Hayek F A 1952 *The Counter-Revolution of Science: Studies on the Abuse of Reason*. The Free Press, Glencoe, IL

Homans G C 1987 *Certainties and Doubts: Collected Papers, 1962–1985*. Transaction Books, New Brunswick, NJ

Lévi-Strauss C 1958 *Anthropologie Structurale*. Plon, Paris [1963 *Structural Anthropology*. Basic Books, New York]

Malinowski B 1944 *A Scientific Theory of Culture and Other Essays*. The University of North Carolina Press, Chapel Hill, NC

Merton R K 1949/1968 Manifest and latent functions. In: Merton R K (ed.) *Social Theory and Social Structure*, enlarged edn. Free Press, New York

Marx K 1859 *Zur Kritik der politischen Oekonomie*. Duncker, Berlin [1904 *A Contribution to the Critique of Political Economy*. Kerr, Chicago, IL]

von Mises L 1949 *Human Action: A Treatise on Economics*. Yale University Press, New Haven, CT

von Mises L 1957 *Theory and History: An Interpretation of Social and Economic Evolution*. Yale University Press, New Haven, CT

Parsons T 1951 *The Social System*. Free Press, New York

Popper K R 1957 *The Poverty of Historicism*. Routledge and Kegan Paul, London

Radcliffe-Brown A R 1935/1952 On the concept of function in social science. In: Radcliffe-Brown A R (ed.) *Structure and Function in Primitive Society*. Free Press, Glencoe, IL

Weber M 1956 *Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie*, 4th edn. Mohr, Tübingen [1978 *Economy and Society: An Outline of Interpretive Sociology*. University of California Press, Berkeley and Los Angeles, CA]

R. Fillieule

Systems Modeling

The term 'system' has many uses. Within the social sciences, one finds it used in several combinations such as: behavioral system, response system, neural system, social system, structural system, functional system, systems therapy, developmental system, cognitive system, etc. It is difficult to provide a general definition of 'system' which would be clear and distinct, covering the manifold uses of the term. The definition given in Klir (1991) is: a system is comprised of a set of things, together with a set of relations defined between these things. According to Klir, the traditional sciences can be understood as being concerned with things, whereas general systems science is concerned with systemhood as reflected by generalized relationships between things. Klir adds an important qualification—a system is not to be considered as a natural category but as a convenient cognitive construction.

A number of fundamental characteristics of systemhood in the social and biological sciences will be discussed. These characteristics concern general types of relationships, in conformance with Klir's (1991) definition. In contrast, specific relationships can be of any kind, like semantic relationships in cognitive systems, associative relationships in behavioral systems, synaptic relationships in neural systems, etc. From a general perspective, such specific relationships can each be conceived as a mapping in an abstract space, where this space is spanned up by generalized coordinates indexed by location and time. In this abstract setting, location-like relationships involve geometrical patterns and forms, timelike relationships involve dynamic regimes, and relationships involving both location and time pertain to spatiotemporal patterns. In this article, only timelike relationships can be considered in detail. However, much of the discussion generalizes rather straightforwardly to spatial and spatiotemporal aspects of systemhood.

Perhaps the most basic characteristic of systemhood is stability. Stability is a multifaceted characteristic. On the one hand, it pertains to the relationships of a system with its environment, how a system reacts to external inputs. This is the more classical variant of stability in which various types of equilibrium are distinguished, associated with distinct dynamic responses to environmental perturbations. On the other hand, stability characterizes the internal relationships between components of a system, how these relationships accommodate to endogenous changes. This more modern variant of stability concerns the changes in within-system types and number of equilibria under slow (so-called adiabatic) variation of system parameters. In between these classical and modern variants of stability we find equilibrium concepts associated with dynamic game theory.

In what follows the focus will be on a modern conceptualization of stability. That is, a consideration of qualitative changes in the equilibria of systems due to slow continuous variation of system parameters. Modern approaches to stability include much more than this; for instance, the recently discovered new equilibrium types (strange attractors associated with deterministic chaotic dynamic regimes). But attention will be restricted to qualitative changes in equilibria, because this has more definite implications for the social and behavioral sciences. A general theory of sudden qualitative changes, including methodological guidelines and some applications, will be outlined.

Another important characteristic of systemhood, closely related to stability, is homeostasis. The reason why it is considered separately is because in theoretical treatises the ubiquity of homeostatic processes in neurological and physiological systems is often emphasized, yet in actual empirical research it is equally often neglected. It will be shown, to the best of our knowledge for the first time in the social scientific literature, what are the expected consequences of homeostasis for empirical studies of, for instance, the biological bases of personality and behavior.

The final characteristic of systemhood to be considered is difficult to summarize with a single label. The characteristic concerned is related to the well-known slogan: 'The whole is more than its parts.' This slogan could be elaborated as a new logic of part–whole relationships for general systems, but that is not our present endeavor. Yet the possibility of a whole transcending its constituent parts also bears on a number of fundamental issues such as unity, individuality, and other questions concerning the lack of comparability of systems that are composed of the same parts, but may differ at some higher level. In the closing sections it will be shown that taking the individuality of systems seriously leads to entirely new psychometric concepts and to a vindication of the so-called Developmental Systems paradigm.

1. *Statistical Systems Modeling*

To provide a proper setting for the discussion of characteristics of systemhood, a preliminary overview of statistical systems modeling will be given. For this purpose a canonical representation of systems, the so-called state–space model, is introduced. This canonical

state–space model will only be used for the purpose of clarification. There exists an extensive literature on system identification, parameter estimation and hypothesis testing (cf. Elliott et al. 1995) which can be consulted for further details.

The state–space model concerned is defined as:

$$\mathbf{s}_i(t+1) = \mathbf{f}_i[\mathbf{s}_i(t), \boldsymbol{\theta}_i(t), t] + \mathbf{w}_i(t+1) \quad (1a)$$

$$\mathbf{y}_i(t) = \mathbf{h}_i[\mathbf{s}_i(t), \boldsymbol{\theta}_i(t), t] + \mathbf{v}_i(t) \quad (1b)$$

In Eqn. (1a) $\mathbf{s}_i(t)$ denotes the state of system i at time t, a vector-valued (hence bold-faced) latent (i.e., not directly observed) process. The difference equation (1a) expresses the dynamics of the state process of system i: $\mathbf{s}_i(t+1)$ is related to $\mathbf{s}_i(t)$ by some function $\mathbf{f}_i$, where $\mathbf{f}_i$ also may depend on time t as well as on a vector-valued parameter $\boldsymbol{\theta}_i(t)$. The latter parameter $\boldsymbol{\theta}_i(t)$ itself may vary in time. The remaining term $\mathbf{w}_i(t+1)$ in Eqn. (1a) denotes random variation in $\mathbf{s}_i(t+1)$ which cannot be predicted on the basis of $\mathbf{f}_i$. In Eqn. (1b) $\mathbf{y}_i(t)$ denotes the observed (manifest) vector-valued process of system i at time t. The function $\mathbf{h}_i$ expresses the dependence of $\mathbf{y}_i(t)$ on the state $\mathbf{s}_i(t)$, where $\mathbf{h}_i$ also may depend on time t as well as on the parameter $\boldsymbol{\theta}_i(t)$. The remaining term $\mathbf{v}_i(t)$ in Eqn (1b) denotes random measurement error.

The state–space model Eqn. (1) is quite general, although an even more general definition could be given. But for our present purposes Eqn. (1) suffices. Through additional specifications, many well-known types of process models are obtained as special instances of Eqn. (1). These specifications mainly concern the measurement scales of the state process $\mathbf{s}_i(t)$ and manifest process $\mathbf{y}_i(t)$, which each can be metrical or categorical; the forms of $\mathbf{f}_i$ and $\mathbf{h}_i$, which can be linear or nonlinear, and the distributional assumptions characterizing $\mathbf{w}_i(t)$ and $\mathbf{v}_i(t)$. Furthermore, the time index in Eqn. (1) can be defined as real-valued instead of integer-valued. Some well-known process models which can be obtained as special instances of Eqn. (1) are the hidden Markov model, the generalized linear state–space model, and the doubly stochastic Poisson model. Statistical analysis of Eqn. (1), in particular parameter estimation of $\boldsymbol{\theta}_i(t)$, can be accomplished according to a uniform Maximum Likelihood technique involving a combination of recursive filtering and the Expectation-Maximization algorithm (cf. Molenaar 1994).

In what follows we will often consider the linear state–space model, a special case of Eqn. (1). Let $\mathbf{s}_i(t)$ and $\mathbf{y}_i(t)$ be metrical processes, $\mathbf{f}_i$ and $\mathbf{h}_i$ be linear functions, and $\mathbf{w}_i(t)$ and $\mathbf{v}_i(t)$ be Gaussian processes. Denoting the linear functions $\mathbf{f}_i$ and $\mathbf{h}_i$ by, respectively, the (q, q) dimensional matrix $\mathbf{F}_i[\mathbf{s}_i(t), \boldsymbol{\theta}_i(t), t]$ and the (p, q) dimensional matrix $\mathbf{H}_i[\mathbf{s}_i(t), \boldsymbol{\theta}_i(t), t]$ yields:

$$\mathbf{s}_i(t+1) = \mathbf{F}_i[\boldsymbol{\theta}_i(t), t]\mathbf{s}_i(t) + \mathbf{w}_i(t+1);$$
$$\mathbf{y}_i(t) = \mathbf{H}_i[\boldsymbol{\theta}_i(t), t]\mathbf{s}_i(t) + \mathbf{v}_i(t) \quad (2)$$

where $\mathbf{s}_i(t)$ is the q-variate factor or state process of system i at measurement occasion t, $\mathbf{F}_i$ is the matrix of autoregression coefficients relating $\mathbf{s}_i(t+1)$ to $\mathbf{s}_i(t)$, and $\mathbf{H}_i$ is the matrix of loadings at t. Specific version of this linear model underlie both longitudinal factor analysis (Jöreskog 1979) and dynamic factor analysis of time series (Molenaar 1994). Statistical analysis of Eqn. (2) proceeds as for Eqn. (1), although the linear state–space system Eqn. (2) allows for an analysis using standard structural equation modeling techniques.

2. System Stability

Modern concepts of system stability characterize the ways in which the internal relationships between the components of a system react to endogenous parametric changes in $\boldsymbol{\theta}_i(t)$. The analysis of this so-called structural stability constituted a milestone in the nonlinear dynamics paradigm which was initiated by three scientific theories: catastrophe theory, non-equilibrium thermodynamics, and synergism. Despite differences between the three theories, they share a basic focus on the explanation of structural stability and self-organization in nonlinear systems.

Consider a system S that can vary along n dimensions. Take, for instance, a network of n artificial neurons, where the activity of each neuron constitutes a dimension along which the network can vary in time. Then the state of S at each time point is described by an n-dimensional vector $\mathbf{s}_S(t)$ composed of the configuration of activity values along the n dimensions. Moreover the set of all possible values of the state across all time points constitutes the state space. Some states, i.e., some locations in state space, are special in that the system stays there if there are no perturbations. These special states are the equilibrium states of the system. From a geometrical point of view each equilibrium state acts as an attractor for those fluctuating states that are located in its domain of attraction in state space. Hence the set of equilibrium states yields a qualitative characterization of S in that it induces an organization of the state space in the form of a geometrical structure of domains of attraction.

The time-dependent changes of S are described by an n-dimensional evolution equation like Eqn. (1) in which the parameter $\boldsymbol{\theta}_S(t)$ is associated with the strength of the connections between each pair of neurons. It is a defining characteristic of $\boldsymbol{\theta}_S(t)$ that it changes in time at a much slower rate than the rate of change of the state $\mathbf{s}_S(t)$. Moreover the standard assumption is that the slow time-dependent changes of $\boldsymbol{\theta}_S(t)$ are smooth, i.e., without sudden jumps.

The equilibrium states of S are defined as the states for which the rate of change as described by Eqn. (1) is zero. Consequently, the equilibrium states constitute the roots of Eqn. (1). As Eqn. (1) depends upon $\boldsymbol{\theta}_S(t)$, it follows that the equilibrium states of S also depend

upon $\boldsymbol{\theta}_S(t)$. Hence it is evident that the set of equilibrium states will change if $\boldsymbol{\theta}_S(t)$ changes. The important question in assessing structural stability is, in what way these equilibria will change. More specifically, given that $\boldsymbol{\theta}_S(t)$ changes in a slow and smooth way, does this cause the set of equilibrium states to change in a similar slow and smooth way? It turns out that this depends upon the functional form of Eqn. (1): if $\mathbf{f}_S[\mathbf{s}_S(t), \boldsymbol{\theta}_S(t), t]$ is a linear function than smooth variation of $\boldsymbol{\theta}_S(t)$ always causes smooth changes in the set of equilibria of S (i.e., smooth changes in the topological pattern of domains of attraction in its state space). But if $\mathbf{f}_S[\mathbf{s}_S(t), \boldsymbol{\theta}_S(t), t]$ is a nonlinear function then smooth variation of $\boldsymbol{\theta}_S(t)$ can give rise to abrupt qualitative changes in the set of equilibria of S which can be likened to sudden earthquakes that completely reorganize the existing topological pattern of domains of attraction in its state space. That is, existing equilibria may disappear or change in character, and new types of equilibria may suddenly emerge. Such abrupt qualitative changes in the equilibria as a result of smooth parameter changes are called bifurcations (in mathematics), phase transitions (in physics), or stage transitions (in biology and psychology).

It might seem that a bifurcation originates from the smooth variation of the system parameter $\boldsymbol{\theta}_S(t)$. But that can only be so in a restricted sense because smooth parameter variation belongs to a different category than the sudden qualitative changes characterizing a bifurcation. Smooth parameter variation is necessary for a bifurcation to occur, but it is not sufficient. Moreover, the action of the parameter variation in $\boldsymbol{\theta}_S(t)$ is nonspecific in that the process according to which it varies does not determine the kind of qualitative changes of the equilibria of S during a bifurcation. Hence parameter variation is not a formative cause of the ensuing bifurcation. In contrast, it is the dynamical behavior of S, described by $\mathbf{f}_S$, that constitutes the true formative cause of the abrupt qualitative changes in its equilibrium states. The dynamical behavior of S defines a trajectory in the same state space to which also the equilibrium states of S belong. Hence the reorganization of the topographical pattern of domains of attraction in state-space associated with a bifurcation constitutes self-organization, i.e., it is due to the formative action of the dynamics of S itself.

Analysis of structural stability of systems like S shows that nonlinear dynamical systems have the capacity to self-organize: the existing organization of their dynamical behavior transforms into qualitatively new organizations with emergent properties. This is a general mathematical result that in principle applies to any nonlinear dynamical system. In the context of developmental systems it is evident that the mathematical theory of structural stability and self-organization matches the characteristics of epigenetic processes. Epigenesis is the context-dependent unfolding of genetic control of development. In theoretical accounts this context-dependent unfolding of genetic control gives rise to stagewise developmental processes (cf. Molenaar and Raijmakers 2000, for elaboration of this point).

In mathematical analyses of bifurcations it is taken for granted that the details in Eqn. (1) describing a system are known. While this requirement will usually be met in applications in physics, it should be acknowledged that at present such detailed knowledge is lacking for most biological, behavioral and social systems. However, for a class of bifurcations called elementary catastrophes (Gilmore 1981) the analysis can proceed without specification of Eqn. (1). This makes catastrophe theory of unique importance to the analysis of structural stability, self-organization and epigenesis in biology and psychology. Van der Maas and Molenaar (1992) discuss in detail the application of catastrophe theory to the detection and classification of cognitive stage transitions. They present a set of inductive criteria, the so-called catastrophe flags, which provide the necessary methodological tools to establish empirically the presence of genuine stage transitions and to distinguish these from alternative developmental scenarios like sudden changes (spurts) in the level of equilibrium states due, for instance, to changes in environmental influences or the switching on or off of genetic influences. Some applications and elaborations of catastrophe analysis can be found in van der Maas and Molenaar (1992) and Raijmakers et al. (1996).

3. *Homeostasis*

Homeostasis refers to a general principle that safeguards the stability of natural and artificial systems, where stability is understood in its more classical sense of robustness against external perturbations. Homeostasis is a fundamental concept in neuropsychology, psychophysiology and neuroscience (Cannon's thesis). In the behavioral sciences, however, the concept of homeostasis is acknowledged, but it rarely fulfils a prominent role in actual analyses. A noteworthy exception is the monograph by McFarland (1971). Yet the mathematical-statistical theory of homeostasis, in particular optimal control theory of systems with feedback (e.g., Goodwin and Sin 1984), shows that homeostasis has important effects on system behavior and hence should be taken into account in statistical system modeling and analysis. Molenaar (1987) gives an application of optimal control theory to the optimization of a psychotherapeutic process.

A concise illustration is given of some of the effects of homeostasis in applied systems modeling. Homeostasis will be defined as negative feedback control (McFarland 1971). It is shown by means of a simple computer simulation that the presence of homeostasis has profound effects on measured intercorrelations

between the behavior of coupled systems. Although the systems are strongly coupled, the manifest intercorrelations between their behavior approaches zero as a direct function of the strength of homeostasis. This has noteworthy implications for applied behavioral systems analysis. Consider, for instance, the long-standing research tradition investigating the biological basis (a physiological system) of personality (a behavioral system). Both systems are assumed to be coupled, but both have also to be considered as being homeostatic. The homeostatic nature of the coupled physiological and behavioral systems investigated in this research tradition therefore can be expected to yield biased measures of their coupling (i.e., intercorrelations close to zero), whereas the actual strength of this coupling may be considerable.

The illustrative simulation study is based on a simple instance of the linear state–space model in Eqn. (2). Only a single set of two coupled systems is considered, hence the subscript i can be dropped in the defining equations. To simplify matters further, it is assumed that the state $\mathbf{s}(t)$ is observed directly, hence the matrix $\mathbf{H}$ is taken to be the identity matrix and the measurement noise $\mathbf{v}(t)$ is absent. Denoting the unidimensional state process of system 1 (e.g., the behavioral system) by $s_1(t)$ and the unidimensional state process of system 2 (e.g., the physiological system) by $s_2(t)$, this yields:

$$\begin{aligned} s_1(t+1) &= f_{11}s_1(t)+f_{12}s_2(t)+w_1(t+1) \\ s_2(t+1) &= f_{21}s_1(t)+f_{22}s_2(t)+w_2(t+1) \end{aligned} \tag{3a}$$

In Eqn. (3a) $f_{12}s_2(t)$ denotes the influence of the physiological system on the behavioral state process $s_1(t+1)$, where f_{12} is the strength of this coupling; $f_{21}s_1(t)$ denotes the reverse influence of the behavioral system on the physiological state process $s_2(t+1)$.

The system Eqn. (3a) does not include homeostasis. In contrast, the following analogue of Eqn. (3a) includes homeostasis:

$$\begin{aligned} s_1(t+1) &= f_{11}s_1(t)+f_{12}s_2(t)+c_1[s_1(t)]+w_1(t+1) \\ s_2(t+1) &= f_{21}s_1(t)+f_{22}s_2(t)+c_2[s_2(t)]+w_2(t+1) \end{aligned} \tag{3b}$$

In Eqn. (3b) $c_1[s_1(t)]$ and $c_2[s_2(t)]$ denote optimal feedback functions which depend on a number of additional parameters that are not displayed explicitly. Molenaar (1987, and references therein) presents a complete description of the computation of these optimal feedback functions.

To assess the effects of homeostasis on the manifest correlation between $s_1(t)$ and $s_2(t)$, time series are generated according to Eqns. (3a) and (3b). This requires that numerical values are assigned to the system parameters in both Eqns. (3a) and (3b). For instance: $f_{11} = 0.6$, $f_{12} = 0.4$, $f_{21} = 0.4$, and $f_{22} = 0.7$. In addition, $c_1[s_1(t)]$ is taken to be zero (only the physiological system 2 in (3b) includes homeostasis). It is then found that the manifest instantaneous intercorrelation between $s_1(t)$ and $s_2(t)$, as determined for a realization of $T = 100$ time points, is 0.85 for Eqn. (3a) and 0.15 for Eqn. (3b). This shows clearly that the intercorrelation for Eqn. (3b), that is when homeostasis is present in the physiological system 2, is substantially biased towards zero and does not reflect the strength of the coupling between the two systems without homeostasis present.

The same result (substantial underestimation of the actual coupling strength by the manifest intercorrelation between coupled systems in case homeostasis is present) is found under all other possible scenarios (negative coupling between systems without homeostasis, homeostasis present in both systems, etc.). This general result also can be proven by means of mathematical-statistical methods. It is concluded that homeostasis has substantial biasing effects on manifest measures of system coupling, even in simple linear systems such as considered above.

4. *Wholeness and Individuality*

The slogan that the whole is more than its parts implies that distinct systems may each be composed of the same set of components, yet still be different from each other in their global characteristics of systemhood. Another way to state this point of view is that the unity characterizing each system gives rise to emergent characteristics, and thus to individuality.

To investigate the individuality of a given system i, it is required to observe its time-dependent behavior $\mathbf{y}_S(t)$, $t = 1, 2, \ldots, T$, at T consecutive time points (where, to simplify the discussion, consecutive time points are taken to be equidistant). Hence S is fixed and the time interval from $t = 1$ to $t = T$ is conceived of as a sample from the time axis. With regard to Eqn. (1) and Eqn. (2), this implies that $i = S$ is taken to be fixed, whereas statistics and parameter estimates are determined by taking appropriate averages over time points, and consecutively are generalized to the whole time axis (prediction and retrodiction). Henceforth this ($N = 1$ time series) design will be called an analysis of intra-individual variation (IAV).

If a set (population) of systems can be taken to be homogeneous in all relevant aspects, then their time-dependent behavior can be analyzed in the following way: take a sample of N systems from the set and observe the time-dependent behavior $\mathbf{y}_i(t_j)$, $i = 1, \ldots, N$ of this sample at a fixed set of time points t_j, $j = 1, \ldots, J$. Hence, with regard to Eqns. (1) and (2) the J time points are fixed, whereas statistics and parameter estimates are determined by taking appropriate averages over systems, and consecutively are generalized to the whole set or population of systems. Henceforth this (longitudinal) design will be called an analysis of inter-individual variation (IIV).

The particular aspect of wholeness and individuality of systems which will concern us in this section can now be stated more specifically as: under which circumstances and conditions will a longitudinal analysis of IIV yield results that are qualitatively the same as a time series analysis of IAV? This question can be raised with respect to all known techniques of statistical analysis, like analysis of variance, regression analysis, factor analysis, etc. In what follows, only factor analysis is considered in the context of the linear model in Eqn. (2), leading to the particular question: given that $\mathbf{y}$ consists of a fixed set of manifest variables (e.g., psychological test scores), under which circumstances will a longitudinal factor analysis of $\mathbf{y}_i(t_j)$ yield results that are qualitatively the same (same number of latent factors, same pattern of factor loadings, etc.) as a dynamic factor analysis of $\mathbf{y}_S(t)$? For more extensive elaboration of this issue, the reader is referred to Molenaar et al. (2001).

The mathematical-statistical ergodicity theory is of direct relevance to this question. Ergodicity is a basic concept in statistical mechanics and equilibrium thermodynamics (e.g., Cornfield et al. 1982). Consider a closed domain containing a pure gas, where each gas molecule traverses a path or trajectory according to standard dynamical laws. The totality of these time-dependent trajectories of all gas molecules within the domain constitutes an ensemble (comparable with the concept of population in statistics). The dynamics of this ensemble is called ergodic if statistics like the mean velocity determined by following a single molecule in time is (asymptotically) the same as the mean velocity determined by averaging across all molecules at a fixed point in time. Notice that this closely parallels the basic issue under consideration: a process is ergodic if analysis of its IAV yields the same results as an analysis of its IIV. In contrast, if a process is nonergodic then this equivalence no longer holds. A heuristic criterion for ergodicity which applies in many cases is that the mean, variance and covariance associated with an ergodic process are invariant in time. With respect to Eqn. (2) this ergodicity criterion implies that $\mathbf{F}_i[\boldsymbol{\theta}_i(t), t] = \mathbf{F}_i[\boldsymbol{\theta}_i]$ and $\mathbf{H}_i[\boldsymbol{\theta}_i(t), t] = \mathbf{H}_i[\boldsymbol{\theta}_i]$, i.e., these matrices (as well as the parameter vector $\boldsymbol{\theta}_i$) have to be invariant in time.

The ergodicity question is an important issue in theoretical discussions in developmental psychology (cf. Molenaar et al. 2001). If developmental psychology is conceived of as the study of ontogenetic trajectories across the lifespan, then the life history of each individual subject constitutes a path in some high-dimensional state–space, while a population of subjects defines an ensemble of such age-dependent paths. The theoretical debate concerns the question whether the pattern of IIV within such an ensemble of life histories at some point in time contains sufficient information to arrive at valid causal explanations of the pattern of IAV characterizing the time course of each individual life history (cf. Wohlwill 1973).

The ergodicity question also is relevant to the foundations of psychometrics. Standard (longitudinal) factor analysis is the classical method to construct psychological tests (Lord and Novick 1968). Now suppose that it has been concluded in a longitudinal factor analysis that a particular psychological test obeys a 1-factor model at each measurement occasion, that the factor loadings stay invariant across measurement occasions, and that the intercorrelations of factor scores between measurement occasions are high. Hence the structure of interindividual variation of the scores on this test is simple: it stays qualitatively the same across measurement occasions and is characterized by high stability. If this particular test is used in a clinical or counseling setting to predict the scores of a single subject belonging to the population from which the longitudinal sample was drawn, can we be sure that this simple structure also characterizes this subject's IAV? More specifically, can we expect to find in a dynamic factor analysis of this single subject's scores also a 1-dimensional dynamic factor which stays qualitatively invariant in time and has high stability? If the test is measuring a nonergodic process then the answer will be negative.

The consequences of nonergodicity can be observed clearly in classical test theory. Lord and Novick (1968) define the basic concept of true score as the mean of the intraindividual distribution of scores of a fixed subject. Then they remark (Lord and Novick 1968, p. 32): 'The true and error scores defined above are not those primarily considered in test theory. They are, however, those that would be of interest to a theory that deals with individuals rather than with groups (counseling rather than selection).' Next, instead of defining true and error scores for a single fixed person tested an arbitrarily large number of times, true and error scores are defined for an arbitrarily large number of persons tested at one or more fixed times. Hence, instead of focusing on IAV as stipulated by the initial definition of true score, classical test theory is based on IIV. The quote from Lord and Novick given above clearly indicates the restricted nature of results thus obtained in classical test theory.

The standard longitudinal factor model is a special instance of Eqn. (2) in which $\mathbf{F}_i[\boldsymbol{\theta}_i(t), t] = \mathbf{F}[\boldsymbol{\theta}(t), t]$ and $\mathbf{H}_i[\boldsymbol{\theta}_i(t), t] = \mathbf{H}[\boldsymbol{\theta}(t), t]$. That is, these matrices as well as the parameter vector are taken to be invariant across systems (subjects), but may vary in time. Comparison with the heuristic criterion for ergodicity given earlier (i.e., $\mathbf{F}_i[\boldsymbol{\theta}_i(t), t] = \mathbf{F}_i[\boldsymbol{\theta}_i]$ and $\mathbf{H}_i[\boldsymbol{\theta}_i(t), t] = \mathbf{H}_i[\boldsymbol{\theta}_i]$), it is immediately evident that the possible time-dependency of $\mathbf{F}[\boldsymbol{\theta}(t), t]$, $\mathbf{H}[\boldsymbol{\theta}(t), t]$ and $\boldsymbol{\theta}(t)$ in the longitudinal factor model implies the possibility of nonergodicity. Hence in general the structure of IIV as determined in standard longitudinal factor analysis (e.g., test–retest designs) will differ from the structure of IAV as determined in dynamic factor analysis. In particular, psychological tests which have been constructed and normed by means of test–retest factor

analysis will in general yield invalid assessments and predictions in individual counseling. This is a straightforward consequence of classical theorems in ergodicity theory.

It also is evident that the assumptions underlying standard (longitudinal) factor analysis regarding the invariance of the **F** and **H** matrices across systems (subjects) would seem to be a odds with the individuality and wholeness of systems as considered in this section. There is strong theoretical and empirical evidence that in many circumstances systems (subjects) do not obey these strong homogeneity assumptions and instead are heterogeneous, i.e., $\mathbf{F}_i$, $\mathbf{H}_i$ and $\boldsymbol{\theta}_i$ vary across systems i. It can be proven that (longitudinal) factor analysis is in general insensitive to this kind of heterogeneity, despite its incompatibility with the homogeneity assumptions underlying this analysis. This insensitivity has been corroborated in several simulation studies using the following design. An ensemble of individual time-dependent trajectories (life histories) $\mathbf{y}_i(t)$, $t = 1, 2, \ldots, T$, $i = 1, 2, \ldots, N$, is generated according to dynamic factor models with heterogeneous system matrices and parameter vectors varying across systems (subjects). Then (longitudinal) factor analysis is carried out on a sample from this ensemble at one or more fixed time points. The consistent finding then is, that 1-factor (longitudinal) models yield satisfactory fits to these data, while the structure and parameter estimates in the fitted models do not match any of the structures and parameters in the dynamic factor model used to generate the data (cf. Molenaar et al. 2001, for details and references).

5. *Conclusion*

Aspects of systemhood can have profound implications for the modeling and analysis of social, behavioral and biological systems. This has been illustrated by a consideration of three such aspects: structural stability (emergent changes in the geometry of (strange) attractors in state–space), homeostasis (negative feedback inducing underestimation of system coupling), and wholeness (nonergodicity inducing a lack of relationship between analyses of IAV and IIV). Each of these aspects can be underpinned by powerful mathematical-statistical theories and has been studied in extensive simulation studies, showing their manifold effects on system modeling. It is concluded that to accommodate the effects of systemhood in valid ways, advanced time series analysis techniques will prove to be indispensable.

See also: Factor Analysis and Latent Structure: Overview; Graphical Models: Overview; Motivational Development, Systems Theory of; Neural Systems and Behavior: Dynamical Systems Approaches; Self-organizing Dynamical Systems; Spatial Statistical Methods.

Bibliography

Cornfield I P, Fomin S V, Sinai Y G 1982 *Ergodic Theory*. Springer, New York

Elliott R J, Aggoun L, Moore J B 1995 *Hidden Markov Models: Estimation and Control*. Springer, New York

Gilmore R 1981 *Catastrophe Theory for Scientists and Engineers*. Wiley, New York

Goodwin G C, Sin K S 1984 *Adaptive Filtering, Prediction and Control*. Prentice-Hall, Englewood Cliffs, NJ

Jöreskog K G 1979 Statistical estimation of structural models in longitudinal developmental investigations. In: Nesselroade J R, Baltes P B (eds.) *Longitudinal Research in the Study of Behavior and Development*. Academic Press, New York, pp. 303–51

Klir G J 1991 *Facets of System Science*. Plenum, New York

Lord F M, Novick M R 1968 *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA

McFarland D J 1971 *Feedback Mechanisms in Animal Behaviour*. Academic Press, London

Molenaar P C M 1987 Dynamic assessment and adaptive optimisation of the therapeutic process. *Behavioral Assessment* **9**: 389–416

Molenaar P C M 1994 Dynamic latent variable models in developmental psychology. In: von Eye A, Clogg C C (eds.) *Latent Variable Analysis: Applications for Developmental Research*. Sage, Newbury Park, CA, pp. 155–80

Molenaar P C M, Raijmakers M E J 2000 A causal interpretation of Piaget's theory of cognitive development: Reflections on the relationship between epigenesis and nonlinear dynamics. *New Ideas in Psychology* **18**: 41–55

Molenaar P C M, Huizenga H M, Nesselroade J R 2001 The relationship between the structure of inter-individual and intra-individual variability: A theoretical and empirical vindication of developmental systems theory. In: Staudinger U M, Lindenberger U (eds.) *Understanding Human Development*

Raijmakers M E J, van Kooten S, Molenaar P C M 1996 On the validity of simulating stagewise development by means of PDP networks: Application of catastrophe analysis and an experimental test of rule-like network performance. *Cognitive Science* **20**: 101–36

Van der Maas H L J, Molenaar P C M 1992 Stagewise cognitive development: An application of catastrophe theory. *Psychological Review* **99**: 395–417

Wohlwill J F 1973 *The Study of Development*. Academic Press, New York

P. C. M. Molenaar

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

T

Taboo

A concept assumed in the Victorian period to be central to the religions of supposedly less evolved societies, the term taboo today generally is used to refer to prohibitions grounded in custom or religion rather than in bureaucratic law or common sense and hence bearing some moral weight. Taboos can be widespread, observed over the long term, internalized, and deeply felt, as in the case of incest taboos, or applied temporarily with respect to specific persons, acts, conditions, and objects. This article reviews the history of the concept which can be seen to reflect the development of anthropological theory more generally.

1. Taboos, Social Order, and Moral Persons

The word taboo is originally Polynesian. Via the writings of explorers like Cook and missionaries like Ellis it rapidly entered ordinary usage in Victorian England. In 1842 the 7th edition of the *Encyclopaedia Britannica* defined taboo succinctly as, 'A word used by the South Sea islanders, and nearly of the same import as prohibited or interdicted. It applies equally to persons and things, and is also expressive of anything sacred, devoted, or eminent' (quoted in Steiner 1999, p. 164). In the hands of Robertson Smith and Frazer, taboo came to epitomize the essence of primitive thought, distinguishing it from civilized religion (i.e., the Judaeo–Christian tradition) in the way it combined holiness with defilement. In a deconstructive work equivalent to that which Lévi-Strauss performed for totemism, Steiner admirably summarized and critiqued the use of the term through the first half of the twentieth century, demonstrating the problems with applying a culturally particular concept in a comparative and evolutionary fashion. Steiner showed that it was impossible to separate defilement from the sacred in any consistent manner. Steiner's critique of the abstraction and reification of taboo has been pursued by MacIntyre, while his insights into its significance in Polynesia have been considerably enlarged by Gell, Sahlins, and Valeri.

The most profound reflections on taboo have concerned its universal social basis. Both Freud and Lévi-Strauss have seen the incest taboo as fundamental to the constitution of society. For Freud, internalizing prohibitions against intra-familial violence and sexual relations establishes the conditions for the emergence of kinship and moral conscience. Lévi-Strauss argued that the incest taboo generates inter-familial relations, providing the basis for the exchange of women (or marriage partners of either sex) that in effect constitutes society. Whereas Freud in *Totem and Taboo* emphasized the temporal significance of the incest taboo (whose phylogenetic origin in actual historical events is largely dismissed by anthropologists) and the implications for intrapsychic and intrafamilial relations (which finds broader though not unanimous support), Lévi-Strauss argued that the incest taboo forms the logical basis of society and is rooted in the universal mental capacity to make categorical binary distinctions, notably those between self and other, culture and nature. In an ambitious attempt to unify and critique the theory, Rubin added the injunction against homosexual relations.

It is Lévi-Strauss's emphasis on classification that has most influenced subsequent analyses of taboos and has proved fruitful in understanding their varied and often ostensibly trivial content. Drawing on their common heritage of Durkheim, both Douglas and Leach emphasized the role of taboos in protecting systems of social classification by banning anomalous or marginal words, objects, and practices that might violate the order of the categories or blur their distinctions. The title of Douglas's major work, *Purity and Danger* (1966), captures the crux of her argument perfectly. Dirt, or that which is viewed with revulsion and hence tabooed, is 'matter out of place' and hence always relative to a system of classification but also dangerous for it. Building on Steiner, Douglas argued convincingly that individual taboos can only be understood as parts of larger social wholes; that taboos are not false or failed attempts at science; that Western lives are replete with taboos which are no more or less rational than those of other societies; and that without principles of restriction, conceptual and moral discrimination would be compromised and the ensuing freedom of both thought and action would be hopelessly chaotic.

Douglas's argument has inspired much subsequent work, especially on gender, the body, and food. Among the merits of her approach is the ability to show structural similarities in taboos, whatever their specific content, both within a given society, between societies, and across both ostensibly secular (e.g., hygiene, courtesy) and sacred realms. Somewhat more controversial have been attempts both to illustrate how taboos in various orders of phenomena (notably

diet) reflect concerns with the maintenance of social boundaries and divisions and, at a greater level of abstraction, to plot the degree of general interest in maintaining the purity of classificatory orders against the underlying nature of the social relations (the type of society, relatively defined). Following one strain of Durkheim's thought, Douglas gives primacy to the form and integration of social groupings and thus, despite her radical demonstration of the contingent nature of particular taboos, the argument appears rather conservative. Purity and impurity are not only relative to a system of categories, but this system may also be arranged with respect to a hierarchy of sanctity within a holistic universe, as epitomized by Dumont's account of the South Asian caste system. Ensuing questions about the relationship of sanctity and encompassment to power have been much debated.

While the presence of taboos is universal, their substance and direction are not, giving rise to culturally specific social orders. In some parts of the world, like eastern Indonesia, prescriptive rules are emphasized, while in others, like Madagascar, proscriptions are pervasive (Lambek 1992). Lévi-Strauss demonstrated that specific positive rules of marriage—made possible by the underlying incest taboo—generate certain patterns of relationship over time. Taboos, by contrast, in that they do not specify what must or should be done, allow greater degrees of freedom. Yet, as Fortes argued, they are also grasped more easily and internalized more deeply than positive injunctions.

While taboos usually are analyzed as discursive elements, it is important to understand that they become realized in practice. Often the most consequential taboos are the least explicit. If in a structural model the semantic contents of taboos distinguish who or what one is not, from a practice perspective it is the observance of the taboo that substantiates who one is. Van Gennep saw taboos as acts of obligation, creating both continuity and separation. Fortes emphasized the morally binding quality of taboos and the personal significance of observing them for the internalization of an enduring identity and the acknowledgment of dependence on, and obligation to, higher authorities or values. Moreover, taboos 'crystallize these abstract norms in concrete objects and precise rules of conduct which are the more effective because they are of no utilitarian or rational value' (Fortes 1987, p. 129).

Fortes observed both the essentially inward orientation of taboos, especially with respect to eating, and their continuous nature. An important consequence of the fact that taboos are phrased in the negative is that their observance is a relatively continuous and private form of action, whereas carrying out a positive order is a discontinuous and relatively discrete and public act. Hence while the power entailed in observing a taboo may be less explicit and forthright than a direct command, the implications may be more profound. Defiance of taboo may challenge subjection, identity, and value at quite a deep level.

Taboos, then, are not merely imposed, lifted, or carried out in discrete acts but observed continuously. As practical disposition the observance of taboos exemplifies Mauss's and Bourdieu's notion of the habitus and Aristotelian virtue that underlies it. Although not always accepted happily, taboos index personal dignity and (self-)respect, and insofar as they are embodied statements, they may be understood as acts which performatively establish and maintain moral states of social commitment and individual conscience. Moreover, in the visceral reactions to the violation of taboos (nausea, disgust, etc.), underlying social conventions are naturalized. Hence, as Douglas also intuits, taboos do not merely signify social or cosmological order but, in acts of separation, commitment, and principled refusal, they constitute it.

Taboos may be understood phenomenologically with respect to a dialectic of subject and object. Gell argues that it is through taboos, applied at varying levels of specificity across social categories, and operating as intentional, if negative, engagement with the world, that the ego or self is constituted. 'To observe a taboo is to establish an identifiable self by establishing a relationship (negative in this case—*not* eating) with an external reality such that the "self" only comes into existence in and through this relationship' (1979, p. 136). 'Taboos on eating, on killing, on sexual intercourse, on looking at, touching, etc., together circumscribe a "hole" in the texture of shareable intersubjective reality which, privileged with respect to a specific individual, constitutes his soul, or ego, or personality. Taboo does more than express the self: it constitutes the self' (1979, p. 137). When Gell's argument that taboos carve out the self is combined with Lévi-Strauss' argument that taboos generate exchange and hence society, it is clear that the self is relationally constituted, simultaneously held apart from and connected to others, simultaneously social and itself. There is a dynamic quality here: Umeda of Papua New Guinea alternate phases of taboo and consumption; in Gell's analysis the taboo phase restores the integrity of the ego by 'reconstituting it as a closed system' (1979, p. 146). Elsewhere Gell (1995) generalizes the logic of taboo in its original Polynesian locus where it produced differentiation in a cosmos in which sacredness was immanent rather than transcendental.

2. *Transgression*

One of Freud's central arguments concerned the essential ambivalence of taboos, understood as the coalescence of the conflict between a desire and its refusal, attraction, and repression. It is clear that anticipation or suspicion of the violation of taboos often produces visceral reactions—whether laughter, lust, or queasiness—disgust rather than mere indig-

nation. Moreover, because the embodied practice of powerful taboos goes without saying, their pervasive presence is only highlighted by the observance of violation. That is to say, taboos are often indicated through instances of their transgression, in effect, a double negation. As Douglas, Gluckman, and Turner showed, suitably framed transgressions may provide central episodes in religious experience, moral education, and social transition, but otherwise they are perceived as highly threatening. Transgressions may establish political dissent or domination, but they are also the source and substance of neurotic illness, social panic, and much fantasy.

Stallybrass and White (1986) conjoined these ideas with Bakhtin's notion of the carnivalesque and Elias's history of manners to explore the significance of taboos in the historical constitution of the European middle class. The rigorous expulsion of what was perceived as disorderly, low, and carnal has produced, in their view, various 'returns of the repressed.' These include nostalgia, longing, fascination, and even the academic preoccupation with the body.

While some people continue to argue for a biological basis to the incest taboo (and possibly certain other widespread prohibitions), the increased acknowledgment of the prevalence of actual incestuous relations supports those theories that find the source of taboos in culture or society and human ambivalence. Indeed, the question has been raised whether incest taboos are any longer necessary and whether they have been *de facto* lifted, or at least limited. On another front, the violence and rape characteristic of twentieth-century warfare, and the torture carried out by repressive political regimes, must be recognized. Indeed, these might be viewed as deliberate and state-sanctioned liftings of taboos in order to maximize personal violation. However, due in part to fears of contamination that accompany any discussion of transgression, research has focussed more on the condition of victims than the acts of perpetration.

Of course, the presence of taboos has always generated fantasies of transgression. These are often difficult to distinguish from actual transgressions and at the same time may produce transgressive acts of their own, as in the 'de-repression' characteristic of accusations of witchcraft and child abuse prevalent in recent years in North America and northern Europe but also globally (Comaroff 1997, LaFontaine 1998). The need for moral discernment is not always matched by epistemological clarity, yet the fantasies and acts, characteristic of sites and moments of moral vertigo and vacuum, underline the near universality of disgust in the face of certain kinds of physical and sexual brutality and invasiveness, the universal need for avenues for the exercise of dignified moral agency, and the place of taboos among such means.

Conversely, it could be argued that the preceding discussion suffers from an overly socialized, moralistic, and tame conception of taboos; that negation and excess are central to religious experience; and that transgression be recognized as 'the ultimate sacred act' (Taussig 1998, p. 361).

Bibliography

Comaroff J 1997 Consuming passions: Child abuse, fetishism, and the 'New World Order'. *Culture* **17**: 7–19

Douglas M 1966 *Purity and Danger*. Routledge & Kegan Paul, London

Dumont L 1967 *Homo Hierarchicus*. Gallimard, Paris (1970 University of Chicago Press, Chicago)

Fortes M 1987/1966 Totem and taboo. In: Goody J (ed.) *Religion, Morality, and the Person*. Cambridge University Press, Cambridge, UK

Freud S 1958 *Totem and Taboo*. Vol. 13 of The Standard Edition of the Complete Psychological Works of Sigmund Freud (Strachey J, ed. and trans.). Hogarth, London

Gell A 1979 Reflections on a cut finger: Taboo in the Umeda Conception of the Self. In: Hook R H (ed.) *Fantasy and Symbol*. Academic Press, London

Gell A 1995 Closure and multiplication: An essay on Polynesian cosmology and ritual. In: de Coppet D, Iteanu A (eds.) *Cosmos and Society in Oceania*. Berg, Oxford, UK

LaFontaine J S 1998 *Speak of the Devil*. Cambridge University Press, Cambridge, UK

Lambek M 1992 Taboo as cultural practice among Malagasy speakers. *Man* **27**(2): 245–66

Leach E 1964 Anthropological aspects of language: Animal categories and verbal abuse. In: Lenneberg E H (ed.) *New Directions in the Study of Language*. MIT Press, Cambridge, MA

Lévi-Strauss C 1949 *Les structures élémentaires de la parenté*, 1st edn. Presses Universitaires de France, Paris (1966 *The Elementary Structures of Kinship*. Beacon, Boston)

MacIntyre A 1990 *Three Rival Versions of Moral Enquiry*. Duckworth, London

Rubin G 1975 The traffic in women: Notes on the 'Political Economy' of sex. In: Reiter R R (ed.) *Toward an Anthropology of Women*. Monthly Review Press, New York

Stallybrass P, White A 1986 *The Politics and Poetics of Transgression*. Cornell University Press, Ithaca, NY

Steiner F 1999/1956 Taboo. In: Adler J, Fardon R (eds.) *Taboo, Truth, and Religion: Franz Baermann Steiner Selected Writings*, Vol. I. Berghahn, New York

Taussig M 1998 Transgression. In: Taylor M C (ed.) *Critical Terms for Religious Studies*. University of Chicago Press, Chicago

Van Gennep A 1904 *Tabou et totémisme à Madagascar*. Ernest Leroux, Paris

M. Lambek

Tacit Knowledge, Psychology of

Tacit comes from the Latin *tacitus* and translates roughly as *silent*, or *unspoken* and, indeed, the sense of the unverbalized, the unuttered, is still carried by contemporary usage. Usually the Hungarian/British

scholar Michael Polanyi is credited with importing the notion of tacit knowledge into the social sciences and the credit is deservedly his (Polanyi 1962, 1966, Reber 1993a). In Polanyi's framework, tacit knowledge was information that was intensely personal, introspectively opaque, and communicatively unresponsive. It was also, importantly, complex knowledge, abstract in nature, and deeply causal. He wrote of knowing more about familiar conceptual domains than we can tell and characterized his own early life as a creative scientist as being largely a struggle to make tacit knowledge available to consciousness and open to introspection and communication.

Within psychology, this notion of a rich, personal knowledge base with direct causal links to behavior has been around for a long time, surfacing and resurfacing in various forums from the psychoanalytic to the neurocognitive, from the study of perception and learning to models of memory and information storage. The flurry of interest in these issues has spawned a host of synonyms, so much so that Polyani's original term *tacit* is actually rather rare. Preferred lexical entries refer to knowledge that is *implicit* or *unconscious*, occasionally to mental processes that are *nonconscious*, *unaware*, *procedural*, *covert*, and, from time to time, to functions which are *incidental* or *automatic*. However, no matter what cover term is employed, the focus remains on the cluster of perceptual and cognitive processes that function largely independently of awareness of both the processes by which such knowledge was acquired and of the mental representations that result from such learning. In short, *tacit knowledge*, by whatever name enjoys momentary currency, is knowledge that is acquired and used with little in the way of the top-down, modulating control of consciousness. And herein lie the big psychological puzzles. How is complex knowledge acquired largely outside of consciousness? How are memories encoded so that their impact on behavior is manifested by indirect measures while their contents remain largely unavailable to recall and recognition? Is the tacit system merely functionally distinguishable from the more cognitively familiar, consciously modulated systems? Or, as many suspect, are the observed dissociations reflective of deeper neurocognitive systems? Is it possible that explicit, retrievable knowledge systems are linked with biological systems that are evolutionarily distinct from the tacit, implicit systems? Is knowledge that was acquired implicitly superficial knowledge that is tied to the specific form of the stimulus displays, or can such knowledge also be abstract and symbolic?

While work on these questions since the 1970s has touched on virtually every area in modern psychology, it is not unreasonable to distill it down to four classical problems. Not surprisingly, these four are the classical topics that have dominated psychological thought since the nineteenth century: perception, learning, memory, and motivation. The difference is that here an exclusionary clause is invoked, the tacit dimension includes only those processes and mechanisms that operate while remaining more or less impenetrable to top-down, conscious access and control.

As noted above, the cover term *tacit knowledge*, despite its currency in earlier work and its central position in Polanyi's writings, no longer appears in many 'keyword' lists. This is too bad. One of the difficulties with much recent work on the cognitive unconscious is that connotatively loaded terms such as *awareness* and *consciousness* pepper the field and fuel disputes that are often unproductive and unsatisfying. In the following, synonyms will be freely interchanged and attempts will be made to resurrect the core term *tacit*. So, with lexicographic inclusiveness in place, the following are the four primary areas of research.

1. *The Four Primary Areas of Research*

1.1 *Tacit or Implicit Perception*

For the most part, the existence of information pick-up that occurs when the stimulus materials are presented under nonoptimal conditions is a generally accepted phenomenon. The last serious critique of the effect (Holender 1986) was countered by a large number of effective rebuttals and it seems safe to conclude that the acquisition of information about stimulus displays that participants are unaware of perceiving is a robust phenomenon. There is also good evidence that such subliminally perceived representations have emotive elements (Kunst-Wilson and Zajonc 1980). For more detail on this issue, see *Perception without Awareness, Psychology of*.

1.2 *Implicit or Tacit Memory*

Implicitly held memories are, in effect, those memorial residues of experience whose contents affect behavior while remaining largely unavailable for recall or recognition. Such memories show up primarily in indirect tests such as word-stem completion tasks (fill in the blank with the first word that comes to mind: MOT_) where participants reliably complete the task using items previously encountered even though they are unable consciously to recall them. One of the more intriguing findings is that implicit memory is relatively intact in amnesic patients who manifest severely disordered explicit memories. For additional information on this issue, see *Implicit Memory, Cognitive Psychology of*.

1.3 *Implicit or Tacit Learning*

Implicit learning is the process whereby knowledge about complex stimulus domains is acquired largely

without involvement of top-down, conscious control (Reber 1993b). Naturally occurring examples of the operation of implicit learning are language acquisition and the process of socialization. Implicit learning yields implicit or tacit knowledge in that not only is the learner unaware of the process of learning, the very knowledge itself is highly resistant to explication.

This notion that deep, abstract knowledge about a complex stimulus domain can be acquired and represented outside of consciousness is the one that is closest to Polanyi's original use of the term. In Polanyi's framework, tacit, personal knowledge played a critical role in everyday life. Moreover, such communicatively resistant knowledge was also viewed as lying at the very core of creativity. In Polyani's epistemology, to engage in a creative act to struggle to make tacitly acquired and held knowledge conscious and communicable. Doing science, in his mind, was following the urging of one's carefully crafted intuitions and slowly shaping the reservoir of initially unavailable, tacit knowledge out into a communicable form.

1.4 Implicit or Tacit Motivation

The focus here is the exploration of unconsciously held representations. The essential argument, which derives from the classical, Freudian approach, is that unconsciously held knowledge plays its primary role by modifying and adjusting perceptions and memories. This approach goes one step further than the above in that the unconsciously held knowledge in assumed actually to change conscious representations in accordance with deeply held knowledge systems. This use is found primarily in analytically and psychodynamically oriented research where unconsciously held knowledge is presumed to have its impact on behavior through disguised, hidden routes (e.g., Horowitz 1988) although hints of it emerge in more recent work on the development of preferences and aesthetic judgment (Manza and Bornstein 1995).

2. Disputes and Controversies

Perhaps not surprisingly, many of the issues raised in these contexts have proven controversial. It is important to appreciate that most of modern experimental psychology is built on a foundation that goes back to Locke and other British empiricists. This approach, which favored introspection as the primary method of inquiry, equated *mental* with *conscious* and rejected out of hand the notion of a cognitively sophisticated unconscious. Although few appreciate (consciously?) these Lockean roots, many of the arguments being waged over the characteristics of the tacit dimension appear to hinge on the notion that the unconscious, as Polanyi maintained, is just flat-out smart. Specifically, we can consider the following.

2.1 Methodological

It has been argued that the observed implicit–explicit distinction is not an ontological one but merely results from the procedures used to measure participants' awareness of held knowledge (Holender 1986, Shanks and St. John 1994). This argument turns on two points: first, that the measures used to evaluate tacit knowledge are more sensitive than those used to evaluate consciously held knowledge, and second, that no other independent evidence exists to support the distinction. There are good reasons for suspecting that the former is (or was) correct. When more effective methods of inquiry are employed, participants often have shown more awareness of held knowledge than first suspected. Contemporary researchers are much more careful to use more sensitive measures of knowledge held by participants in these studies. The argument, however, appears weaker on the second point. There is a large and growing body of evidence that shows that distinct neuroanatomical substrata are involved in the acquisition and storage of tacit knowledge, specifically structures in the medial temporal lobes. Patients with severe damage in these cortical regions show often devastating episodic memory dysfunctions with little or no disruption in the pick-up and acquisition of tacit knowledge. In addition, there are good reasons for suspecting that the two systems have emerged under distinct evolutionary pressures and have very different evolutionary biological histories (Sherry and Schacter 1987, Reber 1992).

2.2 Functional

The point is often made that there is no 'true' cleavage between tacit knowledge and knowledge held at the edges of consciousness, that the observed dissociations are mere functional adjustments made to particular contexts (Whittlesea and Dorken 1993). The argument here is that participants in standard experimental settings shift from conscious to unconscious modes depending upon the constraints placed upon them and the contexts in which they are required to function. Neither system is the default system and, in fact, they should probably not even be thought of as separate systems but as poles on a continuum running from the deeply tacit to the poignantly conscious. From a functionalists point of view, there is much to recommend in this contention. In all likelihood, most of the interesting things that human beings do on a

routine basis involve a synergistic blend of the conscious and the unconscious, the explicit and the tacit, the overt and the implied. However, recognizing this truth of a complex mind operating in a world of flux misses a basic point. The tacit dimension still needs to be viewed as neurologically and evolutionarily distinct from those processes that are modulated by consciousness. Human consciousness is a late arrival on the evolutionary scene and was preceded by a sophisticated unconscious that functioned without benefit of top-down controlling mechanisms that were open to introspection. They may function interdependently but they are almost surely distinct systems.

2.3 Representational

The dispute here concerns the underlying form of tacit knowledge. Beginning with Polanyi's (strictly nonempirical) work, the presumption was that unconsciously held knowledge had all the rich abstract and symbolic features of knowledge that resided within the spotlight of consciousness. However, many theorists have maintained that the evidence tends to support a more concrete and nonsymbolic representational system (Brooks 1978, Perruchet et al. 1990). Ironically, the problem turns on methodological and functional issues. As has been noted (Manza and Reber 1997), the conditions under which the stimuli are presented seems to be the strongest predictor of the nature of the mental representation. When the inputs consist of large numbers of complex stimuli, participants invariably set up abstract memorial systems which reflect rule-governed qualities. When they consist of simple displays or of small numbers of exemplars, concrete representations abound. Tacit knowledge, like conscious knowledge, most likely can take any of a variety of epistemic forms. The critics of the symbolic are almost certainly correct in that much knowledge held tacitly is concrete and tied to the form of the input stimulus. They are also almost certainly wrong that all tacit knowledge has these features.

2.4 Causal

The dispute here stems from the psychoanalytically motivated claim that tacit knowledge can affect behavior through covert means by disguising the meaning of stimulus displays or adjusting the symbolic content of thoughts, dreams, and fantasies. Of all the purported characteristics of the tacit knowledge system, this one is the most controversial, so much so that it is rarely even mentioned in the cognitive literature (see Erdelyi 1985, 1996 for a singular exception). The problems stem from the assumption of a second level of causal action. Not only must the tacit system be capable of representing complex, symbolically rich knowledge, it must also be capable of making adjustments in its representational form and of disguising its functional meaning. Even theorists who are comfortable with the notion of a cognitively sophisticated and abstractly represented tacit knowledge base resist this extension.

3. Conclusion

The intriguing state of affairs at the start of the twenty-first century is that virtually no one in the social sciences doubts the existence of and the importance of a tacit epistemic dimension, and virtually no two can agree on the nature and mechanisms that underlie it. The future looks interesting.

See also: Conscious and Unconscious Processes in Cognition; Consciousness, Cognitive Psychology of; Implicit Learning and Memory: Psychological and Neural Aspects; Implicit Memory, Cognitive Psychology of; Perception without Awareness, Psychology of

Bibliography

Brooks L R 1978 Nonanalytic concept formation and memory for instances. In: Rosch E, Lloyd B B (eds.) *Cognition and Categorization*. LEA Press, Hillsdale, NJ, pp. 170–211

Erdelyi M H 1985 *Psychoanalysis: Freud's Cognitive Psychology*. Freeman, San Francisco

Erdelyi M H 1996 *The Recovery of Unconscious Memories: Hyperamnesia and Reminiscence*. University of Chicago Press, Chicago

Holender D 1986 Semantic activation without conscious identification in dichotic listening, parafoceal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences* **9**: 1–23

Horowitz M J 1988 *Introduction to Psychodynamics: a New Synthesis*. Basic Books, New York

Kunst-Wilson W R, Zajonc R B 1980 Affective discrimination of stimuli that cannot be recognized. *Science* **207**: 557–8

Manza L, Bornstein R F 1995 Affective discrimination and the implicit learning process. *Consciousness and Cognition* **4**: 399–409

Manza L, Reber A S 1997 Representing artificial grammars: transfer across stimulus forms and modalities. In: Berry D C (ed.) *How Implicit is Implicit Learning?* Oxford University Press, Oxford, UK, pp. 73–106

Perruchet P, Gallego J, Savy I 1990 A critical reappraisal of the evidence for unconscious abstraction of deterministic rules in complex experimental situations. *Cognitive Psychology* **22**: 493–516

Polanyi M 1962 *Personal Knowledge: Toward a Post-critical Philosophy*. University of Chicago Press, Chicago

Polanyi M 1966 *The Tacit Dimension*, 1st edn. Doubleday, Garden City, NJ

Reber A S 1992 The cognitive unconscious: An evolutionary perspective. *Consciousness and Cognition* **1**: 93–133

Reber A S 1993a Personal knowledge and the cognitive unconscious. *Polanyiana* **3**: 97–115

Reber A S 1993b *Implicit Learning and Tacit Knowledge: an Essay on the Cognitive Unconscious*. Oxford University Press, New York

Shanks D L, St. John M F 1994 Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences* **17**: 367–95

Sherry D F, Schacter D L 1987 The evolution of multiple memory systems. *Psychological Review* **94**: 439–54

Whittlesea B W A, Dorken M D 1993 Incidentally, things in general are particularly determined: an episodic-processing account of implicit learning. *Journal of Experimental Psychology: General* **122**: 227–48

A. S. Reber

Tactile Maps in Geography

People gather geographic information mainly through sight. Maps and graphics are important in this process and the visually impaired user needs them in tactual format. Types of tactile maps are presented in this article. The design, production, and use of tactile maps are also discussed, including new computer technologies and current research.

1. *Tactile Maps: the Concept and its Relevance in Geography*

Maps are graphic representations of space, and as abstractions of reality, they belong to the world of images. Most maps are not available for the visually impaired person and they need to be translated to a tactual language using raised symbols. Maps are called tactile maps when they are in a format that lets them be 'seen' by touch.

People acquire and develop the concept of space mainly through the eye. Visual perception is fundamental for understanding the environment. People deprived of vision, either totally or partially, depend on the other senses to gather information on the real world and its geography. Tactile maps have raised symbols on a flat surface. They differ from models, which are three-dimensional, though both are perceptible by touch. Tactual perception works with movement (kinesthetic force), feeling raised images, surfaces, and objects by moving the fingers. This is called haptic perception and it allows visually impaired persons to explore the tangible, world. Besides tactile, other terms such as haptic, tangible, and raised relief maps are currently applied to the same concept.

Tactile maps have been used since Egyptian times where they first appeared as stones set in sand. Since then, many materials have been used, including wood, plaster, and even embroidery. With the invention of Braille in the mid-nineteenth century, it became possible to describe map features using a font that blind people could understand. Embossed images and Braille text have been used for 150 years in maps.

2. *Types of Tactile Maps*

Blind people learn about geographical space through direct exploration, verbal explanations, sounds, smell, gestures, models, and tactile images. Tactile maps are essential to navigate, to develop space perception, and to gather geo-spatial data. There are two main groups of tactile maps, based on their purpose, scale, and features:

2.1 *Orientation and Mobility Maps*

These maps are directed only to blind and severe low-vision people who need to be aware of the physical environment in order to navigate in space. An orientation map gives an overview of the area. Mobility maps, as the name suggests, provide the kind of detailed information that is needed if the person is to move safely through the environment. Mobility maps show obstacles to be avoided, as well as elements to use such as sidewalks, or elevators. They can depict indoor and outdoor environments and both are always large-scale representations of buildings, routes, streets, transportation networks, neighborhoods, and recreation areas. Verbal maps, also called sequential maps, can replace tactile mobility maps to explain routes to blind people.

2.2 *Geographical Maps*

Geographical maps represent larger areas in much smaller scale than mobility maps. They are classified as follows:

General reference maps include political and physical tactile maps, ranging from cities, states, and countries to the whole globe. Tactile topographical maps are relief models, also produced in colored plastic for the sighted user, which need adaptation for the blind by having all symbols in raised format.

Thematic maps portray qualitative and quantitative geographical information in tactile form. They show location, pattern, and structure of geo-spatial data on a wide variety of themes such as population, industry, climate, distribution of vegetation, and at all scales from local and regional to global.

3. *Tactile Map Design, Production, and Use*

Many authors have contributed significantly to tactile cartography. Research on tactile map design, production, and use has been carried out by a large number of specialists (e.g., Barth 1987, Edman 1992, Ishido 1989, James 1982, Kidwell and Greer 1973, Tatham and Dodds 1988, Schiff and Folke 1982, Vasconcellos 1996, Wiedel 1983, and Wiedel and Groves 1972). Guidelines for map design, production, and use can be found in the suggested bibliography, including issues like scale, standardization, degree of generalization, and choice of symbols.

3.1 Tactile Map Design

The design of a tactile map involves a cartographic language with tactile symbols such as raised lines, points, and textures. Visual variables studied by graphic semiology have to be translated into tactual form to construct tactile maps. The nature of geographical data, either quantitative or qualitative, has to be considered when selecting the correct variable to represent the particular data element. Graphic variables (e.g., size, value, texture, color, orientation, and shape) can be expressed in point, linear, and area modes. Color is not part of the tactual graphic language, but elevation (height) can be used alone or combined with other variables in multiple ways, to represent features in addition to height of the terrain.

Clarity and simplicity are important qualities in tactile mapping, as is the use of a map legend or key. A collection of maps or a multi-layered tactile map can help avoid complexity and excessive clutter, making tactual perception easier. Tactile maps when designed for low vision users should combine visual and tactile symbols, preferably in color. Many studies have been done on the legibility and discriminative capability of symbols from the perspective of the visually impaired as map users. In general, choice of symbols, scale, and degree of generalization greatly depend on the production method to be used, including master construction and reproduction techniques.

3.2 Production Methods and Materials

There are many techniques and materials used in the construction and reproduction of tactile maps. The advantages and limitations of these production methods might include, for example, cost, available technology, number of copies needed, or the user's ability to read maps. The most common methods are:

(a) hand-made collage maps using a large variety of inexpensive materials, usually prepared by non-specialists;

(b) molded plastic maps done in vacuum forming machines (e.g., Thermoform with Brailon plastic), and made from a master map prepared in a variety of materials such as embossed aluminum, collage, paper, or plaster carved manually or driven by computers;

(c) embossed paper maps done by Braille printing machines, using conventional or digital methods;

(d) silk-screen maps using puff ink, usually printed on paper;

(e) microcapsule paper maps compatible with conventional printers and raised by heat, largely used with digital maps.

Collage has very good discriminatory properties because of its varied textures, and low precision and low-cost reproduction capability. On the other hand, silk-screen and microcapsule paper maps have low discriminatory properties, but high reproduction capabilities. Besides these more common techniques, there are others such as German film, virktyping, or thermocraft. Each has its advantages and disadvantages depending on a variety of factors including costs.

3.3 Tactile Map Use: the Need for Training

There are some rules to be followed regarding tactile map use. In general, more training is necessary for all producers and users. Before using maps, it is advisable to teach basic geographical concepts such as scale, point of view, relative positioning, orientation, location, and distance. Tactual graphic language has to be introduced to the users before tactile map reading exercises that help them to understand physical space and its representation.

It is important to consider the user's degree of sensory impairment, previous experience, knowledge about the environment, and cartographic skills. Tactual sense is coarser and it works by bringing pieces together to form a map. It is different from sight, which is synthetic with high resolution. Despite some limitations, blind people can have an excellent ability to decode map keys by using both hands simultaneously, one on the legend and the other on the map. In most cases, the visually impaired user needs personal assistance or good instructions to understand the message conveyed in a tactile map. Successful cartographic communication depends on different approaches and efficient training since many of the problems this group may encounter in map reading differ from those identified for sighted users.

4. New Technologies and Future Directions

New technologies have introduced great changes and improvements in the area of tactile map production and use. Tactile maps with Braille text have been produced and used in many countries, but only a minority of blind people read Braille. In the late 1980s, a tactile–audio system called Nomad was developed. This system combined touch and sound so that a tactile map, when touched, spoke the information under the finger. Now, much more information can be placed on a tactile map and Braille is no longer required. Other problems related to map size, low resolution, and shape interpretation are removed.

Since the mid-1980s, there have been refinements of the Nomad system, and in the late 1990s a system called Tactile Graphics Designer or TGD, also by the Australian inventors of the original Nomad system, was developed. This new system allows the blind and visually impaired to draw their own maps and graphics using a combination of software and hardware, printer and embosser (Parkes 1998). Research in Japan, the United States, and Germany has also been focusing on dynamic refreshable surfaces using a matrix of solenoid driven pins, piezoelectric cells, and the possible use of dynamic shape memory to create tactile surfaces.

Thus far, these approaches have proved prohibitively expensive and no commercial products are available. This is an exciting field where computers, telecommunications, geographical position systems, and even phone and fax machines connected to the Internet will work together, facilitating the production of multisensory maps combining sounds, images, and tactile graphics. Both visually impaired and sighted users will appreciate this.

Design and production of tactile maps is a task usually done by other specialists. Concepts such as generalization and standardization can facilitate design and use, but they are not easy to deal with and are still a challenge for geographers and cartographers. Orientation and mobility maps are more likely to have a system of conventions established at the international level. In any case, it is important to prepare users and mapmakers to deal with tactile cartography. There is a need to produce and disseminate tactile graphics, especially maps, not only for the blind but also as teaching materials for all schools.

The International Cartographic Association has supported this initiative and a Commission on Tactile and Low Vision Maps was formed in 1984, aimed at the study and dissemination of tactile cartography. It has members from over 20 countries and has organized four major international conferences in the USA, the UK, Japan, and Brazil, and several regional meetings on all continents.

Tactile cartography helps to overcome informational barriers for those who cannot see, facilitating their access to a world dominated by images. It is very important to design, reproduce, and distribute tactile maps, and to teach how to use them. All kinds of maps and graphics should become available in tactual form, promoting graphic literacy and integration of the visually impaired. Future research should move from tactile cartography to a broader field of special needs cartography with maps reaching more people.

See also: Cartography; Spatial Cognition; Wayfinding and Navigation Behavior

Bibliography

Barth J L 1987 *Tactile Graphics Guidebook*. American Printing House for the Blind, Louisville, KY

Edman P K 1992 *Tactile Graphics*. American Foundation for the Blind, New York

Ishido Y (ed.) 1989 *Proceedings of the Third International Symposium on Maps and Graphics for the Visually Handicapped People*. Yokohama Convention Bureau, Yokohama, Japan

James G A 1982 Mobility maps. In: Schiff W, Foulke E (eds.) *Tactual Perception: A Source Book*. Cambridge University Press, New York

Kidwell A M, Greer P S 1973 *Sites Perception and the Nonvisual Experience: Designing and Manufacturing Mobility Maps*. American Foundation for the Blind, New York

Parkes D 1998 Tactile audio tools for graphicacy and mobility: A circle is either a circle or it is not a circle. *The British Journal of Visual Impairment* **16**(3): 99–104

Tatham A F, Dodds A G (eds.) 1988 *Proceedings of the Second International Symposium on Maps and Graphics for Visually Handicapped People*. University of Nottingham, King's College, London

Vasconcellos R 1996 Tactile mapping design and the visually impaired user. In: Wood C, Keller P (eds.) *Cartographic Design: Theoretical and Practical Perspectives*. John Wiley, UK

Wiedel J W (ed.) 1983 *Proceedings of the First International Symposium on Maps and Graphs for the Visually Handicapped*. Association of American Geographers, Washington, DC

Wiedel J W, Groves P 1972 *Tactual Mapping: Design, Reproduction, Reading and Interpretation*. University of Maryland, College Park, MD

R. Vasconcellos

Talk Show in Media

Talk shows can be defined as televised broadcasts of conversation. Usually a host and several guests discuss a topic in a studio. In TV guides a wide variety of programs are listed as talk shows, ranging from outrageous shows like *Ricki Lake* to serious political discussion programs such as BBC's *Question Time* and from therapeutically inclined programs like the *Oprah Winfrey Show* to the light hearted *David Letterman Show*. Talk shows are a hybrid genre, which contains both journalistic and entertainment elements in different combinations. Looking at talk shows one can trace a number of conversational techniques (interview, debate, informal conversation), varying subject matters (news related issues, entertainment, lifestyle, victims of human tragedy, personal relationships and family matters), and different types of guests (experts, politicians, celebrities, lay people). A broad distinction can be made between the night-time talk shows and the daytime talk shows. The night-time shows generally are entertainment shows or political discussions and broadcast once a week. The daytime talk shows are often broadcast more than once a week, aimed at women and deal with issues from people's private lives.

1. History

The precursors of the television talk show are the listener–participation radio shows of the 1930s and 1940s and the 1960s talk radio with its emphasis on political debate, loud controversy, and shock. On radio the talk show still is a popular format. In the US there are radio stations that only broadcast live talk shows such as NBC's Talknet and ABC's Talkradio Network. Earlier roots can be traced back to public traditions like the nineteenth century literary circles, lyceums, and coffee houses where citizens gathered to discuss matters of common interest. Other influences

come from working class public leisure and spectacular mass entertainment such as the carnival, vaudeville, circus sideshows, and tabloids.

Like many television genres, the talk show format originates in the US. Other countries adopted the genre and adapted the format to their own culture. In many countries commercial channels also broadcast several of the popular American shows. Due to a stronger public broadcasting system, West European talk shows, especially those on public channels, are often less extravagant, slightly more serious, and journalistically inclined than their American counterparts. Cheering audiences are mostly lacking as well as staged confrontations between opponents. In newly democratized countries, where state censorship used to prevent open discussions between ordinary people in the media, talk shows are a new, but often immensely popular phenomenon. Talk shows (and other tabloid media) have in many countries had an influence in reshaping the norms of public discourse in other spheres like politics. In some countries talk shows have taken over prime time and it has even been claimed that their conversational style has become the standard of public discourse (Liebes and Katz 1998).

2. *Host and Participants*

The key to most talk shows is the personality of the hosts. Many hosts acquire celebrity status and often shows are named after their presenters. Hosts can take on several roles, such as stand-up comedian (Steve Allen on *Tonight*! Jack Paar in *The Tonight Show*, David Letterman in the *David Letterman show*), therapist (Oprah Winfrey, Jack Kilroy, Donahue), confrontational interviewer (in some political discussion programs), or promoter of ordinary people's interests. Sometimes their role is a more modest one as the moderator of the discussion. In all cases hosts, to a large extent, control the course of the debate.

Guests on shows are celebrities, actors, authors, and politicians promoting new plays, films and books, or defending a cause. Characteristic is the participation of ordinary people as live, cheering, and applauding audiences but also as the main guests on the panel.

Talk shows often break with traditional roles of experts and lay people. In news programs, politicians, academics, and other public figures are the main spokespeople. In talk shows on the contrary, lay people who disclose personal experiences are the most prominent participants. In some talk shows the contribution of experts is used to substantiate, amplify, or balance the contributions of lay people. In others expert knowledge is confronted with real life experiences of ordinary people. The contribution of the latter is often valued better because their informal performance and straightforward vocabulary is better suited to television codes than the more formal attitude of experts.

Talk shows, especially the more extravagant daytime shows, have been accused of misleading the audience because they are said to stage actors instead of real people. This may be true for some guests on some shows; there are also numerous guests who are not actors. In fact, some claim that the production logic of the shows blurs the boundary between actors and real people, since there is so much coaching of the 'real people' to play themselves or to play the most outrageous version of themselves (Gamson 1998). There has also been criticism on the exploitative character of shows and its potentially harmful effects on the talk shows guests. Guests cannot determine the conditions of their appearance on television and generally cannot foresee its consequences. Their interests do not always coincide with those of staff and producers of the talk shows. Nevertheless, research into people's motivations to appear on talk shows indicates that most participants feel positive about appearing on talks shows and feel that their contribution has been valuable and worthwhile. Promoting a cause, sharing experiences with others in similar situations, or trying to get recognition from relatives, friends, or authorities are mentioned most often as motivations for participating (Priest 1995, Gamson 1998).

3. *Audiences*

The composition of talk show audiences varies with the type of talk show. Late night talk shows hosted by the stand-up comedian type of host are aimed at a predominantly male audience. Subject matter and style of daytime talk shows on the other hand are often aimed at housewives. In the late 1990s a shift has taken place in which daytime shows turned to a younger, racially mixed, and urban audience (*Ricki Lake*, *Jessy Jones*, *Jerry Springer*) (Gamson 1998).

Ratings are rarely very high, but considering that many of the daytime talk shows are broadcast daily, they still attract large numbers of viewers over the week and their market share can be relatively high.

Apart from some quantitative data concerning sociodemographic characteristics of viewers, little is known about the meaning of talk shows for their audiences. There is some evidence however that viewers approve of the genre—even though they are often critical about guests, hosts, or topics discussed. It offers them an opportunity to check their own experiences and knowledge against the statements of the guests on the show (Livingstone and Lunt 1994). For people belonging to groups that are socially marginalized such as handicapped people or sexual minorities it is one of the few television genres in which they are visible. Their experiences are acknowledged, even if they are not always approved of. Often talk show hosts and studio audiences adopt a liberal attitude that at least underlines peoples right to express

themselves. For viewers belonging to a social minority the shows thus offer more than just voyeuristic pleasures.

4. Production

Talk shows are a popular and relatively cheap television format. With the economic pressure of ever expanding numbers of channels and broadcast time to fill, the genre has proved to be a versatile and flexible format that can be adapted to all program slots and target audiences. It has been found particularly suitable to fill less favorable time slots in the mornings, late afternoon, or late night. Talk shows are also often used as a channel to promote books, films, plays, shows, and new releases of CDs and are thus an important marketing outlet for the entertainment industry.

Despite their emphasis on lively, spontaneous, and revealing conversations, most talk shows are carefully orchestrated through pre-interviewing guests and preparing questions for the host. Participants are called upon through screened phone numbers and ads in newspapers. Often the quest for guests involves an active search by editors.

Guests are guided in what to say during the interview, audiences are told when to applaud by a master of ceremony, and camera work is cleverly coordinated. However, debates cannot be completely controlled. Especially the shows that are broadcast live still offer space for genuine audience interruptions and unexpected turns in the debate.

Talk shows are regularly accused of being harmful to society. They have been accused of allowing and even provoking the use of indecent language, physical violence among guests or members of the studio audience, and offering a platform for all sorts of freaks. In their attempt to attract large audiences through staging entertaining and controversial subject matters and guests, they are said to violate moral standards, often leading to confrontations among the participants or to mental and verbal abuse.

5. Theoretical Approaches to Talk Shows

The study of talk shows is relatively new and has expanded with their growing share in program schedules. At the same time these studies reflect more general concerns in mass communication research. In the 1950s psychologists studied the effects of talk shows on the emotional well being of supposedly alienated and isolated members of the audience (Horton and Wohl 1956). In the 1960s and 1970s, talk shows were analyzed from an economic perspective. They were considered to exemplify the commercialization of television and its corruption of morals and the public sphere. In the 1980s and 1990s, in the wake of the growing popularity and academic status of media and cultural studies, more serious attention is being paid to the study of popular culture in general and popular television genres such as the talk show in particular. Currently a number of different approaches can be distinguished. An important strand of research looks at talk shows as an alternative public sphere that provides a platform for citizens to communicate about public affairs among themselves or with representatives of government institutions (Livingstone and Lunt 1994, Liebes and Katz 1998, Munson 1993, Gamson 1998). It is considered 'alternative' because it encompasses social groups that are usually excluded or under-represented in the public sphere. In particular, daytime talk shows in which studio audiences participate in the discussion offer opportunities for the participation of marginalized social groups (women, blacks, 'ordinary people') which are not normally heard in the traditional news genres. They deal with issues and perspectives that are usually excluded from the public sphere, but are nonetheless relevant in many people's lives, especially the lives of less powerful groups in society. Another contribution comes from research into the motivation of guests on talk shows (Priest 1995, Gamson 1998, Hoffmann 1998) and into audience gratification and interpretation (Livingstone and Lunt 1994). An in-depth study of certain characteristics of talk shows is a sociolinguistic analysis of the forms of speech in talk shows (Carbaugh 1988).

Finally the genre is labeled as a postmodern genre *par excellence* because of its extreme intertextuality, its ambiguity and incoherence, its combination of both information and entertainment, personal and public experiences, lay people and experts (Munson 1993). Relatively unexplored is the place of talk shows in the media systems of non-Western countries.

Bibliography

Carbaugh D 1988 *Talking American: Cultural Discourses on Donahue*. Ablex, Norwood, NJ

Carpignano P, Andersen R, Aronowitz S, Difazio W 1990 Chatter in the age of electronic reproduction: talk television and the 'public mind'. *Social Text* **25/26**: 33–55

Gamson J 1998 *Freaks Talk Back. Tabloid Talk Shows and Sexual Nonconformity*. Chicago University Press, Chicago and London

Goldberg C 1998 *Theme und Gästestruktur deutscher Talk-Shows* (*Themes and Guests in German Talk Shows*). Universität München, München

Hoffmann A C 1998 *Öffentlichkeit als Therapie? Zur Motivation von Daytime-Talk-Gästen* [*Public Appearance as Therapy. Motivations of Daytime Talk Show Guests*]. KoPäd Hochschulschriften. KoPäd Verlag, München

Horton R, Wohl R 1956 Mass communication and para-social interaction. Observations on intimacy and distance. *Psychiatry* **19**: 215–29

Liebes T, Katz E 1998 Disaster marathons: A danger to the democratic processes. In: Liebes T, Curran J (eds.) *Media, Ritual and Identity*. Routledge, London

Livingstone S M, Lunt P K 1994 *Talk on Television. Audience Participation and Public Debate*. Routledge, London and New York
Munson W 1993 *All Talk. The Talk Show in Media Culture*. Temple University Press, Philadelphia, PA
Priest P J 1995 *Public Intimacies, Talk Show Participants and Tell-All TV*. Hampton Press Inc. Cresskill, NJ

A. G. D. Leurdijk

Taste and Smell, Psychology of

Taste and smell are separate senses anatomically: taste is mediated by receptors in the oral cavity, smell by receptors high in the nasal cavity; their receptors, too, connect to the brain by different nerves. In perception, however, they function in intimate connection with each other and also with a third sense, the 'common chemical sense' (CCS). The CCS is mediated by receptors in both cavities and in the eyes, throat, and larynx, and gives sensations of irritation and pungency. These three 'chemical senses' are indispensable to feeding, sexual, and predatory behavior in most organisms. In human beings they function mainly to promote the intake of healthy substances and to reject harmful ones.

In everyday perception, sensations aroused via the chemical senses are often blended with one another. Eating and drinking stimulate taste and smell together, but both sensations are localized in the mouth, and in everyday parlance we call the complex 'taste' or flavor. Touch, temperature, pungency (as in pepper), even sound (as in crunchy foods) may also participate in the complex 'taste.' The neural (brain) mechanism underlying this referral and melding of disparate sensations is largely unknown, but the phenomenon is nonetheless central to understanding eating and drinking.

In addition, substances like salt and vinegar that stimulate taste may simultaneously trigger pungency as well as true taste. Likewise 'odors,' e.g., ammonia, may trigger pungency as well as true smell.

To evaluate the separate roles of the chemical senses it is useful to compare their psychological and psychophysical properties in several categories, as follows.

1. *Quality*

The number of taste qualities is small, the classical view being four: sweet, sour, salty, and bitter. Occasionally others are debated, such as the savory taste of monosodium glutamate (called '*umani*' in Japan). A complication here is the potential confusion of true taste with other sensations localized in the mouth as mentioned above.

Qualitative discrimination is a hallmark of smell, but the number of qualities is even harder to settle. One can discriminate among hundreds of thousands of odorous compounds, but the roles of quality, intensity, and CCS are hard to disentangle. Descriptive terms for qualities abound, but none of the various schemes for their classification commands general acceptance. When it comes to CCS, little, if any, quality discrimination is to be discerned.

2. *Hedonic Tone*

Tastes and smells are very often experienced as pleasant or unpleasant, even at moderate intensity levels. This characteristic is not surprising given their roles in monitoring intake of foods and airborne substances. Acceptance of sweet and rejection of bitter appear already in neonates, but odor preferences develop over the first decade. Compelling as they are, preferences are modifiable, dramatically by conditioned aversion (pairing of chemical sensation with nausea) or gradually by repeated experience. Even CCS sensations, initially unpleasant, can become appetizing, as in peppered food concoctions. The contexts in which sensations occur help determine their hedonic tone. For example, sour and bitter alone are unpleasant, but in lemonade or chocolate medium, agreeable.

3. *Intensity*

The strength of tastes, smells, and CCS sensations all increase with increasing stimulus concentration, but in strikingly different ways. The following example illustrates these differences. To 'double' subjectively rated taste intensity requires roughly doubling of the concentration. To 'double' subjective smell intensity requires 3 to 100 times (or more) increase in concentration, depending on the particular compound. (This property of smell is called psychophysical compression, and no other sense exhibits compression more impressively.) To 'double' a CCS sensation (nasal irritation from inhaled CO_2) may require only a 50 percent increase in concentration (psychophysical expansion). Thus, as a function of increasing stimulus concentration, odor sensation grows slowly, CCS sensation fast, and taste sensation roughly proportionally to concentration.

4. *Absolute Sensitivity*

Smell exhibits exquisite sensitivity. That is, just-detectable physical threshold levels are often amazingly low. Some compounds, notably the sulphurous mercaptans, require for detection only about one part per trillion of air. This approximates a theoretical limit of only a few molecules absorbed by a few receptors. Across compounds, however, sensitivity varies by over a trillion-fold. To account for this in terms of chemical structure is a challenge for olfactory science, so far met only for limited families of organic compounds. For the same compound, the detection threshold in CCS (evaluated in persons lacking smell) is higher than in

smell, but everyday intensities often stimulate both senses simultaneously.

Taste sensitivity is less remarkable and, like smell, also varies among compounds. Some bitter compounds are notable for high sensitivity. Genetic differences account for large differences among persons' sensitivity to some bitter compounds, such as those known as PTC and PROP.

5. *Differential Sensitivity*

In contrast to their impressive absolute sensitivity, taste and smell exhibit poor differential sensitivity. By this is meant that to make a just-noticeable increase in supra-threshold strength requires a relatively large percentage increase in concentration: estimates range between 15 percent and 60 percent, as compared with 1 percent for vision and 10 percent for hearing. One investigation of factors limiting olfactory discrimination blames random fluctuations in stimulus concentration rather than the sense of smell itself. Discounting stimulus fluctuation, Cain (1988) concluded that smell rivals hearing in differential sensitivity.

6. *Reaction Time*

Relative to hearing, touch, and vision, whose reaction times to moderate stimulus levels are of the order of one or two tenths of a second, the chemical senses respond sluggishly. Weak odors give reaction times longer than a second, and even strong odors rarely less than half a second. CCS reaction times may be even longer. Reaction times to taste are also slow, of the order of half a second, and appear to be quality-specific, with bitter giving the longest reaction times.

7. *Adaptation*

Most if not all of the senses exhibit adaptation, which means that the detection, magnitude, and quality of a given stimulus depend on prior history of stimulation. The chemical senses are no exception. In general, in a chemical-free world, whether of taste or smell stimuli, sensitivity increases with time, that is to say one can detect progressively weaker stimuli, and supra-threshold stimuli appear progressively more intense. In contrast, as these senses are exposed to taste or smell stimuli, detection suffers and supra-threshold sensations decline. An interesting example is the perception of salt. The tongue is normally adapted to saliva, which contains salt. Bathing the tongue with water removes the salt and thereby increases salt sensitivity (i.e., reduces absolute threshold) by about a hundredfold.

When a taste stimulus lies undisturbed on the tongue the sensation will fade with time and often disappear altogether. Smell sensations, too, fade with prolonged exposure, not altogether, but to a steady state of roughly a third of their initial subjective magnitude.

Both smell and taste can exhibit cross-adaptation. This means that adaptation to a given compound can often affect sensitivity to another compound. Sourness exhibits cross-adaptation with the sourness of all compounds that elicit sourness, consistent with the idea that sourness is mediated by a common mechanism. The same is true of saltiness. In contrast, some sweet compounds cross-adapt, but other do not; and some bitter compounds cross-adapt, but others do not. This is consistent with the belief that sweet and bitter have multiple stimuli (and perhaps multiple receptor mechanisms), unlike sourness and saltiness.

As in taste, some smell compounds cross-adapt with each other, others do not, but the data so far suggest no clear-cut underlying receptor mechanisms. Contrary to naive expectation, a pair of similar-smelling compounds may (or may not) exhibit less cross-adaptation than a pair of dissimilar-smelling compounds. And adaptation by one member of a pair on the other may greatly exceed adaptation in the reverse direction.

From time to time it has been suggested that distilled water can have a definite taste. It is now known that water can take on any quality, depending on the stimulus to which the tongue is previously adapted. For example, after adaptation to salt, distilled water tastes bitter–sour.

8. *Stimulus Mixtures*

A single chemical compound of the taste or smell laboratory serves only rarely as a stimulus in everyday tasting or smelling. In eating, especially, normal stimuli are mixtures, often comprising thousands of compounds. Such complexes are difficult to study, so scientists have usually employed mixtures of a few components. Even so, the picture is complicated and much remains to be learned.

Two important phenomena can emerge simultaneously when two or more chemical compounds are tasted or sniffed in mixture. (a) *Masking*. A compound may mask another, either totally (if the masker is sufficiently strong) so as to eliminate any perceptual trace of the second, or partially so as to reduce but not eliminate the perceptual intensity of the second. Masking may be reciprocal so that the second may also mask the first. (b) *Integration*. A compound may add its effect to the effect of another compound so that the mixture of the two is more detectable than either alone, or have somewhat greater overall strength than either alone.

A mixture of taste compounds may be detectable even though all components are below their thresholds measured separately. Contrary to earlier belief, such integration occurs whether the constituents are of like or unlike quality. One can detect taste mixtures even when their components are qualitatively unrecogniza-

ble. Supra-threshold mixtures behave differently. The subjective intensity of a complex is usually less than predicted by simple addition of the intensities of the constituents assessed separately. Nevertheless, much remains to be learned about the rules of integration and masking in such mixtures.

As for taste, the rules governing olfactory mixtures are only partly known. Integration appears to characterize detection of smell in much the way that it does taste, with generous integration across many constituents. Less is known about supra-threshold mixtures but, similar to taste, mixtures smell weaker than the sum of their unmixed components. Also as in taste, smells are able to mask one another, partly or wholly, depending on their relative strengths.

Integration across compounds in the CCS has been shown to be generous. This finding has important implications for understanding air pollution. A mixture of many compounds that unmixed are all below the threshold of perceived irritation (of the nose, throat, eyes) can nevertheless cause significant irritation. Thus, mixtures may constitute a nuisance or danger, even when each constituent by itself is deemed acceptable.

9. *Aging*

All of the senses undergo changes throughout the life span, the chemical senses included. By middle to older age the ability to detect smells and tastes appears to be slightly to moderately impaired in nearly everybody, and more markedly in smell than in taste. Supra-threshold intensity losses are clearly more pronounced in smell than in taste.

In evaluating aging of the sense of smell one must emphasize the cognitive richness of this sense. Of special interest is the ability to identify and remember, over long stretches of time, a large number of different odors. It is not surprising, therefore, that aging takes a toll on odor identification and memory, independent of its effect on sensitivity and magnitude.

See also: Fechnerian Psychophysics; Olfactory System; Psychophysical Theory and Laws, History of; Psychophysics; Taste, Neural Basis of

Bibliography

Bartoshuk L M 1988 Taste. In: Atkinson R C, Herrnstein R J, Lindzey G, Luce R D (eds.) *Stevens' Handbook of Experimental Psychology*, 2nd edn. Wiley, New York, pp. 461–99

Cain W S 1988 Olfaction. In: Atkinson R C, Herrnstein R J, Lindzey G, Luce R D (eds.) *Stevens' Handbook of Experimental Psychology*, 2nd edn. Wiley, New York, pp. 409–59

Getchel T V, Doty R L, Bartoshuk L M, Snow J B (eds.) 1991 *Smell and Taste in Health and Disease*. Raven Press, New York

Green B G, Mason J R, Kare M R 1990 *Chemical Senses, Vol. 2: Irritation*. Marcel Dekker, New York

Murphy C (ed.) 1998 Olfaction and taste XII. *Annals of the New York Academy of Sciences* **855**:

Stevens J C, Marks L E, Lakatos S, Cruz L A 1998 Aging of the senses: A multimodal study of thresholds. *Journals of Gerontology: Psychological Sciences* **53B**: P263–72

J. C. Stevens

Taste and Taste Culture

Taste is the ability to make discriminating judgments about aesthetic and artistic matters. It is one component of symbolic systems of classification whose content and structure both reflect and shape particular states of social relations. Taste is part of the process by which social actors construct meaning about their social world, classifying people, practices, and things into categories of unequal value. It is displayed in conversation, habits, manners, and in the possession of goods, which signal co-membership into communities of wealth or knowledge. Taste serves as an identity and status marker, being used simultaneously as 'fences or bridges' (Douglas and Isherwood 1979, p. 12) in processes of exclusion and inclusion. Displays of taste contribute to the creation of networks and shared identities within groups, but they also allow for the identification and exclusion of outsiders whose standards of taste differ and who do not belong. Taste cultures are clusters of cultural forms which embody similar values and aesthetic standards.

1. *Taste and the Emergence of the Modern Self*

In his 1757 *Essai sur le Goût* [*An Essay on Taste*] written for the *Encyclopédie* published under the direction of Diderot, Montesquieu defined taste as 'the faculty of discovering with quickness and delicacy the degree of pleasure which we should receive from each object that comes within the sphere of our perception' (Montesquieu 1970, p. 260). Montesquieu's definition of taste has a double meaning. Taste refers both to the feelings of pleasure one experiences when confronted with beautiful objects and to the intrinsic standards of beauty embodied in these objects. Taste is an individual and subjective sentiment, but it is also a discriminating faculty, through which individuals *discover* the amount of pleasure that things *ought* to give them by virtue of their objective properties. According to Ferry (1990), the eighteenth-century invention of taste as a subjective faculty mediating the perception of objective beauty represents a radical break with the past, since aesthetic standards are no longer seen as immediately given by nature, tradition, or other principles transcending human experience. It is emblematic of the emergence of the modern self. The individual's ability to feel and to think rationally became the foundation of aesthetic experience at the

same time that it became the basis of political legitimacy, economic life, and scientific truth.

Attempts to reconcile the subjectivity of aesthetic experience with objective standards of taste remained a central problem of modern aesthetics through the eighteenth and late nineteenth centuries. The classical Cartesian view maintained that art imitates nature according to principles which are discovered through reason, much in the same way that science uncovers the eternal laws of nature. In an argument which resonated well with the Romantic spirit, others argued that beauty does not rest on objective properties of nature, but in immutable and eternal characteristics of the human soul. The mid-eighteenth century controversy on the sources of beauty in music, for example, opposed composer Jean-Philippe Rameau, who argued that music is beautiful because its rules of harmony and melody reflect mathematical properties of sound, to philosopher Jean-Jacques Rousseau, who argued that music is beautiful because sounds mirror humanity's primitive passions (see Ollivier 1987). Moving away from both of these positions, Immanuel Kant's groundbreaking *Kritik der Urteilskraftwork* [*Critique of Judgement*], published in 1790, maintained that taste is an autonomous realm independent from external influences, be they rationality or passions, as well as from considerations of utility, morality, or economy. This position paved the way for theories of art for art's sake, according to which art reflects the genius of the artist and should be allowed to develop unfettered by social and economic constraints. Throughout the twentieth century, art increasingly came to be seen as an extension of subjective experience, reflecting the artist's personal interpretation of the outside world rather than mirroring some preexisting and eternal cosmological order. Aesthetic experience thus lost its objective foundations, and the search for universal and immutable standards of taste appeared increasingly illusory.

2. *Taste in the Social Sciences*

The abandonment of the search for objective standards of taste in aesthetic theory is contemporary with the rise of the social sciences, whose purpose is not so much to understand the universality of aesthetic experience, but rather to analyze how tastes are shaped by ever-changing social conditions. Social science perspectives on taste thus stand in sharp contrast both to theories of taste as a direct reflection of universal principles and to neo-Kantian theories of aesthetic experience as entirely detached from external constraints. Rather than speaking of the discriminating faculty of taste, in the singular, social scientists nowadays also refer to tastes, in the plural form, acknowledging the changing, diverse, and often incommensurable standards of taste that co-exist in complex stratified and multicultural societies.

2.1 *Taste and Pecuniary Emulation*

One of the first modern analyses of tastes in the social sciences is Thorstein Veblen's study of conspicuous consumption (Veblen 1939). Veblen argued that invidious distinction and emulation are driving motives of social interaction. The accumulation of wealth is driven not only by a need to ensure basic subsistence, but primarily by the honor attached to its possession. In order to confer honor, Veblen argued, wealth must be displayed either through conspicuous leisure (abstention from work or from tasks considered demeaning) or through conspicuous consumption of costly goods. Standards of taste do not reflect autonomous and eternal standards of beauty, but a 'sense of costliness masquerading under the name of beauty' (Veblen 1939, p. 128). It is the honor attached to the appropriation and consumption of costly objects which determines our standards of taste, even when honorific considerations are not present in the consciousness of the valuer.

2.2 *Taste and Class Domination*

The most influential treatment of taste in the social sciences over the last decades has been Bourdieu's (1984) theory of distinction. Like Veblen, Bourdieu argues that taste is shaped by status competition: 'good taste' is a mark of distinction, in the double sense of setting apart from, and conferring honor to, those who claim to possess it. 'Taste classifies, and it classifies the classifier' (Bourdieu 1984, p. 6). As opposed to Veblen, however, Bourdieu argues that taste confers honor not because it is a signal of economic wealth, but rather as an expression of wealth which is cultural in nature. Taste is an unequally distributed capacity for appropriating, both symbolically and materially, classes of objects and practices. Along with attitudes, preferences, manners, know-how, and educational credentials, it is a component of cultural capital, which is transmitted by a complex process of socialization through the family and the education system. As part of class habitus, it is one element of a set of dispositions and preferences which are perceived as freely-embraced by social actors but which thoroughly reflect their objective class position.

In spite of appearances, taste is thus a cultivated rather than an innate disposition. In a society divided into antagonistic classes and class fractions, class membership determines both the position of actors in social space and the perceptual framework through which they make sense of their world. Since members of the dominant class experience few material constraints in everyday life, their aesthetic outlook reflects an attitude of detachment, an appreciation for abstract ideas, and a concern for formal properties of objects. Distance from economic necessity imposes a style of life and a legitimate way of perceiving the world. By virtue of its closeness to material constraints, popular

culture is oriented towards the practical and functional. It is defined negatively, as the flip side of legitimate taste.

Because it is misperceived as spontaneous and disinterested, taste functions as an effective instrument of class domination and reproduction. The definition of 'good taste' is part of struggles for the monopolization of symbolic violence, which arbitrarily imposes as natural and legitimate the evaluative standards and perceptual categories of the dominant class. Mastering the complex and subtle nuances of good taste requires a long process of familiarization. Since this process is carried out primarily within the family and in elite schools, good taste only comes 'naturally' to upper-class children who have long been exposed to it. In everyday interaction, displays of a taste for 'difficult' objects signal membership in the privileged class while a taste for common, vulgar, or less refined ones betrays membership in the dominated class.

While Bourdieu's theory of taste as a key mechanism of class reproduction has had a major impact in the social sciences, it has also come under intense criticism. Most importantly, Bourdieu has been criticized for postulating the existence of a single, objective space of distinction, in which cultural capital operates as a universally recognized currency of exchange. Although Bourdieu recognizes that cultural distinctions are expressed differently in various fields, he maintains that a unique organizing principle ultimately accounts for these variations. Bourdieu postulates a strict homology between class position and class dispositions. He subsumes under the concept of class the entire spectrum of status distinctions based, for example, on race, ethnicity, gender, language, and age (Hall 1992). As a result, he pays scant attention to non-class-based struggles for distinction in local status markets, where upper-class tastes may be entirely irrelevant. Similarly, Bourdieu has been criticized for treating popular culture as a negative image of dominant culture rather than as an autonomous space of symbolic production, with its own logic and characteristics (Cingolani 1984). Bourdieu's model is thus seen as giving analytical primacy to processes of distinction which most closely reflect the experience and world view of the French upper classes.

2.3 Taste and Symbolic Classification

An interesting extension of Bourdieu's model, designed for understanding how and why tastes vary over time and place, was proposed by DiMaggio (1987). Tastes, he argues, are one component of symbolic systems of classification. These systems are shaped by structural features of societies but they vary along four distinct dimensions: differentiation (the extent to which tastes are separated into few or several taste cultures); hierarchy (the extent to which tastes are arranged hierarchically as opposed to being considered of equal value); universality (the degree to which specific standards of taste are universally recognized and accepted as legitimate); and symbolic potency (the relative strength and permeability of boundaries between tastes).

The degree to which tastes are segmented, are arranged hierarchically as opposed to being considered of equal value, are universally recognized as legitimate rather than being contested, and have strongly bounded rather than permeable boundaries—these features of classification systems vary over time and place according to changing social and historical conditions. If tastes are homologous with certain characteristics of social structure, it is because they are shaped by a complex interplay of social forces, including the extent of social differentiation and inequality in a community, the cohesiveness and capacity for mobilization of status groups and classes, state policy towards the arts and popular culture, the degree of centralization of cultural authority, and the types of organizational structures through which art and culture are produced and distributed. For example, DiMaggio (1982) has shown that the emergence of a sharp differentiation between high and popular culture in nineteenth-century Boston was the result of class action on the part of urban elites, whose status claims were reinforced by exclusive access to high culture. The differentiation and stratification of tastes was sustained through specific organizational structures, including upper-class patronage for the arts and a system of upper education which transmitted canons of taste to upper-class youth. Similarly, Lamont and Lareau (1988) argue that Bourdieu's conception of tastes as highly hierarchized, differentiated, and universally recognized cultural capital more accurately represents France than North America, where social and geographic mobility, ethnic and racial diversity, and political and cultural decentralization account for more fluid and less stable boundaries between tastes.

3. Taste Cultures: Highbrow, Lowbrow, Nobrow

One of the most ubiquitous and ritually potent classification of tastes in modern times is the distinction between high culture, based on formal aesthetic standards and displayed in museums, theaters and symphony halls, and low or popular culture, in its commercialized and folk versions. The ritual strength and universality of the high/low boundary, however, has varied over time, in different societies, and across cultural genres. In the USA, several observers have noted that after a period of relatively low differentiation and hierarchization through much of the nineteenth century, tastes became more sharply polarized at the turn of the twentieth century. In a process whose pace varied in different geographical locations and across cultural genres, high and low culture

became more strongly differentiated and the boundary between them gained in ritual potency (DiMaggio 1982, Kammen 1999).

The debate about mass culture, which culminated in the 1950s, may be understood as one manifestation of the polarization between high and low culture. Critiques of mass culture reflected anxieties about the coming of mass society, in which traditional hierarchies were seen as rapidly tumbling down. Mass society was perceived as an aggregate of atomized and depoliticized individuals, whose passive consumption of standardized cultural goods made them easy preys for totalitarian ideologies. The commercialized forms of popular culture, produced for mass consumption by profit-driven cultural industries, were seen as aesthetically worthless and socially dangerous. Radical (Horkheimer 1986) and more conservative (MacDonald 1964) critics of mass culture cast a pessimistic look at social relations, arguing that declining standards of taste were evidence of a degradation in individual autonomy and independent judgement, capacities seen as essential for participation in an egalitarian and democratic society.

It is in this context that Herbert Gans (1974) developed the concept of taste cultures. In explicit contrast to much of the literature on mass culture, Gans wishes to go beyond any sharp distinction between high and low culture. He identifies several taste levels within high and low culture and encompasses them all within a single theoretical framework. As stated above, taste cultures are clusters of cultural forms (e.g., in art, entertainment, architecture, consumer goods, etc.) which embody similar values and aesthetic standards. The concept of taste cultures applies as much to classical symphonies performed by the best orchestras as to garish landscapes printed on black velvet. Gans thus rejects the negative connotations associated with low, mass, and popular culture. He advocates what he calls 'cultural pluralism,' arguing that all cultures should be considered of equal value and be placed on an equal footing, at least from the point of view of social analysis and public policy. Taste cultures correspond to a diversity of taste publics, defined as unorganized aggregates of people sharing similar aesthetic standards. Gans acknowledges that the conditions for appreciating works of art associated with different taste cultures are unequally distributed across taste publics, but he does not examine how these inequalities both reflect and reinforce social processes of domination and exclusion.

4. *Postmodern Tastes*

The degree to which discrete and homogeneous taste cultures actually exist as empirical categories has long been a topic of debate among social scientists. In spite of disagreements, a strong consensus has emerged that there has been, in the last decades of the twentieth century, an erosion of boundaries between taste cultures. Evidence of a convergence between high and popular culture is found in all aspects of cultural production and consumption, from content and aesthetic standards to organizational structures and publics. With regard to taste publics, surveys of cultural preferences do not support the picture of societies stratified into sharply differentiated and class-based taste levels. Rather, surveys show that high socio-economic status respondents report preferences for, and participation in, the broadest range of cultural genres and activities, while low-status respondents display more restricted tastes and practices (Peterson and Kern 1996).

The erosion of high/low boundaries parallels a series of social transformations which have increased the breadth and anonymity of social relations: high levels of geographic mobility, made possible by the development of means of transportation and communication; the declining significance of physical place and the concurrent extension of loose networks associated with the new communication technologies; high levels of intragenerational and horizontal occupational mobility under the joint effect of technological change and economic globalization; greater accessibility to higher education beyond the restricted circles of traditional elites; and the extraordinary increase in the amount and diversity of information in circulation, as print and electronic media bombard us with images which are used in strategies of self-presentation in everyday life.

Greater breadth and anonymity, in turn, affect processes of identity formation and status competition, as more people come into more frequent and fleeting contact with others in a wide diversity of social situations. In relatively small and stable communities, status and identity are linked to personal knowledge of fixed categories such as family name, reputation, kinship, and place of residence. In enlarged and anonymous social spaces, individuals typically participate in a multiplicity of groupings which do not necessarily coincide. Combined with the availability of a wide range of styles conveyed by mass communications, this expands the possibilities of experimenting with self-construction in various social circles. Elites no longer form cohesive and tightly bounded status groups but more diffuse networks of well-educated and geographically mobile professionals and upper managers. In this context, the ability to manipulate a diversity of cultural symbols becomes a crucial resource in social interaction. Knowledge of the cultural codes appropriate in various social milieus and the ability to culture-switch according to circumstances become more useful than knowledge of a restricted range of high culture symbols.

While the case for a convergence between high and popular culture is well documented, the meaning of this convergence is open to conflicting interpretations.

At one extreme, the erosion of cultural boundaries and the proliferation of aesthetic hierarchies, none of which can claim absolute legitimacy, is taken as an indication that tastes are now strictly an expression of individual preferences and private choices (Lipovetsky 1983). Taste-based inequalities may persist, but they are individual in nature rather than an expression of class domination. A less extreme but related position is that tastes, along with other cultural resources, remain linked to status competition among groups and classes. Contemporary societies, however, are characterized by multiple, heterologous, and incommensurable status markets rather than by a single, objective space of distinction dominated by a legitimate elite culture (Hall 1992, Erickson 1996). Taste-based domination is constantly challenged, as no one group succeeds in imposing as legitimate its own standards of taste. This leads to the co-existence of multiple spaces of distinction anchored in constantly shifting identities and hierarchies.

Without contesting the co-existence of multiple status markets in contemporary societies, still others argue the blurring of boundaries between taste cultures signals a transformation rather than the end of taste-based class domination (Peterson and Kern 1996). The celebration of cultural diversity expressed by high-status respondents does not mean that tastes no longer serve as a means of distinction, but rather that new rules of distinction are emerging. Upper-class position is no longer expressed as an exclusive taste for a limited range of high-status cultural objects. It is above all characterized by an openness to an array of multicultural objects and practices, which include high culture but are not restricted to it. The omnivore/univore pattern is thus consistent with a new class division which separates high-status omnivores, who prefer and consume a wide range of cultural products, from low-status univores, whose tastes and activities are more restricted.

Regardless of the specific patterns and meanings attached to them, it is clear that tastes are no longer considered an expression of universal and eternal standards of beauty originating outside social experience. Tastes are part of symbolic systems of classification which both express and shape social interaction. They are linked closely to processes of identity formation and status exclusion. Tastes are instrumental to the creation of symbolic boundaries (Lamont and Fournier 1992), whose degree of differentiation, hierarchization, universalization, and ritual potency varies over time and space. Whether this social conception of taste necessarily leads to cultural relativism remains an open question. If there are no objective criteria of evaluation and if aesthetic hierarchies simply reflect the arbitrary values of the dominant classes, nothing distinguishes a Shakespeare play from a comic strip or a pair of boots (Finkielkraut 1987). Whether the equalizing of all hierarchies should be celebrated as the triumph of democracy and the end of all cultural domination, whether it should be lamented as a failure to reach beyond cultural particularities and to tap into truly universal aspects of human existence, or whether these questions are outside the reach of social sciences, remains a pressing issue in societies increasingly shaped by economic and cultural globalization.

See also: Cultural History; Cultural Psychology; Cultural Studies: Cultural Concerns; Culture and Emotion; Culture, Sociology of; Disgust, Psychology of; Mass Society: History of the Concept; Taste and Smell, Psychology of

Bibliography

Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA [1979. *La Distinction: Critique Sociale du Jugement*. Les éditions de Minuit, Paris]

Cingolani P 1984 Eppur si muove! Classes populaires et structures de classes dans *La Distinction*. In: *Collectif Révoltes Logiques: L'Empire du Sociologue*. La Découverte, Paris

DiMaggio P 1982 Cultural entrepreneurship in nineteenth-century Boston: The creation of an organizational base for high culture in America. *Media, Culture, and Society* **4**: 33–50

DiMaggio P 1987 Classification in art. *American Sociological Review* **52**: 440–55

Douglas M, Isherwood B 1979 *The World of Goods*. Basic Books, New York

Erickson B H 1996 Culture, class, and connections. *American Journal of Sociology* **102**: 217–35

Ferry L 1990 *Homo Aestheticus: l'Invention du Goût à l'Âge Démocratique*. Grasset, Paris. [De Loaiza R, Ferry L 1993 *Homo Aestheticus: The Invention of Taste in the Democratic Age*. Chicago University Press, Chicago]

Finkielkraut A 1987 *La Défaite de la Pensée: Essai*. Gallimard, Paris. [1995 *The Defeat of the Mind*. Columbia University Press, New York]

Gans H J 1974 *Popular Culture and High Culture: An Analysis and Evaluation of Taste*. Basic Books, New York [1999 2nd revised 25th anniversary edn.]

Hall J R 1992 The capital(s) of culture: A nonholistic approach to status situations, class, gender, and ethnicity. In: Lamont M, Fournier M (eds.) *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago, pp. 257–85

Horkheimer M 1986 Art and mass culture. In: *Critical Theory: Selected Essays*. Continuum, New York, pp. 273–90

Kammen M 1999 *American Culture American Tastes: Social Change and the 20th Century*. Alfred K. Knopf, New York

Lamont M, Fournier M 1992 *Cultivating Differences: Symbolic Boundaries and the Making of Inequality*. University of Chicago Press, Chicago

Lamont M, Lareau A 1988 Cultural capital: Allusions, gaps and glissandos in recent theoretical developments. *Sociological Theory* **6**: 153–68

Lipovetsky G 1983 *L'ère du vide. Essais sur l'individualisme contemporain*. Gallimard, Paris

MacDonald D 1964 A theory of mass culture. In: Rosenberg B, White D M (eds.) *Mass Culture*. Free Press, New York, pp. 59–73

Montesquieu C 1970 An essay on taste. In: Gerard A (ed.) *An Essay on Taste*. Garland, New York, pp. 257–314
Ollivier M 1987 Modern music and reification. *California Sociologist* **10**(1): 62–78
Peterson R A, Kern R 1996 Changing highbrow taste: From snob to omnivore. *American Sociological Review* **61**: 900–7
Veblen T 1939 [1899] *The Theory of the Leisure Class*. Modern Library, New York

M. Ollivier and V. Fridman

Taste, Neural Basis of

The sense of taste is located at the interface between the uncontrolled chemical environment beyond the body and the highly regulated biochemical environment within. It serves to select, from among a vast array of candidates, those chemicals that satisfy the biochemical needs while identifying for rejection those that have proven toxic. This is accomplished through transduction mechanisms at the receptor level that permit the recognition of sugars, salts, acids, alkaloids, amino acids, fats, and starches. Peripheral taste nerves carry the signal to the brain stem where it activates reflexes for swallowing nutrients or rejecting toxins, then on to the cortex for recognition, and to the lower forebrain for an analysis of pleasure or disgust.

1. *Sensitivity to Chemicals*

The skin is only a crude chemical detector of deviations from pH 7 or certain plant irritants. The exceptions to this insensitivity occur at the points where chemicals enter the body: the nose and mouth. Here, chemical sensitivity is refined, complex, and subtle, for decisions must be made about which substances to admit. Taste may be viewed as the beginning of a long chemosensory tube that extends through the intestines, with receptors along its length that are sensitive to the products released by digestion. Thus, taste should be, and is influenced by the visceral signals that are a continuation of its own function. Together, taste and visceral sensations represent a core of chemical sensitivity that permits an animal to detect a subset of nutrients from an array of toxic or useless chemicals that surround them. What distinguishes taste is not that it recognizes chemicals—glucose is detected in the pancreas, sodium in the kidneys, and amino acids in the duodenum by similar mechanisms to those in the mouth—but that it does so before the irrevocable decision to swallow has been made. While preliminary assessments are made by other senses, notably vision and smell, and through familiarity and cultural norms, nonetheless, taste is the final arbiter of what is taken into the body. Every chemical that enters the mouth goes on trial, and the only verdicts are swallow or not swallow.

2. *Taste Projections into the Nervous System*

Taste input is carried through peripheral nerves to the hindbrain of the central nervous system, from which it projects to at least four locations to mediate distinct functions.

2.1 *Somatic Reflexes to Swallow or Reject*

Clusters of taste cells in the brain stem send projections to local motor nuclei that coordinate acceptance–rejection reflexes. These are based on a preliminary analysis of the chemical composition of a potential food, and elicit swallowing if the substance is deemed nutritious or rejection if it is toxic. The reflexes are fully integrated in the brain stem and are stereotypical for each basic taste. Consequently, they are not altered even by the loss of all neural tissue above the level of the midbrain (Grill and Norgren 1978). This chemical monitor is intact in humans by the seventh gestational month (Steiner 1979).

2.2 *Visceral Reflexes to Aid Digestion*

A second set of projections passes to motor nuclei that initiate digestive reflexes. There are several of these 'cephalic phase' reflexes whereby the brain elicits a preemptory digestive response to prepare for the food that it is simultaneously ordering to be swallowed. These reflexes are mediated by the *vagus* (Lat: wandering) nerve that ramifies throughout the viscera.

2.3 *Discrimination of Quality and Intensity*

A third projection from brain stem taste areas passes up to the thalamus and cortex of the forebrain. Neurophysiological recordings from macaques indicate that taste quality and intensity are represented most accurately in primary taste cortex (Scott and Plata-Salamán 1999). Therefore, the capacity to identify a substance in the mouth most likely derives from cortical activity.

2.4 *Hedonic Appreciation*

The degree to which a taste is appealing or repulsive—its hedonic tone—is determined by projections from primary taste cortex to other cortical and subcortical sites. Prominent among these are secondary taste

cortex in the orbitofrontal region, the amygdala, and the hypothalamus. In macaques, neurons in these areas respond briskly to the taste of glucose when the animal is hungry. As he feeds and becomes satiated, the neurons show progressively less response. The appeal of glucose declines in parallel, finally to turn to rejection as these hedonic neurons fail to respond (Rolls 1995). Even then, however, other tastes may activate these cells and restore the macaque's eagerness to eat. Similarly, humans may be surfeited by a main course, yet show renewed enthusiasm as dessert arrives. It is this enduring sense of pleasure that extends the reflexive bite to a meal; a meal to a diet.

3. Basic Taste Qualities

It remains controversial whether there are basic taste qualities, though most scientists accept the notion at least as a convenient means of organizing the system. The number and identity of the basic tastes is also at issue. Criteria for acceptance include (a) having a specialized transduction mechanism for recognition at the receptor level, (b) eliciting a pattern of neural activity that is distinct from, and does not cross-adapt with, those of other basic taste stimuli, and (c) evoking a taste perception that is unique (in humans) and which does not generalize to those of other tastes (in lower animals). Sweet (glucose, sucrose), salty (NaCl), sour (HCl, citric acid), and bitter (quinine) arouse little debate, and are held by some to comprise the complete range of taste perceptions. Others argue for the inclusion of the amino acid taste *umami* (Japanese, roughly translated as *savory*, represented by monosodium glutamate), which they deem to satisfy the criteria above. Less well-developed cases are presented for the taste of fats and for complex carbohydrates or starches.

4. Taste Receptors

4.1 Anatomy

Humans have about 300,000 taste receptor cells, though the number may vary by a factor of 100 across individuals. They are gathered in groups of about 50 into goblet-shaped structures called taste buds, which number some 6,000. Two-thirds of the buds are located on the surface of the tongue, housed in groups of 3 to 250 in small mounds called papillae (Lat: nipple). Each of the three types of papillae covers a discrete area of the tongue. At the front are 200 fungiform (Lat: mushroom-shaped) papillae, each containing from 0–36 taste buds with an average of three. On each side are about six foliate (Lat: leaf-shaped) papillae, each housing some 100 buds. Toward the back are typically nine circumvallate (Lat: surrounded by a trench) papillae each containing 250 buds. This brings the total number of buds on the tongue to about 4,000. Another 2,000 are not gathered into papillae, but are embedded in the epithelium of the soft palate, pharynx, larynx, and epiglottis (Fisher and Scott 1997).

4.2 Transduction Mechanisms

4.2.1 Sweet taste. Sweetness signals the availability of immediate energy. While sugars are not as dense a caloric source as are fats, they are useful to the brain and muscles only minutes after consumption, and so play a crucial role in survival. Consequently, sweetness has a powerful hedonic appeal, and the word itself has become synonymous with pleasure.

Sugars and most other sweet molecules bind with specific—though as yet unidentified—proteins on taste receptor cells. Once activated, these receptors stimulate G-proteins to amplify the signal. This activates the enzyme adenylate cyclase (AC) which converts ATP to cAMP. It is cAMP that serves as the second messenger, causing protein kinase A (PKA) to close the potassium (K) channels of the cell's membrane. The positively-charge K ions that would otherwise exit through the membrane are instead sequestered within it, depolarizing the cell and causing calcium ion channels to open. This is the event that releases neurotransmitter from the taste receptor onto the peripheral nerve terminals, initiating the signal that is carried to the central nervous system.

4.2.2 Salt taste. Sodium (Na, Lat: natrium) is the crucial ion for controlling movement of body fluids by means of the osmotic force it exerts, and for carrying electrical currents through its positive charge. Its loss devastates normal physiological function. As sodium concentration declines from 140 to below 115 mM, cardiovascular and muscular processes are compromised, and with them the animal's speed and endurance. With further loss, sodium is unavailable to carry electrical potentials resulting in neurological symptoms of stupor, convulsions, and coma. The management of sodium stores is, thus, central to an animal's existence. Accordingly, there is a powerful selection pressure to detect sodium, and reward associated with its taste up to isotonic concentrations.

Sodium enters and activates taste cells simply by passing through ion channels along its concentration gradient. The positive charge the ion carries serves to depolarize the cell and so generates action potentials and transmitter release. These ions are then extruded by Na/K transporters to re-establish the resting state of the receptor.

This occurs readily at the top of the cell where its membrane is exposed to saliva. Access to sodium channels along its side is more controversial. There are tight junctions between taste cells that have been

thought to prevent molecules from passing down their sides. This turns out not to be the case for the small, mobile sodium ion. However, the degree to which sodium can gain access to side channels depends on the anion with which it is paired. Chloride is nearly as small as sodium, passes through the tight junction with it, and allows sodium extensive contact with side channels. Larger anions such as acetate or gluconate are blocked by the tight junction. The negative charge that builds up at this barrier holds the positively charged sodium ion close and prevents it from reaching side channels. Accordingly, the larger the anion, the weaker the salt taste.

4.2.3 Sour (acid) taste. Sourness is the recognition of protons (H^+). These ubiquitous ions participate in many physiological processes. They are also plentiful in unripe fruit and spoiled food, and so serve to warn the animal of danger. These two roles make the recognition of protons a central feature of all chemosensory systems, from protozoans to humans.

The small, mobile proton can penetrate ion channels on all parts of the receptor, and once inside they can affect cellular processes in several ways. Thus, the mechanisms of acid transduction should be varied and subtle. One path has been established and another hypothesized. In amphibians, protons block leaky potassium channels, restricting potassium to the cell and so causing depolarization. In mammals, it is hypothesized that protons enter and so depolarize the cell via sodium channels. This raises an issue of sensory coding that has not yet been addressed, namely, how the animal can discriminate between sour and salty tastes if protons and sodium—both monovalent cations—activate the same receptors by means of the same mechanism. The implication is that other transduction processes for acids remain to be discovered.

4.2.4 Bitter taste. Plants defend themselves against foragers with toxins. Some 10 percent of plant species contain alkaloids, and toxic glycosides are still more common. Animals counter by developing detectors of toxins which, when activated, usually generate a bitter taste and rejection response.

Toxins comprise a variety of molecular structures and disrupt animal physiology in many ways. The taste system cannot focus on a particular molecule, but rather must attend to the common physiological outcome of toxicity. Thus, there will be multiple transduction mechanisms for bitter stimuli. Since most toxic molecules are organic, specialized protein receptors and second messenger systems are likely.

Three mechanisms for bitter molecules have been identified. First, like protons, bitter salts and quinine block leaky potassium channels, restricting potassium to the cell and causing depolarization. How the animal can discriminate bitter from sour when protons use the same mechanism is not known. Second, alkaloids are thought to bind with an undiscovered protein receptor whose activation is amplified by a G-protein. This stimulates the enzyme phosphodiesterase to block cAMP production, a process that could disinhibit calcium channels that are normally held in check by cAMP. The entry of calcium would depolarize the cell and cause transmitter release. Third, bitter stimuli are transduced through the inositol triphosphate (IP_3) second messenger system. The protein receptor, coupled through a G-protein, activates phopholipase C (PLC), the enzyme that produces IP_3. This second messenger then causes the release of calcium from stores sequestered within the cell, resulting in depolarization and release of transmitter. Most toxins are transduced by more than one of these mechanisms. Animals build redundancy into their capacity to detect poison (Lindemann 1996).

5. *Neural Coding*

Once transduced, taste information must be encoded in a neural signal that travels to the central nervous system for processing and interpretation. The nature of that code and the organization of the neurons that generate it have been sources of debate for decades.

5.1 *Gustatory Neuron Types*

Most researchers accept that taste neurons may be divided into three or four discrete types according to the profiles of activity they give in response to a battery of stimuli. Sweet cells and salt cells are clearly identifiable. Acid cells and quinine cells overlap more, and whether they are discrete from one another is doubtful. The centrist position is that taste neurons are of three types: sweet, salt, and generalist, with the last showing almost equal sensitivity to all qualities except sweetness (Smith and Shepherd 1999).

5.2 *The Code for Taste Quality*

The nature of the neural code for taste has been the central controversy of the field for decades. The two positions have come to be called labeled-line and across-fiber patterning.

5.2.1 Labeled-line theory. Proponents of labeled-line theory contend that there is a small number of discrete basic tastes, each served by an associated neuron type. All neurons of one type are dedicated to and responsible for the recognition of their assigned basic taste. Hence, their communication

channels—their lines—may be labeled with that quality. In this view, there are four independent systems whose transduction mechanisms, structures, connections, functions, and very reasons for being may be studied in isolation.

5.2.2 Across-fiber patterning. Neuroscientists had anticipated a labeled-line code for taste from the work of nineteenth century psychophysicists, who concluded that taste was reducible to a small number of independent qualities. The first electrophysiological recording from taste nerve fibers violated that anticipation (Pfaffmann 1941). Single fibers of taste nerves were broadly responsive across the basic taste stimuli, rendering it unlikely that any one fiber would bear responsibility for signaling one taste. A code that involved a pattern of activity across fibers emerged as a more reasonable alternative. In its pure form, patterning theory holds that all neurons contribute equally to the code for each taste. A neuron makes its contribution to the code by filling one place in the activity pattern whose shape represents the taste.

While the theoretical distinctions between labeled-line and patterning are clear, the empirical differences are not. Both accommodate most data. Similar taste stimuli elicit highly correlated patterns of activity, but they do so from a common set of neurons. A patterning theorist does not recognize the need for a channel in this scheme, while a labeled-line proponent sees need for *only* the appropriate channel and nothing else. Still, the same set of neurons, channel or not, is generating most of the activity.

The original finding of breadth of sensitivity in individual taste cells is the irrefutable evidence that favors patterning. Because a neuron responds to multiple taste qualities, its activity is only interpretable in the context provided by the responses of other neurons. A salt neuron may give the same response to a weak salt and to a strong acid. Only by comparing the response to that being given simultaneously by acid cells can it be interpreted. There may be taste neuron types and even coding channels, but they cannot operate in isolation. Taste in the central nervous system must be an integrated sense rather than a federation of four independent systems.

6. *The Reward of Taste*

Eating is fundamentally a process of turning 'other' into 'self.' The raw materials for doing so—carbohydrates, proteins, fats, sodium—must come from the diet. The taste system is charged with identifying them and with providing the reward that motivates their consumption. Cloaked though it may be in table manners, the primitive, taste-inspired reward of eating nutritious foods is irresistible to many.

In modern developed societies, however, taste is viewed paradoxically as an enemy to good health. This is not because the taste system has changed, but rather the definition of health. Where once it was sufficient strength through the childbearing years, it is now clear arteries at age 70. Technical advances in recent centuries have provided nutrients in unprecedented quantity. The over consumption that followed revealed a set of pathologies that had played little role in evolutionary selection: cardiovascular disease, adult-onset diabetes, hypertension, and joint problems. Taste, tuned so exquisitely to the chemical environment in which it evolved, now clashes with the very society that has learned to pander to, and profit from it. The biological reward of eating is dulled by the social disgrace of gluttony, and by the deleterious effects on health that appear only at ages our ancestors never achieved.

See also: Food Preference; Hunger and Eating, Neural Basis of; Signal Detection Theory; Taste and Smell, Psychology of

Bibliography

Fisher C, Scott T R 1997 *Food Flavours: Biology and Chemistry*. Royal Society of Chemistry, Cambridge, UK

Grill H J, Norgren R 1978 The taste reactivity test. II. Mimetic responses to gustatory stimuli in chronic thalamic and chronic decerebrate rats. *Brain Research* **143**: 281–97

Lindemann B 1996 Taste reception. *Physiological Review* **76**: 719–66

Pfaffmann C 1941 Gustatory afferent impulses. *Journal of Cellular Composition Physiology* **17**: 243–58

Rolls E T 1995 Central taste anatomy and neurophysiology. In: Doty R L (ed.) *Handbook of Olfaction and Gustation*. Dekker, New York

Scott T R, Plata-Salamán C R 1999 Taste in the monkey cortex. *Physiology and Behavior* **67**: 489–511

Smith D V, Shepherd G M 1999 Chemical senses: Taste and olfaction. In: Zigmond M J et al. (eds.) *Fundamental Neuroscience*. Academic, San Diego, CA

Steiner J E 1979 Human facial expressions in response to taste and smell stimulation. In: Reese H W, Lipsett L (eds.) *Advances in Child Development*. Academic Press, New York, Vol. 13

T. R. Scott

Teacher Behavior and Student Outcomes

1. *Teacher Behavior and Student Achievement*

One empirical approach to developing information about effective teaching involves documenting relationships between classroom processes (especially

teacher behaviors) and student outcomes (especially adjusted gains on standardized achievement tests). Such process–outcome research has identified some of the classroom management and instructional behaviors associated with achievement gains.

2. Process–Outcome Research

Classroom teaching is difficult to study because it is a multifaceted professional practice that takes place in a complex and evolving interpersonal context. Nevertheless, research on teaching has begun to establish a knowledge base capable of informing teachers' planning and decision making. Especially relevant findings come from studies designed to identify relationships between classroom processes (what the teacher and students do in the classroom) and student outcomes (changes in students' knowledge, skills and dispositions that represent progress towards instructional goals). Two forms of process–outcome research that became prominent during the 1970s were school-effects research and teacher-effects research (see *School Effectiveness Research*).

Teacher effects research (reviewed in Brophy and Good 1986) identified teacher behaviors and patterns of teacher–student interaction associated with student achievement gains. Such research initially was limited to correlational studies and focused mostly on basic skills instruction in the early grades. However, eventually it encompassed a wider range of grade levels and subject areas and included experimental verification of some of the causal hypotheses suggested by its correlational findings. Some conclusions established through this research are:

(a) *Teachers make a difference*. Some teachers reliably elicit greater gains than others, because of differences in how they teach.

(b) *Teacher expectations/role definitions/sense of efficacy*. Teachers who elicit strong achievement gains accept responsibility for doing so. They believe that their students are capable of learning and that they (the teachers) are capable of teaching them successfully. If students do not learn something the first time, they teach it again, and if the regular curriculum materials do not do the job, they find or develop others that will.

(c) *Exposure to academic content and opportunity to learn*. Teachers who elicit greater achievement gains allocate most of the available time for activities designed to accomplish instructional goals. They do not schedule many activities that serve little or no curricular purpose.

(d) *Classroom management and organization*. These teachers establish their classrooms as effective learning environments and gain the cooperation of their students. They minimize the time spent getting organized, making transitions, or dealing with behavior problems, and maximize the degree to which students are engaged in ongoing academic activities.

(e) *Active teaching*. Rather than depend solely on curriculum materials as content sources, these teachers interpret and elaborate the content for students and stimulate them to react to it during interactive discourse. They do not merely maximize 'time on task;' most of their time is spent actively instructing their students during interactive lessons featuring teacher–student discourse, so their students do not spend much time working silently on assignments.

(f) *A supportive learning environment*. Despite their strong academic focus, these teachers maintain pleasant, friendly classrooms and are perceived by their students as enthusiastic, supportive instructors.

In addition to these more generic findings, teacher effects research has contributed knowledge about qualitative aspects of instructional methods and classroom processes. Research on teachers' lectures and demonstrations has verified the importance of delivering these presentations with enthusiasm and organizing their content so as to maximize their clarity and 'learner friendliness.' Various studies have shown the value of pacing, gestures, and other oral communication skills; avoiding vagueness, ambiguity, and discontinuity; beginning with advance organizers or previews that include general principles, outlines, or questions that establish a learning orientation; briefly describing the objectives and alerting students to new or key concepts; presenting new information with reference to what students already know about the topic; proceeding in small steps, sequenced in ways that are easy to follow; eliciting student responses regularly to stimulate active learning and ensuring that each step is mastered before moving to the next; finishing with a review of main points, stressing general integrative concepts; and following up with questions or assignments that require students to encode the material in their own words and apply or extend it to new contexts.

Research on teacher–student interaction processes has identified important dimensions of interactive discourse that need to be adjusted to the instructional goals, the students, and various context factors. The difficulty levels of questions need to be suited to the students' levels of ability and prior knowledge. The forms and cognitive levels of the questions need to be suited to the instructional goals. Primarily closed-ended and factual questions might be appropriate when teachers are assessing prior knowledge or reviewing new learning, but most instructional goals require more emphasis on open-ended questions that call for students to apply, analyze, synthesize, or evaluate what they are learning. Questions should be addressed to the class as a whole rather than to an individual designated in advance. This encourages all of the students, not just the one who eventually is called on, to listen carefully and respond thoughtfully to each question. Another important dimension is wait

time. After posing a question, teachers need to pause to allow students time to process the question and at least begin to formulate responses, especially if the question is complicated or high in cognitive level of response demand.

When teachers have posed a question and elicited an answer, follow-up options include providing immediate feedback about correctness, praising or criticizing the answer, asking other students to comment on it or suggest alternatives, accepting it and moving on, or probing to elicit more information When the student answers incorrectly or is unable to respond, follow up options include repeating the question, providing clues or rephrasing the question, shifting to a lower level question, giving the answer, or redirecting the question to the class. Teacher effects research has not produced findings indicating consistent relationships between these follow-up alternatives and student achievement gain, probably because optimal teacher response varies with the nature and goals of the instruction. These response options are worth considering in planning recitation and discussion activities, however, particularly in determining when and why it makes sense to terminate an interaction with the original respondent by giving the answer or calling on someone else versus sustaining the interaction by repeating or rephrasing the question or giving clues. Sustaining is often desirable as a way to scaffold thinking and develop confidence, especially in reticent or self-doubting students. However, teachers often feel the need to terminate and move on owing to time pressures or growing restlessness among onlooker students.

Teacher effects researchers have not had much to say about desirable activities and assignments, except for studies relating homework practices to student achievement gains. Findings have been equivocal for the elementary grades, but tend to show positive relationships between assignment of homework and achievement gain in the secondary grades. However, Cooper (1989) cautioned that attention is needed to qualitative aspects of homework such as the goal relevance and pedagogical value of the assignments, the degree to which they are suited to students' prior knowledge, how homework performance figures into overall grading, and the degree to which the teacher does not merely assign homework but reviews it promptly, gives feedback and requires correction of omissions or mistakes.

Research on learning tasks suggests that assignments should be varied and interesting enough to motivate student engagement, new or challenging enough to constitute meaningful learning rather than pointless busy work, and yet easy enough to allow students to achieve high rates of success if they invest reasonable effort. The effectiveness of assignments is enhanced when teachers explain the work and go over practice examples with students before releasing them to work independently, then circulate to monitor progress and provide help when needed.

Research on strategy teaching indicates that general learning and study skills as well as domain-specific skills (such as constructing meaning from text, solving mathematical problems, or reasoning scientifically) are learned best if taught as strategies to be brought to bear purposefully and implemented with metacognitive awareness and self-regulation. Many students do not develop effective learning and problem-solving strategies on their own but can acquire them through explicit instruction that includes not only demonstrations of and opportunities to apply the strategy but also explanations of when and why it is used.

Research on cooperative learning indicates that collaborative work in pairs or small groups is associated with a variety of positive affective outcomes. If the cooperative learning combines group goals with individual accountability, it may enhance student achievement gains as well (see *Cooperative Learning in Schools*).

3. *Research on Teaching for Understanding and Use of Knowledge*

Early process–outcome research focused on important but very basic aspects of teaching that did not include the more subtle fine points that distinguish the most outstanding teachers. Also, most of this research relied on standardized tests as the outcome measure, which meant that it assessed mastery of relatively isolated knowledge items and skill components without assessing students' understanding of networks of subject-matter content or ability to use this information in authentic application situations. During the 1980s a newer kind of research emerged that emphasizes teaching for understanding and use of knowledge. It focuses on particular curriculum units or even individual lessons, taking into account the teacher's instructional goals and assessing student learning accordingly. This research indicates that clear explanations and modeling from the teacher are important, but so are opportunities to answer questions about the content, discuss its meanings and implications, or apply it in authentic problem-solving or decision-making contexts. Increasingly, research is pointing to thoughtful discussion, and not just teacher lecturing or student recitation, as characteristic of the classroom discourse involved in teaching for understanding.

Researchers have also begun to stress the complementary changes in teacher and student roles that should occur as learning progresses. Early in the process, the teacher assumes most of the responsibility for structuring and managing learning activities and provides students with a great deal of information, explanation, modeling and cueing. As students develop expertise, however, they can begin regulating their own learning by asking questions and by working on increasingly complex applications with increasing

degrees of autonomy. The teacher still provides task simplification, coaching, and other scaffolding needed to assist students with challenges that they are not yet ready to handle on their own, but this assistance is gradually reduced in response to gradual increases in student readiness to engage in independent and self-regulated learning.

Research on teaching school subjects for understanding and use of knowledge supports the following principles, which might be considered components in a model or theory describing good subject-matter teaching:

(a) The curriculum is designed to equip students with knowledge, skills, values, and dispositions that they will find useful both inside and outside of school.

(b) Instructional goals emphasize developing student expertise within an application context and with emphasis on conceptual understanding of knowledge and self-regulated use of skills.

(c) The curriculum balances breadth with depth by addressing limited content but developing this content sufficiently to foster conceptual understanding.

(d) The content is organized around a limited set of powerful ideas.

(e) The teacher's role is not just to present information but also to scaffold and respond to students' learning efforts.

(f) The students' role is not just to absorb or copy input but also to actively make sense and construct meaning.

(g) Students' prior knowledge about the topic is elicited and used as a starting place for instruction, which both builds on accurate knowledge and stimulates conceptual change if necessary.

(h) Activities and assignments feature authentic tasks that call for problem solving or critical thinking, not just memory or reproduction.

(i) Higher order thinking skills are not taught as a separate skills curriculum but instead are developed in the process of teaching subject-matter knowledge within application contexts that call for students to relate what they are learning to their lives outside of school.

(j) The teacher creates a social environment in the classroom that could be described as a learning community featuring discourse or dialogue designed to promote understanding.

Embedded in this approach to teaching is the notion of 'complete' lessons that are carried through to include higher-order applications of content. This implies the need to limit the breadth of content addressed in order to allow for more in-depth teaching of the content that is included. It also implies the need to structure a great deal of thoughtful discourse. Instead of rapid-fire questioning and short answers, there is sustained examination of a small number of related topics, in which students are invited to develop explanations, make predictions, debate alternative approaches to problems, or otherwise consider the content's implications or applications. Some of the questions admit to a range of defensible answers, and some invite discussion or debate (e.g., concerning the relative merits of alternative suggestions for solving problems).

Enthusiasm for recent findings needs to be tempered by some important qualifications and cautions. First, the research base on teaching for understanding is still quite thin, and few studies have included both comparison groups and systematic measurement of student outcomes. Also, some of the best known and most widely respected innovations, such as Reciprocal Teaching, have produced mixed rather than uniformly positive results. Finally, like the process–outcome research that preceded it, this more recent research has finessed basic curricular issues (i.e., identifying the most important content to teach) rather than addressed them. Process–outcome research did it by using standardized tests as the criteria for learning. More recent research has equated the teaching of school subjects with enculturation into the academic disciplines. This leads to problematic curricular decisions, because the school curriculum should reflect deliberations about what constitutes basic knowledge that all citizens need to know. This knowledge should be consistent with disciplinary knowledge, but it should be selected, organized, and taught as general citizen education. Content does not necessarily belong in the school curriculum just because it is currently of interest to one of the disciplines.

4. *Conclusion*

The best teaching is adapted to the context, including instructional purposes and goals, the students, and the subject matter. Techniques associated with active teaching or strategy instruction are most relevant when the context calls for presenting new information, modeling skills or coaching students as they attempt to implement new learning Techniques associated with teaching for understanding are most relevant when one wishes to develop understanding and appreciation of networks of knowledge through shared construction and negotiation of meanings and implications. The principle of transferring responsibility for managing learning from the teacher to the students applies to all teaching contexts, but determining exactly how to apply it (how much modeling, explanation, coaching, and other scaffolding to provide, and how quickly to fade this support) takes experience with the content and the students. As research on teaching progresses, more is being learned about the complexities involved in adapting instruction to students and contexts.

See also: Open Classroom: Learning and Teaching; Teacher Expertise; Teaching and Learning in the Classroom; Teaching as a Profession; Teaching for Thinking; Teaching: Trends in Research

Bibliography

Brophy J, Alleman J 1991 Activities as instructional tools: a framework for analysis and evaluation. *Educational Researcher* **20**: 9–23

Brophy J, Good T 1986 Teacher behavior and student achievement. In: Wittrock M (ed.) *Handbook of Research on Teaching*, 3rd edn., Macmillan, New York, pp. 328–75

Cooper H 1989 *Homework*. Longman, White Plains, NY

Creemers B P M 1994 *The Effective Classroom*. Cassell, London

Denham C, Lieberman A (eds.) 1980 *Time to Learn*. National Institute of Education, Washington, DC

Gage N 1999 Theory, norms, and intentionality in process-product research on teaching. In: Stevens R J (ed.) *Teaching in American Schools*. Merrill, Upper Saddle River, NJ, pp. 57–80

Good T 1996 Teaching effects and teacher evaluation. In: Sikula J, Buttery T, Guyton E (eds.) *Handbook of Research on Teacher Education*, 2nd edn. Macmillan, New York, pp. 617–65

Good T L, Brophy J E 2000 *Looking in Classrooms*, 8th edn., Longman, New York

Knapp M S 1995 *Teaching for Meaning in High-poverty Classrooms*. Teachers College Press, New York

Pressley M, Beard El-Dinary P (eds.) 1993 Special issue on strategies instruction. *Elementary School Journal* **94**: 105–284

Rowe M B 1986 Wait time: slowing down may be a way of speeding up. *Journal of Teacher Education* **37**: 43–50

Shuell T 1996 Teaching and learning in a classroom context. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Macmillan Library Reference USA, New York, pp. 726–64

J. Brophy

Teacher Education

This entry describes the current status of teacher education across several Western countries. It begins with a brief historic note before reviewing policy initiatives, programs and programmatic orientations, teacher educators, and prospective teachers. The entry concludes on a pervasive theme, namely, that while contemporary teacher education has high potential, it remains fraught with problems.

While teacher education varies a great deal across and within most Western countries, common patterns do emerge. The historic evolution of teacher education shows similarities across Europe and North America. Formal teacher education began in the USA in the 1820s and 1830s with the establishment of 'normal' schools that derived their general design from European models. State and large city normal schools spread rapidly in the latter half of the nineteenth century and the first part of the twentieth century. By 1914 every city in the US with a population of at least 300,000 maintained normal schools, usually in conjunction with the public schools and typically integrated with the curriculum of those schools. However by 1940, normal schools had become largely obsolete, evolving into 2-year teachers colleges. In the latter half of the century they became 4-year state or city colleges and in several instances became multi-purpose colleges and even universities.

This pattern of teacher preparation evolving from normal schools to colleges and university campuses occurred in Canada and several European countries shortly after World War Two. The reasons for this move provide a useful context and perspective for understanding the state of contemporary teacher education in Western society. The policy arguments offered for making the move did not vary much from one country to the other. Feiman-Nemser (1990) described the trend in the USA as a means to professionalize teaching. Europeans and Canadians believed that by moving teacher preparation from colleges to universities candidates with higher academic standing would enter the profession, an improved context for research would be created, and new and more innovative programs for teacher preparation would develop (Gwyne 1976, Myers and Saul 1974). The contemporary picture of teacher education presented in this entry can be judged in part at least by how well these expectations have been met. Clearly, achieving those goals has been mixed. For example Sheehan and Fullan (1995) argue, '…teacher education has failed to achieve the place it deserves in the improvement of education' (89). In 1990, Goodlad and his colleagues concluded their study of teacher education with concerns that remain common today:

> To be sure, the crisis that we speak of (in teacher education) is one of degree, affecting education faculty and their organizations variously. This was certainly the case for the 29 institutions in our study. Yet, we did not sense that there was a strong bond of shared identity in purpose and self-confidence sufficient to ward off intrusion and to develop political strategies for long-term survival and organizational renewal. Perhaps other professional schools in multiversities are similarly affected. But it seems to us that schools of education are particularly vulnerable to being pushed around by arts and sciences, by central administration, and by state lawmakers and certification boards. (1990, p. 398)

On the other hand, progress made by many groups and institutions at the ground level frequently goes unnoticed by the critics and policymakers.

1. *The Policy Arena*

Policy in teacher education differs from one country to another both in terms of who determines it and the nature of that policy. The USA and Canada differ from other countries that have Ministries of Education, which develop and implement national education policies, which affect all levels of education from higher or post-secondary to elementary and secondary education. In the USA and Canada teacher education is a state or province responsibility and elementary and secondary education is viewed as

primarily a local responsibility with an overlay of accountability in each of the 50 state agencies and 10 provinces. Unsurprisingly, then, little curriculum articulation exists between the two sectors and standards for pupil achievement in elementary and secondary schools do not interface with emerging standards for prospective teacher performance. Perhaps most disturbing is the broad scale lack of intersection between reforms in the organization and culture of elementary and secondary schools and those initiated in higher education teacher education. In contrast, the Ministry of Education in Britain has total authority over all levels of education including teacher education. In Australia, both national and state Ministries have jurisdiction. But as Tisher (1995) points out, the federal government in Australia has been increasing its authority in recent years.

The situation in the USA, where policy is now focused squarely on the testing of teacher performance, illustrates this trend for more external control of the enterprise. Driven by recent research which underscores dramatic differences in teacher performance as revealed in student test scores (Sanders and Rivers 1998), the testing and accountability movement has moved from the kindergarden to grade 12 (K–12) sector to teacher education. Measuring teacher performance has become the current focal point in teacher education 'reform' with program approval moving from processes to products. While performance assessments have some ways to go in terms of sound psychometric principles, the majority of States have enacted recent legislation with respect to teacher testing and the National Council for the Accreditation of Teaching (NCATE) has also adopted a teacher performance assessment posture. It should be noted that national voluntary accreditation of teacher education has been available since 1927. What is particularly telling is that less than 600 education schools of over 1300 are accredited, although those accredited do prepare more than two-thirds of the United States' licensed teachers. Thus from a policy perspective major limitations exist here.

Beyond performance assessment, the 50 states vary greatly in terms of policies impacting teacher education. Some of these policies appear enabling, but others do not. The teaching professions, research community and learned societies have typically, but not always, provided the input for these policies. However state legislative bodies have become increasingly intrusive and in some instances have curtailed the number of professional courses and mandated what should be taught. Several States have also initiated alternative routes to certification.

The centralized governance systems in Australia and Britain have produced substantial policy changes in teacher education. Pimm and Selinger (1995) described the situation in Great Britain in terms of the 'commodification of teaching.' Under the Thatcher government initial teacher preparation has been nationalized with proposals to give elementary and secondary schools a stronger role in initial teacher preparation, provide more time in these schools for prospective teachers, and require courses to equip them with specific competencies. Apparently, the government hoped to give schools the full role of training teachers with or without the accreditation of higher educational institutions.

Policy developments in Australia followed a period of public criticism of teacher education (Tisher 1995). The critics saw teacher education as having a weak knowledge base, a shallow curriculum, and an inability to attract top students. Hence policy moves included, among others, encouraging stronger partnerships with the schools such that more of the pedagogical studies would occur there, a stronger base in academic subjects for beginning teachers, and a better integration of pedagogy and research. The policy decisions since have led to numerous closures and amalgamations of institutions that provide teacher education.

Policy in Canadian teacher education has been left largely to the faculties of education themselves. Provincial Ministries of Education who have responsibility for teacher preparation typically limit their efforts to insuring that potential candidates have the necessary courses to graduate. Consequently, intrusive policies of teacher testing and radical restructuring do not exist in Canada. Ministries do occasionally undertake reviews of teacher education in Canada. In Nova Scotia, for example, following such a province wide review, four of eight teacher preparation programs were shut down. In two provinces, Ontario and British Columbia, Colleges of Teachers have recently been established to set standards for certification, but their long terms impact on policy has yet to be determined. (Sheehan and Fullan 1995.)

A common historic policy approach across these countries becomes evident. In the 1960s and 1970s, after teacher education had moved to Universities, schools and faculties of education appeared to enjoy considerable freedom in how they undertook the preparation of teachers. But the 1990s have seen numerous initiatives on the part of governments, particularly in the USA, Australia and Great Britain to exercise greater and greater control over teacher preparation in the form of mandating standards, introducing programs of teacher testing, and placing limits on curriculum within schools and faculties of education. The reaction of academics appears to vary from one country to the next. Tisher (1995) notes that the policy changes in Australia went unchallenged by a docile group of teacher educators. Taylor (1994), who finds teacher education in England under attack, proposes alliances to improve the image of teachers and to assist educators to defend teacher education from its enemies.

The motivation of policymakers to undertake such moves appears to rest with economic concerns. Sanders (1995), for example, who examined the

political influences in Germany, perceives that the need for high levels of productivity in the economy drove policy changes in that country.

In sum, themes emerge across these policy initiatives. First, in every country steps have been taken by policy makers to reform teacher education from outside the institution itself. The reform initiatives within institutions, which we describe elsewhere, have either not matched the expectations of policy makers or have been ignored by them.

Second, most policies mandate or support greater involvement of the schools in teacher education either in the form of school/university partnerships or in other cases control by schools over teacher education. However, such policy proposals for restructuring and innovative activity have not typically been followed by a major reconceptualization of what should actually occur as a result of such collaboration.

Third, across these western countries is a move toward greater accountability which too often translates into a narrower, more technical view of teaching. For example, the British experience is built around a national curriculum with the development of competencies for teachers to teach for that curriculum. The apparent simplicity of many of these policies has the potential of 'dumbing down' the curriculum for teacher education. The notion of teaching as a complex endeavor and the beginning teacher as an educated person appears to have been lost in these policies.

2. *Teacher Education Programs*

The program landscape of teacher education takes the form of a common organizational model within which alternative orientations appear. Traditionally, learning to teach has been a process of providing beginning teachers with knowledge and skills about teaching through coursework followed by a school experience where beginning teachers are expected to apply and integrate such knowledge. Britzman (1986) terms this an implicit theory of simple and immediate integration of knowledge about teaching. The university provides the knowledge, the school the field experience and the students integrate the experience. Because it has more in common with the culture of the university than it does to the culture of the school, links to the field are often ignored. As Guthrie and Clifford (1988) note, as Schools of Education have drifted toward the shores of academe they have drifted away from schools. During that period of drift, schools and faculties of education have lost the clear sense of mission that typified the normal schools. What has changed over the years however is the type of knowledge, skills, and expectations that are acquired by beginning teachers during the coursework phase.

These changes have come about with the move of teacher education to the university which has achieved the intended goal of creating a setting for research and scholarly activity to address issues in teacher education. This has produced a rich and varied knowledge base ranging from research on teaching and reflective practice to the philosophical and psychological studies leading to the notions of constructivism. This diverse corpus of knowledge, which Zeichner (1999) calls the 'new scholarship in teacher education,' can potentially equip beginning teachers with the tools and understanding to better approach their teaching than was possible in the days of the Normal schools.

Also, numerous studies of program attributes have contributed to successful teacher learning (Howey and Zimpher 1989, Goodlad et al. 1990, Darling-Hammond 1996). Howey and Zimpher identified in their case studies the following attributes of coherent programs:

> Programs have one or more frameworks grounded in theory and research as well as practice, frameworks that explicate, justify, and build consensus around such fundamental conceptions as the role of the teacher, the nature of teaching and learning, and the mission of schools in this democracy. These frameworks guide not only the nature of curriculum as manifested in individual courses but, as well, questions of scope; developmental sequence; integration of discrete disciplines, and the relationships of pedagogical knowledge to learning how to teach in various laboratory, clinical, and school settings. Programs embedded in such frameworks clearly establish priorities in terms of key dispositional attitudes and behaviors enabled and monitored in repeated structured experiences. Programs reflect consideration of ethos and culture building and the critical socialization of the prospective teacher. (1989, p. 242)

3. *Varying Orientations to Teacher Education*

While a common organizational tradition has remained prevalent over the years, alternative approaches in all countries have developed. These reforms push against this predominant tradition of teacher education. These tend to occur at the ground level, where individual and small groups of faculty members develop and carry out innovations. These alternatives are driven by competing conceptions of teaching and learning including progressive/critical, personalistic, competency, traditional, and academic orientations. For example, the RATE (Research About Teacher Education) study in the United States in 1990 polled teachers and students and found that about two-thirds of both groups identified a discernible orientation guiding the program they were engaged with at that time. Context greatly affected responses, as liberal arts institutions were disposed toward content knowledge, early childhood and elementary preparation programs were driven by views of child development, and programs placing students in schools populated by the underclass tended toward a critical posture. No prevalent orientation appeared, but hybrids were common and inquiry into and

reflection on practice represented a common programmatic theme.

Initiatives from different countries illustrate this plurality of orientations. At the University of Melbourne, Australia, an innovative program based on constructivist principles has been ongoing for a number of years. Building on the needs of the students, the program emphasizes group collaboration and a unique organizational structure that combines campus coursework and school experience (Gunstone et al. 1993). Similar constructivist programs incorporating a personalistic approach can be found in several Western countries.

Nigel (1995) examine efforts in European countries to increase the professionalization of teacher training. He reports that France had recently elevated the status of teachers to civil servants hence increasing their prestige and that Germany had strengthened efforts to retain a high degree of academic excellence in its professional training.

In Canada, alternative approaches to the traditional organization of teacher preparation have been occurring for several years. At the University of Simon Fraser, the Faculty of Education has had over 30 years of experience with an alternative model of teacher preparation where initial coursework has been replaced by combined seminar and school activity and where much of the teaching has been undertaken by seconded teachers (Ellis 1968). The orientation in these programs integrates a clinical/practical and personalistic stance.

Brower and colleagues (1987) report an integrated curriculum of alternating college based seminars and school experience at the University of Utrecht. What began and was studied over time in one institution later became adopted in most Universities in the Netherlands and became endorsed as official government policy. Their work illustrates how ground level reform can be supported by research and influence national policy.

Increasingly, in the USA formalized interinstitutional relationships between schools of education and surrounding school districts appear in the form of professional development schools or partner schools. Seventy-five major institutions working in partnership with local schools and school districts form part of a national reform network known as the Holmes Partnership. Within this confederation, 31 partnerships address the challenges of preparing and retaining teachers in urban settings in a reform effort known as the Urban Network to Improve Teacher Education (UNITE).

The United States, Great Britain and Australia have also been characterized by 'blue ribbon' commissions convened regularly to share their collective wisdom about what they perceive as needed future directions for teacher education. Two examples from the USA illustrate these types of reports. A recent report was prepared by a group with the auspicious designation of the National Commission on Teaching and America's Future (NCTAF) titled *What Matters Most*: *Investing In Quality Teaching* (1996). Rennert-Ariev and Valli (1999) examined nine other reform documents disseminated since 1985 against the NCTAF recommendations and found considerable variability across these reports. What these groups, comprised of leaders in education, government, business and industry, have consistently recommended is more rigorous content preparation and increased and improved clinical experiences. A high level of agreement also existed for improving the awareness of multiculturalism and increasing the competence of beginning teachers to deal with students of diversity.

4. *Teacher Educators*

While the role of teacher educators has attracted little research they do shape programs and their delivery. Just who is involved in preparing teachers, what do they tend to do and why? In addition to the school of education professoriate, clinical faculty and professors in the arts, sciences and humanities also play significant roles in shaping the beliefs and practices of beginning teachers.

Ducharme and Ducharme's (1996) research, corroborated by other studies, revealed the following general portrayal of the teacher education professor:

> He is a white male in his late forties or early fifties, tenured at either the full or associate professor level. He acquired his doctorate while studying part-time, and has been at his current place of employment for more than 15 years, during which time he has published six or seven articles in refereed publications. Nearly 80 percent of elementary and secondary (teacher education) faculty had several years of prior work in elementary and secondary schools. (Ducharme and Ducharme 1996, p. 698.)

The scholarly productivity of teacher educators varies widely depending upon the type of institution in which they are located and when they entered the professoriate. For example, those faculty in research institutions tend to publish as frequently as other faculty across the academy and teacher education research has in fact mushroomed in recent years. Division K, Teaching and Teacher Education, of the American Educational Research Association now has the largest membership in this organization and reflects the greatest number of scholarly presentations. Contrary to the claims of critics, the RATE studies revealed that prospective teachers report that teacher educators on the whole are both competent and accessible. Further, work in Canada (Weber 1990) reveals a professoriate committed to high ideals and to a reformist agenda in teacher education, but a context in which they are frustrated by the institution.

Teacher educators reported 55 hours a week on the average devoted to their various tasks and the demands on faculty in recent years have increased as many struggle to straddle the diverse cultures of higher education and elementary and secondary schooling. Teacher educators report spending more time in schools and especially in the growing movement of Partner or Professional Development Schools. However while these teacher educators generally spend increasing time in schools, they tend to do so largely engaged in traditional functions; that is supervising prospective teachers and providing some school-based professional development for veteran teachers. They rarely, if ever, teach youngsters themselves and since it has often been a considerable number of years since they did so, they are often unaware of many of the problems teachers face.

Increasingly, clinical faculty, teachers seconded from the schools to work in Schools and Faculties of Education, play a significant role in teacher education. They typically bring to the university a strong connection with the field, but often represent the unheralded, forgotten workers in teacher education. Teachers in elementary and secondary schools have always played a significant role in educating beginning teachers by acting as cooperating teachers. However, they typically receive little compensation for assisting novices in the complexities of learning to teach while apprenticing at their schools. Current changes in Great Britain propose that the locus for teacher education shift to the school. If this signals the future, then the role of teachers as teacher educators will become even more significant. It should also be recalled that the attitudes and perceptions of future generations of teachers are being shaped through what is termed an 'apprenticeship of observation' as they go through school themselves.

Obviously, the professors of Arts, Science, and Humanities also have considerable influence on those learning to teach. Seldom mentioned in the literature, their influence remains relatively unstudied and understood.

5. *Prospective Teachers*

An understanding of the large and diverse enterprise of teacher preparation in Western countries must also include an examination of the prospective teachers seeking a teaching license. The context in which they teach adds to that understanding.

Criticism of teacher education derives not only from perceptions of the quality of faculty and programs but that of the education students as well. The prevailing perception is that schooling in general is severely constrained by the pedestrian nature of those who choose to teach or, as has been opined by many, those who have no other choice. The RATE data counter these common assertions. Almost 4 in 10 prospective teachers maintained an A average in high school, the great majority had at least a B average and six of seven teacher candidates reported that they were in the top third of their class. In several Canadian provinces entry standards for prospective teachers have risen steadily to the point that an 'A' standing is typically required to enter teaching.

Prospective teachers are somewhat evenly divided in the structure of programs in which they are enrolled. Some begin in their freshman year continually interfacing with their liberal studies and finish with a baccalaureate. Others begin in their junior or even senior year. In many instances the program of preparation is extended to a fifth and in a few instances even a sixth year. Most teachers take at least 5 years to finish their program of teacher preparation. Other teacher preparation programs are strictly at the post-baccalaureate level and some result in a master's degree which is a requirement in some American States, including California.

The research on the 'beginning teacher' reveals a great deal about teacher education programs and confirms many of the points we identified earlier. Where beginning teachers have been asked to assess their programs of teacher preparation, they typically take a critical stance reporting that their real understanding of teaching began when they entered the school. Wideen et al. (1998) cite several reports showing that the first year of teaching for most beginning teachers became a 'sink or swim' proposition for which they felt poorly prepared. Nonetheless upon graduation the vast majority of teachers report they have acquired entry-level skills and are confident they will succeed at least in traditional teaching settings (RATE 1992).

6. *The Future*

Thus, what can we say about the future of teacher education? Depending on the data one decides to emphasize, the future could be the best of times or it could be the worst of times. Those who see the best of times emphasize and believe in the potential of the ground level efforts occurring in all western countries; the university based research efforts, producing new ways of thinking about teaching and teacher education; and the widespread effort that is being made to reform the schools and prepare teachers differently to teach in these new schools. Those on the optimistic side typically ask, what could be more important then the preparation of those who will teach our young?

On the other hand, if the simplistic solutions of many policymakers win the day, if the weak and diffuse institutional base for teacher education remains, if many in the professoriate remain interested in everything other than teacher preparation, and if the public continues to believe that teaching is a relatively simple activity, which can be done by anyone, then it could be the worst of times.

Achieving the best of future times requires that the collective teacher education community work to move beyond the mandates of policymakers, reduce isolationism, and avoid simple solutions to complex problems. Moving beyond mandates requires a reconceptualization process that sets out improved designs for teacher preparation.

Isolationism, whether it involves research, program initiatives, or policy making, will have little impact if it ignores the holistic nature of teacher education. The tendency has been to offer simple solutions to complex problems. Teacher education is a complex phenomenon; future solutions to problems and policy initiatives must reflect this.

See also: Teacher Expertise; Teaching as a Profession

Bibliography

American Association of Colleges for Teacher Education 1992 *Research About Teacher Education: Teaching Teachers, Facts and Figures*. American Association of Colleges for Teacher Education, Washington, DC

Britzman D B 1986 Cultural myths in the making of a teacher: Biographical and social structure of teacher education. *Harvard Educational Review* **56**(4): 442–56

Brower N 1987 *Cooperation Structures in Preservice Teacher Education Programs and their Effects on Beginning Teachers' Classroom Performance*. A paper presented at the Annual Meeting of the American Educational Research Association. Washington, DC

Darling-Hammond L 1996 Educating teachers for the next century: Rethinking practice and policy. In: Griffin G A (ed.) *The Education of Teachers*. The University of Chicago Press, Chicago

Ducharme E R, Ducharme M K 1996 Development of the teacher education professoriate. In: Murray F B (ed.) *The Teacher Educator's Handbook*. Jossey-Bass, San Francisco

Ellis J 1968 Preparing teachers: The present and the future. *Journal of Canadian Studies* **10**(2): 123–33

Feiman-Nemser S 1990 Teacher preparation: Structural and conceptual alternatives. In: Houston W R (ed.) *Handbook of Research on Teacher Education*. Macmillan, New York, pp. 212–33

Goodlad J I, Soder R, Sirotnik 1990 *Places Where Teachers are Taught*. Jossey-Bass, San Francisco

Gunstone R F, Slattery M, Baird J R, Northfield J R 1993 A case study of development in preservice science teachers. *Science Education* **77**(1): 47–73

Guthrie J W, Clifford J G 1988 *Ed School*. The University of Chicago Press, Chicago

Gwyn R 1976 *Current Trends in Teacher Education*. Council for cultural co-operation: Committee for Higher Educational Research, Stasbourgh, France

Howey K R, Zimpher N L 1989 *Profiles of Preservice Teacher Education: Inquiry into the Nature of Programs*. State University of New York Press, Albany, NY

Larabee D 1992 Power, knowledge and the rationalization of teaching: A genealogy of the move to professionalize teaching. *Harvard Educational Review* **62**(6): 123–55

Myers D, Saul D 1974 How not to reform a teacher education system. In: Myers D, Reid F (eds.) *Educating Teachers: Critiques and Proposals*. OISE Press, Toronto

National Commission on Teaching and America's Future 1996 *What Matters Most: Teaching for America's Future*. National Commission on Teaching and America's Future, New York

Nigel N 1995 Initial teacher education in France, Germany, and England and Wales: A comparative perspective. *Compare* **25**(3): 211–26

Pimm D, Selinger M 1995 The commodification of teaching: Teacher education in England. In: Wideen M F, Grimmett P (eds.) *Changing Times in Teacher Education*. Falmer, London

Rennert-Arien P L, Valli L 1998 A framework for identifying answers: Agreement and disagreement among teacher education reform documents. *Paper Presented at the Annual Meeting of Project 30*. University of Maryland, College Park, MD

Rennert-Ariev P L, Valli L 1999 A framework for identifying consensus: Agreement and disagreement among teacher education reform documents. *Paper Presented at the Annual Meeting of the American Association of Colleges for Teacher Education*. Washington, DC, February, ERIC_NO: ED430941

Sanders T 1995 Quality improvement and austerity measures in teacher education: Lessons from Germany. *European Journal of Teacher Education* **18**(1): 97–113

Sanders W L, Rivers J C 1998 *Cumulative and Residual Effects of Teachers on Future Students Academic Achievement*. University of Tennessee, Knoxville, TN

Sheehan N, Fullan M 1995 Teacher education in Canada: A case study of British Columbia and Ontario. In: Wideen M F, Grimmett P (eds.) *Changing Times in Teacher Education*. Falmer, London

Taylor W H 1994 Allies for action in the search for improving teacher education. In: Hartley D (ed.) *International Perspectives on Teacher Education: Agenda for Action*. (Report on the JET Colloquium, Curia Portugal) Bradford Community College,

Tisher R P 1995 Readjustments, reorganization or revolution? The changing face of teacher education in Australia. In: Wideen M F, Grimmett P (eds.) *Changing Times in Teacher Education*. Falmer, London

Weber S 1990 The teacher educator's experience: cultural generativity and duality of commitment. *Curriculum Inquiry* **20**(2): 141–59

Wideen M, Moon B, Mayer-Smith J 1998 A critical analysis of the research on learning to teach: Making the case for an ecological perspective on inquiry. *Review of Educational Research* **68**(2): 130–78

Zeichner K 1999 The new scholarship in teacher education. *Educational Researcher* **28**(9): 4–15

K. R. Howey and M. Wideen

Teacher Expertise

Research on teacher expertise is concerned with the knowledge and skills that characterize the successful teacher. The term 'teacher expertise' covers both the teacher's professional know-how and an approach to

research that compares experienced and successful teachers (*experts*) with beginning teachers (*novices*) and so-called *intermediates* (teachers who have already gained some experience). This research approach ties in with basic research on experts in experimental psychology. This field has shown that peak performance in experts is due particularly to knowledge gained through training, as well as through many years of practical experience. In teachers, this professional knowledge covers a wide range of different domains: professional knowledge on the curriculum is joined by pedagogical knowledge on how to teach the subject. Moreover, ideas on the philosophy of the particular school subject are also important. Professional knowledge is not only significant as the content of teaching, but also influences the actions of experts by shaping their categorical perception of classroom events and of students. Teacher expertise also includes action routines that permit quick responses in complex situations.

Whereas it may be taken for granted that members of many qualified professions (such as physicians or engineers) can be called experts, this is not the case for teachers. Less than one-half of the respondents in a survey of teacher-training students considered the expert concept to be appropriate for their future profession (Bromme 1997a). One reason behind this is the idea that teaching and training students is—particularly in elementary school—such an eminently personal process of intersubjective communication that it is impossible to describe the critical requirements in terms of the expertise concept. Expertise, according to this argument, is an idea that has more to do with a technological rationality. In contrast, the teacher-as-expert approach emphasizes that the art of teaching is based on knowledge and skills that can be acquired through training and refined through practice, and that is accessible to empirical analysis.

1. Three Paradigms of Teaching Research

An inspection of the recent history of research on teaching and learning will make it easier to understand the origins of the teacher expertise approach. Over the last 50 years, three paradigms of teacher research can be distinguished (Bromme 1997b).

1.1 The Teacher Traits Paradigm

Empirical teaching research in the 1950s and 1960s focused on the search for positive teacher traits. The interest was in finding personal characteristics that would explain differences in the type of educational impact (not necessarily in learning gains) and could be measured with psychometric instruments. However, this paradigm had two main weaknesses: first, the personality traits proved to be either trivial (cruel teachers have an unfavorable impact) or very complex; second, it was unable to explain how these characteristics impact on different classroom situations, different subjects, and different school grades.

1.2 The Process–Product Paradigm of Searching for Effective Teacher Skills

The methodological influence of behaviorism led to the emergence of the process–product paradigm which measured the effect of narrowly defined teacher skills on student achievements in the classroom. It studied individual teacher behaviors, such as the number of questions, praise following student statements, the clarity of the teacher's language, or requests for students to answer questions. These skills were defined as closely as possible to behavior (i.e., with a low level of inference).

When the process–product paradigm first emerged, it was assumed that the teacher exerted a direct influence on the students' learning process. However, it became clear—roughly at the beginning of the 1970s—that the impact of a specific teacher action on student achievement depended just as much on the students' own activities, interpretations, and interactions as on the subject, the timing of the action, and the sequence (not just the aggregated frequency) of prior events. This reveals one of the weaknesses of a research paradigm that searches for isolated teacher skills. One outcome of this weakness was a brusque rejection of the process–product paradigm; another, its programmatic remodeling. The expert paradigm is a product of this remodeling (Weinert et al. 1992).

1.3 The Expert Approach in Teacher Research

Because the art of teaching consists, above all, in selecting appropriate actions for each situation and modifying one's activities according to the type of student, teaching content, the stage reached in the academic year, and the goal of teaching, it also becomes apparent that the teacher's decisions and deliberations take on a particularly important role. Their direct, action-related cognitions are, in turn, embedded in more long-term considerations regarding the syllabus and in lasting goals and values. Furthermore, they call for the perception of student characteristics and classroom events and, naturally, lasting knowledge on the contents of teaching and the teaching process. These cognitions are the focus of research on teacher expertise.

The process–product paradigm broke down the person of the teacher into a bundle of subskills and was not concerned with whether these could all be realized in one person. The expert approach, in contrast, searches for the competence of the teacher in the sense of a coherent ensemble of knowledge and

skills that coalesces in the individual. In line with the personality paradigm, the expert approach brings the person of the teacher back into the focus of research—no longer in terms of generally defined personality traits, but as the necessary knowledge and skills for designing teaching processes, however.

This remodeling was accompanied by a change in the metaphor guiding research and, simultaneously, in the underlying general psychological theories in studies on the teaching-learning process: The (successful) teacher is now viewed as an expert, and this expresses a programmatic reference to the questions, research methods, and perspectives of research on experts in cognitive psychology (Ericsson and Smith 1991). This field investigates the cognitive conditions underlying peak performance on complex mental tasks. One example is the empirical confirmation that the ability of good chess players to memorize rapidly complex constellations of figures and reproduce them exactly is not based on a general (topic-unspecific) higher memory capacity, but on the knowledge about chess positions acquired through many thousands of hours of practice. Further findings on experts concern their problem-solving strategies, metaknowledge, and self-monitoring strategies, as well as their categorical perception of problems.

The expert approach not only focuses on experts as peak performers like chess grand masters. When addressing other domains of expertise (e.g., medicine, physics), it is also concerned with professionals (possessing practical experience and good performance) by comparing them with novices and intermediates. A consistent pattern of findings indicates that it is particularly domain-specific declarative and procedural knowledge that permits expert performance. This does not just involve a quantitative difference (experts simply know more). It is the *qualitative differences in the contents and organization of knowledge* that are more decisive.

2. *Who Is to Say Who Is an Expert?*

Naturally, specifying how experts should be defined for research implies a normative decision. A favorite measure is student-learning gains—although, of course, any computation of this measure has to control statistically for the variety of preconditions present among students. The vast number of other factors in addition to teacher behavior influencing student achievements makes it necessary to use large samples, along with major standardizations of teaching contents, before this criterion can be applied for the selection of experts. Such conditions are not always available—often just for practical reasons. Furthermore, studies on expert knowledge are difficult to carry out with large samples, because tapping the professional knowledge of teachers often also requires intensive individual interviews and observations. Therefore, experts are also identified with further external criteria, such as: ratings by colleagues, superiors, and students; classroom observations; completion of certain training; extent of training; and, finally, length of professional experience. The latter naturally takes it for granted that one can talk about experts only in the sense of professionals. The term no longer implies any particular quality of instructional outcome.

3. *Examples of Studies on Teacher Expertise*

3.1 *Differences in the Categorical Perception of Classroom Events*

Berliner (1992) and his colleagues compared the cognitive processing of photographed classroom situations, videotapes, and written information on students in experts, novices, and intermediates. They found clear differences in the categorical perception of classroom events: whereas the experts' perspective was characterized by concepts about typical classroom events and consideration of the class as a whole, novices tended to perceive the separate individual students in the class. Experts tended to interpret what they saw, whereas novices only described the scenarios (Carter et al. 1987).

It seems that teachers perceive the lesson and structure it for themselves more holistically in the form of typical *classroom episodes*. This perspective also shapes their recall of past classroom events. Bromme (1987), for example, asked experienced mathematics teachers immediately after the lesson whether they could remember problems encountered by individual students and the progress they had made. At first glance, results seemed disappointing: teachers could recall hardly any problems or the learning progress in single students. Instead, their answers revealed that the categorical units of recall (and probably also of situation-specific perceptions) were not so much individual students but more the classroom episodes that provided the communal staging of the teaching content. Putnam (1987) has suggested that these proposed screenplays in teachers' minds should be called *curriculum scripts*. They are part of the teacher's professional knowledge and form a focus of empirical research within the expert paradigm.

3.2 *Stable but Flexible Pursuit of Action Goals*

A further empirical finding on expert teachers addresses the flexibility of their actions. Leinhardt and Greeno (1986) have reported, for example, that their experts were characterized particularly through a situationally appropriate (stable but flexible) pursuit of an elaborated repertoire of goals. Concrete goals in

the classroom were, for example, checking homework in all students, presenting a mathematical algorithm, or carrying out practice exercises. However, expert teachers pursued these concrete action goals in completely different ways depending on which difficulties arose. Moreover, compared with novices, they also pursued a larger number of different subgoals simultaneously.

The idea that the intraindividual breadth of the action repertoire is one of the conditions underlying successful work is also confirmed indirectly through similar findings on the interindividual variance in the classroom activities of expert teachers. The students' learning processes can be supported and encouraged by teachers in a broad variety of ways. Weinert et al. (1989) have shown, for example, that expert mathematics teachers who achieve not only above-average learning gains in their students but also a reduction in the divergence in achievement in their class differ greatly in the teaching behavior they use to attain these goals.

3.3 Routines and Flexible Action

Such flexibility nonetheless also presupposes routines; in other words, many of the behaviors needed to teach the lesson are enacted without any particular conscious decisions. As in any complex activity, practiced routines facilitate action and provide the necessary preconditions for setting targets, carrying out fine-tuning, and simultaneously bearing other aspects in mind. Research on teacher expertise also discusses routines in another sense that has to be distinguished theoretically from the psychological process of automating complex activities. *The concept of routine* is also used to describe the socially established activity patterns that teachers and students pursue during the lesson. It has been confirmed empirically that it is favorable for teachers to establish a set of behavioral rules as quickly as they can after taking over a class. These then provide, so to speak, the framework for spontaneous behavior. Agreement on set behavior patterns also reduces cognitive strains and encourages the formation of routines in the former sense. Furthermore, routine expert activity is strongly domain-specific and is determined by the extent of prior experience: when experienced teachers have to face new demands (e.g., a new teaching method), they tend to behave in similar ways to novices (Rich 1993).

4. Important Variables Related to Expertise

4.1 Expertise and Experience

Basic cognitive research on expertise in narrowly defined domains (e.g., chess) has confirmed empirically that years of practice is an important predictor for the establishment and development of expert knowledge. Can a (linear) relation between job experience and teaching success also be found in teachers? Although empirical findings are inconsistent, indications are that the relationship is curvilinear (Barnes 1985). At the beginning of their professional career, teachers have to accumulate practical experience. However, as the years of teaching go by, stress factors also emerge that, in turn, impact on teacher behavior in a way that generally does not lead to further improvement but to more of a deterioration in instructional outcomes. Emotional exhaustion and its effects due to occupational stress have been summarized under the label 'burnout.' Because *enthusiasm* about the contents of teaching impacts positively on the learning gains of students, it is plausible to assume that any strains impairing such enthusiasm may increasingly mask the positive effect of cumulating experience. Moreover, burnout is a contributing factor in the relatively high proportion of teachers who quit their jobs before reaching retirement age. The resulting selection effects could also be one reason for the inability tofind any statistically significant relations between occupational success and the length of job experience.

4.2 Expertise and the Extent of Subject-matter Knowledge

As well as confirming the role of length of practice—as mentioned above—basic research in cognitive psychology has also revealed that the extent and quality of domain-specific knowledge are important predictors of experts' achievements. In the case of teacher expertise, it has to be noted that, as with years of teaching, moderating factors have to be taken into account when assessing the relation between subject-matter knowledge and the expert's instructional outcomes.

In the 1970s, some surprising empirical studies were published that revealed no measurable connection between the extent of teachers' subject-matter knowledge and instructional outcomes. Druva and Anderson (1983) have summarized the empirically established relationships between the extent of education in the natural sciences and both teacher and student behavior, as well as performance in class. The number of courses the teachers had taken in sciences (as a measure of their subject-matter knowledge) explained about 10 percent of the variance in student performance. Although this may seem rather small, it is actually a good predictor of student performance, because individual teacher-related variables are never able to explain more than a relatively small part of students' learning outcomes. It seems to be self-evident that teachers must possess the subject-matter knowledge they are supposed to teach.

This, however, does not permit the conclusion of a direct linkage between the extent of subject-matter knowledge and students' outcomes as measured by standardized tests. The relation between subject-matter knowledge and student outcomes can be ascertained only by making qualitative differentiations in subject-matter knowledge; in other words, by analyzing more precisely what professional knowledge experienced teachers possess. On the side of teaching activity, there is a need for both a theoretical and empirical clarification of which mediating processes are to be found between the quantity and, above all, the quality of the teacher's professional knowledge and instructional outcomes. Up to now, such mediating processes have been described mostly only in case studies (e.g., Stein et al. 1990).

5. *A Cartography of Teacher Knowledge*

A comprehensive description of the professional knowledge of teachers requires some kind of cartography of teacher knowledge that could define the most important areas more precisely and allow a systematic study of the impact of the proposed variables on the instructional process. It is not enough to take the catalogue of subjects that teachers have to study during their training, because this overlooks areas that develop only implicitly and through practical experience. Extending a proposal from Shulman (1986), the following categories can be proposed.

(a) *Content knowledge*. This covers all elements of knowledge that refer to the field to be taught or the school subject (e.g., knowledge from the domain of mathematics in mathematics teachers).

(b) *Curricular knowledge*. The contents of learning in, for example, mathematics instruction are not just a simplified form of mathematics but make up an autonomous canon of knowledge. The relevant instruction concepts cannot simply be derived from the state of knowledge in the scientific discipline. Instead, ideas about the goal of teaching (e.g., concepts on general education) intermingle with subject-specific meanings. A psychological analysis of professional knowledge requires discrimination between content knowledge and curricular knowledge, because the aspects and focuses essential for implementing the curriculum may well conflict with the self-images of teachers who perceive themselves primarily as competent representatives of a discipline.

(c) *The philosophy of the school subject*. These are the ideas regarding the usefulness of subject-matter contents and how they relate to other areas of human life and knowledge. The philosophy of the school subject is also an implicit content of teaching. During, for example, the mathematics class, students learn whether their teacher considers that the essence of mathematics is operations with a clear, previously defined language, or rather the referential relations of the symbols applied, or that mathematics is primarily a tool for describing reality. Numerous studies on sciences and mathematics have provided empirical confirmation that these general ideas impact on teaching practice (e.g., the type of teacher explanations or the integration of student utterances into the development of the subject matter (Brickhouse 1990)).

(d) *General knowledge on pedagogics and didactics*. This covers that area of knowledge that is important for optimizing teaching-learning situations but is relatively independent from the individual subject. It includes, for example, knowledge on how to generate and maintain the necessary patterns of interaction and working styles for the planned course of instruction, or knowledge about how to deal with difficult situations and problem cases. This field can be differentiated along the same lines as the domains of subject-matter-related knowledge to produce two subdomains. The first domain encompasses knowledge in a narrower sense, such as empirically tested facts, rules, or tricks of the trade. A second domain refers to the aspects of pedagogic philosophy. This includes, for example, normative ideas and issues such as educational goals and attitudes.

(e) *Subject-specific pedagogic and psychological knowledge*. The logical structure of the subject-matter content does not permit any inherent decision on the optimal design of teaching. Empirical analyses of teaching have revealed broad variations in the didactic approach of different but equally successful teachers, even when they are teaching exactly the same curriculum (Leinhardt and Smith 1985). This shows how highly individualized the knowledge of professional teachers can be. It is something that emerges during the course of their careers through an increasing integration of general pedagogic, didactic, and psychological knowledge and specific subjective teaching experiences.

Systematic (empirical) research on the internal structure of teacher knowledge is only just beginning. Any study of the professional knowledge and actions of teachers also has to consider the possibility that research participants accumulate their own experiential knowledge that may well differ from the available theoretical concepts learned during their professional training. If this possibility is not taken into account, the real *wisdom gained through practice* may be underestimated.

6. *The Psychological Structure of Professional Knowledge*

Whereas the cartography described above deals with the semantic content—in other words, the different domains of teacher knowledge—it is also necessary to distinguish the psychological structure of knowledge, i.e., the quality of mental representation. This requires further conceptual differentiations. One widespread but very imprecise differentiation is that between

declarative knowledge (e.g., on facts and relationships) and procedural knowledge (i.e., the know-how for putting actions into effect). However, empirical analyses of teacher expertise require more precise theoretical differentiations of knowledge types like those proposed by, for example, de Jong and Ferguson-Hessler (1996). The authors discriminate between, for example, *situational knowledge* (referring to typical episodes and contexts), *conceptual knowledge* (about facts and principles within one domain), *procedural knowledge* (about actions), and *strategic knowledge* (a metacognitive component). However, the knowledge domains described above do not appear within one knowledge type alone. For example, curricular knowledge includes not only conceptual knowledge, as is immediately apparent, but also episodes containing typical task structures for the lesson. Likewise, the philosophy of the school subject should not be classified to strategic knowledge alone, but also contains elements of conceptual knowledge.

A differentiated discrimination of content domains and knowledge types is important, because it can explain why the formation and development of expertise requires a variety of different types of learning.

7. Teacher Training: The Development of Expertise

The structural ways in which teachers are trained vary greatly from country to country as well as according to the type of school within each individual country. Nonetheless, an inspection of the dimensions of this variation reveals that the organization of teacher training is always based on a—frequently only implicit—different weighting of knowledge domains and knowledge types. In other words, each specific organizational structure of teacher training also reflects different ideas about what makes up the expertise of teachers and how it comes about.

The first organizational dimension deals with the relation between *curriculum-related* and *educational-psychological* training. In many countries, training for teachers in secondary education tends to focus on subject-matter and curricular aspects, whereas elementary school teachers receive more instruction in educational psychology and didactics.

A further dimension is the *proportion of practical elements* in training. There are major differences regarding how far teacher students participate in actual lessons during their training. Likewise, variations are to be found in the relation between teacher-training institutions and the universities. Teachers do not all receive university training for all types of school in all countries. In some cases, the entire course is completed outside the university system in independent academies. In other cases, some phases of training take place at universities; others, in special teacher-training institutions. These practices are associated with different ideas about the best way to impart situation- and context-related expert knowledge.

A further dimension involves the *specialization of courses available for teachers*. In some systems, the decision to aim for a teaching qualification is made before entering tertiary education. In other countries, teacher students are recruited among persons who have already completed part of their university education, bringing different degrees of polyvalence into training. This, once again, harbors the issue of the domain specificity of the knowledge that teachers need to acquire in order to fulfill their professional tasks.

All these differences reflect critical conditions in the development of expertise, and they also weight different components of expertise in specific ways. For example, an intensive, academically oriented training broken down into subject-matter fields and oriented toward the corresponding scientific disciplines may well facilitate the development of subject-matter-related knowledge. However, the above-mentioned cartography of knowledge also shows the need to distinguish between subject-matter domains (e.g., mathematics, history) and the canon of knowledge on the subjects taught in schools. This reveals that there is also a need for special training elements that address cognitive processing of this difference.

However, at the present time, there is no general, empirically and psychologically founded model of the development of individual expertise from which a normative model of teacher training could be derived. Indeed, the wide breadth of potentially effective teacher behaviors would indicate that such an unequivocal model is probably impossible. Nonetheless, a few general findings provide at least indirect orientations regarding questions on the design of ways to train teachers. Some examples are: (a) It is clear that the development of expertise takes a domain-specific path. Therefore, training should be oriented toward the particular field of work at the earliest possible stage. (b) Expertise covers different types of knowledge, and its development requires time: time for an active analysis of the theoretical foundations of the profession (e.g., the subject-matter-related knowledge) and enough time to gather practical experience. (c) Because the mere length of a career does not lead to the development of expertise, metacognitive foundations for processing personal events are also required, so that merely passing thorough one's career actually turns into an experience leading to improved performance. (d) The research on teacher expertise reported above has shown that situation-related, experience-based knowledge (e.g., on appropriate teaching-learning arrangements structured as academic tasks for students) is an important part of expert knowledge. It can also be confirmed that the development of these components of expertise needs to be supported by the teacher students' personal confrontation with real-life situations. Nevertheless, the hidden *curriculum of field experiments* is not enough by

itself to provide an adequate preparation for high performance as a teacher (Putnam and Borko 2000). Immersion in practical experience can also lead, for example, to a reactivation of personal experiences as a student that may tend to trigger rather negative instructional behaviors. Hence, the development of teacher expertise also requires components of teacher education that make it possible for teacher students to distance themselves from and reflect on the prior experience they bring with them from their own school days. This concerns categorical and metacognitive knowledge that has to be conveyed in an academic way and in a form that also promotes a thoughtful detachment from the primary experience.

See also: Expertise, Acquisition of; Teacher Education; Teacher Stress and Burnout: Methodological Perspectives; Teaching as a Profession; Teaching: Trends in Research

Bibliography

Barnes J 1985 Teaching experience and instruction. In: Husen T, Postlethwaite T N (eds.) *The International Encyclopedia of Education*. Pergamon Press, Oxford, UK, Vol. 9, pp. 5124–8

Berliner D C 1992 The nature of expertise in teaching. In: Oser F K, Dick A, Patry J–L (eds.) *Effective and Responsible Teaching*. Jossey-Bass, San Francisco, CA, Chap. 15, pp. 227–48

Brickhouse N W 1990 Teachers' beliefs about the nature of science and their relationship to classroom practice. *Journal of Teacher Education* **41**: 53–62

Bromme R 1987 Teachers' assessment of students' difficulties and progress in understanding in the classroom. In: Calderhead J (ed.) *Exploring Teachers' Thinking*. Cassell, London, Chap. 5, pp. 125–46

Bromme R 1997a The teacher as an expert: Everyday concepts and some facts about the knowledge base of the teaching profession. In: Bünder W, Rebel K (eds.) *Teacher Education—Theoretical Requirements and Professional Reality*. IPN, Kiel, Germany, Vol. 1, Chap. 1, pp. 23–40

Bromme R 1997b Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. [Competencies, functions and instructional activities of teachers.] In: Weinert F E (ed.) *Psychologie des Unterrichts und der Schule*. [*The Psychology of Schooling*.] *Enzyklopädie der Psychologie*. [*Encyclopedia of Psychology*.]. Hogrefe, Göttingen, Germany, Vol. 3, Chap. 4, pp. 177–212

Carter K, Sabers D, Cushing K, Pinnegar P, Berliner D 1987 Processing and using information about students: A study of expert, novice and postulant teachers. *Teaching and Teacher Education* **3**: 147–57

de Jong T, Ferguson-Hessler M G M 1996 Types and qualities of knowledge. *Educational Psychologist* **31**: 105–13

Druva C A, Anderson R D 1983 Science teacher characteristics by teacher behavior and by student outcome. A meta-analysis of research. *Journal of Research in Science Teaching* **20**: 467–79

Ericsson K A, Smith J (eds.) 1991 *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press, Cambridge, UK

Leinhardt G, Greeno J 1986 The cognitive skill of teaching. *Journal of Educational Psychology* **77**: 247–71

Leinhardt G, Smith D 1985 Expertise in mathematics instruction: Subject matter knowledge. *Journal of Educational Psychology* **77**: 247–71

Putnam R T 1987 Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal* **24**: 13–48

Putnam R T, Borko H 2000 What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher* **29**: 4–16

Rich Y 1993 Stability and change in teacher expertise. *Teaching & Teacher Education* **9**: 137–46

Shulman L S 1986 Those who understand: Knowledge growth in teaching. *Educational Researcher* **15**: 4–14

Stein M, Baxter J, Leinhardt G 1990 Subject-matter knowledge and elementary instruction: A case from functions and graphing. *American Educational Research Journal* **27**: 639–63

Weinert F E, Helmke A, Schrader F-W 1992 Research on the model teacher and the teaching model: Theoretical contradiction or conglutination. In: Oser F, Dick A, Patry J L (eds.) *Effective and Responsible Teaching: The New Synthesis*. Jossey-Bass, New York, Chap. 16, pp. 249–60

Weinert F E, Schrader F-W, Helmke A 1989 Quality of instruction and achievement outcomes. *International Journal of Educational Research* **13**: 895–914

R. Bromme

Teacher Stress and Burnout: Methodological Perspectives

1. Teacher Stress

Stress researchers regularly lament the lack of agreement in the field about what stress is and how it should be measured. The absence of a shared conceptual framework is especially evident in the study of teacher stress and has led to tremendous heterogeneity with respect to the measurement strategies used to operationalize the construct. Some investigators have focused on organizational pressures (e.g., unavailability of needed resources, workload, lack of control), others have examined the effects of interpersonal conflicts, and still others have studied teacher stress in terms of occupational dissatisfaction. In many cases, investigators have used idiosyncratic measures of teacher stress constructed specifically for their study, rather than established instruments. This practice needs to be abandoned as it introduces an undesirable source of variance that further obfuscates an already murky picture.

2. Teacher Burnout

The term 'burnout' was first introduced by Freudenberger (1974), who described it as a frequent and serious occupational hazard among professionals working under difficult conditions in alternative insti-

tutions (e.g., free clinics, halfway houses). Despite important methodological difficulties, to be discussed later, the study of teacher burnout has not been plagued by the conceptual and measurement problems that are endemic to the teacher stress literature. To a large extent, this can be attributed to the fact that investigations of teacher burnout have generally adopted the tripartite conceptualization of burnout offered by Maslach and Jackson (1981). According to this model, emotional exhaustion, depersonalization, and reduced personal achievement are the three key components of the burnout syndrome. Widespread acceptance of this formulation has resulted in the virtually uncontrasted choice of the Maslach Burnout Inventory (MBI), an instrument developed to assess the three constituents of burnout.

It is important to note that proper evaluation of the teacher stress and burnout literature is impeded by a pervasive tendency to use those terms interchangeably rather than as separate constructs. This confusion becomes especially evident when one considers the sizeable literature on the relationship between stress–burnout and health in teachers. In some studies stress and/or burnout are investigated as predictors of ill health, in other studies they are used as outcome measures, and in yet other studies they are treated as both. Although stress can be profitably studied as a possible antecedent of poor health, the same cannot be said for burnout. Burnout is generally conceptualized as a coping failure, an outcome that results from one's inability to deal with prolonged occupational stress. The MBI, in fact, is replete with items describing physical and emotional symptoms. Viewed in this way, what sense does it make to ask, for example, whether feeling fatigued and used up accurately predict physical exhaustion?

3. *Teacher Stress, Burnout, and Health*

Much of the current interest in the study of teacher stress and burnout was sparked by the idea that pathogenic occupational factors might compromise the health of teachers. The question of whether job stress is causally linked to poor health outcomes in teachers is obviously timely. Should the empirical evidence point to such a link, our educational system would benefit from worksite intervention and prevention programs. A recent review (Guglielmi and Tatrow 1998) identified 40 empirical investigations of the health effects of teacher stress and burnout. Approximately half of those studies were conducted outside the USA in such countries as the UK, Finland, Canada, Australia, Israel, Sweden, and Northern Ireland. What does this abundant cross-cultural literature indicate? At first glance, the findings would appear to support the presence of an association between stress–burnout and ill health in teachers. Guglielmi and Tatrow's critical analysis of this literature, however, makes it abundantly clear that a definitive answer is not yet possible. With respect to conceptual rigor and methodological sophistication, the study of teacher stress, burnout, and health is still in its infancy. It is useful to outline the numerous serious problems that characterize this area of research because the same inadequacies are ubiquitously present in the teacher stress and burnout literature as a whole. In the following sections, these difficulties are identified and recommendations for future research are offered.

3.1 *Inadequate Measures and the Problem of Self-report*

As indicated earlier, a lamentable measurement anarchy pervades the teacher stress literature. In the specific area of teacher stress and health, well over half of the published studies have used 'homegrown' stress measures whose psychometric properties are unknown. Such practice raises grave concerns about the validity and reliability of measurement, makes it difficult to compare findings across studies, and will ultimately thwart the comparative assessment of occupational stress associated with teaching in relation to other professions. Another important problem centers around the virtually exclusive reliance on self-report as a way to gather information about stress, burnout, and health. Self-report measures are notoriously vulnerable to selective distortion by self-presentational biases which seriously threaten the validity of measurement. Moreover, when the predictor measure (stress or burnout) and the criterion measure (health status) are both assessed by the same method (self-report), a problem known as shared method variance might spuriously produce statistically high but conceptually trivial correlations that, then, can be erroneously interpreted as meaningful and as indicating a causal association between predictor and criterion (Kasl 1978). Independent objective indicators of stress (e.g., teaching load, class size, availability of resources) need to be added to the subjective measures. The same applies to the assessment of the criterion; physiological indicators, behavioral measures, and objectively established clinical endpoints need to be considered in addition to self-reported health status.

3.2 *Unwarranted Causal Inference*

Drawing unwarranted causal inferences from correlational investigations obviously leads to erroneous conclusions about the relationship between teacher stress or burnout and health. Several possible obstacles that make causal inferences unwarranted need to be considered.

3.2.1 Reverse causation. The stress and burnout reported by teachers could be the consequence,

rather than the cause, of ill health. It has been noted repeatedly that people's self-reports of past events or experiences (e.g., stress) are distorted by a type of recall bias known as retrospective contamination. Being ill, in other words, may cause teachers to over report the amount of stress experienced as they try to find logical antecedents for their poor health (Leventhal and Tomarken 1987).

3.2.2 Construct overlap. If the presence of somatic symptoms is considered a defining component of the stress and burnout constructs, a high (but again meaningless) correlation between ill health and stress–burnout is guaranteed. As indicated earlier, for example, the most commonly used measures of burnout, the MBI and the Burnout Measure, contain many physical exhaustion items. If a burned out person is, by definition, also ill, there is obviously little value in demonstrating that this individual suffers from some physical symptoms.

3.2.3 Can the cause follow the effect? If stress or burnout are causes of ill health, they must precede the health problems. In the vast majority of cases, the published literature makes it impossible to establish whether the required temporal sequence of events is satisfied because no attempt is made to determine whether the physical symptoms occurred before or after the research participants started teaching.

3.2.4 Third variables. Although statistically significant correlations between stress–burnout and ill health are often interpreted in causal terms, it is certainly possible that no causal link actually exists and that a third unidentified variable is the cause of both. Psychological disturbance or personal difficulties unrelated to the job, for instance, could produce both a reduced tolerance to occupational demands and a heightened vulnerability to illness (or perhaps a tendency to exaggerate one's physical problems).

Fundamentally, whether or not causal links can be established and whether or not the direction of causal relationships can be clarified are research design issues. The methodological shift advocated in the next section will hopefully go a long way toward dealing effectively with these causal inference problems.

3.3 Inappropriate Research Design and Lack of Attention to Individual Difference Variables

The importance of individual characteristics in mediating or moderating the impact of stressors has long been acknowledged in the occupational stress literature and, more generally, in the stress research field. The teacher stress and burnout literature stands out as a glaring exception in its insensitivity to this issue. Variables that have been repeatedly found to influence the impact of occupational stressors include personality characteristics (e.g., hardiness, Type A behavior pattern in its old and new formulations), gender and other demographic characteristics (e.g., age, race), the availability of supportive networks, as well as other individual factors (e.g., cognitive appraisal). Against this backdrop of complex relationships and interactions, the teacher stress literature has typically failed to adopt appropriate multivariate strategies and has preferred instead to focus on simplistic bivariate relationships between, for example, teacher stress on one side and job satisfaction, or psychological dysfunction, or physical symptoms, or stress reduction approaches on the other side. Cross-sectional investigations of bivariate relationships are unlikely to enhance our understanding of the individual and interactive contribution of multiple variables. More sophisticated research designs, better equipped to test multifactorial models of teacher stress, are urgently needed (Kasl 1987).

The investigation of teacher stress and burnout would benefit from a shift to carefully designed longitudinal studies, in which the predictor is assessed before the criterion. When appropriate multivariate data–analytic strategies are applied (e.g., hierarchical regression, crosslagged panel correlation, structural equation modeling), these research designs can be used to test the fit of causal models of teacher stress. They hold the potential to answer questions about both the direction of causality and about the interactive contribution of multiple factors (Frese and Zapf 1988).

3.4 Empiricism in Need of a Theory

From a conceptual and methodological perspective, the overall quality of the teacher stress and burnout literature trails considerably behind its parent occupational stress literature. Over the last 20 years, several heuristically fruitful theoretical formulations (e.g., the Person–Environment Fit Model, the Demand–Control Model, the Effort–Reward Model) have guided empirical contributions to the general area of occupational stress. Despite the demonstrated value of those theoretical perspectives, the teacher stress–burnout research has remained unfortunately impervious to their influence. The typical teacher stress–burnout study is essentially atheoretical. Dozens of measures are intercorrelated and *post hoc* explanations of statistically significant associations are offered without an effort to interpret those findings within a theoretically meaningful framework. The adoption of a theoretical framework would help organize research findings across investigations, would introduce consistency in the way constructs are operationalized and assessed and, most importantly,

would point to focused research hypotheses to be tested empirically.

In the occupational health literature, for example, the Demand–Control Model has emerged as a particularly influential theoretical perspective which has stimulated a tremendous amount of work, especially in Europe. According to this model, job demands (e.g., workload, deadlines) and decision latitude (e.g., freedom and control) are the key factors that increase and decrease, respectively, job strain (Karasek and Theorell 1990). On the basis of this model, focused hypotheses can be formulated concerning the conditions that are likely to exacerbate or to alleviate occupational stress. The predictive value of the Demand–Control Model has been empirically established in a variety of work settings (Kristensen 1996, Theorell and Karasek 1996); yet, the study of teacher stress has not been inspired by it. The futility and sterility that characterize much of the current research efforts in the field of teacher stress and burnout are especially disconcerting when one considers the ultimate importance of this work. The field is ready for a shift from atheoretical, descriptive, and cross-sectional studies to theory-based prospective evaluations of causal models of teacher stress and burnout. Until this happens, stress-reduction interventions targeted at teachers are obviously premature.

See also: Burnout, Psychology of; Stress at Work; Stress: Measurement by Self-report and Interview; Stress: Psychological Perspectives

Bibliography

Frese M, Zapf D 1988 Methodological issues in the study of work stress: Objective vs. subjective measurement of work stress and the question of longitudinal studies. In: Cooper C L, Payne R (eds.) *Causes, Coping and Consequences of Stress at Work*. Wiley & Sons, Chichester, UK, pp. 375–411

Freudenberger H J 1974 Staff burnout. *Journal of Social Issues* **1**: 159–64

Guglielmi R S, Tatrow K 1998 Occupational stress, burnout, and health in teachers: A methodological and theoretical analysis. *Review of Educational Research* **68**: 61–9

Karasek R A, Theorell T 1990 *Healthy Work*. Basic Books, New York

Kasl S V 1978 Epidemiological contributions to the study of work stress. In: Cooper C L, Payne R (eds.) *Stress at Work*. Wiley & Sons, Chichester, UK, pp. 3–48

Kasl S V 1987 Methodologies in stress and health: Past difficulties, present dilemmas, future directions. In: Kasl S V, Cooper C L (eds.) *Stress and Health: Issues in Research Methodology*. Wiley & Sons, Chichester, UK, pp. 307–18

Kristensen T S 1996 Job stress and cardiovascular disease: A theoretic critical review. *Journal of Occupational Health Psychology* **1**: 246–260

Leventhal H, Tomarken A 1987 Stress and illness: Perspectives from health psychology. In: Kasl S V, Cooper C L (eds.) *Stress and Health: Issues in Research Methodology*. Wiley & Sons, Chichester, UK, pp. 27–55

Maslach C, Jackson S E 1981 The measurement of experienced burnout. *Journal of Occupational Behavior* **2**: 99–113

Theorell T, Karasek R A 1996 Current issues relating to psychosocial job strain and cardiovascular disease research. *Journal of Occupational Health Psychology* **1**: 9–26

R. S. Guglielmi

Teaching and Learning in the Classroom

The typical image of a classroom is misleading. Classrooms are not neat, orderly places with a teacher standing in front of about 25 students listening to what the teacher is saying or working on a well-defined task. Even in quiet, well-behaved classrooms, this image fails to capture the complex, multifaceted nature of classroom teaching and learning.

Classroom life is an active, dynamic, and social experience. Cognitive, cultural, social, affective, emotional, motivational, and curricular issues occur simultaneously. Teachers and students work together in a rich psychological soup, the ingredients of which often are not apparent to a casual observer. Friends, strangers, adversaries, and dating partners (current, future, and imagined) find themselves together for long periods of time. Many agendas are operating in addition to learning the desired curriculum.

Teaching and learning in a classroom occur on many levels at the same time. Although cognitive and curricular factors typically receive the most attention, learning and understanding do not occur in isolation. Nor is cognition the only process influencing what students learn. It is impossible for students to acquire subject-matter knowledge without simultaneously acquiring or reaffirming a variety of non-cognitive outcomes associated with his or her prior knowledge, affective predispositions, and the instructional experience within which the learning occurs.

A student does not merely learn facts about World War II, for example. He or she also learns whether the study of history is enjoyable or boring. Attitudes about Europeans, Russians, Italians, Japanese, and Americans are acquired or reaffirmed, as are the extent to which certain individuals can be trusted and the horrors (or glory) of war. Students also discover whether he or she is a good student with good ideas and the ability to learn history (or school subjects in general). These outcomes are all part of what students learn, *whether or not these issues are an explicit part of the lesson*. Contrary to popular belief, it is impossible to present a lesson in such a way that only cognitive information is acquired.

In their attempt to understand and explain teaching and learning, researchers commonly distinguish among cognitive, affective, social, communicative, and

curricular factors. Distinctions also are made between the external world of the classroom (the environment) and the internal world of the student (their mental processes, prior experiences, and present interpretations). Students, however, do not engage in this type of analysis as they encounter all of these things simultaneously. The above distinctions mean little if anything to them, as they seek to make sense of what they are experiencing and learning (Nuthall 1999). Thus, it is helpful to consider teaching and learning from the perspectives of both external analysis and the internal experience of students.

1. Classroom Characteristics

A variety of classroom characteristics have been suggested in the research literature (Shuell 1996). Several of these were identified in the preceding paragraphs, including the reality that many things occur simultaneously on different levels. Other characteristics include the following.

(a) *Immediacy*. Things happen at a rapid, come-and-go pace with little time for reflection.

(b) *Unpredictability*. Given the many factors that operate during social interactions in a classroom, it is difficult if not impossible to predict how a particular activity will play out, and unexpected turns of events are common.

(c) *Publicness*. Classrooms are public places with most events witnessed by a large portion of the class. Most students are aware of how the teacher treats or reacts to a particular student and how other students react to one another and the teacher.

(d) *History*. Any group that remains together for a period of time develops a history of common experiences, norms, and routines that influence the way they perceive and relate to one another. History is a characteristic of all interpersonal relationships, and it is an important determiner of how people relate to one another as relationships mature. The group history of a class is an important aspect of classroom teaching and learning.

(e) *Academic goals*. Every classroom strives to achieve certain academic goals. Some of these goals are explicit while others are implicit. Although these goals differ considerably among classrooms, every classroom has academic goals that students are expected to achieve. The way in which these goals play out, however, is influenced by differing perceptions among students and between teachers on the nature of academic tasks and the respective role of teacher and students.

(f) *Social norms*. Social norms exist in every classroom. These norms dictate appropriate behaviors for teacher–student and student–student interactions *in that classroom*. 'Appropriate' does not necessarily refer to any particular set of outside guidelines. For instance, the social norms of a particular classroom may permit students to talk back to the teacher or to use profanity.

(g) *Student characteristics*. Students differ in many ways. The unique combination of student characteristics in a particular classroom helps define how the teaching–learning process plays out in that classroom. Students' prior knowledge, prior experiences, affective predispositions (attitudes), and values, as well as the range of characteristics that exist in a particular classroom, are important factors influencing classroom life.

(h) *Philosophy of education*. Some philosophy of education and/or theory of student learning is inherent in every classroom (this reality is discussed at greater length in the next section). Classroom participants (including the teacher and principal) often are not aware of the particular theory on which their activities are based. In a few cases, however, a group of people will work to establish a school or classroom based on a particular theoretical or philosophical perspective, e.g., student centered, back to basics, and case-based learning, to name only a few.

2. Inherent Philosophy/Theory of Education

No two classrooms are the same. There is considerable variation in the social organization, group norms, and specific activities that occur in a particular classroom. There are many reasons for these differences, but one important reason is that the factors identified in the preceding section are held together and orchestrated by different philosophies, beliefs, and theories of education.

Every classroom is based on some theory of teaching and learning. The types of learning activities, instructional practices, and assessment procedures differ, depending on the particular theory being followed (see *Learning Theories and Educational Paradigms* for examples of different theoretical approaches to learning). This theoretical foundation usually is implicit, however, and the participants are not aware that their activities and interactions reflect a particular philosophy of education. Some of these philosophical differences have endured for centuries, and the debate will undoubtedly continue, for they represent fundamental differences in how people view humankind. For example, the 'factory model' of schooling that dominated education in the United States and parts of Europe for many years is based on production and management procedures proven successful during the industrial revolution. This perspective on teaching and learning stands in sharp contrast to the voices of Thoreau, Dewey, and others who advocate discovery, social reform, and freedom as the appropriate means of education (Shuell 1993).

Thus, depending on the theory operating in the classroom, students may engage in memorizing factual information or engage in activities designed to produce

meaningful learning. These different philosophies are evident in the many metaphors used to characterize teaching and learning, e.g., 'ship captain,' 'midwife,' and 'learning community.' Different models of teaching serve as organizational structures for classroom activities and procedures (Joyce et al. 1992).

These different conceptions of teaching and learning influence both the research that is conducted and the practices, procedures, and activities that are implemented in the classroom (Shuell 1996, Nuthall 1997). One conception may need a relatively simple theory to explain what occurs in a classroom and to establish appropriate practices and activities, e.g., having students rehearse factual information so they are able to reproduce it on demand. Another conception may require a fairly complex theory to accomplish the same purpose, e.g., having students explain to one another the implications of the material they are studying.

Several different perspectives on classroom teaching and learning are clearly evident in the current literature. Each perspective typically employs different theoretical models of learning and assesses students using different criteria and different types of assessment instruments. Four common ones are discussed below (more detailed discussions are available in Shuell (1996), Nuthall (1997), and see *Learning Theories and Educational Paradigms*).

2.1 Behavioral Perspectives

Although behavioral conceptions of teaching and learning are not popular in the current education literature, many classrooms are still based on these traditional theories. In these classrooms, the curriculum generally consists of factual information and methods for solving well-defined problems. Learning occurs, according to this perspective, as a result of reinforced practice of predefined material for the purpose of reproducing this information or skill when needed. These classrooms can be recognized from activities that stress factual information (often isolated from other information), repetition (in one form or another), and correct answers.

2.2 Cognitive Constructivist Perspectives

Cognitive constructivist conceptions of teaching and learning focus on students understanding the material being studied rather than on the acquisition of factual information and skills; that is, on mental processes rather than behavior. Knowledge is not seen as a commodity passed from teacher or textbook to students' minds. Rather, this perspective believes that students construct their own understandings as they seek to make sense of the material and the classroom experiences in which it is embedded. Learning is seen as the creation and reorganization of students' conceptual structures (knowledge) as a result of cognitive processing. These classrooms can be recognized by students engaged in sense-making activities and authentic (real-life) learning tasks.

2.3 Sociocultural and Community-focused Perspectives

Sociocultural perspectives see teaching and learning as a social rather than cognitive activity. Based on the work of Russian psychologist Lev Vygotsky (1978) and anthropologists such as Jean Lave (Lave 1988, Lave and Wenger 1991) and Barbara Rogoff (Rogoff 1990, Rogoff et al. 1996), this perspective believes that focusing on individual learners is a mistake, in sharp contrast with behavioral and cognitive views. Acquiring culturally relevant knowledge is a process in which the learner becomes a member of a community of practice or community of learners (Brown and Campione 1996, Rogoff et al. 1996). An apprenticeship model of teaching and learning is often employed (Shuell 1996). Classrooms representing this perspective are recognizable by students working on group projects with individual students working on interrelated subtasks. The ability of individuals to work together in negotiating solutions to problems and conceptual understandings is stressed, with the teacher supporting and helping students, as appropriate, to accomplish these goals.

2.4 Linguistic Perspective

The linguistic perspective is similar to the sociocultural perspective in that language is seen as a cultural artifact. It differs from sociocultural views, however, in that language is seen as playing a far more central role in shaping the social interactions that occur in the classroom. As language communities, classrooms develop their own forms of communication. The negotiations that occur as part of this process create opportunities to learn and give birth to the knowledge and meanings that comprise the curriculum. Narratives and stories, rather than cognitive schemata or social participation, are seen as the primary way in which students 'process, come to understand, and remember experiences' (Nuthall 1997, p. 713). Classrooms based on this perspective can be recognized by the instructional use of student–student and teacher–student discussions in which students explain and negotiate the meaning of the topics being studied.

3. Instructional Experiences in the Classroom

Within the complex social environment of the classroom, teachers and students engage in a variety of activities as they work together to accomplish tasks

with academic, social, and personal consequences. Schools and their classrooms, of course, are organized in many different ways. The classroom community may be one that focuses on intellectual activities and in which the members generally enjoy one another's company. Or the community, not necessarily by design, may focus primarily on social relationships (positive or negative) and personal agendas. Whatever the environment, it sets the stage for the teaching and learning that occur in the classroom.

Classroom teaching and learning, however, are not the unidirectional process they often are thought to be. The teacher's philosophical and theoretical beliefs clearly affect what happens in the classroom, but the teacher is not the only one in the classroom determining what occurs. Students, as well as teachers, hold personal theories of learning and teaching that influence the instructional tone of the classroom. A teacher may believe, for example, that learning results from problem solving and critical thinking and provides activities consistent with this belief. The students, however, may believe (perhaps encouraged by parents and previous classroom experiences) that learning is a process in which information is remembered (memorized) well enough to be reproduced in a test. (The opposite scenario, of course, is just as plausible.) Thus, the instructional process that actually transpires in the classroom must either be negotiated among the participants or dictated by the teacher. In either case, social as well as intellectual consequences are involved.

Given the complexities that exist in classrooms, not all students in the same classroom learn the same thing. Even when students engage in the same activities, they learn different things in different ways, and they learn more than they are taught (Nuthall 1999).

4. *The Ideal Classroom?*

Some people hold a myth that there is a single best way to organize and run a classroom. There are many ways in which classrooms can be compared, but in order to judge one classroom as being better than another, agreement must exist on the criteria used to determine 'best,' i.e., best for what? Given the diversity of perspectives and philosophies discussed above, it is not surprising that different people insist on different criteria.

An effective classroom is one that achieves its goals. When schools or classrooms have different goals, comparisons regarding which set of outcomes is 'best' becomes a philosophical discussion. Differences in desired outcomes can occur even when the curriculum appears to be the same, e.g., a particular school or teacher may implement the same formal curriculum (either intentionally or unintentionally) by having students learn factual knowledge or by engaging them in critical thinking. Different assessment practices and procedures can also result in different interpretations of the formal curriculum.

Different instructional methods are appropriate for accomplishing different goals. Educational research and theory can be of assistance in identifying appropriate classroom activities, once desired outcomes are identified. The latter, however, is largely a philosophical issue. Regardless of the goals and curricula pursued, there is greater likelihood those goals will be attained if the teacher and students consider them to be important and adopt them as their own personal goals. Goals and tasks imposed on someone by others usually are pursued with little enthusiasm.

5. *Summary*

Classroom teaching and learning is a complex, multilayered, and social experience. Many things happen at the same time, and the students and teacher must interpret and process these events with considerable rapidity. Teaching is not a one-way transmission of information, for students influence the process as much as the teacher. The teacher plays an extremely important role in establishing a meaningful classroom environment for students, but in the final analysis, it is the social and psychological activities and responses of each student that determines what he or she learns.

The way in which each student perceives, interprets, processes, and understands classroom activities—not what the teacher does—is the single most important factor in determining the educational outcomes acquired by that student. Teachers and outside observers must be careful when making inferences from observations of student performance (either formal or informal). The students' perceptions, interpretations, and understandings may be very different from what they appear to an external observer. A student's performance on an assessment instrument, for example, may reflect something rather different from the teacher's interpretation, for there are many ways for a student to answer questions and solve problems.

Given the complexity that exists in classrooms, different theoretical perspectives and philosophies of education are common. These different perspectives add richness to our understanding of classroom teaching and learning, but they belie a belief in an ideal classroom experience appropriate for all students. Fundamentally different philosophical beliefs about education and human existence lead to fundamentally different criteria for both the curriculum and classroom practices. Some of these differences have existed for centuries, and the debate is likely to continue. Nevertheless, our understanding of the dynamics and complexity of classroom teaching and learning has grown considerably in recent years.

See also: Classroom Assessment; Classroom Climate; Cooperative Learning in Schools; Environments for Learning; Group Processes in the Classroom; Open Classroom: Learning and Teaching; Teacher Expertise

Bibliography

Brown A L, Campione J C 1996 Psychological theory and the design of innovative learning environments: on procedures, principles, and systems. In: Schauble L, Glaser R (eds.) *Innovations in Learning: New Environments for Education*. Lawrence Erlbaum Associates, Mahwah, NJ. pp. 289–325

Joyce B, Weil M, Showers B 1992 *Models of Teaching*, 4th edn. Allyn and Bacon, Boston

Lave J 1988 *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life*. Cambridge University Press, Cambridge, UK

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK

Nuthall G 1997 Understanding student thinking and learning in the classroom. In: Biddle B J, Good T L, Goodson I F (eds.) *International Handbook of Teachers and Teaching*. Kluwer, Dordrecht, The Netherlands, Vol. II, pp. 681–768

Nuthall G 1999 The way students learn: acquiring knowledge from an integrated science and social studies unit. *Elementary School Journal*. **99**: 303–41

Rogoff B 1990 *Apprenticeship in Thinking: Cognitive Development in Social Context*. Oxford University Press, New York

Rogoff B, Matusov E, White C 1996 Models of teaching and learning: participation in a community of learners. In: Olson D R, Torrance N (eds.) *The Handbook of Education and Human Development*. Blackwell, Oxford, UK, pp. 388–414

Shuell T J 1993 Toward an integrated theory of teaching and learning. *Educational Psychologist* **28**: 291–311

Shuell T J 1996 Teaching and learning in a classroom context. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Simon and Schuster Macmillan, New York, pp. 726–64

Vygotsky L S 1978 *Mind in Society: the Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

T. J. Shuell

Teaching as a Profession

This entry initially outlines the contested nature of the concept profession. There then follows an examination of how professionalization has been a key project of teaching as an organized occupation and the implications of this project for status and power. From the mid-1970s, teaching as a profession began to undergo fundamental changes and these are viewed from the perspectives of proletarianization and deprofessionalization. Finally, the worldwide 'educational reform movement' is shown as predicated on conceptualizations of the new professionalism.

1. Profession

Profession is a much-disputed sociological concept, and the existence of criteria which allegedly distinguish professions from other occupations has been strongly contested. Yet the concept has not fallen into desuetude and continues to constitute the focus of lively debate (see Freidson 1986, for a discussion of the concept profession). The critique of the idea of profession was initially a reaction to a functionalist perspective which dominated much of the earlier theorizing. This depicted professionals as exercising knowledge-based skills as autonomous practitioners in nonroutine contexts according to a client-focused ethic guaranteed by a self-regulative professional body. From this theory were derived sets of distinguishing criteria which Millerson (1964) showed to be highly diverse. However, despite this diversity, some criteria were common to most lists. These core criteria were exemplified in the list proposed by Lieberman (1956) in his early study of teaching as a profession: the provision of an essential social service, intellectual techniques, lengthy specialized training, practitioner and group autonomy, personal responsibility for judgments, an emphasis on service, and a self-governing body.

Critics of the functionalist perspective argue that the idea of a profession, and the alleged distinguishing characteristics, is less an analysis of practice than the theorization of a self-interested ideology developed historically by a small group of elite occupations characterized by independent practitioners (e.g. lawyers, doctors) offering a service for fee. An alternative power perspective presents these occupations as having succeeded over time in gaining control over a market for a service, achieving responsibility for bestowing the license to practice through a self-regulating body, strengthening the occupational boundary through credentialism and social closure and using increments of power to enhance levels of status and remuneration.

Despite the power critique that demonstrates the ideological nature of the idea of a profession and of the alleged criteria, profession retains a heuristic potential in the exploration of issues of autonomy, control, knowledge, practice, etc. This heuristic deployment entails treating profession as a particularistic phenomenon the social construction of which varies over time and space (see Ginsburg 1997 for a wide-ranging review). In this connection it can be noted that many of the key issues addressed in the sociology of the professions are different in the Anglophone countries than in continental Europe where the state is more involved in training and employment, and where status is derived more from the institutions of higher edu-

cation which practitioners attended than from market position. The heuristic deployment of profession also entails an emphasis on professionals—their work relations and practices, their lives and careers (see the study of school teaching as a profession by Lortie 1975). In this entry professionalization provides the major organizing theme.

2. Meanings of Professionalization

Professionalization is a process whereby occupations have become, or seek to become, publicly recognized as professions according to the degree to which they meet the alleged criteria. Professionalization can be seen as having two strands. One strand is concerned with the improvement of status. The other strand is concerned with the improvement of the capacity of members to enhance the quality of service which is provided. It is generally assumed that these two elements proceed *pari passu* but this need not be the case. The terminology in relation to this distinction is confusing, however, recent usage distinguishes between professionalization as the pursuit of status and professionalism as the improvement of skills and hence service. The latter terminology will therefore be used here.

One can perhaps refer to 'early' professionalization as the process whereby the elite professions established their position and 'late' professionalization as the period beginning in the late nineteenth century when with 'the rise of professional society' (Perkin 1989) a large number of occupations, including teaching, began their quest for professional status by taking measures to meet the alleged criteria. In these terms, it can be said that teaching was in the process of professionalization from the mid-1870s to the mid-1970s, increasingly so after World War II. The length of the period of teacher education was extended. An increasingly higher proportion of teachers were graduates. Teacher training institutions became integrated into the universities. The knowledge base of teaching became more extensive. Teachers enjoyed a high degree of pedagogical autonomy within a loosely-coupled school structure. Teachers' organizations increased their influence with government. And in some systems, e.g. Malta, teaching achieved the legal status of a profession. However within this general process of professionalization are acute gender differences (Gitlin and Labaree 1996).

3. Professionalization and the Occupational Status of Teachers

Status is multidimensional, and the success of the professionalization project varies across these dimensions. If occupational status denotes the degree to which an occupation meets the alleged criteria of a profession, then the status of teaching in most countries has been increasing. However, this has not been reflected in any obvious improvement in the occupational prestige of teaching, defined here as the relative position of an occupation in a hierarchy of occupations as ranked by samples of members of the public. Teaching is in most countries ranked in the upper quartile of such hierarchies, but below the major professions of medicine, law, etc., the status group of aspiration for teachers who thereby experience a degree of relative deprivation.

International comparisons of occupational prestige (Treiman 1977) show some differences in the relative ranking of the single occupational title of schoolteacher across societies but the data are too sparse to allow generalization and no obvious patterns suggest themselves. Perhaps of greater significance are the quite marked intra-occupational differences in rank, according to age of the students taught and the social and academic character of types of school. Comparative data show that hierarchies of occupational prestige have a high degree of stability—rank correlations over time and space almost invariably exceeding 0.9, which suggests that there is relatively little scope for an established occupation to enhance its prestige to any substantial degree.

It has been suggested that some inherent characteristics of teaching as an occupation are likely to have an inhibiting effect on enhanced prestige. Leggatt (1970) grouped these into three categories, i.e. characteristics of the practitioner group (e.g. size, social class, background of entrants), characteristics of clients and client relations (e.g. children as clients, compulsory and protracted relationships), and characteristics of work performance (e.g. unclear expertise, performance hard to evaluate).

Perhaps of more significance, because potentially more achievable, is the esteem in which teachers are held, defined here as the noncomparative evaluation of an occupation on the basis of such factors as quality of service, client-orientated values and practitioner commitment. Esteem is relatively independent of both status and prestige. In many countries, especially in Asia and Africa, the teacher—as 'guru'—enjoys high esteem but this is not reflected in the hierarchy of occupational prestige. However, esteem is an amorphous phenomenon shaped by various factors, for example, personal experience, image as conveyed by media representation, etc., and is currently under-researched.

4. Professionalization and Power

The professionalization project of teachers in Anglophone countries emerged in the late nineteenth century, coinciding with the growth of mass public education and the formal training of teachers. It was shaped by teachers in the public sector whose salaries

were ultimately paid from public funds. Teachers were thus *de facto* public employees, though the condition of their employment varied between countries. Their collective power was therefore shaped by their relationship with their employer. Two main strategies of professionalization were open to teachers: a professional body strategy, which focused on improving status on the assumption that this would extend to an enhancement in salaries and conditions, and a trade union strategy which put a priority on improving salaries and conditions and thence status. In most systems the trade unionism emerged as the major strategy though this was also accompanied by a 'professional' rhetoric. In the USA this tension is revealed in the contrasting strategies of the National Education Association and the American Federation of Teachers.

A full discussion of teacher unionism is beyond the scope of this entry. However, attention can be drawn to a group of studies which have explored the state–profession relationship in the UK in socio–historical detail (Ozga and Lawn 1981). These studies show that teacher professionalization was the function of a fluid exchange relationship between teacher unions and the state whereby the state conceded a 'legitimated teacher professionalism' in exchange for teachers' acceptance that their legitimate sphere of professional activity was the classroom and not the political arena.

5. *Critics of Professionalization*

The central criticism of professionalization is that it is a project for self-aggrandisement at the expense of client interests. Critics are to be found across the political spectrum, though their analyses are differently nuanced and they propose competing prescriptions. The targets of all critics tend to be the same, namely the 'criteria' which constitute the quest of professionalization. Just two of these will be considered: practitioner autonomy and theoretical knowledge.

One line of criticism is that teachers' protection of their autonomy is a barrier to collaboration between teachers and their immediate clients: students, their families and the community. Ginsburg (1997) has contrasted a 'professionalization' project, which leads to teachers remaining aloof from their immediate clients, with a 'political' project, whereby teachers collaborate with their communities in order to pursue an equitable educational system. Hargreaves (1999) has envisioned a future whereby the teaching profession in partnership with the public becomes 'a vigorous social movement of acting subjects rather than frogmarched individuals' (p. 71). Another form of critique sees teachers' autonomy as a mechanism for avoiding legitimate accountability to a variety of stakeholders, not least the state. The policies which follow from this view entail teachers accepting a managerially led collective accountability for their professional activities. A third form of the critique is that autonomy is a barrier to the increased professionalism of teachers as defined above. Lortie (1975) drew attention to the strong strain of individualism in the culture of teaching and a strong norm of noninterference which he saw as inhibiting teachers from building a strong technical culture.

The knowledge base of teaching has two components: knowledge of subject matter and knowledge of pedagogical, curricular and educational theory. It is the second of these which has been the focus of much criticism, not least from within the profession, as generations of teachers have questioned the relevance of theory. But there is an ambivalence here as teachers' organizations see a strong knowledge base as facilitating a link with the academy which is seen as a source of prestige. However, the likelihood of the full academic acceptance of educational theory has been doubted. The thrust for building a body of educational theory came largely from teacher educators. These criticisms of autonomy and knowledge—and criticisms directed at codes of ethics, self-government, etc.—have, over the past 20 years, fuelled a political reconceptualization of teaching.

6. *Deprofessionalization or 'The New Professionalism?'*

The decline in professional society began in the mid-1970s as politicians in many countries began to question the perceived monopolistic position, producerism and the weak accountability of the professions, together with rising costs of professional services. In education this contributed to a worldwide 'reform movement.' The patterning of this movement varied across countries but some of the main characteristics are as follows: the decentralization of school management and the creation of local markets; the centralization of policies relating to the curriculum, its delivery, its objectives and its measured outcomes; the introduction of a range of accountability procedures nationally and locally; and an increased expectation of the services which the school would deliver.

Three major perspectives on these changes can be identified and discussed: proletarianization, deprofessionalization and 'the new professionalism.'

Proletarianization focuses particularly on the intensification of teachers' work as they are increasingly required to fulfill a wider range of functions, as much managerial as pedagogical. This process is seen as entailing the deskilling of teachers as their range of knowledge and skill is reduced and routinized and a loss of control over their work. The proletarianization position is a component of a social class perspective which is an alternative to the status perspective which dominates much of the literature on the professions (see Robertson 2000, for a post-Fordist analysis of

teachers' work, and Smyth et al. 2000, for a theoretical account and illustrative case study of the labor process approach to teachers' work). The proletarianization perspective is also critical of professionalization as an attempt to enhance teacher autonomy as this is seen as inimical to the interests of students and their families.

The deprofessionalization perspective focuses on the degree to which teaching has regressed from the levels achieved in meeting the traditional criteria. The knowledge base of teaching is seen as being eroded with a move towards competency-based training and school-based professional development. Teacher autonomy is seen as undermined as curricular content and learning objectives become more highly specified. And the impact of teachers' organizations on policy is seen as having been greatly reduced.

'The new professionalism' is an emerging concept and is not readily defined. Essentially it marks a move away from the idea of a profession and the 'status' dimension of professionalization. It implicitly shifts the emphasis to professionalism with its emphasis on skills and service. In political and administrative discourse it marks a growing emphasis on professional rather than profession and can be over-simply summarized as follows: to be 'professional' is to have acquired a set of skills through competency-based training which enables one to deliver efficiently, according to contract, a consumer-led service in compliance with accountability procedures collaboratively implemented and managerially assured.

In educational discourse 'the new professionalism' embraces a number of perspectives which accept, to a greater or lesser degree, the emphasis on professionalism defined as 'commitment to the improvement of service.' Writers who have delineated 'the new professionalism' include Caldwell and Spinks (1998) from the perspective of school management and Hargreaves (1994) from the perspective of teachers' professional development.

7. *Conclusion*

The increasing worldwide political salience of education has led over the past 25 years to changes in the conceptualization of teaching as a profession. In the terminology adopted in this entry, the shift has been from a focus on professionalization to a focus on deprofessionalization and, latterly, to professionalism, i.e. from a concern with status to a concern with the quality of service—without accepting uncritically political and managerial conceptions of quality. In fact, professionalism may entail a resistance to official policies in giving priority to the perceived needs of immediate clients.

The focus of enquiry has also changed. Of growing significance are studies of the teacher's work (e.g. Little and McLaughlin 1993, Helsby 1999) and of teachers' lives and careers (e.g. Day et al. 2000), which are open to the complexities of the situations. An example of this is an openness to the different meanings and practices conceptualized as collaboration or collegiality. These recent studies explore how collegiality can be professionally constraining or liberating, depending on the forms which it takes. It is the exploration of this complexity in local as well as national contexts, and of the relationship between 'the new professionalism' and 'the new managerialism,' that is beginning to characterize the study of teaching as a profession.

See also: Professionalization/Professions in History; Professions, Sociology of; Teacher Education; Teacher Expertise; Teacher Stress and Burnout: Methodological Perspectives; Teaching: Trends in Research

Bibliography

Caldwell B, Spinks J 1998 *Beyond the Self-Managing School*. Falmer Press, London and New York

Day C, Fernandez A, Hauge T E, Moller J (eds.) 2000 *The Life and Work of Teachers: International Perspectives in Changing Times*. Falmer Press, London and New York

Freidson E 1986 *Professional Power: A Study of the Institutionalization of Formal Knowledge*. University of Chicago Press, Chicago

Ginsburg M 1997 Professionalism or politics as a model for teachers' work and lives? *Educational Research Journal* **12**: 1–15

Gitlin A, Labaree D 1996 Historical notes on the barriers to the professionalisation of American teachers: The influence of markets and patriarchy. In: Goodson I F, Hargreaves A (eds.) *Teachers' Professional Lives*. Falmer Press, London and Washington DC, pp. 55–75

Hargreaves A 1999 Reinventing professionalism: Teacher education and teacher development for a changing world. *Asia-Pacific Journal of Teacher Education and Development* **2**: 65–74

Hargreaves D 1994 The new professionalism: The synthesis of professional and institutional development. *Teaching and Teacher Education* **10**: 423–38

Helsby G 1999 *Changing Teachers' Work*. Open University Press, Buckingham and Philadelphia

Leggatt T 1970 Teaching as a profession. In: Jackson J A (ed.) *Professions and Professionalisation*. Cambridge University Press, Cambridge, UK, pp. 155–77

Lieberman M 1956 *Education as a Profession*. Prentice Hall, Englewood Cliffs, NJ

Little J W, McLaughlin M W 1993 *Teachers' Work*. Teachers' College Press, New York

Lortie D 1975 *Schoolteacher: A Sociological Study*. University of Chicago Press, Chicago

Millerson G 1964 *The Qualifying Associations*. Routledge, London

Ozga J, Lawn M 1981 *Teachers, Professionalism and Class*. Falmer Press, Lewes, UK

Perkin H 1989 *The Rise of Professional Society*. Routledge, London

Robertson S L 2000 *A Class Act: Changing Teachers' Work, Globalisation and the State*. Falmer/Garland, London and New York

Smyth J, Dow A, Hattam R, Reid A, Shacklock G 2000 *Teachers' Work in a Globalising Economy*. Falmer Press, London and New York
Treiman D J 1977 *Occupational Prestige in Comparative Perspective*. Wiley, New York

E. Hoyle

Teaching for Thinking

Is it possible to teach thinking skills that improve a person's cognitive functioning? Can a person learn to become a more effective problem solver? In short, can thinking be taught? These are the kinds of questions examined in this article. After an introduction which contains definitions, a rationale, and an example four issues are examined concerning teaching for thinking—what to teach, how to teach, where to teach, and when to teach.

1. Introduction

1.1 Definitions

Teaching for thinking refers to instruction that is intended to improve the effectiveness of a people's thinking. This definition consists of two parts: (a) *teaching*—teaching for thinking involves instruction, and (b) *thinking*—the goal of the instruction is to improve someone's ability to solve problems. *Teaching* refers to manipulations to a learner's environment which are intended to foster a change in the knowledge of the learner, whereas *thinking* refers to cognitive activity aimed at solving a problem. In sum, teaching for thinking involves instructional manipulations that are intended to foster the learner's knowledge of how to solve problems.

Teaching for thinking is a type of cognitive strategy instruction (Halpern 1992, Mayer and Wittrock 1996, Pressley and Woloshyn 1995). In cognitive strategy instruction learners are taught how to control their cognitive processing such as how to think, how to learn, how to remember, or how to motivate oneself. Thus, cognitive strategy instruction involves fostering metacognitive knowledge—knowledge of how one's mind works—and strategic knowledge—knowledge of how to develop plans for accomplishing goals—rather than procedural knowledge—knowledge of how to carry out a routine—and declarative knowledge—factual knowledge about some topic.

1.2 Rationale

Students are expected to be effective thinkers but rarely are taught how to think. Students are expected to be able to solve problems but rarely are given productive experience that would foster the development of problem-solving skills. Although the formal curriculum in any subject area may emphasize the learner's acquisition of content knowledge, an often neglected part of the curriculum concerns the learner's being able to use what was learned to solve problems. Instruction in how to think is sometimes not part of the formal curriculum but rather is part of the hidden curriculum; although unstated it is expected that learners will be able to use what they have learned in problem solving. For these reasons, there has been a growing call for explicit instruction in how to be a more effective problem solver (Halpern 1992, Mayer and Wittrock 1996, Pressley and Woloshyn 1995).

1.3 Example

The Gestalt psychologist, Wertheimer (1959) offers a classic example of the challenges of teaching for thinking. He describes a classroom in which a teacher presents a lesson on how to find the area of a parallelogram. The teacher draws a parallelogram on the chalk board, drops a perpendicular line from the top base to the bottom base (which is called the height), measures the height, measures the base, multiplies height times base and gives the result as the area (using the formula area = height × base). After the demonstration, the teacher calls on a student to solve a similar problem at the board and the student is successful in carrying out each of the steps. Later the students do very well on a quiz in which they must compute the area of parallelograms.

In spite of the apparent success of the lesson, Wertheimer (1959, p. 15) criticized the teaching for its lack of emphasis on thinking: 'Most people would say, "This is an excellent class; the teaching goal has been reached." But observing the class, I feel uneasy, I am troubled. "What have they learned," I ask myself. "Have they done any thinking at all? Have they grasped the issue? Maybe all they have done is little more than blind repetition."' As a test, Wertheimer presents an unusually shaped parallelogram, and the students respond by saying 'We haven't had this yet.'

In contrast to teaching for rote learning, Wertheimer suggests that students should be encouraged to discover that the triangle on one of the parallelograms could be cut off and placed on the other end to form a rectangle. In this way, students could see that a parallelogram is simply a rectangle in disguise. Since they already know how to find the area of a rectangle, the solution to parallelogram problems now becomes much easier. Wertheimer referred to this method of teaching as learning by understanding, and claimed that it resulted in superior problem-solving transfer, that is, in being able to use what was learned in new situations. According to Wertheimer, rote teaching methods lead to reproductive thinking—being able to repeat what was taught—whereas meaningful teaching

methods can lead to productive thinking—being able to use what was taught in new situations.

When the goal is teaching for thinking—that is the promotion of productive thinking—instructors need to consider four issues: what to teach, how to teach, where to teach, and when to teach. In addressing these four issues, Mayer (1997) has compared four common-sense principles of teaching for thinking with corresponding alternative principles based on research in cognitive science.

2. *What to Teach*

The first issue concerns what to teach. Should instruction focus on thinking as a single monolithic ability or as a collection of smaller component skills? The common-sense view casts thinking as a single ability: the ability to solve problems is a unitary skill so teaching for thinking involves improving one's mind in general through learning certain school subjects that require mental discipline and orderly thinking. An implication of the single ability view of what to teach is that the mind is a sort of mental muscle that needs to be exercised to increase its overall functioning. When the mind is improved through learning a school subject that exercises the mental muscle, then the learner will also show improvements in a wide variety of academic tasks.

In the early 1900s the dominant view among educational scholars was the doctrine of formal discipline, namely, the idea that learning certain school subjects, such as Latin, geometry, and logic, promoted proper habits of mind including mental discipline and orderly thinking. One of the first tasks of the newly established field of educational psychology was to test the doctrine of formal discipline in a scientifically rigorous way. For example, in a convincing series of studies, Thorndike and co-workers found that students who had learned Latin performed no better than students who had not learned Latin in learning a new school subject such as bookkeeping (Thorndike 1924). Similarly, more recent claims that learning programming languages such as Logo would improve children's minds were also put to the test; researchers found that students who learned Logo did not outperform non-Logo learners on problem-solving tasks unrelated to Logo (Mayer 1988). Overall, the research evidence failed to support the common-sense view of thinking as a single ability.

In contrast, a common theme in modern research in cognitive science is that a person's performance on any cognitive task—such as solving a word problem or writing an essay—depends on the degree to which the person possesses relevant component skills (Sternberg 1990). To write an essay, for example, students need to engage in three main cognitive processes—planning, translating, and reviewing (Flower and Hayes 1980); to solve an arithmetic story problem students need to engage in four cognitive processes—translating each sentence into a mental representation, integrating the sentences into a coherent mental model of the situation, planning a set of steps to solve the problem, and executing the solution plan (Mayer 1992).

According to the multiple skills view, teaching for thinking should focus on helping students learn the collection of component skills that are relevant for each important cognitive task rather than on improving the mind overall. For example, Pressley and Woloshyn (1995) have shown how focused instruction in component skills can improve students' performance in reading, writing, maths, and other cognitive tasks. In answer to the question of what to teach, their answer is to teach a collection of smaller component skills.

3. *How to Teach*

The second issue concerns how to teach. Should instruction focus on encouraging students to generate correct answers (i.e., the product of problem solving) or on the methods used to solve problems (i.e., the process of problem solving)? The common-sense view favors the product-of-problem-solving approach: as in learning any skill, the best approach to learning of thinking skills is to get plenty of practice in answering questions along with feedback on whether or not the answers are correct. Although drill-and-practice instructional methods can be effective in improving performance on problems like those presented during instruction, they sometimes fail to help students solve transfer problems that go beyond those presented during instruction.

For example, in a classic study, Brownell and Moser (1949) taught students to solve two-column subtraction problems (such as $76-48=$ ____) by drilling them in the steps of the procedure or by showing how the steps correspond to using sticks that bundled into groups of 10 (i.e., by using concrete manipulatives). Although students who learned with concrete manipulatives did not outperform the traditionally taught students on retention problems—two column subtraction problems like those given during instruction—they did excel on transfer problems, such as solving three-column subtraction problems. Brownell and Moser concluded that students understand the material better when the emphasis changes from product to process.

Apprenticeship approaches offer a straightforward way to focus on problem-solving processes because learners attempt to model the cognitive processes of more skilled performers. For example, in an early study, Bloom and Broder (1950) taught unsuccessful students how to answer exam questions by modeling the problem-solving processes of successful students. First, a successful student described his or her thought process while answering an exam question. Second,

the unsuccessful student described his or her thought process while answering the same question, and then the unsuccessful student compared his or her process with that of the successful student. After 10 to 12 sessions with different questions, the unsuccessful students were much better able to answer new exam questions than were a group of matched students who received no training. Bloom and Broder concluded that teaching for thinking is most effective when instruction focuses on the process of problem solving rather than solely on the product of problem solving.

According to the process-of-problem-solving view, teaching for thinking should focus on helping students learn the methods for how to solve problems rather than solely on practice in getting the right answer. Modern research in cognitive science has demonstrated the value of apprenticeship approaches that emphasize the process of problem solving (Collins et al. 1989). Successful techniques include reciprocal teaching of reading comprehension skills (Brown and Palinscar 1989). In answer to the question of how to teach, their answer is to model the process of problem solving.

4. *Where to Teach*

A third issue concerns where to teach. Should instruction be given as a general, independent, stand-alone course or should instruction in thinking be incorporated into specific subject matter areas? The common sense view favors the independent approach: all students should be required to take a course in problem solving, in addition to their regular content courses. An implication is that problem-solving skills are best learned as an isolated subject, separated from the context in which they will be applied.

The Productive Thinking Program was intended as a stand-alone course in problem solving, aimed at helping elementary school students learn to generate and test hypotheses (Covington et al. 1966). It consisted of a series of 15 cartoon-like booklets, each describing a mystery or detective story. The reader was asked to participate in solving the case by generating and testing hypotheses, along with help from various characters in the booklet who modeled appropriate processes for generating and testing hypotheses. Students who took the course tended to outperform matched students who had not taken the course on solving problems like those in the booklets, but not on problems that were from a different context (Mayer 1992). Thus, in spite of attempts to teach context-free problem-solving skills, students tended to perform best in using the skills within the context they were presented. These results tend to refute the common-sense view that teaching for thinking should occur as a separate course.

In contrast, cognitive science research points to the domain-specificity of learning. For example, people were asked to determine which of two products was the best buy, such as a 10-ounce can of peanuts at 90 cents or a 4-ounce can at 45 cents. In school, the most common approach is to compute unit cost, yielding 9 cents per ounce versus 11.25 cents per ounce. However, in a supermarket, people tended to use a ratio approach in which they reasoned that the larger one costs twice as much but contains more than twice as many ounces. Similarly, Nunes et al. (1993) found that students who could solve arithmetic problems without error in the context of their job as street vendors made many errors when confronted with the same problems in a school setting. In addition, research on expertise demonstrates that people who are expert problem-solvers in one domain (such as physics or chess) are generally not superior in problem solving in other domains (Mayer 1992). These results support the idea that the usability of problem-solving strategies depend on the context in which they are learned.

According to the specific-domain view, teaching for thinking should take place within a specific context rather than in a context-independent environment. Research on the lack of transfer of school-learned mathematics (Nunes et al. 1993) is consistent with the specific-domain view. In answer to the question of where to teach, their answer is to teach cognitive skills within the specific context they will be applied.

5. *When to Teach*

The final issue concerns when to teach. Should instruction focus on teaching higher-order thinking skills only after the learner has mastered lower-level basic skills or should instruction on thinking begin when the learner is still a novice? The common-sense view favors the prior-mastery approach: students cannot understand higher-order skills until they have mastered lower-level ones.

However, cognitive science research on teaching reading, writing, and mathematics shows that students can benefit from learning higher-order thinking skills before they have completely mastered the basics. In reading, for example, elementary school students can benefit from learning how to summarize a passage, how to predict what will come next, and how to recognize portions that need clarification even if they are still not fluent in basic decoding skills (Brown and Palinscar 1989). In writing, students can benefit from writing creative essays even before they have mastered spelling, grammar, and punctuation (Kellogg 1994). In mathematics, students can benefit from direct instruction in how to represent word problems even before they have completely automatized their computational fluency (Sternberg and Ben-Zeev 1996). In contrast to the prior-mastery view, this research provides support for an approach that can be called constraint removal—providing scaffolding support that enables students to work on tasks they have not

yet mastered (such as allowing a student to dictate an essay).

According to the constraint-removal approach, teaching for thinking can begin while the learner is still a novice. In answer to the question of when to teach, supporters of this approach argue that it makes no sense to wait until the student has automated all lower level skills.

6. Summary

In summary, teaching for thinking involves helping students to be able to use what they learn to solve new problems. In contrast to common-sense views, teaching for thinking is likely to be most successful when the focus is on (a) teaching thinking as a collection of smaller component skills rather than as a single monolithic ability, (b) teaching the process of how to solve problems rather than solely the product of problem solving, (c) teaching thinking within a specific subject domain rather than as an isolated domain-free course, and (d) teaching while students are novices rather than waiting until they have mastered all basic skills.

See also: Competencies and Key Competencies: Educational Perspective; Critical Thinking, Cognitive Psychology of; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Learning Theories and Educational Paradigms; Metacognitive Development: Educational Implications; Problem Solving and Reasoning, Psychology of; School Learning for Transfer; Transfer of Learning, Cognitive Psychology of

Bibliography

Bloom B S, Broder L J 1950 *Problem-solving Processes of College Students*. University of Chicago Press, Chicago

Brown A L, Palinscar A S 1989 Guided, cooperative learning and individual knowledge acquisition. In: Resnick L B (ed.) *Knowing, Learning, and Instruction*. Erlbaum, Mahwah, NJ, pp. 393–451

Brownell W A, Moser H E 1949 *Meaningful versus Mechanical Learning: A Study in Grade* 3 *Arithmetic. Duke University Research Series in Education* (No. 8). Duke University Press, Durham, NC

Collins A, Brown A L, Newman S E 1989 Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In: Resnick L B (ed.) *Knowing, Learning, and Instruction*. Erlbaum, Mahwah, NJ, pp. 453–94

Covington M V, Crutchfield R S, Davies L B 1966 *The Productive Thinking Program*. Brazelton, Berkeley, CA

Flower L S, Hayes J R 1980 Identifying the organization of writing processes. In: Gregg L W, Steinberg E R (eds.) *Cognitive Processes in Writing*. Erlbaum, Mahwah, NJ, pp. 3–30

Halpern D F (ed.) 1992 *Enhancing Thinking Skills in the Sciences and Mathematics*. Erlbaum, Mahwah, NJ

Kellogg R T 1994 *The Psychology of Writing*. Oxford University Press, New York

Mayer R E (ed.) 1988 *Teaching and Learning Computer Programming*. Erlbaum, Mahwah, NJ

Mayer R E 1992 *Thinking, Problem Solving, Cognition*, 2nd edn. Freeman, New York

Mayer R E 1997 Incorporating problem solving into secondary school curricula. In: Phye G D (ed.) *Handbook of Academic Learning*. Academic Press, San Diego, pp. 473–92

Mayer R E, Wittrock M C 1996 Problem-solving transfer. In: Berliner D C, Calfee R C (eds.) *Handbook of Educational Psychology*. Macmillan, New York, pp. 47–62

Nunes T, Schliemann A D, Carraher D W 1993 *Street Mathematics and School Mathematics*. Cambridge University Press, Cambridge, UK

Pressley M, Woloshyn V 1995 *Cognitive Strategy Instruction*. Brookline Books, Cambridge, MA

Sternberg R J 1990 *Metaphors of Mind*. Cambridge University Press, Cambridge, UK

Sternberg R J, Ben-Zeev T (eds.) 1996 *The Nature of Mathematical Thinking*. Erlbaum, Mahwah, NJ

Thorndike E L 1924 Mental discipline in high school studies. *Journal of Educational Psychology* **15**: 1–22

Wertheimer M 1959 *Productive Thinking*. Harper and Row, New York

R. E. Mayer

Teaching Social Sciences: Cultural Concerns

The concept of a *social* science is at once more elastic and ambiguous than that of *natural* science, for the content of the latter is essentially fixed by the properties of matter and energy in various degrees and combinations. Accordingly, the first task facing teachers of the social sciences is to make clear the grounds on which events and processes are taken to be *social*, and then identify the subject matter that defines a given discipline within the social sciences. Both of these objectives can be met only approximately and in ways influenced by seasonal shifts in perspective, as well as by developments in other and not always kindred specialties within the social sciences. These provisional solutions to problems of definition and classification serve as an introduction to comparably vexing and interrelated problems of explanation, measurement, modes of inquiry and standards of theoretical adequacy. The following sections consider these and related issues.

1. Defining the Social Sciences

The so-called unity of science was a primary objective of that positivism that was heralded in the Enlightenment and developed systematically during the first half of the twentieth century. Taking the subject matter of any and every science to be that which is in principle

observable, positivists found no principled argument against the possibility of subsuming even the most complex social phenomena under general laws. Illustrative of this perspective is Hempel's *Aspects of Scientific Explanation* (1965) in which such diverse fields of inquiry as psychology, history, and economics are judged to be amenable to scientific study and (ultimately) reducible to scientific laws.

Formidable arguments have been arrayed against this perspective based on what is taken to be unique to genuinely *social* phenomena. Dray (1957) rejected positivistic attempts to equate explanations in history with those in science. Unlike the phenomena of science, historical events are singular, are shaped and defined by specific participants, derive their essential features from the plans and purposes of these participants, and are subject to various but equally compelling and coherent explanatory accounts. Indeed, what makes an event 'social' in the first instance is just this constellation of personalities and purposes interacting in ways that are neither mechanical nor replicable.

The 'Battle of Waterloo' may be used illustratively. Consider first the nature of an event that qualifies it as a 'battle' rather than, say, a gang-war or a riot. An event takes on the character and proportions of a battle only insofar as it is a *military* action with clear political or economic or religious ends; only insofar as it is organized for these ends by persons in authority. Authority itself is grounded in complex social, legal, and cultural foundations, to be exercised in recognized ways, and to be retained or lost according to certain standards of conduct. Such a battle then becomes *decisive*, in just the way that Waterloo was decisive, to the extent that it yields consequences beyond the time and space within which the battle itself is waged. It becomes *historical* to the extent that these consequences then shape later and even larger political destinies, including those of nations that may not even have existed at the time of the decisive battle. There is no 'Battle of Waterloo' in the historical sense unless there is Napoleon, the delicate state of Anglo–European affairs in the wake of the French Revolution, the prospect of England's sea power, in the event of victory over Napoleon, supporting a great expansion of imperial influence, etc.

On a very small scale, the same considerations would apply to star-crossed lovers such as Romeo and Juliet. Their story is not indifferent to complex familial rivalries, to cultural values and ideals, to the unique bond that formed between them. Understanding Juliet's suicide is not achieved by studying suicides and reaching some general conclusion, e.g., unhappy persons see no cause for better days sometimes take their own lives. Rather, Juliet's actions become intelligible only under a set of descriptions unique to her life and her time.

In light of such considerations, it has been argued that the understandings to be sought within the social sciences must be drawn from a domain distinctly different from that in which merely physical interactions take place. The domain in which social phenomena occur is *social* owing to features not revealed by any of the prevailing physical or merely natural conditions. Thus, the very *idea* of a social science is constructed on grounds removed from those staked out by physics or biology (Winch 1958).

To these very counters, however, there are positivistic rejoinders and these, too, must be dealt with if students are to be informed as to the ongoing and keen disputes surrounding research and theory in the social sciences. After all, suicides are not random events, and battles tend to be won and lost for reasons historians are able to discern. This much alone suggests the operation of forces or principles that (ultimately) should give rise to scientific explanations of such events. What is important as regards the teaching of the social sciences is the recognition that these meta-scientific issues are central, not peripheral, to the agenda of the social sciences. The position one takes on them effectively dictates the sorts of events chosen for study, the method of inquiry, the nature of the explanations that will be taken to be successful, and the character of the more general theory that grounds such explanations.

Even as debates such as these unfold, specialists work with undaunted energy on problems that retain their interest no matter how the social sciences are defined. A more or less common sense standard is adopted with which to identify a problem as 'sociological,' 'historical,' 'economic,' 'psychological,' 'political', 'cultural.' One can instruct students on, e.g., 'Issues in Post-Soviet Society,' without taking a firm stand on whether (ultimately) the issues in question will be settled by subsumption under a general law. The issues are understood to be social in virtue of their grounding in a *society* recently liberated from an oppressive regime and now attempting to establish new and effective practices and institutions. Similarly, a course serving as an 'Introduction to Medieval Studies: The Age of Dante' is able to record and respect the artistic genius of Dante even as the themes of the *Divine Comedy* are shown to be influenced by medieval philosophy and wider currents of medieval thought. For all practical purposes, then, a field of inquiry is defined as 'social' when the dominant influences are interpersonal, when the salient occurrences reflect social aims and purposes, and when the causal factors are not readily or even plausibly rendered in physical–natural terms.

2. *Classifying the Social Sciences*

If social psychology is defined as the study of *interpersonal influence*, how does it differ from sociology? Also, if one of the major sources of influence is culture

itself, how are sociology, anthropology, economics, social psychology, and (cultural) history to be distinguished? Following the model of the great German research universities of the nineteenth century, doctoral programs have been developed to provide training in ever narrowing fields of specialization. As a result, the questions raised above tend to be answered by referring to just such specialized training. What makes a field of study anthropology rather than, say, social psychology is that it is populated by anthropologists and, as such, is examined and developed by methods of inquiry and standards of explanation of a distinctly 'anthropological' nature. In general, the anthropologist would seek to explain the nature and purpose of enduring customs within the given community; the social psychologist, the manner in which such prevailing customs influenced relationships between and among those living within the community; the economist, the manner in which such customs worked to the material advantage of the participants, perhaps especially in their relationships with trading partners, etc.

If this were all there was to the matter, there would be no substantive difference between teaching and training. Specialists devoted to a set of issues would simply undertake to train the next generation of workers who would then dutifully add to the data-base and seek evidence to 'verify' current speculations. The venture would be faithful to the received methods but, in the sense intended by Francis Bacon, would yield some fruit, but rather little light.

Surely one who examines the effect of trade on personal wealth must understand that 'wealth' may be construed in significantly different ways from one cultural enclave to another. A society made up of those who share the values of Henry James would regard as wealthy all who could 'satisfy the needs of the imagination.' It is a truism that the *value* of wealth depends on how its deployment or use operates at the level of lives individually lived and mutually influential. This, in turn, is related to core values and beliefs, often of a religious or transcendent or aesthetic nature, not reducible to coins of the realm. The point, of course, is the obvious one: specialized fields of inquiry tend to explain little beyond the artificially contracted domain within which the accepted methodology is (allegedly) applicable. Robust and plausible explanations then, require movement beyond such narrowly worked fields to a distance allowing a wider perspective and a deeper understanding. Those likely to find such a vantage point have been educated and not just *trained*.

3. *Methodologies*

Anthropological and cultural studies are dominated by field research that does not permit systematic control and manipulation of factors judged to be related causally to the events of interest. This does not mean that such studies are devoid of measurement *per se*; only that the mode of inquiry is not *experimental*. Sociological and social psychological research includes both experimental and field studies. The prevailing ethos within the social sciences favors the experimental mode of inquiry, a preference seemingly indifferent to the fact that experimental findings bear little relationship to the actual social phenomena they seek to explain or 'model.' The preference itself records a lingering attachment to scientific positivism.

Statistics is a methodological feature of experimental social science. A separate chapter in intellectual history is called for if students are to recognize the adoption of statistical standards of proof and the quite limited sense in which these standards are 'objective' (Daston 1994; Polanyi 1985). The commitment to general laws is *per force* at the expense of the individuated events giving rise to the collective. Inescapably the possibility is ever present that resulting general laws describe *none* of the participant-events and therefore explain none of them. To the extent that the general law thus derived is a model of anything, it may be a model of the data rather than of the social reality that was the instigating source of the inquiry.

More recently the mounting criticism of mainstream statistical methods has led to a liberalization of methods and to modes of inquiry more realistically connected to social beings in their actual complex and various interactions. So-called 'discourse analysis' avails itself of the relatively unconstrained reports of persons engaged in social practices. It is assumed that language itself, the bearer of meaning, reveals the essential nature of those joint ventures that define the social. So, too, with *narratological* methods of inquiry where participants themselves explain the dynamics of their social participations. If such nonmainstream methods can now claim only a little by way of discovery, they can claim much more by way of revealing the limitations and narrowness of the conventional alternatives.

4. *Explanation*

However defined, a 'science' (social or natural) offers itself as a set of practices and principles capable of explaining events in the real world; events in the world of social interactions, e.g., which would include bartering, voting, warring, cooperating, deceiving, assisting. But what does it mean to say that such occurrences and undertakings have been *explained*? Any adequate preparation of those who would advance or just comprehend the social sciences must reserve a central place for the issue of explanation itself. It is, after all, explanation that is the goal or end for which specialty fields are created, defninitions are crafted, methods are adopted.

Within the natural sciences, an event is judged to be explained when its occurrence can be predicted by or subsumed under a general law. Thus, the tendency of massive objects to move toward the center of the earth is explained by the universal law of gravitation. An object's 'falling' is but an instance of the law such that, if the law is true, the 'falling' is but a logical deduction.

The question thus arises as to whether this is the right model of explanation for the social sciences. Is the best explanation of Jack and Jill marrying, one (ultimately) grounded in a causal law of which this particular marriage is but an instance? Or, instead, are significant social undertakings tied not to causal laws, but to idiosyncratic purposes? Are social phenemona accounted for best by *reasons* explanations rather than *causal* explanations?

Students, to be prepared to understand the nature and the implications of such questions, are best served by studying special topics within philosophy devoted to just these matters. The fields of philosophy of social science, philosophy of psychology (Robinson 1985) and metaphysics include close examinations of what it means to explain events and, indeed, what it is that qualifies an occurrence as an 'event' of the relevant sort (Davidson 1980; Mele 1992).

5. *Theoretical Adequacy*

The social sciences now, owing to the relentless specialization within each of them, can claim a number of well-developed microtheories. This same specialization has, it would seem, retarded progress at the level of macrophenomena and, even more obviously, at the level of reality itself. As used here, 'level of reality' refers to the indisputable fact that actual persons seldom are moved by unitary motives toward stable and singular goals, attentive to but one set of environmental cues, and in a fashion rigidly determined by prevailing cultural standards and local meanings. Indeed, such a person would be a candidate for therapy. In teaching the novice or even the advanced student the methods and prevailing theories in any of the social sciences, it becomes important to distinguish between theories (more or less) supported by findings, and those (more or less) supported by the warp and weave of lives actually lived. Too often the 'findings' arise from methods of inquiry so inextricably bound to theoretical presuppositions as to achieve no more than a rediscovery of the presuppositions themselves. What William James dubbed 'the psychologist's fallacy' might be applied generously across the social sciences. James's version refers to the adoption of an experimental method—say the use of a memory-drum in a paired-associate study of memory—that virtually guarantees the form the targeted process will take, e.g., subjects remember terms that have been associated with other terms! Where a general theory assumes that the prime motive to action is economic, and where the modes of inquiry leave little room for any influence other than economic influences, there will result nothing but jejune 'verifications.'

Before the age of specialization, what would be understood as social science was the speculative field occupied by humanistic scholars and philosophers, by writers in law and politics, by designers of buildings and even entire cities. A 'social' science stood as a systematic inquiry into the social conditions that thwart or realize human potentialities; social conditions including but not limited to parental guidance, systems of law, literature and the arts, material means, modes of employment, religious teachings, and institutions. For Plato and Aristotle, the social and the political were inseparable, and the person was understood to be powerfully shaped by both. A developed 'social' science then, presupposed a civic or political science, itself based on a defensible if imprecise theory of human nature. What the ancient accounts lacked in precision they make up for in breadth, integration, and realism.

At present, and in light of prevailing standards and habits, the social sciences offer no comparably realistic and integrative theory of social life. This is one, but not the only, or even the principal, reason why major figures in the history of philosophy continue to attract the attention of those seeking such integrations. These same major figures are usefully consulted in the teaching of social scientists, if only to alert them to the nature of systematic and integrative thinking on matters of social consequence.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Disciplines, History of, in the Social Sciences; Explanation: Conceptions in the Social Sciences; Normative Aspects of Social and Behavioral Science; Paradigms in the Social Sciences; Psychology: Historical and Cultural Perspectives; Social Science, the Idea of; Theory: Conceptions in the Social Sciences

Bibliography

Daston L 1994 *Classical Probability in the Enlightenment*. Princeton University Press, Princeton, NJ

Davidson D 1980 *Essays on Actions and Events*. Oxford University Press, Oxford, UK

Dray W 1957 *Laws and Explanation in History*. Oxford University Press, Oxford, UK

Hempel C 1965 *Aspects of Scientific Explanation and Other Essays*. Free Press, New York

Mele A 1992 *Springs of Action*. Oxford University Press, Oxford, UK

Polanyi M 1985 *Personal Knowledge*. University of Chicago Press, Chicago

Robinson D 1985 *Philosophy of Psychology*. Columbia University Press, New York
Winch P 1958 *The Idea of a Social Science*. Routledge, London

D. N. Robinson

Teaching: Trends in Research

Research on teaching came into its own as a field of study beginning in the early 1970s when large-scale research on teaching studies attempted to tie teacher behaviors to student achievement. In these studies, teachers who were identified as more or less effective on the basis of their students' scores in achievement tests were observed and their behaviors recorded. Correlational data analysis identified differences in behaviors between the more and less effective teachers. This information, much of it leading to the concept of *time on task*, was used for purposes of teacher evaluation, teacher education, and staff development.

Concerns about the nature of the product measures (reading and mathematics multiple choice achievement tests), research methodology, and potential context effects eventually led to the development of different approaches to studying classrooms. Disciplinary frameworks other than educational psychology were applied in the research. These disciplines included anthropology, socio-linguistics, and sociology. Research employing qualitative methodology and case-study approaches began to be published in the major journals. Also, the cognitive revolution turned the attention of researchers away from teacher behaviors toward teachers' thought processes as they plan, implement their lessons, and reflect on their practices. During the late 1980s and early 1990s, this movement to teacher cognition affected all aspects of research on teaching, including the conception of the effective teacher and the research methodologies used to explore teaching.

During this period of time, an even more significant social theory trend was strongly influencing the nature of research on teaching: the movement from modern to postmodern thinking. Between the previous edition of the *Handbook of Research on Teaching* (3rd edn.) (Wittrock 1986) and the present one, the field has experienced a time of turbulence similar to that in other fields. Postmodernism raises questions that jar the very foundations of our understanding of research. These questions concern the nature of knowledge, who owns it, who produces it, and how it should be used. The dissonance between modern and postmodern ways of thinking are evident in the trends described below, and visible in the current edition of the *Handbook of Research on Teaching* (4th edn.) (Richardson in press). The conflicts go beyond the old qualitative–quantitative methodology controversy as described in the 1986 *Handbook*. They now focus on the very nature of research and knowledge and the uses of research in the improvement of practice.

1. *A Vision of Teaching*

During the 1980s, the vision of teaching changed from that of a teacher standing in front of the class and transmitting information and skills to students to one of the teacher as facilitator of learning. The newer conception of teaching originates from constructivist learning or meaning-making theory that suggests that individuals create their own new understandings based upon the interaction of what they already know and believe, and the ideas with which they come into contact. The teacher facilitates student learning by creating environments, tasks and dialogues in which students access their prior knowledge, consider alternative premises, and alter and add to their existing understandings with information and ways of thinking that are introduced to them by the teacher and other experts, readings, other students, and their own research. Current classrooms have students with diverse backgrounds. Since each student has unique background knowledge and experiences, the teacher is compelled to keep track of individual differences, and to adjust lessons on the basis of this knowledge.

Within this conception, teaching requires complex thought and decision making in situations of uncertainty with diverse student bodies and variable contexts. Notions of the complexity of teaching and the variability of the context work together to help promote the view of the teacher as a thinking, decision-making, reflective, and autonomous professional. Since teaching is complex and contexts vary, teachers themselves need to make decisions and reflect on their situations and teaching in order to act in appropriate ways.

Current research on teaching reflects this change in the conception of the teacher, as will be highlighted in the sections below that reflect the current trends in the field.

2. *Teacher Change*

As the conception of teaching changed from transmission to facilitation, the question of how teachers themselves could be encouraged to change their ways of thinking and approaches to teaching became paramount in the research, policy, and practice literatures. Since the various subject-matter standards, such as the national mathematics standards for students, called for a very different approach to teaching, research and practice began to focus on processes of teacher change. The anticipated changes in teachers and teaching were thought to entail more than the acquisition of a new teaching method, but involved changes in beliefs and practices around the nature of teaching. This form of change, itself, required the development of conceptions of teacher change, and

research on processes that would be effective in helping teachers develop these newer teaching roles.

The newer research on teacher change contests the view that teachers are recalcitrant and do not change. The recalcitrant label stems from a view of change in which individuals or groups outside the classroom mandate or suggest the changes that teachers should make. In fact, research suggests that teachers voluntarily change all the time. They reorganize their classrooms, try different activities and texts, change the order of topics in the curriculum, attempt different interpersonal skills, and so on. When teachers experiment with new activities in the classroom, the new practices are assessed on the basis of whether they 'work.' When these activities work, it means that they engage the students, do not violate the teacher's particular need for control, and help the teachers respond to system-determined demands such as high test scores. If they do work, they are internalized and absorbed into the teacher's repertoire of activities.

This new understanding of naturalistic teacher change has led to a considerable number of studies of processes that would help teachers change their beliefs about learning, the nature of teaching, and the role of the learning environment in student development. These studies often focus on particular subject-matter areas such as mathematics, science, English, and social studies since the learning and teaching standards are usually developed around these subjects. The most productive research-based approach to staff development suggests a long-term effort in which teachers are engaged in inquiry into their own often tacit beliefs and practices. This process is enhanced through dialogue, particularly with those who understand practice and the particular context in which the teacher is working. A certain trust level is important within the community since it is important for the participants to accept and talk about practices that do not seem to work, and to accept responsibility for them. Thus, the development of a discourse community is productive in beginning this process of change.

Research on teacher learning that examines changes in knowledge and beliefs as well as practices has led to the development of a number of different conceptions and forms of teacher knowledge and belief.

3. *The Nature of Knowledge in Teaching*

As the focus of attention in research on teaching shifted from teacher behaviors to teacher cognition, beliefs and knowledge have become the two most important factors in the explanations of teacher practices and in considerations of teacher change. In the traditional philosophical literature, knowledge requires a 'truth condition' that suggests that a proposition is agreed upon as being true by a group of people. Propositional knowledge has epistemic standing, that is, there is some evidence to back up the claim. By contrast, beliefs do not require a truth condition. They have also been found to be powerful in their effects on teaching practices.

However, within research on teaching, the differentiation between beliefs and knowledge is not strongly evident. Many in the field define knowledge as that which is held in the teachers' heads, with or without a truth condition. This psychological view of knowledge has led to the identification of and research on a number of different forms of knowledge. Two forms of knowledge that have been studied extensively are *practical knowledge* and *pedagogical content knowledge*. Practical knowledge differs from *formal knowledge* in that formal or warranted knowledge is much more closely related to the philosophical conception of knowledge. Formal knowledge may be found in textbooks and research articles; whereas practical knowledge may be found in teachers' heads and in their actions and focus on their own classroom contexts. Pedagogical content knowledge, on the other hand, combines formal and practical knowledge.

Beliefs, defined as propositions that are accepted as true by the individual holding the beliefs, are of interest in research on teaching in the investigation of the ways in which beliefs may affect teaching practices. Most current studies of teacher beliefs are conducted within the hermeneutic tradition (that is, they focus on how the individual teacher makes sense of the environment in which s/he is operating). These studies suggest a complex relationship between teachers' beliefs and actions. In most current research-based conceptions, the perceived relationship between beliefs and actions is interactive. Beliefs are thought to drive actions; however, experiences and reflection on action may lead to changes in and/or additions to beliefs.

Beliefs are seen as important in teaching and teacher education in several ways. The first suggests that students of teaching bring powerful beliefs into their teacher education classes. These beliefs must be attended to by both the student teacher and the teacher educator if the student teacher is to move beyond the images of teaching acquired as a student. An additional and related way is instruction that focuses directly on beliefs. As Tom Green (1971) suggested, one goal of teaching is to help students form belief systems that consist of a large proportion of beliefs based on evidence and reason. Thus, the investigation of beliefs in the teacher education classroom should involve their rejection, alteration, or transformation into knowledge with warrant and evidence attached. Without attention to beliefs, transformational changes in teaching practices have a low probability of success.

Practical knowledge is an account of how a teacher knows or understands a classroom situation. Practical knowledge is gained through experience, is often tacit, and is contextual or local. This form of knowledge is not, however, synonymous with beliefs because it is thought of as embodied within the whole person, not just in the mind. Embodied knowledge is more than

cognitive and relates to the ways in which people physically interact with the environment. It is this knowledge that may be used in an improvisational manner in the classroom. This conception of knowledge as practical does not separate the knower from the known, is personalized, idiosyncratic, contextual, and emerges during action.

Pedagogical content knowledge refers to a way of knowing the subject matter that allows it to be taught. It is grounded in the disciplines but adds an understanding of how to transform formal knowledge of the disciplines into the enacted curriculum within a teaching context. This knowledge combines that of the subject matter itself with knowing how students learn the content, students' preconceptions that may get in the way of learning, and representations of the knowledge in the form of metaphors, examples, demonstrations, etc., that allow it to be transformed into material that the students may learn. Inquiry into teachers' pedagogical content knowledge has been active since 1985. This research suggests that teacher education students' formal knowledge of the disciplines is fairly weak, and that formal teacher education programs are a weak contributor to the formation of pedagogical content knowledge. However, there is strong evidence that those becoming teachers benefit from the courses that emphasize pedagogical knowledge as compared with those who enter teaching with subject matter background, but no pedagogical education.

4. *Areas of Interest in Research on Teaching*

The shift from teacher behavior to cognition and from the transmitter to facilitator models of teaching has brought forward a number of areas that are of particular interest in research on teaching.

4.1 *Constructivist Teaching*

Constructivism is a descriptive theory of learning, and most constructivists would agree that the traditional approach to teaching—the transmission model—promotes neither the interaction between prior and new knowledge nor the conversations that are necessary for internalization and deep understanding. The information acquired from traditional teaching, if acquired at all, is usually not well integrated with other knowledge and beliefs held by the students.

While these are common understandings, there are also considerable disagreements. Further, and more importantly in this context, there is not a direct relationship between the theory of learning and the practice of teaching. Constructivism is a theory of individual learning. While constructivist learning theory may speak closely to a tutoring situation, it does not direct teaching practice within a 'normal' classroom. Thus, considerable research attention focuses on the constructivist classroom. Some of the questions that are being addressed are as follows: (a) what is the place of formal knowledge in a constructivist classroom, and how it should be introduced, and (b) at what point during instruction does it make sense to 'tell' students formal knowledge and/or answers to questions? These issues are being worked out within the different subject matters, and evidence indicates that constructivist classrooms vary depending upon the subject matter.

4.2 *Teaching to Individual Differences*

Constructivist learning theory suggests that the background knowledge, beliefs, and understandings that students bring to the classroom strongly affect what they learn within the classroom. Given this frame, and the current attempts to integrate students with differing cultural and language backgrounds as well as learning styles and abilities, suggests that attention to the backgrounds of students is critical in the teaching process.

Research on differences in the ways teachers approach girls and boys in the classroom has a longstanding tradition in research on teaching. There is no doubt that teachers, by and large, treat girls differently to boys and that teachers are not aware of doing so. Examples are the questions asked of boy versus girl students. They are often qualitatively different, and the intellectual follow-up on answers is often different and usually more thorough for boys than girls. However, it is also the case that girls act differently than boys in classrooms, and the effort continues to be made to understand teachers' different behaviors as a reaction both to the differences in classroom behavior of boys and girls and the social forces that may drive these different behaviors.

Multicultural education has received considerable attention in research on teaching as schools and classrooms continue to be more integrated. Further, it is also the case in many countries that the teaching population itself is dominated disproportionately by teachers of the majority culture. Knowledge and understanding of different cultures and ways of approaching the tasks of teaching and learning thus becomes imperative for teachers. Of particular importance in this area is the development of the concepts of *culturally relevant learning* and *teaching*. Based on research of successful teachers of African-American students, this concept suggests that knowledge of the culture and community of the students is brought into these classrooms and guides elements of classroom action such as the discipline system.

4.3 *Moral Development*

For years, research has focused on student achievement in the basic skills of reading and mathematics. More recently, other subject-matter areas such as science, writing, and geography have been included in

this research. It is also the case, however, that teachers may have a strong influence on the moral development of students. Current research in this area is investigating the moral dimensions of classroom life, and indicates that the moral life of the classroom is robust and powerful. Teachers, however, are often unaware of the nature and/or consequence of their actions in relation to the moral development of their students. Further research is needed to examine the effects of these the moral dimensions on students, and will undoubtedly affect the ways in which teachers approach this important aspect of classroom life.

5. *Methodology*

The methodology for research on teaching has shifted dramatically from quantitative to qualitative. Case studies are proliferating, particularly in studies of teacher change that focus not just on documenting the fact that the teachers changed, but also on building a theory of the change process. While quantitative methodology is still being used in large-scale studies of educational reform, most studies that examine the teacher learning and change processes are qualitative in nature. Qualitative methodology, however, is not monolithic. There are many different approaches, and Robert Donmoyer (in press) in a chapter in the *Handbook of Research on Teaching* (4th edn.) has differentiated among them in relation to five different purposes for conducting the research. These purposes range from 'truth' seeking to the praxis/social change purpose. Other trends that are found in research on teaching are outlined below.

5.1 *Teacher Research and Self-study*

Teacher research refers to a form in which the teacher conducts research on or inquires into his/her own practices. Self-study is the same process used by teacher educators. This trend is a part of the action research movement that suggests that practitioners are in the best position to inquire into their own practices for purposes of solving problems in the contexts in which they reside and understand. The purpose of teacher research and self-study is primarily the improvement of the teacher/teacher educator's teaching practice. However, there is some debate about this, as a number of scholars suggest that research conducted by teachers is not only useful for the teacher-researcher but of interest to others. There are also questions about the nature of the research that is conducted, e.g., do teacher research and self-study warrant different methods and procedures to research that leads to formal knowledge?

5.2 *Narrative Research*

Narrative research stems from the sense that narrative is a mode of thinking—an expression of cultures' storehouse of knowledge (Bruner 1986). It is argued that teachers' knowledge and understanding of school and classroom practice is stored in narrative and discourse about practice is often in narrative form. The aim of narrative research is to capture school and classroom practice and tell others about it in such a way as to conform to this natural mode of thinking. As described by Gudsmundsdottir (in press): 'narrative approach moves research on school practice as a field out of the constraints that educational psychology has placed upon our community and enables us to move where we belong—into the realm of human sciences as conceptualized by Dilthey.' The inquiry focuses on mediated action, and it is qualitative and interpretive in nature. As a relatively new approach in research on teaching, it has its strong advocates and detractors. It also means different things to different people. However, the attempt to develop a research methodology of practice will keep it in the limelight for some time.

5.3 *Studies of Teacher Change*

A classical experimental design relies heavily on a linear process that involves the prespecification of the treatment and the desired outcomes. Many current change processes—particularly inquiry-oriented staff-development programs—are not of this nature. They are collaborative with the practitioners, and the process moves along a path marked by constant inquiry and reflection. It is questionable whether the use of a control group adds to our understanding of teacher change. Since so many current change processes are fairly individualized without stipulated goals for changes in practice, the question that remains is whether the teachers in the control and treatment group changed. Since we are operating in a model of change that suggests that teachers change all the time, we would assume that even the control teachers would be changing as well.

Thus, in examining teacher change in depth, the questions of significance, validity, and worth must be addressed. The set of questions are as follows: was there change in teaching practice? can this change be attributed to the change process being examined? is the effected change desirable? This calls for a rethinking of research designs that, for example, leads to the ability to tie observed changes in practices to the substance of the conversations that took place in the staff-development process.

See also: Classroom Assessment; Cognitive Styles and Learning Styles; Educational Research for Educational Practice; Teacher Behavior and Student Outcomes; Teacher Education; Teacher Expertise; Teaching and Learning in the Classroom; Teaching as a Profession

Bibliography

Bruner J 1986 *Actual Mind, Possible Worlds*. Harvard University Press, Cambridge, MA

Cochran-Smith M, Lytle S 1993 *Inside Outside: Teacher Research and Knowledge*. Teachers College Press, New York

Clandinin J, Connelly M (eds.) 1995 *Teachers Professional Knowledge Landscapes*. Teachers College Press, New York

Donmoyer R in press. In: Richardson V (ed.) *Handbook of Research on Teaching*, 4th edn. American Educational Research Association, Washington, DC

Fenstermacher G D 1994 The knower and the known: The nature of knowledge in research on teaching. In: *Review of Research in Education*. American Educational Research Association, Washington, DC, Vol. 20, pp. 1–54

Goodlad J, Soder R, Sirotnik K (eds.) 1990 *The Moral Dimensions of Teaching*. Jossey-Bass, San Francisco

Green T 1971 *The Activities of Teaching*. McGraw-Hill, New York

Gudsmundsdottir S in press Narrative research on school practice. In: Richardson V (ed.) *Handbook of Research on Teaching*, 4th edn. American Educational Research Association, Washington, DC

Ladson-Billings G 1994 *The Dreamkeepers: Successful Teachers of African-American Children*. Jossey-Bass, San Francisco

Oser A D F, Paltry J L 1992 *Effective and Responsible Teaching*. Jossey-Bass, San Francisco

Richardson V (ed.) 1994 *Teacher Change and the Staff Development Process: A Case in Reading Instruction*. Teachers College Press, New York

Richardson V (ed.) in press *Handbook of Research on Teaching*, 4th edn. American Educational Research Association, Washington, DC

Shulman L 1987 Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review* **57**: 1–22

Wittrock M (ed.) 1986 *Handbook of Research on Teaching*, 3rd edn. Macmillan, New York

V. Richardson

Teamwork and Team Training

During recent years, teams have emerged as the cornerstone of many organizations. Team efforts are required in many businesses and industries (e.g., government agencies, aviation operations, military organizations, and hospitals) to meet their missions and goals (Ilgen 1999). When applied correctly, teams may offer several advantages over individuals or groups. For example, teams can perform difficult and complex tasks, motivate their members effectively, and in some cases outperform individuals (Foushee 1984). Specifically, teams can be more productive, make better decisions (Orasanu and Fischer, 1997), perform better under stress, and make fewer errors (Wiener et al. 1993). When organized, designed, and managed correctly, teams can be a powerful organizational tool (Guzzo and Dickson 1996).

The purpose of this article is to provide a brief overview of what we have begun to learn about the nature of teamwork and what can we do to enhance it. Five questions will be discussed: (Sect.1) what is a team? (Sect.2) what are team competencies? (Sect.3) what is teamwork? (Sect.4) what is team training? and (Sect.5) what is team measurement? Several key concepts will be defined.

1. *What is a Team?*

According to Salas and Cannon-Bowers (2000) a team can be defined as 'a set of two or more individuals who must interact and adapt to achieve specified, shared, and valued objectives' (p. 313). Teams must also have meaningful task interdependencies—the job cannot be done by a single individual—and usually task-relevant knowledge is distributed among the team members. Other characteristics of teams are that members: (a) have roles that are very defined and functional; (b) are hierarchically organized; (c) share two or more information sources; and (d) have a limited time span (Salas and Cannon-Bowers 2000).

This definition is important because it highlights the fact that not all teams are created equal. Teams differ from groups (e.g., task forces, committees) with respect to the degree of task interdependency, structure, and requirements (Salas and Cannon-Bowers 2000). In addition, teams differ from groups regarding essential knowledge and skills; necessity of critical team feedback for success; and development of shared goals, expectations, and vision. Since teamwork is characterized by synchronizing goal-directed behaviors, interdependent activities, and functions that are separate from the task itself, teams hold unique features, which most other small groups do not. Therefore, one needs to understand the competencies needed for team members to succeed in a team; they will be discussed next.

2. *What are Team Competencies?*

In recent years, research has uncovered the knowledge, skills, and attitudes (i.e., competencies) necessary for effective team performance. These team competencies allow organizations to establish the appropriate requirements for their teams and strategies to enhance teamwork and performance. Teamwork competencies can be characterized as resources that team members draw from in order to function. As noted, they include knowledge, skills, and attitude-based factors.

Knowledge-based competencies (i.e., what team members 'think') refers to the necessity of understanding facts, concepts, relations, and underlying foundations of information that a team member must have to perform a task. For teams to perform required tasks, they must possess a certain set of knowledge. (see Table 1 for some examples.) It has been noted that team members must have synchronized mental models (i.e., knowledge structures) about the roles of their

Table 1
Knowledge-based competencies examples

Type of knowledge	Definition
Cue-strategy associations	Knowing how and when to use environmental cues to change coordination strategies
Team-mate characteristics	General tendencies, abilities, preferences, strengths, weaknesses of team mates
Accurate and shared task models	Team members must share a common interpretation and understanding of task strategies and demands
Task sequencing	Knowing how to organize tasks inputs according to team and task demands

Source: adapted from Salas and Cannon-Bowers 2000

Table 2
Skill-based competencies examples

Type of skill	Definition
Adaptability	Process by a team to use information from a task environment to alter strategies through compensatory behavior, dynamic reallocations of functions, and flexibility
Situational awareness	Process by a team to develop compatible models (or shared understanding) of internal and external environment of team
Communication	Process by which information is exchanged between two or more team members with clarity and accuracy
Decision making	Process by a team to gather and integrate information, use sound judgment, recognize alternatives, select the best solution, and evaluate the consequences

Source: adapted from Salas and Cannon-Bowers 2000

teammates, the tasks, and the situations that the team encounters (see Cannon-Bowers and Salas 1998). Expectations created by these knowledge structures allow team members individually to generate predictions concerning how to perform during routine and novel situations that the team encounters. It has also been shown that it is necessary for team members to possess knowledge about the purpose and objectives of their mission, the available and needed rewards, and the expected norms to be followed.

An important knowledge-based team competency, cue/strategy associations, has been suggested as important for the functioning of teams in complex, knowledge-rich environments (Salas and Cannon-Bowers 2000). Through cue/strategy associations, team members are triggered to perform in a certain way by a pattern of cues in the environment. Essentially, team members make associations between this pattern of cues and what they should do. Teams are able to adapt, reallocate functions, and occasionally perform without the need for overt communication because of this knowledge. While the cue/strategy associations competency is key, there are other knowledge-based competencies that are useful to teams such as: knowledge of teammate characteristics, accurate and shared task models, and team sequencing (Salas and Cannon-Bowers 2000).

Skill-based competencies (i.e., what team members 'do') are the necessary behavioral sequences and procedures needed for task performance. It is clear that team members must behave in certain ways and do certain things in order to achieve their goals. Specifically, teams perform various actions—coordination, communication, strategization, adaptation, and synchronization of task-relevant information. The information can be verbal, nonverbal, or visual. Team members require skills to perform in a fashion that is accurate and timely. As with knowledge-based competencies, there are also many skill-based competencies required for effective team performance (see Table 2 for more examples).

Attitude-based competencies (i.e., what team members 'feel') refer to the affective components needed to perform the task. Teamwork is greatly affected by how team members feel about each other and the task. Due to considerable research, several attitude-based competencies have emerged (i.e., motivation, trust, shared vision, importance of teamwork) which affect teamwork and team outcomes (see Table 3 for a more in-depth description). For example, collective self-efficacy (i.e., the belief that the team can perform the task) and collective orientation are important in team functioning.

The team competencies described here are the necessary elements for effective teamwork and become the desired target when developing team training. The final goal of team training is to improve and foster effective teamwork. To do this, researchers must

Table 3
Attitude-based competencies examples

Type of attitude	Definition
Motivation	Process by a team in which objectives are defined and members are energized to achieve them
Mutual trust	Team members' attitude regarding mood or climate of team's internal environment
Shared vision	Team members' common attitude regarding the team's direction, goals, and mission
Importance of teamwork	Team members' attitude regarding working as a team
Efficacy	Belief that a team can perform effectively (or be successful) as whole when given specific task demands
Collective orientation	Belief that team approach is better than an individual approach

Source: adapted from Salas and Cannon-Bowers 2000

uncover, analyze, and address the necessary knowledge-, skill-, and attitude-based competencies, while also considering the nature of the team task (Salas and Cannon-Bowers 2000). A discussion of what comprises teamwork follows.

3. *What is Teamwork?*

Teamwork can be defined as the ability of team members to work together, communicate effectively, anticipate and meet each other's demands, and inspire confidence, resulting in a coordinated collective action. However, a clear and direct answer to the question 'What is teamwork?' has not been provided. According to McIntyre and Salas (1995), teamwork is a critical component of team performance and requires an explanation of how a team behaves. There are four key behavioral characteristics that compose teamwork: (a) performance monitoring, (b) feedback, (c) closed-loop communication, and (d) back-up behaviors.

The first requirement of teamwork is that team members monitor the performance of others while carrying out their own task. Monitoring ensures that members are following procedures correctly and in a timely manner, while also ensuring that operations are working as expected. Performance monitoring is accepted as a norm in order to improve the performance of the team in addition to establishing a trusting relationship between members.

Next as a follow-up activity to monitoring, feedback on the effectiveness or ineffectiveness of performance is passed along to members being monitored. Team members must feel at ease when providing feedback for teamwork to be effective. No obstacles (i.e., rank, tenure) should stand between members who are giving and receiving this vital information. Free-flowing feedback exists in the highest level of teamwork.

Teamwork involves the effective communication between a sender and receiver. Closed-loop communication describes the information exchange that occurs in successful communication. There is a sequence of behaviors that is involved in closed-loop communication: (i) the message is initiated by the sender; (ii) the message is accepted by the receiver and feedback is provided to indicate that it has been received; (iii) the sender double-checks with the receiver to ensure that the intended message was received. This type of communication is especially apparent in emergency situations.

Finally, back-up behaviors (i.e., the willingness, preparedness, and liking to back team members during operations) are required for effective teamwork. Team members must be willing to help when they are needed and must accept help when needed without feeling they are being perceived as weak. This requires that members know the expectations of their jobs while also knowing the expectations of others with whom they are interacting.

These four complex behavioral characteristics—performance monitoring, feedback, closed-loop communication, and backing-up behaviors—are necessary for effective teamwork. A failure of one of these aspects could result in ineffective team performance. Next, team training and four training strategies will be defined.

4. *What is Team Training?*

Team training can be defined as a set of tools and methods that form an instructional strategy in combination with requisite competencies and training objectives (Salas and Cannon-Bowers 1997). It is concerned with the design and development of theoretically based instructional strategies for influencing the processes and outcomes of teams. Team training consists of strategies derived from synthesizing a set of outputs from tools (e.g., team task analysis); delivery methods (e.g., information-based); and content (e.g., knowledge, skills, and attitudes (Salas and Cannon-Bowers 1997, Cannon-Bowers and Salas 1998).

Team training has one basic objective—to impact crucial team competencies. More specifically, team training seeks to foster in the team members, sufficient

and accurate mental representations of the team task and role structure, and the process by which members interact.

There are four team training strategies that have emerged in the literature and will be discussed—team coordination training, cross-training, team self-correction training, and team leadership training.

Team coordination training, also known as team adaptation and coordination training, is the most common team training strategy (Salas and Cannon-Bowers 1997, Entin and Serfaty 1999). This training strategy is widely used in the aviation industry and is known in commercial aviation as crew resource management, and in military aviation as aircrew coordination training (Wiener et al. 1993).

First, team coordination training is based on the shared mental model (SMM) theory. SMM theory posits that shared knowledge enables team members to predict the needs of the task and anticipate the actions of other team members, allowing them to adjust their behavior accordingly (Cannon-Bowers and Salas 1998). Expectations of what is likely to happen, are formed by team members, as a means to guide action. Especially in a novel situation (when they cannot strategize overtly) teams must exercise shared models of the task and team in order to maintain performance. Thus, the role of shared mental models in explaining team behavior stems from the team's ability to allow its members to generate predictions about team and task demands in the absence of communication. The idea of SMMs and their relation to team effectiveness provide several implications for the understanding of team performance and training (see Cannon-Bowers and Salas 1998).

Team coordination training was designed in order to reduce communication and coordination overhead within the team. Its development was a result of testing and observing successful teams who adapted to increased levels of stress by changing their coordination strategies. Research has shown that these teams performed significantly better on posttraining scenarios than did teams assigned to control conditions. Changes in communication dynamics of team members under stress were used to enhance team performance. Practice on effective teamwork strategies is important in order to change behaviors so they adapt to increasing task demands, which has been demonstrated through team coordination training (Cannon-Bowers and Salas 1998).

Next, the cross-training strategy, also based on the SMM theory, is necessary to lessen the decline in team performance due to turnover. It should be designed so that commonly shared expectations, called interpositional knowledge, regarding specific team-member functioning can be developed. To accomplish this, all parties can be provided with 'hands-on' experience regarding the tasks, roles, and responsibilities of other members within the team. Cross-training research has shown that it is an important determinant of effective communication, task coordination, and performance (see Cannon-Bowers and Salas 1998).

Third, team self-correction training emphasizes the need for team members to provide each other with feedback regarding their performance on a task and the underlying expectations for effective performance. This helps to clarify correct expectations and intentions between members, while facilitating modifications of expectations and intentions that are incorrect. This approach trains teams to observe both their own and each other's performance in order to provide feedback following task completion. Teams are then able to modify their knowledge, skills, and attitudes without intervention from an outside source. Along with team coordination training and cross-training, team self-correction training is based on the SMM theory (see Cannon-Bowers and Salas 1998).

Finally, team leadership training is another team training strategy. When training team competencies, or KSAs, the training begins with the team leader. However, effective team leadership requires: (a) competence that is both technical and operational, (b) knowledge of how to lead a team, and (c) willingness to listen and use team members' expertise. A critical factor in the success of a team is the team leader. A team leader, through specific instruction, can guide the team members to develop those KSAs needed to perform effectively. The influence of a team leader can impact, either advantageously or otherwise, the learning skills of a team and the instructional value of an experience. However, the team leader may not possess the skills needed to accomplish the task. To give him the necessary skills, team leadership training should be implemented. The team leader should be taught to recognize and target crucial teamwork skills needed for training such as communication, coordination, decision-making, and adaptability/flexibility. In the following section, what needs to be measured to diagnose effective or ineffective teamwork is discussed.

5. *What is Team Measurement?*

Measurement of team performance is a component that is critical in team training. Salas and Cannon-Bowers (2000) state that 'measurement is the basis upon which we assess, diagnose, and remediate performance' (p. 322). They continue by noting that learning is promoted and facilitated through measurement. Measurement also produces and provides information about skill acquisition and a metric upon which capabilities and limitations of human performance can be judged. For team training, the dynamic, multi-level nature of teamwork must be able to be captured by the measurement approach. Measurement not only serves to describe teamwork, but it can also be used in the development and guidance of instruc-

tional events. The multiple pieces of information needed for feedback must be integrated by a complete team performance measurement system. In addition, the measurement system must have the following features to foster team training: be theoretically based; assess and diagnose team performance; address team outcomes as well as processes; capture moment-to-moment behaviors, cognitions, attitudes, and strategies; consider multiple levels of analysis; consider multiple methods of measurement; provide a basis for remediation, and rely on observation by experts (Salas and Cannon-Bowers 2000).

According to Salas and co-workers, team performance measurement should be able to: (a) identify the processes linked to team outcomes, (b) distinguish between deficiencies in individual and team levels, (c) describe interactions between team members to capture occurring moment-to-moment changes, (d) produce assessments that may be used to deliver performance feedback, (e) produce reliable and defensible evaluations, and (f) support operational use (Salas and Cannon-Bowers 2000).

Performance measurement is done for the purpose of inferring something about the knowledge, skills, and attitudes of the team member. Observations of critical team behaviors is also required of team performance measurement in order for inferences to be made about categories of behavior or dimensions such as communication, coordination, and adaptability. Techniques have been established for measuring teamwork skills and for studying the relationship between teamwork and team task performance (Ilgen 1999).

A team task analysis, which is needed to uncover the coordination demands of complex tasks, is generally the first step for measuring team performance and designing team training. These analyses provide information about the necessary team learning objectives and team competencies. Key information necessary for designing and delivering team training is also generated by the team task analysis. In addition, team task analysis provides information about the KSAs, cues, events, actions, coordination demands, and communication flows needed for teamwork to be effective. Typically, subject matter experts (SMEs) are asked to provide information on these aspects of the task. This information is then utilized to specify objectives of team training, to create a task inventory that focuses on those objectives, and to guide training scenario development (Salas and Cannon-Bowers 2000).

There are two methodologies that need to be considered for measuring team skills: (a) team vs. individual measures, and (b) outcome vs. process measures. Individual measures assess the skills required to perform and meet individual tasks and responsibilities. On the other hand, teamwork measures focus on the coordination requirements (e.g., performance monitoring, feedback, communication, and back-up behaviors) that are necessary between team members. Process measures (i.e., how the task was accomplished) are important for determining problems in performance that may prevent successful outcomes. However, successful outcomes can be achieved when processes are flawed. Outcome measures usually rely on opinions from experts or automated recording of performance. Because outcome measures do not provide information that is useful in determining the cause of poor performance or performance improvement, it is necessary to measure both process and outcomes to ensure that team performance is effective.

It is important when measuring team performance and training to follow the theory and methodologies outlined above. The team task analysis, or theory level, will allow the researcher to define the critical dimensions, behaviors, objectives, and competencies required of the team training. Team performance and training should be measured next. Measurement of team skills is usually accomplished using the team vs. individual measures approach and/or the outcome vs. process measures approach. It is important to measure all aspects of a team's performance to ensure its effectiveness.

6. *Conclusion*

The purpose of this article has been to provide information regarding teamwork and team training. The authors have attempted to present an overview of these areas by providing definitions of a team, teamwork, and team training. In addition, the necessary team competencies—knowledge, skills, and attitudes—were briefly described. It should be noted that the KSAs mentioned are neither exclusive nor exhaustive. In addition, several team training strategies necessary for effective team performance were discussed. Finally, team measurement through a team task analysis (i.e., development of team training objectives and competencies) and methodologies were shown to be an important part of effective team performance.

See also: Cooperation and Competition, Psychology of; Cooperation: Sociological Aspects; Group Decision Making, Cognitive Psychology of; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Groups, Sociology of

Bibliography

Cannon-Bowers J A, Salas E 1998 Individual and team decision making under stress: Theoretical underpinnings. In: Cannon-Bowers J A, Salas E (eds.) *Making Decisions under Stress: Implications for Individual and Team Training*. APA, Washington, DC, pp. 17–38

Entin E E, Serfaty D 1999 Adaptive team coordination. *Human Factors* **41**(2): 312–25
Foushee H C 1984 Dyads and triads at 35,000 feet: Factors affecting group process and aircrew performance. *American Psychologist* **39**(8): 885–93
Guzzo R A, Dickson M W 1996 Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology* **47**: 307–38
Ilgen D R 1999 Teams embedded in organizations: Some implications. *American Psychologist* **54**(2): 129–39
McIntyre R M, Salas E 1995 Measuring and managing for team performance: Emerging principles from complex environments. In: Guzzo R, Salas E, et al. (eds.) *Team Effectiveness and Decision Making in Organizations*. Jossey-Bass, San Francisco, CA, pp. 149–203
Orasanu J, Fischer U 1997 Finding decisions in natural environments: The view from the cockpit. In: Zsambok C E, Klein G (eds.) *Naturalistic Decision Making. Expertise: Research and Applications*. Lawrence Erlbaum, Mahwah, NJ, pp. 343–57
Salas E, Bowers C A, Edeus E (eds.) 2001 *Improving Teamwork in Organizations: Applications of Resource Management Training*. Erlbaum, Mahwah, NJ
Salas E, Cannon-Bowers J A 1997 Methods, tools, and strategies for team training. In: Quinones M A, Ehrenstein A (eds.) *Training for a Rapidly Changing Workplace: Applications of Psychological Research*. American Psychological Association, Washington, DC, pp. 249–79
Salas E, Cannon-Bowers J A 2000 The anatomy of team training. In: Tobias S, Fletcher J D (eds.) *Training and Retraining: A Handbook for Business, Industry, Government, and the Military*. Macmillan, New York, pp. 312–35
Turner M E (ed.) 2001 *Groups at Work: Theory and Research*. Erlbaum, Mahwah, NJ
Wiener E L, Kanki B G, Helmreich R L (eds.) 1993 *Cockpit Resource Management*. Academic Press, San Diego, CA, pp. 479–501

E. Salas and J. A. Cannon-Bowers

Technological Change in a Multicultural Context, Applied Psychology of

This article addresses a combined set of issues, namely technological change and multicultural contexts. Each issue has received extensive attention in its own right, but in combination there is comparatively little directly relevant literature. After a brief introduction to the two sets of issues, the article highlights three issues that will become increasingly important as areas of investigation in future years.

1. *Technological Change*

Technological change has a long history, and each new wave of technological development has resulted in changes to the nature of work (Wilpert 2000). Examples include the Industrial Revolution, developments in transport (air, land, and water), and most recently the information technology and communications revolution. The uptake of new technologies has not been uniform across the globe, and it is therefore difficult to speak of this without acknowledging the diversity of stages of technological change in different countries and cultures.

Industrialization has resulted in a particular form of work that emphasizes the division of labor, centralization of functions, and new forms of labor relations. The information technology revolution has added another layer of change with major implications for the location of work in time and space, and for the immediacy of communication irrespective of distance. Machine tooling provides an instructive example of the impact of technology. Early blacksmiths required carefully honed craft skills that involved the use of a range of tools. Lathe and mill technologies used by fitters and turners evolved into computerized numerical control (CNC) machines that have radically altered the skills required. Perceptual motor and hand–eye coordination skills were replaced with a need for the conceptual and cognitive skills required for writing numerical control code in order to program the machines. Retraining was not always possible because of the very different skill set that often required a high level of mathematics.

More recent developments of a similar nature include electronic data exchange (EDE), computer aided design (CAD), and computer supported cooperative work tools. Information technology is now pervasive with e-learning, e-commerce, and e-communication via the Internet rapidly replacing existing approaches. These developments have major implications for the nature of work. In general, many new jobs have been created, existing jobs have changed radically, and some jobs have disappeared. New jobs are usually more complex and require a broader set of skills.

Technological change of the sort described above is best understood within a broader work context. A useful framework has been provided by Emery (1969) who presents a model of organizations as consisting of interacting technological, social, and managerial components. The technological system includes all the equipment, infrastructure, and technology in the workplace. The social system includes individuals and groups and the social processes and informal channels used by these groups in communicating and negotiating in the workplace. The managerial system is concerned with the power and authority within the organization, and formal lines of communication. Cultural issues have been widely discussed within the social and managerial system, but to date there has been comparatively little attention to cultural issues within the technical system. Technological change has an obvious and direct impact on the technical system but also on the social and managerial systems, and hence on any cultural system.

2. *Multicultural Contexts*

Organizations and workplaces are increasingly multicultural. This has arisen because of globalization of industry, education, and many other activities, and increased migration and portability of populations. Developments in transport and communication technologies have provided an enabling basis for these trends. The workforce of large corporations is increasingly multicultural even within single organizations (Bochner and Hesketh 1994). Many large international firms have branches throughout the world, with a consequential requirement placed on their employees to deal with a variety of different cultures. These trends have led to the recognition of a need to understand different cultures as consumers and employees.

Triandis defines culture in terms of the group's characteristic way of perceiving its social environment (Triandis 1972, p. 3). This should be extended to the group's way of perceiving the technological and managerial environment. Values are fundamental to these perceptions, and Schwartz (1994) and Smith and Schwartz (1997) have made a major contribution to thinking in the area through their empirically based typology for evaluating environments cross culturally. Cultures differ in terms of: (a) Conservatism (collectivism), where the emphasis is on group goals, interdependence, and closely knit groups; (b) Hierarchy, where resources and roles are allocated by the hierarchy; (c) Mastery, where the emphasis is on self-enhancement and achievement; (d) Autonomy (individualism), with independence emphasized in the pursuit of individual goals; (e) Egalitarian commitment, where the welfare of others transcends self-interest; and (f) Harmony, for the world as it is and for the preservation of the environment. Nations can be characterized in terms of distinct cultural value pat terns or profiles that map on to a circumplex. Bochner and Hesketh (1994) present data that show how these types of value differences, particularly collectivism and individualism, can be found within one multicultural organization.

3. *Technological Change in a Multicultural Context*

Several themes will be used to illustrate the reason for examining technological change within a multicultural context. First, the ways in which technology and culture might contribute individually and interactively to performance and other outcomes will be examined. Second, the impact of information technologies on work organization will be reviewed with a particular emphasis on the implications of the new communication technologies for trust in a multicultural context. Finally, the importance of training for changing technologies and for working in multicultural contexts will be highlighted.

3.1 *The Technology–Culture Interaction*

Hesketh and Neal (1999), in a chapter on technology and performance, highlighted the confounded relationship between individual differences and technology in understanding the contribution of each to performance. Using a sporting analogy, they illustrated how racquet technology has interacted with individual differences in skill level to greatly enhance the performance capability of tennis players. Unfortunately it is difficult to determine how much of the improvement is due to the technology or to individual skill in interaction with the technology because of the confounded nature of the relationship.

The same type of analysis can be used to examine the contribution to performance of factors such as individual differences, technology and culture, and the various combinations of these. Gelfand and Dyer (2000) draw attention to the concept of the tightness or looseness of a culture. Tight cultures provide strong norms, with little opportunity for individual differences. Loose cultures are more likely to provide opportunities for a wider range of behavior and influences. Similarly, some technologies are extremely constraining, with resultant re-engineering that forces systems and cultures to fit their design criteria. Other technologies are more adaptable, providing the opportunity for individuals or cultures to tailor the technologies to fit their values. Using these analytical tools it should be possible to make predictions about the types of changes likely to emerge from technological developments in different multicultural contexts. Although this research is yet to be undertaken, current trends suggest an overwhelmingly strong main effect for technology that appears to be swamping cultural differences. For example, Jarvanpas and Leidner (1999) found no effect for culture in their research on trust in virtual work teams. In some respects technology is providing the common platform for communication within diverse multicultural groups. Even language translations are likely to be more instant and pervasive with future developments in the use of technologies.

3.2 *Technological Change and Work Organization*

The growth in communication technologies has made it possible to provide flexible work arrangements that are no longer restricted in time and space. These changes have major implications for the nature of supervision, monitoring of performance, and labor relations generally. Employees will need greater self-reliance and remote work self-efficacy (Staples et al. 1999).

Research into the newer forms of work organization is essentially descriptive. Current research has com-

pared various forms (telecommuting, teleworking, satellite work stations, neighborhood work centers, and traditional work arrangements) in terms of hours worked, satisfaction, perceived flexibility, and the impact on work and family balance. Telecommuting is a term that is used to describe a work arrangement where an employee is provided with a computer for use in an office at home. The extent to which this is used varies from arrangements that systematically substitute the home office for a work office for anything between one and five days per week, to simply extending work into 'non-work' hours at home. Teleworking covers a broader range of work arrangements including the use of a car as a mobile office, call centers, and home work. In a large US study, Cree and Sorenson (2000) found that teleworkers worked more hours than traditional workers (56.4 hours versus 47.1 hours). Because of the longer hours, teleworkers perceived that the arrangement had a negative impact on work and family balance. However, when hours worked was controlled, teleworking provided the hypothesized benefit for work–family balance.

Most new forms of work require greater self-direction, and potentially alter the modes of communication and informal hierarchies at work. Given the centrality of trust and absence of distrust for effective performance (Baba 1999), it is important to understand how technology affects relationships. Trust and distrust are aspects of social relations that are found in all societies, although their bases may differ vastly. The ingredients and behavioral manifestations of trust are culturally bound. Technology can either improve or interfere with trust depending upon previous experiences with the technology and the manner in which it is introduced. Jarvanpas and Leidner (1999) report on a preliminary study examining issues of communication and trust in global virtual teams. Virtual teams are permeable, form rapidly, and dissolve quickly when no longer needed. They are often located across time, space, and culture, and make extensive use of e-communication. Jarvanpas and Leidner (1999) grouped their cases according to the levels of trust at the start and later in the team lifespan. Those teams that started with high trust and maintained it were able to function most effectively. Establishing trust early was important. Many employees have developed strong distrust of new technology in part because it has changed the ways in which they work. In examples where the technological change includes the development of multicultural virtual work teams, considerable effort is needed to overcome cultural barriers and to develop trust as this has traditionally depended on personal relationships. Jarvanpas and Leidner (1999) found that groups able to maintain trust without social contact were better able to manage uncertainty and complexity. It appears very difficult to overcome early events that disrupt trust or produce distrust.

Technologies differ in terms of the extent to which they provide employees and groups with opportunities to obtain informational control, something that most groups seek. If trust is lacking, the efforts of the workgroup to gain some form of control over their work tasks may yield counterproductive outcomes. Excellent case examples are provided by Baba (1999).

3.3 Training for Technological Change in a Multicultural Environment

Developing methods of training for adaptable and transferable skills has become essential in the face of ongoing technological change. This need has resulted in a remarkable increase of research on training, much of it drawing on recent developments in cognitive psychology (Quinones and Ehrenstein 1996). A multicultural context also presents a major need for culture training. One needs to both skill and multiskill the workforce. Hesketh and Bochner (1994) reviewed issues associated with culture and training. The chapter addresses issues linked to the impact of social change, the changing ethnic composition of work, work-related contacts between societies, a decrease in work-related discrimination, an aging workforce, and attitudes toward change. Training now needs to develop multiple flexibility, including flexibility in response to continuing technological change and flexibility in working with colleagues from a variety of cultural and religious backgrounds. A key issue relates to the development of trust sufficient to endure the many new forms of work. In achieving this, one aim of training is to develop a shared mental model about whatever task or negotiation is being undertaken, including an understanding of the cultural differences in the group. The early development of a shared mental model helps reduce uncertainty, and allows each individual to predict how others in the group might react. Trust is a component of such mental models in that trust allows one to have expectations regarding the behavior of another. Violation of expectations about behavior can often be the start of the development of distrust which is difficult to overcome. A deep understanding of cultural differences facilitates the development of accurate mental models that underlie trust, particularly where the tasks and communications include a multicultural group. Where ongoing interactions are required, then shared mental models are needed to increase trust and communication.

4. Summary and Comment

Technological change has resulted in major and ongoing change to work organization in a multicultural context. To date there is comparatively little empirical data that is useful in making effective use of

technology in this context, although several of the papers listed in the bibliography provide good analyses of the difficulties and issues. The very nature of the technology makes it possible to do research of a different form with the availability of possibilities for electronic capture of communication and output, and on-line surveys of attitudes and reactions. Such research and any forms of electronic monitoring will need to be handled in ways that maintain trust and provide employees with a sense of control.

See also: Industrial and Organizational Psychology: Cross-cultural

Bibliography

Baba M L 1999 Dangerous liaisons: trust, distrust and information technology in American work organisations. *Human Organisation* **58**(3): 331–46

Bochner S, Hesketh B 1994 Power distance, individualism/collectivism, and job-related attitudes in a culturally diverse work group. *Journal of Cross-Cultural Psychology* **25**(2): 233–57

Cree L H, Sorenson S 2000 Effects of telecommunting on total hours worked, flexibility, work and family balance and performance. Paper presented at the XXVII International Congress of Psychology. Stockholm, July 23–28, 2000

Emery F E 1969 *Systems Thinking*. Penguin, Harmondsworth, UK

Gelfand M J, Dyer N 2000 A cultural perspective on negotiation: progress, pitfalls and prospects. *Applied Psychology: An International Review* **49**(1): 62–99

Hesketh B, Bochner S 1994 Technological change in a multicultural context: implications for training and career planning. In: Triandis H, Dunnette M, Hough L (eds.) *Handbook of Industrial and Organizational Psychology* 2nd edn. Consulting Psychologist Press, Palo Alto, CA, Vol. 4

Hesket B, Neal A 1999 Technology and performance. In: Plulakos E, Ilgen D (eds.) *The Changing Nature of Work Performance: Implications for Staffing, Motivation and Development*. Society for Industrial and Oranizational Psychology New Frontiers Series, Jossey-Bass, San Francisco

Jarvanpas S L, Leidner D E 1999 Communication and trust in global virtual teams. *Organisation Science* **10**(6): 791–815

Quinones M A, Ehrenstein A (eds.) 1996 *Training for a Rapidly Changing Workplace: Applications of Psychological Research*. American Psychological Association, Washington, DC

Schwartz S H 1994 Beyond individualism/collectivism: new cultural dimensions of values. In: U. Kim H C, Triandis C, Kagitcibasi S, Choi C, Yoon G (eds.) *Individualism and Collectivism: Theory, Method and Applications*. Sage Publications, Thousand Oaks, CA, pp. 85–122

Smith P, Schwartz S H 1997 Values. In: Berry J W, Segall M H, Kagitcibasi C (eds.) *Handbook of Cross-Cultural Psychology*. Allyn & Bacon, Needham Heights, MA, Vol. 3, pp. 77–118

Staples D S, Hulland J S, Higgins C A 1999 A self-efficacy theory explanation for the management of remote workers in virtual organisations. *Organisational Science* **10**(6): 758–76

Triandis H C 1972 *The Analysis of Subjective Culture*. John Wiley, New York

Wilpert B 2000 Applied psychology: past and future societal and scientific challenges. *Applied Psychology: An International Review* **49**(1): 3–22

B. Hesketh

Technological Determinism

'Technological determinism' refers to a pervasive, yet controversial, theory about the relationship between technology and society. Although the term has had a variety of meanings, two related claims have been central to discussions of this topic: (a) the development of technology proceeds in an autonomous manner, determined by an internal logic independent of social influence; and (b) technological change determines social change in a prescribed manner (Staudenmaier 1985, Misa 1988, Bimber 1994). The claims address two major questions about technology: how and why is technology developed and what is the relationship between technological change and social change? Arguments against the first claim have been a staple of research in the history and sociology of technology since the 1960s. The second claim has been at the center of debates about Karl Marx's theory of history since the early twentieth century. Weaker versions of technological determinism, sometimes called 'soft determinism,' maintain that technology is a major cause, but not the sole determinant, of social change.

1. History of the Concept

It is important to distinguish between widely held beliefs that an autonomous technology is the primary cause of social change and the use of the term 'technological determinism.' In the United States, the Enlightenment doctrine that improvements in the mechanic and industrial arts fostered social progress was replaced over the course of the nineteenth and early twentieth centuries by a more technocratic concept of progress, in which improvements in technical and economic efficiency became ends in themselves, instead of the means for creating desired social ends. Through the writings of the economist and social theorist Thorstein Veblen and others, the word 'technology' gradually displaced phrases like 'mechanic arts' in scholarly and then popular discourse in the twentieth century. In this period, the older meaning of 'technology' as systematic knowledge of crafts and industry was joined to the anthropological one of the artifacts and related practices of a culture. The term's connotations of abstractness and scientific objectivity served to reinforce the idea that technology was an autonomous social force.

These meanings entered popular culture through such media as utopian novels, exhibits at World's

fairs, the Technocracy movement in the 1930s, and company advertisements that endlessly touted new products as the cure for social ills and the source of all happiness. The organizers of the 1933 Chicago World's Fair encapsulated the gender biased, applied science aspect of technological determinism with the motto, 'Science Finds, Industry Applies, Man Conforms.' Ironically, twentieth-century critics of technology's harmful effects, such as technological unemployment and environmental degradation, have tended to reify the notion that technology is an autonomous agent of change in their attempts to control it (Smith 1994, Marx 1994).

The term 'technological determinism' is more recent, with roots in the turn-of-the century debates about Marx's theory of history. While Europeans used the phrases 'historical materialism' and 'economic determinism' to describe Marx's theory, many American social scientists employed the broader phrase, the 'economic interpretation of history,' popularized in a non-Marxist manner at the turn of the century by Edwin R. A. Seligman, editor-in-chief of the first edition of the *Encyclopaedia of the Social Sciences*. Historian Werner Sombart in Germany and economist Alvin Hansen in the United States argued that Marx's theory should more properly be called a 'technological interpretation of history' (a phrase which Hansen seems to have coined). Although many scholars questioned this view, prominent Soviet theorists established a 'technicist' version of historical materialism as the dominant interpretation in orthodox Marxism during the first half of the century (Bober 1927, MacKenzie 1984, Miller 1984).

During the early Cold War, social scientists and historians, who were increasingly concerned about how to control the effects technology apparently was having on daily life, began using the starker term, 'technological determinism' to criticize Marxist theories of technology and society. They also used the term to criticize the controversial views of such authors as Veblen, historian Lynn White, Jr, and the philosopher Jacques Ellul. White and Ellul said their views were not a strict determinism. Only a few authors, such as economist Robert Heilbroner (1967), have labeled their position a form of 'technological determinism' (soft determinism in Heilbroner's case). Philosopher Langdon Winner criticized scholars for dismissing the issue of technology's social effects because of the flaws in the concept of technological determinism. Saying that the 'idea of determinism is not one that ought to be rejected out of hand,' Winner argued that technology does constrain human activities (Winner 1977, p. 77).

By the mid-1970s, the issue of technological determinism had become a central topic in both Marxist scholarship and in the interdisciplinary study of science, technology, and society (Cohen 1978, Staudenmaier 1985). In the recent turn toward social constructivism in the latter field, the issue of technological determinism remains a focal point around which to criticize a theory of technology and society that is still prevalent in other disciplines and in popular culture.

2. *Was Marx a Technological Determinist?*

Recent scholarship is sharply divided about whether or not Karl Marx's theory of history, the *locus classicus* of technological determinism, is, indeed, technological determinism. Much of the debate centers on the meaning of an often-quoted excerpt from the preface to Marx's *A Contribution to the Critique of Political Economy* (1859): humans enter into 'relations of production [*Produktionsverhältnisse*] which correspond to a definite stage of development of their material productive forces [*Produktivkräfte*]. The sum total of these relations of production constitutes the economic structure [of society], the real basis on which arises a legal and political superstructure' (quoted in Cohen 1978, p. 28).

Those who argue for a technological interpretation of Marx usually equate 'forces of production' with 'technology' and interpret the verb 'correspond' to mean that the relations of production evolve to maximize the productivity of the forces of production. William Shaw, who includes instruments, raw materials, and labor power (including knowledge) in the category of productive forces, calls Marx's 'technological determinism' more properly a 'productive-force determinism.' Shaw acknowledges that Marx allowed for interplay between forces and relations of production, but maintains that productive forces are the 'long-run determinant of historical change' (Shaw 1979, p. 160). Gerald Cohen defines productive forces in a similar, broad manner, but rejects the label of 'technological determinism' in favor of a '"technological" interpretation of historical materialism.' In this functionalist view, social structures adapt to technological change to increase productivity (Cohen 1978, p. 29).

Those who argue that Marx was not a technological determinist typically have concentrated on the manner in which he wrote history. Harry Braverman, Nathan Rosenberg, and Donald MacKenzie point to the dialectical aspects of Marx's writings, especially where he attributed the development of productive forces to social relations. Indeed, MacKenzie characterizes Marx's treatment of the history of machinery in the Industrial Revolution as an 'attempt to develop a theory of the social causes of organizational and technical changes in the labor process' (MacKenzie 1984, p. 492). Richard Miller argues that Marx was not a technological determinist because in his historical work, 'economic structures do not endure because they provide maximum productivity. Productive forces do not develop autonomously. Change in productive forces, in the narrowly technological sense

that excludes work relations, is not the basic source of change in society at large' (Miller 1984, p. 188). Bruce Bimber, who defines technology narrowly as artifacts, argues that, for Marx, human characteristics of accumulation and self-expression drive the forces of production. Thus Marx was an economic, not a technological, determinist (Bimber 1994).

3. *Varieties of Technological Determinism*

Bimber and Thomas Misa effectively analyze varieties of technological determinism. Bimber distinguishes between normative, nomological, and unintended-consequences accounts (Bimber 1994). Normative accounts, evident in the writings of such critics of technology as Lewis Mumford, Jacques Ellul, and Jurgen Habermas, claim that society is relinquishing control over technology, or replacing political and ethical norms with those of the technologist's goals of efficiency and productivity. Nomological accounts, such as that by Heilbroner (1967), evoke the two-part definition of technological determinism given above: technology develops autonomously according to an internal logic and forces a prescribed social change. In unintended consequences accounts (e.g., Winner 1977) technology produces unpredictable social change, a view that challenges determinism, but reinforces the idea that technology is out of control. Bimber argues that the term 'technological determinism' should be reserved for nomological accounts, which are the only ones that satisfy his strict definition of the term.

In his survey of scholarship about the relationship between technology and social change, Misa (1988) found that sweeping, macro-level accounts were more likely to be technologically determinist (using the two-part definition of the term) than micro-level accounts, which focused on specific practices. Philosophers, business historians, urban historians, and historians of physical sciences were generally more deterministic than technological historians and labor historians.

Another form of technological determinism is what Claude Fischer (1992) criticizes as the 'impact-imprint' model of technology. Rather than investigating how consumers use a technology, scholars assume that a technology's capability (e.g., the ability of the telephone to enable people to talk at a distance) leads to a predictable 'impact' (e.g., new, long-distance communication patterns).

4. *Political Artifacts and Systems*

Winner and Thomas Hughes have proposed influential theories of technology and society that address the question of technological determinism. Winner (1986) argues that artifacts have politics (i.e., are associated with specific social changes) in two ways. Technologists can design a (plastic) artifact to promote a certain type of politics, as when Robert Moses built low bridges over the Long Island Expressway in order to prevent inner-city buses (carrying poor minorities) from traveling to Jones Beach. Secondly, the design and building of some systems seems to require a certain type of politics. Nuclear power plants, for example, require a centralized form of government to run the reactor and prevent the proliferation of nuclear weapons. Hughes (1994) states that his concept of technological systems allows for the social construction of a system when it is young, but technological determinism comes into play once it is established and shapes society more than being shaped by society.

5. *Social Construction of Technology*

The advent of the approach known as the social construction of technology (SCOT) in the mid-1980s (Pinch and Bijker 1987), along with a contextualist history of technology (Staudenmaier 1985), has challenged the technical aspect of technological determinism (the first part of the definition given above). Sociologists and historians of technology have published numerous case studies of the development of technologies that argue that their design was the outcome of negotiations between several social groups (such as inventors, engineers, managers, salespersons, and users), rather than the product of an internal, technical logic (see *Technology, Social Construction of*). When Winner (1993) and others criticized this approach for leaving out power relations among social groups, constructivists responded by investigating the social aspect of technological determinism (the second part of the definition). Bijker (1995) employed a semiotic version of Michel Foucault's 'micropolitics of power' to show the mutual construction of artifacts and social groups. Ronald Kline and Trevor Pinch (1996) utilized theories of gender in the original SCOT approach to show how gender relations and the automobile were mutually constructed when farm men in the United States before World War II adapted the car, as a stationary and mobile source of power, to a variety of novel uses, such as grinding grain, washing clothes, and plowing fields. Fischer (1992) employed a similar user-heuristic to argue that callers in the United States used the telephone before 1940 to widen and deepen existing communication patterns, rather than to create new ones.

6. *Conclusion*

Although historians and sociologists of technology have discredited the tenet of technological determinism, so much so that it has become a critic's term and a term of abuse in their academic circles, the idea that an autonomous technology drives social change per-

vades other fields of scholarship and popular culture. This is especially evident in discourses claiming that new information technologies have created an information society in the United States and Europe. It remains to be seen if this scholarship will overturn the widely held belief in forms of technological determinism.

See also: Technology, Social Construction of

Bibliography

Bijker W E 1995 *Of Bicycles, Bakelite, and Bulbs: Towards a Theory of Sociotechnical Change*. MIT Press, Cambridge, MA

Bimber B 1994 Three faces of technological determinism. In: Smith M R, Marx L (eds.) *Does Technology Drive History? The Dilemma of Technological Determinism*. MIT Press, Cambridge, MA

Bober M M 1927 *Karl Marx's Interpretation of History*. Harvard University Press, Cambridge, MA

Cohen G A 1978 *Karl Marx's Theory of History: A Defence*. Princeton University Press, Princeton, NJ

Fischer C S 1992 *America Calling: A Social History of the Telephone to 1940*. University of California Press, Berkeley, CA

Heilbroner R L 1967 Do machines make history? *Technology and Culture* **8**: 335–45

Hughes T P 1994 Technological momentum. In: Smith M R, Marx L (eds.) *Does Technology Drive History? The Dilemma of Technological Determinism*. MIT Press, Cambridge, MA

Kline R R, Pinch T J 1996 Users as agents of technological change: The social construction of the automobile in the rural United States. *Technology and Culture* **37**: 763–95

MacKenzie D 1984 Marx and the machine. *Technology and Culture* **25**: 473–503

Marx L 1994 The idea of 'technology' and postmodern pessimism. In: Smith M R, Marx L (eds.) *Does Technology Drive History? The Dilemma of Technological Determinism*. MIT Press, Cambridge, MA

Miller R W 1984 *Analyzing Marx: Morality, Power, and History*. Princeton University Press, Princeton, NJ

Misa T J 1988 How machines make history and how historians (and others) help them to do so. *Science, Technology, and Human Values* **13**: 308–31

Pinch T J, Bijker W E 1987 The social construction of facts and artifacts: Or how the sociology of science and the sociology of technology might benefit each other. In: Bijker W E, Hughes T P, Pinch T J (eds.) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, MA

Shaw W H 1979 'The handmill gives you the feudal lord': Marx's technological determinism. *History and Theory* **18**: 155–66

Smith M R 1994 Technological determinism in American culture. In: Smith M R, Marx L (eds.) *Does Technology Drive History? The Dilemma of Technological Determinism*. MIT Press, Cambridge, MA

Staudenmaier J M 1985 *Technology's Storytellers: Reweaving the Human Fabric*. MIT Press, Cambridge, MA

Winner L 1977 *Autonomous Technology: Technics-out-of-control as a Theme in Political Thought*. MIT Press, Cambridge, MA

Winner L 1986 Do artifacts have politics In: Winner L *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press, Chicago

Winner L 1993 Upon opening the black box and finding it empty: Social constructivism and the philosophy of technology. *Science, Technology, and Human Values* **18**: 362–78

R. R. Kline

Technological Innovation

Technological innovation is the successful implementation (in commerce or management) of a technical idea new to the institution creating it. Innovations are distinguished from inventions, technology and research, but may arise from any of the three. A variety of models of the innovation process are described, for they are useful in developing public policies for encouraging innovations as well as for managing their creation. The more advanced of these models include consideration of complementary assets and social capital, which helps explain the differences in innovative capacity in different societies. The American society is particularly given to the use of banners under which to rally public opinion to the advance of economic well being. In the middle 1970s, when 'high tech' industries emerged as the key to growth, and American firms were immediately challenged by the technically adroit Japanese, the banner was 'critical technologies' derived from defense and space research. When the economic challenge became serious in the 1980s and early 1990s the banner was 'competitiveness'; even conservative President Reagan launched a White House taskforce to suggest how government could enhance American competitiveness in the face of serious price and quality competition in technology intensive industries, especially in Asia. As we prepare to enter the next millenium, the new banner, in nations rich and poor, is 'innovation.'

1. Definitions

While a firm can become more competitive by cornering a market or slashing worker's wages, innovation implies a transformation in the market—the invocation of imagination and daring in the adoption of new ways of doing things. The word innovation has an old history. The Oxford English Dictionary uses a broad definition: 'A change in the nature or fashion of anything; something newly introduced, a novel practice, method, etc.' and traces its first use back to 1553. In its contemporary usage, Burke is quoted as writing in 1796 'It is a revolt of innovation; and thereby the

very elements of society have been unfounded and dissipated.' In modern usage we must distinguish innovation from invention, technology, and scientific or engineering research. Invention is the conception of a new artifact or process that is useful, original, and non-obvious. Technology is the capability to perform a technical task or process. Research leads to understanding of how things work. Innovation, as defined by Richard Nelson (1993), is 'the processes by which firms master and get into practice product designs and manufacturing processes that are new to them…' Thus, an invention does not become an innovation (and many never do) until the invention is successfully embodied in a product and introduced to the market.

Even then, an innovation does not necessarily make a successful contribution to the economy. As David Teece (1987) observed, many products that embody little novelty are successful in the market place, while innovations that are not supported by a variety of 'complementary assets' may fail. The complementary assets of an innovation are all the requisite ingredients of success that are not provided by the innovator and may not be readily available in society because of the novelty inherent in the innovation. Thus, the first personal computers became wildly successful only after a cottage industry provided a wide variety of software applications. A new drug cannot make money for its innovator until the manufacturer receives approval from government regulators.

Nelson's definition requires that an innovation be new, at least to the innovator, and that it must be in production and use, but it does not have to embody an invention or be the result of any research. Thus, many of the innovations with the greatest impact on society were marketing, institutional, or managerial innovations. The Sears Roebuck catalog revolutionized marketing to families in a predominantly agricultural society. The corporate research laboratory that flowered at GE, DuPont, and ATT early in the twentieth century provided a more productive way to create useful technology. Just-in-time manufacturing is a Japanese managerial innovation that reduces cost and accelerates product cycles and quality control. This article focuses on technical innovation—that class of innovations based on technical novelty.

2. *Technical Innovation Models*

The popular imagination is excited by the creation of new firms by the marriage of daring investors with the dreams of inventors and other would-be innovators. The US economy is noteworthy for the dynamism of its venture capital industry, and new research breakthroughs often capture the imagination with their economic potential. Thus, when materials that could conduct electricity without any resistance and do so at or near room temperature were discovered (and the inventors were rewarded with Nobel Prizes) the potential to destabilize the entire electrical machinery industry was in everyone's mind. After over a decade of dashed hopes, finally a company has mastered a technology to make superconducting motors that operate at temperature far above liquid helium temperature. They have a contract to sell such motors to the US navy.

The innovation process that follows such a discovery is described by the time honored, but oversimplified 'linear model' of innovation, in which technical innovations begin with a scientific discovery or an invention. The idea is then explored through a period of applied research. When the technology is sufficiently understood, the work is transferred to product and process development, followed by production, quality control, distribution and marketing. This is a reasonable description of one way a new invention arising from scientific research may be exploited. Scott Shane (2000) has shown that the most radical or fundamental inventions patented and licensed by MIT are most likely to be successfully exploited in a new firm, created by investors for the purpose. Such radical innovations often must create their own markets. However, technical innovations do not always arise from a scientific breakthrough; they are quite frequently the response of technically clever people to an identified market need. This is reflected in the saying 'necessity is the mother of invention.' The transistor was a discovery in search of a more economical switch for telephone exchanges, and the integrated circuit was the answer to the excessive power consumption, size, and manufacturing cost of computers made of discrete electrical elements. In these examples, as in many other cases of science-based innovations, the new capability not only solves the identified need, but does much more, creating new markets or even entirely new industries. Innovations are compounded both of answers looking for questions and questions looking for answers. In the real world both are intimately linked. This fact gives rise to the 'chain link' or 'spiral' models of innovation in which new ideas are stimulated by new market needs, which in turn are extended when the capabilities of the innovation are better understood (Kline and Rosenberg, 1996). Thus, it is rare that a technically radical innovation is introduced into a totally new market. The more typical trajectory is interative. First a new market is found for an existing technology (a market innovation); then new technology is brought to this new market (a technical innovation), and so forth.

At the level of product development, an analogous interaction takes place, between research to understand the new technology and product development to define product performance and cost. Thus when an invention is being reduced to commercial practice, one must know what the product specifications (function and cost) will be in order to know what tests to apply. But function and cost cannot be known until the

technology is developed. This iterative process describes almost all incremental or evolutionary innovation in existing companies.

3. *Innovation as a Socioeconomic Process*

Futurists have often called attention to the remarkable fact that progress in meeting many of the needs of society does not improve linearly with time, but rather exponentially. Thus, if one plots on semi-log graph paper the speed of commercial transportation against the year of introduction, from the horse-drawn trolley, to the train, the airplane, and the rocket, one gets a surprisingly straight line. The same observation is inherent in Moore's Law, which observes that the density (and hence speed divided by cost) of integrated circuits has doubled every 18 months for many years and continues to do so (Lester 1998). What accounts for this remarkable attribute of technical innovations? It is not inherent in the nature of technical progress, but rather is the result of the socioeconomic dynamics of an economy that rewards innovation. Consider the typical time history of a radical innovation arising, let us say, from new university research. After a promising discovery there is typically a period of intense excitement. The most daring investors leap into the business, hoping to acquire that 'first mover' advantage and intellectual property protection that the economist Schumpeter recognized as necessary if a firm hoped to gain enough profit to repay the risks and costs of innovation. Then after perhaps 5 to 7 years the recalcitrance of nature becomes widely evident; the low hanging fruit has been plucked, and the number of firms decreases sharply—through failure, merger, or switch to a more conventional product opportunity. Those who persevere in pursuit of commercial advantage from the discovery and have the requisite financial staying power and the technical skill and knowledge may then master and reduce to practice the technology required for the innovation. The business may then take off with exponential growth as the technology matures, costs fall, and new applications for the technology are found. Investors and inventors alike observe the exponential growth of the business and invest their money and their imaginations in the expectation that exponential growth will continue.

Sooner or later the chosen technical approach becomes increasingly difficult to improve, given the limitations imposed by real materials and the laws of natural science. Continued rate of performance growth and cost reduction slows; the technology is reaching its natural limits. The curve of progress vs. time starts to become S shaped. The gap between expectation and reality begins to widen, and a huge incentive to find another technology that can satisfy the future expectation for improvements in product function and cost is created. Because science has made such extraordinary advances in the last half century, the potential for invention is enormous. It is the reward of filling the expectation gap that converts that potential to another innovation. Thus, another S-shaped curve becomes tangent to the first one, and the exponential growth in capability and in economic activity continues unabated. In this way performance advances seamlessly from horse to auto, to airplane, to rockets. Similarly Moore's Law continues to be valid despite evolution from vacuum tubes to single transistors, to integrated circuits of every increasing complexity.

4. *National Systems of Innovation: Social Capital*

Because innovation is a socioeconomic as well as a technical process, one might expect different societies to exhibit different strengths and weaknesses in their capacity for innovation. Indeed this is true, as observed by many students of comparative innovation policy. As Patel and Pavitt (1994) pointed out, the economic theory prevailing in the 1960s predicted that buoyant demand and an open trading system would allow the international (and domestic) diffusion of technology and this would lead to equalization of technological performance of the industrial nations at the national level. This prediction, they argued, was based on a flawed model of science-based development, indeed of technological change. It presupposed that technical change would be fully determined by investments in better machinery, the diffusion of public technical information, and the acquisition of 'tacit' knowledge resulting from relatively costless 'learning by doing.' If this model were equally applicable to all countries in a similar state of development, it would follow that the gaps between the US economy, and that of Japan, the United Kingdom, Germany, France should have closed rather rapidly. It has not happened. Japan and Germany have moved ahead; the UK and France have fallen behind. Taiwan, Korea, and Singapore have leapt ahead from a very backward state 30 years ago. Brazil, Mexico, and India, all in better economic health than the Asian Tigers at the end of World War II, have failed to do so to the same extent. These three assumptions in the 1960s economic model have persisted, in more recent years, in the form of a Cold War paradigm of the innovation process. It was widely believed in the US government (as well as in the UK and France) that one could gain commercial advantage by force-feeding industrial technical development in pursuit of government missions. This is the 'spinoff' theory of defense industrial research. Defense research investments, under this theory, would bring into existence new machinery, abundant and accessible new technical knowledge, and the private sector could quickly assimilate the embedded knowledge by 'learning by doing.' The problem with the three post-war economic assumptions is that the efficiency of each of these processes varies strongly from one national

socioeconomic setting to another. Patel and Pavitt conclude that technology diffusion, productivity learning, and transfer of embodied technology have large transaction costs when they encounter cultural, managerial, and institutional barriers. It is the absence of these barriers and the presence of efficiently functioning networks connecting innovators, investors, producers, supporting service providers and consumers that largely govern the benefits that flow from technical innovations. These attributes of the economy are referred to as 'social capital.' Richard Nelson's ambitious study entitled *National Systems of Innovation* (Nelson 1993) examined the policies and institutional settings for innovation in 14 countries. It is clear that the structure of economies, socioeconomic policies, cultural factors, and natural endowments all influence the environment for innovation. In addition, Nelson recognizes that innovation systems may not be national, but rather sectoral or regional. For example, the world oil industry shows little variation in its innovation practices from country to country, while it differs significantly from other industries in all countries. The levels of trust that exist among entrepreneurs, the financial sector, government and industry leaders are as influential in the capacity of a society to innovate as is the role of education and the participation of higher educational institutions in the nation's research and invention activities. As firms turn away from vertical integration and increasingly look to global sources of innovation to keep them competitive, issues of social capital will become more important. The methods of microeconomics are still essential but are no longer sufficient to understand why some societies are more innovative than others. Thus the importance of social capital forces the study of innovation beyond the boundaries of the firm into an understanding of the complex relationships among interdependent players in a high-tech economy.

Henry Ergas addressed this problem by defining a model of the innovation process at the level of the economy, rather than of the firm (Ergas 1989). His model is also a dynamic one, in that it incorporates the response of the economy to the introduction of an innovation, which in turn will affect the next generation of the innovation. Ergas sees four parts to the innovation process:

(a) *Generation*—all the activities from technical and market concepts to introduction into manufacturing, in other words virtually all of the activity calling for research and development skills.

(b) *Application*—commercialization: manufacturing learning, introduction of a complete product line and system of distribution, up to initial customer acceptance.

(c) *Verticalization*—the responses of suppliers, distributors, providers of complementary assets, and of competitors to the introduction of the innovation, responses that will, in turn, change the innovation and the way it is received.

(d) *Diffusion*—the responses of the rest of the socioeconomic environment, especially the way the innovation changes the end user's behavior. One can then expect user organizations, and perhaps society as a whole, to be transformed to the degree the innovation has that power (as for example the personal computer accomplished). Society then may change regulations and laws, introduce new curriculae in schools, and eventually modify the culture. Thus the next round of innovation takes place in a new environment calling for still more creativity.

5. *Role of Governments*

Fostering innovation and entrepreneurship is not a natural role for governments. As Nathan Rosenberg and L. E. Birdzell, Jr. (1985) recognized in their classic book, *How the West Grew Rich*,

> In all well-ordered societies, political authority is dedicated to stability, security, and the status quo. It is thus singularly ill-qualified to direct or channel activity intended to produce instability, insecurity and change.

Nevertheless, promoting innovation and entrepreneurship is increasingly seen as a pro-active responsibility of government. Nations that once relied on science and technology policy, by which they usually meant public investments in scientific research, came to realize that this supply-side policy would probably be ineffective without demand-side policies to put those ideas to work. Thus, there is now widespread acceptance of the idea that Science and Technology Policy should be replaced by a Research and Innovation Policy (Branscomb and Keller 1998).

The goals of an innovation policy are clear enough, even if the means for their attainment are not: full employment, decreased income disparities, rising real incomes, safe, healthy working conditions, and satisfying work. What policy tools, other than investing in research, can one use to encourage innovation? The list is long: economic policy, trade policy, tax policy and accounting rules, competition policy, regulation of equity markets and industry structure, intellectual property law, industrial relations, education and training, and the full range of environmental, health, and safety regulations.

This list of policy tools would seem to offer states a lot of options. In fact these policies must also be seen as a set of potential barriers to innovation, set up one behind the other. Any one of them, if sufficiently mismanaged, can seriously discourage would-be innovators. It might seem obvious that all these economic issues should be dealt with by economic policy makers who should appreciate the fragility of an economy increasingly based on science-based innovation. But science policy experts have for so long insisted (erroneously) that innovations happen auto-

matically as a result of research, that economists have mostly left these issues to 'science policy.' And science policy does not enjoy a priority on political agendas any where near economic, trade, tax, or regulatory policy. Furthermore, only a few of these policy tools (intellectual property law is an example) can be used to create a positive, specific incentive to innovation; most exist for other reasons and must be examined to ensure that suppression of innovation is not an unwanted side effect. To ensure that every government agency is alert to this responsibility, it may be most effective to entrust a policy body at or near the top of government with monitoring agency diligence. The challenge to governments is formidable, not only for the reasons given by Rosenberg and Birdzell but because, as discussed earlier, building an innovative society requires developing the social capital of society, a task that most governments perform poorly or are reluctant to even attempt. Thus, whatever the quality of the social capital, innovations must still come from the creativity, imagination, and determination of individuals who want to change the way things are.

See also: Evolution: Diffusion of Innovations; Innovation and Technological Change, Economics of; Innovation: Organizational; Innovation, Theory of; Technological Change in a Multicultural Context, Applied Psychology of; Technology Assessment; Technology, Anthropology of; Technology Transfer

Bibliography

Branscomb L M, Keller J 1998 Towards a research and innovation policy. In: Lewis M, Branscomb L M, Keller J (eds.) *Investing in Innovation: Creating a Research and Innovation Policy that Works*. National Academy Press, Cambridge, MA, pp. 462–496.

Ergas H 1989 Global Technology and National Politics unpublished. Quoted by permission in Young Hwan Choi Y-H, Branscomb L M 1996 A new technology-based innovation model: The role of government and national strategy. In: Lewis M, Branscomb L M, Choi Y-H (eds.) *Korea at the Turning Point: Innovation-based Strategies for Development*. Praeger Press, Westport, CT, p. 202.

Kline S, Rosenberg N 1996 An overview of innovation. In: Landau R, Rosenberg N (eds.) *Positive Sum Strategy: Harnessing Technology for Economic Growth*. National Academy Press, Washington, DC, pp. 275–36.

Landau R, Rosenberg N (eds.) 1986 *The Positive Sum Strategy*. National Academy Press, Washington DC

Lester R K 1998 *The Productive Edge: How US Industries are Pointing the Way to a New Era of Economic Growth*. Norton, New York, p. 116

Nelson R R (ed.) 1993 *National Innovation Systems: A Comparative Analysis*. Norton, New York, p. 4.

Patel P K, Pavitt K 1994 The nature and economic importance of national innovation systems. *STI Review* **14**: 9–32.

Rosenberg N, Birdzell L E Jr. 1985 *How the West Grew Rich: The Economic Transformation of the Industrial World*. Basic Books, New York, p. 265.

Scherer F M 1999 *New Perspectives on Economic Growth and Technological Innovation*. The Brookings Institution Press, Washington, DC

Schumpeter J 1943 *Capitalism, Socialism, and Democracy*. Harper and Row, New York

Shane S 2000 Technology regime and new firm formation. In: Branscomb L M, Morse K P, Roberts M J (eds.) *Managing Technical Risk: Understanding Private Sector Decision Making on Early Stage, Technology-based Projects*. (Report from NIST-ATP Project 471-1608, US Department of Commerce, Washington, DC, pp. 121–38.

Teece D J 1987 Capturing value from technological innovation: Integration, strategic partnering, and licensing decisions. In: Guile B R, Brooks H (eds.) *Technology and Global Industry: Companies and Nations in the World Economy*. National Academy Press, Washington, DC, pp. 65–95.

Utterback J M 1994 *Mastering the Dynamics of Innovation*. Harvard Business School Press, Boston

L. M. Branscomb

Technology and Organization

The term 'technology' is in common everyday use. It may be popularly visualized in purely physical terms as 'equipment,' 'apparatus,' 'pieces of hardware and software,' and it may be defined by different commentators in different ways. Whatever the case may be, what are much more serious than definitional problems are the practical implications of the development, implementation, and control of technology by *a few* for the use of *the many* in an organization and in wider society. This article addresses the issue of technology and the relationship between technology and 'organization.'

1. Organizing and Organization: An Overview

To initiate a discussion on the relationship between organization and technology it is important first to define what is meant by each of the two terms. First, what is meant by the term 'organization'? Well, to organize something is to 'systemize' it for a certain desired end. 'Organizing' a group of people is a process that aims to give an orderly structure to that group with a view to achieving a goal or fulfilling an objective in mind. In such a process the group in question acquires, or is brought into, a state of being 'organized.' This definition is quite a broad one; it is meant to be so since an encompassing conceptualization of the term is required to grasp the range of possible positions and influences of technology in and on organizations.

From an organizational studies perspective, organizations are defined as social arrangements for the controled performance of collective goals (see Huczynski and Buchanan 1991). In a broad sense, therefore,

examples of an organization may include an African tribe, Juventus Football Club, a consumer electronic goods manufacturer, the European Union, or the University of Harvard.

2. *Defining Technology*

Technology has received increased attention since the late twentieth century, because of the pace of technological advancement and the magnitude of its diffusion in society. Indeed, it is likely to impose itself as a pivotal organizational issue more forcefully than ever before, because of the pervasiveness of 'new' technology and its effects on organizing and on the structure and functioning of organizations as well as the behaviour of individuals and groups within them. Examples of new technologies include information technology, biotechnology, materials technology, space technology, and nuclear technology.

But what is technology? Well, again there is no single answer to this question, although to a layperson it is merely 'equipment,' 'apparatus,' or 'pieces of hardware and software.' Even amongst commentators in the field of organizational studies differences can be found as to how technology is defined. Regardless of the way it is defined, however, technology is an important factor in organizing and organizational life. A suitable working definition for the purpose of this article is that, in essence, technology is a man-made creation of a *system* that is supposed to work logically and consistently and is made operational to produce a desired accomplishment that is valued by its inventor or user. Such an accomplishment is often reproducible, through the system, to a given set of performance criteria or precision.

The creation of such a system involves translating some pure science knowledge into an *input–conversion–output* system format based on engineering principles and composed of elements which are functionally arranged or configured so as to yield a certain output as specified by the designer or user (see Clark et al. 1988). According to this definition, technology is defined not only in terms of its physical objects but also in terms of its methods of operation, use, or production and the technical know-how needed to apply them successfully. Human intervention in the operation of a system usually takes the form of monitoring the system's conversion process and controling its output through a specific feedback mechanism.

3. *Conceptualization and Meaning: From Neutral Rationality to Political Sensitivity*

Conceptualization of technology merely in terms of 'equipment,' 'apparatus,' 'pieces of hardware and software' carries with it shades of pure rationality in mind. This conceptualization depicts technology as nothing more than a physical construction of a proven scientific idea or inference. The emphasis here is on technology as a manifestation of a purely logical effort which produces a neutral 'object' or 'pattern of operation' that helps mankind *per se* rationalize a given work task or improve the quality of a certain aspect of social or work life. This is often achieved by substituting human effort with machine effort, a phenomenon which began historically with physical effort (mechanization) and continues today with human coordination and intellectual efforts being replaced with inputs from 'smart machines' (computerization and automation). This takes place in an information age where information technology has become part of a global culture and knowledge-based artificial intelligence is likely to progress further. Thus far the notion of technology could be described as an objective description of an outer form of a disinterested, 'inanimate' object or mechanistic pattern of operation.

The above chosen *system* definition of technology, however, goes beyond this rather simplistic conceptualization as it takes on board the role of the designer or user in *configuring* a system which gives an output in accordance with the designer's prespecified performance criteria. Therefore, the proposition that technology could be explained merely in terms of presumed neutral rationality grounded in a specific scientific logic appears suddenly to be in question as a credible explanation because of the interference of the subjectivities of those associated with its instrumentalization and operation in the real world.

Just how politically sensitive technology is can be illustrated by explaining how significant it is to organizing and organizations. This is a turning point in conceptualizing technology as a serious organizational development involving actual and perceived impacts of a system's output upon organizing and within organizations. In this sense, technology is capable of being socially divisive. That is, a system's output can and does affect the status quo of any organization which accommodates it or comes to live under its influence. This will have implications for any long-established equilibrium in relationships, balances of power, and any sense of continuity within the organization. It may also have implications for the wider industrial sector or economy in which the organization operates and with which it interacts. Additionally, it may indeed have implications for the environment (e.g., pollution and health hazards in the case of a nuclear technology). Since differences of interpretation among individuals and groups will inevitably occur in the world of meaning, a system's output will signify and symbolize different things to different people; it will be highly valued by some but not by others; seem useful to some as something that improves the quality of working life, for example, and harmful to others who see in it a threat or deterio-

ration. In short, it is liable to being controversial and bringing about a sense of dispute as to why it has been acquired in the first place and, now that is has, how it should be managed and controlled to minimize its potential damage.

Furthermore, what makes technology a politically sensitive issue in organizational life is the fact that invariably technology is decided upon, adopted, designed, implemented, and changed by a minority of people in an organization (e.g., managers) on behalf of the majority of the organization's members (e.g., employees) who have to live with the consequences of operating it as an integral part of their jobs. Nonusers are also likely be affected, in that their jobs will now require different coordination skills to keep them in harmony with a changed rhythm and workflow throughout the workplace, for example, where new production methods are introduced by management. A common concern for users and nonusers alike may be that the introduction of a new *input–conversion–output* system format into the workplace is going to disrupt stability and that they have to change their working methods in order to suit the operational requirements of a new and somehow threatening technology-dictated work system.

By the same token, technology, at a societal level, is often invented by a single individual (e.g., a scientist) or an elitist minority in society (e.g., a technology research centre or a team of top biochemists); it is commercialized, marketed, and diffused by another minority (e.g., entrepreneurs or manufacturing companies) who are driven by a different set of motives (e.g., profit-making); it is regulated and controlled centrally by yet another minority (e.g., a sovereign state's legislature or a federal government). Yet, this technology is highly likely to be consumed by the vast majority in society. Even nonconsumers are likely to be affected by it in either a positive or negative way, or both. Because of this minority–majority situation, technology acquires necessarily social, economic, and political implications as well as rational, logical, and technical significances.

It is no wonder, therefore, that the politics of organizational change are increasingly accomplished through technologies in a digital age. Many of the most radical redesign projects for organizations have been in and through the use of digital technologies: from computer integrated manufacturing (CIM) to process re-engineering in manufacturing industry; from projects for 'virtuality' to extensive teleworking, organizations are being sold on projects of technical rationality as visions for the future. Yet so many of these projects appear to be stillborn: the technologies work but the organizational designs do not. Sometimes people sabotage the technologies; othertimes the technologies sabotage them. In a way, organizations are reduced to criteria of design from which most considerations of the politics of organizational change are absent. This is often the case even where the projects are explicitly political or require explicit politics to manage their changes. It is the politics of organizational change technologies that make the critical difference: how, and whether, they work.

4. *The Relevance of Technology to Managing and Management*

It would be misleading to suggest that management should constitute an homogeneous team making rational decisions. Different levels and functions of management within an organization tend not to be a single social or political decision-making group sharing a single rationality; these will therefore tend to hold different perceptions of technological change and are likely to differ between themselves as to how it should be managed (see Pettigrew 1985). Yet it is clear from the discussion so far that an organization includes people who, because of their membership of a collective social group, interact with one another in certain ways to achieve goals which if they act alone they cannot accomplish. Since it is management's responsibility to have work tasks carried out by employees, in the context of a business organization for example, it follows that the impact of technology is of central concern to managers *per se*, particularly those who are in charge of implementing projects involving the introduction of a new technology. Planning, control, implementation, and evaluation activities relating to such projects require decisions to be made, ranging from individual motivation through job design to the management of change and coordination of a newly designed work organization. Such decisions are made by managers occupying different positions within an organization's management structure. Regardless of how successful or unsuccessful those decisions will come to be judged by others, it is *management* as an institution that is going to be judged in terms of its rationality and political sensitivity as well in terms of its dominant attitudes, values, and ideological persuasions.

Whatever the type of organization or management an external observer chooses to study, technology is relevant to organization because the latter is necessarily a source of the following:

(a) money and physical resources, including technology itself;

(b) meaning and purpose, including how different individuals and groups within an organization perceive technology in relation to these variables;

(c) order and stability, including how stabilizing or destabilizing it is in the context of a change project involving the introduction of a new technology;

(d) security and protection, including the consequences of change for security of employment and health and safety concerns;

(e) prestige and self-confidence, including possible improvement or degradation of job design and sat-

isfaction following the utilization of a certain technology; and

(f) power and control, including technology-related assignment of authority and shifts in the *de facto* balance of power due to empowerment caused by, for example, expert knowledge of a key technology.

Therefore, management's role in, and responsibilities for, technological change should be emphasized as a legitimate concern for an organization's stakeholders who are not members of the *governing* institution of management, because of, among other things, the strategic, operational, and control choices available to managers in planning, design and implementation (see Buchanan 1983, Sorge et al. 1983, Child 1987).

5. *The Technology–Organization Link: Empirical Studies and Theorization*

The relationship between technology and organization could be perceived, explained, and theorized upon differently by different people and groups, depending on many factors as explained above. This is in harmony with a central assumption of this article, which recognizes the legitimacy of a multiplicity of views as to what constitutes this relationship.

Such multiplicity is reflected in the management literature, including major empirical studies theorizing upon the organization–technology relationship. Examples of these studies include Joan Woodward's (1965) investigation of 100 firms in south east England, which used different unit and small batch, large batch and mass, and process production technologies. The main conclusion of this study was that *technical complexity* caused increases in the length of chain of command, in executives' span of control, in the proportion of indirect to direct labor, and in the proportion of managers to the total employed workforce. The technology used to manufacture a product placed certain requirements on those who operated it, which, in turn, were reflected in the organizational structure. Woodward's variable was further studied by Charles Perrow (1976) whose name is also associated with 'technological determinism' since he argued that specific aspects of job discretion, coordination, and group interdependence were determined by the type of technology used, much more so than by other variables such as leadership style or small group dynamics.

The Aston research program in Birmingham, England (e.g., Pugh and Hickson 1976, Pugh and Hinings 1976) contributed to organization theory by blending some of the research methods and assumptions of psychology with conceptions of organizations and their workings from sociology and economics. The empirical work undertaken by the Aston group aimed to link organizational structure and functioning, group composition, and interaction, as well as individual personality and behavior. It did so by researching organizational structures in terms of their degrees of 'specialization,' 'standardization,' 'formalization,' 'centralization,' and 'configuration.' According to the Aston studies, an organization's size and its dependence upon other organizations were found to be much more significant variables than technology and ownership.

Other relevant empirical contributions to theory include that of Eric Trist and his colleagues, of London's Tavistock Institute, who studied the effects of mechanization in coal mining (e.g., Trist and Bamforth 1951, Emery and Trist 1960). Their 'sociotechnical perspective' contributed considerably to the Quality of Working Life movement, particularly in the USA and Canada, where Fredrick Taylor's (1947) 'scientific management' style dominated the industrial scene as many large organizations followed Henry Ford's mass production model, which caused degradation in working conditions in the context of a capitalist economy.

Classical theorizing about organizational technology has produced a set of typologies for categorizing technology, which are often used to develop and empirically test contingency theories, including Woodward's (1965) above mentioned work, giving rise to the notion of 'technological imperative.' Other important issues concerning social structure are associated with the position of technology in modernist organization theory, including task interdependence, uncertainty, complexity, and change in the environment of an organization. Linear models of technological innovation are increasingly being challenged, however, by posmodernist theorists and social constructionists (e.g., Clegg 1990, Gergen 1992, Pettigrew 1985) who emphasize the need for researching the contexts in which technologies are introduced.

6. *Conclusion*

In summary, this article has addressed questions pertaining to the management of technology adoption and organizational adaptation in organizations. It defined the terms organization and technology and discussed the relationship between them. One single point remains at issue: the desire for an 'optimum fit' between the two is a noble concern, but in practical situations involving many users and a few decision makers and, equally, in theoretical situations involving numerous typologies and competing interpretations, on whose terms?

See also: Information Society, Geography of; Organizations, Sociology of; Science and Technology, Social Study of: Computers and Information Technology; Technological Innovation; Technology and Social Control

Bibliography

Buchanan D 1983 Technological imperatives and strategic choice. In: Winch G (ed.) *Information Technology in Manufacturing Processes*. Rossendale, London

Child J 1987 Organizational design for advanced manufacturing technology. In: Wall T, Clegg C, Kemp N (eds.) *The Human Side of Advanced Manufacturing Technology*. Wiley, Chichester, UK

Clark J, McLoughlin I, Rose H, King R 1988 *The Process of Technological Change: New Technology and Social Change in the Workplace*. Cambridge University Press, Cambridge, UK

Clegg S 1990 *Modern Organizations: Organization Studies in the Postmodern World*. Sage, London

Emery F, Trist E 1960 Sociotechnical systems. In: Churchman C, Verhulst J (eds.) *Management Science, Models and Techniques*. Pergamon, Oxford, UK. Vol. 2

Gergen K 1992 Organization theory in the postmodern era. In: Reed M, Hughes M (eds.) *Rethinking Organization: New Directions in Organization Theory and Analysis*. Sage, London

Huczynski A, Buchanan D 1991 *Organizational Behaviour: An Introductory Text*. Prentice-Hall, London

Perrow C 1967 *Organizational Analysis: A Sociological View*. Tavistock, London

Pettigrew A 1985 *The Awakening Giant: Continuity and Change in ICI*. Blackwell, Oxford, UK

Pugh D, Hickson D 1976 *Organizational Structure in its Context: The Aston Programme I*. Gower, Aldershot, UK

Pugh D, Hinings C (eds.) 1976 *Organizational Structure—Extensions and Replications; The Aston Programme II*. Gower, Aldershot, UK

Scarbrough H, Corbett J 1992 *Technology and Organization: Power, Meaning and Design*. Routledge, London

Sorge A, Hartmann G, Warner M, Nicholas I 1983 *Microelectronics and Manpower in Manufacturing*. Gower, Aldershot, UK

Taylor F 1947 *Scientific Management*. Harper & Row, New York

Trist E, Bamforth K 1951 Some social and psychological consequences of the Longwall method of coal getting. *Human Relations* **4**: 3–38

Woodward J 1965 *Industrial Organization: Theory and Practice*. Oxford University Press, London, UK

T. M. K. Tantoush

Technology and Social Control

The last half of the twentieth century has seen a significant increase in the use of science and technology for purposes of social control. Examples include video and audio surveillance, heat, light, motion, sound and olfactory sensors, electronic tagging of consumer items, animals and humans, biometric access codes, drug testing, DNA analysis, and the use of computer techniques such as expert systems, matching and profiling, data mining, mapping, network analysis, and simulation. Control technologies have become available that previously existed only in the dystopic imaginations of science fiction writers. Many technologies developed for the military such as spy and global positioning satellites and night vision scopes have become commercially available. Six technology-based strategies of control and some social and ethical issues they raise are discussed.

As used here social control refers to efforts to enforce norms by preventing violations or discovering and apprehending violators, rather than to other aspects of social control such as the creation of norms, processes of adjudication and sanctioning, or the broad societal guidance and integration which was of concern to early theorists of industrialization and urbanization. The engineering of social control can be differentiated from related control forms such as the creation and manipulation of culture and socialization, the redistributive rewards of the welfare state and interpersonal influences.

The increased prominence of social control via engineering is related to concerns over issues such as crime, terrorism, drug abuse, border controls, AIDS, and economic competitiveness; and to technical developments in electronics, computerization, artificial intelligence, biochemistry, architecture, and materials science. The scale, mobility, and anonymity of mass society and ironically, increased expectations of and protections for privacy, have furthered reliance on external, impersonal, distance-mediated, secondary technical means, and database memories that locate, identify, register, record, classify, and validate or generate grounds for suspicion. The perception of catastrophic risks in an interdependent world relying on complex technologies and the entrepreneurial efforts of the security industry and governments such as the United States with its war on drugs, have helped spread the technologies internationally.

There is a magisterial, legitimacy-granting aura around both law and science (Ericson and Shearing 1986). Too often they serve as their own justification divorced from broader critical questions. Technological controls, presumably being science-based, are justified as valid, objective, neutral, universal, consensual, and fair. This view tends to overlook the fact that results are socially interpreted (and thus potentially disputable) and it overlooks the personal interests of control agents and the sectarian, socially constructed interests agents may represent. This legitimacy is strengthened in free-market societies where the tactics can often be used by citizens (e.g., video cameras to record police behavior or DNA analysis offered by a criminal defendant) and internally by police managers for guarding the guards.

Of course the inventors and builders of the first locks, safes, moats, and castles and the developers of early biometric identification systems (e.g., the Italian criminologist Cesare Lombroso, 1835–1909) were engaged in the engineering of social control. What are new are the scale and relatively greater scientific precision, continual invention and experimentation, and rapid global diffusion. Technical means of control

saturate modern society, colonizing and documenting ever more areas of life.

The roots of contemporary social control lie in the development of large organizations and standardized control technologies (Beniger 1986). They are one strand of broad processes of rationalization, professionalization, and specialization occuring with modernization (Weber 1958, Rule 1973, Foucault 1977, Cohen 1985, Laudon 1986, Gandy 1993, Zuboff 1988, Lyon 1994, Shenhav 1999).

The ratio of machines as protectors, monitors, and controllers relative to humans continues to increase. Control has become softer and less visible partly because it is built-in (e.g., software that sends an Internet message or shuts down if it is misused or that monitors work—such as the number of keystrokes entered or the driving behavior of truckers), and partly because of more sophisticated uses of deception (complex undercover operations and disguised video cameras). Much contemporary control is better symbolized by manipulation than coercion, by computer chips than prison bars, and by remote and invisible tethers and filters than by handcuffs, straitjackets, and physical walls. Being more covert, embedded, and remote, it is often involuntary, occurring without the awareness or consent of its subject. Controllers are increasingly able to know things about subjects that the latter do not know about themselves, and to make decisions affecting their life chances of which they are unaware.

Contemporary social control has become more *extensive* and casts a much wider net than at mid-twentieth century, continuing an expansionary trend that began with industrialization and the rise of the nation-state. Control involves ever more integrated information-sharing networks, blurring many traditional institutional and organizational borders (e.g., within and between agencies and levels of government, banks, insurance, healthcare, education, work, telecommunications, and sales and marketing organizations). This data sharing is taken further in the United States than in Canada or Europe.

Technical controls are capital- rather than labor-intensive and hence the cost of control per unit of information has decreased. More objects and areas are subjected to inspection and control and there is a broadening from the traditional targeting of a *specific suspect* to *categorical suspicion* (e.g., the computer search, video camera, or metal detector capture information on all those within their province) and from *individuals* to *networks* and *organizations*.

Control has also become more *intensive*, probing deeply beneath protective surfaces. Many contemporary controls transcend boundaries of distance, darkness, physical barriers, and time—factors that traditionally protected liberty, as well as malfeasance. Data can be easily stored, retrieved, combined (from different places and in different forms such as visual, auditory, print, and numerical), analyzed, and communicated. Control may be remote and deterritorialized, with buffers between controllers and those controlled. Control and knowledge of others' behavior are no longer restricted to what the senses directly reveal through interaction or to a bounded physical place.

Contemporary social control themes are seen in their most extreme form in the maximum security prison (e.g., continuous surveillance, extensive use of computer dossiers, decisions about individuals based on actuarial data, the engineering of control). However, with the spread of such controls throughout the society, we may ask if we are not moving toward becoming a *maximum security society* (Marx 1988).

The *surveillance society's* more omniscient and omnipresent technical social control is seen in a broad variety of rule enforcement contexts beyond public policing or prisons. The original eighteenth-century French notion of an all-knowing, absorbent political police (Brodeur 1983) to protect the state has become generalized across institutions and applied by new users and for new goals.

A reliance on technology for social control is a hallmark of private police who, in the United States (although not in Europe), far outnumber public police. It is also seen in the workplace, the marketplace, schools, and even among friends and family members. In the case of children, for example, there are at home video and audio room monitors, geographic location devices for tracking teenagers in cars, drug tests, beepers, records of phone and computer use, and internet content filters that censor what can be accessed. Consider also 'intelligent highway systems' or the 'smart homes' that are appearing on the market in which all data flows into, and out of, the home are part of an integrated, continuously monitored system.

In an *engineered society*, the goal is to eliminate or limit violations by control of the physical and social environment. As in other areas of social intervention such as public health, there is a strong emphasis on prevention. Ideally problems are anticipated and simply designed away (Sects. 1 and 4), or where that is not possible, the goal is to create deterrence by reducing the gain or increasing the likelihood of identification and apprehension (Sects. 2, 3, 5, and 6) (Marx 1995).

1. Six Social Engineering Strategies

1.1 Target Removal

The logic of prevention is clearest and most effective here. Something that is not there can not be taken. The move toward a cashless society is one example. Merchants who only accept credit or debit cards, or whose registers never have more than a modest

amount of cash are unlikely to be conventionally robbed. Furniture built into the wall cannot be stolen. Subway and bus exteriors built with graffiti-resistant metals are hard to draw upon. Through software programming, computers and telephones can be blocked from sending or receiving messages to, or from, selected locations.

1.2 Target Devaluation

Here the goal is to reduce or eliminate the value of a potential target to anyone but authorized users. The target remains, but its uselessness makes it unattractive to predators. Examples include products which self-destruct, as with some car radios when stolen, or which leave clear proof of theft, as with exploding red dye packs that stain money taken in bank robberies. Encrypted messages can often be easily intercepted, however, without the decryption code, the data are useless. Telephones, computers, automobiles, and even guns are increasingly available which can only be used with access devices such as a unique biometric identifier (e.g., retinal, voice, or geometric hand pattern), card, or access code.

1.3 Target Insulation

With this ancient technique the object of desire remains, but it is protected. Perimeter-maintaining strategies such as fences, walls, moats, guards, and guard dogs can be separated from more specific protections surrounding an object such as safes, chastity belts, goods that are in locked cases or chained to an immovable object, and the hiding or disguising of valuables. High-security, gated communities in which access and egress is controlled carefully and the use of networked sensors, alarms, and in some cases even Internet video of public areas, are becoming more common. The architectural development of 'skywalks' linking downtown private buildings creates 'sanitary zones' more subject to control than the potentially disorderly public streets below.

1.4 Offender Incapacitation

This classic strategy seeks to render potential offenders harmless with respect to the will or ability to violate the norm in question. The means may act directly on the body by permanently altering it and making certain offenses impossible—literal or chemical castration for sex offenders, cutting off the hands of thieves. Passivity may be created by tranquilizers and other medications such as Depo-Provera or psychosurgery for the violent. A variety of nonlethal restraining or blocking devices, from pepper spray to straitjackets to a net fired over a disruptive person, are available. Related efforts deal not with the body of the offender but with the instrumentalities involved in the offense. The goal is to render useless or unavailable something that is essential for the violation. For example, antidrunk driving interlock systems which require passing a breath analyzer test attached to the automobile ignition system before a car will start, or limiting gun purchases to those who have undergone computer checks for purchase eligibility (e.g., no felony conviction), or not permitting adolescents to purchase magic markers that can be used for graffiti, or mixing a bad-smelling chemical into a product to prevent it from being inhaled for its hallucinatory effects.

1.5 Exclusion

Potential offenders have been kept away traditionally from targets or tempting environments by exile, prison, curfew, and place or activity exclusions (e.g., bars for juveniles or for the home of an ex-spouse of an abusing husband). The ultimate form is capital punishment. A related form is the visible warning offered by a stigma such as the brand or clipped ear of offenders in medieval Europe, which encouraged others to stay away.

Electronic monitoring or location devices based on Global Positioning Satellites are contemporary examples. In one form alarms go off and messages are sent to authorities if an adjudicated person wearing a transmitter gets too close to a prohibited person or area or leaves an areas they are restricted to. With the human genome project completed, eugenics will become a contentious issue. For example, the belief (which ignores interactions with the environment and the socially crafted character of most rules) that DNA is linked to violence and other antisocial behavior could generate another ultimate form of exclusion—requiring a license indicating an 'acceptable' genetic pattern before a child could be born.

1.6 Offense/Offender/Target Identification

Where it is not actually possible to prevent the violation physically, or where that is too expensive, it may be possible to at least know that it took place and perhaps who is responsible and where they are. The goal is to document its occurrence, and identify the violator. A major goal of nineteenth-century forensic science was to develop reliable biometric measures of identity based on the analysis of fingerprints, facial measurements, and chemical properties (Thorwald 1965). One technique used by the former East Germany involved identifying individuals by their unique olifactors. Architectural design emphasizing visibility as a deterrent fits here (Newman 1972), as do video, audio, motion, and heat detection means and access codes that are presumed to document who enters an area, or is using a resource such as a computer. Hand-activated personal alarm systems or a luggage alarm that goes off if a purse or suitcase is illegitimately moved or opened, and the electronic tagging of

consumer items or expensive tools at work which give off an alarm if wrongly removed, are other examples.

New information technologies have made it possible not only to watch everyone, but for everyone to be a watcher. This greater ease of mobilizing the law by involving citizens in social control is one characteristic of the Anglo-American police tradition, although not in the rest of Europe. Citizens are encouraged to use hot lines to report infractions (e.g., erratic highway drivers, drug dealing, poaching, or 'whistle-blowing' regarding organizational malfeasance) via cell and traditional telephones and e-mail. The police use mass communications media to help identify and locate wanted persons via posting warrant information on web sites and crime re-enactments on television.

2. *Some Other Social Control Dimensions*

The six 'ideal type' concepts above are based on combining several aspects such as whether the focus is on the potential offender, a resource used in the violation, the offense, or the target of an offense, and whether the action involves removal/exclusion, devaluation, insulation, incapacitation, or identification. Such ideal types can be useful as a shorthand for classification and comparison. Yet they also distort by combining sources of variation that can be analyzed separately.

Another approach starts with single dimensions which may cut across the different forms. Some other relevant dimensions include classifying based on: visibility or invisibility; openness or secrecy regarding use of the tactic both generally and specifically; control access into or out of a system; the individual, an organization, or a network; a focus on the body, consciousness or the environment; the degree of reliability and validity and relative cost of errors and mistakes; the relative availability and ease or difficulty in neutralizing or using a tactic; and the presence or absence of democratic decision-making and review processes regarding the adoption and application of a tactic.

There has historically been much more research on those who violate rules than on those who enforce them. We know little about the prior social correlates and consequences of the above dimensions. While anyone who travels will see a certain international standardization of technical control (e.g., at airports, national borders, banks, nuclear power plants, or with telecommuncations) there is still much local variation. Distinctive histories, social organization, and cultures mean that sweeping generalizations across all technologies and countries are unwarranted (Marx 1995). Singapore, for example, appears to have taken electronic surveillance the furthest, while Japan and China are far behind (if for different reasons). In the United States drug testing, undercover operations, citizen informers, and consumer and work monitoring are widespread, while in Europe these are seen much less often. Great Britain is the world leader in video surveillance and Germany has been among the most innovative in computer matching and profiling. The Scandinavian countries make the most extensive use of social census data as part of their welfare states. France, Italy, and Spain have been generally slower than Anglo-American countries to adopt the new technologies. While the polygraph is rarely used in Europe because of doubts about its validity, handwriting analysis has credibility in France that is lacking in most other countries.

3. *Some Social and Ethical Implications*

However ideal a technical control system may appear in the abstract from the viewpoint of those advocating it, the world of application is often much messier and more complicated. The search for the perfect technical fix is like the donkey forever chasing the carrot suspended in front of it. To the person with a hammer everything may look like a nail. The technology's narrowing of focus may come at a cost of failing to see larger systemic contexts, alternatives, and longer-range consequences.

The complexity and fluidity of human situations makes this a rich area for the study of trade-offs, irony, and paradox. There are some parallels to iatrogenetic medical practices in which one problem is cured, but at a cost of creating another. Technical efforts to insure conformity may be hindered by conflicting goals, unintended consequences, displacement, lessened equity, complacency, neutralization, invalidity, escalation, system overload, a negative image of personal dignity, and the danger of the means determining or becoming ends.

Bars over windows to keep out thieves may prevent occupants from escaping through the window in the event of a fire. In commercial settings where access to merchandise is important, attaching expensive clothes (e.g., leather jackets) to a rack with a locked cable reduces the likelihood that an item will be stolen, but also complicates trying clothes on and impulse buying.

The discovery that a target has been rendered useless to an offender may increase violence, whether as a resource to gain the needed access, or out of frustration. For example, the appearance of 'car-jacking' is related to more sophisticated antitheft devices on cars. The use of access codes to activate autos and appliances may mean that the crime of burglary is converted to robbery or kidnapping, as thieves confront property owners and demand not only the property, but the code to make it work. A frustrated thief may respond to a self-destruct automobile radio by firebombing the car.

While the rule breaking in question may be blocked, the problem may be displaced rather than disappear or new problems may be created. In efforts to keep the homeless away from heating vents or entrances to

subway systems, a variety of grates and barriers have been developed. These do nothing to solve the issue of homelessness, but move it elsewhere. Some persons manage to get off of heroin but only by becoming addicted to methadone.

Equity issues are raised when displacement involves a shift to new victims. If relatively effective technical solutions are commercialized (as with embedding hidden transmitters in cars which permits locating them by remotely activating the transmitter) or gated communities to keep out would-be thieves, predators may focus greater attention on those unable to afford enhanced levels of security.

A related equity issue involves who has access to the technology and its results. Video cameras and desktop computers are potentially egalitarian in their low cost, ease of use, and widespread availability. But other technologies such as satellites and highly sophisticated computer programs are disproportionately available to the most powerful. There is the possibility of societies becoming even more stratified, based on unequal access to information in which individuals live in glass houses, while the external wall of large organizations, whether government or private, are one-way mirrors.

Whether out of self-interested rule breaking or human contrariness, individuals can be very creative in neutralizing systems of control. Faith in the technology and its routinization may create complacency in control agents and a predictability that undermines effectiveness. That locks open with keys and borders require access points means they are eternally vulnerable.

The initial antidrunk driving car interlock systems could be beaten by saving air in a balloon or by having someone else blow into it to start the car. A variety of means are available for beating drug tests—from contaminating the urine with bleach on one's hand to using a catheter to insert drug-free urine into the body. Dogs in heat have been used as antidotes to male guard dogs and debugging devices help discover hidden surveillance.

In a free-market economy new controls create incentives to develop means of neutralization, whether legally or available through the black market. Police use of radar detectors against speeders was soon followed by antiradar detection devices. Police in turn developed a tool for detecting when a driver was using the latter. The guilty may now face charges for the secondary or derivative offense of possession, even if they were not speeding. Not long after antitheft ignition protection systems appeared on automobiles, a device that permitted bypassing the lock appeared.

When systems cannot be technically defeated, as with very sophisticated encryption, then their human context may be compromised, whether through coercion or deception. For example, a thief who could not break a manufacturer's sophisticated encryption code, nevertheless managed to embezzle millions of dollars through generating fake invoices. He did this by having an affair with the individual who had the decryption codes.

There are also important questions around validity and reliability. Even if a measure is empirically valid, that does not guarantee a socially meaningful result. Thus, a DNA match between material from a crime scene and a suspect cannot reveal if a death resulted from a homicide or self-defense. The sample might also have been planted, or a secure chain of evidence custody not maintained. A computer match between persons on welfare and those with bank accounts may reveal a person over the savings limits, but that is not proof of cheating since funds may be held in trust for a funeral—something legally permitted, but not built into the computer program. Audio and video recordings may reflect what was done and said, but will not necessarily reveal why, or what a suspect intended. Seeing should not automatically mean believing. Thus, a suspect in an undercover scheme may have been threatened or entrapped off camera. A threat or seeming admission of a crime may have been said in jest or as boasting. Nor is a drug test, even if 'valid' indicating the presence of drugs within a person's system, a necessary indication of a violation. Depending on the assessment used, if the standard is set low enough it is possible to have a positive reading as a result of just being in a room where marijuana is being smoked (false positive). Conversely justice issues involving false negatives may also be raised if the threshold is set too high. Drug tests may also be distorted by what an individual has eaten (e.g., poppy seeds) or by some over-the-counter medications. Tests may be of doubtful validity to begin with. Concern over the validity of the polygraph led the United States Congress to greatly restrict its use (although that led to an increase in paper-and-pencil honesty tests whose validity has also been questioned).

New control techniques may be turned against control agents. While authorities may have an initial advantage, this is often short-lived. Thus, more powerful armor, bulletproof vests, and sophisticated communication systems are no longer the sole property of police. There may be something of an escalating domestic arms race in which the interaction becomes more sophisticated, but the fundamental dynamic does not change.

Documentation of *all* infractions may overload the control system. This may lower morale among enforcers who feel overwhelmed, or offer corrupt officials a resource (*nonenforcement*) to market. Since resources for acting on all the information may not be available, authorities may be accused of discriminatory enforcement.

Even if adequate resources for full enforcement action were available, organizational effectiveness could be harmed. Automatic technical solutions developed without adequate appreciation of complexity and contingency run the risk of eliminating the

discretion, negotiation, compromise, and informal understandings that are often central to morale and the effective working of organizations (Dalton 1959, Goffman 1961).

If technical solutions could somehow be effective at eliminating all rule breaking (holding apart the conflict between, and ambiguity in, and lack of consensus on, many rules), there could be some unexpected costs. Systems might become too rigid and unable to change. Much innovation is initially seen as deviance. Experimentation and risk taking can be aided by anonymity and secrecy. A socially transparent, engineered society would be more orderly, but likely less creative, dynamic, and free.

If order depended primarily on technical means of blocking infractions, rather than on legitimacy, how would people behave when the means failed, as at some points they invariably would? A social order based primarily on technical fixes is likely to be as fragile over time as one based primarily on overt repression.

Even if systems could somehow be made fool- and fail-proof with ever more, and more advanced, technology, there is a danger of viewing humans as robots, rather than as creative beings with a conscience capable of making choices about how they behave. The former image is inconsistent with belief in the dignity of the autonomous individual in a democratic society. Whatever a technology is *capable* of, the view of humans as volitional (and hence responsible for their behavior) and beliefs about the inviolability (absent clear justification) of the borders that protect private personal zones around one's body, mind, relationships, communications, physical space, and past history are central to ideas of respect for personhood.

The search for stand-alone mechanical solutions also avoids the need to ask why some individuals break the rules and points away from examining the social conditions which may contribute to violations and the possibility of changing those conditions. Technical solutions seek to bypass the need to create consensus and a community in which individuals act responsibly as a result of voluntary commitment to the rules, not because they have no choice, or only out of fear of reprisals.

A well-known, if often naïve expression given social inequality, holds that *where there is a will there is a way*. This speaks to the role of human effort in obtaining goals. With the control possibilities made available by science and technology this may be reversed to *where there is a way there is a will*. As the myth of Frankenstein implies, we must be ever vigilant to be sure that we control the technology rather than the reverse. As Ellul (1964) argues, there is a danger of self-amplifying technical means silently coming to determine the ends or even becoming ends in themselves, divorced from a vision of, and the continual search for, the good society.

See also: Control: Social; Crime and Class; Crime and Ethnicity (Including Race); Crime and Gender; Crime, Geography of; Crime: Sociological Aspects; Crime, Sociology of; Crime: White-collar; Criminal Justice, Sociology of; Criminal Law and Crime Policy; Law and Society: Sociolegal Studies; Prison Life, Sociology of; Prisons and Imprisonment; Punishment, Comparative Politics of; Punishment: Social and Legal Aspects; Technological Change in a Multicultural Context, Applied Psychology of; Technological Determinism; Technology, Social Construction of

Bibliography

Altheide D 1975 The irony of security. *Journal of Urban Life* **4**: 175–95

Andreas P 2000 *Border Games Policing the US–Mexico Divide*. Cornell University Press, Ithaca, NY

Beniger J 1986 *The Control Revolution: The Technological and Economic Origins of the Information Society*. Harvard University Press, Cambridge, MA

Bennett C, Grant R (eds.) 1999 *Visions of Privacy: Policy Choices for the Digital Age*. University of Toronto Press, Toronto, ON

Brodeur J 1983 High policing and low policing remarks about the policing of political activism. *Social Problems* **30**: 507–20

Byrne J, Lurigio A, Petersilia J (eds.) 1992 *Smart Sentencing: The Rise of Intermediate Sanctions*. Sage, Beverly Hills, CA

Clarke R (ed.) 1997 *Situational Crime Prevention: Successful Case Studies*. Harrow and Heston, New York

Clarke R A 1988 Information technology and dataveillance. *Communications of the ACM* **31**: 498–512

Cohen S 1985 *Visions of Social Control*. Polity Press, Cambridge, UK

Dalton M 1959 *Men who Manage*. Wiley, New York

Ellul J 1964 *The Technological Society*. Knopf, New York

Ericson R, Haggerty K 1997 *Policing the Risk Society*. University of Toronto Press, Toronto

Ericson R, Shearing C 1986 The scientification of police work. In: Bohme G, Stehr N (eds.) *The Knowledge Society: The Growing Impact of Scientific Knowledge on Social Relations*. Reidel, Dordrecht

Foucault M 1977 *Discipline and Punish: The Birth of the Prison*. Vintage, New York

Gandy O 1993 *The Panoptic Sort: Towards a Political Economy of Information*. Westview Press, Boulder, CO

Gibbs J 1989 *Control Sociology's Central Notion*. University of Illinois Press, Urbana, IL

Goffman E 1961 *Asylums*. Anchor Books, Garden City, NJ

Graham S, Marvin S 1996 *Telecommunications and the City: Electronic Spaces, Urban Places*. Routledge, London

Janowitz M 1975 Sociological theory and social control. *American Journal of Sociology* **81**: 82–108

Laudon K 1986 *The Dossier Society Value Choices in the Design of National Information Systems*. Columbia University Press, New York

Lowman J, Menzies R, Payys R (eds.) 1987 *Transcarceration: Essays in the Sociology of Social Control*. Gower, Aldershot, UK

Lykken D 1981 *A Tremor in the Blood: Uses and Abuses of the Lie Detector*. McGraw Hill, New York

Lyon D 1994 *The Electronic Eye*. Polity Press, Cambridge, MA

Lyon D, Zureik E (eds.) 1996 *Computers, Surveillance and Privacy*. University of Minnesota Press, Minneapolis, MN
Manning P 1992 Information technology and the police. In: Tonry M, Morris N (eds.) *Modern Policing*. University of Chicago Press, Chicago
Marx G T 1988 *Undercover: Police Surveillance in America*. University of California Press, Berkeley, CA
Marx G T 1995 The engineering of social control: The search for the silver bullet. In: Hagan J, Peterson R (eds.) *Crime and Inequality*. Stanford University Press, Stanford, CA
Marx G T 1995 Undercover in comparative perspective: Some implications for knowledge and social research. In: Fijnaut C, Marx G (eds.) *Undercover: Police Surveillance in Comparative Perspective*. Kluwer Law International, The Hague, The Netherlands
Newman O 1972 *Defensible Space*. MacMillan, New York
Norris O, Moran J, Armstrong G 1998 *Surveillance, Closed Circuit Television and Social Control*. Ashgate, Aldershot, UK
Office of Technology Assessment 1985 *Federal Government Information Technology: Electronic Surveillance and Civil Liberties*. USGPO, Washington, DC
Perrow C 1984 *Normal Accidents*. Basic Books, New York
Rule J 1973 *Private Lives, Public Surveillance*. Allen-Lane, London
Shenhav Y 1999 *Manufacturing Rationality. The Engineering Foundations of the Modern Managerial Revolution*. Oxford University Press, New York
Sherman L 1992 Attacking crime: Policing and crime control. In: Tonry M, Morris N (eds.) *Modern Policing*. University of Chicago Press, Chicago
Snyder E, Blakely M 1997 *Fortress America: Gated Communities in the United States*. Brookings Institution, Washington, DC
Tenner E 1997 *Why Things Bite Back: Technology and the Revenge of Unintended Consequences*. Knopf, New York
Thorwald 1965 *The Century of the Detective*. Harcourt Brace & World, New York
Web: mit.edu/gtmarx/www/garyhome.html
Weber M 1958 *From Max Weber Essays in Sociology*. Oxford University Press, New York
Zuboff S 1988 *In the Age of the Smart Machine*. Basic Books, New York

G. T. Marx

Technology Assessment

The term 'technology assessment,' often abbreviated as TA, covers a variety of usages. The main meaning is the early identification and assessment of eventual impacts of technological change and applications, as a service to policy making and decision making more generally. The term TA is also used when firms, consultancies and technological institutes, agencies in health care and other social sectors, want to assess the promise, and the profit, of new technological options, and/or carry out a broad version of cost–benefit analysis, sometimes also including risks. In both cases, there is an element of anticipation of future developments (of technology, and its relation to markets and society) and an element of feedback of such anticipations into relevant decision arenas. This combination is the defining characteristic of technology assessment exercises.

1. Context

Practical and political concerns about uncertainties involved in developing and applying new technologies were the breeding ground for TA in the 1960s. One response was to try to manage the uncertainties through a combination of technological forecasting and the development of techniques to address and evaluate impacts. US organizations such as the Rand Corporation pioneered such TA techniques. This managerial response could be broadened to include wider societal impacts; the impacts of the US space program, for example, were explored by making comparisons with the impact of the railroad on American society.

The other response was policy-oriented and political, taking the emerging concern about the negative impacts of technology, for example, on the environment or in relation to defense and security issues, as a starting point to try to evaluate such impacts at an early stage. Ways to do so could include the techniques developed in the managerial approach, but would also build on policy analysis and on action-oriented research.

Special agencies could be made responsible for such a task, as when in 1970 in the USA a federal law established a Congressional Office of Technology Assessment (OTA), as 'an aid in the identification of existing and possible impacts of technological application.' Congress was interested also because such analytic support would increase its power vis-à-vis the Executive. The example of OTA was noticed in European countries, and followed, with some delay, the specific format depending on the political tradition of each country.

By the mid 1980s, TA was a recognized activity, both professionally (cf. Porter et al. 1980) and politically. Gradually, the earlier and almost exclusive emphasis on analytic approaches and the production of reports, was complemented with an interest in linking up more closely with decision making, or at least contributing to setting an agenda. Public debates about energy and environmental issues helped to make this aspect of TA more prominent.

By the 1980s, a philosophy of TA, already recognizable in the 1960s, had become widely accepted: the anticipation of impacts and feedback in order to reduce the human and social costs of learning how to handle technology in society compared to when this happens by trial and error. The experience of a number of chemical and nuclear accidents, as well as uncertainties about the new biotechnology (and life sciences more generally) since the mid-1970s, contributed to the legitimacy of TA exercises. This philosophy is shared by what would otherwise seem to be

a wide variety of approaches, ranging from technology forecasting to the consensus conferences, popular as we enter the twenty-first century. This philosophy is also visible in debates about technology policy (implicit in the USA, explicit in most other countries), with the suggestion that technology policy should include a strong TA component. Some economists have argued that TA is the only justifiable component of government technology policy in a market economy (Freeman and Soete 1997).

2. *Main Strands in the Variety of TA Approaches and Activities*

By the late 1990s, four established, plus one emerging, strands of TA could be distinguished, each with its own style and addressing different audiences:

(a) TA in firms and in technological institutes, oriented toward mapping future technological developments and their value to the firm or institute, and used as an input in strategy development (Hastbacka and Greenwald 1994). 'Picking the winners' (or 'avoiding the losers') used to be the overriding orientation. This strand of TA has developed relatively independently of 'public domain' TA, but links are emerging because of the need of firms to take possible societal impacts and public acceptance into account. The biotechnology sector became a clear example by the late 1990s. The 'insiders' are interacting with the 'outsiders,' and the effect is that prudent firms are receptive to public-interest assessments.

(b) TA in specific sectors, such as environment or health care and regulation, as an instrument to determine acceptability of a project, or a new product or technological option, in relation to the public interest in, and policies about, that sector. Environmental Impact Statements are now required in many countries before projects can be implemented (Vanclay and Bronstein 1995). Medical and health care TA is a recognized specialty with its own international society and meetings, focusing on evidence-based evaluations of concrete medical and health-care activities and options (cost–benefit analysis, risk analysis), not on wider societal impacts.

(c) TA for policy development and political decision making about projects or broad programmes with a strong technological component (for example, the electronic superhighway or modern agriculture) or important technologies (for example, genetic modification). One can call this 'public service' TA, and consider the now-defunct US OTA as the embodiment of this type of TA. OTA, during its lifetime, developed a robust approach to TA studies (see Wood 1997, Bimber 1996). Other TA bodies serving national parliaments and/or national governments tend to include participatory TA methods in addition to expert-based approaches.

(d) TA exercises can be oriented to the public arena more generally, and focus on articulation and building an agenda for handling new technology in society. This most recent strand takes up the increasing calls for participation (at least by so-called new stakeholders such as environmental groups). While it is particularly visible and more or less institutionalized in some European countries (Denmark, the Netherlands), participatory methods such as consensus conferences have been taken up all over the world (Guston and Bimber 1997). Agenda-building TA has a longer history, however, given that controversies over new projects or new technologies (and the studies and documents produced in the course of the controversy) induce learning about potential impacts and articulation of the value of the technology. Agenda-building TA merges into informed consultation processes to reach agreement on the value of new technology. Thus, there is overlap between TA and more general political and policy approaches for articulation and learning (e.g., hermeneutic policy making).

The contrast between private domain TA and public domain TA seems strong, because of the difference in goals and in the actors involved. The scope of private domain TA is less broad than public domain TA, but the assessments try to be more precise, and their outcomes are fed back into strategy development and decision making. Building on the broadening of private domain TA (strand 1) and attempts of agenda-building TA to include various private actors (strand 4), a fifth strand in TA has emerged, even if it is still somewhat programmatic.

Constructive TA emphasizes that impacts of new technology are a joint product of the technology, the actors involved, and wider interactions. The experience with introducing new information and communication technologies within and between organizations was important in this respect (Orlikowski 1992), as well as the attempts of private actors to broaden the scope of their assessment processes, already at an early stage of new product or process development (cf. strand 1). Societal experiments with the introduction of new technologies, such as electric cars, mix private and public actors, and are occasions for societal learning about new technologies and for feedback into further development and uptake. While still programmatic in parts, Constructive TA has also developed generic strategies such as Strategic Niche Management (Schot and Rip 1997).

3. *Methodological Issues*

The combination of anticipation and feedback, characteristic of all varieties of TA, entails a methodology which combines, so to speak, writing a history of the future, supported by judgments of experts and by social-science insights and data, and informing action or preparation for action. Relevant social-science insights derive from sociology and economics of technology (Schot and Rip 1997) and from organiz-

ational sociology (Orlikowski 1992), and emphasize the co-evolution of technology and society. Estimating impacts is speculative, of course, but the speculation is controlled, and there are bodies of relevant experience (Vanclay and Bronstein 1995). Uptake into action remains precarious, exactly because of the fallibility of such estimates: Who could have predicted the further development and impact of the Internet in the early 1990s? On the other hand, efforts at foresight can be productive even when they turn out not to be correct, when they stimulate joint learning.

Methods of TA can emphasize data collection and analysis about present impacts and trends, and some counterfactual speculation, or consultations of experts and refinement of the results, and/or sociotechnical scenarios. The toolkit of TA contains a wide variety of tools, from forecasting broadened with social and sociotechnical mapping, multi-criteria analysis, and integrated assessment, up to mediation and communication techniques. The selection of appropriate tools is related to the sponsor of the exercise (private/promoter, sectorial/regulator, public policy, the public domain more generally), and is often eclectic. While there are recognized TA professionals, TA exercises are carried out by a wide range of consultants, academics, dedicated agencies and their committees, and volunteers.

There is a fundamental dilemma, which has been called the anticipation and control dilemma (Collingridge 1980). At an early stage of technology development, the nature of the technology and the articulation of interests are still malleable—but it is unclear what the effects and impacts will be. By the time these become clear, the technology is entrenched and vested interests make it difficult to change the technology. As with the pesticide DDT, the only possible response is to forbid further deployment of the technology. More is possible, however, than in this stark version of the control dilemma.

Recent economics and sociology of technology have shown how path dependency increases over time in technological development, and how co-production of impacts occurs along those paths. The QWERTY keyboard of typewriters, and now also of computers, is a well-known example of path dependency. From the point of view of TA philosophy, increasing path dependencies are an opportunity: they allow better anticipation of future developments. Furthermore, understanding the dynamics which generate path dependency allows one to modulate them to some extent.

A generic strategy to avoid decision regret is to remain flexible as long as possible, at the risk that none of the various options will be pursued sufficiently far so as to see what they are worth. A concrete example is the 1991 French Law on nuclear waste handling which requires three options to be maintained for a period of perhaps 15 years, and with annual public evaluation of progress and prospects. The Parliamentary *Office d'Evaluation des Choix Scientifiques et Techniques* is involved, and over time, modifications can be made to the trajectories.

When there is an addressee or identifiable audience, feedback of TA exercises on policy and on implementation of technology will occur. The direct impact of public service TA on science and technology policy and decision making has not been large, except when there are controversial issues, and the TA exercise provides ammunition for the contending parties.

Government technology policy most often focuses on promotion of (selected) technologies, as when stimulating the electronic superhighway. The main question for such technology policy is to 'pick the winners,' now for society as a whole. On the other hand, other agencies of the same government may be occupied with reducing the human and social costs of the introduction of new technology, for example, through safety and environmental regulation. This dichotomy between promotion and control of new technology is part of the *de facto* constitution of modern societies, and is reflected not only in the division of labor between government agencies, but also in cultural and political views, as in the assumption that there will be proponents and opponents a new technology (Rip et al. 1995). TA becomes a victim of these views when technology promoters see, and condemn, technology assessment as technology harassment or arrestment. While contestation will continue, there is now more interest in constructive approaches (cf. strand 5).

4. *Situating TA, and Future Directions*

One key point, related to divisions of labor in modern society, is the asymmetry between 'impactors' (those at the source of impacts) and 'impactees.' The asymmetry can be due to a difference in power, but it also always involves a difference in timing. Initiators of technological development know more, and have invested more, at an early stage, and impactees and TA agents have to wait and, in a sense, follow their lead.

The asymmetry has another component: technology developers are insiders and do not necessarily know very much about the outside. Adoption and diffusion, however, are up to 'outsiders' who have other knowledge, interests, and expectations. The story of nuclear energy since the 1960s is in part one of a struggle between insiders and outsiders. On a much smaller scale, the same storyline is visible in the development of cochlear implants for deaf people, where it turned out—unexpectedly for the insiders—that the deaf community was very negative about taking deaf people out of their own culture by providing them with implants (Garud and Ahlstrom 1997).

A second key point is that there are cumulative effects of living in a technological society. Risk

repertoires emerge in which the impacts of a new technological development are treated similarly to those of earlier ones. In the 1970s and early 1980s, the risks of recombinant DNA research were seen as those of a runaway organism and addressed with probabilistic risk analysis, exactly as had become accepted for runaway accidents in nuclear power plants. Other more appropriate framings, such as loss of biodiversity, were slower to emerge .

A third key point is that anticipation and feedback can occur without dedicated TA efforts. Early warnings and controversies over new technologies are examples, and may be starting points for TA exercises. We live in a risk society and modernity has become reflexive (Adam et al. 2000). The rise of TA is then an indicator of reflexive modernity, and a conduit to enhance reflexivity—while also channeling it in certain ways. There were criticisms of the instrumentalization of reflexivity and possible technocratic tendencies of TA in the 1970s and 1980s, and of the symbolic nature of participatory TA exercises in the 1990s. Both the critiques and the concrete TA exercises actually build on the TA philosophy, but emphasize different parts of it.

Uncertainties about new technology and society are associated with risks of the unknown and the need to do something, anything, about them, and with the variety of views and values and interests involved—which make decision making and *de facto* acceptance of new technologies difficult, and difficult to anticipate.

A striking new development is the increasing acceptance of the so-called precautionary principle (since the Rio Declaration of 1992): action is justified even if there is not yet any conclusive evidence about risks, or negative impacts more generally. Precaution is predicated on speculative scenarios about what might happen. Quality control of such scenarios is what TA can offer.

Clearly, TA is not just a new professional and managerially oriented activity since the late 1960s, and an approach to interactive agenda-building since the late 1980s. It is part and parcel of larger developments. The co-evolution of technology with organizations and with society is being taken more seriously than in the earlier 'cannon-ball' approaches to impact assessment. This is particularly visible in assessments of information and communication technologies, and their embedding in society, but also in assessments and debates about energy, transport, and infrastructure, even if vested interests and established sociotechnical regimes can be obstacles to a balanced debate. The rapid developments in the life sciences and their applications, with genetics as the most striking example, are promising and a cause for concern at the same time. There, the 'cannon-ball' storyline is still prominent, and concerns about traditional ethical values reinforce a proponent–opponent dichotomy. It is a major challenge for TA to overcome such limited definitions of the *problematique*, and to do so in productive interaction with key actors in the relevant areas.

See also: Biomedical Sciences and Technology: History and Sociology; History of Technology; Innovation and Technological Change, Economics of; Reflexivity, in Science and Technology Studies; Science and Technology, Anthropology of; Science and Technology Studies: Experts and Expertise; Technological Change in a Multicultural Context, Applied Psychology of; Technological Determinism

Bibliography

Adam B, Beck U, Van Loon J (eds.) 2000 *The Risk Society and Beyond. Critical Issues for Social Theory*. Sage, London

Bimber B 1996 *The Politics of Expertise in Congress*. SUNY Press, Albany, NY

Collingridge D 1980 *The Social Control of Technology*. Pinter, London

Freeman C, Soete L 1997 *The Economics of Industrial Innovation*, 3rd edn. MIT Press, Cambridge, MA

Garud R, Ahlstrom D 1997 Technology assessment: A socio-cognitive perspective. *Journal of Engineering and Technology Management* **14**: 25–48

Guston D, Bimber B (eds.) 1997 Technological forecasting and social change **54**(2–3): 125–308

Hastbacka M A, Greenwald C G 1994 Technology assessment—Are you doing it right? *Arthur D. Little – PRISM* (Fourth Quarter): 35–45

Orlikowski W J 1992 The duality of technology: Rethinking the concept of technology in organizations. *Organization Science* **3**: 398–427

Porter A L, Rossini F A, Carpenter S R, Roper A T 1980 *A Guidebook for Technology Assessment and Impact Analysis*. North-Holland, New York

Rip A, Misa Th, Schot J W (eds.) 1995 *Managing Technology in Society. The Approach of Constructive Technology Assessment*. Pinter, London

Schot J, Rip A 1997 The past and the future of constructive technology assessment. *Technological Forecasting and Social Change* **54**: 251–68

Vanclay F, Bronstein D A (eds.) 1995 *Environmental and Social Impact Assessment*. Wiley, Chichester, UK

Wood F B 1997 Lessons in technology assessment methodology and management at OTA. *Technological Forecasting and Social Change* **54**: 145–62

A. Rip

Technology, Anthropology of

Taking as its point of departure the fundamental sociality of technological activities, the anthropology of technology develops anthropological perspectives on 'techniques' and 'technology.'

A human universal, 'technique' includes any skilled activity (Ingold 2000), including—but not limited

to—the use of a tool, an object that skilled people can use to extend their capacity (Sigaut 1993). What interests the anthropology of technology concerning techniques is the interpenetration of material and social factors in the creation, use, maintenance, and disposal of 'artifacts' (technically modified objects). In contrast, a closely related field, material culture studies, is interested primarily in the artifacts themselves—and the ways people employ them in systems of social signification—rather than the material and social factors of their production.

Of debatable universality and anthropological utility is the second of the two notions, 'technology.' The term is a seventeenth-century neologism. Contemporary definitions of 'technology' that stress 'the application of science to conquer nature' reveal the modernist ideology underlying this term. As will be seen, anthropologists of technology are sharply divided on the wisdom of using the term 'technology' to understand technical activities in nonmodern societies.

1. Objectives of the Anthropology of Technology

Anthropologists who study technology share many goals and concerns with scholars in kindred research areas, such as science and technology studies (STS), material culture studies, ethnoarchaeology, and behavioral archaeology. All these research areas share an interest in revealing the deep interpenetration of technological and social phenomena. What makes the anthropology of technology distinctive is an interest in the way people employ the material/social intersection to create 'technoscapes.' A 'technoscape' is a socially constructed world of meaning and sociality that is the consequence of technological activity (Pfaffenberger 1999). A considerable body of anthropological scholarship, some of it concerned with technology, focuses on specific modes of production.

2. Origin and Development of the Anthropology of Technology

In French anthropology, technology has been accepted as a valid topic for anthropological inquiry since the 1930s, and a considerable body of advanced scholarship has arisen that is, regrettably, almost unknown outside France. In striking contrast, British and American anthropologists have developed perspectives inimical to the anthropological study of technology, and are only now beginning to study the subject.

2.1 The Anthropology of Technology in France

French interest in the anthropology of technology is attributed to Marcel Mauss's pioneering work on the technology of the body (1973 [1935]) and, especially, to the intellectual leadership and charisma of a social anthropologist, André Leroi-Gourhan (1911–86). Leroi-Gourhan launched an interdisciplinary project in the 1950s to comprehend the human technological adventure. This work has been far more influential than most Anglophone scholars realize. In his work on the role of gesture and speech in human evolution, Leroi-Gourhan argued (1993 [1964]) that prehistoric tool use stems not from the 'noble fruit of our thought,' but rather from the development of the hand. With the mouth freed for new uses, humans developed the capacity to externalize and remember what they learned in the form of various modes of representation—including not just speech, which generates an externalized mythology, but also 'gesture,' a parallel development that Leroi-Gourhan believed to be just as expressive as speech. Speech generates mythology, while gesture, united with tools for create techniques, generates 'mythography,' externalized graphic representations with markedly different properties than the representations enabled by speech. By refusing to privilege speech, Leroi-Gourhan provided Derrida with the ammunition needed to launch the poststructuralist attack on 'phonocentrism,' the doctrine that speech is the source and inspiration for all forms of human communication.

Another of Leroi-Gourhan's influential notions is the *chaîne opératoire* (Leroi-Gourhan 1943), a series of habitually learned steps or a 'recipe' for transforming a raw material into a finished artifact. Sketched out only nascently in Leroi-Gourhan's work, this concept has been elaborated by French prehistorians and social anthropologists; they developed a series of powerful theoretical concepts that point to the interpenetration of the material and the social. For example, Pierre Lemonnier (1980) distinguishes between the invariant components of a *chaîne opératoire*, those that cannot be omitted without 'upsetting' the technical process, and the variant components; the latter represent adaptations to the underlying social system. In a wide-ranging study of *chaînes opératoires* in pottery, van der Leeuw (1993) asserts that the social inserts itself into material processes at the cognitive level—specifically, whether the vessel is fundamentally understood (for example) as a transformation of a sphere, cone, or cylinder. What has sometimes been lost in this literature, however, is a point of major significance that follows from Leroi-Gourhan's work on gesture and speech: if gesture is no less meaningful than speech, it follows that techniques themselves, and not just the material culture produced by techniques, are a major locus of meaning-formation in human society.

A recent feminist critique of the *chaîne opératoire* concept points the way toward fulfilling the promise implicit in Leroi-Gourhan's work. Dobres (1999) points out that androcentric biases—such as the belief that technology is about production, not reproduction—often lead observers to exclude activities

involving women. Studies of the technical sequence involved in Iñupiat whale-hunting, for example, depict a male-only picture, because men and men alone hunted whales on the open sea. Yet the Iñupiat captain will say, 'I am not the great hunter; my wife is.' When the *chaîne opératoire* is broadened sufficiently to include the pre-hunt and post-hunt activities in which women figure prominently, one learns that the hunt's success is thought to depend on the wife's comportment before the hunt and her skill with her knife afterwards. Through these means the captain's wife calls forth the whale and makes it possible for the whalers to kill it.

Dobres' work shows that the anthropology of technology should adopt a version of the 'principle of symmetry' in science studies. This principle holds that a *sociological* understanding of science becomes possible only when explanations are no longer phrased in terms of the truth or falsity of scientific theories; 'true theories' and 'false theories' should be treated equally. In the anthropology of technology, the boundaries of a *chaîne opératoire* should not be determined by 'technicist' principles—that is, by the identifying sequences that 'really are technology because they work' as opposed to those that are 'merely symbolic.' At stake is the very possibility of an *anthropological* understanding of technical activity—that is, an understanding that is equally concerned with what techniques mean as with what they do.

Few studies of *chaînes opératoires* employ the principle of symmetry, but they strongly suggest that technical activities are among the most potent of the means by which people make culture. What is more, they help to shed light on key anthropological debates concerning how people make culture. For example, Munn's (1986) work on Gawan canoe-making depicts the *chaîne opératoire* as a kind of moral theater in which 'good selves' and 'bad selves' are represented in such a way that people come to understand the 'type of moral conflict in which they are engaged.' Ridington (1999) shows how Athapaskans employ a flexible repertoire of techniques, not to conquer or control nature, but to reveal and establish bonds with a 'nature' that is understood to be filled with sentient beings. The picture of culture-making that emerges from these studies is consistent with the theory of practice (habitus), which was originally formulated by Mauss (1973 [1935]), advanced by Leroi-Gourhan's *chaînes opératoires* concept, and brought to fruition by Bourdieu (1977): by engaging in the various habitual and nonconceptual skilled behaviors that are sequenced in a *chaîne opératoire*, people discover meanings, but only by 'enacting them' and 'unfolding them in time' (Bourdieu 1977). Yet they do not do so in a field of unbounded subjectivity.

What Munn found to be true of Gawa, namely, that canoe-making takes place within a kind of moral theater in which it matters very much what type of people the participants become, seems to be generally true of nonmodern and modern techniques alike (Pfaffenberger 1999). Charlie Yahey puts the nonmodern Athapaskan perspective this way: 'Animals come closer … to people who sing and dance together' (Ridington 1999). In modern technology, most of the significant technical innovations of personal computing—the graphical user interface, the mouse, the laser printer, and more—were inspired by Douglas Englebart's fervent insistence that people need to deal more effectively with an increasingly complex world, and they could not do so without graphical computer systems capable of simulating complexity on-screen. One could multiply examples *ad infinitum*, but it seems clear that personhood and technology are much more closely associated than most scholars suspect.

2.2 The Anthropology of Technology in Britain and the USA

In contrast to the anthropology of technology in France, the British and American schools of anthropology rejected technology studies as irrelevant to the discipline, and relegated the subject to museums, which were subsequently deprived of the resources needed to advance the field. Twentieth-century Anglophone anthropologists did so, in part, to distance themselves from the unscientific speculations of Victorian anthropology, with its armchair scholars, amateur collectors, and 'wild-guess' theories based on artifacts shorn of their social context. The rejection was accompanied by theoretical developments that served to brand technological activity as theoretically uninteresting. In the UK, Leach portrayed technical and ritual activities as if they were at opposed poles of a continuum ranging from the instrumental (technical) to the expressive (ritual). In the USA, Kroeber asserted that what is cultural about an artifact, and therefore of interest to anthropology, was the idea behind it, not its material incarnation or the methods of its manufacture. Both theories place technical activities into a residual category of nonsocial, mundane activities that appear to anthropologists as uninteresting subjects.

The first inklings of anthropological interest in technology did not appear until anthropologists encountered the Internet and asked whether cyberspace could become a legitimate and productive object of anthropological inquiry (Escobar 1994). This question, in turn, led them to discover the work of scholars in science and technology studies (STS), including feminist scholars who were developing a broad-ranging critique of Western science. Although the anthropologies of cyberculture (e.g., Haraway 1997) and the biosciences (e.g., Rabinow 1992) continue to attract anthropological interest, few British or American anthropologists have broadened this interest to include technological activities in the social settings that anthropologists traditionally study. Fewer still are aware of the achievements of the anthropology of

technology in France, or indeed of the fact that Leroi-Gourhan provided part of the inspiration for the poststructuralist views that are professed by so many Anglophone anthropologists, particularly in the USA.

3. 'Formalist' and 'Substantivist' Views in the Anthropology of Technology

A division has emerged in the anthropology of technology that resembles the 'substantivist' vs. 'formalist' split in economic anthropology. In brief, 'substantivists' believe that the 'laws' of economics are historical constructions of market capitalism and do not apply to noncapitalist economies, while 'formalists' believe that economic concepts can illuminate aspects of human behavior in all human societies. In the anthropology of technology, the 'substantivist' position—as expressly stated in an important work by Ingold (2000)—asserts that the 'technology' concept is a construction of modernism and can only mislead when applied to nonmodern societies. The 'strong formalist' position in the anthropology of technology is characterized by certain American archaeologists who argue, in various ways, that nonmodern techniques are subject to selection processes and therefore embody a tacit rationality (see, for example, Durham 1990). However, there is a moderate 'formalist' position that takes postmodern and anti-essentialist critiques of Western ideology seriously. Although there is no single, formal statement of this moderate 'formalist' position, it can be adduced from the work of behavioral archaeologists (e.g., Schiffer and Skibo 1997) and social anthropologists who have worked with historians of technology (e.g., Pfaffenberger 1999). The moderate 'formalist' position argues that the 'technology' concept is indeed relevant to nonmodern as well as modern societies, but only after the term is redefined in a way that eliminates its normative and ideological components.

3.1 'Technology,' Science, and 'Disembeddedness'

The 'substantivist' position is in accord with the central objective of the anthropology of technology: to reveal the interpenetration of the social and material and show how people use technical activities to construct worlds of meaning. It is for precisely this reason that Ingold refuses to use the term 'technology'—and indeed declares, 'there is no such thing as technology in premodern societies' (2000, p. 314). The 'technology' concept, in this view, is historically specific and utterly misleading when applied to nonmodern societies: it arose in the seventeenth century and is associated with the emergence of a new, science-based 'mechanistic ontology' that became increasingly hegemonic over time. The rise of a new science-based technological order went hand-in-hand with multiple forms of disembeddedness: machines 'exteriorized' production by replacing human skills with automatic mechanical movements, engineers staked out a higher status than technicians by establishing rationalist and scientific canons of technical design, and production became separate from consumption. To apply the term 'technology' to premodern societies, therefore, is to apply to these societies the very idea that destroyed them; 'technology' is the result of a historical process of disembedding and hegemony formation that destroyed the former integration of the material and the cultural.

The moderate 'formalist' response faults the 'substantivist' position for taking the definition of 'technology' at face value. Underlying the 'science discovers, technology applies' notion evident in dictionary definitions of 'technology' is a normative component, which holds, essentially, that science should discover and technology should apply; the normative dimension of the term can be fully grasped only by understanding the historical and social processes by which the 'pure sciences' were elevated to an exalted status in British and American universities and, conversely, the 'applied sciences' were denigrated and marginalized (Weiner 1992). In addition, the 'science discovers, technology applies' mythos inverts history. Historians of technology have demonstrated repeatedly that modern disciplines taken as the very essence of science (such as thermodynamics) resulted from new, widely adopted technologies (such as the steam engine) that were poorly understood theoretically.

Moreover, science and technology have differing goals. Science asks, 'Is it true?' while technology asks, 'Will it work?' Technological practitioners can answer that question only by assembling a heterogeneous ensemble of human and nonhuman actors into a network that, as Law (1987) puts it, is capable of resisting dissociation. Such a network can succeed only to the extent that it integrates the material and the social. For example, Napster, Inc., attempted (unsuccessfully) to dominate the distribution of recorded music on the Internet by uniting the MP3 file format, youths' disdain for copyright law, certain elements of US copyright law that the company believed erroneously would exonerate them from liability, and an initial public stock offering that occurred during a period of inordinate enthusiasm for Internet stocks. Although it is certainly true that modern technologies are 'disembedded' through such means as the emergence of an engineering profession, scholars in science and technology studies have demonstrated that the 'disembedding' of engineering practice is more normative than descriptive. It is only very recently, in areas such as biomedicine and genetic engineering in agriculture, that anything like the union of science and technology envisioned in dictionary definitions of 'technology' has occurred; historians of technology

use the term 'technoscience' to distinguish such fields from the largely unscientific technologies of the nineteenth century and much of the twentieth.

3.2 Technological Knowledge

The 'substantivist' position in the anthropology of technology argue that, under the sway of 'technology' definitions laden with modernist ideology, scholars in many disciplines have fundamentally erred in characterizing the type of knowledge associated with 'techniques.' Viewing the world through the lens of 'technology' leads scholars to see 'technology' anywhere there is 'technique.' In this view, as the historian of technology Edward Layton (1974) argues, all technological activity, ranging from the simplest techniques of premodern societies to twenty-first-century space-age technology, begins with the human conceptual ability to conceive of a way to adapt available means to preconceived ends—in other words, to design. If technique necessarily involves conceptual knowledge, then even the most 'primitive' technology involves not only a set of procedures and skills, but also a body of knowledge that differs only in complexity, extent, and sophistication from that possessed by modern practitioners.

Countering Layton's view, Ingold (2000) argues that skilled technical practice cannot be grasped as the mechanical execution of a preconfigured design, as if techniques represented the unfolding of an interior and conceptual algorithm. As evidence for this, Ingold offers a description of various techniques that employ iterative or recursive procedures, thereby eliminating the need for a preconceived design. To equate the knowledge employed in technical practice with that of industrial design is to overstate the role of verbal knowledge in the acquisition of skills; to be sure, verbal imperatives help novices to get started—beginning bicyclists, for example, are told to refrain from leaning until they wish to turn—but this knowledge is insufficient to fully describe technical mastery, which necessarily involves direct experience and imitative learning. At some point, a skilled practitioner abandons the novice's verbal formulations. To retain or elaborate these verbal formulations, Ingold suggests, would be tantamount to learning the map instead of the country. In short, the 'knowledge-for' that is associated with technical skills is far less systematic and verbalizable than the 'knowledge of' propositions that Ingold (wrongly) assumes to characterize the knowledge systems of modern technology.

In response, a moderate 'formalist' position concurs with Ingold's assessment of the destructive impact of an unexamined definition of 'technology,' through the lens of which indigenous knowledge systems are mistakenly seen to be prototypical sciences (Sillitoe 1998). But modern (nineteenth and early-to-mid twentieth century) technologies are equally misunderstood by likening their knowledge systems to science. Modern technological systems are known to proceed, and are even capable of generating significant advances, despite the lack of a complete theoretical understanding; for example, a century of highly successful aeronautical design has proceeded without a sound theoretical understanding of turbulent flow. What is more, these systems can operate quite successfully on the basis of false understandings; for example, aeronautical engineers assumed for nearly a century—wrongly, as is now realized—that stability in flight should overrule all other design considerations. Technological belief systems—which Bijker (1995) terms 'technological frames'—serve important roles in constituting and uniting communities of practitioners, guiding the formation of the meanings attributed to technical activities and artifacts, and providing ready-to-use recipes for technical design. Technological knowledge in modern technological systems also differs from scientific knowledge in that much technological knowledge is embodied in tacit visual representations or experientially learned skills that are exceptionally difficult to verbalize. For example, NASA has conceded that it cannot rebuild the powerful Saturn V booster rockets used in the Apollo program, and not simply because there is no surviving set of blueprints; more damaging is the loss of the tacit, nonverbalized skills possessed by the thousands of men and women, now retired or deceased, who built and flew the rocket with such stunning success.

Arguably, it is only the moderate 'formalist' position, which acknowledges the similarities between nonmodern and modern technologies, that can fully reveal the interpenetration of the material and the social. Unaware that modern technologies function quite well even though they may be founded on 'unscientific' or 'false' belief systems, anthropologists have never known quite what to do about the 'false' beliefs associated with nonmodern technology. One approach, 'technicism,' simply ignores them. Malinowski (1935) explains away the 'false' beliefs Trobriand Islanders maintain about their gardens by viewing these beliefs as a sign of 'anxiety' over the outcome of production. Ingold's approach is to see such beliefs as the subjective outcome of an individual's skilled, nonverbal interaction with an environment or lifeworld. A moderate 'formalist' perspective is in agreement with Ingold's view, except that—in line with the theory of practice developed earlier—it would point out that the meanings drawn from subjective experience are not formulated in a vacuum. The beliefs associated with technological activities in nonmodern societies, such as the Iñupiat whaler's assertion that his wife is the great hunter, are recounted and shared socially by means of the stories communities tell about their experiences (Ridington 1999). Such stories are analogous to the 'technological frame' of modern technologies (Bijker 1995), from which modern and nonmodern technologists alike are able to locate their history and identity.

4. Technological Change

The anthropology of technology is also concerned with long-term technological change, working with a canvas that includes all of prehistory and history combined. It is difficult to resist the tendency, apparent in modernist views of the technological past, to see all previous technologies as forming a unilinear path of development, characterized by ever-swelling cranial capacities and increasingly sophisticated tools, that culminates in today's technological achievements. Whether approached from the 'substantivist' or moderate 'formalist' perspective, the anthropology of technology suggests that to do so would be a very serious mistake.

4.1 Tool Sophistication, Skill, and Intelligence in Prehistory

By stressing the importance of skilled technique (Ingold 2000) and the habitual sequences by which gestures and tools are integrated in the nonmodern *chaîne opératoire* (Leroi-Gourhan 1943), the anthropology of technology sounds a note of profound warning to prehistorians who would interpret every sign of increasing sophistication in the manufacture of tools as evidence of a corresponding increase in human intelligence or the complexity of language. Very smart people may well prefer simple tools, and compensate by developing spectacularly complex skills that, although inordinately difficult to acquire, enable them to work just as efficiently as people who opt for more complex and specialized tools. Simple tools do not signify simple minds.

4.2 'Evolution' of Technology

The anthropology of technology also calls into question 'evolutionary' conceptions of technological change in archaeology, in which the forces of natural selection are seen to punish 'inefficient' technologies (and the groups that cling to them). For one thing, what might seem like an 'inefficient' technology may have been quite efficient in skilled hands; for example, Bantu ironsmiths are known to have been able to produce consistently good iron blooms from their 'primitive' furnaces, but the skill they once possessed has unfortunately been lost; modern attempts to reconstruct and operate these furnaces produce uniformly disappointing results. Furthermore, microanalyses of technological change in both modern and nonmodern societies demonstrate conclusively that it is not the quest for greater 'efficiency,' but rather demands imposed by changes in underlying social and ideological systems, that leads innovators to modify techniques or invent new ones (cf. Schiffer and Skibo 1997). It is only by stepping back from such microanalyses, ignoring discrepant evidence, and abstracting technological change from its social and ideological context that one sees a 'unilinear evolution' of technology (Basalla 1988). The appearance of cumulative technological advance, in sum, is generated by concealed social strategies and belies the fact that the knowledge underlying abandoned technological systems can be unrecoverably lost.

Bibliography

Basalla G 1988 *The Evolution of Technology*. Cambridge University Press, Cambridge, UK

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK

Bijker W 1995 *Of Bicycles, Bakelites, and Bulbs*. MIT Press, Cambridge, MA

Dobres M 1999 Technology's links and chaînes. In: Dobres M, Hoffman C, *The Social Dynamics of Technology*. Smithsonian Institute, Washington, DC, 124–46

Durham W 1990 Advances in evolutionary culture theory. *Annual Review of Anthropology* **19**: 187–210

Escobar A 1994 Welcome to cyberia: Notes on the anthropology of cyberculture. *Current Anthropology* **35**: 211–23

Haraway D 1997 *Modest-Witness@Second-Millennium. Femaleman-Meets-OncoMouse*. Routledge, New York

Ingold T 2000 *The Perception of the Environment: Essays in Livelihood, Dwelling, and Skill*. Routledge, London

Law J 1987 Technology and heterogenous engineering. In: Bijker Huges W T, Pinch T (eds.) *The Social Construction of Technological Systems*. MIT Press, Cambridge, MA, pp. 111–34

Layton E 1974 American ideologies of science and engineering. *Technology and Culture* **17**: 688–700

Lemonnier P 1980 *Les Salines d l' Ouest. Logique technique, logique sociale*. Presses Universitaires de Lille, Paris

Leroi-Gourhan A 1943 *Evolution et techniques: L'homme et la matière*. Albin Michel, Paris

Leroi-Gourhan A 1993 *Gesture and Speech*. MIT Press, Cambridge, MA

Malinowski B 1935 *Coral Gardens and Their Magic*. American Books, New York

Mauss M 1973 Techniques of the body. *Economy and Society* **2**: 70–88

Munn N 1986 *The Fame of Gawa*. Cambridge University Press, Cambridge, UK

Pfaffenberger B 1999 Worlds in the making. In: Dobres M, Hoffman C (eds.) *The Social Dynamics of Technology*. Smithsonian Institute, Washington DC, pp. 147–66

Rabinow P 1992 Artificiality and enlightenment: From sociology to biosociality. In: Crary J, Kwinter S (eds.) *Incorporations*. Zone, New York

Ridington R 1999 Dogs, snares, and cartridge belts: The poetics of a northern Athapaskan native technology. In: Dobres A, Hoffman C (eds.) *The Social Dynamics of Technology: Practice, Politics, and World Views*. Smithsonian Institute, Washington DC, pp. 167–85

Schiffer M, Skibo J 1997 The explanation of artifact variability. *American Antiquity* **62**: 27–51

Sigaut F 1993 How can we analyse and describe technical actions? In: Berthelet A, Chavaillon J (eds.) *The Use of Tools by Humans and Non-Human Primates*. Clarendon Press, Oxford, UK, pp. 381–97

Sillitoe P 1998 The development of indigenous knowledge. *Current Anthropology* **39**: 223–52

van der Leeuw S 1993 Giving the potter a choice. In: Lemonnier P (ed.) *Technological Choices*. Routledge, London, pp. 238–88
Weiner M 1992 *English Culture and the Decline of the Industrial Spirit*. Penguin, Harmondsworth, UK

B. Pfaffenberger

Technology Districts

Technology districts are areas where activities to create and enhance knowledge for practical (especially commercial) uses cluster and constitute a major source or stimulus of value added to the local economy. They include what geographers refer to as 'high-tech region,' 'science/tech city,' 'technopole,' and 'technopolis' (Castells and Hall 1994, Malecki 1997, Scott 1993), which all contrast with traditional industrial districts based on manufacturing.

1. *Common Features*

Silicon Valley, Route 128, and the Research Triangle in the US, Cambridge in the UK, Sophia-Antipolis in France, Tsukuba and Kansai in Japan, Hsin-chu in Taiwan, and Bangalore in India are examples of technology districts. While their activities do not demonstrate a uniform pattern, there exist some similarities, such as interplay between public and private initiatives, active involvement of local educational and research institutions in innovation, a large proportion of highly educated and skilled personnel in the local labor force, and access to a wide range of urban infrastructural facilities.

2. *Driving Forces*

Technology districts are post-World War II (WWII) phenomena. Their genesis and development have been shaped by diverse factors (Sternberg 1996).

2.1 *State Action*

In some countries (e.g., the US and the UK) extensive defense spending since WWII has provided a powerful boost for corporate R & D in selected regions. It has also led to a significant growth of locally anchored university-industry ties through shared research facilities, joint or commissioned projects, and training. A further result is the increase of local technological diffusion and linkages through subcontracting and spin-offs. On the other hand, domestic politics has driven some local and national governments to make territorial high-tech development a major policy agenda. Efforts to carry it through often involve planning, funding, regulatory support, tax incentives, intermediary services, and improvement of physical, social, and research infrastructure in targeted locales.

2.2 *Changing Milieu of Innovation*

Technological change in the postwar era has been accompanied by concurrent weakening of regulatory barriers to entry, lowering of knowledge threshold for tapping the potentials of new technologies, and rise of venture capital. These developments have intensified competition and shortened the cycle for technology-intensive products. This makes it increasingly difficult to reply on organizationally integrated division of labor as the primary mode of innovation. Synergy from interactions across organizational boundaries, thus, assumes greater importance for technological development. Such synergy may derive from territorial clustering of innovation activities, as spatial proximity is conducive to mutual learning, exchange of information, sharing of infrastructure, pooling of skilled human resources, development of trust, containment of risk and transaction cost, coordination of R & D efforts between interdependent parties, and formation of industry and service linkages.

2.3 *Economic Globalization*

In the last quarter of the twentieth century, the gradual lowering of barriers to cross-national movement of resources, information, products, services, and organizational practices has redefined the spatial character of international division of labor. National and regional competitiveness hinges increasingly upon innovation rather than resource endowment and factor cost. The challenges and opportunities posed by such reality have prompted convergent responses from growing numbers of governmental, corporate, and university decision-makers. A focal strategy is to identify and develop viable niches of locational advantage in which to form critical masses of innovation activities that are globally competitive. This leads to the creation or reformation of regulatory and infrastructural enclaves within or outside old manufacturing belts and metropolitan areas. They are intended to channel and attract resources from domestic and international sources for territorially concentrated technological development. A related goal is to strengthen the ties between the local economy and global conditions of demand and supply.

3. *Variations*

The formation and trajectory of technology districts vary in different historical and local contexts. Some districts evolve mainly out of university–industry interactions, whereas others are created and developed under government auspices. To certain extent, the

level, scope, and scale of innovation activities in different technology districts are conditioned by the level and form of regional and national economic development (Castells and Hall 1994).

Moreover, not all efforts to build technology districts are successful, and some established districts have experienced difficulties, stagnation, and decline. Some studies (e.g., Saxenian 1994) attribute this to the rigidities in path-dependent local corporate structure and business culture that hinder the development of innovation networks. Other factors that may stifle local synergy include poor planning, weak entrepreneurialism in face of tight bureaucratic control, predominance of divergent agendas and interests (e.g., property development), and intradistrict negative externalities (e.g., infrastructural overloading and environmental deterioration).

4. *Consequences*

The full socioeconomic ramifications of technology districts remain to be explored. What are their effects on interregional disparities? How do they stimulate economic growth and institutional change beyond the region? How and to what extent do they foster elitism and increase cleavages in society? These are some of the important issues that have been raised but need to be examined more closely.

Furthermore, technology districts are evolving with the forces that they help unleash. In particular, their spatial organization faces growing impact from the explosive expansion of the Internet and rapid development of information technology, which could significantly reduce the constraints posed by distance and locality-specific institutions on social interaction. How this redefines the role of location in fostering innovation will be a major focus of attention among both practitioners and observers.

See also: Community Economic Development; Local Economic Development; Regional Geography; Technological Innovation; Technology-supported Learning Environments

Bibliography

Castells M, Hall P 1994 *Technopoles of the World: The Making of Twenty-first-century Industrial Complexes*. Routledge, London

Malecki E J 1997 *Technology and Economic Development: The Dynamics of Local, Regional and National Competitiveness*, 2nd edn. Longman, Harlow, UK

Massey D, Quintas P, Wield D 1992 *High-tech Fantasies: Science Parks in Society, Science and Space*. Routledge, London

Porter M E 1990 *The Competitive Advantage of Nations*. Free Press, New York

Saxenian A 1994 *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press, Cambridge, MA

Scott A J 1993 *Technopolis: High-technology Industry and Regional Development in Southern California*. University of California Press, Berkeley, CA

Sternberg R 1996 Regional growth theories and high-tech regions. *International Journal of Urban and Regional Research* **20**(3): 518–38

I. Eng

Technology, Social Construction of

SCOT is one approach among several constructivist ways of studying science and technology that emerged in the 1980s. Here, 'constructivist' means that the truth of scientific facts and the working of technical artifacts are studied as accomplishments—as being constructed—rather than as intrinsic properties of those facts and machines. The term 'social construction of technology' can be used to denote two different things. First it is a research approach to study technical change in society, both in historical and in contemporaneous studies. Second, it is a theory about the development of technology. Historically the development of SCOT is closely linked to the sociology of scientific knowledge (SSK) and to the science, technology, and society (STS) movement in the 1970s. In the 1980s and 1990s, it developed primarily as an academic enterprise, but then also found application in the policy domain, in innovation management, and in discussions about new forms of democracy.

1. *Constructivist Studies of Science and Technology*

The phrase 'social construction' was first used by Berger and Luckmann (1966) in their *Treatise in the sociology of knowledge*. Building on the phenomenological tradition and particularly on the work of Alfred Schutz, they argue that reality is socially constructed and that these processes of social construction should be the object of the sociology of knowledge. Berger and Luckmann focus on the social construction of ordinary knowledge of the sort that we use to make our way about society. They are concerned with the reality of social institutions and their focus is on society at large, rather than on subcultures such as science and technology. Nevertheless, scholarship developed around such themes as the social construction of mental illness, deviance, gender, law, and class. Similarly, in the 1970s the social construction of scientific facts developed (Latour and Woolgar (1986)), followed in the 1980s by the social construction of artifacts (Bijker et al. 1987) (see *Laboratory Studies: Historical Perspectives*; *Scientific Controversies*; *Scientific Knowledge, Sociology of*).

Constructivist studies of science and technology

come in a wide variety of forms, both mild and radical (Sismondo 1993). The mild versions merely stress the importance of including the social context when describing the development of science and technology. Examples of such work are Constant's (1980) account of the turbojet revolution, Douglas' (1987) history of radio broadcasting, Nye's (1998) studies of the electrification of America, and Kranakis' (1997) history of French bridge engineering. The radical versions argue that the content of science and technology is socially constructed. In other words, the truth of scientific statements and the technical working of machines are not derived from nature but are constituted in social processes. Radical constructivist studies of science and technology share the same background, have similar aims, and are even being partly carried out by the same authors. The remainder of this article will focus on technology studies (see *Scientific Knowledge, Sociology of*) and more precisely on the radical versions of SCOT.

2. *The Origin and Development of the Social Construction of Technology*

SCOT grew out of the combination of three distinct bodies of work: the STS movement, SSK, and the history of technology. The first started in the 1970s, mainly in the Netherlands, Scandinavia, the United Kingdom, and the US. Its goal was to enrich the curricula of both universities and secondary schools by studying issues such as scientists' social responsibilities, the risks of nuclear energy, the proliferation of nuclear arms, and environmental pollution. The movement was quite successful, especially in science and engineering faculties, and some of the STS courses were incorporated into degree requirements.

SSK emerged in the late 1970s in the United Kingdom on the basis of work in the sociology of knowledge, the philosophy of science, and the sociology of science (see *Scientific Knowledge, Sociology of*; *Strong Program, in Sociology of Scientific Knowledge*). The central methodological tenets of the 'strong program' (especially its symmetry principle) seemed equally applicable to technology. The symmetry principle called for a symmetrical analysis of true and false scientific statements, explaining their development with the same conceptual framework. In studying technology, this should mean that working and non-working machines should be analyzed symmetrically and in the same terms.

In the history of technology, especially in the US, an increasing number of scholars began to raise more theoretical and sociologically inspired questions (influential were Constant 1980, Hughes 1983, Cowan 1983). Path-breaking advocacy for this body of work in the history of technology was provided by the reader edited by MacKenzie and Wajcman (1985). Citing the title of this volume, the phrase 'social shaping of technology' is used to denote a broader area of research, including work in industrial sociology, technology policy, and the economics of technical change (Williams 1996).

Researchers from these three traditions convened an international workshop in 1984 in the Netherlands. The subsequent volume from that workshop, edited by an STSer, a historian of technology, and a sociologist of scientific knowledge (Bijker et al. 1987), has been heralded as the starting point of SCOT. To understand the role of this workshop and volume, it is helpful to distinguish between a broad and narrow usage of the phrase 'social construction of technology' (but to note that both notions fall within the radical meaning of social constructivism).

When broadly used, the 'social construction of technology' encompasses all the work represented in the 1987 volume, including the actor-network approach by Callon, Latour, and Law (see *Actor Network Theory*), and the technological systems approach advocated by Hughes (see *History of Technology*). Used more narrowly, it refers primarily to the program set out by Pinch and Bijker (1984). For this latter program of the social construction of technology, the acronym 'SCOT' was introduced; but this acronym is now increasingly used in the broad sense too, without reference to one specific approach in constructivist studies of technology. The remainder of this article is primarily about SCOT in the narrow sense, although SCOT also broadened its agenda (for reviews of SCOT in the broad sense, see Bijker 1995b).

SCOT developed like any normal scientific program: its agenda, central concepts, and even unit of analysis shifted in response to research findings and discussions among contributing scholars. In that sense, one can distinguish early and late (or recent) versions of SCOT.

An important, though negative, starting point for SCOT was to criticize technological determinism (see *Technological Determinism*). Technological determinism was taken to comprise two elements: (a) technology develops autonomously, and (b) technology determines societal development to an important degree. This view was seen as intellectually poor and politically debilitating. Technological determinism implies a poor research strategy, it was argued, because it entails a teleological, linear, and one-dimensional view of technological development. In addition, it was considered politically debilitating because technological determinism suggests that social and political interventions in the course of technology are impossible, thus making politicization of technology a futile endeavor. To bolster this critique, it was necessary to show that the working of technology was socially constructed, with the emphasis on 'social.' Key concepts in this program, as will be discussed in the next section, were 'relevant social group,' 'interpretive flexibility,' 'closure,' and 'stabilization.' The

unit of analysis was the single artifact. The choice of the artifact as unit of analysis was a choice of the 'hardest possible case.' To show that even the working of a bicycle or a lamp was socially constructed seemed a harder task, and thus—when successful—more convincing, than to argue that technology at a higher level of aggregation was socially shaped.

The agenda of demonstrating the social construction of artifacts by detailed analysis at the micro level resulted in a wealth of case studies. A few years later, the program was broadened in two ways (Bijker and Law 1992). First, questions were raised at a meso and macro level of aggregation as well: for example about the political construction of radioactive waste, clinical budgeting in the British National Health Service, or technically mediated social order. Second, the agenda was broadened to include again the issue of technology's impact on society, which had been bracketed for the sake of fighting technological determinism. Concepts developed for this agenda were 'technological frame,' and various conceptualizations of the obduracy of technology. The unit of analysis was broadened from the singular technical artifact to the more comprehensive and heterogeneous socio-technical ensemble. The emphasis now was on 'construction' rather than on 'social.' (A similar shift occurred in the social construction of science. Thus, for example, Latour and Woolgar (1986) dropped the word 'social' from the subtitle of their book *Laboratory Life* when it was reprinted in 1986.) The aim became to study the mutual shaping of society and technology, rather than the one being shaped by the other.

Present research in SCOT combines ongoing empirical case studies with more general questions about the modernization of society, the politicization of technological culture, and the management of innovation. It has become increasingly difficult (and unfruitful) to observe the boundaries between the various approaches within the broader social construction of technology. Research collaboration and conceptual combinations have emerged between, for example, the actor-network approach, SCOT in the narrow sense, and gender and technology studies. Connections are developing with studies in the social sciences more generally, including politics, economics, and law.

3. *The Social Construction of Technology as a Heuristic for Research*

As a heuristic for studying technology in society, SCOT can be laid out in three consecutive research steps (Bijker 1995a).

Key concepts in the first step are 'relevant social group' and 'interpretive flexibility.' An artifact is described through the eyes of relevant social groups. Social groups are relevant for describing an artifact when they explicitly attribute a meaning to that artifact. Thus, relevant social groups can be identified by looking for actors who refer to the artifact in the same way. For describing the high-wheeled ordinary bicycle in the 1870s, such groups were, for example, bicycle producers, young athletic ordinary users, women cyclists, and anti-cyclists. Because the description of an artifact through the eyes of different relevant social groups produces different descriptions—and thus different artifacts—the researcher is able to demonstrate the 'interpretive flexibility' of the artifact. There is not one artifact, but there are many. In the case of the ordinary bicycle, for example, there was the unsafe machine (through the eyes of women) and there was the macho machine (through the eyes of the young male ordinary users). For women the bicycle was a machine in which your skirt got entangled and from which you frequently made a steep fall; for the 'young men of means and nerve' riding the bicycle was a means to impress people (including young ladies!).

In the second step, the researcher follows how interpretive flexibility diminishes, because some artifacts gain dominance over others and meanings converge, and in the end one artifact results from this process of social construction. Here, key concepts are 'closure' and 'stabilization.' Both concepts are meant to describe the result of the process of social construction. 'Stabilization' stresses process: a process of social construction can take several years in which the degree of stabilization slowly increases up to the moment of closure. 'Closure,' a concept stemming from SSK, highlights the irreversible end point of a discordant process in which several artifacts existed next to one another.

In the third step, the processes of stabilization described in the second step are analyzed and explained by interpreting them in a broader theoretical framework: why does a social construction process follow this course, rather than that? The central concept here is 'technological frame.' A technological frame structures the interactions among the members of a relevant social group, and shapes their thinking and acting. It is similar to Kuhn's concept of 'paradigm' with one important difference: 'technological frame' is a concept to be applied to all kinds of relevant social groups, while 'paradigm' was exclusively intended for scientific communities. A technological frame is built up when interaction 'around' an artifact begins. In this way, existing practice guides future practice, though without logical determination. The cyclical movement thus becomes artifact, technological frame, relevant social group, new artifact, new technological frame, new relevant social group, etc.

Typically, a person will be included in more than one social group, and thus also in more than one technological frame. For example, the members of the Women's Advisory Committees on Housing in the Netherlands are included in the technological frame of male builders, architects, and municipality civil

servants. This allows them to interact with these men in shaping public housing designs. But at the same time, many of these women are included in the feminist technological frame, which enables them to formulate radical alternatives to the standard Dutch family house that dominates the male builders' technological frame (Bijker and Bijsterveld 2000).

This three-step research process thus amounts to: (a) sociological analysis of an artifact to demonstrate its interpretive flexibility, (b) description of the artifact's social construction, and (c) explanation of this construction process in terms of the technological frames of relevant social groups. The remainder of this section discusses some methodological and critical issues.

It is important to appreciate that SCOT provides a research heuristic for interpretive sociology, not a research machine to churn out case studies as Russell (1986) and Winner (1993) seem to suggest. For example, the adage 'identify all relevant social groups by searching citations of the artifact by a variety of actors' does not diminish the researcher's obligation to decide which groups are important to include in the account, and which groups only obfuscate the picture by adding useless details.

As a research heuristic, SCOT shows its debt to the history of technology by an emphasis on case studies and narrative, and to SSK by a preference for qualitative methods such as ethnography. Besides providing a theory of technology and society (as discussed in the next section), SCOT aims at helping researchers to give a detailed and insightful account of the development of technology in society. Concepts such 'relevant social group' and 'interpretive flexibility' primarily help to open up a case and to ask original and thought-provoking questions; they are less important for telling the story (for a debate on this issue, see Buchanan 1991, Law 1991, Scranton 1991). These case studies may be historical or contemporaneous. In the latter, the researcher will often find herself or himself drawn into the case and become one of the actors (see, for example, Bucciarelli 1994)

It is also a mistake to think that studies of SCOT only focus on the early stages of an artifact's development, the design and prototype phases. If the key idea of SCOT—that all kinds of relevant social groups contribute to the social construction process—is taken seriously, then one should also accept that this construction process takes place in all contexts and all phases. Thus, for example, the rural car in the US acquired its specific functionality (including its use as stationary power supply to washing machines, butter churns, corn shellers, water pumps, and wood saws) only after it had come off the production line as a car (Kline and Pinch 1996).

Two concepts may be fruitfully used to highlight one specific methodological cornerstone of research on SCOT: seamless web and symmetry. The phrase 'seamless web of technology and society' is often used as a reminder that non-technical factors are important for understanding the development of technology, and that these factors are closely related. A second, more sophisticated meaning is that it is never clear *a priori* and independent of context whether an issue should be treated as technical or social. Was the Challenger accident a technical failure, an organizational mistake, or primarily a lack of adequate funding (Vaughan 1996)? The activities of engineers and designers are best described as heterogeneous engineering, network building, or system building, rather than as straightforward technical invention. The recognition that all kinds of social groups are relevant for the construction of technology supports this second usage of the seamless web metaphor. The third interpretation of seamless web extends the symmetry principle into the principle of general symmetry. Not only should we treat working and non-working machines symmetrically; the very distinction between the social and the technical world can be treated in that way. The 'stuff' of the fluorescent lamp's invention is economics and politics as much as electricity and fluorescence. Let us call this 'stuff' sociotechnology. The relations that play a role in the development of the fluorescent lamp are thus neither purely social nor purely technical: they are sociotechnical. The principle of general symmetry casts this into a methodological principle: human and non-human actors should be treated similarly, and the construction of society and the construction of technology should be explained in symmetrical terms. In other words, technical reductionism and social reductionism are both outlawed as explanatory strategies. Instead, new forms of explanation need to be developed. The next section ends with an example. (For the lamp case, see Bijker 1995a; for the general symmetry principle, see Callon 1986.)

4. *The Social Construction of Technology as a Theory of Technology in Society*

SCOT also provides a theoretical perspective on the relation between technology and society, and on their joint developments or co-evolution.

As was discussed, the development of technology can be explained as a social process in which a variety of relevant social groups participate. This social process does not stop when an artifact leaves the factory, but continues when users give the technology its specific usage and meaning. Laws of physics and economics are relevant but insufficient to characterize technology's development. Artifacts can be shown to have interpretive flexibility, due to the different constructions that various social groups make. When these relevant social groups interact, typically one meaning of the artifact will gain dominance and stabilize. At the end of that social process, closure occurs, and interpretive flexibility vanishes.

Analogously, the development of social institutions in modern society cannot be fully understood without taking into account the role of technology. Such institutions as the church, capital, the government, labor, communications, and education may keep the social fabric of society together; but where would all of these be without technology? Social order in modern society can only be explained by reference to technology (Latour 1992). This requires a view of technology's impact on society. In SCOT, conceptualizing the hardness or obduracy of technology does this.

An artifact can be hard in two distinctly different ways. The first form, 'closed-in hardness,' occurs when the humans involved have a high inclusion in the associated technological frame. For example, upon meeting problems with printing a letter, a student with a high inclusion in the PC users' technological frame will check cabling, control panels, and printer definitions. Only after some time will it occur to her that she could also write the letter by hand: she was 'closed-in' by the PC technology. Another student having the same problem, but with a low inclusion in the PC users' technological frame, has no clue about cables or control panels; he can barely locate the power switch. He experiences the second kind of impact by technology: 'closing-out obduracy.' He sees no alternative but to leave the technology aside and pick up his fountain pen. In both cases, the technology has an impact on these students, but in completely different ways.

These two forms of hardness of technology can also be seen on the societal level. Automobile technology, for example, exerts a 'closed-in hardness' on the inhabitants of Los Angeles: much differentiation within auto culture, but no alternatives outside it. The standardization of main power voltage and wall plugs means a 'closing-out obduracy' to most people: accept it by buying the right plug and apparatus, or not use electricity.

This constructivist analysis provides a theoretical perspective on both SCOT and the technological impact on society. It offers a reconciliation of the previously opposed social constructivist and technological determinist views. This reconciliation asks for a theoretical elaboration, extending the conclusion of the previous section: that the distinction between technology and society can be transcended and that the proper subject matter for analysis is sociotechnology. The concept 'technological frame' provides the theoretical linking pin between the two views.

A technological frame describes the actions and interactions of actors, explaining how they socially construct a technology. But since a technological frame is built up around an artifact and thus incorporates the characteristics of that technology, it also explains the influence of the technical on the social. Part of the societal impact of the standard Dutch family house was that it dominated architectural thinking in the 1950s through 1970s, and thus made it very difficult to conceive alternatives: architects and even members of the critical Women's Advisory Committees were 'closed-in' by the builders' technological frame, only seeing the two-parents-two-children-house.

5. *The Politicization of Technological Culture*

We live in a technological culture: our modern, highly developed society cannot be fully understood without taking into account the role of science and technology. SCOT offers a conceptual framework for politicizing this technological culture. 'Politicizing' here means showing hidden political dimensions, putting issues on the political agenda, opening issues up for political debate. SCOT approach not only gives an affirmative answer to Winner's (1986) question 'Do artifacts have politics?' but it also offers a handle to analyze these politics. Technology is socially (and politically) constructed; society (including politics) is technically built, and technological culture consists of sociotechnical ensembles. (See Sclove's manifesto 1995 for designing a democratic technological society.)

Studies that elaborate this agenda do not exclusively extend the SCOT framework, but typically draw on social constructivist technology studies in the broad sense. One of the most fruitful bodies of work is the analysis of gender and technology (Cockburn and Ormrod 1993 and Lerman et al. 1997) (see *Gender and Technology*). Another now quickly expanding domain of research focuses on the politicization of information society and information and communication technologies (see Schmidt and Werle 1998 for a fruitful approach via standardization issues). Studies like MacKenzie's (1990 history of guided missiles, Hecht's (1998) history of nuclear power in France, and Vaughan's (1996) account of the Challenger disaster demonstrate that the new framework can be productive for the analysis of classical political themes too (Jasanoff 1994, Wynne 1988) (see also *Risk, Sociology and Politics of*).

The issue of political decision-making about technological projects acquires a special guise in the light of SCOT (see also *Technology Assessment*). If it is accepted that a variety of relevant social groups are involved in the social construction of technologies and that the construction process continues through all phases of an artifact's life cycle, it makes sense to extend the set of groups involved in political deliberation about technological choices. Thus, several countries experiment with consensus conferences, public debates, and citizens' juries. One of the key issues here is the role of expertise in public debates (see *Science and Technology Studies: Experts and Expertise*). The SCOT approach suggests that all relevant social groups have some form of expertise, but that no one form—for example that of scientists or engi-

neers—has a special and *a priori* superiority over the others.

See also: Actor Network Theory; Gender and Technology; History of Technology; Laboratory Studies: Historical Perspectives; Risk, Sociology and Politics of; Science and Technology Studies: Experts and Expertise; Scientific Controversies; Scientific Knowledge, Sociology of; Strong Program, in Sociology of Scientific Knowledge; Technological Determinism; Technology Assessment

Bibliography

Berger P L, Luckmann T 1966 *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Anchor Books, New York

Bijker W E 1995a *Of Bicycles, Bakelites, and Bulbs. Toward a Theory of Sociotechnical Change*. MIT Press, Cambridge, MA

Bijker W E 1995b Sociohistorical Technology Studies. In: Jasanoff S, Markle G E, Petersen J C, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, London, pp. 229–56

Bijker W E, Bijsterveld K 2000 Women walking through plans: Technology, democracy and gender identity. *Technology and Culture* **41**: 485–515

Bijker W E, Hughes T P, Pinch T J (eds.) 1987 *The Social Construction of Technological Systems. New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, MA

Bijker W E, Law J (eds.) 1992 *Shaping Technology/Building Society. Studies in Sociotechnical Change*. MIT Press, Cambridge, MA

Bucciarelli L L 1994 *Designing Engineers*. MIT Press, Cambridge, MA

Buchanan R A 1991 Theory and narrative in the history of technology. *Technology & Culture* **32**: 365–76

Callon M 1986 Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In: Law J (ed.) *Power, Action and Belief: A New Sociology of Knowledge?* Routledge and Kegan Paul, Boston, pp. 196–233

Cockburn C, Ormrod S 1993 *Gender and Technology in the Making*. Sage Publications, London

Constant E W 1980 *The Origins of the Turbojet Revolution*. Johns Hopkins University Press, Baltimore, MD

Cowan R S 1983 More Work for Mother: *The Ironies of Household Technology from the Open Hearth to the Microwave*. Basic Books, New York

Douglas S J 1987 *Inventing American Broadcasting, 1899–1922*. Johns Hopkins University Press, Baltimore, MD

Hecht G 1998 *The Radiance of France. Nuclear Power and National Identity After World War II*. MIT Press, Cambridge, MA

Hughes T P 1983 *Networks of Power. Electrification in Western Society, 1880–1930*. The Johns Hopkins University Press, Baltimore, MD

Jasanoff S (ed.) 1994 *Learning from Disaster: Risk Management after Bhopal*. University of Pennsylvania Press, Philadelphia, PA

Kline R, Pinch T 1996 Users as agents of technological change: The social construction of the automobile in the rural United States. *Technology and Culture* **37**: 763–95

Kranakis E 1997 *Constructing a Bridge. An Exploration of Engineering Culture, Design, and Research in Nineteenth-century France and America*. MIT Press, Cambridge, MA

Latour B 1992 Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker W E, Law J (eds.) *Shaping Technology/Building Society Studies in Sociotechnical Change*. MIT Press, Cambridge, MA, pp. 225–58

Latour B, Woolgar S 1986 *Laboratory Life: The Construction of Scientific Facts*, 2nd edn. Princeton University Press, Princeton, NJ

Law J 1991 Theory and narrative in the history of technology: response. *Technology and Culture* **32**: 377–84

Lerman N E, Mohun A P, Oldenziel R 1997 Gender analysis and the history of technology. *Technology and Culture* **38**: 1–231

MacKenzie D 1990 *Inventing Accuracy. A Historical Sociology of Nuclear Missile Guidance*. MIT Press, Cambridge, MA

MacKenzie D, Wajcman J (eds.) 1985 *The Social Shaping of Technology. How the Refrigerator Got its Hum*. Open University Press, Milton Keynes, UK

Nye D E 1998 *Consuming Power. A Social History of American Energies*. MIT Press, Cambridge, MA

Pinch T, Bijker W 1984 The Social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science* **14**: 399–441

Russell S 1986 The social construction of artefacts: A response to Pinch and Bijker. *Social Studies of Science* **16**: 331–46

Schmidt S K, Werle R 1998 *Coordinating Technology. Studies in the International Standardization of Telecommunications*. MIT Press, Cambridge, MA

Sclove R E 1995 *Democracy and Technology*. Guilford Press, New York

Scranton P 1991 Theory and narrative in the history of technology: Response. *Technology and Culture* **32**: 385–93

Sismondo S 1993 Some social constructions. *Social Studies of Science* **23**: 515–53

Vaughan D 1996 *The Challenger Launch Decision. Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press, Chicago

Williams R E D 1996 The social shaping of technology. *Research Policy* **25**: 865–99

Winner L 1986 *The Whale and the Reactor. A Search for the Limits in an Age of Technology*. University of Chicago Press, Chicago

Winner L 1993 Upon opening the black box and finding it empty: Social constructivism and the philosophy of technology. *Science, Technology, and Human Values* **18**: 362–78

Wynne B 1988 Unruly Technology: Practical rules, impractical discourses and public understanding. *Social Studies of Science* **18**: 147–67

W. E. Bijker

Technology-supported Learning Environments

The idea of learning environments refers to settings that are purposively designed and implemented to facilitate learning in students. The term 'learning

environments' has only recently become current in the educational literature, and reflects an important shift in the basic conception of education, namely from a predominantly information transmission and absorption model toward a more constructivist view, and, consequently, a shift in perspective from teaching to learning (Collins et al. 1996). Technology-supported learning environments are, then, contexts for learning that involve technological devices that are assumed to foster students' acquisition of the intended educational objectives. In this respect, the expression 'technology-*supported*' is also a relative newcomer in the educational vocabulary; in comparison to the frequently used term 'technology-*based*,' it attributes a less central role to the technology in the learning environment.

Using technological devices as teaching aids to facilitate learning has a relatively long history in education and educational research. A constant runs through this history as a continuous thread: initially there is great optimism and high expectations about the potential of the new technology for the improvement of the quality and the outcomes of learning and teaching, but this is afterwards followed by disappointment because the expectations are not met and the expected positive effects fail to appear. For instance, in an excellent historical overview of the classroom use of technology since 1920, Cuban (1986) quotes the following claims: 'The central and dominant aim of education by radio is to bring the world to the classroom, to make universally available the services of the finest teachers, the inspiration of the greatest leaders' (Darrow 1932, p. 79, quoted in Cuban 1986, p. 19); 'The time may come when a portable radio receiver will be as common in the classroom as is the blackboard. Radio instruction will be integrated into school life as an acceptable educational medium' (Levenson 1945, p. 457, quoted in Cuban 1986, p. 19). Obviously these predictions have remained unfulfilled. On the other hand, these statements echo many similar ones heard afterwards with respect to educational television, and more recently about the educational use of the computer. However, there is now robust evidence that in these cases too the euphoric, but naive expectations were not fulfilled (Hawkins 1996, Salomon and Gardner 1986); they were probably more based on wishful thinking than on well-grounded arguments. In sum, the history of educational technology has more or less repeated itself throughout the twentieth century, and apparently not much has been learned from it.

But paralleling the shift from an information transmission to a constructivist view of education, there has recently also been a trend toward a more productive application of educational technology that moves away from the romantic and naive technological optimism of the past, and is in accordance with the new perspective on education. Mainly focusing on the educational use of the new information and communication technology, this article will briefly document this trend, after first identifying major causes of the past failures.

1. *From Adding-on Toward Embedding Computers in Learning Environments*

A major cause of the relative failure of educational computing—as well as of previous 'novelties' in the instructional technology toolbox—is that since the early 1980s the computer has been mainly introduced in educational settings as an add-on to an existing and unchanged instructional setting (Salomon 1992). In other words, the typical application of computers could better be described as 'computer-added instruction' than as 'computer-supported learning.' In mathematics, for instance, the large majority of the available software fitted into the category of drill-and-practice programs, aiming mainly at exercising computational skills and replacing in this respect traditional worksheets (Kaput 1992). This means that the new information technology was implemented to reproduce and preserve the status quo.

The situation in other subject-matter domains of the school curriculum, but also in nonschool settings was largely similar. In language teaching, for example, programs focusing on practicing rules of spelling and grammar also prevailed, and much less software was available that supported the more essential aspects of reading and writing, namely comprehension and communication. It has now become obvious that this mere add-on strategy of educational computer use cannot produce the improvements in the quality and the outcomes of learning that were originally anticipated. A partial explanation of the inefficacy of this strategy is that the predominant drill-and-practice applications elicit only lower-level mental activity in learners, and do not at all exploit the specific potential of the computer such as its interactive possibilities and its tremendous capacity for data presentation and handling.

A more fundamental reason, however, for the failure of the add-on strategy is that it is based on the erroneous assumption that computers will evoke productive learning by themselves (Hawkins 1996). A representative illustration in this respect relates to the way Logo was often used referring to Papert (1980): it was expected that 'mindstorms' resulting in improved thinking and problem-solving skills, would arise of themselves in children's heads due to the unique characteristics of the Logo environment. Contradicted convincingly by well-designed studies, as well as by practical experience, this viewpoint has been abandoned. The point of view has rapidly gained ground that effective computer support for learning requires that it is embedded in powerful teaching–learning environments, i.e., instructional settings that elicit in students and trainees the acquisition processes neces-

sary to attain worthwhile and desirable educational objectives. Embedding means here that the computer is not just an add-on, but is judiciously integrated in the environment capitalizing on its specific strengths and potential to present, represent, and transform information (e.g., visualization and simulation of phenomena and processes), and to induce effective forms of communication, interaction, and cooperation (e.g., through exchanging data, information, and problems via a network). But, there is also a physical aspect to the embedding; indeed, as argued by Collins (1996), the spatial design of the classroom and the design of the computer are currently incompatible with the appropriate and widespread use of computers in schools.

2. *From Computer-controlled Tutoring to Student-controlled Collaborative Learning*

Paralleling the large-scale introduction of computers in education and training, the cognitive science community interested in learning and teaching has invested a lot of work and effort in the design of intelligent tutoring systems (ITS), which can be characterized as computer-supported learning environments based on artificial intelligence (AI) (for an overview see Seidel and Park 1994, Wenger 1987). It is interesting to ask whether this interdisciplinary research endeavor has contributed to overcome the flaws of educational computing discussed above. The question arises because a major incentive for designing ITS derived from dissatisfaction with traditional computer-assisted instruction (CAI). In fact, educational software involving AI was originally called 'intelligent' computer-assisted instruction (ICAI). The critical distinction between CAI and ITS is that CAI involves static systems that embody the decisions of expert teachers, whereas ITS contain the expertise itself and can use it as a basis for taking decisions about instructional interventions.

The domain of AI and education is an interdisciplinary crossroad and, consequently, the development of intelligent tutors is guided by a substantial and varied body of inquiry-based knowledge. Nevertheless the field is strewn with pitfalls. For instance, one very robust result of research on learning and instruction is that learners' prior knowledge is a very strong determinant of their future learning. Therefore, instruction should explicitly be linked up to prior knowledge, and the ITS community has taken this principle seriously. Indeed, a major component of an intelligent tutor is the student model which, as Wenger (1987, p.16) states : '… should include all the aspects of the student's behavior and knowledge that have repercussions for his performance and learning.' But the same author adds immediately that building such a student model is a very difficult task for computer-based systems. Moreover, it is not clear how far one should go in the construction of student models, and how flexible and diagnostic a system should be in view of providing the most appropriate guidance.

An even more fundamental issue concerns the nature of the guidance that an ITS should provide, taking into account the now well-documented conception that learning is an active and constructive process: Learners are not passive receptacles of information, but they actively construct their knowledge and skills through interaction with the environment and through reorganization of their prior mental structures. Consequently, as argued by Scardamalia et al. (1989), computer-based learning environments should support the constructive acquisition processes in students. The question arises whether the existing ITS are in accordance with this constructivist view of learning. Indeed, intelligent tutors that base their instructional decisions on a detailed diagnosis of student's knowledge can easily lead to a preponderance of highly structured and directive learning situations lacking sufficient opportunities for active learner involvement and participation (Kaput 1992). More recent new approaches in ITS have taken such criticisms into account, and imply a more active participation of the learner in the learning process, especially by involving (learner-controlled) interaction and cooperation between the learner and the system using a kind of built in colearner or learning companion (Frasson and Aimeur 1996).

3. *Toward a New Generation of Computer-supported Learning Environments*

Identification of the deficits of the past approaches to educational computing, together with a better theoretical understanding of the characteristics of effective learning processes, have fostered the emergence of the view that technology-supported learning environments should not so much involve knowledge and intelligence to guide and structure learning processes, but that they should rather create situations and offer tools that stimulate learners to make maximum use of their own cognitive potential (Scardamalia et al. 1989). In this connection Kintsch (1991, p. 245) has launched the idea of 'unintelligent tutoring':

> A tutor should not provide the intelligence to guide learning, it should not do the planning and monitoring of the student's progress, because those are the very activities the students must perform themselves in order to learn. What a tutor should do is to provide a temporary support for learners that allows them to perform at a level just beyond their current ability level.

The major characteristics of productive learning processes that have resulted from research on learning and instruction during the past decades can be summarized in the following definition: Learning is an

active, constructive, cumulative, self-regulated, goal-directed, situated, collaborative, and individually different process of knowledge building and meaning construction (for a more elaborated discussion see *Instructional Psychology*). In line with this conception of learning a new generation of technology-supported learning environments has emerged, characterized by a clear shift toward supportive systems that are less structured and less directive, that are more focused on coaching and scaffolding than on tutoring, that involve student-controlled tools for the acquisition of knowledge, and that attempt to integrate both tools and coaching strategies in interactive and collaborative learning environments (Cognition and Technology Group at Vanderbilt 1996). In those environments the technological tools and scaffolds are embedded in the sense as defined above. Different groups of researchers have designed and implemented computer and/or multimedia programs around which novel, powerful learning environments were built that are in accordance with the characteristics of effective learning processes and thus embody the new conception of technology-supported learning (see e.g., Lehtinen et al. 1999, Vosniadou et al. 1996). By way of illustration a few representative cases are briefly described.

A major example is the 'Jasper project' of the Cognition and Technology Group at Vanderbilt (CTGV 1997), focusing on learning and teaching mathematical problem solving in the upper primary school. The CTGV introduced the notion of anchored instruction representing one approach to implementing technology-supported learning: Instruction is anchored in videodisk-based complex problem-solving spaces that provide a richer, more realistic, and more dynamic presentation of information than textual material; the videodisk (such as 'Adventures of Jasper Woodbury' relating to trip planning) creates an environment for active, cooperative learning and discussion in small groups, as well as for individual and whole-class problem solving.

CSILE (Computer-Supported Intentional Learning Environment) is a domain-independent hypermedia environment that can be used as an educational medium throughout the curriculum (Scardamalia et al. 1994). The system allows students to build and elaborate their own common database consisting of text and graphical material relating to problems and topics under study. All students have access to the database and they can comment on each others' notes. This latter feature of the environment aims at inducing collaborative problem solving and knowledge construction in the classroom. A software tool derived from CSILE called 'Knowledge Forum' (Scardamalia and Bereiter 1998) is available on the World Wide Web.

Since 1993 the two previous projects ('Adventures of Jasper Woodbury' and CSILE) have collaborated closely with Brown and Campione's (1996) 'Fostering Communities of Learners' (FCL) program in the framework of the overarching 'Schools for Thought' (SFT) project (Lamon et al. 1996). Important in this respect is that—in line with the view of technology-supported learning put forward above—the SFT project aims at restructuring and transforming the entire learning environment in schools founded on present-day theory of learning from instruction, and integrating the new technologies (Cognition and Technology Group at Vanderbilt 1996).

A last example, Cabri Géomètre, developed in Grenoble, France (Baulac et al. 1988), consists of a series of computer programs that create a powerful learning environment in which students become active learners and explorers of Euclidean geometry guided by their teacher. The underlying ideas are that students can make their own mathematics, and that formulating and testing conjectures constitute the main activities of doing mathematics. The program elicits and facilitates such activities by offering a tool for constructing, manipulating, and exploring geometric shapes.

These technology applications are excellent examples of what Kaput (1992, p. 548) has called 'implementing technology toward reformed objectives.' Indeed, these programs embody major principles for the design of good technology-supported learning environments discussed above: They stimulate active learning oriented toward higher-order cognitive skills in a collaborative and teacher-guided context, and they exploit optimally the computer's interactive potential as well as its capacity to present and manipulate graphic and symbolic information. While a number of educational software projects, of which the three mentioned cases are representative examples, show promising results in small-scale studies, they do not get widely disseminated. A major obstacle in this respect is that most of the available products are expensive, incompatible, and inflexible. Solving this problem is a main challenge for software research and development in the near future. A promising avenue is the 'component' approach to educational software, the gist of which consists in the construction of a digital library of reusable (i.e., adaptable to new needs without the help of the original developers) software modules that can be easily mixed and matched. The further elaboration of this strategy should make it possible to reduce the cost, to increase the reuse, to decrease incompatibilities, and to enhance the flexibility of educational software (Roschelle and Kaput 1996).

4. *The Potential of the World Wide Web for Designing Powerful Learning Environments*

The end of the twentieth century saw schools rapidly acquiring access to the Internet. For instance, according to a recent survey in the USA over 90 percent of

the public schools have some kind of access to the Internet, and 39 percent of the teachers (grades 4–12) have some sort of connection in their classroom (Becker 1999). This new technological *tour de force* again fuels great enthusiasm and high expectations about the potential of the World Wide Web for enhancing learning and teaching. There is no doubt that the WWW has such a potential; however, a critical approach is indicated in order to avoid once again replicating the naivety and errors from the past with respect to other media.

Obviously the Web is a tremendously rich information source, but this is accompanied by the danger of information overload; also the Web can be used as a channel for communication, but its value as a tool for collaboration should not be overestimated (Roschelle and Pea 1999). Furthermore, making information available and facilitating communication does not yet imply that the Web will foster the quality and the outcomes of student learning; in this respect it is useful to remember that the Web was developed at first as a tool for sharing research information in the community of high energy physics scholars, and not as an educational medium. The following important issues have recently been raised from an educational perspective: the lack of integration of today's Web with the structure and content of K-12 education; the difficulty of using the Web as a medium for constructive learning (because receiving information on the Web is much easier than creating information); the inability of the Web by itself to foster higher-order skills such as problem solving, creative and critical thinking, and teamwork (Roschelle and Pea 1999). All this implies that in view of exploiting the educational potential of the WWW there is a need for thorough research on the impact of its application in classrooms on the quality and the results of students' learning and teachers' instruction, as well as for innovative development aiming at overcoming the educational weaknesses of today's Web.

See also: Cognitive Development: Learning and Instruction; Discovery Learning, Cognitive Psychology of; Environments for Learning; Explanation-based Learning, Cognitive Psychology of; Information Technology; Learning and Instruction: Social-cognitive Perspectives; Learning Theories and Educational Paradigms; Motivation, Learning, and Instruction; Science and Technology, Social Study of: Computers and Information Technology; Technological Innovation; Technology Assessment; Technology, Anthropology of

Bibliography

Baulac Y, Bellemain F, Laborde J M 1988 *Cabri Géomètre*. Cedic-Nathan, Paris

Becker H J 1999 *Internet Use by Teachers: Conditions of Professional Use and Teacher-directed Student Use* (Teaching, learning, and computing: 1998 National survey, Report 1). Center for Research on Information Technology and Organizations. University of California, Irvine, CA and University of Minnesota, MN

Brown A L, Campione J C 1996 Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In: Schauble L, Glaser R (eds.) *Innovations in Learning: New Environments for Education*. Erlbaum, Mahwah, NJ, pp. 289–325

Cognition and Technology Group at Vanderbilt 1996 Looking at technology in context: A framework for understanding technology and education research. In: Berliner D C, Calfee, R C (eds.) *Handbook of Educational Psychology*. Macmillan, New York, pp. 807–40

Cognition and Technology Group at Vanderbilt 1997 *The Jasper Project: Lessons in Curriculum, Instruction, Assessment, and Professional Development*. Erlbaum, Mahwah, NJ

Collins A 1996 Whither technology and schools? Collected thoughts on the last and next quarter century. In: Fisher C, Dwyer D C, Yocam K (eds.) *Education and Technology: Reflections on Computing in Classrooms*. Jossey-Bass, San Francisco, pp. 51–65

Collins A, Greeno J G, Resnick L B 1996 Environments for learning. In: De Corte E, Weinert F E (eds.) *International Encyclopedia of Developmental and Instructional Psychology*. Elsevier Science, Oxford, UK, pp. 687–91

Cuban L 1986 *Teachers and Machines: The Classroom Use of Technology Since 1920*. Teachers College Press, New York

Darrow B H 1932 *Radio: The Assistant Teacher*. R.G. Adams, Columbus, OH

Frasson C, Aimeur E 1996 A comparison of three learning strategies in intelligent tutoring systems. *Journal of Educational Computing Research* **14**: 371–83

Hawkins J 1996 Dilemmas. In: Fisher C, Dwyer D C, Yocam K (eds.) *Education and Technology: Reflections on Computing Classrooms*. Jossey-Bass, San Francisco, pp. 35–50

Kaput J J 1992 Technology and mathematics education. In: Grouws D A (ed.) *Handbook of Research on Mathematics Teaching and Learning*. Macmillan, New York, pp. 515–56

Kintsch W 1991 A theory of discourse comprehension: Implications for a tutor for word algebra problems. In: Carretero M, Pope M, Simons R J, Pozo J J (eds.) *Learning and Instruction: European Research in an International Context*. Pergamon, Oxford, UK, Vol. 3, pp. 235–53

Lamon M, Secules T, Petrosino A J, Hacket R, Bransford J D, Goldman S R 1996 Schools for Thought: Overview of the project and lessons learned from one of the sites. In: Schauble L, Glaser R (eds.) *Innovations in Learning: New Environments for Education*. Erlbaum, Mahwah, NJ, pp. 243–88

Lehtinen E, Hakkarainen K, Lipponen L, Rahikainen M, Muukkonen H 1999 *Computer-supported Collaborative Learning: A Review* (The J.H.G.I. Giesbers Reports on Education, Number 10). University of Nijmegen, Department of Educational Sciences, Nijmegen, The Netherlands

Levenson W B 1945 *Teaching Through Radio*. Farrar and Rinehart, New York

Papert S 1980 *Mindstorms, Children, Computers, and Powerful Ideas*. Basic Books, New York

Roschelle J, Kaput J 1996 Educational software architecture and systemic impact: The promise of component software. *Journal of Educational Computing Research* **14**: 217–28

Roschelle J, Pea R D 1999 Trajectories from today's WWW to a powerful educational infrastructure. *Educational Researcher* **28**(5): 22–5

Salomon G 1992 Effects *with* and *of* computers and the study of computer-based learning environments. In: De Corte E, Linn M C, Mandl H, Verschaffel L (eds.) *Computer-based Learning Environments and Problem Solving* (NATO ASI-Series F: Computer and System Sciences Vol 84). Springer-Verlag, Berlin, pp. 249–63

Salomon G, Gardner H 1986 The computer as educator: Lessons from television research. *Educational Researcher* **15**(1): 13–19

Scardamalia M, Bereiter C 1998 *Web Knowledge Forum. User Guide*. Learning in Motion, Santa Cruz, CA

Scardamalia M, Bereiter K, Lamon M 1994 The CSILE Project: Trying to bring the classroom into world 3. In: McGilly K (ed.) *Classroom Lessons: Integrating Cognitive Theory and Classroom Practice*. Bradford Books/MIT Press, Cambridge, MA, pp. 201–28

Scardamalia M, Bereiter C, McLean R S, Swallow J, Woodruff E 1989 Computer-supported intentional learning environments. *Journal of Educational Computing Research* **5**: 51–68

Seidel R J, Park O C 1994 An historical perspective and a model for evaluation of intelligent tutoring systems. *Journal of Educational Computing Research* **10**: 103–28

Vosniadou S, De Corte E, Glaser R, Mandl H (eds.) 1996 *International Perspectives on the Design of Technology-supported Learning Environments*. Erlbaum, Mahwah, NJ

Wenger E 1987 *Artificial Intelligence and Tutoring Systems. Computational and Cognitive Approaches to the Communication of Knowledge*. Morgan Kaufmann, Los Altos, CA

E. De Corte

Technology Transfer

The concept of technology transfer (hereafter, transfer of technology or TOT) has diverse roots: in myth, in the nineteenth century preoccupation with the diffusion of technological artifacts, and in the economics of knowledge. The Promethean act of stealing fire from the gods could be considered a creation myth of technology transfer. Lynn White's classic essay on the origin and diffusion of the stirrup is paradigmatic of the studies on the diffusion of technological artifacts (1962). But TOT as an everyday concept could only come into being with the idea of what Whitehead dubbed the 'invention of invention.' In *Science and the Modern World*, he observed, 'to understand the modern epoch, we can neglect railways, telegraph, radios, spinning machines and synthetic dyes. We must concentrate on the method itself' (1938, p. 118). Whitehead insisted that 'scientific knowledge is not an abundant storehouse from which one readily helped oneself. An intense period of imaginative design lies between them. One element in the new method is how to go about bridging the gap between a scientific idea and its ultimate product' (1938, p. 18).

The idea of the transfer of technology reflects at once a model of governance, a civics we may say, and a paradigm for the interaction between science and technology (S&T). Also present in this concept is a recognition that science and technology operated as parallel streams throughout most of history. If anything connected them, it was the use of scientific instruments. Till the nineteenth century, science was not a systematic part of the productive process. Economic historians now agree that science played little role in the industrial revolution, which was more the work of craftsman inventors who had little systematic education in science or technology. The concept of technology transfer recognizes the problematic relation between S&T and seeks to increase the reciprocity and frequency of the interaction.

In its most rudimentary form, the TOT concept is founded on the idea of the innovation chain as a sequence of the three stages of invention, innovation, and diffusion, which are non-monotonic. Invention represents the conceptualization of a scientific idea conventionally in a laboratory and its concretization in a sample or prototype. Innovation represents the scaling up of the product or process for commercialization, and diffusion is the wider extension of the product in society. Innovation chains represent the lifecycle of a scientific idea within the technological and commercial domain. The nature of the schema disguises certain ideological and cultural biases. First, it assumes that science and technology axiomatically have separate spheres of operation, which the creativity and purity of science requires. Institutionally, science was seen as operative in laboratories and universities; technology in factories and engineering studios. The innovation chain was a linear sequence proceeding from pure science to applied science to development and production. TOT was conceived as following the same linear pathway.

Second, the schema is associated with a set of conceptual separations that tended to reinforce a ritual and institutional separation among the three stages. The relation between pure science, applied science, and development provided the stuff of science policy in the USA for years. If one seminal book embodied this model, it was Vannevar Bush's *Science, The Last Frontier*, which provided the paradigmatic formulation for the science policy of the Cold War years (1945). Bush's work was representative of the time and reflected the era's static understanding of the categories of research. Bush argued that basic research is the engine of technological progress and that it must be separated from applied research because applied research invariably drives out the pure.

Third, present in Bush's writings was not only an understanding of the dynamics of science but of nation building. Bush argued that nations which depend on others for their pure science would lose their competitive edge in world trade. The TOT model thus embodied the ideological assumptions of a particular subculture of science which equated invention with basic science and superiority in basic science with

supremacy in world politics (see Stokes 1997). This subculture was challenged both by the Japanese successes in competing with America in the 1980s and by a newly revisionist history which saw S&T as a more interactive field, in which scientists like Maxwell, Pasteur, and others derived their pure science questions from industrial and agricultural problems. The linearity was also challenged by the development of the new materials sciences which posited and institutionalized a reciprocal relation between science and technology (see Roy 1970). But there were also ideological reasons for the collapse of the conventional paradigm. If the quarrel of the categories constituted the stuff of the science policy discourse in the Anglo-Saxon postwar world, the challenges to technology transfer models came from a different array of positions in the Third World.

1. *Critical Discourses*

The debates on technology transfer resonated with the wider debates on development. As development produced discursive epicycles of equal development, community development, participative development and sustainable development, technology transfer too produced a spate of critical discourses centering on intermediate technology, alternative technology, and alternative science. These responses were not merely driven by academic critiques but evolved as philosophies of grassroots movements.

Present in all these approaches was a geography of power. The innocent horizontality of TOT was now read across the center-periphery axis. The political economy of technology transfer saw the basic choices of S&T as emerging from the metropolis. The periphery was regarded not as a source of independent innovation but only as a recipient of it. The idea of TOT was embedded in the policy of import substitution; industrialization was seen as a passive civics that pulled its subjects into forms of governance over which they did not have active control.

The recognition of intellectual and economic inequality underlying TOT produced spectrum of responses. From Latin America came the work of the centre-periphery theorists who argued through LAFTA and UNCTAD for better terms of trade. Here technology was absorbed in the wider discourse of political economy. Inspired by his work in Burma, E. F. Schumacher articulated the idea of intermediate technology (1973). The Vietnam War, the energy crisis and the peace movement produced a search for soft-energy paths captured in the idea of appropriate or alternative technology (see Lovins and Price 1975). All these critiques continued to see science as universal knowledge but felt that technology was local and adaptable. They believed that local knowledges, ethnosciences, could be brought into play to produce a more local and creative technology less discordant with culture. A fourth approach, however, sought to challenge the epistemological hegemony of science, seeking an alternative basis for science through critiques of Judeo–Christian cosmology and the ethnocentrism of modern science. The Indian chemist C. V. Seshadri argued that if the notions of efficiency implicit in the laws of thermodynamics were applied to Third World societies operating at ambient temperatures, such decisions would work against tribals in the Third World, or peasants opting for indigenous or traditional agricultural practices (Seshadri 1993). The idea of intermediate technology and the emphasis on local knowledges was gradually adopted in World Bank models of technology transfer. The epistemic critique however faced a more ambiguous fate and was subjected to benign neglect.

All the above critiques tried to modify TOT as an antiseptic managerially limited discourse and to relocate it within the wider politics of knowledge and production. They reworked technology within an ethical worldview, arguing that the transfer of technology was also a transfer of values (see George 1984, pp. 75–6). Ivan Illich's work on education, medicine, and energy was a critique of the basic professionalism implicit in the TOT model (1992). He saw it as a disabling, iatrogenic programme, a cultural code for disempowerment.

The effervescence of these critical movements dimmed by the 1990s, though they continue to inspire the struggle against large dams, the critique of forestry, and the logic of urban planning in Third World countries. Intermediate technology in its heyday produced a flowering of locally sensitive technologies such as biogas and modified kitchen stoves, and also revived traditional technologies for housing (see Warren et. al. 1995). As dissenting imaginations, they were powerful, but their critical edge was blunted or absorbed by the UN sponsored idea of sustainable development. Sustainable development gave the TOT model a new lease on life by offering the possibility of a balance between ecology and development, renewability and productivity, the long run and the short run.

By the 1970s and 1980s, Science and Technology Studies (STS) had advanced beyond marginality to produce a more variegated set of discourses. The resurgence of the economics of scientific knowledge revitalized studies of the R&D system. Till the 1960s, economists thought that knowledge, like waste, was something external to the production system. Knowledge was not treated as a factor of production like land, labor, or capital. The work of Jewkes et al. (1958), Rosenberg (1976), Freeman (1974), and Machlup (1962) had a far-reaching impact in visualizing technology transfer as a policy discipline. They provided systematic studies exploring the relation between R&D expenditure and rates of productivity. They emphasized the systematic role of patents in evaluating the productivity of such systems and

introduced methodologies of cost-benefit analysis to audit TOT processes. In all, they provided a fine-grained set of analytical constructs and empirical studies that helped professionalize science policy as a discipline. They also emphasized that the rationality of the firm was altered by the military-industrial complex.

2. *Transfer of Military Technology*

The conceptual world of technology transfer still operated at a split-level with a tendency to dichotomize civilian and military transfers of knowledge. One link between the two was the concept of 'spin-off,' the idea that investments in military research eventually had civilian potential and therefore legitimized the former. The other was the idea of conversion which suggested the transformation of swords into ploughshares, a notion which became popular after the Cold War. One of the finest critiques of military innovation chains was Mary Kaldor's *Baroque Arsenal* (1982). Kaldor argued that while military technology was once innovative, it had become decadent or baroque after World War II. She traced the baroquization of military technology to the interaction between 'the capitalist dynamic of arms manufacturers and the conservatism that characterizes armed forces and defence departments in peace time' (Kaldor 1982, p. 4). Baroquization is a consequence of the paradoxical dynamic between, on the one hand, armies which cling to designs from the previous war and, on the other, the competition to stay in business and win contracts that demands incessant technological effort.

The consequence is an involuted technology of perpetual improvements within a conservative framework. What we face is a narrowing of options, greater uniformity, and diminishing returns. More and more money is expended on small improvements ultimately to obtain military hardware which is difficult to maintain. Cost factors increase to such an extent that by 2020, Kaldor suggests, the defence department might be able to afford one plane. Built into such technology as well is planned obsolescence. What was created by Alfred Sloan of General Motors for the car becomes a part of weapons systems. The useful life of a weapon is much shorter than its physical life because of the inexorable power of obsolescence. The emphasis is more on luxuries, accessories, and minor improvements.

Kaldor warned that such military technology when transferred is dangerous both for industrialized and industrializing countries. She argued that the idea of 'Atlanticism' is not just a set of treaties, but a shared technological paradigm posited on a bipolar confrontation between Western and the (then) Eastern totalitarian systems (see Kaldor 1987, p. 143). She showed how the security concerns of the UK, France, and Germany make them replicate the baroque technologies of the USA, resulting in similar cost overruns and duplication. Worse, alternative forms of defence, including Precision Guided Munitions (PGM) which implied decentralized and dispersed military organizations, could not be developed.

The impact of arms transfers is predictable in developing countries, as illustrated for instance in the regime of the Shah of Iran. The high incidence of arms transfers impairs the capacity to import what is needed for development and industrialization. Once a weapon system has been imported, a chain of supplementary imports is induced. Modern fighter aircraft, tanks, and naval units require an extensive network of support facilities including spare parts and presence of foreign experts. 'The installation of sophisticated military infrastructure is also suboptimal for the development process and often negative for existing traditional infrastructures. Simply put the stability of triage or a road for a tank has to be much higher than for a bullock cart' (Locke and Wulf 1984, p. 114). The chain of demands appears endless. This pervasive logic of military technology independently suggests that Third World nations can arrive at a self-reliant position in military technology only if they abstain from imitating the military paradigms—both doctrines and weapons systems—developed in the context of the confrontation between industrialized nations. In this context, the Vietnam War was a model confrontation between two modes of technology transfer. Unfortunately the debate did not sustain itself into the new millennium.

Technology transfer as a discourse made incremental responses to new problems. It responded gradually to issues of obsolescence, pollution, proliferation, risk, spin-off on separate grids, while also keeping civilian and military processes on separate registers. There was a need to link concepts such as obsolescence and proliferation. If obsolescence raised problems of value and time in TOT models, proliferation emphasized the question of excess. Military technology represents the classic case, but all too often military technology transfer is restricted to a limited scenario with the legitimate nation-state as actor, without realizing that the smuggler, the mercenary, and eventually even 'rogue states' may be the agents of proliferation or illegitimate transfer.

Proliferation is officially impermissible diffusion. It is a cancerous threat to the ordered chaos of diffusion. As good civics, diffusion reinforces the system of innovation. Nuclear proliferation, however, threatens it, demanding a network of laws, treaties, alliances to restrict its flow. One is simultaneously caught in the debate between atoms for peace and atoms for war, unless one agreed with Robert Jungk (1976) that the distinction was a spurious one, difficult to maintain empirically. The question of proliferation was handled at two levels—technologically as well as legally and politically. Technological measures attempted to control latent proliferation through interventions in the nuclear fuel cycle. Scientists have proposed that

proliferation resistance become a design criterion for nuclear energy systems through such techniques as isotopic denaturing. There was, however, a realization that such technical measures needed to be complemented with the political. Within the political domain, proliferation was met by two general strategies. First was the universalist argument in which, through the Nuclear Non-Proliferation Treaty and other legal measures, a full framework of integrated treaties worked for the elimination of nuclear regimes. The other, more restricted effort seeks to argue that a *controlled* spread of nuclear energy alters the behaviour of adopting states, adding a dose of restraint and confidence to their behaviour.

3. *TOT and the Risks of Development*

Throughout the 1960s to the 1980s, the TOT discourse operated between the axes of disarmament and development. It took four decades of development for experts to realize that development, the chief objective of TOT, might produce more refugees than war. India by some counts has over 40 million refugees from development projects; in fact it arguably has more refugees from development than from the wars it has fought after independence. Not unnaturally, some grassroots groups considered TOT as a continuation of war by other means (see for instance, Visvanathan 1997, p. 41).

The accidents at Windscale, Chernobyl and Bhopal introduced into the public sphere of TOT discourse the concept of risk and the more comprehensive idea of the risk society (Beck 1992). The 1984 gas leak at Bhopal, India, the world's biggest industrial disaster, proved that international transfers of technology needed to be located within a wider ethics of responsibility than mere compensation. One needed not only a more systematic right to information and disclosure but also ideas of justice and rehabilitation beyond monetary compensation. The issues of pollution and the proliferation of hazardous waste facilities introduced into the TOT domain new and enlightened legal ideas, including the polluter pays principle. Actors within the legal system realized that as technology moves from certainty to uncertainty, and even to risk, one has to redefine both the rationality of technology and the notion of responsibility. With the rise of environmentalism, risk assessment became a significant element of the TOT discourse. It also brought to the forefront the latent public aversion to science, as enacted in citizens protests. One had to ask whether citizens' worries about environmental hazards from carcinogenic pesticides, nuclear wastes, loss of global ozone was grounded, and then to develop methodologies for resolving such controversy. The work of Douglas and Wildavsky (1982), Brian Wynne (1987), Kristin Shrader-Frechette (1991) were all attempts in that direction, but with different sensitivities to the requirements of science and the demands of democracy.

The question of risk also prompted inquiry into the rationality behind technology. Ulrich Beck (1992) went so far as to institutionalize the problem. Beck argued that in advanced industrial society the social production of wealth is accompanied by the social production of risks. He claimed one needed a new concept of a risk society with which one not only acknowledged that modernization produces risks, but one also systematically and openly sought to answer how risks could be 'prevented, minimized, dramatized and extended' (Beck 1992, p. 21). The notion of the risk society led to a range of debates in the social sciences, raising interesting questions about the role of experts in technology transfer, the question of trust in science, and the construction of scientific knowledge itself.

Meanwhile a revisionist history of the Green Revolution was altering the perception of one of the paradigmatic successes of the technology transfer discourse. The Green Revolution, a project which involved the transfer of hybrid, fertilizer-sensitive seeds to India, Pakistan, and other developing countries, was originally seen as a benign example of successful TOT which saved these countries from starvation and made India self-sufficient in food. Norman Borlaug, one of pioneers of the Green Revolution, received the Nobel Prize for peace. Robert Anderson challenged the received wisdom of the Green Revolution by arguing that it involved not merely high yielding packages of seed, pesticides and fertilizer but also incorporated social ideas of changing agricultural society. 'The broad economic links between modernized agricultural production and US economic interests were crucial elements in the strategy from the beginning' (Anderson et. al. 1991, p. 361). Francesca Bray, arguing from a different perspective, showed that Green Revolution concentrated on only a few crops, 'wheat, and rice for human consumption, corn and soybean for animal feed and cotton for industry' (Bray 1994, p. 30). The capital-intensive, highly mechanized model hastened soil erosion and environmental degradation. Bray and later Vandana Shiva (1992) particularly attacked the monocultural mindset behind such transfer of technology, which ignored the productivity and sustainability of polycultural systems, which not only used several varieties of crop but had several uses for various parts of the crop. Such crops were also resistant to pests. It was the impoverishment of gestalts that constitutes the most devastating aspect of the critique of the Green Revolution as TOT.

4. *TOT and Biotechnology*

But narratives change and so do battle sites. By 1990, even its pioneers recognized the Green Revolution as a dated narrative. The productivity levels of wheat and

rice had levelled off and the scene shifted to biotechnology as the site for the second Green Revolution. While critiques of the Green Revolution came after the event, the movements against biotechnology were spawned almost simultaneously.

The question of the transfer of biotechnology raises fundamental questions for the innovation process in the Third World countries. Pat Mooney (1983, 1998) documents how many crops developed by farmers over centuries became the property of a multinational at the stroke of a pen. Patenting, which was seen as a legal frame for recognizing original innovation, became a technique for appropriating the legacy of innovation in agricultural societies. Important controversies in the 1990s centred around *neem* and *basmati* rice. The growing opposition to chemical products led to a search for natural pesticides and to the *neem*. By 1985, US and Japanese firms had taken out a dozen patents on *neem* pesticides and *neem* toothpaste. The question one has to ask is whether *neem* patents constitute novel forms of knowledge or are such forms of novelty only based on the ignorance of a 2000-year-old tradition of biopesticides in India?

The controversy over *basmati* rice raised similar issues. In 1997, a patent was granted to Ricetech, Inc., a US company, on *basmati* lines and grains. The *basmat*i variety was derived from the Indian *basmati* and crossed with semi-dwarf varieties. The question is whether Ricetech *basmati* is the Indian *basmati* and novel at the same time. If *basmati* patents were recognized in India, Indian farmers would ironically be forced to pay royalties to Ricetech.

A patent does not recognize the prior custody of farmers over their seeds. There is no recognition of prior art. Culturally, the patent defines agriculture as an activity performed in a laboratory rather than on a farm. It is a derecognition of an entire intellectual commons. The juridical framework of GATT and TRIPS has possibilities of becoming a new enclosure movement—comparable in spirit to the actions of fifteenth century British landowners—which turns a farmer who saves and replants the seed of patented or protected plant into a violator of the law.

There was a second, equally fundamental problem. Hybridization disables the seed from reproducing itself and forces the farmer to return to the breeder for further stock. Farmers who were custodians of the seed now were systematically deskilled both by technology and law. Finally, there is the ethical question of whether life forms can be patented at all as many of these communities locate the seed within the domain of the commons and the gift.

4.1 TOT and Information Technology

If biotechnology appears as a dismal science for the TOT process, information technology, as a narrative, was more celebratory. The rise of technologically innovative groups in Silicon Valley, California, created a new paradigm for industrialism outstripping in its impact Boston's legendary Route 128 (Castells 1996). Silicon Valley created a false belief that distance and location did not matter in the transfer of technology. Over the last decade, governments in India, China, Malaysia have pursued the holy grail of reinventing Silicon Valleys in their countries. Information technology and the legend of the innovative successes of Bill Gates and Sabir Bhatia, of Oracle, Sun Microsystems, and Intel created a myth that innovation chains could reproduce themselves anywhere. As Manuel Castells remarked:

> The development of the information technology revolution contributed to the formation of the milieux of innovation where discoveries and applications would interact, and be tested, in a recurrent process of trial and error, of learning by doing; these milieus required (and still do in the 1990s, in spite of on-line networking) spatial concentrations of research centres, higher education institutions, advanced technology companies, a network of ancillary supplies of goods and services, and business networks of venture capital to finance start-ups. (Castells quoted in Mcgray 1999, p. 153)

Information Technology is symptomatic of the cornucopian myths present in TOT. At its most reductive level, technology transfer is a methodology of hope combining the powers of science and technology. It appears as a discourse on management but, as it unravels, its reaches beyond management to the complexities of law, political economy, ethics and anthropology. What begins as a science policy issue transforms itself into a multidisciplinary exercise. Technology transfer is one of the great myths of the twentieth century, a belief that science and technology, if naturally diffused, can fulfil any democratic dream. It is the task of science and technology studies to show the lags, the gaps, the ironies that sustain the myth. What one needs to chronicle is these very lags, gaps, worries and possibilities as the myth of TOT acts itself out in the form of policy.

See also: Evolution: Diffusion of Innovations; History of Technology; Innovation and Technological Change, Economics of; Military and Politics; Science, Technology, and the Military; Technological Innovation; Technology Assessment; Technology, Social Construction of

Bibliography

Anderson R, Levy E, Morrison B M (eds.) 1991 *Rice Science and Development Politics*. Clarendon Press, Oxford

Beck U 1992 *Risk Society: Towards a New Modernity*. Sage, London

Beer J J 1959 *The Emergence of the German Dye Industry*. University of Illinois Press, Urbana, IL

Bray F 1994 Agriculture for developing nations. *Scientific American* **271**: 30–7

Bush V 1945 *Science, The Endless Frontier*. Government Printing Office, Washington
Castells M 1996 *The Rise of the Network Society*. Blackwell Publishers, Oxford
Donahue D 1973 *Thieves of Fire*. Faber and Faber, London
Douglas M, Wildavsky A 1982 *Risk and Culture*. University of California, Berkeley, CA
Freeman C 1974 *The Economics of Industrial Innovation*. Penguin Books, Harmondsworth, UK
George S 1984 *Ill Fares the Land*. Institute of Policy Studies, Washington, DC
Illich I 1992 *In the Mirror of the Past*. Lectures and addresses. Marion Boyars, New York
Jewkes J, Sawyers D, Stillerman R 1958 *The Sources of Invention*. Macmillan, London
Jungk R 1976 *The Atom Staat* (*The New Tyranny*). Warner Books, New York
Kaldor M 1982 *The Baroque Arsenal*. Andre Deutsch, London
Kaldor M 1987 The Atlantic technology culture. In: Kaldor M, Falk R (eds.) *Dealignment*. The United Nations University and Basil Blackwell, Oxford, UK, pp. 143–62
Krohn W, Layton E J Jr, Weingart P (eds.) 1978 *The Dynamics of Science and Technology. Yearbook of the Sociology of the Science*. D. Reidel Publishing Company, Dordrecht, The Netherlands
Locke P, Wulf H 1984 Consequences of the transfer of military-oriented technology on the development process. In: Pradip G (ed.) *Disarmament and Development*. Greenwood Press, Westport, CT, pp. 104–19
Lovins A, Price J H 1975 *Non-nuclear Futures: The Case for an Ethnical Energy Strategy*. Harper and Row, New York
Machlup F 1962 *The Production and Distribution of Knowledge in the United States*. Princeton University Press, Princeton, NJ
McGray D 1999 The silicon archipelago. *Daedalus* **128**(2) Spring: 147–76
Mooney P R 1983 The law of the seed: Another development and plant genetic resources. *Development Dialogue* **1–2**: 3–172
Mooney P R 1998 The parts of life: Agricultural biodiversity, indigenous knowledge and the role of the third system. Development Dialogue (Special Issue), 6–181
Rosenberg N 1976 *Perspectives on Technology*. Cambridge University Press, Cambridge, UK
Roy R (ed.) 1970 *Materials Science and Engineering in the United States*. Pennsylvania State University Press, University Park, PA
Schmookler J 1966 *Invention and Economic Growth*. Harvard University Press, Cambridge, MA
Schumacher E F 1973 *Small is Beautiful: A Study of Economics as if People Mattered*. Blond and Briggs, London
Schrader-Frechette K 1991 *Risk and Rationality*. University of California Press, Berkeley, CA
Seshadri C V 1993 *Equity is Good Science*. A. M. M. Murugappa Chettiar Research Centre, Madras, India
Shiva V 1992 *The Violence of the Green Revolution*. The Other India Press, Goa, India
Stokes D 1997 *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution Press, Washington, DC
Visvanathan S 1985 *Organizing for Science: The Making of an Industrial Research Laboratory*. Oxford University Press, Delhi, India
Visvanathan S 1997 *A Carnival for Science*. Oxford University Press, Delhi, India
Warren M D, Slikkerveer J L, Brokensha D (eds.) 1995 *The Cultural Dimension of Development*. Intermediate Technology Publications, London
White L 1962 *Medieval Technology and Social Change*. Oxford University Press, Oxford, UK
Whitehead A N 1938 *Science and the Modern World*. Pelican Books, Harmondsworth, UK
Wynne B 1987 *Risk Management and Hazardous Waste*. Springer-Verlag, New York

S. Visvanathan

Teen Sexuality

Since the 1950s, societal views of sexuality have changed dramatically, and with those changes, there has been an increased focus on teenage sexuality. Sexual activity among teenagers has increased while the age that teenagers begin sexually experimenting has decreased. Although the exact nature of the association is unclear, sexual intercourse at an early age has been consistently linked with greater psychological risks (e.g., problem behavior, substance use, poor school performance, and disrupted family and parental relationships). Unsafe sex practices and limited sexual and contraceptive knowledge place adolescents at high risk for sexually transmitted diseases (STDs) and early childbearing (Bingham and Crockett 1996).

1. *Definition*

'Teenage sexuality,' a difficult term to define, includes two complex themes: teenage and sexuality. By teenage, most researchers refer to the adolescent years or adolescence. Adolescence begins at puberty, with physiological and hormonal changes. While entrance is defined largely by biological markers, exit has become less clear. Within contemporary Western society, the legal age of 18 has been the traditional marker for the end of adolescence and the beginning of adulthood. With a changing economy and societal structure, for many, education extends beyond the age of 18 and traditional hallmarks of adulthood (e.g., financial independence, marriage, and childbearing) are postponed. Given that adolescence covers the ages of approximately 12 to 21 years, researchers often talk about early, middle, and late adolescence, each period with its own developmental characteristics.

Sexuality encompasses a wide array of domains including sexual behaviors, attitudes, and knowledge (Miller et al. 1986), sexual identity (see Miller and Simon 1980), sexual orientation (see Savin-Williams and Rodriguez 1993), contraception, and pregnancy (see Piccinino and Mosher 1998). Although the sociali-

zation of sexuality begins in infancy, when gender role expectations and identity are imparted, it is in adolescence—when biological factors interact with social expectations—that sexuality becomes a central issue of development. Within the context of other developmental goals, the adolescent is expected to become a 'self-motivated sexual actor' (Miller and Simon 1980). This article will focus on sexuality among heterosexual adolescents, including current statistics on behavior and attitudes, an overview of theoretical and historical perspectives, measurement issues, and directions for future research.

2. *Current Statistics*

Since the 1950s, we have seen a dramatic increase in the rates of sexual activity among teenagers in the USA. Teenagers of both sexes are having sexual intercourse at younger ages than ever before. By the age of 17, more than 50 percent of teenagers report having had sexual intercourse, and by age 19 that figure jumps to approximately 80 percent (Singh and Darroch 1999). Traditionally, boys begin having sexual intercourse earlier than girls, but that gap has decreased steadily in recent years (AGI 1994). Sexual intercourse before the age of 15 is still relatively rare with approximately 80 percent of teenagers reporting that they are sexually inexperienced at the age of 15. Sexual intercourse prior to the age of 13 is reported by adolescents as mostly nonvoluntary or unwanted (Singh and Darroch 1999).

Encouragingly, pregnancy rates among teenagers have declined 17 percent from 1990 to 1996. Contraceptive use at first intercourse increased during the 1980s from 48–65 percent with the use of condoms doubling. By 1995, 78 percent of sexually active teenagers reported using some form of contraception at first intercourse, with 66 percent of these adolescents using condoms. Overall, 90 percent of sexually active teenagers report using some form of contraception. Unfortunately, a large percentage of this group do not use contraceptive methods consistently or correctly (Piccinno and Mosher 1998). Despite an increase in contraceptive use, 25 percent of sexually active teenagers will still contract an STD during adolescence, and 19 percent of sexually active girls will get pregnant before the age of 19. Seventy-eight percent of teenage pregnancies are reportedly unplanned.

3. *Theoretical and Historical Perspectives*

No single theory has adequately explained adolescent sexuality. In their overview of several theoretical perspectives on adolescent heterosexual behavior, Miller and Fox (1987) focus on two major but conflicting paradigms on sexuality: sexuality as an emergent drive and sexuality as a socially learned behavior. The first paradigm, sexuality as emergent drive, encompasses biologically-based (e.g., Udry et al. 1985), psychologically-based theories (e.g., Jessor and Jessor 1977, Miller et al. 1986), and theories of social control (e.g., Hirschi 1969, Parsons 1951).

Biological influences help explain some of the social changes in adolescent sexual behavior (Dyk 1993). Since the mid-nineteenth century, the onset of puberty has dropped gradually in age for both boys and girls. For example, in the 1850s the average age of menarche was 17 years. By 1990s, the average age was 12 years (Dyk 1993). In support of biological theories, a growing body of research suggests that male hormones play an identifiable role in shaping the sexual behavior of adolescents, both boys and girls. For example, a measure of serum androgenic hormones, known as the 'free testosterone index,' was shown to be predictive of virginity status, sexual behavior, and noncoital sexual behaviors among adolescent virgins (e.g., hand holding, kissing) (see Udry et al. 1985).

The view of sexuality as an inner drive also encompasses psychologically-based theories. Early sexual behavior among adolescents has been portrayed as an attempt to satisfy a strong need for affiliation or intimacy (Miller et al. 1986), or a desire for excitement or novelty (Miller and Fox 1987). Although self-esteem and expectations for achievement have been touted as the best predictors of sexual behavior, results from longitudinal studies have yielded inconsistent and contradictory results (Jessor and Jessor 1977).

Social control theories have sought to explain why many adolescents do not act on their biological impulses and look to processes that serve as restraints. Parsons (1951) and Hirschi (1969) propose models of social constraint centering on the roles of parents. In Parsons' model, parents 'serve as policing agents of the social order' (Miller and Fox 1987, p. 273). Parents deter the sexual behavior of their adolescents through rules, supervision, and surveillance. Hirschi's (1969) model stresses the parent–child bond as the constraint mechanism. Because of the emotional bonds formed with parents, and consequent desire to please, adolescents conform to parental expectations.

In contrast to emergent drive-based theories, socialization paradigms (i.e., the process through which acceptable behaviors in a given society are taught and learned) posits that sexuality is socially shaped more than internally driven. Areas of inquiry include how society shapes the expression of sexual behavior, how certain sexual behaviors come to be viewed as appropriate or inappropriate, and how erotic meanings come to be attached to certain objects, persons, or behaviors. Because parents serve as models through which accepted behaviors and norms of a society are learned, parents are considered crucial socialization agents and their influence has been widely investigated. Communication, particularly discussions about sexual issues between mothers and their young adolescents, has a strong influence on the attitudes adolescents

have about sexuality and sexual behaviors (Jessor and Jessor 1977, Treboux and Busch-Rossnagel 1995) During this age period, peers also serve as strong socialization agents. Adolescents often exert explicit pressure on each other to conform to a standard of behavior and look to their peers as a reference point or standard by which they can measure their own behavior. For example, adolescent girls are less likely to become sexually active if their best friends are virgins as well (Billy and Udry 1985). Research suggests that the relative influence of parents and peers may differ as a function of the age of the adolescent. For younger adolescents, parents appear to have a strong influence on sexual behaviors and attitudes, but this influence declines as teenagers approach later adolescence and peers become more influential (see Treboux and Busch-Rossnagel 1990, 1995).

In addition to proximal influences on behavior, contextual factors (e.g., religion, economy, and culture) shape views of teenage sexuality. Reiss' (1967) sociological theory of premarital sexual permissiveness stresses that changes in individual sexual behavior mirror shifts in social views. Prior to the 1960s, intercourse outside of marriage for adults was widely disapproved, especially for women. In the 1990s there was a greater acceptance of premarital sexual activity among adults (see Downs and Hillje 1993 for an overview of historical perspectives on adolescent sexuality). Although social attitudes regarding adult sexual behavior and preadolescent behaviors are clear across Western subcultures, mixed attitudes regarding adolescent sexual behaviors prevail. Young people are biologically defined as sexually mature while simultaneously defined as socially and psychologically immature. Caught in a transitional stage between childhood and adulthood, the task of integrating differing views from the larger society as well as peer subcultures may be quite confusing for adolescents (Graber et al. 1998).

Some researchers have integrated the two approaches, examining both innate and learned influences. Using a biosocial model, Smith (1989) argues that biological factors (i.e., onset of puberty, the variance in hormonal levels) play a clear role in adolescent sexual behavior, but these biological changes are mediated by social and cultural forces and expectations. Smith further suggests that these two influences may fluctuate, with one force becoming stronger than the other at different times. A late maturing girl may not have a strong biological influence to become sexually active at an early age, but within a social setting that accepts and encourages sexual promiscuity (e.g., a peer group of early maturing sexually active girls), she may be influenced to engage in sexual activity. Within a biosocial framework, both components are consistently interacting to affect sexual behavior and attitudes (Smith 1989). Although useful as a conceptual model, it has yet to be tested empirically.

4. *Methodological Concerns*

The nature of the population (i.e., minors) and the topic lead to inherent methodological problems, namely sample and self-report biases. As required of all minors, parental consent is mandatory and often difficult to obtain. Parents are often concerned that simply being asked questions of a sexual nature will encourage their young adolescents to become sexually active.

Even when parents consent, teenagers may be tempted to overstate behavior if they are embarrassed at their lack of sexual activity. In a longitudinal study, Alexander et al. (1993) examined the issue of consistency in reports along three dimensions: whether or not respondents had ever had sexual intercourse, the number of lifetime intercourse experiences, and the age of first intercourse. One year later, 90 percent of adolescents were consistent in their reports of virginity, and 85 percent were consistent in reporting the number of lifetime experiences. However, only 43 percent were consistent in reporting their age at first intercourse.

Designed to explore the causes of adolescent health behaviors, the National Longitudinal Study of Adolescent Health (Udry and Bearman 1998) uses technology to address such methodological problems. Adolescents, grades 7 through 12, enter their responses directly into a computer. Automatic skip patterns prevent students from viewing questions that they are not meant to answer. For example, if participants respond that they are not sexually active, they will not see questions regarding sexual partners, contraception, and so forth. This is an important feature for several reasons: (a) it helps assuage parental concerns that children will 'get ideas' from sex-related questions; (b) it prevents teenagers from lying to match their peer behavior; and (c) the absence of interviewers in the room and a hard copy at the end of the session promotes assurance of confidentiality of responses.

5. *Directions for Future Research*

To date, most psychological and sociological studies have taken an epidemiological approach toward adolescent sexuality. We know who is having sex, at what age, and with whom. We also know that there is an association between adolescent sexual activity and other psychosocial behaviors. However, we have yet to explore fully the meaning that sexuality holds for teenagers. Until recently, research on adolescent sexuality has not been embedded in a developmental framework. As Graber et al. (1998) stress, the adolescent life phase is a period of dramatic change and developmental challenges. Research that examines sexuality within the context of these other changes would provide more meaningful information.

Although researchers stress the importance of considering the influences of culture and religion, as well as integrating the influence of biology, there is sparse research exploring the roles of these influences and/or examining the contexts of community, school, peer group, family, friends, and romantic partnerships. New models and methods of research designed to take more developmental and holistic approaches to the study of adolescent sexuality are needed (e.g., Cairns et al. 1998).

The Add Health project (Udry and Bearman 1998) is currently attempting a longitudinal study using such an approach. The study is unique in its exploration of the roles various relationships—romantic sexual partners, nonromantic sexual partners, same sex as well as opposite sex relationships—are playing in the lives of teenagers. The longitudinal nature of the sample will allow us to examine the natural progression of these relationships.

Finally, because statistics point us to the middle and late years of adolescence as the age where sexual activity begins (Singh and Darroch 1999), most research naturally focuses on the sexual behavior and attitudes of adolescents over the age of 15. Bingham and Crockett (1996) found that psychological adjustment problems associated with early sexual activity are present prior to first intercourse, not as a result of early sexual intercourse. Thus, a greater understanding of the attitudes and behaviors forming in younger adolescents and middle school age children is of great import.

See also: Gay, Lesbian, and Bisexual Youth; Regulation: Sexual Behavior; Sex-role Development and Education; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Risk Behaviors

Bibliography

Alan Gutmacher Institute (AGI) 1994 *Sex and America's Teenagers*. AGI, New York

Alexander C S, Somerfield M R, Ensminger M E, Johnson K E 1993 Consistency of adolescents' self-report of sexual behavior in a longitudinal study. *Journal of Youth and Adolescence* **22**(5): 455–71

Billy J O G, Udry J R 1985 The influence of male and female best friends on adolescent sexual behavior. *Adolescence* **20**: 21–32

Bingham C R, Crockett L J 1996 Longitudinal adjustment patterns of boys and girls experiencing early, middle, and late sexual intercourse. *Developmental Psychology* **32**(4): 647–58

Cairns R B, Cairns B D, Rodkin P, Xie H 1998 New directions in developmental research: Models and methods. In: Jessor R (ed.) *New Perspectives on Adolescent Risk Behavior*. Cambridge University Press, Cambridge, UK, pp. 13–40

Downs A C, Hillje L S 1993 Historical and theoretical perspectives on adolescent sexuality: An overview. In: Gullotta T P, Adams G R, Montemayor R (eds.) *Adolescent Sexuality: Advances in Adolescent Development*. Sage, Newbury Park, CA, pp. 1–34

Dyk P H 1993 Anatomy, physiology, and gender issues in adolescence. In: Gullotta T P, Adams G R, Montemayor R (eds.) *Adolescent Sexuality: Advances in Adolescent Development*. Sage, Newbury Park, CA, pp. 35–56

Graber J A, Brooks-Gunn J, Galen B R 1998 Betwixt and between: Sexuality in the context of adolescent transitions. In: Jessor R (ed.) *New Perspectives on Adolescent Risk Behavior*. Cambridge University Press, Cambridge, UK, pp. 270–316

Hirschi T 1969 *Causes of Delinquency*. University of California Press, Berkeley, CA

Jessor R, Jessor S L 1977 *Problem Behavior and Psychosocial Development: A Longitudinal Study of Youth*. Academic Press, New York

Miller B C, Fox G L 1987 Theories of adolescent heterosexual behavior. *Journal of Adolescent Research* **2**(3): 269–82

Miller B C, McCoy J K, Olson T D 1986 Dating age and stage as correlates of adolescent sexual attitudes and behavior. *Journal of Adolescent Research* **1**(3): 361–71

Miller P Y, Simon W 1980 The development of sexuality in adolescence. In: Adelson J (ed.) *Handbook of Adolescent Psychology*. Wiley, New York, pp. 383–407

Parsons T 1951 *The Social System*. Free Press, New York

Piccinino L J, Mosher W D 1998 Trends in contraceptive use in the United States: 1982–1995. *Family Planning Perspectives* **30**(1): 4–10

Reiss I L 1967 *The Social Context of Premarital Sexual Permissiveness*. Holt, Rinehart, and Winston, New York

Savin-Williams R C, Rodriguez R G 1993 A developmental, clinical perspective on lesbian, gay male, and bisexual youths. In: Gullotta T P, Adams G R, Montemayor R (eds.) *Adolescent Sexuality: Advances in Adolescent Development*. Sage, Newbury Park, CA, pp. 77–101

Singh S, Darroch J E 1999 Trends in sexual activity among adolescent American women: 1982–1995. *Family Planning Perspectives* **31**(5): 211–9

Smith E A 1989 A biosocial model of adolescent sexual behavior. In: Adams G R, Montemayor R, Gullotta T P (eds.) *Biology of Adolescent Behavior and Development*. Sage, Newbury Park, CA, pp. 143–67

Treboux D, Busch-Rossnagel N A 1990 Social network influences on adolescent sexual attitudes and behaviors. *Journal of Adolescent Research* **5**(2): 175–89

Treboux D, Busch-Rossnagel N A 1995 Age differences in parent and peer influences on female sexual behavior. *Journal of Adolescent Research* **5**(4): 469–87

Udry J R, Bearman P S 1998 New methods for new research on adolescent sexual behavior. In: Jessor R (ed.) *New Perspectives on Adolescent Risk Behavior*. Cambridge University Press, Cambridge, UK, pp. 241–69

Udry J R, Billy J O G, Morris N M, Groff T R, Raj M S 1985 Serum androgenic hormones motivate sexual behavior in adolescent boys. *Fertility and Sterility* **43**(1): 90–4

G. S. Abbott and D. Treboux

Teenage Fertility

In traditional societies, the transition from childhood to early adulthood usually occurs rapidly around the time of puberty. For women this almost always means entry into marriage and childbearing. In industrialized

societies, however, preparation for the assumption of adult responsibilities requires extensive further education and job training for both men and women. Thus, the modern way of life has led to a period of adolescence, covering roughly the teenage years, during which young people are capable of reproduction but have not yet achieved the economic and social independence necessary to support families of their own. In any given society, childbearing at this stage may be seen as more or less problematical, depending on both the level of development and the particular culture. The topic is now of special importance, because of the surge in the numbers of teenagers worldwide. The focus in this entry is almost entirely on females; although the issue clearly involves males as well, men are less likely on the whole to become fathers while they are still teenagers, the consequences of parenthood are apt to fall less heavily on them, and, most importantly, very little information is available on male fertility.

1. Patterns

1.1 Birth Rates for Women aged 15–19 Years

The most commonly used measure of teenage fertility is the age-specific birth rate for women 15–19 years old, i.e., the number of births to women in this age group divided by the number of women in the age group. Although some births occur among women below 15 years of age, relatively few of these younger women are exposed to the risk of childbearing, and the usual practice is simply to add these births to the numerator of the rate for women aged 15–19. The implications of childbearing tend to be quite different for women aged 15–17 from those aged 18–19 and, ideally, the rates for the two groups should be separated, but the necessary data are not often available.

Table 1 shows age-specific birth rates per thousand women aged 15–19 in the 1970s and the 1990s for the major regions of the world and their component subregions, as defined by the United Nations. At the regional level, the rates range from a minimum of around 30 for Europe to a maximum that is about four times as high in Africa, where as many as one woman out of eight in this age group gives birth yearly.

There is considerably more variation in the rates for subregions. North America, Australia and New Zealand and the four subregions of Europe comprise the great majority of developed countries. Within this group, the rates for the 1990s are very low in Western and Southern Europe, and are particularly elevated in Northern America (dominated by the USA) and in Eastern Europe (dominated by the former USSR). The underlying national data show that relatively high rates of teenage fertility are common to all of the English-speaking countries. Among less developed areas, the picture in Asia is extremely varied: in the 1990s, Eastern Asia (dominated by China but also including Japan, a highly developed country) exhibits the lowest rate of any subregion, while teenage fertility in South-central Asia (dominated by India, Pakistan and Bangladesh) is higher than anywhere else outside of Africa. Intermediate levels are found in Latin America and the Caribbean and in Other Oceania. The rates are extremely high in sub-Saharan Africa.

In all regions and most subregions, the rates had declined since the 1970s. At the regional level, the steepest drop (24 percent) occurred in Asia. This is due largely to an 86 percent decline in Eastern Asia and took place despite very little downward movement in South-central Asia. The rates rose substantially in two subregions, however: by one-third in Southern Africa (dominated by South Africa) and by one-quarter in Eastern Europe.

1.2 Teenage Fertility Relative to Overall Fertility

In addition to age-specific fertility rates, it is useful to consider teenage reproductive behavior in the context of that of all women, i.e., to examine the proportion of the total fertility rate (TFR) contributed by women aged 15–19. (The TFR is a hypothetical measure representing the number of children a woman would have if she experienced throughout her reproductive career the actual fertility rates of women in each successive age group at the time and place of observation.) These percentages are also shown in Table 1; in the 1990s, they range from less than one in every 30 births (3 percent) in Western Europe to one birth out of six or seven (16 percent) in Eastern Europe.

Comparison of the percentages between the 1990s and the 1970s reveals that in some cases the trends in teenage childbearing noted above parallel those among older women. In Eastern, Middle and Western Africa and Northern America, the rates for women aged 15–19 represent approximately the same proportion of the total fertility rate in the 1990s as in the 1970s. In many subregions, however, teenage rates changed independently of shifts in overall fertility. Adolescents came to account for a much larger fraction of all births than formerly in Eastern Europe and South-central Asia, and their share also grew throughout Latin America and the Caribbean and in Southern Africa. The opposite is true in other parts of Asia and Europe, especially Eastern Asia and Western Europe and, to a lesser extent, in Northern Africa as well as in Australia and New Zealand.

1.3 The Proximate Determinants of Fertility

To understand these patterns, it is necessary to examine the behaviors immediately antecedent to

Table 1
Age-specific fertility rates for women aged 15–19 and percentage contribution to total fertility rate

Region and subregion	1970s Rate per 1,000	1970s Percentage of TFR	1990s Rate per 1,000	1990s Percentage of TFR
Africa	132	9.9	117	10.6
Eastern Africa	147	10.2	137	10.6
Middle Africa	182	14.0	159	14.0
Northern Africa	114	8.7	44	5.8
Southern Africa	88	8.4	117	12.3
Western Africa	171	12.2	150	12.3
Asia	74	6.8	56	10.0
Eastern Asia	35	3.4	5	1.4
South-central Asia	114	9.9	112	15.1
South-eastern Asia	100	9.6	54	8.6
Western Asia	130	11.1	63	8.1
Europe	34	7.5	27	9.4
Eastern Europe	35	8.2	44	15.5
Northern Europe	43	9.1	24	7.1
Southern Europe	29	5.7	14	5.1
Western Europe	31	7.0	9	3.1
Latin America and the Caribbean	103	9.7	89	14.3
Caribbean	94	10.3	75	13.8
Central America	124	10.0	104	12.9
South America	96	9.4	84	14.6
Northern America	66	13.4	57	14.0
Oceania				
Australia and New Zealand	54	9.3	23	6.1
Other Oceania	92	9.0	80	8.4

Source: Tabulations from the data bank maintained by the Population Division of the Department of Economic and Social Affairs of the United Nations

childbearing. Variation in teenage fertility can be very largely explained in terms of three factors: sexual activity (marriage), contraception and abortion. These are known as proximate determinants of fertility.

A young woman is at risk of pregnancy and childbirth only after she becomes sexually active. Both age at initiation and frequency of intercourse thereafter are important dimensions of sexual activity. There is evidence that age at menarche is now dropping in developing countries as it did earlier in Western countries, creating the possibility of exposure to risk starting sooner than was previously the case. The proportion married among women of reproductive age was considered originally to be an appropriate indicator of exposure to risk and remains so in most of Asia. Elsewhere, however, this measure has become unsatisfactory because nonmarital childbearing is much more common than it was formerly in many parts of the world and because of greater awareness that in some cultures marriage is not an easily identifiable event.

Among sexually active women, the use of contraception can reduce substantially the likelihood that pregnancy will ensue. Contraceptive methods vary, however, in their effectiveness, acceptability and ease of use and thus in their probability of failure. Moreover, young women tend to be less likely than their elders to know about contraception and to have less access to acceptable services, including an adequate choice of methods. In many countries, unmet need for family planning services is concentrated in the early years of reproductive life.

When pregnancy occurs, it does not always lead to a birth. Teenagers are often unready to start childbearing and may seek induced abortion even if the procedure is illegal and unsafe. Because very young women often do not recognize or accept that they are pregnant, however, or because abortion services are not readily accessible, their pregnancies are often relatively advanced before they reach a provider. Reliable data on induced abortion exist for only a few mostly developed countries, making it difficult or impossible to assess its role accurately in many cases.

Differing levels and trends of adolescent fertility are largely a reflection of variation in the balance of these three factors. Very high rates, such as those found in sub-Saharan Africa and South-central Asia, necessarily imply a high level of sexual activity, minimal use

of contraception and infrequent abortion. In the traditional societies of sub-Saharan Africa, early childbearing has typically been widely accepted if not encouraged and often occurs before finalization of a marriage. Recently, growing numbers of young women have attempted to postpone marriage and childbearing, but many are nevertheless sexually active, tending to keep fertility high. In South-central Asia, sexual activity is still confined essentially to marriage, but marriage often occurs shortly after puberty, and childbearing starts soon thereafter. Contraception and abortion do not play a significant role in either of these subregions.

At the other extreme, low rates of teenage fertility may be due to little exposure to risk, extensive use of contraception, widespread abortion, or any combination of these conditions. Late marriage and social pressure to postpone sexual activity until marriage, backed up by easy access to abortion, accounts in large part for the very low rate in Eastern Asia. Although very few women marry below age 20 in most countries of Northern and Western Europe, a substantial majority become sexually active during their teenage years. Effective, modern means of contraception are widely used; abortion is legal, safe and accessible but not often necessary. The elevated birth rate in the USA appears to be due not to greater exposure to risk but rather to less successful pregnancy prevention; there is a high incidence of abortion as well. The situation has been somewhat different in Eastern Europe, where early marriage has been common and abortion the only widely available means of fertility control.

2. *Consequences*

In recent years, increasing attention has been given to the negative consequences of teenage childbearing. There is concern both about health effects and about social and economic effects. Such effects may relate to the individual mother and child, to society at large or, very often, to both. Their impact typically varies with the age and marital status of the mother. Younger adolescents tend to be much more seriously affected than older adolescents. Although issues associated with single parenthood have been most prominent, very young married women may encounter somewhat different but equally grave problems.

2.1 *Health Effects*

Research has shown that teenage mothers and their infants face significantly greater health risks than do older women. Some, particularly those resulting from physical immaturity, are a direct function of age. Declining age at menarche enhances the opportunity for conception among women who are still in their early teens. Social factors also play a prominent role, however. The impact of age itself is often exacerbated by poverty, lack of autonomy and, in some situations, specific cultural customs. Moreover, a high proportion of births to adolescents are first births, which are more likely than subsequent births to incur medical complications.

Pregnant teenagers frequently suffer from hypertension, iron-deficiency anemia and premature labor. Childbirth presents the risk of cephalopelvic disproportion (when the pelvis opening is not wide enough to permit passage of the infant's head, resulting in prolonged labor and/or obstructed delivery), hemorrhage and infections. Female genital mutilation, a practice that is common in parts of Africa, increases the likelihood of obstructed delivery and other complications of childbirth in addition to its negative effects on women's general health and sexual functioning. In addition, the infants of very young mothers exhibit higher rates of morbidity and mortality than those of older women, often associated with low birth weight.

Sexual activity in itself carries potential threats to the health of adolescent women. Sexually active teenagers are at high risk of contracting sexually transmitted diseases (STDs) including HIV/AIDS. In some parts of sub-Saharan Africa, STD infection rates are higher among young women than among either young men or older women. Unsafe abortion is a major factor contributing to high maternal mortality in many countries, and teenagers are over-represented among women treated for complications following such procedures.

2.2 *Social and Economic Effects*

Research in both developed and developing countries indicates that, for most teenagers, childbearing also has lasting unfavorable social and economic consequences. The relationships are nevertheless complex and the causal sequences difficult to disentangle. In addition, the effects are highly variable, and a substantial proportion of women may experience no discernible disadvantage.

Becoming a mother necessarily involves the assumption of certain adult responsibilities. In the traditional societies that still exist in some parts of the world, this is the role expected of females. As long as they are in socially recognized unions, new mothers and their infants take their place in established systems of support that include their partners, other family members, and the community at large. For vulnerable young women, motherhood may offer the only clear path to security and social acceptance. But by the same token, their future options are often limited, while, at the community level, their potential contribution to the growth of human capital and economic output is foregone.

The forces of modernization have nevertheless penetrated even remote corners of the globe. More young women are staying in school beyond the primary years. On the one hand, education delays marriage and childbearing and, on the other, it provides skills, promotes self-reliance and stimulates aspirations for new roles. Unplanned births are likely to arrest this process, however, especially if, as commonly happens, they occur outside of recognized unions and the men involved fail to acknowledge paternity. The mother cannot always count on the support of her family and may be forced to raise the child on her own. In some countries, opportunities for young women to participate in the paid labor force have also opened up with effects that are in some respects similar to those of extended schooling. Moreover, education is an investment that ultimately benefits society at large and employment contributes directly to the economy.

In a few industrialized countries, particularly the USA, the social and economic impact of teenage childbearing has been the subject of considerable investigation. In the USA, the majority of teen pregnancies are unplanned and, relative to other countries, fertility rates are particularly high among women less than 18 years of age. When the issue came to the fore in the 1970s, it soon became clear that early motherhood is linked closely with low education, limited opportunities for employment and poverty later in life. A high proportion of the births occur outside of marriage, and the fathers often provide little or no support. Associations with marital instability, large family size and problems in the children's behavior and cognitive development have also been documented. Although it was initially assumed that having a child at a young age is in itself largely responsible for such detrimental outcomes, it is now widely acknowledged that the mothers tend to be selected from a subgroup of women who were already at a disadvantage beforehand, and that causation typically runs in both directions. Even when prior factors are taken into account insofar as possible, however, a negative impact remains. Ultimately, poor, young single mothers must often depend on public support, placing a burden on community resources and creating a dependence that tends to perpetuate itself from one generation to the next. In most other Western countries, non-marital childbearing has also become more common among women of all ages, but the mothers are nevertheless likely to be in stable sexual partnerships, and this lifestyle has become widely accepted.

Whether in modern or traditional societies and whether or not the birth occurs in an established union, a considerable age difference between the young mother and the father of her child is common. In traditional cultures, men must often wait until they have acquired a certain seniority before they marry. In more modern societies, young women may seek the protection of or be exploited by older men. Such relationships are inherently unequal, hindering free communication, preventing the development of mutually supportive partnerships and depriving women of their essential autonomy.

3. *Policy Responses*

3.1 *International*

The international community has only relatively recently singled out adolescent childbearing as an independent issue and begun to deal explicitly with it. The Programme of Action adopted at the 1994 International Conference on Population and Development in Cairo urges governments for the first time to 'protect and promote the rights of adolescents to reproductive health education, information and care' and to 'greatly reduce the number of adolescent pregnancies.' It recognizes the rights of young women to privacy and confidentiality and their special needs with regard to sexual behavior and reproduction. The recommendations of the preceding conference held in Mexico City in 1984 concerning adolescents were limited to calls for family life and sex education 'within the changing socio-cultural framework of each country' and for an effort to 'reduce maternal mortality and morbidity' due to pregnancy and childbearing at young ages. The World Plan of Action adopted in Bucharest in 1974 did not mention adolescents specifically at all.

3.2 *Developing Countries*

In some countries, especially in Eastern Asia, age at marriage has risen considerably without undermining the traditional function of marriage as the only acceptable context for sexual activity and childbearing, and there is little call for governmental action. In many other parts of the developing world, however, the effect of rising age at marriage has been offset largely by increases in births among unmarried women. Similarly, in the Caribbean and sub-Saharan Africa, where union formation involves complex processes, births outside of socially sanctioned relationships also occur more frequently than formerly. Single parenthood must compete for governmental attention with a host of acute social and economic needs, however, and there is often great reluctance to enter what is seen as a controversial area. Efforts to address the problem have generally been limited to family-life education programs.

Childbearing within recognized unions among women less than 15 years old, which is still common in sub-Saharan Africa and to a lesser extent in South-central Asia, has aroused comparatively little interest. Yet, whether within or outside of marriage, the

medical complications of childbearing soon after puberty constitute a public health issue. Many of the risks could be greatly reduced if not eliminated through modern medicine, including better diet, prenatal care, the attendance of trained personnel at the time of delivery and health services for newborns, but the resources to support such interventions are lacking.

A variety of obstacles prevent sexually active adolescent women, especially those who are unmarried, from receiving the support they need to avoid STDs and childbirth. School curriculae rarely include sex education. Existing contraceptive services are frequently restricted to married women. In many countries, abortion is illegal and consequently dangerous. There is much work to be done to provide adolescents with the skills and services necessary for healthy behavior.

3.3 Developed Countries

The great majority of developed countries have gone a considerable way toward meeting the needs of sexually active adolescents. Almost all of these countries have national health systems that provide basic health care, including reproductive health services, to the entire population at low cost to the consumer. Abortion is typically legal and safe. The extension of publicly sponsored contraceptive and abortion services to unmarried minors has often been taken up as a separate and more controversial issue than their provision to adults, but there is widespread agreement that the consequences of unprotected sex are to be avoided.

Scandinavia and a few other countries of northwestern Europe have moved significantly farther. Public policy there supports a strong social ethic of sexual responsibility, and there are few unintended pregnancies. Adolescent sexuality is accepted as a normal part of the maturation process. Sexual matters are discussed openly, and related information is readily available. In a number of cases, separate arrangements have been made to meet teenagers' special need for sensitive, confidential, readily accessible services.

The situation is quite different in the USA, though the issue receives considerable public attention. The main thrust of publicly funded teenage pregnancy prevention programs is toward delaying the initiation of sexual activity, although the success of such efforts remains open to question. Sexuality education in schools is typically recommended although not required, and the content may be limited. Family planning clinics serve adolescents as well as older, married women, but their location is often inconvenient for young people and most require some payment. Contraceptive and especially abortion services for young people are highly contentious topics and subject to a constant barrage of legislative proposals and judicial decisions that threaten their effectiveness.

See also: Adolescent Health and Health Behaviors; Family Theory: Economics of Childbearing; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility of Single and Cohabiting Women; Fertility: Proximate Determinants; Gender and Reproductive Health; Reproductive Rights in Affluent Nations; Teen Sexuality

Bibliography

Alan Guttmacher Institute 1994 *Sex and America's Teenagers*. Alan Guttmacher Institute, New York

Alan Guttmacher Institute 1998 *Into a New World: Young Women's Sexual and Reproductive Lives*. Alan Guttmacher Institute, New York

Bledsoe C H, Cohen B (eds.) 1993 *Social Dynamics of Adolescent Fertility in Sub-Saharan Africa*. National Academy Press, Washington, DC

Bongaarts J, Cohen B (eds.) 1998 Adolescent reproductive behavior in the developing world: Introduction and overview. *Studies in Family Planning* **29:** 99–256

DaVanzo J, Farnsworth G (eds.) 1996 *Russia's Demographic 'Crisis.'* RAND, Santa Monica, CA

Department of International Economic and Social Affairs of the United Nations 1988 *Adolescent Reproductive Behavior: Evidence from Developed Countries. Population Studies No. 109*. Vol. I. United Nations, New York

Department of International Economic and Social Affairs of the United Nations 1989 *Adolescent Reproductive Behaviour: Evidence from Developing Countries. Population Studies No. 109/Add.1*. Vol. II. United Nations, New York

Furstenberg F F Jr, Brooks-Gunn J, Morgan S P 1987 Adolescent mothers and their children in later life. *Family Planning Perspectives* **19**: 142–51

Hayes C D (ed.) 1987 *Risking the Future: Adolescent Sexuality, Pregnancy, and Childbearing*. Vol. 1. National Academy Press, Washington, DC

Hoffman S D 1998 Teenage childbearing is not so bad after all…or is it? A review of the literature. *Family Planning Perspectives* **30**: 236–39, 243

Jones E F, Forrest J D, Goldman N, Henshaw S, Lincoln R, Rosoff J I, Westoff C F, Wulf D 1986 *Teenage Pregnancy in Industrialized Countries*. Yale University Press, New Haven, CT

McCauley A P, Salter P 1995 *Meeting the Needs of Young Adults. Population Reports, Series J. No. 41*. Johns Hopkins School of Public Health, Population Information Program, Baltimore, MD

Mensch B S, Bruce J, Greene M E 1998 *The Uncharted Passage: Girls' Adolescence in the Developing World*. Population Council, New York

Morgan S P 1996 Characteristics of modern American fertility. In: Casterline J B, Lee R D, Foote K A (eds.) *Fertility in the United States. Population and Development Review*. Vol. 22 (suppl.) Population Council, New York

Population Division of the Department of Economic and Social Affairs of the United Nations 1999 Fertility trends among low fertility countries. In: *Below Replacement Fertility*. Population

Bulletin of the UN Special Issue Nos 40/41. United Nations, New York
Population Division of the Department for Economic and Social Analysis, United Nations 1994 *The Health Rationale for Family Planning: Timing of Births and Child Survival*. United Nations, New York
Population Division of the Department for Economic and Social Information and Policy Analysis of the United Nations 1995 *Population Consensus at Cairo, Mexico City and Bucharest: An Analytical Comparison*. United Nations, New York
Singh S, Wulf D 1990 *Today's Adolescents, Tomorrow's Parents: A Portrait of the Americas*. Alan Guttmacher Institute, New York

E. F. Jones

Telecommunications and Information Policy

1. Network Regulation

Telecommunications and information policy involves balancing public and private interest through local, national, regional, and international regulation of a growing variety of communications technologies. These include: (a) network facilities and equipment; (b) basic telephony; (c) value-added and information services including data, video, and voice transmission (i.e., the Internet); (d) mobile and radio services; (e) integrated digital networks and integrated broadband networks. Regulation is a dynamic political process that distributes costs and benefits throughout these various sectors. Telecommunications regulation has always had a domestic and international or multilateral component. Traditionally, this was composed of a national government body with the most power in setting the rules of operation, and the International Telecommunications Union (ITU), which served to establish common standards and protocols. Today, telecommunications and information policy is in a period of dramatic transition as a result of what some have called the rise of a global 'network society' (Castells 1996). This transition can be explained by both rapid changes in technology since the 1970s, particularly 'digitalization' which has led to the convergence of new and old media, as well as dramatic political changes, most commonly understood as 'globalization.' This notion of network is based on the interactive switched telecommunications network that was originally developed for basic telephony. In contrast to traditional top-down, one-way mass media, new electronic media function through a 'hub-and-spoke' model.

In terms of regulation, these networks are unique because they are based fundamentally on shared resources. In the case of telephone service, it would make little economic sense for each individual subscriber to connect separate cables to every other subscriber in order to make a telephone call, even though it is technologically feasible. Individual subscribers do not make enough telephone calls to warrant investment in hundreds of cables. Instead, phone calls made by individual subscribers are routed through a local exchange, where using a common connection the calls are connected to a bigger regional exchange that uses high-capacity connections that link major exchanges in order to distribute calls. The value of this network grows as each additional user joins the network, precisely because it spreads the fixed costs around a larger number of users and because it expands the numbers of people each already existing subscriber can contact. Economists argue that because the network can enhance social benefits beyond the members of the network, telecommunications should be seen as a 'public good' because of 'positive externalities' (Garnham 2000). Putting this into practice, public policy experts contend that the telecommunications network should therefore be seen as a 'club' based on members with mutual interests, as opposed to a market. These 'members' generally include different government bodies, domestic and transnational firms, labor unions, consumer organizations, and public interest groups. The problem in regulating both broadcasting and telecommunications networks is what happens when this 'coalition of mutual interest breaks down' (Noam 1987). The conventional wisdom was that telecommunications networks functioned most efficiently as a 'natural monopoly' because of the enormous fixed costs required to build and upgrade the network. The rationale for monopoly in telephone manufacturing and services was based on the understanding that centralization of operations would 'increase reliability' and that monopolies could best tap economies of scale and scope to better 'advance considerations of equity' with the objective of universal service provision (Aronson and Cowhey 1993). Given this economic logic, regulators and policy makers have been historically concerned with how to best achieve universal access at the most reasonable prices. Between the 1920s and the mid-1970s, there was little technological change that directly affected the telecommunications sector. In this period, telecommunications networks were regulated at the national level through a system of cross-subsidy, whereby urban areas subsidized rural areas, long-distance rates subsidized local rates, large (corporate) users subsidized residential users, and telecommunications revenues subsidized the postal system. With the exception of the American system of state-regulated private monopoly of AT&T, most other nations relied on government ownership and/or operation of their Post and Telecommunications Operator (PTO).

In the 1980s, advances in technology coupled with political changes dramatically challenged the logic and scope of national regulation of the telecommunica-

tions network. Technological advances stemming from research in the defense-related electronics sector introduced new satellite, cellular radio, fiber optic, and digital exchange technology that would suddenly transform telecommunications services into the new 'nervous system' of the global economy (Mansell 1995). The telecommunications network could now transmit all types of information through these new technologies, and thus became an increasingly vital component of all sectors of economic activity. Expansion and modernization of the telecommunications network was seen as crucial in facilitating transnational production and distribution of goods and services, including the proliferation of financial markets and new media technologies. Experts in national and multilateral agencies like the International Telecommunications Union (ITU) therefore began to expand their scope by considering questions of regulation for telecommunications *and* information policy. The convergence of different kinds of communications technologies has introduced a new set of problems for regulators both domestically and globally. Whereas regulation of the telecommunications or broadcast network has been seen as legitimate in most societies, regulation of content or the messages passing through the network is often seen as illegitimate, raising concerns about privacy, freedom of expression, and censorship. Equally important as the technological changes, political changes in the era of Reagan and Thatcher led to the 'liberalization' of regulation over national telecommunications policy. Traditionally, national networks were joined together at the borders by international gateways. In the 1980s, however, transnational corporations began to demand 'seamless networks' in order to facilitate the transactions of these firms across national borders (Hills 1998). The deregulation of AT&T in 1984 and the privatization of the state telecommunications network in the UK introduced a distinctly new logic of competition in the global marketplace by policy makers. Most significantly, traditional concerns for establishing 'fair' prices and maximum access to services was replaced by a new turn to neo-classical economics where the concern for the global success of home-based transnational corporations was measured by sales, profits, and a favorable balance of payments (Melody 1994).

With the end of the Cold War and the changing ideological climate, policy makers around the world began to reconsider the legitimacy of the state's role in economic development. Reformers argued that telecommunications policy should be an instrument that encourages market competition within and across national borders, driving down prices for consumers while simultaneously expanding the pace of technological innovation.

In a sense, as telecommunications and information policy became a greater concern for governments and large corporations, expertise in policy was increasingly cultivated in Business and Law Schools, most often in the US (Singh 1999). This created a strategic consensus about the failure of state-operated monopolies, promoting instead global integration and staying clear of nonmarket intervention in setting the terms of network operations. As a result, the rhetoric of the 'information revolution' in the early 1990s relegated policy-making to creating conditions for market reforms of the old public PTO networks.

Today, the dominant approach in terms of the scholarship on telecommunications and information policy is prescriptive, assuming that the market and technology are inherently neutral forces. Robin Mansell contrasts this 'idealist model' of policy analysis to a more critical 'strategic model' that takes into account institutional and economic power. The dominant model espouses an 'evolutionary process of technological innovation in telecommunication and computing technologies … left to the forces of the competitive market' (1993, p. 6). William Melody, one of the most prolific communications scholars writing from this 'strategic' perspective, explains that although the national telecommunications model failed to meet anywhere near universal service objectives in most places, competition cannot be seen as an 'adequate substitute to regulation or public ownership.' Despite the proliferation of new communications technologies, without effective and independent regulation we are replacing a tradition of state-owned monopoly with private monopolies and oligopolies across national borders (Melody 1997).

Although the spread of new communications technology has been astonishing since the 1980s, the expansion has been extremely uneven. Policy makers today are finally talking about the 'digital divide' which includes recognizing the fact that some 75 percent of the world's population has yet to gain access to basic telephone service (ITU 1997). The Internet, hailed by many enthusiasts as the ultimate technological equalizer, is the least accessible communications technology in the world.

It should be clear from the preceding discussion that changes in technology and politics have transformed the telecommunications and information policy process as a result of 'digitalization' and 'globalization.' The second half of this article will examine in greater detail the 'globalization' of telecommunications and information policy, focusing on the growing importance of this issue for nations in Africa, Asia, and Latin America.

2. *Telecommunications and Information Policy Reform in the 'Developing World'*

Telecommunications was not recognized as an economic priority in the so-called developing world until the 1980s. This was reflected in very low rates of

telephone density (between 1 and 10 percent) and very slow absorption of new technologies. In most nations, the state owned and operated the PTO, and the network was regulated based on the cross-subsidy principles discussed above. Although nearly all states advocated universal service (meaning access to a community telephone as opposed to residential telephones), other more pressing economic priorities relegated the telecommunications network to low levels of penetration especially in rural and low-income areas and long waiting lists. As telecommunications and information technologies became an increasingly strategic input for transnational corporations and governments in the 1980s, policy reform became a dominant concern for states throughout the developing world.

In 1985, the International Telecommunications Union (ITU) released its influential Maitland Commission Report entitled *The Missing Link?*, which condemned the extreme inequalities of telephone access between rich and poor nations. Although the report drew global attention to the question of information disparity, its recommendations pressed for the need to reform inefficient national public monopolies and promote the transfer of technologies from advanced to developing nations (ITU 1985). The ITU's report argued that investment in telecommunications should no longer be seen as a luxury service for corporate and national elites, but rather as an essential service that directly leads to economic growth. In fact, teledensity penetration was correlated with tangible growth in gross domestic product. This approach is referred to as the 'telecommunications for development' perspective and posits that investment in the newest telecommunications technologies would allow developing countries to actually 'leapfrog' stages of development. For example, instead of having to replace copper cables with fiber optic technology, developing countries would bypass that stage in development altogether and adopt the most advanced technologies. Similar arguments are made today about wireless and satellite technologies.

In this period, the ITU along with the World Bank began to urgently promote the liberalization of infrastructure and the privatization and commercialization of services through a series of multilateral and bilateral conferences and seminars on telecommunications reform. Both multilateral institutions began promoting the expanded role of the private sector in telecommunications development, with transnational equipment manufacturers like Alcatel, NEC, Ericsson, British Telecom, US West, and others eager to enter new markets offering to provide direct foreign investment on attractive financial terms (Lee 1996). While the first phase of reforms in developing countries consisted of the liberalization of the equipment market, the World Bank in particular began to push for the liberalization of national telecommunications networks.

The argument was that the poor performance of state-operated telecommunications networks throughout Africa, Asia, and Latin America significantly impeded economic growth. The solution was to implement a comprehensive reform process that would enable competition and technological modernization and balance the concerns of equity between those of efficiency. This would include the deregulation of the state-operated network with the ultimate goal of privatization, liberalization of the supply of services, and the separation of the government's policy and regulatory arm from its responsibilities as a network operator (Saunders et al. 1994). In the midst of the African and Latin American Debt crises of the 1980s, the World Bank's policy perspective was particularly influential wherever its sister lending agency, the International Monetary Fund (IMF), financed new loans (Hills 1998).

Between 1980 and 1997, some 40 states engaged in the partial or total privatization of their national telecommunications operators raising $140 billion worldwide. Unlike their Western counterparts, states in the developing world have privatized national telecommunications monopolies primarily as a means to reduce debt burdens and invite foreign capital and expertise. In many nations, the transfer of public resources into private, especially foreign hands, has raised political opposition and public protest (Noam 1998). However, this kind of politicization of what is seen as a technical issue has been brushed aside as imminent concerns about the dangers of missing out on the 'information age' subsumes policy debates.

While 'telecommunications for development' quickly became the dominant perspective among policy makers, communications scholars pointed out that the value of access to telecommunications was contingent on issues of power, between national and global interests, urban and rural interests, and intra-community class interests (Samarajiva and Shields 1990).

These critics concede that the public service monopoly model failed to achieve the objectives of both efficiency and broad universal access, and in most cases they agree that the introduction of fair competition is a pressing concern for most developing states (Melody 1997). States did not adequately fund the public telecommunications operators and these bureaucracies were largely unaccountable to the public they were supposed to serve. Thus, bureaucratic reform is high on the agenda for these critics (Mody 1993). However, unlike the proponents of 'telecommunications for development,' these critics point out that the only way to avoid a tradeoff between efficiency and equity is to create effective political institutional arrangements that lead to sustainable development (Melody 1997, Samarajiva and Shields 1990). Demand for telecommunications policy reform in the developing world came from the biggest users of services, which were mostly business users based

nationally as well as internationally. Critical communications scholars have argued that the pace and institutional structure of telecommunications and information policy reforms tends to privilege the interests of these corporate actors (Dunn1998). To put it most simply, the new market principles shape policy along 'the more you spend the less you pay' principle, generally benefiting the biggest spenders, in most cases corporations. For countries with very low rates of telecommunications penetration—ranging from 1 to 10 percent—who picks up the tab for the low-income majority leads to serious concerns about widening the digital divide. This is a valid concern for most nations in the developing world where 80 percent of the revenue is generated from 20 percent of the customer base and demand for most advanced services that require larger network capacity is confined to transnational corporate users.

The paradox of telecommunications liberalization around the world is that the moral superiority that competition is supposed to ensure is undermined by the alarming rate of consolidation of global firms. The justification for domestic markets opening their doors to private firms is that competition brings down prices for consumers and offers more and better services for less. However, today a handful of global giants dominate the telecommunications industry worldwide and have the resources to buy out and raise barriers to entry to smaller competitors when they choose. Moreover, the merger of traditional telecommunications companies with producers of content (software) further reinforces concentration as opposed to competition in overall services. Finally, the economic centrality of telecommunications services for business in today's global economy puts pressures on national governments to separate the needs of corporate users of high-speed networks and services from the public networks, thereby creating new levels of information disparity (Sussman 1997).

In the developing world, the pressure to privatize and commercialize networks often precedes the creation of effective regulation. Consequently, these analysts point to the absolute necessity of effective independent regulatory regimes to 'protect infant industry, control private monopoly power, offset the tendency to concentration, protect employment, and achieve macro-balance' (Mody 1993). However, these domestic concerns often conflict with global pressures as trade in telecommunications and information services becomes increasingly important in a global network society. Domestic policy is increasingly coordinated by regulatory agreements like the 'Global Telecom Pact' and the 'International Technology Agreement' through the World Trade Organization (WTO). These new global rules clearly prioritize the liberalization of markets and protection of intellectual property as opposed to ensuring effective regulation of competition and establishing accountable regulatory procedures.

See also: Broadcasting: General; Broadcasting: Regulation; Information Society; Mass Media, Political Economy of; Privatization of the Economy; Radio as Medium; Regulatory Agencies; Television: History

Bibliography

Aronson J, Cowhey P 1993 *Managing the World Economy: The Consequences of Corporate Alliances*. Council of Foreign Affairs Press, New York

Castells M 1997 *The Power of Identity: Economy, Society and Culture*. Blackwell, Malden, MA

Dunn H 1998 Policy issues in communications technology use: Challenges and options. In: Dunn H (ed.) *Globalization, Communications and Caribbean Identity*. St. Martin's Press, New York, pp. 18–39

Garnham N 2000 *Emancipation, the Media and Modernity: Arguments about the Media and Social Theory*. Oxford University Press, New York

Hills J 1994 Dependency theory and its relevance today: International institutions in telecommunications and structural power. *Review of International Studies* **20**(2): 169–86

Hills J 1998 U.S. rules. OK?: Telecommunications since the 1940s. In: McChesney R, Meiksirs Woods E, Bellamy Foster J (eds.) *Capitalism and the Information Age: The Political Economy of the Global Communication Revolution*. Monthly Review Press, New York, pp. 99–121

International Telecommunications Union 1985 *The Missing Link: Report of the Independent Commission for Worldwide Telecommunications Development*. ITU, Geneva

International Telecommunications Union 1997 *World Telecommunications Development Report*. ITU, Geneva

Lee K 1996 *Global Telecommunications Regulation: A Political Economy Perspective*. Cassell, London

Mansell R 1995 *The New Telecommunications. A Political Economy of Network Evolution*. Sage, New York

Melody W 1991 The information society: Transnational economic context and its implications. In: Sussman G, Lent J (eds.) *Transnational Communications—Wiring the Third World*. Sage, London, pp. 27–41

Melody W 1994 The information society: Implications for economic institutions and market theory. In: Comer E (ed.) *The Global Political Economy of Communication*. St. Martin's Press, New York, pp. 21–36

Melody W (ed.) 1997 *Telecom Reform: Principles, Policies and Regulatory Practices*. Technical University of Denmark, Denmark

Mody B 1993 Telecommunications privatization in the periphery: Adjusting the private–public balance. *International Review of Comparative Public Policy* **5**: 257–74

Noam E 1987 The public telecommunications network: A concept in transition. *Journal of Communication* **37**(1): 28–48

Samarajiva R, Shields P 1990 Integration, telecommunication, and development: Power in paradigms. *Journal of Communication* **40**(3): 84–103

Saunders R, Warford J, Wellenius B 1994 *Telecommunications and Economic Development*. Johns Hopkins University Press, Baltimore, MD

Singh J P 1999 *Leapfrogging Development: The Political Economy of Telecommunications Restructuring*. State University of New York Press, Albany, NY

Sussman G 1997 *Communication, Technology, and Politics in the Information Age*, UNDP Development Report, 1999. Sage, London

P. Chakravartty

Telecommunications: Policy

Telecommunications is a collection of electronic technologies for the movement of data and information. It includes telegraph, telephone, radio, television, satellites, and small wireless devices. Technical standards, extent of deployment, and pricing are all subject to government regulations. Regulatory issues represent major public policy issues for governments, such as how to support or control the use of the Internet. Most research on this topic has been about economic and political issues, normally done within the context of individual nations.

1. *Historic Role of Policy Makers*

Over the past 150 years governments throughout the industrialized world, and today in almost all countries, face two sets of issues concerning telecommunications: what to regulate and why, and how to leverage this technology for national advantages. As new forms of telecommunications appeared, governments have had to make judgments about them. In effect, they have had more forms of technology to regulate, create policies around, and exploit. New technologies will emerge that will create the need for additional policy considerations. While technologies have changed over the years (e.g., adding cell phone where wired telephone systems exist, use of pagers), the issues facing policy makers have remained fairly constant over the past century and a half.

There is a specific set of existing technologies which routinely require the attention of policy makers concerned with communications. They are all electronic, often use computer technology, and are widely deployed in many societies. These include: telegraph, telephone, wireless telephones, radio, television broadcasting, cable television, satellites, wireless information appliances (such as a pager), and most recently, the Internet. This article will not cover paper-based forms of communications, such as newspapers and magazines, because those topics have different issues from those of electronically based communications. The majority of the deployment of electronically based technologies continues to be in the industrialized world, although usage is expanding all over the planet. Some technologies are more ubiquitous than others. For example, radio is virtually everywhere, even where there is no electricity, thanks to batteries and hand cranks, and their relatively low cost. On the other hand, at the beginning of the twenty-first century, the Internet is most widely deployed in North America and in Western Europe. Much of what we know about telecommunications policy issues is drawn from the experience of Europe and North America, seats of the earliest industrial economies.

Historic patterns suggest future ones. Over the past two centuries new technologies emerged, such as the telegraph in the 1840s, the telephone in the 1870s, and satellites in the 1950s. At first, little policy attention was paid to a new technology because its impact was minor. As it deployed within a nation, policy makers then faced several questions. To what extent should the national government make a specific technology available to the majority of its citizens? Should the government support further research and development to improve a device's performance, lower its costs, or to get it further deployed? If the technology is related to an existing one already regulated, what should the regulations cover? Pricing? Access? Monopolistic existence? In the special case of broadcasting (radio, TV, cell phones), governments have to allocate bandwidth for the transmission of signals. That effort requires both technical knowledge on the part of regulators, and also a strategy for allocating frequencies among contending technologies. The process of allocation has increasingly become more complex, particularly in the twentieth century, as additional technologies emerged, especially regarding wireless.

Policy makers have also had to deal with issues of national economic development or defense, thinking through how telecommunications can be leveraged in the interests of the nation. The British government, for example, in the nineteenth century saw underwater telegraphic cable communications as vital to national defense and, therefore, supported the establishment of a global network. The United States did the same thing with satellites in the mid-twentieth century. OECD (Organization for Economic Cooperation and Development) nations in Europe collaborated to create an international standard for cell phone communications as an economic stimulant to the new technology in the last quarter of the twentieth century. At the same time, the United States government made its telecommunications network, originally established for defense purposes, accessible to the public; we know it as the Internet. The reason: primarily to stimulate use of a new technology for economic development. In the last decade of the twentieth century, the American government protected the Internet from practices that would have slowed its deployment, such as the imposition of local sales taxes by state and local governments on transactions conducted over the Internet.

2. *Common Issues Shared by Policy Makers*

Policy makers traditionally face five key issues with each major form of electronic communications. In

addition, new issues are emerging. Some issues are related to national policy considerations, such as defense and economic development, others center on activities of government regulators responsible for implementing legal and policy directives. Each of these five vary in importance, depending on issues of the day (war needs, economic considerations, etc.), but each tends to be constantly evident.

2.1 Economic Development

First, there is the issue of how to use a particular technology for economic development. Can investments in the further development of a technology, such as the use of telecommunications with computers, lead to advances that generate new jobs, sales of additional products internally, and for export? Tax incentives and outright financial subsidies in support of local technology firms and for training of new workers are often the tools of choice of policy makers when they support a new technology for economic reasons.

2.2 Infrastructure Development

Second, there is the issue of how an infrastructure can be built, or should be built, to make the technology either usable or accessible. Telephones are useless unless the majority of a nation has access to the same telephone network. In the 1800s telegraph only became useful when it could reach multiple cities. Postal systems and national highways systems raised the same issues for all nations. A communications technology becomes useful when an infrastructure exists to make it available to the entire nation. Tools of choice are often the same as those used to encourage diffusion of a technology to stimulate economic activities.

2.3 Competition

A third issue concerns the extent to which policy makers wish to encourage competition among service providers or different technologies. In the United States, for example, there is a strong tradition of encouraging competition. That is why there are numerous telephone companies, heavily regulated to ensure competition, but essentially operating in the private sector. European governments have chosen a different path in most cases, namely, to run networks out of national government agencies. In part this was to support a technology that initially would have failed in the marketplace, such as telephone and telegraph services in France in the 1800s. Some governments will thus choose to have a national monopoly, operating in an environment in which competition is not permitted. Either approach still raises the question, however, of what to do when an alternative technology emerges, such as wireless telephones as an alternative to wired phone systems. Are both regulated the same way? Should one be encouraged for other reasons (e.g., national defense or economic development) at the expense of the other?

The economics of implementation prove influential. For example in Africa, wireless telephone systems are seen as less expensive than traditional wired systems because there are lower infrastructure costs involved in deploying such networks. So, African governments frequently support the use of cell phone networks over more traditional telephone systems.

2.4 Access

A fourth issue concerns access to a communications technology by citizens of a nation. To what extent should access to communications technologies be facilitated by the state, and why? These are questions that have long confronted policy makers. In the nineteenth century French policy makers decided to run a national telegraphic network to bring high-speed communications to many parts of France. In the early 1900s the American government implemented policies to encourage use of telephones and, by the end of the twentieth century, was investing tax revenues in providing Internet access to all public schools. In each instance, the nation's interests were considered: Does it help development of the economy? Do we need it for national defense? Is access a fundamental extension of some national basic belief about the open flow of information? Do citizens demand government stimulated access? Is government involvement essential because of the technical infrastructure or complexity involved, as in the case of building satellite networks to transmit information inexpensively even though the initial network is very expensive?

2.5 Affordability and Pricing

A fifth issue concerns affordability and pricing, with primary emphasis on pricing policies. Natural or national monopolies are normally subject to government regulation. One of the central policy issues concerns what prices service providers should charge citizens. If the service provider is a government agency (e.g., a national telephone network), the state can charge whatever it wants, since the concern is not necessarily to turn a profit. The issue in that situation is to what extent does the government want to make access to communications affordable. If communications is provided to a nation through a network of private companies, as increasingly is the case around the world, then the requirement that these companies be allowed to make a profit on their services becomes essential. The issue, then, is how much profit should a firm be allowed to make. Policy makers struggle with understanding the true costs and profits of a firm or

communications industry. They also try to understand the needs of citizens and their ability to pay for such services.

As governments have increasingly evolved into democracies or other forms of representative republics over the past two centuries, the role of information in those societies has become an ever more important issue for policy makers. A fundamental premise of these forms of government is that information about political activities, and hence personal and economic issues, should flow relatively uncontrolled. In more autocratic states, the reverse holds. China is still trying to figure out how to block its citizens from using fax machines and the Internet. Yet, in Central Europe more access to information handling tools and communications is on the rise. In North America, all forms of communications are constantly made available to Americans and Canadians. Communications can also extend to press censorship of books, magazines, and newspapers, or their freedom, as a collateral set of policy considerations implemented at the same time that regulators deal with electronic information handling, such as telephones, television, and radio. The nature of government is, therefore, an integral part of the form of policies implemented toward communications.

3. *Emerging Issues in Telecommunications*

In the last quarter of the twentieth century new issues began to emerge that required the attention of policy makers. Several, in particular, are of growing importance to governments across the world.

3.1 *Governments and Global Economics*

As communications companies became global enterprises, implementing worldwide business strategies and policies, policy makers began asking what role should an individual country play in the affairs of such a firm? For example, in the late 1900s, telephone executives in Spain quietly moved into the Latin American economies, buying up local phone companies, thereby increasing control over telephonic practices from offices in Madrid. American communications companies in the same period implemented global satellite networks used to transmit long-distance telephone calls. In the late 1800s, British cable companies overwhelmingly owned and controlled the underwater cable business. This is not a standards discussion, this is about how a government regulates or influences a multinational company. To what extent can it? National egos, public policies toward access to communications, and issues related to national defense all weigh in as influences. Government agencies maintain that they have a critical right and need to regulate activities within their borders. Many firms argue that economies of scale and scope require that they not be constrained by some local policy or practice. It remains a never-ending debate, but, nonetheless, an increasingly important issue. More frequently, groups of nations establish uniform policies and practices, as is evident in the Common Market nations of Europe. Yet, overwhelmingly, nations still set policies and regulate independently of each other.

3.2 *Control of International Flow of Information*

The international flow of information through a variety of electronically based communications technologies profoundly increased during the second half of the twentieth century. Simultaneously, it has become increasingly difficult for governments to control the inflow and outflow of information. For example, Western European governments began regulating what data could come in and out of a nation via telephone lines linked to computers in the 1960s. That worked for a while but eventually technology breached the walls and today information moves around the globe as if there were no national borders. The Internet is the most recent example of that process at work. While some governments are attempting to control the type of information transmitted over the Internet (e.g., political data in China, pornographic materiel in the United States and in Western Europe), the fact remains that anyone with access to the Internet can send and receive just about any kind of information. Is that a problem for policy makers? It can be if there is a compelling need to control some aspect of the flow of information. But what has happened is that the technology with which to move information around the globe easily, quickly, and inexpensively on a massive basis is now reality. What governments should do about that, and their role in the process, remains unclear.

4. *Special Role of the Internet*

Yet, another emerging reality concerns the specific role of the Internet and computing. Computers and telecommunications are two sets of rapidly converging technologies. They are also becoming ubiquitous, increasingly available to people. At the start of the twenty-first century, nearly half the population of the world had yet to make its first phone call. Half of the others had regular access to Internet-based information appliances, ranging from computers to e-mail machines, to pagers and cell phones, and all relying on the communications backbone of the Internet. Governments now face basic policy questions about the Internet that would have been somewhat familiar to policy makers in the early 1900s puzzling over what to do about radios, telephones, and later television. Local cultural values, issues of national defense, economic development, and historical and legal realities play roles in how policy makers consider the Internet.

What is most different from earlier considerations, however, is the speed with which the Internet became a policy issue. Widespread use of the Internet can be dated to the early 1990s, although access to it in the United States went back to the 1970s, first for the military and their suppliers, and later for the general public. But for all practical purposes, the world began to use the Internet in the mid-1990s and, by the start of the new millenium, approximately a billion users were online across the globe. As of this writing, conservative forecasters called for a doubling of the number of users by about the year 2005. Policy makers tend to react to emerging technologies at a much slower pace than the spread of the Internet. So, uncertainty and tensions exist. In Western Europe, for example, the OECD has run annual conferences and conducted studies and surveys in order to inform policy makers in member nations about what each other is doing concerning the Internet and about its technology. National policies toward the Internet are being created, while public officials are simultaneously coming to realize that this technology is fundamentally changing many economic and societal structures and practices in ways not seen since the arrival of television a half century earlier, or the automobile nearly three-quarters of a century ago.

5. *Common Patterns in Communications Policy*

Are there some fundamental conclusions, therefore, that one can draw about the basics of communications policies? Are the profound changes in communications technologies which have occurred in the past half-century subject to some conclusive approaches? Historical precedence is a guide in these matters because governments have had to deal with various forms of telecommunications for nearly two centuries, relying on prior cases and practices. Often the issues involved are less about the nature of the technology and more about its use and costs. The one major group of exceptions are the myriad of regulations required in response to the physics and scientific realities of a technology, such as how to distribute radio frequencies across different types of transmission, e.g., AM and FM radio broadcasting. Occasionally, national 'champions,' in the form of some locally preferred (or developed) technology, influences these technical decisions, such as the debate which occurred in the last two decades of the twentieth century over analogue vs. digital television transmission standards. That issue is also compounded by the continued evolution of technology which today, for example, makes possible DVD transmissions and also use of the Internet and PCs with which to view programming.

Leaving aside these technical regulatory requirements, which can also be highly politicized, there are some common patterns of behavior by policy makers that can serve to inform contemporary interest in the subject.

5.1 *Policy Makers and Emergence of New Technologies*

First, public policy makers have normally and regularly felt compelled to rule on the development, deployment, and use of every new communications technology that has appeared in the past two centuries. While a new technology might avoid drawing the attention of policy makers for a period of time (e.g., as was the case with the telegraph in the United States for over a decade), eventually the regulators catch up. Related to the inevitable role of policy makers has been an almost universal discussion among public officials about how best to upgrade communications infrastructures within their nations. By the end of World War I, every industrialized nation had wrestled with this issue and by the end of the twentieth century, almost every government in the world had dealt with the topic.

5.2 *Policy Makers and Economic Development*

Second, regulatory practices and public policy normally are linked tightly to economic development, national defense, and the realities of existing political infrastructures. Hitler used television to convey his messages; Roosevelt wanted Americans to listen to him on the radio, while the Belgians worried about American television programming corrupting their local culture in the 1980s. As we move increasingly into a new phase in the economies of the world, one characterized by the extensive use of information and its underlying technologies, the debate increases about what role policy makers should play in fostering use of communications within their national borders. While the debate has a long history, what is new is the intensity and critical nature of the process today because of the increased availability of new communications technologies, which are so rapidly spreading through most nations.

5.3 *Types of Policy Makers and Roles*

Third, policies have been created and influenced by three groups within governments: regulators working inside administrative departments, national and state political leaders, and state and national legislative bodies. Increasingly in industrialized societies with representative forms of government, major initiatives to codify regulations, government policies, and the national will have led to far-reaching legislation on communications. For example, the Telecommunications Act of 1996, passed by the US Congress, fundamentally defined how competition could occur in America for all forms of communications. The OECD is at work on creating overarching guidelines of a similar nature in Europe that all European governments can consistently embrace. Put another way, national legislatures periodically redefine and

codify public policy and practices toward communications, usually about every half century with many, other more narrowly focused, legislation passed in the intervening years. The key pattern, however, is that policy and regulatory matters are always politically charged and invariably involve a nation's politicians.

5.4 Policy Making at National Level

Finally, regulatory actions have normally been most effective when implemented at a national level, although local pricing practices within a state or province exist with reasonable success. However, the pattern evident today is toward national policies, even on pricing issues. The use of international conventions is on the rise because, as with the experience in implementing national practices, the most effective cover large blocks of people. It becomes difficult, even awkward, to implement one policy in a country and a different one in a neighboring state; that tends to constrain effective use of a technology in ways citizens want. So, the impulse, or 'best practice,' has been to expand the implementation of consistent policies and practices to ever-larger portions of the world. The trend toward national and even international policies involve:

(a) Spread of universal technical standards (e.g., as the Europeans have done with wireless telephony).

(b) Policies concerning the type of information that can flow in and out of an economy (e.g., with television programming, use of the Internet, and new forms of copyright and patent protection for software and music).

(c) Trade policies that encourage development and acquisition of new technologies (e.g., computer chips, software, and access to satellites).

(d) Increased interconnections with the communications infrastructures of other nations. No sooner had the Americans invented the telegraph in the 1840s, for instance, people around the world wanted telegraphic infrastructures built so that messages could be sent across a nation and to other countries. We are seeing the same with wireless communications and the Internet.

Perhaps the last word on what policy makers face today in the world of communications should come from William E. Kennard, Chairman of the US Federal Communications Commission (FCC). In communicating his agency's strategy for the future in a written document to the US Congress in August, 1999, he argued that his agency's strategy was designed so that 'it would allow the FCC to enter the next century able to respond fully and quickly to emerging technologies and the inexorable movement from regulation to competition.' While he was addressing the US situation, he could just as easily have been commenting about circumstances in many other countries:

> In five years US communications markets will be characterized predominantly by vigorous competition that will reduce the need for direct regulation. The FCC as we know it today will be very different both in structure and mission.

With the Internet now the major new source of policy concerns, and possibly so large as to dwarf other forms of communications, it is relevant to end this article with Kennard's observation:

> The advent of Internet-based and other new technology-driven communications services will erode the traditional regulatory distinctions between different sectors of the communications industry.

While the four trends listed above still hold in that brave new world, his comments suggest that implementation will evolve over time.

6. Research Considerations

There is considerable research done constantly on all aspects of communications and public policy. Research in academic settings is done primarily by economists and political scientists. A great deal of their focus is on the effects of new technologies on a nation's citizens, and results frequently in policy recommendations. Lewis Branscomb (1993) reflects contemporary examples of this kind of work. The second source of considerable research and debate are the regulatory agencies themselves. In the United States, the FCC conducts much research on all aspects of rates for fees and taxes for each form of electronic communications. Their research is routinely published on its web site. In Western Europe, the OECD does similar research but in addition has taken a special interest in information privacy matters and the role of the Internet. Each year, throughout the 1990s, it sponsored conferences on all aspects of this issue; proceedings of each have been published on OECD's web site.

The Internet is the subject of much research all over the world. In the United States this is primarily being done by academics, with leading initiatives underway at the University of Texas, University of California at Berkeley, and at the Massachusetts Institute of Technology. As with other forms of technology, the issues being looked at focus on extent of deployment in the economy, possible government regulatory options, and effects on national policies. Both the FCC and OECD publish many white papers on narrow issues, written in a formal academic style with footnotes and ample bibliography. In addition to the books in the bibliography, each major regulatory agency in the world has an Internet site with public policy announcements, proceedings of conferences, and other studies. The US FCC and the OECD in Europe are the two best sources for material on a wide range of communications issues and technologies.

Bibliography

Branscomb L M (ed.) 1993 *Empowering Technology: Implementing a U.S. Strategy*. MIT Press, Cambridge, MA

Headrick D R 1991 *The Invisible Weapon: Telecommunications and International Politics, 1851–1945*. Oxford University Press, New York

Saunders R J, Warford J J, Wellenius B 1987 *Telecommunications and Economic Development*. Johns Hopkins University Press, Baltimore, MD

Temin P, Galambos L 1987 *The Fall of the Bell System*. Cambridge University Press, Cambridge, UK

J. W. Cortada

Telegraph

The telegraph was the outgrowth and application of advances in electrical science between about 1800 and 1840, particularly the discovery of galvanic electricity by Luigi Galvani and the invention of the electrical battery by Allesandro Volta around 1800, and research into electromagnetism by Hans Christian Oersted, Andre Marie Ampère, Joseph Henry, and Michael Faraday during the 1820s and 1830s. The first two telegraph lines opened in Great Britain in 1837 and the USA in 1844. The telegraph was the only form of electrical communication until the invention of the telephone in 1876, and it remained the mainstay of rapid long-distance communication until the development of practical long-distance telephony after 1900. After 1920, as long-distance telephone rates dropped and many countries introduced airmail service, telegraphy's share of the long-distance communications market entered a long decline. By 1970 the overland telegraph industry in the industrialized world was dead or dying. Submarine telegraphy survived longer; it was the only medium for rapid overseas communications until the advent of transatlantic radiotelephony in 1927, and it remained the cheapest and preferred medium until the installation of undersea telephone cables in the 1950s. After 1980 other technologies, such as facsimile and electronic mail, took over the role of long-distance record communications formerly supplied by the telegraph.

In technological and scientific terms, the telegraph was important for three major reasons. First, telegraphy was the first major application of discoveries in electrical science, and it was one of the first technologies with a firm scientific foundation. Second, telegraph electricians during the mid-nineteenth century helped to establish the discipline of electrical engineering. Finally, technical problems, especially in submarine telegraphy, stimulated major scientific advances in physics and oceanography.

Telegraphy was the first technology to sever the connection between communication and transportation. Because of the telegraph's ability to transmit information almost instantly, it affected many aspects of society, culture, politics, international relations, and economics after 1840. It helped to create integrated national and international markets, sped the dissemination of news, provided a model for new literary forms, aided Western imperialism, and spurred national governments to develop telecommunications and technology policies.

1. *The Telegraph in Great Britain*

In the mid-1830s William Fothergill Cooke developed a telegraph which indicated letters by using the transmitted electrical current to deflect magnetized needles at the receiver. During the same period, Charles Wheatstone conducted several important experiments on the long-distance transmission of electricity. In February 1837, the two men pooled their efforts and in June they obtained a patent for an electromagnetic telegraph using five wires to deflect five needles in order to indicate letters of the alphabet. They built their first commercial line in 1838 and 1839. In late 1842, Cooke patented an improved telegraph which used two wires and needles, instead of five, to indicate letters; this halved construction and installation costs.

During the early 1840s, railways were the major customers of the telegraph. In 1846 Cooke and a group of investors founded the Electric Telegraph Company to build a commercial telegraph system for the use of the press and businessmen. By 1855 its network encompassed nearly every important town in England and Ireland. For the next quarter century, the Electric company dominated the industry but competed with four other companies for business. During this time the industry shifted from the Cooke and Wheatstone two-needle instruments to various forms of printing or recording telegraphs.

By the mid-1860s, much of the British public, Parliament, and the civil service supported the nationalization of the telegraph industry, claiming that Post Office operation would reduce rates, expand the network, and provide better service. In 1868, Parliament authorized the purchase of the telegraph lines, and the Post Office began operating them in February 1870. Despite initial successes, it was soon apparent that the British postal telegraph faced serious financial obstacles, including an inflated purchase price, unremunerative press and railway rates, commercial rates which barely broke even, and competition from the privately owned telephone industry.

At the turn of the century, telegraph traffic peaked at some 90 million messages. Throughout the following decades, traffic dropped and losses mounted because of competition from the telephone and the letter post which delivered letters within the British

Isles within 24 hours of posting. The telegraph's heyday was over by 1920, and by 1970 the telegraph had ceased to be an important communications medium in Britain.

2. The Telegraph in the USA

In 1837 and 1838, Samuel F. B. Morse publicly demonstrated his electromagnetic recording telegraph to prominent scientists and government officials. Morse's system at this time comprised an alphabetic code of dots and dashes, a telegraph key to encode messages, and a receiving register which recorded signals on to a moving strip of paper. Morse believed that the federal government should own and operate his telegraph, and hoped to convince the American government to purchase his system. In 1843, Congress gave Morse $30,000 to build an experimental line between Washington and Baltimore. This line began operation in May 1844, but the government refused to fund an extension of the line northward to New York. In 1846, the government turned over the Washington–Baltimore line to private investors.

Over the next 15 years, the telegraph network spread rapidly; by 1850, lines reached every important point east of the Mississippi River, and by 1861, a transcontinental line reached California. The major change to telegraph equipment in this period was the elimination of the recording register; operators shifted to sound reception by deciphering the clicks of the receiving magnet. During the late 1840s and early 1850s, a confusing welter of several dozen companies sprang up to build telegraph lines, many of them poorly organized and short-lived. During the late 1850s, a cartel of six companies coalesced and provided stability to the industry. The end of the Civil War in 1865 sparked a final wave of consolidation, and Western Union emerged in 1866 as the country's telegraph monopoly.

During the remainder of the nineteenth century telegraph technology remained relatively static. The only major advance arose in the 1870s, Thomas Edison's invention of the quadruplex which allowed the simultaneous transmission of four messages on a single wire. Between 1866 and 1910, Western Union faced two major challenges to its dominant position in the American communications market. During this period, a vocal and vigorous movement arose to nationalize the telegraph and place it under postal administration, and Western Union officials spent much time and energy defending their private monopoly to the public and to Congress. The telephone, independently invented in 1876 by Alexander Graham Bell and Elisha Gray, was Western Union's second major challenge. Bell offered to sell Western Union his patent, but Western Union refused and established a competing telephone system in the late 1870s using Gray's and Edison's patents. Because of Bell's superior patent position, Western Union withdrew from the telephone market in 1879 in exchange for Bell's promise not to compete with Western Union's long-distance telegraph business. In the 1870s and 1880s, the telephone could not operate over distances longer than a few dozen miles. However, by about 1890 telephone engineers expanded the range of audible conversations to a few hundred miles, culminating in the establishment of Bell's transcontinental telephone service in 1915. As Bell's long-distance network expanded and rates declined, the telephone steadily eroded the telegraph's share of the long-distance communications market.

Under the leadership of Theodore N. Vail, American Telephone and Telegraph (AT&T), Bell's parent company, was powerful enough to acquire working control of Western Union in 1910. AT&T's control of Western Union demonstrated that the telephone had eclipsed the telegraph as the nation's preferred long-distance communications medium. AT&T voluntarily relinquished its Western Union holdings in 1914 to forestall federal anti-trust proceedings. Although short-lived, Vail's administration modernized Western Union's antiquated accounting and management structure and spurred the company to replace its Morse instruments with automatic telegraph equipment.

Between 1915 and the end of World War II in 1945, the telegraph's share of the long-distance communications market continued to decline because of competition from the telephone, government air mail service, and AT&T's teletype service inaugurated in the 1930s. During the War, the Federal Communications Commission forced Western Union to divest itself of its international cable network and to consolidate with its sole remaining competitor, the bankrupt Postal Telegraph Company, two moves which weakened Western Union's long-term financial position.

Following World War II, Western Union's managers attempted to modernize the company's physical plant and to preserve a market niche for record communications. Between 1945 and 1980, the company undertook three major modernization programs, the development of analog facsimile technology, the construction of a microwave beam network to replace its wires and cables, and the launching of communications satellites. Although these systems were technologically successful, they failed to reverse Western Union's declining market share with respect to the telephone; by 1990, Western Union was defunct except for its money-transfer service.

3. Telegraphs in Asia and Africa

The first telegraph in India, constructed by two British army officers, opened in 1851. During the 1850s, the British government encouraged the construction of a

government-owned telegraph network to link the major trading and political centers, and by 1856 India had some 7,000 km of telegraph line and 45 offices.

The French telegraph network in Indo-China was also an instrument of colonial control. France began telegraph construction in 1861 and by the turn of the century the Indo-Chinese network comprised about 12,000 km of land lines and cables connecting over 200 offices. After 1880 the colonial powers occupying Africa built land and cable networks to connect their possessions with the home countries. The Indian, Indo-Chinese, and African telegraph networks were primarily instruments of imperial control and only secondarily commercial systems.

The Chinese government regarded telegraphy suspiciously, as a tool of the Western powers to gain control of their country. During the 1860s, Russian, British, and French entrepreneurs and government officials began pressuring China for telegraph concessions. Chinese officials resisted until the 1870s, when they granted foreign companies limited franchises to build telegraph lines. In 1881, the government established an Imperial Telegraph Administration to build and to operate a national network, but officials continued to resist the wholesale construction of foreign-owned telegraph lines. However, several European countries used the Boxer Rebellion of 1900 as a pretext to extract political and economic concessions, including the right to build telegraphs, from the weakened Imperial Chinese government.

Whereas Chinese government officials sought to shield the country from foreign influences in the second half of the nineteenth century, in Japan the Meiji Restoration of 1868 brought to power an elite eager to modernize the country along Western European lines. Japan's new rulers embraced the telegraph and hired a British telegraph engineer in 1869 to build a line between Tokyo and Yokohama and gave landing rights to a Danish cable company. By 1872, Japan enjoyed direct telegraphic communication with Europe. During the next two decades, the Japanese telegraph network expanded rapidly. By 1891, Japan had over 400 telegraph offices connected by nearly 12,000 km of wire.

4. *Submarine Telegraphy*

The successful development of submarine telegraphy between 1851 and World War I established an international communications network and helped Western nations to pursue their imperialist and commercial ambitions in Africa, Asia, and Latin America. From 1850 to the early 1870s, engineers confronted and solved several problems in submarine telegraphy: insulation, cable-laying ships and equipment, signal distortion and attenuation, and sending and receiving apparatus. British scientists, engineers, and entrepreneurs played leading roles in all four areas.

In 1851, a British firm laid the first successful undersea cable between Dover, England, and Calais, France. British and American entrepreneurs and government officials soon began exploring the feasibility of a transatlantic cable. In 1856, the American Cyrus Field and the British telegraph engineers John Brett and Charles Tilson Bright headed a group of investors who started the Atlantic Telegraph Company. The company received subsidies and naval support from the American and British governments. After two failures in August 1857 and July 1858, the company successfully laid a cable between Ireland and Newfoundland in August 1858. The cable failed, however, after a few weeks of weak and intermittent signals. The 1858 cable failure, the straitened finances of the Atlantic Telegraph Company, and the American Civil War (1861–65) all combined to delay a renewed attempt until 1865. Under the leadership of John Pender, the newly formed Anglo-American Telegraph Company successfully laid two cables in July 1866, marking the start of uninterrupted electrical communication between the two hemispheres. During the last third of the nineteenth century, the Atlantic cables were instrumental in establishing global markets in news, agricultural commodities, and financial securities.

During the late 1850s and 1860s, British telegraph engineers and government officials sought to connect Britain with its possessions in India, southeast Asia, and Australia. In 1870, they successfully laid a cable from England through the Mediterranean, Red, and Arabian Seas to Bombay. In 1872, an extension from India provided direct telegraphic connection to Singapore and Australia. Between 1870 and 1914, several American and British cable companies competed for the telegraph business of the Caribbean and Latin America, a competition which mirrored and fueled the general commercial rivalry between the two countries in the region.

On the eve of World War I, about 500,000 km of undersea cables existed, over half under British control and one-fifth under American control. France, Denmark, Germany, and The Netherlands accounted for the remaining cables. This global submarine cable network helped Western nations to attain economic and political dominance of Asia, Africa, and Latin America. During World War I, British leadership in cable and wireless communications played an important role in the Allied victory over the Central Powers.

During the 1920s, demand for international communications boomed and the submarine telegraph industry reached its peak. Furthermore, AT&T introduced a major technological advance, inductively loaded cables which increased message-handling capacity by a factor of five. Western Union installed the first inductively loaded cable in the Atlantic Ocean in 1924, and the new cables ensured that submarine telegraphy would occupy a central place in global

communications over the next several decades. However, economic depression and World War II between 1930 and 1945 reduced demand for international communications generally.

After World War II, two new technologies, telephone cables and satellite communications, converged to make international telephoning easier and cheaper, and hence to render submarine telegraphy obsolete by 1980. AT&T laid the first transatlantic telephone cable in 1956, and by 1983 seven transatlantic cables provided over 11,000 voice circuits. During the 1970s, satellite communications reached maturity, and by the early 1980s satellites provided about 50,000 voice circuits. The installation of digital fiber-optic cables during the 1980s increased the number of international telephone circuits by an order of magnitude and reduced greatly the cost of overseas telephone calls.

See also: Communication and Democracy; Communication and Social Psychology; Communication and Transportation, History of; Communication: Electronic Networks and Publications; Internet: Psychological Perspectives; Mass Communication: Normative Frameworks; Mass Communication: Technology; Telephone; Television: General

Bibliography

Ahvenainen J 1981 *The Far Eastern Telegraphs: The History of Telegraphic Communications Between the Far East, Europe and America Before the First World War*. Suomalainen Tiedeakatemia, Helsinki, Finland

Blondheim M 1994 *News Over the Wires: The Telegraph and the Flow of Public Information in America, 1844–1897*. Harvard University Press, Cambridge, MA

Carey J W 1989 *Communication as Culture: Essays on Media and Society*. Routledge, New York

Czitrom D J 1982 *Media and the American Mind: From Morse to McLuhan*. University of North Carolina Press, Chapel Hill, NC

Gabler E 1988 *The American Telegrapher: A Social History, 1860–1900*. Rutgers University Press, New Brunswick, NJ

Harlow A F 1936 *Old Wires and New Waves: The History of the Telegraph, Telephone, and Wireless*. Appleton-Century, New York

Headrick D R 1991 *The Invisible Weapon: Telecommunications and International Politics, 1851–1945*. Oxford University Press, New York

Hugill P J 1999 *Global Communications Since 1844: Geopolitics and Technology*. Johns Hopkins University Press, Baltimore

Israel P 1992 *From Machine Shop to Industrial Laboratory: Telegraphy and the Changing Context of American Invention, 1830–1920*. Johns Hopkins University Press, Baltimore

Kieve J L 1973 *The Electric Telegraph: A Social and Economic History*. David and Charles, Newton Abbot, UK

Oslin G P 1992 *The Story of Telecommunications*. Mercer University Press, Macon, GA

Perry C R 1992 *The Victorian Post Office: The Growth of a Bureaucracy*. Royal Historical Society, London

Reid J D 1879 *The Telegraph in America*. Derby Brothers, New York

Schwartzlose R A 1989 *The Nation's Newsbrokers*, 2 Vols. Northwestern University Press, Evanston, IL

Thompson R L 1947 *Wiring a Continent: The History of the Telegraph Industry in the United States, 1832–1866*. Princeton University Press, Princeton, NJ

D. Hochfelder

Telephone

Since its invention in 1876, the telephone has become, literally, a fixture of everyday life throughout the industrialized world and much of the developing world as well. It is so commonplace that only its absence becomes noteworthy (such as in school classrooms and prisons). Too easily forgotten by laypeople and scholars alike is the miracle of being able at any time to place a call to any other subscriber worldwide. The enormous value of the telephone can be appreciated if one only considers the plight of a villager who wants to know if there might be work available in a nearby town, or who needs to summon aid for a sick family member.

Also easily overlooked are the ramifications of this global connectivity. By allowing us to transcend a variety of physical and social barriers, the telephone has led to a complex set of dispersed personal and commercial relationships. By any measure, the telephone has dramatically altered the social landscape. Cars and airplanes were adumbrated respectively by horse-drawn vehicles and birds. Yet, in human imagination the power of real-time interactive oral communication over great distances had been a power reserved exclusively for divine beings. Given this godlike power of the telephone it is little wonder that it has been one of history's most successful inventions. Little wonder, too, that some of earliest experiments with it were designed to ascertain whether the telephone could be used to communicate with the dead. (Perhaps not regrettably, the answer is that it cannot.)

1. Global and Microcoordination

Today, for a billion people, the act of making or receiving a telephone call in their home or business is a normal, often daily, occurrence. (Another billion, at the other end of the development spectrum, have in their entire lives never used a telephone.) For rapidly growing numbers of people worldwide, car driving and cityscape walking are becoming settings for making and receiving phone calls. In the United States, 200 or so billion calls are made annually from 100 million households, with many billions more generated from offices, mobile phones, and computers. Given the commonplace quality of the ordinary

telephone it remains hard for us to delineate, let alone appreciate, the telephone's ramifications throughout global societies and the interior world of individual psyches.

Although the telephone has numerous cumulative macrosocial consequences, most of them are predicated on its ability to support social interaction. An essential aspect of the telephone is its ability to allow coordination among geographically dispersed (or even locally concentrated) people. Numerous studies demonstrate that the telephone promotes business, sociality, democracy and information flow. For those who are physically handicapped or otherwise constrained, it opens vast arrays of contacts. (It is interesting and perhaps surprising to note that the telephone has increased greatly the social contacts and informational resources for the deaf, via TTY interfaces. This quality is particularly gratifying given Alexander Graham Bell's original interest in helping the deaf.)

The telephone allows convenient management over distance, and via voice messages, asynchronously. Upon its basis, extraordinarily large enterprises can be established and effectively controlled. As a corollary, those who would otherwise have greater independence find themselves circumscribed by those (sometimes literally) above them. The telephone has enabled the rapid conduct of large-scale financial transactions and has greatly assisted the formation of capital markets.

A collateral effect is that digital telephonic technology, including voice-mail systems, has been partly responsible for the elimination of huge numbers of middle managers and support staff throughout American industry. Ironically, many of these positions were created in the first place to deal with the communication flood created by the analogue telephone. This picture is also influenced by the answering machines, which of course allow the conveyance of telephone messages without human intervention. In former times, self-respecting managers, professors or other professionals would be reluctant to conduct their business without having a secretary on staff. (Indeed, much effort necessarily went into 'faking' the existence of a secretary to act as an intermediary.) Today, for many professionals, having a secretary would seem to be not only wasteful but also an indication to their clients that they are out of touch with current best practices.

The proliferation of digital answering machines also affects the tenor of professional events. For instance, it used to be the case that at a conference or symposium, the break-time would be used for social interaction and enjoying the 'here and now' aspects of the conference venue. Nowadays, however, attendees rush outside to use their mobile phones or crowd the payphone booths in an attempt to call their remote answering machines, collect messages, and return phone calls. In another context, the mobile phone and pager has led to a new form of social effrontery. While seemingly engaged in conversation, one coolly glances down and scrolls through messages on one's communication device, all the while nodding sagely as if heeding every word uttered by one's interlocutor. These are a few of the many faces of what Kenneth Gergen has termed 'absent presence.' People are seemingly at an event or socially engaged, but their attention and mental focus is elsewhere.

Telephones, especially mobile phones, increase the pace and efficiency of life. They also allow more flexibility and individual efficiency at business and professional levels and at those of family and personal life. Moreover, people can take action to allay the feeling they are not accomplishing enough. People can harness spare time, or time previously spent in tasks that do not require much attention (for instance, waiting on a shopping queue) to plan and coordinate with others, get information, or even shop remotely by phone while they also shop in person. On the other hand, those who treasure respite may find themselves pressured to replace otherwise excusable isolation with productive tasks. Once upon a time, being aboard an airplane excused an executive from having to interact with colleagues. No more, for the fax and phone now follow even at six miles high; nor are the seashore and mountaintop beyond their reach. An age of perpetual contact, at least in potential terms, is dawning.

Despite the many advantages to having the telephone available any time and anywhere, there are drawbacks to it as well at both the individual and social levels. The telephone can be a minor irritant or a conduit for crime. In terms of the former, many social critics of the late nineteenth century expressed irritation at the way the telephone seemed to destroy the orderly pace of life, the considerable pleasures of social propriety and as well those of solitude and reflection; the target has remained lucrative, and the frustrations of telephone use and abuse remain a staple of the screen, both large and small. The telemarketing solicitation, of which there is a rising tide, is also an oft-remarked-upon irritant. Telemarketers themselves use market research, so it is no accident that their calls seem to arrive at dinnertime or other exquisitely inconvenient moments. More serious, though, is that there are a large variety of swindles—some of which even use computer-aided dialing schemes—that prey especially on those with the least social resources. Stalking by telephone and obscene phone calls to both men and women are also serious problems. On the other hand, with advances in technology that give call recipients greater power to control access to themselves (such as increasingly popular caller-identification services) people are better able to manage their communication environment to suit their individual tastes and needs.

The increasingly nuanced technology available to callers and callees demands ever greater sophistication to achieve a flawless performance of the choreography of communication. Think of the dilemma of a shy but sedulous teenage boy who wants to ask a popular girl

to the school prom. Presumably he would not want to leave on her caller-identification box a perusable record of his forty or so attempted calls before finally reaching her. The degrees of freedom and maneuver are reduced with intriguing consequences that cannot be explored here. Suffice it to say that, sadly, it appears that there is an iron law of telephone contact: people generally get the most calls from those from whom they would least like to hear, and vice versa.

In all, though at times a source of irritation and adroit tool of criminal activity, the telephone has proven to be an important factor in increasing the pace, scope and intensity of business and management operations while simultaneously lowering the cost of engaging in these activities. It has had a parallel effect on people's levels of social interaction.

2. *Geographical Distribution*

The telephone has influenced substantially the distribution of people across the physical landscape. On the one hand, the telephone has allowed the massive vertical downtown urban center to survive then thrive. Without the telephone, the skyscraper would have been most difficult to build and manage. Further, it would be difficult for the structure to maintain its usefulness to its denizens without the telephone.

Some claim that the telephone has enabled one particular form of geodemographic dispersion, the suburb. However, detailed studies of transportation system development indicate that the telephone was not an important factor in this remarkable internal migration. Still, the telephone does allow many in rural locations to participate in business and social relations in a viable way, which without the telephone would not be possible. At the same time, those in extremely low-density areas have persistent difficulty getting even minimal telephone service; the situation is exacerbated by the fact that this populace is generally poor and remote from economic opportunities. This, unfortunately, is a social problem that will endure for decades. The telephone, though, has allowed those who choose isolation—and such people are often drawn from the middle and upper strata of society—to enjoy it without having to forego ready communication. Some seek to 'get away from it all' but still want employment as consultants, or at least to stay in convenient contact with friends, family, and emergency services.

3. *Economic Development*

It has long been an article of faith that the telephone promotes economic development and social welfare. This belief has had powerful repercussions on national telecommunications policies worldwide. These have ranged from subsidized rates for the poor, subsidized rates for everyone, and special incentives for telephone companies to modernize their systems or physically extend their services to remote areas. Certainly there has been a strong association between telephone lines and economic growth, particularly, as Hudson (1984) has shown, in developing countries. Researchers have demonstrated a logarithmic relationship between main telephone lines and national income on a per capita basis, although the direction of the causal arrow has not yet been established.

What is not in doubt is that the telephone itself fosters employment growth, not the least to serve itself as an industry. In the early days of mechanical switches, when telephone systems were growing dramatically, it was speculated that within a generation all young boys in America would have to be employed simply to connect phone calls. Although this scenario never materialized, phone companies have traditionally been enormous consumers of manpower, and are usually one of the largest employers of a region. Traditional telephone companies everywhere, though, are for the most part looking for ways to reduce their traditional workforces, heavily concentrated in jobs of installation or as operators. They are replacing vast numbers of these workers with smaller numbers of software engineers and technicians.

Yet even as the workers per 1,000 telephone lines decreases worldwide, employment in the wireless telephone industry has mushroomed. In many cases, the growth in the mobile phone sector has far surpassed the cuts in the traditional wireline companies (many of which were formerly government ministries of Post, Telegraph, and Telephone services or PTTs). As the expense of laying copper wire is far greater than erecting a radio tower, many people in developing countries are gravitating to mobile phones simply because they can get them more quickly than wireline ones. It is my guess that because the mobile phone are more flexible and can be used in more places, the economic impact of growth in mobile phones will be greater than that of the growth of their tethered counterparts.

Telephones can have other economic effects: they ease the job prospecting process directly, increasing economic efficiency by making it easier to change jobs. Also, by using the telephone as a sales tool, consumption is stimulated, which leads to job creation and economic growth. Staffing the sales force alone of course means that jobs will be created. To illustrate by the situation in the US, according to the Direct Marketing Association, spending devoted to telephone marketing is about $62 billion annually and comprises nearly 40 percent of the nation's direct marketing expenditures. The telephone's indirect employment impact in telemarketing is substantial; in 1998, employment in direct telephone marketing was estimated at more than nine million workers.

Although much of the prior discussion focused on macroeconomic effects, it is also worth mentioning

the two-fold microeconomic impact of the telephone. First, it allows individuals to run their own business, often from their home or voice mailbox. Thus, it encourages small enterprises. Yet, it also encourages large-scale enterprises as well since it allows firms to capture the benefits of large size without excessive coordination costs as discussed above.

4. *Polity and Politics*

Due to its ability to connect quickly remote areas, the telephone has helped foster national integration. At a structural level, it also serves to centralize exercise political authority while concomitantly circumscribing regional centers of power. Notwithstanding this structural aspect, the telephone has something of a reverse consequence for individual political empowerment. In this regard, Ball (1968) and Aronson (1971) argue that the telephone has been a politically democratizing instrument. There is much to this viewpoint, and studies have shown that, controlling for levels of economic development, in Communist countries of Eastern Europe there were equivalent numbers of TV sets when compared to their Western European counterparts. However, there were drastically fewer telephones. The presumed reason for this maldistribution is that the telephone enables two-way communication, whereas the TV allows controlled information to be sent in a one-way direction from centralized source to a passive mass audience.

The important role the telephone system could play in controlling the populace has not been lost on governments. In the late nineteenth century, France wanted a centralized, governmentally controlled telephone system because authorities anticipated that rebels might otherwise be able to seize control of it during an uprising. During the Soviet communist regime, the absence of telephone books in Moscow was legendary. Telephone tapping is a long practiced art that has been used to fight crime and enemies of the state.

Revolutionaries as much as police recognize that the telephone is the body politic's nervous system. Hence, it is the universal target for those seeking to subvert or protect the government. Lenin directed that the telephone and telegraph offices be initial targets of his Bolshevik seizure of power. As a contrary example, we point to the August 1991 coup attempt against Mikhail Gorbachev and his government. The plotters failed to seize the telephone system. Consequently, the anti-Communists were able to use it to rally themselves. Boris Yeltsin and his supporters, barricaded in the Parliament building, were able to make and receive calls from around the country and the world; they used this power to marshal and coordinate support. A report on Russianet even claims the parliamentary defenders were able to phone in their food orders to the Moscow Pizza Hut. The telephone facilitates the management of political campaigns and the operation of political organizations.

For well-organized campaigns, the telephone bank has become an important element in any contest. They can also be used to generate otherwise lethargic public interest in issues. In the United States, a neologism has been created to describe the abuse of this process. The term is 'astroturfing,' in honor of the artificial grass of the same name. Astroturfing is the process of using telephone banks to create the incorrect impression among politicians that there is a 'grass-roots' swell of public concern about an issue when in fact there is little or none. Instead, the outpouring of opinion on the issue had been catalyzed by a small group of telephone operators who 'pipe through' a citizen to her representative's office to recite a prepared message. This example provides further demonstration of the principle that telephone technology has an impact that is governed largely by the motive of the user.

5. *Societal Aspects*

Traditional sociological variables of age, social class, socioeconomic status, gender, and race/ethnicity influence people's relationships with the telephone. Young people are generally much more likely to adopt new technology than their older counterparts. Material resources, as measured by income, also have a noteworthy effect. Predictably, those at the very bottom of a society's income distribution are least likely to have telephone service. More surprisingly, as shown by Katz's (1999) research, people in the third income quartile are disproportionately represented among purchasers of enhanced telecom services. This may be because these services are perceived as entertaining or of high prestige. This noteworthy trend occurs in both Europe and the US; whether it obtains in other regions is unknown.

Since its inception, there have been speculations about the small-group and social interactional consequences of the telephone. Marvin (1988), in her path breaking study of the social reception of the telephone (among other nineteenth century electrical technologies), demonstrated that concerns over the impact of the telephone were reproductive of those the larger society. Among these were the erosion of family authority, race and social class barriers, and sexual and social propriety. Also of substantial concern was disrespectful and deceitful behavior. Victorian-era Americans believed, justifiably, that the telephone reduced social cues ordinarily conveyed in the richer channels of face-to-face interaction and written correspondence. By masking cues, such as the location and social status of the caller, the telephone denied important cues to assist in making judgments about the interlocutor. Thus, it would be easier to violate

social and legal codes. These concerns remain substantial even after a century of telephone use, and are born anew with contemporary communication technology such as the Internet.

The telephone plays a role in the conduct of gender relations and in the pursuit of sexual gratification. In contemporary, often suburbanized society, the telephone has become an important means by which teens socialize themselves, build and sustain their relationships, and amuse themselves in an historical period when they are defined as problematical non-adults Wynn and Katz 2000. Commonly, women run the 'social work' of maintaining family links and social schedules via the telephone. (All other things being equal, when both a man and a woman are in a room together, a woman is twice as likely to answer a ringing phone.) Rakow (1992) has documented the important role the telephone plays in women's relationships, and the ways it provides support for their independence despite various constraints and provides them with social resources they need to pursue their relationships and goals.

As in other areas, the telephone can enhance coordination in people's search for partners. The telephone's role in abetting prostitution has long been noted. (Indeed, a term-of-art, 'call-girls,' clearly refers to the role of the telephone.) While it is impossible to capture precisely the temporal, economic, or organizational magnitude of the telephone's role in the commercial sex industry, plainly large sections of the yellow pages and of urban newspapers are devoted presenting this information in a slightly hidden argot to knowing patrons. Phone booths in many countries are littered daily with miniature handbills enticing customers to one service or another.

The telephone can also serve as a less direct conduit for sexual expression. Indeed, the conduct of remote interactive sex is an entirely *de novo* mode of expression enabled by the phone. The remote as well as the disembodied aspects of a telephone call can have its sexual appeal, and in this regard one hardly need mention the candlestick design that characterized phones for a half century. By relying on its aural and fantasy capabilities, it can also serve as a safe and sanitary substitute for in-person sex. Recognizing this pent-up demand, an entire industry of 'pink services' has sprung up in any country where it is allowed. Definitive data as to its size are difficult to obtain, but an instructive dataset can be derived from the Austrian experience. Austrian law had forbidden phone sex historically. However, in 1991, a revenue-sharing arrangement was established so that offshore companies could benefit from premium international telephone tariffs. The Dutch Antilles became a center of such 'pink services', and numerous companies began placing ads in Austrian newspapers giving their Antillian numbers. The results were dramatic and instructive. For all of 1990, the total duration in minutes of calls to the Dutch Antilles from Austria was exactly 13,050. For the first nine months alone of 1991, the number of minutes of calls from Austria to the Dutch Antilles exceeded 1,600,000.

Not all telephone sex is predicated on a financial arrangement. Many millions of obscene calls are reported annually to authorities. Both men and women make obscene phone calls to obtain erotic stimulation from the reaction of their shocked victims. These types of calls seem to occur throughout the world wherever a telephone system is present. As a case in point, within six months after the first automatic telephone network was installed in Papua New Guinea, an obscene phone call was reported to authorities. A third class of sexually oriented telephone call is neither commercial nor inflicted on victims, but rather occurs by mutual consent. For decades, lovers have enjoyed erotic phone conversations, and the topic receives substantial treatment in literary and cinematic formats. Nevertheless, seldom has public attention to this private aspect of human behavior risen to the heights it did in 1998. As part of a larger relentless investigation, an incident of sexual gratification via telephone between US President Bill Clinton and his White House aide, Monica Lewinsky, was given protracted national media attention and related to Congress in an official report. Thus, presidents and perverts alike serve to demonstrate that the telephone, like every other communication technology—from the video camera to the car, and from the printed book to the web-enabled mobile phone—can be used by people as a context for sexual excitement and fulfillment.

Another striking aspect of the telephone is its use as a tool of personal empowerment. With emergency phone boxes set up on campuses, and especially with the mobile phone, people can feel safer in public areas. It has expanded the places and times that people, and especially women, can go to a variety of places, thus increasing people's freedom and mobility. It also permits parents to expand the privileges they give their children in terms of when and where they can go. This has the intriguing result of both increasing and loosening parental control.

The telephone also enhances individual autonomy by allowing people to live more successfully by themselves. Yet it preserves instant access between those experiencing a medical or police emergency and a variety of support services. In addition, a substantial variety of community and psychological services are available at any hour and often with complete anonymity.

As with any technology, the use to which it is put is influenced by the motives of the human user. Just as not all people are good, neither are all purposes to which the telephone is put benign. The telephone has offered new venues for monitoring individuals, invading their privacy, and harming them materially or psychologically. In terms of monitoring, the wireless revolution, particularly as it intersects with the tele-

phone, is increasingly affecting all aspects of social interaction, from the conduct of relationships to the receipt of medical care. These technologies will not only be used to deliver services but as well to monitor individuals. In the past telephonic monitoring has targeted parolees and the elderly. The number of citizens who are regularly and passively tracked will be increasing; this will be done through the data collected by systems used to regulate mobile phones and intelligent highways. Already mobile phone records are being used in civil (e.g., divorce actions) and criminal cases (e.g., the apprehending of kidnappers).

The telephone also gives individuals tremendous powers to affect others' lives at a distance. Though often used in pranks and practical jokes, the telephone has also been used to inflict protracted harassment and terror on others. Such practices include spreading malicious rumors, making harassing calls late at night, and even concocting identities in order to manipulate or frighten someone. In one case, a young girl pretended to be a nurse, and in this guise reported (falsely) to another girl that her lab test results indicated that she had AIDS and was pregnant. The girl victim attempted suicide and disaster was only narrowly averted. The culprit was caught using caller-ID.

Turning our attention to the underpinnings of social life, Fischer (1992) found that a social–psychological effect of the telephone was that it seemed to accelerate the pace of social life. His classic study, *America Calling*, found that the spread of the telephone in the US reduced isolation and increased social contact. A concomitant finding is that a good proportion of telephone users experienced anxiety or 'communication apprehension.' Many citizens also feared excessive social contact, becoming the subject of telephone gossip, and being subjected to unwanted calls. However, he found no support for more extreme claims that it brought people to a constantly heightened sense of alertness and anxiety over the potential summoning outcry of the telephone's ringing bell. Fischer concludes that the telephone cannot be implicated and the emergence of 'psychological modernity,' that is, the contemporary sensation of free-floating anxiety and dehumanization. Rather, he concludes that on the whole, people readily and comfortably assimilated the telephone into their lives. They used it as 'the conscious product of people employing things, not of things controlling people.'

6. *Technological Convergence*

It is impossible to speak of today's telephone exclusively in its formerly traditional wireline configuration. Mobile phones, discussed below, are but one aspect of this telephonic reconfiguration; it extends to all aspects of voice communication. One salient aspect is the way people manage their calls. For instance, in the US, the majority of telephone subscribers also have some form of answering machine. This device changes how people use their phones, e.g., to monitor before picking up the receiver, and which in turn affects the behavior of the caller ('I know you're there! Come on, pick up.'). Having the telephone answered by a machine was until 1987 seen by the majority of the US population as an act of rudeness to callers. It was emblematic of the fear that 'robots would take over,' and that people were losing control over their daily lives. (This also ties into the above point about the importance of having a secretary to handle one's calls.)

Today opinion surveys show that the majority judges it an important courtesy that should be extended to callers. Some even go so far as to say that it is rude not to have an answering machine or voice mail. Still, the answering machine/voice mail is but one of the ever-multiplying ancillary service that are being added to the formerly black 'one-size-fits-all' telephone. These include built-in faxes, caller-ID, three-way calling, call-waiting, and call-trace; together they are changing what a telephone is and how it is used, as well as the subterfuges that are used to work around the telephone's expanded capabilities.

Even as the telephone becomes more powerful and capable, it is becoming tetherless. Not only is the home phone increasingly likely to be cordless with reception to a $\frac{1}{2}$-kilometer from base, but people are increasingly subscribing to mobile telephone services that will allow them to range across continents. Moreover, some people are abandoning their wireline phones entirely; in the US, about four percent of subscribers have done so already and their numbers are increasing. Among other things, this trend means the loss of geographical meaning to area codes. Conversely, the physical telephone line is for most Internet users their primary access mode; but mobile telephones are also becoming Internet enabled. Although currently most users access the Internet via the computer, it seems inevitable that soon the majority will be doing so from enhanced mobile phones.

Computers themselves are increasingly encompassing the features of the telephone, becoming Internet phones and answering machines and voice chat messaging units. Personal digital assistants are beginning to act as both mobile phones and portable computers. Platform capabilities are becoming intermixed, that is, telephone conversations can occur virtually anywhere with anybody via any machine. One consequence of this technological advance is that, without too much hyperbole, it appears that there is arising a new sense of 'perpetual contact.' This new state will allow better co-ordination, improved business operations and enhanced kin and social relations. It will also introduce new conflict over the appropriate norms for, and possible constraints on, people's public behavior. To cite an instance, there are now mobile phone-free rail cars and restaurant sections.

Already, however, the proliferation of mobile phones has created a brouhaha over their use in public places, most especially while driving cars. Though some of the controversy revolves around manners and appropriate levels of sensitivity, the larger issue of the use of and conduct in public places is only beginning to be addressed. It remains an unsolved puzzle for social scientists and the architectural profession to create public spaces that can amenably coexist with privatizing mobile telephonic behavior. However, the much more powerful issue is the question of safety. Mobile phones mean not only that lives can and will be saved but also that—because there are those who elect to use it while driving—lives will be lost. Sadly, there is a growing toll of people killed by mobile phone using drivers. Those who are affected by such tragedies will be fostering legislation to restrict the use of mobile phones in cars.

7. *The Telephone as Synecdoche for the Modern Condition*

The telephone has had substantial impact on the structure and quality of human existence, from business to sex, and on the conduct of war and peace from the international to the intra-family level. Never has so small a device been used so easily by so many to do so much to so many.

The use and abuse of the telephone encapsulates and expresses the dimensions of social scientific perspectives on modern life. What happens in the larger canvas of life is reproduced within the ambit of life on the telephone. The telephone's impact stems from the inherent qualities of a technology that allows easy, inexpensive real-time interaction by voice over both short and great distance. People have investigated aggressively and exploited this technology's potential. Through this unrelenting effort, there has been an isomorphism between what the technology can achieve—as its capabilities become more sophisticated—and the needs and desires that exist within mankind's biological and social framework.

The telephone system has been used in ways as broad as the human imagination, many times in ways that were never envisioned (or were even opposed) by telephone company system designers and marketers. Telephone systems have been used to broadcast operas and provide wakeup, dial-a-joke, and dial-a-prayer services. Psychological and emergency services, astrological advice, and interactive matchmaking services are among its staples. Users themselves have created a variety of 'work-arounds' to beat the telephone system (e.g., tone generating boxes that allowed caller to bypass toll charging mechanisms). None of these uses or innovations was foreordained. Rather, as shown by Coopersmith (1991), far from a calculated manipulation of telephone company executives, some innovations as enormously successful as the fax and answering machine were at various points in their development opposed and thwarted by telephone company executives.

No dictator, the telephone has become a servant to those pursuing profit and pleasure. The picture after more than a century and a quarter is quite unlike those sketched originally by hand-wringing dysutopians, such as Mark Twain. These critics saw the technology as depleting social capital, killing social relationships, and inducing depression (a view similar to today's critics of the Internet).

On balance, the telephone has clearly done far more to liberate humanity than enslave it. This observation is not meant to downplay the harm that can be and is inflicted telephonically. Like two other communication technologies—the airplane and the car—the telephone can be used for both good and evil. It can be used to wreak suffering and terror; it can be used to carry messages of love, hope, and redemption beyond the vast oceans. In this context, then, we should recall the lessons of the Bible and the Greek myths that even the existence of a powerful god is not without frustration.

See also: Communication: Electronic Networks and Publications; Communication: Geographic Aspects; Communication: Philosophical Aspects; Immigration: Public Policy; Internet: Psychological Perspectives; Telecommunications and Information Policy; Telegraph

Bibliography

Aronson S H 1971 The sociology of the telephone. *International Journal of Comparative Sociology* **12**: 153–67

Ball D W 1968 Towards a sociology of telephones and telephoners. In: Truzzi M (ed.) *Sociology and Everyday Life*. Prentice-Hall, Englewood Cliffs, NJ, pp. 59–75

Coopersmith J 1991 The failure of the fax: When a vision is not enough. *Business and Economic History* **23**: 272–82

Fischer C 1992 *America Calling: A Social History of the Telephone to 1940*. University of California Press, Berkeley, CA

Gergen K J 2000 A century of absent presence. In: Katz J E, Aakhus M (eds.) *Perpetual Contact*: *Mobile Communication, Private Talk and Public Performance*. Cambridge University Press, Cambridge, UK

Hudson H 1984 *When Telephones Reach the Village: The Role of Telecommunications in Rural Development*. Ablex, Norwood, NJ

Katz J E 1999 *Connections*: *Social and Cultural Studies of the Telephone in American Life*. Transaction Publishers, New Brunswick, NJ

Marvin C 1988 *When Old Technologies Were New: Thinking About Electric Communication in the Late Nineteenth Century*. Oxford University Press, New York

Pool I de Sola 1977 *The Social Impact of the Telephone*. MIT Press, Cambridge, MA

Rakow L F 1992 *Gender on the Line: Women, the Telephone, and Community Life*. University of Illinois Press, Champaign, IL
Russianet.http://www.russianet.ru/~oldrn/history/coup.html
Wynn E, Katz J E 2000 Teens on the telephone. *Info* **2**(4): 401–19

J. E. Katz

Televangelism

Televangelism refers to what used to be called the 'electric' or 'electronic church,' which is the religious programming sponsored by evangelical Christians using commercial television time with the purpose of bringing individuals to salvation. Televangelism as a term was first developed in *Prime Time Preachers: The Rising Power of Televangelism* (1981), which explores the rise of the phenomenon in the USA. The term is also fairly interchangeably used with 'electric' or 'electronic church' as defined by Ben Armstrong (1979) to define the use of broadcast media, first radio, then television, to deliver church messages. Schultze presents the 'electronic church' as a construct defined by its 'business values, experiential theologies, media-derived formats, faith in technology, charismatic leaders, and spin-off ministries' (1990, p. 42). Either term refers to both the broadcast itself and the institution and norms that are created within it to demonstrate its values (Frankl 1987).

1. *The Development of Televangelism*

The technological development, adoption and regulatory environment surrounding the advent of radio and television in a given society is critical to the development of televangelism and for that reason televangelism varies across nations, depending largely on the degree and type of control over media programming. The creation and tone of televangelism in the USA is very closely tied to the rules governing programming in the early days of radio and the ways in which evangelical Protestants took advantage of shifts in media regulation.

1.1 *Televangelism in the USA*

In the early days of radio regulation in the 1920s, the Federal Communications Commission (FCC) required broadcasters to provide free access to the airwaves in the form of public service broadcasting (Finke and Stark 1992). The networks then turned to the Federal Council of Churches as a way of streamlining the work of picking groups that would be allowed access to the airwaves. The Council of Churches, however, represented 25 mainstream denominations (Bruce 1990) and created guidelines to promote mainline denominations, shut out fundamentalists and encourage local churches to share time in ecumenical broadcasts. Because the churches that made up the council quickly monopolized the available free time, only a couple of the largest evangelical churches were able to obtain time through the public service broadcasting avenue. In addition, little opening existed for purchasing time (Bruce 1990).

Protesting evangelical organizations formed their own pressure group for lobbying the networks; this group eventually became the National Religious Broadcasters (NRB) in 1944 (Bruce 1990), after the Mutual Broadcasting Company announced new policies that made it even more difficult for evangelicals to reach the airwaves (Hadden 1993).

The NRB was successful in obtaining access to paid time for its members. For over a decade following its founding, the NRB made significant strides in increasing the visibility of evangelicals on the air and by the late 1960s 'dominated the religious airwaves' (Hadden 1993, p. 116). The theological emphasis on salvation dovetailed neatly with the necessity of evangelical broadcasters to ask for money to support these programs. Hadden asserts that televangelists offered viewers the 'product' of salvation and then asked them to contribute in order to help the broadcasts to continue to reach new (and unsaved) viewers (Hadden 1993, p. 117). Through being forced to pay for air time and thus offer programming that viewers were willing to support financially, evangelicals were well positioned for the changes to come in FCC policy. In 1960, the FCC ruled that broadcasters did not have to give away time for it to count as public service time; time previously given to the mainline churches could now be sold, not given, to them under this policy (Hadden 1993, p. 118). While the mainline denominations protested this change, they were unable to maintain their access to free airwaves and evangelicals moved in to purchase time from local stations. In addition, videotape made it possible to for broadcasters to distribute their programming inexpensively for simultaneous airing in multiple markets (Roof 1997). These two developments in policy and technology, coupled with the programming savvy of the evangelical broadcaster positioned them for the airwave dominance that they maintained until the scandals and political activism of the late 1980s and early 1990s.

While televangelist scandals are well known for landing one broadcaster in jail and ending several television ministries, the effect of the political activism embraced by the broadcasters is less frequently mentioned. The political stances and eventual presidential bid of Pat Robertson and the activism of Jerry Falwell are correlated with the stagnation of their audience share (Hadden 1993). Since the dramatic expansion of cable wiring in the 1980s many televangelist broad-

casters have embraced cable and broadened their programming to include not just services, but also other 'family friendly' programming such as reruns of early network sitcoms.

1.2 Televangelism in Sweden and African Countries

While the experience of televangelism in the USA has been rather high profile and well documented by journalists and scholars alike, it is worth considering the development of televangelism in other cultures to understand more fully the influence of the intersection between religion, technology, and policy, and the development of media programming. Comparative work on religious broadcasting in the USA and Sweden by Linderman (1996) notes the differences in the religious environments in the two countries. Where the USA has a tradition of pluralism that in the case of religious broadcasting was for a time dominated by mainline churches, the Church of Sweden is the Lutheran national state church within a nation of religious freedom. Other religious bodies are the Pentecostal movement and the Mission Covenant Church, which vie with the national church for airtime. The Swedish Broadcasting Corporation allocated 75 percent of time for morning services and all major holidays to the national church and the rest of the time was allocated to the council representing the free churches. However, the Swedish Broadcasting Corporation eventually came to believe that it should regulate not only the time for religious broadcasting, but also the format and content. Their view was that 'religious radio should be more open, and able to communicate the general Christian message to people outside the churches—thus in a way being more "evangelistic" compared to the view of the churches' (Linderman 1996, p. 102). In addition to the broadcasts from Swedish religious organizations, the expansion in the number of channels received by viewers in Sweden also brought televangelists from the USA to Swedish airwaves, including the 'Hour of Power' with Robert Shuller (Linderman 1996).

Work on televangelism in Africa includes discussion of the early entry of international televangelists from the USA and European countries into the African media and debate over the appropriateness of the format. Hackett finds that early televangelism in Africa consisted of broadcasts of American televangelists with broadcasts in Ghana of Oral Roberts from the late 1970s to 1982, and Nigerian broadcasts of Pat Robertson's CBN news in the 1980s (1998, pp. 262–3). Hackett also notes that the development of televangelism and the dissemination of mediated religious messages is tied with the development of the megachurch through the charismatic leadership of evangelical churches in the region. In addition to the radio and television broadcasts, tapes are marketed and shared among family and friends, and thus further disseminated throughout the community. However, the development of televangelism is criticized by some clergy as inappropriate in a region with so much poverty (Asuzu 1987).

2. Approaches to the Study of Televangelism

Televangelism involves media, religion, sociology, history, political science, and economics, thus, is under examination from a variety of disciplines. Frequently cited for their work in documenting the development of televangelism both as a form of mediated religion and as cultural construct are Hadden and Swann (1981), who coined the term televangelism and extended interest in televangelism beyond religious circles to involve sociology and economics. In addition, Bruce's *Pray TV: Televangelism in America* (1990) revisits the history and the current issues in televangelism and asserts that the audience figures for televangelism are overestimated; this debate will be explored more fully later. Some analysis of televangelism has focused on the confluence of the economic and religious marketplace (Finke and Stark 1992). Historians have chronicled the role of religious broadcasters in political debates (particularly Father Coughlin in the l930s) as well as the struggles between religious factions for control of the airwaves.

2.1 Televangelism as Ritual

Other researchers have considered televangelism as a ritual informing the lives of viewers, connecting them with a broader community of believers. These scholars claim that the ritual transforms the everyday world into the symbolic and through reflection provides a base for guiding one's everyday life (Alexander 1997). Televangelism may serve to legitimize the messages performed in the course of the program. This concept of the ritual within televangelism connects it to the interest in the audience for such programming. Study of the televangelism audience takes many forms, from the interest in how believers approach the form as a ritual to analyses of a market share and ratings. Hoover's *Mass Media Religion* (1988, p. 112) examines *The 700 Club* audience in terms of how viewers' histories intersect with their experience as viewers both as individuals and as a community formed as an audience. Comparative research on the reception of re ligious television in Sweden found that US evangelicals differed greatly from Swedish evangelicals in terms of how they perceived the program's believability, pointing out the possibility for future exploration of how media and religion in differing cultures produce these outcomes (Linderman 1996).

2.2 Debates Over the Size of the Audience for Televangelism

As previously alluded to, questions over the size and scope of the audience for televangelism has fostered a

debate that includes not only scholars from a variety of disciplines, but also religious broadcasters. Early studies of this audience, most notably one supervised by the Yale Divinity School in 1955, attempted to delineate attributes of the audience correlated with religious programming and addressed the potential influence of level of religiosity, socioeconomic status and denominational affiliation (Parker et al. 1955). Hadden and Swann (1981) sought to balance the proclamations of the televangelists themselves and ratings by Arbitron by comparing systems of measurement in order to explain why the televangelists' numbers were so high and Arbitron's so low. They concluded that while Arbitron's figures may have been low because of the failure to include cable data, that the rating company's numbers were probably the most accurate. Hadden and Swann also concluded that while the numbers of viewers were not insignificant, the audience for televangelism had declined from 1978–80 by approximately two million viewers (1981, p. 55). Bruce (1990) noted that a 1987 study using data from both Arbitron and Nielson produced radically different audience figures, 13.3 million and 70 million viewers, respectively. These discrepancies, of course, point out major problems in comparing audience measurements based on differing assumptions, but they also call to question the power of televangelism as a force for mobilization. Hoover's discussion of the myths surrounding religious broadcasting labels the assumption that there is a large and 'significant' audience for televangelism as such a myth (1990). He notes that the funding formulas for televangelism differ considerably from that for standard commercial television and that audience size, donations to meet costs, and the ability of a program to stay on the air are not interdependent. Hoover also points out the importance of televangelism programming from within an evangelical group as opposed to the importance assessed by those from outside that community who may tend to overemphasize the significance of a program being on the air. Obviously, given the differing perceptions of those within and without the televangelism community and the differing means of audience measurement, this debate will not be easily settled and its continuation fuels other questions of secularization and the intersection of media, religion and culture.

3. *Future Directions in the Study of Televangelism*

As the media landscape has changed considerably, not only in terms of forms of media, but also the economic structure of media, the concept of televangelism is probably going to change with its environment. While the dramatic growth of the Internet with its potential for religious communities, dissemination of information and the creation of decentralized networks of individuals is certainly one avenue, more closely tied to the original development of televangelism is the structure of the mass communication industry. It appears that two intertwined avenues are likely for future investigation. The first, the economic structure of mass communication, has produced powerful conglomerates for distributing programming across a variety of media and with these changes comes a synergy, as a conglomerate uses its many venues to promote a certain media product. The second, closely related, change in the media environment comes from the production of messages to be disseminated across formats, from film to video to television to radio and print. Just as Disney promotes films and characters that are then presented in Disney World, Bakker used his television platform to promote his dream for a Christian values theme park. Colorado-based 'Focus on the Family' started with a radio broadcast by Dr. James Dobson and built an organization that now markets tapes, videos, magazines, and books to disseminate the organization's message not only in the USA but in multiple languages throughout the world. In addition, Big Idea Production's 'Veggie Tales' video series for children provides a marketing hook for the merchandising of toys, stuffed animals, workbooks, and tapes in the USA, Latin America and Saudi Arabia. Future studies of televangelism may find it increasingly difficult to address one television program or one televangelist in isolation; instead, work may need to focus on Christian media companies that produce a variety of formats for multiple audiences.

See also: Charisma and Charismatic; Charisma: Social Aspects of; Christianity: Evangelical, Revivalist, and Pentecostal; Media and Social Movements; Media Effects; Media, Uses of; Religion, Sociology of; Religiosity, Sociology of; Religious Fundamentalism: Cultural Concerns; Ritual; Television: Industry

Bibliography

Alexander B C 1997 Televangelism: Redressive ritual within a larger social drama. In: Hoover S M, Lundby K (eds.) *Rethinking Media, Religion, and Culture*. Sage, Thousand Oaks, CA, pp. 194–208

Armstrong B 1979 *The Electric Church*. Nelson, Nashville, TN

Asuzu B N 1987 *Communications Media in the Nigerian Church Today*. Tipografia Ugo Detti, Rome

Bruce S 1990 *Pray TV: Televangelism in America*. Routledge, London

Finke R, Stark R 1992 *The Churching of America: Winners and Losers in Our Religious Economy*. Rutgers University Press, New Brunswick, NJ

Frankl R 1987 *Televangelism: The Marketing of Popular Religion*. Southern Illinois University, Carbondale, IL

Hackett R 1998 Charismatic/pentecostal appropriation of media technologies in Nigeria and Ghana. *Journal of Religion in Africa* **3**: 258–77

Hadden J K 1993 The rise and fall of American televangelism. In: Roof W C (ed.) *Religion in the Nineties*. Sage, Newbury Park, CA
Hadden J K, Swann C E 1981 *Prime Time Preachers: The Rising Power of Televangelism*. Addison-Wesley, Reading, MA
Hoover S M 1988 *Mass Media Religion: The Social Sources of the Electronic Church*. Sage, Newbury Park, CA
Hoover S 1990 Ten myths about religious broadcasting. In: Abelman R, Hoover S (eds.) *Religius Television: Controversies and Conclusions*. Ablex, Norwood, pp. 23–39
Linderman A 1996 *The Reception of Religious Television: Social Semeiology (sic) Applied to an Empirical Case Study*. S. Academiae Upsaliensis, Uppsala, Sweden
Parker E C, Barry D W, Smythe D W 1955 *The Television Radio Audience and Religion*. Harper, New York
Roof W C 1997 Blurred boundaries: Religion and prime time television. In: Suman M (ed.) *Religion and Prime Time Television*. Praegar, Westport, CT, pp. 61–7
Schultze Q 1990 Defining the electronic church. In: Abelman R, Hoover S M (eds.) *Religious Television: Controversies and Conclusions*. Ablex, Norwood, NJ

H. Warren

Television: General

Television has been the dominant mass medium in most Western countries since the 1950s. It has been studied in a great variety of ways, in a number of scholarly disciplines, from psychology and the social sciences to film studies and art history. In the 1970s and 1980s, precisely because the medium invites cross- and interdisciplinary approaches, the study of television was central to the establishment of new forms and institutions of scholarly inquiry and teaching on the borders between the social sciences and the humanities, media studies and cultural studies. So far, there are very few if indeed any university departments exclusively devoted to television studies.

The study of television is, thus, still organized in accordance with the general divisions between scholarly disciplines and between 'schools' within (mass) communication and media studies. But especially since the early 1980s there has also been a trend toward a form of eclecticism where it is recognized that useful work is done from a number of starting points, employing a number of methodologies and a variety of theoretical frameworks.

1. *The Medium and Modernity*

Williams suggested in *Television: Technology and Cultural Form* (1975) an understanding of television that focused on its social organization as a broadcasting medium, arguing that there is a functional fit between this form of communication and the conditions of modern society. Processes of modernization led on the one hand to centralization of resources and power, on the other to increased social and geographical mobility. Williams termed this latter aspect of modernization 'mobile privatisation,' referring to the relative social isolation of individuals and nuclear families literally and metaphorically 'on the move.' Broadcasting was simply a highly adequate institution in this social situation. It makes immediate distribution of necessary information to all citizens possible, and efficiently contributes to the construction of the 'imagined community' (Anderson 1983) of nation–states.

Historical research has demonstrated that television technology in itself had other possibilities as well. It was already envisaged in the nineteenth century as a new form of two-way communication. Companies such as AT&T experimented with 'picture-phones' in the 1920s, and amateurs started constructing simple transmitters and receivers. Also, the idea of video surveillance was explored at this time (Allen J 1983). A more serious alternative to broadcast TV was television as a theatrical medium for collective, public reception. Nazi Germany developed such a system in the 1930s, with hundreds of public television halls throughout the country (Uricchio 1989). Most people who experienced the first telecasts of the BBC in that decade did so in restaurants, at railways stations, and the like (Corrigan 1990).

But already in the 1880s, there is evidence that popular imagination played with the idea of broadcast TV, that is, the transmission of sound and images from some central source directly to people's homes, organized in genres such as news, education, art/entertainment, and advertising (Wheen 1985). In the 1890s, Thomas Alva Edison and others were striving to come up with a successful solution for broadcasting TV (Sklar 1976, p. 11), and the idea circulated throughout the Western world. All of this lends support to the idea that broadcast television to private homes is related to fundamental social structures and interests in modern capitalist societies, such as the division between the domestic and the public realms on the one hand and the need for socially and culturally integrative efforts on the other. Furthermore, processes of democratisation could obviously also be served by broadcasting media since they allow for a more equal access to resources of information, knowledge, and cultural experiences.

The development of multichannel systems of broadcasting via satellites and cable systems since about 1980 may thus be thought of as 'divisive' since they constitute numerous separate viewer-communities within a given nation. So-called postmodernist theorizations of television have consequently seen multichannel television's abundance of programs on offer at any time as a manifestation of a new, presumably 'decentered' social and cultural situation (e.g., Collins 1989). Relatively little empirical work has been done on the larger social and cultural consequences of the transition from one or a few to dozens or even hundreds of television channels. But

most viewers still spend much or most of their viewing time with a handful of channels, and watch discreet programs as much as they just zap around. Hence the continued existence of 'hit shows' and newscasts competing over ratings. Most channels still stick to the same genres of programming, have fairly similar newscasts, and address the same or similar issues. It can, thus, be argued that television has retained much of the socially and culturally integrative functions ascribed to it by Williams. Clearly, though, these functions are modified, first, by multichannel satellite and cable TV, and, second, by the already ongoing digitalisation and convergence with (other) digital media.

2. *Policies*

The history of television policies starts with the attempts to regulate radio broadcasting in the 1920s. The fact that the number of available frequencies was finite led to forms of public regulation that are unheard of for other mass media. While most European countries decided to organize broadcasting institutions as state-owned monopolies, financed by a license-fee, the US from the beginning regarded broadcasting as another business opportunity, where profits would come from advertising revenues. The US also had to regulate broadcasting in ways that made it a quite particular area of private enterprise. Frequencies had to be allocated to broadcasting organizations by some public agency, operating in accordance with a set of politically determined principles. The first comprehensive broadcasting legislation in the US, the 1927 Radio Act, made clear that those who were licensed to use radio waves, were required to serve 'the public interest, convenience, or necessity' (Hoynes 1994, p. 38). The Federal Radio Commission, which in the 1930s became the Federal Communications Commission, was therefore established. It has ever since played a pivotal role in political struggles over the development of broadcasting in America, with agendas shifting as political administrations have come and gone. Key issues have been the ways in which competition is to be ensured with a necessarily limited number of actors in the market place, and whether or how space and resources are to be provided for noncommercial, public TV. Underlying both issues is the question of what the notion of the 'public interest' which broadcasting is to serve actually means.

Regulations and the forms of funding clearly influence program output and thus the medium's social functions. Williams (1975) demonstrated in a simple quantitative comparison the differences both between programming in commercial and noncommercial channels and between the television systems in the UK and the US. In the US, public channels were without broadly popular drama or other entertainment, while network TV was lacking documentaries, one-off dramas, educational programming, etc. The result was a polarized system where the majority of viewers would in practice not encounter certain types of programs that might be important to them as citizens. In the British 'duopoly' system the characteristic differences between commercial and public channels were also observable, but the differences were considerably less pronounced so that the programming menu available to all viewers was more varied.

The BBC generally functioned as a model for European public service television. The definition of public broadcasting's mission formulated by the BBC's first Director General, Lord Reith, was thus programmatic for most or all of these institutions—they were to provide information, education, and entertainment to the entire population at a price affordable to all. Some critical scholars, especially in the 1970s, perceived the BBC as an elitist institution basically affirming the existing cultural and social hegemony in the UK. However, when commercial satellite television arrived in the 1980s such critique partly gave way to a view that emphasized the democratic qualities of the public service television tradition. Since the mid-1980s, the threats to public service television posed by so-called deregulation of broadcasting and the expanding commercial offerings have motivated lots of research into the traditions and conditions of this form of broadcast TV all over Europe, including comparative studies that clarify considerable differences between the various national systems (e.g., Syvertsen 1992, Raboy n.d.).

The European Union now decides much of the premises for broadcasting policies in Europe even if this is in principle an area where member states are to retain national sovereignty. On the other hand, the EU has officially decided that public service broadcasting is to be protected. The European Council of Ministers (which includes the EU member states but also all other European countries) has done the same and also defined 'public service television' in a way that very clearly sets it apart from and in opposition to purely commercial services. Many of the pertinent issues in research on public service broadcasting and broadcasting regulation in general now has to be of a legal and economic nature. Most crucial, however, is the question of what happens when digitalisation does away with spectrum scarcity, the legitimation of traditional broadcasting regulation.

3. *Production*

Research on television production ranges from studies of political regulations, ownership, and other aspects of the economics of the television industry, to participant observation in broadcasting institutions and independent production companies to theoretical discussions of the notion of 'authorship' in the context of the necessarily collective production processes in television.

The so-called 'political economy' tradition in mass communication research has always focused on the importance of economic forces and related social structures. Prominent representatives of this tradition are Schiller (1989) in the US and Garnham (1990) in the UK, both regarding television as part of the larger media system of capitalist societies. Networks and production companies are parts of huge conglomerates encompassing not just media and leisure industries but also any number of other branches of business. These enormous concentrations of capital exert considerable political and cultural power, both domestically and internationally. British scholars of the 'political economy' tradition have, moreover, been exploring the implications of Jürgen Habermas's theory of the 'public sphere' for our understanding of television and the media in general (see *Mass Media and the Public Sphere*; *Mass Media, Political Economy*; *Information Society*).

In-depth studies of television production practices and personnel are not many. Cantor's *The Hollywood TV Producer: His Work and His Audience* (1971) is a classic in the field, a piece of organizational sociology charting the space for creativity in television and describing three types of television producers—the 'filmmakers,' the 'old-line producers,' and the 'writer–producers.' The often well-educated writer–producers appeared as the only group that both take the medium seriously and expect the audience to accept more ambitious shows with social 'messages.' Newcomb and Alley (1983) represent a more traditional humanistic view of textual production in their *The Producer's Medium*, based on interviews with individuals they claim have managed to leave their personal imprint on the series they have produced. Gitlin's *Inside Prime Time* (1983) is a study also of top-ranking executives which demonstrates not least the space for personal agendas and preferences when decisions about programming are made. Executives may use results of audience research that support their intuitive preferences while setting aside those that do not.

A few studies of production are combined with textual analyses of products. An early example is a study of the independent company MTM Enterprises, which contributed to a renewal of US television drama with shows such as *The Mary Tyler Moore Show* and *Hill Street Blues* (Feuer et al. 1984). In Britain, Tulloch and Alvarado (1983) did a similar study of the BBC serial *Doctor Who*. Buckingham (1987) and Gripsrud (1995) are examples of studies that also add an analysis of audiences to the study of production and textual analysis, thus covering the entire communicative process.

4. Programs

Williams (1975) was the first to name and point out the importance of the way in which television's (and radio's) programs are presented as parts of a continuous 'flow.' This means, according to Williams, that the impressions and meanings of individual programs will tend to overlap and merge in the viewer's experience. Flow remains a fundamental feature of broadcast programming. The ordering of programs in certain sequences, so-called scheduling, is of great importance to any television channel. It represents a channel identity and can be use to attract and not least keep viewers for the channel's programs and ads. As Ellis (1983) pointed out, most programs also consist of *segments*, that is, sequences of up to about 5–7 min that are more or less self-contained in terms of meaning. This structure results in part from its practical advantages from a production point of view, but it is also related to the need to make programs accessible to viewers who arrive in the middle of a program. The 'program' unit is thus in a sense a troubled category, threatened by dissolution both by the 'flow' and by internal segmentation.

The limits of 'the program' as an object of textual analysis have consequently been discussed, including the problem of separating the program text from the ads that are inserted into it and the question of the status of a single program or episode in soap operas that run an ongoing story for decades (Feuer 1989, Allen R 1985). In practice, viewers have few problems in identifying their favourite show and distinguishing it from the ads, and a soap opera (as well as newscasts and other long-running programs) may be analysed by way of a more or less representative sample of episodes (Gripsrud 1995).

It is probably correct to say that most textual analyses of television have been preoccupied with questions of representation and ideology, that is, the representation of gender and sexuality, of race and ethnicity, class, social conflicts, political issues, etc. and how they relate to larger ideological patterns and struggles. To these ends, all the theoretical, more or less philosophically grounded approaches in textual analysis have been applied in a variety of blends—semiotics, psychoanalysis, phenomenology, hermeneutics, etc. While fiction and entertainment have provided most of the material for such work, also newscasts, talkshows, and documentaries have been analyzed, often interestingly demonstrating the ways in which such 'reality programming' is actually constructed in accordance with certain textual traditions and principles rather than relating directly the structures of external reality.

The most prominent of social science traditions in the field of program analysis is so-called quantitative content analysis, which defines certain elements of content and then proceeds to count their occurrences, most often across large samples of programs. Typically, the elements counted would be violent acts, sexual innuendo, women in powerful positions, etc. Such research may provide overviews of developments over time in certain dimensions of programming or

suggest the extent to which certain portrayals of social categories or issues have a dominant position in the total program output or in a certain genre. A prominent example of large-scale quantitative content analysis of television programs is that conducted by Gerbner and co-workers in the late 1960s and early 1970s, the results of which were central to the formulation of the so-called 'cultivation theory,' that is, a theory of television's long-term consequences for people's world view (Gerbner 1973). Their analyses indicated that (US) television, across genres and individual programs, presented a systematically distorted view of social realities in a number of areas such as family life, work and social roles, aging and dying, crime, and violence. Correlating content analysis with opinion surveys then seemed to suggest that viewers, especially heavy viewers, accepted this distorted view of social conditions as true.

5. *Audiences*

In line with the general history of research on modern mass media, worries over how the new medium affected especially children and adolescents were important starting points for the first studies of television in the 1950s. For example, Maccoby (1951) concluded that the medium undermined communication within the family, even if it made family members spend more time together. The same researcher later tried to show that television viewing had a negative 'escapist' function for children who were lacking real-life satisfactions (Maccoby 1951). Studies such as these were fraught with both conceptual and methodological problems, and marked by a prejudiced, negative attitude to the medium. Any viewing of dramatic fiction would for instance tend to be counted as 'escapist,' reflecting a somewhat shallow understanding of the uses of fantasy and art in general. Around 1960, large empirical projects both in the UK and in the USA tried to estimate the social and cultural implications, especially for children, of the nationwide introduction of television (Himmelweit et al. 1958, Schramm et al. 1961). The everyday lives of children in communities with and without television were charted in surveys and compared. The findings of these studies were quite complex and did not give a totally negative view of the medium's social and cultural role.

The above mentioned 'cultivation theory'—or, rather, hypothesis—is basically in line with critical theories of the structural, long-term power of the media to define social reality in ways that harm various disadvantaged social groups. Throughout the 1970s such perspectives on television were theorized and significantly renewed in work conducted by researchers at the Centre for Contemporary Cultural Studies at the University of Birmingham, UK. Hall (1973) applied concepts from semiotic theory to the process of communication, speaking of production as a moment of 'encoding' and reception as a moment of 'decoding,' thus theoretically demonstrating the possibility of complications in the 'transfer of messages' that were not simply due to 'noise' or misunderstandings. Hall suggested three main types of 'readings' of television programs—'dominant,' 'negotiated,' and 'oppositional,' that is, acceptance of the program's 'message,' partial agreement, and critical rejection. An empirical study of groups of viewers conducted by Morley and Brunsdon (Morley 1980) demonstrated how social class and political convictions might shape audience attitudes to television programs, making some groups more critical than others. In the wake of this study, audience research within or close to the British cultural studies tradition has repeatedly demonstrated how varied interpretations and evaluations of programs might be, both within and between social classes, races, genders, and nations. They have also described how television viewing is inscribed in people's everyday lives and the routines and power struggles that prevail there (e.g., Morley 1986, Seiter et al. 1989), how this came about historically in a wider social and cultural context (Spigel 1992) and, somewhat more abstractly, how television viewing has deeper social functions related to its psychological functions for viewers (Silverstone 1994, Ellis 1999) (see *Audience*).

6. *Medium*

Much scholarly writing and various sorts of research may be categorized as contributions to an understanding of television's specificity as a medium in terms of the ways in which it represents reality outside of itself and the ways in which it relates to its audiences. One of the first contributions of lasting significance on the latter theme was the article by social psychologists Horton and Wohl (1956) in which the authors coin the term 'para-social interaction' in a discussion of the 'intimacy at a distance' characteristic of the relations between viewers and television characters such as talk-show hosts. The term refers to the way in which viewers tend to establish pseudo-dialogical relations with television characters, 'talking back' to them, etc. These practices set television apart both from film and radio, and reflect the particular type of intimacy associated with this domestic audio-visual medium. They can, besides the domestic location of sets, be tied to the normal size of the screen, which renders close-ups of human faces close in size to those of interlocutors in the same room. Importantly, television's 'talking heads' often address viewers directly ('you'), simulating real conversation also in its preferred styles of speech. These features are shared by radio but have an even stronger effect when combined with moving images.

Furthering this line of thought, Meyrowitz (1985) argues that television has contributed significantly to a series of important changes both in the public sphere

and in the relations between people in general. Television has moved politicians and other public figures closer to the rest of the population by presenting them in close-ups and settings that make them appear more 'ordinary.' A new way of appearing in public has thus developed, in which the public/private distinction has become blurred. Meyrowitz terms this 'middle-stage' behavior, referring to Goffman's (1959) distinction between a 'front stage' and a 'back stage' set of conventions for self presentation. This has radically altered the conditions for all sorts of rhetorical practices. Moreover, television programs have provided children with insights into the world of grown-ups and boys and girls now know more about each other's intimate lives. For better or worse, television has thus reduced distances between various social categories and weakened certain hierarchical and authoritarian structures.

The capacity for simultaneity between a real event and its transmission and reception as audiovisual representation, that is, 'liveness,' is central to television's specificity as a medium and its social role. It shares a form of distribution (broadcasting) with radio, it shares subject matter also with newspapers, magazines, and books, and it shares the audiovisual form of representation with film. Early television was all 'live,' but now nearly all television programming is prerecorded. Television producers and broadcasters actively seek the impression of immediacy or 'liveness'; it is simply a key aesthetic value in television. This is related to what Barthes (1981) termed the 'photograph effect': when we see a photograph we (used to) know that what is in the picture has actually once been in front of a camera. Television wants this, and more: not just 'this really happened,' but 'this really happens, right now!' Studio debates, interviews, and variety shows will either be live—or try to appear to be. Pretaped shows like sitcoms will often inform us that they are 'taped in front of a live studio audience.' Liveness is particularly important to newscasts, since 'news' as a genre is based on getting as close to immediacy as possible. This is why reporters often talk to the camera—'live,' 'on the spot'—at a dark place where nothing is happening (anymore). What is at work here is an equation of 'live' with 'real': liveness means reality, or *truth*.

A medium that can give us 'reality in the raw,' unfolding as it happens, cannot lie, it would seem, and this ability tends to lend a particular quality of immediacy, realism, or truth to much more than live programming (cf. Heath and Skirrow 1977). 'Liveness' as an aesthetic value is, therefore, still important even if new technologies and styles of presentation progressively makes the medium *less* 'live' (cf. Feuer 1983). Caldwell (1995) has argued that television is more and more becoming a medium not about outside 'events' but about itself and its visual style. But there is still much live programming, and liveness as an effect remains important for the above reasons.

'Liveness' is furthermore a key to television's ability to establish experiential communities. It strengthens the general experience of collectivity in the act of viewing (cf. Rath 1989). Televised national events manifest and support national identity, and so the question is whether or to which extent globally televised events—rock concerts, Olympic games, royal weddings, or funerals—can contribute to the establishment of new sense of transnational community. Dayan and Katz (1992) argue that television's live broadcasts of major events function as rituals that not only reinforce existing symbolical orders but also create new ones. This is just one of the ways in which television continues to be of great political, social, and cultural importance (see *Media Events*).

At the start of a new millennium, television as we know it is, according to many, about to disappear. Digitalization leads to convergence with any and all other media of communication. The distributive model of broadcasting may appear to be replaced by that of the World Wide Web, that is, a system without designated centers, presenting unlimited choice of contents to consumers. These consumers must now be thought of as active individuals and hence the term 'audience' is no longer adequate.

It may, however, also be argued that such a total revolution is envisaged mostly on the basis of technological possibilities and that the outcome of an undoubtedly comprehensive process will be modified by social and cultural forces and interests that may sustain the further existence of broadcast television. Total fragmentation and individualization will be restrained by the need for community-forming, shared media experiences and the need for competent, reliable editorial services or guidance in the confusing mass of contents on offer. Commercial interests will no doubt seek to exploit such needs and strive to establish control of gateways to and routes within the new landscapes of electronic information and entertainment. Governments may seek to find ways of securing the production of identity and distribution of cultural resources represented by the classic broadcasting model, particularly in its public service versions. Historical experience with introductions of new media indicates that already existing media survive alongside the newcomers. Theater was not killed by film; film was not killed by television. Even radio survived television, and books and newspapers are still around.

The dynamics of the ongoing process present television research with a number of exciting, difficult tasks, even if the profession of prophets should not be confused with that of serious scholars. The situation calls for studies of the relations between technological developments, economic forces, sociocultural structures, and, not least, the space for political interventions. Such studies could well be motivated by an interest in how democracy is served by various media systems and structures. The link between theories of democracy and media is generally in need of sustained,

philosophically qualified reflection. The process of convergence and increased interactivity also calls for in-depth analyses of the uses and functions of new and old versions of television, not least the ways in which new and old forms relate to each other. Finally, digitalization opens a whole set of new aesthetic and rhetorical possibilities; new forms of representation, new ways of telling stories, new ways of persuading consumers and citizens. While all of these new areas of inquiry will no doubt require new theoretical and methodological tools, the experiences of some 60 years of television research are well worth keeping in mind. Re-inventing the wheel remains a waste of time.

See also: Audiences; British Cultural Studies; Broadcasting: Regulation; Film and Video Industry; Mass Media, Political Economy of; Mass Media, Representations in; Televangelism; Television: Genres; Television: History; Television: Industry

Bibliography

Allen J 1983 The social matrix of television: Invention in the United States. In: Kaplan E A (ed.) *Regarding Television: Critical Approaches—An Anthology*. University Publications of America, The American Film Institute, Los Angeles, pp. 109–19

Allen R 1985 *Speaking of Soap Operas*. University of North Carolina Press, Chapel Hill, NC

Anderson B 1983 *Imagined Communities: Reflections on the Origins and Spread of Nationalism*. Verso, London

Barthes R 1981 *Camera Lucida: Reflections on Photography*. Hill and Wang, New York

Buckingham D 1987 *Public Secrets: 'East Enders' and its Audience*. British Film Institute, London

Caldwell J T 1995 *Televisuality: Style, Crisis and Authority in American Television*. Rutgers University Press, New Brunswick, NJ

Cantor M 1971 *The Hollywood TV Producer: His Work and His Audience*. Basic Books, New York

Collins J 1989 Watching ourselves watch television, or who's your agent. *Cultural Studies* **3**(3): 261–81

Corrigan P 1990 On the difficulty of being sociological (historical, materialist) in the study of television: The 'moment' of English television, 1936–1939. In: Syvertsen T (ed.) *1992 and After: Nordic Television in Transition, Report no.10*. University of Bergen, Norway, pp. 130–60

Dayan D, Katz E 1992 *Media Events: The Live Broadcasting of History*. Harvard University Press, Cambridge, MA

Ellis J 1982 *Visible Fictions: Cinema, Television, Video*. Routledge and Kegan Paul, London

Ellis J 1999 Television as working through. In: Gripsrud J (ed.) *Television and Common Knowledge*. Routledge, London, pp. 55–70

Feuer J 1983 The concept of live television: Ontology as ideology. In: Kaplan E A (ed.) *Regarding Television: Critical Approaches—An Anthology*. University Publications of America, The American Film Institute, Los Angeles, pp. 12–22

Feuer J 1989 Reading dynasty: Television and reception theory. *South Atlantic Quarterly* **88**(2): 443–60

Feuer J, Kerr P, Vahimagi T (eds.) 1984 *MTM: 'Quality Television.'* British Film Institute, London

Garnham N 1986 The media and the public sphere. In: Golding P, Murdock G, Schlesinger P (eds.) *Communicating Politics. Mass Communication and the Political Process*. Leicester University Press, Leicester, UK

Garnham N 1990 *Capitalism and Communication. Global Culture and the Economics of Information*. Sage, Newbury Park, CA

Gerbner G 1973 Cultural indicators—the third voice. In: Gerbner G, Gross L, Melody W (eds.) *Communications Technology and Social Policy*. Wiley, New York, pp. 553–73

Gitlin T 1983 *Inside Prime Time*, 1st edn. Pantheon Books, New York

Goffman E 1959 *The Presentation of Self in Everyday Life*. Anchor, New York

Gripsrud J 1995 *The Dynasty Years: Hollywood Television and Critical Media Studies*. Routledge, London and New York

Hall S 1973 *Encoding and Decoding in the Television Discourse*. Stencilled Occasional Paper no. 7, CCCS, Birmingham

Heath S, Skirrow G 1977 Television: A world in action. *Screen* **18**(2): 7–59

Himmelweit H T, Vince P, Oppenheim A N 1958 *Television and the Child*. Oxford University Press, London

Horton D, Wohl R 1956 Mass communication and para-social interaction. Observations on intimacy at a distance. *Psychiatry* **19**(3): 215–29

Hoynes W 1994 *Public Television for Sale: Media, the Market, and the Public Sphere*. Westview Press, Boulder, CO

Maccoby E 1951 Television: Its impact on schoolchildren. *Public Opinion Quarterly* **15**(3): 421–44

Meyrowitz J 1985 *No Sense of Place: The Impact of Electronic Media on Social Behavior*. Oxford University Press, New York

Morley D 1980 *The 'Nationwide' Audience*. British Film Institute, London

Morley D 1986 *Family Television: Cultural Power and Domestic Leisure*. Comedia, London

Newcomb H, Alley R S 1983 *The Producer's Medium*. Oxford University Press, New York

Raboy M n.d. *Public Broadcasting for the 21st Century*. Acamedia Research Monograph no. 17, John Libbey Media, Luton, UK

Rath C D 1989 Live television and its audiences: Challenges of media reality. In: Seiter E, Borchers H, Kreutzner G, Warth E-M (eds.) *Remote Control: Television, Audiences and Cultural Power*. Routledge, London, pp. 79–95

Schiller H I 1989 *Culture Inc.: The Corporate Takeover of Public Expression*. Oxford University Press, New York

Schramm W, Lyle J, Parker E B 1961 *Television in the Lives of Our Children*. Stanford University Press, Stanford, CA

Seiter E, Borchers H, Kreutzner G, Warth E-M 1989 Don't treat us like we're so stupid and naïve: Towards an ethnography of soap opera viewers. In: Seiter E, Borchers H, Kreutzner G, Warth E-M (eds.) *Remote Control: Television, Audiences and Cultural Power*. Routledge, London, pp. 223–47

Silverstone R 1994 *Television and Everyday Life*. Routledge, London and New York

Sklar R 1976 *Movie-made America: A Cultural History of American Movies*. Vintage Books, New York

Spigel L 1992 *Make Room for TV: Television and the Family Ideal in Post-war America*. Chicago University Press, Chicago

Syvertsen T 1992 *Public Television in Transition*. Ph.D. thesis, University of Leicester

Tulloch J, Alvarado M 1983 *'Dr Who': The Unfolding Text*. St Martin's Press, New York

Uricchio W 1989 Rituals of reception, patterns of neglect: Nazi television and its postwar representation. *Wide Angle* **11**(1): 48–66
Wheen F 1985 *Television*. Century Publishing, London
Williams R 1975 *Television: Technology and Cultural Form*. Schocken, New York

J. Gripsrud

Television: Genres

The most basic definition of genre is a 'type,' 'kind,' or 'sort', and when applied to television, it refers to specific kinds or categories of programming. The categories may distinguish broad program types such as fiction and nonfiction, or may create narrower delineations such as the division of fictional programs into soap operas, cop shows, sitcoms, and hospital dramas, and non-fictional ones into sports, news, religious shows, talk shows, and so on.

Genres forge a strong link between the television audience and the television industry, a direct connection between the consumption and the production of programs. For a commercial television system, such as that of the USA, this link is of the utmost importance.

The phrase 'television genres' has unique meanings and ramifications for different television systems and audience members in different areas of the globe, but the general notion of television genres has had a profound effect on structuring television the world over. Many television systems were born in the throes of enormous struggles over what types of programs would constitute the broadcasting of the nation. Would they be educational and informative or predominantly entertaining? Would they be religious or secular? Would they be high culture or popular? Would they be nationally produced or imported?

Even after the establishment of national television systems, heated debates over program types often emerged in establishing regulatory policies and shaping the future of the system. If systems were commercial, such as that of the USA, what would be the place of news, educational, or politically oppositional programming? If systems were state operated, such as those of many European, Middle Eastern, and Asian nations, what would be the place of entertainment, imported programs, and shows that might run counter to state policies or predominant religious beliefs?

Every national television system and their audiences need to be studied in great detail to determine exactly how genres have worked in their histories. A closer look at US television will detail some of the major ways that genres have participated in defining it along particular lines, in forging its commercial character, and in fashioning the culture of its viewers and nation. Because American television is currently imported in numerous countries of the world, a closer look at the functioning of genres in its operations will aid in understanding a great deal of television programming in circulation today, particularly from the point of view of its production. Such a look should also bring into focus various aspects of genre that can be compared and contrasted with the programming of other television systems in other areas of the globe.

1. Genre and the US Television Industry—A Historical Sketch

1.1 The Transition from Radio to Early Television

From the first flickers of the first cathode ray tubes in the living rooms of postwar America, the principle of genre was already at work in trying to create an audience for the new medium. By the early 1950s many popular radio shows of various generic categories appeared in television incarnations, including the western, *The Lone Ranger*, the quiz show, *Twenty Questions*, the soap opera, *The Guiding Light*, and the domestic comedies, *Beulah*, *The Goldbergs*, and *The Life of Riley*, to name only a few. This was because television inherited many of its programming and operating principles from US radio, which had developed over the previous 30 years into a thoroughgoing commercial system, dependent on advertiser dollars and governed by a ratings system for judging the popularity of shows.

Genres had become a major force in the economic workings of US broadcasting. Throughout their histories, the radio networks (NBC, CBS, and later ABC) and the television networks (NBC, CBS, ABC, and the short-lived DuMont) gradually sought to determine, through ratings mechanisms and statistical sampling (diaries kept by families 'scientifically' selected by ratings companies, and later by a combination of diaries and meters hooked up to home receivers), how many listeners or viewers were tuned in to what programs; to catalog specific listener or viewer demographics (primarily according to age and gender); and to rate programs based on which ones attracted the most audience members. The implications of this phenomenon were more far-ranging than they may at first appear: the programs (and consequently their genres) that attracted the most audience members of the right ages (those aged 18–49 generally came to be the most valued) became the ones that the broadcasting industry wanted to produce and reproduce; programs and genres that seemed popular because of their ratings were reproduced or cloned, while others fell into oblivion.

This dynamic has endured throughout the history of US television. Specific genres have proliferated at different periods (westerns in the 1950s and 1960s, detectives and spies in the 1960s, 'jiggle' shows in the 1970s, prime-time soaps in the 1980s, situation com-

edies in the 1990s, reality programs in the 2000s, and so forth) because television networks could minimize financial risk by renewing contracts for shows with high ratings and choosing new pilot programs that fell within the tried and true category of proven ratings winners. The recoiling from innovation by sticking with highly rated genres became one of the hallmarks of US television.

It took until the late 1950s for genre to achieve its full-blown status in the functioning of the television industry. Two other factors, the influence of the New York theater and the technology of live television, had significant impacts on the emerging television medium in the late 1940s and early 1950s. Each of these factors themselves had strong implications for the development of genres on television.

1.2 The Genre of Anthology Dramas

In the late 1940s and early 1950s, many New York playwrights and directors worked in the emergent medium of TV to produce the spate of anthology dramas that have come to be equated with the phrase of 'Golden Age' of American television. These dramas, such as the well-known *Requiem for a Heavyweight* by Rod Serling, *Marty* by Paddy Chayefsky, and *Thunder on Sycamore Street* and *Twelve Angry Men* by Reginald Rose, were aired on shows sponsored by individual advertisers and called anthology drama series, shows such as *Playhouse 90*, *Goodyear Television Playhouse*, *Studio One*, *Kraft Television Theater*, and *US Steel Hour*. Hundreds of these dramas or teleplays were produced during the 1950s—original plays every week for the numerous anthology programs on the air at the time. They were also produced live, on the sets of the three-camera television studios that came to epitomize one of the thrills of early television—viewers in their own living rooms seeing what was simultaneously happening in a distant space.

Although other programs and genres, some of them not live but shot on film such as the hugely popular comedy *I Love Lucy*, were attracting large groups of viewers, the anthology dramas constituted a genre of enormous critical and audience acclaim. By the middle of the 1950s, however, the anthology drama genre and live prime-time fictional programming as a whole was virtually obliterated.

As commercial enterprises, NBC, CBS, the young ABC, and corporate television sponsors struggled in the early 1950s to rationalize the television industry, to get it working in ways most profitable for their companies. This militated against anthology dramas in four ways. First, the dramas were too costly and too cumbersome to produce: each one required a new director and a new playwright who needed to conceive of an entirely new play. Second, the networks and advertisers began to consider the teleplays too controversial for national television, dealing as they often attempted to (though frequently muted by censorship) with serious social problems. Third, sponsors began to think of these dramas as non-conducive to advertising—their serious subject-matter was often considered too serious and their unglamorous, working-class characters too jarring to the spirit of consumption and the showcasing of products. Fourth, the live productions were not oriented toward future profits—they could not be rerun or exported.

1.3 Formulaic Filmed Genres

These business imperatives, which became more sharply formulated by the industry as the experimental years of the 1950s waned, led to the standardization of generic formulas and a factory-like system in American Television. Most entertainment programs, whether they were westerns, hospital dramas, police programs, situation or domestic comedies, came to develop narrative formulas with well-defined story premises and possibilities. (Even quiz shows, game shows, variety shows, newscasts, and sports programming gradually developed similar formulaic structures.) The programs had standard and repeatable characters (or hosts, reporters, anchors, and announcers), sets, wardrobes, and shot compositions. They were overwhelmingly shot on film (with the exception of news, sports, and daytime soaps) for maximal use in the rerun, syndication, and export markets. Equations were also made between different desired audience groups and different genres: men were drawn in with westerns, spy shows, cop shows, sports; women with comedies, hospital drama, and, of course, daytime soap operas; both men and women with quiz shows, variety shows, and game shows, and so forth. Filmed (and, from the 1970s on, some videotaped) formulaic genres thus came to serve as linchpins of the factory-like production process that dominated Network Era American television. They provided a way to control production costs and predict consumption patterns.

One popular genre, the quiz show, allowed the networks in the late 1950s to complete their corporate rationalization of the TV business by wresting the control of programming away from individual corporate sponsors—when a national scandal burst from the discovery that the producers of corporate sponsored popular quiz shows secretly provided answers to contestants, the networks argued that they (rather than the individual corporations who owned and sponsored the shows) were the only ones that could ensure the integrity of program content. NBC, CBS, and ABC effected a switch from a system in which individual companies sponsored individual programs (and therefore had the final say over a program's content, renewal, or cancellation) to a system in which individual companies could only buy advertising spots before, after, or in the middle of programs controlled

by the networks. From the later 1950s forward, therefore, the three major television networks could produce and reproduce formulaic genres in an assembly-line way with total control over their content, their longevity, and their cloning. Standardized genres came to have a powerful place in the operations of the entire American television system and its popular culture.

2. Genre and Television Audiences

2.1 Genres and Audience Pleasure

During the past 15 years, scholars have begun to analyse the reception by audience members of the type of delimited TV fare that the US system has, in fact, offered; and they have produced many intriguing studies of specific genres and their most avid viewers or fans. In his book *Television Culture*, John Fiske, the cultural theorist, examines the place of many different genres from soap operas to action adventure programs to game shows in the lives, fantasies, and pleasure domains of their many viewers. He strongly advocates (and has been criticized for taking too extreme of a position) looking at the ways viewers make use of television programs for their own enjoyment and even political empowerment, rather than focusing on the relationship of TV programs and audiences to the industry that produces and profits from them. Other scholars have taken a similar approach.

Henry Jenkins (1992), Constance Penley (1991), and Camille Bacon-Smith (1992), for example, have intensively studied fans of the science fiction genre, particularly those of *Star Trek*. They show how these fans take an industrially produced formulaic television genre and turn it to their own ends, making it part of their own lives and creative endeavors. Jenkins, for instance, chronicles the passionate investments of fans in the culture surrounding *Star Trek* conventions. Penley details the ways a group of women viewers rewrite *Star Trek* narratives with their own unique plots for their own illustrated fanzines. Jenkins' notion of 'textual poaching' has, in fact, been taken up by a number of cultural critics to describe the ways that viewers of commercially produced genres can 'poach' or rifle through these programs (programs produced by a capitalist culture industry) in order to derive pleasures that may have nothing to do with the intentions of the television networks and advertisers to generate audiences as consumers.

2.2 Genres and Society—Gender, Race, Sexuality, Nation, and the Family

Various facets of contemporary US society and its citizenry have been examined in relation to television genres. How, for example, do genres contribute to defining femininity and masculinity, race, ethnicity, sexuality, nationhood, and notions of the family? How, in other words, do genres contribute to shaping the norms of those social facets—to defining what 'normal' femininity or masculinity means, what it means to be white or black in a society such as the US? How, on the other hand, might viewers use TV genres to question or challenge dominant societal norms?

Scholars such as Tania Modleski (1984), Robert Allen (1985), and Dorothy Hobson (1982) analyse the ways that many women viewers (particularly those caring for small children in the home) have made soap opera an integral part of their everyday lives, pleasures, and friendship networks. Modleski elaborates upon ways that women viewers derive sometimes unpredictable meanings and pleasures from the soaps. She argues, for example, that some women viewers identify with the ubiquitous villainess characters in soap operas as a way of getting fictional revenge on a social system that minimizes women's power, work, and social contributions. Scholars studying sports programming are currently trying to analyze the ways that sports produce meanings of conventional (macho) masculinity but at the same time may engender in their male viewers nonconventional meanings of men being emotionally, tied to other men in overt displays of friendship, and so forth.

Julie D'Acci (1994), in a study of the popular crime drama *Cagney and Lacey*, demonstrates how TV genres contribute to the ways that masculinity and femininity are defined and actually lived out in society. This cop show, starring two women, became a women's program—a hybrid of cop show, soap opera, and comedy—in order to attract an audience of working women with the kind of genre fare TV executives thought such an audience wanted to see. The show also became part of a social movement for women's rights and equality, and an important rallying point for its many committed women viewers. Szu-ping Lin (2000) uses a similar approach in studying the importance of a Taiwanese genre called *Hsiang-Tu Hsi*. These nightly run melodramas have become an integral part of wide-ranging public debates on women's issues and rights in contemporary Taiwan.

In US television, black characters and stories have historically been confined to the genre of situation comedy and denied the range of representation and seriousness that dramatic shows can afford. Herman Gray (1995) has demonstrated how one of the most innovative programs on American television, *Frank's Place* (about an African American college professor who inherits his father's New Orleans restaurant), met with quick cancellation because it was genre transgressive, combining elements of the comic and the dramatic, a transgression that the TV industry found untenable given its history of corralling black characters and subject-matter in the comic terrain. How-

ever, Crystal Brent Zook (1999) chronicles the importance of African American writers and production teams in producing comedies for the FOX network that, although working within the confines of the stereotypical genre, deal in more progressive ways with black stories and characters. Tricia Rose (1997), for her part, argues that the music video genre has given voice to African American women artists and a sense of empowerment to African American women viewers. Gloria Abernathy-Lear (1992) illustrates the ways some African American women viewers make use of the primarily white-oriented daytime soap operas to, among other things, critique and lay bare the shortcomings of a dysfunctional white-dominated society.

From the point of view of sexuality, Ronald Becker (1998) has examined the spate of gay-themed comedies in the 1990s, discussing how these programs contribute to ongoing public debates about sexuality, and help explore the boundaries of sexual identity for their gay, lesbian, and heterosexual viewers. And from the point of view of nationalism, Michael Kachman (1999) has demonstrated how the television spy genre participated in defining 'America' and the American citizen during the Cold War along particular xenophobic lines.

Finally, researchers such as David Morley in *Family Television: Cultural Power and Domestic Leisure* and Ellen Seiter in *Television and New Media Audiences* have conducted ethnographic studies of TV viewers to show how TV programs affect both family life and children. Morley profiles, for example, how different genres become grounds of power struggle in the daily life of some British families, with possession of the remote control falling along conventional gender based lines—male 'heads' of households have the clicker and consequently the power to decide what genres the family will watch and when. (See *Audiences*.)

The work cited here represents only the tip of the iceberg of studies on the relationship between popular commercial television genres and social life in the world beyond the home screen. Much more research is already in print and under way. The importance of television genres as definers of societal norms and as rallying points for viewers becomes more and more evident.

3. *Television Genres—Theoretical Considerations*

In an influential article on television genre, media scholar Jane Feuer groups most theoretical work on TV genre into three different camps—the aesthetic, the ritual, and the ideological (Feuer 1992). According to her schema, the aesthetic approach includes attempts to define genre solely in terms of the artistic or formal components of a television program or 'text' (apart from its social or historical context), and particularly with deference to the 'author' of that text. The ritual approach, on the other hand, sees genre as intimately tied to its social context. It conceives of genre (and all television) as a 'cultural forum' that helps maintain the social order and aids it in adapting to change (see Newcomb and Hirsch 1983). Genres serve an important, though tacit, cultural function, in that they feed back to viewers important meanings about the culture at large, and thus become 'an exchange between industry and audience, an exchange through which a culture speaks to itself' (Feuer 1992, p. 145). According to this take, some workplace sitcoms of the 1970s (*The Mary Tyler Moore Show* and *Alice*, for example) aid viewers in adapting to the social change involving women's massive entry into the workforce at the time.

In contrast, the ideological approach 'views genre as an instrument of control.' Thus at the level of the television industry genres assure an audience for the networks and advertisers. At the level of the programs themselves genres serve to reproduce the ideology of the capitalist system. At the level of the audience, familiar genres help make it seem that the ideologies expressed in the programs are just plain common sense, the way things naturally are (rather than capitalist-based consumer-creating ideologies) (p. 145).

Feuer goes on to describe the ways that television scholars have begun to combine the ritual and ideological approaches to account for the active production of meanings by the viewers as well as the TV industry, and therefore the constant conflict (rather than the ease) in the reproduction of ideology. She also contends that such a combination has allowed TV and film critics to conclude that by means of genres, audiences unconsciously confront and grapple with many societal contradictions such as law and order vs. individuality (the gangster genre), nature vs. culture (the western), the work ethic vs. the pleasure principle (the musical), and so forth.

Rick Altman (1984) also incorporates a ritual and ideological approach when he uses a linguistic model to conclude that genres are comprised of 'semantic elements' (standard generic components such as cowboys, horses, guns, saloons, for a western) and a 'syntax' (the crystallized generic meaning of a western such as a white male hero as the 'necessary' bridge between nature and culture). This syntax develops over time and may express both a dominant ideology and the changes and challenges that bubble up in a culture. Many scholars stress that no matter which of these positions one espouses, genres develop historically and should be understood as fluid and changing rather than as fixed entities.

These various approaches are not necessarily incompatible. An 'integrated approach,' as exemplified here, conceives genre as operating at the intersection of the television industry, the audience, the text, and

the social context. Jason Mittell (2001) takes the integrated approach further and argues that most previous scholarly work on genre (such as that just recapitulated) proceeds too much from what he calls a 'textualist assumption'—an assumption that places too much emphasis on genre as equivalent to the program itself, as essentially a description of a particular kind of TV text. On the contrary, he argues, instead of examining genres as textual attributes, they can best be understood as cultural categories, belonging to specific historical moments and contexts.

For Mittell, genres are not comprised of the programs that they categorize, but by the practices of industries and audiences that use generic categories to define television programming. Thus when we want to study sitcoms, he says that instead of confining our analysis to the examination of sitcom texts (individual programs over time), we should explore how television networks schedule sitcoms, how audience members define their own tastes in terms of program choices, and how critics use particular standards to judge the genre—all strategies of definition, interpretation, and evaluation that link the sitcom to cultural norms, hierarchies, and relations of power. This shift in analytic focus looks beyond media texts themselves to understand their operation in the culture at large, and parallels a reconfiguration of genre within literary and film studies as well.

4. *Television Genre and the Future*

Since the break-up of the three-network domination over US television and the proliferation of cable channels that began in the mid-1980s, genres are playing a somewhat different role in American commercial culture. Now entire channels are founded on their generic status—a Sci Fi channel, a Romance channel, a Golf channel, a Comedy channel, a Cooking channel, and so forth. The confluence of television and the internet in the coming years, and the possibility of completely individualized regimes of viewing, similarly present many more possibilities for the interaction of genres and viewers' everyday lives. Likewise, the creation of global television markets and communities only underscores the importance of looking more closely at the way genres operate cross-nationally and cross-culturally. Robert Allen's *To Be Continued: Soap Operas Around the World* (1995) furnishes an exemplary model of this type of work.

See also: Consumer Culture; Consumer Psychology; Consumption, Sociology of; Culture, Sociology of; Entertainment; Film: Genres and Genre Theory; Film: History; Mass Media and Cultural Identity; Media and Child Development; Media, Uses of; Television: History

Bibliography

Abernathy-Lear G 1992 African Americans' relationship with daytime serials. Ph.D. dissertation, University of Wisconsin-Madison

Allen R 1985 *Speaking of Soap Opera*. University of North Carolina Press, Chapel Hill, NC

Allen R 1995 *To Be Continued: Soap Operas Around the World*. Routledge, New York

Altman R 1984 'A semantic/syntactic approach to film genres.' *Cinema Journal* **23**(3): 6–18

Bacon-Smith C 1992 *Enterprising Women: Television, Fandom and the Creation of Popular Myth*. University of Pennsylvania Press, Philadelphia, PA

Becker R 1998 Prime time television in the gay nineties: Network television, quality audiences and gay politics. *Velvet Light Trap*, **42**: 36–47

D'Acci J 1994 *Defining Women: Television and the Case of Cagney and Lacey*. University of North Carolina Press, Chapel Hill, NC

Feuer J 1992 Genre study and television. In: Allen R (ed.) *Channels of Discourse*. University of North Carolina Press, Chapel Hill, NC, pp. 138–60

Feuer J, Kerr P, Vasmagi T (eds.) 1984 *MTM: Quality Programming*. BFI, London

Fiske J 1987 *Television Culture*. Methuen, New York

Gledhill C 1997 Gender and genre. In: Hall S (ed.) *Representation*. Sage, London, pp. 339–86

Gray H 1995 *Watching Race: Television and the Struggle for 'Blackness.'* University of Minnesota Press, Minneapolis, pp. 113–29

Hobson D 1982 *Crossroads: The Drama of a Soap Opera*. Methuen, London

Jenkins H 1992 *Textual Poachers: Television Fans and Participatory Culture*. Routledge, New York

Kachman M 1999 Secret agents, civil subjects: Espionage, television and cold war nationalism. Ph.D. dissertation, University of Wisconsin-Madison

Lin S 2000 On prime time television drama and Taiwanese women. Ph.D. dissertation, University of Wisconsin-Madison

Mittell J 2001 A cultural approach to television genre theory. *Cinema Journal* **40**(3): forthcoming

Modleski T 1984 *Loving with a Vengeance: Mass Produced Fantasies for Women*. Methuen, London

Morley D 1986 *Family Television: Cultural Power and Domestic Leisure*. Comedia, London

Newcomb H, Hirsch P M 1983 Television as a cultural forum: Implications for research. *Quarterly Review of Film Studies* **8**(3): 45–55

Oren T 1999 A clenched fist and an open hand: The formation of Israeli television. Ph.D. dissertation, University of Wisconsin-Madison

Penley C 1991 Brownian motion: women, tactics, and technology. In: Penley C, Ross (eds.) *Technoculture*. University of Minnesota Press, Minneapolis, MN, pp. 135–61

Rose T 1997 Never trust a big butt and a smile. In: Brunsdon, D'Acci, Spigel (eds.) *Feminist Television Criticism*. Oxford University Press, Oxford, UK, pp. 300–17

Seiter E 1999 *Television and New Media Audiences*. Oxford University Press, Oxford, UK

Zook C B 1999 *Color by FOX: The FOX Network and the Revolution in Black Television*. Oxford University Press, New York

J. D'Acci

15578

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Television: History

Forecast in novels and science journals since the late nineteenth century, television came to the USA and other developed nations in the decade and a half following World War II. Major manufacturers of radio receivers, led by Radio Corporation of America (RCA), as well as individual entrepreneurs had labored in the 1930s in the USA, Germany and the UK to perfect a device capable of receiving over the air clear images and sound. US companies had been especially anxious to perfect a receiver that could be mass produced. Military research during World War II had advanced their work.

In the late 1940s, the USA was best positioned to experience a 'television boom.' RCA and other companies were offering television sets that, though expensive, were not beyond the means of most consumers. At the same time, RCA established a nationwide network, National Broadcasting Company (NBC), that offered set owners regular programming. Another TV set manufacturer, Du Mont, similarly founded a TV network, as did the radio networks, ABC and Columbia Broadcasting System (CBS) (Du Mont ceased operations in 1955). The four chains provided programming to individual stations, most of which, because of federal regulations, were not owned by the networks. With TV stations operating in most of the nation's larger cities, consumers began purchasing sets. A strong economy allowed such consumption. So did the postwar boom in marriages and child-bearing, which left many young adults housebound, thus making television relatively cost-efficient home entertainment. Although high income was a strong predictor of set ownership in the first years of the television boom, family size soon became the stronger variable. By the late 1950s just under 90 percent of all US homes had one or more televisions. In time, owning a television became an expectation throughout the world. That said, television spread more slowly in poor countries because of the relatively high costs of receivers, the absence of electrical service and uneven TV production standards.

In contrast to what emerged in many other nations, the so-called 'American' system followed the US radio model of privately owned broadcast stations and networks carrying programs or series paid for and sometimes produced by advertisers. Oversight by the US Federal Communications Commission (FCC) was nominal. As the medium's patron, advertisers held enormous control over programs in television's first dozen years, especially because demand favored them over networks and stations. They normally sponsored entire shows, often including plugs of their wares. In most instances, advertisers sought the largest possible audience. This came to mean relying on a set of program 'genres,' most of which owed something to radio and motion pictures. That said, a few sponsors self-consciously cultivated a smaller, elite audience by underwriting musical, dramatic or news programs. Many advertisers and broadcasters initially assumed that the TV viewer was better educated and well to do. Partly because of this misidentification of the audience, which also occurred during the early years of radio, many network programs were culturally ambitious. Original dramas, produced in New York, the nation's theatrical center, appeared every night. Critics later insisted that the USA had experienced a 'Golden Age of Television.' However, the fare most identified with the Golden Age could not, over time, compete for viewer favor with other, less pretentious efforts.

In the late 1950s, US television underwent a series of changes that established industry practice for the remainder of the century. As television spread into smaller communities, demand for dramatic and comedy programs that had followings in the larger cities collapsed. Individual comics like Milton Berle and Sid Caesar, whose humor resonated with eastern viewers, were unable to sustain their early success. The larger, more representative audience preferred filmed programs, including westerns and detective dramas, that were inexpensive imitations of motion pictures. This shift in demand encouraged the movement, already underway, of TV production from New York to California. Television increasingly took on the look of Hollywood as opposed to Broadway. At the same time, a sharp increase in advertiser demand gave the networks and larger stations the upper hand over advertisers in programming decisions. By the mid-1960s, most television programs had multiple sponsors.

Outside of Latin America and the Caribbean, much of the world actively sought to avoid taking up the US system. Much as they had disdained US motion pictures, political and intellectual elites abhorred what they construed to be the vulgarity of US television. As a result, few broadcasters enjoyed the autonomy of their US counterparts. Communist and many third-world countries tightly controlled television's development and production. The Soviet Union and other regimes regarded television, like radio, as state instruments. Other countries, notably in Western Europe, followed the corporation model of the British Broadcasting System (BBC). Under the British model, the government created an independent broadcasting authority that relied on receiver fees rather than advertising. In the mid-1980s, about two-thirds of Europe's television authorities derived their revenues from receiver fees, as did about half of all African and Asian systems. Most of the time, BBC was free of direct government interference. The UK did permit commercial television in creating the Independent Broadcasting Authority (IBA) in 1955. But commercial TV in the UK, as in many other nations, was subject to far more regulations than in the USA. In the UK, non-government TV endured an excess profits tax that had the effect of encouraging more elaborate

cultural programming that rivaled the self-conscious, educational efforts of BBC. The Japanese Broadcasting System (NHK), in contrast, gave commercial TV far greater freedom. Although former British colonies initially embraced the BBC model, political anxieties caused the leaders of most newly independent countries to take up more direct, state control akin to that of the Soviet Union. Then, too, the relatively high cost of starting television service and the absence of domestic commercial investors partly explained the adoption of the statist model in most poor nations. Some third-world leaders sought to put off TV's development indefinitely, only to relent under pressure from urban elites who had encountered television in the West.

Although many if not most nations sought to avoid the US system, few could resist its programming. In the early 1980s, UNESCO reported that most nations, including those with totalitarian or Islamic fundamentalist regimes, offered more entertainment programming than any other type. This tendency created an enormous opportunity for US producers. US-made entertainment series, especially those that emphasized action or broad humor, proved immensely popular. As in the case of motion picture assembly, US television producers had a huge comparative advantage in the emerging global market for programming. The switch of US networks to filmed series created a substantial inventory of exportable entertainment. The USA possessed unequaled production facilities in southern California, as well as the movie industry's large reserve of on-camera and production talent. A large domestic audience and advertising demand, as well as the sale of syndication rights as 'reruns' to individual US stations (and eventually, cable outlets), meant that successful US programs were profitable before they were distributed overseas. They could thus be offered relatively inexpensively. British television production, with somewhat similar advantages, was a distant second as an exporter of TV programming. Most other nations simply could not produce programming of the same standard and appeal. This was especially true in less developed countries where, Elihu Katz and George Wedell estimated in the mid-1970s, imported series constituted just over half of all programming. Indeed, viewers in some countries, Katz and Wedell noted, frequently could select from different US programs aired by competing channels at the same time. One scholar of international communication concluded in 1988, 'A kind of sameness fills TV screens just about everywhere.' Revealingly, a 1999 NATO air strike against a communications center in Belgrade, Yugoslavia, destroyed 123 episodes of three US series.

The Americanization of television throughout much of the world had its critics. One of the first, Herbert I. Schiller, regarded the US's export of programming as part of a conscious attempt by the nation's political and business elites to extend the USA's influence via communication. 'The "have-not" nations,' he wrote, 'stand practically defenseless before a rampaging Western commercialism.' More likely, Katz and Wedell argue, market forces and the law of comparative advantage, and not sinister calculation, accounted for the US's impact on television globally.

In its first decade and a half, US commercial television presented relatively little news programming, very little of which elicited controversy. News programs lost money and drew few viewers. The networks considered informational programming a way of placating influential political and cultural leaders as well as of meeting the FCC's modest public service standard. In receiving TV frequencies or licenses for a fixed term, stations promised to devote a minority of its overall schedule to programs, including news, 'in the public interest.' In theory, license renewal hinged on stations meeting this 'minority interest' standard. In practice, the FCC very rarely scrutinized individual licensees' public service schedule. Nonetheless, station managers were not prepared to ignore the agency altogether. Not surprisingly, television news tended to be uncritical of those in power and, as Nancy Bernhard has argued, supportive of America's Cold War with the Soviet Union.

By the 1960s, US networks and many stations were profiting from certain types of informational programs, specifically early evening newscasts. These were not as popular as some contemporary observers assumed. In the late 1960s, one survey indicated, just over half of all US citizens watched no network evening newscasts during a two-week period. Yet television news did command the attention of many influential US citizens, who assumed its primacy; TV newscasts also drew increasing advertiser interest. Indeed, nightly newscasts became stations' largest single source of revenue. In the 1970s, the network evening 'magazine,' CBS's *Sixty Minutes*, with its mix of exposes and features, started winning viewers For the first time, a network news program actually competed against evening entertainment shows in popularity rankings. Imitations on other networks followed. Although none could generate the following of *60 Minutes*, because they were less costly than entertainment series, the networks earned more money carrying such programs. All told, the realization that news shows could, in fact, be even more profitable, caused broadcasters to seek more viewers by adopting a less serious news agenda. The line separating news and entertainment blurred.

The first great test of US television news came with the US intervention in the Vietnam war. The medium earned at best a low pass. Popular mythology, perpetuated by surviving correspondents (and their critics), insists that TV brought the war's horrors home every evening to a people glued to their sets. Television reporters, according to this wisdom, criticized the conflict itself and thus helped to end US involvement. Subsequent scholarship by Daniel Hallin and others

has largely discredited this self-serving history. As already noted, newscast audiences were never as vast as some imagined. Nor did TV news, for a variety of reasons, show much combat footage. Relatively few reports were carried live. And in the initial years of the US build-up, 1964–66, reporters and news anchors (or readers) passively if not enthusiastically relayed the positions of the US military and civilian leadership. After 1966, some TV correspondents did challenge, ever so mildly, the progress of the war. Yet television was hardly, as some later asserted, leading public opinion or the news media generally. Other news organizations, specifically the hugely influential *New York Times*, had already started to express skepticism about the Vietnam commitment. And television news failed, Todd Gitlin has contended, to take the budding anti-war movement seriously.

As the USA withdrew forces from Southeast Asia, network television had to cover an even larger crisis. These involved criminal activities associated with the June 1972 break-in of the Watergate apartment complex in Washington DC, by figures tied to the re-election campaign of President Richard M. Nixon. The TV networks, especially CBS, covered the Watergate story. Still, they did not do so with nearly the rapidity and thoroughness of some newspapers. Herbert J. Gans' study of the evening network newscast production in the 1970s stressed the extent to which NBC and CBS depended on the *New York Times* for their news agenda.

US television remained primarily an entertainment medium, and TV's staple product began to change. A 1970 FCC regulation restricted the networks' ownership of entertainment programming. This gave producers more freedom, and some sought to convey the more relaxed cultural climate. To many critics, this shift fostered a second 'golden age.' Shows like *All in the Family* and *The Mary Tyler Moore Show*, in violation of industry practice, dealt with contemporary issues like race relations and the sexual revolution. Dramatic series in the 1980s and 1990s, usually involving law enforcement or medicine, similarly took a more realistic tone.

Despite renewed critical respect for some shows, network television's hold over the US viewer slipped in the latter decades of the century. In the 1950s and 1960s, most US citizens received their TV service over the air. The finite size of the radio band limited the number of on-air stations the typical set could receive. And until the late 1970s, the three networks in effect determined what US citizens watched. Most TV stations remained affiliated with a network. On a given evening, nine out of every ten televisions were tuned to a network program. In the 1980s, however, the networks began losing some of their audience. Although the presence of more independent stations cut into the networks' share, the spread of cable proved the far more decisive blow. By subscribing to a cable service, consumers could greatly augment their choices, originally from four to 12 to 30 or more channels. By 1997, the average US cable system provided 47 channels. This much larger array of choices proved too tempting for most consumers. The percentage of homes with cable rose from 7.6 percent in 1970 to 49.4 percent in 1988. By the late 1990s, up to 70 percent of all households had cable. Others had satellite signal receivers that afforded them even more services; such systems were more common elsewhere in the world. The US network audience shrank to about half of the total evening audience.

Cable's rise had several significant effects on the television. Providing many more channels greatly weakened an original justification—scarcity of service—for broadcast regulation, as did a larger disillusionment with the efficacy of business oversight. The FCC relaxed most rules regarding public affairs programming and eventually, network ownership of entertainment programs. The growth of all-news cable channels—along with the end of the Cold War—caused the network news divisions to reduce their coverage of more serious political and international news in favor of stories more likely to attract viewers. At the same, competition from cable channels caused the established networks to forsake their traditional observance of US prudishness. Trying to hold viewers, some network program producers took more risks in their handling of sexuality and crude language. Compared with their predecessors, broadcasters worried less about offending some viewers, who could always turn to a cable channel rerunning an innocuous 1960s situation comedy. In planning a racy police drama, *NYPD Blue*, producer Steven Bochco recalled being told by an ABC official in the early 1990s 'to come up with something that would at least begin to compete with cable.'

Cable channels offered a range of programs and formats. Even so, one of the ironies of cable's growth was that the most popular cable channels relied excessively on reruns of series originally telecast on the networks. These were accompanied on some channels by professional wrestling and variety shows that targeted young men who delighted in bad taste. Many cable channels practiced 'narrow-casting,' that is, they presented programming intended to engage only sports fans (ESPN) or pre-adolescents (MTV). Although the popularity of no one cable channel could compete with any of the networks, combined the cable outlets constituted a regiment of Lilliputians tying down the three Gullivers.

The networks were also wounded by the entry of new rivals in the 1980s and 1990s. None proved more successful than Fox, whose programs usually ignored industry practice and middle class proprieties. 'We were free from the hereditary constraints at the other networks, the decades of doing things a traditional way,' a Fox vice president explained in 1997. Along with still more networks, WB and United Paramount Network (UPN), Fox frequently violated another

industry rule by playing to subgroups of the national audience, that is, younger or African–American viewers.

Although the advent of more TV channels mainly on cable complicated their routines, advertisers in many ways benefited from the more competitive television. Promoters of products more appealing to a specific subgroup, like adolescents or women, welcomed the efficiencies of cable narrow-casting. At the same time, cable's growth restored some power to advertisers in their dealings with the networks. Because younger viewers were thought to be more likely to try new brands and services, product makers in the 1990s were prepared to pay more – sometimes significantly more—to co-sponsor a TV program with the 'right' demographics. That is, they began basing more of their 'media buys' on who rather than how many watched a given show. All in all, this represented an unprecedented shift by the networks' underwriters, most of whom had heretofore prized the size rather than the composition of the audience.

Amid such industry upheaval came dramatic changes in industry proprietorship. After several generations of stable ownership, all three established networks changed hands in the 1980s. Capital Cities, owners of TV stations and newspapers, purchased ABC in 1985 and then sold the network to the Disney entertainment conglomerate in 1995. In 1986, General Electric took over RCA and NBC; investor Laurence Tisch bought CBS. Westinghouse, an electronics company, secured CBS in 1995 only to allow itself to be combined with Viacom, another entertainment concern, in 2000. Meanwhile, individual stations, some of which had been controlled by the same company or family for several generations, were sold and resold through the 1980s and 1990s. Such franchise transfers increased all the more when Congress passed the Telecommunications Act of 1996, which lifted or weakened FCC regulations, including restrictions on the number of stations individual companies could own.

With only with a few exceptions did US television remain the mass medium that it had been in 1960 or 1975. Annual rituals like the National Football League's Super Bowl commanded huge audiences, as did momentarily popular shows like *Who Wants to Be a Millionaire* and *Survivor*. Otherwise, even top-ranked network programs like NBC's comedy *Seinfeld* had to settle for a much smaller percentage of the total television audience than had comparably ranked shows a generation earlier. The future offered more of the same. The spread of the Internet into more US homes at the century's end portended ever more individualized media services.

1. *The Historiography of Television*

The history of television has only begun to command serious scholarship. Until the 1990s, most research on the medium has consisted of highly specialized studies conducted by social scientists testing theories of media influence. The first wave of this research is ably summarized in Leo Bogart's *Age of Television* (1972). Historians, except for those tied to programs focusing on the film or journalism, have all but avoided the medium, despite its immense popularity and seeming influence. Their reluctance is understandable. Few are familiar with the constantly ballooning research on media effects. Moreover, evaluating mass culture itself, as opposed to its influence, has always been a tricky endeavor, much simpler for newspaper columnists than academics. It became all but impossible in the latter decades of the twentieth centuries, as scholars in the humanities and sociology dismissed cultural hierarchies altogether as hopelessly subjective, if not indicators of class and educational attainment. Historians who could pass these obstacles faced more logistical roadblocks. In the USA, corporations have been the major players in the medium's history, and relatively few have opened their records to scholars. Then, too, the sheer volume of TV programming over the years has all but compelled historians to limit their inquiries to case studies (see, e.g., Balio 1990, Spigel and Curtin 1997). Relying on middle-class sources, Lynn Spigel has sensitively examined how US citizens sought to adapt themselves and their families to television. There have been more ambitious histories, notably by Asa Briggs (1961–95)of the BBC and Erik Barnouw's more episodic treatment of US broadcasting (1970). Paul Rutherford has written a very competent history of Canadian television (1990, see also Levy 1999).

Meanwhile, many gaps will need to be closed, including far more examinations of ethnicity and gender. Virtually no systematic investigations have been conducted on advertisers' complex and changing relationships to the medium. Much more research needs to be attempted measuring the relationships between TV news and those in power.

Finally, the tendency of program producers to take up concepts first developed overseas needs to be explored. Successful US series had many imitators over the years. Still, the phenomenal success at the end of the twentieth century of ABC's *Who Wants to be a Millionaire* and CBS's *Survivor*, both modeled after European programs, demonstrated that even US program producers have occasionally engaged in plagiarism. CBS modeled its long-running *Sixty Minutes* after a Canadian program. British programs inspired such popular US series as *All in the Family* and *Sanford and Son* as well as some far less resilient endeavors.

See also: Globalization and World Culture; Media and History: Cultural Concerns; Media Effects; Media Effects on Children; Media Events; Media, Uses of; News: General; Television: General; Television: Industry

Bibliography

Balio T (ed.) 1990 *Hollywood in the Age of Television*. Unwin Hyman, Boston
Barnouw E 1970 *The Image Empire*. Oxford University Press, New York
Baughman J L 1997 *Republic of Mass Culture: Journalism, Filmmaking and Broadcasting in America since 1941*, 2nd edn. Johns Hopkins University Press, Baltimore
Bernhard N 1999 *U S Television News and Cold War Propaganda, 1947–1960*. Cambridge University Press, New York
Bogart L 1972 *The Age of Television*, 3rd edn. Frederick Ungar, New York
Briggs A 1961–95 *The History of Broadcasting in the United Kingdom*. Oxford University Press, New York, 5 Vols.
Gans H 1979 *Deciding What's News*. Pantheon, New York
Gitlin T 1980 *The Whole World Is Watching: Mass Media in the Making & Unmaking of the New Left*. University of California Press, Berkeley, CA
Hallin D 1986 *The 'Uncensored War': The Media and Vietnam*. Oxford University Press, New York
Katz E, Wedell G 1977 *Broadcasting in the Third World: Promise and Performance*. Harvard University Press, Cambridge, MA
Levy M (ed.) 1999 *La Television dans la Republique, less annees 50*. Complexe, Paris
Rutherford P 1990 *When Television Was Young: Primetime Canada 1952–1967*. University of Toronto Press, Toronto
Schiller H I 1971 *Mass Communications and American Empire*. Beacon Press, Boston
Spigel L 1992 *Make Room for TV: Television and the Family Ideal in Postwar America*. University of Chicago Press, Chicago
Spigel L, Curtin M (eds.) 1997 *The Revolution Wasn't Televised: Sixties Television and Social Conflict*. Routledge, New York

J. L. Baughman

Television: Industry

The television industry, by the close of the twentieth century, had grown into an economic force generating mass media's greatest revenues, costs, and profits. In the USA the production, distribution, and broadcasting of television (and later cablecasting and delivery directly by satellite) has long been dominated by a few vertically integrated networks, which produce much of their own programming, and often own broadcast stations, cable systems, and satellite-to-home divisions. While a variety of state-controlled or semi-independent publicly funded television systems have existed around the world, by the end of the twentieth century, more and more TV industries outside the USA were deregulated, that is privatized, and began to look increasingly like the American commercial television industry model.

1. The US Television Industry

From its innovation after World War Two to the middle 1980s, in the USA three national systems of interconnection or networks dominated the TV business. Then came cable TV's 100-plus networks, and in the 1990s even more offerings through satellite-to-home delivered direct TV or DBS. This choice among hundreds of channels seemed to signal an open, competitive industry. Just the opposite was the case.

Seven corporations dominated. A seeming plethora of TV choices collapsed to seven conglomerates all built around organizations producing programming allied with other parts of the same company that determine the program scheduling. In most cases the entertainment programs were made in Hollywood in a deal with a studio such as Disney or Paramount. These corporations also own TV networks which broadcast, cablecast, or deliver shows via satellite to the home. Executives with the same company make programming decisions either aimed at mass audiences, the so-called traditional networks such as Disney's American Broadcasting Company (ABC), or more and more to niche networks, such as Disney's growing number of cable channels.

1.1 US Broadcast TV

The television industry in the USA started, like all others in the world, broadcasting signals over the air. From the late 1940s to the 1980s, three networks dominated American television—NBC, CBS, and ABC. In the 1990s came three more—United Paramount, the WB, and Fox. Five companies owned them as the twenty-first century commenced: Disney's ABC, Viacom's CBS and United Paramount Network, AOL Time Warner's WB, News Corporation's Fox, and General Electric's NBC. These six networks—all parts of vast corporations—defined what most US watchers viewed. The actual local TV stations in the USA simply functioned as spigots, drawing their programming from these networks plus a smattering of syndicated fare, and three or four hours per day of locally produced news.

By the close of the twentieth century, all the networks—save NBC—were tied directly to Hollywood-based companies. Broadcast television became a classic case of vertical integration. The networks had long produced news and sports; by 2000 their Hollywood parents were creating all forms of programming, from comedies to dramas, from reality shows to news reports. They produced the show, distributed it through their own networks, and then showed it to much of the nation on their increasing number of owned and operated stations (by 2000 covering more than a third of all US households).

The Australian media baron Rupert Murdoch introduced the Hollywood vertically integrated TV network system when in 1985 he bought a studio (Twentieth Century Fox), then purchased Metromedia TV stations (for his owned and operated group), and then launched the first new TV network in 30 years—Fox. During the 1990s, Viacom followed

Murdoch's lead and fashioned its United Paramount Network (supplied by its Paramount studio), and Time Warner created the WB network. Disney underscored the importance of this new broadcast network economics when in 1995 it acquired ABC. In 1999, Viacom purchased CBS and so by 2000 owned two broadcast networks.

1.2 Cable TV

As late as 1980, most viewers in the USA watched solely broadcast TV, but then more and more households began to subscribe to cable television. As the twentieth century ended, about two-thirds of Americans paid monthly fees to a local cable monopoly. After the passage of the 1996 Telecommunications Act, cable franchises consolidated, and in 1998 AT&T acquired cable franchises representing about a third of all customers in the USA. AT&T's CEO Michael Armstrong bet his corporation's future on cable, based on two principles. First, cable was a legal monopoly, reminding Armstrong of the old phone company. Second, Armstrong saw that cable's broadband wires could offer not only offer television but also Internet access.

At first glance there seemed to be a plethora of cable networks. There was BET for African Americans, TNN for fans of country music, and ESPN for sports fans, to name but three of an estimated 150 cable networks. Yet of all these choices, most were owned in part (or completely) and/or operated by one of the major TV media conglomerate companies noted above as controlling broadcast television. These companies desired to own networks so that their local franchises could be guaranteed popular programming. In 1999, AT&T controlled through its corporate partner, Liberty Media, the Discovery Channel, The Learning Channel, BET, and a half dozen others, while AOL Time Warner, the second-top owners of cable franchises, owned and operated Superstation TBS, the network piped into most cable homes, plus TNT, and three CNN services.

But not all cable networks were controlled by these vast, vertically integrated media conglomerates. An alternative tactic, reasoned a few Hollywood moguls, concluded that it was not necessary to spend billions of dollars to wire local cable franchises, but simply to produce desirable programming, and TV viewers would find and pay for those channels, regardless of who was their local monopoly cable franchise. Disney, Fox, and Viacom executives, for the three leading examples, concentrated in creating programming to feed cable networks such as ESPN, The Family Channel, and MTV, respectively. They did not vertically integrate.

In sum, cable networks add no dominant new owners to the list of TV powerhouses described above. As a consortium, these very media corporations financed C-SPAN as a public service, as well as to insure access to the US Congress. Indeed, only one of the top 20 cable networks could be called independent—Landmark's Weather Channel. It is more efficient to think of the dozens of cable networks as simply outlets for Disney, Viacom, AOL Time Warner, The News Corporation, AT&T, and General Electric. Indeed, to think of separate broadcast TV and cable TV industries no longer makes sense from an ownership perspective. The aforementioned major companies all own sizable broadcast and cable television production studio properties. They also dominate pay-TV—Viacom's Showtime and the Movie Channel, and Time Warner's HBO and Cinemax—as well as the creation and distribution of movies on rented and sold video.

1.3 Satellite Delivered Television

The 1990s did introduce a whole new means of gaining access to television— direct broadcast satellite (DBS), that is, satellite-to-home direct delivery. After 1994, many new DBS entrants flooded in; one in ten of US television households signed up before the century ended. Paying more, DBS subscribers could access up to 200 channels, principally around the clock pay-per-view movies, and sports broadcasts from all regions of the USA and around the world. DBS pioneers took direct aim at TV's junkies. Until the final month of 1999, however, DBS subscribers could not get local broadcasts; then the US Congress removed this restriction, and in most places in the USA subscribers to DBS were able to access all channels available on cable as well as a hundred more.

By the time the law was changed in November 1999, though, the world of DBS competitors had been winnowed to one dominant player. DirecTV, owned by Hughes Electronics, a division of General Motors, possessed more than 90 percent of the market, with tiny EchoStar and its DISH Network struggling to stay in business. In nearly all markets across the USA DirecTV alone was positioned as the alternative to the monopoly cable company.

DirecTV thus completed the list of the seven dominant TV companies in the USA, and indeed the seven major multinational TV industrial giants around the world. These seven dominant companies—General Electric, Viacom, Disney, News Corporation's Fox, AOL Time Warner, AT&T, and DirecTV—defined TV's oligopoly. The world of the TV industry in the USA had expanded from three broadcast networks to seven key corporations. The big change was that for cable- and satellite-delivered TV one had to pay. But the oligopolistic seven did not offer all forms of programming. Within each genre of TV programming, they picked their spots to compete. So for 24-hour news, for example, a customer had three choices: AOL Time Warner's CNN, NBC's MSNBC, or News

Corporation's Fox News. The other four conceded this genre, and concentrated on becoming the top player in other program categories.

2. *The Rest of the World*

As the twenty-first century began, the major TV conglomerates relied more and more on exporting programming around the world. Yet Hollywood-made TV by no means monopolizes the world's television. For example, in Latin America, Brazilian and Mexican TV industries were successful in regionally exporting TV programming. In Europe, French and German companies did the same thing. Still, overall Hollywood functioned as the leading maker and exporter of TV entertainment. Its influential trade association—the Motion Picture Association of America—smoothly paved the way for export by lobbying both the US Department of State and governments around the world to keep TV trade open and unrestricted. Foreign governments resented US pressure, and formally and informally struggled to protect their domestic TV industries.

While much of the TV industry outside the USA started as an extension of government, by the 1990s deregulation had eliminated most regulations, and the TV industries in wealthy nations in Europe, the Americas, and Asia looked more and more like mini-versions of Hollywood. The TV industries of Japan, Brazil, France, and Canada, as typical examples, for the key variables of industrial analysis would include (a) geography, particularly land mass; (b) economic development, particularly following the collapse of the Soviet Union; and (c) long-term cultural differences, whereby some countries continue to see the TV business as having formidable externalities (i.e., benefits to the national state). In Canada, for example, the Canadian Broadcasting System is strongly supported by the nation–state to ward off the overwhelming of the Canadian culture by US interests. As the twentieth century closed, the differences among larger developed nations—save for the pressure from the Big Seven—had narrowed, as all adopted some amalgamation of advertising-based and pay television, delivered by broadcast, cable, and satellite.

Small countries—such as The Netherlands and Belgium, and any African state—even if economically well developed, more than ever needed strong state support and subsidy. Undeveloped nations—where TV was important, but feeding the nation far more vital—increasingly saw their native TV industries struggle against the vast multinational invasions. Indeed, throughout developed and underdeveloped Asia, Rupert Murdoch's Star TV network brought in news coverage that shattered government monopolies, and introduced US and European cultural products into cultures that could offer only minimally financed alternatives.

European nations banded together as the European Community to challenge the multinational TV industry giants in order to preserve local and native culture. With the percentage of European households owning and using TV sets increasing dramatically as the twentieth century ended, no governmental monopoly could satisfy new channel demands, particularly once cable TV and DBS were introduced. With more TV use, and less national control, concern rose about foreign influences, and a united Europe struggled to set up quotas so Europeans could watch European-made TV shows, not simply more and more from Disney, Viacom, and the other vast multinational TV conglomerates. This sort of nation–state collective action will continue as long as the structure of TV industries around the world insures that economies of scale result in the costs of Hollywood programming being less than costs of local programming.

Some nations have stakes in multinational success, such as the United Kingdom with its Pearson PLC, Germany with its Bertelsmann AG, Australia with its News Corporation, and Japan with its Sony. Spain and Italy have tried following the same model, but with less success. The big unanswered question of this assistance to multinationals comes from emerging national powers such as India, Indonesia, and China, all with vast populations, but all still emerging from colonialism or communism and still finding a place for domestic industries in the face of the world's multinational companies. For all TV industries outside the USA, change will remain the order of the day as media conglomerates merge and states seek the resources to protect their native cultures.

3. *Analytical Perspectives on the TV Industry*

There are several leading approaches to the analysis of television as an industry. Critics on the left have long seen the television industry as an increasingly influential cultural industry no different from other industries in monopoly capitalism. For example, Edward Herman and Robert McChesney (1997) focus on the inevitable rise and continuation of giant corporations, and the lack of true TV alternatives that challenge the dominant political–economic structure. Even with the increasing number of cable channels, one strains to find any progressive alternatives to corporate voices that sing largely in unison. The dominant TV powers have continuously increased their power as they fashion a global media presence. In particular, the innovation of space satellites as the dominant means of distribution—replacing coaxial cables and microwave relays—enables the images and sounds of television to flow easily from anywhere to anywhere else. Critics such as Herman and McChesney argue that the dream of TV as a means of mass education and democratic enlightenment has turned into a system devoted almost wholly to selling

products and avoiding controversial presentations which would inform citizens.

Conservative neoclassical industry analysts, in contrast, deny the cultural homogeneity of which left-wing critics complain. Such analysts as Benjamin Compaine insist that TV shows face competition from many sources of popular culture. Defenders of the status quo, like Bruce Owen and Steven Wildman (1992) agree with TV executives who claim they face a highly contested domain with ever-changing forces of supply and demand. They stress that no global television producer or network can force anyone to tune in. With more channels being offered by cable and directly by satellite, one can hardly argue there exists little diversity or choice in the current configuration of the TV industry. Genres of TV fare—from docudramas to science fiction, from news magazines to news reports 24-hours a day—see their ratings rise and fall. With the rise of new TV technologies like the VCR and DVD, for example, the TV fan can take charge of their set, and program their own favorite shows. As profit-seeking companies making up the TV industry add more and more channels, viewers will continue to be better and better off, argue neoclassical economists and industry owners.

Both camps come to their studies with predetermined assumptions. Leftists know the system does not work; neoclassicists know it does. A third group, industrial organization economists, perhaps start from a less ideological base. They instead seek to show how different industry structures lead to different forms of firm and industry conduct. As Gomery (1993) has pointed out, TV industries are not all the same, nor are their social and political effects either uniformly good or uniformly bad for the public welfare. The industrial organization economists describe and analyze the structure, conduct, and performance of each industry—stating up-front, not simply assuming what values might constitute good or bad performance. If the refusal of industrial organization economists to begin with a normative orientation is a strength, it weakens their grounds to promote particular public policy recommendations, apart from accepting the terms or performance criteria that policy makers themselves bring to the table. Industrial organizational analysts can bring alternative policy recommendations to the table, but cannot argue that they have the best solution to industrial problems critics might raise. They can offer only fixes on the margin, accepting (indeed assuming) that capitalism rarely offers the best media policy left unfettered, and a radical alternative to capitalism seems not to promote a more enlightened, educated citizenry.

See also: Film and Video Industry; Mass Media: Introduction and Schools of Thought; Mass Media, Political Economy of; Media Imperialism; Media, Uses of; Television: General; Television: History

Bibliography

Albarran A, Chan-Olmsted S 1998 *Global Media Economics*. Iowa State University Press, Ames, IA

Alexander A, Owers J, Carveth R 1969 *Media Economics*. Erlbaum, Mahwah, NJ

Browne D 1989 *Comparing Broadcast Systems: The Experience of Six Industrialized Nations*. Iowa State University Press, Ames, IA

Compaine B, Gomery D 2000 Who *Owns the Media?* Erlbaum, Mahwah, NJ

Dizard W 1966 *Television: A World View*. Syracuse University Press, Syracuse, NY

Gershon R 1997 *The Transnational Media Corporation: Global Messages and Free Market Competition*. Erlbaum, Mahwah, NJ

Gomery D 1993 The centrality of media economics. *Journal of Communication* **43**(3): 190–8

Herman E, McChesney R 1997 *The Global Media*. Cassell, London

Hoffmann-Reim W 1996 *Regulating Media: The Licensing and Supervision of Broadcasting in Six Countries*. Guilford Press, New York

Katz E, Wedell G 1977 *Broadcasting in the Third World*. Harvard University Press, Cambridge, MA

Lardner J 1987 *Fast Forward: Hollywood, the Japanese, and the Onslaught of the VCR*. Norton, New York

O'Neil P H 1997 *Post-communism and the Media in Eastern Europe*. Frank Cass, Portland, OR

Owen B, Wildman S S 1992 *Video Economics*. Harvard University Press, Cambridge, MA

Parsons P, Frieden R 1998 *The Cable and Satellite Television Industries*. Allyn and Bacon, New York

Skidmore T 1993 *Television, Politics, and the Transition to Democracy in Latin America*. The Johns Hopkins University Press, Baltimore, MD

Tunstall J 1977 *The Media Are American: Anglo-American Media in the World*. Columbia University Press, New York

D. Gomery

Temperament and Human Development

Because temperament refers to the early-appearing dispositions upon which personality is based, it can be studied in infants and nonhuman animals as well as older children and adults. It offers an important place to begin in understanding human development. Rather than seeing individual differences in personality and social behavior as solely the result of socialization, temperament research identifies basic dispositions toward affect, arousal, and attention that will be further shaped by experience in development.

1. A Definition of Temperament

Temperament has been defined as constitutionally based individual differences in emotional, motor, and

attentional reactivity and self-regulation, demonstrating consistency across situations and relative stability over time (Rothbart and Derryberry 1981). The term 'constitutional' stresses links between temperament and biology. Over the long history of the study of temperament, individual differences in temperament have been linked to the constitution of the organism as it was understood at the time. The term 'reactivity' refers to the latency, rise time, intensity, and duration of responsiveness to stimulation. The term 'self-regulation' refers to processes that serve to modulate reactivity; these include behavioral approach, withdrawal, inhibition, and executive or effortful attention.

2. *History of Temperament Research*

Temperament study has an ancient history: individual differences in temperament were described in the fourfold typology of the Greco-Roman physician Vindician, who linked temperamental characteristics to Hippocrates' model of the humoral constitution of the body (Diamond 1974). The term 'temperament' derives from the Latin *temperamentum*, a proportionate mixture, denoting the relative preponderance of one or more of the body humors. In Vindician's typology, the melancholic person, quiet and moody, was seen as having a predominance of black bile; the choleric person, touchy, aggressive, and active, a predominance of yellow bile; the sanguine person, sociable and easygoing, a predominance of blood; and the phlegmatic person, calm and even-tempered, a predominance of phlegm. Ideas about the fourfold typology of temperament persisted throughout the Middle Ages and the Renaissance, and were found in the writings of Immanuel Kant.

Early in the twentieth century, major schools in Europe and the UK contributed to temperament research (Rothbart 1989). An Eastern European tradition began with I. P. Pavlov's observations of individual differences in his dogs' behavior in the laboratory. On the basis of the dogs' reactions, Pavlov and his followers posited characteristics of the nervous system, including strength or weakness of excitation and inhibition, and nervous system lability. This was nevertheless a conceptual nervous system; the nervous-system characteristics were based on behavioral observations. Work in Eastern Europe continued in Nebylitsyn's studies of individual differences in the laboratory, with adults as subjects. In the late twentieth century, due to a lack of consistency across stimulus modalities and response systems in the laboratory, questionnaire assays of individual differences have been undertaken.

In Central Europe, research on temperament in The Netherlands was to become highly influential. Heymans and Wiersma, in 1908, asked 3,000 physicians to observe a family (parents and children) and to fill out a temperament/personality questionnaire on each person. When the more than 2,500 responses were analyzed, three factors were identified: activity (the tendency to act out what is thought or desired), emotivity (the tendency to show body symptoms, to be fearful and shy), and primary vs. secondary process (the tendency to react immediately or in a postponed and more organized way). This research, which involved a wide range of subjects, was followed by European research on temperament in early development, with intense longitudinal observation of infants. Wallon in France and Meili in Switzerland investigated infants' muscle tension and emotionality, and Meili reported stability across time in infants' irritable or anxious responses to standardized stimuli. Meili's contributions showed a similarity to later research in the USA on temperamental emotions and behavioral inhibition (G. Balleyguier, personal communication, September 8, 1996).

In the UK, factor-analytic studies of individual differences in temperament and personality were carried out using adults' self-reports. Early work yielded factors of introversion–extraversion, emotional stability–instability (later called by Eysenck 'neuroticism'), and volition or will (Rothbart 1989). Eysenck added a physiological base to the temperament factors by relating them to cortical excitation and inhibition, and the functioning of the limbic system. Still later, Gray revised Eysenck's theory by positing individual differences in behavioral activation and inhibition, as well as tendencies to fight and flight. He also related these individual differences to an underlying neurophysiology. At the turn of the twenty-first century, Gray's theory was one of the major psychobiological models of temperament, along with models put forward by Panksepp, Cloninger, DePue, and Zuckerman.

In the USA, there were multiple contributions to the study of temperament. Allport used temperament in his trait-based theory of personality, and Cattell and Thurstone carried out early factor-analytic research on temperament and 'vectors of mind.' These would later be followed by research on the Big Five personality factors in adults (Digman 1996). Nevertheless, in early research on child development, temperamental dispositions were rarely discussed. Instead, social-learning theories stressed the importance of reward and punishment in shaping behavior and accounting for individual differences. The study of cognitive development would also become much advanced before the study of emotional development and temperament became popular. Only in the late twentieth century did research grow and flourish in temperament and development. This renewed interest has been linked to the realization that the direction of parent–child influence is not only from parent to the child, but that the infant and the young child also bring much to shaping the interaction (Bell 1968). Interest in temperament has also been linked to advances in neuroscience, with knowledge of indi-

vidual differences in temperament providing links to genes and neural networks, as well as to social interaction.

3. Current Research

In the USA, the New York Longitudinal Study (NYLS) by Thomas, Chess, and colleagues formed the basis for much of the recent research on temperament in children (Thomas and Chess 1977). Parents of infants between the ages of two and six months were interviewed about their children's behavior across a wide variety of situations. A content analysis of interview information on the first 22 infants yielded nine dimensions of temperamental variability: activity level, rhythmicity, approach-withdrawal, adaptability, threshold, intensity, mood, distractibility, and attention span-persistence. The goals of the NYLS were chiefly clinical, and no attempt was made conceptually to differentiate these dimensions from one another. Nevertheless, numerous parent-report questionnaires were developed based on the nine NYLS dimensions.

As a result of more recent research, however, major revisions to the NYLS list have been proposed (Rothbart and Bates 1998). Both factor-analytic research using NYLS items, and research following a more rational approach such as that of Rothbart, have identified a shorter list of temperament dimensions. These dimensions include positive affect and approach (extraversion or surgency), fearfulness, anger/frustration, attentional orienting, and executive attention (effortful control). These dimensions are particularly interesting because they are evolutionarily conserved affective-motivational and attentional systems. They serve adaptive functions, and can be seen in nonhuman animals as well as in the human species (Panksepp 1998).

In factor-analytic studies of parent-reported temperament in childhood, three to four broad factors are frequently found (Rothbart and Bates 1998). The first of these is surgency or extraversion, which includes activity level, sociability, impulsivity, and enjoyment of high-intensity pleasure. The second is negative affectivity, including fear, anger/frustration, discomfort, and sadness, and the third is effortful control, including attentional focusing and shifting, inhibitory control, perceptual sensitivity, and low-intensity pleasure.

Current research on temperament in childhood makes use of parent-report questionnaires, laboratory assessments of children's behavioral and psychophysiological responses to standardized stimuli, and observations of children's behavior in the home or school. With this information, it has been possible to study the development of temperamental systems, and to relate temperamental functioning to other important outcomes. These include the development of empathy and conscience, the development of behavior problems, children's adaptations to the classroom, their susceptibility to drug and alcohol abuse, and their accident-proneness.

Developmental research also indicates that the emotional systems of temperament and reactive attentional orienting are in place before the development of executive effortful attention (Rothbart and Bates 1998). In the newborn, individual differences in irritability and orienting can be observed, and by two to three months, infants demonstrate clear positive responses to stimulation. Early forms of what will later be called extraversion or surgency are present in the smiling and laughter and rapid approach of infants, and measures of approach tendencies and smiling and laughter at this early age predict children's extraverted tendencies at seven years. Throughout early development, children who are more extraverted also seem to be more susceptible to anger and frustration.

At four months, individual differences in infants' distress and body movement to stimulation predict later fear and behavioral inhibition. By six months, it is also possible to predict seven-year-old children's parent-reported frustration and anger by the infant's responses to toys out of reach or behind a plastic barrier. Infants' behavioral approach tendencies are also manifested at six months in their latency to reach and grasp objects.

The onset of fear or behavioral inhibition in the last quarter of the first year of life appears to work in opposition to the infant's approach tendencies, so that some infants who formerly rapidly approached novel objects are now slowed in their response to novel stimuli, and may not approach at all. They may also show distress to possibly threatening objects. As with approach tendencies, individual differences in fear-related behavioral inhibition show considerable stability across childhood and even into adolescence (Kagan 1998). Longitudinal research has reported stability of fearful inhibition from two to eight years, and from the preschool period to age 18.

Fear-related control of behavior can be seen in the early development of conscience (Kochanska 1997). Fearful children are more likely to show early signs of conscience. In addition, fearful children whose mothers use gentle discipline, presumably capitalizing on the child's tendency to experience anxious states, develop highly internalized conscience. More fearful infants also later tend to be empathetic and susceptible to guilt reactions. Fear can thus be seen as a basic control mechanism that is importantly implicated in socialization.

Beyond the inhibitory control provided by fear, a later-developing attentional system makes a crucial contribution to socialization. As executive or effortful attention develops in the second or third year of life and beyond, individuals can voluntarily deploy their attention, allowing them to regulate their more re-

active tendencies (Posner and Rothbert 2000, Ruff and Rothbart 1996). In situations where immediate approach is not allowed, for example, children can limit their attention to the rewarding properties of a stimulus, thereby resisting temptation and delaying gratification. Similarly, when faced with a threatening stimulus, children can constrain their fear by attending to environmental sources of safety as well as threat. In both of these examples, individual differences in attention allow children to suppress their more reactive tendencies, take in additional sources of information, and plan more efficient strategies for coping. These capacities have been referred to as effortful control, and they refer to temperamental characteristics that are distinct from extraversion/surgency and the negative emotions. Research indicates some stability of individual differences in effortful control during childhood. For example, the number of seconds delayed by preschool children while waiting for rewards that are physically present has predicted children's parent-reported attentiveness and ability to concentrate as adolescents.

In older children and adults, temperament is often studied through self-reports, allowing analysis of internally experienced feelings as well as behavior. The self-report method is also frequently used in the study of adult personality. In adult research, highly differentiated scales assessing temperament have been factor analyzed, yielding factors very similar to those found in Big Five or Five Factor Model research on adult personality (Rothbart et al. 2000). The Big Five personality factors have been derived from research using trait-descriptive adjectives, following the hypothesis that significant individual differences will be represented in the lexicon, that is, in the words people use to describe others and themselves. The Five Factor Model emerged from factor analyses of large numbers of personality self-report items. There is now considerable agreement for five broad factors of per sonality derived from this method, including extraverversion, agreeableness, conscientiousness, neuroticism, and openness to experience.

In research using adults' self-reports to scales assessing temperamental reactivity and self-regulation, three resulting factors are similar to those found in children and to measures of the Big Five personality factors. These include temperamental surgency or extraversion, positively related to personality extraversion; negative affectivity, related to neuroticism; and effortful control, related to conscientiousness. In addition, self-reported temperamental orienting to external and internal stimuli is related to personality openness to experience, and temperamental affiliativeness to personality agreeableness. Thus, the structures emerging from temperament research—using basic psychological processes of affect, arousal, and attention—and from personality research—using a lexical or personality-scale analysis—are closely related. Because temperamental individual differences are based upon fundamental psychological and neural processes and are present early in life, these findings suggest that early temperamental predispositions may form a core around which the later-developing personality will be built.

As noted above, temperamental systems are evolutionarily conserved. It is interesting that recent research has attempted to apply 'personality' constructs to studies of nonhuman species. A review of factors identified in studies of twelve nonhuman species has found support for factors of extraversion, including energy and enthusiasm; neuroticism, including negative affectivity and nervousness; agreeableness, including altruism and affection; and openness, including originality and open-mindedness (Gosling and John 1999). These are seen in the chimpanzee, monkey, hyena, dog, and cat. Several species tested did not give evidence of attentional openness, and evidence for individual differences in conscientiousness was reported only in chimpanzees. It seems likely that these studies, although labeled studies of personality, are getting at the evolutionarily conserved temperamental systems studied in human temperament research. In addition, since not all capacities appear to be shared across species, especially conscientiousness or effortful control, we can recognize important phylogenetic *differences* in temperament. For example, the capacity for effortful control, when linked with language, provides opportunities for self-regulation of reactive systems in humans that are not present in other species.

Advances in neuroscientific imaging technology and in our knowledge of brain networks underlying the emotions and attention have been an additional source of information for students of temperament and development. There is now detailed knowledge of networks subserving fear as well as both reactive orienting and effortful attention (Rothbart and Bates 1998). Because imaging studies allow researchers to identify tasks that activate these brain networks, it is possible to adapt these tasks to children of different ages in order to study the development of a temperamental system (Posner and Rothbart 2000). This marker-task approach has been used in the study of the development of attentional orienting and effortful control, but it is likely that marker tasks can be used to assess many additional individual differences. In late twentieth-century research, marker tasks of effortful attention in the laboratory were positively related to parents' reports about children's ability to control attention and emotion.

4. *Methodological Issues or Problems*

The usual issues and concerns about psychological measurement apply to the study of temperament. Because the focus of study in temperament is individual differences, it is important to design measures that will demonstrate variability. It is also important

that the tests are reliable, giving consistent measurement, and valid, reflecting something meaningful about the child's dispositions. The latter is often demonstrated in a psychological measure's convergence with other measures.

In temperament research, researchers have often been skeptical about using parents as informants about their children's behavior. It has been felt that parental biases or lack of knowledge will yield measures that are invalid, with direct observation seen as a preferable method. However, considerable evidence indicates convergence between parent report and observational measures (Rothbart and Bates 1998). In addition, laboratory observations have their own limitations. For example, it is difficult to collect extensive information about children's emotionality in the laboratory when one emotional experience has carryover effects on another. Fearful children may also behave differently in a strange laboratory situation than less fearful children, because expressions of temperamental dimensions are not independent of one another.

5. *Probable Future Directions of Theory and Research*

Temperament is a basic integrative construct within psychology, and improvements in temperament concepts and methods will be critical to advances in understanding development. One goal will be a thorough understanding of the basic processes of temperament, including the components of positive affect/approach, fear, frustration, affiliation, and attention, and how they develop. Understanding the affects and attention will benefit the understanding of temperament, just as temperament analyses will benefit our understanding of the affects and attention. Many advances will come from areas of affective and cognitive neuroscience, where much progress has been made in understanding the emotions and attention. Progress in the study of brain structure and its relation to molecular genetics will also be of great importance.

At the same time, improved assessment methods are needed at multiple levels: measures of behavior in the laboratory, including marker tasks for the development of brain structures, behavior in naturalistic daily contexts, physiological measures, informant reports, and self-reports. Increased use of the laboratory, using marker tasks, brain imaging, and recording, is likely to provide further links with underlying neural systems. Questionnaire measures involving self-report will continue to be important.

To further understanding, it will also be necessary to continue to develop the theory of temperament and its development. Because temperament encompasses organized systems of emotional and attentional processes, rather than separate traits, studies of relationships among temperament variables allow a much richer view of development. In this approach, continuity and change within systems, as well as developmentally changing links among the systems, must be considered. Studies of developmental pathways between early temperament and later personality outcomes will be complex, because childhood and adult individuality unfolds in the context of social relationships, and continuity and change cannot be understood without considering the developmentally changing impact of social experience. To understand developmental pathways, it will be necessary to disentangle complex interaction effects among early temperament predispositions, socialization processes, relationships, and culture.

Because of those complexities, temperamentally different children may arrive at similar or equivalent outcomes via different pathways. Pathways of temperamentally different children may converge over development. For example, Kochanska (1997) has found that fearful toddlers whose parents used gentle discipline, and fearless toddlers whose parents capitalized on positive motivation in a close relationship, attained comparable levels of conscience. Temperamentally similar children's developmental pathways may also diverge as a result of different experiences within relationships or varying cultures.

We have entered a very exciting era in developmental psychology. With support from flourishing neuroscience, genetics, developmental psychopathology, and behavioral research in development, the twenty-first century promises to offer new understanding of the uniqueness of pathways of individual growth. Temperament research will make major contributions to this understanding.

See also: Extraversion; Neuroticism; Personality and Adaptive Behaviors; Personality Development and Temperament; Personality Development in Childhood; Personality Psychology; Personality Structure; Personality Theories; Shyness and Behavioral Inhibition; Temperament: Familial Analysis and Genetic Aspects

Bibliography

Bell R Q 1968 A reinterpretation of the direction of effects in studies of socialization. *Psychological Review* **75**: 81–95

Diamond S 1974 *The Roots of Psychology. A Source Book in the History of Ideas*. Basic Books, New York

Digman J M 1996 The curious history of the five factor model. In: Wiggins J S (ed.) *The Five-Factor Model of Personality: Theoretical Perspectives*. Guilford Press, New York

Gosling S D, John O P 1999 Personality dimensions in nonhuman animals: A cross-species review. *Current Directions in Psychological Science* **8**: 69–75

Kagan J 1998 Biology of the child. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology: Vol. 3. Social, Emotional, and Personality Development*, 5th edn. Wiley, New York

Kochanska G 1997 Multiple pathways to conscience for children with different temperaments: From toddlerhood to age 5. *Developmental Psychology* **33**: 228–40
Panksepp J 1998 *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, New York
Posner M I, Rothbart M K 2000 Developing mechanisms of self-regulation. *Development and Psychology* **12**: 427–41
Rothbart M K 1989 Biological processes of temperament. In: Kohnstamm G A, Bates J, Rothbart M K (eds.) *Temperament in Childhood*. Wiley, Chichester, UK, pp. 77–110
Rothbart M K, Ahadi S A Evans D E 2000 Temperament and personality: Origins and outcomes. *Journal of Personality and Social Psychology* **78**: 122–35
Rothbart M K, Bates J E 1998 Temperament. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology: Vol. 3. Social, Emotional, and Personality Development*, 5th edn. Wiley, New York, pp. 105–176
Rothbart M K, Derryberry D 1981 Development of individual differences in temperament (eds.) Lamb M E, Brown A L. *Advances in Developmental Psychology* Erlbaum, Hillsdale, NJ, pp. 37–86
Ruff H A, Rothbart M K 1996 *Attention in Early Development: Themes and Variations*. Oxford University Press, New York
Thomas A, Chess S 1977 *Temperament and Development*. Brunner, New York

M. K. Rothbart

Temperament: Familial Analysis and Genetic Aspects

Individuals differ from birth in their reactivity to stimulation and how they regulate their physiological arousal and emotional states. These temperamental differences are expected to be due to physiological differences that, in turn, may be partly caused by genetic differences. This article reviews the findings genetic studies of behavior that have estimated the heritability of temperamental differences in Western cultures by comparing temperamental traits between members of the same family.

1. Temperament

The concept of temperament as a part of personality that is primarily biologically based, observable already in infancy, and stable throughout the life span, is deeply rooted in prescientific views of human nature. For example, the Hippocratic typology of four temperamental types (sanguine, melancholic, choleric, phlegmatic) was highly influential throughout the Middle Ages and the Enlightenment. In the twentieth century, three main strands of temperament research emerged that later became interconnected (see Strelau and Angleitner 1991).

Soviet and Eastern European research began in the 1930s with Pavlov's observation of differences among dogs in his conditioning experiments. Animals with a 'strong nervous system' responded with increased intensity to increases in stimulation whereas animals with a 'weak nervous system' failed to respond above some threshold. The Pavlov school also introduced the concepts of strength of inhibition, balance between activation and inhibition, and mobility (the flexibility of the adjustment to changing conditions). This approach was developed further in England by Gray (1982) and has influenced particularly physiological and clinical research on adults. According to Gray, individuals differ on two main temperamental dimensions—inhibition and impulsivity. Inhibited individuals react more strongly to conditioned signals of novelty and punishment than uninhibited individuals. Impulsive persons react more strongly to conditioned signals of reward than unimpulsive persons.

A second research tradition was based on self-descriptions of temperament in questionnaires. British factor analytic studies, which were begun in the 1930s by Burt, were taken further by Eysenck who identified two major dimensions of temperamental differences, extraversion–introversion and emotional stability–emotional lability (or neuroticism). Eysenck (1967) was of the opinion that these two dimensions of self-described temperament directly reflected physiological differences in arousability. According to this highly influential theory, extraverts need more stimulation than introverts for optimal arousal, and introverts are sooner overstimulated than extraverts (Pavlov's idea of the weak nervous system). Eysenck attributed these assumed physiological differences to an arousal system in the brain stem. A similar, less developed theory attributed differences in emotional stability to the differential arousability of the limbic system. However, attempts to test this theory in psychological experiments have met with only limited success (see Strelau and Angleitner 1991).

A third research tradition is the study of temperamental differences in infants and young children (Rothbart and Bates 1998). The pediatricians Thomas and Chess initiated the highly influential New York Longitudinal Study on early temperamental differences (Thomas et al. 1963). They repeatedly interviewed parents about their infant's behavior in specific contexts. These descriptions were sorted into nine dimensions of temperament: activity level, approach/withdrawal, adaptability, mood, threshold, intensity, distractibility, rhythmicity, and persistence. Subsequent research showed that some of these nine dimensions were strongly correlated, and parent or teacher questionnaires were designed that assess fewer, less correlated dimensions of infants' or children's temperament. For example, the Colorado Childhood Temperament Inventory assesses five dimensions: (negative) emotionality, activity, persistence, shyness, and sociability.

In recent years, these three traditions have become increasingly interrelated. Gray (1987) showed that his

model of inhibition–impulsivity can be translated into Eysenck's model of extraversion–neuroticism at the level of self-description. Windle and Lerner (1986) developed a questionnaire for the assessment of temperamental dimensions from childhood through adulthood. Halverson et al. (1994) pointed out how temperamental dimensions are related to the more general five-factor model of personality description. By and large, temperament refers to the first two factors of this model, extraversion and emotional stability, which were already identified by the British factor analytic tradition, with age-specific additions particularly for infants and younger children.

Therefore the working definition of temperament for this article refers to individual differences in basic psychological processes that constitute the affective, activational, and attentional core of personality (Rothbart and Bates 1998), and which are mainly related to the first two factors of the five-factor model of personality description, extraversion and neuroticism.

2. *Estimation of Heritability*

The genetic influence on temperamental differences in a human population can be studied both directly and indirectly. The direct method tries to establish a statistical relation between the frequency of variants of a gene in the population (its alleles) and temperamental traits. This method provides estimates of the influence of specific genes on specific traits and requires molecular biological knowledge about which genes and alleles can be distinguished. This direct method became feasible towards the end of the twentieth century as the relevant knowledge became increasingly available.

The present article reviews results from familial analyses conducted in the field of behavioral genetics, a subdiscipline of population genetics (see Plomin et al. 1997 for an overview). This indirect method estimates the overall genetic influence on a specific dimension of temperament relative to the overall environmental influence on this dimension through analyses of the temperamental similarity of members of the same family such as twins, adoptive siblings, and their parents. This method requires only basic knowledge about rules of inheritance, not molecular biological knowledge about genes and their alleles. However, the method is based on certain assumptions that may not always be valid.

The indirect method is based on the fact that the genetic similarity of family members varies from 0 percent shared alleles in the case of adopted siblings or adopted children and their nonbiological parents, to 50 percent shared alleles in the case of biologically related siblings, nonidentical twins (dizygotic twins), and their biological parents, and 100 percent shared alleles in the case of identical twins (monozygotic twins). Siblings, nonidentical twins, and their parents share 50 percent of their alleles on average because half of children's alleles stem from the mother and the other half from the father. Identical twins are an exception because they develop from the same original cell that only later splits into two different individuals; they are therefore genetically identical.

Two basic assumptions of the indirect method are: (a) influences on temperamental differences in the population can be either environmental or genetic, and (b) the overall impact of the environment, relative to genes, is identical for siblings, adoptees, and twins. Therefore, differences in their similarity can be fully interpreted as genetic effects. Thus, if identical twins are more similar in a specific temperamental trait than nonidentical twins, or biologically related siblings are more similar than adopted siblings, this difference in similarity can be genetically interpreted. The larger the difference in similarity, the stronger the genetic influence on the temperamental trait. Thus, there exist two main indirect methods of estimating genetic influences: comparing identical twin pairs with nonidentical twin pairs (twin method), and comparing biologically related sibling pairs with adopted sibling pairs (adoption method). As a result it is possible to crosscheck the results of these two methods with one another. Other indirect methods compare the similarity between identical twins that are reared apart as opposes to those reared together, or the similarity between parents and their biological children as opposed to their adopted children.

The strength of the genetic influence on temperamental differences can be quantified by the heritability coefficient h^2. This coefficient can be interpreted as the proportion of variance in a specific temperamental trait in the population that is due to genetic differences: $h^2 =$ genetic variance/total variance. It can be estimated on the basis of differences between intraclass correlations. In the twin method, the intraclass correlation between nonidentical twins (a measure of their similarity) is subtracted from the intraclass correlation between identical twins. Because this difference estimates $100 - 50$ percent $= 50$ percent of the genetic influence (identical twins share 100 percent of their alleles, and nonidentical twins 50 percent), it is doubled; this double difference is h^2. In the adoption method, the difference between the sibling correlation and the adopted sibling correlation is doubled because it estimates $50 - 0$ percent $= 50$ percent of the genetic influence (adopted siblings do not share alleles).

The heritability coefficient can vary between 0 and 100 percent. Zero heritability would be found if the variation of the temperamental trait in the population is only due to environmental differences. In this case, adopted and biologically related siblings are equally similar in this trait, as are identical and nonidentical twins. A heritability of 100 percent would be found if the variation of the trait in the population is only due to genetic differences. In this case, adopted siblings

would be as dissimilar as children from different families, and identical twins would perfectly match each other in the trait; normal siblings would be moderately similar.

The concept of heritability is a relativistic concept because it estimates the amount of genetic influences relative to environmental influences. Therefore, heritability can vary between populations and between different traits within the same population. Thus, the heritability of adolescent shyness can be higher in the USA than in African cultures, and adolescent shyness can be more heritable than adolescent sociability in the USA. Furthermore, heritability depends on the variability of genes and environments in a given culture. If the genetic variability is increased (e.g., through genetically dissimilar immigrants), or if the environmental variance is decreased (e.g., through compensatory education), the genetic influence can increase; if the genetic variance is decreased (e.g., through a virus that kills people with particular alleles), or if the environmental variance is increased (e.g., through an unjust political system), the genetic influence can decrease.

The heritability of a trait depends not only on the trait, the population, and its culture, but also on age. The genetic influence on a trait may decrease with age because more and more environmental influences accumulate. On the other hand the genetic influence may also increase with age because more and more genes become relevant for the trait or because older children and adults are more able than young children to select their environment according to their genetically influenced interests and motives (such selection would decrease the environmental variance). All in all, heritability is a highly relativistic concept with regard to the target trait, culture, and age.

3. *Heritability of Temperamental Traits*

Empirical studies on the heritability of temperamental traits are nearly exclusively restricted to North American and Western European samples. Therefore it is unknown whether the following findings can be generalized to other cultures. Because heritability can vary with age, the findings are discussed separately for different age groups. Because the findings also vary considerably with the estimation method used, twin studies and adoption studies are discussed separately.

In twin studies of older infants and children, two groups of traits can be distinguished with regard to their heritability (see Goldsmith et al. 1997). A first group of traits (activity, emotionality, sociability, impulsivity, approach/withdrawal) is characterized by a high similarity of identical twins (intraclass correlations approximately 0.60) and a dissimilarity of nonidentical twins (correlations close to zero), suggesting the odd result of more than 100 percent heritability for some of these traits. A second group of traits (inhibition to strangers, adaptability, mood, distractibility, rhythmicity, intensity, persistence, and threshold) is characterized by an equally or even higher similarity of identical twins along with a moderate similarity of nonidentical twins, suggesting 40–60 percent heritability. Adoption studies for this age range have resulted in zero or even negative correlations for adopted children and somewhat more positive correlations for biologically related siblings, suggesting 10–50 percent heritability.

Whereas in all these studies a moderate to strong genetic influence on temperamental traits was found, one twin study found no genetic influence for 5 different observed temperamental traits in 4-day-old infants (Riese 1990). Together with findings of a low stability of temperamental differences over the first months of life, this result can be attributed to strong environmental influences around birth that obscure genetic effects (e.g., perinatal and postnatal difficulties that transitionally influence the infants' emotional and attentional reactivity).

For adolescents and adults, most studies have focused on extraversion and emotional stability (or the reverse trait, neuroticism). Loehlin (1992) reviewed these studies. For twin studies, the results were similar for both traits: approximately 65 percent genetic influence, based on moderately high correlations between identical twins and low correlations between nonidentical twins. For adoption studies, extraversion showed a higher heritability of approximately 50 percent than neuroticism (only approximately 15 percent heritable).

In a large study of Finnish twins, Viken et al. (1994) found a decrease in the heritability of both extraversion and neuroticism between late adolescence and the late twenties and no change in heritability thereafter. It should be noted that this study relied on cross-sectional analyses that compared the results for different birth cohorts. Therefore, the age effects may reflect cultural changes rather than age-related changes. Only longitudinal studies that follow the same persons over a long time can disentangle age-related changes from historical changes.

Overall, there is clear evidence for genetic influence on temperamental traits at all ages except close to birth, but the findings are far from being consistent between the twin and the adoption method. Possible reasons for these inconsistencies are discussed in Sect. 4. When the results are compared to the findings for other individual differences such as intelligence, conscientiousness, or agreeableness, the heritabilities are not higher (see Loehlin 1992, Plomin et al. 1997). Thus, variations in temperament do not seem to be more genetically influenced than other variations of personality and intelligence.

4. *Methodological Problems*

Both the twin method and the adoption method are affected by multiple problems that can bias their

results. For temperament-related traits, the main problems seem to be contrast effects and non-additive genetic influences. Contrast effects make twins and siblings more different, and can occur in two different forms. First, they can be due to sibling rivalry (e.g., one becomes more dominant and the other more submissive in their interaction, and these tendencies generalize to other relationships). In this case, twins and siblings are more different than one would expect on the basis of their genes and other environmental influences. Such contrast effects are expected to decrease with increasing age differences between the siblings; thus, they should operate particularly in the case of twins.

Second, contrast effects can affect personality judgments by family members because the parents or the siblings contrast a child with his or her siblings. For example, the parents may overestimate differences between two siblings' aggressiveness if one is moderately aggressive relative to peers and the other one is not aggressive because they perceive one as 'the aggressive one' and the other as 'the non-aggressive one.' These judgmental contrast effects are expected to increase with increasing dissimilarity of the siblings. Therefore, they are expected to underestimate most strongly the similarity of adopted children, and moderately strongly the similarity of biologically related siblings; for identical twins they should be particularly small.

Saudino et al. (1995) studied contrast effects for parental ratings of temperament. They compared the similarity of identical twins, nonidentical twins, siblings, and full, half, and genetically unrelated siblings in stepfamilies. This approach enabled estimation of the strength of contrast effects. They found significant contrast effects for all temperamental traits. These contrast effects may explain both the odd finding of more than 100 percent genetic variance for many temperamental traits in twin studies (the difference between identical and nonidentical twins is overestimated because the nonidentical twins are judged to be less similar than they really are) and the very low or even negative correlations for adopted siblings (they are judged to be less similar than they really are). It should be noted, however, that after controlling for contrast effects, significant genetic effects remained for all traits in this study.

The second main methodological problem is that the influences of different genes on a temperamental trait may not necessarily be additive. For example, the mere presence of the alleles A and B may not in itself increase the risk for high emotionality but the simultaneous presence of both alleles may. Because identical twins share all alleles, they share 100 percent of all genetic effects, including such non-additive effects, and adopted siblings share neither additive nor non-additive effects. However, nonidentical twins and siblings share 50 percent of the additive effects but a smaller percentage of the non-additive effects because these require that they share combinations of alleles. For example, the probability that they share alleles A *and* B is only 25 percent. If non-additive effects exist, the difference between identical and nonidentical twins estimates more than 50 percent of the genetic effects. Therefore, the heritability coefficient *over*estimates genetic influences. Similarly, the adoption method *under*estimates genetic differences because the difference between biologically related and adopted siblings estimates less than 50 percent of the genetic effects.

In the case of non-additive effects, the genetic estimates from the twin method and the adoption method are biased in opposite directions (overestimation vs. underestimation). This principle applies also to other problems such as the genetic similarity of the parents due to partner preferences (their homogamy), or a genetic similarity between adopted children and their nonbiological parents due to nonrandom assignment to families by the adoption agencies (selective placement). Therefore, it is possible to control these problems by the simultaneous analysis of similarity data from both twins and adopted children. This combination method requires complex statistical models (Neale and Cardon 1992) but circumvents most methodological problems of both the twin and the adoption method. It is even possible to estimate the strength of the various biases (e.g., separate estimations for additive and non-additive genetic effects). Thus, although there are many problems with these indirect estimations of genetic influences on temperament, most of them can be controlled.

5. *Misinterpretations of Heritability Estimates*

Often heritability estimates are misinterpreted as general estimates of 'the' genetic effect on temperament. In contrast, it was pointed out above that heritability estimates depend on (a) the specific temperamental trait, (b) the population of interest, (c) the historical period in its cultural development, and (d) the age of the participants. Another serious misinterpretation is the assumption that genetic effects on temperament can be altered only by changing the genes themselves through gene technology. This widely shared view is not correct because genes and environmental influences often interact in affecting personality. For example, the effects of pathological alleles such as the one that causes phenylketonuria can be minimized through environmental intervention (in this case: keeping a particular diet in the first years of life; see Plomin et al. 1997). Although the alleles that are relevant for temperament-related traits are as yet largely unknown, it is possible that their effects on temperament might be modifiable through diet, medication, specific parenting, or education. Heritability does not necessarily imply non-modifiability.

6. Conclusion

Indirect estimations of the overall genetic influence on temperament indicate significant genetic effects. The many methodological problems of these estimations can be controlled to a great extent through adequate study designs and statistical models. The estimates inform us whether temperamental differences in temperament among peers of a particular population are partly due to genetic differences. Because all temperament-related traits show significant genetic influence, at least in Western cultures beginning with later infancy, it seems promising to search for the specific genes that are responsible for these genetic influences.

See also: Behavioral Genetics: Psychological Perspectives; Developmental Behavioral Genetics and Education; Familial Studies: Genetic Inferences; Genetic Studies of Behavior: Methodology; Personality Development and Temperament; Temperament and Human Development

Bibliography

Eysenck H J 1967 *The Biological Basis of Personality*. Thomas, Springfield, IL

Goldsmith H H, Buss K A, Lemery K S 1997 Toddler and childhood temperament: Expanded content, stronger genetic evidence, new evidence for the importance of the environment. *Developmental Psychology* **33**: 891–905

Gray J A 1982 *The Neuropsychology of Anxiety*. Oxford University Press, Oxford, UK

Gray J A 1987 Perspectives on anxiety and impulsivity. *Journal of Research in Personality* **21**: 493–509

Halverson C F Jr., Kohnstamm G A, Martin R P (eds.) 1994 *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. Lawrence Erlbaum, Hillsdale, NJ

Loehlin J C 1992 *Genes and Environment in Personality Development*. Sage, Newbury Park, CA

Neale M C, Cardon L R 1992 *Methodology for Genetic Studies of Twins and Families*. Kluwer Academic, Boston, MA

Plomin R, DeFries J C, McClearn, G E, Rutter M 1997 *Behavior Genetics*, 3rd edn. Freeman, New York

Riese M L 1990 Neonatal temperament in monozygotic and dizygotic twin pairs. *Child Development* **61**: 1230–7

Rothbart M K, Bates J E 1998 Temperament. In: Eisenberg N (ed.) *Handbook of Child Psychology*, 5th edn. Wiley, New York, Vol. 3

Saudino K J, McGuire S, Reiss D, Hetherington E M, Plomin R 1995 Parent ratings of EAS temperaments in twins, full siblings, half siblings, and step siblings. *Journal of Personality and Social Psychology* **68**: 723–33

Strelau J, Angleitner A (eds.) 1991 *Explorations in Temperament*. Plenum Press, New York

Thomas A, Chess S, Birch, H G, Hertzig, M E, Korn S 1963 *Behavioral Individuality in Early Childhood*. New York University Press, New York

Viken R J, Rose R J, Kaprio J, Koskenvuo M 1994 A developmental genetic analysis of adult personality: Extraversion and neuroticism from 18 to 59 years of age. *Journal of Personality and Social Psychology* **66**: 722–30

Windle M, Lerner R M 1986 Reassessing the dimensions of temperamental individuality across the life span: The revised dimensions of temperament survey (DOTS-R). *Journal of Adolescent Research* **1**: 213–30

J. B. Asendorpf

Temporal Lobe

The temporal lobe subserves disparate functions. The dorsal (or superior) part is involved in the perceptual processing of auditory signals including speech. The primary auditory area is located at the central part of the lower bank of the lateral fissure, with higher

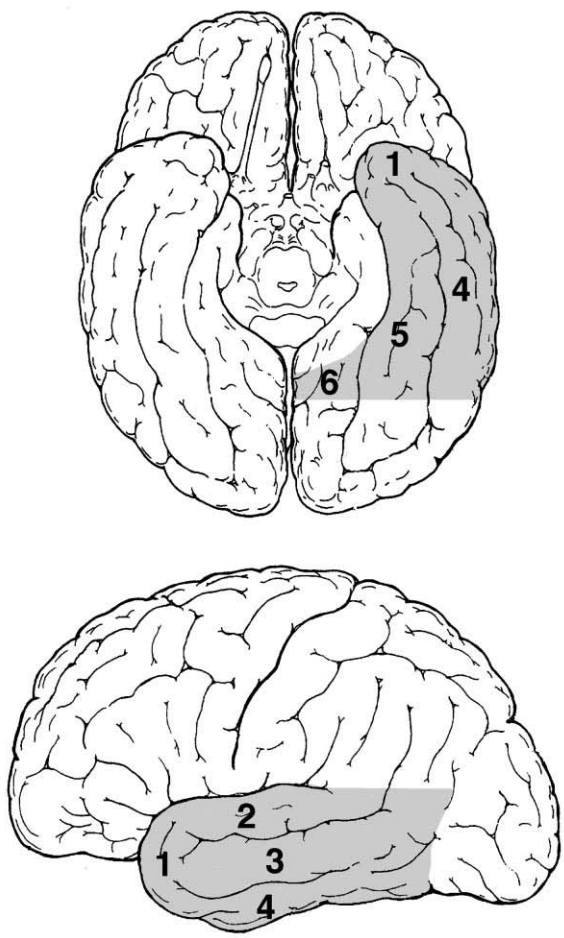

Figure 1
The extent and gyrus composition of the temporal lobe. 1, temporal pole; 2, superior temporal gyrus; 3, middle temporal gyrus; 4, inferior temporal gyrus; 5, fusiform gyrus; 6, lingual gyrus

auditory areas circumscribing it. Visual information processing about objects spreads from the ventral part of the occipital lobe to the posteroventral part of the temporal lobe. The most anterior and anteroventral parts are essential for the memory and retrieval of the semantic knowledge of objects.

1. Borders and Gyrus Composition

The temporal lobe is separated from the frontal lobe and the anterior part of the parietal lobe by the lateral (or Sylvian) fissure. Its borders with the occipital lobe and the posterior part of the parietal lobe do not have clear delimiting sulci. On the lateral surface, an imaginary line drawn from the temporo-occipital incisure (or the occipital notch) at the inferior margin of the hemisphere to the posterior end of the parieto-occipital sulcus defines the anterior border of the occipital lobe. Another imaginary line drawn orthogonally to this first line, from the point at which the lateral fissure turns upward, defines the border between the temporal and parietal lobes on the lateral surface. On the ventromedial surface, an imaginary line drawn from the temporo-occipital incisure to the point at which the parieto-occipital sulcus merges with the calcarine sulcus defines the border between the temporal and occipital lobes. The most medial part of the ventral surface of the hemisphere, medial to the anterior part of the calcarine sulcus and the anterior part of the collateral sulcus, is often treated as a part of the limbic system. The temporal lobe does not border with the parietal lobe on the medial surface.

The anterior end of the temporal lobe, the temporal pole, does not have sulci. The lateral surface of the temporal lobe has three anteroposteriorly elongated gyri separated by two sulci. Going from the more dorsal to the more ventral, these are the superior temporal gyrus, the middle temporal gyrus, and the inferior temporal gyrus. The superior temporal sulcus separates the superior and middle temporal gyri, and the inferior temporal sulcus separates the middle and inferior temporal gyri. The dorsal surface of the temporal lobe, which is concealed in the lateral fissure, is composed, from anterior to posterior regions, of the planum polare, the transverse temporal gyrus (or the transverse gyrus of Heschl), and the planum temporale. From its medial to lateral extent, the ventromedial surface of the temporal lobe is composed of the lingual gyrus, the fusiform gyrus, and the inferior temporal gyrus. The collateral sulcus separates the lingual and fusiform gyri, and the occipitotemporal sulcus separates the fusiform and inferior temporal gyri. The lingual gyrus continues anteriorly to the parahippocampal gyrus, which is the gateway to the hippocampus and related structures implicated in explicit memory (see *Hippocampus and Related Structures*). Both the lingual gyrus and the fusiform gyrus continue from the occipital lobe to the temporal lobe with no marks.

2. Auditory Functions

Because microelectrode recordings in human cortex are possible only in very rare clinical cases, and in vivo tracer studies are not possible, the current knowledge of the structure and function of the auditory cortical areas in the temporal lobe is largely due to results obtained in experimental macaque monkeys. Fortunately, the location and basic organization of auditory cortical areas is very similar in humans and macaques, insofar as the available data indicate.

2.1 Primary Auditory Cortex

The dorsal surface of the monkey temporal lobe does not have a transverse gyrus, but has an anteroposteriorly elongated annectant gyrus. This annectant gyrus in the ventral bank of the lateral fissure shows the histological features of primary sensory cortex such as densely packed granule cells, and dense reactivity to the cytochrome oxidase and calcium-binding protein parvalbumin. Studies with microelectrodes show that cells in this region respond well to pure tones, showing relatively sharp selectivity to a tone's frequency. These studies have permitted the identification of two separate tonotopically organized areas, named A1 and R. A1 occupies a large part of the annectant gyrus, and R occupies a smaller area anteriorly adjoining A1. Frequencies are mapped from lowest to highest in the anterolateral to posteromedial direction in A1, and the reverse in R. Both areas receive strong fiber projections from the main part (ventral division) of the medial geniculate nucleus (MGv). Although the histological features of primary sensory cortex are more pronounced in A1 than R, R is thought to be parallel to A1 in signal flow because responses of cells in R to pure tones survive ablation of A1.

In the human temporal lobe, a region with the histological features of primary cortex is located in the transverse temporal gyrus within the ventral bank of the lateral fissure. MEG (magnetoencephalography) and functional MRI (magnetic resonance imaging) studies consistently show tonotopy in the transverse temporal gyrus; frequencies are mapped from lowest to highest in the anterolateral to posteromedial direction. This mapping orientation is the same as in monkey A1. Some studies have also shown a second tonotopic mapping in the transverse temporal gyrus, which may correspond to R in the monkey. Bilateral damage to these primary auditory areas results in a profound deficit in sound detection, although the deficit usually recovers within several weeks to several months.

2.2 Higher Auditory Areas

The monkey auditory areas in the dorsal part of the temporal lobe are composed of three or four serial

stages (Kaas et al. 1999). The primary auditory areas constitute the 'core.' Axonal fibers from the core strongly project to regions in the 'belt,' which circumscribe the core in all directions except posteromedially. The belt regions are confined to the lower bank of the lateral sulcus. Fiber projections from MGv to the belt are weak, while there are moderate projections from the dorsal (MGd) and medial (MGm) divisions of the medial geniculate nucleus to the belt. There are at least three subdivisions in the belt as determined by distinct tonotopic mapping and histochemical and connectional properties. Cells in the belt tend to respond more strongly to narrow band noises and more complex sounds than pure tones, and their response tuning to pure tone frequencies tends to be broader than in the core (Rauschecker 1998). There is no consensus on the relative contribution of serial inputs from the core and direct inputs from MGd and MGm to the auditory responses of belt cells, but it is at least known that responses to pure tones in a part of the belt disappear after ablation of A1.

The belt strongly projects to the laterally adjoining regions on the lateral surface of the superior temporal gyrus. These regions have been named the 'parabelt.' The projections from the core to the parabelt are weak. Thus, there is a more or less serial projection stream from core to belt to parabelt. Like the belt, the parabelt receives projections from MGd and MGm, but in addition, the parabelt receives projections from other thalamic sites: the suprageniculate, limitans, and medial pulvinar nuclei. The parabelt projects extensively to more rostral and caudal parts of the superior temporal gyrus, adjoining regions in the superior temporal sulcus, prefrontal cortex, parietal cortex, and other brain sites. There are no projections from the core, and only weak projections from the belt, to these fourth-stage regions. Tonotopic mapping has not been found in the parabelt. The parabelt receives projections from the belt in a rostral-to-rostral and caudal-to-caudal manner. Thus, there are at least two serial pathways in parallel. There are some data on the cellular response properties that suggest that the two pathways correspond to the two main pathways in the visual system; the caudal auditory pathway corresponding to the dorsal, or spatial visual pathway, and the rostral auditory pathway to the ventral, or object visual pathway (Rauschecker 1998). Most cells in the rostral parabelt are more selective for sound properties, while cells in the caudal parabelt respond to pure tones of wide frequency ranges. Unlike rostral parabelt cells, most cells in the caudal parabelt, respond only to sounds from contralateral space. Some are more responsive to sounds from moving sound sources, sometimes with sensitivity to the direction of the source motion.

Available data in humans also suggests serial auditory stream (Kaas et al. 1999). Electroencephalogram (EEG) and MEG recordings of auditory evoked potentials show progressively longer response latencies at more distant locations from the primary auditory area. Functional MRI and PET studies have found a few tonotopic maps on the surface of the superior temporal gyrus. As for stimulus selectivity, functional MRI studies have found that complex sounds including human voices more strongly activate these surrounding regions than scrambled voices with matched power spectra (Belin et al. 2000, Scott et al. 2000). A bilateral lesion of the surrounding regions with the primary auditory areas spared is known to cause 'auditory agnosia,' a deficit in perceiving meaningful sounds including phonemes, elements of spoken words, and nonvocal familiar environmental sounds. Left-lateralized damage causes a specific deficit in perceiving phonemes (pure word deafness), while right-lateralized damage causes a less specific deficit in perceiving nonvocal, familiar environmental sounds, for which the term auditory agnosia is sometimes specifically used. In any case, a lesion of the surrounding regions leaves the detection of sounds intact. All these findings suggest that auditory information is processed serially in humans as it is in macaques—from the primary auditory areas located within the transverse temporal gyrus, to the surrounding regions at the planum temporale and the lateral surface of the superior temporal gyrus.

Apart from pure word deafness, there are patients who show a deficit in both the perception and vocalization of words. In the latter case, named 'Wernicke's aphasia,' spontaneous speech may be fluent but includes nonsense words or words with syllabic errors. These patients cannot repeat words or sentences spoken by other people. It is assumed that the phonetic memory of words or its access by phonemic inputs is damaged. The posterior third of the left superior temporal gyrus was originally proposed to be involved in this symptom, thus this region is often referred to as 'Wernicke's area.' However, the lesions in patients with permanent Wernicke's aphasia are larger and include the posterior half of the superior and middle temporal gyri or the white matter under Wernicke's area.

3. *Visual Functions*

In macaques, the middle and inferior temporal gyri are mostly dedicated to the purely visual associations involved in object vision. These regions are the latter stages of the ventral visual pathway. Their bilateral ablation causes a specific deficit in performing and learning tasks requiring visual object discrimination and recognition (Gross 1994). Single cell recordings have shown that cells respond to various complex features of visual object images (Tanaka 1996). Some of the cells specifically respond to views of faces.

In humans, however, purely visual functions appear to be restricted to the posterior part of the ventral temporal lobe. Both 'visual agnosia,' in which the

patient cannot retrieve semantic knowledge including the name of visually presented objects, and 'prosopagnosia,' in which the patient cannot retrieve the identity of visually presented faces, are caused by damage to the regions around the border between the occipital and temporal lobes on the ventral surface. These patients retain semantic knowledge of objects and individual persons, which they can retrieve via inputs of other modalities. More anterior parts of the ventral temporal lobe are thought to be essential for the semantic knowledge of objects and persons, with their names localized mostly in the left hemisphere.

Reading can be also specifically degraded (pure alexia) by damage around the border between the occipital and temporal lobes. Classical cases had combined damage to the left occipital lobe and the corpus callosum splenium (the posterior part of the corpus callosum), in which the deficit was explained by the disconnection of the language centers in the left hemisphere from the spared visual cortical regions in the right occipital lobe. However, a similar deficit is caused by a more localized damage to regions around the border between the occipital and temporal lobes in the left hemisphere. The patients have no deficit in writing words. Many patients can read words by picking up letters one by one. Therefore, the deficit is thought to be due to degraded perception of shapes composing words, especially global features covering the whole word. Among Japanese people, who use both visually complex ideograms (Kanji) and relatively simple syllabograms (Kana), there are rare cases in which reading Kanji is specifically degraded. The damage in these patients involve the most posterior part of the inferior temporal lobe in the left hemisphere.

Functional MRI studies have shown that the posterior part of the ventral temporal lobe is activated by the passive viewing of object images (Tanaka 1998). Parts of the fusiform gyrus are more strongly activated by viewing faces than other objects. The surrounding regions in the rest of the fusiform gyrus and the inferior occipital gyrus are more activated by non-face objects than faces. There are reports that views of towns, including houses, and views of rooms specifically activate the posterior part of the parahippocampal gyrus (Epstein and Kanwisher 1998).

4. *Semantic Knowledge*

There are aphasic patients with symptoms that appear to be due to damage or loss of the semantic knowledge of objects. These patients show impaired understanding of spoken words and sentences. Their spontaneous speech is fluent, but their vocabulary is poor. They also have difficulty in naming visually presented objects. Oral repetition of spoken words and sentences is intact, but without understanding. These patients have left-lateralized or bilateral damage of the middle and inferior temporal gyri and temporal pole. This symptom is a subtype of 'transcortical sensory aphasia,' but it is also called 'semantic dementia.' Semantic dementia is more often observed in 'Pick's disease,' a progressive genetic disease manifesting atrophy of cortical neurons, rather than vascular problems. The atrophy usually starts in both temporal and frontal lobes, but in the limited cases in which the atrophy starts in the temporal lobe, the patients show pure semantic dementia. The knowledge of objects at finer classification levels (individual and subordinate levels rather than categorical levels) tends to be lost first. As the infection progresses to the frontal lobe and the medial aspects of the temporal lobe, the symptoms develop into a more general dementia.

Semantic dementia may show category specificity (Caramazza and Shelton 1998). Three categories may be selectively impaired: animals, tools, and individual persons. Damage to the temporal pole impairs knowledge of familiar persons. In particular, the left temporal pole is important for the association of names with biographical knowledge, while the right temporal pole is important for the retrieval, based on various inputs, of biographical knowledge other than names. Damage to the anteroventral part of the left temporal lobe causes difficulty in naming animals based on various inputs. The conceptual and nominal knowledge of tools is degraded by damage to the posterior region of the lateral surface of the occipital, temporal and parietal lobe junction. Functional imaging studies (PET and functional MRI) show activation patterns in normal subjects consistent with the previously mentioned results obtained from brain-damaged patients (Damasio et al. 1996). In these studies, the subjects named objects from their pictures or made semantic judgments on pictures or visually presented words.

See also: Dementia, Semantic

Bibliography

Belin P, Zatorre R J, Lafaille P, Ahad P, Pike B 2000 Voice-selective areas in human auditory cortex. *Nature* **403**: 309–12

Binder J R, Frost M S, Hammeke T A, Cox R W, Rao S M, Prieto T 1997 Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience* **17**: 353–62

Caramazza A, Shelton J R 1998 Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience* **10**: 1–34

Damasio H, Grabowski T J, Tranel D, Hichwa R D, Damasio A R 1996 A neural basis for lexical retrieval. *Nature* **380**: 499–505

Epstein R, Kanwisher N 1998 A cortical representation of the local visual environment. *Nature* **392**: 598–601

Gross C G 1994 How inferior temporal cortex became a visual area. *Cerebral Cortex* **5**: 455–69

Kaas J K, Hackett T A, Tramo M J 1999 Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology* **9**: 164–70

Rauschecker J P 1998 Cortical processing of complex sounds. *Current Opinion in Neurobiology* **8**: 516–21

Scott S K, Blank C C, Rosen S, Wise R J S 2000 Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* **123**: 2400–6

Tanaka K 1996 Inferotemporal cortex and object vision. *Annual Review of Neuroscience* **19**: 109–39

Tanaka K 1998 Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology* **7**: 523–9

K. Tanaka

Tense, Aspect, and Mood, Linguistics of

The English sentences *He is asleep* and *He was asleep* differ in tense: the former is in the present and the latter in the past tense. In a similar way, *He opened the window* and *He was opening the window* are said to differ in aspect, and the verbs *wish* and *were* in the sentence *I wish I were dead* are said to differ in mood. Tense, aspect, and mood, traditionally labeled grammatical categories, play an important role in the grammar of English and most other languages.

It should be noted from the outset, however, that tense, aspect, and mood do not usually come neatly lined up as separate categories in grammars. Temporal, aspectual and modal properties are commonly interwoven in the same forms. Conversely, distinctions which would traditionally be thought of as belonging to the same category may be expressed in very different ways, for instance, by inflectional forms on the one hand and auxiliary constructions on the other. For this reason, many linguists prefer to see tense, aspect, and mood as one large domain within which various grammatical distinctions may arise.

Recent typologically oriented work has shown that although there is great variation among languages as to their systems of tense, aspect, and mood, the elements of those systems may be reduced to a relatively small number of types, each with its characteristic meaning and typical way of expression, developing historically by processes of grammaticalization. Such processes usually start with phrasal constructions involving lexical morphemes. Future markers may, e.g., develop out of constructions involving motion verbs such as 'come' and 'go,' or modal verbs with the meanings of obligation ('shall, must') and volition ('want'). Later on, inflections may arise through the coalescence of earlier free morphemes (auxiliaries or particles) with the stem of the main verb. The elements of a tense-aspect-mood system in a language may be at different stages in these grammaticalization processes; how far they have advanced determines both their form and their meaning. Thus, only a few of the types described below typically show up in inflectional distinctions, while the others tend to be expressed periphrastically.

The basis for the distinctions between forms traditionally subsumed under the category of tense is the time-dependence that characterizes the overwhelming majority of statements in natural languages. A sentence such as The *baby was asleep* does not make very much sense as long as we do not know what time we are talking about—that is, at what time the baby was asleep. In English, the tense of the verb gives us some indication about the time—we know that the speaker is referring to some point in the past. However, the exact location of that point is not indicated explicitly and has to be retrieved from the context. In many languages, e.g., Chinese, no distinction would be made between *was asleep* and *is asleep* here—in those languages, the identification of the 'topic time' or 'reference time' is left wholly implicit.

Nevertheless, the identification has to be made, in order to make sense of the statement. Thus, the way in which sentences are interpreted is similar whether or not a language makes the grammatical distinction between present and past tense. Notice also that tenses are not complementary to time adverbials, as is often assumed. Languages that lack tense distinctions do not, on the whole, use time adverbials more often than languages that do not. This is because time adverbials are used whenever the contextual temporal information is insufficient. Tenses, on the other hand, being more or less obligatory grammatical elements, are used whether or not the information they contain is necessary for the communication.

In the most straight-forward cases, the choice of tense depends on the relationship between the topic time and the time of speech. Time as experienced by a person is naturally divided into present, past, and future; these are also the three basic tenses of traditional grammar, although it is common for languages to have only a simpler, two-way distinction in their grammar: either between past and non-past or between future and non-future. In English, only the past tense (*he called*) differs from the present (*he calls*) morphologically; what is traditionally referred to as the future (*he will call*) is an auxiliary construction, although inflectional futures are far from uncommon in other languages (e.g., French *il dormira* 'he will sleep'). It is common for languages to have several ways of marking the future grammatically, often differing in subtle modal nuances (cf. English *will* vs. *is going to*). The common modal coloring of future tenses has led many scholars to deny the future status as a tense. Many languages distinguish grammatically degrees of remoteness in the past and sometimes also in the future. In many Bantu languages one thus finds 'today pasts,' 'yesterday pasts' and 'remote pasts' expressed by different verb forms or by auxiliaries.

A distinction is customarily made between absolute and relative tense, depending on whether the point of departure is the time of speech or some other time point. The tenses of finite verbs in main clauses are usually absolute—past tense means that the topic time

precedes the time of speech. Non-finite verb forms such as participles, gerunds and infinitives may also show tense distinctions, but these predominantly mark relations of simultaneity or temporal sequence relative to the time of the next higher finite verb, as in the sentence *She lived in London, working for a large company* where the gerundial clause is understood as being about the same time as the main clause (which, in its turn, is understood as preceding the point of speech). Finite verbs in subordinate clauses may follow varying principles. In clauses expressing reported speech, the verb is in many languages in the same tense as it would have in a direct quote; in English and some other European languages, the tense of such clauses depends on the tense of the matrix main clause. This phenomenon, exemplified in a sentence such as *John said that he was ill* (cf. *John said "I am ill"*) is traditionally referred to as 'sequence of tenses'; there is disagreement among scholars about its nature.

In some languages, special narrative tenses are used to express an event that takes place directly after another.

The sentences *He opened the window* and *He was opening the window* were said above to differ in aspect. Many traditional characterizations of aspect are formulated in terms of the perspective or viewpoint a speaker takes towards what happens or takes place in the sentence. Alternatively, we may say that aspectual distinctions involve different ways in which an event, process or state may be related to a point or period in time. Grammatical distinctions subsumed under the category of aspect are related to the different ways entities such as events, processes and actions are anchored in time, or, if we like, to the different perspectives we may take to these entities. In languages which express aspectual distinctions in their inflectional systems, the most common distinction is between perfective and imperfective aspect. Perfective aspect is used talking about completed events and processes and tends to be restricted to events that took place in the past, or sometimes either in the past or the future. Imperfective aspect, which is more independent of time reference, is used for states and processes that hold or are going on at the topic time but also for habitually occurring events.

A typical example of the intertwining of tense and aspect in an inflectional verb system is found in written French, in which perfective aspect is represented by the so-called passé simple or simple past (*il chanta* 'he sang'), while the present (*il chante* 'he sings/is singing'), the imparfait or imperfective past (*il chantait* 'he sang/was singing') and the future (*il chantera* 'he will sing') represent combinations of the imperfective aspect with various tenses.

Aspect may also be expressed by periphrastic or derivational means. Progressives, as exemplified by the English construction *I am/was/will be singing*, are a common type in languages and are almost exclusively expressed periphrastically. With regard to their meaning, they resemble imperfective aspect but have a more restricted domain of use in that they are only marginally used about states or habitual events. Historically, progressives are a common source of imperfectives.

Special habitual inflections are not uncommon; these may be used for events that take place regularly (*I used to visit him on Sundays*) but also sometimes for generic statements such as *Dogs bark*.

In many languages, the aspectual character of verbs may be modified by derivational processes. In contradistinction to inflectional aspect, these processes tend to be idiosyncratic and restricted to smaller sets of lexical items. A typical distinction expressed by derivational aspect is that between semelfactive or punctual on one hand and durative on the other, as in Hungarian *köhintett* 'he coughed once': *köhögött* 'he coughed (for a longer or shorter time).' The famous aspectual systems of Russian and other Slavic languages are derivational in origin but have over time been regularized and received many features that make them seem more like inflectional aspect.

The English perfect, as in *He has sung*, represents a cross-linguistically common type the semantics of which defies the traditional classification in that it cannot be unambiguously subsumed under either tense or aspect. Perfects are used of events that took place in the past but have a more restricted domain of use than ordinary past tenses. A common way of formulating the meaning of perfects is to say that they imply 'current relevance' of the event related in the sentence. Another way of putting this is to make a distinction between the 'time of event' and the 'topic time' (or reference time): in the perfect, the time of event is in the past but the topic time is equal to the time of speech. In the pluperfect (e.g., *He had sung when I arrived*) the time of event also precedes the topic time but here both points are in the past. Within the semantic domain typically covered by perfects, we also find more specialized forms and constructions. Resultatives express the resulting state of some event (e.g., *He is gone* in English); they are usually restricted to verbs that denote change. Historically, they are often extended in meaning to become perfects. Experientials, such as the Japanese *koto ga aru* construction (*Nihon de hataraita koto ga aru* 'I have worked in Japan') are used to express that some type of event took place one or more times within a period in the past.

The distinctions pertaining to the traditional grammatical category of mood cover only a relatively small part of the wide conceptual domain of modality. Moods are commonly said to express the speaker's attitude to a proposition or to its truth-value. However, a typical grammatical mood, such as the subjunctive in English, French, or German, is rather used to mark certain clause types, which are normally non-asserted and grammatically subordinate. Thus, in French, the subjunctive is obligatory in the complement clauses of verbs meaning 'want, desire, hope.'

The subjunctive thus marks these clauses off against plain main clauses, which are the unmarked, indicative mood (*Il est parti* 'He has left' vs. *J'espère qu'il soit parti* 'I hope that he has left'). Another common function of moods is to distinguish between possible and counterfactual situations in conditional constructions (cf. *If he is ready*, *we can go* vs. *If he were ready*, *we could go*, where the latter sentence contains the obsolescent subjunctive form *were*).

A form usually subsumed under mood is the imperative, whose primary use is to express speech acts such as commands, requests etc.

Many languages make grammatical distinctions on verbs which carry information about the kind of evidence the speaker has for a statement.

See also: Grammatical Gender; Grammatical Number; Grammatical Voice; Linguistics: Overview; Morphology in Linguistics

Bibliography

Bertinetto P M, Bianchi V, Dahl Ö, Squartini M (eds.) 1995a *Temporal Reference, Aspect and Actionality Vol 2; Typological Perspectives*. Rosenberg & Sellier, Turin, Italy

Bertinetto P M, Bianchi V, Higginbotham J, Squartini M (eds.) 1995 *Temporal Reference, Aspect and Actionality, Vol. 1: Semantic and Syntactic Perspectives*. Rosenberg & Sellier, Turin, Italy

Binnick R I 1991 *Time and the Verb: A Guide to Tense and Aspect*. The University of Chicago Press, Chicago

Bybee J L, Perkins R, Pagliuca W 1994 *The Evolution of Grammar. Tense, Aspect, and Modality in the Languages of the World*. The University of Chicago Press, Chicago

Comrie B 1976 *Aspect. An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge University Press, Cambridge, UK

Comrie B 1985 *Tense*. Cambridge University Press, Cambridge, UK

Contini-Morava E 1989 *Discourse Pragmatics and Semantic Categorization. The Case of Negation and Tense-aspect with Special Reference to Swahili*. Mouton de Gruyter, Berlin

Dahl Ö 1985 *Tense and Aspect Systems*. Blackwell, Oxford, UK

Dowty D 1979 *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Reidel, Dordrecht, The Netherlands

Hopper P (ed.) 1982 *Tense-aspect: Between Semantic and Pragmatics*. Benjamins, Amsterdam

Klein W 1994 *Time in Language*. Routledge, London and New York

Montague R 1974 *Formal Philosophy*. Yale University Press, New Haven, CT

Partee B 1984 Nominal and temporal anaphora. *Linguistics and Philosophy* **7**: 243–86

Prior A 1967 *Past, Present, and Future*. Clarendon Press, Oxford, UK

Reichenbach H 1947 *Elements of Symbolic Logic*. Collier-Macmillan, New York

Smith C 1991 *The Parameter of Aspect*. Kluwer, Dordrecht, The Netherlands

Tedeschi P, Zaenen A (eds.) 1981 *Syntax and Semantics 14 Tense and Aspect*. Academic Press, New York

Vendler Z 1967 *Linguistics in Philosophy*. Cornell University Press, Ithaca, New York

Verkuyl H 1972 *On the Compositional Nature of Aspect*. Reidel, Dordrecht, The Netherlands

Ö. Dahl

Territoriality: Geographical

Territoriality in humans, or human territoriality, is the act of delimiting and controlling an area of space—a territory—in order to control people and things. Human territoriality pervades our actions at all geographic scales, and undergirds socio-cultural power. It is a strategy that can be turned on and off. This, and the fact that human territoriality requires particular conceptions about space, make its use in humans different from its use in animals, where territoriality appears as an instinct that is often associated with aggression. Conceiving of human territoriality as an offshoot of animal instinct (Ardrey 1966, Montagu 1968) has resulted in misconstruing the problem. The most fruitful approach to human territoriality is through its relationship to non-territorial strategies; its connections to geographical place and space; and the advantages (and disadvantages) it provides.

1. Geographical Strategies of Control

Consider a hypothetical example of a father and his two small children at home in their kitchen. The father is trying to cook a meal and take care of the children, and they want to help. The father enjoys their company but they are too young to help, and run the risk of injury as they try to touch a knife or dish. The father wants to limit their access to these things, and he has two geographic strategies. One is for the father to rush around the kitchen placing the knife and dish out of reach and to restrain the children physically. Here the father intends to alter or control their spatial actions directly. The other strategy is indirect and territorial. It arises if the parent decides simply to restrict the children's access to things in space by telling them they 'may not enter the kitchen at this time.' That is, the kitchen for now is 'off limits.' Here the children's access to things is being limited by the parent's assertion of control over an area. (Sack 1986, p. 16) In the case, the area or territory was there to begin with; the parent took the opportunity to invoke it, and, after the meal is prepared, the territorial unit may recede into the background. Alternatively, the parent could

have created a new territory by delimiting and asserting control over part of the kitchen. Whether creating a new territory or invoking the power of a pre-existing one, human territoriality is a strategy that can be turned on and off, and when it is not used, we then rely on direct spatial interactions to affect or control the flows and access to things in space.

This example points out that territoriality is found in everyday life and at even the smallest geographic scales. But territoriality is most commonly understood to involve political activity and state control, and leads to concerns of sovereignty and jurisdiction. At an in-between scale, we find territoriality to be conjoined with such things as zoning, or with (often private) ownership of land. These and countless other instances of the control of area are examples of territoriality; not to recognize this is to overlook how territoriality provides humans with advantages that make it a necessary geographical expression of power. These advantages are outlined in Sack's theory of human territoriality (1986).

2. *Theory of Human Territoriality*

The theory builds on the definition of human territoriality 'as the attempt by an individual or group to affect, influence, or control people, phenomena, and relationships, by delimiting and asserting control over a geographic area [where the area is called the territory]' (p. 19) by pointing to three interrelated implications of this definition that reveal how and why territoriality works as an effective strategy. They are that territoriality:

(a) involves a form of classification by area, as opposed to classification by kind;

(b) that it must contain a form of communication by boundary; and

(c) it must rely on power to enforce control over the area. In shortened form, territoriality involves classification, communication, and control.

2.1 *Classification by Area*

Classification by area refers to the fact that when we assert control of an area, we are also using the area as a means of delimiting and defining the things we wish to control. Territorial control could mean that anything and everything inside the area is under someone's or some group's control. The things controlled do not have to be defined or enumerated. The father can tell his children to leave the kitchen or to not cross this particular threshold; in so doing he would be asserting control over any and everything within the area without telling the children what it is he does not want them to touch. In contrast, a non-territorial form of control would have to enumerate or disclose the objects and relationships that are to be controlled.

2.2 *All Territorial Units are Bounded*

The boundary may be thin and precise or thick and vague—more like a zone of transition than a precise line—and boundaries can be more or less porous. Some boundaries let very little in and out, while others allow a great deal to flow through. Regardless of its form, the idea of a boundary must be successfully conveyed to others for territoriality to work. This is because the boundary acts as the primary means for communicating territoriality. The father would not be able to use territoriality in the kitchen if the children were not old enough to know what a boundary means.

2.3 *Territorial Control*

For territoriality to work, the territory must be backed up by authority and power that are recognized and respected. If the signs of the boundary and the authority backing it up are not recognized, then territoriality would vanish and we would have to employ non-territorial direct spatial interaction. If the children were to ignore the father and not leave the kitchen, or promptly re-enter it, the father would have little choice but to pick up the children and remove them, or remove all of the knives and plates.

Each instance of territoriality incorporates all three facets, but the effectiveness of territoriality stems from the fact that one or more of these provides advantages. For example, not enumerating what it is that one wishes to control can be advantageous when we do not know all of the things to which we will need access, or when we do not wish to, or cannot, disclose what we want to control. Not being able (or not wanting) to enumerate the things we need is common for most of our activities. Even a relatively simple place like a building contains mixes of conditions taken for granted such as light, heat, and acoustics, which may be just right for the task but which could not have been enumerated before hand. This is even more the case for complex and open-ended activities such as politics, and explains why political behavior is often the most territorial of all. Territorial sovereignty provided by the nation state is an effective strategy of controlling innumerable and potentially important things for political processes that are open-ended.

Communicating through a boundary is simpler than communicating through a variety of symbols, each indicating a particular relationship to people and objects. In the case of the father in the kitchen, the children may be too young to understand the difference between a sharp and a blunt knife, and so there may be no choice but to use the boundary.

In the case of using power, it may be more efficient to enforce territorial control than to follow all of the interactions between people and objects. Fences and 'no trespassing' signs are often more efficient to deter transgressions than is the non-territorial strategy of following people around. Similarly, it may save effort

to have prisoners behind bars than to assign guards to follow them.

These three aspects of territoriality, which are found in all cultures and at all scales, provide advantages that can be taken up as reasons to use territoriality. However, different combinations of the three generate further advantages which only occur in particular historical contexts. For example, the territorial effect of defining things by area can impose new forms of social relationships when territory is used to classify and control people. An empire or a modern state may establish an administrative district which is then used to tax and conscript. People who before had nothing in common are now part of a new social entity. Colonies and post-colonial nation states are more complex examples. These uses of classification by area to define social relations and identities—a territorial definitions of social relations—is often contrasted with a social definition of the people and the place (Soja 1971).

Another complex consequence of territoriality is that it can lead to a container view of place or space. That is, once a territory is created, and if its contents have not been stipulated, or if its most important contents are not yet in place, it comes to be seen as an 'empty' space, even though there are still things in it. A room without occupants or furniture is 'empty,' even though it still contains air, heat and other aspects of nature. In this way, territoriality can help us to think of space as though it were content-less or empty. This container view of space reinforces and is reinforced by modern mapping which has us think of space as a geometric structure—a series of coordinates that describe space itself, independent of things, and it reinforces the idea that space can be bought, sold, and held as empty lots or parcels for speculation.

These advantages of territoriality can turn into disadvantages when conditions are not right. If the father employed territoriality too often, the children may never learn what it is they are not supposed to do. Stipulating what the right conditions are for taking up a territorial strategy is part of the theory of territoriality.

3. *Territory, Place, and Space*

Rooms, buildings, farms, factories and nation states are all territories and they are also places. What then is the distinction between territory and place? Place is a broader term, but one of its meanings merges with that of territory. 'Primary place,' as it is termed in a theory of place developed in *Homo Geographicus* (Sack 1997) or 'place-as-territory' is virtually the same as territoriality, though the two stress different qualities. Territoriality emphasizes how humans are served by boundaries that control what is in and out of place, while place-as-territory (or primary place) stresses how place is necessary for projects to be undertaken, and so focuses on the way place helps mix nature and culture to promote projects. Both point out that the boundaries and degrees of porosity of place, and the rules of what may or may not take place, affect the flows of things through space. These spatial flows and interactions then are generated and affected by territorial rules; and these rules of 'in and out' are in turn affected by the flows. The idea then of a socially constructed space should be understood to mean not that we construct physical space, but that the territories (and the places-as-territories) and the flows among them are created by us.

4. *Territory, Transgressions, and Identities*

Part of the effectiveness of territories in everyday life is that when they function according to plan, they often recede into the background and are taken for granted. Since territoriality supports and instantiates power, it is important to understand 'who' or 'what' group is in charge. A challenge to power through some sort of territorial transgression can make the territory reappear, drawing attention to and calling into question, the entire fabric of social conventions and power (Cresswell 1996). Sometimes this transgression seems to be playful or even without purpose, as in cases of graffiti. More deliberate transgressions of territorial rules for the sake of disclosing and changing power are found in pickets and sit-ins. Extreme cases of transgression are found in violations of territorial sovereignty and in acts of war.

5. *Territoriality, Globalization, and Scale*

Territories are always being transformed, dissolved, and created anew. However, modern conditions of a world economy and the electronic transmission of information have led to speculation that territories in general, and national territories in particular, are being radically transformed and may even be obsolete (Morley and Robins 1995). Instead of places and interactions through space, perhaps the proper description might be that of a constantly shifting set of nodes and networks. Many believe however, that national territories may not be facing wholesale dissolution (Newman and Paasi 1998). In the global economy they provide varying mixes of opportunities for capital and thus help mute economic crises (the 'spatial fix' according to Harvey 1982). The same mix of opportunities may also be provided by a constantly changing hierarchy of territorial scales within the state (a 'scaler fix' as in Brenner 1998, though he does not use this term.) Others see the nation state as remaining a vital means of expressing cultural and ethnic autonomy. Perhaps the greatest and most persistent challenges to territorial state sovereignty do not come from the economy but from ideas about world community and a common humanity that encourages individuals to care for distant strangers.

Whatever is the fate of nations, countless places-as-territories will exist and continue to be constructed because we cannot undertake projects without them. For now, the nation state is the most powerful one. But to ignore the fact that its powers to contain are permeable and changing is to fall into what Agnew (1994) calls the 'territorial trap.' We run the risk of falling into a different territorial trap when we ignore the power of territories and places-as-territories to mold events, for we might then observe that all kinds of actions seem to have bounded and spatially accordant geographies while being ignorant of the fact that they are shaped by a common territorial structure.

Bibliography

Agnew J 1994 The territorial trap: The geographical assumptions of international relations theory. *Review of International Political Economy* **1**: 53–80

Ardrey R 1966 *The Territorial Imperative*, 1st edn. Anthenium, New York

Brenner N 1998 Global cities, Global states: Global city formation and state territorial restructuring in contemporary Europe. *Review of International Political Economy* **5**: 1–7

Cresswell T 1996 *In Place/Out of Place: Geography, Ideology, and Transgression*. University of Minnesota Press, Minneapolis, MN

Graham S 1998 The end of geography or the explosion of place? Conceptualizing space, place and information technology. *Progress in Human Geography* **22**: 165–85

Harvey D 1982 *The Limits to Capital*. University of Chicago Press, Chicago, pp. 431–8

Montagu M F A (ed.) 1968 *Man and Aggression*. Oxford University Press, Oxford

Morley D, Robins K 1995 *Spaces of Identity: Global Media Electronic Landscapes and Cultural Boundaries*. Routledge, London

Newman D, Paasi A 1998 Fences and neighbours in the postmodern world: Boundary narratives in political geography. *Progress in Human Geography* **22**(2): 186–207

Sack R D 1986 *Human Territoriality: Its Theory and History*. Cambridge University Press, Cambridge, UK

Sack R D 1997 *Homo Geographicus: A Framework for Action, Awareness, and Moral Concern*. The Johns Hopkins University Press, Baltimore, MD

Soja E W 1971 *The Political Organization of Space*. Association of American Geographers, Commission on College Geography, Washington, DC

R. D. Sack

Terrorism

The term 'terrorism' refers to the systematic use or threat of violence to communicate a political message rather than defeat an opponent's military forces. Thus the targets of terrorism are symbolic. Victims of terrorism represent a wider audience. To achieve maximum shock effect, terrorist violence is usually dramatic and provocative. Typically small numbers of extremists who otherwise lack the capacity to challenge those in power resort to terrorism. However, the term is also used to describe clandestine violence in support of the state or by the state. A defining characteristic of terrorism is that its users expect rewards that are out of proportion to both the resources they possess and the risks they assume. Terrorism is, furthermore, a strategy that is not restricted to any particular ideology.

1. Historical Development of Terrorism

As a strategy of resistance to the modern state, terrorism emerged some half century after the French Revolution, when the term originated as a description of the state regime of terror. In the late nineteenth century, 'terrorism' became a strategy of opposition. Russian revolutionaries and anarchists in France, Spain, Italy, and Germany established terrorism as a central mechanism in attempts to overthrow established regimes, most of which were autocratic. In 1881 the assassination of the Czar of Russia horrified European governments while it inspired revolutionaries and nationalists from Ireland to India. Terrorism in Russia focused on selective assassinations of key political leaders. Anarchists introduced the concept of 'propaganda of the deed,' the idea of an outrageous action, such as a bomb thrown indiscriminately into a public gathering, that would frighten the ruling classes and arouse the masses. Irish republicans added to the repertoire by organizing attacks in London, outside the immediate theater of conflict. This formative era of terrorism ended with World War I, sparked by the assassination of the Austrian archduke by a Serbian nationalist.

In the interwar period right wing extremist movements adopted terrorist tactics in their struggles for power. The period is best known, however, for the practice of totalitarian terror from above in Germany and the Soviet Union. After World War II, terrorism figured prominently in some national liberation struggles of the 1950s and early 1960s, such as in Cyprus and in Algeria. As the colonial era ended, social scientists expected terrorism to disappear with it. Yet at precisely this moment terrorism was transformed into a major international phenomenon. In the late 1960s, Latin American revolutionaries and Palestinian nationalists independently conceived of a new tactic of terrorism: the seizure of hostages in order to compel governments to concede to their demands. Violence escalated when foreign states who supported these local regimes, such as the USA, also became targets. Hijackings of aircraft in the Middle East and diplomatic kidnappings in Latin America, for example, became routine. The 1972 seizure of Israeli athletes at the Munich Olympic games demonstrated

the immense publicity value of hostage-taking as well as the spillover of Palestinian terrorism to Western Europe.

By the mid-1970s, left wing and separatist terrorism came to pose serious internal security problems for liberal democracies, especially in Western Europe. In Germany and Italy, for example, terrorism emerged in the wake of student protest movements. In the Basque regions of Spain and in Northern Ireland, ethnonationalist organizations mounted prolonged separatist campaigns, which lasted to the end of the century. Outside of Europe, India and Sri Lanka also faced chronic terrorism from secessionist ethnic minorities in the 1980s (continuing through the 1990s in Sri Lanka). Furthermore, in Latin America, after revolutionary terrorism had ended in Argentina, Brazil, and Uruguay, the Shining Path movement rose to challenge the regime in Peru in the 1980s.

The Iranian revolution of 1979 and the Israeli invasion of Lebanon in 1982 stimulated the growth of yet another type of terrorist actor, as radical anti-Western Islamic groups formed in Lebanon, Palestine, Egypt, and, by the end of the decade, Algeria. The seizure of US hostages in Lebanon, as well as the advent of massive vehicle bombs, such as the attack on the American Marine barracks in Beirut in 1983, continued to focus attention on terrorism in the Middle East. Palestinian groups also remained active, for example, in the hijacking of the Achille Lauro cruise ship in 1985.

State-sponsored terrorism became another new issue of the 1980s, initially because of the Iran hostage crisis. In 1986, the USA, convinced that Libya was responsible for terrorism against Americans, retaliated with military force. This in turn provoked Libya to take revenge in the mid-air bombing of Pan Am 103 in 1988.

Terrorism continued to shock governments and their publics in the 1990s. High-casualty terrorism in the United States was an unprecedented development. Its origins were both domestic, as individuals associated with the US far right bombed the Oklahoma City federal building, and international, with Islamic militants organizing the bombing of the World Trade Center in New York. A second emphasis in the 1990s was the fear of terrorism involving weapons of mass destruction, a fear inspired not only by the 1995 sarin gas attack on the Tokyo subways by a religious cult but also by the prospect of the proliferation of nuclear, chemical, and biological weapons. In the Middle East, radical Islamic groups used terrorism to oppose the peace process initiated between Israel and the Palestine Liberation Organization (PLO), while the Gulf War and the subsequent US military presence in Saudi Arabia inspired Islamic militants to attack American interests. In 1998 simultaneous bombings of the US embassies in Kenya and Tanzania provoked the USA to retaliate militarily against targets in the Sudan and Afghanistan.

2. *Explanations of Terrorism*

Interest in the study of terrorism has paralleled the historical evolution of the phenomenon. Until the 1970s, little was written on the subject. Typically terrorism was treated briefly as the first stage of internal war, to be abandoned as insurgents moved to guerrilla warfare and then on to conventional war. By the 1990s, the literature contained thousands of specific citations to terrorism, including contributions by historians, political scientists, economists, legal scholars, criminologists, sociologists, psychologists, philosophers, and anthropologists. Nevertheless, there is no comprehensive unifying theory of terrorism. Nor is terrorism a central element in major theories of war and conflict. The importance of the concept is determined by the political significance of terrorist actions, rather than by intellectual context.

Terrorism is a contested concept that resists precise definition. Since the term is both elastic and emotionally powerful, it lends itself to subjective interpretations driven by political rather than analytical purposes. It is also difficult to distinguish terrorism from other forms of violence, such as, for example, guerrilla warfare or criminal activity. Must the victims of terrorism be noncombatants, for example? If so, are attacks on military targets acts of terrorism? If terrorism is defined in terms of the intentions behind the action, is it possible to know those intentions? What is the relationship between religion and terrorism? Furthermore, terrorism occurs so frequently and in so many different historical and cultural settings that it is difficult to make comparisons or draw general conclusions. Analysts of the subject even dispute basic questions of fact, such as whether or not terrorism is becoming more lethal or whether it is more likely in democracies.

2.1 *Why Does Terrorism Happen?*

Terrorism occurs in situations of conflict. Divided societies, repressive governments, economic inequalities, or ideological rivalry among states, for example, create the grievances that generate terrorism. In other cases, conditions are permissive. Democracies, for example, may be susceptible to terrorism because they tolerate dissent and allow an open news media. Yet few of the individuals who are dissatisfied or even actively engaged in social protest movements move on to participate in terrorism. Thus understanding the causes of terrorism requires knowledge of psychological motivations, at the individual or small group level. Research points to the dynamics of group relationships rather than specific personality types or mental pathologies. Once established, underground groups may struggle primarily to maintain group solidarity and cohesion. They bind together individuals with different backgrounds, skills, and levels of political commitment.

At the same time, terrorism is purposive behavior, a conscious strategy to gain attention, attract recruits, mobilize popular support, and provoke government repression. The policy of no concessions adopted by the US government in the early 1970s rests on the premise that terrorism is a rationally chosen strategy: if governments deny reward and guarantee punishment, terrorism will decline. When governments blame the news media for publicizing terrorism, they are assuming that terrorism would not exist without the reward of media attention.

The use of terrorism by states against their external enemies may also be related to inequalities of power that force the weaker party to be deceptive rather than open. Internal regimes of terror are due to the desire not only to crush but to forestall resistance. It is worth noting that insurgent movements may also govern by terror in territories they hold.

Terrorism also depends on opportunity, especially the availability of targets and of resources. In a modern world of industrialized states, with sophisticated transportation and communication networks and dense patterns of social, political, and economic interaction, both targets and weapons are in plentiful supply. Government efforts to defend specific points of vulnerability, such as civil aviation or embassies in high-threat countries, may produce a displacement effect, as terrorists substitute easy targets for hard ones. The development of the modern news media, especially television, permits terrorists to reach a worldwide audience almost instantaneously.

2.2 Forms of Terrorism

Bombings constitute the most common form of terrorism. Since the late nineteenth century and the invention of dynamite, terrorists have targeted public buildings and crowded public spaces. Although most bombings do not cause casualties, terrorists are capable of organizing spectacular high-casualty violence, facilitated by technological improvements in both explosives and timing devices. However, the capacity for high-casualty bombings does not depend on sophisticated technology; extremely powerful devices can be constructed with common fertilizer, for example. The same principle applies to potential terrorism using weapons of mass destruction: simple techniques could still be quite deadly.

Armed attacks on unsuspecting civilian targets, such as airline passengers or tourists visiting historic sites, are also part of the terrorist repertoire. Over time assassinations of individual government officials became less frequent, perhaps because of the availability of alternatives or the inaccessibility of public figures, who are now heavily guarded. Nevertheless, the second half of the twentieth century includes the assassinations of important leaders; Anwar Sadat, Indira Gandhi and Rajiv Gandhi, and Yitzhak Rabin number among them. (Not all assassinations can be defined as terrorism; the attack must be linked to an organized conspiracy or oppositional movement.)

The seizure of hostages is an explicit way of bargaining directly with governments. The tactic includes hijackings of airliners and even cruise ships, individual or group kidnappings, and armed takeovers of buildings such as embassies. Hostage seizures as a form of international terrorism had become less common by the end of the twentieth century, but kidnappings have continued, especially in domestic conflicts such as the ongoing insurgency in Colombia.

2.3 The Effects of Terrorism

Does terrorism succeed? It attracts the attention of the news media and the public, and the issue ranks high on the agendas of many governments and international organizations. However, does recognition of the threat or even of the grievances behind terrorism produce favorable political outcomes? Despite widespread publicity, the political effectiveness of terrorism is generally short-term and limited in scope. Terrorism may intimidate jurors or dissuade tourists, but it cannot compel popular majorities to give in to minority demands. Both the Irish Republican Army and the PLO eventually renounced terrorism. Terrorism polarizes opinion by mobilizing support and hardening opposition. In this way, it can inflame existing conflicts and disrupt peace processes, but it rarely changes fundamental attitudes.

See also: Violence, History of; Violence in Anthropology; Violence: Public; War, Sociology of

Bibliography

Crenshaw M (ed.) 1995 *Terrorism in Context*. Pennsylvania State University Press, University Park, PA

Della Porta D 1995 *Social Movements, Political Violence, and the State: A Comparative Analysis of Italy and Germany*. Cambridge University Press, Cambridge, UK

Gilbert P 1994 *Terrorism, Security and Nationality: An Introductory Study in Applied Political Philosophy*. Routledge, London

Hoffman B 1998 *Inside Terrorism*. Columbia University Press, New York

Laqueur W 1987 *The Age of Terrorism*. 1st Am. edn. Little Brown, Boston

Merkl P H (ed.) 1986 *Political Violence and Terror: Motifs and Motivations*. University of California Press, Berkeley, CA

Nacos B L 1994 *Terrorism and the Media: From the Iran Hostage Crisis to the World Trade Center Bombing*. Columbia University Press, New York

Reich W (ed.) 1998 *Origins of Terrorism: Psychologies, Ideologies, Theologies, States of Mind*. Johns Hopkins University Press, Baltimore, MD

Wieviorka M 1993 *The Making of Terrorism*. University of Chicago Press, Chicago

M. Crenshaw

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Test Administration: Methodology

The meaning of an individual's score on an educational or psychological test depends on the test being administered according to the same specified procedures, so that every test-taker faces the same situation. A test administered under specified conditions is said to be standardized. Without standardization, scores from the test will be less accurate, and comparability of test scores will be uncertain. Details of administration depend on the format and content of the test.

1. *Standardization*

Standardized directions help to insure that all test takers understand how much time will be allowed to complete the test, whether it is permissible, or advisable, to omit responses to some test items, how to make the required responses, and how to correct inadvertent responses. Multiple choice items merely require the test taker to choose among a set of alternative answers. Constructed response items may require the test taker to write a short phrase or sentence as a response, or may require a more extended written response. Performance assessments will often require the test taker to perform some task, possibly using some equipment, under standardized conditions, with the performance evaluated by a test administrator, or recorded for later scoring. In general, a quiet environment, free of distractions, will promote appropriate test performance.

Instructions for paper-and-pencil tests are simple; multiple choice items may require the test taker to mark the answer on a separate answer sheet whereas short-answer items require writing a response on the test paper. Test takers are more likely to need practice in responding to tests that are administered by a desktop or similar computer. Items are displayed on the computer display screen, and the test taker responds by means of computer keys, a 'mouse' or similar indicating device. In some instances, the test-taker is asked to respond to constructed response items by typing a word or short phrase with a computer keyboard.

Test takers may need practice in navigating a computer-based test. Usually, the screen shows one item at a time, and does not proceed until some response has been made. One of the response options may be to skip or postpone the item, in which case the computer will offer the item again after the remaining items in a section have been answered. Sometimes the test taker has the option of flagging the item for later reconsideration. Item review, which cannot be controlled on paper-and-pencil tests, is more complex with computer administration and may not be permitted. Test takers prefer to have the option, but those who make use of it seldom change their previous response. When a change is made, however, it is more often a change from wrong to right than from right to wrong. The net effect of item review is an occasional score improvement, weighed against time spent in making the review. When the test is timed, and if further new items are available, the candidate would, in general, be better off attempting the new items rather than returning to items that have previously been answered.

Computer administration permits the use of items that are not feasible with paper-and-pencil tests. Most involve some aspect of temporal control or measurement. Short-term memory can be tested by presenting some material, and then asking questions about it after it is removed from the display. Some aspects of response time can also be measured. Also, test material can include moving displays, including short segments of televised scenes or activity.

1.1 *Accommodations and Test Modifications*

Persons with disabilities may be at a disadvantage in the testing situation. Persons with limited vision may have trouble seeing the test material; others with motor impairment may have difficulty making the required responses. Special test administrations may be offered to accommodate such persons (Willingham et al. 1988). Printed tests with large type, or large fonts on the computer display may help persons with limited visions. A person or a computer system can read verbal items to the test taker, although items with diagrams and graphs pose more challenging problems of communication. Persons with certain motor disabilities may need help in making the appropriate responses. Persons with attention deficits, dyslexia, or similar disorders may easily be granted additional time on timed tests.

When persons are tested under nonstandard conditions, the validity of the score becomes suspect. Much validity evidence is empirical, and depends on standardized test administration. If only a few persons are offered a certain accommodation, it becomes difficult to determine whether their test score has the same meaning as a score from a standardized test. Most such evidence requires substantial numbers of cases in order to be evaluated. Opinions differ about the suitability of using test scores from accommodated tests. In large-scale operations, a score from a special accommodation may be marked, or 'flagged' to indicate its special status. It may be deemed appropriate to indicate the nature of the accommodation, but not the nature of the disability, although it will often be impossible to report one without revealing the other. In some venues, revealing the nature of a person's disability may be seen as a violation of individual rights, and may be illegal.

Persons whose primary language is not the language of the test will probably be at a disadvantage in taking the test. In some circumstances, it may be appropriate

to provide the test in their primary language, although test translation is not an easy matter, and requires extensive test development. An intelligence test or a questionnaire about health may elicit more valid answers if translated. On the other hand, an employment test for a position where a certain language is used as a part of the job would be appropriately given in the language of the job. (For more detail on this and other topics see American Educational Research Association et al. 1999.)

2. *Test Security and Cheating*

Test scores are often used to compare individuals with each other, or with established performance standards, with important consequences to the individual test taker. Some test takers may attempt to gain an advantage by obtaining knowledge of the test questions before the test is given. Thus, procedures for safeguarding the test materials before, during, and after the testing are an important aspect of test administration. Generally, test takers are forbidden from taking notes; any scratch paper or other material that they may have used during the test must be disposed of.

Obtaining advanced knowledge of items is but one way of gaining an advantage over other test takers. Test administrators may be asked to check identities through photographs, fingerprints, or other means of personal identification to preclude the possibility of a candidate arranging for someone else to take a test in their stead.

A variety of statistical checks may be made of test responses in an effort to detect various forms of cheating. If several persons are found to have exactly the same responses to many of the items, further investigation may be warranted. A person who answers many easy questions wrong, but who correctly answers many difficult items has, to say the least, an unusual pattern of responses, triggering further checking. If a test is offered on several occasions, and a person has taken the test more than once, large differences in the person's scores from one time to the next would be suspicious.

2.1 *Test Preparation*

Testing programs with high stakes have spawned a number of organizations offering special tutoring in how to take the test. Such short term preparation ('coaching') may involve having clients take pseudotests that contain items like those that might be expected on the actual test. Methods are taught for coping with particular types of items. Sometimes, a variety of suggestions are offered that are not content-related, but that stress how to detect the correct alternative by special clues, such as 'the longest alternative is probably the correct answer'. Studies of the effectiveness of these commercial test preparation programs have had mixed results, but the best estimate is that the advantage of a two- or three-week course is from about 0.1 to about 0.3 standard deviation units, which is roughly equivalent to one or two additional questions answered correctly. Longer preparation has a larger effect; multiweek courses amount to further instruction in the subject of the test, and may more properly be viewed as education.

Large testing programs, concerned lest commercial preparation courses give well-to-do candidates a special advantage, have made preparation materials available to all test candidates. This may simply be older, retired forms of the test, together with general test-taking advice. Many also offer, for small cost, computer software to provide the sort of instruction that the 'coaching' schools provide. Still, some candidates put themselves at a disadvantage by not using such material, being overly confident of their own proficiency.

3. *Test Scoring, Norms and Performance Standards*

Test standardization extends to test scoring. Responses to paper-and-pencil tests can readily be scored by computer. Constructed responses can sometimes be scored by computer, but these responses usually are scored by human scorers. In some instances, computer programs are being used to score extended written responses and essays. Preparing a detailed scoring rubric, which spells out the detailed criteria by which responses are to be evaluated, can facilitate consistent scoring, by human or computer. Human scorers can be trained on the use of the rubric, and are monitored through a variety of checks, including occasional rescoring of some responses, and statistical comparisons.

Test scores usually consist initially of simply counting up the number of correct responses, or summing the scores given to items when the responses are not scored simply 1 or zero. Sometimes, a penalty is imposed for incorrect answers, in an attempt to correct for the possibility that a test taker might mark the correct response to a multiple-choice item by chance. Despite voluminous evidence, no definitive statement can be made as to whether the penalty for wrong responses improves or harms the validity and reliability of the resulting scores. Omitted responses are generally given no credit, i.e., treated the same way as wrong responses (see *Classical (Psychometric) Test Theory*).

Scores based on item response theory involve combining information from item responses by a more complex statistical method. In effect, more credit is earned for correct answers to more difficult questions, and more penalty is imposed for incorrect answers to

easier questions (an exception is the scoring of fixed-length nonadaptive tests using the one-parameter, or Rasch, IRT model, in which case the number correct is a sufficient statistic for estimating the person's proficiency).

Test scores have little meaning without additional information. Norms are statistical distributions of test scores made by defined groups of test takers, and provide one normative standard of comparison. External performance standards can be set with reference to test content. Some tests are designed to provide scores understandable only by professionals; with only broad categories of performance being revealed to the test taker (Angoff 1971).

4. *Test Equating and Calibration*

Well-established testing programs generally have several equivalent forms of the test. Each test form has different items, but all forms are constructed according to the same set of test specifications, or test blueprint, which indicates topics to be covered, item difficulty, item formats, and other aspects of item content. Still, test development procedures cannot insure score comparability; some adjustments are necessary in the scores from a new form to make them comparable to scores from old forms. The statistical process of making scores comparable is called test equating (Feuer et al. 1999). Linear equating involves giving two forms of the test to equivalent groups of test takers, and then adjusting the score distributions to have the same mean and standard deviation. Equipercentile equating uses the same data but adjusts the score distributions to have the same shape, so that for any given test score, the same proportion of test takers earn scores that are below that score value in both tests. Anchor test equating involves administering a common segment of along with both tests, to form the statistical link.

5. *Adaptive Testing*

The use of computers to administer tests has led to the development of computer-based adaptive testing (CAT) (Wainer 1990, Sands et al. 1997). In CAT, the selection of items presented to a test-taker is shaped by the test-taker's own responses to the test. Normally, the first question is of medium difficulty. Those responding correctly to the item will be given a more difficult item; those giving a wrong response will get an easier item. The same principle continues until a fixed number of items has been administered, or until some other criterion is reached. The net result is that each person is asked questions appropriate to his or her own proficiency level. Time is not wasted asking persons of low proficiency items that are too difficult for them, and high scorers are not asked questions that are too easy for them. The level of accuracy of scores on a traditional test can be achieved in a adaptive version of the test with from 30 percent to 50 percent fewer items.

The idea of adaptive testing can be traced to Binet, the originator of the intelligence test. Some current tests of intelligence for children follow a similar procedure, adapting the level of question to the demonstrated ability of the test taker to respond correctly. Such tests are administered individually, so the test administrator can readily select subsequent items according to a specified protocol. Individual administration is impractical with large-scale testing; group administration is more efficient, but all test takers must take the same test. The use of a computer for administration means that tests can again be tailored to the test taker, without the need for a trained administrator.

CATs get favorable evaluations from test takers. In a CAT, each candidate faces items that are challenging but not overwhelming. Persons who usually do poorly on ability tests are surprised at how easy the test appeared to be. The better candidates find a CAT harder than they expected, but engaging.

5.1 *CAT Procedures*

Different CAT systems use different procedures for administering a CAT. Algorithms differ in how the first item is chosen, how subsequent items are chosen, when to stop the test, and how to score the test. With different test takers receiving tests of different difficulty, the score cannot be the traditional 'number-right' of paper and pencil tests, since most test takers get about the same number of items correct on a well-designed CAT. Current CAT's rely on item response theory (IRT) for scoring the test responses, as well as for choosing items to present. IRT is based on the premise that all the items are measuring the same underlying proficiency. While each item is to some extent unique, the items also share some common elements. The IRT model accounts for item interrelationships with a model that specifies the form of the relationship of each item to the dimension of proficiency being measured.

CAT requires a large set of items, called collectively an item pool, or item bank. Each item is characterized by one or more parameters indicating the item's difficulty, strength of relationship to the proficiency being assessed, and propensity of low proficiency test takers to guess correctly the right answer. The item characteristics, called 'item parameters are estimated statistically from data obtained from administering the items to a group of test takers that serves as the reference group. Since an item bank contains from 200 to 1000 or more items, all items cannot be administered to a single group of persons. Instead, an elaborate design can be used to administer the items to equiv-

alent groups, and to calibrate all the items to a common scale. The statistical calculations necessary to score a CAT can be done in the blink of an eye on modern computers, so a test score is available as soon as a person has completed taking the test. Of course, if constructed response items are included, the computer program must include an algorithm for scoring such responses immediately. Such algorithms are rapidly becoming the state-of-the-art.

Item selection in CAT is usually done one item at a time, but may also be done by selecting groups or items, in stages. Selection in general depends on a preliminary estimate of the test taker's proficiency, based on the item responses already made. The next item, or group of items, is selected to be maximally informative, given the current performance. Often, many other factors must also be considered. Traditional tests that are prepared professionally conform to a detailed set of test specifications, having mainly to do with item content. Such specifications are especially important for achievement tests, which measure the extent of mastery of some domain of knowledge. Each constructed test is expected to have the same balance of content and coverage of topics. By using codes for various content and format characteristics of an item, the item selection algorithm for a CAT can strive to select items so that the content specifications are met as well as possible, as well as finding items of the appropriate difficulty for the test taker.

With CAT, the problem of security is daunting, because persons are scheduled to take the test at different times. Groups of test takers can team up to each remember a few items, and later pool their memorized items. The remembered items can then be offered, possibly for profit, to persons who have yet to take the test. To diminish the effect of foreknowledge of a few items, CATs may be furnished with very large item pools, so that the chance of encountering any particular item will be very small. More elaborate strategies involve using several different item pools, formed from very large collections of items ('item vats'). Studies show that the effect of foreknowledge of a few of the items in a large pool will, on average, provide very small improvement in test scores.

See also: Classical (Psychometric) Test Theory; Computerized Test Construction; Psychometrics; Test Anxiety and Academic Achievement; Test Bias

Bibliography

Angoff W H 1971 *Scales, Norms, and Equivalent Score*. Educational Testing Service, Princeton, NJ

American Educational Research Association, American Psychological Association & National Council on Measurement in Education 1999 *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC

Feuer M J, Holland P W, Green B F, Bertenthal M W, Hemphill F C (eds.) 1999 *Uncommon Measures: Equivalence and Linkage Among Educational Tests*. National Academy Press, Washington, DC

Sands W A, Waters B K, McBride J R 1997 *Computerized Adaptive Testing: From Inquiry to Operation*, 1st edn. American Psychological Association, Washington, DC

Wainer H 1990 *Computerized Adaptive Testing: A Primer*. L Erlbaum Associates, Hillsdale, NJ

Willingham W W, Ragosta M, Bennett R E, Braun H, Rock D A, Powers D E (eds.) 1988 *Testing Handicapped People*. Allyn and Bacon, Boston

B. F. Green

Test Anxiety and Academic Achievement

In contemporary, achievement-oriented societies, test anxiety is a ubiquitous phenomenon. Test anxiety is educationally important because it may influence learning, achievement, health, and personality development. Among a number of emotions relevant for students' academic development, test anxiety has been studied most often. Exam-related anxiety had already been discussed by psychoanalytical writers in the 1930s. Large-scale empirical research started at the beginning of the 1950s (Mandler and Sarason 1952), and since then, more than 1,000 studies have been published (Hembree 1988, Zeidner 1998).

1. Concepts And Measures of Test Anxiety

Test anxiety may be defined as anxiety which subjectively relates to tests and their consequences (the term 'test' being used in a broad way and including exams as well). Typically, being afraid of failing a test, and of the consequences of failing, is at the heart of this emotion. Test anxiety can be experienced both before and during the taking of a test. Seen from the perspective of trait-state conceptions of anxiety, such anxiety is a state emotion when it relates to a specified test and is experienced before or during that test. As it is experienced habitually over extended periods of time and for dispositional reasons, test anxiety is a situation-specific type of trait anxiety. Trait test anxiety may be an important part of a student's personality.

The term 'achievement anxiety' has sometimes been used interchangeably with 'test anxiety.' Achievement anxiety, however, may also relate to situations normally not defined as being tests (e.g., situations at work or in sport which, like tests, imply the possibility of failure, but do not imply test-taking). Anxiety related to such situations may be similar to test anxiety. An example is competitive anxiety in sport.

At the beginnings of empirical research, test anxiety was defined as being one-dimensional. Early measures such as Mandler and Sarason's (1952) Test Anxiety Questionnaire assumed unidimensionality. However, the items of these measures already referred to distinct affective, cognitive, and physiological components of test anxiety, and early factor analyses repeatedly showed that these scales were made up of more than one factor. Therefore, Liebert and Morris (1967) proposed to distinguish affective-physiological components (which they called 'emotionality') from cognitive components (which they called 'worry'). Since then, a number of additional distinctions have been introduced, including I. G. Sarason's (1984) differentiation of affective tension, bodily symptoms, worry, and test-irrelevant thinking. Today, there is common agreement that test anxiety typically comprises (a) affective tension (feeling uneasy, nervous, and tense); (b) physiological activation (including symptoms such as increased heartbeat, sweating, etc.); and (c) cognitive worries relating to threat of failure, to the consequences of failure, to a lack of personal competence, etc. Contemporary measures of test anxiety like I. G. Sarason's (1984) Reactions-to-Tests Questionnaire and Spielberger's (1980) Test Anxiety Inventory (TAI) reflect these distinctions by comprising subscales relating to different test anxiety components.

2. *Test Anxiety, Learning, and Academic Achievement*

2.1 *Empirical Evidence*

The effects of test anxiety on learning and achievement have been studied in a large number of laboratory experiments and nonexperimental field studies (cf. Hembree 1988, Zeidner 1998). Experimental laboratory research has shown the following:

(a) dispositionally test-anxious subjects often perform less well at complex and difficult cognitive tasks than their less anxious counterparts. They tend to perform equally well or even better at tasks characterized by low levels of complexity and difficulty;

(b) the detrimental effects of dispositional test anxiety are more pronounced when an experiment is announced as being test-like and involving ego threat (e.g., when it is declared as being an intelligence test);

(c) state test anxiety experienced when performing complex and difficult tasks is also negatively related to performance. The relation is stronger for cognitive components of test anxiety (e.g., worries) than for affective-physiological components.

In line with experimental findings, test anxiety and academic achievement tend to correlate negatively in normal student populations. Average correlations reported in two meta-analyses were $r = -0.29$ (in grade 4 through postsecondary education; Hembree 1988), and, for mathematics achievement, $r = -0.23$ (Schwarzer et al. 1989). However, the range of correlations across studies also included zero and positive coefficients. Correlations did not differ significantly as a function of gender (Schwarzer et al. 1989). However, they were near zero at the beginning of schooling and then increased over the elementary school years (Hembree 1988).

2.2 *Theoretical Explanations*

One traditional explanation for this pattern of empirical evidence is that cognitive components of test anxiety consume cognitive resources (attention and working memory capacity): worrying about impending failure prevents one from concentrating on the task. This would explain why performance at complex and difficult tasks needing such resources is reduced, whereas performance at easy tasks may be left unaffected. Attention-deficit models using such assumptions have been part of early theories of test anxiety (e.g., Mandler and Sarason 1952) and continue to be one dominant theoretical paradigm.

More recent alternative explanations refer to effects of anxiety on other mediators of achievement, and to alternative ways of causation. Beyond cognitive resources, anxiety may affect students' motivation and learning strategies underlying their achievement. For example, anxiety may reduce interest in the subject matter and intrinsic motivation, since anxiety may be incompatible with positive affective engagement implied by intrinsic motivation. On the other hand, anxiety may well produce extrinsic motivation to invest effort in order to avoid impending failures. Simultaneous operation of such opposing mechanisms would imply that the motivational effects of anxiety on overall motivation and resulting effort may be ambivalent. In addition, effects on learning strategies may be complex as well. For example, anxiety may impede the use of creative cognitive and metacognitive strategies such as elaboration and organization of learning material or flexible time management. On the other hand, it may facilitate the use of more rigid strategies such as simple rehearsal (Pekrun 1992a, Pekrun and Hofmann 1999).

Beyond effects of anxiety on achievement, anxiety/achievement relations may also be produced by the effects of low achievement and low ability on the development of test anxiety. For example, experiences of failures may induce test anxiety, which would imply that negative anxiety/achievement relations may be produced by the effects of achievement on anxiety, instead of the reverse (developmental model; cf. Meece et al. 1989, Pekrun 1992b). In addition, performance deficits in test-anxious students may be caused by

reduced ability, implying that test anxiety might be a reflection of low ability instead of being causal in producing performance decrements (skill deficit model; Everson et al. 1989).

These different explanations may be regarded as being complementary instead of mutually exclusive. In many students, test anxiety may be generated by low ability and failures in the first place. Performance deficits in such test-anxious, low-ability students may then simultaneously be caused by low ability and by high test anxiety implying attention problems, motivational problems, and low-quality learning styles. This would imply multiple mechanisms mediating the effects of anxiety on achievement as well as the reciprocal causation of anxiety and achievement over time.

A number of studies have shown that effects of anxiety on achievement may indeed be due to different cognitive and motivational mechanisms (cf. Pekrun and Hofmann 1999, Zeidner 1998). Multiple mechanisms implying ambivalent cognitive and motivational effects may also explain why some studies found zero or positive correlations between test anxiety and achievement instead of negative relations (Hembree 1988, Schwarzer et al. 1989). Finally, the results of a small number of longitudinal field studies corroborate that anxiety and achievement may in fact be linked by reciprocal causation over the school years (Meece et al. 1990, Pekrun 1992b).

3. Development and Antecedents of Test Anxiety

3.1 Developmental Courses

In many students, the development of achievement-related anxiety can be traced back to the preschool years. Psychodynamic assumptions suggest that high parental expectations and negative parental judgments producing feelings of dependency, guilt, aggression, and anxiety may play a major role at this age (S. B. Sarason et al. 1960). Nevertheless, in early elementary school average levels of test anxiety are still relatively low. Mean values increase over the elementary years, which is congruent with emerging experiences of cumulative scholastic failures in many students, and to the related decline of average self-concepts of ability (Hembree 1988). Simultaneously, correlations between anxiety and achievement tend to become stronger during this time period, indicating that factual performance and affective experiences begin to be linked as suggested by developmental models of anxiety/achievement relations.

After the elementary years, average levels of test anxiety seem to be fairly stable throughout secondary school and college. Nevertheless, there may still be significant positive or negative developments in individual students, thus indicating developmental plasticity. One important mechanism is individual transitions between schools or classes implying changes of relative achievement positions within reference groups, which in turn affect the level of achievement-related anxiety (cf. Zeidner 1998).

3.2 Individual and Educational Antecedents

Empirical research has identified a number of antecedents which can be made responsible for the development of test anxiety (cf. Pekrun 2000, Zeidner 1998). Important individual antecedents are: (a) expectancies of failure, attributions of failures to low ability, and negative self-concepts of ability; and (b) an exceedingly high subjective sense of the importance of failures. The typical test-anxious student believes that it would be of the utmost personal importance to prevent failure at school, but that they are probably not able to do so, implying threats to self-esteem and future life.

Parent, teacher, and peer behaviors influencing such failure-related beliefs and, thereby, the development of test anxiety are: (a) high achievement expectancies and pressure for achievement; (b) competition within classrooms; (c) feedback of failures (as implied, for example, by low grades); and (d) punishment after failure. Parental and teacher support, on the other hand, might theoretically be assumed to be protective factors, but have repeatedly been shown to be uncorrelated with students' test anxiety. This paradoxical finding may be due to the ambivalent effects of support, which may imply hidden messages of pressure for achievement in addition to offering help. Finally, at the level of educational institutions and systems, important factors leading to anxiety may be competitive goal structures, social comparison standards of giving feedback, and educational career structures implying negative long-term consequences of early failures.

4. Therapy and Educational Prevention of Test Anxiety

In numerous studies it has been shown that test anxiety can successfully be alleviated in many students. Test-anxiety therapy may aim at: (a) reducing the experience of anxiety directly (e.g., by methods of relaxation); (b) breaking up habitual ties between situational triggers and the experience of anxiety (as in methods of behavioral therapy such as systematic desensitization); (c) changing belief structures such as expectancies, self-concepts, and achievement values underlying test anxiety (cognitive therapy); and (d) building up competences to learn and perform more efficiently in order to attain success (skills training). Studies suggest that treatments combining different approaches may be most efficient (cf. Zeidner 1998).

In contrast to therapy, ways of prevention have been studied less often. However, prevention might be both more humane and financially more effective than letting test anxiety develop in the first place and then

trying to alleviate it by individual therapy. Judging from the available developmental evidence, important educational strategies suitable for the prevention of test anxiety may be: (a) reduction of excessive demands for achievement; (b) the use of mastery-oriented, individual, or cooperative standards of grading instead of (or in addition to) competitive standards, thus giving low-ability students the opportunity to experience success; (c) adequate help instead of punishment after failure; and (d) adequate design of educational career structures.

Changing parent and teacher behavior, designing educational systems, and fostering school cultures along such lines may produce complex, multiple effects, some of them perhaps unwanted. Therefore, more long-term field studies exploring ways of preventing negative affect in students, as well as fostering positive emotions, are needed.

5. *Directions for Future Research*

Future test-anxiety research should explore in more detail the multidirectionality of causal pathways, and the multiplicity of mediators linking anxiety to learning and achievement. Furthermore, individual differences concerning mediators and causal pathways should be explored, thus producing evidence on students' individual affective life instead of relying on sample statistics as in most test anxiety research to date (Pekrun and Hofmann 1999). In addition, more studies on educational strategies to prevent test anxiety are needed.

Finally, research on students' academic emotions has thus far focused on test anxiety. As a result, not much is known about students' negative emotions beyond anxiety (such as anger, boredom, or hopelessness in achievement situations), or about their positive emotions (enjoyment of learning, hope for success, etc.; cf. Weiner 1985, Pekrun 2000). In many students and academic situations such emotions are probably no less important than anxiety. For example, deactivating negative emotions such as boredom or hopelessness may be even more detrimental to academic achievement than test anxiety has been proven to be, and positive emotions such as enjoyment of learning may be of primary relevance for students' intrinsic motivation and long-term engagement in learning. Therefore, future investigators should transcend test-anxiety research by analyzing other learning-related and achievement-related emotions as well.

See also: Academic Achievement Motivation, Development of; Motivation, Learning, and Instruction; School Achievement: Cognitive and Motivational Determinants; Schooling: Impact on Cognitive and Motivational Development

Bibliography

Covington M V 1992 *Making the Grade*. Cambridge University Press, New York

Everson H T (ed.) 1997 Emotions, cognition and performance.*Anxiety, Stress and Coping* **10**: 1–104

Everson H, Millsap R E, Browne J 1989 Cognitive interference or skills deficit: An empirical test of two competing theories of test anxiety. *Anxiety Research* **1**: 313–25

Hembree R 1988 Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research* **58**: 47–77

Liebert R M, Morris L W 1967 Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports* **20**: 975–8

Mandler G, Sarason S B 1952 A study of anxiety and learning. *Journal of Abnormal and Social Psychology* **47**: 166–73

Meece J L, Wigfield A, Eccles J S 1990 Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology* **82**: 60–70

Pekrun R 1992a The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology—An International Review* **41**: 359–76

Pekrun R 1992b Expectancy-value theory of anxiety: Overview and implications. In: Forgays D G, Sosnowski T, Wrzesniewski K (eds.) *Anxiety: Recent Developments in Self-appraisal, Psychophysiological and Health Research*. Hemisphere, Washington, DC, pp. 23–41

Pekrun R 2000 A social cognitive, control-value theory of achievement emotions. In: Heckhausen J (ed.) *Motivational Psychology of Human Development*. Elsevier, Oxford, UK, pp. 143–63

Pekrun R, Frese M 1992 Emotions in work and achievement. In: Cooper C L, Robertson I T (eds.) *International Review of Industrial and Organizational Psychology*. Wiley, Chichester, UK, Vol. 7, pp. 153–200

Pekrun R, Hofmann H 1999 Lern- und Leistungsemotionen: Erste Resultate eines Forschungsprogramms [Emotions of learning and achievement: First results of a program of research]. In: Jerusalem M, Pekrun R (eds.) *Emotion, Motivation und Leistung*. Hogrefe, Göttingen, Germany, pp. 247–67

Sarason I G (ed.) 1980 *Test Anxiety*. Erlbaum, Hillsdale, NJ

Sarason I G 1984 Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology* **44**: 929–38

Sarason S B, Davidson K S, Lighthall F F, Waite R, Ruebush B K 1960 *Anxiety in Elementary School Children*. Wiley, New York

Schwarzer R, Seipp B, Schwarzer C 1989 Mathematics performance and anxiety: A meta-analysis. In: Schwarzer R, Van der Ploeg H, Spielberger C D (eds.) *Advances in Test Anxiety Research*. Swets and Zeitlinger, Lisse, The Netherlands, Vol. 6, pp. 105–19

Schwarzer R, Van der Ploeg H, Spielberger C D (eds.) 1982–92 *Advances in Test Anxiety Research*. Swets and Zeitlinger, Lisse, The Netherlands, Vols. 1–7

Spielberger C D 1980 *Test Anxiety Inventory: Preliminary Professional Manual*. Consulting Psychologists Press, Palo Alto, CA

Weiner B 1985 An attributional theory of achievement motivation and emotion. *Psychological Review* **92**: 548–73

Zeidner M 1998 *Test Anxiety: The State of the Art*. Plenum Press, New York

R. Pekrun

Test Bias

Of all of the issues in modern testing and assessment, none generates more controversy and heated debate than the issue of 'bias' in testing. Tests are important instruments of public policy. They figure centrally in determining a child's readiness for schooling, in placing children into classes for the educable mentally retarded and into classes for the academically gifted. They have been used for years in European countries and in Japan to determine the class of occupations that students will enter as adults. Tests have become important in the United States in determining who will graduate from high school. For almost a century, professionally developed measures of scholastic aptitude have played an important role in determining who will be admitted to what colleges and who will be employed in what jobs. In the past half-century, tests have played an increasingly determinative role in certifying individuals to practice their chosen profession.

All of the above are serious test uses involving 'high stakes' decisions for individuals. Small wonder that when social, cultural, or ethnic minorities are assigned pejorative classifications or denied life opportunities on the basis of tests, the charge of unfairness and discrimination arises.

Although test bias is discussed generally in the context of bias against cultural and ethnic minorities in a given society, the concept (as well as the methodologies used to investigate it) is quite generic. Following conventional usage, in this article the group against whom a test is suspected of being biased will be referred to as the 'focal group.' The group with which test results of the focal group are compared will be called the 'reference group.' Thus, if a test is suspected of being biased against female examinees, they become the focal group, and males become the reference group. If the test is suspected of being biased against an ethnic minority (e.g., African-Americans), they become the focal group, and European-Americans become the reference group.

1. *What Test Bias is Not*

It is sometimes useful and instructive, although somewhat inefficient, to begin the examination of a concept by describing what it is *not*. On a test of upper body strength, a random sample of men will almost surely have mean scores that are higher than a random sample of women. It would be foolish to label such a test 'biased against women.' In fact, such a difference in performance between the sexes so accords with what we know to be true, that were no such performance difference observed, we would hold the test suspect. A less obvious though equally apt example is the persistent finding in many societies of differences among cultural and ethnic groups on measures of cognitive functioning, scholastic ability, and academic achievement. The existence of such differences, some of which are substantial, is interpreted by many to indicate bias in the measures themselves.

To appreciate the fallacy of such an inference, one need only consider a random sample of, say, African-American and White students in the city of Chicago. When measures of academic achievement are administered to these two groups, the typical finding is that White students as a group score significantly above their African-American counterparts. Again, to conclude that the test itself is therefore biased against African-American students would be erroneous. One would have to systematically rule out a plethora of rival causes of mean score differences (and there are many) before it could be safely concluded that the test itself was in some way deficient.

2. *Internal Sources of Test Bias*

There are many definitions of 'test bias' (cf., Flaugher 1978, Jensen 1980) and in the limited space available we will be able to discuss in some depth only the major sources of concern. We do not cover the extensive literature on possible sources of test bias emanating from individually administered tests of intelligence, such as administrator/examinee differences in race or dialect, nor such factors as differences in motivation and test anxiety. We will restrict treatment here to sources of test bias that are traceable to aspects of the measuring instrument itself and to sources of bias resulting from the use of tests in selection and prediction.

2.1 *Bias as Unintended Multidimensionality*

Most paper and pencil, multiple-choice tests are intended to measure a more or less unidimensional construct. To be sure, this intention is rarely ever met. Few tests actually measure a single ability. Even a test of simple addition involves multiple abilities at some atomistic level. What is meant my 'unidimensional' is that only one ability contributes to score variance. 'Reading comprehension,' for example, entails a certain minimum vocabulary and the ability to decode words. But in most applications these abilities do not contribute significantly to score variance because the vast majority of examinees are sufficiently adept at lower-level skills that such differences as exist do not contribute to test score variance.

A familiar and common example of bias as multidimensionality, also known as construct bias, is a test of quantitative reasoning that contains word problems. For examinees whose first language is not the language in which the test is written, the test may well reflect differences in familiarity with the language of the test, rather than the intended construct, quantitative reasoning. Bond (1981) provides a vivid example of bias as unintended multidimensionality in the verbal domain. Imagine a teacher who wished to assess the verbal analogical reasoning of her class of 14- and 15- year-olds. She constructs an analogies test that contains an unusually large number of terms familiar to rural students, say, but relatively unfamiliar to urban students. A typical item on the test might be:

pig:sty::chicken:_____
(a) barn (b) table (c) turkey (d) coop (e) dinner

Because students raised on a farm are more likely to know the meanings of the words in a given analogy, they would be at a tremendous advantage over urban students on such a test. The typical urban student is penalized because of his or her unfamiliarity with the very words that make up the analogy. For these students, the test is as much a test of vocabulary as it is one of analogical reasoning. For rural students, the test is a 'purer' measure of analogical reasoning, uncontaminated by large differences in exposure to the words composing the analogies.

Bond noted that as a practical measurement matter, it is impossible to separate 'vocabulary' from 'verbal analogical reasoning.' Conceptually, however, the two constructs are different. Although there is no way to avoid this problem in constructing a test, there are ways to control statistically for such differences. The procedure involves giving a separate test of vocabulary, then comparing only those students on the analogies test who have the same vocabulary scores.

An often overlooked source of unintended multidimensionality arises from response processes that a test is intended to assess and those that are actually employed by examinees. For example, if a test is intended to assess mathematical reasoning, it is important to ascertain whether examinees are, in fact, reasoning about the material instead of following a standard, taught algorithm. Different groups may well be differentially exposed to solution routines, or they may systematically employ different strategies of unequal effectiveness. Laboratory studies of the response process by subgroups of the population can assist in determining the extent to which capacities irrelevant or ancillary to the construct may be differentially affecting performance.

Before turning to other conceptions of test bias, brief mention must be made of the continuing debate over the relationship between bias and what has come to be called the 'IQ controversy.' Many people believe that if tests of 'intelligence' are shown to be free of construct bias (i.e., they measure the same construct(s) in different groups), then there is strong evidence for the notion that mean differences on such tests have a substantial genetic component. This is an unfortunate and misguided notion. The demonstration that a test is measuring essentially the same or similar attributes in two groups has no *a priori* implication for the source of any group differences that emerge. The many possible reasons for such differences (including differences in exposure, opportunity to learn, motivation, and so on) must be investigated as a separate matter.

2.2 Bias as Differentially Functioning Items

Not all items on a test are of equal overall quality, reliability, and validity. As noted earlier, the wording of some items may systematically confuse examinees who do not speak, as their first language, the language in which the test is written. In multiple-choice tests, clues as to the correct choice may be inadvertently introduced, so that some items may put certain examinees at a relative advantage over others. In achievement tests, some examinees may have received specific instruction on a particular procedure, while others received no such instruction. Such items are said to exhibit differential item functioning, or DIF, a neutral term that the measurement profession now prefers over the more value-laden term, 'item bias.' (The astute reader will note that one source of DIF is unintended multidimensionality on a smaller scale.)

The methodological and statistical machinery to identify items that function differentially for different groups has reached a state of considerable sophistication over the past 20 years (see Holland and Wainer 1993). In this article, three of the more common approaches (chi-square, Mantel–Haenszel, and item response theory) will be described. Of historical interest, since it was the first attempt to formalize a procedure for identifying biased test items, is the method based on relative item difficulty known as the delta method of Angoff and Ford (1974). The logic underlying the method is that even if groups differ in their overall mean on a test, the rank order of item difficulties within the focal and reference group should be approximately the same. The method involves an inverse normal transformation of item difficulties within the two groups of interest. (The transformation is an inverse one because of the unfortunate historical practice of defining item difficulty, p, as the proportion of examinees who respond correctly to the item, rendering it an item 'easiness' index!). Specifically, the delta (Δ) for the ith group for the jth item is $\Delta_{ij} - 4\pi_{ij} + 13$, where π_{ij} is the z score that cuts off p_{ij} in the unit normal distribution. The arbitrary constants, 4 and 13, are used to place the deltas on the difficulty scale of the Educational Testing Service, with a mean of 13 and a standard deviation of 4. The two sets of

delta values can be plotted against each other and the swarm of points so formed should cluster tightly around a 45° line that passes through the origin. Biased items are indicated by points that are beyond some prespecified distance from the line.

A major shortcoming of the delta-plot method is that excellent items (in the sense of validly discriminating between high and low scorers on the test) will erroneously appear to be biased when the reference and focal groups differ on the trait being measured. A second problem with the Angoff and Ford approach to DIF is that it leaves unanswered the central question in any evaluation of differential item functioning: do individuals of equal proficiency, ability, or knowledge on the construct being measured (regardless of their group membership) have the same probability of answering the item correctly?

Two general methods for investigating the above question (one based on classical test theory principles, the other on item response theory) have evolved. The former uses the total score on the test itself to match members from each group on ability. The summed chi-square methods (Scheuneman 1979, Camilli 1979) first divide the distribution of raw scores into J ability levels. A 2×2 frequency table of right/wrong responses is formed for each item with the groups constituting the rows of the table, say, and the item response (1, 0) constituting the columns. These are then combined in a $2 \times 2 \times J$ contingency table across the J ability levels. The test statistic used to determine the existence of item bias is the familiar chi-square test of association. Scheunamen's χ^2_{correct} statistic takes into account only the correct responses. The more appropriate Camilli full information χ^2_{full} statistic considers both correct and incorrect responses.

By far the most popular method for investigating DIF using classical test theory principles is the measure of association developed by Mantel and Haenszel (1959) and adapted for DIF research by Holland and Thayer (1988). Like the chi-square methods, the Mantel–Haenszel (M–H) procedure divides the score scale into J levels. The procedure uses the *odds* of getting an item correct, p/q (where $q = 1-p$), by members of each group. The ratio of these odds, α_j, is combined across levels and weighted by cell frequency. For computational and interpretive convenience, the natural log of the odds-ratio, rather than the odds-ratio itself, is used. The M–H log odds-ratio, β_j, is a signed index, with positive values signifying DIF in favor of the reference group, and negative values indicating DIF in favor of the focal group. A chi-square test for the M–H index may also be calculated.

Differential item functioning has also been investigated extensively using item response theory (IRT), a powerful and elegant test model that attempts to model, in probabilistic terms, responses of examinees to individual items. Each item on a test can be represented by an item characteristic curve (ICC), an S-shaped item response curve, based upon the logistic function, that specifies the probability that persons of a given ability, θ, will answer the item correctly. The ICCs for each item may be estimated separately for each group. In IRT, ability matching is achieved as a by-product of the rather complex estimation procedures themselves. Once examinees are placed on the same θ scale, if no DIF is present, the two ICCs should overlap within measurement error.

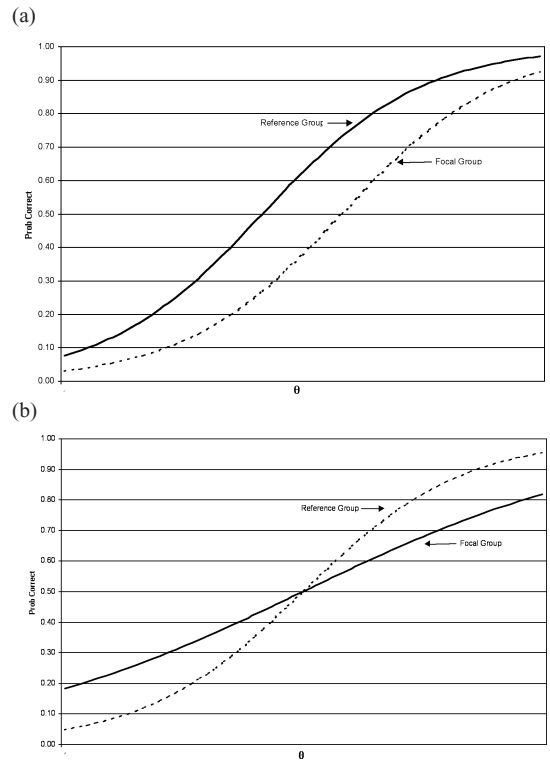

Figure 1
(a) An item exhibiting uniform DIF. (b) An item exhibiting nonuniform DIF

Figure 1 depicts two examples of differential item functioning using item response theory. In both cases, the ICCs are for the same item estimated separately for the two groups. In the first example, the item is biased against the focal group because, for fixed θ, members of that group have a substantially lower probability of getting the item right than members of the reference group. The second example depicts a relatively rare but real occurrence, 'nonuniform' DIF: the item is biased *against* high scoring members of the focal group, but biased *in favor of* low scoring focal group members. The opposite pattern occurs for reference group members. It should be noted that classically-based procedures cannot distinguish between these two forms of DIF.

3. Bias in Selection and Prediction

When a test is used to predict some future criterion such as success in school or performance on the job, the test may not do so equally well for different groups. For example, when the regression line for male candidates or the regression line from combined male and female candidates is used to predict the college performance of female students, the predictions tends to be systematically too low. That is, female students tend to perform better in college than would be expected from their SAT scores. To the extent that an institution's admission policy adopts a 'top-down' strategy, the test would be systematically biased against women.

3.1 The Classical or Regression Model

The definition of selection or prediction bias that has attained some measure of agreement among measurement specialists is that due to Cleary (1968). Known as the 'classical' or 'regression' model of selection bias, the model states that 'a test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test is designed, consistent nonzero errors of prediction are made for members of the subgroup' (p. 115). In this definition, if a given test score is associated with the same level of future criterion performance, regardless of group membership, then the test is 'fair' or unbiased, and to the extent that selection is based solely on predicted performance, the selection system is fair. Statistically, selection bias is indicated when, in the regression of criterion performance on test scores, either the regression lines or intercepts differ for focal and reference groups. When the regression lines coincide, or more precisely, when the regression lines do not differ to a statistically significant degree, the test is said to be unbiased.

Thorndike (1971) was the first to point out an unfortunate difficulty with the regression model. In passing, we note that his reasoning figured prominently in the US National Academy of Science report on fairness in employment testing (Hartigan and Wigdor 1989). The situation is illustrated graphically in Fig. 2. In the figure, we make the usual assumption that there is a 'cutoff' score on the test (defining selection) and a cutoff score on the criterion (defining 'success' and 'failure'). The quadrants of the resulting 2×2 table comprise four distinct groups of individuals. Those who are: *True Positives*: selected and successful (quadrant I); *False Negatives*: rejected but would have succeeded had they been accepted (quadrant II); *True Negatives*: rejected and would have 'failed' had they been accepted (quadrant III); *False Positives*: selected and unsuccessful (quadrant IV).

The regression lines for the focal and reference groups coincide, so that the test is fair according to the regression model. The difficulty with the model arises from two sources. First, the two groups differ in mean scores on the test (a typical finding). Second, the test is never perfectly reliable and is always imperfectly correlated with the criterion. These two conditions together result in the percentage of focal group members who were improperly rejected (quadrant II) being substantially larger than the comparable percentage of reference group members. Hartigan and Wigdor (1989) state the case succinctly, 'The undoubted effect of imperfect prediction when social groups have different average test scores is to place the greater burden of prediction error on the shoulders of the lower-scoring group' (p. 258).

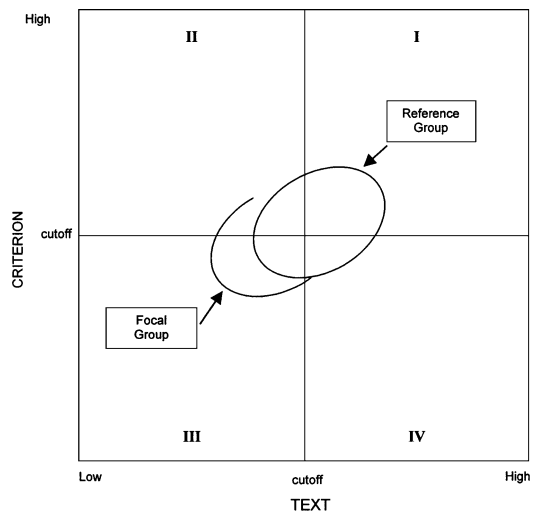

Figure 2
Effects of imperfect prediction on the false negative rate when there is no prediction bias according to the regression model

Before turning to alternative selection models, it should be noted that the apparent unfairness of the regression model, when its conditions for fairness have been satisfied, result not necessarily from any inherent deficiency in the test itself. Nor are these effects inextricably bound up with race, ethnicity, social class, sex, or some other defining feature of the groups. Rather, the 'unfairness' results solely as a statistical result of the two groups having different means on the test. The effects would be as clearly in evidence within the focal or reference group were its members divided into high and low scoring subgroups.

3.2 Alternative Models of Fair Selection

Alternatives to the classical model have been proposed by Cole (1973), Darlington (1971), Einhorn and Bass (1971), Gross and Su (1975), Jensen (1980), Linn (1973), and Thorndike (1971), among others. In a

seminal and influential paper, Petersen and Novick (1976) analyzed and evaluated the models of Cleary, Cole, Einhorn and Bass, Gross and Su, Linn, and Thorndike in terms of the degree to which they were 'internally consistent.' Their analyses of the models will be briefly described here, but a thorough review of all of the models cannot be undertaken in the limited space available. The reader is advised to consult the original sources as well as other in-depth discussions found in Jensen (1980) and the special issue of the *Journal of Educational Measurement* (1976) edited by R. M. Jaeger.

Thorndike drew a distinction between fairness to individuals, which the regression model assures, and fairness to groups. According to Thorndike,

> The qualifying scores on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance (1971, p. 63).

Fairness to the group consists of selecting equitable proportions of members from each group. Different cut scores, x_i, are possible, in fact likely, for the two groups. Assuming a minimum level of acceptable performance on the criterion, y, a selection test is considered fair according to the Thorndike model if the cut-score x^* for the two groups is such that

$$T = \frac{\Pr(X \geqslant x_F^* | \pi_F)}{\Pr(Y \geqslant y^* | \pi_F)} = \frac{\Pr(X \geqslant x_R^* | \pi_R)}{\Pr(Y \geqslant y^* | \pi_R)}$$

where T is a constant for the reference population (π_R) and the focal population (π_F) and x^* and y^* are cut scores for the test and criterion, respectively. Following Peterson and Novick, this model has come to be called the Constant Ratio Model.

Cole (1973) proposed a model of fair selection, the Conditional Probability Model, that assures to any individual, regardless of group membership, an equal chance of being selected, provided that he or she would have been successful if selected. That is, for anyone who can achieve a satisfactory criterion score ($Y \geqslant y^*$), there should be the same probability of selection regardless of group membership. The Cole model specifies that a test is fair when cut scores are set for the two groups such that

$$C = \Pr(X \geqslant x_R^* | Y \geqslant y^*, \pi_R) = \Pr(X \geqslant x_F^* | Y \geqslant y^*, \pi_F)$$

where C is a fixed constant for both populations.

It is noted that in most situations where the focal group mean on the selection test is below that of the reference group, both the Thorndike and Cole models will result in a greater proportion of focal group members being selected than would be the case with the regression model. The downside from the selecting institution's point of view is a less than optimal mean overall performance of those selected.

In the usual selection situation, the 'given' information for each applicant is not his or her future state of being (success or failure) in relation to the criterion variable, but rather his or her present observed standing on the selection test. That is, a fair selection test, it can be argued, should be based upon the conditional probability of success, *given* selection. Such a model has been suggested, though not necessarily advocated, by Linn (1973). The model essentially reverses the Cole model above. For a test to be fair, cut scores for the two groups are set so that the following equality holds, where the notation is as before:

$$L = \Pr(Y \geqslant y^* | X \geqslant x_R^*, \pi_R) = \Pr(Y \geqslant y^* | X \geqslant x_F^*, \pi_F)$$

Associated with each of these ostensibly fair and plausible alternatives to the regression model is a different index which must be constant across groups. Yet, in general, they all result in different cut scores and different proportions of individuals from the focal and the reference groups being selected.

In an influential paper, Petersen and Novick analyzed the competing and contradictory models by applying a test of 'internal coherence.' Briefly, the three models above focus attention on those who would be successful. Fairness is discussed in relation to these individuals only. But the same logic, the same perspective on fairness and bias, can be applied to individuals who were *not* selected and who would have been *unsuccessful* if selected. It turns out that when questions are phrased with these individuals in mind, the above models contain a logical contradiction. For example, the Thorndike model can be restated as follows: the qualifying scores on a test should be set at levels that will *reject* applicants in the two groups in proportion to the fraction of the two groups who *fail* to reach a specified level of criterion performance. The reformulated, converse model then becomes:

$$T' = \frac{\Pr(X \leqslant x_R^* \pi_R)}{\Pr(Y \leqslant y^* | \pi_R)} = \frac{\Pr(X \leqslant x_F | \pi_F)}{\Pr(Y \leqslant y^* | \pi_F)}$$

The models proposed by Cole and Linn may be similarly reformulated. Petersen and Novick prove that only in rare and unlikely circumstances will the three converse models result in the same cut scores as the original models. Different individuals will be selected depending on whether one formulates fairness with selected, successful individuals in mind, or whether one focuses on rejected, unsuccessful individuals. The three models are internally incoherent.

By contrast, the regression model and a model proposed by Einhorn and Bass (1971) are internally coherent. The Einhorn and Bass model sets the cut score for each group in such a way that the criterion score for persons with test scores greater than the cut score will be above the minimum acceptable criterion

with probability at least equal to some specified value. It is then required that this probability (or conversely, risk) be the same for all groups of interest, hence its designation as the Equal Risk Model.

4. Conducting Remarks

Measurement specialists can contribute importantly to discussions about the technical adequacy of tests and the importance of adopting models that meet specific requirements. Of course, matters of test use in education and employment will not be decided by measurement specialists and statisticians. Matters of test use and testing policy will ultimately be decided in the public, legislative, and judicial arenas. It is nevertheless hoped that technical analyses of differential item functioning, of the various sources of internal test bias, and of the consequences of adopting a particular selection model will inform the public, legislative, and judicial debates.

Bibliography

Angoff W H, Ford S F 1977 Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement* **10**: 95–106

Bond L 1981 Bias in mental tests. In: Green B F (ed.) *New Directions for Testing and Measurement: Issues in Testing–Coaching, Disclosure, and Ethnic Bias, No. 11*. Jossey-Bass. San Francisco

Cleary T A 1968 Test bias: Prediction of grades for negro and white students in integrated colleges. *Journal of Educational Measurement* **5**: 115–24

Camilli G 1979 *A Critique of the Chi-square Method for Detecting Item Bias*. Unpublished manuscript. University of Colorado, Laboratory of Educational Research

Cole N S 1973 Bias in selection. *Journal of Educational Measurement* **10**: 237–55

Darlington R B 1971 Another look at 'cultural fairness.'*Journal of Educational Measurement* **8**: 71–82

Einhorn H J, Bass A R 1971 Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin* **75**: 261–9

Fraugher R L 1978 The many definitions of test bias. *American Psychologist* **33**(7): 671

Gross A L, Su W 1975 Defining a 'fair' or 'unbiased' selection model: A question of utilities. *Journal of Applied Psychology*. **60**: 345–51

Hartigan J A, Wigdor A K 1989 *Fairness in Employment Test: Validity, Generalization, Minority Issues, and General Aptitude Test Battery*. National Academy Press, Washington, DC

Holland P, Thayer D T 1988 Differential item functioning and the Mantel–Haenszel procedure. In: Wainer H, Braun H (eds.) *Test Validity*. Lawrence Erlbaum Associates, Hillsdale, NJ

Holland P, Wainer H 1993 *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ

Jensen A R 1980 *Bias in Mental Testing*. Free Press, New York

Linn R L 1973 Fair test use in selection. *Review of Educational Research* **43**: 139–61

Mantel N, Haenszel W 1959 Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**: 719–48

Petersen N C 1976 An expected utility model for 'optimal' selection. *Journal of Educational Statistics* **1**: 333–50

Petersen N C, Novick M R 1976 Towards equalizing educational and employment opportunity selection. *Journal of Educational Measurement* **13**: 77–88

Scheuneman J D 1979 A method for assessing bias in test items. *Journal of Educational Measurement* **16**: 142–3

Thorndike R L 1971 Concepts of culture-fairness. *Journal of Educational Measurement* **8**: 63–70

L. Bond

Test Theory: Applied Probabilistic Measurement Structures

Psychological test theory invites the formulation and testing of methods of assigning numbers to individual's attributes, such as degrees of skills or personality traits. Measurement theory is the formal and logical theory of necessary and sufficient conditions of attributing measures to objects or events (see *Measurement, Representational Theory of*). It is astonishing that test theory is not usually treated as a special area of measurement theory. Test theory (modern test theory, item response theory, IRT) is probabilistic and concerned with the relation of response probabilities to latent traits. It usually neglects the measurement principles (relations and operations, homomorphisms) necessary to represent observations by numbers and their relations (with the exception of the principle of specific objectivity, see below). Usually a representation by an analytical model (see below) is postulated and statistical tests are given to test the adequacy of the representation of empirical data by the model (item characteristic function, parametric IRT). Measurement theory usually is deterministic and axiomatic and concentrates on relations and operations. The relations typically are qualitative or ordinal. How can measurement theory be made probabilistic? (See *Test Theory: Applied Probabilistic Measurement Structures*.) How can test theory be made concerned about logical relations and homomorphisms (structure preserving representations)?

The more recent development of nonparametric IRT (Hemker et al. 1997, Holland and Rosenbaum 1986, Irtel and Schmalhofer 1982, Junker 1993, Junker and Ellis 1997, Mokken 1971, Rosenbaum 1984, Scheiblechner 1995, 1998, 1999, Stout 1987, 1990) is closer to measurement theory. The authors specify empirically (statistically) testable criteria necessary for unidimensional representations of psychological test data.

Typical psychological tests consist of a set of items ($i = 1, 2, \ldots, k$) which are supposed to 'measure' a common (unidimensional) psychological trait (e.g., verbal, numerical, spatial … intelligence; neuroticism,

depression,…), and to which the subject ($v = 1, 2, \ldots, n$) chooses one answer out of several ordered categories ($h = 0, 1, 2, \ldots, m$). In most cases the test performance of a subject is summarized by the Likert score (simple integer score) or the trivial scoring function: all answers are weighted equally with constant weights for categories and the weights are summed over items. A major advance over this naive procedure was made by parametric rating scale models inspired by Rasch's measurement theory. A well-known parametric IRT model for ordered response categories is the rating scale model (RSM).

$$P(x = h \mid v, i) = \frac{\exp(h(\theta_v - \beta_i) + \varpi_h)}{\sum_{x=0}^{m} \exp(x(\theta_v - \beta_i) + \varpi_i)}$$

$$\beta_i = \sum_l a_{il}\eta_l + c$$

(Andersen 1977, Andrich 1978, Masters 1982) where x is the response category of the answer ($x = 0, 1, 2, \ldots, h, \ldots, m$), θ_v the subject (ability) parameter, β_i the item (difficulty) parameter (with possibly a linear model with respect to basic components η_l) and ϖ_h the response category (attractiveness) parameter (with the normalizations for uniqueness of parameterization $\sum_i \beta_i = 0$, $\varpi_0 = \varpi_l = 0$, $c = 0$).

Models like this are an improvement over simple integer scoring because the fit of the model can be tested by a conditional likelihood ratio statistic (and the fit of the linear item parameter model can be tested by an additional likelihood ratio test). Thus we may detect, for instance, that no such psychological trait as conjectured does exist (in the sense of the model) or that not all items belong to the same psychological trait, or that the scoring of the categories may depend on the item and so on. However, the logistic form of the category characteristic curve (CCC, the function describing the probability of choosing category x; Hemker et al. 1996), the existence of three scales (subjects, items, categories) and their additive concatenation, the response category as scoring function (factor h or x in the exponent) are postulated all at once, and if the model does not fit there is no (or little) evidence as to which assumptions are failing (the situation is much better concerning the linear model for the item parameter assuming that the basic model fits).

An essential advantage of the probabilistic measurement theoretic formulation of such models is that no such risky assumptions are necessary simultaneously in advance, that each step can be tested separately, and that those axioms can be identified which are possibly responsible for lack of fit. The test of the logistic form of the CCC and of the adequacy of the scoring function are rendered unnecessary by a measurement theoretic test model because no analytic CCC is assumed and there is no need of a scoring function (beyond an ordinal scale). The existence of ordinal scales for subjects and items can be tested separately from each other and independently of the postulation of an additive conjunction of interval scales for subject and item parameters (see for instance the exponent of the CCC of the RSM). Should the test of additive item and subject parameters be successful, we could then test an additive conjunction with an appropriately constructed scoring function (an interval scale for the response category scale). Independently of the outcome of the latter test, and separately from each other, we can test additive models for the item and subject parameters. In short, the advantages of measurement theoretic test models are the stepwise construction of a hierarchy of gradually more complex models, and an exact specification of the level at which the construction of the models possibly might fail.

1. *Making Deterministic Axiomatic Structures Probabilistic*

A successful translation of the deterministic axioms of a measurement structure into a probabilistic version grants the formal rigor of measurement structures while being at the same time probabilistic, i.e., provides a statistical theory for testing the validity of the model. Some possibilities of probabilization of axiomatic measurement structures are expounded by Heyer and Nideree (see *Measurement Theory*: *Probabilistic*). A version of a probabilistic additive conjoint structure will be given below (ADISOP, CADISOP). The most basic proposition of most measurement structures is that an object a is in some respect larger, more attractive, etc., than an object b, symbolically, $a \gtrsim b$, where the objects and the relation are assumed to be empirically observable and deterministic (for example, two weights and the turning of the scale of an idealized balance). How could this be translated into the single, inconsistent qualitative reactions (right–wrong, strongly refuse–strongly prefer) of subjects to single items? Falmagne (1976, 1979) suggested replacing $a \gtrsim b$ with the probability $P(a \gtrsim b)$ of observing a 'larger reaction' (for example, the judgement a is louder than b) upon repeated presentation of the pair (a, b) (or of observing a larger median reaction time, etc.). This is feasible for some psychophysical tasks, but totally impractical for usual psychological intelligence and personality tests where individually identifiable items reasonably can be presented only once to subjects (a repetition would change the meaning of the item). For these I suggest replacing the larger relation $a \gtrsim b$ or the probability of a larger relation $P(a \gtrsim b)$ with the relation of two probabilities, $P(X \leqslant x \mid a) \leqslant P(X \leqslant x \mid b)$, where $P(X \leqslant x \mid a)$ is the probability of observing a specified or 'smaller' reaction than x (for example, a predominantly right answer) on presentation of a. If there are more than two graded responses (rating scales, polytomous

Table 1
Stochastic dominance of variables, $F \gtrsim T \gtrsim P \gtrsim U$ (cumulative distribution functions: 1, positive; 2, moderately positive; 3, moderately negative; 4, negative)

	Rating			
Category	1	2	3	4
U	0.417	0.778	0.972	1.000
P	0.183	0.563	0.817	1.000
T	0.064	0.278	0.714	1.000
F	0.063	0.175	0.588	1.000

ordered response categories) available for items I suggest using the concept of stochastic dominance (see below and *Majorization and Stochastic Orders*). The probabilities are an intermediate layer intervening between the empirical and numerical relational system of a measurement structure (Roberts 1979, p. 274) and even estimates from single observations turn out to be sufficient for statistical testing of larger restrictive measurement structures (for example, if a subject 'moderately agrees' to an item his cumulative probabilities of 'disagreeing' or 'moderately disagreeing' are set to 0 and his probabilities to 'moderately agree' or 'agree' are set to 1).

A stochastic variable W dominates a stochastic variable V, if the probability that W is smaller/equal than an arbitrary value x is smaller/equal to the probability that V is smaller/equal to x for all x, that is:

Definition. Stochastic dominance $\gtrsim$.

$$W \gtrsim V \Leftrightarrow F_w(x) \leqslant F_v(x) \quad \text{for all } x$$

where $F(x) = P(X \leqslant x)$

Definition. Strict stochastic dominance $\succ$.

$$W \succ V \Leftrightarrow F_w(x) \leqslant F_v(x) \text{ for all } x$$

and

$$F_w(x) < F_v(x) \quad \text{for at least one } x.$$

If $W \gtrsim V$ we also write $F_w(x) \gtrsim F_v(x)$ for the cumulative distribution functions (CDFS) and similarly for $\succ$.

As an example, the students of a lecture on statistics rate their attitude (preference/aversion) towards the usefulness of statistics for the understanding of psychology (U) (high/low), the importance of statistics for professional success (P) (important/unimportant), the time needed to learn statistics (T) (acceptable/annoying), and the work with formulae (F) (fascinating/disgusting) on a four point scale. The students reject formulae most, the time requirements for learning statistics next, and think more positively about the importance of statistics for professional success, and most for understanding psychology (Table 1).

If the preference (aversion) for one option stochastically dominates the preference (aversion) for another option, then the probability that the first option is accepted (rejected) is always greater than or equal to the probability that the second option is accepted (rejected) if the ordinal scale is dichotomised (transformed into a binary scale) at an arbitrary position. This invariance makes the stochastic order relation independent of a great variety of different empirical operationalizations (experimental arrangements for the determination of larger relations).

If the deterministic order relations are made probabilistic by the concept of stochastic dominance, then the concept of empirical verification and falsification of single relations by single observations must be replaced by statistical testing of order relations (Bartholomew 1983, Robertson et al. 1988) and by the likelihood principle of model testing. Suppose we want to test the validity of the above order relation $F \gtrsim T \gtrsim P \gtrsim U$ for all subjects of a specified population. Then there are many nonparametric tests to test the null hypothesis of no differences and we can test the hypothesis of the same (similar) order of items for all subjects by testing samples from all relevant subpopulations of subjects; we can also calculate the likelihood of all possible orders and choose the maximum likelihood order.

2. *The Construction of Ordinal Scales for Subjects and Items: the Isotonic Probabilistic (ISOP) Measurement Structure*

The most basic level of the hierarchy of models for the construction of unidimensional scales for psychological traits is the construction of ordinal scales for subjects and/or items. The construction of two scales simultaneously for two sets of objects (e.g., subjects and items) has been referred to elsewhere as biscalability or biorder (Guttman scale, see *Ordered Relational Structures*). Because these concepts are not probabilistic or appropriate only in psychophysical contexts they are not reported here (but they presumably could be made probabilistic).

In the example of rating different aspects (items i) of a negative attitude towards statistics each subject v generates a response vector $\boldsymbol{x}_v = (x_{v1}, x_{v2}, x_{v3}, \ldots)$ of ratings of aversions for items 1, 2, 3,…. The subjects possibly are characterized by a unidimensional latent trait 'aversion against statistics.' The response distributions of the response variables X_{vi} of subjects $v = 1, 2, \ldots, n$ and items $i = 1, 2, \ldots, k$ can be ordered by stochastic dominance;—but how can the subjects (items) be ordered? The response vectors of the subjects give no simple order of the subjects with respect to aversion against statistics. The order relation on the set of subjects (items) has to be defined by

a new concept. We can say that one subject dominates another (with respect to aversion against statistics) if all his/her ratings are larger than or equal to the ratings of the other subject. Of course, many pairs of subjects will not be ordered by this relation (the order is partial). The subjects cannot be compared directly, but by means of their response vectors, i.e., by means of the items as instruments; and therefore the measurement structure is called an instrumental probabilistic pair comparison system (Irtel and Schmalhofer 1982; bioorder, see *Ordered Relational Structures*).

The set of observed multidimensional polytomous response vectors $\boldsymbol{x} = (x_1, x_2, x_3, \ldots)$ ordered by the relation

$$\boldsymbol{x} \text{ dominates } \boldsymbol{y} \equiv x \gtrsim y \text{ iff } x_i \geqslant y_i \text{ for all } i$$

is a multidimensional coordinatewise partial order.

If several variables are stochastically ordered and their order coincides with a scalar parameter of their distributions then they are said to be *isotonic* (similarly ordered) in that parameter. That parameter may represent a known attribute of the variables or a scale value on a latest dimension (for example, aversion in Table 1).

Definition. Isotonicity. A family of variables X_i or distribution functions F_i depending on a parameter p_i is isotonic in p_i if $X_i \prec X_j \Leftrightarrow F_i(x) \geqslant F_j(x)$ for all x of $p_i \leqslant p_j$.

In order to be able to construct ordinal scales for the set of subjects and items we now extend the concept of isotonicity to a two-dimensional array of variables X_{vi}, for example the aversions of several subjects for a set of attitude items or the preferences of visitors for various holiday arrangements. This is achieved by the axioms of the ISOP measurement structure.

Let A be a set of subjects $v, w, \ldots$, and Q a set of items $i, j, \ldots$,. We associate with the Cartesian product $A \times Q$ a family of random variables X_{vi} with distribution functions $F_{vi}(x) = P[X \leqslant x \,|\, v, i]$ defined on the set $X = \{x, x', x'', \ldots\}$ of categories (values) of an ordered dependent ('achievement') variable X.

Definition. A system $\langle A \times Q, F \rangle$ where F is a family of distribution functions on $A \times Q$ is weakly, independently ordered (an ISOP model) iff the following axioms are valid.

Axiom W1. *Weak subject independence* holds if the response distribution of subject v dominates the response distribution of subject w at item i, then the response distributions of v are not dominated by w at any other item j.

Axiom W2. *Weak item independence* holds if the response distribution of item i dominates the response distribution of item j for some subject v, then the response distributions of i are not dominated by j for any other subject w.

Axiom LI. *Local independence* holds if the probability of the response vector of a subject v is the product of the probabilities of the single responses

$$F(x_{v1}, x_{v2}, x_{v3}, \ldots \,|\, v) = F(x_{v1} \,|\, v)\, F(x_{v2} \,|\, v)\, F(x_{v3} \,|\, v) \ldots$$

(Scheiblechner 1995, 1998).

We can write W1 in the form 'if $F(X_{vi}) \succ F(X_{wi})$ for some i, then $F(X_{vj}) \gtrsim F(X_{wj})$ for all j meaning that the stochastic dominance relations between subjects are similar (isotonic) for all items. Analogously for W2, the stochastic order of items is similar for all subjects and therefore independent of subjects (except for ties: Scheiblechner 1995, 1998). The pair $\langle A \times Q, F \rangle$ is called an instrumental probabilistic pair comparison system. Axiom W1 orders subjects by the stochastic order of their distribution functions of reactions to items and states that the orders are independent from the specific items used, except for possible ties (subjects are similarly ordered by items). This independence of the orders of subjects of the items is the basis for the construction of the latent unidimensional subject ordinal scale or person parameter. Axiom W2 similarly states that the stochastic order of items is independent of subjects and this is the base of the item ordinal scale or item parameter. The axioms are independent and can be statistically tested independently of each other.

If the system of variables is an ISOP, then the distribution functions can be ordered, and there exist ordinal scales $\varphi_A(v)$ on the set of subjects, and $\varphi_Q(i)$ on the set of items, such that the distribution functions of the family X_{vi} of random variables are functions of the two ordinal scales (Scheiblechner 1995):

$$F_{vi}(x) = F(x \,|\, \varphi_A(v), \varphi_Q(i))$$

where $\varphi_A(v)$ and $\varphi_Q(i)$ are ordinal scales.

The two ordinal scales are 'order independent relative to subsystems' (Scheiblechner 1995), that is, the ordinal scale for subjects does not depend on the set of items and the ordinal scale for items does not depend on the set of subjects, both except for ties (the corresponding property is called specific objectivity or sample independence in the logistic Rasch model; Andersen 1995, Rasch 1960)

The following implications of ISOP can be tested:

(a) the validity of axiom W1 for each item and for the total test (= set of items);

(b) the validity of axiom W2 for each response vector (subject) and for the total sample;

(c) the validity of the coordinatewise partial order of response vectors for each subset of items of the test and for the total test;

(d) the validity of the coordinatewise partial order implied by one part of the test for another part of the test; and

(e) the validity of a large number of manifest partial and linear isotonicities.

Especially the facility to test separately the single axioms of the measurement structure and the separate

tests for different parts and aspects (formal relations) of the data are a measurement theoretical advantage of the ISOP measurement structure.

3. The Construction of Interval Scales for Subjects and Items: the Additive Conjoint ISOP (ADISOP) and the Complete ADISOP (CADISOP)

The axioms of the ISOP model are sufficient for the construction of two separate unidimensional ordinal scales on the sets of subjects and items. We now strengthen the ISOP model by additional axioms in order to derive ordered metric/interval scales for subjects and items. First the ADISOP model is considered which is an ISOP model with a cancellation axiom added that relates the subject and item factors.

Definition. A probabilistic pair comparison system $\langle A \times Q, F_{vi}(x)\rangle$, where $F_{vi}(x)$ is a family of cumulative distribution functions defined on $A \times Q$, is an ADISOP iff the axioms of the ISOP model and the following axiom are satisfied.

Axiom Co. Validity of cancellation of subjects and items up to order o.

(Similar structures are the PACOME models, Falmagne 1979). An explanation of cancellation up to order o is given in Scheiblechner (1999) and Scott (1964). It means that a reduction of the item parameter by an amount δ can be compensated by an increase in the subject parameter by the same amount δ. The subject and item parameters 'cancel.' Or if we attribute real values to subjects and items then the order of their sum is isotonic in the order of their distribution functions. For example, if the 'ability' parameter of subject v plus the 'easiness' parameter of item i is larger than or equal to the sum of the ability and easiness parameter of subject w and item j, then the distribution functions are isotonic in the sum, i.e., $F_{vi}(x) \gtrsim F_{wj}(x)$ and the reverse.

If the system $\langle \boldsymbol{A} \times \boldsymbol{Q}, \boldsymbol{P}\rangle$ is an ADISOP model then the order of the CDFs is identical to the order of the sums of the subject and item parameters, i.e., there exist real functions ϕ_A, ϕ_Q on A and Q and $\psi_{A\times Q}(v, i) = \phi_A(v) + \phi_Q(i)$ on $A \times Q$ such that for all (v, i), $(w, j) \in A \times Q$

$$\phi_A(v) + \phi_Q(i) < \phi_A(w) + \phi_Q(j) \Leftrightarrow F_{vi}(x) > F_{wj}(x)$$

for some x or $\Leftrightarrow F_{vi}(x) \prec F_{wj}(x)$

The scales ϕ_A, ϕ_Q are unique up to positive linear transformations, i.e., they are interval scales if Axiom Co is valid up to arbitrary order o, and they are ordered metric scales if Co is valid up to the finite order $(\min\{n, k\} - 1)$ where n and k are the number of elements of A and Q, respectively. Ordered metric scales are unique up to monotonic transformations which leave invariant the order of the (finite number of) intervals. Ordered metric scales are approximations to interval scales. The cumulative response distribution functions on the product set of subjects and items

$$F_{vi}(x) = F(\varphi(x);\ \varphi_A(v) + \phi_Q(i))$$

are antitonic (do not increase) in $\psi_{A\times Q}(v, i) = \phi_A(v) + \phi_Q(i)$ and isotonic (do not decrease) in $\varphi(x)$, where $\varphi(x)$ is an arbitrary strictly isotonic function of x (i.e., is an ordinal scale). The functions $\phi_A(v)$ and $\phi_Q(i)$ are interval scales (ordered metric scales) with an arbitrary origin (additive constant) and a common unit (scale factor).

The ADISOP model can further be strengthened to a model where the response category parameter also is additive to the item and subject parameter. The axiom W3 strengthens Co and extends cancellation from subjects and items to include subjects and instrumental variable (response category) and items and response variable.

Definition. A complete ADISOP (CADISOP) model is an ADISOP with axiom (W3) (and RS) replacing axiom Co:

Axiom W3. Weak instrumental variable independence holds if:

$F_{vi}(x) < F_{vi}(x')$ for some v and some i, then $F_{wj}(x) \leqslant F_{wj}(x')$ for all w and all j

Axiom RS. Restricted solvability. Certain restrictions apply to the solution of stochastic dominances.

The weak instrumental variable independence axiom W3 is an independence axiom like W1 and W2 and makes the factor of the coding of the response variable X independent of the experimental factor of subjects and of the factor of items (for each value x of the instrumental variable the biorder of subjects and items is the same). It makes the coding of the instrumental variable another additive factor. Cancellation is no longer restricted to subjects and items but includes the response variable factor (an increase of the response category parameter by an amount c can be compensated by an equal increase in the subject or item parameter). Axiom RS grants that the system of equations and inequalities representing stochastic dominances has a solution under certain restrictions (e.g., if the CDF of subject v at item i and category x is between the CDFs of subjects u and w at item j and category x, then there is a subject s whose CDF at j and x' is equal to the CDF at (v, i, x); similarly for the item and category factors (Roberts 1979 p. 218).

If the pair comparison system $\langle A \times Q, P\rangle$ is CADISOP then there exist scales ϕ_A on A, ϕ_Q on Q and ϕ_X on X such that for all (v, i, x), $(w, j, x') \in A \times Q \times X$

$$\phi_X(x) - (\phi_A(v) + \phi_Q(i)) \leqslant \phi_X(x') - (\phi_A(w) + \phi(j)) \Leftrightarrow F_{vi}(x) \leqslant F_{wj}(x').$$

The scales are interval scales with a common unit (scale factor). The 'system characteristic function'

$$F_{vi}(x) = F(\phi_X(x) - (\phi_A(v) + \phi_Q(i)))$$

is strictly isotonic in its argument (Scheiblechner 1999).

4. The Integration of Linear Models for Item and Subject Parameters in Additive Models: Linear ISOP, ADISOP, and CADISOP Models

The ADISOP can be combined easily with an additive model for the item parameter:

$$\varphi_Q(i) = \sum_l a_{il}\eta_l$$

where a_{il} are known coefficients specifying the frequency or amount of 'basic components' l present in item i, and η_l are the (easiness, attractiveness) parameters of the basic components.

The matrix A with the coefficients a_{il} is termed the (item) model matrix. The estimation of the linear item model can easily be integrated into the algorithm for fitting an ADISOP model (Scheiblechner 1999).

It is interesting to note that the coordinatewise partial order of items implied by the model matrix A can be tested in ISOP model with only ordinal scales for items (and subjects) available. The full linear model for items with interval estimates for the basic parameters, however, requires interval scales for items (and subjects) and thus an ADISOP. Thus in a model with latent scales and latent concatenations the question of additivity of item components cannot be separated completely form the question of additivity of item and subject parameters.

It is possible also to include an additive model for the subject parameter

$$\varphi_A = \mathsf{B} \cdot \delta$$

where φ_A is the column vector of subject parameters, B the (n, k) matrix of 'dosages' of (basic) treatments which were given to subjects and δ the column vector of 'gains' (increments of the subject parameter) caused by the treatments (the first column of B may be a column of 1s for an additive constant). The necessary supplements of the estimation algorithm are similar to the inclusion of the item model. Both extensions can be performed in a CADISOP model, of course.

The linear ADISOP/CADISOP models with the inclusion of linear models for items and/or subjects can be tested versus the respective unrestricted models by likelihood ratio tests.

5. Discussion and Future Directions

Probabilistic measurement models are known in psychophysics and utility theory, for choice behavior and pair comparisons, for decision making under uncertainty, and measurement of subjective probability. Not much attention is usually placed on estimation of parameters (probabilities and scale values) and statistical testing of fit (a rare but well-known and unsatisfactory (biased) example is Mosteller's test of fit for pair-comparison data). The probabilistic version of the additive conjoint measurement structure by Falmagne (1979) is a noteworthy psychophysical example for the statistical treatment of measurement models. Some statistical principles of estimation and testing could be borrowed and adapted from isotonic regression (Dykstra 1983, Robertson et al. 1988). The tests concern stochastic (probabilistic) transitivity or consistency. The ISOP model is a weak stochastic transitivity model for a product structure ($A \times Q$) and satisfies the statistical standards of probabilistic test theory. There are many statistical tests for various assumptions (axioms) in various parts of the data (see above, end of Introduction). The tests are based on generalizations of Goodman and Kruskal's index for ordinal association (1954, 1959, 1963, 1972). The U-model by Irtel and Schmalhofer (1982) is ISOP for dichotomous items and is its immediate precursor. The authors suggest an optimal contrast test for the bi-isotonic model (based on Robertson et al. 1988).

It is possible to specialize the ISOP model so as to pair comparison data (and to compare it with other probabilistic measurement models; Fishburn 1973). It is possible also to generate probabilistic polynominal conjoint measurement structures (Roberts 1979) by exactly the same steps as with ISOP structures (there is little use for those structures at present because no interesting empirical procedures for gathering corresponding data are known). Finally, it is possible to apply ISOP structures to geometrical measurement structures (multidimensional scaling, Kruskal 1964a, b).

A permanent problem in psychological test theory is the modelling of the speed–accuracy trade-off for items where the speed and accuracy of the response are recorded at the same time. This problem can be attacked along the lines of isotonic probabilistic measurement structures whereby some interesting new problems arise.

See also: Classical (Psychometric) Test Theory; Factor Analysis and Latent Structure: IRT and Rasch Models; Measurement Theory: Probabilistic; Psychometrics

Bibliography

Andersen E B 1977 Sufficient statistics and latent trait models. *Psychometrika* **42**: 69–81

Andersen E B 1995 Polytomous Rasch models and their estimation. In: Fischer G H, Molenaar I W (eds.) *Rasch Models: Foundations, Recent Developments and Applications*. Springer, New York, pp. 271–92

Andrich D 1978 A rating formulation for ordered response categories. *Psychometrika* **43**: 561–73
Bartholomew D J 1983 Isotonic inference. In: Kotz S, Johnson N L (eds.) *Encyclopedia of Statistical Sciences*. Wiley, New York, Vol. 4, pp. 260–5
Dykstra R L 1983 An algorithm for restricted least squares regression. *Journal of the American Statistical Association* **78**: 384, 837–42
Falmagne J C 1979 On a class of probabilistic conjoint measurement models: some diagnostic properties. *Journal of Mathematical Psychology* **19**: 73–88
Fishburn P C 1973 Binary choice probabilities: On the varieties of stochastic transitivity. *Journal of Mathematical Psychology* **10**: 327–52
Goodman L A, Kruskal W H 1954 Measures of association for cross classifications, I. *Journal of the American Statistical Association* **49**: 732–64
Goodman L A, Kruskal W H 1959 Measures of association for cross classifications, II. *Journal of the American Statistical Association* **54**: 123–63
Goodman L A, Kruskal W H 1963 Measures of association for cross classifications, III. *Journal of the American Statistical Association* **58**: 310–64
Goodman L A, Kruskal W H 1972 Measures of association for cross classifications, IV. *Journal of the American Statistical Association* **67**: 415–21
Hemker B T, Sijtsma K, Molenaar I W, Junker B W 1996 Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika* **61**: 679–93
Hemker B T, Sijtsma K, Molenaar I W, Junker B W 1997 Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika* **62**: 331–47
Holland P W, Rosenbaum P R 1986 Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics* **14**: 1523–43
Irtel H, Schmalhofer F 1982 Psychodiagnostik auf Ordinalskalenniveau: Meßtheoretische Grundlagen, Modelltest und Parameterschätzung. (Ordinal psychological diagnostics: measurement theoretic foundation, tests of the model and estimation.) *Archiv für Psychologie* **134**: 197–218
Irtel H 1995 An extension of the concept of specific objectivity. *Psychometrika* **60**(1): 115–18
Junker B W 1993 Conditional association, essential independence and monotone unidimensional item response models. *Annals of Statistics* **21**: 1359–78
Junker B W, Ellis J L 1997 A characterization of monotone unidimensional latent variable models. *Annals of Statistics* **25**: 1327–43
Kruskal J B 1964a Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**: 115–29
Kruskal J B 1964b Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29**: 115–29
Masters G N 1982 A Rasch model for partial credit scoring. *Psychometrika* **47**: 149–74
Mokken R J 1971 *A Theory and Procedure of Scale Analysis*. Mouton, Paris
Rasch G 1960 *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen
Rasch G 1961 On general laws and the meaning of measurement in psychology. *Proceedings of the IVth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, Vol. 4, pp. 321–33
Roberts F S 1979 Measurement Theory. *Encyclopedia of Mathematics and its Applications*, Rota, G-C (ed.), Addison-Wesley, London, Vol. 7
Robertson T, Wright F T, Dykstra R L 1988 *Order Restricted Statistical Inference*. Wiley, New York
Rosenbaum P R 1984 Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika* **49**(3): 425–35
Scheiblechner H 1995 Isotonic ordinal probabilistic models (ISOP). *Psychometrika* **60**(2): 281–304
Scheiblechner H 1998 Corrections of theorems in Scheiblechner's treatment of ISOP models and comments on Junker's remarks. *Psychometrika* **63**: 87–91
Scheiblechner H 1999 Additive conjoint isotonic probabilistic models (ADISOP). *Psychometrika* **64**(3): 295–316
Scott D 1964 Measurement models and linear inequalities. *Journal of Mathematical Psychology* **1**: 233–47
Stout W F 1987 A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika* **52**: 589–617
Stout W F 1990 A new item response theory modeling approach with application to unidimensionality assessment and ability estimation. *Psychometrika* **55**: 293–325

H. Scheiblechner

Text Comprehension: Models in Psychology

Models of text comprehension are concerned with how people understand and remember the verbal information they see or hear. The need for models of text comprehension arises from over 40 years of research on how people comprehend and learn from text. Early work focused on what people remembered from different kinds of texts (e.g., stories, lectures, newspaper articles, and encyclopedia-type articles). This early work showed that some kinds of texts (e.g., stories) were easier to understand and remember than others (e.g., encyclopedia-like articles). Researchers investigated ways to make texts easier to understand and remember, with variable success. These studies brought to light the role of task demands and reader characteristics. For example, memory tasks placed different demands on the comprehender than did problem-solving tasks. Furthermore, comprehension differed depending on reader characteristics such as prior knowledge, reading and listening strategies, attentional resources, purpose, and interest. Current models of text comprehension attempt to account for how text is processed and the relationship between how it is processed, the memory representation that is created, and performance on memory, problem solving, learning, and decision-making tasks. Models of text comprehension are tested by how well they predict behavioral data and through computational modeling. (For further discussion, see Britton and Graesser 1996, Goldman and Rakestraw 2000, Graesser et al. 1997a, 1997b, Kintsch 1998, Weaver et al. 1995.)

Text comprehension involves many cognitive processes that relate the elements in the text to one another and to prior knowledge so that the result is a coherent mental representation. An important issue for models of text comprehension is the contribution of prior knowledge of the text topic and structure to the formation of coherent representations. Texts are seldom completely explicit with respect to the connections among text elements so comprehenders often have to make inferences to bridge these 'gaps.' Inferences sometimes are made just from the text information and sometimes involve elaborative processes that rely on prior knowledge.

The outcome of text comprehension is a memory representation that reflects the meaning the comprehender has constructed. The representation may reflect the surface text itself, the meaning constructed from the text elements (e.g., words, phrases, and sentences that were actually in the text), and/or meaning elements derived from prior knowledge. The representation often goes beyond the meaning of the text (the textbase model) to reflect a model of the situation described by the text. Prior knowledge plays a greater role in the situation model than in the textbase.

In general, models of text comprehension seek to describe, explain, and predict various processing activities and their representational consequences. A lower bound on the adequacy of models of text comprehension is that they provide explanatory accounts of the behavioral data. Currently, models of text comprehension provide more adequate descriptions and explanations than predictions of behavioral phenomena. Furthermore, most models do not deal adequately with all of the empirical findings. Most models selectively focus on some subset of empirical findings, although they increasingly attempt to be comprehensive (see, e.g., Britton and Graesser 1996, Gernsbacher 1997, Kintsch 1998, O'Brien and Myers 1999, van Oostendorp and Goldman 1999). The remainder of this article discusses current and future directions in the study of text comprehension models.

1. Approaches to Studying Models of Text Comprehension

Text comprehension research has required the development of methods for systematically analyzing and describing characteristics of the text, the comprehender, and the processing and representation demands of various comprehension tasks. Researchers examine these 'input' variables in relation to two types of dependent measures, namely those that assess processing as it is occurring (on-line measures) and those that assess the results of the processing (memory representation measures). Computational models have been formulated to give more precise explanatory and predictive power to models of text comprehension.

1.1 Text Analysis

Texts are often described at the level of the individual ideas (microlevel) as well as the more global organization of these ideas (macrolevel). Origins of modern text analysis systems typically represent the ideas as propositions. Although the specific details of propositional notation differ somewhat across systems, a proposition is a theoretical unit that contains a predicate (e.g., a verb or adjective) and associated arguments (e.g., nouns, embedded propositions). Arguments have a functional role such as agent, object, or location (for discussion, see Graesser et al. 1997a, Goldman 1997).

Consider the first sentence in a hypothetical paragraph: Mike drove to the gas station and got out of the car. There are two ideas linked together by an and. Mike is the agent of both driving and getting out of the car and the location of both actions is the gas station. The ideas in subsequent sentences would be similarly described and attempts made to link them. Suppose the next several sentences relate details of Mike's attempt to rob the gas station, including the fact that the attempt was foiled when the car stalled as Mike was making his getaway. The global organization would reflect an attempted robbery theme, provided that the comprehender had prior knowledge of robbery scenarios. Ideas more central to the robbery theme would have greater importance, be more central, and be at a more superordinate level in the representation.

Empirical research has shown that propositions are a useful way to characterize text and prior knowledge. Recent challenges to propositional representations contend that there is more to meaning than the symbolic form, i.e., the word (Glenberg 1997). Clearly this is the case, but it does not seem to deny the usefulness of propositional representations for text research.

Although prior knowledge can guide the organization of ideas in a text, the surface text itself contains cues to the underlying semantic organization. Relationships among ideas may be signaled by repetition of specific concepts, either in exactly the same form (e.g., car) or through synonym (e.g., jalopy for the stalled car in the Mike example). Rhetorical devices (e.g., The key idea is, The first point is) and graphic cues (e.g., paragraph indentations, boldface font) can help readers recognize and create connections among ideas (Goldman and Rakestraw 2000, Zwaan and Radvansky 1998).

1.2 Comprehender Characteristics

Several characteristics of comprehenders provide constraints on text comprehension. The most frequently

examined are prior knowledge relevant to the topic of the text and discourse forms, working memory, and interest.

1.2.1 Prior knowledge. Researchers have developed various techniques to assess the impact of prior knowledge on text comprehension and vice versa. Techniques for assessing domain-relevant concept knowledge include concept relatedness ratings, question answering, sorting, and cued associative responding. Assessments of discourse genre knowledge, and rhetorical structures include judgment tasks, their use in generated text, and patterns of on-line processing measures such as reading time and verbal reports. By comparing groups that differ in prior knowledge measures, researchers can assess the impact of knowledge on text comprehension. Researchers infer the impact of the text on knowledge from changes in measures of prior knowledge before and after comprehension.

1.2.2 Working memory. Although all humans have limited attentional resources, there are individual differences in working memory span for verbal material and these are related to text processing and memory. Daneman and colleagues (see Daneman and Tardif 1987) developed a technique for assessing working memory span. Working memory limitations mean that, in most cases, the whole text cannot be processed at once but must be processed in cycles. Ericsson and Kintsch (1995) proposed that one way of dealing with resource constraints is to construct retrieval structures in long-term working memory.

1.2.3 Interest. Personal interest in particular topic areas is assessed through self-report measures typically. Text comprehension is positively related to personal interest in the topic of the text. Situational, or text-based interest, is associated with specific information in a specific text and often is measured through ratings of each sentence in a text. Information with high situational interest, sometimes called 'seductive details,' is well recalled regardless of its importance to the overall theme of the text (see review by Alexander et al. 1994).

1.3 Analysis of Task Demands

Task demands are characterized in terms of the memory representation needed to perform on post-comprehension tasks and whether the tasks require that the comprehender just retrieve information or retrieve and reason with information. For example, comprehenders can do an excellent job of retelling a story so long as they have a good representation of the surface elements and textbase of the story. However, if they need to decide whether the main character acted ethically, they need to reason about the events in the story. The reasoning involves the situation representation.

Other analyses of task demands examine indices of processing difficulty (e.g., unfamiliar vocabulary or complex syntax) in order to compare different texts or task situations. Under conditions of high task demands, comprehenders should either take longer to process the text or end up with less coherent representations. The effects of task demands will be minor if comprehenders do not interpret them appropriately and strategically respond to them (Goldman 1997).

1.4 Assessing Online Processing of Text

Online techniques allow text processing to be assessed independently of the resulting representation and its subsequent use in memory or problem solving tasks. Online measures indicate what is active in working memory at the same time. This is important because limits on attentional resources constrain the amount of text and prior knowledge that can be active simultaneously. Simultaneous activation is necessary for forming links among text or prior knowledge elements, according to virtually all current models of comprehension (Britton and Graesser 1996, Kintsch 1998, van Oostendorp and Goldman 1999).

Some online techniques (e.g., eye movements, reading times, and inspection patterns) provide estimates of where readers look, for how long, and the order in which they look at specific information. Other techniques provide estimates of what information is active in working memory at specific points during the reading process (e.g., probe memory, priming, and think-aloud techniques). Several models and mechanisms have been proposed to account for observed patterns of processing and activation of prior knowledge and previously processed portions of text, including the structure building framework (Gernsbacher 1997), associative priming (Kintsch 1998), resonance (O'Brien and Myers 1999), constructionist theory (Graesser et al. 1994), and the memory-based processing model (O'Brien and Myers 1999). In addition, think-aloud protocols indicate that there are age- and skill-related differences in strategies that comprehenders use to form coherent representations of text, with more effective comprehenders using explanation-based strategies to a greater degree than less effective comprehenders (Chi in press, Coté et al. 1998, Trabasso and Magliano 1996).

1.5 Assessing Mental Representations

Obviously, mental representations are not directly observable. Researchers make inferences about mental

representations on the basis of observable indices of those representations, such as the information comprehenders produce when asked for the 'most important' information in a text or when asked to recall a text. Assessments of text representations are taken *after* the comprehender has read or heard the text, and thus differ from on-line measures.

Methods for studying comprehenders' representations are typical of those used to study broader issues of memory. A primary distinction is between recognition and recall or free response, techniques. Recognition memory is often used to determine the content and form of elements in the representation (e.g., surface, text, or situation). Recall or free response measures can be more or less open-ended. Comprehenders may be asked to recall the text, recall just the main ideas from the text, summarize the text, create a diagram of the important information from the text, or answer questions about presented and potentially inferred information.

Assessments of mental representations have established a number of important findings, summarized here. Ideas that are more important, more causally central, or more superordinate are remembered better and more frequently included in summaries than are subordinate ideas. Generally, this is the same information that receives more extensive processing. In addition, the more coherent the text, the better the recall, especially for comprehenders who are 'low knowledge' with respect to the topic of the text. Finally, interest affects what is remembered. When interest value is pitted against importance, interest wins.

2. *Computational Models*

Computational models simulate text comprehension on a computer. Computer models must be extremely specific and precise with respect to the hypothesized processes, representations, and interactions. The fit of the simulated output can be compared with the output of humans for a comparable task to provide a basis for validating and updating models of text comprehension. Likewise, mathematical models quantify precise patterns of processing times, memory scores, ratings, and other psychological data. Failures of fit can be as informative as good fits because these discrepancies can lead to new insights about the limitations of existing models (Goldman et al. 1996, Graesser et al. 1997a).

Computational models of text processing have been influenced by symbolic and connectionist cognitive theories. In symbolic theories (Anderson 1983), actions are linked to specific conditions under which they might occur. These are called production rules and take the form of If ⟨condition⟩ Then ⟨action⟩. When the condition side of a production is satisfied, the production 'fires' and the action occurs. When new input enters working memory, all production rules are evaluated to determine which ones have conditions that are satisfied, and one is selected to fire, completing a cycle. On successive cycles, the input to working memory dynamically changes, and different productions are satisfied.

Connectionist theories (Rumelhart and McClelland 1986) distribute representations and processes among a large set of simple units analogous to neural units. Each word, concept, or proposition has a corresponding ensemble of neural units. The units are connected with one another through weights that are excitatory, inhibitory, or neutral and form a neural network. During comprehension, a particular set of units that matches the input is initially activated and activation spreads throughout the network consistent with the excitatory and inhibitory links. Activation continues spreading until there are relatively minimal changes in the activation levels of the units and a stable pattern of activation is achieved. This pattern of activation captures the meaning of the input.

Most models of text comprehension are hybrids of symbolic and connectionist theories. Currently, the two most influential models are those of Just and Carpenter (1992) and Kintsch (1998).

2.1 *Just and Carpenter's Collaborative Activation-based Production System (CAPS) Model*

The CAPS architecture (Just and Carpenter 1992) is a production system architecture augmented by two characteristics that give it a connectionist flavor. First, the collaborative aspect of CAPS results from removing the restriction that only one production fires on each cycle. Rather, all matched productions are simultaneously fired. Second, the activation process is not all or none. The condition sides of productions accrue levels of activation. The collaborative and changing activation characteristics of CAPS combine to generate a process in the connectionist spirit: productions fire in parallel, gradually changing the activations of working memory elements by an amount proportional to the activations of other elements. When the activation of a particular element falls below a threshold activation level, that element is no longer noticed by the system and is unable to participate in further processing. Just and Carpenter constructed a comprehension-simulation model within the CAPS architecture, READER. It was designed to process a short, expository passage on flywheels. Lexical, syntactic, and semantic representations were built simultaneously due to the parallel firing of all productions that matched on a given cycle.

Just and Carpenter (1992) embellished the CAPS model with the assumption of an activation limit ('cap') on working memory. The capacity-constrained architecture (3CAPS) means that productions com-

pete for available activation. When fired production rules request more activation than is available, the 'cap' has been reached. Processing at the 'cap' results in a graceful loss from working memory of productions that are not participating in processing, and an overall system slowdown.

2.2 Kintsch's Construction–Integration model

The Construction–Integration (CI) model (Kintsch 1998) represents the input as propositional nodes in a network of linked nodes. The node is a symbolic representation of the content that has a weighted link in the connectionist manner to other nodes in the network. The model simulates the construction of a text-base representation and the incorporation of prior knowledge to form a model of the situation described by the text. The model proceeds in two phases. During Construction, propositions are linked based on argument (concept) overlap to create an initial representation. Additional links can be specified depending on assumptions about causality, temporality, and logical relations. Prior knowledge is also activated 'indiscriminately' and some irrelevant information may be included as part of the network representation. Kintsch (1998) uses the Latent Semantic Analysis (LSA) model of knowledge representation (Landauer and Dumais 1997) to simulate prior knowledge activation. In LSA, knowledge is represented as vectors in a high-dimensional 'semantic space.' Concepts in the input text are located in the semantic space and neighbors are activated.

During Integration, activation spreads through the network of nodes. Activation flows out of nodes with few or weak connections and to nodes that are richly connected, until the network reaches a stable state. During Integration, relevant knowledge associations tend to maintain activation and become part of the representation, while irrelevant associations tend to lose activation and drop out of the representation. At the completion of a CI cycle, each text node has a particular strength value and link strengths for each node to which it is connected. At the conclusion of comprehension of the entire text, each node and link in the network has a cumulative strength value, reflecting its cumulative activity during all cycles of comprehension. These values are used to create long-term memory strength values that can be used to predict the relative likelihood of remembering a particular node or relationship.

The CI model takes into account working memory limitations in that input propositions are processed in cycles and only the most active are held over for processing in subsequent cycles. The number that are 'held over' is a parameter in the model. Propositions not held over can be reinstated in subsequent comprehension cycles. Kintsch (1998) uses the theory of Long-term Working Memory (Ericsson and Kintsch 1995) to elaborate the relationship between working memory and long-term memory. This theory postulates that expert performance for tasks such as reading is accomplished by having highly organized and easily accessible knowledge structures in long-term memory. During processing, working memory utilizes 'pointers' to these structures, which makes them quickly and easily accessible, thus accounting for the fast and accurate application of prior knowledge under conditions of normal comprehension.

2.3 Other Activation-based, Dynamic Models

The CI and 3CAPS models have spawned a 'family' of models that share assumptions about fluctuating activation of elements and dynamic behavior of the representations. One model, Capacity Constrained Construction Integration (3CI), combines CI and 3CAPS by replacing the fixed-buffer working memory of the CI model with the dynamic, fixed-capacity working memory system of 3CAPS (Goldman et al. 1996). 3CI allows all propositions that are currently active after a CI processing cycle to be carried over to the next processing cycle. Loss of activation is a natural result of processing when the 'cap' is reached. 3CI results in an enhanced ability to establish coherence across sentences by having more propositions potentially participate in the same processing cycle, thereby increasing the link strength between them. Simulation results show that for texts that have hierarchical organizations (such as a premise and examples), the 3CI model is able to assess the global discourse structure and relate specific information to that global structure more effectively than the CI model.

Other models in the activation-based family focus on the nature of the links among text elements (Britton and Graesser 1996, van Oostendorp and Goldman 1999). For example, Langston and Trabasso (1999) implemented a model in which causal relatedness drives the creation of links between clauses and is responsible for text coherence. The model accounts for a variety of narrative processing phenomena such as reading time, importance ratings, and the contents of think-aloud protocols.

Van den Broek et al. (1999) focus on different kinds of information and connections among nodes. They proposed a 'Landscape' model in which a text representation is constructed from information that is activated from the text in the current reading cycle, the text in the preceding reading cycle, information earlier in the text, and from prior knowledge. Different types of connections carry different weights, and a limited capacity working memory provides constraints on the amount of information that can be processed in any particular cycle. The Landscape model can account for both the order and the content of free recall.

3. Issues for the Future

Conceptual advances in models of text comprehension require that they continue to focus on knowledge activation and inference making, especially for learning from text in content and technical domains. Optimal learning from text occurs when individuals who vary in their prior knowledge of the text topic are 'matched' to texts that vary in conceptual difficulty. While this observation is not novel, Wolfe et al. (1998) provided evidence that LSA provides a reasonably convenient way to establish a range of texts on the same topic that differ in conceptual complexity.

Models of text comprehension also need to expand their focus in a number of ways. First, they need to include the processing of the surface text, including parsing, knowledge of discourse structure and genre, and the impact these have on the creation of propositional representations. Advances in techniques for processing natural language make it more feasible for computational models to take surface text input rather than a researcher-parsed list of propositions. This expansion of computational models will permit more detailed understanding of interactions between surface structure characteristics of the discourse and interpretations that comprehenders construct (Goldman et al. 1996).

A second expanded focus concerns the impact of new input on existing knowledge structures (O'Brien and Myers 1999). This is important for learning in content domains and is important in the everyday conduct of human activity. Adults are resistant to updating information they have read in newspapers and magazines when contradictory information is published subsequently. Whether updating occurs appears to depend on the combined impact of strength and explicitness of the contradiction (see chapters in van Oostendorp and Goldman 1999). Future research on updating needs to examine domain-relevant prior beliefs and knowledge because these are likely to affect readers' judgments about the plausibility of the original and updating information.

Models of text comprehension also need to expand the nature of the texts they examine. In response to the ubiquitous nature of new multimedia technologies and the information resources they make available, individuals are comprehending an expanding array of texts. Research is beginning to examine models of text comprehension in some of these new venues, such as hypertext and hypermedia environments, particularly with respect to scanning and inspection strategies. Comprehenders are accessing multiple sources of information quickly and easily, sometimes in response to self-assessment of comprehension. We know a little about the processing, representation, and integration of information from different sources (Perfetti et al. 1999). We need to know more.

Finally, models of text comprehension need to move beyond the single reader and deal with the social construction of meaning. Text comprehension, like other aspects of cognition, is situated in particular social, historical, and cultural contexts that impact the interpretation process (Bloome and Egan-Robertson 1993, Goldman 1997). What individuals regard as important and meaningful in texts is governed by their interactions with others, as well as by their interactions with the text. Models of text comprehension need to examine the impact of such contexts on the psychological processing and representation of meaning.

See also: Comprehension, Cognitive Psychology of; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Language and Thought: The Modern Whorfian Hypothesis; Learning from Text; Literary Texts: Comprehension and Memory; Memory for Text; Narrative Comprehension, Psychology of; Sentence Comprehension, Psychology of

Bibliography

Alexander P A, Kulikowich J M, Jetton T L 1994 The role of subject-matter knowledge and interest in the processing of linear and non-linear texts. *Review of Educational Research* **64**: 201–52

Anderson J R 1983 *The Architecture of Cognition*. Harvard University Press, Cambridge, MA

Bloome D, Egan-Robertson A 1993 The social construction of intertextuality and classroom reading and writing lessons. *Reading Research Quarterly* **28**: 305–33

Britton B K, Graesser A C (eds.) 1996 *Models of Text Comprehension*. Erlbaum, Hillsdale, NJ

Chi M T H in press Self explaining: the dual process of generating inferences and repairing mental models. In: Glaser R (ed.) *Advances in Instructional Psychology*. Erlbaum, Mahwah, NJ, pp. 161–238

Coté N, Goldman S R, Saul E U 1998 Students making sense of informational text: relations between processing and representation. *Discourse Processes* **25**: 1–53

Daneman M, Tardif T 1987 Working memory and reading skill re-examined. In: Coltheart M (ed.) *Attention and Performance 12: The Psychology of Reading*. Erlbaum, Hillsdale, NJ, pp. 491–508

Ericsson K A, Kintsch W 1995 Long-term working memory. *Psychological Review* **102**: 211–45

Gernsbacher M A 1997 Two decades of structure building. *Discourse Processes* **23**: 265–304

Glenberg A M 1997 What memory is for. *Behavioral and Brain Sciences* **20**: 1–55

Goldman S R 1997 Learning from text: reflections on the past and suggestions for the future. *Discourse Processes* **23**: 357–98

Goldman S R, Rakestraw J A Jr. 2000 Structural aspects of constructing meaning from text. In: Kamil M L, Mosenthal P B, Pearson P D, Barr R (eds.) *Handbook of Reading Research*. Erlbaum, Mahwah, NJ, Vol. 3, pp. 311–35

Goldman S R, Varma S, Coté N 1996 Extending capacity-constrained construction integration: toward 'smarter' and flexible models of text comprehension. In: Britton B K, Graesser A C (eds.) *Models of Text Comprehension*. Erlbaum, Hillsdale, NJ, pp. 73–113

Graesser A C, Gernsbacher M A, Goldman S R 1997a Cognition. In: van Dijk T A (ed.) *Discourse as Structure and*

Process. Discourse Studies: A Multidisciplinary Introduction. Sage, Thousand Oaks, CA, Vol. 1, pp. 292–319
Graesser A C, Millis K K, Zwaan R A 1997b Discourse comprehension. *Annual Review of Psychology* **48**: 163–89
Graesser A C, Singer M, Trabasso T 1994 Constructing inferences during narrative text comprehension. *Psychological Review* **101**: 371–95
Just M A, Carpenter P A 1992 A capacity theory of comprehension: individual differences in working memory. *Psychological Review* **99**: 122–49
Kintsch W 1998 *Comprehension: A Paradigm for Cognition*. Cambridge University Press, New York
Langston M, Trabasso T 1999 Modeling causal integration and availability of information during comprehension of narrative texts. In: van Oostendorp H, Goldman S R (eds.) *The Construction of Mental Representations During Reading*. Erlbaum, Mahwah, NJ, pp. 29–69
Landauer T K, Dumais S T 1997 A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**: 211–40
O'Brien E J, Myers J L 1999 Text comprehension: a view from the bottom up. In: Goldman S R, Graesser A C, van den Broek P (eds.) *Narrative Comprehension, Causality, and Coherence: Essays in Honor of Tom Trabasso*. Erlbaum, Mahwah, NJ, pp. 35–53
Perfetti C A, Rouet J-F, Britt M A 1999 Toward a theory of documents representation. In: van Oostendorp H, Goldman S R (eds.) *The Construction of Mental Representations During Reading*. Erlbaum, Mahwah, NJ, pp. 99–122
Rumelhart D E, McClelland J L, P&P Research Group 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA
Trabasso T, Magliano J P 1996 Conscious understanding during comprehension. *Discourse Processes* **21**: 255–87
Van den Broek P, Yang M, Tzeng Y, Linderholm T 1999 The landscape model of reading: Inferences and the online construction of memory representation. In: van Oostendorp H, Goldman S R (eds.) *The Construction of Mental Representations During Reading*. Erlbaum, Mahwah, NJ, pp. 71–97
van Oostendorp H, Goldman S R (eds.) 1999 *The Construction of Mental Representations During Reading*. Erlbaum, Mahwah, NJ
Weaver III C A, Mannes S, Fletcher C R (eds.) 1995 *Discourse Comprehension: Essays in Honor of Walter Kintsch*. Erlbaum, Hillsdale, NJ
Wolfe M B W, Schreiner M E, Rehder B, Laham D, Foltz P W, Kintsch W, Landauer T K 1998 Learning from text: matching readers and texts by latent semantic analysis. *Discourse Processes* **25**: 309–36
Zwaan R A, Radvansky G A 1998 Situation models in language comprehension and memory. *Psychological Bulletin* **123**: 162–85

S. R. Goldmann and M. B. W. Wolfe

Textbooks

The textbook remains today what Altbach calls a 'virtual icon of education' (Altbach 1991, p. 242), yet controversy surrounds nearly every aspect of its production, dissemination, and use, and its prospects for survival are uncertain in a high bandwidth, digitized, and wireless world. Textbooks in modern society function simultaneously as surrogate curricula, cultural artifacts, and commercial objects. Their development over the past two millennia has been driven by the evolution of formal schooling and of the printing arts, as wells as by the development, particularly since the start of the twentieth century, of educational psychology. Important differences in textbook form and use, and of the issues surrounding these areas, exist across the schooling levels—elementary, secondary, and tertiary; curricular areas; and countries. The form that textbooks will assume in the future is projected to change with the fuller integration of information and computer technologies in schooling.

1. Definition and Evolution

1.1 Definition

Textbook usually is interpreted in a general sense, referring to texts that are especially prepared for the study of a subject. The prototypic textbook has as its core a sequence of topics that are arranged to lead a student through a subject or course of study, accompanied on the periphery by occasional guiding statements, summaries, or the like, exercises for practice, questions for self-assessment, and assistance for the instructor. All textbooks possess the core element but differ in representation of the periphery. Primary level textbooks tend to possess all of the peripheral attributes; secondary level ones, most, and tertiary level ones, fewer. Nevertheless, all textbooks are viewed as surrogate curricula, defining a prescribed codification of knowledge.

1.2 Historical Development

Until at least the Enlightenment, textbooks for advanced study were either original works, such as the grammars of Donatus and Pricisian, Boethius's *Geometry*, and Ptolemy's *Mechanisms of the Heavens*, or abridged compendiums of these works, such as Martinus Capella's *The Marriage of Mercury and Philology* (Cubberley 1920). For children, specially developed materials originate with the *Orbis Pictus* of Comenius (1658/1887), which was the first illustrated schoolbook published for introductory learning. By the end of the seventeenth century primers for teaching reading were published in Europe with the alphabet, lists of syllables and words, as well as religious material: the Lord's Prayer, the Creed, and so on. From the influence of Rousseau and Pestalozzi came, by the second decade of the nineteenth century,

carefully sequenced materials for learning, along with occasional suggestions for the instructor. With the adoption of tax supported schooling, particularly in Europe and North America, both the market and the need for textbooks expanded dramatically, leading to changes in both size and content. Coincidentally, improvements in printing technology and paper making throughout the nineteenth century were also reflected in textbooks. With the rise in educational psychology, beginning at the end of the nineteenth century, instructional approaches became more differentiated and more scripted, leading by the post-World War II period, to separate teacher manuals to accompany many elementary and secondary level textbooks.

2. *Surrogate Curriculum*

According to Altbach and Kelly (1988, p. 3), '(Textbooks) constitute the base of school knowledge, particularly in third world countries where there is a chronic shortage of qualified teachers.' But even in industrialized countries textbooks are the primary codification of school knowledge and the principal guide for teachers. Teachers at the elementary levels, who are responsible for teaching a broad curriculum in self-contained classrooms, choose out of necessity to use textbooks, along with workbooks and teacher guides. At higher levels of schooling the textbook is as much a prop for students, a *vade mecum* for a course, as it is a guide for the instructor.

2.1 *Grade Level Differences*

Although textbooks are used at all levels of education and in all subjects, their importance varies not only across countries but also across grade levels and curricular areas. At the primary levels their use is limited primarily to literacy, at least until the students, on average, can read well enough to gain meaning from other content area texts. In many school systems language arts and mathematics texts may be the only ones commonly used prior to the fourth or fifth year of schooling. Then, through the first years of graduate school, textbooks tend to be the basis of instruction for most subject areas.

2.2 *Subject Area Differences*

Thomas Kuhn has pointed out in *The Structure of Scientific Revolutions* that at the college level textbook use varies dramatically by field of study. Music, the graphic arts, and literature typically are taught through direct exposure of the student to artists' works, textbooks playing only a secondary role. In history, philosophy, and the social sciences, textbooks have a more significant role, but students usually are also assigned readings from the classics and from contemporary research reports. In contrast, the natural sciences are taught almost exclusively from textbooks (Kuhn 1970). Within the natural sciences, textbooks are rewritten after each paradigm shift (or 'revolution,' in Kuhn's terms), with the revolution obscured and the past reinterpreted to show only its connection to the issues of the new paradigm.

2.3 *Use of Textbooks by Teachers*

How teachers use textbooks is a matter of concern for educational administrators. In developing countries where teachers may not be well trained, close adherence to the sequence and directions of the textbook is desired. In other countries, reliance upon textbooks, particularly at the primary levels, is not viewed positively. Studies of teaching practices, however, have tended to show that high percentages of elementary teachers do use textbooks as their primary guide for instruction (e.g., Shaver et al. 1979, Weiss 1987). But other studies have shown that teachers, in general, do not follow their textbooks slavishly (e.g., Stodolsky 1989).

2.4 *Control of Textbook Content*

In most countries a central ministry of education determines the general content of textbooks but the actually writing of the texts may be performed by persons outside the ministry, as in Russia, or within the ministry, as in Singapore. In the US, Canada, and Germany control over textbook content is vested at a provincial or state level, although the actual texts are produced by the private sector. In many developing countries, where indigenous textbook industries generally are not yet developed, multinational textbook companies exercise some control over content. French produced textbooks, for example, are used in many West African secondary schools (e.g., Togo); some of these texts portray sex roles and other culturally sensitive matters that are alien to West African cultures (Biraimah 1988).

2.4.1 Imported textbooks. At a postsecondary level and especially in technical subjects, textbooks are relatively expensive to produce, due to the cost of illustrations, complex diagrams, formulas, and the like. Many countries, including the smaller industrialized ones such as The Netherlands, Finland, and Luxembourg, must import textbooks for these levels, and the majority of the these texts are written in English. A further factor that affects the content of textbooks worldwide, but particularly at a secondary and tertiary level, is that a small number of highly industrialized countries dominate knowledge production and dissemination. Countries like the US invest large amounts of money in all fields of research and development, including educational research,

and tend, therefore, to set many of the tertiary level standards worldwide for curriculum and assessment. Through academic institutions, research laboratories, libraries, and professional journals, countries like Britain, Germany, and the US exert both direct and indirect influence over textbooks in most of the rest of the world. That this knowledge is shared freely is offset partially by the reduction in cultural autonomy that results.

2.4.2 International concerns over content. Even within a world that recognizes, at least in principle, the right of each country to design and teach its own curriculum, cross-national conflicts over curriculum content have occurred. One example is the concern expressed by a number of countries to specific national treatments of twentieth-century history and particularly of World War II and the Holocaust. In 1982 The People's Republic of China and South Korea protested changes in Japanese history books that removed references to the Japanese massacre in Nanking and watered down the language previously used to describe other Japanese behavior toward conquered territories (Ketcham 1986). More recently, an Austrian publisher of a high school history series was forced, through pressure from other countries, to stop referring to Nazi concentration camps as 'punishment camps.' History texts in the US have been attacked for over a half century for lack of fairness and objectivity in describing wars (Walworth 1938), for being bland and condescending (Sewall 1987), and for failing to attribute blame to individuals (FitzGerald 1979). Similarly, science texts have been criticized for their handling of evolution (Skoog 1984).

3. *Cultural Artifact*

3.1 *The Latent Curriculum*

The table of contents and index of a typical textbook define, to a large degree, the manifest curriculum offered by the text. But on top of this curriculum are a secondary set of messages, reflected by commission and omission, that constitute the latent or hidden curriculum (Cronbach 1955). In part, this is what Williams calls the selective tradition of society (Williams 1980). A mathematics textbook that constructs its word problems around interests of a consumer society—goods to be purchased, interest on loans, fees paid for services—conveys a different message from one that presents the same mathematical challenges built around revolutionary themes—distribution of land to the poor, triumph in sports over capitalistic societies, and so on. In China, for example, mathematics textbooks during the Cultural Revolution conformed to the prevailing ideology; with a shift in politics, mathematics textbooks by 1980 were stripped of all praise for the Gang of Four and their policies (Kwong 1988).

3.2 *Class Structure*

Texts for teaching reading have been studied worldwide for different aspects of their latent curricula, including gender roles (e.g., Lorimer and Long 1979–1980, Zimet 1972, Biraimah 1988), need for achievement (e.g., McClelland 1961), moral/ethical values (e.g., de Charms and Moeller 1962), and sociocultural ideologies (e.g., Ahier 1988, Apple and Christian-Smith 1991, Luke 1988). Drawing on Paulo Freire's claim that pedagogy for literacy, by definition, requires an educational ideology, some of these latter authors argue that in countries such as Britain, the US, and Canada, literacy textbooks are as much concerned with maintaining class structures and political hierarchies as they are with imparting reading and writing. Although there is a surface validity to these claims, as demonstrated particularly by textbooks from the 1950s and 1960s, they are difficult hypotheses to test on modern textbooks in industrialized countries, given the pressures on publishers to produce textbooks that are unbiased toward any sex, race/ethnicity, age, country of origin, or political belief. Texts today could be attacked as much for failing to promote cultural continuity as for over-representing any particular class or belief in society.

4. *Commercial Object*

In almost every country that has formal schooling, textbooks have some relationship to the commercial sector. In Singapore, for example, textbooks are designed and written by a Ministry of Education, but copublished with private sector firms. In the USSR, prior to its breakup, the Ministry of Education or a State Committee for Public Education designed the content of textbooks but authors were selected through public competitions and paid royalties for their efforts, which were published by state controlled printing houses. In Mexico, where textbooks have been provided free to the six grades of public and private elementary school, government support has been provided to develop what World Bank researchers have called 'the most dynamic private sector publishing industry in Latin America' (Neumann and Cunningham 1982).

4.1 *Concentrations of Expertise and Production Capacity*

Textbook publishing and marketing require expertise in subject matter and pedagogy, along with capital, production and distribution capacity, and access to

paper. These resources are found primarily in the major metropolitan centers of North America, Asia, and Europe, and are increasingly being dominated by the larger European conglomerates—Pergamon and Pearson in the UK, Bertelsmann in Germany, and Hachette in France. Even in the US, few major textbook publishing houses have remained independent. The rest of the world lies in the periphery of textbook publishing, even though many other large countries such as India, China, Mexico, and Egypt produce many if not most of their own textbooks and the latter two export regionally to smaller developing countries (Altbach 1991).

4.2 Physical Characteristics

The appearance of textbooks—the quality of paper, typography, illustrations, bindings, and cover stock—are a function of printing technology, different national economies, grade level, and national (or regional/local) educational policies, such as the adoption cycle. In all school systems primary level textbooks usually are done on the cheapest paper, often with paper covers and limited use of color for illustrations. Russia textbooks, which the former Soviet government subsidized, were produced to last 2–3 years and were done with cheap paper, cover stock, and bindings. In contrast, textbooks in the US are produced to last 5–7 years and due to market factors have far more expensive components. But they also cost almost 10 times as much as their Russian counterparts (Mehlinger 1991).

4.3 Authorship

Authors of secondary and tertiary textbooks generally write the works that bear their names. At the elementary level, on the other hand, writing houses or other anonymous writers often do the majority of the development. The authors usually are involved in planning and review, and may even draft a few model lessons, but their main function is to impart validity to the resulting product, through their professional reputations, affiliations, or their race/ethnicity, age, or sex. In the US, most elementary reading textbooks in the eighteenth and nineteenth centuries were authored by single individuals, with a few having two authors, or (rarely) three. By the early part of the twentieth century, most readers had two or three authors, by the 1950s, four, and by the 1970s, as many as eight authors and instructional consultants (Woodward 1986).

4.4 Marketing and Adoption

In many countries, including some with national ministries of education, textbooks compete in a partially open marketplace. The selection of textbooks for schools, however, is often enshrouded in a complexity of national, regional, and local politics. In the US, where local control of schools and lay control of the school curriculum are fundamental tenets of the educational process, school board prerogatives for textbook selection often clash with particular citizen interests as well as with state and national politics. State-wide adoption practices, initiated originally for both economic and political reasons, have led to several heavily populated states (e.g., Texas, California) exerting unwarranted influence over the entire elementary and high school textbook trade.

5. The Future of Textbooks

In a science fiction classic entitled *A Round Trip to the Year 2000 or a Flight through Time*, written by Cook (1903), news and novels in the year 2000 are no longer printed. Instead central 'bureaus' transmit voice recordings upon request from phonographic records. However, textbooks still exist and specific mention is made of a 'text-book on Applied Psychology which has a place in every school in the nation' (Cook 1903/1974, p. 109). Whether the projected survival of textbooks was considered as a positive or negative factor by the author is not known; however, while daily newspaper reading has declined steadily since the early 1950s and the novel has been declared dead at various times over the past decades, textbooks have continued to thrive at all levels of education and in all countries where formal education exists.

In the 1960s the main threat to the existing form of textbooks was the programmed textbook, a dull, tedious approach to sequencing subject matter in baby-sized bites, based on the behavioral conditioning studies that the psychologist B. F. Skinner performed with pigeons and rats (Reid 1969). In the 1980s computer-assisted instruction was offered as a high-tech alternative to textbooks and to teacher directed instruction, with the promise that each student could progress at his or her own pace—a goal that was seldom challenged by educators. Today the World Wide Web (WWW), hypertext, and inexpensive, book-sized computers that can display texts on high quality displays, are the assumed replacements.

For the developing world, which is still struggling to recruit and train sufficient numbers of teachers and to provide adequate buildings for schooling, the provision of textbooks written in the national languages and appropriate for their cultures remains a challenge. Similarly, smaller industrialized countries would like to find inexpensive approaches for producing or adapting tertiary textbooks to their languages rather than importing them. Finally, for the historian, the preservation of textbooks remains an issue. Few, if any major research libraries are willing to invest resources in preserving and cataloging textbooks,

particularly elementary and secondary ones. Fortunately, a number of newer research centers and projects for the study of textbooks have been funded, including the Japan Textbook Research Center, The Teaching Resources and Textbook Research Unit in the Faculty of Education at the University of Sydney, Australia, the UNESCO International Textbook Research Network (in cooperation with the Georg Eckert Institute in Germany), and the Textbook Colloquium in the Open University Faculty of Arts in the UK.

See also: Environments for Learning; Internet: Psychological Perspectives; Learning from Text; Learning Theories and Educational Paradigms; Learning to Learn; Literacy and Illiteracy, History of; Literacy and Orality; Literary Texts: Comprehension and Memory; Mass Communication: Normative Frameworks; Mass Communication: Technology; Mass Media: Introduction and Schools of Thought; Mathematical Learning Theory; Mathematical Learning Theory, History of; Publishing as Medium; Reading Skills; School Achievement: Cognitive and Motivational Determinants; School Learning for Transfer; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values

Bibliography

Ahier J 1988 *Industry, Children, and the Nation: An Analysis of National Identity in School Textbooks*. Falmer, London & New York

Altbach P G 1987 Textbooks in comparative context. In: Thomas R M, Kobayashi V N (eds.) *Educational Technology—Its Creation, Development and Cross-cultural Transfer*. Pergamon Press, Oxford, UK

Altbach P G 1991 Textbooks: The international dimension. In: Apple M W, Christian-Smith L K (eds.) *The Politics of the Textbook*. Routledge, New York, pp. 242–58

Altbach P G, Kelly G P (eds.) 1988 *Textbooks in the Third World: Policy, Content and Context*. Garland, New York & London

Apple M W 1986 *Teachers and Texts: A Political Economy of Class and Gender Relations in Education*. Routledge and Kegan Paul, New York

Apple M W, Christian-Smith L K (eds.) 1991 *The Politics of the Textbook*. Routledge, New York & London

Biraimah K L 1988 Gender and textbooks: An African case study. In: Altbach P G, Kelly G P (eds.) *Textbooks in the Third World: Policy, Content and Context*. Garland, New York & London, pp. 117–43

Comenius J A 1658/1887 *Orbis Sensualium Pictus*. Bardeen, Syracuse, NY

Cook W W 1903/1974 *A Round Trip to the Year 2000: Or a Flight through Time*. Hyperion, Westport, CN

Cronbach L J (ed.) 1955 *Text Materials in Modern Education*. University of Illinois Press, Champaign, IL

Cubberley E P 1920 *The History of Education*. Houghton Mifflin, Boston

De Charms R, Moeller G H 1962 Values expressed in American children's readers: 1800–1950. *Journal of Abnormal and Social Psychology* **64**: 136–42

FitzGerald F 1979 *America Revised: History Schoolbooks in the Twentieth Century*. Little, Brown, Boston

Ketcham A F 1986 World War Two as viewed through secondary school textbooks: A report on research in process. *Internationale Schulbuchforschung* **8**: 61–5

Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago

Kwong J 1988 Curriculum in action: Mathematics in China's elementary schools. In: Altbach P G, Kelly G P (eds.) *Textbooks in the Third World: Policy, Content and Context*. Garland, New York & London, pp. 227–45

Lorimer R, Long M 1979–80 Sex-role stereotyping in elementary readers. *Interchange* **10**: 25–45

Luke A 1988 *Literacy, Textbooks and Ideology: Postwar Literacy Instruction and the Mythology of Dick and Jane*. Falmer, London

McClelland D C 1961 *The Achieving Society*. Van Nostrand, Princeton, NJ

Mehlinger H 1991 American textbook reform: Can we learn from the Soviet experience? In: Altbach P G, Kelly G P, Petrie H G, Weis L (eds.) *Textbooks in American Society: Politics, Policy, and Pedagogy*. State University of New York Press, Albany, NY, pp. 145–60

Neumann P H, Cunningham M A 1982 *Mexico's Free Textbooks: Nationalism and the Urgency to Educate*, World Bank Staff Working Papers No. 541. The World Bank, Washington, DC

Reid J M 1969 *An Adventure in Textbooks: 1924–1960*. Bowker, New York

Sewall G T 1987 *American History Textbooks: An Assessment of Quality*. Educational Excellence Network, New York

Shaver J P, Davis O L Jr, Helburn S W 1979 The status of social studies education: Impressions from three NSF studies. *Social Education* **43**: 150–3

Skoog G 1984 The coverage of evolution in high school biology textbooks published in the 1980s. *Science Education* **68**: 117–28

Stodolsky S 1989 Is teaching really by the book? In: Jackson P W, Haroutunian-Gordon S (eds.) *From Socrates to Software: The Teacher as Text and Text as Teacher* (Eighty-ninth Yearbook of the National Society for the Study of Education, Pt. 1). National Society for the Study of Education, Chicago, pp. 159–84

Walworth A 1938 *School Histories at War*. Harvard University Press, Cambridge, MA

Weiss I R 1987 *Report of the 1985–86 National Survey of Science and Mathematics Education*. National Science Foundation, Research Triangle Park, NC

Williams R 1980 *Problems in Materialism and Culture*. Verso, London

Woodward A 1986 Taking teaching out of teaching and reading out of learning to read: A historical study of reading textbook teachers' guides, 1920–1980. *Book Research Quarterly* **2**: 53–73

Woodward A, Elliott D L 1990 Textbooks: Consensus and controversy. In: Elliott D L, Woodward A (eds.) *Textbooks and Schooling in the United States* (Eighty-ninth Yearbook of the National Society for the Study of Education, Pt. 1). National Society for the Study of Education, Chicago, pp. 146–61

Zimet S (ed.) 1972 *What Children Learn in School*. Grume & Stratton, New York

R. L. Venezky

Thematic Maps in Geography

As the name implies, a 'thematic map' addresses a specific theme, such as health or climate. The theme can be narrowly focused, as for a map showing infant-mortality rates for a specific year, or comparatively broad and complex, as for a map relating health-care delivery and mortality. In both function and content, thematic maps are fundamentally different from navigation maps, which serve the way-finding needs of drivers, pilots, sailors, and tourists, and general-purpose or reference maps, which portray a diverse set of basic features such as coastlines, rivers, terrain, cities, transport routes, and political boundaries (Robinson and Petchenik 1976). (For a discussion of other types of maps, see *Ethnic Conflict, Geography of*). Comparatively rare at the outset of the nineteenth century, thematic maps as a group have become more numerous, varied, and intellectually richer than either of these other two types (Robinson 1982). This article examines the distinctive elements of thematic maps, including principal strategies and pitfalls.

1. *Elements of Thematic Mapping*

Thematic maps have two main components: a thematic overlay, and a base map. Typically, the base map is already available in a cartographic database or collection of general-purpose maps, whereas the thematic overlay consists of data from a noncartographic source, such as field observations, a government census, or a thematic database. After selecting or specifying a base map with an appropriate geographic scope, map projection, scale, and reference features, the map author adds the graphic symbols and labels that portray the map's unique theme, as shown in Fig. 1. Consisting largely of structural and background information, the base map is usually less visually prominent than the thematic overlay, which focuses on a specific distribution or relationship.

1.1 *An Appropriate Map Projection Is Essential*

Because the earth is spherical and most maps are flat, the base map distorts the geometry of boundaries and other reference features (Maling 1992). The amount of distortion depends in a large measure on the map's 'scale,' defined as the ratio of distance on the map to distance on the earth. On a 'large-scale' map covering a neighborhood, an archeological excavation, or other small area for which the earth's curvature is minimal, the distortion can be negligible. By contrast, a 'small-scale' map covering a continent, a hemisphere, or the entire world must distort large shapes as well as most distances and directions, and might also distort areas and angles. But by selecting an appropriate projection, the map author can avoid distortions likely to mislead readers.

The world map in Fig. 1 illustrates the distortions of the '*plate carrée*,' or 'plain chart,' a map projection for which locations are easily reproduced on a sheet of graph paper. This particular perspective is 'developed'

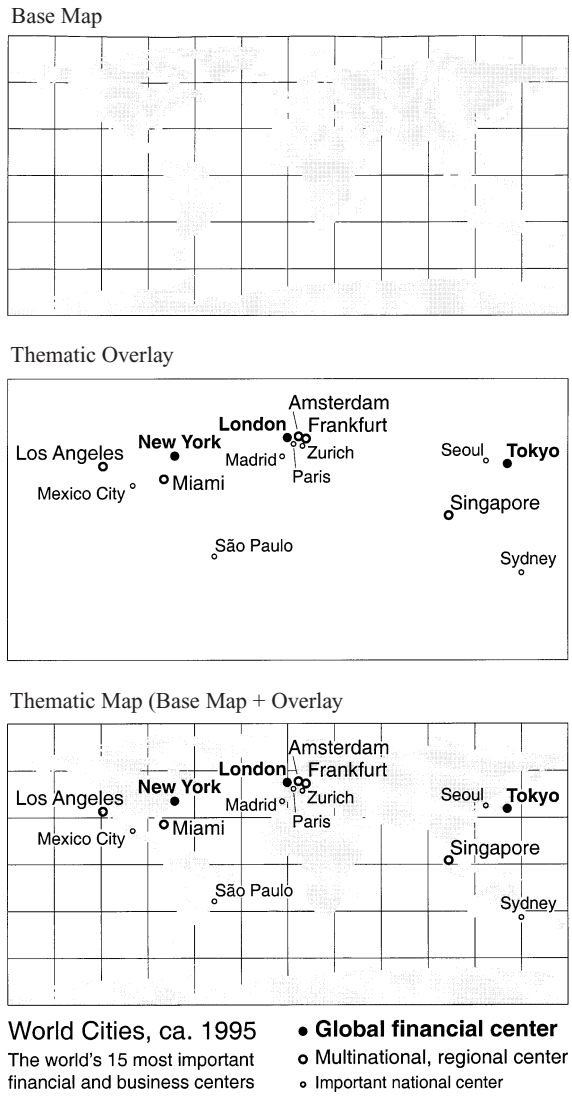

Figure 1
A thematic map consists of a base map and a thematic overlay. On this base map a *plate carrée* projection affords true scale along the equator and all meridians

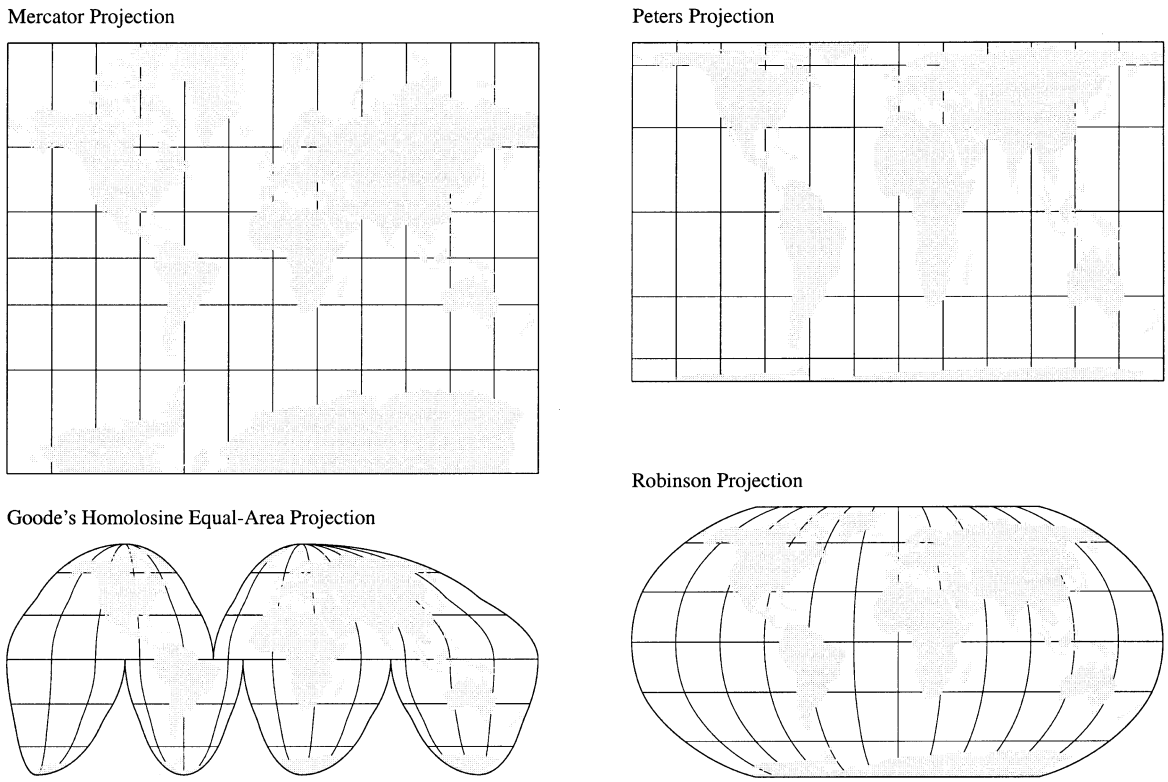

Figure 2
The Peters (upper right) and Goode's (lower left) projections are 'equivalent' projections, which preserve areal relationships. In contrast, the Mercator projection (upper left), which is 'conformal,' preserves angles and small shapes but severely distorts the relative areas of continents and countries. The Robinson projection (lower right), which is neither conformal nor equivalent, offers low distortion of angles and areas, except in relatively unpopulated polar regions

by shrinking the earth to a globe, and then projecting meridians, parallels, and other features onto a cylinder wrapped around the globe and tangent at the Equator. The globe establishes the projection's 'stated scale,' which for most 'cylindrical projections' is valid only where the cylinder is in contact with the globe, in this case along the Equator. Distinctively the *plate carrée* also provides true scale along all meridians, so that north–south scale is constant, and the parallels, like the meridians, are uniformly spaced. Because the meridians do not converge, as on a globe, the east–west scale grows progressively larger with increasing distance from the Equator, so that the map not only inflates the areas of Europe, North America, and other middle-latitude regions, but also portrays Antarctica as twice the size of South America, which is appreciably larger. Even so, the projection offers a comparatively reliable portrait of Africa, which straddles the Equator.

Base maps with severely distorted areas can misrepresent seriously the extent and importance of thematic information (Monmonier 1993). For instance, the *plate carrée*, if used for a map of fertility, automobile ownership, or other phenomenon with marked differences between mid-latitude and tropical countries, will overstate the incidence of patterns occurring in North America, Europe, and Australia, while underrepresenting patterns in Latin America, Africa, and southern Asia.

The Mercator projection (in the upper-left part of Fig. 2) fosters even more egregious areal distortions. Designed to render lines of constant direction as straight lines—sailors and pilots can plot a course as a straight line, and read its bearing with a protractor—the equatorially-centered Mercator projection increases the poleward separation of the parallels steadily so that Greenland looks as large as South America. Because the north–south scale increases to infinity at the poles, which are never shown, the grid in Fig. 2 is truncated at 80° N and 80° S. (More potentially misleading versions of the projection show

only half of Greenland and omit Antarctica altogether.)

Area distortion can be especially troublesome for a 'dot-distribution map,' on which equal-sized dots represent a fixed number of people, hogs, or hectares of wheat. On this type of map, closely spaced dots indicate a higher density than widely separated dots. An accurate portrayal of relative density thus depends upon a base map with little or no areal distortion (Monmonier 1996). For example, if regions A and B, in different portions of the map, each contain a hundred evenly spaced dots within a square centimeter, viewers will assume their densities are similar. But on an inappropriate base map, this apparent similarity could be an artifact of the map's projection: for example, if a globe shows region A, notably farther from the Equator, as only half the size of region B, the exaggerated relative size of region A on the map will have spread its dots farther apart and reduced the apparent density.

Map authors can avoid misleading impressions of density and relative importance by choosing an 'equivalent' projection, designed to preserve areal relationships. 'Equivalency' is one of several properties with which map projections avoid specific kinds of distortion (Snyder 1993). Although all map projections distort most distances, 'equidistant' projections show true distances along straight lines converging on a point at the center (often but not necessarily one of the poles). Although map projections distort most directions, an 'azimuthal' projection displays true relative bearings for all shortest-distance, 'great-circle' routes converging at its center. Although map projections typically distort angles, 'conformal' projections show true angles as well as true shape within very small neighborhoods of every point: a property useful on small-scale weather maps, where angles between fronts, equal-pressure lines, and wind arrows are especially important, and on large-scale topographic maps, which cover small areas in detail and need to minimize distortion of shape and distance. Unfortunately, 'equivalency' and 'conformality' are mutually exclusive. Moreover, equivalent projections distort shape severely on world maps, whereas conformal projections (such as the Mercator projection in Fig. 2) greatly distort area.

The other three projections in Fig. 2 illustrate the trade-offs confronting a map author who needs to choose a world base map for a thematic overlay. The Peters projection (upper-right part of Fig. 2) is one of many equivalent projections with a rectangular grid. Although touted by pro-Third World and human rights groups in the 1970s and 1980s (Crampton 1994), its distinctive combination of severe east–west stretching near the poles and marked north–south stretching near the Equator yields gaunt, noticeably distorted shapes for Latin America, Africa, and most tropical countries, but provides comparatively accurate renderings of the UK, Germany, and other countries closer to 45° N and S. By contrast, Goode's homolosine equal-area projection (lower-left part of Fig. 2) is a composite of 12 locally centered projections that interrupt the oceans in order to describe more accurately the shapes of continents and other land features. A distinctly different trade-off is apparent in the Robinson projection (lower-right part of Fig. 2), which is neither equivalent nor conformal. By tolerating small amounts of areal distortion, mostly near the poles, the Robinson map lessens the distortion of large shapes without interrupting the oceans.

1.2 Appropriate Visual Variables Are also Essential

Verbal and graphic languages are complementary. Words in a thematic map's title, key, and caption link the symbols to natural-language terms and concepts, and are no less important than the lines, 'fill patterns,' and other graphic marks representing features, categories, and data values. But because map readers are almost always more skilled in the conventions and vocabulary of verbal communication, the map's effectiveness often hinges on its author's understanding of fundamental graphic concepts, particularly the appropriate use of visual variables.

Jacques Bertin (1983), a French semiologist who proposed a formal theory of graphic language, recognized eight visual variables: the two 'locational variables' (X-Y coordinates) for describing position on the paper or computer screen, and six 'retinal variables' for portraying a feature's character or category: size, shape, orientation, texture, hue, and value. Of the six, the latter three require clarification: 'texture' refers to the spacing of individual elements in a dashed line or set of parallel lines, 'hue' concerns readily apparent distinctions among blue, green, orange, red, and other named colors, and 'value' pertains to the relative lightness or darkness of a color or gray-tone. Although cartographers employ other retinal variables, such as numerousness, movement, and duration, Bertin's original six account for most symbols on thematic maps (Dent 1999).

Whether a symbol is appropriate depends in part upon whether the data are qualitative or quantitative. For countries or neighborhoods, area symbols varying in hue can be most efficient for qualitative themes such as dominant religion, native tongue, or type of agriculture, whereas a light-to-dark sequence of gray-tones is generally more effective for quantitative differences, as on maps showing local proportions of dairy farms, Roman Catholics, or Francophones. On a map of language regions, for instance, red for English and blue for French underscore differences in kind by looking quite different, at least to viewers with normal color vision. By contrast, shades of gray rely on the ink-intensity metaphor whereby a gray-tone representing 51–65 percent, for example, is at least noticeably darker than a gray-tone representing 36–50

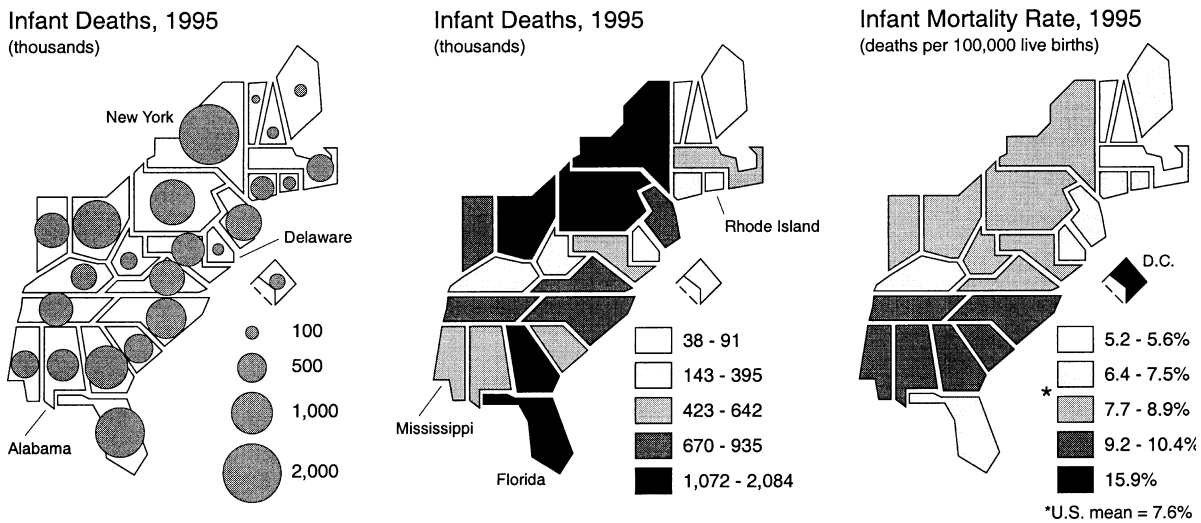

Figure 3
Comparative treatments of count (left and center) and intensity (right) data with symbols varying in size (left) and value (center and right). Misuse of intensity symbols for count data on the center map yields a misleading pattern, In this example, a visibility base map affords clearly readable symbols for Delaware, Rhode Island, and the District of Columbia

percent. Among the retinal variables, hue and shape are better suited for qualitative differences, while size and value are more appropriate for quantitative differences. Orientation is best reserved for showing differences in direction, as with flow lines on a traffic map (Szegö 1987) or wind arrows on a weather map, while texture can reinforce or mimic value, as when closely spaced parallel lines make a fill pattern look darker than a similar symbol with more widely spaced lines.

Quantitative variables require a further distinction between count data and intensity data. Gross national product, tons of steel, and numbers of infant deaths are examples of count data: magnitudes represented most effectively by symbols varying in size so that bigger means more. By contrast, national income per capita, the percentage increase in steel production between 1990 and 2000, and the infant mortality rate are examples of intensity variables, better served by the darker-means-more metaphor of ordered gray-tones or single-hue colors.

Figure 3 illustrates the need to distinguish between count and intensity data. All three maps address infant mortality in the eastern half of the USA. At the left, proportional circles describe differences in the number of infant deaths; a bigger circle unambiguously represents more infant deaths than a smaller circle. In the center, a gray-tone treatment of the same data yields a potentially misleading map with dark, ominous-looking symbols for states with relatively large numbers of deaths and comparatively light symbols for states with few deaths. Stark differences in gray-tone thus suggest a health crisis in New York, Florida, and other relatively populous states, and comparatively benign conditions in moderately populous states like Alabama and Mississippi. By contrast, the gray-tones in the right-hand map offer a far more accurate portrait of geographic variations in the infant mortality rate, which adjusts infant deaths for the number of live births. On this last map, dark symbols for Alabama and Mississippi reflect two of the highest infant death rates in the nation, while New York and Florida register markedly lower (although moderately high) rates.

2. *Principal Cartographic Strategies*

Although thematic maps constitute a rich cartographic genre, any moderately conscientious examination of atlases, government publications, and scholarly books and articles will confirm the prevalence of a small number of graphic strategies, collectively appropriate for most needs (see *Cartography*). This section of the article identifies the more common varieties and their primary uses.

Aside from location maps, which are more topographic rather than thematic, the most common thematic map is the 'choropleth map,' in which area symbols representing specific categories completely fill the boundaries of countries, provinces, census tracts, and other areal units. Although some choropleth maps are qualitative, most are quantitative, largely because censuses and other systematic surveys have generated

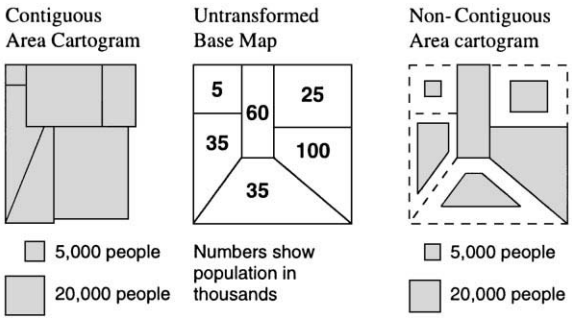

Figure 4
Area cartograms adjust the size of areal units according to a transforming variable such as population. Contiguous cartograms may distort shape severely, whereas non-contiguous solutions preserve shape but lose contiguity

enormous amounts of numerical data for areal units. As the rightmost map in Fig. 3 illustrates, quantitative choropleth maps require intensity data as well as a coherent, light-to-dark series of ordered gray-tones or colors with little, if any, variation in hue. A principal exception is the map showing rates of change, for which contrasting hues, such as red and blue, can distinguish efficiently increases from decreases, while a variation in value differentiates larger rates from smaller rates (Brewer 1997). Because different classifications of the data can yield radically different spatial patterns, the efficiency and effectiveness of a choropleth map depends heavily on its categories (Evans 1977). A wise practice, often ignored, is the use of inherently meaningful category breaks, such as zero on a map of migration rates, or the national mean on a map of per capita incomes by province (Monmonier 1993).

Other common symbols for displaying quantitative data include proportional circles (as in the left-hand map of Fig. 3), 'pie charts' (circles divided into sectors representing subcategories with different hues), and small, equal-sized dots. On a dot-distribution map, each dot represents the same amount, ideally chosen so that the dots barely coalesce in the most dense area. On transportation maps, line symbols describe routes or flows, line width reflects volume of traffic or relative importance, and hues distinguish among highways, railways, and pipelines.

Some thematic maps demand specially tailored base maps (Southworth and Southworth 1982). For London, New York, and many smaller metropolitan areas, 'linear cartograms' of subway or bus systems represent individual routes with highly generalized lines varying in hue. By altering the map projection radically, linear cartograms afford greater detail in the central city, where routes converge and stops are closer. By contrast, on an 'area cartogram', the symbols representing countries or provinces vary in size according to population, wealth, military strength, or some other transforming variable. As Fig. 4 illustrates, contiguous area cartograms may require severe distortions of shape, whereas non-contiguous area cartograms sacrifice contiguity for shape (Olson 1976). In addition to dramatic demonstrations of international disparities in wealth or defense-spending, population cartograms provide a 'demographic basemap' useful for choropleth maps of public health and socioeconomic data (Dorling 1994). Adjusting the base map for differences in population diminishes the exaggerated impact of large units with small populations and ensures visible symbols for small, densely populated places. Another strategy, not based on population, is the 'visibility base map' (illustrated in Fig. 3), which merely provides readable symbols for small, otherwise obscure areal units.

See also: Cartographic Visualization; Cartography, History of; Cognitive Maps; Dynamic Mapping in Geography; Tactile Maps in Geography

Bibliography

Bertin J 1983 *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, WI

Brewer C A 1997 Spectral schemes: Controversial color use on maps. *Cartography and Geographic Information Systems* **24**: 203–20

Crampton J 1994 Cartography's defining moment: The Peters projection controversy, 1974–1990. *Cartographica* **31**(4): 16–32

Dent B D 1999 *Cartography: Thematic Map Design*, 5th edn. William C. Brown, Boston, MA

Dorling D 1994 Cartograms for visualizing human geography. In: Hearnshaw H M, Unwin D J (eds.) *Visualization in Geographical Information Systems*. Wiley, Chichester, UK

Evans I S 1977 The selection of class intervals. *Transactions, Institute of British Geographers* **2**: 98–124

Maling D H 1992 *Coordinate Systems and Map Projections*, 2nd edn. Pergamon Press, Oxford, UK

Monmonier M 1993 *Mapping It Out: Expository Cartography for the Humanities and Social Sciences*. University of Chicago Press, Chicago

Monmonier M 1996 *How to Lie with Maps*, 2nd edn. University of Chicago Press, Chicago

Olson J M 1976 Noncontiguous area cartograms. *The Professional Geographer* **28**: 371–80

Robinson A H 1982 *Early Thematic Mapping in the History of Cartography*. University of Chicago Press, Chicago

Robinson A H, Petchenik B B 1976 *The Nature of Maps: Essays Toward Understanding Maps and Mapping*. University of Chicago Press, Chicago

Snyder J P 1993 *Flattening the Earth: Two Thousand Years of Map Projections*. University of Chicago Press, Chicago

Southworth M, Southworth S 1982 *Maps: A Visual Survey and Design Guide*. Little, Brown, Boston, MA

Szegö J 1987 *Human Cartography: Mapping the World of Man.* Swedish Council for Building Research, Stockholm

M. Monmonier

Theory: Conceptions in the Social Sciences

1. The Main Postwar Ideal—The Deductive–Nomological Notion of Theory

The notion of theory plays a crucial role in social science, as in any science. But few social scientists have discussed the term in a philosophy of science framework. Even fewer have asked whether the term has the same meaning in all social science. Instead, tense debates have raged, with opponents implying—but not explicating—different understandings of theory. Attempting a classification, we find one ideal notion of theory together with four different understandings of theory, all of which have been influential in the social sciences.

The ideal notion of theory dominated postwar science. It originated from early twentieth century positivism. It generalizes the procedures of experimental natural science, physics above all. Two elements are crucial: first, the idea of logical deduction in axiomatic systems: theorems are deduced from a limited number of axioms. Second, a correspondence criterium of truth: empirical sciences depend on correspondence with facts (sense data), which represents a mind-independent external reality (philosophical realism). The deductive system generates hypotheses that are potential universal laws.

Given the laws, it is possible to specify a set of initial conditions under which a predicted result should follow. In an experiment, these conditions are established. If the result confirms the prediction, the hypotheses (laws) are strengthened. The experiment shows that the law represents general knowledge on how the initial conditions cause an outcome. This is also called the covering law model of explanation: the law is the black box that transforms causes into effects. Theory is a set of such laws. Theory is compact knowledge: many regularities can be subsumed under the same general law, making research systematic and cumulative. Such a notion of theory provides a clear-cut criterium of demarcation: testing of law-based deductions. Explanations, which do not satisfy this criterium, are *ad hoc* explanations and must be avoided.

Theory is thus an ever more complete system of universal laws. At the extreme end lies the vision of a complete unity of science possessing nomological knowledge of reality in its entirety. There are nuances between various 'positivist schools,' including very complicated philosophical disputes concerning the notion of causality, which cannot be dealt with here.

In the postwar period, the critical rationalist version of the deductive–nomological ideal, first argued by Karl Popper, has dominated: no theory can ever be verified, only falsified. Theories are not just generalizations; they are hypotheses on universal relations that have so far not been falsified. In later decades, the force of the deductive–nomological notion has been

Table 1
Four notions of theory in social science

	Law oriented	Idealizing	Constructivist	Critical theory
Committed to the deductive-nomological ideal	+	+	÷	÷
Accepts that sciences studying human actions are distinct from generalizing natural sciences	÷	+	+	+
Accepts that ethical fundamentals matter for social science theory	÷	÷	÷	+

weakened, even in the natural sciences. Philosophers of science now discuss much more sophisticated criteria of demarcation; it is also admitted that the notion tells us little about nonexperimental natural sciences like evolutionary biology. However, it seems that in the social sciences, the original and very restrictive notion has retained its status as a regulatory idea.

This is a paradox, since the results of social science's search for universal laws have been disappointing. Merton's (1949, p. 108) statement is as relevant today: 'sociologists (including the writer) may discuss the logical criteria of sociological laws without citing a single instance which fully satisfies these criteria.' If an instance is quoted, Durkheim's statement on the relation between social cohesion and suicide is probably the most frequent choice. But it remains contested whether that 'law' has served social research on suicide in ways that can be compared with the ways in which laws on gravity, thermodynamics, etc. have served natural science. This notion of theory may be sustained if social science could be reduced to natural science. We cannot here specify the various forms of reductionism (physicalism, biologism, psychologism), but if the joys, sorrows, memories, identity, and free will of humans could in some way be analysed as 'no more than the behavior of a vast assembly of nerve cells and their associated molecules' (Crick 1994, p. 3), then social scientific statements could be reduced to natural scientific statements. Although some scholars hold that such a reduction is in principle possible, very few claim that it can be achieved in practice. This topic, however, will not be discussed further here.

As long as no such reduction has been achieved, we must be aware of the paradoxical influence of the deductive–nomological ideal: it has had an enormous influence as a regulative idea, but there is hardly a single example of its successful application in social science. In practice, social scientists adhering to the ideal have applied two different modifications of the ideal: we shall name them the law-oriented and the idealizing notions of theory. Opposed to these notions are the constructivist and critical notions of theory, implied by scholars who do not accept the deductive–nomological ideal.

These four notions have all been influential and involved in applied social science research in the postwar period. Besides their differing commitment to the deductive–nomological ideal, Table 1 shows how they differ on two other counts: on the question of whether the fact that the social sciences study human actions makes any difference, and on the question of whether ethical foundations matter for theory.

2. *The Law-oriented Notion of Theory*

This notion retains a nomological focus at the cost of deductive rigor. This solution can be related to an ambiguity in Popper's original discussion. Despite his support for the covering law model, Popper was critical of the unity of science program. He rather accepted the distinction between generalizing sciences and history. Causal explanations in history relate to singular events, and universal laws here prove empty.

> For the theories or universal laws of generalizing science introduce unity as well as a point of view; they create, for every generalizing science, its problems, and its centres of interest as well as of research, of logical construction, and of presentation. But in history, we have no such unifying theories; or, rather, the host of trivial universal laws we use are taken for granted; they are practically without interest, and totally unable to bring order into the subject matter (Popper 1942, p. 264).

Popper notes particularly that such trivial laws do not define the initial conditions. Popper mostly classified social science together with the generalizing natural sciences. But at times he also regarded social sciences as specifying. The law-oriented notion of theory emerged as a response to this dilemma. The deductive–nomological ideal is retained: social science is seen as generalizing, distinct from the historians' focus on singular processes. But it is admitted that universal laws in the social sciences have proven trivial. Social science must still generalize, but given the absence of truly universal laws, we should focus instead on regularities that apply only within specific contexts. Such a notion of theory as a collection of 'law-like regularities' or 'quasi-laws' is less strict than the deductive–nomological ideal.

In this understanding, the difference between a theoretical hypothesis and inductive generalization evaporates, as does the distinction between 'theoretically based' and '*ad hoc*' explanations. Law-like or bounded regularities are difficult to separate from boundary conditions, since these conditions themselves enter into the definition of the law-like regularities. It is thus hard to devise decisive experiments, even in cases with no ethical obstacles to experiments involving fellow humans. If a theory (a set of quasi-laws) does not explain (is falsified), it will always be possible to argue that slightly different initial conditions applied. In addition, such bounded regularities are also more easy reconcilable with statistical generalizations than the strictly determinist notion of a covering law is.

Merton (1949), one of those who sense the gap between his own methodological ideals and the ways in which social scientists really proceed when they explain something, introduced the notion of middle range theories for such explanations. Middle range theories were legitimated by reference to the deductive–nomological ideal, but allowed contextual explanations unrelated to universal laws. This was the main methodological conviction in all branches of variables-oriented social science in the postwar period.

The dissonance between the deductive–nomological

ideal and real-life explanatory strategies Merton solved by treating middle range theories as temporary. He claimed that the social sciences were still too young. When they matured, as physics already had, the body of middle range theories would converge into a system of universal laws on the model of experimental natural science. Until social science matured, he concluded, it should give priority to middle range theories, deliberately avoiding universal theory. On this notion, the social sciences would be weak on deductive rigor and parsimoniousness, but strong on contextual explanations.

3. *The Idealizing Notion of Theory*

All through its history, the social sciences have been haunted by debates on methods, always returning to the question of whether the subject matter of the social sciences—human action, intention and meaning—require a specific approach to theory. This line of argument made possible another modification of the deductive–nomological ideal, one that retained deductive rigor by replacing law-orientation with idealization.

As a human being, the social scientist has the capacity to understand motives involved in actions and interaction. But these motives cannot be conceived as universal laws without exceptions. Even the simplest economic regularity—that demand increases when the price decreases—has its exceptions. Trying to transcend the dichotomy between nomothetic and ideographic approaches in the German *Methodenstreit*, both Max Weber and Georg Simmel emphasized that laws in the social sciences are ideal types, idealizations of human motives.

These idealizations are often formalized in axiomatic systems. In the law-oriented notion of theory, emphasis was on the experimental criterium of correspondence to facts. In the idealizing notion, the emphasis is rather on the other feature of the covering law model: deduction in axiomatic systems. While law-oriented researchers want to retain a notion of laws by specifying all conditions, idealizing researchers hold that it is impossible to describe 'all the sets of circumstances' (Boudon 1986, p. 65) in which an idealization applies.

They thus focus on the conditions that establish the ideal situation. Under such stylized conditions, human interaction can be understood via mathematical modeling of parametric and strategic rational action (game theory) (Elster 1983). This is the way the most prestigious postwar branch of social science, neoclassical economics, has always gone about. The program of rational choice theory—particularly influential in political science and sociology since the 1980s—strives to generalize such procedures in the other social sciences. This is often formulated as a program for 'methodological individualism,' according to which the human actor is the basic unit in social science explanations.

In these idealized models, the researcher is she who is in extensive command over the initial conditions, notably the preferences that are plugged into the models. These idealized conditions are 'theoretically interesting,' but they do not occur frequently (Boudon 1986, p. 67). It is thus crucial to distinguish social science models from natural science laws: the former yields predictions in an idealized world (classical economists often regarded this as a 'natural order'), but only in a very limited number of real situations. In contrast, the law of gravity applies very widely and in cases of flying objects such as birds or airplanes, the physicist can readily define the conditions for this deviance. 'A law applies to all cases, a model to ideal, that is, to particular cases' (Boudon 1991, p. 68).

Parsimoniousness and deductive power here comes at the cost of a gap between models and the real world. By constructing idealized models, social scientists use their skills in calculating, inferring, idealizing, and guessing with respect to human intentionality. If the model is to be fully idealized, perfect knowledge must be assumed. The scientist assumes perfect knowledge of the knowledge possessed by the actors of her model world. Both the scientist and the actors in the model world are omnipotent. A dilemma in this approach to theory is thus that these very strong assumptions must be relaxed if the model is to claim any empirical relevance. This relaxation makes the models highly complex.

This gap between model and explanation seems even harder to bridge than the theory–explanation gap of the law-oriented approach. But scholars who are concerned about explanatory relevance try to bridge it. They then end up in a methodological compromise quite similar to that of middle range theories (cf. recent discussions on the concept of mechanisms, starting with Elster 1983). The move to the middle range level mostly also implies relaxed assumptions about knowledge, sometimes bringing them quite close to the constructivist position, which shall be presented below.

The two modifications of the deductive–nomological ideal—law-orientation and idealization—have different advantages. But often, the distinction between them is not made, and scholars strive to have it both ways, that is to discover law-like regularities through idealized models. Such attempts (e.g., Elster 1983, Goldthorpe 1996) should be regarded as attempts at synthesis, and be judged with reference to the differences between generalization and idealization, as traced above. In addition, other differences between the two modifications should be taken into account: they also differ with respect to their implied notions of causality. Models of parametric and strategic rationality are mostly simultaneous equation models. While it may be claimed that any equation implies a causal relation, such general interdependence is a looser

notion than the mechanical notion of causality—requiring temporal asymmetry—implied in the law-oriented understanding of theory.

4. The Constructivist Notion of Theory

The term constructivism has many meanings. In philosophy, it is related to a long tradition of skepticism, in our century known under labels such as pragmatism and phenomenology. In the social sciences, more special schools are called constructivist. For the purpose of this survey, a constructivist notion of theory is defined with reference to its questioning of any foundation for the social sciences: Not only is the deductive–nomological ideal rejected, it is also implied that no alternative ethical foundations can be found (Table 1). It should be noted that this constructivist position has not been systematized to the same extent as the deductive–nomological notion of theory. What follows, then, is a synthetic presentation based on several statements that may differ with respect to important details.

Compared with the law-oriented and idealizing notions of theory, the constructivist approach entails a very different notion of knowledge. The deductive–nomological ideal provides a strict separation between scientific knowledge and everyday knowledge. According to the constructivist position, in contrast, the line of demarcation between science and everyday knowledge is 'fuzzy.' Social science theory is not in principle different from everyday knowledge! It may differ in degrees, but since scientists are themselves human beings, their knowledge is also embedded in their everyday life. This is because also science is a human collective, research groups are communities, and their knowledge is the bounded knowledge attained by interacting human beings. Any demarcation of social science is seen as a result of social processes, and thus relative.

Originating in skepticist epistemology, such a notion of knowledge has been implied by many schools in the sociology of knowledge. More recently, interdisciplinary cognitive science has also been a major inspiration to constructivism. Such a notion of theory does not draw its main notions from the routines of experimental natural science, but from the study of knowledge, communication and information in society at large. It also has a strong relationship to applied studies of social problems, which study how social processes define the problems and risk perceptions that create demands for social science research.

Science cannot be easily demarcated, so social processes define (construct) certain realms of knowledge as science. Attempts to apply strict demarcation criteria implies a focus on testing of theories, while constructivists mainly focus on theory formation, studies of the disciplines and professions into which knowledge production is organized, and of continuing struggles to define what science (theory) is. Relations of trust, authority, tacit knowledge, processes of exclusion, and rhetorical strategies influence the scientific community. As Kuhn (1970, p. 21) emphasized, the explanation of scientific discoveries must 'be psychological or sociological … a description of a value system, an ideology, together with an analysis of the institutions through which that system is transmitted and enforced.'

Certainly scientists work in a social subsystem which encourages analytic rigor, systematization, quest for consistence, democratic access to research results so that they can be replicated, etc. But still science is not qualitatively different from everyday knowledge; not something normatively sacred pursued independently of society at large. A sociologist of knowledge may study the deductive–nomological ideal of theory as a powerful rhetorical figure in the contest on how to define science.

We have seen that there was a particularly strong connection between economics and the idealizing notion of theory. The constructivist notion of theory has a particularly strong affinity to anthropology. It is critical of methodological individualism, emphasizing the crucial role of shared knowledge. A strong version of the constructivist position argues that any normative definition of 'science' is in principle not different from the myths found, e.g., in tribal societies studied by anthropologists.

In our everyday interaction, we agree on a view of reality. Such is the case among scientists too: physicists, for instance, use a number of concepts which cannot be 'seen' as facts in reality (force, field, atoms, electrons, quarks). These entities are studied through complex experimental equipment, e.g. particle accelerators. But scientists agree on the reality of these entities. Thereby, they retain their culture, just like other human communities. In tribal societies, people believe in other kinds of entities, but the claim that these entities are real, is quite parallel (Barnes et al. 1996, p. 84). Even in the hardest of sciences, the process of discovery implies visions, analogies and other features that are also found in religious or mythical world views.

Kuhn's famous notion of paradigms is in line with such a view: paradigms are exemplars, synthesizing in them theoretical discussion, experimental practice and instrumentation. It is a 'disciplinary matrix' with three main components: a specific symbolism that allows formalization, specific models for connections in the research field (from heuristic analogies to causal claims) and master examples, exemplary solutions derived from experimentation and research. The members of specific research communities are socialized into such a matrix via education and recruitment procedures (Kuhn 1974).

Theory is here understood in continuity with the formation of everyday knowledge. Observing some-

thing unfamiliar, we try to understand it in the light of what we already know. We use analogies, extending these to new areas. Such an extension always also implies adjustments. A famous case from the history of science is Darwin, who formulated his theory of natural selection by extending an analogy based on a view of the market mechanism implied in Malthus' theory of population.

Social science has often seen debates on which analogies to allow: the idealizing notion of theory prefers game-theoretical and mathematical analogies, and has been quite skeptical of functionalist theories that extend analogies based on the biological organism. Constructivists rather emphasize that a much larger diversity of analogies exists to be extended in the service of better explanations of social affairs. Interestingly, even Merton notes that middle-range theories start from some notion and its 'associated imagery.'

The idealizing notion of theory emphasized the similarities between social science and history. In the constructivist notion, not only history, but all of the humanities are seen as highly relevant to theory formation in the social sciences. History is the specifying science of singular events, tracing complex causal chains involving both intentions and social structural forces. These chains are reconstructed as narrative sequences, unpredictable in advance but understandable in retrospect (Ricoeur 1986). The historian's craftwork is to interpret the motives of the actors in their historical context, that is to imagine the everyday reality of historical actors. Similarly, the majority of anthropologists define their task as 'thick description' (Geertz 1973) of how everyday action is embedded in more or less local cultures. The various phenomenological, ethnomethodological, and symbolic interactionist (Glaser and Strauss 1967) strands within sociology share such a perspective.

In addition, most poststructuralist approaches imply a constructivist notion of theory: the basic structure of everyday knowledge is here conceived as discourses, which are analyzed by means of analogies drawn from linguistics. Social science theories are embedded in such discourses. As parts of such discourses, the social sciences influence the practices of institutions, media, professions and disciplines of our modern society.

Constructivists agree with the idealizing approach that social science study meaningful human action. Meaning is irreducibly linked to human interaction, which is crucially mediated by symbols and language. The social sciences thus face a challenge that does not bother natural science: how to access the symbolically mediated meaning, motives, preferences, identities of interacting human beings. The idealizing position tackles this challenge by making strict assumptions about knowledge. In contrast, the constructivist focus on everyday knowledge implies that knowledge is always limited, and so rationality is always bounded. The notion of theory is not linked to the ideal of perfect knowledge. Theories do not take the form of axiomatic models, but of contextual understanding of interacting motives (as, e.g., in Goffman's notion of an interaction order). Social science theories are social products; they are sometimes drawn into complex social constellations where other individual or collective actors pursue their own strategies. Theories may influence real developments, but they rarely predict them!

In this notion of theory, the gap between theory and explanation, between models and their empirical application, is minimized. Since all knowledge is limited, any theory is 'middle range'! That label is unnecessary since there is no assumption of a complete, full range theory. Since we do not possess coherent deductive theories accusations of *ad hoc* theorizing are misplaced. Theories are processes (Glaser and Strauss 1967), they cannot be accumulated as a stock of knowledge; they are not specific enough for direct application and they change over time. They do not provide clear-cut accounts of an external reality, since—even if we assume that such a reality is 'out there'—any such allegedly unequivocal representations will eventually be contested. At the extreme end of the constructivist spectrum are positions such as ethnomethodology and poststructuralism, that question any effort to establish theories above the level of the specific case. In particular, poststructuralists attempt to deconstruct any theory in order to celebrate relativism. Other scholars that fall under our broad heading of constructivism hold more moderate positions. Some have even adopted the term realism— e.g., 'pragmatic' and 'non-representational realism' (Pickering 1995, p. 181)—to emphasize this difference.

The deductive–nomological ideal, in this view, leads social science astray by implying that science should converge towards complete knowledge of the laws governing an external reality. Quite the contrary, social science has never seen anything like, e.g., the theory of relativity, social scientists tend to use the modest label 'a theory of … .' Constructivists argue that this plurality of theories should not be regretted, more theories would lead to better explanations (Glaser and Strauss 1967).

The constructivist approach—which we have here defined very broadly—does not generate very parsimonious theories, but by resisting any kind of decontextualization, it often proves strong on explanation. The lesson is that careful consideration of context and motives—structure and action—can never be dispensed with.

5. *Critical Theory*

The constructivist notion of theory is frequently accused of relativism: by devaluing the correspondence to an external reality, constructivism allegedly

implies that the most influential groups in the social system define what scientific theories are. This view is a provocation to the covering law tradition, whose notion of theory was actually designed to expose as pseudo-science fascist theories about such things as the 'fate of a people.' But with the role of nuclear physics in developing nuclear arms, and more broadly, the use of science as legitimation of technocratic rule, criticism rather turned against those who wanted to base social science methodology on the model of experimental natural science.

In this context, J. Habermas (1970) and others revitalized the interwar (Frankfurt School) program of critical theory. This position agrees to the rejection of the deductive–nomological approach, but insists that constructivism must be cured of its relativism by reference to an ethical foundation defined by communicative ethics. This internal linkage of theoretical and ethical reflection sets critical theory apart from the other three notions of theory (Table 1).

The social sciences are related to human interaction, involving understanding based on language. Since it should in principle be possible to agree peacefully in disputes between human beings, a communicative ethics is implied in all social interaction, the social sciences included. The evolution of modern society has enlightened the humans, equipping mankind with a set of human rights. Wherever these are not respected, social movements may form, legitimating their strategies with reference to solid ethical principles. Habermas regards this consensus–criterium of truth as an alternative to the mainstream criterium of correspondence. Social science is defined not by its stock of universal laws, but by its commitment to universal ethical principles. Since the reality studied by social science consist of interacting humans, the truth of its theories cannot be determined by correspondence to the present state of affairs, since social mobilization may prove this state of affairs to be unjust. Since also the social scientist is a participating actor, she cannot139 refrain from making her own judgment on this question.

The mobilization of social movements creates increased public attention to the life situation of certain social groups, supporting claims of equality (universal suffrage, equal social rights, etc.). The standard example is the labor movement, which according to classical nineteenth century Marxism was the 'critical subject' whose interests coincided with the universal interests of mankind. Marx himself, in fact, combined a sociology of knowledge with a law-oriented approach: he exposed 'bourgeois' economists by claiming that their theories were constructions based on the special, nonuniversal interests of their class. Only scientists who related to universal interests could provide an 'objective' understanding of the manmade world, of the 'laws of motion' of contemporary society. Later Marxist philosophers of science and others who try to fuse inspiration from the sociology of knowledge with anti-constructivist realism (Bhaskar 1978, Bourdieu 1990) have taken up this idea. In this survey, however, the features that unite critical theory and constructivism are emphasized.

In the postwar twentieth century, several feminist theorists claimed that the women's movement at that time played a role similar to that of the nineteenth-century labor movement. Fusing such an emphasis on women's 'standpoint' with a phenomenological focus on everyday knowledge, Smith (1990) criticized the male bias of all established sociological theories, using the common female experience of housework and child-rearing as a standard.

This position was countered by feminist theorists who link up with poststructuralism (Scott 1988), claiming that the idea of a standpoint serving as a basic foundation of theory is in itself a construction, as relative as others. More generally, the whole dispute between modernists and postmodernists in contemporary social theory—dating at least back to Habermas' (1985) attack on the poststructuralists, and on Foucault in particular—can be seen as a controversy between critical and constructivist notions of theory.

Critical theory has focused more on ethical foundations and less on concrete paradigms involved in the explanatory efforts of applied social science. Habermas (1968) distinguishes the action sciences from empirical–analytical (natural) sciences and historical–hermeneutic sciences. These are driven by different 'knowledge interests': technical dominance (natural science), maintenance and extension of intersubjective understanding (historical sciences), and liberation. The latter, critical knowledge interest implies that to the extent that social sciences discover law-like regularities, these are evaluated through collective reflection on their legitimacy.

Habermas argued that adaptations of the deductive–nomological ideal of universal laws by social engineering-oriented social scientists turned science and technology into ideologies. The law-like regularities are reinterpreted: the very point that social science laws can never be universal is not regretted, but given an 'offensive' meaning. Their nonuniversal nature means that they can be changed by social action. Later, Habermas (1981) linked technocratic world views and also mainstream economics to the systemic imperatives of modern societies, while critical theories were connected to the defense of *Lebenswelt* against the system's 'colonizing efforts.'

Unique to critical theory is its explicit reflection on the ethical foundations of social science theory. Like constructivism, critical theory is weak on parsimonious theories, but in addition it is mostly also weak on applied explanations. In the latter respect, it may at times converge with the constructivist focus on unique historical sequences, at other times with the law-oriented approach, emphasizing the changeable nature of quasi-laws.

6. Conclusion

We can conclude that there are clear ties between substantial theoretical perspectives and notions of theory in the social sciences. A number of the fundamental debates on social science—between constructivists and realists, between critical theorists and constructivists, between methodological individualism and 'collectivism,' between rationalists and empiricists—in fact imply debates between different understandings of theory and of knowledge more generally. This should not surprise us: social science research is after all in itself a social activity.

The discussion of understandings of theory in the social sciences has brought us around a vast rectangular field. Two varieties of a very strict, formal, and demarcation-focused notion occupy the corners at the one end. A ratr loose, substantial, and middle range notion of theory is found in a third corner and an ethically conscious version in the last corner. Representatives of each side and even of the third and fourth corner in between them are mostly so mutually suspicious that possibilities of coexistence and even interaction are seldom considered. Disputes also tend to interact with distinctions between the separate social sciences. For instance, economists almost exclusively work with an idealizing notion of theory, while anthropologists mostly prefer a constructivist notion. Still, it is striking that the law-oriented approach focus on theory testing, the idealizing approach on modeling, the constructivist approach on theory formation and the critical approach on ethical reflection. Possibilities of crossfertilization cannot, however, be specified here. While one must not gloss over fundamental differences, such possibilities should be explored.

See also: Constructivism/Constructionism: Methodology; Critical Theory: Contemporary; Critical Theory: Frankfurt School; Hermeneutics, Including Critical Theory; History of Science: Constructivist Perspectives

Bibliography

Barnes B, Bloor D, Henry J 1996 *Scientific Knowledge. A Sociological Analysis*. Athlone, London

Bhaskar R 1978 *A Realist Theory of Science*. Harvester, Hassocks, Sussex

Boudon R 1986 *Theories of Social Change*. Cambridge University Press, Cambridge, UK

Bourdieu P 1990 *Outline of a Theory of Practice*. Polity, London

Crick F 1994 *The Astonishing Hypothesis*. Schribner, New York

Elster J 1983 *Explaining Technical Change*. Norwegian University Press, Oslo

Geertz C 1973 *The Interpretation of Cultures*. Basic Books, Chicago

Glaser B G, Strauss A L 1967 *The Discovery of Grounded Theory*. Aldine De Gruyter, Chicago

Goldthorpe J 1996 The quantitative analysis of large-scale dataL21 sets and rational action theory: for a sociological alliance. *European Sociological Review* **12**: 109–26

Habermas J 1968 *Technik und Wissenschaft als 'Ideologie'*. Suhrkamp, Frankfurt am Main, Germany

Habermas J 1970 *Zur Logik der Sozialwissenschaften*. Suhrkamp, Frankfurt am Main, Germany

Habermas J 1981 *Theorie des kommunikativen Handelns*. Suhrkamp, Frankfurt am Main, Germany

Habermas J 1985 *Der Philosophische Diskurs der Moderne*. Suhrkamp, Frankfurt am Main, Germany

Hempel C G 1965 *Aspects of Scientific Explanation*. Free Press, New York

Kuhn T S 1970 Logic of discovery or psychology of research. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK, pp. 1–23

Kuhn T S 1974 Second thoughts on paradigms. In: Suppe F (ed.) *The Structure of Scientific Theories*. University of Illinois Press, Urbana, IL, pp. 459–82

Merton R K 1949 *Social Theory and Social Structure*. Free Press, Glencoe, IL

Pickering A 1995 *The Mangle of Practice*. University of Chicago Press, Chicago

Popper K R 1942 *The Open Society and its Enemies*. Routledge, London, Vol. II

Ricoeur P 1986 *Du Texte à l'Action: Essais d'Herméneutique 2*. Éditions du Seuil, Paris

Scott J 1988 Deconstructing equality—versus difference: or the uses of post-structuralist theory for feminism. *Feminist Studies* **14**: 33–50

Smith D E 1990 *The Conceptual Practices of Power. A Feminist Sociology of Knowledge*. Northeastern University Press, Boston

L. Mjøset

Theory in Archaeology

Archaeological theory is a vibrant tapestry of competing approaches and agendas which mutually constrain and enable one another. Some of these approaches have been imported wholesale from the humanities, social, and natural sciences, while others represent novel reformulations, or extensions, of ideas originally developed in these fields. There have even been some attempts to devise theory that is unique to archaeology. Because of this eclecticism, archaeology stands in a special relation to other disciplines as a testing ground for exploring the articulation of different theories and this is a source of archaeology's strength. This view of archaeology as multiplex implicitly challenges the popular idea that the discipline has developed in a simple linear fashion through a series of stages—for example, from culture-historical, to processual, to post-processual (Preucel and Hodder 1996). Indeed, it suggests that more nuanced and historical studies are needed.

1. The Archaeological Record

What is the archaeological record? This seemingly simple question masks a set of rather complicated philosophical issues which have been and continue to be debated within archaeology. Does the archaeological record consist of the material remains of past societies?; Is it comprised of samples of those remains discovered through survey or excavation; is it the written accounts of those discoveries; or does it refer to our conception of our relation to the past in the present? Three influential models have emerged in response to these questions—the fossil model, the behavioral model, and the textual model.

The fossil model is the view that the archaeological record is the repository of the physical traces of past human activity (Binford 1964). Artifacts, monuments, and buildings all potentially reflect the decisions of past actors and the workings of past social systems. This perspective is based upon the principle that human behavior is patterned, and that past behavior can be inferred, given prior knowledge of pattern-generating behavior. The transfer of meaning from a known context to an unknown one is facilitated by means of analogy (usually direct historical, specific comparative, or general comparative versions). This view is shared by advocates of culture historical archaeology as well as the early processual archaeology.

The behavioral model holds that the archaeological record is better understood as a distorted reflection of the past behavioral system. This perspective was developed by Michael Schiffer (1976), who identified two major kinds of distortions: a) n-transforms, or distortions due to natural processes, and b) c-transforms, distortions due to cultural processes. The assumption here is that after controlling for these distortions, explanation can proceed in a manner similar to the fossil model by identifying and making use of established correlates between pattern and behavior. Although this view has been critiqued on the grounds that the archaeological record cannot properly be viewed as a distortion of itself, it has drawn much needed attention to taphonomic processes (how the archaeological record is formed) and to middle range theory (how archaeologists establish logical connections between present statics and past dynamics).

The textual model was originally introduced by Ian Hodder (1986) and is ultimately derived from the philosopher Paul Ricoeur's idea of 'social action as text.' It regards the archaeological record as meaningfully constituted by individual actors in the context of negotiating their personal identities and social roles within society. In this case, artifacts, monuments, and buildings not only serve utilitarian functions, but they can also be seen to be symbolically charged, potentially encoding such social meanings as ethnicity, gender identity, and worldview. Additional implications of the textual approach are that there can be no final reading since each reading has its own validity and that the intent of the past actor is only one among a number of possible meanings (Tilley 1991). This model is primarily associated with advocates of post-processual archaeology.

2. Epistemology

Archaeologists typically make use of three methods of knowing the past: positivism, historical materialism, and hermeneutics. These methods are often represented in opposition to one another, but they actually differ with respect to their basic goals. Positivism has been used to establish causal explanations of cultural adaptation and behavioral change. Historical materialism has been adopted to understand social evolution in terms of changes in the forces and relations of production. Finally, hermeneutics has been used to provide interpretive understandings of social practices and meanings.

The introduction of positivism in archaeology is associated with the emergence of processual archaeology (also called the New Archaeology) in the early 1960s. The agenda of processual archaeology was to formulate and test general laws of human behavior. These laws were to be used to describe, explain, and predict (or retrodict) cultural similarities and differences represented in the archaeological record. Consistent with the tenets of positivism, unobservables such as ideology or belief systems, were deemed metaphysical (or paleopsychological) concepts and, therefore, of no explanatory relevance. In its zeal to ally itself with the natural sciences, processual ar-chaeology moved quickly to espouse the deductive-nomological (D-N) approach (e.g., Watson et al. 1971) and to a lesser degree structuralism. More recently, it has broadened to include statistical-relevance models (S-R), abduction, and neo-positivist models.

Historical materialism, developed by Karl Marx and modified by his followers, focuses upon the dialectics of the mode of production, traditionally defined as the opposition between the forces and relations of production. The productive forces are the elements of the production process. These include the means of production, the objective conditions of labor, and labor power. The relations of production are the conditions under which the productive forces exist. These are fundamentally based upon class relations with capitalists enjoying the ownership of the means of production and the ability to purchase labor power to create a surplus. Although it has never been a dominant perspective in the West, Marxism has a long legacy, particularly in Europe. For example, one of its most significant early advocates was V. Gordon Childe (1936). Several different versions of historical materialism have now been used including structural

Marxism, dialectical Marxism, and neo-Marxism (Leone and Potter 1999, McGuire 1992).

A hermeneutic, or interpretivist, approach was introduced in the early 1980s in reaction to the limitations of positivism. As in the social sciences in general, archaeologists challenged the artificial distinction between the 'context of discovery' and the 'context of justification' and the characterization of ideational factors as epiphenomenal (Hodder 1982, Shanks and Tilley 1989). A hermeneutic approach seeks to understand the meanings behind actions—why certain actions were taken in a particular case—rather than to explain them as an instance of a general process. It is thus an appreciation of context, rather than an attempt at retrodiction. This method is central to post-processual archaeology and today has diversified to include a number competing variants ranging from post-structuralism, to gender theory, to discourse analysis, to phenomenology.

3. *Cultural Change and Social Evolution*

Archaeology does not currently possess a unified theory of cultural change, similar to the theory of evolution in biology, and most practitioners take this to be a consequence of the complexity of archaeology's subject matter which includes the full range of past human behavior from the appearance of the first6 anatomically modern humans to the emergence of the modern world system. There are three basic approaches to explaining long term, cultural change and social evolution: adaptationalism, transformationalism, and selectionism.

3.1 *Adaptationalism and its Variants*

Adaptationalism is closely associated with processual archaeology and its ecosystemic approach. Following the anthropologist Leslie White, Lewis Binford (1962) defined culture as 'man's extrasomatic means of adaptation.' The principal way that cultures adapted to their environments was by means of economic, social, and ideological subsystems mediated by technology (hence his technomic, soeiotechnic, and ideotechnic categories). It was assumed that a culture's natural state was equilibrium and that this stability was achieved through the operation of social strategies which served as homeostatic mechanisms. Any change within the cultural system was viewed as an adaptive response to some external stress, usually seen as environmental degradation.

Optimality theory is based upon the principles of evolutionary ecology as developed by ecologists and modified by evolutionary anthropologists. It regards culture change as the output of individual organisms seeking to maximize the net rate of energy capture so as to enhance their reproductive fitness. In archaeology the most popular version of these theories is optimal foraging theory (see Bettinger 1991). Advocates typically regard this approach as a heuristic approach, useful in modeling evolutionarily successful behavior, but not sufficient, in and of itself, to explain all human behavior. Indeed, the model is often used to highlight deviations from predicted behavior and these deviations are then explained in terms of cultural factors.

A sophisticated version of adaptationalism is coevolution. This approach studies the interaction between social learning, cultural transmission, and biological evolution. These studies take their inspiration from the coevolutionary or dual inheritance theories of biological anthropology. They typically regard genes and culture as two distinct, but interacting, systems of information inheritance within human populations. Culture is seen to have its own mechanisms and properties and, in this sense, it can be said to have a 'life of its own.' On this view, continuity rather than change is the norm because of the inertial properties of inheritance systems. The most comprehensive application of coevolutionary theory has been the study of plant domestication (Rindos 1984).

Cognitive archaeology is a relatively new direction that seeks to understand the evolution of the mind by means of special human ability to construct and use symbols (Renfrew and Zubrow 1994). There is, as yet, no clear consensus on what constitutes cognitive archaeology and a variety of different approaches are being explored. Nonetheless, it is possible to identify two emerging trajectories as particularly promising. These are those approaches based upon cognitive psychology and those based upon technological choice. An example of the former is the work of Steven Mithen (1996) who has proposed that the evolution of human intelligence is characterized by oscillations between increasing modularity and increasing cognitive fluidity.

3.2 *Transformationalism: Marxist and Political Economy*

Transformationalist approaches regard the roots of culture change as internal to society. Some of these approaches borrow liberally from versions of Marxism and the development of society is assumed to occur through the unity of opposites in a dialectical process. Others have taken a political economic perspective, closer to classical economics, that has emphasized competition and finance systems in the emergence of political institutions.

Prestige goods exchange theory was originally developed by Jonathan Friedman and Michael Rowlands (1978) as one of the pathways of their epigenetic model of social evolution. A prestige system is one where social status is based upon access to the prestige goods necessary for social reproduction. A depen-

dency relation emerges between core and peripheral elites such that only the core elites can maintain centralized control over luxury goods. Elites on the periphery can access these goods only through establishing alliances with the core. The result is an increase in the intensity of exchange and in the quantity of goods produced. An increase in production specialization occurs because, as the core elites desire particular kinds of goods, demand is heightened and this can often lead to standardization. Friedman and Rowlands suggest that prestige systems are typically unstable because of the difficulty of maintaining a clear monopoly over long distances.

World Systems Theory (WST), developed by the political economist Immanuel Wallerstein, is a theory of the spread of capitalism. It explores the structural linkages between First and Third Worlds such that development in one area generates underdevelopment in another. The central process is one of exploitation as nation state cores extract goods from peripheries and redistribute them according to specific rules of allocation. This process is both dynamic and complex since the status of the cores may fluctuate over time, resulting in unstable and changing boundaries. WST has been especially attractive to archaeologists because of the scale at which it operates. It places cultural entities within their larger, historical, political, and economic contexts and is sensitive to the spatial dynamics of control. Archaeologists have been sharply critical of Wallerstein's premodern and modern distinction and argued that its applicability needs to be fully explored in particular cases of State formation with special attention paid to the units of analysis (Stein 1999).

Yet another approach is Peer–Polity Interaction (PPI), a model developed by Colin Renfrew (Renfrew and Cherry 1986) to explain the social dynamics underlying the emergence of State political economies. It identifies social structures (political institutions, ritual systems, and non-verbal language) as the products of long term interactions between adjacent early city states. He identifies three processes—competition, symbolic entrainment, and increased trade. Renfrew notes that both warfare and competitive emulation can lead to the emergence of leaders who establish their authority through gift-giving and feasting and to the intensification of production. Symbolic entrainment can lead to the adoption of writing systems and the institution of kingship. Increased trade can result in craft specialization and mass production aswell as new institutions and facilities to manage surplus goods. PPI has been widely used in both old and new world contexts.

Tim Earle, Terrence D'Altroy, and Elizabeth Brumfiel have established what might be called the political finance model. These scholars regard specialization, not as the result of the political economy adapting to a new problem, but rather as the outcome of elites seeking to solidify and strengthen their political and economic interests (Earle and D'Altroy 1989). This reversal of the standard adaptationalist argument shares strong similarities with prestige goods models (see above), but retains an implicit notion of agency. Central to this approach is the distinction between staple and wealth-finance. Staple-finance systems refers to States that collect and redistribute subsistence goods as part of their juridicial rights. Wealth-finance systems are those States that make use of some form of wealth that is paid out as currency and is exchangeable for subsistence goods in the marketplace. Although both systems are characteristic of complex societies, Brumfiel and Earle (1987) emphasize that it is wealth-finance systems that seem to be the primary force in generating political development.

3.3 Selectionism; the New Darwinian Synthesis?

Selectionism is a controversial new approach which minimizes human intentionality and promotes natural selection to a central position in explaining change in the archaeological record. This ambitious project, largely an American phenomenon, seeks to develop a single general evolutionary theory sufficient to encompass both cultural and genetic transmission in humans and other animals. As defined by Robert Dunnell, the leading proponent of this view, change must be understood as a selective, rather than a transformational, process. He further proposes that artifacts are the 'hard parts of the behavioral segment of phenotypes' and that differential frequencies of artifacts are thus explainable by the same process as those in biology (Dunnell 1989: 44). To date, there have been more theoretical expositions than worked case studies. The few case studies that do exist are focusing on engineering design and performance characteristics of material culture in the explanation of social transitions.

4. Theories of Practice

Dissatisfaction with both the adaptationalist and transformationalist perspectives has led some archaeologists to explore elements of a theory of practice. Practice theories are grounded in post-structuralist approaches in anthropology, sociology, and gender studies and they generally emphasize the interplay between conscious, or goal directed, action and social structures which generate routinized action.

One of the most influential theories of practice is structuration theory. This theory was created by the sociologist Anthony Giddens as a means of bridging the limitations of functionalism and structuralism. A key component of this theory is the recursive relationship between structure and agency such that, through their actions, individuals both shape and are shaped by social rules. Structure can thus be concep-

tualized as both the medium and outcome of social action, a property that Giddens (1984) calls 'the duality of structure.' Some archaeologists have critiqued Giddens's conception of agency, arguing that it is not a cross-cultural category and needs to be conceived as particular to a specific time period (Dobres and Robb 2000). This suggests that a historicity of agency is needed. Others have critiqued it on the basis of its implicit male focus. Most studies of agency have, in fact, tended to center on capacities for action of men, rather than women.

A second version of practice theory is that developed by the anthropologist Pierre Bourdieu (1977). This theory is perhaps best known for the term 'habitus,' the unconscious and durable predispositions that underly routinized practice. Structure is seen as the social conditions of the production of the habitus and conjuncture represents a particular state of the structure. Hodder (1986) has characterized habitus as a linguistic and cultural competence mediating structure and practice. He regards it as of special significance for archaeology because it links social practices with the cultural history of society. One problem that has limited its use in understanding processes of social evolution is that it fails to adequately account for the transformation of the habitus. Perhaps for this reason it has seen its greatest use in ethnoarchaeological studies.

Recently, a number of scholars have drawn attention to embodiment as constitutive of social practice. Meskell (1999) has distinguished two trends. The first is the tendency to treat the 'body as artifact.' Here, the body is manipulated in ways similar to material culture in the context of negotiating identities and meanings. This perspective is particularly characteristic of British and European archaeology. The second is the 'body as the scene for display.' This approach is associated with issues of posture, gesture, costume, and sexuality and is linked to Egyptian, Near Eastern, and Mediterranean archaeology. Meskell's own work has both been inspired by and critical of Judith Butler's (1993) work. For example, she critiques Butler's idea of 'gender as performance' arguing against the assumption that performance is always about the replication of a set of norms on the grounds that there can be no single experience of being a certain biological sex. Rather one must consider the materiality of the body, its social setting, the operations of sex/gender and the individual dimension of living one's own body.

5. *The Social Context of Archaeology*

One of the signal contributions of the recent theoretical debates is the foregrounding of the social context of knowledge production. Many archaeologists, following the conclusions of Michel Foucault and Jurgen Habermas, now accept that the idea of neutral or objective knowledge is a fallacy—that is to say, knowledge is always produced to serve social interests. What still remains controversial, however, is the degree to which specific social contexts shape the practice of archaeology. These issues are particularly well illustrated in the cases of nationalist, feminist, and post-colonial archaeologies.

It seems inescapable that the past has always been used for political purposes (Kohl and Faucett 1995, Meskell 1998). In recent history, perhaps the clearest example of this is the use of Gustav Kossina's work to support the Indo-Aryan origins of the Nazis. Another frequently cited example is the interpretation of the site of Zimbabwe as the product of a white race in an attempt to deny indigenous southern Africans their rightful heritage. These examples, both obvious abuses, are sometimes taken to mean that all political uses of archaeology are inappropriate. This perspective, however, is a convenient overgeneralization that seemingly seeks to preserve the status quo. Archaeology can be used to empower people by providing them with a deeper appreciation of their heritage as, for example, can be seen at colonial Williamsburg where new exhibits features slave life alongside that of the plantation owners. Tourism can thus be a vehicle of neocolonialism or one of empowerment.

Feminism denotes a basic political commitment to confronting and eliminating androcentrism in Western society. Since the mid 1970s, a large literature has developed spanning the arts, the humanities, and social and natural sciences that explores the multiple effects of this bias. The first explicitly feminist archaeology emerged in Scandinavia (Norway) in the 1970s. In the United States, the first widely circulated paper on gender and feminism dates to 1984 (Conkey and Spector 1984) and the first professional conference was held in 1988 (Gero and Conkey 1991). The reasons for this late development in archaeology have been traced to the dominance of positivism and its view of objective knowledge. A number of studies have now addressed equity issues, particularly those issues associated with employment, promotion, grants, and publishing (Nelson et al. 1994). Many feminists have adopted Sandra Harding's view that feminist politics is not only a practical necessity, but is, in fact, essential in order to do good science. The best science is, therefore, the science that is produced under conditions where women and men enjoy equal representation.

A convenient date for the emergence of post-colonial archaeologies on the global stage is 1986. This date is the year the World Archaeological Congress was held at Southampton, UK. This conference is particularly important because it showcased sharp ethical conflicts (between anti-apartheid and academic freedom) for an entire profession (Ucko 1987). At the same time, it provided post-colonial scholars with a valuable opportunity to participate in substantial numbers in an international conference. For the first

time, post-colonial and indigenous issues in archaeology were featured in a series of influential publications (Layton 1989a, 1989b). In the United States, the passage of the Native American Graves Protection and Repatriation Act (NAGPRA) in 1990 has empowered many tribes by providing them with a vehicle for the repatriation of human skeletal remains and associated funerary goods, objects of cultural patrimony, and sacred objects. Although there have been some notable disputes, (e.g., the Kennewick case), archaeologists and native peoples have successfully collaborated on numerous projects of mutual interest.

See also: Cognitive Archaeology; Evolution, History of; Evolutionary Approaches in Archaeology; Evolutionary Theory, Structure of; Gender, Feminism, and Sexuality in Archaeological Studies; Genes and Culture, Coevolution of; Marxist Archaeology; Survey and Excavation (Field Methods) in Archaeology; World Systems Theory

Bibliography

Bettinger R L 1991 *Hunter-Gatherers: Archaeological and Evolutionary Theory*. Plenum Press, New York

Binford L R 1962 Archaeology as anthropology. *American Antiquity* **28**: 217–225

Binford L R 1963 A consideration of archaeological research design. *American Antiquity* **29**: 425–441

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge

Brumfiel E, Earle T K (eds.) 1987 *Specialization, Exchange, and Complex Societies*. Cambridge University Press, Cambridge, UK

Butler J 1993 *Bodies that Matter: On the Discursive Limits of Sex*. Routledge, New York

Conkey M, Spector J 1984 Archaeology and the study of gender. *Advances in Archaeological Method and Theory* **7**: 1–38

Childe V G 1936 *Man Makes Himself*. Watts, London

Dobres M-A, Robb J (eds.) 2000 *Agency in Archaeology*. Routledge, London

Dunnell R C 1989 Aspects of the application of evolutionary theory in archaeology. In: Lamberg-Karlovsky C C (ed.) *Archaeological Thought in America*. Cambridge University Press, Cambridge, pp. 35–49

Earle T L, D'Altroy T N 1989 The political economy of the Inka empire: the archaeology of power and finance. In: Lamberg-Karlovsky C C (ed.) *Archaeological Thought in America*. Cambridge University Press, Cambridge, pp. 183–202

Friedman J, Rowlands M J 1978 Notes toward an epigenetic model of the evolution of civilization. In: Friedman J, Rowlands M J (eds.) *The Evolution of Social Systems*. Duckworth, London, pp. 201–276

Gero J, Conkey M 1991 *Engendering Archaeology*. Blackwell, Oxford, UK

Giddens A 1984 *The Constitution of Society*. University of California Press, Berkeley

Hodder I R 1982 (ed.) *Symbolic and Structural Archaeology*. Cambridge University Press, Cambridge, UK

Hodder I R 1986 *Reading the Past*. Cambridge University Press, Cambridge, UK

Kohl P L, Fawcett C 1995 *Nationalism, Politics, and the Practice of Archaeology*. Cambridge University Press, Cambridge, UK

Layton R (ed.) 1989a *Who Needs the Past: Indigenous Values and Archaeology*. Unwin Hyman, London

Layton R (ed.) 1989b *Conflict in the Archaeology of Living Traditions*. Unwin Hyman, London

Leone M P, Potter R B Jr. (eds.) 1999 *Historical Archaeologies of Capitalism*. Kluwer Academic/Plenum, New York

McGuire R H 1992 *A Marxist Archaeology*. Academic Press, New York

Meskell L 1998 *Archaeology Under Fire: Nationalism, Politics and Heritage in the Eastern Mediterranean and Middle East*. Routledge, London

Meskell L 1999 *Archaeologies of Social Life*. Blackwell, Oxford, UK

Mithen S 1996 *The Prehistory of the Mind*. Thames and Hudson, London

Nelson M, Nelson S, Wylie A (eds.) 1994 *Equity Issues For Women In Archaeology. Archaeological Papers Of The American Anthropological Association*, No. 5 , Washington, DC

O'Brien M J 1996 *Evolutionary Archaeology: Theory and Application*. University of Utah Press, Salt Lake City, UT

Preucel R W, Hodder I 1996 *Contemporary Archaeology in Theory: A Reader*. Blackwell, Oxford

Renfrew C R, Cherry J F (eds.) 1986 *Peer Polity Interaction and Socio-Political Change*. Cambridge University Press, Cambridge, UK

Renfrew C R, Zubrow E B W (eds.) 1994 *The Ancient Mind: Elements of Cognitive Archaeology*. Cambridge University Press, Cambridge, UK

Rindos D 1984 *The Origins of Agriculture: An Evolutionary Perspective*. Academic Press, New York

Schiffer M B 1976 *Behavioral Archaeology*. Academic Press, New York

Shanks M, Tilley C 1989 *Re-constructing Archaeology*. Cambridge University Press, Cambridge, UK

Stein G 1999 *Re-thinking World Systems: Diasporas, Colonies, and Interaction in Uruk Mesopotamia*. University of Arizona Press, Tucson

Tilley C 1991 *Material Culture and Text: The Art of Ambiguity*. Routledge, London

Ucko P 1987 *Academic Freedom and Apartheid: The Story of the World Archaeological Congress*. Duckworth, London

Watson P J, LeBlanc S, Redman C 1971 *Explanation in Archaeology: An Explicitly Scientific Approach*. Columbia University Press, New York

R. W. Preucel

Theory of Mind

The term 'theory of mind' was coined by US psychologist David Premack in a famous article (Premack and Woodruff 1978) reporting experiments carried out on the chimpanzee, Sarah. The question of whether and in what sense animals other than humans have a 'theory of mind' ability remains unsettled. Human beings, however, clearly attribute a wide range of mental states to each other, including intentions, hopes, expectations, imaginings, desires and beliefs. Psychologists have come to use the term 'theory of mind' to denote this everyday ability to attribute mental states to other people and to thereby interpret,

explain, and predict their behavior. Theory of mind concerns our ability, not simply to *have* beliefs as such, but to have beliefs *about* mental states, including the recursive ability to have beliefs about beliefs. Of the various mental states, at least three states appear to be basic to our commonsense explanations: *desires* (which identify an agent's goals), *beliefs* (which reflect what an agent takes the state of the world to be), and *pretence* (because people do not always mean what they do or say).

1. Theory of Mind and Cognitive Psychology

It is all too easy to overlook or take for granted our everyday cognitive abilities, even the most spectacular of them. The ability to think about thinking is a case in point. While our knowledge of cause and effect, space, time, number, object, and so forth have been the focus of intense philosophical scrutiny since classical times, the highly abstract concepts that form cornerstones of our *social intelligence* have until recently received much less attention. One of the most surprising aspects of theory of mind ability is just how early in life it begins to develop. One might have thought that concepts as abstract as desire, pretense, or belief could only be acquired late in life, long after many other more concrete concepts, extensive general knowledge, and sophisticated reasoning abilities are in place. It would be very hard to explain to a young child what a belief is, given that one cannot point at a belief or pick one up or produce an intelligible definition. Fortunately, one does not need to do any of these things because young children acquire the relevant concepts spontaneously. In fact, mental state concepts emerge rapidly and without formal tuition during the preschool period (Leslie 1987, 2000, Perner et al. 1987, Wellman 1990), appear to be universal across cultures (Avis and Harris 1991), and can be acquired even by children with a clinically diminished IQ (Baron-Cohen et al. 1985, Tager-Flusberg et al. 1998).

According to US psychologist, Henry Wellman, the concept of desire develops around two years of age (Wellman 1990). A related sensitivity to intention or goal of action seems to emerge even earlier, as shown by the 18-month-old child's disposition to imitate the action an actor intended to perform but, due to error, did not actually perform (Meltzoff 1995). The Scottish psychologist, Alan Leslie, has shown that the concept of pretending begins to be used by children between 18 and 24 months of age. Strikingly, when children first begin to pretend play by themselves, they simultaneously acquire the ability to recognize that another person is pretending and can even share pretenses with others (Leslie 1987, 1994).

The question at what age the concept of belief is first acquired has been more controversial and has attracted a great deal of study. It is generally accepted that to show a grasp of the concept of belief, the child should understand that what someone believes might be false. Working in Austria, Heinz Wimmer and Josef Perner developed a seminal task in which children are required to predict the behavior of a story character by attributing a false belief to the character (Wimmer and Perner 1983). They tested children between three and six years of age and found that it was not until the children were six years old that a majority of their subjects passed their task.

Subsequently, a British team, working in London, and comprising Simon Baron-Cohen, Alan Leslie, and Uta Frith, simplified the Wimmer and Perner task and found that the vast majority of four year olds whom they tested passed by attributing a false belief to the story character (Baron-Cohen et al. 1985). Since then, such simplified tasks have been found to produce successful performance in four year olds. The child is told a short story, with the aid of pictures or props, in which a character, Sally, has a marble. She places this marble in a basket and covers it. Sally then leaves and goes outside. While she is gone, naughty Ann enters the room, finds the marble in the basket, and moves it to a nearby box, and hides it there. The child is then asked two 'control' questions to test for basic comprehension: 'Where did Sally put the marble in the beginning?' and 'Where is the marble now?' Finally, the child is asked either directly about Sally's belief, 'Where does Sally think the marble is?' or the child is asked to predict Sally's behavior, 'Where will Sally look for the marble?' In order to answer either of the latter questions, the child must attribute to Sally a belief that is different from the child's own belief, and that the child considers to be false.

Baron-Cohen et al. (1985) also found that older children with Down's syndrome who had an average IQ of 64 (in the mildly retarded range) mostly passed this test of false belief. However, the majority of a third group of children, children with autism, failed the false belief task. The group of children with autism had an average IQ of 82 and were thus in the low normal range. Their failure could not therefore be accounted for by mental retardation.

Wimmer and Perner developed a further false belief task in which the child is shown a container which normally contains a well-known candy, and asked what is inside (Perner et al. 1987). After the child answers with the name of the candy, the container is opened and the child is shown that there is in fact a pencil inside. The pencil is replaced and the container closed once more. The child is then told that another person will be shown the container in the same way, and asked to predict what the other person will say is inside. Their findings showed that most four-year-olds could pass this task while most three-year-olds failed. Perner et al. (1989) found that a group of children with autism, despite having verbal mental ages in excess of seven years, nevertheless mostly failed this task. Children with specific language impairment, matched on verbal ability with the autistic group, almost all

passed. Subsequently, these patterns have been widely replicated: Normally developing four-year-old children and older moderate to mildly retarded children mostly pass false belief tasks, and most children with autism fail (see Happé 1995 for a review).

2. *Theories of Theory of Mind—Agreements and Disagreements*

The existence of these early developing yet highly abstract concepts raises the question: What is their origin? Are these concepts innate and, if so, how is that possible? Alternatively, are these concepts acquired and, if so, how is that possible? Either way, we have a challenge. If innate, how could such concepts be built into a cognitive system? If acquired, how could such concepts be gained by children with limited experience and ability? The concept of belief seems as abstract as the concept of the electron, and there seems to be as little direct everyday evidence for the existence of beliefs as there is for the existence of electrons. Thus, the child's theory of mind offers an intriguing case study of the nature of abstract concepts and their origins.

So far, two main answers to the above questions have been proposed. The first answer assumes that possession of abstract concepts depend upon possession of abstract knowledge. This reflects the traditional assumption that a concept is really a name for a packet of knowledge. In the case of theory of mind concepts, the knowledge packets are said to be commonsense theories (Gopnik and Meltzoff 1997, Perner 1991, Wellman 1990). This view of abstract concepts is referred to as the 'theory theory.' The second model assumes that possession of abstract concepts can depend upon a cognitive mechanism, rather than upon knowledge. This view is known as the theory of mind mechanism (ToMM) view (Baron-Cohen 1995, Frith and Frith 1999, Leslie 1987, 1992, 2000). The principle focus of debate has been around the concepts of pretense and belief. Before examining the knowledge versus mechanism dispute, we should outline points on which the two views agree.

First of all, while interpretations of the data differ, there are few disputes about the data themselves. Second, it is not disputed that even young children's theory of mind concepts are mentalistic rather than behavioristic in character. Behaviorist philosophers and psychologists have argued that mental states do not really exist and that talk of mental states is really just a confused way of talking about behavior. However, our common sense, rightly or wrongly, is undeniably committed to the existence of mental states and, whether or not mental state concepts are 'sound,' we certainly possess and employ such concepts. We naturally assume that Sally's behavior will be caused by her desire to find the marble in combination with her belief concerning its whereabouts. We use such belief-desire predictions ubiquitously and to good effect everyday. This suggests that our mental states concepts do refer to something real, probably to complex cognitive properties of agents. If concepts like pretending and believing are mentalistic rather than behavioral, then they cannot be a simple summary of sensory experience. And that will have implications for how these concepts can be acquired. A third point that is not generally in dispute is that the child's theory of mind has some kind of innate basis in a disposition to employ mentalistic concepts. And finally, there is no dispute that in the course of development, the child acquires more and more knowledge about people, about mental states, and about the role mental states play in causing people's behavior.

Beyond the four points above, however, there is much controversy. As we have indicated, the disagreements center on the nature of abstract concepts and the role of concepts in development. If a concept is the 'name' of a body of knowledge, then the only way to acquire a given concept is to acquire the appropriate body of knowledge. Another way to look at this is that possession of a given body of knowledge determines possession of a given concept. So, if this body of knowledge changes, then so too will the associated concept. Since everyone agrees that children's knowledge changes in the course of development, then so too, on this view, must their concepts change. Hence, a major topic in the study of development is the study of conceptual change (Carey 1985). A further implication of this widespread view is that if a given concept is innate, then its associated body of knowledge must also be innate. Often, it is assumed that if something is innate then it cannot change with development. However, some theory theorists have proposed that some knowledge (specifically, knowledge of some theories) is both innate and revisable (Gopnik and Wellman 1994, 1995). On this account, both the child's innate and acquired concepts can and do change.

According to the ToMM view, concepts do not name bodies of knowledge but instead directly represent properties in the world. This means that the representational relation between a concept (in the head) and the property (in the world) that it designates is subserved by a causal mechanism of some sort. For example, the concept 'red' is typically established and sustained by the neurocognitive mechanisms of color vision. Without color vision, it is hard to attend to and learn about redness. But nothing you will learn about redness changes the meaning of the concept 'red,' namely, the property of redness. On this view, concepts do not depend upon knowledge; on the contrary, they are a prerequisite for obtaining knowledge. Therefore, conceptual development is not the same thing as knowledge development. Consequently, innate concepts do not require innate knowledge, and acquired

concepts do not depend upon acquiring knowledge. Instead, on this view, we have the job of inquiring into the nature of the neurocognitive mechanisms that establish and sustain conceptual representations.

The above contrasting assumptions about the nature of concepts lead to very different interpretations of theory of mind development.

3. *Early Development, Normal and Abnormal*

As mentioned earlier, 18-month-old infants will imitate the goal of an action rather than simply the movements comprising the action. It is possible that they do so because infants have knowledge of a theory of goal-directed action. If so, the 'theory of goals' they have knowledge of needs to be discovered and specified. Alternatively, infants may not have knowledge of a theory at all; their imitation may instead simply reflect the operation of a neurocognitive mechanism. The 'mirror neurons' discovered by Rizzolatti in his investigations of cells in the ventral premotor cortex (area F5) in monkeys are suggestive in this regard. Neurons with similar properties are apparently involved in human imitation (Iacoboni et al. 1999). Such mechanisms with their sensitivity to the 'hidden' goals of an actor's actions may form the innate basis for the concept of desire.

Early pretending provides another example of an early social sensitivity where it is hard to imagine what the content of the infant's theory, in this case 'theory of pretending,' might be. Leslie (1987) drew attention to a key and long ignored aspect of early pretending, namely, that when the ability for solitary pretending first emerges, between 18 and 24 months, so does the ability to recognize pretending in other people. Leslie went on to argue that pretending should be understood as part of our theory of mind ability. Instead of proposing that the infant possessed innate knowledge of a theory, he postulated the existence of a specialized neurocognitive mechanism, ToMM, that forms the innate basis for theory of mind concepts, including the concept of pretending. A number of empirical predictions followed from the link between pretending and ToMM, including the prediction that autistic children, who do not develop spontaneous pretending, would be specifically impaired in their understanding of belief.

As we saw, this prediction has been confirmed. However, this confirmation in itself does not prove the link between autism and an impaired neurocognitive mechanism like ToMM. One alternative possibility is that children with autism have some kind of general processing difficulty, for example, with language, or with abstract reasoning, or with working memory, or with some other general factor that is demanded by false belief tasks. Any of these possibilities would provide an alternative explanation to the idea of a specific impairment caused by an impaired ToMM.

To test the alternative general processing explanation for autistic performance on false belief, Leslie and Thaiss (1992) adapted tasks devised by Zaitchik (1990). Zaitchik had developed a task in which a Polaroid photograph is taken of an object in a certain location. The object is then moved to a new location. Following control questions about where the object was when the photograph was taken and where the object is now, the child is asked to say where the object is in the photograph. The child does not get to see the photograph. After all, the child in the Sally and Ann task does not get to see Sally's belief! Zaitchik found that preschool children tend to fail or pass both the photographs task and the Sally and Ann task about the same age. This is not surprising given how similar both tasks are from the point of view of general problem solving. When Leslie and Thaiss tested a group of high-functioning autistic adolescents and a group of normally developing four-year-old children on similar pairs of photographs and false belief tasks, they found that the four year olds performed well on both sets of tasks. The autistic subjects, as expected, performed poorly on the false belief tasks, but their performance on the photograph tasks was nearly perfect. This result too has been replicated and extended. The failure of autistic children on false belief tasks cannot be attributed to the general problem-solving demands of the tasks. The photograph tasks make the same or similar demands and the autistic subjects passed these.

So far, we have considered only comparisons between autistic subjects who fail false belief tasks and normally developing four year olds or nonautistic developmentally disordered subjects who pass. But we should also compare autistic subjects with normally developing three-year-olds who also fail. Do the latter two groups fail false belief tasks for the same reason? The ToMM model predicts that they fail for different reasons. Only in the case of autism is ToMM neurologically impaired; in the typically developing three-year-old, ToMM should be intact, limited only by immaturity. One way to test this prediction is to examine whether ways that help three year olds to pass false belief tasks also help autistic subjects. Siegal and Beattie (1991) found a very simple way to help three year olds achieve better performance on the Sally and Ann task. Instead of asking, 'Where will Sally look for her marble?,' they asked 'Where will Sally look *first* for her marble?' Surian and Leslie (1999) found that while this small change did indeed help three-year-olds to calculate Sally's false belief, it had no effect on autistic performance, which remained just as poor. Other manipulations which help typically developing three-year-olds but which do not help older children with autism have been reported by Roth and Leslie (1991, 1998).

Finally, although the existence of the neurocognitive mechanism ToMM was originally postulated on the basis of purely cognitive findings, recent advances in

brain imaging techniques are beginning to reveal the neural systems basis of this instinct (Frith and Frith 1999).

4. Summary

The ability to detect the hidden mental states of other people is a remarkable cognitive power. It is tempting to assume that the idea of mental states must rest upon the same sort of foundation as many other highly abstract ideas, like electrons or inflation—that is, upon the capacity for scientific reasoning and theorizing. However, the speed and precocity with which ideas about mental states emerge in normal development, together with the patterns of normal and abnormal development reviewed above, suggest a different picture. The basic theory of mind that typically emerges during the preschool period is based upon an in-built mechanism that provides an intuitive and instinctual sensitivity to the mental states of other people.

See also: Autistic Disorder: Psychological; Brain Development, Ontogenetic Neurobiology of; Cognitive Psychology: Overview; Concept Learning and Representation: Models; Early Concept Learning in Children; Mind–Body Dualism; Natural Concepts, Psychology of; Prefrontal Cortex Development and Development of Cognitive Function; Schemas, Frames, and Scripts in Cognitive Psychology

Bibliography

Avis J, Harris P 1991 Belief-desire reasoning among Baka children: Evidence for a universal conception of mind. *Child Development* **62**: 460–7

Baron-Cohen S 1995 *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, Cambridge, MA

Baron-Cohen S, Leslie A M, Frith U 1985 Does the autistic child have a 'theory of mind'? *Cognition* **21**: 37–46

Carey S 1985 *Conceptual Change in Childhood*. MIT Press, Cambridge, MA

Frith C D, Frith U 1999 Interacting minds—A biological basis. *Science* **286**: 1692–5

Gopnik A, Meltzoff A N 1997 *Words, Thoughts, and Theories*. MIT Press, Cambridge, MA

Gopnik A, Wellman H M 1994 The theory theory. In: Hirschfeld L, Gelman S A (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, New York, pp. 257–93

Gopnik A, Wellman H M 1995 Why the child's theory of mind really is a theory. In: Davies M, Stone T (eds.) *Folk Psychology: The Theory of Mind Debate*. Blackwell, Oxford, UK and Cambridge, MA, pp. 232–58

Happé F G E 1995 The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development* **66**: 843–55

Iacoboni M, Woods R P, Brass M, Bekkering H, Mazziotta J C, Rizzolatti G 1999 Cortical mechanisms of human imitation. *Science* **286**: 2526–8

Leslie A M 1987 Pretense and representation: The origins of 'theory of mind.' *Psychological Review* **94**: 412–26

Leslie A M 1992 Pretense, autism, and the 'theory of mind' module. *Current Directions in Psychological Science* **1**: 18–21

Leslie A M 1994 Pretending and believing: Issues in the theory of ToMM. *Cognition* **50**: 211–38

Leslie A M 2000 'Theory of mind' as a mechanism of selective attention. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*, 2nd edn. MIT Press, Cambridge, MA, pp. 1235–47

Leslie A M, Thaiss L 1992 Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition* **43**: 225–51

Meltzoff A N 1995 Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology* **31**: 838–50

Perner J 1991 *Understanding the Representational Mind*. MIT Press, Cambridge, MA

Perner J, Frith U, Leslie A M, Leekam S R 1989 Exploration of the autistic child's theory of mind: Knowledge, belief and communication. *Child Development* **60**: 689–700

Perner J, Leekam S R, Wimmer H 1987 Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology* **5**: 125–37

Premack D, Woodruff G 1978 Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* **4**: 515–26

Roth D, Leslie A M 1991 The recognition of attitude conveyed by utterance: A study of preschool and autistic children. *British Journal of Developmental Psychology* **9**: 315–30

Roth D, Leslie A M 1998 Solving belief problems: Toward a task analysis. *Cognition* **66**: 1–31

Siegal M, Beattie K 1991 Where to look first for children's knowledge of false beliefs. *Cognition* **38**: 1–12

Surian L, Leslie A M 1999 Competence and performance in false belief understanding: A comparison of autistic and normal three-year-old children. *British Journal of Developmental Psychology* **17**: 141–55

Tager-Flusberg H, Boshart J, Baron-Cohen S 1998 Reading the windows to the soul: Evidence of a domain specificity sparing in Williams syndrome. *Journal of Cognitive Neuroscience* **10**: 631–9

Wellman H M 1990 *The Child's Theory of Mind*. MIT Press, Cambridge, MA

Wimmer H, Perner J 1983 Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* **13**: 103–28

Zaitchik D 1990 When representations conflict with reality: The preschooler's problem with false beliefs and 'false' photographs. *Cognition* **35**: 41–68

A. M. Leslie

Theory: Sociological

Sociologists share a map of the history of sociological theory in terms of phases. In the earliest phase, in the nineteenth and early twentieth centuries, a few individuals, working independently, produced massively influential works. By the middle of the twentieth century, these works were recognized as a common 'classical' heritage for the expanding discipline of

sociology. A second, synthesizing phase was initiated through efforts to build a body of general theory. However, advocates of alternative perspectives rejected these postclassic theoretical systems. Thus, a third phase of sociological theory emerged in which multiple paradigms continue alongside renewed efforts to consolidate theoretical ideas. In this article, central attention will be given to the postclassical phase, treating two major efforts of generalized theoretical synthesis and, in each case, a major alternative perspective. The works of the classic phase appear here as intellectual resources for these efforts, both the syntheses and the theoretical alternatives. The recent situation in sociological theory is viewed as it emerges out of this postclassic phase.

1. *Classical Theory and the Convergence Thesis*

The very term 'sociological theory' came to prominence only after World War II, especially in the influential writings of Talcott Parsons. In his classic prewar treatise, *The Structure of Social Action* ([1937], 1968), Parsons argued that a convergence had taken place in recent social theory. He analyzed the writings of four influential European social theorists (Marshall, Pareto, Durkheim, and Weber) with a view to demonstrating that these writers had expanded the scope of analytical social theory beyond the positivist and idealist traditions. Any analytical theory, Parsons argued, treats only selected aspects of a complex reality, formulating two kinds of conceptual schemes. One such scheme is structural in the sense of specifying parts and relations among them that are characteristic of a type of empirical system. The other kind specifies analytical relationships among defined 'analytical elements' or variables. Parsons' convergence thesis pertains only to structural analysis. What this means and how it led to a particularly influential body of writings will be explained below.

1.1 *Social Action Systems*

Parsons treats social entities such as groups as empirical systems of the type he calls social action systems (see *Action, Theories of Social*). The analysis of the structure of the general social action system attempts to delineate types of acts and types of relationships among them. Neoclassical economic theory employs a conceptual scheme of minimal structure. It analyzes a type of action system—a market or a system of markets—consisting of an aggregate of independent rational acts (decisions) by producers and consumers in a competitive setting. By contrast, the four writers that Parsons discussed developed more elaborate structural concepts. Pareto, for instance, starting as a neoclassical economic theorist, later embarked on a project devoted to general sociology with an analytical focus on nonrational components of motivation (e.g., 'sentiments'). But it was Durkheim and Weber who, partly through Parsons' analysis of their works, came to be regarded as the key founders of sociological theory.

In his writings, Durkheim defined some of the enduring concerns of sociological theory. In particular, his major works deal with the nature and types of social solidarity; the analysis of social reality as an emergent realm of social facts such as institutions; the explanation of social facts in terms of other such facts, exemplified in the explanation of variable suicide rates by reference to variable levels of social integration; and the genesis of emotion-laden collective representations and their role in creating and maintaining enduring bonds among people (see Giddens 1972). From Parsons' point of view, these works were key contributions in the extension of the scope of analytical social theory while at the same time their common focus is clearly social integration.

In his writings on religion and society, Weber presented empirical evidence for a historical instance—the rise of modern capitalism—of the general idea that a cultural ethic or value system could influence the direction taken by economic forces. Weber also took on the task of constructing the conceptual foundations of sociology. In doing so, he started from the concept of action, defined as behavior with subjective meaning. He noted a variety of types beyond the instrumental-rational type featured in the market model. Sociology is defined as a discipline dealing with the interpretation and explanation of social action, defined as action involving orientation to the actions of one or more other actors (see Heydebrand 1994). This was the most direct source of Parsons' own conception of society as a system of action.

1.2 *The Fundamental Dynamic Theorem of Sociology*

In arguing that there had been a convergence on a conceptual scheme that generalized economic theory, Parsons went on to suggest what he regarded as a core idea of sociological theory. Namely, a stable social equilibrium—social order—is contingent upon the existence of a system of common values that are institutionalized so as to control the play of economic and political interests. The psychological foundations for this idea were developed further in his postwar writings, leading to his 'fundamental dynamic theorem of sociology' in *The Social System* (1951): the stability of social equilibrium requires the institutionalization of a value system that is also sufficiently internalized in the personalities of members. This idea reflects emphases in the writings of Pareto, Durkheim, and Freud concerning the nonrational foundations of social order.

1.3 Functional Analysis

In Parsons' later work, functional analysis became very prominent in the form of the 'AGIL scheme.' (For the most elaborate statement of the AGIL scheme, see Parsons and Platt 1975, see also *Functionalism in Sociology*.) Any action system, the argument goes, has four functional imperatives: adaptation to its environment (A), attainment of its goals (G), integration of its parts (I) and what Parsons calls latent pattern maintenance (L). Functional subsystems of a society include an economy (social adaptation function), a polity (social goal attainment function for society), societal community (social integrative function), and a 'fiduciary' system (latent social pattern maintenance function). At any level of functional analysis, the other functional subsystems form part of the immediate environment of any one such subsystem. A society itself is one of four functional subsystems of an overall action system, namely, the action-integrative (I) functional system, comprised of institutional structures. Its action environment consists of instances of the three other subsystems of general action: cultural systems (L), personality systems (G) and behavioral systems (A). Parsons describes relationships among these systems in terms of theoretical ideas such as the fundamental dynamic theorem of sociology.

1.4 The Fundamental Problem of Sociological Theory

Through elaboration of numerous ideas in the context of the AGIL scheme, Parsonian theory became an extraordinarily comprehensive but also rather complex framework. The complexity arises in part because Parsons was working on at least three levels of theory: general action theory, social theory, and sociological theory. That is, there is an implied 'tree of theory' in which sociological theory is only one among a family of interrelated analytical theories. Call the root of the tree 'the general theory of action.' The idea that action is behavior with subjective meaning belongs at this level, as well as the notion that any action has both rational and nonrational elements. Then there are four branches, corresponding to the AGIL scheme. One branch is social theory, focused on the integrative functional subsystem of action, the social system. This level too has four branches, one of which is focused on the integrative functional subsystem of the social system. It is this branch that constitutes the core of sociological theory, in Parsons' action framework. It relates to the problem of integration of a system whose structure contains instances of collectivities, roles, norms and values. These structural parts may be quite differentiated, e.g., in values, leading to strains in intergroup relations, among other integrative problems.

This analytical view of sociological theory as focused on the problem of social integration leads to a specification of a fundamental problem: what holds society together? Alternatively, it has been framed as the problem of order. It also has a long history in social thought, as was pointed out by Ralf Dahrendorf (1959). He argued that the Durkheimian–Parsonian mode of addressing this question favors a reply in terms of common values, institutional rules, and socialized motivation that stabilizes a system. By contrast, he argued, a very different type of theoretical answer has been given in another stream of social thought, namely conflict theory (see *Conflict Sociology*). This type of theory is exemplified by Marx's theory: coercion and domination of some by others is what holds society together, but with a constant latent conflict relationship that may break out into more or less major episodes of social change. Where Marx focused on property relations as the structured basis of latent or actual conflict and change, Dahrendorf argued that a more general and analytical theory would make power relations central. Any organization, and not just the private enterprise of capitalism, is treated as a structure of power relations between actors with and without authority. Subsequently, Collins (1975) attempted a synthesis of Durkheimian theory with a Dahrendorf type of conflict theory.

Criticism of Parsons' theory came not only from those who favored a conflict theory alternative but also from theorists adopting other perspectives. Prominent among these was George Homans, and his ideas constitute another mode of generalized synthesis in sociological theory that also was met with critical reactions from alternative perspectives.

2. The Behavioral Foundations of Social Bonds

As equally devoted to generality and synthesis as Parsons, Homans focused more specifically on the second of the two types of conceptual scheme mentioned earlier—the analytical element type.

2.1 The Human Group as a Social System

In the first of his two major theoretical works, *The Human Group* ([1950], 1992), Homans argued that the creation of sociological theory should begin with a scope restriction to small groups, those in which each member could interact with every other. Just as Parsons had looked to major works of leading theorists to create his convergence thesis and to launch his own theory program, Homans looked to significant works of empirical research and to the more specific findings emergent from them. Homans asked: What makes customs customary? In other words, how do we account for order? But here, as in all his work, the focus is not on the problem of integration of distinct groups but the more elementary problem of the

construction, maintenance, or change of relationships among persons that constitute a human group.

In the book, Homans develops the theoretical answer in conjunction with a review of empirical cases: a work group, a street gang, the family in a preliterate society, a group of professionals, and even a small town. However, he also cites Durkheim and Pareto, drawing upon the latter's system idea. Order in the form of social integration is explained through an emergent 'internal system,' given external conditions. Social bonds among members and shared norms are generated by mechanisms that are described in terms of specific hypothesized linkages among analytical elements pertaining to activities, sentiments, and interaction. For instance, the more frequently people interact with each other, the more similar their sentiments, normative ideas, and activities become. Each hypothesis qualifies the others, forming a dynamic system. Clearly, Homans' theory is illustrative of the second of Parsons' two types of conceptual schemes and far more clearly than in Parsons' own work.

2.2 Reductionism

In his second major work, *Social Behavior: Its Elementary Forms* ([1961], 1974), Homans asked the question: If these sorts of analytical hypotheses or laws describe group dynamics and the build-up (or dissolution) of a group, what explains the laws? In searching for this more fundamental level of theorizing, Homans invoked a conceptual scheme from behavioral psychology. A typical element is the frequency of reward, and a key mechanism is reinforcement: the greater the frequency of reward, the greater the frequency of the rewarded behavior. As in his earlier system of hypotheses, each such fundamental behavioral proposition is qualified by the others; for instance, a satiation proposition says that as a reward is given repeatedly, each additional unit of it declines in value. As applied to interaction, the sources of the reward (or cost) are other actors. Hence, interaction is an exchange involving material and nonmaterial goods and social approval is a fundamental category of social reward (see *Exchange: Social*).

The basic theoretical aim is the explanation of social life from a nonsocial foundation. This is analogous to the explanation of molecular levels of existence from a purely atomic basis. That is, it is an explicit theoretical reduction program. But what if atoms only could have the postulated properties they have if these properties emerge out of molecular relations? Then there would be no sense to the reduction. Treating persons as social products implies just that conclusion. For this reason, many theorists deny the premise of 'methodological individualism' that Homans is here presupposing. But this denial, in turn, may rest upon a misunderstanding.

What Homans takes as his fundamental conceptual scheme does not involve the socialized individual of everyday life. It involves the concept of a behaviral act and principles about behavioral processes, such as reinforcement. Behavioral acts, not persons, are the basic theoretical units. The socialized individual and the social system are equally outcomes of complex interactive behavioral processes. Thus, Homans has his own implied tree of theory, with behavior theory as the root. Analyzing group processes is one branch of such a tree.

2.3 Mind, Self, and Society

In this interpretation, Homans can agree with the classical sociological theorist Charles Cooley ([1909], 1962) who argued that individual and society are 'twin-born,' in that the person is socially constructed in social interaction and that a society is a system of interaction. However, Cooley stressed the importance of the concept of the self, a notion that Homans employs without theoretical elucidation. It is a branch of his tree of theory that he does not climb out on. Another classical theorist, the philosopher George Herbert Mead—writing in the first third of the twentieth century—had argued that the self can only emerge in a symbolic environment in which interaction involves a common system of meanings (Mead 1934). This is another branch of Homans' tree that he did not explore. Homans presupposes but does not treat the capacity for symbolization that is part of the human condition. And his behavioral focus intentionally avoids any discussion of the 'internalized conversations' that make up the mind in Mead's view. Hence, 'mind' is another branch that went unexplored. In practice, then, Homans simply took mind, self, and symbols as givens in the pursuit of a pure theory that would formulate and explain group processes.

But this was not how other theorists viewed Homans' project. For them, it was a misconceived exercise in sterile behaviorism. They preferred the social behaviorism of Mead, as reinterpreted by Herbert Blumer (1969). Blumer stressed the existence of interpretive processes by which actors define each other's conduct and creatively develop actions in situations rather than merely automatically acting out pre-existing structural roles. This 'symbolic interactionist' approach won a considerable following within sociology after the 1950s, especially through the influential analyses of interaction produced by Blumer's student, Erving Goffman (1967) (see *Interactionism: Symbolic*).

2.4 Sociological Explanation Revisited

In taking the reductionist approach, Homans accompanied his work with a polemical argument. He took aim at Durkheim, who had argued that what explanation means for sociological theory is a causal account that remains at the level of social facts. For

instance, to explain varying rates of deviance in groups, Durkheimian theory would point to varying levels of solidarity: the greater the solidarity of the group, the lower the rate of deviance from its norms. What Homans argued is that an analytical law or proposition stated in terms of social facts itself requires scientific explanation. We cannot be satisfied with purely group-level laws. What Homans argued was that such a proposition, if it is true, could be derived logically from a behavioral foundation. For instance, in a highly solidary group, members experience or can anticipate receiving high costs for undertaking some deviation from the group's norms, while in a less solidary group, such costs are lower. This is so because 'solidarity,' when analyzed in behavioral terms, means that members value each other's social approval to a certain degree. Hence, the mechanism that explains why varying rates of solidarity lead to varying rates of deviance is behavioral and the Durkheimian law is explained.

3. *The Current State of Sociological Theory*

Subsequent to Homans' work, the philosophy of science and social theory itself underwent some basic shifts of viewpoint. Both were strongly influenced by Thomas Kuhn's *The Structure of Scientific Revolutions* ([1962], 1970) in which it was argued that a normal science is characterized by a shared paradigm but that there are also revolutionary episodes in the history of science involving paradigm shifts.

3.1 *Multiple Paradigm Science*

In application of the paradigm concept to sociology, commentators characterized the field as one with multiple paradigms. By the late 1970s, most texts reflected this consensus, featuring separate chapters on functionalism (Parsons), conflict theory (both critical theory and the Dahrendorf tradition), exchange theory (Homans), symbolic interactionism (Blumer), structuralism (French and US versions), and phenomenology (social constructionism and ethnomethodology). (See Turner 1998 for overviews of these and other perspectives in contemporary sociological theory.) To make the picture even more diverse, two other developments occurred. Feminists launched a wide-ranging critique of sociological theory and helped to make the study of gender a key research topic. Postmodernist sociologists attacked the project of sociological theory as a continuation of the Enlightenment's grand narrative with scientific pretensions that could not succeed. Critics respond by attacking the cognitive relativism of this approach (see *Relativism: Cognitive*).

Some commentators argue that there is no possibility of placing these paradigms under a common intellectual framework, thereby seeing the discipline as permanently fractured and at war with itself. Others regard the situation as a positive one, emphasizing the importance of diverse viewpoints that could be brought to bear on any particular feature of social life. Still others recognize the diversity but argue for integrative theorizing, as described below.

3.2 *The Spirit of Unification*

The two most prominent midcentury efforts in sociological theory—each aiming toward generality and synthesis—have failed on the criterion of acceptance as *the paradigm* of general sociological theory. Yet the spirit of what they tried to accomplish is not gone. We can call it 'the spirit of unification,' meaning a value-commitment to generalizing synthesis efforts in episodes of consolidating components of distinct theoretical systems (Fararo 1989). Robert Merton emphasized this idea in his often-cited paper 'On Sociological Theories of the Middle Range' (in Merton [1949], 1968). A middle range theory employs a general conceptual scheme with analytical elements, but it is scope-restricted to some abstractly specified class of empirical systems, e.g., thermodynamic systems. It cuts through folk classification schemes, explaining intuitively very different empirical systems using the same analytical elements and laws that do not exhaust the content of the empirical system. In short, a middle-range theory is an analytical theory. Its scope is limited, not in the sense of dealing only with a class of concrete entities as classified culturally, but in the sense of treating only a certain system of variables.

A value-commitment to the construction of limited scope but abstract theories coupled with a value commitment to a recursive process of unification of such theories may well be a plausible path for the advance of sociological theory. At present, this approach is most strongly institutionalized in the field of research known as group processes, in which theorists elaborate and integrate their theories over time in connection with the construction of experimental situations that provide opportunities for testing the implications of theories. (For examples of group process research programs in relation to theory growth, see Berger and Zelditch 1993.)

3.3 *Mutations and Innovations in the Paradigms*

Recent developments indicate certain mutations in the paradigms and the emergence of new paradigms. Parsons's theory has been transformed into what is now called 'neofunctionalism,' while Homans's paradigm has mutated into part of 'rational choice theory.' Neofunctionalism departs from Parsons in a number of ways that reflect the influence of cross-paradigmatic critiques, attempting to incorporate ideas from conflict theory and interactionism, for instance, so that its advocates regard it as a new synthesis (Alexander 1985). Rational choice theory has departed from the behavioral psychological foundation that Homans

advocated, often favoring a more mathematically tractable rational action model (see *Rational Choice Theory in Sociology*).

One of the new paradigms is social network analysis (see *Networks: Social*). Although social system theorists often employed 'network' as a metaphor for the relational systems they analyzed, they did not employ formal tools. The social network paradigm incorporates a strong mathematical and statistical foundation into a program of cumulative research on the properties of social networks. But critics argue that there is little actual specification of explanatory mechanisms in the research undertaken under this paradigm.

In terms of influence on empirical research, these three paradigms are not on an equal footing. Social network analysis has been incorporated into a great deal of ongoing research. Indeed, its focus on formal techniques for the description of networks implies that it can be employed as a tool along with virtually any other perspective. Neofunctionalism has had a limited following. Most macrosociological research avoids use of the function concept. A return to functionalism as a dominant approach appears unlikely at this point, although some conceptual breakthrough that reinvigorates it cannot be ruled out.

By contrast, rational choice theory created a considerable debate in the late twentieth century. Given that the classic writers and Parsons created ideas that were seen by sociologists as overcoming the scope limitations of economics, a 'return' to rational choice as a postulate seemed regressive and was criticized on that basis. Defenders of rational choice theory can point to its strong contribution to the formulation of micro–macro linkage. This is particularly clear in the formulation set out by Coleman (1990). Macrolevel systemic givens constrain and enable microlevel situations of actors. Making rational choices based on their internal preferences and the situational constraints, actors then collectively shape macrolevel outcomes. This is not equivalent to Homans' reduction program. Instead it is a working strategy for incorporating agency into any sociological explanation, a point emphasized strongly by other recent theorists who are nevertheless critical of rational choice as an explanatory mechanism (Giddens 1984, Bourdieu 1990, Boudon 1998; see also *Macrosociology–Microsociology*).

If the Parsonian notion is accepted, that the problem of social integration defines the core problem of sociological theory within the broader framework of social theory, the route through the middle-range strategy is not inconsistent with the more general project that Parsons and others have pursued. A key contribution of sociological rational choice theorists has been their sharp theoretical focus on variants of the basic problem, treating coordination, cooperation, and trust, as well as solidarity (see *Solidarity, Sociology of*). At the same time, the synthesis of the Durkheimian theory of solidarity with conflict theory undertaken by Collins (1975) provides a different middle-range perspective on the problem of social integration. Thus, given the centrality of the problem and the existence of explicit theories treating it, the topic may well constitute an important locus of episodes involving theoretical unification. Finally, some of this work illustrates the use of mathematical models in bringing precision and deductive fertility to sociological theory (Doreian and Fararo 1998). Such model building can fulfill a variety of goals, including the clarification of concepts, the representation of processes, and the specification of theoretical constructs that explain a variety of phenomena (Berger et al. 1962). Developments such as these are a major hope for the future of sociological theory.

See also: Behaviorism; Behaviorism, History of; Coleman, James Samuel (1926–95); Durkheim, Emile (1858–1917); Marx, Karl (1818–89); Mead, George Herbert (1863–1931); Parsons, Talcott (1902–79); Postmodernism in Sociology; Sociology, Epistemology of; Sociology: Overview; Traditions in Sociology; Weber, Max (1864–1920)

Bibliography

Alexander J (ed.) 1985 *Neofunctionalism*. Sage, Newbury Park, CA

Berger J, Cohen B P, Snell J L, Zelditch M Jr 1962 *Types of Formalization*. Houghton Mifflin, Boston

Berger J, Zelditch M Jr (eds.) 1993 *Theoretical Research Programs: Studies in the Growth of Theory*. Stanford University Press, Stanford, CA

Blumer H 1969 *Symbolic Interactionism*. Prentice-Hall, Englewood Cliffs, NJ

Boudon R 1998 Social mechanisms without black boxes. In: Hedstrom P, Swedberg R (eds.) *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge University Press, New York, pp. 172–203

Bourdieu P 1990 *The Logic of Practice*. Stanford University Press, Stanford, CA

Coleman J S 1990 *Foundations of Social Theory*. Harvard University Press, Cambridge, MA

Collins R 1975 *Conflict Sociology*. Academic Press, New York

Cooley C H [1909] 1962 *Social Organization*. Schocken, New York

Dahrendorf R 1959 *Class and Class Conflict in Industrial Society*. Stanford University Press, Stanford, CA

Doreian P, Fararo T J (eds.) 1998 *The Problem of Solidarity: Theories and Models*. Gordon and Breach, Amsterdam

Fararo T 1989 *The Meaning of General Theoretical Sociology: Tradition and Formalization*. Cambridge University Press, New York

Giddens A (ed.) 1972 *Emile Durkheim: Selected Writings*. Cambridge University Press, New York

Giddens A 1984 *The Constitution of Society*. University of California Press, Berkeley, CA

Goffman E 1967 *Interaction Ritual*. Doubleday, New York

Heydebrand W (ed.) 1994 *Max Weber: Sociological Writings*. Continuum, New York

Homans G C [1961] 1974 *Social Behavior: Its Elementary Forms*. Harcourt Brace Jovanovich, New York
Homans G C [1950] 1992 *The Human Group*. Transaction, New Brunswick, NJ
Kuhn T [1962] 1970 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
Mead G H 1934 *Mind, Self and Society*. University of Chicago Press, Chicago
Merton R [1949] 1968 *Social Theory and Social Structure*. Free Press, New York
Parsons T 1951 *The Social System*. Free Press, New York
Parsons T [1937] 1968 *The Structure of Social Action*. Free Press, New York
Parsons T, Platt G 1975 *The American University*. Harvard University Press, Cambridge, MA
Turner J 1998 *The Structure of Sociological Theory*, 6th edn. Wadsworth, Belmont, CA

T. J. Fararo

Therapist–Patient Relationship

The therapist–patient relationship is accepted widely as a central aspect of psychotherapy, although there is less agreement on how this relationship plays a role in treatment. This article will examine the varying perspectives on the therapist–patient relationship in both historical writings on psychotherapy and in current research and theory. Methodological issues in attempting to study empirically the therapist–patient relationship will also be delineated, and speculations on future directions for research and theory will be offered. But first, the next section will attempt to define the therapist–patient relationship.

1. *Definition of the Therapist–Patient Relationship*

A 'relationship' is something that emerges when two or more people interact. However, with the therapeutic relationship, and perhaps any human relationship, it becomes difficult to think about or observe this 'emergent' property of the interaction independent of the contributions of each of the participants. Thus, the thoughts, feelings, attitudes, and behaviors of the participants, as well as the technical interventions of the therapist, all potentially play a role in the formation and maintenance of the therapist–patient relationship and need to be considered in any conceptualization of this facet of psychotherapy. In the broadest sense, then, the therapist–patient relationship is the emergent entity that is produced by the expectations, attitudes, thoughts, feelings of the participants about each other, as well as their behaviors towards each other.

Despite this broad and elusive definition of the construct, we know a good therapist–patient relationship, and a bad one, when we see it. A good therapist–patient relationship is usually evident when patient and therapist are acting in a warm, caring, empathic, and productive way towards each other. A bad therapist–patient relationship is characterized by distance, hostility, lack of respect, manipulation, or poor capacity to work effectively together.

The positive, 'reality' based aspects of the therapist–patient relationship are subsumed under the concept of the 'therapeutic alliance.' Bordin's (1979) definition of the therapeutic alliance in particular has had considerable influence in theoretical and research writings. The therapeutic alliance, according to Bordin (1979), consists of three elements: (a) agreement between patient and therapist on the goals of therapy, (b) agreement on the tasks of therapy, and (c) the affective bond between patient and therapist. While the bond incorporates the emotional aspects of the patient–therapist attachment, agreement on goals and tasks refers to expectations or cognitive elements of the relationship. To understand how these aspects of the therapist–patient relationship have come to be accepted widely and how they apply to the full range of different psychotherapies requires a historical perspective, given in the next section.

2. *History of Concepts of the Therapist–Patient Relationship*

Modern views of the therapist–patient relationship can be traced directly to Freud's writings. Freud considered the therapist–patient relationship to be a unique laboratory for observing aspects of patient's personalities that may have relevance to the patient's symptoms or psychopathology. Moreover, he utilized the relationship to effect change in the patient (see *Psychoanalysis in Clinical Psychology* and *Psychoanalysis: Overview*).

Freud described several components of the therapist–patient relationship. In terms of the patient's contributions to the relationship, Freud (1958) described both positive and negative transference reactions. Transference refers to the expectations, desires, thoughts, and feelings that are 'transferred' from a previous relationship on to a new relationship (e.g., the therapist). Positive transference was further divided into the reality-based 'friendly and affectionate aspects of the transference which are admissible to consciousness and which are the vehicle of success' (Freud 1958) vs. other positive feelings and perceptions (e.g., sexual feelings, strong dependency) that were not reality-based, originating instead from the patient's unconscious linking of the therapist with significant past relationships. Similarly, negative transference included negative feelings and perceptions that originated from past relationships and were not evident in the therapist–patient relationship. Finally, Freud acknowledged that distortions in the therapist–

patient relationship might result from the therapist unconsciously linking the patient to significant people in the therapist's past ('countertransference') (see *Transference in Psychoanalysis*).

While the reality-based friendly feelings motivated a patient to stay in treatment and engage in therapeutic work, distorted positive and negative transference reactions served as the basis for the therapist's technical interventions. By illustrating these distortions to the patient, previously unconscious conflicts became conscious and such insight lead to patient improvements.

A number of subsequent psychodynamic theorists expanded upon Freud's discussion of the therapeutic relationship. The term 'ego alliance was used by Sterba (1934) to refer to the patient's capacity to work with the therapist in treatment, alternating between experiencing and observing. Subsequently, Greenson (1965) described this capacity as the 'working alliance,' and theorized that it stemmed from the patient's mature ego functioning in conjunction with an identification with the work orientation of the therapist. Zetzel (1956) discussed the patient's attachment to the therapist (labeling it the 'therapeutic alliance'), and postulated that it originated from positive aspects of the mother–child relationship.

Because the analysis of transference became the defining feature of many approaches to psychodynamically-oriented psychotherapy, a relatively large literature has developed on this aspect of the therapist–patient relationship. Esman (1990) has collected a sample of influential papers on transference that illustrate the perspectives of a range of psychoanalytic writers including Freud, Klein, Winnicott, Gill, Kohut, Kernberg, Lacan, and others.

Although there are historical disagreements about whether the positive aspects of the therapist–patient relationship represent patient distortions or realistic appraisals of the treatment situation, the writings of Bordin (1979) served to mark the beginning of consensus on the importance of the reality-based, collaborative therapist–patient relationship (the 'alliance'). This is not to say that other schools of psychotherapy had been silent on the importance of the therapist–patient relationship. In many ways, client-centered psychotherapy, developed by Rogers, can be seen as an approach that relied exclusively on a positive therapist–patient relationship for inducing change (see *Person-centered Psychotherapy*). Unlike the psychodynamic emphasis on reactions of the patient to the therapist, Rogers' (1951) focus was on the therapist's contribution to the relationship. Rogers claimed that if certain 'facilitative conditions' (therapist empathy, genuineness, and unconditional positive regard) were provided to a patient, the patient's natural tendencies for growth and healing would be activated.

The behavioral school of psychotherapy, however, has often been characterized as minimizing the role of the therapist–patient relationship. This school of therapy describes behavior change in terms of principles of learning (i.e., classical or operant conditioning; social learning theory), with no focus on the context in which therapy takes place—i.e., the therapist–patient relationship. Practicing behavior therapists came to realize that this was an oversimplification. Goldfried and Davison (1976), for example, devoted an entire chapter of their book on clinical behavior therapy to the therapist–patient relationship, describing how behavior therapists can use the therapeutic relationship to directly sample aspects of the patient's problematic behavior (see *Behavior Therapy: Psychological Perspectives*). Moreover, a positive relationship was seen as crucial to facilitating favorable patient expectations for change and receptiveness to the behavioral approach, enlisting patient's active cooperation in treatment, and motivating the patient to attempt new behaviors outside of therapy.

Beck's cognitive therapy approach has long acknowledged that the therapist–patient relationship was important to successful outcome, but it was largely considered a 'given,' with only brief explicit attention to the relationship in the original writings on cognitive therapy as a treatment for depression (Beck et al. 1979) (see *Cognitive Therapy*). However, as cognitive therapy expanded its scope to the treatment of other disorders, especially personality disorders, the therapist–patient relationship has come into focus increasingly as a key element of the process of treatment (Newman 1998).

3. *Emphases in Current Theory*

As mentioned, different schools of psychotherapy have largely converged in their agreement that the therapist–patient relationship is of central importance to successful psychotherapy. Even pychodynamic approaches to treatment, which historically have focused attention on patient distortions about the therapist, now highlight the role of the reality-based aspects of the therapist–patient relationship. The consensus about the importance of the therapeutic alliance is evident in a recent edited volume by Safran and Muran (1998) presenting chapters examining the alliance in psychodynamic, cognitive, experiential, strategic, family, and group psychotherapies.

While each approach to psychotherapy tends to stress somewhat different aspects of the therapeutic alliance, Safran and Muran (1998) have identified a number of principles on which there is basic agreement across therapeutic schools regarding what the therapist can do to facilitate a positive alliance. Among others, these include establishing a bond with the patient by conveying warmth, respect, and interest, outlining the tasks and goals of treatment at the beginning of therapy, focusing on realistic goals,

maintaining a clear therapeutic focus, keeping a balance between being active (which helps maintain the treatment focus and the patient's involvement) vs. receptive to patient needs and experiences, and addressing ruptures in the alliances when they occur.

Despite the agreement that has emerged regarding the importance of the alliance, there remains considerable disagreement about how a positive alliance results in more favorable treatment outcome. One view is that the alliance in itself is curative. This perspective proposes that a positive relationship between therapist and patient provides a corrective emotional experience for the patient. Through their relationship with the therapist, the patient emotionally learns over time that other people are not abusive, controlling, manipulative, or otherwise destructive in relationships, as the patient has experienced others to be in the past. Thus, the patient's core expectations or beliefs about people change, and this improves relationships outside of therapy.

The second view on the role of the therapist–patient relationship posits that a good alliance sets the stage for psychotherapy techniques to have their impact. Whether the techniques are cognitive–behavioral, such as homework assignments, or psychodynamic, such as interpretations of interpersonal patterns, the patient's receptivity to such therapist techniques will depend upon the extent to which the patient has a good bond with the therapist and basically agrees with the therapist on the goals and tasks of therapy. In this model, it is neither a good relationship nor good techniques that lead to better outcome, but rather the interaction of the two—a combination of good techniques in the context of a good relationship.

Both of the above models assume that the alliance is a relatively stable aspect of the therapist–patient relationship. A contrasting perspective has been offered that views the alliance as a relatively fluid process with ongoing ruptures occurring frequently during treatment sessions (Safran et al. 1990). It is unclear at this point whether such moment-to-moment fluctuations in the alliance are the norm or whether they occur primarily in the treatments of certain types of patients (e.g., those with severe problems in interpersonal relationships).

4. *Research on the Therapist–Patient Relationship*

Beginning in the 1970's, a variety of studies reported on the development of scales for assessing the alliance and examined its relation to psychotherapy outcome. Several reliable self-report measures and observer ratings scales (applied to tapes of psychotherapy sessions) were constructed. Many of the scales are linked to the theoretical model described by Bordin (1979).

Horvath and Symonds (1991) summarized the emerging literature on the relation of the alliance to psychotherapy outcome. In a meta-analysis of 24 studies, these authors reported that, on average, there was a small to moderate relationship between the quality of the alliance during treatment and eventual treatment outcome. Moreover, the alliance was found to be related to treatment outcome across diverse psychotherapies. The size of the statistical relationship, however, suggests that the alliance is not the only factor, or even a particularly strong factor, in determining psychotherapy outcome. In brief, while the therapist–patient relationship appears critical, there is no reason to suggest that a good therapist–patient relationship is the only curative factor in psychotherapy.

Several studies have examined the common clinical hypotheses that good psychotherapeutic technique has its greatest impact in the context of a positive alliance, i.e., the alliance and technique interact to produce better outcomes. No definitive results, however, have been reported to date, with some studies reporting evidence for an interaction (Gaston et al. 1994), and other studies yielding inconclusive but suggestive findings with small samples (Gaston et al. 1998).

5. *Methodological Issues in Research*

A host of methodological issues need to be considering when conducting research on the relation of the alliance to treatment outcome. Perhaps the most crucial issue relates to whether a positive alliance is actually a causal factor in producing better treatment outcomes. The reverse causal direction needs to be entertained as well: as symptoms improve, patients are more likely to report a positive attitude about the therapist and treatment. Thus, significant correlations between the alliance (measured during treatment) and outcome may be spurious—early improvement in symptom lead to both a positive alliance and good eventual treatment outcome. This potential confound between the alliance and prior improvement was addressed by Gaston et al. (1991), who found large correlations between the alliance and outcome even when prior improvement was controlled; however, because of small sample sizes the findings were not statistically significant. Thus, this important issue of the direction of causality remains to be addressed in future research.

A methodological issue that might limit findings on the relation of the alliance to outcome is the possibility that the alliance may interact with patient dispositional variables in predicting outcome. For example, Horowitz et al. (1984) found that two of four alliance measures interacted significantly with patient pretreatment motivation for psychotherapy in predicting outcome—a good alliance only matters if the patient is also motivated for treatment.

A further consideration is that while most research has examined average levels of alliance over the course

of psychotherapy or sampled alliance at a single session, the patterns in change in the alliance over treatment might be more related to outcome than average levels. This hypothesis was examined by Patton et al. (1997) who speculated that alliance scores would be initially positive, decline somewhat during the difficult working through middle phase of therapy, and then increase again during the termination phase. The results of the study were consistent with the authors hypothesis.

6. *Future Directions*

The almost universal acceptance of the therapeutic alliance as an essential part of psychotherapy points the way to some obvious future directions for scientific study. If a positive therapist–patient relationship is key to successful outcome, can therapists be trained to maximize the quality of the alliance? If not, is this because good therapists are 'born' not 'made'? Resolution of this issue could have profound implications for the training of psychotherapists. If therapists cannot be taught how to establish and maintain a good alliance, perhaps greater effort should be placed on the selection process for training programs in the disciplines (psychology, psychiatry, social work) which conduct psychotherapy. Priority would be given to applicants who have innate interpersonal skills (e.g., empathy) that are associated with positive therapist–patient relationships. If, however, therapists can be trained to maximize the quality of the alliance, such training should become a core element of all training programs, regardless of which treatment modalities are emphasized.

Rather than focusing on the broad question of whether the alliance predicts treatment outcome, future research is likely to continue to explore more refined and specific questions that attempt to unravel the contexts under which a positive alliance is especially important. Larger and more definitive studies, following up on intriguing preliminary findings with smaller samples, can identify the treatment variables and patient dispositional variables that interact with the therapeutic alliance to produce optimal outcome. Research on human relationships, emerging from basic studies in social cognition and interpersonal processes, is likely to provide direction for understanding the nuances of the therapist–patient relationship. Similarly, as intensive empirical examination of the therapist–patient relationship increases our knowledge about the influence of this relationship, we may find applications of these findings to other helping relationships in medicine and elsewhere in society.

See also: Behavior Therapy with Children; Cognitive and Interpersonal Therapy: Psychiatric Aspects; Dialectical Behavior Therapy; Interpersonal Psychotherapy

Bibliography

Beck A T, Rush A J, Shaw B F, Emery G 1979 *Cognitive Therapy of Depression*. Guilford Press, New York

Bordin E S 1979 The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research, and Practice* **16**: 252–60

Esman A H (ed.) 1990 *Essential Papers on Transference*. New York University Press, New York

Freud S 1958 The dynamics of transference. In: Strachey J (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Original work published 1912), Hogarth London, Vol. 12

Gaston L, Marmar C R, Gallagher D, Thompson L W 1991 Alliance prediction of outcome beyond in-treatment symptomatic change as psychotherapy progresses. *Psychotherapy Research* **1**: 104–13

Gaston L, Piper W E, Debbane E G, Garant J 1994 Alliance and technique for predicting outcome in short- and long-term analytic psychotherapy. *Psychotherapy Research* **4**: 121–35

Gaston L, Thompson L, Gallagher D, Cournoyer L, Gagnon R 1998 Alliance, technique, and their interaction in predicting outcome of behavioral, cognitive, and brief dynamic therapy. *Psychotherapy Research* **8**: 190–209

Goldfried M R, Davison G C 1976 *Clinical Behavior Therapy*. Holt, Rinehart, and Winston, New York

Greenson R R 1965 The working alliance and the transference neuroses. *Psychoanalytic Quarterly* **34**: 155–81

Horowitz M J, Marmar C, Weiss D S, DeWitt K N, Rosenbaum R 1984 Brief psychotherapy of bereavement reactions. *Archives of General Psychiatry* **41**: 438–48

Horvath A O, Symonds B D 1991 Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology* **38**: 139–49

Newman C 1998 The therapeutic relationship and alliance in short-term cognitive therapy. In: Safran J D, Muran J C (eds.) *The Therapeutic Alliance in Brief Psychotherapy*. American Psychological Association Press, Washington, DC

Patton M J, Kivlighan D M, Multon K D 1997 The Missouri psychoanalytic counseling research project: Relation of changes in counseling process to client outcomes. *Journal of Counseling Psychology* **44**: 189–208

Rogers C R 1951 *Client-centered Therapy*. Houghton-Mifflin, Boston

Safran J D, Muran J C (eds.) 1998 *The Therapeutic Alliance in Brief Psychotherapy*. American Psychological Association Press, Washington, DC

Safran J D, McMain S, Crocker P, Murray P 1990 Therapeutic alliance rupture as therapy event for empirical investigation. *Psychotherapy* **27**: 154–65

Sterba R 1934 The fate of the ego in analytic therapy. *International Journal of Psychoanalysis* **15**: 117–25

Zetzel E R 1956 Current concepts of transference. *International Journal of Psychoanalysis* **37**: 369–76

P. Crits-Christoph

Thick Description: Methodology

'Thick description' is the term that Geertz (1973) used to describe ethnography in one of the most famous and influential anthropology texts in the second half of

the twentieth century. The phrase was the title of the first essay in the volume of the collected essays, and Geertz explained at the time that he had written it to state his theoretical position as generally as he could. As a result, thick description has come to imply two different things: on the one hand, good ethnography—what all anthropologists have done and do when they write up the results of solid, careful ethnography—and, on the other hand, what Geertz understood to be the intellectual entailments of this particular methodology.

1. As Methodology

Geertz borrowed the phrase 'thick description' from the philosopher Ryle (1949), who used it to contrast two different ways of describing the same piece of behavior. Ryle's example was a boy's twitch and his wink. Both could be described as 'rapidly contracting his right eyelids.' That Ryle called 'thin description.' Thick description was an account of the intentional, communicative, interpretative, meaning of the behavior: why it was done, how it was read, using which social codes. That, Geertz argued, was the object of ethnography: 'a stratified hierarchy of meaningful structures in terms of which twitches, winks, fake-winks, parody, rehearsals of parodies are produced, perceived and interpreted, and without which they would not … in fact exist, no matter what anyone did or didn't do with his eyelids, (1973, p. 7).

Ethnography is the written description of long term participant observation. Professional anthropologists now typically expect that a person doing serious fieldwork (of the sort required for an anthropological dissertation, for example) will spend at least a year living in the 'field,' in the social setting which they have chosen to study. They expect that fieldworkers will speak the dominant language spoken in that setting, and they expect that fieldworkers will learn about that world not only or even primarily by interviewing its members, but by learning how to participate as a member themselves: how to joke, to be in relationship, to be an appropriate person of their type (a woman in her twenties) in this social setting. If you are studying Londoners who practice witchcraft and magic, for example, you learn how to write rituals, to read tarot cards, and to feel the magic running in your blood. If you study Malagasy villagers who plant rice and perform cattle sacrifice, you work in the rice paddies and offer up a cow. Particularly in recent decades, professional anthropologists have decried their pith-helmeted, informant-paying forebears. They are likely to evaluate the student fieldworker by asking whether the student hung back from human contact or really got to know the people he worked among and came to be accepted by them. The point of this emphasis is to make sure that the fieldworker has grasped the realities and the rules of engagement of the world he has come to study. As Stephen Hugh-Jones once remarked, there is little point of having an abstract debate about what it means when the natives say that they 'see' their gods during rituals if you have not participated in those rituals and learned that they involve hallucinogens.

Data that rely on the collector's participation and subsequent recollection can seem hopelessly flawed and subjective to more quantitative researchers. Professional ethnographers use at least three criteria to help them to judge the reliability and validity of the ethnographic report.

First, they look for external corroboration of the ethnography. In part this is provided by pre-publication peer review by experts in the specific area of the world the ethnography concerns. However, the primary responsibility for external corroboration lies with the ethnographer, who is expected to refer to published materials by field subjects (when relevant) and by other ethnographers to support his or her conclusions. At this point in world history, there are few societies remote enough to be completely unknown. Second, they look at the richness of observational detail. Ethnography can be at risk of being dismissed as hearsay or anecdotal. But when the observational detail, reliably reported, is detailed enough, these objections no longer seem plausible. Ethnographic monographs are often weighty and at times tedious for that reason: the ethnographer supports an argument on a foundation of thousands of specific observations about thousands of events.

Third, they look for evidence that the ethnographer knows the ethnographic area well. As in any discipline, the conventions to judge competence are somewhat arbitrary, but still powerful. A year of full-time fieldwork is considered minimal; more years provide more reassurance about the quality of the data. Ethnographies thus present the specific dates and conditions of the fieldwork. To be able to converse in the language is minimal; to command idioms, whether in the spoken language itself or in the metaphorical language of local practice again provides more reassurance. Ethnographies give evidence of language competence and mastery of the local idioms of practice (the sheep trade, the shaman's invocations, the dancers' litany of foot complaints). To be accepted in the target social world well enough to have conversational and observation access to field subjects is also minimal, and so ethnographies typically describe how that access was acquired and maintained; sometimes these accounts are called 'arrival stories.'

There is, however, a hard-to-specify distinction between the competent ethnography and the really good one. Ethnographies do much more than give an account of who did what when. An ethnographer can sit in a courtyard for one full day and describe exactly what he sees, and his observations will fill a monograph; and the monograph will be unreadable and

unintelligible. By contrast, good ethnographies are accounts of a social world that make sense of that world to those who are alien to it. 'What the ethnographer is in fact faced with,' Geertz explained, 'is a multiplicity of complex conceptual structures, many of them superimposed upon or knitted into one another, which are at once strange, irregular, and inexplicit, and which he must contrive somehow first to grasp and then to render' (1973, p. 10). Good ethnography describes what is really important to understand about this world, both to the participants and to its observers; it conveys what people mean when they say, 'you really get it.' When anthropologists use the phrase 'thick description' to refer to the ethnographic method, they mean to imply that the anthropologist does serious, engaged fieldwork; that he really grasps the social process of the world being studied; and that he writes an ethnography so detailed and so observant that it is utterly persuasive.

2. *As an Epistemology*

'If you want to understand what a science is,' Geertz remarks early on in his famous essay, 'you should look in the first instance not at its theories or its findings, and certainly not at what its apologists say about it; you should look at what the practitioners of it do' (1973, p. 5). Ethnographers immerse themselves in another society for long periods of time, and collect far more bits of information than they could ever report or code or convey. Then they write books that attempt to describe the foreignness of that other world in ways that make it comprehensible to their readers.

As a consequence, Geertz argued, ethnographies are by their nature interpretations. They are not hypothesis driven and predictive. This does not mean that they are unconstrained and undisciplined. 'That is like saying that as a perfectly aseptic environment is impossible, one might as well conduct surgery in a sewer' (1973, p. 30). But they are not normal science—normal social science—in the style of psychological experiments and economics theories. The anthropologist, from this perspective, is more like a literary critic who writes about a novel that anyone could in principle read. They may write about the novel in many ways—explain it to those who do not understand, praise, or critique it, argue that other readers have misunderstood it, suggest that it captures certain human experiences that help its readers to see those experiences differently from before, insist that one way of thinking about it is more illuminating than another. But whatever they do, that novel exists as an art form independent of their presence, and their job is to help others to appreciate it more deeply. 'The claim to attention of an ethnographic account does not rest on its author's ability to capture primitive facts in faraway places and carry them home like a mask or a carving, but on the degree to which he is able to clarify what goes on in these places, to reduce the puzzlement—what manner of men are these?—to which unfamiliar acts emerging out of unknown backgrounds naturally arise, (1973, p. 16).

What can be interpreted, however, is only what can be 'read.' The power in the comparison of ethnography and literary criticism lay not only in the reasonableness of the general claim—fieldwork is indeed different from laboratory science—but in the pragmatic usefulness of its central metaphor. Following Paul Ricouer, Geertz not only compared the anthropologist to a literary critic, but described culture and social interaction as a text—the metaphorical 'literature' that the anthropological critic interpreted. More loosely, he described culture as discourse and as language. He argued that culture—'this acted document' (1973, p. 10)—was public because meaning was public. As an ethnographer, a watcher of others, you have no access to people's minds. What you can see is that they communicate, and you can interpret what they communicate, and what you are after is in effect the language or discourse that they are using, not the ideas they have in their mind. Geertz used Wittgenstein's private language argument to support the claim that culture, like language, was a public, social phenomenon, but his primary concern at the time was that you did not need to understand psychological process in order to be a good ethnographer. Understanding the mind was a psychologist's job; understanding culture was the anthropologist's.

And this central metaphor, of culture as a text, gave the ethnographer something specific to do: to identify symbols that somehow represented the particular social process of that group, and to then explain their meaning. In a much quoted definition Geertz stated: 'Believing, with Max Weber, that man is an animal suspended in webs of significance he himself has spun, I take culture to be those webs, and the analysis of it to be therefore not an experimental science in search of law but an interpretive one in search of meaning' (1973, p. 5). Geertz went on to argue that the kinds of things that anthropologists interpret—rituals, myths, cockfights—are themselves interpretations of the society in which they were found, and so the anthropologist is producing second and third order interpretations for a different purpose. 'The culture of a people is an ensemble of texts, themselves ensembles, which the anthropologist strains to read over the shoulders of those to whom they properly belong' (1973, p. 452).

Because (from this perspective) culture is a semiotic system, so that all representation is in effect interdependent, and because the social behaviors that despite everything anthropologists tended to think of as cultural—rituals, stories, songs; the noninstrumental side of life—are also understood as interpretations of their own society, the anthropologist could focus on some practice that on the surface seemed trivial or irrational, and use that practice to understand and reflect upon the nature of the social whole. Geertz did

this most famously in his essay on the Balinese cockfight. The cockfight is a seeming unimportant and irrational piece of Balinese culture in which grown men bet money they cannot afford to watch birds kick each other to death in illegal matches. Geertz suggested that Balinese men were gripped by their cockfights because the matches flung away the restrained, polished social hierarchy that otherwise defined Balinese daily life, and through his discussion, this small social practice became the prism through which you came to understand fundamental features of Balinese society.

In the decade after these two essays were published, Geertz became one of the leading figures of socio-cultural anthropology, if not its dominant presence, and his work precipitated a paradigm shift in the field and influenced other fields as well, most notably history. The new approach was called 'symbolic' anthropology and contrasted both with functionalism (a primarily British approach which sought to understand how all elements of social practice contributed to the effective functioning of that social world) and structuralism (an initially French approach that sought to identify the formal logic underlying cultural representation). David Schneider and the mid-career Marshal (Sahlins were also associated with this new emphasis on the way cultural practices represented social process, and symbolic anthropology tended to be associated with the department of anthropology at the University of Chicago, where Geertz had been before he left for the Institute of Advanced Study at Princeton. Rosaldo's *Knowledge and Passion* (1980), an ethnography which explores the society of the Filipino Ilongot head-hunters through a dominant emotion term, is a good representative of symbolic anthropology.

Now, however, the phrase that is associated with thick description is 'interpretive anthropology.' That phrase tends to mean that the ethnographer values serious (in depth, detailed) fieldwork and understands the ethnography as an attempt to make the foreign world comprehensible to the reader. The phrase also often implies that the anthropologist is self-consciously aware of the literary nature of the ethnographic enterprise: that ethnography is an act of writing, and that successful anthropologists are first and foremost writers. Finally, the phrase often (but not always) implies that this writerly, interpretive act is the primary goal of ethnography, and that anthropology cannot be a hypothesis-testing science, because its methods will not allow it to be so. As a result, the phrase is sometimes used to imply more about what ethnography is not, than about what it is, and the phrase can be used as a battle cry by those who wish to see anthropology associated with the humanities. Those battles however take the phrase beyond the domain that Geertz first mapped out.

See also: Ethnography; Field Observational Research in Anthropology and Sociology; Interpretation in Anthropology; Participant Observation

Bibliography

Geertz C 1973 *The Interpretation of Cultures: Selected Essays*. Basic Books, New York

Ryle G 1949 *The Concept of Mind*. Hutchinson, London

Rosaldo M 1980 *Knowledge and Passion: Ilongot Notions of Self and Social Life*. Cambridge University Press, New York

T. M. Luhrmann

Think Tanks

The term 'think tank' originated in the USA during World War II. It was used to describe secure and 'sealed environments' for expert strategists preoccupied with military planning (Day 1993). Subsequently, the term was applied to the RAND Corporation (initially a division of the US Airforce). By the 1960s, the term was entrenched in the lexicon and was being applied to independent research institutes throughout the English-speaking world, including bodies created much earlier in the century. Consequently, social science characterization of think tanks has been shaped by Anglo-American experience.

The proliferation of think tanks has been most pronounced in the USA, where generous foundation and corporate support, as well as government-contract research opportunities has been a boost to their development. Additionally, the constitutional architecture, federal structure, and relatively weak party system provided a conducive environment for think tanks to engage in policy debates and interact with government. At least 1200 think tanks have emerged in the USA since 1945 (Hellebust 1997). The stronger party systems, corporatist modes of decision making, strong, and relatively closed bureaucracies, and sometimes weak philanthropic sectors are often said to have dampened think-tank development in the parliamentary systems of Great Britain, its former dominions of Canada, Australia, and New Zealand, and most of Western Europe. Nevertheless, think tanks were present in these systems throughout the last century, with around 100 based in Britain and smaller numbers elsewhere.

Research institutes and think tanks are relatively autonomous organizations engaged in the research and analysis of contemporary issues independently of government, political parties, and pressure groups. They are relatively autonomous, although they are often in resource-dependent relationships with these organizations, funding may come from government sources, but institutes attempt to maintain their research freedom and usually claim not to be beholden to specific interests. Think tanks attempt to influence or inform policy through intellectual argument and

analysis rather than direct lobbying; they are engaged in the intellectual analysis of policy issues and are concerned with the ideas and concepts that underpin policy, although not all their research and activity need be policy relevant. Think thanks collect, synthesize and create a range of information products, often directed towards a political or bureaucratic audience, but also for the benefit of the media, civil society groups, and the general public. They are concerned with knowledge creation just as much as political communication and the effective application of knowledge.

Think tanks defy exact definition. They vary considerably in size, legal form, policy ambit, longevity, organizational structure, standards of inquiry, and political significance. Not only is there considerable scholarly difference over how to identify these organizations, but also the directors and senior scholars of these organizations often make fine distinctions between 'research institute' and 'think tank.' Such disputes usually revolve around the role of advocacy on the one hand, and organizational capacity for quality policy research on the other, with think tanks deemed to do the former and institutes the latter. Often this distinction is false.

Some organizations claim to adopt a 'scientific' or technical approach to social and economic problems; others are overtly partisan or ideologically motivated. While some institutes are routinely engaged in intellectual brokerage and the marketing of ideas, whether in simplified policy-relevant form or in soundbites for the media, others are more academic. Many institutes are discipline based—economic policy think tanks, foreign policy institutes, social policy units, etc. Specialization is a contemporary phenomenon. There are environmental think tanks, regionally focused operations, and those that reflect the communal interests of ethnic groups. While most display a high level of social-scientific expertise or familiarity with governmental structures and processes, and usually with international politics, there is considerable diversity in style and output of think tanks. The 'ink tank' can be poised against the 'think-and-do tank.' In other words, there are significant differences between think tanks that are scholarly in focus and geared towards publication of books and reports, and think tanks that are more activist and engage in grass-roots activity.

The Western view that a think tank requires independence or autonomy from the state, corporate, or other interests in order to be free-thinking does not accord with experiences in other cultures. In many countries, the line between policy intellectuals and the state is blurred to such an extent that to talk of independence as a defining characteristic of think tanks makes little sense. Many organizations now called think tanks operate within government; this is evident in countries such as the People's Republic of China. Some institutes have been founded or incubated in government and subsequently made independent; it is not unusual to find institutes with political patrons or formal links to political parties. Many of the German foundations, for instance, have been established by political parties or have strong ties to the *Länder*, and some German institutes have grown to such a large size that they rival some of the major American institutes. In other circumstances research institutes are attached to corporations, as is evident in Japan, South Korea, and Taiwan. Notwithstanding funding dependence or political affiliation, high-quality research and analysis along with critical advice is feasible.

Accordingly, financial independence can be construed as developing an endowment or having numerous sponsors and a diverse funding base, so that an organization is not dependent on any one benefactor. Scholarly independence is reliant upon certain practices within an institute: the processes of peer review and a commitment to open inquiry rather than directed research. Legal or administrative independence comes from constitution as a private organization. Administrative and financial independence may be a facade under authoritarian or illiberal regimes where censorship and control prevail. Alternatively, some institutes may exhibit a high degree of research freedom despite being found within the state architecture. Appreciation of the political culture of a nation is essential to interpreting think tank independence and intellectual leverage.

A further problem of definition is that it is often difficult to establish a precise distinction between think tanks and other organizations with similar objectives; for example, operating foundations. Think tanks could also be regarded as more intellectual variants of pressure groups of nongovernmental organizations (NGOs) since many now base their arguments on relatively sophisticated research that is sometimes conducted in-house. Nevertheless, there are important differences: think tanks do not engage in public demonstrations and they tend to address multiple policy areas (notwithstanding specialization) rather than focus on a single issue as do many pressure groups. There are similarities with consultancy firms, financial institutions, or legal offices that provide contract research and analysis, although these organizations engage in a wider range of activity than research; yet many think tanks only survive through their consultancy services and contract research.

Some think tanks have been described as 'universities without students' (Weaver 1989, p. 564); while the relationship between think tanks and universities has been close in most political systems, important differences are usually observed, for example, think tanks are not normally degree-granting institutions. There are a few exceptions, however, notably RAND in the USA and FLACSO in Latin America. In terms of future think-tank development, the blurring of boundaries and the overlap of objectives and activities

between institutes and other organizations in society may mean that these organizations lose organizational distinctiveness.

In the last two decades of the twentieth century, the number of think tanks has proliferated. Countries where think tanks were already present such as the USA, Britain, Sweden, Canada, Japan, Austria, and Germany witnessed further organizational growth. Increased competition in the think-tank industry has often encouraged policy advocacy and the politicization of institutes, most particularly in the USA. The Heritage Foundation is usually cited as the exemplar. Elsewhere, democratic consolidation, economic development, and greater prospects of political stability in Latin America and Asia provided fertile conditions for think-tank development. The demise of the Soviet Union also opened political spaces for policy entrepreneurs. The think-tank model has been exported around the world. The term 'think tank' has been adopted in its English wording, with all its cultural connotations (Krastev 2000). The global think tank boom development has been prompted by foundations, corporations, and other nonstate actors demanding high-quality research, policy analysis, and ideological argumentation on the one hand, but, on the other, by grants and other funding from governments and international organizations seeking to extend policy-analytic capacities, aid civil society development, or promote human capital development.

Beyond the nation-state, there are strong signs of think-tank adaptation and evolution. The activities of international organizations such as the United Nations agencies, the World Bank, the International Monetary Fund, and the World Trade Organization have drawn think tanks into their ambit. The European Community provides yet another institutional forum for think-tank activity with the emergence of EU-wide think tanks disengaged from specific national identities. Furthermore, with the revolution in information and communication technology, the possibilities for policy research disconnected from specific organizational settings has become increasingly feasible and fashionable. It has also made international research exchange and collaboration between think tanks commonplace. Global and regional think-tank networks have become extensive.

In tandem with this extensive organizational growth, scholarly interest has expanded. Where discussion of think tanks was once limited almost exclusively to Anglo-American systems (see *inter alia*, Weaver 1989, Stone 1996, Smith 1991), interest has expanded to other OECD nations. Wider disciplinary interest has resulted from a core of scholars interested in comparative politics (Thunert 2000) and public policy to include international relations and security studies specialists as well as a greater array of anthropologists, economists, and sociologists writing in this field. Current research trends are increasingly focused on think-tank development in the developing world (McGann and Weaver 2000) and post-communist transition countries (Krastev 2000). Think-tank transnationalization is likely to be a field of future research given the impact of globalization and regionalization on institute research agendas and activities.

Analysis of think tanks has diverged into two branches. Those focused on the organizational form such as Weaver (1989), with McGann and Weaver (2000) and James Smith (1991), were interested in explaining why and how think tanks have emerged and why some think tanks are influential. They distinguished independent public-policy institutes from academic research centers, government research units, or lobbyists. They focused on the organizational ingredients of what makes a think tank successful, how think tanks are managed, who funds them, who quotes them, and whom they try to influence. Writers in the second branch analyzed think tanks as a vehicle for broader questions about the policy process and the role of ideas and expertise in decision making. Writers such as Diane Stone (1996), Ivan Krastev (2000), and Martin Thunert (2000) employed network concepts such as 'policy community,' 'advocacy coalitions,' and 'epistemic communities' to explain why ideas matter. It cuts to a central issue in the study of think tanks—policy influence.

Notwithstanding extensive growth, think tanks do not enjoy automatic political access or unhindered bureaucratic routes to decision makers. Think tanks need to have some kind of engagement with government if they are to have policy relevance; however, their desire to preserve intellectual autonomy means that most think tanks try to strike a delicate balance between dependence on government and isolation from it. Conducting policy-relevant research and attempting to broker it to decision makers does not equate with policy influence. This conundrum has been the major methodological problem in the study of think tanks.

Proof of influence is elusive. There is no clear causal nexus between policy research or an idea espoused by an institute, and political decisions and policy change. It is a rare occurrence that a policy outcome can be attributed to one organization, as there are other intermediary forces; political parties, interests, and the media also shape the content and reception of ideas. Furthermore, the cognitive order of politicians and bureaucrats cannot be accessed by 'think-tanks' researchers, and interviewing has its own problems of faded memories and selective recall in establishing the veracity of perceptions of think-tank policy input. This hiatus is compounded in that the political stature of many institutes waxes and wanes over time, and that the criteria by which influence is assessed vary not only across think tanks but within the social sciences in general.

Nevertheless, there are many countries where overstretched bureaucracies and limited capacities for in-

house governmental policy analysis provide the opportunity for think tanks to serve government needs; however, the degree of incorporation and cooption varies from think tank to think tank, and from one country to another. Latin American institutes, particularly the liberal institutes, were often marginal to the political system until the demise of authoritarian regimes. By contrast, a group of elite think tanks in Southeast Asia have enjoyed a much closer relationship with their governments where the trade-off was accommodation with government concerns and controls. A large number of neoliberal or free-market research institutes eschew government funding. The organizational choices made by think tanks about their relationship to centers of power and authority must be understood by reference not only to the ideological disposition or mission of the think tank, but also by acknowledging environmental conditions such as legal and economic constraints of the political culture in which they are located.

Outside the political sphere and state sector, think tanks have cultivated other audiences. Students and academics in colleges and universities regularly use think-tank publications. Foundation officials, business executives, bureaucrats from various international organizations, university researchers, journalists, and, for want of a better term, the 'educated public' are often engaged by think-tank pursuits. Similarly, third-sector intellectuals such as are found among trade unionists, religious spokespeople, NGO leaders, and social-movement activists can be attracted by the intellectual and ideological sustenance to be found in the think-tank community. Furthermore, think tanks provide an organizational link and communication bridge between their different audiences. They connect disparate groups by providing a forum for the exchange of views, by translating academic or scientific research into policy-relevant publications, and by spreading policy lessons internationally.

Bibliography

Day A J 1993 *Think Tanks: An International Directory*. Longman, Harlow, UK

Hellebust L 1996 *Think Tank Directory: A Guide to Nonprofit Public Policy Research Organizations*. Government Research Service, Topeka, KS

Krastev I 2000 Post-communist think tanks: making and faking influence. In: Stone D (ed.) *Banking on Knowledge: The Genesis of the Global Development Network*. Routledge, New York

McGann J, Weaver R K (eds.) 2000 *Think Tanks and Civil Societies: Catalysts for Ideas and Action*. Transaction Press, Sommerset, NJ

Smith J A 1991 *The Idea Brokers: Think Tanks and the Rise of the New Policy Elite*. Free Press, New York

Stone D 1996 *Capturing the Political Imagination: Think Tanks and the Policy Process*. Cass, London

Thunert M 2000 Players beyond borders? German think tanks as catalysts of internationalization. *Global Society* **14**: 191–211

Weaver R K 1989 The changing world of think tanks. *PS: Political Science and Politics* 563–78

D. Stone

Third Parties: United States

A third party is any political party in the US political system other than one of the two major parties. The term is applied generically to all such minor parties, regardless of reputation or record. There are many third parties in the USA. In 1996, for example, presidential candidates for the Reform and Libertarian parties appeared on ballots in all 50 states and in the District of Columbia, the Natural Law Party's candidate made it onto 45 ballots, the candidate for the US Taxpayers Party was on 41 ballots, and the Green Party's candidate appeared on 24 ballots. Candidates from 11 other parties were listed on the ballot in at least one state. Moreover, 1996 was not an unusual year. In almost every election many candidates appear on the ballot (Rosenstone et al. 1984, pp. 231–5).

Although there are many political parties in the USA, the US political system is correctly classified as a two-party system. The same two political parties, the Democrats and the Republicans, have held dominant positions since 1860. Third-party candidates win remarkably few elections. In 2000, none of the 100 US senators was elected from a third party, and only one of the 435 members of the US House of Representatives was a member of a third party—Bernard Sanders of Vermont, a socialist elected as an independent. Only two of the 50 state governors were elected as something other than a Republican or a Democrat (Angus King, an independent in Maine, and Jesse Ventura, from the Reform Party in Minnesota), and of the 7,524 state legislators, only 20 were elected from third parties.

The USA has a two-party system for several reasons. The most compelling is the single-member, plurality electoral system used in the vast majority of elections. Under these rules the US House of Representatives, for example, is divided into 435 districts. Each district elects only one representative. The winner in each district is the candidate who receives a plurality of the votes. Third-party candidates almost never succeed under such rules, and a stable two-party system emerges. The generalization that a country with a single-member, plurality electoral system develops a two-party system is called Duverger's Law after the person who first noted it (Duverger 1959).

The two-party system's stability has been reinforced over the years by laws passed by institutions controlled by the major parties that limit the ability of third parties to compete successfully for votes. Ballot access

rules, for example, virtually assure that candidates from the two major parties will appear on them, but often make it very difficult for candidates from other parties to be listed. Indeed, at times during the twentieth century some states employed laws that, in effect, guaranteed the two major parties ballot listings in perpetuity. A 1949 law in Florida, for example, stated that any party that had elected a president since 1900 would automatically qualify for the ballot, something that only the Democrats and Republicans had accomplished. That law was repealed in 1977. Most states now grant ballot access to parties based on their candidate surpassing some threshold percentage of votes in a specified earlier election, typically a gubernatorial election. The other route to gaining ballot access is through a petition drive. Such petitions require the signatures of some specific number of citizens or, in some states, registered voters. The number of required signatures varies by state. The higher the number of signatures required, the greater the obstacle to getting a third-party candidate on the ballot. As Lewis-Beck and Squire (1995) have shown, the difficulty in getting a third-party candidate on the ballot increases with the political stakes in a state. Thus, the states with the greatest number of electoral college votes are those with the most stringent ballot access requirements.

Most American states ban fusion candidacies, another restriction that hampers the ability of third parties to become competitive. Fusion is the process allowing a candidate to appear on the ballot as the nominee of more than one party. In the state of New York, for example, a candidate can be the nominee of more than one party. Thus, a candidate nominated by the Republican Party also can be the nominee of the Conservative Party. Fusion enhances the prospects for third parties by allowing them to associate themselves with popular candidates, and perhaps increase their vote totals over what they would otherwise have gotten. Increased vote totals can matter because they are often the standard used to determine ballot access in the next election. The United States Supreme Court in a 1997 decision, *Timmons vs. Twin Cities Area New Party*, upheld the constitutionality of the state of Minnesota's law prohibiting party fusion.

Finally, campaign finance laws work to the benefit of the two major parties and to the detriment of third parties. In presidential campaigns, for example, the law provides direct financing for the general election campaigns of the Democratic and Republican candidates. Third-party candidates, however, are often in the position of not qualifying for any public financing, or only qualifying after the fact—based on the percentage of votes they receive in the general election. Without the same access to campaign funds as the major parties, third-party candidates find it difficult to make their case to the voters.

There also are informal norms in US politics that work against third parties. Third-party candidates are rarely accorded the same level of news coverage as major-party candidates. Similarly, third-party candidates are often excluded from candidate forums and debates sponsored by civic organizations and the media. Exclusion is usually justified on the basis of the lack of public support for or interest in third parties and their candidates. Such exclusion, however, prevents third parties and their candidates from publicizing their policy positions, keeping them from possibly expanding their meager bases of support.

The first third party was the Anti-Masonic Party, formed in 1827 in response to concerns about the supposed influence of Freemasons on US government. The Anti-Masonic Party lasted only a few years before merging with one of the then major parties, the Whigs, in 1838. During its brief life, the Anti-Masonic Party made one enduring contribution to US politics: they were the first party to hold a national convention, an innovation immediately adopted by the major parties and which they continue to use today.

Third parties became more prevalent on the national political scene during the second half of the nineteenth century. Notable among them were the Greenback Party, the Populist Party, and the Prohibition Party. The Greenback Party existed from 1874 to 1888. It was formed to promote the continued issuance of paper currency (popularly referred to as 'greenbacks') as a means to increase inflation and thereby lighten the debt burden carried by midwestern farmers. Although the party managed to win some seats in the US House of Representatives, support for it declined after a different inflationary scheme, one calling for the free issuance of silver coins, came into vogue.

The push for 'free silver' resulted in the establishment of another third party in the early 1890s. Formally called the People's Party, but more commonly referred to as the Populist Party, this group worked to represent the interests of farmers and the emerging industrial working class. Among the policy positions they espoused in their 1892 platform were free silver, government ownership of the railroads, a graduated income tax, the direct election of US senators, and an eight-hour day for industrial workers. The party enjoyed some success in the 1892 election—their presidential candidate, James Weaver of Iowa, won 22 electoral votes—but that proved to be its high-water mark. By 1896 the Democrats had adopted many of the Populists' positions, and the party ended up supporting the Democrat's presidential candidate, William Jennings Bryan. The Populists ran their own presidential candidates in 1904 and 1908, but ceased to function as a political party after that point in time.

Although many third parties existed during the twentieth century, they were, in most instances, qualitatively different from those in the nineteenth century. As Rosenstone et al. (1984) observed, third parties in the nineteenth century were organized around specific issues and ideologies. The two incarnations of the Prohibition Party, for example, were formed to push a

specific issue, the prohibition of production and consumption of alcohol. In contrast, most twentieth century third parties were centered on particular personalities rather than policies. The three Progressive Parties demonstrate this point. The first Progressive Party was created in 1912 to support Theodore Roosevelt's attempt to recapture the presidency. Also called the Bull Moose Party, this version of the Progressives gave its 1916 nomination to the Republican Party's candidate, and then, in effect, disappeared from the political scene. The Progressive Party re-emerged in 1924, uniting with another third party, the Socialist Party, in support of Robert La Follette's campaign for the White House. The last significant appearance of the Progressive Party label was in 1948, when former Vice-President Henry A. Wallace ran as the party's candidate for president. Although there were some common policy themes across the three major Progressive Parties, there were also significant differences, and there was virtually no organizational continuity. All three efforts were driven by well-known political personalities and after each of them failed to win office, the party, for all intents and purposes, ceased to exist.

The major third-party efforts in the second half of the twentieth century were all personality driven. George Wallace's impressive performance in the 1968 election led to the formation of what is now called the American Independent Party. But without the former Alabama governor on the ballot the party has never again been taken seriously. Of all the third-party efforts in the twentieth century, perhaps only the Reform Party, created primarily through the efforts of 1992 and 1996 presidential candidate Ross Perot, has any chance to become established as a significant political force independent of the personality around which it was first built.

What role do third parties play in US politics? There is little evidence that significant third-party challenges bring more Americans to the polls. In 1948, for example, voters were offered two serious third-party candidates (Governor Strom Thurmond, of the States' Rights Democrat (Dixiecrat) Party, and the Progressive Party's Wallace). Voter turnout that year was almost five percentage points lower than in 1944 when no significant third-party efforts were mounted. Similar patterns held in 1968 and 1980 when important third-party candidates ran, but voter turnout fell compared to the previous election. Only Ross Perot's 1992 campaign was associated with an increase in turnout, an almost five-point gain over 1988. But his presence in the 1996 campaign failed again to increase turnout. In fact, participation declined more than five points.

There are, however, two important contributions third parties make to the US political system. First, they offer voters a way to express dissatisfaction with the two major parties and their candidates. Second, third parties are vehicles for promoting issues and policies that the major parties may ignore. If the issues raised by third parties resonate with the public, the major parties are apt to embrace them. Many of the policies espoused by the Populist Party, for example, proved popular and were commandeered by the major parties. Thus, although the Populist Party disappeared from the political landscape, a number of the ideas it brought to the political debate were later enshrined in law.

See also: Democratic Party; Electoral Systems; Party Systems; Political Parties; Political Parties, History of; Political Representation; Republican Party

Bibliography

Duverger M 1959 *Political Parties*. Wiley, New York

Lewis-Beck M S, Squire P 1995 The politics of institutional choice: Presidential ballot access for third parties in the United States. *British Journal of Political Science* **25**: 419–27

Rosenstone S J, Behr R L, Lazarus E H 1984 *Third Parties in America: Citizen Response to Major Party Failure*. Princeton University Press, Princeton, NJ

Timmons, vs. Twin Cities Area New Party 1997 520 U.S. 351

P. Squire

Third World

1. Defining the Third World

The Third World has become a convenient catch phrase that refers collectively to the large number of heterogeneous societies, cultures, and civilizations of the non-Western world most of which lie in the tropics, have experienced colonial rule, whose economies are less industrialized or not completely modernized and a majority of whose population has the lowest standards of living or are considered or consider themselves poor relative to those of the First and Second Worlds. The Third World includes all countries of Africa (except South Africa), Asia (except Japan), and Latin America and the Caribbean, and some states and territories of Oceania.

2. The Origin of the Term, and its Varied Meanings

The term 'Third World' (*le Tiers Monde*) is of French origin. It was first used in France in the early 1950s by a number of French writers like Alfred Sauvy (politician, economist, who is credited with coining the term in 1952) and Claude Bourdel, in the context of apparently increasing polarization of the international system into rival blocs. The term conveys a con-

scious echo of the older concept of '*le tiers etat*' that is, 'the third estate' (the common people, the poor and underprivileged) of the prerevolutionary *ancien régime*. Some claim that it was Frantz Fanon, Martiniquan writer and psychiatrist (author of *The Wretched of the Earth* 1963), who first used the term to refer to the poor decolonized societies of Asia, Africa, and Latin America.

The term has been used in several related senses, substituting for or succeeding other terms used by Europeans to refer to non-European peoples and cultures: 'aboriginal,' 'indigenous,' 'savage,' 'barbarous,' 'traditional,' 'native,' 'primitive,' 'tribal,' 'oriental,' 'Arab,' 'Negro,' 'Indian,' 'Mohammedan,' 'pagan,' 'precapitalist,' 'pre-industrial,' 'non-Western,' 'non-White,' 'premodern.' Such terms have been used sometimes with deliberate contempt and condescension, with pejorative connotations, or even racist overtones. These terms have also been adopted by the educated, more privileged urban-dwellers in the Third World itself.

'Third World' has also been employed as a shorthand euphemism for the social condition or plight of the very poor or powerless or underdeveloped society anywhere in the world, including the First and Second Worlds, even though a vast majority of the people fitting the description live in the Third World. In this sense, the phrase 'America's Third World' refers to America's rural and urban poor: the black underclass, native Americans, and rural white Appalachians. In a similar usage, the former Soviet Union was once described by the *Times* of London as having a 'Third World' economy with First World weapons. South Africa is described as a First World society in which the Third World is ever present.

The term 'Third World' has also been extended to the poorer Eastern European countries whose economies were subjected to the needs of and exploited by Northern Europe in the early modern period. In everyday use, the term has thus acquired loose, but frequently negative, social, cultural, political, economic, ideological, and 'racial' meanings, often with a fine disregard for geography. From 1945 to 1989 a capitalist block (First World) led by the United States and a communist one (Second World) led by the Soviet Union fought a cold war which ended with the collapse of the Soviet Union in 1991. In the Cold War context the Third World consisted mostly of newly decolonized, uncommitted countries not aligned diplomatically with either the capitalist world or the communist world. But the two dominant historic models of development, capitalism and socialism, especially the former, continued to broadly shape the rate and direction of change in the Third World.

3. *Third World as an Analytical Concept*

Since World War II and the invention of the term, the conditions of Third World, and the relationship among the First, Second, and Third Worlds, have been the focus of theory, intensive research, and extensive documentation, aimed at understanding the origin and nature of poverty in the Third World, and of policy concerned with strategies for development and poverty alleviation. The leading theories of Third World development include 'modernization theory,' 'dependency theory,' and 'world-system theory.' Some Marxist scholars have rejected the term 'Third World' as a false concept, masking the contradictions of capitalist development in the former colonies. Third World countries often see themselves as old societies 'lagging behind' economically and even politically and thus trying to 'catch up' with the more advanced industrialized countries of the First and Second Worlds.

The rise and spread of capitalism globally were intimately linked to the growth of large powerful oppressive and exploitative European colonial empires, inspired by the prevailing mercantilist philosophy of the early modern period dedicated to the unbridled pursuit of economic power and wealth of states.

Because of its vast natural resources, which were largely untapped, its considerable potential commercial value, and the importance of some areas to White settlement and occupation, the Third World has been for over five centuries a principal focus of European territorial ambition and rivalry. Most of the states and territories of the Third World took their present form in the nineteenth and early twentieth century when large parts of Asia and Africa were colonized by European imperial powers. By the 1880s the world had been carved up among a handful of European industrial powers. By 1914 European economic and political control covered most of the global land surface, claiming nearly the entire non-European world. Japan successfully resisted European colonization, and China was never formally colonized but for the highly lucrative treaty ports along the coasts, e.g. Hong Kong following the Opium Wars of the 1830s. Latin America, conquered, colonized, and settled by the Portuguese and Spanish crowns three centuries before Asian and African colonization, was transformed into independent republics, along with formerly French Haiti in the Caribbean, in early nineteenth century, after long, bloody and bitter nationalist struggles. Likewise, anticolonial protests and revolts, often violent and bloody, led to the decolonization of Asian, African, Caribbean, and Pacific countries in the late 1940s and after. But in both cases political freedom did not end foreign—mostly European—economic control and ownership, or bring about rapid industrialization and better living standards for most of the new states.

The European metropolitan powers made colonies a principal source of raw materials destined for European industries and consumer markets while suppressing the development of manufacturing industries in the colonies to avoid competition as they

provided markets for European manufactured goods. Arguably the affluence of industrial Europe is literally the creation of the Third World.

Colonialism which followed the transatlantic slave trade in Africa, during which an estimated 10 million or more Africans were exported to provide unpaid labor for European plantations and mines of the Americas, did incalculable damage to black Africa, notwithstanding some lasting beneficial by-products: Christianity, schools, hospitals, and the start of modern administration. The colonial borders which were drawn across Africa divided African peoples and cultures, and along with the authoritarian and repressive colonial political tradition are largely responsible for much of the chronic poverty, cultural identity crisis, and political instability of independent African countries. Despite wide differences in local cultures, historical traditions, and responses to colonialism, colonies came to share several common features: political control from Europe; unchecked economic exploitation; and particularly in the black world, the construction and reproduction of a racist social order of which apartheid of South Africa was the most notorious recent example.

One enduring legacy of European colonial empires is a global economic order with an asymmetrical North–South division between a rich industrial North of mostly former imperial powers and poor, less industrialized South of former colonies. The North–South divide is exacerbated by neocolonialist attempts to influence or exercise *de facto* control over the policies of the South through formal and informal alliances and agreements calculated to ensure the perpetuation of the hegemonic economic power of the North.

4. *The UN and the Rise of the Third World*

The United Nations charter endorsed the principle of self-determination and sovereign equality of all nations. After the late 1940s, the large ex-colonies like India, Pakistan, Egypt, and Indonesia, supported by Latin American countries, used the UN as a forum to press for an end to colonialism and rapid development of former colonies. The term 'Third World' received formal endorsement following the Bandung conference of Asian and African states in Indonesia in April 1955, which affirmed the solidarity of the Third World, and prepared for the birth of the nonaligned movement (NAM). At the conference, China declared herself a 'Third World' country. The term was also adopted in the UN classification system to distinguish between the developed and underdeveloped countries.

The conferences of the NAM, the first one held in Belgrade (former Yugoslavia) in 1961, and the founding in June 1964 of the UN Group of 77, at the conclusion of the first UN conference on Trade and Development (UNCTAD), which was convened at the initiative of Third World countries, marked the growth of the Third World as a potent force in world affairs.

With approximately two-thirds of the total UN membership, the 1960s and 1970s witnessed the rapid rise in the power and sense of purpose of the Third World. The G-77 spearheaded the UN call for a New International Economic Order (NIEO) to reform the economic relations between the North and the South and to rectify the inequality evident in the relative access of the rich and poor countries to markets, to supplies, to science and technology, and to credit, which implied a redistribution of wealth in favor of the South. The issue of North–South relations was made the subject of an international commission which met from 1977 to 1979 under the chairmanship of Willy Brandt.

By the 1990s, with the collapse of the Soviet Union and the end of bipolar politics, NAM and the G-77 which have a joint coordinating committee shifted their focus to problems of underdevelopment, the Third World debt burden, trade liberalization, and aid.

The G-77, whose membership increased to 130 with the admission of post-apartheid South Africa, believes that global shared prosperity through North–South cooperation on a fair and equitable basis is the only long-term remedy for Third World poverty and human misery.

5. *Diversity, Division, and Integration*

Depending on the definition and context, different countries or populations will make up the Third World. Accordingly, the 'Third World' may be further divided into those states that have access to rich natural resources, faster growth rates or better development potential and those which do not. While the latter are sometimes described as the 'Fourth World,' that term is probably more frequently used to refer to marginalized indigenous or tribal peoples. Despite internal divisions and differences, Third World countries share some common features as well as problems.

Most of the peoples of the Third World still suffer from hunger, disease, ignorance, lack of shelter, potable water, and opportunities (especially for women) to realize their potential as human beings. Third World countries are faced with crippling external debt burdens, poor social and physical infrastructure, and inadequate technology, making it difficult or impossible to recover from or respond effectively (without massive external help mostly from the First World) to natural disasters and civil conflicts to which many are prone. Because of the chronic and deepening poverty in the midst of burgeoning populations, diminishing natural resources, and increas-

ingly degraded environments, these countries have also become major producers of massive illegal political and economic migrations to the more prosperous industrial capitalist North, and of economic and political refugees and internally-displaced persons within the Third World itself.

6. *The Third World and the Post-Cold War International System*

The 'Third World' as an ideological concept has been undermined by the end of the Cold War, the collapse of Soviet-style communism, and the increasing cooperation between the leading industrial capitalist nations—the G-77—and Russia. Its economic meaning has also been challenged by the fundamental changes in the world economy, brought about by the forces of globalization. The transition of the former Soviet Union and Eastern bloc countries and some Third World nations to pluralist democracy and market economies has made it increasingly difficult to distinguish between some First and Second World countries, and in some cases between Third World and Second World countries.

Given the wide range in the levels of economic development and standards of living within and between Third World countries, placing prosperous Hong Kong and impoverished Haiti, or industrializing Brazil and hunger prone Bangladesh, in the same category is clearly questionable.

Again, by the end of the twentieth century, new economic groupings made up of selected countries of the First, Second, and Third Worlds, some of them regional (e.g., NAFTA), were formed to promote regional prosperity; and others, cross-regional (e.g., Group of 20), emerged to foster dialogue if not unity of action between the rich industrial countries and countries of the emerging markets. The growing convergence of national and transnational economic and political interests has added a new challenge to the usefulness of the Third World as an economic or ideological category.

In the new strategic global environment, the NAM also seems increasingly anachronistic. But the overall objectives of the NAM which include the pursuit of world peace, the elimination of disease, hunger, and ignorance; the conduct of international relations on the basis of respect for and understanding of other cultures, systems, and values; and the elimination of all forms of neo-colonialism and racism, make it still relevant.

Despite the new redefinition/regrouping of economies, and the cross-cutting interests reflecting the progress of globalization, the industrial countries, mostly European—the First and Second Worlds—still have between them most of the world's economic power and military might. Of the estimated global population of six billion, only about 20 percent live in the rich North; the overwhelming majority (80 percent) live in the impoverished South. Yet 86 percent of global output of goods and services is produced and consumed by the developed countries. Key measures of human well-being and survival such as gross-national product per capita (GNP), per capita energy consumption, life expectancy, infant mortality rates, and literacy levels, attest to the unacceptably wide and growing gap between the world's richest and poorest (Third World) countries.

See also: Capitalism; Colonialism, Anthropology of; Colonialism: Political Aspects; Colonization and Colonialism, History of; Dependency Theory; Fourth World; Globalization, Anthropology of; Globalization: Political Aspects; Imperialism, History of; Imperialism: Political Aspects; Modernization and Modernity in History; Postcoloniality; World Systems Theory

Bibliography

Amin S 1976 *Unequal Development*. Monthly Review Press, New York

Brandt W 1980 *North–South: A Program For Survival. The Report of the Independent Commission on International Development Issues Under the Chairmanship of Willy Brandt*. MIT Press, Cambridge, MA

Fanon F 1963 *The Wretched of the Earth*. Grove Press, New York

Forson A 1998 The non-aligned movement and the Third World. *Daily Graphic (Accra)* **August 26**: 7

McCormick J 1988 America's Third World. *Newsweek* **August 8**: 21–4

Owusu M (ed.) 1979 *Colonialism and Change*. Mouton Publishers, The Hague, The Netherlands

Rodney W 1972 *How Europe Underdeveloped Africa*. Bogle-L'Ouverture Publications, London

Sauvy A 1952 *General Theory of Population* [trans. Campos C]. Methuen, London

Singham A W (ed.) 1977 *The Non-aligned Movement in World Politics*. Lawrence Hill, Westport, CT

Stavrianos L S 1981 *Global Rift. The Third World Comes of Age*. William Morrow, New York

The Times 1983 April 22

Wolf E R 1982 *Europe and the People without History*. University of California Press, Berkeley, CA

Worsley P 1984 *The Three Worlds. Culture and World Development*. University of Chicago Press, Chicago

Yapp ME 1985 What is Third World history? *History Today* **35**: 37

M. Owusu

Thirst and Drinking

Drinking is one of the triad of motivated ingestive behaviors used by animals to ingest those substances essential for maintaining metabolism, growth, and

overall bodily integrity. Feeding controls the intake of nutrients (fat, protein, carbohydrates), vitamins, and cofactors; targeted appetites control the intake of specific ions (principally sodium, but also to a lesser extent other essential ions such as calcium) from a variety of sources; while drinking regulates water intake. These behaviors are controlled by complex neural systems that have components distributed throughout the brain, and transform the need to replenish or anticipate deficits, the desire to find suitable goal objects, and knowledge of their whereabouts into appropriate sets of behavioral and visceral motor output (see also *Motivation, Neural Basis of*).

1. Composition and Physiological Regulation of the Body's Fluid Compartments

Understanding how body fluids are compartmentalized and which of their components are maintained within regulated limits is central to understanding how drinking behaviors are initiated.

The body consists of two fluid compartments: a cellular compartment comprising the internal fluid of all cells, and an extracellular compartment comprised of all interstitial fluid and plasma in the blood. In terrestrial mammals the water in the cellular compartment makes up about 65–70 percent of the body's total water. Fluid in the cellular compartment is separated from extracellular fluid (ECF) by the semipermeable membranes that surrounds all cells. This membrane is completely permeable to water, but the transmembrane movement of many other molecules and ions is actively regulated, making the physiology of this membrane critical for water balance.

One of the key features of the cell membrane is its ability of the enzyme Na^+/K^+ ATPase to promote the active transport of sodium ions out of the cellular compartment. This enzyme is a two-subunit transmembrane protein, is a major utilizer of cellular energy, and exchanges 3 Na^+ out of the cell for 2 K^+ in. This process maintains the ECF Na^+ concentration (150 mM) at about tenfold greater than that inside the cell. This concentration gradient is critical for fluid balance and the control of drinking behavior because it determines the rate of water flux across the membrane.

Two variables of the ECF compartment are regulated tightly by physiological and behavioral process: ECF osmolality and volume. Under conditions of isotonicity the net flow of water across the cell membrane is zero. However, any disturbance of this osmotic balance leads to increased transmembrane water flow. The sodium ion is the most important osmolyte in this regard. In terms of thirst and drinking, increasing the osmolality of the ECF compartment is an important parameter that results in cellular dehydration, which is a net loss of water from the cell. In turn, the kidney attempts to excrete the excess sodium from the blood, but this cannot be achieved without incurring an accompanying loss of water from the body.

ECF volume is important for long-term maintenance of blood pressure and efficient cardiovascular function. Unlike cellular dehydration, which involves a net loss of water but not sodium, loss of fluid from the extracellular compartment (hypovolemia) involves loss of water and all of its other components, including sodium. Therefore, in addition to triggering an appetite for water (i.e., thirst), hypovolemia also initiates the specific appetite for sodium (Grossman 1990).

2. Types of Drinking Behavior

Drinking behaviors can be divided usefully into two categories: those that occur in response to a significant disturbance in ECF composition and result in a classic homeostatic response; and those that occur in the absence of such a disturbance, but instead are initiated to prevent the occurrence of such deficits. Each of these types of drinking is associated with distinct neural mechanisms (Grossman 1990, Rolls and Rolls 1982).

2.1 Homeostatic Drinking

The most well-studied aspects of drinking behavior are those initiated in response to significant changes in ECF composition. A large body of work has identified specific types of dipsogenic (i.e., thirst-inducing) sensory signals that are associated with particular disturbances to the composition of the ECF. In humans, the most common form of homeostatic drinking occurs in response to exercise or heat-stress. Pathological causes of homeostatic drinking include severe hemorrhage or conditions associated with severe diarrhea.

2.2 Anticipatory Drinking

Drinking that occurs without a major disturbance to ECF composition requires regulation by regions of the brain concerned with the anticipatory components of motivated behaviors. Two groups of brain regions are particularly concerned with regulating this type of drinking; those involved with controlling arousal state and circadian timing (particularly the circadian clock in the suprachiasmatic nucleus), and those cortical regions concerned with the complex processing of sensory objects and their representation in processes such as reward/aversion, and learning and memory

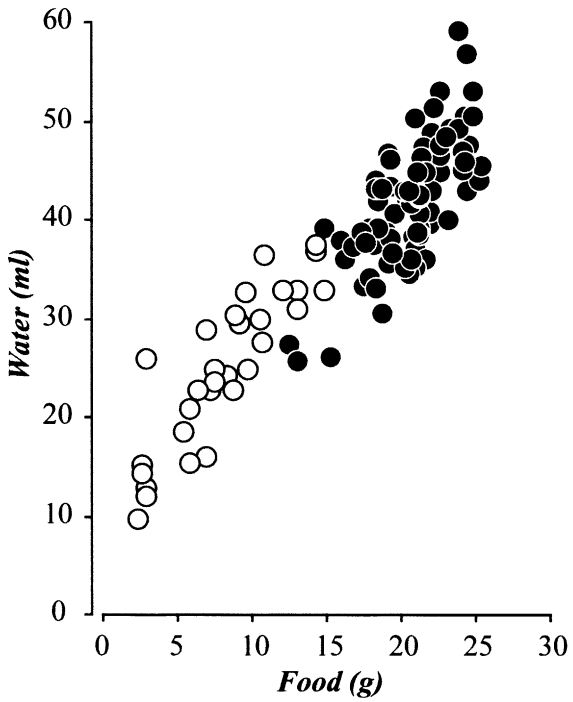

Figure 1
The relationship between nocturnal water and food intakes in individual laboratory rats allowed either unrestricted (closed circles) or restricted access to food (open circles). Note the strong correlation that is maintained during limited access to food

(Watts 2001). In these circumstances, drinking is often closely associated with feeding behavior, both temporally and in quantity (Fig. 1). Food-associated water intake (either as a component of the food itself or as a separate drinking episode) represents the most common route of water entry into the body for humans living in economically developed countries (see also *Hunger and Eating, Neural Basis of*).

3. *Sensory Signals and their Transduction*

3.1 *Extracellular Fluid Volume*

Physiological mechanisms concerned with maintaining ECF volume are evolutionarily older than those that maintain ECF osmolality. This is because the challenge from water deprivation and increasing ECF osmolality did not pose a physiological threat until the evolution of terrestrial life. Indeed in primitive vertebrates, the physiologically more important actions of two of the hormones most closely associated with the response to dehydration—angiotensin II (ATII) and the vasopressin-related peptides—is vasoconstriction driven by AT1 and V1 receptors, respectively, which leads to a reduction in renal blood flow, and consequently reduced fluid loss from the kidneys. The antidiuretic actions of vasopressin are evolutionarily more recent and are mediated by specific water channel proteins called aquaporins. An elevated insertion rate of these proteins into the basal membrane of convoluted tubule is mediated by vasopressin's interactions with V2 receptors. Although vasopressin has significant effects on water conservation, it has no dipsogenic actions. Decreases in ECF volume are transduced into appropriate motor actions by two sensory mechanisms.

3.1.1 Angiotensin II. Reduced blood flow through the specialized juxtaglomerular apparatus in the kidney stimulates the release of an renin, an enzyme that catalyzes the conversion of angiotensinogen to angiotensin I. Once in the general circulation angiotensin I is converted to ATII by angiotensin-converting enzyme in the lungs. ATII has four major actions associated with thirst and fluid balance. First, it acts on vascular AT1 receptors to constrict blood vessels and maintain blood pressure; second it is dipsogenic; third, it stimulates sodium appetite; and finally, it acts on zona glomerulosa cells in the adrenal cortex to stimulate aldosterone release, which then targets the convoluted tubules in the kidney to potentiate sodium retention. To stimulate drinking, circulating ATII binds to AT1 receptors in the subfornical organ (SFO). This structure is the most sensitive ATII responsive site in the brain and many studies place this structure centrally in a circuit that directly and specifically stimulates water intake (Fitzsimons 1998). The SFO provides neural efferent projections—most of which also contain ATII (Swanson 1987)—to a relatively limited set of brain structures, including parts of the prefrontal cortex, substantia innominata, medial preoptic area, bed nuclei of the stria terminalis (BST), zona incerta, paraventricular hypothalamic nucleus (PVH), supraoptic nucleus, and lateral hypothalamic area (Swanson 1987).

3.1.2 Baroreceptors. Inputs from baroreceptors in the walls of blood vessels on low-pressure (venous) side of the circulation (primarily the superior vena cava) provide important neural information that contributes to the thirst generated by hypovolemia. This is demonstrated by the fact that the thirst generated by hypovolemia is inhibited by a small inflatable balloon placed to maintain the output of the baroreceptors in the wall of the superior vena cava as the animal becomes hypovolemic (Kaufman 1984). The afferent information from these receptors is initially

processed by the nucleus of the solitary tract (NTS) in the hindbrain, before being relayed to those regions of the brain responsible for organizing drinking behavior.

3.2 Extracellular Osmolality

All cells are *osmosensitive* in that they lose water and shrink in proportion to increases in the osmolality of the ECF. However, these increases are only transduced into neural signals by select groups of *osmoreceptive* cells located in the brain and periphery. In terms of thirst and drinking behavior, the most important set is located in the vascular organ of the lamina terminalis (OVLT), a hypothalamic structure located at the rostral end of the third ventricle. Supplementary sets of osmoreceptors may also contribute to the motor responses to dehydration, and are located at other sites in the brain (particularly in the SFO and area postrema) as well as in the periphery (the liver and possibly hepatic portal vein).

A cellular mechanism that accounts for the osmotransduction exhibited by some neurons has recently been described by Bourque and co-workers (Bourque and Oliet 1997). In this scheme, osmoreceptive magnocellular neurons in the supraoptic nucleus possess stretch-inactivated cationic channels that allow increased entry of positive charge as they shrink in response to increasing ECF osmolality (Fig. 2). In turn, these neurons depolarize and generate action potentials in proportion to the change in ECF osmolality (Fig. 2). Interestingly, unlike many other types of sensory receptors, the osmoreceptors associated with drinking apparently do not adapt when exposed to a slowly changing stimulus intensity (Fitzsimons 1963).

3.3 Sensory Processing

The information derived from these sensory receptors that encodes alterations in ECF composition requires initial processing before being integrated into the motor control networks that select the most appropriate motor action. As with many other motivated behaviors, the exact mechanisms involved with this function are not clear. However, a recent insight was recently provided into those regions important for generating human drinking behaviors in response to cellular dehydration by a positron emission tomographic study (Denton et al. 1999a, 1999b). Volunteers were infused intravenously with hypertonic saline and asked to gauge the subsequent intensity of thirst. In turn, those regions in the brain showing increased blood flow were correlated with the hyperosmolemia and the reported intensity of thirst. These studies identified parts of the cortex as well as the cerebellum as being specifically activated by hyperosmolemia.

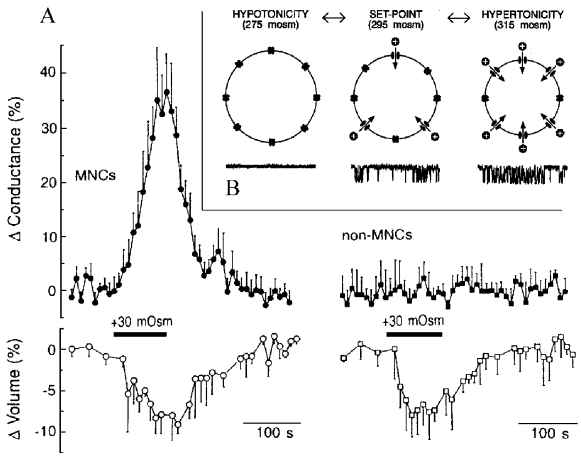

Figure 2
(A) Top traces show average changes (S.E.M.) of input conductance (filled symbols) recorded from magnocellular neurons (MNC; circles, n = 7) and non-MNCs (squares, n = 8) upon exposure to a hypertonic stimulus. Bottom traces plot the corresponding changes in cell volume (percent of control; open symbols) estimated from 6 MNCs and 5 non-MNCs exposed to an identical stimulus (from Oliet and Bourque 1993; with permission from Annual Reviews Inc.).
(B) Stretch inactivated (SI) cationic channels transduce osmoregulation. Under resting conditions (middle panel) a portion of the SI cationic channels is active and allows the influx of positive charge. Hypotonic stimulation (left) provokes cell swelling and inhibits channel activity, thereby hyperpolarizing the cell. In contrast, hypertonic stimulation (right) causes cell shrinking. Activation of an increased number of channels under this condition augments charge influx and results in membrane depolarization. Traces representing changes in the activity of a single SI channel are shown below (from Bourque and Oliet 1997; with permission from Nature Publishing Group)

4. The Components of Drinking Behavior

4.1 Motor Actions

As animals become dehydrated the sensory mechanisms described above are activated. However, which of these is specifically engaged depends on the nature of the dehydrating stimulus. In the case of water deprivation, which is likely to be the most commonly encountered stimulus by most animals, approximately 70 percent of the drinking behavior can be attributed to the sensory mechanisms associated with cellular dehydration, while the remainder is mediated by the mechanisms triggered by hypovolemia (Rolls and Rolls 1982). On the other hand, the thirst and drinking that follows loss of body fluids resulting from heat stress, for example, leads to isotonic hypovolemia and

consequently ATII and neural baroreceptor sensory mechanisms play critical roles in the subsequent thirst and drinking behavior.

Despite the variety of these sensory mechanisms, there is a strong degree of convergence from the brain regions responsible for organizing drinking behavior onto a well-characterized set of neuroendocrine, autonomic, and behavioral motor mechanisms. These responses should not be considered in isolation and can be usefully categorized as protective, adaptive, or restorative to the body's fluid compartments (Grossman 1990, Watts 2001).

The principal protective responses involve the neuroendocrine and autonomic motor systems (see also *Autonomic Nervous System*; *Neuroendocrinology*). They include the very rapidly initiated release of vasopressin and oxytocin from the posterior pituitary that are activated by neuroendocrine reflexes following dehydration. These hormones help maintain blood pressure, and increase renal water retention and natriuresis. Other protective responses occur after hemorrhage or hypovolemia, and include modifications to autonomic control of cardiovascular function, and the rapid activation of all levels of the hypothalamo-pituitary-adrenal axis.

Both types of dehydration generate a set of adaptive autonomic and behavioral motor responses that target gastrointestinal function to conserve water, although cellular dehydration is more closely associated with actions on the gut. A dry mouth is the sensation that is often identified with thirst. Indeed, one of the earliest theories for the generation of drinking behavior proposed that a dry mouth was the most important sensory trigger (Rolls and Rolls 1982). Although many experiments have now disproved the idea that a dry mouth is the primary instigator of deficit-induced drinking, there is no doubt that it provides important sensory feedback signals that may be important for associating drinking behaviors with particular stimuli that could lead to dehydration. Reduced salivation and a dry mouth may occur as a consequence of reduced blood flow through the salivary glands, as well as reduced parasympathetic activation of salivation in response to increasing plasma osmolality. But a dry mouth also develops as a consequence of, for example, extended talking or eating dry food. Finally, if dehydration is maintained for a significant period decreased gastric motility, and eventually anorexia develop as protective responses to maintain body water by reducing the water investment required by digestion and reducing the inflow of osmolytes into the body (Watts 2001) (see also *Appetite Regulation*).

Finally, restorative mechanisms modify behavior to return water and sodium (in the case of ECV challenges) back into the body. The selection, timing, and magnitude of all these responses ultimately depends on the intensity and kinetics of the stimulus. For example, mild hyperosmolemia will stimulate vasopressin secretion but not anorexia.

4.2 Feedback and Inhibition

The ability to inhibit drinking when sufficient water has been drunk to cover for the deficit is as important a function as behavioral initiation; overhydration can be as much of a problem as dehydration. Two types of mechanism are available for this purpose, although each varies in importance depending on the species: alterations in plasma osmolality, and oropharyngeal and gastric sensory feedback. For alterations in plasma osmolality to be a major factor in inhibiting drinking behavior, they have to occur rapidly. In small animals this does not occur, and drinking is terminated before plasma osmolality has begun to respond to the absorption of water. In this instance oropharyngeal mechanisms activated by the act of drinking are important. In larger species however, the volumes of water imbibed during a drinking episode may be sufficient to lower plasma osmolality rapidly and inhibit drinking (Rolls and Rolls 1982).

See also: Amygdala (Amygdaloid Complex); Arousal, Neural Basis of; Circadian Rhythms; Hunger and Eating, Neural Basis of; Neural Representations of Direction (Head Direction Cells); Reinforcement: Neurochemical Substrates; Sleep: Neural Systems; Suprachiasmatic Nucleus

Bibliography

Bourque C W, Oliet S H 1997 Osmoreceptors in the central nervous system. *Annual Review of Physiology* **59**: 601–19

Denton D, Shade R, Zamarippa F, Egan G, Blair-West J, McKinley M, Fox P 1999a Correlation of regional cerebral blood flow and change of plasma sodium concentration during genesis and satiation of thirst. *Proceedings of the National Academy of Science USA* **96**: 2532–7

Denton D, Shade R, Zamarippa F, Egan G, Blair-West J, McKinley M, Lancaster J, Fox P 1999b Neuroimaging of genesis and satiation of thirst and an interoceptor-driven theory of origins of primary consciousness. *Proceedings of the National Academy of Science USA* **96**: 5304–9

Fitzsimons J T 1963 The effects of slow infusions of hypertonic solutions on drinking and drinking thresholds in rats. *Journal of Physiology* (*London*) **167**: 344–54

Fitzsimons J T 1998 Angistensin, thirst, and sodium appetite. *Physiological Reviews* **78**: 583–686

Grossman S P 1990 *Thirst and Sodium Appetite*. Academic Press, San Diego, CA

Kaufman S 1984 Role of right atrial receptors in the control of drinking in the rat. *Journal of Physiology* (*London*) **349**: 389–96

Oliet S H, Bourque C W 1993 Mechanosensitive channels transduce osmosensitivity in supraoptic neurons. *Nature* **364**: 341–3

Rolls B J, Rolls E T 1982 *Thirst*. Cambridge University Press, Cambridge, UK

Swanson L W 1987 The hypothalamus. In: Bjorklund A, Hökfelt T, Swanson L W (eds.) *Handbook of Chemical Neuroanatomy*. Elsevier, Amsterdam, pp. 1–124

Watts A G 2001 Neuropeptides and the integration of motor responses to dehydration. *Annual Review of Neuroscience* **24**: 357–84

A. G. Watts

Thorndike, Edward Lee (1874–1949)

'Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality' (E. L. Thorndike 1918, p. 16). Few careers as diverse as that of Edward Lee Thorndike could be summarized so succinctly. Yet the hallmark of all the work of this extremely prolific psychologist was his abiding faith in the quantifiability of all aspects of human experience. From his first abortive attempts in 1896 to study mind reading through his last publication on shifts of interest with age, he was committed to quantification and, where possible, the controlled experiment. 'Professor Thorndike's major life-purpose was to demonstrate the unrivaled fruitfulness of the scientific method in the solution of social problems and to introduce it in exacting, quantitative forms which had been so productive in the physical fields' (Gates 1949, p. 31).

1. *Life Events*

E. L. Thorndike was born August 31, 1874 in Williamsburg, Massachusetts, the second son of Edward Roberts and Abbie (Ladd) Thorndike. The senior Thorndike was a Methodist minister who served churches in Maine and Massachusetts. The family often spent parts of their summers during E. L. Thorndike's childhood at an evangelical Methodist camp meeting on Martha's Vineyard. Perhaps in part as a reaction to the intensely religious nature of his upbringing, E. L. Thorndike rejected religious practices almost as soon as he left home for study at Wesleyan University. Throughout his life, he maintained a posture that favored science over religion.

After graduating with highest honors from Wesleyan in 1895, Thorndike went to Harvard University to study with William James (see *James, William* (*1842–1910*)). It was under James that he began his psychological studies, first of mind reading (a failure), then of learning (a success). The course of his studies for the next two years took him from William James at Harvard to James McKeen Cattell at Columbia, and from maze learning in chicks to his famous puzzle box studies with cats, in which he traced for the first time the trial-and-error learning process as his hungry subjects attempted to escape from crude cages to reach food. His doctoral dissertation describing these experiments, *Animal Intelligence: An Experimental Study of the Associative Processes in Animals* (1898), is one of the most well-known and influential studies in all of psychology.

Upon completion of his doctorate, Thorndike took a position at Western Reserve University. After one year he gladly accepted an invitation to return to New York City and become a member of the founding faculty at Teachers College in the fall of 1899. Thorndike remained at Teachers College until his retirement in 1940, a period that saw the publication of approximately 450 books, monographs, and papers. Another 70 were published in retirement.

Thorndike married Elizabeth Moulton on August 29, 1900. They had four children who survived infancy. Feeling his own training had been very deficient in mathematics, Thorndike insisted that all his children develop a solid background in mathematics. All of the children earned doctorates in scientific fields, and the second son, Robert L., followed in his father's footsteps so closely that he eventually occupied E. L.'s office at Teachers College.

E. L. Thorndike's work was widely recognized within his own lifetime. He was elected President of the American Psychological Association in 1912, President of the American Association for the Advancement of Science in 1933, and served as Second President of the Psychometric Society in 1936. In 1925 he was awarded the Butler medal in gold by Columbia University. Cattell found him ranked first among American psychologists in a 1921 poll for American Men of Science. Many of Thorndike's friends and colleagues said he was a 'workaholic,' but what he enjoyed more than anything was simply messing around with data. Other than tennis (he built his own court) and golf (he used a putter anywhere within 50 yards of the green), his recreation was analyzing data. Retirement was not a happy time for Thorndike. His health was not good: he had been a heavy smoker and quite overweight much of his adult life, and he suffered from arteriosclerosis. But what troubled him most in his old age was an acute awareness of his own declining cognitive abilities. His mind was no longer capable of the feats of reasoning that had been central to his professional life (R. L. Thorndike, personal communication). He died of a massive cerebral hemorrhage, at his home in Montrose, New York, on August 9, 1949, the day on which plans to celebrate his 75th birthday were to be finalized.

2. *Thorndike's Research and Theory*

As a result of his experiments with learning in cats, E. L. Thorndike formulated a theory of mental associations which he called connectionism. He believed that the central nervous system was the basis of all behavior, that learning and other behavioral modifications took place when connections were formed between neurons, and that there was a continuity

among species and a similar continuity in human development. 'The intellectual evolution of the race consists in the number, delicacy, complexity, permanence, and speed of formation of associations. In man this increase reaches such a point that an apparently new type of mind results, which conceals the continuity of the process' (E. L. Thorndike 1901, p. 65).

It has been suggested that Thorndike was driven to applied work with humans by employment pressures. This may have been true for Robert Yerkes (R. M. Thorndike 1990) but in E. L. Thorndike's case his first interest was humans and he resorted to animal research 'because they wouldn't let him work with children' (R. L. Thorndike, personal communication).

Upon returning to Columbia, Thorndike quickly turned his attention to human problems. He was fascinated by individual differences, trying to measure them, explain them, and use them for prediction. He believed, as did many of his generation, that science could solve all human problems, and he set out to apply scientific method to problems of education. Much of the remainder of his career was spent trying to develop methods to measure individual differences and to use these measurements to make education more efficient and successful.

Thorndike set out immediately to place education on a sound scientific footing. Educational practice of the time, the doctrine of mental discipline, was guided by 'the teachers' faith that mental abilities are so little specialized that improvement in any one of them will help all the rest; manual training is often introduced into schools on the strength of somebody's confidence that skill in movement is intimately connected with efficiency in thinking' (E. L. Thorndike 1903, p. 2).

Curriculum was founded on opinion, not evidence. Based on his transfer of training studies with Woodworth (Thorndike and Woodworth 1901) and similar studies he and his students conducted, he concluded that 'since those who succeed in the study of Latin are better at general discrimination and judgment than those who fail, we conclude that learning Latin vastly improves general discrimination and judgment … It suits the vanity of educational theory to fancy that the changes [that result from study] are wholly due to discipline' (E. L. Thorndike 1903, p. 93).

Thorndike continued his pursuit of quantification and applying scientific method in education with his publication of one of the first books on psychological and educational measurement and statistics in 1904. Here he noted the difficulty of defining interval scales for measuring cognitive activities, advocated the use of norm-referenced measurement as pioneered by Galton, and sounded a strong warning concerning the reliability and validity of measurements.

From the beginning, Thorndike adopted the view that inherited differences in cognitive abilities were important. 'There is no reason to suppose that the brain is less influenced by [immediate ancestry] than the tissues that cause height. … Immediate ancestry is thus the probable cause for original mental nature' (E. L. Thorndike 1903, p. 53). 'The schools always have and always will work to create a caste, to emphasize inequalities. Our care should be that they emphasize inequalities, not of adventitious circumstances, but of intellect, energy, idealism, and achievement' (p. 96). Thorndike was a thoroughgoing intellectual meritocrat, but he opposed discrimination on any other basis.

The rise of behaviorism in the 1920s shifted psychological research away from an interest in 'original nature' for almost 60 years. Many of the findings of modern behavioral genetics and cognitive psychology support positions Thorndike proposed at about the time genes were being rediscovered or the mind was being banished from psychology. For example, he argued that there was an interaction between genotype and environment. Anticipating current thought, he suggested that 'Each nature in some measure selects its own environment, and each nature may get from an environment a different influence' (E. L. Thorndike 1911, p. 48).

In addition to his efforts to make education more efficient and scientifically sound, Thorndike expended considerable effort from 1900 to 1925 in developing measures of what he referred to as intellect. He rejected the notion proposed by Charles Spearman that there was a single mental trait of intelligence, arguing instead that each test must specify what aspects of intellect it was designed to measure (see R. M. Thorndike 1990 for a description of the Spearman–Thorndike debate). Beginning in 1903, and anticipating more popular work by Alfred Binet (see *Binet, Alfred* (*1857–1911*)), Thorndike and his students developed a wide variety of measures of human abilities. The work culminated in the CAVD, a 'test constructed (from 1922 to 1925) as a sample of what a measuring instrument for a mental ability should be' (E. L. Thorndike 1949, p. v). With the CAVD, Thorndike (Thorndike et al. 1927) produced a test that would foreshadow item response theory and the test design methods of the 1990s.

In the 1930s, Thorndike turned his attention to issues of social science and social value. He pioneered the scientific study of environmental quality with his research on urban environments in Your City (1939) and he subjected the domain of values to scientific scrutiny in *Science and Values* (1936). But perhaps Thorndike had his greatest impact during this period through his contributions to lexicography. In the 1930s Thorndike revolutionized the dictionary by placing the design of these reference works on an empirical footing.

For years Thorndike had been collecting data on the relative frequency of words in published English. A list of the most frequently appearing 20,000 words, along with their relative frequencies, was published in 1931. The list was expanded to 30,000 entries in 1944.

Thorndike used word counts to determine which words should appear in his dictionary, and semantic counts (the most common meanings of each word) were used to select the definitions and their order. In addition, Thorndike insisted that each word be defined only in terms of words that would be more familiar than the one being defined. The Thorndike Century Junior Dictionary, published in 1935, was the first dictionary to use these principles and to focus its content on the needs of its most probable users. By 1950, the three Thorndike Century dictionaries dominated the field of dictionaries in education, and every serious competitor was copying his approach. All modern dictionaries follow his pattern.

E. L. Thorndike's influence is so pervasive in modern education and psychology that it is almost impossible to detect. If his contributions to lexicography were not enough, he pioneered the scientific study of education, made major contributions to psychological and educational measurement and testing, and put comparative psychology on a firm experimental footing. There is hardly a specialty in either field, with the possible exception of some clinical areas, that has not been significantly touched by his work or that of his hundreds of students, a list of which would read like a Who's Who of psychology and education in the middle and latter half of the twentieth century, including Abraham Maslow (who rejected his empiricism) and Raymond Cattell (see *Cattell, Raymond Bernard* (*1905–98*)) (who embraced it). A fairly complete list of the publications of E. L. Thorndike is available in Gates (1949); a full biography is given in Joncich (1968).

See also: Animal Cognition; Cognitive Psychology: History; Connectionist Models of Concept Learning; Experimentation in Psychology, History of; Learning Curve, The; Learning Theories and Educational Paradigms; Learning to Learn; Lexicology and Lexicography; Pavlov, Ivan Petrovich (1849–1936); Psychology: Historical and Cultural Perspectives; Quantification in the History of the Social Sciences; Skinner, Burrhus Frederick (1904–90); Watson, John Broadus (1878–1958)

Bibliography

Gates A I 1949 The writings of Edward L. Thorndike. *Teachers College Record* **51**: 28–31

Joncich G 1968 *The Sane Positivist: A Biography of Edward Lee Thorndike*, 1st edn. Wesleyan University Press, Middletown, CT

Thorndike E L 1898 *Animal Intelligence: An Experimental Study of the Associative Processes in Animals*. Columbia University Ph.D. thesis

Thorndike E L 1901 Evolution of human intellect. *Popular Science Monthly* **60**: 58–65

Thorndike E L 1903 *Educational Psychology*. The Science Press, New York

Thorndike E L 1911 *Individuality*. Houghton Mifflin, New York

Thorndike E L 1918 The nature, purposes, and general methods of measurement of educational products. In: Courtis S A (ed.) *The Measurement of Educational Products* (17th Yearbook of the National Society for the Study of Education, Pt. 2. pp. 16–24). Public School, Bloomington, IL

Thorndike E L 1949 *Selected Writings from a Connectionist's Psychology*. Appleton-Century-Crofts, New York

Thorndike E L, Woodworth R S 1901 The influence of improvement in one mental function on the efficiency of other functions. *Psychological Review* **8**: 247–61, 384–95, 556–64

Thorndike E L, Bregman E O, Cobb M V, Woodyard E 1927 *The Measurement of Intelligence*. Teachers College Bureau, New York

Thorndike R M 1990 *A Century of Ability Testing*. Riverside, Chicago, IL

R. M. Thorndike

Time, Anthropology of

Sociocultural constructions of time and time reckoning are now central topics of anthropological inquiry. Many cultures lack an explicit category 'time.' Still, time is everywhere present, in numerous local figurations that may or may not evidence a few universal forms, and thus the anthropological spectrum of research, theories, and contentions is vast (Gell 1992, Munn 1992). I offer here a representative, nonprogrammatic review of the topic that focuses on some of the important concepts, debates, and intriguing new areas of investigation.

1. Universal Time

Certain natural phenomena appear to be universally temporalized in terms of periodicity and, in some but not all cultures, progressive quantification: the human lifecycle and bodily processes (e.g., menstruation), seasonality, celestial patterns, day/night, and so forth. The same is true for the reproduction of social order by generation and various social and kinship groups. All societies regularly coordinate labor, occasionally punctuate the tempo of everyday life with ritual, and envision some type of past, present, and future. Calendrical arrangements of days, weeks, months, and years are also universal—but, again, not everywhere counted. In anthropological theory, time is implicitly understood to be binary. For example, all cultures accommodate conceptual categories and linguistic terms for duration and sequence, which are often said to be the basic forms of time. Leach (1961) proposed that time is a social construction rooted in two basic human experiences: the repetitions of nature, and the irreversibility of life. Many anthropologists contend that all societies possess both incremental/linear/irreversible and episodic/cyclical/repetitive forms of time—which are not necessarily antithetical since time

cycles can return to the 'same logical, not temporal, point' (Howe 1981). While Farriss (1987) claims that any one mode can incorporate the other, Leach proposes that all religions deny the finality of human mortality by subsuming linear time under a cyclical framework. Birth and death become two phases of an eternal sequence. Leach also claimed that time is everywhere a 'sequence of oscillations between polar opposites.' This temporal pendulum is related to basic sociological processes such as reciprocal gift-exchange and marriage. More recently, Gell (1992) argued that all culture-specific patterns of temporality are variations of two cognitive modes: an A-series which orders events according to relative notions of pastness, presentness, and futurity, and a B-series of absolute before/after.

2. *Social Time*

Universal conjectures notwithstanding, most anthropologists focus on local notions of time and subscribe to the Durkheimian thesis that time categories and conventions are 'social representations' that mirror the organization of society. This sociological perspective is best exemplified by Evans-Pritchard's (1939) elegant study of the East African Nuer, which is perhaps the most widely-known anthropological account of a distinctly non-Western style of time. Nuer 'oecological time,' which is cyclical or seasonal, reflects the environmental rhythms of cattle pastoralism. More importantly, Nuer indicate time but they lack an abstract chronological system that is numerical, cumulative, uniformly incremental, and measurable. Beyond the year, Nuer speak about temporal duration, intervals, and sequences with sole reference to 'structural time' or significant genealogical events in the social system—say, the division of a lineage, or the birth of an age set. Since Nuer genealogies are fixed at a depth of five generations, too, the passage of time is contained within a static framework. Similarly, there is considerable debate over whether or not the terminological groupings and repetitions in many kinship systems freeze, reverse, or emphasize the passage of time (Silverman 1997). Here, too, the opposition between cyclical and linear time is a leitmotif of anthropological analysis.

3. *Cultural Time*

The Durkheimian approach to time, enshrined in British social anthropology, anchors temporal conceptions to social organization. An alternative approach to time, represented by the 'interpretivist' tradition of American cultural anthropology, highlights the symbolic dimensions of time. Here, analysis seeks to discern the relationship between general temporal attitudes and the overall emotional emphases and focal symbols of a cultural worldview. Geertz (1973) suggested that the cyclical structure of Balinese kinship, naming, and calendrical systems 'don't tell you what time it is; they tell you what kind of time it is.' In a celebrated passage, Geertz concluded that 'Balinese social life takes place in a motionless present, a vectorless now' in which 'time lacks motion because Balinese social life lacks climax.' Moreover, the Balinese 'immobilization of time' is linked to the 'anonymization of persons.' Ultimately, the formal complexity of Balinese calendars pertains not to social order but to the unique, somewhat aesthetic themes and tones of Balinese culture and personhood. Whereas Evans-Pritchard pioneered the study of how others think about temporality, Geertz has been an inspiration to anthropologists who seek to understand how others experience time.

4. *Political Time*

One key debate in the anthropology of time, then, concerns the locus of analysis—broadly speaking, society vs. culture—and the question of whether or not time configurations can be reduced to a more elemental, causal factor such as human ecology. Another quarrel arises from the materialist division between infrastructure and superstructure. Bloch (1977) criticized Geertz for failing to distinguish between ritual time, which does differ significantly cross-culturally, and universally cognized 'normal' time that organizes practical activities. Hierarchical societies, Bloch continued, tend to elaborate a ceremonial view of timeless time. Consequently, ritual concepts of time are ideologies that sustain everyday productive inequalities. But while Geertz, in this view, may be guilty of foregrounding difference or 'radical temporal alterity,' and thus neglects the politics of time, Bloch himself assumes that people everywhere subscribe to the same 'practical reason,' which remarkably resembles the economic concerns of Western Marxism. Many anthropologists emphasize the lability of time concepts and, more recently, the political manipulation of time. The Northern Thai calendar, which is oriented around a 'fatal quality of time,' enables post hoc explanations of misfortune that exonerate individuals from personal blame (Davis 1976). The Kodi of Indonesia, for whom calendrical knowledge is prestigious, manipulate temporal sequences and durations, question the timing of ceremonial feasts and other public events, and strategically stage disruptions (Hoskins 1993). The Kodi calendar is also a 'debating ground' as traditional forms of time now clash with exoteric calendars associated with the state, Christianity, and modernity. Multiple, or layered, forms of time which are not, moreover, necessarily integrated into a single unifying system are common. One temporal series may regulate markets, while another system organizes lunation. Time is often contextual—

as well as competing and seemingly contradictory. Kabyle precapitalist peasants of Algeria submit to the present but exalt foresight within the temporal horizon of the 'forthcoming.' They also view the future—which capitalist hubris, in this view, seeks to predict—as a void known only to God (Bourdieu 1963). A final note on the politics of time: when competing groups anchor their prestige in a politico-ritual system to ancestral primacy, the mere assertion that a temporal series, or time itself, has a specific beginning point or location is inevitably political since it denies the validity of rival times. This is true for small-scale societies as well as the contemporary USA, which remains embroiled in debate between evolution and creationism.

5. *Ritual Time*

Not only is the differentiation of ritual from mundane time important to many anthropologists, but so is the construction of time within ritual itself. Humans tend, wrote Leach (1961), to conceptualize time as a repeating series of different intervals. Each year is distinctive but they all begin and end with an identical festival. Significant calendric events are everywhere marked by ceremonial antics and attire that are censored from everyday life. These periodic ritual inversions or 'rites of passage' actually create time by introducing intervals in social life. Yet Barnes (1974, pp. 128–9) offered the opposite view for the Kédang of Indonesia: ritual insures precisely that time is not cyclical—that the dead, for example, are irreversibly dead, and that calendric cyclicity is unable to prevail over linear temporality. From a different perspective, Gell (1992, pp. 37–53) recently evoked Lévi-Strauss and argued that the ritual reversal or regression of normal time among the Umeda of Papua New Guinea resolves problems that arise from the imposition of 'cultural' conventions such as temporal classifications onto the 'real' or 'natural' world.

6. *Language of Time*

Unlike spatial deixis, the linguistic study of time remains sparse in anthropology. Yet the famous Sapir–Whorf hypothesis of linguistic relativity was initially popularized through the alleged absence of Hopi verb tenses (now disproved). Still, language shapes our experience of the world, including time. Hence, the Trobriand Islanders of Papua New Guinea are unable to discuss and conceptualize linear, chronological sequences of events (Lee 1950). They have no time line but only temporal points organized as noncausal patterns with 'at most a swelling in value.' Devoid of linguistic temporal connections, Trobriand events are self-contained: 'a series of beings but not becoming.' Related studies of language and time include past vs. future-oriented locutions (e.g., Dundes 1969), orality vs. literacy (Goody 1991), and cognitive 'maps' of spatiotemporal passage (Frake 1985).

7. *Spatial and Bodily Time*

The spatialization of time and the temporality of space are topics of growing anthropological interest (see Munn 1992). Many cultures—e.g., Melanesia and Aboriginal Australia (Silverman 1997)—understand the primordial creation of the universe through a series of ancestral migrations, radiating outwards from an origin place, that simultaneously generated space and time. But other cultures possessing a temporal topology, such as the Iraqw of Tanzania, lack an origin: 'there is no zero point against which to set a regularized chronology' (Thornton 1980, p. 184). Spatiotemporal mythologies occur across lowland South America, where time and space are also embodied. The passage of time in the external world—the flow of a river, the movement of the sun—represents the human lifecycle. Conversely, the temporal unfolding of physiological processes such as sexuality and alimentation are projected onto the local topography (e.g., Hugh-Jones 1979, Chap. 7). The experience of time is thus understood through a bodily schema, specifically, the circulation of substances into and out of the body, and wider cycles of birth and death.

8. *Gendered Time*

The relationship between gender and time is another promising area of anthropological study. The riverine Eastern Iatmul of Papua New Guinea associate water with uterine fertility, death, and the inexorable erosion of land and trees that threatens to regress the cosmos back to the timeless and feminine era that preceded male creation (Silverman 1997). Like the river, female time flows to uncertainty. Male time, often conceptualized as the planting of trees and the establishment of villages, progresses forward against the flow of female temporality. But the arboreal ascendance of masculine structure is illusory. The river and rainy season floods ultimately prevail. All men can do to stem the feminine tide of regressive time is plant trees, build houses, and enact rituals. Weiner (1991) also discusses the gender of time in a Melanesian society, and links temporality to linguistic iconicity and motions through the landscape and body.

9. *The Future and End of Time*

Anthropologists are no longer content merely to bemoan the end of the 'time of the primitive.' Nor do we continue to define non-Western 'Otherness' as an essentially temporal construct, communities who are removed from us in both space and time (Fabian 1983). Instead, anthropologists are increasingly interested in historical conjunctions and contemporary practices that blur the distinctions between local and global times. Through temporal ambiguities, local people exercise creativity and resistance. While all societies allocate time between different activities

(Gross 1984), the spread of global capitalism and wage-labor has imposed on people everywhere computational, monetary, and aggressive notions of temporality such as saving, wasting, and fighting against time. But the future of time, to invoke a common idiom, is not confined to 'time is money.' The millennial turn encouraged many communities to confront the historical dissonances of culture contact by rethinking conventional notions of past, present, and future, cyclical and linear time, tradition and modernity, and the relationship between the biblical apocalypse and ancestral spirits. The millennium was a time to ponder the end of time (Scaglion 2000)—as well as, in the context of Y2K, the opposition between 'virtual time' and 'real time.' Thus the opening decades of this millennium, including its end-of-time scenarios, will doubtless prove to be a propitious era for the anthropological study of time both globally, locally, and in the newest temporal zone: cybertime.

See also: Cycles: Social; Millennialism; Ritual; Time, Chronology, and Periodization in History; Time-geography; Time–Space in Geography; Time-use and Gender

Bibliography

Barnes R H 1974 *Kédang: A Study of the Collective Thought of an Eastern Indonesian People*. Clarendon, Oxford, UK

Bloch M 1977 The past in the present in the past. *Man* (N.S.) **12**: 278–92

Bourdieu P 1963 The attitude of the Algerian peasant toward time. In: Pitt-Rivers J (ed.). *Mediterranean Countrymen: Essays in the Social Anthropology of the Mediterranean*. Mouton, Paris, pp. 55–72

Davis R 1976 The Northern Thai calendar and its uses. *Anthropos* **71**: 3–32

Dundes A 1969 Thinking ahead: A folkloristic reflection on the future orientation in American worldview. *Anthropological Quarterly* **42**: 53–72

Evans-Pritchard E E 1939 Nuer time reckoning. *Africa* **12**: 189–216

Fabian J 1983 *Time and the Other: How Anthropology Makes Its Object*. Columbia University Press, New York

Farriss N M 1987 Remembering the future, anticipating the past: History, time, and cosmology among the Maya of Yucatan. *Comparative Studies in Society and History* **29**: 566–93

Frake C O 1985 Cognitive maps of time and tide among medieval seafarers. *Man* (N.S.) **20**: 254–70

Geertz C 1973 Person, time, and conduct in Bali. In: *The Interpretation of Cultures*. Basic Book, New York, pp. 360–411 (Originally 1966)

Gell A 1992 *The Anthropology of Time: Cultural Constructions of Temporal Maps and Images*. Berg, Oxford, UK

Goody J 1991 The time of telling and the telling of time in written and oral cultures. In: Bender J, Wellbery D E (eds.) *Chronotypes: The Construction of Time*. Stanford University Press, Stanford, CA, pp. 77–96

Gross D R 1984 Time allocation: A tool for the study of cultural behavior. *Annual Review of Anthropology* **13**: 519–58

Hoskins J 1993 *The Play of Time: Kodi Perspectives on Calendars, History, and Exchange*. University of California Press, Berkeley, CA

Howe L E A 1981 The social determination of knowledge: Maurice Bloch and Balinese time. *Man* **16**: 220–34

Hugh-Jones C 1979 *From the Milk River: Spatial and Temporal Processes in Northwest Amazonia*. Cambridge University Press, Cambridge, UK

Leach E 1961 Two essays concerning the symbolic representation of time. In: (ed.) *Rethinking Anthropology*. Athlone, London, pp. 124–36

Lee D 1950 Lineal and nonlineal codifications of reality. *Psychosomatic Medicine* **12**: 89–97

Munn N D 1992 The cultural anthropology of time: A critical essay. *Annual Review of Anthropology* **21**: 93–123

Scaglion R 2000 Zero hour: Reflecting backward, looking forward. *Ethnohistory* **47**: 227–40

Silverman E K 1997 Politics, gender, and time in Melanesia and Aboriginal Australia. *Ethnology* **36**: 101–21

Thornton R J 1980 *Space, Time, and Culture Among the Iraqw of Tanzania*. Academic Press, New York

Weiner J F 1991 *The Empty Place: Poetry, Space, and Being Among the Foi of Papua New Guinea*. Indiana University Press, Bloomington, IN

E. K. Silverman

Time, Chronology, and Periodization in History

1. General Remarks on Historical Time

Homo sapiens is the only species on earth equipped to look back on its past consciously. The past, however, is chaotic. Nevertheless, it is the peculiar ability of humankind, as *homo historicus*, to put some kind of order on this chaos, making it understandable. It is this conscious looking back to the past that identifies *homo sapiens* as human.

The most important element in bringing order to the chaos of the past is some concept of time. Time is a human invention rendering chaos intelligible, rather than something that exists in itself. Among the various notions of time, chronology and periodization are the most closely connected with the organization of our past. Without these conceptual inventions, there would be no way of organizing our past as history.

Chronology is a dating system, or a system of reckoning a series of years from a starting point (year), which gives an era a theoretical definition. In this respect, chronology should be distinguished from a calendar, which is defined as a system of fixing the year, month and date.

Periodization is a form of historiological cognition proposed by the human mind for making the past intelligible and meaningful by dividing the collective past into compartments along time.

On the one hand, chronology is different from periodization in that it is used by a group of people for practical purposes, mainly for counting years. Those who try to make sense of history through rendering the past retrospectively intelligible, on the other hand, mostly use periodization. It is of no practical use for counting time.

Chronology, therefore, is in these terms mainly for constructing time, and periodization for dividing time. However, further inquiry reveals the fact that chronology also acts to divide time by creating a starting point for a new era, for example, 'regnal years,' 'Christian chronology,' and 'East Asian era names.' On the other hand, the very act of the periodization of the past effectively constructs a new time by inventing an arbitrary starting point, though these times are not used by people in everyday life, as for example the 'Victorian period' or the 'Middle Ages.'

2. *Four Aspects of Time in History*

Time in history will be considered from the following four viewpoints: the three types of time, the relation between linear time and cyclical time, the relation between what I call the stream of time, and the stack of time, and finally, what governs time.

2.1 *Three Types of Time*

Time will be classified into three categories: natural time, human time, and metaphysical time. Natural time is a reflection of such natural cycles as seasons, human life, etc. Human time is what *homo sapiens* invented to make the remembered past intelligible and comprehensible, and consists of such historical times as chronology and periodization. Metaphysical time is what *homo sapiens* created to make the remembered past meaningful beyond human experience, as has been proposed by philosophers of history.

2.2 *Linear Time and Cyclical Time*

Among Western historians time in history has been discussed in terms of the established dichotomy of linear and cyclical time. There is a consensus among them that a culture based on non-linear time would not be able to construct historical thinking (Needham 1966). This was based on the assumption that the rise of the historical mind is premised on linear time. The assumption is however contradicted by the historiographical accomplishments of East Asia, which led the pre-modern world in both the quality and quantity of historical production, despite the fact that linear conceptions of time were relatively weak in the historical cultures of the region.

Linear time has been almost universal in both the East and the West, despite variation in styles and manners, in the past and present. India has sometimes been described as being without the historical mind, because it was a culture that lacked linear time and lived in cyclical time. However, ancient Indian genealogies inscribed on tombstones and monuments show a richness of linear conceptions of time. Archeological remains discovered in ancient Mesoamerican civilizations also testify to the employment of linear time. The Mosi kingdom in West Africa still conducts the genealogical narrations and historical recitations of the dynasty through the sounding of a drum, and not by means of the spoken or written word.

2.3 *Stream of Time and Stack of Time*

The examination of what constitutes linear time will lead us to a comprehensive understanding of the nature of historical time.

Linear time proceeds from the viewpoint that time is ordered in a straight line like a stream of water in the river. This time is recorded by a sequence of numbers, as is characteristic in the Western (Christian) way of counting years, and it has given us a view of historical time as proceeding in a linear fashion.

Along with this idea, a frame of regnal years, or East Asian era names as a metamorphosis of regnal years has been used widely. This implies a view of time as being stacked like a pile of bricks, rather than flowing like water in the river. This image of time reflects a fundamental stratum of our consciousness, as when we look back into the past, and remember 'my school days,' 'our Tokyo period' or 'between the wars.'

The Western cultures that now dominate the modern world extended their mode of historical thinking based on linear time, and imposed their sense of time as universal. However, recent studies show that linear time is comparatively artificial, while compartmentalized time is in many ways more 'natural' (Sato 1991).

The fact that periodization was so very important in the development of historical studies, therefore, is really a reflection of European circumstances and the primary role that linear time has in that culture. This very rarely happened in the East Asian tradition, where varied combinations of compartmentalized time developed their own historical culture.

It could be said that the idea of a stream of time is rather artificial, and the idea of a stack of time actually constitutes the basic sense of historical time.

2.4 *Time Belongs to the Supreme Nominal Power*

Throughout history human time belonged to the supreme power. Time was made and was controlled by this power.

Historically, the powers that governed human time can be divided into two different types: religious power and political power. Religious powers in the western

hemisphere, in particular Christianity, Judaism, and Islam have controlled time, in broad terms. However, time in the eastern hemisphere has been governed by political power.

In China, time belonged to the emperor. A new dynasty and a new emperor declared the enthronement of a new power by proclaiming a new era name and a new calendar. The Chinese emperor confirmed the establishment of his power by the introduction of his time. In Japan, since the seventh century, time also belonged to the emperor. It was the emperor's duty and prerogative to declare a new era's name, a privilege retained even during military rule. Even the strong Tokugawa shogunate could not deprive the Emperor of this privilege. After World War II, the Christian chronology-based time came into use in East Asia, along with Greenwich Mean Time. However, Japan and Taiwan still employ the era name system officially. North Korea introduced an era name system in 1997, and the new era name was 'Juche' (Autonomy), to commemorate the late Kim Il-Sung. The era started retroactively from his birth in 1921.

The Western time based on Christian chronology and universal time now occupies the role of common time in the world. The introduction of Christian chronology made the practice of a world historiography possible by giving a synchronicity to the history of the world. The time that now governs the world belongs to the supreme world power called 'Globalization.' The expansion of Internet communications necessitates a world common time. A similar situation was found in early modern Europe. Problems arose for merchants in setting contract dates between different countries that used various regnal years and dating systems. As a result merchants were among the first to promote the use of a common Christian chronology strongly.

3. *Chronology*

Chronology can be classified into the three types according to their features: linear chronology, stratified chronology, and cyclical chronology. The major issue of chronology focuses on the question of why, in spite of its defects, the Christian chronology was able to become a common system.

3.1 *Linear Chronology*

Linear chronology is a system of reckoning a series of years in a straight line from a starting point (year) of an era. Jewish chronology, Christian chronology, Islamic chronology, Buddhist chronology, etc. belong to this type of chronology.

Linear chronology consists of two types: one counts the years serially from the starting year towards the future, and the other counts years both backwards and forward from the starting year. The latter idea of reckoning years both to the future and to the past endlessly is only found in Christian chronology.

This idea of a single era system of counting years is found all over the world. In ancient times, Mesopotamian chronology began with the accession of Nabonassar in 747 BC, and Roman chronology reckoned years from the foundation of Rome in 753 BC Jews have counted years from the foundation of the world (*Anno Mundi*) since the third century.

Dionysius Exiguus, an abbot in Rome in 525, first introduced Christian chronology as we know it, on the grounds that years should be noted from the incarnation of Jesus Christ, instead of the era of Diocletian, the persecutor of Christians in ancient Rome. Dionysius calculated the birth of Christ as the 753rd year from the foundation of Rome, and made this year the first year of *Anno Domini* (AD). However, it was not until the eleventh century that this chronology began to circulate widely in Europe.

Islamic chronology (the era of the *Hijra*) was instituted by Caliph 'Umar I in 638 by introducing the system of counting years from the *Hijra*, that is the flight of Mohammed from Mecca to Medina on July 16, 622, the first day of the lunar year. The Islamic chronology is unique among the other linear chronologies in that it is a purely lunar-based chronology, i.e., the year is counted as the cycle of twelve lunar months, not as the revolution of the earth around the sun.

Buddhist chronology is now used in some Southeast Asian countries. In Thailand and Cambodia years are counted from the putative date of the death of the Buddha in 543 BC, while Myanmar employs the date of the introduction of Buddhism in 638 as the first year of their chronology. In India, the government reformed the calendar in 1957 based on the Saka era (starting from 78), and it is now widely used.

At least ten other single-era linear chronologies have been proposed, including the era of the emperor Huang and the era of Confucius in China, and the era of Tangun in Korea.

3.2 *Stratified Chronology*

Stratified chronology is a system of reckoning years by successively piling eras (sets of years) upon one another in sequence, an idea that originated in the regnal year system. The method of counting years by royal succession has been used throughout the world since ancient times, and it was also used widely in Europe even after Christian chronology was circulated in the sixteenth century. The method of dating by regnal years, though troublesome, satisfies the theoretical requirements of specifying years, and this chronology finds its most significant expression in the era name system of East Asia. The era name system derives its origin from the idea of conferring a significant title to a certain series of years, instead of

using simple regnal years. It was first introduced by the emperor Wu of the Han dynasty in China, in 114 BC. The years of his reign were named according to Heavenly good omens as 'Founding the First Era,' 'The First Light,' 'The First Hunting,' etc.

This idea of chronology later rippled out in all directions to the countries on China's periphery. It was employed in Vietnam and Korea by the middle of the sixth century, and by the middle of the seventh century, at the latest, it had reached Japan. These East Asian countries on the Chinese periphery took the idea of the era name system but made little use of the Chinese names themselves. When used in tandem with the Chinese sexagesimal cyclical system, the era name system fulfilled its function as a method of reckoning years in East Asia.

East Asia was chronologically under the exclusive dominance of the era name and sexagesimal cycle systems. The era name system was premised on the very existence of Chinese letters as ideographic characters. Ideographic characters made possible the development of the East Asian philosophy of chronology, i.e., that the counting of years is not an act of assigning an ordinal number to a year, but rather that of giving to a year a meaningful name.

It originates in the *Weltanschauung* inherent in ideographic characters: that Chinese characters by their very nature assign a meaningful name to every aspect of the world and the era name system could be seen as an extension of this. It never crossed the minds of peoples outside East Asia to assign a meaningful name to a set of years.

For more than two thousand years, until the middle of the twentieth century, this era name system was employed in East Asia, to the virtual exclusion of all other modes of reckoning. The Christian era system was brought into East Asia during the latter half of the nineteenth century as part of Western culture. The system took the place of era names in the subsequent century. Japan, Taiwan, and North Korea still employ the era name system officially, though the Christian era system is widely used as well. It tells us that the idea of reckoning years within the context of meaningful names is deeply rooted and alive in the collective consciousness of East Asian people.

3.3 Cyclical Chronology

In addition to the above two types of chronology, cyclical chronologies have long been used all over the world. Well-known examples of these are the chronology by indiction (*indictio*) in early Medieval Europe, and the sexagesimal cycle in East Asia. In this type of chronology, years were given according to their place in the cycle of 15 years (*indictio*), or 60 years (the sexagesimal cycle), the number of the *indictio* or the sexagesimal cycle itself being ignored. They could be called chronologies in the sense that they turned into absolute chronology by giving ordinal numbers to the cycle, in the same way that the Olympic dating system operates. It was also possible to apply such an ordering to the sexagesimal cycle, but it was rarely used in practice. This method may also be called a relative chronology in the sense that it was not used to designate specified years in comparison with an absolute chronology.

The Hindu idea of chronology is cyclical, but it is unique in its philosophy of time in that cosmic cycles of creation and destruction continue indefinitely. In Hinduism, time belongs to the supreme God, and one chronological cycle (*yugas*) is 4,320,000 years.

3.4 The Christian Era System as the Common World Chronology

The twenty-first century begins in 2001, not 2000. The Christian chronology starts from the 1st Anno Domini. The concept of zero was not yet known in the West when the Christian chronology was created in the sixth century. Zero came to Europe around the tenth century, from India, via the Islamic countries. In addition to this defect, it is now believed that there was an error in assigning the year of Jesus's birth, and he was in fact born a few years before the year 1, Anno Domini.

In spite of these imperfections, we live with a Christian chronology that is now so dominant in worldwide usage that there is no possibility of it being replaced with a new or revised method at any time in the foreseeable future. Christian chronology now occupies the role of a common or neutral chronology in the world. This was theoretically brought into effect by introducing the concept of 'BC' and deciding on the first day of January as New Year's Day; and it had been politically determined by the dominance of Western culture since the nineteenth century.

It is important to inquire here why the invention of 'BC' made the Christian era system the universal chronology. It is because it has the great advantage of locating every historical event on a single time base.

Venerable Bede, in his *Historia ecclesiastica gentis Analorum* (731), first proposed the idea as *anno igitur ante incarnationem Dominicam* (in the year, therefore, before the incarnation of the Lord). The expression became widely used after the late sixteenth century in a period known as the dispute over Christian chronology. This dispute arose because of disagreements between scholars concerning the period between the Creation and the birth of Christ, and because of inconsistencies between Christian and non-Christian ancient histories, as for example in the differences between the Chinese records and the Bible. However, since the seventeenth century, the idea of BC has been used widely in Europe.

It should be remembered that BC was first used, not independently, but as an auxiliary chronology to *Anno Mundi*. The secularization of the Christian idea of history was necessary for the exclusive use of BC,

which started in the eighteenth century. An early example of the transformation of the Biblical chronology into a secular chronology can be found in Voltaire's *Essai sur les moeurs et l'esprit des nations* (1756), where the concept of BC became an indispensable part of modern chronology.

As a result of the developments in archaeology and geology in the scientific revolution of early modern Europe, the method of counting serially backwards from the birth of Christ proved extremely useful in dealing with the debate over 'the creation of the world.' It was, perhaps, accidental that BC came into such use, since it was originally conceived in order to describe the years between the Biblical Creation and the birth of Christ. It was beyond the imagination of its founders to think of a time 'before the Creation.'

It is a curious irony that this Christian chronology was so suited to tackle the scientific revolution, and it spread itself beyond the boundaries of Christendom. The invention of BC transformed the Christian era system from being a specifically Christian chronology to being one with a universal potential. It is now widely used in non-Christian countries as the main chronology. For example, China officially introduced the Christian chronology as the common era (*kungyuan*) in 1949, in order to become synchronized with the rest of the world. Joseph Needham proposed the idea of chronological expression avoiding the writing of dates as 'BC' and 'AD' by the use of '−' and '+' respectively in his *Science and Civilization in China*, because, 'such a convention seems also more suitable for civilizations which were never part of Christendom.' In recent years, Jews started to use the notations 'CE' (common era) and 'BCE' (before the common era).

The idea of a common chronology is premised on the notion that all peoples, cultures, and nations have their own 'year,' which allows the common man to create a balanced sense of history, and acquire the possibility of a free and varied sense of time. Because the reckoning of years is not a mere assignment of numbers in order, but the creation of a cognitive framework for experience and memory, it is in itself an advanced intellectual exercise peculiar to the human species. It is also a political, social, and cultural act of human intelligence.

4. Periodization

Periodization has been practiced since ancient times. Well-known examples are the four metallic ages (gold, silver, bronze, and iron) as introduced by Hesiod in the seventh century BC, Daniel's dream of the four kingdoms and its interpretation (Babylonian, Medo-Persian, Macedonian, and Roman monarchies), the Buddhist idea of the threefold periods (the true law, the reflected law, and the end of the law), and Saint Augustine's proposition of six periods (*infantia*, *puericia*, *adulescentia*, *iuventus*, *gravitas*, and *senectus*, the last named corresponding to the period from the Christ to the end of time). Each had a fundamental influence on the pre-modern people in its respective civilization. In the East, it was a matter of primary concern to know when the period of the end of the law started. Similarly, in the West people were eager to know the advent of Doomsday.

An examination of these periodizations shows us that: (a) at a very fundamental historical level people employ compartmentalized time and not linear time for their historical thinking, and (b) characteristically a people acquire their historical *raison d'être* using their past to place their present in history.

This article, however, focuses on the periodization that appeared with the rise of modern historical scholarship in the seventeenth century, and has become a major historiographical asset.

4.1 Two Types of Periodizing Terms

Time has been worked with in two ways, using both continuous time and compartmentalized time.

It is natural and inherent for us to employ compartmentalized time when we look back into the past. In the same way historians in their research invent compartmentalized time in order to give the past a certain unity. Thus compartmentalized time is a natural and indispensable historiographical tool.

Historians have employed compartmentalized time in two fundamentally different ways; one was to use the historically existing compartments, and the other was to invent and retrospectively apply a structure. Dynastic periodizations widely used in the East and the West are typical of the former; however, the latter represented, for example, by Christophus Cellarius, Karl Marx, Fernand Braudel have almost always been created in the West.

4.2 Periodization as a Historiological Tool

Periodization is a form of historical understanding, designed as a historiological tool for making the past understandable, intelligible, and meaningful by dividing it into compartments. It was developed largely in the West in the early modern period under the dominance of Christian chronology. More generally speaking, it was to be found in civilizations where linear chronology played a central role, for example Ibn Khaldun's distinction between the *umran badawi* (rural culture) and the *umran hahari* (urban culture).

Periodization became a key feature of Western historiography, which developed into a cognitive discipline by proposing new perspectives and varied interpretations. This contrasts strongly with the East Asian experience. Even though a rich historiographical culture flourished in East Asia for more than two thousand years as a major field of learning, it did not develop the type of terminologies used by Western

historians, except some periodizations proposed by Japanese historians because East Asian countries used the combination of many different types of periodization in their written history, such as dynasties, era names, sexagesimal cycles, and periods based on changes in the site of the capital. Modern European history, on the other hand, drew on a long and deep-rooted tradition of annals and chronicles, whose style depended solely on a concept of continuous time. The various periodizations were essential for their establishing history. Even now, a list of titles of recent works by Western historians reads like an inventory of variously compartmentalized time.

The Christian era system had been in use in the West long before the birth of historiography in modern Europe, at which time the historians took an interest in compartmentalizing the past in order to analyze history. In one sense, it could be viewed as a declaration of independence by modern historiography from such traditional forms as annals and chronicles.

In the 1700s, the word 'century' came into use, signifying that the modern Western mind needed compartmentalized time for thinking about history. In Western historiography such statements as 'the first half of a century,' 'the first quarter of a century,' or '1930s,' are used quite freely and this is a sign that compartmentalized time has become indispensable for looking back upon the past, which cannot be grasped simply by employing absolute years.

To summarize in simple terms, in the West the idea of a single era count formed the basis of chronology, i.e., the Christian era system. Given this basis historians employed many different kinds of periodization to compartmentalize time to think about history. By contrast, in East Asia only the compartmentalized time could be found in historiography. Their historiographical style was developed on a stratified or stack-shaped chronology and various other types of compartmentalized time. The expansion of modern European historiography, paralleling the global reach of Euro-American military and economic power in the late nineteenth century meant that historians everywhere are now using the technique of using periodization to understand history. '*Pensare la storia e certamente periodizzarla*' (to think history is certainly to periodize, B. Croce) is a view now shared by all world historians.

Later development of periodization to incorporate such concepts as the 'feudal period,' the 'medieval period,' etc., has given rise to the reinterpretation of human history on an international scale. The idea of dividing history into three periods was the first success of the method.

4.3 Universality of the Tripartite Periodization

The universal acceptance of a tripartite periodization: ancient, medieval, and modern has had a fundamental influence on the development of world historiographies.

This threefold periodization was first employed by Georg Hornius, a Dutch historian in his *Arca Noae* (1666), and became widely known later through Christophorus Cellarius, a German historian, in his *Historia nova, hoc est XVI et XVII saeculorum* (1696). With the transmission of modern German historical scholarship it spread across the world two centuries later and has been taken for granted as a global standard since then.

This threefold periodization found universal acceptance for two reasons: (a) the terminologies were almost free from religious connotations and Euro-Christian signs, even though '*medium Aevum*' was originally a theological notion, and (b) it included a universal attitude towards the past by instituting a negative period between the glorious past and the contemporary period.

There is still no agreement in deciding on the exact border years among the ancient, the medieval, and the modern. Professional historians have concentrated later on subdivisions of the three periods. This lack of consensus encouraged disputes among the historians of the world. As a result of this debate and the development of social and economic histories it became apparent that the very idea of border years was out-of-date.

Periodization as a way of historical thinking was developed still further by new proposals for dividing the human past. The most influential of these ideas of periodization were the 'feudal period,' pioneered by Henri Pirenne, and 'historical materialism,' as introduced by Marxist historians. This periodization was discussed in the historical contexts not only of Europe, but also of the non-Western world, especially China, India, and Japan.

After World War II, Fernand Braudel published a new idea of periodization in 1949, by proposing the notion of three different rhythms of history. The *long dureé* (the long duration) became the main concern of postwar historians. This structural periodization spread over the world on the wind of social history, and became the shared perspective of the historians of the world.

5. Visualization of Historical Time, Chronology, and Periodization

History is the science of making our chaotic past intelligible and meaningful through a variety of interpretive and structural techniques; among these, the imposition of a systematic notion of time has proved to be one of the most important. The construction of time lines, royal lines, genealogies, chronicles, annals, and historical tables, which have developed since ancient times, helps to form our historical knowledge and mind. The further necessity

of knowing the synchronicity of the plural lines was the first step towards reaching a historical consciousness. Peoples have proposed a variety of ideas to synchronize the different times of human history.

The synchronizing of historical events that happened in different places and times can be viewed as the singular most distinguished achievement in the historical consciousness of the human mind. This is reflected in the construction of chronological tables or historical tables.

It is necessary here to distinguish chronological tables from annals and chronicles; they have very different purposes. Annals and chronicles attach importance to the content they record; on the other hand, chronological tables stress the date or sequence of historical events.

The chronological table gives us a visualized temporal image of our past, in the same way that a world atlas shows us a visualized spatial image of the world. It has two important functions. One is to give a bird's eye view of the stream of time in history, the other is to make certain of the synchronicity of historical events in various parts of the world.

In East Asia, the use of chronological tables as a reference for comprehending history was first proposed by Sima Qiam (BC135–93) in his *Shiji* (Records of the Historian). They can be viewed as the beginning of a consciousness of historical time. He arranged 'the chronological table of the twelve dukes' by separating lines for the events of each of the twelve states. Reading down the columns, one sees what happened in each state in any given year by using the sexagesimal cycle and regnal years of that state. Reading across the columns, one can see the sequence of historical events for any given state. We could say that he made the table to show us history in both its temporal and spatial contexts.

The development of the chronological table in Europe was effected by choosing Anno Mundi, or Christian chronology, as the key time scale, around which the various historical events were listed.

The historical time-based Christian chronology was already established in Europe by the sixteenth century. However, at that time it was primarily employed in the field of ecclesiastical history. It was not until the eighteenth century that these two chronological methods were unified into Christian chronology. At that time historical events were rewritten in terms of the Christian chronology and thus were synchronized with events in the rest of the world. This synchronicity laid the foundations of a time consciousness capable of giving birth to the modern method of historical thinking. Benjamin Marchall's *Chronological Tables* (1713) is one of the earliest works of this genre. John Blair published *The Chronology and History* (London, 1754), in which he proposed a table of various regnal years in Europe, which was constructed around a time line of years since the birth of Christ. James Bell established the earliest style of modern historical tables in his *A View of Universal History* (London, 1842). Historical tables were disseminated throughout the world through the works of Karl Ploets (1819–81), a German historian.

Since the latter half of the nineteenth century, Christian chronology as the historiographical 'standard time' has spread throughout the non-Christian world. This chronology has been established as the pivotal time by which the world creates and maintains the possibility of a synchronicity of the various local historical times. It is only with this temporal foundation that it is possible to write a history of the world.

Bibliography

Blackburn B, Holford-Strevens L 1999 *The Oxford Companion of the Year*. Oxford University Press, Oxford, UK

Braudel F 1980 *On History*. The University of Chicago Press, Chicago

Eliade M 1954 *The Myth of the Eternal Return*. Pantheon Books, New York

Hall E 1983 *The Dance of Life: The Other Dimension of Time*. Anchor Press, New York

Finegan J 1964 *Handbook of Biblical Chronology*. Princeton University Press, Princeton, NJ

Fraser J T (ed.) 1972 *The Study of Time*. Springer-Verlag, Berlin

Lippincott K (ed.) 2000 *The Story of Time*. Merrell Holberton, London

Needham J 1966 Time and knowledge in China and the West. In: Fraser J T (ed.) *The Voice of Time: A Cooperative Survey of Man's Views of Time as Expressed by the Sciences and by the Humanities*. George Braziller, New York, pp. 92–13

Prusek J 1963 History and epics in China and in the West. *Diogenes* **42**: 20–43

Ricoeur P 1984–1985 *Time and Narrative*, 2 Vols. The University of Chicago Press, Chicago

Sato M 1991 Comparative ideas of chronology. *History and Theory* **30**: 26–50

Whitrow G J 1988 *Time in History*. Oxford University Press, Oxford, UK

M. Sato

Time-geography

The core of the discipline of human geography is often considered to be space. But, as the early navigators found, time is the inevitable correlate of space and it is therefore no surprise that geographers have made continuous contributions to thinking about time as well as space. In the social sciences, such contributions have been notable since at least the 1920s, with the rise of the school of human ecology which understood time and space as inseparable elements of human activity (Thrift 1977). But the major contributions to geographical thinking on time have been made since the 1960s and it is on this latter period that this article is focused.

The study of time has been a notably interdisciplinary field and this means that both geographers and others have worked on the spatial aspects of times, timing, temporal structures, and time consciousness. This article reflects that porosity in that it starts with specifically geographical contributions but then widens the discussion out from narrowly disciplinary considerations to more general geographical contributions.

1. *Specifically Geographical Contributions*

Geographers have made four main contributions to the study of time since the 1960s.

1.1 *Chronogeography*

Chronogeography was an attempt by Parkes and Thrift (1980) to extend the human ecological paradigm. In line with that paradigm, it was stimulated by essentially biological analogies. In particular, a heavy emphasis was put on rhythm. Rhythms were seen as operating at a number of different time–space scales, thereby providing the time–space backbone to different societies. In turn, these rhythms were produced by '*zeitgeber*' of various kinds, sociotemporal 'lighthouses' ranging from the heartbeat and other chronobiological phenomena at one end of the scale to various kinds of economic cycle at the other end, which produced different kinds of timed spaces and different ways of spacing time. Chronogeography was a phenomenon of the 1970s and early 1980s.

1.2 *Time-geography*

In marked contrast, so-called time-geography has made a sustained and lasting contribution to the geographical literature, notwithstanding considerable debates within geography as to the efficacy of certain of its features. Time-geography emerged as a prominent area of geographical work in the late 1960s and early 1970s in Sweden as a result of the work of the remarkably original thinker Torsten Hägerstrand (1982, Carlstein et al. 1978, Pred 1981). It approached the analysis of social interactions by focusing on the time–space paths of their participants, within temporal and spatial structures of interaction, movement, and communications. Time and space are conceived as resources in short supply on which individuals draw to provide 'room' in which they can realize projects. Time and space are facilitating resources in which collateral processes can occur. The use of time and space is constrained in three ways: by the power, wealth, skills, and knowledge of individual participants (capability constraints); by the conditions or circumstances in which people can interact with one another (coupling constraints); and by the power relations that mould people's access to particular spaces and shape the roles in which they do have access (authority constraints). The stress on constraints to the completion of projects provides a series of ways in which the power relations of a society operate by shaping the time–space patterns of human activities. Put another way, social institutions are structures of repetition through which the time–space coherence of societies can be maintained. They do this either by imposing particular projects or time–space routines within a society, or through internalizing certain projects and routines as 'natural' within people's everyday perceptions.

One of the most distinctive elements of time-geography was the unique graphic language that Hägerstrand deployed in order to describe the time–space structures of society. Often likened to a musical score, time–geographic diagrams have clearly been one of the most distinctive contributions made by geographers to the social sciences, since they are both analytic and evocative. Increasingly, as well, it has been realized that Hägerstrand was well ahead of his time in his attention to the nontextual elements of human life which time–geographic diagrams are so well able to conjure up.

Time-geography has been an enormously productive means of showing the time–space structure of situations because of its ability to be both general and concrete. It has been applied widely, both to consider particular problems of the allocation of time and space (most especially in planning) but also more generally, for example to look at problems of technological change or environmental degradation (Carlstein 1982). It has also become the basis of many important social theoretical ruminations.

Yet, like any approach, it is not without its problems. First, even with the benefit of greater computer power, many of its insights are still difficult to operationalize at the scale of a city or region, as opposed to the smaller communities that time-geography seems to fit so well. It is no surprise then that Hägerstrand developed another level of analysis that worked at this larger scale. Second, time-geography is often reticent in discussing struggle within societies, although researchers like Pred (1986) have very successfully addressed this issue using time–geographic approaches. Third, time-geography has been accused of neglecting the symbolic dimension. If this is true it is, in a sense, a philosophical neglect. Hägerstrand has been concerned to show, rather like actor–network theory, that a good deal of what is regarded as symbolic might be more profitably addressed in different ways (see Gren 1994; *Urban Activity Patterns*).

1.3 *Time–space Compression*

The third distinctive geographical contribution to the study of time and geography has been the concept now

known as 'time–space compression,' which is based on the notion of progressive shrinkage in the resources of time and space. From the 1920s onwards, geographers became concerned with what Marx called the 'annihilation of space by time,' as travel times between locations diminished because of new transport and telecommunications technologies. Until the 1980s, this was usually seen as an interesting empirical phenomenon. Then David Harvey (1989) linked 'time–space convergence' to wider processes of spatial extension and intensification within capitalism: the history of capitalism has been characterized by speed-up in the pace of life, while so overcoming barriers that the world seems to collapse in upon us. In other words, ever more intense time–space compression dislocates everyday practices and understandings. Harvey identified postmodernism as the latest form of dislocation and, in line with a political economy approach, was particularly keen to associate it with new post-Fordist production paradigms.

1.4 Nonrepresentational Theory

The last distinctive contribution to the study of time and geography has been nonrepresentational theory (Thrift 1996, 2000). Nonrepresentational theory is concerned with the flow of practices in time, with the 'presentations' produced by acting in the now rather than the *post hoc* reconstructions of the event which are the stuff of much of the social sciences. It draws inspiration from all the approaches so far considered—chronogeography for its interest in the different rhythmical structures of time and space, time-geography for its attention to detail as not just detail, and time–space compression for its attempt to understand certain characteristics of the modern world as integrally bound up with the construction of particular times and spaces—as well as numerous other elements of the social sciences, and especially various theories of practice. Nonrepresentational theory attempts to do two related things: first, to provide an ontology which takes 'mundane' practices seriously and, second, to provide various means of amplifying the creativity of these practices through various performative methods.

2. Geographical Themes in Work on Time

A key area for early work on time was provided by the simple fact of spatial variations in the time-patterns of life, the institutions and practices that were central to different societies' handling of times, and how different societies thought of time. Much of this material was derived from the anthropological literature. One key geographical concern was the identification of environmental or social factors underlying spatial patterns in times. Another was how these spatial patterns were changing as dominant Western patterns of timing diffused around the world. This last concern clearly derived from the 'modernization thesis' of economic and cultural change that dominated the social sciences from the mid-twentieth century (Thompson 1967).

But standard depictions of people of great temporal innocence have not survived close scrutiny, especially once time is conceived more broadly than as the clock time taken for granted in the twentieth-century West. The complexity and variety of traditional and non-Western views of time is now virtually axiomatic. In general, traditional cultures were not caught up in unchanging, nature-based and cyclical ideas of time. They subscribed to complex and sophisticated calendrical arrangements; were able to conceive of a linear time; did not lack a sense of their own past; and were not imprisoned by tenseless languages and a reliance on oral cultural transmission. Rather, the temporal values present in different traditional cultures display an incorrigible diversity (Gell 1992). This diversity is now recognized to be a crucial feature of contemporary societies as well, as can be demonstrated through four main themes.

2.1 Multiple Times and Spaces

Geographers have been among writers highlighting the coexistence of differing time-senses among social or ethnic subgroups within populations, linked to work on similar differences in senses of space and place.

Everyday experiences of time are structured by many influences besides the determinants of religion, trade, and industrialization so often emphasized (Thompson 1967). General or specific notions of timing and coordination also arise within households; from structures of preindustrialized work; from communications; from civic administration and law; from recreation and consumption; from disciplinary regimes such as prisons, workhouses, and hospitals; from cartography and navigation; and from the military.

Although most of these arise among relatively specialized temporal communities, it is essential to acknowledge that they coexisted in places, especially in towns, where the institutions of government, church, trade, and industry were most likely to come together, increasingly acting to impose artificial schedules on a population which then began to talk about these schedules as a natural occurrence (Thrift 1996). A geographical focus on time in places, rather than time in particular occupations, can hardly avoid drawing attention to the 'spillovers' of temporal practices within urban populations. In other words, temporal frameworks are only rarely the preserve of a single community of expertise, and everyday temporal orderings are commonly hybrids of several ordering frameworks, each arising from particular subgroups.

The point that any given place or space can take on multiple temporal forms for different social subgroups is rarely more clearly evident than for women. For a long time, women's times were not taken seriously by geographers, who drew on implicitly male social features. Though recent work has begun to remedy the neglect of everyday time practices relating to women, the general area of women's times and spaces cannot be said to be over-researched.

2.2 The Continuing Heterogeneity of Timed Spaces

Historians of time began with ideas of a linear progression from timeless to highly coordinated cultures, through a series of intermediate stages, with 'industrial time consciousness' diffusing through society and across space (Thompson 1967). But it is now clear that spatial differences in sense of time were not being obliterated. Notwithstanding considerable homogenization of industrial production techniques and of temporal frameworks (such as the geographical spread of Greenwich Mean Time) (Thrift 1981, Pawson 1992), there is clear evidence of international and inter-regional differences in temporal frameworks and their interpretations.

In turn, this realization made it possible to revalue time derived from 'natural' cues and to understand that senses of time derived from nature were often associated with impulses to disciplined and unstinting work, and hence attention to saving time. Again, it also became clear that industrial time consciousness was itself a heterogeneous creature, with very different practices instituted in different places.

2.3 Timing in Relation to Travel and Communications

The mobility of people, objects, finance, and information has many effects on time measurement and temporal practices, usually involving not only practical questions of coordination (such as the implementation of regular timetables and calendars for specific purposes), but also new techniques of time-calculation (interest, risk costs, and so on) and new conventions of everyday life. Five topics have been closely researched.

First, new temporal practices are often associated with long-distance trade, including various means for deferred payments and futures trading. Second, temporal precision has been central to knowledge of spatial positions at sea and on land, especially in calculating longitude. Third, the close dependence of administration, as well as business, on efficient handling of information over time and (relatively) rapid communication across space has formed an important theme in historical geographies of European world dominance. The emergence of standard times and time zones and the impact of electronic communications have also received particular attention (Kern 1982). A fourth focus has been on how temporal cues flow from one timed space to others. The obvious instance here is clock time, although even this is taken note of and reacted to differently in different timed spaces. Fifth, time has been studied in relation to tourism, where both the coordinated movement of people and the specific character of the 'tourist experience' itself rest on particular ways of handling, and of representing, time(s).

2.4 Temporal Symbolism as Constitutive of Time-senses

The inherently cultural character of any time-keeping system has become a commonplace observation. Since geography underwent a cultural turn in the 1980s, it is no surprise that the variegated cultural character of time-keeping systems has become a focus of research. This research has followed three main avenues of investigation. First, what counts as a 'task' is very much a matter of cultural values. Industrial geographers have shown how a task varies from place to place in how it is constructed and valued (Stein 1995). Second, there has been work on disputed interpretations of time, demonstrating its heavily symbolic character. Such work includes contemporary and historical studies of strikes during restructurings of industries. Parallel work has considered timings linked to national identities, especially involving disputes over calendars. Third, work considering language as language-in-use necessarily incorporates a temporal perspective.

3. Future Directions of Theory and Research

We end by suggesting some of the themes that might constitute a research agenda for future geographical work on times.

The first of these consists of work on time consciousness as the mediated imposition of a particular set of practices upon a population, which in turn affects the way a population makes accounts of practice. Much here echoes Thompson's (1967) approach, but it is pursued rather differently, with shifts in ideologies of everyday time practice and the consciousness of time indexed through three dimensions: discourses and texts; devices and instruments; and disciplines and routines. Each of these dimensions is strongly geographical. The effectiveness of temporal discourses is limited by the spatially uneven availability of texts and devices and by variations among temporal communities, which in turn shape systems of discipline and meaning that reinforce or undermine particular temporal conceptions and practices.

The second area of work is on instantaneity. Writers like Harvey and Virilio have discussed the supposed

effects of the increasing swiftness of many social practices, but none of this writing stems from any detailed ethnographic research. It all consists of assumptions—usually based in technological determinisms—masquerading as theory. Thus, we know remarkably little about the practices that have sprung up which have refigured the world as 'fast' and which are to be acted out as such. These practices are very diverse, and range all the way from different forms of visualization through different forms of economic practice, including those to do with international finance which have involved the evolving use of different forms of telecommunication stretching over the last 150 years, to different forms of warfare, including Virilio's 'chronopolitics'.

The third area of work consists of temporal policy. Since Kevin Lynch's (1972) *What Time is this Place?* geographers have been interested in practices of temporal planning that they might lay alongside those of spatial planning. Now various experiments in countries like Italy and Germany are starting to give these practices real expression. For example, six Italian cities have passed statutes providing temporal planning in matters such as the staggering of work and school hours so as to smooth traffic peaks. These may be minor interventions, but they begin to show what might be possible.

In conclusion, for a discipline raised on space, geography has had a remarkable amount to say about time. Indeed, recent writings on the importance of time–space as the focus of the discipline (e.g., Massey 1992, May and Thrift 2001) rather than simply space can only reinforce this tendency.

See also: Place in Geography; Space and Social Theory in Geography; Spatial-temporal Modeling; Time, Anthropology of; Time–Space in Geography

Bibliography

Carlstein T 1982 *Time, Resources, Society, and Ecology*. George Allen and Unwin, London

Carlstein T, Parkes D, Thrift N 1978 (eds.) *Timing Space and Spacing Time*. Arnold, London

Gell A 1992 *The Anthology of Time*. Berg, Oxford, UK

Glennie P, Thrift N 1996 Reworking E P Thompson's 'Time, work discipline and industrial capitalism.' *Time and Society* **5**: 275–99

Glennie P, Thrift N 2001 *The Measured Heart* (forthcoming)

Gren M 1994 *Earth Writing*. University of Göteborg, Göteborg, Sweden

Hägerstrand T 1982 Diorama, path and project. *Tijdschrift voor Economische en Sociale Geografie* **73**: 323–39

Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK

Kern S 1982 *The Culture of Time and Space 1880–1918*. Harvard University Press, Cambridge, MA

Lynch K 1972 *What Time is this Place?* MIT Press, Cambridge MA

Massey D 1992 Politics and space/time. *New Left Review* **196**: 65–84

May J, Thrift N 2001 (eds.) *Time Space Geographies of Temporality*. Routledge, New York

Parkes D, Thrift N 1980 *Times, Spaces, Places*. John Wiley, Chichester, UK

Pawson E 1992 Local times and standard time in New Zealand. *Journal of Historical Geography* **18**: 278–87

Pred A 1981 Social reproduction and the time-geography of everyday life. *Geografiska Annaler* **63B**: 5–22

Pred A 1986 *Place, Practice and Structure: Social and Spatias Transformation in Southern Sweden 1750–1850*. Barnes and Noble, Totowa, NJ

Stein J 1995 Time, space and social discipline: Factory life in Cornwall, Ontario 1867–1893. *Journal of Historical Geography* **21**: 278–99

Thompson E P 1967 Time, work discipline and industrial capitalism. *Past and Present* **38**: 56–97

Thrift N 1977 Time and theory in human geography. *Progress in Human Geography* **1**: 65–101 (part 1) and 415–57 (part 2)

Thrift N 1981 Owner's time and own time: The making of a capitalist time consciousness 1300–1880. In: Pred A (ed.) *Space and Time in Geography, Essays for Torsten Hägerstrand*. Academisk, Lund, Sweden, pp. 56–84

Thrift N 1988 Vicos Voco. Ringing the changes in the historical geography of time consciousness. In: Young M, Schuller T (eds.) *The Rhythms of Society*. Routledge, London, pp. 53–95

Thrift N 1996 *Spatial Formations*. Sage, London

Thrift N 2000 Afterwords. *Environment and Planning D. Society and Space* **18**: 213–56

P. Glennie and N. Thrift

Time Perception Models

Time perception is a distinct area of study with its own psychophysical methods designed for assessing the perceived duration of a temporal interval. There is now an extensive database to guide the development of models of time perception. Most timing models are firmly rooted in an *internal clock*. Internal clock models postulate a timing-specific mechanism which produces perceived temporal values that bear orderly relations to real time. This timing-specific mechanism, or internal clock, is an interval timer, rather than a periodic or oscillatory mechanism. A periodic clock, such as the kind responsible for circadian rhythms, is continuously running and self-sustaining. In contrast, interval clocks need to be initiated with a signal. Once started, the system goes to completion, comes to a rest on its own, and must be restarted.

The first psychophysical models for time perception were proposed at about the same time by Creelman (1962) and Treisman (1963). Creelman (1962) introduced the signal-detection approach, which was continued by Allan and Kristofferson (1974). The most sophisticated signal detection model is Killeen et al.'s (1997) *Pseudo-logistic model* (PLM). The information-processing approach was introduced by Treisman (1963) and continued by many others, including Eisler (1975), Thomas and Cantor (1978), and Zakay and

Block (1994). The most developed of the information-processing models is Gibbon et al.'s (1984) *Scalar timing* (ST), which has been reformulated as a connectionist model by Church and Broadbent (1991). In direct contrast to information-processing models is Killeen and Fetterman's (1988) behavioral theory of timing (BeT).

Because signal-detection models and information-processing models have been applied to both human and nonhuman timing data, we shall focus on such models in this article, presenting ST and PLM as exemplars of these two classes of models. Specific versions of ST and PLM have been derived for a variety of timing tasks, such as time-left, temporal generalization, and *temporal bisection*. We shall examine ST and PLM in their application to temporal bisection.

In the prototypic temporal bisection task (e.g., Allan and Gibbon 1991), the subject is familiarized with a constant pair of referents, one short (S) and the other long (L). On probe trials, a temporal interval t, $S \leqslant t \leqslant L$, is presented, and the subject is required to indicate whether t is more similar to $S(R_S)$ or to $L(R_L)$. Temporal bisection data are summarized as a psychometric function relating the proportion of long responses, $P(R_L)$, to probe duration t. The *bisection point* ($T_{1/2}$) is the value of t at which R_S and R_L occur with equal frequency, $P(R_L) = 0.5$. One interpretation of $T_{1/2}$ is that it is the point of subjective equality (PSE), identifying the perceived value of t that is midway between S and L. Empirically, it has been found that the location of $T_{1/2}$ is influenced by the size of the L-to-S ratio (r) and by the spacing of the probes (see Allan 1998). For $r \leqslant 2$, $T_{1/2}$ is around the geometric mean (GM) of the S and L referents, but moves toward their arithmetic mean (AM) as r increases. For $r > 2$, $T_{1/2}$ for linearly spaced probes is greater than that for logarithmically spaced probes, with the probe spacing effect becoming negligible for $r \leqslant 2$.

1. Scalar Property

Both ST and PLM incorporate *Weber's law* for the duration values used in most psychophysical experiments. According to Weber's law, for a given clock time $t > 0$, the variability in perceived time (σ_t) is *scalar* with mean perceived time (μ_t); that is, there is a constant $\gamma > 0$ such that

$$\frac{\sigma_t}{\mu_t} = \gamma \tag{1}$$

Also, for both models, mean perceived time is a power function of clock time with an exponent close to 1.0, i.e.,

$$\mu_t = t \tag{2}$$

and therefore Eqn. (1) can be rewritten as

$$\frac{\sigma_t}{t} = \gamma \tag{3}$$

The scalar property in Eqn. (3) results in distributions that superpose when the temporal axis is normalized with respect to the mean of the distribution. This superposition in relative time reflects a rescaling in time, a scale-invariant error distribution for perceived time.

2. Scalar Timing

ST has its origins in scalar expectancy theory (Gibbon 1977), and made its appearance as a timing model in Gibbon et al. (1984). The model consists of three, interrelated information processing stages—clock, memory, and decision—which are associated with timing, storage, and responding, respectively. The clock is responsible for transforming clock time into perceived time. It consists of a pacemaker that emits pulses at some mean rate, a switch that is controlled by a timing signal, and an accumulator that sums the pulses. There are two memory registers for the storage of clock information, *working memory* and *reference memory*. Working memory is loaded from the accumulator and serves as an extended buffer for temporal information from the current trial. Reference memory stores critical temporal information from past trials. The comparator determines a response on the basis of a decision rule which involves a comparison of current time stored in working memory with remembered time sampled from reference memory.

According to ST, the decision to respond R_S or R_L in temporal bisection is made by comparing the similarity of the perceived value of t with memories of the two referents, S and L. This comparison is based on a ratio of the similarity of t to S relative to the similarity of t to L. If that ratio is less than some fixed value β, the response is R_L. In his mathematical derivation of the *bisection function*, Gibbon (1981) combined the similarity ratio decision rule with two locations for the scalar variability, perception or memory. In the referent known exactly (RKE) model, he assumed no variability in the memory for the referents and placed the scalar variability in the perception of the probes. In contrast, in the stimulus known exactly (SKE) model, he assumed no variability in the perception of the probes and placed the scalar variability in the memory for the referents. There are two parameters in these models. The noise parameter γ is the coefficient of variation of time. For SKE, γ is the scalar variability in the memory representations of the S and L referents; for RKE, γ is the scalar variability in the perception of t. For both models, β is a response bias parameter. It

follows from the assumptions of the model that the SKE bisection function is approximately

$$P(R_L) = 1-\phi\left[\frac{1-\beta\left(\frac{t}{\sqrt{SL}}\right)^2}{\gamma\sqrt{\gamma^2+2}}\right] \tag{4}$$

and the RKE function is approximately

$$P(R_L) = 1-\phi\left[\sqrt{\frac{SL}{\beta\gamma^2 t^2}}-\frac{1}{\gamma}\right] \tag{5}$$

where ϕ is the cumulative normal distribution function. For both models, $T_{1/2}$ is proportional to the geometric mean (GM) of the S and L referents:

$$T_{1/2} = \sqrt{\frac{SL}{\beta}} = \frac{GM}{\sqrt{\beta}} \tag{6a}$$

and for unbiased responding ($\beta = 1$), $T_{1/2}$ is at the GM of the S and L referents:

$$T_{1/2} = \sqrt{SL} = GM \tag{6b}$$

Equation (6a) provides for variables such as ratio and probe spacing to have their effect on the location of $T_{1/2}$ through β. For example, with a large r and linear probe spacing, $T_{1/2}$ located at the GM would yield more R_L than R_S responses. To balance the frequency of the two response categories, it would be necessary to set $\beta < 1$, moving $T_{1/2}$ towards the AM. With log probe spacing, however, the two response categories would be equal with $\beta = 1$ and $T_{1/2}$ at the GM.

3. Pseudologistic Model

In addition to deriving the bisection function for the similarity ratio rule, Gibbon (1981) also derived the function for the signal detection likelihood ratio rule. For the likelihood rule, the decision to respond R_S or R_L is based on the odds ratio that the perceived value of t is generated by S or L. For this rule, Gibbon (1981) placed the scalar variability in the perception of the probes, resulting in a signal-detection model for temporal bisection. Killeen et al. (1997) simplified the derivation of the bisection function (and the resulting equation) by approximating the normal distribution with the logistic. The pseudologistic bisection function is

$$P(R_L) = \left[1+\exp\left(\frac{C-t}{0.55\gamma t}\right)\right]^{-1} \tag{7}$$

where C is the signal detection criterion. According to PLM, $T_{1/2}$ is the value of t where the S distribution crosses the L distribution. Killeen et al. (1997) showed that $T_{1/2}$ is proportional to the harmonic mean (HM) of S and L; i.e.,

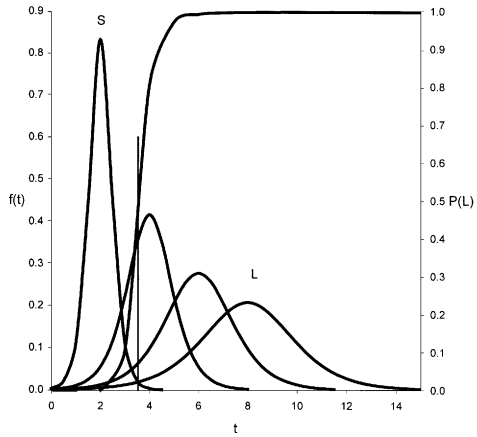

Figure 1
Schematic of the PLM model for $r = 4$ ($L = 8$ and $S = 2$). The logistic density functions [$f(t)$] are for $t = 2$, 4, 6, and 8. The psychometric function [$P(R_L)$] is for C located at $T_{1/2} = (1+k)HM$ and $\gamma = 0.15$.

$$T_{1/2} = (1+k)\frac{2}{\frac{1}{S}+\frac{1}{L}} = (1+k)HM \tag{8}$$

where

$$k = \left(\frac{\gamma\ln\left(\frac{L}{S}\right)}{2}\right) \tag{9}$$

PLM predicts that the L-to-S ratio will affect the location of $T_{1/2}$.

PLM is presented schematically in Fig. 1 for $r = 4$ ($L = 8$ and $S = 2$). The spacing of t is linear, $\gamma = 0.15$, and C is located at $(1+k)HM$. It is clear from Fig. 1 that because of the scalar property a percept located at $C = (1+k)HM$ is more likely to have been generated by L than by S. Killeen et al. (1997) suggested that in order to balance the frequency of the two response categories, C would move toward the AM with increasing r rather than being set at $(1+k)HM$.

4. Pseudologistic Model and the Behavioral Theory of Timing

Killeen et al. (1997) emphasized that PLM is a descriptive rather than a process model in that it resides close to the level of the data and does not

postulate particular underlying processes that generate the data. They do describe, however, how in their view PLM dovetails with BeT.

5. Other Models

Allan and Kristofferson (1974, p. 26) concluded that 'There are few quantitative theories of duration discrimination and few established empirical phenomena to guide theorizing.' Five years later, Allan (1979, p. 340), referring back to Allan and Kristofferson (1974), stated that 'Since that time, there has been a dramatic change. Many articles on the discrimination of brief temporal intervals, and several new quantitative models, have appeared in a period of a few years.' Now in 2000 we have an abundance of quantitative models for time perception. In this short article, I have merely provided the reader with a few gems. To obtain a broader exposure to and appreciation of the current status of quantitative models for time perception, the reader is referred to two special issues of *Behavioural Processes* in 1998 and 1999 edited by Jennifer Higa.

See also: Psychophysics

Bibliography

Allan L G 1979 The perception of time. *Perception and Psychophysics* **26**: 340–54

Allan L G 1998 The influence of the scalar timing model on human timing research. *Behavioural Processes* **44**: 101–7

Allan L G, Gibbon J 1991 Human bisection at the geometric mean. *Learning and Motivation* **22**: 39–58

Allan L G, Kristofferson A B 1974 Psychophysical theories of duration discrimination. *Perception and Psychophysics* **16**: 26–34

Church R M, Broadbent H A 1991 A connectionist model of timing. In: Commons M L, Grossberg S, Staddon J E R (eds.) *Neural Network Models of Conditioning and Action*. Erlbaum, Hillsdale, NJ, pp. 225–40

Creelman C D 1962 Human discrimination of auditory duration. *Journal of the Acoustical Society of America* **34**: 582–93

Eisler H 1975 Subjective duration and psychophysics. *Psychological Review* **82**: 429–50

Gibbon J 1977 Scalar expectancy theory and Weber's law in animal timing. *Psychological Review* **84**: 279–325

Gibbon J 1981 On the form and location of the bisection function for time. *Journal of Mathematical Psychology* **24**: 58–87

Gibbon J, Church R M, Meck W H 1984 Scalar timing in memory. In: Gibbon J, Allan L G (eds.) *Annals of the New York Academy of Sciences: Timing and Time Perception*. New York Academy of Sciences, New York, Vol. 423, pp. 52–77

Killeen P R, Fetterman J G 1988 A behavioral theory of timing. *Psychological Review* **95**: 274–95

Killeen P R, Fetterman J G, Bizo L A 1997 Time's causes. In: Bradshaw C M, Szabadi E (eds.) *Time and Behaviour: Psychological and Neurobehavioural Analyses*. Elsevier, Amsterdam, pp. 79–139

Thomas E A C, Cantor N E 1978 Interdependence between the processing of temporal and non-temporal information. In: Requin J (ed.) *Attention and Performance VII*. Erlbaum, Hillsdale, NJ, pp. 43–62

Treisman M 1963 Temporal discriminations and the indifference interval: implications for a model of the internal clock. *Psychological Monographs* **576**: 1–31

Zakay D, Block R A 1994 An attentional-gate model of prospective time estimation. In: Richelle M, DeKeyser V, d'Ydewalle G, Vandierendonck A (eds.) *Time and the Dynamic Control of Behavior*. Université de Liège, Liège, Belgium, pp. 167–78

L. G. Allan

Time Series: Advanced Methods

1. Introduction

Many social phenomena can be described, at least approximately, by a discrete-time dynamical system, which can be expressed as a difference equation,

$$Y_t = F(X_{t-1}) \tag{1}$$

where $\{Y_t\}$ is a sequence of m dimensional observations and X_t is a d' dimensional state vector, with $t = 1, 2, \ldots$ and $1 \leqslant m' \leqslant m$. In many applications, X_t stands for the past values of Y_t. Depending on the functional form $F(\cdot)$, the difference Eqn. 1 can exhibit a variety of dynamical features such as limit cycles, time irreversibility, and periodic changes (Tong 1983, 1990). However, in the real world, stochastic noise is often present and therefore it is more realistic to consider a dynamical system with noise such as:

$$Y_t = F(X_{t-1}, a_t; \theta) \tag{2}$$

where $\{a_t\}$ is often assumed to be a sequence of independent identically distributed (*i.i.d*) (possibly vector valued) random variables and θ is a vector of parameters. For instance, if $Y_t = X_t$ with $m = 1$ and $F(\cdot)$ is a scalar valued function, say, $f215(X_{t-1}; \theta) + a_t$, then (2) takes the form:

$$X_t = f(X_{t-1}; \theta) + a_t \tag{3}$$

which is a non-linear autoregressive process of order one. Clearly Eqn. 3 can be generalized to include more past values of X_t and then we have

$$X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-p}; \theta) + a_t \tag{4}$$

which is a general nonlinear autoregressive process of order p. An important special case of Eqn. 4 is when $f(\cdot)$ is linear in the past values of X_t and we have then

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t \tag{5}$$

which is the usual linear autoregressive model of order

p. Depending on $f(\cdot)$ or more generally $F(\cdot)$, the number of nonlinear models is clearly unlimited and the choice of a model is typically dictated by the situation and the ability of the model to describe/predict a social phenomenon.

It is clear that the dynamical system formulations Eqn. 2 and Eqn. 4 can be extended by considering a specification of the variance of the noise term a_t. Thus,

$$X_t = f(X_{t-1}; \theta) + g(X_{t-1}; \theta)a_t \tag{6}$$

where $g(\cdot)$ is positive. Now, X_t is conditionally heteroscedastic. In other words, model Eqn. 6 now consists of both specifications for the mean (the drift) and for the conditional variance (the diffusion). Below we consider some advanced models in time series which have been developed and found to be useful since around 1980. For simplicity, we let $Y_t = X_t$ and a_t be a sequence of *i.i.d.* scalar random variables.

2. Nonlinear Time Series for the Mean

2.1 The Class of Threshold Models

Proposed by Tong in the later 1970s, the threshold models are a natural generalization of the linear autoregression Eqn. 5 by allowing different linear autoregressive specification over different parts of the state space. For example, a two-regime threshold autoregressive model of order p_1 and p_2 may be defined as follows,

$$X_t = \begin{cases} \phi_0^{(1)} + \phi_1^{(1)} X_{t-1} + \cdots + \phi_{p_1}^{(1)} X_{t-p_1} + a_t^{(1)}, & \text{if } X_{t-d} \leqslant c \\ \phi_0^{(2)} + \phi_1^{(2)} X_{t-1} + \cdots + \phi_{p_2}^{(1)} X_{t-p_2} + a_t^{(2)}, & \text{otherwise} \end{cases} \tag{7}$$

where $1 \leqslant d \leqslant \max(p_1, p_2)$, $\{a_t^{(i)}\}$ are two *i.i.d.* noise sequences with mean zero and variance σ_i^2, $i = 1, 2$, $\{a_t^{(1)}\}$ and $\{a_t^{(2)}\}$ are also independent of each other. We call c the threshold parameter and d the delay parameter. Eqn. 7 is called a self-exciting threshold autoregressive (SETAR (2; p_1, p_2)) model. Generalizations to more than two regimes are immediate. Stationarity and ergodicity conditions for Eqn. 7 when $p_1 = p_2 = 1$ and $\phi_0^{(i)} = 0$, $i = 1, 2$ have been obtained while a sufficient condition for the general SETAR (2; p, p) model is available (Tong 1990).

Estimation of Eqn. 7 can be easily done using the conditional least squares method given the parameters p_1, p_2, c, and d. Identification of p_1, p_2, c, and d can be done by the minimum Akaike information criterion (AIC) (Tong 1990). For finite samples the corrected AIC or AICC is recommended (Wong and Li 1998). The computer programme STAR 3 accompanying Tong (1990) provides a comprehensive set of modeling tools for threshold models. Diagnostic checking for model adequacy can be done using residual autocorrelations. The appropriate asymptotic distribution was derived in Li (1992). Tsay (1989) suggested an approach in the detection and modeling of threshold structures which is based on explicitly rearranging the least squares estimating equations using the order statistics of X_t, $t = 1,\ldots, n$, where n is the length of realization. (See Tong 1990 for references.) A similar rearrangement was incorporated in the software STAR 3. Chen and Tsay (1993) considered a functional-coefficient autoregression model which has a very general threshold structure.

An easy-to-use statistic for detecting departure from linearity is the port-manteau test based on squared residual autocorrelations, the residuals being obtained from an appropriate linear autoregressive moving-average model fitted to the data (McLeod and Li 1983). Of course, a general test statistic may not be optimal in terms of power when specific alternative hypotheses are considered. Tong (1990) has described other tests for nonlinearity due to Davies and Petruccelli, Keenan, Tsay and Saikkonen and Luukkonen, Chan and Tong. Petruccelli (1990) considered a comparison for some of these tests.

As a textbook-like example (albeit outside the social sciences), we consider the annual Canadian lynx trapping data in the MacKenzie River for the period 1821–1934. The nonlinearity of the data has been extensively documented by Tong (1990). Threshold nonlinearity was confirmed by applying the likelihood ratio test of Chan and Tong (1990) at the 1 percent level. Following other authors we transform the data by taking common log. We use the AICC as a criterion in selecting the best SETAR (2; p_1, p_2) model. The maximum possible value for p_1 and p_2 is 10, and the maximum possible value for the delay parameter d is 6. It is recommended that possible candidates of the threshold parameter can be chosen from a subset of the order statistics of the data. Following Wong (1998) we use 2.4378, 2.6074, 2.7769, 2.9464, 3.1160, 3.2855, and 3.4550, as potential values of the threshold parameter. The best fitting model using the minimum AICC criterion is the following SETAR (2; 4, 2) model,

$$\begin{aligned} X_t &= 1.0303 + 0.8959X_{t-1} - 0.0313X_{t-2} + 0.0406X_{t-3} \\ &\quad - 0.2001X_{t-4} + a_t^{(1)} \\ &\quad \text{if } X_{t-3} \leqslant 2.9464, \\ X_t &= 1.6092 + 1.3487X_{t-1} - 0.8937X_{t-2} + a_t^{(2)}, \\ &\quad \text{if } X_{t-3} > 2.9464 \end{aligned}$$

where $a_t^{(1)}$ and $a_t^{(2)}$ have estimated variance equal to 0.0164 and 0.0642, respectively. The residual autocorrelation and squared residual autocorrelation show no significant values suggesting that the above model is adequate.

Instead of adrupt jumps between regimes in Eqn. 7 a smooth transition threshold autoregression was proposed by Chan and Tong (1986). Teräsvirta (1994)

considered some further work in this direction. Continuous time threshold model was considered by Tong and Yeung (1991) with applications to water pollution data. Brockwell (1994) and others considered further work in the continuous time. Multivariate (mainly bivariate) threshold models were included in the seminal work of Tong in the 1980s and further developed by Tsay (1998). The relation between chaos and nonlinear time series is also treated in some detail in Tong (1990).

2.2 The Class of Bilinear Time Series Models

Bilinear models featured for a while in the early days of nonlinear time series modeling. In its simplest form, a single term bilinear model is given by

$$X_t = \beta X_{t-k} a_{t-l} + a_t \tag{8}$$

where $k, l \geqslant 1$. Note that Eqn. 8 is linear in $X_{t-k}(a_{t-l})$ given $a_{t-1}(X_{t-k})$.

Granger and Andersen (1978) provided the first comprehensive introduction to these models. By deriving the second order autocovariances of Eqn. 8, they demonstrated that if $l < k$ then the time series cannot be distinguished from an uncorrelated noise process. The general bilinear autoregressive moving average model of order (p, q, P, Q) has the form

$$X_t = \sum_{j=1}^{p} \alpha_j X_{t-j} + \sum_{i=0}^{q} b_i a_{t-1} + \sum_{k=1}^{P} \sum_{l=1}^{Q} \beta_{kl} X_{t-k} a_{t-l}. \tag{9}$$

Granger and Andersen (1978, p. 31) denoted this by BARM (p, q, P, Q) and Subba Rao (1981) denoted this by $BL(p, q, P, Q)$. Subba Rao (1981) and Subba Rao and Gabr (1984) gave a general theory for model (9). Further results on causality and stationarity were given by Liu and Brockwell (1988). Identification (in the sense of estimation) of model (9) is in general difficult because of the enormous number of possibilities.

3. Models with Conditional Heteroscedasticity

One very popular class of such models is the autoregressive conditional heteroscedastic (ARCH) models of Engle (1982). The ARCH model of order one (ARCH(1)) is given by

$$X_t = \sqrt{h_t} a_t \tag{10}$$

and $\sqrt{h_t} = g(X_{t-1}; \theta) = (\alpha_0 + \alpha_1 X_{t-1}^2)^{\frac{1}{2}}$. Here $\alpha_0 > 0$, $\alpha_1 \geqslant 0$ and a_t is again *i.i.d.* with mean zero and unit variance. It can be shown that X_t is covariance stationary if $0 \leqslant \alpha_1 < 1$. The unconditional variance of (6) is then $\alpha_0/(1-\alpha_1)$. The fourth moment of (6) exists if $\alpha_1^2 < \frac{1}{3}$. It is then greater than 3. For $r > 1$, the

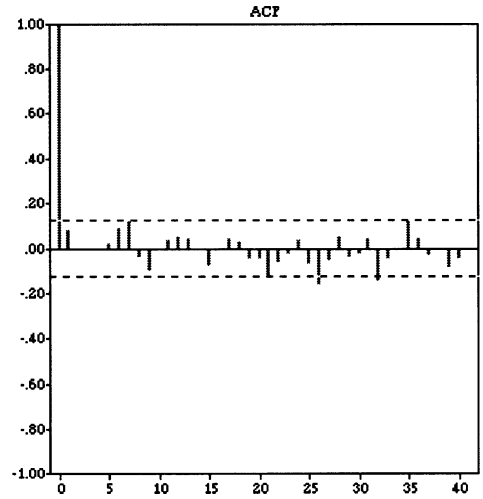

Figure 1
Sample autocorrelations of R_t

ARCH(r) model is defined by including more past values of X_t^2 so that $h_t \equiv E(X_t^2 \,|\, \text{past information}) = \alpha_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_r X_{t-r}^2 (\alpha_0 > 0, \alpha_i \geqslant 0 \quad \text{for} \quad i > 0)$. Again, X_t is covariance stationary if $\alpha_1 + \alpha_2 + \cdots + \alpha_r < 1$. It can be shown that the autocorrelation function of X_t^2 behaves like that of an AR(r) process. Milhøj (1985) derived the condition for the existence of the $2m$-th moments of X_t. Estimation of model 10 can easily be done by assuming a_t to be standard normal so that X_t is conditionally normal with mean zero and variance h_t. The conditional log likelihood l is then given by

$$l = \sum_{t=1}^{n} l_t = \sum_{t=1}^{n} -\tfrac{1}{2} \log h_t - \frac{1}{2} \frac{X_t^2}{h_t}. \tag{11}$$

Maximum likelihood (ML) estimation can be performed using Newton–Raphson or a method that does not require explicit derivatives. Assuming finite 4-th moments, Weiss (1986) showed that the ML estimates are consistent and asymptotically normally disturbed. Many commercial statistical softwares such as the SAS now provide capabilities for ARCH modeling.

Diagnostic checking of the fit of ARCH models using standardized squared residual autocorrelations was considered in Li and Mak (1994). A fast method of detecting ARCH is by means of the portmanteau test of McLeod and Li (1983) using residuals from an autoregressive moving average model. As an illustration, we consider the daily closing price of the Hong Kong Hang Seng Index in the year 1980. The data are transformed by taking natural logarithm and differenced once to achieve stationarity. The transformed data R_t are called the returns of the index. There are 245 observations. The sample autocorrelations of R_t in Fig. 1 show basically no noticeable dependence struc-

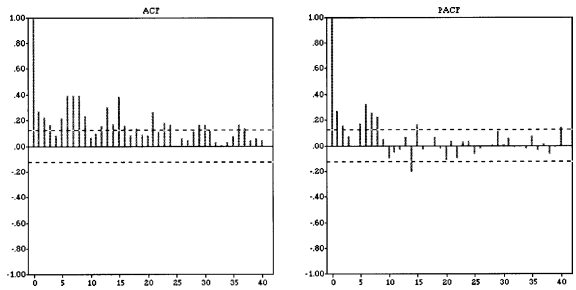

Figure 2
Sample autocorrelations of R_t^2

ture in R_t itself. However, the sample autocorrelations of R_t^2 in Fig. 2 show a very strong dependence structure for R_t^2 suggesting the presence of an ARCH structure. Based on the partial autocorrelations of R_t^2 we consider an ARCH model of order 7 for R_t. Using the method of maximum likelihood we have the following result:

$$R_t = \sqrt{h_t} a_t$$

with

$$h_t = 0.00012 + 0.03997 R_{t-1}^2 + 0.13506 R_{t-2}^2 + 0.12798 R_{t-3}^2 + 0.15475 R_{t-6}^2 + 0.28445 R_{t-7}^2.$$

As reported in Li and Mak (1994) the sample autocorrelations of R_t^2 standardized by h_t show no further significant values and the above ARCH model is thus considered adequate.

Bollerslev (1986) generalized the ARCH(r) model to the GARCH(r, s) model, where

$$X_t = \sqrt{h_t} a_t$$

and

$$h_t = \alpha_0 + \sum_{i=1}^{r} \alpha_i X_{t-i}^2 + \sum_{i=1}^{s} \beta_i h_{t-i} \tag{12}$$

with $\alpha_0 > 0$, $\alpha_i \geqslant 0$, $\beta_i \geqslant 0$.

The condition for covariance stationarity is $\sum_{i=1}^{r} \alpha_i + \sum_{i=1}^{s} \beta_i < 1$. Estimation of model 12 is similar to that of the ARCH models and exogeneous variables can be easily included. The second order moment properties of X_t^2 are like those of an ARMA(p, s) where $p = \max(r, s)$. The GARCH models have been very popular in the modeling of volatility in the finance literature. Bollerslev et al. (1992, 1994) provided two extensive reviews of the GARCH literature. Gourieroux (1997) and Bera and Higgins (1995) gave a more advanced coverage including multivariate ARCH models. The work edited by Engle (1995) has included many of the important contributions to the ARCH literature.

4. *Models with Long Memory*

Many stationary dynamical systems under model 4 in fact lead to $\{X_t\}$ which are short memory in the sense that the moduli of the autocovariances $C(k) = \mathrm{cov}(X_t, X_{t-k})$, $k = 1, 2, \ldots$ sum to a finite value. However, in some applications it is possible that $\sum_{k=0}^{\infty} |C(k)| = \infty$. A time series with this property is said to have the long memory property. One approach to modeling the long memory property is by means of fractional differencing. Suppose that B is the backshift operator, $BX_t = X_{t-1}$. Then a long memory dynamical system can be constructed by considering the process satisfying (Granger and Joyeux 1980),

$$(1-B)^d X_t = a_t \tag{13}$$

where d is a non-integer and

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k B^k.$$

The process is stationary and invertible if $-1/2 < d < 1/2$ and possesses the long memory property if $0 \leqslant d < 1/2$. The fractionally integrated autoregressive moving average (FARIMA) process of order (p, d, q) satisfies (Hosking (1981), and Granger and Joyeux (1980),

$$\phi(B)(1-B)^d X_t = \theta(B) a_t \tag{14}$$

where

$$\phi(B) = 1 - \phi_1 B \cdots - \phi_p B^p,\ \theta(B) = 1 - \theta_1 B \cdots - \theta_q B^q$$

and d is as in Eqn. 13. It is further assumed that ϕ(B) and θ(B) have all roots outside the unit circle, which implies that X_t is stationary and invertible. Hosking (1981) showed that the autcovariances $C(k)$ of X_t are of the order $|k|^{2d-1}$ so that unlike ARMA process, $C(k)$ decays hyperbolically rather than exponentially. The spectral density $f(\lambda)$ of the process can be written in the form

$$f(\lambda) = \frac{\sigma_a^2}{2\pi} |1 - e^{i\lambda}|^{-2d} \cdot \frac{|\theta(e^{i\lambda})|^2}{|\phi(e^{i\lambda})|^2}, \tag{15}$$

where σ_a^2 is the variance of a_t. Beran (1994) contained a detailed exposition of the subject while Brockwell and Davis (1991) gave a good introduction including least squares and maximum likelihood estimation.

5. *Second Generation Models*

Tong (1990) suggested that many of the aforementioned models can be combined to form new nonlinear models. He called these second generation

models. For example, Tong (1990) suggested the possibility of a threshold AR model with ARCH. His idea was applied to real data by Li and Lam (1995). Li and Li (1996) extended this further to the double threshold ARCH or DTARCH models. These models allow for a threshold structure in both the conditional mean and the conditional variance. Clearly the number of possibilities is unlimited and one can consider models like double threshold long memory models with GARCH.

6. Discrete Valued Time Series

In many applications, the time series under consideration may be integer valued or may possess a discrete marginal distribution. An obvious approach to generate discrete random variables is by means of a possibly higher order Markov chain. Such is the approach of Raftery (1985) and Pegram (1980). A more elegant approach is that of McKenzie (1986) which is based on the method of thinning. Böckenholt (1999) considered a mixed integer-valued autoregressive Poisson regression with an application to category purchase data. Discrete-valued time series may also be generated using a conditional generalized linear model as in Zeger and Qaqish (1988). Li (1991) considered the goodness of fit problem for this approach. The recent work of MacDonald and Zucchini (1997) gave a valuable and timely survey of this vast area.

7. Conclusion

There is certainly a myriad of non-linear models to consider when one decides to leave the restriction of the linear domain. What models to consider clearly depends on the goal and nature of the investigation. In many cases, flexibility in describing a variety of features such as time irreversibility, limit cycles, and autocorrelation structure is an important criterion. The availability of handy softwares is another important factor as illustrated by the examples considered. In many cases, automatic model selection criteria such as AIC and BIC, appropriately used, help to reduce the number of choices to a manageable few and so do model checking methods. The final test of a model in time series, in many applications is, of course, its ability to predict the future accurately.

Bibliography

Bera A K, Higgins M L 1995 A survey of ARCH models: Properties, estimation and testing. In: Oxley L, George D A R, Roberts C J, Sayer S (eds.) *Surveys in Econometrics*. Blackwell, Oxford, UK, pp. 215–72

Beran J 1994 *Statistics for Long-Memory Processes*. Chapman and Hall, London

Böckenholt U 1999 Mixed INAR(1) Poisson regression models: Analyzing heterogeneity and serial dependencies in longitudinal count data. *Journal of Econometrics* **89**: 317–38

Bollerslev T 1986 Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**: 307–27

Bollerslev T, Chou R Y, Kroner K F 1992 ARCH modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics* **52**: 5–59

Bollerslev T, Engle R F, Nelson D B 1994 ARCH models. In: Engle R F, McFadden D L (eds.) *Handbook of Econometrics 4*. North-Holland, Amsterdam

Brockwell P J 1994 On continuous time threshold ARMA processes. *Journal of Statistical Planning and Inference* **39**: 291–304

Brockwell P J, Davis R A 1991 *Time Series Theory and Methods*, 2nd edn. Springer-Verlag, New York

Chan K S, Tong H 1986 On estimating thresholds in autoregressive models. *Journal of Time Series Analysis* **7**: 178–90

Chan K S, Tong H 1990 On likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society* **52**: 469–76

Chen R, Tsay S 1993 Functional-coefficient autoregressive models. *Journal of the American Statistical Association* **88**: 298–308

Engle R F 1982 Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrika* **50**: 987–1008

Engle R F 1995 *ARCH: Selected Readings*. Oxford University Press, Oxford, UK

Gourieroux C 1997 *ARCH: Models and Financial Applications*. Springer, New York

Granger C W J, Andersen A P 1978 *An Introduction to Bilinear Time Series Models*. Vandenhoeck and Ruprecht, Göttingen, Germany

Granger C W J, Joyeux R 1980 An introduction to long-range time series models and fractional differencing. *Journal of Time Series Analysis* **1**: 15–30

Hosking J R M 1981 Fractional differencing. *Biometrika* **68**: 165–76

Li C W, Li W K 1996 On a double threshold autoregressive heteroscedastic time series model. *Journal of Applied Econometrics* **11**: 253–74

Li W K 1991 Testing model adequacy for some Markov regression models for time series. *Biometrika* **78**: 83–9

Li W K 1992 On the asymptotic standard errors of residual autocorrelations in nonlinear time series modeling. *Biometrika* **79**: 435–7

Li W K, Lam K 1995 Modeling asymmetry in stock returns by a threshold autoregressive conditional heteroscedastic model. *Journal of Royal Statistical Society D*. **44**: 333–41

Li W K, Mak T K 1994 On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. *Journal of Times Series Analysis* **15**: 627–36

Liu J, Brockwell P J 1988 On the general bilinear time series model. *Journal of Applied Probability* **25**: 553–64

MacDonald I L, Zucchini W 1997 *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman and Hall, London

McKenzie E 1986 Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Advances in Applied Probability* **18**: 679–705

McLeod A I, Li W K 1983 Diagnostic checking ARMA time series models using squared residual autocorrelations. *Journal of Time Series Analysis* **4**: 269–73
Milhøj A 1985 The moment structure of ARCH processes. *Scandinavian Journal of Statistics* **12**: 281–92
Pegram G G S 1980 An autoregressive model for multilag Markov chains. *Journal of Applied Probability* **17**: 360–2
Petruccelli J D 1990 A comparison of tests for SETAR-type nonlinearity in time series. *Journal of Forecasting* **9**: 25–36
Raftery A Z 1985 A model for high-order Markov chains. *Journal of Royal Statistical Society B*. **47**: 528–39
Subba Rao T 1981 On the theory of bilinear time series models. *Journal of Royal Statistical Society B*. **43**: 244–55
Subba Rao T, Gabr M M 1984 *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Lecture Notes in Statistics 24. Springer, New York
Teräsvirta T 1994 Specification, estimation and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* **89**: 208–18
Tong H 1983 *Threshold Models in Non-linear Time Series Analysis*. Springer Lecture Notes in Statistics 21. Springer-Verlag, New York
Tong H 1990 *Non-linear Time Series: A Dynamical System Approach*. Clarendon Press, Oxford, UK
Tong H, Yeung I 1991 Threshold autoregressive modeling in continuous time. *Statistica Sinica* **1**: 411–30
Tsay R S 1989 Testing and modeling threshold autoregressive processes. *Journal of American Statistical Association* **84**: 231–40
Tsay R S 1998 Testing and modeling multivariate threshold models. *Journal of the American Statistical Association* **93**: 1188–202
Weiss A A 1986 Asymptotic theory for ARCH models: Estimation and testing. *Econometric Theory* **2**: 107–31
Wong C S 1998 *Statistical inference for some nonlinear time series model*. Ph.D. thesis. University of Hong Kong, Hong Kong
Wong C S, Li W K 1998 A note on the corrected Akaike information criterion for the threshold autoregressive models. *Journal of Time Series Analysis* **19**: 113–24
Zeger S L, Qaqish B 1988 Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44**: 1019–31

W. K. Li and H. Tong

Time Series: ARIMA Methods

1. Objectives of Time Series Analysis

Data in business, economics, engineering, environment, medicine, and other fields of scientific investigations are often collected in regular time intervals. Such data are called time series. Examples are hourly temperature readings, daily stock prices, weekly traffic volume, monthly money supply, quarterly GNP (gross national product), annual growth rates, etc. Time series data tend to exhibit patterns such as trends, seasonal fluctuations, irregular cycles, and occasional shifts in level or variability. The data are, therefore, dynamically or serially correlated.

In practice, the main objectives of analyzing time series are: (a) to ascertain and extrapolate the dynamic pattern in the data for forecasting future observations, (b) to assess the effect of exogenous interventions known to have occurred at given time points, and (c) to detect unknown and unsuspected interventions. For example, one may use the available weekly sales of frozen orange juice over the last three years to forecast the sales for the next six weeks and to assess the impact of a major promotion campaign initiated four weeks ago. As another example, an air pollution data analyst may discover a sudden drop in the time series of monthly ambient carbon monoxide concentration level and find out that the drop can be attributed to an unsuspected change in the measuring instrument. We here sketch some useful methods for achieving these objectives.

2. Stochastic Structure of Time Series Observations

Let $\ldots, t-1, t, t+1, \ldots$ denote equally spaced time points and y_t the observation at time t. The probability model generating the series $\{y_t\}$ is called a *stochastic process*, and the observed time series data are then a realization of the process. It is important to note a key distinction between time series observations and observations generated from designed experiments. For time series, there is usually one and only one realization at a given time point t, while because experiments can be repeated, there can be many realizations under the 'same' experimental conditions. Thus, statistical inference on time series depends critically on the nature of the serial dependence of the stochastic process under study.

3. Stationary vs. Nonstationary Processes

When the joint probability distribution of any k observations $y_{t+1}, \ldots, y_{t+k}$ from a stochastic process remains the same for different values of t, the process is said to be 'stationary.' In practice, this means that there is a state of equilibrium in which the overall behavior of the time series stays roughly the same over time. When this condition does not hold, the series is said to be nonstationary. The theory of stationary processes has been well developed and plays an important role in time series analysis. Although many of the time series encountered in practice exhibit a nonstationary behavior, it is frequently the case that a suitable transformation such as differencing the data will render them stationary so that the theory of stationary processes can be applied.

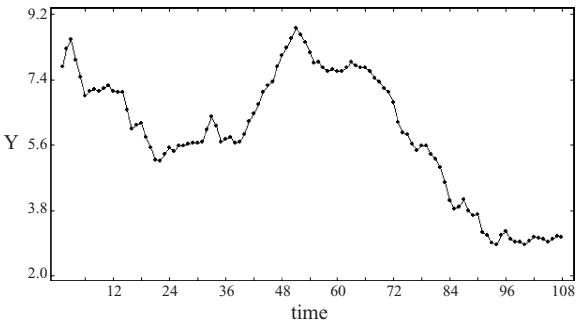

Figure 1
Monthly interest rates of 90-day Treasury bills, 1985–93

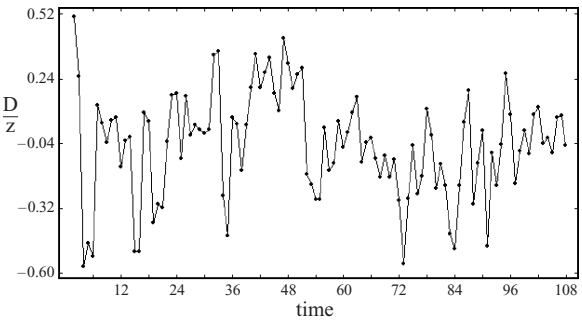

Figure 2
Month to month changes of the interest rates

To illustrate, Fig. 1 shows the time series of monthly interest rates of 90-day Treasury bills from January 1985 to December 1993, and Fig. 2 the series of month to month changes of the rates (first differences). The interest rate series exhibits a drifting or wandering behavior without a stable level. On the other hand, the series of the first differences seems to fluctuate about a fixed mean level with constant variance over the observational period. This example shows that a drifting nonstationary series can be transformed into a stationary one by the differencing operation. In fact, financial time series such as stock prices, prices of derivatives, and exchange rates often behave in this manner.

4. *Linear Dynamic Models*

The serial dependence in the time series $\{y_t\}$ can often be adequately approximated by a linear model of the form

$$y_t = c + \pi_1 y_{t-1} + \pi_2 y_{t-2} + \ldots + \pi_j y_{t-j} + \ldots + a_t \quad (1)$$

where c is a constant, $\pi_1, \pi_2, \ldots$ are parameters relating y_t linearly to its past values, and a_t is the innovation or shock hitting the system at time t. It is commonly assumed that $\{a_t\}$ is a series of identically and independently distributed normal random variables with zero mean and common variance σ^2. The linear dynamic model in Eqn. (1) can also be expressed in the alternative form

$$y_t = c^* + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} \ldots + \psi_j a_{t-j} + \ldots \quad (2)$$

where c^* is a constant, and the two parameter sets $(\pi_1, \pi_2, \ldots)$ and $(\psi_1, \psi_2, \ldots)$ are functionally related. In Eqn. (2) ψ_j measures the effect of the shock a_{t-j} hitting the system j periods ago on the observation y_t. The ψ_j's are called impulse responses or the memory of the system. The two representations in Eqns. (1) and (2), although equivalent, highlight two distinctive features of the linear system.

For forecasting future observations at a given time origin T, let us consider first the next observation y_{T+1}. By putting $t = T+1$ in Eqn. (1), it is easy to show that the best (minimum mean squared error) predictor of y_{T+1} is a linear combination of the available observations $(y_T, y_{T-1}, \ldots)$, $\hat{y}_{T+1} = c + \pi_1 y_T + \pi_2 y_{T-1} + \ldots$, and the associated forecast error is simply $y_{T+1} - \hat{y}_{T+1} = a_{T+1}$ having variance σ^2. It is straightforward to deduce from Eqn. (1) best predictors of observations y_{T+l}, $\ell = 2, 3, \ldots$ periods ahead as linear functions of $(y_T, y_{T-1}, \ldots)$ and their associated forecast errors as linear combinations of future shocks $(a_{T+1}, \ldots, a_{T+\ell})$.

5. *ARMA and ARIMA Models*

In reality, the model in Eqns. (1) or (2) is not known and has to be built from available data. Since y_t can in principle relate to its remote past, or to the remote past of a_t, for parsimony it is useful to make the π's or ψ's functions of a small set of parameters which can then be accurately estimated from the data. A class of such models originally proposed by Yule and Slutsky in the 1920s is the (mixed) autoregressive moving average (ARMA) models of orders (P, Q),

$$y_t - \phi_1 y_{t-1} - \ldots - \phi_P y_{t-P} = c + a_t - \theta_1 a_{t-1} - \ldots - \theta_Q a_{t-Q} \quad (3)$$

This class of models has been found useful to represent a variety of stationary, nonstationary, and seasonal time series in diverse scientific fields. By straightforward substitutions, one can readily deduce the π's and ψ's in Eqns. (1) and (2) as functions of the finite parameter set $(\phi_1, \ldots, \phi_P; \theta_1, \ldots, \theta_Q)$ in Eqn. (3).

To illustrate, consider first the case $(P = 0, Q = 1)$ so that $y_t = c + a_t - \theta_1 a_{t-1}$, which is the first-order moving average model. This model says that when the shock a_t hits the system at time t it will affect not only the observation y_t but also the next observation y_{t+1} (since $y_{t+1} = c + a_{t+1} - \theta_1 a_t$). Thus, the effect of the shock a_t will last over the two periods t and $t+1$ and

dissipate afterward. This short memory stationary model has been found useful to represent time series of various financial returns.

Consider next the case ($P = 1, Q = 0$) so that the model is $y_t = c + \phi_1 y_{t-1} + a_t$, which is the first-order autoregressive model frequently discussed in the economics literature. This model is stationary when $|\phi_1| < 1$. It can be expressed in the alternative memory form $y_t = c^* + a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \mathrm{L}$ showing that the effect or memory of the shock a_t will persist over many periods but will gradually dissipate when the model is stationary. In this case, the l-step ahead forecast will converge to the mean level of the series as the horizon l increases. For ($c = 0, \phi_1 = 1$), this model is a nonstationary 'random walk' widely believed to be appropriate for stock prices in an 'efficient market.'

As a final illustration, for ($c = 0, P = Q = 1, \phi_1 = 1$) we have the model $y_t - y_{t-1} = a_t - \theta_1 a_{t-1}$. For this model, the left-hand side is the first difference of y_t implying that the series $\{y_t\}$ is nonstationary but the first difference series is stationary and follows the first-order moving average model. At the time origin T, the l-step ahead predictor is, for all $\ell \geqslant 1$, $\hat{y}_{T+\ell} = (1-\theta_1)y_T + \theta_1(1-\theta_1)y_{T-1} + \theta_1^2(1-\theta_1)y_{T-2} + \ldots$, which is an exponentially weighted average of current and past observations. Predictors of this kind which give more weight to recent information and less and less weight to information from the remote past seem intuitively appealing because time series data, especially in business and economics, tend to exhibit nonstationary behavior over time. As a result, the 'exponential smoothing' forecast discussed here has been the most widely applied forecasting technique in practice. It will be seen later that the Treasury bills data shown in Fig. 1 can be adequately represented by this model.

To sharpen the distinction between stationary and nonstationary models, we can express the ARMA (P, Q) models in Eqn. (3) alternatively as

$$\begin{aligned}\nabla^d y_t - \varphi_1 \nabla^d y_{t-1} - \ldots - \varphi_p \nabla^d y_{t-p} \\ = c + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}\end{aligned} \tag{4}$$

where $d = 0, 1, 2, \ldots$, $\nabla^d y_t$ is the dth difference of y_t (i.e. $\nabla^d y_t = y_t$ if $d = 0$, $\nabla^d y_t = y_t - y_{t-1}$ if $d = 1$, $\nabla^d y_t = y_t - 2y_{t-1} + y_{t-2}$ if $d = 2$, etc.), $P = p + d$, $Q = q$, the two parameter sets ($\phi_1, \ldots, \phi_P$) and ($\varphi_1, \ldots, \varphi_p$) are functionally related, and we shall assume that the ARMA model for $\nabla^d y_t$ is stationary. This is saying that the time series $\{y_t\}$ is such that its dth difference $\nabla^d y_t$ is stationary and follows a mixed autoregressive moving average model. For $d = 0$, y_t itself is stationary; and for $d > 0$, y_t is nonstationary. The models in Eqn. (4) are called autoregressive integrated moving average (ARIMA) models of orders (p, d, q) for the time series y_t. Here d signifies the order of differencing to make the transformed series stationary, p is the number of lagged values of $\nabla^d y_t$, and q is the number of lagged terms of a_t in the model.

It should be noted that for the ARIMA models in Eqn. (4), the nonstationarity of the series y_t is solely accounted for by the difference operation $\nabla^d y_t$. While this is not the only kind of nonstationarity in the ARMA models in Eqn. (3), we shall, however, focus primarily on this class of models.

6. *Model Building*

To build an ARMA or ARIMA model for the data at hand, Box and Jenkins (1976) have proposed an iterative approach consisting of (a) tentative model specification, (b) efficient estimation, and (c) diagnostic checking. This approach has been widely adopted and in fact has revolutionized the use of time series models for forecasting in practice.

In the tentative specification phase, the goal is to employ computationally simple techniques to narrow down the range of parsimonious models worthy of more detailed further study. Techniques such as the sample autocorrelation function (SACF) and sample partial autocorrelation function (SPACF) have been widely used. The SACF can help determine the order of differencing needed to make the transformed series stationary and pin down the order of a moving average model for stationary series. The SPACF, on the other hand, is useful to help decide the order of a stationary autoregressive model. The more recently proposed sample extended autocorrelation function (SEACF, see Tsay and Tiao (1984)) makes it possible to specify directly the orders of an ARMA (P, Q) model, stationary or nonstationary, and hence simplifies the procedure for model specification. In this case, the question of stationarity will be answered at the estimation stage.

Once a model is tentatively entertained, the parameters can then be efficiently estimated by the method of maximum likelihood. In addition to standard tests on the statistical significance of parameter estimates, unit root tests may provide a useful tool here to help determine the order of nonstationarity in the series if needed. Finally, residuals calculated from the fitted model are then subject to various diagnostic checks to determine the overall adequacy of the entertained model and to search for directions of improvement. Of particular importance are procedures to detect the existence of autocorrelations in the residuals, outliers and structural changes such as level shifts, variance changes, and nonlinearities.

To illustrate this model building approach, we consider the monthly interest rate data introduced earlier. Table 1 shows the SACF of the series y_t, the SACF and SPACF of the differenced series ∇y_t, and the SACF of the residuals $\hat{a}_t$ for lags 1–12.

In each of the four panels in Table 1, the first row gives the sample correlation values and the second their associated estimated standard errors. We now discuss tentative model specification. First, the per-

Table 1
SACF and SPACF for the interest rate data

1.	SACF of y_t												
	Lags 1–12	0.97	0.94	0.89	0.85	0.81	0.77	0.73	0.68	0.64	0.59	0.53	0.48
	ST.E.	0.10	0.16	0.21	0.24	0.27	0.29	0.31	0.32	0.34	0.35	0.36	0.36
2.	SACF of ∇y_t												
	Lags 1–12	0.46	0.06	0.12	0.12	0.07	0.09	0.07	0.09	0.13	0.08	0.05	−0.02
	ST.E.	0.10	0.11	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
3.	SPACF of ∇y_t												
	Lags 1–12	0.46	−0.18	0.22	−0.05	0.07	0.03	0.01	0.08	0.06	−0.02	0.03	−0.11
	ST.E.	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	SACF of $\hat{a}_t$												
	Lags 1–12	0.00	0.03	0.07	0.09	−0.00	0.08	0.03	0.03	0.11	0.02	0.03	0.02
	ST.E.	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10

Table 2
Simplified SEACF for the interest rate data

$(Q \rightarrow)$	0	1	2	3	4	5	6	7	8	9	10	11	12
$(P=0)$	X	X	X	X	X	X	X	X	X	X	X	O	O
$(P=1)$	X	O	O	O	O	O	O	O	O	O	O	O	O
$(P=2)$	X	X	O	O	O	O	O	O	O	O	O	O	O
$(P=3)$	X	X	O	O	O	O	O	O	O	O	O	O	O
$(P=4)$	O	X	O	O	O	O	O	O	O	O	O	O	O
$(P=5)$	X	O	X	O	O	O	O	O	O	O	O	O	O
$(P=6)$	X	O	O	O	O	O	O	O	O	O	O	O	O

Table 3
Estimation results for the interest rate data

	$\hat{\phi}_1$	$\hat{c}$	$\hat{\theta}_1$	$\hat{\sigma}$
Model (5)			−0.597 (0.077)	0.198
Model (6)	0.993 (0.017)	−0.002 (0.107)	−0.592 (0.078)	0.196

sistently high values of the SACF of y_t over long lags suggest that the series is nonstationary, leading to considering the first difference of the series. Second, the SACF of ∇y_t has only one large value at lag 1, and this 'cutting off' property strongly indicates that a simple first-order moving average model would be adequate for the differenced series. On the other hand, the pattern of the SPACF of ∇y_t, showing two large values at lag 1 and lag 3, would suggest a more complex model. Consequently we are led to consider the ARIMA (0, 1, 1) model

$$\nabla y_t = a_t - \theta_1 a_{t-1} \tag{5}$$

The above mode of model specification, while widely adopted in practice, has several drawbacks. First, judgment of whether to difference the data based on 'persistently high values of SACF' can often be difficult in practice. Second, for $d \geqslant 2$, at least three cycles of SACF and SPACF are needed before a model can be determined. Third, both SACF and SPACF are not designed to help specify stationary mixed ARMA (P, Q) models. The SEACF procedure is designed to address these difficulties. Table 2 gives the simplified SEACF for the interest rate data. In this two-way table, the rows correspond to the AR orders P and the columns the MA orders Q. An entry 'X' signifies that the sample estimate is over twice its estimated standard error in magnitude and a '0' otherwise. The goal is to search for a triangle of 0's just like the one shown in Table 2, and the coordinates of the vertex give the AR and MA orders. In our case, we have clear indication of an ARMA (1, 1) model.

$$y_t - \phi_1 y_{t-1} = c + a_t - \theta_1 a_{t-1} \tag{6}$$

which reduces to the model in Eqn. (5) when $\phi_1 = 1$ and $c = 0$. The estimation results for both models in Eqns. (5) and (6) are summarized in Table 3. The values in parentheses are the associated estimated standard errors of the estimates. The results $\hat{\phi}_1 \cong 1$ and $\hat{c} \cong 0$ clearly show that the fitted models Eqns. (5) and (6) are essentially identical. Thus, instead of making the decision to difference the data series at the tentative model specification stage, the same result can be obtained at the estimation phase which seems more appealing. To provide a partial check of the adequacy of the model, the bottom panel of Table 1 shows the

SACF of the residuals $\hat{a}_t$'s from the model in Eqn. (5). All the sample autocorrelations are very small, indicating no dynamic pattern left in the residuals. Some additional residual diagnostics including the Box–Ljung statistic and an iterative outlier test have also been performed without revealing any serious model inadequacy. The fitted model can then be used to predict future observations if desired.

7. Seasonal ARIMA Models

Business, economic, and atmospheric time series data often exhibit a strong, but not necessarily deterministic, seasonal behavior. It seems natural to think that the seasonal pattern arises simply because observations at the same time point of the year are dynamically related across the years. For monthly data, this means that observations 12, 24,… periods apart are correlated. Just as the class of ARIMA (p, d, q) models represent the dependence of observations in successive periods (months), we may construct similar models to account for dynamic dependence of observations in the same month of successive years. As an illustration, corresponding to the regular ARIMA (0, 1, 1) model in Eqn. (5), we may consider a seasonal model of the form $\nabla_{12} y_t = e_t - \Theta_{12} e_{t-12}$, where $\nabla_{12} y_t = y_t - y_{t-12}$ is the seasonal difference and e_t is the noise term. To account for both seasonal dependence as well as month to month dependence, we may make e_t to follow the ARIMA (0, 1, 1) model $\nabla e_t = a_t - \theta_1 a_{t-1}$. Together, we have that

$$\nabla\nabla_{12} y_t = a_t - \theta_1 a_{t-1} - \Theta_{12}(a_{t-12} - \theta_1 a_{t-13}) \qquad (7)$$

where $\nabla\nabla_{12} y_t = y_t - y_{t-1} - (y_{t-12} - y_{t-13})$ is the double, regular and seasonal, difference of y_t. This 'doubly' exponential smoothing model, proposed by Box and Jenkins and motivated by their analysis of airline passenger data, has been widely used to present a variety of seasonal time series in practice. It is known as the 'airline model.' Similar extensions are readily obtainable for other ARIMA models.

8. Intervention Analysis and Outlier Detection

Time series observations are often subject to the influence of special events, or interventions, that occurred at known time points. Examples are the outbreak of armed conflicts, natural disasters, changes in fiscal policy, and major sales promotions. It has been found useful to employ linear dynamic models of the form

$$Y_t - \delta_1 Y_{t-1} - \ldots - \delta_r Y_{t-r} = \omega_0 x_t - \ldots - \omega_s x_{t-s} \qquad (8)$$

where x_t is an indicator (dummy) variable for an intervention and Y_t represents its dynamic impact on the time series observations. This class of models can accommodate a very wide variety of dynamic patterns including isolated large values, level shifts, trend, oscillations, and cycles. We can write the model of the observations as the sum of the intervention component Y_t and a noise component N_t, $y_t = Y_t + N_t$, where N_t follows an ARMA model of the form in Eqn. (3). The parameters in the intervention and the noise components can then be jointly estimated to quantify the effects of these two components.

When the time of occurrence of interventions is not known or the interventions themselves are unsuspected, we have the problem of detecting unusual events or 'outliers' in time series. Estimation procedures developed for interventions can be readily adapted to become iterative procedures for outlier detection, and these methods, especially those for detecting isolated large values and level shifts, have been commonly adopted as part of diagnostic checking.

9. Time Series with Heterogeneous Variances

The ARMA model in Eqn. (3) supposes that the variance σ^2 of the innovation a_t stays constant over time. In reality, this variance homogeneity assumption is often violated, especially in highly disaggregated financial data such as hourly stock returns or exchange rates where clusters of high volatility tend to occur. Several models have been developed to describe this kind of phenomenon. Denoting σ_t^2 as the variance of a_t, one may employ a random step-change model in which there is a nonzero probability λ that σ_t^2 may change at any given point of time. Alternatively, one may relate σ_t^2 to past values of a_t^2 and itself. In particular, the model

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \qquad (9)$$

which is called the generalized autoregressive conditional heteroscedastic (GARCH) model of orders (1, 1) has been extensively used in the analysis of financial time series. Now writing $a_t^2 = \sigma_t^2 + u_t$, the above model can be alternatively expressed as

$$a_t^2 = \alpha_0 + (\alpha_1 + \beta_1) a_{t-1}^2 + u_t - \beta_1 u_{t-1} \qquad (10)$$

It can be readily shown that the u_t's have zero mean and common variance, and are uncorrelated. Thus the model in Eqn. (10) has exactly the same autocovariance structure as an ARMA (1, 1) model in a_t^2. The above correspondence is readily extended to higher order models. Thus the GARCH models can be viewed as an interesting extension of the ARMA structure to the squared innovations a_t^2's.

10. Nonlinear Time Series Models

There is an extensive literature on nonlinear time series models where y_t is related to its past values more generally by a nonlinear function. Such models are

necessarily more complex, but they can represent phenomena such as time irreversibility and limit cycles that cannot be adequately described by linear models. For example, it is well known that the dynamic behavior of economic indicators such as the GNP or the unemployment rate during a recession is quite different from that when the economy is expanding. To describe this kind of nonhomogeneous behavior, one may employ a piecewise linear model for different regimes of economic activity. To illustrate, let y_t represent the quarterly change in GNP. We may postulate a model such that

$$y_t = \begin{cases} c^{(1)} + \phi^{(1)} y_{t-1} + a_t^{(1)}, & \text{if} \quad y_{t-1} < 0 \text{ and } y_{t-2} < 0 \\ c^{(2)} + \phi^{(2)} y_{t-1} + a_t^{(2)}, & \text{if} \quad \text{otherwise} \end{cases} \tag{11}$$

where $c^{(1)} \neq c^{(2)}$, $\phi^{(1)} \neq \phi^{(2)}$, and $a_t^{(1)}$ and $a_t^{(2)}$ have unequal variances. In Eqn. (11) the regime ($y_{t-1} < 0$ and $y_{t-2} < 0$) represents a period of economic decline, and the model says that the dynamic behavior of y_t is different for the two regimes. This model is a special case of a class, known as the threshold autoregressive models, which have been used for time series data in atmospheric sciences, economics, and other scientific fields.

11. Further Reading

The reader interested in a deeper understanding of the basic concepts and methods of time series analysis may consult the books by Box et al. (1994), Brockwell and Davis (1996), Shumway and Stoffer (2000), and Wei (1990), and the book edited by Pena et al. (2001) which also contains several chapters on multivariate ARMA models. For further details on the GARCH models and the general topic of heterogeneous variances and volatility clustering in economic time series, see Hamilton (1994). Nonlinear time series models are discussed in Priestley (1988) and Tong (1990). Finally, for multivariate extension of linear time series models, the reader may study the books by Lutkepohl (1993), Reinsel (1993) and Reinsel and Velu (1998).

Bibliography

Bollerslev T 1986 Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**: 307–27

Box G E P, Jenkins G M 1976 *Time Series Analysis, Forecasting and Control*, rev edn. Holden-Day, San Franscisco

Box G E P, Jenkins G M, Reinsel G C 1994 *Time Series Analysis, Forecasting and Control*, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ

Box G E P, Tiao G C 1975 Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association* **70**: 70–9

Brockwell P J, Davis R A 1996 *An Introduction to Time Series: Theory and Methods*. Springer-Verlag, New York

Chang I, Tiao G C, Chen C 1988 Estimation of time series parameters in the presence of outliers. *Technometrics* **30**: 193–204

Dickey D A, Fuller W A 1979 Distribution of the estimates for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74**: 427–31

Engle R F 1982 Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50**: 987–1008

Hamilton J D 1994 *Time Series Analysis*. Princeton University Press, Princeton, NJ

Lutkepohl H 1993 *Introduction to Multiple Time-Series Analysis*. Springer-Verlag, Berlin

Pena D, Tiao G C, Tsay R S (eds.) 2001 *A Course in Time Series Analysis*. Wiley, New York

Priestley M B 1988 *Non-Linear and Non-Stationary Time Series Analysis*. Academic Press, Orlando, FL

Reinsel G C 1993 *Elements of Multivariate Time-Series Analysis*. Springer-Verlag, New York

Reinsel G C, Velu R P 1998 *Multivariate Reduced Rank Regression*. Springer-Verlag, New York

Shumway R H, Stoffer D A 2000 *Time Series and its Applications*. Springer-Verlag, New York

Tong H 1990 *Nonlinear Time Series: A Dynamical System Approach*. Oxford Science Publications, New York

Tong H, Lim K S 1980 Threshold autoregression, limit cycles, and cyclical data (with discussion). *Journal of the Royal Statistical Society, Series B* **42**: 245–92

Tsay R S, Tiao G C 1984 Consistent estimates of autoregressive parameters and extended autocorrelation function for stationary and nonstationary ARMA models. *Journal of the American Statistical Association* **79**: 84–96

Wei W W S 1990 *Time Series Analysis*. Addison-Wesley, Reading, MA

G. C. Tiao

Time Series: Co-integration

1. Introduction

Co-integration deals with the common behavior of a multivariate time series. It often happens in practice that each individual component of a multivariate time series may be nonstationary, but certain linear combinations of these components are stationary. Co-integration studies the effects of these combinations and the relationships among the components.

A survey on recent developments in unit root test and nonstationary time series is covered by Dickey (see *Time Series: Nonstationary Distributions and Unit Roots*). As far as co-integration is concerned, a large number of texts have been written on the subject since the publication of the seminal paper by Engle and Granger (1987). For example, the book by Johansen (1995) provides one of the most complete coverage on

the statistical theory of co-integrated time series. A succinct survey of some of the results presented in this book is given in the review article by Johansen (1996). Other references about statistical inference for co-integrated systems include Stock and Watson (1988), Phillips (1991), Park (1992), Reinsel and Ahn (1992), and Chan and Tsay (1996). In the econometric front, several monographs and special collections of papers have also been devoted to this subject see, for example, Banerjee et al. (1993), Engle and Granger (1987), Maddala and Kim (1998), and special issues of *Oxford Bulletin of Economics and Statistics* (1990 and 1992), and *Journal of Policy Modeling* (1993).

The notion of co-integration usually is discussed with topics such as nonstationarity, unit roots, and common trends. Two excellent surveying articles on trends and co-integration are Stock (1994) and Watson (1994). Given the large number of papers written on this subject, it seems prudent to limit this article to a few fundamental concepts of co-integration. By adopting such an approach, this article aims at providing a useful tutorial in co-integration for social scientists.

From a statistical perspective, the idea of co-integration stems from the notion of transformations. When the underlying data exhibit certain nonstationary behavior, one usually introduces a transformation to render the data to stationarity. For example, the Box–Cox power transformation is often applied when nonconstant mean or nonconstant variance behaviors are detected. In a time series context, Box and Tiao (1977) discuss the idea of applying canonical correlation transformations on a multivariate time series to identify stationary components.

In economics, it is often the case that although individual components of a multivariate time series $\boldsymbol{X}_t = (X_{t1},\ldots, X_{tk})'$ appear to be nonstationary, the overall behavior of $\boldsymbol{X}_t$ may be modeled by a stationary process when an appropriate transformation is applied to $\boldsymbol{X}_t$. Thus, the idea of co-integration deals with the common behavior of a multiple time series. It stems from the fact that certain linear transformations of $\boldsymbol{X}_t$ may be stationary. Engle and Granger (1987) formulate the idea of co-integration and present statistical procedures to test for co-integration for a multiple time series.

2. Definitions and Examples

To begin, first note that nonstationarity arises when a stationary process is aggregated. For example, consider the process $\boldsymbol{X}_t = \sum_{i=1}^{t} Z_i$ where Z_i is an uncorrelated sequence of random variables with mean zero and variance σ^2, usually known as a white noise sequence and is denoted by $Z_t \sim WN(0, \sigma^2)$. The $\boldsymbol{X}_t$ has variance $t\sigma^2$ so that $\boldsymbol{X}_t$ is nonstationary. In this case, $\boldsymbol{X}_t$ is an example of an integrated order 1, $I(1)$, process and it is the cumulated sum of Z_t, that gives rise to the nonstationarity. Specifically, the notion of integration for a multivariate time series $\boldsymbol{X}_t$ is defined as follows:

(a) The process

$$\boldsymbol{X}_t = \sum_{i=0}^{\infty} \Psi_i \boldsymbol{Z}_{t-i}$$

is said to be integrated of order zero,

$$I(0), \text{ if } \sum_{i=0}^{\infty} \Psi_i \neq 0.$$

Having defined the notion of $I(0)$, we can now define the notion of $I(1)$ as follows. Herein, we define $B\boldsymbol{X}_t = \boldsymbol{X}_{t-1}$.

(b) A k-dimensional $\{\boldsymbol{X}_t\}$ is said to integrated of order *1*, $I(1)$, if

$$\boldsymbol{\Delta X}_t = (I-B)\boldsymbol{X}_t = \boldsymbol{X}_t - \boldsymbol{X}_{t-1} \text{ is } I(0) \tag{1}$$

To understand the real effect of co-integration on a multivariate time series, consider the following three-dimensional example.

Let

$$X_{t1} = \sum_{i=1}^{t} Z_{i1} + Z_{t2},\ X_{t2} = \frac{1}{2}\sum_{i=1}^{t} Z_{i1} + Z_{t3},\ X_{t3} = Z_{t3} \tag{2}$$

where $(Z_{t1}, Z_{t2}, Z_{t3})' \sim WN(\mathbf{0}, I_3)$, and I_3 denotes the identity matrix of order 3.

Clearly, $\boldsymbol{X}_t = (X_{t1}, X_{t2}, X_{t3})'$ is nonstationary since the first two components consist of the common term $\sum_{i=1}^{t} Z_{i1}$, the so-called common trend or the common nonstationary component of $\boldsymbol{X}_t$. On the other hand

$$\begin{aligned}\boldsymbol{\Delta X}_t &= (I-B)\boldsymbol{X}_t = \boldsymbol{X}_t - \boldsymbol{X}_{t-1} \\ &= (Z_{t1} + Z_{t2} - Z_{t-1,\,2},\ \tfrac{1}{2}Z_{t1} \\ &\qquad + Z_{t2} - Z_{t-1,\,2}, Z_{t3} - Z_{t-1,\,3})'\end{aligned} \tag{3}$$

is stationary. It can easily be checked that $\boldsymbol{\Delta X}_t$ is $I(0)$, hence $\boldsymbol{X}_t$ is $I(1)$. Furthermore

$$\begin{aligned}X_{t1} - 2X_{t2} &= \sum_{i=1}^{t} Z_{i1} + Z_{t2} - \sum_{i=1}^{t} Z_{i1} - 2Z_{t3} \\ &= Z_{t2} - 2Z_{t3}\end{aligned} \tag{4}$$

which is stationary as it is expressed as a finite sum of two uncorrelated white noise sequences. Although $\boldsymbol{X}_t$ is nonstationary, $\beta'\boldsymbol{X}_t$, where $\beta = (1, -2, 0)'$ is stationary and we say that $\boldsymbol{X}_t$ is co-integrated with co-integrating vector $\beta = (1, -2, 0)'$. The part $\sum_{i=1}^{t} Z_{i1}$ is called a common stochastic trend of the process $\boldsymbol{X}_t$. In this example, stationarity is attained by either differencing or by taking linear combinations.

As an application of this example, suppose X_{t1} represents the wholesale price index for a specific country, X_{t2} represents a trade-weighted foreign price index, and X_{t3} represents the exchange rates between these two countries, all measured in logarithmic scale. Preceding equations indicate that the two indices are driven by the same stochastic common trend $\sum_{i=1}^{t} Z_{i1}$. The equation

$$X_{t1} - 2X_{t2} - X_{t3} = Z_{t2} - 3Z_{t3} \tag{5}$$

represents a stationary equilibrium relationship between the two indices and the exchange rate with $(1, -2, -1)'$ being an co-integrating vector. This is an example of the so-called purchasing power parity (further details can be found in Johansen 1996). With this example as the background, we can define co-integration precisely.

(c) Let $\{\boldsymbol{X}_t\}$ be an $I(1)$ process. If there exists $\beta \neq 0$ such that $\beta'\boldsymbol{X}_t$ is stationary by a suitable choice of $\beta'\boldsymbol{X}_0$, then $\boldsymbol{X}_t$ is called co-integrated and β is called the co-integrating vector. The number of linearly independent co-integrating vectors is called the co-integrating rank and the space spanned by the co-integrating vectors is called the co-integrating space.

Let $\boldsymbol{X}_t$ be $I(1)$ and let A be a matrix of full rank. Then $A\boldsymbol{X}_t$ is also $I(1)$. Further if β is the co-integrating vector, then $(A^{-1})'\ \beta$ is the corresponding co-integrating vector of $A\boldsymbol{X}_t$. Consequently, the notion of $I(1)$ is invariant under any nonsingular linear transformation.

3. *Error Correction and Granger's Representation Theorem*

Consider a k-dimensional vector autoaggressive (VAR(p)) model $\boldsymbol{X}_t$ which satisfies the equation

$$\boldsymbol{X}_t = \Phi_1 \boldsymbol{X}_{t-1} + \cdots + \Phi_p \boldsymbol{X}_{t-p} + \boldsymbol{Z}_t \tag{6}$$

where $\{\boldsymbol{Z}_t\} \sim WN(\boldsymbol{0}, I_k)$, a k-dimensional white noise process. It is well known that (see, for example, Lütkepohl 1993, p.11), a sufficient condition for $\boldsymbol{X}_t$ to be stationary is that the solutions of the characteristic polynomial to Eqn. (6) have to lie outside that unit circle (no unit roots allowed), i.e.,

$$\det(I - \Phi_1\zeta - \cdots - \Phi_p\zeta^p) \neq 0 \text{ for all } |\zeta| \leqslant 1 \tag{7}$$

Now, define $\Phi(\zeta) = I - \Phi_1\zeta - \cdots - \Phi_p\zeta^p$, $\Phi = -\Phi(1)$, and $\Gamma = -d\Phi(\zeta)/d\zeta|_{\zeta=1} + \Phi$. Then

$$\begin{aligned}
\Phi &= -\Phi(1) = \Phi_1 + \cdots + \Phi_p - I \\
\Gamma &= -\frac{d\Phi(\zeta)}{d\zeta}\Big|_{\zeta=1} + \Phi \\
&= \Phi_1 + 2\Phi_2\zeta\,|_{\zeta=1} + \cdots + p\Phi_p\zeta^{p-1}|_{\zeta=1} + \Phi \\
&= \Phi_1 + 2\Phi_2 + \cdots + p\Phi_p + \Phi_1 + \cdots + \Phi_p - I \\
&= 2\Phi_1 + 3\Phi_2 + \cdots + (p+1)\Phi_p - I'
\end{aligned} \tag{8}$$

With these definitions, we first state a well-known condition under which a stationary VAR(p) process is $I(0)$.

If $\boldsymbol{X}_t$ is a stationary VAR (p) process such that det $(\Phi(\zeta)) \neq 0$ for all $|\zeta| < 1$, then $\boldsymbol{X}_t$ can be given an initial distribution such that it becomes $I(0)$ if and only if Φ is full rank. In this case, there is no unit root for $\boldsymbol{X}_t$ (stationary) and

$$\boldsymbol{X}_t = \sum_{i=0}^{\infty} \Psi_i \boldsymbol{Z}_{t-i} \tag{9}$$

where $\Psi(\zeta) = \sum_{i=0}^{\infty} \Psi_i \zeta^i = (\Phi(\zeta))^{-1}$converges for $|\zeta| < 1 + \delta$ for some $\delta > 0$.

Since $\Phi = \Phi_1 + \cdots + \Phi_p - I$, the full rank condition of Φ in this statement holds if and only if det Φ = det $(\Phi(1)) \neq 0$. In other words, the full rank condition of Φ simply means that det $(\Phi(\zeta)) \neq 0$ for $\zeta = 1$. When this matrix is of reduced rank, then we have a co-integrated system.

Let α be any $k \times r$ matrix of rank r $(r < k)$. Define $\alpha_\perp$ to be $a\ k \times (k-r)$ matrix which is full rank such that $\alpha'\alpha_\perp = 0$. Now assume the following condition holds for $\boldsymbol{X}_t$: $\det(\Phi(\zeta)) \neq 0$ for all $|\zeta| < 1$, i.e., the process $\boldsymbol{X}_t$ is nonstationary since it can have unit roots. Under this assumption, we have the Granger's representation theorem.

3.1 *Granger's Representation Theorem*

Let $\boldsymbol{X}_t$ be a VAR(p) and $\det(\Phi(\zeta)) \neq 0$ for all $|\zeta| < 1$, i.e., unit roots are allowed so that the process $\boldsymbol{X}_t$ can be nonstationary. Then $\boldsymbol{X}_t$ is $I(1)$ if and only if

$$\Phi = \alpha\beta' \tag{10}$$

where α and β are $(k \times r)$ matrices which are of full rank r $(r < k)$, and $\alpha_\perp'\Gamma\beta_\perp$ is of full rank. In this case, $\Delta\boldsymbol{X}_t$ and $\beta'\boldsymbol{X}_t$ can be given initial distributions so that they become $I(0)$. Furthermore, under these circumstances, $\boldsymbol{X}_t$ can be expressed as

$$\boldsymbol{X}_t = \beta_\perp(\alpha_\perp'\Gamma\beta_\perp)^{-1}\alpha_\perp' \sum_{i=1}^{t} \boldsymbol{Z}_i + \Psi(B)\boldsymbol{Z}_t \tag{11}$$

and

$$\Delta\boldsymbol{X}_t = \alpha\beta'\boldsymbol{X}_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta\boldsymbol{X}_{t-i} + \boldsymbol{Z}_t \tag{12}$$

where

$$\Gamma_i = -(\Phi_{i+1} + \cdots + \Phi_p),\ i = 1, \ldots, p-1 \tag{13}$$

Note that Eqn (12) in this theorem is the error correction form expressed in terms of $\Delta\boldsymbol{X}_t$. This theorem can best be interpreted in terms of a specific VAR(2) example

$$\Delta\boldsymbol{X}_t = \begin{pmatrix} -\alpha_1 & 2\alpha_1 \\ 0 & 0 \end{pmatrix} \boldsymbol{X}_{t-1} + \boldsymbol{Z}_t = A\boldsymbol{X}_{t-1} + \boldsymbol{Z}_t \tag{14}$$

In other words

$$X_t = (I+A)X_{t-1} + Z_t = \Phi_1 X_{t-1} + Z_t \tag{15}$$

where

$$\Phi_1 = (I+A) = \begin{pmatrix} 1-\alpha_1 & 2\alpha_1 \\ 0 & 1 \end{pmatrix} \tag{16}$$

Observe that

$$\begin{aligned} \Phi(\zeta) &= I - \Phi_1\zeta = I - (I+A)\zeta \\ &= \begin{pmatrix} 1-\zeta+\alpha_1\zeta-2\alpha_1\zeta \\ 0 \qquad 1-\zeta \end{pmatrix} \\ &= \begin{pmatrix} \alpha_1 & -2\alpha_1 \\ 0 & 0 \end{pmatrix} + (1-\zeta)\begin{pmatrix} 1-\alpha_1 & 2\alpha_1 \\ 0 & 1 \end{pmatrix} \end{aligned} \tag{17}$$

with

$$\Phi = \Phi(1) = \begin{pmatrix} \alpha_1 & -2\alpha_1 \\ 0 & 0 \end{pmatrix} \tag{18}$$

and

$$\det(\Phi(\zeta)) = (1-\zeta)(1+\alpha_1\zeta-\zeta) \tag{19}$$

This characteristic polynomial has roots $\zeta = 1$ and $\zeta = (1-\alpha_1)^{-1}$, if $\alpha_1 \neq 1$. Therefore, for $0 < \alpha_1 < 1$, $\det \Phi(\zeta) = 0$ for $\zeta = 1$ and $\zeta = \frac{1}{1-\alpha_1}$ and as a result, $\det\Phi(\zeta) = 0$ for $|\zeta| \geqslant 1$. Clearly, Φ has reduced rank $(=1)$ and $\Phi = \alpha\beta'$, where $\alpha = (\alpha_1, 0)'$, $\beta = (1,-2)'$, $\alpha_\perp = (0, 1)'$ and $\beta_\perp = (2, 1)'$. Hence

$$\alpha_\perp'\Gamma\beta_\perp = (0, 1)(2\Phi_1 - I)\begin{pmatrix} 2 \\ 1 \end{pmatrix} = 1 \tag{20}$$

By letting $X_0 = \mathbf{0}$, the condition of the Granger's representation theorem is satisfied and X_t is, therefore, $I(1)$. In this case

$$\begin{aligned} X_t &= \begin{pmatrix} X_{t1} \\ X_{t2} \end{pmatrix} \\ &= \begin{pmatrix} 2\sum_{i=1}^{t} Z_{i2} + \sum_{i=0}^{\infty}(1-\alpha_1)^i(Z_{t-i,1} - Z_{t-i,2}) \\ \sum_{i=1}^{t} Z_{i2} \end{pmatrix} \\ &= \begin{pmatrix} 0 & 2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \sum_{i=1}^{t} Z_{i1} \\ \sum_{i=2}^{t} Z_{i2} \end{pmatrix} + \sum_{i=0}^{\infty}\Psi_i Z_{t-1} \end{aligned} \tag{21}$$

where

$$\Psi_i = \begin{pmatrix} (1-\alpha_1)^i & -2(1-\alpha_1)^i \\ 0 & 0 \end{pmatrix}, \; i = 0, 1, \ldots \tag{22}$$

that is,

$$X_t = \beta_\perp\alpha_\perp'\sum_{i=1}^{t} Z_i + \sum_{i=0}^{\infty}\Psi_i Z_{t-i} \tag{23}$$

Note that the quantity $\alpha_\perp'\sum_{i=1}^{t} Z_i$ of Eqn (23) constitutes the common trend. Substituting the definition of $\alpha_\perp = (0, 1)'$ into this expression, the common trend equals to $\sum_{i=1}^{t} Z_{i2}$. Furthermore

$$\begin{aligned} \Delta X_t &= \begin{pmatrix} -\alpha_1 & 2\alpha_1 \\ 0 & 0 \end{pmatrix} X_{t-1} + Z_t \\ &= \begin{pmatrix} \alpha_1 \\ 0 \end{pmatrix}(1,-2)X_{t-1} + Z_t \\ &= \alpha\beta' X_{t-1} + Z_t \end{aligned} \tag{24}$$

In this case, the co-integrating vector is $\beta = (1,-2)'$ and $\Gamma_i = 0$ in the Granger's representation theorem.

4. *Statistical Inference for Co-integrated Systems*

Recall that from the Granger's representation theorem, given that X_t is $I(1)$, the error correction form is

$$\Delta X_t = \Phi X_{t-1} + \sum_{i=1}^{p-1}\Gamma_i\Delta X_{t-i} + Z_t \tag{25}$$

where $\Phi = \alpha\beta'$ is of rank $r(r < k)$ with α, β being $(k \times r)$ matrices. Let $Z_t \sim \mathrm{N}(\mathbf{0}, \sum_Z)$ so that X_t becomes a Gaussian time series. We now show how to construct estimators and tests for co-integrated systems by the following steps. Let the length of the underlying series be T.

(a) Regress ΔX_t on ΔX_{t-i}, $i = 1,\ldots, p-1$ and form the residuals R_{0t}.

(b) Regress X_{t-1} on ΔX_{t-i}, $i = 1,\ldots, p-1$ and form the residual R_{1t}.

(c) Calculate the sums

$$S_{ij} = \frac{1}{T}\sum_{t=1}^{T} R_{it}R_{jt}', \quad i, j = 0, 1 \tag{26}$$

Under the hypothesis H_r: $\Phi = \alpha\beta'$ with rank $(\Phi) = r$, the maximum likelihood estimate for β is given by

$$\hat{\beta} = (\hat{v}_1,\ldots, \hat{v}_r) \tag{27}$$

where each $\hat{v}_i$ is a $k \times 1$ vector obtained by solving the following 'eigenvalues' and 'eigenvectors' problem:

$$|\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}| = 0 \tag{28}$$

for

$$1 > \hat{\lambda}_1 > \cdots > \hat{\lambda}_p > 0 \tag{29}$$

with

$$(\hat{\lambda}_i S_{11} - S_{10} S_{00}^{-1} S_{01}) \hat{v}_i = 0 \tag{30}$$

In this case, the maximum likelihood function becomes

$$L_{\max}^{-2/T} \propto |S_{00}| \Pi_{i=1}^{r} (1 - \hat{\lambda}_i)$$

and an estimator for α is given by $\hat{\alpha} = S_{01}\hat{\beta}$.

Note that by the definition of canonical correlation, $\hat{\lambda}_i$, represents the squared canonical correlation between $\Delta \boldsymbol{X}_t$ and $\boldsymbol{X}_{t-1}$ after adjusting for the effects of $\Delta \boldsymbol{X}_{t-1}, \ldots, \Delta \boldsymbol{X}_{t-p}$. A number of remarks follow from this theorem.

(a) If $r = 0$, $\Delta \boldsymbol{X}_t$ is $I(0)$ but $\boldsymbol{X}_{t-1}$ is $I(1)$ so that we expect that $\Delta \boldsymbol{X}_t$ and $\boldsymbol{\Delta X}_{t-1}$ are 'uncorrelated.' Proof of this fact can be found in Chan and Wei (1988). This result means that all $\hat{\lambda}_i$ should be very small.

(b) If $r = k$, $\Delta \boldsymbol{X}_t$ and $\boldsymbol{\Delta X}_{t-1}$ are both $I(0)$. In this case, we expect significant relationship among them. As a result, the estimates of β are coming from those that correlate most with the stationary process $\boldsymbol{\Delta X}_t$ after adjusting for the effects of $\boldsymbol{\Delta X}_{t-1}, \ldots, \boldsymbol{\Delta X}_{t-p}$.

(c) For r in between, we choose $\hat{\beta}$ by taking the most correlated components in the canonical correlation sense.

After estimating $\hat{\beta}$ as above, we can estimate α by means of $\hat{\alpha} = S_{01}\hat{\beta}$. Turning to testing for the hypothesis of H_r, i.e., there are at most r co-integrating relationships, consider the likelihood ratio statistic as follows

$$Q_r = -2\log Q(H_r|H_k) = -T \sum_{i=r+1}^{k} \log(1 - \hat{\lambda}_i) \tag{31}$$

We reject H_r in favor of H_{r+1} if Q_r is too big. The interpretation for constructing such a statistic can be argued as follows. When H_r is true, there are supposed to be r co-integrating relationships which should be captured by $\hat{\lambda}_1, \ldots, \hat{\lambda}_r$. As a consequence, $\hat{\lambda}_i$ for $i = r+1, \ldots, k$ should then be relatively small. If, on the contrary, they are observed to be too big for any given series, Q_r becomes huge and there is strong evidence of rejecting H_r in favor of H_{r+1}. With this interpretation, the next step is how to compare Q_r with known distributions. In a standard multivariate context, we expect to compare Q_r with some form of a chi-squared distribution. Unfortunately, due to the nonstationarity of $\boldsymbol{X}_t$, standard asymptotic is no longer available. Instead, we need to consider the following.

Let $\boldsymbol{X}_t$ be a VAR(p) with r co-integrating relations. Then

$$\begin{aligned} Q_r &= -2\log(Q(H_r|H_k)) \\ &= -T \sum_{i=r+1}^{k} \log(1 - \hat{\lambda}_i) \\ &\xrightarrow{\mathscr{L}} \xi \end{aligned} \tag{32}$$

where

$$\xi = \text{trace} \left\{ \int_0^1 \mathrm{d}B(s)(B(s))' \left[\int_0^1 B(s)\, B'(s)\mathrm{d}s \right]^{-1} \int_0^1 B(s)(\mathrm{d}B(s))' \right\} \tag{33}$$

with $B(s)$ being the $(k-r)$ dimensional standard Brownian motion, and $\xrightarrow{\mathscr{L}}$ designates convergence in distribution of the test statistics Q_r to that of ξ as the length of the time series $T \to \infty$.

Notice that the limiting distribution depends only on the number $(k-r)$, which is the number of common trends in the underlying series $\boldsymbol{X}_t$. Tabulations of special cases of ξ can be found in a number of places, including Banerjee et al. (1993), Johansen (1995), Tanaka (1996), Maddala and Kim (1998), and Lütkepohl (1993).

5. *Summary*

As indicated in the introduction, co-integration has been an actively pursued area of research, particularly among econometricians. Owing to its simple idea, it is reasonable to expect that the concept of co-integration will be more significant in other areas of social sciences. Recently, the concept of co-integration has found applications in many areas in finance. For example, Hiraki et al. (1996) use co-integration and common trend to explain the term structure of offshore interest rates and, more recently, Lee et al. (1999) use the idea of co-integration to determine the intrinsic value of the Dow Jones industrial index. Another area that is under rigorous development is combining long-memory phenomena together with co-integration, the so-called fractional co-integration (see, for example, Lien and Tse 1999). It is hoped that this survey offers a focused and practical introduction to co-integration.

See also: Time Series: Advanced Methods; Time Series: ARIMA Methods; Time Series: Cycles; Time

Series: Economic Forecasting; Time Series: General; Time Series: Nonstationary Distributions and Unit Roots; Time Series: Seasonal Adjustment

N. H. Chan

Bibliography

Banerjee A, Dolado J J, Galbraith J W, Hendry D F 1993 *Co-Integration, Error Correction, and the Econometric Analysis of Non-stationary Data*. Oxford University Press, Oxford, UK

Box G E P, Tiao G C 1977 A canonical analysis of multiple time series. *Biometrika* **64**: 355–65

Chan N H, Tsay R S 1996 On the use of canonical correlation analysis in testing common trends. In: Lee J C, Johnson W O, Zellner A (eds.) *Modelling and Prediction: Honoring S. Geisser*. Springer-Verlag, New York, pp. 364–77

Chan N H, Wei C Z 1988 Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* **16**: 367–401

Engle R F, Granger C W J 1987 Cointegration and error correction: Representation, estimation, and testing. *Econometrica* **55**: 251–76

Engle R F, Granger C W J 1991 *Long-run Economic Relations: Readings in Co-integration*. Oxford University Press, Oxford, UK

Hiraki T, Shiraishi N, Takezwa N 1996 Cointegration, common factors, and term structure of Yen offshore interest rates. *Journal of Fixed Income* **December**: 69–75

Johansen S 1988 Statistical analysis of co-integration vectors. *Journal of Economic Dynamics and Control* **12**: 231–54

Johansen S 1995 *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, Oxford, UK

Johansen S 1996 Likelihood-based inference for co-integration of some nonstationary time series. In: Cox D R, Hinkley D V, Barndorff-Nielsen O E (eds.) *Time Series Models: In Econometrics, Finance and Other Fields*. Chapman and Hall, New York

Lee C M C, Myers J, Swaminathan B 1999 What is the intrinsic value of the Dow? *Journal of Finance* **54**: 1693–741

Lien D, Tse Y K 1999 Forecasting the Nikkei spot index with fractional co-integration. *Journal of Forecasting* **18**: 259–73

Lütkepohl H 1993 *Introduction to Multiple Time Series Analysis*. 2nd edn. Springer-Verlag, New York

Maddala G S, Kim I M 1998 *Unit Roots, Cointegration, and Structural Change*. Cambridge University Press, Cambridge, UK

Park J Y 1992 Canonical co-integrating regressions. *Econometrica* **60**: 119–43

Phillips P C B 1991 Optimal inference in co-integrated systems. *Econometrica* **59**: 283–306

Reinsel G C, Ahn S K 1992 Vector AR models with unit roots and reduced rank structure: estimation, likelihood ratio test, and forecasting. *Journal of Time Series Analysis* **13**: 353–75

Stock J H 1994 Unit roots, structural breaks and trends. In: Engle R F, McFadden D L (eds.) *Handbook of Econometrics*, Vol. IV. Elsevier, Netherlands

Stock J H, Watson M W 1988 Testing for common trends. *Journal of the American Statistical Association* **83**: 1097–107

Tanaka K 1996 *Time Series Analysis: Nonstationary and Noninvertibel Distribution Theory*. Wiley, New York

Watson M W 1994 Vector autoregressions and co-integration. In: Engle R F, McFadden D L (eds.) *Handbook of Econometrics* Vol. IV. Elsevier, The Netherlands

Time Series: Cycles

Time series data in economics and other fields of social science often exhibit cyclical behavior. For example, aggregate retail sales are high in November and December and follow a seasonal cycle; voter registrations are high before each presidential election and follow an election cycle; and aggregate macroeconomic activity falls into recession every six to eight years and follows a business cycle. In spite of this cyclicality, these series are not perfectly predictable, and the cycles are not strictly periodic. That is, each cycle is unique. Quite generally, it is useful to think of time series as realizations of stochastic processes, and this raises the question of how to represent cyclicality in stochastic processes. The major descriptive tool for this is the 'spectral density function' (or 'spectrum'), which is the subject of this article.

1. *Cycles in a Typical Time Series*

Figure 1 shows a time series plot of new housing authorizations ('building permits') issued by communities in the USA, monthly, from 1960 through 1999. This plot has characteristics that are typical of many economic time series. First, the plot shows a clear seasonal pattern: permits are low in the late fall and early winter, and rise markedly in the spring and summer. This seasonal pattern is persistent throughout the sample period, but does not repeat itself exactly year-after-year. In addition to a seasonal pattern, there is a slow moving change in the level of the series associated with the business cycle. For example, local minima of the series are evident in 1967, the mid-1970s, the early 1980s, and in 1990, which correspond to periods of macroeconomic slowdown or recession in the US economy.

How can the cyclical variability in the series be represented? One approach is to use a periodic function like a sine or cosine function. But this is inadequate in at least two ways. First, a deterministic periodic function doesn't capture the randomness in the series. Second, since several different periodicities (seasonal, business cycle) are evident in the plot, several periodic functions will be required. The next section presents a representation of a stochastic process that has these ingredients.

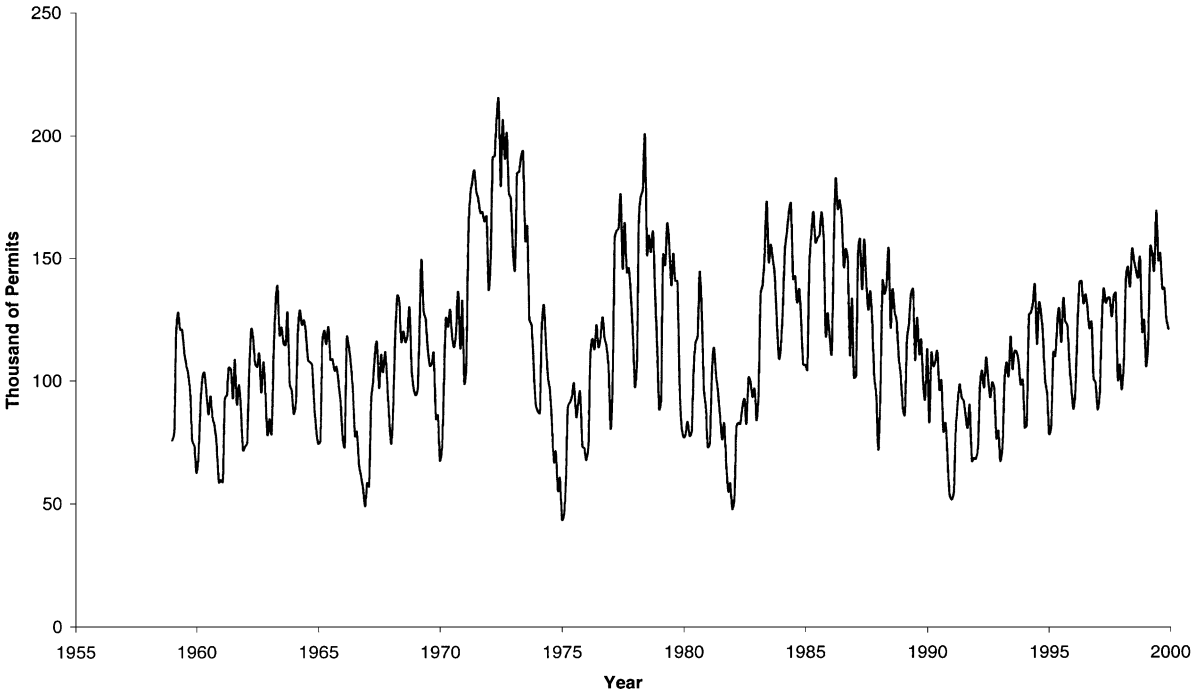

Figure 1
Building permits in the USA

2. *Representing Cyclical Behavior in Time Series: The Spectrum*

2.1 *Spectral Representation of a Covariance Stationary Process*

Consider a sequence of scalar random variables $\{Y_t\}$ where $t = 0, \pm 1, \pm 2, \ldots$, and $E(Y_t) = \mu$, $Cov(Y_t Y_{t-k}) = \lambda_k$. That is, assume that Y_t is observed at regular intervals, like a week, a month, a year, etc. Assume that the first and second moments of the process do not depend on time, that is μ and λ_k are time invariant. To simplify notation, assume that $\mu = 0$.

To motivate the spectral representation of Y_t, it is useful to begin with a very special and simple stochastic process:

$$Y_t = \alpha \cos(\omega t) + \delta \sin(\omega t) \tag{1}$$

where α and δ are random variables with $E(\alpha) = E(\delta) = 0$; $E(\alpha\delta) = 0$; and $E(\alpha^2) = E(\delta^2) = \sigma^2$. While simple, this process has three attractive characteristics. First, it is periodic: since $\cos(a + 2\pi) = \cos(a)$ and $\sin(a + 2\pi) = \sin(a)$, then $Y_{t+j2\pi/\omega} = Y_t$ for $|j| = 1, 2, \ldots$. So Y_t repeats itself with a period of $2\pi/\omega$. Second, the random components α and δ give Y a random amplitude (value of at its peak) and a random phase (value at $t = 0$). Thus, two realizations of Y will have different amplitudes and different timing of their peaks and troughs. Finally, the process is covariance stationary with $E(Y_t) = 0$ and $\lambda_k = \sigma^2 \cos(\omega k)$.

Adding together several components like Eqn. 1 produces a more interesting stochastic process:

$$Y_t = \sum_{j=1}^{n} [\alpha_j \cos(\omega_j t) + \delta_j \sin(\omega_j t)] \tag{2}$$

with $E(\alpha_j) = E(\delta_j) = 0$, for all j; $E(\alpha_j \delta_k) = 0$, for all j, k; $E(\alpha_j \alpha_k) = E(\delta_j \delta_k) = 0$, for $j \neq k$; and $E(\alpha_j^2) = E(\delta_j^2) = \sigma_j^2$. For this process, Y_t is the sum of n different uncorrelated components, each corresponding to a different frequency, and each with its own variance. (The variance of $\alpha_j \cos(\omega_j t) + \delta_j \sin(\omega_j t)$ is σ_j^2.) A calculation shows $E(Y_t) = 0$, $\lambda_k = \sum_{j=1}^{n} \sigma_j^2 \cos(\omega_j k)$, and the variance of Y_t is given by $var\,(Y_t) = \sum_{j=1}^{n} \sigma_j^2$.

So far, this has all been special. That is, two very special processes have been constructed. However, a fundamental result (Cramér 1942, Kolmorgorov 1940) shows that a generalization of this decomposition can be used to represent any covariance stationary process. The result, known as the 'Spectral Representation Theorem,' says that if Y_t is covariance stationary, then it can be represented as

$$Y_t = \int_0^{\pi} \cos(\omega t)\, d\alpha(\omega) + \int_0^{\pi} \sin(\omega t)\, d\delta(\omega)$$

where $d\alpha(\omega)$ and $d\delta(\omega)$ are zero mean random variables that are mutually uncorrelated, uncorrelated across frequency, and have variances that depend on frequency. Like the special example shown in Eqn. 2, this general representation decomposes Y_t into a set of strictly periodic components—each uncorrelated with the others, and each with its own variance. For example, processes with important seasonal components will have large values of the components corresponding to the seasonal frequency, processes for series with strong 'business cycle' components will have large variances at business cycle frequencies, etc.

Sometimes the spectral representation is written using complex numbers as:

$$Y_t = \int_{-\pi}^{\pi} e^{i\omega t} d\gamma(\omega) \tag{3}$$

where $d\gamma(\omega) = (1/\sqrt{2})[d\alpha(\omega) - i d\delta(\omega)]$ for $\omega \geqslant 0$ and $d\gamma(\omega) = d\gamma(-\omega)$ for $\omega < 0$. This representation simplifies notation for some calculations.

The spectral representation Eqn. 3 is a generalization of Eqn. 2 where the 'increments' $d\gamma(\omega)$ have orthogonality properties like those of α_j and δ_j. The variance of each component is frequency specific (like σ_j^2), and is summarized by a density function $S(\omega)$. Specifically, the variance of $d\gamma(\omega)$ is $E(d\gamma(\omega)\overline{d\gamma(\omega)}) = S(\omega)d\omega$. (The use of a density function to summarize the variance of $d\gamma(\omega)$ is not completely general and rules out processes with deterministic (or perfectly predictable) components. The references listed in the last section provide this more general result.)

You may wonder why the spectral representation doesn't use frequencies larger than π. The answer is that these are not needed to describe a process measured at discrete intervals. This discreteness means that periodic components associated with frequencies larger than π will look just like components associated with frequencies less than π. For example, a component with frequency 2π will have period of 1 and, because the series is measured only once every period, this component will appear to be constant—it will look just like (be 'aliased' with) a component that has a frequency of $\omega = 0$.

2.2 Relationship Between the Spectral Density and the Autocovariances

There is a one-to-one relationship between the spectral density function and the autocovariances of the process. The autocovariances follow directly from the spectral representation:

$$\begin{aligned}\lambda_k = E(Y_t Y_{t-k}) &= E(Y_t \bar{Y}_{t-k}) \\ &= E\left[\int e^{i\omega t}\, d\gamma(\omega) \int e^{-i\omega(t-k)}\, \overline{d\gamma(\omega)}\right] \\ &= \int e^{i\omega k} S(\omega)\, d\omega. \end{aligned} \tag{4}$$

The spectrum can be determining by inverting this relation:

$$S(\omega) = (2\pi)^{-1} \sum_{j=-\infty}^{\infty} \lambda_j e^{-i\omega j} \tag{5}$$

(To verify that Eqn. 5 is the inverse of Eqn. 4, use

$$\int_{-\pi}^{\pi} e^{i\omega k} d\omega = \left\{ \begin{array}{lll} 2\pi & \text{for} & k = 0 \\ 0 & \text{for} & k \neq 0 \end{array} \right\} \tag{6}$$

so that $(2\pi)^{-1}\{\int_{-\pi}^{\pi} \sum_{j=-\infty}^{\infty} \lambda_j e^{-i\omega(j-k)} d\omega\} = \lambda_k$.)

A simpler formula for the spectrum is

$$S(\omega) = (2\pi)^{-1}[\lambda_0 + 2\sum_{j=1}^{\infty} \lambda_j \cos(j\omega)]. \tag{7}$$

which follows from Eqn. 5, since $\lambda_j = Cov(Y_t Y_{t-j}) = Cov(Y_{t+j} Y_t) = \lambda_{-j}$.

2.3 A Summary of the Properties of the Spectrum

Summarizing the results presented in Sect. 2.2., the spectrum (or spectral density function) has 4 important properties.

(a) $S(\omega)d\omega$ can be interpreted as the variance of the cyclical component of Y corresponding to the frequency ω. The period of this component is $2\pi/\omega$.

(b) $S(\omega) \geqslant 0$. This follows, because $S(\omega)$ is a variance function.

(c) $S(\omega) = S(-\omega)$. This follows from the definition of $\gamma(\omega)$ in the spectral representation Eqn. 3, or from Eqn. 7, since $\cos(a) = \cos(-a)$. Because of this symmetry, plots of the spectrum are presented $0 \leqslant \omega \leqslant \pi$.

(d) $Var(Y) = \int_{-\pi}^{\pi} S(\omega)\, d\omega$.

2.4 The Spectrum of Building Permits

The time series plot of building permits (Fig. 1) shows that most of the variability in the series comes from two sources: seasonal variation over the year and business-cycle variability. This is evident in the estimated spectrum for the series, shown in Fig. 2. Most of the mass in the spectrum is concentrated around the seven peaks evident in the plot. (These peaks are sufficiently large that spectrum is plotted on a log scale.) The first peak occurs at frequency $\omega = 0.07$ corresponding to a period of 90 months. This represents the business cycle variability in the series. The other peaks occur at frequencies $2\pi/12$, $4\pi/12$, $6\pi/12$, $8\pi/12$, and π. These are peaks for the seasonal frequencies: the first corresponds to a period of 12 months, and the others are the seasonal 'harmonics' 6, 4, 3, and 2 months.

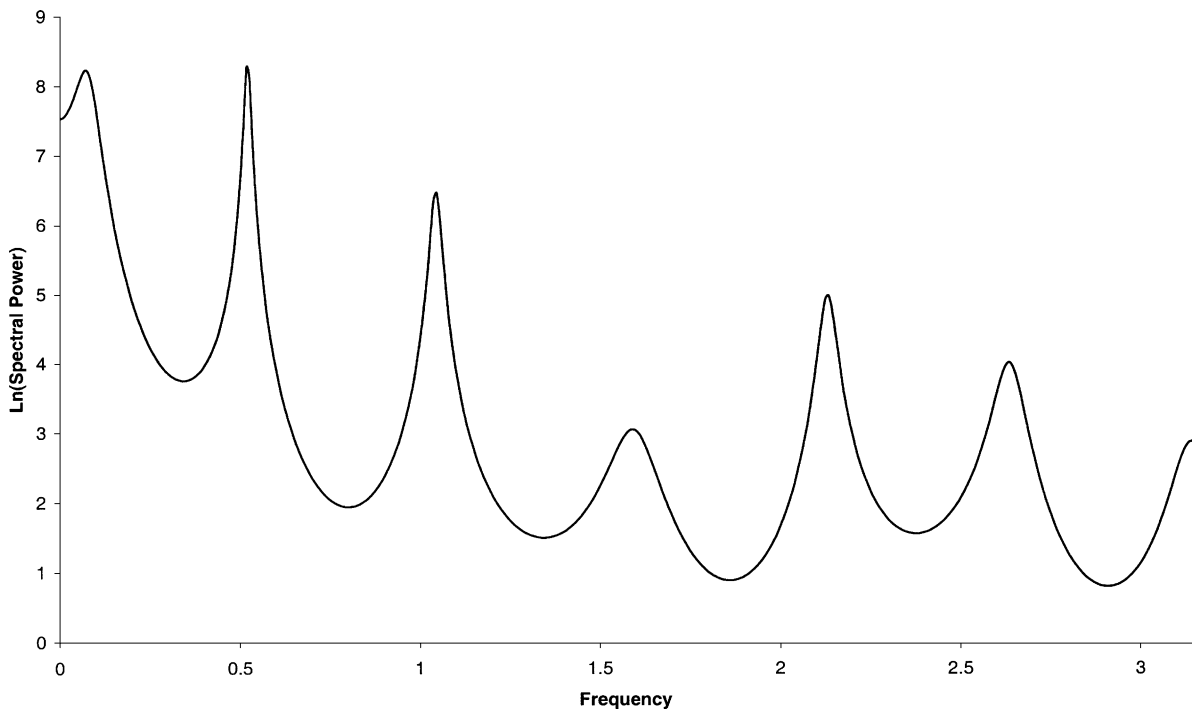

Figure 2
Spectrum of building permits

3. *Spectral Properties of Moving Average Filters*

3.1 *Some General Results*

Often, one time series (Y) is converted into another time series (X) through a moving average function such as:

$$X_t = \sum_{j=-r}^{s} c_j Y_{t-j}$$

For example, X_t might be the first difference of Y_t, i.e., $X_t = Y_t - Y_{t-1}$, or an annual moving average $X_t = \frac{1}{12}\sum_{j=0}^{11} Y_{t-j}$, or one of the more complicated moving averages that approximate official seasonal adjustment procedures (see *Time Series: Seasonal Adjustment*).

Let the 'lag' operator L shift time series back by one period (so that $LY_t = Y_{t-1}$, $L^2 Y_t = L(LY_t) = Y_{t-2}$, $L^{-1}Y_t = Y_{t+1}$, etc.), then this moving average can be represented as:

$$X_t = c(L)Y_t$$

where,

$$c(L) = c_{-r}L^{-r} + \cdots + c_{-1}L^{-1} + c_0 L^0 + c_1 L + \cdots + c_s L^s$$

The operator $c(L)$ is sometimes called a 'linear filter.'

How does $c(L)$ change the spectral properties of Y? If the spectral representation of Y_t is written as:

$$Y_t = \int_{-\pi}^{\pi} \mathrm{e}^{i\omega t}\, \mathrm{d}\gamma(\omega)$$

then X_t can be written as,

$$\begin{aligned} X_t &= \sum_{j=-r}^{s} c_j Y_{t-j} \\ &= \sum_{j=-r}^{s} c_j \int_{-\pi}^{\pi} \mathrm{e}^{i\omega(t-j)}\, \mathrm{d}\gamma(\omega) \\ &= \int_{-\pi}^{\pi} \mathrm{e}^{i\omega t} \left(\sum_{j=-r}^{s} c_j e^{-i\omega j} \right) \mathrm{d}\gamma(\omega) \\ &= \int_{-\pi}^{\pi} \mathrm{e}^{i\omega t} c(\mathrm{e}^{-i\omega})\, \mathrm{d}\gamma(\omega) \\ &= \int_{-\pi}^{\pi} \mathrm{e}^{i\omega \left(t - \frac{\rho(\omega)}{\omega} \right)} g(\omega)\, \mathrm{d}\gamma(\omega) \end{aligned} \tag{8}$$

where the last line uses the polar form of $c(\mathrm{e}^{-i\omega})$: $c(\mathrm{e}^{-i\omega}) = g(\omega)\mathrm{e}^{-i\rho(\omega)}$ where

$$g(\omega) = |\, c(\mathrm{e}^{i\omega})\, | \tag{9}$$

and

$$\rho(\omega) = \tan^{-1}\left[\frac{-\mathrm{Im}[c(\mathrm{e}^{-i\omega})]}{\mathrm{Re}[c(\mathrm{e}^{-i\omega})]}\right] \tag{10}$$

This representation for X_t shows that the filter $c(L)$ 'amplifies' each component of Y_t by the factor $g(\omega)$ and shifts it back in time by $\rho(\omega)/\omega$ time units.

The function $g(\omega)$ is called the filter *gain* (or sometimes the 'amplitude gain.' The function $\rho(\omega)$ is called the filter 'phase' and $g(\omega)^2 = [c(\mathrm{e}^{-i\omega})c(\mathrm{e}^{i\omega})]$ is called the 'power transfer function' of the filter.

3.2 Three Examples

3.2.1 Example one: lag filter. Suppose that $c(L) = L^2$. Then $c(\mathrm{e}^{-i\omega}) = \mathrm{e}^{-2i\omega}$,

$$g(\omega) = |c(\mathrm{e}^{i\omega})| = 1$$

and

$$\rho(\omega) = \tan^{-1}\left[\frac{\sin 2\omega}{\cos 2\omega}\right] = 2\omega$$

Thus, this filter does not alter the amplitude of any of the components, $g(\omega) = 1$, but shifts each component back in time by $\rho(\omega)/\omega = 2$ time periods.

3.2.2 Example two: first difference filter. Suppose that $c(L) = (1-L)$ so that X_t is the first difference of Y_t. Then $c(\mathrm{e}^{-i\omega}) = 1-\mathrm{e}^{-i\omega}$,

$$g(\omega) = (-\mathrm{e}^{-i\omega}+2-\mathrm{e}^{i\omega})^{\frac{1}{2}} = \sqrt{2(1-\cos(\omega))}$$

and

$$\rho(\omega) = \tan^{-1}\left[\frac{\sin(\omega)}{1-\cos(\omega)}\right]$$

Thus, both the gain and phase are frequency specific. For example, $g(0) = 0$ so that the first difference filter eliminates the lowest frequency component of the series (the level of the series is 'differenced out'), and $g(\pi) = \sqrt{2}$ so that the high frequency components are amplified (X_t is 'choppier' than Y_t).

3.2.3 Example three: census X-12 seasonal adjustment Filter. The official monthly seasonal adjustment procedure in the USA and several other countries (Census X-12) can be well approximated by a linear filter $c(L) = \sum_{j=-84}^{84} c_{|j|}L^j$ where the coefficients c_i are given in Wallis (1974). Since the filter is symmetric, $c_j = c_{-j}$, $c(\mathrm{e}^{i\omega})$ is real and so $\rho(\omega) = 0$. The function $g(\omega)$ is plotted in Fig. 3. The gain is nearly unity except near the seasonal frequencies, and so the filter can be interpreted as producing a new series that leaves the non-seasonal components unaltered, but eliminates the seasonal components. For a detailed discussion of seasonal adjustment procedures see *Time Series: Seasonal Adjustment*.

3.3 Band-pass Filters

The seasonal adjustment filter has the desirable properties of (essentially) zeroing out certain frequency components (the 'seasonal') and leaving the other components (the non-seasonals) unaltered. Filters with this characteristic are called 'band-pass filters,' since they 'pass' components associated with certain frequencies.

Band-pass filters for prespecified frequencies are easy to construct. For example, suppose that you want a filter with no phase shift that passes components between $0 \leqslant \omega \leqslant \underline{\omega}$. That is, you want to construct a filter, say $c(L)$ with gain $|g(\omega)| = 1$ for $0 \leqslant \omega \leqslant \underline{\omega}$ and $|g(\omega)| = 0$ for $\underline{\omega} < \omega \leqslant \pi$. (Assume that the corresponding negative frequencies are also passed.) Choosing a symmetric filter, $c_j = c_{-j}$, will make the phase 0 for all ω. In this case, $c(\mathrm{e}^{i\omega})$ is a real number, and so the gain is given by:

$$g(\omega) = \sum_{j=-\infty}^{\infty} c_{|j|}\mathrm{e}^{-i\omega j}$$

Now, using Eqn. 6, it is straightforward to verify the identity

$$\begin{aligned} c_j &= (2\pi)^{-1}\int_{-\pi}^{\pi} \mathrm{e}^{-i\omega j} \sum_{j=-\infty}^{\infty} c_j \mathrm{e}^{-i\omega j}\,\mathrm{d}\omega \\ &= (2\pi)^{-1}\int_{-\pi}^{\pi} \mathrm{e}^{i\omega j} c(\mathrm{e}^{i\omega})\,\mathrm{d}\omega \end{aligned}$$

Replacing $c(\mathrm{e}^{i\omega}) = g(\omega)$ with the target value of the gain for the band-pass filter and carrying out the integration yields

$$c_j = (2\pi)^{-1}\frac{1}{ij}\,\mathrm{e}^{i\omega j}\Big|_{-\underline{\omega}}^{\underline{\omega}} = \left\{\begin{array}{ll} \dfrac{1}{j\pi}\sin(\underline{\omega} j) & \text{for} \quad j \neq 0 \\ \underline{\omega}/\pi & \text{for} \quad j = 0 \end{array}\right\}$$

The resulting filter $c(L)$ passes all components with frequencies between $0 \leqslant \omega \leqslant \underline{\omega}$. This means that the filter $1-c(L)$ passes everything except these frequencies. The filter $c(L)$ is called a low-pass filter (it passes low frequencies), and $1-c(L)$ is called a high-pass filter (it passes the high frequencies). Combinations of high- and low-pass filters can be used to construct band-pass filters for any set of frequencies. For example, a filter that passes components with frequencies between ω_1 to ω_2 (with $\omega_2 \geqslant \omega_1$) can be

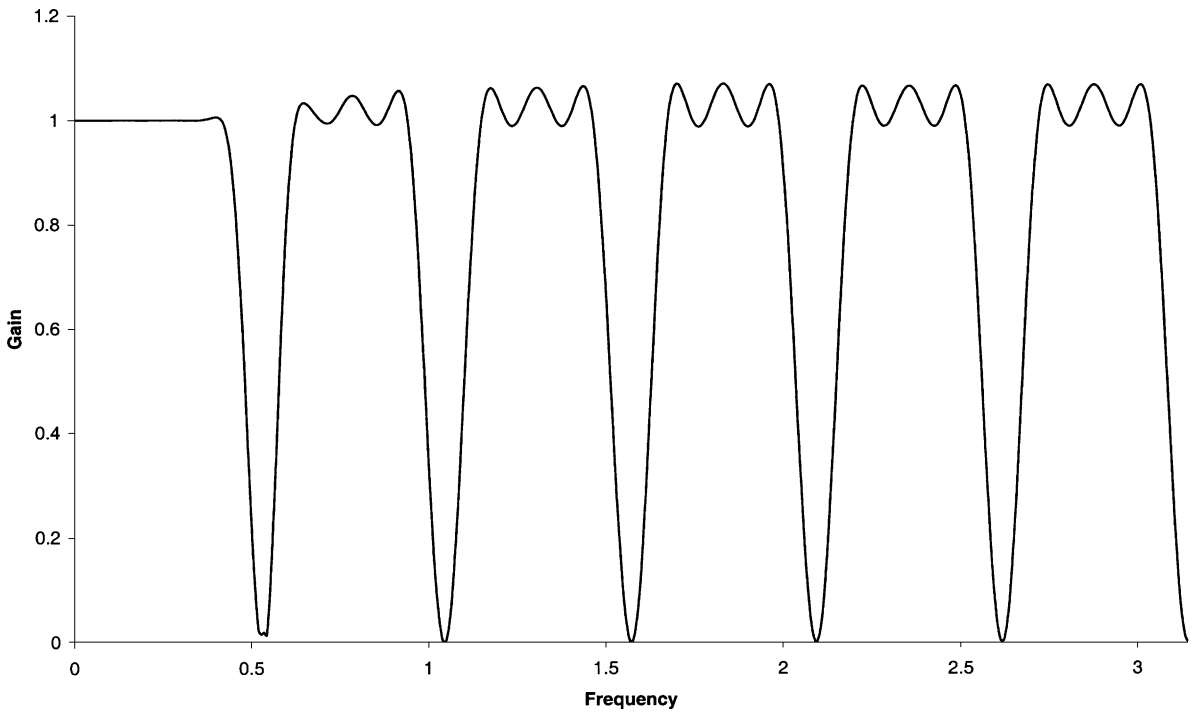

Figure 3
Gain of linear approximation to monthly × 12 seasonal adjustment filter

constructed as the difference of the low-pass filter for ω_2 and the low-pass filter for ω_1.

One practical problem with exact band-pass filters is that the coefficients c_j die out very slowly (at the rate $1/j$). This introduces important 'endpoint' problems when applying these filters to finite realizations of data, say $\{Y_t\}_{t=1}^{T}$. One approach is simply to truncate the filter at some point, for example to use $c_k(L) = \sum_{j=-k}^{k} c_j L^j$ and apply this filter to $\{Y_t\}_{t=k+1}^{T-k}$. An alternative procedure is to construct a minimum mean square error estimate of the infeasible band-pass values:

$$X_t = c(L)Y_t = \sum_{i=-\infty}^{\infty} c_i Y_{t-i},$$

as

$$E(X_t \mid \{Y_j\}_{j=1}^{T}) = \sum_{i=-\infty}^{\infty} c_i E(Y_{t-i} \mid \{Y_j\}_{j=1}^{T})$$

This can be accomplished by using backcasts and forecasts for the missing pre-sample and post-samples values of Y_t. The relative merits of these two approaches for band-pass filters is discussed in Baxter and King (1999), Geweke (1978) and Dagum (1980) discuss this problem in the context of seasonal adjustment.

4. *Spectra of Commonly Used Stochastic Processes*

Suppose that Y has spectrum $S_Y(\omega)$ and $X_t = c(L)Y_t$. What is the spectrum of X? As was shown in Eqn. 8, the frequency components of X are the frequency components of Y scaled by the factor $g(\omega)e^{-i\rho(\omega)}$, where $g(\omega)$ is the gain and $\rho(\omega)$ is the phase of $c(L)$. This means that spectra of X and Y are related by:

$$S_X(\omega) = g(\omega)^2 S_Y(\omega) = |\, c(e^{-i\omega})\,|^2 S_Y(\omega)$$

which follows from $|e^{-i\rho(\omega)}| = 1$ and the definition of $g(\omega)$.

Now, suppose that ε_t is a 'white noise' process, defined by the properties $E(\varepsilon_t) = 0$, $\lambda_0 = \sigma^2$, and $\lambda_k = 0$ for $k \neq 0$. The spectrum of ε is then easily calculated from Eqn. 7:

$$S_\varepsilon(\omega) = (2\pi)^{-1}\sigma^2$$

So, the spectrum of white noise is constant. (Which is why the process is called 'white' noise.)

Now suppose that $Y_t = c(L)\varepsilon_t$. Then:

$$S_Y(\omega) = |c(e^{-i\omega})|^2 S_\varepsilon(\omega) = |c(e^{-i\omega})|^2 (2\pi)^{-1}\sigma^2$$

This result can be used to determine the spectrum of any stationary ARMA process (see *Time Series:*

ARIMA Methods) for a detailed discussion of these models). If Y_t follows an ARMA process, then it can be represented as:

$$\phi(L)Y_t = \theta(L)\varepsilon_t$$

The autoregressive operator, $\phi(L)$ can be inverted to yield $Y_t = c(L)\varepsilon_t$ with $c(L) = \phi(L)^{-1}\theta(L)$. This means that:

$$S_y(\omega) = |c(\mathrm{e}^{-i\omega})|^2(2\pi)^{-1}\sigma^2 = \frac{2(\pi)^{-1}\sigma^2|\theta(\mathrm{e}^{-i\omega})|^2}{|\phi(\mathrm{e}^{-i\omega})|^2} \qquad (11)$$

As an example, consider the AR(1) model:

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

equivalently written as,

$$(1-\phi L)Y_t = \varepsilon_t.$$

Applying Eqn. 11 yields

$$S_Y(\omega) = \frac{\sigma^2}{2\pi}\frac{1}{|1-\phi\mathrm{e}^{-i\omega}|^2} = \frac{\sigma^2}{2\pi}\frac{1}{(1+\phi^2-2\phi\cos(\omega))}$$

This spectrum is equal to $\sigma^2[2\pi(1+\phi^2-2\phi\cos(\omega))]^{-1}$ at $\omega = 0$. When $0 < \phi < 1$, it falls steadily as ω increases from 0 to π. This means that, relative to white noise, the low frequency components of the AR(1) are more important than the high frequency components. Thus, realizations of the series appear smoother than white noise.

5. *Spectral Estimation*

There are two general approaches to estimating spectra. Perhaps the simplest is to estimate an ARMA model for the series (as explained in *Time Series: ARIMA Methods*) and then compute the implied spectrum for this estimated model. Asymptotic approximations can then be used for statistical inference. Since the estimators of the ARMA parameters are asymptotically normally distributed and the spectrum is a smooth function of these parameters (see Eqn. 11), the estimated spectrum at any point ω will be asymptotically normal with a variance that can be computed using the δ-method (a mean-value expansion).

An interesting extension of this procedure is discussed in Berk (1974). He shows that, quite generally, a suitably long autoregression can be specified and the spectrum estimated using the estimated AR coefficients. The precise definition of 'suitably long' depends on the size of the available sample, with more terms included in larger samples. These 'autoregressive spectral estimators' are shown by Berk to be consistent and approximately normally distributed in large samples.

An alternative set of estimators are based on nonparametric methods. These estimators can be motivated by Eqn. 7 which shows the spectrum as a weighted sum of the autocovariances. This suggests an estimator of the form:

$$S(\omega) = (2\pi)^{-1}\sum_{k=-l_T}^{l_T} w_T(k)\hat{\lambda}_k$$

where $\hat{\lambda}_k$ is an estimator of λ_k, l_T are truncation values that depend on the sample size T, and $w_T(k)$ is a weight function that also depends on T. These estimators can be studied using nonparametric methods (see *Nonparametric Statistics: Asymptotics*).

A closely related set of nonparametric estimators are based on the 'periodogram'. The periodgram provides a frequency decomposition of the sample variance of a partial realization of the process, say $\{Y_t\}_{t=1}^T$. The periodgram is computed at frequencies $\omega_j = 2\pi j/T$, $j = 1, 2,\ldots, T/2$ (assuming T is even), and the value of the periodgram at frequency ω_j is:

$$p_j = \frac{2}{T}\left|\sum_{t=1}^{T}(y_t-\bar{y})\mathrm{e}^{-i\omega_j t}\right|^2$$

The periodgram has three interesting asymptotic sampling properties: (a) $E(p_j)/4\pi \to \mathrm{S}_Y(\omega_j)$; (b) $var(p_j/2\pi) \to \mathrm{S}_Y(\omega_j)^2$; and (c) $cov(p_j, p_k) = 0$ for $j \neq k$. Property (a) shows that a scaled version of the periodgram provides an asymptotically unbiased estimator of the spectrum, but since the variance doesn't approach zero (property (b)), the estimator is not consistent. The third property shows that averaging a set of periodgram ordinates around a given frequency reduces the variance of the estimator, although (from property (a)) it may introduce bias. By carefully selecting the averaging weights, consistent estimators with good finite sample mean square error properties can be constructed.

Some recent advances for nonparametric spectral estimators are developed in Andrews (1991).

See also: Nonparametric Statistics: Asymptotics; Time Series: ARIMA Methods; Time Series: Seasonal Adjustment

Bibliography

Andrews D W K 1991 Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**: 817–58

Baxter M B, King R G 1999 Measuring business cycles: approximate band-pass filters for economic time series. *Review of Economics and Statistics* **81**(4): 575–93

Berk K N 1974 Consistent autoregressive spectral estimates. *Annals of Statistics* **2**(3): 489–502

Brockwell P J, Davis R A 1991 *Time Series: Theory and Methods*, 2nd edn. Springer, New York
Cramér H 1942 On harmonic analysis of certain function spaces. *Arkiv für Matematik, Astronomi och Fysik*. **28B**(12): 1–7
Dagum E B 1980 *The X-11 ARIMA Seasonal Adjustment Method*. Research Paper, Statistics, Canada
Geweke J F 1978 The revision of seasonally adjusted time series. *Proceedings of the Business and Economic Statistics Section—American Statistical Association*, pp. 320–25
Granger C W J, Newbold P 1977 *Forecasting Economic Time Series*. Academic Press, New York
Hamilton J 1994 *Time Series Analysis*. Princeton University Press, Princeton, NJ
Harvey A 1993 *Time Series Models*, 2nd edn. MIT Press, Cambridge, UK
Kolmorgorov A N 1940 Kurven in Hilbertschen Raum die gegenüber eine einparametrigen Gruppe von Bewegungen invariant sind. *C. R. (Doklady) de L'Academie des Sciences de l'IRSS, New Series* **26**: 6–9
Nerlove M, Grether D M, Carvalho J L 1979 *Analysis of Economic Time Series*. Academic Press, New York
Priestly M B 1981 *Spectral Analysis and Time Series*. Academic Press, New York
Wallis K F 1974 Seasonal adjustment and relations between variables. *Journal of the American Statistical Association* **69**: 18–31

M. W. Watson

Time Series: Economic Forecasting

Time-series forecasts are used in a wide range of economic activities, including setting monetary and fiscal policies, state and local budgeting, financial management, and financial engineering. Key elements of economic forecasting include selecting the forecasting model(s) appropriate for the problem at hand, assessing and communicating the uncertainty associated with a forecast, and guarding against model instability.

1. *Time Series Models for Economic Forecasting*

Broadly speaking, statistical approaches to economic forecasting fall into two categories: time-series methods and structural economic models. Time-series methods use economic theory mainly as a guide to variable selection, and rely on past patterns in the data to predict the future. In contrast, structural economic models take as a starting point formal economic theory and attempt to translate this theory into empirical relations, with parameter values either suggested by theory or estimated using historical data. In practice, time-series models tend to be small with at most a handful of variables, while structural models tend to be large, simultaneous equation systems which sometimes incorporate hundreds of variables (see *Economic Panel Data*; *Simultaneous Equation Estimation: Overview*). Time-series models typically forecast the variable(s) of interest by implicitly extrapolating past policies into the future, while structural models, because they rely on economic theory, can evaluate hypothetical policy changes. In this light, perhaps it is not surprising that time-series models typically produce forecasts as good as, or better than, far more complicated structural models. Still, it was an intellectual watershed when several studies in the 1970s (reviewed in Granger and Newbold 1986) showed that simple univariate time-series models could outforecast the large structural models of the day, a result which continues to be true (see McNees 1990). This good forecasting performance, plus the relatively low cost of developing and maintaining time-series forecasting models, makes time-series modeling an attractive way to produce baseline economic forecasts.

At a general level, time-series forecasting models can be written,

$$y_{t+h} = g(X_t, \theta) + \varepsilon_{t+h} \tag{1}$$

where y_t denotes the variable or variables to be forecast, t denotes the date at which the forecast is made, h is the forecast horizon, X_t denotes the variables used at date t to make the forecast, θ is a vector of parameters of the function g, and ε_{t+h} denotes the forecast error. The variables in X_t usually include current and lagged values of y_t. It is useful to define the forecast error in (1) such that it has conditional mean zero, that is, $E(\varepsilon_{t+h}|X_t) = 0$. Thus, given the predictor variables X_t, under mean-squared error loss the optimal forecast of y_{t+h} is its conditional mean, $g(X_t, \theta)$. Of course, this forecast is infeasible because in practice neither g nor θ are known. The task of the time-series forecaster therefore is to select the predictors X_t, to approximate g, and to estimate θ in such a way that the resulting forecasts are reliable and have mean-squared forecast errors as close as possible to that of the optimal infeasible forecast.

Time-series models are usefully separated into univariate and multivariate models. In univariate models, X_t consists solely of current and past values of y_t. In multivariate models, this is augmented by data on other time series observed at date t. The next subsections provide a brief survey of some leading time-series models used in economic forecasting. For simplicity attention is restricted to one-step ahead forecasts ($h = 1$). Here, the focus is on forecasting in a stationary environment; the issue of nonstationarity in the form of structural breaks or time varying parameters is returned to below.

1.1 *Univariate Models*

Univariate models can be either linear, so that g is linear in X_t, or nonlinear. All linear time-series models

can be interpreted as devices for modeling the covariance structure of the data. If $\{y_t\}$, $t = 1, \ldots, T$ has a Gaussian distribution and if this distribution is stationary (does not depend on time), then the optimal forecast is a linear combination of past values of the data with constant weights. Different linear time-series models provide different parametric approximations to this optimal linear combination.

The leading linear models are autoregressive models, autoregressive–integrated moving-average (ARIMA) models, and unobserved components models.

An autoregressive model of order p (AR(p)) is written,

$$y_{t+1} = \mu + \alpha_1 y_t + \cdots + \alpha_p y_{t-p+1} + \varepsilon_{t+1} = \mu + \alpha(L) y_t + \varepsilon_{t+1},$$

where $(\mu, \alpha_1, \ldots, \alpha_p)$ are unknown parameters, L is the lag operator, and $\alpha(L)$ is a lag polynomial. If y_t is observed at all dates, $t = 1, \ldots, T$, these parameters are readily estimated by ordinary least squares.

ARIMA models extend autoregressive models to include a moving-average term in the error, which has the effect of inducing long lags in the forecast when it is written as a function of current and past values of y_t. These are discussed in *Time Series: ARIMA Methods*.

In an unobserved components model, y_t is represented as the sum of two or more different stochastic components. These components are then modeled, and the combined model provides a parametric approximation to the autocovariances; see Harvey (1989) for a detailed modern treatment. The unobserved components framework can provide a useful framework for extracting cyclical components of economic time-series and for seasonal adjustment (see *Time Series: Seasonal Adjustment*).

If $\{y_t\}$ is not Gaussian, in general the optimal forecast will not be linear, which suggests the use of forecasts based on nonlinear time-series models. Parametric nonlinear time-series models posit a functional form g, and take this as known up to the finite-dimensional parameter vector θ. Two leading examples of parametric nonlinear time-series models, currently popular for economic forecasting, are threshold autoregressive models and Markov switching models. Both are similar in the sense that they posit two (or more) regimes. In the threshold autoregression, switches between the regimes occur based on past values of the observed data; in Markov switching models, the switches occur based on an unobserved or latent variable.

The threshold autoregressive family of models can be written

$$y_{t+1} = d_t \alpha(\mathrm{L}) y_t + (1 - d_t) \beta(\mathrm{L}) y_t + \varepsilon_{t+1} \qquad (2)$$

where the mean is suppressed, $\alpha(\mathrm{L})$ and $\beta(\mathrm{L})$ are lag polynomials, and d_t is a nonlinear function of past data that switches between the 'regimes' $\alpha(\mathrm{L})$ and $\beta(\mathrm{L})$. Variations of Eqn. (2) interact d_t with only some autoregressive coefficients, the intercept, and/or the error variance. Various functions are available for d_t, either a sharp indicator function (the threshold autoregressive model) or a smooth function (smooth transition autoregression). For example, setting $d_t = (1 + \exp[\gamma_0 + \gamma'_1 \zeta_t])^{-1}$, yields the logistic smooth transition autoregression (LSTAR) model, where ζ_t denotes current or past data, say, ζ_t might equal y_{t-k} for some k. See Granger and Teräsvirta (1993) for additional details. For an application of threshold autoregressions (and other models) to forecasting US unemployment, see Montgomery et al. (1998).

The Markov switching models are conceptually similar, except that the regime switch depends on an unobserved series. In the context of Eqn. (2), d_t is modeled as unobserved and following a two-state Markov process; see Hamilton (1994).

An alternative approach is to treat g itself as unknown, which leads to nonparametric methods. These include nearest neighbor, kernel, and artificial neural network models for the conditional expectation. In practice these nonparametric approaches to economic forecasting have met with success which is mixed at best. There are several possible reasons for this, including that there are insufficiently many observations in typical economic data sets to support the use of these data-intensive methods.

1.2 Multivariate Models

In multivariate time-series models, X_t includes multiple time-series that can usefully contribute to forecasting y_{t+1}. The choice of these series is typically guided by both empirical experience and by economic theory, for example, the theory of the term structure of interest rates suggests that the spread between long and short term interest rates might be a useful predictor of future inflation.

The multivariate extension of the univariate autoregression is the vector autoregression (VAR), in which a vector of time-series variables, Y_{t+1}, is represented as a linear function of $Y_t, \ldots, Y_{t-p+1}$, perhaps with deterministic terms (an intercept or trends). An interesting possibility arises in VARs that is not present in univariate autoregressions, specifically, it might be that the time-series are cointegrated (that is, the individual series are nonstationary in the sense that they are integrated, but linear combinations of the series are integrated or order zero); see *Time Series: Co-integration* and Watson (1994).

Multivariate time-series models involve a large number of unknown parameters, a problem which is greatly exacerbated when nonlinearities are introduced. Conceptually, the extension of univariate nonlinear models to the multivariate setting is straightforward. In practice, however, because of the relatively

small number of time-series observations available to economic forecasters, it is unclear how best to implement nonlinear multivariate models and there are currently no definite conclusions in this area.

1.3 Models for Forecasting Higher Moments

The foregoing categorization of forecasting models has focused on prediction of conditional means. In some economic applications, especially in financial economics, interest is also in forecasting conditional higher moments. The most developed tools in this area are models of conditional variances. Two families of such models are autoregressive conditional heteroskedasticity (ARCH) models (Engle 1982; see Bollerslev et al. 1994) and stochastic volatility models. Both approaches provide parametric models for the volatility of the series, which can then be forecasted given parameters estimated from historical data.

2. When Do Time-series Forecasts Fail?

Regrettably, only by luck will economic forecasts be 'right': there are inevitable sources of uncertainty that make perfectly accurate predictions of the future impossible.

2.1 Shock, Model, and Estimation Uncertainty

Sources of uncertainty that comprise the error term ε_{t+h} in (1) ('shock uncertainty') consist of unknowable future events, random measurement error in the series being forecast, and unforeseeable data revisions. Additional forecast uncertainty arises from model approximation error (that is, error in the choice of g) and from error in the estimation of the model parameters θ.

Sensible use of economic forecasts requires that this forecast uncertainty be accurately communicated. Prediction intervals can be computed either directly based on historically estimated parameters and residuals, or by simulation for more complicated nonlinear models. Ideally these prediction intervals should incorporate parameter estimation uncertainty. There are several ways to do this, but a natural one is to construct forecast error distributions using a simulated out of sample prediction methodology. In this approach, the model is estimated using data through some date, say t', then a forecast is made for the next period ($t'+h$, for h-step ahead forecasts). The model is then reestimated using data through date $t'+1$, and the forecast is made for $t'+h+1$. This is repeated until a series of simulated out of sample forecast errors is constructed, which in turn can be used to construct prediction intervals around the point forecast of interest.

2.2 Overfitting

Overfitting poses a particular threat in economic forecasting given the impossibility of creating new data sets through experiments and given the relatively short number of observations on many economic time-series. Overfitting can be addressed in part by relying on automatic methods for pruning a model. Leading such methods are model selection using information criteria such as the Akaike or Bayes information criteria (see *Model Testing and Selection, Theory of*). Alternatives include Bayes or Empirical Bayes estimation. Combining forecasts by taking averages or weighted averages can also be interpreted as a method for addressing the overfitting that arises in model selection.

2.3 Difficulties Modeling Trends

An important characteristic of economic data is that many economic time-series contain considerable persistence, sometimes resembling a trend. Starting with the seminal work of Nelson and Plosser (1982), there has been considerable debate about whether these trends are best modeled as deterministic or stochastic (that is, arising from unit autoregressive roots), see Stock (1994). If a forecasting model incorrectly incorporates deterministic trends, but the series is highly persistent, spuriously large weight can be placed on an estimated trend component and out-of-sample forecasts can suffer. Such considerations lead to the use of methods that explicitly address the possibilities of stochastic trends, an example being preliminarily testing for a unit root in a univariate time-series and constructing a univariate model in levels or in differences depending on the outcome of the pretest.

2.4 Model Instability

Clements and Hendry (1999) argue that many major failures of economic forecasts arise because of structural breaks to which conventional forecasting models fail to adapt. In some cases, model instability can arise because of slow changes in parameter values, perhaps arising from evolution in underlying technologies in the economy. In other, more dramatic cases, model instability appears as a sudden break, such as a change in monetary policy or the collapse of an exchange rate regime.

Whether structural change is gradual or abrupt, it is critical to monitor forecast performance and to test for model instability. Various statistical methods are available for detecting structural change; these typically involve testing for evidence of instability either in estimated coefficients or in simulated forecast errors. In practice, these methods need to be augmented by expert judgment that draws on information not incorporated in the model, such as knowledge of

particularly severe weather (which could lead to a structural break being spuriously detected) or of a known change in a policy regime (so a break might be imposed even if it is not detected statistically). Unfortunately, expert judgment can be wrong, and knowing when to intervene in an economic forecast is a rather murky art.

3. Which Methods Work Best?

It is perhaps overly ambitious to hope for a simple prescription that provides a good forecasting method for all economic time-series. Nonetheless, a number of studies have attempted to provide some general guidance on this question. Two general lessons emerge: among univariate models, simple linear models often forecast as well or better than more complicated nonlinear models; and the gains from moving from univariate to multivariate models are often (but not always) small.

This is illustrated by forecasts of the rate of price inflation over six months in the United States, as measured by the percentage rate of inflation of the Consumer Price Index (this example is taken from Stock and Watson's (1999) comparison of linear and nonlinear forecasting models for 215 US macroeconomic time-series). Simulated out of sample six-month-ahead forecasts over the period March 1971 to June 1996 (the full sample begins January 1959) produce a root mean-square forecast error (RMSFE) of 2.44 percentage points (at an annual rate) for a univariate AR(4) fit recursively to the annualized rate of inflation. If the autoregressive lag length is selected recursively using the Bayes information criterion and a recursive unit root protest is used to determine whether the AR is fit to the rate of inflation or the change in the rate, the RMSFE drops to 2.05. Forecasts based on artificial neural networks and LSTAR models have RMSFEs exceeding 2.15 and in some cases exceeding 2.5, depending on the specification. A VAR which adds the unemployment rate and short-term interest rate (recursive Bayes information criterion lag selection) produces a RMSFE of 2.31. For these models, then, introducing nonlinearity or adding these other predictors does not improve upon a simple autoregression with a unit root pretest and data-dependent lag length selection.

The future of economic forecasting promises the development of new, computer-intensive methods and the real-time availability of very large data sets with complex patterns of historical dependence. These developments could prove invaluable for advancing economic forecasting, and the lessons from the past four decades of economic forecasting might not extend to these new datasets. Yet those lessons do underscore the importance of having a simple and reliable benchmark forecast against which alternative forecasts can be assessed.

See also: Econometrics, History of; Explanation: Conceptions in the Social Sciences; Quantification in the History of the Social Sciences; Time Series: Advanced Methods; Time Series: General

Bibliography

Bollerslev T, Engle R F, Nelson D B 1994 ARCH Models. In: Engle R, McFadden D (eds.) *Handbook of Econometrics*, Elsevier, Amsterdam, Vol. IV, pp. 2959–3038

Clements M P, Hendry D F 1999 *Forecasting Non-Stationary Economic Time Series*. MIT Press, Cambridge, MA

Engle R F 1982 Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica* **50**: 987–1008

Granger C W J, Newbold P 1986 *Forecasting Economic Time Series*, 2nd edn. Academic Press, Orlando, FL

Granger C W J, Teräsvirta T 1993 *Modelling Non-linear Economic Relationships*. Oxford University Press, Oxford, UK

Hamilton J D 1994 *Time Series Analysis*. Princeton University Press, Princeton, NJ

Harvey A C 1990 *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, UK

McNees S K 1990 The role of judgment in macroeconomic forecasting accuracy. *International Journal of Forecasting* **6**: 287–99

Montgomery A L, Zarnowitz V, Tsay R S, Tiao G C 1998 Forecasting the US unemployment rate. *Journal of the American Statistical Association* **93**: 478–93

Nelson C R, Plosser C I 1982 Trends and random walks in macroeconomic time series—some evidence and implications. *Journal of Monetary Economics* **10**: 139–62

Stock J H 1994 Unit roots, structural breaks, and trends. In: Engle R F, McFadden D (eds.) *Handbook of Econometrics*. Elsevier: Amsterdam, Vol. IV, pp. 2740–843

Stock J H, Watson M W 1999 A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: Engle R F, White H (eds.) *Cointegration, Causality, and Forecasting: a festschrift in honour of Clive W. J. Granger*. Cambridge University Press, Cambridge, UK, pp. 1–44

Watson M W 1994 Vector Autoregressions and cointegration. In: Engle R, McFadden D (eds.) *Handbook of Econometrics*. Elsevier, Amsterdam, Vol. IV, pp. 2844–915

J. H. Stock

Time Series: General

A *time series* is a stretch of values on the same scale indexed by a time-like parameter. The basic data and parameters are functions.

Time series take on a dazzling variety of shapes and forms, indeed there are as many time series as there are functions of real numbers. Some common examples of time series forms are provided in Fig. 1. One notes periods, trends, wandering and integer-values. The time series such as those in the figure may be

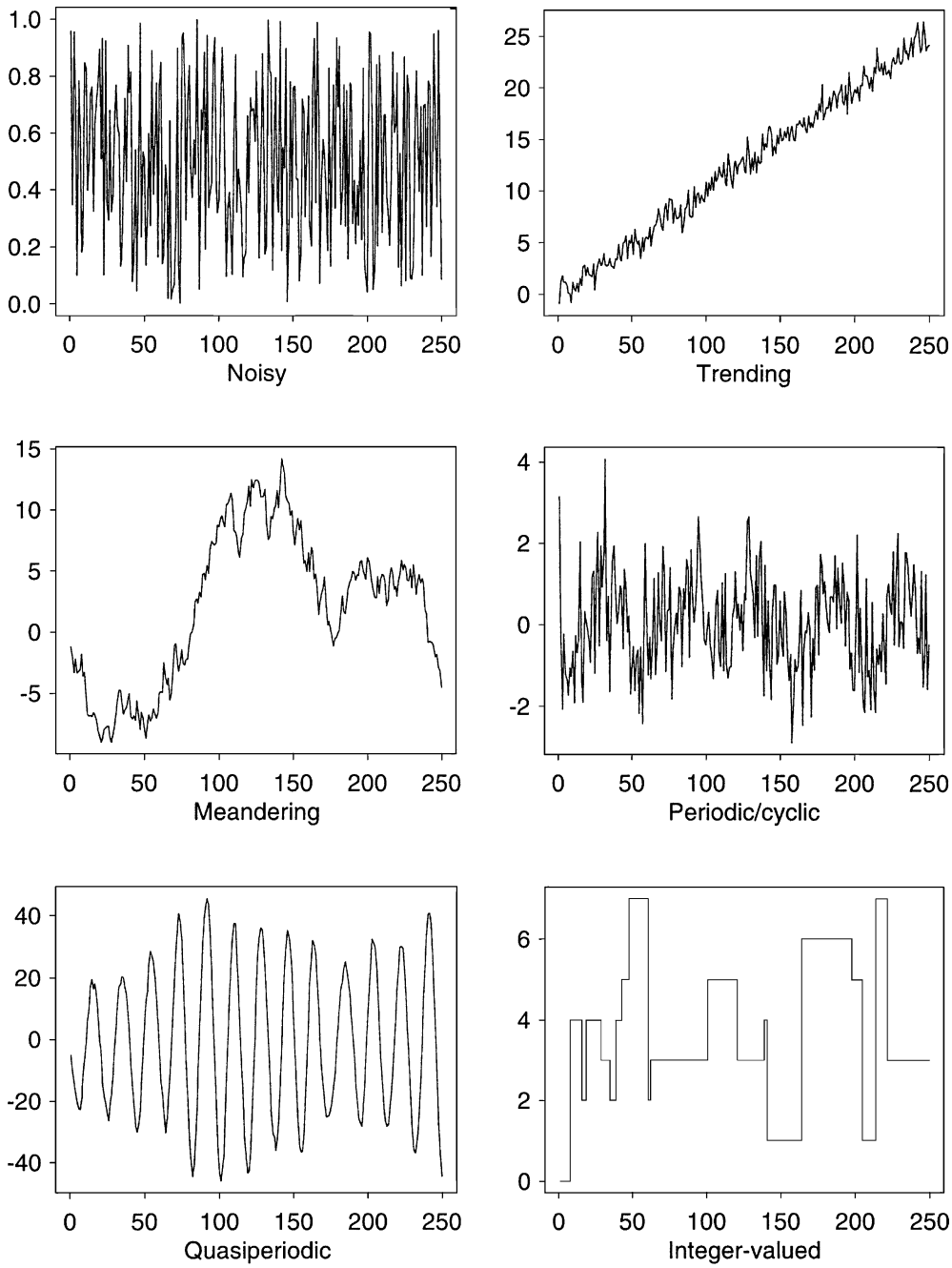

Figure 1
Some different types of time series

contemporaneous and a goal may be to understand the interrelationships.

Concepts and fields related to time series include: logitudinal data, growth curves, repeated measures, econometric models, multivariate analysis, signal processing, and systems analysis.

The field, time series analysis, consists of the techniques which when applied to time series lead to

improved knowledge. The purposes include summary, decision, description, prediction.

The field has a theoretical side and an applied side. The former is part of the theory of stochastic processes (e.g., representations, prediction, information, limit theorems) while applications often involve extensions of techniques of 'ordinary' statistics, for example, regression, analysis of variance, multivariate analysis, sampling. The field is renowned for jargon and acronyms—white noise, cepstrum, ARMA, ARCH (see Granger 1982).

1. Importance

'... but time and chance happeneth to them all' (*Ecclesiastes*).

Time series ideas appear basic to virtually all activities. Time series are used by nature and humans alike for communication, description, and visualization. Because *time* is a physical concept, parameters and other characteristics is mathematical models for time series can have real-world interpretations. This is of great assistance in the analysis and synthesis of time series.

Time series are basic to scientific investigations. There are: circadian rhythms, seasonal behaviors, trends, changes, and evolving behavior to be studied and understood. Basic questions of scientific concern are formulated in terms of time series concepts—Predicted value? Leading? Lagging? Causal connection? Description? Association? Autocorrelation? Signal? Seasonal effect? New phenomenon? Control? Periodic? Changing? Trending? Hidden period? Cycles?

Because of the tremendous variety of possibilities, substantial simplifications are needed in many time series analyses. These may include assumptions of stationarity, mixing or asymptotic independence, normality, linearity. Luckily such assumptions often appear plausible in practice.

The subject of time series analysis would be important if for no other reason than that it provides means of examining the basic assumption of statistical independence invariably made in ordinary statistics. One of the first commonly used procedures for this problem was the Durbin–Watson test (Durbin and Watson 1950). The autocovariance and spectrum functions, see below, are now often used in this context.

2. History

Contemporary time series analysis has substantial beginnings in both the physical and social sciences. Basic concepts have appeared in each subject and made their way to the other with consequent transfer of technology. Historical researchers important in the development of the field include: Lauritzen (1981), Hooker (1901), Einstein (1914), Wiener (1949), Yule (1927), Fisher (1929), Tukey (1984), Whittle (1951), Bartlett (1966). Books particularly influential in the social sciences include Moore (1914) and Davis (1941). Nowadays many early analyses appear naive. For example, Beveridge in 1920 listed some 19 periods for a wheat price index running from 1500 to 1869 (Beveridge 1922). It is hard to imagine the presence of so much regular behavior in such a series. When statistical uncertainty is estimated using latter-day techniques most of these periods appear insignificant; see the periodogram with 95 percent error bounds in Fig. 2. Many historical references are included in Wold (1965).

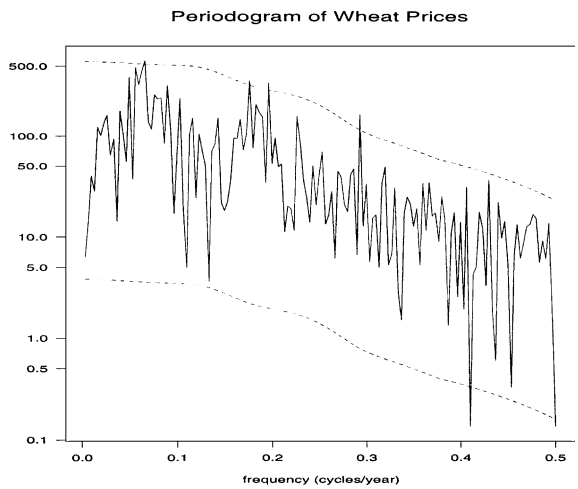

Figure 2
The dashed lines provide 95 percent bounds about a central curve

Historians of science have made some surprising discoveries concerning early work with time series. An example is presented in Tufte (1983). He shows a time series plot from the tenth or eleventh century AD. This graph is speculated to provide movements of the planets and sun. It is remarkable for its cartesian type character. More generally Tufte remarks, following a study of newspapers and magazines, that 'The time-series plot is the most frequently used form of graphic design.' Casual observation certainly supports Tufte's study.

Important problems that were addressed in the twentieth century include: seasonal adjustment, leading and lagging indicators, and index numbers. Paradigms that were developed include:

$$\text{series} = \text{signal} + \text{noise}$$

$$\text{series} = \text{trend} + \text{seasonal} + \text{noise}$$

$$\text{series} = \text{sum of cosines} + \text{noise}$$

$$\text{series} = \text{regression function} + \text{noise}$$

These conceptualizations have been used for forecasting, seasonal adjustment and description amongst other things. There are surprises, for example, ordinary least squares is surprisingly efficient in the time series case (see Grenander and Rosenblatt 1957). Other books from the 1950s and 1960s that proved important are Blackman and Tukey (1959), Granger and Hatanaka (1964), Box and Jenkins (1970). Important papers from the era include: Akaike (1965), Hannan (1963) and Parzen (1961).

An important development that began in the 1950s and continues today is the use of recursive computation in the state space model; see Kalman (1963), Harvey (1990), Shumway and Stoffer (2000), and Durbin and Koopman (2001).

3. Basics

3.1 Concepts

There are a number of concepts that recur in time series work. Already defined is the *time series*, a stretch of values on the same scale indexed by a time parameter. The time parameter may range over the positive and negative integers or all real numbers or subsets of these. *Time series data* refer to a segment of a time series. A time series is said to have a *trend* when there is a slowly evolving change. It has a *seasonal* component when some cyclic movement of period one year is present. (The *period* of a cyclic phenomenon is the amount of time for it to repeat itself. Its *frequency* is the reciprocal of the period.)

There is an algebra of mathematical operations that either nature or an analyst may apply to a time series $\{X(t)\}$ to produce another series $\{Y(t)\}$. Foremost is the *filter*, or *linear time invariant operation* or *system*. In the case of discrete equispaced time this may be represented as

$$Y(t) = \sum_u a(u)X(t-u)$$

$t, u = 0, \pm 1, \ldots$ An example is the running mean used to smooth a series. The functions $X(.)$, $Y(.)$ may be vector-valued and $a(.)$ may be matrix-valued. In the vector-valued case feedback may be present. The sequence $\{a(u)\}$ is called the *impulse response function*. The operation has the surprising property of taking a series of period P into a series of the same period P. The filter is called *realizable* when $a(u) = 0$ for $u < 0$. Such filters appear in causal systems and when the goal is prediction.

The above ideas concern both deterministic and random series. The latter prove to be important in developing solutions to important problems. Specifically it often proves useful to view the subject of time series as part of the general theory of stochastic processes, that is, indexed families of random variables. One writes $\{Y(t, \omega),\ t \text{ in } V\}$, with ω a random variable and V a set of times. Time series data are then viewed as a segment, $\{Y(t, \omega_0),\ t = 0, \ldots, \mathrm{T}-1\}$ of the realization labeled by ω_0, the obtained value of ω. Stochastic time series $\{Y(t),\ t = 0, \pm 1, \pm 2, \ldots\}$ sometimes are conveniently described by *finite dimensional distributions* such as

$$\text{Prob}\{Y(t) \leqslant y\} \quad \text{and} \quad \text{Prob}\{Y(t_1) \leqslant y_1, Y(t_2) \leqslant y_2\}$$

This is particularly the case for time series values with joint normal distributions.

Time series may also be usefully described or generated by stochastic models involving the independent identically distributed random variables of ordinary statistics.

Stochastic models may be distinguished as parametric or nonparametric. Basic parameters of the nonparametric approach include: moments, joint probability and density functions, mean functions, autocovariance and crosscovariance functions, power spectra.

One basic assumption of time series analysis is that of *stationarity*. Here the choice of time origin does not affect the statistical properties of the process. For example, the mean level of a stationary series is constant. Basic to time series analysis is handling temporal dependence. To this end one can define the *crosscovariance function* of the series X and Y at lag u as the covariance of the values $X(t+u)$ and $Y(t)$. In the stationary case this function does not depend on t. In an early paper, Hooker (1901) computed an estimate of this quantity. Another useful parameter is the *power spectrum*, a display of the intensity or variability of the phenomenon vs. period or frequency. It may be defined as the Fourier transform of the autocovariance function. The power spectrum proves useful in displaying the serial dependence present, in discovering periodic phenomena and in diagnosing possible models for a series.

In the parametric case there are the *autoregressive moving average* (ARMA) series. These are regression-type models expressing the value $Y(t)$ as a linear function of a finite number of past values $Y(t-1)$, $Y(t-2)$, ... of the series and the values $\varepsilon(t), \varepsilon(t-1), \ldots$ of a sequence of independent identically distributed random variables. ARMAs have proved particularly useful in problems of forecasting.

Contemporary models for time series are often based upon the idea of *state*. This may be defined as the smallest entity summarizing the past history of the process. There are two parts to the state space model. First, a state equation that describes how the value of the state evolves in time. This equation may contain a purely random element. Second there is an equation indicating how the available measurements at that time t come from the state value at that time. It too may involve a purely random element. The concept of

the state of a system has been basic to physics for many years in classical dynamics and quantum mechanics. The idea was seized upon by control engineers, for example, Kalman (1963) in the late 1950s. The econometricians realized its usefulness in the early 1980s (see Harvey 1990).

A number of specific probability models have been studied in depth, including the Gaussian, the ARMA, the bilinear (Granger 1978), various other nonlinear (Tong 1990), long and short memory, ARMAX, ARCH (Einstein 1914), hidden Markov (MacDonald and Zucchini 1997), random walk, stochastic differential equations (Guttorp 1995, Prakasa Rao 1999), and the periodically stationary.

A list of journals where these processes are often discussed is included at the end of this article.

3.2 Problems

There are scientific problems and there are associated statistical problems that arise. Methods have been devised for handling many of these. The scientific problems include: smoothing, prediction, association, index numbers, feedback, and control. Specific statistical problems arise directly. Among these are including explanatories in a model, estimation of parameters such as hidden frequencies, uncertainty computation, goodness of fit, and testing.

Special difficulties arise. These include: missing values, censoring, measurement error, irregular sampling, feedback, outliers, shocks, signal-generated noise, trading days, festivals, changing seasonal pattern, aliasing, data observed in two series at different time points.

Particularly important are the problems of association and prediction. The former asks the questions of whether two series are somehow related and what is the strength of any association. Measures of association include: the cross-correlation and the coherence functions. The prediction problem concerns the forecasting of future values. There are useful mathematical formulations of this problem but because of unpredictable human intervention there are situations where guesswork seems just as good.

Theoretical tools employed to address the problems of time series analysis include: mathematical models, asymptotic methods, functional analysis, and transforms.

4. Methods

4.1 Descriptive

Descriptive methods are particularly useful for exploratory and summary purposes. They involve graphs and other displays, simple statistics such as means, medians and percentiles, and the techniques of exploratory data analysis (Tukey 1977).

The most common method of describing a time series is by presenting a graph, see Fig. 1. Such graphs are basic for communication and assessing a situation. There are different types. Cleveland (1985) mentions the connected, symbol, connected-symbol and vertical-line displays in particular. Figure 1 presents connected graphs.

Descriptive values derived from time series data include extremes, turning points, level crossings and the periodogram (see Fig. 2 for an example of the latter). Descriptive methods typically involve generating displays using manipulations of time series data via operations such as differencing, smoothing and narrow band filtering.

A display with a long history (Laplace 1825, Wold 1965) is the Buys–Ballot table. Among other things it is useful for studying the presence and character of a phenomenon of period P such as a circadian rhythm. One creates a matrix with entry $Y((i-1)P+j)$ in row i, column $j = 1, \ldots, P$ and then, for example, one computes column averages. These values provide an estimate of the effect of period P. The graphs of the individual rows may be stacked beneath each other in a display. This is useful for discerning slowly evolving behavior.

Descriptive methods may be time-side (as when a running mean is computed), frequency-side (as when a periodogram is computed) or hybrid (as when a sliding window periodogram analysis is employed).

4.2 Parameter Estimation

The way to a solution of many time series problems is via the setting down of a stochastic model. Parameters are constants of unknown values included in the model. They are important because substantial advantages arise when one works within the context of a model. These advantages include: estimated standard errors, efficiency, and communication of results. Often parameter estimates are important because they are fed into later procedures, for example, formulas for forecasting.

General methods of estimating parameters include: method of moments, least squares and maximum likelihood. An important time series case is that of harmonic regression. It was developed in Fisher (1929) and Whittle (1951).

There are parametric and nonparametric approaches to estimation. The parametric has the advantage that if the model is correct, then the estimated coefficients have considerably smaller standard errors.

4.3 Uncertainty Estimation

Estimates without some indication of their uncertainty are not particularly useful in practice. There are a variety of methods used in time series analysis to develop uncertainty measures. If maximum likelihood estimation has been used there are classic (asymptotic)

formulas. The delta-method or method of linearization is useful if the quantity of interest can be recognized to be a smooth function of other quantities whose variability can be estimated directly. Methods of splitting the data into segments, such as the jackknife, can have time series variants. A method currently enjoying considerable investigation is the bootstrap (Davison and Hinkley 1997). The assumptions made in validating these methods are typically that the series involved is stationary and mixing.

4.4 Seasonal Adjustment

Seasonal adjustment may be defined as the attempted removal of obscuring unobservable annual components. There are many methods of seasonal adjustment (National Bureau of Economic Research 1976), including state space approaches (Hannan 1963).

The power spectrum provides one means of assessing the effects of various suggested adjustment procedures.

4.5 System Identification

System identification refers to the problem of obtaining a description or model of a system on the basis of a stretch of input to and the corresponding output from a system of interest. The system may be assumed to be linear time invariant as defined above.

In designed experiments the input may be a series of pulses, a sinusoid or white noise. In a natural experiment the input is not under the control of the investigator and this leads to complications in the interpretation of the results. System identification relates to the issue of causality. In some systems one can turn the input off and on and things are clearer.

4.6 Computing

Important computer packages for time series analysis include: Splus, Matlab, Mathematica, SAS, SPSS, TSP, STAMP. Some surprising algorithms have been found: the fast Fourier transform (FFT), fast algorithms from computational geometry, Monte Carlo methods, and the Kalman–Bucy filter. Amazingly a variant of the latter was employed by Thiele in 1880 Lauritzen 1981), while the FFT was known to Gauss in the early 1800s (Heideman et al. 1984).

5. Current Theory and Research

Much of what is being developed in current theory and research is driven by what goes on in practice in time series analyses. What is involved in a time series analysis? The elements include: the question, the experiment, the data, plotting the data, the model, model validation, and model use. The importance of recognizing and assessing the basic assumptions is fundamental.

The approach adopted in practice often depends on how much data are available. In the case that there are a lot of data even procedures that are in some sense inefficient are often used effectively. A change from the past is that contemporary analysis often results from the appearance of very large fine data sets. The amount of data can seem limitless as, for example, in the case of records of computer tasks. There are many hot research topics. One can mention: exploratory data analysis techniques for very large data sets, so-called improved estimates, testing (association? cycle present?), goodness of fit diagnostics. There are the classical and Bayesian approaches, the parametric, semi-parametric, and nonparametric analyses, the problem of dimension estimation and that of re-expression of variables.

Current efforts include research into: bootstrap variants, long-memory processes, long-tailed distributions, nonGaussian processes, unit roots, nonlinearities, better approximations to distributions, demonstrations of efficiency, self-similar processes, scaling relationships, irregularly observed data, doubly stochastic processes as in hidden Markov, cointegration, disaggregation, cyclic stationarity, wavelets, and particularly inference for stochastic differential equations.

Today's time series data values may be general linear model type (Fahrmein and Tutz 1994), for example, ordinal, state-valued, counts, proportions, angles, ranks. They may be vectors. They may be functions. The time label t may be location in space or even a function. The series may be vector-valued. The data may have been collected in an experimental design.

There are some surprises: the necessity of tapering and prewhitening to reduce bias, the occurrence of aliasing, the high efficiency of ordinary least squares estimates in the correlated case, and the approximate independence of empirical Fourier transform values at different frequencies (Brillinger 1975).

6. Future Directions

It seems clear that time series research will continue on all the existing topics as the assumptions on which any existing problem solution has been based appear inadequate. Further, it can be anticipated that more researchers from nontraditional areas will become interested in the area of time series as they realize that the data they have collected, or will be collecting, are correlated in time.

Researchers can be expected to be even more concerned with the topics of nonlinearity, conditional

heteroscedasticity, inverse problems, long memory, long tails, uncertainty estimation, inclusion of explanatories, new analytic models, and properties of the estimates when the model is not true. The motivation for the latter is that time series with unusual structure seem to appear steadily. An example is the data being collected automatically via the World Wide Web. Researchers can be anticipated to be seeking extensions of existing time series methods to processes with more general definitions of the time label—spatial, spatial-temporal, functional, angular. At the same time they will work on processes whose values are more general, even abstract.

More efficient, more robust, and more applicable solutions will be found for existing problems. Techniques will be developed for dealing with special difficulties such as missing data, nonstationarity, outliers. Better approximations to the distributions of time series based statistics will be developed.

Many have stressed the advantages of linear system identification via white noise input. Wiener (1958) stressed the benefits of using Gaussian white noise input. This idea has not been fully developed. Indeed data sets obtained with this input will continually yield to novel analytic methods as they make their appearance.

The principal journals in which newly developed statistical methods for time series are presented and studied include: *Journal of Time Series Analysis*, *Annals of Statistics*, *Stochastic Processes and their Applications*, *Journal of the American Statistical Association*, *Econometrica*, *IEEE Transactions in Signal Processing*.

See also: Statistics, History of; Time Series: Advanced Methods; Time Series: Seasonal Adjustment

Bibliography

Akaike H 1965 On the statistical estimation of the frequency response function of a system having multiple input. *Annals of the Institute of Statistics and Mathematics* **17**: 185–210

Bartlett M S 1966 *Stochastic Processes*. Cambridge University Press, Cambridge, UK

Beveridge W H 1922 Wheat prices and rainfall in western Europe. *Journal of the Royal Statistical Society* **85**: 412–59

Blackman R B, Tukey J W 1959 *The Measurement of Power Spectra From the Point of View of Communications Engineering*. Dover, New York

Box G E P, Jenkins G M 1970 *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco

Brillinger D R 1975 *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston, New York

Cleveland W S 1985 *The Elements of Graphing Data*. Wadsworth, Belmont, CA

Davis H T 1941 *The Analysis of Economic Time Series*. Principia, Bloomington, IN

Davison A C, Hinkley D V 1997 *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, UK

Durbin J, Koopman S J 2001 *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford, UK

Durbin J, Watson G S 1950 Testing for serial correlation in least squares regression. *Biometrika* **37**: 409–28

Durbin J, Watson G S 1951 Testing for serial correlation in least squares regression. *Biometrika* **38**: 159–77

Einstein A 1914 Méthode pour la détermination des valeurs statistiques d'observations concernant des grandeurs soumises à des fluctuations irrégulières. *Archives of Sciences Physiques et Naturelles Series 4* **37**: 254–5

Engle R F (ed.) 1995 *ARCH Selected Readings*. Oxford University Press, Oxford, UK

Fahrmein L, Tutz G 1994 *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York

Fisher R A 1929 Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London A* **125**: 54–9

Granger C W J 1982 Acronyms in time series analysis. *Journal of Time Series Analysis* **2**: 103–8

Granger C W J, Andersen A P 1978 *An Introduction to Bilinear Time Series Models*. Vandenhoeck und Ruprecht, Gottingen

Granger C W J, Hatanaka M 1964 *Spectral Analysis of Economic Time Series*. Princeton University Press, Princeton, NJ

Grenander U, Rosenblatt M 1957 *Statistical Analysis of Stationary Time Series*. Wiley, New York

Guttorp P 1995 *Stochastic Modelling of Scientific Data*. Chapman & Hall, London

Hannan E J 1960 *Time Series Analysis*, Methuen, London

Hannan E J 1963 Regression for time series. In: Rosenblatt M (ed.) *Time Series Analysis*. Wiley, New York

Harvey A C 1990 *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, UK

Harvey A C, Koopman S J, Penzer J 1998 Messy time series: A unified approach. *Advances in Econometrics* **13**: 103–43

Heideman M T, Jondon D H, Burrus C S 1984 Gauss and the history of the fast Fourier transform. *Archives in History and Exact Science* **34**: 265–77

Hooker R H 1901 Correlation of the marriage-rate with trade. *Journal of the Royal Statistical Society* **64**: 696–703

Kalman R E 1963 Mathematical description of linear dynamical systems. *SIAM Journal of Control* **1**: 152–92

Laplace P S 1825 *Traité de Méchanique Céleste*. Bachelier, Paris

Lauritzen S L 1981 Time series analysis in 1880: A discussion of contributions made by T. N. Thiele. *International Statistical Review* **49**: 319–31

MacDonald I L, Zucchini W 1997 *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman and Hall, London

Moore H L 1914 *Economic Cycles: Their Law and Cause*. Macmillan, New York

National Bureau of Economic Research/Bureau of the Census 1976 *Conference on Seasonal Analysis of Economic Time Series*. Dept. of Commerce, Bureau of the Census, Washington, DC

Parzen E 1961 An approach to time series analysis. *Annals of Mathematical Statistics* **32**: 951–89

Prakasa Rao B L S 1999 *Statistical Inference for Diffusion Type Processes*. Arnold, London

Shumway R H, Stoffer D S 2000 *Time Series Analysis and its Applications*. Springer, New York

Tong H 1990 *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press, New York

Tufte E R 1983 *The Elements of Graphing Data*. Graphics Press, Cheshire, CT

Tukey J W 1977 *Exploratory Data Analysis*. Addision-Wesley, Reading, MA
Tukey J W 1984 *The Collected Works of John W. Tukey I-II Time Series*. Wadsworth, Belmont, CA
Whittle P 1951 *Hypothesis Testing in Time Series Analysis*. Almqvist and Wiksell, Uppsala, Sweden
Wiener N 1949 *Time Series*. MIT Press, Cambridge, MA
Wiener N 1958 *Nonlinear Problems in Random Theory*. MIT Press, Cambridge, MA
Wold H O A 1965 *Bibliography on Time Series and Stochastic Processes*. MIT Press, Cambridge, MA
Yule G U 1927 On the method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philosophical Transactions A* **226**: 267–98

D. R. Brillinger

Time Series: Nonstationary Distributions and Unit Roots

1. Introduction

A stationary time series is one whose mean and variance are constant and the covariance between observations depends only on their separation in time. From a statistics point of view the distinction between stationary and nonstationary series is important because of its implications about the distributions of estimators. From an economics point of view, the distinction is important because certain economic theories imply nonstationarity and can be tested by looking for it in the data. There are questions about the impact of unanticipated shocks to an economic system. Do they produce permanent or transitory shifts in economic variables? This too boils down to a stationarity question. Diebold and Nerlove (1990) provide a long list of references in this area.

2. The Role of Nonstationary Processes and Unit Roots

When data are taken over time, a natural way to forecast future observations is to model the current observation Y_t as a linear combination of previous observations Y_{t-j}. Deviations from some deterministic function, such as a long run mean μ, are often the basis for prediction. Such a prediction is written as

$$\hat{Y}_t - \mu = \alpha_1(Y_{t-1}-\mu) + \alpha_2(Y_{t-2}-\mu) + \ldots + \alpha_p(Y_{t-p}-\mu)$$

For the usual *autoregressive model of order p*

$$Y_t - \mu = \alpha_1(Y_{t-1}-\mu) + \alpha_2(Y_{t-2}-\mu) + \ldots + \alpha_p(Y_{t-p}-\mu) + e_t$$

this predictor is the conditional expected value of Y_t given the past. Here e_t is *white noise*, that is, a sequence of i.i.d. $(0,\sigma^2)$ variates. One must estimate the α_i and μ parameters as well as the number p of lags to use. Early work by Mann and Wald (1943) established the asymptotic normality of parameters estimated by least squares regression of Y_t on $Y_{t-1},\ldots,Y_{t-p}$ for *covariance stationary* time series. A series is covariance stationary if it has a constant mean and the covariance between Y_t and Y_{t+h} is a function only of $|h|$ denoted by $\gamma(h)$ and called the autocovariance function. For an autoregressive series as defined above, covariance stationarity is ensured if the roots of the characteristic equation

$$1 - \alpha_1 m - \alpha_2 m^2 - \ldots - \alpha_p m^p = 0$$

are all greater than 1 in magnitude. In that case the covariance $\gamma(h)$ is bounded by some constant multiple of $(1/|m_{\min}|)^{|h|}$ where $m_{\min}$ is the root of the characteristic equation with minimum (closest to 1) magnitude.

For example if $p = 2$ lags, with $Y_{99} = 150$ and $Y_{100} = 120$, then using

$$(\mu,\alpha_1,\alpha_2) = (100, 0.9, -0.2)$$

the one-step-ahead forecast is $\hat{Y}_{101} = 108$ computed as

$$100 + 0.9(120-100) - 0.2(150-100) = 0.3(100) + 0.9(120) - 0.2(150)$$

while $(\mu,\alpha_1,\alpha_2) = (\mu, 1.2, -0.2)$ gives the forecast

$$\hat{Y}_{101} = \mu + 1.2(120-\mu) - 0.2(150-\mu) = 1.2(120) - 0.2(150) = 114$$

regardless of the value of μ. The characteristic equations factor as $(1-0.4m)(1-0.5m) = 0$ in the first and $(1-0.2m)(1-m) = 0$ in the second case. The *unit root* $m = 1$ in the second case is the reason that the forecast does not involve μ. A unit root series is nonstationary, having no tendency for mean reversion.

The bounding of the autocovariance by an exponentially decaying function in the stationary case makes the data sequence close enough to an uncorrelated sequence so that standard results like asymptotic normality of regression estimates will hold. While convenient from a statistical theory point of view, the exponentially decreasing covariance seems inconsistent with many observed data sets. Furthermore, exponential decay of the correlations implies that forecasts h steps into the future will approach the series mean at a rate bounded by $(1/|m_{\min}|)^{|h|}$.

Such quick reversion to the mean may at times seem unrealistic.

As a motivating example, suppose the price of a stock were represented as

$$Y_t - \mu = \alpha(Y_{t-1} - \mu) + e_t$$

an autoregressive process of order $p = 1$. The deviation of the current price from the historic mean should hold no information about future prices. If it did, the pattern would be known to intelligent market participants who would in turn bid the price up or down in such a way as to nullify the effect of the information. What does this say about α? It eliminates $|\alpha| < 1$, for then the prediction of Y_{t+h}, $\mu + \alpha^h(Y_t - \mu)$ would move toward the historic mean thus implying useful information in the knowledge of $Y_t - \mu$. Assuming that such knowledge is not informative suggests $\alpha = 1$, a testable hypothesis.

For $|\alpha| < 1$, the series can be represented as a convergent infinite sum

$$Y_t - \mu = \sum_{j=0}^{\infty} \alpha^j e_{t-j}$$

in which the effect of the unanticipated 'shocks' e_{t-j} in the distant past is negligible. However, for $\alpha = 1$ the recursive model becomes $Y_t = Y_{t-1} + e_t$, a simple *random walk* in which the mean μ has no apparent role and the only thing moving the stock price is the unanticipated (white noise) shock e_t. When $\alpha = 1$, Y_t becomes $Y_t = Y_1 + e_2 + e_3 + \ldots + e_t$ and the shocks e_t have a permanent effect. Each Y_t is Y_1 plus a sum of es so the deviation of Y_t from the average will not involve Y_1 nor will the residuals from any regression that includes an intercept. The assumptions on Y_1 will be irrelevant for any statistics computable from a regression with an intercept column.

In a more general context, whenever the characteristic equation has a unit root, the term μ drops out of the model, the forecasts do not revert to a historic mean, the forecast error variance grows without bound, and the distribution of the least squares regression estimates are not all normal even in the limit. Just as the reported Dow Jones average is usually followed by its difference, the 'up or down points' on the evening news, series with unit roots are often differenced prior to analysis in order to produce stationary series with nice statistical properties. The nonstationary series

$$Y_t - \mu = 1.2(Y_{t-1} - \mu) - 0.2(Y_{t-2} - \mu) + e_t$$

with a bit of algebra, is re-expressed as

$$Y_t - Y_{t-1} = 0.2(Y_{t-1} - Y_{t-2}) + e_t$$

This is a stationary autoregressive process in the differenced series $Y_t - Y_{t-1}$. The parameter 0.2 can be estimated by regressing the difference on its lag. In general

$$Y_t - \mu = \alpha_1(Y_{t-1} - \mu) + \alpha_2(Y_{t-2} - \mu) + e_t \text{ becomes}$$
$$Y_t - Y_{t-1} = -(1 - \alpha_1 - \alpha_2)(Y_{t-1} - \mu) - \alpha_2(Y_{t-1} - Y_{t-2}) + e_t$$

The coefficient of the lag level term $(Y_{t-1} - \mu)$ is seen to be the negative of the characteristic polynomial $1 - \alpha_1 m - \alpha_2 m^2$ at $m = 1$. This coefficient is 0 if and only if there is a unit root and is negative if the series is stationary.

The expression above gives a general method for testing the hypothesis of a unit root in autoregressions. Regress $Y_t - Y_{t-1}$ on $Y_{t-1}$and enough lags of $Y_t - Y_{t-1}$ to reproduce the autoregressive order of the series, then test to see if the coefficient of Y_{t-1} is 0. Of course to do that, one must have at hand the null distribution of the estimated coefficient and/or its t statistic. The work of Mann and Wald does not apply to a nonstationary series and in fact, the limit distributions are not normal. This does *not* imply that the regression t statistic is an inappropriate test. It does imply that a new distribution will need to be tabulated for it.

3. *Distributions for Unit Root Tests*

Starting very simply, assume $Y_t = \alpha Y_{t-1} + e_t$ or equivalently $Y_t - Y_{t-1} = (\alpha - 1)Y_{t-1} + e_t$ and $\alpha = 1$ is to be tested. If $\alpha = 1$, the regression of $Y_t - Y_{t-1}$on Y_{t-1} for $t = 2,3,\ldots,n$ gives regression coefficient

$$\sum_{t=2}^{n} Y_{t-1} e_t \Big/ \sum_{t=2}^{n} Y_{t-1}^2$$

(because $Y_t - Y_{t-1} = e_t$). Dickey and Fuller (1979) show that this statistic converges to 0 at rate $1/n$. The *normalized bias* statistic

$$n(\hat{\alpha} - 1) = \frac{1}{n\sigma^2} \sum_{t=2}^{n} Y_{t-1} e_t \Big/ \frac{1}{n^2\sigma^2} \sum_{t=2}^{n} Y_{t-1}^2$$

is of order 1. For stationary series $\sqrt{n}(\hat{\alpha} - \alpha)$ is the $O_p(1)$ normalization. Here the numerator and denominator quadratic forms have each been normalized to be $O_p(1)$. Dickey and Fuller derive a representation for the denominator using its eigenvalues. The limit is an infinite weighted sum of squared $N(0,1)$ variables. They derive the joint representation for the numerator and denominator in terms of these $N(0,1)$ variables and a method for simulating from the limit distribution. Empirical

Table 1
Limit-critical values for unit root tests (τ, τ_μ, τ_τ) compared to normal

Regress $Y_t - Y_{t-1}$ on	1 percent critical value	5 percent critical value
Y_{t-1} plus lagged differences	-2.58	-1.95
1, Y_{t-1} plus lagged differences	-3.42	-2.86
1, t Y_{t-1} plus lagged differences	-3.96	-3.41
(normal values)	-2.326	-1.645

percentiles appear in Fuller (1996). Matching of the large n and limit percentiles provides a check on the theory and computer simulations.

An alternative approach to the distribution of this statistic was suggested by White (1958, 1959) and shown in detail by Chan and Wei (1987) and by Phillips and his students in a series of papers; see for example Phillips and Perron (1988). This mathematically elegant approach expresses the limit distribution of the normalized bias statistic $n(\hat{\alpha}-1)$ in terms of a Weiner process $W(t)$ as

$$\frac{1}{2}(W^2(1)-1)/\int_0^1 W^2(t)dt$$

Billingsley (1968) gives the general theory for this type of convergence. The denominator and numerator are both random variables and the limit distribution is the same regardless of whether a quadratic form decomposition or Weiner Process representation is written, so the tables from Fuller (1996) are used either way.

The associated studentized statistic, called τ rather than t by Dickey and Fuller to emphasize its nonstandard distribution, converges to

$$\frac{1}{2}(W^2(1)-1)/\sqrt{\int_0^1 W^2(t)dt}$$

in the Weiner process representation and has a similar quadratic form representation. An exceptionally nice feature of τ is that its limit distribution under the null hypothesis is the same for higher order autoregressive processes.

4. *Beyond Autoregression*

For purely autoregressive processes, it has been seen that regression of the first difference on the lagged level and enough lagged differences gives a unit root test. The test on the lagged level term in the presence of lagged differences has become known as the augmented Dickey–Fuller test or ADF. Said and Dickey (1985) show that in mixed models (ARIMA models involving both lags of Y_t and e_t) direct nonlinear estimation of the parameters results in test statistics having the same limiting distributions as discussed above. Said and Dickey (1984) show that if one ignores the true mixed structure and simply fits a long enough autoregression, the asymptotics again give the previously discussed τ test. This has been quite reassuring for practitioners who often just use autoregressive representations at this stage of analysis. Hall (1994) studied methods for determining the correct number of lagged differences to include based on the data. Overfitting then testing to omit unnecessary lagged differences appeared to be a good way to prepare a model for unit root testing. With that in mind, a nice feature of unit root models in the ADF form is that the limit distribution of the regression t tests for the lagged difference terms have the usual standard limiting normal distributions, as in Mann and Wald (1943).

For higher order mixed models, Phillips and Perron (1988) regress the difference on the lagged level as though a lag one model is appropriate. They then adjust the test statistics based on the autocorrelation function of the differenced series which, of course, must be estimated. Decisions on how to estimate these autocorrelations and how many of them to use in the adjusting process must be made, just as a decision on how many lagged differences to use must be made in the ADF test. The method uses the same critical values as ADF. All methods are justified by asymptotics. The finite sample power and adherence to the claimed significance level will differ. Schwert (1989) provides a comparison.

5. *Models with Deterministic Parts*

In a usual regression problem, under the null hypothesis the t test for a regression coefficient has a t distribution in a properly specified model regardless of what other terms are in the model. Unfortunately the same is not true of τ in unit root testing. For example, a lag 1 model with a trend and autocorrelation might be posed as

$$(Y_t-\beta_0-\beta_1 t) = \alpha(Y_{t-1}-\beta_0-\beta_1(t-1))+e_t$$

or equivalently in two parts as $Y_t = \beta_0+\beta_1 t+Z_t$ where $Z_t = \alpha Z_{t-1}+e_t$. Now if $|\alpha| < 1$ we have stationary errors around a trend, or *trend stationarity* as it is

known in the econometric literature. On the other hand if $\alpha = 1$ our model becomes

$$Y_t = Y_{t-1} + \beta_1 + e_t$$

known as a random walk with drift β_1. In order to estimate the model well under both the null and alternative hypotheses, one regresses $Y_t - Y_{t-1}$ on Y_{t-1}, 1, and t, then tests the coefficient of Y_{t-1} using its studentized statistic τ_t. Even if the process is a simple random walk, this test has neither a limit normal distribution nor the limit obtained when regressing $Y_t - Y_{t-1}$ on just Y_{t-1}. The limit distributions for both normalized bias and studentized statistics change when an intercept and/or trend are added to the regression.

The 1 percent and 5 percent critical values from the limit distribution of the studentized test statistic for cases in which one regresses the first difference on a lagged level and lagged differences (τ) are given in Table 1, along with the corresponding critical values when an intercept (τ_μ) or both an intercept and time (τ_τ) are included. These are tabulated under the unit root null hypothesis. The familiar normal distribution critical values are also shown for comparison.

Clearly the naive use of normal or t critical values in this time series setting will result in far too many spurious declarations of stationarity, even in extremely large samples. From the table it is seen that the problem worsens with the fitting of deterministic mean and trend components. Models containing 1 (and t) are invariant to nonzero means (and deterministic trends) in the data. That is, the calculated test statistics used in rows 2 and 3 will remain exactly the same if an arbitrary constant is added to the data, while that in the third row will be unchanged by the addition of an arbitrary linear trend. In discussing trends, it is important to recall that under the $\alpha = 1$ null hypothesis, a trend is induced by a nonzero β_1 in the model $Y_t = Y_{t-1} + \beta_1 + e_t$.

6. Beyond Regression

In an attempt to increase the power of unit root tests, several alternative estimation methods have been tried. Dickey et al. (1984) introduce a symmetric estimator which combines the regression of Y_t on past Ys with a regression of Y_t on future Ys. Fuller (1996) discusses this and a weighted symmetric estimator. Elliott et al. (1996) base a test statistic on the following clever idea. Pick a point in the alternative that depends on n, say $\alpha = 1 - 0.7/n$. Now find the point optimal test of $\alpha = 1$ vs. the simple alternative $\alpha = 1 - .7/n$. Counting on the continuity of the power curve, that test should give good power over a range of alternative points near that for which it is point optimal.

The estimator that maximizes the stationary normal likelihood function performs slightly better than least squares in the stationary case. Looking at this likelihood function as simply an objective function to be maximized, Gonzalez-Farias (1992) and Gonzalez-Farias and Dickey (1999) compute its maximum for a unit root process defining what she calls the *unconditional maximum likelihood* estimator. She gets representations for it including cases with intercepts and trends in the model. In a comparison study by Pantula et al. (1994) this estimator and Fuller's weighted symmetric estimator have very similar behavior and seem to be the best performers of the tests mentioned here. All tests described in this section require new tables of critical values which the corresponding authors provide.

7. Related Topics

Unit roots can arise in multivariate time series, that is, in cases where a vector of observations is recorded at each time point. In such cases, there may be some linear combinations of the vectors that form stationary time series and other linear combinations that are nonstationary. These results appear under the name 'co-integration analysis' and 'reduced rank regression.' Another recent area of research is 'trend breaks.' There might be one mean during the first period of data collection and a second one later. How do unit root tests perform if the break is undiscovered? Can we discover the break point? Is it possible that series we thought were nonstationary are in fact stationary around two or three different means? Seasonal unit roots and methods of decomposing them are also currently popular topics.

See also: Time Series: Advanced Methods; Time Series: General

Bibliography

Ahn S K, Reinsel G C 1990 Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association* **85**: 813–23

Billingsley P 1968 *Convergence of Probability Measures*. Wiley, New York

Chan N H, Wei C Z 1987 Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* **16**: 367–401

Dickey D A, Fuller W A 1979 Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74**: 427–31

Dickey D A, Hasza D P, Fuller W A 1984 Testing for unit roots in seasonal time series. *Journal of the American Statistical Association* **79**: 355–67

Diebold F X, Nerlove M 1990 Unit roots in economic time series: a selective survey. *Econometrica* **8**: 3–69

Elliott G, Rothenberg T J, Stock J H 1996 Efficient tests for an autoregressive unit root. *Econometrica* **64**: 813–36

Fuller W A 1996 *Introduction to Statistical Time Series*. Wiley, New York
Gonzalez-Farias G 1992 A new unit root test for autoregressive time series. Ph.D. Thesis, North Carolina State University
Gonzalez-Farias G, Dickey D A 1999 Unit root tests: An unconditional maximum likelihood approach. *Boletin de la Sociedad Matematica Mexicana* **5**: 199–221
Hall A 1994 Testing for a unit root in time series with pretest data-based model selection. *Journal of Business and Economic Statistics* **12**: 461–70
Mann H B, Wald A 1943 On the statistical treatment of linear stochastic difference equations. *Econometrica* **11**: 173–220
Pantula S G,, Gonzalez-Farias G, Fuller W A 1994 A comparison of unit root test criteria. *Journal of Business and Economic Statistics* **12**: 449–59
Phillips P C B, Perron P 1988 Testing for a unit root in time series regression. *Biometrika* **75**: 335–46
Said S E, Dickey D A 1984 Testing for unit roots in autoregressive moving average models of unknown order. *Biometrika* **71**: 599–607
Said S E, Dickey D A 1985 Hypothesis testing in ARIMA (p,1,q) models. *Journal of the American Statistical Association* **80**: 369–74
Schwert G W 1989 Tests for unit roots: a Monte Carlo investigation. *Journal of Business and Economic Statistics* **7**: 147–59
White J S 1958 The limiting distribution of the serial-correlation coefficient in the explosive case. *Annals of Mathematical Statistics* **29**: 1188–97
White J S 1959 The limiting distribution of the serial-correlation coefficient in the explosive case II. *Annals of Mathematical Statistics* **30**: 831–4

D. A. Dickey

Time Series Quasi-experiment

Campbell and Stanley (1963), Campbell (1963) recommend the time series quasi-experiment for assessing the causal effect of a discrete intervention on a time series. In the simplest case of this design, a discrete intervention breaks a time series into pre- and post-intervention segments of N_{pre} and N_{post} observations. For pre- and post-intervention means μ_{pre} and μ_{post}, analysis of the quasi-experiment tests the null hypothesis

$$H_0\colon \omega = 0 \text{ where } \omega = \mu_{post} - \mu_{pre} \qquad (1)$$

Rejecting H_0, H_A attributes ω to the intervention. The validity of H_A can be challenged on the two grounds that Cook and Campbell (1979) call 'threats to internal validity' and 'threats to statistical conclusion validity.' Whereas the threats to internal validity apply to quasi-experiments generally, the threats to statistical conclusion validity, which arise because the time series observations are sampled sequentially, distinguish this design from other quasi-experiments. Whereas the threats to internal validity are controlled by design, a statistical model must control the threats to statistical conclusion validity.

1. Internal Validity

Although in principle, all eight of the threats to internal validity discussed by Campbell and Stanley (1963) apply to the time series quasi-experiment, four threats are inherently plausible. The threats of history and maturation are plausible due to the design's relatively long time frame and to the nature of long time series. The threats of instrumentation and regression are plausible due to the nature of planned interventions; for unplanned interventions, a special case that Campbell (1969) calls the 'natural experiment,' neither threat is plausible.

Hennigan et al. (1982) use the time series in Fig. 1 to document an increase in property crime following the introduction of commercial television. In 34 'early' and 34 'late' cities, television broadcasting begins in 1950 and 1954, respectively. If the two time series are considered separately, history is a plausible threat to the internal validity of both effects. The 'late' effect may be due to an economic recession that began in 1955, for example, and the 'early' effect may be an artifact of the Korean War mobilization. As husbands were called away to military service and wives joined the workforce, homes were left undefended against property criminals. Although history is a plausible threat to the internal validity of each effect, however, it is implausible for the joint effect.

The other three common threats to internal validity are less robust to standard designs. Fig. 2 shows monthly burglaries for Tucson before, during, and after an intervention (McCleary et al. 1982). Prior to 1979, burglaries were investigated by uniformed police officers. When this task was taken over by detectives, burglaries dropped abruptly. Two years later, when the task was transferred back to uniformed officers, burglaries returned abruptly to pre-1979 levels. While the on-off component of this design rules out the threat of history, the other threats are still plausible. In fact, this effect is entirely explained by differences in the way detectives and uniformed officers keep records; it is an instrumentation artifact.

Because interventions often affect the way in which a target phenomenon is measured, instrumentation is always a plausible threat to the evaluation of planned interventions. Regression threats are also associated with planned interventions but for different reasons. Following a 1955 traffic 'crackdown' in Connecticut, for example, highway fatalities dropped significantly. Examining a long time series, however, Campbell and Ross (1968) show that the reduction was largely due to an uncontrolled regression threat. The 'crackdown' was implemented in 1955 because traffic fatalities in

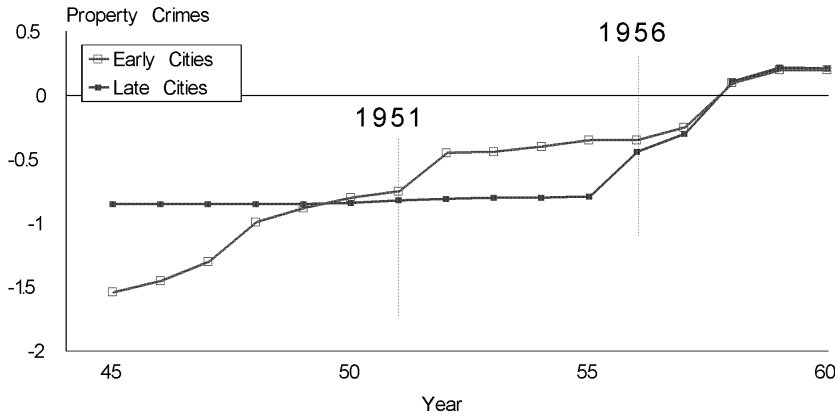

Figure 1
Annual property crimes for cities with commercial TV broadcast service in 1951 and for cities with TV in 1955. Source: Hennigan et al. (1982)

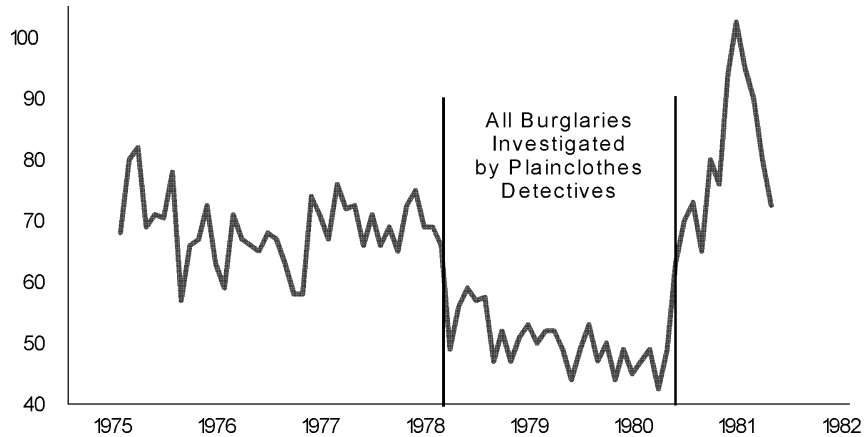

Figure 2
Monthly burglaries for Tucson, 1975–1981. Source: McCleary et al. (1982)

that year were abnormally high; regression to the mean was virtually guaranteed.

Unlike history and instrumentation, which are controlled by design, a statistical model controls regression. The opportunistic timing of the intervention biases the estimates of the pre- and post-intervention means, μ_{pre} and μ_{post}. As N_{pre} and N_{post}, grow longer, this bias approaches zero and, accordingly, authorities on the statistical model recommend $N_{pre} + N_{post} > 50$ (McCleary and Hay 1980, McDowall et al. 1980).

Like regression, the statistical model controls maturation. Because secular trend or stochastic drift generates this threat, and because these properties are found in most long social science time series, maturation is plausible in virtually all time series quasi-experiments.

Fig. 3 shows a weekly productivity series from the Hawthorne experiment. From 1927 to 1932, researchers manipulated daily rest breaks and, though not conducting a formal analysis, claimed that rest breaks had a large, positive impact on productivity. However, formal analyses of these data challenge the original finding. Regressing step functions coded for three 'punishments' regimes, Franke and Kaule (1978) demonstrate that the productivity increases were motivated by fear.

The fundamental problem with this interpretation, as well as with the original interpretation, is that the productivity time series exhibits a systematic trend throughout its history. To the extent that the bank wiring room machine operators were becoming more productive over time, maturation is a plausible threat to the internal validity of both interpretations.

To illustrate the problem, the punishment effects estimated by Franke and Kaule are plotted against the

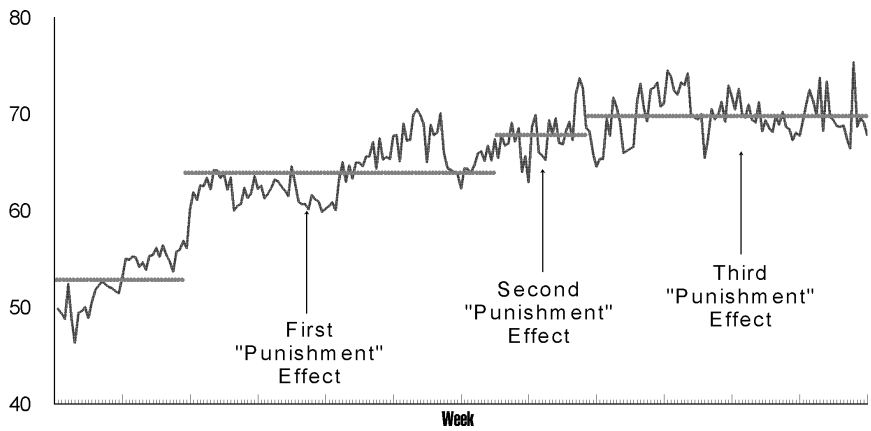

Figure 3
Weekly productivity measures for five machine operators in the first bank wiring room (Hawthorne) experiment. Estimated step function effects are plotted against the series. Source: Franke and Kaul (1978)

productivity series in Fig. 3. The effects are collinear with the natural trend of this series. Reanalysis of these data with a difference equation model controlling for trend finds a modest but significant effect for rest breaks and no effect for punishment (McCleary 2000).

2. *Statistical Conclusion Validity*

Reliance on a sophisticated statistical model distinguishes the time series quasi-experiment from other quasi-experimental designs. Although times series designs appear in the social science literature as early as the nineteenth century, without a statistical model, the design offered little advantage over a simpler, more economical pre-post design. In the year 2000, the most widely used statistical model for the time series quasi-experiment was the *A*utoRegressive *I*ntegrated *M*oving *A*verage (ARIMA) model (Box and Jenkins 1970, McCleary and Hay 1980. A statistically adequate ARIMA model controls the maturation threat to internal validity and all relevant threats to statistical conclusion validity.

The seven threats to statistical conclusion validity discussed by Cook and Campbell (1979, pp. 42–4) refer to misstatements of the Type I and Type II error rates. A Type I error occurs whenever a true H_0 is rejected; at the nominal 0.95 confidence level, the Type I error rate is $\alpha = 0.05$. A Type II error occurs whenever a false H_0 is accepted; at the nominal 0.80 power level, the Type II error rate is $\beta = 0.2$. Threats to statistical conclusion validity distort the nominal values of α and β resulting in false inferences about the hypothetical impact of an intervention; that the intervention had a significant impact in the Type I case, or that the intervention had no impact in the Type II case.

The Type II threats to statistical conclusion validity are relatively straightforward in explication and control. Not withstanding the magnitude of an impact—small effects are difficult to detect in even the best circumstances—the statistical power of the time series quasi-experiment is a function of series length, balance, and model adequacy. Because estimation algorithms rely on large sample properties, analyses of short time series will often fail to achieve the nominal level of statistical power. Given a long time series, say $N \geqslant 50$ observations, statistical power is at its maximum when $N_{pre} = N_{post}$. Given a long time series and a balanced design, however, statistical power assumes a statistically adequate model; i.e., a model that adequately decomposes the series into a stochastic component that is unrelated to the intervention and into a component that is related to the intervention.

The Type I threats to statistical conclusion validity, which have received greater attention in the literature, are associated with common stochastic time series properties, particularly trend or drift and autocorrelation. And since the time series model controls these properties, all Type I threats are a function of a model property called statistical adequacy.

Virtually all of the models used to analyze time series quasi-experiments are based on the *A*uto-Regressive *I*ntegrated *M*oving *A*verage (ARIMA) model (Box and Jenkins 1970, McCleary and Hay 1980. The general ARIMA model can be written as,

$$\varphi(B)\nabla Y_t = \theta(B)a_t + \omega(B)I_t \qquad (2)$$

where Y_t, a_t, and I_t are, respectively, the tth observations of a time series, an independently, identically distributed (*iid*) Normal error series, and a step function defined such that $I_t = 0$ t N_{pre} and $I_t = 1$ thereafter; where $\varphi(B)$ and $\theta(B)$ are polynomial lag operators; where $\omega(B)I_t$ is a transfer function of I_t; and where ∇ is a differencing operator. When the inter-

vention has had no impact on the series, H_0 is true so $\omega = 0$ and the ARIMA model reduces to

$$\varphi(B)Y_t = \theta(B)a_t \tag{3}$$

This simpler form reveals the statistical rationale for the three operators.

The threat to statistical conclusion validity posed by the most common nonstationary properties of time series, trend or drift is illustrated in the productivity series (Fig. 3). In a trending series, any arbitrary intervention biases the estimates of μ_{pre} and μ_{post} and, thus, the test of H_0. Assuming a homogeneous nonstationary time series, however, the differencing operator, ∇, transforms the nonstationary Y_t to a stationary form. If Z_t denotes the differenced time series,

$$Z_t = \nabla Y_t = Y_t - Y_{t-1} \tag{4}$$

Then Z_t retains the fundamental characteristics of the original time series, at least with respect to the hypothetical impact of the intervention, but is flat or trendless.

Although differencing controls the threats to statistical conclusion validity posed by trend or drift, Z_t is not necessarily *iid*. On the contrary, successive values of the differenced time series will ordinarily be highly correlated and autocorrelation poses a separate threat to statistical conclusion validity. Contrary to popular belief, unlike trend or drift, autocorrelation cannot be seen in most time series. Nevertheless, the autocorrelation threats are potent and plausible in most time series quasi-experiments.

In the ARIMA model, the threats due to autocorrelation are controlled by the polynomial lag operators, φ(B) and θ(B). If the parameters of φ(B) and θ(B) are restricted a range of values where the operators have algebraic inverses, the model can be solved for the *iid* error term.

$$a_t = \theta(B)^{-1}\,\varphi(B)Z_t \tag{5}$$

Because the *iid* time series satisfies the assumptions of an ANOVA model, μ_{pre} and μ_{post} can be estimated with conventional maximum likelihood methods.

Although the ARIMA model controls all threats to statistical conclusion validity in principle, in practice the ARIMA model must satisfy two criteria. First, the inverted model must yield an *iid* time series; this criterion has been called statistical adequacy. Second, among all statistically adequate ARIMA models, the final model must have the minimum residual variance. The first criterion pertains to Type I threats and can be tested empirically. Given that there are many statistically adequate models for most time series, the second criterion has no firm empirical sign. If the final ARIMA model does not have a minimum residual variance, statistical power will be less than optimal and a Type II threat will be plausible. This suggests the need for conducting formal analyses of statistical power whenever H_0 cannot be rejected.

3. *External Validity*

Although components of the ARIMA model were available as early as 1920 (e.g., Slutzky 1927, Yule 1927), practical applications of the model required computing resources that were unavailable before 1970. In an appendix to *Time Series Analysis: Forecasting and Control*, Box and Jenkins (1970) described a general algorithm for building ARIMA models from time series. Implementations of the algorithm became available shortly after 1970. Modern time series software packages, which are included in virtually all-commercial statistical analysis systems, use the original Box and Jenkins (1970) modeling algorithm.

As software became more widely available, after 1980 applications of the time series quasi-experiment began to accumulate. In evaluation research, effect estimates of replicated social programs often-supported discrepant substantive inferences about the program's effectiveness. The literature evaluating the effectiveness of gun-control legislation is typical in that the published estimates include both positive, negative, and null effect estimates (Reiss and Roth 1993, pp. 255–87). Whereas a small proportion of randomly distributed null effects can be ignored, a broader range of quasi-experimental effect estimates suggests that the inter-replication variance is more important than the general effect itself.

This raises questions of external validity which, following Cook and Campbell (1979, p. 71) concern 'generalizing across types of persons, settings, and times.' The external validity of several time series quasi-experiments can be analyzed in either of two ways. First, the set of k time series can be defined as vector time series, $\mathbf{Y}_t$:

$$\mathbf{Y}_t' = \{Y_{1,t}, Y_{2,t}, \ldots, Y_{k,t}\} \tag{6}$$

The vector $\mathbf{Y}_t'$ can then be analyzed as a vector *ARIMA* process (e.g., McGaw and McCleary 1985). Second, variance in the set of k reported effect estimates,

$$\omega' = \{\omega_1, \omega_2, \ldots, \omega_k\} \tag{7}$$

can be decomposed into components associated with threats to external validity. Since the second approach is a special, restricted case of fitting the set of k related quasi-experiments into a vector ARIMA model; the first approach is preferred. In practice, unfortunately, few time series satisfy the rigid assumptions required for a vector ARIMA model.

Though less general, the second approach is more practical and flexible in most situations. The statistical

models used to analyze the set of k related effect estimates, which were originally developed in meta-analysis or research synthesis, decompose the effect variance into components that can be attributed to setting (place, historical period, etc.), properties of the intervention (resources devoted to the intervention, mode of implementation, etc.), methodology (time series properties, dependencies, etc.), and other potential threats to external validity. McDowall et al. (1992) demonstrate this method for combining effect estimates from evaluations of gun control laws; McCleary (2000) uses a similar model for combining effect estimates from the Hawthorne experiment plotted in Fig. 3.

4. Future Directions

At this point in its development, solutions to the questions of internal validity have been proposed, tested, and accepted. Although solutions to the questions of external validity are still too novel to enjoy the same level of consensus, it appears that the currently available solutions will become more widely used in the near future. The unresolved problems relate to statistical conclusion validity. The problems of missing data and short time series (e.g., $N < 30$ observations) pose a particular need.

See also: External Validity; Internal Validity; Quasi-Experimental Designs; Time Series: Advanced Methods; Time Series: ARIMA Methods; Time Series: General

Bibliography

Box G E P, Jenkins G M 1970 *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, CA

Campbell D T 1963 From description to experimentation: Interpreting trends as quasi-experiments. In: Harris C W (eds.) *Problems In Measuring Change*. University of Wisconsin Press, Madison, WI

Campbell D T 1969 Reforms as experiments. *American Psychologist* **24**: 409–29

Campbell D T, Ross H L 1968 The Connecticut crackdown on speeding: time series data in quasi-experimental analysis. *Law and Society Review* **3**: 33–53

Campbell D T, Stanley J C 1963 *Experimental and Quasi-experimental Designs for Research*. Rand-McNally, Skokie, IL

Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Houghton-Mifflin, Boston, MA

Franke R H, Kaal J D 1978 The Hawthorne experiments: First statistical interpretation. *American Sociology Review* **43**: 623–43

Hennigan K , Del Rosario M L, Heath L, Cook T D, Wharton J D, Colder B J 1982 Impact of the introduction of television on crime in the United States: Empirical findings and theoretical implications. *Journal of Personality and Social Psychology* **42**(3): 461–77

McCleary R 2000 Evolution of the time series experiment. In: Bickman L (eds.) *Research Design: Donald Campbell's Legacy*. Sage, Thousand Oaks, CA, pp. 215–34

McCleary R, Hay R A Jr 1980 *Applied Time Series Analysis for the Social Sciences*. Sage, Beverly Hills, CA

McCleary R, Nienstedt B C, Erven J M 1982 Uniform crime reports and organizational outcomes: three time series quasi-experiments. *Social Problems* **29**: 361–72

McDowall D, Loftin C, Wiersema B 1992 A comparative study of the preventive effects of mandatory sentencing laws for gun crimes. *Journal of Criminal Law and Criminology* **83**: 378–94

McDowall D, McCleary R, Hay R A Jr, Medinger E E 1980 *Interrupted Time Series Analysis. Vol. 21, University Papers Series: Quantitative Applications in the Social Sciences*. Sage, Beverly Hills, CA, and London

McGaw D B, McCleary R 1985 PAC spending, electioneering, and lobbying: a vector ARIMA time series analysis. *Polity* **17**: 574–85

McSweeney A J 1978 The effects of response cost on the behavior of a million persons: Charging for directory assistance in Cincinnati. *Journal of Applied Behavioral Analysis* **11**: 47–51

Richardson J, Kroeber A L 1940 Three centuries of women's dress fashions: A quantitative analysis. *Anthropological Records* **5**(2): 111–54

Riess A J, Roth J A 1993 *Understanding and Preventing Violence*. National Academy Press, Washington, DC

Ross H L, Campbell D T, Glass G V 1970 Determining the effects of a legal reform: The British 'breathalyzer' crackdown of 1967. *American Behavioral Scientist* **13**: 493–509

Shlaifer R 1980 The relay assembly test room: An alternative statistical interpretation (Comment on Franke and Kaul, ASR. October 1978). *American Sociological Review* **45**: 995–1005

Slutsky E 1927 The summation of random causes as the source of cyclic processes. *Problems of Economic Conditions* **3**: 1

Webb E J, Campbell D T, Schwartz R D, Sechrest L 1966 *Unobtrusive Measures*. Rand-McNally, Skokie, IL

Yule G U 1927 On a method of investigating periodicity in disturbed series with special reference to Wolfer's sunspot numbers. *Philosophical Transactions* **A226**: 267

R. McCleary

Time Series: Seasonal Adjustment

Time series data are often subject to various adjustments among which seasonal adjustment is one of the most common. Seasonal adjustment means the removal of seasonal variations from the original series. Seasonality represents the composite effect of climatic and institutional events which repeat more or less regularly every year. One main reason for seasonal adjustment is the standardization of socioeconomic series because seasonality affects them with different timing and intensity. Hence, the seasonally adjusted data reflect variations due only to the remaining components, trend, cycle, and irregulars. The infor-

mation given by seasonally adjusted series plays a crucial role in the analysis of current economic conditions and provides the basis for decision making and forecasting, being of major importance around cyclical turning points. This article discusses the main characteristics of seasonal variations, common statistical seasonal models and several seasonal adjustment methods with emphasis on those widely applied by statistical agencies.

1. *The Main Characteristics of Seasonality*

The impact of seasonality in socioeconomic activities has long been recognized. Seasonal variations in agriculture, the low level of winter construction and high pre-Christmas retail sales are all well known. The four major causes of seasonality are the weather, the composition of our calendar, timing decisions, and expectations. These causes are mainly exogenous to economic systems but human intervention can modify their extent and nature. Thus, seasonal variations in the automobile industry are affected by manufacturers' decisions regarding the extent of model changeover each year; seasonality in the money supply and related series can be controlled by central banks' decisions on the interest rate. Another characteristic of seasonal variations is that although they repeat annually with certain regularity, they may evolve. Many causes can produce an evolving seasonality, for example, changes in technology alter the importance of climatic factors; customs and habits are affected by different distributions of income and thus, can modify the seasonal demand pattern for certain goods. For most time series, moving seasonality is more the rule than the exception. Depending on the main causes of the seasonal variations, these can change slowly or rapidly, in a deterministic or a stochastic manner or a mixture of both. A third important feature of seasonality is the feasibility of its identification even if it is a latent (non-directly observable) variable. The identification and estimation of seasonality, however, is not done independently of the other components affecting the series under study. Traditionally, three other components have been distinguished, namely; the trend, the cycle, and the irregulars. The trend reflects long term movements lasting many years. It is generally associated with structural causes, for example, institutional events, demographic and technological changes, new ways of organization. The cycle, usually referred to as business cycle, is a quasiperiodic oscillation characterized by periods of expansion and contraction. It lasts from 3–5 years on the average. In much analytical work, the trend and the cycle are combined because, for series covering a short period of time, the long-term trend cannot be estimated adequately. Finally, the irregulars represent unforeseeable movements related to events of all kinds. In general, the irregular component has a stable random appearance but, in some cases, it may include extreme variations or outliers with well identifiable causes, for example, floods, severe winters, strikes.

The seasonal variations can be distinguished from the trend by their oscillatory character, from the business cycle by having annual periodicity and from the irregulars by being systematic.

For series characterized by flows such as sales, production, other variations associated with the composition of the calendar have also been recognized. Of these, the most important are the 'trading-day' variations which represent the 'within-month effects' due to the number of times that each day of the week

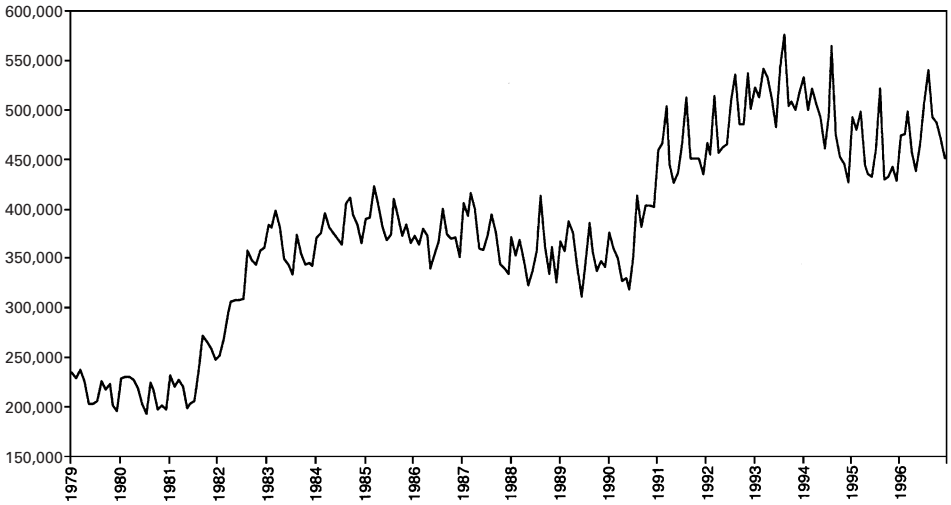

Figure 1
Original series—Canadian unemployed women over 25

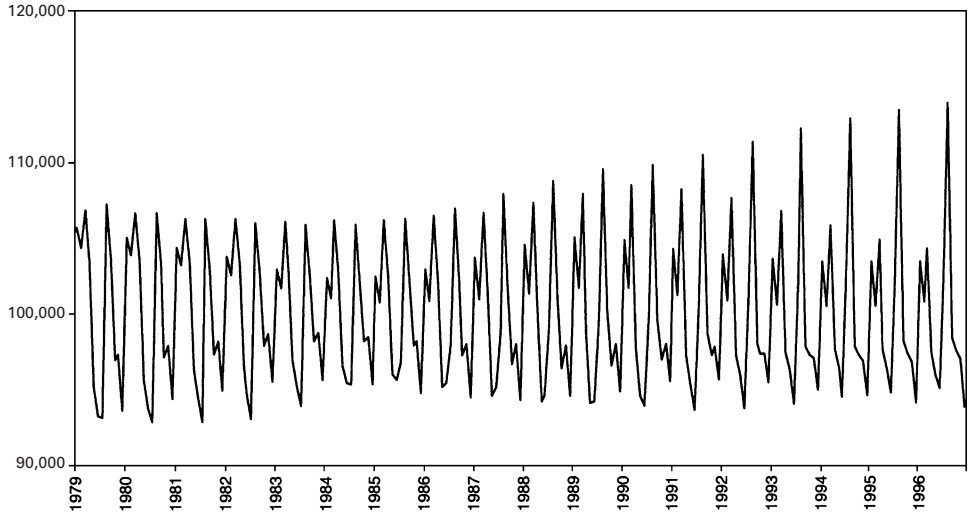

Figure 2
Seasonal component—Canadian unemployed women over 25

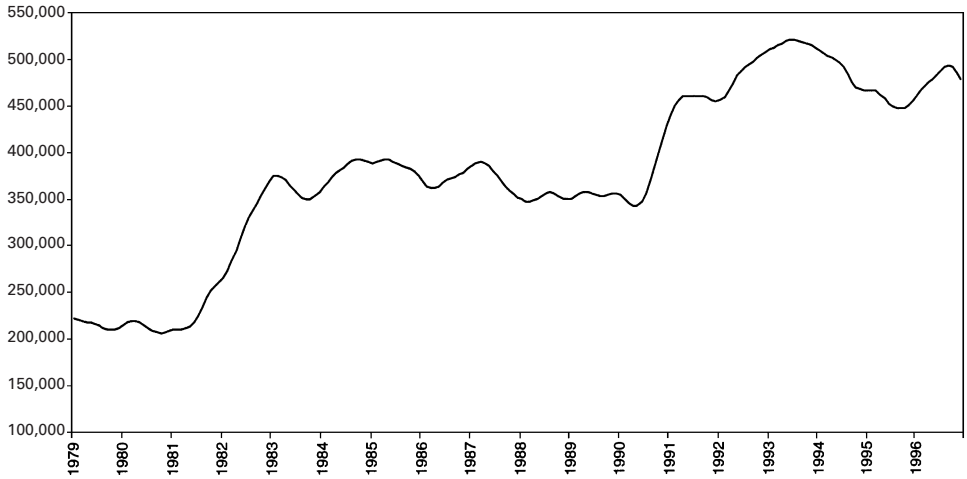

Figure 3
Trend cycle—Canadian unemployed women over 25

occurs in a calendar month; and moving holidays effects, such as those due to Easter which can fall in consecutive months or quarters depending on the year. These variations attributed to the calendar are usually treated as seasonal in character and removed together with the other seasonal variations when producing a seasonally adjusted series.

For illustrative purposes, Fig. 1 shows the pattern of an observed time series (Canadian Unemployed Women, 25 years old and over, from 1979 to 1997) which is affected strongly by both seasonality and the business cycle, but their effects cannot be inferred by simply looking at the graph. The various components estimated with the X11ARIMA method are shown in Figs. 2–4. Figure 2 exhibits the seasonal variations that are characterized by two peaks and two troughs. A large peak is observed in August and a secondary one in March whereas the maximum trough corresponds to the month of December followed closely by another in June. Furthermore, it is apparent that seasonality evolves with increasing amplitude. The effects of the trend-cycle are shown in Fig. 3 with two large jumps in 1981–82 and 1990–91 which are recession years. The irregulars exhibited in Fig. 4 are generally small except for a few cases where extreme values are present. Finally, the seasonally adjusted

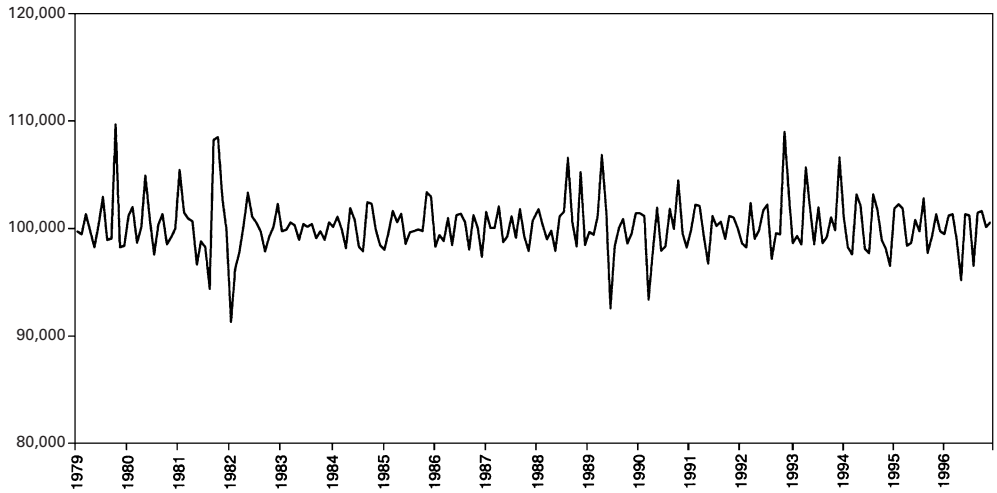

Figure 4
Irregular component—Canadian unemployed women over 25

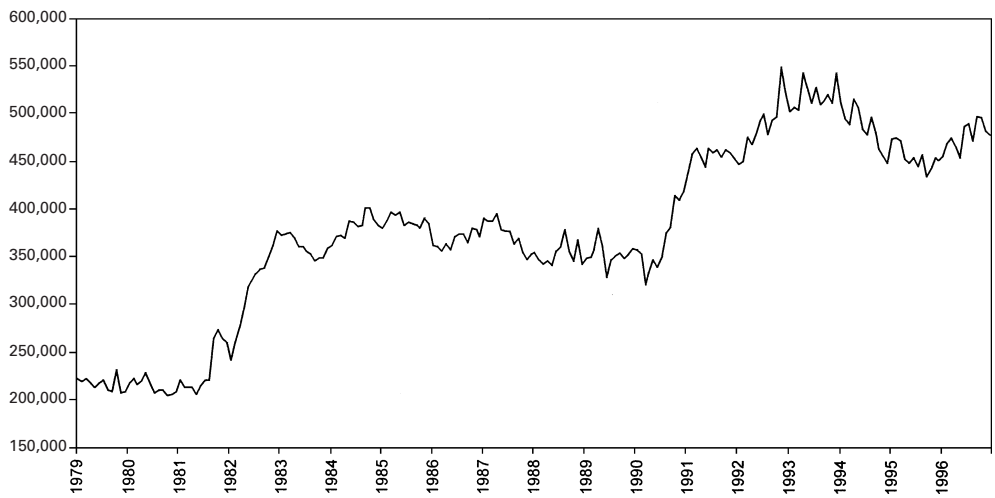

Figure 5
Seasonally adjusted series—Canadian unemployed women over 25

data are shown in Fig. 5 clearly indicating that the impact of seasonality has been removed from the original series. The seasonally adjusted series contains the effects of the trend-cycle and those of the irregular component and, therefore, it is less smooth than the trend-cycle shown in Fig. 3.

2. *Seasonal Models*

The most often discussed seasonal models can be classified as deterministic or stochastic according to the assumptions made concerning the time evolution of the seasonal pattern.

The simplest and often studied seasonal model assumes that the generating process of seasonality can be represented by strictly periodic functions of time of annual periodicity. The problem is to estimate s seasonal coefficients (s being the seasonal periodicity, e.g., 12 for monthly series, 4 for quarterly series) subject to the seasonal constraint that they sum to zero.

This regression seasonal model can be written as

$$Y_t = S_t + \varepsilon_t, \quad t = 1,\ldots, T;$$

$$S_t = \sum_{j=1}^{s} \gamma_j d_{jt} \quad \text{subject to} \quad \sum_{j=1}^{s} \gamma_j = 0 \tag{1}$$

where the d_j's are dummy variables taking a value of unity in season j and a value of zero otherwise; $\{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma_\varepsilon^2)$ is a white noise process with zero mean, homoscedastic, and nonautocorrelated. The seasonal parameters γ_j's can be interpreted as the expectation of Y_t in each season and represent the seasonal effects of the series under question. Since the parameters are constant, the seasonal pattern repeats exactly over the years, that is,

$$S_t = S_{t-s}, \quad \text{for all } t > s \tag{2}$$

Eqn. (1), often included in econometric models, assumes that the effects of the seasonal variations are deterministic because they can be predicted with no error. However, for most economic and social time series, seasonality is not deterministic but changes gradually in a stochastic manner. One representation of stochastic seasonality is to assume in regression Eqn. (1) that the γ_j's are random variables instead of constant. Hence, the relationship, Eqn. (2), becomes

$$S_t = S_{t-s} + \omega_t, \quad \text{for all } t > s \tag{3}$$

where $\{\omega\} \sim \mathrm{WN}(0, \sigma_\omega^2)$ and $\mathrm{E}(\omega_t \varepsilon_t) = 0$, for all t. The stochastic seasonal balance constraint is here given by $\sum_{j=0}^{s-1} S_{t-j} = \omega_t$ with expected value equal to zero.

Eqn. (3) assumes seasonality to be generated by a nonstationary stochastic process. In fact, it is a random walk process and can be made stationary by seasonal differencing $\nabla_s = (1 - B^s)$ where B denotes the backward shift operator. In reduced form, the corresponding regression model is a linear combination of white noise processes of maximum lag s. Since $(1 - B^s) = (1 - B)(1 + B + \ldots + B^{s-1})$, model-based seasonal adjustment methods attribute only $S(B) = \sum_{j=0}^{s-1} B^j$ to the seasonal component leaving $(1 - B)$ as part of a nonstationary trend. Thus, the corresponding seasonal stochastic model can be written as

$$S(B) S_t = \omega_t, \tag{4}$$

Eqn. (4) represents a seasonality that evolves in an erratic manner and it is often included in seasonal adjustment methods based on structural model decomposition (see, e.g., Harvey 1981, Kitagawa and Gersch 1984). This seasonal behavior is seldom found in observed time series where seasonality evolves smoothly through time. To represent this latter, Hillmer and Tiao (1982) introduced a major modification

$$S(B) S_t = \eta_s(B) b_t \tag{5}$$

where the left-hand side of Eqn. (5) follows an invertible moving average process of maximum order $s-1$, denoted by MA($s-1$). Eqn. (5) is discussed extensively in Bell and Hillmer (1984) and can be generalized easily by replacing the MA part with stationary invertible autoregressive moving average (ARMA) processes of the Box and Jenkins (1970) type. Depending on the order of the ARMA process a large variety of evolving stochastic seasonal variations can be modeled.

Other seasonal models used in regression methods where the components are assumed to be local polynomials of time, are based on a smoothness criterion. For example, given $Y_t = S_t + \varepsilon_t$, $t = 1, \ldots, T$, restrictions are imposed on S_t, such that, $\sum_t (S_t - S_{t-s})^2$ and $\sum_t [S(B) S_t]^2$ be small. The solution is given by minimizing a weighted linear combination of both criteria as follows

$$\sum_{t=1}^{T} (Y_t - S_t)^2 + \alpha \sum_{t=1}^{T} (S_t - S_{t-s})^2 + \beta \sum_{t=1}^{T} [S(B) S_t]^2 \tag{6}$$

where the ill-posed problem of choosing the α and β parameters was solved by Akaike (1980) using a Bayesian model.

Finally, mixed seasonal models with deterministic and stochastic effects are given in Pierce (1978).

3. Seasonal Adjustment Methods

The seasonal adjustment methods developed so far are based on univariate time series decomposition models with no causal explanation for any of the components involved. It is difficult to classify existing methods into mutually exclusive categories. However, it is possible to group the majority of seasonal adjustment methods into two main classes, one based on moving average techniques and the other on explicit models with a small number of parameters for each component.

3.1 Moving Average Methods

The best known and most often applied seasonal adjustment methods are based on moving averages or time-invariant smoothing linear filters applied sequentially by adding (and subtracting) one observation at a time. These methods assume that the time series components change through time in a stochastic manner. Given a time series Y_t, $t = 1, \ldots, T$, for t far removed from both ends, say $m + 1 \leqslant t \leqslant T - m$ the seasonally adjusted value Y_t^a is obtained by application of a symmetric moving average (linear smoothing filter) $h_m(B)$

$$Y_t^a = h_m(B) Y_t = \sum_{j=-m}^{m} h_{m,j} Y_{t-j}, \tag{7}$$

where the weights $h_{m,j}$ are symmetric, that is, $h_{m,j} = h_{m,-j}$ and the length of the filter is $2m + 1$. For current

and recent data ($T-m < t \leqslant T$) a symmetric filter cannot be applied and, therefore, truncated asymmetric filters are used. For example, for the last available observation Y_T, the seasonally adjusted value is given by

$$Y_T^a = h_0(B)Y_T = \sum_{j=0}^{m} h_{0,j} Y_{T-j}. \tag{8}$$

The asymmetric filters are time varying in the sense that different filters are applied for the $m+1$ first and last observations. The end estimates are revised as new observations are added because of: (a) the new innovations and (b) the differences between the asymmetric and symmetric filters. The estimates given by symmetric filters are called 'final.'

The development of electronic computers contributed to major improvements in seasonal adjustment methods based on moving averages and facilitated their massive application.

In 1954, Julius Shiskin of the US Bureau of the Census developed software, called Method I, based mainly on the works of Macaulay (1931) already applied by the US Federal Reserve Board for seasonal adjustment. Census Method I was followed by Census Method II and 11 more experimental versions (X1, X2,…, X11). The best known and widely applied is the X11 variant (Shiskin et al. 1967). This method, however, produces poor seasonally adjusted estimates for the most recent observations that are of crucial importance to assess the direction of the short-term trend and the identification of a cyclical turning point in the economy. Dagum (1978) developed the X11ARIMA method to correct for this serious limitation.

The X11ARIMA method consists of: (a) modeling the original series with an autoregressive integrated moving average (ARIMA) model of the Box and Jenkins (1970) type, (b) extending the observed series with 1–3 years of extrapolated values from an ARIMA model that fits and extrapolates the data well according to well-defined acceptance criteria, and (c) estimating each component with moving averages that are symmetric for middle observations and asymmetric for both end years. The latter are obtained from the convolution of the smoothing linear filters of Census Method II- X11 variant and those of the ARIMA model extrapolations.

For flow series, deterministic components such as trading-day variations and Easter holiday effects are estimated with dummy variable regression models and removed from the series, so that only the remainder is subject to steps (a) to (c) above. This sequential procedure is followed by the X11ARIMA/88 and X11ARIMA/2000 computer packages developed at Statistics Canada which are today the most widely applied by statistical agencies for the seasonal adjustment of official statistics. Another recent software called X12ARIMA developed by Findley et al. (1998) of the US Bureau of the Census offers a RegARIMA option that enables the estimation of the deterministic components simultaneously with the ARIMA model for extrapolation. It also includes an option for analysis of revisions to most recent estimates and spectral techniques to assess the goodness of the seasonal adjustment.

All the moving average methods here discussed are *nonlinear* and consequently, the seasonally adjusted total of an aggregated series is not equal to the sum of the seasonally adjusted series that enter into the aggregation. The main causes of nonlinearity result from the use of: (a) a multiplicative decomposition model for the unobserved components, (b) the automatic procedure for the identification and replacement of extreme values, (c) the ARIMA model extrapolations, and (d) the automatic selection of moving averages for the estimation of the trend-cycle and seasonality.

The properties of the combined linear filters applied to estimate the various components were originally studied by Young (1968) for a standard option. Later, Dagum et al. (1990) studied the properties for all the symmetric and asymmetric linear filters of Census X11 as well as those of the X11ARIMA. Although these methods lack explicit models for the unobserved components, Cleveland and Tiao (1976) and Burridge and Wallis (1984) have found ARIMA models which can approximate well some of the linear filters used for the trend-cycle and seasonal components.

3.2 Model-based Seasonal Adjustment Methods

The best known among the model-based methods are: (a) the *regression* methods with globally or locally deterministic models for each component and, (b) stochastic-model-based methods that use ARIMA models.

3.2.1 Regression methods. Seasonal adjustment by regression methods is based on the assumption that the systematic components of a time series can be approximated closely by simple function of times over the *entire* span of the series. In general, two types of functions of time are considered. One is a polynomial of fairly low degree to represent the trend component; the other, linear combinations of sinusoidal functions with different periodicity and fixed amplitude and phase, to represent the cycle and seasonality. To overcome the major limitation of using global deterministic representations for the trend-cycle and seasonality, regression methods were extended to incorporate stochastic representations of the components by means of local polynomials (spline functions) for successive short segments of the series and introduc-

ing changing seasonal amplitudes. The estimated values were still significantly affected at the beginning and end of the series by the choice of the functions for each component. A major breakthrough on this direction was made by Akaike (1980) who introduced prior constraints to the degree of smoothness of the trend-cycle and seasonality and solved the problem by a Bayesian modeling. Another important contribution is a regression method with locally deterministic models estimated by the loess (locally weighted regression) smoother, developed by Cleveland et al. (1990).

3.2.2 Stochastic-model-based methods. Stochastic-model-based methods were mainly developed during the 1980s following two different approaches. One is known as seasonal adjustment by signal extraction (Burman 1980) or as ARIMA-model-based seasonal adjustment (Hillmer and Tiao 1982), and the other referred to as structural model decomposition method (see, e.g., Harvey 1981). The major difference between these two approaches is that in the latter simple Gaussian ARIMA models are directly specified for each unobserved component whereas in the former an overall Gaussian ARIMA model is obtained from observable data and by imposing certain restrictions, models for each component are derived.

Since the components are unknown, to obtain a unique decomposition from the overall ARIMA model fitted to the observed series, Hillmer and Tiao (1982) proposed a canonical decomposition which has the properties, among others, of maximizing the variance of the white noise component and minimizing the variance of the stochastic seasonal model.

ARIMA models are very sensitive to outliers and cannot handle deterministic components such as trading-day variations and moving holidays that can be present in flow series. Therefore, further developments were made by combining dummy variables regression models with ARIMA errors to deal with these cases. In this regard, Gomez and Maravall (1996) developed software for seasonal adjustment using ARIMA models with exogenous regressors.

On the other hand, the structural model decomposition method starts directly with an observation equation (sometimes called measurement equation) consisting of the unobserved components, trend-cycle, seasonality, and irregulars (trading-day variations and moving holidays effects can also be included if present). Simple ARIMA or stochastic trigonometric models are assumed for each component. The structural model is cast in a state-space form and often estimated with Kalman filtering and smoothing. Koopman et al. (1998) developed software for structural model seasonal adjustment, which includes several types of models for each component. Model-based seasonal adjustment methods are seldom used by statistical agencies. The main reason for this is the fact that the models are very restrictive and, thus appropriate only for a limited class of series, particularly, highly aggregated data. Furthermore, they have not yet proved to be superior to official moving average methods concerning the two most important conditions for seasonal adjustment, namely, stability of the estimates and lack of residual seasonality in the seasonally adjusted data.

4. Conclusion

Seasonal adjustment plays a significant role in the analysis of current social and economic conditions, particularly in determining the stages of the business cycle at which the economy stands. Such knowledge provides the bases for decision making to control the level of economic activity and it is of major importance around turning points.

The fact that seasonality and the other time series components change through time leads to major modifications of the traditional methods based on moving averages and to the development of new ones mainly based on ARIMA models.

During the 1990s, major financial, economic, and ecological changes of a global character have introduced more variability in time series and, consequently, in their seasonally adjusted values, making it very difficult to determine the direction of the short-term trend for an earlier detection of a turning point. As a consequence, producers and users of seasonally adjusted data have centered their attention on the need for further smoothing. In this regard, new research in seasonal adjustment has been oriented toward a better estimation of the trend-cycle for current economic analysis, including parametric and non-parametric approaches as well as linear and nonlinear smoothers. The presence of many outlying observations have also stimulated research on nonlinear and non-Gaussian models for seasonal adjustment.

For information on trend-cycle estimation the reader may refer to Kenny and Durbin (1982), Maravall (1993), Dagum (1996), Dagum and Luati (2000), and references given therein. More information on seasonal adjustment theory and practice can be found in Hylleberg (1992) and Zellner (1978, 1983).

See also: Climate Impacts; Time Series: Cycles; Time–Space in Geography

Bibliography

Akaike H 1980 Seasonal Adjustment by a Bayesian Modelling. *Journal of Time Series Analysis* **1**: 1–13

Bell W R, Hillmer S C 1984 Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics* **2**(3): 291–320

Box G E P, Jenkins G M 1970 *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco
Burman J P 1980 Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society Series A* **143**: 321–37
Burridge P, Wallis K F 1984 Unobserved-components models for seasonal adjustment filters. *Journal of Business and Economic Statistics* **2**(4): 350–9
Cleveland R B, Cleveland W S, McRae J E, Terpenning I 1990 STL: A seasonal trend decomposition procedure based on LOESS. *Journal of Official Statistics* **6**(1): 3–33
Cleveland W P, Tiao G C 1976 Decomposition of seasonal time series: A model for the census X-11 program. *Journal of the American Statistical Association* **71**: 581–7
Dagum E B 1978 Modelling, forecasting and seasonally adjusting economic time series with the X11ARIMA method. *The Statistician* **27**(3): 203–16
Dagum E B 1988 *The X11ARIMA/88 Seasonal Adjustment Method. Foundations and User's Manual*. Statistics Canada, Ottawa
Dagum E B 1996 A new method to reduce unwanted ripples and revisions in trend-cycle estimates from X11ARIMA. *Survey Methodology* **22**(1): 77–83
Dagum E B, Chhab N, Chiu K 1990 Derivation and properties of the X11ARIMA and census X11 linear filters. *Journal of Official Statistics* **12**(4): 329–47
Dagum E B, Luati A 2000 Predictive performance of some nonparametric linear and nonlinear smoothers for noisy data. *STATISTICA anno LX* **4**: 5–25
Findley D F, Monsell B C, Bell W R, Otto M C, Chen B 1998 New capabilities and methods of the X12ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics* **16**(2): 127–52
Gomez V, Maravall A 1996 *Program TRAMO and SEATS*. Bank of Spain, Madrid
Harvey A C 1981 *Time Series Models*. Allan, Oxford, UK
Hillmer S C, Tiao G C 1982 An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association* **77**: 63–70
Hylleberg S (ed.) 1992 *Modelling Seasonality*. Oxford University Press, New York
Kenny P B, Durbin J 1982 Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society Series A* **145**: 1–41
Kitagawa G, Gersch W 1984 A smoothness priors—state space modeling of time series with trend and seasonality. *Journal of the American Statistical Association* **79**(386): 378–89
Koopman S J, Harvey A C, Doornik J A, Shephard N 1998 *Structural Time Series Analysis STAMP (5.0)*. Thomson Business Press, London
Macaulay F R 1931 *The Smoothing of Time Series*. National Bureau of Economic Research, New York
Maravall A 1993 Stochastic linear trends: Models and estimators. *Journal of Econometrics* **56**: 5–37
Pierce D A 1978 Seasonal adjustment when both deterministic and stochastic seasonality are present. In: Zellner A (ed.) *Seasonal Analysis of Economic Time Series*. US Department of Commerce Bureau of the Census, Washington, DC
Shiskin J, Young A H, Musgrave J C 1967 *The X-11 Variant of the Census Method II Seasonal Adjustment Program*. Technical Paper 15 (revised). US Department of Commerce, Bureau of the Census, Washington, DC
Young A H 1968 A linear approximation to the Census and BLS seasonal adjustment methods. *Journal of the American Statistical Association* **63**: 445–57
Zellner A (ed.) 1978 *Seasonal Analysis of Economic Time Series*. US Bureau of the Census, Washington, DC
Zellner A (ed.) 1983 *Applied Time Series Analysis of Economic Data*. US Bureau of the Census, Washington, DC

E. B. Dagum

Time–Space in Geography

Geographers have always regarded space and time as fundamental attributes of reality. However, significant variations have existed in the meaning associated with these concepts. The independence of space and time, where each concept provides a separate frame of reference for positioning objects and events, is seen in both regional studies and in attempts to cast geography as a purely spatial science. In contrast, the merger of time and space into the singular notion of time–space has for many researchers become the ideal in concept if not in practice.

1. *Absolute vs. Relative Views of Time and Space*

Immanuel Kant (1724–1804), less noted for instruction in physical geography at the University of Königsberg than for his broader theory of knowledge, was instrumental in laying a philosophical basis for treating time and space as *a priori* separate from each other, with absolute dimensions validated through human experience. In contrast to Kant's view, the inseparability of space and time is embraced in the concept of time–space as a single definable dimension of reality, where one cannot have space without time, and vice versa. This perspective, one of the foundations of relativity theory, conceives time–space in relationship to the movement of material objects and to events. Conceptually, the coincidence of time and space is easy to grasp—any object or event embraces simultaneously positions in time and in space (e.g., tossing a ball and observing its flight).

From an operational and measurement perspective, researchers often return to the Kantian view, treating the separation of time and space as a practical necessity. Geographers have attempted to work around this difficulty in two principal ways. First, explanatory and descriptive models (e.g., Walter Christaller's central place theory) may be used to consider the human settlement system at different points in time as a means of detecting historical changes in the spatial character of the landscape. Second, researchers may abstract the time dimension of change into periods or stages, sometimes in arbitrary correspondence to the processes being investigated. An example of the later is use of census data, where changes might be assessed over a 10-year period,

such as population growth within a region between 1990 and 2000. In contrast, frequently researchers also define periods of time or stages according to the dominance of certain social circumstances (for instance, the baby-boom period) or of prevailing technologies, such as the automobile era.

2. *From Spatial to Time–Spatial Perspectives*

Attempts in the 1950s and 1960s to recast geography from a descriptive to a predictive science offer a clear demarcation in the progression of thinking about time–space in geography. In the traditional paradigm of regional descriptive geography, time and space were treated as separate even though attention was given to documenting landscape conditions at different points in time. Theoretical quantitative geography emerged to counter the descriptive focus of this tradition, but many researchers, nonetheless, retained the separation of time from space in what was termed the exceptionalism of the spatial paradigm. This was most pronounced in the advocacy of Schaefer (1953) and Bunge (1962) for whom geography was a theoretical science in search of spatial morphological laws. The merger of geography's theoretical thrust with practical issues of society, however, necessitated a predictive perspective.

Interest in prediction resulted in a shift of emphasis from the depiction and explanation of geographical patterns to that of process identification to understanding the dynamics of regional economies and the nature of human spatial behavior. Thus, from a fundamental focus on pure spatial form, attention turned to the movements of objects and to the passage of events through space and time. Models to simulate urban growth and other phenomena as processes of spatial diffusion (see *Diffusion: Geographical Aspects*) (represent explicit attempts to mesh the time and space dimensions in geographical research. Significant statements advancing time–space perspectives are found in Hägerstrand's (1973) conceptualizations of spatial diffusion and time-geography, Janelle's (1969) formulation of time–space convergence, and Harvey's (1990) notion of time–space compression.

At the behavioral level, Hägerstrand's time-geography model of society stands out in its attempt to capture the movement and activity participation of people through graphic representation of their simultaneous and sequential space and time locations. The concept of an individual's time–space path portrays the continuity of human existence as a line that occupies space and time simultaneously and continuously, relative to a person's activities in space through the temporal cycles of the day or a lifetime. The inherent Euclidean nature of this space is a consequence of treating time and space as homogeneous media for movement, broken up in response to social constraints that determine who has access to specific spaces at given times and under what circumstances.

3. *Time–Space Convergence and Human Extensibility*

Time–space convergence is a measure of change in the effort required to overcome distance. Generally, it is defined as the average rate of decline in travel time between two places over time. For example, owing to innovations that span the eras of stagecoach travel to automobiles on limited-access highways, Boston and New York have approached each other at an average annual rate of 22 min per year between the years 1800 and 2000. The parallel concept of time–space divergence provides for the tangible measurement of increases in separation between places that result from traffic congestion and failure to maintain infrastructure. Implicit in the convergence/divergence viewpoint is the recognition that physical points (places on the earth) are in relative motion with respect to one another whenever functional measures (such as travel time and cost) are used as the distance metric. Because levels of investment in transportation and communication technologies are not uniform, some places converge much more rapidly with the remainder of the settlement system than others. Some places, bypassed by modern space-adjusting technologies, will diverge, and others will converge and diverge selectively relative to various parts of the urban system. Variations in the set of rates that define convergence among all possible pairs of places create a nonhomogeneous time–space that is relative and dependent on the inferred movement of places with respect to one another. In this time–space, places are seen to compete in their quest to gain relative accessibility over other places, thereby gaining advantages for attracting economic and other opportunities.

Time–space convergence is analogous to the physical concept of velocity and purports to assess the earth-shrinking impact of technological investments to overcoming distance. As a time-rate changes over time, it fits in the domain of metaconcepts. It is not a velocity of sensation; rather it is a process identifiable and abstracted only through measurement. The structure of the resulting space, the frame of reference for human activities, is constantly in flux and is not describable in terms of Euclidean geometry. Even though this abstraction of reality may not be recognized intuitively, the time–space convergence/divergence process embraces explicit recognition of its social origins as the result of deliberative innovation through investment decisions. It is this human behavioral response to time–space convergence that reshapes the space of human geography. Yet, Forer (1974) demonstrates reciprocity in this process, whereby the existing spatial structure of society also

influences emergent patterns of investment to encourage time–space convergence.

As a complement to time–space convergence, Janelle (1973) advanced a reciprocal concept of human extensibility. Developed more fully by Adams (1995), this refers to the ability of an individual or agency to extend their influence and presence over space and through time to distant locations via media and other means of influence over human resources. Human extensibility parallels the construct of time–space distanciation as advanced in Giddens's (1984) structuration theory. Together, these concepts pose questions regarding the closure or spatial boundedness of human societies and systems. They also bring forward issues concerning the differential abilities of individuals, firms, nations, and other forms of human agency to reach beyond their physical locations to shape possibilities for human livelihood elsewhere. Such discrepancies lead Harvey (1990) to advance yet another conception of time–space, one rooted in the historical materialism of the political economy.

4. *Time–Space Compression*

Harvey advances the notion of time–space compression and provides a compelling rationale for imbuing time–space with social content and experiential meaning. This is brought about by attempts within Capitalism to extract as much surplus value per unit of time as possible. Though annihilation of space through time is analogous to time–space convergence, the broad implications of compression may be less amenable to direct measurement. Use of the term compression captures the general speeding up of all kinds of events that intensify the turnover of capital and that implode on human existence as substantive levels of stress. Furthermore, in Harvey's interpretation, the transcendence of global capital over local domains thrives on shortened time horizons and on the ability to eradicate distance as an inhibitor to exchange and dominance. The ability of capital to shift resources to different places easily, and often with impunity to the disruption of life at the local scale, suggests that time-space compression may denigrate the importance of place in human society. It is in the inherent mobility and speed of such possible actions that time–space compression challenges any attempts to define the time–space nature of society too rigidly.

5. *Time–Space, Dynamic Cartographic Visualization, and the Humanist Challenge*

One of the issues that hinders broad application of time–space perspectives in the social and behavioral sciences is the difficulty inherent in representing the underlying complexity in the human organization of space through time. This is both an issue regarding the nature of the tools at the disposal of researchers and of how questions in empirical research are linked within alternative theoretical frameworks. The integration of space-time analyses and modeling within the frameworks of dynamic mapping, cartographic visualization, and Geographic Information Systems (see *Dynamic Mapping in Geography*; *Geographic Information Systems: Critical Approaches*) holds promise that research tools finally are catching up with needs of social and behavioral scientists for assessing alternative theoretical constructs. The transfer of these technologies to the interactive realm of the Internet offers the potential for incorporating simultaneous representation of processes through space and time. Both real-time and simulated processes are possible. Driving the interest in such developments are the needs for planners, policy makers, and scholars to understand the dynamic time–space transformations that are occurring among places and regions as a consequence of continuing advances in information and communication technologies. However, these transformations pose fundamental challenges to how social and behavioral scientists interpret the meaning of place in everyday life and to how they integrate these interpretations within theoretical and decision-making frameworks (see *Place in Geography*; *Space and Social Theory in Geography*).

Do the meanings of space and the meanings of time mean anything aside from their relationship to one another as time–space? The theoretical grounding of answers to this question rest, in part, on the ability of researchers to use the emerging tools of analytical representation in their descriptions and analyses. At the same time, answers rest fundamentally on being able to recognize humanist qualities in the nature of time–space (e.g., temporality and spatiality, as discussed in Kellerman (1989)). The humanization of time–space acknowledges its relativity to different value systems, recognizes its manifold possible metrics, and is able to relate these to human thought processes and experiences. Casting time–space in this manner draws on past formulations and opens the realm of possible investigations to include analytical time–space prediction as well as research formulations grounded in the approaches of critical science and postmodernism.

Bibliography

Adams P C 1995 A reconsideration of personal boundaries in space-time. *Annals of the Association of American Geographers* **85**: 267–85

Bunge W 1962 *Theoretical Geography*. Gleerup, Lund, Sweden

Forer P 1974 Space through time. In: Cripps E L (ed.) *Space-time Concepts in Urban and Regional Models*, Pion, London, pp. 22–45

Giddens A 1984 *The Constitution of Society: Outline of the Theory of Structuration*. Polity Press, Cambridge, UK

Hägerstrand T 1973 The domain of human geography. In: Chorley R J (ed.) *Directions in Geography*, Methuen, London, pp. 67–87

Harvey D 1990 Between space and time: reflections on the geographical imagination. *Annals of the Association of American Geographers* **80**: 418–34
Janelle D G 1969 Spatial reorganization: a model and concept. *Annals of the Association of American Geographers* **59**: 348–64
Janelle D G 1973 Measuring human extensibility in a shrinking world. *Journal of Geography* **72**: 8–15
Kellerman A 1989 *Time, Space and Society: Geographical Societal Perspectives*. Kluwer Academic Publishers, Dordrecht
Schaefer F K 1953 Exceptionalism in geography: A methodological examination. *Annals of the Association of American Geographers* **43**: 226–49

D. G. Janelle

Time-use and Gender

Men and women have to make decisions on how to use their time, as a day has only 24 hours, or rather time is limited. Activities men and women make decisions about can be divided into paid work, domestic work and childcare, personal obligations, and leisure. Time has become an increasingly scarce good; newspapers and other media abound with stories suggesting that time pressure is becoming a real epidemic. The article discusses how time use is measured and the patterns of time use of men and women and provides explanations of gender differences.

1. The Measurement of Time-use

To understand time use of men and women, it can be categorized in several ways. If one is interested in paid work and leisure, a distinction is made between work vs. nonwork activities. With respect to the division of labor between husbands and wives, the most frequently used division is between paid work, domestic work and childcare. It is evident what is meant by paid work, and most authors use more or less the same description. However, household labor can be defined in a variety of ways. Naturally, cooking and cleaning are domestic tasks, but it isn't clear how to classify activities that are part leisure and part work, such as vegetable gardening, gourmet cooking, or playing with the children. Leisure is a broad category, including indoor activities like reading books and watching television, outdoor activities like sporting, but also social contacts and going out for dinner. Personal obligations often form a kind of residual category, including activities like sleeping and eating at home.

The most accurate and reliable way to measure the time use of people is the time diary method (Andorka 1987, Juster and Stafford 1991, Harvey 1993). In this approach, respondents are asked to keep records of what they were doing at all times for a day or more. In some studies the main activity is noted in fixed interval units (e.g., in quarters of an hour) while in others intervals are left free (i.e., starting and finishing times are indicated). The period of time when the diary is kept also differs. In some cases respondents are only required to keep a diary for a single day, while in others they are asked to do so for a whole week. The time diary method prevents respondents from exaggerating the durations of their favorite activities while neglecting the time devoted to day-to-day concerns and trivialities. Thus, the actual subject being studied is not overstated, which sometimes happens when specialized questionnaires are used. A serious limitation of time diaries is the difficulty in dealing with more than one activity performed simultaneously (Warner 1986). For example, people may cook dinner while supervising children, or sew while watching television. Another method to measure patterns of time use is asking direct questions that require men and women to estimate their usual time spent on a list of activities. In order to help the respondent, in these questionnaires often domestic work is distinguished into some clear categories like cooking, cleaning, doing the laundry, and shopping. Those few studies that compare time-diary to direct-question data find that direct questions typically produce higher time estimations than time-diary questions (Juster and Stafford 1991), especially for activities that occur frequently (Marini and Shelton 1993). Both husband and wife overestimate their time devoted to domestic duties.

2. Time-use Patterns

There has been a traditional gender division of time consuming activities within western societies: women used to do most of the household work and childcare, whereas men worked for pay the whole day and were responsible for providing family income. This has changed substantially in recent decades. In the western world, women's paid work has increased vastly. Now a majority of women in many nations are employed for pay, something that would have been hard to imagine just after the Second World War. Despite the overall increase in levels of women's employment large differences exist between countries in the northern continents. Nordic countries, such as Sweden and Denmark, have had higher levels of female labor force participation than other western countries for a long time. The Anglo-Saxon countries, such as the USA and the UK were runners up, while in Italy, Spain and Greece still less than half of the active female population has a paid job. For Eastern Europe, the picture is different. The communist ideology, established after the Second World War in Eastern Europe with the exception of Russia, forced many women to enter the labor market. In the 1960s already three quarters of the women had a paid job. Despite the

overall increase in the level of women's employment there are considerable differences between countries in the number of hours per week that women work for pay. While part-time work is common in Sweden, Denmark, and the Netherlands, employed women in Italy, Greece, Spain, and Portugal almost all work full-time. About 80 percent of US employed women work full-time. In Eastern Europe part-time work simply did not exist before the political turnover. After 1989 full-time work is still the rule in the former socialist countries (van der Lippe and Fodor 1998).

The growth in paid work of women is linked to the often noted decline in domestic work on the part of women in western countries (Gershuny and Robinson 1988, Gershuny 1992). Compared to this large change in women's time use in the west in general, there has been hardly any change in men's time use. Although there has been a small increase in men's domestic work, these changes are trivial compared to the increase in women's employment and the reduction in their household work. The more equal division of domestic work between husbands and wives nowadays is less a result of an increase in men's domestic work and more because of the reduced time women spend on household duties. When men perform domestic tasks, they disproportionately do what are often considered more 'pleasant' tasks (e.g., childcare and cooking rather than cleaning the toilet).

The amount of leisure time differs not so much between the sexes as between those in paid and domestic work. Between the 1960s and the 1990s for both sexes leisure time accounted for between 20 to 30 percent of weekly time (Robinson 1991). Men do rather more outdoor activities such as sport, whereas women spend more time on social contacts. Some authors argue that the workweek in most industrialized countries has been reduced and therefore the amount of free time has increased (Robinson 1991, Gershuny 1993). However, others point to a reduction of leisure and household time as more women are employed and working hours increase for some men (Schor 1998a). According to Schor 'the quality of European life is threatened as relatively short working hours, relaxed daily life and appreciation for the "finer things" are giving way in a number of countries to longer hours of work, a faster pace of life, and American style consumption' (Schor 1998b, p. 7). In two-earner families both partners are exposed to the greed associated with the jobs in which they work, and thus both will bring the restrictions imposed by their work situation to the time allocation in the household.

3. *Explanations for the Different Use of Time by Gender*

A rich body of literature is available on men and women's allocation of time (Szalai 1966, Robinson 1977, Gershuny 1993, Shelton and John 1996). Many study men's and women's division of paid and domestic work (Coverman 1985, Huber and Spitze 1983, Presser 1994) and some of these studies compare several countries.

3.1 *The Individual Level*

An important theoretical explanation of time allocation has been provided by economic theory (Becker 1965). Given their constraints or resources human beings strive for goals and make choices in their time use. An obvious resource within economic theory is human capital. Human capital constitutes the knowledge and skills that increase an individual's productivity. Education is often considered the most important form of human capital. So, for highly educated women and men, it is appealing to devote a great deal of their time to working in the labor market, since the benefits in terms of money and status can be high. This implies less time for domestic work and non-work activities. To understand the division of paid and domestic work between husbands and wives, economic theory is valuable as well. Within the 'new home economics' explicit attention is given to the importance of household activities (Becker 1981). In this model of household production, households are viewed as small factories that, within the time and income limitations, make every effort to maximize their output. If the husband can earn more money and is less productive in household labor than his wife, the husband will specialize in paid labor and the wife in household labor (Becker 1981). In Becker's view, which does not highlight any conflicts of interest between spouses but assumes altruism, each spouse specializes in what will make his or her best contribution to the family.

Sociological exchange theory (Blood and Wolfe 1960, Brines 1994) provides a different account of spousal specialization. Differences in resources, such as education and income, are considered to be a main cause for the different time use by men and women. Assuming that individuals do not like to do domestic work, this theory predicts that the spouse with the greater resources will have more power and will use it to minimize his or her domestic work. Thus the predictions are similar to those of the new home economics theory, that spouses with higher relative earnings will do less household work, but the reasoning is different. Whereas in economic theory, each spouse chooses between market and household work in order to maximize contributions to the family, exchange theory typically treats employment and earnings as exogenous and examines how they affect bargaining power, and sees doing household work as an indication of less power.

Many studies in the US found that men's contribution to housework reduces the higher their earnings relative to female spouses (Ross 1987, Presser

1994), although men who are economically dependent on their wives compensate by adopting traditional behaviors in the household (Brines 1994). With respect to other resources such as education or occupational prestige, results are less consistent with exchange theory. Results with respect to the link between the participation of women in paid labor and the division of household labor are inconsistent as well. Some studies find that men married to employed women spend more time on housework (Coverman 1985), while others did not find significant effects (Shelton 1990).

Another important theoretical perspective emphasizes gender specific socialization (Moore 1985). It is not such a long time ago that women were socialized to be mothers and housewives and men to be wage-earners and support the family. More recently there is a great deal of variation in the degree of gender traditionality of socialization. Research suggests that men who adhere to traditional sex roles contribute less to household tasks than men whose sex roles are nontraditional (Ross 1987, Kamo 1988, van der Lippe and Siegers 1994) and it has also been noted that women with modern role expectations do more paid and less domestic work (Brayfield 1992, Presser 1994). Although gender roles are measured in a variety of ways, most studies indicate that men's attitudes are more strongly associated with the division of domestic work than are women's but that ideology is certainly not the only time use determinant.

These theories can also explain the change in time use of men and women in recent decades. Given that the educational level of women all over the western world has increased and is nowadays comparable to the level of men, according to the resource or productivity argument it has become attractive for women to spend more time on paid work. Consequently they have less time available for domestic work and leisure. Furthermore, attitudes in the western world have changed. Nowadays, working women are accepted in society and fathering has become one of the opportunities for men. According to the sex-specific socialization it has become more likely as well that men carry more of the domestic workload than some decades ago, and women make a career. Indeed we do see an increase in time spent in childcare by men, but rather unexpectedly also by women. Given the sociological and economic explanations, one would expect the opposite for women. It might be that the psychological meaning of children has changed. Children in our modern society are a highly valued 'good,' in which one wishes to invest a great deal of time. Moreover, the growing threat of the environment might play an important factor. The number of cars has increased continuously with the consequential danger of teaching children to play unattended or travel to activities alone. Finally, it might be an artifact of the methodology, due to the fact that only first activities are reported. Full-time housewives in unmechanised households used to do several household and caring tasks simultaneously, but reported only the first activity.

3.2 *The Institutional Level*

In addition to individual factors, differences in the institutional context are important for time use decisions of men and women. Although time use decisions are made on a microlevel, the institutional context of a country can facilitate or restrict a certain time use decision. With respect to western countries and Eastern Europe, one of the most obvious differences has been the political system. In the state socialist period, women were encouraged to have a full-time job irrespective of their individual circumstances. To enable women to have a full-time job, extensive childcare facilities were available. There are differences among western countries as well, which become clear using Esping-Anderson's typology (1990) of Welfare State regimes (liberal, conservative, social democratic), or related more gender friendly typologies. In social democratic countries, such as the Scandinavian ones, public childcare facilities are more extensive and women are stimulated more to have a paid job. In conservative countries, for a long time, women were not encouraged to have a paid job, and consequently childcare was not available. Indeed in countries with less well-developed childcare arrangements women with young children are less likely to be employed than women with older children (Gornick et al. 1998).

4. *Future Research*

Two directions require attention for future research. First, we need to study how paid work, childcare, other domestic work, and leisure compete for people's time and how gender-related patterns are changing. The issue of time competition has to be studied in its full sense, covering both the demands from the work organization and from the private life. Second, in line with the disappearing boundaries between countries and states, due to the increase in communication and information technology, comparative studies on time use become increasingly important. At this moment information is available on many western industrialized countries, but one can only hope to compare their figures with those of Asian, South American, and African countries.

See also: Family and Gender; Family Bargaining; Household Production; Motherhood: Economic Aspects

Bibliography

Andorka R 1987 Time budgets and their uses. *Annual Review of Sociology* **13**: 149–64

Becker G S 1965 A theory of the allocation of time. *The Economic Journal* **80**: 493–517
Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
Blood R O, Wolfe D M 1960 *Husbands and Wives: The Dynamics of Married Living*. The Free Press, New York
Brayfield A 1992 Employment resources and housework in Canada. *Journal of Marriage and the Family* **54**: 19–30
Brines J 1994 Economic dependency, gender, and the division of labor at home. *American Journal of Sociology* **100**: 652–88
Coverman S 1985 Explaining husbands' participation in domestic labor. *The Sociological Quarterly* **26**: 81–97
Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Princeton University Press, Princeton, NJ
Gershuny J 1992 Are we running out of time? *Futures* **24**: 3–22
Gershuny J 1993 Post-industrial convergence in time allocation. *Futures* **25**: 578–86
Gershuny J, Robinson J P 1988 Historical changes in the household division of labor. *Demography* **25**: 537–52
Gornick J C, Meyers M K, Ross K 1998 Public policies and the employment of mothers: A cross-national study. *Social Science Quarterly* **79**: 35–54
Harvey A S 1993 Guidelines for time use data collection. *Social Indicators Research* **30**: 197–228
Huber J, Spitze G 1983 *Sex Stratification: Children, Housework and Jobs*. Academic Press, New York
Juster F T, Stafford F P 1991 The allocation of time: Empirical findings, behavioral models, and problems of measurement. *Journal of Economic Literature* **29**: 471–522
Kamo Y 1988 Determinants of household division of labor: Resources, power, and ideology. *Journal of Family Issues* **9**: 177–200
Marini M M, Shelton B A 1993 Measuring household work: Recent experience in the United States. *Social Science Research* **22**: 361–82
Moore J C 1985 Role enactment and self-identity. In: Berger J, Zelditch M Jr (eds.) *Status Rewards and Influence, How Expectations Organize Behavior*. Jossey-Bass, San Francisco
Presser H B 1994 Employment schedules, gender, and household labor. *American Sociological Review* **59**: 348–64
Robinson J P 1977 *How Americans Use Time: A Social–Psychological Analysis of Everyday Behavior*. Praeger, New York
Robinson J P 1991 Trends in free time: A cross-national comparative analysis for seven industrial countries 1961–1985. In: O'Conhaile W, Kohler G (eds.) *The Changing Use of Time: Report from an International Workshop*. European Foundation, Dublin, Eire
Ross C E 1987 The division of labor at home. *Social Forces* **65**: 815–33
Schor J B 1998a *The Overspent American: Upscaling, Downshifting, and the New Consumer*. Basic Books, New York
Schor J B 1998b Beyond work and spend. *Vrijetijdstudies* **16**: 7–20
Shelton B A 1990 The distribution of household tasks: Does wife's employment make a difference? *Journal of Family Issues* **11**: 115–35
Shelton B A, John D 1996 The division of household labor. *Annual Review of Sociology* **22**: 299–322
Szalai A 1966 The multinational comparative time budget research project. *American Behavioral Scientist* **9**: 1–20
van der Lippe T, Fodor E 1998 Changes in gender inequality in six Eastern European countries. *Acta Sociologica* **41**: 131–49
van der Lippe T, Siegers J J 1994 Division of household and paid labor between partners: Effects of relative wage rates and social norms. *KYKLOS* **47**: 109–36
Warner R L 1986 Alternative strategies for measuring household division of labor: A comparison. *Journal of Family Issues* **7**: 179–95

T. van der Lippe

Time-use: Research Methods

1. *Why Study Time Allocation?*

We all have 1440 min in our days. While we can buy goods, and the services of other people, we cannot directly purchase more minutes for ourselves. However, as we buy goods and services, and sell our own labor, we are trading our own time, and other people's. Furthermore, our position in this market for time reflects aspects of our own past histories of time use—how much we spent in interaction with parents and siblings in our households of origin, time in schools and training, in various sorts of work and leisure activities. These past times accumulate, through the life course, to form that aspect of human capital that determines prospective earnings rates. Those whose hourly wage is greater than the average may wish to sell a relatively large part of their day as paid labor, since their earnings can, in effect, buy a still larger amount of time indirectly, as embodied in goods and services.

So the apparently equally distributed resource is, paradoxically, both a source of and a fundamental expression of social differentiation. How we spent our time in the past provides the human capital that helps to determine our current time allocation, not just between work and leisure, but also, through mechanisms set out by Becker (1965) and Bourdieu (1985), amongst different sorts of leisure consumption. And employment-related human capital in turn effectively redistributes time: those below the average wage who have no financial capital must exchange each hour of their work for goods and services that embody on average less than an hour of the work of others, while those with high wages may exchange many hours of embodied labor for each hour of their work. Once we subtract paid work time from the day, and add back the time embodied in purchased goods and services, the poorest members of a society might have effective control over 15 hours each day and the richest 50 hours or more.

Yet despite this crucial role in the emergence of social structure, time allocation evidence does not at present form part of the core research material of any social science discipline. Why?

One contributing explanation is the paucity of good data. We have limited knowledge of our own time use. Like fish in water, we live in time; we are *in* it, but not aware *of* it. We are not well equipped to mark its passage or duration, and in any case the social scientist is interested not in time itself but in the particular activities to which successive periods are devoted. So if, for example, we ask survey respondents directly how much time they devote to a particular activity over a given period, the answers we get reflect little more than conventional expectations ('I work 40 hours per week' our respondent tells us, because her employment contract states this—but the actual hours worked are only accurately known if there is some externally imposed mechanism for registering this). For many categories of activity (housework, leisure activities, travel) respondents are unable to provide any straightforward answers at all (unless to report complete nonparticipation). There is no particular reason in the course of day-to-day living why we should know such an abstraction as the amount of time devoted to each element in a general classification of activities, so we cannot give good answers to researchers when they ask us about it.

2. *Methodological Alternatives*

There are four distinct approaches to the measurement of time use in a population: 'stylized estimates' as part of questionnaire inquiries, observational methods, the use of administrative records, and diary methods.

Despite the evident fact that people are unaware of their time allocation, much of what currently passes for scientific knowledge of time allocation nevertheless relies on direct questions to survey respondents. In particular, the standard models for the Labor Force Surveys collected by most developed countries routinely ask for estimates of amounts of time, on average, or over a typical week, or during the previous week, spent working in paid employment. Other surveys ask similarly about time devoted to, for example, child care and domestic chores. These produce 'stylized estimate' measures of time use.

This approach is certainly the simplest and cheapest way to collect information about how people use their time, but it has significant drawbacks. Research on the paid work time estimates deriving from this methodology suggests substantial and systematic biases (e.g., Robinson and Godbey 1997, pp. 58–9). Those who claim to work in excess of 40 hours per week progressively overstate the length of their work-week. The reason for this bias seems to be that as work time increases, it become more difficult to accommodate other necessary or desirable activities outside scheduled work times. These exceptional events within conventional work-start and -finish times are often forgotten, whereas exceptional events that lengthen the working week (e.g., overtime) are more likely to be remembered in the context of a question about work time. The longer the actual work week, the more likely it becomes that activities that might otherwise be located in 'free' time (e.g., medical visits, shopping, sports participation) are inserted during periods that the respondent normally expects to spend working, and the greater the proportional overestimate; hence the systematic bias.

Ethnographic research has often involved direct observation of people and the recording of their activities by the researchers. New technology now extends the role of direct observation, including such unobtrusive devices as individual location monitoring systems and closed-circuit or internet camera systems.

Observational methods have certain advantages, in that the researcher can ensure that respondents do not forget or suppress particular activities of interest. In the case of new technology, the option of reviewing tapes and videos can ensure accurate recording of people's activities; ambiguities can be reassessed until a maximally precise determination can be reached. Such methods may be useful for analyzing small samples of special groups (residents of a retirement home or a small rural community) or activities in a specific location (such as the numbers of people who pass by a shop display and react or who engage in various activities in a town center).

On the other hand, direct observation studies are costly, they are in general labor intensive, and they may compromise personal privacy. Moreover, people can be expected to censor or distort their normal patterns when aware of continuous monitoring (the more so if they are conscious of an observer of differing nationality, class, sex, ethnicity, age group, or religion). However, continuous remote monitoring of medical conditions is gaining wide acceptability, and some extension of this, perhaps in conjunction with 'beeper' techniques (or, more formally, 'experience sample methods'), in which respondents record current activities on random occasions during the day indicated by a 'beep' from a portable device, may grow in the future. But for the moment, for cost and other reasons, such approaches are not widely used.

Less direct observational methods, or 'unobtrusive measures,' include programs tracking the use made of computers at work and school (measuring the time people spend word processing, performing data analysis, checking e-mail, browsing the web, and so forth), telephone call and retail transaction records, and the like, recording what people do at particular times and in particular places.

Administrative records, some of which cover considerable periods of time, chiefly concern the official records of hours during which people are in engaged within institutions; most typically this involves employers' records of work hours, but could in principle now involve any of those recreational facilities, hospitals, and prisons that use electronic

methods to monitor individuals' entries and exits. Administrative records typically provide large numbers of cases for analysis. However, they only cover the population which has contact with the institutions keeping the records, and do not indicate if people have simultaneous contact with related institutions at the same time (e.g., working in two jobs). They tend to contain a bias against short spells, and the contextual information about personal and socioeconomic characteristics typically needed for academic or policy research is sparse.

Many of the problems with the foregoing methods are solved if respondents are asked to provide a complete account of the sequence of events, during one or more days, within a framework of clock-times. This 'diary' approach, unlike the stylized estimate method, involves a narrative skill that is often called for in the course of daily life. Diary methods are by far the most widely used techniques for the estimation of time-use patterns: there are at least 400 such studies, from at least 80 countries, and more than half of these countries can provide at least one full national random sample of diary-based time-use evidence.

3. Diary Methods

The diary method provides a number of clear advantages, which stem from the fact that the narratives mirror, in a straightforward way, the structure of daily activities. Diaries require the respondent to engage in an active process of recall, in which they reconstruct the successive events of the day, so that the particular occurrences such as a trip to the doctor that disturbs the normal passage of the working day, emerge naturally as the respondent recollects 'what happened next.' Diary accounts are complete and continuous, so any respondent intending to suppress some activity must make a conscious effort to substitute some other activity. This will also tend to limit exaggeration of the duration of particular activities, since the potential for double counting (e.g., of a period as both work and leisure) that is entirely possible using the stylized estimates, is precluded in diary accounts. There are other advantages in the diary-reporting format for the analyst. Increasingly researchers are concerned to investigate the activity sequences themselves, or the times of the day at which activities occur; simultaneous activities can be properly investigated, and if multiple diaries are collected within a single household, researchers can use them to investigate patterns of copresence, interdependence, and cooperation.

The time-diary method itself is nevertheless not unproblematic. In particular, it produces some 'under-reporting of socially undesirable, embarrassing, and sensitive' behaviors, including personal hygiene, sexual acts, and bathing (Robinson and Godbey 1997, p. 115). Techniques for persuading people to account completely for their days need some refinement.

4. Diary Data Collection

There are wide variations in the design of time-use diaries. The summary results seem fairly stable with respect to the different methods, and comparisons of different methodologies do not suggest any unequivocally dominant methodological choices (Harvey 1999, however, provides general-purpose recommendations). Choices are problematic (cost issues apart) because particular design characteristics affect respondent burden and hence response rates. There is a straightforward trade-off between a 'heavy' design, which yields a fuller picture of individuals' activities but relatively low response, or a 'light' design providing a less complete picture, but with a higher response rate.

For example: diaries may cover longer or shorter periods. The problem, which is unusual in survey research terms, insofar as most personal characteristics are fixed at any historical point, is that individuals' activities differ from day to day. Suppose all respondents do a weekly clothes wash, on one, randomly chosen, day of the week: a single-day diary will suggest that one-seventh of the sample spend a lot of time in laundry, and six-sevenths do none, where the reality is that all do on average a small amount. A short diary gives a spurious impression of interpersonal variability which is in fact largely intrapersonal. A week-long diary would reveal the real situation, but is of course very demanding for respondents.

There are two alternative modes of diary administration, conventionally referred to as 'yesterday' and 'tomorrow' studies, the former involving an interviewer-led and recorded process and the latter a self-completion instrument. The 'yesterday' diary normally does cover the previous day, as recall of events decays fairly rapidly, but it is practicable to collect useable information after a gap of two or even three days, and difficulties in recruiting respondents over the weekend may mean that this is necessary so as to ensure even coverage of the days of the week. The 'tomorrow' self-completion diary is left with the respondent by the interviewer with the instruction to make records as frequently as possible through the diary period so as to maximize the recall quality. Tomorrow instruments have normally covered either one or two days, but experience in The Netherlands and the UK shows that periods up to a full week are possible, although with reduced response rates.

A variety of interview technologies have been employed for yesterday diaries; in addition to conventional pen-and-paper personal interviews, there are examples of computer-aided personal interviews (CAPI), and computer-aided telephone interview (CATI) methods. Studies sometimes use a combi-

nation of yesterday and tomorrow with the self-completion instrument being posted back by the respondent. In this case the tomorrow shows lower and somewhat biased responses. For tomorrow diaries, response is substantially improved if the interviewer returns to collect the diary.

A second substantial design issue concerns the temporal framework within which the activities are recorded. A key decision here is whether to adopt an 'event' basis of recording ('When did you wake up?'; 'What did you do first?'; 'When did you stop?'; 'What did you do next?'). This style is appropriate for a 'yesterday' interview, but not for a tomorrow diary. For self-recording it is more appropriate to use a schedule with regular time slots, of 5–15 min (although slots of up to 30 min have been used, and the length of the slots may vary through the day and night); respondents indicate the duration of each activity on the schedule using lines, arrows or ditto marks.

Activities may be reported using either a 'closed' vocabulary, in which a predetermined set of activity codes is provided to cover the full range of possibilities, or an 'open' approach, in which respondents use their own words to describe the activities, these being coded subsequently. Respondents frequently find the 'closed' approach constraining, however well-designed the categories, and the choice of categories may to a limited degree influence the level of reporting (e.g., a more detailed classification of travel types leads to a small increase in estimates of travel time). However, the open approach has symmetrical shortcomings: respondents report at varying levels of detail, which are difficult to disentangle from genuine interpersonal differences.

For much of the day it is difficult to classify the current state as a single activity: 'writing at the computer' may be accompanied by 'drinking coffee' or 'talking to my child.' The normal practice is to ask respondents to record activities in an hierarchical manner: 'What are/were you doing?' 'Were you doing anything else at the same time?' Some studies ask respondents to record all activities in a nonhierarchical manner; but respondents tend to prefer opportunity to use a narrative mode ('First I did this, and then I did that …'), which provides a hierarchy of primary and secondary activities. The hierarchy also simplifies the process of analysis, although there are ways of analyzing the data that are nonhierarchical but still conserve the 1440 min of the day. Other ancillary information included within the rows or columns of diary forms includes items such as 'Who else is present/engaged in the activity?' and 'Where are you?'—the latter sometimes also requesting information on modes of transport. National income accountants' interest in nonmonetized economic activity has led to attempts to acquire 'Who for?' information, but these have not so far been successful. Specialized diary studies also include interpretative or affective information, (e.g., 'Is this work?' or 'How much did you enjoy this?') attached to each event or activity.

Beyond the design of the diary instrument, there are also various specialized sampling issues. Seasonal effects may be of importance; representative samples of time use require an even coverage across the year. There are serious practical difficulties but most national-scale samples attempt multiple sweeps during different parts of the year; the most developed approach to temporal sample design is found in Poland, whose 1970s and 1980s studies achieved approximately even numbers of diaries for each day of the year. Samples that include simultaneous diaries for all household members greatly extend the analytical usefulness of the evidence. Conventional 'tomorrow' studies include children down to the ages of 12 or even 10 years; and Italian national studies have, by a combination of simplified forms for younger children, and parental proxies for preliterate children, achieved total household coverage.

5. *Time-diary Research*

Table 1 indicates the global extent of diary research. Large-scale national random sample studies have been conducted in each of three or more decades in 20

Table 1
National time-diary studies

Number of studies	Countries
100+	Russia/Soviet Union
20+	France, Japan, UK*, USA
6–19	Australia, Belgium, Brazil, Bulgaria*, Canada, Denmark, Germany*, Hungary*, India, Italy*, Netherlands, Poland*, South Korea, Spain*, Sweden*
1–5	Albania*, Austria, Bangladesh, Botswana, Cote d'Ivoire, Czech Republic, Estonia*, Finland, Gambia, Greece*, Ireland*, Israel, Java, Kenya, Latvia*, Lithuania*, Luxembourg*, Mexico, Macedonia*, Malaysia, Nepal, New Zealand, Nicaragua, Norway, Pakistan, Papua New Guinea, Peru, Philippines, Portugal*, Romania*, Rwanda, Saudi Arabia, Slovak Republic*, Slovenia*, South Africa, Switzerland, Tanzania, Thailand, Turkey, Uganda, Upper Volta, Yugoslavia, Zimbabwe

countries. The first cross-national comparative time-diary study was supported by UNESCO in 1965 (Szalai 1972). Twenty countries (marked with asterisks in Table 1) have participated in Eurostat's 1996–2001 European Harmonized Time Use Study project. Results from the Multinational Time Use Study, providing evidence of change in time use over the period 1961–95 from 20 countries, are reported by Gershuny (2000).

See also: Circadian Rhythms; External Validity; Field Observational Research in Anthropology and Sociology; Questionnaires: Cognitive Approaches; Sample Surveys: Methods; Sample Surveys: Survey Design Issues and Strategies; Time, Anthropology of; Time Perception Models; Unobtrusive Measures

Bibliography

Becker G 1965 A theory of the allocation of time. *Economic Journal* **80**: 493–517
Bourdieu P 1985 *Distinction*. RKP, London
Gershuny J 2000 *Changing Times*. Oxford University Press, Oxford, UK
Harvey A S 1999 Guidelines. In: Pentland W E, Harvey A S, Lawton M P, McColl M A (eds.) *Time Use Research in the Social Sciences*. Kluwer, New York, pp. 19–45
Robinson J P, Godbey G 1997 *Time for Life*. Pennsylvania State University Press, Philadelphia, PA
Szalai A (ed.) 1972 *The Use of Time*. Mouton, The Hague

J. Gershuny

Timing: Psychological and Neural Aspects

Humans and other animals are able to make reasonably accurate estimates of intervals of time, and to adjust their behavior on the basis of these estimates. The first section of this article describes some facts and principles of the perception of time and the ability of humans and other animals to time their responses; the second section describes some psychological explanations of time perception and timed responding; and the final section describes some explanations of these phenomena in terms of neural mechanisms.

The study of time perception was already well developed when William James wrote his textbook, *The Principles of Psychology*. In a chapter entitled, 'The perception of time,' James (1890) provided definitions of subjective concepts such as the duration of the perceived present, and reviewed the extensive psychophysical data on time perception available primarily from European laboratories. With the psychophysical method a person makes a psychological judgment based upon the characteristics of the time interval between two physical stimuli. James describes the use of this method to identify such values as the intervals between two stimuli that are the most accurately perceived, the shortest interval between two stimuli that could be perceived to be two distinct stimuli, and the longest interval between two stimuli that can be perceived as a unit. By the middle of the twentieth century, most research on timing was based on psychophysical research with human participants (Woodrow 1951).

The perception of time was not a main focus of studies of animal learning in the first half of the twentieth century. The first laboratory study of the determinants of the behavior of rats was conducted by Small (1900), and rats became the most frequently used animal in psychological research. In his *Handbook of Psychological Research on the Rat*, Munn (1950) did not describe any experiment that was explicitly about timing.

But many of the experiments on learning involved variations in time intervals between stimuli, responses, and reinforcements that affected behavior. The results of such experiments have been important in the development of the current understanding of time perception.

1. Timing and Timed Performance

Classical conditioning (also called Pavlovian conditioning) is a procedure in which a reinforcement, such as food, is delivered contingent upon the time of occurrence of a previous stimulus or reinforcement.

The simplest example is one in which food is given at a regular interval and the animal develops an anticipatory response to the time of arrival of the next food. This phenomenon, known as temporal conditioning, was demonstrated in experiments on salivary conditioning of dogs described by Pavlov (1927). If food were given to a dog regularly at 30-min intervals, the dog would begin to salivate shortly before the time of occurrence of the next food delivery. This was objectively determined by counting the number of drops of saliva in each 30-s interval. The importance of such observations was that it demonstrated that time since the occurrence of food, or some biological effects that were correlated with that time, could serve as a conditioned stimulus.

A more typical example of classical conditioning is one in which food is given at a regular interval following the onset of a stimulus, and there is a long irregular interval between successive food deliveries. This also results in an anticipatory response to the time of arrival of the next food, but the time since the onset of the stimulus (rather than the previous food) is the relevant interval. This phenomenon, known as delay conditioning, was also demonstrated in the experiments of Pavlov (1927). It provided additional evidence that time from any discriminable event could serve as a conditioned stimulus.

The temporal control of behavior in classical conditioning has been demonstrated with many different species, stimuli, responses, and reinforcers over a wide range of time intervals. It occurs in a wide range of classical conditioning procedures, including those in which the time intervals are not constant. Variations in the time intervals between stimuli and reinforcers produce corresponding variations in the time of conditioned responses.

Instrumental conditioning (also called operant conditioning) is a procedure in which a reinforcement, such as food, is delivered contingent upon a response, although it may also be contingent upon the time of occurrence of a previous stimulus or reinforcement.

A simple example is one in which food is made available at a regular interval after the last food delivery, and it is delivered following the next response. The animal develops an anticipatory response to the time of availability of the next food. This procedure, known as a fixed interval schedule of reinforcement, was described by Skinner (1938). If food is made available to a rat in a lever box one minute after the previous food delivery, the rat will be increasingly likely to press the lever as the time of food availability approaches (as signaled by the last food). Again, the importance of such observations was that it demonstrated that time since the occurrence of food, or some biological effects that were correlated with that time, could serve as a conditioned stimulus.

There are many other examples of timing in operant conditioning. If food is made available a fixed time after the onset of a stimulus and delivered following the first lever response after it is available, the rat will be increasingly likely to press the lever as the time of food availability approaches (as signaled by the last stimulus onset). Many other operant conditioning procedures with rats, pigeons, and other animals also demonstrate temporal control of behavior. These include the ability of animals to delay responding if only long interresponse intervals are reinforced, the ability to hold down a lever for an interval that corresponds approximately to the duration necessary to secure reinforcement, and many others.

Two procedures that have been used explicitly to study the timing ability of animals are the peak procedure and the bisection procedure. Versions of both of these tasks have been devised for human participants, and the results have supported the conclusion that the same principles of timing apply to humans and other animals (Church 1993).

In the peak procedure, reinforcement is delivered following the first response after a fixed interval of time after the onset of a stimulus on some occasions; on others there is no reinforcement and the stimulus continues for a long interval. Response rate increases as a function of time from stimulus onset until approximately the time that food is sometimes delivered, and then it decreases. The location of the peak of the response rate function provides a measure of the expected time of reinforcement, and the spread of the function provides a measure of the certainty. The peak procedure has been regarded as a temporal production procedure in which responses occur at the time of expected reinforcement.

In the bisection procedure, a stimulus is presented for a short time (such as 2 s) or a long time (such as 8 s), and then two responses are made available (such as a left and right lever). Reinforcement is delivered if the rat presses the left lever following the 2-s stimulus or if it presses the right lever following the 8-s stimulus. Then five or more stimuli of intermediate duration are presented, and responses following these durations are not reinforced. The psychophysical function relating the probability of a right response following each of the stimuli is typically an increasing function with an ogival shape. The duration that correspondence to indifference between the two alternative responses provides a measure of the psychological middle. It is usually approximately at the geometric mean of the extreme durations. The bisection procedure has been regarded as a temporal perception procedure in which the animal reports whether a presented interval was short or long.

Estimates of the absolute duration of short intervals are more accurate than of long intervals. But estimates of the relative duration of short and long intervals are about equally accurate. For example, the error in judgment of a 10-s interval and a 60-s interval is approximately the same if the error is calculated as a proportion of the interval. This is an example of Weber's law for timing (Gibbon 1977). In peak procedure, the bisection procedure, and many other procedures, both the mean and the standard deviation increase linearly with the temporal interval. Thus the ratio of the standard deviation to the mean, the coefficient of variation, remains constant. A more general principle has been called 'superposition.' This refers to the fact that the entire functions relating the behavioral measure to time in the peak procedure, the bisection procedure, and many other procedures are very similar if time, and the behavioral measure, are scaled in relative units (Gibbon 1991).

The ability of animals to time intervals in the range of seconds to minutes is now well established, and the quantitative effect of variations in the conditions on performance in different timing tasks is now well established. These results have provided a guide for the development of psychological explanations about the mechanism of timing. For the evaluation of quantitative psychological models of timing, it would be useful if archives of raw data from timing experiments were available. Such resources are becoming increasingly feasible to establish.

2. *Psychological Explanations*

A psychological process model of timing and time perception necessarily involves intervening variables.

These intervening variables can generally be classified as being related to perception, memory, and decision processes. A formal model of the process consists of rules connecting the input to one or more of the intervening variables, one or more of the intervening variables to the output, and, perhaps, the intervening variables to each other. In a quantitative model of the process, the rules are specified mathematically.

The input to a psychological process model of timing is a physical duration of time that begins with some event, such as the onset or termination of a stimulus, response or reinforcement. This event initiates some internal or behavioral process that changes in a regular way with time, a process that is often referred to as an 'internal clock.' Many types of internal representations of physical time have been proposed. One proposal is that a pacemaker emits pulses according to some distribution, the event that begins the timing process closes a switch that permits the pulses to enter an accumulator, and the psychological representation of time is the accumulation of pulses. This leads to an approximately linear relationship between psychological and physical time. Another proposal is that the event begins a continuous, nonlinear process, such as an exponentially decaying memory trace, and the magnitude of the trace provides a nonlinear relationship between psychological and physical time. Other functions to convert physical to psychological time have also been proposed.

The internal clock may consist of multiple processes that are initiated by an event. The multiple processes may be a several periodic or nonperiodic functions with different characteristics that are initiated at the same time, or several processes that are initiated serially. The representation of time with such clocks is usually a vector of heights of each of the functions. There has been no consensus regarding the nature of the internal clock. Any of them can account for the behavioral data qualitatively, and none of them can be ruled out as neurally implausible.

A psychological process model of timing also requires some memory of previously reinforced intervals. For example, the animal must remember that it was reinforced 20 s after the onset of a stimulus, or it must remember that a left lever response was followed by food after a 2-s stimulus. One proposal is that temporal memory consists of an unorganized collection of examples; another is that it consists of a continuum of strengths that correspond with particular times after an event. In addition to a structure, a psychological model of timing requires a rule for storage of new information and a rule for retrieval of information. There has been no consensus regarding the nature of temporal memory based upon either the behavioral or neural data.

Finally, a psychological process model of timing requires some decision about whether or not to respond, or, if multiple responses are available, which response to make. The response decision is based upon a comparison of perception of the current time with the memory of the reinforced time. If the perceived current time is near enough to the remembered time of reinforcement, a response will occur.

Many quantitative models of timing have been proposed that can be described in terms of the three processes of temporal perception, memory, and decision (Church and Kirkpatrick 2000). Some of them are mechanistic, but others consider the animal to be a computational device that operates on symbolic representations (Gallistel 1990). The models differ in their quantitative representation of each of the parts. Scalar timing theory (Gibbon et al. 1984) is based upon a perceptual representation of time that has a pacemaker that emits pulses that are switched into an accumulator, a memory representation of time that is based on an unorganized collection of examples, and a decision that is based on a ratio comparison of the perceptual representation with the a single sample from temporal memory. In contrast, the Behavioral theory of timing and related models (Killeen and Fetterman 1988, Machado 1997) are based upon a perceptual representation of time that is a cascade of functions, a memory representation of time that is based on the strength of each of the functions at the time of reinforcement, and a decision that is based on the product of the perceptual and memory representation. The various quantitative models of timing have been found to produce reasonable fits to the data in different timing procedures, but none of them make accurate quantitative predictions in all procedures. The development of a single quantitative model of timing that makes accurate quantitative predictions for a large number of behavioral measures in a wide range of procedures is essential. Such a model is likely to contain modules for temporal perception, memory, and decision processes that included parts of current theories.

3. *Neural Explanations*

The perception of time is quite different from the perception of physical stimuli such as lights, sounds, and chemicals. In the case of vision, audition, olfaction, taste, pain, and other sensory systems, there are physical stimuli that impinge upon receptors specialized for particular physical stimuli. But time, like location, is an attribute of any stimulus and there are no known specialized receptors. A perceived time refers to the time since the onset of a physical stimulus, the interval between two physical stimuli, or the duration of a physical stimulus. Thus, it is possible that the psychological explanations that propose general mechanisms for timing are incorrect. Instead, there are specialized mechanisms for the perception of the duration of visual stimuli, auditory stimuli, and others, based upon the unique properties of each modality. One argument in favor of some general

processes is based upon the success of the cross-modal transfer experiment. Rats trained to make a temporal discrimination in one modality (such as a 2-s noise vs. an 8-s noise) can be tested on a temporal discrimination in another modality (such as a 2-s light vs. an 8-s light). The test can either use the same response assignments for the short and long stimuli or use a reversed assignment. During testing the rats learned significantly more rapidly with the same assignment of responses to the short and long stimuli than with reversed assignments (Roberts 1982). This demonstrates that there is something general about a 2-s and an 8-s stimulus in different modalities.

Although it is clear that the brain is the biological mechanism responsible for interval timing, the specific neural processes involved are not well understood. To obtain a general understanding of the neurotransmitter mechanisms and their sites of action, the primary methods have been to study the effects of brain lesions and systemic drug administration on behavior in timing procedures. The goal is to obtain some degree of neural and behavioral selectivity. Neural selectivity refers to the specificity of the brain structure or receptors affected, and behavioral selectivity refers to the specificity of the observed effects. With the development of timing procedures with well-established quantitative functional relationships between independent and dependent variables, and quantitative psychological models of the timing process that consist of three or more separate modules, it has become possible to search for biological manipulations that affect one module but not others. On the basis of systemic injection of dopamine agonists and antagonists, the speed of the internal clock appears to be controlled by activity of D^2 dopamine receptors. Similar research with systemic injection of cholinergic agonists and antagonists produced results that suggested that the memory storage appears to be controlled by the level of activity of the cholinergic system (Meck 1996). Other research with neurotoxin lesions has identified the clock effects with structures in the basal ganglia, and the temporal memory and attentional effects with structures related to the frontal cortex.

See also: Time Perception Models

Bibliography

Church R M 1993 Human models of animal behavior. *Psychological Science* **4**: 170–3

Church R M, Kirkpatrick K 2000 Theories of conditioning and timing. In: Mowrer R R, Klein S B (eds.) *Handbook of Contemporary Learning Theories*. Erlbaum, Hillsdale, NJ, pp. 211–53

James W 1890 *Principles of Psychology*. MacMillan, London, Vol. 1

Gallistel C R 1990 *The Organization of Learning*. MIT Press, Cambridge, MA

Gibbon J 1991 Origins of scalar timing. *Learning and Motivation* **22**: 3–38

Gibbon J 1977 Scalar expectancy theory and Weber's law in animal timing. *Psychological Review* **84**: 279–325

Gibbon J, Church R M, Meck W H 1984 Scalar timing in memory. In: Gibbon J, Allan L (eds.) Timing and time perception. *Annals of the New York Academy of Science* **423**: 52–77

Killeen P R, Fetterman J G 1988 A behavioral theory of timing. *Psychological Review* **95**: 274–95

Machado A 1997 Learning the temporal dynamics of behavior. *Psychological Review* **104**: 241–65

Meck W H 1996 Neuropharmacology of timing and time perception. *Cognitive Brain Research* **3**: 227–42

Munn N 1950 *Handbook of Psychological Research on the Rat*. Houghton Mifflin, Boston

Pavlov I 1927 *Conditioned Reflexes*. Oxford University Press, Oxford, UK

Roberts S 1982 Cross-modal use of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes* **8**: 2–22

Skinner B F 1938 *The Behavior of Organisims*. Appleton-Century-Crofts, New York

Small W S 1900 An experimental study of the mental processes of the rat. *American Journal of Psychology* **11**: 133–65

Woodrow H 1951 Time perception. In: Stevens S S (ed.) *Handbook Of Experimental Psychology*. Wiley, New York, pp. 1224–36

R. M. Church

Tip-of-the-tongue, Psychology of

The 'tip-of-the-tongue' (hereafter, TOT) phenomenon is an everyday, near-universal experience characterized by the inability to retrieve a word known to the speaker. Researchers have long been interested in this phenomenon, both in terms of understanding why it occurs and when it is more likely to occur, and because TOT provides a window into memory retrieval. In this article the main findings stemming from these different lines of research are summarized.

At a youth conference in London in 1998, the British Deputy Prime Minister John Prescott, when asked to justify the increasing costs of the Millenium Dome in Greenwich, said in front of thousands of attendees, 'The money came from the…you know…what do they call it…' and finally blurted out in desperation 'the raffles.' In response to laughter from the audience, Prescott tried to show he knew something about the word he could not produce: 'I don't do it myself.' At this point the conference chairperson whispered discreetly 'the lottery,' but it was too late to cover up the retrieval failure which was reported in the next day's *Times* (*Times*, January, 1998, from Schacter 2001).

Like John Prescott, seemingly all speakers occasionally experience the temporary inability of retrieving a word: a TOT. In a survey including 51 languages,

Schwartz (1999) reports not only that all informants across languages knew what a TOT was, but also that all the languages considered had a specific expression for it. Furthermore, in 45 of these languages, this expression includes the word 'tongue,' which may be related to our feeling of immanent retrieval of the word during a TOT.

This feeling was well described by William James (1890, pp. 251–252):

> Suppose we try to recall a forgotten name. The state of our consciousness is peculiar. There is a gap therein: but no mere gap. It is a gap that is intensely active. A sort of wraith of the name is in it, beckoning us in a given direction, making us at moments tingle with the sense of closeness, and then letting us sink back without the longed-for term. If wrong names are proposed to us, this singularly define gap acts immediately so as to negate them. They do not fit into its mold.

This description captures some of the basic facts about TOT. The longed-for word is known and on the verge of coming out. Furthermore, the word is not completely unavailable: its 'mold' is there.

1. *Methodologies Used to Investigate TOT*

Psychological studies investigating TOTs are based on introspection. Using either experimental methods to induce TOTs or naturalistic observations, researchers have capitalized on speakers' ability to identify their TOTs and their ability to reflect on their cognitive state during TOTs.

1.1 *Experimental Induction of TOTs*

Experimental studies started in 1966, with the work of Brown and McNeill (1966). They presented speakers with definitions of rare words such as 'a place where bees are raised for their honey' and then asked them to provide the name. When speakers could not say the name, they were first asked to indicate whether or not they were in a TOT state (unable to say the word but sure they knew it, and feeling that the word was on the verge of coming back). If in a TOT state, they were asked to provide information about the unavailable word (other words that came to mind, the initial letter of the word, etc.). Finally, speakers were provided with the target name (e.g., *apiary*) and were asked whether the word they were attempting to retrieve was the same as the experimenter's target. Cases in which speakers did not recognize the experimenter's target as their target were defined as *negative TOT states* and used as a baseline for guessing, which could be compared with the information reported for *positive TOT states* (i.e., cases in which the experimenter's target matched the speaker's target). This experimental methodology (and its variants) to induce TOT has since been used in a number of studies to investigate, for example, lexical retrieval during speaking.

1.2 *Diary Studies*

A second methodology that has been used is the recording of TOTs as they spontaneously occur during daily activities (see *Speech Errors, Psychology of*). For example, in a study by Burke et al. (1991), speakers of different age groups were asked both general questions about their retrieval failures (e.g., to estimate how often they experience TOTs), as well as specific questions concerning TOTs as they occurred over a four-week period. These specific questions (as for experimentally induced TOTs) included reporting partial information, but also strategies used to resolve TOT states. Such a methodology has provided important information regarding the incidence of TOTs for different verbal materials and for different populations (i.e., young and elderly speakers).

2. *Incidence of TOTs*

TOT incidence is influenced by a number of factors; some of the most important are reviewed below (see Brown 1991 for a comprehensive review).

2.1 *Speaker-related Factors*

In their diary study, Burke et al. (1991) found that TOT states are more common for older (1.65 TOTs per week) than younger (0.98 TOTs per week) speakers. Other studies have also found a higher incidence for older speakers (see also *Memory and Aging, Cognitive Psychology of*)

Dramatic differences in incidence rates are also related to brain damage. Anomia is defined as a language specific disturbance arising after brain damage whose main symptom is the inability of retrieving known words. In some anomic patients, these retrieval failures have the same features of TOTs occurring for nonlanguage-impaired speakers, but they are far more common. For example, patient MS (Vigliocco et al. 1999), failed to retrieve the target name for a picture 32 percent of the time (for the same pictures, 10 nonlanguage-impaired speakers performed 100 percent correct). On these occasions, MS could still provide partial information about the name, indicating that he was indeed in a TOT state.

2.2 *Word-related Factors*

Certain words are more susceptible to TOTs than others. In their naturalistic study Burke et al. (1991) found that 62 percent of the reported TOTs involved

proper names, 12 percent involved object names, and 23 percent involved abstract words. In her catalog of TOTs, Browman (cited in Brown 1991) found that 50 percent of the words inducing TOTs were proper names, 25 percent were concrete and abstract nouns, 14 percent were adjectives and 9 percent were verbs. The predominance of retrieval failures for proper names is also well established in other studies (see Brown 1991 and Schacter 2001). Why? One reason may be that proper names, relative to common names, are not so well integrated with related concepts: they denote individuals but they do not provide any further information about those individuals. This possibility is supported by experiments that suggest a higher incidence of TOTs for proper names which are nondescriptive (e.g., John Baker), than for proper names that are descriptive of some attribute of a character (e.g., John Baker who is a baker) (Schacter 2001). Another possible reason is that proper names cannot be replaced easily by another word, while this is often an option for common names. For instance, if I want to introduce Flavio Monti, I have no option but to say his name; if instead I want to talk about a sofa, if I am experiencing retrieval difficulties, I can use a synonym (couch).

Burke et al. (1991) further reported that the higher incidence of TOT for proper names was magnified by age differences. That is, although there was an overall increase in TOTs from younger to older adults, such an increase was particularly dramatic for proper names. Burke et al. (1991) discuss this result in the framework of a model of the TOT phenomenon in which the difference between proper names and common names is thought of as a difference in the convergence of activation from conceptual to lexical representations. For proper names, a single conceptual representation (denoting the individual) would activate the corresponding lexical representation; for common names multiple conceptual features would converge on a lexical representation. In this framework, the difference between younger and older adults is further accounted for by assuming an age-related weakening of the conceptual to lexical connections.

Differences in frequency and recency also contribute to words' differential susceptibility to TOTs. Low frequency words tend to induce TOTs more often than high frequency words. Rastle and Burke (1996) investigated whether differences in recency of use of a word may account for the different incidence of TOTs for proper names and within this class, for the different performance of younger and older adults. In their study, recency was manipulated by priming target words in an experiment inducing TOTs. They found that while recent use of a word, indeed, decreased the incidence of TOTs, it did not differentially affect proper names and common names and did not differentially affect younger and older adults, suggesting that recency has a global effect on the retrieval process.

3. *Information Available for Report during a TOT State*

Numerous studies have reported that speakers in a TOT state are able to report partial information about the word, which supports the interpretation of TOT as a memory retrieval failure rather than a memory encoding or storage problem (see also *Memory Retrieval*). In experimental studies, speakers' guesses when in a TOT state are compared to baseline guesses. Two types of baselines are usually used; (a) the speaker's rating of being in a 'don't know' (as opposed to a TOT) state; (b) negative (as defined above) TOTs. It has been shown that speakers in TOT states can report both lexicosyntactic and phonological information about the words above baseline.

Regarding syntactic information, researchers have shown that Italian speakers can report above chance whether a noun is masculine or feminine (i.e., a syntactic property that does not have conceptual connotates for nouns referring to objects and abstract entities); English language-impaired (e.g., patient MS) and unimpaired speakers can report whether a noun is countable (i.e., pluralizable, such as *opinion*) or mass (i.e., nonpluralizable, such as *knowledge*) (see Vigliocco et al. 1999) when such information is not semantically transparent. Regarding phonological information, many studies from the seminal work by Brown and McNeill (1966) have shown that speakers can report above chance information such as the number of syllables, the main stress, and the first phoneme or letter in the word.

Often, in both diary and experimental studies, words semantically and phonologically similar to the target are reported by the speakers. This fact has led some researchers to propose that TOTs are blockages caused by the intrusion of these unwanted words; however, this hypothesis has not received much experimental support (Meyer and Bock 1992).

4. *A Psycholinguistic Perspective on TOT*

TOT has been used as a tool to test hypotheses regarding lexical retrieval in production. Theories of speech production (see *Speech Production, Psychology of*) assume that lexical retrieval is achieved in two main steps. During a first step an abstract lexical representation that specifies the meaning and the syntactic properties of the intended word is retrieved; during a second step the phonological make-up of the word is retrieved. Within this framework, TOT states can be accounted for in terms of an inability to retrieve the phonological word form, despite successful retrieval of the abstract lexical representation. The fact that speakers in a TOT state can report information about the syntactic properties of a word is compatible with the specification of this information at the level of abstract representations; while the availability of

phonological information may be accounted for in terms of partial activation of word-forms.

5. Metacognition, Feeling of Knowing, and TOT

TOT is related to the 'feeling of knowing' (FOK), which is the ability to predict the likelihood of later recognition of a word presently unavailable. In both cases, speakers 'know that they know,' that is, they have knowledge about their memory state (see also *Metamemory, Psychology of*). Although related, note that the two differ because while TOTs occur spontaneously, FOK judgments are usually elicited. Furthermore, speakers in TOT states are confident they will eventually recall the word; FOK, instead, assesses recognition likelihood for a word. In TOT and FOK there is a discrepancy between the subjective conviction of knowing the word and the actual inability to produce it. This discrepancy has been taken by some authors to suggest a distinction between the ability to monitor the content of memory and the ability to retrieve from memory. Alternatively, it has been suggested that FOK and TOT arise as a consequence of the amount of accessible information pertaining to the unavailable word (see Koriat 1993 for an overview).

See also: Encoding Specificity in Memory; Incidental versus Intentional Memory; Memory for Meaning and Surface Memory; Memory Problems and Rehabilitation

Bibliography

Brown A S 1991 A review of the tip-of-the-tongue experience. *Psychological Bulletin* **109**: 204–23

Brown R, McNeill D 1966 The 'tip of the tongue' phenomenon. *Journal of Verbal Learning and Verbal Behavior* **5**: 325–37

Burke D M, McKay D G, Worthley J S, Wade E 1991 On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language* **30**: 542–79

James W 1890 *The Principles of Psychology*. Holt, New York, Vol. 1

Koriat A 1993 How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review* **100**: 609–39

Meyer A S, Bock K 1992 The tip-of-the-tongue phenomenon: Blocking or partial activation? *Memory-and-Cognition* **20**: 715–26

Rastle K G, Burke D M 1996 Priming the tip-of-the-tongue: Effects of prior processing on word retrieval in young and older adults. *Journal of Memory and Language* **35**: 586–605

Schacter D L 2001 *The Seven Sins of Memory. How the Mind Forgets and Remembers*. Houghton-Mifflin, Boston

Schwartz B L 1999 Sparkling at the end of the tongue: The etiology of tip-of-the-tongue phenomenology. *Psychonomic Bulletin and Review* **6**: 379–93

Vigliocco G, Vinson D P, Martin R C, Garrett M F 1999 Is 'count' and 'mass' information available when the noun is not? An investigation of tip of the tongue states and anomia. *Journal of Memory and Language* **40**: 534–58

G. Vigliocco

Tocqueville, Alexis de (1805–59)

1. The Unity of a Process

Born into a family of very old Norman nobility on July 29, 1805 in Paris, Tocqueville first studied philosophy and rhetoric at the Collège Royal in Metz, then law in Paris. In 1827, he was appointed to the judiciary and became a probationary judge at Versailles. Throughout the Restoration, he took part in the bitter struggle which set the old aristocracy, enclosed in its traditions, against the new France, led by the middle classes. After the Revolution of 1830, and having sworn—not without hesitation—the oath of allegiance to Louis-Philippe, he gave himself plenty of room as far as events were concerned by requesting a commission to conduct a survey into the American penitentiary system. This trip (April 1831 – February 1832), undertaken with his friend Gustave de Beaumont, resulted first in a report entitled *Du Système pénitentiaire aux Etats-Unis et de son application en France (On the penitentiary system in the United States and its application in France)* (1833), then the first two volumes of *De la Démocratie en Amérique (Democracy in America)* (1835), and finally what is known as the *Second Démocratie (Second Democracy)* (1840). The success of the first part of this work was enough to establish his reputation; in 1838 he was elected to the Academy of Moral and Political Sciences and, three years later, he entered the Académie française.

Tocqueville's liking for philosophical reflection did not cancel out his sense of political action. Elected to the Chamber of Deputies in 1839, and reelected several times in his Valognes constituency, he committed himself resolutely to the parliamentary path on the side of the constitutional opposition. On January 27, in a speech with a prophetic note, he announced the revolution which was to break out one month later. After the February Days, he was witness to the clashes of June 23 to 26, 1848 which put 'the haves' at odds with the 'have-nots.' His *Souvenirs (Recollections)* (1951–, Vol. XII) relate what this 'great battle waged in Paris' was like. A short-lived Minister for Foreign Affairs under the Prince President of the Second Republic, Tocqueville was returned to his studies for good with the coup d'État of December 2, 1851. *L'Ancien Régime et la Révolution (The Ancien Regime*

and the Revolution) appeared in 1856, the second great work which its author was anticipating expanding when tuberculosis cut short his life. He died on April 16, 1859 in Cannes.

Tocqueville's work—which comprises the two great texts of 1835/40 and 1856, the *Recollections*, various papers, speeches, and reports, as well as a vast correspondence—draws its unity from the constant questioning which inspired it: why since 1789, in 1830, and then in 1848, was it always the *same* revolution which starts up again? The unity of the work also stems from the method adopted in order to understand the repetition—in Europe, and in France in particular—of the revolutionary process. Its comparative method, not systematized, but rather applied spontaneously, allowed it to avoid the perils that Eurocentricity and sociocentricity are denounced for today. 'Whoever has only studied and seen France will understand nothing of the French Revolution.' (1951–, Vol. II, 1, p. 94) It was therefore without being obsessed with one singular case, and being free of all aristocratic prejudice, that Alexis de Tocqueville set about penetrating the essence of democratic society; by combining 'holism' and 'individualism,' he posited a diagnosis along with a prognosis on modern society whose value and validity were only recognized much later.

2. *In Search of Democracy*

'It is not only in order to satisfy curiosity—however legitimate—that I examined America,' Tocqueville wrote 'I wanted to find any lessons there from which we might profit […] I confess that in America I saw more than just America; I sought an image of democracy itself, its tendencies, its character, its prejudices, its passions. I wanted to become acquainted with it, if only at least to know what we should hope or fear from it.' (1951–, Vol. I, 1, p. 11) With this definition, the Tocquevillian project was set out as an effort to understand a 'social state' and, secondarily, a political regime. If the first readers of the 1835 work were attracted to the evocation of exotic locations, they were rather more invited to turn the American mirror upon themselves. Three years later, they were to see a different image of democracy which was less suited to charming them.

Tocqueville's confessed search was not, in fact, separable from the recognition of a tendency—considered from the outset irresistible and irreversible—which was specific to modern societies: the march towards the equalization of conditions. Behind the analyses which were put forward, the contrast between aristocratic society and democratic society operated in a virtually permanent state, and it was the transition from the traditional order to the new arrangements in human relations that Tocqueville undertook to study. He approached this by going back to the principles which govern social states and to the institutions which arise out of them. From an early stage, he realized that one principle had to dominate and that the government which was called mixed, that is, which united an aristocratic element with a democratic one, was merely an idle fantasy. The English monarchy, with its strongly dominant aristocratic character, was still able to pass for the model of government; but it was the American system which was called upon to become the model for modern nations. Equally, Tocqueville also developed an 'ideal type' for democratic society, based on the equality of conditions and set against an aristocratic society founded on the hierarchy of status.

If the paths leading to democracy are marked out in different ways, 'the great advantage of Americans is to have arrived [there] without having to undergo democratic revolutions, and to have been born equal rather than having to become it.' While examining the customs—defined as 'the moral and intellectual state of a people'—of Americans, Tocqueville assigned their foundation to a fully accepted equality of conditions. (1951–, Vol. 1, p. 300) In the USA, democracy was in some way 'natural'; its inhabitants had not had to struggle against privileged groups, such as the ones which had resisted in France; they made up an immense middle class within which there was active mobility. Whatever was set down in custom was fostered by institutions—the commune, the jury, and numerous associations. Decentralization and communal democracy allowed citizens to take part in public affairs, and allowed the body social to regulate itself. 'Each individual makes up an equal portion of the ruling body, and plays an equal part in the government of the State.' The close bond between the political and the social was actually realized in this society, a society which acted upon itself and through itself, since power only existed within it. In this way, the negative effects of individualism were limited. As for the ills which could arise out of equality, liberty provided the cure. In this society where one endeavored to bring people together rather than to set them against each other, and where the most serious social antagonisms were excluded (one found rich people, not a rich class), religion made a powerful contribution to social regulation. Tocqueville clearly highlighted the role of religious beliefs in the bolstering of customs and the upholding of Republican institutions. Religion prevented the American citizen 'from imagining everything and forbids him from daring to do everything.' It ordered him to confine himself to the representation of that which was real and to the conception of that which was useful. Through this, the American mentality took shape—pragmatic, not speculative; with a firm grasp of the facts, not prey to ideology.

The general characteristics of the American model are set out clearly in *Democracy in America II*, along with its potential variations such as liberal democracy and revolutionary or democratic despotism; it also

includes the amendment to the relationship between the passion for liberty and the passion for equality. Individualism—the childhood illness of democracy—is defined in this work as 'a deliberate and peaceful feeling which inclines each citizen to isolate himself from the mass of his fellow men and to withdraw from his family and friends. In this way, after having created for himself a small-scale society for his own benefit, he readily leaves wider society to itself.' (1951–, Vol. I, 2, p. 105) In *The Ancien Regime and the Revolution* one again finds the threat this represents illustrated in the case of France, and the dangers signaled in 1840 which were brought up to date by the comparative analysis of the passions of democracy and of revolution. Another model, this time continental, was in place at the time and was in a way the negative of the American model.

3. *The Analysis of Social Change*

'What strikes the European who travels through the United States,' Tocqueville noted in *Democracy in America I*, 'is the absence of what we call government or administration.' (1951–, Vol. I, 1, p. 69) In 1840, he further drew attention to his initial observation: 'Wherever, at the head of a new venture, you see the government in France and a nobleman in England, in America you can be sure to find an association.' (1951–, I, 2, p. 113) If it was true that a good government was one which taught its citizens to manage without it, this was not working in France. The backdrop to the 1856 work, which links the interpretation of an historical crisis with a theory of social change, was provided by feudalism and its networks of dependence and solidarity where masters and servants were united by reciprocal services. It was this system, social and political, with its intermediary groups, corporations, village communities, provincial assemblies, and so on, that the French monarchy had set about destroying by means of administrative centralization. No one benefited from it; neither the nobleman whose preeminence was affected, nor the bourgeois whose vanity was wounded, nor the ordinary man who still remained the underling. Discontent was universal.

By suppressing the intermediary bodies, by condemning the nobility—stripped of its function and devitalized—to appearing merely as the ornament of despotism, while at the same time encouraging the emergence of new elites but without satisfying their statutory ambitions, royalty had converted an organic whole where each individual was bound to a group into an homogenous mass oscillating between apathy and demand, submission and subversion. Such was the new social state: 'men, no longer being bound to each other by any bond of caste, class, corporation, or family, are only too inclined to concern themselves with their own individual interests and too prone to considering only themselves and to withdrawing into a state of narrow individualism where every political virtue is stifled.' Thus, centralization, the restriction of liberties, and the deterioration of public life combine forces. In this way, equality leads to despotism, since it is no longer built on liberty or on participation in communal affairs.

The explanatory variations which Tocqueville identifies therefore relate, as in the 1835/40 work, to characteristic customs, institutions, and circumstances. In the process which they deal with, the Revolution no longer appears as a rupture; the Terror had only prolonged the process of state centralization which had been initiated earlier. There was, in reality, a continuity between royal absolutism, revolutionary Jacobinism, and imperial politics. In the process in question, religion had also played an important role by virtue of the universalist dimension presented by Catholicism. At the same time, the Revolution—itself preaching the abolition of privilege and the equalization of conditions, had 'itself become a kind of new religion, an imperfect religion.' (1951–, Vol. II, 1, p. 89) As a result of this, Tocqueville was in no doubt that 'all the Revolution did would have been done without it; it was only a violent and hasty process with whose help the political state was adapted to the social state, the deeds to the ideas, and the laws to the customs.' (1951–, Vol. II, 1, p. 66)

One should take note of the manner in which Tocqueville analyzes the social conflicts and their origins and exacerbation. They were born in eighteenth century France in a short period of economic recession following a long phase of general prosperity. It was therefore not increased poverty which triggered off the conflicts, nor was it the weight of former coercive institutions which were the cause. The Revolution 'did not break out in the regions where these institutions were better preserved and their discomfort and rigour felt most acutely by the people, but on the contrary in those regions where they were felt the least, in such a way that their yoke had seemed all the more unbearable precisely where it was the least heavy.' (1951–, Vol. II, 1, p. 99) It is a question here of situations that one analyzes today in terms of relative frustrations. Tocqueville recognized them perfectly in the case of the emancipation of the peasantry, just as he had in the case of the ascendancy of the bourgeoisie. They made two strongly felt passions explosive: one was 'the violent and inextinguishable hatred of inequality'; the other 'more recent and less deep-rooted' led individuals 'to want to live not only equally, but also freely.'

These were the passions which the *philosophes* of the Enlightenment had stirred up. They provided a theoretical framework for social protest by showing the necessity of replacing traditional habits with new rules. Far removed from the affairs of state, with no practical experience and therefore inclined to conceive of 'general and abstract theories in government matters,'

men of letters undertook to found cloud cuckoo land. Vain debates of ideas took the place of public life: 'Political life was violently driven back into literature, and writers, taking the direction of public opinion in hand, found themselves from one moment to the next holding those positions which the leaders of political parties usually occupied in free countries.' (1951–, Vol. II, 1, p. 196) With this observation, Tocqueville paved the way to a sociology of intellectuals.

4. *The Originality of the Contribution*

The author of *Democracy in America* was hailed by his contemporaries as a new Montesquieu. The historians of the nineteenth century saw in *The Ancien Regime and the Revolution* a major contribution to historiography. If the glory of Tocqueville knew no eclipse in the USA, his work did in France, however, fall into a kind of oblivion from which it did not emerge until after World War II. It is to Raymond Aron (1967) and André Jardin (1984)—who actively worked on the publication of the *Oeuvres Complètes (Complete Works)*—as well as to Jean-Claude Lamberti (1970, 1983)—the author of the first state-commissioned thesis on Tocqueville—that one principally owes the renewal of interest in a work which from then on took its place among the classic works of sociology. There are many reasons for this long purgatory. Tocqueville was not a maker of systems as one so likes in France; he appeared as the champion of liberalism at a time when this current of thought was being strongly denounced; nor did his interpretation of the French Revolution seem ideologically sound; and then his way of thinking seemed to come under a mixed genre of headings—political philosophy, philosophy of history, and general sociology.

The reasons for Tocqueville's return are by no means separable from the turnaround in the intellectual climate which took effect in the last two decades of the twentieth century. The new tendencies are well known—the decline of hypersocialized conceptions of the individual; the freeing of collective memory previously marked by the rupture of revolution; and the loosening in particular of the vice of monocausal determinism. On the latter, the *Recollections* give this opinion which is consistent with the orientation of the social sciences at the beginning of the twenty-first century: 'I, for my part, hate these absolute systems, which make all the events of history dependent on great first causes, each one being linked to the other by a chain of inevitability, and which, as it were, remove men from the history of mankind.' (1951–, Vol. XII, p. 84) The social protagonists are restored with a power of deliberation and decision, the choice of strategies, and a certain amount of room for maneuver through methodological individualism. They have 'good reasons,' as Raymond Boudon (1977) definitively emphasized, to act as they do, and it was these reasons which, in concrete situations, Tocqueville attempted to bring to light.

This writer was original and modern first and foremost through his method of implementation; it brings together observations, comparisons in space and time, and generalizations; it retains the significant fact without losing sight of the point of view of the collective whole; and the investigation which it guides is situated both on the macrosociological as well as the microsociological level. The rational reconstruction of the significance, in the end ascribed to human actions and historical events, is carried out from objective and subjective elements. The protagonists' reasons are integrated into the explanatory schema along with their passions. Tocqueville, in fact, invites his reader to weigh up the importance of the passion factors; he urges him thus to 'imagine' through a sympathetic understanding the fate of the French peasant in the eighteenth century: 'Imagine this man's condition and calculate if you can the wealth of hatred and envy which has amassed in his heart.' (1951–, Vol. II, 1, p. 106)

Tocqueville's contribution to social science is also original in the way that it poses problems and formulates questions. Why, it is asked in the *Mémoire sur le pauperisme (On pauperism)* (1835), are the countries which appear to be the most poverty-stricken the ones which, in reality, have the lowest number of inhabitants? And why, inversely, is a section of the population of the opulent nations forced to have recourse to the help of other people in order to live? A comparison between Portugal—where poverty was hardly to be seen because the responsibility was assumed by the networks of traditional solidarity—and England—whose social regulations were broken by the Industrial Revolution which created its wealth—provided the answer. It is therefore not simply a question of observing and describing—*Democracy in America* is not a simple monograph resulting from banal field research—but of understanding the social processes arising from 'good questions.'

Tocqueville's way of thinking was modern, in particular because of the subject it was applied to—that is, democracy. A type of entirely new society was establishing itself in the nineteenth century. Emphasizing its principal traits—notably the reign of public opinion—and the dangers it comprised—in particular the tyranny of the majority—Tocqueville assigned its foundation to a dependent work force and the equality of conditions. The development of the egalitarian mentality which accompanies the establishment of democratic society is a source of dissatisfaction that the hope of mobility is not up to appeasing. Inequalities, reduced in one sphere, spring up again elsewhere—there will never be an end to them. Individualism which causes social completeness to disintegrate brings about a displacement of the boundary which separates public space from the

private sphere; the latter extends like a zone of retreat where one concern rules: that of material well-being. At the same time, the author of *Democracy in America* 'saw,' in a splendid passage where prophecy and prediction merge the 'type of despotism that democratic nations have to fear.' This was a vision of a 'vast mass of people all alike and all equal,' seeking trifling pleasures, all separated from each other, and above whom 'rises up an immense guardian power, absolute, intricate, regulated, provident, and gentle,' (1951–, Vol. I, 2, p. 324) which reduces men to the level of children by making them dependent—this tableau really is the Welfare State.

See also: Democracy; Democracy, History of; Enlightenment; French Revolution, The; Public Opinion: Political Aspects; Revolutions, History of; Revolutions, Theories of; Revolutions, Sociology of; Social Movements, History of: General; Social Change: Types

Bibliography

Aron R 1967 *Les Etapes de la pensée sociologique*. Gallimard, Paris

Birnbaum P 1970 *Sociologie de Tocqueville*. Presses universaires de France, Paris

Boudon R 1977 *Effets pervers et ordre social*. Presses universaires de France, Paris

Coenen-Huther J 1997 *Tocqueville*. Presses universaires de France, Paris

Drescher S 1968 *Dilemmas of Democracy—Tocqueville and Modernization*. University of Pittsburgh Press, Pittsburgh, PA

Jardin A 1984 *Alexis de Tocqueville 1805–1859*. Hachette, Paris

Lamberti J-C 1970 *La Notion d'individualisme chez Tocqueville*. Presses universaires de France, Paris

Lamberti J-C 1983 *Tocqueville et les deux démocraties*. Presses universaires de France, Paris

Lively J 1962 *The Social and Political Thought of Alexis de Tocqueville*. Clarendon Press, Oxford

Manent P 1982 *Tocqueville et la nature de la démocratie*. Julliard, Paris

Melonio F 1993 *Tocqueville et les Français*. Aubier, Paris

Nolla E (ed.) 1990 *De la Démocratie en Amérique Alexis de Tocqueville*. 2 vols. Vrin, Paris

Pierson G W 1969 *Tocqueville in America*. Peter Smith, Gloucester, UK (originally published as Tocqueville and Beaumont in America. Oxford University Press, New York, 1938)

Schleifer J T 1980 *The Making of Tocqueville's Democracy in America*. University of North Carolina Press, Chapel Hill, NC

Tocqueville A de 1951– *Oeuvres complètes* (OC). 22 vols. to date. Gallimard, Paris (edition still in progress)

Tocqueville A de 1986 *Bouquins*. Robert Laffont, Paris (a collection of the main works with studies by J-C Lamberti, F Mélonio, J Schleifer)

Tocqueville A de 1991 *Oeuvres, Bibliothèque de la pléiade Vol. 1, L'Ancien Régime et la Révolution*. Gallimard, Paris

Tocqueville A de 1992 *Oeuvres, Bibliothèque de la pléiade Vol. 2, De la Démocratie en Amérique*. Gallimard, Paris

Valade B 1985 Tocqueville le contemporain. In: *Universalia 1985*. Encyclopaedia Universalis, Paris

B. Valade

Tolerance

By 'tolerance,' most contemporary social scientists mean citizens' willingness to put up with others' efforts to participate in public life even if they personally dislike those others' views. In other words, 'One is tolerant to the extent one is prepared to extend freedoms to those whose ideas one rejects, whatever these might be' (Sullivan et al. 1979, p. 784). Tolerance among the mass public has been assumed to be a crucial psychological underpinning of democracy, a form of restraint that makes it possible for a wide range of groups and people to express their ideas and participate in public life (Almond and Verba 1963). It is an orientation toward political and social life that presumes agreement on certain basic procedural norms by which differences of opinion are negotiated. In the USA, this is usually interpreted as support for the basic civil rights of those with whom one disagrees, for example, their right to free speech, free assembly, due process, privacy, freedom to seek political office, and other similar liberties.

Tolerance suggests that despite (or perhaps because of) the diversity of opinion on matters of public interest, there needs to be general agreement on the 'rules of the game' (Neubauer 1967), the procedural norms by which substantive matters are negotiated, and these rules must then be applied equitably to all groups, regardless of how counternormative their views might be. The notion of political tolerance evolved from the idea of religious toleration (Sullivan et al. 1982); the former was a way to live with one's ideological enemies just as the latter was a way to live among those of different religions.

The body of social science research pertaining to political tolerance is closely related to research on support for civil liberties (see *Civil Liberties*) and studies of intergroup conflict, but the concept has evolved in such a way that it now carries a more specific meaning: In order to 'tolerate' another person or group, one must disagree with the views that the person or group represents or espouses (Sullivan et al. 1982). Thus a communist who supports the civil rights of other communists is not demonstrating tolerance because the element of disapproval and disagreement is missing. Only when the same level of support is extended to those with whom one disagrees can the label 'tolerance' be properly applied.

The many definitions of tolerance that have been offered and the corresponding measurement strategies

they have spawned are part of a lengthy progression in social scientific interest on this topic dating back to the McCarthy era in the USA. As Finkel et al. note (1999, p. 204), the key questions and controversies motivating research on tolerance to date have included:

To what extent are citizens willing to extend procedural democratic liberties (such as freedom of speech and association) to unpopular political groups? How widespread is political intolerance?

What are the cultural, demographic, social, psychological, economic, and political sources of political tolerance and intolerance? Are certain types of citizens (e.g., the less educated) more intolerant?

Is the mass public becoming more tolerant than it once was? If so, why?

Does intolerance tend to be concentrated on a few unpopular groups, or is it more widely dispersed? In other words, to what extent is intolerance 'pluralistic?'

Are elites more tolerant than the rank and file? If so, why? And is the elite-mass tolerance gap widening or narrowing?

Are there significant cross-national differences in tolerance, whether in the general public or among elites? How can such differences be explained?

What individual-level behaviors and system-level outcomes does tolerance or intolerance foreshadow?

In this article, I begin by providing a brief description of the seminal, early studies of political tolerance in the USA. Although an overview of all of the areas of subsequent research described above is beyond the scope of this article (see Sullivan and Transue 1999, and Finkel et al. 1999 for two recent and far more extensive literature reviews), this discussion highlights four of these important themes including (a) differences in levels of tolerance held by the mass public as opposed to more politically active segments of the population; (b) debates over the conceptualization and measurement of tolerance; (c) evidence pertaining to whether tolerance has increased in the USA; and (d) cross-national examinations of political tolerance. These four topics serve as a loose organizational scheme for the article, but because these debates are closely interconnected, these topics cannot be neatly divided into four separate discussions.

1. *Early Research on Tolerance*

In what is widely recognized as the first systematic national study of tolerance in the USA, Samuel Stouffer gathered information on the opinions of a large probability sample of Americans in 1954. They were asked a series of questions about support for the civil liberties of a variety of unpopular left-wing groups including socialists, atheists and Communists. His findings, published as *Communism*, *Conformity*, *and Civil Liberties* (1955) suggested that the majority of Americans did not think an admitted communist should be allowed to make speeches in public, to retain American citizenship, to teach in high schools or colleges, or even to work as a clerk in a store. Moreover, the majority felt that books by communists should be taken out of public libraries, that the government should have the authority to tap their personal phone conversations to acquire evidence against them, and that they should be thrown in jail. Although these same responses suggested somewhat greater tolerance when asked with reference to socialists, atheists, or 'suspected communists,' the reaction this study generated was still one of great surprise. Either the USA was not truly a democracy, or the assumption that widespread public commitment to these procedural norms was necessary to democracy was incorrect. The latter interpretation predominated, and subsequent studies seized upon two findings that seemed to explain how democracy could thrive without widespread political tolerance among members of the mass public.

1.1 *'Carriers of the Creed'*

Stouffer's work, along with two important studies published in the early 1960s (Prothro and Grigg 1960, McClosky 1964) illuminated patterns of findings that continue to be central in debates over tolerance today. One explanation that was offered for how democracy persisted without mass tolerance centered on mass-elite differences. Stouffer's original study included a sample of local community leaders, and this group proved to be much more supportive of the rights of left-wing groups than the mass sample. Likewise, in his study of political party activists in the late 1950s, McClosky (1964) found greater support for civil liberties among political activists. Consistent with this pattern, Prothro and Grigg (1960) found greater tolerance among the well-educated and well-to-do. Thus the dominant conclusion that was drawn was that the well-educated and politically active segments of the population were the 'carriers of the creed' who protected the democratic system from an intolerant majority (Prothro and Grigg 1960, p. 281). If the intolerant were inactive, they could do little damage; moreover, their apathy supposedly prevented them from acting on their political views.

1.2 *Principles versus Application*

A second prominent explanation for democratic stability in the face of apparently low levels of support for procedural norms emerged from Prothro and Grigg's (1960) study which asked respondents both about general principles of democracy and about some specific applications of those principles. They found widespread support for the principles but greatly diminished support when questions were posed so as

to apply those principles to specific controversial cases. So, they argued, to say that Americans did not support democratic principles was technically incorrect; they did so in theory but often did not apply those principles when faced with specific applications. Consistent with the elite model of democratic stability, McClosky's study (1964) further confirmed that political activists were more supportive of the rules of democracy both in principle and when they were applied to specific cases.

The general conclusion drawn from these early studies was that democracy could be sustained even with mass intolerance. This was possible because the intolerant were concentrated among the politically inactive segments of the population. The relatively tolerant attitudes of political elites were assumed to be due to their educational and socioeconomic backgrounds and a result of their exposure to democratic norms through actual involvement in the political process (Finkel et al. 1999, p. 208).

That the well-educated and politically-engaged tend to be more tolerant than ordinary citizens has now been widely documented over the years. Nonetheless, the underlying cause of mass-elite differences has not been fully resolved. Moreover, if one argues that tolerance demands support for freedom of expression and assembly, due process and privacy *even for those groups that promote intolerance*, then the thesis of democratic elitism itself can be questioned. As Sniderman et al. (1996) point out, although there is a consensus among elites in support of free expression and assembly for most unpopular groups, when racists are the target group, there is no elite consensus in defense of their rights.

2. *Changing Conceptualizations and Trends in Tolerance Over Time*

In the 1970s, Nunn et al. (1978) updated Stouffer's original work and found Americans far more supportive of the civil liberties of those same groups that had been the target of Stouffer's assessments 20 years earlier. Nunn et al. argued that although mass democracy in America may have been on shaky ground in the 1950s, improvements in education as well as other demographic changes had made it far more tenable by the 1970s. Perhaps political leaders were no longer so essential to sustaining democratic norms.

This optimistic conclusion was soon called into question. Sullivan et al. (1979) countered this assessment by pointing out that the target groups used by Stouffer as well as Nunn and his colleagues to assess levels of tolerance had been almost exclusively groups on the left. If tolerance is defined as putting up with those with whom one disagrees, then support for the rights of groups on the left alone should not be considered an adequate measure. For those who are on the left themselves, such questions do not require tolerance. Sullivan et al. proposed an alternative measurement technique that involved first asking individual respondents which groups they liked the least, and then asking them the standard items addressing support for civil liberties with respect to those specific groups. This two-step technique insured that respondents did, in fact, have negative feelings toward the groups they were asked about.

Based on results using this technique, Sullivan et al. argued that the American public was far less tolerant than Nunn et al. had indicated. Instead, they suggested that Americans were more tolerant only of groups on the left; other unpopular groups, primarily right-wing or racist groups, still elicited high levels of intolerance when people were asked about their rights to demonstrate, run for office, give public speeches, and so forth. Although there were no comparable data from earlier points in time using this same two-step technique, Sullivan et al. suggested that the improvements in levels of tolerance noted by Nunn et al. were 'illusory' at best. On the more optimistic side, however, they argued that intolerance that was focused on a multitude of targets—so called 'pluralistic into lerance'—was safer for democracy because large groups of people did not gang up on particular groups as targets. A multitude of targets was deemed preferable to intolerance focused on one or a very small number of groups.

2.1 *Critiques of the Least Liked Group Technique*

Following Sullivan et al.'s efforts at reconceptualizing tolerance, this two-step, 'content-free' technique has become increasingly popular. However, both the technique and the conclusions that have been based on it have not been accepted without question. In response to Sullivan et al.'s claim about the lack of an trend toward greater tolerance over time, both Abramson (1980) and Immerwahr (1980) pointed out that it is impossible to tell without comparable data from the 1950s whether the increases noted by Nunn et al. were truly 'illusory.' Mueller (1988) analyzed trends using a comparable content-controlled approach, and concluded that there had been a marked increase in tolerance since the early studies, but that this had occurred mainly because Americans had grown less concerned about threats from leftist groups as the communist threat dissipated, and their concern was not redirected toward similarly threatening rightist groups. Although incontrovertible evidence on the shape of the trend over time does not exist, it appears that tolerance in some forms has increased.

A second controversial aspect of the Sullivan approach is its effect-driven view of tolerance (see Finkel et al. 1999). Respondents are asked to name their most disliked group, and this becomes the basis for tolerance judgments, with the assumption that this

should serve as a critical test because it will be most difficult to extend these liberties to one's most disliked group. But surprisingly, intolerance of people's second most disliked group actually turns out to be higher than for the most disliked group (Sullivan et al. 1982), thus raising questions about whether negative effect is the best criterion for group selection.

In addition, as Sniderman et al. (1989) and Gibson (1989) point out, if Sullivan's conclusions about safety in the proliferation of groups that could serve as possible targets for intolerance were correct, then the levels of tolerance that individuals express should vary a lot based on which groups they are asked about, and conservatives and liberals should pick different groups, with liberals more intolerant of conservatives and conservatives more intolerant of liberals. But in their empirical analyses, Sniderman et al. found that neither the group nor the act made much difference to people's responses. And while conservatives are less tolerant than liberals, this did not much affect their choice of target groups; both liberals and conservatives were most likely to choose militarists and racists. Gibson (1989) concurred with findings showing that tolerance of communists and Ku Klux Klan members were highly correlated. Findings in this vein have led to the conclusion that tolerance is adhered to as a general principle rather than as a group-specific construct (see also Duckitt and Farre 1994):

> The person who does not honor and protect the rights of those whose point of view clashes with his own is … a bad bet to protect the rights even of those whose point of view supports his own … It is the racial bigot, not the person committed to racial tolerance, who is more likely to oppose free speech for racists (Sniderman et al. 1991, p. 135).

Although the shortcomings of the Stouffer-type approach to measuring tolerance are widely acknowledged, currently there is no consensus on an ideal measurement technique, and thus researchers continue to use variants of the Sullivan technique (for example, Barnum and Sullivan 1989) as well as Stouffer-like measures that are balanced across the ideological spectrum. Although these techniques result in different distributions of the proportion of the population that is tolerant and intolerant, Gibson's (1992) study suggests that the same set of independent variables predicts both kinds of tolerance measures equally well, thus the structure of relationships remains similar regardless of measurement technique.

2.2 *Further Conceptual Dilemmas*

The concept of tolerance may be religious in origin, but over the past century it has been explored primarily in terms of the tolerance of groups holding particular political viewpoints. Although the term racial tolerance was in general use in the latter half of the twentieth century, the term is now seldom used, at least in part due to the reconceptualization of the term by Sullivan et al.: If tolerance presumes dislike of the target group, then racial tolerance is not what most advocates of racial equality have in mind. Rather, the goal is a lack of discrimination on the basis of race and ethnicity so that these groups would not be disliked any more than any other and thus not need to be 'tolerated.' Instead of racial tolerance, terms such as prejudice, discrimination, or, more generally, racial attitudes are used in discussions of attitudes toward racial and ethnic minorities (see *Prejudice in Society* and *Discrimination, Racial*). Nonetheless, the distinction between attitudes toward racial as opposed to political groups have become increasingly important to understandings of tolerance in contemporary politics.

For example, in the many studies using the least liked group method, much of the intolerance observed in contemporary samples is what might be described as *intolerance of intolerance*. In other words, it involves people's objections to the exercising of these basic procedural rights by groups who themselves promote intolerance. The most commonly mentioned target groups of this kind are racists or Ku Klux Klan members.

Does tolerance require that one tolerate the intolerant? While there has been general agreement that citizens should be free to organize their lives around their own beliefs and values even if they are at odds with the majority, there has been no consensus as to whether people whose beliefs and values are at odds with the majority by virtue of being intolerant should be allowed the same rights of speech, assembly, and so forth (Sniderman et al. 1996). The current popularity of anti-hate legislation in the USA testifies to this important tension in current definitions of what tolerance means. Does tolerance demand that citizens support measures to restrict the rights of intolerant groups, or does it require that they support their rights despite their promotion of intolerant views?

Although this is an old quandary, it brings to the fore the often oversimplified perspective of research on tolerance where the 'correct,' politically enlightened answer was typically assumed to be straightforward and obvious (Sniderman et al. 1996). Discrepancies between belief in the principle of tolerance and willingness to support its application to certain cases typically have been ascribed to citizens' lack of ability to understand the connection between principle and practice. Instead, this discrepancy appears to be due to citizens' efforts to balance support for tolerance with other values, such as the desire to discourage negative racial attitudes.

3. *Cross-national Research*

These same kinds of core democratic values also have been studied in contexts outside of the USA, including established democracies such as Canada, Israel, and

New Zealand, and newly-emerging democracies such as post-World War Two Europe, the former USSR countries, and Latin America in the 1980s. Some of the measures used in these efforts are identical to or adapted from those already described, while others use entirely new measurement strategies tapping a far broader range of attitudes toward democratic norms and values than tolerance *per se*.

Of particular interest to understanding the origins of mass tolerance in democratic systems are studies of the attitudes of citizens in countries without established democracies. Kaase's (1971) study of postwar West Germans found that by 1968 their beliefs were similar to those of people in the USA in their high levels of support for principles involving free speech and assembly and the necessity of political opposition. Also like US citizens, levels of support were much lower when these principles conflicted with public order or other values. Kaase's study, along with others in Spain, Italy, and Japan, suggested that changes in attitudes toward democratic values can occur after a regime change (see, for example, Gunther et al. 1986).

To explain processes of 'democratic legitimation' after a regime change, Weil (1989) developed an influential model suggesting that the longer people live within a responsive, competitive, and democratic system, the more they will develop support for democratic values. Thus, in studies of democratization during the post-Cold War era, what was particularly surprising was the high levels of support for democratic principles found immediately after the demise of communism (see, for example, Klingemann and Hofferbert 1994 on the former DDR (East Germany); Seligson and Booth 1993 on Nicaragua and Costa Rica; Finifter and Mickiewicz 1992, Gibson et al. 1992, Reisinger et al. 1994, Rose and Mishler 1994, and Tedin 1994 on the former USSR and Eastern Europe). Since little time had elapsed for the gradual emergence of democratic values, and citizens had had little experience with a responsive political system, this rush to embrace democratic norms was unexpected.

Post-hoc explanations for high initial support for democratic values have led to caution in interpreting these prodemocracy, protolerance attitudes as indicative of a wholehearted embrace of democratic values. One interpretation suggests that what researchers are observing are sheer novelty effects based on citizens' initial attraction to and excitement about democracy. A second explanation for high initial levels of support suggests that through 'demonstration effects' from observing the West, democratic political values have actually already been exported to these countries (Dalton 1994, Gibson et al. 1992, Weil 1996).

However, it is worth noting that responses to questions asking about commitment to, or overall evaluations of, democracy are less supportive than those asking about specific values associated with democratic government (e.g., Whitefield and Evans 1996) and there is considerable variation from question to question, thus suggesting ambivalence at best in countries formerly within the USSR (see Dalton 1994, Miller et al. 1994).

One of the ongoing measurement problems in studies utilizing broader attitudes toward democratic government is the extent to which these measures are tapping attitudes toward current system performance as opposed to support for democratic values (see Finkel et al. 1999). Moreover, it has been difficult for researchers to discern the stability of these beliefs over time, especially in rapidly changing political circumstances. Still, these cross-national studies have tremendously improved the study of political tolerance and support for democratic norms, not only by providing contexts in which there is greater variance in levels of tolerance than in the USA, but also by providing insight into the process by which tolerant attitudes develop in the mass public.

Until relatively recently, claims about the causes and correlates of tolerant attitudes have been heavily dependent on American data, thus making generalizations about tolerance quite difficult. Levels of tolerance clearly vary by country. For example, Gibson et al.'s studies in the (now former) USSR suggested that even countries without strong democratic traditions may have high levels of support for democratic norms in the abstract, yet their levels of tolerance when applied to specific cases were extremely low, much more so than in the USA and other countries where democracy has had a longer time to take root. Sullivan et al.'s (1985) studies in Israel and New Zealand using the same least-liked group method also found substantial aggregate differences in levels of tolerance with New Zealanders much more tolerant than Americans and Israelis.

But aggregate differences aside, cross-national research suggests some degree of generalizability of findings from the USA. For example, Sullivan et al. found that although the *level* of threat perceived varied by country, its impact on tolerance judgments was roughly the same, and the same was true for the influence of personality characteristics. In their studies of tolerance among the Soviets, Gibson et al. (1992) likewise found that perceived threat was the strongest predictor of tolerance, and personality characteristics such as dogmatism were also influential in promoting intolerance. Commitment to abstract democratic norms decreased intolerance as it did in the USA and New Zealand. However, Sullivan et al. found that the impact of belief in democratic norms was quite different in Israel, and reflected cross-national differences in understandings of what constitutes democracy. Thus although the goal of tolerance may be the same in all places—finding a peaceful and noncoercive way to deal with inevitable political conflict—cross-national research on tolerance must take into account the fact that not all countries will define democracy to include the same norms for adjudicating its differences of political opinion.

See also: Authoritarian Personality: History of the Concept; Civil Liberties and Human Rights; Civil Rights; Civil Society, Concept and History of; Democratic Theory; Discrimination; Ethnocentrism; Freedom: Political; Human Rights: Political Aspects; Pluralism and Toleration; Political Culture; Prejudice in Society; Wisdom: Philosophical Aspects; Wisdom, Psychology of.

Bibliography

Abramson P R 1980 Comment on Sullivan, Piereson, and Marcus. *American Political Science Review* **74**: 780–81

Almond G, Verba S 1963 *The Civic Culture*. Princeton University Press, Princeton, NJ

Barnum D G, Sullivan J L 1989 Attitudinal tolerance and political freedom in Britain. *British Journal of Political Science* **19**: 136–46

Dalton R 1994 Communists and democrats: Democratic attitudes in the two Germanies. *British Journal of Political Science* **24**: 469–493

Duckitt J, Farre B 1994 Right-wing authoritarianism and political intolerance among whites in the future majority-rule South Africa. *Journal of Social Psychology* **134**: 735–41

Finifter A, Mickiewicz E 1992 Redefining the political system of the USSR: Mass support for political change. *American Political Science Review* **86**: 857–74

Finkel S E, Sigelman L, Humphries S 1999 Democratic values and political tolerance. In: Robinson J P, Shaver P R, Wrightsman L S (eds.) *Measures of Political Attitudes*. Academic Press, San Diego, CA, pp. 203–96

Gibson J L 1989 The structure of attitudinal tolerance in the USA. *British Journal of Political Science* **19**: 562–70

Gibson J L 1992 Must tolerance be 'least-liked'? Alternative measures of political tolerance. *American Journal of Political Science* **36**: 560–77

Gibson J L, Duch R, Tedin K 1992 Democratic values and the transformation of the Soviet Union. *Journal of Politics* **54**: 329–71

Gunther R, Sani G, Shabad G 1986 *Spain after Franco: The Making of a Competitive Party System*. University of California Press, Berkeley, CA

Immerwahr J 1980 Comment on Sullivan, Piereson, and Marcus. *American Political Science Review* **74**: 781–83

Kaase M 1971 Demokratische Einstellung in der Bundesrepublik Deutschland. In: Wildenman R (ed.) *Sozialwissenschaftliches Jahrbuch für Politik Vol 2*. Günther Olzog Verlag, Munich, Germany, pp. 119–326

Klingemann H D, Hofferbert R 1994 Germany: A new 'wall' in the mind? *Journal of Democracy* **5**: 30–44

McClosky H 1964 Consensus and ideology in American politics. *American Political Science Review* **58**: 361–82

McClosky H, Brill A 1983 *Dimensions of Tolerance: What Americans Believe about Civil Liberties*. Russell Sage, New York

Miller A, Hesli V, Reisinger W 1994 Reassessing mass support for political and economic change in the former USSR. *American Political Science Review* **88**: 399–411

Mueller J 1988 Trends in political tolerance. *Public Opinion Quarterly* **59**: 1–40

Neubauer, D E 1967 Some conditions of democracy. *American Political Science Review* **61**: 1002–9

Nunn C Z, Crockett H J Jr, Williams J A Jr 1978 *Tolerance for Nonconformity*. Jossey-Bass, San Francisco

Prothro J W, Grigg C M 1960 Fundamental principles of democracy: Bases of agreement and disagreement. *Journal of Politics* **22**: 276–94

Reisinger W, Miller A, Hesli V, Maher K 1994 Political values in Russia, Ukraine and Lithuania: Sources and implications for democracy. *British Journal of Political Science* **24**: 183–223

Rose R, Mishler W 1994 Mass reaction to regime change in Eastern Europe: Polarization or leaders and laggards? *British Journal of Political Science* **24**: 159–82

Seligson M, Booth J 1993 Political culture and regime type: Evidence from Nicaragua and Costa Rica. *Journal of Politics* **55**: 777–92

Sniderman P M, Fletcher J F, Russell P H, Tetlock P E 1989 Political culture and the culture of double standards: Mass and elite attitudes toward language rights and the Canadian Charter of Rights and Freedoms. *Canadian Journal of Political Science* **22**: 259–84

Sniderman P M, Fletcher J F, Russell P H, Tetlock P E 1996 *The Clash of Rights*. Yale University Press, New Haven, CT

Sniderman P M, Fletcher J F, Russell P H, Tetlock P E, Gaines B 1991 The fallacy of democratic elitism: Elite competition and commitment to civil liberties. *British Journal of Political Science* **21**: 349–70

Stouffer S A 1955 *Communism, Conformity and Civil Liberties*. Doubleday, Garden City, NY

Sullivan J L, Piereson J E, Marcus G E 1979 An alternative conceptualization of political tolerance: Illusory increases, 1950s–1970s. *American Political Science Review* **75**: 92–106

Sullivan J L, Piereson J E, Marcus G E 1982 *Political Tolerance and American Democracy*. University of Chicago Press, Chicago

Sullivan J L, Shamir M, Walsh P, Roberts N S 1985 *Political Tolerance in Context: Support for Unpopular Minorities in Israel, New Zealand and the USA*. Westview Press, Boulder

Sullivan, J L, Transue J E 1999 The psychological underpinnings of democracy: A selective review of research on political tolerance, interpersonal trust and social capital. *Annual Review of Psychology* **50**: 625–50

Tedin K 1994 Popular support for competitive elections in the Soviet Union. *Comparative Political Studies* **27**: 241–71

Weil F D 1989 The sources and structure of legitimation in Western democracies: A consolidated model tested with some time-series data in six countries since World War II. *American Sociological Review* **54**: 682–706

Weil F D 1996 Western political culture and the consolidation of democracy in eastern Germany. Paper presented at the Eighth International Conference on Socioeconomics, Geneva

Whitefield S, Evans G 1996 Support for democracy and political opposition in Russia, 1993–1995. *Post-Soviet Affairs* **12**: 218–42

D. C. Mutz

Topographic Maps in the Brain

Topographic maps in the brain are organized collections of neurons, within subcortical nuclei or in cortical areas that isomorphically represent or map the

receptor sheet or muscles of the body. Topographic sensory representations reflect the arrangements of receptors on the receptor sheet, while topographic motor maps reflect the arrangement of the muscles. Maps might be of the mechanoreceptors of the skin or muscles, the photoreceptors of the retina, or the row of hair cells of the cochlea in the inner ear. In a topographic map of the receptors of the skin, touching a sequence of positions on the skin activates a comparable sequence of groups of neurons in the map. The order of representation reflects the order of activating afferents terminating in the map, although the map is rendered more precise by local circuits within the map that tend to inhibit less active and sparse inputs. Thus, neurons have smaller receptive fields, and there is less overlap of receptive fields for adjacent groups of neurons. All vertebrates have topographic maps in the brain. Many of these maps are in the brain stem. Mammals have a specialized part of the forebrain, the neocortex, which varies greatly in size and differentiation across taxanomic groups. All mammals have a few homologous topographic maps in neocortex, but mammals in which neocortex is greatly enlarged have additional maps that are not found in all mammals. Topographic maps appear to be fundamental to the processing of sensory information and motor control.

1. *The Discovery of Sensory and Motor Maps in the Brain*

One of the early advances in the field of neuroscience was the realization that the brain has functionally specialized subdivisions and does not act as a whole or single organ. Instead, the brain consists of a number of interacting and interconnected subdivisions, or organs of the brain, usually termed nuclei for subcortical or brain stem structures and areas or fields for neocortex. Some of the most convincing evidence for the view that functions are localized in specific parts of the brain came from early electrophysiological studies that demonstrated that neocortex has topographic maps of receptors and muscles.

The first signs of the existence of maps in the brain came from careful clinical observations of the impairments of patients with cortical lesions or focal epilepsy. Most notably, in the 1860s the neurologist, John Hughlings Jackson, reasoned that only the existence of orderly maps in the brain would explain the experiences of patients with epilepsy where a sensation or movement would start at some location on the body and then expand systematically to include successively more distant body parts. Compelling experimental evidence for such maps soon came from investigators that exposed the neocortex of dogs and monkeys, and directly stimulated the brain through small electrodes applied to the cortical surface. Remarkably, they found that small, brief currents often evoked movements of some part of the body, especially when the electrodes were placed in the frontal lobe of the brain. While movements could be evoked from many locations in the brain, it soon became obvious that movements could be evoked with the smallest currents in a strip of cortex in the back of the frontal lobe, and that the type of movement depended on the location of the electrode in this strip. In monkeys and in humans, movements of the foot, trunk, hand, and face could be evoked in sequence as the electrode was moved across the strip. Similar organizations were identified in the most posterior portion of frontal cortex in a range of mammalian species, and this cortex became known as primary motor cortex, or M1.

A second systematic representation of body movements was soon identified in cortex along the middle edge of the cerebral hemisphere just ahead of M1, first in human patients undergoing brain surgery and later in other mammals. This motor representation was termed the supplementary motor area, SMA, and it is also sometimes called the second motor area, M2. We now know that monkeys and other primates also have a number of additional motor representations.

The topographic organizations of motor maps were soon revealed in more detail when thin, penetrating microwires or microelectrodes were used instead of surface electrodes. The exposed tip of the microelectrodes could be placed in cortical layer 5, the layer projecting directly to motor neurons in the brain stem and spinal cord, and thus only a small number of the most relevant neurons could be activated. Because the corticospinal projections of the neurons in M1 on the right side of the brain predominantly cross in the lower brain stem to terminate on or near motor neurons projecting to muscles in the left side of the body, neurons in M1 mainly control contralateral body movements. The topographic maps in M1 and other motor areas are largely of contralateral body movements.

Generating small currents to stimulate neurons proved to be technically much easier than recording the small currents generated by neurons as they are activated by sensory stimuli or other neurons. Thus, topographic maps of sensory surfaces were demonstrated later than the first motor maps, initially in the 1930s when the oscilloscope for measuring small currents and good amplifiers came into general use. Clinton Woolsey was in the forefront of such recording, and he soon demonstrated maps of the skin, cochlea, and retina in the cortex of rabbits, cats, and monkeys. A strip of cortex just behind M1 was responsive to light touch and taps on the body surface, and as for M1, the sequence of representation across this strip of cortex was from foot, trunk, hand, and face. This first discovered somatosensory representation became known as S1, or primary somatosensory cortex. A second representation was soon discovered on the lateral margin of S1, and this representation became known as S2. The second representation is

activated by the face along its border with S1, while the foot and hand (forepaw) are more lateral. Presently, there is evidence that most mammals have two to three additional somatosensory representations, and monkeys and most likely humans have several more.

The same recording techniques allowed the discovery of topographic representations of the cochlea and retina. In initial studies, different parts of the cochlea were electrically stimulated while recording electrodes were variously placed in auditory cortex. Later brief tones of different frequencies were used selectively to activate receptors along the long, narrow, coiled row of receptors in the cochlea. Both methods revealed a primary auditory area, A1, where the cochlea is represented from base (low tones) to apex (high tones) in a sequential manner. Inputs from both ears activate auditory cortex, but the stronger input is from the contralateral cochlea. Soon, a second auditory area, A2, was proposed, but this second area, originally described in cats, was later found to include parts of other auditory areas. A redefined A2 was described with a crude representation of tones that parallels A1 in representation of tones. Possibly because of early confusion over the nature of A2, the term is not commonly applied to an auditory area except in cats. Gradually, more topographic maps of the cochlea have been described in cats and monkeys, and monkeys may have as many as 10 areas with at least a crude representation of tones.

The organizations of visual areas were first explored by positioning flashing lights in various parts of the visual field and later by presenting bars and spots of light on a screen in front of the eye. Most of visual cortex can be stimulated by either eye, because both eyes project to the lateral geniculate nucleus of the visual thalamus of each cerebral hemisphere. However, the inputs to each nucleus are only from the portions of retina that are devoted to the contralateral half of visual space, the contralateral half of the visual field. In the visual cortex of the right cerebral hemisphere, neurons are activated by stimuli in the left visual field via the nasal half, or more, of the retina of the left eye and the temporal half, or less, of the right eye. Recordings from visual cortex at the posterior end of the cerebral hemisphere revealed a systematic representation of the contralateral half of the visual field, that was termed primary visual cortex, V1.

In all mammals, V1 corresponds to the classically defined architectonic subdivision of cortex, striate cortex or area 17 of Brodmann. In V1 of cats and rabbits, stimuli along the vertical midline (the zero vertical meridian) of the visual field activate neurons along the lateral border; stimuli progressively further into the lower visual quadrant activate neurons located progressively further anterior, while upper quadrant stimuli activate more posterior neurons. Stimuli in peripheral temporal vision activate more medially located recording sites in V1. In monkeys, V1 is rotated so that the border representing the zero vertical meridian is mediolateral in orientation, and much of V1 is on medial and ventral surfaces of the posterior end of the cerebral hemisphere, as well as within a brain fissure of the midline surface, a fissure that is found in all primates. Soon after the discovery of V1, recordings in cortex lateral to V1 in cats and rabbits revealed a second representation of the visual field, V2. This representation occupied a narrow strip of cortex, sometimes identified as area 18. V2 parallels V1 in representations of lower and upper visual quadrants along a common representation of the zero vertical meridian, but the outer border of V2 corresponds to the zero horizontal meridian. Thus, V2 is a 'split' or second-order transformation of the visual field. More recently, a number of additional representations of the visual field have been discovered, especially in well-studied cats and monkeys.

2. *Brain Maps Today*

After many years of intense study, a number of conclusions about brain maps seem justified. First, brain maps are often associated with anatomically distinct subdivisions of the brain. Recently, a number of histochemical methods have been developed to help reveal these subdivisions. Furthermore, some of these procedures reveal internal patterns within some cortical areas and subcortical nuclei that precisely conform to the representational topography of these structures. Thus, some brain maps have visible morphological correlates of their physiological maps. Most notably, there are maps of skin receptors that have visible subdivisions for each digit, pads of the palm, and often separate parts of the body. The maps contain isomorphs of the receptor sheet.

Second, maps vary from fairly precisely reflecting the order of the receptor sheet or muscle arrangement, to only roughly doing so. Thus, maps range from being highly topographical to being only roughly topographic. In general, the more precise maps are found early in sensory processing hierarchies, and less precise maps occur later. Thus, the brain stem nuclei that directly receive peripheral nerve afferents have the most detailed maps, while subsequent relays of this information to higher brain stem and thalamic nuclei may degrade this order, with further changes occurring in primary sensory cortex and beyond. Most higher-order representations are only globally topographic.

Third, even the gross topography of brain maps may be arranged in various ways. While primary visual cortex topographically represents the contralateral half of the visual field, the superior colliculus of the midbrain represents both the contralateral half of the visual field and much of the ipsilateral half of the visual field, including all that is seen by the contralateral eye. Only primates appear to have a superior colliculus that represents only the contralateral half of the visual field. Another example is the split repre-

sentation of the visual hemifield along the zero horizontal meridian in V2. The reasons for the different types of topographic representations are not clear. However since there can be different types of representations of the same receptor surface in the same individual, the arrangement of the receptor sheet is not the sole factor in the formation of maps.

Fourth, topographic cortical maps are commonly and perhaps always modular in internal organization. Primary visual cortex of most primates segregates inputs related to the right or left eye with so-called ocular dominance columns or bands, separates neurons according to preference for stimulus orientation and direction of movement in so-called orientation columns, and isolates different classes of neurons within and without a dot-like honeycomb pattern of neuron groups that express high amounts of the oxidative enzyme, cytochrome oxidase. The second visual area, V2, of monkeys and other primates has three types of band-like modules that are arranged in repeating sets of three along the length of V2. Thus, V2 actually consists of three fragmented, interdigitated maps of the contralateral visual field. Locally, there are discontinuities and repetitions in the map, but globally a topographic pattern is maintained. Primary somatosensory cortex, area 3b, isolates small patches of cortex devoted to slowly adapting cutaneous afferents from patches activated by rapidly adapting afferents. Auditory maps are unusual in that the receptor array is a row rather than a sheet, and thus a representation of these receptors covers only one dimension of a two dimensional surface of a cortical area. Thus, one dimension is of tone frequencies and the other dimension consists of the same frequency contours. Nevertheless, bands of neurons with similar binaural response properties are segregated from bands of neurons with different binaural properties. Motor maps are also quite modular and fragmented, with variable adjacencies and arrangements of modules of neurons for given movements. Thus, each type of module may be adjacent to a number of different types of other modules, possibly allowing more types of interactions for cooperative movements.

Fifth, there are maps that do not reflect the order of the receptor sheets, but are orderly in some other way. As these maps appear to be derived or computed from topographic maps, they have been called computational maps. As an often given example, a map of auditory space for the systematic location of sounds in space exists in the midbrain of birds and probably other vertebrates. This map does not reflect any arrangement of auditory receptors, but it is constructed from information derived from the differences in how the same sound source in different locations stimulates the two ears. Most higher-order cortical maps appear to reflect both topographic and computational components.

Sixth, species of vertebrates and especially mammals vary in the number of brain maps that they have. Mammals are impressive in that many have greatly expanded sheets of neocortex, providing more space for cortical maps. Most mammals resemble the first mammals in having small brains with relatively little neocortex. These mammals only have a few maps in cortex, usually S1, S2, and two or three other somatosensory maps, V1, V2 and one or more other visual maps, A1 and perhaps a few other auditory maps, and possibly but not always M1 and M2. In contrast, monkeys have a lot of neocortex and they have over 30 visual representations, 12 or more auditory representations, 10 or more somatosensory representations, and six or more motor representations. Possibly humans, with vastly more neocortex, have even more cortical maps. More complex information processing seems to depend on having more maps.

3. *The Development of Maps*

Sensory inputs to brain stem nuclei form orderly arrangements during development, and these nuclei form orderly connections with other nuclei, which in turn project to cortical areas in an orderly fashion and so on. The rough order in at least the early stations of this sequential process seems to depend on intrinsic factors in the brain. More specifically, it has long been hypothesized that afferents are chemically labeled according to some gradient, and target neurons are also chemically labeled with a comparable gradient across the target structure. Thus, the neurons find chemical matches and a topographic map is created. However, chemical or molecular matching is not the only factor in the development of brain maps, since there is much evidence that the development of such maps can be altered by changes in the sensory environment that induce changes in neural activity patterns. The details of sensory maps appear to be fine tuned by an adjustment and sorting of local connections in largely adaptive ways in response to variations in patterns of sensory activation. The claim is that neurons that fire together wire together. Coactive neurons with overlapping connections maintain and strengthen their synapses on shared target neurons, while neurons that do not respond at the same time tend to weaken and lose synapses on their shared target neurons. Thus, neurons with uncorrelated activity patterns tend to segregate their terminations so as to activate separate groups of neurons in the next structure. This allows maps to adjust and compensate for losses or additions of inputs, and maps in a sequence to adjust to each other.

There are many types of evidence that maps adjust to sensory changes during development. Usually, this evidence comes from some sort of experiment where some sensory input is removed or altered. Normal inputs expand to activate neurons formerly devoted to the removed or weakened inputs. Other evidence comes from the variations observed in maps in nature.

Rats and mice have morphological maps of the whiskers of the face in S1. Each whisker activates neurons in a small oval of cortical tissue that resembles a barrel, and is commonly called a cortical barrel. Rats and mice usually have the same number of whiskers, and there is one barrel in S1 for each whisker. Likewise there is one barrel-like structure for each whisker for the termination of afferents from the face in the brain stem, and in the ventroposterior nucleus of the thalamus. Thus, maps at three levels of the somatosensory system precisely match the arrangement and number of whiskers on the side of the face. This match could be due to innate developmental mechanisms, such as the matching of connections between levels as a result of a chemical code, but this appears to be unlikely.

Mice have been found that differ in the number of whiskers by having one or two more or less than the normal number, and in these mice the maps at all three levels match exactly the increase or decrease in whiskers. Thus, somehow the receptors in the skin are informing the central maps about their organizational state. In a similar way, the star-nosed mole has 11 sensory rays or protrusions from each side of its nose, and processed sections of somatosensory cortex reveal 11 corresponding bands of tissue in S1, one for each ray. Eleven bands are also seen in S2, the ventroposterior nucleus, and the somatosensory brain stem. Moles with 10 or 12 rays have been found, and 10 or 12 bands are then seen in all four brain structures in these moles. This matching of maps from one level to the next and with the peripheral receptor sheet would be most easily accomplished by a system that uses natural activity patterns to fine-tune its maps.

4. *Why Have Topographic Maps?*

Much of the brain is devoted to topographic maps suggesting that they are a useful way of organizing brain tissue. One cannot argue that such maps are simply preserving the order of afferents from the periphery, since afferents become mixed and disorderly, but sort themselves out again as they terminate. Instead, topographic maps appear to be adaptive in that they provide a substrate for making the most common types of computations in the nervous system over short, local connections. Neurons in topographic representations are well arranged to make center-surround comparisons between stimuli anywhere on the receptive field sheet, and stimuli in the immediate surround. This type of comparison provides biologically useful information. Neurons rapidly determine if a stimulus at some location is brighter, greener, harder, louder, or moving differently than stimuli in the immediate surround. While it may be possible to make such computations in nontopographic structures and systems, neurons would need to be massively interconnected and over larger distances. This would add unreasonable bulk to the brain, and would be metabolically costly. In complex brains with large amounts of neocortex, it is probably more efficient to have multiple interconnected maps, rather than just making maps larger. Large maps are most suitable for extremely local comparisons of stimuli, while more global comparisons can be mediated over short connections in small maps. Thus, it seems to be more productive to subdivide tasks, and process information in parallel over sequences of interrelated small maps, and use long connections to coordinate the activities of maps within and across processing sequences. Maps vary in size in the same individual because large maps are better at fine-grain local comparisons while small maps are better at more global comparisons.

See also: Cerebral Cortex: Organization and Function; Neural Plasticity in Auditory Cortex; Neural Plasticity in Visual Cortex; Visual System in the Brain

Bibliography

Kaas J H 1997 Topographic maps are fundamental to sensory processing. *Brain Research Bulletin* **44**: 107–12

Kaas J H 2000 Organizing principles of sensory representations. In: Bock G R, Cardew G (eds.) *Evolutionary Developmental Biology of the Cerebral Cortex*. Wiley, New York, pp. 188–205

Knudsen E I, duLac S, Esterly S D 1987 Computational maps in the brain. *Annual Review of Neuroscience* **10**: 41–65

Penfield W, Boldrey E 1937 Somatic motor and sensory representation in the cerebral cortex as studied by electrical stimulation. *Brain* **60**: 389–443

Udin S B, Fawcett J W 1988 Formation of topographic maps. *Annual Review of Neuroscience* **11**: 289–327

Woolsey C N, Marshall W H, Bard P 1942 Representation of cutaneous tactile sensibility in cerebral cortex of monkeys as identified by evoked potentials. *Bulletin of Johns Hopkins Hospital* **70**: 399–441

J. H. Kaas

Torts, in Law

Tort law is a fascinating topic of research for legal sociology generally, when law is observed from the viewpoint of legal sociology, two aspects are involved. One is 'to make observations putting law within society': observing how law operates in fact, within society. Another is 'to make observations of society within law': observing how the imprint of society is found in the law. Both aspects develop a circular, reflexive relationship, whereby law incorporating society impacts on society, and society being affected by law affects law. By then observing a particular society's law from a distance, adopting a comparative

sociological viewprint, one law appreciate the distinctive features characterizing that society's law.

Reading with such awareness the sociological literature on both law, the literature can be understood either as dealing with processes for achieving desirable social consequences with tort law operating in society or as dealing with processes whereby various interests and conceptions within society are read into the law.

1. Functions of Tort Law

Generally, it is said tort law has three goals: compensating harm, preventing accidents, and achieving justice. Tort law adopts the structure of requiring a person to compensate for harm caused to another when the former has been negligent in some way, defined as a breach of a duty of care. This is expected to encourage people to avoid obligations to compensate others, making them more careful, thus reducing accidents. Further, under this structure compensating harm is seen of itself as a desirable objective to the benefit of the injured and imposing this obligation directly on negligent people, rather than on third parties for the state, accords with our feelings of justice.

However, these three functions cannot always be achieved purely by implementing tort law (Sugarman 1989). In fact, various mechanisms are interposed, resulting in the creation of complex systems.

1.1 Compensating Harm

First, concerning 'compensation of harm,' various schemes of the third party assumption of the duty to compensate the injured have been developed. The most familiar one is the liability insurance, which covers the risk of tortfeasors incurring huge debts to pay compensation. It is also beneficial for the injured to receive full compensation. So, quite often the state intervenes for the sake of the injured either to require the tortfeasors to carry liability insurance or to assume the duty to compensate the injured while levying some form of 'tax' on the potential tortfeasors. This scheme of compulsory liability insurance is seen in many countries, especially in the case of automobile accidents compensation. The state managed compensation, scheme is also used when it is judged too serious to leave the matter in the hands of private insurers, as in the case of accidents caused by nuclear power plants. The state takes such an initiative when it is both feasible (easy to calculate the loss and/or to identify and levy the tax on the potential tortfeasors), and politically wise to have such schemes (to show the concern over the plight of the victims, or to assuage the fear of the local residents surrounding the power plant).

The extent to which the state initiates this kind of compensation scheme is dependent upon the political culture of the respective countries, however. The most elaborate initiative along these lines is known as the New Zealand model in which compensation is paid out from the pool of dues collected mainly from the designated firms and drivers. The similar notion of paying out compensation with greater certainty and without costly legal battles is found in the United States as well. But the actual scheme is more fragmented, with legislative interventions mostly creating no-fault compensation for injuries from automobile accidents. In Japan, within the framework of general tort law, still in principle based on fault, compensation payments have been standardized and ever more detailed compensation structures have been elaborated. Especially in automobile accident compensation this system is so refined that adversarial confrontation is virtually eliminated from the majority of cases (Tanase 1990).

Policy intervention in law is, therefore, facilitated by the emergence of insurance. Spreading costs through insurance lessens the difficulties involved, and it becomes generally easier politically to adopt policies oriented towards compensating injured parties. This promotes reform of tort law involving more certain compensation of injured parties and higher duties of care approaching strict liability. Such reform also reflects a general normative consciousness that society should be made safer and less traumatic, just as we have used science to overcome nature and create more amenable environments. This consciousness, along with more mundane political considerations, leads to what have been termed 'liability explosions,' or what Friedman (1985) calls a vision of 'total justice.'

1.2 Struggles over the Meaning of Compensation

Nonetheless, even if a trend does develop towards greater protection of injured parties through compensation, such protection does not come about uniformly and without tensions. Changing the rules creates winners and losers, and political dissension inevitably arises regarding such changes. To counter 'liability explosions,' politics emerge regarding legislative initiatives, and at a more micro level tort law rules are reinterpreted through courts and litigations. Tort reforms specifically to cap compensation amounts and limit rights to bring suit, by legislation, have been an ongoing and vocal point of political debate in the United States. Such political dissension about access to the law is an important topic for sociolegal analysis. In particular, as the political attack on the expansion of tort liability often involves critiques of the 'litigious society' and 'society of lawyers' in the US, it leads to an examination of problems in legal order in a broader sense and of how people relate to law.

In this regard, commentators taking views of tort law reform as being played out on the people's fear of losing control to professionals, for instance, have gained much support. They argued that assertions like 'litigation explosions hurt the US economy,' or frightening reports of 'suits by children for malparenting,' used to criticize the legalization of society, are unfounded in reality and based on hidden political agendas such as curbing the exercise of rights by consumers and minorities vis-à-vis corporations have strong elements of a campaign (Galanter 1983). The reality, they claimed, is that litigation turned out not to have grown so explosively. In fact, compared to the large body of accidents, people mostly acquiesced without asserting legal rights, or if they tried to do so, they were hampered in their attempts.

However, it seems that the roots of the problem lie deeper than suggested by these arguments. The political campaign against tort litigation developed in an era in which the American economy was losing its international competitiveness. Understandably, in these circumstances, the culprit was sought in businesses and individuals engaging in unproductive litigation strategies, rather than seeking to expand the pie, for instance, through technological innovation. At the same time, in this self-criticism of a society relying on litigation there was a clear nostalgia for the vision of an American society which hitherto had not indulged in litigation. One can smile at such anachronistic attitudes, but the shift towards a society in which law formed a central element for social integration and a basic model for social relations, had proceeded in many stages even in the United States. It is understandable that Americans would have felt concern especially over one important period of rapid transition, the 20 years from the 1960s, when the numbers of lawyers tripled. Yet in due time, neighboring institutions were reshaped, and they learned to play the game under the new rules, finding themselves once again surrounded by now familiar environments.This is a reason why criticism of the legalization of society died out, especially as the American economy revived over the 1990s.

Underlying these concerns is an appreciation that society and law operate on different logics. As the roles of law expand and law penetrates society more deeply, tensions and frictions from the collision of these different logics mount, which are reflected in the reaction against law by citizens involved in social life. This is true especially in tort litigation, which derives from events involving traumatic incidents such as loss of life. Torn by anger and sadness, people try to find 'life-world' meaning in a search for justice, and the law played out fully by over-zealous lawyers is resented as neglecting this aspect and transforming grief into greed. Unlike parties to a contract who make explicit legal commitments beforehand, parties in tort actions forge legal remedies out of ordinary contacts. So tort litigation is put through more rigorous tests of community standards before being accepted by people. For example, in a traditional American community, while contract litigation is taken for granted because 'one should naturally return money borrowed,' tort litigation tends to involve the dubious morality of 'getting another to take responsibility, rather than admitting one's own' (Engel 1983). This consciousness may have dissolved now in the United States, as further time has elapsed since the study was conducted. But we should not forget that there are always communities, which resist the control of the law, or that there are everyday interactions in which individuals give their own meanings to accidents, disputes, and solutions, which are not exhausted by legal frameworks. This is particularly important in sociolegal analyses aimed at understanding how law works within society, drawing on areas of law such as tort litigation.

2. *Strains within Tort Law*

2.1 *Deterrence through Tort Litigation*

There are also major corrections to the central mechanism underlying the traditional tort function of 'deterring accidents,' namely that tort law obligations create an incentive for tortfeasors to avoid harm caused by negligent behavior. The most important reason is that deterrence through obligations to compensate is undermined by 'moral hazard' resulting from insurance. Returning to a world without insurance remains too risky for both tortfeasors and injured parties, however, and hence an unrealistic policy option. Thus, deterrence of accidents must be assigned to legal enforcement by criminal or administrative sanctions, rather than tort law.

Yet such enforcement by specialist institutions then involves the problem of tending to lack necessary incentives and information to enforce the law at optimal levels. A system of legal enforcement must incorporate action by those who have suffered the injury, after all, and it is this aspect that is most distinctive about the United States. Known as the 'private attorney general' model (Cappelletti 1978), rights are given to injured parties under specific legislation, and systems are developed to lessen the economic burden of bringing suit, such as class actions, triple damage awards, and award of legal fees to the prevailing party in suits against the government. The possibility of being awarded punitive damages also heightens the capacity of injured parties to bring suit and to penalize the tortfeasors' negligent acts harshly. Although there is criticism that this goes too far, and punitive damages have become the biggest issue in the tort law reform debate, Americans seem to believe in the wisdom of utilizing private pecuniary interests to

advance the public interest in achieving the rule of law.

With legal enforcement capacity heightened in these ways, mass litigation is now frequently brought. In litigation involving drugs, asbestos, and recently tobacco, large companies have been forced into the corner through the pressure of massive compensation obligations, after political action proved ineffective. Particularly in regard to compensation for injuries resulting from smoking, an interesting tension emerges vis-à-vis American consciousness of liberty and dislike of paternalism (Kagan and Vogel 1993). When plaintiffs who had continued smoking claimed damages against tobacco companies, they met resistance from many citizens who asked 'didn't you continue smoking of your own free will?' This not only creates more difficulty in winning against tobacco companies (Rabin 1992) but also strains smokers in assuming 'victim' status. For to call oneself a victim, one must assert either that one is seduced to smoke by tobacco companies' advertisements or that one is too weak to stop smoking once hooked by nicotine. In either case, a smoker-plaintiff presents oneself as a weak person, someone who cannot 'smoke by choice' despite being given the 'freedom to smoke.' In fact, this kind of twist is not peculiar to tobacco litigation. Indeed, it is widely seen in other claim-making situations, wherein one assumes a victim status, whereby one is susceptible of being victimized (Bumiller 1988).

At the same time, however, in the case of tobacco, citizens' anger was directed towards tobacco companies as the greater evil especially when the companies were found to have lied before the public. This sentiment reflected upon jury trials. In one judgment finding tobacco companies liable, the foreman was quoted of saying 'companies must not lie about their products, this is our message.' Especially in these days of 'jury activism,' tort litigation functions to send messages from citizens to firms.

2.2 Morality in Tort Law

Finally, regarding the function of 'achieving justice,' while justice defined as 'just deserts' by classical tort law is not losing its importance, it is no longer enough to appreciate the contours of justice in contemporary tort law. This is connected, first, to the fact that many of the accidents which become the subject of tort law nowadays either arise repeatedly and extensively (like automobile accidents), or relate to more sporadic accidents which nonetheless can have large effects on society (such as product liability cases). In dealing with such cases, it is possible, and necessary in view of its extent, to calculate cumulative damages, and to weigh them against the costs necessary for their prevention. In fact, this risk management is carried out by companies as well as by the government.

In this 'collective justice' approach—'collective' in the sense of taking accidents as an aggregate, and 'justice' in the sense of achieving the maximum safety given a set amount of available resources—to give priority to the standpoint of efficiency in deciding on whom to impose what obligations may sometimes come into conflict with the moral meaning attributed to accidents premised on traditional notions of 'individual justice.' The moral dubiousness of tort litigation in general in the eyes of the self-reliant traditional villagers, or in particular of smokers claiming victim status, is evidence of this.

Furthermore, related to such totalizing tendencies, there is a problem associated with the fact that tortfeasors come to lose a sense of pain in relation to making compensation, as liability insurance, both private and publicly managed, becomes pervasive. Particularly in automobile accident cases, insurance companies take over settlement negotiations in place of the insured, and the matter is resolved without the tortfeasor ever having to face the pain caused by his or her acts to the injured party. This may offend the sense of justice held by at least some people. Gilligan (1982) points out the existence of 'a different voice,' the morality of 'care and responsibility,' the core sensibility is to empathize with the pain of the other caused by his or her act and think through the way to minimize it. To the extent that this moral reasoning is foreclosed from the workings of tort law, the latter is judged to be morally deficient.

In fact, leaving matters to the insurance company in this way still arouses strong condemnation from injured parties in Japan. A tortfeasor is expected to visit the hospital or to join the funeral, enduring the baleful gaze of the injured party and his or her family, as a proper attitude of 'human sincerity.' It would be enough to pay compensation in the form of damages arising from a breach of one's legal obligations, but there is an important truth hidden in this seemingly irrational ceremonial apology on top of the legal duty. That is, the law defining the dispute and resolving it represents only one part of the whole dispute. For an injured party affected by an accident, someone able to empathize with his or her hurt is a necessary part of recovery so aside from monetary compensation, the matter must be viewed 'pre-legally' as involving relations between vulnerable human beings.

This much may be obvious. In the United States, too, this role for apology is presently being revisited. But in the legal world, to make an apology is not easy. For it is deemed admitting responsibility and conceding defeat from the outset. Although some states in the US have passed laws legally not to use an apology as evidence, the 'attribution of responsibility' gave, rejecting one's own responsibility and pushing responsibility onto others, is played out. In these adversarial circumstances, apology itself easily comes to be seen as a mere gesture (Taft 2000). More generally, the sense of liberty abhorring constraints

other than those expressly assumed, or mandated by law, is at the core of the law. So 'moral reductionism,' divorcing issues from nonlegal duties, and forcing them into the framework of legal duties, easily prevails.

The future is opaque. The trend towards 'legalization'—increasing resort to law—is generally proceeding irreversibly, and these 'pre-legal' negotiation processes are gradually being abbreviated. But the counter-ideal emanating from the direct experience of the ways people interact within community may put the validity of law in question. This 'law skepticism' comes to surface most acutely in the case of tort law as it directly touches the daily lives of the common people, and hence is affected by commonsense interpretations of justice. In this sense, 'communitarian justice' should inform the law along with the other two notions of justice.

See also: Civil Liberties and Human Rights; Class Actions: Legal; Deterrence; Guilt; Insurance and the Law; Liability: Legal; Litigation

Bibliography

Bumiller K 1988 *The Civil Rights Society: The Social Construction of Victims*. The Johns Hopkins University, Baltimore, MD

Cappelletti M 1978 Access to justice: The newest wave in the worldwide movement to make rights effective. *Buffalo Law Review* **27**: 181–292

Engel D 1983 Cases, conflict, and accommodation: Patterns of legal interaction in an American community. *American Bar Foundation Research Journal* : 803–74

Friedman L 1985 *Total Justice*. Russell Sage Foundation, New York

Galanter M 1983 Reading the landscape of disputes: What we know and don't know (and we think we know) about our allegedly contentious and litigious society. *UCLA Law Review* **31**: 4–71

Gilligan C 1982 *In A Different Voice*. Harvard University, Cambridge, MA

Kagan R, Vogel D 1993 The politics of smoking regulation: Canada, France, the United States. In: Rabin R, Sugarman S (eds.) *Smoking Policy: Law, Politics, and Culture*. Oxford University Press, New York

Rabin R 1992 A sociolegal history of the tobacco tort litigation. *Standard Law Review* **44**: 853–78

Sugarman S 1989 *Doing Away with Personal Injury Law*. Quorum, New York

Taft L 2000 Apology subverted: The commodification of apology. *Yale Law Journal* **109**: 1135–60

Tanase T 1990 The management of disputes: Automobile accident compensation in Japan. *Law & Society Review* **24**: 651–91

T. Tanase

Torture and its Consequences, Psychology of

1. Introduction

Torture remains a widespread problem despite numerous international conventions prohibiting its practice. There are no reliable estimates of the prevalence of torture in the world. The only data available are those published by human rights organizations such as Amnesty International (AI). The 1998 Report of AI listed over 150 countries known for some form of human rights violation in the previous year (Amnesty International 1998). According to an analysis (Başoğlu 1993) of the 1992 AI report, systematic torture was reported in 93 of the 204 countries in 1991. Although reports of torture were more common from countries affected by political unrest, the problem was not confined to these regions. Systematic torture and/or ill-treatment in police custody or prisons were also reported in 25 percent of the Western European and North American countries. Thus, although torture is more widespread in certain parts of the world, no society or ideological system enjoys complete freedom from gross human rights violations.

Little is known about the prevalence of torture among various at-risk populations. About 5–35 percent of the world's 14 million refugee population (0.7–4.9 million refugees) are estimated to have had at least one experience of torture (Baker 1992). These figures do not reflect the current extent of the problem after the recent developments in Eastern Europe, former Yugoslavia, Middle East, and other parts of the world torn by political turmoil, nationalistic movements, and regional wars.

In this article, the evidence on the long-term mental health effects of torture will be reviewed. Torture is a sociopolitical phenomenon with wide-ranging physical, psychological, social, economic, and political consequences. Our knowledge on these issues is limited owing to the scarcity of systematic research. This review will focus on the psychological effects of torture as relatively more work has been done in this area. Detailed information on the physical effects of torture can be found elsewhere (Skylv 1992). The reader is also referred two books (Başoğlu 1992, Jaranson and Popkin 1998) for more detailed information on other aspects of the problem.

The difficulties in defining torture have been widely recognized. Scientific study of this problem requires a clear definition. The definition adopted for the present review is the one that has been provided by the 1986 United Nations Declaration of Human Rights: 'any act by which severe pain or suffering, whether physical or mental, is intentionally inflicted on a person for such purposes as obtaining from him or a third person information or a confession, punishing him for an act he or a third person has committed or is suspected of

having committed, or intimidating or coercing him or a third person, or for any reason based on discrimination of any kind, when such pain of suffering is inflicted by or at the instigation of or with the consent or acquiescence of a public official or other person acting in an official capacity.'

2. *Psychological Effects of Torture*

Most studies of torture survivors have methodological limitations that preclude definitive conclusions on the psychological effects of torture. These include insufficient description of the interview procedures, assessment instruments, diagnostic criteria, and medical diagnoses. Few studies have examined the relationship between the symptoms and the diagnosis of post-traumatic stress disorder (PTSD) (see *Post-traumatic Stress Disorder*). Most studies used an uncontrolled design and thus did not control for other nontorture, potentially traumatic life events. Furthermore, most of these studies have been carried out on refugees so the additional effects of refugee trauma are not controlled for. In recent years, however, controlled studies have been carried out, attempting to overcome some of the methodological problems of the earlier studies. The evidence will be reviewed in two sections, uncontrolled and controlled studies, with greater emphasis on the latter.

2.1 *Uncontrolled Studies*

In a review of eight studies by Goldfeld et al. (1988), the most common problems in torture survivors were anxiety, depression, irritability/aggressiveness, emotional lability, social withdrawal, disorientation, memory and concentration impairment, impaired reading ability, lack of energy, insomnia, nightmares, and sexual dysfunction. Another review of the literature (Somnier et al. 1992) involving 46 studies, reported similar findings. Although cerebral atrophy, abnormal sleep patterns, somatization, and traumatic brain injury have also been reported (reviewed in Başoğlu 1997), the evidence on these findings is not conclusive.

Some of the evidence comes from studies of former prisoners-of-war (POWs) and concentration-camp survivors. In a review of the literature, Miller (1992) concluded that former POWs continue to have anxiety, depression, and PTSD 40 years postcaptivity and torture during captivity is related to more severe psychological problems in the long term. (Solkoff 1992), in his review of the literature on concentration-camp survivors, pointed to various methodological problems which characterize the research from which the 'survivors' syndrome' evolved. The author concluded that the effects of Holocaust-related trauma on survivors and the second generation have not been adequately investigated.

2.2 *Controlled Studies*

A Danish study (Hougen et al. 1988) compared 14 tortured refugees with 14 nontortured controls. The torture survivors had more psychological symptoms than did the controls. The sample size, however, was too small and matching of controls was not adequate. The measurement of symptoms was not sufficiently detailed. Another controlled study (Petersen and Jacobsen 1985) of Spanish torture survivors found that tortured individuals, relative to controls, had more symptoms of depression, anxiety, emotional lability, sleep disturbance, nightmares, and memory and concentration difficulties. However, the measures were not sufficiently detailed and standardized instruments based on established diagnostic criteria were not used. In addition, the sample size ($n = 10$) was too small for statistical analysis.

A more recent study in Turkey (Başoğlu et al. 1994a) used standardized assessment instruments in comparing 55 tortured political activists with 55 nontortured political activists and 55 nontortured, nonactivist controls, matched for age, sex, marital and socio-cultural status. The torture survivors had significantly more lifetime and current PTSD than did the controls (33 percent vs. 11 percent and 18 percent vs. 4 percent, respectively). Secondary impact of captivity/torture on family, family history of psychiatric illness, and post-torture psychosocial stressors related to long-term psychological status (Başoğlu et al. 1994b). Both torture and subsequent psychosocial stressors contributed to traumatic stress symptoms (Başoğlu and Paker 1995). In an attempt to explain the relatively low prevalence of PTSD in this sample, the investigators pointed to the possible protective role of a strong belief system, commitment to a cause, prior knowledge and expectations of torture, and prior immunization to traumatic stress (Başoğlu et al. 1994b). The majority of the survivors were highly committed political activists with prior expectations of and psychological preparedness for torture. They also had a negative appraisal of the state and had not maintained beliefs concerning a 'benevolent state' before the trauma, another factor that may have protected them against the traumatic effects of torture (Başoğlu et al. 1996).

The study by Başoğlu et al. (1994a), while pointing to the possible protective role of 'psychological preparedness for trauma,' could not test this hypothesis because most study participants were highly committed political activists. A further study (Başoğlu et al. 1997) was carried out to compare 34 less psychologically prepared, nonactivist torture survivors with the previous sample of 55 tortured political activists. Compared with the political activists, the nonactivists were relatively less severely tortured but had significantly more PTSD (58 percent vs. 18

percent) and major depression (24 percent vs. 4 percent). These findings supported the role of prior immunization in reducing the effects of traumatic stress, and of unpredictability and uncontrollability of stressors in exacerbating the effect of torture.

Another study (Maercker and Schutzwohl 1997), which investigated the long-term effects of political imprisonment in the former German Democratic Republic, compared 146 former political prisoners with 75 controls matched for age and sex. In comparison with controls, the former political prisoners had significantly higher rates of lifetime (59.6 percent) and current PTSD (30.1 percent), claustrophobia, social phobia, substance abuse, and higher scores on measures of anxiety, depression, general psychopathology, and dissociative symptoms.

A controlled study by Shrestha et al. (1998) involved 526 Bhutanese refugee survivors of torture in Nepal matched with 526 nontortured refugees on age and sex. The study group was selected randomly from the Bhutanese refugee community in the United Nations refugee camps in eastern Nepal. Compared with the control group, the torture survivors had more PTSD symptoms, higher anxiety and depression scores, and more musculoskeletal and respiratory system complaints. The authors concluded that torture may increase the risk of mental health problems among refugees and that PTSD symptoms appear to be part of a universal reaction to torture.

A further study (Mollica et al. 1998) in the USA involved 62 Vietnamese ex-political detainees and 22 controls. Compared with the controls, the ex-political detainees showed significantly higher scores on depression and PTSD symptoms but the prevalence rates of PTSD (77.3 percent vs. 88.2 percent, respectively) and major depression (36.4 percent vs. 56.9 percent, respectively) across the two groups did not reach statistical significance. Severity of torture was the most significant predictor of depression and PTSD symptoms, particularly of arousal symptoms among the latter. The authors suggested that a small sample size may have accounted for the lack of statistically significant differences in the prevalence of psychiatric diagnoses.

A controlled study (Holtz 1998) in India involved 35 tortured refugee Tibetan nuns and lay students and 35 nontortured controls matched for sex, age, lay status, and years in exile. Relative to the controls, the torture survivors had significantly higher scores on measures of anxiety but the two groups did not differ on ratings of depression. Lack of prior knowledge of and psychological preparedness for torture was a significant predictor of post-torture anxiety symptoms. Other predictors included male sex, longer duration of torture, exposure to solitary confinement, feelings of hopelessness in prison, and recent arrival in India. The author suggested that the survivors' Tibetan Buddhist training as nuns may have contributed to their resilience against the traumatic impact of torture.

In conclusion, there is clear evidence that in some individuals torture has psychological effects independently of other associated stressors such as harassment and persecution by the authorities, unlawful detentions, unfair trials, imprisonment, exposure to mass violence and life-threatening situations, threats of death to self and family, having to go into hiding, loss of employment or educational opportunities, bereavement, uprooting, and refugee status. The distinction between the effects of torture *per se* and those of other associated stressors is important for medical and legal reasons.

3. *Implications of Current Knowledge for Survivor Care Policies*

The research evidence reviewed so far may have important implications for legal and public health policies concerning torture survivors. Immigration, asylum, and refugee care policies need to take into account the fact that a significant proportion of asylum seekers and refugees may have a history of torture and be at risk of disabling psychological problems. Due attention to mental health issues may be as important as providing legal, social, and economic aid for refugees. Research findings (Başoğlu and Paker 1995) suggest that social support may not be sufficient for recovery unless it is complemented with effective psychological treatment. Greater funding needs to be allocated for mental health care services for tortured refugees and mental health professionals need to play a more central role in policy making in refugee care.

Given the prevalence of torture in the world and the mental health hazards posed by its chronic and disabling psychological consequences, there is an urgent need for psychological treatments with demonstrated efficacy. Survivor care policies need to place greater emphasis on developing effective treatments through systematic research and on training of mental health workers in treatment techniques.

Finally, there has been relatively little scientific interest in the study of torture and the treatment of its consequences, despite the serious mental health hazards posed by the problem. Much of the available resources for survivor care is often used for social, legal, and economic aid and for various rehabilitation programs, the efficacy of which has not yet been demonstrated. In policy making, research is either given low priority or sometimes even dismissed as unethical. Scientific research with due regard for ethical issues is possible and essential for effective survivor care. Study of torture survivors also has important implications for human rights, theory, assessment, classification, treatment of traumatic stress responses, and legal issues concerning torture survivors.

4. Directions for Future Research

Future research needs to focus on three main issues: psychobiological mechanisms of traumatization, factors related to long-term psychological functioning, and effective treatments for survivors of torture. Given the evidence on the role of psychological preparedness for trauma as a protective factor, further studies are needed to elucidate the behavioral and cognitive components of this construct. A better understanding of resilience factors is needed to develop more effective treatment programs.

More research is needed to clarify cognitive responses to torture and their bearing on long-term outcome. Among these cognitive responses, sense of injustice arising from impunity for perpetrators is of particular importance. Little is known how various forms of compensation and redress relate to survivors' psychological functioning. Studies of this issue would also be useful in clarifying the psychological effects of 'truth and reconciliation' processes, such as in South Africa, both on survivors and on the community. It could also provide insights into how such processes should be conducted to avoid further traumatization and maximize the psychological well-being of survivors and their community.

Further work is needed to examine how various stressors such as torture, uprooting, refugee trauma, and loss of social support relate to psychological problems in survivors of torture. International collaborative studies are needed to develop standardized and validated assessment instruments for refugee and nonrefugee torture survivors.

There is need for greater emphasis on research in developing effective treatments for torture survivors. Controlled treatment trials are needed to identify the most effective treatments and their mechanisms of action. Given the scarce resources available for the care of torture survivors, the efficacy of current rehabilitation models needs to be evaluated and their therapeutic ingredients clarified. For Western countries facing a serious refugee problem, epidemiological studies are needed to investigate the prevalence of past torture and associated psychosocial problems among refugee populations.

See also: Human Rights, Anthropology of; Human Rights, History of; Human Rights: Political Aspects; Post-traumatic Stress Disorder; Torture: Legal Perspectives

Bibliography

Amnesty International 1998 *Amnesty International Report*, Amnesty International Publications, London

Baker R 1992 Psychosocial consequences for tortured refugees seeking asylum and refugee status in Europe. In: Başoğlu M (ed.) *Torture and Its Consequences: Current Treatment Approaches*. Cambridge University Press, Cambridge, UK

Başoğlu M 1992 *Torture and Its Consequences: Current Treatment Approaches*. Cambridge University Press, Cambridge, UK

Başoğlu M 1993 Prevention of torture and care of survivors: an integrated approach. *Journal of the American Medical Association* **270**: 606–11

Başoğlu M 1997 Torture as a stressful life event: a review of the current status of knowledge. In: Miller T W (ed.) *Stressful Life Events*. International Universities Press, New York, pp. 45–70

Başoğlu M, Paker M 1995 Severity of trauma as predictor of long-term psychological status in survivors of torture. *Journal of Anxiety Disorders* **9**: 339–50

Başoğlu M, Paker M, Paker Ö, Özmen E, Marks I M, Incesu C, Şahin D, Sarimurat N 1994a Psychological effects of torture: a comparison of tortured with nontortured political activists in Turkey. *American Journal of Psychiatry* **151**: 76–81

Başoğlu M, Paker M, Paker Ö, Özmen E, Ŏahin D 1994b Factors related to long-term traumatic stress responses in survivors of torture in Turkey. *Journal of the American Medical Association* **272**: 357–63

Başoğlu M, Paker M, Özmen E, Taşdemir O, Şahin D, Ceyhanli A, Incesu C, Samimurat N 1996 Appraisal of self, social environment, and state authority as a possible mediator of post-traumatic stress disorder in tortured political activists. *Journal of Abnormal Psychology* **105**(2): 232–6

Başoğlu M, Mineka S, Paker M, Aker T, Livanou M, Gök Ş 1997 Psychological preparedness for trauma as a protective factor in survivors of torture. *Psychological Medicine* **27**: 1421–33

Goldfeld A E, Mollica R F, Pesavento B H, Faraone S V 1988 The physical and psychological sequelae of torture: symptomatology and diagnosis. *Journal of the American Medical Association* **259**: 2725–9

Holtz T H 1998 Refugee versus torture trauma: A retrospective controlled cohort study of Tibetan refugees. *Journal of Nervous and Mental Disease* **186**: 24–34

Hougen H P, Kelstrup J, Petersen H D, Rasmussen O V 1988 Sequelae to torture. A controlled study of torture victims living in exile. *Forensic Science International* **36**: 153–60

Jaranson J M, Popkin M K 1998 *Caring for Victims of Torture*. American Psychiatric Press, Washington, DC

Maercker A, Schutzwohl M 1997 Long-term effects of political imprisonment: a group comparison study. *Social Psychiatry and Psychiatric Epidemiology* **32**: 435–42

Miller T W 1992 Long-term effects of torture in former prisoners of war. In: Başoğlu M (ed.) *Torture and Its Consequences: Current Treatment Approaches*. Cambridge University Press, Cambridge, UK, pp. 107–35

Mollica R F, McInnes K, Pham T, Smith Fawzi M C, Murphy E, Lin L 1998 The dose–effect relationship between torture and psychiatric symptoms in Vietnamese ex-political detainees and a comparison group. *Journal of Nervous and Mental Disease* **186**: 543–53

Petersen H D, Jacobsen P 1985 Psychical and physical symptoms after torture. A prospective controlled study. *Forensic Science International* **29**: 179–89

Shrestha N M, Sharma B, van Ommermen M, Regmi S, Makaju R, Komproe I, Shrestha G B, de Jong J 1998 Impact of torture on refugees displaced within the developing world: symptomatology among Bhutanese refugees in Nepal. *Journal of the American Medical Association* **280**: 443–8

Skylv G 1992 The physical sequelae of torture. In: Başoğlu M (ed.) *Torture and Its Consequences: Current Treatment Approaches*. Cambridge University Press, Cambridge, UK, pp. 38–55

Solkoff N 1992 The Holocaust: survivors and their children. In: Başoğlu M (ed.) *Torture and Its Consequences: Current Treatment Approaches*. Cambridge University Press, Cambridge, UK, pp. 136–48

Somnier F, Vesti P, Kastrup M, Genefke I K 1992 Psycho-social consequences of torture: current knowledge and evidence. In: Başoğlu M (ed.) *Torture and Its Consequences: Current Treatment Approaches*. Cambridge University Press, Cambridge, UK, pp. 56–71

M. Başoğlu

Torture: Legal Perspectives

Torture is regarded as an aggravated kind of criminal violence, mainly carried out by state agents, but occasionally by non-state actors or with the toleration of governments. It can be perceived as the strongest form of inhuman or degrading treatment in civil society. It has often been used as a tool in criminal proceedings to force confessions from a suspect. Since the time of the Enlightenment in eighteenth-century Europe, nations increasingly have disapproved of torture and, at least since World War II, have considered torture a serious violation of fundamental human rights. The Law of Nations has universally prohibited torture: In the Universal Declaration of Human Rights of 1948, Art. 5 states 'no one shall be subjected to torture or to cruel, inhuman or degrading treatment or punishment.' Nevertheless, some detainees and prisoners still suffer. Therefore, an urgent need exists for more effective prevention as well as for support and treatment of victim survivors.

1. *Defining Torture*

According to the United Nations Convention against Torture (UNCAT 1984, Art. 1 sect. 1), torture is defined as 'any act by which severe pain or suffering, whether physical or mental, is intentionally inflicted on a person for such purposes as obtaining from him or a third person information for a confession, punishing him for an act he or a third person has committed or is suspected of having committed, or intimidating or coercing him or a third person or for any reason based on discrimination of any kind, when such pain or suffering is inflicted by or at the instigation of or with consent or acquiescence of a public official or other person acting in an official capacity. It does not include pain or suffering arising only from, inherent in, or incidental or (*sic*) lawful sanctions.'

Although this notion is widely accepted, it is still a matter of controversy whether the concept of torture should be enlarged to incorporate similar patterns carried out by non-state actors or organizations, such as armed opposition groups. These groups have used torture as a weapon of war, to terrorize and thereby to control segments of the population. For example, mass rape of women as a tactic of 'ethnic cleansing' in the wars in the former Yugoslavia, combined with a feminist reconceptualization of human rights, made it impossible for the world community to continue to keep gender-based violence outside the circle of urgent anti-torture strategies (cf. Copelon 1993/4, pp. 293 and 367). Because the growing public sentiment could no longer be ignored, rape and sexual abuse of both females and males can now be considered a form of torture.

2. *Historical Development*

There is a long tradition of torture in various countries, although its objectives differ widely. It has accompanied mankind since approximately the beginning of the hierarchical organization of society. In the ancient city of Athens, torture was used to extort confessions from slaves. In the Roman Empire, free men as well as slaves were tortured for the purpose of obtaining or confirming evidence. Torture was also employed by government authorities against detained persons. In Europe, the use of torture as a means of directly extorting evidence dates back to at least the fourteenth century. Although early legal records demonstrate awareness of the dangers inherent in torture, its application was intensified in the so-called trials by inquisition. The developmental significance of these trials hence ought not to obscure the fact that the torture system authorized the commission of crimes far more serious than those for which the torture was imposed. Only sporadic attempts were made to stem these excesses.

The Age of Enlightenment initiated the modification and, sometimes, the abolition of torture and inhumane treatment by the state. This was no doubt due not only to the realization that such a line of action is inhumane, but also to the fact that trials by inquisition (prevailing at that time) were increasingly subject to criticism. As a consequence, the main goal of torture (to extort a confession at any price) lost importance.

Despite these advances, various codes of criminal procedure in the nineteenth century still stipulated that an accused who refused to disclose information or who denied facts was to be either caned or imprisoned in chains. Indeed, some of these codes or their relevant provisions remained in effect until almost the mid-twentieth century. The renaissance of torture and inhumane treatment of criminals under the auspices of authoritarian or totalitarian state systems, particularly under the national socialist regime during World War

II, shocked the entire international community and intensified efforts towards recognition of antitorture standards, protection of vulnerable people and rehabilitation of victim survivors.

3. International Conventions and Control Mechanisms

Although the prohibition of torture has been effective as imperative international customary law since about the end of World War II, and despite the fact that Art. 5 of the Declaration of Human Rights of December 10, 1948 explicitly includes this prohibition, torture and other kinds of serious ill-treatment of detainees are still rather common and, in some parts of the world, an everyday occurrence. Many victims are tortured with modern technology ('high-tech torture'), at times even under medical supervision, while some are subjected to 'isolation torture.' These modern methods of torture are unique because they do not leave visible marks. Torture has taken on a more psychological character, causing permanent mental disorder in the victim. Ill-treatment has therefore become rather difficult to ascertain and to verify. Thus, the need to advocate the prevention of torture and the protection of human rights remains acute.

The international community is developing control mechanisms and counter-measures against torture. The United Nations and the Council of Europe have increased their efforts to combat torture and other forms of ill-treatment by the introduction of special control mechanisms (cf. Bank 1996, pp. 9 et seq., 19 et seq., 381 et seq.). Art. 7 of the International Covenant on Civil and Political Rights of December 19, 1966 and Art. 3 of the European Convention for the Protection of Human Rights and Fundamental Freedoms of November 4, 1950 (European Human Rights Convention, EHRC), which have legal force at least in all European states. Both explicitly guarantee the protection of detainees from torture and from inhuman or degrading treatment or punishment. In order to assure the fundamental freedoms under the conditions of deprivation of liberty, the adoption of a supplementary protocol to the EHRC is currently being discussed. If adopted, this provision is intended to strengthen the legal position of detained persons.

In 1985, the UN Commission of Human Rights appointed a special rapporteur to prepare and submit an investigative report of torture (cf. e.g., Rodley 1994). In 1987, the Convention against Torture and other Cruel, Inhuman, or Degrading Treatment or Punishment (UNCAT) became effective, and was simultaneously backed up by the establishment of the Committee against Torture (CAT), a taskforce set up to oversee implementation of the Convention. The European Convention for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment, and the European Committee for the Prevention of Torture (ECPT) came into being in 1989 (cf. Council of Europe 1989, pp. 2 et seq., Evans and Morgan 1998, p. 1 et seq.). Moreover, in 1987, the organization of American States enacted the Inter-American Convention to Prevent and Punish Torture, with the aim of imposing substantive obligations on member states very similar to those enshrined in the UNCAT. But as the American Convention neither provides for a specialized body to pursue implementation nor introduces any novel procedures, it has had only a small impact. Nevertheless, the Inter-American Court on Human Rights on December 2, 1999 condemned the State of Guatemala for the torture of five street youths in violation of the Inter-American Convention for the Prevention and Punishment of Torture.

To date, the most reliable information on torture, inhuman and degrading treatment is based on the reports of the coalition of International Nongovernmental Organizations (NGOs) against torture (CINUT) and the so-called fact-finding missions of the European Anti-Torture Committee (ECPT). The ECPT carried out more than 90 visits to nearly 40 European countries during the 1990s (cf. Council of Europe 1999, pp. 2 et seq., Kaiser 1995, pp. 2 et seq.), starting in Ireland and ending in Siberia at the Japanese Sea. On this solid basis, enriched by the worldwide reports of NGOs (e.g., Amnesty International 1997, pp. 3 et seq.), one can generalize the findings of present-day torture and ill-treatment practices.

4. Pressing Problems and 'Sensitive' Domains

In the 1990s, much of the 'torture' literature has focused on abuses by police officers and, through solitary confinement and overcrowding of prisons in general, on the totality of inhumane custodial conditions. Complaints of prisoners that came to ECPT's attention ranged from serious bodily torture (hanging up by the arms, electric shocks, contusion of testicles, caning of soles of the feet [*fallaka*], violent penetration of body orifices with truncheons) to considerable mental injuries resulting from solitary confinement in extremely small, unlit and unventilated cells. Many of the complaints were confirmed by further evidence; frequently the alleged acts of ill-treatment were verified medically. However, none of these acts was shown to be systematic. It is hence to be assumed that amendments to laws or even to constitutions are not alone sufficient to prevent the ill-treatment of prisoners. Rather, what is needed is a fundamental change of attitude throughout the entire system of criminal justice: we must refuse to accept that torturers are simply 'ordinary people' whose willingness to carry out acts of torture can be attributed to the brutal socialization techniques of totalitarian regimes.

Observations indicate that the risk of ill-treatment by prison personnel is considerably lower than the risk

posed by local police. At present, however, it seems that inhuman living conditions and a pattern of torturous behavior by the prison staff are not unheard of in some institutions in which prisoners are held prior to trial.

5. *People Particularly Vulnerable to Torture*

Certain categories of people are particularly vulnerable to torture: aliens, minorities, and apprehended persons from the moment of arrest throughout their imprisonment. Detention of asylum seekers and refugees, especially those likely to be expelled, presents considerable difficulties with respect to equal treatment and the prohibition of discrimination. Penitentiaries in European capitals or in large cities with an international airport frequently harbor detainees from more than 30 countries. All in all, native and foreign prisoners speak various languages, engage in a variety of eating habits and religious practices, and have diverging recreational needs. These circumstances render difficult communication, treatment, and assurance of basic rights such as visits, exchange of letters, and telephone contact. Restricted communication with the outside world, denial of education, and denial of privileges are often imposed by the prison administration for security reasons upon the foreign detainee. The dangers of tension in the institution, of interpersonal violence between detainees, and of ethnic conflicts between different groups of foreign inmates are additional aggravating circumstances. In this respect, the prison management must counteract possible conflicts by distributing the prison population sensitively and responsibly.

6. *Overcoming the Restricted Impact of Control Mechanisms*

All international organs and agencies acknowledge the limited impact of their activities. Because fewer than 120 state parties have ratified the UNCAT, it remains one of the least ratified human rights treaties at the level of UN. Additionally, the international organizations are sometimes confronted by obstructive government reactions; their recommendations for structural changes are not always met with positive responses. Of course, the most effective results could be produced with the help of an international control network representing the international bodies themselves. In particular, follow-up visits would constitute an effective monitoring tool. Nevertheless, none of the agencies—including the ECPT—are in a position to provide close and continuous monitoring of new developments. Careful observations by the NGOs might contribute and fill the gaps.

The future of torture control ought not to be confined to Europe. The establishment of a universal visitation system, based on the model of the European Convention, is as necessary as ever. Therefore, the prospects for effective protection against torture throughout the world will depend on the introduction of a global system of visits. Lasting control of torture and effective improvements in the conditions of detention presuppose international recognition of the preventive visitation system. The existence of international conventions and the criminalization of acts of torture do not always succeed in preventing ill-treatment and abuse. Thus, we must attach increasing importance to the training of police and prison officials in accordance with the relevant recommendations of the United Nations.

In addition to strong efforts to repress and prevent torture, the need to provide for the care and rehabilitation of torture survivors is urgent. This is regarded as a new field and is supported by the International Rehabilitation Council for Torture Victims.

See also: Civil Law; Civil Liberties and Human Rights; Common Law; European Union Law; Human Rights, History of; Religion: Peace, War, and Violence; Torture and its Consequences, Psychology of; Violence, History of

Bibliography

Amnesty International (ed.) 1997 *The International Conference on Torture. Final Report*. Amnesty International Publications, Stockholm

Amnesty International (ed.) 1999 *Amnesty International Report*. Amnesty International Publications, London

Bank R 1996 *Die internationale Bekämpfung von Folter und unmenschlicher Behandlung auf den Ebenen der Vereinten Nationen und des Europarates: Eine vergleichende Analyse von Implementation und Effektivität der neueren Kontrollmechanismen*. Edition Iuscrim, Freiburg, Germany

Copelon R 1993/4 Recognizing the egregious in the everyday domestic violence as torture. *Columbia Human Rights Law Review* **25**: 291–367

Council of Europe (ed.) 1989 *Explanatory Report on the European Convention for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment*. Directorate of Human Rights, Strasbourg

Council of Europe (ed.) 1999 *The Prevention of Torture at the Dawn of a New Millennium*. On the occasion of the 10th anniversary of the European Convention for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment (CPT). 'Substantive' sections of the CPT's general reports. Strasbourg

Evans M, Morgan R 1998 *Preventing Torture: A Study of the European Convention for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment*. Oxford University Press, Oxford, UK

Kaiser G 1995 Detention in Europe and the European Committee for the Prevention of Torture. *European Journal of Crime, Criminal Law and Criminal Justice* **3**: 2–17

Lakes G 1995 *An Overview of the Prison Systems of the Baltic States*. European Committee on Crime Problems, Council of Europe, Strasbourg

Rodley N S 1994 *Report of the Special Rapporteur, Mr Nigel S. Rodley, submitted pursuant to Commission on Human Rights Resolution 1994/37: addendum: Visit by the Special Rapporteur to the Russian Federation*. United Nations, Geneva

UNCAT 1984 United Nations convention against torture, inhuman and degrading treatment or punishment. In: Council of Europe (ed.) *Human Rights in International Law. Basic Texts*. Council of Europe, Strasbourg, pp. 68ff.

Walmsley R 1995 The European prison rules in Central and Eastern Europe. *European Journal of Criminal Policy* **3–4**: 73–90

G. Kaiser

Tory

The term Tory is colloquially used as a synonym for a British Conservative. However, it is best defined as a particular tendency within conservative thought that stresses order over liberty. Conservatism is marked by a tension between two of its fundamental principles, the necessity of order and the desirability of liberty (see *Conservatism: Historical Aspects*). The former is concerned with how government is conducted; the latter with the role and limits of government. For example, conservative communists can exist in Russia in the first sense, but not in the second. Often, perhaps usually, these two elements of conservatism are mutually reinforcing, but not always. In this tension, Tories emphasize order. This article will examine six issues: origins; principles; history; contemporary forms; relevance outside the United Kingdom; and future.

1. *Origins*

The origin of the word 'Tory' comes from the Irish *toraighe*, an Irish thief dispossessed of his lands and existing on raids on English settlers. It was used in British politics as a term of abuse during the debate on the succession to Charles II in 1679. His natural heir and brother, James, Duke of York, was a devout follower of the Catholic church, which made him unacceptable to a largely Protestant people. Those who insisted that he should become King, despite his religious views, were called by their opponents 'Tories.' In return they accused their opponents of being 'Whigs,' or Scottish bandits (see *Whig*).

For the seventeenth and eighteenth centuries 'party' is too formal to describe the loosely allied groups and factions based around particular individuals. The Tories were those who favored the Crown, the Church of England ('the Tory party at prayer'), and rural or 'Country' interests, and largely drawn from the local squirearchy of small landowners, in opposition to the aristocratic Whigs. The first identifiably Tory thinker was Henry Bolingbroke. In *The Idea of a Patriot King* (1738), he articulated several recurring Tory themes: a reliance on experience against abstract theory; the significance of social groups such as the family; a constitution based on historical liberties; a suspicion of power; and the need to abandon unsustainable positions that undermine stability, in this case sympathy for the Jacobite cause.

2. *Tory Principles Within Conservatism*

W. H. Greenleaf (1983), in his magisterial study of the British ideological tradition, identified a tension between libertarian and collectivist thinking within all parties. The collectivist element within conservatism is described as Tory paternalism, sympathetic to state intervention in the economy and the welfare state, and associated with humanitarianism, duty, piecemeal concessions, and social and economic reforms. His terminology of dualism exaggerates these differences into deep and irreconcilable divisions.

The most extensive attempt to distinguish between Tory and Whig within conservatism is to be found in Norton and Aughey (1981). They contrast Tory themes of traditionalism, original sin, order, and authority, with Whig themes of capitalism, rationalism, faith in human progress, and the creation of wealth. Toryism is associated with security and stability, concern for the moral and economic well-being of the people, trust, instinct, honor, loyalty, authority, discipline, prejudices, traditions, customs, and obligations. 'Toryism is sceptical, with a touch of fatalism' (Norton and Aughey 1981, pp. 64–5). The problem with this distinction is that Whiggery is given a solely economic orientation, which ignores its perspectives on other issues, while Toryism is left bereft of an economic position.

Norton and Aughey identify four subdivisions within Toryism. Pessimistic or fatalistic Tories accept change as inevitable though undesirable. Paternalists are motivated by a strong sense of *noblesse oblige*, of responsibilities and duties toward others. Progressives seek to update paternalism and demonstrate its contemporary relevance. Combative or reactionary Toryism is opposed to any attempt to undermine the values of society. These categories are primarily ones of style rather than substance.

3. *Toryism in the History of the Conservative Party*

When did the Tory party become Conservative? How has the Tory tendency expressed itself within Con-

servative party politics? The emergence of British parties is usually associated with the divisive reaction to the French Revolution, between those who feared and decried it around William Pitt the Younger, and those who hailed it around Charles Fox. The former were known as Tories, although Pitt always described himself as a Whig.

The founder of modern conservatism, Edmund Burke, was also a Whig, of the Rockingham persuasion. He was the chief critic of the French Revolution, and provided intellectual sustenance to the Pittites. Burke combined Tory and Whig elements into conservative philosophy, the scepticism of the Tory David Hume and the economic liberalism of the Whig Adam Smith. The chief Tory features of his thought were: the limits of reason, the inherent weakness of human nature, the wisdom of institutions, the defence of private property, the necessity of reform for stability, and the protection of the social order. The term Conservative was first used to describe the party in January 1830 by the *Quarterly Review*. The modern Conservative party is usually dated from the 1832 election and the Tamworth Manifesto written by Robert Peel.

Peel came under increasing criticism from a group within his party over his support for religious toleration for Catholics and free trade. While predominantly based on the landed aristocracy, its spokesman was Benjamin Disraeli, probably the first writer to identify distinctively Tory thinking within conservatism. In *Sybil* (1845), he declared 'Toryism will yet rise from the tomb … to bring back strength to the Crown, liberty to the subject, and to announce that power has only one duty—to secure the social welfare of the people' (Blake 1985, p.5). The Conservative party broke over free trade and the repeal of the Corn Laws in 1846, which Peel felt was essential to address the famine in Ireland. In conservative historiography, Peel is the epitome of the Whig tendency of economic rationality and Disraeli of romantic Toryism.

Disraeli proclaimed the principles of the Conservative party as the upholding of established institutions, the maintenance of the empire, and the elevation of the condition of the people. The emphasis on the latter, through social legislation aimed at benefiting the working class, absorbed the Tory Radicals, such as Robert Southey and the poet Samuel Coleridge, who favored an alliance of the aristocracy and the working class against the commercial middle class. This strand of social reform can be traced through Tory Democracy, led by Lord Randolph Churchill, to Prime Minister Stanley Baldwin, and to the contemporary One Nation Tories. In the 1940s, R. A. Butler and the One Nation Group articulated a progressive Toryism, accepting an extensive role for the state. This found expression during the premiership in the 1950s of Harold Macmillan, who earlier had advocated a Middle Way between capitalism and communism. Edward Heath belonged to this tradition of progressive Toryism. The reaction to it led to Margaret Thatcher, who combined Whig and High Tory elements together in the New Right, despite the tensions between them. Two groups of Tory critics, High and One Nation, emerged against the Whig dominance within the New Right, who represent the current state of Toryism.

4. *Contemporary Toryism*

The leading intellectual exponent of High Toryism is Roger Scruton, editor of the *Salisbury Review*. He identifies the features of High Toryism as: legitimacy based on custom, history, popular morality, religion, and patriotism; the value of a ruling class; the validity of private property; suspicion of unbridled capitalism; and a belief in one nation of different classes held together by mutual obligations. He protests that Toryism has been undermined by commitments to democracy, liberalism, and the market economy. He uses the unusual term 'low Toryism' of business and free enterprise, and deplores its influence over the Conservative party (*A Dictionary of Political Thought*, Macmillan, London, 1983, p. 466). Critics see a strong Hegelian and continental influence in his writing, in contrast to Burkean conservatism. The *Salisbury Review* is the main mouthpiece for High Tories, with its fascination with the Established Church, fox hunting, rural life, and the nation-state. With the exception of the issue of the European Union, their problem is the appearance of irrelevance to the concerns of the average person.

The other school is the One Nation Tories, consciously echoing Disraeli. They criticize Thatcherism for promoting economic reform at the expense of a divided nation, and surrendering the tradition of social legislation. They were dismissed by Mrs Thatcher as 'wets.' Their most prominent intellectual was Ian Gilmour who wrote a series of withering attacks on Thatcherism, which he portrayed as abandoning traditional Tory scepticism towards ideology. He became ever more virulent towards Thatcher for dancing with dogma and her successors for drifting with dogma. The weakness of One Nation Tories is the lack of a coherent program, with their most distinctive policy, European integration, clashing with their own terminology of One Nation.

5. *Toryism in the English-speaking Diaspora*

Those colonists who opposed the American Revolution against the Crown were known as Tories, and most were forced to migrate to English Canada. Toryism is associated in the USA with monarchy, aristocracy, class, and hostility to change, ideas incompatible with American values of individualism,

freedom, progress, and natural rights, embodied in the Declaration of Independence and the Constitution. It is viewed as fundamentally alien to the liberal society that is America. However, Russell Kirk provided numerous favorable references to the Tory tradition in *The Conservative Mind* (1952), which invigorated the conservative movement in the 1950s. He happily embraced the term Bohemian Tory. His description of conservatism has been attacked as more Anglo than American, and fails to convey the distinctive nature of American conservatism from its British cousin. George F. Will defends Tory notions of scepticism towards capitalism and an affirmative role for the state in promoting virtue. Beyond the UK, Tory can be found in extensive use only in Canada, as a synonym for the Progressive Conservative party, more collectivist than its British counterpart and close to One Nation Toryism. Historically Toryism was associated with strong links to Britain, Crown, and Empire.

6. *The Future of Toryism: Three Scenarios*

First, Toryism could disappear as an influential political voice. Outside of its use as a synonym, Tory has an old-fashioned flavor. Tory themes of hierarchy, elitism, authority, and deference appear out of step with the Zeitgeist of the twenty-first century. Toryism has always suffered from several weaknesses: uncertainty as to whether it is an ideology; problems in turning instincts into a political program; difficulty in clearly articulating its principles; a tendency to narrow nationalism; and a lack of appeal to the young who believe in the capacity to improve the world. Its deepest flaw, as Hayek (1962) notes, is its inability to provide an alternative direction; it can only slow the process of change, it cannot direct it. In this scenario, Toryism will be seen as an obsolete relic of the past.

A second scenario is that, as conservative parties surrender to neo-liberalism and neglect their concern for order, Tories will seek alliances with the left, especially the green movement, which opposes the destructive trends of capitalism, as the only genuinely conservative, if not reactionary, force in modern politics. This is the conclusion of John Gray in his numerous jeremiads against neo-liberalism. The third scenario is that Toryism will once again become the dominant thinking amongst conservatives and that Whiggery will be recognized as an alien import that should be expunged. This is the aspiration of both High and One Nation Tories, despite their distinctive political goals.

See also: Aristocracy/Nobility/Gentry, History of; Conservatism: Theory and Contemporary Political Ideology

Bibliography

Beer S H 1965 *Modern British Politics*. Methuen, London
Blake R 1985 *The Conservative Party from Peel to Thatcher*, rev. edn. Fontana, London
Butler G 1957 *The Tory Tradition*. Conservative Political Centre, London
Gilmour I 1977 *Inside Right: A Study in Conservatism*. Hutchinson, London
Greenleaf W H 1983 *The British Political Tradition 2: The Ideological Heritage*. Methuen, London
Hayek F A von 1962 *The Constitution of Liberty*. University of Chicago Press, Chicago
Kirk R 1986 *The Conservative Mind*. Regnery, Washington, DC
Norton P, Aughey A 1981 *Conservatives and Conservatism*. Temple Smith, London
Scruton R 1984 *The Meaning of Conservatism*. Macmillan, London

N. Ashford

Totalitarianism

Totalitarianism was a major concept of the twentieth century. Yet both its meaning and its application were contested. While most political scientists believed that the concept summarised well the key features of both Fascism (inc. Nazism) and Communism, historians such as Alan Bullock argued that the difference between Nazi Germany and the Soviet Union (particularly under Stalin) were such that it was misleading to use a single term to describe both. While there unquestionably were important distinctions between Fascism and Communism, particularly in terms of the role played by ideology, the predominant view nowadays is that the term *is* a useful short-hand way of describing a range of systems that, on balance, had more commonalities than differences. As will be demonstrated, while having much in common with authoritarianism, totalitarianism is conceptually distinct.

This article begins by considering the defining features of totalitarianism, and then outlines the development of the concept in terms of its historical and intellectual contexts. Following this is a brief evaluation of the concept. The final section concerns its relevance today and into the future.

1. *Defining Features of Totalitarianism*

As suggested by the root of the word, totalitarianism refers to a system in which control of the population is greater—closer to 'total'—than in any other type of system. It is the most extreme version of the subordination of the individual to the state and to a questionable notion of the collective. Totalitarian and authoritarian dictatorships have much in common,

including a lack of respect for citizens' rights, for the public accountability of political elites, and for the rule of law. But a totalitarian dictatorship is distinguishable from an authoritarian one in several ways (see Linz 1975).

First, ideology plays a far more significant role in the former than in the latter, even though many observers agree that ideology was less important in Fascist Italy and Nazi Germany than in Stalin's USSR and other communist states. While totalitarian ideology often makes use of the past, particularly national traditions, it is basically radical and future-oriented ('chiliastic' or 'teleological'), whereas authoritarian regimes typically focus more on the past and continuity. It is also often overtly exclusionary, and arbitrarily identifies particular minority groups as 'enemies' of the state and/or the people (e.g., Jews and gays in Nazi Germany, or 'the bourgeoisie' or *kulaks* [rich peasants] in Stalin's USSR), for whom life could become literally unbearable.

Another difference is that, in theory at least, the single political party plays a more prominent role in a totalitarian regime than in an authoritarian one. Third, authoritarianism relates primarily to a political system, and there is no necessary implication that the state is involved heavily in the economy. In contrast, most theorists of totalitarianism maintain that, while communist states generally controlled the economy much more than did fascist states, the *overall* level of state interference was much higher in totalitarian states than in authoritarian. Finally but importantly, totalitarianism was seen as a peculiarly twentieth-century phenomenon. This was largely because it was not until then that the technological means for controlling and transporting populations existed. With the rapid development of the means of mass communication, notably the radio and cinema, citizens could be subject to a constant barrage of state propaganda in a manner that was not possible in earlier kinds of dictatorship. This, together with the much greater emphasis placed by totalitarian regimes on state-run education for all, meant that the state's capacity for indoctrination ('thought control') reached qualitatively new levels.

Although there have been several comparative analyses of totalitarianism, the most widely cited are those of Friedrich and Brzezinski (1956, 1965). In the earlier study, these authors identified what they considered to be the six salient features of totalitarian regimes. These were an official, all-embracing, and chiliastic ideology; a single mass party, typically led by one man (the dictator); a system of terroristic police control (high levels of apparently arbitrary coercion); a near-complete monopoly of the means of communication; a near-complete monopoly of the means of armed combat; and central control of the entire economy. In the second edition of their book, this list was supplemented to include a commitment to expansionism and the administrative control of justice. One other distinguishing feature, emphasised particularly by Arendt (1951), is a high level of citizen alienation.

2. *The Development of the Concept*

The concept of totalitarianism can be traced back to the 1920s; from 1925, Mussolini himself used it to describe the Italian fascist system, for instance. But the term was not in common usage in English until after the Second World War. As Brown (1999) points out, one of the first to alert the public to the differences between traditional authoritarian and the new totalitarian regimes was not a social scientist, but the novelist George Orwell. In *Animal Farm* (1945), Orwell focused on the hypocrisy and undemocratic nature of communism, while in *Nineteen Eight-Four* he highlighted the horrors of a society in which the notion of ends justifying means was taken to new extremes, and in which political elites could manipulate mass consciousness through high technology indoctrination. In this nightmare world, the goal of achieving a better society in the future was seen to override the concerns of ordinary individuals in the present.

Within the social sciences, it was primarily a number of scholars of German or East European origin who elaborated the concept and produced comparative analyses of its application in practice. While Arendt and Friedrich were of German origin, Brzezinski had Polish roots. These analysts, or members of their families, had often experienced the horrors of totalitarianism first hand. While fascism appeared to have been defeated by the end of the Second World War, Stalin's USSR became stronger, with much of Eastern Europe coming under direct Soviet influence for the first time. One of the objectives of many scholars of totalitarianism was thus a normative one. They wanted not only to explain what it was and how it had arisen, but also to demonstrate that it was inherently evil, and why the West should resist it. In the climate of the early Cold War between liberal democracy and communism, such normative analyses found considerable resonance and support, especially in the USA.

But the term began to lose some of its popularity by the 1960s, even in the USA (see Friedrich et al. 1969). Many now felt that it was inappropriate to use the same term to describe both Stalin's and Khrushchev's USSR, for instance, especially since the latter was by the early 1960s referring to his predecessor's crimes against the Soviet population. The term became even less fashionable in the early-1970s, when the relationship between the Soviet bloc and the West entered a more constructive phase, known as *détente* (lit. a reduction in tension). At about the same time, scholars such as Jerry Hough in the USA and Gordon Skilling in Canada began publishing articles and books, based on solid research, that suggested there might be a limited form of pluralism emerging in the Soviet Union.

But in the aftermath of the Warsaw Pact invasion of Czechoslovakia of 1968, according to Rupnik (1988), many intellectuals in what is nowadays known as Central and Eastern Europe began to use the term totalitarianism in a different way from that favored by Western (esp. American) observers. These East Europeans tended to focus less on political institutions, and more on the political use of language—not merely to indoctrinate citizens, but to limit the scope of discussion (discourse) and hence the parameters within which thought could roam. They also criticized the methods used by totalitarian—here, basically meaning communist—regimes for suppressing knowledge of the past that did not suit those regimes. Totalitarianism was, thus, seen to be intent on destroying historical memories, and hence culture, that did not suit the rulers.

3. *The Usefulness of the Concept*

As indicated, totalitarianism has been criticized for over-generalization about regimes that some analysts consider too diverse to be grouped together under one heading. To Italy and Germany, for instance, could be added several other countries that were fascist or quasi-fascist at some point during the 1930s and 1940s, including Croatia, Hungary, Japan, and Romania. Among communist states, there were substantial differences between, for instance, the *relatively* liberal and outward-looking Yugoslavia or Hungary and the far more orthodox and isolationist Albania or North Korea. The point also pertains in some cases to significant changes within one country over time. Is the concept, therefore, too broad to be useful?

There is considerable variety among systems usually labeled 'liberal democracies,' yet few would advocate abandoning the term; most accept that there is more in common between these systems than there are differences. Much the same can be said of totalitarian systems. The term is particularly useful if it is used as an ideal type, rather than as a model. The former is taken here to mean a pure form of an idea, against which actual regimes are measured. While the concept of a model has diverse meanings, it refers here to an attempt to encapsulate the salient features of a given system or type of system in a structured, schematic way. If totalitarianism is used as an ideal type, it is possible to say that a given country is more or less totalitarian than another, or more totalitarian at one time than at another. Using totalitarianism in this way overcomes one of the most frequent criticisms of the concept, namely that it is statically descriptive and lacks any internal dynamism.

4. *Totalitarianism Today and Tomorrow*

The concept of totalitarianism has experienced a limited resurgence of popularity since the collapse of communist power in 1989–91, especially among Central and East Europeans. When the communists were in power in that region, generally it was forbidden to suggest publicly that they were totalitarian dictatorships. Indeed, most of these countries officially claimed to be 'socialist' democracies. Since the early-1990s—and in some more liberal communist states even before that—many critics of the old regime regularly have referred to their countries' pasts (until the late-1980s) as totalitarian. Although this usage of the word is often loose and little more than a term of abuse, it has led some scholarly analysts to argue that the concept may have been abandoned prematurely during the Cold War (see Siegel 1998), and that, liberalization phases and certain changes notwithstanding, most communist states remained fundamentally totalitarian until their collapse.

While there are no overtly fascist regimes at present (2000), there remain five communist ones (PR China, Cuba, DPR Korea, Laos, Vietnam), and these are sometimes still described as totalitarian. Moreover, while it might be true that globalization and the Internet are rendering it ever more difficult for individual governments to control their populations in ways they sometimes did in the twentieth century, it is possible that governments will catch up with the technological revolution and harness it to new totalitarian objectives. The rise of the extreme right (neo-fascism) in several European countries—including Austria, France, and Germany—from the 1990s means that it would be ingenuous to consign totalitarianism to the rubbish bin of history. It may yet rear its ugly head again.

See also: Arendt, Hannah (1906–75); Authoritarianism; China: Sociocultural Aspects; Communism; Communist Parties; Ideology: Political Aspects; National Socialism and Fascism; Parties/Movements: Extreme Right; Socialist Societies: Anthropological Aspects; Southeast Asian Studies: Politics; Soviet Studies: Politics

Bibliography

Arendt H 1951 *The Origins of Totalitarianism*, 1st edn. Harcourt Brace Jovanovich, New York

Brown A 1999 The study of Totalitarianism and Authoritarianism. In: Barry B M, Brown A, Hayward J E S (eds.) *The British Study of Politics in the Twentieth Century*. Oxford University Press, Oxford, UK

Friedrich C J, Brzezinski Z 1956, 1965 *Totalitarian Dictatorship and Autocracy*, 1st and 2nd edn. rev. Harvard, Cambridge, MA

Friedrich C J, Curtis M, Barber B 1969 *Totalitarianism in Perspective: Three Views*. Pall Mall Press, London

Linz J J 1975 Totalitarian and Authoritarian Regimes. In: Greenstein F I, Polsby N W (eds.) *Handbook of Political Science*, Vol. 3. Addison-Wesley, Reading, MA

Rupnik J 1988 Totalitarianism revisited. In: Keane J (ed.) *Civil Society and the State*. Verso, London
Schapiro L 1972 *Totalitarianism*. Pall Mall Press, London
Siegel A (ed.) 1998 *The Totalitarian Paradigm after the End of Communism*. Rodopi, Amsterdam

L. Holmes

Totalitarianism: Impact on Social Thought

If it is right to say that the meaning of a concept is at least in part determined by its use, then we can find in the career of the concept of totalitarianism many distinct meanings. They include:

(a) its directly political use in the 1930s for or against the establishment of the Nazi and Stalinist regimes;

(b) its conventional use within postwar political science and political philosophy to define and denounce these regimes from the point of view of liberal democracy;

(c) its more radical use within critical theory to trace the linkages between what happened at Auschwitz and in the Gulag and the 'normal' forms of modern political life;

(d) its rejection, especially by Marxists, as an inherently ideological concept which falsely equates communism with fascism; and

(e) its reappropriation by postmodern social theorists as the prevailing potentiality of political modernity.

1. *Political Origins*

The Italian idealist philosopher, Giovanni Gentile, coined the term 'totalitarianism' in the 1920s to express what he saw as the actuality of 'total freedom' and 'total domination' in which the self-realization of the individual is identified with the universality of the state and when the state itself is 'comprehensive, all embracing, pervasive … total' (Bellamy 1988). Gentile (1961) presented his theory as 'Hegelian' but he criticized Hegel for having an *a priori* view of individual right and morality and argued that the rationalistic conception of atomized individuals still present in Hegel's philosophy of right had to be overcome. He insisted that *society* should be seen as the (one) true subject of right and that our obligation to the *state* is the product of our own will. He argued that for the individual the law is identified with the 'very act of his self-realization' and cannot therefore be opposed to force. When the law constrains us, we are 'forced to be free.' My will is subordinate to the state to the extent that it is identical to it: 'I want what the law wants me to want.'

In the 1930s the concept of totalitarianism was also used in a critical sense by political writers who sought to compare the regimes that ruled over Stalinist Russia and Nazi Germany or who believed that the whole world might be going down this same path. Leon Trotsky, for example, argued that Stalin's regime was like Hitler's except for the 'greater barbarism' of the former. He faced up to the conjunction of socialism *and* barbarism in a way that was largely absent in a Marxist tradition which either denied the barbarism of Soviet socialism or the socialism of Soviet barbarism (Trotsky 1937).

2. *Political Science and Political Philosophy*

After World War II the concept of totalitarianism was used in conventional political science to denote an antiliberal political system which annihilates all boundaries between the state, civil society and individual personality. Totalitarianism was meant to be a distinct type of modern dictatorship which manifests the following characteristics: state power is concentrated into a single party; control is exercised over all areas of social life; terror is employed by the secret police; the mass of the people are mobilized behind the regime; and an official and irrefutable ideology is disseminated which aims either at the construction of a new historical order or at the recovery of a lost order of nature. During the Cold War the concept was often used in an openly polemical sense to indict twentieth-century Marxism and to contrast 'authoritarian' regimes on the right with 'totalitarian' regimes on the left on the ground that only the former were open to democratization. Perhaps the key to the conventional use of the concept is that it leaves liberal ideas, practices, institutions and forms of government intact as the innocent, unfortunate victim of malevolent external forces. No *intrinsic* relation between liberal modernity and the origins of totalitarianism is posited or permitted. Among those political scientists who wrote more or less in this vein were Carl Friedrich and Zbigniew Brzezinski (1967) and Leonard Schapiro (1972).

Gentile's neo-Hegelian idealism misconstrued what Hegel was doing in the *Philosophy of Right*. Textual support for his reading of the text was found in its depiction of the state as 'the divine idea as it exists on earth' and its declaration that the existence of individuals is a 'matter of indifference' compared with the ethical life of the state. However, when Hegel characterized the concept of the state as the march of God on earth, it was in order to display the vanity of its concept and the untruth of 'conceptual thinking' which hives off the concept of the state from its actual social existence. While Hegel insisted that the idea of the state was the unity of its concept and existence, Gentile, by contrast, turned the concept of the state as

an ideal to be actualized. He was not alone in reading Hegel in this way, and from another point of view, this interpretation also encouraged many liberal political philosophers to depict Hegel as a key intellectual forerunner of totalitarianism.

After World War I, L. T. Hobhouse (1918) set the scene in his *Metaphysical Theory of the State* when he described Hegelianism as a 'false and wicked doctrine' which uses so-called 'dialectical logic' to convert the freedom of the individual into the freedom of the state against the individual. After World War II, with the experience of totalitarianism before its eyes, 'English' liberalism became ever more impassioned in its hostility to Hegel's 'doctrine of the state.' Bertrand Russell, Ernst Cassirer, John Plamenatz, Isaiah Berlin and most famously Karl Popper, wrote to the effect that Hegel's political philosophy not only justified every internal tyranny and external aggression, since 'the state is everything and the individual is nothing,' but was the crucial link between old Platonism and modern totalitarianism. It was Hegel's *inversion* of freedom into its opposite which appeared as his greatest offence. There was a *prima facie* case that Hegel's philosophy of right prepared the way for totalitarian movements, but this reading was one-sided and excluded Hegel's repeated insistence on the centrality of the right of subjective freedom in the modern age.

3. *Critical Theory*

A more critical appropriation of the concept of totalitarianism was defined first by its acknowledgement of the implication of liberalism in the origins of totalitarianism. The basic message was that when liberalism presents totalitarianism as the Other of itself, it can understand neither totalitarianism nor itself. Second, it was defined by its challenge to the empiricism of conventional political science. The argument was that 'totalitarianism' should be treated neither as a fixed analytical category nor as a *type* of political system which can be defined according to the presence of certain specifiable criteria, but more dynamically as a movement of thought and practice which has deep roots in the modern political age. Third, critical theories of totalitarianism dismissed the notion of any *direct* connection between the *event* of totalitarianism, that is, the fact that in the twentieth century totalitarian movements came into being and achieved power, and any of the great nineteenth-century thinkers (not just Hegel and Marx but also Kierkegaard and Nietzsche or more recently Weber and Durkheim) who were both heirs to and rebels against the tradition of 'Western' political thought. Among the social theorists who used the concept of totalitarianism in this way were the critical theorists of the Frankfurt School; former Trotskyists/left socialists like Cornelius Castoriadis, Claude Lefort and George Orwell; existentialists such as Albert Camus and Hannah Arendt; and individual thinkers such as J. L. Talmon who perhaps bridged the conventional and radical paradigms.

Against the charges leveled against Hegel and Marx by 'English' liberals, 'German' critical theory stressed the incompatibility of Hegelianism and totalitarianism. It was stressed that the aim of Hegel's political philosophy was not to subordinate the individual to the state but to harmonize the principles of political community derived from the ancient *polis* with principles of individual freedom derived from modern Christianity (Löwith 1967). Franz Neumann argued that Hegel and totalitarianism made impossible bedfellows and that this is why Hegel's doctrine of the state supremacy was abandoned in Germany: 'Today, the doctrines exalting the state, notably Hegelianism, have been thrown overboard ... Hegel's idea of the state is basically incompatible with the German racial myth ... Hegel's theory is rational; it stands also for the free individual.' (Neumann 1942). For the Nazis, by contrast, the state represented not an end in itself but a means for their movement and the idea of the state was subsumed to the authority of the 'true community' bound together by blood and soil and subject to no rational norms. Totalitarian ideologues attacked Hegel's philosophy of right, as Gentile had done, precisely because individual right remained the basis of his rational state.

Critical theory was united, however, in arguing that Hegel failed in his attempt to reconcile individual freedom and political community because of his incapacity to address the social questions that were determining the future of bourgeois society: how to control the poverty brought about by wealth, the increasing claims of the masses who sought to rule by force of numbers, the collision with liberalism (Löwith 1967, p. 241). Löwith concluded that Hegel was, like many philosophers, politically naive. Marcuse and Neumann argued that Hegel's philosophy contained an inexcusable paean to the Prussian state, 'the state of broken promises, of disappointed hopes, a state which cared nothing for free institutions,' and expressed the authoritarian inclinations of liberalism itself. In their eyes Hegel inadvertently revealed something fundamental about the origins of totalitarianism: namely that the only solution *within existing conditions* to the antagonisms of civil society was to turn the state into an ever more 'independent and autonomous power.' From this standpoint it seemed that liberalism was fully implicated in the origins of totalitarianism and that the only alternative was to foster new forms of individualism and political association which transcend this given historical reality. This meant facing up to the 'universal negativity' of bourgeois society and advancing an 'affirmative materialism' which privileges 'the idea of happiness and material satisfaction' over the system of right (Marcuse 1979).

The full force of critical theory's criticism of the totalitarian tendencies present within both Hegel's

philosophy of right and modern liberalism was to be found in the work of Theodor Adorno. In his *Negative Dialectics* (Adorno 1973) he writes of how Hegel turns the state into an object of worship, degrades individual subjects into its mere executors, dissolves the everyday experience of alien and oppressive power from an allegedly higher philosophical vantage point, associates 'fault-finding' with inferior consciousness, disguises decisions of state as democratic procedures emanating from the will of the people, identifies the rational individual with obedience to the state and holds the individual always in the wrong whenever he is 'too benighted to recognize his own interest in the objective legal norm.' According to Adorno, the world which the *Philosophy of Right* sanctified contains 'an endless procession of bent figures chained to each other, no longer able to raise their heads under the burden of what is.' It is a world which compresses the particular 'like a torture instrument.' It was not Hegel, however, who was 'responsible' for the rise of totalitarianism; he was rather the first to admit that the universal prevails over the particular in modern states, to repudiate the individualistic illusions of liberal thought and to demonstrate that liberalism itself possesses totalitarian potentialities.

4. *Hannah Arendt*

Hannah Arendt also counted the failure of liberalism to live up to its own ideals and its inability to resist the rise of totalitarian movements 'among the historical facts of our century' (*Essays in Understanding* 1998). She too insisted that to hold the thinkers of the modern age, especially the nineteenth-century rebels against tradition, responsible for the structure and conditions of the twentieth century is 'even more dangerous than it is unjust.' She argued that the implications apparent in the *event* of totalitarian domination went far beyond the most adventurous ideas of any of these thinkers and that their greatness lay in the fact that they perceived their world as one invaded by 'new problems and perplexities which our tradition of thought was unable to cope with' (*Between Past and Future* 1977). The more immediate culprits were imperialism, racism, anti-Semitism, inter-imperialist wars and the destructive nihilism of a front generation which celebrated violence and cruelty as supreme values. She praised these nineteenth-century rebels for being the first to think 'without the guidance of any authority whatsoever' and for retrieving the concept of political freedom from the buried treasures of ancient Greece, but she believed that they were still held by the categorical framework of the tradition against which they rebelled. In her view they ended up reducing 'divergent values, contradictory thoughts and conflicting authorities to a unilinear and dialectically consistent thread of historical continuity.'

Arendt was not consistent in her writings, but the main thrust of her work was not to treat totalitarianism as a *type* of political system but to understand it as a principle of movement: one that has perpetually to destroy in order to survive and offers an infinity of violence in place of any real power. The totalitarian mind dreams of the identity of total domination and total freedom but the obstacle that this modern megalomania encounters is that the world is not a nullity in relation to its own will. Arendt recognized that there is something profoundly inhuman about an idea which treats human beings *as if* their own self-consciousness was fully identified with that of the state, without conflict or contradiction, *as if* nothing is left over and there is no supplement, and *as if* the victims really are turned into mindless things and deprived of spontaneity, difference and plurality. To be sure, totalitarianism for Arendt was not 'merely an idea' but neither should it be confused with the actuality of social life. It is always an 'as if': an imaginary identity which is actualized in an unsatisfiable will to destroy.

5. *Marxist Criticism*

The use of the concept of totalitarianism was strongly criticized by Marxist scholars who refused to accept that the Soviet Union was ever 'like' Nazi Germany and objected to what they read as its confusion between two different systems of rule. They maintained that Communism held universal values, only denounced democracy because it is merely formal, endeavored at least in theory to establish a true or higher form of democracy, justified violence only if it served as a counter-violence, and had as its final aim the good of humanity. On the other hand, they held that Nazism glorified nationalist passions, sought only to realize the particular destiny of the *Volk*, attributed absolute superiority to a pure race summoned either to subjugate inferior races or to eliminate them, put anti-Semitism and hatred of others at the center of its political program, and treated violence as an expression of life itself. It was argued that Communism referred, in however distorted a fashion, to laws of history while Nazism referred only to illusory laws of nature. In the first case, we have a political order based on the overcoming of private property and capitalist exploitation; in the second, an ethnic order from which private property profited. Claude Lefort has convincingly argued, however, that these objections to the concept of totalitarianism remain on the surface of political life and fail to measure the difference between ideology and practice. They are especially vulnerable to the countercharge that what is as important as the manifest content of these rival ideologies is the comparable function they performed. Thus in both cases, the dominant ideas were tied to the existence of

a party movement whose organization and unity represented the untouchable ground on which 'the power of discourse and the discourse of power became indistinguishable' (Lefort 1998).

6. Totalitarianism and Modernity

In contemporary social theory, especially that which is sometimes dubbed 'postmodern,' the idea of totalitarianism has been expanded to refer to the tendencies within political modernity to elevate the idea of society over the individual, to subject all aspects of social life to surveillance and discipline, and to sacrifice moral choice at the altar of instrumental rationality. Modernist social theories are represented as having a mimetic relation to the society which they theorize. Marx is indicted for submitting subordinating human action to the laws of historical necessity; Durkheim for equating morality with conformity to social norms; Weber for idealizing a rational bureaucracy which limits itself to amoral choices based on criteria of efficiency. The idea that totalitarianism has roots in 'normal' society has been radicalized so that totalitarianism appears as an ever-present, ever-pressing product of 'modernity' or at least of 'high' or 'hard' modernity (Fine and Turner 2000). For Zygmunt Bauman (1990) the 'rationality' of the modern age, or at least the instrumental form of rationality which dominated both the conception and execution of the Holocaust, appears as the main connecting link between modernity and totalitarianism. He writes that this 'period' may be finished now that we have entered what he calls the age of 'liquid modernity,' and that the threat of totalitarianism has accordingly abated. There is, however, an instability in an approach which totalizes the specter of totalitarianism in one period and evaporates it in the next.

7. Conclusion

The event of totalitarianism, marked by Auschwitz and the Gulag, confronts social theory with the question of whether the phantasmagoria of a fascist consciousness should be imported into the conceptual arsenal of critical social theory. The totalitarian mind dreams of the conjunction of total domination and total freedom, but it is less the realization of this dream than its failure that draws it into mammoth displays of power (we may think of those represented cinematically by Leni Riefenstahl) and into what Arendt calls its 'escalating orgies of destruction.' In social theory we are accustomed to the argument that the way people present themselves should not be confused with who they are, but we are inclined to forget that this holds not only when people present themselves as all-good but also when they present themselves as all-powerful. An analogy might be drawn with Foucault's error when he turns the advertising slogans of Jeremy Bentham's *panopticon*—that it would be all-seeing, all-knowing, all-powerful, a perfect machine for 'grinding rogues honest,' etc.—into the actuality of disciplinary power. Foucault successfully used Bentham to devalue the claims of penal reformers who presented their innovations in the language of the 'rights of man' or as advances in humanitarian practice, but he did not apply the same skeptical consciousness to those, like Bentham himself, who claimed that they had finally discovered the means of achieving a rational, complete, pervasive, gapless and seamless power. Somehow, when it comes to power rather than to humanism, social theory is more inclined to miss the shadow that lies between the image and the act.

Bibliography

Adorno T W 1973 *Negative Dialectics*. Routledge and Kegan Paul, London

Arendt H 1951 *The Origins of Totalitarianism*. Harcourt Brace, New York

Arendt H 1977 *Between Past and Future*. Penguin, New York

Arendt H 1998 *Essays in Understanding*. Harcourt Brace, New York

Aron R 1968 *Democracy and Totalitarianism*. Weidenfeld and Nicolson, London

Bauman Z 1990 *Modernity and the Holocaust*. Polity, Oxford, UK

Bellamy R 1988 *Modern Italian Social Theory*. Polity, Cambridge, UK

Camus A 1951 *The Rebel*, trans. 1953 by A Bower. Hamish Hamilton, London

Castoriadis C 1997 *The Castoriadis Reader*, trans. and ed. by D Ames Curtis. Blackwell, Oxford, UK

Fine R, Turner C 2000 *Social Theory after the Holocaust*. Liverpool University Press, Liverpool, UK

Foucault M 1977 *Discipline and Punish*. Allen Lane, London

Friedrich C, Brzezinski Z 1967 *Totalitarian Dictatorship and Autocracy*. Praeger, New York

Gentile G 1961 *I fondamenti della filosofia del diritto*. Sansoni, Florence, Italy

Isaac J 1992 *Arendt, Camus and Modern Rebellion*. Yale University Press, New Haven, CT

Lefort C 1986 *The Political Forms of Modern Society: Bureaucracy, Democracy, Totalitarianism*. Polity, Cambridge, UK

Lefort C 1998 'The concept of totalitarianism.' *Papers in Social Theory*. Warwick University: Social Theory Centre, UK

Löwith K 1967 *From Hegel to Nietzsche*. Constable, London

Marcuse H 1979 (first published 1941) *Reason and Revolution: Hegel and the Rise of Social Theory*. Beacon Press, Boston

Neumann F 1942 *Behemoth: The Structure and Practice of National Socialism*. Gollancz, London

Popper K 1945 *The Open Society and Its Enemies*. Routledge, London

Schapiro L 1972 *Totalitarianism*. Macmillan, London

Talmon J L 1986 *The Origins of Totalitarian Democracy*. Penguin, Harmondsworth, UK

Trotsky L 1937 *The Revolution Betrayed*. Faber, London

R. Fine

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Totemism

There is patently a history of the study of totemism: the eminent French scholar Claude Lévi-Strauss provides good chunks of it in *Totemism* (Lévi-Strauss 1963). Considerably less clear is whether there is any one thing to study here, and in the present scholarly age of 'deconstruction' this is an especially salient issue. Lévi-Strauss says there is, most notably in his subsequent *The Savage Mind* (Lévi-Strauss 1966), but nowhere does he say how much his grapplings with the problem are indebted to those of Franz Boas and his early students, particularly Alexander Goldenweiser, in the second, third, and fourth decades of the twentieth century (Shapiro 1991). This article draws heavily on Goldenweiser's thinking, supplemented here and there by that of Lévi-Strauss. Hence the following definition: totemism is a scheme of thought in which some sort of relationship is posited between classes of people and classes of objectifications. This is deliberately vague, and inconsistent with textbook definitions in three ways: (a) it does not specify the nature of the relationship; (b) it does not insist that the classes of people be situated in spatially bounded 'groups'—the sorts of collectivities with which anthropologists have been most comfortable; and (c) it does not insist that the classes of objectifications be animals, plants, or other natural phenomena.

This last feature was crucial in early theories of totemism, and it nowadays feeds readily into the primitivist assertion of an alleged 'kinship with nature' in the Tribal World. The pervasiveness of this kinship has frequently been overstated, and its metaphorical character—where it *does* exist—unrecognized: e.g., it is neither flattering to some Aboriginal Australians, nor is it literally true, that they consider themselves to be descended from eaglehawks or crows. Moreover, such romantic interpretations of totemic systems are caught short when Aboriginal social classes claim mystic affiliation not only with segments of nature but also with Indonesian ship-masts, the Christian cross, and diarrhea (Keen 1994; Worsley 1967). Still further, recent work in Central Australia, the *locus classicus* of much theorizing on totemism, indicates that the archetypical Creative (Dreaming) beings, even when they do have animal features, blend them with human ones, or change from one to the other (Morton 1987). And finally, there is often a decided chasm posited between animal-like Dreaming beings and their current zoological manifestations. The conclusion here is that 'kinship with nature' and totemism have some overlap but do not co-occur with such frequency as to be considered aspects of the same phenomenon. It will be argued below that, quite to the contrary, they are more often antagonistic.

Similar qualifications apply to the alleged link between local groups and totemism. A clear example—and one much closer to home—is the system of Zodiac signs, which classifies people by 'location' in a calendrical scheme, i.e., in a system of temporal classification. What Aboriginal Australian totemism does, by contrast, is to classify people according to systems of *spatial* classification, and this has mistakenly been rendered by much of the anthropology of the last half-century as a form of residential grouping. Such iconic use of space is of course widespread, and it occurs in traditions which no one would wish to call 'totemic.' Pilgrimage sites in Christianity, Hinduism, and the other Great Religions are exemplary.

This leaves the relationship between classes of people and classes of objectifications. The key—and here the argument follows Lévi-Strauss following the Boasian scholars—is the relationship of analogy, such that

> people A: objectification A(1, 2, 3 …):: people B: objectification B(1, 2, 3 …):: people C: objectification C(1, 2, 3 …) …

Here a distinction can be drawn between those schemes in which the number of peopled segments is clearly subject to historical contingency, such as Aboriginal Australian clan systems or American sports leagues, and those which deny the contingency and keep constant the number of peopled segments, such as the Zodiac system, or the clan system of the Bororo of Central Brazil, wherein a division of the population into eight such units is desperately maintained (Crocker 1985).

Still, there is something lacking in this analysis, for none of the systems noted seems to be entirely satisfied with positing analogy between segments: however imperfectly, they all assert a relationship *within segments*—a bond of *consubstantiality*. But this bond, though like a kinship one, can also be an anti-kinship one, i.e., it can be derived from common participation, real or imagined, in an entity or activity other than sexual generation: location of conception, quickening, birth, or initiation in Aboriginal Australia (Tonkinson 1978), time of birth in the Zodiac system, naming among the Bororo, birth from hermaphroditic Creative Beings in the Northwest Amazon (Hugh-Jones 1979), and real or vicarious co-participation in controlled violence in sports. These are all forms of what has been called *pseudo-procreation* (Shapiro and Linke 1996), and they are by no means limited to totemic systems.

Hence a rethinking is called for of a subject once dear to the hearts of kinship theorists, viz. 'clan totemism.' Objectifications in such systems are not simply emblematic, as Lévi-Strauss would have it, nor is the 'common descent' of clan members even 'assumed,' in some sexual sense. If anything, it is *denied*, at least in the Bororo and some Aboriginal Australian cases (Shapiro and Linke 1996), where, as elsewhere, what *is* asserted is a 'mystical' bond between clanspeople. Thus the suggestion that these units be renamed 'pseudo-procreative corporations' (Shapiro 1988). They—at least many of them—are not 'descent

groups,' as received wisdom would have it, even where membership is based on the mother/child tie ('matrifiliation,' in anthropological parlance), or on the father/child bond ('patrifiliation'). The resultant 'descent' through 'lines' of women or of men, respectively, may prove to be more often in the heads of anthropologists than in those of the people they have worked with; and the key conceptualized common relationship among clanspeople may well turn out to be their shared consubstantial relationship with their totem or totems rather than their shared 'descent.' Christians who find this odd need to be reminded that a comparable relationship of asexual consubstantiality is the cornerstone of their faith, and that similar 'second-birth' ideologies pervade the world's religions (Bloch and Guggenheim 1981).

At stake here is another chunk of primitivist assertion whereby the Tribal World is a universe of libertines, which stands in stark contrast to Western puritanism. From this fresh perspective, totemism—particularly clan totemism—is a means by which human beings deny what Ernest Becker (1973) calls their 'creatureliness,' their embeddedness in sexualized nature. It thus points *not* to a link with nature, as was once thought, but in the opposite direction, i.e., it is an attempt to dissociate from it.

All this calls into question materialist theories of clanship and clan totemism, such as those put forward by Lowie (1920) and many other scholars. It is true, as these theories contend, that clanship tends to occur not among the simplest hunter–gatherer populations but among more sophisticated ones (Woodburn 1982), and those who have crossed the 'neolithic' threshold and taken up farming. But there are numerous instances, probably most common in Amazonia, of farmers without clans. Moreover, the (supposedly) technologically-generated residential alignments—men hunting or raising animals and related through male links, women cultivating gardens and related through female links—of incipient clanspeople are often absent, or as among the Bororo, part of the 'superstructure' of clanship. Correspondingly, the alleged 'corporate' character of clans in domains other than ritual has been shown to be illusory.

Finally, the time-honored theory that a certain kind of kin-classification system—the one, following Murdock (1949), introductory textbook writers label 'bifurcate merging'—is congruent with clanship is in fact nothing of the sort. Advocates of this theory like to point to certain restricted areas of genealogical space, e.g.,

$$F = FB \neq MB$$

To put this more prosaically: unlike English, which lumps both paternal and maternal uncles under a single term ('uncle') and separates out one's genitor ('father'), many languages in the Tribal World classify the paternal uncle (FB) with the genitor (F) and distinguish both from the maternal uncle (MB). Now this much is consistent with a clan system: with 'patrilineal' clans my genitor and his brother are members of my own unit, whereas—assuming, as is usually the case, clan outmarriage or 'exogamy'—my maternal uncle is a member of another unit. And with 'matrilineal' clans, the structure is the same but its composition is reversed: I am in the same unit as my mother and her brother, whereas my father and his brother are in another unit.

There are so many problems with this interpretation that the wonder is that it has survived for so long. For one thing, a very large number of bifurcate merging systems exist in the absence of clans. For another, in many such systems, kin-terms are extended beyond the sphere of close relatives—indeed, sometimes they are extended to anyone an individual may encounter—and the principles of extension, now well-known, usually have nothing at all to do with clan affiliation even where such units exist. But even in the sphere of close kin the 'clan theory' of bifurcate merging breaks down: as in English, a man's children are usually equated with a woman's, paternal grandparents with maternal grandparents, and other cross-clan equations exist. Finally, and with the greatest theoretical import, it has been repeatedly shown that, in bifurcate merging and other systems of kin-classification, the core or 'prototypical' members of such classes are the genealogically closest kin (Shapiro 1995). Hence the 'clanship' interpretation of bifurcate merging is without foundation, and the supposed causal chain—from farming (or 'advanced' hunting–gathering) to certain residential alignments of kin to clan totemism to bifurcate merging kin-classification—is called into serious question. Rather, genealogical space in 'clan' populations seems to have much the same structure as such space elsewhere. We are left with a theory which views clan organization not as the outcome of 'material conditions of existence' but as systems of existential categories. And this leads again to an emphasis on a consubstantiality which is asexual. Clan totemism, like other forms of totemism, is not an epiphenomenon to a way of making a living but a way of living which, in a manner typical of our species, grapples with our uncomfortable sense of ourselves as creatures of biology.

See also: Clan; Freud, Sigmund (1856–1939); Taboo; Tribe

Bibliography

Becker E 1973 *The Denial of Death*. Free Press, New York

Bloch M, Guggenheim S 1981 Compadrazgo, baptism, and the symbolism of a second birth. *Man* **16**: 376–86

Crocker J C 1985 *Vital Souls: Bororo Cosmology, Natural Symbolism, and Shamanism*. University of Arizona Press, Tucson, AZ

Hugh-Jones C 1979 *From the Milk River: Spatial and Temporal Processes in Northwest Amazonia*. Cambridge University Press, Cambridge, UK
Keen I 1994 *Knowledge and Secrecy in an Aboriginal Religion*. Oxford University Press, Oxford, UK
Lévi-Strauss C 1963 *Totemism*. Beacon, Boston
Lévi-Strauss C 1966 *The Savage Mind*. University of Chicago Press, Chicago
Lowie R H 1920 *Primitive Society*. Boni and Liveright, New York
Morton J 1987 The effectiveness of totemism: 'Increase ritual' and resource control in Central Australia. *Man* **22**: 453–74
Murdock G P 1949 *Social Structure*. Macmillan, New York
Shapiro W 1988 Ritual kinship, ritual incorporation, and the denial of death. *Man* **23**: 275–97
Shapiro W 1991 Claude Lévi-Strauss meets Alexander Goldenweiser: Boasian anthropology and the study of totemism. *American Anthropologist* **93**: 599–610
Shapiro W 1995 Fuzziness, structure-dependency, and 'structural anthropology'. *Journal of the Anthropological Society of Oxford* **26**: 197–214
Shapiro W, Linke U (eds.) 1996 *Denying Biology: Essays on Gender and Pseudo-Procreation*. University Press of America, Lanham, MD
Tonkinson R 1978 *The Mardudjara Aborigines: Living the Dream in Australia's Desert*. Holt, Rinehart and Winston, New York
Woodburn J 1982 Egalitarian societies. *Man* **17**: 431–51
Worsley P 1967 Groote Eylandt totemism and Le Totémisme aujourd'hui. In: Leach E (ed.) *The Structural Study of Myth and Totemism*. Tavistock Publications, London

W. Shapiro

Tourism

1. The Subject of Tourism

Mankind has traveled since ancient times, but the phenomenon of tourism is a new kind of travel born of modernity. The word 'tourism' began to be used only in the nineteenth century due to the development of modern transportation technology on the one hand, and the socioeconomic changes caused by the industrial revolution on the other. Before that travel was hard and used to require patience and effort. In fact, the English word 'travel' is related to the French word 'travail', meaning 'labor.'

In the West in the second half of the twentieth century, the work ethic was transformed, and people's desires for making life worth living through leisure and enjoyment became increasingly evident. The length of the working week decreased and paid holidays were lengthened. As a result, tourism as a form of leisure activity became widespread in the developed countries of the North.

Tourism has, then, become a phenomenon of enormous importance in the contemporary world. Against this background of a massive expansion in tourism, many countries, both developed and developing, are now pouring resources into the industry, which they see as a possible option for economic development.

2. The Anthropology of Tourism

However, in the humanities and social sciences of the past, tourism was not taken seriously as a subject for research. Even among anthropologists it made its debut as a focus of research quite recently. The first academic symposium on tourism was held in conjunction with meetings of the American Anthropological Association in Mexico City in 1974. Up to then there had been a few pioneering efforts, but nothing systematic had been carried out. The results of the symposium were published in Smith's (1977) book, *Hosts and Guests*. Focusing on the relationships between the 'hosts' (the society receiving the tourists) and the 'guests' (the tourists themselves), this book was the first edited volume to deal with the anthropology of tourism. During the 1980s, tourism became established as an item on the menu of anthropology, both for research and teaching.

The central task of the anthropology of tourism is to investigate the sociocultural implications of tourism. In fact, in many parts of the world it has become obvious that tourism is playing a culturally dynamic role. Amidst the 'global cultural flows' (Appadurai 1996), social boundaries are weakening and people are enjoying the benefits of the culture which is crossing them. As this happens, expressions of traditional culture on the one hand are being progressively fragmented, while on the other culture is being reinvented, especially in the context of tourism. Therefore, MacCannell (1992) wrote that tourism was a primary ground for the production of new cultural forms on a global base.

These circumstances present a challenge to the student of culture. It is becoming impossible to treat culture in the way that it once was, as a closed system of meaning and symbols. It is also increasingly difficult to use the old-fashioned ethnographic method of recording coherent, patterned cultural systems enclosed within discrete territories and peoples. As Hannerz (1992) has suggested, one must investigate culture in relation to the macrolevel system and the 'global ecumene,' a region of persistent cultural interaction and exchange.

3. The Tourist

Smith (1977) defines a tourist in the following way: 'a temporarily leisured person who voluntarily visits a place away from home for the purpose of experiencing a change.' Here, 'work' is contrasted with 'leisure' and time is structured as follows: tourists are released from

work, this leisure time they devote to themselves, and they return to work refreshed once more. This structure of time is very similar to that of ritual as analyzed by Leach (1961) who suggested that ritual puts a stop to normal and profane time, and creates abnormal or sacred time. The end of the ritual signifies reintegration with normal and secular time. Based on this analogy with ritual, Graburn (1977) treats tourism as a 'sacred journey.' In fact, in Europe and many other parts of the world, pilgrimage, which provided the model for tourism, was simply a ritual journey.

Tourism, therefore, takes the place of ritual in the modern world. Just as rites of passage bring about a change in status or circumstances through the ritual experience, what the tourist is looking for is the experience of a change. The essence of tourism always consists of experiences that are different. As a result a trip is a success if a change can be seen in the self. People refresh themselves by traveling: they are 'reborn' with a new lease on life.

This difference in perception may also apply to tourist space in what Urry has called the 'tourist gaze.' Applying Michel Foucault's concept of the 'gaze,' Urry (1990) states that the tourist experience is 'to gaze upon or view a set of different scenes, of landscapes or townscapes which are out of the ordinary.' Just like tourist time, then, tourist space is marked and 'framed' as discussed by MacCannell. Tourism is, as MacCannell (1992) put it, an ideological framing of history, nature, and tradition; a framing that has the power to reshape culture and nature to its own needs.

4. *The Tourist Experience*

There are two important theories of the tourist experience. The first is the theory developed by MacCannell (1976) that tourism itself consists of the search for 'authenticity.' This starts from Marx's observation that living in the modern world produces a state of alienation in which we cannot realize our real selves. The reason why people travel is, therefore, to escape from the real world of the present day, to experience a different world, and to discover the 'authentic.'

On the other hand, Boorstin (1962), writing before MacCannell, observes that the authentic tourist experience is impossible in the contemporary world. Rather than being a journey in search of authenticity, travel is increasingly based on artificial images. Before setting out on a trip, tourists read the guidebook. They then go to the places written up in the guidebook, and take photographs similar to those in the guidebook. This is an experience that Boorstin has called a 'pseudo-event.' Thus, the tourist experience ends up oscillating between being a 'pursuit of authenticity' and a 'pseudo-event.' Cohen (1979), however, does not regard it as either of these two alternatives. He suggests instead that there are a number of different modes of tourist experience ranging from the recreational mode in search of mere pleasure to the existential mode in search of one's real self. Thus, there are many different types of tourism (e.g., 'historical tourism,' 'ethnic tourism,' 'recreational tourism,' and so forth), and many different types of tourists. Further, tourist experiences may vary with 'who travels, and where.'

Today, where the boundary between an original and a copy is becoming increasingly blurred in many areas, the very basis of the question of whether tourism is a search for authenticity or a pseudo-event has lost its meaning. *Cannibal Tours*, a film by O'Rourke (1987), illustrates this point. The film describes the 'ethnic tourism' to the villages on the Sepik River in Papua New Guinea. As Bruner (1989) says, the tourists already know that in New Guinea things like cannibalism no longer exist. If cannibalism did exist, or if there was no air-conditioned cruiser, they would probably not go to New Guinea. What they really want is to experience going from a safe, 'civilized' place to one that is 'primitive.' What they experience here is not something which is really primitive, but a reconstructed primitiveness. The substance of this 'postmodern tourism' is, then, the experience of a copy or *simulacrum*, and the copy becomes more important than the real thing.

Cannibal Tours also poses the critical question of how this type of tourism is based upon the issue of relations between the North and the South or 'neocolonialism.' In fact, the majority of the world's tourists are from the 'northern' countries: 57 percent from Europe and 16 percent from North America. Eighty percent of all international travelers are nationals of just 20 countries. In terms of expenditure on tourism, the world's top five spending tourist countries in 1995—Germany, the United States, Japan, the United Kingdom, and France—were responsible for 48.6 percent of the total (Mowforth and Munt 1998).

5. *The Host Society*

The anthropology of tourism investigates the interaction between hosts and guests. The central issue here has been the question of how the development of tourism influences the culture of the host society. Much of the debate along these lines has focused especially on whether its effects are beneficial or not.

On the relationship between tourism and culture, tourism has been often seen in a negative light, the general tenor of discussions being that it contributes to the destruction of traditional culture. For example, Greenwood (1977) described the case in Spain of a traditional Basque public ritual, *Alarde*, and discussed how it was commoditized and packaged for sale with

the introduction of tourism, resulting in the destruction of the irreplaceable historical heritage of the Basques.

However, this line of reasoning is not necessarily correct. First of all, tourism is not the only influence on these societies. It is just one of many ways in which the contemporary world system brings about change in the societies within it. Local cultures are also transformed by other factors, including industrialization, urbanization, migration, the mass media, pollution, poverty, and civil war, as Greenwood (1989) himself admitted in his revised version.

It is useful on this point to recall Clifford's distinction between two ways of narrating culture: metanarratives of homogenization and loss on the one hand, and of emergence and invention on the other (Clifford 1988). A narrative of homogenization becomes an 'entropic' narrative or one of loss, because modernization or Westernization means a loss of indigenous tradition. But such an analysis as the disappearance of an 'innocent' tradition that is not affected by modern Western civilization is often incorrect, because it presupposes the idea that a cultural tradition dates back to primordial times, which is actually seldom the case (cf. Hobsbawm and Ranger 1983).

In relation to the theme of creation of culture in tourist contexts, therefore, special attention should be paid to Clifford's narrative of emergence and invention. Culture in this perspective is observed as something emerging: it is not something static but a dynamic, evolving process. An example may be taken from Bali, Indonesia. In Bali, traditional culture, and particularly the traditional dances such as the famous *kecak* dance, have been elaborated and refined since the 1930s under the 'tourist gaze.' 'Traditional culture' reconstructed in this way not only has become 'symbolic capital' for developing tourism but also has been appropriated by the local people to assert their own identity. As Picard (1996) puts it, tourism has neither 'polluted' Balinese culture, nor kindled its 'renaissance,' but made the Balinese self-conscious about their culture.

Therefore, it is simplistic to consider only the 'impact' of tourism as an exogenous force, assessing its effects as either 'good' or 'bad' (Lanfant et al. 1995; Yamashita et al. 1997). What is at stake is not simply the impact of tourism on local cultures, but rather how the local cultures develop during the dynamic process of making use of tourism to redefine their own identities.

6. *Touristic Culture*

Culture in this context is not the unconscious custom with which conventional anthropology has been concerned. Rather it is an object of conscious manipulation within a broader social, economic, and political context. In other words, in the context of tourism, what is involved is not an 'authentic culture' but a newly created 'staged culture.'

Staged culture in the tourist context, or what Picard (1996) calls 'touristic culture,' including examples such as dance performances, cultural exhibitions, or souvenir production for tourists, is usually disparaged, because it is not 'authentic.' However, in the case of Bali, as was seen, it is rather the 'tourist gaze' that has stimulated creativity in their arts. What, then, is the mechanism for the emergence of touristic culture, particularly tourist art?

As an example, *Sepik* carvings, the souvenirs of *Cannibal Tours*, originally were used in ritual contexts. At the end of the nineteenth century, however, they were designated 'primitive art' by western explorers, and were seen as having objective, 'authentic' value by western scholars. Now they have been transformed into 'touristic art,' by being taken up by the tourist guidebooks to be sold to tourists. This kind of transformation occurs in the scheme of what Clifford (1988) called the 'modern art-culture system.' The *Sepik* carvings have, in this way, been transformed from cultural (ritual) to artistic objects, and then to tourist (commercialized) objects (see also Graburn 1976).

What is necessary for describing these phenomena is not an 'entropic narrative' of traditional cultures which are vanishing. Rather, it is a narrative of new cultures emerging in the context of tourism. It is in this point that tourism has become a catalyst for contemporary cultural production.

7. *Tourism and Sustainability*

The notion of sustainability became something of a buzzword from the 1980s, with the increased awareness of environmental and other problems to which particular kinds of tourism could lead. In 1989 in Zakopane, Poland, the theme of the first general meeting of the International Academy for the Study of Tourism was 'alternative forms of tourism' such as 'appropriate tourism,' 'responsible tourism,' and 'ecotourism.'

Sustainable tourism can be defined as development which will benefit the people of today without harming the resources to be shared by future generations and their prosperity. It is promotion of development which maintains a balance between the ecosystem and the society and culture concerned. According to de Kadt (1992), the sustainable development involves the following traits: ecological soundness; small-scale production; recognition of needs other than those of material consumption; equal consideration of the needs of all (including future generations); and political involvement from below.

However, human societies at present are experiencing problems that they are unable to cope with only through local knowledge. What matters most is the way in which the local people handle the process of adapting themselves to the macrosystem surrounding their own region, under the name of development.

8. *The Lesson of the Study of Tourism: Anthropology on the Move*

In reviewing research in the anthropology of tourism, Crick (1989) argues: 'Anthropology has often been defined as the study of human beings in culture and society. Tourism is thus an odd anthropological object, because international tourists are people *out* of culture in at least two senses. First, they do not belong to the culture of the destination country, and second, they have stepped beyond the bounds of ordinary social reality into what has sometimes been referred to as a 'ludic' or 'liminoid' realm.' The lifestyle of international tourists that involves living a double existence, inside and outside culture and society, as well as work and leisure, means that people are separated from their home areas, and move across international boundaries in the transnational age.

Following this, rather than tourism being an eccentric object of study for anthropologists, it is on the contrary an essential object of study if anthropologists are to change the object of their research. Anthropologists formerly used to assume that, in a given area, a given people lived with a given culture. This kind of static anthropological viewpoint cannot deal with the 'realm in between.' A new kind of anthropology that extends to lifestyles on the move must be shaped in the gaps between cultures and between societies. The anthropology of tourism presents one challenging attempt to do just this.

See also: Ecotourism; Leisure, Sociology of; Tourism, Geography of

Bibliography

Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN

Boorstin D J 1962 *The Image: Or What Happened to the American Dream*, 1st edn. Atheneum, New York

Bruner E M 1989 Of cannibals, tourists, and ethnographers. *Cultural Anthropology* **4**: 438–45

Clifford J 1988 *The Predicament of Culture: Twentieth-Century Ethnography, Literature, and Art*. Harvard University Press, Cambridge, MA

Cohen E 1979 A phenomenology of tourist experiences. *Sociology* **13**: 179–201

Crick M 1989 Representations of international tourism in the social sciences: Sun, sex, sights, savings and servility. *Annual Review of Anthropology* **18**: 307–44

de Kadt E 1992 Making the alternative sustainable: Lesson from development for tourism. In: Smith V L, Eadington W R (eds.) *Tourism Alternatives*. University of Pennsylvania Press, Philadelphia

Graburn N H H (ed.) 1976 *Ethnic and Tourist Arts: Cultural Expressions from the Fourth World*. University of California Press, Berkeley, CA

Graburn N 1977 Tourism: The sacred journey. In: Smith V (ed.) *Hosts and Guests: The Anthropology of Tourism*. University of Pennsylvania Press, Philadelphia

Greenwood D 1977 (1989) Culture by the pound: An anthropological perspective on tourism as cultural commoditization. In: Smith V (ed.) *Hosts and Guests: The Anthropology of Tourism*. University of Pennsylvania Press, Philadelphia

Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Lanfant M F, Allcock J B, Bruner E M (eds.) 1995 *International Tourism: Identity and Change*. Sage, London

Leach E R 1961 *Rethinking Anthropology*. Athlone, London

MacCannell D 1976 *The Tourist: A New Theory of the Leisure Class*. Schocken, New York

MacCannell D 1992 *Empty Meeting Grounds: The Tourist Papers*. Routledge, London and New York

Mowforth M, Munt I 1998 *Tourism and Sustainability: New Tourism in the Third World*. Routledge, London and New York

O'Rouke D 1987 *Cannibal Tours* (film). Distributed by Direct Cinema, Santa Monica, CA

Picard M 1996 *Bali: Cultural Tourism and Touristic Culture*. Archipelago Press, Singapore

Smith V L (ed.) 1977 [1989] *Hosts and Guests: The Anthropology of Tourism*. University of Pennsylvania Press, Philadelphia

Urry J 1990 *The Tourist Gaze: Leisure and Travel in Contemporary Societies*. Sage, London

Yamashita S, Din K H, Eades J S (eds.) 1997 *Tourism and Cultural Development in Asia and Oceania*. Penrbit Universiti Kebangsaan Malaysia, Bangi, Malaysia

S. Yamashita

Tourism, Geography of

Tourism geography is concerned with the place, space, and environmental aspects of tourism. While there is a distinctive field of tourism geography studies, with its own specialist journal, *Tourism Geographies: An International Journal of Tourism Space, Place and the Environment* (first published by Routledge in 1999), there are considerable overlaps with the interests of economic and cultural geographers. These have infused tourism geography with greater theoretical content in recent years, counter-balancing a more traditional emphasis on the empirical, allied to a concern for applied research. Tourism geography contributes to, as well as being influenced by, discourses on subjects such as globalization, representation, and sustainability. A review of the contribution

of tourism geography and of its links with neighboring disciplines can be found in a special theme issue of *Tourism Geographies* (2000, Vol. 2, no. 3).

1. *Definition*

An early issue in tourism geography has been the generic definition of tourism. Most definitions emphasize that tourism is characterized by nonpermanent moves, an intention to return home within a relatively short time period, and a purpose other than taking up permanent residence or employment. This gives rise to two main difficulties. First, there is no clear conceptualization of 'permanence,' and instead the pragmatic device of absence from home for at least one night but for less than 12 months is adopted both by academics and by the leading international tourism body, the World Tourism Organization. Second, the definition is chaotic in the way it bundles together diverse motivations and practices, ranging from business tourism, to education to pleasure. The relationship between tourism and recreation constitutes another problem area (Shaw and Williams 1994), for some forms of tourism activities are virtually impossible to differentiate from recreational activities undertaken by excursionists (or day visitors). In this instance, the case for a distinctive tourism geography lies not so much in differences in the content of the recreational practices but in the *implications* which follow from tourism involving overnight stays away from home.

It is the latter concern which, potentially, places tourism geography at the heart of tourism studies. While, technically, overnight stays could involve staying at a nearby hotel for a conference or a training event, or children 'sleeping over' at the homes of friends within the same town, there is an in implicit assumption that tourism involves more profound interactions between the destination and host areas. There are three main dimensions to this. First, there is an economic dimension arising from the expenditures related to the overnight stay (long distance travel, accommodation, meals, and other services) so that there are significant inter-regional transfers of income, and tourism becomes a potential economic development instrument. Second, travel brings individuals into contact with the 'other,' in the form of different cultures, with implications for both the tourists and for the receiving communities; there is a strong overlap here with the work of social anthropologists on host-guest relationships (Smith 1977). Third, tourism brings an additional dimension to the environmental impacts of human activity, resulting from the scale and intensity of travel, the redistribution of human activity across space, and a hypothesized change in attitudes to the environment amongst tourists when freed from the cultural constraints of 'home.'

2. *Changing Perspectives in Tourism Geography*

In the twentieth century, the changing approaches to tourism geography research have broadly mirrored wider shifts in other social sciences, although the applied interests of many researchers has meant that these changes have attracted relatively little methodological debate. Nevertheless, there has been a shift from a largely inductive positivist framework to approaches grounded in political economy or cultural interpretation, although tourism geography remains methodologically diverse.

2.1 *Spatial Models*

Positivist approaches have mostly been inductive, but there are some notable deductive contributions. Foremost amongst these is a relatively neglected early paper by Christaller (1963), who is otherwise best known for his work on central place theory. This paper sought to conceptualize tourism flows from 'central' to 'peripheral' locations in Europe in terms of basic principles of accessibility, and—at least indirectly—was the forerunner of an array of models concerned with the spatial interactions between areas of tourist origin and destination. Pearce (1995) has developed a fourfold typology of such models. Models of tourist travel analyze the alternative routes which link places, and the volume and character of tourist flows along these; they are often informed by a distance decay assumption. Origin-destination models are more sophisticated in that they examine the flows to and from a multiple set of areas, and identify spatial hierarchies of tourist destinations. Structural models are concerned with the evolution of distinctive spatial structures in destination areas, reflecting the asymmetrical power relationships between tourists, the tourist industry, and the destination society. These relationships are often epitomized by the development of resort enclaves which are spatially, culturally, and economically isolated from their local milieu. Finally, evolutionary models—reflecting geographers' time-space concerns—have examined the development of tourism spaces, especially resorts.

The most widely known evolutionary model is Butler's (1980) resort life cycle model, which contends that resorts pass through a series of stages: exploration, involvement, development, consolidation, and stagnation, culminating in a number of alternative trajectories, ranging from immediate decline to rejuvenation. The key factors shaping the evolution of individual resorts are visitor numbers, the roles of local/external entrepreneurs, competition, and the pressures on carrying capacity. The model is largely deterministic, but does acknowledge the role of human agency in the latter stages when alternative futures are possible, depending on the responses of the tourism industry complex. Although the model is more de-

scriptive than explanatory, it has spawned a considerable debate as to the shape of the curve, the determinants of change, and the criteria for identifying turning points between stages.

2.2 The Political Economy of Tourism

Since the 1970s, there has been a more critical approach to tourism geography which, initially at least, drew on the experiences of the less developed countries and on the political economy of development. De Kadt's (1979) '*Tourism: Passport to Development?*' argued for a more critical approach to understanding tourism impacts, but failed to offer a coherent alternative conceptual framework for this. Britton (1991), however, did set out such a framework, based on theories of capital accumulation, cultural capital, and the representation of place. This key paper developed two main streams of analysis—the commodification of culture and of place, and the particularities of tourism production within capitalist systems. These were explored through a number of case studies, such as spectacles and urban regeneration.

Britton's work did inspire a more critical approach in tourism geography, although there has been relatively limited further development of a political economy framework. One major exception to this has been the work of Ioannides and Debbage (1998), whose book sought to integrate studies of tourism production with economic geography. Their own contribution to this volume analyzed changes in tourism production in terms of a shift from Fordism to post Fordism, drawing on regulation theory. Inevitably, they conclude that there are sectoral differences between, for example: airlines and tourism retailing, and that different forms of tourism production coexist at any one time. The underlying concern of this and other research on the tourism industry has been the requirement for flexible accumulation in the face of variable demand and labor intensive production (Shaw and Williams 1994).

2.3 Cultural Interpretations

Tourism geographers have a long-standing interest in the notion of place, and in the way in which tourism behavior is both shaped by and shapes places. They have drawn extensively on the work of sociologists, such as Cohen (1984), who emphasized that all tourists are seeking some degree of novelty and difference, whilst at the same time wanting to retain an element of familiarity. The way in which tourists balance these differing needs was used by Cohen to construct a typology of tourists ranging from mass to more individualistic travelers. This, and other, typologies have been important because of the way they highlight the relationship between the different forms of tourism consumption and their impact on destination areas.

The representation of place is also of central importance in tourism behavior. The search process before booking a holiday is strongly influenced by the media and other sources of images of the destinations, both formal and informal. These are important as they condition expectations, and therefore affect the actual holiday experience and the way in which this is evaluated. The notion of the tourist gaze, which has been developed by Urry (1990, p. 2), 'presupposes a system of social activities and signs which locate the particular tourist practices.' In terms of image, the tourist gaze is constructed through signs and signifiers in the landscape, with tourists being collectors of such signs. Tourist practices do not simply entail the purchase of specific goods and services but their practices involve the consumption of signs, which are constructed around and help to define notions of place.

One of the most interesting aspects of Urry's thesis is that, in the late twentieth century, there was a change in the values held by tourists, and in the signposting of the objects to be gazed upon. The contrast between tourism as a separate sphere of social activity and the practices of every day life has greatly diminished. Tourism has become de-differentiated or merged with other activities, and this has contributed—at least in the UK—to a shift in interest from traditional seaside resorts to urban/industrial heritage sites along with valued rural landscapes and cultures. More recent cultural interpretations have been characterized by broader concerns with the embodiment of tourism experiences (Crouch 1999).

3. Tourism Geography and the Challenges of a Changing World

Tourism geographers are increasingly contributing to debates about the role of tourism in a changing world. Two themes are of particular note: the environment and economic restructuring.

In recent decades (i.e., the last decades of the twentieth century) there has been considerable research on the environmental effects of tourism, and geographers have played a key role in this. Mathieson and Wall (1982) provided an early and enduring review of the impacts of tourism. Most research, subsequently, has involved a refinement of the measurement of these impacts, through such techniques as carrying capacity, rather than advances in conceptualization *per se*. Most of the research has been in wilderness or other protected areas and there has been surprisingly little work on mass tourist environments. In the 1990s tourism geography engaged enthusiastically with the notion of sustainability, and a specialist literature has developed around

sustainable tourism. Much of the geographical contribution has been at the level of advocacy, but a more critical stance is emerging (Hall and Lew 1998). Tourism geographers have argued for a more holistic approach to sustainability, which recognizes the links between the economic, social, and environmental dimensions. They have also become more critical of idealistic attempts to build partnerships of stakeholders, as evidence emerges of unequal power relationships within these.

Tourism geography is also engaging with economic restructuring, and this is particularly evident in urban areas. In the face of major declines in their traditional economies, many urban localities have looked to tourism as one, and often the leading, component in their regeneration strategies. Tourism has the advantages of being labor intensive, requiring relatively limited investment, and having the capacity to refashion place images so as to reposition localities in the wider map of economic opportunities. Tourism geographers have played a role in analyzing the social and spatial impacts of such strategies, whilst demonstrating that their success is in part contingent on place characteristics. Similar research also exists on the role of tourism in other types of place environments, including the rural and the coastal resort.

See also: Consumption, Sociology of; Ecotourism; Environmental Policy: Protection and Regulation; Leisure and Cultural Consumption; Leisure, Sociology of; Tourism

Bibliography

Britton S 1991 Tourism, capital and place: towards a critical geography of tourism. *Environment and Planning D: Society and Space* **9**: 451–78

Butler R W 1980 The concept of a tourist area cycle of evolution: implications for management of resources. *Canadian Geographer* **24**(1): 5–12

Cohen E 1984 The sociology of tourism: Approaches, issues and findings. *Annual Review of Sociology* **10**: 373–92

Christaller W 1963 Some considerations of tourism location in Europe: the peripheral regions—underdeveloped countries—recreation areas. *Regional Science Association Papers* **12**: 95–105

Crouch D (ed.) 1999 *Leisure/Tourism Geographies*. Routledge, London

Hall C M, Lew A 1998 *Sustainable Tourism: A Geographical Perspective*. Longman, Harlow, UK

Ioannides D, Debbage K (eds.) 1998 *The Economic Geography of the Tourist Industry: A Supply Side Analysis*. Routledge, London

de Kadt E 1979 *Tourism: Passport to Development?* Oxford University Press, Oxford, UK

Mathieson A, Wall G 1982 *Tourism: Economic, Physical and Social Impacts*. Longman, Harlow, UK

Pearce D 1995 *Tourism Today: A Geographical Analysis*. Longman, Harlow, UK

Shaw G, Williams A M 1994 *Critical Issues in Tourism: A Geographical Perspective*. Blackwell, Oxford, UK

Smith V (ed.) 1977 *Hosts and Guest: The Anthropology of Tourism*. University of Pennsylvania Press, Philadelphia

Urry J 1990 *The Tourist Gaze*. Sage, London

A. M. Williams and G. Shaw

Trade and Exchange, Archaeology of

The term 'trade' can invoke images of professionals making profits in marketplaces, while 'exchange' implies the reciprocity of goods, such as food, between nonkin, or the movement of valuables within a social group itself. In this article trade and exchange will be treated equally and defined as the movement of materials or goods between individuals, social groups, or organizations. Ancient trade and exchange are often associated with rare and exotic valuable materials known to have restricted distributions, such as lapis lazuli, ostrich egg shell (Africa), amber, spondylus shell, jade, obsidian, copper, and other precious material items. Yet most past trade and exchange would have involved materials or goods used for daily existence, such as food. These, however, would be rarely seen in the archaeological record, being perishable. Thus the archaeology of trade and exchange deals mostly in those resilient artifacts hardy enough to be left in the archaeological record, such as obsidian, stone, pottery, or jade. Even then, these goods are residues 'resulting from a succession of repeated exchange transactions' (Kirch 1997). How and why an archaeologist identifies prehistoric trade and exchange is the subject of this article.

1. The Development of Trade and Exchange in Archaeology

Since the 1980s there has been an upsurge in studies on the archaeology of trade and exchange. The reasons for this are twofold. First, advances in science and technology have allowed and made easily available the sourcing and characterization of many materials. This provides evidence for the physical identification of their movement. Second, archaeologists began to see the usefulness in using trade and exchange in modeling the distribution of materials over space, and in explaining changes in societies over time.

1.1 Technological Developments

Developments in technology since the 1960s have led to an upsurge in chemical 'sourcing' or characterization studies. Yet the identification of traded archaeological material is not a recent phenomenon. Perhaps

the most famous early study was by Anna Shepard, who in the 1930s and 1940s postulated the exchange of pottery over wide areas of the Southwest USA by analyzing mineral inclusions in thin sections to pinpoint the origins of the minerals (Shepard 1965). Chemical techniques to identify the elemental composition of archaeological materials as a means of characterization are also not new. Neutron activation analysis was applied on coins from the Louvre in 1952, while in 1957 oriental ceramics were analyzed using nondestructive methods of x-ray fluorescence spectrometry (XRF) and x-ray diffraction (XRD) analysis (Young and Whitmore 1957). However, access to these techniques was restricted and costly.

A major advance in the chemical analysis of archaeological material came from the initiative of Dr. R. Oppenheimer. On March 8, 1956 he assembled a group of archaeologists and chemists at the Institute of Advanced Studies, Princeton, to discuss the possibility of applying methods of nuclear research to the study of archaeology (Sayre and Dodson 1957). As a result of this meeting, work was undertaken at two laboratories—the Brookhaven National Laboratory in the USA and the Research Laboratory for Archaeology and the History of Art at Oxford, UK. Techniques used included neutron activation analysis (NAA) and spectrographic methods. These early studies were reasonably successful, being able to separate pottery wares from Asia Minor, Greece, and Italy, and different factories of Samian ware. These studies laid the foundations for chemical analyses for the next three decades, in which thousands of analyses using varying techniques were carried out on many types of object including pottery, stone (obsidian, marble, chert, volcanic rocks), amber, and metals including coins. Apart from NAA, XRF, XRD, and spectrographic methods, techniques currently in use include proton induced x-ray emission and proton induced gamma ray emission (PIXE-PIGME), inductively coupled plasma emission spectrometry (ICP), lead isotope analysis, and electron microscopy, to mention a few. Major changes in the instrumentation of these techniques since the 1970s have meant that more elements can be analyzed with a higher precision. The choice of technique depends on availability to the archaeologist and cost. See Harbottle (1982) for a list of materials that can be characterized.

1.2 Theoretical Developments

The importance of trade and exchange in archaeology is the result of a reorientation in the ways archaeologists have perceived trade/exchange operating in the past. The realization that migrationist/invasion models could not adequately explain the distribution of material across the landscape, nor change in societies over time, led archaeologists to look at other mechanisms. Inspired by the works of economic anthropologists such as Polanyi (1957), who defined three types of trade and exchange (reciprocity, redistribution, and market exchange), and anthropologists such as Malinowski (1922), who studied trade and exchange in a Melanesian society, archaeologists realized that trade and exchange were embedded within the social and economic structures of society. By studying trade and exchange, the intention was to reconstruct the economies and organizations of past societies and their changes over time. The access to and control over prestige goods through trade and exchange were seen as prime movers for change leading to ranked societies.

A major proponent for the study of ancient trade and exchange was Colin Renfrew. In a classic paper on trade and culture process in European prehistory, Renfrew (1969) argued that trade was a causal factor for cultural change leading to the beginning of urbanization in the Aegean. An increase in trade with new wealth, craft specialization, and communication, led to change from a 'village subsistence economy into an urban society.' This was a marked shift away from inferring invasions or diffusion from a 'higher culture' to account for the Aegean Early Bronze Age. See Rathje (1971) for a similar case study involving the Maya.

2. Current Theory and Research

From these earlier studies a number of influential edited volumes on the archaeology of trade and exchange appeared: Sabloff and Lamberg-Karlovsky (1975), Ericson and Earle (1982), Earle and Ericson (1977), Renfrew and Shennan (1982), and Renfrew and Cherry (1986). These proceedings covered an eclectic array of topics from the Old and New Worlds, and provided the benchmark for trade and exchange studies for the next decade. These studies were important not only in identifying the movement of materials in the past, but also in cementing links with the chemists, geologists, or physicists who undertook the physicochemical analyses which demonstrated that trade and exchange took place. These studies are also important in initiating innovative research into the archaeology of trade and exchange, such as correlating the distribution of materials with modes of exchange and types of social organization, or exploring the relationship between trade and exchange with production and consumption patterns. These innovations will be discussed below.

Within this research a distinction has been made between two schools of thought about the way in which trade and exchange should be modeled: formalist and substantivist. The formalist approach works within the paradigm of market economics. Here, all exchange transactions can be viewed by economic rationalism, based on cost efficiency. Fall-off models presented in Sect. 2.2 below are based on

cost efficiencies. The substantivist school, on the other hand, views exchange as being embedded in social and economic processes. This approach allows an investigation into social groups, and social change. In reality, archaeologists draw from both approaches to model ancient trade and exchange.

Research mostly followed the following steps:

(a) identifying that exchange/trade took place;

(b) modeling the distribution patterns of the exchanged/traded material.

2.1 Identifying What is Moving and What is Not

Trade or exchange *must be demonstrated* rather than assumed. Archaeologists often view stylistic similarities over large distances as a consequence of exchange. Stylistic similarities in pottery, for instance, have been equated with the trade or exchange of pots from one or a restricted number of centers. Yet, in many instances this is not the case. One pottery tradition, Lapita, for instance, is spread from Papua New Guinea to Tonga, a distance of over 4,000 km. It is stylistically homogeneous, having a suite of shared decorated motifs made by a dentate stamping technique. Before the 1970s it was thought that this pottery was made in one or a handful of manufacturing centers and then exchanged to other communities, thus accounting for its spread within Melanesia. Yet, after extensive petrographic and chemical analysis, it was established that with the exception of a couple of assemblages, most Lapita pottery is locally made. It was people and ideas that were exchanged, not the pots. Thus, if local production of an object is demonstrated then models other than the trade/exchange of the object must be developed to account for stylistic similarity.

Identifying the source of an object and studying the context of production is not easy. The use of a 'resource zone concept of source' is useful here in defining a geographical region from which the object was produced (Clark et al. 1992). For instance, the exact spot where the firing of a pot took place, or where obsidian was extracted is not crucial here, only whether the pottery is of local origin or not, or whether the obsidian originated from a defined area. Attributing an object to a general regional zone (or procurement zone) is often more important than trying to identify a precise spot for manufacture.

The aim of many physicochemical analyses is to discern the geographical origin of the object under study. Ideally, the sources of material with restricted distribution could be characterized and 'fingerprinted' chemically or petrographically. Obsidian, for instance, is chemically homogeneous, and its sources can be chemically fingerprinted. Obsidian found in archaeological contexts can be analyzed and related back to the source. Identifying the potential source areas of a material requires an intensive search within a known region. For instance, in a recent study on obsidian exchange within the western Pacific over a 20,000 year time span, archaeologists have concentrated on the source areas within West New Britain, Papua New Guinea. Over a number of years, obsidian flows were mapped and samples taken for characterization, with the result that a finer discrimination of sources was possible with five source localities chemically defined (Summerhayes et al. 1998). By analyzing obsidian found in archaeological sites (some up to 3,700 km distant), it was possible not only to trace obsidian to its source localities, but also to define changes in the selection of different sources over time. For an overview of archaeological approaches to obsidian studies, see Shackley (1998).

Pottery, on the other hand, is more difficult to characterize or fingerprint (see *Ceramics in Archaeology*). Often archaeologists use what is called a 'criterion of relative abundance,' which infers that most pottery from a production center is to be found locally. Yet if pottery was produced for exchange only, then little would be found locally, thus limiting the application of this concept. The 'criterion of relative abundance' is useful, however, when used with a physicochemical analysis for provenance studies; that is, relating the pottery under study back to its raw materials: non-plastic inclusions (minerals, sands, etc.) and clay. Mineral inclusions found within pottery can be related to local geology and both river and beach sands. Clays, on the other hand, vary greatly in respect of the underlying geology, and a study of the elemental composition of pottery may give a clue to its origin. The chemical analysis of clay is complex. Bishop et al. (1982) note that clay minerals not only depend on their parent material, but are also affected by weathering or hydrothermal activity, climate, and geomorphology. They state that 'even if a region is considered to be homogeneous in its gross geologic characteristics, significant mineralogical and chemical differences may be discerned between clay deposits' (Bishop et al. 1982).

In any provenance study of archaeological objects a few words of caution from Garman Harbottle, a nuclear physicist, should best be heeded:

> archaeologists love the term sourcing, with its upbeat, positive thrust—that you analyze or examine an artifact and, by comparison with the material of known origin, 'source' it. In point of fact, with very few exceptions, you cannot unequivocally source anything. What you can do is characterize the object, or better, groups of similar objects found in a site or archaeological zone by mineralogical, thermoluminescent, density, hardness, chemical and other tests, and also characterize the equivalent source materials, if they are available, and look for the similarities to generate attributions. A careful job of chemical characterization, plus a little numerical taxonomy and some auxiliary archaeological and/or stylistic information, will often do something as useful: It will produce groupings of artifacts that make archaeological sense. This, rather than absolute proof of origin, will often necessarily be the goal. (Harbottle 1982)

For an overview of archaeological approaches to pottery characterization studies, see Neff (1992).

2.2 Distribution Patterns

Archaeologists analyze the distribution of an object from its source or production area to its consumption site in order model the type of trade/exchange that took place. In some instances the nature of the society participating in the trade/exchange system is inferred, such as egalitarian versus chiefly, or nonhierarchical versus hierarchical.

Popular among archaeologists is the use of fall-off curves and modes of trade/exchange. By plotting the frequency of an item against distance from its source (fall-off curve), Renfrew (1975) observed different distribution patterns which he attempted to relate to modes of exchange or access. Some of these modes included reciprocity, down-the-line, and central-place exchange.

How were fall-off curves and modes of exchange related? Down-the-line exchange and reciprocity led to an exponential fall-off in abundance away from the source or production center. The term 'law of monotonic decrement' was coined to account for this fall-off in the frequency of an item away from the source (Renfrew 1977). Attempts were also made to define an object's value by the slope of the curve. A small slope equates with high-value goods, such as spondylus shells in Europe, Maya jade, and Oaxaca iron-ore mirrors. These modes were equated with an egalitarian society (semisedentary agriculturalists and hunter-gatherers; nonranked; nonhierarchical).

Central-place exchange, on the other hand, was seen as directional (no exponential linear fall-off), should lead to multinodal fall-off curves, and is indicative of redistribution. Central place relates to the movement of goods toward one place or center, from which they are then distributed. Redistribution has been related to complex chiefly societies (hierarchical, ranked and/or stratified societies). Examples of such societies given by archaeologists include early states such as the Maya, Egypt, the Near East, or chiefdoms such as Hawaii.

Although problems occur with equifinality—that is, different modes of exchange produce similar distributions—this approach is important as it relates the measurement of archaeological material away from the source to trade and exchange modes and the structures of societies.

3. Future Directions

Approaches to the archaeology of trade and exchange will continue to change on three fronts. First, change will occur in tune with developments in archaeological theory. As archaeologists will be asking different questions about the roles of trade and exchange and society, changes in approaches and methodologies will take place. Second, at the same time, it is reasonable to expect advances in characterizing techniques, making them yet more accessible and cheaper to use. Lastly, archaeologists will consolidate knowledge on the sources and distribution of a variety of goods, giving a more complete perspective on the roles of trade and exchange within a society. As an example, attention is drawn to the Pacific where the production and distribution of a number of materials, such as pottery, obsidian, chert, volcanic adzes, and shell artifacts, are used to show that each class of material has different production and distribution patterns. If any class of artifact were taken by itself, a 'myopic' view would be seen of the roles of trade and exchange within societies. Only by comparing the production and distribution patterns of the complete range of material available can a proper picture of trade and exchange be given.

See also: Archaeometry; Ceramics in Archaeology; Geoarchaeology; Archaeology of Intensification and Specialization; Archaeology of Pacific Islands; Cultural Contact: Archaeological Approaches; Economic Anthropology; Economic Sociology; Exchange in Anthropology; Malinowski, Bronislaw (1884–1942); Polanyi, Karl (1886–1967)

Bibliography

Bishop R, Rands R, Holley G 1982 Ceramic compositional analyses in archaeological perspective. *Advances in Archaeological Method and Theory* **5**: 275–320

Clark C P, Neff H, Glascock M 1992 Neutron activation analysis of the late Woodland ceramics from the Lake Superior Basin. In: Neff H (ed.) *Chemical Characterization of Ceramic Pastes in Archaeology*. Prehistory Press, Madison, WI, pp. 255–67

Earle T K, Ericson J E 1977 *Exchange Systems in Prehistory*. Academic Press, New York

Ericson J E, Earle T K 1982 *Contexts for Prehistoric Exchange*. Academic Press, New York

Harbottle G 1982 Chemical characterization in archaeology. In: Ericson J E, Earle T K (eds.) *Contexts for Prehistoric Exchange*. Academic Press, New York, pp. 13–51

Kirch P V 1997 *The Lapita Peoples: Ancestors of the Oceanic World*. Blackwell, Oxford, UK

Malinowski B 1922 *Argonauts of the Western Pacific*. Routledge and Kegan Paul, London

Neff H 1992 *Chemical Characterization of Ceramic Pastes in Archaeology*. Prehistory Press, Madison, WI

Polanyi K 1957 The economy as instituted process. In: Polanyi K, Arensberg C M, Pearson H W (eds.) *Trade and Market in the Early Empires*. Free Press, Glencoe, IL, pp. 243–70

Rathje W 1971 The origin and development of classic Maya civilization. *American Antiquity* **36**: 275–86

Renfrew C 1969 Trade and cultural process in European prehistory. *Current Anthropology* **10**: 151–69

Renfrew C 1975 Trade as action at a distance. In: Sabloff J A, Lamberg-Karlovsky C C (eds.) *Ancient Civilization and Trade*. University of New Mexico Press, Albuquerque, NM, pp. 3–59
Renfrew C 1977 Alternative models for exchange and Spatial Distribution. In: Earle T K, Ericson J E (eds.) *Exchange Systems in Prehistory*. Academic Press, New York, pp. 71–90
Renfrew C, Cherry J F 1986 *Peer Polity Interaction and Socio-political Change*. Cambridge University Press, Cambridge, UK
Renfrew C, Shennan S (eds.) 1982 *Ranking, Resource and Exchange: Aspects of the Archaeology of Early European Society*. Cambridge University Press, Cambridge, UK
Sabloff J A, Lamberg-Karlovsky C C 1975 *Ancient Civilization and Trade*. University of New Mexico Press, Albuquerque, NM
Sayre E V, Dodson R 1957 Neutron activation study of Mediterranean potsherds. *American Journal of Archaeology* **61**: 35–41
Shackley S 1998 *Advances in Archaeological Volcanic Glass Studies*. Plenum Press, New York
Shepard A 1965 Rio Grande glaze-paint pottery: A test of petrographic analysis. In: Matson F R (ed.) *Ceramics and Man*. Viking Fund Publications in Anthropology 41. Aldine, Chicago, pp. 62–87
Summerhayes G R, Bird R, Fullagar R, Gosden C, Specht J, Torrence R 1998 Application of PIXE-PIGME to archaeological analysis of changing patterns of obsidian use in West New Britain, Papua New Guinea. In: Shackley S (ed.) *Advances in Archaeological Volcanic Glass Studies*. Plenum Press, New York, pp. 129–58
Young W, Whitmore F 1957 Analysis of oriental ceramics by non-destructive methods. *Far Eastern Ceramic Bulletin* **9**: 1–27

G. R. Summerhayes

Trade Unions, Economic Behavior of

1. Trade Unions

A trade union is an organized association of workers formed for the protection and promotion of their common interests. The standard view of trade unions is that they are monopoly organizations that improve the welfare of members, principally by raising wages above the competitive level. This article investigates the conditions under which a union can increase wages, and explores ways of modeling the competing preferences of unions and management.

2. Conditions Under Which a Union can Improve Workers' Welfare

For a trade union to be able to increase wage rates above the competitive level, there must be some surplus that can be shared between the firm and the union, and the union must have some bargaining power to induce the firm to share this surplus. How can a union achieve such power? It may be able to organize all workers in an industry, thereby acting as a monopolist over the supply of labor. The earliest successful unions were craft unions, which became established because of a combination of high demand for skilled labor during industrialization, and the control of trained labor by skilled workers through the apprenticeship system. The emergence of general or industrial unions followed a different path. During industrialization, the huge surplus of unskilled displaced agricultural labor made manipulation of labor supply impossible. A depression, coupled with a readily available pool of substitute workers, could destroy a union. To ensure survival, a general union needed political support or very high levels of membership. Obtaining the latter was difficult in the early stages of unionization, when there were no immediately obvious wages benefits because the union had not yet obtained any power.

Union bargaining power derives from the ability of the union to inflict damage on the firm by withdrawing labor. This works when there is no substitute pool of nonunion workers, for example, if there are high labor turnover costs or if the union controls labor supply through entry restrictions or through having high levels of membership relative to the size of the sector. The view that workers have to battle to acquire sufficient power to induce the firm to concede any surplus accords with the Marxist analysis of the capitalist means of production (see, for example, Bowles 1985). Union membership density may play a part in determining the power of the union to extract from the firm a share of any rents. As membership density increases, there is a shrinking pool of workers available to the firm should it choose not to employ union labor. Thus one would expect to find a minimum critical level of membership below which firms would refuse to recognise and bargain with the union (Osborne 1984). Where membership density is crucial to the union's monopoly power as, for example, with the emergence of the general unions, the union must consider how to attract workers at the margin. The basic problem for economists in explaining why workers join a union relates to the free-rider problem, arising because of the collective nature of many of the goods provided by the union (Olson 1965, Booth 1985). Why join the union when union dues are costly, and when the benefits are not immediately available or apply to all workers regardless of their union status? Of course established unions may also provide incentive private goods that encourage workers to unionize. Indeed, a union can act as the workers' agent, since it can reduce the costs of negotiation and processing of grievances, a role which may also be useful to employers (Faith and Reid 1987).

Even if a union controls labor supply, it will not necessarily be able to negotiate a large wage increase relative to the competitive level. The magnitude of the union wage depends crucially on the elasticity of labor demand in that sector. There can be little doubt that, if unions emerge in competitive markets, high union

wages introduce allocative inefficiencies into the economy through the distortion of factor prices. Without unions, allocative efficiency is associated with the equalization of the marginal products of identical factor inputs across sectors. With higher wages in the union sector, union firms employ fewer workers. Displaced union workers crowd into the nonunion sector, lowering wages there. As a result, too few workers are employed in the union sector where output falls, while too many workers are employed, and too many goods produced, in the nonunion sector. There is a deadweight efficiency loss, because the value of marginal products in the two sectors are not identical. In addition, there are distributional issues to consider, and longer run effects due to substitution of capital for labor. Furthermore, in unionized sectors there may be under-investment in capital through the hold-up problem (see, for example, Grout 1984).

This discussion of allocative inefficiency assumes that the union has emerged in an economy characterized by competitive product and labor markets. But there is considerable evidence that, in modern industrialized countries, many product markets are characterized by imperfect competition. Moreover, nonunion firms may also face an incumbent workforce with a degree of bargaining power (Lindbeck and Snower 1988). Thus, even in the absence of unionization, management and workers may be in a situation of bilateral monopoly. An important question (little addressed in the literature) arises as to whether or not the replacement of individual bargaining by collective bargaining generates additional inefficiencies and misallocation of resources in situations where markets were previously not functioning in accordance with the textbook model of perfect competition. The majority of formal trade union models in the literature assume a perfectly competitive product market to allow the models to focus on wage and employment determination in the simplest environment (Dowrick 1989 *inter alia* is an exception). Nonetheless, it is an empirical regularity that imperfections in product and labor markets are correlated (Stewart 1990).

Even within the competitive model, there are arguments suggesting that, in the presence of imperfect information and uncertainty, unions may enhance efficiency. To the extent that unions reduce labor turnover and negotiating costs, they may increase the available surplus to be shared between parties (Freeman and Medoff 1984). Of course, there may be interdependence between the monopoly and efficiency roles of trade unions: unless the union has some bargaining power, it may be unable to increase efficiency.

3. *Formal Modeling of Union Behavior*

To investigate union behavior, the economist needs to model the objectives of the relevant actors involved in wage setting, as well as the means of resolving their potentially conflicting objectives. Most union models consider the objectives of two agents—the union and management. Typically, the government's behavior is assumed exogenous. Because the objectives of union and management may be conflicting, models need to consider how conflicts might be resolved within the constraints provided by the institutional environment. The range of issues covered by collective bargaining in Europe is narrower than in the USA, which probably reflects a relative lack of US social welfare legislation. (In Europe, legislation typically covers issues such as health insurance, pensions, redundancy pay, and unfair dismissal that are often covered by collective agreements in the US.)

While bargaining may take place over a number of issues, for tractability researchers typically concatenate wages and other forms of compensation, such as fringe benefits, under the single heading of wages. Union preferences appear to include employment as well as wages, but it is rare to find instances of the union and the firm bargaining over both wages and employment. Of course, unions affect employment indirectly, not only through increasing wages but also through negotiation over staffing levels, recruitment, and redeployment.

While unions typically have a number of goals (from increasing remuneration to broad political aims concerning the labor movement as a whole), the orthodox union model assumes unions are concerned only with the unionized sector. But a union typically comprises heterogeneous individuals with varying and possibly conflicting preferences as to the union's strategy. For example, there may be principal–agent problems arising because union leaders (agents) have objectives that may conflict with the membership (the principal). The leadership may be better informed than its membership, or perhaps interested in self-aggrandizement or in wider labor goals. Unions are usually large, with complex organizational structures concerned with representation of members' interests and with administration. Because the internal organization of the union is typically hierarchical in structure, the question arises as to how well members' preferences are transmitted through the various levels, each of which may have conflicting aims. Moreover, there is heterogeneity across unions in terms of their structure and constitution (ranging from the democratic through to the totalitarian) that may determine a union's maximizing strategy.

While there is a small literature allowing for union objectives to be affected by membership heterogeneity or a conflict of interests between leaders and members (Farber 1978, Blair and Crawford 1984), the widely used models of union behavior side-step these problems by supposing that all workers in the union sector are identical and are union members. They also assume that the union cares only about the economic welfare of the union sector. For simplicity, union-negotiated

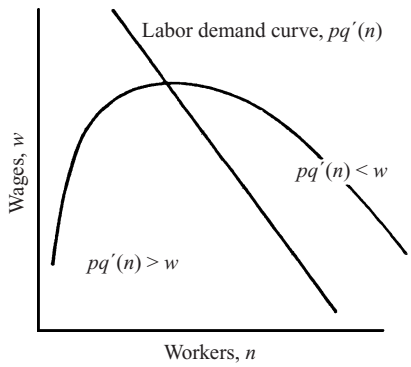

Figure 1
The isoprofit curve

benefits are assumed to be encapsulated in the wages variable. Because the union is also hypothesized to care about the employment of union workers, union objectives can be specified as an increasing function of wages w and employment n.

In most recent union literature, the union is assumed to maximize either a utilitarian objective function, or an expected utility objective function. The utilitarian objective function assumes that the union maximizes the sum of the individuals' utilities. The union comprises identical members treated identically by the union, and therefore in this context no distributional or normative judgments are being made by the utilitarian objective. However, if workers were assumed heterogeneous, the use of this as the union objective would introduce normative judgments (since if individuals were not identical, it would be necessary to consider why the utilitarian social welfare should be adopted instead of some alternatives such as the Rawlsian or Maxi–min criterion, or a more egalitarian approach). If the union's wage-setting behavior results in some union members being unemployed, the utilitarian objective function is:

$$U = nu(w) + (t-n)u(b) \quad w > b;\ 0 < n \leqslant t \qquad (1)$$

where $u(.)$ is a concave (indirect) utility function, w denotes wages, n denotes employment in the union sector, t is the total number of union workers in the sector (and membership by assumption), and b is the alternative sector wage or the unemployment benefit.

The second popular union objective function is the expected utility function. If membership is fixed, Eqn. (1) can be written equivalently in terms of expected utility. Since the union raises wages above the competitive level, each member faces some probability $(1-n/t)$ of being unemployed. If unemployed, a worker receives the alternative sector wage or the unemployment benefit. The expected utility of a representative union member is given by:

$$E(U) = \frac{n}{t}u(w) + \left(1 - \frac{n}{t}\right)u(b) \quad u'(w) > 0;\ u''(w) \leqslant 0 \qquad (2)$$

where $u'(w) = \mathrm{d}u/\mathrm{d}w$. The slope of the union indifference curve in (w, n) space is found by totally differentiating either Eqns. (2) or (3) with respect to w and n, holding (expected) utility fixed, yielding $\mathrm{d}w/\mathrm{d}n = -[u(w) - u(b)]/nu'(w)$

Both the utilitarian and the expected utility objective functions have the implication that identical workers receive different outcomes *ex post*. Employed members receive the union wage, while the unemployed workers receive b. However, if unemployed members effectively leave the union, the leadership will be concerned about future union viability, since there will be a larger pool of nonunion labor that might undercut the union wage, and union income from dues would fall. This is likely to affect union outcomes, and is an argument for inclusion of the level of membership in the union objective function as an endogenous variable. Moreover, b may not be exogenous—the union may affect remuneration for unemployed members through redistribution by benefits or redundancy pay, suitably adjusted for the disutility of work for those in employment. Indeed, historically, many unions paid unemployment benefits to out-of-work members, and some commentators argue that this was important for union growth and survival.

Now consider the behavior of the firm. The necessary conditions for a union to be able to appropriate any surplus in a competitive product market, without driving the firm out of existence, are: very high levels of unionization in a competitive industry, no nonunion foreign competition, entry barriers, and inelastic product demand. Without the first three conditions, domestic firms, or foreign firms, or new firms entering the industry could simply hire nonunion labor and produce at a lower cost than unionized competitors, thereby driving them out of business. Assuming the firm maximizes profits and has no power in the labor market (which means it will take wages as given), it has just one choice variable—employment. The optimization problem facing the firm is therefore:

$$\max_n \pi = pq(n) - wn \quad q'(n) > 0;\ q''(n) < 0;$$
$$q(0) = 0 \qquad (3)$$

where π denotes profits, p denotes product price, and $q(n)$ is the production function characterized by diminishing returns to labor. The firm's wage bill is wn

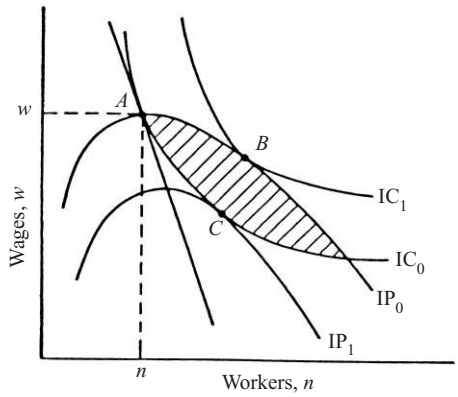

Figure 2
Inefficiency of the Monopoly Union Model

and capital is held fixed. The slope of the firm's isoprofit curve in (w, n) space is found by totally differentiating Eqn. (3) with respect to wages and employment keeping profits fixed, to obtain $\mathrm{d}w/\mathrm{d}n = [pq'(n) - w]n$. The locus of the turning point of each isoprofit curve traces out the labor demand curve where $pq'(n) = w$, as illustrated in Fig. 1. Two isoprofit curves IP_0 and IP_1 are illustrated in Fig. 2, together with two union indifference curves IC_0 and IC_1.

4. *The Monopoly Union Model*

In the monopoly union model, there is no bargaining between the two parties. Instead, the union acts as a monopolist in the supply of labor and imposes a wage rate on the firm. The firm, however, retains its managerial prerogative to determine the number of workers employed at the union wage. Once the union sets the wage, the firm reads off from the labor demand curve the number of workers to hire at that wage. Thus, a union maximizing the expected utility of its representative worker sets wages as follows:

$$\max_{w} E(U) = \frac{n}{t}u(w) + \left(1 - \frac{n}{t}\right)u(b) \quad \text{subject to} \quad pq'(n) = w \tag{4}$$

The first-order condition can be found directly by equating the slopes of the union indifference curve (giving the union's marginal rate of substitution of employment for wages) with the slope of the labor demand curve, illustrated as point A in Fig. 2. The monopoly union seeks its highest level of utility, subject to the constraint of the firm's labor demand curve, where w^* and n^* are the optimal wage and employment levels. It can be seen, from the first order condition in Eqn. (5), that wages will be set by the union such that the percentage increase in a member's utility due to a percentage increase in wages is exactly equal to the wage elasticity of labor demand ε. Intuitively, this result arises because an increase in wages reduces union employment, and thus each worker faces a greater probability of being unemployed. On the other hand, any member lucky enough to be employed receives a higher wage. Thus, in equilibrium, the percentage marginal benefit is equal to the percentage marginal cost:

$$\varepsilon \equiv -\frac{n'(w)w}{n} = \frac{wu'(w)}{[u(w) - u(b)]} \tag{5}$$

This equation simply characterizes the equilibrium situation. To obtain a specific value for wages, it is necessary to choose explicit functional forms for the utility function and the labor demand function that can then be inserted into the generalized characterization and solved for the wage rate. For evaluation of empirical work attempting to test this model, see Pencavel (1991) and Booth (1995).

Using this model, it is straightforward to show how the wage set by the monopoly union alters due to a change in each of the explanatory variables. An isoelastic increase in labor demand has no effect on wages, but employment increases. An improvement in alternative opportunities increases the union wage, but employment is reduced as we move up the labor demand curve. An increase in membership has no effect on wages or the numbers employed. The empirically testable predictions of the monopoly union model are first that the alternative wage and unemployment benefit should affect wages directly. But employment should be determined by the wage rate, and unaffected by variables measuring alternative opportunities. Second, membership changes should have no impact on the union wage.

5. *The Bargaining Approach*

Next consider more sophisticated methods of reconciling the different preferences of union and management, using a bargaining approach. It turns out that the monopoly union model is nested as a special case within a simple bargaining framework. While some researchers argue that the monopoly union model is a reasonable approximation for particular industries at particular times, observation suggests that wages are more often bargained over by unions and firms. When union and firm bargain over wages, they are in a bilateral monopoly situation, in which a single seller confronts a single buyer of labor. A vital element in this bargaining is the ability of each agent to impose costs on the other agent if it refuses to bargain and engages in a strike or lockout. Although strikes are relatively infrequent, this does not mean that the strike or lockout is unimportant. It is the threat to

withdraw that counts. What induces the two parties to bargain is the threat of facing costs imposed by the other party.

Since the union and the firm are in a situation of bilateral monopoly, additional structure must be imposed on the model to generate a determinate solution of the division of the available surplus between the two parties. There are two broad approaches to modeling bargaining behavior—the axiomatic or Nash bargaining approach, and the more recent game theoretical approach (Nash 1950). Under certain conditions, these generate identical predictions in spite of the fact that they proceed from very different assumptions (Binmore et al. 1986).

The solution concept widely used in the union literature is the generalized Nash bargaining solution, according to which wages are determined by maximization of the product of each agent's gains from reaching a bargain, weighted by their respective bargaining strengths. Since management is assumed to retain the right to determine employment unilaterally, this is referred to frequently as the 'right-to-manage' model. Define a status quo or fall-back point for each party as its position if no bargain is reached. For the firm, the status quo is zero, because if it does not reach a bargain with the union workforce, it cannot obtain any other workers since there are no nonunion workers in the sector. So the firm's net gain is simply its profits function. For the union, its status quo for a representative member is $u(b)$, since that is what a member will receive if no bargain is reached. Thus, the net gain to the union is simply the expected utility of a representative member (see Eqn. (2)) less fall-back utility, given by $[u(w)-u(b)]n/t$.

The generalized Nash bargaining solution (the product of the weighted net gains to each party) can be written as:

$$\max_{w} B = \left\{ \frac{n}{t}[u(w)-u(b)] \right\}^{\beta} \{pq(n)-wn\}^{1-\beta} \quad (6)$$

where β measures the bargaining strength of the union and $(1-\beta)$ is the bargaining strength of the firm. The solution to this bargaining problem is obtained from the first order condition of Eqn. (6), and is given by:

$$\frac{\beta w u'(w)}{u(w)-u(b)} = \beta\varepsilon + \frac{(1-\beta)wn}{pq(n)-wn} \quad (7)$$

where ε is the wage elasticity of labor demand. The left-hand side of Eqn. (7) represents the proportional marginal benefit to the bargain from a proportional increase in the wage. Since only the union receives this benefit, it is weighted by the union's bargaining power β. The first term on the right-hand side is the union's proportional marginal cost (the percentage reduction in employment due to the proportionate wage increase) weighted by union power β. The second term on the right-hand side represents the firm's proportional marginal cost, weighted by the firm's power $(1-\beta)$. If the union has all the power, $\beta = 1$ and the union extracts the entire surplus—the right to manage model collapses to the monopoly union model. For $0 < \beta < 1$, the employment and wage outcome will lie on the labor demand curve anywhere between the monopoly union extreme and the competitive outcome. If the firm has all the power, $\beta = 0$ and the wage is then simply the alternative wage b.

The empirically testable predictions of the right-to-manage model are the same as a monopoly union model with the addition of the prediction that variables proxying union strength will affect wages directly.

Is the right-to-manage model Pareto-efficient? (Pareto efficiency is defined by economists as a situation where no one can be made better off without making someone else worse off.) Or can at least one party be made better off by shifting to some other (w, n) pair? In Fig. 2, the isoprofit curves IP represent different levels of profits for the firm, with higher curves denoting lower profits since they are associated with higher wages. IC denotes the union indifference curves, with expected utility increasing as the union moves northeast. Suppose the equilibrium of the right-to-manage model is at point A. The union could be made better off while keeping the firm on the same isoprofit curve, by shifting to point B. Alternatively, the firm could shift to a higher isoprofit curve, while keeping the union on the original indifference curve, by moving to point C. Both parties could be made better off by shifting from A to any point in the shaded area. Positions on the labor demand curve are thus Pareto-inefficient. Either the union or the firm could be made better off by bargaining over employment as well as wages, allowing them to shift from a point such as A to B or C or anywhere in the shaded area. Efficiency in the context of this model is where the marginal rates of substitution of employment for wages, for both union and firm, are equal.

6. *The Efficient Bargaining Model*

Now suppose the union and firm simultaneously determine wages and employment. While evidence suggests that unions and firms do typically not bargain over wages and employment, this approach has been important in the union literature because of its efficiency property beloved by economists. But the fact that this does not appear to happen in practice suggests that there may be something economists are missing in their modeling. (Of course the wage and employment outcome predicted by the efficient bargaining model is inefficient from the viewpoint of society as a whole.)

Efficiency requires only that one party's welfare be maximized subject to an arbitrarily fixed level of welfare of the other. This makes it possible to characterize an efficient outcome simply by consider-

ing union and firm preferences, without imposing a particular bargaining structure. While this does not yield a unique outcome, it illustrates the necessary conditions for efficiency. An efficient bargain over (w, n) solves:

$$\max_{w,n} pq(n) - wn \quad n \geqslant 0; \; w \geqslant b$$

$$\text{subject to} \quad \frac{n}{t}[u(w) - u(b)] + u(b) = \bar{U} \qquad (8)$$

From the first-order conditions is obtained the result that an efficient (w, n) pair is given by equating the marginal rates of substitution of employment and wages for the union and the firm. This means equating the slopes of the isoprofit and union indifference curves, as illustrated as point C in Fig. 2, and shown below:

$$pq'(n) - w = -\frac{[u(w) - u(b)]}{u'(w)} \qquad (9)$$

The contract curve is defined as the locus of all points satisfying Eqn. (9). Two points on the contract curve are B and C in Fig. 2. The contract curve is positively sloped for risk-averse workers (but vertical if workers are risk-neutral). From Eqn. (9), note that the value of the marginal product of labor $pq'(n)$ is less than the wage rate, by an amount that is equal to the marginal rate of substitution of employment for wages. Hence, the contract curve lies to the right of the labor demand curve for $w > b$. If $w = b$, then from Eqn. (9) $pq'(n) = w$, the perfectly competitive result. Thus, the contract curve and the labor demand curve coincide where $w = b$. Since for efficiency we only require that one party's welfare be maximized subject to any arbitrarily fixed level of welfare of the other, any point on the contract curve represents an efficient outcome for a particular arbitrarily fixed level of welfare of one agent.

To determine a unique (w, n) pair, a means of obtaining a particular solution to the bargaining procedure must be added to the model. A common approach is to use the generalized Nash bargaining model of Eqn. (6) modified to allow for bargaining over both w and n. The solution to this maximization problem generates a determinate outcome for wages and employment, which lies somewhere on the contract curve (Leontief 1946, McDonald and Solow 1981). In the 1980s considerable attention was devoted to devising ways of testing which model was appropriate for the union sector (Pencavel 1991).

Many unions are unable to bargain over employment directly, and thus the relevance of this fully efficient model might be questioned. However, unions can often affect employment indirectly, for example, through bargaining over the labor-to-capital ratio, work intensity, or redundancy payments. Johnson (1990) find that, in models in which there is bargaining over wages and the labor-to-capital ratio or work intensity, the wage-employment outcome lies somewhere between the fully efficient and the right-to-manage model.

7. *Conclusions*

This article can only give a flavor of the economics behavior of the trade union, and many important topics have been omitted, for example, the impact of unionization on the macroeconomy (Layard et al. 1991). Other important topics omitted (but covered in Booth 1995) include the economics of strikes (see *Dispute Resolution in Economics*), the relationship between unions and hours of work, and the modeling of endogenous union membership, an important topic in countries such as the UK, in which compulsory union membership is illegal. There is also a vast empirical literature estimating the impact of unions on economic performance (see *Trade Unions: Empirical Analyses*).

What areas are currently being investigated by economists working on unions? Some recent developments include combining ideas about oligopolistic firm behavior with a unionized labor market in a dynamic framework. While it is too early to see where this literature is heading, this approach might possibly be instructive about appropriate forms of collective bargaining structure. Some countries (such as many European economies) have centralized wage bargaining arrangements, while others such as the US are decentralized. It is important that these institutional differences are incorporated into the modeling framework. Calmfors and Driffill (1988) model the impact of collective bargaining structure on economic performance, while Horn and Wolinsky (1988) model choice of collective bargaining structure. These approaches could in principle provide pointers as to what an 'ideal' collective bargaining structure might be.

While union membership and power have declined in many industrialized countries in the 1980s and 1990s (for example, Britain and the US), it is unlikely that unions will vanish all together. Indeed, in the future unions may well become stronger in Eastern European, Asian, and African countries. It is therefore important for researchers working on the economics of trade unions to consider what form of unionism might be least distortionary in a modern economy. Moreover, while many models of union behavior refer to explicit labor contracts between a union and management, the models are also relevant to a wider class of situations than those in which a union explicitly represents workers. The models are potentially applicable, with only a few minor modifications, to nonunion labor markets where incumbent workers have some market power and where the firm has economic rents that might be shared.

See also: Economic Geography; Employment and Labor, Regulation of; Labor, Division of; Spatial Labor Markets; Unemployment: Structural; Work: Anthropological Aspects; Work, Sociology of; Workplace Environmental Psychology

Bibliography

Binmore K, Rubinstein A, Wolinsky A 1986 The Nash bargaining solution in economic modeling. *RAND Journal of Economics* **17**(2): 176–88

Blair D H, Crawford D L 1984 Labor union objectives and collective bargaining. *Quarterly Journal of Economics* **99**(3): 547–66

Booth A L 1985 The free rider problem and a social custom theory of trade union membership. *Quarterly Journal of Economics* **100**(Feb): 253–61

Booth A L 1995 *The Economics of the Trade Union*. Cambridge University Press, Cambridge, UK

Bowles S 1985 The production process in a competitive economy: Walrasian, Neo-Hobbesian, and Marxian models. *American Economic Review* **75**(1): 16–36

Calmfors L, Driffill J 1988 Bargaining structure, corporatism and macroeconomic performance. *Economic Policy* **6**: 13–62

Dowrick S 1989 Union-oligopoly bargaining. *Economic Journal* **99**: 1123–42

Faith R L, Reid J D 1987 An agency theory of unionism. *Journal of Economic Behavior and Organization* **8**: 39–60

Farber H S 1978 Individual preferences and union wage determination: The case of the united mine workers. *Journal of Political Economy* **86**(5): 932–42

Freeman R B, Medoff J L 1984 *What Do Unions Do?* Basic Books, New York

Grout P A 1984 Investment and wages in the absence of binding contracts. *Econometrica* **52**(2): 449–60

Horn H, Wolinsky A 1988 Worker substitutability and patterns of unionisation. *Economic Journal* **98**: 484–97

Johnson G E 1990 Work rules, featherbedding and pareto-optimal union–management bargaining. *Journal of Labor Economics* **8**: S237–59

Layard R, Nickell S J, Jackman R 1991 *Unemployment; Macroeconomic Performance and the Labor Market*. Oxford University Press, Oxford, UK

Leontief W 1946 The pure theory of the guaranteed annual wage contract. *Journal of Political Economy* **54**(1): 76–9

Lindbeck A, Snower D J 1988 *The Insider–Outsider Theory of Employment and Unemployment*. MIT, Cambridge, MA

McDonald I M, Solow R M 1981 Wage bargaining and employment. *American Economic Review* **71**: 896–908

Nash J F 1950 The bargaining problem. *Econometrica* **18**(2): 155–62

Olson M 1965 *The Logic of Collective Action*. Harvard University Press, Cambridge, MA

Osborne M J 1984 Capitalist–worker conflict and involuntary unemployment. *Review of Economic Studies* **51**: 111–27

Pencavel J 1991 *Labor Markets under Trade Unionism*. Blackwell, Oxford, UK

Stewart M B 1990 Union wage differentials, product market influences and the division of rents. *Economic Journal* **100**: 1122–37

A. L. Booth

Trade Unions: Empirical Analyses

A common definition of a trade union is that it is an association of workers who bargain collectively with their employer regarding the terms and conditions of employment. Empirical analysis of trade unions falls largely into four related categories:

(a) Unions as maximizing agents.

(b) The bargaining process: strikes and labor disputes.

(c) The effect of unions on wages.

(d) The determination of union membership.

Category (a) is covered in *Trade Unions, Economic Behavior of*, and category (b) in *Dispute Resolution in Economics*. This article is organized around discussions of categories (c) and (d), the effect of unions on wages and the determination of union membership, respectively. The discussion will rely on the (admittedly atypical) institutional arrangements that exist in the United States, but the key points are applicable more generally to unions in market economies. The discussion is necessarily selective, and it is meant to give an overview of a very substantial literature.

1. The Effect of Unions on Wages

1.1 Unions and Wages: What is Measurable?

Lewis's (1963) classic monograph set the standard for empirical analysis of the effect of unions on wages using aggregate (e.g., industry-level) data. As individual level data became available beginning in the late 1960s, a substantial literature, surveyed comprehensively by Lewis (1986), developed, addressing this same question using the new data. Other useful surveys of this literature are continued in Freeman and Medoff (1984), Hirsch and Addison (1986), and Booth (1995).

There are at least two levels on which unions can affect wages in an economy. The first is the direct effect of unions on the wages of workers in jobs where wages are set through collective bargaining (for nonmembers as well as members). The second, and less well understood, level is that the presence of unions in an economy can change the level and distribution of wages generally. For example, unions can raise the wages of their members, but the wages of nonunion workers can fall or rise as well. A standard competitive market model would imply a fall in the wage of nonunion workers. The mechanism is that unions, in raising wages of their members, cause a decline in employment in the union sector. The result is an increase in the supply of workers to the nonunion sector as displaced union workers move to that sector, implying a drop in the nonunion wage rate, called the 'spillover effect' of unions. A force working to raise

nonunion wages is the so-called 'threat-effect' of unions. The threat effect results from nonunion employers raising the wages of their workers in order to avoid becoming unionized (Rosen 1969, Freeman and Medoff 1981). Thus, based on theoretical considerations, the effects of unions on aggregate wages in an economy are not clear.

In order to be more precise, define the effect of unions on the wages of union and nonunion workers respectively as:

$$\Delta_u = \frac{W_u - W_0}{W_0} \tag{1}$$

and

$$\Delta_n = \frac{W_n - W_0}{W_0} \tag{2}$$

where W_u and W_n represent the union and nonunion wage, respectively, and W_0 represents the wage that would have prevailed in the absence of unions in the economy. It is not possible to estimate either of these wage effects because W_0 is not observable.

The union–nonunion wage differential, called the union wage gap, and defined as:

$$\Delta = \frac{W_u - W_n}{W_n} \tag{3}$$

is estimable and is related to the underlying union and nonunion wage effects by:

$$1 + \Delta = \frac{1 + \Delta_u}{1 + \Delta_n}. \tag{4}$$

As long as all differentials are small (less than 0.25 or so), this expression is usefully approximated by

$$\Delta \approx \Delta_u - \Delta_n. \tag{5}$$

This relationship states that the measured union wage gap is approximately equal to the difference in the proportional effects of unions on the union and nonunion wage.

The union wage gap, defined in Eqn. (3), can be approximated usefully by the difference in log wages. This implies that:

$$\Delta \approx \ln(W_n) - \ln(W_n) \tag{6}$$

and estimation of this quantity is the goal of much of the empirical literature on the effects of labor unions.

The union wage gap reflects a combination of effects. Clearly, it includes the direct effect of unions on the wage of unionized workers, and it includes the offsetting effects, described above, of unions on the wage of nonunion workers. It is not possible to parse estimates of the union wage gap into its component effects, and estimates of Δ have been interpreted broadly as the effect of unions on wages.

A further complication is that workers are heterogeneous and only a single wage is observed for each worker: the union wage for union workers, and the nonunion wage for nonunion workers. If both wages were observed for workers, then calculation of Δ would be straightforward. But workers make choices regarding whether to seek union or nonunion employment, based in part on the wage they would receive in each sector. Thus, union status is endogenous to wages, and the result is a classic selection bias problem that has been treated extensively in the literature. In the presence of this selection bias, the value of Δ estimated from a standard cross-section of workers can be interpreted as the average difference in wages between union and nonunion workers, but it cannot be interpreted as the effect of union membership on the wage of a particular worker.

1.2 Estimating the Union Wage Gap

Trade unions have been characterized variously as organizations whose operational goals have been to 'take wages out of competition,' to impose a 'standard rate,' and to attach wages to jobs rather than to workers. This characterization has important implications for who will be union members and, hence, for the estimation of the effect of union membership on wages.

The standard approach to estimation of the union–nonunion wage gap using individual-level data is to estimate by OLS:

$$\ln W_i = X_i\beta + \delta U_i + \varepsilon_i \tag{7}$$

where i indexes individuals, X_i is a vector of worker characteristics including such skill measures as education and labor market experience, U_i is a dummy variable indicating union membership, and ε_i is a random component. The parameter δ represents the average proportional difference in wages between union and nonunion workers adjusted for measured differences in worker characteristics, and it is the regression adjusted analogue of Δ. Most estimates of δ are in the 0.1–0.25 range.

A central problem with this approach is that it constrains the marginal effect of skill on earnings to be the same in both union and nonunion jobs. Given that unions tend to 'standardize wages' and 'attach wages to jobs,' it is likely that the function characterizing the earnings of union workers will be flatter in skill dimensions than the function for nonunion workers. Specifying separate earnings functions for the two sectors yields:

$$\ln W_{ni} = X_i\beta_n + \varepsilon_{ni} \tag{8}$$

Table 1
OLS estimates of nonunion and union earnings functions 1999 US Current Population Survey, merged outgoinrotation groups (standard errors)

Variable	Nonunion (NU)	Union (U)	U − NU
Constant	1.3829	1.7927	0.4099
	(0.0082)	(0.0157)	(0.0197)
Age	0.0094	0.0088	−0.0006
	(0.0001)	(0.0003)	(0.0004)
ED < 12	−0.2535	−0.2194	0.0341
	(0.0055)	(0.0131)	(0.0159)
ED 13–15	0.1662	0.1148	−0.0514
	(0.0041)	(0.0080)	(0.0100)
ED ⩾ 16	0.5834	0.3616	−0.2218
	(0.0042)	(0.0085)	(0.0106)
Female	−0.1250	−0.1637	−0.0388
	(0.0049)	(0.0108)	(0.0132)
Married	0.2114	0.0958	−0.1157
	(0.0048)	(0.0091)	(0.0114)
Married*	−0.1650	−0.0840	0.0809
Female	(0.0065)	(0.0133)	(0.0164)
Nonwhite	−0.0817	−0.0609	0.0208
	(0.0045)	(0.0083)	(0.0104)
Public	−0.0677	−0.0177	0.0500
	(0.0051)	(0.0071)	(0.0095)
N	88556	17223	
R-Squared	0.350	0.217	
Root MSE	0.475	0.409	

Note: The base group consists of single white males with 12 years of education who are employed in the private sector. Estimates are weighted by the CPS final sampling weights. The third column is the difference between the second and the first columns.

for nonunion workers, and:

$$\ln W_{ui} = X_i \beta_u + \varepsilon_{ui} \tag{9}$$

for union workers. The usual modus operandi of unions implies that $\beta_u < \beta_n$ in skill dimensions such as education and experience. Since wages are tied less closely to individual characteristics in the union sector, and since the X_i vector consists largely of individual characteristics, the R^2 from cross-sectional estimation is generally smaller for the union earnings function than for the nonunion earnings function. It is also the case that the residual variation in earnings is smaller in the union sector than in the nonunion sector, again likely the result of standardization. Bloch and Kuskin (1978) were among the first to estimate separate union and nonunion earnings functions, and to note the systematic differences. Separate earnings functions imply that the estimated union–nonunion wage gap varies with worker characteristics such that the estimated union wage gap is declining in skill.

Ordinary least squares estimates of nonunion and union earnings functions, such as Eqns. (8) and (9) provide evidence consistent with the standardization-of-rates view of union wage setting. Table 1 contains estimates of nonunion and union log-earnings functions based on data from the outgoing rotation groups of the 1999 US Current Population Survey. These regressions control for age, education, sex, marital status, the interaction of sex and marital status, and sector of employment (private vs. public). The general character of these estimates is consistent with the large number of similar estimates that have appeared in the literature since the 1970s.

With respect to education, the most important skill measure, the college–high school wage differential is 22 log points (25 percentage points) larger in the nonunion sector than in the union sector. Thus, skill is rewarded more in the nonunion sector, and the union wage gap is considerably larger for less-skilled workers. This is a consistent finding in the literature on the union wage gap.

Additionally, other personal characteristics are of lesser importance in the union sector. Nonwhites in nonunion jobs earn about 8.2 percent less than whites, while nonwhites in union jobs earn about 6.1 percent less than whites. This confirms the earlier work of Ashenfelter (1972, 1978) that the union wage gap is larger for nonwhites than for whites. Married males in the nonunion sector earn 21 log points (23.5 percent) more than single males, while the marital status differential is only 9.5 log points (9.9 percent) in the union sector.

One apparently anomalous result is that the sex differential in wages is *larger* in the union sector than in the nonunion sector. This is due to the fact that females are substantially segregated in particular occupations, both within and outside the union sector.

The union wage gap varies with other measures of job and worker characteristics that are not included in the estimates in Table 1. The union wage gap differs substantially across occupations, with relatively large wage gaps in the less-skilled occupations. This is a manifestation of the organization of unions in the United States along industrial lines, and the standardization of wage rates in the union sector. It also appears that the union wage gap varies inversely with firm size. Workers in small firms receive a large wage advantage from unionization, while workers in larger firms receive a much smaller advantage. Given that unions are more common in large firms than in small firms, this might reflect the well-known firm-size wage differential where large firms pay higher wages than small firms (Brown and Medoff 1989).

Overall, existing estimates suggest a substantial positive union wage gap for most workers. The raw (not regression adjusted) union wage gap in the data used in Table 1 is 28.5 log points (33 percentage points). The wage gap for a single 40-year-old white male high school graduate predicted by the estimates in Table 1 is 38.6 log points (47 percentage points). This declines sharply as education increases, falling to 16.4 log points (17.8 percentage points) for an otherwise-similar college graduate, but it remains

substantial and positive. The wage gaps are smaller for otherwise-similar married males, falling to 27 log points (31 percentage points) for married high-school graduates, and to 4.8 log points (4.9 percentage points) for married college graduates.

Other evidence consistent with the standardization-of-rates view is that regression mean squared error is larger in the nonunion sector than in the union sector. This suggests that unobservable characteristics of workers are more important in wage determination in the nonunion sector than in the union sector. It is also the case that the R^2 of the estimated earnings function is considerably larger in the nonunion sector than in the union sector, which is also consistent with the view that observable individual characteristics are a relatively more important determinant of wages in the nonunion sector than in the union sector.

The standardization of wages in the union sector, and the concomitant lower variation in wages in the union sector implies that unions reduce the overall dispersion of wages. A force for increased dispersion as a result of unions is the potential decline in the nonunion wage as the union wage increases and the supply of labor to nonunion jobs increases. Freeman (1980a) analyzed the effect of unions on the dispersion of wages, and he concluded that, on balance, unions reduce the aggregate level of wage dispersion. Card (2001) finds that 15 to 20 percent of the increase in male wage inequality in the United States between 1973 and 1993 can be accounted for by the decline in union membership.

The substantial variation in the union wage gap, particularly by worker skill level, has important implications for worker preferences for union representation, and for employer hiring decisions. A reasonable inference, made explicit in the next section, is that the union status of workers is related to skill, both observable and unobservable. This implies that union status is endogenous to the determination of wages, so that OLS estimates of the union wage gap are likely to suffer from selection bias.

1.3 The Non-wage Effects of Unions

In addition to the work of the effect of unions on wages, there has also been substantial research on the effect of unions on non-wage aspects of employment. This literature is surveyed by Freeman and Medoff (1984), and by Hirsch and Addison (1986). Some of the findings are:

(a) union members receive a larger fraction of their compensation in the form of fringe benefits such as health insurance and pensions (Freeman 1981);

(b) workers in union jobs are less likely to quit, reflecting the higher level of wages and benefits. Freeman (1980b) suggests that this is due to union provision of a mechanism for 'voice' as an alternative to 'exit' in addressing workplace concerns; and

(c) workers in union jobs are more likely to be laid off, reflecting the fact that union employers cannot adjust wages in the short run in response to demand shocks (Medoff 1979).

2. The Determination of Union Status of Workers

There are two components to the determination of union status of workers. First, workplaces must become organized by trade unions. In the United States, the presumption is that workplaces are not unionized, and it generally takes a substantial effort on the part of workers and unions to organize the workplace. However, once the workplace is organized and a contract is negotiated, the terms and conditions of employment of all workers, present and future, are determined through negotiation between the union and the employer. This organization process effectively sets the stock of union jobs in the economy, and it is largely determined by past organization activities.

The second component of the determination of a worker's union status is the individual decision regarding whether to seek a job in a union or nonunion workplace. Taking a job in a unionized workplace is essentially costless (other than the payment of a modest initiation fee and the regular payment of dues), and does not involve any investment in organization.

2.1 The Worker-choice Model

Early empirical analyses of the union status of individual workers (e.g., Lee 1978) relied on a model (the worker-choice model) where the stock of union jobs was not taken as a binding constraint, and where individual workers compared their utilities in union and nonunion employment and selected the sector that yielded the highest utility. In order to focus the discussion, consider a worker-choice model where unions affect only the wage received by workers but where the effect on the wage can differ across workers. Let each worker i have a nonunion wage of W_{ni}, and a union wage of W_{ui}, so that the wage advantage to workers i of being in the union is $V_i = W_{ui} - W_{ni}$. Given a cost of union membership of C_i, worker i is unionized if $V_i > C_i$, and is nonunion otherwise.

This simple model has direct implications for the composition of the union sector. Abstracting from differences across workers other than in skill, suppose that the earnings of worker i in a nonunion job are defined by:

$$\ln W_{ni} = \beta_0 + \beta_1 S_i + \beta_2 \mu_i + \varepsilon_{ni} \qquad (10)$$

and that the earnings of worker i in a union job are defined by:

$$\ln W_{ui} = \gamma_0 + \gamma_1 S_i + \gamma_2 \mu_i + \varepsilon_{ui} \tag{11}$$

where S_i is an observable measure of skill, μ_i is an unobservable measure of skill with mean zero, and the εs are random components. The βs and γs are parameters. The worker-choice model of union status determination implies, using Eqns. (10) and (11), that workers are unionized if, and only if,

$$V_i = (\gamma_0 - \beta_0) + (\gamma_1 - \beta_1) S_i + (\gamma_2 - \beta_2)\mu_i + (\varepsilon_{ui} - \varepsilon_{ni}) > C_i. \tag{12}$$

To the extent that unions raise the wages of at least low-skilled workers, it is the case that $\gamma_0 > \beta_0$. The characterization of unions as standardizing pay implies that the marginal effect of both observed and unobserved skill on wages is likely to be greater in the nonunion sector than in the union sector ($\beta_1 > \gamma_1$ and $\beta_2 > \gamma_2$). The clear implication of the worker-choice model is that the union sector will be composed disproportionately of less-skilled workers. Consistent with this, Farber and Saks (1980) find that the likelihood that a nonunion worker will desire union representation is related inversely to the worker's position in the earnings distribution of the potential bargaining unit being organized.

The worker-choice model of union status determination implies a standard binomial discrete choice specification, such as a linear probability, probit, or logit model. While Lee (1978) made the structural formulation explicit, most others who have estimated binomial discrete choice models of union status determination are not clear about the foundations and consider the specification a reduced form. Hirsch and Addison (1986) attempt to interpret a variety of estimates of models such as this in the context of the worker-choice model.

2.2 The Queuing Model

A more realistic model of the determination of the union status of workers takes account of the fact that not all workers who desire union employment can find union jobs. This is because, as noted above, workplaces are costly to organize, but, once organized, remain so into the indefinite future. The result is that there will be workers who would desire employment in an already-unionized workplace but who would not be interested in undertaking organization themselves.

Denote the cost of organizing an unorganized workplace by C_o and the cost of taking a job in a existing unionized establishment by C_e so that $C_o > C_e$. Then a worker will desire a job in a unionized establishment when $V_i > C_e$, but a worker will be willing to invest in organization activity only when $V_i > C_o$. Since the organization decision is what determines the stock of union jobs, there will be workers who would like an existing union job but who are unwilling to invest in organization themselves. These are the workers for whom $C_e < V_i < C_o$. The result is that there will be more workers who desire an existing union job than there are union jobs available. Thus, there is likely to be excess demand by workers for existing union jobs, resulting in an implicit queue. Aside from some unions of skilled trades workers, union employers in the United States are free to hire workers of their choosing to fill vacancies. Thus, the union status of a given worker is the result of decisions by both the worker and union employers (Abowd and Farber 1982, Farber 1983a).

The queuing model also has important implications for estimating models of the union status of workers. A worker will be in a union job only if the worker prefers union employment to nonunion employment *and* a union employer hires the worker. As in the worker-choice model, the most skilled workers are less likely to want a union job. But the standardization of wages reduces the marginal cost of skill to employers, so that union employers will want to hire the most skilled workers from among those workers who desire a union job. Thus, union workers will be those in the middle of the skill distribution. The most skilled will not want union jobs, and the least skilled will not be hired by union employers. This selection process could account for at least part of Freeman's (1980a) finding that there is less dispersion of earnings in the union sector.

Empirical implementation of the queuing model requires a bivariate discrete choice model in order to capture both employer and worker decisions. Identification is a potential problem because it is not known *a priori* whether nonunion workers are nonunion because they did not desire a union job, or because they were not hired by a union employer. Farber (1983a), Farber (1990), and Farber and Krueger (1993) use several US surveys on preferences of nonunion workers for union representation to provide this information.

2.3 Implications for Estimation of the Union Wage Gap

Regardless of which model of union status determination is appropriate, applying standard OLS techniques to the estimation of the union wage gap using either Eqn. (7) or Eqns. (8) and (9) will yield biased estimates of the key parameters. Both models imply that workers' union status is determined at least partly on the basis of unobservable skills (μ_i), so that there will be classic 'selection bias.'

In the context of the worker-choice model, less-skilled workers are more likely to be union members, so that workers in union jobs are more likely to be those with low values of μ_i relative to workers in

nonunion jobs. The result is that OLS will yield a downward-biased estimate of the effect of unions on the wage of a particular worker.

The effects of the double selection implied by the queuing model on OLS estimates of the union wage gap is not clear. The two selection processes appear to have offsetting effects on the estimated wage gap, with the worker selection implying a negative bias, and the employer selection implying a positive bias.

Two general approaches have been used to address the problem of selection bias in estimating the union wage gap. The first approach is to model union status determination simultaneously with earnings, and to estimate an econometric model that takes account of the simultaneity. However, this requires strong assumptions regarding functional form of the distribution of the errors in the earnings and union status determination functions. Virtually all existing work proceeds using a two-step estimator where the earnings functions and the union status determination function are assumed to have errors that are jointly normal. Lewis (1986) surveyed most of the studies that used this general approach, and he concluded that the specific results are unreliable and depend heavily on arbitrary assumptions regarding functional form and untestable exclusion restrictions (variables assumed to affect union status but having no direct effect on earnings).

The second approach relies on repeated observations on wages for individuals who change union status (Freeman 1984). Thus, both union and nonunion wages are observed for these workers, and the difference in wages for these individuals can be thought of as their union wage gap. However, this approach relies on the strong assumption that the probability of switching sectors is not related to unobservable skill.

Another problem with the use of longitudinal data in the context of difference or fixed effects models is that there is error in the classification of workers as union or nonunion. Card (1996), based on data from the January 1977 Current Population Survey (Mellow and Sider 1983) estimates that the misclassification rate of union status in the CPS is approximately 2.7 percent. Even this relatively low rate of misclassification implies that the number of changes in union status will be overstated substantially. It is well known that measurement error is a more serious problem in estimating fixed effect or difference models than in estimating standard levels models. The result is that estimates derived using longitudinal data can be subject to substantial attenuation bias (bias toward zero). Even relatively low rates of misclassification of the key union status change variable imply that estimates of the union wage gap derived using fixed effect or difference models are likely to be biased downward substantially. Lewis (1986) presents a full discussion and analysis of this problem. He finds that the estimates derived using these techniques are substantially smaller than those derived using cross-section data, and he attributes this to measurement error bias.

See also: Class: Social; Comparable Worth in Gender Studies; Dispute Resolution in Economics; Income Distribution; Industrial Relations and Collective Bargaining; Industrial Relations, History of; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Labor Movements, History of; Labor Supply; Labor Unions; Sex Differences in Pay; Trade Unions, Economic Behavior of; Wage Differentials and Structure; Work, History of; Work, Sociology of

Bibliography

Abowd J M, Farber H S 1982 Job queues and the union status of workers. *Industrial Labor Relations Review* **35**: 354–67

Ashenfelter O 1972 Racial discrimination and trade unions. *Journal of Political Economy* **80**: 435–64

Ashenfelter O 1978 Union relative wage effects: New evidence and survey of their implications for wage inflation. In: Stone R, Peterson W (eds.) *Econometric Contributions to Public Policy*. St. Martin's Press, New York

Bloch F E, Kuskin M S 1978 Wage determination in the union and nonunion sectors. *Industrial Labour Relations Review* **31**: 183–92

Booth A L 1995 *The Economics of the Trade Union*. Cambridge University Press, Cambridge, UK

Brown C, Medoff J L 1989 The employer size–wage effect. *Journal of Political Economy* **97**: 1027–59

Card D 1996 The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica* **64**: 957–99

Card D 2001 The effects of unions on wage inequality in the United States. *Industrial Labour Relations Review* **54**: 329–48

Farber H S 1983a The determination of the union status of workers. *Econometrica* **51**: 1417–38

Farber H S 1983b Worker preferences for union representation. In: Reid J D Jr. (ed.) *New Approaches to Labor Unions*. JAI Press, Greenwich, CT

Farber H S 1990 The decline of unionization in the United States: What can be learned from recent experience? *J. Lab. Econ.* **8**: S75–S105

Farber H S, Krueger A B 1993 Union membership in the United States: The decline continues. In: Kaufman B, Kleiner M (eds.) *Employee Representation: Alternatives and Future Direction*. Industrial Relations Research Association, Madison, WI

Farber H S, Saks D H 1980 Why workers want unions: The role of relative wages and job characteristics. *Journal of Political Economy* **88**: 349–69

Freeman R B 1980a Unionism and the dispersion of wages. *Industrial Labor Relations Review* **34**: 3–23

Freeman R B 1980b The exit voice tradeoff in the labor market: Unionism, job tenure, quits, and separations. *Q. J. Econ.* **94**: 643–74

Freeman R B 1981 The effect of unionism on fringe benefits. *Industrial Labour Relations Review* **34**: 489–509

Freeman R B 1984 Longitudinal analyses of the effects of trade unions. *J. Lab. Econ.* **2**: 1–26

Freeman R B, Medoff J L 1981 The impact of the percent organized on union and nonunion wages. *Rev. Econ. Stat.* **63** 561–72
Freeman R B, Medoff J L 1984 *What Do Unions Do?* Basic Books, New York
Hirsch B T, Addison J T 1986 *The Economic Analysis of Unions: New Approaches and Evidence*. Allen & Unwin, Boston, MA
Lee L 1978 Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *Intl. Econ. Rev.* **19**: 415–34
Lewis H G 1963 *Unionism and Relative Wages in the United States: An Empirical Inquiry*. University of Chicago Press, Chicago
Lewis H G 1986 *Union Relative Wage Effects: A Survey*. University of Chicago Press, Chicago
Medoff J L 1979 Layoffs and alternatives under trade unions in US manufacturing. *American Economic Review* **69**: 380–95
Mellow W, Sider H 1983 Accuracy of response in labor market surveys: Evidence and implications. *J. Lab. Econ.* **1**: 331–44
Rosen S 1969 Trade union power threat effects and the extent of organization. *Review of Economic Studies* **36**: 185–96
Tracy J S 1986 An investigation into the determinants of US strike activity. *American Economic Review* **76**: 423–36

H. S. Farber

Tradition, Anthropology of

Few concepts have played a more central role in the development and practice of anthropology than tradition. We will first outline the principal senses in which tradition figures in anthropological thought, then trace the historical foundations of these contemporary usages, and conclude with the major lines of inquiry centering on tradition in contemporary anthropology.

1. *Tradition as Substance, Process, Practice*

In its most general usage, tradition may be defined as the element of historical continuity or social inheritance in culture, or the social process by which such continuity is achieved. More substantively, tradition refers to the collective social inheritance of particular people, culture, society, group, or collectivity, and as such, stands as the referent of its collective identity. In this sense, tradition is generally conceived of as the aggregate mass of cultural forms—cast as knowledge, beliefs, practical routines, or texts—that are the constituent elements of a culture, handed down from one generation to the next. One may speak also of 'a tradition' as any discrete element or trait within the general mass. Here, tradition takes on the sense of an inherent quality of the past which is diagnostic of traditionality; hence, 'a traditional practice,' 'a traditional belief,' 'a traditional tale,' and so on. So conceived, a tradition assumes the guise of a durable natural object, passed down through time by successive generations of 'culture bearers.'

A more concrete and practice-centered understanding of tradition recognizes that the temporal continuity of tradition is instantiated in successive enactments of cultural forms that are guided by precedent and convention and construed as replicating, in some essential sense, what has been done before. When the focus is on the dynamics of tradition, the process of handing down, or intergenerational transmission, the element of social learning is often emphasized, whether under the rubric of enculturation, which foregrounds the acquisition of cultural knowledge, or the now preferred socialization, which foregrounds the social matrix of the acquisition process.

The conjoint consideration of socialization and tradition tends to foster a conception of intergenerational transmission as the replication of uniformity, in Anthony F. C. Wallace's (1970) apt phrase. This orientation points to what is perhaps the most common theme in anthropological considerations of tradition, that is, its normative force, the notion that tradition enjoins conformity. The binding force of tradition is variously attributed to the force of precedent or habit inherent in the routinized quality of replicated practice, the psychological pressure attendant upon deviating from accustomed modes of thought and action, the moral weight of public opinion, and—most strongly—the power of a coercive apparatus of external sanctions sustained by sacred mandate (as from the ancestors) and intensified by ritual enactment. Where such discriminations are made systematically, they are generally marked by terminological distinctions, as between custom and tradition, with tradition reserved for the most strongly sanctioned and ritualized elements of the social inheritance. In general, however, there is little definitional rigor or regularity of terminological usage in the field; as often as not, terms such as tradition, custom, or folkways are employed as loosely equivalent. In such systematization, again, tradition may be distinguished from law, the former deriving its force from 'the sanctity of immemorial traditions,' the latter from 'the legality of enacted rules,' in Max Weber's classic formulation. Others, however, such as Bronislaw Malinowski, take custom and tradition as more or less equivalent, but argue that certain traditions, those surrounded by supernatural sanctions and sustained by coercive machinery, are the functional equivalent of law in 'primitive' societies.

Perspectives also vary in regard to how they frame the locus of traditional authority. One line of inquiry focuses on epistemological issues, emphasizing the binding force of traditional modes of thought, characterized by a purported resistance to or incapacity for rational, reflective, critical, or skeptical stances toward knowledge and belief. In a related vein, the emphasis may fall on the behavioral correlates of such epis-

temological conservatism, manifested in traditional patterns of 'tried and true,' naturalized modes of action and in ritual. A third line maintains a discursive focus, concentrating on authoritative discourse; here, authority is vested in the weight of ancestral and collective truth expressed in oft-repeated, socially resonant texts, such as proverbs, myths, or legends. Finally, we may identify a sociopolitical frame of reference, which is primarily concerned with traditional authority as vested in particular statuses or roles, such as elders (by virtue of their accumulated knowledge and the historical reach of their experience into the past) or ritual specialists (who control the coercive machinery of supernatural sanctions), whose claims to authority rest upon their connection to an authorizing past. These distinctions are analytical; much work, both theoretical and substantive, explores dimensions of inter-relationship among these loci of traditional authority.

2. *Tradition and the Construction of Modernity*

The roots of the anthropological preoccupation with tradition lie in the seventeenth and eighteenth centuries, in the epistemological challenges of enlightenment empiricism, the religious overturning of the Protestant Reformation, and the interpretive explorations of classical and biblical philology. Enlightenment philosophers, such as Bacon and Locke in the UK and Condorcet in France, insisted that the rational, empirical pursuit of true knowledge required a throwing off of traditional authority, which demands the surrender of one's capacity for independent thought and renders one incapable of making new contributions to knowledge.

Baconian natural philosophy and the concomitant enlightenment ideology of progress were foundational to the development of antiquarianism in the late seventeenth and eighteenth centuries. Antiquities were the etiolated remnants of past ways of thinking, acting, and conveying knowledge that were 'derived down' uncritically into the present, shorn of their original meaning and functional coherence. In this sense, the notion of antiquities anticipates E. B. Tylor's concept of survivals. Only by rigorous historical investigation of their origins could they be fully revealed as irrational survivals of 'old customs' and expunged from modern life. For Protestant reformers, the remnants of old 'heathen' and 'Romish' 'superstitions,' handed down to the credulous members of the populace by 'custom,' 'tradition,' and 'hearsay,' were likewise to be revealed by historical investigation and extirpated from modern life. There was a clear sociological component to the antiquaries' and reformers' antiquarian programs as well, identifying the 'vulgar' classes, country people, the elderly, the uneducated, and women (as in 'old wives' tales') as the segments of the population among whom the survivals of past error were most likely to persist, thus, in effect, constructing a set of social 'Others' within contemporary society who were to be the special targets of tradition-oriented antiquarian investigation.

The development of eighteenth-century classical and biblical philology was stimulated by the literary battles of the Ancients and the Moderns, contesting the proper basis of contemporary literary production. This line of inquiry had a textual and interpretive focus: to render ancient and exotic texts, like the Homeric epics and the Hebrew Old Testament, intelligible, coherent, and meaningful to contemporary readers and to secure their canonical status and value. This interpretive approach combined a kind of proto-relativism, reading the texts as shaped by the social, cultural, and linguistic milieu of ancient Greece or Israel—that is, as the inspired expressions of a nation—with a species of conjectural history, in terms of which the ancient texts were characteristic expressions of an early stage in the development of human society. One of the most significant methodological consequences of this approach was to motivate the use of data from living cultures—chiefly peoples of 'the Orient,' such as Arabs, Turks, East Indians—viewed as developmentally behind their modern European contemporaries, to illuminate the archaic cultures reflected in the Homeric epics and the Old Testament. These peoples of the East, then, took on the guise of exotic 'Others,' living exemplars of a 'rude' and 'primeval' age, bound by tradition rather than by an ideology of progress, reliant on oral communication rather than literacy, credulous and emotional rather than rational, and so on.

In a sweeping synthesis and extension of the philologists' program and a critical counterstatement to the social contract theories of enlightenment political philosophers, the German political philosopher, Johann Gottfried Herder argued that the authentic foundation of a viable polity is the poetic tradition of its people, made up of the inspired expressions of its poets—who are also its 'lawgivers'—who give voice to the spirit of the Volk, rooted in its particular time and place. This intertextually constituted tradition became 'the archive of the people,' the principal expression of national identity, the mechanism of its cultural cohesion and continuity, and the charter for its social and political life. In the most general sense of the term, Volk designates a nation, a people in its organic wholeness, but it may also designate that portion of a more complex, stratified society that remains grounded in its inherited language and traditions and still open to poetic feeling, as distinct from those who have been distanced from their traditions and feelings by over-rational refinement, the alienation of writing, or the adoption of foreign languages and alien ways. In Herder's (see Bauman and Briggs 2000) intellectual synthesis, then, a precursor to Boasian anthropology, tradition is most powerfully operative in simpler, premodern, community-based societies and in those

sectors of complex modern society in which a sense of community cohesion, the emotionally resonant ties of locality, kinship, and attunement to the national spirit still prevail.

Herder's thought was influential in the emergence of Romanticism, which was also inflected by the violent upheavals of the French Revolution. One response to the Revolution took the form of a conservative reaction, in which the sanctity of tradition and the cohesive force of kinship, religion, customary law, rootedness in place, and traditionally founded hierarchy were valorized as the essence of an authentic national character and a viable and moral polity. What is noteworthy, however, is that traditionalist conservatives tended to see the trajectory of change in the same terms as the rational reformers, framing an oppositional contrast between tradition and modernity. This contrast, convergent with jurisprudential investigations and symbolic constructions of 'customary law,' became the core of the classic typological tradition in social thought out of which emerged the modern social disciplines, including anthropology, in which the force of tradition served as a defining criterion of social types.

By the latter half of the nineteenth century, the foundational ideas we have summarized coalesced in a series of typological syntheses aimed at comprehending the broad historical trajectory of human social and cultural development and at providing a classificatory framework of social formations based upon stages of evolutionary progress. Whether framed in terms of ideal types, developmental continua, or evolutionary stages, these magisterial syntheses all accorded a defining or diagnostic role to tradition in the construction of contrastive social types. E. B. Tylor (1960 [1881]), for example, in contrasting 'savage' and 'civilized' man, ascribes the former's intellectual conservatism and aversion to change to his 'tendency … to consider his ancestors as having handed down to him the perfection of wisdom.' Ferdinand Tönnies (1988 [1887]), in basing his classic distinction between Gemeinschaft and Gesellschaft on contrastive forms of common will, characterizes the 'natural will' that predominates in Gemeinschaft societies as becoming naturalized 'through practice and tradition.' Emile Durkheim, in outlining the contrast between mechanical and organic solidarity, attributes the authority of the collective conscience in mechanically solidarity societies to 'the authority of tradition.' Max Weber, in constructing his three pure types of legitimate domination, maintains that 'the validity of a social order by virtue of the sacredness of tradition is the oldest and most universal type of legitimacy.' These and related formulations served to crystallize the notion of 'traditional society' as a typological category, intellectually and behaviorally conservative, bound by an unreflective and unquestioning adherence to traditional ways and submission to traditional authority, and slow to change, in contrast to 'modern society,' dynamic, impelled by the exercise of reason and rationally motivated innovation, and open to change. This typological contrast has proven remarkably durable, persisting to the present day as an orienting framework in the social disciplines, not least in anthropology. *Mutatis mutandis*, it underlies Robert Redfield's folk-urban continuum, Karl Popper's closed versus open society, Talcott Parsons' pattern variables, Lévi-Strauss' cold vs. hot society, and a host of other variations on the theme, persisting in the face of trenchant criticisms of its resistance to operational specificity, its lack of analytical utility in the investigation of empirical cases, its masking of the dynamics of tradition in so-called 'modern' society, and its role in sustaining structures of social inequality and policies of domination.

3. *Tradition in Contemporary Anthropology*

The conceptual and critical reach of tradition in contemporary anthropology is pervasive, especially as it implicates the element of continuity in culture. We may, however, outline several foci of empirical, conceptual, and critical interest in which the notion of tradition figures relatively explicitly.

3.1 *The Social Organization of Tradition*

One prominent line of inquiry concentrates on what we might call the social organization of tradition, addressing those aspects of the typological model that cast traditional societies as homogeneous or undifferentiated, with the implication that tradition is collectively shared within the group. This understanding may be called into question in several ways, most directly by the recognition within the group of cultural specialists, who are charged with or have an individual bent toward the cultivation, formulation, or communication of tradition, as ritual specialists, elders, reciters, or other custodians of expert or esoteric knowledge. The foci of investigation here include the definition and organization of specialist roles; the recruitment, socialization and careers of cultural specialists; and the institutional and situational context in which cultural specialists act in their expert capacities.

This latter focus may direct attention to cultural performances or display events, such as rituals, festivals, contests, or dramatic performances in which the knowledge, symbols, and values of the group are embodied, enacted, and placed on public display. The production of such events typically requires the efforts of specialists in various capacities, and an organization of production by which the performance is accomplished. These lines of inquiry have been especially prominent in studies framed in terms of the distinction between the Great Tradition (cf. 'high,' 'classic,'

'cultivated') and the Little Tradition (cf. 'low,' 'folk,' 'popular') developed by Robert Redfield.

Investigation of the social organization of tradition may converge as well with the study of socialization and enculturation: the roles, contexts, and practices by which knowledge and behavioral patterns are conveyed to, and acquired by, succeeding generations. Finally, it is worth noting that the social organization of tradition has critical implications for anthropological method, insofar as cultural specialists figure prominently in ethnographic work as favored sources of data, and that ethnographic encounters, organized around the getting and giving of knowledge, are frequently framed as pedagogical events.

3.2 Tradition and Creativity

A further variation on the categorical opposition between tradition and modernity is the contrast between tradition and creativity, especially prominent in the anthropology of the arts. Classic work in this area, such as Franz Boas's *Primitive Art* (1955 [1927]), tended to view tradition as a constraint on the 'primitive' artist, consistent with the typological conception of traditional society as inherently conservative and Herderian notions of artistic forms as the collective expression of a national spirit. Thus the convention of collective attribution that has dominated the presentation (with a few notable exceptions) of verbal and material art forms from 'primitive' or 'traditional', or 'folk' societies (a Navajo chant, a French folktale, a Yoruba staff).

Since the 1960s, however, there has been an accelerating shift in the anthropology of the arts toward greater attention to the role of the individual artist/performer, and more nuanced investigation of the interplay between convention and creativity in artistic practice. In the anthropological study of the material arts, the distinctive styles and virtuosic achievements of individual artists and the situated practice of artistic production, use, and evaluation have aided significantly in the recalibration of the balance between tradition and creativity, and even further to a re-evaluation of what constitutes creativity, no longer restricted to novelty but understood to include the use of conventional forms in new ways.

There has been a parallel development in the study of the verbal arts, stimulated by Milman Parry and Albert Lord's pioneering studies of the composition in performance of Serbo-Croatian oral epic and by the correlative turn to the ethnography of performance by linguistic anthropologists and folklorists. Central to these investigations is a focus on the situated nature of performance as a mode of communicative practice and recognition of the emergent quality of performance. In linguistic anthropology especially, the ethnography of performance has incorporated the close analysis of formal, functional, and intertextual relationships as a vantage point on the calibration in performance of conventional, 'traditional' orienting frameworks for the production and interpretation of discourse on the one hand, and the exigent and emergent 'creative' qualities of situated performance on the other. Both qualities are always present, and the task becomes one of determining the relative weight and the dynamics of interplay of the two in particular social and historical circumstances.

3.3 Tradition and Communicative Technologies

Closely related to the typological contrast between traditional and modern societies is the distinction between oral and literate societies, which foregrounds specific technologies of communication as diagnostic—even determinative—of the contrastive types. The typological uses of orality and literacy have their roots in eighteenth-century philology, which in turn had a shaping influence on anthropology. The twentieth century revival of this framework by classical scholars has once again infiltrated anthropology and re-energized debates concerning the validity of 'Great Divide' typologies. Stated in very general terms, the 'literacy hypothesis' attributes to the advent of writing, the invention of print technologies, and the associated spread of literacy, a concomitant transformation of modes of thought, specifically the objectification and decontextualization of knowledge, and the encouragement of a critical or skeptical orientation toward received truth. These intellectual (or cognitive) orientations, according to proponents of the literacy hypothesis, had a transformative influence on society and culture, effecting an epochal shift from a blind and unreflective adherence to traditional modes of thought and practice to the rational, scientific openness to change that is diagnostic of modernity.

While some anthropologists have endeavored to find ethnographic corroboration of the typological contrast between traditional and modern modes of thought, the weight of ethnographic evidence has tended to call the literacy hypothesis into question. The most effective and persuasive critiques have come from linguistic anthropology and related work in the anthropology of teaching and learning. A large and growing body of ethnographic studies, focusing on speaking, writing, reading, and the acquisition of communicative competence as social practices embedded in situational and institutional contexts of use, has provided a range of critical correctives to unitary conceptions of orality and literacy studied in relation to abstract systems of thought. These lines of research, illuminating the role of communicative technologies, forms, and practices in the production and reproduction of social life, offer as well an increasingly ramified and nuance understanding of the dynamics of

continuity and change in human society, to supplant the classic categorical distinction between tradition and modernity.

3.4 The Politics of Tradition

The political force of tradition has been a consistent theme in anthropology, built into the very foundations of the discipline and epitomized in Weber's identification of traditional authority as one of the three pure types of legitimate domination. Oriented by classic conceptions of traditional authority and 'traditional' society more generally, foundational studies of traditional authority in anthropology have tended largely to view it as a unifying force in the service of social system maintenance, sustained by the mechanisms of socialization, the mythological charter of institutions, and the rigid formalization of ritual and traditional oratory.

Here again, however, the ethnographic turn to situated practice, and especially to discursive interaction and performance, has led to recognition of the ways that traditional tales, for example, may be instruments of divisiveness rather than unity, insofar as they may be performed in defense of competing claims to status, or that the formal regimentation that marks the performance of oratory as 'the words of the ancestors' may serve to enhance the rhetorical power of disputants in political conflict. That is to say that rhetorical appeals to tradition and the public ideologies of traditional authority that motivate them do not necessarily have the binding force in practice that participants and observers alike may claim for them. Ideologies of traditional authority—like all ideologies—are positioned, interested, and strategically invoked, and tradition as a rhetorical resource may serve as well for conflict as for the 'binding force' of acquiescence.

The ideology and rhetoric of tradition are understandably more apparent under conditions of marked social and cultural change; indeed, the capacity of 'traditional' peoples consciously to intervene in the management of tradition was recognized earliest in anthropology in the study of nativistic movements, 'any conscious, organized attempt on the part of a society's members to revive or perpetuate selected aspects of its culture.' (see Linton 1943) This work anticipated a vigorous line of anthropological inquiry that has burgeoned since the 1980s, devoted to the investigation of tradition and related notions of authenticity as symbolic, interpretive constructions in and of the present through the invocation of links to a meaningful past.

The process of traditionalization—under the rubric of tradition, heritage, patrimony, custom, or the like—is a prominent resource in the service of nationalism, whether cultural nationalism or nation-state formation, and it is a conspicuous aspect of the politics of culture in many parts of the world in which there is an extended history of anthropological research but which are now increasingly determined to shed the vestiges of a colonial past and assert the integrity and authenticity of their own 'cultural traditions.' The readiness of anthropologists to identify the tradition or custom, or patrimony on which such claims stand as selective, strategic, symbolic constructions is a source of tension both within the discipline and between anthropologists and their ethnographic interlocutors: whose interpretations and constructions are to be accepted as authoritative?

This question and the further problems it raises afford an illuminating critical vantagepoint on the historical formation of anthropological theory and practice. The modern conception of tradition was shaped in an intellectual arena that was defined largely in political terms. What are the foundations of a viable polity? A legitimate authority? A national culture? What are the bases of social inequality? Who is authorized to intervene in the management of these formations? Tradition and allied conceptions of culture proved to be foundational to the emergent discipline of anthropology and the definition of its purview. A measure of the success of anthropology may be strength with which its core concepts have come to be accepted by the formerly 'traditional' peoples who have been its central objects of study. One of the challenges that now confronts anthropology is that in asserting their own subjectivity, these peoples should base their claims on those foundational concepts just as anthropologists themselves are increasingly inclined to deconstruct them, to recognize traditionalization as an interpretive process, to challenge the reification of tradition, or to unmask primordial traditions as recent inventions.

See also: Cultural History; Emotions, History of; Social Change: Types; Traditions in Sociology; Traditions: Social

Bibliography

Bauman R, Briggs C L 2000 Language philosophy as language ideology: John Locke and Johann Gottfried Herder. In: Kroskrity P V (ed.) *Regimes of Language: Ideologies, Polities, and Identities*. School of American Research Press, Santa Fe, NM

Ben-Amos D 1984 Seven strands of tradition: varieties in its meaning in American folklore studies. *Journal of Folklore Research* **21**: 97–131

Biebuyck D P (ed.) 1969 *Tradition and Creativity in Tribal Art*. University of California Press, Berkeley, CA

Briggs C L 1996 The politics of discursive authority in research on the 'invention of tradition'. *Cultural Anthropology* **11**: 435–69

Boas F 1955 [1927] *Primitive Art*. Dover Publications, New York

Collins J 1995 Literacy and literacies. *Annual Review of Anthropology* **24**: 75–93

Finnegan R 1988 *Literacy and Orality: Studies in the Technology of Communication*. Blackwell, Oxford, UK
Finnegan R 1990 Tradition, but what tradition and for whom? *Oral Tradition* **6**: 104–24
Firth R 1961 *History and Traditions of Tikopia*. Polynesia Society, Wellington, New Zealand
Handler R, Linnekin J 1984 Tradition, genuine or spurious. *Journal of American Folklore* **97**: 273–90
Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK
Jolly M, Thomas N (eds.) 1992 The politics of tradition in the Pacific. *Oceania* **62**(4) Special issue
Kelley D R 1990 'Second nature': the idea of custom in European law, society, and culture. In: Grafton A, Blair A (eds.) *The Transmission of Culture in Early Modern Europe*. University of Pennsylvania Press, Philadelphia
Linton R 1943 Nativistic movements. *American Anthropologist* **45**: 230–40
Shils E 1981 *Tradition*. University of Chicago Press, Chicago
Singer M B (ed.) 1959 *Traditional India: Structure and Change*. American Folklore Society, Philadelphia
Tönnies F 1988 [1887] *Community and Society*. Transaction Press, New Brunswick, NJ
Tylor E B 1960 [1881] *Anthropology*. University of Michigan Press, Ann Arbor, MI
Wallace A F C 1970 Culture and Personality, 2nd edn. Random House, New York

R. Bauman

Traditions in Sociology

Traditions are broadly a form of received authority giving weight to the past in how social actors in the present orient themselves and act upon their situation. They are ways of acting upon the world transmitted from one generation or cohort to the succeeding one. All societies, all institutions except for some brief charismatic moments when 'all things are born new' are marked to some extent by tradition. Adherence to a tradition provides group awareness and identity; this includes sociology as a general discipline as well as its more differentiated units imparting sociological training to students at different levels.

In the late twentieth century there has been increased recognition of traditions as framing the development of sociology. Indeed, one can speak of an emergent sociological historiography which takes 'tradition' as the central focus for providing accounts of the genealogy of the discipline in terms of, following Anderson (1991), 'imagined communities' of scholars. Representing the development of sociology in terms of one or a few major traditions has become a more analytically challenging approach than earlier listings of names, single topics, or philosophical orientations. Yet, while the invoking of traditions may seem to be the present's salute to the past, the past does not speak for itself 'as it really happened.' Traditions, like ceremonials, rituals, and even like nations are not handed down as inert 'rocks of the ages'; they are activated and reactivated by contemporaries, even 'invented' on occasions for quite novel purposes (Hobsbawm and Ranger 1992). Essentially, then, the topic of sociological traditions interweaves objective referents of the past with subjective constructions or perceptions of these referents.

1. Traditions and Modernity

Without discussing sociological traditions specifically, Shils (1973, 1981) has provided an extensive analysis of tradition in the modern setting, including the tradition of intellectuals. This may be seen in the form of a paradox. On the one hand, social scientists are attached to the concepts and figures that are rooted in the past; on the other hand, they are negligent if not outright unsympathetic to traditionality; major traditions recognized as authentic by the great majority are 'antitraditional intellectual traditions (Shils 1981). This paradox may be traced to the 'enlightenment philosophy' that made change and progress new basic values of modernity. Rational, scientific knowledge, promulgated above all by intellectuals, would replace irrational authority bound by the chains maintained by traditional institutions. This modernist orientation is a constitutive element of a bundle of traditions in sociology deriving from the 'enlightenment's' project of human capability to reconstruct or 'reinvent' the individual, society, and social institutions. Shils also points out that 'no tradition begins *ab novo*.Every tradition grows out of another'; moreover, intellectual traditions though continuous are also replete with innovations, though 'at different rates in the different parts of the same traditions and in different traditions' (Shils 1973).

These observations are well exemplified in the development or evolution of '*Critical Theory*' over the course of several generations (Calhoun 1995). Emerging in late Weimar Germany as the *Frankfurt School* at the Institute for Social Research around such figures as Horkheimer, Adorno, and Marcuse, its chrysalis was in Marxist social thought enriched with cultural and historical materials (partly influenced by Weber) and supplemented by perspectives from psychoanalysis and phenomenology. Further, a major work by a central figure, Adorno's *Negative Dialectics*, likens the Marxist origins to their philosophical grounding in Hegel's dialectical method which Marx modified to dialectical materialism. But even this may be traced back to an older Western esoteric tradition of 'negative theology' and 'negative mysticism' transmitted from Denis the Areopagite through Meister Eckart and Jacob Böhme, a tradition that stretches over a dozen centuries. Although a German component has always been part of this 'critical tradition,' a certain Americanization was introduced, first during the period of exile in America of the Frankfurt School group (some staying permanently, some returning after the war),

and second by the serious reading of Parsons in Habermas's theory of communicative action.

As noted, the tradition of the Enlightenment has been a major pillar of the social sciences. Until the 1970s or so, the *paradigm of progress* formulated in an earlier era, though shaken by two world wars, still carried sufficient weight that the study of development tended to dichotomize between societies weighted down by 'tradition' and those fortunate few (Western) who had emancipated themselves in favor of a 'modern' social order. A more balanced perspective then emerged, viewing tradition and modernity as not so much zero-sum opposites as complementary forces and orientations in the structure and dynamics of societies (Eisenstadt 1973). In fact, to void a Western-centric bias, more recent conceptualization even favors the notion of multiple historical modernities (Daedalus 1998).

2. Traditions as Interpretations

For the most part it is sociologists offering a brush-stroke of the development of the discipline as a whole, and particularly of its general theoretical framework, who have been most articulate in providing accounts of the sociological tradition. Despite overlap, there is no standardized account of traditions, including their principal figures and elements. The accounts, though varying in how traditions are categorized, themselves are constitutive of the traditions since they entail an active and critical interpretation by present authors of the legacy of previous generations of sociologists.

Moreover, unlike the Weberian ideal-type image of 'traditionalist authority (Weber 1958), there is nothing sacrosanct about how sociologists treat their traditions (save Marxists hallowing the writings of Marx but not of themselves). An author who writes about sociological traditions can make substantive changes in depiction. Thus, in one edition of his popular text of readings, Farganis credits Comte with the origins of the 'classic tradition' in sociology but in the following edition he credits Marx with beginning the tradition, and there is no reference to Comte at all (Farganis 2000). Similarly, in 1985 another equally widely adopted text posited three main ongoing core traditions: the *conflict tradition* (in which is grouped Marx, Engels, and Weber), the *Durkheimian tradition* (that includes as primary figures Comte, Spencer, Merton, and Parsons), and the *microinteractionist tradition* embracing the pragmatist orientation of Peirce, Mead, and Dewey, the symbolic interactionist line of Cooley, W. I. Thomas, and Blumer, and the phenomenological derivative of Schutz, Garfinkel, and the later Goffman; in his conclusion, Collins (1985) noted that not all the history of the field can be fitted in one of these three principal traditions, such as the utilitarian tradition, sociobiology or structuralist network theory. However, the follow-up edition elevates the *utilitarian tradition* as a tradition in its own right, acknowledging that in the previous edition, the utilitarians 'figured largely as a foil for the other traditions (Collins 1994).

What is to be retained from these two instances is that *sociological traditions are always (re)interpreted by present authors* (just as all societal traditions are always interpreted by contemporaries).

Related to this, figures from the past may be left out in present accounts or conversely, some figures from the past who were not viewed as belonging to a sociological tradition may come by later sociologists to be so regarded. Thus, Marx in the United States has only in the past 40 years or so become taken as a sociological progenitor of 'conflict theory.' On the other hand, mention rarely is made of the sociologist Ludwig Gumplowicz (1838–1909) as a figure worthy of inclusion in conflict theory, although this *fin-de-siècle* Austrian sociologist gave centrality to ethnic and race conflict, which have been much more salient in recent decades in the West and elsewhere than the class conflict paradigm of Marx. By the same token, most anglophone accounts of the development of sociology in discussing 'classical' formulators of conflict theory leave out Pierre-Joseph Proudhon (1809–65), whose influence was very extensive well into the twentieth century.

A contrary instance of a figure 'moving into' the sociological tradition after his lifetime is Tocqueville, widely accepted in the late twentieth century as a major contributor to the comparative study of institutions and revolutions in the context of modernity. Yet, as Bondon et al. (1997) have indicated, it was not until after World War II and because of the intellectual effort of Raymond Aron that Tocqueville became recognized as a core figure in the sociological tradition.

3. Sociological Traditions: One or Many?

It is particularly in the teaching of theory courses that students are made explicitly aware of sociological traditions, usually in terms of models of social reality that provide normative and cognitive orientations having heuristic significance in the devising of research and the interpretation of established data. This is often in the context of presentations of the 'sociological classics,' including debates as to whether there is merit in discussing the framers and societal context of major classics. Much of the debate hinges around whether sociology at its core (or its major subfields) is (a) a series of discrete *paradigms* around which are organized multiple and discrete communities (as understood by Kuhn 1970), or (b) a *scientific research program* in the positivist tradition, displaying an 'intermittently progressive empirical shift,' which though maintaining a hard core of irrefutable assumptions permits verification of theories, predictions of new facts, and a general increase in empirical content (Lakatos 1970). A 'humanist' perspective of sociology would favor multiple paradigms or a pluralist view of traditions;

a 'positivist' perspective would favor a cumulative tradition.

Most professional sociologists would accept Marx, Weber, Durkheim, and Mead as having provided one or more key ingress to critical sociological aspects of modernity, and traditions of theorizing and research have developed from these early foundations. But whether these are discrete or interactive, and whether they are sufficient today as main frames of inquiry or need recasting with other, perhaps newer, traditions, are matters of controversy.

3.1 Great Traditions

Despite significant variation as to what figures are included (or omitted) from certain traditions, there is a widely-shared sentiment among sociologists of different orientations that there is a common core which provides continuity and identity, a unity in diversity. Derived from the comparative study of civilizations advanced by the anthropologist Robert Redfield and the sociologist S. N. Eisenstadt, sociology in its historical development may be said to be structured by a 'Great Tradition' that endures along with 'Little Traditions.' The former frames the basic identity of the discipline, its core problem areas and assumptions, its major endeavors, and in general what might be taken as a 'mission statement' for the discipline as a whole; the Great Tradition would include exemplars in succeeding generations who have been instrumental in giving viability to the discipline. The notion of Little Traditions pertains either to subfields (e.g., social stratification, demography, sociology of religion, comparative development) or to radical traditions (in effect, sociological heterodoxies) which reject the basic normative and/or methodological structures of sociology: phenomenological sociology, some derivatives of Marxist sociology, 'Christian sociology,' and 'humanistic sociology,' as instances.

An early and influential instance of conjuring a 'Great Tradition' perspective on the development of sociology was provided by Nisbet (1966) writing in a decade of change and acrimony in the discipline itself. He distanced himself from mere historical accounts of the discipline by arguing for a nucleus of five key unit ideas—*community, authority, status, the sacred, and alienation*—which crystallized in the formative period of sociology, 1830–1900. This set of 'unit-ideas' generated poles of antithetical values: on the one hand, *community, moral authority, hierarchy*, and *the sacred*; on the other, *individualism, equality, moral release*, and *rationalist techniques of organization and power*. As underlying the fundamental structural processes of modernization (industrialization and democratization), these antitheses of modernity have given continuity to the sociological tradition. Yet Nisbet (1966) in his epilogue acknowledged that the core ideas may have exhausted their vitality, rendering problematic the 'continuing viability of a tradition, or rather, the concepts which form it.'

A newer version of the Great Tradition of sociology is provided in the account of Levine, who seeks to establish a continuity in sociological thought that can repudiate prevalent intellectual nihilism which 'relativizes truth to matters of gender, race, ethnicity, class or narrow ideology' (Levine 1995). The continuity amidst diversity is located in the 'moral dilemmas of industrial civilization voiced early in the nineteenth century,' whence 'the quest for a rational ethic was the booster that launched the social scientific disciplines into orbit'(Levine 1995). The inspiration for this Great Tradition is a vision of a *secular ethic* that undergirds both different narratives of the history of sociology seeking to lay-out *the* tradition of sociology and national traditions in search of 'the good society dialogue.'

A third approach invoking a Great Tradition theme also seeks to combine analytical thematic materials with a broad geographical differentiation between the 'American' tradition and the 'European' tradition (Alexander et al. 1997, Boudon et al. 1997). Each complementary set of volumes is a congeries of readings integrated by an editor's introduction. In his, Alexander proposes that the American tradition, with a preoccupation on the individual self, has as a paramount orientation as the 'secularization' into scientific language of two utopian movements of reform seeking social salvation: Protestantism and anti-authoritarian democracy (Alexander et al. 1997). Unlike European sociology, American sociology is preoccupied with the motives ('action' in sociological terms) and mutual relations ('interaction') of individuals. Polar orientations of optimism and pessimism are traced to foundational documents in Jefferson and Madison, respectively.

While this volume is an imaginative construction or invention of a Great American Tradition, many of the authors in the selected readings are not sociologists (Madison, Gay, Emerson, Dewey); the 'native' antebellum first generation American Comtean sociologists (Henry Hughes, George Fitzhugh, and George Holmes) are ignored; and there is no reflection on the many sociologists (McIver, Sorokin, Coser, Bendix, Etzioni, etc.) who have come as immigrants, exiles, or refugees imbued or exposed to other traditions and who have thereby modified significantly what was identified as *the* indigenous microtradition. Moreover, sociology as the secularization of utopian movements of salvation based in the Enlightenment is arguably as much an ultimate goal of continental sociology.

Rejecting that there has occurred an 'epistemological break' between our era and the Enlightenment, Cherkaoui in presenting the European classical tradition proposes two paradigms that provide continuity: *holism* and *methodological individualism*. The core identity of the European tradition, he argues, was produced between 1890 and 1914 with the tutelary

figures of that generation providing the themes and theories for subsequent generations (Bondon et al. 1997). Within this Great Tradition umbrella, Cherkaoui gives particular attention to the postwar internal development of sociology in England, Italy, Germany, and especially France. If the last named was imprinted heavily by the dominant paradigm of *structuralism* (drawing from both Marxism and Durkheimian sociology as well as other sources), an important opposition deserves to be noted in figures like Boudon, Crozier, and Touraine who in different respects have an actor-oriented sociology, viewing the actor (either as an individual or as a collective subject) with 'greater or lesser autonomy, rationality, and ability to construct strategies,' (Bondon et al. 1997). Cherkaoui leaves out of his account an important cluster of postwar French sociologists who drew from the Durkheim–Mauss tradition a dynamic, non-structuralist orientation to problems of modernity: Gurvitch, Balandier, Dumazdier, and Duvignaud, among others.

Considering that postwar American sociology has had important structuralist orientations (e.g., from Goffman and arguably from ethnomethodology at the microlevel, and from world-system analysis at the macrolevel) and that recent European sociology has had important action or subject-oriented components (e.g., Boudon and Touraine in France, Joas in Germany), it would appear as if the sharp differentiation of 'American' and 'European' tradition is becoming less tenable.

3.2 An Overlooked Great Tradition

Since the *historiography of sociology* tends to be written by those most familiar and most likely to be specialists in sociological theory, most of the accounts of sociological traditions dispense with what is, in fact, an integral aspect of the discipline's identity: the tradition of *empirical research*. It is this tradition which is the backbone of graduate training. How it developed (or failed to develop) in different settings has to date been documented only fragmentarily.

Yet empirical and quantitative research has had from the beginning of the nineteenth century a parallel but neglected development with sociological theory (Lécuyer and Oberschall, 1968, Oberschall 1972). As part of the forgotten history of this Great Tradition, Lazarsfeld (who at Columbia stimulated research on the research tradition) pointed out that Weber contributed with his research associates, 'at least 1000 pages of research findings which, in style and format, would not be easily distinguished from the pages of our contemporary sociological journals,' (Lazarsfeld in Oberschall, 1965). Oberschall presents in detail why Weber's persistent efforts at establishing a tradition of collaborative empirical research failed to become institutionalized in Wilhelmine Germany (Oberschall 1965).

In contrast to Germany's failure, a research tradition flourished almost from the start in the United States. The evolution of that tradition is documented in the important account of the Chicago School by Bulmer (1984). Bulmer points out that almost forgotten in previous accounts is the quantitative aspect of research at Chicago in the interwar period of its great flourishing, with such figures as Burgess, Ogburn, Thurstone, and Stouffer; equally important, Chicago sociology had a major impact on the profession world-wide by institutionalizing *collaborative* sociological research that stimulated theory and research interacting in the same university department (Bulmer 1984).

3.3 Schools as Strong Traditions

Attempts at categorization in terms of 'schools' were pioneered by Sorokin (1928), who proposed about 11 broad orientations (e.g., the 'mechanistic school,' the 'geographical school,' varieties of the 'sociologistic school,' etc.). Expanding on the concept of school, a *strong tradition* may be established where actual members of a school in or more institutional setting transmit a core set of beliefs and methodological approaches to a new generations of recruits, which in turn will add content and modify some features of the orientation of its predecessor, though retaining the core presuppositions. In this vein, Tiryakian (1979, 1986) has suggested that major advances in the social sciences stem from the collaborative efforts of 'great schools' founded by a charismatic intellectual leader (such as Durkheim, Park, Parsons, Horkheimer, Malinowski, Evans-Pritchard, and Freud) who in fact frames a research program. While schools are not as comprehensive as 'great traditions' in the number of sociologists identifying with them, the density of interaction and the network of associates provide them with a concentrated impact on the development of their respective discipline.

3.4 Little Traditions

3.4.1 National traditions. There are many axes on which traditions of lesser magnitude than those embracing the entire profession may be drawn. One axis frequently used, explicitly or implicitly, is the national context of the development of sociology. Sociology's unstated assumption as the science of society favors a universalistic claim for the conceptual framework of its theories; yet, sociology for the most part develops in a national context in which graduate training is offered.

This is recognized in a volume edited by Genov under auspices of the International Sociological Association, stimulated in the 1980s by growing awareness that the question of national tradition becomes

salient in an era of *internationalization* (or *globalization*) and of the renaissance of social sciences in regions where they have not been in continuous contact with the mainstream (e.g., in countries blocked off from free and easy access to the West during the Cold War). The consequences of internationalization (or globalization) and/or of having been in the periphery raises the problematic of *indigenization* (Genov 1989). Genov's formulation of the question is the extent to which sociological 'newcomers,' those national settings which have only recently begun to participate freely in the development of sociology (including Eastern Europe and Russia), can achieve a balance between being responsive to their local problems and traditions and to the international aspect of sociology as a universalistic science. Stated otherwise, how much autonomy and originality can a national tradition retain with the internationalization of sociology? (Genov 1989). This may not be perceived as a problem for American sociology because of American hegemony in world affairs, including the linguistic hegemony of English, but it is a realized challenge for even established sociological centers like the French- and German-speaking countries.

While a national tradition is a major factor in the institutionalization and reproduction of the sociological enterprise from generation to generation, there are limitations to giving it undue weight. First, a national tradition may obscure significant variation in theoretical and methodological orientations within a national setting, from one sociological center to another. Second, national traditions, even where well-established, may become transformed due to factors exogenous to the sociological enterprise, such as political transformations. This happened (in the 1930s in Western and Central European countries, and in the 1950s and 1960s in various Eastern European ones) by regimes becoming authoritarian or totalitarian.

3.4.2 Other little traditions. Miscellaneous traditions of lesser scope than national ones or general paradigms continually are evoked, invoked, and invented in the course of providing accounts and legitimization to certain orientations. A few illustrations will indicate the range of these traditions.

In their overview of *the sociology of development* Evans and Stephens (1988) observe that:

> 'While Cardoso and Faletto's model (first published in 1969) was primarily concerned with internal dynamics, other work *in the dependency tradition* was much more preoccupied with tracing the connections between the evolution of core countries and developmental sequences on the periphery.'

Since Cardoso and Faletto are credited as having produced the 'founding statement of the dependency approach,' it would indicate that this tradition is of very recent origin.

Similarly, Giele in discussing pioneers in the gender field contrasts among orientations '*liberal feminist theory* (that) drew on Enlightenment ideas of individual rights, justice and freedom,' and '*traditional Marxist feminist theory*' (Giele 1988). Both of these recent cited traditions invoke older established 'great traditions' of liberalism and Marxism, not too dissimilar from new sects seeking legitimization by invoking very established older religious traditions. Also, Olick and Robbins (1998) in their overview of sociological approaches to collective memory and tradition make reference to 'the Mannheimian tradition in the sociology of knowledge and the Mertonian tradition in the sociology of science.' In this instance, a 'little tradition' is that of a specialized field of sociology—the sociology of knowledge and the sociology of science—whose foundation is identified with a recognizable person; presumably, even specialty areas within broad fields may also take on the aura of a tradition, whether the designation is given by the name of a figure or an institutional setting where a certain approach is handed down from generation to generation. Lastly in this vein, a 'little tradition' may be invoked tacitly by former students adducing the contributions of their illustrious teachers, as, for example, the volume of Merton and Riley (1980) dedicated to their teacher Talcott Parsons and having contributors testifying to such mentors as Weber, Mead, Park, Sorokin, Frazier, Lazarsfeld, MacIver, and Stouffer.

4. *The Future of Traditions*

Significant changes in the bearing of tradition(s) to sociology are likely to emerge in the twenty-first century, even if the changes that can provoke new images of tradition can only be partly anticipated.

Although some retention of the Great Tradition(s) is likely, structural, political, and technological transformations already underway will give only vestigial importance to much of the conceptual legacy of a past that was grounded in the dynamics of industrializing and late industrial society. While boundaries between national traditions may become increasingly blurred stemming from globalization and hybridization, emergent regional linkages (e.g., the European Union, NAFTA) will also prod sociologists to acknowledge established neighboring traditions. New subfields and new fields linking sociologists to other disciplines (in the humanities, social sciences, and natural sciences) will emerge and in becoming institutionalized, new traditions will appear evoking *their* founding figures and documents. Already, a new historiography reconstructing the tradition of women in the development of sociology is well under way (Lengermann and Niebrugge-Brantley 1998); it is likely that what had been the covert gender side of the Great Tradition will become recognized in terms of the contributions of

women to both social theory and social research, and in the process, recasting the image of the Great Tradition. By the same token, dramatic breakthroughs in knowledge of the biological bases of social behavior could also well lead to reconceptualizing the Great Tradition.

See also: Conservatism: Historical Aspects; Critical Theory: Contemporary; Durkheim, Emile (1858–1917); Functionalism in Sociology; Interactionism: Symbolic; Macrosociology–Microsociology; Marxism in Contemporary Sociology; Modernity: History of the Concept; Sociology, History of; Structuralism, Theories of; Structure: Social; Theory: Sociological; Traditions: Social; Utilitarianism: Contemporary Applications; Weber, Max (1864–1920)

Bibliography

Alexander J, Boudon R, Cherkaoui M 1997 (eds.) *The Classical Tradition in Sociology: The American Tradition*. Sage, London and Thousand Oaks, CA

Anderson B R 1991 *Imagined Communities*. Verso, New York

Boudon R, Cherkaoui M, Alexander J (eds.) 1997 *The Classical Tradition in Sociology: The European Tradition*. Sage, London & Thousand Oaks, CA

Bulmer M 1984 *The Chicago School of Sociology. Institutionalization Diversity and the Rise of Sociological Research*. University of Chicago Press, Chicago

Calhoun C 1995 *Critical Social Theory*. Blackwell, Cambridge, UK

Collins R 1985 *Three Sociological Traditions*. Oxford University Press, New York

Collins R 1994 *Four Sociological Traditions*. Oxford University Press, New York

Daedalus 1998 *Early Modernities*. **127**(3)

Eisenstadt S N 1972 *Post-Traditional Societies*. Norton, New York

Eisenstadt S N 1973 *Tradition, Change and Modernity*. Wiley, New York

Eisenstadt S N, Graubard S R (eds.) 1973 *Intellectuals and Tradition*. Humanities Press, New York

Evans P B, Stephens J H 1988 *Development and the World Economy*. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA

Farganis J 2000 *Readings in Social Theory. The Classic Tradition to Post-Modernism*, 3rd edn., McGraw-Hill, Boston

Genov N (ed.) 1989 *National Traditions in Sociology*. Sage, Newbury Park, CA

Giele J Z 1988 Gender and Sex Roles. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA, pp. 291–23

Hobsbawm E, Ranger T (eds.) 1992 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Kuhn T S 1970 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Lakatos I 1970 Falsification and the Methodology of Scientific Research Programmes. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK, pp. 91–196

Lécuyer B, Oberschall A R 1968 Sociology: Early History of Social Research. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Vol. 15, pp. 36–53

Lengermann P M, Niebrugge-Brantley J 1998 *The Women Founders. Sociology and Social Theory 1830–1930*. McGraw-Hill, New York

Levine D N 1995 *Visions of the Sociological Tradition*. University of Chicago Press, Chicago

Merton R K, Riley M W (eds.) 1980 *Sociological Traditions from Generation to Generation*. Ablex Publishing, Norwood, NJ

Nisbet R A 1966 *The Sociological Tradition*. Basic Books, New York

Oberschall A 1965 *Empirical Social Research in Germany. 1848–1914*. Mouton, Paris

Oberschall A 1972 (ed.) *The Establishment of Empirical Sociology. Studies in Continuity, Discontinuity, and Institutionalization*. Harper & Row, New York

Olick J K, Robbins J 1998 Social Memory Studies: From "Collective Memory" to the Historical Sociology of Mnemonic Practices. *Annual Review of Sociology* **24**: 105–40

Shils E 1973 Intellectuals, Tradition and the Tradition of Intellectuals. In: Eisenstadt S N, Graubard S R (eds.) *Intellectuals and Tradition*. Humanities Press, New York, pp. 21–34

Shils E 1981 *Tradition*. University of Chicago Press, Chicago

Sorokin P A 1928 *Contemporary Sociological Theories*. Harper, New York

Tiryakian E A 1979 The Significance of Schools in the Development of Sociology. In: Snizek W E, Fuhrman E R, Miller M K (eds.) *Contemporary Issues in Theory and Research: A Metasociological Perspective*. Greenwood, Westport, CT, pp. 211–33

Tiryakian E A 1986 Hegemonic Schools and the Development of Sociology. In: Monk R C (ed.) *Structures of Knowing: Current Studies in the Sociology of Schools*. University Press of America, Lanham, MD, pp. 417–41

Weber M 1958 *From Max Weber: Essays in Sociology*. Gerth H H, Mills C W (eds.) Oxford University Press, New York

E. A. Tiryakian

Traditions: Social

The concept of tradition has been important in the history of sociology in the twentieth century, but its signification has completely changed in contemporary works. At the turn of the century, sociologists and anthropologists utilized this concept to outline a type of society defined by traits placed in opposition to other traits characterizing industrial society, and in a well-known work Max Weber has distinguished three types of authority: legal-rational, charismatic, and traditional. One hundred years later, sociology considers tradition as an invented reference developed by societies in function of the demands of the present (Hobsbawm) and as a form of rationality among others that orients action (Boudon).

The tradition–modernity dichotomy dates back to Saint-Simon who compared the characteristics of the emerging industrial society to those of the *Ancien régime*. This dichotomy became a key element in explaining the social transformations that were oc-

curring in modernizing societies during the first part of the twentieth century. The defining traits of the traditional society are well known: the fusion of family and business, the division of labor dictated by social structure (age, gender, relatives), reference to custom as the dominant value, the privileged attention accorded to saving and accumulation to the detriment of productive investment, the importance of the local community, relative autarky, and self-sufficiency. The American anthropologist Horace Miner (1939) found in Saint-Denis-de-Kamouraska in Quebec a kind of archetype of the preindustrial traditional rural society depicted in a work that remains a classic in social sciences read by generations of American students. At the opposite end of the spectrum, the industrial society described by Raymond Aron, Georges Friedmann, Daniel Bell, and others is marked by the importance of technique, economic growth, calculation and management, productive investment, and the accumulation of capital, separation of family and business, establishment of a labor market, urbanization.

The transition from traditional to industrial society paradigm and the folk–urban continuum formalized by R. Redfield (1956) were criticized sometimes as descriptive approaches of social change and as paradigms with limited explanatory power. Henri Mendras (1976) underlined the fact that this vision of things denied the proper originality of nonindustrial societies and their diversity and he recalled that the concept of the traditional society often embraces different types of societies that are in fact very different (e.g., from primitive societies to peasant societies). He argued that one must not confuse traditional societies and peasant societies. If hunter-gatherer societies are totally independent, peasant societies are relatively autonomous but are embedded in larger, global societies having diverse characteristics. The paradigm of the transition from traditional to industrial society is based on an evolutionist postulate, that has been placed in question, that makes the developed societies the norm to be obtained. In the same vein, we can criticize the thesis of the resistance to social change by tradition because it supposes clear, obtainable goals and also the existence of bad faith on the part of traditional individuals attached to the past and incapable of recognizing the merits of anything new (Boudon 1984).

Consider another manner of conceiving the tradition–modernization opposition. Lévi-Strauss (1962) compares the knowledge and technique of the handyman, traits of a traditional society, with the knowledge and technique of the engineer, attributes of the industrial society. The first type represents empirical and inductive knowledge. It depends on a long tradition of precision which can be complex and efficient but it is also filled with errors, notably as to the explanation of its efficiency. The knowledge of the engineer is of another order. It is based upon experimental science, accumulation of scientific knowledge, verification, technique, manipulation of objects, operationalization, and organization. The classical opposition in economics between protectionism and free-trade also illustrates the opposition between the ancestral or traditional ways of doing business and the modern methods. The authorities in the Republic of Venice jealously protected the ancestral traditions of their glass-makers, sources of important benefits. Contemporary entrepreneurs count on innovation, discovery, free circulation, investment and strategic alliances to support the growth of their activities and their profits.

The examples of the engineer and the entrepreneur clearly demonstrates that tradition and modernization are differentiated in two respects. First, the relationship to time is different. Tradition is oriented towards a legitimate reference to the past, while modernization is oriented towards the mastery of the future, the discovery of the new. Second, tradition and modernization are marked by a strongly different manner of intervening in the world, of producing goods and services; an empirical manner of learning over time and transmission across social relations that are embedded within the family or small communities, and a scientific manner that asserts itself as the result of research and calculation, embedded within complex social networks and formal organizations.

E. Shils (1981) proposed a definition of the concept of tradition which was largely accepted. He defined tradition as anything which is transmitted or handed down from the past to the present. According to Shils, traditions include beliefs, norms, institutions, as well as material objects which all individuals will find when they grow up; traditions refer to transmitted things, ideas, and values. It is now generally accepted that the scope of tradition should be narrowed down. We will retain two aspects—which are hidden in the folk–urban continuum or in the community-society ideal type of Tonnies—which contribute to define tradition in contemporary analyses.

First, to be part of a tradition, objects, norms, or ideas have to be well alive, and never having been destroyed, abandoned, or forgotten. Specifically qualified fragments of the heritage should retain a strong intimate link with the present to be a tradition.

Second, it is generally accepted that traditions change with the passage of time. Traditions are not immutable and they are re-examined and questioned. Tradition is a transmitted heritage which survives and orients contemporary actions, but this heritage is often transformed (Halbwachs 1935/1994), Eisenstadt 1973). Sztompka (1994) stated that a special part from the whole of the historical heritage is elevated to the category of tradition by contemporary people for their own usage. In the United States, national history is being rewritten to take into account the role of Amerindians, women, and ethnic minorities.

Canada delimited the frontiers of a new territory that contains an Inuit majority in the North in

1998—Nunavut—that took into account the testimonies of the oral traditions of the elders. Scientific cartography and satellite images were the privileged work instruments but the life histories and experiences of those who inhabited the land were taken into account, especially to establish its Eastern frontier. Similarly, the oral traditions of Canadian Amerindians and the *wampums* of their material culture were also taken into account as elements of proof in the Canadian system of justice that examined their grievances and ancestral rights. Canadian judges are now obliged to interpret the rights of Amerindians in light of ancestral traditions and the orientations of modern ways of living (equality between men and women, for example). There is no doubt that the resort to tradition will inspire other courts of justice throughout the world where we find indigenous cultures, in Latin America or in Oceania, for example. Here tradition also appears as a legitimate form of knowledge—but different—than the scientific knowledge and cartographic techniques of historians or jurists and it is susceptible to applications in the same way as the latter. Tradition no longer appears as the only or principle source that orients behavior and decisions but as one element among others that is susceptible to being criticized.

Hobsbawm (1983) has contributed to the new examinations of tradition by stressing its *invented* character. He has distinguished three groups of invented traditions. First, there are traditions that express the social cohesion of nations and communities. Second, traditions can legitimize an existing order, status, institutions, or authority. Third, traditions which socialize certain values or rules of behavior into individuals. The traditions of the British crown, for example, leaves the impression that they are rooted deep into the past. Now the work of historians has demonstrated that these traditions are of more recent origins and were invented recently (in fact in the nineteenth century) in order to give greater legitimacy to the institution of the monarchy. Closer to us, the celebration of the new Beaujolais in France is a recent creation, the fruit of a new approach in the marketing of the wine of this region that has been successful in countries where they appreciate French wines. Another example, Australia. At the moment, the Australians are questioning whether or not they should abandon the royal symbols from their colonial period. If they decide to abandon those traditions and symbols, they will probably give themselves another symbolic tradition to mark the birth of their society peopled in large part by immigrants. A new tradition will therefore be created for use in the present.

The protection of the environment, fauna, and flora, has become a daily theme at the end of the twentieth century. In North America, the Amerindians have been characterized as environmentalists before the advent of an environmental movement. Research by anthropologists and historians has demonstrated that the picture is much more complex. Krech (1999) recalls that the Amerindians had overexploited the beaver in the St. Lawrence Valley after the arrival of Europeans in New France in the sixteenth century to the point of extinction. The notions of environmental protection and endangered animals were foreign to the traditional culture of the aborigines, because the way of life of hunter-gatherers is based on maximal exploitation of the immediate environment. On the other hand, the Amerindians have an animist conception of the animal world. Thus, the Plains Indians believed that they could not allow an animal to escape under the pretext that it would warn the others to not venture in that area, they therefore killed all the animals they saw. Because they believed in reincarnation, the majority of Amerindians thought that the dead animals would, at any rate, return the following season. For Krech, the myth of the ecologist Indian essentially reflects contemporary preoccupations for the environment rather than the traditional and historical culture of the Amerindians.

This analysis led us to the following conclusion: contemporary sociology has abandoned the transition paradigm (from traditional to industrial societies) and now integrates tradition into analyses of ongoing social change. In recent works, tradition is used to define a type of society (be it theoretical or empirical), and it also defines a subjective orientation toward action. These different approaches will be developed here.

In his general theory of action, Freitag (1986) defines three types of societies based on three forms of definition of the norms of social action. The first form he proposes—culture—defines mythic and archaic societies. The second one—power—defines two other types: traditional societies and modern societies, the later ones built against tradition. The third form of regulation—control—characterizes postmodern society, in which a large number of organizations contribute to the structuration of society.

Tradition is also used to describe empirically different societies. Following Max Weber's arguments, Huntington (1996) argued that the world is divided into eight major cultural zones which revealed persistent cultural differences caused by different religious traditions which are still powerful despite industrialization and the same forces of economic development. Inglehart and Baker (2000) have studied values in surveys made in 65 countries which cover more than 75 percent of the world population and they observed the persistence of distinctive traditional values.

Tradition is also considered as a normative orientation toward action, a perspective underlying Tocqueville's works (1856/1986, for example). In his works, Raymond Boudon formulates a general theory of action based on the actor's search for meaning. He proposes to enlarge the concept of the rationality of the actor by distinguishing five types of rationality:

utilitarian, teleological, axiological, cognitive, and traditional. All of these types of rationality describe the good reasons why the actor is driven to act in a certain way. Economists and rational choice theorists privilege the utilitarian rationality based on the pursuit of self-interest. However, for Boudon, there are other kinds of reasons/motives that can also influence the way that individuals act, and to limit the content of rationality to the single dimension of utility is an exercise in reductionism. Custom and habit can also incite individuals to act in a certain way. Therefore, when one speaks of traditional rationality—'it has always been that way'—he is speaking of one type of rationality among others.

For Boudon, sociological analysis has to define the context and circumstances in which one type of rationality or another is at work, keeping in mind that one type is no more important than another. This approach permits us to understand the role of tradition in sociological explanations. In peasant societies, the rhythm of the seasons is of major importance and the technologies are not very developed (Weber 1976). Here, the reference to custom and tradition is strong and the individuals do not have good reasons to question that which already works. They are led to value the stability and the family that provides the necessary labor to exploit the land. However, when technological inventions appear to menace the traditional ways of life, tradition remains important for understanding the implantation of these transformations or the difficulties in accepting them. Mendras (1967) has demonstrated that the introduction of inventions in peasant societies cannot be understood independently of the welcoming culture because these innovations have to be first integrated into the traditional culture before being fully accepted.

Boudon (1992) gives a good example that demonstrates the role of tradition in ongoing social change in restating the observations contained in the celebrated monograph by Thomas and Znaniecki (1927/1984) on immigrant Polish peasants in Chicago at the turn of the twentieth century. The immigrants that adapted the most rapidly to the welcoming society were those that manifested the highest degree of traditionalism and attachment to their original milieu. This integration in a familiar milieu gave them support and a frame of reference that slowed down the process of social disorganization that was typical of immigrant communities when faced with a radical rupture with their traditional original community. The Polish immigrants adopted a rational action that permitted them to integrate themselves in the welcoming society, the traditional collectivity serving as a relay.

To illustrate the richness of this approach, here is an example taken this time from French sociology. In his studies of religious sociology, Le Bras (1955) observed a paradox. The Bretons who migrated to Paris during the 1950s rapidly abandoned religious practice. A superficial analysis gave the impression that they practiced a strict conformist religion that they abandoned once they were beyond the traditional way of life structured by family and village. Le Bras proposed another explanation by demonstrating that the logic of behavior was the same in the village and in the city. In the village, behavior was regulated by the social group. Going to mass was not only a way of confirming one's Catholicism, it was also participating in a valued social activity which strongly rhythmed life. In the city, in a pluralist environment, going to mass was seen as expressing an opinion and a behavior that did not attract the adherence of all, and in a way, that was contrary to majority opinion. Thus, it was difficult for the newcomers, who wished to integrate in their new environment, to oppose something which was the norm in their original milieu. In this case, one could say that going to mass in the village and not going once in the city follow the same social logic.

All of these examples demonstrate different ways to refer to the concept of tradition. Tradition describes a type of society, but it also defines an orientation of action. This perspective can be found in numerous works of contemporary sociology and demonstrates how different types of rational behavior can mix together. To illustrate this perspective, we will recall how the reference to tradition is posed in three fields of empirical studies of major importance at the end of the twentieth century.

The market consumption society in the developed world proposed standardized products—fast-food, coca-cola, computers, cars, jeans, CDs, films—and globalization tends to homogenize commercial practices and modes of production. However, sociologists have started to discover the multitude of innovative responses in the national cultures of consumption, as Miller (1995) has illustrated in his research. He argues that local contexts remain extremely diversified and, in each case, outside material influences are reinterpreted. Cultural diversity continues to be reaffirmed and supported by living national and cultural traditions. Coca-cola does not occupy the same place in the dietary culture of the French and the connoisseurs state that American pizza is quite distant from its Italian cousin.

New technologies have a great impact on the character of work, but work is also embedded in culture and traditions. People work, and supervise other's work, within stringent institutional limits marked both by national traditions and organizational innovation.

The empirical work on the integration of immigrants in developed societies demonstrates that they do not form totalities with contours that are well defined by their original traditions. In fact, they are quite diverse and in reconstructing their traditions in the new country, they combine innovative elements with the adopted elements of their new home. Tradition and modernity combine to produce new cosmopolitan identities that can be different from one

country to another as Kymlicka (1995) has suggested. The reference to tradition, therefore, remains quite alive in present sociological analyses.

See also: Industrial Society/Post-industrial Society: History of the Concept; Modernity; Modernity: History of the Concept; Modernization, Sociological Theories of; Multiculturalism: Sociological Aspects; Social Change: Types

Bibliography

Boudon R 1984 *La place du désordre*. Presses universitaires de France, Paris

Boudon R 1992 Action. In: Boudon R (ed.) *Iraite de Sociologie*. Presses Universitaires de France, Paris, pp. 21–55

Eisenstadt S M 1973 *Tradition, Change, and Modernity*. Wiley, New York

Freitag M 1986 *Dialectique et société. Introduction à une théorie générale du symbolique*. Éditions L'âge d'homme, Genève, Switzerland and Éditions Saint-Martin, Montréal, Canada

Halbwachs M 1935/1994 *Les cadres sociaux de la mémoire*. Albin Michel, Paris

Hobsbawm E 1983 Introduction: Inventing traditions. In: Hobsbawm E, Ranger T (eds.) *The Invention of Tradition*. Cambridge University Press, Cambridge, UK, pp. 1–14

Huntington S P 1996 *The Clash of Civilization and the Remaking of World Order*. Simon and Schuster, New York

Inglehart R, Baker W E 2000 Modernization, cultural change, and the persistence of traditional values. *American Sociological Review* **65(1)**: 19–51

Krech S 1999 *The Ecological Indian. Myth and History*. W. W. Norton and Co., New York

Kymlicka W 1995 *Multicultural Citizenship*. Oxford University Press, Oxford, UK

Le Bras G 1955 *Études de sociologie religieuse*, Vols. I and II. Presses universitaires de France, Paris

Lévi-Strauss C 1962 *La pensée sauvage*. Plon, Paris

Mendras H 1967 *La fin des paysans*. S.E.Q.E.I.S, Paris

Mendras H 1976 *Les sociétés paysannes. Éléments pour une théorie de la paysannerie*. Armand Colin, Paris

Miller D (ed.) 1995 *Worlds Apart. Modernity Through the Prism of the Local*. Routledge, London

Miner H 1939 *St. Denis. A French-Canadian Parish*. University of Chicago Press, Chicago

Redfield R 1956 *Peasant Society and Culture*. University of Chicago Press, Chicago

Shils E 1981 *Tradition*. University of Chicago Press, Chicago

Sztompka P 1994 *The Sociology of Social Change*. Blackwell, Oxford, UK

Thomas W I, Znaniecki F 1984 *The Polish Peasant in Europe and America* New York [1927], rev. edn. University of Illinois Press, Chicago

Tocqueville A de 1856/1986 *L'Ancien régime et la Révolution*. Robert Laffont, Paris, coll. Bouquins

Weber E 1976 *Peasants into Frenchmen. The Modernization of Modern France 1870–1914*. Stanford University Press, Stanford, CA

S. Langlois

Training in Clinical Psychology in the United States: Practitioner Model

The practitioner model of training in clinical psychology focuses on preparing students to become highly skilled professional psychologists. A professional clinical psychologist is part of the organized community of scientific psychology; his or her professional efforts focus on helping individuals, couples, or families with the prevention and remediation of psychological problems in human behavior. This work may be with individuals, in families, or in group settings. Although some clinical psychologists who are trained as practitioners work in research or academic settings, most devote themselves to applied work.

Clinical psychologists have traditionally worked in a variety of settings including mental health centers, private practices, prisons, day care centers, psychiatric hospitals, substance abuse programs, and programs for persons with developmental disorders. They serve children, adolescents, adults, and the elderly with a full range of diagnostic labels.

1. A Brief History of the Development of the Practitioner Model

Several streams have fed the development of clinical psychology as a profession. These included the creation of a body of knowledge relevant to practice, the demonstration of the value of psychological treatments to ameliorate human suffering, and conferences that gradually shaped the standards and curriculum for educating clinical psychologists. We can date the establishment of clinical psychology as practice to the spring of 1896 when a psychologist at the University of Pennsylvania, Lightner Witmer, opened the first psychological clinic. Many of his early clients were children with intellectual and academic problems. After completing their education, a number of Witmer's graduate students went on to provide psychological services, establish treatment programs, and do research on clinical problems.

1.1 A Body of Knowledge

As Routh (1994) and others have noted, the science and practice of psychology should be tightly linked to one another. An effective profession requires effective skills. For clinical psychology the initial creation of this scientific knowledge base included such things as the development of tools for cognitive and personality assessment including the Binet and Wechsler tests of intelligence as well as the development of projective and objective tests of personality. Another aspect of the early growth of the profession came from the

establishment of outpatient psychotherapy centers. Many of these early clinics were psychodynamic in orientation. Later clinics offered services from the client centered, family systems, and behavioral approaches. It was important for the growth of the profession of clinical psychology to develop a body of research examining the process and outcome of these various therapies. Much of the knowledge base for clinical psychology was highly international in origin with creative work flowing to the USA from Europe. The creation of professional journals devoted to the publication of research on topics of relevance to clinical practice, including the *Journal of Abnormal Psychology* and the *Journal of Consulting Psychology*, also served to bring together in a focused way the base of scientific knowledge upon which practice could be founded. That initial knowledge base has changed greatly over the years as our understanding of scientific psychology has evolved.

1.2 Tensions Within the Field

Initially, the American Psychological Association (APA) was founded to support the advancement of psychology as a science, not psychology as a profession (Peterson 1997). Another group, the American Association of Clinical Psychologists, first met in 1917 to organize the clinical profession. Their action was highly controversial at the time (Wallin 1961) with considerable opposition within the APA to the creation of a new psychological organization. Although this group only lasted two years, out of it grew a clinical psychology section of the APA and the increasing recognition that clinical practice was a distinctive and legitimate aspect of psychology.

Creating a division within the APA to recognize applied psychology helped to advance the recognition of clinical psychology as a profession, but there remained tensions within the APA about the legitimacy of professional practice. Routh (1994) notes that one consequence of this continuing tension between practitioners and academics was the development of state psychological associations and of the American Association of Applied Psychology. The role of clinical psychologists during World War II brought recognition of the value of clinical services, and had a major impact on the growth of clinical psychology as a profession of practice as well as research. By 1945 there was a reorganization of the APA and the by-laws of that organization were expanded formally to include psychology as a profession as well as a science and to value the promotion of human welfare.

1.3 The Boulder Conference

The first major conference on the training of clinical psychologists was held in 1949 in Boulder, Colorado. This conference, sponsored by the National Institute of Mental Health, had a profound impact on the development of university programs to educate clinical psychologists. It was the Boulder Conference that defined the 'scientist–practitioner' model of professional training that for many years was the dominant model of clinical training in the USA. The goal of this model was to ensure that students were well grounded in research methods but also given applied training through their practicum and internship experiences. Although students felt pulled between these two sometime competing demands for research and practice, the Boulder scientist–practitioner model remained essentially the sole approach to clinical training for several decades (Routh 1994).

Routh (1994) points out that from early in its inception there were challenges to the Boulder model. The majority of clinical psychologists did not enter research careers after training and questions were raised about the value of the time spent teaching in-depth research skills to clinicians who would never use them beyond graduate school. Additionally, many practitioners argued that the Boulder model training programs did not devote sufficient time to teaching applied skills. Given the dissatisfaction with the Boulder model on the part of many practitioners, it is not surprising that an alternative model developed to training clinical psychologists for professional practice.

In 1964, an APA Committee on the Scientific and Professional Aims of Psychology, chaired by Kenneth E. Clark, recommended that clinical psychologists be trained in two tracks. One track would be for those persons who intended to focus on research careers and the other track would be for those who wished primarily to be practitioners. In spite of the Clark Committee recommendation, the growth of programs with primary emphasis on practitioner training proceeded slowly (Peterson 1985). The year Clark's APA committee issued its recommendations for two tracks of training, the only clinical training program focused mainly on professional practice was at Adelphi University. The Fuller Graduate School of Psychology was started in 1965. In 1968, the University of Illinois, which had a respected scientist–practitioner program in clinical psychology, added a professional practice program and granted a Doctor of Psychology (PsyD) degree to those who completed the training. The California School of Professional Psychology, a free-standing school, was established in 1982 and a number of other schools not affiliated with a college or university followed in the next decade.

1.4 Mission Bay Conference and Beyond

As practitioner programs became more prominent for clinical training in the USA, several national conferences on standards of training for practice were

sponsored by the National Council of Schools of Professional Psychology (NCSPP), a group representing the professional schools and programs. From a conference at Mission Bay California this group published *Standards and Evaluation in the Education and Training of Professional Psychologists: Knowledge, Attitudes and Skills* (Bourg et al. 1987). In successive conferences, the NCSPP has focused on such topics as ethnic diversity, curriculum, women's issues, and the role of evaluation in professional education (Peterson 1997)

2. Competencies for the Practitioner of Clinical Psychology

Modern practice in clinical psychology requires a broad base of knowledge in general psychology as well as specialized professional skills. The general base includes a background in biological functioning, cognitive processing, and human learning. The practitioner must understand the overt and covert functions of individuals, and the functioning of groups. A well-trained clinical practitioner is knowledgeable about how to assess and manage individual behavior, and he or she understands interpersonal relationships and group dynamics as well as the regulation of psychobiological factors in human functioning.

Along with their broadly based understanding of human behavior, the practitioner develops expertise in one or more approaches to the assessment and treatment of psychological problems. This expertise often reflects a specific theoretical orientation for understanding behavior, such as a psychodynamic, behavioral, humanistic, or family systems framework for approaching treatment. In addition, the professional will often have special expertise with one or more populations of clients including children, adolescents, adults, people with developmental disorders, individuals with addictive behavior, patients with serious mental illness such as schizophrenia or depression, or persons dealing with post-traumatic stress disorders. Each of these areas has an assessment and treatment literature that must be mastered by the competent practitioner.

Although the practice of psychotherapy is often through of as the core activity of the clinical psychologist, these practitioners do many other things as well. They must be able to do an assessment of an individual, family, or group, identify the best intervention methods to fit the needs of the situation, and then apply those techniques and make modifications in treatment based on feedback from the process of change. Empirically supported treatment techniques have been developed to intervene with some psychological problems and the clinical psychologist must be able to apply these as appropriate. He or she must also now how to modify or extend these treatment methods for patients who do not respond to the treatment procedures as they may be defined in a treatment manual.

Peterson (1997) employs the concept of 'disciplined inquiry' to describe the role of the practitioner in bringing intellectual rigor and a systematic approach to the evaluation of his or her work and decision making. One brings scientific knowledge to bear in solving the problems of one's client, but always with the needs of the client and his or her response to treatment as the guiding principle in choosing interventions.

3. Curriculum

Doctoral education to prepare a student to work as a practicing clinical psychologist combines academic instruction with practicum placements in settings where students apply the skills they learn in the

Table 1
A sample practitioner training curriculum

Year One
Theoretical foundations of intervention—behavioral
Theoretical foundations of intervention—psychodynamic
Theoretical foundations of intervention—organizational
Interviewing skills
Child psychopathology
Adult psychopathology
Individual cognitive assessment
Behavioral assessment
Psychodynamic assessment
Foundations of cognitive/affective psychology
Foundations of social psychology
Year Two
Philosophy and systems of professional psychology
Statistics and experimental design
Community psychology
Professional issues and ethics
A Course in intervention methods, e.g., family therapy, short term dynamic therapy, cognitive behavior therapy
Clinical psychopharmacology
Year Three
Intervention techniques with ethnic and racial minority clients and families
Courses in intervention methods
Begin dissertation research
Year Four
Elective courses
Dissertation
Year Five
Full time internship in clinical psychology

classroom. As shown in Table 1, the first year of such a doctoral training program typically provides foundation courses in such topics as the social, cognitive, and affective bases of behavior, descriptive adult and child psychopathology, basic interviewing techniques for gathering diagnostic and treatment data, and an introduction to the theories that underlie assessment and treatment modalities. Most commonly these theoretical approaches are behavioral, cognitive behavioral, family systems, humanistic, or psychodynamic in nature.

As students progress in training, their courses become focused on specific patient populations and treatment modalities. For example, a student may study family therapy treatment methods, psychoanalytic psychotherapy, or the use of cognitive behavior therapy in the treatment of children or adults. The typical program also requires students to learn basic research methodology and statistics in order to ensure that the practitioner is a sophisticated consumer of the research literature. To permit the systematic evaluation of specific efforts to deal with clinical problems, most programs require the student to do a professional dissertation. These projects often draw upon a broader methodology than the typical dissertation of a student who intends to pursue a career as a research psychologist. Examples of these research methodologies include a sophisticated set of case studies used to develop a theoretical concept, a program evaluation, or a qualitative analysis of interview data. Some practitioner dissertations may also employ a traditional empirical design.

Throughout their training, students will be exposed to various professional practice settings where they can apply their new knowledge under the close supervision of senior psychologists. One year of this training is a full-time internship that is typically done after the student has completed his or her course work and in many cases has also completed a dissertation project. Many newly graduated clinical psychologists will also spend one or two years in postdoctoral training further to develop special expertise under advanced supervision.

4. *The Practitioner Model vs. the Scientist Model*

Two major models have developed to educate doctoral students in the practice and study of clinical psychology. These are the 'scientist–practitioner' and the 'practitioner–researcher' (or practitioner–scholar) models. Table 2 summarizes some of the differences between these models. The differences in practice are sometimes less sharp than the models suggest. Some people trained as scientists are engaged in full-time practice of clinical psychology and some people trained as practitioners work in universities or research settings.

The scientist–practitioner model almost exclusively involves a Ph.D. degree at the completion of training, which is the degree held by most clinical psychologists who work in academic or research settings. These training programs typically provide more research training and less applied training than do practitioner programs. By contrast, the graduate of a practitioner–research or practitioner–scholar program is typically preparing for a career in practice. A few practitioner programs may grant a Ph.D. degree in psychology but most now grant the PsyD degree.

See also: Clinical Psychology in Europe, History of; Clinical Psychology in North America, History of; Curriculum Theory; Professions, Sociology of; Psychotherapy: Legal Issues (United States); Training in Clinical Psychology in the United States: Scientist–Practitioner Model

Bibliography

Bourg E F, Bent R J, Callan J E, Jones N F, McHolland J, Stricker G (eds.) 1987 *Standards and Evaluation in the Education and Training of Professional Psychologists: Knowledge, Attitudes and Skills*. Transcript Press, Norman, OK

Peterson D R 1985 Twenty years of practitioner training in psychology. *American Psychology* **40**: 441–51

Peterson D R 1997 *Educating Professional Psychologists. History and Guiding Conception*, 1st edn. American Psychological Association, Washington, DC

Routh D K 1994 *Clinical Psychologist Since 1917. Science, Practice and Organization*. Plenum, New York

Wallin J E W 1961 A note on the origin of the APA clinical section. *American Psychologist* **16**: 256–58

S. L. Harris

Training in Clinical Psychology in the United States: Scientist–Practitioner Model

Although some argue that the science and practice of psychology is as old as recorded history, contemporary psychology, as an evolving discipline, emerged at the end of the nineteenth century. From the beginning, the science and practice (or application) of psychology were inextricably woven together although few early psychologists embraced this model. Experimental psychology was born in 1879 in Willhelm Wundt's laboratory in Leipzig, Germany and brought to the USA by numerous students who worked with him. They pursued rigorous scientific investigations of animal and human subjects specifically around learning, memory, and sensory processes. At about the

Table 2
A comparison of practitioner–researcher and scientist–practitioner models

	Practitioner	Scientist
Training	Practice & research oriented	Research oriented
Coursework oriented	Application & practice oriented	Research & application
Research	Applied/consumer	Basic & applied
Department culture	Practice oriented	Research oriented
Training goals	Practitioner–researcher	Scientist–practitioner
Career settings	Community mental health centers, colleges counseling centers, clinics, hospitals private practice	Colleges & universities, research agencies, medical schools
Admissions criteria	Human service interest and/or experience, interpersonal skills, intellectual abilities	Research interest and/or experience, interpersonal skills, intellectual abilities

same time in Vienna, Austria, Sigmund Freud was developing his 'talking cure' psychoanalysis, a clinical approach to the influence of trauma or past events on disruptive emotional symptoms often having to do with cognitive processes such as 'repression' or 'transference' (see *Psychoanalysis, History of*). But whatever the overlapping interest in learning, memory, and relearning, proponents of these two approaches had little to do with each other and a reapproachment between the scientific investigations of behavior and the clinical techniques of psychoanalysis was not offered until some fifty years later (Dollard and Miller 1950).

The late nineteenth century was also a time of considerable debate about the etiology of mental illness. Physicians, such as Emil Kraeplin who had also studied with Wundt, embraced an organic model which proposed that manic-depressive psychosis and dementia-praecox (schizophrenia) would be found to have a biological basis similar to syphilis, for which the spirochetes causing the brain degeneration had just been found. Others assumed that psychological processes composed the underlying etiology of emotional and mental disorders. These debates were further couched within discussions about whether investigations of general behavior or a concern with individual differences would lead to new understandings about the nature of human behavior.

Lightner Witmer, the father of clinical psychology, was an extraordinary exemplar of a scientist–practitioner. In his attempts to use psychology to help children who were having difficulty in school, Witmer was the first to write about the need to integrate the basic science of psychology and its application. In fact, he believed that there was no valid distinction between a pure science and an applied science.

Witmer founded the first Psychological Clinic at the University of Pennsylvania in 1896 and the first clinical journal, *The Psychological Clinic*, in 1907 (see *Clinical Psychology in North America, History of*). In the opening article of the journal, Witmer writes movingly of the difficulties he had finding ways to make psychology useful in his treatment of clients. One of his first cases was a fourteen year old boy referred by his teacher because he could not seem to learn to spell. Witmer (1996) looked to basic psychology but could not find 'that the science of psychology had ever addressed itself to the ascertainment of the causes and treatment of a deficiency in spelling. Yet, here was a simple developmental defect of memory; and memory is a mental process of which the science of psychology is supposed to furnish the only authoritative knowledge. It appeared to me that if psychology was worth anything to me or to others it should be able to assist the efforts of a teacher in a retarded case of this kind' (p. 2). Witmer goes on to quote the President of the American Association for the Advancement of Science noting that 'the final test of the value of what is called science is in its applicability … the pure and the applied sciences advance in a single front. What retards the progress of one, retards the progress of the other; what fosters one fosters the other' (Witmer 1996, p. 2).

In an effort to bridge the science and practice of psychology, Witmer linked his psychological clinic to the psychology laboratory at the University of Pennsylvania. His clinical cases also comprised part of the curriculum for the child development courses. Witmer proposed a 'scheme of practical work in psychology' that laid the foundation of clinical, counseling, and school psychology. This consisted of the investigation of the phenomena of mental development by means of statistical and clinical methods, a clinic, supplemented by a training school, practical experience for those working with children, and the training of students for a new profession—the psychological expert. Forty years later these guidelines were adopted almost in their entirety by the Boulder Conference on education and training in clinical psychology.

Numerous other children's psychological clinics were developed during the first few decades of the twentieth century. Usually these were associated with

universities but some were established to provide clinical intervention for juvenile delinquents. Clinical psychologists were also working in facilities for the mentally retarded and became heavily involved in intelligence testing. Both basic and applied psychologists extended psychology's expertise in 'mental' testing to the assessment of recruits into the military during World War I. Unfortunately, this began one of the most tragic episodes in psychology's history—the use of intelligence testing to keep thousands upon thousands of immigrants from relocating in the USA.

Debates about appropriate educational and training models for the proper roles and functions of psychology began in the earliest days. Most psychologists, especially those who had trained with Wundt, envisioned psychology as an evolving science which would explore the intricate processes of consciousness and memory, unlock the mysteries of sensory processes, and answer questions about how people perceive and make sense of the world. Psychological findings would be based on the most rigorous of experimental methods. Other psychologists, faced with practical and poignant human problems, such as a child's failure to learn, believed that psychology could advance knowledge through a consideration of practical issues and problems and be immediately applied to help people in need.

While the parallels are not perfect, these differences in epistemological styles reflect a philosophical conflict of some 2,500 years. A Socratic view of education places emphasis on curiosity and the accumulation of knowledge for its own sake while the Sophists emphasize human welfare and believe 'man is the measure of all things' (Kimble 1984). Within early psychology, William James (1907) described the history of philosophy as a constant battle between these two sharply contradictory intellectual styles. He described the 'tough-minded' skeptical, empirical scientific psychologists who depend only on facts vs. the 'tender-minded' rational and idealistic sort who base their efforts on principles. James commented on the antagonisms between the 'unrefined, callous, or brutal' tough-minded and the 'sentimental, soft-headed,' tender-minded (James 1907). Later, Snow (1964) wrote extensively about the clash of the 'two cultures' of science and humanism and Kimble (1984) extended this observation into psychology. Kimble suggests that contemporary psychology continues to have an identity problem in that the two cultures of the scientist and practitioner cannot agree on a coherent set of values.

This clash of cultures, as exemplified by the different values embraced by the scientific vs. the practitioner psychologist has always been with us. As early as the turn of the twentieth century, applied and clinical psychologists were complaining that the newly formed American Psychological Association (APA) was not meeting their needs. Practicing psychologists were particularly concerned that they be recognized through some formal mechanism as having been trained in psychology. In 1917, an American Association of Clinical Psychology was proposed as a vehicle whereby psychologists who were qualified to perform psychological testing could be identified. With some reluctance, APA agreed to certify some members as 'consulting psychologists' and formed a Clinical Section. The primary mission of APA, however, was to promote psychology as a science.

As practitioners involved themselves in psychological testing, child guidance, and helping individuals with 'adjustment' problems, issues of ethics, licensing, and standardization of training became paramount. In the early 1930s, the New York Psychological Association in an effort to meet practitioner needs at a national level became the Association of Counseling Psychologists. They became so strong that in 1937 the Clinical Section of APA disbanded to join them in the new American Association for Applied Psychology (AAAP).

AAAP began to move across the national scene to meet the needs of practitioners and state affiliates in regard to uniform training, certification, public relations, and ethical conduct. The group proposed a model certification act to gain licensure or registration for qualified psychologists and published a membership directory to identify such practitioners for the public. They proposed an ethical code and pushed for a curriculum, both undergraduate and graduate, that would include applied and clinical psychology. Further, Carl Rodgers (1939) proposed a supervised, one year internship training for clinical psychology graduate students and discussed the idea of a doctor of psychology degree, rather than the Ph.D. (Napoli 1981).

The advent of World War II had a powerful influence on organized psychology. In an effort to present a unified front for responding to the war efforts and the needs of the military, an Emergency Committee in Psychology was established that included representatives from all of the national psychology organizations (now up to six). A subcommittee on Survey and Planning for Psychology, chaired by Robert Yerkes, was formed to begin making long-range plans for psychology's future. Their first report deplored the splintering of psychology into various groups and proposed an increased emphasis on professional activities within the APA. They further proposed a new national organization that would serve both science and practice. When the war was over, the stage was set for the APA to reorganize to serve the needs of psychologists, whether their interests be basic or applied psychology. The two cultures of science and practice in psychology would be subsumed, once again, within one organization.

Psychologists had served in a number of capacities during World War II, ranging from the regular duties of the infantry to preparing propaganda and disseminating disinformation. Basic psychological scientists

became involved in human factors designing user-friendly instruments for vehicles and planes. Clinical psychologists found themselves providing counseling and psychotherapy for soldiers and by the end of the war, some 450 clinical psychologists were serving in the US Army.

At the urging of the National Research Council, the disparate national groups of psychologists began to meet together at the end of the war. Acting on recommendations from the Yerkes committee, the APA revised its by-laws to 'advance psychology as science, a profession, and a means of promoting human welfare' and incorporate other groups and representatives from state associations. The AAAP became the Clinical Division of APA and professional psychologists began to feel more at home in the then 4,000 member organization.

The National Institute of Mental Health and the Veteran's Administration were eager for clinical psychologists to continue to provide services for veterans who were suffering psychological and emotional problems. In fact, they began to subsidize training to increase the numbers of psychiatrists and social workers. They were willing to do the same for clinical psychology if appropriate training programs could be identified. APA provided the names of 22 such university programs and appointed a Committee of Training in Clinical Psychology to determine how best clinical psychologists might be trained (American Psychological Association 1947). The report of this Committee became the basis for a 1949 conference held in Boulder, Colorado which would recommend a revolutionary program of education and training. Clinical psychologists would be trained as both scientists and practitioners. Such an integrative model, unique among the major sciences and professions, would bridge scientific and humanistic traditions.

The Boulder conference was supported by the National Institute of Mental Health and the United States Public Health Service. Seventy-one representatives from universities, mental health service agencies, and allied professions met daily for two weeks to define clinical psychology and develop consensual standards for education and training. A 252 page book, *Training in Clinical Psychology*, reported the recommendations of the conference which were both comprehensive and specific (Raimy 1950). Clinical students' attributes were enumerated, faculty described, and curriculum proposed. Clinical psychologists were to be trained in general psychology and demonstrate research competence culminating in a doctoral (Ph.D.) dissertation. Further, students would engage in field training and a predoctoral internship so that they would be skilled in clinical assessment, diagnosis, and treatment. Such education and training was to take place in universities and internship facilities accredited under the auspices of the APA.

During the second half of the twentieth century, the Boulder or scientist–practitioner model became the preeminent mode of education and training in clinical psychology. Admissions to university-based, Ph.D.-granting programs were, and continue to be, highly selective. Academic requirements include general areas of psychology such as history and systems and psychopathology as well as cognitive, developmental, experimental, personality, physiological, and social psychology. Students are expected to learn to conduct research and be familiar with the various statistical methods necessary to evaluate results. Further requirements include practical training in assessment and intervention under close supervision including a one year predoctoral internship. The scientist–practitioner model has been reaffirmed at each major graduate educational conference held since Boulder although a professional model was also recommended at a 1973 conference held in Vail, Colorado (Strickland 1985, 1987, 1998).

The scientist–practitioner model is not reserved solely for clinical psychology. Other applied areas in psychology such as counseling and industrial-organizational psychology also embrace this paradigm as do the specialty areas of aging and health. For those who believe that psychology is a continually evolving discipline and field, basic training in science would seem essential in producing new knowledge. Further, practical training within such a model should lead to state-of-the-art applications. An integration of science and practice allows the continued interplay between scientific knowledge and clinical application where each informs the other.

Many would maintain that the development of the scientist–practitioner model has been one of psychology's finest achievements. Certainly for a newly evolving science which has clear implications for application this novel educational and training approach gives the field vitality and direction. Embracing the two cultures of science and humanism, the scientist–practitioner model has not been an easy one for individual psychologists to achieve. But, scientist–practitioners have served psychology and society well both through advancing knowledge and applying that knowledge for the greater good.

Bibliography

American Psychological Association Committee on Training in Clinical Psychology (D. Shakow, Chairman) 1947 Recommended graduate training programs in clinical psychology. *American Psychologist* **2**: 539–58

Dollard J, Miller N E 1950 *Personality and Psychotherapy*. McGraw-Hill, New York

James W 1907 *Pragmatism: A New Name for Some Old Ways of Thinking*. Longmans, New York

Kimble G 1984 Psychology's two cultures. *American Psychologist* **39**(8): 833–9

Napoli D S 1981 *Architects of Adjustment*. Kennikat Press, Port Washington, NY

Raimy V C 1950 *Training in Clinical Psychology*. Prentice-Hall, New York
Snow C P 1964 *The Two Cultures: And a Second Look*. The New American Library, New York
Strickland B R 1985 Over the Boulder and through the Vail. *The Clinical Psychologist* **38**: 52–6
Strickland B R 1987 On the threshold of the second century of psychology. *American Psychologist* **42**: 1055–6
Strickland B R 1988 Clinical psychology comes of age. *American Psychologist* **43**: 104–7
Strickland B R 1998 History and introduction to clinical psychology. In: Cullari S (ed.) *Foundations of Clinical Psychology*. Allyn and Bacon, Boston, pp. 1–26
Witmer L 1996 Clinical psychology (reprinted from *The Psychological Clinic* **1**: 1–9, 1907) *American Psychologist* **51**(3): 248–51

B. R. Strickland

Transaction Costs and Property Rights

Transaction costs and property rights are related, important, and expansive concepts. Both move beyond technology to introduce law and organization. Both deal with issues of first-order economic importance (as against nuances). Because both have been difficult to operationalize, both have been used expansively and tautologically.

Property rights is the older concept, going back to antiquity. The use of transaction cost reasoning in the 1960s brought new life to the property rights movement. Not only is the economics of property rights basic, enduring, and important in its own right, but it also helped to usher in and remains an important part of the new institutional economics: 'Modern institutional economics focuses on the institution of property and on the systems of norms governing the acquisition or transfer of property rights' (Furubotn and Richter 1991, p. 7). Of the various categories of property rights, the right of ownership—which consists of the right to use an asset, the right to appropriate the returns from an asset, and the right to change its form, substance, or location (Furubotn and Richter 1991, p. 6)—is the most important.

The article begins with a sketch of the antecedents, then some of the leading property rights successes are examined, followed by the limitations and the conclusion.

1. Antecedents

Frank Michelman (1967) traces property rights concepts and thinking from Aristotle through the Christian Fathers, John Locke, David Hume, Jeremy Bentham, and Oliver Wendell Holmes. Despite this early and continuing interest, the economics of property rights 'has not been discussed analytically until very recently' (Demsetz 1998, p. 144). The introduction of transaction cost reasoning into the study of property rights in the 1960s is responsible for much of the analytical advance.

2. Modern Property Rights

2.1 Externalities

Work on the economics for property rights flourished in the 1960s. Leading contributors included Ronald Coase (1959, 1960), Armen Alchian (1961, 1965), and Harold Demsetz (1967). Much of this work contested the orthodox interpretation of externalities as 'market failures' for which government intervention 'in the form of taxes, or subsidies, or regulation' was presumptively warranted. Rather than take property rights assignments as given, Coase moved the analysis back a step to examine the ramifications of reassigning property rights. His treatment brought out 'the reciprocal nature of the relationship which tends to be ignored by economists who, following Pigou, approach the problem in terms of a difference between private and social producers but fail to make clear that the suppression of the harm which A inflicts on B inevitably inflicts harm on A. The problem is to avoid the more serious harm' (Coase 1959, p. 24). Assessing the more serious harm entailed introducing transaction costs.

Coase proceeded in a two-part way. Starting with the assumption of zero transaction costs, Coase advanced the (then) revolutionary proposition that 'the delimitation of rights is an essential prelude to market transactions; but the ultimate result (which maximizes the value of production) is independent of the legal [award of property rights to one party rather than the other]' (Coase 1959, p. 25). This is because, distributional considerations aside, the parties will bargain to the same efficient result whichever way property rights are awarded.

For Coase, however, this was an expositional device. The real world is one of positive transaction costs, whereupon the assignment of property rights one way or another does matter. Efficiency now turns on the differential transaction costs of assigning property rights one way rather than another. Coase furthermore contested the idea that a 'benign government' could be used to correct market failures, it being the case that all feasible forms of organization are flawed (Coase 1964). He also advised that total as well as marginal conditions needed to be included in a comparative assessment: there is a need to 'compare the total product yielded by alternative social arrangements' (Coase 1960, p. 42). This comparative approach in which transaction costs are featured reshaped the

study of property rights, market failures, and economic organization in the ensuing years, up to and including the present.

2.2 Common Pool Problems

Common pool problems arise in a variety of contexts, of which common grazing land (Hardin 1968), fisheries (Turvey 1964), petroleum extraction (Libecap and Wiggins 1985), and water and air pollution are familiar examples. Whereas the externalities examined by Coase were predominantly of a bilateral kind, the common pool problem poses the need to coordinate many actors. Additional problems arise because (a) different actors are differently situated, (b) the relevant information is dispersed and costly to verify, (c) strategizing and politicking often arise, and consent is needed to, (d) agree on an overarching plan, and (e) craft an organization to implement whatever plan is agreed to. The need for political action and regulation often arises in efforts to reach and enforce collective agreements in the multiactor context.

Recourse to public ordering, is not, however, the only or always the best response. Private ordering can and has successfully mitigated many common pool problems (Furubotn and Richter 1997, p. 102). Government assistance— for unitizing common pools (such as oil) or for creating markets (as in the creation of tradeable property rights for air pollution)—is often the cost-effective way to proceed.

2.3 Variable Access

The simplest transaction is an outright sale. But even that is complicated by the question of how the property right is defined. As Demsetz observes, 'To have the right to till a parcel of land is not equivalent to having all the rights to the land. Others may have rights both to minerals contained in the land and to trespass across the parcel. The identification of the owner of the asset is made more difficult by the virtually infinite number of different rights that can be associated with use of a given asset' (1998, p.145).

Access arrangements other than sale include leasing, tenant farming, inside contracting, and franchising. These can vary in length and protections. Incentives for due care (where the concern is over negligence by the lessee) and assured compensation for unexpired improvements (where the lessor stands to gain by refusing compensation for improvements made by the lessee that extend beyond the length of the lease) are common concerns. Access terms is also an area where the risk of changes in the political rules of the game is a chronic concern, the imposition of rent control being one example and the arbitrary exercise of imminent domain and 'takings' by the state being another (Michelman 1967).

2.4 Theory of the Firm

The theory of the firm poses numerous property rights issues, of which the most famous is the separation of ownership from control to which Adolf Berle and Gardiner Means (1932) made early reference in the context of the large corporation where shareholding is widely dispersed. Thus although shareholders remain *de jure* owners in the large corporation, effective or *de facto* ownership purportedly passes to the management. Berle and Means thereupon questioned the presumption that firms are reliably given to profit maximization.

Analytical reactions to this condition of separation have varied. One response has been to dismiss the issue by assuming that competition in product and capital markets is relentlessly taut. As Tjalling Koopmans has observed, however, 'if this is the basis of our belief in profit maximization, then we should postulate that basis itself and not the profit maximization which it implies in certain circumstances .... It would lead us to expect profit maximization to be most clearly exhibited in industries where entry is easiest and where the struggle for survival is keenest, and would present us with the further challenge to analyze what circumstances give to an industry that character' (1957, pp. 140–1).

Another response to the separation of ownership from control is to reformulate the objective function to reflect *de facto* control rights. The sales maximization hypothesis (Baume 1959), growth maximization hypothesis (Marris 1964), and managerial discretion hypothesis (Williamson 1964) are in this spirit. Still another response is to reconceptualize the problem of organization in contracting terms. The 'nexus of contract' approach (Jensen and Meckling 1976) brings out agency cost features to which monitoring is a response, while the governance structure approach (Williamson 1975, 1985) treats organization form as a decision variable. According to the latter, hierarchical decomposition principles—especially between centralized and decentralized modes of organization—matter. Organization is thus viewed both as a source of the managerial discretion problem and a partial solution, in that the large corporation can be reconfigured to mitigate some of its own failures.

A 'modern' property rights theory of the firm, of which Oliver Hart (1995) is the principal architect, has taken shape more recently. This theory works out of an incomplete contracting setup. The central issue with which this theory is concerned is that of vertical integration: who should own the assets of two related stages of production? This theory illustrates both the strengths and limitations of property rights reasoning. To its credit, the analysis makes clear how the ownership of property rights influences *ex ante* investments. But this exclusive focus on property rights comes at a cost, in that problems of *ex post* mal-

adaptation are suppressed (by assuming that the parties will costly bargain to an efficient result, whatever the assignment of property rights). The idea that organization matters—that the management of some transactions is better done by markets and of others by hierarchy—is sacrificed in the process.

2.5 Rent Seeking

The *de facto* ownership of property rights poses numerous intertemporal issues, of which rent seeking is one. The basic rent seeking proposition is this: the creation of 'rights' to which supernormal returns potentially accrue invites *ex ante* prepositioning, thereby to better qualify the applicant/bidder for the job or contract in question. For example, the award of a natural monopoly franchise or of a favored class of employment (Civil Service) invites parties to politick to receive the award and to make other efforts to improve their eligibility. Rent protective efforts also ensue (Tollison 1998).

In the degree to which such responses went unnoticed or were ignored by economists, such costs were mistakenly omitted from the original calculus. Upon 'discovering' such costs, most economists declared them to be wasteful. That, however, assumes that economics trumps politics (Stigler 1992). If political choices are to be respected, then the best course of action is to recognize that prepositioning is a predictable consequence of efforts to create new classes of rights, whereupon prepositioning effects are folded into the project design calculus. In that event, rent seeking, and its attenuation, are included in the 'approved plan.' Those who condemn rent seeking are loathe to come to terms with this proposition.

2.6 Economic Development and Reform

Early development theory emphasized 'macroeconomic accounting aggregates such as savings and the balance of payments, and the relative balance between broadly defined "sectors" such as "industry" and "agriculture"' (Lal 1983, p. 5). When that prescription proved disappointing, the pendulum swung in the opposite direction. The new imperative was to activate the market and 'get the prices right' (Lal 1983, p. 107), but that too was overly simple.

'Getting the property rights right' seemed more responsive to the pressing needs for reform in Eastern Europe and the former Soviet Union. But while 'an essential part of development policy is the creation of polities that will create and enforce efficient property rights, … we know very little about how to create such polities' (North 1994, p. 366). As discussed in Sect. 3.2.2, privatization is important but is not a panacea.

2.7 De Facto *Property Rights*

Just as the study of contract has benefited by going beyond a legal rules concept of contract to make prominent provision for private ordering (Llewellyn 1931, Galanter 1981), so too does the study of property rights benefit by going beyond *de jure* property rights to consider *de facto*. The recent application of *de facto* reasoning to explain economic development and reform in China (Qian and Weingast 1996) is illustrative. The dilemma was this: *de jure* privatization was politically unacceptable, yet the state-owned enterprises in China had weak incentives and poor performance. What to do?

Decentralization, by creating local township and village enterprises in which local government remained prominent, turned out to be a productive move. As Qian and Weingast put it (1996, pp. 87–8):

> These local government-owned enterprises have remarkably different governance structures from (central) state-owned enterprises and thus face better positive and negative incentives.… Politicians in every political system tend to bail out inefficient firms or spend on wasteful consumption. In a federal system, however, the mobility of resources across regions raises the opportunity costs to local governments of bailing out inefficient firms or wasteful public expenditures.

Not only do local governments share the successes of local enterprises, but they incur the risk that investments will relocate if governments interfere. Federalism is thus an instrument for realizing quasi-privatization and the incentive benefits that accrue thereto. *De facto* property rights thus vary with the political environment, *de jure* property rights unchanged.

There are nevertheless limits to the risks that parties will incur in a regime where *de jure* credibility is lacking. I conjecture that China will experience serious incentive obstacles in its attempts to develop leading edge technologies unless it provides more secure *de jure* supports for intellectual property rights.

3. Limitations

Although property rights concepts and property rights reasoning have gone a long way to reshape our understanding of economic institutions and effect public policy reform, the early ambitions of the economics of property rights have been incompletely realized. Here, as elsewhere, there can be too much of a good thing.

3.1 Oversimplification

According to Coase, 'a private-enterprise system cannot function properly unless property rights are

created in resources, and, when this is done, someone wishing to use a resource has to pay the owner to obtain it. Chaos disappears; and so does the government, except that a legal system to define property rights and to arbitrate disputes is, of course, necessary' (Coase 1959, p. 12).

This property rights solution applied not merely in general but even to such apparently intractable problems as allocating electromagnetic spectrum to radio broadcasters: 'there can be little doubt that, left to themselves, the courts would have solved the problems of the radio industry in much the same way [by defining and enforcing property rights] as they had solved similar problems in other industries' (Coase 1959, p. 30). Other interested parties were less sanguine. Chief Justice Taft, for example, resolutely 'dodged the radio question. I have refused to grant writs and have told the other justices that I hope to avoid passing on this subject as long as possible' (Coase 1959, p. 30).

Indeed, what would the courts, left to themselves, have done? Taft was baffled by the radio question because property rights were ill-defined and the courts did not know how to define them. Awaiting a workable definition of property rights in electromagnetic spectrum—with the help of economists, lawyers, engineers, or managers—a property rights 'solution' remained elusive.

As Arthur DeVany put it, 'the real stumbling-block to a market system of spectrum allocation is the interference problem. The spectrum is a multi-dimensional space; signals propagate indefinitely in the frequency, spatial, and time domain' (1998, p. 169). The team of DeVany, Ross Eckhart, Dan O'Hara, Charles Meyers, and Richard Scott (three economists, a lawyer, and a physicist) turned their attention to this matter in 1969. Their best efforts to define property rights that would be 'exclusive, predictable, capable of subdivision, enforceable and flexible' (1998, p. 169) notwithstanding, DeVany now concedes that their 1969 definition was merely a template and needed further refinements (1998, p. 169).

Indeed, going beyond the definition of property rights, there is a further problem: how to devise an auction? Although DeVany avers 'that our model has been used to privatize the spectrum in New Zealand, Australia, and Latin America' (1998, p. 168), Preston McAfee and John McMillan maintain that 'simultaneous single-round, sealed-bid auctions were used in New Zealand for spectrum licenses and in Australia for satellite-television licenses with disappointing results: low revenues and inefficient license allocations' (1996, p. 162). New developments in auction theory, which did not take shape until the 1980s and 1990s, were also needed. The simultaneously ascending auction was adopted by the FCC in the 1990s to auction spectrum for personal communications services: mobile telephones, two-way paging, portable fax machines, and wireless computer networks. Some judge that auction to be a 'success' (McAfee and McMillan 1996, p. 163), but the actual sale of spectrum turned out to be 'more complicated that anything in auction theory' (McAfee and McMillan 1996, p. 171).

Merely to conduct an auction ought not to be confused, moreover, with full privatization. In the US the FCC remains an active player in the management of spectrum. The upshot is that defining and enforcing property rights for some resources is a daunting exercise. Politics often continue, usually with the active participation of the key players. Realpolitik and partial privatization are evidently bedfellows.

3.2 Overreaching

3.2.1 The modern corporation. Might the salient features of the modern corporation be explained in property rights terms? Harold Demsetz took this position. Upon observing that there are 'significant economies of scale in the operation of large corporations … [and] that large requirements for equity capital can be satisfied more cheaply by acquiring the capital from many purchasers of equity shares' (1967, p. 358), Demsetz described the modern corporation in the following property rights terms: (a) efficiency is realized by delegating effective ownership to the management; (b) shareholders are essentially lenders of equity capital; and (c) limited liability relieves shareholders of the need to carefully examine corporate liabilities and the assets of other shareholders (as they would need to under partnership law).

This property rights interpretation of the modern corporation over-reaches in the first two respects. First, to suggest that diffuse ownership is responsible for delegation, whereupon the 'management group becomes the *de facto* owners' and that ownership is 'legally concentrated in management's hands' (Demsetz 1967, p. 358) is incorrect. For one thing, delegation is observed in large corporations whether there are many or few owners of equity shares. For that matter, delegation is observed in every large organization—public, private, nonprofit. Well-known limits on the span of control (from organization theory) rather than the economics of property rights are responsible for that result. Additionally, of the three rights of ownership—the right to use, transform, and appropriate the income from an asset—the management is never delegated this last. Equity is the residual claimant.

Second, to describe shareholders as 'lenders' of equity capital is misleading. Debt and equity are not merely instruments of finance; they are also instruments of governance. If shareholders are merely lenders, why not just sell collateralized bonds? Or why not loans from the banks? It is not for naught that equity investors, who invest for the life of the firm and are residual claimants, are awarded control of the board of directors (Fama and Michael 1983). The

efficiency logic of debt and equity is usefully viewed as a governance issue (Williamson 1988).

3.2.2 Russian privatization. The architects of privatizing Russian industry in the early 1990s appealed to the modern property rights theory of Sanford Grossman and Oliver Hart (1986) to urge mass and rapid privatization of all sectors of Russian industry (Boycko et al. 1995). The overarching theory was this: 'When property rights over a productive asset are clearly specified, and the person who decides how to employ his asset bears full costs and enjoys the full benefits of employment, he puts the asset to its most productive use' (Boycko et al. 1995, p. 19). Upon privatizing two-thirds of Russian industry in June 1994, Boycko et al. boldly declared that privatization had been brought to a 'triumphant completion' (1995, p. 8).

Subsequent events have been unkind. Although it may be unduly pessimistic to view Russian privatization as a failure (Braguinsky 1999, Black et al. 1999), it is undisputed that Russian privatization has not been a triumphant success.

Errors of two kinds explain the premature pronouncement of success. First, the property rights theory of the firm suppresses problems of *ex post* governance by assuming costless renegotiation. That has analytical advantages, but it is also disconnected from reality. Had Boycko et al. inquired into contract implementation, they would have discovered that the efficacy of privatization varies, working much better in some industries than others (Williamson 1976, Goldberg 1976, Priest 1993). Second, Boycko et al. never faced up to lapses in the laws and, even more, their efficacious enforcement in Russia. Failures to look beyond the immediate award of property rights and examine the needs for *ex post* governance support and effective legal recourse turned out to be fateful.

As with the allocation of electromagnetic spectrum and interpreting the modern corporation, so too with Russian privatization. Thus although property rights reasoning is pertinent to all, the economics of property rights is a self-limiting perspective. Awaiting a unified theory of economic organization, which is not in prospect, the application of several lenses—to include an examination of hazards that beset *ex post* governance and the condition of the institutional environment—has a lot to recommend it.

4. Conclusions

The economics of property rights is unarguably important. Were it that the Socialist Controversy in the 1930s was less preoccupied with technicalities (such as marginal cost pricing) but addressed itself more to an assessment of the property rights differences between capitalism and socialism, the field of comparative economic systems would have developed very differently (Demsetz 1998). But while property rights are crucial, they are not fully determinative of economic organization. To claim that efficiency will be realized upon defining and enforcing property rights simply over-reaches. Not only can property rights be difficult to define, but best efforts at court enforcement can be costly. Also, property operates in combination with contract and organization.

Kenneth Arrow's remarks about externalities, market failures, and transaction costs help to restore perspectives: 'I contend that market failure is a more general category than externality [and, because] market failure is not absolute … , it is better to consider a [still] broader category, that of transaction costs, which in general impede and in particular cases completely block the formation of markets' (1969, p. 48). Transaction cost is an inclusive concept that applies to property, contract, and governance.

See also: Development: Socioeconomic Aspects; Firm Behavior; Property Rights

Bibliography

Alchian A A 1961 *Some Economics of Property*, RAND D-2316. RAND Corporation, Santa Monica, CA

Alchian A A 1965 The basis of some recent advances in the theory of management of the film. *Journal of Industrial Economics* **14** (December): 30–41

Arrow K 1969 The organization of economic activity: Issues pertinent to the choice of market versus nonmarket allocation. In: *The Analysis and Evaluation of Public Expenditure: The PPB System*. US Joint Economic Committee, 91st Congress, 1st Session. US Government Printing Office, Washington, DC, Vol 1, pp. 59–73

Baume W J 1959 *Business Behavior, Value and Growth*. Macmillan, New York

Berle A A, Means G O Jr 1932 *The Modern Corporation and Private Property*. Commerce Clearing House, New York

Black B, Kraakman R, Tarassova A 1999 Russian privatization and corporate governance: What went wrong? Unpublished manuscript. Stanford Law School.

Boycko M, Shleifer A, Vishny R 1995 *Privatizing Russia*. MIT Press, Cambridge, MA

Braguinsky S 1999 Enforcement of property rights during the Russian Transition: Problems and some approaches to a New Liberal solution. *Journal of Legal Studies* **28** (June): 515–44

Coase R 1959 The federal communications commission. *Journal of Law and Economics* **2** (October): 1–40

Coase R 1960 The problem of social cost. *Journal of Law and Economics* **3** (October): 1–44

Coase R 1964 The regulated industries: Discussion. *American Economic Review* **54** (May): 194–7

Demsetz H 1967 Toward a theory of property rights. *American Economic Review* **57** (May): 347–59

Demsetz H 1969 Information and efficiency: Another viewpoint. *Journal of Law and Economics* **12** (April): 1–22

Demsetz H 1998 Property Rights. In: Newman P (ed.) *New Palgrave Dictionary of Economics and Law*. Stockton Press, New York, Vol. 3, pp. 144–55

DeVany A, Eekert R, Meyers C, O'Hara D, Scott R 1969 A property system for the allocation of the electromagnetic spectrum. *Stanford Law Review* **21** (June): 1499–1561
DeVany A 1998 Property rights in electromagnetic spectrum. In: Newman P (ed.) *New Palgrave Dictionary of Economics and Law*. Stockton Press, New York, Vol. 3, pp. 167–71
Fama E F, Michael C J 1983 Separation of ownership and control. *Journal of Law and Economics* **26** (June): 301–26
Furubotn E, Richter R 1991 The new institutional economics: An assessment. In: Furubotn E G, Richter R (eds.) *The New Institutional Economics*. Texas A&M Press, College Station, TX, pp. 1–32
Furubotn E, Richter R 1997 *Institutions and Economic Theory*. University of Michigan Press, Ann Arbon, MI
Galanter M 1981 Justice in many rooms: Courts, private ordering, and indigenous law. *Journal of Legal Pluralism* **19**: 1–47
Goldberg V 1976 Regulation and administered contracts. *Bell Journal of Economics* **7**: 426–52
Grossman S J, Hart O D 1986 The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* **94** (August): 691–719
Hardin G 1968 The tragedy of the commons. *Science* **162**: 1243–8
Hart O 1995 *Firms, Contracts, and Financial Structure*. Oxford University Press, New York
Jensen M, Meckling W 1976 Theory of the firm: Managerial behavior, agency costs, and capital structure. *Journal of Financial Economics* **3** (October): 305–60
Koopmans T 1957 *Three Essays on the State of Economic Science*. McGraw-Hill, New York
Lal D 1983 *The Poverty of Development Economics*. Institute of Economic Affairs, London
Libecap G D, Wiggins S N 1985 The influence of private contractual failure on regulation: The case of oil field unitization. *Journal of Political Economy* **93** (June): 690–714
Llewellyn K N 1931 What price contract? An essay in perspective. *Yale Law Journal* **40**: 704–51
Marris R L 1964 *The Economic Theory of Managerial Capitalism*. Free Press of Glencoe, New York
McAfee R P, McMillan J 1996 Analyzing the airwaves auction. *Journal of Economic Perspectives* **10** (Winter): 159–75
Michelman F 1967 Property, utility and fairness: Comments on the ethical foundations of 'Just compensation' law. *Harvard Law Review* **80** (April): 1165–257
North D C 1994 Economic performance through time. *American Economic Review* **84** (June): 359–68
Priest G L 1993 The origins of utility regulation and the 'Theories of Regulation' debate. *Journal of Law and Economics* **36**: 289–323
Qian Y, Weingast B 1996 China's transition to markets: Market preserving federalism, Chinese style. *Journal of Policy Reform* **1**: 149–85
Stigler G 1992 Law or Economics? *Journal of Law and Economics* **35** (October): 455–68
Tollison R 1998 Rent seeking. In: Newman P (ed.) . *The New Palgrave Dictionary of Economics and the Law*. Stockton Press, New York , Vol.3, pp. 315–22
Turvey R 1964 Optimization and suboptimization in fishery regulation. *American Economic Review* **54** (March): 64–76
Williamson O E 1964 *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ
Williamson O E 1975 *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York
Williamson O E 1976 Franchise bidding for natural monopolies—In general and with respect to CATV. *Bell Journal of Economics* **7** (Spring): 73–104
Williamson O E 1985 *The Economic Institutions of Capitalism*. Free Press, New York
Williamson O E 1988 Corporate finance and corporate governance. *Journal of Finance* **43** (July): 567–91

O. E. Williamson

Transfer of Learning, Cognitive Psychology of

1. *Scope of the Phenomenon and Definition of 'Transfer'*

Imagine someone handling a new text processing program on a PC, driving a car in Australia instead of in the USA, playing the counter-back role instead of the one used to, placing a bond portfolio under changed economic conditions, or meeting the new boss for the first time. In all these cases people are learning some specific new behavior using previously acquired knowledge and skills. For this phenomenon the term 'transfer of learning' was coined. The knowledge used may be of different kinds: (a) conceptual (known facts), (b) procedural (skills), or (c) metacognitive (evaluating or regulating one's own cognitive activities). The corresponding research has always focused on understanding and predicting extent, direction, and 'locus' of transferred knowledge (for overviews, see Cormier and Hagman 1987 and Detterman and Sternberg 1993).

2. *The Quest for a General Perspective of Learning and Transfer*

In order to give an overview, a general perspective for learning and transfer is outlined here. Learning, on the one hand, is thought (a) to occur within an immense body of prior knowledge, and (b) to be a process of continual integration of knowledge by constructions or reconstructions. Such (re-)constructions include:

(a) relating meaningful units of different conceptual levels that represent objects and their features,

(b) organizing these related elements into larger units according to the constraints imposed by the content structure of its concepts (e.g., money and goods have to be related first in order to become organized further to explain the higher-order concept of 'inflation'), and

(c) chunking: the process of *condensing* built-up structures for both memory capacity reasons and ease of subsequent use. The chunks may be related further

to other units of different conceptual levels, organized, and chunked anew. Chunks are indexed for retrieval and further processing, highlighting the top-level or goal concept: 'and all this is called inflation'; or 'and this is how acceleration is calculated in mechanics'; or 'this protection by the bird's color almost guarantees survival.'

Transfer, on the other hand, is thought of as a ubiquitous, continuous, systematic use of selected parts of the immense body of prior knowledge. It starts from a cognitive need for information that has to be (a) retrieved from memory, (b) selected for tentative application, and (c) mapped on a corresponding subprocess of learning as described above.

3. *Transfer of Skills: The 'Common Element' Theory of Transfer—An Old and a Modern Version*

Experimenters in cognitive psychology look for learning or, preferably, problem-solving tasks that can be constructed such that identities or similarities between two tasks (a source and target task) can be controlled. Some of the favorite research domains include learning programming languages, use of text processing programs, solving algebra word problems mainly from physics, and motor ability training.

3.1 *Playing the Violin—An Example*

One of the most obvious transfer situations can be illustrated by the use of common elements in playing a musical instrument, e.g., the violin. When practicing, a sequence of several notes and accordingly a sequence of finger movements of the left hand are learned. Once mastery has been achieved, the required skill can be applied to the whole range of the violin's touch board. This is the violinist's economy to use the practiced motor sequences to generate the very same melody in different pitches. This 'very same melody' was the kind of holistic unit or 'totality' Christian von Ehrenfels (Ehrenfels 1890) was working on, calling it a 'Gestalt,' a term used for a whole movement early in this century: the Gestalt psychology. The underlying common elements in violin playing that are transferred from string to string or from one hand shift to another are complete knowledge structures (chunks) containing the condensed motor program or action plan for performing that invariant melody. The specific parameters of the respective program or plan may be adapted if necessary, e.g., the higher the shift position of the left hand the shorter the absolute distances between the set fingers will be. However, the proportions of the distances between the fingers and the single movements within the sequence remain the same. Motor programs do transfer if the relative timing of a motor sequence is not violated. The music pattern can be speeded up without too high a risk of a loss of transfer.

3.2 *Common Elements—An Early Approach*

A first version of a 'common element' theory of transfer was developed early in the twentieth century by Thorndike. He showed how difficult it is to transfer general abilities from one academic domain to another (Thorndike and Woodworth 1901), except in cases where both share common elements: stimulus–response associations, typical for the explanation of behavior at that time. Transfer increased proportionally with the number of such overlapping associations in the learning and the test tasks (Thorndike 1913).

3.3 *Production Systems—A Modern Approach*

The common element theory of transfer has made considerable progress during the twentieth century. One of the well-known contemporary theoretical approaches for explaining what kinds of common elements are transferred in mastered skills comes from Anderson's ACT theory in its several versions (Singley and Anderson 1989) (the acronym ACT stands for adaptive control of thought). In the beginning of the violin play, each of the single knowledge elements (the printed notes to be transformed into movements) are elements of declarative knowledge (facts), each having both a certain single and a relational musical meaning (regarding rhythm and harmony). Repeating the sequence leads to what Anderson calls knowledge compilation: Some tones get closer to each other, are 'heard together' as a subpart within the sequence until, finally, larger and larger subparts are experienced as forming a totality. With practice, procedural knowledge is gradually built up. It is composed of specific conditions followed by actions (IF … THEN associations, so-called productions). In our example the condition (the IF part of the production) can be defined by the set of notes; it is followed by the corresponding action (the THEN part), namely the finger movements. If learned and practiced in a violin sonata by Händel, the production would transfer to a Mozart sonata if the same class of conditions were met again. Productions are, thus, the common elements that can explain transfer of skills. The progress in transfer theory in the twentieth century consists of (a) the elaboration of the former stimulus and the response side of knowledge use by introducing the production concept and (b) the integration of productions into compound production systems for knowledge acquisition and use.

4. *Analogical Transfer*

The most extensive transfer research since the mid-1980s has focused on so-called analogical transfer whereby true analogies are found in those problems that share a similar deep structure but not necessarily specific content (e.g., the analogy of the atom and the solar system). According to current research, an adequate theory of analogical transfer would have to account for how analogical transfer develops over the processing time span, and for the role that surface and schematic knowledge play. The process would have to follow several stages as the following 'green walnut' example illustrates (Fig. 1).

4.1 *The Green Walnut—Another Example*

Assume that a young girl has acquired the skill to handle several kinds of fruit properly: small apples, bananas, and even a coconut with the help of her father. She is now holding a green walnut in her hand (some time before harvest, as it is still covered by its soft green skin that later dries out), an object she has never seen before.

If we look at the process of analogical transfer the following four stages can be distinguished:

(a) Stage 1 refers to encoding some elements of the target problem. Our young girl encodes her object as something small, round, and green (top right in Fig. 1) which reminds her of some other object with which she is familiar: small, green apples (top left). The corresponding REMINDING 1 subprocess is based on surface similarities between a visible object and an imagined apple (as an exemplar of the class of small apples).

(b) Stage 2: retrieval of one or more source analogues (i.e., elements from specific prior knowledge) when presented with the target problem. The retrieval process can result in a variety of working memory activities: spontaneous spreading-out, systematic decomposing of chunks, or inferencing, leading to a list of features such as … is juicy, … has skin, … is biteable, that are candidate analogs to some corresponding characteristics of the target task.

(c) Stage 3: selection and mapping of the source analogue to the target task. The crucial question is which one of the retrieved items will be adequate for use. There are constraints determined by the target task limiting the selection of candidates to be used for attaining the goal of the target task. To identify the object, the selection will be an action schema or an assimilatory schema (Piaget 1952). Thus, the 'bite' schema is selected (subprocess SELECTION 1) for mapping (subprocess MAPPING 1). The corresponding schema use is an application of abstract schematic knowledge, the schema being a generalized unit of the prior knowledge repertoire of the child. Mapping is the subprocess of establishing a contact with the main learning process where still further processing will take place: the schema application as an operation on that green object to be identified.

(d) Stage 4: knowledge integration. The green object is now assimilated by the 'bite' schema: The object is indeed biteable. This knowledge is easily integrated into the child's knowledge, but this is not necessarily the case, as will be shown below. Although the taste while biting is terribly bitter, there is the important additional recognition that the object has a removable skin. (Status and 'locus' of stage 4 subprocesses will be discussed in detail.)

From here it becomes obvious that transfer is an oscillating process with recurrent subprocesses: The removable skin is encoded (a recurring stage 1 subprocess) and reminds (REMINDING 2) the child of another object with removable skin, a banana. It is certainly a very fast retrieval process followed by a similarly fast selection (since the task's constraints do not leave a great choice) of the 'peel' schema (recurring stage 3 subprocesses SELECTION 2 and MAPPING 2). These are, again, followed by knowledge integration in the form of the application of the 'peel' schema. This 'peel' schema is successful with the recognition of a hard shell covering the object. In a next step the hard shell is encoded (another stage 1 subprocess), leading to REMINDING 3, which, in turn, lets the child recall another object with a hard shell, a coconut. Here, stage 2 subprocesses will recur: memories about opening that coconut when the saw from her father's toolbox had to be used. The list of features of the retrieved coconut is relatively long (… is brown, … has fibers around, … is shakable, etc.) but the 'open-it' schema prevails according to the constraints of the task at hand (SELECTION 3 as another recurrent stage 3 subprocess). The 'open-it' schema will be mapped (MAPPING 3) for tentative use. But the assimilation by the 'open-it' schema will not be realized. Before looking for the saw that was adequate for opening the coconut, her father will explain the inadequacy of the old 'open-it' schema and suggest the use of the nutcracker. This is, in Piaget's terms, an accommodation of the assimilatory schema that would otherwise not satisfy the constraints of the task. So, knowledge integration adopts a specific form in the last phase of the learning process: The 'open-it' schema is accommodated and adapted into a 'crack-open' schema which, then, serves for assimilation. This is, of course, a specific stage, *not* of transfer subprocesses but of the main learning process: the accommodation of a schema enabling it for final assimilation in a whole series of accumulating assimilations (notice the embedded rectangles labeled L_1 through L_3 in Fig. 1), leading finally to the identification of the once unknown object as a 'walnut.' The crucial point of this final knowledge integration stage is that important processes take place: triggering further constructions after the arrival of a mapped element, partially reorganizing knowledge leading to new structures, chunks or insights. These processes certainly do not

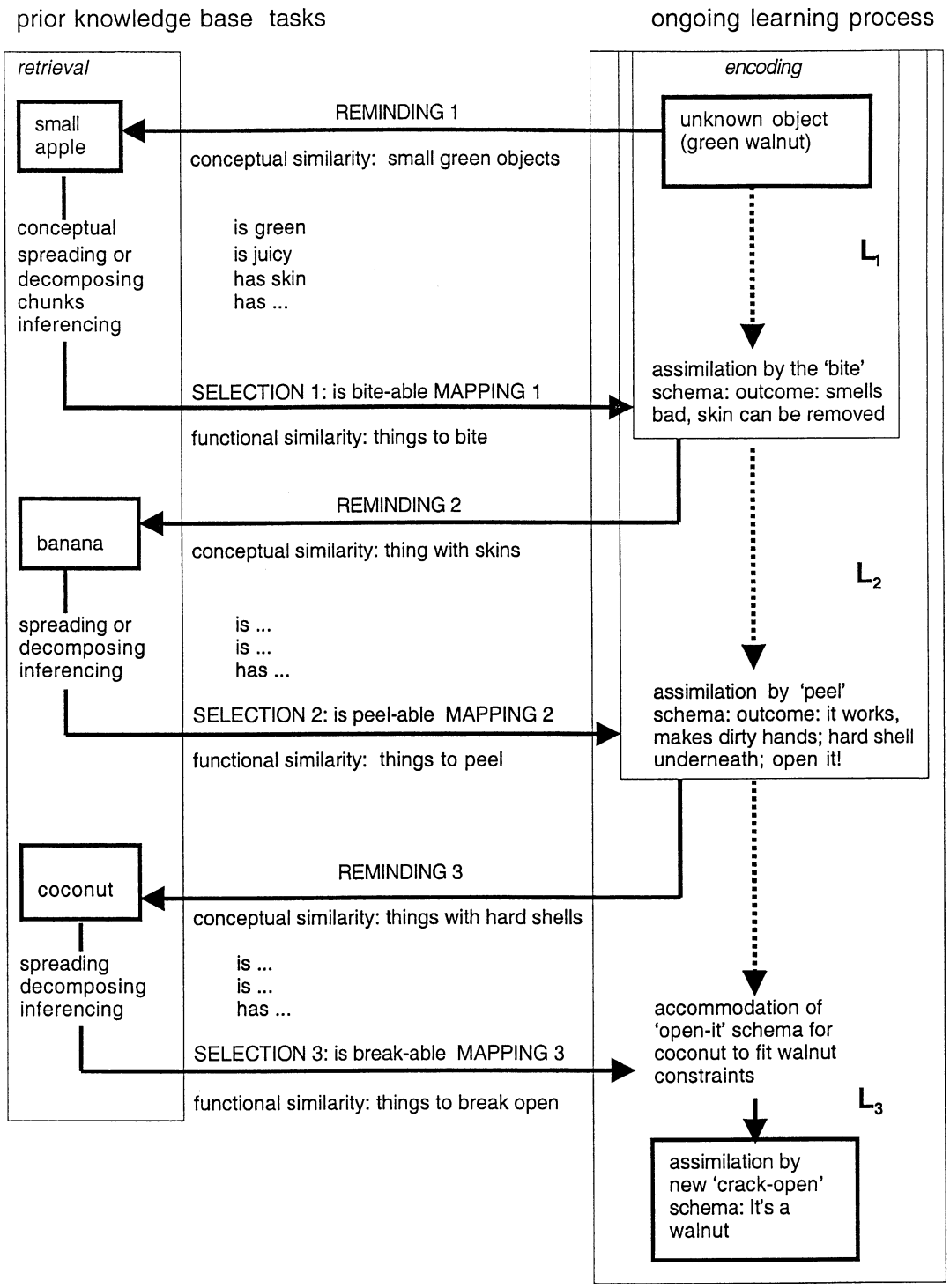

Figure 1
The oscillating process of transfer—an analysis of the subprocesses involved

belong to the set of transfer subprocesses. However, several authors plead explicitly for a fourth stage in analogical transfer focusing on schema induction and structural abstraction (Gentner 1989, Holyoak 1985, Gruber et al. 1995).

4.2 Theoretical Conclusions from the Green Walnut Example

The following conclusions can be drawn.

(a) The interesting part of transfer research is to be found in the subprocesses and not so much in the outcome or the problem solution.

(b) Transfer (and mapping as part of it) has a definite serial time structure. It should be kept in mind that the need for knowledge stemming from a target task depends on the actual stage of the ongoing learning or problem solving process. This fact dictates the time sequence of mapping activities.

(c) Mapping is in itself not a unitary subprocess. Its components are selection for mapping, tentative matching (i.e., looking for fit), evaluation of fit (including feedback), and either return to a next selection or else finishing the mapping process and giving the information over to the ongoing learning process (knowledge integration).

(d) Mapping is an ancillary process in favor of the ongoing learning or problem-solving process. All subprocesses *after* a fit are part of knowledge integration, which is part of the main learning process, in our example the accommodation of the inadequate assimilatory schema.

(e) The transfer subprocesses are oscillating processes. They follow the constraints of the ongoing learning or problem solving process, including possible feedbacks from nonfit situations of some mappings (not exemplified in Fig. 1).

4.3 Analogical Transfer Research

There is a huge research literature on specific details while transferring source knowledge on to target knowledge: (a) the structure-mapping approach stresses the fact that 'analogy is a way of focusing on relational commonalties' (Gentner 1989, p. 201), (b) the pragmatic schemata approach emphasizes the construction of pragmatic schemata by comparing and abstracting the analogues of the two tasks (Holyoak 1985), and (c) the exemplar approach holds that concrete exemplars instead of abstract structures play the dominant role in analogical transfer. All three approaches agree on the above-mentioned stages of transfer but differ in attributing relative importance to content-specific surface analogues on the one hand and abstract schemata on the other across the stages of transfer. No single theory can explain all available data concerning the effects of source analogues on retrieval and application; there is a plea for a hybrid of the structural, pragmatic, and exemplar views for that purpose.

5. Enhancing Transfer Effects: Knowledge Constructions Need Consolidation

Lack of transfer is the rule rather than the exception—at least in every-day learning. It seems that the primary reason for this is the fact that basic learning processes have not been led up to a level of sufficient understanding, the students often remaining on a level of surface processing or mere symbol manipulation (especially in mathematics and physics). Conceptual and procedural constructions (including problem solutions) in the main learning process need consolidation, as follows.

5.1 Strengthening Knowledge Structures and Making them Transparent

Consolidation for strengthening cognitive structures involves modified reconstructions of previous constructions: Basically the same conceptual elements are connected again with each other

(a) in a different sequence (sometimes a backward reconstruction),

(b) under different conditions (e.g., time),

(c) from different perspectives, or

(d) with a different goal (e.g., for a different audience),

providing transparency and coherence to the structure. Much of this kind of consolidation can be implemented by means of self-explanations (Chi et al. 1994, Renkl et al. 1998).

5.2 Flexibilizing Knowledge Structures

Several approaches focus on flexibilizing knowledge structures, one stemming from Spiro et al.'s cognitive flexibility theory (Spiro et al. 1991). Another is the progressive transformation procedure (Steiner and Stoecklin 1997), aiming at structural flexibility by systematically transforming the original construction (mainly mathematical problem solving) forcing the learner to define similarities as well as differences by intensive comparisons. To secure flexibility of structures, it is important that knowledge is abstracted by the learners themselves (e.g., from examples) during the learning process instead of being taught as abstract units by some teacher or tutor.

One specific point refers to retrieval flexibility. In order to optimize stage 1 transfer subprocesses (encoding elements of the target task or problem), it is necessary that chunks formed during the previous construction process are decomposed in due time to keep the elements accessible for further processing (relating). Such decomposing activities can be seen as a specific form of retrieval practice that, at times, come close to self-explanations as mentioned above. In order to speed up retrieval of subroutines for learning and problem solving (e.g., conceptual knowledge in addition to algorithms for mathematical transforma-

tions), 'proceduralized' knowledge (after knowledge compilation; Anderson 1983) may yield an important contribution to transfer. However, the automatic use of proceduralized knowledge may include task-inappropriate effects if disconnected from the original declarative knowledge from which it is built up.

5.3 Transforming the Representational Format—Multiple Knowledge Representations

A specific kind of flexibilizing knowledge structures is transforming one representational format into another: a symbolic representation (e.g., an algebraic equation) into an iconic (pictorial) or an enactive (i.e., action-related) format and vice versa (Bruner 1966). The ability to transform representational formats and to keep knowledge structures accessible in several different formats renders parts of knowledge highly flexible and apt to play a role in several subprocesses in transfer.

5.4 Decontextualizing the Structures

Transfer often fails because knowledge is strongly caught in the training context, so that the common elements of source and target tasks cannot be found. Therefore, an important goal for learning for transfer is decontextualization. This means changing the content of the chosen examples, and thereby replacing contextual facts such as the agents of a problem situation (soldiers attacking a fortress from different directions instead of X-rays attacking a tumor in the same way), the concrete material (sand instead of water) or cultural circumstances (rural instead of urban) in order to have the learners recognize the invariant structural aspects or the abstracted schemata, respectively (Renkl et al. 1998).

5.5 Reviewing—Focusing on One's Own Processes, Not on Outcomes

There is one severe lack in today's teaching of learning for transfer, namely reviewing one's own learning processes: 'How did I acquire this wonderful problem-solving procedure, this marvelous proof in geometry, or this striking insight into a network of economic relations?' Reviewing one's own learning processes means mentally unpacking some of the chunks formed, recalling the relations established during listening to the teacher's explanations or the student-centered discussion; in short, reactivating the *how* of that multistep process of relating and condensing (chunking).

To sum up: the emphasis on consolidation is not an emphasis on automating (and thus not on knowledge compilation; Anderson 1983), but rather on making structures transparent, enhancing the coherence of relations and chunks, but also flexibilizing the elements, relations and larger units of a total knowledge representation including decomposing chunks. Consolidation contributes to the comprehension or understanding of the built-up structures (Kintsch 1998, Steiner 1999).

The crucial point for transfer is not whether or not chunks, productions, schemata, scripts, or mental models are constructed, nor is it their very nature that counts. It is *how* all these knowledge representations were dealt with in the original learning process to fight from the outset against structural deficits. There is a long-term aspect involved in transfer with accumulating learning effects that does not seem to be reflected sufficiently in current research.

6. Social, Emotional, and Motivational Aspects of Transfer

In the 'situated' learning approach (Gruber et al. 1995), the social context plays a significant role; it is encoded and stored as part of episodic memory. Indeed, actual features of target task elements or requirements may remind someone of episodic information that, in turn, may activate some conceptual source task information either to trigger a transfer process or to keep it going. Furthermore, learning, whether individual or social, is permanently accompanied by individual evaluations of both the potentially emotional aspects of the content and some individual aspects of learning such as goal attainment, success, or failure. It is therefore not surprising that emotional effects are encoded along with cognitive contents and activities that may be retrieved later in transfer situations when the target task itself evokes emotional or mood memories (Bower 1981). Moreover, there are motivational and personality-related aspects that contribute to the overall quality of the learning and transfer process and outcome: volitional self-regulations (Kuhl 1985) along with high personal involvement and mental effort.

6.1 External Support or the 'Proximal Zone of Development'

Experimental work and educational field studies suggest that transfer is usually not performed without external support, be it by hints with regard to similarities or analogues in two tasks, tutoring, or other sociocognitive guidance. All these measures correspond to the 'proximal zone of development', a concept introduced by Vygotsky (1978): the psychological space where individuals are able to develop cognitive achievements not by themselves but through the support of a tutor.

6.2 About Future Transfer Research

A lot of admirable experimental work has been done within transfer research but there is a strong bias toward artificial microworlds in the tasks to be learned

that diminishes the usefulness of the results with regard to broader, less experimental settings. As far as the 'analogues' are concerned, the nonexperimental reality often looks different: at the beginning of many problem-solving processes there are no analogues whatsoever in the learner's mind. It would be promising for future transfer research to focus on the subprocesses at the outset of transfer when individuals try to find and define the mental gaps to be filled in their learning or problem-solving tasks or when the retrieval subprocesses are set in motion. There is, furthermore, a definite need for more research focusing on both metacognitive and noncognitive aspects for explaining the lack of transfer so often observed.

See also: Cognitive Theory: ACT; Connectionist Models of Concept Learning; Expertise, Acquisition of; Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Learning Curve, The; Learning to Learn; Production Systems in Cognitive Psychology

Bibliography

Anderson J R 1983 *The Architecture of Cognition*. Harvard University Press, Cambridge, MA

Bower G H 1981 Mood and memory. *American Psychology* **36**: 129–48

Bruner J S 1966 *Toward a Theory of Instruction*. Belknap Press of Harvard University Press, Cambridge, MA

Chi M T H, DeLeeuw N, Chiu M H, La Vancher C 1994 Eliciting self-explanations improves understanding. *Cognitive Science* **18**: 439–77

Cormier S M, Hagman J D (eds.) 1987 *Transfer of Learning. Contemporary Research and Applications*. Academic Press, San Diego, CA

Detterman D K, Sternberg R J (eds.) 1993 *Transfer on Trial: Intelligence, Cognition, and Instruction*. Ablex, Norwood, NJ

Ehrenfels von C M 1890 Über Gestaltqualitäten. (On the qualities of 'Gestalt'). *Vierteljahreszeitschrift fűr Wissenschaftliche Philosophie*. **14**

Gentner D 1989 The mechanisms of analogical learning. In: Vosniadou S, Ortony A (eds.) *Similarity and Analogical Reasoning*. Cambridge University Press, New York, pp. 199–241

Gruber H, Law L C, Mandl H, Renkl A 1995 Situated learning and transfer. In: Reimann P, Spada H (eds.) *Learning in Humans and Machines: Towards an Interdisciplinary Learning Science*. Pergamon, Oxford, UK, pp. 168–88

Holyoak K J 1985 The pragmatics of analogical transfer. In: Bower G H (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory*. Academic Press, New York, Vol. 19, pp. 59–87

Kintsch W 1998 *Comprehension. A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK

Kuhl J 1985 Volitional mediators of cognition–behavior consistency: self-regulatory processes and action versus state orientation. In: Kuhl J, Beckmann J (eds.) *Action Control: From Cognition to Behavior*. Springer-Verlag, Berlin, pp. 101–28

Piaget J 1952 *The Origins of Intelligence in Children*. International Universities Press, New York

Renkl A, Stark R, Gruber H, Mandl H 1998 Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemporary Educational Psychology* **23**: 90–108

Singley M K, Anderson J A 1989 *The Transfer of Cognitive Skill*. Harvard University Press, Cambridge, MA

Spiro R J, Feltovich P J, Jacobson M J, Coulson R L 1991 Cognitive flexibility, constructivism, and hypertext: random access instruction for advanced knowledge acquisition in ill-structured domains. *Education and Technology* **31**(5): 24–33

Steiner G 1999 *Learning. 19 Scenarios from Everyday*. Cambridge University Press, Cambridge, MA

Steiner G F, Stoecklin M 1997 Fraction calculation—a didactic approach to constructing mathematical networks. *Learning and Instruction* **7**(3): 211–33

Thorndike E L, Woodworth R S 1901 The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review* **8**: 247–61

Thorndike E L 1913 *Educational Psychology, Vol. 2: The Psychology of Learning*. Teachers College, New York

Vygotsky L 1978 *Mind and Society*. Harvard University Press, Cambridge, MA

G. Steiner

Transference in Psychoanalysis

1. Definition

Transference, and its counterpart, countertransference, have been recognized as among the most fundamental organizing concepts of the psychoanalysis created by Freud. Basically, transference is the conception that our past, our lifetime of experience, lives on in us and plays a formative role in our interpersonal behaviors in the present. Transference was discovered by Freud quite inadvertently. In accord with his original conception that neurotic symptoms and behaviors were representations of earlier mental conflicts that patients had been impelled to repress from conscious memory, Freud devised the method of 'free association,' requiring that the patient try to verbalize every thought as it arose, without regard to usual considerations of logical and tactful expression, i.e., no matter how seemingly nonsensical, offensive, embarrassing, or shameful. To his dismay, he found that, no matter what the patient's conscious intent, this process of sequential narration would on occasion be blocked, with the patient lapsing into embarrassed silence or claiming that simply nothing came to mind. When pressed, the patient could often acknowledge that unbidden and unwanted thoughts or feelings about the analyst had come to mind, displacing the historical recounting.

Freud at first saw this as an unwelcome intrusion upon, and an impediment to, the joint therapeutic task of uncovering the originating traumatogenic events, and tried therefore to disengage the patient from such distracting impingements. He soon, however, realized

the regularity of this occurrence, that it was not an accidental or inadvertent happenstance, but rather a regular 'transferring' onto the person of the analyst of reactions to the earlier major figures in the patient's life, primarily parental and other family figures, but even more widely, any significant figures from the past who had played formative roles in the past experiences that were under present therapeutic scrutiny. In Freud's words, 'They (transferences) are new editions or facsimiles of the impulses and phantasies which are aroused and made conscious during the progress of the analysis; but they have this peculiarity … that they replace some earlier person by the person of the physician. To put it another way: a whole series of psychological experiences are revived, not as belonging to the past, but as applying to the person of the physician at the present moment' (Freud 1905). To state it simply: a feeling of anger welling up towards the analyst, perceived at the moment as an overbearing and arbitrary authority, would be a re-presentation of a sequence of childhood interactions with the patient's father, experienced often in just that light. Within this line of thinking, Freud designated transference reactions as father or mother transferences (or sibling, or mentor, or whomever), and as positive or negative transferences, depending on their emotional valences, whether affectionate, hateful, or some intermediate admixture.

Conceptualizing this way, Freud reversed his reaction to transference manifestations, from seeing them as impediments to analysis, to rather viewing them as essential vehicles for the analytic work. By 'analyzing' the transferences, i.e., unraveling their meanings in the interactions within the analytic session, and tracing them back, as representations in the present of pathogenic interpersonal interactions from the past, the patient's repressed traumatic past experiences would inevitably be brought to light, and their continuing formative role in the maintenance of neurotic behaviors in the present would be progressively dissipated. To revert to the example already given, the patient's chronic damaging interactions with his work superior, perceived as overbearing and arbitrary, that impeded his job satisfaction and could threaten job promotion or perhaps even the job itself, would be progressively eased, and even eliminated entirely, the more the patient could come to see them as unwitting re-enactments of his earlier experiences with his father (and to see concomitantly his own provocative role in eliciting those repetitive interactions).

Freud shortly came to feel in fact that just because of the emotional intensity experienced in the transference interaction in the immediate therapeutic situation, as compared with the remoteness of describing historical interactions in the long ago past, that the patient was most susceptible in the transference to the jarring emotional impact that could disrupt established, automatized behavior patterns, and lead to the shifts that we call insight and therapeutic change. This effect is described in one of Freud's most famous sentences as, 'For when all is said and done, it is impossible to destroy anyone *in absentia* or *in effigie*' (Freud 1912). In a well-known paper by Strachey, one of Freud's early adherents, interpretation of the transference was described as the only truly 'mutative' interpretation, all other interpretive efforts only leading up to, and supportive of, interpretation in and of the transference (Strachey 1934).

Freud's technical advice from the start was to allow the analytic transferences to unfold gradually, as they inevitably would, and only to interpret them as they clearly crystallized in the analytic situation as resistances to the ongoing work, either as negative transferences of hateful feelings directed at the analyst, or as erotized positive transferences of loving feelings. Freud did make an exception for what he called 'unobjectionable positive transferences,' the benign feelings that led the patient to the best possible conscious cooperation in the analytic work with the analyst, experienced by the patient as expert, helpful, and interested in effecting therapeutic change and cure.

In more modern times, some analysts have split off this so-called 'unobjectionable positive transference' into a separable conception alongside the transference, called the 'therapeutic alliance' (Zetzel 1956) or the 'working alliance' (Greenson 1965), usually operating automatically as a protective framework within which the transference vicissitudes—all of the complicated ambivalent relationships that characterize the human experience as it is lived out over a lifetime with all those who have played significant roles in each of our lives—could safely unfold, but at times, or regularly with some sicker patients, itself needing specific attention and nurture. Other analytic authors have, however, been dubious about this unobjectionable transference, seeing it as a defensive cover, hiding other much more objectionable transference manifestations, that would ultimately themselves need to be interpretively undone in the interest of the completeness of the psychoanalytic work.

As the transference unfolded over the course of the analysis, Freud further conceived of the evolution of what he called the 'regressive transference neurosis,' in which the presenting neurotic symptoms and behaviors dropped away in the patients' outside life, as their psychic life, with all of its neurotic components, came to full expression in the transference interplay within the analytic situation. To Freud, such a fully evolved transference neurosis, focused completely on the person of the analyst, recapitulated what he came to call the 'infantile neurosis,' the set of childhood maladaptive interactions that established the template for the later evolved neurotic illness that was to bring the patient to psychoanalytic treatment. It was the interpretive resolution of this fully-fledged regressive transference neurosis that Freud, and for long, most of

his followers, regarded as the central task of the analytic treatment, which could only be declared properly successful when the transference neurosis was fully resolved, and the various transference distortions were transformed into an objectively realistic relationship of the patient with the analyst.

2. *Transference Psychosis*

A related, and subsidiary, concept to that of the transference neurosis developed a good deal later, in the 1950s (but with roots back to Freud's close collaborator, Ferenczi, in the teens of the twentieth century), that of the transference psychosis. This conception developed along two distinct lines. The one was a precise analogue of the concept of the transference neurosis, i.e., that just as neurotic patients taken into psychoanalytic treatment display 'neurotic' transference reactions to the analyst, culminating in the full-fledged transference neurosis, so do psychotic patients taken into psychoanalytic treatment display 'psychotic' transference reactions which can coalesce into a fully formed transference psychosis. The other conception was that of a neurotic patient, seemingly well-suited for a classical psychoanalytic treatment, who under the regressive pressures of psychoanalysis could erupt (in an unforeseen way) into a psychotic transference response with loss of the usual reality moorings that maintain the transference as an 'as-if' phenomenon. I have reviewed this particular literature in detail elsewhere (Wallerstein 1967).

3. *Understanding of Transference at the End of the Twentieth Century*

Since Freud's day, our understanding of the transference has been altered and transformed significantly. A signal figure in this effort was Gill and Hoffman (1982) who, in a series of papers and a major monograph, squarely challenged the notion that transference manifestations consist just of the projections of the patient's past relationship patterns upon the neutral objective analyst who could be the epistemic arbiter of reality, clearly separating transference distortion from objective perception, and doing this from a position outside the transference display, as neither a participant in it nor a contributor to it. Given the inherent ambiguity in all human relationships, Gill saw the transference as the patient's plausible construction of his or her experience of the analyst and of the analytic interaction, with the analytic task then, the construction between analyst and patient of how they each differently perceive the same interaction, each perception being the product of the individual life experience of the perceiver, with the individually different predispositions then towards preferred ways of endowing meaning to the encounter.

The therapeutic focus would of course be on the psychology of the patient, what in the patient's life experience disposed him or her to construe the therapeutic situation and interaction in the manner manifested, and in that way to link the lived past to the present transference manifestations. Thus, the past is still present in the transference as a formative underlying template of interpersonal expectations, but the focus has shifted to a present-day interaction in the here-and-now of the therapeutic encounter as immediately experienced by the two participants, with again, the stated focus on the light that could be thrown thereby on the patient's prior life experience and the enlarged meanings that could now be ascribed to the life development. The focus has thus been shifted from the discovery of the past emergent in the present (in the transference), a focus on the there-and-then, to the new focus on the interaction in the present, in the here-and-now, as framed and colored by the past.

Along with that has been the concomitant shift from the transference as a purely intrapsychic phenomenon within the patient, projected onto the reflecting and assessing screen of the neutral, objective, and uninvolved analyst (what is now called the embodiment of a one-person psychology), to the transference as an interaction of two subjectivities, each bringing their own built-in predispositions to that interaction, with the focus again of course on the patient's characteristic construal of the situation, embedded in the life experience and consequent neurotic structuring of the patient, whose neurotic difficulties in living are, after all, the *raison d'être* of the therapeutic undertaking. This is called now a two-person psychology, the intersubjective interaction of two participants, albeit focused primarily on the difficulties of the one, the patient. These altered conceptions of the transference have been developed further as a fundament of what is called the object-relational perspective in psychoanalysis as distinct from Freud's perspective, called now drive/structural, or put differently, two-person vs. one-person psychologies. (For a full development of these distinctions, see the volume by Greenberg and Mitchell 1983.)

4. *Analytical Techniques*

These modern alterations in the conception of the transference have led to many consequent alterations in the technique of analysis. The focus is now more on the early interpretation of the transference, and on seeking out implicit allusions to the transference, and rendering them explicit, again, as early as feasible. The conception of the unfolding of the full-fledged transference neurosis as a necessary recapitulation of the infantile neurotic experience has given way to the experience of a great variety of transference mani-

festations in the present, more or less intense, more or less covert or manifest, of great range in emotional valence, and each embedded in uniquely characteristic relationship patterns, with past and present relationship experiences in constant dialectical interplay.

This entire history of transference development has been well summarized in an article by Cooper (1987) in which he contrasted what he called the 'historical' and the 'modernist' models of the transference. In Cooper's words, 'The first idea, close to Freud, is that the transference is an enactment of an earlier relationship, and the task of transference interpretation is to gain insight into the ways that the early infantile relationships are distorting and disturbing the relationship to the analyst, a relationship which is, in turn, a model for the patient's life relationships. I shall refer to this as the *historical* model of transference, implying both that it is older and that it is based on an idea of the centrality of history. The second view regards the transference as a new experience rather than as an enactment of an old one. The purpose of transference interpretation is to bring to consciousness all aspects of this new experience including its colorings from the past. I shall refer to this as the *modernist* model of the transference, implying both that it is newer, in fact still at an early stage of evolution, and that it is based on an idea of the immediacy of experience.' Implicit of course, though not directly stated, is that in the historical concept, transference is an intrapsychic phenomenon in one person, the patient, and in the modernist concept, it is an interpersonal phenomenon, involving two people.

5. *Countertransference*

One cannot talk about transference of course without considering countertransference. Originally, countertransference was conceived by Freud as the analyst's inappropriate and nonobjective responses to the patient's transferences, stemming from unresolved issues in the psychology of the analyst that were stirred up by the patient's transferences. This would necessarily cloud the analyst's proper evaluation of the transference, and was a conception therefore quite parallel to Freud's original viewing of the transference as an impediment to the analytic work. In this conception the recommended remedy was constant watchfulness by the analyst, to monitor such inadvertent and inappropriate responses, and to control as fully as possible their expression, in order to keep them from contaminating the transference display. If such monitoring and control to keep the countertransference from impinging on the transference proved ineffective, the further remedy would be rigorous self-analysis or a return by the analyst for more personal analysis of his or her own—in order to overcome these interfering neurotic residues. This view later came to be called the 'narrow' conception of the countertransference. Later, others, especially analysts working with sicker-than-neurotic patients, came to a wider conception of the countertransference, as the totality of the analyst's emotional responses to the patient, including countertransference proper (the specific neurotic response to the patient's transference), as well as emotional reactions by the analyst to the patient as a whole and to the activity of analyzing itself. This wider conception was called 'totalistic'; some analysts (Winnicott 1949) divided this overall countertransference into 'neurotic' and 'objective' components.

Though Freud came to see the transference, which he first felt to be an impediment to the analytic work, as, rather, the essential vehicle of the analytic therapy, with success hinging upon the successful recognition and interpretation of the transference, he failed to take the same step in his consideration of the countertransference. It was only about a dozen years after his death, in the early 1950s, that psychoanalysts, first in the UK, began to view the countertransference not as an occasional inadvertent and inappropriate intrusion upon the psychoanalytic process that needed to be guarded against, and only dealt with when it nonetheless manifested itself, but rather to the countertransference being an inevitable and constant expression of the analyst's involvement with the patient in the two-person, mutually interactive, treatment process. This was a reflection of the, by that time, almost universal shift to the totalistic conception of the countertransference as the totality of the analyst's emotional responses to the patient, whether in specific reaction to the patient's transferences, or whether stemming from the analysts' own developmental history and their overall (emotional) responses to the patient and to the situation between them.

It was further developed that through careful monitoring of their countertransference thoughts and feelings, the analyst could obtain useful insights into what the patient was trying to get the analyst to think or feel, and thus to obtain a wider knowledge of the patient's psychology and developmental history. This process of the arousal by the patient of emotional responses within the analyst has been variously called 'projective identification' (by the Kleinian school of psychoanalysis) or 'role responsiveness' (by the classical Freudian school; cf. Sandler 1976). Conceived and utilized in this way, the perception of the countertransference has likewise shifted from the initial view of it as an occasional and inappropriate intrusion upon the work of analysis (to be mastered and overcome by the analyst), to an inevitable and vitally necessary vehicle to the enlarged understanding of the patient, in this way, and many years later, following the same path as our understanding of the transference, seen first by Freud as an unhappy impediment to the analytic work, and seen only later as an essential conceptual and technical vehicle for that work.

In this sense, transference and countertransference are completely equivalent, neither being just a re-

sponse to the other, but each stemming from the psychological development of the respective parties to the psychoanalytic process. Some analysts (cf. McLaughlin 1981) have therefore even called for the elimination of the word countertransference, calling both sides of the responsive interaction transference, and privileging neither as just a response to the other.

Countertransference evocations tend to be more subtle and more contained with psychologically well-functioning analysts, and with classically typical neurotic patients, and more severe and difficult to contain with those sicker patients, called narcissistic or borderline in their character organization, seen nowadays with increasing frequency in psychoanalytic and psychotherapeutic practice. In this mutually interactive psychoanalytic situation, the intense transference-countertransference interplay results often in what are called 'role enactments' in which the analyst feels impelled to take on the role (punitive authority, benign rescuer, neglectful parent, or whatever) that the patient, through speech and behavior, is working, more or less unconsciously (usually more) to elicit from the analyst. And it is through the dawning awareness of having inadvertently taken on, or fallen into, the assigned role in the psychological interplay between them, that the analyst often becomes first aware of his or her countertransference involvement, and can then use this to trace out the psychic pressures from the patient that, when identified, become insights into the patient's character predispositions, with their linkages back into the patient's developmental history.

A major issue in this regard that is at present a matter of much professional controversy, is that of the circumstances in which it may be useful for the analyst to communicate to the patient aspects of understanding of the countertransference, when arrived at by the analyst. There are some who, while fully utilizing countertransference as additional avenues of access to the understanding of the patient, would avoid completely any countertransference 'confession' to the patient, as only an unnecessary, and varyingly troubling, burden for the patient to bear. Others advocate, at least at some times, or under certain circumstances, sharing their countertransference responses with the patient, as useful avenues towards the patient's understanding of what kind of reaction they were working to evoke in the analyst (and of course therefore with others in their daily life outside of the analysis), and the kinds of responses (so often consciously not desired by the patient) that are elicited in the other. And of course there are other instances where the patient detects the countertransference response before the analyst is aware of it ('You're angry at me and taking it out on me'), and at such times the analyst does need to deal with this truthfully, albeit tactfully, rather than to try to deny or avoid it.

A last and very important point about transference and countertransference: The psychoanalytic situation is a highly contrived arena in which these phenomena—emotional reactions to each other when individuals are engaged in an intimate and intensely personally meaningful dialogue—are highlighted and can be subjected to mutual scrutiny for the enhancement of the psychological understanding of the neurotic issues (and their developmental history) in the patient seeking relief from mental and emotional difficulties in life. But it is implicit in everything said to this point, and should be made explicit now, that we all live, all our lives, by transferences, in all of our human relationships. We always bring ourselves, our personalities or characters, as shaped and colored by our pasts, into all of our interpersonal encounters, and see our relationships and expectations in the present in very significant part through the dispositions built into us by the whole lifetime of our past experiences. In fact, it can be said that our entire capacity for relationships, for empathy with others, even for creative endeavor in any arena, rests on the transferences that we bring and live by (cf. Loewald 1980).

6. *Summary*

All of the issues covered in this article are explored at great length in Section III (more than 100 pages) in a book by Wallerstein (1995), *The Talking Cures: The Psychoanalyses and the Psychotherapies*, a volume that traces the historical development and unfolding of psychoanalysis and the derived psychoanalytic psychotherapies.

See also: Freud, Sigmund (1856–1939); Interpersonal Psychotherapy; Psychoanalysis: Current Status; Psychoanalysis, History of; Psychoanalysis: Overview; Psychotherapy: Ethical Issues; Therapist–Patient Relationship

Bibliography

Cooper A M 1987 Changes in psychoanalytic ideas: Transference interpretation. *Journal of the American Psychoanalysis Association* **35**: 77–98

Freud S 1905/1953 Fragment of an analysis of a case of hysteria. *Standard Edition* **7**: 1–122

Freud S 1912/1958 The dynamics of transference. *Standard Edition* **12**: 97–108

Gill M M, Hoffman I Z 1982 *Analysis of Transference, Volume I. Theory and Technique*. Psychological Issues Monograph 53. International University Press, New York

Greenberg J R, Mitchell S A 1983 *Object Relations in Psychoanalytic Theory*. Harvard University Press, Cambridge, MA and London

Greenson R R 1965 The working alliance and the transference neurosis. *Psychoanalysis Quarterly* **34**: 155–81

Loewald H W 1980 *Papers on Psychoanalysis*. Yale University Press, New Haven, CT and London

McLaughlin J T 1981 Transference, psychic reality, and countertransference. *Psychoanalysis Quarterly* **50**: 639–64

Sandler J 1976 Countertransference and role-responsiveness. *International Review of Psycho-Analysis* **3**: 43–7
Strachey J 1934 The nature of the therapeutic action of psychoanalysis. *International Journal of Psycho-Analysis* **15**: 127–59
Wallerstein R S 1967 Reconstruction and mastery in the transference psychosis. *Journal of American Psychoanalysis Association* **15**: 551–83
Wallerstein R S 1995 *The Talking Cures: The Psychoanalyses and the Psychotherapies*. Yale University Press, New Haven, CT and London
Winnicott D W 1949 Hate in the counter-transference. *International Journal of Psycho-Analysis* **30**: 69–74
Zetzel E R 1956 Current concepts of transference. *International Journal of Psycho-Analysis* **37**: 369–75

R. S. Wallerstein

Transition, Economics of

The economics of transition is a relatively new field of scientific inquiry dealing with economic phenomena that arise as countries move from a communist system of central planning (also known as a command economy) to a decentralized market system. Transition economics uses as its laboratory the one-third of the world's population that left the market system for central planning in the first half of the twentieth century, and then launched an attempt to return to a market system toward the century's end. Since transition is one of the most important political-economic phenomena of the late twentieth and early twenty-first centuries, transition economics as a field has grown rapidly since 1990 and contains important theoretical, empirical, and policy components.

From an academic standpoint, the field of transition economics is unique because it permits researchers to examine how economies behave when markets are just being created and established phenomena of advanced economies, such as unemployment, are just starting to appear. Analysts of the transition economies are hence able to provide valuable information about the functioning of market economies, as well as to analyze the transition process itself. From a policy standpoint, understanding the transition process is important for the formulation of public policy in and towards these economies. Finally, the transition economies constitute an important part of the world market and an understanding of their problems and performance is essential for the formulation of an effective business strategy in these economies.

The performance of the transition economies to date has been uneven, providing variation that analysts need in order to draw significant conclusions and showing the sensitivity of outcomes to variations in initial conditions, government policies, and exogenous shocks. In particular, the fall of the Berlin Wall in 1989 has led to a rapid demise of the Soviet system of central planning and the launch of a relatively rapid transformation of the Soviet bloc economies into ones that would be based on free prices and newly created markets. Yet, while all the former Soviet bloc countries suffered an unprecedented decline in output at the start of the transition, their subsequent economic performance diverged considerably. The Central European and Baltic countries have performed relatively well, while countries further east, including Russia and the Ukraine, suffered further decline in output and achieved only limited structural transformation. In contrast, China's gradual reforms since 1978 have led to a remarkable transformation of the Chinese (more decentralized) version of the centrally planned system and generated a long-term period of rapid economic growth. These varied outcomes have naturally given rise to a debate as to which types of initial conditions and reforms are more conducive to a successful transition.

The principal theoretical models in transition economics focus on (a) the strategy of transition in light of various political constraints and uncertainty in the determination of optimal sequencing and speed of reforms, (b) the political economy of mass privatization (focusing on issues of possible privatization reversals, giveaway of shares of state-owned firms to the population at large versus managers and workers in these firms), (c) sectoral shifts, labor reallocation, labor market frictions, and capital accumulation, (d) speed and methods of price liberalization (including the Chinese dual-track approach), (e) the large output fall at the start of the transition (including explanation such as disruptions of supplies, inefficiencies of bargaining and search frictions), (f) the effect of weakness or collapse of government on economic performance (covering issues related to the switch from a strong communist government to a weak government plagued by corruption and the inability to collect taxes, enforce laws and prevent the rise of Mafia-type groups and an underground economy), (g) corporate governance, as well as governance at the level of a country, and the related issues of incentives, development of the banking sector, and large-scale bailouts of banks and firms.

The basic models of the economics of transition are similar in nature to the dualistic models of economic development that were developed in the 1950s and 1960s. The latter models describe development as a process of reallocation of labor from the low marginal productivity (traditional) sector to the high productivity (modern) sector. The modern sector generates surplus that is reinvested and leads to the expansion of employment in the modern sector. The traditional sector shrinks over time as its surplus labor is absorbed into the modern sector and the remaining workers adopt productivity-enhancing technology. The development process is finished when the marginal product of the traditional sector rises to the level of the modern sector and the dual economy becomes a single-

sector economy. In the transition models, transition consists of a reallocation of labor from the 'inefficient' state sector, where there is redundant labor, to the 'efficient' private sector that is assumed to pay labor its marginal product. Transition is a process in which these dualistic economies move toward a single-sector economy characterized by the equality of labor's marginal product across uses.

A prototypical model of transition is that of Aghion and Blanchard (1994). The authors assume that the marginal product of labor is higher in the private than in the state sector and that the government maximizes the net present value of output, which is equivalent to maximizing the rate of growth of private sector employment. The government engineers the downsizing of the state sector through the reduction of subsidies to state-owned firms and the creation of unemployment benefits for the released workers. Market forces determine employment in the private sector. The level of wages is a negative function of the unemployment rate and the creation of jobs in the private sector hence positively depends on the speed of downsizing of the state sector. Slow downsizing leads to low unemployment, which in turn puts upward pressure on wages and slows down the growth of the private sector. If the government downsizes the old sector too quickly, it will create high unemployment, which will reduce wage increases. However, the model assumes that private firms pay wage taxes that are used to finance unemployment benefits, and an excessive rate of closure of state firms, which leads to excessive increases in unemployment, and reduces the expansion of the private sector because of an excessive wage tax levied to finance unemployment benefits.

As was the case with the early development models, the question arises as to whether the first generation transition models provide a sufficiently useful approximation of the transition process. A number of features of the transition indicate that they may not be, and one may hence expect that further theoretical work, based on empirical evidence, will yield the next generation of theoretical models. The following two empirical patterns have been documented in empirical studies and are likely to be incorporated into future theoretical models.

First, while the private sector has emerged on a large scale in most transition economies, empirical evidence does not support the assumption that the private firms are more efficient than the state-owned ones. Ownership is also found not to be systematically related to employment and wage setting. Moreover, there are many different combinations of private ownership and corporate governance and they generate different levels of efficiency. In addition, while most theoretical models assume that private firms are newly created, a large number of private firms are privatized state-owned firms or spun-off subsidiaries. The phenomenon of breakups and spin-offs of firms appears to have increased the efficiency of both the spun-off and the remaining 'master' firms. Second, the flows of workers are more complicated than those captured in the first generation theoretical models. There are important job-to-job flows that do not entail unemployment and many workers move from one state firm to another, rather than from state to private firms only. Moreover, private as well as state firms are important in both job creation and job destruction.

The transition process is hence more complex than a shift of labor and other resources from inefficient state-owned firms to more efficient private firms. Existing evidence suggests that various types of firms coexist, compete for resources and markets, and through various forms of restructuring they improve their efficiency over time.

Since the literature on transition is relatively young and yet covers many traditional fields of economic inquiry, there are no large-scale, comprehensive surveys. The best overview of the theoretical literature may be found in Roland (2000), while a recent survey of the empirical literature related to the labor markets is in Svejnar (1999).

See also: Area and International Studies: Development in Eastern Europe; Area and International Studies: Development in Southeast Asia; Democratic Transitions; Eastern European Studies: Economics; Eastern European Studies: Politics; Revolutions of 1989–90 in Eastern Central Europe; Soviet Studies: Economics; Soviet Studies: Politics

Bibliography

Aghion P, Blanchard O 1994On the speed of transition in Central Europe. *National Bureau for Economic Research Macroeconomics Annual*, pp. 283–319

Roland G 2000 *Transition and Economics: Politics, Markets, and Firms*. MIT Press, Cambridge, MA

Svejnar J 1999 Labor markets in the transitional Central and East European economies. In: Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*. North Holland, Amsterdam, Vol. 3B, Chap. 42, pp. 2809–57

J. Svejnar

Transnational Citizenship: Cultural Concerns

Me velle civis totius mundi non civis oppidi. When Swiss Reformation leader Ulrich Zwingli offered Erasmus of Rotterdam citizenship in Zurich, the humanist philosopher is said to have replied that he did not want to become a citizen of this single city, but of the 'entire world.' World citizenship was always one of human-

ity's dreams, one considered as honorable as it is unfulfillable. Now globalization has not only razed the fortress of the nation-state (as progress in ballistics did to the fortified walls of medieval city-states), it has also created opportunity for what is called *transnational citizenship*. In historical perspective, nations provided the framework of modern community and society; they limited the space of communication where political parties and interest groups of all kinds acted, thus creating the prerequisites for equal representation and civic participation. In the welfare state, too, who should be included and who excluded from solidarity, was defined within this framework. But the universalistic principle of inclusion itself helped migrants gain a high-value status as residents, without being formal citizens. A person's national citizenship thus uncoupled from his or her right to have rights, insofar as international laws, for example to protect migrant labor, and human rights conventions became plausible as independent sources of individual and collective rights. The question at hand is: Between local patriotism and global markets, where can *homo politicus* be found?

1. The Term 'Transnational'

The *Oxford Dictionary* dates the emergence of the term to about 1920, documented with a quotation from an economic text that saw Europe after World War I characterized by its 'international or more correctly transnational economy'; interestingly, it added another source that identified the Christian Church as the sole power that could create the conditions for a 'transnational, non-racial democratic polity.' One author, who had 'transnational America' in view as early as 1916, is not even mentioned: Randolph S. Bourne. While the nations of Europe hit out at each other in enmity, this literary critic and pacifist from New York emphasized a special potential of the USA, one he thought represented the core and seed of future world society. The USA, as the 'first new nation' of immigrants, clearly did not have recourse to the European bases of nation-building and collective identity, either ethnic-cultural or state-bureaucratic. On the other hand, Bourne rejected the then-current idea of a 'melting pot': 'We are all foreign-born abroad or the descendants of foreign-born, and if distinctions are to be made between us they should rightly be on some other ground than indigenousness.' One should not seek the foundations of American collective identity in a mystified past, as was the case with the European nationalism, instead '… we must perpetrate the paradox that our American cultural tradition lies in the future.' From this, Bourne drew the conclusion: 'America is coming to be, not a nationality but a transnationality, a weaving back and forth with other lands, of many threads of all sizes and colors' (Bourne 1916, pp. 87, 92, 96).

At the beginning of the twenty-first century, 'transnational' has become a technical, but also a fashionable term, though one seeks it in vain in dictionaries and most social science lexicons. It is used especially in connection with international organizations and the 'multis' (transnational corporations, or TNCs). According to the original meaning of the Latin prefix 'trans,' it points beyond the usual diplomatic dealings among sovereign nation-states and the customary division of labor among 'national economies'; the unit of analysis, not only in international relations, has become *Weltgesellschaft* (world society, see Luhmann 2000). This does not intend to announce an 'end of nation-states'; but state sovereignty as we knew it has become 'anachronistic' (Czempiel 1969). From the accustomed macroperspective or bird's-eye view of the world economy, the glance drifts to the microlevel of *Lebenswelt* (life world), where we can trace the cross-boundary paths on which people and objects, metaphors and symbols, individual life histories and collective biographies are transferred. It is time for social science and political actors to acknowledge this 'paradigm shift' from international relations to transnational that also allows that a new form of world politics is arising: governance and citizenship beyond the nation-state.

2. Economic and Cultural Preconditions of Transnationality

2.1 Global Economy

From the beginning, the most powerful and conspicuous relativizers of the nation-state were the transnational corporations with their world-spanning trade and finance transactions. There has been a world market since sailing and techniques of navigation first permitted intercontinental trade, currency exchange, and investment; the Hanseatic League and the spread of Chinese family households throughout the Pacific demonstrate that the postmodern subversion of national borders has its premodern precursors. Yet, capital's inherent 'lack of patriotism' (Karl Marx) has meanwhile developed beyond the accustomed relations between parent companies and foreign subsidiaries; economic and financial conglomerates have merged so that, aside from having a legal seat, they are no longer anchored in one particular nation. E-commerce, including business-to-business deals, will further radicalize this despatialization.

Certainly, the prospects of the new economy must be viewed with caution, but state and cultural boundaries already play hardly any role anymore in its leading branches, telecommunications and biotechnology. Elimination of boundaries through 'digital capitalism' already have dramatic consequences for taxes and customs—which, after all, are the primary

sources of revenue for nation-states and even supranational economic communities like the European Union. Global standards of management and services are joined by the harmonization of trade and economic law through legal firms, auditing companies, and consultancies. These need not really have global reaches; what we observe is rather a regionalization (in the sense of 'continentalization') of economic domains (North American Free Trade Association, European Union) or a geographic subdivision, as with the executive board of ICANN (Internet Corporation for Assigned Names and Numbers) and other transnational private–public institutions.

2.2 Migrant Communities

These examples show that transnational spaces have arisen between nations. Moving in the interstices is a 'business class,' a loose conglomerate of managers, consultants, and service providers more closely associated with companies and concepts (e.g., the 'Apple culture' and, more recently, the 'Linux movement') than with native countries and native tongues. Border-crossing mobility is not limited to the upper levels of the labor market; even unskilled workers with precarious residence status have meanwhile become highly mobile and multicultural. This *Völkerwanderung* (mass migration), joined by masses of foreign travelers, commuters, and retirees in the sunny South, has especially stimulated thought about 'transnational social spaces' (Faist 2000). Today's nomads differ from their long-familiar precursors, especially the streams of immigration to the USA, Canada, Australia, and France in the nineteenth and early twentieth centuries. Today, 'homeland' and 'host country' are more closely connected and networks between them are much denser. With cheap transportation and easy communication, not to mention the worldwide reception of 'home' television and radio broadcasts via cable and satellite, permanent settlement no longer seems necessary. Bags were always packed and ready in the Diaspora, but usually they remained in the attic; and the back and forth of those times—for example between New York's Lower East Side and southern Italy or between Chicago and Warsaw—also lacked the emotional routine that now characterizes transnational mobility, and above all it did not have the effects on the structure of entire societies that transnational communities are likely to have today (Ong 1999, Leggewie 1999).

We can now define more precisely as transnational a social field, transcending national affiliation, in which a growing number of people lead a kind of double life. The expropriation of farmers, which tore the agrarian population from the 'soil' and drove them into the industrial cities, now finds its counterpart in a kind of 'expropriation of borders.' This casts doubt on sociology's 'container vision'; its spatial metaphors referring to closed national societies now seem exhausted. For transmigrants live in the long term in two or more places, constantly speak two or more languages, possess *en masse* two or more passports, and pass continually in both directions through makeshift households, networks of relationships, and spaces of communication. From this perspective, migration 'routes' count more for cultural studies than the migrants's 'roots' of personal identity in national collectivities (Clifford 1997, Hannertz 1996). All this together, perhaps combined with and crowned by the World Wide Web, a medium of transnationalization *par excellence*, leads to a despatialization that, unlike classic emigration, permits 'virtual proximity' and merely temporary absence. This allows communities to maintain themselves even without constant face-to-face encounters; and with successive long-term presence 'at home' and 'abroad,' both poles of such transitory existence finally become almost interchangeable.

2.3 Religious Pluralism

Wanderers between the worlds generally bring their native gods and rites in their bags, for people who are supposed to have lost their ties are especially likely to seek 'ties back' (Latin *re-ligere*), which is the literal translation of religion. Religious communities have always transcended national borders; spiritual belief is especially suitable for 'stretching' and reconstructing in what was originally called 'the Diaspora' since the scattering of Jewish and early Christian congregations. The supralocal spread of religious ideas and communities is not new in principle (the Catholic Church can be seen as the first-ever agent of globalization), but the transnationalization of religion is no longer identical to Christianization; nor has secularization, which became typical for the Christian societies of Europe, prevailed as a model for the world.

Though religious life (and dominant cultures based on it) are relativized by transnational migration, emigration can thus revive religious feelings of belonging. In regard to world society, religion pushes globalization forward precisely by resisting the profane effects of economic globalization inasmuch as it forms particularist congregations. Diaspora, once experienced as catastrophic, is no longer an exception in the religious pluralism of our time; and at the same time, protection of religious freedom has improved all over the world, so that imported religious symbols are becoming visible again in the public space of secularized societies. Globalization is thus not limited to corporate mergers, Internet communication, and financial transactions. Religious communities, too, travel around the globe, and not only in the organizationally and functionally supported manner of

world churches, but also and especially as decentralized grass-roots movements, as an unofficial, heterodox, self-founding, religious civil society.

2.4 Hybrid Cultures

What can be said about religious coexistence is also true for the diffusion and mixing of multicultural societies in general; a characteristic of this world of hyphens is provided by 'world music' (e.g., in the style of Peter Gabriel or Putumayo)—a collage of disparate materials often fusing completely contrary styles. In such mixtures, called 'hybrid' in allusion to biological crosses, things grow together that would not sound compatible to any purist's ear. But proponents of musical and other mixtures point to quite successful combinations since the earliest origins of the fine arts; and in general we can note that the creativity of cultures is never based on keeping the foreign at a distance, but to 'extraordinary' borrowings, i.e., to mimetic appropriation and the constant exchange of inventions, to innovation from the margins and the acquisition of the seemingly unassimilable. Seen in this way, phenomena of 'world culture' are merely a further stage of the hybridization of hybrid cultures, and all efforts toward canonical restoration are futile while the barriers between high and low culture have been razed, too (Appadurai 1996, Tomlinson 1999).

The 'wild' combining of artifacts, symbols, and identities began with the primary means of expression and communication, language itself, where it 'creolized' at the points of contact between two or more groups of speakers. The process of assimilation was usually asymmetrical and often violent, but, at the same time, hybrid constellations mirror the growing interdependence of world society, which of course excludes as often as it includes. In this process, components of all cultures are uncoupled from local roots and contexts; this has often been interpreted as 'standardization' and, since the origins of hegemonial popular culture can be traced to the USA, as Americanization. Much speaks for this assumption, if we observe the selectivity and modes of influence of the culture industry, whose effects sometimes amount to a kind of cultural extinction; on the other hand, the conformation of the world society's structural prerequisites does not lead to complete cultural uniformity.

In the terminology of cultural anthropology, culture must be conceived as bounded and embedded, and thus as a field that can be cultivated only in a particular time and place. The global economy and technology require and produce universal media like money and standardized systems of experts, which must ignore local origins, because this abstraction is the only thing that permits communication and creates trust; but cultural enclaves resist standardization. It is now rare for this aloofness to achieve the power of a fundamentalist resistance aimed at preserving one's own culture. For however much the desires for distinctions can be agitated and politicized against each other in this 'culture war,' in the end they remain oriented toward a structure of common differences. Popular mass culture permits travels through time and space without requiring viewers and listeners to leave their homes. For the first time, electronic mass media like CNN and MTV can reach overarching audiences with 'major events,' including warlike conflicts, athletic competitions, and appearances by entertainment stars. There are topics that enthrall literally the entire world, and there are growing audiences for global stagings. But this thematic concentration is accompanied by a fragmentation of local audiences, which again can illustrate the dialectics of globalization and localization.

3. Arenas of Transnational Citizenship

If the nation, in combination with government by bureaucratic agencies and popular representation, has been the fulcrum of personal identity and the condition of social affiliation since the nineteenth century, at the beginning of the twenty-first century, flexible forms of belonging and community transcending the nation-state challenge the representivity and legitimation of democratic rule. Again, the question is: Between local cultures and global markets, has the *homo politicus* become a dying species?

3.1 Transnational Governance

Governance without government, i.e., collective binding decisions outside the familiar arena of the nation-state (Rosenau and Czempiel 1992), is increasing, as evidenced by the multilateral regimes of the World Bank, the International Monetary Fund (IMF), and the World Trade Organization (WTO), as well of course as of the 'multilevel politics' of the European Union. These decisions are also effective, often to the dismay of national governments and populations, and they are no longer made solely unanimously, but increasingly also by majority votes. But the chains of legitimation for these supranational and transnational institutions are too long, and supranational and international regimes are suspected of disrespecting essential democratic prerequisites, namely the accountability of representatives and the congruence between those who make decisions and those who are affected by them. Attempts to ameliorate these defects, for example, by parliamentarizing the European Union, only underscore the difficulties of transnational citizen participation; the worldwide elections of the executive board of ICANN (fall 2000) show that there is a certain desire for worldwide participation and codecision, but also how difficult it is to organize such a process outside the nation-state. Only a few thousand out of millions of Internet users took part;

we can speak at best of 'successful failure,' from which conclusions should be drawn for the status of the world citizen in the age of global markets and information (Held 1995).

3.2 Complex Multilateralism

Up to now, participation in the transnational sphere has been mostly informal and unconventional. Analogously to the transition from medieval municipal citizenship to modern national citizenship, a space of communication has grown that transcends the nation-state's sphere. Here, individuals not only engage in economic and scientific exchange or unite culturally and religiously, but can also engage in political activity. One should mention more than 5,000 international nongovernmental organizations (INGOs) and, recently, social protest movements of global reach that articulate political will at conference tables, on the streets, and sometimes in both arenas. Both meanwhile fill the entire political space by devoting themselves to such topics as environmental protection and sustainable development, human rights, and equal rights for women. They can mobilize substantial resources and organize worldwide campaigns, with which they have attained the respect of and, if not a seat, at least a voice among representatives of states and international organizations. Such transnational activities can at least provide an impetus of agenda-setting, but they also extend as far as the sphere of problem solving and decision making. Here, in a broad and still undefined field, where the 'boundaries of the political' are still unstable, transnational citizenship is crystallizing; as a rule, formulation of political demands and decision making in these organizations and social movements are, alas, just as lacking in democratic legitimation and control. Nation-states are still the most effective accumulations of political power. Nevertheless, violations of human rights by dictators and ruling by force are generally made widely public and effectively condemned, ignoring claims to political or cultural autonomy. Basic rights have been bindingly codified in international conventions to the degree that this has resulted in an international jurisdiction, albeit in embryo form and without any means of enforcement.

In particular, the USA blocks any form of 'complex multilateralism' as soon as it relativizes US hegemony. Viewed in this way, it is a pretty irony but no coincidence that, once again a movement that intensely criticizes the lack of democratic citizen participation in international regimes spread from the US West Coast to circle the globe, this time under the slogan 'Seattle.' In December 1999, the WTO met there for a conference to address routine questions of the further liberalization of world trade. The meeting was poorly prepared, but this caused less trouble for all the government representatives than did the surprising fact that thousands of determined demonstrators blocked the entrance to the conference building and that the sometimes violent protests precluded any normal course of conferring. The wisps of tear gas have moved on, and, in response to other summits, the baton of similar protest has passed to other summits in Davos, Washington, Prague, and Seoul.

'Seattle' has meanwhile become the mantra of a new transnational social protest that makes use of the latest communications technology and may thus be regarded as the first virtual protest movement of its kind (O'Brien et al. 2000). Economic globalization from the top has found its counterpart in this movement, an opposition from below that is equally despatialized and that, where it is not protectionist and nationalist-populist, calls for an alternative, and more inclusive, form of globalization and, under the honorable slogan 'No globalization without representation,' sets out to battle the autocracy of international organizations. The movement is very heterogeneous and runs all the risks of short-livedness and fragmentation that social movements have always incurred. It includes environmental protectionists and trade unionists, Christians and secular human rights advocates, workers and academics, reformers and revolutionaries, libertarians and anticapitalists; and these groups find only small areas ideologically and organizationally in common.

Environmental protectionists lambast the destructive effects of industrial methods of production; the religious people denounce its materialistic disposition and indifference to human dignity; socialists of every stripe are bothered by exploitation and private property; others reject the consumerism of McDonald's and Coca-Cola. Ultimately, these groups contrast with each other as sharply as they do with their common enemy, the international organizations: Ecology and the economy can be arch-enemies; protectionist unionists want to close borders that dyed-in-the-wool cosmopolitans want to eliminate; development aid workers in the South attack the environmental protection regulations of the North as ecoimperialism, and so forth.

Here one can not predict whether such fronts will resolve into an opposition inspired by anticapitalism (like that at the end of the nineteenth century) against globalism or, more probably, into another reform movement from within globalism. But the logic of inclusion seems inescapable. In transnational spaces, 'internal affairs' no longer exist. What the humanist thinker Erasmus anticipated, is common sense now: one cannot be a citizen of a city without becoming a citizen of the world.

See also: Citizenship and Public Policy; Citizenship, Historical Development of; Citizenship: Political; Citizenship: Sociological Aspects; Global Governance; Globalization: Political Aspects; Migration, Theory of; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns

Bibliography

Appadurai A 1996 *Modernity at Large. Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN

Bauböck R 1994 *Transnational Citizenship. Membership and Rights in International Migration*. Elgar, Aldershot, UK

Bourne R 1916 Trans-national America. *Atlantic Monthly* **118**: 86–97

Clifford J 1997 *Routes, Travel, and Translation in the Late 20th Century*. Harvard University Press, Cambridge, MA

Czempiel E-O (ed.) 1969 *Die anachronistische Souveränität* (Politische Vierteljahresschrift, Vol. 1) Westdeutscher Verlag, Opladen, Germany

Faist T 2000 *The Volume and Dynamics of International Migration and Transnational Social Spaces*. Clarendon Press, Oxford

Hannertz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London

Held D 1995 *Democracy and the Global Order. From the Modern State to Cosmopolitan Governance*. Stanford University Press, Stanford, CA

Leggewie C 1999 Turcs, Kurdes et Allemands. Histoire d'une migration: de la stratification sociale à la differenciation culturelle, 1961–1990. *Le Mouvement Social* **188**: 103–118

Luhmann N 2000 *Die Politik der Gesellschaft*. Suhrkamp, Frankfurt/Main, Germany

O'Brien R, Goete A M, Scholte J A, Williams M 2000 *Contesting Global Governance. Multilateral Economic Institutions and Global Social Movements*. Cambridge University Press, Cambridge, UK

Ong A 1999 *Flexible Citizenship. The Cultural Logics of Transnationality*. Duke University Press, Durham, NC

Rosenau J N, Czempiel E-O (eds.) 1992 *Governance Without Government: Order and Change in World Politics*. Cambridge University Press, Cambridge, UK

Tomlinson J 1999 *Globalization and Culture*. Chicago University Press, Chicago

C. Leggewie

Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns

The subject of 'transnational religious identity' is unexpectedly complex. It refers to personal or collective religious identities that transcend enclaves, localities, regions, nations, and states to attain a wider unity of belief, practice, or community. Transnational religious identities also imply an inherent tension between claims to universality and the specifics of language, ethnicity, place of origin, locally shared culture, values, and history. The term also implies the existence of the nation-state system—with its own competing claims for personal and collective identity—as the principal worldwide form of political sovereignty. Transnational religious identities are continuously reconstructed in different contexts and ongoing struggles over ideology and authority in which shifting coalitions of elites, counter elites, ethnic groups, or those excluded from local authority structures compete for voice and recognition.

1. Constructing Religious Identities

Transnational religious identities, like other personal and collective identities, are socially constructed. They are not given or 'natural,' although the advocates of some religious identities, like secular nationalists, often make such claims. William Robertson Smith's influential *Lectures on the Religion of the Semites* (Smith 1956) traced the development of Semitic religious practices and institutions, stressing how a people's social and political institutions, and the shifting boundaries between their own and 'other' communities, contributed to shaping their religious and ethical conceptions. Thus the prophecies of Ezekiel and Samuel differed not because of a contradiction in the Bible, but because each prophet addressed himself to different groups and historical contexts.

For Smith and his successors, a key element in the growth and development of religions was the growing awareness that ideas and practices of community are socially constructed even if they are firmly rooted in divine revelation. In this view, a more 'modern' understanding of religious belief and practice implies a self-consciousness of ritual and tradition, places a high value on individual responsibility and commitment, and recognizes the existence of other religions and accords them some measure of legitimacy. In this respect, such Quranic verses as the following, revealed in seventh century Arabia, are distinctly modern: 'To each among you, we have ordained a law and assigned a path for each of you. Had God pleased, he could have made you one nation, but His will is to test you by what He has given you; so compete in goodness' (Quran 5: 48). From its beginnings in seventh century Arabia, Islam as religious tradition and practice has been remarkably open to the outside world. The vast expanse of the Muslim majority world, ranging from Morocco to Indonesia, inevitably meant that it came to encompass a variety of civilizational and cultural forms. In Arabic, the Quran exhorted its initial seventh century audience to break through parochial identities of tribe and kinship and aspire to a wider, more inclusive community of belief and practice.

Other religious traditions, including Judaism and Christianity, show similar diversity of form, commingling local and universal elements. Pilgrimages in all faiths serve to project believers across lines of immediate locality or group. Pilgrims set off from home in the hope of reaffirming contact with a spiritual center, encountering 'others' and returning home with a sharpened awareness of difference and similarity.

Whether to Mecca, Jerusalem, Ayodhya, Rome, the holy places of Shi'a Muslims or to the shrines of Sephardic saints (*zaddiqim*) in North Africa and Israel, pilgrimage creates unity as it highlights differences of locality.

Catholic missionaries and the adepts of Sufi religious brotherhoods also carried the word and practice of their respective faiths across vast spaces before they became nation-states (Rudolph and Piscatori 1997). The missionary activities of the Catholic Church in Latin America date from the Spanish conquests of the sixteenth century. Missionary orders such as the Jesuits and Franciscans implanted complex networks of churches, missions, and schools, transferring people and resources between Spain and the colonies. In the Middle East in the nineteenth century, French Catholic and American Protestant missionaries opened religious schools in the Levant, and, from the 1880s onward, the Alliance Israélite Universelle sent emissaries throughout North Africa and the Ottoman Empire. Although not primarily religious in orientation, Alliance missionaries had a profound impact on the self-perception of the Jewish communities where they worked. The work of these missionary enterprises was aimed, respectively, at Christians and Jews. These activities occurred within a framework of economic hegemony and political domination in which foreign powers actively sought to protect the interests of Middle Eastern Jews and Christians. However, these 'minority' communities were widely distributed geographically throughout the Middle East in the nineteenth century, so that missionary activities aimed at them had a significant indirect impact on the region's Muslim majority.

Sufi orders had a major role in propagating Islam, both in the Islamization of North Africa and in the subsequent consolidation of Islamic domination. Some orders, such as the Bektashiya, attracted as adepts the political elite of the Ottoman Empire. Other orders, such as the Tijaniyya, had a major role in the spread of Islam to West Africa. Some orders were highly structured, but most were only loosely linked through informal but pervasive networks of affinity and trust. Their forms of association were widely understood and respected throughout the Muslim world, and their organizational forms continue to serve as a template for many contemporary religious movements.

2. *Religious Identity and Nationalism in the Nineteenth Century*

In nineteenth century Europe, the idea took hold that peoples could be divided into nations sharing a common language, religion, ethnicity, and culture, and therefore should form separate political units. This notion was complemented by a heightened self-consciousness, systematization, and explicitness of religious traditions. By the 1830s, the plural 'religions' came into common currency, with a similar development in Arabic, to signify the idea of 'religion' as a system of beliefs and practices that could be set alongside nonreligious elements of the human condition and apart from other religions (Smith 1963).

The linkage between nationalism and religion was implicit and the boundaries of religious and national identities rarely coincided. Partially in response to political and ideological developments in nineteenth-century Europe, a pan-Islamic movement developed in the late nineteenth century. It continues in modified form to the present. Although the pan-Islamic movements of the 1880s bore similarities to the pan-Hellenic and pan-German movements of the time, only pan-Islam was based formally on a religious commonality. The pan-Islamic movements were initially confined to the educated elite, although this elite claimed to speak for all Muslims. The combined French and Spanish effort to crush the rebellion of Abd al-Karim al-Khattabi in 1927 generated protests throughout the Muslim world and appeals for pan-Islamic unity, but the results were transient and minimal.

3. *The Re-emergence of Religion in Public Life*

Prevailing theories of modernity and modernization in the mid-twentieth century assumed that religious movements, identities, and practice had become increasingly marginal to modern or modernizing societies, and that only those religious intellectuals and leaders who attached themselves to the nation-state would continue to play a significant role in public life. As Casanova (1994) notes, by the late 1970s this prevalent view rapidly eroded through the impact of several nearly simultaneous developments: the Iranian revolution, the rise of the Solidarity movement in Poland, the role of liberation theology in political movements throughout Latin America, and the return of Protestant fundamentalism as a force in American politics. Religion continues to play a significant role in politics and society worldwide. Even locally rooted movements benefit significantly with transnational sponsors or supporters, whether emigrants, states, or others sharing a group's objectives. Opposition to the Shah's regime was firmly rooted in Iran, but the safe haven provided for opposition leaders in exile and their transnational ties and freedom of movement and action were integral to the revolution's success.

Contemporary religious transnationalism and the identities it supports take many forms. Some are highly structured, while others are only informally organized and related primarily by affinity of goals. Informality and the lack of formal structure can be a strength in contexts of political and religious oppression.

At one end of the spectrum of organizational structure is the vertical transnationalism of the Roman

Catholic Church. In its encounter with the emerging system of sovereign nation-states in Europe, the Church incrementally reached an accommodation with nation states so that the Church hierarchy, although transcending the boundaries of nation-states, also worked within them. The close ties often formed with sovereign states often preserved a large measure of local freedom of action for the Church, but it also checked its ability to serve as a voice for social and political reform and for human rights.

The Second Vatican Council (1962–65) significantly altered the nature of Catholic transnationalism. With the Vatican's Declaration of Religious Freedom, *Dignitatis Humanae*, it unambiguously supported the doctrine of universal human rights and used its voice internationally to secure these rights for all peoples and not just Catholics. Vatican II also intensified interfaith dialog and formally recognized the diversity of Catholic belief and practice. It spurred the translation of liturgies into local languages and promoted lay participation in Church activities, further adapting the Church to local political and social circumstances.

The Church's overall hierarchical structure remained intact after Vatican II, but it was complemented and sometimes overshadowed in importance by decentralized grass roots networks and organizations created for purposes as diverse as Bible study, evangelism, and social and political action. Many organizations and networks had direct ties to the Church hierarchy, while others operated autonomously. In Latin America, as elsewhere, both types of networks benefited from transnational support. Dutch, Italian, French, Swiss, US, Spanish and other organizations acting transnationally contribute funds and personnel to sustain local developmental programs and action networks. Church-related human rights networks (along with groups with no religious affiliation, such as Amnesty International and Human Rights Watch) have played a significant role in sustaining resistance to authoritarianism and repression, giving concrete form to the idea and practice of transnational civil society.

Protestant transnationalism lacks the time depth of the Catholic presence in Latin America, Africa, and elsewhere, but missionaries from mainline Protestant denominations established a stronghold in Africa by the last part of the nineteenth century, often informally dividing spheres of influence with competing Catholic and Protestant groups, establishing schools and hospitals that often had no state counterparts. Schools have been especially important, and often formed the cadres of early nationalist movements (Beidelman 1982). Education, intended to inculcate Christian identity and practice, ironically undermined colonial authority and provided the cadres for African independence movements.

The first Protestant churches in Latin America were established primarily for European immigrants. By the 1960s, however, the mainline denominations were increasingly displaced by evangelical, Pentecostal, and fundamentalist churches and missionaries, trained and supported primarily from North America. Levine and Stoll (1997, p. 71) describe the growth of these groups, especially the 'Religious Right,' mixing evangelism with nationalism and anticommunism, as 'spectacular.' They suggest that these groups, in addition to providing needed community services, hospitals, and schools in areas where state facilities are minimal, appeal primarily to people experiencing intense social change and aspiring to improve their status and lives.

The ease of travel and the greatly enhanced role of the mass media have also enhanced the scope and ability of nongovernmental associations and networks to infuse an inherent transnational dimension to contemporary religious identity and practice. For both Catholic and Protestant groups, the 'horizontal' networks of civic and religious engagement, once established, become durable and self-sustaining. The transnational ties engendered through travel for labor migration, missionary activity, civic action, and human rights work sustains both institutional ties and a sense of moral community for all participants, not just for the elite. Unlike the hierarchical transnationalism of the Church in earlier periods, the loose network organization that characterizes many contemporary groups is more difficult for states to regulate or control. Moreover, the ease with which participants in such networks can spread information across state boundaries and secure external support serves as a brake on the state abuse of authority. This is as much the case for Latin America as elsewhere. Della Cava (1993), for example, notes the influential role of Catholic nongovernmental philanthropic and service organizations in Central and Eastern Europe as a catalyst for change and an alternative, although not a replacement, for attachment to the nation-state.

4. *Transnational Islamic Identities*

Islamic transnational movements occur in a significantly different context than Christian ones because Islam lacks a central organizational hierarchy. Although some Muslim intellectuals call for a spiritual or political unity, the Muslim world has almost always been characterized by a *de facto* pluralism of belief and practice. Even radical Islamic groups use the universal language of Islam but in practice seek only to establish an Islamic state or an Islamic order within their respective national borders even if they secure external support.

Some Islamic transnationalism is formal, organized, and linked to state organizations. After an arsonist's attack on the al-Aqsa mosque in Jerusalem in 1969, Muslim majority states founded the Organization of the Islamic Conference (OIC), which by the end of the twentieth century consisted of 52 member states.

However, this group has never advocated political unity and some member states, such as Nigeria, are anxious not to appear too closely identified with the organization. Nor has the OIC proved effective in the Iran–Iraq war (1980–1988), in defusing the Iraqi–Kuwaiti tension that led to the 1990 to 1991 Gulf War, or other conflict among Muslim majority states. However, its Islamic Development Bank, established in Saudi Arabia in 1974, has become one of the most effective aid agencies in the developing world, and various charitable funds associated with it have had a significant impact in its regions of activity.

Unlike Catholicism, in which vertical and horizontal transnationalism coexist, Islamic transnationalism is thus entirely horizontal. In addition to such formal organizations as the OIC, some clandestine or 'deniable' transnationalism is state-supported. After the Soviet invasion of Afghanistan in 1979, for example, the USA, with the cooperation of Pakistani and Saudi Arabian security services, supported the recruitment and training of mercenaries from other countries, including Egypt, Algeria, the Yemen, and Sudan to operate alongside Afghan refugees from bases in Pakistan. This 'joint venture' subsequently turned against its initial sponsors.

Islamic transnationalism is also manifested in nongovernmental movements such as the Muslim World League (*Rabitat al-'Alam al Islami*). It sponsors educational programs, religious scholarships for students to study at Saudi and Muslim world universities, and a vigorous publication program throughout the world. Such nonstate organizations, particularly through their formal presentation of Islamic issues and standardization of language and approach, help create common ideological communities that transcend state and national frontiers. Such organizations also show an increasing concern over the situation of Muslim minorities throughout the world, bringing their plight to the wider Islamic community worldwide. The concern of these groups with Muslims living in North America, Europe, and Australia is reci-procated by the Muslims', active concern with the situation of Muslims in their countries of origin or heritage and with the situation of Muslims in general. Other movements, such as those associated with the Turkish religious leader, Fetullah Gülen, are based on loosely connected clusters of supporters who lack formal central direction. Gülen's movement sponsors schools throughout Turkey, Central Asia, and Eastern Europe, as well as radio, newspaper, and television stations, notwithstanding the lack of central organization, the movement's educational and media programs have emerged as one of the most enduring forms of Turkish presence in Central Asia.

An even less formal and structured form of transnationalism is represented by *da'wa* (call to Islam) movements such as the Tablighi Jama'at (Masud 2000). Founded in India in 1927, the movement expanded worldwide after 1947 and currently operates in 80 countries. One principle of the movement is that the practice of Islam does not depend on educated elites alone. The movement's main efforts are directed at deepening the faith of other Muslims and, to promote unity, it avoids theological controversies and politics. South Asian in origin, the Tablighi Jama'at is also active in Europe, North America, and the Arab world. The movement offers no formal training institutes, although some institutions of religious learning are attached to Tablighi centers. Unlike other Muslim organizations, the Tablighi Jama'at lacks welfare or educational programs. Since the movement eschews the broadcast and print media as a matter of doctrine, its spread is entirely through personal contact. With its emphasis on strict adherence to rituals and a concern with the development of individual morality and renewal of personal faith rather than collective action and community change, the Tablighi movement has limited appeal to educated Muslims.

5. *Religious Identity and Globalization*

In the last third of the twentieth century, modern forms of travel and communication and mass education have accelerated religious transnationalism—the flow of ideologies, access to information on organizational forms and tactics, and the transformation of formerly elite movements to mass ones—rendering obsolete earlier notions of frontier as defined primarily by geographical boundaries and accelerating in pace the earlier spread of ideas already speeded through the introduction of print technologies. Thus the devotees of the North African Jewish saint, Baba Sali (Rabbi Israel Abouhatzeira) meet not only in Israel and Morocco, but also in New York and Europe.

Especially in the Muslim majority world, the advent of mass higher education since the mid-twentieth century, combined with the proliferation of mass media and the greater ease of communication and travel, has led to a significant reimagining of religious and political identities. One change is that patterns of authority have become more diffuse and open. Religiously committed activists increasingly replace traditionally educated scholars as sources of authority. Second, religion has implicitly been systematized and objectified in the popular imagination, making it self-contained and facilitating innovation. Questions such as 'What is my religion?' 'Why is it important to my life?' and 'How do my beliefs guide my conduct?' have become foregrounded in the lives of large numbers of believers, facilitating individual questioning and encouraging individuals to assume responsibility for their conduct and to participate in movements of social and political action. These transformations also mean that 'authentic' religious tradition and identity must be foregrounded and constructed rather than taken for granted. Finally, the greater access to

education, like access to the mass media, has broadened senses of language and community. Just as the rise of written vernacular languages in Europe since the sixteenth century and the spread of print technology created language communities wider than those of face-to-face interactions, mass education and mass communications play such a role in the contemporary world, transcending the narrow bases of local communities and dialects and creating new bases of religious identity.

See also: Buddhism; Christianity Origins: Primitive and 'Western' History; Citizenship, Historical Development of; Judaism; Religion: Nationalism and Identity; Religiosity: Modern; Transnational Citizenship: Cultural Concerns

Bibliography

Beidelman T O 1982 *Colonial Evangelism*. Indiana University Press, Bloomington, IN

Casanova J 1994 *Public Religions in the Modern World*. University of Chicago Press, Chicago

Della Cava R 1993 'Financing the faith': The case of Roman Catholicism. *Journal of Church and State* **35**: 37–59

Eickelman D F, Piscatori J 1996 *Muslim Politics*. Princeton University Press, Princeton, NJ

Hefner R W (ed.) 1993 *Conversion to Christianity: Historical and Anthropological Perspectives on a Great Transformation*. University of California Press, Berkeley, CA

Landau J M *The Politics of Pan-Islam: Ideology and Organization*. Clarendon Press, Oxford, UK

Levine D H, Stoll D 1997 Bridging the gap between empowerment and power in Latin America. In: Rudolph S H, Piscatori J (eds.) *Transnational Religion and Fading States*. Westview Press, Boulder, CO, pp. 63–103

Masud M K (ed.) 2000 *Travellers in Faith: Studies of the Tablighi Jama'at as a Transnational Islamic Movement for Faith Renewal*. Brill, Leiden, The Netherlands

Roy O 1994 *The Failure of Political Islam*. Harvard University Press, Cambridge, MA

Rudolph S H, Piscatori J (eds.) 1997 *Transnational Religion and Fading States*. Westview Press, Boulder, CO

Smith W C 1963 *The Meaning and End of Religion: A New Approach to the Religious Traditions of Mankind*. Macmillan, New York

Smith W R 1956 *The Religion of the Semites*. Meridian Books, New York

D. F. Eickelman

Transportation Geography

Transportation geography describes, explains, and predicts flows of people, goods, and information across space. Flows occur as a result of some form of exchange: people offer their services to employers in exchange for wages, friends exchange information about husbands and children, manufacturers produce and sell consumer goods to wholesalers and retailers. Flows occur within a network: the journey to work in vehicles on roads or rail, information via telephone lines or the Internet, and goods across roads, rail, shipping lanes, or air route. Other subfields of geography are concerned with places and their spatial organization. Places and flows are interdependent, hence the boundaries of transportation geography are quite fluid. This article describes the major concepts of transportation geography and discusses some theories and issues.

1. Understanding Flows

Transportation is critical to contemporary society. In the USA, the transportation sector represents 11 percent of gross domestic product (GDP), and transportation is the second largest expenditure category for households. Economic growth is closely linked with transportation: movement of people and goods is positively associated with GDP. Why is transportation so critical, and what motivates these flows? Transport is costly, both in terms of time and money. Hence there must be some benefit to justify incurring these costs. Flows occur when the benefits derived from transport exceed transport costs. Flows can be understood in terms of a few basic principles.

1.1 Flow Depends on Transport Costs and Gains from Trade

The first principle of flows can be illustrated with the journey to work. Consider a worker: the cost of commuting is incurred in order to remain employed. How much is a worker willing to pay in order to receive a given wage rate? The average US commute is about 23 minutes (based on 1995 national survey data); and just 4 percent are longer than 1 hour. Since commute costs reduce the effective wage (wages net of commute costs), it follows that as commute costs increase, fewer people will be willing to accept employment at a given wage level. The converse also holds: as commute costs decrease, more workers will be willing to accept employment at a given wage level. The likelihood of flow taking place is a function of transport costs and benefits from trade. This principle applies to goods and information as well as people.

1.2 Declining Transport Costs are Associated with Increased Flows

Transport cost includes both time and money costs. Technological change and infrastructure investment have increased travel speeds and reduced unit costs.

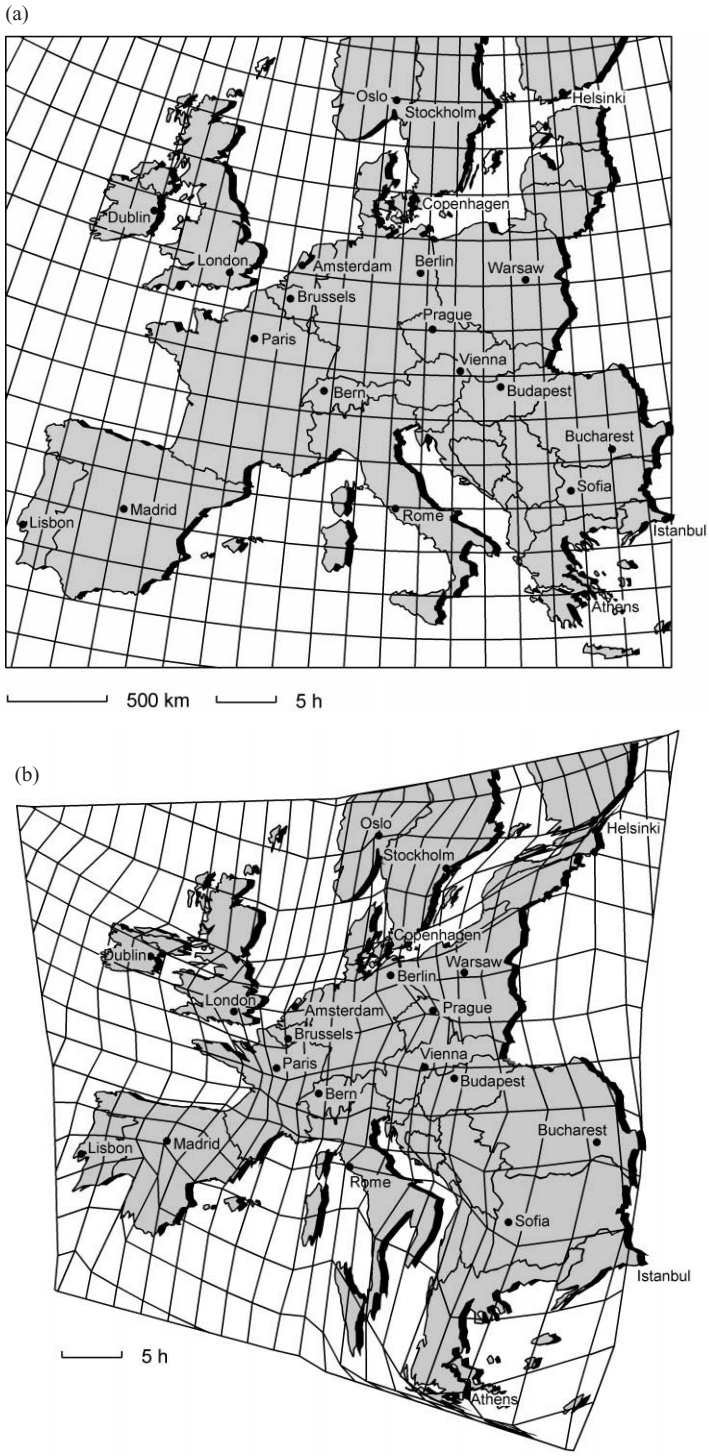

Figure 1
Time/space convergence: (a) map of Europe, constant travel speed; (b) map of Europe, 1993 rail transport network

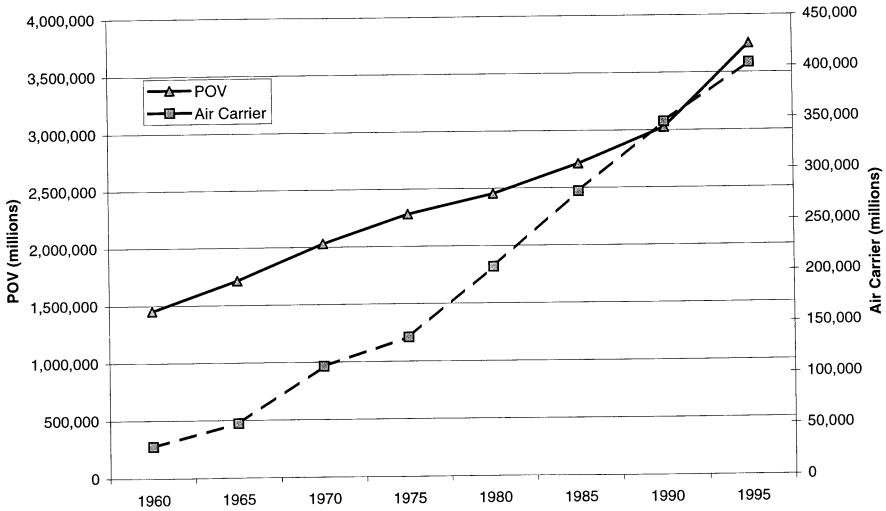

Figure 2
US passenger-miles of travel, air, and POV

The late nineteenth century transcontinental railroad offered trips from New York to San Francisco in 6 days; the trip by air today takes about 5 hours. Similarly, France's high-speed train offers service from Paris to Lyon in 2 hours. Around 1900 the same trip would have taken 6 or 7 days. These improvements also apply to intrametropolitan travel, but to a lesser extent, since intrametropolitan travel has so far been limited to surface transport modes. Transport cost has also greatly declined. For example, it is estimated that the average cost of commuting by streetcar in the USA at the turn of the twentieth century was about 20 percent of the average daily wage. Today the average cost for auto commuters is about 7 percent (Giuliano 1989).

Increases in speed result in 'time/space convergence,' a reduction in friction of both distance and time (Janelle 1995). Time/space convergence is illustrated in Fig. 1(a) and Fig. 1(b) for all of Europe, excluding Russia and the Ukraine (Spiekermann and Wegener 1994). Figure 1(a) provides actual spatial relationships and is calibrated on a homogeneous travel speed of 60 km per hour^{-1}. Figure 1(b) provides a picture of temporal relationships, based on the 1993 rail transport network. Note that Paris is 'closer' to major cities of Switzerland, The Netherlands, and Amsterdam, while in East Europe, Bucharest and Sophia are both 'further' from one another and from Central Europe. Time/space convergence is associated with globalization and the emergence of the time-sensitive economy (Janelle and Beuthe 1997).

Large reductions in transport costs have resulted in equally large increases in transport demand. Figure 2 shows the increase in US passenger-miles for private vehicles and for domestic air service, 1960 through 1995. Figure 3 shows the increase in ton-kilometers of surface freight transport for 10 European Union (EU) countries, 1970 through 1996.

1.3 Declining Transport Costs and Increased Flows are Associated with Development of the Space Economy

Historical development of the space economy is explained in part by declining transport costs. In the preindustrial era, nonmotorized transport was slow and difficult. There was little exchange between cities, because transport costs exceeded benefits at very short distances. Technological changes that reduce transport cost, e.g., 'space adjusting technologies,' allow regions to exploit comparative advantage, engage in regional specialization, and expand their areas of economic influence (hinterlands).

Nonmotorized modes of the preindustrial era were replaced with motorized modes in the nineteenth century: steamships and railroads. These technological advances greatly reduced the cost of intercity travel and hence led to large increases in both passenger and goods transport. The railroads in the US connected vast regions, resulting in the emergence of an integrated national economy. Railroads also contributed to the development of the national economies in Europe. Change accelerated in the twentieth century with the emergence of paved roads, autos and trucks, and the development of commercial air transport.

A famous example of the impact of reduced transport costs is the Erie Canal, built in 1825 to open a waterway between Lake Erie and New York by connecting Lake Erie to the Hudson River. The

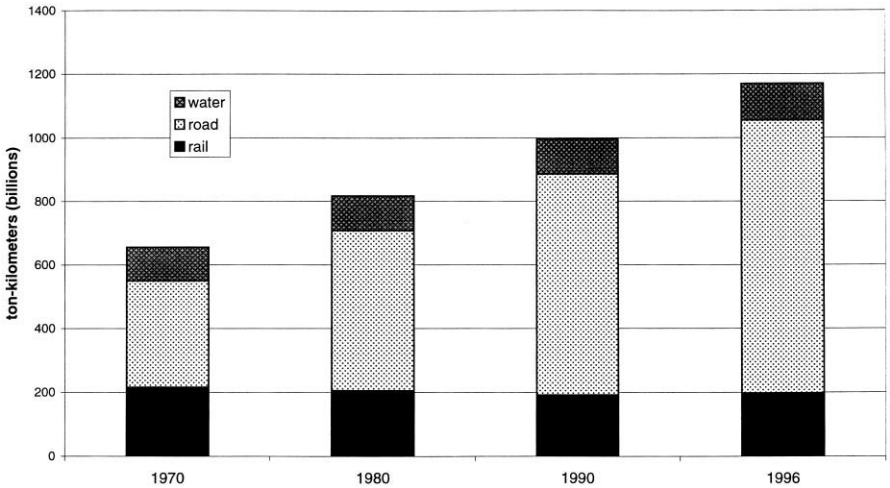

Figure 3
Freight transport trends, 10 ECMT countries

reduction in transport cost was dramatic: over-the-road transport cost 20 times as much per ton-mile as shipping via canal; travel time was 20 days over the road and 6 days via canal. Goods movement greatly increased. New York's hinterland was extended through the entire northeast region and its position as the US center of international trade and the dominant city in the US urban hierarchy was consolidated.

Empirical research on the relationship between transport and economic development is mixed. Research on both developed and less developed countries indicates that transport infrastructure is a necessary but not sufficient condition for economic development (Hoyle et al. 1998). The first links of major networks (rail, highway) have generally had significant impact, while the impact of additional links tends to be marginal (Munnell 1992). There are many failed examples of transport infrastructure investments made to promote trade and economic development, as for example the US efforts to promote economic development in the Appalachian region.

1.4 Transport and Urban Spatial Structure are Interdependent

Urban spatial structure is the outcome of economic, social, political, and cultural forces. Within the complex process of urban evolution, however, transportation has played a major role. The historical relationship between transport and urban form is well-established (Muller 1995). With faster and cheaper transport has come more decentralized and dispersed spatial structure, and as activities have become more spread out, travel has increased. Decreased accessibility has been offset by increased travel speeds.

US cities in the latter half of the nineteenth century were crowded and congested. Industrialization had generated a surge of urbanization. The lack of motorized transport forced population and businesses to co-locate in close proximity to one another. Electricity-powered rail transport proved to be the breakthrough technology. Starting in the late 1880s, urban railways allowed cities to grow outward as workers took advantage of the new mode to seek less crowded residential locations outside the city core, and gave the city center special dominance as the center of the intracity transport network. Urban rail was followed in the 1920s by the automobile and truck. Development of the highway network proceeded rapidly, increasing travel speed, and allowing ever more rapid decentralization. The post-World War II freeway era continued these trends, with both jobs and population decentralizing, and with the emergence of suburban activity centers. European cities have followed a similar trajectory, but under quite different social, economic, and political conditions, and with—to date—less dramatic spatial change.

The relationship between transportation and urban form has been the subject of much theoretical research. The familiar 'standard model' of urban location and spatial structure as developed by Alonso (1964) and others, serves to illustrate the close relationship between transport costs and urban spatial structure. Consider the simple case of a city where all employment is located at the center, transport cost is a linear function of distance, and households trade off housing and commuting costs in making their residential location decisions. How will workers locate themselves with respect to the city center? Clearly net wages will be highest at the city center, where transport

costs are near zero, hence workers will prefer location near the center, and locations near the center will be valued higher than locations away from the center. If workers could bid on locations, the result would be a gradient of both land prices and population density that declines with distance from the center. Declining transport cost results in less steep density gradients, as the relative value of living near the city center declines. Decreases in the density gradient have been documented for metropolitan areas around the world (Mills and Hamilton 1994).

2. Research Issues

Increasing concerns regarding the social and environmental consequences of urban transport, structural changes resulting from development of information and communications technologies, and a changing institutional context in transportation policy decision-making are affecting the research agenda of transport geographers (e.g., Hanson 2000). This section discusses three areas of current research: travel behavior; transportation, land use, and environmental costs; and transportation and information technology.

2.1 Travel Behavior

Transport geographers have examined personal travel at both aggregate and disaggregate levels of analysis.

2.1.1 Aggregate models. Various forms of spatial interaction models have been applied in aggregate analysis, most commonly the gravity model. The gravity model incorporates two basic factors that affect the level of flow between places: the population of each place (or some measure of potential for flow), and the distance between them:

The model in Eqn. 1 states that flow from origin i to destination j is a positive function of (a)

$$T_{ij} = O_i \frac{A_j d_{ij}^{-\beta}}{\sum_{\kappa} A_k d_{ij}^{-\beta}} \tag{1}$$

the potential of the origin zone for generating trips and (b) the relative attractiveness of the destination zone, and is a negative function of the distance between them. Gravity models have been applied to many topics, from intercity transport to local shopping flows. They have been extended to account for alternative travel destinations (intervening opportunities) as in the example above, differences in nodal characteristics, and differences in travel costs across modes or links (Taaffe et al. 1996).

The widespread availability of geographic information systems (GIS) has made possible the use of gravity formulations to measure accessibility patterns of different modes or population segments and hence analyze such policy issues as the distribution consequences of urban transport investment. The gravity model plays an integral role in state-of-practice urban transportation planning models: trip distribution (the allocation of trips to specific zonal destinations) is accomplished using a model that accounts for the relative attractiveness of zones as destinations and expresses distance as a generalized cost.

2.1.2 Disaggregate models. Understanding the travel patterns of individuals is a more complex task. Geographers and economists began to address the issue in the 1970s with the application of utility theory. In this case, travel decisions are discrete choices made on the basis of the utility of each alternative available. The utility of an alternative is a function of its attributes and the attributes of the individual making the choice. Since utility cannot be observed directly, utility-based models estimate the probability that a given alternative will be chosen.

The first applications of discrete choice were made to mode choice problems; later applications included destination choice, route choice, and car ownership. Mode choice models are now routinely used in transportation planning applications. Models have been extended to account for conditional or sequential choices, as for example, mode choice given car ownership. Multilayered models are one approach to developing integrated transportation planning models (Anas et al. 1998).

A major theoretical issue in travel behavior analysis is what should be predicted: the trip or the underlying behavior that motivates the trip. Given that travel constitutes a disutility, it is argued that understanding travel requires understanding activity patterns. Termed activity analysis, this approach to travel behavior predicts travel choices as a function of activity choices (Kitamura 1988).

The roots of the activity analysis approach to travel behavior lie with the Swedish school's time-space geography (Hagerstrand 1970). In this case the individual has a set of resources (time, money, vehicles) and a set of constraints (schedule, physical, and coupling constraints) that define the feasible set of activities they may accomplish over the course of a day. Hagerstrand terms this feasibility area the space-time prism. Activity-based models seek to capture the complex process of choosing a set of daily activities and their location, timing, and sequencing, given feasibility constraints.

Utility-based models assume rational behavior. Travel choices are a function of resources, preferences, and supply characteristics. Some geographers and planners argue that the rational behavior assumption is too strong; people act as much on attitudes and

perceptions as on resources and constraints. Travel choices are joint choices (resources and tasks are allocated among household members) which reflect (and affect) underlying attitudes and perceptions. They therefore borrow somewhat from the disciplines of sociology and psychology.

2.1.3 Travel as having positive utility. Travel has historically been considered a derived demand. However, some recent research suggests that this may not be the case (Mokhtarian and Salomon 2001). People may derive pleasure from travel to a far greater extent than assumed. They may conduct activities outside the home in order to travel, or they may travel to more distant destinations because the travel itself is part of the activity experience (e.g., take a drive and go out to dinner). Transportation planners have consistently underpredicted the growth in travel demand, even accounting for unanticipated changes in demographics, income levels, fuel prices, and other key factors. Positive utility of travel may explain people's greater-than-predicted propensity to travel.

The concept of travel as having positive rather than negative utility has significant policy implications. On the one hand, it helps to explain why travel demand is highly inelastic. On the other, it suggests that efforts to reduce travel for environmental or other reasons will be most difficult to accomplish.

2.1.4 Travel budgets. Finally, many geographers argue that time budgets determine travel. People make about four (one-way) trips per day and spend an average of about 1.3 hours each day traveling. These averages have been observed consistently both over time and across places, ranging from preindustrial agrarian villages to the major developed nations of the world. This consistency has led to the concept of a time budget: people allocate a given amount of time to traveling (Zahavi and Talvitie 1980). Hence distance traveled is a function of transportation technology and infrastructure. As travel speeds increase (say by construction of a new highway or rail line); people will use travel time savings to travel further distances. There are two main criticisms of the time budget. First, taken to its logical extreme, it conflicts with notions of rationality and utility-maximizing behavior. Second, the empirical observation may be an artifact of central tendency, and therefore may have little value as an explanation for human behavior.

2.2 Transportation, Land Use, and Environmental Costs

Transportation provides many benefits to society. As noted earlier, an efficient transportation system is critical to economic growth. People have greatly benefited from cheap and efficient mobility. The value placed on mobility is revealed in willingness-to-pay: households devote a significant portion of income to transportation, and the public has been remarkably willing to tax itself to invest in transportation infrastructure. However, mobility generates an array of environmental costs, notably air and noise pollution, traffic congestion, accidents, damage to the natural environment, and possibly global warming (Greene et al. 1997). In addition, everyone has not realized the benefits of mobility. Those without access to a private vehicle, for example, have suffered a relative decline in accessibility as activities have dispersed and public transport has become an increasingly poor substitute for the car. Concerns about these external costs are intensified by observation of current trends (Fig. 2 and Fig. 3). Extrapolation of these trends leads many to believe that policy actions must be taken to move toward a sustainable transport system (Newman and Kenworthy 1998). Hence transport is a key issue in the sustainability debate.

Sustainable transport has many dimensions. Strategies to increase sustainability range from improving transport technology to reducing travel via regulation, pricing, or land use policy. Transport geographers have focused on land use. The relationship between accessibility and travel is well known. Some geographers make a distinction between accessibility and mobility (e.g., Hanson 1995). Accessibility measures potential—the relative location of opportunities—and is a function of spatial form and transportation supply. Mobility measures outcomes—how much a person travels. Mobility is a function of both accessibility and travel resources. It follows that as accessibility increases, activities can be conducted with less travel, all else being equal. It is therefore argued that one promising strategy for sustainable transport is to promote higher density development. Proponents argue that such a strategy preserves accessibility while reducing mobility: activities can be accomplished closer to home, and greater access may allow more non-motorized travel. The overall effect is to reduce private vehicle travel and its associated external costs.

A corollary to this strategy is to expand public transport while imposing restrictions on the private vehicle. Public transport is seen as less damaging to the natural environment and contributing to urban quality of life. Effective public transport depends upon concentration; supportive land-use policies will create the trip densities required to support transit. Conversely, the presence of transit will promote higher density development (Newman and Kenworthy 1998).

Whether such strategies can succeed in reversing current trends in transport and urban structure is the subject of many debates. A positive correlation between development density and measures of private vehicle use has been observed in several empirical studies, but their validity is questioned on several

grounds, e.g., whether less car use is offset by congestion, whether the relationship is significant at densities that realistically could be achieved, or even whether a causal link exists. Studies that control for household characteristics and other relevant factors show a much weaker (and sometimes insignificant) relationship between density and private vehicle use for all but extremely high densities (Pickrell 1999). Other research has examined neighborhood level characteristics and its relationship to transit use or walks trips, with very mixed results.

The empirical research examines whether travel patterns would change, if land use and transport policies were used to promote reconcentration and less private vehicle use. The more interesting question is whether reconcentration and reduced private vehicle use at a scale sufficient to generate significant environmental benefits is feasible. Views on this question are decidedly mixed. Some argue that long-term economic and technological trends suggest continued decentralization and dispersion, while rising incomes will generate increasing demand for private vehicle travel (e.g., Morrill 1991, Giuliano 1999). Hence environmental costs should be addressed via technology, pricing, and regulation. Others argue that the social and environmental consequences of not changing these trends are enormous, and it is assumed that the political will necessary to achieve major change will therefore be found (e.g., Newman and Kenworthy 1998).

2.3 Transportation and Information Technology

The emergence of the information economy is an increasingly important theme of transport research. Advances in information and communications technology (ICT) are affecting all aspects of society and the economy. Transport geographers have focused primarily on direct effects of ICT: to what extent does ICT serve as a substitute for transport? Most of this research addresses telecommuting (working at home or at a nearby work center by communicating electronically with the office) (Mokhtarian 1998).

Telecommuting eliminates or shortens the daily commute. Remote offices, home offices, and mobile offices also reduce the work trip. Empirical research shows that travel savings from fewer trips to and from work are not entirely offset by nonwork travel. People are more likely to telecommute if they have very long commutes, generating travel savings in the short run, but suggesting that the opportunity may also motivate people to live even farther from their jobs. In addition, social and institutional constraints have so far prevented telecommuting from becoming a common practice.

Far more widespread than telecommuting is the use of telecommunications in place of personal contacts. Of course the telephone has served this purpose for a century. The difference now is in the variety and efficiency of communications devices, notably the now ubiquitous email. Interaction is growing; business and personal networks are increasingly complex and extensive. It seems reasonable to expect that more interpersonal connections would result in more rather than less physical interaction.

The impacts of ICT on work patterns are potentially far more extensive than these direct substitution effects. ICT is fundamentally changing the structure of the workplace and the organization of work. Labor statistics show that temporary work and self-employment are increasing more than proportionately with job growth. There is some evidence of declining job tenure. Changes in the structure of work will no doubt affect commuting and residential location patterns. Will households increasingly choose residences on the basis of neighborhood quality and other amenities, given uncertainty of future job location? Will access to job opportunities become more important, promoting location close to large job centers? Research on these questions is just beginning.

The impacts of ICT on firm location and on the structure of production also have enormous implications for transport. Outsourcing, customized production, flexible production, and just-in-time management of production and distribution characterize restructuring. The emergence of spatial production networks implies increased goods movement and growing dependence on an efficient global transport network.

Preliminary research indicates strong complementarity between ICT and freight transport (Plaut 1997). The freight transport sector has rapidly adopted ICT to increase speed and efficiency of freight flows. The growth of multinational firms and global trading networks also suggests more long distance passenger transport. While ICT allows for rapid and cheap electronic communications, the need for face-to-face communication among global managers remains. Hence the rapid increase of commercial air transport observed since the 1980s.

Time/space compression is bound to increase as the boundaries between transportation and communication break down. Indeed, ICT calls into question the very notion of accessibility as an essentially physical concept (Hodge and Janelle 2000). The direct and indirect effect of ICT will be a major area of future research in transport geography.

See also: Demography, History of; Development and Urbanization; Development, Economics of; Globalization: Geographical Aspects; Migration, Economics of; Migration, Theory of; Place in Geography; Space and Social Theory in Geography; Tourism, Geography of; Transportation Planning; Transportation: Supply and Congestion

Bibliography

Alonso W 1964 *Location and Land Use*. Harvard University Press, Cambridge, MA

Anas A, Arnott R, Small K 1998 Urban spatial structure. *Journal of Economic Literature* **36**: 1426–64

Giuliano G 1989 New directions for understanding transportation and land use. *Environment and Planning A* **21**: 145–59

Giuliano G 1995 Land use impacts of transportation investments. In: Hanson S (ed.) *The Geography of Urban Transportation*, 2nd edn. Guilford Press, New York, NY

Giuliano G 1999 Land use and transportation: Why we won't get there from here. *Transportation Research Circular* **492**: 179–98

Greene D, Jones D, DeLucchi M 1997 *The Full Costs and Benefits of Transportation*. Springer-Verlag, Berlin

Hagerstrand T 1970 What about people in regional science? *Papers, Regional Science Association* **24**: 7–21

Hanson S 1995 Getting there: Urban transportation in context. In: Hanson S (ed.) *The Geography of Urban Transportation*, 2nd edn. Guilford Press, New York

Hanson S 2000 Transportation. In: Sheppard E, Barnes T (eds.) *A Companion to Economic Geography*. Blackwell, Oxford, UK

Hodge D, Janelle D 2000 *Measuring and Representing Accessibility in an Information Age*. Springer-Verlag, Berlin

Hoyle B, Leinbach T, Smith J, Spencer A 1998 The role of transport in the development process. In: Hoyle B, Knowles R (eds.) *Modern Transport Geography*, 2nd edn. Wiley, Chichester, UK

Janelle D 1995 Metropolitan expansion, telecommuting, and transportation. In: Hanson S (ed.) *The Geography of Urban Transportation*, 2nd edn. Guilford Press, New York

Janelle D, Beuthe M 1997 Globalization and research issues in transportation. *Journal of Transport Geography* **5**(3): 199–206

Kitamura R 1988 An evaluation of activity-based travel analysis. *Transportation* **15**: 9–34

Mills E, Hamilton B 1994 *Urban Economics*, 5th edn., Harper Collins, New York

Mokhtarian P 1998 A synthetic approach to estimating the impacts of telecommuting. *Urban Studies* **35**(2): 215–41

Mokhtarian P, Salomon I 2001 Driven to travel: The identification of the mobility-inclined market segments. In: Black W, Nijkamp P (eds.) *Social Change and Sustainable Transport*. Indiana University Press, Bloomington, IN

Morrill R 1991 Myths about metropolis. In: Hart J (ed.) *Our Changing Cities*. Johns Hopkins University Press, Baltimore, MD

Muller P 1995 Transportation and urban form: Stages in the spatial evolution of the American Metropolis. In: Hanson S (ed.) *The Geography of Transportation*, 2nd edn. Guilford Press, New York

Munnell A 1992 Infrastructure investment and economic growth. *Journal of Economic Perspectives* **6**(4): 189–98

Newman P, Kenworthy J 1998 *Sustainability and Cities: Overcoming Automobile Dependence*. Island Press, Washington, DC

Pickrell D 1999 Transportation and land use. In: Gomez-Ibanez J, Tye W, Winston C (eds.) *Essays in Transportation Economics and Policy*. Brookings Institute, Washington, DC

Plaut P O 1997 Transport-communications relationships in industry. *Transportation Research A* **31**(6): 419–29

Spiekermann K, Wegener M 1994 The shrinking continent: New time-space maps of Europe. *Environment and Planning B* **21**: 653–73

Taaffe E, Gauthier H, O'Kelly M 1996 *Geography of Transportation*, 2nd edn. Prentice-Hall, Upper Saddle River, NJ

Zahavi Y, Talvitie 1980 Regularities in travel time and money expenditures. *Transportation Research Record* **750**: 13–9

G. Giuliano

Transportation Planning

Planning for the movement of people, goods, and ideas about cities and regions requires a systems framework. This reflects the intimate, two-way relationship that exists between transportation facilities and the social, economic, and environmental quality and character of places that are served. It is in this broader, more inclusive context of transportation planning that the field finds itself in the midst of a paradigm shift.

1. A Balanced Approach to Transportation Planning

Historically, transportation planning has been about ensuring efficiency in movement—that is, getting from point A to point B as quickly and as safely as possible. In the industrialized world, expansive overpasses and monumental expressways that criss-cross metropolitan areas stand as edifices of this philosophy. However, with time the cumulative impacts of planning for speedy, effortless movement, or for mobility, have revealed themselves—in some ways, positively, such as increased economic productivity and modernization, but in other ways, negatively, in the form of air pollution, worsening traffic congestion, wasteful sprawl, severance and degradation of neighborhoods, and isolation of people by class, race, and physical abilities. For this reason, some critics call for a more balanced, ecological approach to the field, one that accepts limits and acknowledges co-dependencies and connections (Ewing 1995). Other objectives besides mobility that need to enter the mix include accessibility, livability, and sustainability.

The risks of planning principally for mobility are underscored by US experiences. The United States averages 750 automobiles per 1,000 inhabitants, making it the world's most automobile-dependent society. The nation's economic prosperity is intimately tied to systems of highways, railways, and air transport that allow for efficient collection and distribution of people, raw materials, goods, and ideas. However, with under 5 percent of the world's population, the United States consumes 25 percent of petroleum-based fossil fuels and produces a comparable share of all energy-related carbon dioxide emissions. Cities like Los Angeles and Denver suffer from some of the worst photochemical

smog anywhere. As poor countries of Africa and Asia prosper and begin to mimic the vehicle ownership patterns of the United States and other industrialized nations, the world's finite resource base and capacity to absorb the astronomical increases in greenhouse gases without adverse climate change, many fear, will be pressed to the limits.

1.1 Mobility Planning

Time is a much-valued, irretrievable resource, thus planning for speedy movement has always been, and will always be, central to transportation planning. The standard methodology for mobility-based planning is the four-step travel-demand forecasting model. This involves a series of sequential, independent mathematical steps which lead to bottom-line forecasts of future vehicle trips on major highways and transit lines in a metropolitan area. The programming and construction of capital improvements like new highways or rail-transit extensions are based on the outputs of this four-step model.

Step one, called 'trip generation,' estimates numbers of trips per household by purpose (e.g., to work, for shopping) based on factors like household size, income, and number of motor vehicles. Step two, called 'trip distribution,' projects the flows of estimated trips to competing destinations, whether employment centers, shopping plazas, entertainment venues, or elsewhere. Traditionally, a gravity model is used to distribute estimated trips spatially—for example, a shopping mall with a million square feet of floor space can be expected to draw twice as many trips as one with a half million square feet, adjusted for impedance effects of travel time or distance. The third step, called 'modal split,' divides forecasted travel flows, by purposes, among competing modes, like drive-alone car, carpool, public transit, or walking. Discrete choice models that estimate the likelihood of selecting a particular mode based on its service-price attributes and characteristics of the traveler (e.g., income) are normally used. The final step, called 'traffic assignment,' involves computerized loading of predicted flows by mode on planned highway and transit networks to evaluate performance, with performance defined mainly in terms of average speeds.

In recent years, transportation planners have sought to refine the four-step model in several ways. One is to represent the dynamic, synergistic relationship between transportation and urban settlements patterns—to reflect, for example, that while new growth generates traffic, as roadways start filling up, growth starts migrating elsewhere (Putman 1991). Another is to model household activities as opposed to vehicle trips. Called 'activity-based micro-simulation,' this approach uses nonlinear mathematical programming tools to simulate activity patterns across a metropolitan landscape and the vehicle trips as well as tailpipe emissions that will result. In the United States, the Los Alamos National Laboratory has developed a micro-simulation package, called TRANSIMS, that projects movement of individuals in Portland, Oregon second-by-second according to where future activities are most likely to be sited. The model aims to capture the many behavioral adjustments that are made to changing traffic conditions—e.g., when a trip takes too long, people seek out other routes, switch from car to bus, leave at different times, or opt to forego the trip altogether.

1.2 Accessibility Planning

An alternative to planning for movement is to plan for places by making them more accessible. This can be achieved by bringing origins and destinations (e.g., homes and workplaces) closer to one another or by increasing speeds on the channelways that connect them. Accessibility planning embraces the notion that transportation is a secondary activity—it involves people going to places, be they places of work, of recreation, of worship, of retailing, etc. People do not consume transportation for its own sake, but rather for purposes of reaching oft-visited places. If anything, people want to minimize the time spent moving so that they can spend more time at desired destinations, rather than on the road. Focusing on accessibility as opposed to just mobility leads to a whole different set of planning strategies—for example, instead of laying more asphalt to expedite cars and trucks, mixed-use communities can be designed that reduce the spatial separation of activities. Many in-vogue land-use planning approaches, like jobs-housing balance, transit villages, and compact, neo-traditional development, are centrally about making cities and regions more accessible (see Calthorpe 1994, Bernick and Cervero 1997). The difference between planning for mobility versus accessibility is the difference between planning for cars versus planning for people and places.

So far, the Netherlands has made the most headway in reforming regional transportation planning to give equal emphasis to accessibility and mobility (Tolley and Turton 1995). Dutch planners draw mobility profiles for new businesses which define the amount and type of traffic likely to be generated. They also classify various locations within a city according to their accessibility levels. For example, sites that are well served by mass transit, that are connected to nearby neighborhoods by bike paths, and that feature mixes of retail shops receive high accessibility marks. Such locations are targeted for land uses that generate steady streams of traffic, such as college campuses, shopping plazas, and public offices. More remote areas that can be conveniently reached only by motorized means are assigned land uses that need not be easily accessible to the general public, like warehouses and factories. Thus, to make sure the 'right

businesses' get the 'right locations,' Dutch planners see to it that the mobility profile of a new business matches with the accessibility profile of a neighborhood.

1.3 Planning for Sustainability and Livability

As a rapacious consumer of natural resources and emitter of pollutants, the transportation sector must also be judged on the basis of sustainability—maintaining or improving, as opposed to harming, the natural environment (see Newman and Kenworthy 1999). Sustainability argues for resource-efficient forms of mobility, such as metrorail systems that link planned urban centers in the case of big cities and dedicated carpool and bus lanes that reward efficient motor-vehicle use in smaller ones. Congestion pricing, parking restraints, and the development of alternative-fuel vehicles are other strategies that embrace sustainability principles.

Livability applies principles of sustainable planning to the scale of the neighborhood. It is about the human need for social amenity, health, and well-being. Many European cities have brought livability to the forefront of transportation planning, opting for programs that tame and reduce dependencies on the private car (see Pucher and Lefèvre 1996). Traffic calming is one such approach, pioneered by Dutch planners who have designed speed humps, realigned roads, necked down intersections, and planted trees and flowerpots in the middle of streets to slow down traffic. With traffic calming, the street is viewed as an extension of a neighborhood's livable space—a place to walk, chat, and play. Automobile passage becomes secondary. After traffic calming its streets in the early 1990s, the city of Heidelberg, Germany witnessed a 31 percent reduction in accidents and a 44 percent reduction in casualties (Beatley 2000).

Car-sharing and car-limited communities have also gained popularity in European cities as ways of enhancing the quality of in-city living. Car-sharing relieves residents of the financial burden of owning one or more automobiles, allowing them instead to lease cars on an as-need basis through a cooperative arrangement. Car-share members can efficiently tailor car usage to specific purposes (e.g., small electric cars for in-neighborhood shopping and larger ones for weekend excursions). They also economize on travel by opting to move around by foot, bike, and mass transit more often. Presently, over 50,000 members share some 2,800 vehicles in 600-plus cities across Germany, Switzerland, Austria, Denmark, the Netherlands, Norway, Sweden, and Great Britain. More extreme is car-limited living, which is being actively promoted in parts of Amsterdam, Bremen, Germany, and Edinburgh, Scotland. In Amsterdam's Westerpark district, only 110 permanent parking spaces are provided for residents of 600 housing units. The project has attracted a number of middle-class families with children who value living in a setting where space is used for common greens and play areas, rather than for parking cars.

2. Coming Challenges and Policy Responses

Many powerful mega-trends are working against the broader objectives of transportation planning—namely, in favor of increased automobile dependence and usage at the expense of concerns like accessibility, sustainability and livability. The postindustrial world has given rise to 'travel complexity'—the trend toward irregular, less predictable work schedules, the geographic spreading of trips, and increased trip-chaining (e.g., linkage of trips, such as from work to the child-care center to the grocery store to home). All of these trends favor drive-alone, 'auto-mobility.' The share of trips by carpools and vanpools has fallen throughout US cities over the past two decades, despite the construction of dedicated high-occupancy vehicle (HOV) lanes and aggressive rideshare promotion (Pisarski 1997). Ridesharing's decline in metropolitan Washington, DC, traditionally one of America's strongest vanpooling markets, is rooted in the shift from predominantly government to increasingly high-technology employment; many of the region's software engineers and Internet-industry workers keep irregular hours and rely on their cars during the midday break, making it nearly impossible to share a ride to work. The entry of women into America's workforce, which soared from 26 million in 1980 to 63 million in 1997, has fueled trip-chaining—nearly two-thirds of working women stop on the way home from work, often to pick-up children at day-care centers (Sarmiento 1998). Telecommunication advances continue to diminish the need for spatial proximity, hastening the pace of new growth on the edges of metropolitan areas and in far-flung rural townships. As growth continues to spread out, there is a widening mismatch between the geography of commuting (tangential and suburb-to-suburb) and the geometry of traditional transportation networks, which tend to be of a radial, hub-and-spoke design. Circuitous trip patterns and mounting traffic congestion, especially in the suburbs and exurbs, have resulted.

Collectively, evolving economic, demographic, and urbanization trends have formed new space-time arrangements, conspiring against all forms of movement except the private car. Figure 1 portrays this evolution along a space–time continuum. The traditional monocentric city with concentrated activities (e.g., downtowns and 8-to-5 work schedules) supported point-to-point rail services reasonably well. As technology advances gave rise to polycentric settlements and less regular time schedules, more flexible forms of collective-passenger transport, like bus tran-

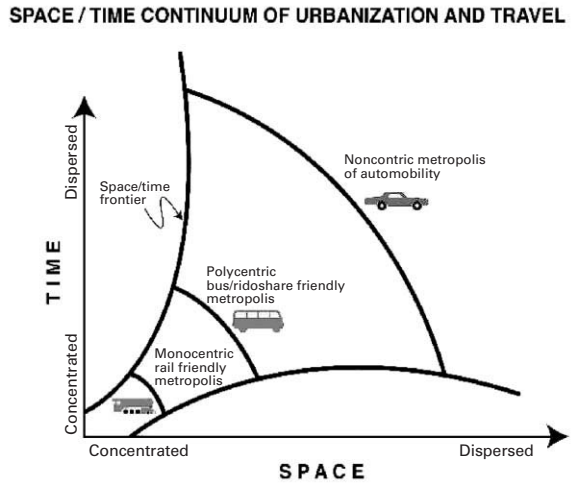

Figure 1
Travel complexity in a space–time continuum

sit and carpools, prospered. As cities and the regions of the future become increasing 'noncentric' and time schedules less certain and predictable, the frontier of space-time possibilities has expanded considerably. An immense challenge faced by the transportation planning profession is how to pursue the balanced agenda of mobility, accessibility, sustainability, and livability in light of these protracted, complicating trends.

2.1 Induced Demand

One of the most compelling reasons for more balanced, ecologically based transportation planning is the well-documented phenomenon of induced travel demand. Time and again, experiences show that building new roads or widening existing ones, especially in fast growing areas, provides only ephemeral relief—in short time, they are once again filled to capacity. A study using 18 years of data from 14 California metropolitan areas found every 10 percent increase in highway lane-miles was associated with a 9 percent increase in vehicle-miles-traveled four years after road expansion, controlling for other factors (Hansen and Huang 1997). Similar findings have been recorded in the United Kingdom (Goodwin 1996). In the United States, regional transportation plans, such as in the San Francisco Bay Area, have been legally contested by environmental interest groups on the very grounds that they failed to account for the induced travel demand effects of road investments and expansions.

Expanding road supply produces both near- and long-term shifts. In the short run, expanded capacity prompts behavior shifts—people change modes, routes, and times-of-day of travel. For example, once a new road is opened, many who previously commuted on the shoulders of the peak shift to the heart of the peak. Over the longer term, structural adjustments occur, such as people and firms locating on parcels near interchanges to exploit the increases in accessibility provided by a new freeway.

The challenge of transportation planning is to create constructive channels for the induced shifts prompted by new transportation infrastructure. To encourage 'green' modes of transportation, like vanpools and shuttles, for instance, planners can set up HOV lanes along corridors which are the recipients of shifted trips. Time-of-day swings in trip-making can become more efficient by relaying real-time information on travel conditions and least-congested paths via Internet web sites and on-board navigational systems. Locational shifts can be of a more compact, mixed-use, walking-friendly form through zoning, tax policies, and other instruments that reward efficient and sustainable patterns of development.

2.2 Technology

Worldwide, billions of dollars are being spent on making roadways and cars smarter, under the guise of the Intelligent Transportation System (ITS). These initiatives seek to increase the efficiency of automobile movements many orders of magnitude, relying on the kind of technology and intelligence gathering once reserved for tactical warfare. In Great Britain, Germany, and France, over 200,000 motorists currently subscribe to Trafficmaster, a nationwide traveler information system. Infrared cameras located every six kilometers automatically record the license plates of passing vehicles and transmit the data via radio to a centralized data center. By matching license plates across different camera locations, a megacomputer calculates the average speed of travel for each roadway segment and relays the information in encrypted form to small receivers and visual displays mounted on the dashboards of subscribers' cars. Alternative routes are suggested for avoiding traffic jams. One concern of making automobiles smarter is that travelers will flee buses and trains in droves. Mass transit must therefore also become smarter, through advanced technologies which optimize routing for door-to-door shuttles and which relay information on schedule-adherence to drivers and on expected times of arrival to waiting customers.

The one area where all sides agree that technology could yield important social, environmental, and economical benefits is by allowing for marginal-cost pricing, such as the collection of congestion tolls (Gómez-Ibáñez 1999). This is happening in Singapore, where an electronic road pricing (ERP) scheme was introduced in 1998 (Cervero 1998). The system applies

a sophisticated combination of radio frequency, optical detection, imaging, and smart card technologies. With ERP, a fee is automatically deducted from a store-valued smart card (inserted into an in-vehicle reader unit) when a vehicle crosses a sensor installed on overhead gantries at the entrances of potentially congested zones. The amount debited varies by time and place according to congestion levels. Cameras mounted on gantries snap pictures of violating vehicles to enforce the scheme. The system can identify a vehicle, deduct a charge, and capture a clear photo of its rear image at speeds up to 130 kilometers per hour. During its first year of implementation, traffic volumes declined by 15 percent on a major freeway along Singapore's eastern shoreline, matched by comparable patronage increases along a parallel mass rapid transit (MRT) route.

Another promising direction for new technologies which supports sustainability and livability objectives is the modification of vehicle propulsion systems. Hybrid electric propulsion can boost energy efficiency by an estimated 30 to 50 percent by electronically recovering braking energy, temporarily storing it, and then reusing it for hill climbing and acceleration. Visionaries see a future of ultralight hybrid vehicles that combine the advantages of regenerative electronic braking with on-board fuel cells (Von Weisäcker et al. 1997).

2.3 Institutional Reforms

Even with the most sophisticated technological breakthroughs, little progress can be made unless there is the institutional capacity to coordinate and implement new transportation programs and ideas successfully. Worldwide, transportation planning is mired by bureaucratic inertia and redundancies. In Bangkok, Thailand, for instance, more than thirty government agencies are responsible for the city's transportation planning, management, and operations. Many economists promote privatization of public transportation as a means to cut costs and spur innovations, however even increased competition can backfire without a workable institutional arrangement. This is underscored by experiences in Rio de Janeiro, where more than 60 private bus companies serving the city have independent timetables and fare structures, penalizing many users who must transfer across bus routes.

Germany and Switzerland have formed regional federations for coordinating mass transit services, called *verkehrsverbunden*. Greater Munich's federation, for example, functions as an umbrella organization for planning routes, services, and fares across the 5500-square-kilometer region's many rail and bus service offerings (Cervero 1998). Besides coordinating schedules and timetables, the federation pools farebox receipts and reallocates them to ensure socially equitable fares are charged while also rewarding individual operators for being productive and cost-effective.

Another noteworthy institutional reform in Europe has been the functional separation of roadbed ownership and maintenance (including highways and rail tracks), which remain in the public domain, from service delivery and which is open to public and private providers through competitive tendering (see Pucher and Lefèvre 1996, Cervero 1998). Within the confines of standards set by government sponsors (e.g., regarding timetables, fares, and routing), the lowest-cost provider delivers the service. Thus, the public sector retains control over how services are configured, leaving it to market forces to determine at what price.

In the United States, the eyes of the transportation planning profession are on the state of Georgia, where an all-powerful regional transportation authority has recently been formed. Called the Georgia Regional Transportation Authority (GRTA), the organization not only oversees the planning and expenditure of funds for all urban transportation improvements in the state, but also has broad control over regionally important land uses, like shopping malls, industrial parks, and sport stadia. Local land-use decisions must conform to regional transportation and development goals, otherwise GRTA can effectively veto decisions by threatening to cut off all state infrastructure funds. GRTA's formation was largely in reaction to decades of poorly planned growth in metropolitan Atlanta, matched by ever-worsening traffic congestion. The announced plan of a large high-technology employer to relocate out of Atlanta because of unsustainable traffic congestion and a declining quality of life was a political wake-up call. The region's new planning philosophy—one of balancing urbanization and transportation investments—aims to enhance mobility while also placing the region on a pathway that promises a more sustainable and livable future.

See also: Air Pollution; Location Theory; Planning Issues and Sustainable Development; Sustainable Transportation; Transportation Geography; Transportation: Supply and Congestion; Urban Activity Patterns; Urban Places, Planning for the Use of: Design Guide; Urban Policy: North America; Urban Sprawl; Urban System in Geography

Bibliography

Beatley T 2000 *Green Urbanism: Learning from European Cities*. Island Press, Washington, DC

Bernick M, Cervero R 1997 *Transit Villages in the 21st Century*. McGraw-Hill, New York

Calthorpe P 1994 *The Next American Metropolis: Ecology, Community, and the American Dream*. Princeton Architectural Press, Princeton, NJ

Cervero R 1998 *The Transit Metropolis: A Global Inquiry*. Island Press, Washington, DC
Ewing R 1995 Measuring transportation performance. *Transportation Quarterly* **49**: 91–104
Gómez-Ibáñez J 1999 Pricing. In: Gómez-Ibáñez J, Tye W, Winston C (eds.) *Essays in Transportation Economics and Policy*. Brookings Institution Press, Washington, DC
Goodwin P 1996 Evidence on induced demand. *Transportation* **23**: 35–54
Hansen M, Huang Y 1997 Road supply and traffic in urban areas: A panel study. *Transportation Research* **31**: 205–18
Newman P, Kenworthy J 1999 *Sustainability and Cities: Overcoming Automobile Dependence*. Island Press, Washington, DC
Pisarski A 1997 *Commuting in America II: The Second National Report on Commuting Patterns and Trends*. Eno Transportation Foundation, Washington, DC
Pucher J, Lefèvre C 1996 *The Urban Transport Crisis in Europe and North America*. Macmillan Press, Houndmill, UK
Putman S 1991 *Integrated Urban Models 2*. Pion, London
Sarmiento S 1998 *Household, Gender, and Travel*. Federal Highway Administration. US Department of Transportation, Washington, DC
Tolley R, Turton B 1995 *Transport Systems, Policy and Planning: A Geographical Approach*. Wiley, New York
Von Weisäcker E, Lovins A, Lovins L 1997 *Factor Four: Doubling Wealth, Halving Resource Use*. Earthscan Publications, London

R. Cervero

Transportation, Psychology of

The psychology of transportation is a field of applied psychology. It entails the study of human behavior and well-being in conjunction with the regularized movement of people, goods, and services regarding work, personal life, and community activity systems. In addition, it involves the interface of humans with transport vehicles and facilities, including their design for optimal operation, provisions for human needs, forms of utilization, and social impacts.

1. Overview and Scope

The domain of transportation has great relevance for psychology, precisely because of its salience in the routines of daily life and its importance for meeting personal, family, and organizational goals. Because various elements of transport systems are operated by people, psychology has an important role in system development, efficiency, and safety. Human cognitive, personality, and performance factors affect vehicle operation. Moreover, transport systems shape the structure of our communities and have a significant impact on both personal and community well-being.

The progress of human societies has always been associated with mobility and transportation. People respond to environmental constraints on mobility and strive to enhance well-being. What happens on a particular journey depends on many intertwined psychological elements, including personality disposition, attitudes about the origin and destination of the trip, and resources for choosing alternative travel modes and schedules, including telecommuting. This provides a vast terrain for psychological research and application of psychological knowledge. The creation and use of transportation systems, the experience and behavior of people in transportation conditions, and the effects of mobility constraints on personal and community health, are broad areas entailed by the psychology of transportation field.

Human factors have long been recognized as being intrinsically involved in the operation of vehicles; however, the psychology of transportation is much broader in scope than the interface between the human operator, the vehicle, and the road design (Rothengatter and Vaya 1997). It entails travel-mode decisions, driver attitudes and performance, the interplay between work and family, emotional well-being, and major societal issues such as the needs of special populations, the definition of the workplace, and the provision of transportation facilities.

The field of transportation has been dominated by engineers, urban planners, and economists; but the contributions of psychologists and sociologists are increasingly integrated into this arena (Altman et al. 1981, Wachs and Crawford 1992). Discussions of the social costs of transportation are now common, and psychological elements are given weight in such considerations. Because individuals, families, and work organizations recognize that they absorb hidden costs associated with traffic congestion, air pollution, and traffic noise, there is considerable merit for the role of psychology in understanding how to optimize transportation systems and to assess the consequences of travel constraints (Stokols and Novaco 1981).

2. Historical Development of the Field

Involvement by psychologists in the field of transportation began in the 1920s when transport companies and government agencies sought to improve the selection of drivers of public transport vehicles, and psychometric testing began to be used in choosing the operators of buses, trams, and trains. A central concern was to minimize accident risk, and psychological tests were used to screen applicants who could be anticipated to be accident-prone. Wartime studies of adjustment among military personnel and the early psychological formulations of stress, adaptation, and coping that resulted from observations made by psychological researchers working with the military extended nicely to the subject of driver performance demands and safety risk. In addition,

industrial psychology research on fatigue at the workstation transferred to the issue of vigilance in professional train and road drivers (Barjonet 1997).

As the motor car came to dominate travel in Western societies in the 1950s and 1960s, car ownership grew dramatically and accident rates and road deaths became serious public health problems. For example, in most European countries the number of deaths on the road doubled between 1955 and 1970. Public safety policies, such as speed limits, penalties for drunk driving, and ultimately seat belts, were instituted to curtail casualties, and research into the human causes of accidents came into play. Social psychological knowledge was used in public persuasion efforts, and theories and techniques of learning found fertile ground in driver education and rehabilitation. As the advent of the computer augmented the capacity to analyze and integrate data, the approach in the 1970s and 1980s to understanding driver-based safety risk often took the form of psychological modeling that incorporated personal background, attitudinal, motivational, and emotion variables, as well as ergonomic factors.

The growth in automobile commuting that occurred between 1960 and 1980, especially in the USA, prompted interest in modeling how trips were generated, how they were distributed across travel zones, and how trips were split across available transportation modes (car, bus, train). The boom in automobile commuting followed from substantial growth in the number of jobs, in the working age population, and in the number of women in the labor force. In addition, homes and jobs shifted to suburban locations. Because neither highway nor transit systems were designed for suburb-to-suburb commuting, traffic congestion in metropolitan areas became a salient issue. In the USA, the use of private vehicles between home and work almost doubled from 43 million users in 1960 to 83 million in 1980. At the same time, alternative modes of travel and vehicle occupancy declined and national investment in highway infrastructure fell far behind the demand for road use (Pisarski 1987).

European cities, with their treasured historic centers having limited road space and parking facilities, became concerned about the detrimental effects of traffic congestion and pollution and were keen to impose access restrictions. These trends spurred interest in the travel behavior of individuals. Computer-analytic methodology spurred development of travel demand modeling, which blended consumer theory in economics with choice theory in psychology, which assumes that people use some utility or value function in assessing choice alternatives.

Attention to the transportation needs of special populations, such as working mothers, the elderly, and those affected by physical disabilities, also came into focus during the 1970s and 1980s. An important component of the auto-commuting boom was a large increase in the number of women in the labor force (Grieco et al. 1989, Pisarski 1987). Single women with children who lived in suburbs were very dependent on the automobile (Rosenbloom 1992). In the USA, most trips made by working mothers and their children cannot be made by transit, and the impact of travel constraints on the quality of family life is an important topic. In the 1990s, telecommuting emerged, partly as a way to reduce strain on family life. The convergence of technological, economic, social, and psychological factors made work less place-dependent, and telecommuting was heralded as a viable alternative to auto-commuting (Mokhtarian 1990).

In the 1990s, the information age had many effects on transportation systems. The development of intelligent transport systems, such as in-vehicle devices to aid navigation (e.g., route guidance and congestion warnings), as well as the increased sophistication of driver panel displays, raised the importance of driver control, visual search efficiency, and mental load on driving proficiency. On the horizon of advanced technologies are Intelligent Vehicle Highway Systems, which control steering, speed, and spacing automatically. However, having a vehicle travel under the control of a computerized system raises psychological issues of personal control and freedom, which have been themes intrinsic to car ownership and driving.

3. *Human Factors and Driver Performance*

Psychological variables affect the operation of transport vehicles, as can be seen in studies of driver error and performance decrements (cf. Rothengatter and Vaya 1997). Fatigue has been the most researched topic in this regard. Ordinarily understood as physical exhaustion, it can also involve conditions characterized by boredom, task aversion, uncontrollable sleepiness, hypnotic trance-like states, or loss of alertness. Importantly, driving performance decrements occur before the driver is aware of the fatigue subjectively. Fatigued drivers become complacent and underestimate the effort needed to maintain driving performance. Hours of service limitations are implemented for professional drivers and pilots to guard against fatigue-related safety errors. Physiological correlates of fatigue have been studied in both driving simulator studies and open-road tests. Heart rate decreases with prolonged driving time, but results across studies have not been uniform. Focal fixation, detected by eye-tracking, is associated with fatigue, but it also occurs in conjunction with hazard perception.

Driver alertness is clearly related to collision avoidance. The driver, as an information processor, must be proficient in making time-to-collision judgments—e.g., time to collision is underestimated for smaller vehicles. Raising the rear brake light signal was a vehicle design modification to reduce rear-end collisions. The use of reflectors, warning signs to reduce speed on curves, and edge lines are road-design features intended to

enhance the input of safety information. Generally, the more enriched the visual environment, the lower the underestimate of time-to-collision.

Driver expectations and interpretations, however, interact with visual information. Expectancies developed from driving experience may have adverse effects. The person may omit perceiving relevant cues or fail to give them appropriate weight in estimating a hazardous condition. 'Driving experience' does not guarantee expertise, as an experienced driver may have fixed expectations about how a driving situation will progress, which may interfere with accurate perceptions and interpretations of relevant cues. Driver behavior is a function of what drivers believe and feel they can do, as well as what they see and are told.

4. *Driver Risk-taking, Attitudes, Motivation, and Emotion*

The study of driver performance lends itself to experimental lab analyses of perception, decision-making, and reaction time. In contrast, the research on driver risk-taking has been conducted largely by questionnaires. Driver attitudes and personality have been studied as indicators of the propensity to commit driving errors and traffic-code violations, thereby raising accident liability. Thus, rather than focus exclusively on the driver's psychomotor or perceptual skills, attention is given to higher-order cognitive processes and motivation (cf. Rothengatter and Vaya 1997)

Self-reported violations and risky driving behavior, as assessed by driver behavior questionnaire measures, have been found to predict involvement in accidents, independent of the amount of yearly driving. Because self-evaluations of driving habits are longitudinally predictive of driving violations and accidents, it then becomes feasible to identify problematic drivers and augment their safety motives. Since these relationships hold strongly for young men, they are commonly a target group for intervention.

The most consistent finding about driver risk-taking is its association with a sensation-seeking personality, which is thought to have a biological basis. High-sensation seekers are more likely to drive drunk, exceed speed limits, take risks in passing and lane changing, not use seat belts, receive citations for traffic violations, and have crashes. These findings have been obtained in studies conducted in several European countries, Canada, and the USA.

Human behavior is highly adaptive, and drivers adjust to changes in the intrinsic safety of the driving environment. In hazardous weather, drivers slow down and sharpen alertness. Alternatively, as safety in the driving environment is enhanced through improvements in roadway conditions and vehicle design, drivers may be inclined to drive faster, shorten the distance in following, overtake more readily, or be less attentive to hazards. Therefore, improved vehicle safety features may lead some drivers to overcompensate for perceived risk reduction, driving less safely because they feel more insulated from collision or injury.

Aggressive driving and its alias 'road rage' have become prominent topics in the traffic-safety field in many countries (Brewer 2000, Shinar 1998). However, aggression has been part of driving since chariots careened around the *Circus Maximus* in ancient Rome. Indeed, the symbolization of the automobile has been infused with themes of aggression, which is evident in vehicle names and designs. Uncontrolled anger and aggression on roadways do constitute a significant driver behavior issue, but neither the magnitude of the problem nor its societal trend is clear. While human aggression springs from many causes, aggression on roadways is a product of various disinhibitory influences (Novaco 1991), such as anonymity, opportunity to escape, increasingly fortified vehicles, and strong elevations in physiological arousal produced by driving situations. The personalized territoriality of the vehicle and diffusion of roadway aggression scripts in the media also facilitate antagonism in driving when the behavior of other drivers is viewed as provoking.

5. *Driver Selection, Training, and Rehabilitation*

Psychologists have helped to develop aptitude tests to insure public safety on road, rail, and air travel. Driving safety has been viewed classically as involving the three 'Es': engineering, enforcement, and education. Accident liability is primarily due to underestimation of risks, negligence or carelessness, personality, attitudinal disposition, or substance misuse, rather than to roadway or vehicle engineering. Hence, reduction in risky driving hinges on detection of problematic drivers and on driver training and rehabilitation programs. There is wide variability across countries in how systematically this is approached.

Reconviction for traffic offenses and for serious traffic offenses have been found in New Zealand national data, gathered prospectively, to be a function of age at first conviction, being male, number of prior convictions, and having a traffic conviction as the most serious offense; moreover, statistical models have achieved high predictive power in identifying traffic offenders at high risk for reconviction for an alcohol-related driving offense. In Germany, an association of psychologists has formed a commission of experts for quality management regarding driver diagnosis and improvement. In addition to developing psychological models for the prediction of re-offending drivers, their rehabilitation courses have reduced significantly drunk-driving recidivism. Such specifically targeted driver-education programs have been successful, but it has generally been found that secondary-school driver-

training programs do not reduce accidents, because it is inconsequential whether young people learn to drive from relatives, friends, or schools.

6. Travel Demand and Alternative Modes of Travel

Transport-systems planners use mathematical theories about choice in modeling travel demand, and psychological variables add to what is otherwise offered by economic models. Because choices are often grounded in perceptions and subjective valuations, psychological research methods help to ascertain the relationship between travel time and the desirability of travel-mode options or goodness-of-fit with personal needs. Travel time, cost of alternative modes, availability of mass transit, activity plans, work demands, and personal preferences are weighed by the traveler in making transportation choices.

Psychological theory has been applied to change travel behavior, such as using reinforcement theory to boost transit-ridership. Transit use has been promoted by offering cash and travel prizes through lottery-type drawings, as well as distributing value coupons and tokens (for food, goods, and services) with bulk purchases of transit passes or dispensing them to people on boarding a bus. Similarly, inducements for ridesharing have been given by providing special lanes on congested freeways for multiple-occupancy vehicles. Such schemes are more effective when the special lane has been added to the road, as opposed to when an existing lane has been converted. Other efforts to reduce solo driving have involved the provision of priority parking facilities and/or reduced parking rates for carpools and vanpools.

Embedded in the issue of transit use is its social image, particularly with regard to buses. In many countries, persons of high social status commonly do intercity travel by rail for work commutes and discretionary trips, but bus travel typically has the image of being less than prestigious. In addition, people may be disinclined to use the bus because of a desire to avoid crime victimization on the bus and at bus stops. Psychological knowledge can be utilized in modifying unfavorable images of transit use and to reduce unwarranted fears about safety risk.

7. Commuting Stress

Exposure to traffic congestion is stressful by virtue of impedance—the blocking or thwarting of movement and goal attainment. Travel impedance has both physical and perceptual dimensions (Novaco et al. 1990). 'Physical impedance' can be objectively measured by the distance and time of the journey, along with the number of road exchanges—each representing a node of congestion. Time of day is relevant, because congestion is greater in the evening, and commuting stress is compounded by strains from the workday. 'Subjective impedance' is the person's perception of commuting constraints, assessed by questionnaire items or structured interview. Because the commute is sandwiched between the home and the job, the demands and affordances of each of these domains, as well as personality factors, bear on the commuting experience.

Commuting stress has been found to be associated with elevated blood pressure, negative mood on arrival at work, negative mood at home in the evening, lowered frustration tolerance, cognitive performance impairments, illness occasions, work absences, job instability, lowered residential satisfaction, and decrements in overall life satisfaction (Koslowsky et al. 1995). Female solo drivers in high impedance commutes are most negatively affected, independent of job-involvement and household income. Because women tend to have more logistical responsibilities for home and family, they are perhaps more sensitized to stress on high-impedance routes (Novaco et al. 1991).

A useful concept in understanding commuting stress is 'inter-domain transfer effects,' whereby the psychological consequences of environmental conditions in one life domain (home, commuting, work, or recreation) transfer to another, either positively or negatively (Novaco et al. 1990, 1991). An example of a positive transfer effect is that a high level of residential choice buffers commuting stress. Examples of negative transfer effects are that exposure to commuting stress affects negative mood at home in the evening, controlling for a great many other relevant variables, and that commuting stress prospectively leads to job location change, independent of job satisfaction.

Dissatisfaction with the commute is a cost the employee absorbs, often with delayed effects on health and family life. However, commuting stress affects work organizations, as well. People will change jobs to cope with traffic congestion, but commuting stress is more often manifested in illness-related absences from work and with impaired work performance.

8. Important Future Topics for Transportation Psychology

Transportation mobility restrictions present important psychological issues for a number of special populations, such as single working mothers with children, people who are physically or developmentally disabled, and the elderly. Perhaps the most important problem area will be the needs and functioning of older adults, because of the growth in that sector of the population. In the USA in 1900, people aged 65 and over accounted for 4 percent of the population. By 1990, this percentage had grown to 12.5 percent, and it is estimated that by 2030 it will be over 20 percent.

Transportation problems for the elderly are exacerbated by suburbanization and by the vast majority of older adults being accustomed to driving automobiles. Decrements in visual, cognitive, and psychomotor skills that occur progressively with aging present societal challenges in scrutinizing the driving proficiency of older adults, for whom the loss of the driving license constitutes a major stressful life event (Sandeen 1997). Without transit alternatives, older people become a transportation-deprived group. Many older adults adjust progressively to the decline in driving skills by ceasing to drive at night or avoiding turns across oncoming traffic; but such adaptations do not suffice for long, leaving a great many elderly people unable to be licensed to drive.

Advanced technologies, such as automated roadways, are expected to ease transportation difficulties for the general population, but if automotive vehicles become significantly more cognitively complex to operate, this may be disadvantageous to the elderly. It can be expected that there will be substantial societal pressure to develop innovations in transportation to meet the needs of older adults who cannot drive and who do not have access to normal transit services. The development of flexible and perhaps individualized transportation services for the elderly to facilitate grocery shopping, receiving health care, and going to religious, recreation, and cultural events is an important societal agenda that beckons for psychological input.

9. *Adaptation: An Overarching Concept*

A core concept in the psychology of transportation is behavioral adaptation. Drivers and transport users will adjust to conditions in the transportation environment, whether that be in their overall commuting scheme, roadway or transit conditions, vehicle conditions, or changes in skill level or state of functioning. Thus, adaptation occurs at different levels of analysis, ranging from 'high-level' trip and travel-mode decisions to 'low-level' vehicle control behaviors that maintain safety margins. When behavioral adjustments involved in adaptation are unsuccessful in the long term, more substantial behavioral coping can be expected, as when someone changes jobs to reduce commuting strain or stops driving because of age-related driving skill decrements. Innovations in information technology, by decreasing the need for vehicle trips, may progressively facilitate adaptation to travel constraints imposed by congested roadways and to unwanted reductions in mobility for some populations.

See also: Sustainable Transportation; Transportation Geography; Transportation Planning; Transportation: Supply and Congestion

Bibliography

Altman I, Wohlwill J F, Everett P B 1981 *Transportation and Behavior*. Plenum, New York

Barjonet P 1997 Transport psychology in Europe: A general overview. In: Rothengatter T, Carbonell-Vaya E C (eds.) *Traffic and Transport Psychology: Theory and Application*. Pergamon, New York, Chap. 2, pp. 21–30

Brewer A M 2000 Road rage: what, who, when, where, and how? *Transport Reviews* **20**: 49–64

Grieco M, Pickup L, Whipp R 1989 *Gender, Transport, and Employment: The Impact of Travel Constraints*. Avebury, Aldershot, Hampshire, UK

Koslowsky M, Kluger A V, Reich M 1995 *Commuting Stress: Causes, Effects, and Methods of Coping*. Plenum, New York

Mokhtarian P L 1990 A typology of relationships between telecommunications and transportation. *Transportation Research-A* **24**: 231–42

Novaco R W 1991 Aggression on roadways. In: Baenninger R (ed.) *Targets of Violence and Aggression*. Elsevier, North Holland, The Netherlands, Chap. 7, pp. 253–326

Novaco R W, Stokols D, Milanesi L 1990 Objective and subjective dimensions of travel impedance as determinants of commuting stress. *American Journal of Community Psychology* **18**: 231–57

Novaco R W, Kliewer W, Broquet A 1991 Home environmental consequences of commute travel impedance. *American Journal of Community Psychology* **19**: 881–909

Pisarski A E 1987 *Commuting in America: National Report on Commuting Patterns and Trends*. Eno Foundation for Transportation, Westport, CT

Rosenbloom S 1992 Why working families need a car. In: Wachs M, Crawford M (eds.) *The Car and the City: The Automobile, the Built Environment, and Daily Urban Life*. University of Michigan Press, Ann Arbor, MI, Chap. 4, pp. 39–56

Rothengatter T, Vaya E C (eds.) 1997 *Traffic and Transport Psychology: Theory and Application*. Pergamon, New York

Sandeen B A 1997 Transportation experiences of suburban older adults: Implications of the loss of driver's license for psychological well-being, health, and mobility. Ph.D. thesis, University of California

Shinar D 1998 Aggressive driving: the contribution of the drivers and the situation. *Transportation Research part F* **1**: 137–60

Stokols D, Novaco R W 1981 Transportation and well-being. In: Altman I, Wohlwill J F, Everett P B (eds.) *Transportation and Behavior*. Plenum, New York, Chap. 4, pp. 85–130

Wachs M, Crawford M (eds.) 1992 *The Car and the City: The Automobile, the Built Environment, and Daily Urban Life*. University of Michigan Press, Ann Arbor, MI

R. W. Novaco

Transportation: Supply and Congestion

1. *Congestion*

Transportation facilities are prone to congestion because of peaks in travel demand, inflexibility of supply and inability to store transport services. Urban road traffic congestion and congestion in air travel impose significant economic costs. Traditional policies to alleviate congestion such as building new roads

have had at best, limited success. Congestion pricing has the greatest theoretical potential because it can affect all dimensions of travel behavior.

2. Introduction

Technological advances in transportation since the industrial revolution have spurred the growth and spread of cities, enabled people to live far from where they work, and made long-distance business and recreational travel commonplace. Yet growth in travel has been accompanied by increasing adverse side effects: noise, air pollution, accidents, and congestion—the subject of this article. To be sure, congestion is not a recent phenomenon. Two thousand years ago chariot riding was banned in Rome during peak hours because of traffic jams (Gibbs 1997). But the automobile and other motorized forms of transportation have increased the prevalence of congestion in surface travel, and congestion in air travel has become a serious problem.

The most widespread and costly form of congestion is travel delay. Delay on the roads is manifest as reduced speeds, and queues at intersections and entrances to freeways. Air travelers suffer delay in checking in, passing through security, boarding an aircraft, waiting to take off and land, clearing customs and immigration, and retrieving baggage at the destination. A fishing vessel on busy waters endures delay in procuring the catch when other vessels get in its way (Walters 1987). These examples feature delay systems in which users experience delays but eventually complete their trips. Delay systems differ from loss systems, which have a finite capacity to accommodate users and reject further demand once their capacity is reached so that service becomes unavailable. For example, a flight cannot be booked if all seats have been reserved. A given transportation facility can sometimes be a delay system and sometimes a loss system. A bus is a delay system when it stops for passengers and waits for them to board and alight and a loss system at full occupancy. And an airport is a delay system if arriving aircraft have to queue prior to landing, and a loss system if aircraft get diverted to another airport.

Crowding is a third form of congestion that is commonly experienced on buses, subways and waiting platforms, in using luggage compartments on trains and aircraft, etc. Delay, unavailability of service, and crowding are problems not only in transportation but also in facilities associated with travel including hotels, restaurants, sports stadiums, recreational areas, and telecommunications. These facilities will however not be treated explicitly here.

3. Why Does Congestion Occur?

Transportation facilities are prone to congestion because of three characteristics of travel demand and supply: demand varies over time, supply is relatively fixed over long time periods, and output is not storable.

3.1 Temporal Variation in Demand

Travel demand has pronounced temporal fluctuations. Systematic daily and weekly variations are driven by work and school schedules, and by the operating hours of businesses, shopping, and entertainment establishments. Annual variations in travel are induced by the timing and length of school holidays and religious festivals, by sporting event schedules, and by the seasonal nature of outdoor recreational activities. Travel demand also fluctuates irregularly due to sporting meets, fairs, and other 'special events.'

3.2 Rigidity of Transportation Supply

Transportation supply consists of infrastructure and mobile plant. Infrastructure is generally long-lived and very costly and time-consuming to alter. Mobile plant comes in smaller and cheaper units than infrastructure, and has a shorter lifetime, but it still lasts much longer than the time scale of demand fluctuations. The scope to alter the supply of vehicle services is limited by the mobility of the plant itself (little or none for subways and trams) and by the degree to which differences between regions and traveler groups in the timing of peak demands can be exploited.

3.3 Output is Not Storable

A traveler or freight shipment must be present to be transported. Thus, like other services such as haircuts and concerts, but unlike commodities such as grain and steel, production cannot be stored to smooth imbalances between supply and demand.

As a consequence of demand fluctuations, rigid supply and the impossibility of storage, the utilization rate of transportation facilities varies over time. Variations in utilization are magnified by intermittent reductions in capacity due to accidents, strikes, bad weather and so on. Because land is scarce and capacity is expensive, it would be prohibitively expensive to build enough capacity to prevent congestion at all times.

Congestion is exacerbated by two further considerations. First, travelers usually lack incentive to account for the congestion costs they impose on other users. Usage fees are not charged on most roads, and the costs of registration, licensing and insurance do not vary with distance traveled—let alone traffic volumes. Access fees are levied at facilities such as waterways and airports, but the fees are usually either fixed or

vary at best crudely with usage. Thus, congestion is a negative externality because its adverse side effects are not priced in markets. Second, transportation supply tends to lag growth in population and travel. Many countries experienced substantial increases in vehicle use per kilometer of road from 1977 to 1987 (Small and Gómez-Ibáñez 1999, Table 1). Air travel volume in passenger-kilometers grew at an average annual rate of 8.9 percent from 1960 to 1995 (ICAO 1999), spurred by airline deregulation in the US in the late 1970s and subsequently in Canada, Australia, New Zealand, and Europe. Airport planning authorities, meanwhile, failed to invest adequately in runway capacity and air traffic control, or to price usage of airports efficiently.

4. *The Costs of Congestion*

Congestion imposes a range of costs on travelers: increased fuel consumption and vehicle wear and tear; extra travel time; the inconvenience of using an alternative travel mode or rescheduling, combining or canceling trips; disutility from crowding; and (in the longer run) the costs of relocating jobs and residences. Delays are also costly for freight shippers and unreliable deliveries can upset just-in-time manufacturing and increase the need for inventories.

Because delays in passenger travel are a dominant component of the total cost of congestion it is important to measure accurately the cost of travel time—known as the Value of Time (VOT). VOT depends on mode of travel, trip purpose, time of day, household time pressures, and other individual and trip-specific factors. As for as automobile commuting trips, Small (1992, p. 44) concludes from an extensive literature survey that: '… a reasonable average value of time for the journey to work is 50 percent of the gross wage rate, while recognizing that it varies among different industrialized cities from perhaps 20 to 100 percent of the gross wage rate, and among population subgroups by even more.' Estimates of VOT for air travel are generally higher than for driving, in part because many air travelers are on business trips with tight schedules and in part because individuals with high VOT prefer to take fast modes of transport. There is limited evidence that VOT is higher for congested than uncongested travel (Calfee and Winston 1998).

Various estimates of the economy-wide costs of congestion have been compiled. For traffic congestion in the US Schrank and Lomax (1999) report a value of $72 billion for a set of 68 large urban areas in 1997. This figure includes time delay valued at $12 per hour (88 percent of the total) and extra fuel consumption. For Western Europe the European Commission in 1995 reported the costs of traffic congestion to be two percent of Gross Domestic Product (Prud'homme 1997)—more than double the equivalent US figure just mentioned. An extreme case is Bangkok, where congestion is alleged to consume 35 percent of gross annual output (Gibbs 1997). Aggregate statistics on the cost of air traffic congestion are patchy. In 1995, the US Federal Aviation Administration estimated that delays cost the industry approximately $2.5 billion per year in higher operating expenses. This excludes costs borne by passengers and freight shippers. In Europe a third of all flights are delayed by more than 15 minutes, at a cost to airlines and passengers of about $10 billion.

For several reasons aggregate statistics such as these should be interpreted with caution. First, they usually include only the more readily measured costs of congestion. Second, valuing time delay with a single number for VOT ignores differences across travelers as well as possible dependency on trip duration and the severity of the congestion encountered. Third, because eliminating congestion is not a practical option, measuring the total costs of congestion under actual travel conditions is not indicative of what might be gained from policies aimed at congestion relief (on this point see also Sect. 4).

These caveats notwithstanding it is instructive to compare the costs of congestion with the costs of other transportation externalities. Estimates from various countries reported in Greene et al. (1997) suggest that for road traffic, congestion costs are considerably smaller than accident costs, but at least as important as noise, air pollution and global warming costs. For air travel Levinson et al. (1998) estimate that noise is the largest cost, followed in order by congestion, air pollution, and accident costs. Thus, the costs of congestion appear to be appreciable both in absolute terms and relative to other externalities.

5. *Modeling Congestion in Transportation Facilities*

Transportation congestion modeling has been taken furthest for road traffic. Various approaches are described in Small (1992), Daganzo (1997), Hall (1999), and Lindsey and Verhoef (2000). Modeling the behavior of individual drivers is a daunting task not only because of their sheer number but also because drivers—being self-interested, fallible, and heterogeneous—do not behave like atoms according to universal and immutable physical laws. Computer simulation models have been developed that track the second-by-second motion of thousands of vehicles on complex traffic networks (Howard 1997). Useful insights into the economics of traffic congestion can nevertheless be obtained by using conventional supply–demand diagrams in which the dynamics of traffic are ignored and vehicles are treated as a homogeneous continuum.

Consider Fig. 1, but ignore all its details for now—which will be covered later in Sects. 4 and 5. The

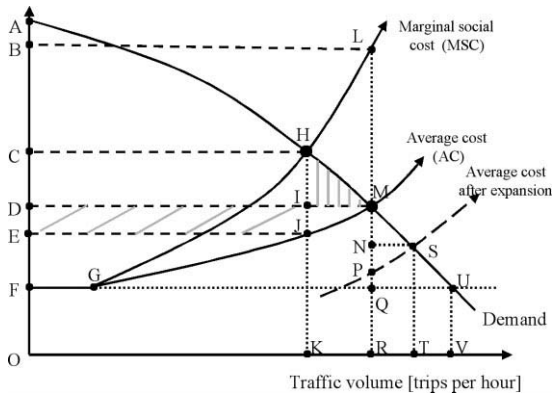

Figure 1
Trip costs, equilibrium and optimal usage in a delay system

Average Cost (*AC*) curve passing through points *F, G, J* and *M* depicts the cost borne by each driver to take a particular trip as a function of the traffic volume; i.e. the number of vehicles that make the trip per hour. Average cost includes vehicle operating cost plus travel time, which is assumed to be valued equally by all drivers. At low volumes vehicles are far enough apart that they can proceed under free-flow conditions, so that the trip cost curve is horizontal between *F* and *G*. At higher volumes drivers are forced to slow down to maintain a safe following distance. Travel time rises and *AC* slopes upwards. The shape of *AC* depends on road design, speed limits, vehicle size and handling characteristics, weather conditions, and driver reaction times and preferences for risk-taking. Due to changes in these factors over time, driving speeds have increased and following distances have decreased, thereby boosting effective road capacity and shifting the *AC* curve down.

The *Marginal Social Cost* (*MSC*) curve in Fig. 1, passing through *F, G, H* and *L*, is derived by adding to *AC* the cost of the congestion that each driver imposes on others. The *MSC* curve coincides with *AC* over the interval from *F* to *G*, and lies above *AC* elsewhere. To determine how many trips are taken, a demand curve can be added as shown by the downward-sloping curve passing through *A, H, M, S* and *U*. In the absence of traffic control or other means of regulation, equilibrium is established at point *M* where the marginal willingness to pay for a trip as measured by the demand curve matches the cost, *AC*. The unregulated equilibrium number of trips is *OR*. By contrast, the socially optimal number of trips, *OK*, is found at point *H* where the demand curve intersects *MSC*. Fewer trips per hour are taken than in the unregulated equilibrium, although travel is still congested because trip cost, *JK*, exceeds free-flow cost, *OF*. One way to support the optimum is to impose a toll equal to *HJ*: the difference between *MSC* and *AC* with traffic volume at *OK*. This is the basis behind congestion pricing, to be discussed in Sect. 5.2.

At the unregulated equilibrium the congestion cost imposed by each driver is *LM*. Various studies have estimated this quantity empirically. Some find it to be about equal to the average cost that drivers bear themselves: *MR* (Small and Gómez-Ibáñez 1999 p. 1942). Other studies estimate *LM* to be several times as big as *MR* during peak periods in large cities. To get a figure for aggregate congestion costs it is tempting simply to multiply *LM* by the number of trips, *OR*, to get area *LMDB*. But this is inappropriate since the number of excessive trips taken in equilibrium relative to the social optimum is only *KR*. Similarly, it does not make sense to evaluate the congestion cost at the optimum, *HJ*, and multiply it by *OK* to get area *HJEC*. Yet as Prud'homme (1997) points out, some studies have identified either *LMDB* or *HJEC* as the total cost of congestion.

A more practically useful measure of congestion costs is the difference between the net social benefits of travel at the optimum (area *AHJE*) and at the unregulated equilibrium (area *AMD*). This difference can be decomposed into the reduction in costs on the *OK* trips that continue to be taken at the optimum (diagonally shaded area *IJED*) minus the loss of benefits on the *KR* trips that are no longer taken (vertically shaded area *HMI*). The difference in net social benefits can also be determined by computing the net social cost of the *KR* excessive trips taken in equilibrium. This quantity is equal to area *HLM*, which must therefore equal *IJED–HMI*.

Note that the cost of travel under free-flow conditions plays no role in the calculations. One could of course calculate the difference in cost between equilibrium and free-flow conditions, *MQ*, and multiply it by *OR* to get area *MQFD*. But this would be relevant only if it were somehow possible to maintain free-flow conditions at any traffic volume. Traffic volume would then increase from *OR* to *OV*, and a better measure of the gains would be the increase in net social benefits: area *MUFD*.

Supply–demand diagrams such as Fig. 1 are applicable not only to automobile traffic but also to other delay systems such as airport runways, bus service, waterways, and pedestrian walkways. In the case of loss systems the *AC* curve is approximately horizontal up to capacity flows and it is therefore optimal to use the facility to capacity while avoiding queuing or other wasteful forms of rationing access. In practice, demand cannot be predicted perfectly and a (probabilistic) tradeoff exists between unused capacity and congestion. Airlines deal with this tradeoff by using sophisticated yield management software that set fares according to the average fraction of seats that are occupied on a given flight.

Though useful, the simple supply–demand approach does abstract from various real-world features

of transportation congestion. First, roads, airports and other facilities are components of transportation networks. Costs at one facility can depend on usage at other facilities because of conflicts between traffic flows, spillovers of queues and so on. Second, users differ in the congestion they create. Traffic engineers make allowance for differences between road vehicles in size, speed, and maneuverability by using passenger car equivalents that depend on grade and other road characteristics. Third, the static supply–demand approach ignores fluctuations in travel demand over time and does not address trip-timing decisions. Queuing theory is often useful for handling these aspects (Newell 1982).

6. *Ways to Alleviate Congestion*

Many policies have been adopted to relieve congestion. Some—such as land-use controls—have additional objectives, whereas policies such as road pricing are targeted at congestion. A comprehensive review of policies in urban transportation is found in Small and Gómez-Ibáñez (1999, Sects. 2.2–2.5). Discussion will be limited here to three policy instruments: capacity investment, congestion pricing, and information technology.

6.1 *Investment*

The traditional approach to congestion relief is to expand capacity. Investment shifts the *AC* curve in Fig. 1 down, as shown by the dashed line passing through points *P* and *S*. A new unregulated equilibrium is established at *S*. Two benefits result: a reduction in travel costs for the *OR* pre-existing trips; and the benefit to travelers from the *RT* new trips. But because the new trips contribute to congestion, travel costs decline by only *MN* rather than by *MP*, and this partly undoes the cost savings to previous users. Generation of new trips is a manifestation of so-called latent (hidden) demand that was not apparent prior to the investment. Though the increase in usage can be appreciable if the demand curve is relatively flat, the effect of latent demand has frequently been overlooked in engineering and cost-benefit studies of investment. Even worse than latent demand, capacity expansion on one link of a transport network can induce shifts of demand to more congestion-prone routes or travel modes that actually raise overall travel costs (Arnott and Small 1994).

Unlike the unregulated equilibrium, investment always yields a positive gross benefit for the social optimum. The effects can be assessed by tracing the rightward shift of the *MSC* curve (not shown in Fig. 1). Depending on the strength of latent demand, the benefits may be greater or smaller than for the unregulated equilibrium. More generally, the benefits of investment, and hence optimal capacity too, depend on how traffic is regulated and priced.

6.2 *Congestion Pricing*

Economists have promoted congestion pricing since the 1950s. It makes use of the price mechanism, with all its advantages of decentralized decision-making. And, unlike other policies, congestion pricing restrains latent demand. The biggest potential gains from congestion pricing in transportation appear to be for road traffic and airport runways.

6.2.1 Road pricing. Road pricing technology has come a long way since drivers were obliged to stop and pay cash at tollbooths. Thanks to advances in automatic vehicle identification and electronic toll collection, tolls can be levied anonymously using prepaid smartcards while vehicles maintain highway speeds. The biggest virtue of road pricing is that it can influence all dimensions of travelers' choice. It is also consistent with the 'user pay principal.'

The optimal toll in the static supply–demand framework of Fig. 1 is *HJ*. Imposition of this toll raises travel cost by *HI* relative to the no-toll unregulated equilibrium. That road pricing tends to make drivers worse off is one of the reasons for longstanding opposition to it. Another is that road pricing is often seen as 'just another tax.' There is growing recognition that to generate political support for road pricing, toll revenues must be used to benefit travelers without undoing incentives for them to curb their travel. Possibilities include improving public transit, reducing gasoline taxes, and investing in roads themselves. New self-financed private toll roads are showing promise as a market-oriented approach to alleviating congestion (Gómez-Ibáñez and Meyer 1993).

The simple supply-demand approach of course abstracts from various real-world complications in the implementation of congestion pricing, including setting tolls on a network, varying tolls over time, charging for parking, and avoiding harm to lower-income drivers. These and other aspects are discussed in Button and Verhoef (1998) and McDonald et al. (1999).

6.2.2 Pricing airport runways. Traditional means of allocating runway space are a far cry from congestion pricing. Access to runways has been determined by scheduling committees that may be dominated by representatives of major airlines operating at the airport. And landing fees have been based on aircraft weight, which reflects congestion cost only crudely. Congestion pricing of runways would be administratively easier to implement than road pricing because takeoff and landings are recorded at airports

anyway, and since the transactions costs for a flight are spread over many travelers.

The virtues of congestion pricing of runways were recognized in the late 1960s. Pricing encourages airlines to schedule flights off peak, to use lower-cost airports, and to deploy larger aircraft, which create less delay per passenger. The potential gains are large. Morrison and Winston (1989) estimate that optimal pricing at US airports would reduce delays by up to 50 percent, and yield nearly $4 billion in net benefits. An alternative to charging for every use of runways is to auction tradable take-off and landing slots. Slot trading was introduced in 1986 at Washington National, Chicago O'Hare, and New York's La Guardia and Kennedy airports, and has worked well. Runway pricing and slot trading, as well as pricing of airport terminal services are described in Doganis (1992).

6.3 Information Technology

Development of Intelligent Transportation Systems (ITS) began in the 1970s, and expensive research programs are under way. One of the long-term goals of ITS is to assemble information about travel conditions and make it available to travelers at public places, by phone or Internet, or directly to vehicles equipped with on-board computers and Global Positioning Systems receivers. Computer simulations suggest that information can help drivers avoid hot spots of congestion by switching routes, changing departure times, and canceling trips. Travelers without access to information may also benefit from the adjustments made by informed travelers.

Despite these encouraging research findings, the potential of ITS is limited by the fact that they do not make drivers pay for the congestion they create. Indeed, provision of information can make congestion worse if it induces drivers to overreact; e.g. by switching *en masse* to Route 2 after learning about an accident on Route 1 (Ben-Akiva et al. 1991). And, similar to the effects of capacity expansion, congestion can be exacerbated if information induces people to travel more, or to concentrate their travel on congestion-prone routes and at peak times. These drawbacks can in principle be avoided by implementing road pricing in tandem with ITS.

6.4 General Policy Conclusions

From a theoretical perspective congestion pricing is the most effective means of controlling congestion because it influences all travel choices and also contains latent demand. But for various reasons—including the cost of pricing all links on travel networks, equity concerns and the need to maintain safety—nonprice constraints on travel such as air traffic control and regulations on the movement of heavy road vehicles will persist. The battle against traffic snarls, delays in the air, and congestion on other transport modes must continue to be waged on several fronts.

7. Looking Ahead

Despite the hazards of forecasting it is tempting to speculate about how congestion in transportation will evolve in the years to come. Consider first traffic congestion in the US, which is costly in an absolute sense and receives regular attention in the press. Aggregate congestion indexes suggest that congestion is getting worse in major urban areas. But average speeds and commuting times have remained fairly constant over the last 25 years. After growing for decades, auto ownership per capita is near saturation. And average commuter vehicle occupancy, which fell from 1.18 in 1970 to 1.09 in 1990, cannot fall much further (Schrank and Lomax 1999). These and other factors have led analysts such as Meyer (1999) to conclude that radical policies to affect sweeping changes in automobile travel patterns are not warranted.

Technological innovations such as air conditioning, car stereo systems, books on tape and cellular phones have made driving more pleasant and allow people to make productive use of time in the car. Similarly, laptops, cellular phones and executive lounges at airports make delays in air travel less onerous. And people can adapt to congestion by adjusting departure times, destinations, residential location and other choices. As a result, their tolerance for delay increases (Calfee and Winston 1998). Congestion costs are also held in check by the trend toward flexible work times, late night and weekend shopping, and extended hours of other services. And, most fundamentally, the costs of congestion are borne largely by travelers themselves. Congestion is a self-limiting phenomenon in the sense that the private cost of travel can never exceed the private benefit derived from it.

Travel is governed by resource constraints and other influences that vary greatly from region to region, and evolve over time. Two regularities in travel data have nevertheless been documented. One is that the amount of time spent traveling per person has remained relatively fixed at 1–1.5 hours per day. The second is that people spend a relatively constant fraction of their income on travel. These observations led to the constant travel time budget hypothesis and constant travel income budget hypothesis (for a summary see Schafer and Victor 1997). The two hypotheses are not logically inconsistent because as incomes rise people can switch to faster and more expensive modes of travel. To the extent that the hypotheses are valid they are useful for forecasting general trends in travel and congestion. Rising living standards will induce people to travel further each year. Faster economic growth in

developing countries will mean faster expansion in travel there. Shifts toward faster modes of travel suggest that congestion in air travel will grow most quickly. Schafer and Victor (1997) forecast that by 2050 high-speed transport (air and high-speed rail) will account for about 50 percent of total distance traveled.

To what extent might new technology alleviate congestion by reducing travel demand? Attention has been focused on teleconferencing, teleshopping, FAX, electronic mail, the Internet and other telecommunications technology. Spread of computers into homes has made telecommuting perhaps the most promising application. Nevertheless, it has some limitations (Mokhtarian 1997). People may value the journey to work as a buffer between their home and work roles, desire interaction at the workplace, or fear being overlooked for promotion if they stay at home. The option of telecommuting may encourage people to live in more remote locations, so that when they do commute they have further to go. Part of the time freed up from commuting may be used to make other trips, as the constant travel time budget hypothesis predicts. And any reductions in travel by telecommuters may be partially offset by latent demand from the rest of the population. Still, by reducing peak-period travel, telecommuting may alleviate congestion and permit savings in expenditure on infrastructure. And even if the increased use of telecommunications as a whole does not reduce travel, it may yield benefits in increased work productivity and greater personal flexibility. Perhaps the safest prediction is that congestion and efforts to alleviate it will persist for the foreseeable future.

See also: Ecotourism; Regulation Theory in Geography; Spatial Pattern, Analysis of; Transportation Geography; Transportation Planning; Urban Geography; Urban System in Geography

Bibliography

Arnott R, Small K A 1994 The economics of traffic congestion. *American Scientist* **82** September–October: 446–55

Ben-Akiva M, de Palma A, Kaysi I 1991 Dynamic network models and driver information systems. *Transportation Research* **25A**(5): 251–66

Button K J, Verhoef E T (eds.) 1998 *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*. Edward Elgar, Cheltenham, UK

Calfee J, Winston C 1998 The value of automobile travel time: Implications for congestion policy. *Journal of Public Economics* **69**: 83–102

Daganzo C F 1997 *Fundamentals of Transportation and Traffic Operations*. Pergamon, New York

Doganis R 1992 *The Airport Business*. Routledge, London and New York

Gibbs W W 1997 Transportation's perennial problems. *Scientific American* **277**(4): 54–7

Gómez-Ibáñez J A, Meyer J R 1993 *Going Private: The International Experience with Transport Privatization*. The Brookings Institution, Washington, C

Greene D L, Jones D W, Delucchi M A (eds.) 1997 *The Full Costs and Benefits of Transportation*. Springer-Verlag, Berlin

Hall W R (ed.) 1999 *Handbook of Transportation Science*. International Series in Operations Research and Management. Kluwer Academic Publishers, Boston, MA

Howard K R 1997 Unjamming traffic with computers. *Scientific American* **277**(4): 86–8

International Civil Aviation Organization 1999 Outlook for air transport to the year 2005, Circular 270-AT/111. www.icao.org

Levinson D M, Gillen D, Kanafani A 1998 The social costs of intercity transportation: a review and comparison of air and highway. *Transport Reviews* **18**(3): 215–40

Lindsey R, Verhoef E T 2000 Congestion modelling. In: Hensher D A, Button K J (eds.) *Handbook of Transport Modelling*. Elsevier Science, Oxford, UK, Vol. 1

McDonald J F, d'Ouville E L, Liu L N 1999 *Economics of Urban Highway Congestion and Pricing*. Kluwer Academic Publishers, Boston, MA

Meyer M D 1999 Demand management as an element of transportation policy: using carrots and sticks to influence travel behavior. *Transportation Research* **33**: 575–99

Mokhtarian P L 1997 Now that travel can be virtual, will congestion virtually disappear? *Scientific American* **277**(4): 93

Morrison S A, Winston C 1989 Enhancing the performance of the deregulated air transportation system. *Brookings Papers on Economic Activity, Microeconomics* : 61–123

Newell G F 1982 *Applications of Queuing Theory*. Chapman and Hall, London

Prud'homme R 1997 On transport-related externalities. *Jornadas de Finanzas Publicas* **30**: 139–56

Schafer A, Victor D 1997 The past and future of global mobility. *Scientific American* **277**(4): 58–61

Schrank D, Lomax T 1999 The 1999 Annual Mobility Report Information For Urban America, Texas Transportation Institute, Texas A&M University System, College Station, TX http://mobility.tamu.edu

Small K A 1992 Urban transportation economics. In: Lesourne J, Sonnenschein H (eds.) *Fundamentals of Pure and Applied Economics*. Harwood Academic Publishers, Chur, Switzerland

Small K A, Gómez-Ibáñez J A 1999 Urban transportation. In: Cheshire P, Mills E S (eds.) *Handbook of Regional and Urban Economics 3*. North-Holland, Amsterdam, pp. 1937–99

Walters A A 1987 Transportation. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave: A Dictionary of Economics*. Macmillan, New York, Vol. 4, pp. 695–701

A. de Palma and R. Lindsey

Transsexuality, Transvestism, and Transgender

'Transexuality' or 'transsexuality' (with two 's's'—for an explanation of these orthographic differences, see below) and 'transgender' are terms often used inte-

rchangeably in contemporary political activism and social scientific writing, as is the associated term 'transvestism.' Despite the overlap, these terms have different histories and a complex relationship with one another.

The donning of garments of another gender (so-called 'cross-dressing') has a long recorded history. Only in the late nineteenth and early twentieth centuries, however, did cross-dressing come to be viewed specifically as a psychosexual disorder. At this point, cross-dressing became 'transvestism,' a term first used by the German sexologist and sexual rights activist Magnus Hirschfeld in his study, *Die Transvestiten* (*Transvestites*), published in 1910. Hirschfeld coined the term from the Latin *trans* ('across') and *vestis* ('dress'), to refer to individuals whose principle source of erotic pleasure arose from dressing in the clothes of the opposite sex. Hirschfeld's understanding of cross-dressing and his term 'transvestism' have been debated in the sexological and psychoanalytic literature ever since his study appeared (the English sexologist Havelock Ellis, for example, argued that Hirschfeld focused too narrowly on clothing and should also have considered the appeal of the social role of the opposite sex. He suggested the term 'eonism,' which he invented, using the name of the eighteenth century cross-dresser, the Chevalier d'Éon; see Bullough and Bullough 1993, pp. 203–25). However, contemporary psychiatric discourses have largely followed Hirschfeld in characterizing transvestism in terms of (primarily heterosexual male) erotic cross-dressing.

In contemporary psychiatric classification, transvestism is distinguished from transexualism, which is regarded as a separate pathology. Whereas transvestism is considered a psychosexual disorder characterized by an erotic focus on clothing, transexuals are defined as individuals who have a 'gender identity' that is at odds with their assigned gender. In other words, transexuals are considered to be individuals whose core identification of themselves as male or female is at odds with the gender by which they are socially desginated. Transexualism involves a wide range of psychiatric and medical interventions, including the surgical alteration of genitals and secondary sex characteristics, and hormonal supplements to attain the physical appearance of the desired sex/gender.

The development of transvestism and transexualism as medical categories must be understood as an outgrowth of the social and cultural discourses that developed in the late nineteenth and early twentieth centuries, when the dominant paradigm for understanding and regulating gender and sexuality shifted from the various Christian churches to the state and to various branches of the developing medical sciences. Modern discourses on transexualism have their historical roots in nineteenth century debates about the management of hermaphrodism, which is the medical term for a range of conditions which give rise to non-normative genitals in a small segment of the population (the term 'intersexuality' has recently been coined by intersex activists to replace 'hermaphrodism,' which they regard as saturated with medical stigma). Michel Foucault has argued that the centuries-old understanding that intersexed individuals had two sexes came to be replaced during the nineteeth century with the view that they must have a single, 'true' sex; one that once it became identified by medical science and certified by juridical institutions could never be changed. This concern to 'strip the [hermaphrodite's] body of its anatomical deceptions' (Foucault 1980, p. viii) and discover its truth was part of a much broader move that occurred during this period towards the medical and juridical management of the sexual behavior and bodies of all individuals, especially those who engaged in nonprocreative sexual acts, or who transgressed gendered boundaries.

Current understandings of categories like hermaphrodism/intersex, homosexuality, transvestism, and transexualism all arose during this historical period, and are linked to one another and to the medical and legal classifications that created them as social categories. Transexualism itself, however, is essentially a mid-twentieth century phenomenon. It arose together with the modern, surgical techniques that made it possible. As such, 'transexuality' only became available as a medical category, a surgical possibility, and a personal identity around the middle of the twentieth century.

While there were attempts at surgical changes to sex (from male to female) prior to the 1950s (the first attempt to create an artificial vagina through surgery was in 1904), only 31 cases of surgical alteration or change of sex are known prior to the 1950s. It is likely that these were mainly early attempts to intervene in cases of intersexuality or hermaphrodism in the nascent field of plastic surgery. Sensationalist media reports of early intersexual surgeries provided the impetus for many individuals who would be described as male-to-female transexual in contemporary medical terms to demand surgical reassignment. The case of Christine Jorgensen, who received *sex reassignment surgery* (SRS) in Denmark in 1952, brought this phenomenon to the attention of the public and the medical community in the United States. However, it was mainly through the efforts of a New York endocrinologist, Harry Benjamin, that the term 'transsexual' (spelled with two 's's') gained force as the term to describe the growing number of people appealing to surgeons and doctors for SRS following the publicity surrounding the Jorgensen case and others that followed. The term 'transexualism' was now used to distinguish these people from transvestites who cross-dressed but did not want SRS. It was also through Benjamin's efforts that the first 'gender identity clinic' opened in the US at Johns Hopkins in 1965. The first SRS was completed there in 1966, the same year

Benjamin's classic in the field, *The Transsexual Phenomenon*, was published.

In the late twentieth century, some of the categories of 'dis-order' that were part of what Foucault calls the 'deployment of sexuality,' such as that of homosexuality, have, to an extent, extricated themselves from purely medical discourses. Transexualism, however, because of its links to surgery, is by definition a 'condition' that requires a close relationship between a transexual and the medical profession. SRS requires skilled surgeons, endocrinologists, urologists, and others, and the path to this surgery is controlled by psychiatrists and counselors. Furthermore, SRS is still regarded as a cure which aims at managing a pathology. Access to this cure, though, is predicated on an appropriate diagnosis. However, the history of the development of such diagnoses has not been an even or unproblematic development.

Within the medical community, the period 1953–65 was characterized by resistance both to the adoption of transexualism as a diagnostic category and to performing SRS. However, by the 1970s, an entire subfield of specialty had developed in the area, along with new terminology. The term 'gender dysphoria' (and the associated 'gender identity disorder') came increasingly to dominate the literature on transexualism and transvestism, to a certain extent incorporating and replacing these latter terms. The term was partly an attempt to include a growing number of phenomena that psychiatrists and doctors were seeing in patients who did not conform to the classic descriptions of transexualism, in particular children who displayed non-normative gender behaviors and individuals who claimed a nonassigned gender identity but did not desire surgical transition. Indeed, Norman Fisk, who coined the term, saw 'gender dysphoria' as a liberalizing concept, allowing surgery to more people (Fisk 1974). By 1979, gender dysphoria was the primary working concept of those in the field, although 'transsexual' was, and still is, used as a diagnostic and descriptive category.

King (1993, p. 63) notes that there are several consequences of this change in terminology, the most important being that it shifts the focus from the individual to a condition and invokes a disease model. In doing so, the concept of gender dysphoria reaffirms medical authority. It also widens the scope of diagnosable people to include the physically intersexed (hermaphrodites) and even gay men and lesbians. The definition is so broad that 'indeed possibly all of us suffer from it in a mild form' (King 1993, p. 64).

Another consequence of the adoption of the term 'gender dysphoria' is that it encourages the idea that such a disorder existed long before it was 'discovered' and named by doctors and psychiatrists. A quotation from a 1986 review of the collaboration between various branches of the medical establishment is indicative: '[t]he concept of gender dysphoria and its specific forms have been known since antiquity and have been described in the classic literature from Herodotus to Shakespeare' (Pauly and Edgerton 1986, pp. 316–17).

Despite the widening use of 'gender dysphoria' as a trans-historical category, transexualism itself, as noted above, is a medical category of the latter half of the twentieth century. However, until 1980, when transexualism (and the associated category of 'gender identity disorder') was included in the *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn., (hereafter *DSM-III*), and when the Harry Benjamin International Gender Dysphoria Association published its document on Standards of Care (Harry Benjamin International Gender Dysphoria Association 1985 [1980]) for treating transexuals and 'gender dysphoric' persons, psychiatrists had no standardized criteria by which to judge whether a person was a 'good candidate' for SRS. Hence, they did so on the basis of a variety of fluctuating guidelines set by a variety of writers in the field.

The basic preconditions for SRS that became established in 1980 are as follows: a person seeking SRS should, prior to surgery, have the recommendation of at least two psychologists/psychiatrists; have been on hormonal treatment for periods varying from six months to one year; and have lived exclusively in the role of the chosen gender for periods varying from one to two years. Alongside this are other factors which are considered to make a 'good candidate,' including lack of sexual activity and no history of prostitution.

The publicity surrounding the Jorgensen case led many people to identify themselves as transexuals and seek SRS. This is indicative of one of the most striking aspects of transexualism in psychiatric terms: virtually all transexuals are 'self-diagnosed.' Transexuals challenge the medical establishment at the very root of its claims to authority, the privilege of diagnosis. This has set into motion a chain of related themes: distrust of transexual self-reports of their conditions and an emphasis on third party confirmation, attempts to come up with a stable list of criteria by which access to SRS can be regulated, and a push toward developing a psychiatric diagnosis for transexualism (or the broader term, gender identity disorders) that does not rely on self-reports, but on perceived objective criteria.

Despite the consistencies and patterns that have emerged, however, the history of the establishment of a stable diagnosis for transexualism and its related terms is neither simple, nor even. Some writers have denied that there is a diagnostic category at all, or that transexualism should not be a diagnosis but rather is a 'wish' that is a feature of other, underlying psychopathologies. The self-diagnosis of transexuals complicates the field. Benjamin wrote in 1966 that 'it is true that the request for a conversion operation is typical only for the transsexual and can actually serve as a definition' (1966, p. 21). However, this definition came increasingly under scrutiny as follow-up studies

in the 1970s and 1980s showed a significant (between 5 and 15 percent) 'dissatisfaction' with or 'negative outcomes' of surgery. There are a number of factors here, but significantly, medical practioners associated dissatisfaction primarily with giving the 'wrong' people SRS. In other words, clinicians no longer regarded requests for SRS as a definition or a diagnosis of transexualism. In fact, they tended to treat all such requests with suspicion, potentially as expressions of psychosis or even schizophrenia.

It was only in 1980, in the third edition of what is often referred to as the 'psychiatrists' bible,' the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), that transexualism was included as a mental disorder (transvestism was included from the original publication of the *DSM* in 1952). Some writers link the formal inclusion of transexuality or gender identity disorder in *DSM-III* to the exclusion of homosexuality from the *DSM-II* in 1973 after pressure brought to bear on the American Psychiatric Association (APA) by gay and lesbian activists. Scholars have noted that the work and writing of some APA members (particularly those who opposed the deletion of homosexuality) indicates the extent to which the diagnosis of gender identity disorder has become a justification for psychiatric interventions in cases of young children who are seen to be in 'danger' of becoming homosexual. Since 1980, two more editions have been prepared: the revised third edition in 1987, and the fourth edition in 1994. An analysis of the shifting categorization of transexuality and transvestism over these three editions complicates claims that gender dysphoria or gender identity disorder is a transhistorical, cross-cultural phenomenon with a certain universal stability and truth. The changes across three editions within 14 years indicate a far less stable set of concepts which are malleable, constructed, and negotiable.

This movement toward categorization and diagnostic criteria should not be seen to represent a monolithic set of attitudes and practices. Instead, it should be understood as the outgrowth of attitudes towards, and practices around, psychiatric diagnosis, relationships with patients, and the production of data and empirical evidence/knowledge in a field of power relations. This field of power has been generated by, and in turn has generated, a whole subdiscipline of medicine with a coalition of clinicians from numerous disparate fields coming together in a new category of theory and practice. Over the past 40 years, these clinicians have debated, produced, and to an extent consolidated a series of meanings and practices that have had specific, though varying, effects on the lives of individual people. Alongside the development of diagnostic criteria, there has also been a concern with follow-up and assessment, which has been central to generating the data upon which diagnostic criteria are based. In addition, clinicians implemented 'charm schools,' which train successful (male-to-female) candidates in culturally-specific gender behaviors and ideals.

Critics of such 'charm schools' have observed that they are almost entirely focused on male-to-female transexuals. This fact was seized upon by an early and still-influential line of radical feminist thought which argues that transexualism 'constitutes a sociopolitical program that is undercutting the movement to eradicate sex-role stereotyping and oppression' (Raymond 1994, p. xxi). In an era when women are increasingly rejecting rigid sex roles, this criticism maintains, patriarchal social structures are responding by 'molding and mutilating *male* flesh, but for the purpose of *constructing women*' (Raymond 1980, p. xvi, emphasis in original). From this point of view, individuals identifying as transexual need consciousness raising, not surgery, which these critics regard as socially sanctioned mutilation that should be outlawed.

That most medical attention has focused on managing male-to-female transexuality also highlights the broader issue of the category 'transexual.' *DSM* claims there are larger numbers of male-to-female transexuals than female-to-males. Many female-to-males/transexual men dispute this, and argue that medical discourses and statistics obscure a large population of transexual men. These men, like a number of scholars, give a variety of reasons for the medical concern with male-to-female transexuality, including the cost and unsatisfactory outcomes of surgeries for transexual men, the statistically lower socioeconomic position of female-born people, the relative ease of passing as a man with only hormonal interventions, cultural unease with the possibility of surgically constructing men (Judith Halberstam [1994, p. 215] quips 'If anyone could have [a penis], who would want one?'), and a broader social anxiety surrounding femininity in male-born people.

A significant development in conceptualizations about transexuality has been the development during the 1990s of the term 'transgender.' The phrase, adopted from a usage coined by Virginia Prince (1971), originally signified a person who chose to live full time in a nonassigned gender without surgical or, in some cases, without hormonal interventions. While 'transgender' is still often opposed to 'transexual' in this sense, it has become more frequently used as a collective term to describe a range of disparate groups which had previously been seen as distinct: transexuals, transvestites or crossdressers, drag queens, and even intersex people. Other terms, such as Richard Ekins's (1997) 'male-femaling' or Judith Halberstam's (1998) 'female masculinity,' have been proposed in the social science literature as a way of analyzing this range of people who cross or transgress gendered boundaries. However, 'transgender' has increasingly become the accepted generic term.

The category 'transgender' arose primarily in the northeastern and western United States (New York and California) in the early 1990s as a way of achieving

a diverse set of political, social, and cultural goals: the depathologization of gender difference in psychiatric and popular discourses, the recognition of gender variant people as a political constituency and a social group, and the achievement of social justice around such issues as exposure to violence and discrimination. In the brief years since then, 'transgender' has become ubiquitous in progressive community-based organizations, identity-based political movements, popular media accounts, international human rights discourses, academic debates, cross-cultural descriptions of gender difference, and is even finding its way into medical discourses, the very discourses that 'transgender' has been opposed to. Like transgender, 'transexual' with a one 's' spelling (the usage we adopt here), was coined by activists in opposition to medical categories, as in the New York-based direct action group 'Transexual Menace.' Other linguistic innovations abound, including formulations like 'transmen' or 'transwomen,' 'transcommunities' or 'person of transgender/transexual experience,' all of which attempt to describe phenomena without being constrained by standard medical terms.

The collectivity of 'transgender' also takes into account complexities of lived identities and experiences in a way that medical categories such as 'transsexual' or 'transvestite' do not. While *DSM* diagnostic criteria, for example, insist that transvestism is particular to heterosexual men, in fact self-identified heterosexual women, lesbians, and gay men also engage in erotic cross-dressing. Likewise, the line between cross-dressers and transexuals is not clearly drawn, nor is the line between butch lesbians and transmen. A cross-dresser or transvestite, who in psychiatric diagnostic criteria is a heterosexual man who cross-dresses for erotic pleasure, may also take hormones, and in some cases, transition to live full-time as a woman. 'Transgender' is useful, therefore, in bringing together a range of individuals who share an experience of gender variance.

However, 'transgender' as a category is not without its complexities either. While some activists use it in the collective sense described above which includes postoperative transexuals, others insist on the narrower definition which includes only those who choose not to have surgery. Further, while some self-identified transgender activists profess a desire to remove 'gender identity disorder' from *DSM* in an attempt to depathologize gender variance, others have demanded that some diagnostic criteria remain in place in order to allow for insurance reimbursement of the costly surgeries involved in sex reassignment. Similarly, many of those who are grouped under 'transgender' as a category are opposed to the inclusion of others, claiming that their particular experiences are diluted or effaced by their inclusion in such a broad category. Further, some transexual men argue that transgender privileges a male-to-female model which elides the particular experiences of transexual men. Most interestingly, medical professionals themselves are beginning to use the term, at least partly in deference to the growth of an active and vocal activism which seeks to overturn the power of medical professionals in the field. Indeed, 'transgender' works in some ways like the medical concept 'gender dysphoria' in that it groups together a wide array of individuals who may not necessarily see themselves as being part of the same category. The important distinction, however, is that as a community-based term, 'transgender' contains none of the implications of pathology apparent in formulations such as 'gender dysphoria' or 'gender identity disorder.' 'Transgender' then is characterized as much by disputes and contradictions as it is by its efficacy in highlighting issues of social justice in the fields of politics, medicine, law, and social service provision.

A related point of complexity is the insistence in many activist contexts on the distinction between 'transgender' issues on the one hand and 'gay and lesbian' issues on the other. In contemporary US politics and theory, for example, gay and lesbian identities are seen to be rooted in sexual orientation, whereas transgender identities are understood to be rooted primarily in gender identity. This has been the source of some tension in contemporary sexual and gender politics, and some gay and lesbian organizations resist the inclusion of transgender issues in their political and social activism. However, as some transgender activists have pointed out, much of the discrimination and violence experienced by gay- and lesbian-identified people is directed at a perceived gender transgression. These activists often prefer to use constructions like 'gender variance' or 'gender difference' to include both transgender-identified and gay- and lesbian-identified people in order to stress these connections and to avoid the kinds of essentializing identity-based discourses which often characterize progressive sexual and gender politics.

See also: Fashion, Sociology of; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Foucault, Michel (1926–84); Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender Differences in Personality and Social Behavior; Gender Identity Disorders; Gender Ideology: Cross-cultural Aspects; Heterosexism and Homophobia; Masculinities and Femininities; Queer Theory; Rationality and Feminist Thought; Regulation: Sexual Behavior; Sex-role Development and Education; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexual Preference: Genetic Aspects; Sexuality and Gender; Sexuality and Geography

Bibliography

American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1987 *Diagnostic and Statistical Manual of Mental Disorders-DSM-III-R*, 3rd rev. edn. American Psychiatric Association, Washington, DC

American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders-DSM-IV*, 4th edn. American Psychiatric Association, Washington, DC

Benjamin H 1966 *The Transsexual Phenomenon*. The Julian Press, Inc., New York

Bullough V L, Bullough B 1993 *Cross Dressing, Sex, and Gender*. University of Pennsylvania Press, Philadelphia

Ekins R 1997 *Male Femaling: a Grounded Theory Approach to Cross-Dressing And Sex-Changing*. Routledge, London and New York

Fisk N M 1974 Gender dysphoria syndrome: The conceptualization that liberalizes indications for total gender reorientation and implies a broadly based multi-dimensional rehabilitative regime. *The Western Journal of Medicine* **120**: 386–91

Foucault M 1980 *Hercule Barbin, Being the Recently Discovered Memoirs of a Nineteenth-Century French Hermaphrodite*. Pantheon Books, New York

Halberstam J 1994 F2M: The making of female masculinity. In: Doan L (ed.) *The Lesbian Postmodern*. Columbia University Press, New York

Halberstam J 1998 *Female Masculinity*. Duke University Press, Durham, NC

Harry Benjamin International Gender Dysphoria Association 1985 [1980] Standards of Care: The hormonal and surgical sex reassignment of gender dysphoric persons. *Archives of Sexual Behavior* **14**(1): 79–90

King D 1993 *The Transvestite and The Transsexual: Public Categories and Private Identities*. Avebury, Aldershot, UK

Pauly I B, Edgerton M T 1986 The gender identity movement: A growing surgical-psychiatric liaison. *Archives of Sexual Behavior* **15**(4): 315–29

Prince V 1971 *How to be a Woman, Though Male*. Chevalier, Los Angeles

Raymond J 1980 *The Transsexual Empire: The Making of the She-Male*. The Women's Press, London

Raymond J 1994 *The Transsexual Empire: The Making of the She-Male*, 2nd edn. Teachers College Press, New York and London

D. Valentine and D. Kulick

Treatment of the Repetitive Criminal Sex Offender: United States

Designation of a specific subclass of sex offenders as 'repetitive' derives from sexual psychopath legislation that has sought, over the past half of the twentieth century, to confine sex offenders for indeterminate periods under civil commitment. In response to horrendous crimes committed by a few sex offenders recently released from prison, states passed a series of laws providing for the involuntary, civil commitment of individuals identified as 'sexually dangerous persons,' 'sexual psychopaths,' or 'mentally disordered sex offenders.' Thirty states passed such laws, but over time they fell into disuse or were successfully challenged in court and most were legislatively repealed. In 1990, however, a sex offender recently released from a prison in Washington state abducted, raped, and castrated a 7 year old boy. An outraged public demanded a response, and involuntary commitment was resurrected. The legality of the Washington statute, as well as other similar laws, were immediately challenged, and a Kansas statute, identical to Washington's statute, reached the United States Supreme Court.

1. *The* Hendricks *Era*

The single most influential court decision in the decade of the 1990s has been *Kansas vs. Hendricks*, -U.S.-, 117 S. Ct. 2072 (1997). The Kansas statute established procedures for civil commitment of persons who, due to a 'mental abnormality' or a 'personality disorder' are likely to engage in 'predatory acts of sexual violence.' *Kan. Stat. Ann. Sec.* 59-29a01 1994; *Kansas v. Hendricks*). The state invoked the Act for the first time to commit Leroy Hendricks, a 60 year old man with a long history of taking 'indecent liberties' with children, and who was scheduled for release from prison shortly after the Act became law.

Hendricks had a 'serious mental disorder' (pedophilia) that rendered him dangerous because of a 'specific, serious, and highly unusual inability to control his actions' (from Justice Breyer's dissenting opinion) (Cornwell 1998, p. 397). The clear implication is that pedophiles are incapable of controlling their impulses to molest children. Indeed, in those instances where diagnostic criteria may be inferred, the predominant behavioral criterion for the old sexual psychopath statutes and the new sexual predator statutes involves some form of volitional impairment (i.e., an inability to control impulses to engage in acts of sexual aggression). Impaired impulse control is, at least in principle if not in reality, associated with 'repetitive' offenses. Thus, the 'modal' individual committed under these statutes is a repetitive or habitual sex offender. Importantly, the use of the word 'repetitive' in statutory lexicon tells us relatively little about the individual, clinically or dynamically. All offenders with multiple known sexual crimes may be classified as 'repetitive,' resulting in a highly heterogeneous group.

2. *Treatment of the Institutionalized Sex Offender*

To improve the constitutional viability of civil commitment for 'sexual predators,' treatment had to be

offered. The model for treating these 'repetitive' sex offenders is no different, however, from the model used to treat sex offenders not classified as predators or as repetitive. The cognitive-behavioral model is an adaptation of relapse prevention, and the many elaborations and revisions of this model have been the subject of considerable commentary (cf., Laws 1989, Schwartz and Cellini 1995, 1997; see *Sex Offenders, Clinical Psychology of*). The rationale for the modification of any unwanted behavior stems from the informed consideration of those factors that are most importantly associated with the emergence and the sustenance of the behavior. From the ample research on addictive disorders, particularly alcoholism, smoking, and obesity, there has emerged a substantial literature on preventing relapse (Brownell et al. 1986). Relapse prevention, as a model, has been revised and adapted for use with sex offenders (cf. Laws 1989, Laws et al. 2000). Although sexual aggression derives from a substantially more complex amalgam of factors than most addictive disorders and typically reflects a chronic pattern of maladaptive behaviors, the underlying principles remain the same. That is, before designing strategies for modifying sexually aggressive behavior, we first must identify those factors that are most importantly related to the behavior (i.e., determinants of relapse). Although the factors discussed here are, by no means, inclusive, they do represent the core components of most contemporary treatment programs.

3. *Core Elements of Most Treatment Programs for Sex Offenders*

3.1 *Impaired Adult Relationships*

Impaired relationships with adults reflect a broad dimension of social competence that is temporally stable and multiply determined. Although social skills training is one of the most frequently included components of treatment for sex offenders, its role in relation to sexual aggression remains unclear. Because there are many facets to social competence, specifying what aspect one is assessing is critical. For instance, although there is reasonable support for the conclusion that rapists have assertiveness problems, differences in other areas of social competence are less evident. Likewise some child molesters, such as 'pedophiles,' have a variety of social and interpersonal inadequacies.

Although the question of the etiologic importance of social competence remains uncertain, the inclusion of techniques to address social skills deficits is still clinically appropriate. In addition to social skills training, the modalities that are most often used to improve relationships with adults include assertiveness training, relaxation training, systematic desensitization, sex education, and self-esteem enhancement.

3.2 *Lack of Empathy*

In all domains of interpersonal violence, a general lack of empathic relatedness for one's victim can be regarded as a powerful disinhibitor. Alternatively, the presence of empathic concern may serve to inhibit aggression. There is a clear clinical rationale for assuming the presence of such deficits and targeting interventions to enhance empathy.

Hildebran and Pithers (1989) described the importance of developing empathy for all victims of sexual abuse as an essential first stage of treatment, followed by the development of empathy for one's own victims. At this point most sex offender treatment programs include a separate component for increasing victim empathy. Indeed, Knopp et al. (1992) found in their survey of treatment programs in North America that 94 percent included victim empathy training. In addition to the standard exercises and tapes (video and audio) used in victim empathy training, expressive therapy may be used to increase the offender's emotional or affective response to the distress of the victim. Some programs introduce victim advocates, victim counselors and occasionally victims to increase further the emotional impact. Moreover, increasing the offender's affective appreciation of his own childhood experiences of victimisation can instill a greater awareness of his victim's experience of abuse.

3.3 *Degree and Nature of Anger*

The degree and nature of anger evidenced in the assault has long been assumed to differentiate between offenders who were—and those who were not—willing to use extreme force to gain victim compliance (cf. Prentky and Knight 1991). Sex offenders differ considerably in the nature and extent of their anger and the ways that anger manifests itself in their crimes. Sex offenders who intend only to force victim compliance are likely to vary widely in the amount of aggression evident in their offences. When the aggression clearly exceeds what was necessary to force victim compliance, the motivation and manifestation may also vary considerably. In some cases the anger may be undifferentiated with respect to victims (i.e., the offender may be globally angry and express his anger at any available target), and in other cases the offenders may be focally angry at women (i.e., misogynistic anger) and displace their anger at women in general or on specific groups of women (e.g., prostitutes or the elderly). Similarly, child molesters who intentionally inflict considerable physical injury on their victims may be focally angry at children.

The recognition of the critical importance of anger as a driving force in sexual offenses has resulted in the inclusion of treatment techniques to reduce and contain anger. The most commonly employed of these techniques is an anger management group which uses

cognitive-behavioral strategies to increase self-control as well as the timely and appropriate expression of angry feelings. In addition, Relapse Prevention, which also focuses on increasing self-management skills, and Stress Management can assist the offender to gain control over chronic and situationally-induced anger. Finally, early life experiences of victimization can fuel lifelong anger that is periodically triggered by real or imagined provocations. A group that focuses on childhood victimization can help the offender to master these traumatic events.

3.4 Cognitive Distortions

Cognitive distortions are 'irrational' ideas, thoughts, and attitudes that serve to: (a) perpetuate denial around sexually aggressive behavior, (b) foster the minimization and trivialization of the impact of sexually aggressive behavior on victims, and (c) justify and sustain further sexually aggressive behavior. Cognitive distortions are presumed to be learned attitudes that are instilled at an early age by caregivers, reinforced by peers during childhood and adolescence, and further strengthened in adulthood by the prevailing social climate.

The social and cultural forces that have been hypothesized to contribute to sexual violence include the permissive responses of a wide variety of social systems and institutions that function to perpetuate rape myths and misogynistic attitudes, the objectification and exploitation of children and women in pornography, and the often similar but more subtle messages conveyed in advertising that support, or at least condone, sexual harassment (cf. Stermac et al. 1990). The importance of cognitions in moderating sexual arousal has been repeatedly demonstrated. Moreover, clinical observations have suggested that many sexual offenders harbor offense-justifying attitudes and that these attitudes are importantly related to the maintenance of the 'sexual assault cycle.' Thus, the modification of irrational attitudes has been a major focus of treatment intervention. Although there are a variety of treatment modalities that may impact these distortions, the most commonly employed technique is cognitive restructuring. For cognitive restructuring to be most effective, it is critical that cognitive and affective components be addressed. That is, it is insufficient merely to confront the 'distorted' nature of the attitudes, to discuss the role that such distortions play for the individual, or to provide accurate information about sexual abuse (all cognitive components). It is equally important to create discomfort by focusing on the victim's response (e.g., fright, pain, humiliation)—the affective component. This latter exercise is also integral to victim empathy training.

In addition to cognitive restructuring, a group that focuses on childhood victimization can also be very helpful. Since the origin of these distorted attitudes is often a primary caregiver who was an influential role model, as well as exposure to peer role models, often in institutional settings, a group that focuses on these early life experiences can help to trace the cognitions to their source, thereby challenging their generality and diminishing their 'truth' or their 'reality.'

3.5 Sexual Fantasy and Deviant Sexual Arousal

Sexual fantasy refers to cognitive activity that focuses on thoughts and images having sexual content, and deviant sexual arousal refers to an arousal response that is prompted by thoughts, stimuli (visual or auditory), or behaviors that are defined as unconventional or antisocial by society. Studies investigating the relation of sexual fantasy to sexual aggression have typically employed plethysmography to assess sexual arousal to auditory and visual stimuli that purportedly tap specific sexual preferences. The guiding premise of these studies has been that deviant sexual fantasy is highly correlated with deviant sexual arousal and that both deviant fantasy and deviant arousal patterns are important precursors of deviant sexual behavior (cf. Prentky and Knight 1991).

The frequent targeting of such fantasies for therapeutic intervention reflects the widely held belief that the modification of deviant sexual fantasies and deviant sexual arousal patterns is critical for the successful treatment of sex offenders. Indeed, the presence of deviant sexual fantasies does appear to increase the likelihood of subsequent deviant sexual behavior. Moreover, the moderate success at increasing nondeviant arousal and behavior by applying techniques aimed solely at modifying arousal to deviant sexual fantasies supports the hypothesis that deviant fantasies not only lead to and maintain deviant sexual behavior but also impede normal sexual adaptation.

Behavioral techniques for modifying sexual arousal patterns are grouped into two categories, those that decrease deviant arousal (e.g., covert sensitization, aversion, masturbatory satiation) and those that increase appropriate arousal (systematic desensitization, fantasy modification and orgasmic reconditioning). Although numerous behavioral techniques have been reported in the literature, the most widely used method has involved some variant of aversive therapy.

In addition to the repertoire of behavioral interventions, somatic treatment has become increasingly popular as a complement to psychological treatment. These organic or drug treatments consist primarily of antiandrogens (e.g., Bradford 1989) and antidepressants (e.g., Kafka 1991). The antiandrogens (e.g., medroxyprogesterone acetate and cyproterone acetate) reduce sexual drive by reducing the level of testosterone. The antidepressants that are used are primarily the selective serotonin reuptake inhibitors, such as fluoxetine. Although the neuroregulation of

sexual drive remains unclear, there is some evidence that enhanced central serotonin neurotransmission inhibits sexual arousal. In addition, there is clinical evidence that uncontrollable sexual urges and compulsive sexual behaviors are associated with dysthymia and major depression.

3.6 Antisocial Personality/Lifestyle Impulsivity

In research on sexual offenders, impulsive, antisocial behavior has proven to be a critical risk factor for recidivism (e.g., Hall 1988, Prentky and Knight 1986). There is increasing evidence even within noncriminal samples (e.g., college students) that the likelihood of engaging in sexually aggressive behavior is greater among those who are more impulsive (Prentky and Knight 1991).

Clinicians have long recognised the importance of impulsivity for relapse and have introduced self-control and impulsivity management modules into treatment. In addition to groups that focus specifically on impulse control, most treatment programs include components of Relapse Prevention. Relapse Prevention begins by identifying the chain of events and emotions that lead to sexually aggressive behavior. Once this 'assault cycle' is described, two interventions are employed: (a) strategies that help the offender avoid high risk situations, and (b) strategies that minimize the likelihood that high risk situations, once encountered, will lead to relapse. This is an 'internal self-management' system (cf. Pithers 1990) that is designed to interrupt the seemingly inexorable chain of events that lead to an offense. Relapse Prevention is potentially helpful for interrupting patterns of behavior that eventuate in specific outcomes, such as sexual assault, as well as patterns of behavior that are more global, such as impulsive, antisocial behavior.

4. Conclusion

Given the failure of more traditional correctional remedies, such as deterrence and incapacitation, for reducing the level of sexual violence in society, other interventions must be actively sought. One potentially effective intervention for known offenders is treatment. The verdict as to the efficacy of treatment for sexual offenders will inevitably be a complex one that addresses: (a) optimal treatment modalities for specific subtypes of offenders, (b) optimal conditions under which treatment and follow-up should occur, and (c) selection (or exclusion) criteria for treatment candidates. At the present time, the most informed and dispassionate conclusion must be that the jury is still out. The evidence submitted thus far, however, is encouraging.

The essential point, of course, is the legislative intent of laws that commit sexual offenders for indefinite periods of treatment. If the primary intent of the law is preventive detention, as Pacht (1976) argued 25 years ago, and treatment is merely thrown in so that the law will pass constitutional muster, then it is very unlikely that the law will promote a marked diminution in recidivism among released offenders. The task of effecting behavior change in seriously characterologically disordered individuals is, in itself, considerable. When these individuals have not chosen treatment but are placed in treatment by the court (as is often the case), and when the treatment occurs in the context of a dangerous prison environment where the goal is survival not therapy (as is often the case), then the task becomes immeasurably more difficult. When the treatment provided is suboptimal, the conditions (i.e., prison) are highly problematic and the therapists are inadequately trained and supervised, the task may well be impossible. Under these onerous conditions it would be an egregious error to draw any conclusions about the presumptive inefficacy of treatment with sex offenders.

See also: Rape and Sexual Coercion; Rehabilitation of the Offender

Bibliography

Bradford J M W 1989 The antiandrogen and hormonal treatment of sex offenders. In: Marshall W L, Laws D R, Barbaree H E (eds.) *Handbook of Sexual Assault: Issues, Theories, and Treatment of the Offender*. Plenum, New York, pp. 297–310

Brownell K D, Marlatt G A, Lichtenstein E, Wilson G T 1986 Understanding and preventing relapse. *American Psychologist* **41**: 765–82

Cornwell J K 1998 Understanding the role of the police and *parens patriae* powers in involuntary civil commitment before and after *Hendricks*. *Psychology, Public Policy, and Law* **4**: 377–413

Hall C C N 1989 Criminal behavior as a function of clinical and actuarial variables in a sexual offender population. *Journal of Consulting and Clinical Psychology* **56**: 773–5

Hildebran D, Pithers W D 1989 Enhancing offender empathy for sexual abuse victims. In: Laws D R (eds.) *Relapse Prevention with Sex Offenders*. Guilford Press, New York, pp. 236–43

Kafka M P 1991 Successful antidepressant treatment of non-paraphilic sexual addictions and paraphilias in men. *Journal of Clinical Psychiatry* **52**: 60–5

Kan. Stat. Ann. Sec. 59–29a01 (1994)

Knopp J H, Freeman-Longo R E, Stevenson W 1992 *Nationwide Survey of Juvenile and Adult Sex-offender Treatment Programs*. Safer Society Press, Orwell, Vermont

Laws D R (ed.) 1989 *Relapse Prevention with Sex Offenders*. Guilford Press, New York

Laws D R, Hudson S M, Ward T (eds.) 2000 *Remaking Relapse Prevention with Sex Offenders*. Sage, Thousand Oaks, CA

Pacht A R 1976 The rapist in treatment: Professional myths and psychological realities. In: Walker M J, Brodsky S L (eds.) *Sexual Assault: The Victim and the Rapist*. Lexington Books, Lexington, MA

Pithers W D 1990 Relapse prevention with sexual aggressors: A method for maintaining therapeutic gain and enhancing external supervision. In: Marshall W L, Laws D R, Barbaree H E (eds.) *Handbook of Sexual Assault: Issues, Theories, and Treatment of the Offender*. Plenum, New York, pp. 343–61

Prentky R A, Knight E A 1986 Impulsivity in the lifestyle and criminal behavior of sexual offenders. *Criminal Justice and Behavior* **13**: 141–64

Prentky R A, Knight R A 1991 Identifying critical dimensions for discriminating among rapists. *Journal of Consulting and Clinical Psychology* **59**: 643–61

Schwartz B K, Cellini H R (eds.) 1995 *The Sex Offender. Corrections, Treatment and Legal Practice*. Civic Research Institute, Kingston, NJ

Schwartz B K, Cellini H R (eds.) 1997 *The Sex Offender. New Insights, Treatment Innovations, and Legal Developments*. Civic Research Institute, Kingston, NJ

Stermac L E, Segal Z V, Gillis R 1990 Social and cultural factors in sexual assault. In: Marshall W L, Laws D R, Barbaree H E (eds.) *Handbook of Sexual Assault: Issues, Theories, and Treatment of the Offender*. Plenum, New York, pp. 143–59

R. A. Prentky and B. K. Schwartz

Trials: Cultural

1. Introduction

It is tempting to adopt the definition of trials offered by Lon Fuller, leading modern authority on adjudication, who defines a trial as a process of adjudication which permits disputants 'an opportunity for the presentation of proofs and reasoned arguments' in order to give 'formal and institutional expression to the influence of reasoned argument in human affairs.' As such, trials 'assume a burden of rationality not borne by any other form of social ordering' (Fuller 1981, pp. 93–4). Although this description probably captures cultural assumptions about the contemporary Western trial, it fails to encompass the idea of a trial more generally. It treats human rationality as a given rather than a variable, and it assumes that the primary focus of trials is to resolve disputes, rather than to proclaim truth or to assert authority.

I suggest an alternative: a trial is a process by which proof is authoritatively established. Rationality can be but is not necessarily involved in the process, parties affected may or may not have an opportunity to offer proofs or to make reasoned arguments, and there may or may not be a 'dispute.' This more general definition is useful in accounting for a greater variety of processes that are generally understood to be trials, although in widely different settings, including so-called simple or primitive societies, intensely religious societies, and modern Western societies. It also includes processes in which there are no 'disputes,' and where disputes are not 'resolved' so much as they are processed (see *Courts and Adjudication*).

A good place to start in thinking about this more generic definition of trial is Max Weber's analysis of rational and irrational forms of legal decision-making (Weber 1954, Trubek 1972). For Weber, *forms* of legal decision-making vary along two dimensions: the nature of their rationality and their degree of formality. By rationality Weber meant the presence or absence of 'general rules' that are to be invoked in decision-making. A rational legal system (and decision), according to Weber, involves use of abstract, calculable, and general principles that are invoked when deciding a particular case. In contrast, an 'irrational' system does not rely on the application of general rules to particular facts or cases, but is guided by a holistic process that may not be determinable in advance of the particular decision. Similarly, Weber pointed out, legal decision-making varies in its degree of formality. At one end of the continuum is a formal legal system that is autonomous; it has a distinct set of 'legal' norms and procedures to guide legal decision-making, and perhaps also a distinct set of offices and officials to make decisions. At the other end of the continuum are not casual, informal norms but rather an array of substantive norms, a mixture of legal, religious, moral, and practical precepts that are intermixed and undifferentiated, and so do not constitute a separate and distinct set of (formal) legal norms. In this way Weber distinguished *formal* justice from *substantive* justice. In proceedings dispensing substantive justice, religious, traditional, societal, ethnic, and family norms may blend with distinctively legal norms; here, strictly legal decision-making and decision-makers are not always distinguishable from other realms of social behavior and social actors (see *Weber, Max* (1864–1920)).

Legal decision-making can vary independently along both dimensions, so Weber developed a set of ideal types that has often been used to characterize distinct forms of legal decision-making across time and across cultures. They are presented in Table 1.

Table 1 generates four identifiable types of 'trials': (I) informal irrational (substantive irrationality); (II) informal rational (substantive rationality); (III) formal irrational; and (IV) formal rational. The two extremes on the diagonal (cells I and IV) are familiar to many students of law and the social sciences. Readers of Weber are familiar with the example for cell I—the decision by a *khadi* (an Islamic clerical judge), who after hearing a case hands down a ruling without explicit reference to any general rules. Weber believed that the outcome is understood in Islamic culture as an expression of divine inspiration transmitted through the *khadi*, and that the ruling applies only to the particular case in hand. By contrast, cell IV represents the trial in the modern, rule-of-law state. Facts are introduced and testimony is taken according to a clear and precise set of procedures, and this evidence is judged in light of predetermined rules. It is this type of decision-making that Fuller had in mind in his

Table 1
Weberian typology of legal decision-making (adapted from Inverarity et al. 1983)

	Degree of rationality (presence or absence of general rules) Irrational	Rational
	I	II
Informality (substantive justice)	Substantive irrationality khadi justice	Substantive rationality cost/benefit analysis
Degree of formality		
	III	IV
Formality (formal justice)	Formal irrationality trial by ordeal	Formal rational modern Western trial

discussion of the forms and limits of adjudication. Turning to Weber's two other types of decision-making, cell II represents substantive rationality, a form of decision-making that does not entail a distinct and separate legal sphere but melds law with norms in other social realms (or, more precisely, separate realms are not differentiated). Examples include legal decisions based upon cost-benefit analysis or religious precepts. Such decisions are 'substantive' not 'formal,' since they do not derive legitimacy from a distinct and separate legal sphere. Yet they are 'rational' in that they specify clear norms, rules, and procedures. Cell III describes a process that is formal in that it depends upon specialized and distinct legal procedures. But because decisions flow from an ineffable process rather than from general rules, the process is irrational. A familiar example is the establishment of proof through appeals to divine intervention—ordeal by fire, ordeal by water, or other attempts to invoke supernatural signs. Whatever its defects, this last set of procedures for establishing proof had one virtue, if only for true believers: certainty of outcome (see *Law and Development*).

The discussion above suggests that what Fuller had in mind as the essence of the trial is in fact only one of several forms of legal decision-making. However, he is not so far from the mark. Weber's purpose in developing the typology outlined above was not only to emphasize structural variability in legal decision-making, but also to set the stage for his theory that modernization requires the adoption of formal legal rationality. He saw formal rational adjudication, like bureaucratization, as one of the essential features and defining characteristics of modernity. It is therefore important to understand what is lost by abandoning the other three forms of decision-making in his typology, and what is gained by adoption of the fourth form. According to Weber, modern law undermines competing forms of authority. To the extent that modern law triumphs—and law is notoriously good at triumphing—legal precedent, legal decision-making, and hence the modern trial overshadow tradition, custom, the family, and religion as sources of authority. In this sense it simplifies and rationalizes autonomous legal authority. The modern idea of the rule of law also aspires to precision; it identifies legal entities and separates them from other forms of control. Thus it offers both a sense of freedom and predictability at the same time. In a modern system the ruling at a trial trumps other commitments, other relationships, and other rule systems. That is, the modern trial privileges the state: as officials of the state, modern judges are part of the state-made legal system, and must apply only state-made law. Modern trials undermine other legal systems and challenge the idea of legal pluralism. However, there is a continuing debate about how autonomous modern law is or must be (Nonet and Selznick 1977).

2. *The Multiple Functions of the Trial*

Privileging state law and trials, as formal rational forms of decision-making do, does not mean that there are no differences in the nature and function of trials in modern states. Forms and functions of trials vary widely in modern societies. Damaska (1986) argues that states with traditions of strong government rely on trials not only as means to handle disputes between the parties, but also as opportunities to reassert and to implement state policies. In contrast, weak states tend to view trials in conflict-resolution terms. European countries with traditions of strong central government and civil law systems are examples of the former, while Great Britain and the United States, less-centralized states and common law systems, are examples of the latter. Predictably, settlements in lieu of trial are formally discouraged in the former set of countries and encouraged in the latter. This doubtless helps to explain why in England and the United States a much higher proportion of cases commenced in courts are eventually settled short of trial, relative to most countries with civil law systems (Jacob et al. 1996). Political scientist Martin Shapiro, who has explored some of the same themes developed by Damaska (Shapiro 1981), maintains that always and everywhere, courts—including trial courts—in addition to being institutions for conflict resolution, are important

agents of social control. He means that in all societies courts are used by the dominant political regime to help promote its agenda. Trials, therefore, always serve two simultaneous and symbiotic functions: they resolve conflicts among the immediate parties, and they provide an opportunity for a type of micromanagement, in the pursuit of regime interests. When trial courts strive to assert independence and do not or will not serve the latter function, Shapiro argues, their jurisdiction will be curtailed. For example in the mid-twentieth century, when English courts failed to rule in favor of labor unions frequently enough at trial, their jurisdiction was curtailed and special labor relations courts, more favorable to labor, were established.

In a more generalized analysis, Durkheim's ideas of social solidarity rest in part upon his understanding of the social functions served by the legal process, including trials (Durkheim 1964). Although his writings focus on punishment and the criminal process, his overall theory emphasizes the solidifying function of a variety of legal institutions, including the trial. Much of the dramaturgy associated with the process of punishment is also found in the rituals of the trial. Trials, and public reactions to them, represent an important and visible expression and reassertion of Durkheim's 'conscience collective.' Labeling theorists in the social sciences have seized upon this symbolic function, and have explored its relevance for the criminal trial. In a classic article, Garfinkel described the criminal trial as possessing all the necessary 'conditions of a successful degradation ceremony' (Garfinkel 1956). Similarly, Stuart Scheingold (1977) and others have argued that trials, and particularly high-profile trials, function to foster a 'myth of rights,' which serves to undermine disruptive political struggles. Symbolic victories at trial give the appearance—but not necessarily the reality—of change, and as a consequence they foster political quiescence (Scheingold 1977). Others have explored the more overtly political functions of some trials. Otto Kirchheimer has identified a particular type of trial, the 'political trial,' which, he argues, serves the primary purpose of communicating political lessons rather than resolving conflicts or holding people accountable (Kirchheimer 1961). The Nuremberg trials for the war crimes of Nazi officials in the aftermath of World War II, and other more recent war crimes trials, are good examples. So too are Stalin's 'show trials' of the 1930s, and those of some of the American protesters of the Vietnam War in the 1970s (see *Political Trials*).

Clearly one of the basic functions of trials is to resolve conflicts. Yet, some social scientists question even this seemingly obvious notion. Unlike mediation, where the parties participate by mutual consent, a trial is a process in which at least one party is an unwilling participant. Consider, for example, the criminal defendant whose appearance must usually be coerced. Moreover, modern law does not necessarily express the deeply valued norms of any tightly-knit community or even widely held norms. As a consequence, both the institutional framework and the norms applied may be regarded with suspicion, if not hostility, by one or more of the participants at trial (Shapiro 1981). Various scholars of legal pluralism emphasize just how problematic these challenges are, and often criticize governments and other scholars for giving pride of place to the formal legal trial as the preferred institution for dealing with conflicts (Galanter 1974). In light of these and related concerns, it is conceivable that a verdict by a *court* may not always be an effective way to *resolve* conflicts, especially in the eyes of the losing party. Consequently, many people who lose at trial continue to press their claims in other forums (Abel 1982, 1983). Others do not even bother to seek redress in courts (Felstiner et al. 1980–1) or seek out formal alternatives to adjudication (see *Parties: Litigants and Claimants*, *Disputes, Social Construction and Transformation of*, *Law, Mobilization of*).

3. *Trial Complexity and its Consequences*

One distinctive feature of modern trials is the replacement of reliance on divine intervention with modern empirically based decision-making. Factual evidence subjected to scrutiny has replaced the evidentiary value of supernatural signs. Eyewitness testimony has superseded oath-taking by the accused. And gathering, sifting, and pondering evidence has been transformed from an informal process into a disciplined arrangement governed by detailed procedures and formal rules. Consider that the earliest-known juries convened in criminal cases in Anglo-Saxon England consisted of men from the local community who had firsthand knowledge about the case and/or the character of the accused. By the nineteenth century, a jury was a group legally barred from any firsthand knowledge of the facts of the case before it, and could receive only evidence screened by highly complex rules and introduced according to complicated procedures. In most ordinary cases, the medieval trial was a relatively informal and brief affair handled by the parties involved in the dispute. But the modern trial is dominated by lawyers—highly-trained and highly-paid professional advocates for the parties to the dispute—who must contend with the proliferating subtleties of modern substantive law, and who must practice according to increasingly complex rules of evidence and procedure.

Differences in the structure of the trial process produce a number of important differences in effects. In systems with jury trials, proceedings are likely to rely heavily on oral evidence. But in European and other legal systems that do not have or do not rely so heavily on juries, the trial often proceeds by written

submissions to the court. The court confronts a dossier containing statements of facts and conclusions and sworn testimony (depositions), not a series of witnesses telling their stories in open court. Additionally, in systems with juries, the trial once begun is likely to run continuously until each side has rested its case. By contrast, a trial by dossier is likely to be intermittent and extend over a much longer period of time. These differences also affect the pace of true closure to the litigation. Cases in courts with active juries are almost always concluded within a few days of commencement, and verdicts are typically handed down very shortly thereafter. But in systems without juries, with trials by dossier, evidence-gathering is far more protracted, and final judgments are not necessarily issued shortly after submission of the final bit of evidence. Thus, ironically, although the jury-trial system is often criticized as inefficient, it may actually foster expeditious trial proceedings (see *Procedure: Legal Aspects*, *Juries*).

The nature and form of different trial systems have other consequences as well. The conventional understanding is that Anglo-American legal systems have developed elaborate rules of evidence and procedure in order to restrict the power of jurors, who are untrained lay decision-makers. Since civil law systems have had no such need, they have simpler rules. Although these claims have been vigorously disputed, there is no doubt that rules of evidence and procedure in civil law systems with no juries or weak juries are far simpler than those in common law systems with predominant juries. The 'quest for perfect justice' invoked to justify the complexity of common law trial proceedings may be counterproductive (Pizzi 1999). Trials are swifter, cheaper, and perhaps fairer in many civil law countries than they are in the United States and England. Feeley and Lester (1994) have traced increases in length and complexity of trials in England in the eighteenth and nineteenth centuries, and attribute the changes to the growing complexity of rules of evidence and procedure as well as the greater use of representation by lawyers. Examining the comparative effects of European and American trials in the late twentieth century, Kagan (2001) has shown that the expansion of 'adversarial legalism' in the United States dramatically increased the cost and length of trials without offering any corresponding increase in benefits (see *Litigation*, *Legalization*).

This tendency to increase trial complexity and adversariness in the quest for judicial perfection led legal historian John Langbein to liken the modern adversarial criminal trial to Medieval European proceedings (Langbein 1976, 1978). Both systems, he argued, aspired to 'perfect justice' but in fact precipitated disastrous results. Impossibly high standards of proof in medieval law led jurists to transform torture into a legal expediency in order to fulfill their need for certainty (which confession could provide). Contemporary jurists have developed a functional equivalent, the coercive practice of plea-bargaining, to facilitate a similar objective (see *Plea Bargaining*). Certainty of outcome in both cases, Langbein argued, becomes a mere legal nicety, since both arrangements depend upon coercion. Both torture and plea bargaining are little more than loopholes to circumvent unrealistically high standards of proof. Simpler proceedings and more realistic expectations, Langbein and many others have maintained, would result in fairer and more workable trials (see *Legal Formalism*, *Civil Law*, *Torture: Legal Perspectives*).

Finally, in a very real sense the term 'trial' is misleading and is perhaps an outmoded concept. It implies a single coming together of the parties in a forum where facts are presented and issues resolved. Although a great many cases are handled in a single session, most cases involve a series of hearings and/or decisions that may extend over a significant period of time. Consider a putatively 'simple' criminal case; a suspect is arrested and taken to court for an arraignment, at which time she is asked to enter a plea and bail is set. If pretrial release is not effected, there may be subsequent bail hearings. There may be separate hearings on such issues as a request for representation by a legal aid attorney, probable cause, a motion to suppress evidence, a motion to dismiss, or separate trials for co-defendants; there may be a hearing to determine guilt, a sentencing hearing, and appeals on any of the issues just mentioned. Each one of these separate proceedings can be considered a trial in that it can result in an adjudicated determination of an issue. Contrast this simple criminal case with complex civil cases, which can involve dozens of such issues and hearings. Under such conditions, it does not make much sense to talk about either a 'case' or a 'trial,' since contested issues have been authoritatively resolved at any number of different hearings; issues may have moved from trial courts to appellate courts and back several times, and indeed the name of the 'case' may have changed (Feeley and Rubin 1998). European legal scholars are more likely to understand the term 'trial' in this greatly elaborated sense than are Anglo-Americans, who are more familiar with the idea of concentrated, continuous oral proceedings. Yet, the ambiguity of the word is likely to persist in the context of most modern, complex systems of litigation. In light of this, it may be more useful to use the term 'dispute-processing.' This term better captures the dynamic, multi-step systematizing of most contemporary judicial proceedings, and it implicitly acknowledges that legal conflicts may at best only be 'processed,' rather than 'resolved,' at trial (Abel 1982, 1983).

See also: Dispute Resolution in Economics; Disputes, Social Construction and Transformation of; Justice and its Many Faces: Cultural Concerns; Justice and Law; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Legal Process and Social

Science: United States; Norms; Procedure: Legal Aspects; Rule of Law; Rules in the Legal Process

Bibliography

Abel R 1982 *The Politics of Informal Justice: The American Experience, vol. 1*. Academic Press, New York

Abel R 1983 *The Politics of Informal Justice: Comparative Experiences, vol. 2*. Academic Press, New York

Damaska M 1986 *The Faces of Justice and State Authority*. Yale University Press, New Haven, CT

Damaska M 1997 *Evidence Law Adrift*. Yale University Press, New Haven, CT

Durkheim E 1964 *The Division of Labor in Society*. Free Press, New York

Feeley M, Lester C 1994 Legal complexity and the transformation of the criminal process. In: Gouron A, Mayali L, Schioppa A P, Simon D (eds.) *Subjektivierung des justiziellen Beweisverfahrens*. Klostermann, Frankfurt am Main, Germany

Feeley M, Rubin E 1998 *Judicial Policy Making and the Modern State: How the Courts Reformed American Prisons*. Cambridge University Press, New York

Felsteiner W, Abel R, Sarat A 1980–1 The emergence and transformation of disputes: Naming, blaming, and claiming. *Law and Society Review* **15**: 631–53

Fuller L 1981 Forms and limits of adjudication. In: Winston K (ed.) *The Principles of Order: Selected Essays of Lon Fuller*. Duke University Press, Durham, NC

Galanter M 1974 Why the 'haves' come out ahead: Speculations on the limits of legal change. *Law and Society Review* **9**: 95–160

Garfinkel H 1956 conditions of successful status degradation ceremonies. *American Journal of Sociology* **61**: 420–4

Jacob H, Blanckenburg E, Kritzer H, Provine D, Sanders J 1996 *Courts, Law, and Politics in Comparative Perspective*. Yale University Press, New Haven, CT

Inverarity J, Lauderdale P, Feld B 1983 *Law and Society: Sociological Perspectives on Criminal Law*. Little Brown, Boston

Kagan R 2001 *Adversarial Legalism*. Harvard University Press, Cambridge, MA

Kirchheimer O 1961 *Political Justice: The Use of Legal Procedure for Political Ends*. Princeton University Press, Princeton, NJ

Langbein J 1976 *Torture and the Law of Proof: Europe and England in the Ancient Regime*. University of Chicago Press, Chicago

Langbein J 1978 Torture and plea-bargaining. *University of Chicago Law Review* **46**: 3–24

Merry S, Silbey S 1984 What do plaintiffs want? *Justice System Journal* **9**: 151–83

Nonet P, Selznick P 1978 *Law and Society in Transition: Towards More Responsive Law*. Harper Row, New York

Pizzi W 1999 *Trials without Truth*. New York University Press, New York

Scheingold S 1977 *The Politics of Rights*. Yale University Press, New Haven, CT

Shapiro M 1981 *Courts: A Comparative and Political Analysis*. University of Chicago Press, Chicago

Trubek D 1972 Max Weber on law and the rise of capitalism. *Wisconsin Law Review* **3**: 720–53

Weber M 1954 *On Law in Economy and Society* [ed. Rheinstein M]. Harvard University Press, Cambridge, MA

M. M. Feeley

Triangulation: Methodology

1. Introduction

Triangulation refers to the strategy of using multiple operationalizations of constructs to help separate the construct under consideration from other irrelevancies in the operationalization. At its simplest level, triangulation refers to the use of multiple measures to capture a construct. The triangulation strategy, however, also can be applied to multiple operationalizations of treatments and manipulations and to the use of multiple theories, analyses, analysts, methodologies, and research designs, to name but a few. At a basic level, the triangulation strategy requires an acknowledgment that no one measure in the social sciences is a perfect measurement of the construct under consideration. With only one measure of the construct, the error and biases inherent in the measure are inextricably confounded with the construct of interest. By using different measures with different irrelevancies, a researcher can bring into better focus the construct of interest. Although this triangulation logic was initially explicated in the social sciences in terms of measures (e.g., Campbell and Fiske 1959), the same logic can be extended to treatments, settings, populations, and many other aspects of the research enterprise. Although simple in concept, triangulation 'carries with it a host of implications regarding the proper conduct of social research, the effects of imperfection or unreliability of measurement operations on the development of theory, and the manner in which our field might gradually attain the status of the more developed sciences' (Crano 1981, p. 320).

2. History of Triangulation Concept

Over 40 years ago, social scientists called for triangulation of research measures as a means of more adequately capturing the essence of the construct under consideration (Campbell and Fiske 1959, Webb et al. 1966). The term 'triangulation' was probably borrowed from surveying, where multiple measures from different vantage points are used to locate an object in space, and explanations based on physical analogues are instructive. (See, for example, Crano's 1981 example of locating a radio signal by using multiple directional antennae.) Social scientists suggested that social and psychological constructs, as well as physical objects, could be brought into clearer focus

by using multiple measures from different vantage points with unique sources of error. The uses of triangulation quickly expanded to include triangulation of methods, sources, and theories, as well as measures (e.g., Denzin 1978, 1989).

Some researchers adopted the triangulation strategy whole-heartedly and launched large-scale, multi-investigator, multimethod, multisite, multimeasure studies of social problems (e.g., Lewis 1981). The need to contend with conflicting results that arose from different methods and measures in the same study led to extensive discussions of the use of qualitative and quantitative research methods and data, across and within the same project (e.g., Reichardt and Cook 1979). Recently the focus of discussion has shifted to triangulating among multiple paradigms (Greene and Caracelli 1997, Greene and McClintock 1985), multiple stakeholders, multiple analysts, multiple theoretical and value frameworks, and multiple targets within a study, a concept Cook calls 'multiplism' (1985). Developments in analysis techniques have also promoted the use of triangulation, either at the measurement level, as in causal modeling (Joreskog and Sorbom 1984), or at the method level, as in meta-analysis (Glass et al. 1981, Rosenthal 1991).

3. *Logic of Triangulation*

Triangulation strategies can be contrasted with what Cook and Campbell (1979) termed 'mono-operation bias.' Whenever a construct is operationalized in only one way in a research study, the error attached to the method of operationalization, as well as any random error in the operationalization, is inextricably linked with the true essence of the construct under consideration. This relationship is captured by the simple equation $X = T + e$, meaning that every data point (X) is composed of part True Score (T) and part error (e). (See, for example, Crano 1981 for further discussion). With only one measure of a construct, the error caused by method bias, response sets, or investigator bias, to name but a few, cannot be disentangled from the true measure of the construct. (See Cook and Campbell 1979, Webb 1978 for discussions of myriad types of error). By triangulating, or operationalizing the construct in several ways, the sources of influence on the data point can be disentangled, leading to clearer understanding of the construct under consideration. Campbell and Fiske's classic discussion of the multi-trait, multimethod matrix (1959) lays out a strategy for such specification.

The acknowledgment of the necessity of triangulation represented a maturing of the social sciences beyond the paradigm envy which had previously gripped the field (Crano 1981). Early social scientists had attempted to gain acceptance and legitimacy by adopting much of the research trappings of the physical or 'hard' sciences. Operating from the positivist view of science (Cook 1985), social scientists attempted to achieve parsimonious explanations of events and circumstances that would hold across times and settings, much as physical scientists were claiming to do. They sought the precise, well-calibrated measures, the elegant, critical experiments, and the straightforward analyses of the hard sciences. In some areas of research, this approach could be fairly successful. For example, a social scientist could operationalize 'response time' as 'the amount of time that is recorded on this particular calibrated recording device between stimulus and response,' with stimulus and response both being defined in terms of their operationalizations (e.g., 'the onset of the light' and 'pushing the button,' respectively). But as soon as the social scientists ventured beyond the laboratory and basic research, they discovered that mono-operationalization was inadequate to capture the true essence of the constructs, that simple research methods were inadequate to tap underlying relationships accurately, and that simple analysis strategies were inadequate to describe the complicated interactive patterns in the data.

For example, social scientists quickly discovered that they could not define 'family size' as 'the number of people listed on the interview schedule under the heading Family' and be confident that they had captured the true essence of the construct 'family size.' Beyond the problems with deciding whom to list as 'family' (such as deciding what to do with people who are not biologically related but may be committed to the unit and people who are biologically related but rarely present), social scientists realized that some people are not going to report accurately, either because they misunderstand the question, because they are not aware of the true situation, or because they choose to report incorrectly.

To compensate for these problems with error and bias, researchers began to use multiple-operationalizations of key constructs, and many claims of triangulation in research design are based solely on the use of multiple items to tap a construct. But, as Webb (1978) and others pointed out, using multiple instances of measures that share critical components of error does not improve the research much beyond identifying some random error. To be most useful, multiple measures must have differing sources of error, including method error, and errors due to response bias and self-presentation, among many others. Campbell termed this important qualification the 'heterogeneity of irrelevancies' (Brewer and Collins 1981).

So, for example, having the same interviewer ask the same respondent 'how many people live in this house?' in a dozen different ways during the same face-to-face interview does not advance the construct validity of 'family size' very much. A respondent who really does not know that her daughter's boyfriend has been living upstairs will report incorrectly every time. A respon-

dent who does not want to reveal that a cousin who is in the country illegally lives there will consistently not report that person as living in the house. An interviewer who consistently reverses threes and fives will do that every time. So repeated questioning only reduces the error associated with not understanding a particular question or with random errors made by the respondent or the interviewer. This is not to imply that reducing such error is unimportant, but triangulation can be much more powerful when used across types of measures instead of within one method of measurement.

Partly in response to the shared error inherent in verbal responses, (Webb et al. 1966) suggested the use of unobtrusive measures as a means of tapping constructs without the error inherent in verbal responses. Unobtrusive measures are not without error, but their usefulness comes from the fact that many of their error sources are different from the error sources that plague verbal responses (e.g., social desirability, response set). In one famous example, researchers measured the popularity of different museum exhibits by measuring the wear and tear on floor tiles surrounding different displays (Webb et al. 1966). Another researcher measured level of activity among nursing home residents by measuring how dirty adhesive tape attached to their wheelchair wheels became. And Mosteller (1955, cited in Webb 1978) measured the usage of various sections of the *International Encyclopedia of the Social Sciences* by noting how worn and dirty various sections of the book were. (Note to reader: Feel free to smudge, bend, and dog-ear this page.)

4. *Extensions of Triangulation*

The logic of triangulation was quickly extended to methods, treatments, theories, settings, and other aspects of the research enterprise. Denzin (1978) compared the pictures of reality obtained from different researchers using different methodologies at different times to the changing images within a kaleidoscope, sharing some common elements but greatly influenced by the circumstances of viewing. This is not to imply that there is no objective reality, but rather, even if there is an objective reality, it is filtered through the researcher's assumptions, research choices, and the factors unique to that point in time. In much the same way that Heisenberg pointed out to physical scientists that the very act of observing or measuring a physical object can alter the reality of that object, social scientists were coming to appreciate the extent to which their own assumptions, choices, historical time frame, and even mere presence influenced the data they collected.

Denzin (1978) outlined four basic types of triangulation: data, investigator, theory, and methodological. He further divided data triangulation into the three subtypes of time, space, and person. Data triangulation across time refers to both time of day as well as historical setting, which has been referred to as population changes associated with time (Campbell and Stanley 1963). Some time effects are profound and obvious. For example, attitudes toward national security in the middle of the Gulf War could be seriously effected by historical time frame. Other time effects are less obvious. For example, attitudes toward daycare among women who engage in telephone interviews during the day might be very different from attitudes of women who respond in the evening (and presumably might have been at work during the day). Similarly, space could have obvious or nonobvious effects. That gun attitudes might vary by geographic locale is fairly obvious, but the geographic effects on altruism might be less apparent. Similarly, with person effects, bosses and employees might have different views of the workplace, but so might workers near and far from the elevator.

A strength of triangulation is that the researcher does not need to understand completely the variation inherent in the data collection methods. In the same way that random assignment distributes error randomly across conditions in a true experiment even if the researcher is unaware of the importance of the error, triangulation captures and makes somewhat tractable irrelevancies of the research situation, even if the researcher is unaware of the importance of those irrelevancies. Of course, in the same way that true experiments can be strengthened if the researcher has identified problematic outside variables and blocked on those variables (i.e., randomized within levels of those variables), research designs can be strengthened if researchers can specify which outside factors might be problematic and can triangulate in such a way as to capture those irrelevancies. For example, if a researcher suspects that age of respondent will relate to the variables of interest, the researcher should specifically select for a variety of ages among respondents and analyze the data specifically to examine age effects, instead of just hoping that triangulating over samples will capture those irrelevancies.

Beyond data triangulation, Denzin (1978) also advocates investigator triangulation, whereby multiple investigators bring their unique perspectives and biases to the research project. This is an acknowledgment that people, no matter how well trained, are not robotic recorders of social events. Again, if the researcher has *a priori* reason to suspect that some characteristic of the investigator, such as gender, might introduce error into the observations, care can be taken to select equal numbers of male and female investigators and to analyze for differences between the observations of men and women. But the use of multiple investigators should reduce effects of irrelevant observer factors, even in the absence of the researcher's awareness that such factors could be

important. Of course, only irrelevant factors that vary among the investigators will be ameliorated.

In a similar vein, Denzin (1978) proposed that researchers approach projects from the perspectives of multiple theories, taking care to extend the scope beyond the theoretical orientation most comfortable for the researcher. He further suggested that rather than pit only two theories against each other in research, projects should be specifically designed to capture as many theoretical orientations as possible. Hypotheses derived from all of these theories could then be tested, and conclusions could be drawn about the relative strengths and weaknesses of the theories. The logic of triangulation is the same as with previous examples: by using multiple theoretical views, the researcher is better able to bring the research question into focus.

Finally, Denzin (1978) described both within-method and between-method triangulation. Within a single method, such as a telephone survey, multiple items that are intended to tap the same construct can provide much clearer measures of the construct under consideration. Multiple items can be combined via a wide assortment of scaling techniques, ranging from simple additive scales to elaborate factor analytic scales. Again, the triangulation strategy does not result in items measured without error, but, rather, it allows the error that is not shared among the items to be identified and removed. The resulting scale should reflect only the 'true' component of the construct measured plus any error shared by the various items. For example, items that all cause the respondent to respond in a socially desirable manner will not allow the researcher to separate out social desirability components from components that actually tap the construct under consideration. But if some of the items have a response set bias and others do not, the use of multiple measures will allow the response set bias to be identified (within limits).

Because within-method triangulation can leave varying amounts of error inextricably linked to the key construct, researchers are encouraged also to triangulate between methods (Brewer and Hunter 1989, Campbell and Fiske 1959, Cook 1985, Cook and Campbell 1979, Denzin 1978, Reichardt and Cook 1979, Webb et al. 1966). By combining methodologies as different as participant observation and survey research, for example, researchers can study processes under vastly differing assumptions, biases, and errors. For example, the Reactions to Crime Project (Lewis 1981) involved participant observation, content analysis, telephone surveying, and in-person interviewing, all within one large framework. The approach used by the Reactions to Crime Project was sequentially multimethod, with the participant observation being completed first, followed by the telephone survey, and finishing with the in-person interviewing. With sequential multimethod work, researchers can use results from previous to stages to inform design at later stages. Although this sequential strategy has strengths, Greene and McClintock (1985) point out that concurrent multimethod work, in which all methods are implemented at the same time, avoids contamination of one method with results from another, and produces a purer triangulation.

Probably the most extensive elaboration of triangulation logic was provided by Cook (1985) who described research that involves multiple stakeholders, multiple data analyses, multiple definitions, multiple methods, multivariate causal modeling, multiple hypotheses, multiple populations, multiple settings, and multiple times. He states that such multiplism allows the researcher to approximate the 'ultimately unknowable truth through the use of processes that critically triangulate from a variety of perspectives on what is worth knowing and what is known. Triangulation requires generating multiple instances of some entity and then testing whether the instances converge on the same theoretical meaning despite the irrelevant sources of uniqueness that any one instance may contain' (p. 57). Like previous writers, however, Cook is quick to point out that 'multiplism is no panacea, for we can multiply replicate the same mistake or the same parochial set of assumptions' (p. 58). Although triangulation can be vastly expanded in scope, its strengths and weaknesses remain the same.

5. *Limitations of Triangulation*

As Flick (1992) and Fielding and Fielding (1986) have suggested that proponents of triangulation mistakenly believe triangulation solves all construct validity problems. To the contrary, early proponents of triangulation were sounding a warning about the inherent weaknesses of all measures, and the proposal to use multiple measures with nonoverlapping irrelevancies was an acknowledgment of the profound problems inherent in measuring. In a discussion of the distance between desired and achieved measures of social science constructs and hence of the hypotheses the operationalized measures test), Campbell (1969, p. 52) concluded that 'direct knowing, nonpresumptive knowing, is not to be our lot.' Therefore, one important limitation of triangulation is that it does not produce perfect measures.

Another limitation inherent in the triangulation strategy is that time, financial resources, and respondent goodwill conspire to prevent a researcher from triangulating on all the constructs in the research program. One can easily imagine how attempting to triangulate on theoretical orientations, research designs, and measurement strategies could quickly lead to a program of research that could never produce a single result, because it could never be completed. The successful researcher must therefore choose where to invest his/her triangulation resources, and the soundness and usefulness of the research outcomes depend

to a large extent on the wisdom of those choices during the research design phase.

The amount of effort a researcher chooses to devote to triangulating on a construct should reflect both the importance of that construct for the research question and the level of difficulty in achieving a good measurement. For example, if age is a crucial variable for a research question, investigators might invest extra resources in accessing birth records as well as asking for age and birthrate in a survey. If age is less crucial to the question at hand, the researchers might decide to rely on the verbal replies only. Similarly, a true experiment performed in a laboratory might provide adequate for testing a hypothesis about visual processing, but researchers investigating hypotheses about reactions to stress might need to invest in multiple methods to achieve clarity about the processes under investigation.

The researchers' choices about constructs for triangulation resources can have significant effects on the results of the research. For example, if a researcher believes Family Income to be a more important construct in explaining school achievement than Parental Education Level and consequently devotes more effort to triangulating on the measure of Family Income than Parental Education Level, analyses might well reveal a stronger relationship between School Achievement and Family Income than School Achievement and Parental Education. If the researcher had made the opposite assumption and had devoted more resources to measuring Parental Education than to Family Income, the results might well have shown the opposite effect. The utility of triangulation, therefore, is limited by the soundness of the researchers' initial decisions about resource allocation.

6. *Advances in Triangulation*

The logic of triangulation is inherent in both causal modeling (Joreskog and Sorbom 1984) and meta-analysis (Rosenthal 1991). Causal modeling is an analysis technique that relies on multiple measures (manifest variables) of some underlying construct (latent variable). Researchers specify which manifest variables relate to which latent variables, and the analysis strategy pulls out common variance among the manifest variables that feed into a latent variable. Variance in the manifest variable that is not shared with the other manifest variables for that latent variable is identified as 'error.' Error terms can be allowed to correlate with other error terms. This is a sophisticated extension of the logic used by Campbell and Fiske (1959) in their explication of the multitrait multimethod matrix. They pointed out that only part of what researchers are measuring is the key construct they are intending to measure. The rest of the measure includes method variance, random error, and possibly other types of error and bias. Causal modeling requires researchers to identify error explicitly in their models, which makes very clear the frailty of any one measure in isolation. In addition, causal models show clearly how error terms from similar measures correlate, increasing researchers' awareness of such measurement effects.

In a similar vein, meta-analysis exploits multiple studies, with multiple operationalizations, multiple investigators, multiple methods, multiple sites, multiple populations, and many other differences to arrive at a summary statistic that reports the strength of underlying relationships across studies. Instead of requiring one researcher or research team to approach a research question from multiple perspectives, theories, and methodologies, meta-analysis capitalizes on natural variations among research programs to achieve triangulation. A main benefit of meta-analysis is that methodological variation can be explicitly built into the analysis. For example, in a recent review of the research on environmental and genetic effects on aggression, Miles and Carey (1997) reported that observational measures in laboratory studies showed no genetic effects and strong environmental effects, while self-report and parental rating measures showed both genes and environment to have strong relationships with aggression. Because variance associated with particular methods and measures can be explicitly analyzed in meta-analyzes, the increased understanding afforded by triangulation strategies becomes apparent. And the conflicting results from different methods, far from being a nuisance or drawback to triangulation approaches, are rich clues to contingencies that need to be brought to light if social research is ever to reveal fully how these processes operate (Cook 1985).

The importance of triangulation is also revealed by the prevalence of statistical interactions in social research. Evaluation researchers and others who study social problems rarely find simple, main-effect effects capture the richness (some would say messiness) of the world beyond the research laboratory (Cook 1985). Evaluators no longer ask 'Does this work?' but, rather, 'For whom does this work, in what settings, with what constraints?' The prevalence of interactions pushes researchers to branch out to different populations, different time frames, different methods, in other words, to triangulate across many factors in the research setting.

Social research has come a long way from the days when researchers sought pure measures, elegant, simple designs, and straightforward analysis strategies. The norm for social research now is multiple measures, elaborate multimethod design, and analysis strategies that clearly indicate the multiple causation and interactive patterns apparent in the social world. In no small part, this march toward increasing complexity of the research enterprise was begun by early explications of the logic of and necessity for triangulation.

See also: Aggregation: Methodology; Content Validity; Generalizability Theory; Indicator: Methodology; Meta-analysis: Overview; Unobtrusive Measures

Bibliography

Brewer J, Hunter A 1989 *Multimethod Research: A Synthesis of Styles*. Sage, Newbury Park, CA

Brewer M, Collins B E 1981 Perspectives on knowing: Six themes from Donald T. Campbell. In: Brewer M, Collins B E (eds.) *Scientific Inquiry and the Social Sciences*. Jossey-Bass Publishers, San Francisco, CA

Campbell D T 1969 A phenomenology of the other one: Corrigible, hypothetical, and critical. In: Mischel T (ed.) *Human Action: Conceptual and Empirical Issues*. Academic Press, New York

Campbell D T, Fiske D W 1959 convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* **56**: 81–105

Campbell D T, Stanley J C 1963 *Experimental and Quasi-experimental Designs for Research*. Rand McNally, Chicago

Cook T D 1985 Postpositivist critical multiplism. In: Shotland L, Mark M (eds.) *Social Science and Social Policy*. Sage, Beverly Hills, CA

Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand McNally College Publishing, Chicago

Cook T D, Reichardt C S (eds.) 1979 *Qualitative and Quantitative Methods in Evaluation Research*. Sage, Beverly Hills, CA

Crano W 1981 Triangulation and cross-cultural research. In: Brewer M, Collins B E (eds.) *Scientific Inquiry and the Social Sciences*. Jossey-Bass Publishers, San Francisco, CA

Denzin N K 1978 *Sociological Methods: A Sourcebook*. McGraw Hill, New York

Denzin N K 1989 *The Research Act: A Theoretical Introduction to Sociological Methods*, 3rd edn. McGraw Hill, New York

Fielding N G, Fielding J L 1986 *Linking Data*. Sage, Beverly Hills, CA

Flick U 1992 Triangulation revisited: Strategy of validation or alternative? *Journal of the Theory of Social Behavior* **22**: 175–197

Glass G V, McGaw B, Smith M I 1981 *Meta-analysis in Social Research*. Sage, Beverly Hills, CA

Greene J, Caracelli V J 1997 Defining and describing the paradigm issue in mixed-method evaluation. In: Greene J, Caracelli V J (eds.) *Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms*. Jossey-Bass Publishers, San Francisco, CA

Greene J, McClintock C 1985 Triangulation in Evaluation: Design and Analysis Issues. *Evaluation Review* **9**: 523–45

Joreskog K G, Sorbom D 1984 *LISREL VI, Analysis of Linear Structural Equation Systems by the Method of Maximum Likelihood: User's Guide*. International Educational Services, Chicago

Lewis D 1981 *Reactions to Crime*. Sage, Beverly Hills, CA

Miles D R, Carey G 1997 Genetic and environmental architecture of human aggression. *Journal of Personality and Social Psychology* **72**: 207–17

Mosteller F 1955 Use as evidenced by an examination of wear and tear on selected sets of ESS. In: Daris K et al. (eds.) *A Study of the Need for a New Encyclopedic Treatment of the Social Sciences*. Unpublished manuscript

Reichardt C S, Cook T D 1979 Beyond qualitative versus quantitative methods. In: Cook T D, Reichardt C S (eds.) *Qualitative and Quantitative Methods in Evaluation Research*. Sage, Beverly Hills, CA

Rosenthal R 1991 *Meta-analytic Procedures for Social Research*. Sage, Newbury Park, CA

Webb E J 1978 Unconventionality, triangulation, and inference. In: Denzin N K (ed.) 1978 *Sociological Methods: A Sourcebook*. McGraw Hill, New York

Webb E J, Campbell D T, Schwartz R D, Sechrest L 1966 *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Rand McNally, Chicago

L. Heath

Tribe

The notion of tribe emerged in the seventeenth and eighteenth centuries in the course of colonial expansion. Abandoned increasingly by the majority of anthropologists from the middle of the twentieth century onward, the term nevertheless lived on in certain anthropological constructions, until waves of criticism dealt final blows to a concept that is now largely defunct as a general comparative category. Those anthropologists and historians who retain the contested concept of tribe use it mainly as an equivalent to vernacular terms referring to certain local conceptions of collective sociopolitical identity, while agreeing that a precise general definition is almost impossible. This results in working definitions which see tribes as medium-sized, centralized, or acephalous entities displaying a combination of basic characteristics:

(a) They are usually associated with a territory, homeland, or tribal area, while using nonterritorial criteria to distinguish between members and nonmembers. The degree of control of the tribal territorial base, and its extension and location, tend to change as a result of struggles with states or other opponents. To the outside world, access to territorial usage appears to be confined to members of the tribe, although neighboring groups and superior authorities may also enjoy rights of usage, of tribute, or of taxes. Internally, access to territorial usage may be differentiated according to gender, age, status, wealth, and other aspects of power.

(b) Ideologically and socially, tribal members usually share some dominant idiom of common origin, such as descent from a single ancestor, emphasizing group cohesion over outside interests and internal differentiation. This may also serve to legitimize a preference for 'marrying within the tribe,' which in practice, however, is observed far less often than it is upheld in theory. On the contrary, some amount of intertribal and interethnic marriage is indispensable to ensure all kinds of (political, economic, social) alliances with other groups.

(c) Although part of the tribal population live more or less permanently in the tribal territory, a considerable portion may live in the world outside. Ritual and kinship ties and obligations are enacted to ensure and strengthen intimate networks among 'home' groups and 'outside' groups, serving as one among several inventories of group adherence within a wider world. In the course of these interactions, tribal group membership may change in many ways: individuals or groups may enter or leave a specific tribe; and whole tribes may also collapse or be created.

1. *Tribes and Colonialism*

The concept of tribe dates from the earliest phases of anthropology's history, carrying assumptions that are now being criticized and questioned inside and outside academia. Most governments of colonial times used the term for their specific practical interests of administration and dominance, while often constructing such 'tribes' themselves for that purpose. Beyond these practical interests, colonial settler regimes also valued the term ideologically, since it tended to portray local groups as 'backward' and 'primitive,' which helped to justify a colonial 'mission' of development and civilization. For basically the same reason, many post-colonial governments and national elites reject the term as being counterproductive to their own claims toward national unity, progress, and modernity.

During decolonization, the term tribe therefore attained a largely negative and pejorative meaning in dominant opinions in many parts of the world. In public discourses concerning these areas, the terms 'ethnic groups' or 'indigenous peoples' are usually preferred to 'tribes,' although the latter usually designates smaller groups. Legal terminology nevertheless may continue to apply the term, and in many regions even native speakers may use it with dignity.

In parts of central and western Asia, however, as well as in North Africa, particularly in those areas influenced by Islam, the term tribe and its local equivalents have never had a predominantly pejorative meaning. On the contrary, the term here is usually associated with notions of pride and autonomy, of honor and of independence from colonial and local states. Also, notions of 'indigenous' or 'ethnic' groups do not apply well to many of these specific ethnographic contexts. Both of these factors are therefore responsible for the continuing, stronger importance of the term tribe in academic and public discourses inside and about these regions.

2. *Tribes in Anthropology*

Before anthropology's emergence, travelers, missionaries, and scholars had used the concept since the seventeenth century in describing those indigenous peoples whom European powers were about to colonize. Ancient Middle Eastern and Graeco-Roman literature served as sources of reference to depict the newly colonized peoples as representing some kind of forerunners to the 'advanced nations,' equivalent to the *tribus* of the ancient Mediterranean world. The construction of this 'forerunner' image could follow either more sympathetic lines, by emphasizing the enlightened aspect of a noble savage who basically was 'like us,' or it could enhance the more discriminating image of backwardness.

Both ideas presupposed a search for origins, taking European culture as the climax of humanity. In its basic conceptualization, a tripartite world had thus already been constructed, which early anthropology inherited and further elaborated. 'Primitive' cultures consisted of foragers, and pastoral and agrarian 'tribes,' whereas 'oriental' cultures were associated with agrarian states, cities, and with writing. Anthropological evolutionism (e.g. Morgan 1877, Smith 1903) elaborated and specified this conceptual inventory, in which, by and large, tribes constituted the decisive transitory stage between early hunters and gatherers and the rise of agrarian states. In a long-term archeological perspective, this may still serve as a valid distinction for certain regions and periods, if liberated from its earlier ideological underpinnings.

Classical anthropological evolutionism collapsed, however, because the increasing body of empirical data showed the failure of any general laws of development, including those regarding tribes. By contrast, the new school of 'historical particularism' or cultural relativism that came to dominate US anthropology around 1900 emphasized the situated, fluid contextualization of each local and tribal group. While this prepared the ground for an empirical orientation that is still valid at the time of writing, the inherent relativism rejected almost any comparative meaning for the concept of tribe. The comparative aspect, by contrast, was elaborated explicitly by anthropological functionalism, largely of British and Northern European provenance. In an earlier variant, Evans-Pritchard (1940) outlined the 'structural relativity' of many tribal group relations around a basic social organization of 'nested' group segmentation.

In fact, such segmentary relations do represent a dominant model among many local (tribal) groups in Africa and Asia, often being expressed in the prevailing idiom of unilineal genealogy. In these terms, the 'descendants' of one ancestor are contradistinguished, under some conditions, against those from another ancestor, while under different conditions the same group of descendants may split up and oppose each other. Mutual threat, equality of group status, and reproduction of social equilibrium are inherent to this indigenous model, taken up and reconstructed as an anthropological model. Major critiques, however, pointed to implications of timeless stability

and secluded cohesion, which to some degree corresponded well with the ideologies of late colonialism.

Gellner (1981, 1990), who was most influential in transferring the segmentary model from Sub-Saharan Africa to North Africa and the Middle East, reacted to some of that criticism with an increasingly more refined variant. According to him, Middle Eastern tribal societies interact constantly out of the peripheries with each other and with the states in their centers, sometimes being absorbed or reshaped in the course of these interactions, and sometimes themselves destroying and recreating central states. These processes always promote tribal centralization, inequalities, and tendencies to fusion and fission, which are simultaneously more or less countered by applying the local segmentary model. Jamous (1981) has shown in more detail than Gellner how these tendencies to feuding and conflict may be embedded in idioms of 'honor and shame.' His account illustrates how these tensions, which favor the emergence of ephemeral 'big men,' are countered through mediation by third parties, who in many cases are of superior, holy status.

While Marxist, late functionalist, and neo-evolutionist critics (for example, Fried 1975, Godelier 1973, Southall 1970) were the first to radically criticize current ideas about tribes in the 1960s and early 1970s, a few anthropologists held on to a general, neo-evolutionist conception of tribe. Sahlins (1968) elaborated a distinction between acephalous 'segmentary lineage' tribes and more centralized 'chiefdoms.' On the other hand, in Middle Eastern studies the concept gained new respect in the Gellner/Jamous variant, by its consideration of formerly neglected aspects, such as historical dynamism; ideological, political and social factors; and some hierarchies and contradictions. A second wave of academic criticism throughout the 1980s and early 1990s came from feminist, postmodern, and postcolonial writers, who deconstructed profoundly the colonial legacy in the concept, and exposed the complete negligence of economic factors and of female lives in the 'tribal' model (Abu-Lughod 1986, Gilsenan 1996).

3. *Current Readings*

In present discussions, the debate is therefore oscillating between complete rejection of the term 'tribe', and its more or less restricted, critical elaboration for specific times and regions. Most scholars, however, would agree that the concept is obsolete as a general comparative category outside particular areas. In addition, historical and archeological research might continue, for parts of Africa and Asia at least, to find some conception of tribe to remain useful. Such a restricted, critical elaboration of the concept would, of course, have to take into account seriously previous critiques.

Some of the more recent criticism was creatively synthesized by Bonte and Conte (Bonte et al. 1991). Their late structuralist and post-structuralist approaches basically privilege historical change over repetition, interaction over system, status rivalry over group equality, and, most importantly, marriage and other forms of alliance over descent and ancestry. This helps to outline internal hierarchies among local tribal groups, including the constraints and fields of agency encountered by women, while demonstrating simultaneously how tribal organization is used to defend states, to support struggles for central power, or to engage in civil war for various interests (Gingrich 1995).

The postcolonial condition seems in fact to encourage such a restricted, critical elaboration of the concept, rather than discarding it totally. Under the conditions of advanced globalization after the Cold War, and of increased competition within as well as among weakened local states (ranging from Mauritania, Algeria and Somalia to Afghanistan and Bangladesh), ideologies of ethnicity, but also of tribalism, are gaining significance in many local arenas. It would amount to a strange paradox if anthropologists limited themselves defensively to deconstructing completely the concept of tribe, while many parts of the world in fact go through retribalization of some sort. In view of these developments, an active reformulation of the concept seems to be a more productive alternative. With this alternative, ideological movements of 'tribalism' will have to be distinguished more clearly from 'tribes,' parallel to concepts of nationalism and of nations. Tribalism and its activists strive more or less successfully to create imagined communities of tribes, with ideological pretexts of 'origins' and of 'cohesion' that serve competition over power and resources. In this perspective, anthropologists are as critical and skeptical about tribalist ideology as they are with regard to nationalism. All the same, they would not deny from the outset that nation states and tribal groups exist: rather, they would insist on exploring the interrelations between such local entities, their ideological activities, and the world around them.

See also: Chiefdoms, Archaeology of; Clan; Ethnic Identity/Ethnicity and Archaeology; Ethnicity: Anthropological Aspects; Ethnoarchaeology; Nationalism and Expressive Forms; Nationalism: Contemporary Issues

Bibliography

Abu-Lughod L 1986 *Veiled Sentiments: Honor and Poetry in a Bedouin Society*. University of California Press, Berkeley, CA

Bonte P, Conte E, Hamès C, Ould Cheikh A W 1991 *Al-ansâb: la quête des origines. Anthropologie historique de la société tribale arabe*. Éditions de la MSH, Paris

Evans-Pritchard E E 1940 *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People*. Oxford University Press, London
Fried M H 1975 *The Notion of Tribe*. Cummings, Menlo Park, CA
Gellner E 1981 *Muslim Society*. Cambridge University Press, Cambridge, UK
Gellner E 1990 Tribalism and the state in the Middle East. In: Khoury P S, Kostiner J (eds.) *Tribes and State Formation in the Middle East*. University of California Press, Berkeley, CA
Gilsenan M 1996 *Lords of the Lebanese Marches: Violence and Narrative in an Arab Society*. Tauris, London
Gingrich A 1995 The Prophet's smile and other puzzles. Studying Arab tribes and comparing close marriages. *Social Anthropology* **3**: 147–70
Godelier M 1973 Le concept de tribu. Crise d'un concept ou crise des fondements empiriques de l'anthropologie? In: Godelier M (ed.) *Horizon, Trajets Marxistes en Anthropologie*. Maspero, Paris
Jamous R 1981 *Honneur et Baraka: Les Structures Sociales Traditionelles dans le Rif*. Cambridge University Press, Cambridge, UK and Éditions de la MSH, Paris
Morgan L H 1877 *Ancient Society*. Holt, New York
Sahlins M D 1968 *Tribesmen*. Prentice-Hall, Englewood Cliffs, NJ
Smith W R 1903 *Kinship and Marriage in Early Arabia*. Black, London
Southall A 1970 The illusion of tribe. *Journal of Asian and African Studies* **5**(1–2): 28–50

A. Gingrich

Tropospheric Ozone: Agricultural Implications

While stratospheric O_3 protects life by absorbing damaging solar UV radiation, ground-level O_3 can harm life by oxidizing living tissue. Under appropriate meteorological conditions, anthropogenic emissions of nitrogen oxides, volatile organic compounds, and carbon monoxide can trigger O_3 pollution episodes with ground-level concentrations enhanced by factors of 2 or more. One consequence of these enhancements can be a reduction in agricultural crop yields.

1. *Atmospheric Ozone*

Ozone is a 3-atom allotrope of oxygen and thus has the chemical formula: O_3. It is gas with a pungent odor and hence its name, which is derived from the Greek word *ozien* meaning smell. Christian Fredrick Schobein generally is credited with first identifying O_3 as a compound in the laboratory in 1840, and John Hartley is credited with demonstrating its ubiquitous presence in the atmosphere in 1881.

O_3 is one of a myriad of atmospheric *trace species*, so called because their atmospheric abundances are of the order of one part per million (ppm) or less, where 1 ppm of a gas signifies an atmospheric abundance of 1 molecule for every million atmospheric molecules. In fact, in the case of O_3 there are on average about 3.5 O_3 molecules for every 100 million atmospheric molecules. If the entire atmosphere were compressed to a pressure of 1 atmosphere (i.e., the pressure at the earth's surface), it would have a height of about 8.5 km. If the same were done to atmospheric ozone (O_3), it would have a height of only 0.3 cm.

Despite its seemingly minute concentration, atmospheric O_3 has an enormous environmental impact. As a constituent of the stratosphere, where about 90 percent of the atmosphere's O_3 resides, it protects life from harmful uv radiation from the sun. In the troposphere, O_3 plays a key role in determining the oxidizing capacity of the atmosphere, facilitating the process by which the atmosphere is flushed of a wide range of pollutants. O_3 at ground level represents a small but important component of the overall tropospheric O_3 burden. Because of its properties as a powerful oxidant, ground-level O_3 can have adverse effects on human health, agricultural crops, and forests

As illustrated in Fig. 1, O_3 concentrations generally increase with increasing altitude throughout the troposphere (the lowest 10–15 km of the atmosphere) and much of the stratosphere. O_3 concentrations typically range from about 30 ppb near the ground to about 100 ppb or more near the top of the troposphere, where 1 ppb O_3signifies an abundance of 1 molecule of O_3 for every billion molecules of air. However, ground-level O_3 concentrations can exceed 100 ppb during air pollution episodes and especially in urban areas (see *Air Pollution*). The highest hourly-averaged O_3 concentration on record in North America is apparently 680 ppb measured in downtown Los Angeles in 1955. When ground-level O_3 concentrations reach these levels, they threaten rather than sustain life. In particular, they can damage agricultural crops, ultimately causing a reduction in the crop yields grown in affected regions.

2. *The Generation of Tropospheric O_3*

The basic chemical mechanism responsible for the generation of O_3 in the stratosphere was first worked out by Sidney Chapman in the 1930s. However, it would take a few more decades before scientists began to understand the chemical reactions that produce tropospheric ozone.

Stratospheric O_3 production involves two reactions (see Fig. 1). The first is the photolysis and dissociation of molecular oxygen (O_2) by uv radiation from the sun. The second reaction involves the combination of an oxygen atom produced from the first reaction with another O_2 molecule.

Because ultraviolet radiation from the sun capable of photolyzing O_2 does not penetrate into the tropo-

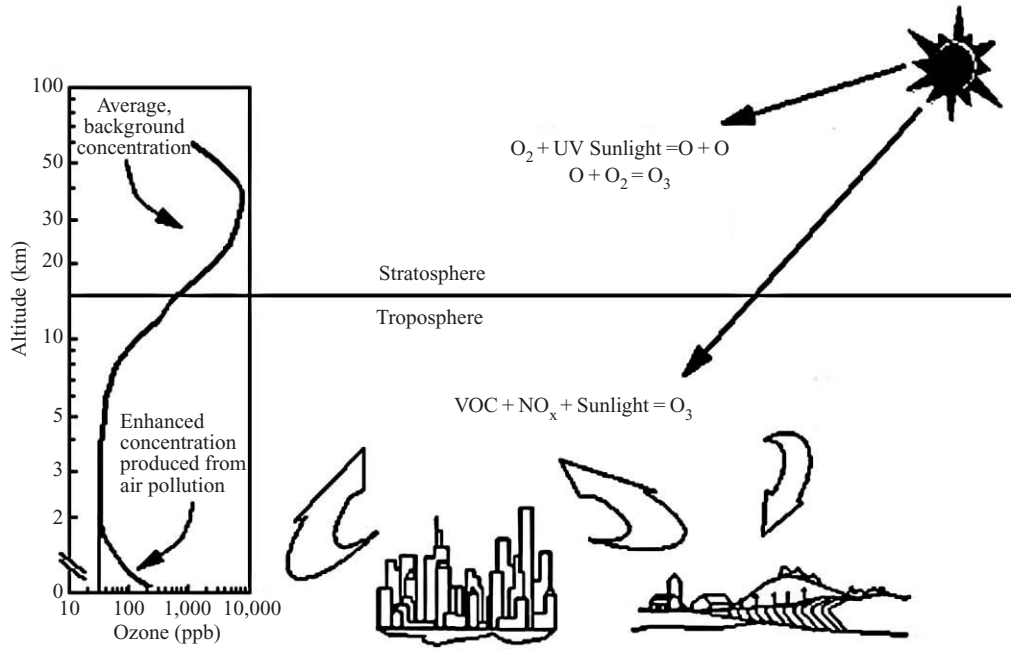

Figure 1
Tropospheric and stratospheric ozone (O_3). Graphic on left illustrates average vertical distribution of O_3 with the O_3 abundance expressed in units of parts per billion (ppb), that is, molecules of O_3 per billion molecules of air. The central schematic illustrates the photochemical processes that produce O_3 in the stratosphere and troposphere

sphere, it was believed for a number of decades following Chapman's work that O_3 is an inert constituent of the troposphere. It was argued that the downward transport of O_3-rich air from the stratosphere is the source of tropospheric O_3 and reaction with material on the earth's surface (a process referred to as dry deposition) removes O_3. However, the appearance of photochemical smog in Los Angeles in the 1950s, and the realization that O_3 was a major product of this pollution forced a gradual reappraisal of this simple picture. Today, it is recognized widely that chemical production of O_3 occurs throughout the troposphere, at relatively modest rates within the more pristine regions of the atmosphere and at significantly accelerated rates in polluted regions.

2.1 The Chemical Reactions that Produce Tropospheric Ozone

A scientist by the name of A. J. Haagen-Smit first unraveled the chemical mechanism that produces O_3 in the lower atmosphere. Working in Los Angeles in the early 1950s, Haagen-Smit became interested in understanding how photochemical smog was formed and what was in the smog that made it noxious. He irradiated mixtures of volatile organic compounds (VOC) and nitrogen oxides (NO_x) in a reaction chamber. (Both VOC and NO_x are emitted into the atmosphere by automobiles and other industrial processes and NO_x represents the sum of nitric oxide or NO and nitrogen dioxide or NO_2.) After a few hours, Haagen-Smit observed the appearance of cracks in rubber bands stretched across the chamber. Knowing that ozone (O_3) can harden and crack rubber, Haagen-Smit correctly reasoned that photochemical smog was caused by photochemical reactions involving VOC and NO_x, and that one of the major noxious constituents in this smog was O_3 (see Fig. 1).

Since Hagen-Smit's work, more sophisticated laboratory experiments, along with advanced chemical monitors and computer simulation models have made it possible to further elucidate the chemical reactions that form O_3 in the troposphere. A typical reaction sequence might include:

(a) $RH + OH \rightarrow R + H_2O$
(b) $R + O_2 + M \rightarrow RO_2 + M$
(c) $RO_2 + NO \rightarrow RO + NO_2$
(d) $RO + O_2 \rightarrow HO_2 + \text{carbonyl}$
(e) $HO_2 + NO \rightarrow OH + NO_2$
(f) $2 \times (NO_2 + h\nu \rightarrow NO + O)$
(g) $2 \times (O + O_2 + M \rightarrow O_3 + M)$

NET: $RH + 4O_2 + 2\ h\nu \rightarrow 2O_3 + \text{carbonyl} + H_2O$

In the above, *RH* represents a hydrocarbon (i.e., a type of VOC), *R* is a hydrocarbon radical or compound fragment (e.g., CH_3CH_2 for RH = ethane), *hν* represents a photon from the sun and *M* is a nonreactive, energy absorbing molecule referred to by chemists as a 'third body.' In the atmosphere, this third body is most often molecular nitrogen (N_2) or molecular oxygen (O_2). Carbon monoxide (CO) can replace the VOC. In this case the reaction sequence is initiated by

(h) $CO + OH \rightarrow CO_2 + H$

(i) $H + O_2 + M \rightarrow HO_2 + M$

and then followed by Reactions (e), (f), and (g).

In the above reaction sequences, VOC and CO are oxidized and consumed, while NO_x acts as a catalyst. Collectively these three compounds are often referred to as *O_3 precursors*. The by-product labeled 'carbonyl' is itself a VOC and can react and produce more O_3. The OH radical, which triggers the sequence by reacting with VOC or CO, is generated initially in the atmosphere from O_3:

(j) $O_3 + h\nu \rightarrow O^* + O_2$

(k) $O^* + H_2O \rightarrow 2OH$

where O* is an electronically excited atom of oxygen. Because the OH consumed by Reaction (a) or (h) is regenerated by Reaction (e), and, because more OH can be generated during the breakdown of the carbonyl produced by Reaction (d), the reaction sequences listed above are net sources of O_3.

In clean regions of the troposphere, O_3 is produced primarily by the oxidation of isoprene (C_5H_8) and/or methane (CH_4), hydrocarbons that are emitted into the atmosphere by biogenic processes, as well as CO. The catalyst NO_x largely is supplied by lightning as well as soil emissions. In polluted regions, VOC from automobiles, solvents, and various industrial processes become important, and the NO_x arises largely from automobiles and power-plant emissions.

2.2 *Ozone Pollution Events*

Ground-level O_3 concentrations tend to be largest during specific periods or episodes when meteorological conditions favor the production and accumulation of pollutants within the *atmospheric boundary layer*. (The atmospheric boundary layer denotes the first 1–2 km of the atmosphere where winds are slowed by frictional resistance with the earth's surface. The frictional resistance, in turn, creates turbulence within the layer that tends to keep pollutants and other chemical constituents uniformly mixed. For this reason the atmospheric boundary layer is also referred to as the *mixed layer*.) These conditions are often associated with summertime, slow-moving high-pressure systems that can extend over thousands of kilometers. Ground-level O_3 concentrations during these pollution episodes tend to be highest in and around urban areas. However, suburban sprawl, growing numbers of automobiles and expanding roadway networks (see *Land Use and Cover Change*; *Development and Urbanization*) coupled with the large spatial scale of the meteorological conditions that foster the pollution, can lead to enhanced O_3 concentrations over thousands of kilometers, encompassing agricultural as well as urban areas. To distinguish this type of pollution from more localized or urban-scale pollution, it is often referred to as *regional air pollution*. The repetition of regional air pollution episodes in agricultural areas over a growing season can produce a pattern of chronic exposure that can, in principle, reduce crop yields. The extent of crop-yield reduction depends upon both the degree of exposure and the sensitivity of the crop to O_3, as discussed below.

3. *Ozone Effects on Crops*

The first documentation of O_3 damage to crops arose out of studies of the effects of photochemical smog in Los Angeles in the 1940s and 1950s. Subsequent research by scientists throughout the world have provided a rudimentary understanding of the mechanisms by which O_3 can affect vegetation and the development of semiquantitative relationships between exposure to O_3 and crop yield.

O_3 damage to vegetation most often occurs as a result of penetration into the leaf interior through the stomata and then reaction with the cell's plasmalemma (i.e., cell membrane) and/or mesophyl (i.e., cell interior). Field studies using open-top chambers to control the O_3 concentrations in contact with crops growing under otherwise ambient conditions have established that crop sensitivities to O_3 vary from crop to crop and, within a given crop, from variety to variety. In general: (a) O_3 damage to crops becomes observable when concentrations exceed a threshold of about 50–70 ppb; (b) the effects of repeated intermittent exposures above this threshold tend to accumulate over a growing season; and (c) the higher the O_3 concentration above the threshold the greater the damage.

Because O_3 damage to crops accumulates over a growing season, any metric used to quantify crop exposure to O_3 must integrate over all pollution episodes occurring in that season. One such metric is the so-called 'SUM06' that represents the sum of all daytime, 1-hour-averaged O_3 concentrations in excess of 60 ppb over a 3-month period. Analysis of data from the US National Crop Loss Assessment Network suggests that the average yields of crops grown in North America are reduced by about 10 percent when SUM06 exposures are about 15–25 ppm-hours, where 1 ppm-hour is equivalent to 1,000 ppb-hours. Expo-

sures of this magnitude are not uncommon in rural areas of many industrialized and developing nations.

3.1 Implication for World Food Production

O_3 damage to crops is likely to be most severe in those areas where there is an intermingling of the urban and industrial activities that lead to the emissions of VOC, CO, and NO_x and agricultural activities on spatial scales of thousands of kilometers. It turns out that such an intermingling of the urban and industrial with the agricultural is quite common in the developed world and is most pronounced in three regions of the northern mid-latitudes: (a) Eastern North America (25–50°N, 105–60°W); (b) Europe (36–70°N, 10°W–90°E); and (c) eastern China and Japan (25–45°N, 100–146°E). Within each of these regions, urban–industrial and agricultural activities tend to cluster together into a single large network or plexus of human-impacted land-use categories. The term Metro-Agro-Plex (or MAP) is sometimes used as a shorthand to describe this intermingling of agricultural and urban-industrial activities within a developed or developing geographical area. The term Continental-Scale Metro-Agro-Plexes (or CS-MAPs) is in turn used to specifically refer to the three large MAPs of the northern mid-latitudes.

Collectively, the CS-MAPs represent a major force in the global economy and also in environmental change. Although they comprise only 23 percent of the earth's land surface, the three CS-MAPs account for about 75 percent of world's consumption of commercial energy and fertilizers and about 60 percent of the food crop production and food exports. They are also major source regions for atmospheric pollutants including the O_3 precursors. Estimates using a global atmospheric chemistry model indicate that, because of these emissions, about 10–35 percent of the world's cereal crops currently are exposed to potentially damaging O_3 pollution within the CS-MAPs. The net effect of this exposure to the world's total production of cereals and other food staples probably amounts to few percent decrease in total production. However, the presence of other pollutants that accompany O_3 in photochemical smog most likely exacerbate the crop losses. For example, the haze associated with photochemical smog and O_3 pollution reduces the amount of sunlight that reaches the earth's surface and this likely also reduces crop yields. Moreover, as pollutant emissions increase in developing nations the impacts on agriculture within these areas will become more severe. The economic implications of such crop losses have yet to be assessed. However, in countries like China, where growing populations and rising standards of living may soon begin to strain food production capacities, the effect of O_3 and related pollutants on agriculture could conceivably have adverse or even dire consequences (see *Food Security*).

4. Tropospheric Ozone and Global Change

The spread of O_3 pollution may extend beyond the regional scale to the continental, and even hemispheric scale. Analyses of late nineteenth and early twentieth century O_3 datasets indicate that tropospheric O_3 in the northern hemisphere has undergone significant variations over the past century and continues to vary in some regions of the globe today. The evidence for long-term variations in free tropospheric O_3 is strongest in Europe where measurements of O_3 date back to the 1800s. For example, German scientists Volz and Kley (1988) used current techniques for measuring O_3 to recalibrate O_3 measurements made in Paris in the late nineteenth century. Their analysis indicates that tropospheric O_3 concentrations at the end of the 1800s were about a factor of 2 less than current concentrations over Europe. Subsequent analyses have reached similar conclusions. The implications of such a long-term O_3 trend for understanding O_3 pollution effects on humans and ecosystems have yet to be fully explored but could be considerable.

Bibliography

Brasseur G P, Orlando J J, Tyndall G S 1999 *Atmospheric Chemistry and Global Change*. Oxford University Press, Oxford, UK

Brown L R 1995 *Who Will Feed China? Wake-up Call For a Small Planet*. Norton, New York

Chameides W L, Kasibhatla P S, Yienger J, Levy H II 1994 Growth of continental-scale metro-agro-plexes, regional ozone pollution, and world food production. *Science* **264**: 74–7

Chameides W L, Li X, Tang X, Zhou X, Luo C, Kiang C S, St. John J, Saylor R D, Liu S C, Lam K S, Wang T, Giorgi F 1999 Is ozone pollution affecting crop yields in China? *Geophysical Research Letters* **26**: 867–70

Crutzen P J, Zimmermann P H 1991 The changing photochemistry of the troposphere. *Tellus* **43AB**: 136–51

Fishman J, Kalish R 1990 *Global Alert: The Ozone Pollution Crisis*. Plenum, New York

Finlayson-Pitts B J, Pitts J N 1997 Tropospheric air pollution: Ozone, airborne toxics, polyclic aromatic hydrocarbons, and particles. *Science* **276**: 1045–51

Haagen-Smit A J 1952 Chemistry and physiology of Los Angeles smog. *Industrial and Engineering Chemistry* **44**: 1342–6

Heck W W, Cowling E B 1997 The need for a long term cumulative secondary ozone standard—An ecological perspective. *Environmental Management* **23**: 33

Lefohn A S 1992 *Surface Level Ozone Exposures and Their Effects on Vegetation*. Lewis, Chelsea, MI

Levy H II, Kasibhatla P S, Moxim W J, Klonecki A A, Hirsch A I, Oltmans S J, Chameides W L 1997 The global impact of human activity on tropospheric ozone. *Geophysical Research Letters* **24**: 791–4

Logan J A 1989 Ozone in rural areas of the United States. *Journal of Geophysical Research* **94**: 8511–32

Seinfeld J H, Atkinson R, Berglund R I, Chameides W L, Cotton W R, Demerjian K I, Elston J C, Fehsenfeld F, Finlayson-Pitts B J, Harriss R C, Kolb C E, Lioy P J, Logan J A, Prather M J, Russell A, Steigerwald B 1991 *Rethinking the Ozone Problem in Urban and Regional Air Pollution*. National Academy Press, Washington, DC

Seinfeld J H, Pandis S N 1998 *Atmospheric Chemistry and Physics*. Wiley Interscience, New York

Turco R P 1997 *Earth Under Siege: From Air Pollution to Global Change*. Oxford University Press, Oxford, UK

Volz A, Kley D 1988 Evaluation of the Montsouris series of ozone measurements in the nineteenth century. *Nature* **332**: 240–2

W. L. Chameides

Trust: Cultural Concerns

1. *The Concept of Trust in Modern Sociology*

The roots of the concept of trust are to be found in philosophy, theology, sociopolitical thought, and ethics (cf. Silver 1985, Misztal 1996). In this article the discussion will be limited to the domain of sociology, where during the last two decades of the twentieth century the new wave of theoretical concern with trust has emerged.

In 1979 Niklas Luhmann related the phenomenon of trust to the growing complexity, uncertainty, and risk characterizing contemporary society. For the first time, there was a suggestion that trust is not an obsolete resource typical of traditional society, but just the reverse, it gains in importance with the development of modernity. In 1983 Bernard Barber reviewed the manifestations of trust in various institutional and professional settings introducing the category of 'fiduciary trust.' In 1984 Shmuel Eisenstadt and Louis Roniger identified trust as a core ingredient in the patron–client relations, as they appear in various guises from antiquity to modernity. In 1988, Diego Gambetta brought together a number of authors looking at trust and distrust in various domains, from various perspectives, and later himself presented the analysis of trust in closed, exclusive communities, like the mafia (1993). In 1990, James Coleman provided an exemplary analysis of trust as a purely rational transaction, within the framework of rational-choice theory. This avenue was followed in a number of contributions in the 1990s by Hardin (1991, 1993). From a macrosociological perspective Anthony Giddens approached trust as the characteristic feature of late modernity, elaborating on Luhmanian themes of complexity, uncertainty, and risk. In 1995, Francis Fukuyama provided the comprehensive exposition of trust as an indispensable ingredient of viable economic systems, basing his argument on the experience of China, Japan, and other South-East Asian societies. In 1997, Adam Seligman presented an interpretation of trust as a specifically modern phenomenon linked with the division of labor, differentiation, and pluralization of roles and the consequent indeterminacy and negotiability of role expectations. In 1999, Piotr Sztompka offered a synthetic treatment of trust as a cultural resource necessary for viable functioning of society, illustrating his argument with the vicissitudes of trust in postcommunist societies of Eastern Europe.

The importance of trust derives from some fundamental qualities of human action. Interacting with others we must constantly articulate expectations about their future actions. Most often we lack the possibility of precise and accurate prediction or efficient control. Facing other people we remain in the condition of uncertainty, bafflement, and surprise. And yet, most often we cannot refrain from acting—to satisfy our needs, to realize our goals. Then we have to face risks that others will turn against us.

Trusting becomes the crucial strategy to deal with an uncertain, unpredictable, and uncontrollable future. Trust consists of two main components: beliefs and commitment. Placing trust in others, people behave 'as if' they knew the ways in which other people will act. But trust is more than just contemplative anticipation. People must also face the future actively, by committing themselves to action with at least partly uncertain and unpredictable consequences. Thus, people gamble, making bets about the future actions and reactions of partners. To sum up, *trust is a bet about the future contingent actions of others*.

2. *Targets of Trust*

We vest trust in various objects. First, there is trust in the members of our family, pervaded with strongest intimacy and closeness. Then comes the trust toward people we know personally, whom we recognize by name, and with whom we interact in face-to-face manner (our friends, neighbors, co-workers, business partners, etc.). Here trust still involves considerable degree of intimacy and closeness. The wider circle embraces other members of our community, known at most indirectly, by sight, and directly only through some individual representatives (inhabitants of our village, employees of our firm, professors at our university, members of our political party). The widest circle includes large categories of people, with whom we believe we have something in common, but who are mostly 'absent others,' not directly encountered, and constructed as a real collectivity only in our imagination ('imagined communities' of our compatriots, members of our ethnic group, of our church, of our race, of our gender, of our age cohort, of our generation, of our profession, etc.). Here trust in

concrete persons shades off imperceptibly into trust in more abstract social categories.

The next target of trust is social roles. Independent of the concrete incumbents, some roles evoke *prima facie* trust. Mother, friend, doctor of medicine, university professor, priest, judge, notary public—are just some examples of the trusted personal roles, or offices of 'public trust.'

An even more abstract case is the trust directed at institutions and organizations. The school, the university, the army, the church, the courts, the police, the banks, the stock exchange, the parliament, are typical targets for this type of trust. An interesting variant of trust in institutions may be called procedural trust. It is trust vested in institutionalized practices or procedures. A good example is trust in science as the best method for reaching the truth, or trust in the democratic procedures (elections, representation, majority vote, etc.) as the best ways to reach reasonable compromise among conflicting particular interests.

The next important category of objects endowed with trust are technological systems ('expert systems,' 'abstract systems,' cf. Giddens 1990). In modern society people live surrounded by them: by telecommunications, water and power systems, transportation systems, air-traffic control systems, military command networks, computer networks, financial markets, etc. The principles and mechanisms of their operation are opaque and cryptic for the average user. People usually take them for granted and do not even notice their pervasive presence. And everybody has learned to rely on them, to the extent that their failure produces a major crisis.

Finally, the most abstract objects of trust are the overall qualities of the social system, social order, or the regime. Trust in them engenders feelings of existential security, continuity, and stability.

The various types of trust reviewed above operate according to the same logic. Most importantly, behind all of them there looms the primordial form of trust—in people and their actions. Appearances notwithstanding, all of the above objects of trust, even most abstract, are reducible to human actions. Ultimately we trust human actions, and only derivatively their conglomerates, effects, or products.

3. *The Substance and Grounds of Trust*

The trusting expectations can be arranged along a sort of scale: from least demanding to the most demanding, and respectively from the weakest, least risky bets, to the strongest, most risky bets of trust. First, we may expect only some instrumental qualities of actions taken by the others: (a) regularity (orderliness, consistency, coherence, continuity, persistence), (b) reasonableness (giving grounds, good justification for actions, accepting arguments), (c) efficiency (competence, consistency, discipline, proper performance, effectiveness).

The second class of expectations is more demanding. We may expect some moral qualities of actions performed by the others: (a) we expect them to be morally responsible (i.e., engaging in principled, honest, honorable conduct, following some moral rules, showing integrity), (b) we expect them to be kind, gentle toward ourselves, treating us in human fashion, (c) we expect them to be truthful, authentic, straightforward, (d) we expect others to be fair and just (applying universalistic criteria, equal standards, due process, meritocratic justice). Generally speaking, betting on moral virtues of others is more risky than believing merely in their basic rationality.

We may also make the strongest bets and expect from others what Barber called the 'fiduciary' conduct and defined as 'duties in certain situations to place others' interests before our own' (Barber 1983). This category is exemplified by: (a) disinterestedness (i.e., acting without consideration of one's own interests or even against such interests), (b) representative actions (acting on behalf of others, displaying concern for the welfare of others, serving their interests, cf. Dahrendorf 1979), (c) benevolence and generosity (caring, helping, protecting, expressing sympathy, sensitive to the sufferings of others). This is the strongest, most risky bet because the probability that most people will be disinterested is low, and that they will take representative duties and engage in altruistic help is even lower.

There are three grounds on which decisions to trust (to place the 'bets') may be based: reflected trustworthiness, personal trustfulness, and trust culture.

As far as trust is a relationship with others, granting trust is based on the estimate of their trustworthiness. Trust in this case may be considered as 'reflected trustworthiness' of the partners: our perception of their reputation, performance, or appearance. The probability of well-placed trust rises with the amount and variety of true information about the trustee. Without such knowledge trust is blind and the chances of the breach of trust are high.

But trust is not only a calculating relationship, but also a psychological impulse (Wilson 1993). Innate trustfulness may push people to trust quite independently of any estimate of trustworthiness. This has nothing to do with knowledge about the partners of future engagements. Rather, the impulse derives from past history of diverse relationships pervaded with trust, primarily in the family and later in other groups, associations, or organizations.

People may also be encouraged to trust by the surrounding cultural rules. Normative rules may push toward trusting, defining trust as proper. If the rules demanding trust are shared by a community, and perceived as given, and external by each member, then they exert strong constraining pressure on actual acts of granting trust. They may modify significantly the

rational estimates of trust, as well as inherent trusting impulses.

4. Trust Culture

Trust culture is a system of such rules—norms and values—regulating and granting trust as well as reciprocating trust. There are the normative obligations to trust and there are the normative obligations to be trustworthy, credible, reliable. One locus of both types of obligations is the social roles. There are social roles that refer to trusters and include a normative imperative to trust others. This is true of 'helping professions': the doctor of medicine, the defense counsel, the social worker, the priest, etc. There are other social roles that refer to trustees and place strong emphasis on trustworthiness (the demand for meeting trust, i.e., acting reliably, morally, caringly). For example, university professors are expected to be truthful and responsible for their words, judges to be fair and just in their verdicts, football referees to be impartial. The more general rule of '*noblesse oblige*' demands exemplary conduct from those who have attained elevated positions in social hierarchy endowed with high esteem.

All those are role-specific rules of trust. But there are also more diffuse expectations to trust, which become pervasive in some societies at some periods of time. Fukuyama (1995) makes a distinction of 'high-trust cultures,' where he includes several countries of the Far East, and 'low-trust cultures,' where he includes some countries of the West. Putnam and Stivers complain about the demise of high-trust American culture of the nineteenth century and the emergence of the culture of cynicism in our time (Putnam 1995, Stivers 1994).

There are also culturally diffuse rules demanding and enforcing general trustworthiness. Mediaeval guilds, firms with long tradition, famous corporations, gold and diamond dealers, elite newspapers and journals, established publishing houses, etc, put great emphasis on fulfilling the obligations and meeting the trust of their clients. The 'pride of the profession' or the 'honor of the firm' become general normative guidelines embracing various sorts of activities.

Once the trust culture emerges and becomes strongly rooted in the normative system of a society, it becomes a powerful factor influencing decisions to trust, as well as the decisions to reciprocate trust. It may become the strong stabilizing force guaranteeing persistence and continuity of trust.

5. The Etiology of Trust Culture

Several macrosocietal circumstances may be postulated hypothetically as conducive to the emergence of a trust culture. The first is normative coherence, as opposed to normative chaos, or anomie. The norms—of law, morality, custom—provide the solid skeleton of social life and their effective enforcement assures their binding nature. This makes social life more unproblematic, secure, orderly, and predictable, as there are fixed scenarios indicating what people should do and will do. Such normative ordering of social life raises the likelihood that other people will meet our expectations. The resulting feeling of existential security and certainty encourages the bets of trust. But apart from that there are enforceable norms more immediately relevant for trust, demanding honesty, loyalty, and reciprocity. Their presence raises the likelihood of such conduct, and assures us that our partners will fulfill obligations and extend mutual trust.

The second structural condition is the stability of the social order, as opposed to radical change. If the network of groups, associations, institutions, organizations, and regimes is long lasting, persistent, and continuous, it provides the firm reference points for social life, a feeling of security, support, and comfort. Repeated routines which people follow allow us to predict their conduct. Similarly, meeting obligations and reciprocating trust becomes not so much a matter of duty, but rather an unproblematic, habitual response ('second nature'). People simply do not entertain the possibility that one could act otherwise. Trust may therefore be more easily offered, as the chances that it will be met, repaid, or mutually extended are high. By implication, social change is compatible with trust only if it proceeds gradually, regularly, predictably, in a slow rhythm, and consistent direction.

The third contextual, macrosocietal factor relevant to the propensity to trust is the transparency of the social organization, as opposed to its pervasive secrecy. The easy availability of information about the functioning, efficiency, levels of achievement, as well as failures and pathologies of groups, associations, institutions, organizations, and regimes provides the feeling of security and predictability. People are apt to relate to them with trust, because they are assured about what they may expect.

The fourth factor is the familiarity or its opposite the strangeness of the environment in which people operate. We mean by the environment the immediate 'life-world,' natural, technological, and civilizational milieu which surrounds the people. It includes various components: landscapes and topography, architecture, interiors, designs, colors, tastes, smells, images, etc. This factor, like the earlier one, has to do with accustomed routines. The experience of familiarity provides one with the feeling of security, certainty, predictability, and comfort. In effect, it produces a trust-generating atmosphere, where it is easier to believe that trusting predictions will be borne out, that entrusted values will be cared for and returned, and that others will reciprocate with mutual trust.

The fifth condition is the accountability of other people and institutions, as opposed to arbitrariness and irresponsibility. If there is a rich, accessible, and properly functioning set of institutions, setting standards and providing checks and controls on conduct, the danger of abuse is diminished, and the regularity of procedures safeguarded. If people can resort to such institutions when their rights are not recognized, or the obligations of others toward them not respected, then they acquire a kind of insurance or backup option and therefore feel safer. Everybody is confident that standards will be observed, departures prevented, and that even if abuse occurs it will be corrected by recourse to litigation, arbitration, restitution, etc.

6. *Functional Substitutes for Trust*

Trust has generally beneficial consequences for the partners in social relationships, the groups to which they belong, as well as the peaceful, harmonious, and cohesive quality of the wider social life. We may suspect that when trust is missing the resulting vacuum will be filled with some alternative arrangements providing similar functions and meeting universal cravings for certainty, predictability, and order. These will be the functional substitutes for trust.

The typical and widespread ways of coping with deficiencies of trust may acquire a normative sanction, turn into cultural rules prescribing certain conduct, or even into complex institutions. The danger is that some of these practices, strategies, and institutions may be clearly pathological. Appearing as functional substitutes to correct for the unfulfilled functions of trust, they themselves produce dysfunctional consequences for the wider society.

The first adaptive reaction is providentialism; invoking of supernatural or metaphysical forces—God, destiny, fate—as anchors of some spurious certainty. They are thought to take care of a situation about which nothing seemingly can be done. This may bring some psychological consolation, but at the social level it produces disastrous effects—apathy, passivism, and stagnation.

The second, quite perverse substitute for trust is corruption. Spreading in a society it provides some misleading sense of orderliness and predictability, some feeling of control over chaotic environment, some way to manipulate others into doing what we want them to do. The sane tissue of social bonds is replaced by the net of reciprocal favors, 'connections,' the cynical world of mutual manipulation and exploitation.

The third mechanism is the overgrowth of vigilance, taking into private hands the direct supervision and control of others, whose competence or integrity is put into doubt, or whose accountability is seen as weak, due to the inefficiency or lax standards of enforcing agencies. Private security forces, the walled communities with sentries, private possession of weapons, burglar alarms in cars and apartments, private agencies for debt collection—all those are clear indicators that trust has collapsed.

The fourth mechanism is excessive litigatiousness. If trust is missing, the handshake will no longer do. People will try to safeguard all relationships formally: draw meticulous contracts, insist on collaterals and bank guarantees, employ witnesses and notaries public, and resort to litigation in any, even miniscule, event of breaching trust by their partners.

The fifth mechanism may be called ghettoization, i.e., closing in, building impenetrable boundaries around a group in an alien and threatening environment. The diffuse distrust in the wider society is compensated by strong loyalty to local tribal, ethnic, or familial groups, matched with xenophobia and hostility toward foreigners. People close themselves in ghettoes of limited and intimate relationships, isolated and strictly separated from other groups, organizations, and institutions. By cutting the external world off, they reduce some of its complexity and uncertainty.

The sixth reaction may be called paternalization. When trust is missing, people seek protection in a father-like figure, a strong autocratic leader, a charismatic personality who would restore, if necessary by force, the semblance of order, predictability, and continuity in social life. When such leaders emerge they easily become a focus of blind, substitute trust. A similar craving for abdication of responsibility is also satisfied by other institutions: spreading cults, sects, 'voracious communities,' demanding full loyalty and total undivided commitment. They become quasi-families, with a strong substitute father taking full care of the members.

The seventh reaction may be called externalization of trust. In the climate of distrust against local politicians, institutions, products, etc., people turn to foreign societies and deposit their trust in their leaders, organizations, or goods. Such foreign targets of trust are often blindly idealized, which is even easier because of the distance, selective bias of the media, and lack of direct contrary evidence.

See also: Anomie; Conventions and Norms: Philosophical Aspects; Corruption: Political and Public Aspects; Interpersonal Trust across the Lifespan; Law: Overview; Rechtsstaat (Rule of Law: German Perspective); Solidarity: History of the Concept; Trust: Philosophical Aspects; Trust, Sociology of

Bibliography

Barber B 1983 *The Logic and Limits of Trust*. Rutgers University Press, New Brunswick, NJ

Coleman J C 1990 *Foundations of Social Theory*. Harvard University Press, Cambridge, MA

Dahrendorf R 1979 *Life Chances: Approaches to Social and Political Theory*. University of Chicago Press, Chicago
Eisenstadt S N, Roniger L 1984 *Patrons, Clients and Friends*. Cambridge University Press, Cambridge, UK
Fukuyama F 1995 *Trust: The Social Virtues and the Creation of Prosperity*. Free Press, New York
Gambetta D (ed.) 1988 *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell, Oxford, UK
Gambetta D 1993 *The Sicilian Mafia*. Harvard University Press, Cambridge, MA
Giddens A 1990 *The Consequences of Modernity*. Polity Press, Cambridge, UK
Giddens A 1991 *Modernity and Self-Identity*. Stanford University Press, Stanford, CA
Hardin R 1991 Trusting persons, trusting institutions. In: Zeckhauser R J (ed.) *Strategy and Choice*. MIT Press, Cambridge, MA, pp. 185–209
Hardin R 1993 The street-level epistemology of trust. *Politics and Society* **21**(4): 505–29
Luhmann N 1979 *Trust and Power*. Wiley, New York
Misztal B A 1996 *Trust in Modern Societies*. Polity Press, Cambridge, UK
Putnam R D 1995 Bowling alone: America's declining social capital. *Journal of Democracy* **6**(1): 65–78
Seligman A 1997 *The Problem of Trust*. Princeton University Press, Princeton, NJ
Silver A 1985 Trust in social and political theory. In: Suttles G D, Zald M N (eds.) *The Challenge of Social Control*. Ablex Publishers, Norwood, NJ
Stivers R 1994 *The Culture of Cynicism*. Blackwell, Oxford, UK
Sztompka P 1999 *Trust: A Sociological Theory*. Cambridge University Press, Cambridge, UK
Wilson J Q 1993 *The Moral Sense*. Free Press, New York

P. Sztompka

Trust: Philosophical Aspects

Trust is agreed to be essential for securing the goods of cooperation; however, philosophers disagree about how trust is best analyzed. On the simplest account, trust just is reliance. On more restrictive accounts, trust also requires something else: that the truster be disposed to respond to a breach of trust with reactive attitudes such as resentment; that the one trusted proves trustworthy because of being trusted; or that the reliance involved in the trust is grounded in an assumption of goodwill. In addition to the question of how to analyze the concept of trust, other central questions include how to generate trustworthiness, when trust is justified and whether trust and trustworthiness are virtues.

1. The Importance of Trust

Trust makes it possible for us to enter, willingly and with confidence, into relationships of dependency and, as Williams points out, 'if there are to be continuing practices of cooperation, then people must be motivated, one way or another, to enter into dependent positions' (1988, p. 8). Such dependencies may be economic (for example, when goods must be paid for prior to receipt), political (when power is entrusted to elected representatives), professional (when patients depend on their physicians), social (the mutual dependence of voluntary associations), personal (dependencies among family members), or epistemic (when laypersons defer to experts). Dependency brings with it risk: one may be left worse off than one would have been had one not depended. Trust makes such risks seem worth the taking.

Recent work in the social sciences (Putnam et al. 1993, Fukuyama 1995, among others) suggests that trust is an important form of social capital, the absence of which hampers economic development. Societies with high levels of trust present more economic opportunities than societies with low levels of trust. Moreover, trust enables transaction costs to be lowered by reducing the need for expensive substitutes for trust such as legal remedies and close monitoring of performance. Monitoring and legal remedy presuppose a relatively determinate understanding of what counts as an acceptable discharging of the responsibilities of a dependency relationship. Trust allows the truster to give trustees discretionary powers with respect to how they fulfill that trust (Baier 1986, p. 237). Discretionary powers greatly enhance the usefulness of dependency relationships: often the truster cannot precisely specify in advance what it is that they are counting on the other to do either because the truster does not have the ability to predict the contingencies that must be planned for (think of the discretion given to babysitters), or because they do not have the knowledge to specify what would count as an appropriate response to such contingencies. Subjective benefits also follow from assigning discretionary powers—trust lets us inhabit a simpler and less threatening world in which we need not plan for every contingency (Luhmann 1973). The limits of the trustee's discretionary powers are only vaguely specified and this vagueness opens up the possibility for abuse—an untrustworthy trustee might conceal subtle breaches of trust as appropriate uses of her discretionary powers. Assigning discretionary powers thus brings with it additional risks as well as benefits.

Trust is required for the proper functioning of divisions of economic and epistemic labor, divisions that can greatly increase the speed with which goods, including knowledge, are produced and, thus, the complexity and scope of the projects that can be undertaken. In these ways, trust is shown to have instrumental value insofar as it contributes to bringing about the goods of cooperation. However, trust has more than instrumental value since it is also constitutive of relationships that are valuable in themselves, such as the relationships between family members and friends.

This brief catalogue of some of the ways in which trust is important points to the significance of developing an account of what trust is, how trustworthiness can be elicited, and when trust is justified. These are the topics of the next three sections.

2. Accounts of Trust

While trust might appear to be a dyadic relation of the form 'A trusts B' (for example, the citizens trust their government), trust is best analyzed as a three place relationship: A trusts B to do Z (Baier 1986, Holton 1994). Even among close friends, trust seldom extends to all domains of interaction and there is typically some background assumption in place about that with respect to which the other is trusted. The triadic nature of trust shows that trust need not be an all or nothing affair—one can trust with respect to the performance of some actions, but not others. This opens up the possibility of incrementally increasing trust by gradually extending the range of actions that the other is counted on to do. Paradigmatically, the 'A' and 'B' of the schema 'A trusts B to do Z' stand for persons. However, non-natural agents, such as corporations and governments can also fill either role, as can perhaps nonhuman animals. Furthermore, on some accounts of trust, but not others, things such as machinery and ropes can occupy the role of B.

The single most important question that occupies the philosophical literature on trust concerns how and whether to separate trust from reliance. On the simplest model, trust just *is reliance*. On this model agents can trust machinery and ropes as well as other agents, natural and artificial. However, agents will count as trusting even when they do not expect a favorable outcome. For example, someone may decide to rely on a frayed rope which they do not believe strong enough to hold them—perhaps it is better to risk probable death or injury from a fall than certain death from the oncoming fire. A trust-as-reliance model conforms to ordinary linguistic usage, for it would not be incorrect to say that such a person preferred to trust the rope than to take their chances with the fire. Nevertheless, the trust-as-reliance model identifies too heterogeneous a class of dependencies to support useful generalizations and thus does not provide a useful classification for the purposes of social scientific and other theorizing. For example, the claim that trust contributes to social capital and reduces transaction costs is false if trust is understood as reliance with or without confidence. Someone who relies while expecting the worst will be on the lookout for alternatives to such reliance and will cease to rely once another option is open to them. Such wary and reluctant dependence cannot be thought of as social capital. If we are to have a category that supports useful generalizations, we need a richer conception of trust.

According to Thomas Hobbes, '[t]rust is a passion proceeding from the belief of him whom we expect or hope for good, so free from doubt that upon the same we pursue no other way' (1994: IX 9). Call this the *trust-as-confidence* model. Confidence can be grounded in all sorts of properties of that which is trusted. An agent might trust someone to do Z because they are motivated by self-interest, goodwill, fear, or just because they are in the habit of doing Z. Trust-as-confidence remains a comparatively thin conception of trust and can be enriched in one of several ways.

First, notice that on this model someone can trust when the other is unaware of their trusting and when the other acts from reasons not in anyway connected with the fact that they are being relied on. Consider Kant's habit of taking a walk at exactly the same time everyday (Blackburn 1998). Knowing his habits, his neighbors could use Kant to find out the time. Should we say that Kant's neighbors trust Kant for the time? Again, such usage would not be improper, but it seems to group together rather different phenomenon. Philosophers have tried different strategies to capture the difference between the kinds of phenomenon grouped together by this account. Hollis (1998) uses the distinction between predicative and normative expectations, or the distinction between expecting *that* and expecting *of*. I may expect that the rope will hold and may build that expectation into my plans. I may thus be said to trust, in the predicative sense, that the rope will hold. I may also expect of you that you will keep your promise and may build that expectation into my plans. In this case I may be said to trust, in the normative sense, that you will keep your promise.

Holton seeks to capture the same distinction in a different way: according to Holton, trusting essentially requires taking the 'participant stance' towards the one-trusted (1994, p. 66). The participant stance is characterized by the readiness to have reactive attitudes such as resentment or gratitude towards the other. While we might be annoyed at, or frustrated by things that disappoint us, we will not feel betrayed and resentful should they let us down. According to this account, trusting someone 'involves relying on them to do something, and investing that reliance with a certain attitude' (1994, p. 67). To return to the example of Kant: on the reliance-with-participant-stance account, as long as Kant's neighbors are treating Kant's walking habits as a mere sign of the time of day, much like they would treat a clock, and have no reason to believe that Kant is aware of their use of his clockwork schedule, then they rely on Kant for the time, but do not trust him to provide it. Suppose, however, that Kant came to believe that his neighbors were counting on his regular habits for their time-keeping, and the neighbors knew this—perhaps Kant promised to continue his helpful habits—then they might come to invest their reliance with the attitudes appropriate to the participant stance and become disposed to feel

resentful should Kant let them down. If so, then they would be trusting Kant for the time.

Hardin seeks to distinguish cases of trust from mere reliance in the following way: 'it is trust I have, and not merely inductive expectations *if* I think that you have a reason in some way grounded in me for fulfilling my trust. For example, I might well expect you to rise every morning at 7:00 because I know your habits very well, but I do not trust you to do so if I have no reason to suppose that you do so somehow on my behalf.' (1998, p. 12) The reference to 'on my behalf,' should not, however, be taken to indicate that Hardin thinks trust requires attributing to the one-trusted a direct concern for the truster's welfare. Hardin thinks that typically the one-trusted will have self-interested reasons for not letting the truster down.

Other accounts see reliance on self-interest as incompatible with trust. Baier analyses trust as reliance on the goodwill of the other, rather than on other features of their psychology or on external constraints (1986, p. 234). However, this account is vulnerable to the objection that, while the conman relies on his victims' goodwill in order to manipulate them, he cannot be said to trust them (Holton 1994, p. 67). The account would need to explicate the sort of reliance on goodwill that is distinctive of trusting reliance. A somewhat different account that shares the intuition that goodwill is central but that stresses the importance of affect or emotion, is offered by Jones (1996) who defines trust as an affective attitude of optimism that the goodwill and competence of the other will extend to cover the domain of your interaction, together with the confident expectation that the other will be directly and favorably moved by the thought that you are counting on them. Although many cases of trust undoubtedly do involve optimism about the competence and goodwill of another—for example, the trust between friends and family members takes this form—it seems that not all instances of trust involve an element of affect and not all presuppose goodwill. If accounts of trust that do not distinguish it from bare reliance are too broad, will-based accounts appear too narrow and appear to take as paradigm trust in interpersonal relationships rather than trust in governments and institutions. Thus, it is preferable to have a more general model as an account of trust in all its forms, while continuing to note that there can be important differences in the kind of trust that results from confidence in the good will of the one trusted and in their responsiveness to being relied on.

3. *Generating Trustworthiness*

If trusting is to bring about the goods of cooperation, it must be met with trustworthiness. Trust in the untrustworthy breeds exploitation and abuse. Moreover, the costs of misplaced trust are frequently borne by others more vulnerable than the truster—think, for example, of child abuse. For this reason, being trusting is not a virtue. Likewise, the endorsement of trustworthiness must be qualified: sometimes one is required to respond to trust with judicious untrustworthiness. Nevertheless, trustworthiness in the service of unobjectionable or praiseworthy ends is a good thing, and trustworthiness is almost always something it is valuable to increase.

The problem of generating trustworthiness can be brought into focus by considering the following games: consider two agents, A and B. A must choose whether or not to rely on B. If A chooses to rely, B must choose whether to exploit or fulfill that reliance. Suppose the payoffs are, and are known to be, as represented in Fig. 1 (Brennan 1998, p. 199). If B is rational, then, if A chooses to rely on B, B will choose to exploit that reliance, since the payoff of exploiting is higher than the payoff of fulfilling. But if A believes B to be rational, then A will know that B will choose the option with the highest payoff and thus will choose to exploit. Since the payoff for A of relying and being exploited (-1) is worse than the payoff of not relying at all (0), A, if rational, will choose not to rely. Thus, both players will be left worse off than they would be had they been able to rely and respond to that reliance with reliability, since they both miss out on the goods of cooperation (the (2,2) payoff).

Consider an alternative game which, following Brennan (1998, p. 200), we can call the 'dependence game.' As before, A must choose whether or not to rely on B and, if A chooses to rely, B must choose whether to exploit or fulfill that reliance. The difference between the two games lies in the payoffs, which are as illustrated in Fig. 2. With these payoffs, if A believes B to be rational, it will be rational for A to rely on B, and for B to respond to that reliance with reliability, since the payoff for B of fulfilling the reliance is greater than

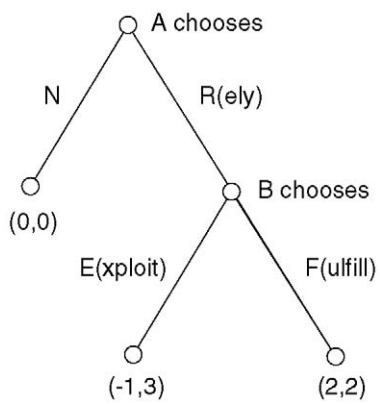

Figure 1
The reliance predicament

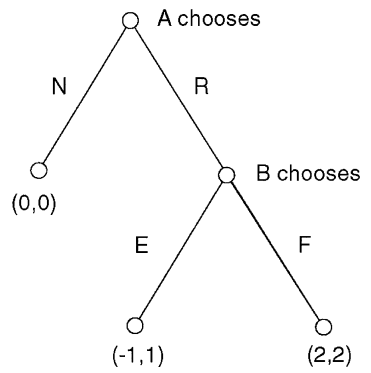

Figure 2
The dependence game

the payoff of exploiting it. Both agents can now achieve the goods of cooperation.

The problem of how to convert a reliance predicament into a dependence game is the problem of how to make B trustworthy with respect to the exchange in question. There are three ways to do this: by decreasing the payoff for B of exploiting (this is the strategy illustrated in Fig. 2), by increasing the payoff of cooperating, or by increasing and decreasing the relevant payoffs at the same time. Any of these solutions will work provided that B's payoff for fulfilling is greater than B's payoff for exploiting. The payoffs can be divided into subjective and objective components (Brennan 1998, p. 201). The subjective component refers to features of the agent's motivational set that will increase or decrease the utility she assigns to the outcome—examples include feelings of guilt at letting someone down and feelings of satisfaction at fulfilling their trust. The objective component refers to institutional incentives that increase or decrease objective utilities. Thus, each of the three strategies can be pursued either by adjusting the agent's subjective or objective payoffs or by pursuing a mixed strategy that affects both.

Among the strategies that can be pursued to structure the agent's objective payoffs are mechanisms for detecting and penalizing exploitative behavior. Thus, one way of building trustworthiness is to have in place mechanisms of distrust; mechanisms, that is, that provide checks against untrustworthy behavior.

Incentives are often cheaper than penalties. An important incentive is provided by the fact that, should B prove untrustworthy, A will not trust in future encounters. When there is the possibility of repeated plays of the game, the cost of losing the benefits of future cooperation can in itself be enough to adjust the payoffs towards trustworthiness. Thus Hardin (1991, 1998) can claim that trustworthiness is produced by self-interest; including especially by the trustee's interest in having an on-going relationship with the truster. Moreover, reputation for trustworthiness can encourage the reliance of others. For the reputational effect to be significant, background conditions must be met. In particular it must be possible for potential trusters to have information about the trustworthiness of potential trustees. Institutional design (such as structures that ensure openness and accountability) contribute to determining whether these background conditions are met. In these ways, institutional design can support (or erode) individual trustworthiness. While trust in individual actors cannot be entirely eliminated and replaced with trust in social institutions, social institutions can create a confluence of motives for trustworthiness. Working out which structures support, and which undermine, trustworthiness is thus a question of the first importance for institutional design.

It is widely observed (for example, Baier 1986, Jones 1996) that trust and distrust are self-confirming. Trusting is often met with perceived trustworthiness and distrust with perceived untrustworthiness. In part, this is to be explained by perception—someone who trusts will interpret the one-trusted's actions through the lens of that trust. An action that might have been seen as indicative of untrustworthiness will be explained away as compatible with trustworthiness. Equally, someone who distrusts will tend to interpret actions that could be given a favorable interpretation as nevertheless revealing motives indicative of untrustworthiness ('what does she want from me that she is being so nice?'). In part, it is to be explained by opportunities to prove oneself trustworthy: where possible, one steers clear of those one does not trust and so they are never given the chance to demonstrate their trustworthiness. But, in large measure, the self-fulfilling nature of trust and distrust is to be explained by trust-responsiveness (Pettit 1995) and its corollary distrust-responsiveness.

Trust-responsiveness is the 'disposition to prove reliable under the trust of others' (Pettit 1995, p. 203). Trust-responsive agents assign higher degrees of utility to performing actions that they are being counted on to perform than they would ascribe to those actions were they not being so counted on. The very act of relying, and conveying that reliance, itself contributes to adjusting the payoffs in the way needed to move from a reliance predicament to a dependence game. Trust-responsiveness can be grounded in a variety of elements in an agent's motivational set, including positively valued traits such as loyalty, virtue, and prudence. But it can also be grounded in agents' desire for the good regard of others. By openly relying, a truster can convey their assumption that the other is trustworthy. Since the desire to be thought well of by others is widely shared, conveying good-regard, or the promise of good-regard should the other fulfill the trust, is likely to be motivating (Pettit 1995). In this way, trust can be self-fulfilling.

Trust-responsiveness has a negative corollary in distrust-responsiveness. If others become aware of our

distrust, either because we refuse dependency, or because we engage in intrusive surveillance, we risk undermining the positive effect of trust-responsiveness and, further, we risk alienating the one being monitored and thus activating untrustworthiness. Surveillance generates resentment, undermines internal motivations towards trustworthiness, and encourages even the honest to attempt to beat the surveillance system (Kramer 1999, p. 591). Thus, strategies for converting the reliance predicament into the dependence game must be carefully crafted least adjustments in objective payoffs have unintended consequences on subjective payoffs and so offer no solution at all.

4. *Rational Trusting*

If the problem of trustworthiness can usefully be thought of as the problem of how to convert reliance predicaments into dependence games, the problem of rational trusting can usefully be thought of as arising when the assumption that the truster automatically knows what the trustee's payoffs are is abandoned, as it surely must be in real life. No matter how trustworthy the trustee would prove to be, the goods of cooperation will remain elusive unless it is reasonable for the truster to think that trust will in fact be met with a trustworthy response. Trust-responsiveness provides a mechanism which allows trusters to bootstrap their way to rational trust. Any means of communicating information about the potential trustee's motivations can also serve to ground rational trusting. Transparent social structures that reveal what incentives are in place can serve this communicative function, as can mechanisms, from gossip, to watchdog authorities, which spread abroad information about past performance with respect to trustworthiness. Would-be trustees have reason to convey that they can be trusted. Often, however, this cannot be conveyed directly: '"Trust me!" is for most of us an invitation which we cannot accept at will—either we do already trust the one who says it … or it is properly responded to with, "Why should I and how can I until I have cause to?"' (Baier 1986, p. 244). Trusted third parties can thus be important brokers for new trust relationships.

How much, and what sort of information agents need before it is rational for them to trust depends on the climate of trustworthiness they inhabit, on what it is that the other is being trusted to do, and on the cost of misplaced trust (Jones 1996). When the climate is sufficiently bad it might be reasonable to begin from a default stance of distrust and move towards trust only when there is considerable evidence of the trustworthiness of the other. In such climates, trust-responsiveness is undermined, for a truster cannot credibly convey to the trustee that she thinks them trustworthy if hardly anyone ever thinks another trustworthy (Pettit 1995, p. 221). Some actions, such as refraining from harming others as they go about their business, require only minimal decency and no particular competence and thus, unless the climate of trust has broken down completely, it can be assumed that trust is warranted. Other actions will be such that their performance demands competencies or motivational resources that cannot be assumed to be widely distributed—in such cases, the truster will require specific evidence about the trustworthiness of the other. Such evidence may take the form of marks that indicate that the trustee belongs to some subset of the population in which trustworthiness with respect to the required action will be the norm (for example, board-certified surgeons), or it may take the form of evidence about motivations and competencies specific to the trustee. The consequences of misplaced trust also affect how much evidence is required for rational trust—holding constant climate and kind of action to be performed, it will take more information before I am justified in trusting when I risk a grave loss should the other prove untrustworthy than when the loss is less.

5. *Trust and Epistemic Dependency*

Some of our most significant dependencies are epistemic: laypersons depend on experts for information about everything from the best cancer treatments, to the probable effects of an interest rate hike. Moreover, we rely on a vast array of technologies and intermediaries to convey this information. Nor is such trust limited to the relationship between experts and laypersons: experts must trust other experts in the design and maintenance of experimental equipment and must rely on the results of others in constructing and testing their theories. Hardwig reports the case of an experiment measuring the life of charm particles that took 280 person-years to perform and was reported in a paper with 99 authors (Hardwig 1991). No one person has, or could acquire, the expertise necessary to vouch for all the parts of the experiment even if it were physically possible for the experiment to have been conducted single-handedly. Such epistemic dependency was always present—consider day-to-day reliance on the testimony of others for directions—but the explosion of knowledge and of technologies for its dissemination, brings questions of epistemic dependency to the forefront of epistemology and raises questions about the design of knowledge generating and certifying institutions, such as the academy and the peer review process. Interdisciplinary inquiry into which structures support the truth-seeking aims of science and which undermine them is thus of the first importance (Goldman 1999) and can be seen as the extension to the realm of the epistemic, of the sorts of inquiries that are being undertaken into the trustworthiness of social and political institutions.

See also: Decision Theory: Bayesian; Decision Theory: Classical; Prisoner's Dilemma, One-shot and Iterated

Bibliography

Baier A 1986 Trust and anti-trust. *Ethics* **96**: 231–60

Blackburn S 1998 Trust, cooperation and human psychology. In: Braithwaite V, Levi M (eds.) *Trust and Governance*. Russell Sage Foundation, New York

Brennan G 1998 Democratic trust: a rational-choice theory view. In: Braithwaite V, Levi M (eds.) *Trust and Governance*. Russell Sage Foundation, New York

Fukuyama F 1995 *Trust: The Social Virtues and the Creation of Prosperity*. The Free Press, New York

Goldman A I 1999 *Knowledge in a Social World*. Clarendon Press, Oxford, UK

Hardin R 1991 Trusting persons, trusting institutions. In: Zeckhauser R J (ed.) *Strategy and Choice*. MIT Press, Cambridge, MA, pp. 185–209

Hardin R 1998 Trust in government. In: Braithwaite V, Levi M (eds.) *Trust and Governance*. Russell Sage Foundation, New York

Hardwig J 1991 The role of trust in knowledge. *The Journal of Philosophy* **88**: 693–708

Hobbes T 1994 [1650] Human nature. *The Elements of Law, Natural and Politic*. Oxford University Press, Oxford, UK, pt.1

Hollis M 1998 *Trust Within Reason*. Cambridge University Press, Cambridge, UK

Holton R 1994 Deciding to trust, coming to believe. *Australasian Journal of Philosophy* **72**(1): 63–76

Jones K 1996 Trust as an affective attitude. *Ethics* **107**: 4–25

Kramer R M 1999 Trust and distrust in organizations: emerging perspectives, enduring questions. *Annual Review of Psychology* **50**: 569–98

Luhmann N 1973 Vertrauen [Trust: a mechanism for the reduction of social complexity]. *Trust and Power*. Wiley, Chichester, UK, pp. 4–103

Pettit P 1995 The cunning of trust. *Philosophy and Public Affairs* **24**: 202–25

Putnam R, Leonardi R, Naretti R Y 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ

Williams B 1988 Formal structures and social reality. In: Gambetta D (ed.) *Trust: Making and Breaking Cooperative Relations*. B. Blackwell, New York

K. Jones

Trust, Sociology of

Trust is an expectation, a disposition, or a virtue. It is cognitive, emotional, moral, or some combination of the three. There is, in fact, no universally consensual conception, but by any definition trust involves a social relationship in which one person makes themself vulnerable to another who can do them harm if the trust is misplaced. That second person is either trustworthy or not, and much of the literature revolves around the evaluation of the trustworthiness of the trusted by the trustor. The two terms, trust and trustworthiness, are distinct but so implicated with each other that it is impossible to consider the first without also considering the second.

1. The Term 'Trust'

Trust is a word so common in everyday parlance that any effort to transform the term into a useful social science concept is fraught with difficulties. Yet, given the seeming importance of trust in interpersonal relations at home, at work, and on the street, in the relationships between citizens and their governors and governments, and even, possibly, among states, it is useful to clarify not only what we mean by trust but also to specify what work it actually does. To what extent trust is essential to well-ordered and productive relationships in contemporary societies is an empirical question as yet unresolved.

There are several distinct literatures on trust and several ways to categorize them. One typology differentiates trust among those who know each other intimately (personal trust) from trust among strangers or relative strangers (interpersonal, social, or generalized trust), which in turn is distinct from trust of government, firms, and other institutions (political, institutional, or organizational trust). Another involves the distinction between dyadic and embedded relations of trust. Yet another is according to disciplines, of which the most contributory to the discussion are philosophy, law, sociology, psychology, economics, and political science. There are approaches to trust that rely on rational choice, others that call upon structural functionalism, and yet others that evoke social identity theory. There are surveys of trust, experiments on trust, and formal models of trust. Yet, the concept remains elusive.

Some minimal consensus does exist, nonetheless. Trust is relational, involving at least one individual making themself vulnerable to another individual, set of individuals, or institution that possesses the potential to do harm or betray. Trust implies a judgment of risk in conditions of uncertainty. Moreover, trust is seldom unconditional; it is given over specific domains. A client may entrust a banker with money but never consider leaving a child in the banker's care; citizens may entrust their lives to their government during wartime but not new taxes to legislators who promise improved public services.

Trustworthiness is also relational but in a more limited sense. Even when there is no call for trust, a person or institution can possess the attributes of trustworthiness, attributes that assure potential trustors that the trusted will not betray a trust. Trustworthiness has two dimensions. The first involves the commitment to act in the interests of the trustor due to moral values about promise-keeping, caring about the

trustor, or incentive compatibility, or some combination of all three. When we call someone trustworthy, we often mean only this commitment, but there is in fact a second dimension, competence in the domain over which trust is being given. The trustworthy will not betray the trust as a consequence of either bad faith or ineptitude.

This minimal consensus does not resolve the conceptual ambiguity that surrounds the social science usage of the terms trust and trustworthiness. Confidence, reliance, and assurance are related concepts that some scholars conflate and others distinguish albeit in a variety of ways.

2. Research on Trust and Trustworthiness

Trust has a long intellectual history and a flourishing contemporary industry of research. In political philosophy, Thomas Hobbes argued that where there is trust, there is no reason to search for a better way to achieve a good, and John Locke emphasized the importance of mutual trust between the governors and the governed within legitimate polities (Dunn 1990). In sociology, it is a concept used by scholars as various as Talcott Parsons, Peter Blau, Harold Garfinkel, Niklas Luhmann, and Bernard Barber in their efforts to understand the construction of social order, cooperation, institutions, and organizations, and even everyday interactions. To the economist Arrow (1974), 'Trust is an important lubricant of a social system' because it can enhance efficiency. There are two major strands of work on trust, one that emphasizes trust as reigning where the market, contract, and modern institutions do not and the other emphasizing the underlying importance of trust in modern exchange, societal, and institutional relations. What ties these perspectives together is the recognition of trust as a means to cope with uncertainty; what distinguishes these perspectives is how trust comes about, when it should be relied upon, and what kinds of interactions it enhances.

The recent burgeoning of the social science literature on trust is a response to the perception that political and social trust is in decline and in part to an argument that trust is essential to a good society. The fact that there are now decades of responses to the same survey questions and that they indicate a reduction in both political and social trust in the advanced industrial democracies has generated an effort to explain and correct this trend. The emphasis on the importance of social capital and the claim that it, too, has decreased over time is another factor producing a focus on trust. Both Coleman (1990) and Putnam et al. (1993) included trust as a critical element in their original accounts of social capital.

Economists and organizational theorists have come to share Arrow's and others' view that more than formal incentives or credible commitments may be necessary to account for productive internal relationships and innovative change within firms and organizations (Kreps 1990, Miller 1992, Kramer and Tyler 1996). Social theorists attempting to explain variation in the capacity for collective action and cooperation (Gambetta 1988) and better governance (Braithwaite and Levi 1998) are also reconsidering the roles trust and trustworthiness play. And they are hardly alone in discovering, or rather revisiting, these concepts, as recent books by Fukuyama (1995), Seligman (1997), and others attest.

2.1 Motivational Assumptions About Trust

Why does someone trust? There are at least four possible kinds of motivations: dispositional, moral, social, and instrumental.

In one widespread perspective, trust is a disposition, that is, some individuals are more inclined to trust or be trustworthy than others due to factors of personality or character. A dispositional view of trust leads to investigations of variation among the population in the propensity to trust and be trustworthy. There is little agreement in the trust literature, however, on the sources of such a disposition. It could be a result of innate character, childhood experience, learning and updating, instrumental manipulation, knowledge, social networks, social bonds, or social norms. The assumption of trust as a disposition fails to capture important elements of trust: Trust always has both an object, the person, persons, or institution being trusted, and a domain, what the person, persons, or institution is being trusted to do.

By some accounts, trust is motivated by morality. Individuals are trusting because they believe trust is the act of a good person. Trust as a virtue is a very unsatisfying account of motivation, however. To trust a conman or to trust a corrupt politician is hardly virtuous by any common language sense of virtue. Trust may reflect a normative ordering that demands deference and obedience to certain kinds of authorities, but it is important to maintain the distinction between the normative features of trust and the claim that trusting behavior is either morally motivated or a moral good. What may well have moral motivation is trustworthiness, which generally implies commitments, promise keeping, and obligations. We are more likely to consider someone trustworthy, ceteris paribus, when we believe they have ethical values that make it unlikely for them to betray a trust.

A third possible motivation is social in the sense that individuals derive their identity and self-esteem from their relations to others and thus trust or are trustworthy in keeping with their sense of self or to enhance social status, or both. Although the social motivation is intentional, the end sought has more to do with identity than with the achievement of some material

benefit. This leads many of the adherents of social motivation to claim that it is neither interest-based nor calculative.

A final kind of motivation is, in fact, instrumental and calculative. Trust is given when it is a means to a particular end and when there is reason to believe the trust will be rewarded. There is a positive expectation that the payoffs from taking a risk are worth the possible costs. It is easiest to trust when the risks are sufficiently low, and it may well be worth the risk when the returns are sufficiently high. There are some circumstances, however, when the cost of misplaced trust is so high that trust is rarely given or given only when there is considerable insurance. Trusting one's child to the care of another or investing one's life savings are two such cases.

The development of a reputation for trustworthiness can also be instrumental. The trustworthiness of an individual may enlarge her possibilities for valued exchange and cooperation, and the trustworthiness of a government or organization may enhance the voluntary compliance and consent of those subject to its rules. Trustworthiness is also likely to increase the range of discretion given to the trusted, which is an issue of particular importance to agents, be they legislators or union officials or managers, acting on behalf of principals who lack the same intimate knowledge of the situation and the alternatives.

2.2 Conceptualizing Trust

In conceptualizing trust in order to engage in empirical research, two crucial distinctions exist: cognitive vs. noncognitive trust; dyadic vs. embedded trust. There is no empty cell in this matrix; a scholar can be concerned with either of the first and either of the second.

Assumptions about motivations are closely linked to whether the conceptualization of trust that informs empirical research is noncognitive or cognitive. Those who argue for motivations based on morality or social bonds tend to have a noncognitive approach, and those who argue motivations are instrumental generally have a cognitive view. The dispositional motivation can fit with either, depending on one's view of disposition formation.

The noncognitive conception of trust draws on the work of psychologists concerned with social identity, particularly Henri Tajfel and John C. Turner. It is advocated by contemporary social psychologists, such as Robyn Dawes, Tom Tyler, Rod Kramer, and Marilyn Brewer, and by social theorists, such as Allan Silver, who perceive trust as having a moral basis that contrasts with, although often complementing, the world of contracting and impersonal transactions. Individuals trust because of their attributions of motives to the person or institutions that must be trusted. If they attribute trustworthiness, they are more likely to trust, and they are most likely to attribute trustworthiness when there are social bonds or group identification between the trustor and the trusted. This perspective is rooted more in the emotions and morals of an individual than in her capacity to reason. It reflects personal relations and cannot be easily regulated by the impersonal relationships embodied in bureaucracy, market, or formal law. The cognitive conception of trust builds on the work of economists, political economists, and social exchange theorists, such as Kenneth Arrow, James S. Coleman Russell Hardin, and Diego Gambetta. The act of trust, of making oneself vulnerable, requires a thoughtful assessment of risk, informed by a rational expectation or a rational belief that the trusted will be trustworthy. Such an expectation or belief is generally grounded in knowledge of the other's past behavior or present motives and constraints. It can be affected by the existence of institutions, sanctions, and incentives that affect the probabilities that the trusted will act in the interest of the trustor.

The second distinction is between dyadic and embedded trust. Trust is relational in either case, but embedded trust permits conceptualization of generalized trust and institutional trust, and not just trust at the individual level. Dyadic trust depends upon expectations about or social bonds with a particular other. Embedded trust rests in social networks and institutional arrangements that permit expectations about or social bonds based on the social roles, categorization, and rules affecting others.

2.3 Measuring and Observing Trust

The most common methods for studying trust include surveys, experiments, and case studies, sometimes but not always informed by formal models. The surveys attempt to measure either political or social trust, but there is considerable controversy about how well they measure either. This is particularly true for the questions in the National Election Studies, which seem to be more about confidence in or positive assessment of elected officials than of political trust, and for those in the general social surveys, which seem to be tapping issues of social anxiety rather than social trust. The World Value Survey asks about both institutional and interpersonal trust, but its questions have similar limitations as the American versions. These major surveys do have the advantage of asking the same questions over the course of several decades, but there remains the issue of whether the time range covered is adequate to capture real changes in trust. Moreover, it is possible that people asked at one point in time understand a question differently than when they are asked at a later point in time. Surveys designed explicitly to capture the relationships of trust are even less likely than the major surveys to enhance understanding of changes in trust relations over time, and

their questions still tend to be open to multiple interpretations.

Some survey work does offer insight into trust relations, nonetheless. For example, Toshio Yamagishi uses survey evidence, often in combination with experimental work, to provide some interesting observations on variations in trust within and across populations and countries. Comparing the United States and Japan, Yamagishi and Yamagishi (1994) conclude that the embeddedness of the Japanese, at least until recently, in such tight social networks meant that they did not need to develop the capacity for discerning who and what to trust that has long been characteristic of Americans.

Experimental work on trust has different strengths and weaknesses. At their best, experiments can control for factors that might influence or be confused with trust and trustworthiness, they can investigate choice behavior and not just the attitudes captured by surveys, and they can be designed so as to clarify the conditions under which individuals are likely to trust. There have been interesting experiments linking trust to issues of exchange and collective action, and there has been some work attempting to get at the effect of dispositional variations within a population. Most of the experiments, however, focus on dyadic relationships, although there are experiments that investigate the operation of social networks. Moreover, much of the experimental work to date has conflated cooperation and trust. For these experiments to be more compelling, they need to clarify the roles that trust plays in producing cooperation rather than assuming that where there is cooperation, there is trust.

Several other approaches to the study of trust also deserve mention. The trust game and other formal models have proved useful to organizational theorists and economic historians attempting to understand the role trust plays in the relationships between principals and agents and in long-distance trade. There have been statistical analyses of indicators of trust in order to determine variation over time, and there have been case studies that attempt to capture the attributes of trustworthy government or trust that affects economic interactions and productivity in and across firms, industries, groups, regions, and countries. However, this research, although multiplying rapidly, is still in fairly early stages of development.

3. *Is Trust a Useful Concept?*

There is some question of whether a concept like trust is even necessary if the purpose is to explain phenomena that rely on impersonal, multi-person, and anonymous interactions. Williamson (1993) makes a case that it is not; to him trust exists only in personal relations with incentives, interests, and institutions regulating political, market, and large-scale social interactions. Silver (1985), coming from a totally different perspective, offers a similar argument. Hardin (1998) claims that trust, as he defines it, can exist only in the rather limited set of cases where the trustor has sufficient knowledge of the trusted to believe that their incentives are compatible; he is wary of such concepts as trust in government. Levi (2000) claims that trust is not always the solution to distrust. There are, after all, numerous functional alternatives to trust that rely more on assurance or coercion than on trust, such as: third-party insurers; contracts and other legal enforcement devices; professional ethics; fictive kinship; and a variety of social networks and institutional arrangements.

The argument of the social identity theorists and those who come out of the sociological tradition of Durkheim or the liberal political tradition of Locke is that trust most certainly matters and is, in fact, essential to a well-ordered society. It is constitutive, in this view, of the moral and personal glue that enables modern societies to function. These scholars tend to decry the loss of trust and seek to halt its further decline by advocating the maintenance and creation of community, republicanism, and more personalized interactions, even in large-scale organizations. On the other hand, those relying on instrumental and cognitive theories treat the work trust does as an empirical question. Relationships of trust are but one of many means to achieve given ends, and whether they are the best means is still open to considerable research and debate.

See also: Belief, Anthropology of; Control: Social; Emotion: History of the Concept; Emotions, Sociology of; Interpersonal Trust across the Lifespan; Organizations, Sociology of; Psychoanalysis in Sociology; Rational Choice Explanation: Philosophical Aspects; System: Social; Trust: Cultural Concerns; Trust: Philosophical Aspects; Value Pluralism; Values, Sociology of

Bibliography

Arrow K 1974 *The Limits of Organization*. Norton, New York

Braithwaite V, Levi M (eds.) 1998 *Trust and Governance*. Russell Sage Foundation, New York

Coleman J S 1990 *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge, MA

Dunn J 1990 Trust and political agency. In: Dunn J (ed.) *Interpreting Political Responsibility*. Princeton University Press, Princeton, NJ pp. 26–44

Fukuyama F 1995 *Trust: the Social Virtues and the Creation of Prosperity*. Free Press, New York

Gambetta D (ed.) 1988 *Trust: Making and Breaking Cooperative Relations*. Blackwell, New York

Hardin R 1998 Trust in government. In: Braithwaite V, Levi M (eds.), *Trust and Governance*, Russell Sage Foundation, New York, pp. 9–27

Kramer R M, Tyler T R (eds.) 1996 *Trust in Organization: Frontiers of Theory and Research*. Sage Publications, Thousand Oaks, CA
Kreps D M 1990 Corporate culture and economic theory. In: Alt J E, Shepsle K A (eds.) *Perspectives on Positive Political Economy*, Cambridge University Press, Cambridge, pp. 90–143
Levi M 2000 When good defenses make good neighbors. In: Menard C (ed.) *Institutions, Contracts, and Organizations: Perspectives from New Institutional Economics*, Edward Elgar, Northampton, MA, pp. 137–57
Miller G J 1992 *Managerial Dilemmas*. Cambridge University Press, Cambridge
Putnam R, Leonardi R, Nanetti R Y 1993 *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton, NJ
Seligman A 1997 *The Problem of Trust*. Princeton University Press, Princeton, NJ
Silver A 1985 Trust in Social and Political Theory. In: Suttles G D, Zald M N (eds.) *The Challenge of Social Control: Citizenship and Institution Building in Modern Society: Essays in Honor of Morris Janowitz*. Ablex, Norwood, NJ, pp. 52–67
Williamson O E 1993 Calculativeness, trust, and economic organization. *Journal of Law and Economics* **36**: 453–86
Yamagishi T, Yamagishi M 1994 Trust and Commitment in the United States and Japan. *Motivation and Emotion* **18**(2): 129–66

M. Levi

Truth and Credibility: Science and the Social Study of Science

The natural sciences occupy a special place in modern culture. Scientific knowledge substantially defines our culture's notion of truth, and the overwhelming credibility of that knowledge in our society is widely taken to be sufficiently explained by its truthfulness. The social sciences traditionally have concurred in these assessments and have paid the natural sciences the compliment of emulation. Yet that emulation largely has proceeded by way of textbook idealizations of natural scientific knowledge and how it is produced, while the idealizations themselves—produced by both academic epistemologists and reflective scientists—are much more variable than generally is supposed. It has not been until fairly recent times that social scientists have attempted wholly naturalistic inquiries into how scientific knowledge is produced and how it comes to be believed. The sociology of scientific knowledge that emerged from the early 1970s purposely set aside epistemological evaluation and took as its primary task understanding scientific credibility. How is it that scientific claims succeed, or fail, in being believed? This article surveys a number of considerations that recent social studies of science have found to be widely relevant to the interpretation of scientific credibility.

1. *Scientific Truth: As Evaluation and Resource*

'Truth' appears only occasionally as a vernacular term, nor does the special notion of intellectual quality the word connotes in academic philosophy figure much in everyday usages. Professional epistemologists worry how to reliably sort genuine knowledge from belief, opinion, or mere knowledge-claims, and this is the major context in which academic theorizing about truth takes place. Many theorists are agreed that knowledge (properly so-called) is 'true belief,' but there is no philosophical consensus about the definition of truth, about the domains in which we can expect to have genuine knowledge, about the grounds for establishing truth, or, indeed, about what it is that can be said to be true (propositions, facts, theories, thoughts, concepts, beliefs, etc.). The major battle pitches correspondence theories against coherence theories of truth, while skirmishes are fought along the margins between conventionalist, pragmatist, performative, redundancy, disquotational, and miscellaneous other theories (or antitheories) of truth.

In ordinary life, the word truth often crops up as a special-purpose term whose employment indicates that some out-of-the-ordinary guarantees are being sought for promises, professions, or claims to knowledge, or that there is some specially occasioned doubt about their genuine status, as in legal proceedings when witnesses are ceremonially sworn 'to tell the truth, the whole truth, and nothing but the truth' (Lynch and Bogen 1996), or in intimate social relationships when one has misgivings about one's lover's sincerity and looks for explicit warrants of affection: 'When my love swears that she is made of truth,/I do believe her, though I know she lies.' Perhaps even more commonly, everyday actors, and even philosophers in many circumstances, use the words true or truth just as a way of giving added emphasis or value to assertions of what is the case, the lexical equivalent of thumping the table.

The category of everyday actors may be taken, for the most part, to include professional natural scientists engaged in their ordinary knowledge-producing practices. While their expert employments generate knowledge whose content, methods of production, and value are widely said to differ markedly from lay beliefs about the world and lay modes of cognition and evaluation, epistemological reflection does not occupy much of scientists' time, nor is it customary for them to engage in metascientific inquiries into the nature and status of their knowledge. The result is a situation in which the bodies of expert knowledge that are widely taken as the touchstone of truth in our culture, and the object of vast amounts of epistemological reflection by philosophers, tend to be produced

by practitioners who rarely concern themselves reflectively or systematically with the nature of their knowledge, or its standing vis-à-vis nonexpert knowledge, or with the methods that produce their so-highly-valued knowledge. The paucity of practitioners' reflection on the nature and quality of their knowledge is an important feature of the modern natural sciences, and the social sciences and humanities are quite different in this respect.

Practicing natural scientists seem to take scarcely any notice of epistemologists' debates or dicta. Indeed, in the natural sciences, epistemological and metascientific musings, as well as sociological reflections, are often taken as a sign of emeritus (or, unkindly, 'past it') standing. They are sometimes explicitly condemned as irrelevant, unprofessional, or even pathological: it may be what philosophers or sociologists do, but (it is said) it is not what practicing scientists ought to do. In the nineteenth century the experimental physiologist Bernard (1865/1957, p. 49) vigorously disengaged the scientific production of truth from scientists' possible reflections on truth or on how truth was produced: 'To find scientific truth, we … have little need to know how our mind reasons; it is enough to let it reason naturally …'. Bridgman (1955, p. 157) was one of many modern physicists taking a jaundiced view of philosophers' attempts to characterize both scientific truth and scientific method: philosophers' 'abstractions' tended to attract a 'veneration' that was wholly undeserved. C. P. Snow—a one-time chemist and latterly a spokesman for science—bumptiously admitted that 'By *truth*, I don't intend anything complicated … I am using the word as a scientist uses it. We all know that the philosophical examination of the concept of empirical truth gets us into some curious complexities, but most scientists really don't care' (Snow 1961, p. 247). Albert Einstein famously celebrated, rather than condemned, what he took to be the fact 'that the man of science is a poor philosopher.' If, Einstein said (1936/1954, p. 296, p. 318), 'you want to find out anything from the theoretical physicists about the methods they use, I advise you to stick closely to one principle: don't listen to their words, fix your attention on their deeds.' Moreover, the small number of empirical studies we now have of natural scientists' grasp of the principles of formal logical reasoning strongly suggests that the power and reliability of scientific knowledge do not spring from the superior reasoning capacities of individual scientists (Mahoney 1979, Mahoney and DeMonbreun 1977).

When one considers the content of such scientists' musings about truth as are occasionally produced, further problems are posed for any social science wishing to model its epistemic goals on those of the natural sciences. Many academic, and some lay, views of the natural sciences point to scientific facts, laws, and theories as enjoying the status of accredited truth, and to the proper goals of scientific inquiry as the production of such truth. Yet important strands of practitioners' metascientific stipulations identify the goal of producing truth as illegitimate, and such special status for scientific representations as a lay misunderstanding of what genuine science is about. Some scientists do indeed say that science aims at, or even arrives at, one universal truth. However, others say that the truths of sciences are plural, or that science is just 'what works' and that truth—or even correspondence with the world—is none of their concern—just 'what stands the test of experience,' or 'what seems to be the case to the best of our current efforts and beliefs.' Oppenheimer (1954, p. 4) followed the nineteenth-century Scientific Naturalists in warning science to shun 'metaphysical' notions like 'real' and 'ultimate' in talking about its concepts and theories: if we say that physical theories are true, we must say so in a carefully qualified non-metaphysical sense. Einstein (1929/1954, pp. 261–2) found it 'difficult even to attach a precise meaning to the term "scientific truth." Thus, the meaning of the word "truth" varies according to whether we deal with a fact of experience, a mathematical proposition, or a scientific theory.' Bridgman (1955, p. 158) noted that 'In physics we have been forced to recognize … that there are different kinds of truth, depending on the method used for establishing "truth."' Bridgman's student Oppenheimer (1954, p. 95) vigorously agreed: 'Perhaps only a malignant end can follow the systematic belief that all communities are one community; that all truth is one truth.' Just as epistemologists espouse widely varying conceptions of truth, so too do those natural scientists occasionally drawn to philosophizing, while most practicing scientists seem well content to leave philosophy to the philosophers.

2. *From Truth to Credibility: Science as a Topic*

For these reasons alone, the posture of the social sciences vis-à-vis the natural sciences, and, especially, with respect to emulating their truth-criteria, is problematic. Should social scientists really wish to pattern themselves on the natural sciences in order to solve problems of intellectual identity, authority, and quality, they would find it no easy thing to do, such is the diversity of natural scientific practices and sensibilities, and of professional metascientific stipulations about the natural sciences (Mills 1959, pp. 119–31). This diversity is seldom adequately reflected in textbook accounts of scientific method and scientific knowledge, nor is it apparent when one particular natural science—typically some version of modern physics—is made by fiat to stand for all the rest (Shapin 1999a). Yet once this variability is properly appreciated, it becomes clear that the social sciences will have to make their own claims to knowledge and support them as best they can, using whatever resources and warrants are available and are deemed legitimate.

However, misguided emulation is not the only possible mode of social scientific engagement with the natural sciences. Social scientists can study the cultures of the natural sciences, seeking to understand how a range of natural scientists collectively produce their knowledge, how they make it persuasive, and how they render their intellectual evaluations. This has been the task taken up by the sociology of scientific knowledge over the past 30 years or so. Indeed, the founding gesture of modern sociology of scientific knowledge was the rejection of epistemological evaluation as the proper basis for sociological accounting. Traditional sociology of knowledge—descended from Marx and Mannheim—had reckoned, almost as a matter of course, that true and false beliefs should be explained differently. False beliefs might customarily be referred to distorting social and psychological factors, while the truth of true beliefs—their adequate evidential base, or their logical coherence, or their production and assessment using transparent and impersonal rational methods—could count as its own sufficient explanation (Mannheim 1936, pp. 239, 261, Phillips 1974). It was not considered that much of sociological interest could be said about true beliefs, and therefore about the body of currently accepted scientific knowledge. That is why traditional sociology of knowledge explicitly exempted accepted scientific beliefs from its scope.

Bloor's (1976/1991) *Strong Programme* in the sociology of scientific knowledge enjoined sociologists professionally to set aside external judgments of truth and falsity and to explain the acceptance of both reputed true and reputed false beliefs using the same resources. It was not that sociologists were to account all beliefs of equal value—that would itself be an act of epistemological assessment—rather it was that sociologists, in doing sociology, should be impartial with respect to the reputed truth or falsity of beliefs, as true and false seemed to the individual sociologist or as the beliefs were routinely sorted in the culture of which the sociologist was an everyday member. Bloor here identified generally acknowledged academic virtues that would flow from adhering to this maxim of method and academic vices that had traditionally stemmed from rejecting it. On the one hand, it was not the proper business of the social scientist to engage in epistemological assessment; on the other hand, inscribing external judgments of truth and falsity in social scientific accounts illegitimately blocked the free inquiry and rendered explanation circular. In Bloor's view, the sociologist could and should be as curious about why some cultures—in this case, our own—classify the cassowary as a bird as they had traditionally been about such cultures as the New Guinea Karam that classify what we call cassowaries and what we call birds in different taxa (Bulmer 1973, Barnes 1973, Barnes and Bloor 1982).

The *Strong Programme* did not mean to ban consideration of truth and falsity but, rather, to encourage sociologists to take a fuller naturalistic interest in the problem of how actors themselves sorted beliefs as between true and false. Actors' truth-judgments, or judgments of what is the case about the world, were not to be just an asymmetrical resource but a central topic of sociological inquiry. Or, put another way, this newer form of the sociology of knowledge placed the credibility of beliefs at the center of its intellectual agenda. Suppose, taking up the toy taxonomical example, one asked a member of the Karam why it was that they classified cassowaries in a different taxon than the rest of what we call birds. After translating the question into their own vocabulary, the Karam might reply in the prevalent folk-epistemological idiom, rolling their eyes, or saying that the matter is self-evident, or citing all sorts of obvious and important empirical differences between the big, ground-dwelling cassowary and the small, aerial bird-of-paradise. Pressed further, the Karam might point to the views of legitimately respected elders and experts, as sanctioned by custom and use. And so it might plausibly strike the Western social scientist that the grounds of credibility of Karam taxonomy had a lot to do with authority, custom, socialization, and with the vehicles that routinely transmit and maintain culture, including classificatory culture.

Patterns of routine socialization are not the only considerations in explaining structures of credibility—indeed, the anthropologist who most closely studied the particular case in question laid greatest stress upon the supposed functions of Karam taxonomy in sustaining social order (Bulmer 1973)—but socialization is everywhere pertinent to giving accounts of what people find credible, whether the socializing agencies are those of Karam agricultural society or departments of zoology in modern American universities (Barnes 1972, p. 271). So, if the sociologist were to put a parallel classificatory question to an academic practitioner of modern Linnaean taxonomy, it is arguable that the responses received would structurally resemble those supplied by the Karam, and, again, patterns of socialization would be major aspects of any explanation the sociologist might offer of the grounds of credibility of taxonomic beliefs in our own expert culture. Note that the empirical adequacy of Linnaean classification is not thereby disparaged—though there is substantial controversy about its theoretical and evidential adequacy within expert taxonomic culture—but the sociologist of scientific knowledge holds that standards of empirical adequacy, and the particular goals towards which classificatory schemes are oriented, are themselves objects of socialization (Dean 1979). All institutionalized beliefs about nature are causally connected to reality, and 'all beliefs are on a par with one another with respect to the causes of their credibility. It is not that all beliefs are equally true or equally false, but that regardless of truth or falsity the fact of their credibility

is to be seen as equally problematic' (Barnes and Bloor 1982, p. 23, Barnes 1992). The 'paradigms' of what Kuhn called 'normal science' are exemplary items of technical knowledge, and they are also learned bits of culture. Since this culture is transmitted in authoritative institutions, the sociologist's whole range of resources for dealing with socialization and its consequences is, therefore, pertinent to explaining the credibility of technical scientific knowledge among expert practitioners (Polanyi 1958, Kuhn 1962, Barnes 1982).

3. *Accounting for Credibility: Mundaneness and the Moral Order*

So the sociologist of scientific knowledge studies credibility. What actors find credible, and how they account it credible, may include specially occasioned judgments of truth or they may not. The distinction is not crucial, since the credibility of any such reflective truth-judgments (by way of correspondence or coherence or any other criterion) may also be a matter for inquiry. The sociologist, qua sociologist, here differs from the philosopher (and maybe from the layperson): 'Instead of defining (knowledge) as true belief, knowledge for the sociologist is whatever people take to be knowledge,' what they find credible (Bloor 1976/1991, p. 5). And patterns of socialization are everywhere pertinent to giving a sociological account of the credibility of beliefs. These are, so to speak, ground-conditions for the sociological study of science, and, as such, they are not themselves answers but necessary conditions for arriving at particular answers to particular questions about particular bodies of knowledge—answers that are sociologically proper, constructive, and interesting. For sociologists to say that they are interested in credibility is not to say that they are concerned with a specific aspect or 'factor' bearing on knowledge. Credibility is not a factor in the sense in which, for traditional sociology of knowledge, 'class-allegiance' or 'religious affiliation' counted as 'social factors' contingently bearing upon knowledge—sometimes pertinent, sometimes not (e.g., Barber 1952, p. 60). Credibility is the state achieved when knowledge-claims become knowledge, and the means for achieving credibility are the assemblage of means by which knowledge-claims become knowledge. From a pertinent point of view, credibility is the only topic for the sociologist of knowledge. As the philosophers might say, credibility has a transcendental character for sociologists of knowledge: no credibility, no knowledge, no object of sociological study (Shapin 1995a, pp. 257–8).

It might also be said, no credibility, no moral order. To give an account of what a community finds credible, and how it comes to find it credible, is tantamount to giving an account both of what that community is and how it experiences the world (Douglas 1975, p. 238). So to say that sociologists of science should shift attention from truth to credibility is a way of orientating practitioners to the proper domain of inquiry, a way of discouraging them from arbitrarily limiting inquiry, and, finally, a way of reminding them of the potential pertinence of a range of social scientific and everyday resources that are available to give accounts of what people find credible.

No matter how highly valued or how technical the form of belief, the sociologist can usefully draw attention to the mundane means by which belief is made credible. In the ethnomethodological tradition, this was a maxim of method derived from the phenomenologists (especially Alfred Schutz) and from the later writings of Wittgenstein. Attention was to be directed to the 'behavior of seeking truth … the institutional and public character of truth, in contrast to the usual psychological and semantic descriptions that depict private disembodiments of that behavior.' For the sociologist, there is no other proper way of engaging with truth save through the study of what people collectively do: 'Truth resides in the rule-guided institutional procedures for conceding it.' In doing sociology, one should accept 'that there are no adequate grounds for establishing criteria of truth except the grounds that are employed to grant or concede it' (McHugh 1970, pp. 320–1, 333–5). This sensibility strongly influenced the ethnographic study of science that began to emerge in the 1970s, especially with the work of Collins and co-workers (Collins 1975, 1983, 1985/1992), but it also left a mark on historical studies of science concerned with how past communities established facticity, proof, and legitimate inference (e.g., Shapin 1994, Dear 1995, Porter 1995, Poovey 1998, Burke 2000).

If, indeed, the practices for making and maintaining credible knowledge are co-extensive with the moral life of a community, then there are very few aspects of that life which do not potentially bear upon the credibility of belief. It follows that there can be no simple and coherent master-recipe for how credibility is to be achieved or model for understanding it. Much will depend upon the items of knowledge in question, the precise scenes in which the knowledge is assessed, and the state of transmitted background belief. Truth—like God and the devil—is in the details, and the sociologist wanting to understand why it is that a specific individual, or group of individuals, finds the special creation of species, or Darwinian evolution, credible may find it pertinent to consider a long list of mundane, and even some apparently trivial, features of cognition and social interaction (Shapin 1999b, pp. 7–8, Lynch 1993, pp. 278–88), while much work in the sociology of scientific knowledge during the 1970s and 1980s pointed to differences in collective goal-orientations (or social interests) as pertinent considerations in judgments of the credibility of scientific claims (e.g., Shapin 1982, 1995b).

4. Accounting for Credibility: Economies of Trust

A large body of recent historical and sociological work has studied the means by which the veracity of scientific claims or the legitimacy of inference has been made persuasive. Because most natural knowledge—that possessed by experts as well as by the laity—is held on the basis of trust, deferring to others whose knowledge is understood to be more direct or competent, even belief in matters of fact has to be actively secured: how persuasively to communicate the experience to be credited? how linguistically (and gesturally) to display one's entitlements to trustworthy knowledge? Studies of both historical and present-day episodes have drawn attention to the rhetorical constitution of authorial transparency, virtue, and intellectual entitlements; to the rhetorical invocation of allies leagued with the author and of the consequences of giving or withholding assent; and to the range of rhetorical means by which authors in general seek to assist the translation of individual claims into collectively credible knowledge (Latour 1987, Shapin 1984, 1994; Dear 1995, Golinski 1998, Chap. 4).

Recognizing the ineliminable role of trust in the credibility of knowledge is consequential for two diametrically opposed reasons. First, many situations present themselves—both within science and between scientific experts and the laity—in which the relative trustworthiness of sources is in doubt or is contested. The role of scientific expert testimony in courts of law, for example, is one important setting in which lay jurors are set the task of finding the signs of credibility in conflicting scientific accounts, and some studies emphasize the role of common-sense procedures in that task—monitoring testimony for manner and its sources for biasing interest (Brannigan and Lynch 1987, Lynch and Jasanoff 1998). Wynne's (1995) work on the public understandings of science, and specifically on the credibility of environmental radioactivity risk-assessments, stresses the complex array of considerations that bear upon lay attitudes towards which scientists, and which scientific pronouncements, are trustworthy. Epstein's (1996) study of the relations between AIDS researchers and AIDS activists is a richly detailed investigation of the grounds of trust in a situation where the very identity of laypersons and experts has been made problematic. Porter (1995) has asserted the crucial role of 'trust in numbers' in modern, complex, and highly differentiated societies, in which, as he argues, personal and familiar warrants for credibility are lacking or deemed inadequate. Jasanoff has examined how trust in expertise is secured in modern environmental science where conflicts over credibility are endemic (Jasanoff 1992), and Hilgartner's (2000) work on a series of highly contested nutritional reports by the US National Academy of Sciences has drawn attention to the performative aspects of trust in modern expertise.

In the passages of modern 'extraordinary science' studied by Collins (1975), where the existence of specific physical phenomena and the competence of detecting instruments were simultaneously contested, members of a scientific 'core-set' reached out for a diverse array of contingent considerations to locate trustworthiness, and, therefore, to assess the credibility of technical claims: including the personal characteristics of claimants, the reputation of the institutions in which they worked, their allies, their track-records, etc. Historical studies of scientific controversy (Shapin 1994, esp. Chap. 6) have drawn attention to the imputed trustworthiness of the early modern gentleman in managing scientific controversy, in deciding the credibility of diverging claims, and even in deciding which controversies ought to be left undecided in favor of maintaining the interaction order.

In such lines of research, reposing trust in scientific claims and claimants appears as an actor's practical problem. Looking to see how and why actors reckon trustworthiness is a way of understanding how they deal with problematic credibility when experts disagree or when it is unclear who indeed is to count as an expert. There are plenty of such occasions in modern science, and especially in those sectors of science in which public participation in decision-making is mandated or pertinent. Is global warming a fact, and, if so, what is to be done about it? Who is to be believed about the risks of genetically modified foods and what policies should be framed for producing and labeling such foods?

Yet the second reason why trust is important in connection with scientific credibility is that the scientific community very often does speak—if not with one voice—then in a comparatively concerted way. Scientists in modern society are regarded as enormously trustworthy, and the credibility of what they concertedly deliver about the natural world is very great. Scientists are masters at trusting each other and they are masters at generating public trust in themselves. That is just another way of saying that consensual scientific knowledge is regarded in our society as the touchstone of truth. Scientific credibility may be an occasional problem, and a vitally important one, but it is also interesting to note how often in our culture the credibility of science does not manifest itself as problematic. Sociologists do not yet adequately understand how scientists achieve and maintain this trustworthiness and the resultant powers of professional credibility (Barnes 1985, Chaps. 2–3), but the consequences of this trustworthiness for the social and intellectual order of modern society cannot easily be exaggerated (Bauman 1992, p.71, Shapin 1995a). To understand the grounds of scientific credibility is to understand quite a lot about modern society. Scientific credibility is so blended into modern society's institutional life—its schools, universities, government, medicine, the military, and the media—that one cannot readily imagine credible science without the

institutions which maintain and transmit it, nor credible institutions without the knowledge which gives them much of their legitimacy.

See also: Clinical Psychology: Validity of Judgment; Empiricism, History of; Experimental Design: Overview; Experimentation in Psychology, History of; Expert Systems in Cognitive Science; Explanation: Conceptions in the Social Sciences; History of Science; Kuhn, Thomas S (1922–96); Scientific Knowledge, Sociology of; Social Science, the Idea of; Theory: Conceptions in the Social Sciences; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

Barber B 1952/1962 *Science and the Social Order*. Free Press, Glencoe, IL

Barnes B 1972 On the reception of scientific beliefs. In: Barnes B (ed.) *Sociology of Science: Selected Readings*. Penguin, Harmondsworth, UK, pp. 269–91

Barnes B 1973 The comparison of belief systems: Anomaly versus falsehood. In: Horton R, Finnegan R (eds.) *Modes of Thought: Essays on Thinking in Western and Non-Western Societies*. Faber & Faber, London, pp. 182–98

Barnes B 1982 *T. S. Kuhn and Social Science*. Macmillan, London

Barnes B 1985 *About Science*. Blackwell, Oxford, UK

Barnes B 1992 Realism, relativism and finitism. In: Raven D, van Vucht Tijssen L, de Wolf J (eds.) *Cognitive Relativism and Social Science*. Transaction, New Brunswick, NJ, pp. 131–47

Barnes B, Bloor D 1982 Relativism, rationalism and the sociology of knowledge. In: Hollis M, Lukes S (eds.) *Rationality and Relativism*. Blackwell, Oxford, UK, pp. 21–47

Bauman Z 1992 *Intimations of Postmodernity*. Routledge, London

Bernard C 1865/1957 *Introduction to the Study of Experimental Medicine* (trans. Greene H C). Dover, New York

Bloor D 1976/1991 *Knowledge and Social Imagery*, 2nd edn. University of Chicago Press, Chicago

Brannigan A, Lynch M 1987 On bearing false witness: Credibility as an interactional accomplishment. *Journal of Contemporary Ethnography* **16**: 115–46

Bridgman P W 1955 *Reflections of a Physicist*, 2nd edn. Philosophical Library, New York

Bulmer R 1973 Why the cassowary is not a bird. In: Douglas M (ed.) *Rules and Meanings: The Anthropology of Everyday Knowledge*. Penguin, Harmondsworth, UK, pp. 167–93

Burke P 2000 *A Social History of Knowledge: From Gutenberg to Diderot*. Polity, Oxford, UK

Collins H M 1975 The seven sexes: A study in the sociology of a phenomenon, or the replication of an experiment in physics. *Sociology* **9**: 205–24

Collins H M 1983 The sociology of scientific knowledge: Studies of contemporary science. *Annual Review of Sociology* **9**: 265–85

Collins H M 1985/1992 *Changing Order: Replication and Induction in Scientific Practice*, 2nd edn. Sage Publications, Beverly Hills, CA

Dean J 1979 Controversy over classification: A case study from the history of botany. In: Barnes B, Shapin S (eds.) *Natural Order: Historical Studies of Scientific Culture*. Sage, Beverly Hills, CA, pp. 211–30

Dear P 1995 *Discipline and Experience: The Mathematical Way in the Scientific Revolution*. University of Chicago Press, Chicago

Douglas M 1975 *Implicit Meanings: Essays in Anthropology*. Routledge & Kegan Paul, London

Einstein A 1929/1954 Scientific truth. In: Einstein A (ed.) *Ideas and Opinions*. Crown Publishers, New York, pp. 261–2

Einstein A 1936/1954 Physics and reality. In: Einstein A (ed.) *Ideas and Opinions*. Crown Publishers, New York, pp. 318–56

Epstein S 1996 *Impure Science: AIDS, Activism, and the Politics of Knowledge*. University of California Press, Berkeley, CA

Golinski J V 1998 *Making Natural Knowledge: Constructivism and the History of Science*. Cambridge University Press, Cambridge, UK

Hilgartner S 2000 *Science on Stage: Expert Advice as Public Drama*. Stanford University Press, Stanford, CA

Jasanoff S 1992 Science, politics, and the renegotiation of expertise at EPA. *Osiris* **7**: 195–217

Kuhn T S 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Latour B 1987 *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, Cambridge, MA

Lynch M 1993 *Scientific Practice and Ordinary Action: Ethnomethodological and Social Studies of Science*. Cambridge University Press, Cambridge, UK

Lynch M, Bogen D 1996 *The Spectacle of History: Speech, Text, and Memory at the Iran-Contra Hearings*. Duke University Press, Durham, NC

Lynch M, Jasanoff S 1998 Contested identities: Science, law and forensic practice. *Social Studies of Science* **28**: 675–86

McHugh P 1970 On the failure of positivism. In: Douglas J D (ed.) *Understanding Everyday Life: Toward the Reconstruction of Sociological Knowledge*. Aldine, Chicago, pp. 320–35

Mahoney M J 1979 Psychology of the scientist: An evaluative review. *Social Studies of Science* **9**: 349–75

Mahoney M J, DeMonbreun B G 1977 Psychology of the scientist: An analysis of problem-solving bias. *Cognitive Therapy and Research* **1**: 229–38

Mannheim K 1929/1936 *Ideology and Utopia: An Introduction to the Sociology of Knowledge* (trans. Wirth L, Shils E). Kegan Paul, Trench, Trubner & Co., London

Mills C W 1959 *The Sociological Imagination*. Oxford University Press, London

Oppenheimer J R 1954 *Science and the Common Understanding: The BBC Reith Lectures 1953*. Simon and Schuster, New York

Phillips D L 1974 Epistemology and the sociology of knowledge: The contributions of Mannheim, Mills, and Merton. *Theory and Society* **1**: 59–88

Polanyi M 1958 *Personal Knowledge: Towards a Post-critical Philosophy*. University of Chicago Press, Chicago

Poovey M 1998 *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. University of Chicago Press, Chicago

Porter T M 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ

Shapin S 1982 History of science and its sociological reconstructions. *History of Science* **20**: 157–211

Shapin S 1984 Pump and circumstance: Robert Boyle's literary technology. *Social Studies of Science* **14**: 481–520
Shapin S 1994 *A Social History of Truth: Civility and Science in Seventeenth-century England*. University of Chicago Press, Chicago
Shapin S 1995a Cordelia's love: Credibility and the social studies of science. *Perspectives on Science* **3**: 255–75
Shapin S 1995b Here and everywhere: Sociology of scientific knowledge. *Annual Review of Sociology* **21**: 289–321
Shapin S 1999a Être ou ne pas être antiscientifique. *La Recherche* **319**(April): 72–9
Shapin S 1999b Rarely pure and never simple: Talking about truth. *Configurations* **7**: 1–14
Snow C P 1961 Address by Charles P. Snow (to Annual Meeting of American Association for the Advancement of Science). *Science* **133**: 256–9
Wynne B 1995 Public understanding of science. In: Jasanoff et al. (eds.) *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA, pp. 361–88

S. Shapin

Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

The point of an inquiry is to get at the truth—not the whole truth, but the whole truth of some matter at hand. This truism generates the central problems of epistemology: What is truth? What kinds of things are true? How can we get at truth, or know when we have attained it? Will any old truth do, and is any old falsehood a failure?

1. *Truth and its Bearers*

Truth seems to be a simple affair. That snow is white is *true* if … well, snow is white! That snow is hot is *true* if … snow is hot! We know what it would take for snow to be white. A certain kind of stuff—flaky frozen water—has to have a certain appearance. So we know what it takes for it to be *true* that snow is white, and since the stuff has the appearance, it is true that snow is white. One could be forgiven for thinking that there's an end of it. The so-called redundancy theory of truth elaborates this insight (Ramsey 1927, Horwich 1990) although it is also the main motivation behind the so-called *correspondence* theory.

But what exactly are the bearers of truth and falsity? One answer is *sentences*. Sentences (grammatically correct strings of letters like 'Snow is white') appear familiar, unproblematic, concrete. Another answer is *propositions* (or whatever it is that sentences *mean*). These appear more abstract and remote than sentences. The 'linguistic turn' in philosophy, a widespread presumption against abstract entities, and Tarksi's ground-breaking definition of sentential truth (Tarski 1969), have all favored sententialism.

How many sentences are there on the next line?:

Snow is white. SNOW IS WHITE. *Snow is white*.

The answers, *one* and *three*, are both defensible. We distinguish *sentence types* (of which there is just one here) from *sentence tokens*, individual instances of a sentence type (of which there are three here). A type is not identical with any of its instances, or with a set of them. A set cannot change its members—its identity is given by its members—whereas the number of instances of a sentence type is in constant flux. A type, even a sentence type, is an abstract entity. That 'Snow is white' is true is a claim about the sentence *type*. So eschewing meanings in favor of sentences does nothing to rid the universe of abstract entities (Church 1956).

That 'Snow is white' is true is a fact about English, but the fact that snow is white is not. 'Snow is white' expresses a truth which can be expressed in other languages, suggesting *language-independent truths* (Frege 1984). Given continuous space–time the trajectory of a single particle generates nondenumerably many truths. Since any learnable language contains at most denumerably many sentences, the totality of truths outstrips the totality of sentences of any learnable language. So the *primary* bearers or truth cannot be sentences. Without meaning, it is obscure what the truth of a sentence would consist in. 'Snow is white' is true *in virtue of* its expressing something which in fact obtains. A meaningful sentence expresses a meaning and that, in conjunction with the state of the world, determines whether it is true or false.

2. *Verifiability*

To determine whether 'Snow is white' is true one must know what it means. And to know what it means one must know what the world would look like were it true. So meaning seems related to an ability to *ascertain* truth. The *verificationist theory of meaning*, the battle cry of the logical positivists, articulates this conviction: that a sentence is meaningful if and only if it can be verified by observational means (Schlick 1979).

The positivists took their inspiration from Einstein's analysis of simultaneity. In his 1905 relativity paper, Einstein argued that claims about simultaneity are physically meaningless until we agree on a procedure for determining simultaneity. Einstein supplied a verification procedure for 'A is simultaneous with B,' thereby precipitating a fundamental revolution in the physics of space–time. The positivists co-opted the approach to effect a similar revolution in philosophy (Schlick 1979). (Einstein later repudiated his early verificationism with the now famous put-down: 'One shouldn't tell a good joke twice.')

The verificationist theory serves the twin desiderata of positivism: endorsing science while disenfranchising

metaphysics, religion, and (perhaps embarrassingly) ethics. Sentences reporting observations ('There is a red patch in my visual field') are paradigmatically meaningful. Sentences like 'The Nothing nothings,' 'God is omnipotent,' and 'Torture is wrong' are meaningless, since there are no possible observations which could verify them.

Verificationism dovetails nicely with an uncompromising epistemology. A true sentence must be meaningful, hence verifiable, and so knowable. The facts cannot outstrip our means of discovery. This consequence attracted devastating criticisms, mostly leveled by the rigorously self-critical positivists themselves. For example, the theory deems scientific theories themselves meaningless, since they make wildly general claims. Newton's gravitational theory, for example, claims that any two masses anywhere in space–time attract each other. All the observational data that we will ever have is but a tiny fragment of the plethora of predictions the theory makes. Even a simple claim like 'metals expand when heated' goes beyond any finite amount of data. For this reason, scientific theories are unverifiable (Popper 1963).

This is a variation on Hume's notorious problem of induction. Whenever we make a prediction about the unobserved we go beyond what has been verified. But it is precisely because it does this that science is so interesting and useful. Science at its best gives us *depth*—profound and contentful theories which go far beyond the actual and possible data—and depth pulls away from the positivists' goal of *certainty* (Watkins 1984).

The positivists made special provisions for logical and mathematical truths, as well as philosophical claims like verificationism itself. These are not established by observation, for they are not *factual* claims at all. Rather, they are *analytic*—true in virtue of the meanings of constituent terms—like 'Bachelors are unmarried males.' Some can be obtained by analysis of meaning, others by deduction from established analytic truths. Deduction is justified by appeal to the meanings of logical terms such as 'and,' 'not,' 'all.'

A first cousin of verificationism is thus the doctrine that mathematical truth and provability coincide. The refutation of this also came from the ranks of the movement. Gödel proved that any consistent axiomatization of elementary arithmetic fails to yield some arithmetic truths. Given any such axiomatization, we can construct a true arithmetic sentence not provable in that system. Strengthen that system (perhaps by adding that very sentence as a further axiom) and you have a new system generating its own unprovable truths. Mathematical truth outstrips provability (Gödel 1986).

The verificationists tried weakening their core doctrine, but the accounts failed to endorse science while repudiating metaphysics. Verificationism has, somewhat surprisingly, experienced a recent revival in the guise of *antirealism* (Dummett 1978). This version of it is also motivated by the desire to cut content down to verifiable size, eschewing depth in favor of certainty, and it faces the same kinds of objections.

3. *Truth Conditions*

The standard counterexamples to verificationism are sentences with this feature: we know what it would take for them to be true, even though we have no way of verifying them. We do know what 'Metals expand when heated' *means* although it is unverifiable. The meaning of a sentence is given by the condition which would *render* it true, rather than to any procedure which would *verify* its truth. The *truth-conditional* theory of meaning is that the meaning of a sentence is its truth condition, not its verification condition.

One objection to abstract entities, like propositions or truth conditions, is that they do not have clear-cut criteria of identity. There is no criterion telling us when we are dealing with one or two or more such entities. Quine's slogan, 'no entity without identity,' demands we make do without them (Quine 1981). Carnap, perhaps the most important verificationist, laid the foundations for verificationism's successor: the possible-worlds account (Carnap 1947).

As it happens, it snowed in Boulder, CO, on 1 January 1999, but it might not have. The actual world is one complete way for the world to be, but it is not the only way it could be. It is not *logically necessary* that the world be the way it is. We countenance a space of different possibilities just one of which is actual. A claim divides the class of complete possibilities into two. It induces a *mapping* from the collection of complete possibilities (the logical space) to the two truth values. A set can be thought of as just such a mapping—it maps its members to a positive indicator (True), and nonmembers to a negative indicator (False). So a proposition is associated with a set of worlds. The possible-worlds account says there is no more or less to a truth condition than that (Lewis 1986). The account utilizes Quine's paradigm of a clear-cut criterion of identity—the identity condition for sets. Sets are identical just in case they contain the same members. Propositions (sets of worlds) are identical if and only if they contain the same worlds. We can now tap set-theory for a unifying account of many logical phenomena.

For example, entailment is standardly understood in terms of truth preservation: P entails Q if the truth of P guarantees the truth of Q. This becomes: P is a subset of Q. Transitivity of entailment falls out of the transitivity of the subset relation. Disjunction is set-theoretic union, conjunction is intersection, and negation is complementation. A proposition is (logically) possible if it is true in some world; impossible if true in no world; necessary if true in all worlds. Modal principles—for example, that the negation of a necess-

ary proposition is impossible—also fall out of set theory. Logical relations between sentences thus piggyback on the propositions they express. A proposition is true *simpliciter* if it contains the actual world, and a sentence is true if it expresses a true proposition.

Despite its explanatory power, this account cannot be the whole story. A necessary proposition is true in all worlds, but only one set of worlds contains all worlds. 'One plus one is two' expresses a necessary truth, as does 'Two is two.' Hence, on this account, these two sentences express one and the same proposition. However, the former expresses a mildly interesting truth about addition, while the latter expresses something trivial. Since what they express is different, what they express cannot be just a class of worlds. The account is too coarse-grained. While the sentential approach can make these fine-grained distinctions, it is *too* fine-grained. ('*Une piu une fa due*' expresses the same truth, in Italian, as 'One plus one is two,' but they are distinct sentences.)

There are a range of fine-grained theories of meanings which accommodate these distinctions without resorting to sentences (Bealer 1982, Tichý 1987). These theories do not need to dispense with possible worlds and some are built on them. Provided a proposition *induces* a mapping from worlds to truth values, even if it can not be *identified* with such a mapping, we can reap the benefits of the possible-worlds framework.

4. *Falsifiability*

That metals expand when heated is false in any world in which one piece of heated metal does not expand. Since we cannot check every piece of heated metal, the proposition cannot be shown to be true. By contrast, it *can* be shown to be false, or *falsified*. The observation of one nonexpanding piece of heated metal will do the trick.

Karl Popper (1963) was one of the first to stress the fallibility of scientific theories and the asymmetry of verification and falsification. If the positivists took inspiration from Special Relativity, Popper took it from General Relativity—specifically, its prediction of the bending of light near massive bodies. He was impressed that Eddington's eclipse experiment of 1919 could verify neither Newton's nor Einstein's theory, but could have refuted either, and actually refuted Newton's. Thenceforth he took falsifiability to be the hallmark of genuine science. Pseudoscientific theories (Popper's examples were psychoanalysis and astrology) are replete with 'confirming instances.' Everything that happens *appears* to confirm them only because they rule nothing out. Popper argued that genuine theories must forever remain *conjectures*. But science is still rational, because we can submit falsifiable theories to severe tests. If they fail, we weed them out. If they survive? Then we submit them to more tests, until they too fail, as they almost assuredly will.

There are three serious problems with falsificationism. First, it does not account for the apparent epistemic value of confirmations. The hardline falsificationist must maintain that the appearance is an illusion. Second, it cannot explain why it is rational to act on the unrefuted theories. Confidence born of experimental success reeks of inductivism. Third, pessimism about the enterprise of science seems obligatory. Although truth is the goal of inquiry, the best we can manage is to pronounce a refuted theory false. We return to these in succeeding sections.

Other criticisms stem from Duhem's problem (Duhem 1954). Duhem noted that predictions can be deduced from theories only with the aid of auxiliary assumptions. One cannot deduce the path of a planet from Newton's theory of gravitation without additional hypotheses about mass and position of the Sun and the planets, and the absence of other forces. It is not Newton's theory alone, but the conjunction of the theory with these auxiliaries, which faces the tribunal of experience. If the conjunction fails it is not clear which conjunct should be blamed. Since we can always blame the auxiliaries, falsification of a theory is impossible.

Quine generalized Duhem's point to undermine the positivist's distinction between factual truth and analytic truth. If no sentence is 'immune to revision' in the face of anomalies, no sentence is true simply by virtue of its meaning. Not even the law of excluded middle s sacrosanct, and meaning itself becomes suspect (Quine 1981).

Kuhn used Duhem to undermine the rationality of scientific revolutions. Since a theory cannot be refuted by experiment, it is not irrational to stick to it, tinkering with auxiliaries to accommodate the data. 'Normal science' consists precisely in that. A revolution only occurs when a bunch of scientists jump irrationally from one bandwagon to another. In Kuhn's famous phrase (which he later regretted) a revolution is a 'paradigm shift' (Kuhn 1962).

Feyerabend followed Kuhn, maintaining that proponents of different paradigms inhabit incommensurable world-views which determine their perception of the so-called observational 'facts' (Feyerabend 1993), thereby setting the stage for *constructivists* who maintain that science is a human construct (obviously true) with no controlling links to reality (obviously false). From constructivism it is a short leap into postmodernist quicksand, the intellectual vacuity of which has been brutally but amusingly exposed (Sokal 1996).

The more extreme lessons drawn from Duhem's simple point may sound sexy, but they do not withstand sober scrutiny (Watkins 1984). It is just false that each time a prediction goes awry it is one's whole global outlook that is on trial, and that one may rationally tinker with any element at all. In the nineteenth century, Newton's theory was well confirmed by myriad observations and applications, but it

faced an apparent anomaly in the orbit of Uranus. Because the theory was well confirmed, scientists looked to the auxiliary assumptions, such as the assumption that there are no unobservable planets affecting Uranus's trajectory. It was far more probable, in the light of the total evidence, that there should be an as yet unobserved planet in the solar system than that Newton's theory be false. It was eminently reasonable to postulate such a planet and search for it. The subsequent discovery of Neptune was a resounding success for the theory.

This rejoinder, of course, requires a positive account of probability and confirmation, one which falsificationism does not supply.

5. *Confirmation*

Having jettisoned verifiability, the positivists turned to the theory of probability to ground a rigorous logic of confirmation and induction, comparable to deductive logic (Carnap 1950).

The logical probability of a proposition is a measure of its area in logical space. It is a normalized measure, so the probability of the whole space (the necessary proposition) is 1; the probability of the impossible proposition, the empty set, is 0. Other propositions have a probability somewhere between 0 and 1, depending on the proportion of the space they cover. Conditional probability—the probability of P *given* Q, written $\Pr(P|Q)$—is a kind of generalized entailment and is defined as a ratio of unconditional probabilities: $\Pr(P\&Q)/P(Q)$. If Q logically entails P then $\Pr(P|Q)$ is maximal, 1. If Q is logically incompatible with P then it is minimal, 0. In between these cases, the closer P is to entailing Q, the closer $\Pr(P|Q)$ is to 1. Evidence E *confirms* hypothesis H just in case conditioning on E *raises* the probability of H: $\Pr(H|E) > \Pr(H)$. If H entails E, $\Pr(E) < 1$ and $\Pr(H) > 0$, then if E is verified, E confirms H.

Where do the logical probabilities come from? Carnap searched for plausible *a priori* constraints on logical probability. The most obvious is given by the principle of indifference: all possible worlds have an initially equal probability. Logic does not play favorites. Unfortunately, this measure yields counterintuitive results. If the universe is infinite, and the evidence consists of a finite number of reports of the expansion of heated metals, it does nothing to raise the probability of the hypothesis that metals expand when heated. Even if the universe is finite, no amount of such data will raise our confidence that the *next* piece of heated metal will expand.

The search was on for a measure that yields more palatable results. However, there is one rather pressing problem. If the universe is infinite, then there will be an infinite number of mutually exclusive possible theories pretty much on a par in the absence of evidence. So they should all receive the same initial logical probability, which must therefore be zero. Now no evidence can ever raise the probability of an hypothesis if its initial probability is zero. So confirmation is impossible.

Inductive logicians turned to the task of motivating initial probability distributions which would give differential finite weight to generalizations, but the upshot was an embarrassment of riches, a plethora of 'inductive methods' (Hintikka and Suppes 1966). The failure to motivate a single preordained measure of logical probability is not a fatal blow to confirmation theory, however. The probabilist can allow a range of distributions, reflecting distinct *subjective* degrees of confidence. *Bayesianism*, as the position is called, places two objective constraints on rational belief (Howson and Urbach 1989). First, one's degrees of belief must obey the probability calculus. Second, one must update degrees of belief by conditionalization on incoming evidence. Together these constraints motivate the convergence of rational opinion. Given suitable initial conditions, in the long run differences in initial distributions will be overwhelmed by incoming evidence.

The basic problem with Bayesianism, however, is that probability, like verifiability, strongly favors certainty over depth. The less content a theory has, the larger the region of the logical space it covers, and the more probable it is. The whole truth is just a tiny speck in logical space—the proposition which entails all true propositions is clearly the singleton of the actual world. Any deep theory which comes anywhere close to the truth will likewise be of very low probability. We want deep theories that are close to the truth. How can that be reconciled with the demand for evidential support?

6. *Truthlikeness*

Newtonian physics superseded Aristotelian physics, but it in turn was superseded by Einsteinian physics. Newton's theory is now known to be false. Does it follow that Newton did no better at realizing the aim of physical inquiry than Aristotle? If truth is the goal, are all falsehoods alike in being total failures?

Take a simple query: 'What is the number of the planets (Π)?' The ancients gave the false answer $\Pi = 5$, but the answer has been steadily revised upwards to $\Pi = 9$, and let us suppose that is the truth of this matter. $\Pi = 5$, $\Pi = 6$, $\Pi = 7$ and $\Pi = 8$ are all false, but each makes some progress towards the truth over its predecessor. We have here a sequence of four falsehoods which increase in their *truthlikeness* or *verisimilitude*.

Probability and truthlikeness are easily conflated. Subjective probability is, if you like, *seeming like* truth. Truthlikeness is *being similar to* truth. However, the two concepts are completely different. Probability is an epistemic concept, truthlikeness is not. $\Pi = 8$ is

closer to, more *like*, the truth ($\Pi = 9$) than $\Pi = 5$, completely independently of the evidence. Moreover, it remains closer to the truth even after the probability of both falls to 0 with the discovery of Pluto. So having minimal probability is compatible with great truthlikeness. Or consider the logically necessary truth. This is not only true but certain, but it is not close to the truth about Π. Having maximal probability is thus compatible with low truthlikeness. Now consider the following sequence of truths:

$$5 \leqslant \Pi \leqslant 9$$
$$6 \leqslant \Pi \leqslant 9$$
$$7 \leqslant \Pi \leqslant 9$$
$$8 \leqslant \Pi \leqslant 9$$
$$\Pi = 9$$

This sequence exhibits decreasing probability, but increasing truthlikeness. So probability and truthlikeness, like certainty and depth, tend to pull in opposite directions. A proposition cannot be close to the whole truth without being deep. It cannot be probable without being shallow. The whole truth (the proposition which is true in just one world, the actual world) is undoubtedly closest to the whole truth. It is the deepest, correct proposition. But on any actual measure of subjective probability it will have probability 0 or close to 0.

Popper was the first to realize the importance of the concept of truthlikeness (Popper 1963). Popper wanted to be a *realist* (truth might outstrip our means of discovery), a *fallibilist* (any general theory is likely to be false), and an *optimist* (we can and do make epistemic progress). If all we can say for sure is 'missed again!,' a miss is as good as a mile, and the history of inquiry is a sequence of such misses, then epistemic pessimism follows. But if some false hypotheses are closer to the truth than others, then realism and fallibilism are compatible with optimism. The history of inquiry could turn out to be one of steady progress towards our goal, and there may be empirical indicators of that.

Popper also formulated the first rigorous account of the notion, in terms of the true and false consequences of theories. According to Popper, a proposition P is closer to the truth than Q if P entails more truths than Q, and no more falsehoods (or, fewer falsehoods, and no fewer truths). Unfortunately, this elegant theory implies that no false theory is closer to the truth than any other. (Surprisingly it took nearly 10 years before this defect was noticed (Miller 1974, Tichý 1974).) Let G be some truth implied by P but not by Q, and let F be any falsehood implied by P. Then F&G is a falsehood implied by P, but not by Q. (If Q implied F&G then Q would also entail the second conjunct, G—contradiction.) We cannot add truths to P without adding falsehoods (also, we cannot subtract falsehoods without subtracting truths).

A different approach takes the *likeness* in truthlikeness seriously, meshing this with a possible-worlds approach to propositions (Tichý 1974, Hilpinen 1976, Oddie 1986).

Suppose possible worlds exhibit degrees of similarity to one another, and distance between worlds is a measure of dissimilarity. If we could find a natural and plausible way of measuring the distance between subsets of the space we would have a measure of distance between propositions. Suppose T is the whole truth about some matter (as $\Pi = 9$ is the whole truth about the number of the planets). Answers to a query, whether partial (like $7 \leqslant \Pi \leqslant 9$) or complete (like $\Pi = 7$), would then be ordered by their distance from the truth ($\Pi = 9$). This provides a schema for a theory, the details of the similarity metric and of propositional distance to be filled in.

There are two important criticisms of this likeness approach. First, any similarity metric is sensitive to the set of properties deemed salient (Miller 1974). If all possible attributes are salient, then distinctions of similarity collapse. So if the selection of attributes is to be objective, then some properties must 'carve nature at the joints.' That involves a robust version of realism about properties, but one the scientific realist might not mind embracing (Armstrong 1978).

Second, a dilemma: if we know the truth we do not need truthlikeness; and if do not know the truth we cannot know which theories are truthlike. Either way the concept is a useless addition to our intellectual armory.

Both horns are weak. First, even once we have attained the truth it is interesting to trace our history of attempts to reach it, and characterize the progress. Second, even though degree of truthlikeness is typically an unknown magnitude, with a measure of probability we can *estimate* its value in the usual ways. Expected degree of truthlikeness is epistemically accessible. Truthlikeness would then dovetail nicely with decision theory. Suppose there are cognitive utilities, just as there are noncognitive utilities. The point of inquiry would be to maximize cognitive utility. The best cognitive state to be in is, Godlike, to embrace the whole truth—the proposition with maximal truthlikeness. Embracing propositions far from the truth is cognitively bad. So it is natural to identify cognitive utility with degree of truthlikeness. However, the injunction to maximize objective cognitive value does not provide a *selection* procedure since, as in the practical case, we have no direct epistemic access to states with high utility. So we must be guided by *expected* cognitive utility preferring those theories with higher *expected* truthlikeness (Niiniluoto 1987).

7. *Conclusion*

This decision–theoretic approach is an interesting compromise between the positivists's concern with

truth and probability and the falsificationist's apparently conflicting concern with depth. But as so often happens, this offspring exposes defects in both its parents. The positivists were misguided in concentrating on truth and probability. A theory can be true and have high probability and yet very low truthlikeness. The tautology is an extreme example, but trivial contingent truths can be very well confirmed, and still be far from the truth. Being true, or being likely to be true, is not where the action is. Such theories may lack depth. But the falsificationists were wrong to demonize falsity and exhalt falsification. A false and falsified theory can still be close to the truth, with good evidence that it is so (Newton's theory). Falsifying a theory shows it is not the best, but the best theory is not the only good theory. A falsehood may not hit the bullseye, but it can come very close.

Coherent defensible accounts of truth, probability, and truthlikeness can be combined. The goal of inquiry is the whole truth. Actual progress occurs when truthlikeness increases. The epistemic indicator of truthlikeness is expected truthlikeness. This can be high even when a proposition is refuted, low even when it is confirmed. The conflict between the depth and certainty poles, between which epistemologists have been oscillating for 400 years, can thus be resolved.

See also: Bayesian Theory: History of Applications; Constructivism/Constructionism: Methodology; Hume, David (1711–76); Kuhn, Thomas S (1922–96); Logical Positivism and Logical Empiricism; Popper, Karl Raimund (1902–94); Positivism, History of; Truth and Credibility: Science and the Social Study of Science

Bibliography

Armstrong D M 1978 *Universals and Scientific Realism*. Cambridge University Press, Cambridge, UK

Bealer G 1982 *Quality and Concept*. Clarendon Press, Oxford, UK

Carnap R 1947 *Meaning and Necessity*. Chicago University Press, Chicago

Carnap R 1950 *Logical Foundations of Probability*. Chicago University Press, Chicago

Church A 1956 Propositions and sentences. In: Bochenski J M (ed.) *The Problem of Universals*. University of Notre Dame Press, Notre Dame, IN, pp. 1–25

Duhem P 1954 *The Aim and Structure of Physical Theory* [trans. Wiener P]. Princeton University Press, Princeton, NJ

Dummett M A E 1978 *Truth and Other Enigmas*. Harvard University Press, Cambridge, MA

Feyerabend P K 1993 *Against Method*. Verso, London

Frege G 1984 The thought: a logical inquiry. In: McGuinness B (ed.) *Collected Papers on Mathematics, Logic and Philosophy*. Blackwell, Oxford, UK, pp. 351–72

Gödel K 1986 On formally undecidable propositions of *Principia Mathematica* and related systems I. In: Feferman S, Przelecki M, Szaniawski K, Wójcicki R (eds.) *The Collected Works of Kurt Güdel*. Oxford University Press, New York, Vol. I

Hilpinen R 1976 Approximate truth and truthlikeness. In: Przelecki M, Feferman S, Dawson J W, Kleene S C, Moore G H, Solovay R M, van Heijenart J (eds.) *Formal Methods in the Methodology of the Empirical Sciences*. Reidel, Dordrecht, The Netherlands, pp. 144–95

Hintikka J, Suppes P (eds.) 1966 *Aspects of Inductive Logic*. North-Holland, Amsterdam

Horwich P 1990 *Truth*. Blackwell, Oxford, UK

Howson C, Urbach P 1989 *Scientific Reasoning: the Bayesian Approach*. Open Court, La Salle, PQ

Kuhn T 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago

Lewis D 1986 *The Plurality of Worlds*. Blackwell, Oxford, UK

Miller D 1974 Popper's qualitative theory of verisimilitude. *British Journal for the Philosophy of Science* **25**: 166–77

Niiniluoto I 1987 *Truthlikeness*. Reidel, Dordrecht, The Netherlands

Oddie G 1986 *Likeness to Truth*. Reidel, Dordrecht, The Netherlands

Popper K R 1963 *Conjectures and Refutations*. Routledge, London

Quine W V O 1981 *Theories and Things*. Harvard University Press, Cambridge, MA

Ramsey F P 1927 Facts and propositions. *Proceedings of the Aristotelian Society* **7**: 153–76

Schlick M 1979 Form and content: an introduction to philosophical thinking. In: Mulder H L, Van de Velke-Schlick B (eds.) *Philosophical Papers*. Reidel, Dordrecht, The Netherlands, Vol. II, pp. 285–387

Sokal A 1996 Transgressing the boundaries: toward a transformative hermeneutics of quantum gravity. *Social Text* **46/47**: 217–52

Tarski A 1969 The concept of truth in formalized languages. In: Woodger J (ed.) *Logic, Semantics and Mathematics*. Clarendon Press, Oxford, UK, pp. 152–268

Tichý P 1974 On Popper's definitions of verisimilitude. *British Journal for the Philosophy of Science* **25**: 155–60

Tichý P 1987 *The Foundations of Frege's Logic*. de Gruyter, Berlin

Watkins J W N 1984 *Science and Scepticism*. Princeton University Press, Princeton, NJ

G. Oddie

Tversky, Amos (1937–96)

1. Amos Tversky

Amos Tversky was born in Haifa, Israel, on March 16, 1937. His father was a veterinarian, his mother a member of Israel's Parliament, the Knesset. He fought in Israel's 1956, 1967, and 1973 wars, and received its highest honor for bravery. He received his BA in psychology at the Hebrew University of Jerusalem, in 1961. In 1965, he received his Ph.D. from the University of Michigan's mathematical psychology program. He worked there with Clyde Coombs (his

doctoral advisor), Ward Edwards, and David Krantz, among others. He returned to Jerusalem as a faculty member in 1967, moving to Stanford in 1978. A year at the Center for Advanced Study in the Behavioral Sciences provided Tversky and Kahneman with concentrated time to develop their approach to judgment under uncertainty.

Tversky died of melanoma, on June 6, 1996, in Stanford, California. At his death, he was Professor of Psychology at Stanford University, where he was also a Principal of its Center on Conflict and Negotiation. He held positions at Tel Aviv University, as Visiting Professor of Economics and Psychology and Permanent Fellow of the Sackler Institute of Advanced Studies. His wife, Barbara, is Professor of Psychology at Stanford, studying perception. He had three children, Oren, Tal, and Dona. He was known for his great energy, joy of life, and sense of humor.

Tversky made fundamental contributions to the understanding of human thought processes, and the mathematical foundations of the behavioral sciences. His work was distinguished by elegant formal models, tested in simple, illustrative experiments. Throughout much of his career, he worked together with Daniel Kahneman, on studies of judgment and decision making under conditions of risk and uncertainty. On these topics and others, he also collaborated with other scholars from psychology, economics, political science, law, statistics, philosophy, and mathematics. These projects, and the personal interactions surrounding them, helped to integrate these disciplines, at a time of increasing specialization. Among his many honors were the Distinguished Scientific Contribution Award of the American Psychological Association, a MacArthur Fellowship, and honorary doctorates at Yale, Chicago, Göteborg (Sweden), and New York at Buffalo. He was a member of the US National Academy of Sciences. It served as one venue for pursuing his life-long commitment to encouraging peace and understanding among diverse people.

2. *Axiomatic Theories of Choice*

An early paper (Tversky 1969) typifies his approach. Normative theories of choice provide rules that individuals should follow, in order to make effective decisions. The most prominent of these is utility theory, founded on a set of intuitively appealing axioms. It describes options in terms of a set of attributes, or features, that an individual might like or dislike about them. When buying a car, the attributes might be (price, size, styling, reliability, comfort). According to utility theory, an individual should consider all relevant attributes, then rank the options according to their attractiveness (or 'utility'). That ordering should have the property of *transitivity*. Someone who prefers Car A to Car B and Car B to Car C should also prefer Car A to Car C.

Transitivity is a decision-making norm that most people would endorse. If it accurately described their behavior, they would be better off. So would scientists and policy makers. The former would find it relatively easy to predict behavior, knowing that there is a stable overall ranking, which can be identified without examining every possible paired comparison. The latter would be able to create policy options (e.g., parks, housing, insurance schemes, transportation systems), knowing that the public's response will be relatively predictable and consistent.

Tversky showed systematic violations of transitivity, with experimental tasks, similar to choices encountered in everyday life. Moreover, he went beyond merely showing that utility theory described behavior imperfectly. Such a limited demonstration could have been easily attacked as a destructive curiosity that does little to help scientists do their job of explaining and predicting behavior. However, Tversky showed that intransitivity arose from essential psychological processes, which he captured in a simple, but insightful model. These processes reflect the limits to human information processing capacity. In situations without the opportunity for trial-and-error learning, people can at best be boundedly rational, working in an orderly way on a subset of the problem (Simon 1957).

The coping process that Tversky proposed has people comparing options initially on the most important attribute (e.g., car price), then checking whether differences on the second attribute (e.g., reliability), and perhaps others, are so great as to change the choice. For example, imagine sequential comparisons between overlapping pairs of cars (A with B, B with C, C with D, and so on). The sequence is arranged so that each new car is both cheaper and less reliable that the previous one. Looking primarily at the price would make each new option more attractive, while the reductions in reliability slowly mount up. At some point, though, if the first car were compared with the last car, a buyer might decide that reliability has gotten so low that the reduction in price is insufficient compensation. In that case, the first car, which was long ago rejected, would suddenly be preferred.

Such intransitivity could, in principle, turn people into 'money pumps,' willing to pay something to move from one option to the next, but eventually going back to the first option and starting over again. As Tversky noted, car salespeople may exploit this cognitive weakness by adding options (and costs) to a favored car. Unlike the experimenter, though, they hope to stop just short of the point where customers balk at the increased price and revert to the stripped-down model. Hanson and Kysar (1999) discuss the implications of such exploitation for marketplace regulation.

Two themes in this research recur in Tversky's work. One is that choices, especially novel ones, can be very sensitive to how the options are presented. That insight is central to prospect theory (see below), which

also provides expression for results from a long tradition of psychological research into 'context effects.' The second theme is that descriptively valid formal theories of human behavior should be assembled from psychologically plausible assumptions, rather than derived from normatively postulated principles.

The latter insight is expressed in many of Tversky's projects, which articulate the formal structure of particular choice situations in ways that allow the expression of plausible behavioral responses (e.g., Tversky 1967, 1972). In some cases, these theories adapted existing mathematical approaches. In other cases, they drove the mathematics. The fundamental mathematical research associated with this research is summarized in three landmark volumes (Krantz et al. 1971/1989/1990). The study of alternative choice rules for dealing with multiattribute options has become a cornerstone of a *conjoint measurement*, a family of procedures used widely in applied research (e.g., marketing multiattribute consumer products).

3. *Prospect Theory*

These strands of research came together in Kahneman and Tversky's (1979) 'prospect theory.' Published in a central economics journal, it offered an alternative to utility theory, the core of that discipline. It claimed to achieve greater descriptive validity—but at the price of abandoning the normative aspirations of economics. It depicted individuals as thinking in an orderly, sensible way when faced with choices. However, the rules that people follow differ from those prescribed by the axioms of utility theory, in ways that produce systematic violations. People are sensitive to features of choice tasks that have no representation in utility theory, and insensitive to features that are in the theory.

Superficially, prospect theory resembles utility theory. 'Prospects' (i.e., options) are represented as sets of possible consequences. The value (rather than 'utility') assigned to a prospect is the sum of the values assigned to each consequence, weighted by the chances of it happening. However, the components of this evaluation differ from their utility theory analogs in ways designed to make the theory more realistic psychologically (Tversky and Kahneman 1991):

(a) The evaluations are made relative to a *reference point*, rather than with respect to an individual's *net asset position* (or entire worth). That reference point is typically the status quo. However, it could be another salient value (e.g., evaluating a raise relative to what one expected to get or to what others got, rather than relative to one's current salary). This feature of prospect theory reflects people's widely observed sensitivity to changes.

(b) People care less about a given absolute difference in gains or losses as the overall change increases. The idea the people care less about a fixed gain as they have or win more is shared with utility theory. However, the idea that people care proportionately less as losses mount up is not. Utility theory explains insurance buying (among other things) in terms of people's special aversion to large losses. Prospect theory reflects the psychological principle that the just-noticeable difference between situations is greater when the overall change is larger—whatever the direction of change.

(c) Before being combined with value judgments, the probabilities of the consequences are subjected to a 'probability weighting function.' It reflects the observation that people pay particular attention to sure things, while being relatively indifferent to probability differences in the middle range (e.g., the difference between 0.35 and 0.45 matters less than that between 0.90 and 1.00).

As combined in the theory, these principles (along with some others) predict (among other things) *framing effects*. Changing the reference point can reverse choices, in ways that have no reflection in utility theory (which has no such point). For example, an emergency medical intervention that could save 200 of 600 possible lives is more attractive when described in terms of the 200 lives saved than the 400 lives lost. Prospect theory's challenge to utility theory evoked vigorous responses. One extreme involved radical methodological skepticism. Partisans attacked any aspect of prospect theory's research method that could, conceivably, have contributed artifactually to behaviors offered as evidence supporting the theory. Although some of these critiques were ad hoc claims, disconnected from the research literature, others raised fundamental questions with theoretical and methodological implications.

One such question concerns the conflicting norms of experimental economics and experimental psychology regarding the provision of financial incentives. Economists typically believe that what people say and do has little meaning, unless they are rewarded for their performance. In this view, even if they are systematic, responses to hypothetical question could reflect no more than acquiescence to implicit task demands. In contrast, psychologists typically view all behavior as 'real,' with the attendant challenge of determining how individuals have construed their task. Even a hypothetical choice can induce high intrinsic and extrinsic motivation, as individuals try to impress themselves and those watching them. Camerer and Hogarth (1999) summarize the evidence, regarding the varied impacts of monetary rewards on decision-making performance. The other extreme involves economists who accept the validity of prospect theory's results and attempt to reshape their science in its light. Some have examined its implications for the axiomatic basis of utility theory, either by refining those axioms to fit the data better or by replacing them with ones expressing different normative principles.

Others have examined the expression and implications of non-rational behavior in real-world circumstances, prompting new fields like behavioral finance, which questions widely accepted efficient market theories.

Still others have sought to help people reduce such inconsistencies in their preferences. These 'constructivist' approaches assume that people's inconsistencies arise because evaluating options requires an inferential process, whereby people derive a specific valuation from more general, basic values. If that process is incomplete, then people will be unduly influenced by the initial problem presentation. Reducing that risk requires a more interactive process, explicitly considering alternative perspectives—trying to balance those suggestions, so that people come to know what they want. Fischhoff and Manski (1999) show the convergence among economists and psychologists concerned with these questions.

4. *Judgment under Uncertainty: Heuristics and Biases*

For economists and psychologists concerned with decision making, values are a matter of personal preference. As a result, showing them to be suboptimal means demonstrating inconsistencies. With beliefs, suboptimality can be found by comparing judgments to one another (looking for inconsistency) and to an accepted external standard (looking for error). In their first collaboration, Tversky and Kahneman (1971) demonstrated errors in the judgments of mathematical psychologists and statisticians, when estimating the statistical power of research designs. These individuals knew how to perform the relevant calculations. However, when forced to rely on intuition, they systematically exaggerated the chances of observing anticipated results—acting as though the law of large numbers applied to samples with small numbers of observations.

Tversky and Kahneman (1974) proposed that these judgments arose from relying on the *representativeness* heuristic. According to this rule of thumb, an event is judged likely to the extent that it captures (or 'represents') the salient features of the situation that might produce it. Conversely, reliance on representativeness should lead to neglecting factors lacking a salient role in such processes, even if they are normatively relevant. Sample size is such a factor. If a situation can produce an event, why should it matter how much one observes? A second normatively relevant factor with no role in representativeness is the relative frequency of an event in the past. Such *base-rate* (or *a priori*) information should guide predictions unless there is strongly diagnostic *individuating* information, showing why a specific case is different from the norm. (Medical students are sometimes told, 'When you hear hoof beats, think horses, not zebras,' at least in North America.) However, studies have found that even flimsy evidence can attract the lion's share of attention. It is natural to think about how well someone fits the image of an engineer (or how much a mammogram image looks like a tumor). It is not so natural (or easy) to combine that vision or theory with a statistical summary of how frequent engineers (or tumors) are in the population. There have been many subsequent demonstrations of underweighting base rates (including in the interpretation of mammograms), as well as vigorous debate (sometimes with data, sometimes not) regarding the conditions determining its extent (Gilovich et al. in press, Kahneman et al. 1982).

Analogous reasoning led to predicting that reliance on representativeness would encourage underestimating regression to the mean and exaggerating the predictability of random sequences. The latter was demonstrated in studies showing an illusory belief in being able to identify streaks in basketball free-throw shooting. Both collegiate and professional players (and fans) interpreted patches of random sequences as reflecting transient 'hot hands' (Gilovich and Tversky 1985).

Tversky and Kahneman (1974) proposed the representativeness heuristic as one exemplar of a general strategy that people use when confronted with unfamiliar tasks (and no opportunity for systematic study or calculation). A second such heuristic, *availability*, judges the probability of an event by the ease with which examples come to mind. That rule has some validity (in many domains, things that one hasn't seen tend to be less likely than things that one has). However, it can predictably lead one astray, as when the news media disproportionately report one risk (e.g., homicide vs. asthma) (Slovic 1987). Subsequently, Tversky and Koehler (1994) formalized the relationship between evaluations of evidence and probability judgments in 'support theory.' Like its predecessors, it predicts the conditions under which simple, intuitive thought leads to good and bad judgments. A third heuristic, also with historical roots in psychology, is *anchoring and adjustment*. Users begin an estimation process from a salient value (the 'anchor'), then adjust from there, in response to other concerns that come to mind. Typically, that adjustment is insufficient. If the anchor is far from the correct value, then the resulting estimate will be as well.

One cross-cutting theme in Tversky's judgment research (as with that on choice) is questioning the normative standards used to evaluate people's performance. Here, too, the result has been reevaluating those standards and their domains of applicability (Shafer and Tversky 1985). Another recurrent theme is the importance of comparison processes in judgment. Representativeness asks how well an event fits the process that would have to produce it. Availability asks how well an example fits the archetype of the category whose probability is being predicted. In a parallel research program, Tversky (1977) created a

theory of similarity judgments, based on measure-theoretic concepts and, again, simple, elegant experiments.

See also: Bayesian Theory: History of Applications; Behavioral Economics; Bounded and Costly Rationality; Cognitive Psychology: Overview; Decision and Choice: Economic Psychology; Decision Biases, Cognitive Psychology of; Decision Making, Psychology of; Decision Research: Behavioral; Decision Theory: Bayesian; Decision Theory: Classical; Heuristics for Decision and Choice; Risk: Theories of Decision and Choice; Utility and Subjective Probability: Contemporary Theories

Bibliography

Camerer C F, Hogarth R M 1999 The effects of financial incentives in experiment. *Journal of Risk and Uncertainty* **19** 7–42

Fischhoff B, Manski C F (eds.) 1999 Preference elicitation. *Journal of Risk and Uncertainty* **19** (1–3)

Gilovich T, Griffin D, Kahneman D (eds.) in press *The Psychology of Judgment: Heuristics and Biases*. Cambridge University Press, New York

Gilovich T, Tversky A 1985 The hot hand in basketball. *Cognitive Psychology* **17**: 295–314

Hanson J D, Kysar D A 1999 Taking behavioralism seriously: Some evidence of market manipulation. *Harvard Law Review* **112**: 1420–1572

Kahneman D, Slovic P, Tversky A (eds.) 1982 *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York

Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263–81

Krantz D, Luce R D, Suppes P, Tversky A 1971/1989/1990 *Foundations of Measurement*, Vols. 1–3. Academic Press, New York

Shafer G, Tversky A 1985 Languages and designs for probability judgment. *Cognitive Science* **9**: 309–39

Simon H A 1957 *Models of Man: Social and Rational*. Wiley, New York

Slovic P 1987 Perception of risk. *Science* **236**: 280–5

Tversky A 1967 A general theory of polynomial conjoint measurement. *Journal of Mathematical Psychology* **4**: 1–20

Tversky A 1969 The intransitivity of preferences. *Psychological Review* **76**

Tversky A 1972 Elimination by aspects: A theory of choice. *Psychological Review* **79**: 281–99

Tversky A 1977 Features of similarity. *Psychological Review* **84**: 327–52

Tversky A, Kahneman D 1971 Belief in 'law of small numbers'. *Psychological Bulletin* **76**: 105–10

Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* **185**: 1124–31

Tversky A, Kahneman D 1991 Advances in prospect theory. *Journal of Risk and Uncertainty* **5**: 297–323

Tversky A, Koehler D J 1994 Support theory: A nonextensional representation of subjective probability. *Psychological Review* **101**: 547–67

B. Fischhoff

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

U

Unconscious: History of the Concept

The history of the concept of the unconscious in western thought will be discussed in three phases of its development:

(a) from as early as the sixth century BC to the seventeenth century, when the concept was not usually distinguished from other, wholly metaphysical concepts in which it was embedded;

(b) the beginning of the attempt, under the influence of Descartes, to distinguish the concept of the unconscious from its metaphysical moorings, and to submit it to scientific examination;

(c) the modern period, in which the concept very commonly is assumed by researchers in the social sciences despite doubts by some about whether the concept's separation from its metaphysical origins is, or ever can be, complete.

1. Metaphysical Beginnings

The idea that people are sometimes unaware of their own mental processes while they are going on and may even be incapable of later retrieving them introspectively, has had a long life of its own independent of and prior to its uses in modern experimental psychology, psychoanalysis, cognitive science and linguistics. However, the idea's early occurrences do not appear tagged by a single term translatable as 'the unconscious'; rather, the idea commonly occurs as a corollary of (a) the rationalist doctrine of innate ideas, and more broadly, of (b) the even more ancient idea of the microcosm, the traditional world-view according to which each mind contains the universe, in miniature, within itself (Levy 1967). Since the microcosm concept can be thought of, at least in part, as the view that all ideas are innate, the rationalist doctrine, for example, of Descartes, for whom only some ideas, e.g., of God, or of thought, are innate, is implied by, but does not imply the microcosm view.

Since, presumably, human subjects are never consciously aware of having all ideas within their minds, the concept of the unconscious is implicit in the idea of the microcosm; and to the extent that one might never consciously reflect on (what Descartes meant by) the idea of God or of thought, the concept of the unconscious is implicit in the doctrine of innate ideas, as well. However, rationalists often accepted the microcosm model, too, as for example, Leibniz, who writes of souls that they are 'perpetual living mirrors,' each representative of the entire universe; and Plato, who used the example of a young slave untrained in geometry, led only by questioning to the discovery of a mathematical truth, to show the boy 'recovering the knowledge out of himself' where the possibility of this 'recovery of knowledge,' (*Meno*, 85D), which Plato called recollection (*anamnesis*), rests on the claim that '… as all nature is akin, and the soul has learned all things, there is no reason why we should not, by remembering but one single thing … discover everything else (*Meno*, 81C-D). Since neither Plato nor Leibniz meant to say that any human subject could consciously experience themselves 'mirroring' the universe or consciously recall having known the recovered mathematical truth at some earlier time, the concept of the unconscious implicitly employed by them is a thoroughly metaphysical one.

2. Descartes and the Beginning of Scientific Scrutiny

Descartes's views require separate comment, since, unlike Plato and Leibniz, he had no attachment to the microcosm concept; yet although he had no difficulty with the existence of unconscious mental states insofar as his doctrine of innate ideas implied them (Cottingham 1998), he has nevertheless gained the reputation of having defined them out of existence as a consequence of his mind-body dualism and his supposed identification of conscious awareness as the essential attribute of mind. Whyte goes so far as to write, 'Descartes … by his definition of mind as awareness, may be said to have provoked as reaction, the European discovery of the unconscious mind … Descartes … had no clear conception of the unconscious mind, for to him all that is unconscious is physiological' (Whyte 1962, p. 80). But it is thought, not awareness that Descartes defines as essential to mind (*Meditation II*), and Descartes takes seriously the possibility, concerning his ideas of the external world, that there is some 'faculty not yet fully known to me, which produces these ideas without any assistance from external things; this is, after all, just how I have always thought ideas are produced in me when I am dreaming' (*Meditation III*). Such a faculty in him, by implication unconsciously active, apparently would need an intellectual substance to inhere in (*Meditation VI*), so it cannot, presumably, be purely physiological in nature. What finally persuades Descartes that such a deceptive and unconsciously acting

mental faculty does not exist is not the thought that the idea of it is incoherent in itself (as would be natural if his view were as Whyte supposed), but rather that once God's existence has been proved, the existence of that faculty is incompatible with God's perfect goodness (*Meditation V*).

What it was in Descartes's philosophy capable of provoking the somewhat desperate efforts to defend the as yet unnamed (Whyte 1962, pp. 25, 60) concept of the unconscious against him can more properly be found in his rejection of the comforts of the microcosm model, and his insistence on accepting as true only things that are clearly and distinctly perceived, things for which there are 'sure signs' (*Meditation I*). For when such a standard is applied to the concept of the unconscious, its defenders as well as its detractors are at least temporarily at a loss, since no-one then knew, nor is there complete consensus even now, as to where the defining line separating conscious from other mental states is to be drawn. How exactly it can be proved that an unconscious part of the mind exists in reality, i.e., that it is more than a mere conceptual possibility (if even that), and what sure signs ought to be required of specific contents in it, remain subjects of debate. The contrast on these points with the conscious mind and its states appears quite stark. It was Descartes who first made necessary the project of attempting to separate the concept of the unconscious from the metaphysical moorings that had until that time shrouded it from scrutiny. The concept of the unconscious was itself about to enter consciousness. The search was on for its manifestations in empirical phenomena that could be examined in a scientific way.

3. *The Modern Period*

By the end of the nineteenth century, a wide range of psychological phenomena had been identified as seeming to need unconscious mental activity for their explanation; ordinary perception and habit-formation, hypnosis and post-hypnotic suggestion, dreaming and neurotic symptom-formation, to name a few. According to James (1890), however, all such 'proofs' of the existence of the unconscious would fail, since the very idea of the unconscious mind is an 'unintelligible and fantastical' notion, and more parsimonious explanations of the phenomena in question are in all cases available, e.g., explanations in terms merely of (a) 'brain-processes to which no ideation seems attached,' (b) 'perceptions and volitions … performed consciously, only so quickly and inattentively that no *memory* of them remains,' or (c) 'split-off' consciousness, without any reference to unconscious mental activity having to be invoked. By the end of the twentieth century, the range of psychological phenomena offered in support of the concept of the unconscious had expanded to include language acquisition and 'subliminal perception,' 'blindsight,' and cognitive dissonance and self-perception attribution processes (Nisbett and Wilson 1977). However, skeptics have continued to doubt the existence of unconscious mental activity entirely, for reasons roughly similar to those laid down by James, though with more sophisticated arguments (Searle 1992, Strawson 1994).

There is here an evident lack of prior agreement about what conditions a phenomenon must meet for it to be a genuine instance of unconscious mental activity; one side of the debate assumes that more (and more exotic) examples are needed to make their case, while the other side is already committed to the view that no possible example can succeed. At least it can be said of those who use the sorts of cases mentioned to support the existence of unconscious mental activity that they agree about what criteria determine the contents of the unconscious in any given case, and of what those contents are; there is, e.g., no dispute among them about what stimulus is unconsciously perceived in any given subliminal perception experiment, or what forgotten order the subject is obeying in posthypnotic suggestion even if there are many questions about the way these stimuli are unconsciously processed.

By contrast, the psychoanalytic evidence raises different problems on this point, since there is no agreement concerning what Freud thought the criterion of truth in psychoanalysis is, and of what the consequences of adopting one criterion or another would be. It is widely supposed (e.g., by Grünbaum 1984, Rey 1998) that Freud thought therapeutic success is the only (or best) evidence of the correctness of psychoanalytic interpretations. The acceptance of such a criterion of correctness seems to leave Freud vulnerable to the well-known criticism that therapeutic success could as easily be due to suggestion or placebo effects as to interpretive truth. By contrast, Wittgenstein (1966) and Levy (1996) attribute to Freud the view that assent by the subject of an interpretation is ultimately the sole, or main criterion of its correctness. Which criterion Freud accepted, and which he ought to have accepted, are questions impossible to settle here; however, in favor of the assent criterion as the answer to both questions, it can be said first that the assent criterion seems more in accord with Freud's own most extensive discussion of the matter (Freud 1937). Indeed, he even considers there the possibility that in some cases the mark of a true 'construction' will be that the patient's symptoms and general condition are not improved but rather aggravated (Freud 1937, p. 265). In addition, that criterion also allows interpretations to be true outside the therapeutic situation, where the danger of suggestion and placebo effects can more easily be avoided. Of course, in cross-cultural anthropological testing of non-individual psychoanalytic hypotheses (e.g., on the Oedipus complex, Whiting et al. 1958, Stephens 1962, Spiro 1982), neither the assent nor therapeutic success cri-

terion is directly applicable. Lastly, it is wishes and ideas, for Freud, that inhabit the unconscious, not desires and beliefs (Freud 1893, p. 251, 1895, p. 10, 1915a, p. 152, 1915b, p. 178). However, ideas and wishes, whether conscious or unconscious, lack the tight connection with behavior to be found in desire and belief; it is plausible to say, in many cases at least, that the character of a subject's ideas and wishes is ascertainable only by way of the assent and self-ascription the subject provides, whereas the comparable claim about desires and beliefs would have very little plausibility. This provides further evidence that Freud accepted the assent criterion, after all.

Regardless of which criterion is accepted, the question remains whether unconscious ideas and wishes as referred to psychoanalytically provide genuine exceptions to James's requirements. Without trying to resolve this matter here, it should be noted that the presence or absence of a brain-process cannot hinge on the assent of the person whose brain it is; so if the assent criterion were assumed, explanations of James's type (a) would be foreclosed. How, if at all, unconscious contents as psychoanalytically interpreted, might also elude his explanation types (b) and (c) remains to be seen. More broadly, whether psychoanalytic interpretations are capable of being tested experimentally, and how (if at all) to resolve the difficulties that arise when such testing is attempted are also unanswered questions (Edelson 1984, Erdelyi 1985, Erwin 1996).

See also: Conscious and Unconscious Processes in Cognition; Consciousness and Sensation: Philosophical Aspects; Consciousness, Cognitive Psychology of; Consciousness, Neural Basis of; Freud, Sigmund (1856–1939); James, William (1842–1910); Psychoanalysis: Overview

Bibliography

Cottingham J 1998 *Philosophy and the Good Life: Reason and the Passions in Greek, Cartesian and Psychoanalytic Ethics*. Cambridge University Press, Cambridge, UK

Edelson M 1984 *Hypothesis and Evidence in Psychoanalysis*. University of Chicago Press, Chicago

Erdelyi E 1985 *Psychoanalysis: Freud's Cognitive Psychology*. Freeman, New York

Erwin E 1985 *Final Accounting: Philosophical and Empirical Issues in Freudian Psychology*. MIT Press, Cambridge, MA

Freud S 1893 *The Complete Letters of Sigmund Freud to Wilhelm Fliess 1887–1904*. Harvard University Press, Cambridge, MA

Freud S 1895 Studies on hysteria. In: Breuer J, Freud S (eds.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (trans. Strachey J et al. 1958). Hogarth Press, London

Freud S 1915a Repression. In: Breuer J, Freud S (eds.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (trans. Strachey J et al. 1958). Hogarth Press, London

Freud S 1915b The unconscious. In: Breuer J, Freud S (eds.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (trans. Strachey J et al. 1958). Hogarth Press, London

Freud S 1937 Constructions in analysis. In: Breuer J, Freud S (eds.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (trans. Strachey J et al. 1958). Hogarth Press, London

Grünbaum A 1984 *The Foundations of Psychoanalysis: A Philosophical Critique*. University of California Press, Berkeley, CA

James W 1890/1983 *The Principles of Psychology*. Harvard University Press, Cambridge, MA

Levy D 1967 Macrocosm and Microcosm. In: *The Encyclopedia of Philosophy*. Macmillan and the Free Press, New York, pp. 121–5

Levy D 1996 *Freud among the Philosophers: The Psychoanalytic Unconscious and Its Philosophical Critics*. Yale University Press, New Haven, CT

Nisbett R E, Wilson T D 1977 Telling more than we can know: Verbal reports on mental processes. *Psychological Review* **84**(3): 231–59

Rey G 1998 Unconscious Mental States. *Routledge Encyclopedia of Philosophy*. Routledge, London, pp. 522–7

Searle J 1992 *The Rediscovery of the Mind*. MIT, Cambridge, MA

Spiro M E 1982 *Oedipus in the Trobriands*. University of Chicago Press, Chicago

Stephens W N 1962 *The Oedipus Complex Hypothesis: Cross-cultural Evidence*. Free Press, Glencoe, NY

Strawson G 1994 *Mental Reality*. MIT, Cambridge, MA

Whiting J W M 1958 The function of male initiation ceremonies at puberty. In: Maccoby E E et al. (eds.) *Readings in Social Psychology*. Holt, New York

Whyte L L 1962 *The Unconscious before Freud*. Doubleday Anchor Books, Garden City, NY

Wittgenstein L 1966 *Lectures and Conversations on Aesthetics, Psychology, and Religious Belief*. University of California Press, Berkeley, CA

D. Levy

Underclass

The term 'underclass' has found its way into the common vernacular to describe a subset of the poor whose cultural traits are thought to be at odds with the rest of society. Its use, however, has been the subject of much scholarly debate, not only the meaning of the term and its value as a social category, but also the social processes that generate and reproduce those segments of the poor population that fall under the label. The outcome of these debates has important implications for finding a lasting solution to the problems of those at the very bottom of the urban social structure.

1. The Poor are Different

Distinguishing between the poor based on cultural differences has a long history in industrialized countries. For decades, the British establishment

raised public concerns about whether 'the rabble,' 'dangerous classes,' 'lumpen,' or 'vagrants,' were morally unfit and a potential threat to social order. They were often contrasted with others who, despite their poverty, more or less conformed to the social expectations of the upper and middle classes. Similarly, in the USA, early poor laws distinguished between the incapacitated poor and the able-bodied. This distinction restricted the amount and type of public aid the latter could benefit from and also differentiated the poor in terms of 'deserving' and 'undeserving' subgroups (Katz 1989).

With the advent of the industrial revolution, European social scientists began to think of the poor as victims of the technological revolution, which was changing the way work, especially manual labor, was being performed. These notions found their way to the USA and resulted in a great deal of thoughtful inquiry about the causes and consequences of persistent poverty. The recent 'underclass' label was first used in the USA by Gunner Myrdal, a European economist, in the context of theorizing about the increasing polarization of American society (Myrdal 1962). Myrdal observed that a growing segment of the poor, because of a paucity of marketable skills and lack of access to institutional supports, such as adequate schools and government support, were being consigned to the bottom of the class structure, if not outside it altogether.

Myrdal's initial insight into the structural underpinnings of the underclass was given new life by Wilson (1978), who used the term to call attention to the growing disparity between the middle class and chronically poor, urban blacks who suffered the brunt of the deindustrialization visited on northern US cities in the preceding years. Wilson (1987) later refined his definition of the underclass to describe a 'heterogeneous grouping of families and individuals who are outside the mainstream of the American occupational system … individuals who lack training and skills and either experience long-term unemployment or are not members of the labor force.' Without denying the existence of maladaptive culture, Wilson argued that this was 'a response to social structural constraints' which proliferate in areas of concentrated poverty. Apart from Glasgow's provocative book, *The Black Underclass* (1980), which affirmed the complex relationship between economic restructuring, long-term unemployment, and prosocial orientations, few liberals sought to highlight this relationship.

Reviving acrimonious debates between liberal and conservative scholars over the relative merits of cultural versus structural explanations of poverty which raged in the USA in the 1960s and 1970s, conservative explanations for the existence of an underclass focused on the cultural deficiencies of the poor. In *The Underclass*, journalist Ken Auletta (1982) underscored the character flaws of this subgroup of the poor. Other publications such as George Gilder's (1981) *Wealth and Poverty*, Charles Murray's (1984) *Losing Ground*, and Lawrence Mead's (1986) *Beyond Entitlement* also characterized the underclass in terms of cultural inadequacies and a predilection for idleness, dependency, and deviance.

Undoubtedly, these publications resonated with the public's difficulties in understanding the escalating rates of personal and property crimes in the major cities of the USA, despite the hundreds of millions of dollars that were spent to combat urban social problems during the 1960s and 1970s. Public discomfort was also fueled by journalistic accounts with lurid descriptions of depravity, squalor, and crime among the urban poor (*Time* 1977). Several empirical assessments of the size and growth of the underclass did appear during this time, but for the most part, rather than help to draw attention to the underlying causes of chronic poverty, these studies contributed to the perception that the underclass were, as a group, synonymous with deviant practices (Ricketts and Sawhill 1988, Lehmann 1986, Nathan 1987). Growing public opinion that the poor were incapable of self-regulation gave credence to those who argued for more punitive policies rather than incentives and assistance (Katz 1993).

Concern about the existence of an underclass has not been limited to the USA. Recent social, economic, and immigration trends occurring in the UK have resulted in a great deal of speculation about whether they portend the emergence of a British underclass. Initial attempts to chronicle the rise of a highly disadvantaged, heterogeneous group focused on issues of class inequality and citizenship rights. Dahrendorf (1985) argued that the defining characteristic of the underclass in that country is the marginal status of the group in terms of its representation by, or involvement in, social and political organizations, a marginalization which has largely occurred because of macroeconomic shifts and contractions in the welfare state. Unlike in the USA, however, the British underclass is less likely to be spatially segregated in high-poverty urban areas.

More recently, the discussion has centered on amoral and deviant behavior as the root cause of persistent poverty, a turn influenced in part by American conservative thought, especially that of Murray (1990). As in the USA, research on the effects of structural barriers and economic marginality must compete with postulations about whether the chronically poor are different in character and comportment.

2. *The Relationship Between Economic Restructuring and the Urban Underclass*

Proponents of the structuralist approach claim that culture-centered explanations of the underclass fail to sufficiently acknowledge the role played by the social

transformation occurring in many US inner cities during the 1960s and 1970s. New social and economic conditions in the nation's oldest cities were creating a confluence of forces which were effectively isolating a subsection of the poor from the rest of society. Economic restructuring in urban areas in the northeast and midwest brought about a dramatic decline in demand for low-skilled workers in the goods-producing sector of the economy, which disproportionately hurt poor unskilled black male workers and helped to elevate poverty rates in inner-city neighborhoods.

The out-migration of nonpoor families also contributed to the growth of concentrated inner-city poverty. Encouraged by technological advances, improvements in transportation, and increased employment opportunities outside of metropolitan areas, whites, by 1968, had already begun their out-migration from the inner cities. In addition, because of the Civil Rights Acts of 1964 and 1965 and the Fair Housing Act of 1968, oppressive restrictions on black employment opportunities and residential patterns were outlawed. This increased the access of black skilled and semiskilled workers to higher paying jobs in the broader economy. Greater labor market opportunities, a concomitant improvement in their standard of living, and a reduction in the overt discriminatory practices of banks and real estate brokers, which had long mired blacks of all socioeconomic groups in some of the most dilapidated areas of the cities, resulted in a significant out-migration of middle class blacks from urban ghettos beginning in the early 1970s.

The declining economic fortunes of black unskilled workers and the neighborhoods they resided in resulted in the increasing 'ghettoization.' In other words, poor urban blacks who had weak attachments to the labor force, had become more spatially, socially, and economically segregated from more advantaged areas (Wilson 1987). Accordingly, any discussion of dysfunctional cultural traits of a ghetto underclass ought to be conjoined with a discussion of the structural forces which undergrid the social and economic life of central cities.

3. *Structure and Culture in Poor and Jobless Neighborhoods*

A more nuanced view of underclass or ghetto-related cultural traits is slowly taking shape, one that avoids some of the unproductive debates surrounding the culture-of-poverty thesis. In this view, the existence of a ghetto cultural milieu which influences the behavior of the poor in ways that are often counter-productive, is explicitly acknowledged. However, it also highlights the complex interrelationship between structural circumstances and behavioral and attitudinal responses by showing how both the cultural milieu and dysfunctional behavior are themselves shaped by the social constraints faced by poor ghetto residents (Wilson 1996).

A considerable body of literature now exists which posits that economic stratification in high-poverty neighborhoods has deleterious effects on the social aspects of community by corroding cohesion among neighbors, lessening the likelihood that people will find work through personal acquaintances, and undermining the institutional framework which anchored these neighborhoods (Brooks-Gunn *et al.* 1997a, 1997b). These conditions create the seedbed for what Wilson (1996) later referred to as 'ghetto related' behavior or the cultural adaptations of the ghetto poor to their spatial and social exclusion from mainstream America. Building on his earlier conception of the underclass, Wilson now explicitly associates the concept of 'underclass' with the dual problem of marginal economic position and social isolation in concentrated poverty areas (Wilson 1991). In other words, what distinguishes members of the underclass from those of other economically disadvantaged groups is that their marginal economic position or weak attachment to the labor force is uniquely reinforced by their neighborhood or social milieu. The dual problem of weak labor force attachment and neighborhood disadvantage is not captured by the standard designation 'lower class.'

Despite this attempt to formulate discussions of the term 'underclass' in more theoretical terms to increase its scientific import—that is, its role in the description, explanation, and prediction of behavior—both journalists and scholars alike continue to appropriate the label and imbue it with meanings that serve their own ideological ends. Many of the scholars who used the term in earlier writing, including Gunner Myrdal, never intended it to be a generic label for a host of traits that supposedly characterize a deviant, often criminal, faction of the poor who disavow the normative orientations of mainstream society. To be sure, crime is a characteristic feature of high-poverty neighborhoods, but criminal activities are as abhorrent to the majority of the chronic poor who must live in these depressed communities as it is to the rest of society that disparages them. Furthermore, welfare dependency and unemployment are as much an anathema to the poor who must endure them as they are to those who see these conditions as self-inflicted.

4. *Social Policy*

Changes in the global economy are creating daunting challenges for governments in the USA and Europe to improve or sustain existing employment rates at a time when national labor markets are increasingly subject to the vicissitudes of a global economic system. Growing numbers of people, including those who are socially isolated in concentrated poverty areas, are being excluded from the workforce because of lack of

minimal skills and have to rely on the state to maintain even the most meager economic subsistence. In the USA, the growing conservative backlash against swelling welfare rolls has resulted in punitive welfare reform aimed at curbing a so-called dependency mentality, while the inequality and social dislocation brought on by global economic processes is largely ignored.

As in America, efforts to address social and economic inequities in other parts of the world are becoming mired in debates about whether joblessness should be addressed at the individual or the national level. It is imperative that whatever terminology is used to describe the heterogeneous amalgam of people who reside in impoverished areas and are forced to operate outside the labor force, that efforts be made to increase public understanding of their economic and social woes. Support for public policy options will very likely depend on how effectively we communicate the following message: the plight of the underclass has much more to do with dwindling opportunities for unskilled and semiskilled people to eke out a living than with individual choices to remain poor, unemployed, or dependent on state handouts.

See also: Class: Social; Crime: White-collar; Discrimination; Discrimination, Economics of; Discrimination: Racial; Minorities; Poverty, Culture of; Poverty in History; Poverty Policy; Racial Relations; Racism, History of; Racism, Sociology of; Residential Segregation: Sociological Aspects; Social Stratification; Spatial Equity; Urban Geography; Urban Poverty in Neighborhoods; Urban Sociology; Urban Studies: Overview

Bibliography

Auletta K 1982 *The Underclass*. Vintage Books, New York

Brooks-Gunn J, Duncan G J, Aber J L (eds.) 1997a *Neighborhood Poverty Vol I: Context and Consequences for Children*. Russell Sage Foundation, New York

Brooks-Gunn J, Duncan G J, Aber J L (eds.) 1997b *Neighborhood Poverty Vol. II: Policy Implications in Studying Neighborhoods*. Russell Sage Foundation, New York

Dahrendorf R 1985 *Law and Order*. Westview, Boulder, CO

Gilder G 1981 *Wealth and Poverty*. Basic Books, New York

Glasgow D G 1980 *The Black Underclass: Poverty, Unemployment and Entrapment of Ghetto Youth*, 1st edn. Jossey-Bass Publishers, San Francisco

Katz M B 1989 *The Undeserving Poor: From The War on Poverty to the War on Welfare*, 1st edn. Pantheon, New York

Katz M B (ed.) 1993 Reframing the 'underclass debate'. In: Katz M B *The 'Underclass' Debate: Views from History* Princeton University Press, Princeton, NJ

Lehmann N 1986 The origins of the underclass. *Atlantic Monthly* June: 31–55 (Pt. 1) and July: 54–68 (Pt. 2)

Mead L M 1986 *Beyond Entitlement: The Social Obligations of Citizenship*. Free Press, New York

Murray C 1984 *Losing Ground: American Social Policy 1950–1980*. Basic Books, New York

Murray C 1990 *The Emerging British Underclass*. Institute of Economic Affairs, London

Myrdal G 1962 *Challenges to Affluence*. Pantheon, New York

Nathan R P 1987 Will the underclass always be with us? *Society* **24**: 57–62

Ricketts E R, Sawhill I V 1988 Defining and measuring the underclass. *Journal of Policy Analysis and Management* **7**: 316–25

Time 1977 The American underclass. (August 29): 14–27

Wilson W J 1978 *The Declining Significance of Race: Blacks and Changing American Institutions*. University of Chicago Press, Chicago

Wilson W J 1987 *The Truly Disadvantaged: The Inner City, The Underclass, and Public Policy*. University of Chicago Press, Chicago

Wilson W J 1991 Public policy research and the truly disadvantaged. In: Jencks C, Peterson P E (eds.) *The Urban Underclass*. The Brookings Institution, Washington, DC

Wilson W J 1996 *When Work Disappears: The World of the New Urban Poor*, 1st edn. Knopf, New York

W. J. Wilson, J. M. Quane, and B. H. Rankin

Unemployment and Mental Health

1. The Term Unemployment

How unemployment is defined and operationalized determines who is included (and excluded) as participants in empirical unemployment research and thereby to whom findings generalize as well as having important administrative, political and ideological assumptions.

Until recently, in so-called 'developed' industrialized countries, unemployment has been widely operationalized in terms of eligibility for State funded unemployment-specific financial support. However, although this embodies politically and administratively useful criteria, it is arbitrary from a social scientific perspective. Changes in the number of people who cease to be recorded as unemployed are not always the same as changes in the numbers of people at risk of psychological consequences of unemployment.

The International Labour Organisation counts people as unemployed if they have not undertaken any paid employment in the survey reference week, have looked for a job within the previous four weeks, and are available to start one within two weeks. This is preferable to the claimant count, but should jobless people who cease to actively look for employment due to hopelessness, depression, and/or demoralization, be considered any the less unemployed for social and behavioral purposes, especially if the psychological states which militate against active job search turn out to be consequences of unemployment?

The Broad Labour Force Survey counts all those who do not have but want a paid job and are available to start within two weeks but who have not necessarily been actively looking for a job. However, preferable though this operationalization is, it may still greatly underestimate the number of people who do not self-perceive themselves as satisfactorily employed and who are looking for employment. Many unemployed people take places on employment and training schemes only through fear of losing their state financial support if they decline. Others take part-time, short-term, or insecure jobs or become 'underemployed' only quasivoluntarily.

These people, along with those indicated by the Broad Labour Force Survey, constitute the so-called 'slack' labor force.

2. *The Size of the Unemployment Problem*

These different ways of operationalizing unemployment suggest very different numbers of unemployed people. For example, in late 1999 in the UK, unemployment as indicated by 'claimant eligibility' stood at 1,209,600. As indicated by the International Labour Organisation the figure was 1,714,000, while the Broad Labour Force Survey gave a figure of 2,396,000 for the same period. The slack labor force, on the other hand, gave a figure of 4.6 million. Beyond the UK, the International Labour Organisation calculated unemployment to stand at 34 million in the countries of the Organisation for Economic Cooperation and Development in 1996.

3. *The Relationship Between Unemployment and Mental Health*

There have been valid grounds for concern about the social, physical, and mental health consequences of unemployment for at least 200 years, and this has been beyond reasonable doubt since the 1930s (Eisenberg and Lazarsfeld 1938, Lazarsfeld-Jahoda and Zeisel 1933). In the early days, fear, frustration, irritability, declining self-respect, decreasingly meaningful sense of time, loneliness, apathy and resignation, and poorer physical health were noted. These were broadly confirmed by substantial empirical research and scholarship in the 1980s but standardized measures checked for reliability and validity were preferred. Anxiety, depression, positive and negative self-esteem, and a variety of measures of affect were studied repeatedly. Rather than 'demoralization,' researchers wrote of 'psychological well-being' and the General Health Questionnaire was used widely. That there are negative mental health consequences of unemployment is being reiterated currently by sophisticated reviews and meta-reviews (see *Depression, Hopelessness, Optimism, and Health*.

Research clarifying the relationship between unemployment and mental health has been carried out in Australia, New Zealand, USA, and many European countries over many decades. The resulting large literature has operationalized mental health in a wide range of ways, has been carried out using just about every viable research method, has been supported by diverse funding arrangements, and has been undertaken from a variety of ideological starting positions. Of course the impact of unemployment is not homogenous: some are affected more negatively than others, and a number of moderator variables have been identified. Effective coping with unemployment is also widespread. However, the degree of consensus is startling. Groups of unemployed people almost always have poorer mental health than groups of otherwise similar but employed people and unemployment is overwhelmingly found to be the cause of poor mental health rather than the result of it.

For many, the well-designed, longitudinal studies, which have tracked large, carefully matched, samples of people in and out of paid jobs using standardized measures, are the key studies. Some of the most persuasive longitudinal quantitative studies of all were done with young people. Typically, these studies measure the mental health of large groups of young people in school and follow them out into the labor market, periodically measuring the mental health of those who get jobs and those who do not, and comparing group mean scores cross-sectionally and longitudinally. Groups of unemployed youngsters repeatedly are demonstrated to have poorer mental health than their employed peers but statistically significant differences are seldom found between the scores of the same groups when at school.

Many such studies have been done, but three research programs have been particularly influential. The first was the 1980s research program based at the Social and Applied Psychology Unit in Sheffield, England (see Warr 1987 for an overview). The second program was based at The Flinders University of South Australia (see Feather 1992 for an overview). The third was based at the University of Adelaide (see Winefield et al. 1993 for an overview). In brief, such studies provide powerful evidence that unemployment causes, rather than merely results from, poor psychological health.

Recently, sophisticated meta-reviews have reiterated that social causation is involved. For example, Murphy and Athanasou (1999) conducted a meta-review of 16 longitudinal studies using valid and reliable measures and published in the last 10 years in English-language scientific journals. They concluded that 'the results from the 16 longitudinal studies suggest that unemployment has reliable (negative) effects on mental health' and that 'effect size information … suggests that moving from unemploy-

ment to employment not only produces a reliable change to mental health, but a change that is "practically significant."'

It was never remotely plausible that mass unemployment was caused by mass, sometimes organizationally confined, epidemics of mental illness, but of course, in some individual cases, pre-existing mental health problems do predispose people to loss of, and/or failure to get, a job. However, even here social causation processes may be hard to discount completely. For example, during economic recession some employers ratchet up the criteria they use to select employees with the result that people are excluded from employment on mental and physical health grounds at one stage of the economic cycle, who would have been employed in other economic circumstances. Other conceptual and methodological concerns suggesting the stark dichotomy between social causation and individual drift may be unhelpful are discussed more fully in Fryer (1997).

In conclusion, research has persuaded most researchers in the field that unemployment has mental health consequences that are negative and widespread.

4. *Disguised Unemployment*

According to Dooley and Catalano (1999), in the US 'under-employment (involuntary part-time or poverty wage) both appears increasingly common and seems to have health effects more like those of unemployment than adequate employment.' Part-time work, on-call contracts, fixed-term contracts, seasonal work, agency work, home-working, tele-working, freelancing, self-employment, and informal work are all on the increase. In 1992, 9.3 percent of workers in the European Union were on temporary contracts, with more than three out of five of such jobs being in low-skilled occupations. About 15 percent of the European workforce was employed part-time in 1991, and the percentage is increasing year on year. However, there is massive variation between countries: Nearly a third of employment in the Netherlands is part-time as opposed to less then 4 percent in Greece, for example. More than half of part-time work in the European Union is done by low-skilled, low-paid workers. Twice as many women as men are employed part-time. The trend looks set to continue. In May 1999, the OECD recommended member countries to make working practices more flexible: abolishing employment protection policies; ensuring minimum wages for young people are not too high; switching from unemployment to in-work benefits.

What are the psychological consequences of disguised unemployment? As long ago as 1938, it was recognized that 'just having a job itself is not as important as having a feeling of economic security. Those who are economically insecure, employed or unemployed, have a low morale' (Eisenberg and Lazarsfeld 1938). Advances in clarifying the concept of job insecurity came with Hartley et al. (1991). Studies in Israel, the Netherlands, and the UK showed that job insecurity is associated with experienced powerlessness, impaired mental health (depression and reported psychosomatic symptoms), reduced job satisfaction, reduced organizational commitment, reduced trust in management, resistance to change, and poorer industrial relations. Burchell et al. (1999) reported that job insecurity has increased gradually since the 1960s for both men and women in both full-time and part-time employment and is now higher than at any time since 1945. While job insecurity traditionally has been accepted as problematic for blue-collar workers, professional workers, who in 1987 were the most secure workers, were found to be the most job insecure 10 years later. Job insecurity was associated with poorer psychological wellbeing, poorer family relations, and organizational demotivation. Both physical and mental wellbeing was found to continue to deteriorate rather than stabilize with continuing job insecurity.

Evaluation of mental health consequences of participation in employment and training schemes reveals a complex picture. Some cross-sectional research found trainees' mental health was not significantly worse than that of an employed comparison group and was significantly better than an unemployed comparison group. However, other research found that trainees were significantly more depressed than an employed comparison group and not significantly less depressed than an unemployed comparison group. Still other research revealed a multifaceted experience of 'community program' with participants positive about some aspects and negative about others. The nature of the labor market context, quality of training and job placement, and degree of experienced compulsion are probably critical. Some research has found that unemployed people, of any age (male or female), who were coerced into employment through workfare-type programs used medical facilities more frequently. Their use of primary care decreased if they became unemployed again

Re-employment is likely to be at a lower level, yet re-employment in poor-quality jobs can be as psychologically corrosive as unemployment. Kaufman (1982) reported that 20 percent of his re-employed professionals were under-employed (in terms of salary, type of work, and skill use), and less than 50 percent said their lives had returned to normal after re-employment. Fineman (1987) found 'those re-employed in jobs which they felt to be inadequate were experiencing more stress and even poorer self-esteem, than they had during their period of unemployment.' Graetz (1993) showed that 'the benefits of employment are confined to those, albeit a majority, who manage to find a satisfying job. In contrast, those who do not end up in

a satisfying job—approximately one in every five workers—report the highest levels of health disorders. At the same time, the adverse consequences of job loss are confined to those who were satisfied with their former jobs.' More recently still, Leana and Feldman (1995) report that 'unsatisfactorily re-employed workers reported significantly higher levels of anxiety and psychological distress than those who were satisfactorily re-employed; they also reported significantly lower levels of life satisfaction than those who were unemployed.'

5. *Methodological Issues*

The method of choice of most recent research on unemployment and mental health has been to contrast the better mean mental health of employed with the worse mean mental health of unemployed people in both cross-sectional and longitudinal designs. In the past the mental health of employed groups has been reliably better than that of otherwise comparable unemployed groups. However, should the mental health of the majority of those in employment deteriorate, the formerly reliable finding that unemployed respondents have poorer mean mental health than otherwise similar employed people would no longer be found, without it being the case that unemployment was any the less pathogenic. With the apparently inexorable march of the 'flexible labor market,' it may now be only a matter of time until that happens. The realization that comparison of labor market transition outcomes is unsatisfactory in certain economic circumstances should alert researchers to the fact that it has always relied upon ideologically problematic assumptions about the salutogenic nature of employment.

Rather than relying on discovering differences in outcomes, researchers must increasingly try to uncover the detailed multilevel processes through which the social causation of mental health problems by unemployment occurs. This is likely to require use of a much wider range of ecologically contextualized and collaborative research methods, including qualitative, action research, and participatory methods. Given the vast scale of socially caused health problems attributable to unemployment, training clinicians and treating negatively affected individuals one at a time is practically impossible as well as ineffective and offensive. Prevention and structural intervention must be priorities.

6. *Future Directions of Theory and Research*

Radical rethinking of the alleged dichotomy between employment and unemployment is required. The problems of unemployment are still widely and mistakenly seen as merely the problems of the absence of employment. For many unemployed people the search for employment and the hoops through which they have to jump in order to keep body and soul together is a stressful low-status, insecure, part-time, poorly paid job in itself, carried out in poor working conditions, with virtually no negotiating rights or scope for collective action under potentially punitive state surveillance and with a high risk of occupational strain. Unemployment is ready for job redesign.

See also: Unemployment: Structural

Bibliography

Burchell B et al. (eds.) 1999 *Job Insecurity and Work Intensification*. Routledge, New York

Dooley D, Catalano R 1999 Unemployment, disguised unemployment and health: The U.S. case. *International Archives of Occupational and Environmental Health* **72** (Suppl. S): S16–S19

Eisenberg P, Lazarsfeld P F 1938 The psychological effects of unemployment. *Psychological Bulletin* **35**: 258–390

Feather N T 1992 *The Psychological Impact of Unemployment*. Springer Verlag, New York

Fineman S 1987 Back to employment: Wounds and wisdoms. In: Fryer D, Ullah P (eds.) *Unemployed People: Social and Psychological Perspectives*. Open University Press, Milton Keynes, UK

Fryer D 1997 International perspectives on youth unemployment and mental health: Some central issues. *Journal of Adolescence* **20**: 333–42

Graetz B 1993 Health consequences of employment and unemployment: Longitudinal evidence for young men and women. *Social Science and Medicine* **35**(6): 715–24

Hartley J, Jacobson D, Klandermans B, van Vuuren T, Greenhalgh L, Sutton R 1991 *Job Insecurity: Coping with Jobs at Risk*. Sage, London

Kaufman H G 1982 *Professionals in Search of Work: Coping with the Stress of Job Loss and Underemployment*. Wiley, New York

Lazarsfeld-Jahoda M, Zeisel H 1933 *Die Arbeitslosen von Marienthal. Psychol. Monographien*, 5 (1971 *Marienthal: The Sociography of an Unemployed Community*. In: Jahoda M, Lazarsfeld P F, Zeisel H 1971 Aldine Atherton, Chicago)

Leana C R, Feldman D C 1995 Finding new jobs after a plant closing: Antecedents and outcomes of the occurrence and quality of reemployment. *Human Relations* **48**(12): 1381–1401

Murphy G C, Athanasou T A 1999 The effect of unemployment on mental health. *Journal of Occupational and Organizational Psychology* **72**: 83–99

Warr P B 1987 *Work, Unemployment and Mental Health*. Clarendon Press, Oxford, UK

Winefield A H, Tiggemann M, Winefield H R, Goldney R D 1993 *Growing Up with Unemployment: A Longitudinal Study of its Psychological Impact*. Routledge, London

D. Fryer

Copyright © 2001 Elsevier Science Ltd. 15951
All rights reserved.

 ISBN: 0-08-043076-7

Unemployment: Structural

1. Definitions

It is an old tradition in labor economics to distinguish between structural, frictional, and cyclical unemployment. Structural unemployment is then envisioned as a result of the institutional set up of the economy, including private and government organizations, types of market arrangements, demography, laws, and regulations. In the literature, the importance of these institutional features for structural unemployment is tied particularly to their implications for demand for and supply of labor, price and wage formation, and the efficacy of search and matching processes in the labor market. Frictional unemployment may be regarded as a subset of structural unemployment, mainly reflecting temporary unemployment spells as the result of job search and matching difficulties in the connection with quits, new entries to the labor market, and job separation because of the employers' dissatisfaction with individual workers.

Cyclical unemployment differs from structural and frictional unemployment by basically being tied to short-term economic fluctuations. An empirical illustration of the importance of structural unemployment as compared to cyclical is that variations in actually measured unemployment rates have turned out to be much larger between cycles than within cycles, presumably reflecting differences in structural unemployment (Layard et al. 1991). For example, the average unemployment rate over the business cycle in Western Europe has moved from about 3 percent in the mid-1960s to 6 percent in the mid-1970s and to 10 percent from the mid-1980s to the late 1990s. But as pointed out below, various unemployment-persistence mechanisms blur the distinction between cyclical and structural unemployment.

In economic theory, structural and cyclical unemployment usually are regarded as disequilibrium phenomena in the sense that they reflect excess labor supply at existing wages and hence that the labor market does not clear. Then, individual employers informally ration jobs. Nevertheless, technically (analytically) structural unemployment often is analyzed in terms of the concept of *equilibrium unemployment*. This means that the aggregate-unemployment level is in a 'state of rest': existing excess labor supply is assumed to last as long as certain characteristics (parameters) of the economy are unchanged. If this equilibrium is dynamically stable, unemployment equilibrium may also be described as an unemployment level toward which the economy moves as long as new disturbances do not emerge. (The term 'equilibrium' is used in several different ways in economics. It refers sometimes to the equality between demand and supply in a market, i.e., to traditional 'market-clearing,' in other cases to a state that tends to continue over time regardless of whether the market clears or not. The concept of 'equilibrium unemployment' is an example of the latter.)

It is useful to distinguish between two main analytical approaches to equilibrium unemployment: *stock approaches* and *flow approaches*. Stock approaches focus on the difference, at a given point in time, between the workforce desired by firms (aggregate *stock demand* for labor) and the number of workers willing to work (aggregate *stock supply* of labor). Flow approaches deal with the difference between the flows in and out of the unemployment pool during a certain period. Flow-demand for labor is represented then by the supply of 'job slots,' and flow-supply by offers of workers to fill such slots. These two approaches are related because stock demand and stock supply may also be derived in flow approaches. Let us start with stock approaches.

2. Stock Equilibrium

Milton Friedmans's celebrated 'natural unemployment rate' is a stock-equilibrium concept of structural unemployment: 'the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demand and supplies, the costs of gathering information about job vacancies and labor availabilities, the costs of mobility and so on' (Friedman 1968). Apparently, this concept of equilibrium unemployment covers structural and frictional unemployment as described earlier.

The nonaccelerating inflation rate of unemployment (NAIRU) is close to Friedman's natural rate, though it emphasizes (more than Friedman does) the non-clearing character of the labor market. The NAIRU is defined explicitly as the unemployment rate at which inflation is constant. It is derived from an expectations-augmented Phillips curve, i.e., a function that assumes that the price-inflation rate is a decreasing function of the unemployment rate and an increasing function of the expected rate of future inflation (Phelps, 1967). In the simplest possible terms (with linear relations):

$$\pi = -au + b + c\pi^e \qquad (1)$$

where π is actual inflation, u is the unemployment rate, and π^e is expected inflation. For given values of the (positive) parameters a, b, and c, the equation may be depicted as a set of downward sloping short-term Phillips curves (SR_i), one curve for each set and for each value of π^e (see Fig. 1).

In the context of this model, usually (and realistically) it is assumed that economic agents ask for full compensation for expected inflation in the *long run*;

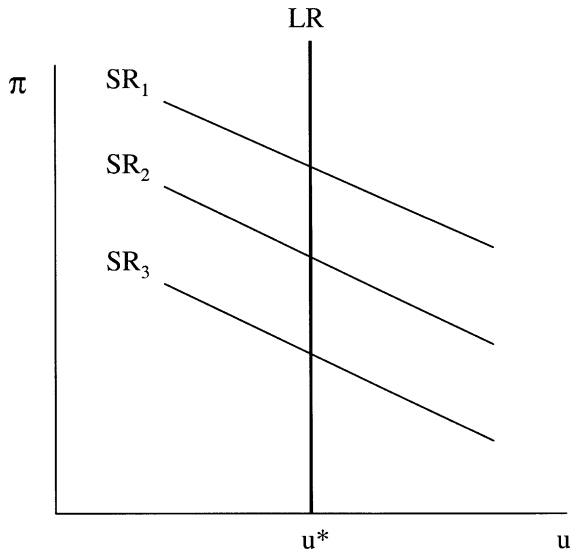

Figure 1
Phillips curves and the NAIRU

thus $c = 1$ in that time perspectives. It is usually also assumed that economic agents do not make *systematic* expectational errors in the long run; thus $\pi = \pi^e$. The only unemployment rate consistent with these two assumptions is, from Equation (1), $u^* = b/a$, where * denotes the equilibrium unemployment rate. It is often also assumed that inflation in this situation is constant over time (and not just that actual and expected inflation coincide), because agents have no reason to revise their price and wage setting behavior when inflation is exactly what they had expected when they formed their decisions. u^* is then the NAIRU (the unemployment rate at which inflation is constant), geometrically represented as a vertical long-run Phillips curve, *LR* in Fig. 1. Inflation increases to the left of that level and falls to the right. Because inflation in this framework in the long run is independent of unemployment, there is no long-run trade-off between these variables. But it has been argued by Lucas (1972) that there is not even a *systematic* short-term trade-off if economic agents have 'rational expectations' (thus, if they never make systematic expectational errors), and if various agents are free to revise their prices and wages in response to changes in expectations. Such a trade-off will then only arise due to random (non-systematic) expectational errors, e.g., due to random shocks or randomized government policy.

An alternative exposition of stock-equilibrium in the labor market, and hence structural unemployment, gained ground in the 1980s and 1990s; it emphasizes the requirement of consistency of price- and wage-setting behavior (Shapiro and Stiglitz, 1984, Layard and Nickell, 1986). Figure 2 gives a simple diagrammatic illustration, with aggregate employment (N) on the horizontal axis and the aggregate real-wage level (w) on the vertical. The *LS* curve is the traditional aggregate labor supply curve, i.e., the sum of labor supplied by households. The *PS* curve is the price-setting curve of firms: for alternative unemployment levels and given nominal wages, the curve defines the price level desired by profit-maximizing firms that operate in imperfectly competitive product markets. Or the *PS* curve can be described as a labor-demand relation, because it also expresses the combination of employment and real wages desired by profit-maximizing firms. Unemployment is then the horizontal difference between the *PS* and *LS* curves.

The *WS* curve defines wage-setting behavior. It depicts the influence of the (un)employment situation on the real-wage rate that the wage-setting process generates. The intuition is that firms feel compelled to offer higher real wages and that workers and unions demand higher real wages when the labor market is tight (low unemployment) than when it is slack. So by contrast to models with perfectly competitive labor markets, a distinction is made between labor supply (by households) and wage setting (by firms, unions, or more realistically a bargaining process).

The equilibrium real-wage rate is now w_0, equilibrium employment N_0 and equilibrium unemployment U_0, the latter being the only unemployment level at which there are no incentives for agents to change prices relative wages and vice versa (corresponding to the equilibrium unemployment rate u^* in Fig. 1). This model will be called the *PS–WS* model of equilibrium (structural) unemployment.

As an illustration, suppose that the real-wage rate in Fig. 2 is initially w_0 but that aggregate employment happens to be above N_0, e.g., as a result of positive demand shocks to which nominal magnitudes have not yet adjusted. Firms and workers are then dissatisfied with the existing real wage w_0. Workers and their unions try to raise nominal wages (at existing nominal prices), as can be read off from the *WS* curve, and firms try to raise prices (at existing nominal wages), as can be read off from the *PS* curve. The result is a wage-price spiral, as described by the expectation-augmented Phillips-curves model. Firms also start cutting their work force because their desired level is lower than the actual one. Aggregate employment continues to fall until it reaches the unique equilibrium level N_0. The same reasoning, *mutatis mutandi*, holds if aggregate employment is initially lower than the equilibrium level.

3. *Flow Equilibrium*

Flow approaches to structural unemployment instead emphasize the flow of workers in and out of un-

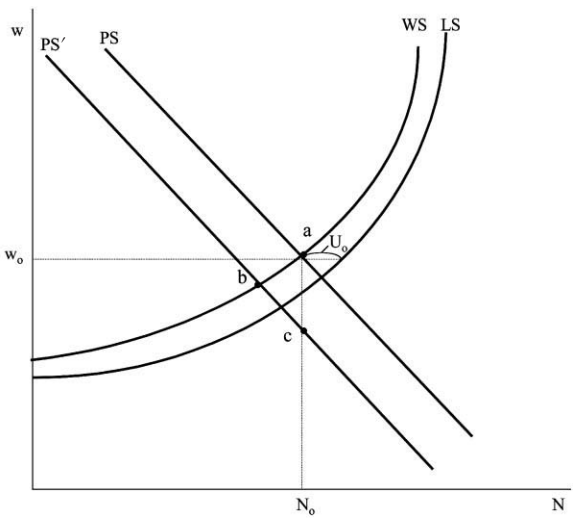

Figure 2
The PS—WS model of equilibrium unemployment

employment (Phelps 1970, Hall 1979, Diamond 1982). These approaches explicitly grant the heterogeneity of jobs and workers, which means that agents in the labor market must devote time and costs to search to match up idiosyncratic preferences, skills, and skill requirements. In the process, unemployment and vacancies coexist because the latter are not eliminated immediately by the hiring of workers in connection with decentralized negotiations of wages between firms and work applicants.

For simplicity, let us regard the size of the labor force (L) as constant. Further, let s denote the job separation rate, i.e., the fraction of employed workers (N) who lose their jobs in each time period. Let f denote the rate of job finding, i.e., the fraction of unemployed workers (U) who find a job in each time period. We assume that both fractions are constant. The equilibrium unemployment rate (u^0) is then defined as a situation when the number of individuals finding jobs equals the number of individuals who are separated from jobs; hence $fU = sN$. Noting that $N = L - U$, i.e., that the number of unemployed workers equals the labor force minus the number of employed, we get $fU = s(L - U)$. Rearranging terms and dividing both sides by L, we obtain the equilibrium, or structural, unemployment rate $u^0 = U/L = s/(s + f)$. As expected, the higher the rate of job separation (s), the higher the equilibrium unemployment rate. And the higher the rate of job finding (f)—due to greater intensity and efficiency of job search and hence better job matching—the lower the equilibrium-unemployment rate.

Though the model basically is designed to highlight frictional unemployment, as described earlier, unemployment in the connection with job search and matching may also be due to continuing changes in the composition of demand and supply of different types of workers. This means that flow models of search and job matching also cover elements of what usually is regarded as structural unemployment, which illustrates the previous assertion that frictional unemployment may be regarded as a subset of structural unemployment.

It is natural and feasible to integrate flow and stock approaches to (un)employment determination. For instance, the efficacy of search and matching processes influences the positions of the *PS* and *WS* curves in Fig. 2 (and the positions of the Phillips curves in Fig. 1). The longer it takes firms to find the right workers for their vacancies, and the longer it takes work applicants to acquire suitable job offers, the more costly it is to hire workers, and the lower will be the labor-demand curve—the *PS* curve in Fig. 2. A poorly functioning search and matching process will force recruiting firms to offer higher wages than otherwise; the result is also a higher *WS* curve. In this way, flow models are related to previously-discussed stock models. Search and matching processes may also be seen as part of the short-term *dynamics* by which stock equilibrium is reached, with positive gross flows in and out of unemployment, netting to zero when there is flow and stock equilibrium.

4. *Microeconomic Foundations*

All these macroeconomic models of structural unemployment assume that unemployed workers are not able or willing to get jobs by underbidding the prevailing wages of incumbent workers. The most obvious microeconomic explanation of the absence of wage underbidding is perhaps *minimum wage laws*. But there seems to be rather general agreement among labor market economists that minimum wages have not been high enough in developed countries to explain much of aggregate structural unemployment in that part of the world. But in some countries, minimum wage laws are binding for specific types of workers, such as inexperienced, unskilled, and physically or mentally handicapped people. (In the special case when a firm has a monopsonistic position in a local labor market minimum wages may, up to a certain wage level, increase rather than reduce employment. In this case, the minimum wage simply prevents the firm from keeping down wage costs by holding back its demand for workers.)

There may also be informal *social norms* against the underbidding of the wages of incumbent workers. In other words, wage underbidding may not be socially acceptable behavior among workers, and the acceptance of such offers may not be socially acceptable

behavior among firms. Legislation may support such norms.

But more elaborate microeconomic foundations of structural unemployment have been developed since the 1980s. In the *union monopoly model*, unions are assumed to strive for a trade-off between real wages and unemployment. Often, this trade-off is derived formally from a model where unions maximize the expected utility of union members—including those who are employed and those who wind up unemployed due to the union's wage policy. So labor unions are assumed conscientiously to choose a real-wage rate that generates some structural unemployment.

In *union-bargaining models*, the real wage is instead set in a bilateral bargaining process between a union and a firm or an association of firms. Such models qualitatively yield the same result as the union-monopoly model, though the real wage will normally be lower and hence the employment level higher, because the firm is assumed to strive for a lower real-wage rate than the union. But neither type of union model answers the questions why there is no underbidding of existing wages by nonorganized workers and why unemployed union members do not leave the union.

There are also theoretical reasons and empirical evidence why the organization of bargaining, such as the degree of coordination and centralization of bargaining, would influence the outcome of the bargaining process and hence also the unemployment level. The most generally accepted hypothesis is probably that both highly centralized and highly decentralized bargaining are more conducive to employment than wage bargaining on levels in between, such as industry-level bargaining (Calmfors and Driffill 1988).

The *insider–outsider* theory of unemployment emphasizes the asymmetric market powers of incumbent workers, *insiders*, and disenfranchised workers without jobs in the formal labor market, *outsiders* (Lindbeck and Snower 1988). Because hiring and firing costs protect incumbent workers, they can push up their wages above the reservation wage of outsiders without losing their jobs. Another way for insiders to prevent wage underbidding is by refusing to cooperate with underbidding outsiders in the production process and by threatening to harass them if underbidding workers are actually hired. With the help of such noncooperation and harassment threats, insiders (possibly with the support of unions) may also be able to create and sustain social norms in their own interest against wage underbidding.

In *efficiency wage* models, by contrast, the microeconomic foundations of structural unemployment are based on the incentive structure of firms (Yellen 1984). Firms, which are assumed to have all market powers in wage formation, set wages high enough to get a good selection of work applicants and to discourage workers from shirking and quitting. More specifically, firms set wages to balance the cost of higher wages against the productivity increase of having a more efficient work force. Only by accident would the wage level that is optimal for firms clear the labor market.

All these microeconomic theories are rather silent about the employment situation in *informal* labor markets, where minimum wages are not always respected, unions may not exist, collective bargaining agreements are not binding, efficiency wages are not important, and labor turnover costs are very low and hence insider-outsider mechanisms not important. Extreme examples are sellers of flowers or newspapers on street corners, shoeshine boys, and workers in the underground economy. One conceivable explanation why some workers who are excluded from the *formal* labor market remain unemployed rather than choose to work in informal labor markets may be an unwillingness to take jobs that are regarded as highly insecure and demeaning. Generous unemployment benefits help them to make this choice. It is significant that informal labor markets are particularly important in countries without unemployment insurance.

In the popular discussion of structural unemployment, it is often believed that labor-saving technological progress and related increases in labor productivity will necessarily result in higher unemployment. Workers are asserted to become redundant because machines replace workers—a vision of *technological unemployment* as a special variation on structural unemployment. A usual conclusion is that some kind of work sharing is necessary to avoid ever-increasing unemployment. An obvious flaw of this argument is that a gradual rise in productivity will normally be accompanied by higher real wages and profits in the long run, and hence of real income. In this sense, productivity growth creates the foundation for higher demand for output, though deficiencies in aggregate demand may occasionally occur and contribute to recessions. In fact, no long-term trend of the unemployment rate can be detected during the twentieth century.

So work sharing—in the form of a lower general pension age, more generous rules for early retirement, or fewer working hours—is not necessary to prevent gradually rising labor productivity from resulting in rising unemployment. Indeed, the literature in the field clarifies that work sharing is not even likely to contribute to reduced unemployment except in the very short run; one reason is that a reduction of the labor force tends to push up wages. But there is nevertheless a potential problem with continuing productivity growth. Some national economies may fail to adjust to changes in the *composition* of the demand for labor in the connection with uneven rates of productivity growth among sectors and types of workers (sector shifts or skill shifts in the labor market). This is a respectable explanation as to why

unemployment rates among low-skilled workers have increased so much (in percentage points) in Western Europe, where relative wages are quite rigid.

5. Unemployment Persistence

One serious limitation of the concept of unemployment equilibrium as an expression of structural unemployment is that this concept is difficult to distinguish empirically from unemployment persistence, i.e., inertia of unemployment after temporary shocks that have shifted the unemployment rate. Estimated equilibrium rates of unemployment tend to shadow the actual rate, which is an indication of unemployment persistence. Important examples of unemployment persistence are the persistently high unemployment rates after the deflationary policy pursued in several developed countries in the early 1920s and early 1930s; after the oil-price hikes in mid- and late 1970s; and after the restrictive monetary (and in some cases also fiscal) policy in the late 1980s and early 1990s. These experiences raise the suspicion that attempted statistical measurements of equilibrium unemployment also reflect the consequences of various persistence mechanisms. It is hard to find good reasons why the equilibrium unemployment rate in Western Europe should have increased as dramatically as the actual unemployment rate in the 1980s and 1990s.

There are also reasonable microeconomic explanations of unemployment persistence. While some explanations emphasize the behavior of outsiders, others point instead to the behavior of insiders. Examples of persistence mechanisms of the former type are the loss of skill among individuals who have been unemployed for a long time; endogenous changes in preferences in favor of leisure or household work; and the breakdown of social norms in favor of work and hence the emergence of unemployment cultures. We would expect reduced job search in all these cases. Long spells of unemployment may also functions as a negative signal to prospective employers about the quality of individual workers. An example of a persistence mechanism that operates via the behavior of insiders is that they (possibly supported by unions) may exploit subsequent business upswings to push up their wages, hence discouraging firms from expanding their workforce. All this suggests that unemployment persistence is an important factor behind what is usually regarded as structural unemployment. A basic conclusion is that endogenous changes of private behavior after a temporary unemployment-creating shock make the concept of *equilibrium unemployment* too narrow as a description of structural unemployment. For instance, it is realistic to regard prolonged high unemployment in Western Europe after the mid-1970s as the combined result of macroeconomic shocks and institutional rigidities that contribute to unemployment persistence. Hence, cyclical unemployment is asserted to have developed into structural unemployment. The result is strong 'history dependence' of the unemployment rate. So it is reasonable to widen the concept of structural unemployment to also include unemployment persistence.

6. Policies that Influence Structural Unemployment

Leaving the issue of legislated work sharing aside, what are the influences of government policy on structural unemployment? In the schematic *PS–WS* model presented earlier, the positions of the *PS*, *WS*, and *LS* curves reflect this influence. The discussion below is confined to taxes, transfer programs, regulations of job-protection, union powers, education/training, and other types of active labor market policy.

Taxes on labor insert a wedge between the labor costs paid by firms, the product wage $(1 + t_w)w$, and the return to workers, the consumption wage, $(1-t_i)w$. t_w and t_i then denote the wage tax (payroll tax), and the income tax. *The tax wedge*, expressed as the ratio between the product wage and the consumption wedge, is then

$$(1+t_w)/(1-t_i) \tag{2}$$

A consumption tax would enter the tax wedge in the same way as the wage tax. For example, in terms of Fig. 2, an increase in t_w shifts the labor-demand relation (the *PS* curve) down, e.g., to *PS*': at every real-wage rate (w), the demand for labor will be lower than before because the firm must now pay $(1 + t_w)w$ for its labor. As long as the *WS* curve is unchanged, the equilibrium position shifts from point *a* to point *b* in the figure, and the equilibrium employment level (N) and the real-wage rate (w) fall. These effects are relevant in the short run. The long-run effects are less obvious, and they depend on the extent to which wage setting adjusts to the tax wedge, which in turn depends on what happens to alternative incomes of workers, such as unemployment benefits and the return on household work. But for low-skilled workers, wages are often tightly regulated, which means that the wage costs for firms are likely to rise and labor demand will fall for this specific group of workers.

The results of empirical studies on these matters are somewhat ambiguous. But on balance, it seems that aggregate employment is negatively affected by wider tax wedges not only in a short- and medium-term perspective but also in the long run (Nickell and Layard 1999).

More generous *unemployment benefits* enable workers to be more selective about offered jobs and hence search for longer time periods—as highlighted by the earlier discussed search approach. Frictional unemployment (or *wait unemployment*) will then increase. Because the income loss in connection with

unemployment becomes smaller, firms must offer higher wages to prevent shirking and to keep down quits, as highlighted by efficiency wage models. Moreover, workers and unions become less hesitant to ask for high wages, as emphasized by union and insider–outsider models. This tendency is accentuated by the fact that unemployment benefits are, as a rule, financed by taxes rather than by insurance fees directly tied to specific workers. In terms of Fig. 2, the net effect of all this is that the *PS* curve is shifted down and the *WS* curve is shifted up, with rising equilibrium unemployment as a result. Unemployment persistence would also be predicted to increase.

Casual observation and systematic empirical studies, including cross-country regressions, suggest that generous unemployment benefits that may be kept for long periods (in some cases for several years or indefinitely), in fact, have effects of this type (Layard et al. 1991). Other evidence is that the exit rate for individuals from unemployment increases substantially in the US as soon as the benefit period expires after 26 weeks (Katz and Meyer 1990).

Job-security legislation is often singled out as another factor that tends to raise structural unemployment. Basically, such legislation raises the costs of hiring and of firing workers, the former because firms must consider the possibility of future firing costs when hiring workers. Firms will be less anxious to fire workers in recessions but also less interested in hiring workers in business upswings. A predicted effect is that employment tends to be stabilized at the actually existing level whatever this happens to be. In other words, (un)employment persistence would be expected to increase.

Since such legislation increases the market powers of insiders, this helps them to push up their wages. The predicted effect is an increase in the average unemployment rate over the cycle. Moreover, the market power of insiders and unions depends also on other legislation. Important examples are rules that extend the coverage of bargaining agreements to nonunion firms and give large freedom for unions to use conflict weapons, such as the rights to sympathy strikes, secondary actions and blockades, and the right to take out small 'key groups' in conflicts.

Most government policies discussed earlier tend to raise structural unemployment (including unemployment persistence). Of course, the other side of the coin is that a reversal of these policies is likely to reduce structural unemployment. But there are also well known examples of government policies directly aiming to lower structural unemployment. Competition policy, which accentuates competition in product markets, induces firms to increase production and employment at given real wages; in terms of Fig. 2, the *PS* curve shifts to the right. Increased competition in product markets also makes product demand more sensitive to changes in wage costs. This is likely to restrain the wage demands by insiders; the *WS* curve shifts down. The net effect would be lower equilibrium unemployment.

Active labor market policy is designed to keep workers out of open unemployment in the short term and to subsequently help them to get jobs. In particular, labor-exchange systems and retraining may improve matching between work applicants and vacancies. In terms of Fig. 2, the *PS* curve shifts up and the *WS* curve down with falling structural unemployment as a result. Moreover, subsidized education and training that raises labor productivity for (previously) low-skilled workers are likely to raise demand for them and reduce the upward wage pressure for skilled workers. This tends to generate similar shifts in the *PS* and *WS* curves as in the case of improved labor market exchanges. Unemployment persistence would also be expected to recede.

On balance, empirical studies suggest that labor-market exchange and properly conduced training programs, in fact, are useful for reducing structural unemployment (Layard et al. 1991); in the short term, also public works programs have such effects. But it is natural that the marginal return falls, possibly to zero, when the programs reach a large size (such as several percent of the labor force), and in particular when the programs are used largely to keep down statistically recorded figures on open unemployment.

See also: Labor Supply; Market Structure and Performance

Bibliography

Calmfors L, Driffill J 1988 Centralized wage bargaining and macroeconomic performance. *Economic Policy* **6**: 13–61

Diamond P 1982 Aggregate demand management in search equilibrium. *Journal of Political Economy* **90**: 881–94

Friedman M 1968 The role of monetary policy. *American Economic Review* **58**: 1–17

Hall R E 1979 Theory of the natural unemployment rate and the duration of employment. *Journal of Monetary Economy* **5**: 153–69

Katz L F, Meyer B D 1990 Unemployment insurance, recall expectations and unemployment outcomes. *Quarterly Journal of Economy* **105**: 973–1002

Lindbeck A, Snower D J 1988 *The Insider–Outsider Theory of Employment and Unemployment*. MIT Press, Cambridge, MA

Layard R, Nickell S 1986 Unemployment in Britain. *Economica* **53**: 121–69

Layard R, Nickel S, Jackman R 1991 *Unemployment: Macroeconomic Performance and the Labour Market*. Oxford University Press, Oxford, UK

Lucas R 1972 Expectations and the neutrality of money. *Journal of Economic Theory* **4**: 103–24

Nickell S, Layard R 1999 The labour market consequences of technical and structural change. In: Ashenfelter O C, Card D (eds.) *Handbook of Labour Economics*. North Holland Press, Amsterdam

Phelps E S 1967 Phillips curves, expectations of inflation and optimal unemployment over time. *Economica* **34**: 254–81

Phelps E S 1970 (ed.) *Microeconomic Foundations of Employment and Inflation Theory*. Norton, New York
Shapiro C, Stiglitz J E 1984 Equilibrium unemployment as a workers disciplinary device. *American Economic Review* **74**: 431–44
Yellen J 1984 Efficiency wage models of unemployment. *American Economic Review* **74**: 200–5

A. Lindbeck

Uneven Development, Geography of

'Uneven development' refers to the systematic processes by which social and economic change in capitalist societies enhances the wealth of some places at the expense of others. The resulting geographical differentiation in types and levels of development is not arbitrary but a systematic spatial expression of the social differences, inequalities, and processes that lie at the heart of capitalist societies. According to theories of uneven development, unevenness in space and time does not result simply from an accidental or statistical lack of evenness, nor does it result from local failures to adopt politically correct models of 'development' as might be suggested by contemporary doctrines of 'modernization' or 'globalization.' Rather, uneven development is a hallmark of capitalist societies, their geographical signature, a direct spatial translation of the logic of capital accumulation. The results of uneven development are easily understood in such contrasts as First World (centered on Europe, North America and Japan) vs. Third World (Africa, Asia, South America), city vs. suburb, and developed vs. underdeveloped regions. But processes and patterns of uneven development are highly dynamic and transform themselves as an integral expression of the ever-changing and evolving social relations of capitalism.

1. *Theories of Uneven Development*

Although a recognition of the unevenness of capitalist development can be found in various intellectual traditions, this notion is most fully developed in marxist and related critical traditions (see *Marxist Geography*). For Marx, the social relations of production and reproduction defined the specificity of any mode of production and the uneven geography of capitalist development emanates from the same source; uneven development is the elemental spatial form taken by the geography of capital accumulation. 'Capital grows in one place to a huge mass in a single hand because it has in another place been lost by many,' Marx famously argued (1987, p. 586). Marx displayed an evident sensitivity to the spatiality of social and economic change, indicating that specific geographies both facilitate and are produced by the accumulation and reproduction of capital, but he did not pursue the geography of capitalist expansion *per se* in any systematic way.

In the hands of later theorists, uneven development took a more political complexion. From his analysis of emerging capitalism in Russian industry and agriculture and its effects on the peasantry to his theories of imperialism, Lenin's work was infused with a sense of capitalism's systematically uneven geography (Lenin 1967, 1975), but it was Leon Trotsky who more than anyone crystallized the notion of 'uneven and combined development.' His focus was less on the economic geographic differentiation within and between societies than on the ways that such differentiation were bound up with and helped shape the possibilities for political change (Trotsky 1969, Löwy 1981). More specifically, what were the chances, he asked, of a transition to socialism directly from feudal relations of production. Accordingly, theories of uneven and combined development played a crucial role in Trotsky's rejection of Soviet aspirations for 'socialism in one country'—an oxymoron, he believed—but with Stalin's successful seizure of power and suppression of all opposition in the 1920s, any elaboration of the theory, now branded as counterrevolutionary, was blunted.

The question of uneven development re-emerged powerfully after the 1960s, this time in the West. Its recrudescence came with the need to develop critical analyses of the USSR and, in the context of the cold war, its relationship to global capitalism. Much more, however, it reflected a need to understand the intense geographical differentiations of global capitalism itself in the postwar world as Western social and political economies slouched toward crisis in the late 1960s and early 1970s (Mandel 1975) and as 'development' became the central means of relating to so-called Third World countries (see *Development Theory in Geography*). Decolonization, the Vietnam War, the precipitous decline of a previously hegemonic British economy, deindustrialization and the emergence of so-called newly industrializing countries, urban disinvestment and suburban expansion, and challenges by student, feminist, and civil rights movements, all put questions of uneven development on the political and economic agenda. In retrospect, this period from 1968 to 1991 initiated a significant transition from what some have called a Fordist to a post-Fordist regime of capital accumulation and regulation, but it simultaneously marked a crucial breakdown in established postwar geographies. Erstwhile differences between a First, Second, and Third World no longer made sense, deindustrialized regions retooled, and previously declining Western cities gentrified. Geographies that had previously been assumed as relatively fixed were now revealed as quite fluid and malleable, the products of

social, political, and economic change, and this cried out for explanation.

If the marxist political tradition was often indifferent or even hostile to geographical questions, treating them as displacements from the central issues of class struggle and the contradictions of capitalism, Western academic geography in the first seven decades of the twentieth century was equally immune not just to marxism but to social theory in general. It took a rapprochement between these separate political and academic traditions to make manifest the theories of uneven development that the transformations after the 1960s called out for (see *Space and Social Theory in Geography*).

2. *Sources of Uneven Development*

Uneven development is at root an expression of the contradictory tendencies toward differentiation and equalization of levels and conditions of social production (Smith 1991). On the one hand the stimulus toward differentiation stems directly from the creation and continual reproduction of the social division of labor. The logic of continued divisions of labor, whether at the scale of individual workplaces or transcontinental 'worlds' (First World, Third World), is inherently geographical. Thus Samir Amin (1976) distinguishes between a developed world which enjoys a balanced, self-centered mode of capital accumulation based on the preponderance of the world's industrial production and mass consumption, and an 'underdeveloped' world which specializes in the supply of raw materials and cheap labor and therefore experiences 'dependent development.'

Marx was equally explicit about the historico-geographical importance of the division of labor, albeit at a different scale: 'The foundation of every division of labour that is well developed, and brought about by the exchange of commodities, is the separation between town and country. It may be said, he continued rather enigmatically, 'that the whole economic history of society is summed up in the movement of this antithesis' (1987, p. 333). The development of the division of labor and attendant questions of social reproduction, consumption, and the differential power of nation-states and other political entities stimulates continuous social differentiation which is simultaneously spatial. The division of labor is the product of the competition between capitals which lies at the definitional heart of the capitalist mode of production, and it perpetually divides places as much as enterprises and people on the basis of their ability to differentiate their systems of production from those of neighbors and competitors.

But this powerful catalyst of social and spatial differentiation, enkindled by competition, does not go uncontested. There are powerful forces leading in the opposite direction, toward the equalization of conditions and levels of production, and these too emanate from the competitive impulse. 'Capital is by nature a leveller,' Marx once commented, exacting 'in every sphere of production equality in the conditions of the exploitation of labour' (1987, p. 375). Whatever its origins, this equalizing impulse is most intensely conveyed via the financial system where capital is most fluid, most mobile, and competition over decimal point differences in interest rates is most intense. It is precisely because the pressure of equalization is so intense in the financial sector that ideologies of 'globalization' and the 'end of geography' sprang from here in the later years of the twentieth century (Omae 1990, 1995, O'Brien 1992). Much writing about globalization (see *Globalization: Geographical Aspects*) at the end of the twentieth century captured very well the effects of this equalizing tendency in capitalist development while ignoring the equally profound counter-tendency toward differentiation.

These contradictory economic tendencies toward social differentiation and equalization find resolution in geographical space: a social problem is resolved, however temporarily, by fixing certain differences geographically in the form of discrete nation states, counties, regions, cities, and neighborhoods in such a way that the fluidity of commerce can continue with the least impediment. Uneven geographical development therefore represents a certain spatial fix (Harvey 1981) for the contradictions inherent in the capitalist mode of production. Far from being an ephemeral spatial result of capitalist growth and change, uneven development is integral to its constitution. Underdevelopment, therefore, is neither accidental nor simply a question of slower development along a pre-existing path. Rather, underdevelopment is an integral result of capitalist development.

At the same time, however much it represents a momentary fixing of social relations into specific spatial forms, uneven development is itself highly dynamic. The spatial fix is only ever temporary, and specific geographies that facilitate capital expansion and reproduction in one place in one period can become barriers to future development. To take an obvious example, the urban structure suitable for decentralized commercial and small-scale industrial activity in the seventeenth century is not only obsolete but was also an economic and political hindrance to the larger-scale industrial production of the nineteenth century. This might lead to wholesale urban 'renewal' as for example in Haussmann's Paris, or it may lead to a continued respecialization of activity leaving the old spaces largely intact, as for example in Venice where tourism is now the major industry.

A similar dynamism characterizes other spatial scales. Industrial regions owing their growth to nineteenth and twentieth century industrial expansion, such as New England or central Scotland, experienced economic decline in the second half of the twentieth

century as capital moved in search of cheaper and more pliant labor. But the very fact of disinvestment in those regions, the increase in unemployment, the cheapening of labor, and the breaking of previously powerful unions, also eventually made them attractive again for capital investment. There is a spatial correlate, therefore, to the classic characterization of capitalism as a system that lurches through cycles of expansion and crisis. Indeed, there is a powerful tendency for the cycles of investment and disinvestment inherent in capital accumulation to translate into a 'seesaw' geography of uneven development (Smith 1991). Capital concentrates in specific places but the fact of that investment continually erodes the attractiveness of these same places (higher labor costs, urban congestion, technological obsolescence, etc.) leading to disinvestment. Yet disinvestment itself creates the conditions and opportunities for new rounds of investment once sufficient devaluation has taken place. Over the long term, therefore, there is a powerful tendency toward a geographical seesaw.

Uneven development is every bit as fluid and dynamic as the logics of capital accumulation of which it is the spatial expression. A central theoretical issue here concerns the ways in which places come to be defined as discrete places vis-à-vis other places. Henri Lefebvre (1974) brilliantly taught us to see the world as a certain 'production of space' but there is little sense in his work of the specific social mechanisms and processes by which places and spaces become differentiated one from the other. This raises the question of geographical scale (see *Scale in Geography*) for not only is space 'produced' as an integral dimension of social change but the scaling of social activities is equally produced. Geographical scale, in fact, can be conceived as a metric—a vital sociospatial infrastructure—that organizes geographical differences into a coherence of different places. There is nothing in any way natural about the separation of different spatial scales: the scale of the global, the nation-state, regions and cities, the home and the body, all represent the ever-temporary spatial resolution of competing social forces.

This is easily illustrated with reference to the origin and functions of nation-states. Although nations and states separately preceded the flowering of nation-states in the eighteenth century, a good case can be made that the scale of nation-states emerged in response to contradictory impulses toward competition and cooperation that surfaced powerfully with the fruition of capitalism (Smith 1991). On the one hand, separate private capitals were driven by the necessity of competition, yet on the other it was imperative for them to find ways of cooperating in order to provide vital social and physical conditions for profit-making; the control of labor, a transportation infrastructure, the defense of private property, the provision of legal and financial institutions, systems of ideological mandate, all required a state apparatus—a structure of governance—to establish and regulate a modicum of self-interested cooperation among competing enterprises.

The territorial expression of this requirement was far from inevitable. From West Africa to South Asia to the Mediterranean, city states had previously performed this function, but the rapid expansion of capital accumulation in Europe by the seventeenth and eighteenth centuries now took place on a geographic scale that was less and less controllable from city states (Smith 1991, Arrighi 1994). The nation-state represented a very specific geographical compromise between capitalist competition and cooperation. Separate capitals cooperated within the nation-state over the provision of the vital collective conditions of production and reproduction while intensifying their competition in the market. Nor, of course, was this new territorial scale of governance uncontested or easily evinced (Tilly 1990). Rather it involved a dramatic mobilization of political as well as economic power, ideological persuasion establishing and expanding national identities, and military conflicts. The result was a system of national states and national economies which, while never entirely stable, represented a rescaling of the world such that national states for the first time became the basic building blocks of the global system.

3. *Recent History of Uneven Development*

The unevenness of the geographical landscape can therefore be read as a kind of palimpsest in which old social inscriptions are still quite evident in the landscape, albeit written in quite different registers—a kind of invisible socio-economic and political ink perceptible to the discerning geographical eye amidst later inscriptions. The decades spanning from postwar expansion (1950s and 1960s) to economic crisis (late 1960s to early 1980s) and the expansion of a restructured capitalism (1980s and 1990s) have witnessed a significant evolution in the patterns and processes of uneven development and although the tendency toward a seesaw motion of capital is far from complete, there are clear signs of it especially at lower scales.

At the urban scale, for example, the sustained disinvestment in many central cities throughout the Western world represented the geographical corollary of suburban expansion. Uneven urban development for most of the twentieth century was precisely about this differentiation of suburb from city, an expression of class as well as race and gender differences inscribed dramatically in the landscape. However, by the later decades of the century two new processes led a further evolution of urban unevenness. In the first place, suburbs became much more integrated functionally, adding retail, recreational, cultural, and service ac-

tivities to existing residential and industrial land uses, giving rise to what one commentator has called 'edge cities' (Garreau 1991). On the other hand, the widespread gentrification of central and inner city areas was spearheaded by reinvestment in previously disinvested neighborhoods, the seesaw of uneven development at the urban scale (Smith 1996). This reworking of the differences between suburb and city is driven by the continued geographical expansion of metropolitan areas and is mediated by the structure of ground rent within what is, effectively, a single (territorial) labor market.

At the regional scale we have already indicated a quite different if commensurate set of processes. Many major industrial regions that had been the economic backbone of national economies in Europe and North America—from the Midwest to the Rhine-Ruhr to the English Midlands—experienced disinvestment and deindustrialization after the 1950s as declining profit rates in traditional industrial sectors were matched by the expansion of new sectors. New regional agglomerations such as Silicon Valley, the Third Italy, or the high tech concentration around Taipei, emerged as the leading edge of regional change while by the 1980s some of the older deindustrialized regions (New England, central Scotland) experienced considerable reinvestment at the behest of the 'new economies.'

At the national scale, however, something rather different has been taking place. By the end of the postwar expansion in the early 1970s, the scale of capital accumulation had increasingly superceded the system of national states and economies that had, for two centuries, nurtured it. The pursuant 'globalization,' as it is misleadingly labeled, has eroded the rationale for discrete national economies and many systems of national state governance have, as a result, restructured themselves, significantly withdrawing from the support of social reproduction (the dismantling and/or privatization of social welfare and social services) while hewing more closely to the market logics of capital accumulation. If this 'hollowing out' of the state is only partial, it none the less has had the geographical effect of eroding the importance of national boundaries in economic terms. In general, both capital and laborers move more freely at the beginning of the twenty-first century than they did three decades earlier. There is a crucial caveat here, however. The increasing economic porosity of national boundaries does not in any way imply the end of the nation-state, and in fact the cultural and political responses to 'globalization' can equally lead to quite dramatic, even violent reassertions of national identity and difference (e.g., Serbia, Rwanda, various ex-Soviet states, or, in a softer vein, Scotland).

While previously hegemonic national powers—especially, perhaps, the UK—have clearly experienced comparative economic decline there is no real evidence of national economies undergoing a full 'seesaw' from developed to underdeveloped and back to developed economic status. The power of national boundaries over the last two centuries has entrenched highly resilient forms of inequality via colonialism, imperialism, and underdevelopment. Amidst the fluidity of capital and commodities in the global economy, a certain geographical fixity has prevailed at the national scale, accentuating patterns of unequal development.

This brings us to the global scale, where indeed we see that some regions and economies that were long underdeveloped have suddenly become major centers of global economic production. In other words, a 'new international division of labor' emerged in the wake of postwar reconstruction (Fröbel et al. 1980). Since the 1960s, the economies of South, South-east, and East Asia have experienced the most intense industrial revolution of all time, culminating in the momentous economic crisis between 1997 and 1999. Economies like South Korea, Singapore, or Taiwan, which in the 1960s were definitively consigned to Third World status, are now significant powers in the global economy. Together with the financial integration of European, North American, and Japanese economies after 1971, it was the inclusion of the industrial economies of Asia that spawned what is now known as globalization. Globalization is at root a recalibration of political economic power between transnational capital and national states. For apologists of globalization, it ushers in an equalization of conditions across the globe, while for critics it accentuates the differences. The theory of uneven development suggests that both visions are partially correct and that the contradiction between equalization and differentiation is resolved geographically. While much of Asia, even including China, which only nominally remains a socialist state, is drawn toward the capitalist core, Sub-Saharan Africa is increasingly distanced from any benefits of a new globalism. To borrow a language from the urban scale, Sub-Saharan Africa has, since the 1970s, been increasingly 'redlined' by global capital: a red line has been drawn around the region and international banks, governments, and aid organizations have minimized investments there, accentuating social poverty (see for example Bond 1998).

We have stressed that the geographical unevenness of capitalist development emanates from the structured social relations of capitalism itself, but this too is a historical result. Not until far flung economies were significantly interlinked through labor, finance, and commodity markets could uneven development evolve into the systematic expression of the inequalities at the heart of capitalism. While this was inevitably a gradual process, the years from the 1880s to 1919 marked a period of most intense transition. Radically divergent levels and types of economic development prior to this period may still have owed as much to divergent political economies, colonialism and imperialism notwithstanding. For the remainder of the twentieth century, however, such differentials in development were largely internal to a global capitalism that was

increasingly organized around US market dominance rather than around systems of European colonialism (Smith 1991).

Viewed this way, globalization after the 1980s represents not so much a wholly new phase of capitalist expansion but an accentuation of existing processes and patterns of uneven development. In this light it is important to stress that specific patterns of uneven development, while driven by powerful processes endemic to capital accumulation, also express the results of political struggles, social and cultural differences, and relict features and processes that have adapted to channel capitalist social and economic processes. Just as capitalism is a historically discrete mode of production, therefore, there is nothing inevitable about uneven development either in its broad outlines or specific details. Different modes of production make different kinds of social geographies.

See also: Dependency Theory; Development, Economics of; Development: Social-anthropological Aspects; Development: Socioeconomic Aspects; Development Theory in Geography; Globalization: Geographical Aspects; Marxist Geography; Space and Social Theory in Geography; Urban Geography; World Systems Theory

Bibliography

Amin S 1976 *Unequal Development*. Monthly Review Press, New York

Arrighi G 1994 *The Long Twentieth Century*. Verso, London

Bond P 1998 *Uneven Zimbabwe*. Africa World Press, Trenton, NJ

Fröbel F, Kreye O, Heinrichs J 1980 *The New International Division of Labour*. Cambridge University Press, Cambridge, UK

Garreau J 1991 *Edge City: Life on the New Frontier*. Doubleday, New York

Harvey D 1981 The Spatial Fix—Hegel, von Thünen and Marx. *Antipode* **13**(3): 1–12

Lefebvre H 1974 *La Production de l'Espace*. Anthropos, Paris

Lenin V I 1967 *The Development of Capitalism in Russia*. Progress Publishers, Moscow

Lenin V I 1975 *Imperialism, the Highest Stage of Capitalism*. Foreign Languages Press, Beijing, China

Löwy M 1981 *The Politics of Combined and Uneven Development*. New Left Books, London

Mandel E 1975 *Late Capitalism*. New Left Books, London

Marx K 1987 *Capital*. International Publishers, New York, Vol. 1

O'Brien R 1992 *Global Financial Integration: The End of Geography*. Council on Foreign Relations, New York

Omae K 1990 *The Borderless World*. Harper Business, New York

Omae K 1995 *The End of the Nation-state*. Free Press, New York

Smith N 1991 *Uneven Development: Nature, Capital and the Production of Space*. Blackwell, Cambridge, MA

Smith N 1996 *The New Urban Frontier: Gentrification and the Revanchist City*. Routledge, London

Tilly C (ed.) 1990 *Coercion, Capital and European States: AD 990–1990*. Blackwell, Cambridge, MA

Trotsky L 1969 *Permanent Revolution and Results and Prospects*. Merit, New York

N. Smith

Unfolding and Vector Models

1. Individual Differences in Preference

In this article a family of multidimensional models for preferential choice (or other 'dominance') data is considered, focussing on accounting for individual differences in relative preferences (or dominance judgments) of different subjects (or other data sources) for a number of stimuli (or other objects). The family of models involved comprise what is called the 'linear-quadratic hierarchy' of multidimensional preference models (Carroll 1972, 1980), beginning with the simplest case of the fundamentally linear 'vector model' first proposed by Ledyard Tucker, in which stimuli are represented as points in a multidimensional space, while subjects are represented by different vectors, or directed line segments, in the same multidimensional space—with the preference scale values (and thus the preference order) for a given subject being determined by projections of stimulus points onto that subject's vector. Next, in terms of complexity, is the 'simple' unfolding (or ideal point) model, originally proposed by Clyde Coombs. This inherently quadratic model assumes that each subject (or data source) is represented by a unique point, called the subject's 'ideal point,' representing the most preferred stimulus for that subject. That subject's preference scale values and orders (i.e., interval or ordinal scale values on what Coombs called the 'I scale' for that subject) for the stimuli are defined via a decreasing monotonic function of the (Euclidean) distances between the stimulus points and that subject's ideal point. Stated most simply, the closer a stimulus is to a subject's ideal point, the better that subject *likes* that stimulus. Other more complex models, generalizing the unfolding/ideal point model to include an idiosyncratic set of dimension weights for each subject, with distance being defined via weighted, rather than simple (unweighted) Euclidean distances, and another allowing each subject an idiosyncratic rotation of coordinate axes followed by differential weighting of the *rotated* axes, are also considered. The 'simple,' weighted, and generalized unfolding/ideal point models are all inherently quadratic models, entailing increasingly complex quadratic functions of the dimensions of the multidimensional space underlying the stimuli. These models, together with the vector model, comprise a hierarchy, since each simpler model is a special case of each of the more complex models—and is referred to as the linear-quadratic hierarchy of preferential choice models because they

all are either linear or quadratic, in the strongest (metric) case, or are *monotonically* related to linear or quadratic models in the weaker 'nonmetric' case.

Methods are discussed for fitting these models to appropriate individual differences preference data, including use of multidimensional scaling (MDS) methodology for 'internal' analyses—in which both stimulus *and* subject parameters are simultaneously estimated based on the preference data alone—as well as for 'external' analyses—in which stimulus parameters (e.g., coordinates of stimuli in a multidimensional perceptual space) are predefined, fixed *a priori* values (e.g., derived from MDS of proximity data), while subjects are 'mapped' into this *a priori* stimulus space by fitting subject parameters (vector termini, ideal points, and/or other parameters) to account optimally for the preference data *conditional* on the fixed dimensional coordinates of the stimuli.

2. Types of Preference Analysis

First we consider some important different types of analysis, beginning with internal vs. external analyses.

An *internal analysis* of preference data utilizes only the preference data to develop parameterization of both stimuli and subjects. An *external analysis* utilizes a stimulus space defined à priori; that is, *externally* to the preference data, for example, by multidimensional scaling (MDS) of independently defined proximity data, or from a physical, theoretical, or other 'rational' dimensionalization of the stimuli.

A mixed (ambiguous) case is that in which the stimulus space for an external analysis is defined via MDS or other structural analysis of the preference data.

Another important distinction is that between metric and nonmetric analyses. In a *metric analysis* the functions F_i (defined below) generally are assumed linear (or, at least of some well-defined parametric form) while in a nonmetric analysis they are assumed to be merely *monotonic*. It is important to keep in mind that this distinction between metric and nonmetric analysis has nothing to do with the metric (or distance function) defining distances between stimuli and ideal points, but with whether the basic data are ratio or interval scale (the 'metric' case) or merely ordinal scale data (the 'nonmetric' case).

In some cases the functions F_i may be assumed all to equal a single function F (so the subscript is dropped). This is a very strong assumption, since it assumes comparability among different individuals' preference data (thus assuming *unconditional* data). As will be seen below, this can lead to degenerate solutions in the nonmetric case. In most cases the data will be assumed to be noncomparable for different subjects, and thus to be *conditional* data.

3. Models for Multidimensional Preference Analysis

We now consider the '*linear-quadratic hierarchy*' of multidimensional preference models—ranging from the (linear) vector model through increasingly complex quadratic models corresponding to the 'unfolding' or ideal point model and its generalizations.

We begin with the vector model illustrated in Fig. 1, with an example due to Wish (see Carroll 1980 for details) shown in Fig. 2, from an intensive study of perception of and preferences for nations. This is a result of application of an *internal metric* analysis in terms of the vector model (via the MDPREF procedure, to be described below) to data on judgments of 'similarity to ideal' of 21 nations by 75 subjects, who were students selected from a wide range of nationalities at Columbia University's International House, in summer 1968 (at the height of the Vietnam war, and when, of course, the political and economic status of many of them was quite different than is now the case). The two dimensions can be interpreted as economic development (the first, or horizontal, dimension) and communist vs. noncommunist (or 'political alignment and ideology'), dimensions which were also found in independent MDS analyses of similarity data collected in this same or related studies. In the present case the subjects were given the labels 'dove,' 'hawk,' and 'moderate' based on responses to a one item questionnaire regarding US involvement in Vietnam; 'doves' being those preferring less US involvement in Vietnam, while 'hawks' wanted more involvement,

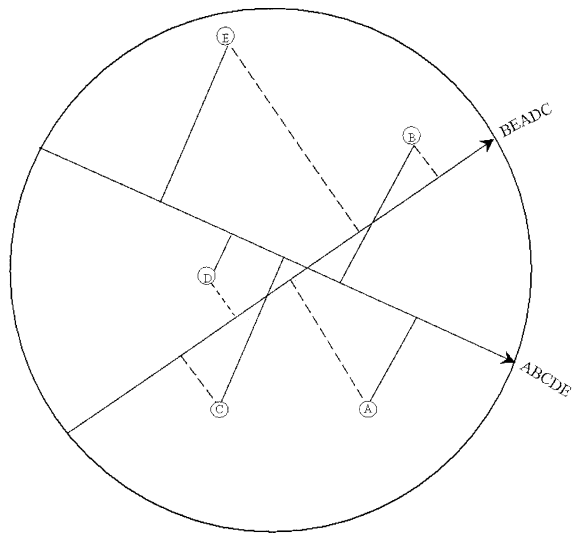

Figure 1
Vector model for preference illustrated. Projections of stimulus points onto a subject's vector are assumed to define preference scale for that subject

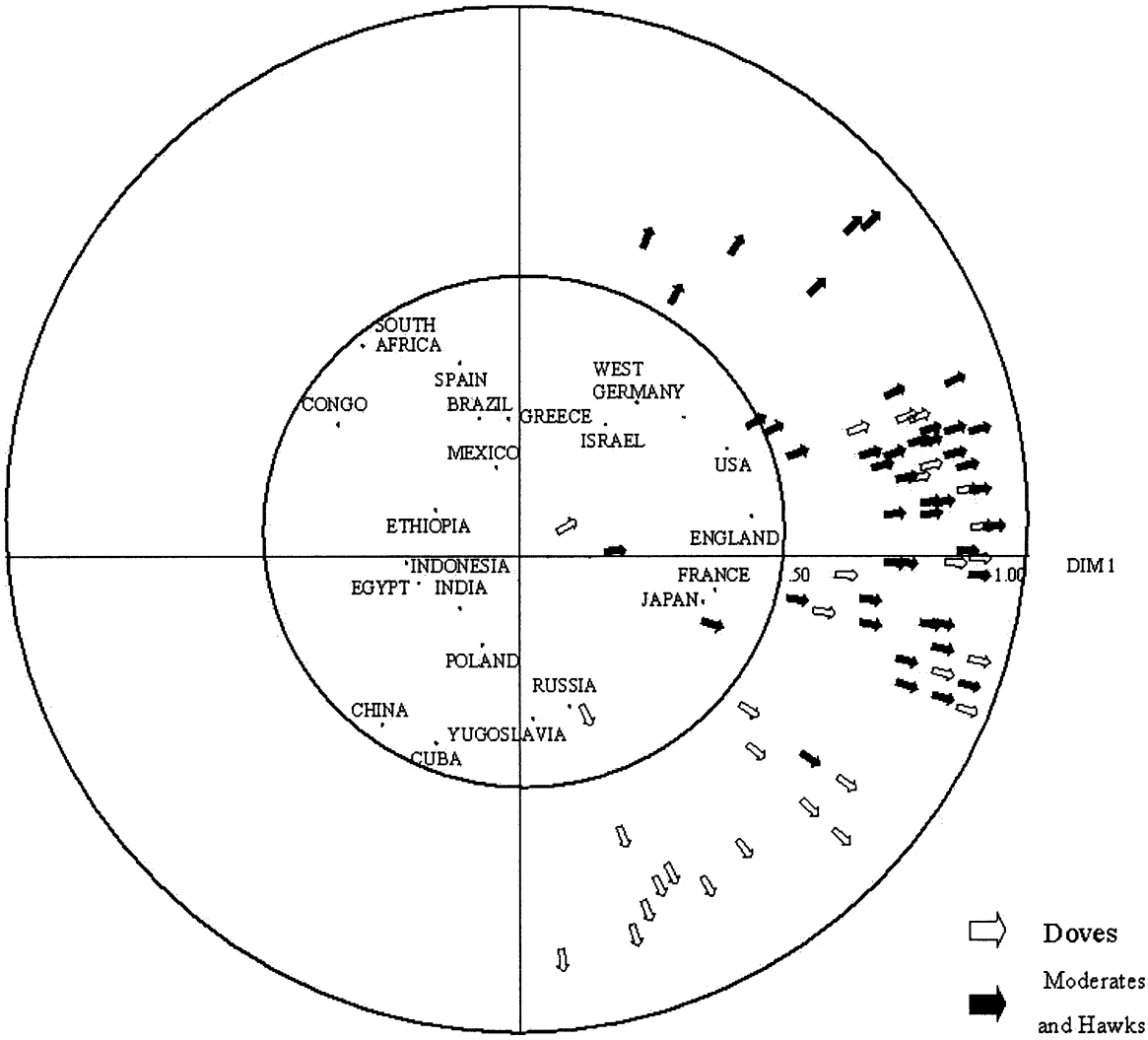

Figure 2
Two-dimensional solution from Wish's MDPREF analysis of 'Similarity to Ideal' judgments of 75 subjects on 21 nations

and 'moderates' preferred maintenance of the status quo. Since, however, many of these subjects were students from (then) communist countries, many of those classified as 'doves' on this basis might have been communist 'hawks,' that is, strong supporters of their own countries' positions opposing US involvement in Vietnam. For this reason we can see that many (but not all) of the 'doves' have vectors whose directions indicate preference for the communist countries, unlike what we might expect if these were 'doves' on the Vietnam issue from the US or other generally anti-Communist countries. On the other hand, all the 'moderates' and 'hawks' have vectors indicating either preference for the highly economically developed countries (as measured by judgments of 'similarity to ideal'), with relative indifference to political orientation, or preference for the noncommunist countries.

The vector model is mathematically of the form:

$$\pi_{ij} = F_i(p_{ij}) \cong \sum_{r=1}^{R} b_{ir}\, x_{jr} \qquad (1)$$

where R = number of dimensions, while $\cong$ means 'equals, except for (unspecified) error terms.' x_{jr} = value of the jth stimulus on r^{th} dimension ($j = 1 \ldots n$), b_{ir} = *importance* of dimension r for subject i (it is,

technically, a direction number, or is proportional to the direction cosine of the angle subject i's vector makes with the rth dimension or coordinate axis). p_{ij} = preference scale value of the jth stimulus for the ith subject ($i = 1 \ldots m$), π_{ij} = transformed preference scale value for subject i and stimulus j, F_i is a (monotonic, linear, or other) function transforming p_{ij} into π_{ij}.

In matrix form

$$\Pi \cong \mathbf{BX}' \tag{2}$$

where $\Pi \equiv (\pi_{ij})$ is the matrix of transformed preference values, π_{ij}, while $\mathbf{B}$ is the $m \times R$ matrix (b_{ir}) of importances, and $\mathbf{X}$ is the $n \times r$ matrix (x_{jr}) of coordinates of the stimulus space. Geometrically, we can think of this as *projecting* stimulus points onto a subject vector; the greater the projection the greater is that subject's (predicted) preference for the stimulus. These projections are multiplied by the *length* of the subject vector, which can be regarded simply as a scaling factor proportional to the square root of the variance accounted for in the transformed preference data for that subject.

In an *internal* analysis, both $\mathbf{B}$ and $\mathbf{X}$ are derived from data. A *metric internal* vector model analysis can be accomplished by a singular value decomposition (SVD) analysis—which amounts to a kind of 'direct' principal components analysis (PCA) in which both the components (corresponding to the vectors representing the subjects) and the component scores (corresponding to the *stimulus* dimensions) are derived directly from the Π matrix (which, in the metric case, is derived from the $\mathbf{P}$ matrix by simply normalizing rows/subjects—usually to zero mean and unit variance). This is the basis of the MDPREF method (see Carroll 1972, 1980)—perhaps the most widely used method of metric internal multidimensional analysis of preference data.

A *nonmetric internal* analysis can be accomplished by Kruskal and Shepard's *nonmetric* factor analysis of the $\mathbf{P}$ matrix (see Carroll 1972, 1980), which provides a nonmetric equivalent of MDPREF. Monte Carlo studies described by Kruskal and Shepard show that the parameters estimated using this approach are virtually indistinguishable from those fitted via the metric MDPREF approach—except in the case of the most extreme (but unrealistic) nonlinear, but order preserving, transformations of the data—so that they recommend using the metric procedure instead. For this reason, nonmetric factor/components analysis is never used in practice.

In an *external* analysis $\mathbf{X}$ is given and $\mathbf{B}$ is to be estimated. A *metric external* analysis can be accomplished by straightforward ordinary least squares (OLS) regression techniques—specifically multivariate multiple linear regression. The stimulus dimensions play the role of independent variables, while the preference scale values play that of dependent variables, while the regression coefficients predicting a subject's data estimate that subject's vector. This method is available in Carroll and Chang's PREFMAP (Carroll 1972) or PREFMAP-2 (Carroll 1980) procedures, or in PREFMAP-3 (Meulman et al. 1986).

A *nonmetric external* analysis would be accomplished by *nonmetric* multiple linear regression. Here one seeks a monotonic function of the dependent variable (the preference scale values) and the regression coefficients that *jointly* optimize an overall OLS criterion of fit. This method is available in all the PREFMAP procedures, as described in detail in Carroll (1972, 1980), or in Meulman et al. (1986).

An intuitively unattractive property of the vector model is that it assumes that preference changes monotonically with all dimensions. That is to say, it assumes that if a certain amount of a given attribute is good, even more is better (or, at least as good), which is unrealistic for most variables or attributes in the real world.

4. *The Unfolding or Ideal Point Model*

Coombs introduced the *one-dimensional unfolding* (or ideal point) *model*, which is illustrated in Fig. 3. Different subjects correspond to different '*ideal points*' representing their respective optimal values on the stimulus continuum. The farther a stimulus is from a subject's ideal point, the less the subject will like that stimulus. In Fig. 3, three hypothetical subjects are represented by ideal points, and the preference orders generated by these different ideal points are indicated. It is clear that, even in one dimension, quite different orders can be generated by this model. Coombs called this the 'unfolding' model because the preference order for a given individual can be generated by *folding* the stimulus scale at that individual's ideal point. To recover the stimulus scale from the preference data, it is necessary *simultaneously to unfold all these preference scales for individuals* (what Coombs calls the I scales) to find the common, or joint, stimulus scale (the J scale). The three curves drawn above the abscissa in Fig. 3 show *hypothetical preference functions* that go with the three individuals. Of major importance is that the curves are, to use Coombs's terminology, *symmetric* and *single peaked functions* (the latter meaning having a single maximum, corresponding to the ideal point). The particular shape is not important, so long as these two conditions are met, nor do they all have to have the same shape. Bennett and Hays generalized this model to the multidimensional case. The two-dimensional case of this *multidimensional unfolding* (*or ideal point*) *model* is illustrated in Fig. 4. As in the unidimensional case, the farther a given stimulus point is from an individual's ideal point in the multidimensional space the less that individual likes that

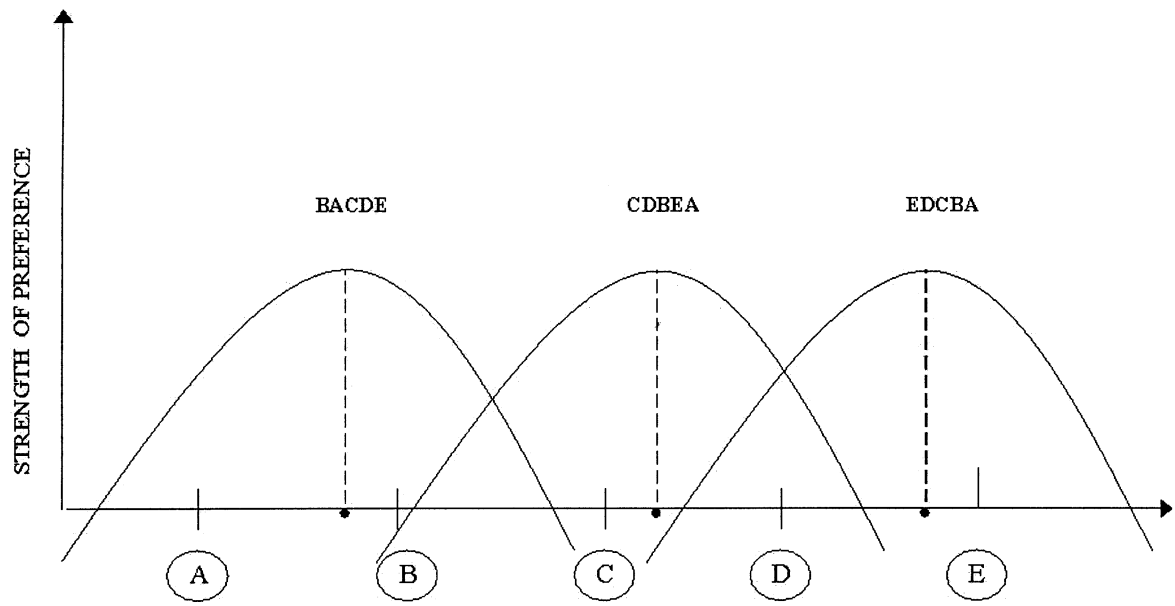

Figure 3
Unidimensional unfolding model illustrated. Distance of stimulus from subject's 'ideal point' (their optimal value on the unidimensional scale) is assumed to define (inversely) the preference scale for the subject

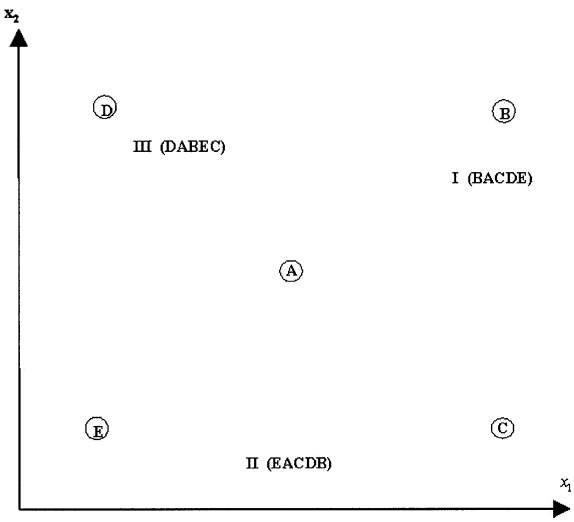

Figure 4
Multidimensional unfolding model illustrated. Distance from subject's 'ideal point' (inversely) defines that subject's preference scale. Isopreference contours are concentric circles centered at 'ideal points.' They would be spheres or hyperspheres in higher dimensions

stimulus. This, however, implies a metric on the space, which Bennett and Hays assumed to be Euclidean. This means the *isopreference contours* (or loci of equal preference) are, in two dimensions, a family of concentric circles centered at the individual's ideal point. They are concentric spheres in three dimensions and hyperspheres in higher dimensions. The basic algebraic structure of the (simple or unweighted) unfolding/ideal point model is defined in Eqns. (3) and (4) below

$$\pi_{ij} = F_i(p_{ij}) \cong d_{ij}^2 \tag{3}$$

$$d_{ij}^2 = \sum_{r-1}^{R} (y_{ir} - x_{jr})^2 \tag{4}$$

where $y_{ir} = r^{\text{th}}$ coordinate of the ideal point for the ith subject (and other parameters are as before).

While vector and unfolding models appear superficially to be quite different, the vector model can be shown to be a special case of the unfolding model. As the ideal point moves farther out in a fixed direction the preference orders are asymptotically identical to those for a vector in that direction. An algebraic and geometric proof is possible, but is beyond the scope of this article (see Carroll 1972, 1980). Thus, it follows that data appropriate for an R-dimensional vector model can always be fitted (arbitrarily closely) by an R-dimensional unfolding/ideal point model—with the preference orders for the ideal point model approaching those for the vector model as the ideal point moves toward infinity in the direction corresponding to that of the vector. This is also true for a *strong metric* version of the unfolding model, as described by Carroll (1972, 1980). Conversely, it can be shown that data

appropriate for either a *nonmetric* or *metric* unfolding/ideal point model can always be fit via a vector model with one extra dimension, as originally conjectured by Coombs and Kao and confirmed for the strong metric version of multidimensional unfolding by Ross and Cliff (see Carroll 1972, 1980). Thus, while an R-dimensional vector model is a special case of an R-dimensional ideal point model, an R-dimensional ideal point model is a special case of an $(R+1)$-dimensional vector model.

Any nonmetric or metric internal unfolding problem can be formulated as a nonmetric or metric multidimensional scaling problem of an $m \times n$ (e.g., subjects $\times$ stimuli) preference matrix (see *Multidimensional Scaling in Psychology*; Carroll and Arabie 1998). As first argued by Coombs the preference matrix **P** is interpretable as a *conditional off-diagonal proximity* matrix, that is, Coombs interpreted preference data as *proximities* between *stimuli* and subjects' *ideal points*, but *off-diagonal* since proximities among stimuli or among subjects are not provided, and *conditional* since data (e.g., preference orders or scale values) for one subject (row of the matrix) cannot be compared with those for a different subject/row.

Kruskal's MDSCAL-5, as well as various versions of KYST have options allowing a preference matrix (or another off-diagonal proximity matrix) to be treated as the above special form of proximity matrix and analyzed via either nonmetric or metric MDS. When doing this, for reasons discussed in Kruskal and Carroll (1969) and by Carroll (1972, 1980), it is essential to use the 'split-by-rows' option, that enables the preference data to be treated as conditional proximities, and an alternative measure of STRESS, called STRESS-2 (as contrasted to the earlier version of STRESS originally defined by Kruskal, now called STRESS-1). The details, and reasons for this, are discussed in Kruskal and Carroll (1969), or Carroll (1972, 1980). Suffice it to say that there are various totally meaningless *degenerate solutions* that are virtually guaranteed to occur, and which *appear* to yield perfect fit (STRESS = 0), yet recover essentially *no* information in the data, and are essentially totally *independent* of the data, if the guidelines stated in these papers are *not* followed. *Unfolding–External analysis* (*metric and nonmetric*).

Carroll and Chang's PREFMAP (for PREFerence MAPping) or PREFMAP-2, or, most appropriately, Meulman et al's. PREFMAP-3, accomplish an external unfolding analysis. Metric external unfolding analysis can be translated into a quadratic regression problem, while the nonmetric case can be treated as a nonmetric generalization of this special case of quadratic regression (Carroll 1972, 1980, Meulman et al. 1986). It is also possible to fit two generalizations of the unfolding/ideal point model via these programs—a weighted generalization in which a weighted Euclidean distance (analogous to that assumed in the INDSCAL approach to individual differences MDS) between stimuli and ideal points replaces the unweighted Euclidean distance assumed in the 'simple' unfolding/ideal point model, and a generalized model assuming a generalized Euclidean distance in which each subject is assumed to effect an idiosyncratic rotation of the coordinate system as well as having idiosyncratic weights for these (rotated) dimensions. At the other extreme, the simplest model in this class of models, called the *linear-quadratic hierarchy* of models; namely the *vector model*, can also be fit, by solving a metric or nonmetric least squares linear regression problem.

All four of these external models can be fit metrically or nonmetrically, while statistical tests based on the fact that these are, formally, nested models, can be used, at least in the metric case, to decide, for each subject, which model is most appropriate. In the nonmetric case, while not strictly valid statistically, the resulting 'pseudo-F's' can be useful in conjunction with other information (e.g., interpretability of the various representations) in deciding on the best model for each subject's preference data.

Other approaches to (internal or external) unfolding include a metric unfolding method proposed by Schöneman, and Srinivasan and Shocker's LINMAP, which fits an external weighted ideal point model, using linear programming techniques (minimizing a 'least absolute deviations' measure of fit) but with nonnegativity constraints on subject weights (see Carroll 1980). Other approaches include one based on 'homogeneity analysis' (closely related to correspondence analysis) proposed by Heiser, and DeSarbo and Carroll's three-way generalization of unfolding analysis (see Carroll and Arabie 1998). A general approach called GENFOLD is discussed by DeSarbo and Rao (1984).

In addition to spatial unfolding/ideal point models, tree-structure based unfolding models have been proposed by De Soete, DeSarbo, Furnas, and Carroll and generalized to the three-way case by De Soete and Carroll (see Carroll and Arabie 1998). Also, stochastic (or probabilistic) vector and (spatial) ideal point models as well as stochastic tree unfolding models have been fit to paired comparisons preference data (see Carroll and De Soete 1991, De Soete and Carroll 1992).

See also: Multidimensional Scaling in Psychology; Preference Models with Latent Variables; Psychometrics; Test Theory: Applied Probabilistic Measurement Structures

Bibliography

Carroll J D 1972 Individual differences and multidimensional scaling. In: Shepard R N, Romney A K, Nerlove S B (eds.) *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, Vol. 1. Seminar Press: New York.

(Reprinted in *Key Texts on Multidimensional Scaling*, Davies P, Coxon A P M (eds.) 1984. Heinemann, Portsmouth, NH)
Carroll J D 1980 Models and methods for multidimensional analysis of preferential choice (or other dominance) data. In: Lantermann E D, Feger H (eds.) *Similarity and Choice*. Hans Huber, Bern
Carroll J D, Arabie P 1998 Multidimensional scaling. In: Birnbaum M H (ed.) *Handbook of Perception and Cognition. Volume 3: Measurement, Judgment and Decision Making*. Academic Press, San Diego, CA
Carroll J D, De Soete G 1991 Toward a new paradigm for the study of multiattribute choice behavior: Spatial and discrete modeling of pairwise preference. *American Psychologist* **46**: 342–51
De Soete G, Carroll J D 1992 Probabilistic multidimensional models of pairwise choice data. In: Ashby F G (ed.) *Multidimensional Models of Perception and Cognition*. Erlbaum, Hillsdale, NJ
DeSarbo W S, Rao V R 1984 GENFOLD2: A set of models and algorithms for the GENeral unFOLDing analysis of preference/dominance data. *Journal of Classification* **1**: 147–86
Kruskal J B, Carroll J D 1969 Geometrical models and badness-of-fit functions. In: Krishnaiah P R (ed.) *Multivariate Analysis II*. Academic Press, New York
Meulman J, Heiser W J, Carroll J D 1986 *PREFMAP-3 User's Guide*. Unpublished manuscript. Bell Laboratories, Murray Hill, NJ

J. D. Carroll

Unitary State

1. Definition

The nation state is the conventionally accepted form of territorial organization which originated in Western Europe. The term 'unitary state' is a black hole in the political science literature; a taken-for-granted notion for which it is rare to find a definition (an honorable exception is Rose 1982, pp. 51–2). It is all too often treated as a residual category, used to compare unitary with federal states to highlight the characteristics of the latter (Elazar 1997). It is used in this article to refer to the politically sovereign, centralized governments of unified nation states. It is an institutional and constitutional notion; politically sovereignty refers to international recognition that a government rules a specified territory. It is centralized because the decision to decentralize can be revoked by the central authority.

Following Rokkan and Urwin (1982, p. 11), I distinguish between the unitary state and the union state. The unitary state is 'built up around one unambiguous political centre which enjoys economic dominance and pursues a more or less undeviating policy of administrative standardisation. All areas of the state are treated alike, and all institutions are directly under the control of the centre.' In the union state 'integration is less than perfect … (and) … While administrative standardisation prevails over most of the territory … in some areas pre-union rights and institutional infrastructure … preserve some degree of regional autonomy and serve as agencies of elite recruitment.'

2. Types of Unitary State

There are at least three types of unitary state among liberal-democracies: the Anglo-Saxon, the Napoleonic, and the Scandinavian.

The Anglo-Saxon state draws a clearer boundary between state and civil society, there is no legal basis to the state, and 'the idea of the state as a thing in itself, an institution independent of and superior to members of society … is alien to British political thinking' (Rose 1982, p. 47). The Crown represents the sum of formal political authority. The classical example of a unitary state is the British Westminster model characterized by parliamentary sovereignty, strong cabinet government, accountability through elections, majority party control of the executive (i.e., prime minister, cabinet, and the civil service), elaborate conventions for the conduct of parliamentary business, institutionalized opposition, and the rules of debate (Gamble 1990, p. 407, Rhodes 1997, Chap. 1). Combining parliamentary sovereignty and a strong executive makes the British unitary state one of the most centralized in Western Europe.

Britain can also be seen as a union state. Scotland, Wales, and Northern Ireland have distinctive administrative arrangements because the English center chose an operating code which was flexible and accommodating, stressing indirect control of the periphery and representation of the periphery in the center (Bulpitt 1983). Constitutional reform in the 1990s reinforced functional decentralization by devolving political authority to the constituent territories of the United Kingdom.

The Jacobin–Napoleonic tradition sees the French state as 'the one and indivisible republic,' exercising strong central authority to contain the antagonistic relations between state and civil society (on French political traditions, see Hazareesingh 1994, Chaps. 2, 6) 'France is a unitary state superimposed on a multinational society.' Thus the Jacobin tradition stresses the need for a monolithic state apparatus to coerce the fragments of French society into a semblance of order. National uniformity lies at the heart of the link between state and citizen; there is 'a comprehensive system of central administrative control based on a hierarchical chain of command.' Regulations multiply. In debates about regional reform, fears are ever present that they will be 'a real threat to the unity of the state and nation.' The continuing challenge of government is to find unity out of chaos. So the state moulds society into a whole (all quotes from Hayward 1983, Chap. 2).

Knudsen (1991, p. 96) identifies homogeneity, continuity, a strong state tradition, a high integration between state and society through strong local governments, a tradition for consensual democracy, social democratic parties, and a public sector with universal bureaucracies, little corruption, and comparatively high efficiency as key characteristics of the Scandinavian model of the unitary state. Denmark owes its strong bureaucratic and legal tradition to the influence of its neighbor, Germany. The defining characteristic of a *Rechtsstaat* is that it is a legal state vested with exceptional authority but constrained by its own laws (Dyson 1980). Civil servants are not just public employees, but personifications of state authority. So Knudsen (1991, p. 37) describes the dominant tradition as 'a highly bureaucratic, rule-bound and hierarchic structure designed to promote objective decision making on the basis of appropriate laws or regulations in obedience to the current government, whatever its political complexion.' Denmark differs from the German *Rechtsstaat* in being decentralized with a strong participation ethic. So local authorities have the right 'to manage their own affairs under the control of the state.' They have a high degree of constitutional autonomy; described as 'spectacular' by Knudsen (1991, p. 74). The complexities of constitutional reform are an intimidating safeguard of devolution to local government. Denmark is a decentralized unitary state.

3. *Centralization*

No unitary state is wholly homogeneous. No unitary state is completely centralized. Decentralization is a common strategy for coping with diversity. There are types and degrees of decentralization: deconcentration (prefectoral and functional), delegation, and devolution to both regional and local governments (Rhodes 1992).

3.1 *Decentralization*

This refers to the distribution of power to lower levels in a territorial hierarchy whether the hierarchy is one of governments within a state or offices within a large-scale organization (Smith 1985, p. 1). Or, more briefly, it refers to the areal division of powers (Maass 1959). So defined, the terms encompasses both political and bureaucratic decentralization, federal and unitary states, and decentralization between levels of government and within units of government.

3.2 *Deconcentration*

Sometimes referred to as field administration, deconcentration involves 'the redistribution of administrative responsibilities … within the central government' (Rondinelli and Cheema 1983, p. 18). A broad distinction can be drawn between prefectoral and functional systems. In the prefectoral system, a representative of the center—or the prefect—located in the regions supervises both local governments and other field officers of the center. This person is the superior officer in the field, embodying 'the authority of all ministers as well as the government generally and is the main channel of communication between technical field officials and the capital' (Smith 1967, p. 45). The classical examples are the French prefect and the Collectors or District Commissioners of India. In the functional system, field officers belong to distinct functional hierarchies. The administration of the several policy areas is separate. There is no general, regional coordinator; coordination occurs at the center. This system of multifarious functional territories is typified by England (see Hogwood and Keating 1982).

3.3 *Delegation*

This term refers to 'the delegation of decision-making and management authority for specific functions to organisations that are not under the direct control of central government ministries' (Rondinelli and Cheema 1983, p. 20). Such organizations are referred to as parastatal organizations, nondepartmental public bodies, or quangos (quasi-autonomous non-governmental organizations). They include public corporations and regional development agencies. This category is also used to cover the transfer of functions to the private sector or voluntary bodies through marketization, privatization, or contracting-out, cumbersome neologisms which refer to the various ways of delivering 'public' services using markets or quasi-markets. Decentralization understood as managerial delegation and marketization fueled major reforms of the public sector throughout the world in 1980s and 1990s (Kickert 1997).

3.4 *Devolution*

Devolution refers to the exercise of political authority, by lay, mainly elected, institutions within areas defined by community characteristics (Smith 1985, p. 11). Thus, 'local units are autonomous, independent and clearly perceived as separate levels of government over which central authorities exercise little or no *direct* control' (Rondinelli and Cheema 1983, p. 22). Historically, the *locus classicus* of devolution is said to be British local government, but the most significant trend in the decentralization of political authority is the growth of regional government in Europe (Jones and Keating 1995).

4. *For and Against Centralization*

Centralization is the root of all evil for advocates of decentralization. The arguments for and against are clear even if there is no foreseeable end to the debate.

So centralization promotes territorial justice and equality. Central authorities uphold service standards, rationalize resource allocation, and coordinate local development. There is a need for national plans, especially when resources are scarce, and only the center can ensure territorial equality by the central provision of funds and supervising the uniform implementation of national policies. Centralization is encouraged by financial weakness, national elites, including the bureaucracy, eager to protect their interests, and political instability.

On the other hand, decentralization is often said to be the counterweight to central power. Liberal-democratic theory assumes that decentralization promotes democratic participation, especially local self-government. Nationally, decentralization is said to promote political education, training in political leadership, and political stability. In local government, it promotes the values of equality, accountability, and responsiveness (Smith 1985, p. 20, Sharpe 1970). It is also said to have many managerial or administrative advantages. First, it is seen as a way or surmounting the administrative incompetence of the center; that is, the limits of national planning, by getting closer to problems, cutting through red tape, and meeting local needs. Second, it improved central 'penetration' of rural areas, spreading knowledge of, and mobilizing support for, the plan and bypassing obstructive local elites. Third, it encouraged the involvement of various religious, ethnic, and tribal groups, promoting national unity. Fourth, it increased the speed and flexibility of decision-making, encouraging experimentation, and reducing central control and direction. Fifth, it increased the efficiency of the center by freeing top management from routine tasks and reducing the diseconomies of scale caused by congestion at the center. Sixth, it increased the administrative capacity of the localities and regions, and improved the coordination of service delivery. Finally, it institutionalized participation, provided opportunities for many interests to get a 'stake' in the system, trained citizens for democracy and politicians for government, and promoted political maturity and democratic stability (paraphrased from Rondinelli and Cheema 1983, pp. 14–16 and Smith 1985, pp. 186–8). Unfortunately, decentralization has not lived up to its promises. The strongest argument for centralization is the failure of decentralization, especially in developing countries, where the main outcome is chaotic inefficiency.

5. *Conclusion*

To raise the topic of centralization is to raise an emotive issue. Centralization is 'bad'; decentralization is 'good.' There is a clear implication that a federal state is a more decentralized form of government than a unitary state (see *Federalism*). Any such conclusion should be resisted. A federal state can devolve restricted powers to its constituent governments. The theory and practice of federalism can diverge markedly and, as the case of Denmark shows, a unitary state can devolve considerable powers to local and regional governments. The unitary state cannot remain a taken-for-granted notion. The degree of centralization, and its varied political and administrative consequences, must become matters of empirical inquiry.

See also: Delegation of Power: Agency Theory; Nations and Nation-states in History; Rechtsstaat (Rule of Law: German Perspective); Sovereignty: Political; State and Society; State: Anthropological Aspects; State Formation; State, History of; State, Sociology of the

Bibliography

Bulpitt J G 1983 *Territory and Power in the United Kingdom*. Manchester University Press, Manchester, UK

Dyson K 1980 *The State Tradition in Western Europe*. Martin Robertson, Oxford, UK

Elazar D J 1997 Contrasting unitary and federal systems. *International Political Science Review* **18**(3): 237–51

Elder N, Thomas A H, Arter D 1982 *The Consensual Democracies? The Government and Politics of the Scandinavian States*. Martin Robertson, Oxford, UK

Gamble A 1990 Theories of British politics. *Political Studies* **38**: 404–20

Hayward J E S 1983 *Governing France. The One and Indivisible Republic*. Weidenfeld and Nicolson, London

Hazareesingh S 1994 *Political Traditions in Modern France*. Oxford University Press, Oxford, UK

Hogwood B W, Keating M (eds.) 1982 *Regional Government in England*. Clarendon Press, Oxford, UK

Jones B, Keating M (eds.) 1995 *The European Union and the Regions*. Clarendon Press, Oxford, UK

Kickert W J M 1997 (ed.) *Public Management and Administrative Reform in Europe*. Edward Elgar, Aldershot, UK

Knudsen T 1991 State building in Scandinavia: Denmark in a Nordic context. In: Knudsen T (ed.) *Welfare Administration in Denmark*. Ministry of Finance, Copenhagen, Denmark

Lijphart A 1984 *Democracies*. Yale University Press, New Haven, CT

Maass A (ed.) 1959 *Area and Power*. The Free Press, Glencoe, IL

Rhodes R A W 1992 Intergovernmental relations in unitary political systems. In: Hawkesworth M, Kogan M (eds.) *The Routledge Encyclopaedia of Government and Politics*. Routledge and Kegan Paul, London, pp. 316–35

Rhodes R A W 1997 *Understanding Governance: Policy Networks, Governance, Reflexivity and Accountability*. Open University Press, Buckingham, UK

Rokkan S, Urwin D S 1982 *The Politics of Territorial Identity*. Sage, London

Rondinelli P A, Cheema G S 1983 Implementing decentralisation policies: An introduction. In: Rondinelli P A, Cheema C S (eds.) *Decentralisation and Development*. Sage, London

Rose R 1982 *Understanding the United Kingdom*. Longman, London

Sharpe L J 1970 Theories and values of local government. *Political Studies* **18**: 153–74

Smith B C 1967 *Field Administration*. Routledge and Kegan Paul, London
Smith B C 1985 *Decentralisation: The Territorial Dimension of the State*. Allen and Unwin, London

R. A. W. Rhodes

United Nations: Political Aspects

The United Nations is an international organization of which almost every state in the world is a member. An international convention, the UN Charter, in force since 1945, constitutes its basic law. Its paramount purposes comprise the maintenance of international peace and security and the promotion of human welfare through international cooperation. As the UN agenda as well as its organizational design involves definition and resolution of issues of public policy, its establishment implies limited governance. Its decisional processes exhibit relative power among participants as well as attempts to change it. Shifting coalitions of members aggregate support or opposition for various ends, including allocation of resources. Some conclusions are converted into binding law. The United Nations maintains relationships with a variety of constituencies and other organizations.

1. *Conventional and Novel Capacities*

The organizational design, largely planned by American officials during World War II with some British and other contributions, owes much to the League of Nations, especially the concept of *collective security* that underlies enforcement against violation of the UN Charter. In promoting human welfare, the activist, reforming, government-led approach of the New Deal in the United States and the doctrine of *functionalism* offered important models.

The organization nevertheless rests on foundations of conventional diplomatic experience with the concept of *state sovereignty* as a backdrop. This heritage includes bargaining among diplomats to propose binding international law stated in agreements that governments ratify. The UN structure supports routine diplomatic, conciliatory, arbitrary, and judicial methods for the settlement of disputes. Such techniques set most boundaries to UN governing capacities.

The United Nations was endowed with some novel governing attributes. In interstate conflicts that breach the prohibition in the UN Charter on using force in international relations, the Security Council, as principal organ for international security, was empowered to coerce violators of the peace. This coercion included the use of a military force as well as diplomatic and economic sanctions. That coercive measures would both deter and arrest aggression was underpinned by the requirement that majority voting in the Security Council must include concurrence of the permanent members, the Great Powers of 1945. Moreover, arms control would bring the Security Council closer to a military monopoly.

Less directly coercive, but savoring of governmental functions, were broad duties in setting international standards. These included the novel, policy-pregnant topic of human rights. Related were supervisory functions pertaining to self-government or, in limited cases under the surveillance of the Trusteeship Council, independence of colonial territories. Less novel were recommendations related to international economic and social issues, put in the purview of the Economic and Social Council (ECOSOC) which was also to oversee other international organizations with technical agendas. The General Assembly, where all members had equal votes, would adopt overall policies.

Governments retained the main responsibility for executing policies adopted in the UN organs. Some central executive capacity, however, was given to the Secretary General, the head of an international secretariat. He carries out instructions of the intergovernmental organs, prepares documentation for meetings, and serves as the public representative. He thus acquires a role in decision-making. The Secretary General has the overtly political right to bring matters affecting peace and security to the attention of the Security Council.

2. *Political Processes and Outcomes*

The harmony of aims that planners assumed and acceptance of a formal process for deciding on policies, making recommendations, and reacting to crises never materialized. Instead, the organization has only partially functioned as a centralized policy source and provides an arena for competitive as well as cooperative policies among members.

2.1 *The Effects of the Cold War*

The lengthy antagonism between the United States and the Soviet Union and their respective allies made the United Nations an arena of confrontation and conflict. It soon crippled attempts to bring nuclear arsenals under international control. The planned armed force for the Security Council aborted amidst controversy. In a long line of disputes, e.g., the conflict between India and Pakistan over Kashmir, the Security Council operated more as sounding board than a settlement device. In the early years, the opposition of the Soviet Union to many proposed actions led it to much use of the veto. Led by Washington, the General Assembly tried to recompense the Security Council's lameness, but as the membership increased, the United States abandoned its position and emphasized the Security Council, where its opposition could

block action. Later the United States became a leading user of the veto.

Nevertheless, the Cold War neither canceled some use of the United Nations to maintain international security nor blocked organizational development. By reacting immediately to the attack on South Korea in 1950 by the Communist regime in the North, the United States gathered a coalition under UN auspices to check the invasion. But UN recommendations for the reunification of Korea remained unfulfilled during the next 45 years.

The UN organizational cluster was extended with the establishment of the World Health Organization, the Food and Agriculture Organization, the International Civil Aviation Organization, the UN Educational, Scientific and Cultural Organization, and the General Agreement on Tariffs and Trade. The Soviet Union and its allies opposed much of this construction and the programs that emerged.

2.2 Functionalism

This wave of organizational construction had its roots in functionalist doctrine (Mitrany 1966). It reflected the work of the League of Nations and the International Labor Organization to extend human welfare by international cooperation. Supporters of this organizational principle futilely hoped that technical agencies would have a nonpolitical character.

The use of ECOSOC as a policy leader and coordinator soon was frustrated in part by the Cold War antagonism and in greater part by resistance from the specialized agencies. These technical bodies had separate legal status, their own contact points with governments, and controlled their own budgets. Some of the organizations, especially the International Monetary Fund and World Bank, kept their relations with the council and other organizations at a distance. Yet the obvious need for coordination to prevent duplication and promote synergy led to a network of interorganizational relationships, alliances, and coalitions that became a standing feature of UN politics.

Charting interorganizational relationships was further complicated by the General Assembly's creation of agencies under its wing and by the notion of regionalism which was incorporated the in the UN Charter. To UN Children's Fund (UNICEF) and UN High Commissioner for Refugees (UNHCR), which by 1950 the General Assembly established as short-term agencies to which contributions were voluntary, were added the UN Population Fund; the large UN Development Program; the UN Conference on Trade and Development (UNCTAD); and more. It also created four regional economic and social commissions.

In matters of peace and security, both the Security Council and the General Assembly made sporadic and usually superficial use of independent regional organizations, such as the Organization of American States. Not until the disintegration of Yugoslavia and subsequent conflicts in Bosnia and Kosovo in the 1990s was a link forged between the United Nations and the North Atlantic Treaty Organization (NATO) and the European Union. NATO troops served for enforcement of policies given legitimacy by the Security Council.

Despite elaborate coordinating structures, reliance on *ad hoc* executing coalitions in both welfare and security activities hardly diminished. They reached new high points in the armed response to the Iraqi invasion of Kuwait and in several humanitarian disasters from the 1980s onwards.

2.3 Increasing Membership

Formation of new and shifting coalitions for both policy formation and execution connected directly to the near quadrupling of the original 51 UN members. First held in check by Cold War maneuvers, the membership greatly increased with the rapid collapse of the colonial system. Poor, weak, and small countries could coalesce in the General Assembly to restrict the powerful. Formal coalitions represented the views of the nonaligned states, which hoped to shun the Cold War; the developing countries organized as the UNCTAD group of 77; and several geographically-based groupings. While coalitions had no formal role in the Security Council, they conditioned its deliberations.

When security matters, including arms control, and such issues as human rights and development come before the General Assembly, the powerful governments necessarily seek support from a broad segment of the membership. This leads both to confrontation and broader cooperation. An always cumbersome decisional process, occasioning diplomatic bargaining, requires coalition formation. It offers a platform for nationalistic gestures and even for dissident ethnic minorities. It also impels some governments, especially the wealthy and powerful, to restrain UN involvement on certain security issues as well as on cooperation on development, trade, some arms control arrangements, and financial contributions.

2.4 Resource Allocation

Formally, the United Nations can only allocate resources put at its disposal by members who decide together on levies or offer voluntary contributions. Generally, financial constraints have pinched hard. They also reflect national policies such as when the Soviet Union and France declined to pay for some peacekeeping activities or the United States objected to the budgetary share assigned to it.

From early on, the General Assembly and ECOSOC sought to steer resources to governments seeking economic and social development. This effort included

various forms of technical assistance, capital grants which were very limited, and emergency relief. All of these found financial support in voluntary contributions, mainly from governments with highly developed economies. Together these efforts produced substantial programs that, however, failed to satisfy either needs or majority demands.

Broad approaches with a more compulsory tone emerged during the years of argument that led to creation of UNCTAD in 1964 as a global bargaining arena on terms of trade. This was followed by the General Assembly's attempt in the wake of the restriction of oil supplies during the Arab–Israeli war of 1973 to mandate a New International Economic Order. These stabs at resource reallocation fell well short of their high goals and were rejected by developed members. However, they affected various operations of the UN organizational cluster.

3. *Global Developments*

By its fifth decade, the United Nations had only a diminishing effect on international power relationships and warlike attempts to change them. Nevertheless, it assumed a central role in marking out and partially applying international cooperative agendas that reflected profound long-term global developments and in stimulating some popular opinion.

3.1 *Power Ranking*

The assumption of 1945 that the permanent membership of the Security Council—China, France, Soviet Union, United Kingdom, and United States—including the governments most able to muster effective force for UN missions would work harmoniously soon failed to match actuality. After 1948, China was represented by the exile government on Taiwan and stirred controversy and pressure from Beijing, which took over the place following a General Assembly action in 1971. France and the United Kingdom were by the 1980s challenged by the renaissance of Germany and Japan as well from large developing countries, including India, Nigeria, and Brazil. Yet finding a formula for reform that would be sustained by the existing permanent members, as is necessitated by the UN Charter, remained a paralyzing puzzle.

Even after the end of the Cold War and a newly cooperative spirit in Moscow, finding requisite majorities for enforcement and pacific settlement actions by the Security Council usually proved difficult. The weaker Russian successor to the Soviet Union and a China jealous of its national prerogatives implied use of the veto. The United States, obviously the most capable member, increasingly treated the Council with skepticism based on domestic developments and emphasized bilateral and alliance security arrangements. Elected members sometimes pursued national policies rather than cooperation.

The permanent members of the Security Council often ignored the General Assembly, dominated by lesser powers, and sometimes even its elected members. The Council accordingly was favored primarily for endorsing policies that had been negotiated independently of public debate and usually had little role in steering the resulting programs despite a UN stamp. This development was illustrated by the Council's activity in the 1990s in the Bosnia-Herzegovina and Kosovo as well as in its hesitant approach to the Rwanda genocide.

3.2 *Intrastate Conflicts*

Conflicts that originated in repressive governmental policies that ignored human rights standards, challenges to existing authorities that clung to power, and nationalistic movements that echoed the UN emphasis on independent nation-states increasingly came on the UN agenda, whether in the Security Council or the General Assembly. With rapid communications, the resulting human costs could not be hidden. Relief for victims of humanitarian disasters became a central and novel UN concern. Reconstruction of failed states as in Cambodia and Somalia included creation of democratic regimes. These aims expanded the scope of peacekeeping, even to include forceful intervention in such later cases as Kosovo and Bosnia-Herzegovina.

3.3 *New Global Issues*

Topical intergovernmental conferences, stimulated and prepared by the Secretariat and endorsed by ECOSOC, broadened the international agenda or sharpened attention to older items. Familiar agenda topics were population, arms control, food, and narcotic drugs control. Newly identified as global issues included the status of women; protection of the environment and climate change; refugees in Africa and from Cambodia; exploration of outer space; nuclear energy; development of the least developed countries and debt relief; human habitations; and an international criminal court.

While the form of topical intergovernmental conferences followed conventional diplomatic patterns and ended with communiqués that urged governments to act, those on human rights, women, and the environment attracted massive observation and some actual participation by nongovernmental organizations (NGOs). This added an unprecedented and controversial popular tinge to the otherwise formal conferences. The NGOs, some of which had consultative status with ECOSOC, claimed to voice the views of large groups. They effectively supported actions on human rights and the status of women

which many governments rejected as dilution of state authority. Some of the NGOs skillfully lobbied governments to apply the conclusions of the conferences and by using electronic communications acted as pressure groups undeterred by national jurisdictions. After the 1992 UN Conference on Environment and Development, a much larger number of NGOs were offered admission to ECOSOC consultancy. This suggested diminution of the governmental membership principle, but NGO participation remained controversial and fluid.

4. Scholarship and Research

Academic study of the United Nations and allied agencies has varied greatly in intensity and method (see *International Organization*). Realist and neorealist theorists of international relations generally ignore or give glancing attention to the UN system. Their concentration on national interest and foreign policies, moreover, fits with diplomatic rhetoric heard in UN fora. More specialized research that involves the formation of governmental policies has usually treated the UN system as one of many factors involved in international approaches to global topics, such as human rights, economic development, the environment, and chemical and biological weapons.

A cascade of theories seeking explanations for interstate cooperation have been employed as main tools or incidentally in studying the United Nations as an institution and as a locus of political processes. These include a variety of structural approaches, including Marxism and *dependencia*. Other intellectual paths that reflect nonstate as well as governmental actors comprise functionalism and derivatives; international regimes; organizational learning; neoinstitutionalism; and social networks. A few of these approaches have permitted statistical analysis and mathematical modeling.

On the whole, no single approach dominates research on the United Nations. The most detailed studies, which however aim at normative conclusions or the formal development of UN recommendations, generally come from legal or policy-oriented scholars. Publications of NGOs and interest groups focus on specific policy issues. Official documentation, e.g., topical reports, statistical collections, and records of deliberations, is voluminous and widely used in specific enquiries.

See also: Cold War, The; Globalization: Political Aspects; International Organization; Peace; Peacemaking in History; Regional Government; Regional Integration; Second World War, The; Sovereignty: Political

Bibliography

Archer C 1992/1999 *International Organizations*. Routledge, London and New York

Baehr P R, Gordenker L 1999 *The United Nations at the End of the 1990s*, 3rd edn. Macmillan, London

Claude I L Jr 1964 *Swords into Plowshares*, 3rd edn. rev. Random House, New York

Claude I L Jr 1984 *Power and International Relations*, 4th edn. Random House, New York

Cox R W 1987 *Production, Power, and World Order: Social Forces and the Making of History*. Columbia University Press, New York

Goodrich L M, Hambro E, Simons A P 1969 *Charter of the United Nations: Commentary & Documents*, 3rd edn. rev. Columbia University Press, New York

Jacobson H K 1984 *Networks of Interdependence*, 2nd edn. Knopf, New York

Mitrany D 1966 *A Working Peace System*. Quadrangle Books, Chicago

Roberts A, Kingsbury B (eds.) 1994 *United Nations, Divided World*, 2nd edn. Clarendon Press, Oxford, UK

Rosenau J N 1992 *The United Nations in a Turbulent World*. Lynne Rienner, Boulder, CO

Ruggie J G 1996 *Winning the Peace: America and World Order in the New Era*. Columbia University Press, New York

Russell R B 1958 *A History of the United Nations Charter*. Brookings Institution, Washington, DC

Simma B 1995 *The Charter of the United Nations: A Commentary*. Oxford University Press, New York

Weiss T G (ed.) 1993 *Collective Security in a Changed World*. Lynne Reinner, Boulder, CO

Weiss T G, Gordenker L (eds.) 1996 *NGOs, the UN & Global Governance*. Lynne Rienner, Boulder, CO

Weiss T G, Coate R A, Forsythe D P 1997 *The United Nations and Changing World Politics*, 2nd edn. Westview Press, Boulder, CO

L. Gordenker

Universities and Science and Technology: Europe

The role of science and technology at European universities has been undergoing tremendous changes (Etzkowitz and Leydesdorff 1997, Zinberg 1991). Given the increased importance of information technology and the emerging global knowledge-based economy, scientific research and development, as well as academic training, form the basis of national policies and institutional strategies to promote economic growth. As a consequence, universities are expected to respond to the need for well-trained scientists and engineers, new knowledge, and applied science (Zinberg 1991). The industry and technology-oriented academic fields like engineering and business

management find it easier to combine science (as basic research) with technology (as a product of applied research) in order to meet the new requirements of the economy and to raise additional funds. Other disciplines in the social sciences or humanities may have greater difficulty in maintaining a key position within universities and obtaining necessary resources (Slaughter and Leslie 1997).

In order to understand the current trends in academic science and technology in Europe, it is important to describe the institutional varieties of universities. Trends in different European countries serve as a basis for analyzing new challenges for higher education systems. Consequences for the production and management of knowledge show the new imperative for the future, which emphasizes a redefinition of the relationship between universities, industry, and the state in Europe (Etzkowitz and Leydesdorff 1997, Neave and Van Vught 1991).

1. *Academic Science and Technology in Europe*

European systems of higher education consist of a variety of universities and colleges (Gellert 1993, Teichler 1996). On a very general level, two types of universities prevail: one of a comprehensive character and the other specializing in one field or area. The comprehensive university represents a conglomerate of different faculties, academic programs, and consequently a larger number of students. Specialized universities are focused on a particular academic discipline or field, such as business administration or engineering, with only one faculty and fewer students. Specialized universities tend to focus on applied research and professional training, thereby responding more easily to external needs of the labor market and private industry. Comprehensive universities are more likely to be dedicated to basic research and training (Teichler 1996).

The modern European university which combines research and teaching emerged in the early nineteenth century and went through a lively period of transitions up to the present. The proliferation of fields (e.g., biotechnology, telecommunications, electronic commerce) and new institutional types (e.g., business schools, vocational colleges) have been based on the increasingly diverse research and education needs of national economies. In retrospect, this differentiation of functions appears to have been a consequence of the new requirements of the labor market and changes in the knowledge infrastructure (Etzkowitz and Leydesdorff 1997). At the same time, the role of higher education shifted emphasis from social mobility to wealth creation (Nowotny 1995). In other words, universities had to find new forms to translate science into technology, moving away from earlier concerns with liberal education and curiosity-driven research (Slaughter and Leslie 1997). Many European governments and ministries have played a pivotal role in this shifting of emphasis in higher education systems by supporting universities through research money and other financial incentives (Etzkowitz and Leydesdorff 1997).

These new trends in the social and economic context of universities have also changed the role of knowledge (Gibbons et al. 1994). Traditionally, institutions of higher education across the world have performed core missions and functions: knowledge production (research), knowledge dissemination (teaching), and knowledge application (service–outreach). For universities operating in a social context of application, research has to accumulate not only disciplinary-based knowledge but also useful results for society. The reason for the increasing application–orientation of higher education systems, especially in Europe, has been mass higher education leading to new forms of knowledge production focusing more on a link between science and technology. As a consequence, the university has moved from an ideal model (i.e., social change) to a pragmatic model (i.e., wealth creation) (Gibbons et al. 1994, Trow 1973).

In Europe, mass higher education not only entails a sharp increase in students, faculty, and institutions. Generally, it also implies increasing institutional diversity, organizational complexity, and academic heterogeneity. Also, public pressure has challenged the ideal of autonomous universities as 'ivory towers' and has called for more accountable institutions (Dill and Sporn 1995). In the following section these challenges are analyzed more closely.

2. *New Challenges for European Systems of Higher Education*

Colleges and universities in Europe have become part of mass higher education systems because of expansion and diversification. Expansion refers to the large increases in student numbers in most European countries caused by public policies to provide education to a larger portion of the population. Diversification means mostly a binary system of higher and vocational education translating into different types of institutions for different target groups (Sporn 1999). In this process, knowledge production and dissemination have become very important. 'The driving force behind accelerated supply and demand of marketable knowledge lies in the intensification of international competition in business and industry. In many cases in-house research is no longer sufficient to meet competitive demands. In order to commercialize knowledge, firms have to look for new types of links with universities, government laboratories as well as with other firms' (Gibbons et al. 1994, p. 46).

An additional factor aggravating the socioeconomic challenges for higher education has to do with the trend to 'scientification' of employment and work

(Teichler 1996). The service and production industry has become increasingly dependent on the results of research. Additionally, a growing number of positions are labeled as 'knowledge work' and call for high skills and solid academic training. Management positions, especially in large companies with a wide reach, require a university degree (Teichler 1996).

As many observers of higher education have outlined, major challenges within European universities have occasioned the following changes (Gibbons 1994, Sporn 1999): diversification of functions, different social profiles of student population, more professional education, diverse funding of higher education, shifting power between academic and administrative roles, focus on efficiency and effectiveness, and new forms of knowledge production. In this realm, science and technology have gained importance by being more tightly linked with each other and with external constituencies. Scientific research has been translated into application-oriented results, thereby raising public credibility and legitimacy. For this amalgamation of science and technology, higher education scholars employ to the widely used term 'technoscience,' implying a combination of 'basic and applied research, discovery and innovation, science and product, knowledge and commodity' (Slaughter and Leslie 1997, p. 38).

As outlined earlier, these changes in university science and technology have been based on new economic conditions in many European countries caused by budgetary constraints and new patterns of public spending. Additionally, restructured job markets call for constant relearning of the workforce. These different areas display a general cross-national dynamic and include five major trends: scarce financial resources, changing role of the state, new technologies, continuing education, and increasing globalization (Sporn 1999).

State budgets for higher education have been restructured in many European countries. Although public funds for higher education are stable or even increasing in some European countries, they cannot respond to the rising number of students. As a consequence, new sources of funding are needed. Public spending is very often based on output measures like the number of graduates or publications. As a consequence, universities have acquired a more diversified public and private funding base and have concentrated their strategies on enhancing efficiency and effectiveness of teaching and research (Clark 1998, Williams 1995).

Closely linked to the problem of limited public funds are new mechanisms of coordination and control in higher education systems, especially regarding the role of the state. Many European systems formerly characterized by 'state control' have moved to a state 'supervising' model, with increased institutional self-regulation and autonomy while the state steers the system from a distance. In many cases like Sweden or The Netherlands, leadership structures have been strengthened and quality control is handled through external audits and evaluation (Neave and Van Vught 1991). As national economies in Europe have become more volatile and dynamic, the life-cycle of knowledge has shortened and employees need to refresh their expertise more frequently. Hence, continuing education and lifelong learning have become key elements of national policies for enhancing economic productivity. Universities responded by expanding their programs, offering adult education, interdisciplinary programs, and part-time studies with a more vocational orientation. This has expanded the number of continuing, postsecondary, and professional education programs. In Europe, this trend has been labeled as the 'professionalization' of general education (Gellert 1993, Teichler 1996).

The new knowledge economy is mainly based on the widespread use of technology. Information technology has also changed university teaching, research, and management. The expansion of the Internet and affordable hardware and software have led to computer literacy of large parts of the student population. The impacts on how students learn, professors teach and conduct research, and administrators manage the institution are complex. Recent trends like virtual universities and distance education programs, international research projects, and administrative information systems are a few examples of changes influenced by information technology (Sporn 1999). Globalization and international competition have put pressure on colleges and universities. The importance of research and development, the creation of common markets, rising mobility of students and staff, and the free movement of capital have increased the need for universities to contribute to national wealth by making the economy more competitive (Etzkowitz and Leydesdorff 1997). Consequently, market-oriented university practices and national policies have been pushing for university-industry research and development programs which aim at stimulating regional economic development (Slaughter and Leslie 1997).

These trends in European higher education systems have caused a process of restructuring of academic science and technology. Developments include new technology transfer strategies, shifting relations between applied and basic research, new funding patterns, and rising need for management capacity within universities.

3. *Consequences for Academic Science and Technology*

3.1 *Knowledge Transfer*

Government and universities have been designing strategies which should create and sustain competitive

advantage in industry. The critical role of knowledge transfer either through research or through teaching has triggered new alliances between industry, government, and academic institutions. Trends include partnerships between universities and businesses, technology transfer units, incentives for a combination of basic and applied research, and for professors contributing to spin-off or start-up companies (Etzkowitz and Leydesdorff 1997).

Knowledge transfer is often enhanced through funding for research that supports economic development. At the same time, industry relationships help universities to raise additional funds and support companies in their research and development. In this process, academic science and technology are increasingly linked to form the basis for this relationship. For example in The Netherlands, the University of Twente moved from a third-class technology-oriented institution to a successful university through the creation of new entrepreneurial practices focusing on transfer, research, and development (known as TRD activities) (Clark 1998).

International organizations like the European Union (EU) have a high interest in knowledge transfer as well. EU projects like EUREKA reflect new state interests by establishing large supranational networks of scientists and practitioners to stimulate technological innovations. EUREKA promotes 'market-driven' collaborative research and development through the use of advanced technologies thereby strengthening European competitiveness.

3.2 Problem Orientation

As higher education systems all over Europe have expanded, research and teaching have shifted their emphasis as well. Given the complex requirements of the job market, of knowledge-based industries, and of society at large, research and teaching have to be oriented more to problems than to disciplines. This orientation helps universities to better meet the expectations and needs of external constituencies and offer relevant programs to their students. Increasingly, universities define fields of competence (e.g., environmental economics, social medicine) which combine different disciplines and areas to provide solutions to problems like pollution or AIDS (Gibbons et al. 1994, Dill and Sporn 1995).

As a consequence, new forms of knowledge production have become necessary. Gibbons et al. (1994) coined the term 'Mode 2' knowledge production, ascribing five basic attributes to it: application context, transdisciplinarity, heterogeneity and organizational diversity, social accountability and reflexivity, and quality control. 'Mode 1' refers to the traditional discipline-oriented research of tenured faculty members. 'Mode 2' knowledge production includes a wider, more temporary, and heterogeneous set of researchers and practitioners collaborating on a problem defined in a specific and localized context (Gibbons et al. 1994).

3.3 Market-based Funding

In Europe, scarce resources for higher education have given rise to new claims for public accountability and performance-based funding of universities. These developments have increased the importance of market mechanisms for financing higher education (Williams 1995). Departments or groups which work in a field that is highly relevant to industry and society have more opportunities for raising additional funds and for gains in the process of resource allocation. They are also better able to contribute to the university budget through their share of overheads. These additional funds can then be used to cross-subsidize other less 'marketable' areas or programs (Clark 1998). Mechanisms of market-based funding have also created a more entrepreneurial culture within many European universities. Faculty members responsible for their own research institute or center at universities can form a periphery which obtains additional income and also serves the public need for applied research, teaching, and service (Clark 1998, Sporn 1999). Again, fields which can more easily combine science and technology, like telecommunications, are especially inclined to get involved in these activities. Conflicts arise if less successful areas which belong to the tradition of university research and teaching get scrutinized for downsizing or closure (Slaughter and Leslie 1997).

3.4 Managerial Capacity

The new relationship between universities, government, and industry requires institutional management capacity. Given the need to translate external demands into internal responses and strategies, faculty members in Europe—who have been steering universities for decades—have become overburdened. A new type of professional university manager is emerging, who is responsible for institutional efficiency and effectiveness (Gumport and Sporn 1999). Reforms in many countries like Austria, The Netherlands, Sweden, or Spain were aimed at strengthening the leadership structure of universities.

European universities historically were dominated by strong state control and little decision-making power. But with ongoing reforms in many European countries, professional administrators and university managers have gained importance and governments have delegated many tasks to the institutional level. Consequently, in countries like Norway or Sweden

(Gornitzka et al. 1998) an increase in the number of university administrators can be observed while the size of the faculty stagnates. The rising interest in knowledge transfer based on science and technology is adding to the need for management support for new activities.

Tensions arise from these trends as to the new status of employees and the relationship between basic and applied research. While in many continental European countries university professors have been state employees, new trends call for more flexible and decentralized forms of employment. Also, at the institutional level, basic and applied research will have to be redefined. Although the basic mission of the university includes basic research a more market-oriented system requires a larger amount of applied and problem-oriented research projects.

To summarize, the link between science and technology has played a critical role within universities in Europe. National policies have converged around increasing national economic competitiveness through product and process innovations, channeling students and resources into well-funded curricula that meet the needs of a global marketplace, preparing more students for the workplace at lower costs, and managing institutions more effectively and efficiently (Slaughter and Leslie 1997).

The expansion of European higher education systems will continue to aim at high participation rates. European economies will need more qualified specialists as well as valid research results in applied fields in order to stay abreast of major technological and scientific developments. Governments will move from a control to a supervising role, thereby trying to enhance new partnerships between higher education and industry to boost the economy. Universities have to emphasize their core commitment to a service function in order to meet these high expectations and needs. Science and technology will continue to play a pivotal role in this development.

See also: Academy and Society in the United States: Cultural Concerns; Higher Education; Humboldt, Wilhelm von (1767–1835); Policy Knowledge: Universities; Research and Development in Organizations; Scientific Academies in Asia; Universities and Science and Technology: United States; Universities, in the History of the Social Sciences

Bibliography

Clark B R 1998 *Creating Entrepreneurial Universities: Organizational Pathways of Transformation*. Pergamon, Oxford, UK

Dill D D, Sporn B 1995 The implications of a postindustrial environment for the university: An introduction. In: Dill D D, Sporn B (eds.) *Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly*. Pergamon, Oxford, UK, pp. 1–19

Etzkowitz H, Leydesdorff L 1997 Introduction: universities in the global knowledge economy. In: Etzkowitz H, Leydesdorff L (eds.) *Universities and the Global Knowledge Economy*. Pinter, London, pp. 1–8

Gellert C 1993 Structures and functional differentiation—remarks on changing paradigms of tertiary education in Europe. In: Gellert C (ed.) *Higher Education in Europe*. Jessica Kingsley, London, pp. 234–46

Gibbons M, Limoges C, Nowotny H, Schwartzmann S, Scott P, Trow M 1994 *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. Sage, London

Gornitzka A, Kyvik S, Larson I M 1998 The bureaucratisation of universities. *Minerva* **36**: 21–47

Gumport P J, Sporn B 1999 Institutional adaptation: demands for management reform and university administration. In: Smart J (ed.) *Higher Education: Handbook of Theory and Research*. Agathon, New York, pp. 103–45

Neave G, Van Vught F A 1991 *Prometheus Bound: The Changing Relationship Between Government and Higher Education in Western Europe*. Pergamon, Oxford, UK

Nowotny H 1995 Mass higher education and social mobility: A tenuous link. In: Dill D D, Sporn B (eds.) *Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly*. Pergamon, Oxford, UK, pp. 72–89

Slaughter S, Leslie L 1997 *Academic Capitalism: Politics, Policies, and the Entrepreneurial University*. The Johns Hopkins University Press, Baltimore, MD

Sporn B 1999 *Adaptive University Structures: An Analysis of Adaptation to Socioeconomic Environments of US and European Universities*. Jessica Kingsley, London

Teichler U 1996 Higher education and new socio–economic challenges in Europe. In: Burgen A (ed.) *Goals and Purposes of Higher Education in the 21st Century*. Jessica Kingsley, London, pp. 96–111

Trow M 1973 Problems in the transition from elite to mass higher education. In: OECD (ed.) *Policies for Higher Education*. OECD, Paris, pp. 51–101

Williams G L 1995 The 'marketization' of higher education: reforms and potential reforms in higher education finance. In: Dill D D, Sporn B (eds.) *Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly*. Pergamon, Oxford, UK, pp. 170–93

Zinberg D S 1991 *The Changing University: How Increased Demand for Scientists and Technology is Transforming Academic Institutions Internationally*. Pinter, London

B. Sporn

Universities and Science and Technology: United States

The American research university's history is inseparable from the growing importance of science and technology to the state in the late nineteenth and twentieth centuries. From the passage of the 1862 Morrill Land Grant Act to the establishment of the National Science Foundation in 1950, government and university have established deep and long-lasting

connections, although the nature and character of those connections has changed profoundly over time. Once a ward of private contributions and philanthropic foundations, the American university, like American science and technology, became in some respects a ward of the state. By the beginning of the twenty-first century, the American research university was busy recasting its social relations and its position in the nation's institutional and intellectual ecology.

In many respects, this history repeats itself. At the beginning of the twentieth century a new kind of institution—the private research organization—emerged as a potential threat to the university as the primary site for the production of scientific and technical knowledge. Although such institutions survived, they were incapable of mounting a successful challenge to the university's intellectual hegemony. Research and teaching were powerful resources that other institutions could not match whatever their potential virtues to researchers. Today, universities and government-sponsored research are challenged by private corporations intent on capitalizing upon new discoveries, especially in the biomedical disciplines such as genomics. Ironically, the intellectual property issues that ruled out significant pre-World War II government patronage of academic science and technology have returned in a similar guise.

Equally important, the university's internal structure has altered profoundly over the past century, a change perhaps most visible in the shrinking role of university presidents. The presidents of the late nineteenth and early twentieth centuries were dynamic individuals willing to risk reputation and even wealth for their beliefs as they sought to shape and mold institutions. Charles Eliot at Harvard and Daniel Coit Gilman at Johns Hopkins invented the research university in the US, bringing German ideals to bear in a distinctly American fashion. Later, Harvard University president James Bryant Conant played a critical role in building the atomic bomb and developing postwar national security policy. Conant also led a vigorous public debate over the nature and character of the postwar American state. Presidents in the year 2000 have become largely fundraisers, anonymous even within their own institutions, let alone—with some exceptions—on the wider stages of national policy.

1. *The American College*

The antebellum American college was an institution devoted to the production of civic leaders, ministers, and schoolteachers. At heart, it was a character development institution complete with a psychological model of the mind legitimizing its pedagogical routines (Guralnick 1975, Veysey 1965). 'Faculty psychology,' the belief that distinct physical areas of the brains were the locations of different emotional and intellectual functions, was the early theory of cerebral localization undergirding the college's character-building program. Whether mandatory Latin or Greek memorization or mathematical reasoning, specific kinds of intellectual work were thought to strengthen the faculties to produce better, more capable citizens. The extent to which science and technology played a role in the antebellum college rested upon their ability to lay claim to this process of disciplining and shaping young minds. Social or economic utility was not yet an acceptable reason for the addition of science and technology to the college curriculum.

Passage of the 1862 Morrill Act granted each state federal lands for the establishment of a college dedicated to the mechanic and agricultural arts. These new institutions, including Cornell University, the University of Wisconsin, the University of Illinois, and others became sites for the development of new academic roles and new kinds of knowledge. Because graduate education in the sciences was largely unheard of in the US, the majority of instructors had studied abroad, mainly in Germany, and to a lesser extent in England and France. The land grant and its equally important sister institution, the agricultural experiment station, were hothouses for the production of a distinctive vision of American science. Utility had been the driving force in the creation of the new institutions, but many of the researchers hired at these new sites desired to transform these new domains into institutions resembling the laboratories of their German teachers, such as the famous chemist Justus Liebig. As Charles Rosenberg (1997, [1976]) argued, the agricultural experiment station yielded the first American hybrid, the scientist entrepreneur. Obtaining resources for research demanded that academics develop techniques and practices capable of satisfying multiple constituencies, ranging from the local farmer who wanted to know why crops would not grow in area soil, to the dean who wanted researchers to solve problems of interest to potential donors, and the politician who wanted results to justify re-election. Land grant colleges and agricultural experiment stations provided reasons for the support of scientific research so long as such research delivered useful results (Veysey 1965, Geiger 1986).

2. *Research*

The American research university was the product of two distinct intellectual trends—the increasing number of Americans travelling abroad for graduate education, and the problems posed by the rapidly industrializing US polity following the Civil War. First, American researchers who studied in Germany came to believe that the German university was the model of a research institution. This was an illusion fostered by the American's peculiar status in the

German university system—they were as much sophisticated tourists as students and few, if any, pursued the full course of study. American students in German laboratories worked with outstanding teachers, such as Carl Ludwig, but they failed to observe that the German university system trained more professionals—physicians, lawyers, and bureaucrats—than pure academic researchers. When American students returned to the US to take up academic posts they brought only part of the German university system back with them (Geiger 1986, Owens 1985).

The founding of the Johns Hopkins University in 1876 is often taken as the origin of the American research university. Although President Daniel Coit Gilman promoted the university as a center of '*wissenschaft*' in Baltimore, Hopkins represented less Göttingen on the Chesapeake than another form of hybridization: the meshing of research with the antebellum college's character-building function. Gilman benefited from the fact that Johns Hopkins had provided in his will for the foundation of a new university and medical school. However, the meaning of Gilman's innovations—laboratory- and seminar-based instruction—are often lost in the promotion of research. We should see Gilman and the staff he handpicked—including the physicist, Henry Rowland; the chemist, Ira Remsen; the mathematician, J. J. Sylvester; and the biologist, Henry Newell Martin—as men who brought new resources and practices to bear upon old problems. Gilman and other contemporaries such as Harvard University President Charles Eliot, believed that post Civil War America was a nation in crisis, suffering from effects of unprecedented technological change and the breakdown of local social orders in the building of a national culture. Education, that most universal remedy for social ills, was once again called upon to protect and reinvent social order. Research, manifested in the labors of the laboratory and the seminar room, was the perceived antidote for a hyperkinetic, market culture. It was in the laboratory, whether working on the cutting of diffraction gratings, as in Rowland's laboratory, or in the seminar room engaged in the philological study of classical texts, that the student would encounter a new form of discipline in which steady and methodical intellectual work would prepare the young mind for the task of living and working in a new social world. More importantly, such students would themselves work as teachers in the nation's colleges and secondary schools, transferring this new form of discipline and understanding to a wider populace (Kargon 1986, Hannaway 1976).

In the hands of educational reformers, the laboratory and the seminar room became new technologies for the (re)production of social order. They were also sites of intellectual innovation whose products might have value independent of their pedagogical origins. The *fin de siècle* saw the emergence of two new types of research institutions—the private, philanthropic research institution, such as the Carnegie Institution of Washington and the Rockefeller Institution for medical research (now the Rockefeller University), and the industrial research facility, such as those established at General Electric and AT&T. Both institutions sought to break the university's stranglehold on the production of knowledge and each was modestly successful. In the new organizations, research had a value for its own sake unknown in the university. Part of the problem was that many American academics were unable to understand why one would want to separate research and teaching. Andrew Carnegie charged his private research facility, the Carnegie Institution of Washington (CIW), with the mission of seeking out the 'exceptional man' and providing him with funds and resources to complete his work free of distractions. While the CIW consulted with academics on how to put Andrew Carnegie's generous gift to work, they were hard pressed to keep academics from applying for funds to relieve them of their teaching burden or pay for graduate research assistants. Neither was an acceptable use of CIW funds. Robert S. Woodward, the CIW's second president, established a variety of research divisions to insulate 'exceptional' men from both the university and other demands on their time (Reingold 1979).

Industrial research facilities were rather different and represented a wholesale transfer of the laboratory as a knowledge production site to the factory. Instead of students and theses, the new laboratories produced patents and processes necessary for corporate survival. Of the high-tech industries of the early twentieth century, electricity and communications led the way, integrating their corporate structures into the very manufacture of knowledge that might have industrial implications. Despite early press that these laboratories were universities in exile, later studies recognized that industrial research was always basic to corporate needs and goals. Successful laboratories integrated themselves into the firm's business; unsuccessful laboratories fought for independence that quickly evaporated when the firm suffered in the marketplace (Beer 1958, Reich 1985, Wise 1985, Noble 1977, Dennis 1987).

3. *The Visible Hand*

World War I brought about a fundamental reorganization of the university with respect to the production of technical knowledge. Although American academics made no great contributions to the development of weapons technology (save for the peripheral development of intelligence tests (Carson 1993), they organized themselves in new and powerful ways that allowed for the development of new forms of patronage. In particular, the National Research Council (NRC) established by Caltech president and astrophysicist George Ellery Hale, to resuscitate the

National Academy of Sciences, proved an invaluable mechanism for the support of university-based research. The early twentieth century had seen the emergence of incredibly wealthy private philanthropic foundations, including the gargantuan Rockefeller Foundation (RF), the Carnegie Corporation of New York, and the Guggenheim Foundation (Kohler 1991, Kay 1993). The RF easily dwarfed all its competitors and would forever alter the topography of American higher education. In the wake of World War I, the foundation supported a postdoctoral fellowship program administered by the NRC. As the historian Robert Kohler (1987) argued, the NRC was a 'trade association' for American science, providing philanthropies with a means of dispensing funds without having to support mundane university needs. That is, the foundations were not interested in supporting undergraduate education, or in building facilities; instead, the RF emphasized a policy of 'making the peaks higher.' Those who had excellent resources and professors became sites where NRC postdoctoral students clustered. In turn, those fellows left for jobs and attempted to build institutions capable of attracting such fellows.

We should note that the RF did not distribute its monies across the disciplines. The biomedical and social sciences were its primary targets of patronage; even during the Great Depression, the RF maintained its funding of these sciences, especially through efforts to import the techniques and instrumentation of the harder sciences into the biomedical and social sciences. It is to an RF officer, Warren Weaver, that we owe the term 'molecular biology;' and the RF's interest in genetics, human biology, and social engineering were to have lasting repercussions in another *fin de siècle*'s most public work, the Human Genome Project (Kohler 1976, Kay 1993).

Private support of the sciences was not matched by government support before World War II. Indeed, efforts to have the federal government pay for university-based research ran up against long-standing beliefs that government money invariably brought with it government control. Curiously, American academics did not seem to fear RF funds for analogous reasons. Various efforts to organize federal support of the sciences during the Great Depression foundered on this belief (Kargon & Hodes 1985, Genuth 1987). Equally important was a fundamental inability of university administrators and government bureaucrats to determine who would own the fruits of government-sponsored research (Owens 1990). Such disagreements made federal support of university-based research nearly impossible, save for research lacking any potential commercial value. Nor was industry a powerful patron of interwar academic science and technology. In large part, corporations sought such tight control over sponsored academic research that researchers almost welcomed the war and the possibility of relatively unconstrained government support (Galison et al. 1992). In postwar America, the state emerged for science as a patron of unprecedented wealth and power.

4. *Wards of the State*

World War II's effect upon the American university and American science and technology was profound. American researchers, organized by their leader, Vannevar Bush, of the Massachusetts Institute of Technology (MIT) and the Carnegie Institution of Washington, instituted a virtual revolution in warfare, bringing about a permanent revolution in munitions, whether it was the atomic bomb or attempts at the mathematization of strategic thinking through game theory and operations research. Scholars, especially David Noble (1984), Stuart W. Leslie (1993), and Rebecca Lowen (1997), have begun the task of examining how the war and postwar military patronage transformed specific universities. Others, such as Paul Forman (1987, 1992), have led an effort to understand how the very content of technical knowledge changed to meet the needs of new patrons. The rise of the military as the dominant patron of American university-based science had far-reaching effects upon the university, not the least of which was a strengthening of the faculty with respect to administrators. It was the faculty with whom military officers established relationships and it was the faculty that produced both the new technologies and the students that the military wanted. In turn, administrators attempted to reassert their authority through the development of new means of accounting, most notably the invention of indirect overhead as a means of capturing revenue from grants and contracts made to specific individuals (Dennis 1994).

Postwar military patronage focused upon four major research areas: guided missiles, continental defense, the managerial sciences, and intelligence-related research in areas as diverse as area studies and cryptography. These four domains capture the range of military interest and share a common tool, the digital computer. The development of the ballistic missile brought with it the development of miniature electronic components, including the integrated circuit and the ubiquitous microprocessor. Continental defense, the effort to develop a shield against a possible Soviet aerial attack, produced a range of technologies, including the modem and magnetic core memory. The US continental defense system, semi-automatic ground environment (SAGE), provided corporations like IBM with the technological experience to engage in massive commercial software developments, including the Semiautomatic Business Research Environment (SABRE), the airline reservation system. The managerial sciences encompass a multitude of fields, ranging from the quantitative social sciences, like operations research and systems analysis, to the

emergence of management science. In turn, interest in the managerial sciences, work which ranged from Simon and Ackoff's views on information processing and management to decision support systems, flourished in the reformed business school of the postwar era (Schlossman et al. 1994). That management might be taught as a science, in the most positivistic sense, was one reason for the growth and development of graduate business education, among the most notable stories of Cold War education. Finally, the effect of the US intelligence services—the CIA and the NSA—remains among the most secret effects of the Cold War on the American university. While it is clear that Area Studies and various departments of communications research were adjuncts to the Cold War effort, what remains unclear is the extent to which research on computers, artificial intelligence, and mathematics was driven by the ever increasing needs of national security. The digital computer and its associated academic departments—applied mathematics, linguistics, computer science, electrical and communications engineering, information theory, etc.—owe much to the nation's cryptographic needs for their development. Given that such work is still considered among the most sensitive, we may never know the depth of the state's penetration into the university.

What lies ahead for the American university and the technical research practiced there? The Cold War's end brought back fundamental issues related to commercial rights and production of knowledge, one reason among many for the enhanced visibility of intellectual property issues. At the same time, university administrators are eager to embrace the marketplace and its potential riches. In doing so, university presidents and faculty will continue to negotiate an increasingly fuzzy boundary separating students from consumers and research from commodities.

See also: Academy and Society in the United States: Cultural Concerns; Educational Systems: North America; Higher Education; Scientific Academies, History of; Scientific Disciplines, History of; Universities and Science and Technology: Europe; Universities, in the History of the Social Sciences

Bibliography

Beer J J 1958 Coal tar dye manufacture and the origins of the modern industrial research laboratory. *Isis* **49**: 123–131

Carson J S 1993 Army alpha, army brass, and the search for army intelligence. *Isis* **84**: 278–309

Dennis M A 1987 Accounting for research: New histories of corporate laboratories and the social history of American science. *Social Studies of Science* **17**: 479–518

Dennis M A 1994 Our first line of defense: Two university laboratories in the postwar American state. *Isis* **85**(3): 427–55

Forman P 1987 Behind quantum electronics: National security as basis for physical research in the United States, 1940–1960. *Historical Studies in the Physical Sciences* **18**: 149–229

Forman P 1992 Inventing the maser in postwar America. *Osiris* (2nd series) **7**: 105–34

Galison P 1997 *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, Chicago

Galison P B, Hevly B, Lowen R 1992 Controlling the monster: Stanford and the growth of physics research, 1935–1962. In: Galison P, Hevly B (eds.) *Big Science: The Growth of Large Scale Research*. Stanford University Press, Stanford pp. 46–77

Geiger R L 1986 *To Advance Knowledge: The Growth of American Research Universities*. Oxford University Press, New York

Genuth J 1987 Groping towards science policy in the 1930s. *Minerva* **25**: 238–68

Guralnick S M 1975 *Science and the Ante-bellum American College*. American Philosophical Society, Philadelphia

Hannaway O 1976 The German model of chemical education in America: Ira Remsen at Johns Hopkins (1876–1913). *Ambix* **23**: 145–64

Kargon R H 1986 Henry Rowland and the physics discipline in America. *Vistas in Astronomy* **29**: 119

Kargon R H, Hodes E 1985 Karl Compton, Isaiah Bowman, and the politics of science in the Great Depression. *Isis* **76**: 301–8

Kay L E 1993 *The Molecular Vision of Life: Caltech, the Rockefeller Foundation, and the Rise of the New Biology*. Oxford University Press, New York

Kohler R E 1976 The management of science: The experience of Warren Weaver and the Rockefeller Foundation Programme in Molecular Biology. *Minerva* **14**: 279–306

Kohler R E 1991 *Partners in Science: Foundations and Natural Scientists, 1900–1945*. University of Chicago Press, Chicago

Leslie S W 1993 *The Cold War and American Science: The Military–Industrial–Academic Complex at MIT and Stanford*. Columbia University Press, New York

Lowen R S 1997 *Creating the Cold War University: The Transformation of Stanford*. University of California Press, Berkeley, CA

Noble D 1977 *America by Design: Science, Technology, and the Rise of Corporate Capitalism*. Knopf, New York

Noble D 1984 *Forces of Production: A Social History of Industrial Automation*. Knopf, New York

Owens L 1985 Pure and sound government: Laboratories, playing fields, and gymnasia in the 19th-century search for order. *Isis* **76**: 182–94

Owens L 1990 MIT and the federal 'Angel:' Academic R&D and federal–private cooperation before WWII. *Isis* **81**: 188–13

Reich L S 1985 *The Making of American Industrial Research: Science and Business at GE and Bell, 1876–1926*. Cambridge University Press, New York

Reingold N 1979 National science policy in a private foundation: The Carnegie Institution of Washington. In: Oleson A, Voss J (eds.) *The Organization of Knowledge in Modern America, 1860–1920*. Johns Hopkins University Press, Baltimore, pp. 313–41

Rosenberg C E 1997 [1976] *No Other Gods: On Science and American Social Thought. Revised and Expanded Edition*. Johns Hopkins University Press, Baltimore

Schlossman S, Gleeson E, Sedlak M, Allen D G 1994 *The Beginnings of Graduate Management Education in the United States*. Santa Monica Graduate Management Admission Council, Santa Monica, CA

Veysey L R 1965 *The Emergence of the American University*. University of Chicago Press, Chicago

Wise G 1985 *Willis R. Whitney, General Electric, and the Origins of US Industrial Research*. Columbia University Press, New York

M. A. Dennis

Universities, in the History of the Social Sciences

Since they first appeared in the twelfth century, universities have been the main if not exclusive setting for the development of the social sciences. No two universities are or have ever been absolutely alike, but all possess common features. A division of intellectual labor is foremost. The curriculum, composed of building blocks of knowledge called 'disciplines,' is organized into multiple teaching, training, or research structures called variously (depending upon time, place, and national language) 'faculties' (the oldest), 'institutes,' 'laboratories,' 'departments,' 'schools,' 'boards of studies,' 'colleges' (which can be freestanding), or, more recently, 'programs' and 'organized research units,' or 'centers.'

1. *The Problem of Knowledge*

Neither the absence of a universal measure of 'discipline' nor the variety of knowledge delivery systems alters the elementary cognitive principle that unclassified information is relatively useless. Knowledge and understanding are acquired through classification and specialization. According to the philosopher and logician Alfred North Whitehead, specialism is the 'natural' path to meaning. Specialism underlies all civilizations. Increasing specialism is a salient feature of modern life, an indispensable requirement for scientific investigation, intellectual discovery, and problem solving.

2. *Institutionalizing Knowledge*

Disciplinary specialization affects the primary internal structure of universities and their systems of teaching, examining, funding, staffing, and even the arrangement of buildings and spaces. Specialties seek what some scholars call 'autonomy,' consisting of control over their own destinies. Accordingly, specialties desire institutional guarantees of disciplinary self-regulation. But the opposite is equally true. The institutional forms themselves influence the content, development, and application of academic specialties. In turn, the missions of educational institutions such as universities are affected by outside pressures generated by governments, the judiciary, churches, the professions, public opinion, the media, and, increasingly, the commercial demands of industrial or high-technology economies. The 'modern' or 'nation-state' has been an important element in the formation of all contemporary disciplines but particularly in the sciences and social sciences. The story of the interrelationships between the social sciences, their host institutions, and all other disciplines, as well as society and government, is consequently complex, unstable, and unpredictable. The rate of knowledge change and production is however more rapid in some historical periods than in others. The resulting disjunctions between the generation and absorption of knowledge can produce social consequences.

2.1 *The Organization of Knowledge in the Medieval University*

In the universities of the Middle Ages, disciplines were incorporated into large divisions known as 'faculties.' Some faculties, as at Paris, were virtually quasi-independent corporations within a single university, but others were less settled in their administration. 'Nations,' groupings of students from European regions, homes away from homes, were often sub-administrative and teaching units too, with separate identities, as were colleges. The 'inferior' (or arts) faculty,' with its broad divisions into language and mathematical studies, was preparatory to the second, the 'superior' or 'higher faculties' in which the three great professional training programs were located—medicine, law (canon or civil), and theology.

Within the inferior faculty were the disciplines that contributed to the foundations of a general education. The seven liberal arts, inherited from Rome, were divided into two other general classifications. One was the language-based *trivium*—logic, grammar, and rhetoric. The second was the mathematically based *quadrivium*—music, astronomy, geometry, and arithmetic.

The basic organization of knowledge into binary arts and faculty teaching divisions was rendered more complicated by the teaching of the 'three philosophies'—natural, moral, and metaphysical—also derived from ancient classical sources. The analytical methods, as well as the substance of the philosophies, informed and interpenetrated all faculties and disciplines in a process of knowledge generation and debate that produced gradual and continuous revisions in the curriculum.

2.2 *The Relationship of Disciplines to One Another*

Therefore disciplinary taxonomy does not by itself adequately explain the uses or content of disciplines. The study of literature, for example texts such as Virgil's *Aeneid*, was included within grammar and thereby subordinated to the teaching of syntax.

Architecture could be found within the *quadrivium*, allied to the mathematical sciences. Rhetoric was a particularly elusive discipline. In some universities (or periods) it was subservient to grammar or logic and often limited to letter writing and religious preaching, but in other circumstances it was an important ingredient in the training of lawyers and public servants. The character of political science (based initially on Aristotle's *Politics*, translated into Latin in the mid-thirteenth century) changed significantly when it was later associated with law and employed in the task of building bureaucratic governments to replace medieval feudalities. Religious questions regarding the existence of the soul were not confined to the faculty of theology but were debated within logic, using such assigned texts as Cicero's or Aristotle's *Topics*.

Science, in the modern sense of the study of nature, and medicine were part of metaphysics. Methods and subjects commingled. In a scribal-based (and before the invention of printing, also oral) academic culture dominated by the writings of a relatively few classical authorities, with their Arabic and scholastic glossators and commentators, disciplinary crossover was readily achieved. The need to focus attention on a limited number of issues, notably those relating to the spread and administrative growth of the universal Church of Rome or the fiscal and legal problems of princely and imperial courts, contributed to a unity of academic purpose.

Yet the differences and principles of knowledge production between the medieval and modern university should not be exaggerated. The general rules about knowledge acquisition are similar. Disciplines absorb conceptions and methods from one another—content overlap—and the direction of investigation is influenced by the uses to which knowledge is put. The capacity of a discipline to describe its own trajectory is constrained (positively or negatively) by the faculty or disciplines (or professional associations) to which it might defer.

Today's universities have characteristics similar to earlier universities. The placement of contemporary disciplines like sociology, political science, ethical philosophy, or even cultural anthropology within a school of law determines the approaches and problems that receive primary attention. The study of economics, when situated within business schools, inclines more towards problem solving than towards the laws or principles underlying economic behavior viewed as a science. Historical inquiry, when located within departments of economics, political science or policy studies, music, or education, becomes the study of the evolution of those fields, or is oriented toward questions of interest to the supporting departments.

Anthropologists whose appointments lie within medical schools and hospitals address issues arising from the treatment and care of patients, as do ethical philosophers, psychologists, or social workers in the same environment. Architecture is affected differentially when associated with engineering, schools of urban planning, colleges of the environment, schools of design, or when viewed as a member of the liberal arts. The study of music in research universities favors composition or musicology or the history of music more than performance. Examples of disciplinary plasticity and allegiance are endless.

3. *Continuities and Discontinuities in Knowledge Acquisition*

The medieval university was not ideologically designed for intellectual discovery but for the dissemination of received knowledge. Nevertheless, significant knowledge growth and development transpired, if often unrecognized by practitioners or credited by polemicists. Change was advanced by the normal process of argument, by the discovery of new texts, by the reinterpretation of older sources, by the evaluation of existing systems of logic and proof, by theological controversies such as the Great Schism of the fourteenth century, and by issues concerning ecclesiastical reform eventually culminating in the Protestant heresies.

An essential difference between the medieval or later universities with respect to knowledge production is arguably the much higher degree of self-conscious disciplinary promotion within the latter as driven by professional and external imperatives. A second difference is the larger amount of ambiguity and relativism tolerated by modern cultures.

3.1 *Scholasticism and Humanism*

The period termed the 'Renaissance' (now usually called the 'early modern' period) or 'rebirth' of classical learning, beginning with fifteenth-century Italy and spreading northwards in Europe in the centuries following, has long occupied a central place in discussions concerning the differentiation of disciplines and the arrival of 'modern' conceptions of learning. Historians conventionally divide the learned into two groups: the medieval schoolmen and the early modern humanists. The issues dividing them were broadly three in number. First was the type and variety of classical sources later available when the Greek Byzantine empire collapsed in Asia Minor. Authors and texts unknown or known only by hearsay were brought into Western Europe. Sources could now be compared, stimulating an interest in variorum texts and the detection of forgeries. Historical criticism, whose place was hitherto minimal, developed as a cardinal feature of early modern scholarship and was applied relentlessly to canonical texts and the dogmatic foundations of Western Christianity. A second issue concerned the role of observation and experiment

in the conduct of scientific work. The humanists and practitioners of the 'new learning' charged that inductive reasoning was neglected in favor of deductive systems of logic derived from Aristotle and Arab scholars. A third issue was the purpose of the university itself. Detractors alleged that the medieval university was isolated from practical affairs, its members preoccupied with abstract knowledge and shopworn or irrelevant ideas. A corollary point involved personal style. Humanists stressed public speaking, appearance, and the *vita activa*, being interested in careers at the courts of state-building aristocracies in an age of conspicuous political symbols and ceremonies. They contrasted the philosopher-statesman style with the torpid and pedantic manner of university teachers.

Stereotypes, however plausible, are notoriously dangerous, especially when shaped by polemical intent. Not all humanists were outside the university system. Many remained within, attempting, with success, to alter scholastic habits. Some scientists—Galileo is the *ur*figure—borrowed heavily from scholastic methods of interrogating nature and were not as dependent on experiment and observation as they or encomiasts claimed. And some humanists, in their zeal to attract the patronage of rulers, became hired propagandists, sacrificing their integrity as scholars to hypocrisy and superficial elegance. Furthermore, whatever directions humanistic scholarship took, it is now evident that the world of new learning was first framed within the intellectual structures and methods created by the medieval universities.

3.2 Alternatives to Universities

Nevertheless, in the centuries that followed, criticisms of the university as hidebound and tied to an antiquated system of teaching were convincing enough to result in the introduction of different institutional forms and new sources of funding in which innovation, particularly of an applied nature, received a more immediate reception. In fifteenth-century England the teaching of the common law was given a home in institutions designated 'inns of court,' leaving Oxford and Cambridge universities with the teaching of canon law, Roman law, and jurisprudence. In the centuries following, academies, botanical gardens, royal libraries, observatories, museums, military schools, and new colleges were established outside university systems. In France in the course of the eighteenth century the state's need for experts resulted in the creation of specialized engineering schools,starting with the *École des Ponts et Chaussées* in 1747. The Jacobins of the French Revolution adopted this form for their own instrumental purposes. Their *École Polytechnique* of the 1790s was followed in the next century by other *grandes écoles*, institutions of high quality whose mission was to train specialists for appointments in government, industry, engineering occupations, or education. At the same time the French university system was dismantled by Napoleon as irrelevant to the tasks of a revolutionary and secular society. That system was only gradually rebuilt but never attained the prestige of the leading schools. Only recently has it acquired a few of the research functions typical of universities in other countries, high-level research in France being normally assigned to a different system.

The French *grandes écoles* are unique institutions in their patterns of student selection, purpose, and prestige. They are also unique because their prestige is not based on a research mission as in other countries (although now several research functions have been added). Few educational institutions anywhere have equivalent elite connections to government, bureaucracy, industry, and the leading professions. However, other countries have attempted to emulate some aspects of the French engineering model, most notably in the establishment of upper-level polytechnical establishments in European capitals such as London, Berlin, Vienna, Stockholm, and Copenhagen, and in major US cities such as Cambridge, Massachusetts. Many of these have been evolving towards a university form. Relevant social science departments, foreign language teaching and, in some cases, humanities courses were added, especially in the twentieth century.

In Germany and Austria other specialized universities were created: *Wirtschaftsuniversitäten*, for example—institutions where disciplines could be grouped and united with a focus on economic problems. In the UK regional universities and university colleges were established that combined applied with liberal arts studies but with special attention to the economic or social problems of their regions.

Medical schools and hospitals, freestanding or associated with universities, became umbrella or 'incubator' institutions for an astonishing range of clinical and scientific specialties. Also in the nineteenth century, geodetic surveys and sponsored explorations supported certain kinds of limited intellectual activity. Although professional education as contained in the 'superior faculties' was the featured curriculum of the medieval university, and especially law in the later centuries, the admission of other forms of vocational or occupational training was never guaranteed. Law, medicine, and theology were acceptable because they suited the needs of urban societies in which religion and secular life interpenetrated. Vocational latecomers had to fight for entry. Rivals for income and teaching appointments, for students, space, and a role in governance, they represented the interests of outsiders. Hence it was that newer professionalizing occupations of the later centuries, especially the nineteenth, such as engineering, commercial studies, schoolteaching, public administration, and social work, were associated with separate schools where they might develop an applied mission according to their own rules and methods. Several occupations—notably architecture,

accounting, and some legal and medical activities—were controlled by practitioners recruiting through guild apprenticeship methods. But guild and academic forms of education are not necessarily incompatible. The academic profession itself is the supreme example of guild-directed education. Early twenty-first-century universities tend to be comprehensive. Subjects or emphases once deemed unacceptable to universities built on the German research model or influenced by it are now more likely to be welcome. The general terms of inclusion have been mentioned. New recruits to the university community are normally expected to conform to its prevailing styles of intellectual achievement. However, the rapidly increasing fragmentation of the knowledge base everywhere evident suggests that the universities of the future will be less able than predecessors to impose uniform standards of work or a common scholarly ethos on their numerous departments, laboratories, and programs of instruction.

4. *The Transformation of the University*

Applied and experimental science owed more to medieval systems of logic and natural philosophy than critics granted, yet there was indeed a 'scientific revolution' of the seventeenth century that transformed the medieval worldview, produced a materialist conception of the cosmos, departures in physiology, and eventually the treatment of disease, and revised theological speculation concerning human understanding and potential. Mathematical measurement developed as a formidable scientific tool; for despite the medieval *quadrivium* with its mathematical base, the acceptance of Aristotle's physics by the schoolmen had produced a science that was heavily qualitative and metaphysical in character. Scientific instruments such as the microscope and the telescope, and later the Leyden jar, greatly enhanced the scope of intellectual achievement. Sophisticated machines have continued to be indispensable for science (and certain fields of social science), and never more than in the early twenty-first century.

4.1 *The 'Knowledge Revolution' of the Nineteenth Century*

Historical criticism, the development of the laboratory, and new systems of biological classification underlay the eventual adoption of a research ethic that completely reoriented the pantheon of disciplines inherited from the past.

Teaching and professional training were the two educational missions of the medieval university. Now there was a third, to be called 'research.' The ideas that knowledge could be created, that discovery was an aim of all scholarship, and that the new was preferable to the old were revolutionary assumptions that profoundly affected all educational institutions, and especially universities. The research spirit, while commencing in the scientific revolution and noticeable here and there in later Enlightenment Scottish, German, Dutch, or Swedish universities, did not dominate university culture itself until the nineteenth century. Historically, it is most closely associated with the German university model.

The general acceptance of an academic culture oriented towards original inquiry gave enormous impetus to the establishment of new or subdisciplinary fields. Pressures for the autonomous representation of those fields within universities increased. But two lasting versions of a research ethic emerged: one advocating discovery as an end in itself, the other advocating the use of research results to solve social, economic, or political problems, or to serve the needs of a democratic, consumer society. Proponents of the first feared that applied research funded mainly from commercial sources or results-minded governments would transfer direct and certainly indirect control over discipline formation and knowledge production to agencies outside the universities.

'Pure' research would suffer. Advocates of the second noted the disappearance of any useful distinction between 'pure' and 'applied' knowledge in the age of high technology. In any case, contemporary society, they maintain, requires applied knowledge as a condition of survival and prosperity. Arguments remain intense; and the available evidence suggests that both sides have a case.

4.2 *The Effect of the Social Sciences on Universities*

As observed, the 'social sciences' were embedded in the scholastic arts curriculum. The bureaucratizing states of the early modern period found a place for some forms of applied social science, notably *Kameralwissenschaft*, within the law faculties. The humanists, however, strengthened belles lettres within the faculty of arts (later, in the German university, to be called the faculty of philosophy). Accordingly, the maturation of the academic social sciences was a slow, two-stage affair. First came new methodologies for the study of social change and process developed by thinkers both inside and outside universities. Second was the emancipation of social science from domination by arts and professional curricula. Later Enlightenment Scottish university historians and philosophers developed 'conjectural history,' incorporating some of the insights of political economy and early theories of evolutionary change. Nineteenth-century French intellectuals such as the St. Simonians and Auguste Comte developed some of these beginnings into an overarching scheme of human development based on the 'march of mind,' the progress from primitive religion through metaphysics to 'positive science.' While the separation of subjects such as sociology or anthropology or economics, or psychology from such parent fields as law, history, religion,

and philosophy was a nineteenth-century development, the flowering of social science as a full-fledged set of university disciplines belonged to the twentieth century. The timing was not accidental. A focus on the comparative, normative (and deviant) behavior of groups, communities, and classes; on the functioning of institutions such as schools, bureaucracies, professions, religious organizations, trade unions, industrial firms, and family structures; and on the nature of belief and value systems of societies reflected the dominant features and problems of modern society with its mobile, plural populations and global reach.

Besides concentrating on issues and problems not conventionally incorporated into the inherited model of a university, the social science disciplines claimed the prestige gained by the physical sciences. As in the natural order, the human or social order could also be explained by scientific laws. And although in time some of the rigor and promise of words such as 'scientific laws' would be mitigated and exchanged for softer terms such as 'patterns' or 'regularities,' the claim to scientific objectivity served as a major legitimizing strategy. It was also the basis for an instrumental approach to the resolution of social problems and the means by which the social sciences secured access to external funding sources.

The number of social science disciplines and the different strategies pursued by social scientists to become citizens of the republics of letters and science in varying national contexts is a highly complex story. Yet some insight into the issues may be gained from a cursory look at the history of the London School of Economics and Political Science (LSE), founded in 1895 by socialists and left-wing liberals on the model of the *École Libre des Sciences Politiques* in Paris (the LSE was later incorporated into the federal structure of the University of London).

Two issues confronted the founders. The first was the independence of the LSE, solved initially by the creation of a wholly new establishment where the disciplines could function on their own terms. This conforms to the pattern indicated earlier, where alternative institutions provide the means of disciplinary entry into the knowledge universe. The second issue concerned the LSE's primary mission, and here the founders disagreed on basic aims and methods. Should the institution's mission be applied or scholarly? Should attention be focused mainly on metropolitan regions such as London or spread more widely? Were ideological commitments to social amelioration compatible with standards of objective research? These differences of purpose have never been completely resolved within the LSE's academic divisions.

Indeed, the differences cannot be resolved since they are inherent in the history of the social science subdisciplines. The structural flexibility of the modern research university has provided a practical solution. The applied social sciences have found homes in the professional schools, a number of which—social work, urban studies, education, public policy, and perhaps now journalism and media studies—are in fact based on the social sciences. But law, medicine, public health, and architecture all have room for social science subjects and methods. Elsewhere in the university the 'autonomous' social sciences are separated into distinct disciplines according to the history of each. They are either located in departments within 'colleges of letters and science' (the typical US case), traditional arts faculties or new faculties of social science (in some European universities). They are also to be found in crossover and multidisciplinary programs and research units.

Ideas, concepts, and problem-solving methods developed by the social sciences are now commonplace elements of any university curriculum. Historical research is greatly influenced by psychology, anthropology, economics, sociology, statistics, and political science; but history, as a 'universal recipient,' has often if not invariably demonstrated intellectual flexibility. Yet literary studies have also become social scientific as, *inter alia*, critics incorporate interpretations derived from the sociology of class relationships, the psychology of the individual, and the anthropologist's interest in symbol and ritual.

In concert with the technical and physical sciences (and the revolutionary biological sciences), the social sciences have played a prominent role in tying the university to social or national objectives, strengthening its place in the formulation of policies affecting the distribution of wealth, educational opportunities, the provision for social services, and other polemical issues that are the concern of modern, mass democratic societies. The entanglements with powerful and influential outside interests are not to everyone's liking, as they compromise the university's decision-making independence. However, the university has never been a wholly autonomous institution, not even within free and liberal societies. Such relative independence as it enjoys is the result of negotiation or 'social contracts' that each generation of university leaders, teachers, and researchers is destined to conduct.

5. *Controls over Disciplinary Formation*

The organization of the disciplines varies in different national and university settings, affecting the kinds of disciplinary emphases receiving attention.

5.1 *The Chairholder System*

The medieval faculties were the ultimate internal determinants of the curriculum, with alternative institutions such as colleges providing some opportunities for independent innovations. Centuries later, with the evolution of the professorial or 'chair' system of academic leadership, best known in its German form,

individual chairholders administered 'institutes' and laboratories and set the standard for teaching and research, especially at advanced levels. Although the institute structure was embedded in the traditional faculty structure, changes had occurred that made chairholders and deans the effective authorities, and they respected one another's independence.

This system, which became standard in continental Europe, Latin America, Israel, Japan, and wherever the German model of a university took hold before 1939, led the way in world scholarship and science. But an essential drawback was that younger academics with newer specialties could not easily find room within a structure dominated by seniors. Their career alternatives were to await chair vacancies at home or elsewhere, or to persuade institutions (and the government ministers who largely controlled financing, especially after the later nineteenth century) to establish new or different chairs in developing subdisciplines. Before unification in 1870, 'Germany' was a collection of independent *Länder* competing against one another for professorial talent. Thus intra-German rivalry itself was a variable in the formation of new fields of inquiry, along with the adoption of radical critical methods (the 'Higher Criticism'), and philosophical conceptions concerning intellectual objectives (*Bildung*) and research ('Humboldtianism').

5.2 Alternatives to Faculties and Institutes

Many of the features of the organization of knowledge that had evolved in Europe also existed in the USA, whose higher-education system, first established in the colonial period, reflected English and Scottish influences, then French and finally German ones. The chairholder system was much in evidence throughout the USA as well, but by the end of the nineteenth century it was losing its authority as the arbiter of the curriculum. It was gradually supplanted by the 'department,' the dominant US form of curricular organization. A century later its main features were being adopted in most university systems elsewhere. The essence of a department is that it is based on a discipline rather than a collection of disciplines (as in a faculty structure), is run by committees, and normally features a rotating head. Chairholders or their equivalents, while senior and influential, share authority with other teaching titles. The relatively 'democratic' decision-making apparatus of the department is broadly in keeping with medieval legacies of guild self-government.

Another US innovation, a credit-unit system of instruction that combines teaching and examining in a single person, was a departure from UK and continental European external examining practices, and underscored the intellectual freedom of individual instructors and researchers. In this context, subdisciplines were able to flourish. While it is still possible for an entire department to be built or rebuilt along narrow and exclusionary disciplinary lines, such solutions are generally short-lived given the independence of individual members and the variety of available funding sources.

One other US innovation of the twentieth-century deserves mention: the advent of the graduate school. The continental European chairholder system was suited to advanced instruction and the award of advanced degrees because it relied on a selective system of elite schooling—the *lycée* in France, the *gymnasium* in Germany and Scandinavia, their counterparts elsewhere—and gave the professoriate discretion in choosing the best candidates for creative work. The US university, which for historical reasons was not well supported by an excellent system of secondary school instruction, recruited undergraduates of mixed preparation. The solution to the problem of how to offer research-level instruction was found in the establishment of a wholly separate administrative division freed from the necessity of providing remedial and general teaching. This innovation contributed greatly to the building up of strong research traditions within the social sciences and other disciplines. Versions of the US graduate school appeared in Europe during the 1980s and 1990s in response or reaction to governmental policies encouraging the development of mass forms of secondary and higher education.

Before and during the introduction of departmental and credit-unit systems of instruction, the US pattern of disciplinary formation was similar to the European one. Subjects that did not find an immediate reception within traditional colleges and universities, especially subjects stigmatized as 'practical' (which could include modern languages), sought a more hospitable environment in alternative institutions. These had been springing up ever since the American Revolution.

By the end of the nineteenth century liberal arts colleges, comprehensive universities (combining professional and liberal education), specialized institutes (such as the West Point Military Academy), prestigious polytechnics (such as the Massachusetts Institute of Technology), a German-style 'research university' (the Johns Hopkins University), and state institutions with applied but also liberal education missions (the 'federal land grant' universities) were members of a differentiated system that had expanded in connection with national ambitions and market conditions.

Within the older US colleges and universities, internal reconfiguring took place alongside external differentiation. In the earlier nineteenth century, academic subjects such as the natural sciences or modern languages that were not necessarily applied but were regarded as inimical to prevailing teaching values were able to create alternate degree tracks, if not of equivalent prestige. Within the senior English universities of Oxford and Cambridge, subject-based 'schools' (Oxford) or 'boards of studies' (Cambridge)

representing specialties effectively replaced the medieval faculties structure. As in the German case, professorial chairholders controlled the university system of lecturing. However, as Oxford and Cambridge were collegiate universities or federations of relatively autonomous societies, the professorial system did not constrain the whole system of instruction. The collegiate structure with its small classes and tutorial arrangements in fact liberated the professors from routine or introductory teaching.

5.3 Criticisms of the Departmental Organization

The flexibility inherent in a departmental organization and its comparative openness to innovation reduce the temptation for subjects to migrate toward alternative institutional structures. Departments are therefore one of the factors that have brought the 'university' institutional form into the center of modern society as the principal source of knowledge generation and dissemination. Others—and they are outside the scope of this entry—are national imperatives, such as war and industrial competition, mass higher education requiring extensive private but especially public resources, professionalism, consumer economies, and the high division of labor so typical of contemporary society.

Yet it would be strange if the departmental organization did not have its critics. Discipline-based departments are said to be narrow in their own right, protective of privileges and unreceptive to cross-disciplinary fields of study. A similar criticism but with a different purpose arises from the defenders of US traditions of liberal education. Departments are attacked as the chief obstacles to the development of general education studies for undergraduates. In both instances the response has been the creation of interdisciplinary or multidisciplinary departments, or cooperative programs of teaching, as well as special appointments in broad fields such as 'Humanities' or 'Social Sciences.' Such experiments are now typical of newer European universities.

But an opposite kind of blame has also been voiced, most notably by two famous US *enfants terribles*: the maverick Thorstein Veblen at the turn of the twentieth century and the more establishment 1930s critic Abraham Flexner. Both excoriated US departments for being far too open to innovation, to fads, and to market demand and far too eager to attract the financial support of philistine parvenus. Flexner in particular admired the German university tradition with its high intellectual standards and *Kultur*. For him, the elite characteristics of the European higher-educational system—the alliance between government bureaucrats who decided educational budgets and chairholders who controlled institutes—meant that demands for the teaching of subjects unworthy of incorporation into the noble inheritance called a 'university' could be successfully resisted.

6. Universities of the Twenty-first Century

In the twentieth century the views of critics like Flexner proved to be shortsighted. Totalitarian governments in Nazi Germany and the communist states of Central and Eastern Europe, as well as their copies elsewhere, provided vivid lessons of how knowledge could be manipulated from above and turned from its natural channels of curiosity and negotiation to unimaginable and terrifying applications. The university systems of those countries had to be rebuilt after World War II, and the task remained unfinished as the twenty-first century commenced. While it is nearly an axiom of twenty-first-century higher-education policy in all democratic societies that universities must be flexible in providing access and instruction, disagreement exists on the institutional forms best suited to promote those ends. Technologies associated with the arrival of the 'virtual' or 'cyberspace' university and the possible competition of degree-granting 'corporate universities' have issued new and formidable challenges to all existing forms of higher education. Given numerous choices, it is safe to predict that the longstanding tensions between outside interests, institutional structures, and the formation of disciplines are certain to continue.

See also: Centers for Advanced Study: International/Interdisciplinary; Disciplines, History of, in the Social Sciences; Higher Education; History of Science; History of Science: Constructivist Perspectives; History, Professionalization of; Knowledge Societies; Knowledge, Sociology of; Policy Knowledge: Universities; Professionalization/Professions in History; Science, Sociology of; Scientific Academies, History of; Scientific Knowledge, Sociology of; Universities and Science and Technology: Europe; Universities and Science and Technology: United States

Bibliography

Ben-David J 1991 *Scientific Growth*. University of California Press, Berkeley, CA

Clark B R 1987 *The Academic Life, Small Worlds, Different Worlds*. Carnegie Foundation for the Advancement of Teaching, Princeton, NJ

De Ridder-Symoens H (ed.) 1992 *A History of the University in Europe, I*. Cambridge University Press, Cambridge, UK

De Ridder-Symoens H (ed.) 1996 *A History of the University in Europe, II*. Cambridge University Press, Cambridge, UK

Flexner A 1994 *Universities, American, English, German*. Transaction Publishers, New Brunswick, NJ

Geiger R L 1986 *To Advance Knowledge. The Growth of American Research Universities, 1900–1940*. Oxford University Press, Oxford, UK

Graham L, Lepenies W, Weingart P (eds.) 1983 *Functions and Uses of Disciplinary Histories*. Reidel, Dordrecht, The Netherlands

Kelley D R (ed.) 1997 *History of the Disciplines. The Reclassification of Knowledge in Early Modern Europe*. University of Rochester Press, Rochester, NY

Kerr C 1995 *The Uses of the University*, 4th edn. Harvard University Press, London
Oleson A, Voss J (eds.) 1979 *The Organization of Knowledge in Modern America, 1860–1920*. Johns Hopkins University Press, London
Rothblatt S 1997 *The Modern University and Its Discontents. The Fate of Newman's Legacies in Britain and America*. Cambridge University Press, Cambridge, UK
Soffer R N 1978 *Ethics and Society in England: The Revolution in the Social Sciences, 1870–1914*. University of California Press, Berkeley, CA
Stocking Jr. G W 1995 *After Tylor, British Social Anthropology, 1888–1951*. University of Wisconsin Press, Madison, WI
Veblen T 1965 *The Higher Learning in America. A Memorandum on the Conduct of Universities by Businessmen*. A. M. Kelley, New York
Wagner P 1994 *A Sociology of Modernity, Liberty and Discipline*. Routledge, London
Wagner P, Wittrock B, Whitley R (eds.) 1991 *Discourses on Society. The Shaping of the Social Science Disciplines*. Kluwer Academic, Dordrecht, The Netherlands
Weisz G 1983 *The Emergence of Modern Universities in France, 1863–1914*. Princeton University Press, Princeton, NJ

S. Rothblatt

Unobtrusive Measures

1. Introduction

All measures must be considered to be subject to error. We would have no way of knowing if, perchance, a measure turned out to be absolutely accurate. For example, the National Bureau of Standards has a 10-gram weight which, because of a manufacturing error, weighs just less than 10 grams by about 400 micrograms (the weight of a grain or two of salt) (Freedman et al. 1991). Despite the most careful weighing, done on a weekly basis, the values actually obtained for the standard weight vary by about 15 micrograms (one microgram is about the weight of a speck of dust) either way from the mean. Presumably the mean of a long series of measures is the best *estimate* of the true weight of the standard, but there is no way of being sure of the true weight. Obviously, the way to deal with such errors, usually thought of as *random* errors, is to do the same measure several times. That is why careful carpenters measure their boards more than once before cutting them—especially if the wood is expensive. Random errors are associated with the concept of *reliability*—the expectation that if a measurement is performed twice under exactly the same conditions, very close agreement should be obtained if the measurement instrument and procedures for using it are dependable.

Another type of measurement error, though, is *bias*, i.e., consistent deviation from the 'true' value of the object or phenomenon of interest. For example, a recent news story recounted speculation that, in order to enhance the prospects of players for careers in professional football, some collegiate coaches might have arranged for the running tracks on which the players' speed is tested to be slightly downhill. If that is true, then the running speeds of players from some universities may be biased toward the fast end of the scale. If mothers are asked to estimate the intelligence (or good looks!) of their children, we might expect some bias to be evident, i.e., higher estimates than would be expected from other means of assessment. The bias is not reduced by replicating the measurement. Downhill is still downhill, and mother's love is constant.

The problem of bias has to be dealt with either by knowing the degree of bias so one can allow for it, or by using multiple measures that do not share sources of error, i.e., bias. A football scout who knows that a player has been tested on a track with a downhill slant can 'allow' for that in interpreting reported running speed. We may expect that mothers will exaggerate a bit in describing their children's skills or other virtues and discount the glowing adjectives in the descriptions by some amount. But an even better way to deal with bias is to use other measures that are less biased, or not biased at all in the same way as the original measure. Of course, if a scout knows that a track has a downhill bias, that scout may simply retest the athlete on a track known to be quite level and disregard the first, presumably biased, report. A scout might also use game films to arrive at a judgment of the speed of the player.

In social and behavioral sciences we usually have no absolute standard that can be applied, and we often should not assume that any one measure is less biased than another or unbiased altogether. The best solution is to use multiple measures of constructs and pick or devise those measures in such a way that they have minimally overlapping sources of error. A sports scout may not have any way of knowing that other running tracks are level, but if an athlete is tested on three different tracks, the average speed of those three runs is likely to be less biased (as well as having less random error) than any one of the runs. And, if all three tracks have a downhill slant, then maybe most tracks do, and the estimate of the athlete's speed is not biased with respect to estimates for other athletes.

2. Problems Associated with Single and Reactive Measures

Social science measures are susceptible to many sources of bias, but a few of those sources of bias are particularly important because they are common and may involve relatively large degrees of bias. Different sources of bias are inherent in every measurement procedure. The biases can pose threats to the valid

interpretation of a measure, in a manner akin to plausible rival hypotheses as threats to validity of experiments (Campbell and Stanley 1963).

3. Reactive Measurement Effects

A problem in measurement stems from the fact that the very act of measurement may produce a reaction on the part of the measured object that changes that object, whether temporarily or permanently. One cannot determine the tensile strength of a wire without breaking it, and one cannot determine whether a cake is as delicious as it looks without spoiling it to some degree. Asking to measure a person's height usually results in that person assuming a very erect posture that produces an estimate of height that is taller than their 'walking around' height. Similarly, giving a group of people a 'bigotry' scale is likely to result in an underestimate of the mean level of bigotry in the group, even if the scale is not explicitly labeled 'Bigotry Scale.' Some reactive measurement effects may be of limited duration or generalizability. The effect of bigotry measurement is unlikely to persist, but others may be long lasting, e.g., as when asking a person his or her opinion about something results in reflection leading to crystallization of an opinion that may have previously been ephemeral.

Scientists and others involved in measurement in social sciences often attempt to reduce reactivity in various ways, including telling respondents that their responses will be confidential, or even anonymous, or that their responses will be used only for scientific purposes. Often such generic efforts to reduce reactivity are not enough, or at least not enough to be completely reassuring, in which case the response of the investigator may be to try to identify and employ a nonreactive, or at least less reactive, measure. Although all measures must be considered reactive to some degree, they are not all reactive to the same measurement issues or to the same degree. Other measures may be thought to be relatively low in reactivity because the data on which they are based were collected in spatially and temporally remote ways that should have freed them from contemporary biases. One example would be the use by historians of personal letters in order to diagnose the state of mind of the letter writer at some earlier period in his or her life.

The most obvious and attractive way of reducing reactivity is to carry out measurement activities under conditions that do not require the subject to know that he or she is being measured. Such measures have come to be known as *unobtrusive measures* (Webb et al. 1966). A young businesswoman might, for example, be invited out for lunch without knowing that she was actually being 'sized up' for a promotion. Or, computers in a school might be programmed to record sign-ons by individual students so as to provide a check on students' reports of computer use and study time. Measures may be unobtrusive because they are embedded in or derived from ongoing activities that do not make measurement a salient feature of the activity or because they are concealed in some way. The term *unobtrusive* has come to be used generically for nonreactive measures, but it would be better used to refer specifically to measures obtained without the necessity for persons being measured to be aware of the fact of being measured. It is for that reason that the original publication, *Unobtrusive Measures: Nonreactive Measures in the Social Sciences* (Webb et al. 1966) was retitled *Nonreactive Measures in the Social Sciences* (Webb et al. 1981) in the revised version of the book. Some measures, e.g., many physiological measures, are not highly reactive (susceptible to bias) because they are not under voluntary control; other measures are not particularly reactive because the person being measured is mistaken about the purpose for which they are being measured; and still others because the measurement context is conducive to unbiased responding.

When people are aware of being measured, a frequent consequence is bias, even if the participant is cooperative and well intentioned. People being 'tested' often, very naturally, want to make a good impression, although in some specific instances, a person being measured may want to make what would ordinarily seem to be a bad impression, e.g., malingering. People being measured may adopt specialized roles that reflect their ideas about how they ought to behave, and those roles may not be characteristic of them when they are in other situations. Individuals will differ, of course, in the degree to which their characteristic responses are affected by knowledge of being measured. Bias may be limited in some instances because people believe their responses are completely appropriate. Criminals, for example, often believe that their behaviors are quite justified and therefore feel no need to disguise their statements about them.

Biases can be reduced with the implementation of multiple measures that do not share the same sources of bias. Unobtrusive measures are not free from their own sources of error, but they may not reflect the same type or degree of bias or reactivity as other measures. Unobtrusive measures, because they reflect different approaches to measurement, can often get around the limitations of reactive measures.

4. The Varieties of Unobtrusive Measures

Webb et al. (1966, 1981) provide an extensive review of unobtrusive or nonreactive measures in social science research and many fine examples can be found throughout the literature. Generally speaking, unobtrusive measures can be usefully categorized as follows:

(*a*) *Simple observations*. Many interesting behaviors can be directly observed, often without the necessity of the observed actor being aware of the fact. Observa-

tions may be made of individuals and groups. Targets of observation may include objects and events as well as persons. For example, the nature of ceremonial events may be of great interest. Also included here are observations of physical location and clustering of people, expressive movements, language behavior in the media or as overheard in public areas, the amount of time individuals spend gazing at public displays, or time sampling of observations to determine whether certain occurrences are linked temporally. A recent news story reported an observation that military dictators who begin to detect resistance and who desire to establish their legitimacy tend to abandon their military uniforms and don civilian attire, with recent appearances of Saddam Hussain of Iraq in regular business suits cited as an interesting example.

(*b*) *Contrived observations*. Simple observation can be tedious as one may need to wait a long while for behavior of interest to occur. Under such circumstances, researchers may contrive situations likely to produce relevant responses. The TV show 'You're on Candid Camera!' was based on responses of people to contrived situations. Social psychologists make extensive use of confederates in contrived situations in which the subject is unaware that he or she is a participant in an experiment. Also included in this category of contrived situations are the experiments in which unsuspecting individuals are deceived into doing things and studied in the process. Real estate companies are routinely tested for racially discriminatory practices by persons pretending to be clients. Bochner (1979) has made 'wrong number' phone calls and staged fake collapses on trains to assess individual's helpfulness in common situations. These types of investigations can be classified as nonreactive or unobtrusive because the participants are unaware that they are participating in an experiment. This lack of awareness implies that their reactions will be natural. Contrivances may include the use of hidden hardware such as audio- and videotapes.

(*c*) *Physical traces*. These include both erosion and accretion measures. Measures of erosion include measurements of floor tile or carpet wear in front of various museum exhibits to determine their popularity, or the wear and tear of library books to examine selective reader interest. Measures of accretion include indicators of inscriptions in public restrooms, the amount of debris left from a ticker-tape parade, or the number of cigarette packs that were thrown out with the trash in selected residential areas. Detectives regularly rely on physical traces of responses they have never seen to solve crimes.

(*d*) *Archives*. These include actuarial records. Materials supplied by the mass media, industrial and institutional records, sales records, and private written documents are archival records. Archival records are regularly used to track behaviors that might not otherwise be detected with reactive measures. Episodic personal and private records are in heavy use in hearings involving activities in the White House to check on the veracity of self-reports by those whose actions are in question.

It is important that unobtrusive measures not be regarded as substitutes for other kinds of reactive measures. That is the antithesis of the rationale behind unobtrusive measures. Unobtrusive measures are properly thought of as complementary to other measures such as questionnaires and interviews.

5. *The Uses of Unobtrusive Measures*

The countering of bias in measurement is of focal concern in considering the usefulness of unobtrusive measures, but such measures may actually be adopted for other purposes. That is in part because an investigator may follow different strategies for dealing with bias: verification, adjustment, and avoidance. An investigator may believe that measures being used are not greatly biased, but he or she may also believe that caution requires sensitivity to possible bias. That investigator may elect to use one or more nonreactive measures in order to probe for bias in case it exists. For example, in a workplace study of attempts to induce exercise, a researcher may ask respondents how often they walk up stairs rather than take an elevator, and reports of workers might be regarded as generally unbiased. Nonetheless, the investigator might still elect to probe the accuracy of those reports by using occasional observers to determine whether the proportion of people climbing stairs is consistent with self-reports. If it should appear that workers tend to exaggerate their reported use of stairs, the investigator might change the questionnaire to try to elicit better data or abandon self-reports in favor of reliance on observers. A prudent investigator would, at the very least, be cautious in interpreting self-report data. Probing might not require large numbers of observations or observations of all subjects of interest. A second strategy is to collect sufficient data by alternative means to make it possible to estimate by how much the numbers obtained by a primary measure (likely to be a questionnaire) are out so that estimates can be appropriately discounted. Observational data might indicate that reported use of stairs is exaggerated by 25 percent, in which case the researcher might deflate his estimates of total physical activity by an appropriate amount. Obviously, unobtrusive measures may be used in conjunction with other measures to arrive at a summary assessment that is likely to be less biased than would have been the case for a single measure. If trainees in a program claim to be spending 10 hours per week in a computer laboratory, but actual counts of people working in the laboratory are not consistent with the claimed level of use, a researcher might adjust downward estimates of exposure to the exercises involved.

It is not always easy, however, to obtain unobtrusive measures that can be used directly and exactly in order to adjust for biases in other measures. One major problem is that more common measures such as questionnaires and unobtrusive or other nonreactive measures are not in the same metrics, so that combining them is not easy. What would one do, for example, with questionnaire data indicating that people claim to eat four servings of vegetables every day and observational data indicating that many people fail to finish eating and discard portions of vegetables served to them? Such discrepant data may lead one to question the accuracy of one set of data or the other, but the discrepancies may not translate easily into quantitative adjustments.

Which data should one trust? Certainly our inclination would likely be to distrust questionnaire and other self-report data. Campbell (1969) suggested as much in discussing the role of qualitative data in program evaluation. Emily Dickinson once wrote the line in a poem 'I like a look of anguish because I know it's true.' That is likely to strike all of us as obviously true. Try to imagine the alternative statement 'I like a self-report of anguish because I know it's true.' And yet, there certainly must be times when self-reports are better assessments of underlying dispositions than overt behaviors. Think, for example, of the pressures toward conformity that lead people who may not be at all religious to bow their heads when other people pray, or the politeness that may prevail between politicians who dislike each other. The use of multiple and different measures cannot guarantee anything.

Under some circumstances, nonreactive measures may be available and be used alone. That may be because they are sometimes compelling, but also because they are sometimes inexpensive. In particular, archival records may be exploited, often at fairly low cost. If records are kept for reasons unrelated to any particular policy use, or at least for reasons unrelated to the purposes for which they are used in research, they may be of great value and characterized by very little bias. The identification by Dr. John Snow of a contaminated water supply as the cause of a cholera outbreak in London in the 1850s was facilitated by records kept of the location of individual cases of the infection, making it possible for Snow to map those cases onto alternate water systems (Freedman 1991). Barthel and Holmes (1968) were able to use information in high school yearbooks to show that persons who later became schizophrenic had low levels of social activity before the onset of their illness.

6. *Advantages of Unobtrusive Measures*

Aside from the possibility of reducing bias in measurement, several other advantages are often associated with unobtrusive measures (Rathje 1979, Babbie 1989). Typically, unobtrusive measures require little, if any, effort on the part of persons being assessed, and often no physical contact with, or even close proximity to participants is required. Unobtrusive measures tend to focus on behaviors and obviate the problems that may stem from inaccurate reporting on the part of respondents. Yet another advantage is that the employment of nonreactive measurement procedures is often relatively inexpensive, e.g., simple observations, physical traces, and archival records. They can be of great value in longitudinal studies.

7. *Limitations of Unobtrusive Measures*

Unobtrusive measures, like all other measures, have limitations. In the first place, it is not always easy to identify unobtrusive measures. Unobtrusive methods are typically limited in certain areas that can be open to interrogation with more reactive methods such as interviews and questionnaires. Methods that utilize verbal communication, such as structured and unstructured interviews, have 'an ability to reach into all content areas' (Webb et al. 1966, 1981). Webb et al. (1981) provided a 'generative taxonomy' for nonreactive measures in order to facilitate thinking about them, but very often coming up with good ones is more an act of creative thinking than of straightforward science. No rules govern the process, and hence, no limitations exist either. That means, however, that one cannot guarantee that an unobtrusive measure will be accepted by reviewers or readers of one's work. Unobtrusive measures are often novel and have no history of use in a field by means of which the case for them can be buttressed. Moreover, many unobtrusive measures do not readily fit the requirements of conventional psychometric analyses so that the usual indicators of difficulty level, variance, and so on are not available, and reliability cannot be directly computed. In the end, unobtrusive measures must have face validity, i.e., readers must see immediately and intuitively that they make sense. That is not always likely.

Unobtrusive measures may sometimes raise troubling ethical questions, for they may be obtained under conditions that at least appear to violate usual expectations about informed consent, confidentiality, and so on. Records may have been assembled with no expectation that they would ever be used for anything other than their original purposes, and even if confidentiality is protected at the public level, some people might feel that their personal confidence is breached when researchers gain access to their records. People may feel that even public behavior is in some sense private if they have no expectation of being systematically observed. In fact, some courts have ruled that people riding in automobiles have expectations of privacy that should protect them against at least some types of observation. Ethical concerns may not rule

out many unobtrusive measures, but at least they require careful consideration by researchers.

Furthermore, unobtrusive measures are not always inexpensive. Observers, for example, are expensive to train and support in the field, and unless they can produce information that is markedly better than what can be derived from questionnaires, they may not be affordable. Archives may be readily available, but the cost of mining them for usable data may be quite high. Many other types of unobtrusive measures may turn out to be surprisingly expensive. Questionnaires and interviews may be performed by mail, phone, or web sites, and the expenses associated with them, especially if multiple sites are involved, can prohibit them from being favorite in the array of methods potentially available. As noted earlier, however, unobtrusive measures may still be useful at modest cost when used as probes into the quality and bias of more traditional measures.

See also: Archival Methods; External Validity; Field Observational Research in Anthropology and Sociology; Secondary Analysis: Methodology

Bibliography

Babbie E R 1989 *The Practice of Social Research*, 5th edn. Wadsworth Publishing, Belmont, CA

Barthel C N, Holmes D S 1968 High school yearbooks: A nonreactive measure of social isolation in graduates who later become schizophrenic. *Journal of Abnormal Psychology* **73**: 313–16

Bochner S 1979 Designing unobtrusive field experiments in social psychology. In: Sechrest L (ed.) *Unobtrusive Measurement Today*. Jossey-Bass, San Francisco, CA

Campbell D T 1969 Reforms as experiments. *American Psychologist* **24**: 409–29

Campbell D T, Stanley J C 1963 Experimental and quasi-experimental designs for research on teaching. In: Gage N L (ed.) *Handbook of Research on Teaching: A Project of the American Education Research Association*. Rand McNally, Chicago, IL

Freedman D A 1991 Statistical models and shoe leather. In: Marsden P V (ed.). *Sociological Methodology* **21**: 291–313

Freedman D, Pisani R, Purves R 1991 *Statistics*, 2nd edn. Norton, New York

Rathje W L 1979 Trace measures. In: Sechrest L (ed.) *Unobtrusive Measurement Today*. Jossey-Bass, San Francisco, CA

Sechrest L (ed.) 1979 *Unobtrusive Measurement Today*. Jossey-Bass, San Francisco, CA

Sechrest L, Phillips M 1979 Unobtrusive measures: An overview. In: Sechrest L (ed.) *Unobtrusive Measurement Today*. Jossey-Bass, San Francisco, CA

Webb E J, Campbell D, Schwartz R, Sechrest L 1966 *Unobtrusive Measures: Nonreactive Measures in the Social Sciences*. Rand McNally, Chicago, IL

Webb E J, Campbell D, Schwartz R, Sechrest L, Grove J 1981 *Non-reactive Measures in the Social Sciences*, 2nd edn. Houghton Mifflin, Boston, MA

L. Sechrest and R. J. Hill

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Ur

Urban Activity Patterns

1. Introduction

'Urban activity patterns' refer to spatial and temporal patterns of human activities in an urban area. Time and space are thus the most fundamental dimensions in the analysis of urban activity patterns. Studies have examined spatial movements of people and transitions of their activities over time. There are certain regularities in activity patterns due to physiological, cultural, institutional, and other reasons. There are, at the same time, substantial variations in activity patterns among individuals, or over time; while some activities are repeated with high regularities, some are pursued with irregular frequencies and intervals. Investigations into urban travel patterns have attempted to reveal their regularities and variations across individuals and over time.

Analyses of urban activity patterns may be traced back to Chapin (1974) and his colleagues, who attempted to explain differences in activity patterns using person and household characteristics, in particular, 'stage in the family life cycle.' Another important concept is Hägerstrand's space–time prism (Hägerstrand 1970). Jones et al. (1983) combined these two approaches and extended them to form an analytical framework, which incorporates both needs and desires that motivate engagement in activities, and constraints that govern activity engagement and travel.

Since engaging in activities implies using time in certain ways, activity patterns are synonymous with patterns of time use. Urban activity patterns have also been examined from time use perspectives. Studies have examined the allocation of time to different types of activities, activity duration, time-of-day characteristics of activity engagement, and sequencing of activities.

Additional elements are spatial aspects of activity patterns, which shall be termed 'travel patterns.' Travel patterns have been examined extensively in the field of transportation planning. Of particular interest is trip chaining, which refers to the linking of trips to visit a series of activity locations. (A trip refers to the movement from an activity location to the next activity location. In cases where movement itself is the activity (e.g., walking the dog), the entire movement is a trip.) Also of interest from both time use and travel perspectives are interpersonal linkages, day-to-day variations in activity patterns, and behavioral dynamics.

Characteristics of urban activity patterns are discussed below on each of the aforementioned subject areas. For other reviews of the field and additional references, see Damm (1983), Jones et al. (1983, 1990), Kitamura (1988), Axhausen and Gärling (1992), Hanson and Hanson (1993), Gärling et al. (1994), Pas (1997), Pas and Harvey (1997), and Bhat and Koppelman (1999).

2. Time Allocation

The total of 24 hours each individual has in a day is allocated to different types of activities. Some activities are mandatory due to prior social commitments (e.g., paid work), some are needed for physiological reasons (e.g., sleep) or for subsistence (e.g., grocery shopping), and some are discretionary (e.g., recreation). There are activities whose location, starting time, and ending time are predetermined, such as paid work. These activities constitute 'pegs' around which other activities are arranged.

The amounts of time allocated to sleeping, meals, or personal care are surprisingly similar across data from different regions and times (e.g., see Szalai 1972). Once the population is broken down into segments, data reveal meaningful differences that reflect the characteristics of respective subgroups and regions. For example, American women spend more time on paid work than Dutch women, simply because more American women are employed and more are full-time employees among those who are employed; working American women also spend more time on personal care than their Dutch counterparts.

Time allocation has been studied with theoretical microeconomic models and more descriptive structural equation models (SEMs). A microeconomic theoretic basis can be found in Becker (1965) on which many empirical studies have been based. Because the total amount of time available is fixed, the amounts of time allocated to different types of activities tend to be negatively correlated with each other. For example, as one might expect, those who work more hours spend less time on out-of-home maintenance activities or in-home recreation. A common finding of previous studies indicates that individuals with higher income tend to spend more time on out-of-home recreational activities. Effects of life cycle are evident; individuals

living alone and couples without children are more likely to pursue out-of-home recreational activities; single parents and those in extended families are least likely to do so.

3. Activity Duration

The time allocated to a type of activity may be divided into segments. Unfortunately, there seems to be no consensus on how such a segment should be defined. One candidate is the activity 'episode,' which shall be defined as a block of time during which none of the dimensions of activities (what is being done, where, with whom, for whom, etc.) change. Another construct is a 'project,' which shall be defined as a contiguous block of time in which a sequence of actions is taken to fulfill a purpose. Yet it is typical that the individual engages in multiple activities simultaneously, but not necessarily continuously, that serve a multitude of purposes, making the definition of projects difficult, or arbitrary, to a large extent.

Some projects have fixed durations while others have variable durations that are affected by a number of factors. For example, the project duration at an out-of-home location is positively correlated with the length of the trip to the location. Commute duration affects people's use of time; people with commutes longer than 60 minutes tend to spend most of their evenings indoors. Reductions in commute time induce more travel, i.e., a portion of travel time savings is invested in more travel. Consistent with the results of time allocation studies, older individuals and individuals with lower incomes tend to have longer home-stay durations.

4. Time of Day Characteristics of Activity Engagement

Many activities repeated daily have regular starting and ending times, for physiological and institutional reasons. Sleeping, eating meals, and work activities are repeated highly regularly. Californian and Dutch time-use datasets indicate that personal care has a peak in the morning, then distributed throughout the day, with another peak in the late evening. Activities such as childcare, on the other hand, have no peak. Running errands is distributed over a period between 8:00 a.m. and 10:00 p.m. with no clear peak. Social activities are distributed over a similar period, with higher intensities in the evening. Engagement in entertainment, reading, and TV viewing tends to increase toward the end of the day and peaks in the evening. These time-of-day characteristics of activity engagement have been examined with the concept of time-varying utilities of daily activities, and time-of-day effects have been quantified. Past studies have also found that characteristics of activity location choices differ by time of day.

5. Activity Sequencing

A set of multiple activities tends to be sequenced in certain ways by individuals. Those activities that are more mandatory appear to be pursued first as people tend to pursue out-of-home activities in a day in the following order: serve-passenger activities, personal business, shopping, then social/recreational. Similar tendencies are found within a trip chain. These sequencing patterns are correlated with the time-of-day characteristics of activity engagement discussed above.

Activities are engaged under a set of constraints. Most important constraints are work schedules, which typically define three space–time prisms for a worker. The intensity of constraints varies across individuals. For example, the concentration in time of commute trips is interpreted in a study as an indicator of how constrained trip making is, and it is noted that unmarried women with children have the least flexible, and singles have the most flexible, work trips. Several studies have attempted to explicitly incorporate constraints into models of activity engagement.

6. Activity and Travel

A trip is generated each time the activity location changes. Although the spatial and temporal characteristics of trips have been extensively studied, the mechanism of trip generation is not well understood. A theoretical model of shopping trip frequency has been developed by applying the inventory theory. This theoretical framework, however, is not applicable to all types of trips with nonwork purposes. The concept of time-dependent utility has been adopted to show when a trip will take place, and later some empirical evidence was offered in support of the model. These are among a very few examples in which the mechanism of trip generation is directly addressed.

From microeconomic viewpoints, more travel is done—by making more trips, traveling longer distances, or both—when the cost of travel is less. Then individuals residing in areas with higher accessibility to opportunities should be traveling more. Empirical evidence is not conclusive on this point, however. For example, a study found the 'degree of urbanization' is not associated with travel; another study reports that household location classification variables based on density measures are not significantly associated with the number of stops by nonworkers. On the other hand, very significant effects of GIS-based accessibility measures are reported in a study.

Attempts have been made to treat daily travel behavior in its entirety. Approaches taken in these studies include: applications of discrete choice models and applications of mathematical programming concepts. The former treats daily travel behavior as the choice of a travel pattern from among all possible alternative patterns. With the spatial and temporal dimensions, the number of alternative patterns can be astronomical, whose enumeration may impose computational difficulties. In addition, it is not compelling that the tasks of enumerating and evaluating numerous alternatives are manageable with human cognitive capacity. Likewise, treating the decision process underlying daily behavior as a mathematical programming problem does not appear to offer realism either. Furthermore, if viewed as an optimization problem, the decision process deals with an extremely complex problem of formulating and finding a solution for, leading to prohibitive computational requirements. An alternative approach is to decompose the decision associated with a daily travel pattern into a series of sequential decisions. Sequential approaches offer a better representation of cognitive processes underlying scheduling behavior. Yet they may not adequately depict individuals' responses to changes in the travel environment.

7. *Typologies of Activity–Travel Patterns*

Yet another approach to daily travel patterns is to develop typologies of travel patterns. An early work proposed a seven-group classification of travel patterns, obtained by applying a grouping procedure to factor scores that comprise measures of travel patterns, e.g., number of trips and total travel time. A later study developed 'geometrical similarity indices' based on a hierarchy of activity pattern attributes, and proposed travel pattern typologies that include five patterns of trip chaining. Application of pattern recognition theory and a variety of similarity measures can be found in the literature.

8. *Urban Travel Pattern Characteristics*

There is a substantial body of literature on urban travel patterns. A majority of individuals make 'simple' daily patterns of visiting one out-of-home location with two trips, i.e., a trip from home to the location, then a trip from the location back home. According to the 1990 National Personal Transportation Survey (NPTS) data, 30.4 percent of 48,385 respondents were reported to have made just two trips on the survey day, 8.6 percent three trips, 15.3 percent four trips, and 23.1 percent more than four trips. On the other hand, 21.2 percent of these respondents reported no trips at all. The average number of trips per person per day is 3.09; the average for those who made at least one trip is 3.92.

Gender roles and family life cycle are strongly associated with travel patterns. A number of studies have found that women engage in shopping more than men. An analysis of a 1990 dataset from the San Francisco Bay area indicates that women make about 75 percent more grocery trips than men, regardless of race/ethnicity, do more child chauffeuring, and make more household-serving trips.

The organization of nonwork travel, especially linking work and nonwork trips, varies with household structure. For example, singles tend to be away from home; two-person households have a positive propensity to conduct activities out of the home; and households with children tend to have more nonwork activity time, more nonwork trip chains, and more travel time to nonwork activities.

Household auto ownership is another factor that is strongly associated with household members' travel patterns. It is associated not only with the choice of the means of travel (or 'travel mode') but also with the number of trips itself. Study after study has shown a negative association between residential density and auto ownership, while household auto ownership is the principal explanatory variable of household trip generation that has traditionally been used in transportation planning studies.

9. *Trip Chaining*

Trip chaining is another subject area that has been extensively studied (e.g., see Hanson 1979, Golob and Golob 1981). Early models of trip chaining are Markovian that attempt to replicate linkages between activities pursued at successive out-of-home locations. This was later extended to include a continuous time dimension, and the stationarity and history dependence of activity transitions were statistically examined.

Many studies have indicated that individuals with higher incomes tend to chain trips more frequently and make complex trip chains. There are, however, results that income is not associated with forming complex commute chains, or indications that households with higher incomes have smaller propensities to chain trips.

Females appear to be more active than males, with more trips, more trip chains, and more trips per trip chain. On household structure and trip chaining, it has been observed that households of young working adults without children tend to have chains involving work trips; households with preschool children tend to make simple one-stop chains and simple commute trips; households with school-age children are inclined toward complex chains; while elderly households tend to have simple chaining patterns. Life cycle is also associated with the generation of serve-passenger trips,

and making more serve-passenger trips implies more trip chains.

A study of the 1983 NTPS data indicates that nonwork travel accounted for just over half of all person trips in the a.m. peak and two-thirds in the p.m. peak, and nonwork trips grew considerably faster than work trips between 1977 and 1983. This may be due to the linking of nonwork travel to commute trips. Importance of commute trips as a base to form trip chains has been noted by several researchers. Reductions in household size coupled with increases of multiworker households result in a greater tendency of linking nonwork trips to work trips.

10. Interpersonal Linkages

Most analyses of urban activity and travel patterns are concerned with the behavior of each individual. Although the focus on life cycle stage is in part based on the understanding that interactions among household members affect each member's activity and travel, characteristics of interpersonal interactions have been relatively unexplored. Subjects for investigation include how resources are allocated to household members, how tasks are assigned or performed on somebody else's behalf, and how activities are pursued jointly. Also of interest is how interpersonal interactions motivate the engagement in an activity, either jointly or separately. Such interactions are expected to take place through various social networks.

Inter-relationships among household members may be classified into substitutable relationships, complementary relationships, and companionship relationships. A study indicates that relations between household members are mostly complementary. Evening space–time paths of male–female couples have been examined and yielded findings that those with preschool or school-age children have higher propensities toward evening out-of-home activities, and young couples without children are oriented toward joint activities, often after meeting each other outside the home. Studies have shown that male travel demand is sensitive to female work and maintenance activities.

11. Day-to-day Variability

Another limitation of most of the studies is that they are concerned with activity and travel patterns for just one day, mostly a weekday. This reflects the fact that most datasets are based on survey periods of just one day. Such data contain patterns on a randomly sampled day, carrying no information on how typical or unusual that day might have been.

Examining day-to-day variability is important for several reasons. First, without information on day-to-day variability, it is impossible to determine exactly how much of the interperson variability in travel patterns is genuine, and how much is the artifact of intraperson, day-to-day variability. Likewise, it is not possible to detect how much behavior has changed over time. Second, statistical analysis of one-day data may not yield a correct depiction of behavioral change if it is history dependent. Third, one-day data do not indicate how many individuals exhibit certain behaviors over a span of time. For example, suppose one-day data indicate that 15 percent of commuters travel by public transit. This may mean a specific group of commuters, which comprises 15 percent of all commuters, travel by public transit every day, or every commuter commutes by public transit 15 percent of the time.

The level of repetition and variability in daily travel patterns has been studied using 35-day travel diary records, yielding findings that weekday and weekend patterns are different, and each individual has more than one typical weekday pattern. It is further shown that the level of repetition varies from population subgroup to subgroup. It has been reported that role-related constraints and commitments are associated with the variability in the individual's travel patterns. A study postulates a two-stage process: weekly behavior is selected first, then daily behavior is selected. Employed people with low incomes tend to select 'simple' weekly travel activity patterns. In another study, a two-stage model system is adopted in which a 'latent' daily pattern is selected, then travel time expenditure is determined.

12. Behavioral Dynamics

Most studies of urban activity patterns assume that patterns are stable or in equilibrium. Underlying this is the assumption that individuals adjust their activity patterns immediately and precisely whenever a change takes place in their travel environments or in their preferences (Goodwin et al. 1990). This of course is a very strong assumption. In fact, responses to changes often involve time 'lags.' Learning, or trial-and-error search for adaptation schemes, takes time, also producing lags. On the other hand, behavioral responses may precede changes in the environment because of individuals' planning actions, resulting in 'leads' in responses.

It is often the case that there are no responses to changes because of habit and behavioral inertia. Some of the issues addressed in day-to-day variability pertain to behavioral dynamics. The accumulation of empirical results is still in progress in this subject area, and it is premature to assess common dynamic characteristics of activity and travel behavior at this time. It is noted that there are pieces of empirical evidence which demonstrate the presence of behavioral intertia, show discrepancies between longitudinal elasticities and cross-sectionally evaluated elasticities,

and demonstrate dynamic properties in many aspects of activity and travel behavior.

See also: Community, Social Contexts of; Community Sociology; Neighborhood: General; Spatial Analysis in Geography; Spatial Interaction; Spatial Interaction Models; Time-geography; Transportation Geography; Transportation Planning; Urban Anthropology; Urban Sociology; Urban System in Geography

Bibliography

Axhausen K W, Gärling T 1992 Activity-based approaches to travel analysis: Conceptual frameworks, models, and research problems. *Transport Reviews* **12**(4): 323–41

Becker G S 1965 A theory of the allocation of time. *Economic Journal* **75**: 324–41

Bhat C R, Koppelman F S 1999 A retrospective and prospective survey of time-use research. *Transportation* **26**(2): 119–39

Chapin F S 1974 *Human Activity Patterns in the City: Things People Do in Time and Space*. Wiley, New York

Damm D 1983 Theory and empirical results: A comparison of recent activity-based research. In: Carpenter S, Jones P M (eds.) *Recent Advances in Travel Demand*. Gower Publishing, Aldershot, UK

Gärling T, Kwan M-P, Golledge R G 1994 Computational-process modelling of household activity scheduling. *Transportation Research B* **28B**(5): 355–64

Golob J, Golob T F 1981 Classification of approaches to travel-behaviour analysis. In: *Travel Analysis Methods for the 1980's*, Special Report 201, Transportation Research Board, Washington, DC

Goodwin P, Kitamura R, Meurs H 1990 Some principles of dynamic analysis of travel behaviour. In: Jones P (ed.) *Developments in Dynamic and Activity-Based Approaches to Travel Analysis*. Avebury, Aldershot, UK, pp. 56–72

Hägerstrand T 1970 What about people in regional science? *Papers of the Regional Science Association* **24**: 7–21

Hanson S 1979 Urban travel linkages: A review. In: Hensher D A, Stopher P (eds.) *Behavioural Travel Modelling*. Croom Helm, London

Hanson S, Hanson P 1993 The geography of everyday life. In: Golledge R, Garling T (eds.) *Behavior and Environment: Psychological and Geographical Approaches*. Elsevier Science, Amsterdam, pp. 249–69

Jones P, Koppelman F, Orfueil J P 1990 Activity analysis: State-of-the-art and future directions. In: Jones P (ed.) *Developments in Dynamic and Activity-Based Approaches to Travel Analysis*. Gower Publishing, Aldershot, UK

Jones P M, Dix M C, Clarke M I, Heggie I G 1983 *Understanding Travel Behaviour*. Gower Publishing, Aldershot, UK

Kitamura R 1988 An evaluation of activity-based travel analysis. *Transportation* **15**: 9–34

Pas E I 1997 Recent advances in activity-based travel demand modeling. In: Texas Transportation Institute (ed.) *Activity-Based Travel Forecasting Conference, June 2–5, 1996, Summary, Recommendations and Compendium of Papers*, Travel Model Improvement Program, US Department of Transportation, Washington, DC, pp. 79–102

Pas E I, Harvey A S 1997 Time use research and travel demand analysis and modelling. In: Stopher P, Lee-Gosselin M (eds.) *Understanding Travel Behaviour in an Era of Change*. Pergamon, Oxford, UK, pp. 315–38

Szalai S (ed.) 1973 *The Use of Time*. Mouton, The Hague

R. Kitamura

Urban Anthropology

Urban anthropology attends to the social relations, symbols, and political economies that are most manifest in the city (Lynch 1994, Zenner 1994). In earlier overviews, Fox (1972, 1977), Jackson (1985), and Gulick (1989) advocate an anthropology *of* the city, rather than *in* the city, and argue that this distinction 'is not trivial or hairsplitting' (Gulick 1989, p. 14). Both approaches make important contributions: *in* studies focus on urban ethnography and foreground people, while *of* studies emphasize peoples' relationships to political and economic forces and the social and physical structure of the city. Urban anthropology as a subdiscipline combines the strengths of participant observation and intensive research with small groups with holism and political economy within a comparative framework.

Theorizing the city is a necessary part of understanding the changing postindustrial, advanced capitalist, postmodern world in which we live. The city as a site of everyday practice provides valuable insights into the linkages of these macroprocesses with the texture and fabric of human experience. The city is not the only place where these linkages can be studied, but the intensification of these processes—as well as their human outcomes—occurs and can be understood best in cities. Thus, the city is not a reification, but the focus of study of cultural and sociopolitical manifestations of urban lives and everyday practices.

1. Historical Overview

The historical trajectory begins with the Chicago School in the 1920s and 1930s and the development of an urban ecological perspective (Park and Burgess 1974). The city is theorized as made up of adjacent ecological niches occupied by human groups in a series of concentric rings surrounding the central core. Class, occupation, world view, and life experiences are coterminous with an inhabitant's location within this human ecology. Social change occurs through socioeconomic transitions of these areas in an ever downward spiral toward the inner city. Research strategies focus on participant observation as a method of uncovering and explaining the adaptations and accommodations of urban populations to these microenvironments. Urban ethnographies suggest that familiarity, avoidance, and surveillance play important roles in allaying fears that emerge from heterogeneity in the city. Sally Merry (1982) documents the interactions and perceptions of what she terms

'Black,' 'White,' and 'Chinese' residents in a high-rise, low-income housing project in a large Midwestern city, and concludes that lack of familiarity plays an important role in the perception of danger. Eli Anderson (1990) documents avoidance as a coping strategy in his study of 'streetwise' behavior of Philadelphians, in which residents cross the street when faced with oncoming young 'black' males. Wacquant (1994) portrays how institutional avoidance of young 'black' males can create a 'hyperghetto' of isolated families in Chicago's Black Belt, where the streets are deserted and no longer patroled by police; and Philippe Bourgois (1995) dramatizes the fear and sense of vulnerability experienced by residents of El Barrio and depicts their strategies of avoidance and surveillance used to deal with street crime. These studies describe how fear is spatially managed in urban contexts, and how avoidance and streetwise behavior are used by low- to middle-income people to mitigate their fears, while abandonment of the inner city by police and local institutions exacerbates these residents' feelings of fear and isolation.

A second major influence is a series of community studies undertaken as part of the Institute of Community Studies' program of policy and planning research on the slum clearance and replacement of housing in London, England and Lagos, Nigeria. These studies, beginning in the 1950s and continuing throughout the second half of the twentieth century (Young and Willmott 1957, Marris 1962, 1995), theorize the city as made up of a series of urban communities, based on extended family relations and kinship networks. Coincidentally, the Tavistock Institute of Human Relations published Bott's (1957) study of the social networks of middle class English families that drew upon discussions with anthropologists at the University of Manchester (Gluckman 1971). The methodological contribution of network analysis as the basis for studying the social organization of city residents is widely used to understand the rapidly urbanizing populations of Africa (Mitchell 1969) and Latin America (Lomnitz 1977), as well as by North American researchers interested in the interconnections and interdependencies of family and household relationships among the urban poor (Stack 1974, 1996). Network studies still provide an important methodological strategy and theoretical model for urban researchers (Laguerre 1994, Kadushin and Jones 1992, Liebow 1989).

Studies of planned physical and social change in Latin American low-income residential neighborhoods (Lobo 1983, Logan 1984), as well as studies of the planning and design of new towns such as Ciudad Guyana (Peattie 1972) and Brasilia (Epstein 1973) provide further ethnographic examples of local as well as national/international conflict over planning goals. These studies identify foreign capital investment (Peattie 1987) and the power/knowledge of the technologies of planning and architecture (Rabinow 1989) as antithetical to producing a humane environment for local populations and workers. Studies of urban renewal (Greenbaum 1992) and community rebuilding after natural disasters (Oliver-Smith 1986) further contribute to understanding how the dynamics of redevelopment processes often exclude the psychosocial needs of residents. Anthropological analyses of the conflicts that emerge among government institutions, planning experts, and local communities set the stage for contemporary poststructuralist studies of urban struggle for land tenure rights (Holston 1991) and adequate housing (Low 1988), as well as for studies of planning and architecture as instruments of social control (McDonogh 1991, Plotnicov 1987).

Another theoretical force has been the cumulative work of Leeds (1973) recently published in a posthumous volume (Sanjek 1994). Although Leeds' work concentrated on supralocal and local linkages and the nation/state level of analysis, the majority of his fieldwork dealt with the city as the point of articulation of these complex relationships. Anthropologists continue to utilize Leeds' theoretical model of the flow of goods, cash, labor, and services between metropolis and countryside in their analyses of the city (Guildin 1992).

The most important theoretical transition, however, occurred in the 1980s with the introduction of the study of the political economy of the city. Susser's (1982) landmark ethnography of a Brooklyn working class neighborhood, Hannerz's theoretical monographs (1980, 1992), and Mullings' (1987) critique of the study of cities in the USA ushered in a decade of critical studies of the structural forces that shape urban experience. The social organizational paradigm that dominated earlier studies was superseded by a political economy paradigm (Sanjek 1990). These studies theorize the city by examining the social effects of industrial capitalism and deconstructing the confusion of urbanism with inequality and alienation (Mullings 1987, Ong 1987).

The final development in this theoretical trajectory is what Jacobs (1993) has called representational cities—an approach in which messages encoded in the environment are read as texts. Jacobs argues that 'ethnographic studies were commonly prescribed the role of rendering more real the exotic and marginalized, but were seen to have little value in terms of the modern project of theory-building' (1993, p. 828). Radicalized urban ethnography, however, makes possible a link between everyday practices and the broader processes of class formation. According to Jacobs, new cities require new forms of analysis in which the urban built environment becomes a discursive realm. Within anthropology this representational approach is reflected in Holston's (1991) analysis of the planning and architecture of Brasilia, in which the city is read as an ideological tract of an imagined socialist utopia, and in Dorst's (1989) postmodern ethnography of the recreation of the history and landscape of Chadd's Ford.

The correspondence of academic training with the geographical area of study has also contributed to continuities of research and theory within culture regions. The tradition of British social anthropology in Africa has created a history of studies that focus on social relations in the city—exchanges, political alliances, market relationships, and network analyses that form the core of contemporary theoretical work (Barnes 1986, MacGaffey 1987, Peil 1991). Other continuities include studies of *favelas*, shantytowns, and *turgurios* in the urban periphery, and the informal economy in Latin America (Safa 1986); Japanese studies that focus on work organization (Hamabata 1990, Kondo 1990, Sumihara 1993, Bestor in press); and Chinese studies that emphasize urban hierarchies (Guldin 1993, Jankowiak 1993).

Thus, the historical development of the anthropological study of the city has produced a number of theoretical approaches that continue to be drawn upon by urban anthropologists. These approaches include: urban ecology models; community, family, and network analyses; studies of the power/knowledge of planning and architecture; supralocal/local linkage analyses; and political economic, representational, and discursive models of the city.

2. *Theorizing the City: Space, Knowledge, Time, and Aesthetics*

The view that contemporary cities not only pose problems that are intrinsic to the metropolitan experience, but also underscore and transform many of the most traditional concerns of the discipline—the social organization of space, the meanings of knowledge, group, and power, and the intricacies of commodity, exchange, and political economy—has been taken up by a group of scholars involved in the theoretical revitalization of the field (Low 1999). The concern is to articulate understandings of particular cities—Vienna, Barcelona, Valencia, Savannah, Atlanta, New York, Toronto, Lagos, Shanghai, Tokyo, Belize City, and San José (Costa Rica)—with broader anthropological concepts of space, knowledge, time, and identity. The inquiry begins not with imagining the city as a metaphorical object, but with imagining urbanites—residents, homeless people, planners, municipal bureaucrats, and architects—experiencing the city through the social relations, political economic, and planning processes.

Rotenberg and McDonogh's influential collection *The Cultural Meaning of Urban Space* reflects an attempt to understand 'the meaning of urban spaces through the knowledge of the people who live within them' (1993, p. 197, 11). The studies are characterized by the search for the underlying social and cultural values and power politics that give form and meaning to the cityscape and the urban built environment. McDonogh's (1993) exploration of 'emptiness' provides an evocative category that marks not an absence of urbanness, but a zone of intense competition that betrays the imposition of urban power. Low (1993, 1995) focuses on the historical emergence of spatial meanings of power relations in the Spanish American plaza, while others are concerned with the symbolic mapping of contested arenas of urban social interaction such as privacy (Pellow 1993), neighborhood (Bestor 1993), and schooling (Rutheiser 1993). Research on the design of housing (Cooper and Rodman 1992, Low and Chambers 1989) and place attachment to urban space (Altman and Low 1992) also contribute to this ongoing venture.

Pellow's work *Setting Boundaries* is more concerned with boundaries and perimeters, that is, the way in which 'physical and conceptual boundaries are integrally tied to the creation, maintenance, transformation and definition of social and societal relations—of sociocultural behavior and action' (Pellow 1996, p. 179, 3). The concept of boundary is dealt with both spatially and metaphorically, creating a link between the materialist and metaphorical analyses of social and spatial demarcation.

Many of the ideas have been elaborated by asking how meaning is created through both the social production of space and the social construction of space (Low 1996), and how power is represented in the history and evolution of the built forms (Leone 1995, Lawrence 1992). An ethnographic approach to the study of urban space includes four areas of spatial/cultural analysis—historical emergence, sociopolitical and economic structuring, patterns of social use, and experiential meanings—as a means of working out this theoretical agenda.

Rotenberg proposes a second project, that of identifying forms of metropolitan knowledge as a subset of the knowledge people gain from their lived experience and value socialization. City dwellers share the knowledge because they live in dense and specialized concentrations of people, information, built form, and economic activity. Rotenberg (1993) speaks of the 'salubrity of sites' as a way of understanding how metropolitan knowledge is made manifest on the urban landscape. In *Landscape and Power in Metropolitan Vienna* (1995), he traces the history of open spaces and gardens in Vienna and documents how these spaces have become a spatial template of urban symbolic communication.

An additional project, the study of time, directs urban researchers to consider the way in which schedules coordinate the circulation of people in the city as a means for studying ebedded power relations. Rutz argues that 'a *politics* of time is concerned with the *appropriation* of the time of others, the *institutionalization* of a dominant time, and the *legitimation* of power by means of the control of time' (1992, p. 7). Rotenberg's studies (1992) demonstrate the tyranny of urban schedules over individuals subject to their control. Issues of both time and space

emerge in the study of homeless people who live in urban public spaces (Hopper 1991).

The city also structures residents' urban experience, adding urban identity to place and time as universal sources of metropolitan knowledge (Rotenberg 1996). This proposal resonates with Sennett's (1994) interest in embodied urban experience. These universal characterizations bring us temptingly close to earlier essentializing discourses, but at the same time provide provocative material for perceiving how the city as a set of processes links experience and structure.

Another project that holds great promise is the study of ethnoaesthetics. Although much of this literature focuses on indigenous media and aesthetic and political sensibilities, the implications of this work transform our notions of the urban, global, transnational, and marginal. In the city, where culture-making often takes place, 'performing aboriginality' takes on new aesthetic and identity meanings (Ginsburg 1991). Bright and Bakewell (1995), in *Looking High and Low*, reposition art and ethnoaesthetics to redefine notions of cultural identity. Urban murals and low-rider cars provide examples of how the city constitutes an important dimension of the aesthetic. Although relatively undeveloped as part of the urban discourse, the study of ethnoaesthetics and cultural identity, and the demystification of art and artistic creation provide important insights for the analysis of the culture of cities (King 1996, Zukin 1995).

Contemporary anthropological studies of the city focus predominantly on the center, producing ethnographies of culturally significant places—such as markets, housing projects, gardens, plazas, convention centers, waterfront developments, and homeless shelters—that articulate macro and micro urban processes. These studies illuminate both the material and metaphorical power of spatial analysis for theorizing the city. Studies of the relationship of the urban to globalization, economic restructuring, and transnational migration have transformed traditional research concerns. The Society for Urban Anthropology responded to this transformation by changing its name in 1998 to the Society for Urban, National, and Transnational/Global Anthropology (SUNTA). As a subsection of the American Anthropological Association with a flourishing journal, *City and Society*, SUNTA provides a forum for the exchange of ideas within the subdiscipline and across the urban social studies.

3. Conclusion

The anthropology of the city incorporates a number of paradigms from other disciplines. The influence of political economy, architectural and planning theory, cultural studies, urban sociology, and cultural geography can be seen in the increasing attention to economic, political, and discursive models of the city. At the same time poststructural and postmodern perspectives have recast the kinds of questions and modes of inquiry used to study the city. The dominant research trends in urban anthropology are currently poststructural studies of race, class, and gender in the urban context; political economic studies of transnational culture; and studies of the symbolic and social production of urban space and planning.

Some areas of anthropological theory are more influential within the broader discourse of urban studies and urban policy. Many anthropologists contribute actively to theory and research on urban poverty, racism, globalization, and architecture and planning. (Stack 1996, Bourgois 1995, Susser 1991, Newman 1993.) The anthropological take on globalization is to focus attention on the transnational aspects of migration, culture-making, and identity management, and on the shifting cultural environments and meanings that contextualize and decontextualize behavior (Smart and Smart 1991, Pellow 1991). Urban anthropologists offer an experience-near critique of city life that provides a more complex understanding of the differences between cities' and residents' responses to racial segregation and class inequality (Gregory 1992, Susser 1991). Anthropological critiques of planning and design projects provide a methodology and theoretical framework for decoding the ideological intentions and material consequences of architectural plans and landscape designs (Holston 1989, Rabinow 1989, Rotenberg 1995, Rutheiser 1996). And anthropological studies link the macroprocesses of cultural hegemony with the microprocesses of resistance and social transformation within the context of deindustrialization and globalization demonstrating that these linkages can be made through a combination of sophisticated theorizing and empirical fieldwork. The contributions of anthropological fieldwork still retain the power to demonstrate the how, why, and when of urban processes, but are even more effective when linked to theoretical frameworks that provide a grounding for further study and discussion.

Bibliography

Altman I, Low S M 1992 *Place Attachment*. Plenum, New York

Anderson E 1990 *Streetwise: Race, Class and Change in an Urban Community*. Chicago University Press, Chicago

Barnes S 1986 *Patrons and Power: Creating a Political Community in Metropolitan Lagos*. Indiana University Press, Bloomington, IN

Bestor T C 1993 Rediscovering shitamachi: Subculture, class and Tokyo's 'traditional' urbanism. In: Rotenberg R, McDonogh G W (eds.) *The Cultural Meaning of Urban Space*. Bergin and Garvey, Westport, CT, pp. 47–60

Bestor T C in press *Tokyo's Marketplace: Culture and Trade in the Tsukji Wholesale Fish Market*. Stanford University Press, Stanford, CA

Bott E 1957 *Family and Social Network: Roles, Norms and External Relationships in Ordinary Urban Families*. Tavistock, London

Bourgois P 1995 *In Search of Respect: Selling Crack in El Barrio*. Cambridge University Press, Cambridge, UK

Bright B J, Bakewell Z (eds.) 1995 *Looking High and Low: Art and Cultural Identity*. University of Arizona Press, Tucson, AZ

Cooper M, Rodman M C 1992 *New Neighbours: A Case Study of Cooperative Housing*. University Toronto Press, Toronto, ON

Dorst J D 1989 *The Written Suburb: An American Site; An Ethnographic Dilemma*. University of Pennsylvania Press, Philadelphia, PA

Epstein D 1973 *Brasília, Plan and Reality: A Study of Planned and Spontaneous Urban Development*. University of California Press, Berkeley, CA

Fox R 1972 Rational and romance in urban anthropology. *Urban Anthropolology* **1**: 205–33

Fox R 1977 *Urban Anthropology: Cities in their Cultural Settings*. Prentice Hall, Englewood Cliffs, NJ

Ginsburg F 1991 Indigenous media: Faustian bargain or global village? *Cultural Anthropology* **6**: 95–114

Gluckman M 1971 Preface. In: Bott E (ed.) *Family and Social Network* 2nd edn., New Press, New York, pp. 13–30

Greenbaum S 1992 Housing abandonment in the inner-city black neighborhoods: A case study of the effect of the dual housing market. In: Rotenberg R, McDonogh G W (eds.) *The Cultural Meaning of Urban Space*. Bergin and Garvey, Westport, CT, pp. 127–49

Gregory S 1992 The changing significance of race and class in an African-American community. *American Ethnologist* **19**: 255–74

Guildin G E 1992 *Urbanizing China*. Greenwood Press, New York

Guildin G E, Southall A 1993 *Urban Anthropology in China*. EJ Brill, Leiden, The Netherlands

Gulick J 1989 *The Humanity of Cities: An Introduction to Urban Societies*. Bergen and Garvey, Granby, MA

Hamabata M M 1990 *Crested Kimono: Power and Love in the Japanese Business Family*. Cornell University Press, Ithaca, NY

Hannerz U 1980 *Exploring the City*. Columbia University Press, New York

Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York

Holston J 1989 *The Modernist City: An Anthropological Critique of Brasília*. Chicago University Press, Chicago

Holston J 1991 The misrule of law: Land and usurpation in Brazil. *Comparative Studies in Society and History* **33**: 695–725

Hopper K 1991 Symptoms, survivial and the redefinition of public space. *Urban Anthropology* **20**: 155–75

Jackson P 1985 Urban ethnography. *Progress in Human Geography* **10**: 157–76

Jacobs J 1993 The city unbound: Qualitative approaches to the city. *Urban Studies* **30**: 827–48

Jankowiak W 1993 *Sex, Death, and Hierarchy in a Chinese City*. Columbia University Press, New York

Kadushin C, Jones D 1992 Social networks and urban neighborhoods in New York City. *City and Society* **6**: 58–75

King A D 1996 *Re-presenting the City*. Macmillan, London

Kondo D 1990 *Crafting Selves: Power, Gender, and Discourse*. Chicago University Press, Chicago

Laguerre M S 1994 *The Informal City*. St Martin's Press, New York

Lawrence D 1992 Transcendence of place: The role of La Placeta in Valencia's Las Fallas. In: Altman I, Low S M (eds.) *Place Attachment*. Plenum, New York, pp. 211–30

Leeds A 1973 Locality power in relation to supralocal power institutions. In: Southall A (ed.) *Urban Anthropology*. Oxford University Press, New York, pp. 15–41

Leone M P 1995 A historical archaeology of capitalism. *American Anthropologist* **97**: 251–68

Liebow E D 1989 Category or community? Measuring urban Indian social cohesion with network sampling. *Journal of Ethnic Studies* **16**: 76–100

Lobo S 1983 *A House of my Own: Social Organization in the Squatter Settlements of Lima, Peru*. University of Arizona Press, Tucson, AZ

Logan K 1984 *Haciendo Pueblo: The Development of a Guadalajaran Suburb*. University of Alabama Press, Birmingham, AL

Lomnitz L A 1977 *Networks and Marginality: Life in a Mexican Shantytown*. Academic Press, New York

Low S M 1988 Housing organization and social change: A comparison of programs for urban reconstruction in Guatemala City. *Human Organization* **47**: 15–24

Low S M 1993 Cultural meaning of the plaza: The history of the Spanish American gridplan-plaza urban design. In: Rotenberg R, McDonogh G W (eds.) *The Cultural Meaning of Urban Space*. Bergin and Garvey, Westport, CT, pp. 75–94

Low S M 1995 Indigenous architecture and the Spanish American Plaza in Mesoamerica and the Caribbean. *American Anthropologist* **97**: 748–62

Low S M 1996 Spatializing culture: The social production and social construction of public space. *American Ethnologist* **23**(4): 861–79

Low S M 1999 *Theorizing the City: The New Urban Anthropology Reader*. Rutgers University Press, New Brunswick, NB

Lynch O 1994 Urban anthropology, postmodernist cities, and perspectives. *City and Society Annual Review 1994*: 35–52

MacGaffey J 1987 *Entrepreneurs and Parasites: The Struggle for Indigenous Capitalism in Zaire*. Cambridge University Press, Cambridge, UK

Marris P 1962 *Family and Social Change in a African City: A Study of Rehousing in Lagos*. Northwestern University Press, Boston

Marris P 1995 Knowledge and persuasion: Research at ICS. In: Gench G, Flower T, Gavron K (eds.) *Young at Eighty: The Prolific Public Life of Michael Young*. Carcanet Press, Manchester, UK, pp. 75–83

McDonogh G W 1991 Discourses of the city: Policy and response in post-transitional Barcelona. *City and Society* **V**: 40–63

McDonogh G W 1993 The geography of emptiness. In: Rotenberg R, McDonogh G W (eds.) *The Cultural Meaning of Urban Space*. Bergin and Garvey, Westport, CT, pp. 3–16

Merry S 1982 *Urban Danger*. Temple University Press, Philadelphia, PA

Mitchell C 1969 *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns*. Manchester University Press, Manchester, UK

Mullings L 1987 *Cities of the United States: Studies in Urban Anthropology*. Columbia University Press, New York

Newman K S 1993 *Declining Fortunes: The Withering of the American Dream*. Basic Books, New York

Oliver-Smith A 1986 *The Martyred City: Death and Rebirth in the Andes*. University of New Mexico Press, Albuquerque, NM
Ong A 1987 *Spirits of Resistance and Capitalist Discipline*. SUNY Press, Albany, NY
Park R, Burgess E 1974 *The City*. University Chicago Press, Chicago
Peattie L R 1972 *The View from the Barrio*. University of Michigan Press, Ann Arbor, MI
Peattie L R 1987 *Planning: Rethinking Ciudad Guyana*. University of Michigan Press, Ann Arbor, MI
Peil M 1991 *Lagos: The City is the People*. GK Hall, Boston
Pellow D 1991 Chieftaincy and the evolution of an Accra *zongo*. *Ethnohistory* **38**: 414–50
Pellow D 1996 *Setting Boundaries: The Anthropology of Spatial and Social Organization*. Bergin and Garvey, Westport, CT
Plotnicov L 1987 The political economy of skyscrapers: An anthropological introduction to advanced industrial cities. *City and Society* **1**: 35–51
Rabinow P 1989 *French Modern: Norms and Forms of Missionary and Didactic Pathos*. MIT Press, Cambridge, MA
Rotenberg R 1992 *Time and Order in Metropolitan Vienna*. Smithsonian Institution Press, Washington, DC
Rotenberg R 1993 On the salubrity of sites. In: Rotenberg R, McDonogh G W (eds.) *The Cultural Meaning of Urban Space*. Bergin and Garvey, Westport, CT, pp. 17–30
Rotenberg R 1995 *Landscape and Power in Metropolitan Vienna*. Johns Hopkins Press, Baltimore
Rotenberg R, McDonogh G W 1993 *The Cultural Meaning of Urban Space*. Bergin and Garvey, Westport, CT
Rutheiser C 1993 Mapping contested terrains: Schoolrooms and streetcorners in urban Belize. In: Rotenberg R, McDonogh G W (eds.) *The Cultural Meaning of Urban Space*. Bergin and Garvey, Westport, CT, pp. 103–20
Rutheiser C 1996 *Imagineering Atlanta: Making Place in the Non-place Urban Realm*. Verso Press, New York
Rutz H 1992 *The Politics of Time*. American Anthropological Association, Washington, DC
Safa H I 1986 Urbanization, the informal economy, and state policy in Latin America. *Urban Anthropology* **15**: 135–63
Sanjek R 1990 Urban anthropology in the 1980s: A world view. *Annual Review of Anthropology* **19**: 1151–186
Sanjek R 1994 *Cities, Classes, and the Social Order: Anthony Leeds*. Cornell University Press, Ithaca, NY
Sennett R 1994 *Flesh and Stone: The Body and the City in Western Civilization*. W W Norton, New York
Smart J, Smart A 1991 Personal relations and divergent economies: A case study of Hong Kong investments in South China. *International Journal of Urban and Regional Research* **15**(2): 216–33
Stack C 1974 *All Our Kin: Strategies for Survival in a Black Community*. Harper and Row, New York
Stack C 1996 *Call to Home: African Americans Reclaim the Rural South*. Basic Books, New York
Sumihara N 1993 A case study of cross-cultural interaction in a Japanese multinational corporation in the United States. In: Sims R R, Denneby R F (eds.) *Diversity and Differences in Organizations*. Quorum Books, Westport, CT, pp. 136–47
Susser I 1982 *Norman Street*. Oxford University Press, New York
Susser I 1991 The separation of mothers and children. In: Mollenkopf, Castells, pp. 207–224
Wacquant L J D 1994 The new urban color line: The state and fate of the ghetto in postfordist America. In: Calhoun C (ed.) *Social Theory and the Politics of Identity*. Blackwell, Oxford, UK, pp. 231–76
Young M, Willmott P 1957 *Family and Kinship in East London*. Penguin, Middlesex, UK
Zenner W 1994 Nominalism and essentialism in urban anthropology. *City and Society Annual Review 1994*: 53–66
Zukin S 1995 *The Cultures of Cities*. Blackwell, Cambridge, UK

S. M. Low

Urban Ethnography

Urban ethnography is the close and systematic study of urban life and culture, relying both on first-hand observation and careful interviews with informants, and on available records. Its roots can be traced to the early British social anthropologists. A peculiarly American variant emerged at the beginning of the twentieth century, most notably through the fieldwork of Jane Addams, W. E. B. DuBois, and Robert E. Park, all of whom wrote in the interest of social reform.

Their concern was to inform the wider citizenry of the conditions of the urban poor as well as the nature of racial relations. Concerned particularly with the social challenges of industrialism and urbanization, Park and his students conducted seminal ethnographic work on the city, effectively establishing the premier American school of urban sociology in the early part of the twentieth century. The urban world of the twenty-first century presents new challenges to the ethnographer, who must now deal with the social impact of de- and reindustrialization, increased urbanization, more complex immigration patterns, and the local manifestations of such global economic and cultural processes, including structural poverty.

1. *The Chicago Tradition*

At the University of Chicago, Park and his students produced a series of important and detailed ethnographic case studies of the cultural patterns of the peoples of Chicago. Prominent among these were Anderson (1923), Wirth (1928), Zorbaugh (1929), Thrasher (1927), Shaw (1966), and Drake and Cayton (1945). These studies tended to focus on immigrants, the poor, racial relations and the various social problems of the day, providing a treasure trove of local knowledge about the city, particularly its neighborhoods, creating a mosaic of social scientific work, and establishing effectively the field of urban ethnography.

After World War II, a new generation of Chicago ethnographers emerged, most notably Everett C. Hughes, whose most prominent students included Howard S. Becker and Erving Goffman. Jointly, they

shaped not only the field of urban ethnography but also American sociology more generally. Important examples of urban ethnography also appeared from other settings, such as Boston (Whyte 1943, Gans 1962), Newburyport, Mass. (W. Lloyd Warner's Yankee Studies Series), and Muncie, Indiana (Lynd and Lynd 1929). But as time passed, these efforts were overshadowed by quantitative methods of sociology.

By the late 1960s and early 1970s, the Chicago School was being reinvigorated by Park's students' students, with Morris Janowitz, Gerald D. Suttles, and Howard S. Becker as prominent new teachers. Short and Strodtbeck's (1965) classic study of gangs in Chicago was followed shortly after by influential works on the urban black ghetto. Though not of Chicago, Liebow (1967) and Hannerz (1968) conducted path-breaking ethnographic analyses on the black ghettoes of Washington DC. And Rainwater (1968) added to this work with his impressive study of a failed housing project in St. Louis.

In the mid-1960s, Suttles took up residence in the 'Addams area' of Chicago for three years as a 'participant–observer.' He analyzed and described the social worlds of four major local ethnic groups—blacks, Italians, Mexicans, and Puerto Ricans—and the ways they shared the social spaces of an area undergoing significant 'urban renewal' at the hands of the local government. The groups sorted themselves out in an 'ordered segmentation' created among themselves in a kind of territorial ballet. Residents distinguished their own values and social rules by knowing by whom they were opposed, and thus conflict was kept at a minimum.

During the late 1960s, William Kornblum took a job in a steel mill in South Chicago for two years and involved himself in the social world of the mill employees. They accepted him and his family in ways that became a profound learning experience for him. Among his chief findings was the surprising degree of comity and goodwill in the workplace in spite of the ethnic competition, much of it achieved through political sharing, which provided a certain meaning to the lives of the workers. Contrary to widely held assumptions, the people were quite conservative politically. Getting to know the workers through Kornblum's rich ethnographic experience makes such political views understandable.

In the early 1970s, Elijah Anderson spent three years studying black street-corner men at a Southside Chicago bar and liquor store. He socialized with them closely, drinking, hanging out, visiting their homes and places of work, and he came to know them very well. Contrary to the view of those who are inclined to see this world as monolithic, there were in fact three groups of men at this place. They called themselves 'regulars,' 'wineheads,' and 'hoodlums,' the latter two being somewhat residual, and subject to labeling or name calling. The study sought to understand the ways in which these men came together on this street corner to make and remake their local stratification system.

Around this time, Ruth Horowitz moved into a Mexican neighborhood in Chicago, and over three years affiliated herself with a male street gang, the young women who often spent time with them, and upwardly mobile youth, learning about the issues facing such groups at first hand. Her work represented an early document in the sociology of gender, but she also found that as gang members went about their daily lives in both the community and the wider society, they would experience tensions and conflicts between their efforts to pursue the American dream, and their commitment to a code of honor that demanded actions with a high risk of compromising these efforts.

2. *Ethnographic Fieldwork*

Like the recent Chicago researchers presented above, urban ethnographers typically involve themselves in a social setting or community with the express purpose of learning about the people residing there. Of particular interest is how residents meet the exigencies of life, group themselves socially, and arrive at their shared understandings of the rules of everyday life—conventions, prescriptions, and proscriptions of life peculiar to their world. The answers to the researcher's questions about solving immediate problems of living reveal much about the social order, or what Clifford Geertz labels 'local knowledge.' In particular, key events and people's reactions to them can alert the ethnographer to the subtle expectations and norms of the subjects, and so to their culture.

In penetrating such local cultures, the ethnographers must not only engage in intensive fieldwork, cultivating subjects, and experiencing their social world, but also keep copious field notes—a journal of the lived experience. In developing questions and hypotheses about the nature of the local setting, ethnographers must also deal with their own world view: their 'own story' or set of working conceptions about their own world as well as the world of the subjects. Depending on how the ethnographer treats them, such presuppositions can be problematic or advantageous. The subjectivity inherent in the process of fieldwork is often considered to be a strength, for with it can come profound sensitivity to the core concerns of the people being studied.

In this connection, a useful distinction may be drawn between the 'participant–observer' and the 'observing participant.' The former may be in an early, tentative process of negotiating a relationship with the group under study, and may be satisfied with this position, while the latter has become close to the subjects, effectively empathizing with them, and, it is hoped, able to articulate their point of view. Both positions have their drawbacks and strengths, re-

quiring the ethnographer to remember constantly the primary goal: to provide a truthful rendition and analysis of the social and cultural world of the subjects. To see the world from their point of view requires learning their vocabulary, their concerns, and even their prejudices. It is from such a position that the ethnographer may be able to raise the most penetrating questions, questions that focus on the subjects' core issues of social organization. In this respect, the most effective questions blend both the 'problems' confronted by the subjects in their everyday lives, and the conceptual 'problem'—the answers to which would presumably advance the field theoretically.

The ethnographer's formal response to such questions, once formulated, can be considered a hypothesis, which in turn may serve as the tentative organizing principle for the ethnographic representation and analysis to follow. Here, the critical task is to advance the hypothesis toward a tenable proposition, or a plausible argument. The ethnographer's accumulated field notes will likely include either positive or negative cases, requiring revision of hypotheses to take the case into account. Through this style of analytic induction, the goal is always to develop an accurate account of the world of the subjects, while at times knowingly generating ever more penetrating questions. Such questions, by provocation and stimulation, trial and error, help to advance the ethnographer's case to surer ground. In this sense, the questions can be, and often are, more important than the 'answers.'

In the effort to apprehend, understand, and ultimately represent the social setting, the researcher becomes a kind of vessel, a virtual agent of the subjects themselves, serving as a communication link to the uninformed. Such a task is not accomplished easily. Not only does it require a certain amount of empathy in addition to impressive conceptual and observational skills, but the audience, including other social scientists and the 'lay public' to whom the setting is represented may have such strong presuppositions that no amount of evidence will be convincing. This is one of the inherent difficulties and challenges of doing and presenting worthwhile ethnographic work, particularly in socially, politically, or racially charged environments.

3. *The Challenge for Urban Ethnography in the Twenty-first Century*

In recent years, deindustrialization, reindustrialization, increased urbanization, immigration, and economic globalization have made urban areas increasingly complex, both geographically and ethnically. Boundaries, including national ones, are, at the start of the twenty-first century, less important as a barrier to the movement of people, goods, capital, and culture. Los Angeles, for instance, with its ethnic diversity, sprawl, and lack of a single center may be an anticipation of the shape of future cities. So may 'edge cities' (Garreau 1991), such as the Valley Forge–King of Prussia area northwest of Philadelphia, Pennsylvania—urban centers that develop near, but not within, existing cities. And New York, as a global hub with an increasingly international population, epitomizes the tensions between 'local' and 'global' ways of life.

To understand the new global immigration, for instance, ethnographers must now come to appreciate and learn more about the lives of the British Sikhs of California who travel back and forth between New York and extended families in India, as well as Bombay elites and Punjabi farmers arriving in the United States at the same time, and low- and high-caste Indians in Chicago sharing utter confusion toward suburbanites. Also important is the manner in which such 'new' people impact established ethnic and racial populations. The black street vendor's story is important, as are the stories of the New York Haitian taxi driver, who finds his own identity by actively distancing himself from the African American, and who visits periodically his cousins who reside in suburban Paris (see Duneier 2000). Of no less importance is the social situation of the Taiwanese middle-class immigrants to Philadelphia who assimilate to 'get along,' but who are strongly ambivalent about 'losing' their Chinese heritage (Tsai 1998).

Moreover, the connections between urban poverty and culture become more acute and ever more complicated in these new environments. Park, DuBois, Addams, and other pioneers addressed the effects of industrialization, urbanization, and immigration early in the twentieth century. Modern ethnographers must come to terms with both positive and negative human consequences of deindustrialization, and cybernation of industry in the context of the new global realities, noting particularly the implications for living standards in the local urban environment often beset by ethnic competition. These socioeconomic forces have brought about increasing structural poverty, in which many people are unable to develop the human and social capital necessary to rise from destitution (see Wilson 1987, Anderson 1990). The process of reindustrialization in the areas of light industry, cybernetics, and service must be studied, with its attendant issues of hard and soft skills. 'Brown racism' (Washington 1990) must also be addressed, with its sources and its implications for local urban life and culture. The social world of illegal immigrants from China and from Mexico must be rendered, as well as that of the former peasant from the Ukraine who now makes his living brokering rental properties in New York. In many respects, Thomas and Znaniecki's (1918) early studies of Polish peasants in Europe and America anticipated the kind of ethnographic work on immigrant 'flows' being done at the start of the twenty-first century.

The shifts in social theory that accompany the growing complexity of the empirical world create new lenses with which to see, and therefore present new challenges for conducting a faithful ethnography. Increasingly, 'local' social processes are influenced by supra-local forces that must be studied to illuminate the connections among race, class, power, deindustrialization, and pluralism. To be effective, ethnography, then, must be holistic.

Ethnographers themselves, their audiences, and other consumers are also becoming increasingly diverse. The articulate voices of African-American, Native American, Asian-American, Latino, and gay and lesbian ethnographers as well as local residents who have become anthropologists and sociologists are being heard. Such diversity raises obvious questions about the politics of representation. Increasingly, as never before, ethnographers want to render their own stories, their own realities and local knowledge, and in doing so, make competitive claims on intellectual turf. In these circumstances, some stories get heard, others are silenced, and some interested parties want only the most flattering stories of their 'own' represented.

These are some of the more pressing challenges for urban ethnography today, and they are well worth the effort. As these challenges are met, urban ethnography will become more complex, meaningful, and it is hoped, effectual. David Riesman once likened worthwhile ethnography to a conversation between classes. In this sense, each ethnographic case study can be viewed as an important part of a dialogue for understanding between and among those of diverse backgrounds, a dialogue that becomes steadily more urgent.

See also: Du Bois, William Edward Burghardt (1868–1963); Ethnography; Ethnology; Goffman, Erving (1921–82); Grounded Theory: Methodology and Theory Construction; Participant Observation; Poverty, Culture of; Qualitative Methods, History of; Qualitative Methods in Geography

Bibliography

Abbott A 1999 *Department and Discipline: Chicago Sociology at One Hundred*. University of Chicago Press, Chicago

Abu-Lughod J 1994 *From Urban Village to East Village: The Battle for New York's Lower East Side*. Blackwell, Oxford, UK

Addams J 1910 *Twenty Years at Hull House*. Macmillan, New York

Addams J, Lasch C (eds.) 1965 *The Social Thought of Jane Addams*. Bobbs Merrill, Indianapolis, IN

Anderson E 1978 *A Place on the Corner*. University of Chicago Press, Chicago

Anderson E 1990 *Streetwise: Race, Class, and Change in an Urban Community*. University of Chicago Press, Chicago

Anderson E 1999 *Code of the Street: Decency, Violence, and the Moral Life of the Inner City*. W. W. Norton, New York

Anderson N 1923 *The Hobo: The Sociology of the Homeless Man*. University of Chicago Press, Chicago [reprinted 1961]

Becker H S 1963 *Outsiders: Studies in the Sociology of Deviance*. Macmillan, New York

Becker H S 1970 *Sociological Work: Method and Substance*. Aldine, Chicago

Becker H S 1998 *Tricks of the Trade: How To Think About Your Research While You're Doing It*. University of Chicago Press, Chicago

Borawoy M 2000 *Global Ethnography: Forces, Connections, and Imaginations in a Postmodern World*. University of California Press, Berkeley, CA

Codere H (ed.) 1966 *Kwakiutl Ethnography*. University of Chicago Press, Chicago

Deegan M J 1988 *Jane Addams and the Men of the Chicago School, 1892–1918*. Transaction Books, New Brunswick, NJ

Drake S C, Cayton H R 1945 *Black Metropolis: A Study of Negro Life in a Northern City*. Harcourt Brace, New York [reprinted, Harper & Row, New York, 1962]

DuBois W E B 1899 *The Philadelphia Negro: A Social Study*. University of Pennsylvania Press, Philadelphia, PA [reprinted 1999 as *The Philadelphia Negro: Centennial Edition*]

Duneier M 2000 *Sidewalk*. Farrar, Straus & Giroux, New York

Faris R E L 1970 *Chicago Sociology: 1920–1932*. University of Chicago Press, Chicago

Feagin J 2000 ASA Presidential address, Annual Meetings in Washington, DC

Fine G A, Gusfield J R (eds.) 1995 *A Second Chicago School? The Development of a Postwar American Sociology*. University of Chicago Press, Chicago

Friedman J 1994 *Cultural Identity and Global Process*. Sage Publications, New York

Gans H J 1962 *The Urban Villagers: Group and Class in the Life of Italian Americans*. Free Press, New York [reprinted 1982]

Garreau J 1991 *Edge City: Life on the New Frontier*. Doubleday, New York

Geertz C 1983 *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York

Geertz C 1988 *Works and Lives: The Anthropologist as Author*. Stanford University Press, Stanford, CA

Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY

Guterbock T, Janowitz M, Taub R 1980 *Machine Politics in Transition*. University of Chicago Press, Chicago

Hall K 1997 Understanding educational processes in an era of globalization. In: *Issues in Education Research: Problems and Possibilities*. Jossey-Bass, San Francisco, CA

Hannerz U 1968 *Soulside: Inquiries into Ghetto Culture and Community*. Columbia University Press, New York

Horowitz R 1983 *Honor and the American Dream: Culture and Identity in a Chicano Community*. Rutgers University Press, New Brunswick, NJ

Hughes E C 1943 *French Canada in Transition*. University of Chicago Press, Chicago [reprinted 1973]

Hughes E C 1945 Dilemmas and contradictions of status. *American Journal of Sociology* **50**: 353–59

Hughes E C 1973 *Sociological Eye: Selected Papers*. Aldine, Chicago

Hunter A 1974 *Symbolic Communities: The Persistence and Change of Chicago's Local Communities*. University of Chicago Press, Chicago

Junker B 1960 *Field Work: An Introduction to the Social Sciences*. University of Chicago Press, Chicago

Katz J 1999 *How Emotions Work*. University of Chicago Press, Chicago

Kornblum W 1974 *Blue Collar Community*. University of Chicago Press, Chicago
Ladner J A (ed.) 1998 *The Death of White Sociology: Essays on Race and Culture*, 2nd edn. Black Classic Press, Baltimore, MD
Lidz V 1977 The sense of identity in Jewish-Christian families. *Qualitative Sociology* **14**(1)
Liebow E 1967 *Tally's Corner: A Study of Negro Streetcorner Men*. Little, Brown, Boston, MA
Lofland L 1998 *The Public Realm: Exploring the City's Quintessential Social Territory (Communication and Social Order)*. Aldine de Gruyter, New York, NY
Lynd R S, Lynd H M 1929 *Middletown: A Study in American Culture*. Harcourt Brace, New York [reprinted 1956]
Matza D 1969 *Becoming Deviant*. Prentice-Hall, Englewood Cliffs, NJ
Nyden P, Figert A, Shibley M, Burrows D (eds.) 1997 *Building Community: Social Science in Action*. Pine Forge Press, Thousand Oaks, CA
Park R E 1918 The city: Suggestions for the investigation of human behavior in the urban environment, reprint 1967. In: *The City*. University of Chicago Press. Chicago
Park R E, Burgess E W, McKenzie R D 1928 *The City*. University of Chicago Press, Chicago [reprinted 1967]
Portes A, Rumbaut R G 1996 *Immigrant America: A Portrait*. University of California Press, Berkeley, CA
Rainwater L 1968 *Behind Ghetto Walls: Black Families in a Federal Slum*. Aldine, Chicago
Rohner R P (ed.) 1969 *The Ethnography of Franz Boas*. University of Chicago Press, Chicago
Sassen S 1991 *The Global City: New York, London, Tokyo*. Princeton University Press, Princeton, NJ
Sassen S 2000 *Cities in a World Economy*, 2nd edn. Pine Forge Press, Thousand Oaks, CA
Shaw C R 1938 *The Jack-roller: A Delinquent Boy's Own Story*. University of Chicago Press, Chicago [reprinted 1966]
Short J F, Strodtbeck F L 1965 *Group Process and Gang Delinquency*. University of Chicago Press, Chicago
Suttles G D 1968 *The Social Order of the Slum*. University of Chicago Press, Chicago
Suttles G D 1972 *The Social Construction of Communities*. University of Chicago Press, Chicago
Thrasher F M 1927 *The Gang: A Study of 1,313 Gangs in Chicago*. University of Chicago Press, Chicago
Tsai G 1998 Middle class Taiwanese immigrants' adaptation to American society: Interactive effects of gender, culture, race, and class. Unpublished Ph.D. dissertation. University of Pennsylvania, Philadelphia, PA
Thomas W I, Znaniecki F 1918 *The Polish Peasant in Europe and America*. G. Badger, Boston
Wacquant L J D 2000 A fleshpeddler at work: Power, pain, and profit in the prizefighting economy. *Theory and Society* **27**: 1–42
Warner W L, Lunt P S 1941 *The Social Life of a Modern Community*. Yale University Press, New Haven, CT
Washington R L 1990 Brown racism and the formation of a world system of racial stratification. *International Journal of Politics, Culture, and Society* **4** (2): 209–27
Whyte W F 1943 *Street Corner Society: The Social Structure of an Italian Slum*. University of Chicago Press, Chicago [reprinted 1993]
Wilson 1987 *The Truly Disadvantaged: The Inner City, The Underclass, and Public Policy*. University of Chicago Press, Chicago
Wirth L 1928 *The Ghetto*. University of Chicago Press, Chicago
Zorbaugh H W 1929 *Gold Coast and Slum: A Sociological Study of Chicago's Near North Side*. University of Chicago Press, Chicago
Zukin S 1991 The hollow center: US cities in the global era. In: Wolfe A (ed.) *America at Century's End*. University of California Press, Berkeley, CA, pp. 245–61

E. Anderson

Urban Geography

For the first time in human history, the majority of the world's population lives in cities. Metropolitan areas will soon become the immediate sphere of human existence and experiences, and this sphere is changing rapidly. The megacities of low-income countries are subject to stark urban population growth and as a result are increasingly difficult to command and to control. Urban/metropolitan areas of the highly industrialized nations are likewise undergoing a historical change: in the light of global trends in economics, society, and politics, they are facing challenges which require new and extensive economic and local policies in order to meet increasing intra-urban competition for investments and taxpayers.

Theoretically informed social science research is absolutely necessary to monitor trends of urban development, to provide basic information for the optimization of local strengths, to design urban development concepts which adequately provide for local needs and demands, and for urban planning which equally respects collective decisions, increased competition, modern urban structures, and the individual 'feeling' of the city. Social science research defines and analyzes control factors of urban development, brings to light basic data and information about urban structures, and processes, perceptions and the attractiveness of the cities. This is where urban geography has an important role to play.

1. *Urban Geography: Tasks and Approaches of a Social Science Discipline*

1.1 *The Scope of Urban Geography*

Urban geography is a specialized discipline within human geography. It deals with the analysis, explanation, and prognosis of urban forms, urban social fabric, and economic structures and functions. Urban geography addresses research questions from economic, political, social, and ecosystem geography in

their urban contexts at various scales. The scientific results of urban research increasingly serve as a basis for decisions on public investment, allocation of resources, and socioeconomic and urban development planning. Urban research can be broken down into interurban (system of cities) and intra-urban analyses, which address the processes going on within cities (see *Cities, Internal Organization of*).

1.2 Conceptual Approaches and Types of Research Areas in Urban Geography

A large number of conceptual approaches are utilized in urban geography. Each addresses different types of research questions that may contribute to the understanding of the complexity and complementarity of factors influencing the urban system. The approaches pertain to urban form and urban morphology, the urban social fabric and economic structures, urban development, and urban policy.

1.2.1 Urban form and urban morphology. This approach deals with the analysis of the built environment. The changing urban morphology, including historic/heritage preservation, the creative reuse of landmark buildings, the architectural and technological upgrading of dilapidated buildings and urban revitalization, have gained importance in recent years. These measures can improve a city's attractiveness for investors and high-income earners, hence the city's fiscal base, by increasing the emotional and intellectual attachment of the target population to the city.

Whereas revitalization has been shown to increase the attractiveness of the townscape and urban image, this form of urban redevelopment is not without social costs. Gentrification and urban sub-area revitalization, incumbent upgrading, and the dislocation/displacement of the lower-income population are all interrelated. The study of changes in urban form and the built environment and the effects on local population groups and businesses is but one aspect of understanding urban decline or urban fortunes, social restructuring, and population and neighborhood change.

In European countries, urban morphology is of particular relevance in historic cities, for example, in cities built during the medieval period. Here, the knowledge of the historic urban structure and design aids in finding and identifying historic sites that have either been buried by construction activities of later centuries, or that need to be preserved and protected from contemporary rezoning and/or construction activity.

Urban form and urban morphology have also been taken up by Neo-Marxian/Neo-Marxist urban geographers. The regulationist school, for example, tries to uncover the power structures and the functioning of urban regimes. According to this school of thought, urban form and the 'turning of skylines and CBD's is a means of accumulating capital at the cost of other urban sub-areas, a process accompanied by policies of neglect and disinvestment. Social and welfare disparities as are manifest in differing morphologies across urban space have thus become another major area of urban research.

1.2.2 Urban function and urban social structure. Analysis of urban land use development, urban functions, and specializations: specialized economic activities seek locations with the greatest competitive advantage. This approach helps to understand urban land use development and to determine those land use patterns that most competitively provide basic support for the city. The internal structures of specialized urban functions (wholesale, retail, service activity, etc.) within the city and their market orientation, i.e., urban locations that are best suited to meet the city's needs, are points of focus. On the micro scale of the city, the analysis of central business districts and neighborhood business/convenience centers have received priority. On the meso and macro scales of urban systems analysis, the degree of specialization in and between cities is considered the base for identifying a city's market potentials, strengths, and niches within a regional or national hierarchy of cities.

Empirical examination of factors organizing space and megalopolitan structure deals with the metropolis as a changing configuration and the processes by which metropolitan areas grow and expand beyond their rigid corporate limits. Studies of metropolitan patterning examine the functional differentiation of suburbs, the interrelation of the city and its surrounding area through economic linkages and commuting, and the way metropolitan communities reorganize themselves into supra-metropolitan areas. The development of megalopoles, the increasing functional specialization of metropolitan centers and communities, and their growing interdependency are key research issues.

Urban social geography and the factorial ecology of cities, social status differentiation, and segregation: urban areas are highly differentiated complex systems (see also *Cities, Internal Organization of* in this volume). These approaches deal with urban sub-areas, urban subpopulations, social and economic characteristics of urban neighborhoods, the social patterning of local residents by race, ethnicity, and class, and the behavior of subpopulations as a function of their social group, race, and class, or as mediated by the characteristics of the neighborhood, urban sub-area, and the urban system.

Urban sub-area characteristics and urban social and spatial differentiation are commonly analyzed

using factorial ecology. This term refers to various statistical approaches using factor analytic methods. Small area analyses with a limited set of variables grounded in social theory are common. Factor analysis, on the other hand, depends on a large number of variables that are then reduced in an exploratory way to essential properties of a particular phenomenon of urban sub-areas or urban space. It determines urban sub-areas according to common social characteristics or in terms of households or individuals with common characteristics.

The urban ecologist or social geographer does not focus on spatial differentiation *per se*. Rather, social and spatial patterns are regarded as the manifestation of a social process. Indeed, the urban social geographer or urban ecologist studies such patterns with a view to uncovering the social, political, economic, or cultural processes that may be responsible for these patterns. Residential segregation of different social status groups, for example, occurs in many different cultural settings and reveals the most residentially segregated social groups, for example, as those at the top and at the bottom of the social status hierarchy. In increasingly multicultural societies, patterns of differentiation among social groups and neighborhoods and their particular spatial geometry are becoming a growing problem. Factorial ecological investigations that use a variety of computational techniques help identify the differences between social, demographic, and economic characteristics in urban space. They help to monitor processes of social distance as reflected in the degree of physical distance and residential separation.

Perception of the urban environment: this approach sees age and social status-related perceptions of the urban environment, personal activity radius, and spatial behavior as related variables. Individual perceptions of reality affect spatial behavior, be it shopping trips or intra-urban migration. For example, characteristics of the automobile society, such as the monotonous urban landscape, can negatively affect an individual's identification with the physical urban environment and may induce him/her to move, thus contributing to the erosion of a city's tax base and concomitant inner city decline.

Differing social and age groups, for example, the elderly, the youth, the poor, have distinct ranges of activities and different patterns of spatial behavior in urban space. Market research utilizes the results of studies on the perception of the urban environment and spatial behavior. At the micro scale of individual urban geography, studies concentrate on activities that a person does regularly and urban places that are visited regularly, i.e., daily or weekly, as they form the basis of the personal contact field and average information field. These may be analyzed using the methods of time geography and indicate an individual's capabilities to overcome distance by means of mobility and communication. At the macro scale, the collective spatial behavior of social groups and the spatial geometries of social group behavior are focused upon. Decisions regarding distances between shopping centers and other central urban functions, for example, make use of empirical findings of collective spatial behaviors (see *Behavioral Geography*; *Urban Activity Patterns*).

Personal mental/cognitive pictures of space are also referred to as images. The analysis of these images focuses on the subjective evaluation of urban space by residents, visitors, business people, and potential investors. Perceptual geographic urban analysis upholds the theory that spatial behavior, like intra-urban migration or shopping and recreational activity, is often limited to a closed field of perception and reference, and it is affected by personal spatial evaluation. The overall image of a city from the point of view of a commercial interest is affected by the 'soft' locational factors (e.g., amenities, attractiveness of the urban environment, social and demographic processes in the city, or business climate). Image analysis can contribute significantly towards finding reasons for the exodus of population, companies, and enterprises, or disinvestment. Moreover, measures for the improvement of a negative city image or for instigating desired development can be recommended (see *Spatial Cognition*).

1.2.3 Urban development. The analysis of urban development as related to historical, political, and geographic peculiarities: this approach deals in particular with the development dynamics of the city and the determinants of that development. Different historical–political systems of society, socioeconomic factors and urban politics (see Sect. 1.2.4) fall into this category. Urban development analysis takes the positive consequences (e.g., concentration of social and economic performance, importance as center of innovation/innovation hub) and negative consequences of development (e.g., environmental problems or urban blight) into consideration. This approach can offer important impulses for the political, economic, and social components of urban development planning.

Urban cultural geography, intercultural comparison of cities and urbanism: this approach places the historical and cultural context of cities in the foreground. Processes of urbanization and the current physical urban structures of cities, including the inner differentiation, are assumed to be products of a series of sociopolitical systems. In cultural genetic urban geography, persistent historical urban structures are identified according to the appropriate historical social system and cultural sphere of influence. The understanding of current structures and processes that have evolved through construction, reproduction, renovations, and rebuilding during several cycles of social change, is easier if seen in the context of the past.

However, society changes faster than physical urban structures. Thus, the reflection of changes in society in old buildings and in the urban fabric, as well as the intercultural comparisons in society in old buildings and in the urban fabric, as well as the intercultural comparisons of these generation influences are amongst the most interesting topics of research in cultural genetic urban geography.

1.2.4 Urban policy. Globalization and urban policy: cities are social systems in space and as systems they react to any changes that result from both local and supralocal or global developments. Such supralocal trends include:

Increasing competition between metropolitan areas due to globalization of the economy and international competition.

Deregulation, separation of social politics from trade and industrial politics, leading to a decline of the welfare and possible intensification of social problems.

Economic and political integration, political change and crises that initiate migration flows. These result in increasing polarization of society and affect the perceived quality of the urban environment.

Change of societal values and pluralism of life styles. These encourage urban exodus and the concentration of marginal groups in the inner city.

Global economic restructuring, the increased competition between cities, and the decline of the welfare state force cities to pursue entrepreneurial urban development strategies. These may include economic incentives, the promotion of investments or megaprojects to increase urban attractiveness, and the decision to promote growth rather than redistributive urban development. Also, new entrepreneurial public–private partnerships or rigid urban regimes that place a great emphasis on private sector planning have emerged.

Urban regimes follow different approaches to development; some implicitly exclude or include social policies as part of their overall strategy, with different levels of impact on long-term development processes. On the micro scale of the individual city, the analysis of the interrelationships between urban policy, urban development, urban sub-area development, and the changing geometries of segregation helps to understand how and why planning may become unintentionally a determinant of social and spatial polarization. Given the adoption of Western urban policy and planning models in other countries, it is important to understand the social, spatial, and economic implications of each type of policy for urban development as such and in different cultural settings.

Government devolution, the state and local response: on the macro scale of the national urban systems, the relationship between national/federal urban policy and local policy and local urban development is of interest to urban geographers. Their main interest is what happens in cities when federal policies change. For several years now, one of the more noticeable national trends has been the shifting of financial responsibility away from the federal government toward state and local governments. This trend has been observed particularly in the US and Western European federal systems. Known as 'devolution,' it has led scientists and observers to believe that social problems may worsen if the federal/national role of counterbalancing unfavorable social and economic developments is restricted. To place theoretical concerns and real world developments under a policy of devolution into proper perspective, it is important to look at the actual situation and local responses. Federal/national policies may at first glance adversely affect inner city and other poverty areas, but one must also look at local policies and responses to federal/national policy in order to get a more balanced view of long-term adjustments and trends in societal development.

Policy-oriented urban geography deals with urban problems, such as social and economic disparities, caused by policies at various levels. In research on attempts at restoring run-down inner city neighborhoods in the US, one can find many examples of innovative community concern expressed by both the public and private sectors with the potential to counterbalance public sector devolvement to some extent. European nations, amongst others, have developed interesting local policy mechanisms to deal with the ills and imbalances of society manifest in urban areas. Policy-oriented urban geography thus studies and uncovers the problems and potentials of local planning responses to current problems of urbanism and federal and state polices.

2. *The Operational Methods of Urban Geography*

It is clear, then, that the differing conceptual approaches largely focus on two types of research issues:

those related to urban space and urban sub-area characteristics, the basic dimensions of city patterning, and the determinants and processes behind changing structural characteristics of urban space over time. The attributes may refer to physical, social, economic, or demographic phenomena and their differentiation (their concentration, densities, distribution, growth) in urban space;

those related to people in urban space, i.e., individual and collective spatial behavior (for example shopping trips, commuting, daily activity ranges), moreover, perceptions, images of urban space, and planning.

Concomitantly, there are two major methodological approaches: those of empirical regional/urban sub-area analysis, and those of empirical social science, here in particular the survey methods. These methodologies are, of course, not mutually exclusive, but

rather, complementary. Each sheds light on issues that would otherwise be impenetrable. Not only urban geography but also urban research in general makes use of these methodological approaches.

2.1 Methods of Empirical Regional/Urban Sub-area Analysis: 'Urban Social Monitoring'

Social and societal developments have their real world manifestations in urban space, and social and economic developments in urban areas are reflected in the structural characteristics of urban sub-areas. Urban geography is well suited to examine attributes and developments of structural characteristics by means of analytical techniques. These analyze the areal structure of urban communities in terms of attributes. Empirical urban research is both regional research specifically in urban areas and social or socio-spatial research. The methods correspond to those found in regional geography in so far as they are utilized for delineation and observation of structural change of agglomerations, city centers/cores, urban expansions, and suburban areas. Within the city itself, the research units are districts and neighborhoods as well as other 'official' spatial units of division, be they for planning, political, or statistical purposes (e.g., planning units), school and electoral districts, street rows and blocks. Urban sub-areas may be of any scale: census tracts are commonly used as statistical reference areas. Micro scale urban social geography also makes use of block level data to characterize the increasing differentiation of urban social milieus.

The methods of empirical urban sub-area analysis allow for urban social monitoring. This refers to the inventory, documentation and analysis of detailed socioeconomic structural patterns and processes of change. Complex spatial processes are broken down into individual components.

Statistical techniques aim to characterize and analyze urban space, urban sub-areas, and urban structural developments comprehensively. Three approaches are important:

(a) For descriptive purposes, methods include computer-assisted cartography and the refined cartographic and analytic methods enabled by Geographic Information Systems (GIS). These may produce unidimensional or multidimensional maps of social, demographic, or other phenomena as differentiated in urban space. GIS, of course, is also suited for the establishment of a long-term statistical cartographic database, which can be periodically updated. Such a database would simplify thematic longitudinal onsite analysis of the target urban region with regard to social, economic, and demographic processes and forecasts.

(b) In a more exploratory sense, factorial ecological investigations use a number of multivariate descriptive statistical techniques (the methods of factor analysis) to identify the essential dimensions that characterize and differentiate one urban sub-area from another in terms of social science variables. Underlying the concept of urban social areas is the assumption that societal processes reflect natural processes in that they have a competitive dimension that can lead to processes of selection. Social structures and social change in space are seen as the result of mutual adaptation of competing species. According to R. E. Park (1936), socio-ecological studies deal with processes that either uphold an existing social balance or that disturb the existing order in order to reach a new, relatively stable existence. One specific type of factorial ecology is social area analysis. Social area analysis is based on the theory of Shevky and Bell who understood urban social space as being primarily characterized by social rank, urbanism, and ethnicity. As such it only works with a limited set of input variables. Cluster analyses subsequently performed on factor analyzed urban sub-areas can help identify groups of sub-areas with common patterns of variability.

(c) In order to understand the determinants of and processes responsible for such patterns, one may combine descriptive and analytical statistical techniques. Factor scores from factorial analyses may, for example, be used as input data in multiple regression analyses that relate these aggregate characteristics to explanatory variables.

Social monitoring of urban sub-area characteristics over time enables a scientifically sound evaluation of the current structural change: urban geography falls back on existing statistical data collected by public and private institutions or public welfare organizations. As the data reflects institutional norms and goals, urban geography has no influence on either the exact questions, the survey method, or the aggregation and systematization of the indicators. Consequently, theoretically informed urban research is limited by the quality of these (secondary) data sources. However, the quality of official data banks and the methodology of secondary research in the field of spatial and thematic aggregation of data are improving continually.

2.2 Survey Methods—Empirical Social Science Research

Urban space is not only an attribute of space but also an area of activity. In order to analyze activity systems/patterns of individuals or households, as well as preferences, felt needs, perceived quality of the urban environment, locational/residential behavior, or general spatial behavior in a specific urban context, the urban geographer has to use survey methods.

Empirical social science research collects social data by means of survey methods, experiments, 'pure' observation, or 'participative' observation. In human geography, empirical social science research is applied to identify and investigate areas of perception and activity, as well as opinions and attitudes to values and

norms in society, as these form the framework of spatial behavior. Survey research methods in social geography concentrate on the interrelationship/correlation between spaces and social group behavioral patterns, designation and appraisal, group perception, and evaluation of spaces.

The advantages of incorporating surveys into the initial stages of research are that important issues can be identified at an early stage, and the survey allows for a differentiation and consolidation of research aspects. At the same time, preparation of standardized methods for covering the chosen topic continues unhindered or alternative and appropriate research methods (i.e., not from within empirical social science research) may be sought. Generally, the social surveys of empirical social science research open the door to a greater spectrum of human geographic research issues than would be possible with official statistics.

Research questions of urban geography thus deal with different sets of issues:

either the physical attributes of spaces or the indicators of their functional interconnection; or

immaterial aspects (spatial perception or space-related activity); or

both; or

those aspects of spatial change that result from the interaction between immaterial and physical elements (e.g., the physical results of political visions for the urban area and the consequent measures, or socio-spatial disparities caused by power relations in society).

The selected database depends on the research question. Studies of objective reality (e.g., socio-geographic structural analysis) use data that reflect numerically definable spatial characteristics (as opposed to subjectively perceived spatial attributes). This data is mostly found in official statistical material, published and unpublished, the basis of which is extensive censuses (e.g., of population) or surveys (e.g., of enterprises). Personal, small-scale, and tailored surveys are necessary if the time lapse between official censuses is too great or if certain details are missing (e.g., in order to incorporate structural, socio-spatial, and economic aspects, entrepreneur surveys or mappings of building conditions in different neighborhoods, of spatial use, or overall inner city differentiation may be necessary). However, if the research focus lies in the perceived surface and activity area of individuals and groups, then interviews are necessary if spatially related behavior, behavioral background, and possible consequences of behavioral patterns are to be dealt with.

Methodological procedures in urban geography stand out for their broad range. Some procedures overlap, thus hindering the exclusivity of subfield methodology. For this reason, it is not possible to refer to a particular methodology for economic geography, social geography or, for that matter, urban geography. However, within human geographic research of urban areas, the methodical–methodological approach of dealing with different issues is basically the same. Geographic scale, spatial delineation, and weighing of indices are the main differences. Between factorial ecology and survey methods, only the data sources and the scale of the evaluation differ. Even though the goals may vary, the basic principle of preparing and evaluating official or specifically collected social data from survey research in a descriptive and/or quantitative analytical manner stays the same.

3. *Outstanding Features of Urban Geography*

The standing of urban geography within the social sciences in general draws essentially on two things that the other disciplines dealing with urbanism do not share to the same extent.

3.1 The Combination of Cognitive and Methodical Capabilities

(a) Empirically and theoretically informed research from an integrational/holistic perspective.

(b) Relevance to society.

(c) Diversity of urban research topics in social, economic and environmental fields, amongst others. The range of urban research topics includes, for example:

(i) problems of increasing social differentiation in urban sub-areas and the development of an urban underclass, urban ethnic-cultural and lifestyle milieus, long-term urban unemployment, the welfare poor, the working poor;

(ii) population change (socio-demographic, socio-economic and socio-cultural), infrastructural, and policy adjustments;

(iii) housing and urban labor market developments;

(iv) urban environmental quality and urban sub-areas, questions of environmental racism;

(v) functional interrelatedness of cities and their suburban areas;

(vi) the territoriality of each context, the unique and unrepeatable set of social and economic relations, local networks in urban areas and development; the territorial embeddedness of global impacts;

(vii) local policy responses to supralocal trends;

(viii) entrepreneurial urban policy and micro-scale urban social development;

(ix) cross-cultural urban development; and

(x) urban systems development.

3.2 The Specific Methodological Competence and Diversity of Analytical Responses

(a) Empirical research: primary survey methods, such as monitoring, mapping, or field surveys;

(b) statistics: the scientific approach for drawing conclusions about research questions/hypotheses by means of processing spatial primary data and official-statistical secondary data;

(c) cartography: tool for working with and presenting spatially located and differentiated phenomena;

(d) remote sensing: interpretation of aerial photos and satellite photos as a further important source of information; and

(e) utilization of geographic information systems (GIS) to present and analyze data and developments in space.

4. Conclusions

Global economic restructuring, political change, government devolution, and social restructuring alter urban and regional fortunes and give rise to new socio-spatial patterns of development. Fragmentation, polarization, and new local spatial identities form new urban realities. Formerly prosperous cities may lose their comparative locational advantage. In order to retain competitiveness and build up local (endogenous) potentials, or to govern uncontrollable cities, new models, modes, and mechanisms of planning are being developed. In cities of the Western world, public–private partnerships—coalitions of major local economic and political players—have been outlining/developing and implementing grand visions of development. Megacities of low-income countries have yet to cope with the more basic infrastructural needs of their populations.

Thus increasingly, in cities worldwide, both urban planning and the prevailing urban regime, its orientation, and complex interaction with the regional, national, or supranational economic and political systems, have a significant effect on the development of a city. In order to deal with contemporary urbanism, decision makers will need to rethink urban processes, structures, and policies. What is needed is an understanding of the city from a systems perspective. This views urban developments, form, and social and economic structure as interrelated with the systems of society, economy, and politics. It also understands the city as an organism/entity with enormous local capacities and strengths that might successfully counteract the local or regional effects of globalization or other supralocal forces.

Urban geography is a systems-oriented social science discipline with great relevance to interdisciplinary solutions of problems of urbanism and urban areas. The systems perspective, the skills required, and the variety of research questions and applied research in urban geography make this discipline preeminently suited to understand and deal with problems of contemporary urbanism.

See also: Behavioral Geography; Cities, Internal Organization of; Development and Urbanization; Globalization: Geographical Aspects; Migration, Economics of; Spatial Cognition; Urban Activity Patterns; Urban Anthropology; Urban Growth Models; Urban History; Urban Poverty in Neighborhoods; Urban Sprawl; Urban Studies: Overview; Urban System in Geography

Bibliography

Bingham R D, Bowen W M, Amara Y A 1997 *Beyond Edge Cities. Garland Reference Library of Social Science Vol. 1180*. Garland, New York

Bridge G, Watson S (eds.) 2000 *A Companion to the City*. Blackwell Publishers, Oxford, UK

Cadwallader M (ed.) 1996 *Urban Geography: An Analytical Approach*. Prentice Hall, Upper Saddle River, NJ

Carter H (ed.) 1995 *The Study of Urban Geography*. Arnold, London

Dear M J (ed.) 1999 *The Postmodern Urban Condition*. Blackwell, Oxford, UK

Drakakis-Smith D W (ed.) 2000 *Third World Cities*, 2nd edn. Routledge, London

Fincher R, Jacobs J M (eds.)1998 *Cities of Difference*. Guilford Press, New York

Gugler J (ed.) 1997 *Cities in the Developing World: Issues, Theory and Policy*. Oxford University Press, New York

Hall T (ed.) 1998 *Urban Geography*. Routledge, New York

Hall T, Hubbard P (eds.) 1998 *The Entrepreneurial City: Geographies of Politics, Regime and Representation*. John Wiley & Son, Chichester, UK

Hayden D (ed.) 1995 *The Power of Place: Urban Landscape as Public History*. The MIT Press, Cambridge, MA

Herbert D T, Thomas C J (eds.) 1997 *Cities in Space: City as Place*. Fulton, London

Knox P L (ed.) 1994 *Urbanization: an Introduction to Urban Geography*. Prentice Hall, Englewood Cliffs, NJ

Knox P L, Pinch S (eds.) 2000 *Urban Social Geography: An Introduction*. Prentice Hall, Harlow, UK

Le Gates R T, Stout F (eds.) 2000 *The City Reader*. Routledge, London

Lyon L 1999 *The Community in Urban Society*. Waveland Press, Prospect Heights, IL

Massey D, Allen J, Pile S (eds.) 2000 *City Words* (*Understanding Cities*). Routledge, London

Miles M, Borden I, Hall T (eds.) 2000 *The City Cultures Reader*. Routledge, London

Paddison R (ed.) 1999 *Handbook of Urban Studies*. Sage, London

Pile S, Brook C, Mooney G (eds.) 2000 *Unruly Cities? Order/Disorder* (*Understanding Cities*). Routledge, London

Potter R B, Lloyd-Evans S (eds.) 1998 *The City in the Developing World*. Longman, Harlow, UK

Pugh C (ed.) 2000 *Sustainable Cities in Developing Countries*. Earthscan, Sterling, VA

Roberts G K, Steadman P (eds.) 1999 *American Cities and Technology: Wilderness to Wired City* (*The Cities and Technology Series*). Routledge, London

Short J R 1996 *The Urban Order: An Introduction to Cities, Culture and Power*. Blackwell, Cambridge, MA

Soja E W (ed.) 1996 *Thirdspace: Journeys to Los Angeles and Other Real-and-imagined Places*. Blackwell, Cambridge, MA
Yeates M (ed.) 1997 *The North American City*. Longman, New York

R. Schneider-Sliwa

Urban Government: Europe

Urban governments are defined first as political arena and instrument for enhancing democracy, participation, and steering local societies and second in terms of services provision and public policy. They are back on the political agenda of Europe, not as the old medieval city, but as more autonomous political authorities within a European governance in the making. Urban governments usually are related to the nation–state in terms of democratization and legitimation of forms of territorial management.

1. Urban Government and the making of the Nation–State in Europe

Weber famously emphasized what he saw as the distinctive characteristics of European societies, that is, the medieval occidental city defined in terms of 'sworn confraternisation' based upon a fortress, a market, bourgeois associations, specific rules in terms of land ownership and tax, and sometimes courts and armies (Weber 1978). The importance of the urban government was stressed as the city developed upon the movement of medieval 'communalisation' (i.e., communes being formed through acquiring a *charter*). Cities became institutionalized associations, autonomous and autocephalous, active territorial corporation characterized by autonomy and capacity for action towards the outside (the lord, the prince, the state, the emperor, rival cities) and led by urban officials. Medieval urban governments developed democratic institutions, and commerce left its mark through the edification of monuments symbolizing this power: squares, town halls, belfries or bell-towers.

In due course, 'voracious states' consolidated with or against 'obstructing cities' (Tilly and Blockmans 1994). The making of the modern state and the coming age of the second (industrial capitalism) marked the end of the golden age and autonomy of occidental cities.

Later, in industrial cities, for instance in the UK but also in Germany, France, and Scandinavia, the scope of social problems became such that elites in urban government pioneered policy programs in housing, planning, basic elements of welfare, education (De Swaan 1988) and hygienist concerns led to the 'Haussmanisation' movement of city rebuilding, that is, the emergence of local public goods. Urban governments played a key role in providing basic utilities and services such as water, sewage, street lighting, and, later, gas and electricity, firemen, and transport, not to mention slaughterhouses. This development was diverse, fragmented, contested between a conservative petty bourgeoisie and the municipal socialism movement, and more consistent in the North of Europe than in the South. Most local government in Europe gained legal recognition in the second part of the nineteenth century. Gradually, a professional local bureaucracy emerged to deal with those developments. The rise of urban government was not just a local or national phenomenon. Exchanges of experiences of ideas, for instance in planning and social housing, were crucial.

Later, in most of the post-1945 period, the category 'European urban government' did not make much sense, and was not an issue. Within the social democratic compromise of most European states, the role of urban government was understood within the center-periphery paradigm (Mény and Wright 1985), that is, in national terms and as part of local government. Urban governments differed in Europe because each country had a different constitutional setting, different rules, different public finances systems, different political systems and traditions, and different organizations to provide services. Sometimes, variations within a country were also important (Germany or Italy).

Urban governments were understood either as a functional entity to deliver services, in particular welfare services (hence the long-lasting debate on size and amalgamation), or as a political unit. In their classic comparative research, Goldsmith and Page (1987) have suggested that local government autonomy in Europe should be analyzed in terms of autonomy through two major criteria which encompass or are closely related to other dimensions: legal status and political status).

That analysis clearly stressed the differences between the welfarist northern European urban governments and the more political (sometimes clientelistic) southern European urban governments.

2. A New Pattern of Constraints and Opportunities for European Urban Governments

These classic distinctions are now under question because a common set of pressures and opportunities (Europe, fragmentation, state reorganization, economic restructuring, social tensions) tends to blur the frontiers between existing national models of urban governments and to reinforce differences within nation–states. Several pressures for changes are put forward.

Urban governments were contested in the 1970s and 1980s by urban social movements. The bureaucratization, hierarchies, urban regeneration projects, com-

plex and fragmented decision-making process of urban government were attacked in cities all over Europe. Conflicts entered the realm of urban politics in relation to housing, planning, large infrastructures projects, economic crisis, and cultural issues. New groups, beyond class basis, organized to raise new issues (quality of life, democracy and participation, economic development, and culture) to promote urban change against elected urban leaders. New middle classes gradually were incorporated within political parties (social democratic and green) and played an important role in many European cities to promote a new set of urban policies to deal with those issues. In the most radical cases, squatters in Amsterdam or Berlin for instance, urban government officials have learned to cooperate, to provide sources of funding, and to incorporate those groups in more loosely-defined structures of governance (Mayer 2001). Preventing large social conflicts and including various groups has become the norm for urban governments.

Those movements and changing patterns of governing elites led both to structural changes and experiences in urban governments all over Europe such as the 'free communes' in Scandinavia. More significantly, most urban governments have initiated management reforms including neighborhood councils and the decentralization of services management, that is, new mechanisms for citizen participation in decisions despite the twofold difficulty of giving up power and budgets (uneasy for councillors) and of sustaining citizen interests in running day-to-day business. Beyond the UK, market friendly ideas associated to 'new public management' (Dunleavy 1994) are having an impact in urban government, in particular in the north of Europe (Baldersheim and Stählberg 1995). Issues of citizen participation in urban governments are associated with growing issues of management efficiency in the delivery of services to customers. The restructuring of the public sector leads to increased confusion in public policies and the fragmentation of urban governments (Pierre 1998); hence the growing interests for issues of leadership, management, coordination, and governance (Borraz et al. 1994, Stoker 2000).

The fragmentation of urban government was also accelerated by the large privatization movement which is slowly making its way in western Europe. Lorrain and Stoker (1997) suggest that variation depends upon the intensity of political ambition and the public/private dimension with, for instance: (i) the UK (reform of urban government plus the privatizing of a large part of social housing, major utilities, not to mention the introduction of competitive tendering for most social services); (ii) Scandinavia where municipal public bodies are still dominant (gradual change) despite the pressure of competition; (iii) France where, in the post decentralization process, urban mayors wanted to increase their autonomy vis-à-vis the state by privatizing a large amount of services but remaining active in their regulation, and (iv) slow change in the south of Europe and in oriental Europe because of uncertain changing institutional patterns. Different models of regulations and contracts of those privatized or semiprivatized services emerge in Europe and are in competition both in Brussels and in former Eastern European cities. Large utilities firms are now competing all over Europe and beyond to provide new services.

Urban governments in Europe are also facing political pressure related both to the restructuring of the state and to questions about representative democracy and changes of political culture (Clark and Hoffman-Martinot 1998). First, there is a transnational movement to raise issues of urban democracy and accountability (King and Stoker 1996). This has translated, most importantly in Germany and in Italy, but even in the UK to a limited extent, in the direct election of urban mayors. Mayors have often gained a higher political profile in most European countries, most notably in Italy but even in small centralized countries such as Portugal, Sweden, or Finland. Second, relatively small green parties are now established in most continental countries and have an impact on urban politics. Since the mid-1980s onwards, they have pushed two kinds of issues: direct consultation (local referendum, experiences of deliberative jury) to go beyond the elected representative on major issues (for instance, the choice of public transport) and procedural democracy. They also benefited from the move against corruption in Italy, France, or Germany. Third, at a time when some urban government elites have espoused the discourse and less often the policies of urban entrepreneurialism, thus building sorts of growth coalitions, new urban social movements have appeared specifically opposing flagship urban projects supposedly crucial to enhance cities' image. Fourth, the relative erosion of national parties' capacity to mobilize and increase resources at the disposal of urban government, gives more prominence to the urban leaders and apparatus of parties which are, in some cases, also the aggregation of local parties. Fifth, the complexity of public policies together with the perceived needs to organize at a larger scale than old cities has led to a resurgence of attempts ('*trompe l'œil* renaissance of metropolitan governments' Lefevre 1998, p. 16) to move from urban governments to metropolitan governments for instance in Italy, Rotterdam, or London; hence the salience of the governance issue again and the role of urban governments. Sixth, the restructuring of the economy also had some direct impact on urban governments and planning. In the postwar period, most urban governments had little to do with economic development beyond planning, land, and infrastructures provision. All that was to change in the 1970s. Rising unemployment, state retreat, and gradually more firms' internationalization have put urban governments at risk (tax, social groups, social problems). In the early 1980s, a large movement of

creation of urban economic policy was noticed in particular in the UK, the Netherlands, Germany, France, Belgium, and also in the most dynamic Italian and Spanish cities, but less so in northern Europe (Judd and Parkinson 1990). Supporting firms and, above all, attracting outside capital became standard policies in most European cities; hence strengthened forms of cooperation between private interests and urban governments, but rarely American-like urban regimes or growth coalitions. Seventh, the decentralization of social policy and the European integration are also leading towards specific social and political urban rights (Garcia 1996). Differentiated patterns of political participation (elections, procedural democracy, associations) are leading towards more diverse forms of urban citizenship already obvious in the case of quasifederal states. Eighth, most European countries are still involved in some reform of the legal and tax powers of various levels of subnational government, regions here, cities there, quasifederal state or provinces. The rise of meso-government in Europe can be seen within investment figures. The share of local (but massively urban) government in the national public investment now amounts to around 50 percent in small centralized countries (Portugal, Finland) to around 60 percent (UK, Italy, Germany, Spain, Denmark), or even beyond 70 percent (The Netherlands, France).

3. *Conclusion: Urban Government and Governance in the European Polity in the Making*

Urban governments are part of wider social and political processes such as the making of Europe or the restructuring of the economy. Urban governments in Europe reflect both well-established institutional arrangements and more strategic behaviors with the European governance.

The making of Europe goes together with a vast movement of reallocation of authority. The changing scale does not suppress the narrow legal and financial relations between urban government and nation–states but includes them within a wider set of intergovernmental relations, networks, interdependent bodies. Urban governments have to deal with an increasing amount of actors and policy tools (contracts and partnership, for instance) together with more diverse networks and actors. The making of a European polity implies not just a more complex structure of vertical intergovernmental relations but also horizontal ones. Networks of cities are flourishing in Europe, including policy best practices exchanges financed by the European Commission, or old twinning arrangements modified in multidimensional cooperation between cities in Europe. In this multilevel governance in the making, the constraints imposed by the state are often lessened, urban governments tend to get more discretion (not always) but also a different set of constraints.

The institutionalization of the European Union has also concrete implications for urban governments: access to funding for specific programs (from structural funds to the environment) goes together with the learning of norms, new methods, buzz words (partnership), accounting rules, evaluation. The gradual institutionalization of different policy domains gives more and more importance to EU rules. Urban governments, all over Europe, gradually are developing strategies to cope with those opportunities and constraints, sometimes to influence them, for instance, in the case of structural funds (Goldsmith and Klausen 1997).

Rather than looking for convergence or divergence in urban governments, one can first notice the blurring of national frontiers to identify common modes of governance in groups of cities, usually articulated in part by national patterns. What is interesting in the European case is the attempt made by urban government elites to deal with increased fragmentation, to balance the politics of growth with social issues, and to strengthen governments and processes of governance. Urban governments in most European countries have become more complex organizations, more fragmented, and to some extent, more responsive to the demands of local groups and neighborhood but also more dependent upon firms, and utility firms, in particular.

In other words, the issue of articulating urban government and urban governance in different ways is central. That attempt to bridge the gap between politics and policies and to define some collective urban interest is at heart a political process. The erosion of the nation–state does not mean the end of politics in Europe—it reappears at every floor of the multilevel government—but urban politics may be less defined in terms of domination and more in terms of coalition and institution building, or even identity formation. Cities (together with regions) have returned to the central stage of European politics and urban governments are key actors within the emerging pattern of European governance.

See also: European Integration; European Union Law; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Urban History; Urban Policy: Europe; Urban Sociology; Weber, Max (1864–1920)

Bibliography

Bagnasco A, Le Galès P (eds.) 2000 *Cities in Contemporary Europe*. Cambridge University Press, Cambridge, UK

Baldersheim H, Stahlberg K (eds.) 1995 *Towards the Self-regulating Municipality*. Dartmouth, Aldershot, UK

Balme R, Faure A, Mabileau A (eds.) 1999 *Les Nouvelles Politiques Locales*. Presses de Sciences, Paris

Borraz D, Bullman U, Hambleton R, Page E, Rao N, Young K 1997 *Local Leadership and Decision-Making*. Joseph Rowntree Foundation, LGCC, London
Clark T, Hoffman-Martinot V (eds.) 1998 *The New Political Culture*. Westview Press, Boulder, CO
Cole A, John P 2000 *Urban Networks: A Comparison of Four Cities*. Routledge, London
De Swaan A 1988 *In Care of the State*. Polity Press, Cambridge, UK
Dunleavy P 1994 The globalization of public services production: Can government be 'best in the world.' *Public Policy and Administration* **9**(2):
Garcia M 1996 Cities and citizenship. *International Journal of Urban and Regional Research* **2**:
Goldsmith M, Klausen K (eds.) 1997 *European Integration and Local Government*. Elgar, London
Goldsmith M, Wolman H (eds.) 1990 *Urban Politics and Policies*. Blackwell, Oxford, UK
Harding A, Le Galès P 1998 Cities and states in Europe. *West European Politics* 2. F. Cass, London
Healey P, Davoudi S, Graham S, Manipour A, Marvin S 1995 *Managing Cities*. Wiley, Chichester, UK
Judd D R, Parkinson M (eds.) 1990 *Leadership and Urban Regenerations: Cities in North America and Europe*. Sage, London
Judge D, Stoker G, Wolman H (eds.) 1995 *Theories of Urban Politics*. Sage, London
Keating M 1991 *Comparative Urban Politics*. Elgar, London
King D, Stoker G (eds.) 1996 *Rethinking Local Democracy*. Macmillan, Basingstoke, UK
Lefevre C 1998 Metropolitan government and governance in western countries: A critical review. *International Journal of Urban Regional Research* **22**(1): 9–25
Mayer M 2000 *Urban Social Movements*. Blackwell, Oxford, UK
Mény Y, Wright V (eds.) 1985 *Centre Periphery Relations in Western Europe*. Allen and Unwin, London
Lorrain D, Stoker G (eds.) 1997 *The Privatization of Urban Services in Europe*. Pinter, London
Page E, Goldsmith M (eds.) 1987 *Central and Local Government Relations: A Comparative Analysis of Western European Unitary States*. Sage, London
Pierre J 1998 *Partnerships in Urban Governance*. Macmillan, Basingstoke, UK
Pickvance C, Préteceille E (eds.) 1991 *State Restructuring and Local Power*. Pinter, London
Sharpe J (ed.) 1993 *The rise of Meso Government in Europe*. Sage, London
Stoker G (ed.) 2000 *The New Politics of British Local Governance*. Macmillan, London
Tilly C, Blockmans W (eds.) 1994 *Cities and the Rise of States in Europe*. Westview Press, Boulder, CO
Weber M 1978 *Economy and Society (chapter XVI: The city)*. University of California Press, Berkeley, CA

P. L. Galès

Urban Growth Models

An urban model is a statement about the interrelationships between parts of an urban system, which not only describes these interactions, but makes possible the prediction of the effects of some kinds of changes in the system or in its environment. This is the essence of any scientific theory.

Viewed as a part of a scientific discipline, urban models provide an organizing framework for bringing together many aspects of knowledge about urban life, for testing subsidiary theories, and for identifying gaps in knowledge and guiding future research. Viewed as an aid to professional practice in urban planning or engineering, a model which is operational at a relevant level of detail can provide insights into the future results of new policies, and guidance in their formulation and improvement. To serve these two divergent purposes, a model must be understandable and acceptable to both the scientist and the professional.

Prior to the mid-1950s, many disciplines and professions engaged in disjointed efforts to analyze urban affairs. In Geography and Sociology, there were efforts to achieve an overall view in largely descriptive terms. Examples include Burgess (1927), Hoyt (1939), and Harris and Ullman (1945). Starting in a period of optimism about scientific management, the last 50 years have seen substantial growth in urban modeling, but there are still many imperfections in current models, and obstacles to their filling their dual role. We will try to expose the bare bones of both the progress and the difficulties as they have developed over five decades.

1. Basic Considerations

An urban metropolitan area comprises a set of organized activities and residential spaces, occupying land and buildings, and using infrastructure. These activities and their containers are at higher densities than their nonurban surroundings, and are more closely connected with each other than with their ambient environment. Their existence and development require the proper functioning of markets or allocation methods for land, for housing, offices, and factories, for labor, and for public and private services, including those of the transport and communications systems which knit these other functions together.

Because of the paramount importance of resource allocation, and the role of real or artificial markets, these interrelationships are largely governed by economic considerations—but with many differences from received economic dogma. There is a division of decision-making between the private and public sectors, necessitated in part by economies of scale and indivisibilities. In the ultimate political sense, economic efficiency is only one of several means to a bundle of humanistic ends. In analyzing behavior, especially that of households and their members, economic forces are greatly influenced and modified by social and psychological ones. Economic analysis depending on equilibrium tendencies ignores or belittles the lags in personal and organizational adjustment to change,

and the durability of investment in infrastructure, buildings, and human capital.

There is another and fundamental schism between economic descriptions of 'rational' behavior and the realities of urban life. This divergence has led to two trends in urban modeling which are only now being resolved. In economics, similar actors in similar conditions make similar decisions—people shop at the nearest center, traffic follows the shortest route, and employees seek out the nearest jobs. In reality, some actors make long trips, and roundabout ones (but with diminishing frequency) for the same purposes which might be pursued using shorter and more direct trips. The 'rational' behavior is defined in economic models of optimizing choices, while diverse behavior is described, but not explained, in a number of ways by 'gravity models', 'entropy maximization', and 'discrete choice'—three very similar approaches.

By far the most important issue in model development has been this antinomy between an economic view that posits uniform behavior and a humanistic view which posits diversity. These two views underlie different lines of model development, and we will follow their reconciliation, though a recognition of ignorance—a reconciliation which clears the way for the resolution of other but less important difficulties.

2. *Inventing a Paradigmatic Model*

In the early 1950s a number of factors favoring the discovery of a new paradigm in urban modeling came together: on the ground, through the suburban redirection of American urbanism, and planning and building urban expressways; in technique, through the optimism of instrumental rationalism, seen in Operations Research and Management Science and buttressed by the commercial availability of computers; and in planning, through the widespread influence of the University of Chicago Program. In the middle of the decade, urban land planning began a slow movement in the direction of modeling (see Voorhees 1959), and the Chicago Area Transportation Study (CATS 1959) developed an innovative approach, drawing on many sources, which in retrospect must be considered a new paradigm for urban modeling.

A new paradigm in scientific or professional matters involves a framework for studying phenomena and defining their relationships. The importance of such innovation does not imply perfection; implementing and extending the paradigm and overcoming its imperfections may require decades. The essential novelty of this paradigm for analyzing the critical aspects of urban settlement systems extends far beyond transportation itself, and embodies the following elements:

(a) Examine the system in all relevant detail, with massive data collection and thorough analysis of user behavior and system responses. (This was done in transportation studies by large-scale sampling of travel behavior and traffic conditions.)

(b) Simulate user behavior over the entire system or subsystem, together with the system response, and with feedback in which the system response influences the user behavior. (The simulations in CATS were inventive and completely novel.)

(c) Draw heavily on other disciplines that can contribute to the accuracy and realism of the simulations; use exact formulations wherever possible, and implement data management and simulation through computers where this is necessary and relevant. (CATS employed sociologists, planners, economists, engineers, and computer experts—and used computers for its most advanced innovations.)

Similar simulations could be imagined, and indeed have been undertaken, for the housing market, the labor market, the location of economic activity and public services, and the environment. Clearly, these subsystems are mutually interdependent, and in the final analysis their simulations would have to interact.

Despite its seminal influence, the CATS study has often been denied its paradigmatic role because of its shortcomings, most obviously inadequate attention to transit and failure to take into account the reaction of land uses to transportation. It shares many other problems with land-use modeling, as we will see in exploring them. It also embodied the schism between economics and behavior: in connecting residences and workplaces by trips it used a gravity model of dispersed behavior, but in assigning trips to the highway system it used the deterministic idea of taking only shortest routes. As a paradigm for transport analysis, it has been and is being reworked; in principle its methods can be traced even in divergent but related urban models.

3. *Residential Models: Two Directions and Their Linkages*

The major steps in land-use modeling are easily summarized. A series of theoretical models, initially of residential location, were generated by Wingo (1961), Alonso (1964), and Mills (1972)—with other contributions. These models all assumed a monocentric city (having a single employment center), which produced rings of settlement, each occupied by a single residential income class. Density declined and income rose with distance from the center. Residents were trading off commuting costs against space, under the constraints of their income. Mills's model dealt with zones in a linear programming framework, while the other models used continuous space.

The rigor and simplicity of these models inspired efforts to make them operational, hindered by their restrictive assumptions. Herbert and Stevens (1960) developed very early the idea of bid-rents (i.e., speci-

fications of willingness to pay for various types of housing in various locations), and postulated a multicentered city with competition for land in a linear programming format. They did not operationalize the measures of preferences that would underlie a bidding process, or the connection between place of residence and place of work. The National Bureau of Economic Research (Ingram et al. 1972) extended this work in a larger operational form, with actual employment centers, a diversity of choice of spatially distributed housing submarkets, and a final determination of residency by linear programming based on the cost of work-to-home travel. Building, remodeling, and redeveloping housing was in a separate supply model.

The alternative route of replicating consumer choice in housing and shopping through the use of a gravity model (that specified a declining probability of choice with increasing distance) was more realistic and more appealing to planning practitioners, but was initiated at the cost of disregarding some economic realities. The Lowry Model (Lowry 1964) thus gave the appearance of modeling the behavior of all locators (including residents and their choices), and gave room for diversity through a gravity model. The same approach was widely adopted; for a review, see Goldner (1971). It was popularized by Putman (1982) in the USA, and in structure planning and by Echenique in the UK (reviewed in Echenique and Owers 1994). Most of these efforts went beyond modeling residential choice to consider other locational trends, and other improvements were introduced to make the Lowry Model still more realistic. This model was thus a major contribution to the paradigm shift, by its introduction of a partially unified system of activity location.

The stage was set for a new type of advance that tended to bring these two trends much closer together. The necessity for multicentered models in applied analysis was firmly established. The NBER had introduced a certain amount of diversity through a probabilistic choice of housing markets. Echenique had realized the necessity for constraints on residential location, and began to introduce rents as an incentive for supply and a brake on choice in the face of income limitations. Three important events in the 1970s solidified these advances. In Geography, Wilson (1970) explained the gravity model by analogy with entropy (or constrained randomness) in physics. In Economics, MacFadden (1973) formalized the idea of 'discrete choice', which made diverse behavior acceptable to economists; the gravity model became a special case of this accepted theory. Independently, Anas (1975) showed in a seminal paper that interaction under a gravity model could be constrained by ability to pay in a competitive housing market, leading to an economic equilibrium solution, with market clearing.

All three models—gravity, discrete choice, and maximum entropy—can produce mathematically identical results. There is a common recognition that there are aspects of behavior which are not represented in models, and which contribute to a kind of statistical error in any deterministic explanation. The degree of error varies, and may or may not be susceptible to resolution by further analysis.

4. *Complete Models*

While residential choice models are the most important and perhaps the most interesting, a complete model as described at the outset would be a model suite containing numerous submodels. First, land development and its uses are determined by way of locational decisions that in turn are strongly influenced by transport considerations. Putman and Echenique have taken the lead over the last three decades in driving this lesson home—at least to the transport planning community. We will assume that any complete land-use model will embody or interface with an appropriate transportation model and use a forecast of transportation system changes. Second, even the most advanced residential models suffer from some of the general difficulties discussed below, and in particular, models of residential housing supply are weaker than those of demand. Third, models are needed for all land uses, most particularly manufacturing industry and trade and services. With the increasing shift in the overall economy toward business and consumer services, these sectors and their representation in models need thorough review.

There is a modest number of complete model suites publicly available worldwide (see Wegener 1994 for a review). Putman's ITLUP and EMPAL models have been extended to METROPOLIS, which is widely used, and maintains the Lowry tradition of noneconomic constraints. Its employment location is somewhat descriptive in style. Echenique has developed an elaborate input–output framework to generate goods traffic as well as personal trips, with extensive economic constraints. Anas (1987) is continuing to develop a range of economically rigorous and ambitious models, most recently for the New York Metropolitan Transportation Commission (in progress). New work is being undertaken which we cannot elaborate here, but special mention must be made of de la Barra, who follows Echenique in the Americas, and Waddell (2000), who is approaching residential location from a microeconomic viewpoint in direct collaboration with state planning and transportation studies. Most model sets are provided on a made-to-order basis, which raises their cost and severely limits their availability.

In our judgment, the characteristic nonresidential models have various weaknesses in almost all model sets. The manufacturing sector is undergoing changes that are not yet fully understood and which may upset conventional wisdom. Business services have barely begun to be studied in the locational literature, let

alone embodied in models. Retail trade is the best understood of these three fields, but inadequate use is being made of existing knowledge.

5. *Comments on the State of the Art*

It is apparent that considerable advances have been made in urban models, but that results are still short of what might be desired. There are still no established scholarly standards in the field, nor is there a clear link between research and practice. Limited tests of models on the same or similar data sets show little unanimity in their conclusions.

There are many social and equity issues being newly faced in planning of all sorts. Some of these have appeared in the USA in the implementation of the clean air act and its amendments, and related issues in transport planning have been litigated. It appears that there is inadequate disaggregation in most models, so that the behavior of minority groups and their reactions to policy inputs could be studied, while some aspects of disaggregation affect the evaluation of transport and planning policy. These needs conflict with the needs for simplicity, transparency, and speedy computation which arise in public discourse.

With increased computer capabilities, microsimulation is appearing as a possible approach to the disaggregation problem, but its relation to the speed of computation must be studied. Indications in the transportation planning field are that it will extend greatly the time required for analysis—and even today transportation models are the biggest bottleneck in operational land-use models.

The growth of urban modeling is linked closely with the expansion of computational power through both mainframe and personal computers. The first steps in modeling were conceptual, and computationally limited in their demands. Growing computer power has made it possible to develop and implement more accurate and more elaborate models. Computing for planning has been decentralized to the desks of planners and administrators. Computer users have readier access to large data sets, now organized in Geographic Information Systems. Research workers of all types have greatly increased power to investigate localized and sectoral phenomena, having better access to computation, data, and modeling methods. All of these investigations contribute to a web of knowledge that expands understanding of urban phenomena, and undergirds improved modeling. Examples include, the work of Anas and Waddell discussed above, as well as of specialists like Landis, Cevero et al. (1999), and Shen (2000).

Bibliography

Alonso W 1964 *Location and Land Use*. Harvard University Press, Cambridge, MA

Anas A 1975 The Empirical Calibration and Testing of a Simulation Model of Residential Location. *Environment and Planning A* **7**: 899–920

Anas A 1987 *Modeling in Urban and Regional Economics*. Harwood, Chur, Switzerland

Burgess E W 1927 The determinants of gradients in the growth of the city. *Publications, American Sociological Society* **21**: 178–84

CATS 1959 *Final Report, Vol. 1*. Chicago Area Transportation Study, Chicago

Cevero R, Rood T, Appleyard B 1999 Tracking accessibility: Housing opportunities in the San Francisco Bay Area. *Environment and Planning A* **31**(7): 1259–78

Echenique M, Owers J (eds.) 1994 Research into practice: The work of the Martin center in urban and regional modeling. *Environment and Planning B* **21**(5)

Harris C D, Ullman E L 1945 The nature of cities. *Annals of the American Academy of Political and Social Sciences* **242**: 7–17

Herbert J S, Stevens B H A 1960 Model of the distribution of residential activity in urban areas. *Journal of Regional Science* **2**: 21–36

Hoyt H 1939 *The Structure and Growth of Residential Neighborhoods in American Cities*. Federal Housing Administration, Washington, DC

Goldner W 1971 The Lowry model heritage. *Journal of the American Institute of Planners* **37**: 100–10

Ingram G K, Kain J F, Ginn J R 1972 *The Detroit Prototype of the NBER Urban Simulation Model*. National Bureau of Economic Research, New York

Lowry I S 1964 *A Model of a Metropolis, RM-4035-RC*. The Rand Corporation, Santa Monica, CA

McFadden D 1973 Conditional logit analysis and quantitative choice behavior. In: Zarembka P (ed.) *Frontiers in Econometrics*. Academic Press, New York, pp. 105–42

Mills E S 1972 *Urban Economics*. Scott Foresman, Glenview, IL

Putman S H 1982 *Integrated Urban Models: Policy Analysis of Transportation and Land Use*. Pion, London

Shen Q 2000 Spatial and social dimensions of commuting. *Journal of the American Planning Association* **66**: 68–82

Voorhees A (ed.) 1959 *Journal of the American Institute of Planners* **25**(1)

Waddell P 2000 A behavioral simulation model for metropolitan policy analysis and planning: Residential location and housing market components of UrbanSim. *Planning and Design*, (*Environment and Planning B*) **27**: 247–64

Wegener M 1994 Operational urban models: The state of the art. *Journal of the American Planning Association* **60**: 17–29

Wilson A G 1970 *Entropy in Urban and Regional Modeling*. Pion, London,

Wingo L Jr 1961 *Transportation and Urban Land Use*. Johns Hopkins University Press, Baltimore, MD

B. Harris

Urban History

Urban history started as early as the nineteenth century in Western Europe. Italian scholars had a deep concern with the Commune, the governing body of the medieval city–state in connection with the unification process of the country. French historians had a similar concern with the commune movement

which, over the centuries, had developed into town autonomy against the rule of a feudal lord, and also with freedom of citizens of medieval towns. German scholars showed interest in freedom and equality of townspeople of the medieval towns as a forerunner of democracy. In England during the modernizing process of local government, especially in connection with the reform of municipal corporations, historians had some interest in historical privileges of citizens given in the charters granted either by the King or by feudal lords, lay or ecclesiastical. From the nineteenth century to the early twentieth century, the historians' main interests were in political, legal, and/or constitutional problems.

1. Regional Studies

The new trends of urban historians' concern after World War II are more social, economic, physical, or topographic. There were two opposing opinions as to how to reconstruct historic towns which suffered serious damage by bombing during the War, either rebuilding them exactly as they were before or rebuilding them entirely anew according to new town planning and a new scheme. Moreover, as built-up areas are impossible to excavate, this offered great possibilities for archaeological excavation in areas destroyed by the bombing, especially in German and English towns and cities. Archaeological investigation explored not only medieval towns but also the deeper stratum of ancient Roman cities.

During the 1960s, either construction of new towns or redevelopment of the slum areas of nineteenth century industrial towns became urgent matters, and urban historians were concerned with the problems of new towns or the process of building towns during the Middle Ages, as clearly observed in Maurice Beresford's *New Towns of the Middle Ages, Town Plantation in England, Wales, and Gascony*, published in 1967.

Organized research in urban history began in Germany at the end of the 1950s, in England during the 1960s, and in Japan and the USA in the early 1970s.

In West Germany, two Study Groups were organized: *Arbeitskreis für landschaftliche deutsche Städteforschung* as a national organization and *Arbeitskreis für südwestdeutsche Stadtgeschichtsforschung* as a local organization. They were started in 1959 and 1960, respectively. In Münster, *Institut für vergleichende Stätdegeschichte* was established by H. Stoob, and began to publish its journal, *Städteforschung*, in 1970.

In England, the 'Urban History Group' was organized in 1963 at the General Meeting of the Economic History Society, and H. J. Dyos issued the *Urban History Newsletter* until 1973. From 1974 to 1991, 18 volumes of the *Urban History Yearbook* were published by Leicester University Press. Since 1991, as the continuation of the *Urban History Yearbook*, *Urban History* has been published semi-annually by the Cambridge University Press.

In 1986, The Centre for Urban History was established at Leicester University and the new second series of the *Urban History Newsletter* was published until 1999. The third series of the *Urban History Newsletter* started in Autumn 1999, when Richard Rodger was appointed Professor of Urban History at Leicester University.

In London, in 1975, the Metropolitan Society was founded and a new Journal, *The London Journal, The Review of Metropolitan Society Past and Present*, was launched. This journal was 'planned to combine the work of historians and of scholars in other fields, for the first time presenting a more complete study of the whole region in its historical, its societal, and its cultural framework.' Two years later, at the Institute of Historical Research, University of London, the Centre for Metropolitan History was founded by Derek Keene and *The Newsletter of the Centre for Metropolitan History* was issued in 1992.

In 1987, The Urban Morphology Research Group was founded at the Department of Geography, University of Birmingham, and the *Urban Morphology Newsletter* was issued until 1997 when it was succeeded by *Urban Morphology, Journal of the International Seminar on Urban Form*.

In Japan, The Tokyo Study Group in Comparative Urban History was formed with six charter members in 1971, and issued *Urban History Newsletter* from 1975 to 1981. Since 1982, *The Comparative Urban History Review* has been published semi-annually. In 1991, the Study Group published two volumes of essays entitled *Toshi to Kyodotai [Stadt und Gemeinde]* in commemorating its twentieth anniversary, and also *Edo and Paris, Urban Life and the State in the Early Modern Era* was published by Cornell University Press in 1994, and *Edo and Paris*, the Japanese edition, was also published in 1995. These publications were outcomes of the Symposium held in Tokyo in 1990 to commemorate 400 years of the beginning of *Edo*. In 1978, the Society of Hokuriku Urban Historians was launched as a local organization by the initiative of T. Toyoda, and the Society holds annual meetings at three historic towns; that is, Kanazawa, Toyama, and Fukui, alternately. *The Hokuriku Urban History Review* appeared from 1979 to 1989, and *The Journal of the Society of Hokuriku Urban Historians* has been published since 1990.

In September 1974, the *Journal of Urban History* was established 'to provide urban scholars in the USA with a forum for interpretation and new research.'

2. Urbanization

There are four types of cities in history.

The first historical type of city appeared in the ancient Orient, namely in Mesopotamia and Egypt, in

ancient India, and China. Walled towns with the surrounding territories were the center of theocratic rule of oriental despots. Officials of despots, clergy of temples, and wealthy merchants lived in those cities, which usually were located near palaces or temples, and were surrounded by quarters of subordinating poor tribes. They flourished with long-distance trade. However, there were no citizens as such.

The second type of city appeared in the ancient Greek and Roman worlds. They were city–states called either *polis* or *civitas*. Cities were formed by land-owning warriors settling in militarily defensible places, that is, acropolis. Rome was founded on seven hills, and by conquering the neighboring *civitas* became finally the Roman Empire ruling the Mediterranean World. Colonial cities were founded at strategic points in various provinces with a similar city plan. Agricultural estates or mines organized with slave labor, supported by long-distance trade, brought prosperity and peace, that is, *Pax Romana* to the Empire.

The third type of city, the medieval city, appeared in the area beyond the Alps; namely, Germany, France, the Low Countries, and England. Max Weber characterized ancient cities as consumers' towns, and medieval towns as producers' towns, because citizens of medieval towns consisted mainly of tradesmen and craftsmen. They were enjoying the privileges of freedom of citizenship, and they were enabled by feudal lords to form guilds and to constitute a self-governing body; that is, a mayor and the town council. These medieval cities located in areas that had facilities for water transportation, such as ports on estuaries or river banks. Emperors, kings, and feudal magnates, lay or ecclesiastical, granted these towns privileges of holding weekly markets and yearly fairs, as nexuses of long-distance trade as well as local trade.

The fourth type of city is called a modern or industrial town. As the result of the Industrial Revolution, a new type of city appeared. In England, at the earlier stage of industrialization, factories with machine systems were set up in dales where water power was available. However, after the steam engine was invented, factories were established at places near coal and iron mines, and/or near cities which could serve for the domestic market as well as overseas trade. The earliest of factories were abandoned very quickly, but remained as objects for industrial archaeologists' research. Around the new type of factories, developers constructed cheap houses which accomodated wage laborers and their families. These houses were called 'back to back,' with poor sanitation systems, and later became poor slum areas. On the other hand, in the West End of London, in the suburbs of provincial towns, and in the resort towns such as Bath, new developments of landed estates started during the latter half of the eighteenth century, and constituted fashionable town houses not only for nobilities but also for the new middle-class. Although the size of these new type towns were limited by the system of transport within them, the population of cities increased significantly. This process was called the first stage of urbanization, and lasted until World War I.

After World War II, the so-called second urbanization process started not only in the highly industrialized countries but also in 'developing countries.' As the result of the rapid motorization and efficient communication systems, not only the size of cities expanded, but also the population of cities increased enormously. Huge cities appeared connected with satellite cities constituting a megalopolis.

Thus, urbanized society offered people amenities. However, the rapid urbanization brought pollution problems as well as poverty and misery, causing urban historians deep concerns.

3. *International Cooperation*

One of the distinct characteristics of recent research in urban history is international cooperation. In 1955, when the International Congress of Historical Sciences was held in Rome, European urban historians established *La Commission Internationale pour l'Histoire de Villes*. The first president of the Commission, Hermann Aubin [1955–8] was succeeded by Hektor Amman [1958–67], Phillipp Wolff [1967–81], and Sergi Vilfan [1981–91]. Adriaan Verhulst has been the president since 1991.

The Commission presents a report to the Congress as the result of five years' preparation by the members of the Commission. The overall theme at the XIIIth Congress in Moscow, 1970 was 'Urban Autonomy,' at the XIVth Congress in San Francisco, 1975, 'Social Structure and Urban Morphology,' at the XVth Congress in Bucharest, 1980, 'Central Government and the Towns in Europe,' at the XVIth Congress in Stuttgart, 1985, 'Town and Country,' at the XVIIth Congress in Madrid, 1990, 'New Towns in Europe (XII–XX century),' at the XVIIIth Congress in Montreal, 1995, 'Markets and Fairs,' and at the XIXth Congress in Oslo, 2000, 'Destruction and Reconstruction of Towns.' The list of the themes shows not only the results of international cooperation but also the trends of historians' interest. Along with internationally organized common research and reports, the Commission has undertaken three joint projects—the compilation and publication of urban bibliographies by country, of selected urban history texts (*Elenchuses*), and historical atlases of towns.

The urban bibliographies were published during the late 1960s and early 1970s. The earliest bibliography was by French historians Philippe Dollinger, Philippe Wolff, and Simon Guenée, *Bibliographie d'histoire des villes de France* in 1967. It was followed by German historians Erich Keyser's *Bibliographie zur Städtegeschichte Deutschlands (Acta Collegii historiae urbanae Societatis historicorum internationalis)* in

1969. The last one was by English historians G. H. Martin and Sylvia C. McIntyre, *A Bibliography of British and Irish Municipal History*, Vol. I in 1972.

Elenchus is a collecion of selected sources to illuminate the history of towns. The first volume was published in 1967, *Elenchus fontium historiae urbanae, Volumen primum, quem edendum curaverunt C. Van de Kieft et J. F. Niermeijer (Acta Collegii historiae urbanae Societatis historicorum internationalis)*. In 1977 a guide book was published, *Guide international d'histoire urbaine, I Europe, préparé par la Commission sour la direction de Philippe Wolff, Membre de l'Institute de France*.

Publication of a historical atlas has become a more enterprising and widespread project. Although there was fundamental agreement on a unified concept, the actual outcome has been very diverse, and is unique country by country.

The earliest of the atlases to reach fruition was the British one, with a first volume of *Historic Towns. Maps and Plans of Towns and Cities in the British Isles, with Historical Commentaries, from the Earliest Times to 1800* in 1969. This was the outcome of very close collaboration of M. D. Lobel, urban historian, and Colonel W. H. Johns, cartographer of the Ordnance Survey. The first volume consisted of historical atlases of seven British historic towns bound in one volume of folio size.

The fundamental process of drawing historic town maps is as follows: on the contemporary map of scale 1:2500, modern buildings and constructions were eliminated piece by piece from a map and the town map reconstructed as it was in 1800; that is, the state of the town before industrialization. Then demolished buildings and perished constructions are added on a map based on archaeological findings as well as other information obtained from the old cadastre maps and other literary sources. As a result of these painstaking processes, medieval towns and even Roman towns were shown on large-scale colored maps, with commentaries by historians.

In the preface to this masterpiece (*Historic Towns*), Professor Wolf, as the president of the Commission, made the comment:

> The legitimate curiosities of scholars are not, however, the only interests to be involved. Nor is it only lovers of the past who will be glad to get a faithful picture of the ancient 'heart' of towns. Ours is an era of extraordinary town development. Tomorrow's 'urban textures' run a great risk of losing in charm and in human value what they gain in expansion. Our children run a great risk … of growing up joyless in soulless cities. The old centres of our towns, where the most precious of the ancient buildings cluster, make up the element peculiar to each town, by which it is differentiated from every other. This is a touristic treasure-house, … What is more, it is a human treasure-house, and one which any concern for the future bids us reconstitute and save: atlases of urban history provide an indispensable tool for this work—and only an exceedingly short-sighted realism could prevent us from appreciating the work's importance.

The second volume was published in 1975 with the same format for four more English towns. The third volume was published in 1989 for the City of London from prehistoric times to c.1520.

The second winner of the international competition of publication of atlases of historic towns was German. Four series of historic town atlases were organized in Germany: (a) *Deutscher Städteatlas* published five volumes to 1993; (b) *Hessicher Städteatlas* published one volume in 1998; (c) *Rheinischer Ständteatlas* published 12 volumes until 1996; and (d) *Westfälischer Städteatlas* published five volumes to 1997. Each set consisted of a reconstructed color map on a scale of 1:2500, an old map of the nineteenth century, an air photograph of the present day towns, old prints and historic photographs of towns with present photographs from the air, and a leaflet of the commentaries by historians, making the atlases more attractive to general as well as local readers.

In Austria, the *Österreichischer Städteatlas*, five volumes [1982–97] and *Historischer Atlas von Wien*, five volumes [1981–94] have been published under joint editorship of *Wiener Stadt und Landesarchiv* and the *Ludwig-Boltzmann Institut für Stadtgeschichtsforschung*.

In France, the *Atlas historique des villes de France* has been published under the editorship of Centre Charles Higounet, *Centre de recherches sur l'occupation du sol et du peuplement*, at the Université Michel de Montaigne, Bordeaux since 1982 containing 42 towns in various provinces.

In Italy, two titles of historic atlases of towns have been published; that is, *Atlante storico della cittá italinae (Italia settentrionale et Sardegna)* and *Atlante storico della città italiane (Italia centrale e meridionale)* since 1986, containing four and 14 towns, respectively.

Other European countries are also editing and publishing atlases of historic towns of their countries, so that more than 300 towns in Europe had an atlas by the end of the millennium.

In 1989, the European Association of Urban Historians was organized and the Association is holding biennial conferences. The first conference was held in Amsterdam in 1992, the second conference in Strasbourg in 1994, and the third conference in Budapest; the overall theme was Cities in Eastern and Western Europe. The fourth conference was in Venice in 1998 when the theme was Cities in Europe: Places and Institutions. In 2000, the fifth conference was held in Berlin and the theme was European Cities: Networks and Crossroads. The sixth conference will be held in Edinburgh in 2002.

4. *Multidisciplinary Research*

Another feature of recent trends of urban history is inter- or multidisciplinary research. The Society for Medieval Archaeology was founded in 1957, and publishes a Journal, *Medieval Archaeology*, annually.

In 1969, a small number of British historical geographers formed a group in order to collect material for an International Glossary of Agrarian Terminology, and in 1973 the name of the Historical Geography Research Group was adapted and formally recognized by the Institute of British Geographers. Since 1975, the *Journal of Historical Geography* has been published by the Academic Press, London, under the joint editorship of John Patten (Oxford) and Andrew Clark (for the Americas).

A Planning History Group was established in the Centre for Urban and Regional Studies at the University of Birmingham in 1974, and issued the *Planning History Bulletin* from 1979 to 1987, but when the editorial team changed, the title of the Bulletin changed to *Planning History, Bulletin of the Planning History Group*. Then, from 1993, *Planning History* became the *Bulletin of the International Planning History*.

In 1989, Gordon E. Cherry and Anthony R. Sutcliffe launched a new journal, *Planning Perspectives, An International Journal of History, Planning and the Environment*.

In Italy, since 1976, *Storia Città, Rivista Internationale di storia urbana e territoriae* [International Review of Town Planning History] has been published irregularly.

Another discipline is architectural history. It is deeply connected with town planning and also townscapes of different periods. In Japan, the Society of Architectual Historians of Japan was established in 1983, and is publishing the *Journal of the Society* semi-annually.

Interdisplinary research is organized between different disciplines and is flourishing. However, the most remarkable contribution to urban history has been medeival archaeology. Not only in European countries but also in Japan, excavations of ancient and medieval towns have been carried out extensively.

In Japan, not only Kyoto and Nara, ancient capitals, but also Ichijo-dani, a medieval castle town constructed in 1472 and destroyed by war in 1573, and so-called One Thousand Houses in Kusado, an ancient market town lost by flood in 1673, are distinguished examples of excavation by medieval archaeologists. Archaeological findings were examined with contemporary literary sources, and clarified the details of town plan, economic activities, social life of the day, and even the mentality of townfolks.

In Europe, Germany, England, France, and Scandinavian countries, the results of research on medieval archaeology are remarkable. However, the most distinguished case of cooperation between archaeologists and historians must be the investigation of Anglo-Saxon and medieval Winchester. *Winchester Studies 1*, the first volume of the series of publications was published in 1976 entitled *Winchester in the Early Middle Ages: An Edition and Discussion of the Winton Domesday* [surveyed in 1148]. Two volumes of *Winchester Studies 2, Survey of Medieval Winchester* were published in 1985 and part two of *Winchester Studies 3, Pre-Roman and Roman Winchester* was published in 1979. Still more volumes are to be published.

5. Conclusion

Research into urban history has tended to be concentrated on Western or European countries. China has a long and unique history of towns and cities. However, these never were given any autonomy of their own, and therefore they remained as mere centers of central or local administration. So it makes comparison with European towns difficult. As to other Asian countries, the processes of building new towns under colonial rule such as New Delhi, Rangoon, or Bangkok have been much clarified. However, the urban history of their predecessors, Old Delhi, the Mogul city, Yangon, a port town of Burma, or Ayuttha, the old capital of Siam, should be further explored. The other unexplored field of urban history is towns and cities within the Islamic world.

See also: Cities: Cultural Types; Human Ecology; Industrialization; Urban Sociology; Urban System in Geography

Bibliography

Barley M W (ed.) 1977 *European Towns, Their Archaeology and Early History*. Academic Press, London/New York/San Francisco

Barry J (ed.) 1990 *The Tudor and Stuart Town: A Reader in English Urban History, 1530–1688*. Longman, London

Beier A L, Finlay R (eds.) 1986 *London 1500–1700: The Making of the Metropolis*. Longman, London

Borsay P, McInnes A (eds.) 1990 *The Eighteenth Century Town: A Reader in English Urban History 1688–1820*. Longman, London

Clark P (ed.) 1981 *Country Towns in Pre-Industrial England*. Leicester University Press, Leicester, UK

Clark P (ed.) 1984 *The Transformation of English Provincial Towns*. Leicester University Press, Leicester, UK

Clark P 2000 *Cambridge Urban History of Britain*. Cambridge University Press, Cambridge, UK, Vol. II

Corfield P, Keene D (eds.) 1990 *Work in Towns 850–1850*. Leicester University Press, Leicester, UK

Daunton M 2000 *Cambridge Urban History of Britain*. Cambridge University Press, Cambridge, UK, Vol. III

Engeli C, Matzerath H (eds.) 1989 *Modern Urban History Research in Europe, USA and Japan. A Handbook*. Berg, Oxford/New York/Munich

Ennen E 1979 *The Medieval Town* (trans. Fryde N). North-Holland, Amsterdam/New York

Holt R, Rosser G 1990 *The Medieval Town: A Reader in English Urban History 1200–1540*. Longman, London

Morris R J, Rodger R (eds.) 1993 *The Victorian City: A Reader in English Urban History 1820–1914*. Longman, London

Palliser D M 2000 *Cambridge Urban History of Britain*. Cambridge University Press, Cambridge, UK, Vol. I

Pirenne H 1925 *Medieval Cities, Their Origins and the Revival of Trade*. Princeton University Press, Princeton, NJ
Platt C 1976 *The English Medieval Town*. Secker & Warburg, London
Reynolds S 1977 *An Introduction to the History of English Medieval Towns*. Clarendon Press, Oxford, UK
Whitehead J W R (ed.) 1981 *The Urban Landscape*. Academic Press, London

K. Ugawa

Urban Life and Health

In Western countries where tabulations have been made, such as the USA and the UK, city residents face shorter life expectancies than their rural cousins. The differential was great in the early days of urbanization, when the leading causes of death were infectious diseases that spread readily in densely crowded environments (Wiehl 1948). In 1900, death rates exceeded birth rates in many cities, with urban population growth occurring only as a result of migration from rural areas (Fox et al. 1970). By the mid-twentieth century, with sanitary and public health measures in place, antibiotic treatments and vaccines averting deaths from infectious disease, and financial prosperity far-reaching, the urban mortality disadvantage ran at about 5 percent, after adjustment for differences in the age, race, and gender composition of urban compared to rural populations (Kitigawa and Hauser 1973). However, by the late twentieth century, a sizeable urban mortality disadvantage re-emerged, especially in high-poverty areas, with urban mortality rates running about one-and-a-half to three times those in rural areas or small towns, even after adjustment for age, race, gender, and socioeconomic status (House et al. 2000, Geronimus et al. 1999). Currently, urban residents have higher prevalence rates of many health problems, among them preterm birth, asthma, infectious disease, cancer, cardiovascular disease, and lead poisoning. City residents also suffer disproportionately from pervasive ambient stressors such as noise, traffic, pollution, and crowding, as well as social problems such as homelessness, suicide, and violence. Ironically, urban areas have also been the place of origin for major medical and public health initiatives, including great advances in public sanitation, clean water systems, and medical technology (Freudenberg 1998).

1. *Understanding Urban Health Disadvantages*

1.1 *Methodological Limitations*

Social scientific consensus on the underlying reasons for apparent urban health disadvantages remains elusive. Part of the ambiguity stems from methodological issues. Empirical estimates generally suggest urban residents are worse off than others, at least in terms of mortality, yet the body of research is largely descriptive. Most research on this topic has been cross-sectional, much of it ecological, and investigators have had to make unfortunate trade-offs between analyzing samples with national scope and focusing on important subgroups of the urban population. This data limitation along with another—the inability to control for many potential confounders of the association between urban life and health—severely restrict causal inference making.

1.2 *Early Theories*

Conceptually, the study of urban life and health has changed markedly over the past century. Initially, rapid urbanization in the early twentieth century stimulated theories reflecting broader societal concern that city life was dangerous and unhealthy compared to pastoral country life. Wirth (1938), posited that characteristics of cities—such as their size and population density, their fast pace, their heterogeneity—were negative influences on psychological well-being and group life. The requirements of urban industrial organization and technology were also viewed as deleterious. In these theoretical perspectives, urban life alienated individuals from one another, their roots in nature, their labor, and their core values; it promoted superficial and transitory relationships; and it provoked anomie, alienation, and insecurity. These influences devalued community, threatened social cohesion, and placed individuals in peril.

Empirical studies often found weak endorsement for the working hypotheses that followed from this conceptual emphasis (Fischer 1976). Researchers continuing in the sociopsychological tradition in the mid- to late twentieth century saw urban characteristics in a new light. Milgram (1970) reasoned that brusque behavior and anonymity might demonstrate an adaptive response to the sensory overload produced by urban population density and a fast-paced life, rather than imply that urban residents were without intimate, supportive relationships within their personal social networks. Studies documented the existence and salience of local social bonds within cities among 'urban villagers' (Gans 1962) embedded in kin 'networks' (Stack 1974). Cities were reinterpreted more positively as mosaics of individual and cultural differences, even a refuge for individuals who faced stigma or persecution in small parochial towns (Fischer 1976). House and Wolf (1978) observed that the effects of urban residential environments on behavior grew out of particular social settings at different historical times, not necessarily urbanism, *per se*. This observation was consistent with their empirical findings and with the variation in magnitude of urban health disadvantages in different time periods.

1.3 Recent Developments

In the 1980s, as urban homicide rates increased and were highly publicized, urban life again became viewed as unhealthy and dangerous. By this time, poverty was concentrated in central cities, urban decay left poor neighborhoods as staging grounds for drug trafficking and related homicide deaths, while infectious diseases such as tuberculosis, and in specific cities HIV/AIDS, re-emerged as important causes of death (Geronimus 2000). Though less publicized, some environmentally induced and stress-related diseases also were increasing, such as asthma, reproductive disorders, some cancers, and cardiovascular disease. Variations on early sociopsychological theories took shape in scholarship on urban life and health. For example, the 1990s witnessed renewed interest in the effects of depleted social cohesion or social capital on health, with social epidemiologists joining social psychologists in studying these topics, particularly among urban residents. The actual significance of social capital to health disparities is currently subject to debate (Lynch et al. 2000) and agreement has grown that it alone cannot explain the severely disproportionate ill health and early mortality experienced by some segments of the post-industrial urban populace. In addition, as the growth of edge cities and globalization has shrunk the urban labor market, the notion that the impersonal nature of industrial organization is a key cause of urban health disadvantages has been superseded by new concerns about the health effects of unemployment, segregation, poverty, ambient psychosocial stresses, and the physical environmental hazards of urban life.

1.4 Social Inequality

Evidence suggests that urban areas confer some disadvantages, at least in terms of mortality, on all of their residents, but this penalty is inequitably divided among the haves and have-nots within cities (Geronimus et al. 1996, Shaw et al. 1998). This is unsurprising since the social and psychosocial factors that have been linked to morbidity and mortality, including material hardships, acute and chronic stress, and overburdened or disrupted social supports, are suffered disproportionately by those in poverty (Williams and Collins 1995). Those living in poverty also have more difficulty accessing information, services, medical care or technologies that could abate health risks (Link and Phelan 1996). Much current research is informed by this social patterning of the distribution of health risks or protective factors. Even the ambient stressors to which all residents of major cities are exposed are not equitably distributed among urban neighborhoods. Household crowding and environmental toxicity are more severe in poor, ethnic enclaves than in affluent segments of cities, for example (Mohai and Bryant 1992). Such findings signal the need to broaden the intellectual context for the study of urban life and health by looking at the socioeconomic and demographic landscape of modern cities, and coming to understand the historical and structural underpinnings of that landscape. For example, some investigators now seek to describe how economic restructuring combined with the segregation of ethnic minority groups into urban enclaves served to concentrate poverty in neighborhoods characterized by urban decay, with housing stock, municipal services, and public spaces all substandard (Massey and Denton 1993, Kelley 1997). This new context will permit examinations of how social and economic policies affect the physical, social, and built environment in ways that impact health.

In the USA, low-income African Americans segregated into central cities suffer the worst health disadvantages of any identifiable group (Hart et al. 1998). African American residents of high-poverty urban areas suffer rates of early mortality that are staggering. For example, male youths in Harlem or Chicago's South Side face a lower chance of survival to age 45 than their counterparts nationwide enjoy of surviving to age 65 (Geronimus et al. 1999). Such mortality profiles are similar to those found in less developed nations (McCord and Freeman 1990). In European cities, minority and low-income populations are often fragmented and dispersed along city peripheries where they, too, suffer demonstrable health disadvantages compared to the more affluent who reside in vital urban centers (Shaw et al. 1998). These observations suggest that any theory that hopes to illuminate the association between urban life and health in the post industrial period must explicate the historical and structural factors that produced modern ghettos with predominantly minority populations (Thompson 1998).

Moreover, although homicide deaths account for an important share of deaths among urban African American men, chronic disease deaths have been the primary influence on these rates for urban African American men and women. Indeed, homicide rates declined in the 1990s for this population, yet the influence of chronic disease on excess urban mortality appears to be growing. No one understands why. However, these findings place new emphasis on understanding the particular stresses faced by and the coping strategies available to ethnic minority residents of high-poverty urban areas. They also suggest the prospect of intensifying environmental hazards. Inequitable distribution of environmental hazards among neighborhoods may contribute to increasingly excessive rates of cancer, asthma, and reproductive disorders in high-poverty urban areas. Another possibility under consideration is that the adverse health effects of an ambient stressor or environmental toxin may be increased in interaction with other insults to health, as most often occurs in the poorest urban

communities where residents face multiple, pervasive, and cumulative exposures, with few mitigating resources.

2. Future Directions of Theory and Research

As the study of urban life and health progresses, investigators are likely to continue their examinations of social inequality, augmenting research on material conditions and poverty with improved research on the role of race/ethnicity in two fundamental aspects: (1) the role of racial discrimination (in housing, medical care, employment, and environmental exposures, for example); and (2) the ability of ethnic networks and cultures to mitigate adversarial health impacts. The 1990s witnessed the greatest influx of immigrants to the USA in almost 100 years, many into high-poverty urban areas. Documentation that the extent of ill health among low-income urban residents varies by ethnic group and nativity status has stimulated interest in developing dynamic understandings of the potential for ethnic culture to preserve health. James (1993) speculates that as a group's economic strength diminishes, its ability to supply the protection conferred by mutual aid, social support, and identity-affirming cultural frameworks may be especially critical to preserving the health of its members. According to this model, current urban health problems may derive from a combination of significant material adversity (the result of joblessness, and skyrocketing housing and medical care costs, for example) that is, at least in part associated with racial discrimination, and the concomitant overloading and weakening of protective networks (Geronimus 2000). These remain empirical questions.

More broadly, the question of what exactly it is about living in cities that may adversely affect health remains unanswered. Toward addressing this basic question, several additional empirical directions are anticipated. These include understanding how urban residence affects health across the life course both in terms of its effects on people at different life stages and also the net effect of the accumulation of health insults over time, particularly among those segments of the urban population who are further disadvantaged by race/ethnicity or low income. Another is distinguishing the effects of urban residence on morbidity and life quality separately from mortality, while developing full models of how social background factors and social experience work through physiological processes to exert an impact on morbidity and mortality. Also likely is increased attention to applied research, including demonstration projects, participatory action research, and community-based public health approaches (Freudenberg 1998). The motivations for this are diverse. Current policy interest in the labor force participation of the urban poor prompts an interest in their health and disability status. Similarly, policymakers and citizens concerned about urban sprawl may invite input on the anticipated health impact of social, economic, and urban planning policies meant to revitalize urban centers. The broader environmental movement has become concerned with environmental justice for urban populations. The appeal of applied research in the academic public health community stems from the mandate to reduce as well as understand urban health disadvantages.

See also: Community Health; Epidemiology, Social; Health Interventions: Community-based; Public Health

Bibliography

Fischer C 1976 *The Urban Experience*. Harcourt Brace Jovanovich, New York

Fox J P, Hall C E, Elveback L R 1970 *Epidemiology: Man and Disease*. Macmillan, New York

Freudenberg N 1998 Community-based health education for urban populations: An overview. *Health Education and Behavior* **25**: 11–23

Gans H J 1962 *The Urban Villagers: Group and Class in the Life of Italian Americans*. Free Press of Glencoe, New York

Geronimus A T 2000 To mitigate, resist, or undo: Addressing structural influences on the health of urban populations. *American Journal of Public Health* **90**: 867–72

Geronimus A T, Bound J, Waidmann T A, Hillemeier M M, Burns P B 1996 Excess mortality among blacks and whites in the United States. *New England Journal of Medicine* **335**: 1552–8

Geronimus A T, Bound J, Waidmann T A 1999 Poverty, time and place: Variation in excess mortality across selected US populations, 1980-1990. *Journal of Epidemiology and Community Health* **53**: 325–34

Hart K D, Kunitz S J, Sell R R, Mukamel D B 1998 Metropolitan governance, residential segregation, and mortality among African Americans. *American Journal of Public Health* **88**: 434–8

House J S, Wolf S 1978 Effects of urban residence on interpersonal trust and helping behavior. *Journal of Personality and Social Psychology* **36**: 1029–43

House J S, Lepkowski J M, Williams D R, Mero R P, LantzP M, Robert S A, Chen J 2000 Excess mortality among urban residents: How much, for whom, and why? *American Journal of Public Health* **90**: 1898–1904

James S A 1993 Racial and ethnic differences in infant mortality and low birth weight: A psychosocial critique. *Annals of Epidemiology* **3**: 130–6

Kelley R D G 1997 Playing for keeps: Pleasure and profit on the postindustrial playground. In: Wahneema L (ed.) *The House that Race Built*. Pantheon, New York

Kitigawa E M, Hauser P M 1973 *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Harvard University Press, Cambridge, MA

Link B G, Phelan J C 1996 Understanding sociodemographic differences in health—The role of fundamental social causes. *American Journal of Public Health* **86**: 471–3

Lynch J W, Due P, Muntaner C, Davey Smith G 2000 Social capital: Is it a good investment strategy for public health? *Journal of Epidemiology and Community Health* **54**: 404–08

Massey D A, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA
McCord C, Freeman H P 1990 Excess mortality in Harlem. *New England Journal of Medicine* **322**: 173–7
Milgram S 1970 The experience of living in cities. *Science* **167**: 1461–8
Mohai P, Bryant B 1992 Environmental injustice: Weighing race and class as factors in the distribution of environmental hazards. *University of California Law Review* **63**: 921–32
Shaw M, Dorling D, Brimblecombe N 1998 Changing the map: Health in Britain 1951–91. *Sociology of Health and Illness* **20**: 694–709
Stack C B 1974 *All Our Kin*. Harper and Row, New York
Thompson J P 1998 Universalism and deconcentration: Why race still matters in poverty and economic development. *Politics and Society* **26**: 181–219
Wiehl D G 1948 Mortality and socio-environmental factors. *Milbank Quarterly* **26**: 335–85
Williams D R, Collins C 1995 US Socioeconomic and racial differences in health: Patterns and explanations. *Annual Review of Sociology* **21**: 349–86
Wirth L 1938 Urbanism as a way of life. *American Journal of Sociology* **44**: 1–24

A. T. Geronimus

Urban Life: Future Developments

The future of urban settlements and the lives lived therein will depend on a mix of predictable trends and human contingencies. History is the best guide to those future trends, and three are discussed here: technology, the limits to growth at urban peripheries, and the attractive power of urban centers. Caution is perhaps the best guide to discussing future contingencies, but three major issues are so apparent that they will likely confront urban decision makers in the coming generation: large versus dense settlement structures, local versus regional decision-making processes, and automobiles versus other means of mobility and access to urban resources.

1. *Three Determinants of Urban Futures*

Since the beginning of the urban era, technologies in various forms and the relative 'gravitational' pull of urban centers and rural hinterlands have been powerful determinants of urban life. These fundamental mechanisms are surely the product of human action and are never completely beyond a human capacity to intervene (Mumford 1961). Yet, the power of such mechanisms (for example, the technology of production and the related attractiveness of center versus peripheral locations) often exerts a defining influence on urban settlements in ways that can overwhelm individual or collective action in a given time or place. Introduced below are three such large trends that may be expected to influence the future of urban life.

Waves of innovation in communications and transportation technologies occurred throughout the twentieth century. At present, these roiling changes are understood and discussed in a summary way as the emergence of an Economy (Mitchell 1995). A new digital geography will reshape the possibilities of urban life over the next generation (Kotkin and Siegel 2000). In particular, the Economy changes many of the incentives (e.g., economies of scale and various costs that increase with distance) that led to concentration in urban centers in the nineteenth and twentieth centuries. By weakening and sometimes inverting such incentives, the New Economy is characterized by much more footloose (or non-location-specific) production and consumption. It is now both possible and increasingly common to participate in employment, education and many other daily activities well beyond the commuting radius of a daily urban system. Under such conditions, the nineteenth and twentieth century comparative advantages of specific locations (e.g., entrepot for commodities, large physical agglomerations of workers and consumers) are being eclipsed by locational amenities such as weather and natural beauty as well as cumulative advantages that accrue to particular locations that happened to innovate first in some industry or activity.

Second, the limits of post-1945 development patterns, especially the low density suburban form prevalent in the United States for 50 years, will be tested and some expect those limits to be reached. Technological changes (with, in the United States, the full support and encouragement of several national policies) have maximized the opportunity for low-density choices and this has led to a half-century of suburbanization in urban settlements, in which the increase in built-up land area has far exceeded the increase in population or employment. The resulting development patterns have been associated with a long list of social costs, including increased traffic congestion and lengthening commutes, loss of both open space and agricultural land uses, a rise in social and ethnic segregation as well as in aggregate housing costs, and the decline of communal participation and social capital (Downs 1995, Kunstler 1996, Duany et al. 2000). These critiques have sought to portray as unsustainable the current low-density development regime, often labeled sprawl. Policy recommendations range from government controls on new peripheral growth to the modification of cultural and aesthetic values related to suburbanization.

Third, the attractive power of centers will wax or wane; given the diminished power and autonomy of large industrial urban centers over the preceding century, this third mechanism is likely to be closely related to changes in the first two. Paradoxically, the decentralizing trends of both the digital economy and a low-density development regime have created a

powerful role for centralized command and control exploited by a select group of 'global cities' (Sassen 1991). This centralizing function is also present on a smaller scale within regional systems, somewhat reviving a role played for smaller urban centers. In addition to this emerging economic function, urban centers may also gain a new importance derived from the limits on further growth at the urban periphery (Rusk 1999), whether those limits are imposed by technology or by governments.

2. Three Issues Facing Future Urban Decision Makers

The multiple contingencies of these and other mechanisms will likely frame the important issues facing decision makers over the coming decades. Without knowing how the possibilities will turn out (a placeless city of bits or not? More growth on the undeveloped urban edge or less? A sustained re-centering of urban settlements or their permanent de-centering?), it is imprudent to list specific issues to be grappled with. But three critical issues are already apparent in the foreseeable (or more seeable) future.

First, if largeness is a consequence of a passing industrial era, then how will the more fundamentally urban quality of density be manifest in the coming decades? While the emergence of a New Economy is neither fully understood nor even universally accepted, the technological framework within which urban settlements are constructed and maintained seems to have changed profoundly. While industrialism's appetite for physical agglomeration has weakened if not collapsed, there appears to be a powerful revival of demand for density on a very large scale in a select set of global cities and on a smaller scale in centers throughout the urban system. This demand for density appears to be related less to production and distribution, unlike during the industrial era, and more toward lifestyle amenities. Breaking the link between density and largeness in the minds of urban dwellers and decision makers, which seems fundamental to the suburban reform agenda discussed above, will be a major challenge in coming decades. In many ways, of course, this leads back to the future and to a revival of Ebenezer Howard and Garden City and New Town ideas.

Second, if the well known mismatch between formal and informal organization (local–state–national versus neighborhood–regional–global, respectively) is accurate, then how resistant or malleable will formal structures be in the service of informal imperatives? Many of the changes demanded in the preceding paragraph require collective action and governmental capacity. That the ideas being debated are neither new nor complex suggests the difficulties of actually implementing a desirable future. Whether or not the public apparatus of decision–making related to urban life can be adapted to match the new and emerging scales at which urban life is being lived through new technologies and relationships is perhaps the single most important issue framed by observable trends.

Third, urban life will continue to confront the issue of automobiles versus every other form of mobility. While epitomized by US settlements, this tension is rapidly becoming global in both its prevalence and its consequences. Regardless of the scale and governance issues outlined above, physical access to some resources (perhaps a different set than those of today) will continue to be a fundamental rationale for urban settlements. The access problem is solved through mobility of some form. The worldwide trend toward individual automobile mobility (whether that trend moves from the large base of car use in the United States or the smaller but rapidly growing base found in the developing world) will likely join if not eclipse other specific critical priorities facing urban decision makers. Ironically, the negative environmental impact of increasing automobile use is as global an urban issue as the positive relationships of the New Economy. The intersection of combustion engines, urban settlement patterns, and cross-national environmental degradation will challenge urban life in powerful and perhaps decisive ways.

See also: Mumford, Lewis (1895–1990); Technological Innovation; Technology, Anthropology of; Urban Geography; Urban History; Urban Places, Planning for the Use of: Design Guide; Urban Sociology; Urban Sprawl; Urban Studies: Overview

Bibliography

Downs A 1995 *New Visions for Metropolitan America*. Brookings Institution Press, Washington, DC

Duany A, Plater-Zyberk E, Speck J 2000 *Suburban Nation: The Rise of Sprawl and the Decline of the American Dream*. North Point Press: Farrar, Strauss and Giroux, New York

Hall P 1988 *Cities of Tomorrow*. Blackwell, London

Kotkin J, Siegel F 2000 *Digital Geography: The Remaking of City and Countryside in the New Economy*. The Hudson Institute, Indianapolis, IN

Kunstler J H 1996 *Home from Nowhere: Remaking our Everyday World for the 21st Century*. Touchstone: Simon & Schuster, New York

Lynch K 1981 *Good City Form*. MIT Press, Cambridge, MA

Mitchell W J 1995 *City of Bits: Space, Place, and the Infobahn*. MIT Press, Cambridge, MA

Mumford L 1961 *The City in History: Its Origins, its Transformations, and its Prospects*. Harcourt, Brace & World, New York

Rusk D 1999 *Inside Game, Outside Game: Winning Strategies for Saving Urban America*. Brookings Institution Press, Washington, DC

Rybczynski W 1995 *City Life*. Touchstone: Simon & Schuster, New York
Sassen S 1991 *The Global City*. Princeton University Press, Princeton, NJ

M. A. Hughes

Urban Places, Planning for the Use of: Design Guide

'Life Takes Place on Foot,' is the aphorism of Jan Gehl, a Danish architect who is an acute and sympathetic observer of the way that people interact with each other and with their surroundings. Gehl's theory is that, no matter what technical innovations are taking place, human beings have not changed. They still need the casual contact with other human beings that used to be built into daily life, but is becoming more and more unusual in countries with advanced industrial economies. Gehl has concluded that this complicated concept can be reduced to a relatively simple mechanism. People pursue necessary activities that take them through public spaces. If the spaces are a poor physical environment, people will get through them as quickly as possible. If the environment is attractive, people will linger and engage in what Gehl calls optional activities, like sitting down for a few minutes in a cool place in summer, or a sheltered, sunny spot in winter, or just slowing down and enjoying life, stopping for a cup of coffee or tea, or looking at a statue or a fountain. The more optional activities there are in a public place, the more likely that there will be what Gehl calls 'resultant activities,' that is, sociability, people meeting accidentally or striking up a conversation with strangers (Gehl 1980).

Gehl argues for the sociability created by traditional streets and squares, as opposed to the open spaces in most of Denmark's modern housing projects or the big institutional parking lots where there are trafficways but few streets next to buildings. He believes that designers can make new groups of buildings that have similar characteristics to the traditional townscapes that his research finds are still successful in fostering sociability.

Gehl's contemporary, the late William H. Whyte, makes comparable arguments about streets and public spaces in New York and other big American cities, mapping the places where people congregate and interact, and comparing them to public places that are empty and sterile (Whyte 1980, 1988). He also argues that design can make a big difference in behavior in public places, comparing spots that are taken over by antisocial activities with those that are safe and popular. Whyte adds two elements to Gehl's conclusions: one is the usefulness of providing movable chairs in public places, so that people can create temporary environments for themselves; the second is the importance of animation: multiple activities fronting on the public space and programmed for the space. The more that is going on in and around a space, the more likelihood that people will be attracted, their paths will cross, and Gehl's 'resultant activities' will take place.

1. *The Origins of Modern Public Spaces*

Traditional public spaces were not designed for leisure, so the issue of whether they were inviting or not did not arise. Market streets and squares were crowded, functional places, and other separate plazas in front of palaces or religious buildings were intended for ceremonies. Public fountains, became a meeting place as people went there to draw water. In large parts of the world these functional aspects of public space continue as they always have.

During the Renaissance period in Europe, public spaces began to be embellished to create compositions in perspective; comparable to those already created in paintings, gardens, and stage scenery. Buildings were designed to be symmetrical around a central axis, arcades and colonnades modulated the facades, and often the same design system was applied to the frontages around a square. Long, straight streets were designed to terminate at the center of a public building, or at a plaza with a monument. Most of these design innovations were about ceremony: an enhanced street environment for the carriages of the aristocracy, a place where the arrival of important people and their entrance into a palatial building would be appropriately stately and impressive.

The opening of Vauxhall Gardens and other pleasure gardens in London in the eighteenth century and the conversion of the Palais Royal in Paris during the 1780s into what we might today call a shopping and entertainment center, were some of the earliest public places dedicated to leisure-time activities. The Tivoli Gardens in Copenhagen is a descendant of this type of public space whose most recent evolution is the theme park. Public parks open to everyone, like Central Park in New York, are a nineteenth century innovation, so are village greens with a public bandstand. The New England common in the seventeenth and eighteenth century was a staging area for cattle, overgrazed, dirty, and full of manure—not a green oasis with a while-painted pavilion in the center. The street as a place for a leisurely public promenade is also a mid-nineteenth century invention, the Parisian boulevards being the archetype. The European café with tables on the sidewalk or in the public square is a relatively recent innovation. The pedestrian street or pedestrian district is mostly a new phenomenon since around 1950, the Stroget in Copenhagen, Denmark, being one of the

first shopping streets that were closed to all but pedestrians. In US cities many of the pedestrian malls have been taken out again, as closing the streets to vehicles caused them to disappear from the mental maps of the people who once used them. In the historic centers of European cities, where public transit is much more a part of everyday life, the pedestrian street and plaza remains successful.

2. *Making Public Space Inviting*

The idea that cities should have many urban outdoor spaces, open to all, without an admission charge or the need to patronize a café, owes something to the modern movement in architecture. Tall buildings need space around them, and architects began designing plazas to provide an appropriate setting for towers. These plazas were not always intended to encourage people to spend time there. The idea was to create an impressive approach to the building. The owners often placed spikes on ledges so that people would not sit or lie on them, and the microclimate was frequently unfriendly, with strong downdrafts from the building when it was windy, blinding sunlight on hot days, and large unsheltered expanses in winter.

Uninviting spaces in highly accessible locations became places for 'anti-social elements' to hang out, they attracted drug dealing, they detracted from the atmosphere of corporate dignity that the owners and architects had hoped to create. In addition, some of these public spaces had been built in response to government incentives; they were supposed to be a public benefit, not a public nuisance.

Research by William H. Whyte, Jan Gehl and others has now given us some simple principles for the design of these new kinds of public spaces, which can be summarized under the following seven headings.

2.1 *Legibility*

Savvy pedestrians are not going to enter a public space if they cannot see how they can get out of it again. Sunken and raised plazas are usually not attractive for this reason, neither is any kind of enclosed dead end. A successful public space needs several clearly visible entrances and exits. The same considerations apply to sub-components of the public space. There should not be any secluded pockets or inaccessible corners.

Interior public spaces need openings to the outdoors so that people can figure out where they are. Bridges in skyway systems generally have windows that provide orientation. Underground concourse systems should look out on occasional sunken courtyards. Works of art and distinctive storefronts can help people find their way, but they do not provide any reference to familiar landmarks outside of the building.

2.2 *Microclimate*

As the public open space is sometimes a residual element in the design of the building, it can end up having an unfavorable orientation. A north-facing plaza, which gets almost no sunlight, is not likely to be attractive, even in a warm climate. New York City's plaza regulations require a southern exposure whenever possible. Wind tunnel tests are usually done for tall buildings to evaluate special conditions on the exterior walls that might cause windows to pop out or other structural failures. At the same time, tests could be done to see if there are any unusual wind conditions created on the plaza. The shape of the tower can interact with winds from a certain direction to create augmented air currents across plazas. These locally created winds can be strong enough to knock people over.

People like to sit in the sun on a cold day and be protected from the wind. On hot days they look for shade and a cool breeze. Good public-space design should be able to create both conditions. Well-placed trees and a fountain can create a pleasantly cool microclimate on hot days; chairs or benches in front of a south facing wall that stores heat and blocks wind can produce a favorable environment when the weather is cooler.

2.3 *Seating*

The essence of an inviting space is the ability to sit down and linger as long as you wish. The seating does not have to be benches or chairs. Ledges on buildings or the edges of planting boxes—without spikes—will do very well if they are placed at three feet or less (but more than a foot) above the plaza, a height where sitting is easy and comfortable. Flights of wide shallow stairs lend themselves to all kinds of impromptu seating arrangements.

Whyte recommends individual movable chairs, such as are rented out for a small sum in French parks, over fixed arrangements of benches. The chairs can be moved about to create all kinds of groups, pulled into the sun or shade, and just generally rearranged to give individuals a feeling of comfort. If fixed, backless benches are used, Whyte suggests making them wide enough that strangers can sit on either side without making each other uncomfortable.

Furnishing a public space with movable chairs means that there must be a guard or attendant present, which is desirable for other reasons: even the best designed spaces need someone to keep order and deal with emergencies.

2.4 *People Watching*

One of the principal attractions of being in a public space is the opportunity to watch other people. Seating

areas should face out over circulation paths. Studies have shown that benches or other fixed seating oriented away from traffic flow will go unused, or will be used in what Jan Gehl calls 'untraditional ways.' He illustrates his point with a photograph of two Danish matrons seated back to front on a park bench, with their legs through the gap between the seat and the backrest, watching the crowds walk by.

2.5 Food

A concession that sells sandwiches, salads, soft drinks, and desserts enhances a public space and is an economical way of making sure that there are responsible people around to keep an eye on what is going on. The only problem is if the concession is permitted to take over the space and make it uncomfortable for anyone who has not bought food or drink.

2.6 Lighting

Even cities that have given a zoning bonus for a public plaza should allow it to be closed from the late evening to the early morning. However, lighting is still an important factor in extending the use of the space and making sure it is not misused at night. Floodlights glaring down on the plaza from tall poles or surrounding buildings are efficient, but can create a penitentiary atmosphere. Although harder to maintain, some of the light should be reflected off trees, come from built-in fixtures near ground level, and from street lamps in scale for pedestrians, with light sources that give a pleasant illumination.

2.7 Surrounding Activities

The most important factor in the design of an urban plaza is what happens around it. New York City's regulations governing plazas that qualify for a zoning bonus require that 50 percent of the building frontage on the plaza be devoted to retail—defined as not including banks, brokerages, airline offices, and travel agencies. This is a shrewd requirement as the typical New York zoning plaza, before these regulations were written, belonged to an office building. If the office tower had any ground-floor uses at all, they were likely to be a bank, a brokerage, or a travel agency, dignified activities that were useful for office tenants but did not do much to enliven a plaza, compared to a restaurant, a shop selling take-out food, or a book store.

Jan Gehl goes farther, saying that tall buildings should not directly adjoin a public space or even, if at all possible, a sidewalk. His reasons have to do with keeping buildings and their activities in scale with pedestrians. Someone leaning out a third-floor window is still recognizably a human being to a person walking by at ground level. Of course, most modern office towers are sealed buildings, even on the lower floors, but that is exactly the point Gehl is making. A street or a public square is a three dimensional place, for it to be at its best it should relate to activities on the second floor and perhaps above. Shopping malls are now designed this way, so that, if you are walking along on level one, you can see that there is a food court on level three. On the same principle, retail streets and public places should be designed with a liner of buildings of low buildings with active uses, with the taller buildings set back within this matrix of traditional relationships.

In residential neighborhoods, the same principle suggests that houses be designed in relation to streets and public spaces. This is the basis for the front porches and small front yards characteristic of many 'new urbanist' communities.

2.8 Walking

According to Whyte's observations, the distance that people were likely to walk in New York City was five north–south blocks, which adds up to about 1250 feet, or a little less than a quarter of a mile. This finding agrees with the experience of shopping center developers who have found that 1200 feet is about the practical limit to the distance between major destinations, like department stores, in a shopping mall.

It also agrees with the diagram published by Clarence Perry in his famous article, *The Neighborhood Unit*, showing a residential neighborhood lying within a circle a half-mile in diameter—that is no more than a five-minute walk from a central point (Perry 1929). At the center of the circle were an elementary school and other neighborhood institutions.

These are modern figures applying to people who are used to driving a car or taking public transit. The five-minute threshold is not about physical stamina but about the onset of boredom, and the fact that people today feel that they need to use time efficiently. According to Whyte, New Yorkers, living in a transit-rich environment, will start thinking about taking a bus, the subway, or a taxi if a trip is going to take more than five minutes on foot. In a more typical city, someone leaving the office to go to lunch will start thinking about getting the car out of the garage if the distance is more than a short walk.

Whyte found that to get people to walk for even five minutes meant that you had to keep them interested. In environments where buildings flanking the sidewalk had long stretches of blank walls, Whyte found that pedestrian traffic was low. Again, this confirms the experience of shopping-center owners that even a few vacant storefronts near a prominent corner could be enough to discourage shoppers from entering a whole section of a mall. The interest factor turns out to be far more important than ease of pedestrian movement.

For years city planners and designers believed that the best way to promote pedestrian movement was to remove 'conflict' between pedestrians and vehicles and to protect pedestrians from the weather. This is the theory behind the skyway systems in Calgary, Minneapolis, St Paul, Charlotte and other cities in the USA and Canada, and also the underground pedestrian concourses in Montreal, at Rockefeller Center in New York, and in downtown Dallas and Houston. Whyte measured the pedestrian traffic in skyways and concourses and found that the numbers of people dropped off sharply as he went farther from the center of the system. Not only did a skyway or concourse system not overcome the five-minute limit, but also, unless there were shops along the way, people did not even walk that far.

The weather protection provided by the concourse and skyway system is clearly useful. In Canada, Calgary has sub-zero winter temperatures, although even in Calgary Whyte found that people prefer the streets when the weather is good. On a winter day in Montreal that was so cold that his camera froze, William H. Whyte found as many pedestrians on Ste. Catherine Street as in the nearby underground concourse of Place Ville Marie.

Separating people from cars is more useful for drivers, by getting rid of delays at crosswalks, than it is for pedestrians, who generally prefer a direct route at street level to the bridges or underpasses placed for their safety at busy urban intersections.

The effect of a skyway or underground concourse system is to divide the pedestrians into numbers that are too small to support retail at both street level and the skyway or concourse. Except in very busy cities the protected level will become dominant, leaving the street level without much in the way of shopping, which means it is too boring to attract pedestrians, and does not look safe enough.

Whyte's studies left him much impressed by the efficiency of walking and the instinctive skill with which people avoid collisions on busy sidewalks. He was also bemused to note that the places where people stopped to talk were generally at street corners, right in the middle of the pedestrian traffic stream.

3. *The Empty Square*

Both Gehl and Whyte, working in the 1970s, present these issues as questions of design, a follow up to Jane Jacobs' famous comparison of the deserted and dangerous lawns in a public housing project with the active social life on her street in Greenwich Village in the 1950s (Jacobs 1961).

Today in some parts of the world this kind of livability has become a larger design issue than the configuration of individual streets and public spaces. An article, 'The Empty Square' by Alan Ehrenhalt in a recent *Preservation*, asks the question: 'If casual social encounters are at the heart of civic life, where did everybody go?' (Ehrenhalt 2000). In most US cities and towns necessary activities take place using a car, and the opportunities for any kind of casual interaction are much diminished. The commuter goes from the garage at home to the garage or parking lot at work. Only the journey from the car space to the lobby takes place on foot. Shopping and errands are done by car to individual destinations. Schools, churches, country clubs, restaurants, movies—each are a separate destination reachable only by car. The Court-House Square is empty. People drive to the health club and do their walking on a treadmill.

Architects, landscape architects, and city planners are currently debating whether it is still important to create public places that permit the face-to-face contact found in traditional cities and towns, or whether life today requires something completely different. One side, which includes the people who call themselves New Urbanists, believes that traditional streets, squares, promenades, and parks are still the essence of city design (Charter of the New Urbanism 1996, 2000). Their opponents, among whom the Dutch architect, Rem Koolhaas, is perhaps the most articulate, believe that the idea that cities can be designed at all is based on unexamined philosophical assumptions, and that modern transportation and communication, particularly the Internet, have made traditional urban spaces obsolete. One of Koolhaas's characteristic statements: [we need] 'to accept what exists. We were making sandcastles. Now we swim in the sea that swept them away' (Koolhaas and Mau 1995).

An expert on computer technology, William J. Mitchell, has written a book, *e-topia: 'Urban Life, Jim—But Not as We Know It,'* which observes that the interactions of technology and behavior are much too complex to allow simple apocalyptic conclusions (Mitchell 2000). Mitchell points out that you have to look at the totality of people's work and personal lives if you want to make good estimates about where new technology is taking us. Most of the people who have chosen to leave cities and suburbs because the computer has freed them to work where they choose are not living on mountain-tops but in smaller, rural communities, precisely because of the face-to-face communication possible in such places. Many email messages are about establishing face-to-face meetings. The demand for hotel rooms and convention centers seems to be going up. Mitchell notes the rising interest in living in or near places that are rich in social interaction: older urban centers and suburban downtowns, older mixed-use neighborhoods and their modern imitations, as well as resorts. He concludes that the power of place will still prevail, that people will still gravitate to settings that offer particular cultural, scenic, and climatic attractions—'those unique qualities that cannot be pumped through a wire'—and will continue to care about face-to-face interactions.

Assuming that William Mitchell is correct, and the design of urban public places will continue to matter, how can such spaces be incorporated into new development?

4. *Designing Livable Communities as well as Public Spaces*

Planning workplaces and residences so that they relate more closely to each other, designing retail as park-once districts, following economic development policies to create self-reinforcing clusters of businesses and related activities, building neighborhoods of diverse housing types instead of instead of housing tracts, and integrating school buildings back into their communities would go a long way towards creating more livable communities where personal interaction is easier and more meaningful.

The amount of time spent commuting is clearly one of the most important livability issues, because excessive commuting time takes up the part of the day that could otherwise be used for more leisurely activities or spent at home with children. Attempts to solve this problem by building new roads are seldom successful because the roads induce new traffic. New patterns of working at home and commuting electronically will diminish the problem. But the real answers will come from creating places where home and work are more closely related.

What made the traditional main street a place where people could meet unexpectedly was that they were walking from one destination to another. Now, designers of retail development are trying once again to create park-once districts, partly to foster interaction and communication, but also to create synergy among the different retail tenants. This is another trend that will make the design of public spaces more important, as people walk from one destination to another.

5. *Bryant Park*

How some of the design principles discussed in this article work out in practice can be understood from the example of once derelict Bryant Park, a public space behind the New York City Public Library between 40th and 42nd Streets. William H. Whyte was consulted about the redesign of the park. A restaurant was built on the Library's back terrace, in keeping with Whyte's prescription for animating public spaces, but against the instincts of people who do not believe that public spaces should be used for commercial purposes. Landscape architect Laurie Olin also followed Whyte's precepts in making subtle adjustments to the terracing in the park, making it easier for people to see into the park, and getting rid of dead-end locations. Lush gardens, good maintenance, a constant security presence, and lots of movable chairs completely transformed the park, which is now a popular place during the day and well into the evening.

The story of Bryant Park and similar successes elsewhere demonstrate that derelict places can be made livable by following a few simple principles, which can also be applied to the design of new neighborhoods, workplaces, and shopping districts.

See also: Development and Urbanization; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Planning Theory (North American); Public Goods: International; Strategic Planning; Urban Growth Models; Urban Life and Health; Urban Sprawl.

Bibliography

Arendt R et al. 2000 *Charter of the New Urbanism*. McGraw-Hill, New York

Barnett J 1982 A public open space policy. *An Introduction to Urban Design*. Harper and Row, New York, Chap. 11, pp. 163–86

Ehrenhalt A 2000 The empty square. *Preservation*, March /April 2000 Washington, DC, pp. 42–51

Gehl J 1980 *Livet mellem husene*, Arkitektens Forlag, Copenhagen, Denmark, translation by Jo Koch 1987, Van Nostrand Reinhold, New York

Jacobs J 1961 *The Death and Life of Great American Cities*. Vintage Books, New York

Koolhaas R, Mau B 1995 *S, M, L, XL, Office of Metropolitan Architecture*, Sigler J (ed.). Monacelli Press, New York

Mitchell W 2000 *e-topia: 'Urban Life Jim – But Not as We Know It*. MIT Press, Cambridge, MA

Perry C A 1929 The neighborhood unit. In: *Neighborhood and Community Planning, Vol. VII of the Regional Survey of New York and Its Environs*, Committee on Regional Plan of New York and Its Environs, New York

Whyte W H 1980 *The Social Life of Small Urban Spaces*. The Conservation Foundation, Washington, DC

Whyte W H 1988 *City, Rediscovering the Center*. Doubleday, New York

J. Barnett

Urban Places: Statistical Definitions

Almost a century ago, Walter Lippmann observed that a great gap exists between 'the world outside and the pictures in our heads.' The gathering of facts through the creation of governmental censuses and statistical reporting represents a major attempt to remedy this condition. Statistical information provides important data that are critical to understanding

and planning for the economic and social development of society. It aids in the assessment of local and national patterns of development and the allocation of resources to reflect changing needs. It supports the articulation and coordination of goals for national, local, and regional development. A major obstacle to the ability to accomplish this is the presence of dynamic societal change and the emergence of new facts that challenge the accuracy of existing concepts and statistical definitions. Faced with conceptual obsolescence, governmental statistical agencies often resist change uncertain about how to accommodate to it and concerned about the integrity and comparability data. Conflict emerges between the need for stability in concepts and definitions and the desire for relevance. Attempts to develop information about a society's urban population, which changes so rapidly that it quickly renders obsolete the conceptual underpinnings of data definitions, especially susceptible to this dilemma.

All developing societies create some model of an information-gathering national census, albeit with great differences in quality and extensiveness of coverage. The impetus for statistical definitions and data stems from a variety of factors, such as the need for knowledge about a nation's development patterns, the desire to support local and national planning functions, the projection of future development, and the planning and managing of an allocation of resources responsive to the changing location distribution of population need. In addition, in many societies there is a desire to determine the ways in which racial and ethnic differences relate to patterns of settlement and place.

This article focuses on ways in which nations have developed urban statistical definitions. It emphasizes the United States experience, the historical evolution of the urban definition, the current assessment of statistical practices and the prospects for change in the statistical definition of urban. But first, it will consider contrasting practices throughout the world in order to develop an understanding and perspective of how different nations have responded to the need to develop urban statistical definitions.

1. *Contrasting International Definitions for Urban*

No one statistical definition captures the entire story about urban definitions in countries around the world. In fact, the collection of data regarding urban population is perhaps the least standardized and comparable of worldwide census activities. Individual national approaches are highly differentiated and particular in terms of data collection systems and analysis, thereby rendering difficult attempts at cross-national comparisons. These differences stem from contrasting levels of development, diverse cultures, and changing societal needs. They also reflect differing historical traditions, practices inherited from colonial regimes, fiscal and human resources available for census activities, and levels of national integration. The resulting variety of urban definitions rely on such factors as minimum population size, levels of population density, the distinctiveness of a governmental or administrative political unit, the percent of agricultural and nonagricultural working population, and other political, social, and economic factors distinct to particular societies and nations. For example, the meaning of a governmental jurisdiction, such as town or village, is often very different among nations. Or, the difference between *de facto* and *de jure* population has a great effect on the classification of urban and rural populations, especially at times of turmoil and instability which result in large movements of population.

A survey of national census practices reveals widely varied definitions of 'urban'. The Swedish define it as those built-up areas with less than 200 m between houses. Bulgaria uses the term to refer to localities legally established as urban, regardless of size. Denmark employs a minimum population requirement of 200 while Japan has a minimum population requirement of 30,000. France, Nigeria, Iran, and Turkey have definitions of a minimum population of 2,000, 5,000, 5,000, and 10,000 respectively. Many countries, largely concentrated in Africa and Asia, have either not yet developed or made available estimates of urban population in their national censuses.

India presents an interesting example of the complexities involved in urban definitions. It includes towns (places with municipal incorporation, municipal area committee, town committee, notified area committee, or cantonment board), as well as all places having 5,000 or more inhabitants, a density of not less than 1,000 persons per square mile, or 390 persons km^{-2}, pronounced urban characteristics, and at least three-quarters of the male population employed in pursuits other than agriculture. Clearly Indian demographers have recognized the difficulties in developing accurate terminology that reflects the complexity of their country's economic and social development.

Once established, statistical categories and definitions are increasingly codified and difficult to modify or change. While the goal of consistent comparability of data among different nations is a distant vision, nations are continually expending resources to develop refined population data reflecting their unique patterns of development. We shall now turn our attention to the example of the United States, a nation that has tried to adapt its definitions of urban to dynamic economic and social change.

2. *The American Quest for an Urban Definition*

The American quest for a statistical definition of urban reflects a tortuous pattern of adaptation to

changes in the structure of society over this last century. Demographers of an earlier period could not foresee the dynamic changes that were to affect how Americans live and work. They differentiated between urban and rural by contrasting the preponderance of agriculture or manufacturing, the size and density of the population, and the concentration of commercial activity. These characteristics informed their assessment of urban and rural places. In practice, the distinctions between urban and rural were easily made because the concept urban was applied only to a few large cities that were easily considered as an urban population.

This approach is described with great detail by Adna Weber (Weber 1899) in his statistical study of nineteenth-century cities. He classified rural and urban populations by size of place, specifying rural as 'scattered' or 'agglomerated' (hamlets and villages with fewer than 2,000 population and towns with from 2,000 to 10,000 population) and urban as 'cities' (more than 10,000 population) and 'great cities' (more than 100,000 population). The Census Bureau's current urban and rural classification consists of two different urban categories. One is any area that is part of an Urbanized Area (UA) and those places outside the UA that have a population greater than 2,500. Territory not included in these two categories is classified in the residual category of rural.

The US Bureau of the Census officially adopted the 2,500 level for the 1910 Census, a figure that survives till this very day, but now has a more limited application. Between 1910 and 1940 the decennial censuses had an additional requirement that the urban place be part of an incorporated territory. The Bureau made some allowances for minor civil divisions (MCDs) having large population size and density. The place size definition of urban was soon considered to be inadequate because of its exclusion of many population concentrations located in unincorporated places, resulting in their inappropriate rural classification. To remedy this problem, census demographers continued to grapple with the challenge of modifying the concept to reflect the changing realities of urbanism, especially the increasing spread of the population to new areas. There was a need to create a concept that would be based on additional criteria other than place size and MCD status. This led to the development of the idea of an 'urbanized area' (UA) in 1950, a concept that exists, albeit in a modified form, till this very day.

Although first defined in 1950, the Census Bureau has modified the UA concept over time. A UA is a continuously built-up area with a population of 50,000 or more, comprised of one or more places and the adjacent settled surrounding area, the urban fringe. The UA includes only urban populations, consisting of a central city and incorporated places with a minimum population size of 2,500, or unincorporated places with a population density of at least 1,000/mile^2. Since 1990, there has been no minimum size stipulated for a central place and the requirement of central city has been changed to central place. A Census Designated Place (CDP) that contains a densely settled unincorporated area of 2,500 or more is now classified as urban, thereby correcting a major gap in previous urban definitions.

3. *The Emergence of the Metropolitan Idea*

Even at the time of its formulation, the UA concept failed to provide an accurate measure of contemporary developments. Its focus was on the contiguous connection between the central place and the urban fringe excluded population residing in noncontiguous urban enclaves that were economically and socially related to the center. Demographers searched for a concept to reflect urban relationships in the new era of metropolitan dominance. They developed new terms to refer to the metropolitan concept, beginning with the Standard Metropolitan Area (SMA) in 1950, the Standard Metropolitan Statistical Area (SMSA) in 1959, the Metropolitan Statistical Area (MSA) in 1983, and most recently the Metropolitan Area (MA) in 1990.

Their work paralleled what Hans Blumenfeld described with great insight in 1967 as a new stage in the 'urban revolution,' the transition from the 'modern city' to the 'modern metropolis.' He stated that 'this change of name reflects the fact that from its long, slow evolution the city has emerged into a revolutionary stage. It has undergone a qualitative change, so that it is no longer merely a larger version of the traditional city but a new and different form of human settlement' (Blumenfeld 1967). While Blumenfeld's view of the metropolis was in some ways similar to the 'conurbation' idea originally described by Patrick Geddes at the turn of the nineteenth century, and later elaborated by Lewis Mumford, it was distinctly different from the 'megalopolis' of Jean Gottmann 1961, and the 'ecumenopolis' of Constantinos Doxiadis 1967, which were also conceived in the 1960s. Attempts to portray the realities of contemporary urbanization by referring to an urban settlement that is 'neither city nor country but a complex of urban districts and open areas' have much precedence in the development of urban ideas.

The need for a concept that better reflected the spread of the urban population beyond city boundaries dates back to before the Civil War. This perception is evidenced in the New England Gazetteer's entry on Boston which noted that

> owing to the almost insular situation of Boston, and its limited extent, its population appears small. But it must be considered that the neighboring towns of Quincy, Dorchester, Milton, Roxbury, Brookline, Brighton, Watertown, Cambridge, Charlestown, Medford, Malden, and Chelsea, although not included in the city charter, are component parts

> of the city, and are as much associated with it in all its commercial, manufacturing, literary, and social relations and feelings, as Greenwich, Manhattanville, and Harlem are with the city of New York; or Southwark and the Northern Liberties with Philadelphia (United States Department of Commerce 1994).

By 1905, the Census of Manufacturers utilized a metropolitan concept when it replaced its 'industrial district' classification with a 'metropolitan district' concept, retaining it with modifications through 1940. Metropolitan districts were defined in terms of minor civil divisions (MCDs), with their boundaries based primarily on population density. In 1950, the Bureau of the Census incorporated the metropolitan concept in its term Standard Metropolitan Area (SMA). The journey to work as measured by commutation patterns is the major characteristic defining the boundaries of the metropolitan area. The SMA includes at its center a city, (or in a few instances a census-designated place or county subdivision with a minimum population of 50,000), and its county and those contiguous counties that meet specified requirements such as the 15 percent commutation rate to the central county. Other characteristics relate to population density, percent of the population classified as urban, percentage of growth in the population between the two previous decennial censuses, and percentage of, or absolute number of, inhabitants within the UA that qualifies for inclusion in the metropolitan area. A contiguous county is included in its entirety or it is not included. This can result in a significant number of rural persons being included in a metropolitan count.

The Metropolitan Area (MA) is meant to capture the high degree of economic and social interdependence among a central city and its surrounding counties. It contrasts with the urbanized area that focuses on the densely settled agglomerations surrounding large cities. The MA concept provides a means of defining a functional area based on the activities of the population while the UA measures the extent of urban agglomeration around a central core. The census demographers viewed metropolitan areas and urbanized areas as complementary. Together they present an accurate depiction of the urbanization process.

Dynamic and new forms of growth often render both the MSA and UA ineffective as measures of the new realities. Population decentralization away from the central city to new centers in outlying areas has resulted in a form of urbanization far different from that envisaged in the original metropolitan concept. As metropolitan areas grow, their boundaries expand to and often overlap the boundaries of other adjacent urban areas. In 1983, the Bureau of the Census responded to these developments by introducing two new designations, the Primary Metropolitan Statistical Area (PMSA) and the Consolidated Metropolitan Statistical Area (CMSA). An area with a population of one million or more is recognized as a CMSA if it has component parts that are eligible to be PMSAs. These component parts must meet certain statistical criteria that demonstrate strong internal economic and social links as well as ties with the center of the larger area. The support of local public opinion is an additional criterion for inclusion. The boundaries of the CMSA are defined as the area resulting from the aggregation of all of the PMSAs.

After a century of classifying urban data, the Bureau of the Census has formulated a number of statistical definitions, each relevant to different aspects of the urbanization process. Their development reflects the efforts of government statisticians and demographers to describe and understand different patterns of urban development resulting from the dynamics of technological and social change. Together these definitions form a conceptual framework that accounts for population size, density, economic and social interdependence, and jurisdictional units of government in portraying urbanization and metropolitan growth.

This framework incorporates minimum place size, density characteristics, core–periphery relationships involving cities, suburbs, and metropolitan regions, and monitors the increasing decentralization and deconcentration of urban areas. By moving beyond the more limited notion of urban place they provide an accurate picture of a complex and continually evolving reality. Nonetheless, a gap remains between how American urbanization is currently developing and how it is portrayed in statistical concepts. Attempts to remedy this situation are discussed in the final section of this article.

4. *Prospects for the Future*

Cities are no longer the dominant form and entity of urbanization, and suburbs are generally considered to be that part of the urban population residing outside the central city, referred to as the 'urban fringe.' While there is no formal census statistical definition of a suburb, there is recognition of the fact that suburbs have been the most rapidly growing part of the urban population. A half a century ago, 60 percent of the residents of metropolitan areas lived in central cities. At the current time, 60 percent live outside the central cities, that is, for the most part in suburbs.

This dramatic change in American life styles led the historian Jon C. Teaford (Teaford 1993) to conclude that traditional concepts of urban have become so outdated that:

> at the close of the twentieth century, then, the age of urbanization in the United States had come to a close. The concept was no longer relevant for a nation that was metropolitan but not urban. Unimaginative data collectors in the Bureau of the Census still embraced the outdated concepts of urban and rural, but such concepts masked rather than

> illuminated reality. The urban frontier had closed, and the term *urbanization* should have been retired to that conceptual attic frequented primarily by students of the past.

Similarly, Robert Fishman (Fishman 1990) stated that:

> for most Americans, the real center of their lives is neither an urban nor a rural nor even a suburban area, as these entities have been traditionally conceived, but rather the 'technoburb,' the boundaries of which they can conveniently reach in their own cars. The true center of this new city is not in some downtown business district but in each residential unit. The old central cities have become increasingly marginal, while the technoburb has emerged as the focus of American life.

While in the past dissatisfaction with existing urban definitions led to the introduction of such new statistical measures as the Urbanized Area, and the Metropolitan Statistical Area, many, like Teaford and Fishman, believe that they are not fully adequate in their depiction of contemporary America. The question remains as to what will be the next stage of conceptual change in the statistical definition of urban.

Today's urban statistical definitions have a strong bias towards interpreting urbanization within the framework of a core-periphery model emphasizing the central city. Critics argue that the emphasis on central city-suburban relationships ignores what is going on in the rest of the metropolitan region, especially the increasing decentralization of population and job location that results in an increase of suburb-to-suburb journeys to work and other functional relationships. They also maintain that dependence on the county as a building block results in the inclusion of non-urban population in the metropolitan count and the neglect of numerous and varied subcounty interrelationships.

Further, advances in computer technology allow for data collection and analysis disaggregated at levels far smaller than the county. This phenomenon allows for additional change, never before contemplated, in the statistical definition of urban. The Bureau of the Census is currently reviewing proposals involving utilization of such alternative geographic building blocks, including the Minor Civil Division (MCD), Census County Division (CCD), Census Tract, Zip Code, and Grid Cell. All involve different levels of consistency, stability across boundaries, accuracy of depiction over time and space, data availability and familiarity, and identifiability. It is expected that the current review will result in a number of modifications in existing practices, but it is as of yet uncertain how extensive these recommendations will be.

The center-core orientation of the definition of urban had great validity in the past, but is inadequate at a time when the percentage of the nation's urban population living in suburban area continues to increase. Structural factors (such as population size and density) combined with functional factors (patterns of economic and social linkages such as the journey to work) still remain the best indicators for metropolitan inclusion. In addition, the emergence of less expensive telecommunications, and the increasing pervasiveness of the information superhighway associated with the use of the Internet, are changing functional relationships in ways that will lead to the increasing decentralization of the urban population.

To appreciate the significance of these changes, it is not necessary to agree with the views of Adna Ferrin Weber who forecast in the late nineteenth century that suburbanization 'furnishes the solid basis of a hope that the evils of city life, so far as they result from overcrowding, may be in large part removed. If concentration of population seems destined to continue, it will be a modified concentration which offers the advantages of both city and country life' (Weber 1968). Much has transpired since Weber predicted the growth of the suburbs and its impact upon cities. Today, urbanists focus on the dynamics of urbanization. Flow and movement and the 'space of flows' (Castells 1996) have become a dominant distinguishing urban characteristic of our time. This is true whether one emphasizes the influences of technology upon transportation and production or is more concerned with the impacts of new communications technology on forms of urban development.

Speculations about the future of the city, involving concepts such as 'edge city,' 'technoburb,' 'transactional city,' 'informational city,' megalopolis,' global city,' or 'post-suburbia,' represent attempts to understand how the concept of 'urban' is changing. Nevertheless, while these new forms involve an important and broad range of conjecture and imagination, there still remains the central importance of space as human beings construct their lives primarily in relation to where they live and work. While statistical definitions of urban have undergone a complex and changing history, it remains clear that much is yet still to come, as the dynamic facts of economic and social change intrude themselves upon established statistical categories of urban definition. The search for new concepts and definitions continues. The outcome of this struggle will be clearer to the readers of the fourth edition of the *International Encyclopedia for the Social Sciences*.

See also: Development and Urbanization; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Statistical Systems: Censuses of Population; Strategic Planning; Urban Geography; Urban Growth Models; Urban Life and Health; Urban Places, Planning for the Use of: Design Guide; Urban Sprawl; Urban Studies: Overview

Bibliography

Bird J 1977 *Centrality and Cities*. Routledge and Kegan Paul, London
Blumenfeld H 1967 *The Modern Metropolis: Its Origins, Growth, Characteristics and Planning*. MIT Press, Cambridge, MA
Borgotta E F, Borgotta M L 1992 *Encyclopedia of Sociology*. Macmillan, New York, Vol. 4
Castells M 1996 *The Rise of the Network Society*. Blackwell, Oxford, UK
Dahmann D C, Fitzsimmons J D (eds.) Metropolitan and nonmetropolitan areas: New approaches to geographical definition. Population Division, US Bureau of the Census, Washington, DC, Working Paper No. 12, September 1995
Doxiadis C A 1967 *Ecumenopolis: The Settlement of the Future*. Doxiadis Associates, Athens, Greece
Fishman R 1990 America's new city: Megalopolis unbound. The Wilson Quarterly, Winter. The Woodrow Wilson International Center for Scholars, Washington, DC
Garreau J 1991 *Edge City: Life On the New Frontier*. Doubleday, New York
Gottman J 1961 *Megalopolis*. The Twentieth Century Fund, New York
Goyer D S, Draaijer G E 1992 *The Handbook of National Population Censuses, Europe*. Greenwood Press, Westport, CT, pp. 18–19
Gross B 1966 *The State of the Nation: Social Systems Accounting*. Tavistock, London
Sassen S 2001 *The Global City: New York, London, Tokyo*. Princeton University Press, Princeton, NJ
Teaford J C 1993 *The Twentieth-century City*. The Johns Hopkins University Press, Baltimore, MD
United Nations 1998 *1996 Demographic Yearbook*. United Nations, New York
United Nations 1997 *Statistical Yearbook 1995*. United Nations, New York
United States Department of Commerce 1994 *Geographic Areas Reference Manual* Economics and Statistics Administration, Bureau of the Census, Washington, DC
Weber A F 1968 *The Growth of Cities in the Nineteenth Century: A Study in Statistics*. Originally published in 1899. Reissued by Columbia University Press, Ithaca, New York

S. Moses

Urban Planning: Central City Revitalization

1. Definition

As cities age, their physical features gradually deteriorate. Buildings, public facilities, vegetation, open spaces, and other elements may be involved. Careful maintenance and preservation can minimize this process, but with the passage of time, nearly all structures, spaces, and other features will have to be restored, rebuilt, or replaced. This process, usually accompanied by negative economic and social conditions, is termed central city (or urban) decline. Ameliorative responses to these forces are known as central city revitalization. While improving the condition and appearance of buildings, parks, streets, and the other elements of the urban environment is the most obvious goal, closely linked are desires to improve economic and social circumstances for those who live in, work in, or visit the city.

2. Historical Evolution

From the very earliest human settlements to modern world metropolises, the revitalization of cities is an enduring theme in world history. Cities have been rebuilt in response to the depredations of warring invaders and the destruction from volcanic eruptions, hurricanes, and earthquakes. More commonly, the forces of urban decline act less dramatically and over many decades. For example, during the Renaissance the medieval sections of several European cities were replanned. The clergy or nobility demolished ancient walls, straightened and widened streets, removed aged buildings, and added parks, monuments, and aesthetically pleasing visual features. The rebuilding of Rome under a sixteenth-century plan prepared by Pope Sixtus V and his assistant Domenico Fontana is a most notable example. In the early nineteenth century, royal architect John Nash redesigned a series of narrow London streets into today's Regent Street, which connects Piccadilly Circus to Regent's Park (also designed by Nash). Napoleon III's Prefect of the Seine, Baron Georges-Eugene Haussmann, tore out thousands of Parisian slum buildings and built wide boulevards with landscaping, arches, and statuary in the mid-nineteenth century (Barnett 1986). The City Beautiful movement in the USA brought plans to revitalize San Francisco, Chicago, Denver, Cleveland, Indianapolis, Minneapolis, Washington, DC, and other central cities in the early twentieth century.

3. Contemporary Central City Revitalization

Central city revitalization is associated here with a more modern expression of this idea. A central city is generally thought to be a more densely populated urban core surrounded by a metropolitan area of lower density but larger land area. Many central cities in the advanced industrial nations of Europe, Australia, New Zealand, and Canada have experienced decline, primarily since the middle of the twentieth century. There is little question however that of these, the USA has suffered the most severe and widespread central city deterioration. There is general agreement among social scientists and urban planners that suburbanization (the relocation of businesses and predominantly middle- and upper-income households from central cities to suburbs) has been the primary dynamic influencing these outcomes. While experts

disagree on precisely how and why rapid suburbanization occurred, most concur that it was a combination of the attractiveness of suburban life and the desire to escape the noise, congestion, crime, declining schools, and changing racial, ethnic, and social profiles of city residents.

As urban decline progressed in the USA, once middle-class city neighborhoods became lower-income enclaves. Landlords reduced maintenance due to diminished income and city agencies cut public services in response to lower tax receipts. In most cases, overcrowding occurred, crime increased, tensions arose between older and newer residents, and more households departed for the suburbs. Downtown department stores and shops followed their customers outward, establishing suburban outlets at shopping centers and regional malls. Corporations sought more modern factory and office space in the suburbs. Maritime industries such as fishing and shipping relocated their operations away from the historic city waterfronts due to higher costs, eroding dockage, and the introduction of new cargo-handling technologies. National and state government programs built more highways connecting cities to outlying areas, making it easier for more affluent families to live in the suburbs and commute to central city employment. By the latter half of the twentieth century other advanced industrial nations experienced many of these urban forces, though generally with less severity than the USA. Cities such as London, Paris, Amsterdam, Munich, Buenos Aires, Sydney, Melbourne, Auckland, and Toronto began to experience demographic shifts in their centers, rising suburbanization, and disinvestment in their downtowns, some neighborhoods, industrial areas, and waterfronts. In most cases private market interests were insufficient to reverse these trends and it was necessary for national, state, and local governments to step in.

The first era of central city revitalization was under way by the 1950s and is generally termed 'urban renewal.' While the details varied from nation to nation and city to city, urban renewal usually involved financing and rules provided by the national government, with program implementation by the municipal and/or state governments. Real property in declining central city enclaves was acquired, occupants were moved to other locations in the city, decaying structures were demolished, large areas were cleared and master plans for redevelopment were prepared. Land was sold cheaply, supportive public facilities were upgraded, and other incentives were provided. In the USA in many cases, hospitals and universities expanded into these areas. In downtowns, pedestrian walkways, parking garages and office buildings were constructed. Often, public or social housing for low- and moderate-income households was erected, although fewer units were built to replace the number destroyed, lowering population densities. By the 1970s a combination of the high public costs of urban renewal, the politically and legally contentious issues of land acquisition through condemnation and eminent domain, and the social disruption of compulsory household relocation brought urban renewal to a halt in several nations.

Gradually, national governments replaced urban renewal with a more selective approach. This process, never known by a singular identifying phrase, could be called 'urban conservation.' The emphasis in urban conservation was the rehabilitation and restoration of older buildings, their re-use for new purposes, isolated or 'spot' land clearance and relocation, and greater efforts to involve indigenous residents, businesses, and institutions in planning for the future of their communities. Instead of erasing older neighborhoods, concern was for extending their utility and, where necessary, adapting them to new uses. Closely allied with this trend was the growing popularity of historic preservation and an awakened interest (primarily among young adults) in architecture of the nineteenth and early twentieth centuries. These urban 'gentry' fed a process now known as 'gentrification.' While some of the gentry were the beneficiaries of government subsidies designed to spur urban conservation, most received no public assistance for renovating their own homes (Davies 1980, Gale 1984).

Urban conservation sought to revitalize other central city sections as well. Downtown storefronts were rehabilitated, streets were repaved and new lighting, vegetation, signage, and other features were added. Public spaces were built or expanded and new or upgraded mass transit systems were installed. Rotting piers and dilapidated factories and fish processing operations were removed from waterfronts and replaced by marinas, parks, pedestrian ways, restaurants, housing, hotels, and entertainment facilities (Breen and Rigby 1994).

Obviously, central city revitalization is a complex process, unfolding over several years, even decades, with multiple participants, complicated financing mechanisms, and a multiplicity of people who are affected in various ways. It may be understood from several disciplinary perspectives.

3.1 Physical Dimensions

Nothing conveys the impression of decline more readily than the sheer visual impact of deteriorating architecture, bridges, streets, street furniture, parks, playgrounds, schools, public transit systems, and the other physical manifestations of urban life. These many signs of deterioration contribute to an ethos of disrespect among residents, often resulting in poor maintenance, vandalism, and destruction. Those forced to live in slums, dispirited by their surroundings, are left to measure their personal worth in terms of the sights and sounds that greet them each day. Visitors to the city avoid these enclaves, if possible,

further isolating their residents and limiting their economic opportunities. Another impact of decline is the sheer ineffectiveness with which such conditions serve their residents and others. Uninhabitable dwellings and apartments offer no shelter, vacant factories and offices provide no jobs, empty shops offer no goods or services, deteriorating schools undermine educational processes, and broken playground equipment provides no recreation. The physical dimensions of urban decline usually result in a lack of confidence among property owners, banks, and developers which contributes to further disinvestment, thus compounding the dilemma.

3.2 Economic Dimensions

Closely related to these conditions are the economic dimensions of central city decline. Until the last quarter of the twentieth century, central city decline was blamed mainly on rising suburbanization in Western societies and the attendant devaluation of urban locations. Since then, the world has undergone increased economic restructuring, resulting in firms relocating business activity not merely to the suburbs but to other regions and nations altogether. As traditional employment in manufacturing industries declined, especially in central cities, employment in fields requiring higher education accelerated. Some cities 'captured' more of these jobs than others. Aggressive investment followed and in cities such as San Francisco, Boston, Sydney, Singapore, Hong Kong, Vancouver, London, Frankfurt, Geneva, Brussels, Madrid, and Paris, downtown, waterfront, and neighborhood revitalization has been widespread.

Questions have been raised of cities that have been less successful regarding how much financial support governments should supply to prop up urban economies. While nearly all cities advance revitalization programs and projects with government incentives such as loan, grant, and tax relief schemes, some have been more effective at attracting private sector investment than others. Should billions of dollars of scarce government resources continue to be invested in these less promising communities? Or, should they be spent on other social needs such as health care or environmental protection? Does it make more sense economically to help needy people in city centers to move to the suburbs, where job, housing, and education opportunities are often better? Economic issues such as these have accompanied the late twentieth-century phenomenon of central city revitalization.

3.3 Social Dimensions

The primary social cost associated with revitalization is the displacement or dislocation of low-income households, often racial or ethnic minorities, as the prices of housing appreciate due to increased investment activity. Rents and property taxes may be raised in response to these forces and tenants are forced to make way for more affluent successors. The socioeconomic composition of a residential neighborhood is likely to change as a result of these dynamics, and religious, ethnic, and cultural organizations may lose members. Small businesses in downtowns and neighborhoods may lose their lease and vacate their premises, leaving low-skilled employees jobless. In short, the positive economic outcomes of revitalization do not redound equally to the benefit of all who live or work in the city (Sanders 1980).

3.4 Political Dimensions

Whenever there are threats to the existing social and demographic composition of central cities, there are political implications, as well. While city administrators may promote revitalization activities, some elected officials with constituents in the affected areas are likely to resist. This is especially true in ethnic and racial minority voting districts that have elected a city councilperson, for example, from their own ranks. Fearful that the councilperson will not be re-elected if the composition of the neighborhood electorate changes significantly, they may promote organized resistance to reinvestment and displacement activities.

On the other hand, pro-growth political administrations in city governments, keen to attract revitalization activity, may overlook or disregard the negative externalities (i.e., unintended outcomes), such as household displacement, which result. Collaborating with developers, banks, corporations, and investment firms, so-called 'growth machines' or 'regimes' result, which foster economic growth with little regard for the negative effects on low-income and minority households, as well as their businesses and institutions (Logan and Swanstrom 1990, Teaford 1990).

3.5 Historic/Cultural Dimensions

As central city urban renewal and highway construction removed thousands of older structures during the 1950s and 1960s, a growing appreciation arose for the historic and cultural legacies being destroyed. Historic preservationists organized and local governments enacted historic landmark and district ordinances to protect older neighborhoods and sections of downtowns and waterfronts. Urban conservation replaced urban renewal as government officials and business interests recognized that a community's historic identification could enhance the local tourism economy. Railroad stations, public schools, warehouses, factories, hotels, office buildings, movie theaters, churches, and other structures, depending on their age and

stylistic identity, qualified for special protective measures. Revitalization schemes often adapted the older structures to new uses. In other cases only the street-facing facades of aged buildings were preserved, while the interiors were demolished and replaced with modern spaces. Notable examples of historic districts include the Vieux Carré in New Orleans, the Old City of Quebec, the Altstadt in Bonn, the Lace District of Nottingham, the Marais in Paris, and Gamla Stan in Stockholm.

3.6 Design Dimensions

Closely related to the historic significance of urban spaces are the challenges to architects and designers of integrating modern structures into historic settings. New office buildings and hotels, for example, can tower over cathedrals or public libraries built in the last century, casting shadows. Newer materials such as glass, concrete, and steel do not necessarily blend in with the brownstone, brick, or granite surfaces of another century. Larger structures generate more vehicular and pedestrian traffic, causing access and parking problems. Some successful design projects have created ensembles of existing and new buildings, joining them in complexes linked by glass-covered atria, pedestrian walkways, bridges, and other devices. In other cases, new spaces have been placed partially below-ground under and adjacent to historic buildings, thus presenting a less obtrusive profile at street level. A second challenge is the alteration of historic buildings for new uses. Sometimes called adaptive use, these conversions tax architects' design skills as they try to accommodate modern demands for elevators, access for the handicapped, heating and cooling equipment, fiber optic cable, and office devices such as computers, photocopiers, facsimile machines, and other communication tools. Equally challenging are the constraints imposed on older buildings by local building codes, especially in settings where earthquakes, cyclones, and hurricanes occur.

3.7 Behavioral Dimensions

At least since the late nineteenth century, questions have arisen about the effects of urban environments on human behavior. Leaders of social reform movements often cited the pernicious effects of crowding, inadequate sunlight, and poor air circulation on immigrants. Alcohol abuse, mental illness, prostitution, street violence, and other forms of social deviance were sometimes linked to the debilitating effects of the slums in which they occurred. Later, urban renewal advocates implied that tearing down these blighted areas and relocating their inhabitants to newly built public or social housing would substantially reduce the incidence of social deviance. Thus, a connection—still unsupported by empirical research—was suggested between urban revitalization and behavioral change. If there is a consensus among social and behavioral scientists, it is that urban environments have moderate impacts on human behavior but that other conditions such as poverty, racial prejudice, and crime have greater explanatory power. Ironically, studies of those who were forced to move due to urban renewal or revitalization processes found that some people experienced greater personal hardships, at least over the short run, when taken out of their earlier neighborhood (Gans 1982, Nasar 1988, Schill and Nathan 1983).

3.8 Environmental Dimensions

The notion of environmental sustainability has achieved prominence since the 1970s. It argues that there are certain natural limits on the capacity of individual ecosystems to support human life without significant environmental degradation. Suburban sprawl is viewed as unsustainable, due to its extravagant consumption of space, prime agricultural soils, water supplies, water pollution, and other effects. Urban decline is seen as a consequence of suburbanization and sprawl. Central city revitalization has been advocated as a means to redirect growth from suburbs to central cities. Environmentalists and some urban planners have argued that older city centers should be revitalized and reused to meet continued human needs for urban space, rather than allowed to become underutilized, decaying slums.

4. Methodological Issues and Future Research

As a public policy issue, central city revitalization raises questions that are generic to public policy research. Are the costs of revitalization outweighed by the benefits accruing to the public? Or, in a somewhat different frame of reference, how effective are revitalization expenditures in achieving designated social objectives such as increasing property tax receipts, lowering unemployment rates, and increasing household income? Are there less expensive options for achieving the same level of effectiveness? Given the complexities of revitalization programs and the duration of time over which they must unfold to show measurable effects, these questions are difficult to answer with certainty.

Another difficult methodological issue is sometimes termed 'the substitution dilemma.' It asks, to what extent do public expenditures in revitalization programs engender benefits that would have been forthcoming even in the absence of those expenditures? For example, if a city government offers tax incentives to developers to invest in downtown rebuilding projects, would they have made those investments without the incentives? If market conditions were quite negative

prior to commencement of the downtown program, the incentives may have been necessary. However, as conditions improve (especially after the first few developer projects have been completed), continued tax incentives for subsequent projects may not be necessary. Yet, the city government may continue to 'substitute' public subsidies (e.g., tax incentives) for private investment when, in fact, the investment climate has improved to the point that developers would be willing to undertake projects even without those subsidies.

A third research issue questions the impact of new technology, especially telecommunications, on the physical, economic, and social character of central city life. Suburbanization and global economic restructuring have reduced the importance of many central cities as centers of culture, religion, and commerce. With the introduction of telephones, radios, televisions, and automobiles within the past century, the necessity for central places began to unravel. Today, with the internet, computer networks, facsimile machines, cellular and digital telephones, and other devices, businesses and individuals have more locational freedom than at any time in the past. Large headquarters buildings in downtowns may become less important, especially among global industries with facilities and clients in several nations. As retail sales increasingly take place over the Internet, the need for downtown (and suburban) department stores and speciality outlets will probably diminish. With newer telecommunications equipment, white-collar workers in many fields are able to carry out their duties from their homes or other remote locations. Thus, research will be needed to understand the evolving purposes of central cities for meeting residential, commercial, cultural, and entertainment needs in the future. To be sure, there will always be human need for contact with others. Whether those needs will be sufficient to sustain large central cities into the twenty-first century, especially in their present configurations and organizational structures, remains to be seen. Urban revitalization initiatives sympathetic to the evolving future of central cities must be carefully examined.

See also: Cities: Capital, Global, and World; Community, Social Contexts of; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Race and Urban Planning; Urban Geography; Urban Growth Models; Urban Places, Planning for the Use of: Design Guide; Urban Sociology; Urban Sprawl; Urban Studies: Overview; Waterfront Planning.

Bibliography

Barnett J 1986 *The Elusive City*. Harper and Row, New York

Breen A, Rigby D 1994 *Waterfronts: Cities Reclaim Their Edge*. McGraw-Hill, New York

Davies H W E 1980 Neighborhood revitalization: the British experience. In: Rosenthal D (ed.) *Urban Revitalization*. Sage, Beverly Hills, CA

Gale D E 1984 *Neighborhood Revitalization and the Post-industrial City: A Multinational Perspective*. Lexington Books, Lexington, MA

Gans H 1982 *The Urban Villagers: Group and Class in the Life of Italian-Americans*. Free Press, New York

Kostof S 1991 *The City Shaped: Urban Patterns and Meanings Through History*. Little, Brown, Boston

Levitt R L (ed.) 1988 *Cities Reborn*. Urban Land Institute, Washington, DC

Logan J R, Swanstrom T 1990 *Beyond the City Limits*. Temple University Press, Philadelphia, PA

Morris A E J 1994 *History of Urban Form: Before the Industrial Revolution*, 3rd edn. Longman, New York

Nasar J L 1988 *Environmental Aesthetics: Theory, Research and Applications*. Cambridge University Press, New York

Redstone L G 1976 *The New Downtowns: Rebuilding Business Districts*. McGraw-Hill, New York

Sanders H 1980 Urban renewal and the revitalized city. In: Rosenthal D (ed.) *Urban Revitalization*. Sage, Beverly Hills, CA

Schill M H, Nathan R P 1983 *Revitalizing America's Cities: Neighborhood Reinvestment and Displacement*. SUNY Press, Albany, NY

Teaford J C 1990 *The Rough Road to Renaissance*. Johns Hopkins, Baltimore

Van der Ryn S, Calthorpe P 1986 *Sustainable Communities*. Sierra Club Books, San Francisco

D. E. Gale

Urban Planning (Chinese), History of

The principles of Chinese urban planning grew out of the heartland of Chinese history and culture. In broad historical perspective, Chinese city planning can be divided into millennia-old imperial planning and twentieth century modern planning. Chinese urban planning history ultimately reflected Chinese cultural visions, political ideologies, and power relations. Chinese imperial urban planning, characterized by its basic design principles of centrality, axiality, and rectilinearity, was based on a syncretic world view which emphasized the centrality of the emperor as the pivot of the four quarters of the universe, and combined traditional Chinese cosmology, the *yin* and *yang* and five-element theories, and *fengshui* geomancy. Modern city planning is a product of the interaction and fusion between Western concepts of city development and Chinese indigenous urban traditions, which reflects tensions, conflicts, and accommodations between impulses of global forces or ideological and political monuments to reduce Chinese cities into decontextualized uniformity and the heightened appreciation for Chinese cultural roots and local sensibilities in spatial arrangements. The future of Chinese urban planning is destined to transcend both modernism driven by the pursuit of rationality and

functionality and rigidified socialist politics and be based on a pluralistic and multidimensional sense of Chinese history and culture.

1. *Imperial Urban Planning*

The depths of the planning and design principles of traditional cities have been explored by scholars such as Paul Wheatley, Arthur Wright, and Sen-dou Chang. In a comparative study, Wheatley analyzed the origins and nature of ancient Chinese urban planning, arguing that the ceremonial center represented the earliest stage of Chinese urban development and the cosmo-magical symbolism played a decisive role in defining the functionality of the ancient Chinese city (Wheatley 1971). While some scholars, based on new archaeological data, have recently challenged the ceremonial-theocratic or 'city as temple' thesis of urban genesis, the cosmic symbolism in the design of cities that perhaps found a more explicit expression in China than in many other civilizations was further emphasized. The symbolic and cosmological elements associated with the design and location of ancient Chinese cities, which included early cosmological traditions, Taoist principles, *yin* and *yang* and Five Elements schools, and the geomantic (or *fengshui*) considerations were idealized and systematized as general urban planning principles by Confucianist philosophers in the Han dynasty (206 BC–AD 222) (Wright 1977). At the same time, these rules were appropriated to symbolize and sanctify the centrality of the emperor and Confucian social order that emphasized power, authority, and hierarchy.

Consequently, the uniqueness of premodern Chinese city planning lies in the fact that the buildings and their locations within the city were pre-established by age-old traditions set forth in revered literature or documents of the classical age (Steinhardt 1990). These ritual and divination texts include *Zhou Li* (*The Rites of Zhou*) and *Yi Jing* (*The Books of Changes*). According to *Zhou Li Kaogongji*, a planning and construction text, a city should be built in a square with three gates on each side. Preferably surrounded by hills in the distance with a river system running through, the wall- and moat-enclosed city should have an area of about seven square miles, a good size for a pedestrian city. Inside the city, there should be thoroughfares running north–south intersecting with other east–west thoroughfares. Major buildings in the city should face south, the best of the four cardinal directions. On the east side of city an ancestral temple should be built, while on the west side there should be a state shrine. The market place is usually situated in the north, while the imperial palaces lie in the southern part of the city. Built in the fifteenth century, Beijing, with its layout and structure epitomized by a seemingly endless maze of walled compounds within walled compounds, best exemplifies a highly regimented city conditioned by the prescriptions of the imperial cosmology and its geomantic modularity (Wu 1986, Meyer 1991).

Traditional city plans were almost always square or rectangular rather than circular or irregular, reflecting Chinese cosmological belief that heavens were round while the earth was square. Nearly all cities were walled to protect imperial palaces, temples, granaries, and residences from barbarian invasions, tribal uprisings, and peasant rebellions. Huge wall-gates were usually constructed, connecting between different parts of the walled city and between the city and the outside world. Walls were so important to Chinese cities that the characters for city and wall are in fact identical (Chang 1977). Another distinctive feature of traditional city planning was the axial balance and interdependence of urban spaces. Compared with the European Renaissance palace, as at Versailles, where the open vista is concentrated upon a single building and the palace itself was detached from the city, the Chinese conception was much grander and more complex. The central axis of Beijing, for example, was composed of hundreds of buildings, and the palace itself was only part of the larger organism of the whole city. Although so strongly axial, there was no single dominating center or climax, but rather a series of architectural experiences. This Chinese form of the great architectural ensemble, 'combined a meditative humility attuned to nature with a poetic grandeur to form organic patterns unsurpassed by any other culture' (Needham 1975).

Traditional Chinese cities were planned and designed to reinforce important power relations, especially the cultural hegemony and political dominance of imperial monarchs. City planners were determined to uphold imperial power through shaping the design and configuration of the built environment. The urban spatial hierarchy represented powerful testimony to and physical proof of the imperial order. Maintaining a highly structured and carefully planned city form was an important means through which Chinese emperors legitimized their position as both rulers and guardians of tradition. The alteration of an accepted design was therefore considered a challenge to the imperial order.

2. *Modernist Urban Planning*

China's modern planning experience is primarily a twentieth century phenomenon; it has undergone three important phases of development: the modernist (1900–50), socialist (1950–84), and global capitalist (1984–2000).

The principles of imperial urban planning came under attack in the nineteenth century, when Western industrial countries transformed a number of coastal Chinese cities into treaty ports, which became enclaves of foreign political domination as well as Western

trade and commerce. Chinese cities elsewhere were also pressured to adopt Western ways (Dong 1989). Westerners brought along a host of entrepreneurial, technical, and modernist ideas about urban life and industrial prosperity, as well as a body of political, social, and economic values. In addition, the urban transformation was a direct result of domestic political changes in early twentieth century China. Modern urban planning reflected a shift in indigenous ideological discourse from emphasizing the supremacy of the imperial power to one that stressed the primacy of a civil society and people's rights. Both Western and modern Chinese ideology confronted the integrity of the Imperial-Confucian world system and its cosmology with a series of powerful secular trends: modernization, Westernization, and later on, revolutionary nationalism.

As a result, the early twentieth century witnessed a significant change in planning principles and the transformation of Chinese urban landscape. A variety of political, intellectual, and commercial leaders sought to 'modernize' their cities according to their understanding of what modernity meant. The quest for modernization took a number of different forms, but there was also a number of common elements. The newly created Chinese republic required new types of public spaces—parks, public squares, and meeting halls. Municipal governments sought to improve public health, with such measures as building modern waterworks and sewer systems. Many cities sought to improve urban transportation with streetcars and railroads. Through both public and private efforts, the architecture of Chinese cities was transformed, with new-classical Western architecture in such treaty ports as Shanghai and Canton and a variety of attempts to blend modern materials and covenience with distinct Chinese styles. One striking aspect of this effort at urban renewal was the extent to which the old Chinese cities, often viewed as the prototype of the pre-industrial city, were now subjected to a new generation of planning and new attempt to discipline the urban population.

The Republican government was bent on tearing down city walls, remodeling city gates, and constructing or widening roads to allow passage of railroads, streetcars, and general traffic. Proponents for dismantling city walls argued that since the introduction of modern warfare, the ancient city walls and gates became useless. Moreover, the walls blocked mobility and the flow of traffic, and prevented cities from growing spatially to meet the needs of population increase and economic development. Based on the modernist ideology of rationality, the republican government demolished city walls in many Chinese cities. Opponents of modern city planning, however, emphasized the role monumental architecture such as the walls and gates had played in Chinese culture. They argued that the walls and gates served to anchor the collective memory of Chinese cityscape, reminding residents and visitors alike of the enduring power of culture and tradition. They were skeptical of the modernist government's effort to replace traditional monumental legacy with monuments that intentionally express a 'modern' society. Some even evoked the traditional ideas of *fengshui* to resist the modernists' attempts to introduce modern technology to reconfigure the urban order. In short, they opposed modern reconstruction not as defendants of the *ancien regime*, but as advocates of cultural sciences and in the secular effort to preserve and draw inspiration from the one heritage that was uniquely their own.

3. *Socialist Urban Planning*

The second phase of modern Chinese urban planning covers roughly the first 30 years of the history of the People's Republic of China. Shortly after taking over power in China in 1949, the Chinese Communists embarked on a course of urban reconstruction based on a socialist agenda. Swayed by radicalism and the utopian revolutionary ideology of Mao Zedong, Soviet-style planning and design dominated Chinese cities. During this period, Chinese urban planning focused on industrial production and socialist reconstruction. Many Chinese cities were transformed into industrial centers, and capitalist consumer cities like Shanghai were criticized as being exploitative and parasitic (Kirkby 1985). The government cleaned out old teahouses and pleasure zones, and lined major streets with official monuments, government offices, standardized state-run stores, and utilitarian apartment buildings. Mao envisioned Beijing to be a symbol of the Chinese socialist state, a role to be enforced with the buildings of Soviet-style public structures. In a symbolic break with the past, Tiananmen Square replaced the former embodiment of imperial power—the Forbidden City—to become the new political center of the nation. The socialist monuments of this period tossed together ingredients of Soviet, classical Chinese, and modernist architecture to create a new socialist identity for revolutionary China.

Ideas of cultural conservation were not completely ignored during this period. Liang Sicheng represented a group of architectural experts and city planners who became cultural advocates during the radical years of Maoism. Liang himself was an architectural historian who had been trained at the University of Pennsylvania in the 1920s. As early as 1950, he drafted a plan to keep all of old Beijing enclosed in city walls, turning it into a great museum and cultural center, and to build a new, modern city next to it. He insisted that the large structures of Beijing's street grid, its system of enclosures and its axial symmetry—all landmarks in the world's history of urban planning—must all be preserved to maintain the historical and cultural identity of Beijing. As for the city walls, he had an idea of lining the route atop the walls with flowers, benches,

transforming the ring of walls into an elevated park. This plan, he believed, would allow Beijing to leapfrog through time while avoiding all the mistakes Western cities had made: a modernization that would contain the explosive effects of rapid development and keep an ancient masterpiece for aesthetic appreciation (Fairbank 1994). Liang's proposal bore a striking parallel to what Eric Hobsbawm calls 'the invention of tradition' (Hobsbawm and Ranger 1983). Sensitive to both the homogenizing force of modernity and the deep-rooted sophistication of Chinese culture, Liang sought to restore and recreate an older urban fabric and rehabilitate it to new uses and attempted to repackage the old infrastructure of urban space for new style recreation and entertainment. To him, tradition and modernity could go together, and modern transformation and cultural conservation caused no apparent conflict. However, Liang's alternative planning vision was ignored by the government in favor of demolition and other radical urban planning principles.

4. *Contemporary Urban Planning*

In China today, people are no longer concerned with the revolutionary and political ideology of Maoist urban planning, but become preoccupied with modernization along Western lines. While urban planning is still largely controlled by the government what is increasingly evident is the participation of global, profit-driven, capitalist and market forces from Hong Kong, Taiwan, and the Western countries. Modernism returned with a vengeance. Global capitalistic forces are competing with localized cultural traditions, giving rise to an identity crisis in contemporary Chinese cities. Many old cities are being quickly stripped of their character as ancient buildings are torn down for office buildings, apartment blocks, freeways, and shopping centers, edifices which are seen by city officials and real estate developers as a symbol of modernization but which have alarmed culturally sensitive citizens.

Conflict between global and local cultures and the problems of modern economic development vis-à-vis cultural conservation are at the heart of the debate over the future of Chinese urban planning. In the feverish race to transform Chinese cities into world-class metropolises, many urban planners are also struggling to come to terms with China's rich urban heritage. The major urban planning question has become how to excel in the global community without losing China's national identity. Some began to lament the loss of the old urban fabric of walls and gates, alleyways, and courtyard-style residential houses, and especially the close-knit community atmosphere and neighborly connectedness. There is growing complexity in Chinese popular attitudes toward cultural heritage and local identity. Many people realize that in the face of ever-stronger global forces for homogenization through technological change, increased mobility, and improved communication, a primary goal of Chinese urban planning policy must be to retain culturally essential features of China's urban past and to ensure the continuity of Chinese history and culture.

See also: China: Sociocultural Aspects; East Asian Studies: Culture; East Asian Studies: Economics; East Asian Studies: Gender; East Asian Studies: Politics; East Asian Studies: Society; Urban Geography; Urban Growth Models

Bibliography

Boyd A 1962 *Chinese Architecture and Town Planning*. 1500 BC–1911 AD. University of Chicago Press, Chicago

Chang S D 1977 The morphology of walled capitals. In: Skinner G W (ed.) *The City in Late Imperial China 75–100*. Stanford University Press, Stanford, CA

Dong J 1989 *Zhongguo Chengshi Jianshe Shi* (A history of urban construction in China). Zhongguo jianzhu gongye chubanshe, Beijing, China

Fairbank W 1994 *Liang and Lin: Partners in Exploring China's Architectural Past*. University of Pennsylvania Press, Philadelphia, PA

Hobsbawn E, Ranger T 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Kirkby R J R 1985 *Urbanization in China: Town and Country in a Developing Economy, 1949–2000 AD*. Columbia University Press, New York

Meyer J F 1991 *The Dragons of Tiananmen: Beijing as a Sacred City*. University of South Carolina Press, Columbia, SC

Needham J 1975 The planning of towns and cities. *Science and Civilization in China* **4**: 71–9

Skinner G W 1977 *The City in Late Imperial China*. Stanford University Press, Stanford, CA

Sit V F S 1995 *Beijing: The Nature and Planning of a Chinese City*. Wiley, New York

Steinhardt N S 1990 *Chinese Imperial City Planning*. University of Hawaii Press, Honolulu, HI

Wheatley P 1971 *The Pivot of the Four Quarters: A Preliminary Enquiry into the Origin and Character of the Ancient Chinese City*. Aldine, Chicago

Wright A 1977 The cosmology of the Chinese city. In: Skinner G W (ed.) *The City in Late Imperial China*. Stanford University Press, Stanford, CA, pp. 33–73

Wu Liangyong 1986 *A Brief History of Chinese City Planning. Urbs et Regio* 38. Kasseler Schriften zur Geographie und Planning, Kassel, Germany

M. Shi

Urban Planning: Competitions and Exhibitions

Competitions and exhibitions are means of stimulating the production and exchange of ideas and techniques in urban planning. Competitions aim to elicit ideas

from various people on a defined topic, exhibitions permit the comparison of projects and realizations from several locations. Both are aimed at enhancing good design and prompting public discussion and awareness, possibly consensus.

1. Competitions

Competition is one of several options available for the choice of an urban plan and a planner. Others include: direct commissioning, holding of workshops or planning seminars, the appointment of an accompanying committee or a consultant. Competitions are good ways to find solutions, if the program and the constraints are clearly defined, so that the participants can make meaningful statements within a limited period of time. Competitions allow a selection from a variety of proposals, even if no project is realized as such, and may help to stimulate further insights into the planning problem and ideas for realization.

Competitions can take several forms. They can be open to the general public or to members of the relevant professional institute, or they can be restricted to a limited number of participants, appointed by the organizer or chosen after an application process. These procedures may also follow upon each other. Restricted competitions, which include a payment for all participants, have been criticized for being limited to famous names or big companies. Anonymous, open competitions allow unknown and young planners to participate and present their ideas, and may result in new concepts. This can, however, bring about large numbers of participants, or lead to the choice of an author who lacks the means and experience to realize the proposal. As only prizewinners are paid, much unpaid work is done, particularly in the case of complex urban competitions, which therefore are sometimes considered to be unproductive. However, even unbuilt contributions are part of the architectural culture and history of a city and a period. In some cases unrealized competition projects have had a stronger impact on planning history and theory than built ones.

Competitions are often seen as particularly democratic, and are used in democratic societies where the governments have to justify their expenses and choices and where they try to include citizens in the decision-making process. The democratic character of a competition depends primarily on whether its preparation and the ensuing realization are part of a public and open discussion. Competitions can also be used in a dictatorial situation. Thus, the outcome of a competition depends on the participants and is strongly influenced by the aims and quality of the competition brief and the organizers. Their character—public institutions, private sponsors, professional organizations, architectural schools or magazines—and concrete power over the competition site determine whether the competition leads to realization or is limited to creating ideas, whether it is supposed to adapt to the existing context or provide a long-term vision for the future.

1.1 History of Major Large-scale Competitions

Architectural competitions have been held for many centuries, particularly for the design of religious and public buildings. City planning competitions are more recent. Early examples are the international idea competitions for urban embellishment opened by the Russian empress Catherine II for the city of Saint Petersburg (1763) or the contests for the Place Louis XV (today's Place de la Concorde) in Paris (1748). Urban planning was usually seen as a part of architecture, and the competitions held by professional and educational architectural institutes included urban design projects, such as for a 'Seaside resort for a sovereign on the Mediterranean coast' a topic at the Paris Ecole des Beaux Arts in 1872. Competitions were used at educational institutes for grading and awards such as the 'Rome prize' which included travel and a several year stay in Rome. These projects were not made for realization, but nevertheless reflected the design spirit of their time. The 'Rome prize' winners were assured an honored part in public building projects on their return.

The history of urban competitions reflects the development of the planning profession: Following industrialization and rapid city growth, early competitions tackled large areas on the outskirts or even the whole city. With the growing scale and complexity of cities, their compounds became the main topic of later competitions. A comprehensive and international analysis of urban planning competitions is still lacking and the number of competitions is too large to be analyzed here: analysis of the late 1920s, for example, shows between 15 and 30 urban planning competitions annually in Germany alone (Becker 1992). Only some major examples can be presented here. In spite of the large number of competitions held over the last century, major transformations and extensions of cities, including the work of Eugène Haussmann in Paris (1850s–70s) or the extension of Berlin under James Hobrecht (1862), were and are based on direct commissions or procedures other than competitions.

One of the most interesting examples of nineteenth-century urban planning, the transformation of the Vienna fortification zone, did result from an international urban competition in 1857/1858 (Albers 1997, Hall 1997); it requested ideas for traffic planning, plot division and urban beauty for a zone both sides of which were already built. None of the prize-winning or recommended designs was realized, but they were taken as references for a basic urban plan—a pro-

cedure often used in urban competitions which is not faithful to the original ideas. In 1892, Vienna was once again the site for a major urban competition that, like the competition for Munich in the same year, was aimed at obtaining ideas for a master plan of the whole city. As such, they can be seen as a milestone in urban planning. What is more, both competitions showed the impact of major urban planning theoreticians of that time, such as Reinhard Baumeister, Camillo Sitte, and Josef Stübben.

Another landmark in urban planning history was the Berlin competition of 1908/1910. Initiated by architects' associations and organized by several independent municipalities, it concerned a large urbanized area of more than 5 million inhabitants. Although this competition was not followed by immediate realization, it helped clarify major problems and led to several competitions for city parts. Its results were used for promoting general awareness of urban problems and educating the public through a large urban planning exhibition which also featured foreign examples. Through exhibitions, publications in magazines and private brochures the competition formula and its results were disseminated and became major references, stimulating other large-scale competitions in major and minor European cities in the late nineteenth and early twentieth century including Göteborg, Strasbourg, Paris, Barcelona, and Zagreb (Albers 1997).

The early European competitions concentrated on remodeling and extending existing cities. Meanwhile, new nations around the world had a strong desire for cultural representation. This led to the planning and construction of new capital cities. Competitions held for Canberra in 1911/1912 or for Brasilia in 1956/1957 were means to obtain a state-of-the-art urban plan ready to impress the nation and foreign countries. The competitions gave the authorities the opportunity to consult with leading planners and, as in the case of Canberra, eventually appoint a foreigner for the realization of the plan, albeit over a long period. For the evaluation of competitions it is particularly interesting to analyze the process from the competition design to the actual city.

In Europe, competitions for a completely new city are rare, although some proposals have been put forward following private initiatives. The few exceptions were characterized by a utopian touch from the very beginning. The search for a capital of Europe, in particular, led to several city competitions and projects for new cities. Due to the lack of large tracts of land, and the dominance of nationalist thought, most proposals were located close to existing cities. Thus, they could profit from existing infrastructures, but were not free to develop an independent identity. These competitions were, generally, flawed by the lack of political and financial power, as well as by the unclear future of the new European communities. What is more, the largely academic design concepts reflect the lack of conviction of the organizers, they include few innovative solutions (Hein 1996).

Reconstruction after World War II stimulated urban competitions of various scales in several countries, particularly Germany, leading to intense and often controversial discussions. The most political of the large-scale competitions was probably the international urban idea competition 'capital Berlin' of 1957/58 (Hein 1991). Organized by the governments of West Germany and West Berlin it covered an area which was largely under East German administration. A preparatory committee including practitioners from major Western cities outlined a plan for inner-city development that guided urban development for the Western part for many years and served as the basis for the competition. The competition itself was a political statement to the world that West Germany had developed into a democratic and modest nation, and also that the West German government was not abandoning West Berlin. The East German government reacted to this provocation with a competition of its own for the city center, another extension of the Cold War into planning.

1.2 Competitions for Limited Areas and Special Functions

Projecting an entire city area brings with it numerous difficulties and demands extensive knowledge that cannot be expected from competition participants. The twentieth century therefore concentrated on public and private competitions for limited areas and clearly defined tasks such as campuses, government centers, commercial or recreational centers or urban parks. Examples can be found worldwide and include the privately sponsored competition for the University of California (1897/1899), the city-held competitions for educational districts in Tokyo (1940s), the design competitions for civic centers in the USA, or the urban design competition for the government districts in the Spreebogen and Spreeinsel areas in Berlin (1993/94), as well as the competitions for the Volkspark in Hamburg (1906) or the Peace Park in Hiroshima (1949) (Zwuch 1993, 1994). Recent attempts at recreating a multifunctional city have led to competitions combining office, commercial and housing functions as in the case of the Potsdamer Platz or Alexanderplatz competitions in Berlin (1991/1993).

The largest number of competitions, however, has probably been organized for residential layouts. Examples can be found in various countries; examples from Germany are given here for illustration. The country has been a major organizer of competitions over the last century and their topics and results reflect the main concerns of housing over this period. Thus, early competitions focused on small-city and company housing such as for the Krupp colony Altenhof in Essen (1892). Metropolitan housing, economic land

use and the creation of a rational block structure became central topics in the competitions of 1911 for Tempelhofer Feld and Schöneberg South which followed on the Greater Berlin competition of 1910. Convinced that competitions led to the best solutions for metropolitan housing, the city of Hamburg promoted several competitions for social housing areas through the 1920s. The search for an economical solution for residential areas found its climax in the competitions for the experimental district Berlin-Haselhorst (1928) and the Dammerstock-Siedlung in Karlsruhe (1929), both of which featured strict parallel housing blocks. Competitions for new towns dominated the 1950s and 1960s, while the return of housing functions to the city centers has become a major topic since the 1980s.

2. *Exhibitions*

Exhibitions are an important means to present urban planning theories and concepts to the specialists and the general public. They can follow on competitions or accompany conferences but can also be events on their own. In the early years of the profession, topics of urban planning were addressed by various specialists, ranging from the hygienists, to architects and the garden city organizations. These issues were also featured in exhibitions accompanying major conferences. Two forms of exhibitions can be distinguished: the exhibition of projects and plans, or the realization of an urban district.

Major exhibitions on urban planning have been held since the turn of the century, reflecting the ongoing debate. In 1903, Dresden was the site of the first German Municipal Exhibition, presenting the achievements in engineering and urban planning of several German cities. In 1911, an exhibition on hygiene was held in the same city, which included a large display of urban planning. Other exhibitions were held in parallel with congresses of the International Garden City and Town Planning Association or with international architectural congresses, such as that of London in 1906 which displayed works of major planners including Raymond Unwin, Josef Stübben, Charles Buls, and Eugène Hénard.

Key events in urban planning history were the exhibitions held in Berlin and London in 1910, which featured projects from numerous countries. They aimed to transmit political, cultural and esthetic statements. The London exhibition was held in parallel with an international conference on town planning which had an impact as far away as Japan. This effect was further promoted by major publications by Werner Hegemann based on the exhibitions of Berlin (1910), Düsseldorf (1912), and Göteborg (1923) (Hegemann 1913, 1925). The interest in urban planning at that time is revealed by the fact that three major urban planning exhibitions were held in 1913 alone: in Nancy, Ghent and New York. Others followed, often inspired by local planning problems as was the case for the 'Coventry of Tomorrow' exhibition in 1939 or the reconstruction exhibitions 'Berlin Morgen' of 1946.

Planning exhibitions today fall into two major categories. They can have practical tasks, such as the public presentation of concrete proposals. As such, they are part of the planning process of many countries which aim at promoting democratic participation. They can also be primarily cultural events, displaying the history, present or future of planning. These exhibitions have increased over the last decades together with the proliferation of unbuildable projects and the creation of architectural archives and museums. These institutions' exhibitions present historic, present and future planning, and are instruments for the promotion of particular concepts or styles. They may even intervene in planning as was the case with the 'Berlin Morgen' exhibition in Frankfurt am Main in 1991 where proposals were presented that later reappeared as competition entries (Lampugnani and Mönninger 1991).

Exhibitions can also take built form, provisional or permanent. Built exhibitions are a means to try out visions and promote urban concepts. Thus, the American world fairs in Chicago (1893) or in St. Louis (1904) were manifestations of City Beautiful concepts and helped to promote the new style. Several international building exhibitions were held, the most famous of them in Germany, where each epoch was represented through its own exhibition on the function and form of residential design. They include such diverse examples as the artists' colony at Mathildenhöhe near Darmstadt (1901), modernist projects such as the Dammerstock mentioned above or 'traditionalist' examples illustrated by the Stuttgart-Kochenhof Siedlung (1933). Building exhibitions with an urban concept can be found also after World War II. Well-known examples are the reconstruction of an inner-city district as a high-rise housing area, the Hansaviertel of the Interbau (1957), the International Building Exhibition (IBA) in Berlin of the 1980s, which aimed at a critical reconstruction of the city (Faskel and Barth 1988), or the International Building Exhibition Emscher-Park in the Ruhr region (1990s).

3. *A Cultural Phenomenon*

Competitions and exhibitions have furthered the development of modern planning through the promotion of exchange, particularly between European countries and North America at a time when there were no examples to follow. They were helpful tools in the search for the best functional and esthetic solution.

There are, however, cultural differences in the use of these instruments. Germany, for example, has a long tradition of competitions and exhibitions, while other countries with different planning traditions and aims, Japan, for example, has realized only a limited number of competitions with limited impact. The country favors instead the invitation of foreign experts, study trips, the presentation of foreign projects in magazines or book translations for the introduction of foreign ideas.

See also: Planning Ethics; Planning, Politics of; Planning Theory (North American); Strategic Planning; Urban Growth Models; Urban Places, Planning for the Use of: Design Guide; Urban Sprawl

Bibliography

Albers G 1997 *Zur Entwicklung der Stadtplanung in Europa*. Bauwelt-Fundamente. Vieweg, Braunschweig

Becker H 1992 *Geschichte der Architektur und Städtebauwettbewerbe*. Verlag W. Kohlhammer, Stuttgart

Cramer J, Gutschow N 1984 *Bauausstellungen. Eine Architekturgeschichte des 20. Jahrhunderts*. Verlag W. Kohlhammer, Stuttgart u.a.

De Jong C, Mattie E (eds.) 1994 *Concours d'Architecture 1792 à nos jours*. Taschen Verlag, Cologne, Germany

Faskel B, Barth E 1988 *Ideen für Berlin: Die städtebaulichen Wettbewerbe seit 1982*. Senator für Stadtenwicklung und Umweltschutz, Berlin

Haan H de, Haagsma I, Sharp D, Frampton K 1988 *Architects in Competition; International Architectural Competitions of the Last 200 Years*. Thames and Hudson, London

Hall T 1997 *Planning Europe's Capital Cities: Aspects of Nineteenth Century Urban Development*. E and F N Spon, London

Hegemann W 1913 *Der Städtebau nach den Ergebnissen der Städtebau-Ausstellungen in Berlin und Düsseldorf 1910–1912*. 2 vols. E. Wasmuth, Berlin

Hegemann W 1925 *Der Städtebau nach den Ergebnissen der internationalen Städtebau-Ausstellung Gothenburg*. E. Wasmuth, Berlin

Hein C 1991 *Berlin Städtebaulicher Ideenwettbewerb 1957/58*. Gebr-Mann Verlag, Berlin

Hein C 1996 European Unification and the Planning Issue. *Comprehensive Urban Studies 59*. Center for Urban Studies, Tokyo Metropolitan University, pp. 85–112

Lampugnani V, Mönninger M (eds.) 1991 *Berlin Morgen. Ideen für des Herz einer Großstedt*

Miyake R 1991 *Toshi to Kenchiku Competition*. Kodansha Tokyo, V and K Publishing, Laren

Omi S 1986 *Kenchiku sekkei kyôsô/A History of Architectural Competitions in Japan*, Kajima Shuppankai, Tokyo

Speiregen P 1979 *Design competitions*. McGraw Hill, New York

Zwoch F (ed.) 1993 *Capital Berlin: Parliament District at the Spreebogen: International Competition for Urban Design Ideas/Hauptstadt Berlin: Parlamentsviertel im Spreebogen: Internationaler Städtebaulicher Ideenwettbewerb*. Bauwelt/Birkhäuser, Berlin

Zwoch F (ed.) 1994 *Capital Berlin: Central District Spreeinsel: International Competition for Urban Design Ideas/Hauptstadt Berlin: Stadtmitte Spreeinsel: internationaler Städtebaulicher Ideenwettbewerb*. Bauwelt/Birkhäuser, Berlin

C. Hein

Urban Planning: Growth Management

'Growth management' aims to influence the location, magnitude, and timing of urban development through land use and capital facility planning, and development permitting. Unlike conventional planning whereby land is designated for particular land uses and facilities are extended when market demand dictates, growth management is a more conscious effort to shape development patterns during discrete intervals of time based on explicit urban form objectives.

1. Historical Context

The earliest growth management effort might have been recorded in the Old Testament:

> *The Lord said to Moses … Command the people of Israel, that they give to the Levites … cities to dwell in; and pasture lands round about the cities … . The pasture lands of the cities … shall reach from the wall of the city outward a thousand cubits all around* (*Numbers* 35: 1–4)

This command is clearly aimed at containing urban development within a boundary so that farming can occur around it. Britain's 1947 Planning Act extended this concept nationally by nationalizing greenbelt land combined and directing local government to contain urban development within stoplines.

American experience is less draconian. The earliest growth management effort appears to have came out of Milford, Connecticut in the early 1950s, which mapped high and low priority development areas. Development in low priority areas needed to demonstrate that adequate public facilities existed to serve it without prematurely directing development away from existing urban areas. Moreover, approved projects could only be constructed in increments of 20 percent annually.

More famous is the plan devised by Ramapo township, New York, in the mid-1960s. It provided for township buildout over an 18-year period. A capital improvements program helped to pace buildout during that period. Developers had to demonstrate that their project was consistent with the buildout plan (through a point system) and, if not, either wait for infrastructure to reach their site or install it themselves at their own cost.

A version of growth management, called 'growth control,' came from Petaluma, California, in the early 1970s which capped development at 500 dwelling units annually. The cap was suballocated to different housing types to assure reasonable diversity of housing opportunities.

2. Statewide Efforts

Only a few American states engage in growth management on a statewide level, although that number is rising.

Hawaii, the nation's newest state, has the nation's oldest statewide growth management and for good reason: it is composed of islands for which there are very few alternatives to sharing development needs with agricultural and environmental needs. Through efforts by a state planning agency pursuant to the Hawaiian Land-Use Law of 1961, all land on major islands is designated for urban, agriculture, low density rural, and conservation land uses. Counties (individual islands) have very little say about those allocations.

The state with the longest history of statewide growth management among the contiguous 48 states is Oregon. The Land Conservation and Development Act of 1973 established a state commission and an administrative department. The commission's primary role is articulating statewide planning goals (the original 15 goals have grown to 20) and deciding whether locally prepared land use and facility plans meet them. Like Hawaii, land in the state is designated essentially for urban, agriculture/forestry, low density rural, and conservation uses. Unlike Hawaii, local governments have a significant say in how those allocations are made. Communities devising plans that provide more land for urban and low density rural development than needed are instructed to eliminate excess supply or face suspension of development permitting authority. Today, Oregon's land use pattern is a system of urban centers contained within urban growth boundaries (UGBs) surrounded by open space.

In the early 1990s, Washington state adopted its Growth Management Act which has many of the same features as Oregon, including UGBs and open space preservation. In the late 1990s, Tennessee adopted similar legislation.

Florida's growth management efforts, stemming from the 1970s and revamped in the mid-1980s, are relatively weak in containing urban development or preserving open space from low density rural development. Its 'concurrency' provisions, however, which limits new development to only those areas where adequate public facilities exist or are planned, appears to have shifted development away from many open spaces to infill and redevelopment sites.

Maryland's 'smart-growth' initiative, launched in the late 1990s, provides state infrastructure funds for only those projects that redirect development to infill and redevelopment sites and away from open spaces.

3. Regional Efforts

There are very few regional growth management efforts, mostly because they require state-level legislation and commitment among competing local governments to cooperate. The most notable regional growth management effort is that of the Twin Cities Metropolitan Council, coordinating planning, infrastructure investment, and development permitting among the 187 municipalities and seven counties composing the Minneapolis–St. Paul metropolitan area. The Council works to contain urban growth within 10-year urban service limits that are adjusted about every 5 years. The Council was formed in 1967 but not until the late 1990s did it have the authority to manage the region's water, sewer, and transportation operations.

4. Local Efforts

Local government growth management efforts are quite numerous. Early efforts include Milford, Ramapo, and Petaluma reviewed above. Other notable areas include Sacramento, California, Boulder, Colorado, Sarasota, Florida, and Lincoln, Nebraska, all of which employ UGBs with open space preservation. It is possible that every state has at least one example of a local municipal or county growth management program. The primary limitation of all such programs is that unless neighboring jurisdictions also engage in growth management, there is confusion on long-term development and investment expectations.

5. Primary Elements

Growth management is composed of six elements: (a) overall form; (b) growth guidance; (c) infrastructure provision; (d) protecting open spaces; (e) enhancing community life; and (f) coordination between governments.

5.1 Overall Form

A common feature of growth management programs is the containment of urban development within American-style urban growth boundaries, British-style development stoplines, or more flexible urban service limits. Collectively they can be called 'urban containment boundaries' or UCBs. UCBs are complemented by urban facility service area plans that control

the timing of utility extensions and infrastructure improvements. Land within the UCB may be subdivided into different categories for the phasing of services, depending on development suitability, proximity to existing public facilities, contiguity with existing development, and other factors. Sufficient land must be included within a UCB to accommodate market demand within the planning horizon. If too little land is included, land prices may increase beyond the ability of the market to accommodate development needs. Too much land defeats the purposes of encouraging compact, contiguous development and usually results in excessive and premature capital outlays for public services.

5.2 Growth Guidance

Growth guidance relates to the timing, phasing, and dimensioning of development within UCBs. Within UCBs may be intermediate containment boundaries (ICBs) which, as their name suggests, are short-term development boundaries within the long-term UCB. In contrast, outside UCBs, there may be urban development reserves (UDRs) where future urban development will be targeted once the UCB is filled in.

Growth management usually results in 'upzoning' land within UCBs and 'downzoning' land outside. The reason is simple. Land intended for urban development should accommodate more intense activity and land not intended for urban development should be prevented from urban development.

The 'upzoning/downzoning' requirement permeates through other land use policies. For example, conventional planning uses transitional zoning to buffer land-use activities from each other and to facilitate a concentric-ring urban form wherein uses farther away are at lower densities than closer in. Transitional zoning is not generally compatible with growth management, however. The result of transitional zoning is that low- and very-low-density land uses expand into rural areas and into resource lands, thereby perpetuating urban sprawl and undermining the long-term economic viability of resource production. Instead, growth management programs use nontransitional zoning strategies which, among other things (a) establish moderate to high-density and -intensity land-use categories throughout much of the urban area, (b) facilitate nodal and corridor development especially along transit routes and major multimodal intersections that are themselves surrounded by moderate to high-density-zoning, (c) reduce the scale of low and very low urban densities within urban areas, (d) eliminate low- and very-low-density development in areas that are predominantly resource lands or other environmentally sensitive lands, and (e) restrict the size and location of low- and very-low-density development outside urban containment boundaries and away from resource or environmentally sensitive lands.

Growth management also challenges conventional 'Euclidian' zoning—so called after the US Supreme Court's decision upholding zoning of the sort challenged in *City of Euclid, Ohio, vs. Ambler Realty Company*. Euclidian zoning is a pyramid of allowable land uses. At the top is residential land use; land zoned for residential is restricted to residential uses. Commercial zoning allows commercial and residential uses. Industrial zoning allows industrial, commercial, and residential land uses. Agricultural zoning allows all land uses. The problem is that many land uses allowed in some zones are incompatible with the primary purpose of that zone, such as a residential development in an industrial zone. The trend is now towards exclusive use zoning of open spaces and inclusionary zoning allowing the mixing of land uses in urban areas. Growth management thus reshapes Euclidian zoning.

Growth management also aims to eliminate exclusivity. Many communities use zoning strategies to exclude certain classes of people, especially those occupying low-income housing. A typical strategy is to zone the community for large lot sizes and large minimum floor areas. Low-income households will, thus, be unable to afford to live in the community. Growth management recognizes the need to provide housing to all income levels. Many growth management communities have 'inclusionary zoning' policies that require large-scale projects have housing opportunities affordable to a wide range of households.

Finally, it is one thing to designate land for moderate to high density and give developers the impression that they have flexibility in meeting market demands but it is quite another to actually allow such densities. Unwary developers who propose high-density developments can get stalled in decision-making, often because existing residents oppose such developments. Housing density usually is reduced but this defeats the purpose of growth management. The solution is 'minimum density zoning.' For example, if the maximum residential density allowed is 8.0 units per acre of land, the minimum may be 5.0 and the community average target for this category of zoning may be 6.5 units.

5.3 Infrastructure Provision

Growth management cannot succeed without infrastructure planning and financing that makes urban development more attractive than development in the countryside. On the other hand, development must not exceed the ability of existing infrastructure to accommodate it. 'Adequate public facilities' standards assure that, in addition to meeting applicable zoning and subdivision standards, facilities and services are able to serve it. 'Concurrency', whereby facilities needed to serve new development are in place or committed concurrent with the impacts of new development, is another version of the same principle.

Under growth management, infrastructure is a shared responsibility between government and developers. One form of shared responsibility is 'development exactions' in which the developer provides a certain range of infrastructure at their cost. Infrastructure within the development is normally the developer's responsibility but the responsibility for financing offsite roads, water, sewer, schools, parks, and other facilities is complex.

'Special taxing or assessment districts' are one way in which onsite and offsite infrastructure costs can be absorbed via long-term, below-market interest rates financed by new development. 'Impact fees,' which are one-time assessments on new development to pay for new or expanded facilities needed to serve new development, help cover a share of the offsite costs.

5.4 *Protecting Open Spaces*

What happens outside UCBs? Growth management is not effective unless it preserves open spaces and manages rural density development. Some states, such as Hawaii and Oregon, are very aggressive in preventing any development of open spaces. In those states, 'exclusive farm/forest use' regulations require owners of such land to manage it for farming and forestry if they wish to live on it.

Economic mechanisms can be used to encourage development inside UCBs and prevent development outside. 'Purchase of development right' programs involve public acquisition of the development rights of open spaces. 'Transfer of development right' programs involve developers acquiring development rights from open space owners in rural areas for transfer to areas within UCBs at higher density.

Not all land outside UCBs is appropriate for farming, forestry, or other open space uses. Indeed, there is some demand for rural living that if not accommodated would result in urban households acquiring farms and then maintaining only the yard they wish, typically one-quarter to one-half acre. 'Clustered' development in rural areas is one solution. For example, a 100-acre farm may be subdivided into 10 parcels of which nine are one-acre and one is 91 acres, which in many parts of the country is sufficiently large to accommodate reasonable commercial farming operations.

5.5 *Enhancing Community Life*

Growth management changes existing neighborhoods. Oddly, many neighborhoods view growth management anathema to their quality of life, especially if infill and redevelopment activities locate there. 'Design review processes' help assure that new development in stable neighborhoods enhances neighborhood character. 'Supplemental housing' such as mother-in-law, over-garage, basement, or attic living units allowed in established neighborhoods improve property values, increase safety (more 'eyes-on-the-street'), make more efficient use of infrastructure, and generally assure neighborhood stability thereby increasing neighborhood vitality.

5.6 *Coordination Between Governments*

Growth management is not effective in isolation. Because a number of governmental units interact with each other at various levels, effective growth management results in all relevant governmental units sharing the same vision for a preferred built and natural environment given the need to accommodate projected development. Cross-acceptance is one way in which this is achieved. Conceptually, cross-acceptance means that all units of government affecting a particular community—including state, regional, local, and special district governments (such as school districts, water districts, fire protection districts and the like)—agree to each other's plans and they all agree to share the vision of the region's future built and natural environment.

6. *Benefits*

For a long time, the benefits of growth management in the United States seemed illusory. Recent studies indicate, however, that growth management of the kind described here reduces tax burdens, reduces air and water pollution, reduces vehicle-miles-traveled per household, increases incomes, and improves quality of life across many measurable dimensions.

7. *Challenges*

Growth management is not without challenges. Growth management policies *per se* do not solve problems associated with rapid growth or urban sprawl. Growth management requires that development be encouraged in urban areas and discouraged in rural areas. This can engender opposition from stable neighborhoods because of potential threats. It also requires patience. The positive effects of growth management take about 20 years to realise. Unless there is political will to sustain growth management efforts, benefits will remain illusory.

See also: Community Economic Development; Development and Urbanization; Ecology, Political; Human–Environment Relationships; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Planning Theory (North American); Urban Activity Patterns; Urban

Geography; Urban Growth Models; Urban History; Urban Places, Planning for the Use of: Design Guide; Urban Planning: Central City Revitalization; Urban Sprawl; Urban Studies: Overview

Bibliography

Abbott C, Howe D (eds.) 1993 *Planning the Oregon Way*. Oregon State University Press, Corvallis, OR

DeGrove J M 1992 *Planning and Growth Management*. Lincoln Institute of Land Policy, Cambridge, MA

Juergensmeyer J, Juergensmeyer C, Roberts T 1998 *Urban Planning and Land Development Control Law*, 3rd edn. West Publishing Co., St. Paul, MN

Kelly E D 1993 *Managing Community Growth*. Praeger, Westport, CN

Knaap G J, Nelson A C 1992 *The Regulated Landscape*. Lincoln Institute of Land Policy, Cambridge, MA

Nelson A C, Duncan J B 1995 *Growth Management Principles and Practices*. American Planning Association, Chicago, IL

Nelson A C 1999 Comparing states with and without growth management: Analysis based on indicators with policy implications. *Land Use Policy* **16**: 121–127

Nelson A C 2001 *Estimating Land Use and Facility Needs and Impacts*. American Planning Association, Chicago, IL

Porter D R 1997 *Managing Growth in America's Communities*. Island Press, Washington, DC

Porter D R 1996 *Profiles in Growth Management*. Urban Land Institute, Washington, DC

A. C. Nelson

Urban Planning: Methods and Technologies

The collection, interpretation, and dissemination of the information needed to improve public and private decision making is one of the major justifications for urban planning. As a result, planning methods and technologies have played a central role in planning practice and education since the profession's inception. This article reviews the evolution of planning methods and techniques, considers the methods and technologies that are most widely used in contemporary planning practice, and speculates on the future of planning methods and technologies.

1. Evolution of Planning Methods and Technologies

Three relatively distinct eras can be traced in the evolution of planning practice, methods and techniques. The first era, 'planning as design,' characterized planning from its inception until the decade of the 1960s. The second, 'planning as applied science,' characterized the profession in the 1960s and 1970s. The third, 'planning as communication and reasoning together,' characterized the profession during the 1980s and 1990s. The following sections examine each of these eras and their corresponding methods and techniques in turn.

1.1 The Early Profession

The planning profession emerged in the USA at the beginning of the twentieth century in response to the unregulated urban growth and inadequate public facilities of the new industrial cities. Guided by a rather naive form of environmental determinism, the early planners assumed a professional responsibility for improving society through changes in the physical environment and protecting an overarching public interest from the presumably self-interested and uninformed actions of elected officials and private citizens. Planners' methods were largely intuitive methods of manual design adapted from the profession's institutional homes in architecture and landscape architecture, taught primarily by planning practitioners in the ubiquitous studio workshop.

By the 1950s these largely intuitive methods had evolved into a standard land use planning approach codified in and F. Stuart Chapin's *Urban Land Use Planning* (1957) and T. J. Kent's *The Urban General Plan* (1964). This model defined planning as the preparation of a formal document laying out a long-term, comprehensive, but general, vision for a community's future physical development, including public and private uses of land and related public facilities. These plans included a summary of existing and emerging conditions and needs, a statement of planning goals, a 20- or 30-year development plan expressed in map form, and policies for implementing the plan. Variations on this general approach are reflected in contemporary land use planning practice (Kaiser and Godschalk 1995).

1.2 The 1960s and 1970s

The planning profession's attention turned in the 1960s from the design of the physical city to new concerns with improving transportation access, reducing urban blight, and improving social and economic conditions on both the urban and regional scales. The dominant perception of the field changed from an image of 'planning as design' to 'planning as an applied science' as new students, faculties, and departments joined the profession from outside the traditional schools of design. An increased emphasis on quantitative methods and social science theory led to the abandonment of the once ubiquitous design studio for new courses in research design, statistics, and quantitative techniques represented by Isard et al.'s

monumental *Methods of Regional Analysis* (1960) and Krueckeberg and Silvers' widely used *Urban Planning Analysis* (1974). Planners' professional role was now seen as delivering unbiased, 'objective,' and politically neutral advice to elected officials and the public and collecting and disseminating more and better information that could inform and improve the policy-making process.

Computers were first used in planning in the 1960s, as part of a general faith in science and technology and the emergence of new academic fields such as operations research, urban economics, and regional science. The most notable early attempts to use computers in planning were the extremely ambitious and highly visible federally-funded efforts to build large-scale metropolitan land use–transportation models and integrated municipal information systems. These efforts were seen as providing the foundations for a new 'science' of planning by helping to understand and guide the urban development process and improve the information base for public and private decision making (Harris 1965, Webber 1965).

Planners' optimistic faith in computer technology was severely tested in the 1970s. The pioneering efforts to develop large-scale urban models and municipal information systems failed spectacularly, largely as a result of their overly ambitious goals, the limitations of available computer equipment, the lack of required data, and an inadequate understanding of the urban development process. Sophisticated analytic tools such as mathematical programming, which were once assumed to provide the foundations for 'scientific planning,' were found to be largely inappropriate for most planning applications which are hard to define or solve, have outcomes that are difficult to predict, and involve considerable judgement (Lee 1973, Brewer 1973, Laudon 1974).

The practical failure of the early computer-based urban models reflected a more fundamental rejection of the 'rational planning' model that was assumed to underlie all attempts to use computers in planning. Planning methods and technologies that were once assumed to be 'objective' and politically neutral were found inevitably to increase the power of administrators and technical experts at the expense of those who lacked the expertise to use them effectively. Information that once appeared to be more accurate, credible, and objective, simply because it has been generated by a computer, was found to reflect fundamental choices in the selection of data, the application of computational procedures, and the distribution of results that helped answer the fundamental political questions of 'who gets what, when, and how' (Wachs 1982). By the decade's end, most planners had no access to—or perceived need for—the expensive, fragile, and hard-to-use mainframe computers of the era and continued to rely on manual computational techniques that had changed little since the 1950s.

1.3 The 1980s and 1990s

Planners' faith in computing technology was reborn in the 1980s with the emergence of microcomputers that made computer-based methods and techniques readily available to planners the world over. By the end of the century, nearly all planning agencies used computers for general-purpose office functions such as processing documents, monitoring budgets, and maintaining records. Many of the larger planning agencies routinely used geographic information systems (GIS) (see Sect. 2.2) for planning management tasks such as processing permits, collecting and maintaining land-related information, and preparing thematic maps. The maturation of municipal GIS and the dramatic growth of the internet allowed planners to access vast quantities of spatially related information that could only have been dreamed of by earlier planners.

Underlying these developments were new images of 'planning as communication and reasoning together.' Planners increasingly realized that planning involves not only the collection and distribution of information but, more importantly, face-to-face communication with colleagues and clients involving other nontraditional types of 'information': stories, advice, and personal experiences. As a result, planners increasingly realized that the *ways* in which they transmit information is often more important than *what* they say.

For example, planning forecasts could be described either as the result of sophisticated computer-based modeling efforts too complex for laymen to understand, or as implementing a few easy-to-understand assumptions. The same analysis techniques and results could be employed in both cases. However, in the first case planners' communications are likely to mystify their audience and exclude nonexperts from the planning process. In the second, they will invite the public to participate by critically examining the assumptions that underlie the forecasts.

The 1990s also witnessed the emergence of new professional models which suggested that planners should develop and guide interactive and ongoing processes of community learning and debate that allow, public officials, stakeholders, and planners to jointly define issues, collect information, and share their ideas, perspectives, and interests. These strategies not only help promote public participation and understanding, increase public support for the proposals that are developed, but often help resolve issues before the formal decision-making processes begin (Innes 1998).

2. Current Planning Methods and Technologies

Before the advent of computers, planning data were available only in voluminous printed volumes and paper maps that were incomplete, inconsistent, and

could only be updated manually. Planning analysis and communication required manually copying the relevant information, laborious and error-prone hand computations, typing text and tables, and drafting maps and graphs.

Today all of this has changed. An almost overwhelming array of data are readily available in digital form from national, regional, and local databases. Planners' computer-based tools allow them to analyze these facts quickly and use them to produce attractive documents, charts, and maps. As a result, it is almost impossible to conceive of planning *without* computers and related advanced information technologies. The following section briefly reviews the most methods and technologies that are most widely used by contemporary planners.

2.1 Spreadsheets

An electronic spreadsheet (or simply a 'spreadsheet') is an electronic version of the accountant's traditional paper spreadsheet, which stores numeric data in two-dimensional tables that display the results of calculations performed on these data. They have been called, with some justification, 'God's gift to planners.' They are easy to learn, forgiving of errors, and provide an intuitively logical structure for examining any quantitative problem that can be formulated as a two-dimensional table. They are easily adaptable, allowing users to copy data and computational formulas easily from one location to another and develop predefined 'spreadsheet models' that for many years provided the most widely used software tools designed specifically for planning applications (see, e.g., Klosterman et al. 1993). And, most importantly, spreadsheets are ideal for examining the 'what if' questions that permeate planning by allowing planners quickly and easily to determine the effects of alternative policy choices and different assumptions about the future.

2.2 Geographic Information Systems

GIS are computer-based systems for capturing, storing, analyzing, and displaying spatially related data. GIS combine: (a) sophisticated mapping capabilities which allow attractive and informative maps to be prepared quickly and easily; (b) highly developed data base management tools for storing, modifying, and manipulating descriptive data describing the features displayed on a map; (d) topological data structures which explicitly describe the spatial relationships between map features; and (e) extensive data describing anything that can be displayed on a map.

GIS has proven extremely useful for a full range of planning applications such as:

(a) displaying spatially related information;

(b) maintaining land-based inventories and organizing spatial data previously maintained by different organizations;

(c) identifying the best (or worst site) for locating facilities and activities;

(d) conducting spatial analysis tasks such as defining buffers around selected map features; and

(e) dealing with network-related transportation and infrastructure issues.

2.3 Planning Support Systems (PSS)

The popularity of GIS is due largely to its ability to serve as a 'chameleon technology' which provides basic functions that are extremely useful for a wide range of public- and private-sector applications. However, this general applicability comes at the expense of the particular needs of specialized fields such as planning. In response to these limitations, at the end of the twentieth century, planners began to develop PSS that adapted the concept of decision support systems to planning. PSS combined (a) an array of spatial and nonspatial data with (b) a broad range of models and methods for determining the implications of alternative policy choices; and (c) visualization tools for displaying model results in easily understandable formats, ideally over the Internet. More importantly, PSS provided interactive and participatory processes for group interaction and decision making for dealing with non-routine and poorly structured decisions (Brail and Klosterman 2001).

2.4 Visualization and Multimedia Systems

Planners use not only printed reports, charts, and maps to describe and understand neighborhoods, cities, and regions, and relate their descriptions, analyses, and plans to others. They also use computer-based visualization and multimedia systems to combine these traditional types of information with aerial and eye-level photographs and video and audio clips to provide a more comprehensive understanding of an area and alternative plans for its future.

For example, all types of information—statistical, textual, mapped, audio, video, and computer-generated—on a vacant industrial site and alternative proposals for developing it could be compiled in digital form and made available for face-to-face group meetings or 'virtual' meetings held over the web. These different types of information can help provide the foundation for collaborative processes of collective design by allowing meeting participants to explore current and past conditions, review issues and concerns raised by community residents, and jointly explore alternative development possibilities and their impacts. Computer-aided visualization can also help

promote community participation by rendering information more understandable, credible, and usable to different segments of the public, particularly to those who have not had experience with more traditional forms of information (Shiffer 1992).

2.5 The Internet and the Worldwide Web

The internet and the worldwide web (or 'web') provide a 'network of networks' that connects millions of computer users around the world into a seamless web of electronic communication and information dissemination. The internet has been extremely useful by providing planners and the public with 'anytime, anyplace' access to large databases such as the entire contents of the national census, local property information, and interactive audio and video displays. It also allows plans and proposals—along with their supporting documents, information, and analytical models—to be distributed at very little cost, allowing people with different skills, viewpoints, and objectives to contribute to the collective decision-making process, independent of time and/or place.

3. Future of Planning Methods and Technologies

Planning methods and technologies have changed dramatically, due largely to the tremendous advances in information technology that occurred during the 1980s and 1990s. Future technological advances promise to be even more dramatic, making it extremely difficult to forecast what the future will hold. However, some trends are relatively clear. Computers and related advanced information technologies will continue to get faster, more powerful, cheaper, and easier to use. Nearly universal, high-speed access to the internet will allow planners to retrieve, analyze, and distribute vast quantities of spatially referenced information quickly and easily. Planners' software tools for processing and analyzing all types of information—numerical, textual, graphic, audio, and video—and for conveying it to others in readily understandable ways, will become increasingly powerful, available, and easy to use.

The effects that these advances will have on planning practice are much less clear. The increased availability of spatially referenced information, improved data dissemination and visualization technologies, and the ability to conduct meetings unconstrained by location or time, will facilitate increased public access to government information. Continually improving technology and the increased sophistication of planners and the public they serve will allow computer-based planning, analysis, and evaluation models to take their place at the center of planning practice. Together they will provide the foundations for new modes of planning which more directly involve the public in the fundamental public policy choices that help shape their communities. Only time and professional practice will tell whether this potential is achieved.

See also: Computers and Society; Environmental Determinism; Geographic Information Systems; Geographic Information Systems: Critical Approaches; Internet: Psychological Perspectives; Planning Issues and Sustainable Development; Planning Theory: Interaction with Institutional Contexts; Urban Growth Models; Urban Sprawl; Urban System in Geography

Bibliography

Brail R K, Klosterman R E 2001 *Planning Support Systems: Integrating Geographic Information Systems, Models and Visualization Tools*. ESRI Press, Redlands, CA

Brewer G D 1973 *Politicians, Bureaucrats, and the Consultant: A Critique of Urban Problem Solving*. Basic Books, New York

Chapin F S Jr 1957 *Urban Land Use Planning*. Harper and Brothers, New York

Harris B 1965 Urban development models: New tools for planning. *Journal of the American Institute of Planners* **31**: 90–183

Innes J E 1998 Information in communicative planning. *Journal of the American Planning Association* **64**: 52–63

Isard W et al. 1960 *Methods of Regional Analysis: An Introduction to Regional Science*. MIT Press, Cambridge, MA

Kaiser E J, Godschalk D R 1995 Twentieth century land use planning: A stalwart family tree. *Journal of the American Planning Association* **61**: 365–85

Kent T J 1964 *The Urban General Plan*. Chandler, San Francisco

Klosterman R E, Brail R K, Bossard E G (eds.) 1993 *Spreadsheet Models for Urban and Regional Analysis*. Center for Urban Policy Research, New Brunswick, NJ

Krueckeberg D A, Silvers A L 1974 *Urban Planning Analysis: Methods and Models*. John Wiley, New York

Laudon K C 1974 *Computers and Bureaucratic Reform: The Political Function of Urban Information Systems*. John Wiley, New York

Lee D B Jr 1973 Requiem for large-scale models. *Journal of the American Institute of Planners* **39**: 163–87

Shiffer M J 1992 Towards a collaborative planning system. *Environment and Planning B: Planning and Design* **19**: 709–22

Wachs M 1982 Ethical dilemmas in forecasting for public policy. *Public Administration Review* **42**: 562–67

Webber M M 1965 The roles of intelligent systems in urban-systems planning. *Journal of the American Institute of Planners* **31**: 289–96

R. E. Klosterman

Urban Planning (Western), History of

The evolution of urban planning is inextricably linked with the processes of modernization and the nature of modernity. The emergence of planning was the most

literal way in which the role of government expanded in response to the problems of accelerated urbanization from the end of the nineteenth century. The agenda and scope of urban planning have expanded and changed dramatically. At the start of the twenty-first century, urban planning is a multifaceted set of policy and administrative processes which resists easy categorization. Its forms have ebbed and flowed in concert with myriad social crises, economic changes, environmental and demographic pressures, political opportunism, and technological innovation, notably new forms of transportation and communication.

While an early focus on reforming the physical framework for urban life has braided in various social, economic, and legalistic directions, some common threads have characterized the mainstream history of planning theory and practice. These include a pervasive ideology of holism and the benefit of the 'birds eye view' in directing urban development; a practical commitment to reconciling the diverse interests of stakeholder groups; an emphasis on order, scientific analysis, and rationality in public policy decision making; belief in the power of environmental improvement to invoke social change; and an ongoing faith in the potency of enlightened state intervention to secure the 'public interest' in the face of market failure.

1. *Antecedents and Beginnings*

For such a quintessential twentieth-century activity, urban planning has a remarkably long history. But this premodern planning invariably meant a rudimentary form of spatial layout or architectural set-pieces disengaged from the broader policy vectors of the state. Indeed, planning was all too frequently coopted as a mere technical device for the imposition of authoritarian will. Planning as civic design stretches back to antiquity—to ancient Egypt, Greece, and Rome—with the organization of urban spaces manipulated to reflect the dominant political and social order. This thread continues through the grand public squares and piazzas of the renaissance city and into the planning templates for imperial expansion, notably the Vitruvian-inspired Spanish Laws of the Indies of the sixteenth century.

Planned towns were centers for government, displays of autocratic power, military footholds, and central places for land settlement schemes. Rectilinear, gridiron or chessboard street plans were the most time-honored morphology, but by the start of the nineteenth century these were studded with open spaces and enveloped by distinctive parkland belts. Back to antiquity, ideal cities assumed a variety of alternative geometries. Radial plans would decisively influence the first generation of early modern planners. Dating from the beginning of the eighteenth century, Karlsruhe in Germany became an enduring icon with its rigidly-ordered street pattern fanning out from a central palace.

A more modern urban planning emerged in the late nineteenth century to address challenges unleashed by industrial urbanization. Existing narrowly focused administrative procedures and primitive building regulations were ill-equipped to address unprecedented rates of urban population growth, new forms of economic enterprise, and accompanying housing demands and environmental pollution, a new geography of accessibility producing a chaotic competition for urban space, and a transition toward new social structures.

An amazing array of modes of thought and practical innovations was tapped to help shape the new discourse of planning. At the theoretical end of the spectrum, for example, a nexus was forged between a longer tradition of social and economic utopian thought and the notion of better cities through classic works like James Silk Buckingham's *National Evils and Practical Remedies* (1849) and Edward Bellamy's *Looking Backward* (1887). Other thinkers, like the influential anarchist Peter Kropotkin, looked toward a more decentralized society aligning village life with new modes of production. The most substantial experiments in industrial decentralization were more socially conservative. These were the planned industrial villages expressing an emergent form of welfare capitalism geared to educating and disciplining a productive labor force. What Walter Creese in *The Search for Environment*(Yale University Press, 1966) called the 'Bradford–Halifax' School of model village builders in the English Midlands developed into full-blown community building in places like Bournville, Port Sunlight and George Pullman's eponymous suburb near Chicago. More romantic were planned low-density suburbs for the elite (also fleeing the city) such as Riverside (1869) near Chicago.

More utilitarian issues consumed reformers within the cities proper, as the management of established infrastructure (transport, water supply, sewerage) showed the strain. Poor housing conditions in the form of high density slum and tenement dwellings became a major concern. Exposes by reformers like Jacob Riis in New York City in *How the Other Half Lives*(Scribner's Sons, 1890), shocked middle-class sensibilities. Philanthropic model lodging houses and stricter regulations over space standards for rooms, dwellings, allotments, and, in turn, streets ensued. Provision of parks and open spaces was another preoccupation. Utilizing the organic metaphor so prevalent in early planning propaganda, parks were the 'lungs of the city' and needed to be secured in advance of development. Citizens improvement associations, the precursor to civic art and 'tidy towns' movements, provided a 'bottom up' chorus of concern. A major theme through all of this was the need to secure higher standards of public health and hygiene,

a crusade which extended to cognate missions of social, moral and cultural cleansing.

The crucial intellectual breakthrough made simultaneously around the Western world in the 1880s and 1890s was a recognition of the limitations of piecemeal improvement, the interconnectedness of urban life, and the necessity for more coordinated intervention to avoid the undermining of the social order through outbreaks of disease, community dissent, and loss of economic competitiveness. The coining of the terms 'town' and 'city planning' reflected the emergence of the new approach.

Synthesizing ideas then in the air, the basic precepts of planning were expressed through a rich palette of modernist thought. An American contribution was the park system, the concept of interlinked networks of open spaces mixing formal and informal elements and parkways developed by Frederick Law Olmsted Sr., which has given cities like Boston, Denver, and Kansas City a green legacy enjoyed by many generations. The enormously influential garden city idea was similarly associated with an individual, British shorthand reporter Ebenezer Howard, and advanced in his book *To-morrow: A Peaceful Path to Real Reform* (1898), reprinted as *Garden Cities of Tomorrow* (1902). At its most visionary, the manifesto called for clusters of new, self-contained 'slumless, smokeless cities' of limited size on community-owned land as an alternative to the sprawling urban-industrial city. Alternative new urban structures were mooted in the same era. In 1892, Spanish engineer Arturo Soria y Maya invented the idea of the linear city (*La Cuidad Lineal*) which resurfaced in twentieth-century forms such as Edgar Chambless' Roadtown (1910), the ideas of the Soviet deurbanists in the 1930s, and corridor city planning in the 1960–70s.

A string of seminal contributions signposting ways ahead emerged in Europe in the late nineteenth century. Town extension planning and land-use zoning were already accepted municipal practices in Germany before the turn of the century. Important texts were Reinhard Baumeister's *Stadterweiterungen in technischer, baupolizeilicher und wirtschaftlicher Beziehung* (Town Expansions Considered with Respect to Technology Building Code, and Economy, 1876) and Joseph Stubben's *Der Stadtebau* (1890). Baron Haussmann's extraordinary transformation of Paris under the patronage of Napoleon III from a largely medieval city into a modern masterpiece of boulevards, star-points, and strategically located public buildings had begun in the 1870s. A competing aesthetic was that of Viennese architect Camillo Sitte whose *Der Städtebau* (1889) helped codify more artistic, intimate, and historically sensitive techniques for the modern 'art of designing cities and suburbs.'

International exhibitions provided unique opportunities to showcase new urban planning ideas as planned cities in microcosm. The most sensational synthesis was the 1893 World's Columbian Exposition in Chicago. This collaborative integration of grand architecture, landscape design, public art and early amusement park was overseen by local architect Daniel Burnham, who was to emerge as the leading American architect-planner of the early 1900s.

2. *The First Half of the Twentieth Century*

At the turn of the last century, modern planning goals crystallized around recurrent themes of convenience, efficiency, healthfulness, beauty, and order. The actual means to these ends were quite diverse, from totally planned landscapes through selective interventions in the urban fabric to new regulatory frameworks for private development. Certain kinds of planning assumed a cultural distinctiveness. House-and-garden were at the center of British concerns, France was known for its monumentalism, planning advocates looked to Germany for its innovative land management, and to the United States for the business-minded aestheticism of city beautiful thought. The planning movement had become genuinely international. Even a relatively closed Asian country like Japan, in acknowledging the importance of modernization in economic and political advancement, had discovered Western planning ideas through study tours and private visits. The colonial ties of European powers saw many ideas imposed abroad as much as imported.

Major conferences were important agents in the diffusion of knowledge. An arm of Howard's Garden City Association (1899, and renamed the Town and Country Planning Association in the early 1940s) was the International Garden Cities Association (1913) which would evolve into the International Federation for Housing and Town Planning. Information sharing would be damaged by World War I, and never again capture the early excitement, but planning was well established as a global movement with astonishing commonalities of purpose and method in diverse settings. With that came moves to institutionalize the activity as a genuine profession occupying the interstices between the traditional built environment professions of architecture, surveying, and engineering. Urban planning was also developing its own methodology, with the accent on scientific civic surveys, an extension of settlement worker investigations. Scottish polymath Patrick Geddes supplied the requisite mantra: 'survey-analysis-plan.'

In the English-speaking world, several interconnected events in Britain at the end of the first decade of the twentieth century would have reverberations around the world, and especially in the British Empire. In 1909, the first formal town planning course started with establishment of the William Lever-endowed Department of Civic Design at the University of Liverpool, and Raymond Unwin's classic *Town Plan-*

*ning in Practice*appeared. In 1910, the Royal Institute of British Architects organized an international conference which attracted the leading names in world planning, and the first planning legislation was passed in the form of the *Town Planning, Housing etc Act*. The British Parliament was told of the government's intention to realize 'the home healthy, the house beautiful, the town pleasant, the city dignified, and the suburb salubrious.'

The British law was initially restricted to zoning for new development but was later applied comprehensively. Land-use zoning also eventually established itself as the standard form of microregulation in the United States and elsewhere. A popular early mode of zoning was the residential district, a concept which was grafted readily onto existing local government regulations and widely invoked as a tool of social exclusion. A 1916 New York City zoning ordinance introduced novel controls on building height and shape that led to the Manhattan 'wedding cake' skyscraper form. It was not until the historic 1926 US Supreme Court case *Euclid, Ohio vs. Ambler Realty Co*. that the legal validity of planning was finally confirmed in America. British and continental models interconnected planning and zoning, but in the United States city planning commissions and zoning boards could pursue independent agendas.

The early twentieth century is distinguished by some landmark plans which codified modernist planning ideas. The 1901 plan for Washington DC by the Senate Park Commission headed by Daniel Burnham reconfigured Pierre L'Enfant's eighteenth-century plan into a monumental triumph for city beautiful ideas. Many plans for American cities would follow, but the most impressive was the privately sponsored *Plan of Chicago* (1909) by Burnham and Edward Bennett. Artistic in presentation and regional in scope, it set a new standard for comprehensive planning. Edward Lutyens' New Delhi (1912) transplanted grand manner planning to an imperial context. The 1911–12 international competition for the new federal capital of Australia produced a unique survey of early modern planning thought. The winning scheme by Walter Burley and Marion Mahony Griffin, both former associates of Frank Lloyd Wright, was a brilliant landscape synthesis of garden city, city beautiful, and deeper democratic ideals. The most idealistic plan of its day was the footloose *World City for Peace* (1912) by American sculptor Hendrik Anderson and French architect Ernest Hebrard. The multivolume Regional Plan of New York masterminded by Thomas Adams in the late 1920s was the archetypal pragmatic, professional, growth-oriented plan of the interwar period. A celebrated critique by urbanist Lewis Mumford branded the scheme as dinosauric in entrenching rather than curbing trends toward megalopolitan development.

Mumford was deeply influenced by the garden city movement and advocated a more organic, environmentally sensitive variant of regional planning. While stopping short of Howard's dream of a 'better, brighter civilization,' garden city advocates campaigned tirelessly for new towns. Their successes were modest in the overall scheme of things but resulted in impressive showpieces of planning in practice. Two British garden cities were built: Letchworth (1904) designed by Raymond Unwin and Barry Parker, and Welwyn Garden City (1919) by Louis de Soissons. Welwyn was conceived as an overspill satellite community for London, and this concept was readily exported. An early German response was the fringe *grossiedlung* with apartment blocks of social housing. In the United States, echoes of the garden city notion can be read into 1920s speculative sunbelt 'new towns' such as those designed by John Nolen in Florida; the New Deal towns of the late 1930s such as Greenbelt, Maryland sponsored by Rexford Guy Tugwell's Resettlement Administration; and Clarence Stein and Henry Wright's Radburn, New Jersey, with its cul-de-sacs and 'superblocks' representing an experimental 'garden city for the motor age' stifled by the Depression.

The 'garden city' label and concept was translated to many countries: *cite-jardin*, *gartenstadt*, *cuidad jardin*. The most ubiquitous application was to suburban rather than new town settings, becoming more popular and prolific through the prosperous 1920s with privately master-planned elite communities leading the way. These ventures refined further an enduring amalgam of site planning ideas revolving around house-and-garden living, hierarchical street systems, provision of open space and public facilities, and attention to landscaping and streetscape qualities. The low density suburban ideal reached its zenith in Frank Lloyd Wright's *Broadacre City*dream of the early 1930s.

Other expressions of modernism were more vertiginous. New York City's Rockefeller Centre in the 1930s foreshadowed the penchant for large-scale urban redevelopment. The most startling theoretical schemes were those of European theorists like Auguste Perret, Ludwig Hilbersheimer, and most famously the messianic Swiss architect Le Corbusier. Through the 1920s, Le Corbusier worked his way through a succession of urban utopias—'La Ville Contemporaine,' 'City of Three Million Inhabitants,' and the 'La Ville Radieuse'—to perfect a design vocabulary of towers-in-parkland, multilevel transportation systems, and functional land-use segregation. His *Plan Voisin* applied these ideas to a portion of central Paris, sweeping away narrow streets and 'diseased quarters' for a forest of cruciform high-rise tower blocks. The legacy of these ideas can be seen in large-scale state housing programs and planned capital cities such as Chandigarh and Brasilia. The guiding spirit here was architectural modernism, and emerging out of the deliberations of CIAM (*Congress International d'Architecture*) was the 'Athens Charter,' the definitive town

planning manifesto of the 1930s for 'the functional city' distinguished by specialized transportation and freeway systems, high rise buildings, extensive open spaces, and planned dormitory communities.

3. *Post-war Master Planning*

World War II delayed implementation of this blueprint but unleashed a universally strong planning impulse under the banner of 'postwar reconstruction.' In many European cities, the necessary reconstruction was as much physical as ideological. But in most Western nations cross-political agreement on the relevance of at least some state involvement in urban development processes moved planning decisively beyond the rhetoric, experimentation and broad theoretical acceptance attained by the 1930s into the realm of implementation. Through the 1950s the scale of planning activity grew significantly as planning settled comfortably into the policy machinery of the welfare state. The bureaucratization of planning was underpinned by a steady professionalization and growth in tertiary education to train the larger numbers of experts required.

The dominant discourse was of urban planning as an unambiguously beneficent activity of government: comprehensive, technocratic, scientific, and socially progressive. More was better if cities were to remain healthy, efficient, prosperous, and beautiful. The major focus was on land-use planning. The elements of modernist planning had surfaced before the war and were now synthesized into a total package embracing master plans; comprehensive redevelopment of the innercity; mono-functional land-use zoning; high rise social housing towers, low density middle-class suburbs; planning and replanning of residential areas as neighborhood units; freeways, ring roads and pedestrianized town centres; green webs of open space, industrial decentralization, and new towns. From mid-century, these planning ideas, variously interpreted and combined, but broadly accepted by architects, developers and government, made decisive impacts on the urban landscape.

All-encompassing metropolitan blueprints assumed a central status. Three early master plans from the mid-1940s which set moulds for many which followed were those for London, Stockholm, and Copenhagen. Patrick Abercrombie's definitive plans for the city and county of London 1943–44 were built around a cellular community structure, an encircling green belt to contain suburban sprawl, and a ring of new towns. Sven Markelius' general plan for Stockholm 1945–46 skillfully integrated transportation, land use and environmental planning, with new subcenters developed around railway stations. And the first Copenhagen finger plan of 1947 also anticipated the transportation corridor planning which would gain popularity from the 1960s when the *cordon sanitaire* of the green belt concept was doomed by escalating suburbanization.

Postwar city strategies targeted both 'obsolete' inner areas and planning opportunities on 'greenfields' sites. Urban renewal schemes involving high-rise housing, upgraded transportation infrastructure, private and public sector offices, and other new uses sought to modernize the late industrial city, frequently doing so at the expense of existing community structure and historic building fabric. The garden city dream was kept alive in new towns. The most heroic program was in the United Kingdom, with the *New Towns Act* 1946 facilitating construction of over 30 new towns by public corporations between 1947 (Stevenage) and 1970 (Milton Keynes). This model of growth management also found favor in the Third World, while private sector new towns (such as Reston, Virginia and Columbia, Maryland) were more common in the USA. Everywhere, however, and even if it did not look like it, considerable planning energy was devoted to the processes of suburban development.

4. *Postconsensus Urban Planning*

The highpoint of technocratic master planning allied to 'scientific' methodologies and modernist architectural solutions was the late 1960s. But the new world order which emerged in the 1970s constituted a markedly different context. The notion of the centralised, benevolent state crumbled. Economic policy was forced to look beyond the industrial fordism which had been the touchstone of social progress in the first half of the twentieth century. Dissent about unpopular wars and civil rights spilled onto the streets. More frequent calls were heard for greater community participation and self-help. The environmental movement awakened. In the face of these challenges, even if planning seemed to be realizing prewar dreams in physical terms, it was not always delivering attractive, equitable, and environmentally sound outcomes.

As the world changed, urban planning came under attack on many fronts: for standardized formulaic solutions, freeway proposals slicing through established neighborhoods, destruction of historic buildings, minimal attention to social issues, facilitation of suburban spread, and so on. A famous critique of mainstream planning ideology was Jane Jacobs' *The Death and Life of Great American Cities* (Random House, 1961). Robert Goodman portrayed planners as the 'soft cops' in *After The Planners* (Simon and Schuster, 1972). If the demolition of the award-winning planned housing project Pruitt-Igoe (1972) in St Louis was symbolic of the end of architectural modernism, it also represented the loss of faith in the 'brave new world' international-style planned landscapes of the postwar era. Many unpopular grand redevelopment plans for Western cities were scuttled by citizen action during these years.

The political consensus which had been the platform for the rise of urban planning after the World War II had broken down. This did not so much destroy planning as fragment it into many specialties: advocacy planning, participatory planning, cultural planning, economic development planning, transportation planning, historic preservation, urban design, etc. A growing complexity and uncertainty about planning's mission mirrored wider social tensions and crises. In recent decades, the contours of urban planning have shifted still more, as cities have become more competitive, socially fragmented, and unmanageable. It is surely not coincidental that the scholarly pursuit of urban planning history burgeoned in the last quarter of the twentieth century, with an International Planning History Society formed in England in 1974.

5. *The Achievements and Legacy of Urban Planning*

Urban planning ranks as a significant social movement which came of age in the twentieth century. Its modern history is a complex and ongoing story of achievement, failure, resilience, and redemption. Over time, chameleon-like, its role has been constantly adapted to place, culture, and circumstance. It is a history of both continuities (the quests for health, justice, efficiency, environment, amenity) and discontinuities (in response to global technological, economic, and political shifts).

Planning thought has experienced its greatest periods of innovation in response to periodic social crises—explosive population growth, war, global economic shifts. Planning's aspirations have always soared above the practicality of their comprehensive realization, being buffeted and reshaped continually by broader social constraints. At times, planning has been a deeply flawed practice, with outcomes often the reverse of those intended (destruction rather than enhancement of community, promotion of segregation rather than social cohesion, dreary townscapes rather than inspired urban design). This is planning's *noir* side. Always more political than mere technical, it has been exploited as a tool of social control to pursue policies of exclusion, segregation, and marginalization.

At the same time, there is a raft of value-adding achievements in such areas as open space provision, affordable housing, mass transit, public infrastructure, the development of liveable, new communities, the injection of sanity into myriad land use determinations, and the protection of natural and cultural resources. There is general community acceptance that constraints on individual freedoms and property rights may be warranted for collective benefits. On balance, urban planning initiatives have meant that a significant proportion of humanity on every continent found itself living in better circumstances at the end of the twentieth century than at the beginning. World Town Planning Day, an initiative launched in the late 1940s, is now celebrated worldwide every November.

See also: Cities: Capital, Global, and World; Community, Social Contexts of; Crowding and Other Environmental Stressors; Modernization and Modernity in History; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Planning, Politics of; Urban Geography; Urban Growth Models; Urban Places, Planning for the Use of: Design Guide; Urban Planning: Central City Revitalization; Urban Sociology; Urban Sprawl; Urban Studies: Overview

Bibliography

Bosma K, Hellinga H (eds.) 1997 *Mastering the City 1: North-European City Planning 1900–2000*. NAI Publishers/EFL Publications, Rotterdam/The Hague, The Netherlands

Freestone R (ed.) 2000 *Urban Planning in a Changing World: The Twentieth Century Experience*. Routledge, London

Hall P 1996 *Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century*. Blackwell, Cambridge

LeGates R, Stout F (eds.) 1998 *Early Urban Planning: Selected Essays*. Routledge, London, 9 Vols.

Sies M C, Silver C (eds.) 1994 *Planning the Twentieth-Century American City*. Johns Hopkins University Press, Baltimore

Sutcliffe A 1981 *History of Urban and Regional Planning: An annotated bibliography*. Mansell, London

Whittick A (ed.) 1974 *Encyclopedia of Urban Planning*. McGraw Hill, New York

R. Freestone

Urban Policy: Europe

1. *Introduction*

The first question to be addressed here is: is there such a thing as a European urban policy? There is clearly a wide variety of urban policies in the different countries of Europe, so many that it is beyond the scope of this entry to review them all in any insightful way. It is also true that there are a number of European urban policies, in the sense of policies of the European Union that address urban problems. During the 1990s there have been significant moves towards the establishment of a European Urban Policy although it is not yet possible to say that this has been achieved in a comprehensive manner. It is possible, however, to identify the emergence of certain common policy themes and priorities.

The unifying theme of this entry, therefore, is the emergence of a European Urban Policy during the 1990s, the establishment of urban policy instruments in early 2000, and the broad policy objectives that this is designed to achieve. This is related to the broadly parallel development of European spatial policy over the same period, the focus of which will in the early years of the current century be on the development of a balanced and polycentric city system and the forging of a new rural-urban relationship.

The concept of a European policy is applicable specifically to the European Union (EU), which is a jurisdiction whose laws are superior to those of the 15 member states. Of course the EU does not include the whole of Europe but it does cover a sufficiently large number of member states to be representative, particularly taking into consideration the association agreements with certain Western non-members and the expectation that 11 countries in Central, Eastern, and Southern Europe will join and adopt the same policies by 2005–2010.

This entry will begin by discussing the dimensions and variety of urban policies and cultures to be found in Europe and the extent to which certain overall policy objectives are generally accepted. It will then go on to outline the development of urban policy in the 1990s and the state that policy development has reached at the time of writing. The place of urban policy within the wider structure of the EU and of related policy sectors, and the extent to which a European policy community can be identified is then explored. Finally some directions in which policy development may be expected in the next few years are indicated.

2. *Common Themes within Urban Policy in Europe*

Although the urban form and cultures vary greatly between different parts of Europe (Newman and Thornley 1996), there are a number of policy themes that can be discerned very widely, if not universally. It is the exception in Europe for cities to be still growing in the traditional manner as a result of migration from surrounding rural areas. This was the dominant form of urban population growth in the nineteenth century in countries that industrialized early (e.g., Britain, Germany, and Belgium) and in much of the twentieth century in most of Southern Europe. Urban change in the last quarter of the twentieth century was often associated with factors such as migration into Europe of ethnic minority populations from former colonies, economic restructuring and the loss of large numbers of traditional manufacturing jobs, extensive suburbanization, longer distance commuting, and new home–work relationships. Many cities are, of course, attracting substantial population growth. This is especially true of those with the most favorable economic structure or position within Europe's territory. A major concern in recent years has been the extent to which growth is concentrated in cities in the core of Western Europe, the so called 'blue banana' or 'pentagon' (CSD 1999). Many other cities are experiencing growth in the number of households and of the area of urbanized land associated with them without experiencing overall population growth.

National urban policies were widely adopted to address the social and economic consequences of migration, unemployment, and poverty and to promote urban economic development and regeneration. This period of urban policy was also characterized by considerable growth in acceptance of the idea of competition between cities. Several cities that were undergoing economic and social decline adopted policies of image promotion as a significant element in their regeneration strategy. Competition could take different forms. At times it was directed at potential visitors or inward investors for whom a very positive and attractive image was essential. At other times, the competition took the form of seeking to qualify for some form of assistance under national or European policy, for which it was necessary to assert a degree of need greater than that of the competition.

Concern for the urban environment has grown dramatically in the same period. One aspect of this relates to the growth in the use of private car transport and policy towards the construction of urban motorways. Growth in car ownership has been a universal phenomenon in Europe as elsewhere but the 1980s and 1990s have seen significant growth in investment in rail-based urban transit systems and in acceptability of policies to restrain or even prevent the use of cars in urban centres. This latter dimension shows very distinct North–South variations. The Nordic countries accept a very high level of restriction on the use of cars in urban areas while in more Southern countries, such policies are politically very difficult to introduce. Nevertheless several major cities including Athens, Paris, and Rome have suffered such extremes of atmospheric pollution that they have experimented with no-car days and other means of restricting car access.

Cities in the former communist countries of Central and Eastern Europe present a distinct set of urban policy issues. Their existing urban structure is often heavily influenced by the experience of central planning and the lack of any urban land market. Substantial investment and transformation has occurred during the 1990s but, in the context of European urban policy, they will continue to present a distinct set of issues for some years to come, especially in the context of the enlargement of the EU.

A number of identifiable policy themes are widely adopted throughout Europe or are being promoted specifically as European policies. These include policies for urban regeneration, directing development on brownfield, rather than greenfield, locations, urban

economic development, property-led development, policies to combat social exclusion and the problems of peripheral social housing estates, policies to accommodate the growth in household numbers without population growth due to the greater number of one and two person households, policies to counteract urban sprawl and promote the concept of the compact city, and policies to minimize private car commuting and promote the use of energy efficient transit systems. Core–periphery issues are recognized at the European level as well as within several member states and the concept of a balanced and polycentric urban structure has been developed as a central objective of European spatial policy to counteract this.

3. *The Place of an Urban Policy in the EU*

Since the EU originated as an economic community, it is by no means generally accepted that it should play a role in urban policy. Constitutionally, the EU cannot engage in activities for which there is no provision made in the founding treaties, i.e., for which there is no legal competence. Strictly speaking, this limits the scope of EU legislation but is not so restrictive in respect of funding and networking programmes and policy advocacy. In practice, the European Commission, as the main policy-making institution, regularly pushes beyond the strict limits of existing competences, often in anticipation of future treaty revisions. The latest revision, the Treaty of Amsterdam, entered into force in May 1999 (SO 1997).

Many aspects of urban policy have been introduced within the framework of programs for which a treaty basis previously existed, such as regional and environment policies, which are specifically intended to address urban problems although they may not be labelled urban policy. During the 1990s, there was growing political support for the idea that an urban policy should be added to EU competence as part of one of the periodic revisions of the treaties. The overall basis for this is the argument that over 80 percent of Europe's population is urban, and that there is no specific channelling of funding in this direction although a substantial proportion of EU funding is directed at agriculture.

To date the objective of adding an urban title to the treaties has not been achieved. The latest revision, the Treaty of Amsterdam, amended the statement of overall duties and objectives of the Union by including *inter alia*, the phrases 'to achieve balanced and sustainable development ... strengthening of economic and social cohesion,' and 'promoting social and territorial cohesion' (Articles 2 and 7D, Treaty of Amsterdam). These indirectly provide a treaty basis for urban policy, which can be conceived as policies to promote social cohesion. It is a common feature of EU policy development that significant policy initiatives take place before any explicit treaty competence is agreed by the member states. This was true of both regional and environment policies, both now and well established with a clear treaty basis, and it is reasonable to expect that the same trajectory is being followed by urban policy. Meanwhile it is possible to identify several distinct elements of urban policy derived from existing EU policy sectors.

4. *The Evolution of EU Urban Policy*

European urban policy took distinct shape during the 1990s but the foundations in respect of both research studies and innovative policy initiatives were laid during the 1980s. Studies were undertaken into urban problems and into the comparability of urban data from different member states. One of the most meticulous attempts at overcoming data comparability problems and arriving at a consistent analysis was that undertaken by Cheshire and Hay (1989) who developed a concept of the functional urban region for which data was adjusted to eliminate the distorting effects of different administrative boundaries and definitions. Policy initiative during the 1980s included development of the concept of urban black spots, designed to focus the attention of the regional and social funds (European Social Fund [ESF] and the European Regional Development Fund [ERDF]), on what were essentially inner city problems. Up to this time, it was not considered appropriate for EU policy to focus on spatial designations below the regional level. However, the widespread recognition of inner city problems led to political pressure to incorporate policy measures within the structural funds that could address them.

Another element of urban policy thinking that came to the fore by the late 1980s was concern for the urban environment. This moved to the top of the political agenda following the 1989 European Parliament elections where a historically high vote for green parties was recorded in a number of member states and a significant if small green party group was elected.

4.1 *Structural Funds*

Before discussing the outcome of these initiatives, it may be helpful to explain the term structural funds. This term came in to use during the 1980s to refer to three major funding programmes, ESF, ERDF, and the guidance section of the European Agricultural Guidance and Guarantee Fund (EAGGF). The latter, part of the Common Agricultural Policy, is not of further concern here. ESF and ERDF, on the other hand, are very important in the context of European urban policy. The ESF is not a social fund in the welfare sense, but in the sense of providing for the so-called social partners, i.e., employers and those employed. It is a fund designed to promote and steer

Objective 1	Economic adjustment of regions whose development is lagging behind
Objective 2	Economic conversion of declining industrial areas
Objective 3	Combating long-term unemployment and facilitating the integration into working life of young people and of persons exposed to exclusion from the labour market
Objective 4	Facilitating the adaptation of workers to industrial changes and to changes in production systems
Objective 5a	Adjustment of the processing and marketing structures for agricultural and fisheries products
Objective 5b	Economic diversification of rural areas

Figure 1
Structural fund objectives 1994–1999

employment policy. During the 1980s, reforms of the ESF allowed it to focus very directly on the type of job creation and wage subsidy schemes that were being developed in the UK and some other member states as an integral part of inner city policy. The ERDF originated in the 1970s as an augmentation to national regional policies and it was only during the 1980s that it began to operate according to policy objectives set at the European level (Williams 1996).

During the later 1980s agreement was reached on the coordination of the structural funds. This came into operation in 1989. In simple terms, the basis on which it operated was that five overall objectives were set (see Fig. 1). In the case of those with a primarily regional or spatial purpose, the area of benefit was designated. Thus the poorest areas, broadly speaking those with under 75 percent of the EU average per capita GDP, were designated Objective 1 areas and received maximum funding. Objective 2, which focused on areas undergoing industrial restructuring, was also subject to territorial designation and substantial funding. The majority of urban areas suffering economic or social problems were to be found within Objective 1 or Objective 2 areas. There were exceptions, generally whenever poor districts occur within economically successful city regions, such as Hamburg, the assumption being that the necessary resources were available locally.

4.2 Urban Environment

Building upon these foundations, the first major initiative of the 1990s was in respect of the urban environment. This was the preparation of the *'Green Paper on the Urban Environment'* (CEC 1990), known as the Green Book, by the environment directorate of the Commission. This attempted firstly to establish the place of the city in European culture and to argue in a continent where approximately 80 percent of the population lives in urban or sub-urban areas, the quality of urban life and cultural identity are crucially dependent on the quality of the urban environment. More specific urban policy themes were developed, including the link between urban social cohesion and rural poverty as a result of rural–urban migration, the connection between urban development and air pollution, concern for conservation of the cultural heritage and the need to introduce traffic calming, to counteract pressures of urban sprawl and unrestricted tourist development. The Green Book went on to develop more specific policy proposals that could be implemented within the framework of EU powers as they then existed, particularly in respect of urban transport and planning, energy management, management of urban waste, promotion of industry, and heritage conservation.

The Green Book did not lead directly to the development of a European urban policy although it did capture the imagination of many urban planners and have an influence on policy thinking during the following years. One specific outcome was the establishment of the European Academy for the Urban Environment in Berlin (Williams 1996, pp. 204–10). Another was the adoption in 1994 of the Charter of European Sustainable Cities and Towns, whose signatories pledged themselves to work towards sustainable urban development, and establishment of the Car Free Cities Club, whose members were cities prepared to make a political commitment to the promotion of traffic reduction policies and the exchange of experience.

4.3 Urban Pilot Projects

Within the ERDF, provision is made for innovative or pilot projects and for community initiatives, funding programs targeted at specific problems often of an urban policy nature. An important step forward in the development of urban policy was the program of Urban Pilot Projects. This was a series of projects designed to test out new ideas for the delivery of urban policy. The intention of this program was to identify new policy measures and demonstrate the potential they had, not only in the context in which they were developed, but also in respect of transferability to cities in other parts of Europe. Projects were grouped around four major themes:

(a) economic development projects in areas with social problems, such as peripheral and some inner city residential areas with high unemployment levels and limited access to skills and training,

(b) areas where environmental actions can be linked to economic goals,

(c) projects for the revitalization of historic centres to restore economic and commercial activity,

(d) exploitation of the technological assets of cities.

As this was a pilot program, normally only two or three cities in each member state were designated for assistance.

A key feature of the Urban Pilot Programme was the requirement to publicize and disseminate the experience of each participating city. Many conferences and exhibitions were held with this in mind, but it is nevertheless very difficult to identify the extent to which lessons were transferred and applied to non-participating cities with comparable problems. An evaluation of the whole program was carried out in order to identify good practice and inform the development of a more substantial program known as URBAN, funded as a community initiative of the ERDF.

4.4 URBAN Community Initiative

URBAN was launched in 1994, before the pilot projects had finished or been evaluated, as part of the 1994–1999 agreement on the structural funds. This is perhaps an indication of the strength of the political coalition between the Commission and city authorities supporting greater intervention at the urban scale. It was designed to overcome serious urban social problems by supporting schemes for economic and social regeneration, renewal of infrastructure, and environmental improvement. It supports schemes to overcome serious social problems in inner areas of designated cities of over 100,000 population through economic revitalization, job creation, renovation of infrastructure, and environmental improvement. This program had a substantial budget, two-thirds of which went to cities within Objective 1 regions, most of the rest going to Objective 2 areas. One subtext for both these initiatives was to support moves to add an urban competence to the EU treaties at the Amsterdam summit in 1997. Although this was not achieved, the new emphasis on tackling unemployment will have a strong urban dimension, and urban policy development is continuing within existing competences. At the 1998 Urban Forum (see Sect. 4.5), Commissioner Wulf-Mathies emphasized that urban policies would be 'mainstreamed' within the structural funds.

Individual projects normally run for four years, must have a demonstrative character and potential to offer lessons to other cities, and form part of a long-term strategy for urban regeneration. The original intention was to designate around 50 projects, for which a funding partnership between the European Commission and the national authorities was proposed. The pressure for designation from cities was such that a total of 85 projects were approved by January 1997, many of which did not actually start until August 1998 (CEC 1998a), so evaluation has not yet been possible. By the end of the decade, approximately 120 projects had been designated. URBAN has thus proved to be a very popular program, so much so that it was maintained in the post-2000 funding period, contrary to the wishes of the European Commission, due to very strong political support form local and regional politicians throughout Europe.

4.5 Urban Agenda and the Urban Forum

Although the European Commission had initiated these various components of a European urban policy, it was politically necessary for the Commission to proceed with great caution in respect of any suggestion that it was proposing a new role for itself, thus the title it gave to an official communication was 'Towards an Urban Agenda' (CEC 1997). The stated purpose of this document was therefore to assess the extent to which existing policies affected urban areas and to examine the possibility of bringing these together to form a more coherent body of urban policies. It was also designed to stimulate debate among the urban policy professionals and, to some extent, generate pressure on the Commission itself in the form of advocacy of a more coherent EU urban policy from this group. Stimulating pressure from external lobbying groups in this way is not an unusual method of seeking to generate momentum for a particular policy direction. The focus of this pressure was twofold: firstly, to seek to influence governments participating in the discussions concerning the proposed amendments to the European treaties in preparation for the Treaty of Amsterdam; secondly, to contribute to preparations for the Urban Forum in 1998.

There was considerable pressure, especially from political representatives of major cities and urban policy experts, to add a so-called Urban Title to the European treaties in the Treaty of Amsterdam. In the event, this proved unsuccessful in that no Urban Title was incorporated in the Treaty, although the lobby did achieve some indirect success. The phrase 'territorial and social cohesion' was added to the section dealing with overall objectives and duties (Article 7D). This phrase can be taken to refer to the need for coherent spatial and urban social policies. Additionally, the Amsterdam Treaty added an Employment Title, which also provided a basis for intervening in issues of urban economic development and social exclusion.

The Urban Forum was an event held on the initiative of the then Commissioner, Dr. Monika Wulf-Mathies, in Vienna in November 1998. This meeting was attended by over 800 representatives of major cities and urban policy professionals at which a major document outlining the direction in which European Urban policy might go was launched. This was entitled *Sustainable Urban Development in the European Union: A Framework for Action* (CEC 1998b). This document aimed to create the foundation for a more highly developed and targeted European urban policy. The first part set out the general principles, political context, and rationale for the

policy, noting also the applicability of the principal of subsidiarity. The key policy development takes the form of a set of 24 Actions grouped under four broad policy aims:

(a) strengthening economic prosperity and employment in towns and cities,

(b) promoting equality, social inclusion, and regeneration in urban areas,

(c) protecting and improving the urban environment: towards local and global sustainability,

(d) contributing to good urban governance and local empowerment (CEC 1998b, p. 11).

Some of the Actions propose the use of existing policy instruments to meet urban policy objectives more explicitly. For example, Action 8 makes proposals for Structural Fund support for urban regeneration. In order to be implemented, it is necessary for the Structural Funds' operational guidelines to incorporate similar principles and this is being negotiated where appropriate. Other Actions anticipated the new powers, especially in respect of social exclusion and employment promotion, available following ratification of the Treaty of Amsterdam. Certain Actions are directed towards matters within the responsibility of urban authorities, for example, Action 11 on implementation of existing environmental legislation, while others seek to apply to the urban context international agreements, such as the Kyoto strategy on climate change.

5. *The Urban Dimension of European Spatial Policy*

The policy framework for the network of cities and urbanized areas in Europe is not to be found in the urban policy instruments discussed above. Instead it is to be found within the framework of European spatial policy. Spatial policy has been developed over a similar period, reaching a degree of coherence with the adoption of the European Spatial Development Perspective in 1999 (CSD 1999, Williams 1999, 2000). Although this is better understood as a separate field of policy development, concerned with the territorial cohesion of urban and rural areas, it does incorporate certain guiding principles which provide a framework for urban development.

The ESDP was based on certain principles by the ministers responsible for spatial planning in 1994: development of a balanced and polycentric urban system and a new urban/rural relationship; securing parity of access to infrastructure and knowledge; sustainable development, prudent management and protection of nature, and cultural heritage. These principles have no mandatory effect: there is no European spatial plan that must be adhered to. However, these are intended to provide a framework for policy thinking at national, regional, and local levels.

In several member states, most notably France and the UK, the national urban system is dominated by one primate city (Paris and London), and in some others, such as Spain and Italy, two cities jointly dominate the urban structure (Madrid and Barcelona, Rome and Milan). There were fears that, at the European level, a similar situation may come about especially after the creation of the single market in 1992. More specifically the fear was that a core region of Western Europe, from London through Brussels and Amsterdam, including Bonn and Frankfurt and ending with the Milan region of Italy, would become so intensely developed and become so dominant both politically and economically that it would unbalance the overall spatial structure of Europe and consign cities and regions in the periphery to a perpetual state of relative underdevelopment. An influential French study into the spatial structure of Europe gave the term 'Blue Banana' to this core region (Brunet 1989), and the image remains in use.

The principal exception among the major member states to this core–periphery model of urban development was Germany. The pre-1990 German Federal Republic was an especially good example of a polycentric urban structure. Several cities had major national or international functions (e.g., Frankfurt and Bonn) or were internationally significant commercial centres (e.g., Hamburg, Düsseldorf, and Munich). This was perceived to be a more appropriate model for application at the continental scale, and has been incorporated in EU spatial policy as a guiding principle.

6. *The State of EU Urban Policy Development in 2000*

In 2001, EU policy towards cities and urbanized areas can, for the most part, be found in two particular bodies of policy-making. One is the funding program for urban projects and initiatives designed to combat social exclusion, the other is the spatial policy framework for the network of cities and city regions. Policies directed at urban areas can of course also be found within the context of energy, transport, and environment policies, for instance in respect of pilot schemes for energy-efficient transit systems and policies for landfill and waste disposal.

At the Urban Forum in 1998, the Commissioner was advocating that urban policies and their associated funding should be 'mainstreamed' within the overall structural funding program when this was renewed for the funding period 2000–2006. The intention was that cities would not be regarded as dependent upon one relatively small fund for any special initiatives. In fact this proposal was strongly opposed by the representatives of the cities, and it proved necessary to propose the continuation of a dedicated funding program, known as URBAN, in

order to satisfy these objections. This is one of only four Community Instruments, compared with 14 in the earlier funding period and is substantially better funded than its predecessor. It is not possible at the time of writing to comment on what this program may achieve, since the guidelines have only recently been negotiated and have not yet received final approval.

The spatial policy framework is now established in the form of the European Spatial Development Perspective (ESDP), adopted by the Ministers at Potsdam in May 1999. Although this is in no sense a legally binding policy statement, it is anticipated that it will form significant guidance for urban development, especially in respect of networks of cities and links between them, and in respect of the need to minimize urban sprawl and maximize sustainability through compact urban development and efficient transit systems. An action program to follow up the ESDP was agreed in October 1999, which included further study programs designed to develop and operationalize concepts such as a balanced polycentric urban structure.

7. *Urban Networking and the Development of a Policy Community*

A noteworthy feature of urban policy-making during the 1990s has been the extent to which cities have attached importance to networking activities. The term 'policy community' is sometimes applied to the entire group of people and organizations that play a role in any given sector of policy-making. This is readily recognizable phenomenon within the national context. What is new is the extent to which it is becoming possible to identify a European policy community in a number of sectors including urban policy.

The European Commission has sought to stimulate networking and partnership activities through a number of programmes requiring participant cities to form partnerships with cities in other member states in order to qualify for funding. Some partnerships were fairly short-lived but several small networks were established which have since grown into quite substantial organizations. One of the most significant is a body called 'Eurocities,' whose members are all regional cities of over 250,000 population. The original core group was formed by certain 'second cities,' Birmingham, Barcelona, Lyons, and Frankfurt, which were seeking to support each other in their efforts to achieve a higher European profile, outside the shadow of their respective national capitals. Eurocities has members from throughout the EU, and now includes capital cities. It is the major non-government organization currently active in the field of European urban policy and has a wide range of working parties concerned with different policy sectors—urban environment, transit systems, telecommunications, economic development, social exclusion, etc.—which not only react to proposals and consultations from the Commission but also prepare and submit their own policy proposals. The URBAN Community Instrument is itself to some extent the product of this process. Eurocities has also been instrumental in promoting its own specific networks, for example the Car Free Cities Club, a network of cities that have committed themselves to implement measures to reduce the impact of cars in city centers.

There are a number of other networks representing smaller cities, or cities with particular functions or situations, such as historic cities that have difficulty reconciling heritage conservation and tourism pressure, or cities sharing a particular manufacturing industry.

8. *New Policy Concepts and Agendas*

At the beginning of the twenty-first century, very substantial progress has been made, compared with the situation a decade earlier, in developing a coherent body of European urban policy. The principle that the EU should contribute substantially to meeting urban policy objectives is now accepted and for urban policymakers in city administrations, this is a major factor in their daily professional lives. On the other hand, the ambitions of several of those who were actively seeking the incorporation of urban policy within EU competences have by no means been fulfilled. There is still no Urban Title in the treaties, and therefore there are only weak powers to legislate, and only on the basis of unanimous political support.

The present situation is not likely to remain static. The spatial policy aspect is continuing to evolve, associated with study programs which will seek to develop policy instruments to promote a polycentric urban structure and more integrated urban–rural relationships. It is anticipated that the link between the European Spatial Development Perspective and the Structural Funds will be made explicit for the 2006 revision of the Funds. Urban environment and transport initiatives can be expected to continue, with strong political support especially from Northern Europe. It was always envisaged that the Urban Forum should become a regular event. Although this has not yet happened, it would be inconceivable for this not to happen, given the political strength of the cities of the EU. Finally, the next round of enlargement of the EU, to incorporate former communist countries of Central Europe, will take place during the period 2005–2010. The impact of this is yet to be fully appreciated among existing EU members, in spite of many studies and technical assistance programs that have taken place. However, the urban regeneration and social cohesion problems that these will bring will have a more profound influence than

anything hitherto experienced on the future development of EU urban policy.

See also: Europe: Sociocultural Aspects; European Integration; European Union Law; Migration: Sociological Aspects; Urban Policy: North America; Western European Studies: Environment; Western European Studies: Geography

Bibliography

Brunet R 1989 *Les villes européennes*, La Documentation français. RECLUS, DATAR, Paris

CEC 1997 *Towards an Urban Agenda*, COM (97) 197 final, Brussels, Belgium

CEC 1998a *URBAN Community Initiative Summary Descriptions of Operational Programmes*, Directorate-General for Regional Policy and Cohesion, Brussels, Belgium

CEC 1998b *Sustainable Urban Development in the European Union: A Framework for Action*, Directorate-General for Regional Policy and Cohesion, Brussels, Belgium

Cheshire P, Hay D 1989 *Urban Problems in Western Europe*. Unwin Hyman, London

CSD 1999 *European Spatial Development Perspective*, Committee of Spatial Development, Potsdam, Germany

Newman P, Thornley A 1996 *Urban Planning in Europe*. Routledge, London

SO 1997 *Treaty of Amsterdam*, The Stationary Office, Cm 3780, London

Williams R H 1996 *European Union Spatial Policy and Planning*. Paul Chapman Publishing, London

Williams R H 1999 Constructing the ESDP–Consensus without a Competence. *Regional Studies* **33**(8): 793–7

Williams R H 2000 Constructing the European Spatial Development Perspective—For Whom? *European Planning Studies* (forthcoming)

R. H. Williams

Urban Policy: North America

In general usage, 'urban policy' takes on two closely related meanings. The word 'policy' may denote a definite course of action selected from among alternatives and intended to guide specific decisions. In that case, 'urban policy' refers to such courses of action in relation to urban areas and issues, especially in the public realm. 'Policy' may also indicate a more general and high level guide to decision and action, particularly by governmental or other public entities. In this case, urban policy should be seen as comprising general principles that guide public action in urban areas and on the issues associated with them. In practice, it is difficult to disentangle the two, especially in advanced societies, such as those in North America, where most of the population lives in urban areas. Furthermore, the diversity of urban areas within a national political system, no matter whether they are metropolitan regions, cities, or smaller urban places, engenders specific forms of complexity that are reflected in public policy. Thus, to understand urban policy, it is necessary to see how public decisions are formulated, as well as to examine the character of urban policies in both their specific and general senses.

1. *Public Action in Relation to Urban Areas*

The organization of public action reflects a profound dilemma associated with geographically differentiated places whose populations have political power. Metropolitan areas, cities, and other urban entities are the most dominant such places in our time. Considerations of technical efficiency often suggest that public action should be shaped by functional requirements, rather than in accord with governmental structure (examples are transportation, public utilities, and public safety). Such function-based structures might be seen as a 'vertical' form of organization, in which a particular function is managed coherently to take advantage of economies of scale. Most national states have ministries that reflect such a functional division, as do lower levels of government. In strongly unitary states, such national ministries extend their reach to virtually all aspects of life, even at the most local or individual level.

Society, however, comprises much more than the nation state, being a vastly complex structure of social, economic, and political relationships among individuals and groups. The reality of everyday life engenders strong relationships in a 'horizontal' form, crosscutting functional concerns and focusing on place itself. Such 'place based' concerns are accentuated where people in subnational communities have the political power to express their collective desires, even though they may run counter to national concerns expressed through vertical forms of organization. Democratic societies in which legislatures are elected on a geographic basis are particularly likely to express such horizontal pressures. However, the fundamental social organizational forces that create subnational communities are present everywhere, with concomitant tensions between the vertical and horizontal, the functional and the communal, at all levels of government.

One solution to this dilemma is a federal system, in which both legislative and executive responsibilities are allocated explicitly to different levels of government, usually associated with accompanying geographic disaggregation of the national space. This is so in all three countries in North America. Conventionally, systems of this type exhibit at least three formal levels of government: national, state (also called regional or provincial), and local, with the latter often further subdivided. Urban areas are rarely national in scale or

scope, though Singapore provides an important exception. Singapore is, of course, a unitary state, albeit with local forms of administration within itself. Urban areas at the first level below the national in a federal system are also relatively infrequent, though Mexico City and Washington DC are principal examples. Even there, the capital is a unique political entity rather than a state within the federal system. Thus, most urban areas in explicitly federal systems fall under the jurisdiction of local governments.

Although urban places are geographically local in nature, urban policy is not also essentially local. To the contrary, cities are intricately involved in national politics in urbanized societies. Many strands of national policy affect urban areas or are intended to respond to the concerns of their inhabitants. The formation and execution of urban policy occurs at all levels of government. In addition, many policies that affect urban areas and their inhabitants deeply are couched in terms of other policy domains. Major governmental ministries or agencies whose explicit mission is to shape urban policy are rare. Nonetheless, urban places are the targets of many explicit policies, while also being the recipients of the consequences of others. It is often difficult to distill the urban policy implications of policies formulated in other fields.

2. *Urban Policy in North America*

North America is a continent comprising three separate nation states, each with a distinctive history, its cultural and demographic makeup, and its own economic, social, and political system. Nonetheless, Canada, Mexico, and the United States do have much that is common in their forms of urbanism, their governance, and the policies they have evolved. All three countries are highly urbanized, although Mexico has a somewhat lower proportion of its population living in urban areas, reflecting its status as a less-developed economy. In each country, market forces drive urbanization, though the extent to which they are regulated varies substantially. Cities in each are shaped by the dominant technologies of our time, especially by the automobile as the primary means of transportation. Demographic changes, especially immigration and aging, are affecting urban areas in each country. Nonetheless, there are important themes of urban policy in each. Critical to understanding those themes are the categories of public policy involved and the governments that address them.

Although 'public policy' is a widely used term, which now drives graduate and professional education and research, it is seldom voiced in governmental decrees. In relation to urban policy, it is especially important to differentiate the ways in which public policy may be understood. In the first meaning noted above, it is the expression of a particular public course of action that has been decided or is in the process of debate. More generally, in line with the second definition above, public policy also expresses a higher level of generality in the principles that guide public, especially governmental, action. This second sense, however, opens some important distinctions of its own. Much public policy, in fact, may be settled, broadly accepted, and challenged only when conditions change sufficiently to lead to questions being raised about its validity. Urban policy, in this view, is something like an iceberg, much of which is virtually invisible.

2.1 *Urban Policy as Land Use and Development*

Consider the fundamental question of how urban areas in most of North America are designed and built. Current policy debates concerning land use planning, urban sprawl, and transportation infrastructure investment are shaped by long-agreed public policy that urban development should be carried out by private investors, regulated by government and, in part, supported by public investment. Europeans established this policy in the early years of settlement of North America, even though the first settlements were far from pure capitalist enterprises. As colonies of Britain, the Netherlands, France, and Spain, they were intended to benefit their parent countries, and to varying degrees they were rigorously regulated and their forms prescribed. Nonetheless, in the United States, especially, there had emerged by the early nineteenth century a distinctive form of settlement that was to define the American form of urbanism, namely the grid-plan.

The urban grid, a rectilinear structure of streets, was the physical expression of a fundamental policy that urbanization would be market driven, privately organized, and minimally regulated. It enabled two key elements necessary for such a process to work. The first was the creation of easily laid out and marketable lots that could be sold speculatively to buyers at a distance. The second was the easy expandability of the urban structure, simply by adding further streets and subdivisions as the market expanded. Thus, a distinct and recognizable form of American city emerged, speculative in its origins, highly adaptable to changes in growth and technology, and based on the conception of land as a marketable commodity. Even when the grid was replaced in the twentieth century by the curvilinear streets of the suburbs, development remained a function of the private market. The suburbs, with their stronger appeal to the consumer's sense of aesthetic quality and lifestyle choice, necessarily expanded the realm of public or quasi-public regulation, but they were equally the product of private, speculative development. With this shift, however, came major public policies that have supported this form of urbanization, notably the federal funding of major highways, and the deductibility of

mortgage interest, the latter justified by an explicit policy of encouraging home ownership rather than rental tenure. The most recent challenge to suburban forms of urbanism, by neo-traditional or 'New Urbanist' critics, calls only for a change in the way in which urbanization is treated by regulation and taxation. Only relatively rarely in modern American history has government attempted directly to create urban places, even though national, state, and local laws have fundamentally shaped urban settlement patterns. US efforts to build new towns during the Great Depression of the 1930s were unusual in this respect and they failed.

Clearly, this account simplifies a complex, continental-scale reality. The legacy of Spain in Mexico was one of a much more centralized, class defined, and bureaucratic society. Mexican urban development reflected those facts in its slower growth and more constrained forms of urbanism. Similarly, development in Canada reflected both the earlier French centralizing influence and the British legacy of unitary parliamentary government, albeit grafted onto a federal system in which the provinces were relatively strong. It is no accident that public control of urban development in both Mexico and Canada remained historically much stronger than in the United States, yet in both countries cities are not built by governments.

The idea that urbanization is market driven has been powerfully consequential. Paradoxically, the market itself implies public constraints. No market functions without some prior structure of behavior or public order; some constraints must be established for markets to function at all. For urban policy, this implies public regulation of the urban development process. The nature and locus of that regulation has defined many of the key urban policy issues in North America during the twentieth century. In the United States, control of public policy is ultimately vested either at the level of the states or of the federal government, depending on the Constitution and its interpretation by the Supreme Court. With respect to land use, control is generally agreed to rest legally at the state level, although the basic legal doctrines that support land use regulation rest on constitutional decisions that are national. In fact, states have delegated land use regulation to local governments under a variety of names. California, for example, delegates the power under the doctrine of 'Home Rule.' On this legal basis, aided by a long history of piecemeal settlement and independent formation of urban places, US urban policy toward development and change is played out at the most local level.

This situation contrasts sharply with that in many other advanced countries, notably in Europe and Asia, where state or central governments play more centralized roles in land use planning, regulation, and development. Canada and Mexico also present substantially more centrally regulated versions of urban development. In Canada, cities in the central and western provinces tend to look more like those in the USA. The eastern provinces have a stronger tradition of central control, though always empowered at the provincial level. Toronto's ability to control its suburban development contrasts markedly with any US city, but a change in political complexion at the Ontario provincial level has had substantial implications for the city's policy. In Mexico, until very recently, with a single-party state, local control of development was limited. Although state or central governments did not always have the capacity to enforce their will more than intermittently over economically or politically powerful local groups, the tradition of central government finance and fiat remained powerful. How this will evolve in the first decade of the twenty-first century, as the political monopoly of the PRI ends, remains to be seen.

2.2 Urban Policy as Functional Management

Local power over development has major implications for urban development, particularly when viewed from a functional perspective. For decades now US observers and advocates have argued that urban problems of transportation, energy, utilities, and environmental quality require the political ability to act at a regional level. Again and again, such arguments have been turned back as localities have refused to form larger regional governments or to cooperate in accepting land uses that they deem undesirable. That refusal has been encapsulated in the term 'NIMBY' ('Not in my back yard'), which has become both a rallying cry for control by local neighborhoods in the face of larger economic and political forces, and a term of opprobrium conveying selfishness and exclusion. Nonetheless, metropolitan areas cannot be built and cannot operate at a high level of amenity and income without necessary functional efficiency. One widely employed means around the problem has been the creation of single purpose special districts; for example, to handle water supply. These are typically limited in their range of action and their ability to override local decision-making. They are established by state law, but with the clear requirement that local political representatives support their formation. In recent decades, some entities of this type have also emanated from federal governments in all three countries.

Management of air quality is a good example of such urban policy, illustrating both its urban character and its placement within another organizational realm. Awareness of the dangers of air pollution has a long history in urban areas, but its modern manifestation in the USA may be dated from the passage of clean air legislation by the city of Pittsburgh in the 1950s and the enactment of state and federal air quality standards and legislation in the 1970s. At first, efforts to improve urban air quality focused on

stationary sources of pollution, especially industrial sources such as steel mills and oil refineries. As they were controlled, however, it became apparent that the single greatest source of air pollution was from mobile sources, especially private automobiles. In metropolitan areas and cities such as Los Angeles, dirty air was identified as a contributing cause of serious public health problems and reduced quality of life associated with smog. Clearly, these were regional and in some cases inter-state problems. The federal government's response was to treat environmental quality by regulatory legislation and the formation of the US Environmental Protection Agency in 1970. State governments reinforced the effort with similar legislation.

Air quality degradation by automobiles can be countered in many ways, including regulating the emissions permitted, increasing the cost of driving, providing incentives for use of cleaner vehicles, and changing the way in which urban areas are structured so as to reduce the number of miles driven. Any of these policies are conceivable, but the specifics of implementation differ among them. Furthermore, the mode of intervention might have a great deal or very little relationship to urban areas themselves. Regulation of automobile emissions through national specification of permissible technologies is only indirectly urban in character, and its effectiveness will vary with the standards applied from place to place. On the other hand, consider the problem of regulating development as a means to control emissions in metropolitan Los Angeles. Although both federal and state agencies have regional offices, neither would typically be able to carry out such a policy effectively at the local level. To do so for the metropolitan area required also a regional agency, in this case, the South Coast Air Quality Management District. This agency, established directly by the state, legally has the power to override local development decisions. However, in the face of overwhelming political opposition to such actions, which would violate home rule, it has been reluctant to do so, preferring to employ its regulatory powers in other directions. Alternative air quality policies may have similar goals, but their urban manifestation may be quite dissimilar.

This story of policy directed towards urban problems, defined on a functional basis, and legislated and implemented at multiple levels of government can be found again and again in North America, as it can everywhere in the urbanized and urbanizing world. Apart from being much worse than in Los Angeles, air pollution in Mexico City presents much the same set of problems. In transportation, public health, housing, water supply, wastewater disposal, garbage removal, and public safety, governments have been attempting to find solutions within their own financial and political situations. It is not a new phenomenon. From the beginning of the Industrial Revolution, much of what we recognize as urban policy has been driven by attempts to deal with the dysfunctions of the industrial process. This process occurred a little earlier in Britain than in most of Europe and North America, but once on course it has barely slackened. Some urban concerns, of course, are much older. For example, laws to prevent disastrous fires by prohibiting wooden buildings may be found in the very earliest records of urbanization. However, the emergence of systematic urban policies really dates from the nineteenth century.

For early industrial cities in North America, the most urgent issues were public health, public safety, and transportation. The previous section indicated how the use of a planned grid permitted both a functioning land market and lines of transportation. Although the cores of cities were often congested, especially where, as in Boston, their cores were essentially unplanned, they could function with pedestrian and horse-drawn transportation before the advent of the railroads. However, such cities were necessarily densely built, outran their water supplies, and had primitive sewage disposal. Frequently, they endured epidemics, notably of cholera and yellow fever. As epidemiology evolved, the importance of clean water and sewerage became evident and the first great modern age of urban policy began. Although the delivery of pure water was, at first, a privately financed and run form of enterprise, it soon became a municipal monopoly, as did sewerage. Provision of electric power and gas, on the other hand, primarily remained private enterprises in North America, with a few minor exceptions. These utilities were regulated after the turn of the twentieth century in order to limit their monopoly power. Public safety became a key municipal function with the creation of police and fire forces in the nineteenth century and has generally remained so.

Beginning with the development of urban trolley lines and railroads in the nineteenth century, transportation investments have played an especially critical role in urban policy over the past 200 years. Although streets and street maintenance have primarily been managed by local governments, the provision and regulation of transportation has involved both private markets and public entities and governments. In the United States, most public policy in this realm was enacted by state and local governments throughout the nineteenth and early twentieth centuries. The ability of dense but growing cities to expand depended on fast and reliable forms of transportation so long as employment remained centralized. Privately run trolley lines, buses, subways, and commuter railroads, operating under local or state charters, provided a workable, if often uncoordinated solution. Later, many were consolidated under public ownership. Together, they enabled development of the suburbanized city of the early twentieth century, brought to full flower by the automobile. With the advent of automobiles, the focus of public

policy shifted from private investment in infrastructure to public investment in highways, funded in large part by taxes on the sale of gasoline. National government engagement in transportation within cities remained minor until the 1950s. By that time, the invention of the freeway had raised the cost of new investment beyond the capacity of both local and state governments.

Large-scale federal entry into the process began with the creation of the Interstate Highway Program in the 1950s. Despite its formal purpose of enhancing national defense by providing high-speed, controlled-entry highways linking the entire country, the Interstate system played a substantial role in urban development through the second half of the twentieth century. By the year 2000, urban freeways ran both through and circumferentially around every major urban area in the USA, fostering suburban decentralization of cities across the continent. Similar freeway development took place in Canada and, more recently, has been occurring in Mexico. Although there has been much debate about the precise intent and impact of the Interstate system on US cities, there can be little doubt that the availability of federal funds has enhanced suburban development significantly since the 1960s. Whether these were conscious policy choices and whether this was the primary cause of the decline of central cities, discussed below, cannot be demonstrated. Given the rapid growth of population, productivity, and income, together with consumers' desires for single-family housing and automobile transportation, it seems likely that economic and political pressures would have had similar effects, whether or not the federal government were involved. Nonetheless, this period saw the entry of federal policy into urban transportation in a major way. It remains there today, although the nexus has shifted. Since the 1980s there has been a resurgence of political support for urban mass transportation, driven by the perception that the freeway solution is either too costly or too destructive of environmental quality to be sustained. This perspective parallels the critique of the suburb and urban sprawl discussed above. As a result, federal transportation funds now include substantial subventions to mass transportation—buses, light rail, and commuter rail in metropolitan areas. While the effectiveness of this shift is being debated, it now appears to be an entrenched element of national policy towards urban areas.

2.3 Urban Policy as Social Policy

Since the 1960s, urban policy in the United States has not been associated in the media and public debate primarily with the issues discussed so far, though the recent resurgence of concern about urban sprawl has shifted attention back to the effects of urban growth. Rather, for several decades, urban policy has been identified primarily with problems of poverty, racial discrimination, segregation, and the decline of central cities. As is true in so many other instances, such a definition of urban policy is not new. It emerged in tandem with the industrial city in the nineteenth century, when low wages, intense overcrowding, dilapidated housing, crime, social disorganization, and disease combined to generate a new form of urban slum. The public health effects of industrial urbanism first engendered policy responses that were engineering solutions, especially in the form of potable water supplies and waste disposal. Epidemics and infectious diseases, such as tuberculosis, also required medical responses, which took the form of public hospitals and health services. Crime problems engendered police forces and prisons.

However, by the late nineteenth century, the fact that so many of the new urban populations in the United States and Canada were immigrants, led to concerns for public safety in a larger sense, namely the stability of the society itself. Public education became a key element of local urban policy, its intent being both to provide a literate and capable labor force for the expanding economy, and to ensure that children were able to speak English and become acculturated into society at large. Fears that immigrants would bring with them European socialist ideas and form labor parties were not unjustified, for all that they were not realized. The momentum for these policies was certainly national in scope and linked to the prevailing sense of the United States as an emergent imperial power on a world scale. However, their realization was almost entirely at the state and local levels, where governments and influential voluntary groups were at times almost indistinguishable.

Poverty presented not only a risk to the health and social position of the affluent in American cities; it also affronted their belief in the United States as a place of freedom and opportunity. Efforts to help the 'worthy poor' raise themselves out of poverty took many forms, of which the Settlement House movement is the best known, evolving into forms of assistance associated with the social work profession. This movement also provided an early means for women of higher social status to become active in professional and social realms. Thus, by the early twentieth century, many believed that urban poverty called for effective governmental and voluntary actions.

A key realm of response was housing, and it is here that the largest challenge to the idea of urban development through the market emerged. The slum was perceived in many ways, but the vision of grossly overcrowded, unhealthy, dilapidated housing was intrinsic to all of them. Observers and critics of capitalism had noted the inability of markets to provide adequate urban housing from early in the Industrial Revolution. Public efforts to regulate construction or management were largely ineffectual, however. By the end of the nineteenth century in the

USA, as in Britain and Europe, there had emerged two competing policy responses. Public regulation of housing quality took the form of building and housing codes, enforced by local bureaucracies. Those who saw regulation as insufficient or ineffective called for direct public or non-profit ownership and rental of housing for low-income families. Market advocates opposed both.

The twentieth century saw the parallel evolution of these approaches to low-income housing policy in all three countries of North America. Mexico, which was undergoing a revolution and installation of a one-party state, probably exemplifies the highest level of commitment to public provision. However, the lack of resources due to underdevelopment, and the mistaken effort to adhere to unrealistic standards, ultimately ensured that public housing would become only a very limited part of its mass urbanization. In common with other developing countries, urbanization for the poor primarily meant informal housing on the margins of major cities. Canada, with a stronger tradition of government intervention, though varying greatly across the provinces, has also adhered to policies that give greater weight to provision of low-income housing. In both the United States and Canada, the tradition of local regulation continues, with rising standards and incorporation of new areas of oversight. In the United States, a determined effort to create a public housing program in the 1930s saw its culmination in the Housing Act of 1949, which led to construction of federally funded rental housing for the poor in most cities. However, the program became discredited, in part because of social problems emanating from its increasing racial segregation and focus on the very poorest populations, and in part because of insufficient funding to maintain and control high-rise housing that was ill-designed for its purpose. In a major shift of federal policy affecting cities, the program was largely replaced after 1974 by a low-income rental subsidy, which remains the basis of federal policy in this area. As a result, direct provision of new housing for the poor has switched towards the nonprofit sector.

Entangled with housing and poverty policy, especially in the United States, is the issue of race. American cities have long been segregated, both by race and income. African-Americans were confined by economics and social pressure to limited residential areas within cities. This phenomenon was exacerbated by major migration from the rural South to northern cities throughout the first half of the twentieth century. While that migration resulted at first in opportunity and rising incomes, by the 1950s there had also emerged ghettoes of African-American and other minority groups in poverty, in which serious social problems persisted. This process was exacerbated by the out-migration from older cities of white middle- and working-class populations seeking a suburban lifestyle and impelled also by race-based fear and prejudice. As a result, cities' financial capacity to address the needs of their populations declined. Education and public service quality diminished, which reinforced the migration. With improvements in civil rights during the 1960s, yet with racial discrimination still widespread, the contrasts between racial groups' aspirations and achievement burst out into urban civil disorder on a massive scale. Large areas of cities including Detroit, Los Angeles, and Washington DC were devastated.

The question of appropriately responsive policy is almost paradigmatic of urban policy in general. On the one hand, efforts were made to improve incomes and opportunity through broad programs, such as those for education and job training. On the other, were efforts to empower local communities through a variety of supports and incentives for the formation of community organizations. The so-called 'War on Poverty' of the 1970s exemplified both kinds of policy. It certainly led to the formation of new political networks and organizations among the urban poor, with long-term results on the racial and ethnic makeup of local politics. However, the position of the poorest and most socially disadvantaged populations improved little, if at all, in the face of a widening gap in the income distribution at large and a decline in the demand for unskilled labor. Increasing crime and social disorganization associated with rising drug use also contributed to the problem. In the face of falling national rates of economic growth, budget stringency led to the withering away of urban poverty programs. In the mid-1990s, these issues were once more addressed by national policy, operating this time through the medium of the welfare program, and especially its support for unmarried parents. Insistence on entry into the labor market has reduced the level of dependency substantially, but the results have yet to be tested outside a boom economy.

3. *Conclusion*

Urban policy in North America remains something of an anomaly. National governments have few policies that are identified explicitly as urban, yet most of their people live in urban areas and most national domestic policies have powerful urban implications. Urban areas contain their own local governments within larger federal and state structures, and those governments pursue the welfare of their inhabitants and other organizations within their boundaries. States and provinces occupy an intermediate role within the federal structure of each nation state, establishing the ground rules for much that urban governments may do while intervening directly. Although urban policy has come to be identified particularly with some sets of issues, notably those to do with poverty and racial segregation, in fact it is much broader than this. Indeed, it may be argued that many of the most

important components of urban policy in the countries comprising North America are scarcely visible. Cities are the most dynamic form of human social organization. Urban policies reflect fundamental social agreements about how cities and towns will continue to be built and transformed and how their inhabitants will relate to each other. Only when those understandings are challenged does the visible portion of urban policy emerge, usually in the context of political and social conflict. Resolution of those conflicts has been a part of urban life throughout history and it will continue to be so.

See also: Land Use and Cover Change; Planning, Politics of; Planning Theory: Interaction with Institutional Contexts; Planning Theory (North American); Residential Segregation: Geographic Aspects; Urban Geography; Urban Places, Planning for the Use of: Design Guide; Urban Policy: Europe; Urban Politics: United States; Urban Sociology

Bibliography

Bobo L, O'Connor A, Tilly C 2000 *Urban Inequality in the United States: Evidence from Four Cities*. Sage Foundation, New York

Burchell R W 1998 *The Costs of Sprawl—Revisited*. National Academy Press, Washington DC

Cullingworth B 1997 *Planning in the USA: Policies, Issues, and Processes*. Routledge, New York

Duany A, Plater-Syberk E, Speck J 2000 *Suburban Nation: The Rise of Sprawl and the Decline of the American Dream*. North Point Press, New York

Dye T R 1998 *Understanding Public Policy* (9th edn.). Prentice-Hall, New York

Fischler R 1998 Health, safety, and the general welfare: Markets, politics, and social science in early land-use regulation and community design. *Journal of Urban History* **24**(6): 675–719

Frisken F (ed.) 1994 *The Changing Canadian Metropolis: A Public Policy Perspective*. Institute of Governmental Studies Press, Toronto

Hall P 1996 *Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century*. Blackwell, New York

Hall P 1998 *Cities and Civilization*. Pantheon, London

Hayes R A 1995 *The Federal Government and Urban Housing Policy; Ideology and Change in Federal Policy* (2nd edn.). SUNY Press, New York

Judd D R 1998 *City Politics: Private Power and Public Policy*, 2nd edn. Longman, New York

Keating W D 1996 *Revitalizing Urban Neighborhoods*. University Press of Kansas, Lawrence, KS

O'Flaherty B 1996 *Making Room: The Economics of Homelessness*. Harvard University Press, Cambridge, MA

Newman P, Kenworthy J 1999 *Sustainability and Cities: Overcoming Automobile Dependence*. Island Press, Washington, DC

Stein R M 1991 *Urban Alternatives: Public and Private Markets in the Provision of Local Services*. University of Pittsburgh Press, Pittsburgh, PA

US Department of Housing and Urban Development 1997 *The State of the Cities* Washington DC

Wilson W J 1990 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago

M. B. Teitz

Urban Politics: United States

By the time Banfield and Wilson's major text, *City Politics*, was published in 1965 the book and its title already seemed inadequate to describe the processes of conflict, competition, and negotiation taking shape in the public sphere of America's large metropolitan centers (Banfield and Wilson 1965). According to Banfield and Wilson, 'city politics,' which 'rarely occupy the serious attention of practical men,' simply concerned who got elected to office, how to apportion the local tax burden, and where to locate various public facilities in the city's ethnic checkerboard (p. 18). But American cities in the 1960s had begun to change since the immediate postwar decade, as blacks and other racial minorities began to assert themselves at the ballot box and in the streets, as population and economic activity migrated to suburban locales, and as neighborhoods suffered under the destructive onslaught of drugs, crime, and family disintegration. Hence the emergence in common use of the term 'urban politics,' a more resonant, grittier description of the processes by which people and groups in big cities sought to mobilize and allocate public resources to deal with these new developments.

If the term 'urban politics' was originally meant to dramatize the emergence in metropolitan centers in the 1960s of an altered civic landscape, it has now come to conventionally refer more simply to patterns of politics in large cities. These patterns are in fact constantly evolving. The texture and scope of urban politics in the US are best understood by examining three closely intertwined patterns: the shifting nature of political conflict in big metropolitan centers, the changing preoccupations of local politics, and the functional evolution of city government as the institutional center of the urban political process.

1. Shifting Patterns of Political Conflict

In the first half of the twentieth century, politics in big American cities was typically defined by the conflict between working class European immigrant groups and middle class native Protestant interests. This was a period of Irish Catholic ascendance in cities as diverse as Boston, Chicago, and San Francisco, but other ethnic and religious minority groups also emerged in local politics in these decades, including Eastern Europeans, Italians, and Jews. The charac-

teristic vehicle for promoting ethnic political interests was the 'political machine' a disciplined, hierarchical organization designed to capture local elective office for the purpose of distributing city jobs and spending to coethnics and allies. As the power of ethnic machines took shape, Protestant political and business elites responded to the threat of displacement by promoting municipal reform, an agenda of structural changes, such as city manager government and the nonpartisan ballot, designed both to make city government more businesslike and to undercut machine dominance of electoral competition.

By the 1960s the last vestiges of machine organization had crumbled in most cities, in no small part because of their inability to reconcile the demands of newly militant African American voters, who the machines had historically exploited, and their traditional white ethnic base, fearful of housing and school integration (Grimshaw 1992). It was in this context that urban political conflict evolved from class and ethnic divisions within the white community to competition based predominantly on race.

In the two decades after 1967 black mayors came to power in more than 300 towns and cities in the US, including nearly every one of the largest places. In the typical contest for city hall, the electorate polarized along racial lines, although skillful black politicians often managed to win just enough white votes to carry them into city hall (Sonenshein 1993, Kleppner 1985). In cities like Chicago and New York the politics of race sundered the traditional alliance of Jews and blacks, who had come together in the early civil rights movement, leaving a stark racial divide. In cities that became majority black, such as Atlanta, Cleveland, and Detroit, black challengers eventually emerged to run against black incumbents. These black-on-black elections have sometimes featured one candidate who is the clear choice of white interests, generating divisions in the electorate identical to the earlier white-on-black contests. However, with time, the prospect of a black mayor became less threatening to most whites, and many municipal elections with one or more black candidates have now lost their sharp racial edge. Blacks began to win the executive office in cities where black voters constituted less than a majority, such as Seattle, Minneapolis, and Denver.

Racial interests continue to define much of urban politics—police conduct and public school reform are examples—but new patterns of conflict have emerged around a broad set of development issues. In the old downtowns of major cities, the issues are how to use central space and how to apportion the costs of development between the public and private sectors. Developers of office towers compete with developers of loft housing for the attention of mayors and their economic development agencies, and both in turn compete with casino and hotel investors, professional sports teams, preservationists, and public parkland advocates. Another line of cleavage has opened between those who seek to invest capital in commercial and entertainment properties and those seeking to maintain neighborhood residential communities that occupy desirable development locales (Logan and Molotch 1987). Conflicts over development often dominate public discourse in major cities, transcending racial and ethnic cleavages or cutting across them in unexpected ways.

2. *Changing Preoccupations of Urban Politics*

In the heyday of the ethnic political machine, the object of politics was to win control of local government in order to get access to its spoils. Politics was a means of livelihood. Thus, the Irish, skillful practitioners of the politics of patronage, came to dominate the municipal workforce early in the twentieth century (Erie 1988). Eventually, the spread of civil service merit systems diminished the number of jobs that could be handed out by urban bosses.

The chief preoccupations of urban politics began to shift with the political emergence of African Americans. Certainly many of the political goals of black Americans recall those of the Irish and other white ethnic groups of the earlier period: blacks want a fair share of municipal employment and city contracts. But they brought to the streets and the ballot box, a political agenda whose instrumental ends were shaped by broader social justice concerns. Urban politics in the decades of black emergence revolved not simply around demands for jobs, but for an end to police bias, housing discrimination, and school segregation (Sugrue 1996).

By the 1990s black mayors were commonplace. Most urban police departments had been racially integrated, and many were administered by black police chiefs. School integration as an end in itself had receded in importance for many black parents, replaced by a determination to improve public education through vouchers or charter schools. As racial justice goals lost their urgency, a new generation of black mayors and their white colleagues as well, turned their energies to the development agenda.

Urban development takes a variety of forms, and choices are inevitably contested. Many cities have pursued a strategy designed to attract tourists and business visitors, investing in convention centers, stadiums, and riverside entertainment centers. Advocates for neighborhood housing, public schools, and social welfare programs tend to oppose such projects as frivolous. Other cities pin their development aspirations on technology parks; yet others on creating the infrastructure—world trade centers, port and airport facilities—to launch themselves into the global economy.

A central responsibility of the public sector in the development process is to create conditions conducive to profitable private investment. Such an undertaking

has typically inhered in crafting a favorable tax climate, adopting flexible land use regulations, and building good infrastructure. One means by which urban mayors integrated these tasks in the 1990s was to embrace a management reform agenda that proponents claimed would 'reinvent government' by introducing such initiatives as privatization and competition and by empowering public employees to act as problem-solvers rather than cogs in the bureaucratic wheel (Goldsmith 1999). Faithfully pursued, such a strategy was expected to impress and reassure investors by reducing the size and reach of local government, making it more efficient and responsive, and easing the tax burden. 'Government reinvention' became a preoccupation that transcended the lines of partisan and racial divisions, reducing some portions of urban politics to a technocratic debate over the nature of public management.

3. *Evolving Municipal Functions*

Through the first half of the twentieth century the singular function of city government was to offer a set of basic public services to promote public safety and health and to facilitate local commerce. These traditionally local responsibilities included public safety, streets, sanitation, libraries, parks, water systems, land use regulation, and occasionally, mass transit systems. City governments received little fiscal assistance and little guidance from either the state or federal governments. However, at mid-century the US Congress passed a series of housing laws that put an 'urban renewal' program into place, a combination of public housing subsidies and support for slum clearance and redevelopment. The result was to usher in a new functional role for city governments as partners with the federal government in administering a variety of national programs designed to address the cities' physical and social deterioration. Federal resources raised the stakes of urban politics.

This partnership lasted in full force until the election of XXX to the presidency in 1980. For three decades city governments planned the specific implementation of general federal programs and often provided matching funds in areas that ranged from housing and community development to policing, and from job training to the construction of sewage treatment plants (Gelfand 1975). In these years the city evolved from a self-financing and relatively autonomous level of local government to a major, though fiscally dependent, partner in a three-tiered (local, state, and national) federal arrangement. Urban politics in this age of federal partnership was often animated by disagreements not only over basic strategies of program implementation, but over where to locate certain federally funded facilities within the city (e.g., public housing projects) or who should control certain federal programs (the war on poverty programs in the 1960s and federal economic development empowerment zones in the 1990s are good examples). Some mayors eventually came to believe that as long as cities were held in the thrall of the federal government that their program priorities and spending decisions would always be distorted (Norquist 1998), but most relished the expansion of local capacity and responsibilities that federal programs underwrote.

With the substantial federal aid cuts that began in the Reagan presidency, cities entered a period of greater fiscal self-reliance. Although the primary function of city government remained the provision of basic local services, its role as a player in the federal system receded in favor of an emerging responsibility as steward of the local economy. This function called for the development of new public expertise in a range of financing, real estate, and entrepreneurial skills that were once the exclusive domain of business people. Cities established or greatly expanded their economic development departments, and city councils put into place an array of financing subsidies and business incentives to lure investors. Mayors became brokers of deals between developers and the financial community. Mayors and other local public officials led trade missions abroad to tout local industries and seek foreign investment, and cities established trade offices in other states as well as foreign capitals. It became common to speak of 'entrepreneurial cities' (Sbragia 1996).

4. *Conclusions*

The participants, preoccupations, and institutional setting of urban politics in the US give it a distinctive cast. Urban politics is not national politics written small. It is not party politics, nor is it shaped by regional conflicts. It is seldom animated by ideological disagreements, which may explain why Republican mayors have been easily elected in traditionally Democratic cities like New York and Los Angeles. Race and class divisions are common in the contests over development and the administration of public services that dominate public debate in big cities, although these issues are as likely to be fought out along other lines of cleavage such as newcomers versus old residents, downtown versus neighborhood actors, or corporate versus local business interests. Urban politics is an arena in which all sorts of crucial, if undramatic, decisions are made: the length and conditions of the daily commute to work, the level of security citizens feel on the streets, the quality and diversity of the venues that vie for people's leisure time and money, the availability of affordable housing, and the esthetic character of the built environment and urban open space. In all respects urban politics has become a major setting for collective debate over the quality of daily life for most people in America.

See also: Taste and Taste Culture; Urban Ethnography; Urban Geography; Urban Government:

Europe; Urban Growth Models; Urban History; Urban Policy: North America; Urban Poverty in Neighborhoods; Urban Sprawl

Bibliography

Banfield E C, Wilson J Q 1965 *City Politics*. Harvard University Press and the MIT Press, Cambridge, MA

Erie S P 1988 *Rainbow's End: Irish-Americans and the Dilemmas of Urban Machine Politics, 1840–1985*. University of California Press, Berkeley, CA

Gelfand M I 1975 *A Nation of Cities*. Oxford University Press, New York

Goldsmith S 1999 *The Twenty-First Century City*. Rowman and Littlefield, Lanham, MD

Grimshaw W J 1992 *Bitter Fruit: Black Politics and the Chicago Machine, 1931–1991*. University of Chicago Press, Chicago

Kleppner P 1985 *Chicago Divided: The Making of a Black Mayor*. Northern Illinois University Press, DeKalb, IL

Logan J R, Molotch H L 1987 *Urban Fortunes*. University of California Press, Berkeley, CA

Norquist J O 1998 *The Wealth of Cities*. Addison-Wesley, Reading, MA

Sbragia, A M 1996 *Debt Wish: Entrepreneurial Cities, US Federalism, and Economic Development*. University of Pittsburgh Press, Pittsburgh, PA

Sonenshein R J 1993 *Politics in Black and White: Race and Power in Los Angeles*. Princeton University Press, Princeton, NJ

Sugrue T J 1996 *The Origins of the Urban Crisis*. Princeton University Press, Princeton, NJ

P. Eisinger

Urban Poverty in Neighborhoods

Opinions differ regarding what causes poverty in America and why it seems to be concentrated in certain areas. Regardless of these differences, there is a relatively new focus on urban poverty and an agreement that the number of persons in poverty is growing as is the concentration of poverty in urban neighborhoods.

1. *Poverty in America*

National interest in poverty was first triggered when John F. Kennedy, while running for president, was taken on a tour of poor areas in the rural south. What he and his entourage saw was staggering: children with bloated bellies, families living in tar paper shacks, high rates of unemployment, and illiteracy. In short, the other America—families living in a state of poverty, the likes of which most Americans had never seen. This experience so affected the young presidential aspirant that he vowed, if elected, his administration would do everything within its power to address issues of poverty in America.

During the 1960s, politics and the publication of Michael Harrington's *The Other America* (1963), helped to focus the nation's attention on the plight of the poor in urban areas. President Johnson, taking office after Kennedy's assassination, launched the 'War on Poverty' in an attempt to have the federal government significantly address problems of poverty. However, due to the Vietnam conflict this effort was short lived.

It wasn't until the publication of James Wilson's *The Truly Disadvantaged* (1987) nearly two decades later, that interest in poverty was rekindled. Wilson's book sparked a flurry of research projects seeking to provide answers to such questions as: how widespread is poverty in America, what causes poverty, where is poverty most severe, what types of individuals and/or households are likely to wind up in poverty, what does being raised in poverty do to people, can people escape poverty, and perhaps most important, what can be done to reduce the incidence of poverty, and who should be responsible for addressing poverty issues? The results of most of these studies have revealed that poverty in America is not limited to rural areas. Every major city in the USA contains poverty neighborhoods and, in spite of growing economic prosperity, more Americans find themselves increasingly concentrated in poverty neighborhoods.

A person in poverty is defined by the United States Bureau of the Census as one whose total family income falls below a federally defined level. This level varies according to family size and is adjusted annually for inflation. In 2000, for example, the poverty threshold for a family of four was $17,463. The poverty rate for a particular neighborhood is determined by dividing the number of poor persons who live there by the total number of poor and nonpoor persons. Census tracts are used as the geographic basis for identifying the location of poverty neighborhoods. Researchers differ on the cut-off level for defining poverty neighborhoods. The range is from 20 to 40 percent, with the higher level being used most commonly. While different names have often been ascribed to poverty neighborhoods, Jargowsky (1997) refers to African-American high-poverty areas as ghettos, while barrios describe Latino high-poverty neighborhoods, and slums are poverty neighborhoods comprised primarily of whites.

2. *What Causes Poverty and Why Does it Seem to be Concentrated in Certain Neighborhoods?*

A number of different explanations for what causes poverty and why poverty appears to be concentrated have been put forth. Some of these include: discrimination and segregation, immigration of rural farm workers from other countries to cities in America,

the poor quality of inner-city schools, crime and the prevalence of drugs, white flight, and the emergence of a global economy.

2.1 The Creation of Ghettos and Barrios: The Role of Segregation

During the 1940s and 1950s, the USA underwent a rapid expansion of its industrial and manufacturing sectors. African-Americans seeking to improve their quality of life migrated in large numbers from the south to major urban growth centers throughout the USA. Segregated housing, restrictive covenants, and zoning practices forced many of these African-American migrants to seek shelter in older sections of cities. Segregation and discrimination also limited employment and advancement opportunities, which in turn impacted earning power. As a consequence, African-Americans were disproportionately concentrated in low-income and often poverty neighborhoods called ghettos (Drake and Cayton 1945, Clark 1965, Dubois 1967). Even after the passage of Title VI of the Civil Rights Act of 1964 and Title VIII of the Civil Rights Act of 1968, real estate agents, mortgage brokers and lenders continued to make it difficult for African-Americans to acquire housing outside of inner city areas. As a consequence, African-Americans, unlike their white counterparts, have much lower home ownership rates, one of the key ways Americans accumulate wealth (Bullard et al. 1994, Oliver and Shapiro 1995). Massey and Denton are two scholars who maintain that racial segregation and not economic segregation have played a major role in the continued growth of ghetto poverty (Massey and Denton 1993).

During the 1970s, 1980s and 1990s, a large number of immigrants from Latin America, in particular from Mexico, have come to the USA. To a certain extent, these immigrants find themselves relegated to poverty areas or barrios, much like African-Americans did before them. More recently, immigrants from various Asian countries, particularly Vietnam, Cambodia, and Laos have been moving to American cities. In 1990, foreign-born residents constituted 10 percent of all those living in concentrated poverty areas, up from 3.5 percent two decades earlier (Mincy and Wiener 1993, Table 6). As of yet, no single term has emerged to describe Asian poverty neighborhoods. Although references to China Towns, Little Tokyos, and Little Vietnams often imply similar conditions to ghettos and barrios. Like African-Americans, the new immigrants also find that their poverty neighborhoods suffer from inferior schools, high crime rates, inadequate housing, and limited employment opportunities. Language barriers and low educational achievement seem to be key factors contributing to the concentration of these new immigrants in poverty neighborhoods.

While researchers have easily identified poverty neighborhoods with a predominance of racial minorities, that is ghettos and barrios, few predominately white poverty neighborhoods or slums are evident in metropolitan areas. This is not to say that there are no whites in poverty; quite the contrary, the largest number of individuals in poverty are whites. It is just that poor whites are distributed throughout cities and not particularly concentrated in certain neighborhoods.

2.2 Social Mobility

Social mobility, or as Wilson (1987) calls it, "neighborhood sorting," is another reason often put forward to help explain the growing concentration of urban poverty. There are two dimensions of this concept. The first focuses on white flight; that is, as more and more minorities move into urban areas, better educated and more affluent whites move to the suburbs. Reasons most often given for this out-migration of whites from central city areas include: a desire to provide better educational opportunities for their children, safer neighborhoods, better and cheaper services, more neighborhood amenities, and the ability to realize greater return on their investment in housing.

These same interests are propelling the move of middle-class African-Americans from inner city areas, leaving poor African-Americans behind. The consequence is, further concentration of poor people in central city poverty neighborhoods (Wilson 1987).

2.3 The Impact of the Global Economy

More recent literature has focused on what role deindustrialization, brought on by a growing global economy, has had on both increasing the number of persons in poverty and the growing spatial concentration of poverty (Bluestone and Harrison 1982). The growth of new jobs is occurring faster in suburban areas than in inner-city areas. This makes it increasingly difficult for residents of poverty neighborhoods to acquire employment. Kain (1968, 1992) calls this the spatial mismatch phenomena.

Rapid job growth in the service sector has reduced the impact of some losses of industrial employment. However, these jobs tend to pay relatively low wages with few, if any, benefits. This often means that, even though individuals residing in poverty neighborhoods can secure work, the jobs they obtain do not provide adequate income or sufficient benefits to lift them or their families out of poverty. Thus, there are a growing number of working poor residing in poverty neighborhoods (Pastor et al. 2000 and Berstein et al. 2000).

2.4 The Significance of Regional Economic Growth and Poverty Neighborhoods

Some of the more recent literature suggests that there is a link between the size and growth of poverty

neighborhoods and the robustness of a region's economy. Jargowsky, for example, shows that a rise in overall metropolitan mean income resulted in sharp declines in ghetto poverty concentration among African-Americans (Jargowsky 1997). Similarly, Pastor et al. found in 74 regions throughout the USA that regions which performed best economically were also ones that either reduced inner-city poverty or reduced the rate of growth of poverty neighborhoods (Pastor et al. 2000).

To date, the most comprehensive work examining the causes of poverty can be found in *Poverty and Place: Ghettos, Barrios, and the American City* by Paul Jargowsky. Unlike previous studies which are locale specific, this study examines national poverty data. Jargowsky finds that structural changes in the economy have in fact adversely affected those with lower educational levels and job skills by reducing their wages. He also finds that deconcentration of employment further reduces the possibility of these individuals finding work and that the flight of higher-income individuals from poverty neighborhoods is an important factor in increasing neighborhood poverty. However, he could not find any evidence that strongly supports the notion that segregation plays a direct role in the recent growth of poverty neighborhoods. Jargowsky concludes by suggesting that neighborhood poverty is the

> predictable result of the economic status of minority communities and the degree to which minorities are residentially segregated from whites and from each other by income (Jargowsky ,1997 p. 19).

3. *Why is it Important to be Concerned About Urban Poverty?*

One of the reasons for the concern with urban poverty is that place and neighborhoods matter. Therefore what happens to people living in these neighborhoods is of concern to policy-makers. Since the majority of residents of poverty neighborhoods are children often in female-headed households, they are of particular concern. What happens to children residing in poverty neighborhoods is the subject of considerable debate. There is ample evidence, for example, that many of these children are not adequately nourished, are in need of more regular medical care, are not properly supervised, and are constantly exposed to drugs and criminal acts. But there is no consensus on what should be done to improve the quality of life for these children, let alone who should do it, and how it should be financed. At the same time, most urban dwellers believe that something needs to be done. Passage of the 1996 Welfare Reform Act may have a significant impact on many of these households as mothers are required to transfer from public assistance to work.

To a lesser extent, some of these concerns apply to the elderly who are often trapped in poverty neighborhoods. Even though municipalities, for the most part, have been conscientious in providing some services to the elderly (most notably transit, housing, and nutrition), environments in which these elderly reside, by most accounts, can be hazardous.

A second reason for concern about poverty neighborhoods revolves around the question of equity, particularly in the provision of public services. Poverty neighborhoods contain the oldest housing stock most in need of repair, the most deteriorated infrastructure, have the highest levels of toxic waste concentrations, and the least amount of open space, particularly in comparison with more affluent neighborhoods. Given these conditions, how should public resources be allocated to address these problems?

A third reason for concern is that poverty neighborhoods are believed to be the cause of a number of undesirable social conditions, high crime rates, family disintegration, incidences of major health problems including high infant mortality, tuberculosis, HIV/AIDS, alcoholism, and other antisocial behavior. These conditions if left unchecked could spread to other neighborhoods.

Finally, the concern with poverty neighborhoods centers around the widely held belief that these neighborhoods adversely impact other neighborhoods. Here the concern is often with crime and violence, the impact on property values, and how much of a tax drain poverty neighborhoods cause for other areas of the city. There is also the concern, as was evidenced in Los Angeles in 1992, that intolerable conditions can lead to civil unrest which has proven to be extremely costly (Orfield 1997, Pastor 1993).

4. *Conclusion*

After nearly four decades of research and efforts at policy reform and program intervention, one thing is abundantly clear—the number of persons in poverty is increasing and persons in poverty are becoming more and more concentrated. Between 1970 and 1990, the number of poor people in high-poverty neighborhoods nearly doubled, from 1.9 million to 3.7 million. This rate of growth exceeded the overall rate of increase in the total population in poverty neighborhoods during this same time period, in spite of the fact that during the 1980s the nation experienced good economic times. Most perplexing is the growing poverty rate among households where one or both adults are working. A robust economy might have provided jobs but not at wage levels that moved people out of poverty.

What is wrong? One might conclude that problems of poverty cannot be solved. Or one might suggest that those who are responsible for providing policy leadership on this issue simply have not found the right solution yet. More disturbing, however, is the thought

that the vast number of Americans not in poverty simply do not care much about those who are.

Clearly, there is no simplistic solution to the question of poverty. Indeed this may be one of the problems. Politicians in their zeal to provide solutions respond to constituents or press accounts which usually focus on symptoms or sensational exposes. Researchers tend to focus on individual policy issues. What is needed, perhaps, is a more comprehensive policy approach which takes into account the structural sources of poverty, its multigenerational roots, and the link between individual neighborhoods and regions.

See also: Community Sociology; Neighborhood: General; Neighborhood Revitalization and Community Development; Poverty, Culture of; Poverty in History; Poverty Policy; Poverty, Sociology of; Prostitution; Residential Segregation: Geographic Aspects; Residential Segregation: Sociological Aspects; Underclass; Urban Geography; Urban Sociology

Bibliography

Banfield E 1970 *The Un-Heavenly City*. Little, Brown and Company, Boston

Bernstein J, McNichol E C, Lawrence M, Zahradnik R 2000 *Pulling Apart A State-by-State Analysis of Income Trends*. Center on Budget And Policy Priorities Economic Policy Institute, Washington, DC

Bluestone B, Harrison B 1982 *The Deindustrialization of America: Plant Closing, Community Abandonment, and the Dismantling of Basic Industry*. Basic Books, New York

Boger J C, Wegner J W (eds.) 1996 *Race Poverty and American Cities*. The University of North Carolina Press, Charlotte, NC

Boston T D, Ross C L (eds.) 1997 *The Inner City Urban Poverty and Economic Development in the Next Century*. Transaction Publishers, New Brunswick, NJ

Bullard R D, Grigsby J E, Lee C 1994 *Residential Apartheid: The American Legacy*. CAAS Urban Policy Series Vol. 2, University of California Press, LA

Clark K B 1965 *Dark Ghetto: Dilemmas of Social Power*. Harper & Row, New York

Denton N A, Massey D S 1998 Residential segregation of blacks, Hispanic, and Asians by socioeconomic status and generation. *Social Science Quarterly* **69**: 797–817

Drake S C, Cayton 1945 *Horace R. Black, Metropolis: A Study of Negro Life in a Northern City*, Vol. 1. Harper Tourch Books, New York

Dubois W E 1967 *The Philadelphia Negro: A Social Study*. Schocken Books, New York

Harrington M 1963 *The Other America Poverty in the United States*. Penguin, New York

Jencks C 1993 *Rethinking Social Policy: Race, Poverty, and The Underclass*. Harper Perennial, New York Reality. Working Paper, The Urban Institute, Washington, DC. Unpublished manuscript

Jargowsky P A 1977 *Poverty and Place Ghettos, Barrios, and the American City*. Russell Sage Foundation, New York

Kain J E 1968 Housing segregation, Negro employment, and metropolitan decentralization. *Quarterly Journal of Economics* **82**: 175–97

Kain J E 1992 The spatial mismatch hypothesis: three decades later. *Housing Policy Debate* **3**: 371–460

Massey D S, Denton N A 1993 *American Apartheid Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA

Mincy R, Wiener S 1993 *The Under Class in the 1980s: Changing Concept*. Constant Reality, Working Paper, The Urban Institute, Washington, DC. unpublished

Oliver M L, Shapiro T M 1995 *Black Wealth White Wealth: A New Perspective on Racial Inequality*. Routledge, New York

Ong P (ed.) 1994 *The State of Asian Pacific America: Economic Diversity, Issues & Policies*. LEAP Asian Pacific American Public Policy Institute and UCLA Asian American Studies Center, Los Angeles

Orfield M 1997 *Metropolitics: A Regional Agenda For Community and Stability*. Brookings Institute Press, Washington, DC

Pastor M Jr 1993 *Latinos and the Los Angeles Uprising: The Economic Context*. The Thomas Rivera Center, Claremont, CA

Pastor M Jr, Drier P, Grigsby J E III, Garza-Lopez M 2000 *Regions that work. How Cities and Suburbs can Grow Together*. University of Minnesota Press, Minneapolis, MN

Sunquist J L (ed.) 1969 *On Fighting Poverty*. Basic Books, Inc., New York

Wilson W J 1987 *The Truly Disadvantaged: The Inner City, The Underclass, and Public Policy*. The University of Chicago Press, Chicago

J. E. Grigsby III

Urban Sociology

Urban sociology concerns itself with the social and cultural forms assumed by the urban phenomenon in the past and in the present. It endeavors to understand the worldviews of the various cultures that have produced cities, and the coherence or incoherence with which these worldviews have been given concrete form.

The city is the entity to which the Greek philosopher Aristotle attributed a capacity 'to suffice unto itself.' He argued in *Politics* that 'the perfection of existence' was possible in the city and he concluded that 'man is a political animal created to live in cities.' The history of cities displays wide typological differentiation: cities vary according to the manner in which they began and subsequently developed, and according to their geographical and cultural area. But there seems to be one feature shared by them all: a capacity to dominate and organize the space surrounding them by virtue of their concentration of power in its various forms—political and administrative, military, economic, cultural, scientific, and religious. There have indeed been societies and cultures without cities as their political, administrative, economic, and cultural points of reference, or which possessed them only to a limited extent—consider the Europe of the High Middle Ages—yet it is nevertheless true that the

organization and development of systematic control over a territory has always centred on a capital city, and therefore on a hierarchy of cities and settlements.

1. The City in History

Urban sociology seeks not only to describe and explain the rich variety of the phenomenon 'city,' but also to interpret the causal connections between the elements that constitute a city and the factors that first gave rise to them. This approach furnishes understanding of the complex as well as profound meaning of every urban reality. To study the history of cities is to study the history of civilizations, of societies, and of cultures: a far-reaching enterprise which involves numerous disciplines and is constantly enriched with the results of scientific research.

One might enquire whether urban sociology is able to identify a single interpretative theme linking what by now constitutes a vast complex of urban phenomena. A simple list of the best-known urban forms demonstrates the difficulty of the undertaking. The history of the city, in fact, displays a rich variety of forms, which may be summarily listed as follows: the Sumerian cities; those of the Mesopotamian kingdoms; the Greek *polis*; the Phoenician cities; Rome and the cities of the Eastern and Western Roman Empire; the cities of the Asiatic empires; the cities of the Central American cultures; the Muslim cities; the city communes of the Middle Ages; the cities of the seigniories; those of the European sovereigns; the cities of the industrial revolution; the cities of Africa, Latin America, and Asia that arose as a result of European colonization; the metropolises of the twentieth century, and those that have now become city region and global cities; as well as the metropolises of the Third World. Finding an interpretative theoretical paradigm that comprises this huge variety of urban forms is an arduous undertaking, unless one resorts to a set of generic statements which risk lapsing into triviality.

One approach might be to single out a set of shared features and then flank them with those specific to each individual form of urban society and culture. As will be shown, every urban form, in its birth and its development, is geared to the production and distribution of goods, and to the management of the surpluses that engender the city as the hub of functions, services, and political power, administrative and organizational. The city also develops and disseminates patterns of urban planning and organization, models for the construction of supply and service infrastructures. It is the engine of technological innovation and knowledge, the center of political, religious, and military activity. These features, however, vary considerably across the manifold historical forms of the city, and it would be a gross oversimplification to impose a single sociological framework on such diversity. The general considerations to be drawn from comparative sociological analysis take the form of a social-urban philosophy, albeit one grounded on analysis of specific sociological urban types. Examples are the (important) assertions that can be made concerning the relationship between the forms assumed by the city and freedom, democracy, and the reciprocal exchange of urban political power and other cultures.

2. The Territorial Stabilization of Social Life

The city was at first a permanent settlement of a population on a particular territory, a settlement which developed specific forms and functions. The advent of humankind began the process known as *civilization* (from the Latin noun for 'city,' *civitas*) with the birth of the first settlements devoted to farming and cattle-raising but also to a sacred element: cemeteries and shrines. 'The city of the dead antedates the city of the living. In one sense, indeed, the city of the dead is the forerunner, almost the core, of every living city' (Mumford 1961, p. 7). The villages of the Mesolithic age surrounded themselves with palisades, they 'specialized' space, and they organized themselves into a hierarchy, with a political, economic, and social system that the city would later develop and strengthen.

3. The Rise of a Space Symbol System and Culture

It was this revolution that gave origin to space as a symbolic system. By choosing to live in settlements, humankind opted for a form of social life that gave priority to social features expressed through increasingly rich forms of communication: language, tools, communal meeting-places, the recognition of a leader, obedience to the rules of the community, places and forms of worship. Language and the home were the distinctive features of the Palaeolithic village, while the inhabitants of the Neolithic village clung to their burial places and their vital significance for the social group, while continuing to develop agriculture and cattle raising.

The origin and evolution of human settlements produced the sociological meaning of the city through the stages outlined in the following sections.

4. Antiquity and Urban Revolution in the Near East

The origin and evolution of human settlements with their diverse geographical and temporal features were replicated in the birth and rise of the cities. At the end

of the last Ice Age, the alluvial plains of the Nile, Tigris, Euphrates, and Indus yielded abundant crops for the peoples who moved in from the South and organized themselves as farmers. Surplus produce was used to maintain a group of artisans, scribes, functionaries, priests, and soldiers who introduced the spatial distinction between city and countryside. The first known urban conglomeration arose five thousand years ago (Clark 1977).

Space organization was connected with structured social functions: the beginning of the ambivalent relation between the city and its surrounding territories and settlements. The city became the locus of power: the place where the allocation of surplus goods was established, where reliable communication over long distances was made possible by the written word, where construction techniques were developed, where stable patterns of behavior were formed, and where the law was formalized (the fundamental law and rules regulating everyday life which supplanted the family and kinship bonds). An entirely new social and spatial dimension arose, which opened the way for a more formal and objective legal order, one legitimated by the urban community and which, in some cases, acquired the form of a state. The production of material equipment was enormously increased by the advent of models fixed by writing and registration. Thus, culture as a set of symbols produced and reinforced technology, exerting a powerful influence over the material production of goods which, in the absence of that culture, would only have developed sparsely and slowly. City and civilization, therefore, were closely bound up with each other in their birth and development.

5. *The Development of a Governmental Hierarchy*

The city came to reproduce, in historically different ways, the structural elements of the first human settlements. Various factors induced settlements to grow into cities, but the most notable of them were the centralization of the population, the advent of market, the court and administration, the temple, science (writings), and the army.

The clay tablets unearthed during excavations of the Sumerian cities (Ur, Erech, Lagash) document the corporative government exercised by priests who considered themselves the stewards of land entrusted to them by God. This belief was reproduced in the physical form of the city dominated by the temple, with its dwellings, storehouses, and workshops, and around which the city articulated itself according to social and spatial differentiation. When the Sumerian cities came under the sway of a king with sacred status and power, they manifested a power and an administrative organization arranged in a hierarchy between the capital city and those of lesser status in authority and function. Regulating this hierarchy of cities and localities was a set of laws that ensured the consignment of surplus or appropriated goods to the more central cities, and then developed into a system of tax collection and military organization which produced a veritable state. The Egyptian city pivoted on the power of the god-like Pharaoh, who wielded absolute control over the economy, which he utilized to build temples, pyramids, and tombs that raised him above common mortals. In the second millennium BC, the cities of the Far East developed according to a similar hierarchical pattern whereby the provincial cities were dominated by the capital.

The Greek city was the template for Western urbanism. Born of an alliance among communities and their kings (synoecism), from the sixth to the fourth centuries BC the *polis* became a reality in which the Greeks developed a culture devoted to reproducing in the city a vision of the world whose central component was not absolute power and the conquest of neighboring territories, but rather human harmony, body and intellect expressed, both physically and spiritually, in literature and the figurative arts, and also in the social life reproduced concretely and organizationally in the *polis*. The term 'city' thus came to coincide with 'politics,' a city organized as a city-state, the place of orderly and individual freedom. The Greek *polis* of this period was a city whose walls protected the country people when attacked and which granted freedom to its citizens. The *polis* carefully regulated its size (surplus population moved away to found new colonies), the purpose being also to preserve harmony and to protect the system of self-government constituted by the *prytaneum*, the place where the community's divine protector was worshiped, the *agora* or citizens' assembly, and the *bulè*, the council of nobles and representatives of the assembly.

6. *Roman and Greek Urban Development*

The Greek city expressed in admirable harmony the relationship between the city and its rural surroundings, but it was a relationship that was always ambivalent. On the one hand, the city comprised a complex set of powers which subjugated the surrounding territory; on the other, it developed a political, legal, economic, and organizational order, and a network of services, which incorporated the smaller communities into a system of which the city was the center and the administrative seat. The Alexandrine city that arose from Alexander the Great's expansionary policies and conquests was symmetrical in form, ordered according to a Euclidean geometry which ensured control and the orderly deployment of troops; a cultivated, refined and monumental city. The concentration of powers in Alexandria gave it decisive dominance over nonurban territory (Mumford 1961).

Rome, the capital city that ruled the territories around the Mediterranean basin, invested its Senate,

consuls, praetors, aediles, tribunes, and quaestors with powers that produced one of the most stable empires of history. As *caput mundi*, the headquarters of the state, Rome undertook the division of urban and agricultural territory, the founding of new cities, the building of aqueducts, sewers, basilicas, baths, amphitheaters and circuses, and the construction of an immense network of roads and military outposts. Besides being the capital of the state and the empire, it was the locus in which political powers concentrated and evolved, and where the urban models of the empire's cities were developed (e.g., Th. Mommsen). Rome established an organic, almost biological, relationship with the world of the empire; but it was a relationship which, amid the city's estimable political, administrative, legal, organizational, and military system, permitted no form of political, and especially cultural, reciprocity. When Rome's powerful inner impetus subsided, its grand design imploded because it was unable to draw on resources arising from interaction with other cultures regarded as equals. One single legal order was admitted, and no other order was allowed that might introduce innovations.

7. *The Disappearance of the Ancient City*

The barbarian invasions destroyed the political and organizational structure of the Western Roman Empire, while the empire in the East fell to the advancing Muslim Turks. The urban structure of the Roman empire was replaced in the West by the Christian territorial organization based on large rural landholdings, while the cities and the road network lapsed into decay through disuse. In the East, the Muslim cities represented an extension of the power wielded by princes and merchant corporations: without specific autonomies, they served to protect citizens and to safeguard trade routes and the agricultural market, while for other commodities they added further urban components.

8. *European Cities in the Middle Ages*

The medieval city in Europe was the result of revival in the economic and commercial development of the coastal cities of the Mediterranean and northern Europe, and the resumption of agriculture and artisan industry. The cities were revivified and became the centers of corporations, guilds, and confraternities, while their citizens developed a legal order different and distinct from that of the medieval lordship. This was a 'nonlegitimate and revolutionary' legal order (Weber [1921] 1958) endowed with city walls, its own civil law (which regulated relations among citizens according to rules established independently), its own lawcourts, and an armed corps of citizens called upon to defend the new order; all of which were guaranteed by the oath-swearing (*conjuratio*) of those who perceived themselves as brothers in a common enterprise. The medieval city, with its public spaces for artisan, commercial, religious, and social activity, assumed a complex urban form: the social configuration of the new notion of an autonomous community, with a political and legal system that gave birth to the constitutional and civil law of the modern state. The medieval structure of many European cities resisted the assaults of the lords and princes of the Renaissance, and those of the seventeenth- and eighteenth-century kings. The latter decided to forgo their endeavor to change the city, and instead set about building royal palaces with parks and estates outside the cities (Versailles, Schönnbrunn, Aranjuez, Caserta). This separation between royal-noble power and the urban classes made up of shopkeepers and professionals was the prelude to, and ground for, the bourgeois revolution.

9. *The Modern City*

The bourgeois and industrial revolutions engendered the subsequent great urban revolution. Population growth, immigration from the countryside, the more articulated organization of the city, heightened demands for comfort: these factors upset the balance between traditional and new urban space. Flanking such ecological problems were the inequalities among the urban classes (bourgeoisie and proletariat) that socialist (e.g., F.-N. Babeuf, C. Fourier) and liberal theory analyzed, and which the urban policies of the constitutional states addressed, but only by creating public and social spaces and services, while giving free rein to speculation and the market under rules that applied to all the social groups admitted to the competition for social ascent.

Sociology arose in concomitance with the problems that attended the birth and growth of industrial society, and it analyzed the urban phenomena that embodied the urban revolution. Marx conducted his critique of capital by starting from the radical inequality between the owners of the means of production and an exploited proletariat forced to sell its labor in the city called Coketown. Max Weber sought to understand the birth and crisis of capitalism by studying the medieval European city as the origin of the free economic enterprise that gave rise to the new and free legal order; the city that gave liberty to all those who decided to join it, and whose motto was '*Die Stadtluft macht frei*' (The city air makes us free).

However, although this city produced a political and legal system by which citizens were free under the state and the modern legal order, it was not a city of equals: it may have been made up of formally equal citizens but it was marked by frequently profound social and spatial disparities.

The Chicago School analyzed the phenomena and problems of a modern industrial city under pressure from multiethnic immigration and with spatial and

social segregation according to income, ethnicity, and profession (Park et al. 1925), its conviction being that urban dynamics were similar to those of a biological organism. The modern city has developed this segregation into the new and dramatic form exemplified by the cities of the Third World. These urban conglomerations are unable to drive development because they consist of enormous concentrations of immigrants in search of sustenance and with little hope of finding stable and decent employment. They are cities, with their almost boundless demographic and spatial expansion, that lack ancient centers of historical importance with which their inhabitants can identify. The modern installations of the cities that have developed in the ex-colonial countries, their business districts and quarters restricted to the middle and upper classes, have forced their poor and marginalized inhabitants into the periphery.

The metropolises of the European and American West have become the world centers of political, economic, and financial power, of services and the organization of production. Extending as they do across the territory of several states, and incorporating the entirety of politics, science, technology, business, and leisure, for some time they have constituted the symbol, and become the Utopia, of the 'new order of ages' described by Jean Gottmann (1961) in *Megalopolis*, which suggests that the world is controlled and receives its meaning from an urban culture which acts as the engine of all development, material, social, and cultural.

10. The World as a City System

The contemporary city has burst through the boundaries that used to separate it from the countryside and surrounding space. It pervades them as does Gottmann's megalopolis, or it penetrates them and binds them with a network of communication and transport structures which do not stand in a reciprocal relation with the countryside and external space; on the contrary, it is thus that the city asserts its ramified supremacy. On this view, the city determines every type of production and the ways in which surpluses are utilized; it decides who should receive economic rewards and professional recognition; it establishes which spaces are to be allocated to services and which should be reserved for different social classes, according to a meritocratic hierarchy based on a scale of values determined by the elites of the city itself.

11. Urban and Regional Sociology

Cities do not change, even less do they dissolve in the regional dimension. On the contrary, the latter demonstrates that the urban dimension of power has not freed itself from the restrictions imposed by city walls, by administrative or jurisdictional boundaries, by the attraction of a market spatially identified with an area called 'city.' Finance and the organization of production and work do not rely solely on the automation and advanced technologies that require increasingly skilled workforces. One feature of the 'global city,' in fact, is its indifference to the spatial allocation of the production process. The increased production of nonmaterial goods and information technologies has led to the greater use of teleworking, which enables closer and more efficient monitoring of the quantity and quality of output.

The city continues to control the production of services and access to them, and has thus become dominated by tertiary activities. It ensures and refines its supremacy over increasingly broader territories, using instruments of ever greater sophistication but of decreasing visibility to public opinion. The city, in fact, has developed a new technological and professional perspective: that of communications, which is now one of the most highly developed and influential of production sectors, able to construct images which convey lifestyles, symbols of prestige and gratification, and, indeed, visions of the world. The global city displays the features of a market on the world scale which nation-states and international bodies are unable to regulate with effective policies. The city, precisely because it represents and coincides with the power centers of the global market, also produces the symbols and mechanisms that legitimate this planetary power; symbols and mechanisms of legitimation which the legal and political orders of states are unable to counteract or to replace.

See also: Cities: Capital, Global, and World; Cities: Cultural Types; Urban Anthropology; Urban Growth Models; Urban History; Urban Studies: Overview

Bibliography

Clark G 1977 *World Prehistory in a New Perspective*, 3rd edn. Cambridge University Press, Cambridge, UK

Dickinson R E 1962 *The West European City: A Geographical Interpretation*, 2nd rev. edn. Routledge, London

Duncan O D, Reiss A J 1956 *Social Characteristics of Urban and Rural Communities*. Wiley, New York

Dyos H J, Wolff M 1973 *The Victorian City: Images and Reality*. Routledge, London

Eade J (ed.) 1997 *Living the Global City: Globalization as Local Process*. Routledge, London

Fustel De Coulanges N D [1864] 1956 *The Ancient City: A Study on the Religion, Laws, and Institutions of Greece and Rome*. Doubleday, Garden City, NY

Geddes P 1968 *Cities in Evolution: An Introduction to the Town Planning Movement and to the Study of Civics*, 3rd edn. Benn, London

Gottmann J 1961 *Megalopolis: The Urbanized Northeastern Seabord of the United States*. MIT Press, Cambridge, MA

Harloe M (ed.) 1996 *The Sociology of Urban Communities*. Edward Elgar, Cheltenham, UK

Jacobs J 1961 *The Death and Life of Great American Cities*. Random House, New York
Marcuse P, Van Kempen R 2000 *Globalizing Cities: A New Spatial Order*. Blackwell, Oxford
Mommsen Th 1854–1856 *Römische geschichte*, 3rd edn. Weidmann, Berlin
Mumford L 1938 *The Culture of Cities*. Harcourt, New York
Mumford L 1961 *The City in History: Its Origins, its Transformations and its Prospects*. Harcourt, New York
Park R E, Burgess E W, McKenzie R D 1925 *The City*. University of Chicago Press, Chicago
Scaglia A 2001 *Max Webers Idealtypus der nichtlegitimen Herrschaft. Von der revolutionär-autonomen mittelalterlichten Stadt zur undemokratischen Stadt der Gegenwast*. Leske & Budrich, Opladen, Germany
Sjoberg G 1960 *The Preindustrial City: Past and Present*. Free Press, Glencoe, IL
Weber M [1921] 1958 *The City*. Free Press, Glencoe, IL

A. Scaglia

Urban Sprawl

Urban sprawl is an imprecise but highly evocative, usually derogatory, term used primarily in the economically advanced countries of the English-speaking world to describe certain kinds of low-density urban landscapes with little comprehensive public planning. The chief arguments against this kind of settlement pattern are that it is inefficient, inequitable, environmentally damaging, socially irresponsible and aesthetically ugly. To halt sprawl, reformers have called for more highly planned, compact, transit-oriented cities. Among the remedies proposed to stop sprawl are political changes including the move of planning functions upward in the governmental hierarchy, the removal of subsidies supposedly supporting sprawl, the implementation of a system of incentives intended to create higher densities, growth restrictions, growth boundaries, greenbelts, open space conservation and historic preservation. Although the term was used by planners and other experts for at least a half century, it became a major public concern only in the economic boom period of the last years of the twentieth century. At the same time, however, it became clear that even the most basic assumptions about the definition of urban sprawl, its causes, and the effect of various proposed remedies were in dispute.

1. *History of the Term*

Described by various words over the years, the kinds of condition the term 'sprawl' would later refer to have almost certainly been matters of concern as long as towns have existed. By the nineteenth century, certainly, complaints about buildings straggling out from towns and villas marring picturesque landscapes were quite common in Britain. These complaints seem to have entered the political arena in a major way following the great prosperity of the 1920s. One of the most important critics was planner Thomas Sharp, who issued vitriolic attacks on the new 'semi-suburbia,' on 'ribbon developments' along roads leading out of town, and on suburban villas marring the countryside (Sharp 1932). His cause was taken up by groups such as the Council for the Preservation of Rural England with its heavy constituency of large landowners, and several bills passed in the 1930s aimed at restricting ribbon development and creating greenbelts around cities.

It appears to have been during the 1930s that the use of the word sprawl as a noun with something like its current sense came into general usage. One prominent use of the term occurred in the famous Greater London Plan of 1944 (Abercrombie 1945). It was also used on the other side of the Atlantic by Lewis Mumford in *The Culture of Cities* (1938). The original criticisms of sprawl were almost always based on aesthetic and symbolic considerations. It was ugly, and blurred the traditional distinctions between the city and the countryside (Bruegmann 2000).

Complaints about sprawl have continued to accompany every era of major prosperity since World War II, but they have changed considerably in content and have been modified depending on the time and place. Immediately after World War II, sprawl became less of an issue in Britain because it, like many continental European countries, had been exhausted economically by the war and at the same time had adapted the kinds of stringent controls on peripheral development around large cities that planners had advocated in wartime reconstruction plans. In the United States, on the other hand, the prosperity of the postwar years without any strong growth control planning led to a tremendous suburban boom. Consequently the most conspicuous negative reactions to sprawl came from the United States. At the same time, the focus started to expand to include objections based on notions of equity and quality of community. These paralleled closely the criticisms that were being made at that time by a number of prominent sociologists of suburbia in general. According to this analysis, the suburbs were conformist, alienating, homogenous, and boring (Riesman 1950, Whyte 1956).

The most conspicuous use of the term sprawl in the postwar decades came from William H. Whyte, who organized a conference for *Fortune Magazine* on the subject and then published a jeremiad against sprawl in *The Exploding Metropolis*, a book that he edited (Whyte 1958). For Whyte, as for the British, sprawl was primarily an aesthetic problem, but he made scattered comments that suggested that sprawl might also be inefficient economically and have negative social repercussions. These were arguments that would be expanded, starting in the 1960s, and combined with

new elements from the environmental movement. Whyte's essay also started what would become a common practice of using the Los Angeles metropolitan area as a prime example of low-density sprawl.

As the various strands of the sprawl argument were worked out and debated in the following decades, the term became more common but it never acquired a clear definition. By the late twentieth century, the term could mean many different things to different people using it. Almost everyone agreed that it referred to low-density development, but low density in a European city could still be denser than relatively high density in a city in the United States. Sprawl commonly referred to settlement that was scattered, discontinuous, leap-frog, or ribbon, but as many observers pointed out, with additional development the low-density discontinuous settlement of one era often becomes the highly admired, compact, and high-density settlement of the next (Lessinger 1962). Sprawl could also refer to automobile-dependent development or to development with poor access from home to work or to public open spaces, but many other observers have pointed out that, as cities have become wealthier, they have almost inevitably seen more automobile usage and, in many cases, commuting times have remained the same or fallen. The state of the term at the turn of the century can perhaps best be seen in a pair of articles published in the *Journal of the American Planning Association* in 1997, one by Harry Gordon and Peter Richardson, who challenged the idea that sprawl was a major problem; and the other by Reid Ewing who maintained, along with most mainstream planning thought, that it was indeed a problem. In some ways, the most remarkable part of the debate was not the conclusions about the problems and the proposed remedies but the lack of agreement about what sprawl was. The two essays, with the ensuing letters, made it clear, for example, that even as basic an issue as whether Los Angeles was a high-density, relatively compact city or a low-density sprawling city was in dispute (Ewing 1997, Gordon and Richardson 1997).

By the late 1990s, important coalitions to counter scattered, low-density, automobile-oriented growth at the periphery of urban areas were in evidence in virtually every affluent nation in the world, although the term sprawl was mostly used in the United States, Britain, Canada, and Australia. In the United States, the term 'smart growth' became important in the 1990s to designate efforts to combat sprawl. In England, Canada, and Australia, 'the compact city' tended to be the term used to refer to the opposite of sprawl. In all of these places, sprawl also tended to be contrasted with 'sustainable' development, another term with a large literature and extremely divergent definitions (Jenks et al. 1996). In continental Europe, in the face of a vast increase in automobile ownership, suburban development, and peripheral shopping centers, the European Economic Community declared itself in favor of compact, transit-friendly cities with single vibrant centers (Commission for the European Community [CEC] 1990, Nivola 1999). In Japan, the term *supurôru* was coined to designate what English commentators called sprawl, although the target was more haphazard growth and assurance of adequate infrastructure than any specific desire to make development more compact (Hebbert 1994). The concept of sprawl has also been used in Israel and elsewhere (Razin 1998).

2. *Objections to Sprawl*

Objections raised by reformers to sprawl have changed dramatically over the years and according to the viewpoint of the observer. They can, however, be classified under five major headings that range from the apparently quite objective to the obviously very subjective. These categories might be designated efficiency, equity, environmental sustainability, social responsibility, and aesthetic and symbolic suitability.

2.1 *Efficiency Concerns*

Although the earliest and most constant objections to sprawl have been aesthetic, the debate about public policy to curb sprawl has primarily been conducted in more objective terms. The earliest and most extensive segment of this literature has targeted the alleged inefficiencies of sprawl. The most important early summary of these issues in the United States in the postwar decades appeared in the costs of sprawl issued by the Real Estate Research Corporation (1974). Using hypothetical models of various kinds of development, this report concluded that master-planned communities at fairly high densities were much more efficient in terms of private development costs and public infrastructure costs than what they called 'unplanned' communities at low densities. This report was attacked immediately on several fronts. Before the report had appeared, several researchers had noted that what appeared to be scattered low-density development was often a high-density compact neighborhood in the making. In fact, scattered development often produced even higher densities in the end, than development built all at one time (Lessinger 1962). Another attack on the conclusions in the report came from reviewers who stated that the apparent savings actually were primarily due to the failure of the report to hold conditions constant in the various models (Altshuler 1977, Windsor 1979). Other researchers suggested that master-planning, as opposed to the hypothetical models used by the Real Estate Research Corporation, did not in practice yield the savings that should have been expected and was often more expensive (Peiser 1984). The supposed efficiencies of high density, likewise, were disputed (Ladd

1992). More recently, a new report on the costs of sprawl has tried to respond to these challenges (Burchell et al. 1998). According to this report, the dire predictions of previous researchers had not materialized but, it opined, the higher costs of low density settlement, particularly road and other infrastructure costs, would soon become intolerable.

2.2 Equity Concerns

The second charge against sprawl, that it is blamed by some for exacerbating economic and social inequities, was, until recently, much more common in the United States than elsewhere, because it was primarily in the United States that affluent individuals moved in great numbers from the central city to the far suburbs. Concerns about inequality became common in the wake of the civil unrest in American cities in the 1960s and concerns about the survival of the inner city. According to the view of many reformers, sprawl was not inevitable or even a freely made choice by most of the population.

Instead, it was a condition that was, at least in part, induced artificially by massive federal subsides in favor of homeownership, notably homeowner insurance guarantees, mortgage interest deductions that supposedly favored new suburban home building, and federal funding for the massive postwar highway building program. As affluent citizens moved out of the central city, this tended to maroon poor and minority families who were unable to move at the same pace. At the same time, it tended to deplete the tax base of the older municipalities as, according to this line of thought, they lost sales taxes and property taxes to outlying municipalities. Other observers have noted that none of these programs was designed to produce massive decentralization, and that any of them might have had quite different effects absent a high level of affluence and a strong underlying desire to decentralize.

2.3 Environmental Concerns

Environmental concerns started to play an increasing role in anti-sprawl polemics following the ecological activism of the late 1960s. According to this line of thought, low-density settlement led to excessive energy usage, particularly automobile usage (Newman and Kenworthy 1989). This in turn led to the depletion of resources and to increased pollution. Low-density settlement also led to the loss of farmland, most seriously prime farmland near urban areas, and to forests, wetlands, and other open lands. Attempts to stop sprawl for environmental reasons have often been grouped under the heading of sustainability (Jenks et al. 1996, Brotchie et al. 1995).

Clearly, energy usage has been shown to be much higher in the affluent countries of the world than in the developing world. Cities in the United States, moreover, use much more energy than their counterparts in Europe. For many environmentalists, this constitutes a fundamental human challenge that can only be solved by stringent conservation measures and replanning cities to be more energy-efficient. Other scholars have noted a history of decades of dire warnings about looming energy shortages and ecological disaster, none of which has come to pass, and have suggested that technology itself will be much more effective in finding answers to problems than efforts to turn the clock back.

Perhaps the most contentious area involving both energy consumption and environment remains the use of the private automobile. Sprawl reformers are almost all united in the belief that private automobile usage must be lessened and public transportation strengthened to reduce both congestion and pollution. On the other hand, other scholars have pointed out that both congestion and pollution levels are often higher in denser metropolitan areas than they are in areas of lower density. The idea that it will be possible to counter the rise in automobile ownership and per capita vehicle miles traveled (VMT), which is a feature of virtually every city in the world as it has become more affluent, or to counter the drop in the proportion of citizens riding public transportation, which has also been true in virtually all of these places, is, according to some observers, both unlikely and undesirable (Gordon and Richardson 1997).

One of the most emotional parts of the debate about sprawl has concerned the loss of farmland. According to many anti-sprawl reformers, cropland, particularly prime cropland, is an irreplaceable asset which is being fast eroded by urban expansion. Those who take an opposing view counter that very little of the farmland lost since World War II has been lost to development, at least in the United States, and that the pace of farmland loss is moderating in any case. The major reason for loss of farmland, according to this point of view, is that productivity has soared and there is less need for land planted in crops. The increase in land for urban uses, moreover, has been less than the amount of additional land added to the supply of permanently protected open spaces (Staley 1999).

2.4 Social Concerns

Another charge against sprawl coming out of the activism of the late 1960s, and most common in the United States, was a belief that sprawl leads to impoverished social relationships and communities. Antagonism to sprawl on social grounds has often been bound up with attacks on suburbs generally. These were seen as alienating, soulless places in the famous sociological studies of David Riesman and William H. Whyte. This literature reached a crescendo

in the postwar decades before it was rebutted by other sociologists, for example, Herbert Gans (Riesman 1950, Whyte 1958, Gans 1967). The equation of suburbia and sprawl with social alienation was revived in the 1990s. One of the most potent areas of concern has been the so-called privatization of public spaces, for example, the decline of Main Street shopping centers in the wake of suburban shopping malls (Sorkin 1992). Other observers have suggested that what is important in social relationships and community has not changed: they argue that only the outward forms of it have changed, and this should not be unduly troubling. A particularly thorny problem related to sprawl in the United States concerns the factor of race. Some writers have believed that sprawl has been caused in great part by racial tension, and that the low-density suburbs are an attempt by middle class citizens to distance themselves from social problems in the city. Other observers have noted that sprawl has occurred in much the same way in metropolitan areas with few minorities as it has in areas with large minority populations, suggesting that race, while important, may not be an underlying cause. They also note that suburban areas have always been and are increasingly diverse.

2.5 Aesthetic Concerns

The final complaint about sprawl, that it is ugly and symbolically inappropriate, has inspired the most constant and abundant literature throughout the years. The sprawling city has been said to be unplanned, chaotic, monotonous, neither urban nor rural but a combination of the worst aspects of each. One of the chief problems with this kind of judgment, however, has always been that various groups differ widely in their aesthetic sensibilities, and that, even within a single group, responses can change dramatically over the years. It has also been conspicuous, from the interwar complaints of British landowners down to the beginning of the twenty-first century, that there are class-bound aesthetic interests at work in this kind of critique.

3. Proposed Remedies for Sprawl

The most obvious and dramatic remedy to sprawl would be to stop new development outside existing urban areas altogether. This could be done by outright governmental prohibitions or by limiting the development of necessary infrastructure, notably roads, water delivery systems, and wastewater treatment facilities. By whichever means, stopping peripheral growth altogether has been tried but has rarely worked if population was growing. Immediately after World War II, zero growth appeared to be a possibility in European cities but it has since proved illusory. In the United States, on the other hand, growth was always considered a given. In the wake of the postwar boom period, the 'slow growth' and 'no growth' movements became prominent in a number of US cities. These were, in many ways, the precursors of turn-of-the-century anti-sprawl movements. In most cases, growth was stopped or slowed in a given municipality, but subsequent research has suggested that the growth merely migrated to adjacent communities (Porter 1997, Garvin 1996).

3.1 Channeling or Managing Growth

Attempting to manage or channel growth, rather than stop it altogether, on the other hand, has often been implemented over the years with at least limited success. One of the earliest and most important attempts to do this was the greenbelt system instituted around British cities, notably around London. In this system, a growth boundary was established marking the edge of a greenbelt in which development was forbidden. Any additional growth that could not be accommodated in the existing developed area was to be confined as much as possible to existing towns and villages beyond the greenbelt, and to newly established 'new towns.' This system, which borrowed many features from Ebenezer Howard and his garden city concepts, did in fact stop much, although not all, of the growth that otherwise might have invaded the greenbelt around London. It was not nearly as successful in containing growth beyond the belt. In fact, growth beyond the greenbelt eventually scattered across much of southeast England. Moreover, by restricting the land available for new development adjacent to the already built-up area, it very likely raised land prices considerably, making it more difficult for people of modest means to afford housing, and perhaps driving them further out than would have been the case without the restrictions (Hall et al. 1973). Still, the London model has been very influential. A similar kind of greenbelt, together with a set of policies encouraging compaction and public transportation, has been put in place by the regional government of Portland, Oregon. The experience of Portland, which has been America's most conspicuous exemplar of anti-sprawl planning, has been watched closely and debated hotly (Abbott et al. 1994, Mildner et al. 1996).

In other cities, other kinds of devices were used to channel growth. In Paris, for example, new development was to be channeled into specific growth corridors organized around transportation lines and large new towns. The same kind of system was proposed for the Washington DC metropolitan area. One virtue of these proposals was that they mirrored the way development had occurred historically along transportation routes. In practice, however, it proved difficult to enforce these corridors in the face of major

demand for new land to develop, and, because they tended to be based on rail transportation, they are thought by some to be inadequate for areas at the periphery where most people drive.

3.2 Open Space Conservation

Another measure sometimes counted as an anti-sprawl technique involves the creation of permanent open space. This can be done by legislative fiat, by the purchase of land or of development rights by private or public bodies, or some combination of these mechanisms. Although these measures are often counted as anti-sprawl measures, they do not, in themselves, stop sprawl but merely channel whatever growth would otherwise occur in specific directions. They also tend to reduce gross densities, arguably fostering sprawl.

3.3 Governmental Structures and Incentives

Other techniques for curbing low-density settlement and promoting high-density growth involve governmental policies and structure. In the United States many sprawl reformers blame competition rather than cooperation among units of local government. The American tradition of deciding land use policies at the local level, combined, in many cities, with a very fragmented system of municipal government, has been seen by reformers as a major reason for lack of coherence in planning, and as a major incentive for development to sprawl (Gottdiener 1977). The remedy, for reformers, is to form metropolitan-wide governments with strong planning responsibilities, or to push responsibility for planning up to the state or federal level, so that local competition for tax dollars will be less an influence in planning decisions. Moreover, tax revenues would be shared more evenly, making it difficult to move to avoid taxes, and allowing the possibility of a more equitable distribution of income for schools and other public services (Orfield 1997). Among the most important metropolitan governments in North America are those of Portland, Minneapolis-St. Paul, and Toronto. States that have initiated strong growth management laws include Oregon, Florida, and New Jersey. Whether or not these policies will have any substantial effect in stopping the outward decentralization of people and jobs in the long run is still unknown.

Another area targeted by reformers involves tax policies and other governmental fiscal policies. Sprawl was subsidized by federal mortgage insurance policies and by the mortgage interest deduction on the federal income tax, according to reformers. Low-density settlement was also subsidized by heavy federal subsidies provided for roads, and the lack of subsidies at the same scale for public transportation. Many reformers would like to see the end of the federal mortgage interest deductions, heavy taxes on gasoline and vehicles, and much greater funding for transit. One influential group has justified these moves as correctives to make the market reflect more adequately the externalities caused by things like automobile use (Downs 1994). Other observers suggest that any major program to introduce sharp increases in fees on private transportation would be highly regressive; and they question whether, given the apparently inexorable trend toward low density, the higher fees on private transportation and more funding for projects such as rail transit would actually lead to a corresponding rise in transit use, or just to yet higher levels of highway congestion, pollution, and economic hardship (Gordon and Richardson 1997).

3.4 Design Solutions

A final set of recommendations to deal with sprawl has been made at the level of urban design. Proposals for transit-oriented design (TOD), in which high density uses are planned near transit stations, or 'New Urbanist' designs, which attempt to create a mix of uses to promote community and transit use, have all been advocated, and some building has actually occurred (Katz 1994). It is again too early to judge the success of these efforts, which to date have been very modest, but, in the last years of the twentieth century, there was little evidence that these developments, mostly suburban and at quite low densities, had had any effect on overall urban patterns.

3.5 The Sprawl Debate

Beyond the question of problems supposedly created by sprawl, or the efficacy of any particular remedy, is a larger set of questions about whether sprawl is recent, whether it is accelerating or is actually declining, and whether it is peculiarly American or whether American cities merely represent an advanced stage in a process that is visible everywhere in the world. There has actually been very little research on many of these questions but there does seem to be some evidence that in the United States, where the debate at the end of the twentieth century was hottest, decentralization and sprawl might have had their greatest per capita development in the period between the 1920s and the end of the postwar decades, and might actually be declining. There is also a fundamental question about whether the denser cities of Europe and elsewhere are actually different from American cities, or exhibiting all of the same characteristics but merely evolving at a slower rate and less obviously because of the large mass of existing urban fabric (Bruegmann 1996).

In the United States and Britain, the chief opposition to sprawl since World War II has come from the left side of the political spectrum and tends to unite academic planners, land conservationists, and historic preservationists. The anti-sprawl efforts have been opposed primarily by individuals from the right side of

the political spectrum, particularly land economists and those who favor an increased role for the private market. Elsewhere in the English-speaking world, particularly in Australia, the situation appears to be somewhat different, with reformers to the left of center objecting that efforts to enforce more compact development for efficiency reasons will inevitably have a negative impact on the less affluent part of the population, and exacerbate existing inequalities in levels of service and amenity (Troy 1996).

See also: Metropolitan Growth and Change: International Perspectives; Planning, Politics of; Residential Segregation: Sociological Aspects; Sustainable Development; Sustainable Transportation; Urban Geography; Urban Growth Models; Urban History; Urban Places, Planning for the Use of: Design Guide; Urban Planning: Growth Management; Urban Poverty in Neighborhoods; Urban Sociology; Urban System in Geography

Bibliography

Abbott C, Howe D, Adler S (eds.) 1994 *Planning the Oregon Way: A Twenty Year Evaluation*. Oregon State University Press, Corvallis, OR

Abercrombie P 1945 *Greater London Plan 1944*. His Majesty's Stationery Office, London

Altshuler A 1977 Review of the costs of sprawl. *Journal of the American Institute of Planners* **43**: 207–9

Brotchie J, Battey M, Blakely E, Hall P, Newton P (eds.) 1995 *Cities in Competition: Productive and Sustainable Cities for the 21st Century*. Longman, Melbourne, Vic, Australia

Bruegmann R 1996 The American city: Urban aberration or glimpse of the future? In: Cohen M, Rubble B, Tulchin J, Gaylord A (eds.) *Preparing for the Urban Future: Global Forces and Local Pressures*. Johns Hopkins University Press, Baltimore, MD

Bruegmann R 2000 The paradoxes of urban sprawl reform. In: Freestone R (ed.) *Urban Planning in a Twentieth Century World*. Routledge, London

Burchell R, Listokin C, Shad N, Phillips H, Downs A, Seskin S, Davis J, Moore T, Helton D, Gall M 1998 *Costs of Sprawl Revisited, Transportation*. Research Board, National Research Council, Washington, DC

Downs A 1994 *New Visions for Metropolitan America*. Brookings Institution, Washington, DC, and Lincoln Institute of Land Policy, Cambridge, MA

European Commission 1990 *Green Paper on the Urban Environment*. European Commission, Brussels, Belgium

Ewing R 1997 Is Los Angeles-style sprawl desirable? *Journal of the American Planning Association* **63**: 107–26

Gans H 1967 *The Levittowners: Ways of Life and Politics in a New Suburban Community*. Allen Lane, London

Garvin A 1996 *The American City: What Works, What Doesn't*. McGraw Hill, New York

Gordon H, Richardson P 1997 Are compact cities a desirable planning goal? *Journal of the American Planning Association* **63**: 95–106

Gottdiener M 1977 *Planned Sprawl: Private and Public Interests in Suburbia*. Sage Publications, Beverly Hills, CA

Hall P, Gracey H, Drewett R, Thomas R 1973 *The Containment of Urban England*. PEP, London, 2 Vols.

Hebbert M 1994 Urban sprawl and urban planning in Japan. In: Shapira, Masser I, Edington D (eds.) *Planning for Cities and Regions in Japan*. Liverpool University Press, Liverpool, UK

Jenks M, Burton E, Williams K 1996 *The Compact City: A Sustainable Urban Form?* E & FN Spon, London

Katz P 1994 *The New Urbanism: Toward an Architecture of Community*. McGraw-Hill, New York

Ladd H 1992 Population growth, density and the costs of providing public services. *Urban Studies* **29**(2): 273–95

Lessinger J 1962 The case for scatteration. *Journal of the American Institute of Planners* **28**(3): 159–69

Mildner G, Dueker K, Rufolo A 1996 *Impact of the Urban Growth Boundary on Metropolitan Housing Markets*. Center for Urban Studies, Portland, OR

Mumford L 1938 *The Culture of Cities*. Harcourt, New York

Newman P, Kenworthy J 1989 *Cities and Automobile Dependence: An International Sourcebook*. Gower Publishing, Aldershot, UK

Nivola P S 1999 *Laws of the Landscape: How Policies Shape Cities in Europe and America*. Brookings Institution, Washington, DC

Orfield M 1997 *Metropolitics: A Regional Agenda for Community and Stability*. Brookings Institution Press, Washington, DC, and Lincoln Institute of Land Policy, Cambridge, MA

Peiser R 1984 Does it pay to plan suburban growth? *Journal of the American Planning Association* **50**(3): 419–33

Porter D 1997 *Managing Growth in America's Communities*. Island Press, Washington, DC

Razin E 1998 Policies to control urban sprawl: Planning regulations or changes in the rules of the game. *Urban Studies* **35**(2): 321–40

Real Estate Research Corporation 1974 *The Costs of Sprawl*. United States Government Printing Office, Washington, DC, 2 Vols

Riesman D 1950 *The Lonely Crowd: A Study of the Changing American Character*. Yale University Press, New Haven, CT

Sharp T 1932 *Town and Countryside: Some Rural Aspects of Urban and Rural Development*. Oxford University Press, Oxford, UK

Sorkin M (ed.) 1992 *Variations on a Theme Park*. The Noonday Press, New York

Staley S 1999 *The Sprawling of America: In Defense of the Dynamic City*. Reason Public Policy Institute, Los Angeles

Troy P 1996 *The Perils of Consolidation: A Discussion of Australian Housing and Urban Development Policies*. The Federation Press, Annandale, NSW, Australia

Whyte W 1956 *The Organization Man*. Simon and Schuster, New York

Whyte W 1958 Urban sprawl. In: *The Exploding City*, Fortune Magazine, pp. 115–39

Windsor D 1979 A critique of the costs of sprawl. *Journal of the American Planning Association* **45**(3): 279–92

R. Bruegmann

Urban Studies: Overview

Urban studies is the interdisciplinary academic field concerned with understanding cities. Scholars from the social and behavioral science disciplines of history, sociology, geography, economics, political science,

and anthropology and the professional field of urban planning have produced most substantive urban studies knowledge. Urban studies scholars primarily use social science research methodologies. They have developed a body of theory about cities. The field of urban studies has its own academic courses and degree programs, professional associations, and scholarly journals.

While cities have attracted scholarly attention for more than two millennia, rapid urbanization during the industrial revolution stimulated the first systematic urban studies research and writing in the nineteenth century. In the early twentieth century urban studies entered academic curricula through programs in urban planning, sociology, and then other social science disciplines. The field grew slowly through the end of the 1950s. The 1960s saw a huge increase in urban studies research, courses, and academic degree programs.

Today urban studies is sometimes pursued in interdisciplinary urban studies programs; sometimes as part of curricula in geography, sociology or other academic disciplines. Urban studies material is taught as part of professional urban planning curricula. There are many niche subfields within urban studies where two or more disciplines or fields intersect. Urban political economy, urban planning history, urban economic geography, and urban political sociology are examples.

1. The Field of Urban Studies

1.1 Urban Studies, Academic Disciplines, and Professional Fields

The categories into which universities divide human knowledge are ever changing and their boundaries are not clear. Universities usually organize academic disciplines such as sociology and economics separately from professional fields such as urban planning and architecture. Most urban studies material comes from social and behavioral science disciplinary perspectives; some from other humanities, natural sciences, creative arts and elsewhere.

Scholars in the social science disciplines of geography, sociology, economics, anthropology, political science, and history (which is sometimes classified as one of the humanities) have generated most of the urban studies literature. Urban planners, architects, urban designers, and landscape architects have contributed to the field. Journalists like Robert Caro, Neil Peirce, and Joel Garreau have written important urban studies books, as have self-taught polymaths like Jane Jacobs and Lewis Mumford. Key writings by scholars for different disciplines have been collected in an anthology edited by Richard LeGates and Frederic Stout (LeGates and Stout 1999).

Since the object of study is so large and the perspectives that can and have been brought to bear on it so diverse, the boundaries of the field of urban studies are harder to define than disciplines such as paleontology and botany. The field is partly descriptive—seeking to describe the physical form and social characteristics of cities. It is partly explanatory—concerned with explaining physical and social attributes of cities. Description and explanation lead urban studies scholars into prediction and prescription.

1.2 How Urban Studies is Organized Within Universities

Sometimes departments or programs in urban studies hire faculty trained in sociology, political science, economics, history, anthropology, urban planning and other disciplines and professional fields to teach urban studies courses. Often they have relationships with faculty-based disciplinary departments. For example, an urban studies program which does not have an urban economist on its faculty may have an urban economist teach a course for their students.

Urban studies material is part of the curricula of geography, sociology, and other social science disciplines. It is taught in professional urban planning, architecture, landscape architecture, and urban design programs. Faculty based within disciplines emphasize scholarship from their respective disciplines, but almost all disciplinary-based courses incorporate writings by scholars from other disciplines.

1.3 Urban Studies and Related Degree Programs

Many universities offer BA degrees in urban studies. Brown, Cleveland State, Cornell, Columbia, Hunter, Minnesota, MIT, Pennsylvania, Pittsburg, Rutgers, San Francisco State, Stanford University, and Vassar are examples (Urban Affairs Association 1995). Michigan State, New Orleans, Ohio, Ohio Wesleyan, Portland State, and Temple offer masters degrees in urban studies. Portland State, Michigan State, and the University of New Orleans offer Ph.D. degrees in urban studies.

Many universities offer degrees with names like urban studies and planning, urban affairs, and metropolitan affairs. Graduate urban planning programs at MIT, Portland State, Virginia Commonwealth University, the University of Maryland and elsewhere offer masters degrees through departments of urban studies and planning (ACSP 1998).

Two academic associations organize conferences, set standards, publish academic journals, and work to advance scholarship in urban studies. In North America the Urban Affairs Association (UAA) has an annual meeting, publishes the *Journal of Urban Affairs* and a newsletter, and has published a *Guide To Programs in Urban Studies and Affairs* (UAA 1995). A European counterpart organization, the European

Urban Research Association (EURA), was created in 1998.

The *American Planning Association Journal* in North America, the *Town Planning Review* in the UK, *The International Journal of Urban and Regional Research*, the journal *Urban Studies*, and many more specialized journals publish scholarly urban studies material.

1.4 Urban Studies Research Methodology

Urban studies scholars employ both quantitative and qualitative research methods, primarily from the social sciences. Scholarly writing about how to do urban studies research is sometimes explicit; often implicit in the methodologies urban studies scholars have actually employed in their research.

Training in applied statistics is a regular part of most urban studies curricula. Urban studies students are trained to use computerized statistical packages to do quantitative analysis. Because many urban phenomena have a spatial dimension geographical information systems (GIS) software is increasingly important in urban studies education and practice.

Urban studies scholars also employ qualitative methods. William Whyte's study of how people use urban parks and plazas is a notable example of observation. Urban anthropologist Oscar Lewis conducted exhaustive field research. Sam Bass Warner did archival research on the records of ordinary developers of ordinary Boston streetcar suburbs. Planner John Forester conducted interviews of dozens of urban planners to understand planning in the face of conflict. Urban studies research often employs both quantitative and qualitative research methods to triangulate on the object of study.

2. The Evolution of Urban Studies

While scholars have studied, thought, and written about cities for more than two millennia, systematic studies of cities did not begin until the nineteenth century. Academic recognition of the field occurred only in the twentieth century, and most urban studies scholarship has taken place since 1960.

2.1 Early Urban Studies

Cities have excited the interest of scholars since the earliest times. In *The Republic*, the Greek philosopher Plato envisioned the ideal polis in the fourth century BCE. The Roman architect Vitruvius formulated principles for planning cities. Monks worked out a detailed plan for the physical and social organization of the monastery of St. Gall, Switzerland, and by extension other mediaeval communities, in the eighth century AD. Florentine architect Leon Battista Alberti reminded his peers of what cities could contribute to the rebirth of humanism during the renaissance.

2.2 The Industrial Revolution, Nineteenth Century Industrial Cities, and Urban Studies

Prior to the industrial revolution there were few cities and—with the exception of imperial Rome—cities were very small. As water- and then steam-driven machinery fueled the industrial revolution, urbanization proceeded with unprecedented speed and on a scale never before seen in human history. Industrial cities arose in textile towns of the English midlands, then through much of Europe and North America. By the middle of the nineteenth century unplanned, congested, polluted slums jammed with a new impoverished urban proletariat dominated nineteenth century industrial cities. Friederich Engels, Edwin Booth, and Asa Briggs conducted empirical investigations of nineteenth century cities. Robert Owen, Charles Fourier, and Henry de Rouvreoy, Saint-Simon and their followers planned, and sometimes built, utopian communities. Anarchist Petr Kropotkin agitated for cities built around factories in fields to replace the nation state. Karl Marx and Friedrich Engels called for the liberation of the urban proletariat and the elimination of distinctions between city and countryside. Landscape architects Andrew Jackson Downey and Frederick Law Olmsted led an urban parks movement. German land economists Johann Heinrich von Thünen and August Lösch developed models that blossomed into twentieth century location theory. German sociologists Ferdinand Tönnies and Max Weber and French sociologist Emile Durkheim developed theories of societal organization that prefigured urban sociology. Civil engineers like Baron Haussman in Paris and Alberto Soria y Mata in Madrid, public health officials like Benjamin Ward Richardson in London, and settlement house workers like Jane Adams in Chicago studied, theorized about, and agitated for the improvement of nineteenth century industrial cities. But it was not until the twentieth century that academic courses and degree programs in urban studies were created.

2.3 Academic Urban Studies in the Twentieth Century

The University of Liverpool in England and the Massachusetts Institute of Technology in the United States began the first formal courses devoted to cities (Hall 1996). Liverpool established a degree program in town and country planning in 1909 and MIT instituted a course taught by urban planner Thomas Adams the same year. These courses emphasized physical planning and design, but, fortunately, both Thomas

Adams at MIT, and Stanley Adshead and Patrick Abercrombie at Liverpool were well-read interdisciplinary academics who incorporated historical, economic, social, and political material into their urban planning courses, textbooks, and academic writings.

Systematic research and urban studies teaching took a giant leap forward beginning a decade later when the University of Chicago created the world's leading center of urban sociology starting in the 1920s. The Chicago school of sociology attracted creative and talented scholars who used neighborhoods surrounding the campus to study urban society. They developed the field of urban sociology and inspired development of the urban studies content of other disciplines.

Coursework with cities as the unit of analysis slowly penetrated other social science disciplines and the number of urban planning programs grew during the 1930s, 1940s and 1950s. Still, by the end of 1950s, there were only a handful of courses in urban politics, urban history, and urban anthropology, not one text in urban economics, and just a few dozen small design-oriented urban planning programs in the entire world.

Ghetto riots, and a massive increase in federal aid to cities in the United States in the 1960s greatly expanded urban studies. The 1960s saw works by Lewis Mumford in urban history (Mumford 1961), Oscar Lewis in urban anthropology, Edward Banfield in urban politics, Herbert Gans (Gans 1962) in urban sociology, and Wilbur Thomson in urban economics. Jane Jacobs' *The Death and Life of Great American Cities* was a best seller. Discipline-based and interdisciplinary urban studies courses proliferated during the 1960s and early 1970s and many of today's academic urban studies programs were created at that time.

3. Core Urban Studies Concerns

Following is a list of 12 core concerns in the field of urban studies. Notable scholars and key scholarship in each area are discussed.

3.1 The Evolution of Cities

Historians have made the greatest contribution to understanding the evolution of cities. They have been joined by archeologists, geographers, political scientists, economists, architects, and other scholars.

Archaeologist V. Gordon Childe in Mesopotamia, Mortimer Wheeler in the Indus River Valley, James Mellaart in Antolia, and Dame Kathleen Kenyon in Israel excavated Ur in present day Iraq, Mohenjo-Daro and Harrapa in Pakistan and India, Catal Hüyük in Turkey, and Jericho in Israel. These scholars developed descriptions of what the earliest cities were like and conflicting theories of why they arose.

Historians have approached the study of cities from many different perspectives. Raymond Mohl (Mohl 1997) has proposed an 11-category typology of American urban history which provides a useful introduction to all historical urban studies. Mohl's categories are: urban political history, suburbanization, city and region, sunbelt cities, technology and the city, planning and housing, the urban working class, immigration and ethnicity, African-American urban history, urban policy history, and urban culture.

Urban planners such as John Reps (Reps 1965), Eugenie Birch, Christine Boyer, and Mellior Scott in the United States and Sir Peter Hall and Steven Ward in England have developed the subfield of urban planning history. They have been joined by historians Carl Abbot, Mary Corbin, Christopher Silver, William H. Wilson, David Schuyler, and Stanley Buder, geographer Jay Vance, architectural historian Spiro Kostof (Kostof 1991) and many others studying topics as varied as the building of Athens, Michaelangelo's work in renaissance Rome, Baron Hausmann's rebuild of Paris in the mid-nineteenth century, city beautiful planning for early twentieth century Chicago, and the vision and reality of planned garden cities.

3.2 Urban Culture

Cities have always been centers of both high and low culture. Historian Arnold Toynbee, geographer/planner Peter Hall (Hall 1998), polymath Lewis Mumford and many others have studied urban culture.

Greek drama, poetry, art, and sculpture flourished in Athens in the fifth century BCE. Imperial Rome produced both culture for elites and gladiatorial spectacles for the masses. Florence was an incubator of renaissance art, architecture, sculpture, poetry, and literature. Elizabethan London generated the theaters and dramatic troops that made Shakespeare's plays possible.

Lewis Mumford pioneered the study of urban culture in the late 1930s in *The Culture of Cities* (Mumford 1938). Historians Warren I. Susman, Gunther Barth, Neil Harris, and others have studied the history of urban culture. Richard Sennett has collected studies of urban culture by sociologists, geographers, and others (Sennett 1969).

The dominant strand in current culture studies of cities is postmodernism. Geographers David Harvey and Edward Soja, sociologist Sharon Zukin, social critic Mike Davis and many other scholars are fascinated and often horrified by the clash of cultures and values most manifest in contemporary cities.

3.3 Urban Society

The Chicago school of sociology encouraged faculty and students to study Blacks in Bronzeville, prostitutes in the Gold Coast, hobos, polish peasants, gangs, and many other Chicago groups. Urban ethnographies—detailed field research on urban subcultures continue this tradition. Anthropologist Elliot Liebow's poig-

nant descriptions of poor, black streetcorner men and homeless women are exemplary of this tradition.

Another urban studies concern is with how rural migrants adapt to city settings. There are many studies of urban immigration, acculturation, and assimilation such as Oscar Handlin's studies of the Boston Irish and Ron Takaki's studies of Asian immigrants to the West coast of the United States. Chicago school sociologists pioneered urban ecology to describe the structure and processes of urban ethnic neighborhood change, and the study of urban ecology continues (Berry and Kasarda 1977).

Problems of poverty are central to urban studies. Anthropologist Oscar Lewis' controversial theory that a distinct culture of poverty exists stimulated scholarly debate in the 1960s and 1970s. Harvard sociologist William Julius Wilson's studies of the urban underclass continue the debate.

Yale Professor Dolores Hayen and sociologists Lynn Lofland and Janet AbuLughod study the way in which humans use public space. Anthropologists and sociologists have studied the way in which human beings use the space around them. Edward T. Hall studied how people respond to intrusions into their space. Robert Sommers has observed behavioral aspects of how people use the space in such diverse settings as prisons, airports, and jury rooms. These behavioral and social science studies provide guidance to planners and designers.

Another enduring concern in urban studies is the nature of community. Emile Durkheim and Fernand Tönnies regretted a loss of community in large, impersonal nineteenth century cities. Louis Wirth concluded that big-city life created a distinct—and not very attractive—human personality type (Wirth 1938). Other social scientists have identified positive networks and supportive communities within cities. For example, sociologists Herbert Gans (Gans 1962) characterized Boston's West End as an urban village where migrants from southern Italy retained many of the relationships and folkways of their villages of origin. Michael Young and Peter Willmott's classic study of family and kinship in East London found positive networks in a London slum and far less community in new housing estates built for the relocated slum dwellers. The nature of community is one strand in the large and growing urban studies literature on suburbia by scholars such as historian Kenneth Jackson, political scientist Robert Wood, and sociologist Mark Baldassare.

3.4 Urban Politics and Governance

The study of how cities are governed is a central concern for political scientists. A recent review of the literature on how to study urban politics by New York University political scientist John Mollenkopf summarizes the evolution and current status of the subfield of urban politics (Mollenkopf 1992).

Historians like Amy Bridges, Leo Hershkowitz, and Martin Schefter have provided historical studies of urban political bosses and political machines. Samuel Hayes and Roy Lubove, have studied urban progressives and reformers. Economist James O'Connor, geographer David Gordon, social work professor Francis Fox Piven and other urban political economists have brought Marxist and related critical theory to bear on understanding urban politics. Public choice theorists—mainly based on economics—have developed theories about how local governments do (and should) make the economic decisions that constitute the core of urban governance. Scholars like David Osborne and Ted Gaebler from public administration study how cities are (and should be) managed. Journalist Robert Caro won a Pulitzer prize for his monumental study of Robert Moses and the politics of planning the twentieth century New York region.

Sociologist Floyd Hunter and political scientist Robert Dahl in classic studies of governance in Atlanta, Georgia and New Haven, Connecticut developed elitist and pluralist models of urban community power respectively. Other pluralists like Edward Banfield, Herbert Kaufman, Martin Meyersen, Wallace Sayre, and Douglas Yates conflict with Peter Bachrach, Morton Baratz, John Manley, William G. Domhoff and other theorists who continue to advance elitist explanations of community power.

Marxist political theorists like political scientists Ira Katznelsen, William Tabb, and David Gordon, economist James O'Connor, and French geographer Henri Lefebre posit structuralist theories of urban political power. Structuralist interpretations argue that underlying economic and class relationships pre-determine most of what is important to decide at the city level. Structuralists see city governments as unable to control national and international capitalist forces. For them, urban politics is largely irrelevant.

Regime theory is an approach to studying community power, developed by Political Scientist Clarence Stone. Stone, Frank R. Baumgartner, Richard DeLeon, Steven Elkin, Brian D. Jones, Gerry Stoker, and other regime theorists conclude that complex, public/private regimes involve not only local elected officials, but also people involved in business, church leaders, unions, the media, and other interest groups. Regime theorists do not deny the structural constraints local governments face, but they argue that complex, shifting public/private regimes are important to local political outcomes.

3.5 Urban Economics, Urban Public Finance, and Regional Science

The economics of cities is primarily the province of economists. Since Wilbur Thomson's *Preface to Urban Economics* urban economics has been recognized as a subfield of economics. Economic base theory, input–output analysis, location theory and other constructs

as well as numerous studies of specific city economies enrich scholarly understanding of cities.

Economic geographers have studied the production and exchange of goods within and between cities. Historians interested in both economics and cities—like Belgian historian Henri Pierenne and French historian Fernand Braudel—have developed the subfield of urban economic history and have illuminated the relationship of capitalist institutions and trade to the rise of cities.

Two other interdisciplinary subfields closely related to economics help inform urban studies. Urban public finance is sometimes taught in economics departments; sometimes is business schools. It contributes to an understanding of how local governments raise revenue and the nature of their expenditures. The subfield of regional science includes economists, geographers, and others who study regions rather than cities. Regional scientists generally employ quantitative methodologies. They have their own professional associations and scholarly journal.

3.6 Urban and Metropolitan Space and City Systems

Another core area of concern in urban studies is the physical organization of cities and regions and relationship among cities. This work is often classified into studies of the internal structure of cities and studies of systems of cities. Geographers are particularly interested in these issues, but sociologist Ernest W. Burgess and economist Homer Hoyt originally did pioneering work on the internal structure of cities. Writings about the internal structure of cities have been collected in an anthology edited by University of Toronto geographer Larry Bourne (Bourne 1982).

German geographer Walter Christaller's careful studies of the size and function of settlements in southern Germany during the 1930s caused him to formulate central place theory—a unified explanation of the way in which cities of different sizes are related to each other in a distinct hierarchy or system (Christaller 1993). Christaller's work led to many empirical studies of systems of cities. University of Toronto geographers Larry S. Bourne and J. W. Simmons have collected systems of cities studies (Bourne and Simmons 1978).

3.7 Mega Cities and Global City Systems

The rise of enormous cities with populations of 10 million or more people has captured the attention of urban studies scholars. So have the increasingly integrated world economic system and the close connections between cities and regions around the globe.

One strand of research and writing concerns the rise of mega cities by Janet AbuLughod, Manual Castells, Susan Fainstein, Peter Hall, Alan Gilbert, John Kasarda, Fuchen Lo, Janice Perlman, Saskia Sassen (Sassen 1994), Yue-man Yeung and authors of UN-commissioned city-specific mega city studies. The nature, causes, and appropriate policy response to mega cities is particularly important in Asia and South America, where most of the mega cities are emerging.

A related area concerns development of huge urbanized regions—what Scottish biologist Patrick Geddes in the early twentieth century termed conurbations and French geographer Jean Gottman in a study of the eastern seaboard of the United States named megalopolis.

A third area of inquiry involves the increasingly close relationship among cities caused by economic and political integration and by information technology. Spanish sociologist Manuel Castells uses the term informational city to describe advanced cities today and the term network society to describe the current global system in which information-rich multinational corporations are increasingly beyond the control of city government, citizens, and even nation states. University of Chicago and London School of Economics, urban planning professor Saskia Sassen (Sassen 1994) has documented the current global system of cities.

3.8 Technology and Cities

A number of urban studies scholars are interested in aspects of technology in cities. Transportation has received the most attention, but public infrastructure has also been of interest. The impact of the Internet and other information technologies on cities is attracting increasing attention and promises to be even more important in the twenty-first century.

There is a large literature on the impact of transportation technology on cities. Historians like Sam Bass Warner, Charles W. Cheape, and Clifton Hood have documented the way in which the electric streetcar directly affected first the spatial structure of cities and metropolitan regions and, as a consequence, their social and economic structure. Mark S. Foster, William Issel, Kenneth Jackson, Clay McShane, Mark H. Rose, and others have studied the impact of automobiles on the rise of suburbia. They are interested in public policy regarding freeways and automobiles and its impact on urban and regional physical form, social structure, and culture. Policy analysts and planners like Anthony Downs and Robert Cervero write about transportation policy and planning issues.

3.9 Urban Planning, Urban Design, Landscape Architecture, and Architecture

Urban studies overlaps the professional fields of urban planning, urban design, architecture, and landscape architecture. Writings about cities and urban planning have been collected by Jay Stein (Stein 1995).

Professors of urban planning in departments that belong to the Association of Collegiate Schools of Planning (ACSP) in North America, the Association of European Schools of Planning (AESOP), and the Asian Planning Schools Association (APSA) study cities in order to help plan them. Twentieth century land use planning books by Patrick Abercrombie, Thomas Adams, Alan Altschuler, Edward Bassett, F. Stuart Chapin, Anthony Catanese, Barry Cullingworth, Alexander Garvin, Peter Hall, Jack Kent, John M. Levy, Edward J. Kaiser, David Gotschalk, and James C. Snyder combine material related to geography, economics, politics, sociology, and the design professions with applied material on plan preparation.

Ever since nineteenth century Austrian architect Camillo Sitte derived urban design principles from his careful observation of public spaces in European cities and how people use them, urban designers have been close observers of city space and contributors to urban studies. Kevin Lynch blended history, sociology, psychology, and humanistic material in *The Image of the City* (Lynch 1961). Sociologist William Whyte's studies of urban park and plaza design, planner Allan Jacobs' studies of great streets, Spiro Kostoff's studies of urban form (Kostoff 1991) and works by urban designers Edmund Bacon, Gary Hack, Clare Cooper Marcus, Wendy Sarkissian and other design scholar/practitioners have made major contributions to the field of urban studies.

Ever since the pioneering work of Frederick Law Olmsted in the nineteenth century, landscape architects have contributed to understanding the relationship between the natural and built environments. Systematic study of how to plan human settlements in harmony with the natural environment were pioneered by Scottish biologist Patrick Geddes before World War I and revitalized by Scottish landscape architect Ian McHarg's approach to design with nature in the 1960s. This approach resonates with advocates of ecological design, sustainable urban development, and the new urbanism today.

Architects like Frank Lloyd Wright in the USA, Raymond Unwin in England, and LeCorbusier in France moved far beyond design of individual structures to think and write about the design of entire communities and of society itself. Peter Calthorpe, Andres Duany, Peter Katz, Elizabeth Plater-Zyberk, Daniel Solomon, Sim Van der Ryn, and other new urbanists have made notable contributions to the field of urban studies. Key new urbanist writings have been collected in a recent anthology (Katz 1994).

3.10 Race, Ethnic, and Gender Relations in Cities

William Edward Burghardt Du Bois' 1899 study of the Philadelphia Negro began a tradition of studies by African-American scholars about urban black society. Horace R Cayton, Kenneth Clark, St Clair Drake, E. Franklin Frazier, and William Julius Wilson have continued that tradition.

Historian Oscar Handlin, political scientist Daniel Patrick Moynihan, and sociologist Nathan Glazer's early studies of the urban experience of Irish and Jewish immigrants to cities on the East Coast of America have been followed by many studies of Irish, Jewish, central European, southern European and other ethnic groups in cities. Studies of the urban immigrant experience of Asians and Hispanics are now appearing. Ron Takaki's study of Asian immigrants and writings by and about Latino barrios edited by Joan Moore and Raquel Pinderhughes (Moore and Pinderhughes 1993) are important contributions to urban studies.

Sociologists, anthropologists, historians, planners, and women's studies scholars are generating an interdisciplinary literature about women in cities. Yale professor Dolores Hayden pioneered this subfield with studies on the relationship between gender, work roles and the built environment highly critical of the male-built environment. Eugenie Birch, Ann Forsyth, Jacqueline Leavitt, Lynn Lofland, Claire Cooper Markus, Ann Markusen, Risa Palm, Marsha Ritzdorf, Leonie Sandercock, Wendy Sarkissian, Catherine Simpson, Daphne Spain, and others have helped inject feminist concerns into urban studies and planning. Writings about women in cities have been collected into an anthology (Stimpson et al. 1980).

3.11 Urban Issues and Policy

Urban studies is also concerned with urban issues and policy. It was poverty, overcrowding, congestion, disease, crime, vice, and other urban problems that inspired many nineteenth century studies. Urban pathology was a concern of the Chicago school of sociology. The focus on urban problems dominated urban studies after the ghetto riots of the 1960s.

Current urban studies curricula have broadened the urban problems orientation to also consider community assets and strengths. Today applied coursework in urban studies programs balances studies of positive neighboring contributions, and potential of urban community development with the study of racism, poverty, crime, and other enduring urban problems.

Most urban studies programs include applied coursework in policy areas such as land use, housing, community development, environment, transportation, and economic development. Some have coursework on crime, health, homelessness, education, welfare and other policy areas. A tiny sampling of urban studies scholars with applied interests include John DeGrove (land use), Anthony Downs (housing),

Sim Van der Ryn (the environment), Robert Cervero (transportation), Michael Porter (economic development), James Q. Wilson (crime), Leonard Duhl (health), Talmadge Wright (homelessness), James Coleman (education), and Francis Piven (welfare).

3.12 Urban Futurism

A final strand in urban studies is reflection on probable, possible, and desirable urban futures. Virtually all demographers anticipate massive population growth and continuing urbanization—particularly in the third world. There will be more congestion, sprawl, pollution, exhaustion of natural resources, and proliferation of megacities and vast urban conurbations unless governments intervene to plan and regulate city development far more strictly than they have ever done in the past.

Studies of sprawl and ex-urban development raise the possibility that the rate of growth in cities may slow or even that counter urbanization may occur in the future—at least in the developed world. Some writers believe information technologies will make cities unnecessary and a few even predict a withering away of cities, though there is no evidence that this is occurring.

In addition to prediction, urban futurists have formulated different normative visions for alternative urban futures. Ebenezer Howard (Howard 1898), Patrick Geddes, Lewis Mumford, and Raymond Unwin opposed further growth of large cities and proposed systems of human-scale garden cities. Frank Lloyd Wright went further and advocated Broadacre city—a neo-Jeffersonian future in which each family would live on an acre. Soviet-era anti-urbanists took Marx's ideas of eliminating distinctions between the city and the countryside as a call for de-urbanization. Megastructuralists like Le Corbusier and Paolo Soleri took the opposite view. They proposed—and in Soleri's case are actually building—huge, self-contained structures that would house the human population on a small portion of the Earth's surface.

See also: Development and Urbanization; Neighborhood Revitalization and Community Development; Race and Urban Planning; Urban Anthropology; Urban Ethnography; Urban Geography; Urban History; Urban Poverty in Neighborhoods; Urban Sociology

Bibliography

Association of Collegiate Schools of Planning (ACSP) 1998 *Guide to Graduate Programs in Urban and Regional Planning*, 9th edn.

Berry J L, Kasarda J D 1977 *Contemporary Urban Ecology*. Macmillan, New York

Bourne L S (ed.) 1982 *The Internal Structure of the City*, 2nd edn. Oxford University Press, New York

Bourne L S, Simmons J W (eds.) 1978 *Systems of Cities: Readings on Structure, Growth and Policy*. Oxford University Press, New York

Christaller W 1933 *Die zentralen Orte in Süddeutschland*. Gustav Fischer, Jena, Germany

Gans H 1962 *The Urban Villagers*. Free Press, New York

Hall P 1996 *Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century*, 2nd edn. Blackwell, Oxford, UK

Hall P 1998 *Cities in Civilization*. Pantheon, New York

Howard E 1996 *Tomorrow: A Peaceful Path to Real Reform*, (Sonnenschein, London 1898), reprinted in Richard LeGates and Frederic Stout, *Early Urban Planning 1870–1940*, Routledge/Thoemmes, London

Katz P 1994 *The New Urbanism*. McGraw-Hill, New York

Kostof S 1991 *The City Shaped*. Little Brown, New York

LeGates R, Stout F 1999 *The City Reader*, 2nd edn. Routledge, London

Lynch K 1961 *The Image of the City*. MIT, Cambridge, MA

Mohl R 1997 New perspectives on American history. In: Mohl R (ed.) *The Making of Urban America*, 2nd edn. Scholarly Resources, New York

Mollenkopf J 1992 How to study urban political power. In: *A Phoenix in the Ashes: The Rise and Fall of the Koch Coalition in New York City Politics*. Princeton University Press, Princeton, NJ

Moore J, Pinderhughes R (eds.) 1993 *In the Barrios*. Russell Sage, New York

Mumford L 1938 *The Culture of Cities*. Harcourt Brace, New York

Mumford L 1961 *The City In History*. Harcourt Brace, New York

Reps J 1965 *The Making of Urban America*. Princeton University Press, Princeton, NJ

Sassen S 1994 *Cities in a Global Society*. Pine Forge Press, Thousand Oaks, CA

Sennett R (ed.) 1969 *Classic Essays on the Culture of Cities*. Appleton-Century-Crofts, New York

Stimpson C, Dixler E, Nelsen M, Yatrakis K (eds.) 1980 *Women and the American City*. University of Chicago Press, Chicago

Stein J 1995 *Classic Readings in Urban Planning: An Introduction*. McGraw Hill, New York

Urban Affairs Association 1995 *Directory of Programs in Urban Studies and Urban Affairs*. UAA, Cleveland

Wirth L 1938 Urbanism as a way of life. *American Journal of Sociology* **44**: 11–24

R. LeGates

Urban System in Geography

In geography the term 'urban system' has a rather precise meaning: it refers to attempts to order and classify cities according to some underlying criterion or criteria. Thus, most simply, cities within any given geographical area—ranging from the local to the global—may be arrayed in terms of population size. More commonly, cities are classified in terms of the

economic functions they perform; in particular, the services they provide for themselves and their surrounding areas or spheres of influence. From this comes the concept of a 'central place system': a system in which cities are arrayed in terms of the range of services they provide and the areas over which they provide these services. In turn this has been extended into a general urban system that forms an important part of general location theory (q.v.).

1. Two Central Concepts: Rank-size and Central Place

1.1 The Rank-size 'Rule'

It has long been noticed that in many parts of the world, cities and towns are quite regularly spaced. In fact, for any two adjacent towns of similar size, the distance between them is given quite well by the formula $(P1/P2)/D = A$, in which $P1$ and $P2$ stand for the populations of the two towns, D is the distance between them, and A is a constant for any given region reflecting general population density; thus, it is greater in Iowa than in the United Kingdom, and greater in Iowa than in Indiana (Lösch 1954), but the distance between the smallest towns nevertheless seems constant for a wide range of densities. From this, it has also been shown that the distribution of towns by size follows a so-called rank-size rule: if cities are arrayed by population in descending order, then the size of the nth town is $1/n$ the size of the largest city, following a series: 1, 1/2, 1/3, 1/4 … $1/n$ (Stewart 1958).

Zipf, who first demonstrated this rule, produced a great deal of empirical justification for it, but also showed that there were many divergences from it (Zipf 1949): it works robustly for highly self-contained regions, but less so in regions where there is a great deal of interaction with other regions. This means not only that in many highly-developed parts of the world it works less well, but also that it may tend to break down over time as international trade increases. Many major cities, like London, Paris, and Tokyo, appear larger than they should do according to the 'rule' because they provide services for very large areas of the world, even the whole world; such cities are said to be 'primate' and the resulting population distribution is said to show 'primacy.' In other countries primate functions may be shared among two or more cities (for instance, New York and Washington within the United States, or Amsterdam, The Hague, and Rotterdam in the Netherlands, or Rome and Milan in Italy); such countries often exhibit a flattening of the top of the curve of city size distribution. Stewart's empirical examination (Stewart 1958) showed that the five largest cities in every country showed considerable divergence from the 'rule,' and that even in countries which fitted better (such as the United States and Brazil), smaller subdivisions (such as states) showed major deviations from it. Other work showed that France and Mexico showed considerable divergences from the rank-size distribution, while Canada and Italy were much closer to it (Browning and Gibbs 1961).

1.2 The Central Place Hierarchy

The idea of a size–class hierarchy is related to the rank-size concept, but in a very indirect way. First, it tends to have been derived theoretically, and to have been tested empirically only later. Second, by definition 'There is no coincidence, even theoretically, between the rank-size rule for cities and a functional size–class hierarchy of the type postulated by Christaller and Lösch. The former yields a smooth descending curve of town sizes; the latter a ladder of successively lower size classes of towns, grouped around a limited number of discrete normal values corresponding to discrete functional constellations' (Stewart 1958). Third, since the hierarchy is derived from the performance of functions, there is no necessary relationship between this measure and population size; the rather casual assumption seems to be that so many employees will be needed to provide the services, and that this total will relate generally to population numbers.

The real key to a possible relationship lies in the work of Christaller, the German geographer who first developed central place theory (Christaller 1966/1933). His basic assumption was that there was a constant relationship between the numbers of cities found at each level of the hierarchy. He suggested that the commonest relationship was one derived from service (market) areas around cities, which yielded the constant $k = 3$; thus, there would be one city at the topmost level, three at the next, nine at the next, and so on. But the theory also suggests a similar relationship in the sizes of cities: thus, second-level cities are one-third the size of the first-order city, and so on. Further, the market areas of all levels of city, which were hexagonal in shape because a hexagon is the figure derived from 'packing' circles to eliminate overlap or blank space, nested precisely within the larger hexagonal market areas of higher-order cities; the theory simultaneously established the sizes of cities, their relationships, and the sizes and shapes of their market areas.

Therefore, at $k = 3$, there is exact accordance with the rank-size rule. Berry and Garrison (1958a) comment that 'The Christaller scheme is consistent with Zipf's empirical observations, but consistency requires a rigorous choice of relationships.' In fact, Christaller also suggested that in some cases a different value for k might apply: thus along strong transport corridors, where the key was to maximize the number of settlements along transport lines, he suggested that a ratio of $k = 4$ might be more appropriate, while

Table 1
The Christaller central place system (1933)

Type	Market area Radius, km	Population of town	Population of market area	Centrality (Telephones/popn)
M (Marktort)	4.0	1,000	3,500	0.5–2.0
A (Amtsort)	6.9	2,000	11,000	2.0–4.0
K (Kreisstadt)	12.0	4,000	35,000	4.0–12.0
B (Bezirkstadt)	20.7	10,000	100,000	12.0–30.0
G (Gaustadt)	36.0	30,000	350,000	30.0–150.0
P (Provinzstadt)	62.1	100,000	1,000,000	150.0–1200.0
L (Landstadt)	108.0	500,000	3,500,000	150.0–1200.0

Source: Christaller (1966), 67; Dickinson (1967), 51.

political–social (administrative) centers were governed by the principle of separating regions administratively, suggesting a ratio of $k = 7$. So the correspondence with the rank-size rule applies only in the one special case, albeit the one that appears to be the most common one.

Christaller derived the theory from pure principles, but then sought to confirm it empirically. Curiously, he based this work largely on the number of telephone connections in each city, particularly their ratio to the city's population. He developed a sevenfold hierarchy (Table 1). His map shows that over wide areas the marketing principle appears to work satisfactorily, but it is modified in practice by transport lines and by basic variations in natural resources which affect population density. Further, especially at higher levels, political history may be significant: thus, Munich, which is set at the highest theoretical level (Landstadt), is in fact intermediate between this level and an even higher one, not represented in this region, Reichstadt or national capital, because Munich was for centuries the capital of the independent state of Bavaria before being embodied in a unified Germany in 1871.

Subsequent to Christaller, but before translation of his work into English, British and North American geographers made a number of studies of national urban hierarchies. In Great Britain the most notable are by Smailes (1944, 1946, 1947), Carruthers (1967) and Smith (1968, 1970, 1978). Below London, regarded as a unique case, Smailes distinguished only five hierarchical levels (major cities, cities, minor cities, towns, and subtowns; Carruthers distinguished eight, and Smith (who used a similar scheme to Carruthers') 10). All these studies were strictly empirical, and interestingly made no reference to Christaller. The North American studies tended to concentrate on low-density agrarian regions such as Iowa, South Dakota, Washington State, and Saskatchewan, and therefore did not distinguish a complete range of upper-order settlements; comparing Southwestern Wisconsin and Southern England, Brush and Bracey were able to distinguish only two levels of service center (Berry and Garrison 1958a, Berry and Parr 1988, Brush and Bracey 1955, Hodge 1965). However, Berry extended the principle to Chicago and its suburbs, where he found a systematic hierarchy of centers and subcenters, albeit strongly arrayed along major radial transportation lines (Berry 1963, Berry and Parr 1988).

1.3 Towards a General Urban-location Theory: The Work of August Lösch

Six years after Christaller, a German economist sought to extend his work (and that of others, such as Alfred Weber on industrial location and Johann Heinrich von Thünen on agricultural location) into a general theory of economic location. Specifically, he secured an ingenious combination of Christaller's market and transport principles: starting with a distribution of hexagonal market areas around a single (first-order) metropolitan center, he rotated them so as to obtain six sectors with many production sites, and six with only a few, their boundaries established by six through transport lines that radiated through the central metropolitan city. He could, thus, argue that city distributions and sizes are different in transport-rich and transport-poor sectors. (He argued that Christaller had been mistaken in distinguishing the market and transport principles; this was due to an accident, that when $k = 3$ all places lie symmetrically with respect to the six through transport lines, so that no thickly or thickly settled areas appear.) Depending on the basic assumptions made, the middle or two outer segments in transport-poor sectors would have few towns. But in addition, the immediate environs of the metropolis would be city-poor, because only a few local goods could be profitably produced there (Lösch 1954).

Lösch described the resulting arrangement as an economic landscape. He gave only two sketchy empirical illustrations, based on two midwestern American cities (Indianapolis and Toledo), and nowhere attempted a rigorous proof. And, because his system is much more complex than Christaller's (though superficially similar; Lösch does not accept Christaller's basic principle that lower-order areas nest within larger ones, but derives them all independently), it is inherently more difficult to demonstrate empirically.

However, subsequent work, notably that of Berry on Iowa and by Hodge on the northern Great Plains of Canada, suggests that in these quintessential agricultural areas, the central place system essentially developed along rail lines as they were built (Berry and Parr 1988, Hodge 1965).

2. *The Dynamics of Urban Systems*

Christaller himself recognized that central place patterns would change over time, for instance, because of new transport connections, especially in the modern era (Christaller 1966); the pattern of cities would be set by the original transport mode, but their relative significance might change as new modes emerged, and also as wider ranges of goods came to be traded. He wrote: 'We observe a very great frequency of places called towns and market spots; and we may consider the central places of ancient times, of the Middle Ages, and of modern times, the results of the economic conditions, values, and production of each period' (Christaller 1966). However, he did not try to demonstrate this rigorously. Subsequent work has suggested that urban hierarchies can change in three possible ways: in their initial establishment (top-down or bottom-up); later, through changes in the level at which an individual function is supplied; and through changes in the basic structure of the hierarchy, either in terms of the numbers of levels or the numbers of centers belonging to a given level (Berry and Parr 1988). Empirical work has shown, for instance, that in North America, improved transportation—especially the spread of automobile use—caused small trade centers to decline faster than large ones, though the decline did not increase regularly down the scale (Hodge 1965). Some work has been done to incorporate observations of this kind into formal models (Parr and Denike 1970, Mulligan 1980).

3. *From Central Place Systems to General Systems Theory*

The final question has to be whether central place systems can be generalized into General Systems Theory. A system can be regarded as containing: a set of objects (central places), their attributes (e.g., population); inter-relations among objects (e.g., locations); inter-relations among attributes (e.g., logarithmic relationships); and interdependencies between objects and attributes (e.g., a hierarchy). Systems can be closed and open. In a closed system, available energy is dissipated gradually until it becomes random through the system; it will have reached a state of maximum entropy, in which all trace of hierarchy had disappeared. Several attempts have been made to analyze central place systems in this way. An open system in contrast will settle into an equilibrium between the tendency to maximize entropy and the need for organization; this is known as a 'steady state.' Most central place systems correspond to this model, and it is expressed by rank-size regularity, which will tend to persist over time (Berry and Parr 1988).

However, the central place system is only a part of the wider urban system; 'there are many aspects of the economic and social organization of urban systems that are wholly beyond the scope of central-place theory' (Berry and Parr 1988). In 1977 Pred argued that 'In conventional geographic and planning literature on systems of cities excessive emphasis is placed on the hierarchical structure of interdependencies', which'is highly unrealistic as a description of the *total* structural composition of city-system interdependencies in advanced economies' (Pred 1977). This, Pred argued, was because in the Christaller system the largest urban unit was required to be completely self-reliant, because there could be no interdependence between large cities in the form of lateral exchanges of goods and services between second- or third-order cities, and because finally each 'nesting' arrangement below a single first-order places is effectively a closed system. The Lösch system performed rather better because it did allow lateral exchanges, but it too required the top-order city to be totally self-supporting. In the real world, 'the most critical growth-generating city-system interdependencies in economically advanced countries are largely the product of input-output relationships, or of job-control and decision-making relationships within multilocational business and government organizations' (Pred 1977).

Pred sought to demonstrate in detail how such interdependencies had developed historically in the American urban system between 1790 and 1910. In his account, access to specialized information plays a crucial and increasingly important role. Because the largest cities early acquire privileged access, even in the era before the telegraph or railroads, they tend to maintain it with each successive technological advance. This sets up a system of positive multipliers and positive feedback loops which tend to consolidate and enhance their position. Pred shows that to a remarkable degree, in advanced industrial countries the rank ordering of cities has remained stable over a long period; in the United States, where new cities arose during the western expansion, the rule is nevertheless true for each major regional division (Pred 1977). He concludes that there is a 'tendency of early established major channels of interdependence, or interurban growth transmission, to be self-reinforcing, thereby becoming increasingly entrenched and influencing subsequent channels of interurban growth transmission' (Pred 1977). However, city-systems in different countries do demonstrate very different patterns and trajectories, depending on different histories; there is an instructive contrast here between France, with

Table 2
The Loughborough group 'GAWC' inventory of world cities

A. ALPHA WORLD CITIES
12: London, Paris, New York, Tokyo
10: Chicago, Frankfurt, Hong Kong, Los Angeles, Milan, Singapore
B. BETA WORLD CITIES
9: San Francisco, Sydney, Toronto, Zurich
8: Brussels, Madrid, Mexico City, São Paulo
7: Moscow, Seoul
C. GAMMA WORLD CITIES
6: Amsterdam, Boston, Caracas, Dallas, Düsseldorf, Geneva, Houston, Jakarta, Johannesburg, Melbourne, Osaka, Prague, Santiago, Taipei, Washington
5: Bangkok, Beijing, Rome, Stockholm, Warsaw
4: Atlanta, Barcelona, Berlin, Buenos Aires, Budapest, Copenhagen, Hamburg, Istanbul, Kuala Lumpur, Manila, Miami, Minneapolis, Montreal, Munich, Shanghai
World city-ness values produced by scoring 3 for prime center status, 2 for major center status, and 1 for minor center status

Source: Beaverstock, Taylor and Smith 1999, Table 6.
Cities are ordered in terms of world city-ness values ranging from 1 to 12.

its highly centralized history, and Australia, whose city system essentially derives from its colonial past.

A key question concerns change at the top of the hierarchy. Most studies within the central place paradigm, including Christaller's, do not even provide specifically for a national metropolitan centre. But globalization appears to result in centers with a global range (Castells 1989, Sassen 1991). It would appear that centers are in effect dropping out at the bottom of the hierarchy and are being added at the top. Hall (1999) has suggested that the two bottom levels of the Christaller hierarchy have in effect disappeared from the hierarchy, becoming mere villages, while a global and a subglobal level have been added at the top. Research by Beaverstock and Taylor suggests that global-level cities can be distinguished both by key attributes (such as concentrations of specialized services) and by relationships (such as exchanges of personnel). The latter are more significant but are more difficult to measure, since data are often not available. But they use data on branch offices of specialized financial and business services to establish a new hierarchy based on an index of 'globality' (Table 2) (Beaverstock et al. 1999). The significance is that this research rather precisely illustrates the Pred hypothesis: that, in a world increasingly based on the exchange of specialized information, the relationships are increasingly horizontal in terms of the hierarchy of cities, and global in scale. Without doubt, central place systems have been a useful conceptual device, and they still illustrate important features of the modern world (for instance local trading and commuting systems). But, for a fuller understanding, the need for the future is to integrate them into a wider systems framework.

See also: Development and Urbanization; Urban Geography; Urban Growth Models

Bibliography

Beaverstock J V, Taylor P J, Smith R G 1999 A roster of world cities. *Cities* **16**: 445–58

Berry B J L 1963 *Chicago Commercial Reference Handbook*. University of Chicago Press, Chicago

Berry B J L 1964 Cities as systems within systems of cities. *Papers, Regional Science Association* **13**: 147–63

Berry B J L, Garrison W L 1958a Alternate explanations of urban rank-size relationships. *Annals of the Association of American Geographers* **48**: 83–91

Berry B J L, Garrison W L 1958b The functional bases of the central-place hierarchy. *Economic Geography* **34**: 145–54

Berry B J L, Parr J B 1988 *Market Centers and Retail Location: Theory and Applications*. Prentice-Hall, Englewood Cliffs, NJ

Browning H L, Gibbs J P 1961 Some measures of demographic and spatial relationships among cities. In: Gibbs J P (ed.) *Urban Research Methods*. Van Nostrand, Princeton, NJ, pp. 436–59

Brush J E, Bracey H E 1955 Rural service centers in Southwestern Wisconsin and Southern England. *Geographical Review* **45**: 559–69

Carruthers W I 1967 Major shopping centres in England and Wales, 1961. *Regional Studies* **1**: 65–81

Castells M 1989 *The Informational City: Information Technology, Economic Restructuring and the Urban-Regional Process*. Blackwell, Oxford, UK

Christaller W 1966/1933 *Central Places in Southern Germany* (Trans. Baskin C W). Prentice-Hall, Englewood Cliffs, NJ

Hall P 1999 The future of cities. *Computers, Environment and Urban Systems* **23**: 173–85

Hodge G 1965 The prediction of trade center viability in the Great Plains. *Papers and Proceedings of the Regional Science Association* **15**: 87–115

Lösch A 1954 *The Economics of Location*. Trans. Woglom W H, Stolper W F. Yale University Press, New Haven, CT

Mayer H M, Kohn C F (eds.) 1959 *Readings in Urban Geography*. University of Chicago Press, Chicago, pp. 230–39

Mulligan G F 1980 The effects of multiplier shifts in a hierarchical city-size model. *Regional Science and Urban Economics* **10**: 77–90

Parr J B, Denike KG 1970 Theoretical problems in central place analysis. *Economic Geography* **46**: 567–86
Pred A 1977 *City-System in Advanced Economies: Past Growth, Present Processes and Future Development Options*. Hutchinson, London
Sassen S 1991 *The Global City: New York, London, Tokyo*. Princeton University Press, Princeton, NJ
Smailes A E 1944 The urban hierarchy in England and Wales. *Geography* **29**: 41–51
Smailes A E 1946 The urban mesh of England and Wales. *Transactions of the Institute of British Geographers* **11**: 84–101
Smailes A E 1947 The analysis and delimitation of urban fields. *Geography* **32**: 151–61
Smith R D P 1968 The changing urban hierarchy. *Regional Studies* **2**: 1–19
Smith R D P 1970 The changing urban hierarchy in Wales. *Regional Studies* **4**: 85–96
Smith R D P 1978 The changing urban hierarchy in Scotland. *Regional Studies* **12**: 33–351
Stewart C T Jr 1958 The size and spacing of cities. *Geographical Review* **48**: 222–45
Zipf G K 1949 *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA

P. Hall

User Interface Design

This article discusses those design principles that enable the interface between a user and a device to form a synergistic partnership that is user centered and task oriented. The original term for this work was 'man–machine interface design.' However, the term 'man–machine' has been replaced by the more inclusive 'human–machine.' In turn, this term has given way to 'user interface design.' 'User' emphasizes the role the human plays in the system and also highlights the dominant framework for developing and applying interface design, namely user-centered design. 'User interface design' is concise and has currency in recent books by Barfield (1993), Cooper (1995), and Shneiderman (1998). The article is organized around four key concepts: affordances, user models, feedback, and error handling.

1. *Affordances*

When an animal perceives an object it will perceive fundamental properties that signal possible interactions. The size, shape, and solidity of a chair affords sitting for a person, but not for an elephant. A vine affords swinging for a monkey or Tarzan, but not for me. The concept of affordances originates with Gibson (1979), who was interested primarily in actions that take place in natural environments. The term affordance was extended to users interacting with artifacts by Norman in his influential books on user-centered design (1988, 1993, 1998). For Norman, affordances are the set of action–outcome pairings perceived by a user. Affordances derive from a user's past experience and provide obvious clues for how things operate. If the designer and user share the same knowledge, then actions can be signaled naturally, not arbitrarily: switches are for flipping and knobs are for turning.

Affordances are very important in the design of artifacts that users will walk up and use, or are manipulated directly. Users must resolve quickly the issues of what actions are possible, and where and how the selected action is performed. The closed door of an emergency exit affords opening, but the hardware should tell you to push, not to pull or slide. A broad, flat plate that affords little to grasp on the right side broadcasts where to push. In contrast, imagine a sliding door with the type of handle that usually affords grasping, twisting, and pulling. Harry Potter or Harry Houdini would have difficulty escaping such a malevolent design.

Constraints reduce the number of possible actions. Affordances and constraints, working in tandem, allow the user to discover the correct sequence of actions, even in a novel situation. Constraints can be physical, semantic, cultural, or logical (Norman 1988). If you want to leave our courtyard via the gate, the vertical handle affords gripping and the thumb lever just above it affords pressing. However, the crossbar rests and catches on the indented bracket and it is clear that the crossbar must be raised, by pressing on the thumb level, before pulling will do any good. This simple, two-step process is much harder for guests to get right when they first arrive because the physical constraint, the crossbar, is not visible from the opposite side of the opaque door.

Affordances offer a range of possible actions and should be made visible, not hidden. An actual affordance that is not also a perceived affordance contributes nothing to usability. The design of the human–computer interface often relies on conventions to convey affordances. The scrollbar on the side of a computer window affords vertical movement, but a novice may fail to perceive this affordance. These conventions enable experienced users to transfer their user models positively from one application to the next, but one must stretch to think of these conventions as visible and natural affordances to the novice user.

2. *User Models*

User models refer to the user's understanding of how the device can be used in order to perform tasks and accomplish goals. Good user models enable the user to determine the optimal procedures for accomplishing routine tasks, and to infer the solution to somewhat different or even novel tasks. Designers of user

interfaces rely on different techniques for revealing good models to users. A user model may emphasize the underlying technology, a metaphor, or the user's task (Cooper 1995).

2.1 Implementation Models from the Technology Paradigm

The technology paradigm is based on an understanding of how things work. If the mechanism is simple and visible this is an obvious choice. If the affordances lead a person to the right place, suggest the right action, and it is easy to see the consequences of that action, then the user has a very handy mental model. An implementation model even enables the user to repair the device, as I recently had to do with our courtyard gate. Unfortunately, the human–computer interface is often expressed through the technology paradigm in terms of how the software works. Thus, the user must learn how the program works in order to use it effectively. Once acquired, this is also a very handy mental model. The user could even modify the program to suit new needs or preferences. However, for most of us, the cost of gaining an implementation model will be horrendous. It is better to emphasize a functional model of how to use the tool to accomplish specific tasks.

2.2 Borrowed Models from the Metaphor Paradigm

A quick way to provide a user with a model of how a tool works is to draw an analogy to a tool the user already knows how to use. The similarities between the known and unknown allow the user to intuit use without understanding the mechanics of the hardware or software. The most famous metaphor of them all is the desktop metaphor invented by Xerox PARC and successfully implemented by Apple Macintosh. By definition a metaphor involves an implicit comparison, in which the tenor of comparison (e.g., computer screen) resembles the vehicle (e.g., desktop). The set of attributes shared between the tenor and the vehicle form the basis for the metaphor.

The metaphor paradigm for conveying a user model initially received reviews as stellar as the pig who built a brick house, but interface design pundits have huffed and puffed to flatten the metaphor paradigm (Cooper 1995, Norman 1998). Metaphors have several weaknesses. The set of features that are shared between tenor and vehicle may include irrelevant similarities that just clutter the user model. The useful ones are not necessarily the salient ones. There may also be dissimilarities that encourage inappropriate actions. There may be novel features of the tenor (computer interface) that are outside the scope of the vehicle that actually impede the user's discovery of these new features. Finally, metaphors may not scale as size and complexity increases. For example, keeping track of three or four documents on a messy desktop or computer screen is not too difficult, but when I'm working with dozens of articles, books, tables, figures, and printouts my desktop resembles a pigsty and there may be a better way to organize and use a similar electronic collection.

2.3 Task-oriented Models from the User Centered Paradigm

The important idea behind task-oriented models is that if there is a good tool with a good interface one can quickly, easily, and naturally learn how to use it. Interface objects and actions focus on the user's task regardless of how the device actually works.

Cooper's chapter 'Lord of the Files' provides one of the very best detailed examples of the tension between designing an interface based on the underlying technology vs. designing the interface from the viewpoint of the user's task. Consider the 'file' menu from any standard graphical-user interface. The menu options displayed for the user are focused on managing files, particularly the managment of the copy stored on the disk and the copy active in RAM. Many users of application programs have never developed a mental model that correctly reflects the relationship between the yin and yang of these two 'copies.' Nor should they, because the user's task is to generate a usable document. These users would benefit from a revised interface that hides the file management task (a computer implementation problem) and presents a single document that the user 'appears' to control from initiation, through revision, to completion. From a task oriented perspective, the 'file' menu should be replaced by a 'document' menu, where users can find options that correspond to current goals that are a frequent and integral part of document construction (e.g., options to rename a document, make a snapshot copy, move the document to a new location, or make a milestone copy).

3. Feedback

A basic principle of good design is that every user action should be followed by a visible reaction. When you push on the hand plate of an emergency door you hear it click, feel the bar move, and see the door open. Several feedback distinctions are useful: (a) real vs. abstract (Barfield 1993), (b) articulatory vs. semantic (Hix and Hartson 2000), (c) delayed vs. immediate (Shneiderman 1998), and (d) positive vs. negative (Cooper 1995).

When I rotate the switch that turns my car lights on at night I get real feedback when the lights come on. I receive concurrent, but abstract feedback, when the indicator light on the dash shows that the headlights

are on. Redundancy of real and abstract feedback can be good, because abstract feedback can exaggerate real changes that are subtle such as the differences between high- and low-beams, particularly at dusk. Abstract feedback is, of course, necessary when the real effect is remote or, by necessity, hidden from the user.

Articulatory feedback tells users their hands worked correctly, while semantic feedback tells them their heads worked correctly. For example, when the user selects 'open' from a pull-down menu, the choice may blink briefly just before the menu disappears and the dialogue box appears. Paap and Cooke (1997) recommend that menu panels should provide feedback indicating: (a) which options are selectable, (b) when an option is under the pointer and, therefore, can be selected, (c) which options have been selected so far, and (d) the end of the selection process. The latter three provide articulatory feedback that helps the users select the intended options. The first is more semantic: if the desired option is not available for selection then the user has probably made a navigation or mode error.

In an ideal world our interfaces would guide us with continuous and immediate feedback. In the real world some delays are inevitable. Whenever the system is performing a lengthy process the user should be given feedback, a status indicator, that processing is ongoing and not suspended. It is most useful to display an indicator showing what portion of the task has been completed.

Both Cooper (1995) and Norman (1988, 1993) are strong advocates of using more positive feedback and less negative feedback. They argue that designers should try to see users as making imperfect progress toward their goals, and that error messages should be friendly, make clear the scope of the problem, and provide alternative solutions (not just OK buttons). Don't think of the user as making errors; think of their actions as approximations to what is desired. Even the quintessential error recovery mechanism, the undo command, can be reconceptualized as a method of exploration.

4. Error Handling

The distinction between positive and negative feedback brings us to the special topic of errors. An interface with a good design will help users to detect errors, diagnose the cause of errors, assist error recovery, and prevent errors in the first place. In a dynamic system, many dangerous states can be avoided by forcing the operations to take place in proper sequence. Opening the door of a microwave oven switches the oven off automatically. Norman (1988) and Barfield (1993) have useful discussions of different types of forcing functions that prevent dangerous errors from occurring.

4.1 Slips: Errors of Action

A slip occurs when the user has formed the correct intention, but the appropriate action is waylaid on the way to execution. Slips often occur in the context of actions that are highly automatic and under unconscious control (Norman 1988). As such, slips are more likely to emanate from highly skilled users. Slips should be easy to detect because there will be an obvious discrepancy between the goal and result. However, detection is likely only if there is effective and immediate feedback.

Slips include the infamous mode error. Mode errors occur when devices have different modes of operation and the same action produces different results depending on the mode. Mode errors can be benign in digital watches or catastrophic in automatic pilots. They are more likely to occur when tasks are interrupted and time elapses between mode setting and action selection. Mode errors can be prevented either by eliminating modes or by making the mode visible and attention catching. Cooper (1995) offers a cogent defense of well-designed modes.

Description errors are slips caused by performing the right action on the wrong object. When driving in Australia, I usually signal my intention to turn by flipping the left-hand lever, thereby turning on the windshield wipers. Consistency may be the hobgoblin of little minds, but violations of consistency will constantly frustrate the ingrained habit of skilled users. Relative frequency also matters. A high-frequency object may surreptitiously replace the intended target, particularly if they are similar. For example, the frequently summoned file 'mybook' accidently replaces the newer file 'mydoodle', resulting in the unintentioned request to 'delete mybook.' Potentially critical errors like this usually are 'prevented' by confirmation boxes. However, when confirmations are issued routinely, users get used to approving them routinely, even when the confirmation warns of an impending disaster. When experts are working on automatic pilot, it is sometimes better to make sure they can recover from slips easily. The Macintosh trash basket and the Windows recycle bin serve such a purpose.

4.2 Mistakes: Errors of Intention

Mistakes, unlike slips, are errors of intention and are often caused by an incorrect user model. Affordances and feedback should help form correct user models, but even error detection is difficult for general purpose technology that is powerful and flexible. When there are a large number of ways of accomplishing a large array of goals, there will be countless numbers of action sequences that are perfectly legal, but most of them will not move the user closer to task completion. If I turn right on Triviz, it is an error because I will eventually run into a construction detour, but nothing

on my car's dash flags that error, because it has no knowledge of where I want to go and no knowledge of current route conditions. Specialized technologies, such as GPS navigation systems, can have user-centered, task-oriented designs. The user and the interface share goal information and, consequently, errors can be detected and solutions prompted.

The general-purpose PC and its powerful applications programs are plagued with the same problem in error detection. Having evolved within the technology paradigm, the interface detects when user actions violate the program, but the interface offers little or no help as users perform a complicated series of legal actions that the program slavishly follows. The user must detect errors by inspecting the document to see if the intended changes have been made. A specialized program for preparing psychology research articles in the format prescribed by the American Psychological Association, could detect and prevent far more errors and offer more useful guidance because it would have far more knowledge of the task at hand. These considerations play a large role in recent calls to shift from general-purpose devices to information appliances (Norman 1998, Mohageg and Wagner 2000).

5. *Evaluating Usability*

Usability engineers and designers should get involved early and often. Landauer (1995) makes an impressive case for the economic benefits of usability in product development. Guidelines such as those discussed in this article need to be augmented by expert reviews and usability testing. There are a variety of expert-review methods (Shneiderman 1998), of which, Nielson's (1994) heuristic evaluation method may be the most popular. Heuristic evaluation involves having three to five usability experts examine an interface with respect to ten usability heuristics or guidelines. Heuristic evaluation is vastly superior to no review, but not as effective as user testing (Desurvire 1994). User-centered design should include tests with real users doing real work. The benefit of such activities is simply enormous, almost always over 20 to 1, and often in the 200 to 1 range (Landauer 1995).

See also: Computer-assisted Instruction; Ergonomics, Cognitive Psychology of; Human–Computer Interaction; Human–Computer Interface; Intelligent Tutoring Systems; Mental Models, Psychology of

Bibliography

Barfield L 1993 *The User Interface: Concepts and Design*. Addison-Wesley, Reading, MA

Cooper A 1995 *About Face: The Essentials of User Interface Design*. IDG Books Worldwide, Foster City, CA

Desurvire H W 1994 Faster, cheaper! Are usability inspection methods as effective as empirical testing? In: Nielsen J, Mack R L (eds.) *Usability Inspection Methods*. Wiley, New York, pp. 173–202

Gibson J J 1979 *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston

Hix D, Hartson H R 2000 *Developing User Interfaces*. Wiley, New York

Landauer T K 1995 *The Trouble with Computers: Usefulness, Usability and Productivity*. MIT Press, Cambridge, MA

Mohageg M F, Wagner A 2000 Design considerations for information appliances. In: Bergman E (ed.) *Information Appliances and Beyond*. Academic Press, San Diego, CA

Nielson J 1994 Heuristic evaluation. In: Nielsen J, Mack R L (eds.) *Usability Inspection Methods*. Wiley, New York

Norman D A 1988 *The Psychology of Everyday Things*. Basic Books, New York

Norman D A 1993 *Things That Make Us Smart*. Perseus, Reading, MA

Norman D A 1998 *The Invisible Computer*. MIT Press, Cambridge, MA

Paap K R, Cooke N J 1997 Design of menus. In: Helander M, Landauer T K, Prabhu P (eds.) *Handbook of Human–Computer Interaction*, 2nd edn. Elsevier Science, Amsterdam

Shneiderman B 1998 *Designing the User Interface*. Addison-Wesley/Longman, Reading, MA

K. R. Paap

Utilitarian Social Thought, History of

Standard utilitarian theory encompasses two essential elements: (a) the rightness/wrongness of an action is determined by the goodness/badness of its consequences, and (b) the only thing that is good in itself is pleasure and the only thing bad in itself is pain, and happiness is taken to be the sum of particular pleasures (Quinton 1988, p. 1). Implicit in this ethical hedonism is the premise that the motives to action are rooted in self-interest. Based on these elements, the doctrine is then expressed in the form of the greatest happiness principle or the principle of utility. In *A Fragment on Government* (1776) the legal philosopher Jeremy Bentham (1748–1832) declared it 'a fundamental axiom' that 'it is the greatest happiness of the greatest number that is the measure of right and wrong,' and that 'the obligation to minister to general happiness, was an obligation paramount to and inclusive of every other' (Bentham 1977, pp. 393, 440–1). In 1781 Bentham coined the name utilitarian for this theory.

1. *Origins of the Theory*

Although Bentham is usually cited as the founder of utilitarianism, the antecedents of utilitarian principles have a far older vintage in the philosophy of Plato, Aristotle, and Epicurus, and in early Christian thought

(Scarre 1996, Chap. 2). Other significant dimensions of the theory can be traced to the seventeenth century writings of Thomas Hobbes, John Locke, and Richard Cumberland. It was the 'moral sense' philosopher Francis Hutcheson who first spoke of 'the greatest Happiness of the greatest Numbers' as a principle of moral conduct in *An Inquiry Concerning the Original of Our Ideas of Virtue or Moral Good* in 1725 (Shackleton 1972).

1.1 Religious Utilitarianism

In the eighteenth century, utilitarian social thought began to assume its classic form in the moral philosophy of the religious advocates of the utility principle. Advocates of religious utilitarianism included John Gay (1699–1745), John Brown (1715–66), Edmund Law (1703–87), Abraham Tucker (1705–74), and William Paley (1743–1805). The ethical hedonism developed by these moralists was premised on the existence of a God who would punish and reward in the afterlife according to the degree to which individuals acted to enhance general happiness in the present life (Albee 1902, Crimmins 1998). Among the religious exponents of the theory, only Paley offered a comprehensive and systematic social philosophy founded upon the utility principle, encompassing morals, government, law, and political economy, in his popular *Principles of Morals and Political Philosophy* (1785).

1.2 Secular Utilitarianism

Bentham's secular version of the theory was derived from a more progressive strain of enlightenment thought: empirical, analytical, and antagonistic towards organized religion. Unaware of Hutcheson's formulation of the utility principle, Bentham attributed it to the Italian law reformer Cesare Beccaria (1738–94) and sometimes to the dissenter and natural philosopher Joseph Priestley (1733–1804). Other progenitors to whom he traced his intellectual inheritance included David Hume (1711–76) and Claude Helvétius (1715–71). In Hume's secular moral philosophy the source of the rules of justice were located in general utility, but he explained the psychology of moral conduct in terms of a theory of 'moral sentiments' rather than motives supplied by pleasures and pains. Helvétius provided a systematic if rudimentary explanation of human motivation in terms of utility perceptions and expectations and some useful suggestions for the application of this knowledge to law and other social questions.

2. Bentham's Science of Morals and Legislation

In *An Introduction to the Principles of Morals and Legislation* (1789), Bentham delineated the component parts of his 'science' of morals and legislation founded upon the utility principle. The general purpose of the science was 'to rear the fabric of felicity by the hands of reason and of law.' (Bentham 1970, p. 11) The *Introduction* was then considered by Bentham to be the foundational text of his philosophy and his plans for future works were intended to elucidate the practical application of the utility principle in the areas of civil and penal law, judicial procedure, legislative reward, constitutional law, legislative procedure, international law, finance, political economy, and 'universal jurisprudence' (Bentham 1970, pp. 5–6).

2.1 The Principle of Utility

At the base of Bentham's theory was the assumption of 'the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do' (Bentham 1970, p. 11). As a first principle used to establish everything else, the utility principle was not itself subject to proof, but Bentham claimed there was sufficient testimony to the fact that it had always been generally accepted as the standard of right and wrong (Bentham 1970, p. 13). This foundational principle referred to a feeling or sentiment which served to approve the utility of an action. As Bentham explained, 'The principle here in question may be taken from an act of the mind; a sentiment; a sentiment of approbation; a sentiment which, when applied to an action, approves of its utility, as that quality of it by which the measure of approbation or disapprobation bestowed upon it ought to be governed' (Bentham 1970, p. 12). Utility also meant public utility, and the utility of the individual was a part of the public utility in which the individual shared. Thus a person approved of an act or law to the degree to which the happiness of all the individuals affected by it was advanced. For Bentham, the relationship between happiness and pleasure and pain was straightforward and did not require elaboration: pleasure contributes to happiness, while pain detracts from it. In these terms the public interest is constituted of the aggregate of individual interests, construed in terms of the balance of pleasures and pains experienced by each person.

2.2 Motives and Consequences

All motives have their source in the anticipation of pleasure or pain. Bentham delineated four *sanctions* or sources of pain and pleasure, which he may have learnt from John Gay: physical, political, moral, and religious. These sanctions are available to the moralist and to the legislator in guiding and determining an individual's moral conduct, and explain how an essentially self-interested individual can be directed to perform actions which enhance general happiness. In Bentham's theory of motives the dictates of utility were neither more nor less than the dictates of the most

extensive and enlightened benevolence. Other motives were categorized either as 'semisocial' motives (such as love of reputation, desire of amity, and religion), which may or may not coincide with the principle of utility, and purely 'self-regarding' motives (such as physical desire, pecuniary interest, love of power, and self-preservation), which were the least likely to coincide with utility. All motives, including the most extensive benevolence, are rooted in self-interest in one form or another; in these terms, the notion of a *dis*interested motive is implausible.

While the consequences of an act are partly dependent on the motives which give birth to the act (and partly on the empirical circumstances in which the act is performed), in Bentham's theory the utility of an act is not dependent on its originating motive(s). The relative utility of an act is determined solely by its consequences: the advantages and/or disadvantages that result. When deciding whether to act or which act to perform the individual must calculate as best as he or she can the pleasures and pains which may reasonably be expected to accrue to the persons (including him- or herself) affected by the acts under consideration. Bentham's *hedonistic* or *felicific calculus*, by which the value of pleasure and pain may be measured was constituted of the following elements: intensity, duration, certainty or uncertainty, propinquity or remoteness, fecundity, purity, and extent. A similar calculation should guide the legislator in formulating laws. However, Bentham recognized that it was not normally feasible for an individual to engage in such a calculation prior to acting, hence he spoke of the general tendencies of actions to enhance happiness as a sufficient guide. The consequences of acts were central to Bentham's theory of punishment, in which the utilitarian objective was to ensure that the punishment was in proportion to the mischief produced by a crime and sufficient to deter others from committing the same offence. To this end Bentham developed a series of rules to guide the lawmaker in the construction of a utilitarian penal code, including the elements involved in the calculation of the mischief caused by an offence and the appropriate punishment.

3. *John Austin*

Although sometimes described as an act utilitarian, it is generally acknowledged that for Bentham properly constituted *rules*, based on calculations of general utility, ought to be the guides for human action (Baumgardt 1966). John Austin (1790–1859), a disciple of Bentham's jurisprudence, is usually cited as the first explicit rule utilitarian. However, in *The Province of Jurisprudence Determined* (1832), he made it clear that he learnt the significance of rules from Paley, from whom he also borrowed the theological elements of his theory.

4. *James Mill*

James Mill (1773–1836) has often been credited with influencing Bentham's conversion to democratic principles in 1808–9; however, Bentham had already developed utilitarian arguments for democratic reform in manuscripts written around the time of the French Revolution. What Mill certainly brought to the political cause of utilitarianism was clarity of thought and expression, which he demonstrated to telling effect in his *Essay on Government* (1819). Based on the premise that government has the purpose of producing the greatest happiness of the greatest number, in this essay Mill made a succinct case for the radical reform of British political institutions (universal manhood suffrage, secret ballot, and annual parliaments). The essay was roundly criticized by T. B. Macaulay in the *Edinburgh Review* in 1829, largely for its narrow view of human nature and the *a priori* science of politics that it contained.

5. *John Stuart Mill*

It was largely through the sympathetic criticism of John Stuart Mill (1806–73) that Victorian moralists became acquainted with Bentham's utilitarian theory, though Mill's own moral philosophy was also influenced by Austin. Partly in response to the challenge of Macaulay's criticism of his father and partly the result of his own reflections on his intellectual inheritance, Mill developed a revised version of utilitarian theory. This revisionism took concrete form in his essay 'Bentham' (1838) and received definitive expression in *Utilitarianism* (1861). In the latter, Mill sought to defend the basic hedonism of utilitarian theory while offering refinements designed as responses to critics of the doctrine. The most problematic of these is the distinction between the *qualitative* differences of types of pleasures. While adhering to the basic Benthamic analysis of motives, Mill introduced the concept of higher pleasures, by which he meant the pleasures of the intellect, which he claimed were intrinsically more desirable than other pleasures. However, this qualitative distinction undermined the aggregative dimension of the theory and introduced questions concerning the appropriate judge of which pleasures were preferable. Mill only raised further questions about the integrity of the theory by arguing that the constituent elements of an individual's happiness might encompass objectives such as the pursuit of truth, beauty, and virtue, without consideration for the utility they may bring. This was a view later adopted by G. E. Moore (1873–1958) in the *Principia Ethica* (1903). Rejecting the hedonism of Bentham and Mill, Moore set forth a 'consequentialist' theory now known as *ideal* utilitarianism, in which the 'good' to be attained is more expansive than the summation of pleasures, and contains within itself

certain objective features of the universe, pre-eminently the beauty of its constituent parts (Scarre 1996, p. 114).

To the criticism that utilitarianism sacrificed principles for expediency, Mill replied that it was expedient to observe the *rules* associated with general utility and implied that only saints could be utilitarians, able to calculate continually the consequences of their actions. Finally, the supposed conflict between utilitarianism and Mill's commitment to individualism and individual liberty in the essay *On Liberty* (1859) is resolved when liberty is defended, not for its own sake but on the utilitarian grounds of the benefits which accrue to the individual and to the society of which the individual is a member.

6. *Liberal Utilitarianism*

Utilitarianism has enjoyed a sustained if ambiguous presence within the liberal tradition. Bentham's 'liberalism' is largely based on his writings on civil law and political economy, where he introduced subordinate principles to guide the legislator in giving effect to the utility principle: security, subsistence, abundance, and equality, in this order of priority. The stress on the *security* of the individual and *security of expectations* gave a priority to individual liberty when crafting law and tended to introduce a conservative element into the legislator's reform agenda, while the egalitarian principle held out the prospect of redistributive policies. Which of the subordinate principles was most in accord with the dictates of utility at any one point, and therefore should be acted upon by the legislator, depended on the particular context and particular issue. To the 'security-providing' principle Bentham later added the 'disappointment-preventing' principle, based on which disappointed expectations (loss of public office, for example) were to be compensated in the reform process.

Commentators who emphasize the importance of these subordinate principles tend to view Bentham as a *liberal* utilitarian, who intended that laws should be modeled to facilitate individuals in the pursuit of happiness in ways they, rather than the legislator, thought appropriate (Kelly 1990). This interpretation contrasts with the traditional view of Bentham as an aggregative legislative utilitarian (Crimmins 1996). Mill is also interpreted as a *liberal* utilitarian, whose theory of justice stressed the utility value of *security* (Riley 1998). This placed definable limits on the functions of law and implied a gradualist approach to achieving optimal utilitarian outcomes through social and political reform. Nevertheless, in a society of literate people neither Bentham nor Mill saw any barriers to the introduction of democratic institutions, including the extension of the suffrage to women on equal terms with men. Moreover, in later life Mill inclined toward collectivist solutions to social problems (Riley 1996).

In *Man versus the State* (1884) and other writings Herbert Spencer (1820–1903) employed the utility principle to enhance the liberty of the individual, defend the existing social order, and attack the drift towards socialism and 'slavery.' James Fitzjames Stephen (1829–94), Henry Sidgwick (1838–1900), and A. V. Dicey (1835–1922) all advocated versions of utilitarian individualism, although Sidgwick occasionally gave voice to 'socialist' sentiments in developing his *intuitional* utilitarian theory in *The Methods of Ethics* (1874). Reform-minded liberals such as J. A. Hobson (1859–1940) and L. T. Hobhouse (1864–1929) found in the protean nature of utilitarianism a justification for distinctly nonindividualist policies. Thus Hobhouse in *Liberalism* (1911) incorporated utilitarianism into a new liberal discourse that, in addition to laissez-faire economics (and notwithstanding Herbert Samuel's quasi-Benthamic motto 'the greatest liberty of the greatest number') also bridged elements of socialism, social Darwinianism, and the idealism of T. H. Green, and out of which emerged the beginnings of Britain's welfare state.

7. *The Modern Debate*

One of the most effective critiques of utilitarianism was mounted by John Rawls in *A Theory of Justice* (1971). Rawls argued that in counting *only* interests and aggregating *across* persons to arrive at the appropriate action or policy the utilitarian fails to 'take seriously' the distinction between persons. In reply it could be argued that at the level of private ethics utilitarian calculations implicitly acknowledge that different and distinct persons are involved in most situations about which individuals have to make moral judgements. This means that a consistent utilitarian will treat the interests of different people affected by his or her actions as of equal importance. However, the problem with this reasoning is that it risks making utilitarianism primarily a theory of equal consideration and effectively relegates the aim of maximizing utility to the status of a byproduct, not the ultimate ground of the theory.

The implications of an aggregate-utilitarian calculation for societies in which unpalatable sacrifices appear to be determined by simple calculations of optimal utility are frequently noted by critics of the theory. In an indiscriminate calculation where only total quantity of pleasures count, minorities are especially vulnerable to risk. Although Bentham was aware of the problems associated with the potential exclusion of minority interests under the phrase 'the greatest happiness of the greatest number,' the literature on utilitarianism abounds with discussions of the

potentially negative consequences of employing a rigid act utilitarian calculation to determine the individual's appropriate course of action or the legislator's policy choice. This debate is represented in the exchange between the modern act utilitarian J. J. C. Smart and his critic Bernard Williams in *Utilitarianism: For and Against* (1973). Other modern utilitarians who defend the theory as a practical guide to moral conduct and to public action include Robin Barrow in *Utilitarianism: a Contemporary Statement* (1991) and Richard B. Brandt in *Morality, Utilitarianism, and Rights* (1992).

See also: Aristotelian Social Thought; Bentham, Jeremy (1748–1832); Consequentialism Including Utilitarianism; Eudaemonism; Liberalism: Historical Aspects; Marxist Social Thought, History of; Mill, John Stuart (1806–73); Nostalgic Social Thought; Pragmatist Social Thought, History of; Utilitarianism: Contemporary Applications; Weberian Social Thought, History of

Bibliography

Albee E 1902 *A History of English Utilitarianism*. Swan Sonnenschein, London and Macmillan, New York

Barrow R 1991 *Utilitarianism: A Contemporary Statement*. Edward Elgar, Aldershot, UK

Baumgardt D 1966 *Bentham and the Ethics of Today*. Princeton University Press, Princeton, NJ

Bentham J 1970 *An Introduction to the Principles of Morals and Legislation*. Athlone Press, London, UK [*The Collected Works of Jeremy Bentham*. In progress: Athlone Press, London and Clarendon Press, Oxford, UK]

Bentham J 1977 *A Comment on the Commentaries and a Fragment on Government*. Athlone Press, London [*The Collected Works of Jeremy Bentham*]

Brandt R B 1992 *Morality, Utilitarianism, and Rights*. Cambridge University Press, Cambridge, UK

Crimmins J E 1996 Contending interpretations of Bentham's utilitarianism. *Canadian Journal of Political Science* **29**: 751–77

Crimmins J E (ed.) 1998 *Utilitarians and Religion. Part I. Religious Utilitarians Part II. Secular Utilitarians*. Thoemmes, Bristol, UK

Halévy E 1972 *The Growth of Philosophic Radicalism* [*La Formation du Radicalisme Philosophique*, 1901–4], trans. Morris M 1928. Augustus M. Kelley, London

Kelly P J 1990 *Utilitarianism and Distributive Justice: Jeremy Bentham and the Civil Law*. Oxford University Press, Oxford, UK

Mill J S 1962 *Utilitarianism*. In: Warnock M (ed.) *Utilitarianism, on Liberty*. Essay on Bentham. Together with selected writings of Jeremy Bentham and John Austin. Fontana Press, London

Quinton A 1989 *Utilitarian Ethics*, 2nd edn. Open Court, La Salle, IL

Riley J 1996 J. S. Mill's liberal utilitarian assessment of capitalism versus socialism. *Utilitas* **8**: 39–70

Riley J 1998 Mill on justice. In: Boucher D, Kelly P (eds.) *Social Justice: From Hume to Walzer*. Routledge, London

Scarre G 1996 *Utilitarianism*. Routledge, London

Shackleton R 1972 The greatest happiness of the greatest number: the history of Bentham's phrase. *Studies on Voltaire and the Eighteenth Century* **90**: 1461–82

Smart J J C, Williams W 1973 *Utilitarianism: For and Against*. Cambridge University Press, London, UK

J. E. Crimmins

Utilitarianism: Contemporary Applications

Utilitarianism is a moral and political theory that can be characterized as a combination of a theory of welfare and a concern with welfare as the good. Although there are many forerunners, utilitarianism arose in the Scottish Enlightenment. Thereafter, it was forcefully articulated by Jeremy Bentham (1970), who applied it to the design of legislatures and other institutions and to legislation and law; it was applied to law by John Austin (1954); and it flourished in the writings of John Stuart Mill (1969), Henry Sidgwick (1907), and many nineteenth- and twentieth-century economists. The focus on welfare as the good has been virtually constant from the beginning, although the scope of that concern sometimes expands to cover animals other than humans. The theory of welfare, however, has undergone frequent, substantial changes, largely because it is central to the discipline of economics.

Surprisingly often, sophisticated academics suppose that utilitarianism is solely about self-interest, that it is the moral theory of egoism. Among anti-utilitarian moral theorists, it has been characterized as virtually the opposite of egoism: as beneficence run wild. Sidgwick (1907) supposed that egoism and utilitarianism were two great, contrary divisions of moral theory.

Utilitarianism had been the chief moral theory in law for more than a century until recent decades, which have seen the rise of competing rights talk (Hart 1983). The movement of law and economics, which arose at about the same time as the move toward rights talk and which has come to dominate many areas of the civil law, has continued the utilitarian tradition and has even displaced rights talk in many contexts.

Utilitarianism is the principal moral stance of contemporary economists, although the more common term may now be welfarism. The focus of much economic thought since before the time of Adam Smith has been on what the economy does for the welfare of people in a market society. From the time of the Scottish Enlightenment through the nineteenth century, economics and utilitarianism developed together in Anglo–Saxon thought, with many authors contributing to both. During the twentieth century, philosophical utilitarianism became disconnected

from economic reasoning, and it therefore lost contact with further developments of value theory in economics as—some would say—it also lost contact with morality in the real world.

1. Value Theory

The split from economics was a misfortune for utilitarian thought in philosophy. The sometime theory of welfare of Bentham and many utilitarians, including such economists as F. Y. Edgeworth (1881), tended to conceive welfare or utility as cardinal and additive across individuals. From Vilfredo Pareto through the 1930s, there was a revolution in value theory in economics that jettisoned the interpersonal comparisons required for summing across individuals and then jettisoned cardinal-value theory altogether. Contemporary economics is largely grounded in an ordinal-value theory that merely ranks states of affairs and does not allow interpersonal comparisons.

If we cannot aggregate across individuals, then we are limited to making ordinal Pareto comparisons in which we can say that one state of affairs, X, is better than another, Y, only if at least one person is better off in X than in Y and no one is worse off in X than in Y. The beauty of a Benthamite additive utility is that, in principle, it would allow us to give complete rankings of all states of affairs. In Pareto's world of no interpersonal comparisons or aggregations, we cannot rank two states of affairs if some people are better off in one of them and others are better off in the other. Since it is relatively implausible that any significant social or economic policy could fail to make some worse off than in the status quo, we seem to be restricted to living with the status quo.

Contemporary policy analysts solve this restriction to the status quo by using cost–benefit analysis to commend many policies, such as public expenditures on highway programs. In such analyses, the costs and benefits to all are added, and a policy is adopted only if it has a positive sum—especially if that sum is relatively large. On the surface, this sounds like a cheat, as though a century of value theory were being ignored. Even those who accept only the more refined ordinal-value theory without interpersonal comparison, however, might agree *ex ante* that the status quo is worse for us individually than the *expected value* of an alternative. It would therefore be mutually advantageous for us to create institutions to make policy choices that would allow us to reach a new state, the expected value of which is greater. From an ordinal stance, we can choose *ex ante* to have an institution that makes some decisions on the basis of some kind of cardinal assessment, even though we could virtually always be sure of any project that it would cause harm to some.

Often we could recognize directly that a policy would serve mutual advantage *ex ante*. For example, vaccination against a grim disease might be expected to cause harm to a few but to protect virtually everyone else from the disease. *Ex ante*, I do not know whether my child will be one of those harmed or one of those benefited, but I might have a good sense of the probable advantages to my child, and those advantages might weigh heavily in favor of adopting the vaccination policy. If the economic cost of vaccination is negligible, as it generally has been, this approximates a simple cost–benefit analysis in which both the costs and benefits are measured in lives. In the case of a disease that has only human carriers, as is true of smallpox, it is even possible to eradicate the disease altogether. This has been done with smallpox and may soon be done with polio.

2. Institutions

More generally, we create mutual-advantage institutions, which are utilitarian in an ordinal welfarist sense (Hardin 1988). The policy-making institutions we create could adopt a policy of using cost–benefit analysis for many classes of policy problems. In this two-stage account, the use of cost–benefit analysis serves mutual advantage *ex ante* and is therefore ordinally welfarist, despite being based on cardinal, interpersonally comparable values. Any application of the method will lead to policies that harm some while benefiting others, but the adoption of cost–benefit analysis by the institution will be a mutual advantage move *ex ante*. We can make similar arguments for the creation of constitutional arrangements, a criminal justice system, and civil law. *Ex ante*, they can all serve mutual advantage and can therefore be ordinally utilitarian.

Even economists have difficulty with ordinal thinking in many contexts, despite the facts that they think ordinally regularly in their microeconomic theorizing and that they sometimes insist that cardinally additive reasoning is philosophically incoherent. Yet they commonly fall back on cardinal thinking when they do policy analysis. Since much of their policy analysis addresses questions of how to allocate money to various programs, they have devices that make their problems approximate merely the addition and comparison of numbers of dollars. Despite massive criticism of these devices by many non-economists and some economists, it is hard to imagine doing policy analysis or justifying policies at all in some realms without such additions. It is typical to criticize standard cost–benefit analysis for disregarding non-monetary values but to give no alternative scheme for handling the pressing issues such analysis is used to address (Gutmann and Thompson 1996). It is a great advantage of a theory that has real content that it can be criticized in such specific ways. One wishes the alternative theories were sufficiently articulate to be susceptible to criticism.

An inherent implication of an ordinal, as opposed to a Benthamite cardinal, utilitarianism is that we must be more modest in our claims about the good we might do in the world, because our theory often yields ambiguous implications—or, essentially, no implications at all—for what policy we should adopt. This is a conclusion that all moral theorists might well take to heart. If moral debate allowed greater play for ambiguity and indeterminacy, it would have a less assertive and dismissive quality. It would then seem more moral. It would also be more relevant. In any case, on our present understanding of personal welfare, there is little prospect of doing cardinal comparisons and aggregations of welfare over populations in many important policy contexts. It is therefore unfortunate that, although we regularly do it in daily life at the personal and small-number level, thinking ordinally is especially difficult when we think of broad social policy. There are easy cases, such as vaccination, but when the values at stake are not so simply uniform across an entire population, we have much greater difficulty making the comparisons.

For some policies, we can identify the losers *ex ante* and therefore cannot say that these policies are ordinally utilitarian. We can meet this objection by compensating anyone who evidently is made worse off by a policy. There are two substantial problems with such compensation. First, the promise of compensation is likely to provoke exaggerated claims that one is made very much worse off. Compensations, therefore, tend to be standardized as though all have similar values. *Ex ante*, this might be acceptable, and we can therefore make serious policy choices. *Ex post*, such standardization may seem unfair for anyone with unusual preferences or beliefs (Calabresi 1985).

Second, compensation to anyone who is made worse off by any change at all cannot be a plausible program, because it would entail compensating people for the economic failures that are a natural implication of technological and other economic changes. If the state must compensate every bankrupt, the economy will stagnate. One might suppose that the economy is more nearly a force of nature than a matter of public choice and that therefore we need not be concerned with compensation for the harsh results of the market. If, however, we could compensate for market harms and yet do not, we implicitly permit trade-offs across individuals for the sake of greater market productivity overall. Anyone who thinks that economic progress is good in at least some states of development must countenance losses to some for the perhaps greater gains of others.

The chief contemporary use of utilitarian reasoning is in *ex ante* specification of institutional structures and devices, not in direct application to immediate policy choices. Even cost–benefit analysis, which might seem to be the preeminent use of utilitarian reasoning is, in fact, justified not directly as though it were directly a utilitarian calculation. Rather, it is justified only *ex ante*. For a very simple case such as vaccination, direct calculation works because we can make ordinal judgments. For more complex problems, it does not work, because we cannot make such judgments.

See also: Bentham, Jeremy (1748–1832); Mill, John Stuart (1806–73); Pareto, Vilfredo (1848–1923); Rational Choice and Organization Theory; Rational Choice in Politics; Rational Choice Theory: Cultural Concerns; Rational Choice Theory in Sociology; Utilitarian Social Thought, History of

Bibliography

Austin J 1954 *The Province of Jurisprudence Determined*. Noonday, New York

Bentham J 1970 *An Introduction to the Principles of Morals and Legislation*. Methuen, London

Calabresi G 1985 *Ideals, Beliefs, Attitudes, and the Law: Private Law Perspectives on a Public Law Problem*. Syracuse University Press, Syracuse, NY

Edgeworth F Y 1881 *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. C Kegan Paul, London

Gutmann A, Thompson D 1996 *Democracy and Disagreement*. Harvard University Press, Cambridge, MA

Hardin R 1988 *Morality within the Limits of Reason*. University of Chicago Press, Chicago

Hart H L A 1983 *Essays in Jurisprudence and Philosophy*. Oxford University Press, Oxford, UK

Mill J S 1969 Utilitarianism. In: Robson J H (ed.) *Collected Works of John Stuart Mill*. University of Toronto Press, Toronto, ON, Vol. 10, pp. 203–59

Okun A M. 1975 *Equality and Efficiency: The Big Tradeoff*. Brookings Institution, Washington, DC

Rawls J 1955 Two concepts of rules. *Philosophical Review* **64**(April): 3–32

Sidgwick H 1907 *The Methods of Ethics*, 7th edn. Macmillan, London

R. Hardin

Utility and Subjective Probability: Contemporary Theories

Contemporary theories of utility and subjective probability are motivated by decision under uncertainty where a person must choose a course of action from a set of possibilities whose outcomes or consequences depend on which state of an uncertain world actually occurs. Utilities assess the person's preferences for outcomes; subjective probabilities measure the person's beliefs about the states' relative likelihoods. A theory of utility and subjective probability makes assumptions about preferences that lead to measures

of utility and subjective probability that combine to evaluate courses of action. The rule of combination is usually a form of mathematical expectation.

1. Decision Under Uncertainty

The standard formulation of decision under uncertainty (Savage 1954) evolved from a theory of expected utility for decision under risk (von Neumann and Morgenstern 1944), where outcome probabilities are assumed to be known, and a theory of subjective probability (de Finetti 1937). This section outlines Savage's formulation. The next provides background on utility theory, with and without explicit mention of probabilities, and the third section discusses quantification of uncertainty by subjective probability. Utility and subjective probability are then combined in Savage's theory of subjective expected utility, followed by alternative theories motivated by perceived shortcomings of his theory.

Savage's formulation envisions an individual faced with choosing one course of action from a set of courses of action. The chosen action will result in exactly one outcome from a set X of outcomes. It is presumed that outcomes are the primary sources of value for the individual.

The chosen action's outcome depends on which state of an uncertain world is the true state, or 'obtains.' The individual knows the set S of states and the outcome that will occur for every action–state pair, but is uncertain about the state that obtains. It is presumed that all relevant uncertainty resides in S and that the chosen action does not influence the true state.

Given S and X, a course of action, or *act* for short, is a map f from S into X. If f is chosen and state s obtains, the outcome is $f(s) \in X$. The set F of acts is a subset of the set X^S of all maps from S into X. Savage's axioms impose conditions on the individual's preference relation $\succsim$ on F, where $f \succsim g$ signifies the act f is preferred or indifferent to act g. Relations of strict preference $\succ$ and indifference $\sim$ are defined from $\succsim$ by $f \succ g$ if $f \succsim g$ and not $(g \succsim f)$, $f \sim g$ if $f \succsim g$ and $g \succsim f$. A constant act assigns the same outcome to every state. When F includes all constant acts, $\succsim$ extends to outcomes by $x \succsim y$ if $f \succsim g$ when $f(s) = x$ and $g(s) = y$ for all $s \in S$. Ordering and dominance axioms illustrate conditions for $\succsim$. The primary ordering condition, transitivity, says that if $f \succsim g$ and $g \succsim h$ then $f \succsim h$. One dominance axiom asserts that if $f(s) \succ g(s)$ for every s, then $f \succ g$.

The family S of subsets $A, B, \ldots$ of S over which subjective probabilities are defined is assumed to be a Boolean algebra, that is,

(a) $S \in S$
(b) $A, B \in S \Rightarrow A \cup B \in S$
(c) $A \in S \Rightarrow S \setminus A \in S$

Members of S are events; event A obtains if some $s \in A$ obtains. When A is an event and $x, y \in X$, let xAy be the act with outcome x if A obtains, and y if $S \setminus A$ obtains. If $x \succ y$ and $xAy \succ xBy$, the individual prefers to take his or her chances on A rather than B for the preferred outcome. According to de Finetti (1937) and Savage (1954), this means that the individual regards A as more probable than B. It is then natural to assume that, for all $x \succ y$ and $x' \succ y'$, $xAy \succ xBy \Rightarrow x'Ay' \succ x'By'$, so that the underlying comparison between events depends only on the direction and not the degree of preference.

2. Utility

Utility theory, broadly construed, concerns quantitative representations of qualitative preference relations. The set on which preferences are defined might be finite or infinite, it may or may not involve probabilities or outcomes with multiple attributes, and so forth. A few basic definitions follow. Let $\succsim$ be a binary preference relation on an arbitrary nonempty set X. Usually it is assumed that $\succsim$ is an ordering relation. The most common assumption is that $\succsim$ is a weak order, which means that it is transitive and complete (for all $x, y \in X$: $x \succsim y$ or $y \succsim x$). When $\succsim$ is a weak order, the indifference relation $\sim$ is an equivalence relation (reflexive: $x \sim x$; symmetric: $x \sim y \Rightarrow y \sim x$; transitive) that partitions X into equivalence classes, with x and y in the same equivalence class if and only if $x \sim y$. A more empirically realistic assumption is that $\succ$ is asymmetric ($x \succ y \Rightarrow$ not $(y \succ x)$) and transitive. This admits nontransitive indifference, as when $x \sim y$, $y \sim z$, and $x \succ z$. Another assumption, which allows preference cycles such as $x \succ y$, $y \succ z$, and $z \succ x$, presumes only that $\succsim$ is complete.

2.1 Ordinal Utility

When X is finite, there is a real-valued function u on X that preserves $\succsim$ in the sense that, for all $x, y \in X$,

$$x \succsim y - u(x) \geqslant u(y) \tag{1}$$

if and only if $\succsim$ is a weak order. In this case, u is an ordinal utility function. An ordinal utility function also exists when $\succsim$ is a weak order and X is countably infinite, but not necessarily when X is uncountable. The lexicographic order $\succ_L$ on $[0, 1] \times [0, 1]$ defined by $(x_1, x_2) \succ_L (y_1, y_2)$ if $x_1 > y_1$ or $(x_1 = y_1, x_2 > y_2)$ illustrates the problem. If u were an order-preserving utility function for $\succ_L$, there would be uncountably many nondegenerate and mutually disjoint real intervals, which is impossible.

An additional condition is needed for Eq. (1) when the set $X/\sim$ of equivalence classes in X is uncountable.

Given a weak order $\succsim$ on X, define $>$ on the equivalence classes $a, b, \ldots$ in $X/\sim$ by $a > b$ if $x \succ y$ whenever $x \in a$ and $y \in b$. A subset T of $X/\sim$ is $>$-order dense in $X/\sim$ if, whenever $a, b \in (X/\sim)\backslash T$ and $a > b$, there is a $t \in T$ such that $a > t > b$.

Theorem 1. *Suppose $\succsim$ is a weak order on X. Then there is an ordinal utility function for $\succsim$ if and only if some countable subset of $X/\sim$ is $>$-order dense in $X/\sim$.*

Proofs are included in Fishburn (1970) and Krantz et al. (1971). Although countable order denseness is not intuitively appealing, other conditions that ensure real-valued utilities are more intuitive. They are referred to as *Archimedean axioms*.

Other utility theorems for partially ordered preferences, comparable preference differences, multiattribute outcomes, and so forth, are included in the preceding references (see also *Measurement, Representational Theory of*; *Measurement Theory: Conjoint*).

2.2 Expected Utility

The preceding subsection applies to preferences in any situation. The present one considers the special case of decision under risk, where preferences apply to decisions with known outcome probabilities. Expected utility (von Neumann and Morgenstern 1944) applies $\succsim$ to a convex set P of probability distributions $p, q, \ldots$ on a set X of outcomes. Convexity means that $\lambda p + (1-\lambda)q \in P$ whenever $p, q \in P$ and $0 \leqslant \lambda \leqslant 1$. With $p(x)$ the probability that p yields outcome x, the probability that $\lambda p + (1-\lambda)q$ yields x is $\lambda p(x) + (1-\lambda)q(x)$.

The axioms of expected utility are: for all $p, q, r \in P$ and all $0 < \lambda < 1$,

A1. $\succsim$ on P is a weak order;

A2. $p \succ q \Rightarrow \lambda p + (1-\lambda)r \succ \lambda q + (1-\lambda)r$; and

A3. $(p \succ q, q \succ r) \Rightarrow \alpha p + (1-\alpha)r \succ q \succ \beta p + (1-\beta)r$ for some $\alpha, \beta \in (0, 1)$.

A2 is an independence axiom that preserves $\succ$ under similar convex combinations. A3 is an Archimedean axiom.

A utility function u on P is linear if, for all $p, q \in P$ and all $0 \leqslant \lambda \leqslant 1$

$$u(\lambda p + (1-\lambda)q) = \lambda u(p) + (1-\lambda)u(q) \tag{2}$$

When u satisfies certain conditions, it is unique up to a positive affine transformation if another function v on P satisfies the same conditions if and only if there are constants $a > 0$ and b such that $v(p) = au(p) + b$ for all $p \in P$.

Theorem 2. *There is a linear, order-preserving real-valued function u on P if and only if* A1, A2, *and* A3 *hold, and such a u is unique up to a positive affine transformation.*

Proofs appear in Fishburn (1970, 1988). When every degenerate probability distribution on X is in P and u on P is extended to X by $u(x) = u(p)$ when $p(x) = 1$, linearity implies $u(p) = \Sigma p(x)\, u(x)$ to give the expected-utility form

$$p \succ q - \sum_X p(x)\, u(x) > \sum_X q(x)\, u(x)$$

for probability distributions with finite supports in X. Its extension to general probability measures is discussed in Fishburn (1970, 1982) along with other topics for expected utility.

One of these (Hausner 1954) generalizes Theorem 2 by dropping the Archimedean axiom A3 and adding the indifference comparison of A2 ($p \sim q \Rightarrow \lambda p + (1-\lambda)r \sim \lambda q + (1-\lambda)r$). These changes lead to a notion of linear utility vectors ordered lexicographically as follows. Let $(T, <_o)$ be a linearly ordered set ($<_o$ is transitive and $t \neq t' \Rightarrow t <_o t'$ or $t' <_o t$) and let K be the set of all real-valued functions on T that are nonzero on at most a well-ordered subset of $(T, <_o)$, with addition and scalar multiplication defined pointwise. Define the lexicographic order $>_L$ on K by $k >_L k'$ if $k \neq k'$ and $k(t) > k'(t)$ for the first t in $(T, <_o)$ at which $k(t) \neq k'(t)$.

Hausner proves that A1, A2, and its $\sim$ companion hold if and only if there is such a $(K, >_L)$ and a map U from P into K such that, for all $p, q \in P$, and all $0 \leqslant \lambda \leqslant 1$, $p \succ q - U(p) >_L U(q)$, $U(\lambda p + (1-\lambda)q) = \lambda U(p) + (1-\lambda)U(q)$. The latter expression is the linearity property for multidimensional utilities.

2.3 Nonlinear Utility

Many studies, including Allais (1953), MacCrimmon and Larsson (1979), and Kahneman and Tversky (1979) demonstrate systematic violations of A2. In one monetary case from Kahneman and Tversky (1979) with $r(\$0) = 1$ and

p: \$3,000 with probability 1;

q: \$4,000 with probability 0.8, nothing otherwise;

$p' = \frac{1}{4}p + \frac{3}{4}r$: \$3,000 with probability 0.25, \$0 otherwise;

$q' = \frac{1}{4}q + \frac{3}{4}r$: \$4,000 with probability 0.20, \$0 otherwise;

a majority of 94 respondents violated A2 with $p \succ q$ and $q' \succ p'$. Consequently, alternatives to expected utility theory that weaken A2 have been proposed.

An early alternative (Allais 1953) assesses u on X by preference–difference comparisons that do not involve probabilities and defines u on P by

$$u(p) = \sum_X p(x)\, u(x) + \theta(p^*) \tag{3}$$

where p^* is the probability distribution induced by p on the differences of outcome utilities from their mean and θ is a real-valued function. A later alternative by

Machina (1982) for the same context returns to the von Neumann–Morgenstern paradigm and assumes in place of A2 that u on P is smooth and linear locally. Machina's generalization has economically interesting implications that generalize those of expected utility.

Kahneman and Tversky (1979) outline a theory, called prospect theory, that modifies expected utility by transforming P's probabilities by an increasing function τ from [0, 1] on to [0, 1]. This responds to the observation that people often overvalue small probabilities and undervalue large probabilities. The number $\tau(\lambda)$ is a subjective equivalent for the objective probability λ, and τ is referred to as a probability weighting function (Wu and Gonzalez 1996, Prelec 1998). Other behavioral findings have been integrated into prospect theory.

Quiggin (1993) also uses an objective-to-subjective probability transformation τ but applies it to cumulative probabilities in a theory of rank-dependent expected utility. Quiggin assumes weak order with $p \succsim q - u(p) \geqslant u(q)$. When the possible outcomes of p are linearly ordered from worst to best as $x_1 \prec x_2 \prec \ldots \prec x_n$

$$u(p) = \sum_{k=1}^{n} u(x_k) \left\{ \tau \left[\sum_{j=1}^{k} p(x_j) \right] - \tau \left[\sum_{j=1}^{k-1} p(x_j) \right] \right\} \quad (4)$$

This reduces to ordinary expected utility when $\tau(\lambda) = \lambda$ for all λ.

Another weak-order theory, weighted utility theory (Chew 1983, Fishburn 1988), retains A1 and A3 and replaces A2 by a convexity axiom and a weak independence axiom: for $p,q,r \in P$ and $0 < \lambda < 1$

convexity: $(p \succ q, p \succsim r) \Rightarrow p \succ \lambda q + (1-\lambda) r$
$(q \succ p, r \succsim p) \Rightarrow \lambda q + (1-\lambda) r \succ p$;

weak independence: $p \sim q \Rightarrow$ for every $0 < \alpha < 1$ there is a $0 < \beta < 1$ such that, for every $r \in P$, $\alpha p + (1-\alpha) r \sim \beta q + (1-\beta) r$

These axioms imply the existence of real-valued linear functions u and w on P with $w \geqslant 0$ such that, for all $p,q \in P$

$$p \succsim q - u(p)\, w(q) \geqslant u(q)\, w(p) \quad (5)$$

When $w > 0$, this gives $p \succsim q - u(p)/w(p) \geqslant u(q)/w(q)$. Fishburn (1988) axiomatizes generalizations of weighted utility, including a theory that accommodates strict preference cycles in P. It uses a skew-symmetric bilinear real-valued function ϕ on $P \times P$ in the representation

$$p \succsim q - \phi(p,q) \geqslant 0 \quad (6)$$

where skew symmetry means that $\phi(p,q) + \phi(q,p) = 0$ and bilinearity means that ϕ is linear separately in each argument. The function ϕ is unique up to multiplication by a positive constant. Fishburn (1988, chap. 6) describes implications of the theory, including the fact that the convex hull of every finite subset of P contains a distribution that has $\succsim$ to all others.

3. Subjective Probability

Let $\succsim$ be a comparative probability relation on a Boolean algebra $\mathscr{S}$ of subsets of a set S of states of the world. A finitely additive probability measure π on $\mathscr{S}$ agrees with $\succsim$ if, for all $A,B \in \mathscr{S}$, $A \succsim B - \pi(A) \geqslant \pi(B)$. The relation $\succsim$ is a qualitative probability if it satisfies the following for all $A,B,C \in \mathscr{S}$:

S1. $\succsim$ on $\mathscr{S}$ is a weak order;
S2. $S \succ \emptyset$, where $\emptyset$ is the empty set;
S3. not $(\emptyset \succ A)$;
S4. $(A \cup B) \cap C = \emptyset \Rightarrow (A \succ B - A \cup C \succ B \cup C)$

These conditions are necessary for agreement.

3.1 Agreeing Measures

De Finetti asked whether every qualitative probability on a finite $\mathscr{S}$ has an agreeing measure. This was resolved negatively by Kraft et al. (1959), who proved that agreement is possible for a qualitative probability if and only if it satisfies an extension of S4. When $\mathscr{S} = 2^S$, the algebra of all subsets of S, their new axiom says that if $m \geqslant 4$, if $A_1, \ldots, A_m, B_1, \ldots, B_m \subseteq S$, and if $|\{j: s \in A_j\}| = |\{j: s \in B_j\}|$ for every s, then it cannot be true that $A_j \succsim B_j$ for all j with $\succ$ for some j. If this fails and an agreeing π is presumed, addition over j and cancellation over S give $0 > 0$.

Savage (1954) considers infinite S and shows that the following Archimedean axiom (S5) obviates the need for the extension of S4.

Theorem 3. *Suppose $\mathscr{S} = 2^S$ and $\succsim$ on $\mathscr{S}$ is a qualitative probability that satisfies the following for all $A,B \in \mathscr{S}$:*

S5. *$A \succ B \Rightarrow$ there is a finite partition $\{C_1, \ldots, C_m\}$ of S such that $A \succ (B \cup C_j)$ for $j = 1, \ldots, m$.*

Then there is a unique probability measure π that agrees with $\succsim$, and for every $A \in \mathscr{S}$ with $\pi(A) > 0$ and every $0 < \lambda < 1$, there is a $B \subset A$ for which $\pi(B) = \lambda \pi(A)$.

Proofs also appear in Fishburn (1970, 1988). Given S1–S4, S5 is equivalent to the assumption that, for every m, S can be partitioned into m equally likely events. Necessary and sufficient conditions for agreement are given in Chateauneuf (1985).

3.2 Nonadditive Subjective Probability

Many other contributions to the theory of subjective probability are discussed in Savage (1954), Fishburn (1988, 1994), and Machina and Schmeidler (1992) (see also *Subjective Probability Judgments*). Of special interest here are approaches in which subjective probabilities are not additive. This was previously considered for decision under risk through the probability weighting function τ. In decision under uncertainty,

failures of additivity were highlighted by examples in Ellsberg (1961) critique of the approach advocated by de Finetti (1937) and Savage (1954).

Consider an urn filled with 90 balls, 30 of which are red (R). The others are black (B) and yellow (Y) in an unknown proportion. One ball is to be drawn at random with a payoff of \$0 or \$2,000 depending on the act chosen and the color of the drawn ball. Four potential acts, shown according to their payoffs for R, B, and Y, respectively, are

f: \$2,000, \$0, \$0
g: \$0, \$2,000, \$0

and

f': \$2,000, \$0, \$2,000
g': \$0, \$2,000, \$2,000

Many people prefer f to g because of R's specificity (30 balls) relative to B's (an unknown number of balls between 0 and 60). By Savage's definition, $f \succ g$ implies $\pi(R) > \pi(B)$. Many people prefer g' to f' for a similar reason. Then $g' \succ f'$ gives $\pi(B \cup Y) > \pi(R \cup Y)$. Under additivity, $\pi(B \cup Y) > \pi(R \cup Y) \Rightarrow \pi(B) > \pi(R)$, which contradicts $\pi(R) > \pi(B)$ via $f \succ g$. The preceding judgments violate S4 when $\succ$ in S4 is interpreted through preference. The theories of subjective expected utility described in the next two sections forbid such violations, but they are allowed by some theories in the final section.

4. Savage's Theory of Subjective Expected Utility

Savage's theory applies $\succsim$ to the act set $F = X^S$ with event set $\mathscr{S} = 2^S$. For any event $A \in S$, write

$f = {}_A x$ if $f(s) = x$ for all $s \in A$,
$f = {}_A g$ if $f(s) = g(s)$ for all $s \in A$, and
$f = xAy$ when $f = {}_A x$ and $f = {}_{A^c} y$,

where $A^c = X \backslash A$. Event A is null if $f \sim g$ whenever $f = {}_{A^c} g$; $\mathscr{N}$ denotes the set of null events. For each $A \in \mathscr{S}$, define a conditional preference relation $\succsim_A$ on F by

$f \succsim {}_A g$ if, for all $f', g' \in F$,

$$(f' = {}_A f, g' = {}_A g, f' = {}_{A^c} g') \Rightarrow f' \succsim g'.$$

This reflects the notion that preference between f and g should depend only on states where their outcomes differ. Similar definitions pertain $\succ_A$ and $\succ_A$.

Savage (1954) uses seven axioms. They apply to all $f, g, f', g' \in F$, all $x, y, x', y' \in X$ and all $A, B \subseteq S$:

P1. $\succsim$ on F is a weak order,
P2. $(f = {}_A f', g = {}_A g', f = {}_{A^c} g, f' = {}_{A^c} g') \Rightarrow (f \succsim g - f' \succsim g')$,
P3. $(A \notin \mathscr{N}, f = {}_A x, g = {}_A y) \Rightarrow (f \succ {}_A g - x \succ y)$,
P4. $(x \succ y, x' \succ y') \Rightarrow (xAy \succ xBy - x'Ay' \succ x'By')$,
P5. $z \succ w$ for some $z, w \in X$,
P6. $f \succ g \Rightarrow$ [given x, there is a finite partition of S such that, for every event E of the partition, $(f' = {}_E x, f' = {}_{E^c} f) \Rightarrow f' \succ g$, and $(g' = {}_E x, g' = {}_{E^c} g) \Rightarrow f \succ g'$],
P7. $(f \succ {}_A g(s)$ for every $s \in A) \Rightarrow f \succsim {}_A g$; $(f(s) \succ {}_A g$ for every $s \in A) \Rightarrow f \succsim {}_A g$

Axioms P2 and P3 are Savage's sure-thing principle. P2 says that preference should not depend on s for which $f(s) = g(s)$, and P3 ties conditional preferences to outcome preferences. P4 is a consistency postulate for subjective probability because $\pi(A) > \pi(B)$ corresponds to $(x \succ y, xAy \succ xBy)$. P6 (cf. S5) is Savage's Archimedean axiom. P7 is a dominance axiom used to extend expected utility for finite-image acts to arbitrary acts.

Theorem 4. *Suppose* P1 *through* P7 *hold for* $\succsim$ *on* F. *Then there is a probability measure* π *on* 2^S *that satisfies the conclusions of Theorem 3 when* $A \succ B$ *corresponds to* $\{xAy \succ xBy$ *for all* x *and* y *with* $x \succ y\}$, *and for which* $\pi(A) = 0 - A \in \mathscr{N}$, *and there is a bounded real-valued function* u *on* X *such that, for all* $f, g \in F$

$$f \succ g - \int_S u(f(s))d\pi(s) > \int_S u(g(s))d\pi(s),$$

with u *unique up to a positive affine transformation.*

A complete proof is included in Fishburn (1970).

Ellsberg's (1961) criticism is directed at P2. Other dissatisfactions include the implication that $\mathscr{S}$ is infinite (Anscombe and Aumann 1963, Pratt et al. 1964) and the assumption that F contains all constant acts (Karni 1985). Related axiomatizations are noted in Fishburn (1982, 1988, 1994).

5. Lottery Theories of Subjective Expected Utility

This section describes two alternatives to Savage's theory for finite S that use extraneous scaling probabilities to provide the scaling power inherent in P6.

5.1 Lotteries and Real-valued Utilities

Let P_X denote the set of all lotteries or finite-support probability distributions on X, and let $\boldsymbol{F} = P_X^S$, the set of lottery acts, so each $\boldsymbol{f} \in \boldsymbol{F}$ assigns a lottery in P_X to every stage. The convex combination $\lambda \boldsymbol{f} + (1-\lambda)\boldsymbol{g}$ is defined by $(\lambda \boldsymbol{f} + (1-\lambda)\boldsymbol{g})(s) = \lambda \boldsymbol{f}(s) + (1-\lambda)\boldsymbol{g}(s)$, so $\boldsymbol{F}$ is convex and A1–A3 of Theorem 2 can be used for $\succsim$ on $\boldsymbol{F}$. Also take $\boldsymbol{f} = {}_A P$ if $\boldsymbol{f}(s) = p$ for all $s \in A$, define $\succsim$ on P_X by $p \succsim q$ if $\boldsymbol{f} \succsim g$ when $\boldsymbol{f} = {}_s p$ and $\boldsymbol{g} = {}_s q$, and define the set $\mathscr{N}$ of null events in 2^S by $A \in \mathscr{N}$ if $\boldsymbol{f} \succ \boldsymbol{g}$ whenever $\boldsymbol{f} = {}_{A^c} \boldsymbol{g}$.

Theorem 5. *With* $S = 2^S$ *and* $F = P_X^S$, *suppose* $\succsim$ *on* $\boldsymbol{F}$ *satisfies* A1–A3 *along with the following for all* $A \subseteq S$, *all* $\boldsymbol{f}, \boldsymbol{g} \in \boldsymbol{F}$, *and all* $p, q \in P_X$

A4. $p' \succ q'$ *for some* $p', q' \in P_X$ *and*
A5. $(A \notin \mathscr{N}, \boldsymbol{f} = {}_A p, \boldsymbol{g} = {}_A q, \boldsymbol{f} = {}_{A^c} g) \Rightarrow (\boldsymbol{f} \succ \boldsymbol{g} - p \succ q)$

Then there is a unique probability measure π on $\mathscr{S}$ with $A \in \mathscr{N} - \pi(A) = 0$, and a real-valued linear function u on P_X, unique up to a positive affine transformation, such that, for all $\boldsymbol{f}, \boldsymbol{g} \in \boldsymbol{F}$

$$f \succsim \boldsymbol{g} - \sum_S \pi(s)\, u(\boldsymbol{f}(s)) \geqslant \sum_S \pi(s)\, u(\boldsymbol{g}(s)).$$

Axioms A4 and A5 correspond respectively to Savage's P5 and his sure-thing principle. Proofs of the theorem appear in Fishburn (1970, 1982) where it is noted that the representation extends to infinite S under addition of a dominance axiom (cf. P7) which says that $(\boldsymbol{f}(s) \succ \boldsymbol{g})$ for all $s) \Rightarrow \boldsymbol{f} \succsim \boldsymbol{g}$ and $(\boldsymbol{f} \succ \boldsymbol{g}(s)$ for all $s) \Rightarrow \boldsymbol{f} \succsim \boldsymbol{g}$.

5.2 *Subjective Expected Lexicographic Utility*

The theory of subjective expected lexicographic utility described in LaValle and Fishburn (1996) applies $\succsim$ to the set P_F of all lotteries on $F = X^S$. Lotteries in P_F are act lotteries. They are denoted as $\boldsymbol{p}, \boldsymbol{q}, \ldots$. With $(\lambda \boldsymbol{p} + (1-\lambda)\boldsymbol{q})(f) = \lambda \boldsymbol{p}(f) + (1-\lambda)\boldsymbol{q}(f)$, P_F is a convex set, so A1–A3 and similar conditions can be used for $\succsim$ on P_F.

The marginal probability distribution induced by $\boldsymbol{p}$ on X for state s is denoted by $\boldsymbol{p}_s : \boldsymbol{p}_s(x) = \Sigma_F\{\boldsymbol{p}(f) : f(s) = x\}$. For $\boldsymbol{p}, \boldsymbol{q} \in P_F$, $p \in P_X$, and $A \in 2^S$

$\boldsymbol{p} =_A p$ means that $\boldsymbol{p}_s = p$ for all $s \in A$ and

$\boldsymbol{p} =_A \boldsymbol{q}$ means that $\boldsymbol{p}_s = \boldsymbol{q}_s$ for all $s \in A$.

When $\boldsymbol{p} =_s \boldsymbol{q}$, $\boldsymbol{p}$ and $\boldsymbol{q}$ can assign different probabilities to acts so long as they have identical marginals for every state. Event A is null if $\boldsymbol{p} \sim \boldsymbol{q}$ whenever $\boldsymbol{p} =_{A^c} \boldsymbol{q}$.

For outcome x, $\bar{x}$ denotes the constant act in P_F with $\bar{x} =_s x$. The relation $\succsim$ on P_F is extended to the set P_X of lotteries on X by the definition

$$p \succsim q \text{ if } \boldsymbol{p} \succsim \boldsymbol{q} \text{ whenever } \boldsymbol{p}(\bar{x}) = p(x) \text{ and } \boldsymbol{q}(\bar{x}) = q(x) \text{ for } \textit{all } x \in X.$$

The ensuing axioms imply that $p \succsim q$ if $\boldsymbol{p} \succsim \boldsymbol{q}$ whenever $\boldsymbol{p} =_S p$ and $\boldsymbol{q} =_S q$. The axioms of the present theory are: for all $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r} \in P_F$, all $p, q \in P_X$, all $A \in 2^S$, and all $0 < \lambda < 1$,

L1. $\succsim$ on P_F is a weak order; $\bar{x} \succ \bar{y}$ for some $x, y \in X$,

L2. $\boldsymbol{p} \succ \boldsymbol{q} \Rightarrow \lambda \boldsymbol{p} + (1-\lambda)\boldsymbol{r} \succ \lambda \boldsymbol{q} + (1-\lambda)\boldsymbol{r}$; $\boldsymbol{p} \sim \boldsymbol{q} \Rightarrow \lambda \boldsymbol{p} + (1-\lambda)\boldsymbol{r} \sim \lambda \boldsymbol{q} + (1-\lambda)\boldsymbol{r}$,

L3. $\boldsymbol{p} =_S \boldsymbol{q} \Rightarrow \boldsymbol{p} \sim \boldsymbol{q}$,

L4. $(\boldsymbol{p} =_A p,\ \boldsymbol{q} =_A q,\ \boldsymbol{p} =_{A^c} \boldsymbol{q},\ \boldsymbol{p} \succ \boldsymbol{q}) \Rightarrow p \succ q$, and

L5. $(P_F, \succsim)$ is finite-dimensional.

L1 combines weak order with nontriviality, L2 corresponds to A2 and its indifference companion, L3 is a state independence condition, and L4 relates $p \succ q$ to $\boldsymbol{p} \succ \boldsymbol{q}$. L5 replaces the Archimedean axiom A3 for P_F by a much weaker condition which says that a multidimensional linear utility function U on P_F of the type described for Hausner's (1954) theory need have only a finite number of dimensions in its lexicographic structure.

Let J denote the minimum lexicographic dimension of $(P_F, \succsim)$ under L5, and let U be a J-dimensional linear lexicographically order-preserving function on P_F. Hausner's theory applies also to $(P_X, \succsim)$ to yield a linear lexicographically order-preserving u on P_X of dimension $K \leqslant J$. Under admissible transformations, U on P_F and u on P_X can be related as

$$U(p) = \sum_S \Pi(s)\, u(\boldsymbol{p}_S) \tag{7}$$

where $\Pi(s)$ is a $J \times K$ real matrix that premultiplies the column vector $u(\boldsymbol{p}_s)$. The $\Pi(s)$ are state matrix probabilities which can be constructed so that $\Pi(s)$ begins with c_s nonzero columns whose leading nonzero entries are positive and in successively lower rows, followed by $K - c_s$ columns of zeros. In addition, state s is null if and only if $\Pi(s)$ is the zero matrix, and $\Sigma\Pi(s)$ is a 0–1 matrix with exactly one 1 in each column. When the 1's of $\Pi(S)$ are in cells $(j_1, 1)$, $(j_2, 2)$, …, and (j_K, K), we have $1 = j_1 < j_2 < \ldots < j_K \leqslant J$.

6. *Other Theories for Decision under Uncertainty*

Excepting Allais (1953), this section discusses theories motivated by Ellsberg (1961) and subsequent behavioral research. Allais's approach uses a preference-free view of subjective probability. He maintains that a person's probability for event A ought to be assessed by having the person identify a chance event A^* of a well-calibrated random device whose occurrence is judged to be equally likely to the occurrence of A. Then $\pi(A)$ equals the objective probability of A^*. An act's utility is then given by Allais's form for the utility of a risky prospect with probabilities given by π.

6.1 Choquet *Expected Utility*

Choquet expected utility offers a generalization of Savage's theory which replaces π by a monotonic set function σ, called a capacity, and computes the utility of an act by Choquet integration rather than ordinary integration. Choquet summation for finite S is similar to the summation of rank-dependent utility (Quiggin 1993) for the monotonic probability weighting function τ.

The initial axiomatization of Choquet expected utility (Schmeidler 1989) uses Theorem 5's lottery approach. Subsequent developments, including modifications of Savage's axioms, are described in Sarin and Wakker (1992). An outline of Schmeidler's theory follows.

A capacity σ on $\mathscr{S}$ is a monotonic $(A \subseteq B \Rightarrow \sigma(A) \leqslant \sigma(B))$ real-valued function with $\sigma(\emptyset)$

$= 0$ and $\sigma(S) = 1$. The Choquet integral $\int_S w(s)d\sigma(s)$ of a real-valued function w on S with respect to the capacity σ is defined by

$$\int_S w(s)\, d\sigma(s) = \int_{t=0}^{\infty} \sigma(\{s: w(s) \geqslant t\})\mathrm{d}t - \int_{t=-\infty}^{0} [1-\sigma(\{s: w(s) \geqslant t\})]\mathrm{d}t \quad (8)$$

If $S = \{s_1, s_2, \ldots, s_n\}$ and $w(s_1) \leqslant w(s_2) \leqslant \ldots \leqslant w(s_n)$, then evaluation of the integrals at the change points gives

$$\begin{aligned}\int_S w(s)\, d\sigma(s) &= w(s_1) + \sum_{i=1}^{n-1} [w(s_{i+1}) - w(s_i)]\sigma(\{s_{i+1}, \ldots, s_n\}) \\ &= \sum_{i=1}^{n} w(s_i)[\sigma(\{s_i, \ldots, s_n\}) - \sigma(\{s_{i+1}, \ldots, s_n\})] \quad (9)\end{aligned}$$

This is analogous to Quiggin's rank-dependent form (Sect. 2.3). Schmeidler (1989) applies $\succsim$ to $\boldsymbol{F}$. Lottery acts $\boldsymbol{f}, \boldsymbol{g} \in \boldsymbol{F}$ are comonotonic if no $s, s' \in S$ have $\boldsymbol{f}(s) \succ \boldsymbol{f}(s')$ and $\boldsymbol{g}(s') \succ g(s)$. His axioms retain A1 and A3, weaken A2 to a comonotonic independence axiom A2*, and assume nontriviality and a dominance condition: for all $\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h} \in \boldsymbol{F}$, and all $0 < \lambda < 1$

A2*. ($\boldsymbol{f}, \boldsymbol{g}$, and $\boldsymbol{h}$ are mutually comonotonic, $\boldsymbol{f} \succ \boldsymbol{g}) \Rightarrow \lambda \boldsymbol{f} + (1-\lambda)\boldsymbol{h} \succ \lambda \boldsymbol{g} + (1-\lambda)\boldsymbol{h}$,

A4*. $\boldsymbol{f}' \succ g'$ for some $\boldsymbol{f}', g' \in \boldsymbol{F}$, and

A5*. ($f(s) \succsim g(s)$ for all $s \in S) \Rightarrow f \succsim \boldsymbol{g}$.

Theorem 6. *Suppose A1, A2*, A3, A4*, and A5* hold on the subset of finite-image lottery acts in* $\boldsymbol{F}$. *Then there is a unique capacity* σ *on* $\mathscr{S}$ *and a real-valued linear function* u *on* P_X, *unique up to a positive affine transformation, such that, for all finite-image* $\boldsymbol{f}$ *and* $\boldsymbol{g}$

$$\boldsymbol{f} \succsim \boldsymbol{g} - \int_S u(\boldsymbol{f}(s))\, d\sigma(s) \geqslant \int_S u(\boldsymbol{g}(s))\, d\sigma(s) \quad (10)$$

A similar result holds for all bounded lottery acts (Schmeidler 1989).

6.2 A Nontransitive Representation

Another generalization of Savage's theory (Fishburn 1988, pp. 208–17) retains his structure and additivity but weakens transitivity and replaces u by a skew-symmetric function ϕ on $X \times X$. The representation is

$$f \succsim g - \int_S \phi(f(s), g(s))d\pi(s) \geqslant 0 \quad (11)$$

The main axiom change weakens P1 to: on $\succsim$ F is complete, and $\succsim$ on $\{f \in F: f(s) \in \{x, y\}$ for all $s \in S\}$ is a weak order for every pair $\{x, y\} \subseteq X$. The representation holds for all finite-image acts with π unique and ϕ unique up to multiplication by a positive constant.

6.3 Rank- and Sign-dependent Utility

Many people have different attitudes toward outcomes perceived as favorable or positive, and those perceived as unfavorable or negative. Utility theories designed to account for this are referred to as sign-dependent theories. Two recent examples are a theory based on a notion of joint receipt (Luce and Fishburn 1991) and a revision of prospect theory called cumulative prospect theory (Wakker and Tversky 1993). Both also have a rank-dependent aspect.

The joint-receipt theory and cumulative prospect theory assume that $\succsim$ is a weak order and that there is a neutral reference outcome x_0 that divides positive from negative outcomes and creates a three-part partition of X with parts

$X^+ = \{x \in X: x \succ x_0\}$
$X^0 = \{x \in X: x \sim x_0\}$
$X^- = \{x \in X: x_0 \succ x\}$

The reference outcome is assigned utility 0, so members of X^+ have positive utilities and members of X^- have negative utilities. In the theories' representations, outcome utilities are unique up to multiplication by a positive constant. The representations treat X^+ and X^- differently and have ways of decomposing an act into positive and negative parts. Each part has its own subjective probability or capacity weighting procedure, and an act's utility is the sum of a positive-part utility and a negative-part utility. In addition, each theory makes assumptions about the relationship between the positive and negative domains for overall scaling consistency.

The cumulative prospect theory axioms in Wakker and Tversky (1993) take $F = X^S$ with S finite and X a connected topological space. For each $f \in F$, $f^+(s) = f(s)$ if $f(s) \succsim x_0$ with $f^+(s) = x_0$ otherwise, and $f^-(s) = f(s)$ if $x_0 \succsim f(s)$ with $f^-(s) = x_0$ otherwise. It is assumed that $\succsim$ on F is a continuous weak order and that some f has both $f^+ \succ x_0$ and $x_0 \succ f^-$. In addition, there is an axiom for gain–loss consistency (if $f^+ \sim g^+, f^- \sim g^-$ and $f \sim x_0$ then $g \sim x_0$) and a more complex condition referred to as sign-comonotonic tradeoff consistency. The representation uses capacities σ^+ and σ^- associated with positive and negative outcomes respectively and takes

$$u(f) = \int_{t \geqslant 0} \sigma^+(\{s: u(f(s)) \geqslant t\})\mathrm{d}t - \int_{t \leqslant 0} \sigma^-(\{s: u(f(s)) \leqslant t\})\mathrm{d}t \quad (12)$$

with u continuous on X, $u(x_0) = 0$, $f \succ g - u(f) \geqslant u(g)$. The capacities are determined uniquely.

The theory in Luce and Fishburn (1991) takes $\mathscr{S}$ as an algebra of events in S and applies $\succsim$ to a set G that

contains g and h with $g \succ x_0 \succ h$. G is endowed with a binary operation $\oplus$ of joint receipt such that $x_0 \oplus x_0 \sim x_0$, $g,h \in G \Rightarrow g \oplus h \in G$ and, for all $g,h,k \in G$, $g \sim h \Rightarrow g \oplus k \sim h \oplus k$ and $k \oplus g \sim k \oplus h$, $g \succ h \Rightarrow g \oplus k \succ h \oplus k$ and $k \oplus g \succ k \oplus h$. The interpretation of $g \oplus h$ are received.

For any $A \in \mathscr{S}$, a map from a finite event-partition of A into X is a first-order gamble, and a map from the same partition into the union of X and the set of first-order gambles that uses at least one of the latter is a second-order gamble. G equals X in union with all first-order gambles and some second-order gambles.

The axiomtization has two phases. The first proposes axioms for $(G, \succsim, \oplus)$ which imply the existence of a real-valued u on G and scaling parameter such that $u(x_0) = 0$ and, for all $g, h \in G$, $g \succsim h - u(g) \geqslant u(h)$ with

$$u(g \oplus h) = \begin{cases} a^+u(g) + b^+u(h) + c^+u(g)u(h) & \text{if } g,h \succsim x_0 \\ a^+u(g) + b^-u(h) & \text{if } g \succsim x_0 \succsim h \\ a^-u(g) + b^+u(h) & \text{if } h \succsim x_0 \succsim g \\ a^-u(g) + b^-u(h) + c^-u(g)u(h) & \text{if } x_0 \succsim g,h \end{cases}$$

where a^+, b^+, a^- and b^- are positive, $c^+ \geqslant 0$ and $c^- \leqslant 0$.

The second phase considers gambles' internal structures in more detail. For any event-partition $\{A_1, \ldots, A_m\}$ of $A \in \mathscr{S}$ and first-order gamble g: $\{A_1, \ldots, A_m\} \to X$, let A^+ be the union of the A_j for which $g(A_j) \succ x_0$; define A^0 and A^- similarly. Also define first-order gambles g^+ on A^+, g^0 on A^0, and g^- on A^- by $g^i(A_j) = g(A_j)$ if $A_j \subseteq A^i$, for $i \in \{+,0,-\}$. These define the sign-partitioned second-order gamble g_2 on partition $\{A^+, A^0, A^-\}$ of A by $g_2(A^i) = g^i$ for $i \in \{+,0,-\}$. The g_2 thus defined are the second-order gambles in G.

A key second-phase axiom says that, under independent realizations for A^+ and A^-, $g_2 \sim [(g^+ \text{ if } A^+, x_0 \text{ if } A \setminus A^+) \oplus (g^- \text{ if } A^-, x_0 \text{ if } A \setminus A^-)]$. Given u from phase one, the axioms imply that for every nonnull $A \in \mathscr{S}$ there are unique real-valued functions $\sigma^+(\cdot|A)$ and $\sigma^-(\cdot|A)$ from $\mathscr{S}$'s subalgebra on A into [0, 1] such that, for g from the partition $\{A_1, \ldots, A_m\}$of A into X

$$u(g) = a^+u(g^+)\sigma^+(A^+|A) + b^-u(g^-)\sigma^-(A^-|A) \quad (13)$$

A second formula for constant-sign gambles allows $u(g^+)$ and $u(g^-)$ to be reduced to rank-dependent expressions in outcome utilities. In each iteration of the reduction procedure, the remaining outcome nearest to x_0 is separated out and 'subtracted' from other outcomes. The weights that multiply the outcome utilities in the rank-dependent expression depend on the order of reduction in a manner related to Choquet summations for σ^+ and σ^- in cumulative prospect theory. Further contributions along the lines of the preceding developments as well as other important areas of utility theory are presented in Luce (1999).

See also: Utility and Subjective Probability: Empirical Studies

Bibliography

Allais M 1953 Fondements d'une théorie positive des choix comportant un risque et critique des postulats et axiomes de l'école américaine. Colloques Internationaux du CNRS: XL. *Econométrie* 257–332. Translated and augmented 1979 The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American school. In: Allais M, Hagen O (eds.) *Expected Utility Hypotheses and the Allais Paradox*. Reidel, Dordrecht, Holland

Anscombe F J, Aumann R J 1963 A definition of subjective probability. *Annals of Mathematics and Statistics* **34**: 199–205

Chateauneuf A 1985 On the existence of a probability measure compatible with a total preorder on a Boolean algebra. *Journal of Mathematical Economics* **14**: 43–52

Chew S H 1983 A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica* **51**: 1065–92

de Finetti 1937 La prévision: ses lois logiques, ses sources subjectives. Annals de l'Institute Henri Poincaré **7**: 1–68 Trans. Kyburg H E 1964 Foresight: its logical laws, its subjective sources. In: Kyburg H E, Smokler H E (eds.) *Studies in Subjective Probability*. Wiley, New York

Ellsberg D 1961 Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* **75**: 643–69

Fishburn P C 1970 *Utility Theory for Decision Making*. Wiley, New York

Fishburn P C 1982 *The Foundations of Expected Utility*. Reidel, Dordrecht, Holland

Fishburn P C 1988 *Nonlinear Preference and Utility Theory*. Johns Hopkins University Press, Baltimore, MD

Fishburn P C 1994 Utility and subjective probability. In: Aumann R J, Hart S (eds.) *Handbook of Game Theory*. Elsevier, Amsterdam, Vol. 2

Hausner M 1954 Multidimensional utilities. In: Thrall R M, Coombs C H, Davis R L (eds.) *Decision Processes*. Wiley, New York

Kahneman D, Tversky A 1979 Prospect theory: an analysis of decision under risk. *Econometrica* **47**: 263–91

Karni E 1985 *Decision Making Under Uncertainty: The Case of State-dependent Preferences*. Harvard University Press, Cambridge, MA

Kraft C H, Pratt J W, Seidenberg A 1959 Intuitive probability on finite sets. *Annals of Mathematics and Statistics* **30**: 408–19

Krantz D H, Luce R D, Suppes P, Tversky A 1971 *Foundations of Measurement*. Academic Press, New York, Vol. 1

LaValle I H, Fishburn P C 1996 On the varieties of matrix probabilities in nonarchimedean decision theory. *Journal of Mathematical Economics* **25**: 33–54

Luce R D 1999 *Utility of Certain and Uncertain Alternatives: Measurement, Theoretic and Experimental Approaches*. Erlbaum, Mahwah, NJ

Luce R D, Fishburn P C 1991 Rank- and sign-dependent linear utility models for finite first-order gambles. *Journal of Risk and Uncertainty* **4**: 29–59

MacCrimmon K R, Larsson S 1979 Utility theory: axioms versus 'paradoxes.' In: Allais M, Hagen O (eds.) *Expected Utility Hypotheses and the Allais Paradox*. Reidel, Dordrecht, Holland
Machina M J 1982 'Expected utility' analysis without the independence axiom. *Econometrica* **50**: 277–323
Machina M J, Schmeidler D 1992 A more robust definition of subject probability. *Econometrica* **60**: 745–80
Pratt J W, Raiffa H, Schlaifer R 1964 The foundations of decision under uncertainty: an elementary exposition. *Journal of the American Statistical Association* **59**: 353–75
Prelec D 1998 The probability weighting function. *Econometrica* **66**: 497–527
Quiggin J 1993 *Generalized Expected Utility Theory: The Rank-dependent Model*. Kluwer, Boston
Sarin R, Wakker P 1992 A simple axiomatization of nonadditive expected utility. *Econometrica* **60**: 1255–72
Savage L J 1954 *The Foundations of Statistics*. Wiley, New York
Schmeidler D 1989 Subjective probability and expected utility without additivity. *Econometrica* **57**: 571–87
von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ
Wakker P, Tversky A 1993 An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty* **7**: 147–75
Wu G, Gonzalez R 1996 Curvature of the probability weighting function. *Management Science* **42**: 1676–90

P. Fishburn

Utility and Subjective Probability: Empirical Studies

In the last few decades, empirical studies of choice, probability assessment, and certainty equivalents have demonstrated that utilities are *not* independent of subjective probabilities. People are often pessimistic about the odds of a possible outcome if the outcome is bad. Rank-dependent theories and configural-weight theories can describe these interactions by allowing decision weights to vary with the rank orders of outcomes. Holding objective probability constant, lower-ranked outcomes typically receive greater weight than higher-ranked outcomes. (See also *Utility and Subjective Probability: Contemporary Theories*.)

Despite the enthusiasm with rank-dependent theories, recent studies have shown that these theories have difficulties describing choices when options have more than two outcomes. Choices are consistent with configural-weight theories. However, configural-weight theories can not account for interactions across options. Regret theories describe these interactions in emotional terms, but there are probably many reasons for the interactions. An important challenge for future theorists is to describe both types of interactions within a single theoretical framework. For further reading on utility theories and empirical tests, see Edwards (1992) and Luce (2000).

1. Prospect Theory

In a now classic paper on risky choice, Kahneman and Tversky (1979) proposed an alternative to expected utility theory called prospect theory. Rational rules were replaced with psychological principles. The value function in prospect theory applied to changes in wealth, not total wealth. Furthermore, changes had diminishing marginal value, such that a change from \$0 to \$10 had greater impact than an identical change from \$1000 to \$1010. Finally, negative changes had greater impact than positive changes of equal magnitude, an asymmetry known as loss aversion.

Prospect theory also made psychological assumptions about decision weights. Decision weights were nonlinearly related to objective probabilities; weights for small probabilities were larger than objective probabilities, and weights for large probabilities were smaller. In addition, weights at the endpoints were discontinuous. An event with *no* chance of occurring was psychologically different from an event with a one percent chance of occurrence (e.g., contracting a horrible disease), and a sure thing was psychologically different from an event with a 99 percent chance of occurring (e.g., winning a million dollars in the lottery).

Prospect theory accounted for a variety of empirical phenomena. One example was the reflection effect, which suggested that risk attitudes varied around the status quo. Consider a choice between a gamble with an 80 percent chance to win \$4,000 and a sure win of \$3,000. Most people prefer the sure thing. Now consider a choice between a gamble with an 80 percent chance to lose \$4,000 and a sure loss of \$3,000. Most people prefer the gamble. Although expected utility theory assumed that risk attitudes were constant across all levels of wealth, prospect theory asserted that the shape of the value function differed in the gain and loss domains. Preferences were risk averse in the gain domain and risk seeking in the loss domain. Despite its success, prospect theory had a major drawback. It predicted that people would violate stochastic dominance. Stochastic dominance implies that a decision maker will select the alternative that is as good or better in all respects than other alternatives. Prospect theory predicted that, in some cases, people would choose the inferior option. Despite evidence that people violate stochastic dominance, there was a strong desire to find a general representation that served both normative and descriptive needs. Interest turned to cumulative rank-dependent theories.

2. Cumulative Rank-dependent Theories

In cumulative rank-dependent theories, the decision weights associated with outcomes can vary with the rank orders of outcomes. Furthermore, weights are cumulative or decumulative probabilities, an assump-

tion that makes violations of stochastic dominance impossible. Several versions of cumulative rank-dependent theories were proposed, three of which are presented in more detail.

2.1 Rank- and Sign-dependent Utility Theory

Luce (1991) and Luce and Fishburn (1995) proposed an account called rank- and sign-dependent utility theory. This version was built on the assumption of joint receipt (the simultaneous receipt of two or more objects, such as gambles). Luce and his colleagues tested several of the axioms, including segregation, duplex decomposition, and the additivity of certainty equivalents over joint receipt. Segregation implied that the value of a gamble with positive outcomes was equal to the value of the joint receipt of the gamble with the lowest outcome subtracted from all outcomes and the value of the lowest outcome. Duplex decomposition asserted that the value of a gamble with mixed outcomes was equivalent to the joint receipt of a subset of the gamble with all positive outcomes and a subset of the gamble with all negative outcomes. Finally, additivity of certainty equivalents over joint receipt implied that the certainty equivalent for positive outcomes was combined additively with that of negative outcomes.

Tests were conducted with judged certainty equivalents and choice certainty equivalents. Judged certainty equivalents are indifference prices between a sure thing and a gamble, and choice certainty equivalents are indifference points that are inferred from choices between sure things and gambles. Segregation was supported by one measure, duplex decomposition was supported by both measures, and additivity of certainty equivalents was supported by neither measure. In sum, the tests yielded mixed results.

But not all of the evidence was mixed. Tests of an axiom called event commutativity were supportive. Event commutativity implies that the order of events should not matter as long as the outcomes arise under the same conditions. Results were consistent with rank- and sign-dependent utility theory. Camerer and Ho (1994) found more supporting evidence with an axiom called betweeness. Betweeness is a weakened form of the independence axiom and states that a preference for a probability mixture of two gambles should fall between the preferences for the two separate gambles. Violations of the axiom were robust and consistent with cumulative rank-dependent theories.

2.2 Cumulative Prospect Theory

Tversky and Kahneman (1992) developed a theory called cumulative prospect theory. It was axiomatized differently from rank- and sign-dependent utility theory, and it also made different assumptions about values and decision weights. Decision weights had an inverse S-shape form. The value function was a concave power function for gains and a convex power function for losses with a steeper slope for losses than gains.

Empirical tests of cumulative prospect theory were done with global fits to data, and results were generally supportive. One-parameter versions of the S-shaped weighting function provided good fits to median data (Camerer and Ho 1994, Wu and Gonzalez 1998, Prelec 1998). More recently, two-parameter weighting functions that describe individual differences in discriminability and gambling preference were proposed.

2.3 Security Potential/Aspiration Theory

Lopes (1990) developed a cumulative rank-dependent theory called security potential/aspiration theory. She argued that participants making risky choices were both security minded (i.e., they wanted to avoid the worst outcomes) and potential minded (i.e., they wanted to achieve the best outcomes). An S-shaped weighting function reflected the 'optimism' that people feel with smaller probability outcomes and the 'pessimism' they feel with larger probability outcomes. Security potential/aspiration theory described choice as a combination of rank-dependent utilities and consideration of an aspiration level. Results were generally supportive.

3. Configural-weight Theories

Birnbaum and his colleagues turned to another class of theories called configural-weight theories. In these accounts, decision weights were noncumulative and permitted violations of stochastic dominance. Birnbaum and McIntosh (1996) and Birnbaum and Navarrete (1998) provided additional evidence of such violations by showing that people select the dominated option when making choices with certain three-outcome gambles. Configural-weight theories predicted the results.

Birnbaum (1999) also demonstrated that choice behavior violated a property called coalescing. Coalescing, also called event splitting, applies to probability-outcome combinations with identical outcomes. Cumulative rank-dependent theories predict that probability-outcome combinations with identical outcomes should be equivalent regardless of whether they are presented separately or as a sum. For example, a 20 percent chance of \$10 and a 15 percent chance of \$10 should be equivalent to a 35 percent chance of \$10. Birnbaum showed that choices differed with the two modes of presentation. Identical outcomes presented separately had greater impact than identical outcomes presented together. Again, configural-weight theories predicted the results.

Cumulative rank-dependent theories and configural-weight theories differ in several respects. Configural-weight theories allow decision weights to vary

with additional factors, including the value of an outcome, the number of outcomes, the spacing of outcomes, and the decision maker's point of view. The flexibility of the decision weights implies that these theories can describe violations of independence, including branch independence, cumulative independence, and interval independence. However, this flexibility has a downside; it is not easy to reject configural-weight theories. The key assumption of these theories is that utilities are invariant across all factors that influence weights. Birnbaum and his colleagues argue that tests of this assumption have generally supported the theories.

4. *Interactions across Options*

Not only do utilities and subjective probabilities interact within an option, but several studies show that utilities and subjective probabilities interact across options. Chechile and Cooke (1997) and Chechile and Luce (1999) presented participants with a reference gamble and two outcomes of a comparison gamble. Participants adjusted the probability of the better outcome until the comparison gamble and the reference gamble seemed equivalent. Chechile and his colleagues showed that rank-dependent theories could be represented as linear functions with invariant slopes across reference gambles. Results showed that the slopes systematically varied; the utilities of reference gambles depended on the comparison gambles with which they were paired. Mellers and Biagini (1994) also demonstrated interactions across options in the form of strong stochastic transitivity violations and similarity effects. They represented the interactions in terms of attribute weights and argued that people shifted their attention from attributes with similar levels to those different levels to best discriminate among options.

Regret theories also describe interactions across options. People often anticipate the regret they will feel if their outcome turned out to be worse than the outcome of another choice. Theories of regret assume that people avoid such feelings when making choices (Loomes and Sugden 1982, Bell 1982). Numerous developments have occurred along these lines (Mellers 2000), and results are promising. Although no single unifying theory has been suggested yet, future theorists will undoubtedly provide a general representation of the interactions both within and across options.

See also: Characterization Theorems in Random Utility Theory; Decision and Choice: Paradoxes of Choice; Decision and Choice: Random Utility Models of Choice and Response Time; Decision Research: Behavioral; Heuristics for Decision and Choice; Utility and Subjective Probability: Contemporary Theories

Bibliography

Bell D 1982 Regret in decision making under uncertainty. *Operations Research* **20**: 961–81

Birnbaum M H 1999 Testing critical properties of decision making on the Internet. *Psychological Science* **10**: 399–407

Birnbaum M H, McIntosh W R 1996 Violations of branch independence in choices between gambles. *Organizational Behavior and Human Decision* **67**: 91–110

Birnbaum M H, Navarrete J 1998 Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty* **17**: 49–78

Camerer C, Ho R-H 1994 An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty* **8**: 167–96

Chechile R, Cooke A C 1997 An experimental test of a general class of utility models: Evidence for context dependency. *Journal of Risk and Uncertainty* **14**: 75–93

Chechile R, Luce R D 1999 Reanalysis of the Chechile–Cooke experiment: Correcting for mismatched gambles. *Journal of Risk and Uncertainty* **18**: 321–5

Edwards W 1992 (ed.) *Utility Theories: Measurements and Applications*. Kluwer, Boston, MA

Kahneman D, Tversky A 1979 Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263–91

Loomes G, Sugden R 1982 Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal* **92**: 805–24

Lopes L L 1990 Re-modeling risk aversion. In: Von Furstenberg G M (ed.) *Acting Under Uncertainty: Multidisciplinary Conceptions*. Kluwer, Boston, MA

Luce R D 1991 Rank- and sign-dependent linear utility models for binary gambles. *Journal of Economic Theory* **53**: 75–100

Luce R D 2000 *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*. Erlbaum, Mahwah, NJ

Luce R D, Fishburn P C 1995 A note on deriving rank-dependent utility using additive joint receipts. *Journal of Risk and Uncertainty* **11**: 5–16

Mellers B A 2000 Choice and the relative pleasure of consequences. *Psychological Bulletin* **126**: 910–24

Mellers B A, Biagini K 1994 Similarity and choice. *Psychological Review* **101**: 505–18

Prelec D 1998 The probability weighting function. *Econometrica* **66**: 497–527

Thurstone L L 1927 A law of comparative judgment. *Psychological Review* **34**: 273–86

Tversky A, Kahneman D 1992 Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* **5**: 297–323

Wu G, Gonzalez R 1998 Common consequence conditions in decision making under risk. *Journal of Risk and Uncertainty* **16**: 115–39

B. A. Mellers

Utopias: Social

'Utopia' refers to fictions or essays which purport to describe an ideal and feasible community at some, generally undefined, future date. 'Utopia' is a doubly

ambiguous word coined by Sir (and Saint) Thomas More as the title of a book he published in 1516. First of all, the neologism itself has a Greek etymology which can be read as either 'no place' (*U-topia*) or 'good place' (*Eu-topia*). Second, the label 'utopian' can be restricted to utopias as pictures, namely the detailed narrative portrayals of ideal communities, or it can be stretched to include also blueprints which reflect utopian tendencies, namely any systematic exposition of the underlying principles of such communities.

The position taken hereafter will be both comprehensive and restrictive. Comprehensive, as it includes both interpretations under the general heading of utopian impulse—the conviction that 'a definitive and insurpassable condition is attainable' (Kolakowski 1990, p. 132). From Plato on, the attainment of the perfect society has been a constant preoccupation in the West, and it would be unduly restrictive to rule out nonfictional utopian writings. However, some limitation seems suitable in order to retain only those visions of the ideal society which are meant to be implemented by human agency. In other words, the necessary and sufficient condition to be let into the Utopian club is to have been dreaming or speculating about a more perfect social order (or, more often than not, doing a bit of both), starting from the presupposition that human life takes place within a societal framework which is plagued by countless evils, but the cure, a radical one, is at hand, to be followed by indefinite perfection on this earth.

1. *Visiting Utopias*

The utopian quest for a definitive answer to the hypothesized human predicament looks very much like an extended multiple-choice questionnaire with items such as where is this place of bliss located, who are its residents, and what rules do they follow?

1.1 *From No-place to the Earth and Beyond*

The Platonic tradition which conceived good societies as necessarily small has been resumed by Thomas More and countless other writers who placed their utopia on imaginary islands or some similar remote sites, waiting to be visited by admiring travelers from the outside world who have been shipwrecked on their shores. With Mercier's *L'An 2440* (1971) the discovery is made through a lengthy process of sleep and reawakening, Rip van Winkle-style, in the utopian future of a whole nation-state, namely France's 700 years ahead, a narrative convention which has been used in countless utopias. With the passage of time the scope of utopias has become more embracing, to include developed post-capitalist nations in Marxist utopias, the whole Earth complete with a single world government and English as universal language in some H. G. Wells writings, and eventually the solar system through the space colonies as has been anticipated by Freeman Dyson (see Carey 1999, p. 463).

1.2 *Not Everybody (nor Everything) is Allowed into Utopia*

Radical utopian change is primarily aimed at faulty social institutions but as books are the repository of wrong beliefs and/or immoral fictions, they also can be fated to destruction as is the case for poetry in Plato's *Republic* (1941, Book X), in Mercier's twenty fifth century France or in Cabet's communist *Icaria*. Some utopias even provide for the destruction and renewal of much of the urban structures of the past, as in the Boston of Bellamy's *Looking Backward* (1951, chap. 5) or the big murky places which were, as we know, the centres of manufacture in England in *News from Nowhere* by William Morris (1970, chap. 10).

The admittedly extreme case of Wells' *In the Days of the Comet*, where several of the world's big cities are obliterated by a comet whose green vapors simultaneously induce a hypnotic slumber which alters human nature beyond recognition, would suggest that the actual aim of these beneficial disasters is less the physical structures than the human beings, for the latter have to change too in order to deserve the benefits of life in utopia. These changes concern either the existing population or its newborn component.

A first requirement, it seems, is that the raw human material has to be cleansed, either biologically or socially. More's Utopians secretly hope that all the Swiss they use as mercenaries will eventually be 'wiped off the face of the earth completely' as they are unfit to live. A much more significant case is provided by Wells, who considered that in the world state he anticipated, 'people who cannot live happily and freely in the world without spoiling the lives of others are better out of it,' a prescription made in 1901 which, most unfortunately, was taken deadly seriously some 40 years later. An as yet uncommon variant concerns gender cleansing, which aims at a deliberate change in the sex ratio. According to Wells, genetic engineering will allow 'if woman is too much for us, [to] reduce her to a minority' (Carey 1999, p. 371). More radically, but with here again some pretence of scientific veneer, feminist utopianism deals with this value-laden issue by, unsurprisingly, anticipating an all-woman community based on parthenogenesis (Gilman, in Carey 1999, p. 382), or a flower children paradise where some technology has been used to allow men (if that is the right word) to grow breasts and to suckle babies (Piercy, in Carey 1999, p. 475).

However, social cleansing of class opponents, while massively performed during and after the 1920s, is a most rare occurrence in the utopian literature except as a thought experiment performed by the French

utopian Saint-Simon (what if the leading nobles and officials of the bureaucracy died?, the answer being: no impact in social welfare terms), or as an aside in the Marxian literature dealing with the 'dictature of the proletariat.'

Beyond taking care of the initial stock, selective breeding is of great concern in some utopias, with a variety of patterns such as state control and astrological supervision of matings in Campanella's *The City of the Sun* (Carey 1999, p. 61) on the one hand, and reliance on unhindered sexual free market as a means toward race purification in *Looking Backward*, on the other. Once again H. G. Wells is quite clear that in his world utopia nonprocreational sex is all right, but parentage shall be subject to strictly enforced age, health, education, and income standards.

1.3 The Rules of the Utopian Game: A Difficult Tradeoff Between Equality, Compulsion, and Welfare

One crucial commonality underlying all utopian designs, fictional or otherwise, is that the source of what goes wrong in human affairs lies not in the original sin, which nobody can do about, but to faulty social institutions which can be dealt with by human agency. The 'fault' in social institutions lies in their giving free rein to the worst instincts of humankind, such as greed, aggressiveness, and jealousy and/or in stifling its best inborn instincts. Then, the right social arrangements are those that promote maximum equality among members of the relevant community. In the words of one typical exponent, the French revolutionary Babeuf (Carey 1999, p. 177), 'society must be made to operate in such a way that it eradicates once and for all the desire of a man to become richer, or wiser, or more powerful than others.' The positive fallout to be expected from this equality maximizing purpose are the usual ones of plummeting crime rates, eradication of (relative) poverty, and a stronger social fabric. As for the negative aspects, such as drabness and enregimentation in daily life, they figure quite explicitly in some works including *Utopia* ('when you have seen one town, you have seen them all') or Cabet's *Icaria* ('everything concerning food has been regulated by the law [which] accepts or rejects any type of nourishment').

Interestingly enough, perfectly consistent egalitarian utopias are in a minority, as Kolakowski (1990) points out. The majority is made of inconsistent schemes which allow having one's utopian cake and eating it by mixing egalitarianism with other, qualitatively different elements. Saint-Simon and Marx (in his infrequent utopian moments), Bellamy and, incredibly enough, even Cabet, are all evidence of such a compromise, with its unfailing reference to the utmost development of individual potentialities, to the concomitant productive efficiency of the economic infrastructure which allows egalitarianism to coexist with an unlikely upper-middle class affluence, but very little mention of what happens to delinquent or work-shy members of the community.

On the whole, there are few utopias which succeed in convincingly combining (a) a fair amount of equality, (b) minimal state coercion, and (c) nonascetic lifestyles. *News from Nowhere* (Morris 1970) is strong on the first two counts but unconvincing on the third, as all pastoral fairyland are. Huxley's *Island* (1962) is another case of equality *cum* small government, but a very short-lived one, whose indisputable ascetism is moreover counterpoised by a tropical paradise-type environment and recurrent use of mind-changing pills. The aforesaid feminist utopias are too clearly conceived for a restricted female readership to deserve discussion except as one curious brand of science fiction among others. However, Fourier deserves a special mention for the quite imaginative trade-off he devised between the three poles of the (a), (b), (c) triad for he is, as he rightfully argued, 'the only reformer who has rallied round human nature by accepting it as it is and devising the means of utilizing it with all the defects which are inseparable from man' (Manuel 1965, p. 207). Unfortunately, in spite of his efforts, neither the Russian Czar nor the Rothschilds accepted to invest some seed-money to check whether Fourier was right in claiming that all the passions created by God could work in harmony when afforded maximum expression in Fourierist phalansters. As for those experimental utopian communities created in North America by some followers of Fourier, Cabet, or Owen, they were at best honorable failures.

2. Utopias Appraised

2.1 The Birth of Utopias: A Well-kept Secret

Each utopia is in itself a permanent paradox: how an ideal society can be established by a gathering of hopelessly imperfect individuals? As has been shown above, makers of utopias are generally quite willing to let us know how their pet ideal society was discovered; they are also tireless in the description of their commendable arrangements. However, how a given utopia came into being looks like a well-kept secret (Ferns 1999). In most cases the problem is just swept under the carpet. Utopias, as Chesterton (1905) sarcastically pointed out, 'first assume that no man will want more than his share, and then they are very ingenious in explaining whether this share will be delivered by motor car or balloon.' The implicit reasoning seems to be that the merits of the 'nowhere' are vastly more important than the account of how it came about. There are also some cases, especially in the nineteenth century, where utopian changes were vindicated as the obvious response to the necessary laws of historical development which move the world, but except some evasive references from the Marxist

side to the inevitable transition through a dictatorship of the proletariat, William Morris was the only anticapitalist utopian to acknowledge that the new imagined society did not come about peacefully: 'it was war from beginning to end, bitter war.' As for Wells he found a third way of explaining away the emergence of a world utopia by ascribing it to the destabilizing impact of a disaster, either natural (the above-mentioned comet) or human-made (a nuclear war in *The World Set Free*, Wells, 1914).

2.2 Are Utopias Such a Good Place?

While many people have adhered to, and would still subscribe to Oscar Wilde's dictum that 'progress is the realization of Utopias,' reaction against utopia's blandishments is visible from the very beginning, as evinced by Aristophanes (1924), whose *The Parliament of Women*, with its satire of 'utopian' sexual relations, was contemporaneous with Plato's *Republic*. Very often the word 'dystopia' is used to convey the idea of a utopia turned sour. Unfortunately, it is too often applied to any kind of work, fictional or otherwise, which anticipates a dismal future for some existing society or for the world at large. If one agrees with the statement that 'to count as a utopia, an imaginary place must be an expression of desire. To count as a dystopia, it must be an expression of fear' (Carey 1999, p. xi), a case can be made for a distinction between two kinds of fear.

The first one, which is as old as the prophets of doom (with Malthus as its first modern voice) concerns the downfall which humankind might be heading for given its underlying, presumably lethal logic, a message for which the phrase 'disaster scenario' suggested by Carey seems quite appropriate. The second fear has been voiced by the Russian philosopher Berdiaev as quoted in the epigraph to Huxley's *Brave New World* (1963): 'Utopias seem much more realizable than once was thought. We find ourselves faced today with a problem distressing in a very different way: how can we prevent their realization? Utopias are realizable, and towards utopias we are moving. But it is possible that a new age is already beginning, in which cultured and intelligent people will dream of ways to avoid ideal states and to get back to a non utopian society, less 'perfect' and more free.' Dystopia, etymologically 'a bad place,' which makes it a kind of mirror image of utopia, can thereby serve to characterize those anticipations which seek to warn their readers that the world is evolving dangerously closer to utopia, the latter being then no more a solution but the problem.

In some of the dystopias, the target is clearly identified: socialist egalitarian utopias Bellamy-style in the first, and quite funny, example known (Jerome 1891, see Beauchamp 1983); the revulsion against Bellamy and his 'cockney paradise' was a source of inspiration for *News from Nowhere* (Morris 1970), *The Machine Stops* (Forster 1965) was written as a 'counterblast' against *A Modern Utopia* (Wells 1987), while *Men Like Gods*, a utopian novel by the same Wells provoked Huxley into writing a parody famously known as *Brave New World* (1963).

Unsurprisingly, one finds in the 'patented' dystopias from Jerome's prototype on, all the major features that are typical of mainstream utopias, but with reversed connotations: what was described in the typical utopian package as benevolent government care, strong social ties, true equality, and full development of individual potentialities is now disclosed as just the opposite, namely constant public scrutiny, compulsory uniformity, and lack of privacy; in other words the embodiment of the rule that 'everything is allowed for the sake of the society,' an 'unholy maxim' designed, Tocqueville said, 'to legitimize all the oppressions to come.'

If the typical dystopia is utopia taken upside down (so much so that it is sometimes unfairly interpreted as just a pro-Establishment reassertion of reactionary values), *Brave New World* (Huxley 1963) is on the other hand remarkably difficult to classify. As Carey put it, 'medical science has eliminated virtually all diseases, together with the debilitating effects of old age. There is no pain, no hunger, no want. [*Brave New World*] can be read as an ambivalent text, dystopian from one angle, utopian from another.' Hence the society described therein holds more attraction for readers today (particularly young ones) than Huxley expected (Jennings 1996); small wonder that the dystopian human hatchery in *Brave New World* (Huxley 1963) appears as a major positive highlight in Marge Piercy's feminist community (Carey 1999, p. 476). Oddly enough, the same ambivalence has been detected in *Utopia* (More 1965) which can be read, its modern, progressive features notwithstanding, as an ultimately dystopian admonition against relying on economic and social arrangements that are functionally efficient but at odds with a Christian and humanist set of values (Abrash 1977).

3. Some Concluding Remarks

Is there a future for utopias? If the latter are meant as comprehensive and hopefully feasible institutional schemes where happiness prevails once for all through consensual equality, the answer is probably 'none.' However, one can bet on the resilience of a special brand of utopian mentality, which did not leave any book worth remembering and whose followers literally believe in the imminent coming of paradise on earth after a cataclysmic confrontation between the true believers, generally under the guidance of a Messiah, and the forces of evil (Campion 1994). In addition, even if large-scale utopias seem to be a thing of the past, one should not forget the persistence of partial

utopias, namely the widespread conviction that in some specific areas of human endeavor, some ultimate state of (bounded) perfection is attainable, health being the most obvious candidate.

As far as dystopias are concerned, they may be threatened with obsolescence by the vanishing of non-millenarian utopian constructs but, after all, *Nineteen Eighty-Four*, Orwell's dystopian masterwork (1983), was not aimed at a specific utopian fiction but at a too real target. Above all, one must keep in mind that the key *raison d'être* of dystopias is to be a reminder of the ubiquitousness of unintended (dismal) effects of presumably well-intended social action (Boudon 1982). Hence, there is some reason to expect that the major issues on the dystopian agenda might include the dual threats of (a) the subversion from weapon-rich, single-minded millenarian groups; and (b) the disruption of the social fabric through human genetic re-engineering brought on, not by some benevolent Big Brother as in conventional utopias (and dystopias), but through decentralized parental decision making (Silver 1998, in Carey 1999, pp. 514–17).

See also: Action, Theories of Social; Collective Beliefs: Sociological Explanation; Ideology, Sociology of; Knowledge, Sociology of; Marx, Karl (1818–89); Marxist Social Thought, History of

Bibliography

Abrash M 1977 Missing the point in More's *Utopia*. *Extrapolation* **1**: 27–38

Aristophanes 1924 *The Ecclesiazusae*. Heinemann, London and Putnam, New York

Beauchamp G 1983 The proto-dystopia of Jerome K. Jerome. *Extrapolation* **2**: 170–81

Bellamy E 1951 *Looking Backward*. Modern Library, New York

Boudon R 1982 *The Unintended Consequences of Social Action*. Macmillan, London

Cabet E 1970 *Voyage en Icarie*. Anthropos, Paris

Campion N 1994 *The Great Year: Astrology, Millenarianism and History in the Western Tradition*. Penguin, Harmondsworth, UK

Carey J (ed.) 1999 *The Faber Book of Utopias*. Faber and Faber, London

Chesterton G K 1905 *Heretics*. The Bodley Head, Oxford, UK

Ferns C 1999 *Narrating Utopia: Ideology, Gender, Form in Utopian Literature*. Liverpool University Press, Liverpool, UK

Forster E M 1965 'The Machine Stops' in *Collected Short Stories of E. M. Forster*. Sidgwick and Jackson, London

Huxley A 1962 *Island*. Chatto & Windus, London

Huxley A 1963 *Brave New World*. Bantam Classic, New York

Jennings L 1996 Dystopias. In: Kurian G T, Molitor G T T (eds.) *Encyclopedia of the Future*. Macmillan Library Reference, USA, New York

Kolakowski L 1990 The death of utopia reconsidered. In: Kolakowski L (ed.) *Modernity on Endless Trial*. Chicago University Press, Chicago

Manuel F E 1965 *The Prophets of Paris*: *Turgot*, *Condorcet*, *Saint-Simon*, *Fourier and Comte*. Harpertorch Books, New York

Manuel F E (ed.) 1966 *Utopias and Utopian Thought*. Houghton Mifflin, Boston

Mercier L S 1971 *L'An deux mille quatre-cent quarante*. Ducros, Paris

More T 1965 *Utopia*. Penguin, Harmondsworth, UK

Morris W 1970 *News from Nowhere*. Routledge, London

Orwell G 1983 *Nineteen Eighty-Four*. Penguin, Harmondsworth, UK

Plato 1941 *The Republic*. Oxford University Press, Oxford, UK

Wells H G 1914 *The World Set Free*. Dutton, New York

Wells H G 1923 *Men Like Gods*. Macmillan, New York

Wells H G 1985 *In the Days of the Comet*. Hogarth, London

Wells H G 1987 *A Modern Utopia*. University of Nebraska Press, Lincoln, NE

B. Cazes

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

V

Valency and Argument Structure in Syntax

The purpose of this article is to describe the valency and argument structure of the verb or predicate. The notion of argument is known to have been adapted from mathematics. In mathematics, a function may take one or more arguments, which will be used to determine the value of that function. In linguistics, argument is used as a technical term for syntactic elements that combine obligatorily with the verb (or the predicate as the verb is usually referred to in relation to arguments). For instance, the English verb *die* must take at least one syntactic element, e.g., *the mayor*, to form sentences, e.g., *The mayor died*. This syntactic element is taken to be the (sole) argument of the predicate *die*. The verb *kill*, on the other hand, requires two arguments to form sentences, one representing the killer (*the man*) and the other the person killed (*the mayor*), as in *The man killed the mayor*. The number of arguments that are required by a predicate is then specified in the argument structure of that predicate. Additional syntactic elements, e.g., *in the office* or *in 1878*, can also potentially combine with the predicate *die*, e.g., *The mayor died in the office, The mayor died in 1878*, or *The mayor died in the office in 1878*. But these are regarded only as optional, because their absence has no bearing upon the grammaticality of the 'minimal' sentence, *The mayor died*. Such optional syntactic elements are referred to as adjuncts. The distinction between arguments and adjuncts, albeit not always easy to make, is a useful one (see Comrie 1993 for discussion as to how arguments and adjuncts should be distinguished).

The argument structure also contains information concerning the relation between the predicate and its argument(s). For instance, in the sentence *The man killed the mayor*, the predicate *kill* and its arguments *the man* and *the mayor* have logical (or grammatical) relationships between them with the effect that *the man* is the actor (or the subject) who carried out the killing, and *the mayor* is the undergoer (or the direct object) who was affected by the actor's killing (see *Grammatical Relations*). Adjuncts, e.g., *in the office* in *The mayor died in the office*, also bear logical (or grammatical) relationships, e.g., location (or oblique), with the predicate. Information on adjuncts, however, is not included in the argument structure of the predicate itself.

The term valency (or valence in the USA) is said to have been borrowed from chemistry into syntax. In chemistry, the valency of a chemical element is defined as the capacity of that element for combining with a fixed number of atoms of another element, for instance, two hydrogen atoms combining with an atom of oxygen in the case of water (i.e., H_2O). The notion of valency, when applied to verbs, refers to the number of arguments that are required for a given verb to form sentences. For example, the verb *die* is regarded as having a valency structure of a single argument. Thus, verbs such as *die* are characterized as one-place (or monovalent) verbs. The verb *kill*, on the other hand, is a two-place (or divalent) verb, because it requires two arguments.

The rest of the article is concerned principally with illustrating how argument structure can be altered or manipulated with or without a concomitant change of valency. How languages may keep the valency of the derived verb from exceeding that of the basic verb by manipulating argument structure will also be illustrated (see *Morphology in Linguistics* for derivation). Finally, how valency-changing operations are handled in current grammatical theories will be discussed briefly.

1. Change of Argument Structure with Change of Valency

There are two main mechanisms through which the argument structure of the predicate can be altered, together with a change of valency: argument reduction, and argument addition. In argument reduction, one of the arguments of the basic predicate may be demoted to adjunct status. In argument addition, on the other hand, adjuncts may be promoted to argument status, or new arguments may be introduced into the argument structure. There are five important grammatical devices utilized in argument reduction or argument addition: passivization, antipassivization, noun incorporation, advancement, and causativization. The first three belong to argument reduction, whereas the last two fall under argument addition. (But for examples of noun incorporation consisting of argument reduction and addition, see Mithun 1984, and Sect. 2 for causativization involving both argument addition and reduction.)

1.1 Passivization

In the passive clause, one of the arguments of the basic two-place predicate (i.e., the actor) is demoted to an adjunct with the other argument (i.e., the undergoer) retaining its argument status (but being promoted to the subject of the passive clause). This operation is referred to as passivization (Siewierska 1984). The number of arguments of the basic predicate is thus reduced from two (i.e., the actor and undergoer) to one (i.e., the undergoer) in the corresponding passive clause. This is substantiated by the fact that it is possible to omit the newly created adjunct. This change of valency (or valency decreasing) is signaled on the verb either by a verbal affix or by a periphrastic element in the verb phrase. Korean provides a good example of the passive:

(1) Korean
cuin-i totuk-lil cap-əss-ta
owner-NOM thief-ACC catch-PST-IND
'The owner caught the thief.'

(2) Korean
totuk-i (cuin-eke) cap-hi-əss-ta
thief-NOM (owner-by) catch-PASS-PST-IND
'The thief was caught (by the owner).'

[Abbreviations used to gloss the data: ABS = Absolutive; ACC = Accusative; APASS = Antipassive; APPL = Applicative; ASP = Aspect; CS = Causative Suffix; DAT = Dative; DO = Direct Object; ERG = Ergative; F = Feminine; FUT = Future; IND = Indicative; INTR = Intransitive; IO = Indirect Object; LOC = Locative; M = Masculine; NFUT = Non-Future; NOM = Nominative; PASS = Passive; PERF = Perfective; PRS = Present; PST = Past; TR = Transitive.]

1.2 Antipassivization

Antipassivization is often regarded as the counterpart of passivization in languages with the ergative–absolutive case system. But it should not be inferred from this that the passive is never found in languages with the ergative–absolutive system; both the passive and the antipassive may occur in the same language. In the antipassive, the actor argument of the basic two-place predicate remains the (sole) argument of the corresponding antipassive (albeit marked by the absolutive case), with the undergoer argument demoted to adjunct status or, optionally, omitted. If and when expressed, however, the undergoer is marked accordingly by an oblique case. This suggests that the antipassive is as much intransitive or detransitivized as is the passive: a decrease in valency from two (i.e., the actor and undergoer) to one (i.e., the actor) argument. The verb in the antipassive also bears an affix as a marker of valency decreasing. Dyirbal provides a good example of the antipassive:

(3) Dyirbal
balan $d_y$ugumbil baŋgul yaṛa-ŋgu
F-ABS woman-ABS M-ERG man-ERG
balga-n
hit-NFUT
'The man is hitting the woman.'

(4) Dyirbal
bayi yaṛa bagun $d_y$ugumbil-gu
M-ABS man-ABS F-DAT woman-DAT
balgal-ŋa-n_yu
hit-APASS-NFUT
'The man is hitting the woman.'

1.3 Noun Incorporation

Noun incorporation is another type of valency-changing operation found in some languages of the world, whereby one of the arguments of the predicate becomes part of the predicate, thereby losing its argument status (Mithun 1984). The resulting predicate may function as a single word. In Chukchi, for instance, vowels of incorporated nouns must participate in word-internal vowel harmony in conjunction with host verbs:

(5) Chukchi
tumg-e na-ntəwat-ən kupre-n
friends-ERG set-TR net-ABS
'The friends set the net.'

(6) Chukchi
tumg-ət kopra-ntəwat-gʔ at
friend-NOM net-set-INTR
'The friends set nets.'

Note that, with the undergoer incorporated into the predicate, the actor argument of the basic two-place predicate in (5) loses its ergative marking, and is instead marked by the nominative case in (6). In Chukchi, noun incorporation also motivates the verb to shed its transitive marking in favor of the intransitive marking. These changes indicate that noun incorporation involves reduction of two arguments of the basic predicate to one (i.e., valency decreasing).

1.4 Advancement of Adjuncts to Arguments

So far, argument reduction has been illustrated. In many languages, it is also possible to do the opposite, that is, to increase the valency of the basic verb by promoting an adjunct to argument status. In order to signal such promotion or advancement, the verb is normally marked by a so-called applicative affix. In Kinyarwanda, for instance, instrumental adjuncts can be promoted optionally to argument or direct object status. Compare the following two sentences:

(7) Kinyarwanda
úmwáalímu a-ra-andik-a íbárúwa
teacher he-PRS-write-ASP letter
n'ííkárámu

with-pen
'The teacher is writing a letter with the pen.'

(8) Kinyarwanda
úmwáalímu a-ra-andik-iish-a íbárúwa
teacher he-PRS-write-APPL-ASP letter
íkárámu
pen
'The teacher is writing a letter with the pen.'

The instrumental adjunct in (7) is promoted to direct object in (8), evidence for which can be adduced from the fact that the instrumental nominal, which is marked by the instrumental prefix in (7), appears as a bare nominal (i.e., no case marking) in (8). This advancement must also be registered on the verb by the applicative suffix-*iish*, as in (8). Both the old direct object and the newly created direct object in (8) exhibit grammatical properties of objecthood. This suggests that the valency of the derived verb in (8) is greater by one argument than that of the basic verb in (7) (see Sect. 3). In (7), there are two arguments (i.e., *teacher*, and *letter*) and one adjunct (i.e., *pen*). In (8), on the other hand, there are three arguments (i.e., *teacher*, *letter*, and *pen*). Moreover, the argument structure of the predicate in (7) specifies the relation between the actor and the undergoer, and the argument structure of the derived predicate in (8) the relation between the actor, the undergoer, and the instrument. To put it differently, (7) without the instrumental phrase *n'ííkárámu*, will remain grammatical, whereas (8), with *íkárámu* deleted, will be rendered ungrammatical.

1.5 Causativization

In common with advancement of adjuncts, causativization adds an argument to the argument structure of the basic predicate. However, it does so not by converting an existing adjunct into an argument, but by introducing a completely new argument into the argument structure. This new argument is the causer argument. Thus, if the basic verb is a one-place verb, the corresponding causative verb will be a two-place verb (the newly added causer argument and the sole argument of the original basic verb). If the basic verb is a two-place verb, the corresponding causative verb will be a three-place verb (three arguments including the causer argument). This is illustrated by the following pair of Turkish sentences. As opposed to the basic predicate in (9), the causative predicate in (10), with the causative suffix -*t*, has the causer argument, *dişçi*.

(9) Turkish
müdür mektub-u imzala-dı
director letter-DO sign-PST
'The director signed the letter.'

(10) Turkish
dişçi mektub-u müdür-e imzala-t-tı
dentist letter-DO director-IO sign-CS-PST
'The dentist made the director sign the letter.'

The valency of the causative verb thus exceeds that of the basic verb by one argument as can be neatly expressed in:

(11) NONCAUSATIVE: n/CAUSATIVE: $n+1$
[n: number of arguments]

The formula in (11) indicates that causativization is a case of argument addition, with the status of the argument(s) of the basic predicate remaining intact. This may indeed be true of one-place verbs; causativization only increases the valency of the basic one-place verb from one to two arguments. Causativization of two or three-place (or trivalent) verbs, on the other hand, would give rise to three or four arguments, respectively, more than many languages can cope with—regardless of whether verbs are basic or derived. Indeed, there is much cross-linguistic variation in causativization of two- or three-place verbs. For instance, there are languages which derive causative verbs only from one-place verbs (Song 2001, Dixon and Aikhenvald 1997, Dixon 2000). Moreover, in many languages, argument structure is manipulated in such a way that the valency of the derived verb is prevented from exceeding that of the basic verb (Comrie 1989, Song 1996, Dixon 2000).

2. Change of Argument Structure without Change of Valency

There are languages which introduce a new argument into the argument structure of the causative verb, only if one of the arguments of the basic two- or three-place verb is demoted to an adjunct, or omitted completely. In other words, causativization of the basic two- or three-place verb in such languages results in no increase in valency, contrary to the formula in (11), bur rather the valency of the derived causative verb is equal to that of the basic verb. In Babungo, for instance, when the causative suffix -*s* is added to two-place verbs, the direct object of the basic verb must be omitted or expressed as an adjunct. Thus, the direct object in (12) appears as an optional prepositional phrase in (13).

(12) Babungo
ŋwɔ́ fɛ̀e zɔ̃
he fear-PERF snake
'He was afraid of a snake.'

(13) Babungo
mə̀ fɛ̀sə ŋwə (nə̀ zɔ̃)
I fear-CS-PERF him (with snake)
'I frightened him with a snake.'

This phenomenon suggests that, regardless of whether the predicate is basic or derived, languages allow only a certain number of arguments per predicate (i.e., two or three) (Song 1996, Kemmer and Verhagen 1994 from a cognitive–linguistic perspective). Other devices drawn upon in causativization for this purpose include antipassivization and noun incorporation. The net effect of the use of these devices, in conjunction with

causativization, is no change of valency, although there may be a change of argument structure between the basic predicate and the derived predicate: addition of the causer argument offset by demotion to adjunct status (or omission) of one of the arguments of the basic predicate.

3. Doubling of Grammatical Relations

One theoretical issue in argument addition concerns the question of whether or not identical coding of nominals entails doubling of grammatical relations. In some languages, two nominals in the causative construction may be coded identically, for instance. In Kinyarwanda causatives, both the actor argument (or subject) and the undergoer argument (or direct object) of the basic predicate are coded identically as objects. Comrie (1989) interprets such identical coding to be indicative of doubling of the direct object. This position, however, has been challenged by Polinsky (1994), who demonstrates that the two identically-coded nominals in Kinyarwanda causatives behave differently in terms of syntactic properties. In fact, only one of the identically-coded 'direct objects' is syntactically a direct object, the other being an indirect object, identical coding notwithstanding. Kozinsky and Polinsky (1993) describe more or less the same situation in Korean and Dutch causatives. Similarly, the old direct object and the newly created direct object in the Kinyarwanda applicative sentence (8) are bare nominals (that is, coded identically). Moreover, both nominals exhibit grammatical properties of objecthood. This also led to a controversy as to whether or not (8), as opposed to (7), has two direct objects, i.e., doubling of the direct object. Gary and Keenan (1977) take the view that (8) has two direct objects. Perlmutter and Postal (1983), on the other hand, argue that (8) does not have two direct objects, but rather one direct object and one indirect object. Further evidence against doubling of direct object in Kinyarwanda applicative constructions is provided by Dryer (1983) and Polinsky (1994).

4. Theoretical Approaches to Valency-changing Operations

Current grammatical theories have proposed different approaches to representation of valency-changing operations; more than one approach may also be possible within one and the same grammatical theory (Brown and Miller 1996). These approaches can be characterized broadly as lexical or syntactic. Lexical Functional Grammar regards valency-changing operations as productive lexical processes giving rise to alternative assignments of grammatical functions (or grammatical relations) to arguments. In mainstream Government and Binding Theory, valency-changing operations are treated as syntactic. In causativization, for instance, the basic verb undergoes movement to combine with the causative affix at an abstract syntactic level. This movement is then claimed to comply with the same principles that are obeyed by other movements in syntax. Relational Grammar invokes revaluation rules to handle valency-changing operations. These revaluation rules are regarded as syntactic, because they operate on grammatical relations between initial and noninitial strata (or different syntactic levels) of the relational structure.

See also: Constituent Structure: Syntactic; Linguistics: Incorporation; Syntactic Control; Syntax; Syntax–Semantics Interface

Bibliography

Brown K, Miller J (eds.) 1996 *Concise Encyclopedia of Syntactic Theories*. Pergamon, Oxford, UK

Comrie B 1989 *Language Universals and Linguistic Typology*. Blackwell, Oxford, UK

Comrie B 1993 Argument structure. In: Jacobs J, von Stechow A, Sternefeld W, Vennemann T (eds.) *Syntax: An International Handbook of Contemporary Research*. Walter de Gruyter, Berlin

Dixon R M W 2000 A typology of causatives: Form, syntax and meaning. In: Dixon R M W, Aikhenvald A Y (eds.) *Valency Changing: Case Studies in Transitivity*. Cambridge University Press, Cambridge, UK

Dixon R M W, Aikhenvald A Y 1997 A typology of argument-determined constructions. In: Bybee J, Haiman J, Thompson S A (eds.) *Essays on Language Function and Language Type*. John Benjamins, Amsterdam, The Netherlands

Dryer M S 1983 Indirect objects in Kinyarwanda revisited. In: Perlmutter D M (ed.) *Studies in Relational Grammar 1*. University of Chicago Press, Chicago

Gary J O, Keenan E L 1977 On collapsing grammatical relations in Universal Grammar. In: Cole P, Sadock J (eds.) *Syntax and Semantics 8: Grammatical Relations*. Academic Press, New York

Kemmer S, Verhagen A 1994 The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* **5**: 115–56

Kozinsky I, Polinsky M 1993 Causee and patient in the causative of transitive: Coding conflict or doubling of grammatical relations? In: Comrie B, Polinsky M (eds.) *Causatives and Transitivity*. John Benjamins, Amsterdam, The Netherlands

Mithun M 1984 The evolution of noun incorporation *Language* **64**: 847–94

Perlmutter D M, Postal P M 1983 Some proposed laws of basic clause structure. In: Perlmutter D M (ed.) *Studies in Relational Grammar 1*. University of Chicago Press, Chicago

Polinsky M 1994 Double objects in causatives: Towards a study of coding conflict. *Studies in Language* **19**: 129–221

Siewierska A 1984 *Passive: A Comparative Linguistic Analysis*. Croom Helm, London

Song J J 1996 *Causatives and Causation: A Universal-Typological Perspective*. Addison Wesley Longman, London

Song J J 2001 *Linguistic Typology: Morphology and Syntax*. Pearson Education, Harlow, UK

J. J. Song

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

 ISBN: 0-08-043076-7

Value of Life, Economics of

The economics of the value of life is about what individuals and societies are willing to sacrifice to get longer expected lifetimes. Value of life is about private choices that individuals make implicitly and explicitly about their own health and safety. Value of life is also about collective, public choices that societies make concerning tax and expenditure programs which affect expected lifetimes and regulations that shape the environment in which individual decisions are made. Inferences from individual choices inform public decisions and expert-based public policy is offered to improve the quality of individual decisions and consequences.

1. The Term 'Value of Life'

Value of life typically refers to an amount of money an individual is willing to trade for a small change in his or her own probability of survival. This definition is typical because this situation is typically what individuals face in life and what decision makers face in public policy. Exceptions exist, of course, and they will be addressed below.

1.1 Value of a Statistical Life

The typical situation is illustrated by thinking about an individual who is one of a group of 10,000 people. Everyone in the group is identical and knows that during the next year nine people in the group will die. Everyone also knows that the number of people who are going to die could be reduced to eight if sufficient funds can be raised. If somehow it is known that each individual is willing to pay \$500 for the reduction of one death in the group, then the value of life is \$500 per person times 10,000 people, or \$5 million. Sometimes this amount is referred to as the 'value of a statistical life' because the identity of the individual who would have been the ninth death, yet lives, is unknown at the beginning of the period when the decision is made.

1.2 Value of a Change in Mortality Risk

The numerical example illustrates another way of interpreting value of life. The reduction in the number of deaths implies that the probability of death faced by each individual decreases from 0.0009 to 0.0008. The \$500 that the individual is willing to pay reduces the probability of his or her death by 0.0001. Because information is available from similar situations with small changes in the risk of death, but the changes are not exactly the same, for convenience, the value is standardized to a unit change (1 to 0). The standardized 'value of a change in mortality risk' is \$500 times the 10,000 needed to standardize 0.0001 to 1, or \$5 million. Both the 'value of a statistical life' and the standardized 'value of a change in mortality risk' represent what is typically meant by value of life. For further illustration, see Viscusi et al. (1995, Chap. 20).

2. Characteristics of Value of Life

The nature of value of life as defined above can be described further by considering a simple equation of individual expected utility. Let

$$E(U) = PU(C) \tag{1}$$

where $E(U)$ is expected lifetime utility for an individual. U is a well-behaved utility function with marginal utility $U' > 0$ and diminishing marginal utility $U'' < 0$. C is consumption and P is the probability of survival. Differentiating Eqn. (1) keeping the individual at the same level of $E(U)$, and solving for $\mathrm{d}C/\mathrm{d}P$ yields the rate at which the individual is willing to trade off current consumption for a small change in the probability of survival:

$$\mathrm{d}C/\mathrm{d}P = -U(C)/PU' \tag{2}$$

Equation (2) says that the rate at which an individual is willing to trade consumption for a small change in mortality risk equals (the negative of) the utility of consumption divided by the expected marginal utility of consumption. The negative sign indicates a tradeoff, i.e., less C for more P. If consumption were expressed in monetary terms, U' would be the marginal utility of income, and PU' would be expected marginal utility of income. This would mean that the value of a change in mortality risk, $\mathrm{d}C/\mathrm{d}P$, would be the expected monetary value of the utility of consumption. It is the willingness to pay for risk reductions.

While the simple equation above captures the essence of the risk-dollar tradeoff, formal models of consumer and worker behavior include budget constraints, behavior which affects the probability of survival, insurance, workers' compensation, bequests, multiple periods, and multiple risks. For additional reading, see the book by Jones-Lee (1976), article by Rosen (1981), review by Jones-Lee (1985) and comprehensive book chapters by Cropper and Freeman (1991, pp. 165–211) and Freeman (1993, pp. 314–66).

2.1 Relationship with Earnings

Within the context of intense interest in economic growth and national income accounting and from an emphasis on manufacturing and machinery following World War II grew an appreciation for the potentially high returns from investment in human beings. In a special conference issue of the *Journal of Political Economy* in 1962, Mushkin (1962) articulated the case for investment in human beings through promoting better health. She argued that the return

would be through future increases in labor earnings. If one more person were to live because of an investment in health, then national income would be greater by the amount that person would earn in the labor market. Expected future or forgone labor earnings became a standard measure of the value of life. Notwithstanding the fundamental contribution made by establishing human health as an important factor in economic growth, reference back to Eqn. (2) yields no clear, precise connection between the value of life and labor earnings. P may be close to 1, and C may be positively related to labor earnings, but the primary difference is that the value of $U(C)$ is not likely to equal C. No allowance is made for consumption which occurs outside the formal market, and no allowance is made for the value of living. Schelling (1968, pp. 127–76) clearly distinguished between earnings, which he referred to as livelihood, and value of life or living. What matters is the value of the statistical life, or value of change in mortality risk, *to the individual who is affected*. The value to the individual influences individual behavior and should be the starting value for public decisions which affect risks to the individual's life. This point is a reminder that either gross national product, or national income measured with traditional accounting, is a highly imperfect and sometimes misleading indicator of well being.

2.2 Costs of Illness

In valuing investments in health, Mushkin (1962) added to earnings the saving of future health expenditures. The sum of these health expenditures saved (direct costs) and the labor earnings not forgone (indirect costs) became known as the 'cost of illness' avoided. Cost of illness was used extensively because of the wide acceptance of the importance of investment in human beings and because estimation was manageable. The shortcoming that homemakers would be valued at zero was overcome by estimating the value of the services performed or the amount of earnings the person would have had if the individual had worked in the labor market. The low values for infants whose future earnings were heavily discounted and zero values for retirees who are not in the market were troublesome. The most troublesome aspect of cost of illness, however, is the lack of a theoretical basis which includes the individual's own value of living. Further discussion can be found in Tolley et al. (1994), who provide a comparison of the cost of illness and the value of life based on an individual's willingness to trade money and risk.

2.3 No Single Value

Equation (2) is useful for demonstrating that there is no single value of life which applies to all individuals in all circumstances. The amount an individual is willing to trade for a small change in his or her probability of survival depends on P, the base level of the probability. Because P is in the denominator, the value of life can be expected to increase as P decreases. For example, adult aging reduces the probability of survival and can be expected to increase the value of life, holding constant consumption. Automobile drivers who must travel alone at night on winding roads covered with snow and ice can be expected to have higher values of life than at other times.

The value of life depends on idiosyncratic factors that are reflected in the utility function. $U(C)$ can vary among individuals with the same measured consumption, C, owing to differences in family or environment. Individuals enjoying close, supportive family relationships and living in clean, ecologically rich natural environments and societies with high-quality civil institutions and public services can be expected to have higher values of life.

Lastly, from Eqn. (2) the value of life depends upon C. Since C can be expected to increase with income, $U(C)$, and the value of life will be higher with higher income and consumption.

2.4 Ex Ante and Unidentifiable

Equation (2) displays another characteristic which is inherent in the situations illustrating the value of life. The situations for the value of a statistical life with one fewer person of 10,000 people dying and the value of reduction in mortality risk by 0.0001 share an *ex ante* perspective. The situation is confronted, the life lottery must be played, but the outcome is not yet known. The identity of the individual who would have been the ninth death among the 10,000 is not known. Or, whether the individual facing a lower probability of death survives or dies is not known. Decisions which must be made by individuals and by societies about risks to their health, safety, and environment are often made in a similar manner.

Situations in which the individual is identifiable are inherently different. If the individual is thrust into a life-threatening situation by an unforeseen natural event, such as a tornado which causes entrapment in a collapsed building, the individual and society will often spend whatever is available or do whatever is possible to save the life. The situation involves a potentially large change in survival for a known individual. Saving the individual's life from almost certain death is usually valued very highly. If the trapped person is saved, all may end well. If the person cannot be found in the collapsed building after some time, then the probability of saving the person becomes small and the rescue effort decision is transformed into one involving the more typical value of a change in mortality risk. Medical situations involving identifiable patients with known diagnoses have some similarities and are considered below.

Court cases involving wrongful death of a specific individual are also inherently different because they too are *ex post*. Nothing can change the fact that the individual is dead. Nothing can be done to compensate the deceased for death. The value relevant to forensic decisions is the amount that can compensate the survivors or the estate for the individual's death. The value depends on the deceased individual's contribution to the well being of the other members of the family, an amount that is related to labor market earnings. The amount that fully compensates the family for loss of a loved one can be practically infinitely large. From an efficiency perspective, the value relevant for forensic situations should consider generating sufficient incentive to influence future decision makers who deal with similar risks. The typical value of life may well be relevant, but the differences between the values in court cases and the values in benefit–cost analysis tend to be large (see Jones-Lee's (1982) book). The value of life in wrongful death cases is likely to be an area of great interest in future research. Experiences in natural resource damages assessment under Superfund in the USA could be relevant. They too are values for *ex post* compensation. See Kopp and Smith (1993) for issues involving damage assessment.

2.5 *Willingness to Pay and Willingness to Accept*

In contrast to the court cases about *ex post* wrongful death of a specific individual, Eqn. (2) shows the tradeoff an individual is willing to make between consumption and a small change in mortality risk. For small changes Eqn. (2) suggests that it makes no difference whether the individual is paying for small reductions in mortality risk or accepting compensation for small additions to mortality risk. For example, for small, equal (and opposite) changes in mortality risk, the amount the individual is willing to pay for a slightly safer job is approximately the same as the amount the individual is willing to accept for a job which is slightly riskier.

3. *Criticisms of a Value of Life Based on Individual Willingness to Trade Off*

A value of life based on individual preferences as defined above is not without criticism. Criticism ranges from the ethical to the technical.

3.1 *Deontology*

One saying goes that economists know the price of everything and the value of nothing. In this spirit, one objection to the value of life as defined above is that it is vulgar and ethically wrong to place a monetary value on human life. A deontologist is one such as Plato or Kant whose moral theory is based on obligate moral rules independent of the desirability of the consequences of acting on the rules. If it is morally wrong that people are exposed to risks of death, then it is wrong to construct a value of life to be used to guide decisions involving risks of death. Systematic choices involving risk are practically impossible.

In contrast, a teleologist such as Bentham or Mill whose moral theory is based on the goodness of the consequences is highly interested in a value of life which might guide decisions so as to generate the greatest good for the greatest number. While strident deontology can lead to paralysis in decision making when faced with risky situations, ardent teleology can produce policy recommendations, such as euthanasia, which a majority often deems morally wrong. Because teleology is the ethical basis of benefit–cost analysis which uses the value of life, its criticism by deontologists is understandable. For more on a deontological objection, see Brandt-Rauf and Brandt-Rauf (1980), who analyze the ethical conflict over occupational safety and health.

3.2 *Ex Ante Becomes Ex Post*

Another ethical criticism is that neither *ex ante* nor *ex post* willingness to pay is acceptable because at some time the lottery is played the individual knows what is going to happen and will either pay everything to avoid death or demand an infinitely large amount in compensation for death. This situation, Broome (1978) argues, is morally unacceptable. Broome's criticism is complex, but it seems to have a deontological dimension to it. While this may be a quandary, the fact individuals frequently make decisions in which they are tradeoff risk suggests that the tradeoffs are not completely reprehensible to individuals.

3.3 *Murdering Statistical Lives*

Recall that there is no single value of life and that one factor which determines the value is the base level of the probability of survival, P. As P decreases, reflecting the increased risk of death faced by the individual, the value of a specified amount of risk increases. If decision makers for public health and safety programs use the values of the individuals as a guide, then programs will be directed at individuals in high risk situations. With a fixed budget for managing risks, the prospect of spending everything on a few high-risk individuals at the expense of numerous individuals facing lower risks is unattractive. This murdering of statistical lives is something that Linnerooth (1982, pp. 229–61) believes should be made explicit in social decisions which affect the distribution of mortality risk in society. She argues that society should care about the distribution of risk just as it cares about the distribution of income. Another concern about too little attention to low-risk situations is based not on ethics, but on perceptions. If individuals perceive risk and changes in risk less accurately when faced with

low risk of death, then they will not make efficient decisions about their own health and safety. It follows that public programs which target reductions in risk for low-risk situations would be especially important for improving the well being of society. Jones-Lee (1985) provides a review of these and other criticisms of value of life defined as the amount of money an individual is willing to trade for a small change in his or her own probability of survival.

4. Estimates of Value of Life

Value of life is about private choices that individuals make implicitly and explicitly about their own health and safety. Bland, healthy diets are traded for tasty food and drink which increase blood pressure and cholesterol levels. Comfortable, safe dwellings are left for skiing and rock climbing. Jaywalking is done to save valuable time. Time and sweat are spent exercising to improve physical condition. Travel is postponed in order to avoid treacherous conditions and increase the chances of a safe trip. Jobs are given up to pursue a safer, healthier lifestyle. If the changes in risk and the values of whatever is traded are known by the individuals and can be known by researchers, then values of life can be estimated.

Values of life are implied by observable behavior. If realistic risk tradeoff scenarios can be constructed, then values of life can be elicited directly. Values of life have been estimated using these approaches at least since the early 1970s. The models upon which the estimation is based typically use a utility function which in some way resembles the one shown in Eqn. (1), make P endogenous so that individuals can change the probability of survival through their behavior, and finally introduce a resource constraint to reflect limited money and time. An implication of these models is that individuals will engage in risk-changing activity to the extent that the marginal value of the activity equals the marginal cost. Letting V be the value of life as defined in Eqn. (2), P' the change in the probability of survival, and K the cost of changing P, the condition of interest is

$$VP' = K \tag{3}$$

If P', the change in risk, and K, the cost of changing P, can be determined, then V, the value of life implied by individual behavior, can be estimated. For comprehensive presentations of models for estimation of the value of life, see Cropper and Freeman (1991), Freeman (1993), Johansson (1995) and Johannesson (1996).

4.1 Risk-compensating Wages

The labor market has proven to be a fertile area for producing estimates of individuals' willingness to trade off safety and money. Faced with an array of jobs with different characteristics, workers will choose jobs that suit them best. If two identical workers are confronted with two jobs which are identical except for risk of death associated with the job and pay, then equilibrium implies that the worker who chooses the riskier job must be compensated by a higher wage. This risk-compensating wage difference implies a value of life. In Viscusi's (1992) survey of the literature on the value of risks to life and health, he lists 24 studies which report estimates of the value of life based on analysis of labor markets. The studies use data from various years from 1967 to 1986 and from several countries including the USA, UK, Canada, Australia and Japan. The risk variables are based on actuarial data, industry and occupation data collected by the Bureau of Labor Statistics, and worker assessments. He judges that the most reasonable estimates of value of life are clustered in the \$4–9 million range, in 1999 dollars.

Several concerns exist about the reliability of these estimates. The fatality rates used to measure risk may not reflect actual risks on the job. Worker information about the risks may be lacking or worker perception of the risks may not reflect actual risks. Risks of nonfatal injury risks may be correlated with fatality risks and may not be measured well or at all. Unmeasured and omitted job disamenities may be correlated with fatality risks. Estimates of values of life may not be representative because of self-selection of more adept or less risk averse individuals into risky jobs.

4.2 Risk Tradeoffs in Consumption

Markets other than the labor market and activities other than work have also produced estimates of individuals' willingness to trade off safety and other things they value. By estimating the effect of air quality on housing values and isolating the effect of air quality on mortality risks, the value of life implicit in the housing market can be estimated. By estimating the effect of the occupant fatality rate on the prices of automobiles, the value of life implicit in the automobile market can be estimated. These situations match well with Eqn. (3) in that numerous options are available from which the consumer can choose. Other consumption decisions are less continuous, but still imply values of life. Wearing a seat belt while driving in a motor vehicle involves a tradeoff between the gain in safety versus the time and discomfort associated with belt use. Installing smoke detectors involves a tradeoff between a reduction in risk of a fatal fire and installation costs. The value of life estimates from consumption studies tend to be less than the estimates from the labor market and fall in the range of \$1–5 million.

The concerns about the reliability of these estimates are similar to the estimates from the labor market. Some concerns can be readily addressed. For example,

if risks are misperceived according to known relationships, the implied values of life can be adjusted. Two other concerns may not be addressed as easily. One is about the ability to separate the risk of fatality from other characteristics of the product. The other is the ability to estimate the value of disutility associated with the consumption.

4.3 Contingent Market Values

Early in the emergence of value of life based on individual willingness to pay, Mishan (1971, p. 174) encouraged economists to employ a questionnaire method to estimate values of life rather than rely upon the forgone earnings or cost of illness measures. Since Mishan's exhortation, economists working primarily in environmental, safety, and health economics have developed the contingent valuation method. Mitchell and Carson's (1989) book on using surveys to value public goods presents the method after 18 years of development. This method elicits the tradeoff choice through direct questioning of the individual. The hypothetical setting or market is described, the choice is presented, and information about the respondent is collected. The survey might be in person, by mail, by phone, or some combination. Individuals have been presented with more rapid-response ambulance service for higher taxes, changes in risk on the job for compensating changes in wages, safer highway travel for more money, and safer medications for more out-of-pocket expense. Risk–dollar, risk–risk, and hybrid risk–risk and risk–dollar tradeoffs have been designed to elicit values of life. Contingent valuation estimates tend to fall in the range of the estimates from the labor market.

A great advantage of contingent valuation is the ability to provide information and describe the tradeoff precisely. Concerns exist about the reliability of estimates based on contingent valuation, and they are debated in the symposium in the Fall 1994 issue of the *Journal of Economic Perspectives*. Two of the most worrisome for estimating the value of life is the hypothetical bias due to not actually having to pay and the insufficient sensitivity of values to the scale of the change in risk. Future research on 'cheap talk' and statements of intensity of commitment to pay and research on ways to communicate risk may increase confidence in these estimates.

5. Policy Considerations Beyond the Typical Values of Life

5.1 Values for Various Types of Individuals

Benefit–cost analysis of environmental, health, and safety proposals value anticipated changes in mortality risks using the average of values of life for all members of society. In this way the values which individuals place on their own lives inform public policy. For public policies such as improvements in traffic safety which have broad effects across many types of individuals in society, average values of life are appropriate. However, some policies primarily affect specific types of individuals who may have values of life which differ from the typical individual. Whether it is socially acceptable to value lives differently is a sensitive issue. A consensus seems to be emerging that it is unacceptable to use different values for groups which differ by income; an average value of life across all of society is normally used. While no consensus exists yet, there is interest in estimating values of life for vulnerable groups such as the elderly and children. Future research is likely to focus on values of life for types of individuals who get special attention in health, safety, and environmental policy. Related is the question about discounting changes in risk which occur in the future. Discounting future risks can be relevant to people of all ages, but it can matter greatly for the very young and old. Contingent valuation studies in the USA and Sweden which address age and discounting find lower values of life for the elderly and that future changes in mortality risks are discounted substantially (see Cropper et al. (1994), Johannesson et al. 1997).

5.2 Values of Others' Lives—Altruism

Values which individuals place on changes in their own probabilities of survival may be less than the value to all of society because the values that others place on the individual's life are not counted. The strongest case for including others' values is within the individual's immediate family. Since people sort themselves into households based on mutual caring it makes sense that interdependencies are greatest within the household. Bonds within the household are thought to be strong enough that decision making is sometimes modeled at the household level and transfers are made within the household. Differences between individual value and a household value are likely to be dominated by the other factors which influence estimates of the value of life and not produce much change in the range of values relevant for public policy.

Beyond the household it is not clear that the values others have for an individual's change in the probability of survival should be included in benefit–cost analysis. In a general model of a society constrained by limited resources, it can be demonstrated that the value of a statistical life is the same with universal pure altruism with caring about the overall well being of others and universal pure self-interest with no caring at all for others. In other words, the individual's value of his or her change in risk need not be supplemented

for benefit–cost analysis. This result follows from the recognition that adding others' values causes the individual to consume more health and safety and less of other desirable things than the individual would choose to consume. The individual's value is sufficient as long as there is not substantial concern with the distribution of income. With impure, safety-focused, paternalistic altruism, however, it can be appropriate to augment the individual's value by others' willingness to pay. Future research is likely to explore ways in which safety-focused altruism should augment individuals' values of their own lives. For further readings, see Jones-Lee (1992) and Johansson (1995).

5.3 Fiscal Externalities and Burden

Another reason for considering adjusting the individual value of life is the impact that an individual's death has on others through public tax and expenditure programs. In Martin Bailey's (1980) influential policy book, he advocated the use of individual value of life for guiding policy decisions. One adjustment he made was for the individual's contribution to the public tax and insurance system. The amount of adjustment is potentially greater the more extensive is the welfare state in the society of which the individual is a part. A topic for future research is whether the adjustments average out to zero for individuals living in a society with a balanced budget constraint where net contributions are presumably zero.

More recently, the idea of fiscal externalities has motivated something which might be called social burden analysis. The fiscal impact on others of individuals who smoke cigarettes, drink heavily, or have sedentary lifestyles has been estimated to quantify the costs which individuals who do these things impose on others. The motivation for calculating such costs is that they are potential pecuniary benefits of policies which reduce these activities. Interestingly, if this adjustment were made to individual values of life for benefit–cost analysis of, say, environmental programs which reduce the risks of individuals who engage in these activities, they would have lower than average values of life and the benefits of the environmental program would be less than if it affected groups with healthier lifestyles. However, the fiscal burden adjustment is rarely used this way. Also interesting is the variance in fiscal consequences which are included in the burden analysis. For example, a RAND Corp. study estimates that smokers, roughly speaking, pay their own way because of the excise taxes they pay on cigarettes and the amounts they bequeath to survivors through net contributions to pension plans. Others who do burden analysis, including those who determine what is admissible in court, are more selective in what they are willing to accept as relevant. They will count increases in some medical costs financed by others such as those associated with hospitalization but exclude other medical costs which would be saved, such as costs of long-term care. Social burden analysis awaits a theoretical grounding to put it on equal footing with benefit–cost analysis. For further reading on burden analysis, see Manning et al. (1991).

5.4 Quality-adjusted Life Years

Early evaluation of medical outcomes was in terms of life years added. The demand to recognize quality of life lead to the development of adjusting a life year added by a quality weight to yield quality adjusted life years, or QALY. The quality weight for a health state associated with an illness (or injury) can be determined by a direct questioning method. The questioning may be as straightforward as picking a point representing the health state on a straight-line scale between death and full health. It might be deciding what probability of full health in a standard gamble with full health and death as the possible outcomes is equivalent to the health state. QALYs place a value on the expected number of years of life.

QALYs emerged to be used in cost utility, or cost effectiveness in evaluating medical interventions for several reasons. One reason was the deontological repugnance to health professionals of putting a monetary value on changes in health and, especially, life. Another was the relative paucity and suspected unreliability of estimates of individual willingness to pay in the 1980s when QALYs were being developed. With the continued development of value of life estimation and the emergence of the contingent valuation method, future practice is likely to be a blending of values of life, QALYs, and related measures. For further reading, see Gold et al. (1996), Johannesson (1996), Sloan (1996) and Drummond et al. (1997).

6. What is Known

What may seem on the surface to be impossible, place a value on life, for practical purposes is straightforward. People, individually and collectively, make choices all the time in which they implicitly make tradeoffs between changes in their probability of survival and valuable time and money. Estimates of these values of statistical lives, or, alternatively, values of changes in mortality risks, come from analysis of jobs with different wages and risks, consumption decisions involving changes in risk and time and money, and from direct questioning involving risk-money tradeoffs in constructed markets. The evidence from a large number and variety of studies suggests the typical value of life falls in a range from $1 million to $9 million. The values from the consumption decisions tend to be in the lower end of the range and the best estimate of an average value is $5 million. Studies which compare values of life with expected changes in labor earnings (or cost of illness) invariably find the

value of life is greater. Clearly, value of life estimates based on willingness to pay are considerably more reliable than in the 1980s and a growing number of decision makers are finding values of life which are reliable enough for benefit–cost analysis and regulatory impact analysis.

Questions about risk information, perception, econometric estimation, contingent valuation, and the values of life estimates remain. Risk–risk analysis developed as an alternative to benefit–cost analysis in analysis of regulations. The central relationship is the connection between changes in disposable income and individual mortality. Valuing life directly is avoided. For examples, see the collection of studies by Graham and Wiener (1995) or the special issue of the *Journal of Risk and Uncertainty* in January 1997. As long as there are concerns about the ethical propriety of valuing life and about the reliability of value of life estimates, the incentive to develop alternatives to benefit–cost analysis based on values of life exists. As long as progress continues to be made in estimating values of life, their use in making policy will grow.

See also: Children, Value of; Consumer Economics; Cost–Benefit Analysis; Health Economics; Health Risk Appraisal and Optimistic Bias; Human Capital: Educational Aspects; Life Course: Sociological Aspects; Safety, Economics of; Wage Differentials and Structure

Bibliography

Bailey M J 1980 *Reducing Risks to Life: Measurement of the Benefits*. American Enterprise Institute for Public Policy Research, Washington, DC

Brandt-Rauf S I, Brandt-Rauf P W 1980 Occupational health ethics: OSHA and the courts. *Journal of Health, Politics, Policy and Law* **5**: 523–34

Broome J 1978 Trying to save a life. *Journal of Public Economics* **9**: 91–100

Cropper M L, Aydede S K, Portney P R 1994 Preferences for life saving programs: How the public discounts time and age. *Journal of Risk and Uncertainty* **8**: 243–65

Cropper M L, Freeman A M 1991 Environmental health effects. In: Braden J B, Kolstad C D (eds.) *Measuring the Demand for Environmental Quality*. North-Holland, New York, pp. 165–221

Drummond M F, O'Brien B, Stoddart G L, Torrance G W 1997 *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, New York

Freeman A M 1993 *The Measurement of Environmental and Resource Values: Theory and Methods*. Resources for the Future, Washington, DC

Gold M R, Siegel J E, Russell L B, Weinstein M C 1996 *Cost-effectiveness in Health and Medicine*. Oxford University Press, New York

Graham J D, Wiener J B 1995 *Risk vs. Risk: Tradeoffs in Protecting Health and the Environment*. Harvard University Press, Cambridge, MA

Johannesson M 1996 *Theory and Methods of Economic Evaluation of Health Care*. Kluwer, Boston, MA

Johannesson M, Johansson P-O, Löfgren K-G 1997 On the value of changes in life expectancy: Blips versus parametric changes. *Journal of Risk and Uncertainty* **15**: 201–19

Johansson P-O 1995 *Evaluating Health Risks: An Economic Approach*. Cambridge University Press, Cambridge, UK

Jones-Lee M W 1976 *The Value of Life: An Economic Analysis*. University of Chicago Press, Chicago

Jones-Lee M W 1985 The value of life and safety: A survey of recent developments. *The Geneva Papers* **10**: 141–73

Jones-Lee M W (ed.) 1992 *The Value of Life and Safety*. North-Holland, New York

Kopp R J, Smith V K (eds.) 1993 *Valuing Natural Assets: The Economics of Natural Resource Damage Assessment*. Resources for the Future, Washington, DC

Linnerooth J 1982 Murdering statistical lives …? In: Jones-Lee M W (ed.) *The Value of Life and Safety*. North-Holland, New York, pp. 229–61

Manning W G, Keeler E B, Newhouse J P, Sloss E M, Wasserman J 1991 *The Costs of Poor Health Habits*. Harvard University Press, Cambridge, MA

Mishan E J 1971 *Cost–Benefit Analysis: An Introduction*. Praeger, New York

Mitchell R C, Carson R T 1989 *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Resources for the Future, Washington, DC

Mushkin S J 1962 Health as an investment. *Journal of Political Economy* **70**: 129–57

Rosen S 1981 Valuing health risk. *American Economic Review* **71**: 241–5

Schelling T C 1968 The life you save may be your own. In: Chase S B (ed.) *Problems in Public Expenditure Analysis*. Brookings Institution, Washington, DC, pp. 127–76

Sloan F A (ed.) 1996 *Valuing Health Care*. Cambridge University Press, New York

Tolley G S, Kenkel D, Fabian R (eds.) 1994 *Valuing Health for Policy: An Economic Approach*. University of Chicago Press, Chicago

Viscusi W K 1992 *Fatal Tradeoffs: Public and Private Responsibilities for Risk*. Oxford University Press, New York

Viscusi W K, Vernon J M, Harrington J E 1995 *Economics of Regulation and Antitrust*. MIT Press, Cambridge, MA

G. C. Blomquist

Value Pluralism

'Value pluralism' has traditionally been understood as a metaphysical thesis about what values there ultimately are: there are many values that cannot be 'reduced' to a single supervalue. Although value monism has an impressive pedigree of proponents (e.g., Bentham, Mill, and, arguably, Aristotle and Nietzsche among others), it is now widely assumed to be false. The arguments for value pluralism, however, are surprisingly inconclusive. Indeed, they critically depend on the resolution of a neglected question in value theory: how are values properly individuated?

The metaphysical plurality of values has been thought to be important in two main ways. First, if values are plural, any theory that relies on value

monism is mistaken. So, for example, hedonistic utilitarianism and most forms of preference-utilitarianism must be rejected. Second, the plurality of values is thought to raise problems for rational choice. If justice, for instance, is irreducibly distinct from mercy, how can there be rational choice between them? Without reduction, it seems there is no common ground that justifies choosing one value over another.

On closer inspection, however, it turns out that metaphysical or 'reductive' value pluralism does not have the implications for rational choice that many have supposed. A charitable interpretation of value pluralist writings suggests a second, 'nonreductive,' form of value pluralism. Nonreductive pluralism is neutral on the metaphysical question of plurality but insists that in the context of choice, there are differences between values—whether or not those values reduce to a single supervalue—that have important implications for rational choice. Since the differences claimed to hold between values vary from author to author, nonreductive value pluralism is not itself a particular view about values but merely a convenient rubric under which a loose collection of different views about values may be grouped.

This article examines the main arguments for reductive value pluralism, argues that reductive value pluralism does not have certain implications it is widely thought to have, and outlines three forms of nonreductive value pluralism.

1. *Reductive Value Pluralism*

Value pluralists maintain that whatever values are, there are ultimately many of them: they do not all *reduce* to a single ultimate value. Exactly how this metaphysical thesis is to be understood depends on how the notion of reduction is to be understood.

Reduction in the context of values is best understood as an explanatory relation: if one value reduces to another, what it is to bear the one value is fully explained by what it is to bear, promote, or respect the other value. This reduction is neutral on the general ontological question of whether there 'really' are any values at all. The pluralist maintains only that there are many values, whether or not they are to be regarded as entities in their own right.

Two paradigmatic relations of explanatory reduction are 'is merely instrumental to' and 'is wholly constituted by.' (Others include 'is merely symbolic of,' 'is merely contributory to,' 'is merely a part of'). If one value is merely instrumental to another, there is nothing more to having the one value than promoting the value it is a means to. For example, if beauty is merely instrumental to pleasure, what it is to be beautiful is fully explained by the pleasure it brings. If one value is wholly constituted by another, there is nothing more to having the one value than being a way in which the other value is borne. For example, if the evening's pleasure is wholly constituted by the thrill at the gaming tables, what it is to have the evening's pleasure is fully explained by the thrill it is constituted by.

If all values reduce to a single value, that value is the only ultimate value, and value monism is correct. If however there is something more to two or more values than the values they are instrumental to, constituted by, etc., then those values are irreducibly distinct, and value pluralism is correct.

2. *Three Arguments for Reductive Value Pluralism*

Although there has been no agreement on which is the one value to which all others reduce, the view that there is such a value has two main attractions. If all values reduce to one, then values can be neatly systematized as instruments to, constituted by, etc., a single supervalue. Value monism appears to ensure a simple and elegant axiology, with a supervalue at the trunk of a structure that branches out to the other values that each derive from the supervalue in some way.

Moreover, if value monism holds, it seems that all conflicts between values are only apparent. For if there is ultimately only one value, then options for choice can be tidily arrayed according to how much of the supervalue they bear, promote, or respect. Any evaluative choice would ultimately be a choice between two amounts of the supervalue. Choosing between values would always be like choosing between two lumps of coal or three. Thus, value monism seems to dissolve the threats to practical rationality posed by tragic choices, moral dilemmas, and 'incommensurable' options.

Despite these attractions, most contemporary value theorists assume that value pluralism is true. There are three main arguments for pluralism. First is the intuitive implausibility of value monism: given the apparent diversity of values, how could there be a single value 'common to' all valuable items? Second is the thought that *akrasia* can be explained only if values are plural: how can it make sense for someone to choose something that she believes is worse overall unless there is something attractive about the worse option that is not 'included' in the better one? A closely related third argument maintains that some choice situations involve unavoidable loss: no matter which alternative one chooses, something valuable will be forgone, and thus, there must be plural values at stake in the choice.

2.1 *Ordinary Intuition*

Value monism seems to run afoul of common sense. Two intuitions suggest that any monistic account will

be vulnerable to counterexample. First, how could there be just one value that runs through valuable items as diverse as, for example, achieving philosophical insight and eating a slice of delicious cheesecake? While perhaps the value of both options is instrumental to or constituted by some value (e.g., pleasure), it is hard to believe that bearing, promoting, or respecting that value is ultimately all there is to their being valuable. Second, if there were a single value that exhausted the value of all valuable things, the evaluative difference between things could always be given by some *amount* of the supervalue. But how could the difference in value between achieving philosophical insight and eating delicious cheesecake be a matter of quantity of some one thing? It is incredible to think that evaluative differences among diverse goods are just a matter of more or less of a single value.

To these charges, monists offer two replies. The first is to suggest that the troublesome intuitions rely on an uncharitable view of monism, one according to which what it is to bear the supervalue is to bear or bring about a kind of feeling or experience like pleasure or happiness. It is, indeed, hard to believe that all valuable things involve having or producing a pleasurable or happy experience. But monism need not be so crude. Some monists propose instead that the ultimate value is given by the satisfaction of one's 'fully-informed' or 'rational' desires or preferences. The one property that runs through all valuable items, then, is the property of satisfying constrained desires, and it is the value of *that* that is the supervalue. Evaluative differences between items are a matter of the strength or number of desires they satisfy.

Whether desire-satisfaction accounts of the supervalue are plausible depends on what it is for a desire to be 'fully-informed' or 'rational.' 'Full-information' is usually understood as 'having all the relevant facts and being free from logical error.' 'Rational' desires might be procedurally rational (passing certain evaluatively-neutral tests) or substantively rational (passing certain evaluative tests). Insofar as 'full-information' or 'rational' is a value-neutral constraint, there is no guarantee that such constrained desires or their strengths will track what is intuitively valuable or more valuable: it is perfectly possible that everyone have fully-informed or procedurally rational desires to perform evil and malicious deeds, and that the desires for those deeds be stronger than desires for angelic ones. If, in an effort to secure this tracking, one places evaluative constraints on desires, as one might do by insisting that desires be substantively rational, then one must give up monism. For the values that operate as constraints on desires must be distinct from the value they constrain, and so there are ultimately many values, not one.

Monists have a second, more promising, reply. This reply was introduced by Mill in an effort to defend utilitarianism against counterintuitive consequences of Bentham's quantitative hedonism. There are, Mill insisted, not only different *quantities* of pleasure, but different *qualities* of pleasure. The pleasure of philosophical insight is a 'higher' pleasure than the pleasure of eating delicious cheesecake. Thus, while there is ultimately only one value, there are different qualities of it that explain the seeming diversity of values. Moreover, the evaluative differences between valuable items are not simply given by some *amount* of the supervalue; achieving philosophical insight and eating cheesecake may differ in the *quality* of the supervalue they bear or instantiate—one pleasure is 'higher.'

This reply gets to the heart of the dispute between monists and pluralists. The monist insists that the common-sense belief in multiple values is in fact a belief about different qualities or aspects of a single value. Whether this position is plausible depends on how values are to be individuated. What makes two considerations two distinct values as opposed to two qualities or aspects of a single value? Unfortunately, this question has received almost no philosophical attention. Until the question of how values are to be individuated is settled, ordinary intuitions cannot provide good reasons for thinking value pluralism is true.

2.2 Akrasia

Some philosophers (Wiggins 1978, Nussbaum 1986) have argued that for *akrasia* to be explained as a coherent phenomenon, plural values are required, and since *akrasia* is a coherent phenomenon, it follows that there are plural values. This argument is often presented as establishing the 'incommensurability' of values, but 'incommensurability' is used as synonymous with or as entailing plurality.

If one judges that one has most reason to choose one alternative but instead chooses the other, one is weak of will. Suppose one must choose between going to a party and staying home to work. Although one believes that staying home to work has the greater value—and therefore that one has most reason to stay home—one chooses to go to the party. How can such a choice be explained? If monism is correct, it seems that weakness of will must be an incoherent phenomenon. For suppose that there is a single ultimate value, say, pleasure. It seems, then, that there is nothing about the lesser option that could possibly attract one to it. For everything the lesser option has going for it—a certain amount of pleasure X—is included in what the greater option has going for it—X plus more pleasure to boot. One cannot choose to stay home for a reason, no matter how bad, for there is no reason for choosing the one option that is not already a reason for choosing the other.

The argument, however, does not succeed in establishing value pluralism. For akratic choice can be explained simply by attributing to the akrates a *belief* that there are plural values at stake; there need not

actually *be* plural values. One might mistakenly believe that the value of going to the party is irreducibly distinct from the value of staying home and thus be attracted by the allure of the former. One's reason for going to the party, while based on a false belief, could nevertheless be a reason that makes one's weakness of will coherent.

But perhaps for *akrasia* to be coherent in an 'objective' sense, that is, coherent for agents with, *inter alia*, relevant beliefs that are true, plural values *are* required. This version, however, also fails. One possible difficulty is that what grounds the akrates' attraction to the lesser option may be not the plurality of the values at stake but rather some contingent feature of the circumstances in which the value of the lesser option is instantiated or realized. So, for example, although there might ultimately be only pleasure at stake, the fact that the lesser pleasure of going to the party occurs in a seedy part of town has a special allure for the akrates. There are arguably no plural values here, only circumstances extrinsic to the supervalue whose special appeal to the agent might provide a reason for choosing the lesser option.

A more significant problem is the fact that attraction to the lesser option can be explained by appeal to different qualities or aspects of the supervalue. Take, for example, a choice between two glasses of wine, one laced with pepper and one not (Stocker 1990). Although one judges that the unadulterated wine will provide the greater pleasure, one is charmed by the particular pleasure of the wine laced with pepper and so chooses the lesser pleasure for the reason given by its particular charm. Similarly, one might choose to go to the party while judging that staying home is better for the reason that the party option provides a particular quality of the supervalue which staying at home lacks.

The success of this reply depends on an account of the individuation of values. Is what attracts about the lesser option a value distinct from the value of the greater option or some distinctive quality of an ultimate supervalue?

2.3 Unavoidable Loss

In some choice situations, it seems that no matter which alternative one chooses, something valuable will be lost. If, for example, one must choose between attending a lecture on Kant's ethics and having cheesecake with friends, one is bound to lose out on value no matter which alternative one chooses—and not simply because one cannot have both. If one chooses to attend the lecture, one will forgo the gustatory pleasure of eating cheesecake; if one chooses to go out for cheesecae, one will forgo the philosophical insight one would have gained at the lecture. The fact that any choice in such situations entails value loss shows that the choice involves a conflict between distinct values. Therefore, there are plural values. (Essentially the same argument is sometimes put in terms of the possibility of rational regret over having foregone a lesser good.)

Monism, it seems, must deny that such cases exist. For if monism were true, then choice would always be a matter of choosing more or less of a single value. If one chooses the option with more value, there is no value loss. If one chooses the option with the lesser value, then there is value loss, but it is not unavoidable.

Both the argument from *akrasia* in its 'objective' form and the argument from unavoidable loss rely on the conviction that values are sometimes *incompatible* with one another: pursuit of (respect for, instantiation of, etc.) one excludes pursuit of (etc.) the other. It is the incompatibility of values that grounds the akrates' attraction to the lesser option and that ensures that no matter which value one chooses, there will be value loss. This incompatibility of values might be conceptual, as Isaiah Berlin believed was true of value concepts like *justice* and *mercy*, or an intrinsic feature of values given the circumstances in which they arise. If values are incompatible, it seems that they cannot be reduced to a common value.

The monist reply here is already familiar from the previous two arguments. Sophisticated monists of the Millian variety might insist that there can be incompatibilities among different *qualities* or aspects of a single value. In cases of unavoidable loss, there need not be more than one ultimate value. The languorous pleasure of basking in the sun might be incompatible with the piquant pleasure of hearing unexpected good news: having the one pleasure now rules out having the other now though they are nevertheless instances of a single value, pleasure (cf. Stocker 1990). Similarly, the qualitative value of achieving philosophical insight may be incompatible with the qualitative value of eating delicious cheesecake. Choosing between the lecture and the cheesecake involves a loss in value no matter how one chooses, but there is only one ultimate value.

All three arguments for value pluralism crucially depend on the answer to the question, How are values properly individuated? The fundamental idea behind the common monist response to pluralist arguments is that there is a *nonultimate* sense of 'value' according to which it can be true that there are distinct 'values' and yet those values cannot be values in the robust sense implied by the claim that there are two distinct *ultimate* values.

Exactly how this is to be worked out remains to be seen. Hurka (1996) has suggested that one pleasure is not a different ultimate value from another pleasure if it does not differ in any of its 'intrinsic' features. In explaining what it is for a feature to be intrinsic to a value, Hurka urges that the location of a value is *ex*trinsic to it. Thus, John's pleasure and Joanna's pleasure are two instances of the same value, pleasure. Stocker (1990), however, has suggested that if pleasure

is located in different people, there are different ultimate values: there is the pleasure-for-John and the pleasure-for-Joanna. Stocker (1990, 1997) maintains that if it is rational to care about any difference between two evaluative considerations, it follows that those considerations are different ultimate values. Thus, there is sharp disagreement about how values are properly individuated.

Monism, if it is to be plausible, must hold that the correct account of value individuation leaves room for a tenable notion of a value quality or aspect that serves to undermine intuitions that values are plural, accounts for weakness of will, and explains away choices that seem to involve unavoidable loss. If there is such a notion, then value monism may yet win the day. But any such victory for monism may be hollow. For if monism is to be cluttered with complex relations among value qualities or aspects, the two chief attractions of monism—in short, the tidiness and simplicity it seemed to offer—fall by the wayside. In the end, then, the debate between value pluralists and value monists may be much ado about nothing.

3. *Implications of Reductive Value Pluralism*

Assuming that there are ultimately many values, what, if anything, of interest follows? There are two main implications that are thought to be important.

First, if reductive value pluralism is true, then reductive value monism and any theory that relies on it must be rejected. So, for example, Benthamite utilitarianism has to go. Perhaps more significant is the result that preference utilitarianism, or, more generally, desire-satisfaction theories of value must go. (This is not to say that desire-satisfaction theories of what it is for something *to be* a value must be rejected. One might hold that every value must satisfy some or other desire to be a value while denying that all values reduce to the value of desire satisfaction.) Many economists and some philosophers favor desire-satisfaction accounts of value, and their accounts at the very least require modification if value pluralism is true. Economists who model rational choice on satisfaction of preferences, for example, must relinquish any claim that *goodness* is a matter of satisfaction of one's preferences and instead maintain only that the rationality of choice depends on one's preferences conforming to certain axioms (cf. Broome 1991). Philosophers who think that individual well-being is the only value there is and that what makes one's life go best is satisfaction of one's desires must give up either the claim that well-being is the only value there is or the claim that well-being is simply a matter of desire-satisfaction.

Some have thought that value pluralism shows that any form of utilitarianism is mistaken. But what is essential to utilitarianism is compatible with value pluralism (Sen 1981). Others have thought that value pluralism is the foundation of political liberalism: if values are metaphysically plural, then liberalism is the correct theory of justice. But the metaphysical plurality of values is plausibly neutral between liberalism and nonliberalism. The fact that there are ultimately many values does not entail that a state should not compel its citizens to pursue one value over others, nor does monism entail that a state should not protect its citizens' choices to pursue nonultimate values in ways that do not best promote or respect the supervalue. As some political theorists have pointed out, liberalism itself might be understood as monistic about value: the ultimate value is the value of permitting people to pursue different nonultimate values. Whether there is, in the end, some indirect way in which value pluralism supports liberalism remains an open question.

Value pluralism has been thought to have a second implication. If value pluralism is correct, then difficulties supposedly follow for rational choice. The difficulties alleged are various, but the reason for thinking that none follows from value pluralism is the same. Many have supposed, for example, that alternatives for choice bearing plural values are *incomparable*. How can the option of respecting someone's right to free speech be compared with an option that doubles everyone's pleasure? Insofar as the value of a right to free speech is irreducibly distinct from the value of pleasure, it seems that there is no common basis on which to make a comparison between their bearers. The argument goes, if options for choice bearing plural values are incomparable, rational choice between them is precluded.

There is good reason, however, to think that the plurality of values does not entail the incomparability of options that bear them. Take any two putatively ultimate values such as the value of the right to free speech and the value of pleasure. One can always imagine some option that bears the one value in a notable way that can be compared with another option that bears the other value in a nominal way. So, for instance, an option that involves violating everyone's right to free speech is worse than an option that involves reducing one person's pleasure by a small amount. Given any two putatively irreducibly distinct values, there will always be some comparison between a notable bearer of the one value with a nominal bearer of the other value. The existence of 'nominal-notable' comparisons demonstrates that if alternatives are incomparable, it is not the plurality of values *per se* that entails their incomparability (Chang 1997). Assuming that rational choice depends on the comparability of alternatives, the plurality of values, then, does not preclude rational choice in this way.

Although value pluralism does not entail incomparability, perhaps the issue of value pluralism is important because if value monism is true, complete comparability follows. If there is ultimately only one value, evaluative differences between items must al-

ways reduce to differences in amount or quality of the supervalue, and quantities or qualities of the same thing can always be compared. On some views, if alternatives are comparable, the existence of a rational choice is assured. Thus, monism seems to guarantee that there is a rational choice in every choice situation, no matter how intractable the conflict may seem.

The inference from monism to comparability, however, is mistaken. It is not clear that different qualities of a single value can be compared. How can one compare the luxurious, wallowing pleasure of lying in the sun and the intense, sharp pleasure of quenching a fierce thirst? Making the case for comparability is arguably just as hard under the assumption of monism as it is under the assumption of pluralism. Moreover, even assuming that the supervalue does not admit of different qualities, it is a mistake to assume that all *quantities* of a single supervalue are comparable. Pleasure can be greater in amount without the value of that pleasure being thereby greater. The five-course feast that ends with petit fours involves a larger quantity of pleasure than the five-course feast without, for the petit fours add a quantity of pleasure to the pleasure already had. But the additional quantity of pleasure may make the meal less valuable with respect to pleasure—the petit fours induce pleasure 'overkill.' Since a larger quantity of pleasure is not always better with respect to pleasure, it seems possible in principle that different quantities of a single value are incomparable. Assuming that there are such genuine cases of monistic conflict, the issue of value pluralism cuts across the question of whether bearers of value are comparable or incomparable.

4. 'Nonreductive' Value Pluralism

In general, the question of how many ultimate values there are is irrelevant to the question of how conflicts between them are to be resolved. What matters in conflict resolution is how the values relate to one another in the context of choice, not how they stand to one another metaphysically. As already suggested, a monist might happily allow conflicts between value qualities, while insisting that, as a metaphysical matter, there is ultimately only one value.

Given that reductive value pluralism does not have the implications for practical reason that many value pluralists seem to think it has, it is natural to think that there is some other notion of value pluralism besides the metaphysical one. There is, however, no single specific idea that would account for the concerns of all value pluralists who think the 'plurality of values' has implications for rational choice. At best, the cluster of views can be generally characterized as holding that there is some or other difference among values—something apart from mere metaphysical numerical difference—that has important implications for rational choice. The loose collection of views about the ways in which differences between values can have practical implications might be called 'nonreductive' value pluralism.

Nonreductive value pluralism comes in many different varieties, but there are three leading kinds (for general discussion see Griffin 1997). First is the view that some values are *incompatible*, that they cannot be pursued (instantiated, etc.) together. The incompatibility of values raises the possibility of unavoidable loss of value and, more alarmingly, the specter of moral dilemmas in which the unavoidable loss of value is expressed by a moral requirement.

Second is the view that some values are so different that they—and therefore their bearers—are *incomparable*. Suppose you must choose between a career as a clarinetist and one as a lawyer (Raz 1986). If the values of those careers are incomparable, and if their incomparability precludes rational choice between them, it seems that whichever career one chooses, the choice cannot be rational. Thus, this kind of nonreductive value pluralism has important implications for practical reasons. It suggests that there is perhaps a wide swathe of choices that are beyond the scope of practical reason.

A third view is that some values are so different that their bearers cannot be measured by the same cardinal scale of value (cf. Sunstein 1994, Anderson 1993, 1997). Suppose you must choose between guns and butter. Since there is no cardinal scale along which the value of guns and butter can be measured, the alternatives are *incommensurable*. Many pluralists have thought that rational choice between incommensurable options is impossible. But it is clear that the lack of cardinal comparability does not preclude rational choice. One alternative might be better than another though there is no cardinal scale according to which it is better. Lexically superior options, for instance, cannot be measured on the same cardinal scale as the options they are lexically superior to, but they are nevertheless better. Moreover, the failure of *precise* cardinal comparability need not entail incomparability. Items may be *imprecisely* comparable (Parfit 1984, Griffin 1986). Although incommensurable options do not preclude rational choice, the fact that they are incommensurable entails that any justification for choosing one over another cannot appeal to the idea that there is some amount of one that provides the equivalent *measure* of some amount of the other.

See also: Consequentialism Including Utilitarianism; Economics and Ethics; Relativism, Pluralism, and Skepticism (Philosophical Perspective): Cultural Concerns; Relativism: Cognitive; Utilitarianism: Contemporary Applications

Bibliography

Anderson E 1993 *Value in Ethics and Economics*. Harvard University Press, Cambridge, MA

Anderson E 1997 Practical reason and incommensurate goods. In: Chang R (ed.) *Incommensurability, Incompatability and Practical Reason*. Harvard University Press, Cambridge, MA
Berlin I 1950 *Concepts and Categories*. Oxford University Press, Oxford, UK
Berlin I 1959 *The Crooked Timber of Humanity*. Random House, New York
Broome J 1991 *Weighing Goods: Equality, Uncertainty and Time*. B. Blackwell, Cambridge, MA
Chang R 1997 Introduction. In: Chang R (ed.) *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press, Cambridge, MA
Griffin J 1986 *Well-Being: Its Meaning, Measurement and Moral Importance*. Clarendon Press, Oxford, UK
Griffin J 1997 Incommensurability: What's the problem? In: Chang R (ed.) *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press, Cambridge, MA
Hurka T 1996 Monism, pluralism, and rational regret. *Ethics* **106**: 555–75
Kekes J 1993 *The Morality of Pluralism*. Princeton University Press, Princeton, NJ
Nagel T 1979 *Mortal Questions*. Cambridge University Press, Cambridge, UK
Nussbaum M C 1986 *The Fragility of Goodness*. Cambridge University Press, Cambridge, UK
Parfit D 1984 *Reasons and Persons*. Oxford University Press, Oxford, UK
Raz J 1986 *The Morality of Freedom*. Oxford University Press, Oxford, UK
Richardson HS 1994 *Practical Reasoning About Final Ends*. Cambridge University Press, Cambridge, UK
Sen A 1981 Plural utility. *Proceedings of the Aristotelian Society* **81**: 193–215
Stocker M 1990 *Plural and Conflicting Values*. Oxford University Press, Oxford, UK
Stocker M 1997 Abstract and concrete value: Plurality, conflict and maximization. In: Chang R (ed.) *Incommensurability, Incompatability and Practical Reason*. Harvard University Press, Cambridge, MA
Sunstein C 1994 Incommensurability and valuation in law. *Michigan Law Review* **92**: 779–851
Taylor C 1982 The diversity of goods. In: Sen A, Williams B (eds.) *Utilitarianism and Beyond*. Cambridge University Press, Cambridge, UK
Wiggins D 1978 Weakness of will, commensurability and the objects of deliberation and desire. *Proceedings of the Aristotelian Society* **79**: 251–77

R. E. Chang

Values, Anthropology of

The importance of values in anthropology was recognized by the mid-twentieth century by the designation of two chapters on this topic in Alfred Kroeber's *Anthropology Today: An Encyclopedic Inventory* (1952). Both chapters were written by philosophers, F. S. C. Northrop and David Bidney, rather than practicing anthropologists, which perhaps suggests a sense that the topic should be grounded in a humanistic tradition. Both authors trace philosophical underpinnings while also discussing issues still current such as the debate between cultural relativism and human rights. While the present essay is much briefer than the two noted, they serve as a baseline against which we can note new developments.

Values, in the sense of 'conceptions of the desirable,' (Kluckhohn 1951) definitions of the good, the moral, enter anthropology both as an object of study and as an aspect of the anthropologist's own experience. Treating values as the object of study, anthropologists have taken either a cultural or a sociological approach, either emphasizing the patterning of values in themselves, and the power of values as cultural conceptions which shape experience or, *mutatis mutandis*, emphasizing the interrelationship between values and the sociological context.

Examples of the first approach come largely from cultural anthropology in the American tradition (which has philosophical roots in German idealism). Early examples are based on configurationalism: that is, holistic analysis of values which define broad themes of a particular culture. Benedict (1934), *Patterns of Culture*, is the classic example. She distinguished Appolonian and Dionysian values among plains compared with pueblo Indians. Later, during and following World War II, Benedict, Mead and others applied the same approach to whole nations, depicting dominant values and psychological patterns of Japan, Germany, and Russia; thus was created the national-character school in anthropology. Recognizing the somewhat undisciplined, sweeping holism of these anthropological studies, Kluckhohn, Vogt, and Albert (1966), Kluckhohn and Strodbeck (Kluckhohn 1951, Kluckhohn and Strodbeck 1961) refined the approach, explicitly centering on values, defining dimensions for comparison (for example, past vs. future orientation in relation to time or dominant vs. submissive attitudes toward nature) and made systematic comparisons of particular ethnic or regional groups within America. Clifford Geertz adopted a Parsonsian definition of culture (values, symbols, and ideas) that further refined the older configurationalistic notion, then pursued a Weberian project in his *Religion of Java* (1960) where he distinguished the major divisions within a complex culture according to distinctive values. *Habits of the Heart* (Bellah 1985) applied this approach to America, producing an influential critique of American individualism and calling for a return to earlier values that recognized societal responsibility. As with the early configurationalists, German influences remained important in these later, Weberian-inspired works; *the leitmotif* is that values, as a manifestation of culture, have meaning and significance not reducible to any societal context.

The second approach gives more importance to the societal context, which is viewed as a crucial source

and ground of values. Examples of the second approach come classically from the British social anthropological tradition which had philosophical roots in France especially Emile Durkheim. In *Elementary Forms of Religious Life* (1915), he argued that thought generally, including religious belief and values, derive from society in both its normative and experiential aspects. An excellent application of this thesis is Campbell (1964), *Honour, Family, and Patronage*. Campbell demonstrates cogently how bilateral family patterns explained conceptions of honor expressed by the Sarakatsani of Greece. Honor is an ultimate value, for which the Sarakatsani man will sacrifice his life. Yet honor is grounded in kinship; in this society where seven out of ten name a kinsman as best friend, and a bastard child, lacking kin, is left on the mule path to die, one's honor depends on one's kin. The values derive from and sustain the family. Campbell's analysis is one of the best in showing the intricate functional integration of value, and social organization. Building on this particular analysis Pitt-Rivers (1983), Peristiany (1992), and others also showed how the concept of honor can be seen as a Middle Eastern and southern European value, somewhat transcending particularities of social organization but also grounded in a regional cultural and social pattern they termed 'Mediterranean.'

An effort to reduce the first approach to the second was Douglas' *Cultural Bias* (1978). Douglas distinguishes two dimensions of social life, group and grid, which (though varying in definition) pertain to setting boundaries between in-group and out-group, and defining the inner networks and classifications internal to a group (grid). Utilizing her social variables grid/group, Douglas argues that many cultural values—conceptions of time, of health and disease, of nature—depend on which particular social configuration is dominant. Thus, by knowing the social pattern, one can predict which cultural value will dominate; the societal explains cultural values.

Since the peak of the cultural and social anthropological structural-functionalist approaches to value, anthropology has developed in a myriad of directions. One is pursuing the nature of human rationality (see Overing *Reason and Morality* (1985)) and even returning to questions about human cognition (as with Quinn and Holland (1987)). Further exemplifying the sociological emphasis, but with a Marxist twist, one anthropological approach treats political economy as the driving force of culture. This approach tends to diminish the role of values. Values are treated as false consciousness, as ideological constructions motivated by class interests or a drive to power and profit. Studies of colonialism and postcolonialism illustrate this approach, for example, Comaroff's *Body of Power, Spirit of Resistance* (1985), where religious movements and their accompanying values are contextualized within a colonial and postcolonial framework in South Africa.

Studies of transnationalism and globalism likewise tend to ignore or diminish the place of values, emphasizing the capitalistic and corporate forces that construct global culture. However, human-rights formulations of value loom as one of the crucial forces tempering the global market process that has exploded since the Cold War.

While anthropologists who emphasize political economy moved values from center stage, the topic has migrated to other disciplines, especially those concerned with globalism and international relations. Huntington's *The Clash of Civilizations* (1996), for example, reifies certain value configurations, notably the Islamic and the Western, then foresees the clash between them, replacing the ideological conflict during the Cold War. Scott's *Seeing Like a State* (1998) also focuses on the power of values, notably rationalization, when coupled with the state, to shape human life and thought.

Turning to the place of values in anthropology and for the anthropologist, the basic debate is between engagement and objectivity. While Weber's 'value-free social science' concept converges with positivism to express at least an ideal of objectivity, his '*Verstehen*' converges with a Gadamerian postmodernism to acknowledge the role of values (what Gadamer terms 'prejudice') as projected by the analyst into the process of analysis. These Weberian nuances can still inform the anthropological debate (see for example, the debate between D'Andrade and Scheper-Hughes (1995)). Like Weber, anthropologists agree in acknowledging the role of values in their research and lives but they debate whether the analyst should attempt to control the influence of values, i.e., strive toward objectivity, or should embrace their influence, i.e., become engaged. Engagement means that the analyst should express rather than suppress values, whether those of one's informants, one's own, or others' (e.g., of some ideology or vision of the good). Objectivity means that the analyst should control or suppress expression of values (other than that of scientific or accurate description and analysis).

At an organizational level, two formulations by the American Anthropological Association (AAA), one on professional ethics, the other on human rights, illustrate the issue. The ethics statement boils down to a version of the Hippocratic oath 'Do no harm.' That is, above all else, do not endanger the people one studies. The Human Rights Declaration takes a more engaged and activist stance, calling for the anthropologist to intercede if human rights of a group are violated. In either instance, the anthropologist is engaged, but in the professional ethics statement the frame is set as within the research. Concern is especially with the group among whom one is doing fieldwork. With such a group of individuals, one has a 'covenental' relationship and an ethical commitment not to do harm to them as a result of carrying out the research. However, the ethics statement abstains from

a positive injuction 'Do good' or even from an activist call to 'prevent harm' to the group, in case they should come into harm's way.

As proposed by the AAA Committee on Human Rights for vote by the AAA membership, a more activist commitment is implied. The Declaration reads as follows:

> As a professional organization of anthropologists the AAA has long been and should continue to be concerned whenever human difference is made the basis for a denial of human rights, where 'human' is understood in its full range of cultural, social, linguistic, psychological, and biological senses.

The Committee further comments as follows:

> Support of the Declaration will allow anthropologists to participate more centrally in one of the most significant human-values debates of our time. Suppression of diversity by powerful states or factions with a claim to superior values has led to such horrors as ethnic cleansing and genocide. Yet, 'culture' has been used to justify not taking a stand on actions which human rights NGOs consider gross violations of human freedoms.

The history of the AAA attitude toward human rights intertwines with the history of the anthropology of values. In the mid-twentieth century when the Universal Declaration of Human Rights was being formulated, AAA declined to join the process of formulation on the ground that the ethics of human rights violated cultural relativism by claiming universal values. By the 1990s, AAA was re-engaging the conversation, acknowledging and affirming basic human rights while also emphasizing collective and cultural identities and their concomitant dangers e.g., genocide and ethnic cleansing.

While the complex issues of objectivity and engagement cannot be plumbed further here, this example illustrates part of the spectrum between a model of objectivity (analyst detached from object of research), a model of engagement (analyst obligated to intercede on behalf of endangered groups) and shades in between, illustrated by the 'do no harm,' injunction, which presumes a research objective tempered by an ethical constraint.

In the late twentieth century, efforts toward defining the place of values in the discipline of anthropology, display three major streams. The first affirms positivism and objectivity. The second affirms subjectivity, an interpretive, post-modernist stance leading toward reflexivity—explicit acknowledgement of one's values as part of one's analysis. The third also affirms the place of one's values, but leads not so much inward as outward, minimally toward ethical behavior in relation to those among whom one does research, maximally toward engagement and practice: a mission of reform and change of the world.

This debate about the proper place of values in the experience of the anthropologist is important within the discipline, but insofar as the debate is within the discipline, it is limited in its direct impact on the wider world. Arguably of greater impact is the first topic of this essay: the anthropological study of values as expressed by humankind.

Anthropology should count as one of its major contributions, to scholarship and to the world, the recognition of, study of, and affirmation of the worth of human values, both in their common aspects and in their distinctive emphasis by varied groups or cultures. This is an abiding contribution, worthy of perpetuation into the next epoch of the discipline.

See also: Benedict, Ruth (1887–1948); Cultural Psychology; Culture as Explanation: Cultural Concerns; Globalization and World Culture; Human Rights, Anthropology of; Human Rights in Intercultural Discourse: Cultural Concerns; Values: Psychological Perspectives; Values, Sociology of

Bibliography

Bellah N, Madsen R, Sullivan W, Swidler A 1985 *Habits of the Heart: Individualism and Commitment in American Life*. University of California Press, Berkeley, CA

Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston

Benedict R 1946 *The Chrysanthemum and the Sword: Patterns of Japanese Culture*. Houghton Mifflin, Boston

Bidney D 1952 The concept of value in modern anthropology. In: A. L. Kroeber (ed.) *Anthropology Today*. University of Chicago Press, Chicago, pp. 682–99

Campbell J K 1964 *Honour, Family, and Patronage: A Study of Institutions and Moral Values in a Greek Mountain Community*. Clarendon, Oxford

Comaroff J 1985 *Body of Power, Spirit of Resistance: The Culture and History of South African People*. University of Chicago Press, Chicago

D'Andrade R, Scheper-Hughes N S 1995 Objectivity and militancy: A debate. *Current Anthropology* **36**: 399–420

Declaration of Human Rights 1998 *Anthropology Newsletter* **9**

Douglas M 1978 *Cultural Bias*. Royal Anthropological Institute, London

Durkheim E 1965 *Elementary Forms of Religious Life*. Free Press, New York

Geertz C 1960 *The Religion of Java*. Free Press, Glencoe, IL

Huntington S 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon and Schuster, New York

Kluckhohn C 1951 Values and value-orientation in the theory of action: An exploration in definition and classification. In: T. Parsons (ed.) *Toward a General Theory of Action*. Harvard University Press, Cambridge, MA, pp. 338–433

Kluckhohn F, Strodbeck F 1961 *Variations in Value Orientations*. Row, Peterson, Evanston, IL

Kroeber A L (ed.) 1952 *Anthropology Today*. University of Chicago Press, Chicago

Mead M 1943 *And Keep Your Powder Dry: An Anthropologist Looks at America*. William Morrow, New York

Northrop F S C 1952 Cultural values. In: Kroeber A L (ed.) *Anthropology Today*. University of Chicago, Chicago, pp. 668–81

Overing J 1985 *Reason and Morality*. Tavistock, London

Peristiany G J 1992 *Honor and Grace in Anthropology*. Cambridge University Press, Cambridge, UK
Pitt-Rivers J (ed.) 1983 *Mediterranean Countrymen: Essays in the Social Anthropology of the Mediterranean*. Mouton, Paris
Quinn N, Holland D (eds.) 1987 *Cultural Models in Language and Thought*. Cambridge University Press, Cambridge, UK
Scott J 1998 *Seeing Like a State: How Certain Schemes to Improve the Human Condition have Failed*. Yale University Press, New Haven, CT
Vogt E, Albert E M 1966 *People of Rimwock: A Study of Values in Five Cultures*. Harvard University Press, Cambridge, MA
Weber M 1949 *The Methodology of Social Sciences* [trans. Shils E A, Finch H A]. The Free Press, New York

J. Peacock

Values, Development of

Values are patterns of regulation accepted as desirable by persons in a given culture or family environment and serve as guiding principles in their lives. Research on value development often centers around the concept of internalization which attempts to explain how values assumed to be initially external to an individual, become incorporated into the individual's thoughts and actions. This article describes four perspectives that guide current research on the development of values and three new research directions. Early research on value development was guided by psychoanalytic and social learning theories and did not distinguish among types of values (Grusec 1997).

Social learning studies, for instance, focused on single values such as obedience or altruism without considering connections or differences between types of values. This content-free approach to values implicitly assumed that the development of any values could be understood by very general learning or identification processes. There has been an increasing trend to study classifications of two or more values because it is thought that developmental, regulatory, and socialization processes differ depending on the content of a value. Distinctions have been made between individualistic vs. collectivistic values (Grolnick et al. 1997), the short-term (immediate compliance in parents' presence) vs. long-term nature (enduring compliance in parent's absence) of the parent's socialization goal (Kuczynski 1984), and the personal, conventional, moral, or prudential domain of the value (Smetana 1997). A comprehensive category system includes 10 values: stimulation, self-direction, universalism, benevolence, conformity, tradition, security, power, achievement, and hedonism (Schwartz 1992). These 10 values represent an integrated system that predicts how individual values may be compatible with and conflict with other values in the system. Such research on the content of values suggests that theories of value development will be contextualized increasingly in the future.

1. Four Theoretical Perspectives

Socialization perspectives emphasize the unidirectional transmission of values from an older generation to a younger generation. Older models are based on psychoanalytic (Hoffman 1970) and social learning (Bandura and Walters 1963) theories and generally are limited to an analysis of dyadic parent–child interactions. In general, this approach focused on the products of value socialization to the relative neglect of the process of value socialization. The focus on parents has been on pre-existing parental behaviors, traits, and styles that are conceptualized as antecedents of children's socialization. The focus on children has been on outcomes conceptualized as passive conformity with socialization pressures or internalization of values. Such models have been much criticized by psychologists and sociologists because of their static, homogenious assumptions concerning the nature of society, and the passive and overly conforming view of parents to the previous generation and of children to their parents (Wrong 1961). More recent socialization perspectives (Grusec and Goodnow 1994) retain a focus on the parental role in socialization, but incorporate an interest in how parents accommodate to child characteristics, abilities, and motives in designing flexible strategies for transmitting values to children.

The cognitive developmental perspectives of Piaget and Kohlberg represented early challenges to the socialization perspective. These theories emphasized children's cognitive construction of values and children's sequenced development of their capacity for moral judgement. The parent's authoritative role in teaching values was de-emphasized and a greater focus was placed on the superiority of egalitarian peer relations in promoting moral development. More recent social domain theory (Smetana 1997) balances an interest in the child's activity in constructing values in different environmental contexts with an interest in how parents, among others, provide contextually nuanced, affective, and cognitive material from which children can construct their understanding of value systems in their social environments.

Early socialization theories acknowledged the importance of a positive parent–child relationship context in such constructs as parental warmth and nurturance that were considered to enhance the effectiveness of a parent's socialization techniques. In recent relational perspectives, the long-term parent–child relationship moves to the foreground as a primary source of parental influence. Maccoby and Martin's (1983) review of the research on parent–child interaction presaged many elements of a relational perspective on socialization. In contrast to earlier

perspectives that emphasized parental discipline strategies they traced the origins of socialization to the beginnings of relationship formation during the first year of life. Research supporting the relational origins of children's co-operation include correlations with measures of maternal sensitivity during feeding, secure attachment, and measures of reciprocal parent–child responsiveness during play (Kuczynski and Hildebrandt 1997). The relational perspectives emphasize that far from having to be coerced into accepting parental values, children share a common relationship history with their caretakers that leads them to be receptive to parental influence and to adopt similar social goals.

Bilateral and coconstructionist perspectives are emerging integrative frameworks that incorporate concepts of bidirectional causality, agency of parent and child, and a changing cultural context into a comprehensive understanding of the process of value development (Kuczynski et al. 1997, Lawrence and Valsiner 1993). As with most new theories, these perspectives emphasize that values are actively constructed, or interpreted, by the child from the parental and cultural environment rather than transmitted passively unchanged from one generation to the other. What is distinctive is their explicit interest in the process of bidirectional socialization. For example, the bilateral model of parent–child relations (Kuczynski et al. 1999) focuses on relatively unexplored arenas of parent–child relations such as the constructions and strategic actions of both parents and children, power transactions, and relationship dynamics. These models also add complexity to the parental side of value development. They propose that the values that parents initially bring into the child rearing context also undergo change as the result of self-reflection, exposure to different and changing values outside the family, and the continual opportunities for problem solving and challenges that parents experience from their children during the lengthy process of childrearing.

2. *New Research Directions*

Older socialization perspectives focused on a small set of parental socialization strategies (modeling, punishment, reward, and reasoning). A generalization from this literature was that parents who had a warm relationship with their children, who were firm but not overly restrictive, and who used reasoning and persuasion more than power assertion are the most effective in enabling their children's internalization of values. Critiques of this literature are that it ignores impact of culture, and of a multitude of child characteristics (temperament, age, mood) on parental actions. It also is not consistent with findings that parents use different methods and adopt different goals depending on the varying demands of situations (Grusec and Goodnow 1994). Newer approaches that maintain a focus on socialization strategies of parents include an interest in parental flexible and responsive use of strategies and on an expanded range of socialization strategies. These are strategies that accommodate the agency of the child, attend to the construction, maintenance and repair of the relationship context of parenting interventions, and strategies that prevent or modify the impact of societal influences on children's socialization. There is a growing interest in the impact of unconscious, unplanned processes in the development of values (Bugental and Goodnow 1998). These include cultural practices in which values are learned in the doing of everyday routines and interactions. These practices are difficult to resist because they are taken for granted as 'the way things are' and are rarely open for examination or challenge. A second arena concerns the automatic processing of interactions where, for example, a parent's behavior during parent–child conflict occurs in a script-like, reactive manner, such as in mutually coercive interactions (Patterson 1997). When considered in the context of the conscious processes, a new question emerges: what aspects of value socialization develop from conscious vs. unconscious socialization processes? It is possible, for instance, that some cultural and familial values have nothing to do with specific parenting practices but are transmitted passively from parent to child as a consequence of everyday immersion in cultural practices. Other values may be inculcated by the deliberate activity of socializing agents. Conscious socialization might occur especially when parental values are not unambiguously supported by the surrounding culture or when parents deliberately prepare their children for a social context that differs substantially from the circumstances of their own socialization. Parents may also cycle through 'automatic' vs. conscious interactive processes during daily interchanges with their children (Kuczynski 1984). For example, parents may use unsophisticated power assertive strategies, such as unexplained commands and prohibitions during routine interactions when no more than a loose immediate control of a child's behavior is required. At other times, parents may shift to a more deliberate style of interacting, planning their communications or using effortful reasoning strategies, when the situation has implications for the child's long-term well being or invokes a strongly held parental value.

Most contemporary theories emphasize active, constructive processes of parents and children in value socialization and internalization. However, research is just beginning to describe these processes empirically. Research is needed on parental constructive processes concerning their own values. For instance, how do parents think about their socialization experiences and how do they select which values to pass on, abandon, or modify when they come to socialize their own children? Research is also required on children's

perspectives on their parent's socialization efforts. How do children interpret, resist, and negotiate parental messages? Lastly, the phenomenon of bidirectional influence needs to be better understood. This is because the processes underlying children's receptivity and resistance to parental influence and of parental receptivity and vulnerability to children's influence are still largely unexplored.

See also: Altruism and Prosocial Behavior, Sociology of; Ethics and Values; Kohlberg, Lawrence (1927–87); Moral Development: Cross-cultural Perspectives; Moral Education; Moral Reasoning in Psychology; Piaget, Jean (1896–1980); Piaget's Theory of Human Development and Education; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values; Values: Psychological Perspectives; Values, Sociology of

Bibliography

Bandura A, Walters R H 1963 *Social Learning and Personality Development*. Holt Rinehart and Winston, New York

Bugental D B, Goodnow J J 1998 Socialization processes. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*, 5th edn. Wiley, New York, pp. 389–462

Grolnick W S, Deci E L, Ryan R M 1997 Internalization within the family: A self-determination theory perspective. In: Grusec J E, Kuczynski L (eds.) *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory*. Wiley, New York, pp. 135–61

Grusec J E 1997 A history of research on parenting strategies. In: Grusec J E, Kuczynski L (eds.) *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory*. Wiley, New York, pp. 3–22

Grusec J E, Goodnow J J 1994 Impact of parental discipline methods on the child's internalization of values: A reconceptualization of current points of view. *Developmental Psychology* **30**: 4–19

Hoffman M L 1970 Moral development. In: Mussen P H (ed.) *Carmichael's Manual of Child Psychology*, 3rd edn. Wiley, New York, pp. 261–360

Kuczynski L 1984 Socialization goals and mother-child interaction: Strategies for long-term and short-term compliance. *Developmental Psychology* **20**: 1061–73

Kuczynski L, Harach L, Bernardini S C 1999 Psychology's child meets sociology's child: Agency, influence and power in parent–child relationships. In: Shehan C L (ed.) *Through the Eyes of the Child: Revisioning Children as Active Agents of Family Life*. JAI Press, Stanford, CT, pp. 21–52

Kuczynski L, Hildebrandt N 1997 Models of conformity and resistance in socialization theory. In: Grusec J E, Kuczynski L (eds.) *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory*. Wiley, New York, pp. 227–56

Kuczynski L, Marshall S, Schell K 1997 Value socialization in a bidirectional context. In: Grusec J E, Kuczynski L (eds.) *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory*. Wiley, New York, pp. 23–50

Lawrence J A, Valsiner J 1993 Conceptual roots of internalization: From transmission to transformation. *Human Development* **36**: 150–67

Maccoby E E, Martin J A 1983 Socialization in the context of the family: parent-child interaction. In: Hetherington E M (ed.) *Handbook of Child Psychology: Socialization, Personality and Social Development*. Wiley, New York, pp. 1–101

Schwartz S 1992 Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: Zanna M (ed.) *Advances in Experimental Social Psychology*. Academic Press, Orlando, FL, pp. 1–65

Smetana J 1997 Parenting and the development of social knowledge reconceptualized: A social domain analysis. In: Grusec J E, Kuczynski L (eds.) *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory*. Wiley, New York, pp. 162–92

Wrong D H 1961 The oversocialized conception of man in modern society. *American Sociological Review* **26**: 183–93

L. Kuczynski

Values: Psychological Perspectives

1. Overview

This article asks the following questions: what are values and how are values distinguishable from related concepts like motives, goals and attitudes? Are values located within individuals or social structures?

Values are difficult to study and persistent questions arise as to whether they are 'real,' whether they actually can be shown to have causal influence on behavior. Yet much of everyday life is cast in terms of values—think of ethics, law, religion, politics, art, child rearing, and more. Abstract value judgments are embodied in seeming gut reactions that something is right, moral, or natural vs. wrong, immoral, or unnatural. Another way to 'see' values in action is to contrast cultures or subcultures in what seems right, natural, or moral. One of the great contributions of cultural and cross-cultural research is the way that it brings Western cultural values into sharp relief.

Americans are said to value life, liberty, and the pursuit of happiness. But what does value mean? Implicitly or explicitly we evaluate or assign value to everything—regarding things as good or bad, a truth or falsity, a virtue or a vice. How do we know? One important means is through values. Values can be thought of as priorities, internal compasses or springboards for action—moral imperatives. In this way, values or mores are implicit or explicit guides for action, general scripts framing what is sought after and what is to be avoided.

2. Definitions

Modern theories of values are grounded in the work of Kohn (class and values), Rokeach (general value systems), and Kluckhohn (group level). Values can be

conceptualized on the individual and group level. At the individual level, values are internalized social representations or moral beliefs that people appeal to as the ultimate rationale for their actions. Though individuals in a society are likely to differ in the relative importance assigned to a particular value; values are an internalization of sociocultural goals that provide a means of self-regulation of impulses that would otherwise bring individuals in conflict with the needs of the groups and structures within which they live. Thus, discussion of values is intimately tied with social life. At the group level, values are scripts or cultural ideals held in common by members of a group; the group's 'social mind.' Differences in these cultural ideals, especially those with a moral component, determine and distinguish different social systems. In this sense Weber's Protestant 'ethic' and 'spirit' of capitalism describe value systems.

Values, to which individuals feel they owe an allegiance as members of a particular group or society, are seen as the glue that makes social life possible within groups. Yet, they also set the stage for frictions and lack of consensual harmony in intergroup interactions. Values are thus at the heart of the human enterprise; embedded in social systems, they are what makes social order both possible and resistant to change. Values are not simply individual traits; they are social agreements about what is right, good, to be cherished.

What is common to all value phenomena? At the individual level, values contain cognitive and affective elements and have a selective or directional quality; they are internalized. Preference, judgment, and action are commonly explained in terms of values. Individuals take on values as part of socialization into a family, group and society. Once taken on, values are assumed relatively fixed over time (see *Values, Development of*). Indeed, values that are individually endorsed and highly accessible to the individual do predict that individual's behavior. Conversely, even personally endorsed values won't influence action when they are not made salient to the individual at the time of action. Moreover, in any given situation more than one personally endorsed value may apply, and the behavioral choice appropriate for one value may conflict with the behavioral choice appropriate to another value.

Values are codes or general principles guiding action, they are not the actions themselves nor are they specific checklists of what to do and when to do it. Thus, two societies can both value achievement but differ tremendously in their norms as to what to achieve, how to achieve, and when pursuing achievement is appropriate. Values underlie the sanctions for some behavioral choices and the rewards for others. A value system presents what is expected and hoped for, what is required and what is forbidden. It is not a report of actual behavior but a system of criteria by which behavior is judged and sanctions applied. Values scaffold likes and dislikes, what feels pleasant and unpleasant, and what is deemed a success or failure. Values and value systems are often evoked as rationales for action; for example, values of freedom and equality were evoked to elicit American support for the Civil Rights movements. Values differ from goals in that values provide a general rationale for more specific goals and motivate attainment of goals through particular methods.

3. *History and Current Developments*

Initially viewed with suspicion by Western social scientists as too subjective for scientific study, the concept of values found increasing use beginning with *The Polish Peasant in Europe and America* (Thomas and Znaniecki 1921). Impetus for the study of cultural values comes from the work of Alfred Kroeber, Clyde Kluckhohn, Talcott Parsons, Charles Morris, Robert Redfield, Ralph Linton, Raymond Firth, A. I. Hallowell, and more currently Milton Rokeach and Shalom Schwartz.

Kluckhohn and Strodtbeck (1961) proposed that cultural value systems are variations of a set of basic value orientations that flow from answers to basic questions about being: (a) What is human nature—evil, neutral, mixed, or good? (b) How do we relate to nature or supernatural—subjugation, harmony, or mastery? (c) What is the nature of time—past, present, future? (d) What is the nature of human activity—being, being-in-becoming, doing? (e) What is the nature of our relationship to others—are we joined vertically, horizontally or are we simply separate individuals? They also organized a system for comparing values in terms of their level of generalization and function in discourse and conduct, proposing that values fit into a pyramid of ascending generalization. For each society, a few central or focal values were proposed to constitute a mutually interdependent set of what makes for the 'good life.' These include the unquestioned, self-justifying premises of the value system and definitions of basic and general value terms; for example, happiness, virtue, beauty, and morality.

Since American researchers dominated values research, much early work focused on documenting American values. Need for achievement as an American value, and concern over decline in the centrality of this orientation appear as early as 1944 (Spates 1983). Values studies documented the influence of education, age, type of employment, and socioeconomic status on value preferences of Americans, adding to Weber's thesis of the influence of religion (Protestantism vs. Catholicism) on achievement and work values in Europe. Kohn (1977) was responsible for a number of important values surveys documenting that in various European countries and the US, parents of higher socioeconomic status value self-direction in their

children more than parents of lower educational and occupational levels. These findings have been verified cross-nationally in 122 societies.

Extending the documentation of American values, Rokeach (1973) validated empirically 36 values related to preferred end states and preferred ways of behaving. Using Rokeach's scale, value differences tied to class, age, race, subculture, and level of differences were documented in many countries. Building on Rokeach, Schwartz (1992) delineates values as ways of articulating universal requirements of human existence—to survive physically, have social interchange, and provide group continuity. For Schwartz, values represent operationalizations of these needs as goals that fit together in meaningful clusters (achievement, self-direction, stimulation, hedonism, universalism, benevolence, tradition, conformity, security, and power). Some clusters are compatible (e.g., stimulation and hedonism) and others compete (e.g., self-direction and conformity).

Using mostly data from teachers and college students in 20 primarily Western countries, Schwartz shows that, with the exception of China, specific values mostly do 'cluster' and 'compete' as expected. Thus, 'honest,' 'forgiving,' and 'helpful' cluster together as 'benevolence,' and 'self-direction,' and 'stimulation' cluster far from 'conformity,' 'tradition,' and 'security.' These data suggest important universality to how values are organized cross-culturally and that societies differ in which clusters of values predominate public life.

4. *Controversies*

Key tensions in the values literature focus on the conditions under which they may influence behavior, and the appropriate level of analyses for seeing values in action. Interest in values as a research focus has ebbed in the past as each paradigm for studying values has been criticized for lack of specificity of findings as due to values and not other social norms, attitudes or situational constraints. Current cultural psychology focuses attention on social structures as the repository of values such as personal freedom, group harmony, personal happiness, and duty or filial piety.

How do we know that values exist? A number of options are available: (a) Individual testimony—people say what values they hold.Yet, self-reports of values are subject to pronounced context effects (see *Attitudes and Behavior*). (b) Behavioral choices—either in naturalistic or laboratory settings, value differences may be imputed from behavior. Yet, behavior is influenced by many variables other than values. At the individual level, values themselves are assumed to link to behaviors via their influence on norms and attitudes, but people may infer their values from their behavior, reversing the causal relationship. (c) Cultural and social structures—expenditure of resources, time, energy and structuring of the natural environment; cultural products can be seen as concrete residues of value-based choices (see *Cross-cultural Psychology*; *Cultural Psychology*). (d) Social interchange—observation of behavior in situations of conflict, and more generally observation of what is rewarded or punished, praised or vilified provides data for identifying what is socially valued. Here, too, the question of appropriate evidence arises. To what extent is it appropriate to assume that differences in social structures and societies are evidence of value differences? Political and economic influences and simple inertia may set the stage for behaviors, without a causal influence of necessarily values.

5. *Future Directions*

Cross-cultural perspectives are currently becoming increasingly central to values discussion. For example Inglehart (1990) documented values and value change in a large multinational study, and a large number of two-nation comparison studies has emerged. Another important topic of research is the connection between values of individuals, values of subcultural groups, and values of larger cultural systems and methods for identifying and studying each of these. Perhaps in addition to identifying value vocabularies at each level, it is time to begin to ask whether values appropriately are studied as fixed traits of individuals or as embodied in groups, and to what extent values research is synonymous with cultural and cross-cultural research.

Given that any particular behavior importantly is influenced by context effects that make certain information salient at the moment of action, it is not surprising that the effects of individual value endorsements on behavior have a 'sometimes you see it, sometimes you don't' quality about them. But focusing on individual endorsement of values may miss much of the power of value systems to influence everyday life. That is, individuals may not need to personally endorse or have salient particular values in order for their influence to be felt. The most profound influence of values may be through the ways that they influence rules, norms, procedures within a society, and in this way structure the everyday life choices for individuals within a society.

Thus, whereas previous researchers have documented values using survey techniques in which individuals rated the extent to which various values were important to them, future assessment of values may need to consider more indirect approaches such as what services a society provides its members, what behaviors are rewarded or sanctioned and so on.

See also: School Outcomes: Cognitive Function, Achievements, Social Skills, and Values; Values, Anthropology of; Values, Development of; Values, Sociology of; Vocational Interests, Values, and Preferences, Psychology of

Bibliography

Darley J, Shultz T 1990 Moral rules: Their content and acquisition. *Annual Review of Psychology* **41**: 525–56

Inglehart R 1990 *Culture Shift in Advanced Industrial Society*. Princeton University Press, Princeton, NJ

Kluckhohn F, Strodtbeck F 1961 *Variations in Value Orientations*. Row, Peterson, Evanston, IL

Kohn M 1977 *Class and Conformity*. University of Chicago Press, Chicago

Rokeach M 1973 *The Nature of Human Values*. Free Press, New York

Schooler C 1996 Cultural and social-structural explanations of cross-national psychological differences. *Annual Review of Sociology* **22**: 323–49

Schwartz S 1992 Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology* **25**: 1–66

Spates J 1983 The sociology of values. *Annual Review of Sociology* **9**: 27–49

Sullivan J, Transue J 1999 The psychological underpinnings of democracy: A selective review of research on political tolerance, interpersonal trust, and social capital. *Annual Review of Psychology* **50**: 625–50

Thomas W, Znaniecki K 1921 *The Polish Peasant in Europe and America*. Chicago University Press, Chicago

D. Oyserman

Values, Sociology of

Everything social actors appreciate, appraise, wish to obtain, recommend, set up or propose as an ideal, can be considered as a value. Ideas, emotions, moral deeds, acts, attitudes, institutions, material things, etc. may possess this special quality by virtue of which they are appraised, desired, or recommended. But what is attractive for some, can be repulsive to others. Thus, to values correspond countervalues which are underrated, disapproved, rejected. Nationalism and internationalism, private and public property, freedom and equality, etc., may be, according to diverse actors, values or countervalues.

1. *Dimensions of the Concept*

Four main dimensions of the concept can be distinguished:

(a) Each value has an *object*, i.e. what is valued, prized. The nation, Moslem faith, work, profit, instruction, leisure, honesty, the family, etc., may become values. Any element of social reality, of the spiritual and moral world, can have a 'value aspect' insofar it is praised or refused, advocated, or condemned.

(b) This object is qualified by a *judgment* as valuable or contemptible, as good or bad, as useful or useless, as true or false, as desirable or not, as beautiful or ugly, etc. The sentence expressed is a value judgment. One will say, e.g., that one's country is inviolable and its enemies are unkind, that Moslem faith is true and unbelievers are mistaken, that work is sacred and profits are unjust, that honesty is a virtue while robbery is dishonest, and so on. Value judgments answer to a large set of principles and criteria whereby opinions, beliefs, convictions are shaped, choices are made.

(c) Values become *norms* when they command and/or regulate conducts, prescribe a course of action. Norms tend to conform behavior and commitments to the values confessed. If your country is inviolable, you should defend it, if Islam is true, you must comply with its prescriptions, if profits are unjust, you must fight against them, if instruction is important, you must learn, if honesty is a virtue, you are not allowed to misappropriate funds. Values provide the grounds for accepting or rejecting particular norms, and norms are standards for actual conduct.

(d) The *value holders* are either individual or collective actors or social groups. Therefore one can speak of the values of such and such person, of the liberals, of the middle class, of the teenagers, of the Russians, or of Bantu culture.

The concept of value is inseparable from the notion of preference. To value one object rather than another (e.g. to prefer a party of cards to the theater) means that in a given situation the value inducing the choice was adopted or inculcated to the detriment of another.

2. *Value Systems: Definition*

The values of an individual or a collectivity do not appear as sharply separated and independent units. Instead, they are bound together, are interdependent, they form a *system*. When a new value is acquired or an old one is lost, when a value is weakening (lowering) or strengthening (rising), the whole system will be affected.

A system of values is hierarchically built up. It is also a scale of values. Often the difference between actors does not proceed from the content of their systems, but from their difference in ranking their values. For example, in the abortion debate, all participants may highly prize the value of life, but some will emphasize the conceived child's future while others will take into account the mother's decision.

The actor is more or less tied to certain values than to others. Values contain not only cognitive elements, they involve strong affective components too. The more a value is deeply rooted, the more it takes a central place in the system and the more it is lived intensely, arouses emotions, and mobilizes vehement energies. There are values men are ready to die for.

The mode of organizing a system of values varies from one culture to another. Its inner logic does not obey the same rules. This fact is undoubtedly the main reason why misunderstanding prevails between peoples pertaining to different cultures, each one interpreting the world in its own terms.

3. Harmony and Contradictions

At a first glance, the word 'system' suggests that the values upheld by the actors have their inner rationality and that their logic can be discovered. But a closer examination reveals that values are not necessarily clear and cannot be easily stated. A system is not perforce transparent. Often it comes to light only by action.

Moreover, values of the same actor can be in contradiction, i.e. push towards incompatible goals. The tension may remain covert if the actor is unconscious of this inconsistency, but overt contradiction in crucial matters may produce trouble of conscience, crisis, even anomy.

Another type of contradiction occurs when behavior and acts belie values. Criminals are likely to subscribe to principles and priorities similar to those of the man of the street. They only transgress a few of them by their deviant conduct. Ordinary people also regularly experience noticeable discrepancies between their fine words and their deeds.

Consequently, a set of values is a living system, very complex, open to seesaw motion and variations.

4. How Values are Acquired and Maintained

Values are shaped and built in, receive their meaning, and are structured primordially by the process of socialization whose main agents are the family, the school, the media, the peer-group, etc. Men progressively learn the rules of their society, their knowledge, their know-how, are educated, i.e. appropriate their 'cultural equipment.' Everybody confronts his/her life with the values which are accessible to him/her. Persons born in the same social milieu or at the same time (thus pertaining to the same generation) and having been socialized similarly are likely to embark upon their human destiny sharing the same values. Those who share the same values are likely to assemble and to act together.

If the system of values proves to be an efficient guide, a good map for the travel through life, there is no manifest reason to modify it. It is consolidated by experience. It is in congruity with the conditions of real existence.

Social approval or disapproval constitutes another way of reinforcing values. Opinions and behavior are socially controlled: conformity with the predominant values is rewarded, deviance is more or less severely penalized.

5. How Values Change

Nevertheless, a value system is a 'living concern' and it is regularly challenged by an evolving social environment. It responds by a slow, adaptive change, its main features persisting, but sometimes it undergoes more dramatic change, a complete upheaval.

Change has several meanings: some values are being upgraded or downgraded in the actor's value system; some values enter into the system, others exit; some gain in intensity, others weaken; values elicit greater or less commitment from their holders; some expand, the number of their adherents grows, others lose their attractiveness; the scope of validity of a value can vary (e.g. fraternity may be limited to the members of the family, but it may be extended to the nation or to mankind); etc.

Value changes can be induced by a large range of causal factors. It would be impossible to draw up a complete list. Some examples: value change may result from economic crisis or from prosperity, from demographic collapse or from a population explosion, from ecological alterations (such as pollution). Learning, knowledge, increased information are propitious to questioning values. External cultural influences may give rise to imitation or suggest that some institutions, customs, and ways of thinking are inadequate. Technological inventions play an important role: television has successfully replaced or diminished other leisure activities in the preference scale of citizens. If an organization is at odds with the surrounding world, it is challenged to change its mind, to introduce reforms or perish. The past and present of all political regimes or 2000 years of Christian history tend to show this.

Scarcity entails the rise of a value (in an anonymous world people are likely to cherish the value of conviviality). A durable conditioning by either advertisement or propaganda may provoke value shifts. Sometimes dormant values are awoken: generally solidarity is not very highly ranked in opinion polls, but some disaster may suddenly 'warm up' the value and mobilize people to help the victims.

In all these cases external pressure or influence initiate the process of value change. If change begins and continues in the same orientation, it can speed up and/or become cumulative. But the movement between values and the social environment is not unidirectional, as in the examples mentioned where values are dependent variables. Values are also independent variables. If, for example, technological inventions influence values through modifying ways of living, it is also true that valuing inventiveness, novelty, creativity, scientific research, will powerfully stimulate technological progress. The relation between values and economic, social, political, or cultural factors is governed by a dialectical swinging. Change is the product of the combined operation of the external world and the mental and psychic universe.

To explain dramatic value changes, reference must be made to a theory of social problems. In this perspective, as long as the itinerary of a society or a group, or the trajectory of an individual are unfolding without a major accident, there is no need for a change of course. But if suddenly a rupture happens (e.g. the

death of a beloved person, the loss of a job, a divorce … for the individual; deep dysfunctioning, a foreign attack, a revolutionary situation … for the society)—imbalance, disorganization, deadlock, frustration, alienation come about. The overall guiding capacity of the value system is cruelly put to the test.

It becomes rapidly apparent if the existing values are adapted or not to define the problem and suggest solutions. If this operation is successful, the value system of the actor will be consolidated. On the contrary, it may occur that the 'cultural equipment' of the actor is insufficient to recognize and evaluate the ins and outs of the situation. His/her perceptions, ways of thinking, and values which inform and command his/her feelings and reasoning lose grip on the reality. He/she seems to be culturally blind. He/she risks to succumb to a crisis or fall into anomy. At the macrolevel, whole civilizations may witness decline and final collapse.

Another eventuality is particularly interesting from the point of view of value change. The actor has perceived and formulated the problem, but he/she lacks means for tackling it. His/her value system is then submitted to a partial decomposition: the elements which do not respond to the requirements of the situation (i.e. to understand what is at stake and conceive a project) are dismantled, abandoned. For example, a society desirous to undertake economic development is forced to eliminate such values as conformity, respect for inherited hierarchies, or the acceptance of fatality.

A stage of recomposition is undertaken when the actor is trying to integrate new values able to suggest objectives and means for a suitable action. In this search, the total number of appropriate values is not infinite. Only those will be retained which offer a satisfactory solution to the original problem, correspond to the intimate aspirations of the actor, and can be bound up with other existing values. In the case of a development oriented society, entrepreneurship, the notion of exact time, or middle-term prevision will be values to be promoted.

The process described above is by no means a mechanical one, following a logical step-by-step operation or a fixed chronological order. It takes place in the very heart of the conscience, in the midst of impressions, reflexions, and hesitations; the values, as well as the project, which finally will prevail, will emerge from a more or less long incubation period. The path followed will of course differ in the case of an individual actor as for a collective one.

6. *Production of Values*

Values are not impersonal, everlasting principles, existing since the first mists of time, floating in the air. They are, in the last instance, shaped by historically situated outstanding figures or institutions. These latter can be called 'value producers,' 'standard setters,' 'sense makers.'

Who are they? Mostly, great moral personalities, prophets, philosophers, ideologists, intellectuals, scientists, artists, novelists, film directors, and institutions such as Churches, clubs, learned societies, research centers, universities, etc.

Their tasks consist of answering the society's existential questions, in giving a meaning to events, in elaborating systems of thought, in proposing new horizons. They supply society and the actors with ideas, objectives, programs, ethical principles, critical judgments, etc.

The expansion of values is assumed by 'value transmitters,' 'opinion leaders,' 'popularizers,' 'teachers,' etc. and takes the usual channels of cultural communication such as newspapers, radio, television and other media, books, music, the school system, etc.

Public opinion generally lags behind with regard to the value producers and the opinion leaders. The speed and the scale of the diffusion of a value depends on a complex interplay between the needs, the receptivity, and the resistance of the public and the powerholders, as well as the opportunities for innovation.

7. *Classifications*

Many typologies of values have been put forward. Here are some of the most significant and useful:

(a) The *core values* are those endorsed by the whole population. The *specific values* are shared by some segment of people: social class, age group, a region, an ethnic minority, etc. Social consensus is built upon the core values. They make sure that the members of a society can live together, understand each other, communicate and enjoy a minimal cohesion. Core values correspond to what are called the 'ethos of a civilization' or the 'dominant values' of a people.

(b) The *structuring values* are those in the center of a value system, ordering it, commanding the arrangement of the whole. They are the most important for the actor, explain his/her choices. They resist change while the *peripheral values* are more open to external influences. For some individuals, their structuring value is family, for others love or professional success or religion or soccer, etc., as well as any combination of two or three values.

(c) A similar distinction opposes *final* and *instrumental values*. The final values are at the top of the hierarchy in the system while the instrumental values are necessary to achieve them (like hard work as a means for professional success).

(d) *Global values* transcend the different spheres of social life, their area of validity embraces all human relations. For example, justice may be a requirement for the allotment of an inheritance, the fixing of a price, a legal verdict or the international relations.

Moral values are naturally global. *Sectorial values* are confined to a special sphere of society. Political, economic, religious etc. values like parliamentary democracy, competition, or piety, are sectorial.

(e) Totalitarian societies impose *unanimity* in the field of values. Many traditional communities tend to be unanimous. If constraint loosens, society becomes progressively *pluralistic*. The coexistence of divergent value systems calls for tolerance. But they can clash too. Value conflicts are sharper than conflicts of interests because values arouse emotions and passions. When values like nationalism or religion (or both) motivate the contenders and define their very identity, the struggle threatens with destruction and death.

8. *Place, Role, and Function of Values in Sociological Analysis*

Values are present practically in all social process and historical transformation. It would be impossible to list all situations and evolutions where they play an active role. To mention some of them must suffice.

Values assume the function of identity building. The individual aware of his/her values, can consider him/herself as a person, find his/her place in the world, feel confident, interpret and evaluate his/her environment. The group sees in its values one of the major reasons of its members' adhesion, the common denominator of the participants: to share the same values is cementing their union.

Values are constituents of the system of action of individual and collective actors. They motivate them, influence their needs and aspirations (for one strives to obtain what one is valuing), order their perceptions (for one looks on the world through the glasses of his/her values), provide criteria for their judgments on the society, their situation, on others, on themselves.

Values hold an intermediary position between the reality and the actor's will to preserve or to modify it. Indeed, values suggest projects. They command the selection of objectives: one goal is preferentially chosen because it is more attractive, more desirable, more appreciated. They help to determine the means: which one is better, cheaper, more appropriate, more efficient, etc.

Values supply the 'raw materials' of ideologies. Ideologies rationalize the values, the aspirations, and the interests of the actors. Values also intervene in any speech whose aim is to persuade or promote nourishing a more or less occult argumentation. They serve to justify unavowable initiatives, attitudes, or feelings: an aggression, e.g., will not be presented as such, but legitimated by values such as national security or the liberation of a neighboring country.

Values generate attitudes and steer behavior. The meaning of the patterns of behavior is not understandable without referring to the values of a given culture.

Values are often a mobilizing force for social change. Large-scale shifts like revolutions or reforms are generally preceded by transformations in the value systems and guided by new values. They are embodied by new symbols: flag, colors, emblems, statues, revered places or monuments, songs, etc.

Values contribute to maintain and regulate society, being the fundamentals of legitimacy, law, and social control. (Following the scientific ideal, it is claimed that social sciences must be value-free, which means that ideologies, doctrines, and beliefs can never impede, interfere with, or distort the impartiality and desirable objectivity of scientific method.)

9. *Study of Values: Methodological Devices*

Talleyrand said: 'God gave us the faculty to speak in order to conceal what we think.' This is the reason why it is so difficult to study values. Furthermore, even the most honest persons meet with difficulties in accounting for their values. A lot of these remain implicit, taken for granted. Another problem: how to identify values? They are embedded in verbal, behavioral, and situational contexts. We often have to infer them constructing indicators and being attentive to the signs which reveal them. This careful study requires a critical apparatus and a well-constructed method.

A large array of research instruments has been developed to recognize and index values. Allport, Vernon, and Lindzey used a values test. M. Rokeach listed 18 terminal values ('preferred end-states of existence') and 18 instrumental values ('preferred modes of behavior'). Respondents ranked the items by importance as guiding principles for daily life.

Interviews with standardized questionnaires are another path to follow. If they are conceived in an international comparative perspective, the selection of the vocabulary requires a preliminary contextual analysis. In case of free, nondirective interviews or autobiographies, the respondents choose their own words and produce a richer material, but less easy to generalize.

The researcher has at his/her disposal a great variety of sources: newspapers, magazines, advertisements, school textbooks, novels, doctrinal works, speeches, songs, pictures (photos, films), symbols, etc. He/she needs adapted content analysis or different kinds of semiological approaches to extract the values held by the authors and actors.

Values can also be observed in their natural setting focusing on men's activities and behavior which embody their motivations, desires, aspirations. In certain circumstances experimental designs and techniques are advisable.

The approach to adopt depends on several considerations: the nature of the values to be discovered, the seeking of rapidity of results, extensivity or

intensivity of the knowledge targeted, etc. A combination of procedures is always recommended, for none of the instruments covers the entire field alone.

10. Past and Recent Research

For philosophers and moralists, 'value' has been a familiar notion for a long time. For economists, from the outset, it was a central concept. The founding fathers of sociology approached the study of cultural phenomena without necessarily using the word 'value.' Tocqueville, for example, speaks of the 'spirit' of the Americans and attributes to religiosity the main characteristic of their society. For Karl Marx, ideas, feelings, ethics, law are superstructural facts determined by the infrastructure. According to Max Weber protestant ethics (values) play a central role in the birth of capitalism. With *The Polish Peasant in Europe and America* (1918–1921), W. Thomas and F. Znaniecki make of values an important factor.

Finally, the concept becomes the object of a book with C. Bouglé, one of Durkheim's followers (*Leçons de sociologie sur l'évolution des valeurs* 1922). Anthropologists were particularly well qualified at focusing their interests on field studies of cultural phenomena in small communities. Following their dominant values, Ruth Benedict divided Indian communities in the South-West of the US into Apollonians and Dionysians (*Patterns of Culture* 1934).

A new and decisive trend opens with the Cornell Values Study Group under the impulse of Clyde Kluckhohn (1949). It initiated a long-term comparative project which led to the theory of value orientations of Florence Kluckhohn and Fred Strodtbeck. These investigations owe much to Talcott Parsons' functionalism which places the actor and value guided action at the heart of the sociologists' interest. Seymour Lipset was inspired by Parsonian categories in comparing American values with those of other Anglo-Saxon societies (1963) and those of Latin America (1967).

In the meantime, the theory of development founded by economists was extended to the cultural field. Everett Hagen (1964) and others emphasized the decisive contribution of values to the transition towards economic growth.

At the end of the 1960s and the beginning of the 1970s, a deep cultural turmoil surprised the Western world. Old traditional values were refused and new ones adopted and diffused by osmosis. These profound and fast changes stimulated sociological research on values. The new era was baptized, following the authors, individualistic, narcissistic (Christopher Lasch), hedonistic, secularized, postindustrialist (Daniel Bell), postmaterialist, etc.

The postmaterialist hypothesis of Roland Inglehart (from 1971 on) opened a broad current of empirical investigations. Another pioneering effort was due to Milton Rokeach (1973).

In 1981 a first international comparative survey was launched by the *European Value Systems Study Group* headed by Jan Kerkhofs and Ruud de Moor, and covering 10 countries (the same questionnaire being used). Jean Stoetzel synthetized the results in his *Les valeurs du temps présent: une enquête européenne* (1983). A new wave of surveys took place in 1990 with all European countries as well as the US and Canada participating. The enterprise has become the largest research program ever carried out. It was decided to replicate the study at 10-year intervals. The 1999 survey encompassed 45 countries. A rich academic literature abounds around the investigations (between 1990 and 1997 alone, 194 recorded titles). A similar work was done by the European Union's *Eurobarometer*, publishing recurrent data on values too. More recent investigations have penetrated specific fields such as political values (Rezsohazy 1996).

See also: Benedict, Ruth (1887–1948); Ethics and Values; Marxist Social Thought, History of; Moral Economy and Economy of Affection; Moral Education; Moral Sentiments in Society; Norms; Religion: Morality and Social Control; Social Influence, Psychology of; Values, Development of; Weberian Social Thought, History of

Bibliography

Abrams M, Gerard D, Timms N 1985 *Values and Social Change in Britain. Studies in the Contemporary Values of Modern Society*. Macmillan, London

Abramson P R, Inglehart R 1995 *Value Change in Global Perspective*. University of Michigan Press, Ann Arbor, MI

Albert E M, Kluckhohn C 1959 *A Selected Bibliography on Values, Ethics, and Esthetics in the Behavioral Sciences and Philosophy, 1920–1958*. The Free Press, Glencoe, IL

Allport G W, Vernon P E, Lindzey G 1960 *A Study of Values*, 3rd edn. Houghton Mifflin, Boston

Arrow K J 1951 *Social Choice and Individual Values*. Wiley, New York

Ashford S, Timms N 1992 *What Europe Thinks. A Study of Western European Values*. Dartmouth, Aldershot, UK

Baier K, Rescher N (eds.) 1969 *Values and the Future. The Impact of Technical Change on American Values*. The Free Press, New York

Barrett D N (ed.) 1961 *Values in America*. The University of Notre Dame Press, Notre Dame, IN

Bell D 1973 *The Coming of Post-industrial Society. A Venture in Social Forecasting*. Basic Books, New York

Benedict R 1934 *Patterns of Culture*. Houghton Mifflin, Boston

Boudon R 1999 *Le sens des valeurs*. Les Presses Universitaires de France, Paris

Bouglé C 1922 *Leçons de sociologie sur l'évolution des valeurs*. Armand Colin, Paris

de Moor R (ed.) 1995 *Values in Western Societies*. Tilburg University Press, Tilburg

Ester P, Halman L, de Moor R 1993 *The Individualizing Society. Values Change in Europe and North America*. Tilburg University Press, Tilburg

Gross F 1985 *Ideologies, Goals, and Values*. Greenwood Press, London

Hagen E E 1964 *On the Theory of Social Change*. Tavistock, London
Hall E T 1959 *The Silent Language*. Doubleday, Garden City, NY
Harding S, Phillips D, Fogarty M 1986 *Contrasting Values in Western Europe. Unity, Diversity, and Change*. Macmillan, London
Inglehart R 1977 *The Silent Revolution. Changing Values and Political Styles Among Western Publics*. Princeton University Press, Princeton, NJ
Kluckhohn C 1951 Values and value orientations in the theory of action: An exploration in definition and classification. In: Parsons T, Shils E (eds.) *Toward a General Theory of Action*. Harvard University Press, Cambridge, MA
Kluckhohn F R, Strodtbeck F L 1961 *Variations in Value Orientations*. Row, Peterson, Evanston, IL
Lasch C 1979 *The Culture of Narcissism*. W. W. Norton, New York
Lipset S M 1963 *The First New Nation*. Basic Books, New York
Lipset S M 1967 Values, education and entrepreneurship. In: Lipset S M, Solari A (eds.) *Elites in Latin America*. Oxford University Press, New York
Maslow A H (ed.) 1959 *New Knowledge in Human Values*. Harper, New York
Maziarz E A 1979 *Value and Values in Evolution*. Gordon and Breach, New York, London
Morris C W 1956 *Varieties of Human Value*. University of Chicago Press, Chicago
Mukerjee R 1949 *The Social Structure of Values*. Macmillan, London
Pepper S C 1958 *The Sources of Value*. University of California Press, Berkeley
Perry R B 1954 *Realms of Value: A Critique of Human Civilization*. Harvard University Press, Cambridge, MA
Reich B, Adcock C 1976 *Values, Attitudes, and Behaviour Change*. Methuen, London
Rezsohazy R 1996 *Pour comprendre l'action et le changement politiques*. Duculot, Louvain-la Neuve, Belgium
Riffault H (ed.) 1994 *Les valeurs des Français*. Les Presses Universitaires de France, Paris
Rokeach M 1973 *The Nature of Human Values*. The Free Press, New York
Stoetzel J 1983 *Les valeurs du temps présent: une enquête européenne*. Les Presses Universitaires de France, Paris
Thomas W I, Znaniecki F W 1958 *The Polish Peasant in Europe and America (1918–1921)*. Dover Publications, New York
Wilbur J B (ed.) 1979 *The Life Sciences and Human Values*. State University of New York, Genesco, NY
Zulehner P, Denz H 1993 *Wie Europa lebt und glaubt*. Europäische Wertstudie, Patmos, Düsseldorf

R. Rezsohazy

Venture Capital

1. Introduction

Apparently, 'venture capital' was first used as a term in a public forum by Jean Witter in his presidential address to the 1939 Investment Bankers Association of America convention (Reiner 1989). Since then venture capital has evolved into a specialized form of finance supporting small privately owned companies judged to have the potential for fast growth. With few exceptions, the overriding goal of venture capitalists is to reap capital gains in excess of 10 times their original investment in less than seven years. A most strict definition of venture capital excludes buyouts, loans, and other financial transactions. Venture capitalists invest in equity in the form of common stock, or preferred stock, convertible debentures, or other financial instruments convertible into common stock when the small company is sold either through a merger or a public equity offering. At this liquidity event venture capitalists realize their profits in the form of capital gains.

2. The Venture Capital Industry

Since the 1950s a number of the fastest growing firms, particularly in the technology area, in the US economy received venture capital. These include Amazon.com, America Online, Amgen, Apple Computer, Cisco Systems, Compaq, DEC, Federal Express, Genentech, Intel, Lotus, Netscape, Oracle, Seagate, Sun Microsystems, 3Com, Yahoo!, and countless others. These and other venture capital-supported firms were responsible for the establishment of entirely new industries such as biotechnology, semiconductors, database software, the Internet, hard disk drives, minicomputers, workstations, and data communications. In this way venture capital became an important part of the US national system of innovation. Since the emergence of formal venture capital organizations there has been significant institutional experimentation seeking the best format to support venture investing. Not surprisingly, the organizational forms have varied through time. Though the independent private venture capital firm is the dominant organizational form, it is not the only one. Financial institutions particularly banks using their small business investment corporations (see below) have also been significant participants since the early 1960s. Industrial corporations have had significant venture capital operations since the mid-1960s. There have also been other formats including publicity-traded closed-end funds, other publicly traded vehicles, and family funds. In the early 1980s the private firms became the dominant format for venture investing. In the private firm the venture capitalists manage the limited partnership funds, each of which was denoted by a Roman numeral in increasing order. In these funds the venture capitalists were the general partners and investors were the limited partners. The investors included pension funds, foundations, endowments, corporations, wealthy individuals, and various other entities. The duration of a particular fund was typically between seven and ten years. During the first one to

three years the fund actively invests, then in the final years the fund harvests its investments. Compensation for the general partners has two components. The first component is an annual management fee paid by the limited partners; this can range from two to three percent of the limited partner's investment. These fees pay the costs of operating the firm and the salaries of all the employees including the venture capitalists. Very often, successful firms have two or more funds operating simultaneously, all of which provide management fees. The other component is the carried interest, which refers to the percentage of the capital gains the venture capitalists retain from the investments after reimbursing the initial investment of the limited partners. The carried interest varies by fund from 20 to 30 percent of the total returns. In the case of successful funds this can be very large.

3. *History*

Venture capital investing has always been cyclical, and largely governed by the receptivity of the stock market to new stock offerings. However, since the inception of formal venture capital investing in 1946 there has been a continuing growth in the amount of venture capital available. In 1965 venture capitalists invested less than $100 million in new ventures. In 1999 investments had increased to $48 billion (Venture Economics 1999). For centuries, wealthy individuals have invested in risky ventures in the hopes of securing an outsize return. Such investments can be traced back at least as far as the involvement of merchants in organizing consortia to finance long-distance trading ventures. The financing for firms participating in the Industrial Revolution came from merchants and other wealthy individuals willing to back entrepreneurs building manufacturing facilities or exploiting a new technology. These informal investors are now known as 'angels,' i.e., individuals who are not professional investors in startups.

During the 1930s in both the US and the UK there were discussions of the lack of investment capital for small businesses. In the UK the 1931 Macmillan Report identified a shortage of private sector funds for supporting small business. The report contributed to a contemporaneous debate in the US about how to provide funds for the small businesses severely affected by the Great Depression. There was little agreement about what types of small businesses should receive funding and the proper vehicles for providing financing. Prior to World War II, there were a few wealthy families, such as the Mellons, Phipps, Rockefellers, and Whitneys, investing in young ventures. Shortly after World War II in New York City these families established offices and hired professional managers to discover, evaluate, and invest in small companies with superior growth prospects. These were the first family venture capital funds. Though each of them changed many of their organizational characteristics, the Rockefeller (Venrock), Whitney (J. H. Whitney), and Phipps (Bessemer Ventures) funds continued to be active in 2000. The other significant private sector initiative was the establishment of American Research and Development (ARD) in Boston in 1946. ARD was important, not only because of the firms it funded, but also because of the Harvard Business School (HBS) faculty member, General Georges Doriot, who established it. Doriot influenced a number of HBS students who later became venture capitalists. As important, Doriot through his writing, speaking, and example defined many of the practices of venture investing. Finally, ARD's $70,000 investment in Digital Equipment Corporation in 1958 was worth $350 million in 1971—ample proof that significant return was possible. The role of the US government in the development of venture capital has been limited with a few important exceptions. In 1958 the US government passed legislation creating the Small Business Investment Corporations (SBICs) to encourage investment in small businesses. Despite being plagued by abuse during its first decade, the SBIC program made two important contributions to the development of the industry. The first contribution was its loan guarantees, which permitted participants to leverage their private funds. This made low-cost capital available to investors, a number of whom soon left the SBIC program and established important independent venture capital firms. The SBIC Act also loosened Depression-era laws forbidding banks from owning more than 5 percent equity in any company. By having an SBIC the banks were permitted to own large equity stakes in their venture-type investments. Particularly in the 1960s and 1970s banks were a significant source of venture capital. During the 1980s the independent firms displaced them.

There was other Federal legislation which had a significant impact on venture capital. The most important was a series of changes in employee pension fund laws during the 1970s. The first change was dramatic tightening by the Employment Retirement Investment Security Act of 1974, which operated to nearly cut off pension fund investment in venture capital. The legislation was gradually loosened from 1977 through 1982 and new pension fund monies flooded into the independent firms. The lowering of capital gains tax rates, may also have encouraged venture investing (Bygrave and Timmons 1992). With these exceptions the federal government has only had a limited role in creating the industry.

4. *Investment Process*

Venture capitalists are not passive investors. Because of their equity stake, they have an interest in the firm's success, which is defined in terms of a rapid increase in the firm's value. Invariably, one or more venture

capitalists 'lead investors' join the firm's board of directors. From this position they actively monitor the firm's progress. However, they do more than monitor. They often assist and counsel management. This involvement comes in many forms and is situationally dependent, but includes counseling on major strategic decisions, assisting in recruiting key personnel, providing contacts in the legal, investment banking, and other business service communities. Finally, the investment by a prestigious venture capital firm creates the perception that the small firm will be a success, an important asset for convincing customers and suppliers that the startup is viable and worthy of patronage.

Entrepreneurship depends on the structure of investment opportunities; in the typical venture capital investment scenario, an entrepreneur or entrepreneurial team approaches a venture capitalist with a business plan describing the firm, the management team, potential competitors, and the market (Martinelli 1994). Since venture capitalists receive hundreds of business plans per annum, though most receive at least a cursory examination, the plans recommended by someone the venture capitalist knows receive the greatest attention. If one of the venture capitalists is interested, then the entrepreneurial team is invited to present its business plan in person. Should the venture capitalist's interest continue, the entrepreneurs' references and others who might be able to speak of the candidate's qualifications, personal and technical, are contacted. If there are technological unknowns the venture capitalist gathers information on the project's feasibility from technical experts. They also interview prospective customers trying to gauge their need or desire for the product. In other words, they gather as much information as possible about the investment and its growth potential. If these hurdles are surmounted, then the team is invited to present the plan to a meeting of all of the partners. Should the partners agree, then a contract detailing the terms of an offer is extended to the entrepreneurs. Upon acceptance the venture capitalist invests the firm. With this investment the relationship between the entrepreneur changes as the venture capitalist now is committed to the success of the firm. Venture capital investments are staged, and, if the firm meets preset milestones, then it receives new investments at a higher valuation. Should the firm not meet its targets or the market changes at the next stage the firm's valuation can decrease. Usually, venture capitalists syndicate their investments with other venture capitalists, thereby spreading risk, providing multiple contacts and sources of advice for the startup firm, and a cross-check on the investment decision (Gompers and Lerner 1999). There are three exit strategies for venture investments: sale of the stock through an initial public offering, sale of the firm through a merger, or bankruptcy. Upon any of these liquidity events the proceeds are returned to the limited partners as either cash or marketable securities. As a rule of thumb, three or four investments are expected to fail at a near total loss. Another three of four neither fail nor are easily liquidated. These are termed the 'living dead' and are judged failures by the venture capitalists. For the successful funds it is the final two or three that determine the return. In these cases one or two investments provide returns of 10 to over 100 times the initial investment. It is these successes that compensate for the failures and complete the venture capital cycle.

5. *The Globalization of Venture Capital*

Venture capital as a formal institution was an US innovation. In other countries and other historical periods, merchant banks, governmental bodies, and even family networks performed some similar functions, especially with regard to money raising. Beginning in the late 1970s a few of the US funds looked abroad for funding opportunities; also, in other countries' governments, financial institutions, and individuals wishing to reproduce the US experience established funds targeting local opportunities. Most of these non-US efforts floundered because of a dearth of entrepreneurs, an inadequate institutional infrastructure to nurture and support entrepreneurship, and a lack of experienced venture capitalists. As a result, if these venture capital operations survived, it was, with only a few exceptions, by becoming what in the US would be called private equity capital firms concentrating upon management buyouts, bridge financing, purchase of equity in the stock market, and an occasional startup. The exceptions were a few firms in the UK, and notably, Israel, which became a center of venture capital-financed startups in the 1990s. In the late 1990s, due to the opportunities for new firm formation to exploit the Internet, there was a proliferation of venture capital firms in, not only Europe, but also Latin America and Asia.

The impressive returns achieved by US venture capitalists attracted investors from around the world. In some cases, these funds were routed through venture capital firms based in the US and managed by foreign professionals or corporate venture capital investment operations operated by foreign multinationals. More frequently, European and Asian funds were invested in US funds and in US startups. Despite these investments, non-US capital has only played a minor role in financing US venture capital.

6. *Conclusion*

There are continuing debates regarding the efficacy of a venture capital-based innovation system. There have been recurring criticisms of venture capitalists for encouraging the exodus of high-level managers and

even entire research teams from existing firms, thereby disrupting product development teams (Florida and Kenney 1990). In Europe some have seen in venture capital the harbinger of rampant individualism and greater income inequality and social differentiation. The emphasis on large capital gains is thought by some to bias the economic system against incremental process innovation and large-scale systems innovation such as, for example, in transportation systems. Despite these concerns, venture capital as an institution for supporting innovation has experienced a pattern of cyclical but massive secular growth since the 1940s.

See also: Capitalism; Capitalism: Global; History of Technology; Information Society; Institutional Investors; International Marketing; Internet: Psychological Perspectives; Stock Market Predictability; Stockholders' Ownership and Control; Technological Innovation

Bibliography

Bygrave W D, Timmonsd J A 1992 *Venture Capital at the Crossroads*. Harvard Business School Press, Boston

Florida R, Kenney M 1990 *The Breakthrough Illusion*. Basic Books, New York

Gompers P A, Lerner J 1999 *The Venture-Capital Cycle*. MIT Press, Cambridge, MA

Martinelli A 1994 Entrepreneurship and management. In: Smelser N J, Swedberg R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ

Reiner M L 1989 The transformation of venture capital: a history of venture capital organization in the United States. Ph.D. Dissertation. University of California, Berkeley, CA

Venture Economics 1999 *1999 National Venture Capital Association Yearbook*. National Venture Capital Association, Washington, DC

M. Kenney

Verbal Communication, Psychology of

Although linguists think of language as an abstract structure—a set of principles that specify the relations between a sequence of sounds and a sequence of meanings—to its users, what is most significant about language is its versatility as a medium for communication. The ability to communicate is vital to a species' survival, and all animal species communicate, some in ways that are impressively proficient. But none achieve the precision and flexibility that characterizes human communication, a capacity due in large part to the uniquely human ability to use language (Deacon 1997, Hauser 1996).

1. Elements of Communication Systems

All communication systems, regardless of how simple or complicated they may be, operate on the same principle: Signals transmit messages from a source to a destination. The distended belly of the female stickleback signals the male to initiate an elaborate courtship routine that culminates in fertilization of her eggs (Tinbergen 1952). Upon returning to its hive, a foraging honeybee communicates the direction and distance of a source of nectar by engaging in an elaborate waggle-dance (von Frisch 1967). Vervet monkeys (native to East Africa) have three distinctive vocal alarm calls that signal the presence of leopards, eagles, and snakes, their three main predators. Upon hearing one or another call, a Vervet will respond appropriately—climbing a tree in response to the leopard call, scanning the ground when the snake call is sounded (Seyfarth et al. 1980). Each of these systems represents the species' adaptation to the exigencies of a particular ecological niche in which communication facilitates survival. Language can be thought of as a similar sort of adaptation.

Communication systems use two kinds of signals: signs and symbols. Signs are signals that are causally related to the message they convey. We say that blushing means someone is embarrassed because we know that embarrassment is a cause of blushing. Symbols, on the other hand, are products of social conventions. Because of an implicit agreement among speakers of English, the sound pattern we recognize as the word *dog* denotes the familiar category of furry, four-legged creatures. There is no intrinsic reason that *dog*, rather than some other sound pattern, should convey that message, and in languages other than English, of course, very different sound patterns represent the concept DOG. Verbal communication often involves both signs and symbols. The tremulous voice that tells us a speaker is experiencing distress is a sign, i.e., a direct product of the distress it signals. But it is the symbolic content of verbal communication that accounts for its extraordinary effectiveness.

2. Semanticity, Generativity, and Displacement

Language is only one of the symbol systems humans use to communicate. The 'thumbs-up' gesture conveys the message of success, approval or hope; the wedding ring and the mourner's ribbon publicly proclaim the wearer's current status; a facial grimace in response to the question 'How did you like the movie?' symbolically and effectively expresses the person's assessment. Notwithstanding the utility of such symbolic displays, language endows human communication with three properties, semanticity, generativity, and displacement, that collectively distinguish it from other sorts of symbolic displays and from the forms of communication observed in other species.

Semanticity: In human communication, signals stand for things, which is to say that they have meaning. An overheated dog will pant to dissipate heat, and an astute observer may understand the panting to indicate that the dog is hot, but panting cannot be said to stand for overheatedness in the same way that the word 'overheated' does.

Generativity (sometimes called Productivity): All languages are capable of generating an infinite number of meaningful messages from a finite number of linguistic signals. Languages allow symbols to be combined and recombined in ways that yield novel meanings, and any competent language user will regularly produce and comprehend utterances that have never been uttered before, but are immediately comprehensible to all competent language users. Even so sophisticated a system as the Vervet's alarm calls is limited to a fixed set of messages, and lacks the ability to generate novel ones. For the Vervet, there is no way to signal the presence of predators other than eagles, leopards, or snakes.

Displacement: Language makes it possible to communicate about things that are remote in space or time, or indeed exist only in the imagination. Bertrand Russell once remarked that 'No matter how eloquently a dog may bark, he cannot tell you that his father was poor but honest.' Although the observation is self-evident, even banal, it points to a fundamental difference in the expressive capacities of language and other communication modalities. Vervets can signal the presence of a predatory eagle, but even the most articulate Vervet cannot refer to the eagle that attacked a week ago; their communication is limited to what is immediately present. Perhaps more than any other feature, it is the capacity of language to convey displaced messages that distinguishes it from other communication modalities.

The ability of language to generate an unlimited number of meaningful novel messages that are not bound to the here and now, combined with the cognitive capacity to exploit these properties, allows human communication to be extraordinarily effective and versatile. The full extent of this effectiveness and versatility is revealed by an examination of how communication systems work.

Regardless of the type of signal, all communication involves the transfer of information between a source and a destination. Anthropologists Daniel Sperber and Denise Wilson characterize it as:

> …a process involving two information-processing devices. One device modifies the physical environment of the other. As a result, the second device constructs representations similar to the representations already stored in the first device (Sperber and Wilson 1986 p. 1).

The description applies equally well to the transmissions of a fax machine, the gesticulations of a policeman directing traffic, or a conversation among intimates. In verbal communication the 'information-processing devices' are people, the 'representations' are mental representations or ideas, and the 'modifications of the physical environment' are the uniquely human disturbances of the acoustic surround called speech.

3. Four Communication Paradigms

Precisely how is language used to convey information? There are many characterizations of the ways language functions as a medium for communication, but the major ways are captured by four models or paradigms (Krauss and Fussell 1996). Each paradigm focuses on a different dimension of language use, and might be thought of as a necessary but incomplete description of the process. The four paradigms are: the encoding–decoding paradigm, the intentionalist paradigm, the perspective-taking paradigm, and the dialogic paradigm.

3.1 Encoding and Decoding

Language often is described as a code that uses words, phrases, and sentences to convey meanings. A code is a system that maps a set of signals onto a set of significates or meanings, and in the simplest kind of code, the mapping is one to one: for every signal there is one and only one meaning; for every meaning, there is one and only one signal. In Morse code, a familiar example of a simple code, the signals are sequences of short and long pulses (dots and dashes) and the significates are the 26 letters of the English alphabet, the digits 0–9, and certain punctuation marks. The encoding–decoding approach to language conceives of communication as a process in which speakers encode their ideas in words, phrases and sentences, and listeners decode these signals in order to recover the underlying ideas.

Words, phrases and sentences do convey meanings, of course, but viewing linguistic communication simply as encoding and decoding doesn't do justice to the subtlety of the process by which language communicates. In order to appreciate this, we need to consider the role of the speaker's communicative intention and how it is related to comprehension.

3.2 Communicative Intentions

In Morse code, the sequence · · · · designates the letter H, and only that letter, and the letter H is represented by the sequence · · · ·, and only that sequence. In contrast, a speaker who says 'Can you close the door?' might be understood to be: (1) requesting that the auditor shut the door; (2) asking whether the auditor is physically able to shut the door; (3) asking whether the door's physical condition is such that it could be shut; or (4) asking whether shutting the door is permitted. Typically, by saying 'Can you close the door?' the

speaker intends only one of those meanings to be understood. and in all likelihood that is the meaning that will be understood.

A listener who has understood an utterance has grasped the particular sense in which the speaker intended the words to be understood. Considering the number of meanings even the simplest utterance is potentially capable of conveying, the ability of addressees to identify the intended meanings of the vast majority of the utterances they encounter is truly remarkable. This ability is the foundation of human communication. As the linguist Stephen Levinson puts it: '…communication is a complex kind of intention that is achieved or satisfied just by being recognized' (Levinson 1983 p. 18).

There is abundant evidence to support an intentionalist view of language use. For example, bank clerks, when asked 'Can you tell me what the interest rate is?' are likely to respond to the utterance as a request ('Tell me the interest rate') rather than to its literal force ('Are you able to tell me the interest rate'). Listeners demonstrate considerable mental agility in divining speakers' communicative intentions. In one study, people were shown a picture of then-President Ronald Reagan seated alongside David Stockman (the not-well-known director of the Office Management and Budget), and asked one of two questions: 'You know who this man is, don't you?' or 'Do you have any idea at all who this man is?' As reflected in their answers, people asked the first question overwhelmingly assumed the question was about President Reagan, while none of those asked the second question did (Clark et al. 1983). That people strive to understand the communicative intention that underlies the utterance is not in doubt. What is less clear is how they go about doing so. Identifying another's communicative intention is not always a simple or straightforward matter, in part because people do not always perceive the world in the same way. This difficulty is usually formulated as a problem of Perspective-taking.

3.3 Perspective-taking

The meaning of even the most prosaic utterance is grounded in a set of implicit assumptions about what the communicators know, believe, feel, and think. People experience the world from different vantage points, and the totality of each individual's experience is unique to the particular vantage points he or she occupies. To accommodate discrepancies in perspective, communicators must take each other's perspectives into account when they formulate and interpret utterances. The process is most readily illustrated by spatial perspective-taking. Two people viewing the same layout of objects from opposite sides of a table will perceive the objects' spatial relations somewhat differently. The cream pitcher that is to the right of the sugar bowl for one person, will be to the left of it for the other. In talking about such things, or in giving directions, speakers tend to formulate spatial relations from the addressee's point of view ('It's the first door on your left'), even when it differs from their own (Schober 1993).

But even when they communicate about matters that lack spatial content, speakers must take others' perspectives into account. A statement like 'John is the fellow wearing the ugly tie' assumes that speaker and addressee share the same esthetic sensibility. For the utterance to communicate, the speaker must have in mind the addressee's conception of what constitutes an ugly tie. As the social psychologist Roger Brown put it, effective communication 'requires that the point of view of the auditor be realistically imagined' (Brown 1965).

However, the content of another's point of view is not always obvious. In his classic studies of childhood egocentrism, the Swiss developmental psychologist, Jean Piaget, demonstrated that the ability to apprehend others' perspectives represents a major milestone in the child's intellectual development. Young children are unable to detach themselves from their own points of view, and, in effect, seem to assume that the world appears to others as it does to them (Piaget 1952). This reduces their effectiveness as communicators (Krauss and Glucksberg 1977). Although adults do better, they are far from perfect, and like children their judgments of others' perspectives tend to be biased by their own points of view. Under time pressure or when preoccupied, adults are likely to formulate messages that neglect their addressees' perspectives (Keysar et al. 1998). Moreover, adults' perspective-taking efforts display an egocentric bias similar to that found in children. Asked to estimate the likelihood of someone identifying an individual pictured in a photograph, people who themselves recognized the person tended to overestimate his recognizability, while people who did not recognize him underestimated it (Fussell and Krauss 1991).

Such limitations in the ability to appreciate others' perspectives might make verbal communication considerably less effective than it is were it not for the fact that communication is a collaborative rather than an individual process. The collaborative nature of communication is the focus of the paradigm we call Dialogic.

3.4 Dialogism

The encoding-decoding, intentionalist, and perspective-taking paradigms all characterize communication in terms of individual acts of production and comprehension. In this view, participants are viewed as 'autonomous information processors' (Clark and Brennan 1991). The speaker must generate utterances that will convey a particular meaning; the addressee

must process those utterances, and by so doing identify the speaker's intended meaning. Communicative exchanges that relied exclusively on such an arrangement would impose a heavy cognitive burden on the participants. Probably because we do it so well, we seldom appreciate how complicated a process communication is. Producing spontaneous speech requires the speaker to perform two cognitively demanding tasks simultaneously: conceptualizing the information to be conveyed, and formulating a verbal message that is capable of conveying it. The number and complexity of the factors that must be taken into account is dauntingly large (Levelt 1989). The addressee's task is equally challenging. Speech is evanescent; once it has been articulated, it must be processed and comprehended in real time. Conversational speech is produced at a rate of about 2.5 words per second, often in noisy environments and with less-than-perfect articulation. Production and comprehension could pose formidable problems for two completely autonomous information processors. Yet participants typically come away from conversations believing they have communicated successfully, and objective evidence probably would indicate that they have.

One reason people are able to communicate as well as they do in such adverse circumstances is that the exquisite responsiveness of conversation (and similar highly interactive forms) permits them to formulate messages that are closely attuned to each others' immediate knowledge and perspectives, which reduces the cognitive demands of production and comprehension. The participant who at a given moment occupies the role of speaker can determine virtually instantaneously whether the addressee has identified his or her communicative intentions correctly. Simultaneously, the addressee can reveal the nature of his or her understanding as it develops, and in this manner guide the future production of the speaker.

As Herbert Clark and Susan Brennan (1991) have observed, certain activities by their very nature involve joint or collective action, and it makes little sense to think of participants' actions as individual events.

> It take two people working together to play a duet, shake hands, play chess, waltz, teach, or make love. To succeed, the two of them have to coordinate both the content and process of what they are doing…Communication, of course, is a collective activity of the first order.

Close examination of what actually is said in conversations reveals that it is seldom an orderly process in which participants alternate in the roles of speaker and listener. Rather, sentences often trail off inconclusively or are left dangling incomplete; listeners interrupt to ask questions, interject comments, and finish sentences; topics change abruptly and unpredictably; and what is left unsaid may convey more than what is explicitly stated. It would be a mistake to regard such conversational speech as a defective version of some ideal form. Rather, these apparent aberrations reflect the way conversation operates as a communicative process.

In the dialogic view, conversational speech is the model for communication, and a communicative exchange represents a joint accomplishment by the participants, who have collaborated to achieve some communicative goal. As a result, meaning is 'socially situated'—deriving from the particular circumstances of the interaction—and individual contributions are not fully meaningful apart from that situation.

4. *Future Directions*

Although language must have evolved in the setting of face-to-face communication, human inventiveness has long sought ways to transcend the limitations of time and space that physical co-presence imposes. The most important technological development in this quest was the invention of word-syllabic systems of writing, probably in Mesopotamia around 4,000 years ago. Phonogramic representational systems made it possible to transform the ephemera of speech into a tangible record that exists independent of the person who generated it—a development with profound consequences at both the individual and the societal level (Goody 1977).

Writing systems, combined with more recently invented technologies for recording, transmitting, and reproducing signals, make it possible to transmit a dizzying variety of materials (data, text, graphics, facsimile, to name only a few) over virtually limitless distances. Surely the enhanced communicative capacities that such technologies make possible have wide-ranging potential consequences for the way human life is constituted.

A species' survival depends critically upon its ability to communicate effectively, and the quality of its social life is determined in large measure by how and what it can communicate. Human social life as it is presently constituted is predicated upon an extraordinary level of communicative virtuosity—a level of virtuosity that the uniquely human ability to use language confers. Absent this, our lives would be quite different.

See also: Communication and Social Psychology; Communication: Philosophical Aspects; Communicative Competence: Linguistic Aspects; Evolution and Language: Overview; Language Acquisition; Language and Animal Competencies; Language and Thought: The Modern Whorfian Hypothesis; Language Development, Neural Basis of; Syntactic Aspects of Language, Neural Basis of

Bibliography

Brown R 1965 *Social Psychology*. The Free Press, New York
Clark H H 1996 *Using Language*. Cambridge University Press, Cambridge, UK

Clark H H, Brennan S E 1991 Grounding in communication. In: Resnick L B, Levine J M, Teasley S D (eds.) *Perspectives on Socially Shared Cognition*. APA, Washington, DC, pp. 127–49
Clark H H, Schreuder R, Buttrick S 1983 Common ground and the understanding of demonstrative reference. *Journal of Verbal Language and Verbal Behavior* **22**: 245–58
Deacon T W 1997 *The Symbolic Species: The Co-evolution of Language and the Brain*. W.W. Norton, New York
Fussell S R, Krauss R M 1991 Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology* **21**: 445–54
Goody J 1977 *The Domestication of the Savage Mind*. Cambridge University Press, Cambridge, UK
Hauser M D 1996 *The Evolution of Communication*. The MIT Press, Cambridge, MA
Keysar B, Barr D J, Horton W S 1998 The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science* **7**: 46–50
Krauss R M, Glucksberg S 1977 Social and nonsocial speech. *Scientific American* **236**: 100–105
Krauss R M, Fussell S R 1996 Social psychological models of interpersonal communication. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology: A Handbook of Basic Principles*. Guilford Press, New York, pp. 655–701
Levelt W J M 1989 *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA
Levinson S C 1983 *Pragmatics*. Cambridge University Press, Cambridge, UK
Piaget J 1952 *The Language and Thought of the Child*. Humanities Press, New York
Schober M F 1993 Spatial perspective-taking in conversation. *Cognition* **47**: 1–24
Seyfarth R M, Cheney D L, Marler P 1980 Vervet monkey alarm calls: Semantic communication in a free-ranging primate. *Animal Behavior* **28**: 1070–94
Sperber D, Wilson D 1986 *Relevance: Communication and Cognition*. Harvard University Press, Cambridge, MA
Tinbergen N 1952 Derived activities: their causation, biological significance, origin and emancipation during evolution. *Quarterly Review of Biology* **27**: 1–32
von Frisch K 1967 *The Dance Language and Orientation of Bees*. Belknap Press of Harvard University Press, Cambridge, MA

R. M. Krauss

Verstehen and Erklären, Philosophy of

Verstehen (interpretative understanding) and *erklären* (law-governed explanation) are two ways to make scientifically respectable sense of a phenomenon. The scientist who engages in *erklären* tries to make explanatory sense of the phenomenon by finding the laws that govern it, whereas the scientist who engages in *verstehen* tries to make empathetic sense of the phenomenon by looking for the perspective from which the phenomenon appears to be meaningful and appropriate.

The originally German tradition of *verstehen* insists that the social sciences are unlike the natural sciences in that making merely *explanatory* sense of social phenomena can never lead to comprehensive knowledge of these phenomena, even if explanation would be enough, and indeed is all there is to comprehensive knowledge of natural phenomena. The reasons for this claim will be reviewed in this article. A first set of reasons, discussed in Sect. 1, derives from an investigation of the differences between natural and social phenomena. A second set, discussed in Sect. 2, derives from a survey of a number of controversies that are high on the agenda of those who assume that *verstehen* and *erklären* are incompatible methods, only one of which can be the proper method of the social sciences.

1. *The Autonomy of the Geisteswissenschaften*

The German verb *verstehen* is, from a philosopher's point of view, closely connected with Dilthey (1833–1911) and his concern to secure the autonomy of the *Geisteswissenschaften*. A salient feature of the development of German intellectual culture in the nineteenth century was a growing awareness of the fact that philosophy and metaphysics are themselves historical phenomena. An important consequence of this insight was that thoughts and ideas and such like came to be seen as part of the furniture of the world. Thoughts and ideas are not outside the world in some timeless conceptual realm; they are objects in the world that could, in principle, be studied by empirical science. Of course they are peculiar objects, these mental appearances ('*geistige Erscheinungen*'); they are objects with content, quite unlike natural phenomena, but are nonetheless empirical objects. An excellent example is a text printed in a book. That is obviously an empirical object—ink on paper—but is very unlike the ink spilled on blotting paper. Crucial to the text is that it means something, and this feature of the text is so important to its identity that we can easily make sense of the claim that a text which physically consists of nothing but ink on paper can still be numerically identical to a text which physically consists of sounds produced by a human voice, or even (as we now know) to a text which physically consists of a long series of bits on a computer disk.

The fact that the meaning of mental appearances such as texts is so fundamentally important to their identity motivated Dilthey to defend the view that there are two radically different kinds of empirical objects: natural and mental phenomena. According to Dilthey, this dualism within the empirical realm asks for and justifies a dualism on the level of scientific methods. Natural science already proved to be successful in making explanatory sense of natural phenomena by means of *erklären*. But the *Geistes-*

wissenschaften could not and should not try to make sense in a similar way (by *erklären*) of mental phenomena such as thoughts and ideas, even though it finally becomes commonsensical—or so Dilthey believed—that mental phenomena, like natural phenomena, are part of the furniture of the empirical world. If the *Geisteswissenschaften* were to be successful, Dilthey argued, they should not look for explanations in terms of covering laws, but should concentrate on the attempt to make empathetic sense of mental phenomena by means of *verstehen*. After all, the only way to acquire scientifically respectable knowledge of a phenomenon is to gain comprehensive insight into what is of crucial importance to the essence of the phenomenon; in the case of the mental phenomena, this means grasping their meaning.

In order to appreciate the relevance of this reasoning to the social sciences, it is important to point out how likely it is that one will take a second step after having pointed out the relevance of a distinction between mental and natural phenomena. This second step consists in claiming that what is crucial to social phenomena such as human actions is that they mean something, which makes them much more similar to mental appearances such as texts than to natural phenomena such as leaf-fall in autumn. It is relevant in this context to note that the twentieth-century French philosopher Ricoeur (1981, pp. 197–221) has argued that it is almost literally true that human actions are texts.

According to philosophers who defend the autonomy of the *Geisteswissenschaften* (e.g., Dilthey), the paradigmatic objects to be investigated by means of *verstehen* are the social and cultural manifestations of human life, characterized not only by being (a) meaningful, but also by being (b) purposive and (c) essentially relational. As will be shown in the remainder of this section, some of the details of the method of *verstehen* can be brought to the fore by discussing these three aspects of the manifestations of human life.

However, in this connection it should be stressed that although the distinction between *erklären* and *verstehen* was originally made in close connection with that between natural and mental phenomena as well as that between natural science and the *Geisteswissenschaften*, there is no need to accept these three distinctions as one package. That is, one can defend the view that there is a real and important distinction to be made between *erklären* and *verstehen*, independent of whether or not it is (a) plausible to defend that a neat split can be made between natural and mental phenomena, and (b) useful to divide the academic disciplines into the natural sciences and the *Geisteswissenschaften*. In the remainder of this article, it will be taken for granted that the social and behavioral sciences have no interest in making or accepting the latter two distinctions, although it is important for them to accept and take notice of the methodological distinction between *erklären* and *verstehen*.

1.1 Meaningfulness

Things have meaning (at least to a very large extent) in virtue of the fact that people ascribe meaning to them. These ascriptions can be diverse and numerous and can display a multitude of perspectives, motives, and biases. This has two consequences for the scientist engaged in *verstehen*. The first is that it is pretty absurd—if not impossible—to try to capture a phenomenon's meaning independent and irrespective of the meaning ascribed to it by all those people in whose life this phenomenon has a role to play. The meaning of, for example, a gesture such as a smile cannot be found if you start by excluding the meaning ascriptions of all the people involved. This should not, of course, mislead the scientist into thinking that a particular meaning ascription is as good as any other. Nor should he simply assume that particular persons (e.g., the author of a text) have a principled authority over the meanings to be ascribed. *Verstehen* is a delicate undertaking that requires careful attention. A scientist engaged in *verstehen* faces the important challenge of deciding on the relevance and import of each and every meaning ascription (see Schutz 1970).

The second consequence of the prominence of meaning ascriptions is that a double hermeneutics necessarily forms part and parcel of *verstehen*. Trying to grasp the meaning of a phenomenon is always a matter of trying to grasp the meaning of a pre-interpreted phenomenon. And it is one thing to grasp the meaning ascribed to the phenomenon by an earlier act of interpretation, but quite another to grasp the meaning this earlier act of interpretation attempted to grasp. This can lead to a conception of *verstehen* as a quest for the original meaning of a phenomenon, a conception that was dominant in the hermeneutic tradition of Dilthey. The German philosopher Gadamer (1960), however, argued that such a quest beyond all prejudices would be futile. A prejudice is an original meaning ascription, and—rather than blocking the road to empathetic understanding—these prejudices open up the very possibility of *verstehen*. There are no original meanings, according to Gadamer. There are only original meaning ascriptions. The intentions of the person who, for example, makes a gesture, share this character of being a prejudice (i.e., a meaning ascription) with the social scientist's attempt to make empathetic sense of this gesture. Meaning exists, in Gadamer's view, in the multifaceted meeting of different meaning ascriptions to the same phenomena. This leads to a conception of *verstehen* in which the growth of self-understanding is always as central to the scientific enterprise as is the growth in interpretative understanding of the phenomenon investigated.

1.2 Purposiveness

According to many, teleology—or purposiveness—is a distinctive feature of human action. Human actions take place *in order to* bring certain future states of affairs about, not merely *because* certain past states of affairs happened. I go to the hairdresser *in order to* have my hair cut, not merely *because* my hair has grown. This emphasis on goal-directedness and final causes in the description of human action is foreign to the spirit of modern science, which seems so successfully engaged in redescribing purposiveness in terms of functionality and functionality in terms of organizational structures that arise through natural selection. Human action seems to be a last vestige of teleology, and the adherents of the *verstehen* tradition took pains to defend it by arguing for a radical distinction between human behavior and all other natural events. *Erklären*, they argue, is a matter of making explanatory sense of natural events by citing their cause, but there is more to be done than just that in making sense of human actions. Part of what is needed in making sense of actions is to rationalize them, that is, to make empathetic sense of them by citing their reason. And reasons are essentially normative: if I were to try to explain my visit to the hairdresser by merely citing the fact that my hair has grown, I would fail to explain my action, unless I explicitly mention the implicit normativity, that is, that my hair has grown *too long*. The normativity that figures conspicuously in the rationalizing of actions is the tough nut to crack for those who wish to remove purposiveness from any description of such mental phenomena as human actions.

A lot of extremely interesting work done on the relation between explaining by citing causes and explaining by citing reasons can be found in the Anglo-American literature (Davidson 1980, Taylor 1964, Malcolm 1968, Wright 1971, Dennett 1973, Macdonald and Pettit 1981). This work, although mainly written without engaging with the tradition of *verstehen*, is very helpful in that it elucidates what it means to make empathetic sense of an action by looking for the reasons that make the action meaningful and appropriate.

1.3 Essential Relationality (Holism)

According to adherents of the *verstehen* tradition, we pick out natural and mental phenomena in very different ways. We can pick out, or individuate, natural phenomena without engaging in an attempt to make explanatory sense of them. That is precisely what creates the demand for natural science: here is a natural event (say the collapse of a bridge) and we can identify it in such a way that we know we are talking about this event, even though we have not the faintest idea of what caused it to happen. Thus, we can fail to understand a phenomenon we can describe quite accurately, and this is what motivates us to engage in *erklären*, that is, to look for causal and law-governed explanations.

We cannot, however, pick out mental phenomena in a similar way. Because their meaning is crucial to their identity, it is impossible to individuate a mental phenomenon without, by the same token, engaging quite seriously in an attempt to make empathetic sense of it. That is, the atomism that seems possible with respect to natural phenomena (each of which can be described in total isolation from the others) is completely foreign to the realm of mental phenomena. Getting acquainted with these phenomena requires holistic assumptions. In order to pick out such a mental phenomenon as a vote, for example, we need a tremendous amount of detailed knowledge about a whole background in which events could count as votes. This entails knowledge of institutions, of means to justify authority, of decision procedures, etc. This semantic background contains more than just the conceptual relations that make it possible for something to be a vote. It also contains the intentional relations that make it true that some bodily behavior constitutes a particular vote; the intentional relations that determine, for example, that this arm being raised at this particular moment counts as a vote for X, a vote that could also have been expressed by this particular option box being ticked, or by this name written down on this particular slip of paper, etc.

The holism that is essential to the individuation of mental phenomena has important consequences for the method called *verstehen*. Whereas it might be true that a law-governed explanation of a particular natural phenomenon urges us to completely revise our ordinary description of the phenomenon (e.g., contrary to what we always thought, we now know that solid bodies consist mainly of the void), it is inconceivable, according to adherents of the tradition of *verstehen*, that a scientifically respectable, interpretative understanding of the manifestations of human life could radically replace our ordinary understanding of them. The *Geisteswissenschaften*, it is claimed, are bound to elaborate on the conceptual scheme of common-sense (nowadays often labelled folk psychology), because they can only begin to make empathetic sense of mental phenomena such as human actions in virtue of accepting the commonsensically sound understanding of these actions implied by our means to pick them out in the first place.

This is known as the hermeneutic circle, or the inescapability of each individual's own traditional semantic horizon. It raises many of the most important questions discussed in the *verstehen* tradition, for example, how to understand the manifestations of a foreign culture, or texts written in bygone times.

The upshot of the discussion so far is that there is a useful distinction to be made between a scientific interest in the causal laws that govern phenomena and a scientific interest in the perspective from which

phenomena appear to be meaningful and appropriate. The former interest can be satisfied by engaging in *erklären*, the latter by engaging in *verstehen*. *Verstehen* involves (a) the careful assessment of the relevance of meaning ascriptions, (b) a double hermeneutics, (c) a concern for self-understanding, (d) a rationalization of phenomena by citing reasons, (e) a disclosure of implicit normativity, (f) an acceptance of semantic holism, (g) of the inseparability of description and interpretative understanding, and (h) of the inescapability of semantic horizons.

Whether or not these features disqualify *verstehen* as a scientific method, or force us radically to revise our conception of (social) science, has been the topic of serious controversies.

2. *Two Ways to Do Social Science? Some Controversies*

The German verbs *verstehen* and *erklären* almost immediately suggest the image of an opposition and, therefore, of a controversy. Although there were periods in nineteenth- and twentieth-century German intellectual life during which *verstehen* and *erklären* were conceived to coexist peacefully—the first being at home in the humanities, the second in the natural sciences—the most disturbing, and fruitful, interest in the concepts arose during two periods (the first and second *Methodenstreit*) in which people fought over which of the two methods would be most appropriate to the social sciences. During the first period (roughly 1900–30), neo-Kantians such as Rickert, Cassirer, and, most influentially, Weber tried to come to terms with the methodological dualism that seemed characteristic of the social sciences that emancipated at the time from philosophy. During the second period (roughly 1955–85), Wittgensteinians, phenomenologists, and critical social theorists mainly discussed the impact of the ontological commitments connected with a *verstehen* approach to social reality on our conception of (social) science.

In this section, four themes that were high on the agenda during those periods will be discussed. The discussion will be conceptual; no attempt will be made to provide a picture of the historical development of these themes (see *Hermeneutics, History of*).

2.1 *Unity of Science*

Positivism (see *Comte, Auguste (1798–1857)*; *Logical Positivism and Logical Empiricism*) followed in the wake of the natural sciences, endorsing the view that science is to be defined by the methodological procedures that enable us to explain and predict natural phenomena so successfully. If there are to be social sciences, the demand of the unity of science will imply that the method used in natural science (*erklären*) must also be applicable to human action. If human action allows for scientific knowledge this will mean, according to the positivist credo, that we can and should try to explain human actions by looking for the causal laws that govern them.

This bold statement evoked worries, misunderstandings, and refinements with respect to the place of *verstehen* in the social sciences. One plausible reaction (developed by neo-Kantians during the first *Methodenstreit*) is to defend the unity of science but deny the defining role of *erklären*. Science, one could say, is an attempt to make systematic and generally accessible sense of all phenomena. But there is more systematic and generally accessible sense to be made of what happens than merely explanatory sense. We can—and in the case of the manifestations of human life, should try to—make empathetic sense of what takes place by looking for the perspective from which these manifestations appear to be meaningful and appropriate.

Part of the positivistic project, however, was to show that there is no need and no place for empathetic sense in science. If the meaning and the appropriateness of human actions are to be accounted for by science in a systematic and generally accessible framework, this should be done, according to the positivists, by investigating objective patterns of behavior and not by telling stories about how these patterns look from the necessarily perspectivistic point of view of the people on the scene. The main problems with this line of reasoning are discussed under the three following subsections.

2.2 *Intentional Explanations*

According to the positivists, actions are to be explained by subsuming them under general, contingent, empirical, causal laws that connect mental states of individual agents (in particular, their beliefs and desires) to their actions, that is, their overt bodily movements. Acknowledging that reasons are in fact sets of beliefs and desires that function as causes is—according to the positivists (see *Hempel, Carl Gustav (1905–97)*)—the starting point for a social science that can make progress by engaging in *erklären*.

Opponents of this view argue, along three different lines, that one can make empathetic sense of human actions by producing intentional explanations (i.e., explanations citing beliefs and desires), but that such explanations are not a matter of *erklären*, because intentional explanations are neither causal nor nomothetic (involving laws).

First, a Wittgensteinian line of argument says that pointing out the relevance of the agent's beliefs and desires in order to explain the apparent appropriateness of an action is not a matter of citing its cause or of hinting at an underlying causal mechanism. Intentional explanations do nothing more than provide us with useful redescriptions of actions that surprise us (Taylor 1964, Winch 1958, Wright 1971).

Second, the Logical Connection argument says that intentional explanations cannot be causal, because the relation between the mental states of the agent and the action they are supposed to have caused is not contingent but logical. To correctly describe an action as, for example, a vote, logically requires ascribing to the agent the intention to vote. Entities that cannot really be distinct (e.g., actions and their accompanying intentions) cannot be related causally. This conclusion does not deprive intentional explanations of their role in science. Their import is not one of *erklären*, but of uncovering normative relations holding between the agents and institutions involved in the genesis of the phenomenon to be made sense of (Stoutland 1970).

A last line of argument concentrates on the nomothetic (or law-involving) character of *erklären*, arguing that things are different when explaining action: the principles of behavioral and attitudinal rationality that we use in intentional explanations do not function as contingent laws but as a priori principles. That is, *erklären* requires that we take pains to fill in the contingent, endlessly revisable details of the causal laws that govern the phenomena we investigate. *Verstehen*, however, requires us to take pains to redescribe occurring human phenomena over and over again until they make empathetic sense, and we can do this only if we accept the following principles as a priori true: agents tend to (a) act in accordance with what they believe and desire, and (b) believe what they have reason to believe to be true, and desire what they have reason to value (Davidson 1980, 1984, Macdonald and Pettit 1981).

2.3 Subjective Meanings

Many argue that one of the most important features of science is that it is a systematic attempt to transcend distorting effects of particular, biased, subjective points of view. That is, science aims at objectification, and this leaves absolutely no room for a necessarily perspectivistic *verstehen* approach.

The best reply to this line of thought from adherents of the *verstehen* tradition is to point out a problem science should have with subjective meanings, a problem one fails to recognize (let alone address) if one takes an *erklären* approach to human action. The problem was mentioned above in discussing the hermeneutic circle (Sect. 1.3). Subjective meanings are part and parcel of intentional phenomena. Part of what such phenomena are is what they are taken to be from the commonsensical point of view by those engaged in them. Understanding such phenomena requires us to investigate subjective meaning ascriptions (Sect. 1.1), being both a first step towards interpretative understanding (*verstehen*) and an essential building block of the manifestations of human life. The problem can be described as a paradox: is it possible to acquire scientific knowledge of subjective meanings if such knowledge requires transcendence of subjective points of view?

The point can also be stated in another way: to identify a phenomenon as an action implies reference to the contentful mental states (beliefs and desires) of the agent. These states must, according to the principles of behavioral and attitudinal rationality, be roughly in tune with the action and with what the agent believes to be true and desirable. We cannot change the last part of the previous sentence into 'with what *is* true and desirable,' as though this would be something radically independent of the agent's interpretation of it. That is, we cannot sidestep the attempts of agents to ascribe meaning to their actions. In speaking of actions we have to take seriously the commonsensical and folk-psychological interpretations of those involved in the actions. This does not, however, commit us to accept that the agent's interpretation of the action has a privileged authoritativeness. But neither has ours. It is in the careful and sensitive confrontation between the various ways of ascribing meanings that two things happen: (a) the manifestations of human life take shape, and (b) interpretative understanding (*verstehen*) becomes possible.

2.4 Deliberation

Rational choice theory (see *Rational Choice Theory: Cultural Concerns*) is the most serious attempt to fully acknowledge the normativity that is crucial to intentional explanation, and the subjective meanings that play their indispensable role in behavioral and attitudinal rationality, without giving up the aim of a unified science based on *erklären*. The hope of rational choice theory is to explain and predict human action in terms of laws that causally relate expected utility numbers and ensuing actions. The idea is that for an agent whose preferences conform to various constraints, it is possible to represent him as having a utility function and a probability function such that every action maximizes expected utility as computed by those functions (Eells 1982).

A strong reply is open to those favoring a *verstehen* approach. It starts from the observation that a choice is not merely an event that occurs, but is always also a decision made by an agent capable of deliberation. Consider a case where a couple wanders down the street, looking for a place to eat, and deciding to opt for that small Thai restaurant next to the popular pizzeria. It is quite likely that explanatory sense can be made of this choice in terms of the calculation of utility numbers and some additional game theoretic considerations about successful co-operation. But such a story is of no use in trying to make sense of their conversation as a matter of deliberation (Pettit 1991). If we want to make sense of the way in which the woman replies to a suggestion made by the man, and

vice versa, and of how this conversation is their way of making up their minds about where to eat, then we need something rational choice theory does not give us: interpretative understanding. That is, rational choice theory could perhaps make explanatory sense of events that we would ordinarily describe as choices, but making sense of them as choices, as decisions reached by agents who reason their way to these choices, is something that requires a *verstehen* approach. Making a choice or forming an intention is done by deliberative agents, capable of recognizing and responding to considerations of evidence ('Look! They serve Teriyaki!') and valuation ('Oh, you know, you always hate those crowded places.').

Agents capable of deliberation should be in epistemic contact with the normative import of the contents of their beliefs and desires. Deliberation requires an intelligibility related to the first-person perspective of a rule-follower. Deliberative agents must be able to exercise their care for the rationality of their beliefs—given what they have reason to believe to be true—as well as their care for the rationality of their desires—given what they have reason to value—and they must be able to decide what to do on the basis of the beliefs and desires they come to accept as rational. Deliberation is not something that looks like explaining and predicting what to do given the expected utility numbers one assigns to possible outcomes. In deliberation, one does not predict one's own actions. Of course, deliberation might involve the weighting of utility numbers, but if it does, it does not do so from the explaining and predictive standpoint of rational choice theory, but from the first-person standpoint of appreciating the normative import of these numbers as reasons on the basis of which one will choose to do what one does. This is so, even in those cases where habits seem to rule ('Yes, we always end up in this restaurant—but of course, we cannot discuss such predictions *as predictions*.').

If we are to think of choices and actions as events brought about by deliberating agents who care to be rational (i.e., who try to determine and follow the appropriate rules) and who ascribe meaning to what they do, we cannot do without *verstehen*. Although this does not rule out rational choice theory altogether, it does rule out the possibility of a completed social science based on merely an *erklären* approach.

Bibliography

Apel K O 1955 Das Verstehen. Ein Problemgeschichte als Begriffsgeschichte. *Archiv für Begriffsgeschichte* **1**: 142–99

Dallmayr F R, McCarthy T A 1977 *Understanding and Social Inquiry*. University of Notre Dame Press, Notre Dame, IN

Davidson D 1980 *Essays on Actions and Events*. Clarendon Press, Oxford, UK

Davidson D 1984 *Inquiries into Truth and Interpretation*. Clarendon Press, Oxford, UK

Dennett D 1973 Mechanism and responsibility. In: Honderich T (ed.) *Essays on Freedom of Action*. Routledge and Kegan Paul, London

Eells E 1982 *Rational Decision and Causality*. Cambridge University Press, Cambridge, UK

Gadamer H-G 1960 *Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik*. Mohr, Tübingen, Germany

Habermas J 1967 *Zur Logik der Sozialwissenschaften*. Mohr, Tübingen, Germany (1988 *On the Logic of the Social Sciences*. MIT, Cambridge, MA)

Henderson D K 1993 *Interpretation and Explanation in the Human Sciences*. State University of New York Press, Albany, NY

Hollis M 1994 *The Philosophy of Social Science*. Cambridge University Press, Cambridge, UK

Macdonald G, Pettit P 1981 *Semantics and Social Science*. Routledge and Kegan Paul, London

Malcolm N 1968 The conceivability of mechanism. *The Philosophical Review* **77**: 45–72

O'Hear A (ed.) 1996 *Verstehen and Humane Understanding*. Royal Institute of Philosophy Supplement **41**. Cambridge University Press, Cambridge, UK

Outhwaite W 1975 *Understanding Social Life*. Allen and Unwin, London

Pettit P 1991 Decision theory and folk psychology. In: Bacharach M, Hurley S (eds.) *Essays in the Foundations of Decision Theory*. Blackwell, Oxford, UK

Ricoeur P 1981 *Hermeneutics and the Human Sciences*. Cambridge University Press, Cambridge, UK

Riedel M 1978 *Verstehen oder Erklären? Zur Theorie und Geschichte der hermeneutischen Wissenschaften*. Klett-Cota, Stuttgart, Germany

Schutz A 1932 *Der Sinnhafte Aufbau der Sozialen Welt, eine Einleitung in die verstehende Soziologie*. Springer, Vienna (1967 *The Phenomenology of the Social World*. Heinemann, London)

Schutz A 1970 *Reflections on the Problem of Relevance*. Yale University Press, New Haven, CT

Stoutland F 1970 The logical connection argument. *American Philosophical Quarterly* **7**: 117–30

Taylor C 1964 *The Explanation of Behaviour*. Routledge and Kegan Paul, London

Taylor C 1985 *Philosophy and the Human Sciences*. Cambridge University Press, Cambridge, UK

Weber M 1922 *Gesammelte Aufsätze zur Wissenschaftslehre*. Mohr, Tübingen, Germany

Weber M 1949 *Max Weber on the Methodology of the Social Sciences*. Free Press, Glencoe, IL

Winch P 1958 *The Idea of a Social Science and its Relation to Philosophy*. Routledge and Kegan Paul, London

Wright G H von 1971 *Explanation and Understanding*. Routledge and Kegan Paul, London

J. Bransen

Vestibular System

1. Introduction

Throughout the evolution of life on the planet Earth the phenomena of light and gravity have played a compelling role in the development of many biological

mechanisms. While the intensity of light changes from day to night, gravity remains constant and omnipresent at all times. For any being on Earth it remains impossible to escape from its influence for more than a few seconds. Indeed such transitions into weightlessness, or free-fall, provide pleasant sensational experiences as when the baby is thrown up in the air, the child playing on the swing or the more courageous who test their limits while bungee jumping or sky diving.

The gravitational force, which is directed to the Earth's center and is equivalent to an acceleration with a magnitude of $9.81\,\mathrm{m\,s^{-2}}$ corresponding to the acceleration of a formula I racing car with 700–900 hp, represents a unidirectional force for all biological organisms, no matter how primitive, and thus provides the fundamental reference for 'spatial orientation.' For the neurophysiological circuitry of the sensory-motor system, gravity may be considered as a constant background offset or bias signal.

2. *Anatomy of the Vestibular Organs*

Any movement in three-dimensional space involves accelerational components that are compounded with gravity. In order to cope with such continual reorientation, each organism requires adequate sensory organ to transduce the magnitude and direction of the concomitant acceleration. In single cell organisms and plants, this organ is situated within the cell, whereas in higher organisms more sophisticated sensory organs have been developed, namely the vestibular organs in the inner ear.

The basic principle for the measurement of acceleration is that of a heavy body, lying either on a sensitive surface, for example, the endolymphatic reticulum of the cell or the inner cell membrane, or embedded in a gelatinous membrane coupled with

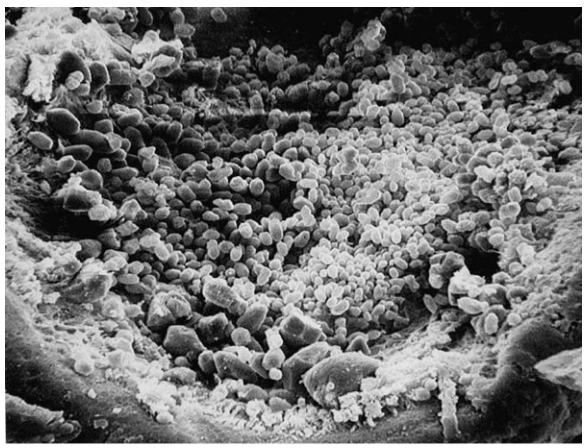

Figure 1
Otoconial layer in a otolith organ of man

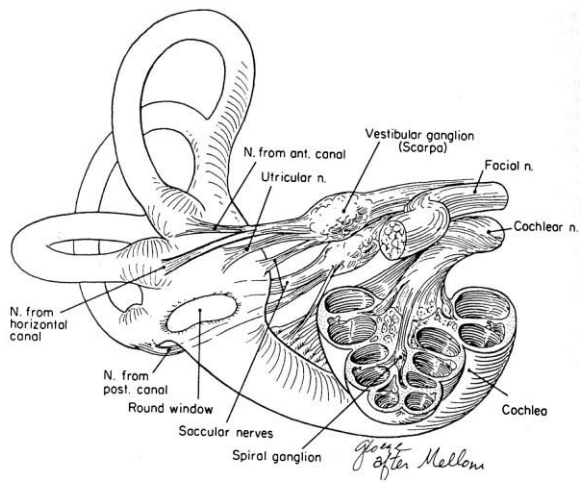

Figure 2
Schematic drawing of the inner ear of man with three semicircular canals, the two otolith organs and the cochlea (Baloh and Halmagyi 1996)

sensory hair bundles that are sensitive to flexion. The tip of a plant root possesses a relatively primitive sensory organ. This provides the root with information on the direction of gravity, enabling correct growth—usually deeper into the earth (a figure for such a type in a cress root can be found in Volkmann et al. 1986). More complex examples of sensory organs for the measurement of linear accelerations are the open systems, which employ foreign bodies (e.g., sand particles used by crabs), and sealed systems with self-made bodies, typically calcium carbonate crystal structures. In the fish these consist of single bodies (otoconia) in each organ, while birds and mammals usually have multiple particles (otoconia) embedded in a gelatinous membrane (Fig. 1).

While a sensory organ that signals linear acceleration (otolith organ) may be adequate for those organisms that move relatively slowly and in a linear fashion, for faster moving species and particularly those that utilise three-dimensional space the transduction of linear acceleration must be complemented with the measurement of rotation. Accordingly, in the vertebrates a system of three, approximately orthogonal, semicircular ducts or canals has evolved to signal rotations around any arbitrary axis (Fig. 2). These canals are filled with liquid, so-called endolymph, and are sealed in a domelike enlargement (the ampulla) by a membrane-like structure (the cupula) projecting from a sensory epithelium on a crest (crista ampullaris) into the cupular structure are a multitude of sensory hairs (Fig. 3). When the head is rotated the inertial moment is counteracted by the viscosity of the semicircular canal liquid, producing pressure on the cupula and thus causing the sensory hairs to be sheared in the opposite direction. This, in turn, results in

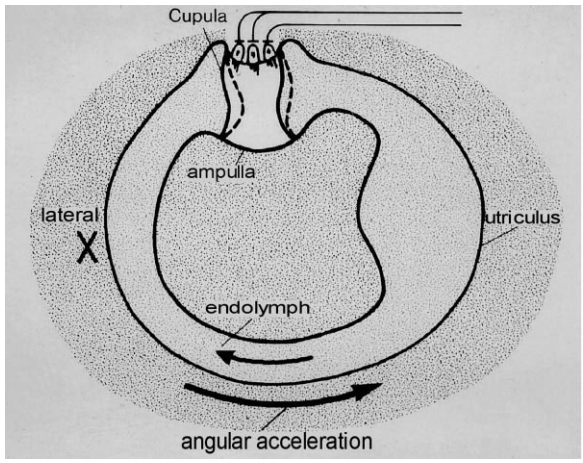

Figure 3
Schematic drawing of the ampulla and cupula of a semicircular duct

hyper- respectively hypoactivity of the sensory cell, according to the direction of hair deflection.

The human vestibular organs are embedded in a fluid-filled bony labyrinth in the temporal bone (Fig. 4) adjacent to the hearing organ, or cochlea. Common to both vestibular and hearing organs are two fluid compartments containing respectively the perilymph and the endolymph. The vestibular labyrinth consists of two organs for the transduction of linear acceleration, the macula utriculi and macula sacculi, and a set of three semicircular canals, namely the horizontal, anterior, and posterior, which function as sensors of rotation. In the upright position the macula utriculi is more or less aligned to transduce any linear acceleration in the horizontal plane while the macula sacculi is effectively aligned to the vertical plane.

The function of this part of the vestibular organ is not yet exactly defined; it may also be responsible for the sensory transduction of very low frequency sound waves. From an ontogenetic viewpoint, the cochlea is an evagination of the sacculus. In principle the sensory hair cells in the cochlea function identically to those of the vestibular organ, although the general structure is optimised for acoustic transduction.

The vestibular labyrinths in the right and left inner ear can be regarded as mirror images. While each labyrinth is able to measure all naturally occurring linear and angular accelerations, they are neurally interconnected to complement one another, to some extent analogous to a push-pull amplifier. While the head is stationary a steady signal of typically 80–100 impulses per second is present on the afferent nerve connecting the vestibular organ to the vestibular nuclei in the central nervous system. During a head rotation to the right, this impulse rate from the right labyrinth increases while the impulse rate from the left labyrinth

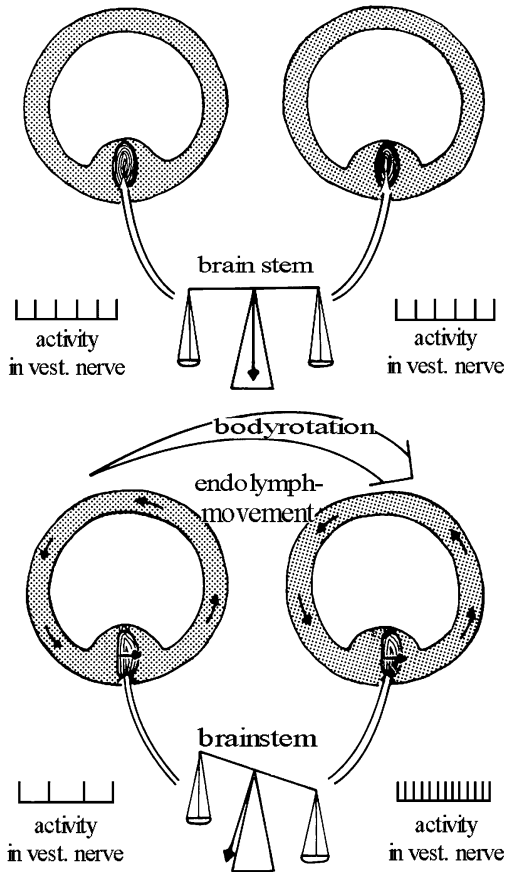

Figure 4
Mechanism of cupula stimulation and frequency modulation of signal transfer in the afferent nerve (push-pull) by rotation of the head (Scherer 1997)

decreases (i.e. pulse frequency modulation). This is illustrated in Fig. 4. This push-pull principle provides increased sensitivity and permits accurate bi-directional measurement.

Such information is transferred via the bipolar vestibular nerve to the central vestibular nuclei in the brainstem. The ganglion Scarpae of this nerve is situated within the inner auditory canal. Its peripheral branch synapses to the primary vestibular sensory cells, the central branch to the cells of the vestibular nuclei, where the information is relayed to many efferent fibers.

3. *Physiology of the Vestibular System*

The neural circuitry in the vestibular nuclei of the brainstem have three principal functions, namely the integration of vestibular information with that from

the visual and the proprioceptive systems related to motion in space, the stabilization of the visual field, and third, the stabilization of body posture and locomotion. All of these tasks function autonomously without the need of our conscious intervention. Only in those situations where there is a disorder in the sensory information, occurring through sensory mismatch or through disease, do we become aware by way of motion sickness or dizziness with symptoms such as ataxia or tendency to fall to one side, or disturbance of the visual field. In the following chapters the major physiological functions—and the associated dysfunctions—of the central vestibular system, and thereafter of the peripheral vestibular system, are described.

4. *Integration of Motion-related Sensory Information*

Several sensory systems—namely the visual, the vestibular, the proprioceptive, and the auditory—contribute to the maintenance of spatial orientation. In principle one can say that all stereo systems responding to motion in three-dimensional space are involved to some degree in the integrative process in the central vestibular nuclei. This multisensory flow of information is essential for coordinated motion in space. Afferent nerve fibres from each of the named sensory systems converge to the nerve cells in the vestibular nuclei. This means, for example, that motion perception in the vestibular area in the cortex (parieto-insular vestibular cortex, PIVC) is induced by real motion or by regarding a moving visual field.

One can play with this phenomenon by looking at a fast running river from a bridge, or by watching clouds on a windy day while lying on a meadow. It takes about 20 seconds until one has the impression that the water, respectively the clouds are standing still and the observer is moving. This effect is due to the fact that when different sensory information converge to a single cell there is no possibility at this level to distinguish a optical from a vestibular stimulation.

Visual, or optical, motion perception is mainly mediated from the peripheral retinal field. Thus, a stronger intensity is reached when the entire field-of-view is stimulated. For clinical testing a stripe pattern or other strongly contrasting presentation (e.g., large format projection of city skyline) moving at angular velocites of up to $90\,\mathrm{deg\,s^{-1}}$ is employed. This is known as 'optokinetic stimulation.'

In the human there is a second component of the optical system for motion perception which must be taken into account. This is mediated by the central area in the retina known as the fovea, which provides high-resolution focussed imaging. The afferent information from this area is relayed to the vestibular nuclei via separate pathways. Clinical stimulation of the fovea is performed with more slowly moving objects ($< 40\,\mathrm{deg\,s^{-1}}$), typically by means of visible point targets or a pendulum.

Brain diseases affecting any of the described systems generally produce disturbances in spatial perception or orientation.

Somatosensory perception of motion or gravity originates from receptors distributed throughout the body, even from the intestines. The neural information from these receptors is relayed either directly, or via mossy or climbing fibres to Purkinje cells in the cerebellum, and in turn via inhibitory connections, to the vestibular nuclei.

One feels gravity with the receptors of the plantar surface of the foot when standing upright, with the sensors in the back when lying down or when accelerated in a car. One can feel the change of gravity with the receptors in the intestinum, e.g. when a cable railway runs over a pillar.

The main sensory inputs converging to cells in the vestibular nuclei stem from the visual, vestibular and somatosensory receptors. When we move actively in our habitat all of these components are well matched and concur with a correct sensory perception. However, the situation can be very different during passive motion when the sensory inputs no longer necessarily match correctly, that is, a sensory mismatch occurs.

Such a sensory mismatch can be experienced by a car passenger who, while driving along a hilly and curved road, simultaneously reads a newspaper. The optical system will indicate that one is stationary (stable image of the newspaper) while the vestibular system, the somatosensory, and the acoustic systems will indicate the movement of the vehicle. The same situation occurs in an aircraft. The cabin compartment of the plane presents a stable optical environment for the passenger, but the vestibular organs transduce the plane's motion during curves and landing manoeuvres.

When such a sensory mismatch, or conflict, cannot be resolved quickly and correctly a crisis known as 'motion sickness' occurs, which is accompanied by a series of symptoms, namely paleness, tiredness with yawning, stomach awareness, outbreak of sweat, nausea, and vomiting.

The vestibular system uses nausea and vomiting as an emergency outlet instead of pain as is the case for all other sensory systems. The reason for this is unknown, but we do know that there would be no fairground machines if the vestibular system would elicit pain when being overstimulated.

Strictly speaking, to avoid motion sickness one must prevent the occurrence of the precipitative sensory mismatch. One should not make fast head movements while the whole body is subject to passive motion. The combination of fast active and passive motion is associated with the outbreak of motion sickness. There is also a wide interindividual variability in susceptibility to motion sickness within the same provocative situation.

J. W. v. Goethe was hypersensitive to motion sickness. He described his problems during his Italian journey on a ship cruise from Napoli to Palermo on the island of Sicily. He was very angry, not only because of his own motion sickness, but also about the fact that his companion was not affected and furthermore was enjoying his food.

One reason for this variability may lie in the difference in mass of otoconia between the right and left labyrinths, which lead to asymmetrical afferent input to the central vestibular nuclei.

Normally, during our active movement repertoire this difference is well-compensated on a neural level. However, during passive motion this asymmetry could potentially introduce a sensory mismatch and therefore play a role in the elicitation of motion sickness. While it has proved difficult to measure the effective mass of otoconia in mammals, it has been possible to measure the mass of the single otoconial stones in fish. In a measurement series performed with the trout, salmon, and xiphophoris Helleri right-to-left differences of up to 76 percent were found (Scherer et al. 1997).

In the case of a sensory conflict (mismatch) in the area of the central vestibular nuclei hierarchical structures of the brain come into play. As we know the visual system dominates over the vestibular and within the vestibular system the evolutionary older otolith organs dominate over the semicircular canals.

One can notice this hierarchical principle when sitting in a stationary train in the railway station. When a neighbouring train starts to move one immediately gets the impression that one's own train has started to move and that the other train is stationary. In this case of different sensory input (optical: moving; vestibular and somatosensory: stationary) the dominant visual system determines the percept.

5. *Stabilization of Upright Body Posture*

Together with the cerebellum, the vestibular system contributes substantially to the stabilisation of body posture during locomotion. The vestibulo-spinal tract emerges from the lateral vestibular nucleus in the brainstem and runs unilaterally to the spinal motoneurons of the extensor muscles (Fig. 5). When a person turns to the right, the stimulus to the right vestibular organ leads to an increase of the afferent firing rate on the nerve fibres leading to the vestibular nuclei; simultaneously the firing rate on the nerve fibres from the left vestibular organ decreases. At the same time the visual system signals the rotation of the surrounding, that is, to the left during a head turn to the right. The afferent signal from the retina converges with that from the vestibular organs and enhances their effect, that is, stimulating the right vestibular nuclei positively and the left negatively. Via the vestibulo-spinal tract the tonic signal to the extensor

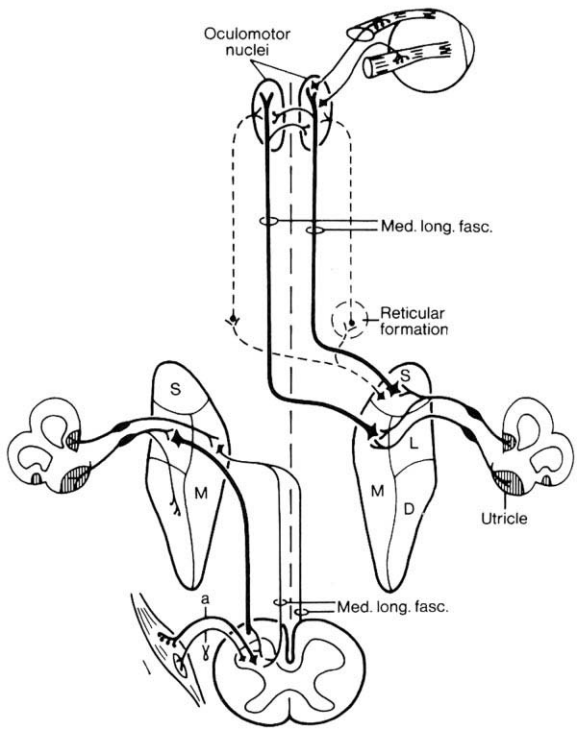

Figure 5
Simplified diagram of the two reflex arcs involving the vestibular nuclei in the brain stem, the vestibulo-spinal and the vestibulo-ocular reflex (Brodal 1981). S, D, L, M = superior, descending, lateral, and medial nucleus, respectively

muscles on the right is increased, which has the effect of counteracting the increased body mass on this side during the turn. At the same time there is a corollary increase of tonus to the left flexor muscles.

When the function of one vestibular labyrinth (right or left) is suddenly reduced due to an accident, an infection, or an operation the person feels a tendency to fall to the injured, or diseased, side. However, this symptom lasts only for a number of days, even in those cases where the function of the diseased side does not recover. The reason for this recovery and for the similar reduction of other symptoms lies with the powerful self-reparatory capability, or adaptability, of the vestibular system. A major therapeutic approach to vestibular deficits is based on self-training. In the first weeks after the sudden loss of vestibular function the patient undergoes training involving general natural body movements; in the following weeks specific training of impaired functions and symptoms is necessary.

The phenomenon of height vertigo can also be explained within the context of sensory mismatch or conflict. This 'physiological vertigo syndrome' (Brandt

1991) occurs when a person stands on the top of a high tower or on the brink of a cliff. In every normal person a certain degree of random body sway can be measured during upright stance. This body sway is regulated by the multisensory convergence in the vestibular nuclei. In this context spatial vision plays a major role, particularly the perceptual relationship between foreground to background. While standing on the edge of a high tower or cliff the relative motion of the background—now located many metres distant in comparison to normal—is perceived visually as much greater. This discrepancy in visual perception destabilises the regulatory mechanism for postural stability and body sway increases. On continued exposure this 'positive feedback' can lead to serious instability. How this fundamentally physiological response is processed cognitively and psychologically can vary enormously among individuals, in some people leading to severe height anxiety with vegetative symptoms. Such physiological instability can be counteracted immediately by stepping back from the brink so that the discrepancy is resolved by viewing a normal foreground at foot level, or by providing more contact to the ground by grasping a handle or rail, sitting down or kneeling. Extreme forms of height vertigo can be cured by behavioural therapy.

J. W. v. Goethe treated his height vertigo by repeatedly climbing up the tower of Strasburg's Gothic cathedral and standing on exposed platforms (similar to contemporary 'flooding' therapy). Another kind of training of height anxiety is shown in Alfred Hitchcock's film *Psycho*.

6. *Stabilization of the Visual Field During Locomotion*

The stabilization of eye position in the head during locomotion is supported by the so-called vestibulo-ocular reflex. Afferent nerve fibres project from the ipsilateral superior and the contralateral medial nuclei upwards via the medial longitudinal fasciculus (MLF) to the oculomotor nuclei, which in turn govern the input to the motor neurons driving the eye muscles (Fig. 6). This reflex is described here briefly since it plays a major role in the diagnosis of vestibular disorders.

Again taking the example of a rotation to the right, the surrounding visual field rotates too the left. If the eyes were to remain fixed in the head during this rotation, the image of the surroundings would slip on the retina, causing a blurring of perception until the head was stationary. This would obviously deteriorate spatial orientation. A major purpose of the vestibulo-ocular reflex is to rotate the eyes compensatorily to the head movement, thus retaining a focussed image on the retina. Under normal conditions the vestibulo-ocular reflex facilitates compensatory eye movement equivalent to the concomitant head movement, regardless about which axis the head is rotated. In this manner, head movement in three-dimensional space is compensated for by appropriate reflex eye movement. However, the mechanical range of eye rotation is limited anatomically by the extraocular muscle and optical nerve structures. Thus, during a full head- and body-rotation, eye position is compensatory, but is regularly reset by means of a quick phase in the anticompensatory direction. This leads to a sawtooth pattern of eye movement, known as 'nystagmus,' during full body rotation. The nystagmus eye movement pattern is defined by the sawtooth pattern consisting of the vestibularly induced slow phase and the quick reset movement (Fig. 6). Such nystagmus patterns of eye movement are elicited by natural movements or by unilateral or bilateral stimulation of the vestibular organs. A similar pattern, known as optokinetic nystagmus, can also be elicited by rotation or translation of the visual field while the subject is held stationary.

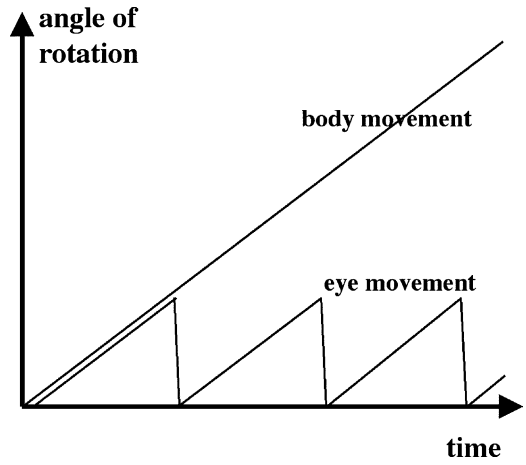

Figure 6
Body movement (rotation) and compensatory sawtooth eye movement (nystagmus)

Optokinetic nystagmus can be seen easily when a person looks out of a train's window or out of the side window of a car. The moving optical pattern causes a slow deviation of the eye in the direction opposite to the train's motion with a fast phase in the direction of the train's motion.

Since the visual system is hierarchically superior to the vestibular system, visual fixation of a target point suppresses these nystagmus responses. A practical consequence is the fact that we are able to read during active or passive motion, although this may lead to a sensory mismatch which may cause motion sickness in sensitive individuals. On the other hand we have to prevent fixation in order to be able to measure nystagmus. This is done usually by means of thick lenses in a goggle (Frenzel glasses), by electric or

magnetic nystagmography or by videoimaging of the eyes in infrared light.

Observation and measurement of eye movements represents an important inroad for the examination of the function and dysfunction of the vestibular system, and therefore provides an important diagnostic tool for the clinician who has to deal with patients suffering from dizziness)

7. *Encyclopaedia of Typical Vestibular Disturbances*

7.1 *Peripheral Disturbances*

(a) Sudden unilateral loss of function of the vestibular organ.

Pathophysiology: unknown, virus infections of Scarpae's ganglion, disturbed blood supply, trophic disorders of the cupula in the ampullae of the semicircular canals (Helling et al. 2001, 2000) are discussed.

Symptoms: heavy turning sensations in the first days, the symptoms decline gradually.

(b) Meniere's disease.

Pathophysiology: osmotic disturbance within the two components of inner ear fluids, leading to a hydrops of the endolymphatic space and a sudden poisoning of nerve fibers with potassium.

Symptoms: sudden attacks of severe vertigo and vomiting in combination with low tone hearing loss and tinnitus in the affected ear.

(c) Benign paroxysmal positioning vertigo (BPPN).

Pathophysiology: aberrated otoconia from the otolith organs lying in the vertical semicircular canal initiating a unphysiologically endolyph wave.

Symptoms: heavy torsional sensation, which lasts for 10 to 20 seconds after performing a fast body motion.

(d) Acoustic neuroma.

Pathophysiology: benign tumor growing in the vestibular part of the eight cranial nerve in the inner auditory canal.

Symptoms: unilateral, slowly intensifying loss of vestibular and acoustic function. While the vestibular defect usually is well compensated and asymptomatic, the acoustic defect brings the patient to the doctor.

(e) Psychogenic (phobic) vertigo.

Vertigo is a frequent symptom of psychiatric illness (Brandt 1991). Alternatively, a transient labyrinthine dysfunction may initiate development of neurosis in which anxiety, panic attacks and depression are most common, and which typically preserves subjective vertigo and postural imbalance (Brandt 1991).

7.2 *Central Vestibular Disorders*

A lot of central disorders induce vertigo or postural instability. Most common disorders are mentioned (for details see Brandt 1991)

(a) Basilar artery migraine
(b) Vestibular epilepsy
(c) Down- or upbeat vertigo syndromes
(d) Central positional vertigo
(e) Posttraumatic vertigo (central or peripheral)
(f) Familial vertigo
(g) Drug vertigo (central or peripheral).

See also: Audition and Hearing; Auditory System; Binocular Space Perception Models; Eye Movements and Perception; Motion Perception Models; Navigation in Virtual Space: Psychological and Neural Aspects; Vestibulo-ocular Reflex, Adaptation of the; Visual Space, Geometry of

Bibliography

Baloh R W, Halmagyi G M 1996 *Disorders of the Vestibular System*. Oxford University Press, Oxford, UK

Brandt T 1991 *Vertigo*. Springer, New York

Brodal 1981 *Neurological Anatomy in Relation to Clinical Medicine*. Oxford University Press, New York

Goebel J A 2001 *Practical Management of the Dizzy Patient*. Lippincott, Williams, Wilkins, Philadelphia

Helling K, Clarke A H, Watanabe N, Scherer H Investigation on the morphology of the cupula in semicircular canal ampulla. *HNO* **48**: 822–7

Helling K, Watanabe N, Jijawa H, Mizuno Y, Watanabe S, Scherer H in press Altered cupular mechanics—A cause for peripheral vestibular disorders? *Acta Otolaryngology*

Scherer H 1997 *Das Gleichgewicht*.. Springer, New York

Scherer H, Helling K, Hausmann S, Clarke A H 1997 On the origin of interindividual susceptibility to motion sickness. *Acta Otolaryngology* **117**: 149–53

Sharpe J A, Barber H O 1993 *Vestibulo-ocular reflex and vertigo*. Raven, New York

Volkmann D, Behrens H M, Sievers A 1986 Development and gravity sensing of cress roots under Microgravity. *Naturwissenschaften* **73**: 438–44

H. Scherer and A. H. Clarke

Vestibulo-ocular Reflex, Adaptation of the

The vestibulo-ocular reflex (VOR) produces eye movements in the direction opposite to head movement, thus serving to automatically stabilize vision relative to space. VOR displays a remarkable capability of adaptation. It modifies its dynamics so as to secure visual stability under any circumstance in which mismatched movements of the head and the visual surroundings impair visual stability. The major site of this adaptation has been shown to reside in evolutionarily the oldest area of the cerebellum, namely

the flocculus. In recent years, VOR has been studied extensively as a model system for investigating the learning mechanisms in the cerebellum.

1. *VOR*

The reflex arc for the VOR is a three-neuron pathway connecting the primary vestibular afferents arising from the semicircular canals, to the VOR relay neurons in vestibular nuclei, and finally to motoneurons innervating the extraocular muscles. VOR relay neurons are either excitatory or inhibitory in their synaptic action, and induce either contraction or relaxation of the extraocular muscles via excitation or inhibition of motoneurons. There are a number of parallel pathways connecting the three semicircular canals (horizontal, anterior, and posterior) and the two otolith organs (saccule and utricle) in each labyrinth to the six extraocular muscles in each eye (medial and lateral rectus, superior and inferior rectus, superior and inferior oblique). These pathways operate in concert under conditions of free head movement, but in an experimental setup they are stimulated separately by giving yaw, pitch, roll or linear acceleration.

The horizontal VOR is tested by sinusoidal or velocity-step head rotation on the horizontal plane, and the VOR gain is measured as the ratio of the attained eye velocity to the applied head velocity. Sinusoidal rotation is convenient for measuring the gain and phase of the VOR separately, while velocity steps enable us to separate the VOR responses, which arise with different latencies, into their constituent components. Measurement of the VOR is performed in the dark or with the eyes closed while the whole body is rotating.

2. *VOR Adaptation*

Long-term visuovestibular mismatching (for days to months) can be imposed in a subject (either animal or human) with the aid of a pair of prism or lens goggles. Short-term mismatching (1–4 hours) is created by combined rotation of the turntable on which the subject is positioned and a screen representing the visual surroundings. Both wearing of Dove prism goggles, which reverses the right–left axis of the visual field, or miniaturizing lenses which reduce the visual field, and in-phase rotation of the turntable and screen in the same direction, cause adaptive reduction of the horizontal VOR gain. Likewise, both wearing of 2x magnifying lenses and out-of-phase rotation of the turntable and screen in opposite directions cause adaptive increase of the horizontal VOR gain. These adaptive changes in VOR gain act to reduce the retinal slips that occur during the VOR, which otherwise are augmented by the imposed visuovestibular mismatching.

A cross-modal adaptation occurs between the horizontal and vertical VORs; when yaw rotation on the horizontal turntable is combined with vertical optokinetic motion, the VOR measured in the dark acquires a vertical component even when the head is rotated only horizontally.

3. *The Flocculus*

In rabbits and cats, the flocculus has five to six major folia, while in rats it usually has only one folium. In classic anatomical studies of the monkey cerebellum, ten folia have been described in the flocculus, but recent studies on neuronal connectivity have revealed that the rostral five folia belong to the ventral paraflocculus and not to the flocculus.

A proportion of Purkinje cells in the flocculus of either side supply inhibitory synapses to both excitatory and inhibitory relay neurons in the VOR arc arising from the ipsilateral semicircular canals. Purkinje cells related to the horizontal VOR arc are localized to a narrow zone (H-zone, about 1 mm across) in the flocculus. The H-zone is flanked by other zones related to vertical (V-zone) or rotatory (R-zone) VORs, and a simple method to identify these zones in physiological experiments is to apply local electrical stimulation to these zones, which induces distinct horizontal, vertical, or rotatory eye movements.

These microzones in the flocculus are also characterized by their climbing fiber inputs relayed by the inferior olive. Purkinje cells in the H-zone receive climbing fiber signals encoding retinal errors, that is, slips of the retinal images on the horizontal plane. Purkinje cells reported to receive climbing fiber signals encoding retinal slips approximately in the vertical and rotatory directions, may be located in the V- and R-zones, respectively.

Branches of mossy fibers arising from the semicircular canals are distributed throughout the flocculus. Canal signals could also be relayed to the flocculus via 4–6 mm-long parallel fiber axons of granule cells. Yet, in the H-zone Purkinje cells, simple spike discharges are modulated specifically in relation to the horizontal head velocity (see below).

4. *Flocculus Hypothesis for VOR Adaptation*

Based upon the abovementioned relationships between the flocculus and the VOR, it is suggested that the H-zone originally receives mossy fiber information from all directions, and through learning by referring to climbing fiber signals, canal-plane-specific simple spike responses are established. It is further suggested that the established simple spike response patterns are updated daily by learning, whenever retinal errors are sensed through the climbing fibers.

Learning in the flocculus would occur at the synapses formed between the axons of granule cells

relaying mossy fiber signals and Purkinje cells. This type of interaction was suggested in Marr's and Albus' theories of the cerebellar circuit proposed around 1970, later identified as long-term depression (LTD). LTD occurs in granule cell axon-to-Purkinje cell synapses, when these are conjointly activated with climbing fibers. These hypothetical views have been tested experimentally in recent decades, as reviewed below.

5. *Deprivation of Flocculus Function*

That the flocculus plays a role in VOR adaptation has been confirmed by showing that impairment of flocculus function results in abolition of VOR adaptation. Previously, surgical ablation of and injection of toxic amino acids in the flocculus were used. More recently, based upon newly acquired knowledge about LTD, various pharmacological reagents and genetic means have been used to block LTD in the flocculus. Injection of hemoglobin into the subdural space over the flocculus in rabbits and monkeys abolished VOR adaptation, presumably due to the scavenging action of hemoglobin on nitric oxide (NO) that is required for the induction of LTD. Injection of an inhibitor of NO synthase into goldfish cerebellum inhibited adaptive increase of VOR gain. Mice transfected with the gene of a pseudosubstrate inhibitor of protein kinase C (PKC), did not exhibit VOR adaptation, as expected from the well-known requirement of PKC for LTD induction.

6. *Behavior of Flocculus Purkinje Cells*

Whether the behavior of H-zone Purkinje cells is consistent with the flocculus hypothesis has also been examined by recording two types of spikes generated by the Purkinje cells: namely, simple spikes, reflecting their mossy fiber inputs, and complex spikes, evoked by climbing fiber inputs.

During sinusoidal head rotation, VOR relay neurons are driven by signals from the ipsilateral horizontal canal, which are modulated in phase with the head velocity (increased during ipsilateral rotation). In the H-zone, Purkinje cells exhibit modulation of simple spikes either in phase or out of phase (decreased during ipsilateral rotation). The in-phase-modulated inhibitory signals of Purkinje cells should depress VOR by canceling the effect of similarly modulated canal signals. On the other hand, the out-of-phase discharges of Purkinje cells should enhance VOR by adding inversely modulated inhibitory signals to the canal signals.

Under the condition that sustained visuovestibular mismatching enhances the VOR, the simple spike responses of H-zone Purkinje cells to head rotation became predominantly out of phase. Conversely, under the condition that sustained mismatching depresses the VOR, the simple spike response became predominantly in phase. These changes in simple spike responses can be interpreted as being causally related to the VOR gain changes.

The two types of simple spike discharges in H-zone Purkinje cells may be induced by mossy fiber inputs from the horizontal canals of the two sides. Signals from the ipsilateral canal would drive Purkinje cells in phase, while signals from the contralateral canal would drive Purkinje cells out of phase. The shift of dominance of the type of simple spike modulation during VOR adaptation can be explained as being caused by LTD, depressing either the ipsilateral or contralateral canal inputs to the Purkinje cells. The climbing fiber discharge patterns observed in H-zone Purkinje cells are consistent with this notion.

7. *Memory Site for VOR Adaptation*

No doubt, the flocculus plays a crucial role in the induction of VOR adaptation. However, there is a dichotomy of opinions regarding its role in retention of the memory of VOR adaptation. To determine whether or not the memory of VOR adaptation is retained in the flocculus, studies have been carried out on the effects of blocking of flocculus functions—either surgically or functionally—after VOR adaptation develops. In goldfish, no less than 30 percent of the altered VOR gain was still maintained after ablation of the cerebellum, suggesting that the cerebellum is the major, but not the sole, site responsible for the maintenance of VOR adaptation. Microdialysis of lidocaine into goldfish cerebellum, however, abolished both the induction and short-term maintenance of VOR adaptation, suggesting that the memory for VOR adaptation is retained at least for a few hours in the cerebellum.

Another strategy to determine whether the memory underlying VOR adaptation is retained in the brainstem or in the cerebellum is to analyze the component eye movements elicited by velocity-step head stimuli at different latencies. The measurements in goldfish indicated that adaptive changes occurred in the earliest component of the eye movement elicited by a stimulus with a latency of 18 milliseconds, which could be mediated via the three-neuron VOR arc. However, the measurements in monkeys remain inconclusive.

8. *Oculomotor Signals in the Flocculus*

A complication in the flocculus hypothesis arises from the possible existence of mossy fibers that relay information regarding oculomotor activities to the flocculus. The existence of such mossy fibers has frequently been assumed in modeling the flocculus–VOR system (Lisberger and Sejnowski 1992).

However, it is difficult to accurately estimate such mossy fiber signals in isolation from the vestibular or

retinal mossy fiber signals. In the previously used visual suppression method, the oculomotor signals were eliminated by training a monkey to fix its gaze on a spot moving with the head, and the resultant changes induced in the Purkinje cell behavior are assumed to represent the oculomotor signals. However, such gaze fixation is executed by central command signals, which would be a source of another mossy fiber input to the flocculus. Oculomotor signals in the flocculus have also been measured during nystagmic eye movements evoked by the stimulation of the optic tract. This measurement yielded a much smaller estimate of the oculomotor signals to the flocculus as compared with the visual fixation method.

It is notable that the simple spike responses of H-zone Purkinje cells to head rotation in the dark and those to screen rotation that evokes optokinetic eye movement responses varied independently of each other. Since oculomotor components should be shared between the two types of simple spike responses, this observation in rabbits favors the view that H-zone Purkinje cells receive few oculomotor signals.

The possibility so far overlooked is that learning forms simple spike responses reflecting oculomotor activities. Since retinal errors represent the deviations of actual eye movements from the ideal eye movements fully compensating for head movement, these should contain information regarding eye movements, which could be reflected in simple spike responses through the LTD mechanism.

See also: Cerebellum: Associative Learning; Cerebellum: Cognitive Functions; Eye Movement, Control of (Oculomotor Control); Long-term Depression (Cerebellum); Vestibular System

Bibliography

Ito M 1998 Cerebellar learning in the vestibulo-ocular reflex. *Trends in Cognitive Science* **2**: 313–21

Ito M 2001 Cerebellar long-term depression: Characterisation, signal transduction and functional rates. *Physiological Reviews* **81**: 1143–95

Lisberger S G, Sejnowski T J 1992 Motor learning in a recurrent network model based on the vestibulo-ocular reflex. *Nature* **360**: 159–61

M. Ito

Vietnamese Revolution, The

In one and a quarter century since the French conquest of southern Vietnam in the 1860s, Vietnam underwent a radical transformation from a Confucian monarchy to a Marxist–Leninist state. The Vietnamese revolution involved a violent clash with two major Western powers, France and the United States, as well as a fundamental transformation of the Vietnamese socio-economic system.

1. Historical Background

Citing the Vietnamese monarchy's persecution of Catholics and refusal of a trade request, France launched a series of attacks on Vietnam from 1858 onwards. By 1867, France had seized as a direct colony the frontier southern third of Vietnam (Cochin China or *Nam bo*) from the Vietnamese monarchy. By 1884, the northern and central Vietnam (respectively Tonkin or *Bac bo*, and Annam or *Trung bo*) had fallen under French control, although the Nguyen monarchy nominally still retained its direct administration of the latter.

In the aftermath of the French conquest, the resistance to French colonialism, mainly in the name of an 'Aid-the-King' movement, was led primarily by Confucian literati. It lasted the longest in central and northern Vietnam. By the time the Union of French Indochina was established in 1897, with Laos and Cambodia as the remaining two of the five components, the armed anticolonial resistance led by Confucian literati had virtually collapsed (see Marr 1971).

Many members of the Vietnamese elite collaborated with the French rulers, while a number of others searched for a more effective path of resistance, strongly influenced by Chinese reformist and revolutionary thoughts and inspired by the rise of Japan as an Asian power. At one end of the collaboration–resistance spectrum in the first quarter of the twentieth century was the Confucian scholar Phan Boi Chau, a native of northern Central Vietnam, who organized a movement of students to Japan and inspired isolated armed resistance against French colonialism (Marr 1971). In the context of a strong French repression, Thailand and southern China served as important bases for exiled resistance leaders (Goscha 1999). At the other end were many collaborators, such as the northern neo-traditionalist Pham Quynh, as well as many southerners who were rewarded with large tracts of land sold at low prices as the cultivated acreage in Cochin China increased five times from 1870 to 1920 thanks to the French water control work. Many of the latter acquired French citizenship, experimented with the more open politics in the direct French colony of Cochin China, or even played a role in the southern millenarian movements of Cao Dai and Hoa Hao (Tai 1983, 1992), co-opting many poor peasants in the sharply stratified Cochin Chinese society.

With the last Confucian examination being held in 1919, a new Western-educated elite began playing a salient role in the anticolonial resistance from 1925 onwards. The major uprisings took place in northern Vietnam in 1930 under the leadership of the Vietnamese Nationalist Party (VNP), and in central and

northern Vietnam in 1930–1 under the Indochinese Communist Party (ICP). The ICP-led movement in 1930–1 represented the most serious challenge to French colonialism since the Aid-the-King movement of the nineteenth century, as it involved the participation of half a million Vietnamese and threatened the colonial order in 25 provinces. At the peak of this movement, the colonial administrative apparatus virtually collapsed in many villages in the provinces of Nghe-An (birthplace of Ho Chi Minh) and Ha Tinh in northern central Vietnam. The ICP local leadership spontaneously seized village administrations, abolished taxes, reduced rents, and redistributed up to 8,500 acres of communal land in these two provinces alone.

In the short run, both the VNP and ICP-led movements in 1930–1 collapsed under the strong repressive measures by the French colonial regime with its superior military forces. Many revolutionary leaders were imprisoned, while Thailand and southern China again became important bases for exiled political activists, including Ho Chi Minh. It was not until the release of a number of political prisoners and the introduction of a more liberal political atmosphere in colonial Indochina under the Popular Front regime in the metropolis (1936–9) that revolutionary activism gained a significant momentum. In the context of the Japanese military advance into Indochina in September 1940 and the French colonial government's subsequent collaboration with Japanese military authorities, the Vietminh (abbreviation of *Viet Nam Doc Lap Dong Minh Hoi*, or Vietnam Independence Alliance League) under Ho Chi Minh began building a base of operation in northern mountainous areas, working with many members of ethnic minorities, against both the Japanese and Axis-aligned French colonial administration (Marr 1995, Chap. 3).

2. *Wars and Vietnamese Revolutionary Momentum*

The major turning point was 1945, when French colonial power was humiliated by the Japanese supported coup of March 9, 1945, and when in the context of a power vacuum with the Japanese surrender to Allied forces in August 1945, the Communist-led Vietminh organized the seizure of power throughout the country. The last Vietnamese monarch abdicated under revolutionary pressure. Ho Chi Minh's Declaration of Independence from France in September 1945 evoked a powerful nationalist sentiment from all social strata, and garnered widespread support for the first independent government of Vietnam (Marr 1995, 537–9). In the Vietnamese Marxist orthodox discourse, this seizure of power has been referred to as the August Revolution (*cach mang thang tam*). It was a revolution only in the narrowest sense of the term, that means, as a seizure of administrative power and official marking of the end of the monarchical era. However, even the maintenance of this administrative power in a territorially unified Vietnam came under a serious challenge by France's desire to regain its influence and control of Vietnam and the complications of post-WWII international politics with the post-WWII rise of US power and the Cold War. France's attempt to regain control of Vietnam and its underestimation of the support for Vietnamese independence led to the Franco-Vietnamese war from 1946 to 1954. The US's greater concern with Europe and Soviet influence after WWII, in combination with the Chinese communists' control of mainland China in 1949, shifted US policy towards a support for France and away from Franklin Roosevelt's support for the aspiration for independence of colonized peoples in Asia and Africa. The violent Franco-Vietnamese war ended in 1954, after the Vietminh defeated the French at Dien Bien Phu in the northern highlands. Vietnam was temporarily divided into two parts at the 17th parallel, with the north under Communist control, and the south under American influence.

Ho Chi Minh's vision of a unified Vietnam was not achieved until 1975, after a long and violent war which, in the context of the Cold War, involved a massive US ground involvement in South Vietnam and US bombing of the north and parts of Laos and Cambodia. The Vietnamese Communist Party began lending support to the armed resistance against the US-allied government in Saigon in 1959, helped to establish the National Liberation Front of South Vietnam in 1960, and increased the flow of personnel and arms from the north to the south from 1960 onwards (Cao van Luong et al. 1995, pp. 285ff.). The United States responded by increasing military advisers in South Vietnam to 15,000 by 1963, bombing North Vietnam from 1964 onwards, and introducing ground combat troops to South Vietnam in 1965. At its peak, the US had half a million of combat troops in South Vietnam. In the context of the Cold War, North Vietnam in turn received significant military and economic aid from the Soviet Union, China, Eastern Europe, and Cuba. The intense warfare led to the migration of millions of peasants to urban areas in South Vietnam, seriously damaged the South Vietnamese economy, and increased the dependence of the US-supported government in South Vietnam on US aid.

In the US, the draft system and the protracted warfare in Vietnam, exposed daily in the mass media, intensified the antiwar movement, especially among the young. This contributed to the US policy of handling the war to South Vietnamese troops, and to the withdrawal of US ground troops and the end of bombing of North Vietnam by 1973 in the aftermath of the Paris Peace Agreement. South Vietnam enjoyed no peace, as the Paris Agreement was violated by both

the Saigon-government troops on the one hand and the North Vietnamese troops and their southern allies on the other. The US-supported government in South Vietnam eventually collapsed in April 1975, under the attack of North Vietnamese troops and their southern allies.

Vietnamese Marxists' seizure of political control in the north by 1954 and in the entire country by 1975 strengthened the foundation for the enactment of their socioeconomic vision. In North Vietnam, the socioeconomic revolution accelerated throughout the 1950s under the Communist vision of a collectivized economy where the state played the dominant role in the industrial and agricultural production process, as well as in the distribution of raw materials and finished products. By 1965, the northern economy was based mostly on the state and cooperative sectors, and sustained by aid from the socialist bloc, as North Vietnam became directly involved in the war against the US and the American-supported government in Saigon. The Communist party's economic vision was imposed on the south after 1975. The economic crisis in the late 1970s, fueled by the resistance of peasants in the southern Mekong Delta to collectivization and the disenchantment of northern cooperative farmers, led to a gradual shift towards a market economy. This process of economic reforms (*doi moi*) started in agriculture in 1981. The economic reform was reaffirmed at the 1986 Communist Party Congress, and implemented in the rest of the economy by 1989. Cultivators consequently have long-term leases on land, and regain control over production and distribution processes. Private enterprises have been allowed to reemerge in industries and commerce, and foreign direct investments since 1988 have helped to fuel a generally strong economic growth in the 1990s.

3. *The Vietnamese Revolution: Theoretical Debates*

Many an important theoretical model on revolution has been constructed partly on the basis of the scholarly knowledge of the Vietnamese revolution. These models are embedded in the major traditions of contemporary Western social theory (John Stuart Mill, Karl Marx, and Emile Durkheim).

In the literature on revolution and agrarian unrest, the political scientist S. Popkin's rational choice model offers a good example of a theoretical model within the Millean tradition of inquiry (Popkin 1979). It considers the behavior of both the peasantry and the elite as the goal-directed acts of self-interested individuals. Popkin argues that modern revolutionary movements such as the Vietminh succeed because, as political entrepreneurs, their leaders can offer peasants concrete welfare improvements and effective organization against the marauding lords and notables of the old order. He emphatically rejects the view that modern agrarian movements involve defensive reactions to the violation of pre-capitalist normative structures by the colonial and capitalist order. In Popkin's argument, the incorporation of indigenous agriculture into the capitalistic world market, the establishment of the colonial regime, and the expansion of state power are not necessarily deleterious to peasant welfare. According to Popkin, colonialism brings vital stability and an improved communication system which keeps prices from fluctuating widely and thus keeps peasants alive in times of local famine.

Intriguing as it is, Popkin's model encounters its share of empirical anomalies in the Vietnamese case. For example, on which matrix of material costs and benefits did a large number of Vietnamese from different social strata volunteer for the Vietminh armed forces in the 1946–50 period, when Vietminh troops faced an enemy of overwhelmingly technological superiority, and before the poor recruits could foresee any substantial socioeconomic reforms in their favor?

In contrast to the Millean tradition of inquiry, the Marxist tradition seeks to situate economic transformation and revolutionary movements within the structure of conflict-ridden class relations. Both Western and Vietnamese Marxist analysts consider Western capitalist expansion as the root of the Vietnamese revolution. In this perspective, French colonialism in Vietnam as the representation of Western capitalist interests sought the maximal return on capital at the expense of the native working classes. The sociologist Martin Murray argues that capitalist exploitative measures included the concession of indigenous land to colonial settlers for the development of major cash crops; the introduction of direct and indirect taxes, and the conversion of tax payments from kind to cash to facilitate capitalist growth; and the oppressive appropriation of labor through the corvee system, the maintenance of rural labor reserves in Tonkin and Annam, and repressive labor laws to hold down labor costs for capitalist agro-mineral ventures.

Murray suggests that the impact of world capitalism was far from uniform in Vietnam. It was most striking in Cochin China which emerged as one of the top three rice exporting areas in the world, where approximately two-thirds of the rural population were separated from the means of agricultural production, where land and labor became strictly commodified, and where classes became increasingly differentiated. In contrast, in the north and center (Tonkin and Annam), although the rural population was increasingly drawn into market relations, the majority still tenaciously held on to their minuscule land holdings, buffered in many communities by their small shares of communal land. In North and Central Vietnam, the proportions of inalienable communal land averaged respectively 21 per cent and 25 per cent of the cultivated acreages.

Murray suggests that the persistence of communal land in the north and the center contributed to the formation of non-capitalist labor reserves for the plantations, mines, and factories within French Indochina and beyond, reserves where workers could rely on kinship and communal ties for survival in times of recession and during the pre-productive and post-productive periods (youth and old age). It has been suggested that the cost of capitalist production was consequently held down because capitalist firms had to pay only the cost of short-term labor reproduction (the cost to sustain workers physically during the period of employment), and not the cost of long-term labor renewal (the cost of replacing a physically debilitated work force with a new generation of laborers and of maintaining a temporarily idle work force during economic downturns). In the long run, Murray suggests, '[t]hese 'blockages' to the homogeneous spread and deepening of the extended reproduction of capital coincided with economic stagnation, declining productivity, and the vicious cycle of rural poverty associated with the 'underdeveloped' regions of the globe' (Murray 1980, p. x).

In the context of the capitalist world system, Marxist analysts consider the proletariat to be either largely located in the periphery of world capitalism, or composed of 'ethnicities' originating in the periphery; the bourgeoisie is heavily concentrated in, although by no means restricted to, the Western core. Within the Marxist world-system framework, class conflict takes on an international dimension, and revolution is seen as originating in the periphery where proletarian class interests emerge most clearly.

Within the same Marxist framework, the sociologist Jeffrey Paige suggests that 'The strength of colonial and imperial political controls long prevented the political expression of these conflicts, but with the decline of colonial power in the postwar era, the commercial export sectors of the underdeveloped world have become centers of revolutionary social movements' (Paige 1975, p. 3). Assuming inherent conflicts between agricultural producers and the dominant class, Paige attempts to construct an overall model of the forms which class conflict takes (Paige 1975, 1983). Paige deduces these actors' behavior patterns and the form of agrarian unrest from their income sources. More specifically, he proposes, on the one hand, that the economic base of the landed dominant class is narrower, more static, and less efficient than that of its commercial and industrial counterparts. The former's economic weakness gives rise to zero-sum class conflict as well as to its tyrannical political control of the means of production (both land and labor). On the other hand, the broader, more dynamic and efficient capital base of the commercial or industrial upper class can expand non-producers' resources. This factor facilitates the adoption of more compromising solutions by the dominant class. In this context, class conflict involves the economic distribution of goods and products rather than the political control of labor and land. Adopting Marx's theoretical arguments in his analysis of the French peasantry (Marx 1964), Paige further hypothesizes that cultivators' dependence on land as their main source of income gives rise to a conservative, competitive, and structurally dependent peasantry. In contrast, their dependence on wages would result in a more radical, solidary, and structurally interdependent cultivator class. In the Vietnamese case, Paige's model predicts that agrarian unrest will take the form of rebellion when both cultivators and non-cultivators derive their incomes from land, as is the case in North and Central Vietnam. Furthermore, in Cochin China, 'A combination of non-cultivators dependent on income from land and cultivators dependent on income from wages leads to revolution' (Paige 1975, p. 71, 1983, p. 706).

The Marxist framework has its share of empirical anomalies. First of all, the active resistance to French colonialism did gain support among numerous members of the native elite whose organizational efforts provided a crucial link among the rural communities. It is reflected in the necessity of the special category 'resistance-supporting landlord' in the land reform campaign in the north after independence. The active support of these members of the indigenous elite for anticolonial movements cannot be mechanically explained in terms of the struggle by the working classes of the periphery against the capitalists of the core. It is their nationalist aspirations that propelled towards a path of political activism (see also Woodside 1976, p. 303, Emerson 1960).

The regional variations in revolutionary strength and socioeconomic conditions in Vietnam present the Marxist models with a second major anomaly. In colonial Vietnam, the peasantry was most heavily dispossessed of the means of production in Cochin China (South Vietnam). However, from the mid-nineteenth century to 1975, the armed resistance to French colonialism and capitalism was generally more limited in the south than in the north and the center. With the sole exception of the 1940 armed insurrection in certain areas of Cochin China (Toan et al. 1985, pp. 318–9, cf. Paige 1975, pp. 323–4), at no point did the anti-colonial armed movement in the south exceed or even approach the intensity of the unrest in the center and the north. In the pre-World War II period, it was in Annam (Central Vietnam) that the 1930–1 unrest posed the greatest threat to the colonial order. During the first Indochinese war (1946–54), the French encountered considerably greater resistance in Annam and Tonkin. During the period of direct and massive American intervention (1964–75), it was not in the former Cochin China, but in the southern part of Central Vietnam that the National Liberation Front developed its strongest roots (Mitchell 1968, Paige 1975, pp. 326–32, Luong 1992, pp. 227–8). It is an historical fact that north and central Vietnam, despite the lesser degree of class polarization, constituted the

areas with both stronger resistance to Western intervention and greater socialist revolutionary potential.

While influenced to varying degrees by the Marxist approach, anthropological analyses of agrarian revolutions, and of historical events in general, have paid closer attention to the dynamic interplay between capitalism and indigenous non-capitalist social formations as sociocultural systems (Wolf 1969).

In James Scott's moral economy model, agrarian unrest is examined in terms of the reactive response of the peasantry in a pre-capitalist social formation to capitalism and colonialism. In the context of the world market and the colonial state, the pre-capitalist order's subsistence ethic, by which lords and the state are supposed to respect peasants' right to subsistence in exchange for legitimacy, is more often violated. The expansion of Western capitalism and the establishment of European colonies led to the violation of the subsistence ethic through the erosion of patron–client ties and the existing welfare mechanisms in the relation between cultivators and non-cultivators. Scott considers peasant movements quintessentially defensive reactions of a moral nature: defensive reactions against the violation of the subsistence ethic and the increasingly serious threat to peasants' survival during subsistence crises. Under this threat and violation, any significant natural or human induced disaster potentially sparks off agrarian unrest. The prominent role of Third Word peasants in twentieth-century political movements, Scott argues, involves this dynamic interplay between a pre-capitalist socio-cultural order on the one hand and colonialism and capitalism on the other. Unrest is thus defined, within the Durkheimian tradition of social thought, in terms of a breakdown of the normative framework during a period of transition.

In the context of the greater socialist revolutionary potential in north and central Vietnam, Luong hypothesizes that the pre-capitalist tradition of the north and the center provided stronger ideological and organizational resources for the active resistance to French colonialism in particular, and to capitalist imperialism in general. Despite considerable internal tension, this tradition nurtures values conducive to the growth of nationalism, to collectivism in a relatively mild form, and to a hierarchical organizational framework (Luong 1992).

On one level, the indigenous pre-capitalist framework was characterized by both class cleavages, and a male-oriented and kinship-centered hierarchy. Luong suggests that the hierarchical structure of the local tradition, maintained partly through the socialization process, accounts for numerous features in Vietnamese anticolonialism and the Vietnamese revolution: the male-centered elitism in the VNP recruitment and organizational methods, the Vietminh's reliance on male members of the intelligentsia for local leadership in the early period of anti-French resistance, and the male dominance in both domestic and public relations in postcolonial Marxist Vietnam.

On another level, the growth of nationalism and a collectivistic vision of political economy is facilitated by important communal institutions and corresponding conceptual categories in the local tradition of the north and the center. Northern and central villages maintained sharp boundaries through villagers' participation in the collective worship of tutelary deities, an extraordinary degree of village endogamy, the institution of communal land, and their nucleated settlement pattern. In contrast, group boundaries were considerably less rigid in southern Vietnamese communities where settlements were non-nucleated, and where there existed significant geographical mobility due to the frontier environment. Luong suggests that through a sharp distinction between insiders and outsiders in the native conceptualization of kinship, communality, ethnicity, and nationality, the northern and central tradition has intensified the negative reaction of many Vietnamese to any foreign intrusion on the indigenous landscape and to the French domination in the colonial racial diarchy. Luong also suggests that in northern and central Vietnam, the relatively strong institution of communal land in particular, and the subsistence ethic of the pre-colonial period in general, nurtured a collectivist vision of political economy even among many VNP members in the 1920s, well before their exposure to Marxism. The existence of communal lands and the subsistence ethic also facilitated the process of land collectivization in northern and central villages, in comparison to its slow progress in the south since 1975. Finally, within the bamboo hedges of the corporate rural community, the extensive kinship and communal ties facilitated the mobilization of resources by local revolutionary leaders, especially the generally well-regarded members of the native intelligentsia. Luong suggests that the indigenous non-capitalist tradition thus cannot be ignored in any analysis of the Vietnamese encounter with French colonialism in particular and the capitalist world system in general (see Marr 1971, Woodside 1976).

See also: Area and International Studies: Development in Southeast Asia; Colonization and Colonialism, History of; Southeast Asia: Sociocultural Aspects; Southeast Asian Studies: Politics; Southeast Asian Studies: Society; War: Causes and Patterns

Bibliography

Cao van Luong, Tao V, Cuong T D, Cuc D, Huu Binh T, Nhat N V 1995 *Lich su Viet Nam, 1954–1965* (Vietnamese History, 1954–1965). Nha Khoa hoc xa hoi, Hanoi, Vietnam

Emerson R 1960 *From Empire to Nation*. Harvard University Press, Cambridge, MA

Goscha C E 1999 *Thailand and the Southeast Asian Networks of the Vietnamese Revolution, 1885–1954*. Curzon Press, London

Luong H V 1992 *Revolution in the Village: Tradition and Transformation in North Vietnam, 1925–1988*. University of Hawaii Press, Honolulu, HI
Marr D G 1971 *Vietnamese Anticolonialism, 1885–1925*. University of California Press, Berkeley, CA
Marr D G 1995 *Vietnam 1945*. University of California Press, Berkeley, CA
Marx K 1964 *The Eighteenth Brumaire of Louis Bonaparte*. International Publishers, New York
Mitchell E J 1968 Inequality and insurgency: A statistical study of South Vietnam. *World Politics* **20**: 421–38
Murray M J 1980 *The Development of Capitalism in Colonial Indochina*. University of California Press, Berkeley
Paige J M 1975 *Agrarian Revolution: Social Movements and Export Agriculture in the Underdeveloped World*. Free Press, New York
Paige J M 1983 Social theory and peasant revolution in Vietnam and Guatemala. *Theory and Society* **12**: 699–737
Popkin S 1979 *The Rational Peasant: The Political Economy of Rural Society in Vietnam*. University of California Press, Berkeley, CA
Scott J C 1976 *The Moral Economy of the Peasant*. Yale University Press, New Haven, CT
Tai H T H 1983 *Millenarianism and Peasant Politics in Vietnam*. Harvard University Press, Cambridge, MA
Tai H T H 1992 *Radicalism and the Origin of the Vietnamese Revolution*. Harvard University Press, Cambridge, MA
Toan N K, Binh N C, Tao V, Nam P X, Thanh B D 1985 *Lich su Viet Nam* (History of Vietnam), Vol. 2 Nha xuat ban Khoa hoc xa hoi, Hanoi, Viet Nam
Wolf E R 1969 *Peasant Wars of the Twentieth Century*. Harper and Row, New York
Woodside A B 1976 *Community and Revolution in Modern Vietnam*. Houghton Mifflin, Boston

H. V. Luong

Violence and Effects on Children

1. Violence in America: History

The common dictionary definition of violence runs something like the following: 'exertion of physical force so as to injure or abuse.' The inclusion of 'abuse' in the definition is important, because it opens up the issues of force being used to induce negative psychological states and to violate basic human dignity. Thus, violence includes the range of behaviors that assault or batter, that induce fear or terror, and that demean and humiliate.

Understanding the effects of violence on children demands an interdisciplinary perspective, one that incorporates psychological, biological, sociological, anthropological, and historical insights within the context of a larger ecological model. Anchoring this issue is the fact that each year in the United States nearly 5,000 children and youth commit acts of lethal violence, half against others (murder) and half against themselves (suicide). The numbers of children and youth who commit potentially lethal acts is, of course, many times greater, what with the tens of thousands of 'serious assaults' and 'suicide attempts' that take place each year involving minors. For each of these kids, and for the families, communities and societies of which they are a part, this problem of violence constitutes a crisis.

The current crisis of these kids is not historically new, however. Writing in *Juvenile Justice Update* in 1997, Robert Shepherd points out that the *New York Times* 140 years ago editorially bemoaned the fact that 'the number of boy burglars, boy robbers and boy murderers is so astoundingly large as to alarm all good men.' He offered the following analysis from a child psychiatrist working in Manhattan for the Juvenile Court in the 1950s: 'At first it comes as a shock to meet youngsters under 16 who rob at the point of a gun, push dope, rape and kill. I've seen boys of 7 so small they could barely clear the desk who had sold themselves to sex perverts. Others had shot out kids eyes or had clubbed or knifed them, just for the fun of it.' And the statement is from an article entitled *Manhattan's Child Criminals Are My Job* in the *Saturday Evening Post* of March 27, 1954. Recall that in the mid-nineteenth century, the governor of New York had to call out the militia to deal with rampant youth gangs in New York City.

2. Violence in America: Today

While there are historical dimensions to the current situation, there are features of the contemporary situation that are distinctive as well. For one thing, we must recognize that the magnitude of the problem is different today. There has been a seven-fold increase in per capita aggravated assault rates among youth in the United States since 1956. Also, kids are being tried as adults and incarcerated correspondingly, in what seems like a macabre return to the medieval concepts of children as simply little adults. But the parallel is even more subject to documentation than that. The homicide rate in the city of Amsterdam in the year 1450 is estimated to have been 150 per hundred thousand. By 1850, when Amsterdam had been 'civilized,' the rate dropped to 2 per hundred thousand. This rate might be taken as a kind of base line for the minimum expectable homicide rate, the rate we find in most of the modern and relatively affluent countries in the world. But in the US we continue to find rates of 20 per hundred thousand and in the most oppressed social environments within our society we find rates of 160 per hundred thousand!

Homicide rates provide only an imprecise indicator of the overall problem of violence in the lives of American children and youth, however, for behind each murder stand many nonlethal assaults. This ratio varies as a function of both medical trauma technology (which prevents assaults from becoming homicides)

and weapons technology (which can increase or decrease the lethality of assaults). An example from Chicago illustrates this: the city's homicide rate in 1973 and 1993 was approximately the same, and yet the rate of serious assault had increased approximately 400 percent during that period. Thus, the ratio of assaults to homicides increased substantially—from 1:100 in 1973 to 1:400 in 1993 (Garbarino et al. 1992).

Class, race, and gender exert important influences on exposure to community violence. The odds of being a homicide victim range from 1:21 for Black males, to 1:369 for white females—with white males at 1:13 and Black females at 1:104 (Bell and Jenkins 1993). Being an American itself is a risk factor. As noted above, the US far exceeds all other modern industrialized nations in its youth homicide rate (even for young whites, where the rate of 11.2 per 100,000 is far more than the second place country, Scotland, with 5 per 100,000 (Richters and Martinez 1993).

3. *Exposure to Violence as an Influence on Development*

Having established this historical and social context, what next becomes clear is that understanding the effects of violence on children is very much a matter of recognizing the developmental links between early experiences of violence as a victim and later experiences of violence as a perpetrator. This developmental relationship exists both within the lives of individual children and across the generations in families. Research indicates that inside virtually every dangerously violent youth is an untreated traumatized child, a child with experiences of violent victimization, and that the there is substantial cross-generational continuity in violence within families.

Research by psychiatrists Perry et al. (1995) and van der Kolk (1996) has begun to illuminate the neurological processes that translate early trauma into later dangerous behavior through brain development and neurochemistry. Dodge et al. (1997) have further elaborated this link by documenting the adaptive processes that link the experience of violence as a young child to the development of a pattern of aggressive antisocial behavior as an elementary school-age child. Their research reveals that the problems comes when abused children develop four particular psychological adaptations (hypersensitivity to negative social cues, obliviousness to positive social cues, a readily accessible repertoire of aggressive behavior, and a belief that aggression is a successful strategy in social relations). Such children are seven times more likely than abused children who do not develop these patterns to end upon diagnosed with 'conduct disorder' (chronic aggressive and antisocial behavior). Abused children who do not develop these patterns are no more likely that children in general to develop conduct disorder.

4. *Children in War Zones*

This is highly significant because it is from the ranks of children with conduct disorder that most chronic violent delinquents come (about a third of such children emerge as chronic violent delinquents according to a variety of studies—cf. Loeber and Farrington 1997). Child abuse and community violence are rampant in the lives of kids who become violent youth. This is one of the most important effects of violence on children. My own research illuminates these links through intensive, multiple interviews with boys involved in lethal violence (Garbarino 1999), and highlights the importance of geographic concentrations of violence that create 'war zone' like settings for children. These are the social settings in which the rates of child victimization is highest (with child maltreatment rates many times higher than in other neighborhoods—Garbarino et al. 1992) and from which most lethal youth violence comes (Snyder and Sickmund 1999).

Children living in neighborhoods which simulate war zones are exposed to high levels of violence. A survey of sixth to 10th graders in New Haven, Connecticut, revealed that 40 percent had witnessed at least one incident of violent crime within the previous 12 months (Marans and Cohen 1993). In three high risk neighborhoods in Chicago, 17 percent of the elementary school age children had witnessed domestic violence, 31 percent had seen someone shot, and, 84 percent had seen someone beat up (Bell and Jenkins 1993).

Some 30 percent of the kids living in high crime neighborhoods of major metropolitan areas like Chicago have witnessed a homicide by the time they are 15 years old, and more than 70 percent have witnessed a serious assault. These figures are much more like the experience of kids in the war zones we have visited in other countries (Garbarino et al. 1991) than they are of what we should expect for children, living in 'peace.' Richters and Martinez (1993) have amplified these results. In their study, 43 percent of the fifth and sixth graders had witnessed a mugging in a 'moderately violent' neighborhood in Washington, DC. Other researchers echo these findings (e.g., Taylor et al. 1994).

5. *Children Adapt: For Better and For Worse*

Children adapt to their perception of community safety in many important ways. One is their view of the future. A Harris poll of sixth to 12th graders in 1992 revealed that 35 percent worried they would not live to old age because they would be shot (Harris et al. 1994). In our interviews with families living in public housing

projects in Chicago we learned that virtually all the children had first-hand experiences with shooting by the time they were five years old (Dubrow and Garbarino 1989). A six-year-old girl once said that her job was to find her two-year-old sister whenever the shooting started and get her to safety in the bathtub of their apartment. 'The bathroom is the safest place,' she said. Interviews with school-age children and youth confirm that the 'gun culture' is a potent factor in the life of children in diverse settings in the US (Garbarino 1995, 1999). Teen use of guns in homicides increased from 64 percent in 1987 to 78 percent in 1991. During the same period, teen arrests for crimes involving weapons increased 62 percent (Snyder and Sickmund 1995).

When children understand that adults cannot protect them they are left with what we might call juvenile vigilantism, the impulse to protect themselves, to take up weapons and relationships that substitute for adult protection. This is evident in comments from children and youth such as 'If I join a gang I'm 50 percent safe. If I don't join a gang I'm zero percent safe' (Garbarino 1999). Adults don't enter into the equation.

6. *Risk Accumulates*

The key to understanding the effects of living in a war zone is the recognition that these children do not simply face a single threat to development, a solitary risk factor. Rather, they face multiple risk factors, a fact of overarching importance in light of research by Sameroff et al. (1987) documenting the cumulative effect of risk factors. In Sameroff's study, while the presence of one or two risk factors was generally manageable for children, the presence of three or more risk factors was associated with significant developmental impairment. These are the children of greatest concern to us, the children who face major accumulations of risk factors, and who are thus most at risk for the psychological effects of trauma due to violence.

Tolan (1996) examined this phenomenon in detail in a study in a Chicago-based study in which he asked the question, 'What percent of kids are resilient if we measure resilience as neither requiring mental health intervention or remedial education?' When he asked that question of data from boys growing up in the most afflicted war zone neighborhoods of Chicago, living in abusive and impoverished families, contending with minority status in a racist society, and looked at this kids over a two-year period, the answer was zero percent. None of the children by age 15 were resilient in Tolan's terms. This testifies to the effects of violence when they occur within a larger context of social and psychological risk accumulation. The relentless pressure imposed on kids who come from that nexus of community violence, family disruption, and personal experience of trauma is uniformly overwhelming.

7. *Future Directions*

As research on the impact of violence on children matures it turns increasingly to studies of the conditions under which alternative pathways are taken by affected children. Why do some child victims become adolescent perpetrators while others do not? How can early intervention change these pathways? How significant are the neurobiological effects of early trauma? These and other questions will require many years of sustained research to address adequately. Unfortunately, the world shows no signs of preventing childhood exposure to violence, so these questions must be answered.

See also: Childhood Sexual Abuse and Risk for Adult Psychopathology; Domestic Violence: Sociological Perspectives; Early Childhood: Socioemotional Risks; Violence in Anthropology; Violence: Public; Street Children: Psychological Perspectives; War, Political Violence and their Psychological Effects on Children: Cultural Concerns

Bibliography

Bell C C, Jenkins E J 1993 Traumatic stress and children. *Journal of Health Care for the Poor and Underserved* **2**(1): 175–85

Dodge K A, Pettit G S, Bates J E 1997 How the experience of early physical abuse leads children to become chronically aggressive. In: Cicchetti D, Toth S L (eds.) *Developmental Perspectives on Trauma: Theory, Research, and Intervention*. University of Rochester Press, Rochester, NY, pp. 263–88

Dubrow N F, Garbarino J 1989 Living in the war zone: Mothers and young children in public housing development. *Journal of Child Welfare* **68**: 3–20

Garbarino J 1995 *Raising Children in a Socially Toxic Environment*. Jossey-Bass, San Francisco, CA

Garbarino J 1999 *Lost Boys: Why Our Sons Turn Violent and How We Can Save Them*. Free Press, New York

Garbarino J, Dubrow N, Kostelny K, Pardo C 1992 *Children in Danger: Coping with the Consequences of Community Violence*. Jossey-Bass, San Francisco, CA

Garbarino J, Kostelny K, Dubrow N 1991 *No Place to be a Child: Growing up in a War Zone*. Jossey-Bass, San Francisco, CA

Harris et al. 1994 *Metropolitan Life Surveys of the American Teacher: Violence in America's Public Schools: Part II*. Metropolitan Life Insurance, New York

van der Kolk B A 1996 The body keeps score: Approaches to the psychobiology of posttraumatic stress disorder. In: van der Kolk B A, McFarlane A C, Weisaeth L (eds.) *Traumatic Stress: The Effects of Overwhelming Experience on Mind, Body, and Society*. Guilford Press, New York, pp. 214–41

Loeber R, Farrington D P 1997 Strategies and yields of longitudinal studies on antisocial behavior. In: Stoff D M, Breiling J, Maser J D (eds.) *Handbook of Antisocial Behavior*. John Wiley & Sons, New York, pp. 125–39

Marans S, Cohen D J 1993 Children and inner-city violence: Strategies for intervention. In: Leavitt L A, Fox N A (eds.) *The*

Psychological Effects of War and Violence on Children. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 281–301
Snyder H N, Sickmund M *Juvenile Offender and Victims: A National Report*. Office of Juvenile Justice and Delinquency Prevention, Washington, DC
Perry B D, Pollard R A, Blakley T L, Baker W L, Vigilante D 1995 Childhood trauma, the neurobiology of adaptation, and "use-dependent" development of the brain: How "states" become "traits". *Infant Mental Health Journal* **16**(4): 271–91
Richters J E, Martinez P 1993 The NIMH Community Violence Project: I. Children as victims of and witnesses to violence. *Psychiatry: Interpersonal and Biological Processes* **56**(1): 7–21
Sameroff A, Seifer R, Barocas R, Zax M 1987 Intelligence Quotient scores of 4-year-old children: Social environment risk factors. *Pediatrics* **79**: 343–50
Shepherd R 1997 *Juvenile Justice Update*. Office of Juvenile Justice and Delinquency Prevention, Washington, DC
Snyder H N, Sickmund M 1999 *Juvenile Offenders and Victims: 1999 National Report*. Office of Juvenile Justice and Delinquency Prevention, Washington, DC
Taylor L, Zuckerman B, Harik V, Groves B M 1994 Witnessing violence by young children and their mothers. *Journal of Developmental and Behavioral Pediatrics* **15**(2): 120–23
Tolan P H 1996 How resilient is the concept of resilience? *Community Psychologist* **29**(4): 12–15

J. Garbarino

Violence and Media

Public and academic concern about media's contribution to real world violence are about as old as the mass media and the social sciences themselves (Wartella and Reeves 1985). Despite frequent framing of the matter as 'controversial,' extensive research—an estimated 3,000 (Donnerstein et al. 1994) to 3,500 (Wartella et al. 1998) studies in the United States alone—have examined the impact of media violence, and a number of recent major reviews (Huston et al. 1992, Murray 1994, see also Potter 1999, Paik and Comstock 1994, Comstock and Paik 1991), have concluded that media violence plays a measurable role in real-world violence. A variety of US agencies, including the Centers for Disease Control of the US Public Health Service (1991), and medical and public interest organizations, including the American Medical Association, the American Academy of Pediatrics, and the National Conference of Parent-Teacher Associations, have identified media violence as a public health problem. The review below focuses most heavily on US research and US media, most notably American television, primarily because a large majority of the published social science research on media and violence is US research on American audiovisual media. Potter (1999, pp. 44–5), for example, reports 42 published content analyses of US television since 1954, and just 19 from the rest of the world. Moreover, American media are among the world's most violent—and most exported—and real-world violence is a recurring public policy issue. Further, much of this literature concerns impacts of media violence on children and adolescents, for the inter-related reasons that young audiences are considered the most impressionable and most vulnerable. Adults are generally viewed to be more resistant to the deleterious influences of violence, and, as some would argue (cf. Huesmann 1997), violent behaviors in adulthood may be traced to media use during childhood.

1. *Theories of Effect*

Three models have been proposed to describe the process by which such learning and imitation of media violence occurs: social learning theory, priming effects theory, and a social developmental model of learning (Wartella et al. 1998).

First proposed by Albert Bandura in the 1960s, social learning theory is the best known theoretical account of violence effects. Bandura asserts that through observing television models, viewers come to learn behaviors which are appropriate, that is, which behaviors will be rewarded and which punished. In this way, viewers seek to attain rewards and therefore imitate these media models. When both children and adults are shown an aggressive model who is either rewarded or punished for their aggressive behavior, models who are positively reinforced influence imitation among the viewers. Even research in the field has demonstrated that aggression is learned at a young age and becomes more impervious to change as the child grows older. In a longitudinal study to examine the long-term effects of television violence on aggression and criminal behavior, Huesmann et al. (1984) studied a group of youth across 22 years, at ages 8, 18, and 30. For boys (and to a lesser, though still significant extent for girls), early television violence viewing correlated with self-reported aggression at age 30 and added significantly to the prediction of serious criminal arrests accumulated by age 30. These researchers find a longitudinal relationship between habitual childhood exposure to television violence and adult crime and suggest that approximately 10 percent of the variability in later criminal behavior can be attributed to television violence.

Priming effects theory serves to augment the more traditional social learning theory account of television violence effects. In the work of Leonard Berkowitz and his colleagues, this theoretical account asserts that many media effects are immediate, transitory, and short-term (Berkowitz 1984). Berkowitz suggests that when people watch television violence, it activates or 'primes' other semantically related thoughts which may influence how the person responds to the violence on television. Viewers who identify with the actors on television may imagine themselves like that character carrying out the aggressive actions of the character on television, and research evidence suggests that exposure to media aggression does indeed 'prime' other

aggressive thoughts, evaluations, and even behaviors such that violence viewers report a greater willingness to use violence in interpersonal situations.

Only Rowell Huesmann's (1986; see also Huesmann 1997) theoretical formulation of the social developmental model of violence effects offers a true reciprocal theoretical account of how viewers' interest in media violence, attention to such violence, and individual viewer characteristics may interact in a theory of media violence effects. Using ideas from social cognition theory he develops an elaborate cognitive mapping or script model. He argues that social behavior is controlled by 'programs' for behavior which are established during childhood. These 'programs' or 'scripts' are stored in memory and are used as guides to social behavior and problem solving. Huesmann and Miller (1994, p. 161) submit that 'a script suggests what events are to happen in the environment, how the person should behave in response to these events, and what the likely outcome to those behaviors would be.' Violence from television is 'encoded' in the cognitive map of viewers, and subsequent viewing of television violence helps to maintain these aggressive thoughts, ideas, and behaviors. Over time such continuing attention to television violence can thus influence people's attitudes toward violence and their maintenance and elaboration of aggressive scripts.

This theory suggests that while viewing violence may not cause aggressive behavior, it certainly has an impact on the formation of cognitive scripts for mapping how to behave in response to a violent event and what the outcome is most likely to be. Television portrayals, then, are among the media and personal sources that provide the text for the script which is maintained and expanded upon by continued exposure to scripts of violence.

Huesmann has demonstrated that there are key factors which are particularly important in maintaining the television viewing–aggression relationship for children: the child's intellectual achievement level, social popularity, identification with television characters, belief in the realism of the TV violence, and the amount of fantasizing about aggression. According to Huesmann, a heavy diet of television violence sets into motion a sequence of processes, based on these personal and interpersonal factors, that results in many viewers becoming not only more aggressive but also developing increased interest in seeing more television violence.

It must be emphasized that all serious scholars of the impacts of media on violence are careful to note that media are not the only, nor perhaps among the most important, contributors to real-world violence. Violent behavior is a complex, multivariable problem, formed of many influences. Racism, poverty, drug abuse, child abuse, alcoholism, illiteracy, gangs, guns, mental illness, a decline in family cohesion, a lack of deterrents, the failure of positive role models, among others, all interact to affect antisocial behavior. As Huesmann has argued, aggression is a syndrome, an enduring pattern of behavior that can persist through childhood into adulthood. The impact of media violence appears strongest as a predictor of real-world antisocial behavior as one facet of a 'culture of violence.'

2. *Types of Effects*

As Potter (1999, Chap. 9) notes, media-violence effects fall into five categories—physiological, emotional, cognitive, attitudinal, and behavioral—and both immediate or short- and long-term effects have been studied. While some attention has focused on direct, short-term imitative or modeling effects (cf. Phillips 1980, 1982, but see Hessler and Stipp 1985), more attention and public policy concern has focused on the long-term impact of repeated exposure to violence. More generally, three overarching categories of effect receive most attention: learning of aggression, desensitization to real-world violence, and the cultivation of fear in repeated exposure to media violence (Wilson et al. 1997).

Clearly, not all violent depictions should be treated equally, nor all viewers. The (US) National Television Violence Study (Wilson et al. 1997) identified several contextual factors within a representation that may influence audience reactions to media violence which include the following.

2.1 *The Nature of the Perpetrator*

Where individuals perceive perpetrators of violence as attractive, as heroes, and/or as similar to themselves, the likelihood of stimulating attention (Bandura 1986) and aggression (Paik and Comstock 1994) increases.

2.2 *The Nature of the Victim*

While the commission of violence on an attractive character with which an audience member identifies might serve to inhibit aggressive behavior, its principal impact would seem to be in arousing fear among the audience members.

2.3 *The Reason for the Violence*

Wilson et al. (1997, p. 24) note that violence viewed as justified likely heightens aggression, while violence viewed as unjustified arouses fear. The impact of justification has been documented with fictional as well as realistic programming (Meyer 1972), and with adult as well as child viewers (Liss et al. 1983). In fact, a recent meta-analysis of 217 media studies documents

that a justified portrayal of violence can enhance aggressive behavior among viewers (Paik and Comstock 1994).

2.4 The Presence of Weapons

A number of studies, including a meta-analysis of 56 published experiments (Carlson et al. 1990) have demonstrated that the presence of weapons, either pictorally or in the natural environment, can enhance aggression among subjects. While, for ethical reasons, the large majority of such research involves adult subjects, in at least one study (Frodi 1975) the presence of weapons enhanced aggression among adolescents. 'Conventional' weapons such as guns and knives are more likely than unconventional means for priming the effect, social learning theory would suggest, because their use as a means of aggression are stored in memory (Berkowitz 1990, Leyens and Parke 1975).

2.5 The Extent and Graphicness of the Violence

A review (Wilson et al. 1997) for the National Television Violence Study suggested that more research is needed, but several tentative conclusions about extent and graphicness could be reached: (a) extensiveness of violence *within* media presentations should be associated with increased desensitization to violence, at least in the short- to medium-term; (b) graphicness of violence should be associated with increased cultivation of fear; (c) longitudinal studies clearly suggest that extensiveness of *viewing*violent media presentations heightens the likelihood of engaging in aggressive behavior.

2.6 The Degree of Realism of the Violence

In brief, realistic violence has been found to induce aggressive behavior, and to induce fear, more than violence believed to be less realistic or more fantastic. An extremely important qualification deals with younger children, who may be unable to distinguish realistic from fantastic characters, behaviors, and situations. In one study, however, where *perceived realism* was manipulated for older children (9 to 11 in Feshbach 1972; 10 to 13 in Atkin 1983), those subjects who were led to believe that footage was realistic news were more likely subsequently to behave aggressively than those led to believe it was taken from an entertainment program.

2.7 Whether Violence is Rewarded or Punished

Rewarded violence is more likely to be imitated than violence which is punished. Significantly, and particularly for children (since, as we will show below, television programming most frequently presents violent actions that are neither rewarded nor punished), the absence of punishment may enhance imitation, even in the absence of explicit reward (Bandura 1965, Walters and Parke 1964). Paik and Comstock's meta-analysis (1994) suggests that rewarded violence stimulates aggression among both child and adult audiences. One study suggests that punishment of criminal violence decreases fear (Bryant et al. 1981).

2.8 Consequences of Violence

In general (and exceptions are noted in Wilson et al. 1997, p. 30), mediated depictions of violence which show either pain cues or other short- or longer-term negative effects or consequences of violence are likely to depress the learning of aggression. There is little research on the effects of pain cues or violence consequences on desensitization and the cultivation of fear.

2.9 Presence of Humor

As the National TV Violence Study review also noted, further research is needed here as well, but the present state of knowledge suggests, other things being equal, that violence coupled with humor is more likely to heighten aggression, and to increase desensitization, than violence without the presence of humor:

> Several mechanisms can be used to explain such a facilitative effect of humor on aggression. Humor might elevate a viewer's arousal level over that attained by violence alone, and increased arousal has been shown to facilitate aggression. … Humor could serve as a reinforcement or reward for violence, especially if the perpetrator is funny or admired or his or her wit. And humor may diminish the seriousness of the violence and therefore undermine the inhibiting effects of harm and pain cues in a scene. … However, we should underscore that our conclusion about the facilitative effect of humor on aggression is tentative until more systematic research … is undertaken (Wilson et al. 1997, p. 32).

3. Young Viewers

As noted, research indicates that certain factors may be processed differently by young viewers. First, children below about age 8 have more difficulty distinguishing reality from fantasy and often imitate superheros with magical powers such as the *Power Rangers* (Boyatzis et al. 1995). Second, young children may have difficulty connecting scenes and drawing inferences from the plot. Timing of punishments and rewards becomes important in this instance. In many programs, the crime or violent behavior may go unpunished until the end of the program. Young children may have difficulty connecting the ending punishment with the initial violent act and may, therefore, believe that the violence went unpunished (Wilson et al. 1997). Thus, learning of aggressive attitudes and behaviors from television varies by both the nature of the portrayals and the nature of the

viewers. The presence of contextual factors in the portrayals which may inhibit young children's social learning of aggression decreases the negative consequences of such portrayals and should be encouraged. Not all violent portrayals are the same and the context of violence is clearly quite important. Similarly, young children, those under the age of seven or eight, may be particularly susceptible to learning from exposure to television violence because of differences in how they make sense of television compared to adults.

4. *The Media Environment*

Television's role as the central mass medium in much of the world for the past half-century, and its ubiquity and ability to enter almost every home, often without parental supervision, has meant that more public and scholarly concern has focused on its contents than on any other medium's, and this concern has accompanied its diffusion into every corner of the earth.

Unfortunately, cumulative and comparative research on television's violent content is hampered by a lack of consistency in defining violence and especially defining the population and sampling frame in studies of television.

The most extensive single content study of US television was the 1994–7 National Television Violence Study (National Television Violence Study 1997, Center for Communication and Social Policy 1997, 1998). Examined were the 6 a.m.–11 p.m. contents of a multistage probability sample constructed sample week of programming on 23 network-station, independent-station, and basic-cable and premium-cable channels; thus about 8,000 programs were analyzed over the 1994–5, 1995–6, and 1996–7 television 'seasons.' Certain programs, including 'hard news,' religious shows, sporting events, quiz shows, and educational shows, were sampled but not analyzed. About three-fifths of the remaining programs contained some visual violence, a figure that like most summary statistics remained stable over the three years of the study. In descending order, premium cable, basic cable, independent-station, broadcast network station, and public broadcast stations' programming were likely to contain violence. By content genre, in decreasing order, movies, dramas, children's shows, music videos, and reality-based and comedy programs were likely to contain violence. Violence was far more prevalent during prime-time than during daytime hours.

Of signal concern to the NTVS researchers was the *context* of televised violence; it was often glamorized (more than a quarter of all violence was perpetrated by 'good' or attractive characters, and some 40 percent by characters with at least some good qualities); sanitized (about 7/8 of violent scenes show no blood and gore; almost half show no harm to victims of violence, although more than half of violent interactions show infliction of harm that would be lethal in 'real life,' and about half depict no pain cues in victims of violence); and unsanctioned (in a majority of scenes, violence perpetrators were neither rewarded nor punished ['punishment' was considered any noticeable sanction, including a perpetrator's oral expression of remorse], and among other scenes, rewards and punishments were about equally likely; in three-quarters of cases, characters perpetrating violence were either never punished, or were punished only at the program's conclusion). Moreover, only three percent of programs with violence had any antiviolence theme (Center for Communication and Social Policy 1998, Chaps. 3–4). Potter (1999) has an extensive discussion of definitions of media violence and the results of content analyses from a variety of studies.

5. *Media Violence and Public Policy*

As noted, the consensus in the social scientific community regarding media violence is that it serves as a contributor to aggression in the real world, and virtually all public opinion surveys confirm that wider publics believe this as well. Nonetheless, such findings are 'controversial' in the media industries, and among a minority of academics (see, e.g. Fowles 1999 and sources cited therein). In the United States in particular, since the 1950s the media industries—television in particular, but also recorded music, motion pictures, pictorial comic books, and video games—have responded to public and governmental outcries against violent content by promises to reform under self-regulation. Motion pictures, video-games, and recorded music all list 'ratings' for their products which suggest age-groups for which the industry self-regulatory groups think the content for these products is appropriate. The US Telecommunications Act of 1996 mandated that the television networks create a 'voluntary' ratings system or face creation of one by the federal government. The system the broadcast industry has created suggests appropriate age categories and levels of sex, language, and violence to allow consumers and parents to make program choices. The same act mandated a 'v-chip' in newly manufactured television sets to allow parents to screen or filter out violent programs. There is to date insufficient research to indicate whether and what sorts of parents and other viewers are using either the ratings or the v-chip technology to screen violence. However, experimental research by Cantor and her colleagues (Cantor and Nathanson 1998, Cantor et al. 1997) suggests that young children may use aged-based ratings systems to shield themselves from violent content while for older children and adolescents, there may be a boomerang or 'forbidden fruit' effect whereby ratings *attract* them to more 'adult' violent or sexually explicit material. In the United States Congress, a moratorium on discussion of television violence is in effect, pending further information on the effects of ratings and the v-chip, but the legislature is

focusing its attention on violent video-games, the target of significant public criticism in the wake of a number of firearms murders in public schools.

6. Conclusion

It is clear that where children and television violence are concerned, the question that remains is not whether media violence has an effect, but rather how important that effect is in comparison with other factors in bringing about the current level of crime in the United States and other industrialized nations. Future research should also aim to establish who precisely is most susceptible to media violence, and, most importantly, what sorts of intervention might help diminish its influence. At the same time, any interventions that help establish policies and practices to reduce the socially inappropriate ways of portraying violence and increase the socially responsible ways (such as using violence to assert antiviolence messages) should be encouraged as well. Long-term solutions to problems caused by violence in the real world, however, will require attention to a much wider variety of causal agents.

See also: Media Effects; Media Effects on Children

Bibliography

American Psychological Association 1993 *Violence and Youth: Psychology's Response. Volume 1: Summary Report of the American Psychological Association Commission on Violence and Youth*. American Psychological Association, Washington, DC

Atkin C 1983 Effects of realistic TV violence vs. fictional violence on aggression. *Journalism Quarterly* **60**: 615–21

Bandura A 1965 Influences of models' reinforcement contingencies on acquisition of imitative responses. *Journal of Personality and Social Psychology* **1**: 589–95

Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ

Berkowitz L 1984 Some effects of thoughts on anti-social and pro-social influences of media events: A cognitive-neoassociationistic analysis. *Psychological Bulletin* **95**(3): 410–27

Berkowitz L 1990 On the formation and regulation of anger and aggression: A cognitive-neoassociationistic analysis. *American Psychologist* **45**: 494–503

Berkowitz L, LePage A 1967 Weapons as aggression-eliciting stimuli. *Journal of Personality and Social Psychology* **7**: 202–7

Boyatzis C, Matillo G, Nesbitt K 1995 Effects of 'The Mighty Morphin Power Rangers' on children's aggression with peers. *Child Study Journal* **25**(1): 45–55

Bryant J, Carveth R A, Brown D 1981 Television viewing and anxiety: An experimental examination. *Journal of Communication* **31**(1): 106–19

Cantor J, Harrison K, Nathanson A 1997 Ratings and advisories for television programming, Part III. In: Center for Communication and Social Policy, University of California, Santa Barbara (ed.) *National Television Violence Study*. Sage, Thousand Oaks, CA, Vol. 2, pp. 267–322

Cantor J, Nathanson A 1998 Ratings and advisories for television programming, Part III. In: Center for Communication and Social Policy, University of California, Santa Barbara (ed.) *National Television Violence Study*. Sage, Thousand Oaks, CA, Vol. 3, pp. 285–321

Carlson M, Marcus-Newhall A, Miller N 1990 Effects of situational aggression cues: A quantitative review. *Journal of Personality and Social Psychology* **58**: 622–33

Center for Communication and Social Policy, University of California, Santa Barbara 1997b *National Television Violence Study*. Sage, Thousand Oaks, CA, Vol. 2

Center for Communication and Social Policy, University of California, Santa Barbara 1998 *National Television Violence Study*. Sage, Thousand Oaks, CA, Vol. 3

Centers for Disease Control 1991 *Position Papers from the Third National Injury Conference*. US Department of Health and Human Services, Atlanta, GA

Comstock G, Paik H 1991 *Television and the American Child*. Academic Press, San Diego, CA

Donnerstein E, Slaby R, Eron L 1994 The mass media and youth violence. In: Murray J, Rubinstein E, Comstock G (eds.) *Violence and Youth: Psychology's Response*. American Psychological Association, Washington, DC, Vol. 2, pp. 219–50

Feshbach S 1972 Reality and fantasy in filmed violence. In: Murray J P, Rubinstein E A, Comstock G (eds.) *Television and Social Behavior: Television and Social Learning*. US Government Printing Office, Washington, DC, Vol. 1, pp. 318–45

Fowles J 1999 *The Case for Television Violence*. Sage, Thousand Oaks, CA

Frodi A 1975 The effects of exposure to weapons on behavior from a cross-cultural perspective. *International Journal of Psychology* **10**(4): 283–92

Gerbner G, Gross L, Morgan M, Signorielli N 1980 Living with television: the dynamics of the cultivation process. In: Bryant J, Zillmann D (eds.) *Media Effects*. Lawrence Erlbaum, Hillsdale, NJ, pp. 17–41

Hessler R C, Stipp H 1985 The impact of fictional television suicide stories on United States fatalities: A replication. *American Journal of Sociology* **90**: 151–67

Huesmann L R 1986 Psychological processes promoting the relation between exposure to media violence and aggressive behavior by the viewer. *Journal of Social Issues* **42**(3): 125–39

Huesmann L R 1997 *Screen Violence and Real Violence: Understanding the Link*. Huesmann, Ann Arbor, MI

Huesmann L R, Lefkowitz M M, Walder L O 1984 The stability of aggression over time and generations. *Developmental Psychology* **20**(6): 1120–34

Huesmann L R, Miller L 1994 Long-term effects of repeated exposure to media violence in childhood. In: Huesmann L R (ed.) *Aggressive Behavior: Current Perspectives*. Plenum, New York, pp. 153–83

Huston A C, Donnerstein E, Fairchild H, Feshbach N, Katz P, Murray J, Rubinstein E, Wilcox B, Zuckerman D 1992 *Big World, Small Screen: The Role of Television in American Society*. University of Nebraska Press, Lincoln, NE

Leyens J P, Parke R D 1975 Aggressive slides can induce a weapons effect. *European Journal of Social Psychology* **5**: 229–36

Liss M, Reichardt L, Fredriksen S 1983 TV heroes: The impact of rhretoric and deeds. *Journal of Applied Developmental Psychology* **4**: 175–87

Meyer T P 1972 Effects of viewing jusitified and unjustified real film violence on aggressive behavior. *Journal of Personality and Social Psychology* **23**: 21–9
Murray J P 1994 Impact of televised violence. *Hofstra Law Review* **22**: 809–25
National Television Violence Study 1997 *National Television Violence Study*. Sage, Thousand Oaks, CA, Vol. 1
Page D, O'Neal E 1977 'Weapons effect' without demand characteristics. *Psychological Reports* **41**: 29–30
Paik H, Comstock G 1994 The effects of television violence on antisocial behavior: A meta-analysis. *Communication Research* **21**: 516–46
Phillips D P 1980 Airplane accidents, murder and the mass media. *Social Forces* **58**: 1001–24
Phillips D P 1982 The impact of fictional television stories on United States adult fatalities. *American Journal of Sociology* **87**: 1340–59
Potter W J 1999 *On Media Violence*. Sage, Thousand Oaks, CA
Turner C W, Layton J F, Simons L S 1975 Naturalistic studies of aggressive behavior: Aggressive stimuli, victim visibility, and horn honking. *Journal of Personality and Social Psychology* **31**: 1098–107
Walters R H, Parke R 1964 Influence of response consequences to a social model on resistance to deviation. *Journal of Experimental Child Psychology* **1**: 269–80
Wartella E 1995 Media and problem behaviours in young people. In: Rutter M, Smith D J (eds.) *Psychosocial Disorders in Young People: Time Trends and Their Causes*. J. Wiley, Chichester, UK, pp. 296–323
Wartella E, Olivarez A, Jennings N 1998 Children and television violence in the United States. In: Carlson U, von Feilitzen C (eds.) *Children and Media Violence: Yearbook from the UNESCO Clearinghouse on Children and Violence on the Screen*. UNESCO International Clearinghouse on Children and Violence on the Screen, Göteborg, Sweden, pp. 55–62
Wartella E, Reeves B 1985 Historical trends in research on children and the media: 1900–1960. *Journal of Communication* **35**(2): 118–33
Whitney C, Wartella E, Lasorsa D, Danielson W, Olivarez A, Lopez R, Klijn M 1996 Part II: Television violence in 'reality' programming. In: University of Texas, Austin (ed.) *National Television Violence Study*. Sage, Thousand Oaks, CA, Vol. 1, pp. 269–360
Wilson B, Kunkel D, Linz D, Potter W J, Donnerstein E, Smith S, Blumenthal E, Gray T 1997 Part I: Violence in television programming overall. In: Center for Communication and Social Policy, University of California, Santa Barbara (ed.) *National Television Violence Study*. Sage, Thousand Oaks, CA, Vol. 1, pp. 1–268

C. Whitney and E. Wartella

Violence as a Problem of Health

In the early 1990s, many policy makers and social scientists related the problem of community violence to neighborhoods trapped in struggles among gangs competing to sell and distribute illegal drugs, especially cocaine and crack cocaine. From this perspective, the challenge for communities was to locate and restrain perpetrators and respond to victims. Gradually, it has been recognized that community violence, however it began, affects many segments of the population. Seemingly clear distinctions between victim and perpetrator seem to have faded. Heightened exposure to violence is assumed to increase acts of aggression in anticipation of violence from another. Effectively, the timing rather than the nature of an act may differentiate victim from victimizer. In such an environment, the sense of unease may seem constant, the need for vigilance continuous, and suspicion of others required. That state may, in turn, disrupt sleep, upset vegetative functions, impair interpersonal relationships, and disable effective concentration on work and studies. Together, these characteristics produce an environment marked by 'pervasive community violence' (PCV), an ecological situation with significant impact on emotional, behavioral, and psychological health.

PCV, therefore, negatively alters the essence of community life, broadly disrupts its functioning and presumably thereby impairs residents' general health. The potential for such broad impact must be appreciated and respected *before* one can enlist communities to join in investigating PCV's nature and consequences. Investigators who do not appreciate that quality may be surprised by the reluctance of community leaders (e.g., school administrators, county health officer, and representatives of the mayor's office) to participate in its study. The leaders may deny the problem's extent or hesitate to make public the negative quality of their neighborhood or community. For them genuine partnerships require time accumulated through historical involvement and future commitment. History enables disclosure and trust. Intimate aspects of a community are rarely shared with strangers. Assurance that the investigator will be there in the future (even after external funds are gone) is important, given community leaders' appreciation of the intransigence of social problems and the fickleness of social scientists. Thus, from the outset it must be appreciated that to study and alter the nature and consequence of PCV requires a genuine and long-term partnership between communities and social scientists.

1. *Definition of Terms*

As one reviews the growing PCV literature, the need for defining relevant terms becomes apparent. *Violence* refers to both acts and consequences of the *intentional* use or threat of use of force to cause injury, harm, or death to another. The qualifier 'intentional' excludes unintended or accidental acts and consequences. Determining intent represents an as yet unresolved methodological challenge. For PCV, the *consequences of violent acts* refer to the rippling health and mental health effects experienced by: (a) individuals directly victimized by an act or threat of violence; (b) indivi-

duals indirectly victimized as a bystander, witness, acquaintance, or loved one of a direct victim; and (c) individuals indirectly victimized by their awareness of, and anxiety about, the occurrence of acts or threats of violence. *Community violence* locates the acts and their consequences in a geographic setting (e.g., a neighborhood or a school), an identifiable social grouping (e.g., a grade cohort, a gang, a nuclear or extended family) or both. *Pervasive community violence* (*PCV*) refers to the chronic and widespread occurrence of acts or threats of violence within such settings or social groups.

Thus, PCV exists when *substantial* acts and consequences of violence are encountered across *multiple* settings and circumstances over *extended* time. To investigate PCV systematically, the qualifiers 'substantial,' 'multiple,' and 'extended' must be operationally defined and rigorously measured. Information must be gathered about a community's rate of violent acts and about rates of individual encounters with violent acts. For methodological and political reasons, PCV measures vary in their consideration of violence occurring within or outside of the home. The majority of such measures limit their focus to outside encounters. Yet, by definition, PCV implies cross-setting exposure. At its most extreme, PCV represents the absence of any sanctuary or setting in which relief and a sense of safety is possible. Thus, the most serious impacts on physical and psychological health are likely to be observed in those exposed at home, at work or school, and en route between the two.

In a series of papers and chapters, I have proposed consideration of *PCV as an environmental contaminant* with measurable effects on health and mental health (Cooley-Quille and Lorion 1999, Lorion 1998, 1999, 2001, Lorion et al. 1998, Lorion and Saltzman 1993). From that perspective, PCV contaminates settings and thereby the actions and interactions of at least some inhabitants. *Contamination* refers to the corrosive or pathogenic effect of PCV exposure on community climate. Slowly investigators are documenting PCV's negative effects on: (a) individual assumptions about the threatening quality of others, (b) the need to limit one's contact with others, and (c) the urgency of responding aggressively to actual or anticipated attacks. This ecological contamination appears to vary in its pace as well as intensity across individuals, situations, and settings. Understanding the determinants and mechanisms of such patterns and their malleability within and across individuals should inform the design of interventions to limit PCV exposure and reduce its consequences.

2. *Measuring PCV*

Investigation of PCV's environmental influence requires the availability of measures to estimate its presence and to confirm its impacts. Applying a common measurement approach across diverse populations and settings, Richters and associates (Lorion and Saltzman 1993, Martinez and Richters 1993, Osofsky et al. 1993, Richters and Martinez 1993) scientifically documented the extent of children's exposure to violence. They also provided early evidence of the relationship between PCV exposure and children's emotional and behavioral functioning. Subsequent investigations of such relationships (e.g., Hill and Madhere 1996, Singer et al. 1995) confirmed the seriousness of such links.

Measures of exposure to community violence share many similarities with the Survey of Children's Exposure to Community Violence (Richters and Saltzman 1990). Instructed explicitly *not* to include media exposure (i.e., what was seen or heard about on radio, television, or in movies), respondents indicate along a five-point Likert scale (from 0 = *never* to 4 = *a lot of the time*) how often they have been victimized by, heard about, or witnessed violent events or related activities (e.g., shooting, stabbing, mugging, seeing a dead body). For each instance of reported exposure, respondents provide information about: (a) where the exposure occurred (i.e., in or near school, in or near home, in the neighborhood); (b) who perpetrated the violence (e.g., a stranger, a family member); (c) who, if not the respondent, was victimized (e.g., a stranger, a family member); and (d) when the incident occurred (e.g., one week ago, five years ago).

Variations of this measure were applied by Martinez and Richters (1993) and Osofsky et al. (1993) in studies of mothers of elementary graders and by Richters and Martinez (1993) and Saltzman (1992) in their studies of fifth and sixth graders. A cartoon-based version of this measure was designed by Richters and Martinez (i.e., Things I Have Seen and Heard, 1990) to assess PCV exposure in primary graders. Test–retest reliability was a most acceptable $r = 0.81$. A second cartoon-based measure, Levonn, assessed distress symptoms associated with exposure (Richters et al. 1990). Subsequently, Fox and Leavitt (1995) developed the Violence Exposure Scale for Children (VEX-R) as a cartoon-based measure of PCV exposure in children from preschool through the elementary grades. Using this measure, Shahinfar et al. (2000) documented PCV's impact on preschoolers. The negative behavioral consequences of such exposure was confirmed in a sample of Israeli elementary school children by Raviv et al. (2001).

A further variation of the Richters and Saltzman (1990) measure was applied by Saltzman (1995) to examine PCV exposure's link to adolescent involvement in antisocial and delinquent behaviors. Building on these variations, Singer et al. (1995) designed the Recent Exposure to Physical Violence and the Past Exposure to Violence surveys to obtain adolescent reports of exposure during the prior 12 months or 'while growing up, not including the past year,' respectively. The acts of violence surveyed by Singer et

al. included threats, slapping, hitting, punching, beatings, knife attacks, and shootings. Separate reports for each (other than knife attacks and shootings) were obtained for home, school, and neighborhood. Respondents were also asked how often they had witnessed someone else victimized at each of the aforementioned sites. Two additional items inquired about respondents' having been 'made to do a sexual act' or having witnessed someone else 'being made to do a sexual act.' Overall, measurement efforts to date support the conclusion that psychometrically acceptable reports of PCV exposure can be obtained from youth ranging in age from preschoolers to adolescents.

In addition to the above work, Jenkins and associates (Bell et al. 1988, Jenkins and Thompson 1986) designed and conducted qualitative studies of African-American children's encounters with violence. Results from four studies are summarized in Bell and Jenkins (1993). Using related measures, Jenkins and colleagues obtained information about youth's exposure in terms of: (a) knowing someone who was a victim of violence; (b) having witnessed violent events *in real life*; (c) having been a victim of a violent incident; and (d) having perpetrated a violent act. Comparable to the work mentioned above, information about such forms of exposure was obtained from students enrolled in grades 2 through 12.

Although not a universal experience, victimization occurs at levels approaching 50 percent for threats and 40 percent for forced entry into their home. Indirect victimization (e.g., being a witness to a violent act) levels were consistently higher for incidents ranging from drug use to shootings, stabbings, and muggings. Reports of seeing a dead body in the neighborhood ranged from 24 to 40 percent. Using an extension of the Richters and Saltzman measure, Singer and colleagues (Singer et al. 1995, Song et al. 1998) validated these reports in their surveys of high-school students ($N = 3{,}735$). As expected, considerable variability is present across settings. Nevertheless, levels of victimization and witnessing (both at home and in school) justify recognition of PCV as a major environmental contaminant. Apfel and Simon's (1996) description of communal violence as referring to settings in 'which every child has witnessed or expects to witness violence and has been or expects to be violated' (pp. 4–5), seems to apply to many of the nation's neighborhoods and schools.

3. *Assessing PCV's Health Impact*

The aforementioned studies also report on PCV's toxicity. Saltzman (1992), for example, observed significantly higher levels of generalized emotional distress in PCV-exposed fifth and sixth graders. This pattern is also reported by Richters and Martinez (1993), Martinez and Richters (1993), Osofsky et al. (1993), and Bell and Jenkins (1993). Across these studies of preadolescents, PCV exposure relates positively to affective and vegetative signs of depression and indices of stress-related disorders. With both preadolescents and adolescents, Singer et al. (1994) linked PCV exposure and increased aggression, conduct disorder, running away, and decreases in self-esteem and social competence. Support for the link between PCV exposure and involvement in violent acts is reported by Saltzman (1995). In Saltzman's study, PCV exposure significantly related to adolescent involvement in antisocial and violent acts independent of peer, familial, and other recognized predictors of aggression during that developmental period. Rubinetti (1996) linked PCV exposure and violent activity most strongly in youth with decreased empathy, hopefulness, and self-esteem. Provocatively, Song et al.'s (1998) findings suggest that indirect victims are more likely than direct victims to aggress *in anticipation of* violent acts. Although further study is needed, PCV may create interpersonal distance among the residents of an affected setting, desensitize them to the consequences of violence on others, while heightening the need to defend oneself against assault. Overall, there appears to be substantial evidence that heightened exposure to PCV increases the likelihood that one will engage in violence. In turn, the resulting acts of violence increase levels of PCV exposure, providing a mechanism for a continuing cycle.

Two recent studies (Cooley-Quille and Lorion 1999, Scarpa et al. 2000) have begun to examine the physical impacts of PCV exposure. Each study found physiological concomitants of PCV exposure, including increased heart rate, respiration, and cortisol levels, long-established indices of heightened stress and arousal. Cooley-Quille and Lorion (1999) also reported preliminary evidence of a link between PCV and sleep disorders. Anecdotal reports from youth living in PCV-heightened neighborhoods report difficulty falling asleep, repeatedly waking up in response to gunshots, loud noises, and police activity and, generally, only rarely sleeping throughout the night. If these reports are confirmed along with further information about the vegetative effects of PCV exposure, new understandings may emerge which change existing assumptions about the lack of impulse control and academic failure, often associated with urban poverty, from attributions about personal failure to questions about environmental toxicity. Thus, although examination of the physiological consequences of living or working within settings characterized by high PCV levels is just beginning, preliminary results suggests that these impacts are real, extensive, and potential triggers for further pathogenesis across multiple parameters of functioning.

Such findings support the hypothesized toxicity of PCV exposure which appears to impair cognitive, emotional, behavioral, and biological functioning simultaneously, sequentially, or synergistically. As yet

unexamined systematically, however, is the breadth of these 'PCV-consequences' (assuming they are causally linked) within settings and communities. In essence, we have yet to study how the toxin spreads and the transmission routes by which the occurrence of violent events (presumably their nature, frequency, and intensity would influence the spread of effects) contaminates the environment. Longitudinal information is necessary to causally link exposure and its observed concomitants. It is also necessary to determine at what point interventions can be introduced to protect against or, if necessary, to treat such effects.

It should be recognized that what is presently known about PCV relates almost entirely to school-aged youth. We know little about the effect of maternal PCV exposure on neonates, of exposure on toddlers and preschoolers, and even less about the nature, extent, and consequences of exposure on adults. What of the parents, grandparents, and others who live in such neighborhoods? Within schools, what of the staff and teachers who may live in such neighborhoods, travel through such neighborhoods, or hear the tales of their students? Work reported to date pays little attention to these members of exposed communities, except perhaps to suggest that they could buffer youth against the impacts of PCV. Whether we can assume the capacity of exposed adults to do so merits careful examination.

We must therefore examine exposure across the entirety of a community's population and settings. Anecdotal reports, informal focus groups, and clinical encounters suggest that exposure reduces the quality of teachers' instructional effectiveness, capacity to apply classroom management strategies, and availability to students before or after the school day. Reasonably, one might expect that PCV similarly disrupts parent–child, parent–teacher, and teacher–teacher interactions. As reported above, Song et al. (1998) found a significant link between indirect exposure and anticipatory aggression. Ambiguous interpersonal cues may be interpreted by youth (and perhaps adults?) as threatening, leading to what is assumed by the perceiver as a defensive act. Presumably (although we do not really know), when self-defense seems unlikely to succeed, ambiguity leads to retreat, avoidance, or perhaps acquiescence with stated or unstated demands.

If this explanation of Singer's finding is valid, would it not extend to adult interactions within and across generations? Might it not explain, at least in part, what from the outside appears as hypersensitivity to insult or disrespect, the immediacy with which one asserts one's right, or the reluctance to give way, lest courtesy be mistaken for weakness or fear? Might it not explain urban parents' use of physical discipline or its reverse, that is, seeming inability or unwillingness to control their children's behavior? What of neighborhoods in which the elderly limit, and the young have limited, time outside of the home because of concerns about danger and vulnerability? Would not isolated streets facilitate violence with immunity? Would not the biological concomitants related to stress, sleep deprivation, and concentration impair adults' (be they teachers, storekeepers, bus drivers, or neighbors) capacity to interact comfortably with each other and to neutralize or at least dilute PCV's toxicity and thereby protect youth? To ignore the status of adults in such settings implies that they are unaffected by PCV or other community toxins.

Albeit at levels of concern, PCV is linked to increased but not universal pathology. Not only does a majority of those exposed not manifest pathology, but most residents do not report exposure. PCV exposure appears to relate not only to the occurrences of violent acts, but also to awareness of those occurrences. Somehow, many exposed are seemingly unaware of that and, potentially, for them there is some bliss in ignorance. What individual and ecological factors determine sensitivity remains to be investigated. That such differences exist, however, argues for attending to issues of adaptation, coping, and wellness. As noted recently:

> By implication it seems reasonable to view wellness as representing the outcome of normative developmental processes, in programming terms, as the default option. In the absence of that relatively rare combination of individual and situational circumstances, individual development appears to proceed within socially expected and appropriate boundaries reflecting the general capacity to adapt to situational demands and cope with stressful events. Just as canalization accounts for the limits within which genetic processes emerge within species (Gottlieb 1992), an analogous process may account for the human capacity to adapt to a multitude of settings and circumstances, and to cope with seemingly debilitating and pathogenic stressors. (Lorion 2000, pp. 13–14)

If wellness, adaptation, and adjustment represent the default option in human development, and individuals, families, neighborhoods, schools, and communities in general cope with PCV, economic (dis)advantage, familial instability, critical life events, inadequate school budgets, etc. (the length of one's list merely underscores the point), understanding those mechanisms may inform the design of interventions to protect those who cannot cope. Answers to questions about how to optimize development and coping must be out there, in the homes, the schools, the classrooms, and the neighborhoods. To find those answers, we in the social sciences must enter the communities to observe, inquire, and learn. I propose we go not as experts bringing esoteric knowledge, but as students and collaborators, hopefully bringing tools and perspectives whose utility will be determined relative to what already exists, and what can be sustained over time and across levels of resources.

See also: Community Health; Ecology and Health; Environmental Health and Safety: Social Aspects

Bibliography

Apfel R J, Simon B (eds.) 1996 *Minefields in Their Hearts*. Yale University Press, New Haven, CT

Bell C C, Hildreth C J, Jenkins E J, Levi D, Carter C 1988 The need for victimization screening in a poor, outpatient medical population. *Journal of the National Medical Association* **80**: 853–60

Bell C C, Jenkins E J 1993 Community violence and children on Chicago's southside. *Psychiatry: Interpersonal and Biological Processes* **56**: 46–54

Cooley-Quille M C, Lorion R P 1999 Adolescents' exposure to community violence: Sleep and psychophysiological functioning. *Journal of Community Psychology* **27**: 367–76

Fox N A, Leavitt L A 1995 *The Violence Exposure Scale for Children (VEX)*. University of Maryland, College Park, MD

Gottlieb G 1992 *Individual Development and Evolutions: The Genesis of Novel Behavior*. Oxford University Press, New York

Hill H M, Madhere S 1996 Exposure to community violence and African American children: A multidimensional model of risks and resources. *Journal of Community Psychology* **24**: 26–43

Jenkins E J, Thompson B 1986 Children talk about violence: Preliminary findings from a survey of black elementary school children. Paper presented at the 19th Annual Convention of the Association of Black Psychologists, Oakland, CA

Lorion R P 1998 Exposure to urban violence: Contamination of the school environment. In: Eliott D S, Williams K, Hamburg B (eds.) *Violence in American Schools*. Cambridge University Press, New York, pp. 293–311

Lorion R P 1999 Community prevention and wellness. In: Herson M, Ammerman T (eds.) *Advanced Abnormal Child Psychology*. Erlbaum, Hillsdale, NJ, pp. 251–66

Lorion R P 2000 Theoretical and evaluation issues in the promotion of wellness and the protection of 'well enough.' In: Cichetti D, Rappaport J, Sandler I, Weissberg R (eds.) *The Promotion of Wellness in Children and Adolescents*. Sage, Thousand Oaks, CA, pp. 1–28

Lorion R P 2001 Exposure to urban violence: Shifting from an individual to an ecological perspective. In: Schneiderman N, Tomes H, Gentry J, Speers M, Silva J (eds.) *Integrating Behavioral and Social Sciences With Public Health*. American Psychological Association, Washington, DC

Lorion R P, Brodsky A E, Cooley-Quille M 1998 Exposure to pervasive community violence: Resisting the contaminating effects of risky settings. In: Biegel D E, Blum A (eds.) *Innovations in Practice and Service Delivery Across the Life Span*. Oxford University Press, New York, pp. 124–44

Lorion R P, Saltzman W 1993 Children's exposure to community violence: Following a path from concern to research to action. *Psychiatry: Interpersonal and Biological Processes* **56**: 55–65

Martinez P, Richters J E 1993 The NIMH Community Violence Project: II. Children's distress symptoms associated with violence exposure. *Psychiatry: Interpersonal and Biological Processes* **56**: 22–35

Osofsky J D, Wewers S, Hann D M, Fick A C 1993 Chronic community violence: What is happening to our children? *Psychiatry: Interpersonal and Biological Processes* **56**: 36–45

Raviv A, Erel O, Fox N A, Leavitt L A, Raviv A, Dar I, Shahinfar A, Greenbaum C W 2001 Individual measurement of exposure to everyday violence among elementary school children across various settings. *Journal of Community Psychology* **29**

Richters J E, Martinez P 1990 *Checklist of Child Distress Symptoms: Self-Report Version*. National Institute of Mental Health, Rockville, MD

Richters J E, Martinez P 1993 The NIMH Community Violence Project: I. Children as victims of and witnesses to violence. *Psychiatry: Interpersonal and Biological Processes* **56**: 7–21

Richters J E, Martinez P, Valla J P 1990 *Levonn: A Cartoon-based Structured Interview for Assessing Young Children's Distress Symptoms*. National Institute of Mental Health, Washington, DC

Richters J E, Saltzman W 1990 Survey of exposure to community violence—Parent report version. Unpublished measure. Child and Adolescent Disorders Research, National Institute of Mental Health

Rubinetti F 1996 Empathy, self-esteem, hopelessness. and belief in the legitimacy of aggression in adolescents exposed to pervasive community violence. Unpublished doctoral dissertation. University of Maryland, College Park, MD

Saltzman W 1992 The effect of children's exposure to violence. Unpublished master's thesis. University of Maryland, College Park, MD

Saltzman W 1995 Exposure to community violence and the prediction of violent antisocial behavior in a multi-ethnic sample of adolescents. Unpublished doctoral dissertation. University of Maryland, College Park, MD

Scarpa A, Fikretoglu D, Luscher K 2000 Community violence exposure in a young adult sample: II. Psychophysiology and aggressive behavior. *Journal of Community Psychology* **28**: 417–26

Shahinfar A, Fox N A, Leavitt L A 2000 Preschool children's exposure to violence: relation of behavior problems to parent and child reports. *American Journal of Orthopsychiatry* **70**: 115–25

Singer M I, Anglin T M, Song L, Lunghofer L 1994 *The Mental Health Consequences of Adolescents' Exposure to Violence*. Case Western Reserve University, Cleveland, OH

Singer M I, Anglin T M, Song L, Lunghofer L 1995 Adolescents' exposure to violence and associated symptoms of psychological trauma. *Journal of the American Medical Association* **273**: 477–82

Song L, Singer M, Anglin T M 1998 Violence exposure and emotional trauma as a contributor to adolescents' violent behavior. *Archives of Pediatric Adolescent Medicine* **152**: 531–6

R. P. Lorion

Violence, History of

Violence in its various forms has been an aspect of historical descriptions since Thucydides, but to this day there is no independent specialty in historical research on violence. The history of violence is thus written as part of military and war history, of research on revolution and protest, and of the history of gender and the body. In Europe and elsewhere, it cuts across the institutional and disciplinary logic of the study of history. This lack of independence has two consequences for historical research on violence. First, it is always interdisciplinary and includes approaches

proper to political science and social history, as well as issues of cultural history. Second, it cannot look back on a unified development, but on disparate traditions in the individual disciplines. While revolutionary violence, for example, is discussed primarily in the context of the French Revolution, research on social protest has developed under the influence of 1950s Anglo-American social history and of the reactions to the 1960s urban riots in the USA. The historiography of research on violence is thus many-stranded and international.

1. Definitions of Violence

The definitions of violence are as varied as its manifestations. The term itself takes on disparate coloring in the context of each respective national language. English, French, and Italian reflect the Latin and emphasize the injuries caused by violence and its violation of corporal integrity. By comparison, the German term '*Gewalt*' is less negatively charged, also being used as a synonym for 'power' or '*pouvoir*'. Only its extension '*Gewalttätigkeit*' captures the injurious character of the Anglo-American term. A comparative examination of the history of terms is a desideratum for research.

The spectrum of scholarly definitions of violence is broad. The distinctions between direct and indirect, collective and individual, legitimate and illegitimate, concrete and structural, physical and psychological, and manifest and symbolic violence reflect varying accents in the discussion. Depending on the definition, the focus can be on the manifestations, the justifications, or the effects of violence. The term has been specified further in accordance with its intended use for qualitative or quantitative studies. Modern research on protest sees collective violence when at least 20 persons take part in acts in which persons or objects are damaged.

In the current discussion among historians, violence is widely understood as injury to people's physical integrity, caused by various historical actors in various contexts. Here, violence is not seen as an anthropological constant, nor as a universal historical trait held in common, but tied to the actions of specific groups and conditions that are subject to change in various national societies and epochs. The German sociologist, Heinrich Popitz, has advanced this view (Popitz 1986). He sees violence as 'an act of power that leads to the intentional bodily injury of others'. In this view, corporality distinguishes violence from other means of domination, such as orders, though it contributes to their effect.

Since violence injures bodily integrity, it possesses a massive potential for threat and evokes fears. These can contribute to the avoidance of acts of violence, but can also amplify them. The pair of terms 'physical' and 'psychological' violence touch on this connection. Not only does the experience of physical violence, such as torture or rape, have psychological consequences; psychological violence, such as brainwashing, can also have physical effects. Popitz' definition also points to the concept of power. It thus designates the content of violent conflicts as an often unequal structural situation of victims and perpetrators. This underscores that the history of violence is also always a part of societal processes in which the distribution and resources of power are up for discussion.

Beyond its physical core, the term violence has taken on various meanings in the course of the modern age.

(a) In the process of forming states, legitimate violence is distinguished from illegitimate private violence, independent of whether the latter is wielded in revolutions or in defense of the status quo. The state has defended and extended its monopoly on violence as a constituting component of being a state since early modern times. This monopoly prevailed through the disarmament of the feudal rulers. With the army and (since the eighteenth, and especially the nineteenth, centuries) the police, state organs exercised the monopoly on legitimate violence. This monopoly required justification in its form and degree, but not in its existence. Lindenberger and Lüdtke saw the hallmark of modernity in the network of relationships between physically suffered violence, on the one hand, and the state monopoly on violence, on the other (Lindenberger and Lüdtke 1995). Thus they not only claim an increase in the state's practice of control and repression, but also touch upon the everyday experience of violence. In the modern age, citizens can come into conflict with state regulations and sanctions, which in turn provoke violent actions. In modern protest research, Charles Tilly, among others, has held state reactions to unrest responsible for the extent of counterviolence (Tilly 1975).

(b) The experiences of dependency and inequality between the centers of power, colonies, and the so-called 'Third World' led the Norwegian social scientist, Johan Galtung, to develop the concept of 'structural violence' (Galtung 1975). From the perspective of peace politics, he drew attention to a discrepancy. He saw a chasm between the existing potential for development based on technology and available societal wealth, on the one hand, and the respective levels of development of disadvantaged groups and countries, on the other. His concept emphasizes, not individual perpetrators, but social structures, and not primarily physically-experienced violence, but poverty and exploitation. He was not so much concerned with legitimating counterviolence, as vehemently defended by the Algerian sociologist, Frantz Fanon, but with social peace, which he thought would be secured for all through equality and opportunities for development (Fanon 1961).

The term 'structural violence' flourished in the 1970s, but also met criticism. It was accused not only

of failing to define 'structure,' but also for lacking specificity in its analysis of violence. 'By tendentiously deciphering all relations as relations of violence, these relations are leveled,' wrote Wolf-Dieter Narr (1973); i.e., the analyses of violence and of society are indistinguishable. Galtung was able to demonstrate that the term 'violence' can be applied meaningfully in the analysis of acts in the context of the forcible relations effective in acts of exchange, in inculcating literacy, and in migrations.

(c) The concept of symbolic violence is beset with similar problems of distinction. Pierre Bourdieu and Claude Passeron used it in their theoretical analysis of the French school system and thus designated 'that form of violence' that 'is exercised against a social actor with the complicity of this "actor"' (Bourdieu and Passeron 1970). The experience is violent in that the actors are subjected to a code of language and behavior alien to them; it is symbolic through the verbal and semiotic form of its mechanisms. These shape the school's demands and enable or prevent scholastic careers. According to Bourdieu, by accepting this code, the actors contribute to the lasting legitimation of the scholastic system of selection. This term rightly touches on the linguistic and semiotic dimension of violence, but expands it to the point that it no longer has clear contours.

(d) In the course of the development of gender history in the 1990s, the concept of violence has been extended to private forms of violence within the family. Feminist literature has used the term 'sexual violence'. This refers to actions that appear in a form tying societies to sexuality. They are carried out against the will of persons who do not have the same power resources as the perpetrators. Sexual violence is thus analyzed as part of the inequality effective in the relationship between the sexes. It shifts from an exceptional phenomenon to one in terms of which the paradigmatically prevailing ideas of power relations and sexuality, and the experiences of women and girls can be grasped. This extension of the concept of violence not only opens new fields to scientific research, but, since the 1980s, has also been taken up in the penal codes of individual societies.

2. *Historical Research on Violence*

In the past, experience with violence was subsumed again and again in theoretical sketches. Among them, the theories of Thomas Hobbes, Max Webber, and Hannah Arendt were doubtless the most influential, though they arose under disparate conditions (Hobbes 1914, Weber 1919, Arendt 1970). But, especially since the 1950s, their influence on historical research on violence has been superseded by that of Norbert Elias, Michel Foucault, and most recently Clifford Geertz (Elias 1976, Foucault 1975, Geertz 1973).

Thomas Hobbes was the primary inspirer of historical research on violence (Hobbes 1914). In seventeenth-century English society, which was no longer integrated by religion, he demonstrated in terms of contract theory the necessity of the state's use of violence. Hobbes postulated that, to pacify an English society composed of private persons hostile to each other, these persons relinquished their rights to the Leviathan state, i.e., in the English case, to the absolute monarch. In return, they received protection from hostile others and the guarantee of societal peace. The double function of the state's monopoly on violence—on the one hand, to protect the interests of individuals, and on the other, to limit their possibilities of self-realization and their rights—has become the common possession of most analyses of violence since Hobbes. The individual's relationship to state violence thus often develops from his specific experiences either with the state's protection or with its limitation of his freedom.

Norbert Elias has provided important stimuli for historical research on violence (Elias 1976). Elias described European history since the Middle Ages as a history of civilization. This hypothesis is based on two central assumptions. First, Elias says that, in the course of the expansion of the trades, commerce, transportation, and the activity of the state, the interaction among members of society became more frequent and more dense. To the degree that they competed for influence and shares of power, agreements on behavior, and behavioral norms were required. In this process, civilized manners prevailed. On the other hand, the state monopoly exercised an increased external compulsion over the behavior of the members of a society. This found its most complete form in courtly society. In the realm regulated by courtly etiquette, the individual's success depended on new civilized modes of behavior. To the degree that peoples' definition of themselves conformed to this framework, forms of self-compulsion and control of feelings resulted. Violent structures of interpersonal societal relations receded as the model of civilized and inner-directed acting spread from the upper strata to the rest of society. Elias' historical periodization has been criticized by Hans-Peter Duerr, but his secular-based hypothesis has nonetheless been very fruitful for research (Duerr 1988–96).

Michel Foucault emphasized the shift in form and content in state practices of punishment, which shaped both societal reality and the individual's life practice lastingly (Foucault 1975). In his view, the practices of domination targeting the body have shifted since 1800 to strategies targeting the mind. The disciplinary measures carried out on the human body in clinics, psychiatric facilities, and prisons took on a more total character in the modern age. As in Jeremy Bentham's 'Panoptikum,' the individual is now subjected to control by institutions and the effects of techniques of power, which in turn are provided by the fields of

study that dominate discourses (Bentham 1787). In Foucault's model, the direct violence of the exercise of power made way for the violence of discourses—at least after the threshold between the epochs, which he locates around 1800. Historical studies have relativized this hypothesis and made it more specific, as they did with Elias' theories, but it too has retained its importance as a challenge to detailed studies.

While Elias and Foucault locate the changing use of bodily violence in epochal contexts and societal macrosituations, cultural anthropological studies have gained importance over the years, and they are no longer devoted to the causes, but to the forms of violence. Attention to the 'how' of the practice of violence, a maximum of 'dense description' of constellations of violence, and even an 'ethics of precision' (von Trotha 1997) is demanded for the description and explanation of acts of violence and their results. In referring to Clifford Geertz, recourse is taken to anthropological approaches or variants of symbolic interactionism are tested (Geertz 1973).

3. *Emphases and New Fields in Historical Research on Violence*

3.1 *An International Comparison of Violence*

Historical research on violence has so far privileged the national context and has seldom conducted international comparisons. This concentration on the nation state suggested itself, since the state monopoly on violence stood in the center of research on violence. But there are still some comparative works, based partly on national research or inspired by the latter's questions. Among these is the long-term comparison of the period from the thirteenth to the nineteenth centuries, in which P. Spierenberg describes behavior toward victims, the weak, and animals to demonstrate a general increase of empathy in Europe (Spierenburg 1991). Charles Tilly formulated a typology of collective violence in Europe, in which he names the parochial conflicts—the feuds between geographical communities (Tilly 1975). He uses the term 'reactive violence' to designate resistance against state interference in the citizens' way of life. According to Tilly, proactive violence is exercised within those organizations that acquire, and seek to defend, their political power.

All these investigations face the problem of having to work with unequal and incomplete national data. Since forms of protest and criminality take different cultural shapes in different societies, views of reality are often grasped as reality itself. In comparative research on revolution, violence has often not been accorded central importance. The only exception is Arno Mayer's study that contrasts the violence of the French and the Russian Revolutions (Mayer 2000). Finally, research on totalitarianism has pinpointed state terror as a central instrument of totalitarian systems. Dietrich Beyrau has presented a critical, detailed analysis contrasting and comparing violent state measures (Beyrau 2000). It elaborates the similarities as well as the differences between the violence of the National Socialist and that of the Stalinist systems of rule. Comparative studies permit discussion of the charged question of modern research on violence: whether private and public violence dwindles in or is compatible with the functioning of modern society.

3.2 *Collective Violence as Social Protest*

Research on protest is one of the most flourishing branches of research on violence. The extent and significance of acts of violence are examined in it. It has taken on new contours since the 1950s, under the influence of English, often Marxist-inspired social history and of the 1960s American research on urban riots. Insight into the inherent logic of popular action has supplanted the condemnation of rage-blinded, aggressive mob and crowd activity. Their violent nature has been interpreted less as a defining characteristic than as an aspect of social protests and less as irrational than as a rational strategy. Georges Rudé and Edward P. Thompson in particular have carried out this revision (Rudé 1964, Thompson 1971). Thompson has also drawn attention to the 'moral economy'. With this term, he designates the model of violence as a fitting behavior by peasants, artisans, and workers in response to the markets in times of economic crisis or famine. Acts of popular violence in Great Britain during the transition from the eighteenth to the nineteenth centuries are set in relation to this model and contrasted as a traditional alternative to the developing market society. Thompson's continuing hypotheses that these protests were directed against the gentry have meanwhile been revised. Nevertheless, building on Thompson, attention has shifted to the significance of violent protests in a strategy based primarily on negotiating and compromising with the authorities. In this context, John Bohstedt speaks of a 'protocol of riot' that the protesters strictly adhered to, in order to initiate negotiations, rather than destructive activities (Bohstedt 1988). When violence was nonetheless employed, its form was often borrowed from traditional rituals or it was applied symbolically against objects. Between 1770 and 1850 in Great Britain, France, and Germany, the majority of protests against famine, tax revolts, and resistance against conscription were usually conducted nonviolently. Following the American urban unrest, Graham and Gurr later questioned this hypothesis of violence's positive significance for negotiating processes (Graham and Gurr 1969). They underscored that violence is, as a rule, unproductive, while in Charles Tilly's interpretation of violent means as part of power struggles, they can indeed be successfully employed (Tilly 1975).

Since the mid-1980s, this still influential view of protest has been accused of attributing too much rationality to it and of denuding it of its emotionality. Alain Corbin in particular has pointed out the degree to which fears, hopes, and rumors have been responsible for acts of violence (Corbin 1990). George Lefebvre already traced plundering and acts of violence against the landed nobility in France in 1789 to the effects of alarming news and fears in the rural population (Lefebvre 1970). For his part, Corbin describes how, on a market day in August 1870, a crowd of 300–800 people tormented a young noble and then burned him alive. He had been accused of being a Prussian and of calling out 'Vive la République.' These acts, like the Grande Peur in the French Revolution, can be traced to the spread of rumors, but that does not explain their cruelty. For historical research on protest, this finding suggests that violent activity should be understood as part of a history of the emotions.

The investigation of the history of violence has not been limited to the extent and significance of acts of violence, but has also asked about its instigators and causes. The mass psychologist Gustav Le Bon has attributed mass violence primarily to women (Le Bon 1908). Historical research on protest has shown that women have taken part in acts of violence. But women remained in the minority and were more intensely active when the well-being of the family was at stake than, for example, in conflicts with the military. Beyond that, the social profile of people arrested for manifestations of violence has been worked out. They usually do not come from the 'Lumpenproletariat,' but are rural and urban wage-earners, sometimes from tradesmen milieus. The people, more than the mob, carried out these actions. Bohstedt, in particular, has pointed to disturbed societal relations as one of the causes of violence in Britain, especially where social contact and dependencies between the upper classes and the masses were attenuated (Bohstedt 1988). But this hypothesis cannot be confirmed for other protests. Whether acts of violence always followed social disintegration, and whether rituals of violence do not in turn presuppose relatively unified communities is a question for additional research. In the past, historical research on protest has concerned itself most with the phase in which civic society was formed, but now it is also examining the period between the world wars of the twentieth century and is including the problem of political violence to a greater degree. Its approaches and methods must now prove themselves in this field.

3.3 Historical Research on Violence as Crime

Violent crime in its long-term historical development, societal distribution, and meaning has been investigated primarily in terms of Europe in the Middle Ages and early Modern Age, and in terms of the twentieth-century USA. Modern historical studies of crime that also address nineteenth- and twentieth-century Europe are rare. Quantitative international comparative studies of violent crime often exhibit methodological deficiencies, because the data base in the individual countries and through time is uneven, and above all because the prevalence of recorded acts of violence determined also depends on the strength and severity of prosecuting authorities. Additionally, Eastern and Central Europe have been almost completely omitted from almost all long-term studies, due to the lack of research there or to language barriers. Nevertheless, trend analyses from a bird's-eye perspective can provide hypotheses that can then be worked out in detail or modified by historical microresearch. Spierenberg hypothesized a drop in the rate of homicide in Europe before 1850 (Spierenberg 1991); in contrast, studies of Western Europe and the USA from 1800 to the beginning of the twentieth century by Gurr and Lane find a rising rate of crime and homicide (Gurr et al. 1977, Lane 1997). These rates allegedly fell in the century from 1840 to 1940, only to rise again in the middle or end of the 1950s, but individual studies of crime in discrete nation states or cities do not confirm this chronology. American research has confirmed the hypothesis that the USA was, and remains, the most violent nation among the Western democracies. But it is still debated whether this can be traced primarily to 'frontier violence' (Hollon 1974) and its consequences.

Modern research on crime focuses on the relationship between delinquents, the penal system, and prosecution. Social history concerns itself with the social origin and situation of criminals; legal history and sociology investigate penal 'labeling,' which criminalizes specific behaviors; and institutional and political history is devoted to the construction of prosecuting bodies. Nicole Castan and Gerd Schwerhoff, among others, have conducted historical studies of cities or regions, examining premodern societies' approach to violent delinquency (Castan 1980, Schwerhoff 1991). The following general finding resulted. Until the early modern age, even homicide was an infraction that was not punished automatically by death, but it could be atoned for and expiated through rituals and payments of money. In sixteenth-century Germany, laws called for more severe punishments, but court practice was orientated more toward requiring expiatory measures benefiting the surviving relatives than toward retaliation. This assessment of violence was not surprising in a time that still accepted it as a definite part of societal practices and rituals in the premodern period. Among youth, villagers, and journeymen and masters in trades, not only were conflicts resolved by violence; initiation rites were also organized around violence.

The criminalization and limitation of violence is a process that began in the early modern age, but did not fully prevail until the full-blown modern age. For

France, and based on the discourse of elites, Muchembled reports on rapid advance of the process of civilization, locating the criminalization of homicide in the eighteenth century (Muchembled 1988). But even for the eighteenth century, doubts have been raised whether the draconian penalties prescribed can be regarded as an indicator that more severe norms prevailed, or if these penalties were intended to compensate for the weakness of prosecuting authorities. For it has been pointed out that, precisely in premodern times, legal jurisdiction and police control had not thoroughly penetrated into rural communities in a number of European societies. For the nineteenth century, Peter Gay has advocated the hypothesis that, despite the criminalization of violence, it was indeed tolerated in specific societal areas, to the degree that it remained limited to and ritualized in them (Gay 1993). He adduces the examples of the duel and students' fencing to gain scars, corporal punishment in schools and the army, and violence within marriage. The question of when which practices were prosecuted judicially and considered illegitimate violence has not yet been investigated adequately in international comparison.

3.4 Recent Approaches of Historical Research on Violence

Historical research on violence has taken on new dimensions in the newer style of the history of war and the military. The latter is concerned with military perpetrators of violence that exceeds normal military action: participation in mass executions, acts of annihilation, torture, or rape. Browning and Bartov have been exemplary in pursuing this approach in studying National Socialist police troops and the *Wehrmacht* on the Eastern front respectively (Browning 1992, Bartov 1985). Concentrating on military perpetrators opens up the analysis of wars that accompany decolonization, the Vietnam War, and warlike conflicts accompanying the collapse of the bipolar world order. Beyond the discovery of formerly unknown or repressed atrocities committed by soldiers and officers, attention has focused on analyzing the causes of such inhumane behavior. Current discussion addresses the question of whether specific images of community and/or group cohesion motivate the perpetrators.

Beyond soldiers' and officers' motivation, interaction, and acts, research is also concerned with how they deal with experiences with violence. To the degree that interest in the history of memory and experience increased, the effects of violence on postwar societies has also been researched. World War I has been more thoroughly considered than World War II. Even today, it is becoming clear that societal discourse on wars and on the legitimacy of the violence exercised in them has as large a role as direct experience of violence in dealing with the past. It turns out that there is a need to correct the one-dimensional assumptions that experiences on the front are transformed directly into a brutalization of postwar behavior.

Since violence is defined as physical injury, its analysis is also part of the history of the body. The experience of pain is not historically constant and unchanging, but is influenced by respectively specific bodily experiences and views of the body. The conditioning of the human body by the military, prisons, and insane asylums, whose importance has been underscored above all by Foucault, as well as the feelings felt and expressed thereby, are relevant as parts of the history of violence (Foucault 1975). Just as the awareness of the body, its integrity, and its injuries are interpreted culturally, the views and assessments of violent intrusions into the body—such as bodily injury, homicide, rape, or torture—depend on the respective common cultural patterns. In the future, these must be examined not only in terms of their shaping in specific epochs and by gender, but also in terms of their disparate shaping in different cultures.

All this already indicates the degree to which research on violence has profited in the late twentieth century from interest in questions of cultural history. In this context, simple attributions have become problematical. The taboo against violence is no longer seen as the dominant characteristic of the modern age, whose ambivalence Baumann has underscored (Baumann 1991). The media's conveying and staging, which depicts and overemphasizes acts of violence, have contributed to this ambivalence of the modern age. These connections are often mentioned in current diagnoses, but there is a lack of comparative historical research devoted to the respective cultures of violence in various milieus and at various times. For most epochs, not even the discursive boundaries between violence and nonviolence are well enough known. Special attention should be paid to the degree to which state authorities and the upper strata of societies have intervened in the justification of forms of violence and in its condemnation. For the respective concepts of violence all have a mediating link between the state monopoly on violence and everyday experience.

See also: Arendt, Hannah (1906–75); Elias, Norbert (1897–1990); Foucault, Michel (1926–84); Hobbes, Thomas (1588–1679); Rape and Sexual Coercion; Social Movements, History of: General; Social Movements, Sociology of; Terrorism; Torture and its Consequences, Psychology of; Totalitarianism; Violence in Anthropology; Violence: Public; War, Sociology of; Warfare in History; Weber, Max (1864–1920)

Bibliography

Arendt H 1970 *On Violence*. Harcourt Brace, New York

Bartov O 1985 *The Eastern Front, 1941–45. German Troops and the Barbarisation of Warfare*. Macmillan, London

Baumann Z 1991 *Modernity and Ambivalence*. Cornell University Press, Ithaca, NY
Bentham J 1787/1962 Panopticon or the Inspection-House. In: Bowring J (ed.) *The Works of Jeremy Bentham*. Russell & Russell Inc., New York, pp. 37–172
Beyrau D 2000 Schlachtfeld der Diktaturen *Osteuropa im Schatten von Hitler und Stalin*. Vandenhoeck & Ruprecht, Göttingen, Germany
Bohstedt J 1983 *Riots and Community Politics in England and Wales, 1790–1810*. Harvard University Press, Cambridge, MA
Bohstedt J 1988 Gender, household and community politics: Women in English riots 1790–1810. *Past and Present* **120**: 88–122
Bourdieu P, Passeron J P 1970 *La Reproduction. Élements Pour Une Théorie Du Système D'ensignement*. Editions de Minuit, Paris
Brockhaus U, Kolshorn M 1993 *Sexuelle Gewalt gegen Mädchen und Jungen, Mythen, Fakten, Theorien*. Campus, Frankfurt, Germany
Brown R M 1993 Western violence. Structure, values, myth. *The Western Historical Quarterly* **24**: 5–20
Browning C 1992 *Ordinary Men. Reserve Police Battalion 101 and the Final Solution in Poland*. Harper Collins, New York
Castan N 1980 *Justice et répression en Languedoc a l'époque des lumières*. Flammarion, Paris
Corbin A 1990 *Le village des cannibales*. Flammarion, Paris
Duerr H-P 1988–96 *Der Mythos vom Zivilisationsprozeß*, 4 Vols. Aubier, Frankfurt, Germany
Elias N 1976 *Über den Prozeß der Zivilisation, Soziogenetische und psychogenetische Untersuchungen*, 2 Vols. Suhrkamp, Frankfurt, Germany
Emsley C, Knafla L A (eds.) 1996 *Crime History and Histories of Crime. Studies in the Historiography of Crime and Criminal Justice in Modern History*. Greenwood Press, Westport, CT
Fanon F 1961 *Les Damnés de la Terre*. Maspero, Paris
Foucault M 1975 *Surveiller et punir. Naissance de la Prison*. Paris
Gailus M 1990 *Straße und Brot. Sozialer Protest in den deutschen Staaten unter besonderer Berücksichtigung Preußens*, 1847–1849. Vandenhoeck & Ruprecht, Göttingen, Germany
Galtung J 1975 *Strukturelle Gewalt. Beiträge zur Friedens- und Konfliktforschung*. Rowohlt, Reinbek, Germany
Gay P 1993 *The Cultivation of Hatred*. New York
Geertz C 1973 *The Interpretation of Culture*. Basic Books, New York
Graham H D, Gurr T R (eds.) 1969 *The History of Violence in America*. Praeger, New York
Gurr T R et al. 1977 *The Politics of Crime and Conflict*. Sage, Beverly Hills, CA
Hobbes T 1914 *Leviathan*. J. M. Dent & Sons, London
Hollon W E 1974 *Frontier Violence. Another Look*. Oxford University Press, New York
Julia D 1995 La violence des foules; peut-on élucider l'inhumain? In: Boutier P, Julia D (eds.) *Passés recomposés. Champs et chantiers de l'histoire*. Editions Autrement, Paris, pp. 208–23
Kocka J, Jesson R 1990 Die abnehmende Gewaltsamkeit sozialer Proteste vom 18. zum 20. Jahrhundert. In: Albrecht P A, Backes O (eds.) *Verdeckte Gewalt*. Suhrkamp, Frankfurt, Germany, pp. 33–57
Lane R 1997 *Murder in America; A History*. Ohio State University Press, Columbus, OH
le Bon G 1908 *Psychologie des Foules*, 13th edn. Alcan, Paris
Lefebvre G 1970 *La Grande peur de 1789*. Colin, Paris
Lindenberger T, Lüdtke A 1995 Physische Gewalt—eine Kontinuität der Moderne. In: Lindenberger T, Lüdtke A (eds.) *Physische Gewalt*. Suhrkamp, Frankfurt, Germany
Mayer A J 2000 *The Furies. Violence and Terror in the French and Russian Revolution*. Princeton University Press, Princeton, NJ
Muchembled R 1988 *L'invention de l'homme moderne. Culture et sensibilités en France du Xve au XVIIIe siècle*. Fayard, Paris
Narr W-D 1973 Gewalt und Legitimität. *Leviathan* **1**: 7–41
Popitz H 1986 *Phänomene der Macht. Autorität—Herrschaft—Gewalt—Technik*. Mohr, Tübingen, Germany
Rudé G 1964 *The Crowd in History, France and England*. Wiley, New York
Schumann D, 1997 Gewalt als Grenzüberschreitung. Überlegungen zur Sozialgeschichte der Gewalt im 19. und 20. Jahrhundert. *Archiv für Sozialgeschichte* **37**: 366–87
Schwerhoff G 1991 *Köln im Kreuzverhör. Kriminalität, Herrschaft und Gesellschaft in einer frühneuzeitlichen Stadt*. Bouvier, Bonn, Germany
Spierenburg P 1991 *The Broken Spell*. Macmillan, London
Thompson E P 1971 The moral economy of the English crowd in the eighteenth century. *Past and Present* **50**: 76–136
Tilly C 1975 *The Rebellious Century*. Harvard University Press, Cambridge, MA
von Trotha T 1997 Zur Soziologie der Gewalt. In: von Trotha T (ed.) *Soziologie der Gewalt* (KZfSS, 37). Kölner Zeitschrift für Soziologie und Sozialpsychologie, Opladen/Wiesbaden, Germany, pp. 9–56
Weber M 1919 *Politik als Beruf*. Duncker & Humblot, Munich, Germany

H.-G. Haupt

Violence in Anthropology

Across the Cold War and the collapse of great-power socialism, anthropological studies of violence have focused on international patterns of revolution, counterinsurgency warfare, state terrorism, militant ethnic nationalism, communal violence, and organized crime (Wolf 1969, Das et al. 2000). The field has also reconsidered the violence of Western colonialism, the epistemic violence of states that marginalize devalued groups, and the structural violence inherent in systems of inequality that consign great numbers of people to unstable incomes and life-threatening poverty, hunger, and preventable disease (Schepper-Hughes 1992, Bourgois 1995).

This array of issues contrasts with social anthropology's earlier functionalist focus on the internal regulation of social tensions, leadership transitions, and intergroup hostilities in lineage-based societies, imagined as isolates. The turn to examining the interplay of local communities and state politics, and the political consequences of globalization has revitalized the subfield of political anthropology.

Although the discipline has appropriated issues from political science and comparative sociology

throughout this scholarly transformation, it has also critiqued central assumptions in these fields and reasserted the importance of cultural issues in the study of violence. As a result, anthropology is contributing to a rethinking of the state in the context of globalization, and reconsidering the interface between state and civilian populations. Most importantly, it has taken on the cross-cutting question of how communities in different political situations cope with their own patterns of internalized violence.

The scope of the anthropological project has grown as researchers questioned the conventional social science emphasis on violence as the state's monopoly over the use of legitimate force by the armed forces, police, and judicial systems (Giddens 1987). By contrast, cultural anthropology has chosen not to privilege uniquely formal institutions or physical harm in its studies, but rather additionally to examine the social fragmentation and suffering created by organized, but often informally instituted, structures of violence generated in situations of civil war, state terrorism, fragmented states, and globalized economies.

In Latin America, Asia, and Europe, a central issue has been the ways in which local communities both resist and become complicit with state violence. In regions of weak states where alternative power structures have assumed state functions, as in the case of the former USSR and sub-Saharan Africa, a major issue is tracing the growing significance of illicit trading networks and mafias.

Anthropology has embraced notions of culture that include the diverse discursive practices that states have used to legitimize the use of violent force by portraying others as inherently violent or not fully human, and thus deserving of preemptive control rather than full citizenship. The discipline has examined alternative forms of resistance to state violence, the politicization of ethnicity, and the impact of counterinsurgency states on civil populations. It has also begun to analyze popular representations of violence in the media.

1. *Violence and Resistance*

Rather than ignoring those without power and status, one well-established anthropological approach examines patterns of cultural resistance to domination and oppression. As has become clear in this literature, resistance in oppressive circumstances takes many highly coded forms. Resistance perspectives examine the way groups configure ideologies and political actions to reject the terms established for subordinates by state authorities and regional power brokers, even in extreme circumstances when they are political prisoners (Aretxaga 1997). A tactical dual consciousness allows submissive subordinates to feign submission yet hold subversive beliefs in their behind-the-scenes 'hidden transcripts' (Scott 1990).

This approach is especially concerned with cultural forms through which local populations keep their veiled social critiques alive: such as oral histories of martyrdom, ritual parodies of power structures, and political art. These cultural forms assert alternative moralities, unveil abuses of power, and describe forms of militancy, great and small. Such codes of resistance cannot be read from afar, that is without knowledge of particular cultural circumstances and access to how people rework cultural forms to deal with the particular existential dilemmas they face.

2. *Violence and Leadership: A Critique of Rational Choice Reductionism*

Another line of analysis, influenced by political science, pursues the motivation of rivals for state power, particularly those who defy political systems and seek to undermine the integrity of the state by organizing separatist rebellions. Researchers have sought to understand political violence in these circumstances by using rational choice theory, and studying the competition of groups for resources and the self-interests of political leaders. Analysts now routinely question political leaders' self-declarations, and unmask opportunists willing to politicize ethnicity and inflame populations to their personal advantage. Leaders exhibit a willingness, sometimes a frightening eagerness, to promote the radical fracturing of society into antagonistic groups with the goal of unifying their own group through strategic antagonisms.

Anthropology suggests that violence and conflict should not be seen as exceptional moments of crisis and rebellion, or simply as the consequences of competitions over resources, but as integral parts of a social fabric already fragmented in innumerable ways. Ethnographic research shows how important it is to comprehend people's understandings of the contradictory tensions, and the heterogeneous personal and collective interests in their lives. Struggles over memory and history, the importance of culturally specific narratives for the expression of grievances, and the local constructions of 'choice,' 'competition,' and 'opportunism,' are all doorways to a fuller cultural analysis. At issue in moments of crisis are the intensification and redirection by leaders of what is already there—the systems of meaning that actively generate cultural understandings.

The mix of opportunism and resistance approaches chosen for interpreting a particular violent conflict depends on a range of factors, including the historically and politically conditioned gaze of researchers working in regional studies traditions. With its long colonial history, decades of military rule from the 1950s to the 1980s, anti-imperialist guerrilla movements that sought to topple elite-driven states, revolutionary socialist experiments, and current strug-

gle to institutionalize democracy, Latin America has drawn the attention of a variety of analysts.

In these contexts, ethnic identities have been shaped by racially inflected modes of stratification, deep cultural resistance, and political mobilization that has had complex relations and tensions with capitalist and socialist ideologies (Hale 1994). In the case of the Zapatista uprising in Mexico—which protested entrenched regional violence and the subsistence shocks of the 1994 neoliberal trade agreements between the USA and Mexico that established the North American Free Trade Agreement (NAFTA)—politicized indigenous identity has been the source of alternative visions of social justice, combining international human rights discourses with the communal politics of Maya communities (Nash 2001). It is hard to find the reductionist explanations of rational choice theory convincing in these complex circumstances.

3. *Death Squads*

Government-mandated counterinsurgency violence that seeks to extinguish political opposition is often an unstable solution to social conflict, yet it has been the common practice of regimes across the political spectrum. Violent states have persisted in labeling wide ranges of their citizens as dangerous and undeserving of full rights. Nationalisms have been crafted by political elites not only to create an imagined community for their country's citizens but additionally to rationalize mass repression much wider in scope than any immediate uprising.

Anthropologists have documented an often hidden and always denied aspect of violent states: death squads operate at the margins, engaging in torture, kidnap, rape, and assassination outside the rule of law. Paramilitary groups specialize in what the police and armed forces officially appear to eschew: the extrajudicial torture and killing of dissenters, the disappearances of local leaders, students, and journalists, and the retaliatory abuse of activists' families, including the rape of women to humiliate their parents and to pollute family lines. The political goal is to silence critics and intimidate the general public through preemptive violence, while maintaining plausible deniability of government involvements in violence against civilians. Some militaries have instituted community-based civil patrols, forced to engage in death squad activities and community surveillance, all in the name of self-defense. There are many variations of this violent mode of control, responding to the particular political and cultural histories of countries as diverse as Northern Ireland, Spain, India, Indonesia, Argentina, and Guatemala (Montejo 1999, Sluka 2000).

It is clear from these studies that the police and army are not bounded institutions set apart from society. Rather, the state, armed forces, paramilitaries, insurgents, and civil society form interpenetrating social fields. Civilians fill the ranks of armed forces on all sides and, often as forced recruits, experience state-centered indoctrination to fight these wars. The general population, whatever its sympathies, frequently finds no feasible exit from the conflict between insurgents and the state (Stoll 1993). Individuals become complicit in state violence by informing on the activities and putative alliances of their neighbors. Whether coerced, offered in exchange for personal gain, or solicited to enhance one's power at the expense of others, these complicities involve betrayals that splinter family and community trust, and create situations of internalized violence that add new tensions to old divisions (Warren 1993, Das et al. 2000). These dynamics create powerful existential dilemmas for people who must endure heightened insecurity over long periods of time. As E. Scary (1985) observed, the deeper cost of violence is not its physical harm but its capacity to 'unmake the world.'

In the face of state terrorism, anticolonial opposition movements pose particular challenges for anthropologists. In instances where cultural minorities or grassroots movements are repressed by overwhelming state power, it has been difficult for anthropologists not to sympathize with those who denounce and resist injustice. For some researchers, the fact of war justifies counterviolence, whatever form it takes. For others, the issue is how repression potentially recreates a sense of common purpose and unites resistance across ethnic, geographic, and class lines.

For still other analysts, resistance to repression involves a painful blurring of motives and allegiances over time, with individuals betraying their established loyalties rather than bridging them, sometimes voluntarily or in other circumstances to put off their own death sentences. Anthropologists have traced the contradictions, ironies, and the mimesis of state violence in anticolonial separatist movements. As Ortner (1995) argues in her critique of resistance approaches, anthropologists are called upon to contextualize and historicize (rather than ignore) the contradictory impulses of such social movements. A return to richer ethnographic portrayals has become necessary to counteract the earlier tendency to idealize opposition movements by focusing on their activities as exemplars of resistance.

Beyond the horror of state violence, there is no neutral subject position from which to narrate these histories. That counternationalisms can generate their own authoritarianism, death squads, and corruption-mirroring state power as they critique it—and that grassroots insurgents may lose rather than gain civilian support over time create special challenges for anthropological writing in politically charged situations. Another compelling issue is the criminalization of former combatants where peace processes have not coped with the extreme social fragmentation that accompanies protracted violence. Former com-

batants, death squad members, civil patrollers, and families displaced by violence need new sources of employment and community involvement if they are not to continue patterns of predatory behavior.

Perhaps the triumphalist discourse of democratic modernity—especially the post Cold War end-of-history variant of this discourse—has inhibited closer examinations of violence in democratic societies. Tambiah (1996) argues that paramilitary violence may be much more intrinsic to democratic states than is usually suspected. In practice, democracy channels intense competition for power through electoral politics which is often accompanied by occult violence. Thus, in India, groups frequently provoke riots at election time for their own political ends. Echoing these patterns, elections in Guatemala and Colombia bring waves of assassinations of candidates for local and national office. Although these riots and cyclic killings are condemned, electoral violence becomes an anticipated event in these political landscapes. The underside of democracy includes spaces of violence that are integral aspects of state politics, even as they are, in effect, public secrets.

4. *Interpenetrating Explanations of Violence*

How does anthropology weight the factors ultimately accountable for contemporary violence? First of all, it is important to consider international influences such as the highly volatile global economy and neoliberal economic policies promulgated by the international development community as part of democratic modernity that creates and tolerates dramatic gaps between prosperous and impoverished citizens. Policies and investment decisions at great distance from their effects continue to bring subsistence shocks that threaten the lives of the poor. The Asian financial crisis of 1997, for example, which began in the realm of international currency speculation, precipitated the collapse of several national currencies and generated riots and continuing ethnic tensions in Indonesia.

Political transnationalism would be another issue. Counterinsurgency and insurgency training, and the sale of arms and surveillance equipment have long been global industries that are increasingly decentralized in practice. The School of the Americas, a US-based training center for Latin American military officers who have been implicated in many human rights abuses in their own countries, has been the conduit for techniques of violent control developed over the longer history of Western colonialism. Strategies first developed in colonial Asia were brought to Latin America in the 1950s, redeployed in Vietnam during the war, and later put back into circulation in Latin America, moving in the last decade from Guatemala to Mexico, where they have been used most recently against the Zapatistas. British colonial models of control developed in Kenya and Malaya have been redeployed in Northern Ireland. The post Cold War repositioning of militaries as partners in national development has left unchallenged the discourses of 'national security,' in which state terrorism can be used to combat whatever is discursively defined as a political threat.

Another crucial element for the explanation of violence is the national institutions and political actors who, for their own reasons and interests, seek to foster authoritarian, militarized solutions to social conflicts. Finally, it is important to include the interplay of local and national formations of violence to trace the intimate interplay of transcommunity organized violence with community conflicts.

To neglect any of these arenas, or to locate the moral or causal explanation of death squads in particular international arenas so that it displaces all other considerations, would appear to be a serious misreading of a complex political field of actors and institutions. Nevertheless, it is clear that the study of patterns of violence shows the direct linkage of very local events with wider power structures and globalized currents of change. Capitalist economics, which has led to the search for ever cheaper sources of labor without wider obligations for workers' social welfare, and the transnational sex trade, which transforms adults and children into commodities, exemplify the social costs of globalization.

5. *Violence and Media Representation*

Current research challenges the usefulness of media accounts of 'ancient ethnic hatreds' to explain conflicts leading to the fragmentation of states. There is a consensus now that Western media and official media elsewhere promote dangerously simplified ethnic readings of conflicts, ones that reinforce primordialist readings of 'the other.' Such presentations in mass media deflect attention from the reality that the West has been directly involved in many 'localized' conflicts through economic and political interventions in the internal affairs of other nations (Allen and Seaton 1999).

In contrast to the overdetermined language of ethnic antagonism, recent research in anthropology demonstrates that many social systems involve the coexistence of diverse groups and hybrid identifications, and that politics in various parts of the world cross-cuts ethnic, religious, and linguistic difference. An important deconstruction of ethnicity as a fixed or essential entity is occurring across these analyses. The classical critique has been to discredit ethnicity in order to uphold class as the prime factor to explain social conflict. A more recent countertrend is to widen the field of identities and identifications under consideration, and to understand class as a creation of particular cultural and historical circumstance, rather than a preestablished or transcultural set of categories.

A fuller cultural analysis, which pursues the multiplicity of identities and the flow of ideas and signs across borders, is central to interpreting conflict and the directions it takes. The issue is not just meanings that mediate reality, but the images that make imagination—in this case violence, fragmentation, and/or coexistence—possible. Instead of assuming that people internalize the ethnic hatred and opportunist politics sold by corrupt leaders and the media—an assumption left unchallenged in US journalism by repeated photographs of armed mobs without any sense of people's reasons for their participation—it is important to understand the playing out of contradictory consciousness in people's daily lives, and the political consequences of these understandings. To make sense of low-intensity peace or nationalist crises, one must historicize and contextualize the forms of violence and its denial.

Culturally informed analyses call for an array of strategies for studying representation, control, resistance, and complicity—the violence within and without. Significantly, these same strategies can be used to explore unexpected alliances and unexpected avenues for negotiation that might successfully challenge violence and its grammar of oppositions.

See also: Conflict Sociology; Ethnic Cleansing, History of; Ethnic Conflicts; Terrorism; Violence and Media; Violence, History of; Violence: Public; War, Sociology of; Warfare in History

Bibliography

Allen T, Seaton J 1999 *The Media of Conflict: War Reporting and Representation of Ethnic Violence*. Zed Books, London

Aretxaga B 1997 *Shattering Silence: Women, Nationalism, and Political Subjectivity in Northern Ireland*. Princeton University Press, Princeton, NJ

Bourgois P 1995 *In Search of Respect; Selling Crack in El Barrio*. Cambridge University Press, New York

Das V, Kleinman A, Ramphele M, Reynolds P (eds.) 2000 *Violence and Subjectivity*. University of California Press, Berkeley, CA

Giddens A 1987 *The Nation-state and Violence*. University of California Press, Berkeley, CA

Hale C R 1994 *Resistance and Contradiction: Miskitu Indians and the Nicaraguan State, 1894–1987*. Stanford University Press, Stanford, CA

Montejo V 1999 *Voices from Exile: Violence and Survival in Modern Maya History*. University of Oklahoma Press, Norman, OK

Nash J C 2001 *Mayan Visions: The Quest for Autonomy in an Age of Globalization*. Routledge, New York

Nordstrom C, Martin J (eds.) 1992 *Paths to Domination, Resistance and Terror*. University of California Press, Berkeley, CA

Ortner 1995 Resistance and the problem of ehnographic refusal. *Comparative Studies in Society and History* **137**(1): 173–93

Scary E 1985 *The Body in Pain: The Making and Unmaking of the World*. Oxford University Press, New York

Schepper-Hughes N 1992 *Death Without Weeping: The Violence of Everyday Life in Brazil*. University of California Press, Berkeley, CA

Schirmer J 1998 *The Guatemalan Military Project: A Violence Called Democracy*. University of Pennsylvania Press, Philadelphia

Scott J C 1990 *Domination and the Arts of Resistance; Hidden Transcripts*. Yale University Press, New Haven, CT

Sluka J (ed.) 2000 *Death Squad: The Anthropology of State Terror*. University of Pennsylvania Press, Philadelphia

Stoll D 1993 *Between Two Armies in the Ixil Towns of Guatemala*. Columbia University Press, New York

Tambiah S 1996 *Leveling Crowds: Ethnonationalist Conflicts and Collective Violence in South Asia*. University of California Press, Berkeley, CA

Taussig M 1992 *The Nervous System*. Routledge, New York

Warren K B (ed.) 1993 *The Violence Within: Cultural and Political Opposition in Divided Nations*. Westview Press, Boulder, CO

Wolf E R 1969 *Peasant Wars of the Twentieth Century*. Harper Torchbook, New York

K. B. Warren

Violence: Public

Visibly violent social processes simultaneously attract and repel analysts of human behavior. War, rape, genocide, assassination, assault, football, looting, official execution, collective suicide, mutual flagellation, automobile collisions, airline crashes, gang fights, and struggles pitting police against demonstrators all qualify by some definitions as public violence. Most of them regularly attract headlines when they occur—the more extensive, public and grisly their consequences the bigger the headlines. Violence makes the news.

In these circumstances, reporters and readers commonly turn to very simple explanations: some humans are violent in nature, others peaceable, so violent events occur when violent people congregate and enjoy free rein. Unfortunately for that sort of explanation, the boundary between violent and peaceable people actually blurs badly. Even individuals who participate energetically in violent events generally spend most of their time and energy on nonviolent activities. Meanwhile persons who live mostly peaceful lives occasionally get involved in organizing or perpetrating violent activities, for example, by doing military service, serving on juries, or engaging in contact sports. Other people who flee violent interactions in their daily routines nevertheless support one variety of public violence or another in the proper circumstances, for example, by voting for candidates who favor the death penalty, by cheering one group of nationalists in their competition with rivals or by cooperating in what they call a just war. Indeed, most citizens of most states distinguish sharply between coercive acts committed lawfully by duly constituted

agents of their own states (legitimate force) and similar acts committed by other people (illegitimate violence). Nonviolence is a beguiling principle, but one honored more in the breach than in the observance.

In general discussions of the subject, three competing views of violence prevail: as propensity-driven behavior, as instrumental interaction, and as cultural form. Treatments of violence as propensity-driven behavior locate its causes within the actor, calling attention to genetic, emotional, or cognitive peculiarities that incline a given individual, group or category of persons, to damaging behavior more than others. Portrayals of violence as instrumental interaction characterize everything from petty assaults to all-out warfare as means (however inefficient and self-defeating) to power, wealth, prestige, or other ends. To call violence a cultural form—as in the claim that because of the frontier, slavery, or capitalist competition the USA has an exceptionally violent culture—argues that the ready availability of certain ideas, practices, models, and beliefs itself promotes violent action.

The three views suggest very different approaches to inhibiting or promoting violence. Propensity-driven violence calls for the alteration or inhibition of motives, instrumental violence for shifts of incentives and outcomes, cultural violence for teaching new ideas, practices, models, and beliefs. Alas, none of these three broad approaches to violence has yet yielded systematic explanations of the genesis, trajectory, outcome, or sheer variety of public violence as a whole. Perhaps we await the social scientific Isaac Newton who will detect underlying regularities in all the diverse phenomena people call violent. More likely the notion of violence resembles such other widespread but diffuse terms as disorder and goodness: morally powerful categories that do not refer to any causally coherent domain.

In any case, the word 'violence' almost always arrives with baggage of disapproval. The distinction of violent from nonviolent social interactions usually depends on a moral boundary, or at least activates one. Precisely because of that moral charge, advocates of one cause or another frequently extend the term 'violence' to varieties of harm far outside the range of direct, short-term infliction of physical damage on one person by another. They label hate speech, pornography, poverty, unemployment, malnutrition, and other ills as forms of violence. To call a phenomenon virtual violence is to condemn it, to claim that it damages something valuable even if it produces no physical destruction in the short run. As a consequence, every social scientific attempt to delimit, describe, classify and explain public violence generates controversy. This attempt will be no exception.

We can nevertheless discipline the inquiry by adopting a restrictive definition of the phenomenon under analysis. Public violence, for present purposes, refers to episodic social interaction that:

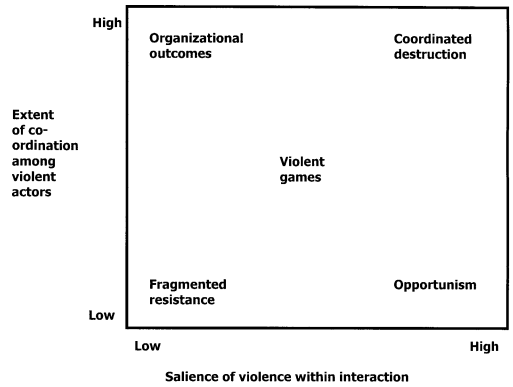

Figure 1
Crude typology of public violence

immediately inflicts physical damage on persons and/or objects

results at least in part from coordination among persons who perform the damaging acts

occurs at least partly in publicly accessible spaces.

(The ensuing discussion will, however, extend 'physical damage' to forcible seizure of persons or objects over restraint or resistance.) At one edge, such a definition excludes strictly individual, private, impulsive, and/or accidental damage to persons or objects. At the other, it excludes long-term, incremental damage such as communication of infectious disease, cumulative wear and tear, exposure to toxic substances, and death hastened by neglect or social pressure. Nevertheless, it includes an immense range of social interactions, from minor scuffles among rival groups of voters to wholesale civil war. Despite a century or more of effort, social scientists have not so far assembled any reliable, important and systematic body of knowledge that applies to the whole range thus defined.

We therefore have no choice but to differentiate, reviewing analyses of narrower phenomena that bear some family resemblance to each other, probably share some causal processes, and have attracted sustained scholarly attention. Assuming episodic interaction in publicly accessible space, at least a little communication among actors, and some minimum of physical damage to persons or objects, Fig. 1 combines the extent of coordination among damage-doers and the salience of damage in the overall pattern of interaction to produce a crude typology of public violence. ('Low' coordination means individualized action with little collective planning and signaling, while 'High' coordination means activation of differentiated and bounded organizational structure with extensive planning and signaling. 'Low' salience of violence means that most of the social interaction involved produces no short-term physical damage,

while 'High' salience means interaction strongly concentrated on damage.)

Other distinctions come immediately to mind. Separation of symmetrical from asymmetrical interaction, for example, usefully distinguishes relatively even contests from situations in which a gang or an oppressor visits violence on a victim having no means of retaliation. Whether agents of a constituted government participate in public violence, as we shall see, significantly affects its character; thus we might also classify public violence by the presence or absence of governmental agents. Again, public violence probably incorporates scale effects, so that damaging interaction involving just two or three people differs qualitatively from otherwise similar interaction involving thousands of people at a time. The present classification gives priority to coordination and salience on the hunch that those two dimensions capture a significant part of the systematic variation about which social scientists have produced useful ideas and findings.

Within the coordination–salience space we can conveniently distinguish five locations:

(a) organizational outcomes: various forms of collective action generate resistance or rivalry to which one or more parties respond by actions that damage persons and/or objects; examples include demonstrations, protection rackets, governmental repression, and military coups, all of which frequently occur with no more than threats of violence, but sometimes produce physical damage;

(b) coordinated destruction: persons or organizations specialized in the deployment of coercive means undertake a program of damage to persons and/or objects; examples include war, genocide, collective self-immolation, and some kinds of terrorism;

(c) opportunism: as a consequence of shielding from routine surveillance and repression, individuals or clusters of individuals use immediately damaging means to pursue generally forbidden ends; examples include looting, gang rape, revenge killing, and some sorts of military pillage;

(d) fragmented resistance: in the course of widespread small-scale and generally nonviolent interaction, a number of participants respond to obstacles, challenges, or restraints by means of damaging acts; examples include sabotage, clandestine attacks on symbolic objects or places, assaults of governmental agents, and the more immediately damaging forms of what James Scott (1985) calls 'weapons of the weak';

(e) violent games: two or more relatively well-defined and coordinated groups follow a known interaction script entailing the infliction of damage on others as they compete for priority within a recognized arena; examples include gang rivalries, contact sports, some election battles, and some struggles among supporters of teams or stars.

Of course these crude types overlap, for example in the activities of the South Asian thugs who sometimes act as enforcers for landlords in disputes with squatters (organizational outcomes) but at other times follow their patrons in large-scale raids on their caste, religious, or political enemies (coordinated destruction). Similarly, the various actions known as terrorism show up in all four corners of Fig. 1—as clandestine resistance, as a by-product of peaceful claim-making in highly repressive regimes, as part of genocide, and as opportunistic elimination of old enemies. It is nevertheless useful to retain the five-part division, because within each area thus delineated social scientists have accumulated at least a modicum of systematic knowledge. Let us therefore review them one by one. The following discussion gives extensive attention to organizational outcomes and coordinated destruction, then uses analogies and comparisons with the first two as it treats opportunism, fragmented resistance, and violent games much more summarily.

Organizational outcomes. A significant share of public violence occurs in the course of organized social processes that are not in themselves intrinsically violent. That is notably the case in collective political struggle. Political regimes differ dramatically in the scope they allow for nonviolent collective making of claims, for example by petitioning, shaming, marching, voting, boycotting, striking, forming special-interest associations and issuing public messages. On the whole, democratic regimes tolerate such claim-making more readily than do their undemocratic neighbors; that is one way we recognize a regime as democratic. Even in democratic regimes, nevertheless, such forms of collective claim-making occasionally generate open violence. That occurs for three main reasons.

First, every regime empowers agents—police, troops, headmen, posses, sheriffs, and others—to monitor, contain and, on occasion, repress collective claim-making. Those agents always have some means of collective coercion at their disposal, and always enjoy some discretion in the use of those means. In one common sequence, claimants challenge repressive agents, occupy forbidden premises, attack symbolically significant objects, or seize property, then agents reply with force. Because variants on that sequence frequently occur, when repressive agents are at hand they actually perform the great bulk of the killing and wounding that occurs in public violence.

Second, collective claim-making often concerns issues that sharply divide claimants from regimes, from powerful groups allied with regimes, or from rival groups; examples are campaigns to stop current wars, outlaw abortion or expel immigrants. In these circumstances, offended parties often respond with counterclaims backed by force, whether governmental or nongovernmental.

Third, in relatively democratic regimes an important share of collective action centers not on specific programs but on identity claims: the public assertion that a group or a constituency it represents is worthy, united, numerous, and committed (WUNC). Asser-

tions of WUNC include marches, demonstrations, mass meetings, occupations of plants or public buildings, vigils, and hunger strikes. Even when the means they adopt are currently legal, all such assertions entail implicit threats to direct WUNC energy toward disruptive action, implicit claims to recognition as valid political actors, and implicit devaluation of other political actors within the same issue area. These features sometimes stimulate counter-action by rivals, objects of claims, or authorities, with public violence the outcome.

Organizational outcomes also include some encounters that do not begin with concerted making of claims. Border guards, tax collectors, military recruiters, census-takers, and other governmental agents, for example, sometimes generate intense resistance on the part of whole communities as they attempt to impose an unpopular measure. Similarly, audiences at theatrical performances, public ceremonies, or executions occasionally respond collectively to actions of the central figures by attacking those figures, unpopular persons who happen to be present or symbolically charged objects. By and large, organizational outcomes connect with issues over which groups are also currently contending in nonviolent ways.

One subclass of organizational outcome, however, displays a rather different pattern. Some organizations specialize in controlling coercive means, threatening to use those means if necessary, but seeking compliance without violence if possible. Examples include not only established agents of repression but also *mafiosi*, racketeers, extortionists, paramilitary forces, and perpetrators of military coups. When such specialists in coercive means encounter or anticipate resistance, they commonly mount ostentatious but selective displays of violence. Their strategy resembles that of many Old Regime European rulers, who lacked the capacity for continuous surveillance and control of their subject populations, but often responded to popular rebellion with exemplary punishment—rounding up a few supposed ringleaders, subjecting them to hideous public executions and thus warning other potential rebels of what might befall them. The strategy is most successful, ironically, when specialists in coercion never actually have to deploy their violent means.

Coordinated destruction differs in precisely that regard. It refers to those cases where persons or organizations specialized in the deployment of coercive means undertake a program of actions that damage persons and/or objects. Coordinated destruc tion overlaps with organizational outcomes because specialists in coercion participate in both and because threats of force sometimes escalate into struggles between coercive organizations. Here, however, the organizations' strategies center, however temporarily, on the production of damage. Examples include war, genocide, collective self-destruction, public penance, and government-backed terror. The major distinctions separate (a) lethal contests, (b) campaigns of annihilation, and (c) ritual harm.

In lethal contests, at least two organized groups of specialists in coercion confront each other, each one using harm to reduce or contain the others' capacity to inflict harm. War is the most general label for this class of coordinated destruction, but different variants go by the names of civil war, guerrilla, low-intensity conflict, and conquest. Although lethal contests of various sorts stretch back as far as humanity's historical record runs, the standard image of two or more disciplined national armies engaged in destroying each other within generally accepted rules of combat only applies to a very small historical segment: roughly 1650 to 1950 for Europe, a few much earlier periods for China, even rarer intervals elsewhere in the world. Outside of those exceptional moments, autonomous raiding parties, temporary feudal levies, mercenary assemblages, bandits, pirates, nomads doubling as cavalry, mobilized villages, and similar conglomerate or part-time forces have fought most historical wars.

Lethal contests shade over into campaigns of annihilation when one contestant wields overwhelming force or the object of attack is not an organization specialized in the deployment of coercive means. Analysts have employed the term genocide for those campaigns in which attackers define their victims in terms of shared heritage and politicide for those in which victims belong to a common political category; so far no commonly accepted term has emerged for similar campaigns aimed at members of religious or regional categories. The usual stakes in campaigns of annihilation are collective survival, on one side, and recognition as the sole party with the right to territorial control, on the other. Because of those stakes, such struggles tend to generate vast mobilizations of support extending far beyond the specialists in coercion who initiate them.

Ritual harm plays a much smaller part in contemporary public violence than do lethal contests and campaigns of annihilation. At times, however, they have assumed considerable importance, for example in the potlatch (aggressive, competitive destruction of the performer's own symbolically charged property) that once marked struggles for precedence along America's northwest coast. Other well known instances include the processions of flagellants once organized regularly by Renaissance Italy's penitent confraternities. Today self-immolation, collective suicide, and hunger strikes occasionally occur as gestures of protest, serving chiefly to dramatize the positions of groups that lack other political resources.

Within the broad category of coordinated destruction, a remarkable change occurred during the half century following World War II. Where interstate wars among well-identified national armies had predominated for a century or more, shifts toward a much wider variety of coercive forces and toward campaigns of annihilation elevated civil war, broadly defined, to

the chief setting of coordinated destruction. Decolonization, expansion of world trade in arms and drugs, reappearance of mercenary forces, and the weakening of central state capacity in many world regions all contributed to that change.

Opportunism combines low levels of coordination with high salience of violence within the total interaction. Opportunism benefits from some of the same processes that have favored recent moves away from interstate war toward violent struggle within state boundaries. Looting, gang rape, revenge killing, and some sorts of military pillage all occur more frequently in interstices, peripheries, and lapses of central state control. In fact, opportunistic public violence often occurs as an outgrowth of coordinated destruction. Disbanded but still armed troops, for example, frequently use their arms to prey upon civilians. The destruction wrought by mercenaries during truces, marches, and aftermaths of wars made them the bane of European civilian life during the sixteenth and seventeenth century heyday of mercenary armies. Similarly, when major protests or disasters divert the attention of repressive forces, looters and killers often take advantage of relaxed control to pursue their own ends.

Fragmented resistance involves segmented or weakly coordinated damage to persons or objects in the course of interactions that are not predominantly violent. When governmental agents, landlords, employers, religious authorities, or other powerful people impose innovations, exactions or constraints on subject populations, members of those populations often resist indirectly and on the small scale. They engage in sabotage, ambushes, and clandestine attacks on symbolic objects or places, and temporary acts of resistance that either inflict damage on authorities or incite violent reprisals from authorities. Fragmented resistance differs from organizational outcomes chiefly in the balance of forces among parties and the scale of organization among the weaker parties. Otherwise, the same regularities that apply to organizational outcomes apply to fragmented resistance.

In its most stylized forms, such as the medieval tournament, coordinated destruction comes to resemble violent games. Likewise, such violent games as ritual fights between youths of neighboring villages edge over into coordinated destruction. Still, we can usefully distinguish a class of public violence in which two or more relatively well defined and coordinated groups follow a known interaction script entailing infliction of damage on others as they compete for priority within a recognized arena. Age grades of local youths, organized gangs, sports teams, supporters of powerful or popular figures, electoral factions and fraternal orders inducting new members all sometimes engage in violent games. Frequently the outcome of the contest produces confirmation or alteration of a rank order that in turn entitles winners and their followers to prestige or privilege. In that sense, violent games stand somewhere between (a) WUNC performances (whether violent or not) and (b) the public displays of wealth, power and patronage by which warlords, crime bosses, political magnates, heads of influential families, *caciques*, and entertainment stars maintain their credibility.

In all its forms, public violence interacts closely with nonviolent politics. Violence as organizational outcome or coordinated destruction clearly entails struggles for power and takes shape from the character of power struggles spanning violent and nonviolent action. Violence as opportunism and fragmented resistance often occurs as a by-product of grand political processes: state formation, revolution, war, military conscription, taxation, patron–client realignment. Violent games follow their own rules, but frequently intersect with or mutate into other forms of public violence. At first glance, violence may seem the antithesis of orderly public politics. At second glance, most public violence connects closely with politics in general.

See also: Conflict Sociology; Ethnic Conflicts; Genocide: Historical Aspects; Military History; Military Sociology; Police, Sociology of; Revolutions, Sociology of; Revolutions, Theories of; Sport, Sociology of; Terrorism; Violence and Media; Violence, History of; Violence in Anthropology; War: Anthropological Aspects; War, Sociology of; Warfare in History

Bibliography

Beissinger M R 1998 Nationalist violence and the state. Political authority and contentious repertoires in the former USSR. *Comparative Politics* **30**: 401–33

Brass P R (ed.) 1996 *Riots and Pogroms* New York University Press, New York

Brubaker R, Laitin D D 1998 Ethnic and nationalist violence. *Annual Review of Sociology* **24**: 423–52

de Waal A 1997 *Famine Crimes. Politics & the Disaster Relief Industry in Africa*. Indiana University Press, Bloomington, IN

Dudley L M 1991 *The Word and the Sword. How Techniques of Information and Violence Have Shaped the World*. Blackwell, Oxford, UK

Fogarty B E 2000 *War, Peace, and the Social Order*. Westview Press, Boulder, CO

Gambetta D 1993 *The Sicilian Mafia. The Business of Private Protection*. Harvard University Press, Cambridge, MA

Gould R V 1999 Collective violence and group solidarity: evidence from a feuding society. *American Sociological Review* **64**: 356–80

Grimshaw A D 1999 Genocide and democide. In: *Encyclopedia of Violence, Peace, and Conflict* Academic Press, San Diego, CA, Vol. 2, pp. 53–74

Gurr T R (ed.) 1989 *Violence in America*. Sage, Newbury Park, CA, 2 Vols.

Gurr T R, Harff B 1994 *Ethnic Conflict in World Politics*. Westview Press, Boulder, CO

Kaldor M 1999 *New and Old Wars. Organized Violence in a Global Era*. Polity Press, Cambridge, UK

Licklider R (ed.) 1993 *Stopping the Killing. How Civil Wars End.* New York University Press, New York
della Porta D 1995 *Social Movements, Political Violence, and the State. A Comparative Analysis of Italy and Germany.* Cambridge University Press, Cambridge, UK
Reiss A J Jr., Roth J A (eds.) 1993 *Understanding and Preventing Violence.* National Academy Press, Washington, DC
Reyna S P, Downs R E (eds.) 1999 *Deadly Developments. Capitalism, States and War.* Gordon and Breach, Amsterdam
Rule J B 1988 *Theories of Civil Violence.* University of California Press, Berkeley, CA
Scott J C 1985 *Weapons of the Weak. Everyday Forms of Peasant Resistance.* Yale University Press, New Haven, CT
Scott J C 1990 *Domination and the Arts of Resistance. Hidden Transcripts.* Yale University Press, New Haven, CT
Scott J C 1998 *Seeing Like a State. How Certain Schemes to Improve the Human Condition Have Failed.* Yale University Press, New Haven, CT
Stanley W 1996 *The Protection Racket State. Elite Politics, Military Extortion, and Civil War in El Salvador.* Temple University Press, Philadelphia
Thomson J E 1994 *Mercenaries, Pirates, and Sovereigns. State-Building and Extraterritorial Violence in Early Modern Europe.* Princeton University Press, Princeton, NJ
Zolberg A R, Suhrke A, Aguayo S 1989 *Escape from Violence. Conflict and the Refugee Crisis in the Developing World.* Oxford University Press, New York

C. Tilly

Virtual Reality and Spatial Cognition

In its broadest sense the term virtual reality is currently being used to describe everything from three-dimensional photorealistic rendering through non-graphical simulation to artificial intelligence. In its narrower sense virtual reality is a technology that makes use of computer-generated interactive graphics which give the user the sensation of being in a virtual, that is, computer-generated world. Special hardware devices such as projection screens or head-mounted displays (HMDs) aid the perception of the synthetic world. Another important feature is the possibility to interact with this world through various devices, for example haptic devices to experience force feedback by the virtual world, bicycles to move through the world, or pointing devices.

In this article, we give an overview of applications of virtual reality in behavioral experiments on human cognition. We focus on spatial cognition, i.e., navigation and spatial memory.

1. Applications

The very first application of virtual reality was the training of air force pilots in flight simulators. The first computer-graphics based flight simulator was built in the 1970s. By the 1980s, with better software and hardware, the market for virtual reality technology had grown far beyond military use. As we enter the twenty-first century, virtual reality is a standard method for manipulating the action–perception cycle underlying normal behavior in all autonomous agents (see *Autonomous Agents*).

Contemporary virtual reality systems are widely used. Examples include:

1.1 Visualization

Visualization is of high importance for interpreting data for many technical or scientific applications. An excellent overview is given in vanDam et al. (2000). Examples come from the fields of art, medicine, architecture, and archaeology. For example, members of the Computer Science Department at Brown University have been collaborating with archaeologists to develop the Archave system which makes it possible to visualize the results of excavations. On the basis of this visualization, further analysis of the data is possible. Another example comes from the field of car design. In virtual realities car-designers and engineers are not only capable of presenting future models but can also examine functional aspects. For example in a packaging simulation, one can study the assembly of an engine or the possibility to dismantle single pieces of the engine for repair. Furthermore, crash tests can be observed from unusual perspectives: One can place a virtual camera right in the middle of the engine during a crash test.

1.2 Training and Education

There are many applications of virtual reality to training and education, especially in the field of medicine. For example, many practising surgeons completed their training before the so-called laparoscopic era of the 1990s. Laparoscopic or minimal invasive surgery has dramatically changed the way surgery is performed, which makes it necessary to teach even very experienced physicians. For this reason, there are many training simulations in the field of endoscopic surgery, for example of the knee (Muller et al. 2000). In this simulation students can act on virtual reality tendons and bones with realistic instruments.

A further application is the various driving simulators—for example, for aeroplanes, helicopters, motorcycles, trucks, and even wheelchairs. In the case of aeroplanes and helicopters, the purpose of training in a flight simulator is not to substitute for the training in the real world completely but to familiarize students, for example, with the handling of the cockpit, instrument displays, navigational tasks, or modifications of existing instruments.

1.3 Behavioral Experiments

Computer graphics has been used for many years in the generation of experimental stimuli to examine the visual system. With virtual reality, the sensory system and related behavior can be investigated, that is, it is possible to examine a whole action–perception cycle. If one focuses on perception alone, it is always hard to decide which feature of the measured sensor performance is responsible for a certain behavior. The following are easily measurable behaviors in virtual worlds: reaction-time measures, pointing direction, grasping behavior, eye movements, or navigation ability. For the experimenter, the advantage of the virtual world compared to the real world lies in the possibility of controlling the environment very strongly, in the flexibility and ease with which different environments can be constructed, and in the ease with which experimental data can be saved and visualized.

2. Spatial Cognition: Experiments in Virtual Reality

Many fruitful experiments in the field of spatial cognition have been performed with virtual reality. A comprehensive overview of the advantages and disadvantages of performing navigation experiments in a virtual world is given in Péruch et al. (1998). On the one hand, it is possible to build up totally new environments, which are unknown to subjects. Thus unbiased measurements of navigational ability are obtained. Mallot et al. (1998) describes a maze—like city called 'Hexatown,' where each junction is a connection of three streets, resulting in a hexagonal structure (Fig. 1). The regularity of the ground plane has the advantage that the structure of the city does not give the subject any cues in a navigational task. During the experiments the subjects sit in front of a computer screen and move through the town by pressing the buttons of a computer mouse. The results suggest that the subjects acquire a cognitive map during exploration even in the absence of physical motion. On the other hand, it is possible to reconstruct existing cities or buildings to investigate in virtual reality the spatial memory acquired in real environments or vice versa (Fig. 2).

Witmer et al. (1996) investigated the transfer of knowledge based on training in a virtual world to knowledge acquired in the real world. They concluded from their experiments that the resulting configuration knowledge is very similar.

As mentioned previously, one advantage of experiments in a virtual world is the high flexibility and controllability they provide. In older navigational studies the experimenters prepared videofilms or slide presentations in order to control the environment that the subjects had to explore. In other studies model towns were built using toy houses and the perspectives from within the town were caught in a camera drive. This latter method resembles the fashion in which virtual worlds are built up today. With software very similar to the computer-aided design (CAD) programs used by architects, the surroundings can be modeled. The interaction with a user resembles a camera drive moving through the model. Using this

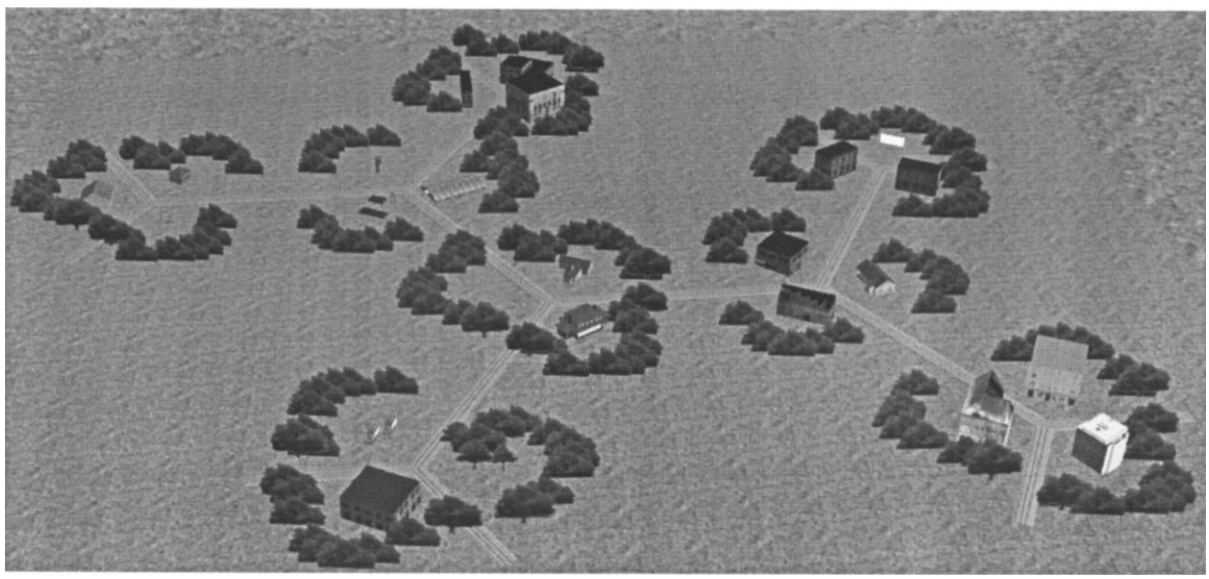

Figure 1
Bird's eye view of the synthetic city called 'Hexatown' used in Mallot et al. (1998)

Figure 2
Single view of the virtual reconstruction of the historic centre of Tübingen used in Mallot et al. (1998), courtesy of H H Bülthoff

type of software it is much easier to make subtle changes. For example, it is possible to move objects during the experiments (Aginsky et al. 1997), to change the lighting conditions (Mallot et al. 1998), or to change the topography of the environment, for instance the slant (Mallot et al. 2000).

One set of navigational experiments has focused on the question of the comparability of the results obtained in the virtual world to results obtained in the real world. Ruddle and colleagues (1997) replicated a distance-estimation and pointing task in a virtual world that was originally carried out by Thorndyke and Hayes-Roth (1982) in the real world. The results turned out to be very similar. However, Péruch et al. (1997) found in a path-integration task that subjects in a virtual world made a different systematic error compared to subjects in the real world (Loomis et al. 1993). The latter tend to compress the range of actually produced turns and distances in such a way that large values are underestimated and small values are overestimated. Subjects in the virtual world underestimated all values. One has to keep in mind that the sensory input in the virtual world is different from the input in the real world. For instance, in the Loomis et al. (1993) study the subjects were blindfolded during the experiments, that is, perception was based on vestibular and proprioceptive input alone. On the other hand, in the Péruch et al. (1997) study the behavior was based only on visual input.

Recently, there have been efforts to combine the visual input in the navigation experiments with adequate proprioceptive and vestibular stimuli. Most of the studies carried out so far have taken place in 'desktop virtual reality,' where subjects sit in front of a computer screen or a projection wall and press the buttons of a computer mouse or move a joy stick. By doing this, they navigate through the virtual world. In a study of Chance et al. (1998) subjects performed a triangle-completion task in which they walked two sides of a triangle and were then requested to return to the starting point, completing the third side. During walking, the visual input was presented on a head-mounted display. The performance was better than in 'just turn' group. The latter could only physically turn and carried out translatory movements in the virtual world.

A very interesting paradigm is the combination of virtual reality with functional imaging studies (PET or

fMRI) to identify regions in the brain involved in spatial cognition. One exciting finding is that there is an activation of the *hippocampus* proper (besides several other regions in the brain), which has already been found in animal experiments. In studies of rat hippocampal pyramidal cells, so-called 'place cells' have been identified, that fire when the animal is in a specific location (O'Keefe and Dostrovsky 1971). In similar fashion, Maguire et al. (1998) found in a PET study with human subjects that the right hippocampal formation shows increased activity during exploration of a virtual town (see *Navigation in Virtual Space: Psychological and Neural Aspects*).

3. Simulation Sickness

A major problem with virtual reality is simulation sickness, which was initially documented in a helicopter flight simulator. The most common symptoms are: general discomfort, sweating, disorientation, nausea, and vomiting. Some people exhibit all of these signs whereas others exhibit only a few or even none. In an analysis of data from 10 US Navy and Marine Corps flight simulators, it was found that approximately 20 to 40 percent of the pilots experienced at least one of the symptoms (Kennedy et al. 1989). The main theory explaining simulation sickness is the 'Cue Conflict Theory,' which implies that the main causes are discrepancies between the senses concerning information about body orientation and motion.

See also: Autonomous Agents; Ethnic Conflict, Geography of; History of Technology; Navigation; Navigation in Spatial Environments, Neural Basis of; Navigation in Virtual Space: Psychological and Neural Aspects; Spatial Cognition; Virtual Reality, Psychology of; Wayfinding and Navigation Behavior

Bibliography

Aginsky V, Harris C, Rensick R, Beusmans J 1997 Two strategies for learning a route in a driving simulator. *Journal of Environmental Psychology* **17**: 317–31

Chance S S, Gaunet F, Beall A C, Loomis J M 1998 Locomotion mode affects the updating of objects encountered during travel: The contribution of vestibular and proprioceptive inputs to path integration. *Presence* **7**(2): 168–78

Kennedy R S, Lilienthal M G, Berbaum K S, Baltyley D R, McCauley M E 1989 Simulator sickness in US Navy flight simulators. *Aviation, Space and Environmental Medicine* **60**(1): 10–16

Loomis J M, Klatzky R L, Golledge R G, Cicinelli J G, Pellegrino J W, Fry P A 1993 Nonvisual navigation by blind and sighted: Assessment of path integration ability. *Journal of Experimental Psychology General* **122**(1): 73–91

Maguire E A, Burgess N, Donnet J G, Frackowiak R S I, Frith C D 1998 Knowing where and getting there: A human navigation network. *Science* **280**: 921–4

Mallot H A, Gillner S, vanVeen H A H C, Bülthoff H H 1998 Behavioral experiments in spatial cognition using virtual reality. In: Freksa C (ed.) *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*. Springer, Berlin, pp. 447–67

Mallot H A, Mochnatzki H F, Steck S D 2000 Geographic slant improves navigation performance in virtual environments. *Investigative Ophthalmology & Visual Science, Suppl.* **41**(4): 44

Muller W, Bockholt U, Lahmer A, Voss G, Borner M 2000 VRATS – Virtual reality arthroscopy training simulator. *Radiologe* **40**(3): 290–4

O'Keefe J, Dostrovsky J 1971 The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat. *Brain Research* **34**: 171–5

Péruch P, May M, Wartenberg F 1997 Homing in virtual environments: Effects of field of view and path layout. *Perception* **26**: 301–11

Péruch P, Gaunet F 1998 Virtual environments as a promising tool for investigating human spatial cognition. *Current Psychology of Cognition* **17**(4–5): 881–99

Ruddle R A, Payne S J, Jones D M 1997 Navigating buildings in 'desktop' virtual environments: Experimental investigations using extended navigational experience. *Journal of Experimental Psychology: Applied* **3**(2): 143–59

Thorndyke P W, Hayes-Roth B 1982 Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology* **14**: 560–89

vanDam A, Forsberg A S, Laidlaw D H, LaViola J L, Simpson R M 2000 Immersive VR for scientific visualization: A progress report. *IEEE Computer Graphics* **20**(14): 26–52

Witmer B G, Bailey J H, Knerr B W, Parsons K C 1996 Virtual spaces and real world places. Transfer of route knowledge. *International Journal of Human–Computer Studies* **45**: 413–28

S. Gillner and H. A. Mallot

Virtual Reality, Psychology of

The term 'virtual reality' refers to a computer program that gives its users the sense of being in a world that exists outside physical space. The experience may be created through visual images and/or text. In some virtual realities we are alone as we navigate virtual oceans, unravel virtual mysteries, and engineer virtual skyscrapers. In others, we are in the company of other people, all connected to the same program and to each other through the Internet, the rapidly expanding global system of networks that links millions of personal computer users together.

The psychology of virtual reality includes a broad set of questions, both because of the diversity of virtual environments (everything from video games to Internet chatrooms) and the wide range of psychological frameworks that may be brought to bear on their exploration. The field includes cognitive psychologists who study sensorimotor adaptation in immersive virtual environments, such as flight simulators, and clinical psychologists who study video game playing as a form of addiction. Current questions in the field include the role of simulated spaces on people's experience of the physical 'real.' This issue is

raised when a seven-year-old child looks at a seashore jellyfish and comments 'It looks so realistic.' It is raised when we compare a rafting trip down the Colorado River to an adolescent girl's use of an interactive CD-ROM to explore the same territory. And it is raised in the contemporary practice of science, highly dependent on simulation, and in science education. Some scientists are concerned that once students have seen an experiment unfold perfectly in a simulation, the messiness of a real experiment—the imperfections of measurement, the crack in the equipment that means you have to repeat it—can come to seem like a waste of time (Turkle 1995).

The perspective taken in this article necessarily narrows the field, focusing on the experience of the self in social virtual realities, environments where people interact with each other through their self-representation in computational worlds. The current practice of social virtual reality enables people to recast their notions of personal identity in more flexible terms, because they are able to use relationships in cyberspace to explore multiple aspects of self. Virtual reality is not the only factor that is moving our psychological culture in this direction (Gergen 1991; Lifton 1993), but it is one of the most potent.

1. *Virtual Personae*

A key element of the psychological impact of on-line life is the possibility of creating and projecting constructed personae into virtual space. This can be done through the use of visual materials (virtual worlds may allow users to design their own virtual 'bodies') as well as through textual description. In cyberspace one's constructed identity can diverge from physical realities: the obese can present themselves as slender, the beautiful can present themselves as plain. And text-based on-line 'conversations' allow time to reflect upon and edit one's 'composition.' This makes it easier for the shy to be outgoing, the 'nerdy' sophisticated. The relative anonymity of Internet identities—in most virtual settings, one has the choice of being known only by one's chosen 'handle,' or on-line name—gives people the chance to express often unexplored aspects of the self. The simplest form of this multiplication of selves takes place in the simple act of naming oneself when people take out multiple e-mail accounts on the web or multiple names or 'handles' on their on-line service. It is not unusual for one person to have three accounts such as, for example, armani_boy@hotmail.com, motorcycle_man@hotmail.com, pinstripe_guy@hotmail.com. It is not unusual for one person to be in one chatroom as MrSensitive, a second as RollingStone, and a third as GayPal. And, of course, it is possible for the physical person behind armani_boy or MrSensitive to be a woman rather than a man.

The on-line exercise of playing with identity and trying out new ones is perhaps most explicit in role-playing virtual communities and in massively multi-player games, where participation begins with the creation of a persona (or several), but it is by no means confined to these somewhat exotic locations. In bulletin boards, newsgroups, and chatrooms, the creation of personae may be less explicit, but it is no less psychologically real.

The creation of site-specific on-line personae is not only dependent on adopting a new name. Personae-shifting happens with a change of virtual place. Cycling through virtual environments is made possible by the existence of what have come to be called 'windows' in modern computing environments. Windows are a way of working with a computer that makes it possible for the machine to place you in several contexts at the same time. As a user, you are attentive to only one of the windows on your screen at any given moment, but in a certain sense, you are a presence in all of them at all times. Your identity on the computer is the sum of your distributed presence.

Technology has always provided psychology with metaphors for mind; perhaps the best-known example is Freud's use of hydraulics to represent the play of forces in the psyche. In our time, the computer has become a model of mind in the academic world (Dennett 1991) and in everyday practice, the windows technology that we use when visiting cyberspace provides a potent metaphor for thinking about the self as a multiple, distributed, 'time-sharing' system. The self is no longer simply playing different roles in different settings—something that people experience when, for example, one wakes up as a lover, makes breakfast as a mother, and drives to work as a lawyer. The windows metaphor suggests a distributed self that exists in many worlds and plays many roles at the same time. It suggests a life experience in which one 'cycles through' and is able to experiment with different 'aspects of self.'

2. *Identity, Moratoria, and Play*

Cyberspace, like all complex phenomena, has a range of psychological effects. For some people, it is a place to act out unresolved conflicts, to play and replay characterological difficulties on a new and exotic stage. For others, it provides an opportunity to work through significant personal issues, to use the new materials of cybersociality to reach for new resolutions. These more positive identity effects follow from the fact that, for some, cyberspace provides what Erik Erikson (1963) would have called a 'psychosocial moratorium,' a central element in how Erikson thought about identity development in adolescence. Although the term 'moratorium' implies a 'time-out,' what Erikson had in mind was not withdrawal. On the contrary, the adolescent moratorium is a time of intense social and intellectual activity. It is a time of

passionate friendships and experimentation. The adolescent falls in and out of love with people and ideas. Erikson's notion of the moratorium was not a hold on significant experiences but on their consequences. It is a time during which one's actions are, in a certain sense, not counted as they will be later in life. They are not given as much weight, not given the force of full judgment. In this context, experimentation can become the norm rather than a brave departure. Relatively consequence-free experimentation facilitates the development of a 'core self,' a personal sense of what gives life meaning that Erikson called 'identity.'

Erikson developed these ideas about the importance of a moratorium during the late 1950s and early 1960s. At that time, the notion corresponded to a common understanding of what 'the college years' were about. Today, the idea of the college years as a consequence-free time-out seems of another era. College is pre-professional, and AIDS has made consequence-free sexual experimentation an impossibility. The years associated with adolescence no longer seem a time out. But if our culture no longer offers an adolescent moratorium, the possibility of a virtual life often does. This is part of what gives virtual reality its psychological holding power.

Erikson's ideas about the importance of a psychosocial moratorium for successful identity development were part of a larger theory of the stages of the lifecycle. Erikson's ideas about stages did not suggest rigid sequences. His stages describe what people need to achieve before they can move ahead easily to another developmental task. For example, Erikson pointed out that successful intimacy in young adulthood is difficult if one does not come to it with a sense of who one is, the challenge of adolescent identity building. In real life, however, people usually have to move on in their lives with significant deficits. With incompletely resolved 'stages,' they simply do the best they can. They use whatever materials they have at hand to get as much as they can of what they have missed. Now virtual social life plays an important role in these dramas of self-reparation. Time in cyberspace reworks the notion of the moratorium, because it may now exist in an always-available 'window.'

3. *Virtual Reality and the Psychology of Everyday Life*

Recent psychological theory has stressed the importance of appreciating multiplicity or 'decentering' as a cornerstone of identity (Bromberg 1994, Dennett 1991, Gergen 1991, Lifton 1993). But despite the prevalence of such critiques of traditional, unitary notions of 'the ego' or 'identity,' it is usually difficult for people to make a personal connection with these ideas—hard, for example, to accept a challenge to the idea of an autonomous ego. The normal requirements of everyday life exert strong pressure on individuals to take responsibility for their actions and to see themselves as unitary actors. And our very language gives us a unified self in our first-person assertions: 'I do,' 'I want,' I desire.' This disjuncture between theory (the unitary self is an illusion) and lived experience (the unitary self is the most basic reality) is an important reason why, in the history of psychology, multiple and decentered theories have been slow to catch on—or when they do, why we tend to settle back quickly into older, centralized ways of looking at things. But today, the everyday practice of people using personal computers and modems to participate in on-line life provides experiences that bring theories of multiple selves 'down to earth.' When virtual reality enables an individual to construct multiple personae that can explore different aspects of the self, the notion of a decentered identity is concretized. In this way, cyberspace becomes an 'object to think with' for thinking about identity.

The popular appropriation of Freudian ideas had little to do with scientific demonstrations of their validity. Freudian ideas passed into the popular culture because they offered robust and down-to-earth objects-to-think-with (Turkle 1992). The objects were not physical but almost-tangible ideas, such as dreams and slips of the tongue. People were able to play with such Freudian 'objects.' They became used to looking for them and manipulating them, both seriously and not so seriously. And as they did so, the idea that slips and dreams betray an unconscious started to feel natural. In Freud's work, dreams and slips of the tongue carried the theory.

Today, virtual reality—life on the computer screen—carries psychological theory. People decide that they want to interact with others on a computer network. They get an account on a commercial service. They think that this will provide them with new access to people and information, and it does. But it does more. When they log on, they may find themselves playing multiple roles; they may find themselves playing characters of the opposite sex. In this way they are swept up by experiences that enable them to explore previously unexamined aspects of their sexuality or that challenge their ideas about a unitary self. To use the psychologist Philip Bromberg's (1994) language, on-line life may enable people to 'stand in the spaces between selves and still feel one, to see the multiplicity and still feel a unity.' To use the computer scientist Marvin Minsky's (1986) language, on-line life may enable people to experience themselves as cycling through their 'society of mind,' a notion of identity as distributed and heterogeneous.

4. *A Heightening of Projection*

The psychological power of virtual reality is enhanced by the fact that phenomena associated with projection and transference are heightened in virtual space. As in

Thomas Mann's *The Magic Mountain*, which takes place in the isolation of a sanitarium, relationships in cyberspace become intense very quickly because the participants feel themselves to be in a remote and unfamiliar world with its own rules. Electronic meeting places breed an easy intimacy. In the first phase of electronic friendships, people describe the excitement of a rapidly deepening relationship and the sense that time itself is speeding up. In virtual encounters, the lack of information about the physically real person to whom one is talking, the silence into which one types, the absence of visual cues, all encourage projection. This can lead to disappointment when, in a second stage of friendship, people who have had only virtual encounters move to bring them into the physical 'real.' However, it also has a positive side. Virtual reality offers a new tool for self-reflection. In close virtual friendships, it is not unusual to 'log' one's interaction with an intimate. It is possible to compare the text of the 'relationship' with the depth of feeling it has evoked. Very often, the gap between the two provides a rough but potentially useful approximation of the 'transference.' More frequently, people do not have an actual transcript of an on-line relationship to work with, but as they become more experienced in virtual social interaction, they are in a good position to notice recurrent patterns of on-line behavior. In this sense, it is possible to use the virtual to reflect constructively on the real. Otherwise put: people who make the most of their lives on the screen are usually those who are capable of approaching it in a spirit of self-reflection; they examine what they are doing on-line and ask themselves what it says about their desires, perhaps unmet, about their needs for connection, perhaps unfulfilled. They make efforts to use what they learn in virtual reality to improve the quality of their lives in the physical 'real.'

The study of the psychology of on-line life requires a better understanding about how to instill and nurture this discipline of self-reflection. Among psychologists, this enterprise has met some resistance by a movement to label the medium 'addicting' and requiring control much as one would control a drug (Young 1998, Greenfield 1999). There is a lively debate among psychologists (Young 1999) about the usefulness of the concept of Internet addiction. For some, the analogy is justified by the compelling hold of virtual life and computer games and by the fact that the term 'addiction' is employed by many who are drawn to chatrooms, on-line communities, and computer games and who find themselves spending more and more time using them. For others, the addiction analogy seems misplaced (Turkle 1995, Grohol 1999). The argument against the addiction metaphor is that it closes down the important psychological questions that discriminate among computer users and the use to which they put virtual reality. Cyberspace may be compelling, but it is only a medium and can be used for 'acting out' or 'working through' for both constructive and destructive psychological ends. From this point of view, a parent whose child is on heroin needs to help the child stop taking the drug, whereas a parent whose child is on-line needs to be curious about what he or she is doing there. Is the child forming on-line relationships that are serving important developmental purposes? Are the specific on-line experiences likely to serve as stepping-stones for important emotional or intellectual growth? From this perspective, when we try to understand the psychological state of a person at a networked computer, it is most constructive to think of the Internet and its virtual spaces as a Rorschach, rather than as a narcotic.

5. *Future Directions*

Early writings about virtual reality stressed the possibilities of fully immersive environments, as well as the possibilities of full-body suits that would open up erotic possibilities for long-distance sexual activity and, less sensational but equally exotic, experiences for world exploration (Rheingold 1993). One could thrash one's way through a Brazilian rain forest and feel the thickness of the brush or walk through the Latin Quarter of virtual Paris and feel the roughness of the cobblestones under one's feet. These possibilities may well be in the future. But it is significant to note that the most robust applications of virtual reality, even after more than a decade of technical progress, remain some version of role-playing community creation. When hundreds of thousands of people join on-line gaming communities, they appreciate the technical innovations that make their 'worlds' increasingly well rendered. But what keeps them engaged is the way they become taken up by roles in which they can be somewhat 'other' and play on teams, where they can socialize with other players. Sociality and the pleasures of talking to other real people remains the 'killer application' for the virtual.

In on-line games set in virtual medieval landscapes, players do not only make friends and have romances, they join guilds, develop strategic alliances, and have political disagreements that lead to the formation of political parties and the drafting of political constitutions and bylaws. But it is one thing to raise a virtual child in a simulation game and quite another to confront the problems of real children, who so often face lives without adequate parenting or schooling, or health care or hope.

I have interviewed 20- to 30-year-olds who consider themselves politically active in virtual communities but who do not bother to vote in local or national elections. They say that they feel more powerful in their virtual communities than in the politics of the physical world, which they see as cynical and unresponsive. One might fear that virtual reality, to borrow a concept from Marx, is like religion, a kind of opiate. It is important to confront the seductions of simu-

lation: our real communities need us badly, and we can ill afford to abandon them for virtual citizenships, no matter how gratifying. This essay began with the thought that having literally written our on-line personae into existence, they can be a kind of Rorschach. We can use them to become more aware of who we are and what still feel we need. The same can be true of our lives in virtual communities. Hopefully, the future will bring greater emphasis on learning how to use the virtual to reflect constructively on the physical real.

See also: Computers and Society; Identity in Childhood and Adolescence; Identity: Social; Internet: Psychological Perspectives; Science and Technology, Anthropology of; Social Identity, Psychology of; Virtual Reality and Spatial Cognition; Visual Imagery, Neural Basis of; Visual Images in the Media

Bibliography

Bromberg P 1994 Speak that I may see you: Some reflections on dissociation, reality, and psychoanalytic listening. *Psychoanalytic Dialogues* **4**(4): 517–47

Dennett D C 1991 *Consciousness Explained*. Little, Brown, Boston

Erikson E H 1963 *Childhood and Society*, 2nd edn. Norton, New York

Gergen K J 1991 *The Saturated Self: Dilemmas of Identity in Contemporary Life*. Basic Books, New York

Greenfield D N 1999 Psychological characteristics of compulsive internet use: A preliminary analysis. In: Young K S (ed.) *Cyberpsychology and Behavior: The Impact of the Internet, Multimedia, and Virtual Reality on Behavior and Society. Special Issue on Internet Addiction*, Vol. 2, no. 5. Mary Ann Liebert, Larchmont, New York, pp. 403–12

Grohol J M 1999 Too much time on-line: Internet addiction or healthy social interactions. In: Young K S (ed.) *Cyberpsychology and Behavior: The Impact of the Internet, Multimedia, and Virtual Reality on Behavior and Society. Special Issue on Internet Addiction*,. Vol. 2, no. 5. Mary Ann Liebert, Larchmont, New York, pp. 395–401

Lifton R J 1993 *The Protean Self: Human Resilience in an Age of Fragmentation*. Basic Books, New York

Minsky M 1986 *The Society of Mind*. Simon and Schuster, New York

Rheingold H 1993 *The Virtual Community: Homesteading on the Electronic Frontier*. Addison-Wesley, Reading, MA

Turkle S 1984 *The Second Self: Computers and the Human Spirit*. Simon and Schuster, New York

Turkle S 1992 *Psychoanalytic Politics: Jacques Lacan and Freud's French Revolution*, 2nd edn. Guilford, New York

Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon and Schuster, New York

Young K S 1998 *Caught in the Net: How to Recognize Signs of Internet Addiction and a Winning Strategy for Recovery*. Wiley, New York

Young K S (ed.) 1999 *Cyberpsychology and Behavior: The Impact of the Internet, Multimedia, and Virtual Reality on Behavior and Society. Special Issue on Internet Addiction*, Vol. 2, no. 5. Mary Ann Liebert, Lachmont, New York, pp. 381–4

S. Turkle

Virtue Ethics

1. Definition

It has often been said that the development of virtue ethics (in modern dress) is in its infancy. For this reason, though virtue ethics is a type of substantive normative ethical theory opposed to consequentialism and Kantian ethics, it resists definition. The problem of definition is compounded by the fact that though much contemporary virtue ethics is neo-Aristotelian, it need not be, and has not always been. Increasingly, David Hume is being read in a way more sympathetic to virtue ethics, Nietzschean virtue ethics is developing, and Plato and Martineau have inspired 'agent-based' virtue ethics. Virtue ethicists, then, need not, and do not agree on such central questions as the definition of right action, and what makes a trait of character a virtue. Common to virtue ethicists, however, is the view that character rather than the motive of duty, rules, or consequences is central to ethics. This is not to say that rules, consequences, and the motive of duty are not important, but a character-based ethics will produce virtue-centered conceptions of moral rightness, objectivity in ethics, impartiality, demandingness or stringency of ethics, and practice, which reconfigure the moral universe.

The major source of inspiration for the development of modern virtue ethics is Aristotle (1955), particularly his *Nicomachean Ethics*. Two central features of his view have influenced definitions of modern virtue ethics. The first is his idea that the judgment of a virtuous agent is the rule and the measure of the right. This feature has spawned the view that virtue ethics is opposed to moral rules and principles, and that these must be entirely replaced by the judgment of a qualified judge (a virtuous agent). However there are 'virtue-rules' such as requirements to be kind, considerate, fair, courageous, and so on: rules which we use to educate our children into the moral life, and which we may use in morally charged situations where we might ask for example 'What would be the kind or tactful thing to do here?' Virtue ethicists, however, follow Anscombe (1958) in believing that moral rules should be couched in the 'thick' virtue concepts.

The second feature sometimes thought to be defining of virtue ethics is Aristotle's view that virtues

are essentially connected with the flourishing of their possessor. David Hume, however, thought that there are four potential grounds for status as a virtue: agreeableness and usefulness to self and agreeableness and usefulness to others. What constitutes right action and what gives status as a virtue are points of disagreement among virtue ethicists, so particular conceptions of right action and virtue status should not be employed to define virtue ethics as such.

2. *Intellectual Content of Virtue Ethics*

Discussion of the intellectual content of virtue ethics can center usefully on three issues: (a) virtue ethical conceptions of moral rightness of action, (b) the nature of virtue in virtue ethics, and (c) different conceptions of what makes a trait of character a virtue, within virtue ethics.

2.1 *Virtue Ethical Conceptions of the Right*

It is often thought that the virtue ethical conception of rightness, indeed one that is defining of virtue ethics is the following. (a) An act is right if and only if it is what a virtuous agent would characteristically (i.e., acting in character) do in the circumstances (Hursthouse 1996). However, other possibilities within virtue ethics exist, and have been canvassed. According to 'agent-based' virtue ethics rightness is understood thus. (b) An act is right if and only if it exhibits good motivation (Slote 1996). Statman (1997) alludes to another possibility. (c) An act is right if and only if it is expressive of virtuous character.

Finally, we might say (d) an act is right if and only if it is overall virtuous, and (d) differs from (a) in that it is possible for a virtuous agent acting in character to do in the circumstances something which is wrong according to (d). For example, according to (d), an act expressive of a benevolent disposition, and benevolently motivated, may fail to be a benevolent act. For a virtuous agent, though endowed with practical wisdom, is not omniscient, and the act may turn out to be extremely harmful. For the same reason (d) differs from (b) and (c). Furthermore, virtuous acts, as opposed to actions from or expressive of a state of virtue, need not possess all the features required for action from a virtuous state (as Aristotle recognized). In certain contexts, they may not express fine character or even good motives. In short a virtue ethicist may share the common conception of right action as one that is successful in hitting its target. Where virtue ethicists who share this view differ from, for example, consequentialists, is in their conception of 'hitting the target.' The target is set not by consequentialist criteria of promoting the most good, but by the ends of the various virtues. These ends will of course differ from virtue to virtue, so unless one subscribes to a strong doctrine of the unity of the virtues, virtue ethicists will possess a pluralistic view of rightness of action.

Where a virtue ethicist does not require fine inner states for rightness (as opposed to goodness) of actions, they may distinguish between rightness and excellence of acts, where the criteria for the latter are more stringent than those for the former.

2.2 *The Nature of Virtue in Virtue Ethics*

A virtue is an excellence of character. It is a disposition of acknowledging or responding to items in the field of the virtue in an excellent way. According to virtue ethical conceptions of virtue, the standards of excellence point to states inside the agent, as well as to the world outside the agent. In short, for a virtue ethicist, the realization of the end of a virtue must express fine inner states, if that realization is to display virtue.

Virtues have a specialized focus, that is, they have a field. The field of a virtue may be human beings in general, one's children, property, friends, honors, pleasure, the natural world, cultural and artistic objects, money, abstract objects such as knowledge, and states of affairs such as dangerous situations, interpersonal conflict, and so on. There is a wide variety of virtues which focus on different fields or domains—virtues such as friendship, courage, justice, environmental virtues, appropriate assertiveness, honesty, politeness and tact, liberality, generosity, thrift. Some virtues are universal such as courage, temperance, and justice, while others are relative to roles, such as professional virtues and virtues of parenthood. Yet others are relative to situation, for example, according to Aristotle, virtues whose field concerns (the expenditure of) money vary according to how much money one has, and range from thrift to 'magnificence.'

Insofar as 'virtue' is a success word, the form of acknowledgment of items in the field of a virtue required by a virtue must be appropriate or successful. What constitutes success or appropriateness will vary according to two types of things: first, according to the different kinds of morally significant feature in those items which it is the end of the virtue to recognize, sustain, further, or bring about (such as their value or status), and second according to the differing forms or modes of moral acknowledgment of items required by the various virtues (such as promoting (value or good) or respecting individuals in ways befitting their status). Virtue ethicists will differ on what are the basic morally significant features in the items acknowledged, and what are the basic forms of moral acknowledgment inherent in virtue. The former may include value, status, bonds, and good; the latter promotion, honoring (Pettit 1997), respect, love, appreciation,

and creativity. The important point is that the criteria of success in meeting the various ends of the different virtues will vary, depending on which forms of moral acknowledgment are involved in their exercise. We may ask, for example, whether such virtues as justice, benevolence and friendship are centrally concerned with all, or some of, the following: the promotion of good, the promotion of value, the honoring of value, respect for individuals in ways befitting their status, or love for individuals in virtue of bonds between those individuals and the agent.

According to virtue ethicists, an exercise or manifestation of virtue requires the successful expression of fine inner states, that is, in displaying virtue, the agent must both have fine inner states and successfully express them in her behavior. For example, virtuous creativity does not merely involve meeting the standards for success in the external manifestation of creativity that confers artistic merit in the created product. The creative response must also be expressive of fine inner states, such as the expression of artistic passion, a desire to create something of value to society or of high artistic quality, practical wisdom in the production of value, originality, a desire to create something novel, interesting, or original. Similarly, love as a virtuous response or form of moral acknowledgment is not merely the promotion of the good of an individual, or the inculcation of a belief in the beloved that she is loved. Virtuous love is expressive of genuinely caring motivation: motivation that is not pathological, such as passively or morbidly dependent. Again, respect for status as a virtuous response is not merely the formal trappings of respect according to societal convention.

Though virtue ethicists share the view that virtue requires excellence of inner states, what kinds of inner states need to be excellent for virtue to be displayed is a matter of some controversy. Neo-Aristotelians emphasize the necessity of practical wisdom, but this view has been challenged by work on so-called 'virtues of ignorance' (Driver 1989), by those who emphasize the sufficiency of admirable motivation for virtue (Slote 1996), and by those who emphasize the pivotal role of passion in the life of virtue (Solomon 1998). Say, for example, that benevolence as a virtue is concerned centrally with promoting the good of individuals, we still need to know answers to the following questions. Is it enough for benevolence to be displayed that the agents display good motives in her attempt to promote good? Is it necessary that she display practical wisdom? Is it sufficient that she display practical wisdom or is she also required to display knowledge in cases where there is large-scale ignorance concerning good? To what extent does virtue require rising above the general ignorance of a population or culture?

Some progress on the issue of what constitutes excellence in inner states appropriate to human beings may be secured if Nietzsche's emphasis on excellence in depth psychological states is followed up. For Nietzsche, virtuous action is action expressive of fine depth states, and this view enables him to draw a distinction between virtuous and nonvirtuous altruism, virtuous and nonvirtuous asceticism, and so on. Forms of beneficence, sublimation, and self-discipline, which are expressive of self-contempt and resentment lack virtue, for Nietzsche. Post-Nietzschean analytic-oriented psychologists also distinguish virtue and vice on the basis of depth states expressed in habitual behavior. For example, genuine and passive dependent or 'morbid' dependent love is distinguished on that basis by, for example, Horney (1951). Fromm (1977) too, distinguishes between healthy (benign) and unhealthy (malignant) forms of aggression. Rather than aggressiveness being simply characterized as a vice, Fromm's account distinguishes between malignant and benign forms based on the human's characterological orientation to the satisfaction of basic needs such as the needs for excitation or stimulation, and relatedness.

2.3 *What Makes a Trait of Character a Virtue*

On the issue of what makes a trait of character a virtue, virtue ethics can be divided into eudaimonistic and noneudaimonistic forms.

2.3.1 Eudaimonism. According to eudaimonistic virtue ethics, which characterizes Greek and neo-Aristotelian virtue ethics in its modern form, virtue is at least partially constitutive of the flourishing of the possessor of the virtues. Common to all versions of eudaimonistic virtue ethics is the view that what constitutes the flourishing of the agent cannot be understood entirely independently from the idea of virtue itself. That is, in determining what the virtues are, we cannot simply ask, first, what is good for a person, and then define the virtues in terms of traits that promote that good. For example, according to Aristotle, fineness of character is the greatest good for a person. Pleasure without temperance, wealth without justice, the possession of friends without the virtue of friendship is not good for an individual, that is, does not benefit an individual. However, unlike the Stoics, Aristotle did not believe that virtue is sufficient for eudaimonia or flourishing: a modicum of good fortune and money, and children that do not disgrace one, are also necessary.

Due to the fact that eudaimonistic virtue ethics has been the dominant form of virtue ethics, it has been the main target of opponents of virtue ethics. The most important of their criticisms are the following. Because eudaimonism does not understand flourishing to be characterizable wholly independently of virtue, it has been accused of circularity. However, a coherentist

epistemology, which is indeed adopted by Aristotle and many neo-Aristotelians, may remove the sting from this criticism. It does so provided the network of concepts and background theories of human nature in which conceptions of virtue and flourishing are embedded, is sufficiently rich, informed by the social sciences, and broad-ranging, and does not rely on a problematically narrow, prejudiced, and uninformed set of intuitions. A second important criticism is the charge of egoism occasioned by the belief that, according to eudaimonism, the rationale for the virtues is their benefiting the possessor of the virtues. This criticism is met by the claim that virtuous action need not be, and is not standardly, motivated by the self-interest of the possessor of the virtue. This reply in turn begets the criticism of indirection, namely the criticism that a paradoxical and unmotivated dissonance opens up between the rationale for status of a trait as a virtue and the motivation and rationale of virtuous action.

2.3.2 Noneudaimonistic virtue ethics. Criticisms such as these have motivated, at least in part, the development of noneudaimonistic vrtue ethics, which does not require that status of a trait as a virtue is grounded in the flourishing of the possessor of the virtues. Two noneudaimonistic forms are currently being explored.

(a) Agent-based virtue ethics. According to 'agent-based' views, developed by Slote (1996, 1998a) and inspired by Plato and Martineau (1891), a virtue is a trait that is admirable, and its admirability need not be based on its contribution to further value, such as the welfare of human agents or the environment generally. Nor need a virtue be even partially constitutive of the flourishing of a human agent. Such a view relies for its development on aretaic intuitions about admirability, rather than the more standardly appealed to axiological intuitions about what is valuable.

(b) Response views. Although all virtues are dispositions of appropriate responsiveness to or acknowledgment of items within their fields, according to response views this feature is what makes a trait of character a virtue. It is not also necessary that a virtue characteristically constitute (in part) the flourishing of the agent. Response views may be more or less pluralistic about both the bases of moral acknowledgment (such as bond, value, status, and good) and the forms of moral acknowledgment (such as promotion, honoring, love, respect, appreciation, and creativity). Response views may allow that some goods and states (such as pleasure, money, friends) cannot be understood as goods entirely independently of virtue, but some theorists, such as Hurka (1992) require that there be base-level valuable states to which virtues are responses and which must be defined wholly independently of virtue.

3. *Intellectual Context of Modern Virtue Ethics*

Two main concerns have motivated the rise of virtue ethics in recent times:

(a) increasing emphasis on the importance of good character and character education for a healthy society, and

(b) dissatisfaction with rival moral theories—particularly consequentialism and Kantianism.

The emphasis on character, and the character training of children is increasingly thought to be important, indeed central, for an ethically sound society. Moral rules, where important, need to be made, interpreted, and applied in complex situations, and in ways that are not distorted by such vices as greed, impatience, cowardice, lack of perseverance, racism and other forms of hostility, callousness, oversensitivity, lack of concern for or over-concern with, fairness, and so on. A difficulty for a reorientation of ethical thought towards virtue ethics, in line with the above concern, is the apparent linkage of virtue ethics with social conservatism. This view, however, is based on several misunderstandings. First, virtue ethics as such is open as to how we delineate the virtues. Specific versions of virtue ethics will differ on this question. Second, though it would be agreed by all that it is imperative to bring up children to possess character traits which are not personally and socially destructive, virtue ethics as such is open as to whether a correct version of virtue ethics provides a monistic or pluralistic answer to the question 'What is a good life for a human being?' Third, it is sometimes thought that virtue ethics is too 'individualistic' and fails to appreciate the limitations of a virtuous agent, and her embeddedness in society. However, a virtue ethicist could agree that even a virtuous human agent necessarily has limited experience, possesses limited perspectives, and has limited expertise and knowledge. Accordingly, it is open to a virtue ethicist to adopt a dialogical view of virtues of practice, and moral epistemology generally.

The philosophical/theoretic push to virtue ethics has been driven by several factors, including the following. One factor is dissatisfaction with the lack of emphasis on, or a denigration of, the 'inner'—feelings, emotions, passions, and practical wisdom—in the life of the moral agent. Virtue ethicists emphasize the importance of outer action as expressive of inner states: benevolence, for example, is not just a matter of promoting value or good. Fine emotions, for example, are important for the situational appreciation required by practical wisdom. Emphasis on the emotions also brings into view forms of moral acknowledgment not sufficiently emphasized in analytic philosophical ethics, such as (virtuous) caring and love. It also highlights the centrality in ethics of certain partialistic virtues such as parental love and friendship.

A second factor in the intellectual context of modern virtue ethics is its objection to decontextualized under-

standings of moral theory. For example, for a virtue ethicist, ethics is neither essentially impartial nor essentially partial. Rather there are both virtues and vices of impartiality, and virtues and vices of partiality relative to different fields. An understanding of partialistic vices, such as nepotism and favoritism (towards one of one's children, for example), and impartialistic vices (such as dispositions to seriously neglect near and dear in the interests of 'the common good'), requires a highly nuanced and contextualized understanding of the boundaries between these vices and allied virtues.

Related to the above feature of virtue ethics is a third and much-emphasized factor: skepticism about the codifiability of ethics. Virtue ethicists favor a conception of ethical expertise that emphasizes experience and practical wisdom. Because of their emphasis on detail, particularity, and complexity, they recognize that moral expertise, which relies on knowledge of, for example, medicine, business, educational, and other practices, does not generalize to all contexts.

A fourth factor is a desire to make the concept of living well central to moral evaluation. Commonly for virtue ethicists, the supposed gap between the good life and the moral life is reduced via the desideratum of living a virtuous life. Living a virtuous life reduces this gap primarily through a reconfiguration of the contrast between altruism and self-interest. A virtuous agent, characteristically and in normal circumstances, enjoys the moral life as Aristotle points out, because her capacities for strength, love, concern, respect, and lack of resentment, enable her to see her own concerns as not sharply in contrast with the interests of others.

4. *Virtue Ethics and Applied Ethics*

The application of virtue ethics to social practice raises the question of an epistemology suitable for virtue ethics. Such an epistemology may be monological or dialogical. Commonly it is thought that virtue ethics is inherently monological because the arbiter of rightness of action is thought to be the (or a) virtuous agent, that is, an agent of good moral character, practical wisdom, fine sensibilities, and the ability to appreciate a situation and context in fine detail. Nonvirtuous agents simply emulate virtuous ones. However, even virtuous agents are necessarily limited because of their human condition (e.g., a virtuous man cannot have the experiences of women who are contemplating an abortion). Arguably, a suitable virtue ethical epistemology for the world of practice, for example, education, medicine, business, aid work, family relations, must be properly socially grounded. Participants in a practice, recognizing their epistemic and other limitations, exercise the moral/epistemic virtues of practice, which include the dialogical virtues. These of course are legion, including imaginative deliberation, insight, perseverance, respect in various forms, open-mindedness, facilitativeness, and an absence of those vices which distort practical wisdom and communication, for example, cowardice, rigidity, hostility, callousness, excessive love of power, squeamishness, and lack of concern for or excessive concern for fairness. For a virtue ethicist, determining whether a participant is, for example, excessively concerned with fairness or insufficiently concerned with fairness is often a matter of fine judgment applied in contexts where many values are at stake, and knowledge and appreciation of relevant facts, including the feelings, preferences and so forth of affected parties, commonly is gleaned only through well-conducted dialogue.

5. *Skepticism About Character*

One of the most interesting recent criticisms of virtue ethics derives from the social sciences, specifically 'situationist' psychology (Ross and Nisbett 1991). According to some philosophers who make use of this psychology, virtue ethics is inadequate because it is based on a characterological psychology that is empirically inadequate. According to situationist psychology:

> Behavioral variation across a population owes more to situational differences than dispositional differences among persons. Individual dispositional differences are not as strongly behaviorally individuating as we might have supposed … (Doris 1998, p. 507)

The claim that experiments such as Milgram's (1963) demonstrate the empirical inadequacy of virtue ethics because there is a lack of cross-situational consistency in behavior, betrays a lack of understanding of the virtue ethicist's concept of virtue. For a virtue ethicist, a virtue is not simply a disposition to perform acts of a certain type (e.g., beneficent acts). The possession of virtue requires also the possession of fine inner states. Admittedly, Milgram's experiments show that a remarkable number of subjects administer electric shocks of considerable severity, in experimental situations of a certain type. The tendency to perform beneficent acts is arguably not as robust as one might hope. However, there was considerable variation in the mental states of those prepared to administer such shocks, and these differences may point to character traits of, for example, compassion, benevolence, respect for authority and commitments, which manifest themselves in various ways in dilemmatic situations. According to Milgram, the subjects differed markedly in their emotional reactions to increasingly pressuring requests to administer shocks. Milgram claims that an unanticipated effect 'was the extraordinary tension generated by the procedures.' He says: 'One might suppose that a subject would simply break off or continue as his conscience dictated. Yet this is far from what happened. There were striking reactions of tension and emotional strain.' While some subjects reported and manifested intense strain and

anguish, however, 'some subjects (who had administered maximum shocks) had remained calm throughout the experiment, and displayed only minimal signs of tension from beginning to end' (Milgram 1963, p. 376).

For the virtue ethicist, emotional reactions are crucial for the attribution of virtue and vice, and the mark of the virtuous is the propensity to violate a 'virtue-rule' such as 'Be benevolent' only with reluctance, anguish, distress. The degree to which these emotions are present will vary according to the circumstances. Furthermore, it is clear from Milgram's account that the anguished saw themselves as operating under a rather severe virtue-dilemma. A benevolent agent may also possess the virtues of respect for authority, trust, honoring commitments, and appreciation of (the value of) knowledge. All of these virtues were in play during the experiment (1963). Subjects rightfully assessed the experimental situation as warranting trust, the honoring of a commitment, and as promoting a worthy objective. Now where virtues of, for example, trust and fidelity appeared in acute dilemmatic relation to another—nonmaleficence—most subjects arguably gave insufficient weight to the latter. But virtue ethicists are not committed to the view that many actual agents are capable of resolving such dilemmas correctly when under severe stress. The degree of virtue in agents is however able to be determined by their emotional reactions to their admittedly faulty resolutions of dilemma (e.g., reactions of anguish or indifference). And those differences were evident. It is open to the virtue ethicist to argue that while trust was not misplaced, and though obedience to authority is a virtue, in this case most subjects displayed this virtue to excess. Arguably, this failure hindered the exercise of the virtue of nonmaleficence present in those who suffered anguish.

According to virtue ethicists, good character is a centrally important component of the good life for an individual in society, and virtue ethics is a philosophically rich and fruitful approach to moral theory. The further development of this theory requires cooperation with the social sciences, if a richer understanding of virtue, as an excellence of character in human beings, is to be obtained.

See also: Ethical Behavior, Evolution of; Ethics and Values; Kantian Ethics and Politics; Moral Reasoning in Psychology; Moral Sentiments in Society; Religion: Morality and Social Control; Values, Sociology of; Wisdom, Psychology of

Bibliography

Aristotle 1955 *Nicomachean Ethics* [trans. Thomson J A K]. Penguin, London

Anscombe E 1958 Modern moral philosophy. *Philosophy* **33**: 1–19

Annas J 1993 *The Morality of Happiness*. Oxford University Press, New York and Oxford

Baron M W, Pettit P, Slote M 1997 *Three Methods of Ethics*. Blackwell, Oxford, UK

Crisp R, Slote M (eds.) 1997 *Virtue Ethics*. Oxford University Press, Oxford and New York

Doris J M 1998 Persons, situations, and virtue ethics. *Noûs* **32**: 504–30

Driver J 1989 The virtues of ignorance. *The Journal of Philosophy* **LXXXVI**: 373–84

Driver J 1996 The virtues and human nature. In: Crisp R (ed.) *How Should One Live? Essays on the Virtues*. Oxford University Press, Oxford and New York

Foot P 1978 *Virtues and Vices and other Essays in Moral Philosophy*. Blackwell, Oxford, UK

Fromm E 1977 *The Anatomy of Human Destructiveness*. Penguin, London

Horney K 1951 *Neurosis and Human Growth: The Struggle for Self Realization*. Routledge and Kegan Paul, London

Hume D 1978 *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd edn. Clarendon Press, Oxford, UK

Hurka T 1999 Three faces of flourishing. *Social Philosophy and Policy* **16**: 44–71

Hurka T 1992 Virtue as Loving the Good. *Social Philosophy and Policy* **9**: 149–68

Hursthouse R 1991 Virtue theory and abortion. *Philosophy and Public Affairs* **20**: 223–46

Hursthouse R 1995 Applying virtue ethics. In: Hursthouse R, Lawrence G, Quinn W (eds.) *Virtues and Reasons: Philippa Foot and Moral Theory*. Clarendon Press, Oxford, UK, pp. 57–75

Hursthouse R 1996 Normative Virtue Ethics. In: Crisp R (ed.) *How should one Live? Essays in the Philosophy of Virtue*. Clarendon Press, Oxford, UK

Hursthouse R 1997 Virtue ethics and the emotions. In: Statman D (ed.) *Virtue Ethics: A Critical Reader*. Edinburgh University Press, Edinburgh, UK

Hursthouse R 1999 *On Virtue Ethics*. Oxford: Oxford University Press.

Kant I 1996 The doctrine of virtue. In: MacGregor, M (ed. and trans.) *The Metaphysics of Morals*. Cambridge University Press, Cambridge, UK

Louden R 1984 On some vices of virtue ethics. *American Philosophical Quarterly* **21**: 227–36

MacIntyre A 1984 *After Virtue*. University of Notre Dame Press, Notre Dame, IN

Martineau J 1891 *Types of Ethical Theory*, 2 vols. Oxford University Press, Oxford, UK

McDowell J 1979 Virtue and Reason. *Monist* **62**: 331–50

Milgram S 1963 Behavioral study of obedience. *Journal of Abnormal and Social Psychology* **67**: 371–8

Milgram S 1974 *Obedience to Authority*. Harper and Row, New York

Nietzsche F 1973 *Beyond Good and Evil* [trans. Hollingdale R J]. Penguin, London and New York

Nietzsche F 1996 *The Genealogy of Morals* [trans. Smith D]. Oxford University Press, Oxford, UK

Oakley J 1996 Varieties of virtue ethics. *Ratio* **9**: 128–52

Pettit P 1997 The consequentialist perspective. In: Baron M W, Pettit P, Slote M (eds.) *Three Methods of Ethics*. Blackwell, Oxford, UK

Ross L, Nisbett R E 1991 *The Person and the Situation*. Temple University Press, Philadelphia, PA

Slote M 1983 *Goods and Virtues*. Oxford University Press, Oxford, UK
Slote M 1992 *From Morality to Virtue*. Oxford University Press, New York
Slote M 1996 Agent based virtue ethics. In: French P A, Uehling T E Jr, Wettstein H K (eds.) *Midwest Studies in Philosophy XX. Moral Concepts*. University of Notre Dame Press, Notre Dame, IN
Slote M 1998a The justice of caring. *Social Philosophy and Policy* **15**: 171–95
Slote M 1998b Nietzsche and virtue ethics. *International Studies in Philosophy* **XXX**: 23–7
Solomon D 1988 Internal objections to virtue ethics. In: French P A, Uehling T E Jr, Wettstein H K (eds.) *Midwest Studies in Philosophy XIII. Ethical Theory: Character and Virtue*. University of Notre Dame Press, Notre Dame, IN, pp. 428–41
Solomon R C 1998 The virtues of a passionate life: Erotic love and 'the will to power'. *Social Philosophy and Policy* **15**: 91–118
Statman D (ed.) 1997 *Virtue Ethics: A Critical Reader*. Edinburgh University Press, Edinburgh, UK
Swanton C 1995 Profiles of the Virtues. *Pacific Philosophical Quarterly* **76**: 47–72
Swanton C 1997 The supposed tension between 'strength' and 'gentleness' conceptions of the virtues. *Australasian Journal of Philosophy* **75**: 497–510
Swanton C 1998 Outline of a Nietzschean virtue ethics. *International Studies in Philosophy* **XXX**: 29–38
Trianosky G V y 1990 What is Virtue Ethics All About? *American Philosophical Quarterly* **27**: 335–44
Watson G 1990 On the primacy of character. In: Flanagan O, Rorty A (eds.) *Identity Character and Morality: Essays in Moral Psychology*. MIT Press, Cambridge, MA pp. 445–69

C. Swanton

Vision for Action: Neural Mechanisms

Sensory input received by the visual receptors from an object in space results in a complex pattern of excitation from which object properties relevant to different functions have to be extracted. This selection is the basis for a specific mechanism called visuomotor transformation, the shaping of the motor effectors into configurations precisely adapted to capturing, manipulating, using, and transforming visual objects. In primates, the remarkable evolution of the hand into a grasping apparatus represents an ultimate acquisition for skilled interaction with objects and use of tools. This article will focus on anatomical and functional analysis of neural mechanisms involved in visuomotor transformation.

Beginning in the 1970s, the anatomical description of cortical visual pathways in the monkey has led to the formulation of the 'two visual systems' model (Ungerleider and Mishkin 1982). According to this model, the primary visual cortex in the occipital lobe projects along two main streams that reach, respectively, the posterior parietal lobe (the dorsal pathway) and the inferotemporal cortex (the ventral pathway). Although the two pathways are densely interconnected with each other, they are now viewed as relatively independent functional entities.

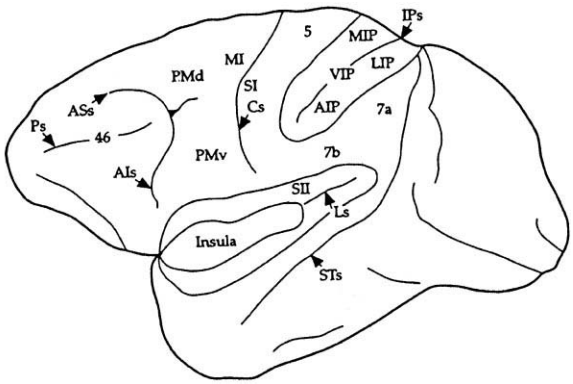

Figure 1
Lateral view of the left hemisphere of a monkey brain. Two of the main sulci have been opened to show the 'buried' areas: the intraparietal sulcus (IPs) and the lateral sulcus (Ls, which separates the parietal lobe from the temporal lobe). The other main sulci are the central sulcus (Cs) which separates the parietal lobe from the frontal lobe, the arcuate sulcus (As) and the principal sulcus (Ps) within the frontal lobe, and the superior temporal sulcus (STs) within the temporal lobe. The anterior part of the parietal lobe, which includes somatosensory areas SI and SII (in the upper bank of the lateral sulcus), is not concerned with visuomotor transformation. The posterior part of the parietal lobe includes areas 7 and 5 on the cortical surface, and the intraparietal areas (medial, lateral, ventral, anterior intraparietal areas, respectively). The motor strip of the frontal lobe includes the primary motor cortex (MI, the command area), and the dorsal and ventral premotor cortex (PMd, PMv, respectively).

The dorsal pathway, which will be under scrutiny in this article, was once considered as a space channel. Indeed, its experimental destruction in the monkey produced spatial disorientation of the animal. Later on, a more complete description of the anatomical organization of the posterior parietal cortex, which is outlined in Fig. 1, led to a new conception of its functional role. First, a number a specialized areas, many of them located in the intraparietal sulcus (e.g., LIP, MIP, AIP), were added to the classical cytoarchitectonic areas (e.g., areas 7 and 5). Second, these newly described areas were found to be connected with motor regions of the frontal cortex. These connections appear to be selectively organized: parietal area MIP is preferentially connected with the upper part of the premotor cortex (PMd in Fig. 1), whereas AIP is preferentially connected with its lower part (PMv).

Both PMd and PMv project to the motor cortex itself, where the commands to the relevant muscles are generated. Thus, the dorsal pathway, whereby visual information is transferred from the primary visual cortex to the premotor and motor cortex, can better be conceived as a pathway specialized for visuomotor transformation.

1. A Mechanism for Reaching Objects in Visual Extrapersonal Space

Reaching for a visual object requires that the set of coordinates in which the object-oriented movement is computed must have its origin on the agent's body (egocentric coordinates). This constraint stems from the fact that object position in space must be reconstructed by adding signals from eye, head, and body positions for the arm to be ultimately directed at the proper location. Determining object position in space by computing its retinal position alone would be inappropriate, because this locus in space would correspond to different retinal positions according to respective eye, head, and body positions (Jeannerod 1988). The following paragraphs describe some of the mechanisms which have been identified in recent experiments in monkeys and which could account for transforming retinal (eye centered) coordinates into (body centered) egocentric coordinates. These mechanisms are primarily located in the intraparietal sulcus.

An influential hypothesis proposes that the basic information about where to reach for an object is the motor command sent to the eyes to move in its direction (the eye position signal). Neurons in area 7a have been shown to encode visual stimuli for a specific position of the eye in the orbit, i.e., they fire preferentially when the object stimulates a given location on the retina (the neuron's receptive field), and when the gaze is fixating in that direction (Andersen and Mountcastle 1983). These neurons would provide a representation of visual space that could subsequently be projected to the premotor cortex for guiding the arm reach. This mechanism is apparently compatible with the organization of visuomotor behavior: the gaze systematically moves first to the location of the desired stimulus, followed by the arm movement within about 150 ms. A closer look, however, shows that, because the motor command to the arm muscles anticipates arm displacement by about the same amount, the motor commands to the eye and to the arm are in fact more or less synchronous (Biguer et al. 1982). Duhamel and his colleagues found the solution to that enigma. They discovered in area LIP neurons which fire in response to a stimulus briefly presented, not within their receptive field, but within the area of space where their receptive field will project after an eye movement is made. This effect begins some 80 ms before the eye movement starts (Duhamel et al. 1992). The retinal position of receptive fields is thus modified prior to the occurrence of an eye movement. This remapping of the representation of space in the parietal cortex accounts for perceptual stability of the visual scene during eye movements. In addition, the existence of this mechanism demonstrates that signals related to the intended eye position are available to the motor cortex. In area VIP, neuron discharges testify to a different form of remapping. Their receptive fields remain in the same spatial position irrespective of eye position in the orbit. In this case, therefore, the receptive fields appear to be anchored, not to the eye, but to the head, i.e., the receptive fields move across the retina when the eyes move (Duhamel et al. 1997).

Eye and head position signals, however, should not be sufficient in themselves for generating a directional signal for guiding the arm. Besides the position of the stimulus on the retina and the position of the eye in the orbit, the position of the arm with respect to the body must also be available to the arm motor command. This function is achieved by neurons presenting receptive fields in two sensory modalities. Neurons in area MIP, for example, respond to both a localized visual stimulus and a cutaneous stimulation applied to one of their hands. These bimodal neurons are strongly activated when the monkey reaches with that hand to the visual target within their receptive field (see Andersen et al. 1997, Colby 1998).

Neurons located in motor areas outside the parietal cortex also present bimodal activation. Graziano and his colleagues found bimodal neurons in the ventral premotor cortex: these neurons, which in addition encode the direction of arm movements, have tactile receptive fields and visual receptive fields that extend from the tactile area into adjacent space. For those neurons which have their tactile receptive field on the arm, the visual receptive field moves when the arm is displaced (Graziano et al. 1994). These visual receptive fields anchored to the arm encode visual stimulus location in body (arm) coordinates and thus may provide a solution for directional coding of reaching. In what way does this premotor mechanism relate to parietal mechanisms? What is the role of connections between parietal areas and the premotor cortex? These points remain unclear. It is likely that, instead of a single directional signal used for all purposes, multiple transformations and multiple spatial maps are generated and adapted to each purpose: guiding the eye, guiding the arm, stabilizing the visual world, etc.

Data from single cell recordings are supported by experimental data in human studies which clearly point to the parietal lobe as the site of visuomotor transformation for reaching. Although human data are less direct than monkey data, they provide deeper insight into the function due to the unique human ability to execute tasks without having to learn them. Functions like visually guided reaching can be examined in more naturalistic situations. Observation of patients with focal posterior parietal lesions in one

hemisphere show deep alterations of reaching movements, usually limited to the hand contralateral to the lesion: kinematics are altered, and large reaching errors are observed. This misreaching is more marked in the condition where the visual target is presented outside the fixation point and increases when the target distance from the fixation point increases (e.g., Milner et al. 1999). These clinical findings are corroborated by brain imaging studies (based on changes in localized cerebral blood flow) in normal subjects during pointing at visual targets. The activated area, in the parietal lobe contralateral to the pointing hand, corresponds closely to the location of lesions that produce misreaching (Faillenot et al. 1997). Finally, the involvement of the posterior parietal cortex in visuomotor control is confirmed by transcranial magnetic stimulation (TMS). This technique allows the application, via magnetic induction, of very brief single shocks to a given cortical zone. When TMS was applied to posterior parietal cortex shortly after onset of a reaching movement directed to a visual target, it prevented online corrections needed to reach the target. TMS only affected the arm contralateral to the stimulated cortex, which excludes the possibility that the stimulation perturbed visual processing of target location: only the adaptation of the motor command to the visual position of the target was perturbed (Desmurget et al. 1999).

2. *The Case of Hand and Finger Movements*

Grasping is often the ultimate achievement of reaching. For this reason, the two functions can hardly be dissociated. Grasping, however, relies on a quite different system than reaching: it involves distal segments, multiple degrees of freedom, and a high degree of reliance on tactile function. It is an ensemble of visuomotor transformations which preshape the hand while it is transported to the object location. Preshaping deals with object features which are not relevant to reaching, like object size, shape, or orientation (Jeannerod et al. 1995).

Indeed, anatomical data suggest that the cortical pathway connecting posterior parietal and premotor finger areas and controlling grasping is separate from the pathway controlling reaching. Hand movement-related neurons were disclosed by Sakata and his colleagues (1995) in a small zone within the anterior part of intraparietal sulcus, corresponding to an area (AIP) closely connected with the ventral part of the premotor cortex (PMv in Fig. 1). Neurons from AIP were recorded in monkeys trained to manipulate various types of objects, which elicited from the animal different motor configurations of the hand: most of them were selective for grasping objects of a given shape. Task-related neurons were classified as 'motor dominant' or 'visual dominant' according to whether their discharge related more to the action of grasping or to the inspection of the object. Most importantly, neuron activity was not influenced by changing the position of the object in space, which suggests that they were related to distal hand and finger movements (for which spatial position is irrelevant) rather than to proximal (reaching) movements of the arm (Sakata et al. 1995). Finally, a distinct population of neurons sensitive to the 3-D orientation of the longitudinal axis of visual stimuli was found in the caudal part of the intraparietal sulcus. It is therefore likely that these neurons process the 3-D characteristics of the object and that the output of this processing is sent to AIP.

Transient inactivation of the AIP region, by locally injecting a GABA agonist (muscimol), produces a subtle change in the performance of visually guided movements during grasping. Lack of preshaping and grasping errors are observed in tasks requiring a precision grip. In addition, there is a clear-cut dissociation of the muscimol effects on grasping, which is altered, and reaching, which is preserved. These results (Gallese et al. 1994) support the view that these parietal neurons play a specific role in the visual guidance of grasping.

In humans, clinical observations also stress the role of the posterior parietal cortex as a structure controlling the visuomotor action of grasping. During prehension of objects, parietal lesioned patients show little or no preshaping with their contralateral hand; they tend to open their finger grip too wide with respect to object size and to close it only when their fingers come into contact with the object; the opposition axis (the axis which passes through the two opposing fingertips and the center of gravity of the object) is improperly oriented, which results in misplacement of fingertips on the object surface (Jeannerod et al. 1994). In normal subjects, the above brain imaging study (Faillenot et al. 1997) revealed an area specifically involved in grasping, at the ventral extremity of the intraparietal sulcus, also including part of the somatosensory area S II. It is likely that this cortical zone includes the human homologue of the monkey AIP area.

An intriguing question thus remains to be discussed: in which system of coordinates are finger movements during grasping coded? The main aspect of grasping, grip formation, has been described using parameters such as grip size, orientation of opposition axis, number of fingers involved, etc., which, arguably, should not depend on the object position in space. This is also the case for another aspect of grasping, the encoding of grasp and lift forces: these parameters depend on visual cues for object weight, for which object location is irrelevant. There are limits to this reasoning, however. Reaching and grasping cannot be considered in isolation, as they represent two aspects of the same action, reaching to grasp an object. The need for coordination between these two aspects becomes apparent when measuring the orientation of the opposition axis during grasping. The position of

the fingertips on the object surface should only be determined by object shape and 3-D characteristics in order to achieve a stable grasp. Paulignan et al. (1997), however, using upright cylinders as targets (e.g., objects affording multiple possible positions for the fingertips) showed that orientation of the opposition axis changed as a function of the cylinder position in the workspace. This result indicates that the action of grasping must encode not only finger movements but also movements of the proximal joints: due to biomechanical limitations of the arm, the configuration of the upper limb cannot be the same for grasping the same object at two different locations.

Even though reaching and grasping cannot be considered as functionally independent (in spite of segregation of their anatomical pathways), they process different types of information and operate at different rates. Experiments using visual 'perturbations' synchronized with movement onset clearly demonstrate this difference (Paulignan et al. 1991a, 1991b). If, for example, the target object to be reached by the subject jumps (by means of an optical device) by 10 degrees to the right as the subject starts reaching in its direction, a corrective movement is produced so that the new target is adequately reached: the first kinematic evidence for correction occurs around 100 ms following the perturbation. If, in another experiment, the object keeps the same location, but changes size (e.g., increases) as the subject starts to move, it takes at least 300 ms for the finger grip to adjust to the changing object size. This sharp contrast between the timing of the two types of corrections accounts for the functional differences between the two components of the action. Although fast execution during reaching has an obvious adaptive value (e.g., for localizing and catching moving targets), this is not the case for grasping, the function of which is to prepare fine manipulation of the object once it has been caught.

3. Conclusion

Posterior parietal cortex thus appears as a critical zone for organizing object-oriented action, whether movements are executed by the proximal or the distal channels. The specialization of the dorsal visual pathway for visuomotor transformation is complementary with that of other brain regions, mostly located within the ventral pathway, and specialized for object identification and recognition. This suggests the existence of a dual mode of processing for object properties like spatial location, orientation, shape, or size. Perception requires that object properties are processed within an allocentric frame of reference, that is, independent from the perceiver and where the relationships between objects within the visual array are preserved. Action, by contrast, as has been shown in this article, requires that information is processed within an egocentric frame of reference where everything in the visual array that is not the target for the movement is ignored.

See also: Attention and Action; Eye Movements and Perception; Motion Perception Models; Object Recognition: Theories; Perception and Action; Perceptual Organization

Bibliography

Andersen R A, Mountcastle V B 1983 The influence of the angle of gaze upon the excitability of light-sensitive neurons of the posterior parietal cortex. *Journal of Neuroscience* **3**: 532–48

Andersen R A, Snyder L H, Bradley D C, Jing Xing 1997 Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* **20**: 303–30

Biguer B, Jeannerod M, Prablanc C 1982 The coordination of eye, head and arm movements during reaching at a single visual target. *Experimental Brain Research* **46**: 301–4

Colby C L 1998 Action-related spatial reference frames in cortex. *Neuron* **20**: 15–24

Desmurget M, Epstein C M, Turner R S, Prablanc C, Alexander G E, Grafton S T 1999 Role of the posterior parietal cortex in updating reaching movements to a visual target. *Nature Neuroscience* **2**: 563–7

Duhamel J R, Bremmer F, BenHamed S, Graf W 1997 Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature* **389**: 845–8

Duhamel J R, Colby C L, Goldberg M E 1992 The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**: 90–2

Faillenot I, Toni I, Decety J, Grégoire M C, Jeannerod M 1997 Visual pathways for object-oriented action and object recognition. Functional anatomy with PET. *Cerebral Cortex* **7**: 77–85

Gallese V, Murata A, Kaseda M, Niki N, Sakata H 1994 Deficit of hand preshaping after muscimol injection in monkey parietal cortex. *Neuroreport* **5**: 1525–9

Graziano M S A, Yap G S, Gross C G 1994 Coding of visual space by premotor neurons. *Science* **266**: 1054–6

Jeannerod M 1988 *The Neural and Behavioural Organization of Goal-Directed Movements*. Oxford University Press, Oxford, UK

Jeannerod M, Arbib M A, Rizzolatti G, Sakata H 1995 Grasping objects. The cortical mechanisms of visuomotor transformation. *Trends in Neuroscience* **18**: 314–20

Jeannerod M, Decety J, Michel F 1994 Impairment of grasping movements following a bilateral posterior parietal lesion. *Neuropsychologia* **32**: 369–80

Milner A D, Paulignan Y, Dijkerman H. C, Michel F, Jeannerod M 1999 A paradoxical improvement of misreaching in optic ataxia. New evidence for two separate neural systems for visual localization. *Proceedings of the Royal Society London B* **266**: 2225–9

Paulignan Y, Frak V G, Toni I, Jeannerod M 1997 Influence of object position and size on human prehension movements. *Experimental Brain Research* **114**: 226–34

Paulignan Y, Jeannerod M, MacKenzie C, Marteniuk R 1991a Selective perturbation of visual input during prehension movements. I. The effects of changing object size. *Experimental Brain Research* **87**: 407–20

Paulignan Y, MacKenzie C, Marteniuk R, Jeannerod M 1991b Selective perturbation of visual input during prehension

movements. II. The effect of changing object position. *Experimental Brain Research* **83**: 502–12
Sakata H, Taira M, Murata A, Mine S 1995 Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. *Cerebral Cortex* **5**: 429–38
Ungerleider L, Mishkin M 1982 Two cortical visual systems. In: Ingle D J, Goodale M A, Mansfield R J W (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA, pp. 549–86

M. Jeannerod

Vision, High-level Theory of

The concept of 'high-level vision' is used to denote aspects of vision that are influenced by information stored in memory such as an object's identity, its name, or the context in which it belongs. It is usually distinguished from low-level visual processing, which is driven primarily by the visual input. High-level vision involves functions such as object recognition, face perception, mental imagery, picture naming, visual categorization, or scene perception.

1. *Background*

In hierarchical models of vision (e.g., Marr 1982, see also *Marr*, *David* (1945–80)), higher levels of visual processing operate on the building blocks delivered by more primitive visual mechanisms. In Marr's approach to perception, each stage has its own algorithms and its own format of representing processing output. The primal sketch represents the basic features and their spatial relations (low-level vision) and the 2,5-D sketch represents surfaces and shapes from the observer's viewpoint (intermediate-level vision). Recognition, that is, matching of object descriptions stored in visual memory with visual input, Marr believed only possible when the full 3-D structure of an object and its components is represented in a so-called 3-D object model (high-level vision).

Marr's approach has been highly influential as a framework to integrate disparate work on the processing of, among others, luminance, texture, shading, color, stereo, or motion (see Palmer 1999). However, his assumption of strictly bottom-up processing in sequential stages has gradually lost attraction. One reason for this theoretical evolution has been the increased popularity in the 1980s of so-called interactive activation models in which different levels of processing interact with one another (see also *Connectionist Approaches*; *Artificial Neural Networks*: *Neurocomputation*): Partial outputs from lower-level processes initiate higher-level processes and the outputs of higher-level processes feed back into the lower-level processes.

When different levels are distinguished in more recent theories, great care is taken to avoid the assumption of strict sequential processing. However, the issue of where early vision ends, where high-level vision starts, and how to demarcate the two has remained important in current theorizing. Pylyshyn (1999) has proposed the notion of cognitive impenetrability as the major criterion for separating early (or low-level) vision from later (or higher-level) vision. Early visual processes run autonomously; they cannot be influenced by what we know about the world or expect in a specific situation. Most aspects of visual processing which seem cognitive (e.g., recovering unique 3-D percepts from ambiguous 2-D inputs) are attributed to embodied natural constraints derived from principles of optics and projective geometry, internal to the visual mechanisms themselves. Only effects of object-specific beliefs on earlier visual processing would count as cognitive penetration. A large number of so-called 'top-down effects' of higher-level processes on early vision are then explained as effects of attention (preceding early vision) or as post-perceptual effects on the decision stage based on the perceptual output. For example, perceptual learning, perceptual expertise and effects of categorization on discrimination might be attributed to learning to focus attention to the important features or attributes of a stimulus. Semantic effects of the context of the whole scene (e.g., a kitchen) on the identification of the objects embedded within the scene (e.g., bread) have been shown to result from response bias rather than from genuine perceptual effects (e.g., Henderson and Hollingworth 1999).

Implicit in Pylyshyn's (1999) effort to delineate early vision from cognitive influences is his belief—perhaps shared by many vision scientists—that real scientific progress is possible only in early vision. Arguably, however, in recent years considerable progress has been made in the area of high-level vision as well (e.g., Edelman 1999, Ullman 1996). Most of this progress has been made in object recognition, and, thus, this topic will be covered predominantly (see also *Binocular Space Perception Models*).

2. *Object Recognition*

2.1 *Biederman (1987)*

Object recognition concerns the identification of an object as a specific entity (i.e., semantic recognition) or the ability to tell that one has seen the object before (i.e., episodic recognition). Interest in object recognition is at least partly caused by the development of a new theory of human object recognition by Biederman (1987). Building on Marr's recognition ideas, where perceptually derived 3-D volumetric descriptions are matched to those stored in memory, Biederman in addition incorporated Gestalt principles of perceptual

grouping and some newer ideas from computer vision into his theory.

Biederman proposed that objects are identified at their entry level on the basis of a structural description of their components and their spatial relations. All object parts are modeled by a limited number of simple primitives called 'geons,' a specific subset of generalized cones. A generalized cone is the volume swept out by a cross section moving along an axis. Geons are defined by the categorical attributes on the following dimensions: the axis (straight or curved), the shape of the cross-section (straight or curved, symmetric or asymmetric), and the size of the cross-section (constant, expanded, or expanded and contracted).

One major claim in Biederman's Recognition-by-Components (RBC) theory is that geons themselves are viewpoint-invariant because the underlying dimensions can be distinguished on the basis of so-called 'nonaccidental properties' or NAPs (a notion developed in computer vision). These are special features in the image (such as collinearity, curvilinearity, cotermination, parallelism, and symmetry) that are reliable in the sense that they are most likely caused by similar characteristics in 3-D space (under the assumption of a general viewpoint). For example, rarely do nonparallel structures in 3-D space project to parallel structures in the image. Biederman argues that Gestalt principles of grouping support the extraction of these NAPs and that they are thus detected fast and reliably enough to support bottom-up identification of geons.

The spatial relations between object components also have to be specified by a number of parameters and in later versions of the theory, a few coarse metric attributes of the geons themselves must be specified too. The combinatorial possibilities of even only two or three geons are then enormous. This allows Biederman to argue that the recovery of two or three geons is enough to recognize complex objects quickly even when they are occluded, rotated in depth, degraded, or lacking color or texture.

This claim is supported empirically by experiments in which parts of objects (in fact, line drawings of objects) were deleted until only two or three parts were left. For most moderately complex objects consisting of six or nine parts (e.g., an airplane, an elephant) three parts were still enough to recognize the object. Other supporting evidence comes from recognition of contour-deleted line drawings (which is possible as long as the geons are recoverable), priming with contour-deleted pictures (which is situated at the level of the parts, not the image fragments and not the whole picture), and viewpoint-invariance of 3-D object recognition (under certain limited conditions).

While Biederman and his associates were accumulating evidence for his RBC theory, others were accumulating evidence for strong effects of viewpoint changes on 3-D object recognition (see, e.g., Bülthoff et al. 1995). This has given rise to a lot of controversy between viewpoint-dependent and independent theories of object recognition, often centered around the relevance of specific types of objects used as experimental stimuli (e.g., line drawings vs. rendered 3-D objects, all sorts of novel objects such as paperclips, amoebas, and 'greebles') or specific types of experimental tasks (e.g., matching, naming, priming). Because the focus in this review is on theory instead of empirical results, this debate will not be reviewed here (but see also *Binocular Space Perception Models*). Instead, two promising new theoretical developments will be highlighted.

2.2 *Ullman (1996)*

Ullman's (1996) book *High-level vision: Object recognition and visual cognition*, brings together his achievements in three diverse sub-domains of high-level vision: (a) object recognition (alignment, combination of views), (b) visual cognition (classification, visual routines), and (c) sequence-seeking and counter-streams.

In the area of object recognition, Ullman proposed two different theories. According to his alignment theory, object recognition consists of a search for a particular object model in memory, M_i, and a particular transformation, T_{ij}, that will maximize the fit between (M_i, T_{ij}) and the viewed object, V. The set of the allowed transformations, such as changes in scale, position, or 3-D orientation, can be applied either to the incoming image (more bottom-up) or to the stored model (more top-down). The simplest version of this theory works well for simple, rigid objects. Because the set of relevant transformations is restricted in that case, only three points are needed to be able to establish a correspondence between the image and the model. This so-called 'three-point alignment scheme' explicitly uses 3-D models but works well only when the features that are used for alignment are visible from different views. For all other cases, his second theory, recognition by the combination of views, works better. This does not explicitly use 3-D models but a representation of 3-D objects as a small collection of 2-D views. Object recognition then consists of matching the viewed object against a linear combination of these 2-D views. Extensions of this method can even handle cases with partial views due to occlusion and cases with nonrigid transformations.

Object recognition is concerned with specific shapes undergoing well-defined transformations, which are relatively easy to specify and compensate for. Object classification, on the other hand, is more difficult because it is not obvious what the critical shape features of object classes like 'dog' or 'chair' are. Ullman demonstrated that relatively simple pictorial comparisons based on extensions of the alignment method can begin to achieve useful classification. Another contribution of Ullman in the domain of

visual cognition concerns the visual analysis of shape properties and spatial relations. For example, judging whether an *X* lies inside or outside a closed arbitrary curved shape, which of two ellipses is most elongated, whether two dots lie on a common curve, etc., is notoriously difficult for artificial vision systems, while humans can make such judgments remarkably well. Ullman believes that the flexibility of human spatial perception and cognition depends on so-called 'visual routines,' efficient sequences of basic operations that are hardwired into the visual system. For different spatial tasks, different visual routines are compiled consisting of different sequences of the same basic operations such as shifting the processing focus, boundary tracing, and marking.

Ullman's most general contribution to the theory of high-level vision is his proposal for a general model of information flow in the visual cortex. At the heart of the model are a process (called 'sequence-seeking') and a structure (called 'counter-streams'). Sequence seeking is a search for a sequence of transformations that link an input image with a stored object description. It works in two directions, bottom-up as well as top-down, and it explores a large number of alternative sequences in parallel. Counter-streams are two complimentary pathways, an ascending one from low- to high-level visual areas and a descending one from high- to low-level visual areas. The basic idea is that bottom-up processing (starting from the input image) is performed in the ascending pathway and that top-down processing (starting from the stored models) is performed in the descending pathway. The interactions between the two complementary pathways support the integration of bottom-up and top-down processing.

With this model it is possible to match images and models by simultaneously exploring multiple alternatives instead of testing them sequentially. The model also does justice to many well-know anatomical and physiological facts about the visual cortex. For example, the visual cortex has a massively parallel architecture and descending or feedback projections are equally abundant as their ascending or feed forward counterparts. Finally, Ullman's model offers a detailed account of many well-known psychological findings such as priming, context effects, and perceptual learning.

2.3 Edelman (1999)

Another important theoretical step forward in high-level vision is Edelman's (1999) book on *Representation and recognition in vision*. Edelman argues that the attempt to base recognition on geometrically reconstructed representations of distal objects (like Marr's 2,5-D sketch) suffers from serious theoretical and practical problems. First, it would lead to the homunculus problem and, more pragmatically, it appears, at least up till now, not to work, not even when using the most sophisticated computer vision techniques. The major reason for this failure might be that previous approaches have tried to achieve veridicality of object representation by aiming for isomorphisms between objects in the world and their corresponding internal representations (i.e., first-order isomorphism or representation *by* similarity). Instead, Edelman proposes to aim for the formally just as good isomorphism between the relations among several external objects and the relations among their corresponding internal representations (i.e., second-order isomorphism or representation *of* similarity).

Edelman then introduces the notion of a shape space, a formalism borrowed from mathematical statistics. Shape space is a metric space in which each point corresponds to a particular shape and in which geometric similarity between shapes can be defined rigorously as proximity, a quantity inversely related to distance. Edelman argues that shape similarity can be made not to suffer from the objections commonly raised against a metric-space approach to similarity (e.g., arbitrary choice of features, context dependence, and asymmetries of comparisons; see also *Categorization and Similarity Models*). Although a large number of parameters is needed to represent distal shape space fully, fewer may be sufficient if, as in second-order isomorphism, only relations between objects must be represented (e.g., a blending coefficient when morphing the shape of a cow and a pig).

A novel shape (or a novel view of a previously seen shape) can be localized in shape space based on its similarity to known objects (or stored views of known objects). A tuned unit (or module of units) that responds optimally to some shape (i.e., the landmark) and progressively less to progressively less similar shapes is enough to implement this mechanism. Edelman uses a radial basis function approximation network that can be trained (see also *Artificial Neural Networks: Neurocomputation*) and generalizes its response from a series of given views of an object to other views of that object. As a byproduct of this learning, the network will also respond progressively less to progressively less similar objects, precisely what is needed for it to work as an active landmark in internal shape space. This scheme not only works for known objects but also for novel objects. The trick is that a new object must be represented by a vector of proximities to several reference shapes or landmarks. For example, even if one has never seen a giraffe before, it is possible to compute the proximities to known animals such as a camel, a goat, a pig, or even a leopard.

Edelman calls this system a 'Chorus of Prototypes' to stress that the prototypical landmarks act together in representing shape, in contrast to a winner-take-all scheme such as Selfridge's Pandemonium (see also *Statistical Pattern Recognition*). Initial simulation results with this system are encouraging, especially in

light of the wide variety of tasks: identification of novel views of familiar objects, categorization of novel object views, discrimination among views of novel objects, local viewpoint invariance for novel objects, recovery of a standard view and of pose for novel objects, prediction of a novel view for a novel object, etc. None of its theoretical competitors (Biederman's RBC, Ullman's alignment) can deal with this variety of tasks.

There is also considerable neurobiological and psychophysical evidence that is compatible with Edelman's system. In terms of neurobiology, Edelman discusses properties of the receptive fields at lower levels of processing (broad tuning and graded response) and at higher levels of processing such as inferotemporal cortex (for a review, see Logothetis and Sheinberg 1996): selectivity to objects, ensemble encoding (i.e., sparse distributed representations) selective invariance, plasticity and learning, and speed of processing (which is important because Chorus is essentially a feed forward model). In terms of psychophysics, Edelman discusses the evidence for the well-known effects of viewpoint on object recognition (see Sect. 2.1).

3. *Other Topics*

A number of other topics could be covered under the umbrella notion of 'high-level vision' but some are addressed more extensively elsewhere in this encyclopedia. Because faces are 'objects' with special significance, face perception has become a research topic of its own, somewhat isolated from more general object perception (see also *Face Recognition Models*). Indeed, some argue that faces are processed differently, even in a separate area in the brain (i.e., the so-called 'fusiform face area'), and subject to a specific disorder, prosopagnosia, after brain damage to this and surrounding regions (see also *Face Recognition: Psychological and Neural Aspects*).

A second topic of special interest, which could be regarded as belonging to high-level vision, is mental imagery, the ability to create mental images and operate on them in tasks requiring visual or spatial judgments in the absence of visual input (see also *Mental Imagery*, *Psychology of*). In his *Image and Brain*, Kosslyn (1994) has defended the analogue nature of mental images and the use of visual processing mechanisms in mental imagery tasks on the basis of classic psychological experiments and more recent neuroimaging evidence obtained with Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI). In addition, he has developed a detailed psychological and neuroanatomical theory of high-level vision, consisting of seven major subsystems: visual buffer, attention window, encoding of object properties, encoding of spatial properties, associative memory, information look-up, and attention shifting. Each of these components is worked out further into its more detailed operations and then localized neuroanatomically on the basis of neuropsychological case studies of brain-damaged patients with specific deficits and available evidence from PET and fMRI studies.

Two other topics of high-level vision, not addressed elsewhere in this encyclopedia, are picture naming and inattentional blindness. Picture naming has always been an interesting task for cognitive psychology because it obviously involves a number of distinct representations and processes. Traditional cognitive theories have proposed different so-called 'box-and-arrow' models to chart the intermediate representations and information flow between them, and traditional experimental methods have been used to test hypotheses derived from these models on the basis of performance measures like error rates and response times. More recently, as with mental imagery, the models have become more detailed and neuroanatomically specific thanks to neuroimaging and patient work (e.g., Humphreys et al. 1999).

In the typical experimental paradigm to study inattentional blindness (e.g., Mack and Rock 2000), subjects receive a rather demanding task at fixation (e.g., reading a small letter) while other visual stimuli are present in the visual field. When asked about the irrelevant (distractor) stimuli afterwards, subjects recall very little about them, except perhaps some basic features like color or orientation. In a more spectacular demonstration along the same lines, sudden changes to even large and central parts of a rich visual display like a colored picture of a scene, go unnoticed when attention is drawn away from the change. This so-called 'change blindness,' as well as the other forms of inattentional blindness, suggest that our conscious visual experience of the world is really quite limited, despite the illusion that it is rich and visually detailed. These phenomena constitute interesting examples of how visual attention, visual perception, visual memory, and visual consciousness interact with one another in everyday circumstances, but little theoretical progress has been made in understanding this intricate interplay of separate psychological functions.

4. *Conclusion and Future Directions*

In the area of high-level vision a number of trends emerge. One is the application of computational methods (previously applied with some success to lower-level visual processing) to higher-level visual processing. A second trend is the increased consideration of the neural basis of high-level vision. Theories about high-level vision should be constrained by what we know about their possible implementation in the brain on the basis of monkey work, neuropsychological studies with brain-damaged patients, and neuroimaging studies with healthy subjects.

High-level vision is an active area of interdisciplinary research and much progress may be expected yet.

See also: Artificial Neural Networks: Neurocomputation; Binocular Space Perception Models; Categorization and Similarity Models; Color Vision; Color Vision Theory; Computational Imagery; Darwin, Charles Robert (1809–82); Face Recognition Models; Face Recognition: Psychological and Neural Aspects; Mental Imagery, Psychology of; Neural Plasticity in Visual Cortex; Occipital Lobe (Visual Cortex): Functional Aspects; Spatial Pattern, Analysis of; Statistical Pattern Recognition; Topographic Maps in the Brain; Vision for Action: Neural Mechanisms; Vision, Low-level Theory of; Vision, Psychology of; Visual Imagery, Neural Basis of; Visual Perception, Neural Basis of; Visual Space, Geometry of; Visual System in the Brain; Visual Systems: Dorsal and Ventral

Bibliography

Biederman I 1987 Recognition-by-components: A theory of human image understanding. *Psychological Review* **94**: 115–47

Bülthoff H H, Edelman S Y, Tarr M J 1995 How are three-dimensional objects represented in the brain? *Cerebral Cortex* **5**: 247–60

Edelman S 1999 *Representation and Recognition in Vision*. M.I.T. Press, Cambridge, MA

Henderson J M, Hollingworth A 1999 High-level scene perception. *Annual Review of Psychology* **50**: 243–71

Humphreys G W, Price C J, Riddoch M J 1999 From objects to names: A cognitive neuroscience approach. *Psychological Research* **62**: 118–30

Kosslyn S M 1994 *Image and Brain: The Resolution of the Imagery Debate*. M.I.T. Press, Cambridge, MA

Logothetis N K, Sheinberg D L 1996 Visual object recognition. *Annual Review of Neuroscience* **19**: 577–621

Mack A, Rock I 1998 *Inattentional Blindness*. M.I.T. Press, Cambridge, MA

Marr D 1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco

Palmer S E 1999 *Vision science: Photons to Phenomenology*. M.I.T. Press, Cambridge, MA

Pylyshyn Z W 1999 Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences* **22**: 341–423

Ullman S 1996 *High-level Vision: Object Recognition and Visual Cognition*. M.I.T. Press, Cambridge, MA

J. Wagemans

Vision, Low-level Theory of

1. Introduction

The primate visual system comprises over 30 cortical areas in addition to the retina and other subcortical structures. To model the lower levels of form vision or motion perception, it therefore becomes necessary to develop neural network theories of retinal, striate, and extrastriate cortical function that are consistent with known anatomy and physiology and that are capable of predicting relevant psychophysical data. The mathematical techniques necessary for this enterprise include convolution and 'nonlinear dynamical systems' theory (see also *Self-organizing Dynamical Systems*). This article will review current theories of retinal and striate cortical function followed by a discussion of neural models related to areas 'V2' and 'V4' in the form vision hierarchy.

2. Theories of Retinal Function

Retinal neurons transform the optical image in order to extract biologically relevant visual information concerning light intensity changes in space (i.e. contours and edges), changes in time, and chromaticity. Furthermore, the retina must shift its operating point so as to signal useful information over more than a brightness range, a process termed light adaptation. These transformations are carried out by networks comprising at least five different types of neurons: photoreceptors, bipolar cells, horizontal cells, amacrine cells, and ganglion cells (see Wässle and Boycott 1991). As the extraction of color information is covered elsewhere (see *Color Vision Theory*), let us focus here on light adaptation and the extraction of contours.

Light intensities in natural scenes cover more than a 10^6 range, yet the ganglion cells providing the retinal output to the brain can only respond over about a 300:1 range. Thus, the retina must compute the mean level of illumination at any given time and adjust the circuitry to respond to about a 300:1 range around that operating point. Key experiments on retinal light adaptation have demonstrated that there are two major processes involved: one subtractive and one divisive (see Hood 1998). The direct path through the retina involves photoreceptors activating bipolar cells, which in turn activate ganglion cells. In addition the retina incorporates two major feedback loops: (a) from photoreceptors to horizontal cells then back to photoreceptors, and (b) from bipolars to amacrine cells back to bipolars. A recent retinal model has tentatively identified each of these loops with one of the light adaptation processes (Wilson 1997). In particular, loop (a) was conjectured to provide subtractive feedback, while loop (b) was suggested to be the site of divisive feedback. In a subtractive loop, the feedback signal is subtracted from the loop input, while in a divisive loop, the feedback signal divides the loop input signal thus reducing its gain. (More likely, each feedback loop incorporates both subtraction and division, suggesting a somewhat more complex formulation.) It is significant that both horizontal cells and the amacrine cells, the key units in these two feedback loops, have a relatively broad horizontal spread across the retinal surface and thus provide

feedback that is averaged over an extended area. Subtractive feedback is well understood in the context of linear systems theory, but divisive feedback moves into the realm of 'nonlinear dynamics'. Designating the bipolar cell response as B, the amacrine feedback response as A, and the light-derived signal received from the photoreceptors as L, the simplest divisive feedback equations are of the form (Wilson 1997):

$$\frac{\mathrm{d}B}{\mathrm{d}t} = -B + \frac{L}{1+A}$$
$$\frac{\mathrm{d}A}{\mathrm{d}t} = -A + B^N \qquad (1)$$

These equations may be analyzed using standard techniques (Wilson 1999) of nonlinear dynamical systems theory. In particular, one can solve for the steady state or equilibrium value at which $\mathrm{d}B/\mathrm{d}t = 0$ and $\mathrm{d}A/\mathrm{d}t = 0$. The results for very small and very large values of L are approximately:

$$B \approx L \quad \text{for} \quad L \leqslant 1$$
$$B \approx L^{1/(N+1)} \quad \text{for} \quad L \gg 1 \qquad (2)$$

Thus, at very low light intensity levels the bipolar response B is approximately equal to its input, so there is no attenuation of weak signals. At high light intensity levels, however, B is compressed to approximately the $(N+1)$ root of L. For $N = 2$, a physiologically plausible figure, the divisive feedback loop in Eqn. (1) will compress an $L = 10^6$ input into a $B = 100$ output, easily compatible with the response range of the ganglion cells. A further property of Eqn. (1) is that the temporal response becomes more transient as L increases (Wilson 1999), and this change is well documented in the literature on flicker sensitivity and light adaptation (Wilson 1997).

As is well known, retinal receptive fields have an excitatory center and inhibitory surround (or vice versa) spatial profile (see *Visual Perception, Neural Basis of*). Direct bipolar input to the ganglion cells produces the receptive field center, while the lateral spread of both horizontal and amacrine cells creates the antagonistic surround. Thus, the neurons engaged in light adaptation via feedback also produce the spatial surround of ganglion cell receptive fields. As is evident from the discussion above, divisive feedback has almost no effect at low light levels (Eqn. (2)), so one would expect receptive field surrounds to be very weak at low luminances but much stronger at higher luminances, and this has been demonstrated experimentally. For some simulation purposes the spatial characteristics of retinal receptive fields may be approximated as a circularly symmetric difference of Gaussians:

$$G(R) = A\exp(-R^2/\sigma_c^2) - B(L)\exp(-R^2/\sigma_s^2) \qquad (3)$$

where R is radius, and σ_c and σ_s are the space constants of the center and surround, respectively. The inhibitory gain $B(L)$ would vary from $B(0) \approx 0$ to $B(L) \approx A\sigma_c/\sigma_s$ for $L \gg 0$, which makes the integral of the receptive field $\int G\,\mathrm{d}R = 0$. For further discussion of retinal models see Wilson (1997) and Hood (1998).

3. *Theories of Striate Cortical Function*

Hubel and Wiesel (1968) categorized striate (V1) cortical cells as simple or complex (see also *Visual System in the Brain* and *Visual Perception, Neural Basis of*). Both of these cell types show a preference for a particular orientation and also for a range of spatial frequencies. The now classical model for a simple cell receptive field (RF) is just an oriented linear filter with parallel excitatory and inhibitory zones defined by horizontal (x) and vertical (y) coordinates:

$$RF(x, y) = A\left\{\exp(-x^2/\sigma_c^2) - \frac{\sigma_c}{\sigma_s}\exp(-x^2/\sigma_s^2)\right\} \times \exp(-y^2/\sigma_y^2) \qquad (4)$$

Equation (4) describes a vertically oriented receptive field, but other orientations (varying in about 15° increments between 0° and 180°) can be produced by coordinate rotation. Based on physiological data, $\sigma_y \approx 3.2\sigma_c$ produces appropriate orientation bandwidths. The activation $R(x, y)$ of a cortical array of simple cell receptive fields is produced by convolution:

$$R(x, y) = \iint_{-\infty}^{\infty} RF(x-x', y-y')\, I(x', y')\, \mathrm{d}x'\mathrm{d}y' \qquad (5)$$

where $I(x', y')$ represents the distribution of light in the image. Simulated neural responses are then produced by passing the results of Eqn. (5) through a nonlinear function exhibiting both a threshold and saturation such as the Naka–Rushton function

$$S(R) = \left\{\begin{array}{ll} \dfrac{MR^N}{\beta^N + R^N} & \text{for} \quad R \geqslant 0 \\ 0 & \text{for} \quad R < 0 \end{array}\right\} \qquad (6)$$

where M is the maximum response level and N $\approx$ 2–3 for cortical neurons. As $S(R) = M/2$ when $R = \beta$, β is referred to as the semisaturation constant. Dynamics may now be added to this formulation as follows (Wilson 1999):

$$\tau\frac{\mathrm{d}E}{\mathrm{d}t} = -E + S(R) \qquad (7)$$

where E is the excitatory firing rate and τ is the time constant for the rate of exponential approach to the asymptotic firing rate.

The simple cell model in Eqns. (4)–(7) can be used to generate complex networks of interconnected model neurons for simulations of cortical function (Wilson 1999). However, contemporary research indicates that orientation selectivity is more complex than this. In particular, the linear filtering in Eqn. (4) suggests that orientation tuning originates entirely from a linear superposition of nonoriented lateral geniculate (LGN) filters similar to that in Eqn. (3). While this was Hubel and Wiesel's (1968) original suggestion for orientation selectivity, the preponderance of evidence now supports additional factors in orientation tuning. First, convergence of as few as two nonoriented LGN filters conveys a modest orientation bias on cortical neurons. This very broad orientation bias is then sharpened considerably by two operations occurring in neural networks at the cortical level: contrast gain control and collinear facilitation (Sompolinski and Shapley 1997).

Contrast gain controls were first modeled by Heeger (1991) as divisive feedback operations. He suggested that the summed activity of many different oriented units provided a divisive feedback signal to each of the units in turn. The divisive feedback loop in Eqn. (1) can be altered to function as a cortical gain control as follows:

$$\tau \frac{\mathrm{d}E_\theta}{\mathrm{d}t} = -E_\theta + S\left(\frac{R_\theta}{1+A}\right) = \frac{MR_\theta^N}{\beta_\theta^N(1+A)^N + R_\theta^N}$$
$$\frac{\mathrm{d}A}{\mathrm{d}t} = -A + \sum_\theta E_\theta \qquad (8)$$

where θ designates the preferred orientation of each unit. Substitution of the Naka–Rushton function S from Eqn. (6) and rearrangement shows that this cortical divisive feedback circuit has the effect of scaling β by $(1+A)$. This network has been implemented and shown to sharpen cortical orientation tuning (Wilson 1993). Due to the dynamics of the network, this orientation sharpening evolves over time as the divisive feedback takes effect, and this has recently been shown to occur in most cortical neurons (Sompolinsky and Shapley 1997).

The final contribution to cortical orientation selectivity is provided by collinear facilitation (Sompolinsky and Shapley 1997). Recent evidence indicates that neurons with the same preferred orientation are interconnected if their receptive fields are situated on a common line defined by their preferred orientation. Such reciprocally linked receptive fields are typically separated by 1–2 receptive field lengths so that their mutual facilitation will only be triggered by stimulus edges of sufficient length. The excitatory connections are probably mediated by NMDA receptors so that both cells in a collinear pair must be independently activated by the stimulus for facilitation to occur. Obviously, collinear facilitation will sharpen orientation tuning when measured with extended lines or gratings. The function of collinear facilitation is to enhance the salience of extended contours, thus making them prominent candidates for further processing.

4. *Global Processes in Extrastriate Form Vision*

The complex cells of Hubel and Wiesel can be readily modeled by spatial pooling of simple cells with the same preferred orientation. More interesting are end-stopped complex cells, which are sensitive to both contour length and curvature. As illustrated in Fig. 1(a), an end-stopped complex cell can be modeled by full-wave rectification of simple cell responses followed by second-stage pooling using a 2–3 times larger version of the oriented filter defined by Eqn. (4). Full-wave rectification is mathematically equivalent to summing responses from both on-center and off-center simple cells that have the same spatial location and preferred orientation. The first- and second-stage filters in this filter–rectify–filter sequence usually have different preferred orientations, with the orthogonal case being illustrated in Fig. 1(a). The optimal line stimulus for this end-stopped unit will obviously be a horizontal line with length equal to the width of the second filter's excitatory zone as shown superimposed on the diagram. In addition to lines of fixed length, end-stopped cells will respond well to curves that are tangent to the preferred orientation of the first-stage simple cells, and they respond well to the orientation of texture boundaries (Wilson 1993). Although Hubel and Wiesel (1968) reported the presence of end-stopped neurons in V1, subsequent work by von der Heydt and Peterhans (1989) indicates that they predominate in the second visual area, V2.

Following V1 and V2 the cortical form vision pathway projects to V4, and finally to inferior temporal cortex (IT). Mean receptive field dimensions increase by a factor of approximately 2.5–3 from area to area within this pathway, reaching diameters of 10°–30° in IT. It was noted above that the receptive fields of end-stopped neurons must be approximately 2–3 times larger in each dimension than the simple cell receptive fields that provide their inputs in order to function optimally. Thus, the increasing receptive field size throughout the form vision pathway indicates that pooling operations are combining information over increasingly large areas. The nature of pooling in V4 was first investigated by Gallant et al. (1993). They studied macaque V4 responses to concentric and radial gratings as well as to conventional sine wave gratings. One subclass of neurons responded best to concentric gratings, while a second subclass responded best to radial gratings. Although many neurons responded to conventional sine wave gratings, few were optimized for these patterns.

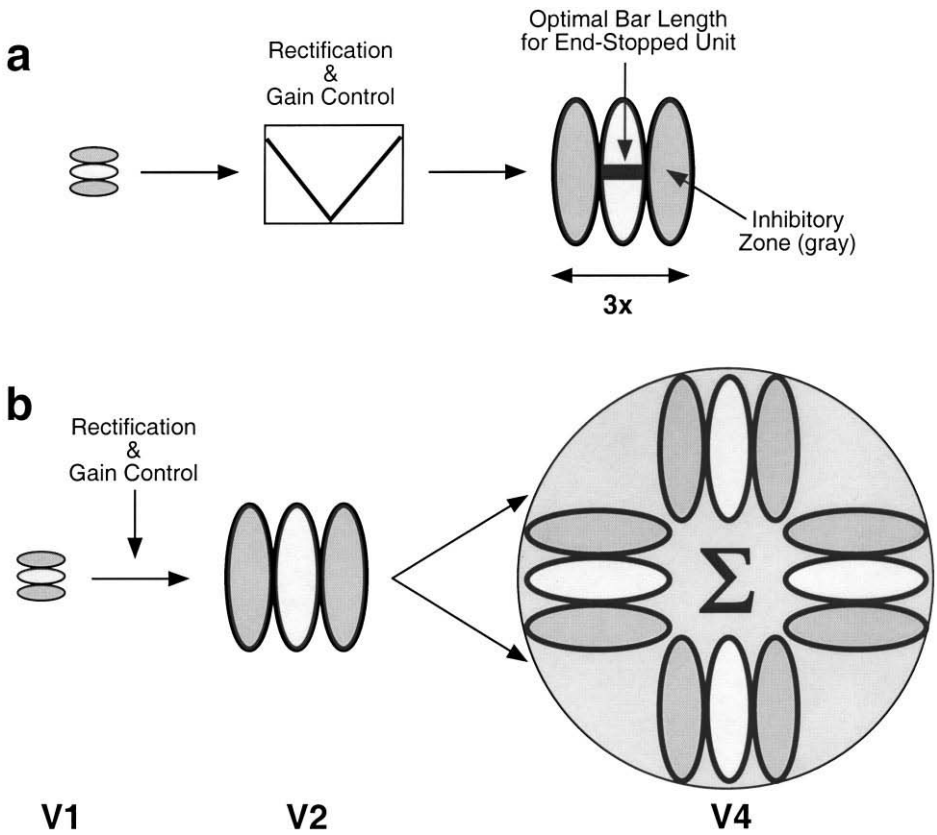

Figure 1
Models of (a) an end-stopped complex cell and (b) a V4 concentric unit. The end-stopped cell (a) is produced by oriented simple cell filtering followed by rectification and subsequent filtering at a different orientation. The optimal bar length and orientation for stimulation of this unit is shown. The V4 concentric model in (b) builds upon the hierarchical theme in (a) by adding a final stage of configural pooling in which responses of a concentric array of end-stopped units are summed. The three stages of the V4 model are hypothesized to occur in V1, V2, and V4

Humans are extremely good at detecting concentric structure in glass patterns, and this results from concentric summation of orientation information over considerable regions (Wilson et al. 1997). A glass pattern is produced by randomly positioning pairs of signal dots such that each pair is oriented tangentially to the desired concentric structure. This produces a random dot pattern appearing to contain concentric swirls of dots. Furthermore, fMRI studies have shown that human V4 is more sensitive to concentric or radial gratings than to conventional gratings (Wilkinson et al. 2000). A quantitative model for V4 concentric receptive fields is depicted in Fig. 1(b) (Wilson et al. 1997). The initial two stages of the model are oriented simple cells followed by rectification and orthogonal second stage filtering. As mentioned above, these model end-stopped complex cells map onto V1 and V2 physiology, and they respond to curved contours that are locally tangential to the orientation of the simple cell filter. The final stage of the V4 model consists of linear summation ($\sum$) of concentrically arranged curvature information from V2. This represents an example of configural pooling, as each curvature response is pooled only if it arises in a spatial subregion of the receptive field consistent with the concentric global pattern. The final V4 receptive field size is approximately three times the diameter of its V2 subunits, thus conforming to the observed receptive field size increase in the form vision pathway. Similar V4 configural pooling models may be constructed to extract radial and other structures from the stimulus, and one challenge to future research will be an enumeration of configural pooling mechanisms.

What might be the function of V4 concentric units? Several observations suggest that they may be involved in the perception of faces and other ellipsoidal shapes. First, computer simulations have shown that model V4 units can detect faces in complex scenes (Wilson et

al. 2000). Second, fMRI studies have shown that the face-selective area in the fusiform gyrus responds not only to faces but also to circular or ellipsoidal contours (Wilkinson et al. 2000), thus supporting the hypothesis that analysis of head shape is a major ingredient in face perception.

5. *Perceptual Oscillations*

There are many instances of instabilities resulting in perceptual oscillations, two of the most common being the Necker cube and binocular rivalry. A third example is the Marroquin illusion in which a static, periodic dot pattern triggers the percept of flashing illusory circles (see Wilson et al. 2000). A Marroquin pattern is constructed by superimposing three square grids of dots, each rotated $\pm 60°$ with respect to one another. It is natural to conjecture that competitive interactions among V4 concentric units might be the cause of the illusory circles triggered by this stimulus. Indeed, there is physiological evidence that competition among V4 neurons forms a basis for spatial and object based selective attention (Reynolds et al. 1999).

A neural network model incorporating spatially regional competitive interactions among V4 concentric units can explain the illusory flashing Marroquin circles (Wilson et al. 2000). In this simulation V4 neural responses were represented by a variant of Eqn. (7):

$$\tau_E \frac{\mathrm{d}E_n}{\mathrm{d}t} = -E_n + \frac{100\, P_+^2}{(\beta + H_n)^2 + P_+^2}$$

where

$$P = S_{Marroquin} - k \sum_{n \neq k} I_k \exp\left(-R_{nk}^5/\sigma^5\right) \qquad (9)$$

$$\tau_I \frac{\mathrm{d}I_n}{\mathrm{d}t} = -I_n + E_n$$

$$\tau_H \frac{\mathrm{d}H_n}{\mathrm{d}t} = -H_n + g E_n$$

In these equations the input P to each neuron is the response of the V4 model in Fig. 1(b) to the Marroquin pattern, $S_{\mathrm{Marroquin}}$ minus a spatially weighted sum of inhibitory activity I_k generated by neighboring V4 neurons, the inhibitory neurons being driven by the E cells as described by the second equation. Parameters k and σ are respectively the gain and space constant of the inhibitory region. Mathematical analysis of equations such as these shows that one further ingredient is necessary for the network to produce limit cycle oscillations: the firing rate of active excitatory neurons must decline over time (Wilson 1999). Excitatory cortical neurons are known to contain slow hyperpolarizing currents that cause the firing rate to drop about threefold over the course of several hundred milliseconds, and this is incorporated into the mathematical model via the H or hyperpolarizing variable (Wilson 1999). The effect of this variable, governed by g, is to increase the semisaturation constant of any active E neuron, thereby reducing the firing rate. Simulations of this network provide a good quantitative explanation of the illusory flashing circles generated by Marroquin patterns (Wilson et al. 2000). Extensions of this work suggest that models analogous to Eqn. (9) can also explain oscillations and traveling wave phenomena in binocular rivalry.

6. *Conclusions and Future Directions*

The low-level vision theories developed here are based on nonlinear differential equations describing firing rates (rather than individual spikes) of retinal and cortical neurons. Large-scale neural networks incorporating these principles can now be used to simulate visual processing from the retina up to V4. Although this approach certainly represents but a first-order approximation to the underlying neural activity, it nevertheless provides a solid foundation for further progress.

Many issues, of course, remain to be addressed. Key among these is the ubiquitous cortical feedback: IT back to V4, V4 back to V2, etc. One conjecture is that these connections mediate a rapid dynamical optimization of the processing circuitry based upon initial feed-forward estimates of the stimulus being analyzed: a sophisticated example of a self-organizing system (see also *Self-organizing Dynamical Systems*). A second likelihood is that this feedback mediates both top-down influences on selective attention and figure-ground segmentation, which appears to involve extrastriate feedback to V1 (Lamme and Roelfsema 2000). The future promises a cooperative interaction between theory and experiment in elucidating these and many other exciting issues.

See also: Color Vision Theory; Perception and Cognition, Single-/Multi-neuronal Recording Studies of; Vision, High-level Theory of; Visual Perception, Neural Basis of; Visual System in the Brain

Bibliography

Gallant J L, Braun J, VanEssen D C 1993 Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science* **259**: 100–3

Heeger D J 1991 Nonlinear model of neural responses in cat visual cortex. In: Landy M, Movshon J A (eds.) *Computational Models of Visual Processing*. MIT Press, Cambridge, MA, pp. 119–33

Hood D C 1998 Lower level visual processing and models of light adaptation. *Annual Reviews in Psychology* **49**: 503–35

Hubel D H, Wiesel T N 1968 Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology* **195**: 215–43
Lamme V A F, Roelfsema P R 2000 The distinct modes of vision offered by feedforward and recurrent processing. *TINS* **23**: 571–7
Reynolds J H, Chelazzi L, Desimone R 1999 Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience* **19**: 1736–53
Sompolinsky H, Shapley R 1997 New perspectives on the mechanisms for orientation selectivity. *Current Biology* **7**: 514–22
von der Heydt R, Peterhans E 1989 Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *Journal of Neuroscience* **9**: 1731–48
Wässle H, Boycott B B 1991 Functional architecture of the mammalian retina. *Physiology Reviews* **71**: 447–80
Wilkinson F, James T W, Wilson H R, Gati J S, Menon R S, Goodale M A 2000 An fMRI study of the selective activation of human extrastriate form vision areas by radial and concentric gratings. *Current Biology* **10**: 1455–8
Wilson H R 1993 Nonlinear processes in visual pattern discrimination. *Proceedings of the National Academy of Sciences USA* **90**: 9785–90
Wilson H R 1997 A neural model of foveal light adaptation and after image formation. *Vision Neuroscience* **14**: 403–23
Wilson H R 1999 *Spikes, Decisions, and Actions: Dynamical Foundations of Neuroscience*. Oxford University Press, Oxford, UK
Wilson H R, Krupa B, Wilkinson F 2000 Dynamics of perceptual oscillations in form vision. *Nature Neuroscience* **3**: 170–6
Wilson H R, Wilkinson F, Asaad W 1997 Concentric orientation summation in human form vision. *Vision Research* **37**: 2325–30

H. R. Wilson

Vision, Psychology of

Vision research began more than 100 years before experimental psychology was founded. Over the past 50 years there have been four major developments that have greatly increased our understanding of how vision works. The first was the appreciation of the role that internal noise and uncertainty plays in limiting visual performance. The second was the resolution of a century-long conflict between two theories of color vision. The third was the discovery that vision, like audition, performs a frequency analysis of the visual stimulus. And the fourth was the realization that there are many parallel streams of visual processing being carried out simultaneously and that different characteristics of the stimulus are analyzed by different modular mechanisms.

The purpose of vision is to provide us with information about objects in the world around us. This information is obtained from photons radiated from objects. When these photons enter the eye and are absorbed by photopigments in the photoreceptors, the resulting photoreceptor signals influence the activity of many processes, ultimately giving rise to the experience of seeing. Providing an incontrovertible definition of 'seeing' is not so easy. *Seeing* usually refers to the generation of conscious experiences through stimulation of the visual system by light. But visual experiences can also be generated as a result of direct electrical or chemical stimulation in the eye and at many locations in the brain. Conscious experiences are taken to mean experiences that can be verbally reported, although some experts find this definition too restrictive, and some find it too broad. Disagreement occurs because there is no clear definition of 'conscious' and 'consciousness.'

Vision involves four realms of human knowledge: the realm of physics (energy and matter); the realm of physiology, and especially neurophysiology (the behavior of the nervous system); the realm of overt behavior (what people say and do); and the realm of the mind and subjective experience. A complete understanding of vision necessarily involves the understanding of these realms and their interactions. In this essay, however, we will focus on the psychological mechanisms that make up the mind. The characteristics of these mechanisms cannot be determined by direct observation but must be inferred from the outcome of appropriately designed experiments.

1. The Role of Uncertainty

1.1 Physical Nature of Light

Light is electromagnetic radiation. It exists in discrete packets of energy called quanta, or photons. Quanta cannot be subdivided, and consequently light always consists of an integer number of quanta. Quanta differ from each other in the amount of energy they contain. Quanta may also be quantified by two other properties that are directly related to the energy: frequency and wavelength. The energy, frequency, and wavelength of a quantum are interrelated by two fundamental equations: $E = h\nu$ and $c = \lambda\nu$, where E is the energy in Joules, h is Planck's constant, ν is frequency in cycles per second (Hertz), λ is wavelength in meters, and c is the speed of light in a vacuum. A single quantum therefore has only one unique characteristic, which may be described as energy, frequency, or wavelength.

Light sources are characterized by the distribution of quanta that they produce. Hot sources produce broad spectral distributions, while some sources (lasers, for example) produce only a single wavelength. When a quantum from a source strikes matter (gas, liquid, or solid), one of three things occurs to the quantum: it is reflected from the surface; it passes through the medium; or it is absorbed and its energy is transformed into heat. The probability that a quantum is reflected, transmitted, or absorbed depends on the energy of the quantum and the physical properties of the material. The distribution of quantal energy from a source is therefore modified by the material with

which it interacts. The human eye takes in quanta from sources, and quanta reflected from or transmitted through objects, and forms an optical image on the retina at the back of the eye.

1.2 Beginnings of Vision

Vision begins when photoreceptors in the retina absorb quanta of light and generate electrochemical signals. Collectively, the human photoreceptors can absorb quanta in the narrow range of wavelengths from approximately 400 to 700 nm ($1\,\text{nm} = 10^{-9}$ meter). The probability that a photoreceptor will absorb a quantum varies as a function of quantal wavelength, with a maximum probability occurring at a specific wavelength, λ_{max}. There are four types of receptors: one rod and three cones. The receptor types differ in their absolute sensitivity (rods are more sensitive than cones), and in their relative sensitivity, to different wavelengths. Rods have a λ_{max} at 510 nm, while the three cone types have a λ_{max} at approximately 420 nm (S-cones), 530 nm (M-cones), and 560 nm (L-cones). The normalized sensitivity of each receptor is plotted as a function of wavelength in the upper panel of Fig. 1.

Photoreceptors signal the rate at which quanta are being absorbed. They do not and cannot signal the wavelength of the quanta that are absorbed. The voltage change that occurs in a receptor when a quantum is absorbed is independent of the wavelength of the quantum. This important principle is known as the 'Principle of Univariance.' The visual system operates over a physical intensity range of 10^{12}. The upper limit of 10^6 is equivalent to a white cumulus cloud illuminated by the noonday sun. The range from 10^6 to 10^{-2} is handled by the cone receptors and is called photopic vision. The lower part of the whole range, from 10^{-2} to 10^{-6} is handled by the rod receptors and is called scotopic vision. The intensity of 10^{-2}, the boundary between photopic and scotopic vision, corresponds to the illumination provided by the full moon. The lowest level at which rods function is about 100 quanta of 510 nm light striking the front of the eye. The minimum amount of energy for detecting each wavelength with the scotopic and photopic systems is plotted in the lower panel of Fig. 1. The difference between the energy needed for photopic detection and that needed for scotopic detection is called the photochromatic interval. At short wavelengths, scotopic vision is much more sensitive than photopic vision, but at long wavelengths the two systems are equally sensitive.

1.3 Internal Representation

The component of the mind called 'the sensory processes' builds internal representations of the external world by transforming physical stimulation into internal activity. This activity plays an important role in shaping the final representation, but memory, attention, and motivation also influence the final outcome. The internal representation is not simply the physical stimulus in a new form. To equate the stimulus with the sensation that it produces is to make the so-called *stimulus error*. Light, as a physical phenomenon, has no color: it has wavelength distribution and intensity. Color is produced by chromatic mechanisms in the sensory processes, and there are many wavelength combinations that give the same color (see Sect. 1.4).

1.4 Uncertainty

Internal representations are characterized by uncertainty. This uncertainty derives from many sources but originates in the quantum nature of matter and energy. If short flashes of weak light are used in a vision experiment, it is not possible to know exactly how many photons of light are delivered to the eye on each trial. This uncertainty stems from the random nature of photon generation. The probability of obtaining a specific number of quanta on a flash can be computed using the Poisson probability distribution if we know the average number of quanta delivered to the eye on each flash:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{1}$$

where x is an integer number ranging from 0 to infinity and λ is the mean number of photons per flash. On some presentations of these stimuli there will be many photons, and on some there will be few, and on some there will actually be zero photons. This uncertainty in the physical stimulus combines with the inherent uncertainty of the many neural synapses and processes involved in constructing internal representations. The result is that the internal representation is not a fixed quantity but fluctuates with time. It is therefore necessary to represent internal representations by probability distributions rather than fixed quantities. The most common one is the Gaussian probability distribution, which is specified by a mean, μ, and a standard deviation, σ:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5((x-\mu)/\sigma)^2} \tag{2}$$

The use of Gaussian probability distributions to describe internal representations has a long history that began with Gustav Theodor Fechner's landmark book *Die Elemente der Psychophysik* in 1860. It was

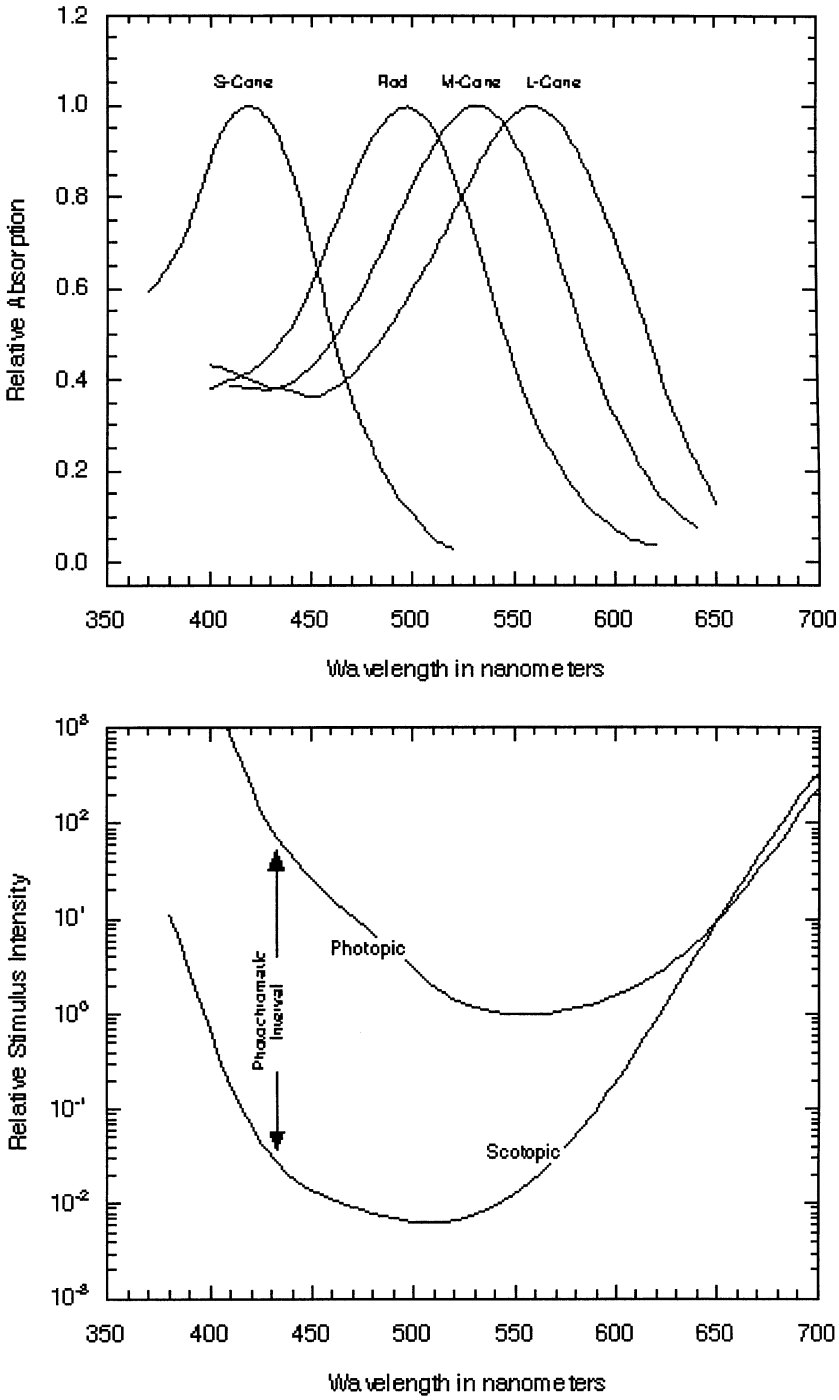

Figure 1
The normalized absorption functions of the rod and three cone photopigments (upper panel), and the minimum stimulus intensity for detection by the scotopic (rod) visual system and for the photopic (cones) visual system (lower panel)

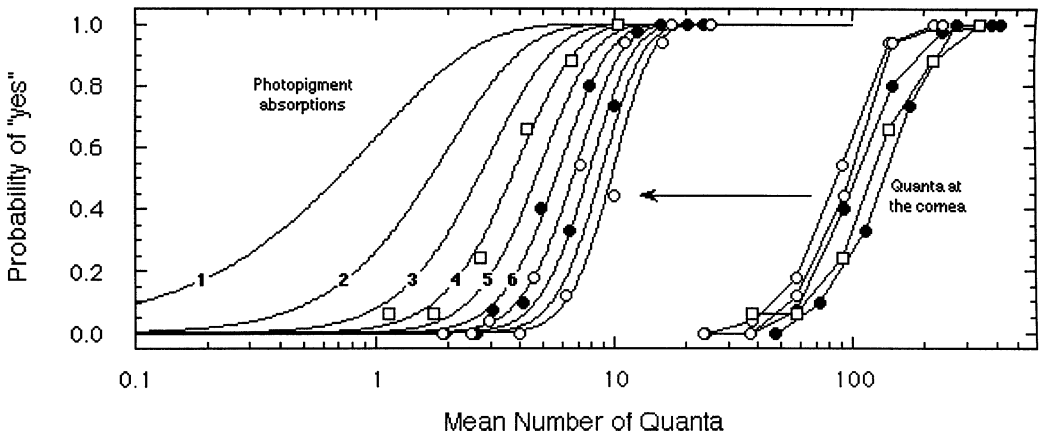

Figure 2
Five psychometric functions from the three authors of the classic 1942 experiment by Hecht, Shlaer, and Pirenne. The smooth functions on the left are the psychometric functions predicted by assuming that internal representations of the stimuli are Poisson probability distributions and that the observer needs a fixed number of quanta or more in order to say 'yes'

assumed by Galton when he introduced percentile analysis in 1885 and became more widespread in the early twentieth century through the writings of Thurstone and Urban. Signal detection theory, developed in the 1950s, has introduced this use of probability distributions to a much wider range of applications.

1.5 Psychometric Function

The most immediate consequence of noisy internal representations is that there is no single stimulus intensity above which one will always see a stimulus and below which one will not see it. Rather, as stimulus intensity is increased, the probability that an observer will say that the stimulus is visible increases. This relationship between the probability of a response and the stimulus intensity is called a *psychometric function*. Psychometric functions are S-shaped (ogive) when stimulus intensity is plotted on a logarithmic axis, as is illustrated at the right side of Fig. 2.

The psychometric function is a powerful tool for inferring the properties of the sensory process and the internal representation. In 1942, Hecht, Shlaer, and Pirenne sought to answer the question 'how many quanta have to be absorbed by rod receptors in order to say "yes, I saw the stimulus"?' Since the processes of generating quanta and of absorbing quanta are described by Poisson probability distributions, and since the theoretical Poisson psychometric functions have shapes that depend on the number of quanta required to 'see' (smooth curves on the left side of Fig. 2), and since only a fraction of the quanta striking the cornea of the eye get absorbed by photoreceptors (quantum efficiency), both the number of quanta required for seeing and the quantum efficiency may be estimated by finding the theoretical psychometric function that best fits the observed psychometric functions. In Fig. 2, the data of Hecht, Shlaer, and Pirenne have been shifted to the left by an amount corresponding to the quantum efficiency of each observer so that the data coincide as well as possible with a Poisson psychometric function. Each set of data fits extremely well with one, and only one, of the theoretical Poisson functions. It is concluded that from four to 10 quanta are needed, and that the quantum efficiency is about 5 percent.

How one interprets the probability distributions depends on the model of the detection process one adopts. All models of detection have a sensory process and a decision process. One widely held, but now thoroughly discredited, model of the sensory process is the high threshold model, in which the sensory process has a threshold that must be exceeded by the stimulus before the sensory process generates an internal representation of the stimulus. With such a model, the psychometric function represents the integral of the underlying probability distribution of the fluctuating threshold values.

The decision process of the high threshold model says 'yes' on test trials where an internal representation was generated and says 'yes' when no representation was generated by guessing some of the time. In order to estimate the guessing rate, experimenters introduced two types of detection trials: those in which a stimulus was presented (signal trials) and those without any stimulus (blank trials). Performance on such a detection task requires two measures to characterize it fully: the hit rate (probability of saying 'yes' when the signal is present); and the false alarm rate (probability of saying 'yes' when the signal was absent). The relationship between the hit rate and the false alarm

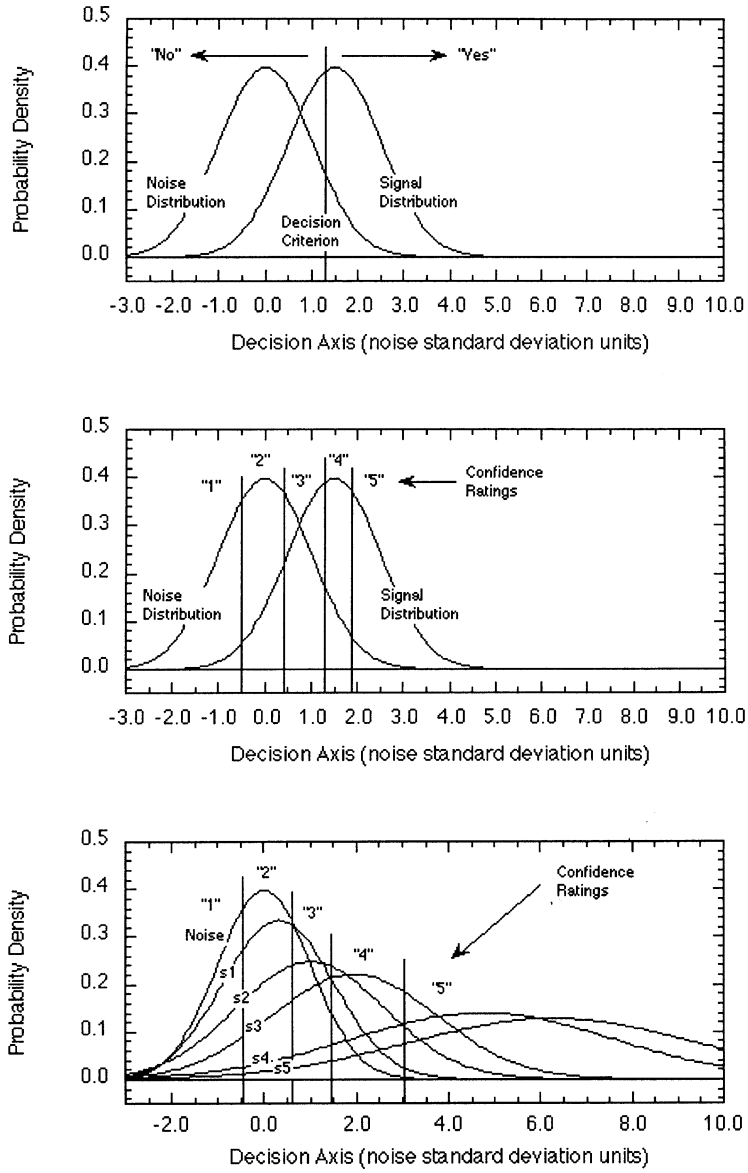

Figure 3
Three variations of the dual-Gaussian, variable-criterion signal detection model. The Gaussian distributions are the internal representations generated by the sensory process. The vertical lines are the decision criteria used by the decision process to generate responses. Upper panel: single criterion generates two detection responses. The insert shows the receiver operating characteristic (ROC). Middle panel: four decision criteria generate five responses expressing confidence that the signal was present. Lower panel: four decision criteria and five different stimuli

rate as the decision process changes its decision strategy is called the receiver operating characteristic (ROC). The high threshold model predicts that the ROC will be a straight line, illustrated on the ROC inserted in the upper panel of Fig. 3. Hit rate–false alarm rate pairs from actual experiments in which decision strategy is manipulated form a bow-shaped ROC, shown in the upper panel of Fig. 3 by the nine filled circles. The fact that the high threshold model fails to predict the correct ROC is one reason it has been rejected as a viable model of the sensory process.

The widely accepted replacement for the high threshold model is the signal detection model. The sensory process of this model has no sensory threshold

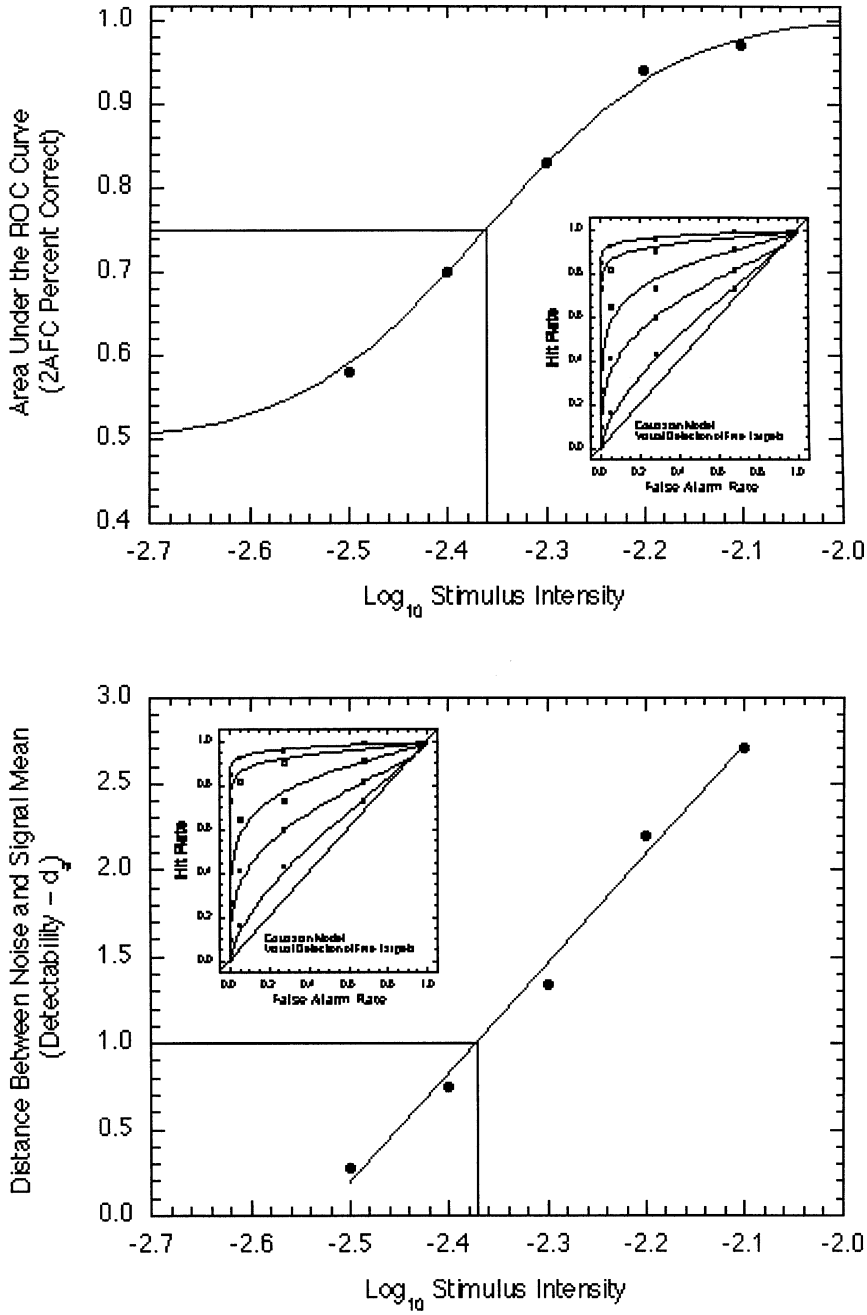

Figure 4
Two forms of the psychometric function of the sensory process derived from the signal detection model. The upper panel shows that stimulus detectability increases as a linear function of log stimulus intensity. The lower panel shows that the classic S-shaped psychometric function is obtained when the area under the ROC is used to characterize the sensory process. The inserts show the ROC for each stimulus intensity

and is always generating an output even with no stimulus: any nonzero stimulus adds to this output. The decision process uses one or more decision criteria to decide what response to generate. The ROC predicted by the dual-Gaussian, variable-criterion signal detection model is plotted as the smooth curved line in the insert in the upper panel of Fig. 3 and provides an excellent fit to the data.

Three variations of the Gaussian model are illustrated in Fig. 3. Two measures of sensory process sensitivity are widely used: the distance between the mean of the no-signal distribution and the mean of the signal distribution, d_a, and the area under the ROC generated by the two distributions, A_z. The model in the bottom panel of Fig. 3 describes detection data from an experiment in which on each trial an observer was presented with one of five visual targets, each of different intensity, or no stimulus at all. The observer rated their confidence that a visual target had been presented. The ROCs for detecting these stimuli are shown in the inserts on the graphs of Fig. 4. As stimulus intensity increases, the ROC for that target becomes increasingly bowed. The psychometric function of the sensory process is formed using either d_a or A_z (Fig. 4), since in the model they represent the sensitivity of the sensory process to the stimuli. The variability contributed by the decision process to actual performance is effectively removed.

In the past 20 years, the application of the signal detection model of noisy internal representations has been expanded well outside its humble beginnings as a model of the sensory and decision processes of vision and audition. Fields in which this model has been shown to provide extremely good descriptions of observed data include recognition memory, medical diagnosis, weather forecasting, lie detection, clinical evaluation, drug detection, releasing prisoners on parole, and computer-guided decision making.

2. *Color Vision*

The field of color vision has been marked by controversy and conflict that began with Isaac Newton's discoveries about the nature of light and of color mixing published in his *Opticks* (1704). By the end of the nineteenth century, two theoretical camps had formed: one based on the ideas of Thomas Young and Hermann von Helmholtz; and the other grounded on the ideas of Ewald Hering and Johann Wolfgang von Goethe. Helmholtz argued that colors are created by the pattern of activity of three types of photoreceptors, each generating a specific color: red, green, and violet (the trichromatic theory). The Hering camp held that there were three processes that generated opponent pairs of colors: red–green, yellow–blue, and black–white (opponent process theory). These two seemingly incompatible theories coexisted for almost 100 years.

The simple resolution came in the late 1960s with the demonstration that each theory described only part of the color vision process. A complete theory of color vision having a Helmholtz component and a Hering component was published by Hans Vos and Pieter Walraven in 1972. To understand the resolution of this conflict, one must carefully distinguish between color matching and color appearance.

Two colors form a metameric match when, in spite of different wavelength composition, they appear identical and indistinguishable in a side-by-side comparison. For people with normal color vision, all colors can be matched by mixing three suitably chosen primary lights. There are many possible primaries: the only constraint is that each of the three primaries cannot be matched by a mixture of the other two. This restriction means that one of the primaries will be reddish, one will be greenish, and one will be bluish. All colors can be specified by the amounts of the three primaries that are needed to match them. The specifications based on one set of three primaries can be changed into any other set of three primaries by simple linear transformation.

When lights are mixed, the amount of the three primaries required to match the mixture is the sum of the amounts of the primaries required to match the individual components of the mixture. This linearity of color matching is remarkable and occurs because the receptors act like quantum counters. Two lights will match if the S-, M-, and L-cones absorb exactly the same number of quanta per second when exposed to the one light as they do when exposed to the others. The three receptors form the Helmholtz zone and govern color matching.

In the above discussion of color matching, no mention was made about the color *appearance* of the match. This omission is deliberate, because color appearance is not caused directly by cone activity. Visual sensations vary along three fundamental dimensions: hue, saturation, and brightness.

Brightness and saturation are so-called *prothetic* dimensions: they can be rank ordered from least to greatest (Stevens 1951). For example, white is brighter than gray, which is brighter than black. Hue is a *metathetic* dimension: different hues cannot be rank ordered from least to greatest; one cannot say that red is greater than green, it is just different. Color appearance is generated by the activity of two chromatic opponent mechanisms and one luminance mechanism. Each of these mechanisms receives input from the three cone types. The red–green chromatic mechanism can generate red or green experiences but not both at the same time. The yellow–blue chromatic mechanism can generate yellow or blue experiences but not both at the same time. In normal color vision, the hue and saturation of a color experience is determined by the joint activity of these two chromatic mechanisms. A color can be reddish-blue but not reddish-green. A color can be yellow-green but not yellow-blue.

The chromatic mechanisms function by comparing the weighted activity of one cone type with the weighted activity of the other two cone types. The cones' connection to the chromatic mechanisms and to luminance are estimated to be (from the work of John Werner and Billy Wooton (1979a, 1979b):

$$(+r-g) = 1.89\ LCone - 2.79\ MCone + 0.45\ SCone$$

$$(+y-b) = (0.94\ LCone + 0.21\ MCone)^{1.65} - 1.65\ SCone$$

$$Luminance = 0.85\ LCone + 0.15\ MCone + 0.015\ SCone$$

There are two important points to be noted: the S-cones make almost no contribution to luminance and to the red–green chromatic mechanism; and the yellow–blue mechanism is nonlinear. The activity of the chromatic mechanisms in one part of the visual field is altered and influenced by the activity of the mechanisms in spatially adjacent areas. A 'neutral gray' test patch will appear greenish when surrounded by a red field.

It is common to read in introductory textbooks that the color opponent processes are located in the ganglion cells of the retina. A careful analysis of the chromatic properties of chromatically sensitive ganglion cells, however, reveals that although many retinal ganglion cells show opponent characteristics by inhibiting their activity to some wavelengths and exciting their activity to other wavelengths, their actual properties do not match those required for the Hering zone chromatic mechanisms. The way the brain actually creates the chromatic mechanisms involves both retinal and cortical mechanisms.

3. Spatial and Spatial Frequency Analysis

Visual stimuli may be described quantitatively in two quite different but interchangeable ways: in terms of spatial features, or spatial frequency content. Simple spatial features are, for example, the diameter or height and width of a target, typically measured in degrees of visual angle on the retina of the eye. Description of more complex stimuli, such as letters of the alphabet or human faces, have feature sets to measure the properties of these stimuli. In the early 1950s, the powerful engineering methods of linear systems analysis were introduced into vision. Linear systems analysis characterizes systems by the degree to which sine wave stimuli of various frequencies are passed through it: the modulation transfer function (MTF). In vision, a physical stimulus does not pass through the observer but is transformed into an internal representation; but by finding the amplitude of a sine wave frequency that will give rise to a constant internal effect (a constant level of detectability, for example) investigators approximated the MTF of the human visual system using the modulation sensitivity function (MSF), or equivalently, the contrast sensitivity function (CSF). A typical photopic contrast sensitivity function is plotted in Fig. 5.

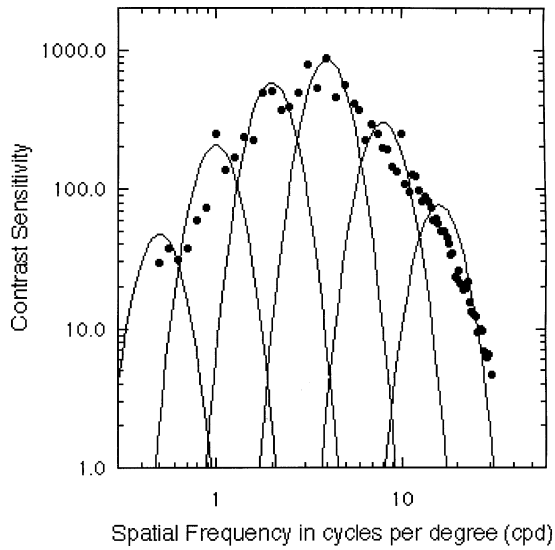

Figure 5
Contrast sensitivity function. Sensitivity is the reciprocal of the contrast required to detect a sine wave pattern at a fixed level of sensory process detection performance, $d_a = 1.0$. The individual solid lines show the sensitivity of underlying Gabor mechanisms tuned to subsets of the whole spatial frequency range to which the human visual system is sensitive

In the late 1960s, Fergus Campbell and colleagues at Cambridge University reported in two landmark papers that visual detection of complex grating patterns was better predicted by their spatial frequency properties than by their spatial properties (e.g., Campbell and Robson 1968). There ensued an intense period of research investigating the spatial frequency properties of stimuli (as computed by means of the Fourier theorem and transform) and their relationship to various aspects of visual perception. There was also intense debate about whether the visual system carried out a spatial analysis or a spatial frequency analysis of the retinal image.

The controversy was resolved in 1980, when John Daugman (1980) and Stephan Marcelja (1980) pointed out that cells in the visual cortex had spatial-response properties that closely resembled so-called Gabor functions. Denis Gabor had proved in 1946 that products of Gaussian envelopes and sine and cosine waves were optimal for minimizing the joint uncertainty of specifying location in space and spatial frequency simultaneously. Numerous experiments that followed these two papers found that the visual

system carries out a simultaneous spatial and spatial-frequency analysis of visual input by means of mechanisms that are like Gabor functions. Thus, the visual system has developed an optimal way of carrying out a spatial analysis and a spatial-frequency analysis simultaneously.

4. Modular Processing

The second discovery of the landmark Campbell papers was that the entire span of spatial frequencies to which the visual system responds (see Fig. 5) is handled by numerous mechanisms, each tuned to a different and smaller subrange of spatial frequencies. The mechanisms that detect high spatial frequencies act independently from the mechanisms that detect lower spatial frequencies. The spatial-frequency tuning properties of a few of these mechanisms are also shown in Fig. 5. Each of these mechanisms is also tuned to a narrow range of contour orientation. To cover the range of spatial frequencies and orientations to which we are sensitive requires approximately 90 different mechanisms at each area of the retinal image, each tuned to a specific range of frequency and orientation.

The idea of modularity of visual processing, with different mechanisms analyzing different properties of the visual input, is at the heart of our current concept of the visual system. In fact, based on both anatomical and psychophysical evidence, we seem to have two visual systems, the M-pathway and the P-pathway (named after the magno- and parvo-cellular layers of the lateral geniculate nucleus). These two systems carry quite different kinds of information from the retina to the brain. The M-pathway, the older of the two, has high-contrast sensitivity but low spatial resolution. It carries no color information but does carry information about motion and depth and responds rapidly to visual stimulation. The P-pathway has high spatial resolution, but responds slowly and has lower spatial resolution. The P-pathway carries color information encoded in an opponent-process manner.

5. Summary

A complete understanding of vision involves an understanding of physics, physiology, behavior, and the mind. At its simplest, the mind contains a sensory process that converts external physical stimuli into an internal representation and a decision process that uses information available in the internal representation to make decisions about what to say and what to do.

Four major advances in the psychology of vision took place in the last half of the twentieth century. The first is the recognition that the internal representation is noisy and unreliable, which has led to the development of powerful models and methods to deal with the resulting uncertainty. The second has been a resolution of the battle between two conflicting theories of color vision by combining both into a single unified theory. The third development has been the discovery that vision, like audition, carries out a frequency analysis on the visual input. Unlike audition, however, it carries out a simultaneous spatial analysis and does it in the theoretically most efficient way. The final advance has been the appreciation that in the early stages of visual processing different types of information are handled by different types of mechanisms, producing a modularity of visual processing.

6. Recommended Reading

There are many oversimplifications in this article. Please consult these works for more detailed information about the material: Sekuler and Blake 2002, Coren et al. 1999, Goldstein 1999, Levine 2000, Schiffman 2000. Texts more specialized in specific topics covered above include: Swets and Picket 1982, de Valois and de Valois 1988, Macmillan and Creelman 1991, Shapley and Lam 1993, Wandell 1995, Backhaus et al. 1998, Rodieck 1998, Oyster 1999.

See also: Binocular Space Perception Models; Color Vision; Color Vision Theory; Eye Movements and Perception; Feature Representations in Cognitive Psychology; Motion Perception: Psychological and Neural Aspects; Neural Plasticity in Visual Cortex; Object Recognition: Theories; Perception: History of the Concept; Perception of Extrapersonal Space: Psychological and Neural Aspects; Perceptual Organization; Psychophysical Theory and Laws, History of; Signal Detection Theory; Vision for Action: Neural Mechanisms; Vision, High-level Theory of; Vision, Low-level Theory of; Visual Development: Infant; Visual Imagery, Neural Basis of; Visual Memory, Psychology of; Visual Perception, Neural Basis of; Visual Space, Geometry of; Visual System in the Brain

Bibliography

Backhaus W G K, Kliegl R, Werner J S 1998 *Color Vision: Perspectives from Different Disciplines*. Walter de Gruyter, Berlin

Blakemore C, Campbell F W 1969 On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology* **203**: 237–60

Campbell F W, Robson J G 1968 Application of Fourier analysis to the visibility of gratings. *Journal of Physiology* **197**: 551–66

Coren S, Ward L M, Enns J T 1999 *Sensation and Perception*. Harcourt Brace, New York

Daugman J G 1980 Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research* **20**: 847–56

de Valois R L, de Valois K K 1988 *Spatial Vision*. Oxford University Press, New York

Fechner G T 1860 *Elemente der Psychophysik*. Breitkopf and Härtel, Leipzig, Germany
Gabor D 1946 Theory of communication. *Journal of the Institute of Electrical Engineers* **93**: 429–57
Goldstein E B 1999 *Sensation and Perception*. Brooks/Cole Publishing, Pacific Grove, CA
Hecht S, Shlaer S, Pirenne M H 1942 Energy, quanta, and vision. *Journal of General Physiology* **25**(6): 819–40
Levine M W 2000 *Levine and Shefner's Fundamentals of Sensation and Perception*. Oxford University Press, Oxford, UK
Macmillan N A, Creelman C D 1991 *Detection Theory: A User's Guide*. Cambridge University Press, Cambridge, UK
Marcelja S 1980 Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America* **70**(11): 1297–300
Newton I 1704 *Opticks, or, A Treatise of the Reflexions, Inflexions and Colours of Light; also Two Treatises of the Species and Magnitude of Curvilinear Figures*. Printed for Sam Smith and Beni. Walford printers to the Royal Society, London
Oyster C W 1999 *The Human Eye: Structure and Function*. Sinauer Associates, Sunderland, MA
Rodieck R W 1998 *The First Steps in Seeing*. Sinauer Associates, Sunderland, MA
Schiffman H R 2000 *Sensation and Perception: An Integrated Approach*. Wiley, New York
Sekuler R W, Blake R 2002 *Perception*. McGraw-Hill, New York
Shapley R M, Lam D M-K (eds.) 1993 *Contrast Sensitivity: Proceedings of the Retina Research Foundation Symposia*. MIT, Cambridge, MA
Stevens S S 1951 Mathematics, measurement, and psychophysics. In: Stevens S S (ed.) *Handbook of Experimental Psychology*. Wiley, New York, Chap. 1, pp. 1–49
Swets J A, Picket R M 1982 *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York
Vos J J, Walraven P L 1972a An analytical description of the line element in the zone-fluctuation model of color vision: I. Basic concepts. *Vision Research* **12**: 1327–44
Vos J J, Walraven P L 1972b An analytical description of the line element in the zone-fluctuation model of colour vision: II. The derivation of the line element. *Vision Research* **12**: 1345–65
Wandell B A 1995 *Foundations of Vision*. Sinauer Associates, Sunderland, MA
Werner J S, Wooten B R 1979a Opponent chromatic mechanisms: relation to photopigments and hue naming. *Journal of the Optical Society of America* **69**(3): 422–34
Werner J S, Wooten B R 1979b Opponent chromatic response functions for an average observer. *Perception & Psychophysics* **25**(5): 371–74

L. O. Harvey, Jr.

Visual Anthropology

1. Definitions

No short phrase will catch the two main streams of thought and practices which 'visual anthropology' implies. One sense refers to attempts to communicate anthropological observations and insights through photography, film, and video. In this sense, many early anthropologists who used these media were practicing forms of visual anthropology before the approach had been made explicit. Margaret Mead and Gregory Bateson used cine film and still photography as central to their studies of culture and personality, and later Jean Rouch and Timothy Asch both filmed extensively and argued for the recognition of film in anthropology. John Collier did the same for photography, as did Paul Hockings, for both mediums, working to formalize the importance of such contributions to academic and humanistic anthropology.

The second definitional approach to visual anthropology is more recent and is concerned with the understanding of visual systems in cultures and societies. It shares common concerns with the *anthropology of art* and with *technology and material culture*. By convention, the subject of literacy and its significance has not been prominently included within visual anthropology, but is normally discussed in wider debates about the transition from the neolithic to large-scale states and empires, and debates about power, hierarchy, and authority. Although visual anthropology is of increasing interest to students, and younger academics, it as yet shows few signs of making serious inroads into the hegemony of classical political economy and other mainstream concerns. The approach claims that special attention to visual systems may lead to reassessment of mainstream assumptions and analytic givens. 'Understanding visual systems' can be broken up into three related elements.

First, the study of the significance of visual perception in human cognition, generally: how far do humans, as a species rely on what is seen, and how do they integrate what they see into their deeper and wider understandings and practices? Psychologists have studied the dominance of one sense over another in experimental situations; anthropologists are increasingly interested in nonlinguistic and prelinguistic cognition, some of which depends in part on visual perception, and are more interested in psychological work than they have been for many years. In any specific society, how great a part do visual symbols, signals, emblems, and sign systems play in the construction of social order? Traffic lights, national flags, systems of body modification, ritual objects, house styles, animal and plant classifications are clearly all necessarily dependent on visual information. How important is this? Do systems of visual signification have any autonomy from the cultural and international systems in which they are necessarily situated?

The second element in this approach concerns the significance of the visual in any particular society or civilization and, related to this, how a given history affects the use made of a visual technology, such as photography. Do some societies place a much greater emphasis on the visual communication of knowledge,

and religious truths than do others? It has been suggested, for example, that in particular societies, the young learn how to do things, particularly artisanal production, by watching experts at work, rather than by explicit verbal and conceptual instruction. How far are singing, dance styles, facial expressions, body postures, clothing forms and usages, the whole range of physically performed elements of cultural behavior, deeply and implicitly shaped by visually led cultural learning? Are there austerely 'unvisual' societies in which speech 'leads' and visibly performed action is less strongly marked? Once, industrial societies relied on reading, memorizing, classroom and radio listening for knowledge transfer, but now television and digital memory storage are major inputs into many lives. What difference does this shift in culturally packaged techno-learning make?

The third element concerns what is broadly termed issues of (visual) representation. The word *represent* has several senses—to bring something into existence, to make visible, to exhibit or reproduce an action, or a person, as upon the stage, to take the place of something, and to stand as an example of something. French anthropology, through the work of Durkheim and Mauss promoted the idea of rituals as 'collective representations,' but the term influenced visual anthropology more through Edward Said's *Orientalism*: 'The idea of representation is a theatrical one: the Orient is the stage on which the whole East is confined. On this stage will appear figures whose role it is to present the larger whole from which they emanate. The Orient then seems to be not an unlimited extension beyond the familiar European world, but rather a closed field, a theatrical stage affixed to Europe' (Said 1978, p. 63). The paragraph goes on to list some of the stock characters on this stage including the Sphinx, Cleopatra, Babylon, Babel and Mahomet. Said's deeper concern is with the production of reductive, and often negative stereotypes of peoples and cultures. Simply put, then, the issue of representation in visual anthropology has become a many-sided discussion of how far a painting, a postcard, a photograph, a film, a video, can be understood to have reasonably and sensitively portrayed the persons or issues with which they are concerned. A good starting point is Roland Barthes' account of his difficulty in finding a photograph of his mother which really did justice to his memories of her. A sensitive discussion of more impersonal representation issues in film-making appears in Ruby (1991): ownership, collaboration, informed consent, censorship, and the right to make representations of other people are searchingly discussed.

2. *Intellectual Contexts*

Visual anthropology can be said to begin when, in the middle of the nineteenth century, photographs are first used as ethnological records of human diversity, as images of colonized peoples for the appreciation of metropolitan publics, and, before long, for portraiture by all, metropolitan or colonial, who had access to the technologies. We discuss these issues in more detail below. However, in the modernist sense, visual anthropology could, and perhaps should, begin with a consideration of all visual systems employed in depicting human action, before photography became a dominant mode. This implies a discussion of the cave drawings, etchings, woodcuts, lithographs, paintings, and many other forms of image employed in early African, Arab, Chinese, European, Hindu, Japanese, Moghul, and many other sources, whenever faced with the task of communication of cultural differences. As yet, such interests are mainly pursued by archeologists and historians of art, but their preoccupations are inevitably somewhat different from those of anthropologists. Not only would illustrations in texts be of interest, but so would objects used in ritual, and the decorations of objects in daily use.

Visual anthropology has in fact developed differently. First, there has been the practical use of photographs from 1860 to the twenty-first century at first untheorized, taken-for-granted as unproblematic additional field research tools. The second component has been an interest in the use of film as a scientific recording medium. In the 1950s and 1960s, articles started to appear which argued forcefully for the importance of 'ethnographic film' as a complementary field research tool for mainstream anthropology. This approach was summarized in Hockings (1995). In the spirit of recording for posterity, several large collections of ethnographic film were created (see below). The third major development, in the 1980s, was a growth of intellectual interest in how photography had been used by anthropologists in the colonial period. In the 1990s, visual anthropology includes many new areas for inquiry, and there is increasing boldness in its claims to raise questions ignored by visually nonliterate, concept-fixated anthropologists.

3. *Photography*

The first great achievement of scholars engaging with photography has been to reverse the truism that 'Every picture tells a story' so that it reads 'Every picture needs its history.' An example is H. L. Seneviratne's essay on a photograph taken at the Temple of the Tooth, Kandy, Sri Lanka (then Ceylon), at some point between 1870 and 1900. The photograph shows the Temple building, with three monks in the top right-hand corner, looking down upon a circle of some 30 persons, many of whom are monks in positions for formal reverence grouped around a table, at a lower level. The author argues that what is visually central in the photograph is in fact politically and semantically peripheral. The author is able to identify the 'central'

figures as numerous Burmese monks, and the three elevated ones as Sri Lankan monks, dressed relatively casually. The fact that an electric light fitting is visible helps date the picture. However, the analysis raises serious doubts about the probability that the object on the 'central' table is the Sacred Tooth relic which the ritual postures of the 'central' monks seems to imply. Three reasons are offered: first, a lack of supporting attendants; second, the relative simplicity and austerity of the table on which the objects are displayed; and third, the Relic is never brought down from the higher level of the building. Seneviratne concludes 'The occasion of the picture thus represents a variety of possibilities, the more so for want of precise information' (Edwards 1992, p. 156). It should be noted that this analysis depends on significant historical knowledge external to the photograph itself: every photograph is historical from the moment it has been taken, and every photograph needs its historian.

The second achievement has been to question implicit realist assumptions about how photography 'works.' That the camera permitted 'stern fidelity' and 'penetrating certainty' (Pinney 1997) was a dominant assumption of many European colonial photographers. Pinney suggests that one use of the camera in colonial India was linked to the task of ethnological classification: photographs were taken of exemplary 'representatives' of cultural groups. Faces, clothes, hair-styles, and adornments were held to characterize named castes and ethnic groups decisively. Foucault's concern with definitive classification and state surveillance helps us see that the evidential fingerprint and the police arrest photograph are inherently realist, too. However, Pinney also describes the enthusiasm with which elite Indians under British rule took up this European invention and developed it for their own purposes, at first employing European conventions of formal portraiture. He also shows how some early British photographers in India used portraiture in a more sensitive way than Foucault's focus would allow, with empathic regard for the qualities of the person portrayed.

The third achievement is to identify specific 'ways of seeing.' Pinney's account of how contemporary Indian villagers in Nagda, Madhya Pradesh, use photographs is illuminating. They opt for portrayals of persons not 'as they are,' spontaneous, informal, and unadorned, but rather 'formalized and theatrically staged images … are most desired.' This may include adorning a photographic image with watercolors, the juxtaposition of a villager with a famous film star, palaces or temples as backgrounds, or wearing special clothes as if a rajah, a bandit, a warrior, or a princess. The 'Indian Eye' is not primarily a naturalistic eye but is concerned to make 'traces of the person endure, and to construct the world in a more perfect form than is possible to achieve in the hectic flow of the everyday' (Pinney 1997, p. 149) (see also the film by David and Judith MacDougall, *Photowallahs*, 1991).

4. *Ethnographic Film*

Film was used by ethnologists as early as 1898, and although travelers and ethnologists continued to make filmed records of native peoples, such material hardly received serious anthropological attention until the 1950s. Enthusiasts for film tended to stress its assumed objectivity and its capacity for contributing to anthropology as a behavioral science, and fitness for tasks of ethnographic salvage, conserving records of rituals, techniques, and practices among peoples threatened with extinction or radical transformation. Certain film makers in the naturalistic documentary tradition undertook large-scale filming projects among particular peoples, such as Ian Dunlop's work with members of the Mandjindjara and Ngadajara groups in the Australian Western Desert, and his subsequent work with the Baruya of East Highlands, Papua New Guinea, Timothy Asch's collaboration with Napoleon Chagnon, among the Yanomami, Christopher Curling and Melissa Llewelyn-Davies' work among the Loita Maasai in Kenya, David Turton and Leslie Woodhead's films among the Mursi of southwest Ethiopia, and some of Jean Rouch's work among the Dogon and Songhay of West Africa.

Such attempts to film naturalistically and objectively, for the historical or archival record, were soon challenged by a different approach, in which the aims were not scientific, but more exploratory, celebratory, observational filming without authoritative commentary, and with a sense of intimacy between film makers and subjects. This has been described as 'participatory' or 'collaborative' cinema (MacDougall 1998) in which the film maker's aim was to allow the personal qualities of individuals to be conveyed, as well as characteristic themes and emphases from the particular society in which they lived. It also encourages what has been termed 'reflexivity' (Ruby 1980), in which the film maker made the intellectual premises on which the film was being made transparent to the audience.

This approach required a decisive break with any suggestion that films were made by detached, all-seeing, and all-knowing scientists presenting 'objective truths' about those filmed. 'Reflexivity' was sometimes understood superficially to mean that the film-makers should introduce themselves and the filming project as the framework within which the film developed. While some anthropological film-makers still work in the observational tradition, more experimental forms of narrative are gaining adherents.

5. *Changes in Emphasis and Future Developments*

Visual anthropology has been influenced by the wider theoretical changes which have affected mainstream anthropology in North America, and Europe—skepticism about objectivity, successive impacts of French structuralism, Marxism, feminism, and post-modernism upon choice of subjects and their treat-

ment. The simplest way to sum up the current state of thinking is to say that visual anthropology, like mainstream anthropology itself, is becoming ever more inclusive, in several different senses.

First, there is an inclusiveness of subjects for description and analysis. As the world continues to change under the impact of global technology transfer and market forces, so new social relations and new social issues emerge, and classical assumptions need reconsideration. Labor migrations and the formation of diaspora communities who keep in touch with their recent origins by cheap air travel, the telephone, and the internet are examples. The wedding video, enthusiastically adopted in India, as in Europe, represents a new field of study, and the styles, the emphases, the inclusions and exclusions, and the contexts of use and reuse are being described and analyzed. The consumption of media products, particularly television soap operas, and national cinema productions are attractive to visual anthropologists because they tend to miniaturize and bring into relief deep cultural themes, and this is doubly the case when the diffusion, both positive flows and negative obstructions, can be included. For example, Greek Cypriots who in the 1970s were likely to watch Greek and Turkish fiction films and US-originated soap operas, in the 1980s started to favor a Mexican soap. In the UK, Australian soaps aimed at teenagers are now more successful than US ones. In Africa, Hindi films were once much viewed, and Australian Aboriginal peoples have enjoyed Kung-Fu films.

Second, the inclusiveness involves what has come to be termed 'indigenous media' activity. This takes several forms. First, there are broadcasting systems run by and for ethnic minorities. Australia and Canada have led the way in these matters, with aboriginal peoples producing local materials in their own languages, to celebrate cultural survival and resistance to national cultural impositions of uniformity. Indigenous media may show films by ethnic minorities to promote their political interests in the face of national systems. Thus films about land rights and the connection between a group and their territory through ancestral religious cults are prominent, as are films celebrating the tribal past, documenting the tribulations of previous generations, and movements for revitalization.

Third, indigenous visual anthropologists came forward to 'speak about' (make films about) cultures with which they are intimately connected by various links. Such work may implicitly or explicitly challenge the expertise of the 'outside' observer. Once there was a widespread assumption (underpinned by ideas about the necessity of scholarly detachment) that the best anthropology had to be done in a society to which the anthropologist came as a cultural outsider. Most scholars would now say that as anthropology has shifted from being primarily a Euro-American discipline to being a more international one, it has become clear that it is the nature of the intellectual training and not the anthropologist's originating culture which makes illuminating cultural analysis possible. Visual anthropology has been at the cutting edge in these matters.

The major claim from postmodern visual anthropology which has yet to be accepted by mainstream practitioners is that a great intellectual role for 'the visual' will challenge conventional thinking, and refresh it (Devereaux and Hillman 1995, Banks and Morphy 1997, MacDougall 1998). Since all postmodern anthropology questions all taken-for-granted classifications, hierarchies, and orthodoxies, the pro-visual anthropologists are unlikely to 'come to power' intellectually, but rather 'the visual' will probably be quietly and painlessly incorporated into mainstream anthropology courses. There will be little resistance to proposals to think freshly about 'visual systems,' although there will always be scholars who continue to argue in support of writing as the best mode of intellectual communication and the written ethnography as superior to any other form of discourse (see, for example, Hastrup in Crawford and Turton 1992). However, just as the ethnographic museum can be reconfigured to become a place which makes us think actively about history, power, tolerance, creativity, and core values, so too can a photograph, a Benin bronze, a fish trap, an MTV pop video, a British imitation of a Japanese garden, and a set of 'dirty' postcards be appropriately re-presented and freshly analyzed.

See also: Art: Anthropological Aspects; Art, Sociology of; Ethnography; Photography as a Medium; Photojournalism; Visualization: Representation in Science

Bibliography

Banks M, Morphy H (eds.) 1997 *Visual Anthropology*. Yale University Press, New Haven, CT

Barbash I, Taylor L 1997 *Cross-cultural Filmmaking: A Handbook for Making Documentary and Ethnographic Films and Videos*. University of California Press, Berkeley, CA

Collier J Jr, Collier M 1986 *Visual Anthropology: Photography as Research Method*. University of New Mexico Press, Albuquerque, NM

Crawford P I, Turton D 1992 *Film as Ethnography*. Manchester University Press, Manchester, UK

Devereaux L, Hillman R 1995 *Fields of Vision: Essays in Film Studies, Visual Anthropology and Photography*. University of California Press, Berkeley, CA

Edwards E 1992 (ed.) *Anthropology and Photography 1860–1920*. Yale University Press, New Haven, CT

Hastrup K 1992 Anthropological visions. Some moves on visual and textual authority. In: Crawford P I, Turton D (eds.) *Film as Ethnography*. Manchester University Press, Manchester, UK, pp. 8–25

Hockings P 1995 (ed.) *Principles of Visual Anthropology*, 2nd edn. (1st edn. 1975). Mouton de Gruyter, Berlin

Loizos P 1993 *Innovation in Ethnographic Film: From Innocence to Self-consciousness 1955–1985*. Manchester University Press, Manchester, UK and Chicago University Press, Chicago
MacDougall D 1998 *Transcultural Cinema* (edited and with an Introduction by Taylor L). Princeton University Press, Princeton, NJ
Pinney C 1997 *Camera Indica: The Social Life of Indian Photographs*. Reaktion Books, London
Ruby J 1980 Exposing yourself: reflexivity, anthopology and film. *Semiotica* **3**: 153–79
Ruby J 1991 Speaking for, speaking about, speaking with, or speaking alongside—an anthropological and documentary dilemma. *Visual Anthropology Review*, Fall **VII**(2): 50–63
Ruby J 2000 *Picturing Culture: Explorations in Film and Anthropology*. Chicago University Press, Chicago
Russell C 1999 *Experimental Ethnography: The Work of Film in the Age of Video*. Duke University Press, Durham, NC
Said E 1978 *Orientalism*. Pantheon, New York
Stoller P 1992 *The Cinematic Griot: The Ethnography of Jean Rouch*. University of Chicago Press, Chicago

P. Loizos

Visual Development: Infant

Immediately after birth, the visual system (see *Visual System in the Brain*) has access to structured visual input from the environment that could be used to guide behavior in, and perception of, the external world. Developmental events prior to birth establish rudimentary neural mechanisms that enable some limited perceptual capacities, but further neural developments as well as interactions with the visual environment lead to substantial improvements in these capacities after birth. This article covers the emergence of these perceptual capacities during the early postnatal period in human infants (see also recent monographs by Atkinson 2000, Daw 1995, Kellman and Arterberry 1998, Simons 1993). After considering methods that have been used to assess infant vision, the most fundamental aspects of visual development will be reviewed.

1. *Methods for Assessing Infant Vision*

Prior to 1960, most anecdotal and published accounts of the visual capacities of human newborns concluded that they were blind or, at best, severely visually impaired. These assessments, however, were based largely on insensitive methods, such as moving a single, small spot of light to elicit from the newborn a change in eye alignment (an ocular following or tracking response). Four methodological breakthroughs (see Gottlieb and Krasnegor 1985) occurred in the early 1960s which revealed that the visual capacities of young infants, although considerably less mature than in adults, were nevertheless quite sophisticated.

1.1 *Preferential Looking*

In the late 1950s, Robert Fantz revolutionized the study of infant vision by modifying a two-alternative preference technique originally developed for testing visual sensitivity in chickens. Fantz substituted the infant's looking response for the chicken's pecking response and presented infants with two side-by-side stimulus locations, only one of which contained a set of black-and-white stripes. Even newborns have a natural tendency to fixate (look at) the side of the display containing the stripes over the side of the display devoid of stripes. An observer viewing the infant's eyes, and estimating the direction of gaze through a peephole in the stimulus display, judges which of the two stimulus locations is fixated more often or for a longer duration. To guard against bias, the observer is 'blind' (i.e., unaware of the location) of the stripes. By systematically varying the width of the stripes, the smallest stripe-width that is just barely resolvable (significantly greater than chance, or 50 percent) can be determined across a series of trials in which the location of the stripes is randomly varied.

Teller (1979) established rigorous standards for using infant looking preferences to assess stimulus *detection* (a choice between a stimulus and no-stimulus). Teller coined the term forced-choice preferential looking (FPL) to describe this technique. FPL can also be used to assess stimulus *discrimination* by presenting two stimuli simultaneously. Evidence of a significant preference allows for the conclusion that infants can discriminate the difference between the two stimuli. However, the absence of a significant preference is ambiguous because the two stimuli may be discriminable but fail to elicit a clear preference in looking behavior.

1.2 *Visual Evoked Potentials*

Surface electrodes attached to the skull can record small changes in synchronized electrical activity from various brain regions. The amplitude and latency of several peaks and troughs in the visual evoked potential (VEP) vary with the size and intensity of a visual stimulus. As with preferential looking, VEPs can be used to assess the smallest stripe-width that is just detectable. VEPs can also be used to assess visual discrimination by measuring the averaged response to a change from one stimulus to a different stimulus and comparing it to the averaged response to no-change control trials.

1.3 *Fixations and Eye-movement Scanpaths*

Newborns show poor evidence of steady and consistent fixations of a small visual target, and their ability to track (follow) a small moving target is very poor. They will, however, follow a set of moving stripes. This following response is called optokinetic

nystagmus (OKN) because it consists of alternations between a smooth tracking response in the direction of stripe movement and a rapid saccadic eye movement in the direction opposite to stripe movement (as the eyes move from one extreme orbital position to the other). As with FPL and VEP, OKN can be used to assess the smallest stripe-width that is just detectable. OKN can be measured by direct observation or by one of a variety of eye-tracking devices, such as electro-oculography (EOG) or video-based images of the pupil and reflections on the front surface (cornea) of the eye. These latter corneal reflection techniques were developed by William Kessen and his colleagues (see Haith 1980) and provided the first detailed assessments of moment-to-moment sequences of visual fixations (scanpaths) by newborns as they viewed a simple visual stimulus.

1.4 Habituation

Fantz noted in his early studies of preferential looking that when the same two stimuli were presented side by side, fixation durations declined across repeated trials. This decrement in looking duration is called *habituation* and indicates that the infant has processed and retained some information about the stimulus across time. Subsequent elaborations on the use of habituation as a measure of visual discrimination (Bornstein 1985) have resulted in an 'industry standard' that consists of two phases: habituation and test. During habituation, a single visual stimulus is presented and an observer records how long the infant sustains fixation (interrupted by no more than two seconds of distraction) to that stimulus on a given trial. Identical trials are repeated, tallying fixation duration until there is a preset decrement (typically 50 percent) from the initial level of looking on the first several trials. When this criterion of habituation has been met, the infant enters a test phase in which the same habituation stimulus and a novel stimulus are presented on alternating trials. Continued low levels of looking to the habituation stimulus and significant increases (recovery) of looking to the novel stimulus are taken as evidence of discriminating the habituation (familiar) from the novel stimulus.

2. Resolving Fine Detail

2.1 Acuity

Perhaps the most basic question one could ask about infant vision is how well they can see small objects. This ability to resolve fine detail, called visual acuity, is what is assessed by an ophthalmologist or optometrist using an eye chart. Because infants cannot verbally report the smallest letter-size on an eye chart, one of the foregoing non-verbal techniques must be used to estimate their visual acuity.

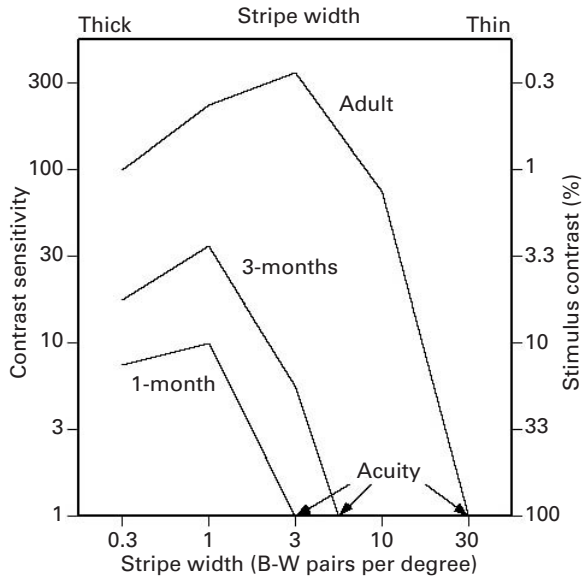

Figure 1
Schematized contrast sensitivity functions for one-month-old and three-month-old infants and for normal adults. The x-axis represents stimulus size (e.g., the number of black-and-white stripes per degree of visual angle) and the y-axis represents sensitivity to contrast.

FPL, VEP, and OKN have been used to assess visual acuity in young infants. In each case, black-and-white stripes (or checks) are systematically varied in size to determine the smallest size that elicits a reliable response. Although there are some differences in estimates of visual acuity using these three methods, two facts are quite clear. First, visual acuity in newborns is approximately five octaves (an octave is a doubling) worse than in normal adults. Using the Snellen notation, where 20/20 refers to normal visual acuity (average performance at a viewing distance of 20 feet), newborn visual acuity is approximately 20/640. This means that a newborn can see at 20 feet what a normal adult can see at 640 feet. Alternatively, a newborn must be five times closer to a visual stimulus than an adult (7.5 inches vs. 20 feet) to resolve it equally well. Second, there is a rapid improvement in visual acuity between birth and six months of age, with average acuity rising to 20/40 within this postnatal period (Norcia and Tyler 1985).

2.2 Contrast Sensitivity

Most objects are not composed of features at the limit of resolution, but rather contain features sufficiently large to be well above the acuity threshold. For these features, size is less important than contrast. Contrast refers to the difference in luminance between adjacent object features. For example, a pair of black-and-

white stripes has very high contrast because the luminance difference is large. As luminance is added to the black stripe and subtracted from the white stripe, the stripes become different shades of gray and have lower contrast. At the limit, when the stripes are equal shades of gray, contrast is zero and the stripes are no longer visible. Thus, for any feature above the acuity threshold, visibility is determined by sensitivity to contrast.

Figure 1 illustrates a typical contrast sensitivity function (CSF) obtained from an adult, as well as CSFs obtained from young infants (Kellman and Banks 1998, Norcia et al. 1990). Notice that the adult CSF peaks at features with medium size and falls off for both smaller and larger sizes. The point on the far right of the CSF, where size is smallest and contrast sensitivity is least (i.e., stimulus contrast is maximal), is the estimate of visual acuity. The lower heights of the infant CSFs indicate that they are much less sensitive to contrast than adults at all feature sizes, and of course at maximal stimulus contrast visual acuity is far poorer. Importantly, even at medium feature sizes where both infants and adults can resolve a stimulus at high contrast, infants require much more contrast for minimal detection of that stimulus. Because many features of objects have low contrast, these features are invisible to the infant's visual system.

2.3 Limiting Factors

The two most obvious limitations on acuity and contrast sensitivity in early infancy are optical quality and neural immaturity. Adults with an optical error requiring spectacles or contact lenses show deficits in both acuity and contrast sensitivity when they fail to wear their correction. Although a small proportion of infants have optical errors (myopia: nearsightedness; hyperopia: farsightness; astigmatism: errors that vary by stimulus orientation), the five-fold deficit in newborn acuity cannot be accounted for by optical errors (Banks and Bennett 1988). Another potential contributor to optical quality is accommodation (change in shape) of the lens in the eye, which optimizes the focus of the retinal image for different viewing distances. Although newborns show poor accommodative control, this deficit is the product of poor visual acuity; that is, the immature visual system is unable to detect subtle changes in stimulus blur that trigger an accommodative response (Kellman and Banks 1998).

These results suggest that the fundamental limitation on acuity and contrast sensitivity is neural, not optical. Classic data on the developmental anatomy of the visual cortex in human infants (Conel 1939) documented substantial elaborations of neural structures (e.g., density of synapses) during the first two postnatal years. However, more recent data on the developmental anatomy of the retina suggest that much of the five-fold improvement in acuity can be accounted for by the increasing density of photoreceptors in the center (fovea) of the retinas and the inefficiency of photoreceptors (the outer segments) in capturing light (Banks and Bennett 1988).

Both maturational and experiential factors influence the development of acuity. The timecourse of developmental improvement in acuity is more closely linked to the infant's age post-conception than to its age post-birth, suggesting that normal visual experience after birth plays a relatively minor role in this developmental process. However, the absence of normal visual experience after birth leads to substantial deficits in acuity. Visual deprivation in the form of cataracts (opacities in the eye that prevent a clear retinal image) slows down the timecourse and eventual level of acuity unless the cataract is removed very early in infancy (Maurer and Lewis 1993).

3. Color

Chromatic discrimination requires at least two classes of photoreceptors (cones) that are sensitive to slightly different wavelengths of light (see article *Color Vision*). FPL, VEP, and OKN assessments confirm that the medium-wavelength (green) and long-wavelength (red) photopigments are functional by two months of age. However, the immaturity of the photoreceptor outer segments (see Sect. 2.3) limits the number of different hues that young infants can discriminate (Banks and Bennett 1988). Clear evidence that infants are trichromats (i.e., have the red, green, and short-wavelength sensitive blue cones required for normal color vision) is not present until four months of age (Teller and Bornstein 1987). Color constancy, the ability of infants to perceive the correct hue despite variations in lighting (e.g., ignoring the reddish tint at sunset), emerges between two and five months of age (Dannemiller 1989).

4. Motion

Moving stimuli are more effective at capturing infants' attention than stationary stimuli. This natural bias to attend to moving stimuli has been used in preferential looking studies to estimate motion thresholds: the minimum stimulus speed required to discriminate a moving from an otherwise identical stationary stimulus. Motion thresholds improve dramatically in early infancy despite nearly adultlike temporal sensitivity at birth, suggesting that these improvements in motion sensitivity rely critically on improvements in spatial resolution (Roessler and Dannemiller 1997).

Discrimination of different directions of motion is the definitive test of a motion mechanism because the responses of single neurons in primary visual cortex and many extrastriate cortical areas in the dorsal visual pathway (see *Neural Plasticity*) have a preferred direction of stimulus motion. However, both FPL and habituation techniques have failed to demonstrate

infants' discrimination of opposite directions of stimulus motion until two months of age. This is surprising because even newborns show directionally appropriate eye movements to stripe motion during OKN. Asymmetries in the direction of OKN (temporalward stripe motion is ineffective in eliciting monocular OKN in newborns) are mirrored in the VEP. Thus, the processing of motion direction appears to be mediated by at least two different neural mechanisms (see *Motion Perception: Psychological and Neural Aspects*). One of these mechanisms is presumably subcortical and is capable of controlling directional eye movements (i.e., OKN) in newborns, whereas the other mechanism is presumably cortical and does not support direction discrimination until several weeks after birth (Banton and Bertenthal 1997).

5. *Orientation and Vernier Acuity*

Another fundamental property of neurons in the visual cortex is orientation sensitivity, and this property is present in rudimentary form in newborn cats and monkeys. Human newborns are also sensitive to orientation, but only for gross differences (e.g., horizontal vs. vertical stripes). A VEP masking technique suggests that orientation sensitivity continues to improve throughout the first postnatal year (Skoczenski 2001).

Vernier or displacement acuity refers to the ability to discriminate the spatial misalignment of two line segments that are abutted end to end. In adults, vernier acuity is far superior to grating (stripe-width) acuity, whereas in infants prior to four months of age vernier acuity is poorer than grating acuity. This developmental difference in the relation between vernier and grating acuity has been documented using both FPL and VEPs (Skoczenski 2001). One potential cue for solving a vernier acuity task is the presence of a slight orientation change at the point where the two line segments meet. Thus, the protracted development of orientation sensitivity may acount for the developmental crossover in the relation between vernier and grating acuity.

6. *Depth and Binocular Rivalry*

The relative distance (depth) of objects can be appreciated using three different sources of information: motion, retinal disparity, and pictorial cues. A rapidly approaching (looming) stimulus elicits a blink response in one-month-olds, and motion parallax (more rapid image speed for near than for far objects) enables three-month-olds to discriminate small differences in object distance. Thus, sensitivity to depth from motion is present in very early infancy and does not require the use of both eyes (Kellman and Arterberry 1998).

Retinal disparity refers to the subtle differences in the images projected to the two retinas from an object at near (less than five meters) viewing distances. FPL and VEP studies have demonstrated that sensitivity to retinal disparity does not emerge until three to four months after birth. Moreover, the smallest retinal disparity that is just discriminable by infants improves very rapidly between three and five months of age, progressing from no sensitivity to nearly adult values (less than one minute of arc) in this age range (Birch et al. 1985).

During this same age range, infants become sensitive to binocular rivalry: the perceptual conflict induced by presenting grossly different images to the two retinas (e.g., horizontal stripes in one eye and vertical stripes in the other). Binocular rivalry occurs when the discrepant retinal images cannot be fused into a single percept. Prior to three months of age, infants appear to have a much greater tolerance for fusing discrepant images than adults, perhaps because of their poor acuity and contrast sensitivity (Birch et al. 1985).

In adults, failure to align both foveas onto a stimulus typically leads to binocular rivalry and prevents stereopsis (the appreciation of depth from retinal disparity). Some individuals, including some infants, have an ocular misalignment (strabismus) that eliminates fusion and stereopsis. If uncorrected in infancy, this misalignment can result in a permanent loss of the capacity for stereopsis, even if the eyes are surgically realigned in childhood (Banks et al. 1975). Thus, there is a sensitive period during which a normally developing neural mechanism for stereopsis, present by four months of age, can be permanently disabled by subsequent abnormal binocular experience (strabismus). These same processes were earlier demonstrated in the visual cortex of cats and monkeys (see *Neural Plasticity in Visual Cortex*), and subsequently confirmed behaviorally.

Pictorial cues to depth are contained in flat (two-dimensional) representations of actual (three-dimensional) scenes. These cues include shading, occlusion, and linear perspective (e.g., receding railroad tracks that converge in the picture plane). A preferential reaching technique was developed by Albert Yonas to determine, under monocular viewing conditions, whether infants perceive the depth information in pictures. At approximately seven months of age, infants under these testing conditions become sensitive to a variety of pictorial cues to depth, as indicated by their reliable reaching for the *apparently* nearer picture (Yonas and Granrud 1985).

7. *Object Perception*

As visual stimuli, objects are bounded three-dimensional collections of features. Even newborns fixate contours (high-contrast edges), suggesting that they can segregate a rudimentary object from its background. However, many objects are comprised of embedded sets of features, such as faces which have external (e.g., hairline) and internal (e.g., eyes) con-

tours. Newborns fail to fixate small, stationary internal contours of embedded-feature objects such as faces, presumably because of limitations in acuity and contrast sensitivity. After two months of age, infants both fixate and discriminate (as indicated by habituation studies) internal and external features, and often prefer to fixate internal features (Maurer and Salapatek 1976; see *Cognitive Control (Executive Functions): Role of Prefrontal Cortex*).

Objects can be extracted from a camouflaged background using differences in texture or motion. While adults can easily extract a single discrepant target element from an array of background elements, infants require highly salient element differences for such texture segregation, including target motion or multiple discrepant target elements (Sireteanu 2000). Orientation differences are particularly difficult for infants to use in texture segregation, consistent with the evidence (see Sect. 5) that sensitivity to orientation has a lengthy developmental timecourse.

By three months of age, the two-dimensional shape of an object can be extracted from coherent motion of a group of elements in a background of identical, but incoherently moving, elements. The three-dimensional shape of a rigid object can be extracted by four months of age from rotational motion that reveals the hidden perspectives of stationary views of the object. And the three-dimensional shape of non-rigid objects can be extracted by five months of age from displays consisting solely of disconnected points undergoing relative motion that simulates a human walker (Bertenthal 1993).

The actual size and shape of an object must be derived from properties of the retinal image and from object distance. For example, the retinal image of a familiar object (e.g., a dollar bill) varies in size when viewed at arms length or across a room, and in shape when viewed directly or from an oblique angle. Yet in all these cases we perceive the dollar bill to retain a constant physical size and shape. Size and shape constancy have been demonstrated in infants as young as newborns, suggesting that mechanisms for combining retinal image properties and object distance (presumably from motion cues) are functional with little or no postnatal experience (Kellman and Arterberry 1998).

Objects are often partially occluded by other objects, yet adults perceive objects as 'complete' despite the absence of various parts behind the occluder. Newborns do not appear to have an intact mechanism for perceptual completion, whereas four-month-olds do (Johnson 2000). Two-month-olds show perceptual completion, but only when given additional information for the continuity of the occluded object (e.g., a smaller occluder or one with gaps in it). And four-month-olds fail to show perceptual completion when the occluded object is stationary, suggesting that coherent motion of the visible object-parts binds them together into a single object.

8. Developmental Mechanisms

The development of mature visual perception during early infancy is influenced by both maturational and experiential mechanisms. Maturational factors include neural developments, such as the migration of photoreceptors (increasing the packing density of cones in the fovea), the elongation of cone outer-segments, and cell death (apoptosis). One result of such maturational factors is a reduction in the intrinsic neural noise that limits stimulus detection and discrimination (Skoczenski 2001). Experiential factors include periods of susceptibility to altered visual input. Although the range of visual inputs sufficient to enable 'normal' visual development is quite broad, visual deprivation (e.g., cataracts or strabismus) during a sensitive period can lead to permanent deficits in visual development.

See also: Binocular Space Perception Models; Linear and Nonlinear Programming; Perception and Action; Perceptual Organization; Vision for Action: Neural Mechanisms; Vision, High-level Theory of; Vision, Low-level Theory of; Visual Perception, Neural Basis of; Visual System in the Brain

Bibliography

Atkinson J 2000 *The Developing Visual Brain*. Oxford University Press, New York

Banks M S, Bennett P J 1988 Optical and photoreceptor immaturities limit the spatial and chromatic vision of human neonates. *Journal of the Optical Society of America A* **5**: 2059–79

Banks M S, Aslin R N, Letson R D 1975 Sensitive period for the development of human binocular vision. *Science* **190**: 675–7

Banton T, Bertenthal B I 1997 Multiple developmental pathways for motion processing. *Optometry and Vision Science* **74**: 751–60

Bertenthal B I 1993 Infants' perception of biomechanical motions: Intrinsic image and knowledge-based constraints. In: Granrud C (ed.) *Visual Perception and Cognition In Infancy: Carnegie Mellon Symposium On Cognition*. Erlbaum, Hillsdale, NJ, pp. 175–214

Birch E E, Shimojo S, Held R 1985 Preferential-looking assessment of fusion and stereopsis in infants aged 1 to 6 months. *Investigative Ophthalmology and Visual Science* **26**: 366–70

Bornstein M H 1985 Habituation of attention as a measure of visual information processing in human infants: Summary, systematization, and synthesis. In: Gottlieb G, Krasnegor N A (eds.) *Measurement of Audition and Vision in the First Year of Postnatal Life: A Methodological Overview*. Ablex, Norwood, NJ, pp. 253–300

Conel J L 1939 *The Postnatal Development of The Human Cerebral Cortex*. Harvard University Press, Cambridge, MA

Dannemiller J L 1989 Computational approaches to color constancy: Adaptive and ontogenetic considerations. *Psychological Review* **96**: 255–66

Daw N W 1995 *Visual Development*. Plenum Press, New York

Gottlieb G, Krasnegor N A 1985 *Measurement of Audition and*

Vision in the First Year of Postnatal Life: A Methodological Overview. Ablex, Norwood, NJ
Haith M M 1980 *Rules That Babies Look By*. Erlbaum, Hillsdale, NJ
Johnson S P 2000 The development of visual surface perception: Insights into the ontogeny of knowledge. In: Rovee-Collier C, Lipsitt L P, Hayne H (eds.) *Progress in Infancy Research*. Ablex, Mahwah, NJ, Vol. 1
Kellman P J, Arterberry M E 1998 *The Cradle of Knowledge: Development of Perception in Infancy*. MIT Press, Cambridge, MA
Kellman P J, Banks M S 1998 Infant visual perception. In: Damon W, Kuhn D, Siegler R S (eds.) *Handbook of Child Psychology: Cognition, Perception, and Language*. Wiley, New York, Vol 2
Maurer D, Lewis T L 1993 Visual outcomes in infant cataract. In: Simons K (ed.) *Early Visual Development: Normal and Abnormal*. Oxford University Press, New York, pp. 454–84
Maurer D, Salapatek P 1976 Developmental changes in the scanning of faces. *Child Development* **47**: 523–7
Norcia A M, Tyler C W 1985 Spatial frequency sweep VEP: Visual acuity during the first year of life. *Vision Research* **25**: 1399–405
Norcia A M, Tyler C W, Hamer R D 1990 Development of contrast sensitivity in the human infant. *Vision Research* **30**: 1475–86
Roessler J S, Dannemiller J L 1997 Changes in human infants' sensitivity to slow displacements over the first 6 months. *Vision Research* **37**: 417–23
Simons K 1993 *Early Visual Development: Normal and Abnormal*. Oxford University Press, New York
Sireteanu R 2000 Texture segmentation, 'pop-out,' and feature binding in infants and children. In: Rovee-Collier C, Lipsitt L P, Hayne H (eds.) *Progress in Infancy Research*. Ablex, Mahwah, NJ, Vol 1
Skoczenski A M 2001 Limitations on visual sensitivity during infancy: Contrast sensitivity, vernier acuity and orientation processing. In: Rovee-Collier C, Lipsitt L P, Hayne H (eds.) *Progress in Infancy Research*. Ablex, Mahwah, NJ, Vol 2
Teller D Y 1979 The forced-choice preferential looking procedure: A psychophysical technique for use with human infants. *Infant Behavior and Development* **2**: 135–53
Teller D Y, Bornstein M H 1987 Infant color vision and color perception. In: Salapatek P, Cohen L (eds.) *Handbook of Infant Perception: From Sensation to Perception*. Academic Press, New York, Vol. 1, pp. 185–236
Yonas A, Granrud C E 1985 Reaching as a measure of infants' spatial perception. In: Gottlieb G, Krasnegor N A (eds.) *Measurement of Audition and Vision in the First Year of Postnatal Life: A Methodological Overview*. Ablex, Norwood, NJ, pp. 301–22

R. N. Aslin

Visual Imagery, Neural Basis of

Human thought makes use of different forms of mental representation; some language-like and some more visual or spatial. The term 'mental imagery' refers to the latter. It plays a central role in many cognitive abilities, from creative problem solving to strategies for improving memory. This article will review the current state of knowledge on the neural bases of mental imagery, beginning with the controversies in cognitive psychology that motivated the neuroscience research.

1. *Imagery, Verbal Thought, and Perception*

Much of the early history of imagery research was devoted to discriminating imagery from verbal thought, and characterizing some of its differences in functional information-processing terms. Allan Paivio (e.g., 1971) addressed these issues within the context of memory research. He demonstrated that, in ostensibly verbal learning tasks, imagery affected memory encoding. Though a program of research in which subjects learned words, which either lent themselves to concrete images or not (e.g., banana vs. vacation), with or without imagery instructions, he gathered evidence in support of a 'dual coding hypothesis.' According to this hypothesis, imagery and language are two distinct forms of mental representation. Imagery was therefore helpful in verbal memory tasks because, in effect, it doubled the number of representations being stored.

A challenge to this view came from Zenon Pylyshyn (1973), who questioned the computational feasibility of storing visual images, and raised a number of issues concerning the role of imagery phenomenology in the information processing that underlies thought. This challenge brought an empirical response from Stephen Kosslyn (e.g., 1978), who devised experimental tasks in which the visual-spatial properties of images could be shown to affect subjects' information processing. For example, when subjects focus their attention on one location within an image, and then shift to a different location, the time taken to shift is directly proportional to the imagined distance (Kosslyn et al. 1978). Findings from the laboratory of Roger Shepard also supported the visual-spatial nature of imagery. In mental image rotation experiments, images took more time to rotate through larger angles (Cooper and Shepard 1973).

As evidence accumulated that imagery is a distinct form of mental representation from linguistic or propositional thought, the question of its relation to visual perception came to force. Early work on this topic was pioneered by Ron Finke (e.g., 1980), who demonstrated many detailed and striking similarities between mental images and percepts. For example, he reported that mental images had similar-shaped fields of resolution to the perceptual visual field.

2. *Image Representation: Insights from Neuroscience*

The relation between imagery and perception can also be assessed in terms of their respective neural substrates. Indeed, many of the alternative interpretations that plagued the purely behavioral approaches to

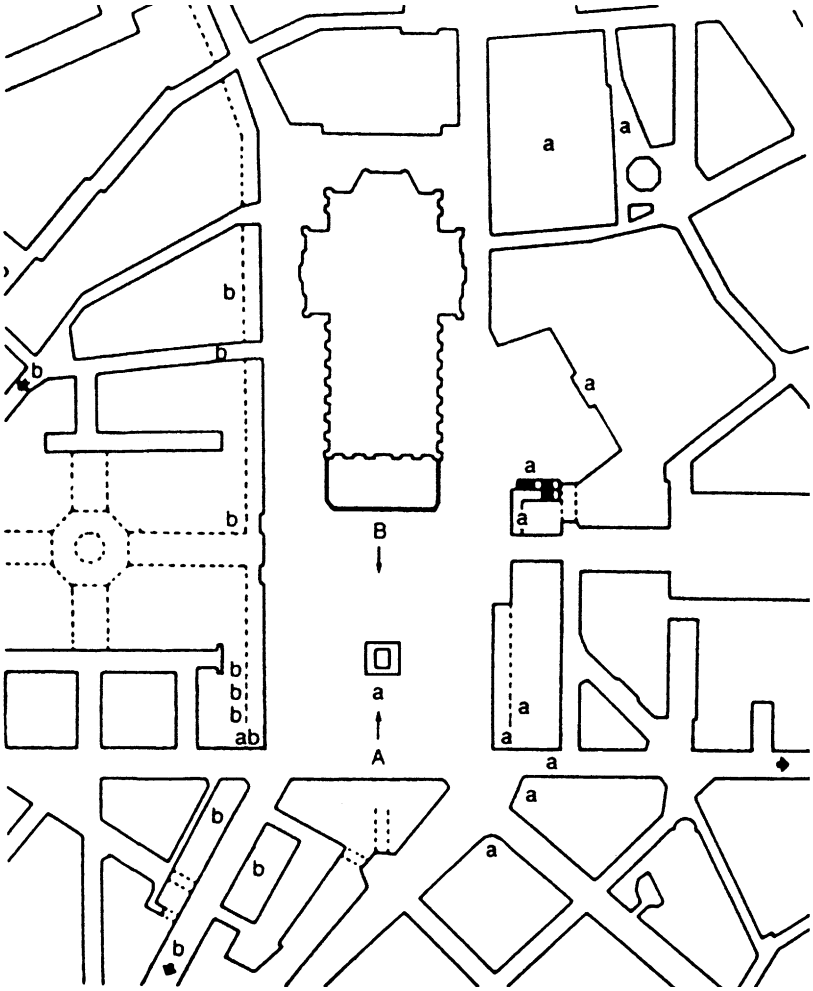

Figure 1
A map of the Piazza Del Duomo in Milan. When patients with left neglect were asked to imagine themselves standing at point A looking toward the cathedral, and to report what they saw in their 'mind's eye,' the locations they mentioned were those marked with 'a.' When they repeated the procedure from the vantage point B, they then mentioned the locations marked with a 'b.' (Source: Farah M J 1999 The neural basis of mental imagery. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA.)

mental imagery did not apply to the neural data. An initial review of the literature on imagery and the brain, motivated by the need for more decisive evidence concerning the imagery–perception relation, uncovered a variety of relevant findings (Farah 1988). These included numerous findings of parallel impairments of imagery and perception after brain damage, which suggest that the same underlying representations are needed for both.

In one of the best-known demonstrations of parallel impairments in imagery and perception. Bisiach and Luzzatti (1978) found that patients with hemispatial neglect for visual stimuli also neglected the contralesional sides of their mental images (that is, the side of their image opposite the side of their brain lesion). Their two right-parietal-damaged patients were asked to imagine a well-known square in Milan, shown in Fig. 1. When they were asked to describe the scene from vantage point A on the map, they tended to name more landmarks on the cast side of the square (marked with lower case a's in the figure); that is, they named the landmarks on the right side of the imagined scene. When they were then asked to imagine the square from the opposite vantage point, marked B on the map, they reported many of the landmarks previously omitted (because these were now on the right side of

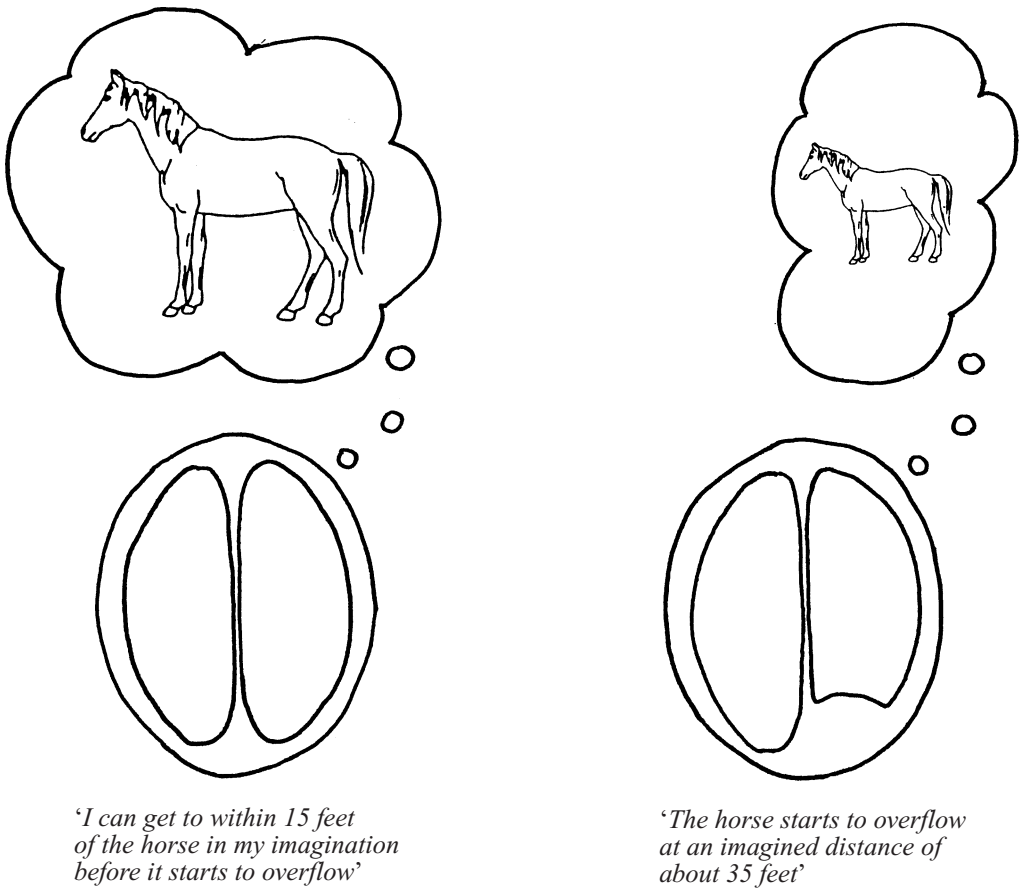

Figure 2
Depiction of the effects of unilateral occipital lobectomy on the visual angle of the mind's eye. (Source: Farah M J 1999 The neural basis of mental imagery. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA.)

the image) and omitted some of those previously reported.

Imagery and perception also share representations at relatively early stages of visual processing. This point was demonstrated by a patient with blindness in half of her visual field following occipital resection (Farah et al. 1992). If mental imagery consists of activating representations in the occipital lobe, then it should be impossible to form images in regions of the visual field that are blind due to occipital lobe destruction. This predicts a reduced maximum image size, or visual angle of the mind's eye, in this patient. By asking her to report the distance of imagined objects such as a horse, breadbox, or ruler when they are visualized as close as possible without 'overflowing' her imaginal visual field, one could compute the visual angle of that field before and after her surgery. It was found that the size of her biggest possible image was reduced after surgery, as represented in Fig. 2. Furthermore, the separate measurement of maximal image size in the vertical and horizontal dimensions showed that only the horizontal dimension of her imagery field was reduced significantly. These results provide strong evidence for the use of occipital visual representations during imagery.

The results of functional neuroimaging studies in normal subjects have been, on the whole, supportive of shared representations for visual mental images and percepts. Early studies with event-related potentials (e.g., Farah et al. 1989) and SPECT (e.g., Goldenberg et al. 1989) implicated a cortical visual locus and, as newer methods allowed better localization, it has been possible to test far more specific hypotheses concerning the precise areas within visual cortex that are recruited for mental imagery (Kosslyn and Ochsner 1994; Roland and Gulyas 1994).

3. Brain Systems for Image Generation

If imagery consists of activating some of the same cortical visual areas used for perception, this raises the question of how these representations become activated in the absence of a stimulus. Whereas one cannot see a familiar object without recognizing it, one can think about familiar objects without inexorably calling to mind a visual mental image. This suggests that the activation of visual representations in imagery is a separate voluntary process, needed for image generation but not for visual perception and object recognition.

Neuropsychological evidence for such a process comes from patients whose perception and general memory are preserved, but who cannot visualize objects or scenes from memory. This is the profile of abilities that would be expected given an impairment of image generation *per se*. A number of such cases have been reported in the neurological literature, with lesions generally being located in the posterior left hemisphere. In one such case (Farah et al. 1988), the imagery impairment was demonstrated in a number of ways, including a sentence verification task developed by Eddy and Glass (1981). Half of the sentences required the use of visual imagery to verify them (e.g., 'A grapefruit is larger than a cantaloupe'), and half did not (e.g., 'The US government functions under a two-party system'). Eddy and Glass had shown that normal subjects find the two sets of questions to be of equal difficulty (as did right-hemisphere-damaged control subjects tested by Farah et al. 1988), and that performance on the imagery questions was impaired selectively by visual interference, thus validating them as imagery questions. RM showed a selective deficit for imagery on this validated task: He performed virtually perfectly on the nonimagery questions, and performed significantly worse on the imagery questions.

RM was also tested on imagery for the colors of objects, using black and white drawings of characteristically-colored objects (e.g., a cactus, an ear of corn) for which he was to select the appropriate colored pencil. His imagery was further tested with drawing tasks. By these measures, too, his imagery was poor, despite adequate color perception and object recognition ability.

In subsequent years, a small number of additional cases of image generation deficit have been reported (e.g., Goldenberg 1992, Grossi et al. 1986, Riddoch 1991), as well as similar but weaker dissociations in subgroups of patients in group studies (Bowers et al. 1991, Goldenberg 1989, Goldenberg and Artner 1991, Stangalino et al. 1995).

What parts of the brain carry out image generation? This question has evoked controversy. Although mental imagery was for many years assumed to be a function of the right hemisphere, Ehrlichman and Barrett (1983) pointed out that there was no direct evidence for this assumption. Most of the patient-based research supports a left-hemisphere basis for image generation (see Tippett 1992, Trojano and Grossi 1994 for reviews; see Sergent 1990 for a different position). However, for questions of localization, we can also turn to functional neuroimaging.

Not all neuroimaging studies of visual imagery are appropriate for localizing image generation *per se*. The study must be designed in a way that can isolate this process, separated from perceptual and memory processing more generally. A large number of studies do meet this criterion, including those done with ERP, SPECT, PET and fMRI (see Farah 1999 for a review); in most cases, but not all, foci of activity are observed in the left inferior temporo-occipital region.

Although the bulk of evidence favors a left-hemisphere superiority for image generation, it also suggests that in most individuals both hemispheres possess some image generation ability. This hypothesis explains both the rarity of cases of image generation deficit after brain damage, as well as the asymmetries observed in neuroimaging studies, and the fact that when impairments are observed after damage, the left or dominant hemisphere is implicated.

See also: Imagery versus Propositional Reasoning; Mental Imagery, Psychology of; Perception: Philosophical Aspects; Perceptual Organization; Vision, Psychology of; Visual Perception, Neural Basis of; Visual System in the Brain

Bibliography

Bisiach E, Luzzatti C 1978 Unilateral neglect of representational space. *Cortex* **14**: 129–33

Bowers D, Blonder L X, Feinberg T, Heilman K M 1991 Differential impact of right and left hemisphere lesions on facial emotion and object imagery. *Brain* **11**: 2593–609

Cooper L A, Shepard R N 1973 Chronometric studies of the rotation of mental images. In: Chase W G (ed.) *Visual Information Processing*. Academic Press, New York

Eddy P, Glass A 1981 Reading and listening to high and low imagery sentences. *Journal of Verbal Learning and Verbal Behavior* **20**: 333–45

Ehrlichman H, Barrett J 1983 Right hemispheric specialization for mental imagery: A review of the evidence. *Brain and Cognition* **2**: 55–76

Farah M J 1988 The neural basis of mental imagery: Converging evidence from brain-damaged and normal subjects. In: Bellugi E A U (ed.) *Spatial Cognition: Brain Gases and Development*. Erlbaum, Hillsdale, NJ

Farah M J 1999 The neural basis of mental imagery. In: Gazzaniga M S (ed.) *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA

Farah M J 2000 *The Cognitive Neuroscience of Vision*. Blackwell, Oxford

Farah M J, Peronnet F 1989 Event-related potentials in the study of mental imagery. *Journal of Psychophysiology* **3**: 99–109

Farah M J, Levine D N, Calvanio R 1988 A case study of mental imagery deficit. *Brain and Cognition* **8**: 147–64
Farah M J, Soso M J, Dasheiff R M 1992 Visual angle of the mind's eye before and after unilateral occipital lobectomy. *Journal of Experimental Psychology: Human Perception and Performance* **18**: 1–6
Finke R A 1980 Levels of equivalence in imagery and perception. *Psychological Review* **87**: 113–32
Goldenberg G 1989 The ability of patients with brain damage to generate mental visual images. *Brain* **112**: 305–25
Goldenberg G 1992 Loss of visual imagery and loss of visual knowledge–a case study. *Neuropsychologia* **30**: 1081–99
Goldenberg G, Artner C 1991 A Visual imagery and knowledge about the visual appearance of objects in patients with posterior cerebral artery lesions. *Brains and Cognition* **15**: 160–86
Goldenberg G, Podreka I, Steiner M, Willmes K, Suess E, Dseoke L 1989 Regional cerebral blood flow patterns in visual imagery. *Neuropsychologia* **27**: 641–64
Grossi D, Orsini A, Modafferi A 1986 Visuoimaginal constructional apraxia: On a case of selective deficit of imagery. *Brain and Cognition* **5**: 255–67
Kosslyn S M 1978 Measuring the visual angle of the mind's eye. *Cognitive Psychology* **10**: 356–89
Kosslyn S M, Ochsner K N 1994 *Trends in Neuroscience* **17**: 289–91
Kosslyn S M, Ball T M, Reiser B J 1978 Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology*: *Human Perception and Performance* **4**: 47–60
Paivio A 1971 *Imagery and Verbal Processes*. Holt, Rinehart and Winston, New York
Pylyshyn Z W 1973 What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin* **80**: 1–24
Riddoch J M 1990 Loss of visual imagery: A generation deficit. *Cognitive Neuropsychology* **7**: 249–73
Roland P E, Gulyas B 1994 Visual imagery and visual representation. *Trends in Neurosciences* **17**: 281–7; 294–7
Sergent J 1990 The neuropsychology of visual image generation: Data, method, and theory. *Brain and Cognition* **13**: 98–129
Stangalino C, Semenza C, Mondini S 1995 Generating visual mental images: Deficit after brain damage. *Neuropsychologia* **33**: 1473–83
Tippett L J 1992 The generation of visual images: A review of neuropsychological research and theory. *Psychological Bulletin* **112**: 415–32
Trojano L, Grossi D 1994 A critical review of mental imagery defects. *Brain and Cognition* **24**: 213–43

M. J. Farah

Visual Images in the Media

Effects in the media of the visual image (i.e., a reproduced sight) are central to a debate which typically contrasts the properties of image and word. Here, that debate is tracked across three key areas: child literacy and education, commercial and political persuasion, and issues surrounding news.

1. *Visual Images vs. the Word*

The world is filled with visual images. Images appear on television and movie screens, billboards, posters, magazines, newspapers, and computer monitors. Increasingly, it is argued that the word, once the dominant medium, is being supplanted by the image. Civilization has taken, in Mitchell's phrase, a 'pictorial turn' (Mitchell 1994).

The visual image has been the subject of research across a range of academic disciplines. In recent years, the emphasis has moved away from theories of art towards a psychosocial perspective that draws on research from semiotics, branches of psychology, sociology and communications, and cultural theory (see *Semiotics*; *Mass Media: Introduction and Schools of Thought*; *Media Effects*; *British Cultural Studies*).

In part, this research has been driven by a long-standing debate over the social implications of the image. Often, this has centered on controversies that surround certain visual media; most prominently, racial and gender stereotyping, sexually-explicit material, and violence as depicted on television, film, and in computer games (see *Mass Media, Representations in*; *Pornography*; *Violence and Media*; *Media Effects*; *Computers and Society*). However, this debate has also been advanced as an opposition between the qualities and social effects of the word or text, and those of the image, with the former typically being privileged over the latter.

1.1 *History of the Debate*

The premise on which this debate turns is the superiority of text over image, a bias informed by an intellectual tradition that stretches back centuries (for a detailed discussion, see Stephens 1998). The Bible, in associating the Word with creation in Genesis, while forbidding the production of the 'graven image or any likeness', has long been used to support this presumption. Even as the image has established itself at the center of Western culture, writers and philosophers from Plato to contemporary thinkers have remained critical of the qualities and effects of the image compared with those of the word.

This intellectual antipathy is prompted by a series of overlapping concerns. Partly, it is founded on a long-standing suspicion of verisimilitude. In one view, such a capturing of likeness evokes a high degree of 'truth.' It is this perspective that typically reflects a traditional, although now contested, history of art as the progress of visual reproduction towards an ever more accurate reflection of the natural world, an ideal established by the Romans, and one that finds its most perfect expression in the photographic process.

However, verisimilitude has also driven a more negative perspective; that what the visual image offers is not truth but seductive illusion. A painter, Plato

wrote, can deceive 'children and fools' with the 'imitation of appearance', instead of bringing 'truth' or 'real things' (1956, p. 463). Substantively, this argument remains the root of one strain of criticism laid against television and movies, that their illusionary realism may offer an ideal vehicle for propaganda.

A second charge leveled against the visual image is that, with the growth of technical reproduction, images devalue the objects they represent. The uniqueness of the original is lost in the deluge of visual copies. This argument is perhaps most acute when what is being reproduced are images of cherished buildings, statues, and landmarks, and most especially, works of art. When the Mona Lisa becomes a common sight in the everyday world beyond the *Louvre*—for example, on T-shirts and advertising hoardings—the 'quality of its presence,' in the words of Benjamin is 'always depreciated' (1968, p. 221).

A third charge involves the long-held belief that the image speaks more directly to the emotions than to the logical mind. Aquinas argued that images 'excite the emotions more effectively aroused by things seen than by things heard' (cited in Freedberg 1989, p. 162). This rousing of emotions is seen as making the problems surrounding pornography and the visual depiction of violence all the more acute. Finally, the image's accessibility may distract the populace from the more cerebral merits of the word, be it the religious tract, improving essay, or newspaper text. It is a view that remains popular today, with some thinkers warning that the reach of words will shrink and that the era of the book is passing (see Nunberg 1996). Such a trend, however, has been difficult to substantiate, with sales of books in the US, for example, steadily increasing since 1998 after a slight fall in the mid-1990s (Schwartz 2000). Evidence for the decline of newspaper readership is more substantial, however. Daily circulations in the US fell yearly between 1987 and 1999, losing nearly 7 million in sales (Newspaper Association of America 2000).

2. *Visual Images and Child Literacy*

If there is unease that the image is replacing the word for society in general, this concern is heightened in debates over child literacy. At different times, anxiety has centered on the presumed ill-effects on children's reading abilities and habits of a series of visual media, including movies and comic books. However, most commonly criticized is television. Beentjes and van der Voort (1989) and Neuman (1995) are among researchers who have offered a number of hypotheses concerning the effects of television on child literacy, and weighed the empirical evidence for each.

A stimulation hypothesis argues that there is no conflict between images on television and words in print because watching television stimulates or improves reading in the young. This theory, although not widely held, rests largely on studies concerned with the reading of a book based on a television show after watching the program, or more generally, on children being prompted by television to research an interest. However, in a study of the juvenile book market, only 4 percent of children reported reading a book on the basis of seeing or hearing of it through television (Book Industry Study Group 1984). Similarly, the findings from other studies imply that any interests that are aroused by television tend to be short-lived (see Schramm 1961).

More systematically studied are variations of a reduction hypothesis, where the consumption of television is predicted to have a negative impact on reading. There are five common variations of this hypothesis. First, a displacement hypothesis argues that television viewing takes time away from reading. However, studies have shown no conclusive evidence of this, indicating instead that time given to reading, although always scant, has remained remarkably stable.

The passivity hypothesis argues that television viewing causes children to become mentally indolent and so less inclined to read. Although studies have shown that television does require less mental effort than reading, it is argued that synthesizing visual and audio information engages children in a process that is active, not passive. The retardation hypothesis, an associated theory that argues television viewing deteriorates the brain, is seen as implausible for the same reason.

A fourth variety of reduction hypothesis is concentration deterioration, which suggests that a child's ability to concentrate is weakened by television. This too, although popular among critics, has found little scientific support.

Finally, anti-school hypotheses argue that children come to expect school to be as entertaining and immediately gratifying as television programs like 'Sesame Street,' and when it fails to meet these expectations, pupils lose motivation (Cohen and Rudolph 1977, Trelease 1982). Once more, the evidence is inconclusive.

2.1 *Visual Images in Education*

In contrast to the attribution of negative effects to television viewing, some educationalists have called for the visual image to play a larger role in school curriculums which, traditionally, have been dominated by the word.

This move, in part, expresses a belief in the benefits of acquiring what is termed 'visual literacy.' The view is that irrespective of content, engagement with visual media may lead to the development of mental skills applicable beyond the understanding of the medium itself. To some extent, this argument is predicated on the widely influential work of Howard Gardner who, in developing a theory of 'multiple intelligences,'

argued that different modes of communication are beneficial to the development of different types of intelligence.

However, the theory that visual imagery may encourage both specific cognitive development and a distinctive worldview is more generally a reflection of similar beliefs regarding textual and verbal communication, as expressed, most notably, by linguists Edward Sapir and Benjamin Lee Whorf. The perspective may also be traced to theories associated with Marshall McLuhan, who links cognition to the technical characteristics of varying media.

In this view, it has been suggested that television-based teaching may have a positive effect on a number of areas of cognition, including the formation of a distinctive worldview, analytic reasoning, and spatial intelligence: the ability to envision mentally, for example, what a shape might look like from various angles. It is also suggested that exposure to certain visual media, such as advertisements that employ metaphors, may encourage and develop abstractive and analogical thinking.

However, it is argued that as images primarily function as analogs of reality and are so less inclined to divide the perceived world into particular categories in the manner characteristic of language, the development of a specific worldview through exposure to visual imagery appears unlikely. (This is not to say that the *uses* to which the visual image may be put, e.g., to depict violence on television, may not have a significant impact on one's worldview.)

Similarly, it is judged improbable that engagement with visual forms will enhance a viewer's skills in analytic thinking, since images lack many of the propositional elements of language (e.g., it is difficult to make statements in the negative through images, e.g., 'there is no ice-cream'). While a stronger case may be made for the connection between visual imagery and spatial intelligence, evidence for these benefits is also weak. Finally, it is argued that aspects of abstractive or analogical thinking may indeed profit from viewing certain visual material, although the argument remains controversial (see Messaris 1994 for a full discussion.)

More generally, there is a growing body of opinion that, rather than teaching new cognitive skills, the representational conventions of images are founded on informational cues that people learn routinely to interpret through their visual experiences in real-life. This calls into the question the very existence of a distinctive 'visual literacy' based on cognitive abilities acquired through exposure to mediated imagery.

3. *Properties of Visual Images*

Overall, there is scant evidence then, that visual images—most typically in the form of television—are either detrimental to child literacy, or significantly contribute to distinctive cognitive development.

However, it is clear that the image does possess distinctive qualities. Peirce's classification of signs into iconic, indexical, and symbolic (Peirce 1991) is seen as an important key to these properties. Images that bear a resemblance to some aspect of reality are examples of iconic signs. At the same time, the photograph may be defined as an index; a sign that has some physical connection to the object or event to which it refers, (such as fingerprints or footprints) since it is produced by the physical effect of reflected light on photographic emulsion.

These features clearly set the image apart from the word. Iconic signs have little role in language, and indexicality is entirely absent. Conversely, the visual image would normally not rely on the type of arbitrary conventions that link signifier (representation) with signified (what is represented) in the manner that constitutes the symbolic mode most characteristic of language.

Another crucial distinction can be made between words and visual images. Whereas verbal communication possesses detailed conventions governing word order (or syntax) that facilitate the linking of concepts, there is no such formal convention for combining images into larger meanings, although those who work with images, such as artists, may develop their own. Images exhibit what is termed 'syntactic indeterminacy' that increases their ambiguity as carriers of meaning (Messaris 1997). These distinct features have important consequences for the roles that the visual image plays within society. While it is difficult for images to convey precise statements about the objects they represent, (e.g., 'the man no longer enjoys sarcasm'), still the image's iconicity, together with less pronounced but flexible symbolic properties, have allowed the medium to convey information on physical characteristics with a precision and accessibility unmatched by words. For example, a range of scientific fields have been advanced through the use of detailed depictions and diagrams.

Equally, the visual image has been a boon in the development of user-friendly computers, where iconic representations, such as scissors and folders, have been used to replace lines of text. At the same time, photographic indexicality has become a vital component in the creation of historical records. However, these properties may also have troubling consequences.

3.1 *Visual Images and Persuasion*

Iconicity, indexicality, and syntactic indeterminacy bring distinctive properties to visual persuasion, be it set to commercial or political purposes. In an extensive study of these issues, Messaris (1997) offers a broad range of devices through which the advertising image may engage attention and mold attitudes. Iconic representations of physical appearance and interpersonal behavior may reproduce real-world visual

cues associated with a series of emotional responses; including sexual interest and excitement. This prompting of emotion is central to advertising strategies that seek to associate positive feelings with the possession or consumption of products (see, e.g., Williamson 1978). Further, iconic devices that manipulate the viewer's point of view (such as close-ups or high and low angles) can incite a series of other responses, including curiosity and trust.

The ability of iconicity to communicate information quickly and with relatively little mental effort may contribute to another advantage the image may have in persuasion. Some cognitive psychologists believe that visual images are typically processed peripherally—rather than centrally—in ways that may decrease the likelihood of viewers engaging their full critical and analytical faculties, especially when the speed of the message leaves little time for reflection (see Petty and Cacioppo 1986).

The 'truthfulness' implied by indexicality is also a critical ingredient in visual persuasion whenever a photographic image is offered as proof of an advertiser's claims. However, visual images may readily be made to lie through the application of a wide range of techniques, from the staging of photographs to airbrushing to improve appearances. Moreover, digital technology has become a powerful aid to those who would practice such visual deception.

Finally, the syntactic indeterminacy of the image may also have important consequences that make it more useful in persuasion than the explicit meanings offered by words. For example, advertisers may take advantage of this property to imply associations or statements that they would be unwise to state verbally (such as the association of cigarette smoking with healthy activity).

3.2 *Visual Images and Politics*

Television has been accused of having a number of detrimental consequences for the political process. These include, in a political equivalent to the displacement theory associated with child literacy (see above), distracting the populace from civic involvement. However, the area of most concern with regard to the visual image is its effect within political advertising and televisual campaigning.

Political advertisements use many of the same visual strategies as their commercial counterparts, although often exhibiting a strong reliance on iconic visual effects. Patriotic appeals are common in the political advertisement or campaign, and the use of certain visual symbols—including the national flag—has been found to be especially prevalent. Equally, the use of family imagery, such as images of spouses and of children, remains common.

The iconicity of visual imagery has been held responsible for other political strategies that accentuate appearance over content. These make extensive use of devices that appear to be modeled on people's real-world experiences of interpersonal space (see Meyrowitz 1986). For example, politicians interviewed on television will typically turn towards the camera, and therefore towards the viewer, in order to imitate the real-world appearance of a direct and candid approach.

The potency of the visual image in political persuasion is thought to have been most graphically displayed through a debate between Richard Nixon and John F. Kennedy. It is reported that the majority of those who saw the 1960 debate on television judged Kennedy the victor, while most who listened on radio thought Nixon had won. Typically, this result has been taken to indicate that the more attractive image that Kennedy presented was the deciding factor. However, there may be a number of reasons for such an outcome, including the possibility that it was radio that may have favored Nixon's patterns of delivery, while exaggerating Kennedy's Boston accent in a manner that repelled listeners (Schudson 1995). This and other uncertainties surrounding the visual effects of television has led some thinkers to argue that the persuasive power of the medium has been overstated.

4. *Visual Images and the News*

Equally controversial has been the influence of the visual image on news. The entry of the image into an arena that traditionally had been the province of the word was viewed with consternation, both by journalists and by many outside of the profession, as the use of visual material developed from artwork through wood-cut engraving to early photographs. Visual images continued to be regarded as a medium that bypassed the reasoning mind and, as such, appeared to threaten journalism's Enlightenment heritage. While the photograph was held in higher esteem in the picture magazine, this had little impact on its critics.

Yet at the same time, the photographic record, as metaphor as well as in practice, came to underpin the journalistic enterprise. By the mid-nineteenth century, there was widespread typification of the newspaper as a 'daguerreotype' of the social and natural world, with the indexicality of the process being used to exemplify, and give credence to, claims of journalistic objectivity (see Zelizer 1998). More recently, however, scholars have pointed to the use of photographs, in both print and broadcast news, as being a practice based on ideological values, as well to the connotative properties inherent in the practice itself. Barthes (1977) has argued that the photograph fulfills not only a denotative (or referential) but also a connotative (or interpretive) function, which undermines its claim to be unmediated truth. Moreover, by embedding the photograph within the narratives characteristic of

news, it is argued, its meaning is colored by that association. In this way, the news-photograph, be it on the page or in broadcast film, may be far from being either 'neutral' or 'objective.'

5. *Visual Images and the Word: Future Prospects*

Much consideration has been given here to the way the cultural impact of the visual image has been traditionally expressed through a rhetorical opposition favoring the 'rationality' of word over visual image. Yet many of the presumed negative effects of the image have little basis in scientific fact, while critical linguists and others have shown that the multiple levels of implicit meaning generated by text can make the word equally suspect.

Moreover, in an era when media has long been multimodal, it is increasingly rare to encounter the visual image separate from the word. Consequentially, discussion has been moving beyond the opposition of the two modes to a consideration of how, when conjoined, words and visual image may work to fix, reinforce and subvert the meanings of one another, as well as to generate new meanings (see *Critical Inquiry* Spring 1980, special issue: Language and Image). Critics engaged in contesting photojournalism's appeal to the 'objectivity' of indexicality have long argued that photographs attain meaning only in relation to the settings in which they are encountered, which invariably include text. Similarly, those studying broadcast news have emphasized that viewing typically involves the simultaneous processing of words and images, with the voice-over commentary representing a particularly interactive relationship between the two modes. This interaction may be utilized to 'anchor' verbal discourse through a visual image, or vica versa, in ways that encourage viewers towards a particular reading of a depicted event (see, e.g., Hartley 1982).

The study of word and image combinations and their consequences has been given new emphasis with the advent of computer technology and the Internet. The ability to digitally combine and manipulate the two modes—in ways that may affect the credibility of both—means that many of the traditional arguments over word and images may soon be relegated to history. As Lester (2000, p. 306) puts it, the next decade promises to be 'known for the way that words and pictures are used together as equal partners in the communication process,' a new Information Age in which 'all forms of communication are included through the World Wide Web.'

See also: Advertising, Psychology of; Communication: Electronic Networks and Publications; Imagery versus Propositional Reasoning; Internet: Psychological Perspectives; Mass Communication: Normative Frameworks; Mass Communication: Technology; Media Effects; Media Effects on Children; Media Literacy; Media, Uses of; Symbolism (Religious) and Icon; Television: General; Television: Genres; Visual Imagery, Neural Basis of; Visual Memory, Psychology of; Visual Methods in the Social Sciences; Visualization: Representation in Science

Bibliography

Barthes R 1977 *Image—Music—Text*. Fontana, London

Beentjes J, van der Voort T 1989 Television and young people's reading behaviour: A Review of research. *European Journal of Communication* **4**: 451–77

Benjamin W 1968 *The Work of Art in the Age of Mechanical Reproduction Illuminations* (English translation). Schocken Books, New York

Book Industry Study Group 1984 *Reading and book purchasing: Focus on juveniles*. Author, New York

Cohen D H, Rudolph M 1977 *Kindergarten and Early Schooling*. Prentice-Hall, Englewood Cliffs, NJ

Freedberg D 1989 *The Power of Images: Studies in the History and Theory of Response*. University of Chicago Press, Chicago

Hartley J 1982 *Understanding News*. Methuen, London

Lester P L 2000 *Visual Communication: Images with Messages* 2nd edn.. Wadsworth Thomas, Belmont, CA

Messaris P 1994 *Does TV Belong in the Classroom: Cognitive Consequences of Visual 'Literacy,' Communication Yearbook 17*. Sage, Thousand Oaks, CA

Messaris P 1997 *Visual Persuasion*. Sage, Thousand Oaks, CA

Meyrowitz J 1986 Television and Interpersonal Behavior: Codes of Perception and Response. In: Gumpert G, Cathcart R (eds.) *Inter/Media: Interpersonal Communication in a Media World*, 3rd edn. Oxford University Press, New York, pp. 253–72

Mitchell W J T 1994 *Picture Theory: Essays on Verbal and Visual Representation*. University of Chicago Press, Chicago, IL

Neuman S B 1995 *Literacy in the Television Age*, 2nd edn. Ablex, Norwood, NJ

Newspaper Association of America 2000 *Facts about Newspapers 2000*. Editor & Publisher and NAA [http://www.naa.org]

Nunberg G (ed.) 1996 *The Future of the Book*. University of California Press, Berkeley, CA

Petty R E, Cacioppo J T 1986 *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer-Verlag, New York

Peirce C S 1991 In: Hoopes J (ed.) *Peirce on Signs: Writings on Semiotics by Charles Sanders Peirce*. University of North Carolina Press, Chapel Hill, NC

Plato 1956 *Great Dialogues of Plato*. New American Library, New York

Schramm W L 1961 *Television in the Lives of our Children*. Stanford University Press, Stanford, CA

Schudson M 1995 *The Power of News*. Harvard University Press, Cambridge, MA

Schwartz E G 2000 *The 1999 Consumer Research Study on Book Purchasing*. NPD Group, Inc., for the American Booksellers Association and the Book Industry Study Group, New York

Stephens M 1998 *The Rise of the Image, the Fall of the Word*. Oxford University Press, New York

Trelease J 1982 *The Read-aloud Handbook*. Penguin, New York

Williamson J 1978 *Decoding Advertisements: Ideology and Meaning in Advertising*. Calder & Boyars, London

Zelizer B 1998 *Remembering to Forget: Holocaust Memory Through The Camera's Eye*. University of Chicago Press, Chicago

J. Huxford

Visual Memory, Psychology of

Human memory is not a single mental faculty or cognitive system; the cognitive processes which we collectively call memory are composed of a number of independent and specialized cognitive systems that encode and store information in different formats. Some of these memory systems express themselves directly in the form of conscious recollection of general knowledge and previous personal experiences, other systems express themselves indirectly in behavior. The concept of visual memory has two different, yet related, meanings.

1. *Visual Memory as the Channel of Communication*

In the most general textbook usage, visual memory refers to the storage of information transmitted through the visual system. According to this definition, it is a form of memory which may be compared with memory for information transmitted through other sensory systems such as auditory, tactile, or olfactory memory. The type of information that is processed (colors, music, smells) and which cannot be translated between modalities captures important characteristics of modality-specific memories. Differences between modalities in the organization of information in memory are also found. For example, when subjects are requested to remember a series of items presented in succession, in auditory presentation memory is better for the first and last items than the middle items (primacy and recency effects), whereas in visual presentation only the former effect is observed. While the basis for such modality effects is not known, they probably reflect differences in the way information is processed by the perceptual systems rather than characteristics of specific auditory and visual memory systems as such.

The channel-of-communication definition of visual memory misses the point that vision is a gate to many memory systems, and conversely that information in different sensory channels may be processed by the same amodal memory system.

2. *Visual Memory as a Code*

In a more restricted usage, visual memory refers to a code by which visual perceptions are inscribed and retrieved from memory in the form of images or sequences of images. According to this definition, visual memory is not a separate memory system but a representation of information that is processed by different memory systems. This definition also acknowledges that the channel of communication does not specify the memory format; what is viewed by the eye may be stored amodally, e.g., in memory systems relying on verbal codes.

2.1 *Object Recognition*

A direct expression of visual memory is the perceptual process itself, in the recognition of colors, textures, shapes, and objects in the world. This is achieved by some process by which on-line information from the visual system is matched to stored descriptions. Perceptual recognition is immediate, representing a fast, automatic and implicit (nonconscious) process (see *Implicit Memory*, *Cognitive Psychology of*), and it precedes semantic and episodic memory retrieval which deal with the identification of particular, familiar objects. The vast collection of memories that allow perceptual recognition is at least partly associated with a separate memory system, called the perceptual representation system (PRS) (Tulving and Schacter 1990). The PRS was formally proposed as a separate memory system in 1990, and is currently under intense investigation. It is believed to consist of several subsystems, one of which is a structural description system which stores descriptions of coherent or meaningful objects. A more recent theoretical proposal identifies a series of specialized memory processes at an earlier and more primitive level of representation, which stores information about elementary attributes of the visual world, such as the orientation, color, motion, and texture of visual patterns (Magnussen and Greenlee 1999). The function of this latter form of visual memory is not known but it might store information across fairly short time spans to help building permanent object representations.

2.2 *Visual Representation in Episodic and Semantic Memory*

Visual recognition goes beyond the recognition of classes of objects. The picture in the morning paper is not just a sidewalk café, it is the Deux Magots in Paris, and the person in the front row is not just a female figure, it is the movie star Catherine Deneuve. The conscious recognition of particulars is a visual expression of semantic memory—a memory system that stores information about general facts. And if you are a Parisian or happened to be a lucky tourist, you might also remember seeing Catherine Deneuve walking by as you were sitting outside the very same café, and you might remember her looks at that particular day. This production of a memory image to the 'inner eye' is a visual representation of episodic memory, a

memory system that stores information about personal facts and episodes.

Semantic and episodic memory represent declarative or explicit forms of memory. Both personal observation and experimental research testify to the human capacity for visual episodic memory. In a series of classical studies in cognitive psychology (Standing 1973) subjects viewed large numbers of black-and-white and color slides of natural scenes, objects, persons, interiors, and landscapes, and memory was later probed by presenting pairs of slides containing one previously seen and one novel picture and asking the subjects to pick from each pair the picture they remembered. Human subjects are amazingly good at this task. With a learning set of 2,500 pictures the score was on average 90% correct after a 36 h memory interval, and with a learning set of 10,000 pictures the score was 84% correct in the test trials. This easily outperforms the memory for words tested with similar methods (see *Recognition Memory, Psychology of*).

The excellent recollection of pictures and visual scenes has been explained in terms of the dual coding theory, which distinguishes between visual and verbal memory codes (Paivio 1995). Picture memory has an advantage because pictures are effectively coded in both memory systems whereas words are coded primarily, but not exclusively, in verbal systems. The effectiveness of dual coding is illustrated by memory-enhancing techniques, e.g., in the 'method of loci' where items to be remembered are attached to places in a visualized familiar environment and later retrieved by mentally traveling the same path in memory (see *Mnemonics, Psychology of*).

The distinction between visual and verbal coding systems is also central in theory of working memory. The working memory is a system for maintaining new information or information retrieved from long-term memory, and this maintenance of information is performed by two so-called slave systems, one of which is the 'visual–spatial sketch pad' a hypothetical rehearsal mechanism that keeps refreshing visual and spatial perceptual information (see *Working Memory, Psychology of*).

3. *Are Visual Codes Necessary?*

The necessity for postulating visual codes has been questioned by cognitive psychologists who favor a unitary memory system based on propositional, language-like codes. As an illustration, studies of memory expertise have shown that the memory of chess masters is vastly superior to that of ordinary players for real game positions, but not for random positions of chess pieces on the board (Chase and Simon 1973). The expertise memory is explained by the fact that masters have a long-term storage of a large number of visual 'chunks,' where a chunk is a constellation of actual game positions involving a few pieces, and the estimated number chunks is about 50,000, which is comparable to the number of words commanded by persons with a college education. Obviously, such chunks may be effectively represented either in terms of image codes or in terms of propositions—or in both (see *Mental Imagery, Psychology of*; *Expert Memory, Psychology of*).

The arguments for the existence of visual codes in memory may be summarized under four headings, as follows.

3.1 *Subhuman Primates Have Detailed Visual Memories*

Experiments with monkeys have shown that these animals can learn to discriminate between large numbers of complex abstract fractal or snow-flake type visual patterns, store these in long-term memory, and utilize the information in single-trial short-term memory tasks. Few, if any scientists have claimed elaborate propositional representation in monkeys.

3.2 *High-precision Storage of Sensory Information*

Experiments carried out in the 1990s have shown that sensory information, i.e., texture, orientation, motion, and similar attributes of the visual image, is extremely well retained in short-term memory. In fact, attributes such as motion and spatial frequency are represented in memory with the fidelity of the sensory image, at least for memory intervals of the order of minutes, but there is also some evidence for long-term storage of detailed sensory information. In one famous experiment, an eidetic subject was able to identify the three-dimensional figure hidden in random-dot stereograms when the left and right-eye patterns were viewed on different days (Stromeyer and Psotka 1970). Such high-fidelity images cannot reasonably be stored in terms of propositions.

Long-term memory changes in the form of perceptual learning are also well established. Subjects get better at performing fine discriminations between orientations or textures across large number of training sessions. The learning is specific for the training conditions and does not transfer to other conditions; for example, when subjects are trained in discriminating visual textures in one part of the visual field, the training effect does not transfer to other parts of the visual field (Karni and Sagi 1991). Some forms of visual memory must be due to local neural changes closely associated with the perceptual process itself.

3.3 *Dissociation Between Visual Memory and Other Forms of Memory*

Lesions of the brain which disturb semantic and episodic memory may leave the perceptual recognition

process intact. The patient's perception of the world is normal but he or she does not remember seeing these objects or persons before. Recognition memory may be demonstrated though perceptual priming (see *Priming*, *Cognitive Psychology of*): If the patient is shown a drawing of an object and at a later time is shown the same drawing again, there is a gain in the time required to identify and name drawings that were previously seen compared with novel drawings, and the gain is comparable for amnesic and normal subjects.

If, on the other hand, the link between visual memory representations and on-line perception is disrupted, amodal (verbal) episodic and semantic memory may be intact but the person can no longer recognize what he or she sees. In this condition of *visual agnosia* there may be nothing wrong with perceptual process *per se*, the person navigates safely in the geographic environment and may even copy diagrams and drawings, but the information which comes from vision is meaningless.

3.4 Visual Processing Areas of the Brain Are Recruited in Memory Tasks

Studies of memory processes in the human brain with modern brain imaging technology such as positron emission tomography (PET) and functional magnetic resonance tomography (fMRI) have shown that there is no special region of the brain where memories are stored. Memory involves widely distributed networks and all major regions of the cerebral cortex have been implicated in the various forms of memory. Since visual memory is a form of representation, a code rather than a system, one would expect it to share processing space with several memory systems, in addition to regions of the occipital lobe that are specific for the processing of visual signals.

Memory researchers now generally agree that memories may involve many of the neural ensembles that were involved in the processing of the information in the first place. In vision a close link between perception and memory is illustrated by the observation that certain forms of perceptual deficits due to local brain lesions appear to affect memory as well. For example, persons who lose their ability to see colors owing to bilateral lesions of a region of the occipital lobe known as V4 usually also lose their ability to remember or imagine colors.

The current debate on the neural mechanisms of visual memory does not consider whether sensory processing areas are recruited in memory, but how far down the processing hierarchy recruitment extends. Some brain imaging studies claim that the primary visual cortex (region V1), i.e., the very earliest level of sensory analysis receiving direct input from the eye, is activated in visual imagery tasks, but other studies do not find this activity (Cabeza and Nyberg 1997).

A logical argument may be directed against the involvement of early visual processing areas and a unification of visual perception and memory processes: Since memory images are normally much less vivid and detailed than perceptual images, and memory is rarely confused with perception, they must activate at least partly different neural circuits. The precise relationship between visual perception and visual memory is a major task for future research.

See also: Encoding Specificity in Memory; Memory Psychophysics; Working Memory, Psychology of

Bibliography

Cabeza R, Nyberg L 1997 Imaging cognition: an empirical review of PET studies with normal subjects. *Journal of Cognitive Neuroscience* **9**: 1–26

Chase W G, Simon H A 1973 The mind's eye in chess. In: Chase W G (ed.) *Visual Information Processing*. Academic Press, New York, pp. 215–81

Farah M 1990 *Visual Agnosia*. MIT Press, Cambridge, MA

Karni A, Sagi D 1991 Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the USA* **88**: 4966–70

Magnussen S, Greenlee M 1999 The psychophysics of visual memory. *Psychological Research* **62**: 81–92

Paivio A 1995 Imagery and memory. In: Gazzaniga M S (ed.) *The Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 977–86

Standing L 1973 Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology* **5**: 207–22

Stromeyer C F, Psotka J 1970 The detailed texture of eidetic images. *Nature* **225**: 346–49

Squire L R, Kandel E R 1999 *Memory. From Mind to Molecules*. W. H. Freeman, New York

Tulving E, Schacter D L 1990 Priming and human memory systems. *Science* **247**: 301–6

Zeki S 1993 *A Vision of the Brain*. Blackwell, Oxford

S. Magnussen

Visual Methods in the Social Sciences

Visual methods, relative to other methods, are underdeveloped in the social sciences. The probable explanation is that social science reasoning generally relies on the analysis of aggregate data, and photography and other visual methods locate their attention on the particular. Thus, it requires a reversal of typical social science reasoning to generalize from the particular, especially when the particular is a visual image.

The nature of visual evidence is also problematical for social science. Photographic (film or video) images

appear to represent the world 'objectively,' that is, without interpretation. At the same time, it is recognized that photographs are socially and technically constructed, thus embodying points of view and interpretive influences. The tension between the competing qualities of objectivity and subjectivity influence the use of visual images in the social sciences.

1. Forms of Visual Methods

1.1 Empirical Evidence

Photographs (or photos linked together in film or video) are the most common form of visual sociology. The most common use of visual images is to catalog aspects of material culture or social interaction (see Collier 1967), which was a common aspect of articles published in the first issues of the *American Journal of Sociology* (see Stassz 1979). In the first 20 volumes of the *American Journal of Sociology*, until 1916, sociologists used photographs to make arguments about the character of rural and urban life; the working conditions in offices and factories, and living conditions in industrializing cities. As sociology sought a more scientific posture, the visual research method (which was not well developed in these early examples) was largely abandoned.

Anthropologists, working in the first decades of the twentieth century, also used photography extensively (Edwards 1992). Anthropology's primary mission during its early eras was to document and classify human groups. Photography's ability to gather a large amount of information in a single document made it an important data-gathering method. As anthropology abandoned paradigms based on the taxonomic organization of racial classifications, photography became less important as a research method.

The full development of photography as part of ethnographic analysis is generally attributed to Gregory Bateson and Margaret Mead's (1942) study of Balinese culture. In this project, the researchers used 759 photographs (selected from more than 25,000 made during their fieldwork) to support and develop ethnographic observations and analysis. Bateson and Mead used sequences of photographs to show how the Balinese performed social rituals or engaged in routine behavior. These images typically record the unfolding of events, and thus are similar to short movie clips. The researchers also used images to survey material culture such as dwellings and agricultural techniques. This promising expansion and redefinition of visual anthropology has been followed by contemporary examples such as Danforth and Tsiaras' study of death rituals of Greek villagers (1982), and several other examples in sociology and anthropology.

Howard Becker was the first modern sociologist to suggest a link between photography and sociology (1974). Much of Backer's argument referred to the ability of photographs to record the world as observed. For example, Becker noted that photography and sociology, since their inception, had both been concerned with the critical examination of society. Becker noted that the early twentieth century photographers, such as Lewis Hine and Jacob Riis, were concerned with topics such as urbanization and industrialization, the exploitation of labor, and the conditions of immigrants, all of central concern to sociologists. Becker suggests that a modern visual sociology should emerge from documentary practice engaged with a sociological theory, and the sociological use of visual data developed in the context of common concerns with validity, reliability, and sampling. Becker's article, reprinted in several anthologies, is credited with the rebirth of visual sociology in the US.

In the meantime, sociologists have used photographs as evidence in the study of social change (Rieger 1996), in sociological classification of farmsteads (Harper 1997), and other topics that reflect conventional sociological research. Visual ethnography, however, remains a relatively small movement within the field methods tradition in sociology and anthropology.

1.2 Photos as Cultural Narratives

Anthropologists have used film and video (and, to a lesser extent, the 'photo essay') to communicate typical events of a cultural scene. This tradition is largely attributed to Robert Flaherty's seminal ethnographic film, *Nanook of the North*, completed during the 1920s (see Rotha 1980). The resulting film tradition in anthropology has both embraced and challenged the capacity of visual narrative format to communicate cultural information. Nevertheless, the use of film (and photography) to tell 'cultural stories' remains a viable part of visual anthropology, and constitutes the single most influential demonstration of visual methods in social research. In anthropology, ethnographic films are presented at professional meetings, scholarly film festivals, and as a part of normal pedagogy.

1.3 Photographs in the Research Process

Collier (1967) first described how researchers may employ photographs to elicit cultural information. The process, now referred to as 'photo-elicitation,' uses images to stimulate culturally relevant reflections in interviews. Researchers may take the photos prior to the interview, or the photographs may be from a personal collection or professional archive. Photo-elicitation has become a common method in visual sociology, used in a wide range of studies, including public responses to social problems (Faccioli and Zuccheri 1998), cultural definitions of an urban neighborhood (van der Does et al. 1992), the 'working knowledge' of a rural artisan (Harper 1987), and

ethnic recognition among Asian immigrants (Gold 1991). The promise of photo-elicitation lies in the understanding of how photographs communicate different meanings to the researcher and the subjects; thus the images become a basis of interviewing that explores the taken-for-granted assumptions of both researcher and subject.

Social science film-makers have integrated imagery into research in a different manner. Beginning in the early 1970s, anthropologists have taught cultural informants to use film and video cameras in order to make films that tell cultural stories from the vantage point of the informant. The first example of this method was Sol Worth and John Adair's research among the Navajo, which resulted in seven films made by native informants (described in Worth and Adair 1972). With the advent of easy to use video cameras, native-produced visual ethnography has become more common. While controversy continues to surround the redefinition of theoretical authority that native-produced films challenge, most visually oriented social scientists find the native-produced film or video a viable source of ethnographic information.

2. *Problems in Visual Methods*

2.1 *Ethics*

There are at least three aspects to the question of ethics and visual methods in research. The first concerns the matter of anonymity; the second concerns the matter of informed consent; and the third represents the burden of 'telling the visual truth.' These three issues are inter-related.

Because social science is generally concerned with the patterns of social behavior, individuals are represented by words, or more commonly, by responses they have made to a questionnaire. Thus the identity of individuals are protected. The protection of the privacy of subjects is, indeed, codified in the ethical canons of the social science disciplines. In the case of visual research, however, it is often impossible to protect the privacy of subjects. All discussion of ethics and visual methods rests on this fundamental fact; the photograph produces a likeness which is often identifiable.

Visual researchers thus face a more pronounced problem of informed consent. Subjects must be made aware of implications of publishing their likeness in social science research, and agreement to participate should be made in writing and prior to the research.

The matter of ethics in visual research is unique among research methods for one final reason. Because it is often possible to identify subjects in research images, visual researchers must be absolutely certain that these images 'tell the truth.' Visually oriented social scientists recognize that it is unethical to alter images to change their meaning, or to present an image that connotes a reality that was not consistent with the understood reality of a situation. Most visual researchers feel that the ethical considerations of visual research most separates it from that of their peers.

3. *Technology*

Like all social science methods, visual research depends upon technology. Like all technologies, visual technology makes it possible to record some types of information, but not others. Steiger (1995) detailed how specific cameras, lenses, film types, and camera settings create different kinds of filmic statements, which, in turn, make possible different types of sociological thinking. The specific connection between 'technologies of seeing' (whether relating to photography, film, video, or virtual reality) and social ideas remains an area of enormous potential and scant attention. Steiger's article raises the important point that to be a skilled researcher using visual methods is to master cameras, video and film cameras, and possibly photo printing. Now that many of these technologies have an electronic dimension, skilled visual researchers also need to become adept at software that catalogs visual images, that allows for the improved presentation of images, and that facilitates the production and distribution of images.

New methodologies, such as computer-based multimedia, the World Wide Web, CD publication, and virtual reality, all offer possibilities for the development of innovative visual methods. Each offers an approach for recording, analyzing, and participating in the visual dimensions of society. The future of visual methods in social research will depend, to a large extent, on how well these new technologies are integrated into social science methods and thinking.

4. *Publication: Organizational Basis*

In the past two decades, visual research in anthropology and sociology has been institutionalized in intellectual organizations such as the Society for Visual Anthropology, a part of the American Anthropological Association, and the International Visual Sociology Association, an independent scholarly organization with ties to several academic disciplines. Publication of visual methods requires special consideration, and the journals such as *Visual Anthropology* and *Visual Sociology* feature the presentation of images. The full development of visual methods in social research will require that the mainstream journals and books accept the publication of images as a normal part of social science research, no different than the publication of tables or graphs.

Bibliography

Bateson G, Mead M 1942 *Balinese Character: A Photographic Analysis*. New York Academy of Sciences, New York

Becker H S 1974 Photography and sociology. *Studies in the Anthropology of Visual Communication* **1**(1): 3–26
Collier J Jr. 1967 *Visual Anthropology: Photography as a Research Method*. Holt, Rinehart and Winston, New York
Danforth L M, Tsiaras A 1982 *Death Rituals of Rural Greece*. Princeton University Press, Princeton, NJ
Edwards E 1992 *Anthropology and Photography 1860–1920*. Yale University Press, New Haven, CT
Faccioli P, Zuccheri N 1998 The double vision of alcohol. *Visual Sociology* **13**(2): 75–90
Gold S 1991 Ethnic boundaries and ethnic entrepreneurship: a photo-elicitation study. *Visual Sociology* **6**(2): 9–22
Harper D 1987 *Working Knowledge: Skill and Community in a Small Shop*. University of Chicago Press, Chicago
Harper D 1997 Visualizing structure: reading surfaces of social life. *Qualitative Sociology* **20**(1): 57–77
Rieger J 1996 Photographing social change. *Visual Sociology* **11**(1): 5–49
Rotha P 1980 Nanook and the north. *Studies in Visual Communication* **6**(2): 33–60
Stassz C 1979 The early history of visual sociology. In: Wagner J (ed.) *Images of Information: Still Photography in the Social Sciences*. Sage Press, Beverly Hills, CA
Steiger R 1995 First children and family dynamics. *Visual Sociology* **10**(1–2): 28–49
Worth S, Adair J 1972 *Through Navajo Eyes*. Indiana University Press, Bloomington, IN
van der Does, Edelaar S, Gooskens I, Liefting M, van Mierlo M 1992 Reading images: A study of a Dutch neighborhood. *Visual Sociology* **7**(1): 4–68

D. Harper

Visual Perception, Neural Basis of

Visual perception requires that patterns of light energy are encoded into neural signals in the retina of the eye, and that these signals are transformed as they pass through the neural network of the retina and are transmitted to a series of interconnected brain areas. These transformations create patterns of neural activity that can represent the visual information needed for recognizing objects, people, and events; for location in the environment; and for guiding actions.

1. *Structure of the Visual Pathway*

Encoding begins in the photoreceptor cells (rods and cones) in the retina, which signal the local light intensity falling on each receptor by their levels of electrical polarization. A network of horizontal, bipolar, and amacrine cells connect these receptors to the retinal ganglion cells. The axons of these cells form the optic nerve, and so the signals that they carry constitute the information that the eye sends to the brain; the nature of this information is discussed in Sect. 3. At the optic chiasm, these axons are routed so that signals from the two eyes can be transmitted to each side of the brain, making binocular integration at a higher level possible.

The lateral geniculate nucleus (LGN) of the thalamus provides a relay station where visual signals can be modulated, but the next major reorganization occurs when the signals arrive in the occipital lobe of the cerebral cortex (see *Cerebral Cortex: Organization and Function*). The primary receiving area, at the posterior pole of the brain, is known as striate cortex, or cortical area V1. It is surrounded by a series of extra-striate visual areas. These areas can be distinguished both by specializations of function (see Sect. 2.4) and by the fact that many of them form distinct, ordered maps of the visual field. Area V2 receives much of the input from V1, and sends information on to areas V3, V4, and V5. Pathways through these areas send information to the temporal and parietal lobes of the brain.

Within each cortical area, most connections between neurons are quite local. However, information is passed between cortical areas in fiber tracts, in an orderly way. Specific layers of the cortex receive input from areas at a lower hierarchical level, while other layers either transmit to higher levels, or receive feedback from higher levels. Thus, cortical areas are commonly thought of as a set of modules, each performing particular visual computations and passing the results of these computations to other modules.

A minority of fibers in the optic nerve transmit information not to the cerebral cortex, but to midbrain structures, in particular the superior colliculus which also incorporates a topographic map of the visual field. In lower vertebrates, the homologous structures provide the main central visual pathway, but in mammals, certainly in primates, this pathway serves the function of orienting to stimuli by eye movements rather than any complex analysis of the visual image, and probably does not contribute directly to conscious visual perception. However, there are connections running both ways between this subcortical visual pathway and the much more elaborate cortical network for visual information processing.

2. *Key Concepts of Neural Representation and their Evolution*

2.1 *Specific Nerve Energies and Place Coding*

In developing modern concepts of the neural basis of vision, an early key idea was that messages from the eye to the brain could convey distinct information according to which element of the nervous system carried them. Thomas Young, around 1800, proposed that different colors were represented according to the degree that three forms of 'sensitive filament' in the eye were stimulated, a view that in its essentials is accepted today. Developments of this idea in the nineteenth century were that each separate fiber in the optic nerve

is associated with a 'local sign' (*Lotze*)—that is, we can know which point in space stimulated our eyes from which particular fiber is active—and the 'doctrine of specific nerve energies' (*Müller*) that the quality of stimulation (e.g., sound versus light) is conveyed by which fibers convey a signal, not by any distinct quality in the signals themselves. This principle (sometimes called 'place coding' or 'labeled line') is now recognized as fundamental for understanding how visual information is represented within the brain, as well as in the sensory pathways leading to it.

2.2 Univariant Coding by Impulse Frequency

The work of Adrian in the 1920s established that sensory information is carried by a series of nerve impulses (action potentials). Since the impulses are identical all-or-none events, quantitative variations, e.g., in light contrast, have to be encoded by the rate at which impulses occur. In principle, more complex statistical properties of the train of impulses could carry rich information, but visual neurophysiology has not yet yielded convincing examples of this. Thus, the activity of a single neuron is univariant—it can encode only a single variable in the visual image. The classic example is in color vision: a particular receptor may respond relatively more strongly to green light than to red light. However, because its response is univariant, it can give indistinguishable responses to a dim green light and a bright red light. Thus, to represent even a single variable, within a multidimensional domain such as color, requires information to be combined or compared from two or more neural signals.

2.3 Receptive Fields and Tuned Filters

These principles of neural coding have been particularly applied in considering how the visual system represents spatial information. A concept which developed early was that each visual neuron had a receptive field—a region of the field of view within which light could influence the activity of the neuron. Further work showed that receptive fields had internal structure, such that a particular spatial pattern of light and dark within this region was optimal for activating a given neuron. This led to the idea that neurons could be 'feature detectors,' signaling the presence of a specific pattern element, such as a small black moving spot that might be a fly in the visual field of a frog. However, in the mammalian brain at least, it currently seems unlikely that any neuron on its own can provide such a signal; in all cases that have been investigated, the distribution of activity across some population of neurons would be needed to determine reliably the presence of a particular feature or object. This idea of population coding should not obscure the fact that, in visual brain areas, coding is sparse. That is, for any particular visual stimulus, only a small fraction of the available neurons are active, and so it is appropriate to think of particular subsets of neurons as being involved in representing particular visual attributes.

Neurophysiological and psychophysical studies in the 1960s and 1970s concentrated on how populations of neurons in V1 encoded particular visual properties, such as contour orientation. Each neuron was found to be a 'tuned filter,' which responded to a limited range of orientations. This makes it possible (a) to derive orientation information from the peak of activation across a distribution of differently tuned filters; and (b) to segregate visual information according to the local orientation, which may be valuable for segmenting and grouping the scene into distinct objects and surfaces. Cortical neurons also act as arrays of tuned filters for other properties, such as direction of motion and color.

2.4 The Hierarchy of Multiple Visual Areas

The finding, that neurons in area V1 were tuned to a range of stimulus properties, led to exploration of higher levels of cortical processing. Experiments with macaques, from the mid-1970s onwards, showed that there is a multiplicity of distinct visual areas beyond V1. As well as progressive transformations through a hierarchy of visual areas, there was found to be parallel processing by areas such as V4 and V5, which receive information from V1 in parallel via separate routes. These areas are specialized, so that, for example, V5 contains a high proportion of neurons tuned to specific directions of motion but few that are wavelength-tuned, while V4 shows the opposite balance. This has led to a picture of functional segregation, within the cortex, of representations which emphasize different aspects of visual information and which presumably serve distinct purposes.

It was recognized that this picture of functional segregation was consistent with findings in human patients who had localized brain damage; for instance, such damage could abolish color discrimination with little effect on other aspects of visual perception. In the 1990s new methods of functional brain imaging made it possible to explore localization of function in the normal human brain (see *Functional Brain Imaging*). These have confirmed the multiplicity of visual areas, each responding to particular stimulus properties and activated in different visual tasks. The fact that brain activation depends on the particular task the subject is set, and on his/her attention to the stimulus, indicates that vision cannot be thought of as a purely 'bottom-up' neural process, in which information originating in the receptors is progressively transformed. Rather, the results of high-level processes can modulate how low-level processes operate. An anatomical basis for this is provided by 'descending' connections from higher to

lower visual areas, which are at least as rich and complex as those transmitting information from lower to higher processing.

These concepts of neural information processing inform current research on all aspects of vision. In this article they are illustrated primarily with respect to the representation of pattern, shape, and spatial layout. They also underlie current understanding of other perceptual functions which, however, will not be developed here (see *Motion Perception: Psychological and Neural Aspects*; *Color Vision*).

3. Representation of Spatial Pattern and Shape

3.1 Receptive Field Structure and Specificity

The richness of vision as a source of information comes largely through the detailed pattern of light and dark that arises from the spatial structure of the environment. Encoding of this pattern begins in the retina. Receptive fields of retinal bipolar and ganglion cells have a center–surround opponent organization. In an 'on-center' cell, light in the receptive-field center activates the cell, but this is countered by inhibition from any light in the surround. Off-center cells show the opposite relationship. In either case, the activity of the cell signals the presence of spatial contrast, with little or no response to a uniform field of light covering the receptive field.

The signals leaving the retina in the optic nerve therefore highlight the field locations where spatial structure is present, but carry little explicit information about the nature of this structure. For instance, ganglion cell receptive fields are circular, and so respond equally to the contrast in an edge crossing the field whether it is horizontal, vertical, or oblique. Neurons in area V1 also respond to local contrast, but mostly have receptive fields with a preferred orientation. For example, if a vertical contour yields the maximum response, this response falls off sharply for a contour oriented 30° either side of vertical. Neurons with different preferred orientations ensure that contours at any angle around the clock are represented by activity within V1. Orientation seems to be a key organizing principle for V1, indicated by the layout of neurons in a semi-regular array of 'orientation stripes' across the cortical surface; presumably this provides the initial basis for the neural representation of shapes and textures.

A second important organizing principle is that of multi-scale representation. Some cortical neurons respond to very fine-scale spatial detail; others, covering the same region of space with larger receptive fields, respond to similar features on a much coarser scale. Mathematically, this scale variation has the effect of a local analysis in terms of component spatial frequencies of the image. Computational studies show that this is an efficient way to capture the significant structure of scenes.

3.2 From Measurements to Spatial Primitives

The multi-scale array of orientation-selective neurons in V1 can be considered as making a comprehensive set of measurements of the image. To develop a useful representation of scenes and objects, these measurements need to define primitives—elements of an 'alphabet' of spatial structure through which more complex entities can be represented. Thus, a measurement of contrast in a particular orientation and scale does not itself unambiguously define an edge segment; but in conjunction with nearby measurements at a range of scales, it can provide the basis for an assertion of a local edge-segment primitive. While such primitives are important in the computational description of what the visual system is doing, it is unclear whether neurons in visual cortex explicitly represent primitives, or rather whether they are implicit in the pattern of connections amongst V1 cells and from V1 to higher levels.

3.3 Grouping Processes

To yield descriptions of object shapes and bounded surfaces, local edge primitives must be grouped together to form extended contours and areas with a common texture—effects whose importance was first recognized by the Gestalt psychologists. A plausible neural basis for grouping exists in the horizontal connections between V1 cells tuned to a similar orientation. Contours, in turn, define the boundaries of surfaces and objects. Neurons have been reported in V2 which respond to the characteristic alignments of elements where one surface occludes a background surface. Local neural activity in V1 and V2 also reflects the key perceptual distinction between regions that form a distinct figure, and those that form the ground.

3.4 Recognition

One important end point of the processing of spatial pattern is the visual recognition of objects, faces, and scenes. Object recognition can be specifically impaired (agnosia) by damage to the temporal lobe of the brain, and a specific loss of the ability to recognize faces (prosopagnosia) may also occur. Functional imaging studies shows an area that is selectively activated during the processing of faces (the fusiform face area), and another that is active in visual recognition of locations (the parahippocampal place area). Human studies cannot show how individual neurons are responding, but in the temporal lobe of the monkey, single neurons are found to respond specifically to faces. The pattern of activation over a group of such cells conveys sufficient information to distinguish one individual from another. In the inferotemporal region, cells respond to features that can define different types of object. In both cases, the response shows position generalization; that is, it is selective for a particular

class of stimulus, irrespective of its position within a wide receptive field.

Thus, the principles of a sparse population code, found in the representation of local pattern in V1, appear to apply also to the higher level representation of visual identity. However, the hierarchy of neural transformations which connects the former with the latter is very poorly understood.

4. Spatial Location and Layout

As well as allowing the identification of objects and scenes, vision defines the space within which objects are located and are the targets of actions. Disturbances of spatial abilities are found with damage to the parietal lobe in both human and monkey brains. These disturbances are quite distinct from the impairment of object recognition, leading to the proposal of a broad division between a ventral stream of cortical visual processing, from V1 to temporal lobe and concerned with what we are looking at, versus a dorsal stream from V1 to parietal lobe which conveys the information needed to know where it is and how we can act upon it. Different specialized visual areas contribute to the two streams; in particular the motion area V5 (MT) sends information primarily via the dorsal stream (see *Visual Systems: Dorsal and Ventral*).

Spatial information requires a frame of reference. Area V1, and other early visual cortical areas, appear to be organized within a retinotopic frame—that is, the receptive fields, and their topographic arrangement across the cortex, are defined in terms of image locations on the retina. Since the eyes are constantly moving in the head, a stable representation of space requires more than retinotopic information. Eye direction must also be factored in, to establish a head-centered frame. Such combination of retinal and eye-direction information occurs in a number of parietal areas. However, the properties of cells in different parts of parietal cortex suggest that they are preparing visual information in an appropriate form for controlling different action systems—e.g., eye movements, reaching and grasping by the hands, and locomotion. Each of these may require a different frame of reference. An alternative frame of reference is allocentric, i.e., referred to the layout of the environment rather than any part of the body of the observer. It has been suggested that this is a feature of the ventral stream which has to identify objects independently of any particular viewpoint.

Parietal areas transmit information to motor-related areas of the frontal lobe, both directly and via the cerebellum. Comparison of cells in the parietal lobe with those in the premotor frontal areas to which they project emphasizes the difficulty of defining any boundary between visual and motor processing. For instance, activity in some parietal neurons is associated both with viewing a particular three-dimensional shape (e.g., cylinder) and with the grasping action required by that particular shape. 'Mirror neurons' in premotor cortex respond both in association with the hand action of the monkey, and when the monkey is passively viewing another hand executing the same action. It is best to think in terms of visuo–motor circuits which embody the spatial information needed to control and define a particular type of action. It must be recognized that the brain contains multiple circuits for different action types, each of which may work within a spatial frame of reference appropriate for its own requirements.

5. The Binding Problem

Research on visual neuroscience has emphasized the fragmentation of brain processing. Many different neurons are activated by any visual stimulus. Distinct brain areas and subsets of neurons process different attributes such as shape, color, and motion, and there are multiple spatial representations in different visuo–motor circuits. In contrast, our experience of the visual world appears unified, and the different attributes of a visual object are bound together—if a white cup and a green book are together on the table, we do not confuse which color belongs to which object. How perceptual unity arises from a fragmented neural representation is known as the binding problem (Treisman 1996).

Two, not necessarily incompatible, answers have been widely discussed. One is based on Anne Treisman's feature integration theory of attention. This theory, supported by evidence from visual search, proposes that a 'spotlight' of focal attention modulates processing within distinct 'feature maps' for different attributes, and so multiple feature attributes are accessed only for the object that is the current focus of attention. The second idea is that cortical neurons show oscillations in their activity at frequencies around 40 Hz, and that temporal coherence between these oscillations acts as a label for different signals contributing to the representation to a common object. Such coding by temporal synchrony has attractions; it could provide a basis for understanding how features at different locations within a single map can be linked or segregated in object-based representations, as well as how binding occurs across different maps. However, the evidence for it remains quite limited.

6. Neural Basis of Visual Sensitivity and Thresholds

Much detailed knowledge of vision comes from the analysis of psychophysical thresholds—measures of how sensitive the visual system is in making fine discriminations. If these data are to be integrated with knowledge from neurophysiology, the relation be-

tween psychophysical threshold and patterns of neural activity needs to be understood.

'Threshold' conveys the suggestion of a stimulus intensity below which no neural activity is elicited. However, in practice, neurons are rarely silent, and their activity, even with a fixed stimulus, shows variability or 'noise.' Thus, the ability to detect weak stimuli is determined by the statistical problem of distinguishing a stimulus-driven change in activity from those due to noise.

Several studies have determined the 'thresholds' of single neurons in monkey visual cortex, in terms of the stimulus level which achieves a statistically defined level of reliable response. These can be compared with behavioral performance in detection; in some cases behavioral and neural measures can be made simultaneously. The information carried by single neurons proves to be remarkably close to that implied by the behavioral responses. Remarkable, because the animal has the opportunity to pool the information from thousands of neurons, which would be expected to yield much more reliable information. The data can be reconciled by realizing that although large numbers of neurons may be involved, the random variations in their activity are correlated so the number of genuinely independent neural signals is much smaller (Newsome et al. 1995).

Whether a particular judgment depends on pooling activity in tens, hundreds, or thousands of neurons, there must be many more neurons in the same brain area whose activity is irrelevant because they do not respond selectively to the stimulus concerned. If activity of such neurons was included in the pooling process, it would introduce noise and degrade sensitivity. Thus sensitive discrimination, either in the laboratory or in real-life situations, must depend on selecting a small minority of neurons whose output is informative for the task in hand.

7. *Modulation of Neural Activity by Attention and Task Demands*

It could be conceived that neural activity, relevant to a particular task, was channeled to appropriate decision and response mechanisms through a purely 'bottom-up' process. However, Sect. 3 has mentioned the ample evidence for 'top-down' control of visual processing, modulating activity at many levels according to the central decisions about what is relevant. For instance: (a) neural activity, as early in the pathway as the LGN, is strongly modulated according to the sleeping–waking state of the organism; (b) the response of single neurons in areas such as V4 and V5 in monkeys can be changed radically according to whether the preferred stimulus of the neuron is behaviorally relevant or irrelevant on a particular occasion; (c) functional brain imaging shows that the overall activation of an area such as V5 in humans is enhanced by attention to the property that it analyses (in this case, motion), even if the stimulus is unchanged; and (d) within a topographically organized area such as V1, activation of a particular region is enhanced when attention is directed to the corresponding region of visual space (see *Attention, Neural Basis of*).

8. *Plasticity and Learning in Visual Neural Organization*

Section 7 discussed how the function of the visual system can be dynamically modulated from moment to moment. Visual neural function shows long-term adaptive changes, in response to the inputs it receives. For instance, if the input to visual cortex has an anomalous distribution across the retina (e.g., as a result of a congenital absence of cones in the central fovea) the usual retinotopic map in V1 is distorted to match this distribution. When the balance of input from the two eyes is disturbed (e.g., when the input of one eye is reduced in a critical period during early development) the cortex adjusts to receive less input from the less active eye. The plasticity seen in these abnormal cases must also play a role in normal development: development of the normal neural organization of visual brain areas must depend on the regularities and correlations in the visual input that these areas receive.

Visual abilities show strong learning effects; almost any visual discrimination becomes strikingly more sensitive with practice. This learning can be highly specific, for example showing improvement for stimuli only at the orientation that has been practiced, suggesting that changes occur at relatively low levels of visual processing such as V1. The discussion in Sect. 6 suggests that such learning may be based on establishing highly specific connections to decision mechanisms from the neurons most sensitive to the stimulus concerned. However, while changes with learning have been demonstrated in the responses of visual cortical neurons, the basis of these changes in modified connections is still speculative.

9. *The Neural Basis of Visual Awareness*

The detailed understanding of how visual information is represented in the nervous system does not answer the question: what aspect of neural activity corresponds to the conscious awareness of the subject? Many aspects of vision (for example, which eye received a monocular stimulus) are not accessible to consciousness, even though the relevant information exists within the nervous system. Is there some subset of visual neurons whose activity contributes specifically to awareness?

Individuals with damage to V1 report that they are blind in the region of the visual field corresponding to

the damage. They can sometimes make reliable responses to stimuli they do not report seeing. Thus there must be routes for visual information, probably through the superior colliculus in the midbrain, whose activity has no conscious correlate. However, it is not known whether V1 activity is itself a necessary correlate of conscious visual experience, or whether V1 simply serves as the only access route to higher processes that have such a correlate.

In binocular rivalry, discrepant patterns are presented to the two eyes, and alternate in perception. Neuronal activity can be monitored in monkeys who are also signaling which pattern is perceptually dominant. For some cortical neurons, activity fluctuates in correspondence with perceptual dominance, but for others there is no correlation. The proportion in the former group increases in going from lower to higher cortical areas (Logothetis 1998). These results suggest that awareness is not uniquely associated with particular cortical areas, but may emerge cumulatively as information is processed through the system. The true relation between brain and consciousness will remain speculative for many years, but visual processing is likely to provide the key test cases for this central problem of the human sciences (see *Consciousness, Neural Basis of*).

See also: Color Vision Theory; Eye Movements and Perception; Face Recognition: Psychological and Neural Aspects; Neural Representations of Direction (Head Direction Cells); Neural Representations of Objects; Occipital Lobe (Visual Cortex): Functional Aspects; Perception of Extrapersonal Space: Psychological and Neural Aspects; Vision, High-level Theory of; Vision, Low-level Theory of; Vision, Psychology of; Visual Memory, Psychology of; Visual Space, Geometry of; Visual System in the Brain

Bibliography

Barlow H B 1995 The neuron doctrine in perception. In: Gazzaniga M S (ed.) *The Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 415–35

Churchland P S, Sejnowski T J 1992 *The Computational Brain*. MIT Press, Cambridge, MA

De Valois R L, De Valois K K 1988 *Spatial Vision*. Oxford University Press, New York

Gazzaniga M S, Ivry B, Mangun G R 1998 In: *Cognitive Neuroscience: The Biology of the Mind*. W. W. Norton, New York, Chap. 4–6

Logothetis N K 1998 Object vision and visual awareness. *Current Opinion in Neurobiology* **8**: 536–44

Newsome W T, Shadlen M N, Zohary E, Britten K H, Movshon J A 1995 Visual motion: linking neuronal activity to psychophysical performance. In: Gazzaniga M S (ed.) *The Cognitive Neurosciences*. MIT Press, Cambridge, MA, pp. 401–3

Treisman A 1996 The binding problem. *Current Opinion in Neurobiology* **6**: 171–78

O. Braddick

Visual Space, Geometry of

Human orientation in space is based primarily on vision: we are able to judge the size of an object even from a distance, its form and shape, and its egocentric distance. The set of all objects in the visual field together with the 'perceived' spatial relationships between them is called 'visual space.' To some extent, our perceptions are veridical. There are, however, marked and stable deviations from the physical relationships. This observation raises two questions: (a) are the regularities of perceived spatial relationships characterized by some particular geometry? and (b) if so, what are the psychological functions relating this geometry to the well-known Euclidean geometry of physical space? This article is only concerned with the first question (for the second one see *Binocular Space Perception Models*). It has four main parts. After some introductory and historical remarks on first observations and perceptual phenomena leading to the notion of non-Euclidean geometry of visual space in the early 1900s (Sect. 1) a formal characterization of what is meant by a geometry is given in Sect. 2. Section 3 presents a discussion of theoretical developments in the field of geometries for visual space, in particular the theory of Luneburg (1947). Section 4 reviews related empirical investigations and results.

1. Is Visual Space Non-Euclidean?

The question of whether visual space is Euclidean was not formulated explicitly until the beginning of the twentieth century. Answers to this question, however, can be traced back as far as the ancient Greek philosophers. Suppes (1997) presents a list of classical contributions to this problem, starting with Euclid's theory of perspective (rated as a 'Yes' answer). His list contains a majority of 'No' statements, some undecided cases rated as 'not sure,' and only a few 'Yes' conclusions, mainly from authors using philosophical arguments rather than empirical results. Strong evidence against Euclidean geometry as an appropriate description of the structure of visual space arose from observations in so-called alley experiments. Investigating the phenomenon of size constancy, Hillebrand (1902, cited in Lukas 1996) asked subjects to arrange small light points in the horizontal plane along two lines such that they should form an alley-like configuration, running from two fixed points towards the observer. He used two similar tasks but with different instructions. The first was to match the 'horizontal distances' of points to make sure that, perceptually, the width of the alley should be perceived as equal at any egocentric distance (D-alley).

The second was to pay attention to the alley's right and left border and make sure that they are perceived as 'parallel straight lines' (P-alley). To Hillebrand's

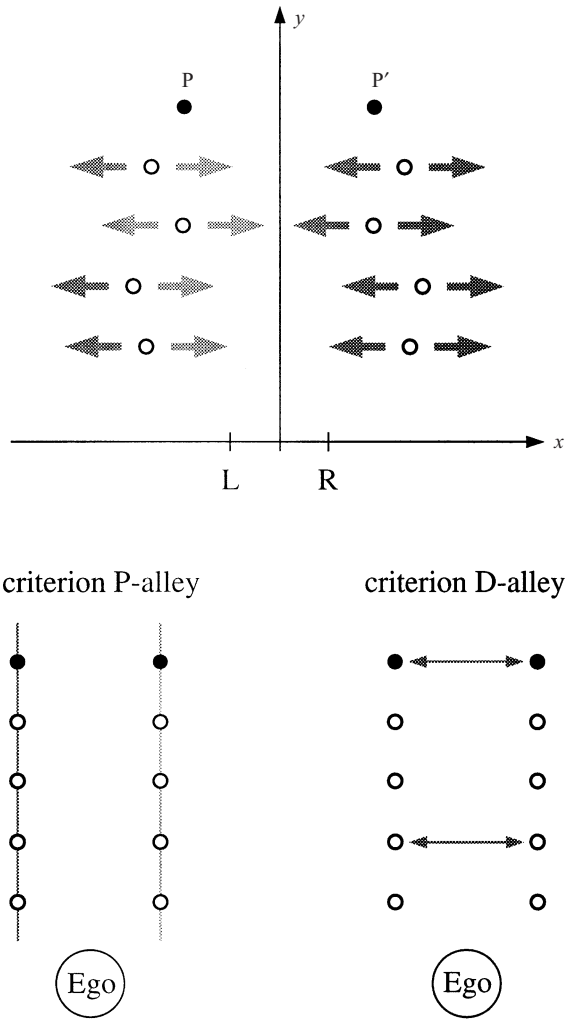

Figure 1
Classical setting for alley experiments. P and P′ are fixed points at eye level, L and R denote the observer's left and right eye, respectively. The subject adjusts movable points according to one of the criteria shown below (left: parallel alley, right: distance alley)

surprise, the two instructions led to different configurations. Both alley types were curved, with D-alleys consistently lying 'outside' the corresponding P-alleys (see Fig. 1). Blumenfeld (1913, cited in Lukas 1996) confirmed this finding and was the first to mention a far-reaching consequence: since in Euclidean geometry lines with constant distance everywhere are called parallel (and vice versa), visual space has to be non-Euclidean. Blumenfeld's geometric reasoning elicited a vivid discussion, but for more than 30 years thereafter, it was not worked out mathematically in any substantive way. Luneburg (1947) was the first to present both a complete geometrical framework for visual space, and explicit psychophysical functions that related physical and visual coordinates in order to explain the striking results of the alley experiments and other empirical facts in binocular space perception known at that time.

2. *Geometry as a Mathematical Language for Spatial Structure*

Within mathematics there are at least two different approaches to geometry, the 'synthetic' and the 'analytic'. In the spirit of Euclid's axiomatic approach, synthetic geometry is defined over a set of points and a set of straight lines as primitives by specifying a number of relations considered to be true. Over many centuries Euclid's axioms have been considered to be the only reasonable description for spatial relationships. They include incidence axioms (for example, any two distinct lines have at most one common point), order axioms, axioms concerning motion and congruity, metric axioms, and, finally, the famous parallel axiom stating that for any point P and any line g not containing P there is one and only one line g' through P parallel to (not intersecting) g. At the beginning of the nineteenth century, however, Gauss, Bolyai, and Lobachevski discovered independently that the parallel axiom can be replaced by the postulate that there exist at least two parallels, yielding a different type of geometry (later called 'hyperbolic'). Moreover it has been shown that even a third version of the parallel axiom, stating that any two lines have at least one point in common, is compatible with the remaining Euclidean axioms, and leads to an 'elliptic' geometry.

The properties of both types of non-Euclidean geometry are demonstrated easily by models that fulfill the respective set of axioms. A well-known model for an elliptic plane is the surface of a sphere in Euclidean space $I\mathbb{R}^3$, where the set of elliptic lines is defined as the set of all (Euclidean) great circles. Similarly, models for hyperbolic geometry can be constructed on surfaces with constant negative curvature in $I\mathbb{R}^3$ (for example, saddle-shaped surfaces). Hyperbolic geometry is therefore said to have constant negative curvature, whereas elliptic geometry is characterized by its constant positive curvature.

3. *Theories on the Geometry of Visual Space*

Let V be the set of all points in the three-dimensional physical space which a fixed observer can see without moving the head. If x, y, z denote the Euclidean coordinates in physical space, then V is a subset of the positive half space $y > 0$ right in front of the observer. The origin of the coordinate system is interpreted as

the egocenter of the observer and is assumed to lie in the closure of V, the x-axis denotes the direction from left to right, and the z-axis corresponds to the vertical axis. We restrict the following discussion to the horizontal plane $z = 0$ at eye level. The well-known Euclidean metric d_{phys} represents the physical distances in V.

In his famous attempt to represent the perceptual structure of binocular visual space geometrically, Luneburg (1947) starts out with the basic assumption that in addition to d_{phys} there exists some metric d_{vis} on V that represents the 'perceived' spatial structure. If γ and ϕ denote binocular parallax and bipolar latitude, respectively (see *Binocular Space Perception Models* for a definition), his theory can be summarized by the following basic hypotheses:

- LH1 There exists a Riemannian metric d_{vis} on the visual field V that represents perceived distances. The geodesics of d_{vis} are perceived as straight lines.
- LH2 The visual space (V, d_{vis}) has constant Gaussian curvature K.
- LH3 The perceived distance d_{vis} $(0, P)$ of point P to the observer is a function of the bipolar parallax only.
- LH4 The ϕ-translations $(\gamma, \phi) \to (\gamma, \phi + \text{const})$ are visual isometries (leave all perceived distances unchanged).
- LH5 The γ-translations $(\gamma, \phi) \to (\gamma + \text{const}, \phi)$ are angle-preserving.

Assumption LH1 provides the basis for the remaining assumptions: In a Riemannian metric space a coordinate system can be introduced that allows numerical representations for any geometrical objects such as lines, angles, triangles, or distances. LH2 restricts the discussion to geometries that are constantly curved everywhere. The remaining assumptions LH3 through LH5 define the psychophysical relationship between (physical) stimulus space and (perceptual) visual space. LH3 implies, for example, that points lying on a Vieth–Müller circle (a circle containing the nodal points of the observers eyes) are perceived as equidistant to the observer because they have the same bipolar parallax γ. These latter assumptions are explained and discussed extensively in *Binocular Space Perception Models*.

Eschenburg (1980) gave a complete proof of Luneburg's conjecture that LH1 through LH5 uniquely determine the metric d_{vis} by two parameters K and σ up to equivalence. More specifically, the line element of the metric is represented in terms of the bipolar coordinates ϕ and γ by

$$ds^2 = \frac{2\sigma^2(d\gamma^2 + d\phi^2)}{e^{\sigma\gamma} + Ke^{-\sigma\gamma}}$$

(Eschenburg 1980 p. 7). K denotes the Gaussian curvature of visual space, and σ is an individual scaling factor.

4. Empirical Results on Geometrical Structures of Visual Space

Whereas LH3 through LH5 are closely connected to the psychophysical relationship between physical and visual coordinates, LH1 and LH2 define the inner geometry of visual space. Empirical investigations have concentrated on two questions: (a) is K constant? (b) if so, is $K = 0$ indicating Euclidean geometry, or $K < 0$ (hyperbolic geometry) or $K > 0$ (elliptic space)? The early observations of distance alleys lying consistently outside the corresponding parallel alleys are only compatible with hyperbolic geometries (Indow and Watanabe 1984). Direct numerical estimates for K can be derived from alley experiments or from so-called 3- and 4-point experiments using double Vieth–Müllercircles (VMC) suggested by Luneburg (1947) (see Fig. 2). Using these methods K has mostly been found to be negative (Hardy, Blank, Boeder, Rand, and Rittler, 1953, cited in Lukas 1996). However, numerical estimates of K cannot always reliably be reproduced in repeated measurements, estimates resulting from 3- and 4-point experiments are not valid predictors for data obtained from alley experiments, and experiments over large egocentric distance ranges

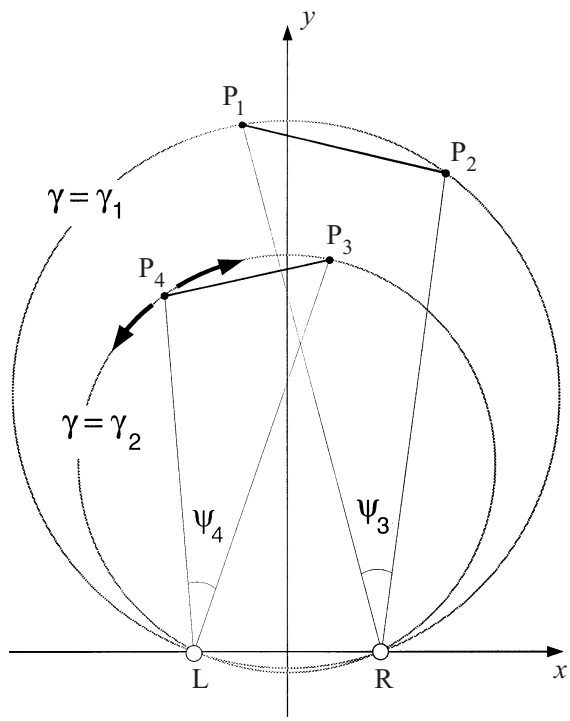

Figure 2
Experimental procedure in 4-point experiments. P_1, P_2, and P_3 are fixed on the Vieth–Müller circles $\gamma = \gamma_1$ and $\gamma = \gamma_2$, respectively. P_4 is adjusted by an observer such that P_4P_3 matches P_1P_2 perceptually.

coincide qualitatively but not quantitatively (Indow 1991).

Moreover, for a test of LH 2, numerical estimates of K are only of limited utility, because they presuppose the psychophysical mapping functions. Several methods have been suggested to test LH 1 and LH 2 independently from the remaining assumptions. In the triangle method used by Blank (1961) subjects had to bisect two sides of a given triangle ABC. Six of a total of seven subjects gave judgments indicating a negative curvature. However, the results seem to depend on the size of the triangle. Using the same method, Higashiyama (1981) found strong variations in curvature depending on lateral positions of the triangles and their distance from the observer. Battro et al. (1976) used stakes to produce experimental alleys of variable size in large open fields in natural environments. The curvature of visual space did not show a negative constant value. In addition, the results of experiments with equidistant lines and visual triangles under these viewing conditions also invalidated the constant curvature assumption.

Luneburg (1947) used quite a different argument and referred to the fact that Riemannian spaces with constant curvature are often characterized geometrically by a property called free mobility (any configuration can be moved to any position in space without changing the shape). Eschenburg (1980) showed that, contrary to Luneburg's argumentation, the existence of visual isometries alone does not guarantee free mobility in a strict sense. In Blank's (1958, cited in Blank 1961) axiomatic version of Luneburg's theory, free mobility is guaranteed by an axiom that implies both the so-called Desarguesian property, and constant curvature in Riemannian spaces. Empirical consequences of this axiom have been tested successfully by Foley (1964, cited in Foley 1972). However, a much stronger test in subsequent experiments failed (Foley 1972). Lukas (1996) gave detailed arguments as to why the first test might have been too weak and the latter one unreasonably strong. From an empirical point of view Eschenburg (1980) provides the most powerful and sensitive test of the constant curvature hypothesis. He shows that an appropriate combination of the double VMC-experiments produces necessary and sufficient conditions for LH2. Experimental tests of these conditions reported in Lukas (1996) reflect a situation that seems to characterize most results in this field: the data are nicely fitted qualitatively by the predictions using constant curvature, yet there are numerically small, but statistically significant deviations.

Nevertheless, there are few models in the literature that try to generalize Luneburg's geometrical assumptions (contrary to his psychophysical assumptions, which have been relaxed in several attempts, see *Binocular Space Perception Models*). Yamazaki (1987) developed a concrete non-Riemannian geometry and suggested its application to visual space. Because this model does not use the concept of length, its utility for empirical research seems limited.

A different approach is reported in Lukas (1996). Starting out with an axiomatic analysis of size constancy, he defines 'apparent distance' in terms of modern measurement (see *Measurement*, *Representational Theory of*) and shows that a very weak invariance condition resembling Luneburg's iseiconic transformations (defined in LH 4 and LH 5) forces the geometric structure of visual space to be Euclidean.

5. *Conclusion and Future Directions*

Critical reviews of empirical results seem to indicate that there is little hope of characterizing visual space globally by one particular geometry. Suppes (1995) concludes that visual space is context-dependent and no reasonable simple set of axioms in the spirit of synthetic geometry can be given for the structure of visual space. This view seems to be supported by recent results from Zimmer (1998). In a series of experiments, subjects failed in a test of very basic requirement of 'betweenness' indicating that not even simple order structures on a straight line can be assumed to hold *a priori*. Nowadays, geometric models in visual perception are actively discussed almost exclusively in the context of mathematical psychology and foundational aspects of science. The main stream of research in vision has obviously shifted to computational models (see *Vision*, *Low-level Theory of*) and the analysis of cognitive processes underlying human performance in spatial perception.

See also: Binocular Space Perception Models; Motion Perception Models; Navigation in Spatial Environments, Neural Basis of; Object Recognition: Theories; Perception of Extrapersonal Space: Psychological and Neural Aspects; Vision, High-level Theory of; Vision, Low-level Theory of; Vision, Psychology of

Bibliography

Battro A M, di Pierro Netto S, Rozestraten R J 1976 Riemannian geometries of variable curvature in visual space: Visual alleys, horopters, and triangles in big open fields. *Perception* **5**: 9–23

Blank A A 1961 Curvature of binocular visual space. *Journal of the Optical Society of America* **51**: 335–9

Eschenburg J H 1980 Is binocular visual space constantly curved? *Journal of Mathematical Biology* **9**: 3–22

Foley J M 1972 The size–distance relation and intrinsic geometry of visual space: Implications for processing. *Vision Research* **12**: 323–32

Higashiyama A 1981 Variation of curvature in binocular visual space estimated by the triangle method. *Vision Research* **21**: 925–33

Indow T 1991 A critical review of Luneburg's model with regard to global structure of visual space. *Psychological Review* **98**: 430–53

Indow T, Watanabe T 1984 Parallel- and distance alleys with moving points in the horizontal plane. *Perception and Psychophysics* **35**: 144–54
Lukas J 1996 *Psychophysik der Raumwahrnehmung*. Psychologie Verlags Union, Weinheim, Germany
Luneburg R K 1947 *Mathematical Analysis of Binocular Vision*. Technical Report, Princeton, NJ
Suppes P 1977 Is visual space Euclidean? *Synthese* **35**: 397–421
Suppes P 1995 Some foundational problems in the theory of visual space. In: Luce R D, D'Zmura M, Hoffman D D, Iverson G J, Romney A K (eds.) *Geometric Representations of Perceptual Phenomena: Papers in Honor of Tarow Indow on his 70th Birthday*. Lawrence Erlbaum Associates, Mahwah, NJ
Yamazaki T 1987 Non-Riemannian approach to geometry of visual space: An application of affinely connected geometry to visual alleys and horopter. *Journal of Mathematical Psychology* **31**: 270–98
Zimmer K 1998 *Experimentelle Untersuchungen zur Geometrischen Struktur des Binokularen Sehraums*. Shaker, Aachen, Germany

J. Lukas

Visual System in the Brain

The visual system in the brain refers to both subcortical and cortical neural structures. In primates and other mammals these structures include various targets of the retina in the midbrain and thalamus as well as a large number of cortical areas. Several parallel visual pathways can be identified that are based on the output of different types of ganglion cells in the retina. Visual cortex is comprised of a large number of cortical areas that can be distinguished by a number of criteria. The patterns of connections among these cortical areas allow them to be organized into a hierarchy of visual processing. Many of these cortical areas are organized into repeating modular subunits that appear to processing all the visual information within a small portion of visual space.

1. Subcortical Visual Structures

1.1 Midbrain Structures

The retina projects to a variety of nuclei in the midbrain that are involved in reflexive visual functions. These pathways are best understood in cats, but similar observations have been made in non-human primates. Specific classes of ganglion cells are known to project to the pretectum (PT-sparse and PT-others), where their information is used for the control of pupillary muscles (Rodieck and Watanabe 1993). Other ganglion cells project to the superior colliculus that supports reflexive movements of the eyes and head, especially in response to moving targets. In cats these ganglion cells are known as Y cells (alpha morphological type), but in macaque monkeys only a small proportion are parasol cells (similar to the Y cells of the cat), with the majority of cells coming from a diverse group with varying morphologies (Rodieck and Watanabe 1993).

The superior colliculus gives rise to pathways that ascend into the thalamus to innervate the pulvinar nucleus. In addition, the superior colliculus receives descending projections from a large number of visual cortical areas. These pathways linking the superior colliculus with extrastriate cortical areas describe a second cortical visual system that is distinct from the geniculo-striate system described later. This second pathway is thought to subserve the visual ability known as blindsight, where judgments about objects and especially motion can be made in the absence of true conscience experience.

1.2 Lateral Geniculate Nucleus

The vast majority of retinal ganglion cells project to the lateral geniculate nucleus of the thalamus. The lateral geniculate nucleus is a multilayered structure that receives input from both eyes to build a representation of the contralateral visual hemifield. The basic pattern of the LGN in primates consists of four layers, two each for each eye. The upper two layers become interleaved to form four layers, two each per eye. In primates, three or more anatomical classes of ganglion cells innervate the LGN. The parasol cells of the retina project to the lower two large cell containing layers, known as the magnocellular layers of the LGN. These ganglion cells and their LGN targets have large, concentrically organized receptive fields, lack color selectivity, have high contrast sensitivity, and give transient responses to stimuli. The midget ganglion cells of the retina project to the upper four small cell, or parvocellular layers of the LGN. These cells and their LGN targets have small receptive fields, which are organized with color opponent centers and surrounds, based largely on red and green cone inputs. These cells provide the majority of input to visual cortex, but their signals unambiguously represent stimulus color since different responses are observed for different size stimuli. The third class of ganglion cell that projects to the lateral geniculate nucleus is known as the bistratified cells which project to cells located in the intralaminar zones between the main LGN layers (Rodieck and Watanabe 1993). Physiological investigations have identified these cells to be of the blue/yellow type and their receptive fields are not concentrically organized. These cells project to distinct layers in striate cortex where they are though to subserve the beginning of a cortical color processing system.

1.3 Pulvinar Complex

The pulvinar is a collection of nuclei in the thalamus that are largely related to visual processing in higher

cortical areas. In phylogeny, the pulvinar nuclei have increased dramatically in size in parallel with the grown of these higher cortical areas. There remains some controversy about the number of nuclei within the visual pulvinar. On the basis of connections, a collicular recipient zone and a striate cortex recipient zone have been recognized. In addition, area MT projects to a distinct zone within the inferior pulvinar. The remaining subdivisions are less clear, but at least three subdivisions can be distinguished on the basis of their staining for the calcium binding protein parvalbumin (Cusick et al. 1993).

2. *Visual Cortical Areas in Nonhuman Primates*

2.1 *Criteria Used for the Characterization of Cortical Areas*

Visual cortex of all mammals consists of at least two and often a much larger number of distinct cortical areas. However, the specific criteria used to identify these areas vary considerably. The earliest views of visual cortical organization were based on anatomical criteria that compared the size and distribution of cell bodies and fibers across the cortical mantle. The most prominent of these architectonic studies were those performed by Brodmann (1909) and von Economo and Koskinas (1925) who subdivided the occipital cortex into three distinct zones. Primary visual cortex, also known as striate cortex, V1, area 17 of Brodmann, or area OC of von Economo and Koskinas has a very distinctive appearance, due largely to the sublamination of layer 4 and the presence of a dense fiber band in Brodmann's layer 4B, the stria of Gennari. Although the architectonic method allowed the identification of two additional belts of cortex surrounding striate cortex, subsequent studies have indicated that these belts contain more than one functional unit. More recent studies have used various histochemical and immunocytochemical markers to define distinctive anatomical features of several cortical areas.

The next most useful criterion for subdivision of cortical areas is the presence of a distinct topographic map of the contralateral hemifield. Typically, microelectrode mapping was used to define topographic maps of visual cortex and thus to delineate areas. Recently, functional imaging techniques have been used to define the topographic organization of visual areas, especially in humans. The third criterion used for the characterization of visual cortical areas has the specific pattern of intracortical and subcortical connections that a given field makes. These patterns studied through various pathway tracing techniques that label the origins or terminations of specific pathways following the intracortical injection of various tracers. The areas that are innervated and the specific laminar patterns of their connections thus help define specific cortical areas.

The next criterion used to identify cortical areas utilizes physiological methods to describe the functional properties of their neurons. Typically microelectrode recording techniques are used to study the receptive field properties of cells in a given area, which are then compared to other cortical areas. This approach has been most useful for the identification of area MT in primates, whose neurons are nearly all selective for the direction of object motion. Recently, functional imaging techniques have allowed physiological investigations in the visual cortex of humans. These studies first used positron emitted tomography (PET) and more recently functional magnetic resonance imaging (fMRI) has allowed the identification of human cortical areas specialized for motion processing (MT), color processing (V4 or V8), or for the processing of faces or objects.

The last criterion used to study the organization of visual cortex is distinctive behavioral deficits following localized lesions or transient inactivation. This criterion was first used to identify occipital cortex as the locus of visual processing. Subsequent studies have revealed specific deficits in motion or color processing following injury to area MT or ventral occipitotemporal cortex, respectively (Zeki 1990, 1991).

Although each of these criteria is useful for the identification of visual cortical areas, most areas have only been characterized by one or two criteria. Furthermore, the results of different studies have led to different conclusions, so some areas are better established than others. Therefore, the described organization of visual cortex should be considered a work in progress rather than the final word on visual cortex in primates.

2.2 *Cortical Areas in Macaque Monkeys*

Traditional views of the organization of visual cortex in monkeys and man were based on the idea of a primary visual area, V1 or area 17, which was surrounded by two belts containing visual association cortex and 'visual psychic cortex.' As additional anatomical and physiological criteria became utilized, it became clear that visual cortex occupies the whole of the occipital lobe and extends into the parietal, temporal, and frontal lobes. Figure 1 illustrates the location and extent of visual cortical areas of the macaque monkey displayed on an unfolded cortical map and on medial and lateral views of the cerebral hemisphere. As discussed above, many of these cortical areas, such as V1, V2, and MT are well characterized, appear in all primates, and are well accepted by the research community (Kaas 1997). Other areas are less well described and their precise borders and names remain somewhat controversial. As a whole, these visual areas occupy more than half of the surface area of the cerebral cortex of the macaque. They range in size from over 1100 mm^2 for areas V1 and V2 to less

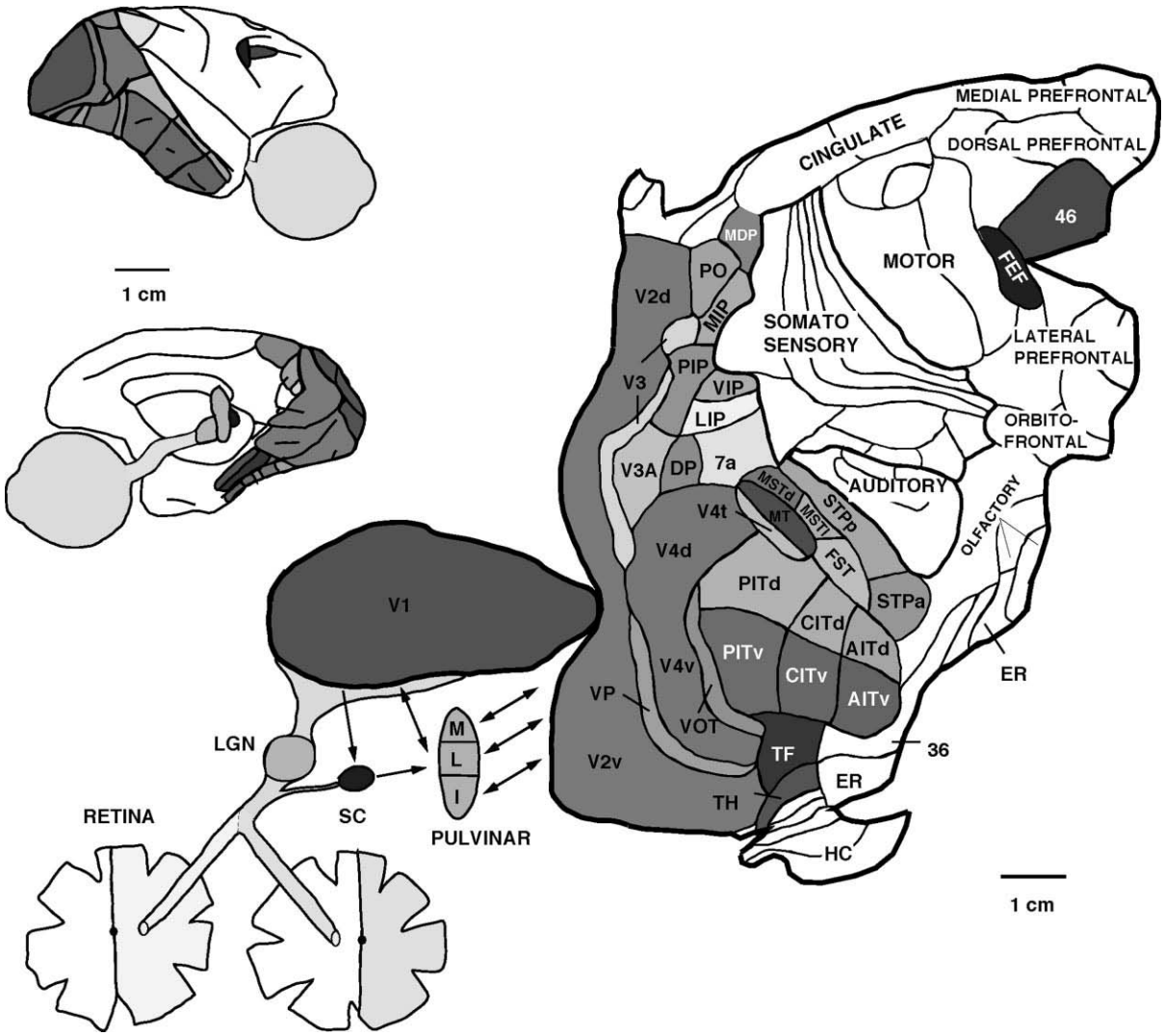

Figure 1
Visual areas in the macaque monkey. (Right) Two dimensional map of the macaque monkey cerebral cortex illustrating the location and extent of 32 visual cortical areas. Shaded regions are visual cortical areas located primarily in the occipital, parietal, and temporal lobes. Two visual areas, FEF and area 46 are located in the frontal lobe. (Left) Lateral (upper) and medial (lower) views of the macaque cerebral cortex illustrating the location and extent of visual cortical areas. Felleman and Van Essen 1991. Distributed hierarchical processing in the primate cerebral cortex. © Oxford University Press. Reproduced with permission

than 40 mm^2 for areas MSTd, MSTl, and V4t. Recent studies have extended this view of the organization of visual and visuo-motor areas in occipital parietal cortex (e.g., Lewis and Van Essen 2000). Specifically, this investigation used architectonic criteria to subdivide area LIP into LIPd and LIPv, VIP into VIPl and VIPm, and to redefine the extent of area PIP through the distinction of area LOP in its lateral margin.

Visual cortical areas vary in their functional specializations within perceptual and perceptual-motor categories. Overall, the visual cortex has been subdivided into dorsal and ventral processing streams (Ungerleider and Mishkin 1982). The dorsal stream begins in areas V1 and V2 and proceeds through area MT into parietal cortex. This functional stream, often called the 'where' stream, utilizes stereoscopic and direction specific signals in area MT to feed parietal areas that

are involved in the planning of eye and limb movements to visual targets. The ventral stream also begins in areas V1 and V2 and proceeds through area V4 into temporal cortex. This functional stream, often called the 'what' stream, utilizes color, shape, motion, and binocular disparity information to identify objects and to enable their encoding into visual memory. Although the dorsal and ventral streams are often described as separate and parallel processing streams, there are many interconnections between them that emphasize the interdependence of these various visual functions.

3. Hierarchical Organization of Visual Areas

Many of the visual areas described above have been identified primarily using their patterns of connections with other cortical areas. While earlier views of cortical pathways viewed each area as having a few inputs and a few outputs, current studies have greatly modified this view. Furthermore, cortical pathways differ in their laminar pattern of inputs and origin of outputs. Pathways that terminate in layer 4 of a cortical area are generally recognized to be of the feedforward type, characteristic of the thalamic input to primary visual cortex. In contrast, so-called feedback pathways tend to avoid layer 4 to terminate in layer 1 and 6. Finally, the lateral pathway has been observed which terminates in a column extending across all cortical layers. These different laminar patterns of terminations have been correlated with different laminar patterns of origin. Some feedforward pathways arise exclusively from cells located in the upper cortical layers, while some feedback pathways arise exclusively from the infragranular layers. Many pathways arise from cells located in both supragranular and infragranular layers and they may be of the feedforward, feedback, or lateral type.

A large number of cortical pathways have been identified interconnecting the 32 visual areas described in Fig. 1. As of 1990, more than 300 pathways had been recognized and since then additional areas and many additional pathways have been identified. By comparing the laminar patterns of inputs and outputs of these 32 areas it has been possible to derive an anatomical hierarchy of visual areas in the macaque (Felleman and Van Essen 1991, Maunsell and Van Essen 1983). In the hierarchy illustrated in Fig. 2, each visual area was assigned to a level based upon pairwise comparisons of each area with all other areas that it connects to. According to this scheme, nearly 90 percent of the cortical pathways can be unambiguously assigned as the feedforward, feedback, or lateral type. Most of the ambiguous connections are among visual areas making connections with frontal cortex. Although Felleman and Van Essen were able to derive a hierarchy based on the reported connections of a large number of visual areas, other investigators suggest that these data are not conclusive and thus it is not possible to derive a single cortical hierarchy (Hilgetag et al. 1996)

Recently, quantitative methods have been used to characterize the laminar patterns of origin for cortical pathways among the lower visual areas (Barone et al. 2000). This scheme introduces an anatomical distance metric between areas based on the proportion of supragranular cells in a given pathway. This analysis yields a hierarchy that in many ways is similar to the qualitative hierarchy described above, but differs in the assignment of a few areas. Future work will extend these quantitative studies to additional visual areas.

4. Modular Organization and Processing Streams in Visual Cortex

4.1 Modular Organization of Striate Cortex

A variety of anatomical and physiological evidence indicates that striate cortex is organized in modular fashion. The first module identified was that of ocular dominance columns. In macaque monkeys, inputs from the two eyes are segregated in layer 4 into ~ 0.5 mm wide zones that appear like zebra stripes when viewed tangentially. These stripes can be identified using anatomical methods using tracer transport or through histochemical staining for cytochrome oxidase following reduction in activity in one eye through disease or lid suture.

Histochemical staining for cytochrome oxidase reveals an array of dense, metabolically active blobs separated by pale, interblobs (see Fig. 3a). Anatomical studies have revealed that these metabolically active blobs are located in the centers of the ocular dominance stripes for each eye, and they are repeated across the cortical surface. Physiological investigations indicate that cells located within the blobs tend to lack orientation selectivity and tend to be color selective. In contrast, cells located in the interblobs are orientation selective and tend to lack color selectivity (Livingstone and Hubel 1984). Furthermore, orientation selectivity is organized in a specific pattern such that the preferred orientation of cells progresses systematically around a non-oriented blob center. Studies of circuitry from the thalamus and within V1 indicate that cells in the blobs receive a direct input from the konicellular (K) cells of the LGN and receive input from the parvocellular and magnocellular streams via inputs from layer 4Cβ and 4B, respectively. The region of cortex containing one blob and the corresponding orientation 'pin-wheel' that surrounds it has been described as a modular hypercolumn that contains all the processing elements for a small patch of visual space.

4.2 Modular Organization of Areas V2 and V4

Area V2 also contains a modular organization that can be revealed through staining for cytochrome

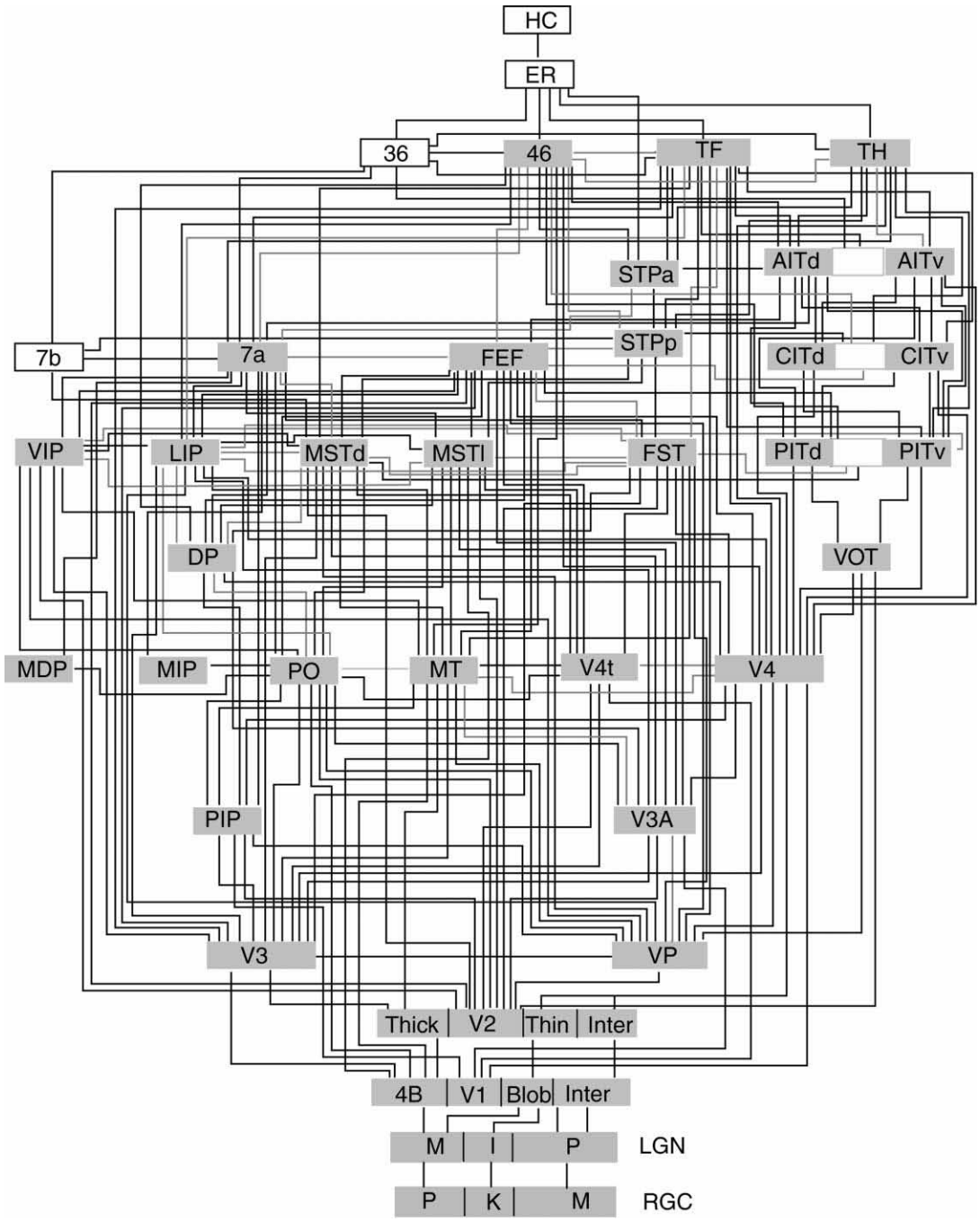

Figure 2
Hierarchy of visual areas in the macaque monkey. Anatomical hierarchy of the cortical areas illustrated in Fig. 1. This hierarchy reflects 305 feedforward, feedback, and lateral pathways among the 32 visual areas. Two subcortical stages are illustrated at the bottom corresponding to retinal ganglion cells (RGC) and the lateral geniculate nucleus (LGN). In general, occipital areas are located near the bottom of the hierarchy, parietal areas on the left, and temporal areas on the right. Felleman and Van Essen 1991. Distributed hierarchical processing in the primate cerebral cortex. © Oxford University Press. Reproduced with permission

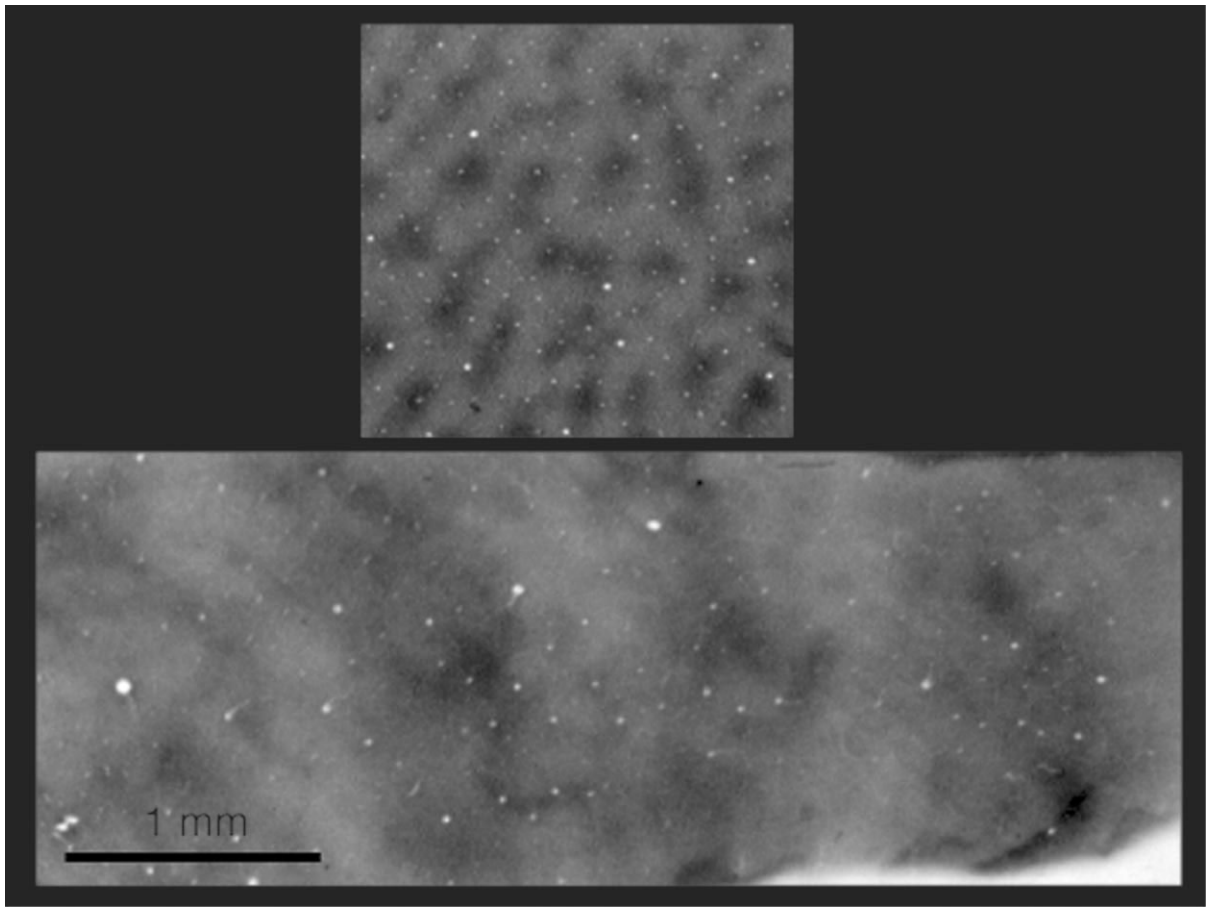

Figure 3
Cytochrome oxidase staining in visual cortex. A. Tangential section through the upper layers of area V1 illustrating the cytochrome oxidase dense blobs and pale interblobs. B. Tangential section through upper layers of area V2 illustrating the cytochrome oxidase dense stripes. Although not readily apparent in macaque monkeys, alternating thick and thin dense stripes receive input from layer 4B and the cytochrome oxidase blobs of V1, respectively

oxidase. Unlike the punctate blob-like staining seen in V1, area V2 is characterized by alternating thick and thin stripes of high metabolic activity separated by pale interstripes (see Fig. 3b). Furthermore, these V2 compartments have specific patterns of connections indicating that the parallel processing streams seen in V1 continue into V2. Thus, the magnocellular stream exits V1 from cells in layer 4B to terminate within the cytochrome oxidase thick stripes. Similarly, cells in the V1 blobs terminate in V2 thin stripes, while cell in the V1 interblobs terminate in V2 interstripes. Physiological investigations of receptive field properties using microelectrode and optical recording techniques have indicated that these different V2 compartments contain neurons with different receptive field properties. Thin stripes contain a high proportion of color selective cells that lack orientation selectivity, while thick stripes and pale stripes contain a regular organization of orientation selectivity. These three V2 compartments are repeated approximately 17 times within V2 of each hemisphere and thus each can be viewed as a modular subunit that contains the processing machinery for a small patch of visual space.

The parallel cortical processing streams that begin in V1 are maintained in V2 and continue through the projections to higher cortical areas. The magnocellular stream, exits V2 to continue on to areas V3 and MT. Both of these areas have a high proportion of cells selective for the speed and direction of object motion. These areas provide input to several parietal areas of the 'where' functional stream. The blob and interblob streams of V1 remain largely segregated in their projections to V2 thin stripes and interstripe compartments, respectively. Both of these compartments have

dense projections to area V4 (DeYoe and Van Essen 1985, DeYoe et al. 1994, Felleman et al. 1997). V4 does not contain a modular organization that can be revealed through cytochrome oxidase histochemistry. However, V2 thin stripes and interstripes project to segregated regions within V4 (Xiao et al. 1999). Furthermore, optical recording indicates that color selective and orientation selective cells are clustered in V4 in a pattern that is reminiscent of the segregated inputs from V2 cytochrome oxidase compartments (Ghose and Ts'o 1997).

5. *Topographic Organization of Human Visual Cortex*

Functional magnetic resonance imaging (fMRI) has been used to identify and map the representation of visual space in the occipital cortex of man (DeYoe et al. 1996, Sereno et al. 1995). The general strategy used in these studies was to present counter phased checkerboard stimuli at the vertical and horizontal meridia and different eccentricities, and to record the voxels that show maximum signals that are correlated with the visual stimulus. Since the visual cortex is highly convoluted in man, the visualization of activated voxels is facilitated greatly by mapping onto an unfolded cortical map. Using this approach topographically organized visual areas V1, V2, V3, VP, and V3A have been identified in human occipital cortex (see Fig. 4). As in macaque monkeys, V1 is a large area localized to the calcarine sulcus and its outer border is formed by the representation of the vertical meridian. Area V1 corresponds to area 17 of Brodmann or area OC of Von Economo and Koskinas. Area V2 adjoins V1 at the representation of the vertical meridan and nearly totally surrounds it, except perhaps at the anterior tip of the calcarine sulcus. Area V2 is split along the representation of the horizontal meridian and forms dorsal and ventral halves that represent the lower and upper fields, respectively. Area V2 only occupies the posterior portion of the architectonically defined area 18 of Brodmann. Thus, cytoarchitecture is insufficient and perhaps misleading, in the identification of area V2. Areas V3 and VP border dorsal and ventral V2, respectively, at the representation of the vertical meridian. Unlike macaque monkeys in which these areas are narrow, human V3 and VP are as large as area V2. Based on position, areas V2, V3, and VP together occupy the previously defined architectonic field, area 18. Area V3A is located anterior to V3 in dorsal extrastriate cortex and unlike the adjacent areas V2d and V3, contains a representation of the full visual field. In the ventral cortex, an additional area has been localized using topographic criteria. This area, V4v, is located anterior to area VP and adjoins VP at the representation of the vertical meridian. This region only contains a representation of the upper visual field and no corresponding lower field representation has been found in dorsal cortex, which would correspond to area V4d. It remains unclear whether this reflects a true dorsoventral difference in cortical processing, or whether area V4d has yet to be found. Finally, the human homologue of the MT complex (MT+) has been localized in the lateral occipital sulcus, anterior to area V3A. These studies thus identify a number of topographically organized areas in humans that appear homologous to those seen in macaque monkeys.

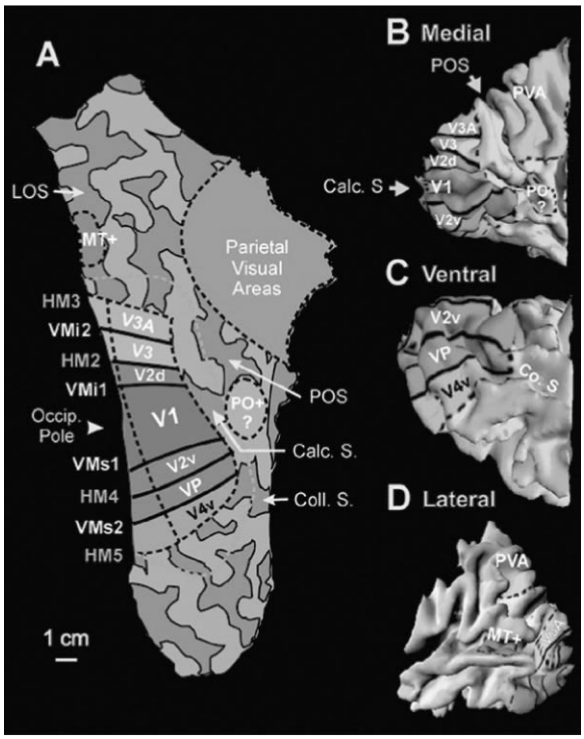

Figure 4
Topography of human visual areas. Location and organization of visual areas in humans derived from topographic mapping using fMRI. A. The borders of identified visual areas are displayed on an unfolded, two-dimensional map of the occipital lobe. B. medial, C. ventral, and D. lateral views of the three-dimensionally reconstructed cortex. DeYoe et al. 1996. Mapping striate and extrastriate visual areas in human cerebral cortex. ©PNAS publisher. Reproduced with permission

6. *Conclusion*

The visual system of the brain consists of a large number of subcortical and cortical components that contribute to different aspects of visual processing. Visual processing begins in the retina where specialized

Bibliography

Barone P, Batardiere A, Knoblauch K, Kennedy H 2000 Laminar distribution of neurons in extrastriate areas projecting to visual areas V1 and V4 correlates with the hierarchical rank and indicates the operation of a distance rule. *Journal of Neuroscience* **20**: 3263–81

Brodmann K 1909 Beitrage zur histologischen lokalisation der grobhirnrinde VI. Mitteilung Die cortexgliederung des Menschen. *J. fur Psychologie und Neurologie* 231–46

Cusick C G, Scripter J L, Darensbourg J G, Weber J T 1993 Chemoarchitectonic subdivisions of the visual pulvinar in monkeys and their connectional relations with the middle temporal and rostral dorsolateral visual areas, MT and DLr. *Journal of Comparative Neurology* **336**: 1–30

DeYoe E A, Carman G J, Bandettini P, Glickman S, Wieser J, Cox R, Miller D, Neitz J 1996 Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Science* **93**: 2382–6

DeYoe E A, Felleman D J, Van Essen D C, McClendon E 1994 Multiple processing streams in occipito-temporal visual cortex. *Nature* **371**: 151–4

DeYoe E A, Van Essen D C 1985 Segregation of efferent connections and receptive field properties in visual area V2 of the macaque. *Nature* **317**: 58–61

Felleman D J, Van Essen D C 1991 Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* **1**: 1–47

Felleman D J, Xiao Y, McClendon E 1997 Modular organization of occipito-temporal pathways: Cortical connections between visual area V4 and visual area 2 and posterior inferotemporal ventral area in macaque monkeys. *Journal of Neuroscience* **17**: 3185–200

Ghose G M, Ts'o 1997 Form processing modules in primate area V4. *Journal of Neurophysiology* **77**: 2191–6

Hilgetag C C, O'Neill M A, Young M P 1996 Indeterminate organization of the visual system. *Science* **271**: 776–7

Kaas J H 1997 Theories of visual cortex organization in primates. In: Rockland R S, Kaas J H, Peters A (eds.) *Cerebral Cortex, Extrastriate Cortex in Primates*. Plenum, New York, Vol. 12, pp. 91–125

Lewis J W, Van Essen D C 2000 Mapping of architectonic subdivisions in the macaque monkey with emphasis on parieto-occipital cortex. *Journal of Comparative Neurology* **428**: 79–111

Livingstone M S, Hubel D H 1984 Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience* **4**: 309–56

Rodieck R W, Watanabe M 1993 survey of the morphology of macaque retinal ganglion cells that project to the pretectum, superior colliculus, and parvicellular laminae of the lateral geniculate nucleus. *Journal of Comparative Neurology* **338**: 289–303

Sereno M I, Dale A M, Reppas J B, Kwong K K, Belliveau V W, Brady T J, Rosen B R, Tootell R B H 1995 Borders of multiple visual areas in human revealed by functional magnetic resonance imaging. *Science* **268**: 889–93

Tootell R B H, Silverman M S, DeValois R L, Jacobs G H 1983 Functional organization of the second cortical visual area in primates. *Science* **220**: 737–9

Ungerleider L G, Mishkin M 1982 Two cortical visual systems. In: Ingel D G, Goodale M A, Mansfield R J Q (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA, pp. 549–86

von Economo C, Koskinas G N 1925 *Die cytoarchitektonik der hirnrinde des erwachsenen mencshen*. Springer, Vienna

Wong-Riley M T T, Carroll E W 1984 Quantitative light and electron microscopic analysis of cytochrome oxidase-rich zones in VII prestriate cortex of the squirrel monkey. *Journal Comparative Neurology* **222**: 18–37

Xiao Y, Zych A, Felleman D J 1999 Segregation and convergence of functionally defined V2 thin stripe and interstripe compartment projections to area V4 of macaques. *Cerebral Cortex* **9**: 792–804

Zeki S M 1990 A century of cerebral achromatopsia. *Brain* **113**: 1721–77

Zeki S M 1991 Cerebral akinetopsia (visual motion blindness): A review. *Brain* **114**: 811–24

D. J. Felleman

Visual Systems: Dorsal and Ventral

1. Our Visual Inheritance

Evolution has provided primates with a complex patchwork of visual areas occupying the posterior 50 percent or so of the cerebral cortex. But despite the complexity of the interconnections among these different areas, two broad 'streams' of projections have been identified in the macaque monkey brain, each originating from the primary visual area (V1): a ventral stream projecting eventually to the inferior temporal cortex (ITC), and a dorsal stream projecting to the posterior parietal cortex (PPC). These regions also receive input from a number of subcortical visual structures, such as the superior colliculus (via the thalamus).

2. Evidence from Private Neuro-Behavioral Studies

In their seminal paper, Ungerleider and Mishkin (1982) argued that the two streams of visual processing play complementary roles in the perception of incoming visual information. According to their account, the ventral stream governs the recognition of objects, while the dorsal stream mediates the localization of those same objects. Apparent support for this idea came from work with monkeys. Lesions of ITC produced deficits in the animal's ability to discriminate visually between objects, but did not affect its performance on a task requiring it to judge the location of a 'landmark.' Conversely, lesions of the PPC produced deficits in performance on the landmark task, but did not affect object discrimination learning. Although the evidence available at the time was consistent with this proposal, recent research fits better with a distinction not between sub-domains of perception, but between

perception on the one hand and the guidance of action on the other (Jeannerod 1997, Milner and Goodale 1995).

3. *Evidence from Visual Neurons in the Ventral and Dorsal Streams*

Neurons in areas such as ITC are tuned to the features of objects, and many of them show remarkable categorical specificity; some of these cells maintain their selectivity irrespective of viewpoint, retinal image size, and even color. These neurons are affected little by the monkey's concurrent behavior, but many are affected by how often the visual stimulus has been presented, and by its association with reward. Evidently the ventral stream is more concerned with the enduring characteristics and significance of objects than with moment-to-moment changes in the visual array.

Neurons in the dorsal stream show quite different properties from those in the ventral stream. Neurons in the PPC are activated by visual stimuli according to the response the monkey makes to those stimuli. For example, some sells respond when the stimulus is the target of an arm reach; others when it is the object of a grasp response; others when it is the target of a saccadic eye movement; others when the stimulus is moving and elicits a slow pursuit eye movement; and still others when the stimulus is stationary and becomes the object of fixed gaze (Jeannerod 1997, Milner and Goodale 1995).

These findings can readily explain why posterior parietal lesions impair landmark task performance: the monkey simply fails to orient toward the landmark. They also explain one of the most striking effects of PPC lesions, namely an inability to accurately reach for a moving or stationary food morsel, and why they fail to shape and orient their hands and fingers appropriately to pick it up. Evidently visual networks in the dorsal stream compute more than just spatial location.

4. *Visuomotor Deficits Following Brain Damage in Humans*

The first systematic description of a patient with bilateral PPC damage was published by Bálint in 1909 (See Harvey 1995). Bálint's patient was unable to accurately reach toward objects with his right hand. In many respects, this disorder, which Bálint called optic ataxia, resembles the problems of the PPC-lesioned monkey. Patients with optic ataxia not only fail to reach in the right direction but also have difficulty orientating their hand and forming their grasp appropriately to pick up target objects (Perenin and Vighetto 1988). Often they are unable to use visual information to form their grip as they reach toward an object, and have to grope for it. For example, they make errors in hand rotation as they try to reach toward and through a large oriented slot. Also, they often fail to show the normal proportional opening of the hand when reaching out to grasp objects of different sizes. Yet despite their visuomotor failures, these patients have comparatively little difficulty in giving perceptual reports of the orientation and location of the very objects they fail to grasp.

5. *Perceptual Deficits following Brain Damage in Humans*

On the other side of the equation, an impairment of ventral stream function occurs in humans who suffer from the condition known as visual form agnosia. These patients not only fail to recognize faces or objects, they fail even to identify simple geometric shapes. One such patient, DF, has been extensively studied during the 1990s. The value of these studies depends on DF's uniquely fortuitous pattern of brain damage, which gives rise to severe visual form agnosia as an isolated symptom, in the absence of any motor impairment or mental deterioration. All other similar patients in the literature either have a less severe form agnosia, or an impaired ability to cooperate or to make precise limb and hand movements.

What DF can do is as important as what she can't. Her attempts to make a perceptual report of the orientation of an oriented slot show little relationship to its actual orientation, whether her reports are made verbally or by manual means. However, when she is asked to insert her hand or a hand-held card into the slot, she moves her hand or the card toward the slot in the correct orientation and inserts it accurately. Indeed her hand begins to rotate in the appropriate direction as soon as it leaves the start position. In short, although she cannot report the orientation of the slot, she can act upon it with considerable skill.

Similar dissociations between perceptual report and visuomotor control were also observed when DF was asked to deal with properties of objects such as their size and shape (Goodale et al. 1991). Thus, she showed excellent visual control of anticipatory hand posture when reaching out to pick up blocks of different sizes that she could not distinguish perceptually. Just like healthy subjects, DF adjusted her finger-thumb separation well in advance of contacting the object, in a perfectly normal and linear fashion in relation to the target width. Yet when she was asked to use her finger and thumb to give a perceptual indication of the object's width, her responses were unrelated to the actual stimulus dimensions, and showed high variation from trial to trial. Functional MRI data have shown that the ventral stream is severely disconnected from visual inputs in DF, so we infer that the calibration of these various residual visuomotor skills must depend on intact mechanisms within the dorsal stream.

Various studies of DF show that she is able to govern many of her actions using visual information of which she has no awareness. But it is clear that this is only true of actions that are directly targeted at the visual stimulus. She cannot successfully use the same visual information to guide an identical response but displaced to another location. Presumably the difference is that the displaced response is an arbitrary or symbolic one. DF seems to depend on a visual processing system dedicated to motor control, which will normally only come into play when she carries out natural goal-directed actions. She has no explicit awareness of the shapes and sizes that she is able to grasp using this system.

There are temporal as well as spatial limits on DF's ability to drive her motor behavior visually. In one study, DF was shown a rectangular block, which was then removed while her eyes were closed, and then she reached out and pretended to grasp it. The preparatory grip size of normal subjects still correlates well with object width even after a 30 second delay. In DF, however, all evidence of grip scaling during her reaches had evaporated even after a delay of 2 seconds. The problem seems to be that a delayed reach is not a natural movement, requiring subjects to imagine the object and then 'pantomime' an appropriate grasp. This strategy would not have been open to DF: she could not generate a visual image of something that she didn't perceive in the first place. Apparently her preserved visuomotor brain can only respond to present or imminent states of the visual world, and disregards past states that may no longer be relevant.

Recently, a complementary 'improvement' in both reaching and grasping accuracy has been recorded in two bilateral optic ataxic patients after a 5-second delay (Milner and Dijkerman 2001). Presumably these patients were able to throw off the dominance of the damaged dorsal stream under the delay condition, allowing them to make use of their better-preserved ventral system.

6. *Operation of the Two Streams in Normal Vision*

It may be that like DF, healthy observers also control their actions using visual information that is not present in their awareness. In providing such visual guidance the dorsal stream may act pretty much in isolation, and independently of any acquired 'knowledge base.' Our implicit assumptions about the perceptual stability and constancy of the world are based on such stored knowledge. These assumptions sometimes cause us to misjudge what we see, as when a shift in a large frame makes a stationary spot inside it seem to shift in the opposite direction. In contrast, the dorsal system, by and large, is not deceived by such spatial illusions. Instead, it directs our saccadic eye movements and hand movements to where a target really is, instead of where our perceptual system tells us it is (e.g., Bridgeman et al. 1997). Similarly, under appropriate circumstances geometric illusions can be seen to affect visually-guided reaching and grasping (Haffenden and Goodale 1998) far less than they affect perceptual judgments. In other words, you may perceive an object as bigger than it really is, but you open your finger-thumb grip correctly when reaching for it.

It seems that the processing accomplished by the ventral stream both generates, and is in turn informed by, stored abstract visual knowledge about objects and their spatial relationships. It may be that these 'top-down' influences are what render our perceptual representations accessible to awareness (Lamme and Roelfsema 2000). This would fit with the idea that coded descriptions of enduring object properties, rather than transitory egocentric views, are needed for mental manipulations like the planning of action sequences and the mental rehearsal of alternative courses of action.

7. *Integration of Visual Function*

Although the dorsal and ventral streams have separate functions, there are multiple connections between them, and indeed adaptive goal-directed behavior in humans and other primates must depend on a successful integration of their complementary contributions. Thus the execution of a goal-directed action might depend on dedicated control systems in the dorsal stream, but the selection of appropriate goal objects and the action to be performed depends on the perceptual machinery of the ventral stream. At the level of visual processing, however, the visuomotor modules in the primate PPC function quite independently from the ITC mechanisms generating perception-based knowledge of the world. Only this latter, perceptual, system can provide suitable raw materials for thought processes to act upon. In contrast, the other is designed to guide actions purely in the 'here and now,' and its products are consequently useless for later reference.

See also: Occipital Lobe (Visual Cortex): Functional Aspects; Vision, High-level Theory of; Vision, Low-level Theory of; Visual Perception, Neural Basis of; Visual System in the Brain

Bibliography

Bridgeman B, Peery S, Anand S 1997 Interaction of cognitive and sensorimotor maps of visual space. *Perception and Psychophysics* **59**: 456–69

Goodale M A, Milner A D, Jakobson L S, Carey D P 1991 A neurological dissociation between perceiving objects and grasping them. *Nature* **349**: 154–6

Haffenden A M, Goodale M A 1998 The effect of pictorial illusion on prehension and perception. *Journal of Cognitive Neuroscience* **10**: 122–36
Harvey M 1995 Translation of 'Psychic paralysis of gaze, optic ataxia, and spatial disorder of attention' by Rudolph Bálint. *Cognitive Neuropsychology* **12**: 261–82
Jeannerod M 1997 *The Cognitive Neuroscience of Action*. Blackwell, Oxford, UK
Lamme V A F, Roelfsema P R 2000 The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences* **23**: 571–9
Milner A D, Dijkerman H C 2001 Direct and indirect visual routes to action. In: De Gelder B, De Haan E H F, Heywood C A (eds.) *Varieties of Unconscious Processing: New Findings and New Comparisons*. Oxford University Press, Oxford, UK
Milner A D, Goodale M A 1995 *The Visual Brain in Action*. Oxford University Press, Oxford, UK
Perenin M-T, Vighetto A 1988 Optic ataxia: a specific disruption in visuomotor mechanisms. I. Different aspects of the deficit in reaching for objects. *Brain* **111**: 643–74
Ungerleider L G, Mishkin M 1982 Two cortical visual systems. In: Ingle D J, Goodale M A, Mansfield R J W (eds.) *Analysis of Visual Behavior*. MIT Press, Cambridge, MA

A. D. Milner

Visualization: Representation in Science

Scientific knowledge is represented in many forms: spoken and printed words, graphs, formulae, tables, drawings, engravings, diagrams, photographs, charts, maps, models, diorama, demonstrations, and many other more specialized renderings. Visual and graphic renderings are deployed at every stage of scientific production, from initial data collection through final publication. Often, diverse forms of rendering are combined in complex assemblages. For example, figures in scientific texts frequently incorporate mathematical, verbal, and pictorial expressions in a single frame, so that cross-referential relations are at least as significant as the particular graphic forms and formats. Since the mid-1970s, a growing body of research in science and technology studies (S&TS) has examined the production of visual evidence, and this has led to a radical revision of standard epistemological conceptions of observation and representation.

Two related senses of the term 'visual representation' are important for current historical, literary, and ethnographic studies of science. A passive or static sense of the word 'representation' refers to an image, inscription, or sign that depicts, signifies, or symbolizes something else. A large body of recent studies of the forms, interpretive meanings, and cultural implications of particular representations pursue this sense of the concept. A more active or dynamic sense of the word refers to the practice of making images and the action of doing representing. The term 'visualization' signals these more active connotations. Studies of visualization in laboratory and field science have contributed some of the most influential and challenging ideas in the S&TS literature, and these are emphasized in the remainder of this article.

1. *Representing and Visualizing*

The study of visual representation in science crystallizes lines of work from diverse academic disciplines. In the 1970s and early 1980s historians of science began to pay close attention to nonverbal representations and nonpropositional forms of reference (Rudwick 1976). At about the same time art historians took greater interest in the broad array of what Elkins (1995) calls 'images that are not art': pictures and models that are crafted as products and by-products of scientific and medical work, but which are rarely considered as art (Edgerton 1976, 1991, Alpers 1983). This led to a convergence of scholarly attention on the role of artisan skill, craft technique, and visual technology in the scientific revolution, and on the continuing place of visual arts and crafts in day to day laboratory work.

1.1 *Visual Representation*

Studies of visual representation in science include analyses of published and unpublished images, and of the historical crafts and technologies responsible for their production. Currently, there are numerous disciplinary and cross-disciplinary approaches to visual representation. Representation is simultaneously a philosophical, historical, and cultural concept. Social, cultural, and literary studies of science include a bewildering array of historicist, semiotic, feminist, cognitivist, epistemological, pragmatist and other lines of interpretation. All of these appear, often in combination, in studies of visual representations. As many cultural and literary studies have emphasized, visual representations do not simply resemble natural objects: they express normative assumptions about natural and cultural worlds; they incorporate historically and culturally specific artistic conventions; and they mediate political and cultural relations of patronage, commodity exchange, and subjugation (Newman 1996, Haraway 1997).

1.2 *Visualization*

When considered in a more active sense, representation includes a broad array of practices and technologies for making, producing, and shaping images and relations among images. In the S&TS literature, this active sense of representation is signaled by the term 'visualization' (Latour 1990). This means more than 'seeing' or 'depicting,' as it refers to any of the broad array of research tools and practices for making natural objects visible: practices for uncovering hidden

features; magnifying objects and enhancing analyzability; facilitating measurement and comparison; and making things publicly witnessable, intelligible, and accountable. Visualization includes efforts to convince as well as to exhibit. Consistent with a programmatic aim to investigate the content of natural science knowledge (Bloor 1976), ethnographers, ethnomethodologists, and social-historians closely examine the situated practices, inscription devices, and textual forms through which scientific data and pictorial representations are exhibited, analyzed, stabilized, objectified, and disseminated (Latour and Woolgar 1979, Shapin and Schaffer 1985, Lynch and Woolgar 1990). These studies stress the significance of visualization as a crucial aspect of a continuing scientific revolution. This revolution involves the incessant invention, adaptation, and use of novel instrumental and literary technologies for revealing invisible objects, expressing abstract relations, and disseminating knowledge.

Renaissance innovations in print and engraving technology, together with novel optical devices and artistic conventions, enabled natural philosophers to inscribe and reproduce 'faithful' representations of observational experience. Latour (1990) uses the expression 'immutable mobiles' to capture the dual sense of stability and mobility attributed to imprinted representations. Scientific matters of fact, while still controlled by those who had the means to purchase, build, and operate equipment like the telescope, microscope, and Robert Boyle's air pump were communicated widely in expanding networks of co-investigators, patrons, and other 'virtual witnesses' (Shapin 1984, Shapin and Schaffer 1985). In most cases, virtual witnesses who read Boyle's experimental reports were in no position to replicate the original experiments. Boyle tried to get around this communicative problem by composing prolix descriptions illustrated by engravings of observed phenomena and laboratory instruments. This literary technology enabled Boyle to disseminate experimental knowledge widely, while sparing all but a few well-equipped adepts from the difficulties and uncertainties involved in actually trying to replicate the experiments. The dissemination of printed and engraved representations facilitated, but did not guarantee, the stability (or 'immutability') of the knowledge in question, since printing, engraving, and publication compounded the contingencies of knowledge production, enlarging the circle of potential disputants as well as adherents. The stability and mobility of images were no better than the instruments and persons involved in their production and distribution. Instruments, like the microscope, and persons, like engravers, printers, and publishers, were subject to variable degrees of trust and distrust (Johns 1998). Consequently, whether located in a gradual historical process or in the temporal production of a late-modern laboratory project (Knorr-Cetina and Amann 1990), 'immutable mobiles' depend upon incessant efforts to control sources of distortion, disruption, and misinterpretation at every point of production and distribution.

2. *Image and Object*

Numerous S&TS works have examined the historical, conceptual, and interpretive problems associated with the production of visual images, and they have also described the communicative and community-building work facilitated by such images. By elucidating the practical handiwork and collective labor process hidden behind the grand polarity of image and object, these studies furnish detailed cases that challenge existing conceptions of the scientific image, the natural world, and the relations between image and object.

Studies of visualization address classic philosophical questions about observation, representation, and objectivity, but they give these questions an original twist. Instead of addressing these questions abstractly, S&TS researchers investigate specific cases in which scientists encounter practical dilemmas and devise provisional solutions. Hanson (1958, p.4) furnishes an ideal-typical version of such a dilemma:

> Consider two microbiologists. They look at a prepared slide. When asked what they see, they may give different answers. One sees in the cell before him a cluster of foreign matter: it is an artifact, a coagulum resulting from inadequate staining techniques. This clot has no more to do with the cell, in vivo, than the scars left on it by the archaeologist's spade have to do with the original shape of some Grecian urn.

In this example, the observational instrument is both a material thing and a transparent medium through which other things appear. Depending upon what the microscopist sees on the slide, the instrument will be judged a source of 'subjective' illusion or a means for revealing an 'objective' cellular structure. One of Hanson's microscopists sees an artifact, but in the particular research context it is less like an artifact uncovered at an archaeological site (a Grecian urn) and more like the unwanted traces on that artifact left by the spade that uncovered it. This example suggests that the distinction between objective structure and artifact can be contentious and difficult to resolve. In their restricted domain, the microscopists are faced with a standard philosophical question: 'How do you know that the image stands for a real object?' A nonstandard philosophical answer is given by Hacking (1983). Hacking rejects the picture of a static image standing for a natural object, and points to the practical interventions performed in the laboratory. These interventions include crosschecks, calibrations, and probes that compare and contrast the outputs from different instruments and techniques. Convictions about reality inform, and develop from, a microscopist's competent handling of the laboratory instruments and materials.

Hacking remains focused on the philosophical problem of knowing natural reality, but he places the solution in the competent hands of the microscopist. The classical confrontation between object and subject is supplemented and mediated by instruments and skills. They are part of the means of perception, and like our eyes and our minds they sometimes are deemed untrustworthy, as sources of 'subjective' illusion. In microscopy and other laboratory sciences, correct judgment involves a mastery of technique as much as clear eyes and an educated mind. A whole chain of preparatory actions precedes and sets up the dramatic confrontation between microscopist and cell. These preparatory practices include cell culture, fixation, staining, counterstaining, thin-sectioning, mounting on a slide, and many other techniques (Lynch 1985a). If the microscopist sees a 'foreign' body—an artifact—in the cell, one or another of the preparatory techniques may be held responsible for the illusion. What the microscopist sees (or does not see) is thus bound up with a labor process that includes much more than looking down the tube of the microscope and focusing the eyes on an object. Hacking treats instrumental design and technique as sources of practical solution to philosophical questions about the reality of 'invisible' entities, whereas ethnographers and social historians tend to defer the question of reality while examining situated practices for enhancing or undermining the credibility of visual evidence.

2.1 Chains of Industry and Credibility

The S&TS treatment of visualization as temporal, embodied, material labor opens up a domain of investigation that is not covered by mentalistic and perceptual accounts of scientific observation. It ties in with a sociological, rather than cognitive or idealist, legacy, by treating the appearance of the scientific object as being bound up in a temporal process, typically involving many hands in a division of labor. Marx famously credited Feuerbach with having stood Hegel on his feet, meaning that he placed Hegel's abstract dialectic on a firm footing in historical materialism. Less famously, Marx suggested a related way to bring history of science down to earth:

> …historiography pays regard to natural science only occasionally, as a factor of enlightenment and utility arising from individual great discoveries. But natural science has invaded and transformed human life all the more *practically* through the medium of industry; and has prepared human emancipation, however directly and much it has had to consummate dehumanization. *Industry* is the *actual*, historical relation of nature, and therefore of natural science, to man (Karl Marx, *Economic & Philosophic Manuscripts* 1844).

A further inversion from feet to hands can be inspired by studies of visualization in science. For Marx, industry, rather than history of ideas, embodies Western 'man's' actual, historical relation to natural science. For S&TS, natural science not simply the historical basis for modern industry and its relations of production. The laboratory's own labor process is itself a site of collective handiwork facilitated by instruments and exhibiting variable degrees of complexity, automation, and historical development. Marx speaks of the 'working up' of raw materials into the form of a product: an artifact that embodies human purposes and provides a résumé of the labor that produced it: 'The product is after all but the summary of the activity of production.' Among the products of laboratory work are data—carefully framed and emplotted observations and measurements which stand to nature much in the way that processed food does to raw ingredients. Scientists sometimes speak of 'raw data' collected in the field or from electronic instruments, and they distinguish such data from processed images that result from 'working up' the data into presentable form.

Likening the visualization of scientific data to a manufacturing process may seem to miss an essential point: that the object of scientific investigation is not an artifact, even though it may be difficult to establish the difference between the two in particular cases. Described in the vernacular of laboratory science, the object is found, or found out. It is not made, and it is distinguished from both deliberately constructed tools and inadvertently produced scratches, residues, and deformations. Close studies of day to day laboratory work have furnished many examples of how research materials are cultivated, handled, and reconfigured, but to conclude, for example, that all cellular structures are artifacts collapses a vernacular distinction that provides laboratory work with a fundamental task and tension. Such a conclusion also invites the sterile polarity and tired polemics of the realist–contructivist debate. Latour tries to avoid the polemics by devising a neologism that straddles the polarity. Where, according to Marx, the industrial product becomes a 'fetish'—a commodity whose exchange value hides the work of its production —according to Latour (1999, p. 272) the laboratory product becomes a 'factish'—a fact whose solidity and epistemic significance hides the day to day work of establishing and maintaining that solidity and significance. For Latour, the 'factish' is never independent of its manufacture, but this does not reduce it to an empty screen for projecting arbitrary beliefs. The robust credibility of 'fact' remains in place, not because it exists in an eternal space remote from human activity, but because of the way it is fashioned.

Latour's creative vocabulary is less instructive than his examples. In an ethnography of a Brazilian field project, Latour (1995) closely examined the series of practices through which a field site was made graphically analyzable. The project was concerned with a boundary between rainforest and savanna. The team

of researchers addressed the question of whether the forest was advancing into or retreating from the grassland. As scientific projects go, this one was simple. It involved only a few researchers, a down-to-earth problem, no complex equations, and relatively commonplace technology. Nevertheless, it involved a complicated series of representational practices for marking out a field site into a systematic grid, selecting and tagging plant specimens, collecting and analyzing soil samples, plotting data on paper graphs, and amassing files for every stage of the project. The field site, the tagged specimens, the binned and color-coded soil samples, and countless other materials used in the project were hybrid objects. None of these things were purely natural or artificial, concrete or abstract, material or symbolic, real or imaginary, but a great deal of care was taken to preserve selected aspects of specimens, keep track of identities, and cross-check measurements and records. The project resulted in a publication in which a graphic figure represented the forest–savanna boundary, and the data collected during the project were used to support conjectures about the movement in question. Again, this was not a massive or powerful project, and it resulted in tentative findings, but it served well enough to illustrate an important point about the credibility of the project's evidence. The researchers staked the credibility of their data on a complex network of accountability in which no single representation corresponded to an independent natural object, and yet all of their materials, in different ways, contained, encoded, traced, or preserved a tenuous relation to the natural field site. The credibility of the production was more like the strength of a rope, weaving together separate strands, and less like the compelling realism of a photographic image. Perhaps another example will help elucidate this point.

The police in the United States use the expression 'chain of custody' to refer to the continuity of evidence as it is collected and transported from one place to another. So, for example, blood evidence collected at a crime scene by a police employee or pathologist is placed in a tube and transported to one or more labs, where it is analyzed with DNA profiling methods. Such evidence typically travels a complicated route, and it is transformed and repackaged several times in the process. The potential consequences of such evidence, and the possibility of courtroom challenge, has led police organizations and forensic labs to use a whole series of administrative procedures for checking, cross-checking, and validating the identity of evidence samples. Standardized forms, tamper-evident bags, stickers with bar codes on them, automated scanners, filing systems, laboratory protocols for recording and 'witnessing' the movement and handling of evidence, and many other mundane bureaucratic artifacts create an administrative envelope that protects the credibility of evidence against courtroom challenge. In addition, courtroom presentations make use of some of the techniques of visualization that have been documented in scientific labs and field studies (Goodwin 1994). The credibility of DNA profile evidence stands only as long as the devices of visualization and their administrative envelope resists challenge. None of the forms, bar codes, and protocols that make up the chain of custody guarantee that it will resist adversary efforts to expose significant gaps. Even an elaborately documented and impressively consistent chain of custody does not guarantee the truth of the evidential result against all possibility of error, mishap, fraud, and conspiracy, but when the numerical indices, signatures, and analytic results tell a story consistent with public and judicial conceptions of good 'scientific' practice, they are more likely to resist adversary challenge (Lynch and Jasanoff 1998)

It often seems natural to employ metaphors of vision, perception, and cognition when discussing scientific observation and representation, but the analogy with an industrial labor process suggests a more material, and less ocular and mental, conception of visualization and its natural products. Consistent with the paradigm of the assembly line, a scientific project is organized into a chain of tasks performed by a series of technicians. Specimen materials are collected, selected, prepared, rendered, and reorganized to expose, enhance, and encode relevant features, to upgrade analyzability, and to frame data with instrumental and graphic fields (Lynch 1985b, Latour 1995, Knorr-Cetina 1999). Nature and culture pervade every link in the chain, and there is never a pure confrontation between and object and its representation.

See also: Knowledge Representation; Visual Anthropology; Visual Images in the Media; Visual Methods in the Social Sciences

Bibliography

Alpers S 1983 *The Art of Describing: Dutch Art in the Seventeenth Century*. University of Chicago Press, Chicago

Bloor D 1976 *Knowledge and Social Imagery*. Routledge and Kegan Paul, London

Cambrosio A, Jacoby D, Keating P 1993 Ehrlich's 'beautiful pictures' and the controversial beginnings of immunological imagery. *Isis* **84**: 662–99

Collins H M 1985 *Changing Order: Replication and Induction in Scientific Practice*. Sage, London and Beverly Hills

Drake S 1975 The role of music in Galileo's experiments. *Scientific American* **232**: 98–104

Edgerton S Y 1976 *The Renaissance Rediscovery of Linear Perspective*. Harper and Row, New York

Edgerton S Y 1991 *The Heritage of Giotto's Geometry: Art and Science on the Eve of the Scientific Revolution*. Cornell University Press, Ithaca, NY

Elkins J 1995 Art history and images that are not art. *The Art Bulletin* **77**: 553–71

Giedion S 1948 *Mechanization Takes Command: A Contribution to Anonymous History*. Oxford University Press, Oxford, UK

Gilbert G N, Mulkay M 1984 *Opening Pandora's Box: An Analysis of Scientists' Discourse*. Cambridge University Press, Cambridge, UK
Goodwin C 1994 Professional vision. *American Anthropologist*. **96**: 606–33
Hacking I 1983 Microscopes. In: Hacking I *Representing and Intervening*. Cambridge University Press, Cambridge, UK, pp. 186–209
Hanson N R 1961 *Patterns of Discovery*. Cambridge University Press, Cambridge, UK
Haraway D J 1997 *Modest Witness@Second Millenium. FemaleMan© Meets OncoMouse*™. Routledge, New York
Johns A 1998 *The Nature of the Book: Print and Knowledge in the Making*. University of Chicago Press, Chicago
Knorr-Cetina K 1999 *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press, Cambridge, MA
Knorr-Cetina K, Amann K 1990 Image dissection in natural science inquiry. *Science Technology & Human Values* **15**: 259–83
Latour B 1990 Drawing things together. In: Lynch M, Woolgar S (eds.) *Representation in Scientific Practice*. MIT Press, Cambridge, MA
Latour B 1995 The 'pédofil' of Boa Vista: A photo-philosophical montage. *Common Knowledge* **4**: 144–87
Latour B 1999 *Pandora's Hope: Essays on the Reality of Science Studies*. Harvard University Press, Cambridge, MA
Latour B, Woolgar S 1986 *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, NJ
Lynch M 1985a *Art and Artifact in Laboratory Science*. Routledge and Kegan Paul, London
Lynch M 1985b Discipline and the material form of images: An analysis of scientific visibility. *Social Studies of Science* **15**: 37–66
Lynch M, Jasanoff S (eds.) 1999 Contested identities: Science, law and forensic practice. *Social Studies of Science* **28**, 675–868
Lynch M, Woolgar S (eds.) 1990 *Representation in Scientific Practice*. MIT Press, Cambridge, MA
Newman K 1996 *Fetal Positions: Individualism, Science, Visuality*. Stanford University Press, Stanford, CA
Rudwick M 1976 The emergence of a visual language for geological science, 1760–1840. *History of Science* **14**: 149–95
Shapin S 1984 Pump and circumstance: Robert Boyle's literary technology. *Social Studies of Science* **14**: 481–520
Shapin S, Schaffer S 1985 *Leviathan and the Air Pump*. Princeton University Press, Princeton, NJ

M. Lynch

Vocabulary Acquisition, Psychology of

1. Why Study Vocabulary Acquisition?

Having a vocabulary means knowing words: how to pronounce them, what they mean, the different forms they can take, and how they can be used in sentences. Children are phenomenal word learners—by the age of six, English-speaking children are estimated to have learned over 10,000 words; undoubtedly, children learning other languages are equally adept. This feat of rapid vocabulary acquisition is remarkable because it is not as straightforward as it might seem. Clearly children learn the words they hear in their environment, but they do not simply imitate everything they hear; imitation cannot explain how they extract words from the speech stream nor how they relate sounds (or signs) to meanings. Children often attain knowledge of a word piecemeal. They might know what kind of word it is (e.g., that it can take certain forms like -s or -ing or that it is allowed in certain sentence positions) before knowing what it means, or they might know its meaning without knowing how to pronounce it properly. Thus, studying how children acquire words—how they represent the sounds, meanings, and formal characteristics of words—is a part of the study not only of language development but also of the human ability to extract information from the environment over time, organize it, and use it. Because vocabulary acquisition is a life-long phenomenon, with estimates of adult vocabularies ranging over 100,000 words, the insights gained studying children's vocabulary acquisition may help in understanding why vocabulary knowledge remains a robust skill among the elderly, word-finding problems notwithstanding.

2. The Beginnings of Word Learning

By the time children start to use words, they have already exercised many of their inborn capacities in the service of language learning. Research shows that sensitivity to sound and memory for sound begin even before birth, and that the capacities needed to segment the speech stream are in place early (see Hoff 2001 for details).

2.1 What Infants Bring to Word Learning

Recent advances in assessing the capacities of infants have revealed remarkable abilities that are recruited to the task of vocabulary acquisition (see Werker and Tees 1999). Newborn infants can detect minimal differences between sounds, and the differences they detect most easily are those occurring in the sound systems of the world's languages. Newborns show remarkable sensitivity to the acoustic cues that specify word boundaries: they can discriminate between a two-syllable sequence extracted from a single word and such a sequence extracted from two separate words. Using acoustic cues, they also can discriminate the two basic kinds of words found in languages, i.e., content or 'open class' words like nouns and verbs and grammatical or 'closed class' words like prepositions and articles. Such sensitivities, plus an ability to keep track of statistical differences in the occurrence of different patterns of acoustic cues, enable infants by 8 months to segment a speech stream into words.

2.2 Experience and Development in Early Word Acquisition

During the first year, infants' general capacity to discriminate speech sounds becomes more attuned to the sound system of the specific language in their environment. By 10–12 months, infants discriminate best the syllabic sound differences of their native language. Shortly thereafter, they begin associating sound strings with meaning, but at the same time their ability to distinguish between words on the basis of sound becomes more holistic and in some sense less discriminating. First word productions are different from later ones in that the first words often label whole scenes or events, or they are used to make requests; one 10-month old would call, 'Dada mi(lk)' from his crib when he wanted his morning bottle.

There are many similarities across children in the first 50 or so words produced: toddlers have an array of labels for common items and familiar people in their environment, along with words like 'more,' 'up,' and 'all gone,' that indicate changes of state and place. About half of all children appear to undergo a vocabulary spurt around 18 months of age, learning words at a rapid rate. Other children may show a more gradual increase in the rate of their word acquisition or experience a later spurt.

Nouns predominate over verbs in the early vocabularies of children learning English. This tendency toward early noun predominance was thought to be universal on cognitive grounds: relating sounds to real-world objects should be easier than grasping the relational nature of verb meanings (Gentner 1982). Although this may be true, which languages caregivers speak and what they, as members of their culture, find important to talk about also play a role in early vocabulary. For example, children learning Asian languages that typically place verbs in perceptually salient sentence-final positions tend to have a higher ratio of verbs to nouns in their early vocabularies than do children learning languages like English (Tardif 1996).

3. Working on Words

By 2 years of age, children typically are adept word learners, using just about everything they know to help them map sounds to meaning. Given how rapidly vocabulary grows even in the early years, it is hard to see how it could be otherwise.

3.1 Constraints on Word Meaning

Because in language meaning is arbitrarily related to sound, there are an infinite number of hypotheses learners could make about the meaning of any string of sounds they hear. Vocabulary would never be acquired if the learner had no way of constraining possible meanings. Researchers have proposed various constraints that can be thought of as learning strategies or default hypotheses that children use to make first guesses at the mappings between sound and meaning. One strategy suggested for infants is that they expect speakers to be talking about what they are looking at. Beliefs that a new label should refer to a whole object and not part of an object, that objects should have only one label, and that objects with the same label are alike in some fundamental way are examples of other suggested strategies for toddlers. Moreover, there is considerable evidence that preschoolers use what they know about language structure to inform their word learning (see Bloom 1994 and Gleitman and Landau 1994 for details and discussions of these sorts of constraints).

3.2 Vocabulary Organization

The evidence that children use information about language structure to help them acquire word knowledge is consistent with other data suggesting that children consider word–word relations and not just word–world relations as they acquire vocabulary. For example, even before they know what a particular color term means, or even which perceptual dimension the word *color* refers to, they regularly answer questions about the color of objects with color terms, albeit incorrect ones (Shatz and Backscheider 1999). They show similar behavior with letter- and number-term categories. At least by the age of two, then, children are not just associating labels with objects or events, but they are relating words to one another, creating word categories based on common grammatical properties and on patterns of use in discourse.

3.3 Practice, Practice, Practice

The numerous reports of young children practicing their language skills (e.g., Kuczaj 1983) provide more evidence supporting the idea of early organization of word–word relations. One 20-month-old was observed reciting to himself a triad of antonym pairs, 'in-out, up-down, hi-bye.' At the age of 33 months, the same child, after talking to himself about family members' possessions, spewed out a string of pronouns, 'I, me, my, you, your, he, him, her, she, we, ours' (Shatz 1994). Thus, throughout their early years, children reveal in their private speech that they think about various kinds of word-word relations.

4. Vocabulary Acquisition Later in Life

4.1 Sources of Innovation

Vocabulary is the aspect of language that continues to show striking development beyond the early school years. This is because not only does an individual's

knowledge base grow, but there are many sources of vocabulary innovation in a speech community that an individual draws from throughout the life-span. New words or new meanings for old words are often introduced by groups such as adolescents eager to create a separate identity through language use, and the innovations often are adopted over time by the larger community. Other sources of vocabulary growth include borrowings from other languages, phrases used in public life or literature that take on idiomatic meaning and words created to describe new technologies. Little is known about how vocabulary grows throughout life, but it is likely that the patterns children display of partial knowledge about meaning organized along with facts about grammatical and discourse uses also characterize adult states of vocabulary knowledge as new words are acquired.

4.2 Vocabulary and the Brain

Although the rapid vocabulary growth of childhood undoubtedly slows over time, even seniors acquire new words, and vocabulary knowledge, like much linguistic knowledge, remains a robust skill among the elderly (Schum and Sivan 1997). Age-related difficulties in vocabulary use, such as word finding, word substitutions, and word loss can often be ascribed to particular brain insults like cerebral-vascular accidents resulting in area-specific brain damage. Others are likely related to decline in more general cognitive abilities like memory or speed-of-information processing (see Tun and Wingfield 1997). The occurrence of specific deficits for certain kinds of words (e.g., nouns) supports the view that words are indeed organized into a variety of categories, some of which are based on quite abstract characteristics.

5. Avenues of Future Research

Future research will build on the discoveries of infants' discrimination and segmentation abilities to address questions about the nature of early representations of sound-meaning relations. Why do children show decreased sensitivity to the specific sound characteristics of words just when they have recognized them as words? Why does a 2-year-old who has said 'crain' for *train* not register their own error but does so when an adult imitates 'crain'? Other questions about early vocabulary acquisition concern the development of word understanding and use. Given that early vocabulary use reveals some organization based on word–word relations, how do children integrate information about word–world and word–word relations over time into comprehensive representations?

Finally, future research must address important questions about the generality of the picture of vocabulary acquisition presented here. What is the relation between early and later vocabulary acquisitions? Can this view of vocabulary organization as one involving integrated aspects of sound, meaning, form, and use profitably be related to difficulties with word usage in aging or disordered populations? Second, can this account of vocabulary acquisition be appropriately applied to instances of acquisition in more agglutinative languages, i.e. those in which words themselves are made up of many more different pieces of meaning than they typically are in a language like English? The question of the generality of this account of vocabulary acquisition will undoubtedly require more than a one-word answer.

See also: First Language Acquisition: Cross-linguistic; Language Acquisition; Language Development, Neural Basis of

Bibliography

Bloom P 1994 *Language Acquisition Core Readings*. MIT Press, Cambridge, MA

Gentner D 1982 Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In: Kuczaj II S A (ed.) *Language Development: Syntax and Semantics*. L. Erlbaum Associates, Hillsdale, NJ

Gleitman L, Landau B (eds.) 1994 *The Acquisition of the Lexicon*. MIT Press, Cambridge, MA

Hoff E 2001 *Language Development*, 2nd edn. Wadsworth, Belmont, CA

Kuczaj II S A 1983 *Crib Speech and Language Play*. Springer-Verlag, New York

Schum R L, Sivan A B 1997 Verbal abilities in healthy elderly adults. *Applied Neuropsychology* **4**: 130–4

Shatz M 1994 *A Toddler's Life*. Oxford University Press, New York

Shatz M, Backscheider A 1999 *Toddlers Create Lexical Domains: The Case for Color*. Unpublished ms University of Michigan, Ann Arbor, MI

Tardif T 1996 Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental Psychology* **32**: 492–504

Tun P A, Wingfield A 1997 Language and communication: Fundamentals of speech communication and language processing in old age. In: Fisk A D, Rogers W A (eds.) *Handbook of Human Factors and the Older Adult*, Academic Press, San Diego, CA pp. 125–49

Werker J F, Tees R C 1999 Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology* **50**: 509–35

M. Shatz

Vocational Education and Training

We define 'vocational education,' as any set of learning activities aimed at developing individuals for personally and socially significant life roles; and 'training' as a confined aspect of vocational education, namely

any set of learning activities aimed at equipping individuals with the knowledge needed for pre-defined tasks or roles. That is, vocational education and training is not just a particular sector of education, or particular activities preparing people for work, but also any other education, which recognizes explicitly the centrality of individuals' life purposes (including work) and their social significance in an individual's construction of meaning.

Our definition adopts Dewey's (1916) inclusive idea of *vocation*, as personally significant life activities. It is consonant with the Latin derivation of the term, *callings*, and the Greek idea that all people have a *telos*. Thus, a vocation might involve such disparate pursuits as motor mechanic, doctor, parent, rock-star, athlete, or street vendor. Some societies afford higher social status to some vocations, especially those taught in universities (e.g., medicine) and call them professions, leading to the professional/vocational distinction.

1. Classical Constructions of Vocational Knowledge

Differentiations of kinds of knowledge (and their acquisition) can be traced to the ancient Greeks who divided society into:

> … those who had to labor for a living [slaves, artisans and women] and those who were relieved from this necessity [free class worthy of liberal education] (Dewey 1916, p. 251).

Aristotle differentiated 'five ways in which the soul arrives at truth' (Thomson 1976, p. 206)—art or technical skill (*techne*); science or scientific knowledge (*episteme*); prudence or practical wisdom (*phronesis*); intelligence or intuition (*nous*); and wisdom (*sophia*). He regarded art and technical skill as concerned with *making* or *production*, but as essentially different from prudence or practical wisdom (*phronesis*), which is concerned with deliberating 'rightly' what is good and advantageous, thereby constituting *action* rather than production. Thus, vocational knowledge is often constructed in opposition to other knowledge, based on its supposed *technical* or *applied* nature. This segregation from 'general' knowledge reflects a social segregation of culture and utility, and of class. Separation of education for *techne* from more valued education (so-called general or 'liberal' education of the 'person') underlies the organization of education in Western industrial countries, with polarization of the constructions of liberal/academic, and vocational knowledge.

Based on classical Greek ideas, later authors have differentiated knowledge in two ways: (a) theoretical vs. applied knowledge, and (b) a knowledge that enables control over causes vs. knowledge that enables 'right' action (to secure what is appropriate or good).

1.1 Theoretical vs. Applied

This distinction is sometimes caricatured as a division of head and hand. Ryle (1949) cautioned against such conceptualizations of thinking that attributed to knowledge of 'how' a form and structure additional to its intrinsic nature, arguing that the capacity to demonstrate expertise was a single phenomenon, not driven by separate internal processes. Separation of the theoretical and the applied in vocational education is exemplified in formal college or school supplementation of on-the-job apprenticeship/traineeship training in Western society. However, the German Dual Apprenticeship system attempts to develop both theory and practice in both school and work settings. Institutional separation of vocational education from 'general' schools and universities also reflects this differentiation of theoretical and applied concerns or of culture and utility. Consider also the construction of educational pathways, which involve personal choice, but at the same time labor market segmentation and class allocation.

1.2 Technical Control vs. 'Right' Action

Critical theorists emphasize the distinction between 'technicist' knowledge for technical control and 'practical' knowledge for 'right' action. Such concepts can generate views of occupational roles as atheoretical and non-moral; and vocational education as separable from individual and societal needs. They help legitimate movements like competency-based training, which separates technical mastery and control for workplace performance from individual construction of meaning and the critical development of society.

Popkewitz (1998) has argued that both Dewey and Vygotsky were social constructivists bringing 'new democratic political rationalities into the governing of individual conduct' (p. 535). They, in a period of modernity, saw that 'social sciences [education] would not only provide a cognitive knowledge but also discipline the capabilities, values, dispositions, and sensitivities through which individuals problematized their participation in the world' (p. 537). They compared their 'problem-solving individual' with the images and constructions of social determinism (Popkewitz 1998) as the basis of state provision of vocational education and training, i.e., self-empowering education for democratic participation or technical/market control.

2. Psychological Ideas of Training

The psychological idea of training has its origins in early behaviorist research on classical and operant conditioning. Often training referred to the shaping of animal behavior, construing systematic mastery of

confined and pre-determined tasks. It places behavior in opposition to internal cognitive processes and externally determined goals in opposition to individual needs. This psychological idea of training has been extended to all work-related learning where learning goals are pre-determined, and individual behavior shaped towards those goals—hence workplace training, on-the-job training, etc. Moreover, the term has been further extended to any kind of learning for any kind of work e.g., teacher training, doctor training, presuming an equivalence between vocational learning and animal tricks.

Modern cognitive psychology views vocational knowledge in terms of expertise. These ideas challenge the separation of physical and intellectual skills, as both kinds of knowledge consist in the same kinds of cognitive representations. Routine action is accomplished through automatic execution of specific cognitive procedures, developed through extensive practice. Experts conceptualize unfamiliar situations in terms of principles, structured at various levels of abstraction, and linked to cognitive procedures, under the control of higher order cognitive procedures.

3. *Social Patterning of Vocational Knowledge*

Cognitive and anthropological studies indicate that knowledge is highly situated. The idea of situated knowledge also challenges the tradition that there is vocational and 'other' knowledge (e.g., knowledge derived from academic disciplines). Rather, as in the tradition of Vygotsky, all knowledge is regarded as socioculturally based, with patterning of concepts and procedures arising from social interactions and social contexts.

Sociocultural considerations impel the idea of normative judgement also challenging mechanistic notions of vocational knowledge. The idea of normative judgement helps to explain why some expected cases of transfer do not occur—transfer in the work place depends on normative judgements about what is appropriate in the setting.

4. *Life-span Centered Constructions of Vocational Education*

Career theory addresses vocational education and training in terms of personal traits, values and goals, with such personal meaning constructs embedded in sociohistorical contexts. For example, in World War II, women were recruited into manufacturing, and farming and transport, re-defining vocational pathways and self-definition.

Life-span approaches argue that understanding any part of a life requires reference to the whole life course—lives/vocations are constructed in a particular social and historical context. Vocation may involve several work and other social roles over the life-span, and the term 'life career' may better reflect such dynamic processes. A life-long learning philosophy of education and training has developed in recognition of changing contexts, individual aspirations and the potential of vocational education and training for human capital (see later) development.

Education is used as a precursor to vocational choice, and as a context for vocational orientation and training. Curriculum, streaming, assessment and various other processes allocate young people to vocational pathways and labor markets through credentials. This stresses the polarity of education for individual and social empowerment (democracy/equity) vs. education for production/class segregation (social differentiation). Vocational education and training become a form of state intervention in pursuit of economic goals. These state measures are evident in the contractual arrangements of apprenticeships and traineeships.

Adult learning theory can also emphasize humanist or socially critical perspectives on vocational education. In the first case, the learner determines the vocational nature of their own goals and the required knowledge. Critical theorists seek to empower learners in understanding how their vocations are constituted by society; how such forces can be oppressive, and how to take action. Thus, both constructions often see work-related discourses in opposition to the needs of the individuals or to society as a whole.

5. *Economistic Constructions of Vocational Education*

In competency-based education/training, learning goals are set against industrial standards. The economic goal is cost-effective use of resources in shaping learner behavior towards pre-determined learning outcomes (Stevenson 1993). This perspective is less concerned with individuals' callings, ways of constructing meaning, and socially appropriate practice, than with potential *human capital*.

Human Capital Theory assumes that education 'helps develop skills of work, that is, improves the capacity of the worker to be productive' (Sweetland 1996, p. 354). Such conceptualizations can confine human development to increasing workplace capital. The learning problem is human resource development, i.e., shaping individuals to increase their potential as economic resources. Such perspectives can separate work from life, reduce the time horizon of vocational education and training to immediate and predictable aspects of work, and give weight to those verbally expressible aspects of human capacities, which can be detailed in advance of learning activities.

6. Sociohistorical Influences on Constructing Vocational Education

The knowledge which counts as vocational differs across cultures and time. Where production consists in the application of knowledge handed down through generations (as in early agricultural economies or in contemporary non-western cultures), the knowledge is holistic and normative, and associated with forming one's cultural identity. Even in Ford's early factories, there was a sense of community, with housing provided for workers, just as in traditional Japanese enterprises. However, production assembly lines and associated boredom have produced a gulf between working and being human. This gulf extends to the requisite knowledge, giving rise to such confined vocational education curriculum development techniques as task analysis (Berryman 1993).

Social crises such as wars have seen recurring regressions to behaviorist/technicist notions of education (Stevenson 1996). These reversions have been resisted by humanist and socially critical movements, with distinctions being emphasized between educative processes and products, and between expressive, and other objectives. However, this resistance has not prevented the re-emergence of behaviorism in the 1980s in the form of outcomes-based education. Notwithstanding, the democratization of education is creating 'greater freedom of choice and a fight against the socio-cultural determinism which predestines some for management positions prepared for through academic training and others for operational tasks requiring vocational training' (Pair 1998, p. 23).

7. Contemporary Concerns

Contemporary concerns in conceptualizing vocational education and training arise from:

(a) (economic) changes in the nature of productive activity,

(b) (societal) changes in attitudes to the role of industrial development,

(c) (cognitive) changes in conceptualizations of knowledge,

(d) (philosophical) changes in views about different ways of knowing and rendering that knowledge, and

(e) (pedagogical policy) changes in the organization and conceptualizations of vocational education and training as programs of pathways evolving over the life-span, with the need for bridges.

Economic constructions of knowledge are moving from narrow short-term capacities to those more characteristic of dynamic workplaces. Socioeconomic needs challenge the separation of individual development from social development and from the capacity for productive work; or, to use cruder terms, of culture from utility. Work, in a knowledge-based society, is clearly reliant on thinking skills, innovation, and capacities to displace time, space, and financial constraints on design, production, delivery, marketing, and sales. A new terminology for productive capacities has emerged: intellective and connective skills; mobile knowledge workers, symbolic analysts (e.g., Harvey 1989). Work has become more continuous with other aspects of life. The OECD (1980) listed 'competencies' for working life, which have now been replaced by key/core competencies/skills/qualifications (see Hodgson and Spours 1997, Nijhof and Streumer 1998). Economic changes highlight further the arbitrariness of the separate denotation of high status vocations (e.g., medicine, law) as professions, simply because their preparation occurs in universities.

The idea that knowledge can be produced outside of 'intellectual' institutions has now also been advanced by Gibbons et al. (1994), who argue that ('Mode 2') knowledge exists within the situational and time characteristics of teams brought together to solve such problems as those generated in modern innovative workplaces

Social changes include increased participation of young people with plural characteristics, in schooling, creating the need to overcome pedagogical complexity in provision of education and learning pathways (Achtenhagen and Grubb 2001). Challenges also include renewed concerns about quality of life in and outside work, the environment, inter-personal relationships, and global harmony. These various social concerns call into question the ways in which educational provision is institutionally separated.

Concerns arising from cognitive psychological and anthropological research include understanding the relationships between tacit and expressible knowledge, situated and generic knowledge, dispositions, values and action, knowing and transferring, and apparently different kinds of intelligence. Philosophical movements challenging philosophical structuralism also challenge the hierarchy and polarization of knowledge developed from ancient Greek society.

8. Methodological Issues

In researching vocational education, an immediate problem is in defining the phenomenon. Discourses have traditionally polarized such ideas as knowledge and experience, theory and practice, life and work, and academic and vocational concerns. Such polarizations often confound knowledge with the capacity to render knowledge in verbal (declarative) form. This arises partly because declarative knowledge has uncriticized currency amongst researchers whose training comes from universities. For example, in researching vocational knowledge, due weight is not routinely given to nonverbalizable (undeclarable) knowledge conceptualized variously as knowing how, tacit knowledge, implicit knowledge, multiple intelligences, and

non-verbalizable intelligence; or to the plurality of representational encoding and memory systems. Nor is weight given to conceptualizing training structures and contexts as the interaction of many actors in sociohistorical time (employers, unions, people, communities, public authorities, training and education institutions) (Pair 1998).

Theories, concepts, research approaches and research tools developed from research into early human development, schools and universities are often applied unproblematically to research in vocational education and training. For example, a high value is automatically assumed for verbalizable propositional knowledge in relation to capacities for expert performance. Recognition that knowledge is not to be confounded with verbalizable renditions places an onus on researchers to re-conceptualize theoretical developments that have been based on tools and perspectives that privilege the verbalizable.

Immediate research goals include the epistemological ones of determining the nature of vocational knowledge and its relationship with other ways of knowing; the illumination of the relationships among epistemological frameworks, research tools and methods; and the development of new approaches that capture and give weight to plural ways of knowing and rendering that knowledge, as well as illuminating non-verbalizable knowledge. These methodological developments are fundamental to reconciling different perspectives on vocational education and training and in researching such matters as the relationships among learning, personal choice, work, social change, institutional structures and learning pathways. Ifn addition, more work is needed on teaching–learning processes within vocational education settings. The pathways concept (Raffe 1998), in particular, may be a tool for research and a metaphor for travel over the life-span. In particular the networks of interconnected pathways afforded by vocational and technical training, universities, and enterprises will be significant for innovation, flexibility and life-long-learning. How well various countries address such issues may well determine their economic competitiveness and their social democracies.

9. *Future Directions: Addressing Polarizations*

The directions of social, technological and economic change compel reconstructions of the ideas of the vocational and vocational education, in such a way as to reconnect ideas that have become polarized. Change will come from a number of sources: more continuities in the workplace between innovation and execution (e.g., flatter hierarchies, pace of change, working in teams); more continuities between being at work and other aspects of life (e.g., time and space displacement such as telecommuting); more continuities between the social regard for those in and out of work (fewer core jobs (Hirsch 1991)); greater continuities between industrial and wider societal needs (a resurgence of values for the environment, quality of life, global harmony); more pressure for innovation in training pathways and their interconnectedness (Pair 1998, Raffe 1998), and greater attention to socialization/exclusion through education and work (Poole 1992, Popkewitz 1998).

A view of knowing more connected with doing will arise also because few cases of doing will be confined to the routine execution of skills. The pace of change will highlight the necessity for the continuous construction and reconstruction of meaning based on changing experiences, reconnecting knowledge with the individual's developing sense of vocation, and with productive/generative activity, in pursuit of vocation. Reconnections will need to link the educational needs of individuals, the development of society, and productivity and generativity in industrial, and other vocations. These links will be contextual, changing over the life-span. We suggest that theoretical conceptualizations will increasingly inter-relate the ideas of knowing, doing, feeling, judging, across the Aristotelian divides, and across the derived academic/vocational divide. We also suggest that the idea of generative/productive activity in all societal roles will serve as a basis to build a bridge across the cultural and utility divide, and that vocational constructions will be social/contextual and dynamic across the life-span.

See also: Adolescent Work and Unemployment; Adult Education and Training: Cognitive Aspects; Apprenticeship and School Learning; Career Development, Psychology of; Education: Skill Training; Human Capital: Educational Aspects; Situated Learning: Out of School and in the Classroom; Vocational Interests, Values, and Preferences, Psychology of

Bibliography

Achtenhagen F, Grubb W N 2001 Vocational and occupational education: Pedagogical complexity, institutional diversity. In: Richardson V (ed.) *Handbook of Research on Teaching*, 4th edn. AERA, Washington, DC

Berryman S 1993 Learning for the workplace. In: Darling-Hammond L (ed.) *Review of Research in Education*, Vol. 19, Chap. 8, pp. 343–401

Dewey J 1916 [1966] *Democracy and Education*. Collier-Macmillan, Toronto, ON

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M 1994 *The New Production of Knowledge. The Dynamics of Science and Research in Contemporary Societies*. Sage, London

Harvey D 1989 *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Blackwell, Oxford, UK

Hirsch J 1991 Fordism and post-Fordism: The present social crisis and its consequences. In: Bonefield W, Holloway J (eds.) *Post-Fordism and Social Form. A Marxist Debate on the Post-*

Fordist State. In the series, *Capital and Class*, Clarke S (ed.). Macmillan, London, pp. 8–34
Hodgson A, Spours K (eds.) 1997 *Dearing and Beyond: 14–19 Qualifications, Frameworks and Systems*. Kogan Page, London
Nijhof W J, Streumer J N (eds.) 1998 *Key Qualifications in Work and Education*. Kluwer, London
OECD 1980 *The Competencies Needed in Working Life*. Report by a Group of Experts. Organisation for Economic Co-operation and Development, Paris, France
Pair C 1998 Synthesis of country reports. In: *Pathways and Participation in Vocational and Technical Education and Training*. OECD, Paris, pp. 9–25
Poole M E 1992 *Education and Work*. Australian Council for Education Research, Hawthorn, Victoria
Popkewitz T S 1998 Dewey, Vygotsky, and the social administration of the individual: Constructivist pedagogy as systems of ideas in historical spaces. *American Educational Research Journal* **35**(4): 535–70
Raffe D 1998 Conclusion: Where are pathways going? Conceptual and methodological lessons from the pathways study. In: *Pathways and Participation in Vocational and Technical Education and Training*. OECD, Paris, pp. 375–94
Ryle G 1949 *The Concept of Mind*. Hutchinson, London
Stevenson J C 1993 Competency-based training in Australia: An analysis of assumptions. *Australian and New Zealand Journal of Vocational Education Research* **1**(1): 87–104
Stevenson J C 1996 The metamorphosis of the construction of competence. *Studies in Continuing Education* **18**(1): 24–42
Sweetland S R 1996 Human capital theory: Foundations of a field of inquiry. *Review of Educational Research* **66**(3): 341–59
Thomson J A K (trans.) 1976 *The Ethics of Aristotle*, rev. edn.Penguin, London

M. E. Poole and J. Stevenson

Vocational Interests, Values, and Preferences, Psychology of

This article describes how interests, values, and preferences are conceptualized in vocational psychology, how they are measured, and what reliable findings have been established by research using these measures. Such findings are the scientific basis for career counseling and career development.

Research in applied psychology has focused on two types of variables: ability and motivation. Psychologists succeeded early in developing measures of ability (e.g., *g*, verbal, numerical, spatial), drawing on the psychology of individual differences (DuBois 1970). Ability tests, *g* in particular, have been shown to be useful in a variety of school and work contexts (e.g., predicting grade-point average, job performance ratings).

Developing measures of motivation constructs proved to be more difficult. Many constructs were tried out and among the more successful ones were interests, values, and preferences. Research has since shown that interests, values, and preferences can contribute to the prediction of performance independently of the contribution of ability (Dawis 1991).

1. Definitions

Interest, value, and preference (and ability as well) belong to the class of constructs called 'dispositions.' Such constructs denote (typically stable) states of persons that indicate their behavioral tendencies. Thus, knowing the interests (or the values or preferences) of a person can indicate the likelihood that the person will engage in a particular activity given the opportunity.

Preference is the fundamental construct, interest and value at bottom involving preference. Preference means a more positive attitude ('leaning') toward one alternative over another (or other) alternative(s) expressed as the choice of the alternative over the other(s). Preference implies discrimination among the alternatives and judgment (or evaluation) of the alternatives. Preference implies comparative judgment, but it can also mean absolute judgment (requiring an absolute threshold for choice).

Interest and value are more complex constructs. Interest is manifested as sustained (long-term) attention involving cognition of the interest object and accompanying positive affect (pleasure). Interest can move beyond affective cognition to motor response (approach behavior or sustained activity), that is, interest can influence behaviour. Value is a similar construct, also involving cognition of the value-object and accompanying affect. It is manifested as affective valuation, and can also influence behavior. Both constructs are expressed as preferences, but they differ in their judgment basis.

When judgment in preference is based on liking (i.e., attraction), it is an interest. When the basis is importance (i.e., significance or meaning), the preference is a value. Liking and importance are differing dimensions of affective response as shown by the absence of empty cells in a two-way cross-classification (liked/important, not-liked/important, liked/not-important, not-liked/not-important).

The scaling of liking extends to its opposite, disliking; likewise, scaling importance extends to unimportance. Importance is readily conceptualized as one dimension, with unimportance as zero point. Liking can be conceptualized as one dimension (with disliking as the negative side of the scale) or as two dimensions (with disliking as a separate dimension, each dimension having 'indifference' as its zero point).

The remainder of this article will focus on liking preferences and importance preferences, that is, on interests and values. As described earlier, interests and values have a cognitive component (content, direction) as well as an affective component (intensity). Cognitive content can refer to persons, data, things, activities,

events, environmental conditions, and even internal conditions and events. In vocational psychology, measures of interests typically have work activities as their cognitive content. For values, the cognitive content is ends (goals or consequences of work) and means (instrumentalities) that have become ends in themselves. Such ends and means-ends are 'reinforcers' because, as behavioral consequences, they maintain behavior.

Besides having content, interests and values can be more, or less, intense. Measures of these constructs have to incorporate both direction (content) and intensity.

2. *Measurement*

Much of the research on vocational interests and values has been devoted to their measurement. Interests and values are commonly measured through self-report instruments (inventories, questionnaires). A typical self-report instrument consists of many items of varied content, with instructions on how to respond to the items. Response can take the form of a check-off (indicating what is liked or is important), a ranking of alternatives (sometimes in sets of pairs or triads), or a rating on a point scale (usually three or five points). All three types of response can be quantified by using any of several scaling methods (Dawis 2000).

Two kinds of scales are commonly used in interest and value instruments. Internally consistent scales are composed of highly intercorrelated items that measure the same principal component. Externally relevant scales consist of items that correlate with a variable external to the scale (the 'criterion variable'). The same interest or value instrument may include both internally consistent and externally relevant scales. Typically, many scales are included in an instrument in order to increase its usefulness.

Internally consistent scales are semantically comprehensible and thus are favored in theory-driven research. Scores on such scales are interpreted as reflecting intensity of interest or value. Externally relevant scales, by contrast, are often composed of items of heterogeneous content and may not 'make sense' theoretically, but are robust in their correlation with the criterion variables and hence are useful in professional practice. Scores on such scales are best interpreted as reflecting the similarity of the respondent to the 'criterion group' (representing the high pole on the criterion variable).

2.1 *Sample Measures*

Arguably the best-researched measure of vocational interests is the *Strong Interest Inventory* (*SII*; Harmon et al. 1994). It contains 320 items requiring response on a three-point, like–indifferent–dislike scale. The *SII* is scored on five sets of scales: general theme, basic interest, occupational, personal style, and administrative. Scores on the (externally relevant) occupational scales indicate the likelihood of a respondent's choosing an occupation (given the opportunity), and of remaining in the occupation (if entered). Such prediction is based solidly on numerous validity studies that demonstrate the occupational scales' usefulness in vocational counseling. Scores on the other sets of scales are useful in elaborating on and providing context for interpretation of the occupational scale scores.

The *Kuder Occupational Interest Survey* (*KOIS*; Kuder and Zytowski 1991) is another well-researched measure of vocational interests. It is composed of 100 items, each consisting of triads (three different activities) requiring a choice of 'most preferred' and 'least preferred.' The *KOIS* is scored on four sets of scales: dependability indices, vocational interest area estimates, occupational scales, and college major scales. Scores on the occupational and college major scales indicate similarity of interests with the corresponding referent group, which can then be the basis for career counseling.

The most researched measure of work values is the *Minnesota Importance Questionnaire* (*MIQ*; Rounds et al. 1981). The *MIQ* is a paired comparison questionnaire in which 20 'work needs' are compared, each with every other. Work needs are workers' requirements for work reinforcers (e.g., ability utilization, compensation, recognition), identified in studies of job satisfaction. Responses to the *MIQ* are scored for the 20 work needs and six work values. The *MIQ* is also scored by comparing a person's *MIQ* profile with commensurate measures for occupations (occupational reinforcer patterns, or ORPs). Where the *MIQ* indexes what work reinforcers are important to the person, the ORP indexes what work reinforcers are important to the occupation. How well the *MIQ* and ORP profiles match is predictive of occupational satisfaction.

Another measure of work values is the *Work Values Inventory* (*WVI*; Super 1970), which consists of 45 items, each requiring a response on a five-point importance scale. It is scored on 15 values (e.g., altruism, creativity, prestige, economic returns).

Two widely used measures of personal values are the *Study of Values* (Allport et al. 1960) and the *Value Survey* (Rokeach 1973). Both are used largely in descriptive studies.

2.2 *Structure of Content Domains*

The content domain of interests is best represented by the well-corroborated Holland (1997) model that uses a circular structure to organize six basic categories of interests: realistic, investigative, artistic, social, enterprising, and conventional. Numerous studies have

shown the circular order to be robust, with people's interests tending to be located in adjacent categories. This circular structure can be simplified by referring to two orthogonal axes: people vs. things, and ideas vs. data.

The content domain of values has not been studied as thoroughly as that of interests. One structure (Dawis and Lofquist 1984) proposes six categories organized by three bipolar orthogonal axes: achievement vs. comfort, status vs. altruism, and safety vs. autonomy.

2.3 Relations Among Interests, Values, and Other Disposition Variables

Interest and value scale scores correlate only marginally with each other, the largest correlations rarely exceeding 0.30. The highest correlations are between enterprising interests and status values, and between social interests and altruism values. Factor analysis yields separate interest and value factors.

Interest and value scales scores correlate slightly more with scores on personality trait scales. Correlations with personality trait scores reach and occasionally exceed 0.40. However, factor analysis shows that the three domains are relatively independent, with separate interest, value, and personality trait factors.

Even less related are abilities. Correlations of interest and value scores with ability scores rarely exceed 0.20. Interests correlate with abilities more than values do, but factor analysis shows separate ability, interest, and value factors. As currently measured, the content domains of ability and motivation are quite independent.

3. Research Findings

The reliable findings of research on vocational interests and values (as measured above) are reported below. More detail can be obtained from Dawis (1991).

3.1 Stability of Interests and Values

For most people, interests are relatively stable over long periods after age 25. Interests are unstable before adulthood (with notable exceptions), but they increase in stability after age 18. The exceptions are early childhood interests in mechanics, science, mathematics, art, and music, which tend to be rather stable. Twin studies show that interests have a significant heritable component—a likely determinant in the stability of interests.

Values are stable during adulthood, although not as much as interests. Little is known about the stability of values during the developmental years. However, twin studies also show that values have a significant heritable component.

3.2 Predicting Satisfaction and Satisfactoriness

Job satisfaction is best predicted by the degree to which the work reinforcers of the job match the incumbent's values. Close behind as a predictor is the degree to which the interests characteristic of the occupational group match the incumbent's interests. At a gross level, the higher the social status or intelligence status of an occupation, the higher is the job satisfaction of its incumbents—but also, the more likely its incumbents will have the appropriately matching values and interests.

The best predictor of worker satisfactoriness is still the degree to which the worker's abilities match the ability requirements of the job. Such prediction can be improved modestly by adding as predictors the degree to which the worker has the interests and values appropriate to the job.

3.3 Predicting Occupational Membership and Occupational Change

Follow-up studies (over 3- to 21-year intervals) of adults with interest measures have found that about half of them had appropriate interests, another fifth had nearly appropriate interests, and the remainder had inappropriate interests. This is the most reliable finding in research on interests, justifying the use of interest measurement in career counseling and career development. Similar results, though not as well documented, have been reported for values.

Studies of people who have changed their occupation have found that their interests and values tend to be more appropriate for the occupations to which they changed than the occupations from which they changed.

4. Problems and Future Directions

4.1 Research Problems

Practically all of the research on interests and values has used self-report instruments. The extent of bias in the findings that result from such use of a single method is unknown. Findings such as those reported above need validation by other methods.

Another problem with self-report measures is the assumption that scores on such measures index intensity of interests or values. By themselves, scores (number of items endorsed) index not intensity but rather, coverage of content domain. Often, intensity is 'built into' the rating scale anchors (e.g., 'Not Important,' 'Important,' 'Very Important'). The assumption that intensity is thereby measured requires validation.

Problems arise when an instrument has scales with the same names constructed by different methods (e.g., the SII). Studies of same-named scales in the same instrument, one internally consistent and the other

externally relevant, have found low correlations (median of 0.30) between the two. Examination of item overlap shows less than 50 percent commonality. Differing inferences from the two kinds of scales are therefore possible.

A similar problem arises for same-named scales in two different instruments. Studies show low correlations and loadings on separate factors in factor analyses even for scales constructed by the same method. Again, these can lead to differing, even contradictory, inferences. Items of such scales should be compared to see how similar they are in content.

4.2 Future Research Directions

The most important problem for future research is to examine how vocational interests, values, and preferences originate and develop. Such development is likely linked to emotional development (as well as cognitive development), involving both biological and cultural aspects. It is clear that vocational interests, values, and preferences are learned, but a more detailed account of how such learning takes place is needed.

However, in the near term, research will probably continue to focus on predictive validation studies using available or improved instruments as current professional practice requires more such research.

See also: Career Development, Psychology of; Interest, Psychology of; Values, Development of; Values: Psychological Perspectives

Bibliography

Allport G W, Vernon P E, Lindzey G 1960 *Manual for Study of Values*. Houghton-Mifflin, Boston

Dawis R V 1991 Vocational interests, values, and preferences. In: Dunnette M D, Hough L M (eds.) *Handbook of Industrial and Organizational Psychology*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA, pp. 833–71

Dawis R V 2000 Scale construction and psychometric considerations. In: Tinsley H E A, Brown S D (eds.) *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Academic Press, San Diego, CA

Dawis R V, Lofquist L H 1984 *A Psychological Theory of Work Adjustment*. University of Minnesota Press, Minneapolis, MN

DuBois P H 1970 *A History of Psychological Testing*. Allyn and Bacon, Boston

Harmon L W, Hansen J C, Borgen F H, Hammer A L 1994 *Strong Interest Inventory Applications and Technical Guide*. Stanford University Press, Stanford, CA

Holland J L 1997 *Making Vocational Choices*, 3rd edn. Psychological Assessment Resources, Odessa, FL

Kuder F, Zytowski D G 1991 *Kuder Occupational Interest Survey General Manual*. CTB/McGraw-Hill, Monterey, CA

Rokeach M 1973 *The Nature of Human Values*. Free Press, New York

Rounds J B, Henly G A, Dawis R V, Lofquist L H, Weiss D J 1981 *Manual for the Minnesota Importance Questionnaire*. University of Minnesota Department of Psychology, Minneapolis, MN

Savickas M L, Spokane A R (eds.) 1999 *Vocational Interests*, 1st edn. Davies-Black, Palo Alto, CA

Super D E 1970 *Manual for the Work Values Inventory*. Houghton-Mifflin, Boston

Thurstone L L 1959 *The Measurement of Values*. University of Chicago Press, Chicago

R. V. Dawis

Voluntary Associations, Sociology of

The term voluntary association has a long history in sociology, reaching back to the founding period of the discipline. Although never at the core of sociological theorizing, the term has figured prominently in a number of approaches (see *Conflict Sociology*; *Institutions*). Moreover, it is an important concept in a number of sociological specialties. Prominent examples are organizational sociology, where voluntary associations are seen as one of the three basic forms next to the business firm and the government agency; urban sociology, where voluntary associations have been identified as vehicles of local integration and sources of community power (see *Community Sociology*); political sociology, where voluntary associations are related to social movements (see *Ethnic and Racial Social Movements*), interest groups and status politics (see *Political Sociology*); or the sociology of development, where voluntary associations play an important role in building a social infrastructure for economic growth by generating networks of mutual trust (see *Modernization, Sociological Theories of*).

1. Definition

Not surprisingly, given the different contexts in which voluntary associations figure in sociology, definitions vary somewhat by theory and field. However, while numerous definitions have been offered, they share a common core in the sense that voluntary associations are seen as private, membership-based organizations in which membership is noncompulsory. What is more, the organization should have identifiable boundaries to distinguish members from nonmembers, be self-governing, and noncommercial in objective and behavior (Salamon and Anheier 1997).

Definitions differ mostly in emphasis and 'at the margins,' i.e., along the demarcation lines to related forms such as business (partnerships, cooperatives, mutual organizations), compulsory organizations (guilds, bar associations, and, in some countries, chambers of commerce), political organizations (parties, political action committees, interest groups), and quasigovernmental institutions (mass membership organization in the former Soviet Union, state churches). Some of the definitional complexities arise from significant overlaps with related terms such as

nonprofit organization, nongovernmental organization or third sector organization. In contrast to these forms, voluntary organizations have a membership focus, whereas many nonprofit organizations like hospitals, social service agencies, or art museums may have a governing board but no broad membership base as such. The same applies to nongovernmental organizations like Oxfam or Save the Children, although some like Amnesty International, the Red Cross, or Greenpeace are membership-based.

2. *Development*

Classical sociological thought saw voluntary associations as an indicator of social evolution in the development of undifferentiated to differentiated societies and voluntary associations as expressions of development supplanting traditional forms of organizing such as the extended family or compulsory groups such as guilds and caste-type institutions. Maine's distinction of status vs. contract, Toennies' Gemeinschaft vs. Gesellschaft, and Durkheim's mechanical vs. organic solidarity are prominent exemplars of this type of thinking.

In traditional societies, voluntary organizations fulfilled a number of functions. They served as integrative mechanisms cross-cutting family and clan structures as well as age groups, thereby avoiding potential divisions within the community; they operated as alternative mobility strata in societies with rigid, often hereditary status systems; they were part of the division of labor and provided mutual assistance and economic benefit to members (farmers, craftsmen, traders); or they were in fact precursors of social classes grouping similar economic statuses, occupations, and property owners around a joint interest.

Researchers emphasized the contributions voluntary associations make to the integration of migrants and minorities. African village associations in urban areas facilitate the acculturation of rural–urban migrants from the same area by providing a buffer zone between village life and the more diverse city culture. Other examples along the same principle include ethnic associations among Italian, Greek, or German immigrants to the United States in the nineteenth century, or Turkish clubs in Germany the late twentieth century. Next to social integration, voluntary associations served a complementary economic function as well. They provided a mechanism for mutual self-help among marginal population groups. Savings and credit associations among the poor, cooperative-type organizations among small-scale farmers and small-scale producers etc. are examples of how excluded members of society pool resources to improve their economic well being.

Modern sociological thought about voluntary associations has its in roots de Tocqueville's *Democracy in America* (1990 [1840]) where he observed the importance of voluntary associations for the working of democracy at the local level. In modern, democratic societies, voluntary associations serve a dual function: they organize people from diverse backgrounds around a common purpose, thereby contributing to social cohesion; and they build an intermediary sphere between the political center of power and the electorate, thereby allowing for the expression of minority preferences against a potential 'tyranny of the majority.'

With de Tocqueville's work, the study of voluntary associations began to put more emphasis on state–society relations, the political system, and social structure. A brief comparison of France, Germany, and US may illustrate this central point.

Tocqueville's analysis of the US was meant as a critique of France and its postrevolutionary political order. France had been a centralized nation-state long before the revolution of 1789, and this very centralization made the revolution possible and facilitated its effectiveness in replacing the *ancien regime* with a new ruling class. In line with the highly individualistic and anticorporatist ideology of the revolution, the influential *Loi de Chapelier* decreed that no intermediary associations were allowed to exist between the individuals as *citoyens* and the state as their supreme and unambiguous representative. As a result, the legal and political space for voluntary association became very limited. Only in the twentieth century did voluntary associations gain momentum as their number increased significantly in the 1980s onwards, particularly in the field of leisure, recreation, and culture (Archambault 1996).

In the US, no French *Jacobin* tradition restricted the development of voluntary associations. Indeed, while both countries emphasized individualism over collectivism, the US remained quasi stateless for a prolonged period, developed both a strong pragmatic orientation in policy-making, and a general mistrust of centralized power. As a result, voluntary associations flourished, and soon created a complex network of social clubs, religious, and church-based organizations, and political, professional, and scholarly associations. Voluntary associations and charities continued to occupy a prominent place throughout US history: as service providers in the absence of a welfare state, as an important source of political mobilizations (e.g., civil rights), and as a platform for status competition in a formally egalitarian society (DiMaggio and Anheier 1990). Coser (1956) suggested that multiple memberships in voluntary associations created overlapping social circles and criss-crossing social conflicts, thereby avoiding the emergence of significant class cleavages and other fault lines that could potentially divide society into two antagonistic groups.

In contrast to 'state-based' France and 'association-based' US, Germany reveals a very different role of voluntary associations in political development. In the

absence of an antifeudal revolution, the German case is a series of compromises between a self-modernizing feudal system and an emerging, co-opted bourgeoisie (Seibel 1992).

When elements of a bourgeois culture and civil society first evolved in the form of literary societies, music clubs, and educational associations, the political system and state administration remained under aristocratic control. Moreover, there existed a widespread system of self-governing organizations, in particular the mediaeval system of guilds and unions, with their compulsory membership and comprehensive coverage, underwent a modernization push. In the early nineteenth century, the mandatory system of corporatist organizations was transformed into modern voluntary associations as part of a modernization policy championed by the state. As a result, the German voluntary associations were not seen as the antithesis to the state, as in France, nor did they develop parallel to it, as in the US, but in interaction with an emerging welfare state.

Characteristically, whereas religions in both France and the US are voluntary associations constitutionally separate from the state, the German Catholic and Protestant Churches in particular continue to enjoy the status of public institutions with their own legal system (canonical law), signifying the incomplete separation of religion and political power as a consequence of historical compromises.

3. Scope

Regardless of the strikingly different political and social contexts in which voluntary associations developed historically, recent decades have witnessed a significant expansion in the number of associations and memberships. The number of voluntary associations in the US is well over 1.5 million, with 57 percent of the population being members of at least one association. In France, an associational boom has increased the number of associations to 700,000–800,000, with 35 percent of the French population having at least one membership. Associational density in Germany has tripled since 1960, and nearly two-thirds of Germans belong to associations. Social democratic countries like Sweden have among the highest membership rates among developed societies, with over 80 percent. On average, about half of the adult population in developed countries belongs to at least one voluntary association.

In the developing world, this proportion is somewhat lower with regard to formal, i.e., registered associations. Estimates of membership in indigenous forms of associations are very incomplete, as many of these organizations are informal and not registered. Moreover the legacy of authoritarian rule in many developing countries over the last 50 years also contributed to somewhat lower membership rates in formal associations. At the same time, there are significant variations. India shows a very rich tapestry of associational forms, with estimates running over 2 million. By contrast, developing countries with prolonged histories of authoritarian rule and a tradition of restrictive laws have much lower rates of associations. Formal registered associations number about 800 in Ghana, 13,000 in Thailand, 20,000 in Egypt, and 200,000 in Brazil (Anheier and Salamon 1998).

In former state socialist countries, mass organization linked to political parties accounted for very high membership rates; in the transformation process, however, many have lost significant numbers of members or otherwise faced dissolution. As a result, the membership structure in such societies in reorganizing and remains at lower levels than in the West. At the same time, important differences exist across countries: in the Czech Republic, Hungary, and Poland, an association boom immediately after the initial regime transition signaled the emergence of new voluntary sector, whereas in other countries like Romania and Bulgaria, such processes are much slower, suggesting a continued underdevelopment of civil society.

4. Theoretical Approaches

Current approaches group voluntary associations frequently with other nonstate/nonmarket forms in an attempt to develop theories that explain the existence of a third sector in modern economies that is located between the market (first) and the state (second). Several such theories have been introduced.

The 'heterogeneity' theory is based on the notion of 'market failure/government failure' in economics. This line of thought assumes markets have an inherent limitation in producing 'public goods,' i.e. goods that are available to all whether or not they pay for them. This 'market failure' justifies the presence of government, which exists to satisfy the demand for goods left unsatisfied by the market system. In a democracy, government can perform this role only when a majority of voters support the production of a particular public good. Where considerable differences in voter preferences exist about which public goods to produce, and which ones not to produce, it may be difficult to generate such majority support, leaving unsatisfied demand for public goods as a consequence. Such 'government failure' is most likely where considerable heterogeneity (religious, linguistic, or ethnic) exists in a population. In such circumstances, the establishment of voluntary associations and other types of nonprofit organizations serves both as a vehicle of collective action in the formation of interests and as a mechanism to supply the public goods neither state nor market provide.

A related body of theory sees voluntary associations not only as a reflection of demand heterogeneity; instead, the theory emphasizes the role of social and political entrepreneurs who seek to maximize nonmonetary returns, i.e., followers, believers, or members (James 1989). According to this supply side theory, the presence of such entrepreneurs is most likely in societies with high levels of religious competition or basic ideological splits within the population. In such circumstances, activists have an incentive to form voluntary associations as a way to attract adherents to their cause by offering services that these potential adherents might find attractive, such as health care, education, or cultural events.

A third approach is based on the notion of information asymmetries in market economies (Hansmann 1996). In many transactions, consumers lack the information they need to judge the quality of the goods or services they are purchasing. This can occur because the purchaser is not the same person as the consumer (e.g., the purchase of nursing home care by children for an elderly parent), because the service in question is inherently complex and difficult to assess, or the inputs of some are difficult to match with the benefits procuring to many. In such cases, purchasers or members seek alternative bases for trust in leadership performance or the quality of the resulting service. Voluntary associations offer a solution to this trust dilemma. Because of the 'nondistribution constraint,' i.e. the prohibition on distribution of profits to owners, voluntary associations may be more trustworthy and more likely to serve client or member needs. Accordingly, the nonprofit status of voluntary associations functions as a proxy for the market, signaling trust in the quality of output.

A fourth approach emphasizes the embeddedness of voluntary associations in the social, economic, and political dynamics of societies (Salamon and Anheier 1999). As such, their evolution cannot be attributed to any single factor, such as the unsatisfied demand for public goods or the supply of social entrepreneurs. Rather, the emergence of voluntary associations in modern society is rooted in the broader structure of class and state–society relations. Salamon and Anheier (1997) suggest four regime types of voluntary sectors, each with a particular constellation of social forces, specifically the social and economic importance of the state, measured in terms of social welfare spending, and the overall scope of civil society, measured as the size of voluntary sector activity. The *statist* regime combines limited state activity and a weak voluntary sector. It is found in societies where the state assumes a controlling role, and where elites co-opt the middle class, thereby reducing the level of organizing outside state structure and the market, with Japan being the most prominent example. The *liberal* model characterized by limited state activity but a well-developed system of voluntary association is found in countries like the United States and Great Britain. By contrast, the *social democratic* pattern in countries like Sweden is characterized by relatively high levels of government social welfare spending and relatively small nonprofit sectors in economic terms, but with a very active voluntary component in terms of number of associations and memberships. Finally, in the *corporatist model* both the state and the voluntary sector work in cooperation with each other, whereby the state assumes the role of financier of activities offered by voluntary associations. A typical example is Germany, where a vast network of religious organizations helps implement government-funded programs.

5. *Future Directions*

Despite some progress, the study of voluntary association continues to face considerable methodological and theoretical challenges. In terms of methodology, the measurement of central variables, such as membership, voluntary activities, and formality/informality remains a problem because of significant cross-cultural variations in meaning. Moreover, the development of adequate databases remains underdeveloped because major economic and social statistical systems exclude nonmarket/nonstate activities from regular reporting. In terms of theory, research on voluntary associations will increase by addressing the emerging field of civil society studies (Keane 1998), the notion of social capital (Putnam 1993), and work in the field of world society (Meyer et al. 1997).

See also: Action, Collective; Tocqueville, Alexis de (1805–59)

Bibliography

Anheier H K, Salamon L M 1998 *The Nonprofit Sector in Developing Countries*. Manchester University Press, Manchester, UK

Archambault E 1996 *Le sectteur sans but lucratif*. Economica, Paris

Coser L 1956 *The Functions of Social Conflict*. Free Press, New York

de Tocqueville A 1990 [1840] *Democacry in America*. Vintage, New York

DiMaggio P J, Anheier H K 1990 The sociology of nonprofit organizations and sectors. *Annual Review of Sociology* **16**: 137–59

Hansmann H 1996 *The Ownership of Enterprise*. Harvard University Press, Cambridge, MA

James E 1989 *Non-Profit Organizations: Comparative Perspectives*. Oxford University Press, Oxford, UK

Keane J 1998 *Civil Society*. Polity Press, Cambridge, UK

Little K 1965 *West African Urbanization: A Study of Voluntary Association and Social Change*. Cambridge University Press, Cambridge, UK

Meyer J W, Boli J, Tomas G M, Ramirez F O 1997 World society and nation state. *American Journal of Sociology* **103**: 144–81
Putnam R 1993 *Making Democracy Work: Civil Traditions in Modern Italy*. Princeton University Press, Princeton, NJ
Salamon L M, Anheier H K 1997 *Defining the Nonprofit Sector: A Cross-national Analysis*. Manchester University Press, Manchester, UK
Salamon L M, Anheier H K 1999 Social origins of civil society: Explaining the nonprofit sector cross-nationally. *Voluntas* **9**(3): 213–47
Seibel W 1992 *Funktionaler Dillettantismus*. Nomos, Baden-Baden, Germany

H. K. Anheier

Voluntary Organizations

Voluntary organizations are groups within which individual members associate of their own volition with others in the pursuit of common objectives. Participation in such organizations is structured according to agreed norms and is normally intended for public benefit. Voluntary organizations engage in a wide range of activities: those that are most often associated with voluntary endeavor include healthcare, education, religion, the arts, sports, and a wide range of social welfare services.

Voluntary organizations are self-governing and are not controlled by government. In a particular jurisdiction, depending on the nature of charity law, a voluntary organization may be categorized as charitable, in which case it may be eligible for specified tax privileges. The work of voluntary organizations may be supported by contributions provided by individual donors, corporations, or philanthropic foundations. Unpaid board members normally govern voluntary organizations; their work, however, may be accomplished either by volunteers or paid staffers.

The main forms of voluntary organization activity include service philanthropy (helping others); social solidarity (helping self and others in the same situation); changing (altering the basis of society); mobilizing (rallying other people or resources to support a cause); creating (bringing something new into play); and intermediary (coordinating or organizing activity as research and development, provision of administrative service, or representation) (Knight 1993). Voluntary organizations are 'voluntary' in a dual sense: (a) much of their human resources are contributed by members as volunteers and (b) they are nongovernmental and not-for-profit (cf. Van Til 1988). Thus, voluntary organizations may be seen as mainstays within what is increasingly referred to as civil society (noncommercial, nongovernmental) and key components of what has come to be called society's 'third sector' (business and government being the other two sectors).

1. *Intellectual Antecedents and the Theory of the Third Sector*

Voluntary organizations are central to a number of important theoretical perspectives in social science (Sills 1968, Van Til 1988). Writing at the end of the nineteenth century, the French sociologist Emile Durkheim, following in the footsteps of his countryman Alexis de Tocqueville, gave a major role to associations in his studies of social solidarity, in turn giving birth to the influential theories of sociopolitical pluralism. The observations of Max Weber laid the basis for later neocorporatist thought, and the studies of Robert Michels illuminated aspects of elitism and populism within voluntary organizations. Karl Marx, although hardly an enthusiast for voluntarism, nevertheless relied upon the power of voluntary social movements to accomplish the revolutionary tasks he predicted and prescribed. And, in the twentieth century, social theorist Talcott Parsons envisaged voluntary organizations playing a major role in the 'integration' of societal institutions, a task extended by the contemporary communitarian theories of Amitai Etzioni, Roger Lohmann, and Jennifer Wolch.

Distinctions between formal and informal organizations, nonprofit and for-profit groups, and public and private purposes (the boundaries between each being often unclear or 'blurred') suggest that voluntary organizations may occupy space where they act very much in 'hybrid' or 'multi-sectoral' ways, as Victor Pestoff has demonstrated visually (in Kramer 1998). Voluntary organizations may assume specific characteristics of organizations in the other sectors. Rudolph Bauer has observed that volunteers tend to experience voluntary organizations as providers of charitable service, board members tend to see them as though they were political organizations, and staff may behave as though the same organization is primarily a business.

2. *Voluntary Organizations: Scope and Scale*

Around the world in diverse societies and in contrasting political cultures the final two decades of the twentieth century saw a remarkable growth in the range and number of voluntary organizations, leading some commentators to speak of a 'global associational revolution of extraordinary scope and dimensions' (Salamon 1995). This pace of development shows no sign of slowing. In the developing world and throughout Eastern Europe locally based voluntary organizations work in partnership with international agencies and philanthropic foundations whose policies increasingly stress empowerment and the participation of local people in development projects. Throughout the developed world the retrenchment of public welfare regimes and pressures on public funding have encouraged the formation of partnerships between public authorities and voluntary organizations to

deliver social welfare and other services with increased flexibility and less bureaucracy.

The development of close relations between government agencies and voluntary organizations has promoted concerns about the 'incorporation' of voluntary organizations by the state, and the danger that they would lose their role as advocates. Although generally represented as a part of a 'private' nongovernmental sector, many voluntary organizations increasingly are described as becoming 'interdependent' with government. This interdependence most often has been linked to the service provider role that voluntary organizations play under contract to government programs (Kramer 1998). Billis and Harris (1996) have identified the linkages created through voluntary organizations' roles as advocates for public policies and as representatives of special interests.

Prominent among the contemporary ranks of voluntary organizations in a number of Western nations stands a subcategory that is governmentally identified as exempt from corporate taxation. These are conventionally identified as 'nonprofit organizations,' that are certified to perform a wide range of functions in society, and in return for their exempt status are required to provide public accounting of their actions.

Contemporary estimates identify over 1.2 million nonprofit organizations in the United States, many of which employ paid staff as well as utilize volunteer employees and board members. In the United States approximately half the population engages in regular volunteering for the benefit of others outside their families; and nine percent of all meaningful employment (and seven percent of all salaries) emerge from the third sector (Hodgkinson et al. 1996). Much additional activity is conducted within grassroots associations, which need not be formally chartered, and whose number has been estimated at 7.5 million or more in the United States (Smith 2000).

3. *The Growth of Voluntary Action Research*

Since 1970, the study of voluntary organizations has become an increasingly active arena for scholarly research. Professional associations like ARNOVA (Association for Research on Nonprofit Organizations and Voluntary Action) and ISTR (International Society for Third Sector Research) have grown rapidly; scholarly journals like the *Nonprofit and Voluntary Sector Quarterly*, *Voluntas*, and *Nonprofit Management and Leadership* have emerged; and organizations like Independent Sector and CIVICUS have grown to play key roles in asserting the role of voluntary activity in society.

The phenomenon of global associational development has been charted in an extensive series of country studies under the aegis of The Comparative Nonprofit Sector Project directed at the Johns Hopkins University in the United States. These studies have established that the numbers of voluntary organizations and the importance of this sector are much greater than previously thought, both in the developed and developing world. Academic centers for the study of voluntary organizations now exist in many countries. In the United States there are many well-established academic research centers such as those at Indiana University (where the Nonprofit Academic Centers Council is located), and at Case Western Reserve, Duke, Harvard, Georgetown, and Yale Universities and the City University of New York. In the United Kingdom, academic centers exist at the London School of Economics, London's South Bank University, Aston University in Birmingham, and the University of Ulster. Elsewhere there are university research centers or academic research concentrations in many countries including Australia, Belgium, Brazil, Canada, Egypt, India, Israel, the Republic of Ireland, South Africa, and Spain.

4. *Challenges and Future Directions*

Most, but not all, voluntary organizations are benevolent or benign in their objectives. Whereas all voluntary organizations tend to feature common bonds of interest, solidarity, and commitment among their members, sometimes described as 'bonding social capital,' not all voluntary organizations generate or facilitate the building of 'bridging social capital,' which, by its nature is broad and reaches out to other groups. Bridging social capital encourages linkages with networks and tends to be tolerant of differences (Putnam 2000). In some organizations the common interest is either so narrowly cast or pursued by such dubious means that many members of the wider society find their actions inimical to the public good or even to be a danger to public safety. Members of ethnic minorities or groups which consider themselves to be oppressed may form voluntary organizations to pursue social and/or political objectives, sometimes using peaceful, legal, and nonviolent means, and at other times adopting paramilitary forms of organization and discipline and using violent methods (Gidron et al. 1999). The Middle East and Northern Ireland provide examples of violent political movements closely linked to voluntary welfare organizations established to provide much-needed social, cultural, and educational services to disadvantaged populations. Voluntary organizations thus serve two purposes: they respond to social need as they assist political movements in building and strengthening their constituencies. But the purposes of voluntary organizations can also be twisted into the service of those in governmental power, as was the case in Nazi Germany (Bauer 1990).

The increasing commercialization of modern life has affected nonprofit organizations, which have become increasingly dependent upon income from fees and, particularly in the case of hospitals and higher education institutions, have adopted many

'business-like' structures and practices (Powell 1986). Furthermore, in an era in which 'everything is for sale,' for-profit corporations often insist that their charitable donations must benefit themselves as well as others, by practicing what has been called 'cause-related marketing.' Philanthropic giving is itself encouraged by tax-deductibility, and individual programs are often required to support themselves as 'profit centers' within specific nonprofit organizations.

In the climate of resource uncertainty that prevails among many paid-staff voluntary organizations, fund-raising (increasingly referred to as 'development' or 'advancement') often becomes a central staff activity, and a high emphasis is paid to the quality of 'nonprofit management' credentials and skills, often provided by the many university centers that have emerged since 1980.

The futurist J Rifkin (1995) has argued that the twenty-first century is likely to see a continuing decline in the amount of paid work made available to individuals by both governments and corporations. This restructuring of economy and society will provide increasing scope, he contends, for participation in voluntary activity. By providing shadow wages (e.g., tax deductions) to support voluntary action and a social wage to facilitate volunteering by under-employed people, the social economy provided by the third sector may play an important role in maintaining social peace and order in an age threatened by the divisiveness of increasing socioeconomic inequality.

In such a time of change and challenge, it is important to recall that it is voluntary action that powers the third sector, and that the third sector is relied upon by society as a crucial developer of its distinctive goals. Voluntary organizations resemble, in part, businesses, governments, and families—and interact directly with all three of these organizational forms.

Voluntary organizations receive their unique stamp, however, from their shared value of advancing the common good by means of reflective and significant voluntary group activity. The core values of voluntary organizations distinguish these organizations from firms that operate in the marketplace and from other kinds of social service organization. These values should be reflected both in the objectives of the organizations and in their operational policies. For example, a typical volunteer or paid practitioner within a voluntary organization would likely consider as of central importance values such as caring, social justice, fairness, and the involvement of users in planning and delivery of services. Similarly, voluntary organizations normally will prioritize marginalized and minority individuals, people in need, and those who feel excluded or disadvantaged.

A particular challenge for voluntary organizations is to interpret and apply their values and to ensure that in their day-to-day work, management and care practices, and personnel and other policies, they are consistent with the values they say they stand for. When voluntary organizations succeed in defining their goals and meeting needs responsively, they can contribute significantly to the quality of life in modern society.

See also: Civil Society, Concept and History of; Communitarianism: Political Theory; Communitarianism, Sociology of; Durkheim, Emile (1858–1917); Lobbying; Pluralism; Social Capital; Tocqueville, Alexis de (1805–59); Voluntary Associations, Sociology of

Bibliography

Bauer R 1990 Voluntarism, nongovernmental organizations, and public policy in the Third Reich. *Nonprofit and Voluntary Sector Quarterly* **19**: 199–214

Billis D, Harris M 1996 *Voluntary Agencies: Challenges of Organisation and Management*. Macmillan, London

Gidron B, Katz S, Meyer M, Hasenfeld Y, Schwartz R, Crane J K 1999 Peace and conflict resolution organizations in three protracted conflicts: Structures, resources and ideology. *Voluntas: International Journal of Voluntary and Nonprofit Organizations* **10**: 275–98

Hodgkinson V A, Weitzman M S, Abrahams J A, Crutchfield E A, Stevenson D 1996 *Nonprofit Almanac 1996–1997: Dimensions of the Independent Sector*. Jossey-Bass, San Francisco

Knight B 1993 *Voluntary Action*. Home Office, London

Kramer R 1998 *Nonprofit Organizations in the 21st Century: Will Sector Matter?* Aspen Institute Nonprofit Sector Research Fund, Washington, DC

Powell W W 1986 *The Nonprofit Sector: A Research Handbook*. Yale, New Haven, CT

Putnam R D 2000 *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, New York

Rifkin J 1995 *The End of Work: The Decline of the Global Labor Force and the Dawn of the Post-market Era*. Tarcher, New York

Salamon L M 1995 *Partners in Public Service: Government Nonprofit Relations in the Modern Welfare State*. Johns Hopkins University Press, Baltimore, MD

Sills D 1968 Voluntary associations: Sociological aspects. *International Encyclopedia of the Social Sciences* **16**: 363–76

Smith D H 2000 *Grassroots Associations*. Sage, Thousand Oaks, CA

Van Til J 1988 *Mapping the Third Sector: Voluntarism in a Changing Social Economy*. Foundation Center, New York

J. Van Til and A. P. Williamson

Volunteerism, Psychology of

Every year, millions of people volunteer substantial amounts of time and energy to helping others. Among other services, volunteers provide companionship to the lonely, tutoring to the illiterate, counseling to the

troubled, and health care to the sick. According to one estimate, 93 million adults in the USA engaged in some form of volunteerism during 1995, with their combined length of service totaling 20.3 billion hours. These 93 million volunteers represented 49% of the adult population, a quarter of whom devoted five or more hours per week to volunteering (Independent Sector 1996). Moreover, volunteerism is not limited to the USA and can be found in many other parts of the world (Curtis et al. 1992).

From a scientific perspective, the study of volunteerism provides opportunities for inquiring into the nature of helping and of prosocial action more generally. Two distinct traditions of research and theorizing in psychology can be identified, each investigating a specific form of helping and each focusing on issues relevant to that form of helping.

1. *Varieties of Helping and Prosocial Action*

One research tradition focuses on situations in which potential helpers are confronted with unexpected opportunities to help strangers (the classical example being 'bystander intervention,' e.g., Latané and Darley 1970). The help that occurs here is typically unplanned and spontaneous, is usually confined to relatively brief encounters, and generally entails neither prior contact nor future contact between helper and recipient. To answer the question of why this spontaneous helping occurs, some researchers have proposed that it reflects purely altruistic motives to benefit others (e.g., Batson 1998). Other answers involve claims that such helping reflects selfish concerns, such as to feel good about oneself (e.g., Cialdini et al. 1987).

A second research tradition focuses on people providing long-term care and support to those suffering with serious illness or chronic conditions. This research examines 'obligated' caregivers who, because of marital bond or blood relationship, provide assistance to a spouse, parent, child, or sibling suffering from ailments, including Alzheimer's disease, arthritis, or the aftermath of a stroke or serious accident (e.g., Thompson and Pitts 1992, Folkman et al. 1994). In such relationships, the impetus for assistance may reasonably be seen to flow from the legal, ethical, or familial obligations that tie the helper to the recipient, thereby rendering questions of personality and motivation all but moot. These helpers help because they have to, doing so for reasons related to moral, legal, or personal obligation. This literature has focused on the nature of such caregiving: it is difficult, demanding, and stressful, and may exact considerable psychological and even physical tolls on those who provide such ongoing, long-term care.

Although it shares many features with other forms of prosocial action, several features of volunteerism mark it as a distinctive form of helping (Piliavin and Charng 1990, Clary and Snyder 1991, Omoto and Snyder 1995). Rather than happening upon situations in which they are pressed to help, volunteers typically seek out opportunities to help, and may deliberate long and hard about the precise nature of their involvement; these features distinguish volunteerism from spontaneous helping. Volunteerism often entails commitments that extend over long periods of time and entail sizable personal costs; this feature of volunteerism is shared with the long-term helping of obligated caregivers. However, unlike obligated caregivers, volunteers typically do not know those they help in advance and have no prior bonds of obligation to help them; instead, volunteers often are matched with the recipients of their help by service organizations.

As a hybrid strain of helping, volunteerism engages many of the questions associated with the spontaneous helping and obligated caregiving traditions of research on helping. Volunteerism raises questions of personality (is it enacted only by individuals with altruistic dispositions?), of motivation (why, in the absence of obligation, do people ever volunteer?), and of mechanisms for sustaining it (in what ways do personal and social resources promote long-term helping?). At a theoretical level, then, understanding volunteerism requires broaching questions that are also of fundamental concern in other forms of helping.

However, in spite of its potential to engage a wide range of theoretical questions, there has been (until relatively recently) remarkably little psychological research on volunteerism (see Piliavin and Charng 1990, Clary and Snyder 1991). Also, because of its defining and characteristic features, volunteerism is something of a curious phenomenon. For a variety of reasons, it simply should not occur. Unlike the helping that occurs in response to emergencies, there is no press of circumstances. Unlike the helping that occurs in families and in existing relationships, there are no bonds of obligations. Volunteerism is effortful, time consuming, and involves opportunity costs to volunteers—all features that should serve as barriers to volunteering. Yet people do seek out opportunities to volunteer, and they do sustain their volunteer efforts over extended periods of time. The question is, 'Why?' Why do some people get involved in helping others as volunteers and in participating in society? What is it that moves people to seek out opportunities to help? What is it that guides them toward some helping opportunities and away from others? And what is it that sustains their efforts over time and through adversity?

2. *Motivational Foundations of Volunteerism*

The questions of why people volunteer to help others are fundamentally questions about motivation. Answering these questions in terms of the forces that move people to action and that guide and sustain

action is, therefore, to provide answers in terms of the processes of motivation. Theories of motivation are, by definition, concerned with what moves people to action, what channels their action in some directions rather than others, and what sustains their actions. One approach to understanding the motivations behind volunteerism is a 'functional' approach concerned with the needs being met, the motives being fulfilled, and the functions being served by engaging in volunteer service (e.g., Omoto and Snyder 1995, Snyder et al. 2000).

Several considerations recommend functionalist theorizing (Snyder and Cantor 1998). First, the functional approach is a motivational perspective that directs us to inquire about the personal and social processes that initiate, direct, and sustain action. Second, a core proposition of functionalist inquiry is that people can and do perform the same actions in the service of different psychological functions, with different people engaging in the same volunteer activity but doing so to fulfill different motives. Third, the functional approach suggests that important psychological events, such as embarking on a course of volunteer activities and then maintaining those activities over extended periods of time, depend on a matching of the motivational concerns of individuals with situations that can satisfy those concerns. Finally, research stimulated by motivationally oriented analyses of a wide variety of cognitive, affective, behavioral, and interpersonal processes supports key functionalist themes.

Guided by the broad directive provided by functionalist theorizing, researchers have sought to identify the precise motivations that can be fulfilled through volunteer service. Based on findings from diverse empirical investigations of volunteers involved in a wide variety of helping activities, it has been possible to identify and operationalize a diversity of personal and social functions served by volunteering (e.g., Omoto and Snyder 1995, Clary et al. 1998, Okun et al. 1998), including the expression of personal values, the quest for understanding, the desire for enhanced esteem, the strengthening of social relationships, the building of community ties, and the search for career advancement.

3. *The Processes of Volunteerism*

In their efforts to understand volunteerism as a form of sustained, ongoing prosocial action, researchers have examined the ways in which people construct agendas for voluntary action, identify their own motivations for volunteering, seek out and pursue service opportunities that they believe have the potential to fulfill their motivations, and sustain and maintain their involvement in these volunteer activities. Much of this research concerns helping and prosocial action as they occur in the real world, focusing on 'real' individuals involved in 'real' acts of volunteerism in 'real' world settings.

A guiding theme of research on volunteerism is that critical events in the 'life history' of a volunteer over the course of his or her volunteer service depend importantly on the match of that volunteer's motivations to the opportunities afforded by the environment in which volunteering occurs. Relevant to the conditions under which volunteerism is initiated, it has been found that persuasive messages motivate people to initiate volunteer service to the extent that the messages are tailored to the specific motivations important to individual recipients of the message (e.g., Clary et al. 1998). The importance of matching an individual's motivations to events in the situation does not end with recruitment. As participants in an ongoing and sustained activity, volunteers whose motivational concerns are served by their participation appear to derive greater satisfaction than those whose concerns are not met (e.g., Clary et al. 1998). Moreover, actual intentions to continue serving as volunteer helpers can also be linked to the matching of experiences with motivations such that volunteers who receive functionally-relevant benefits also express greater intentions to continue as volunteers, both in the short-term and in the long-term, than do volunteers who do not receive functionally relevant benefits or who receive functionally-irrelevant benefits (Clary et al. 1998). Finally, longitudinal investigations have found that fulfilling important motivations predicts actual commitment to volunteering (Omoto and Snyder 1995).

4. *Promoting Volunteerism*

Many societies view participation in volunteer activities as highly desirable, and people are encouraged to become volunteers. In fact, some American schools actually require students to participate in community service. However, requiring people to engage in community service may prove counterproductive in the long run, as there is considerable evidence that applying extrinsic pressures to perform an activity may actually undermine the motivation to continue to perform the activity once the pressure is removed (e.g., Lepper et al. 1973).

Stukas et al. (1999) examined this process in a 'mandatory volunteerism' program in which university students were required to perform community service. Of particular concern were students' intentions to continue to volunteer after the conclusion of the program, and the relation of these intentions to the students' preservice perceptions of control and prior volunteering experience. For those students who felt that their participation was under their own control (and thus perceived less external control to volunteer), greater future intentions were associated with greater previous experiences as a volunteer. However, for

those who felt external pressure to volunteer, their future intentions to volunteer were less than those who felt that they had retained personal control, and a slightly negative relationship between prior experience and intentions was observed.

Thus, volunteerism provides an opportunity to explore the empirical debate over the immediate and long-term impact of compelling people to act. Certainly, immediate volunteering can be induced by requiring students to serve, but it seems that future choices to volunteer may become less likely if the initial experience is accompanied by perceptions of external control.

5. *Theoretical and Practical Implications*

The study of volunteerism delivers potential messages, both theoretical and practical. Theoretically, the study of volunteerism is informative about the nature of helping. As exemplified by volunteerism, it informs us about forms of helping and prosocial action that are planful, sustained, and that occur in the absence of bonds of obligation. Also, from a theoretical perspective, it illustrates the benefits of conceptualizing motivation as the planful pursuit of agendas for action.

Practically, it teaches lessons about the practice of volunteerism, specifically about the ways that organizations dependent on the services of volunteers can enhance the recruitment, placement, and retention of volunteers by systematic attention to the experiences and the motivations of individual volunteers. As suggested by the functional approach, matching of recruitment appeals, volunteer tasks, and volunteer experiences to volunteers' motivations can influence the satisfaction, effectiveness, and longevity of service of volunteers.

Further, at an even broader level, studying volunteerism is also likely to yield valuable information of societal significance. It has the potential to inform about the actual and potential roles of volunteers and volunteer organizations in confronting and surmounting many of the problems that confront and challenge society.

See also: Architecture; Community Environmental Psychology; Environmental Psychology: Overview; Residential Environmental Psychology

Bibliography

Batson C D 1998 Altruism and prosocial behavior. In: Gilbert D, Fiske S T, Lindzey G (eds.) *Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, Vol. 2, pp. 282–316

Cialdini R B, Schaller M, Houlihan D, Arps K, Fultz J, Beaman A L 1987 Empathy-based helping: is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology* **52**: 749–58

Clary E G, Snyder M 1991 A functional analysis of altruism and prosocial behavior: the case of volunteerism. In: Clark M (ed.) *Review of Personality and Social Psychology*. Sage, Newbury Park, CA, Vol. 12, pp. 119–48

Clary E G, Snyder M, Ridge R D, Copeland J, Stukas A A, Haugen J, Miene P 1998 Understanding and assessing the motivations of volunteers: a functional approach. *Journal of Personality and Social Psychology* **74**: 1516–30

Curtis J E, Grabb E S, Baer D E 1992 Voluntary association membership in fifteen countries: a comparative analysis. *American Sociological Review* **57**: 139–52

Folkman S, Chesney M A, Cooke M, Boccellari A, Collette L 1994. Caregiver burden in HIV+ and HIV− partners of men with AIDS. *Journal of Consulting and Clinical Psychology* **62**: 746–52

Independent Sector 1996 *Giving and Volunteering in the United States: Findings from a National Survey 1996*. Independent Sector, Washington, DC

Latané B, Darley J M 1970 *The Unresponsive Bystander: Why Doesn't He Help*? Appleton-Century-Crofts, New York

Lepper M R, Greene D, Nisbett R E 1973 Undermining children's intrinsic interest with extrinsic reward: a test of the 'overjustification' hypothesis. *Journal of Personality and Social Psychology* **28**: 129–37

Okun M A, Barr A, Herzog A H 1998 Motivation to volunteer by older adults: a test of competing measurement models. *Psychology and Aging* **13**: 608–21

Omoto A M, Snyder M 1995 Sustained helping without obligation: motivation, longevity of service, and perceived attitude change among AIDS volunteers. *Journal of Personality and Social Psychology* **68**: 671–86

Piliavin J A, Charng H W 1990 Altruism: a review of recent theory and research. *Annual Review of Sociology* **16**: 27–65

Snyder M, Cantor N 1998 Understanding personality and social behavior: a functionalist strategy. In: Gilbert D T, Fiske S-T, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, Boston, Vol. 1, pp. 635–79

Snyder M, Clary E G, Stukas A A 2000. The functional approach to volunteerism. In: Maio G R, Olson J M (eds.) *Why We Evaluate: Functions of Attitudes*. Lawrence Erlbaum, Mahwah, NJ, pp. 365–93

Stukas A A, Snyder M, Clary E G 1999 The effects of 'mandatory volunteerism' on intentions to volunteer. *Psychological Science* **10**: 59–64

Thompson S C, Pitts J S 1992 In sickness and in health: chronic illness, marriage, and spousal caregiving. In: Spacapan S, Oskamp S (eds.) *Helping and Being Helped: Naturalistic Studies*. Sage, Newbury Park, CA, pp. 115–52

M. Snyder

Voting: Class

The idea of class voting appears straightforward: it refers to the tendency for voters in a particular class to vote for a particular party, political candidate, or

groupings of these, rather than an alternative option when compared with voters in another class, or classes. In other words, it describes a pattern of association between class and vote. Yet this simplicity hides considerable intricacy and ambiguity. The definition of social class in and of itself introduces much confusion—class is a much debated, even 'essentially contested,' concept. The issues raised become more complex still when class concepts are translated into measures of class position that allocate voters to classes and when attempts are made to summarize statistically the class–vote association. Explaining patterns of class voting and understanding their implications raises many more unresolved questions. All of these theoretical and methodological issues have at times been the focus of dispute. In addition, however, there has been one major substantive argument that has received more attention than all other questions and which has touched on all of them: has class lost its importance as a source of voters' political preferences?

Historically, interest in class voting arose in response in part to the failed agenda of Marxism. For classic texts in political sociology electoral politics was an expression of 'the democratic class struggle' (Anderson and Davidson 1943) that replaced, at least for the short-term, the predicted class-based revolution. This in turn led to a focus on class voting as a dispute between just two classes, the working class and the middle class, and their representatives, parties of the left and right. Early research also relied on data obtained at the level of electoral constituencies. Given the well-known problems with inferring individual vote choices from aggregate data this work had to make strong assumptions about how voters in different classes actually voted. To help overcome this limitation, modern social scientific studies of class voting have focused on its measurement through surveys of voters. From its beginnings in a few national election studies this data base has expanded through time and space to create a large body of evidence on class voting that now informs academic debate.

1. *The Two Class/Two Party Era*

The baseline for many current studies was set by Robert Alford's (1967) analysis of trends in class voting in four Anglo-American democracies (Australia, the UK, Canada, and the USA) between 1936 and 1962 using a measure of social and economic position in which occupations were aggregated to form a dichotomous manual/nonmanual class division. He also introduced the most commonly used, cited, and criticized measure of the level of class voting: the 'Alford index.' The Alford index is simply the difference between the percentage of manual workers that voted for left-wing political parties on the one hand and the percentage of nonmanual workers that voted for these parties on the other. This became standard practice in the cross-national and over-time analyses that followed, and which usually concluded that class voting is in decline. Lipset (1981), for example, presented evidence of a downward trend in the UK, Germany, and the USA between 1945 and 1980; while Inglehart (1997) points to evidence of a continuing decline from the early 1980s onwards. Indeed, whether or not class voting in modern industrial societies has declined is, for many commentators, no longer an issue. The question is *why* has it declined? A plethora of explanations have been proposed to account for this outcome (see Goldthorpe 1996, Manza et al. 1995): some assert that it has lost its importance as a determinant of life-chances and in consequence its role as a source of divergent political interests because the working class has become richer, white-collar workers have been 'proletarianized,' or because of extensive social mobility between classes. Others argue that 'postindustrial' cleavages, such as gender, race, ethnicity, public versus private sector, and various 'identity groups,' are emerging and replacing class-based conflict. Postindustrialism also impacts on class voting through the rise of post-material values as a basis of party preference, which has led to the 'new left' drawing its support from the middle classes thus weakening the class basis of left-right divisions; while rising levels of education and 'cognitive mobilization' have supposedly produced voters that are calculative and 'issue oriented' rather than driven by collective identities such as class.

Whatever their strengths and weaknesses—and there are many questions and doubts that remain unresolved in each case—all of these explanations assume that there is indeed a widespread, secular decline in class-based political action which then has to be explained. However, answers to questions about change have been strongly conditioned by choices of method and measurement. The literature on the decline of class voting that followed Alford's work has been characterized by shortcomings in both these respects.

For example, one problem with the manual/nonmanual distinction is that it obscures variations in the composition of the manual and nonmanual classes. Any changes in their composition may lead to spurious change in estimates of class voting. Thus, if skilled manual workers are more conservative than unskilled workers and the number of skilled workers increases, the difference between manual and nonmanual workers will decline even if the relative political positions of skilled, unskilled, and nonmanual workers remain the same. With this and similar indices the measurement of the class–vote association is open to confounding by changes in the shape of the class structure and in the general popularity of the parties. In other words, this type of index confuses differences in the marginal distributions of the variables with differences in the association it is supposed to measure.

The same problem applies to the apparently more sophisticated ordinary least squares regression techniques used in, among others, Franklin, Mackie and Valen's (1992) ambitious 16 nation study that represents, arguably, the culmination of this tradition.

2. *Recent Innovations*

In recent years, therefore, the rather atheoretical manual/nonmanual representation of class voting has been to a large degree superseded as researchers have taken into account more nuanced ideas about class structure. Most influential has been the class schema developed by John Goldthorpe and his colleagues (Erikson and Goldthorpe 1992). First used to study patterns of class voting over time in Britain by Heath *et al.* (1985) this instrument, unlike many other measures of class, has been empirically validated. The main classes identified by the Goldthorpe schema are the petty bourgeoisie (small employers and self-employed), the service class, or salariat (professional and managerial groups), the routine nonmanual class (typically lower grade clerical 'white-collar workers'), and the working class (foremen and technicians, skilled, semi-, and unskilled manual workers). The principal distinction underlying the distinction among the employee classes in the schema is between a service contract in which employees receive not only salaried rewards but also prospective elements—salary increments, job security and pension rights, and, most importantly, well-defined career opportunities—and a labor contract, in which employees supply discrete amounts of labor, under supervision, in return for wages which are calculated on a 'piece' or time basis. As the employment relationship of the service class in relatively advantageous in terms of employment and payment conditions, occupational security, and promotion prospects, its members have a stake in preserving the status quo. In contrast, the disadvantages of the labor contract can explain why the working class provides a basis of support for the redistributive programs of the left.

This shift to greater complexity in the measurement of class has been accompanied by a similar move away from the measurement of political choice as a dichotomy of left versus right (or left versus non-left) to a fuller representation of the voters' spectrum of choice at the ballot box. Apart from its simplicity, the main reason for the use of a dichotomy to represent voter choice seems to have been a desire to make systematic cross-national and over-time comparisons. Unfortunately, the selective nature of what is being compared undermines any true comparability. The problem is analogous to that faced in the analysis of class position. Changes in the composition of composite categories such as 'left' or 'non-left' may lead to spurious changes in estimates of class voting. The use of dichotomies to represent vote choices and social classes also precludes from observation any processes of class–party realignment. The concept of class *re*alignment in voting implies a change in the pattern of association between class and vote without any change in the overall strength of this association—i.e., without class *de*alignment (or, of course, increased alignment). But this cannot be discerned if the distinction between realignment and dealignment is prevented by restricting the numbers of parties and classes to two.

The final innovation of the 1980s and 1990s is in the statistical measurement of the class–vote association, with a move from Alford type indices to ones that are based on odds ratios, which measure the strength of the relationship between class and vote independently of the general popularity of political parties or changes in the sizes of classes. The use of logistic modeling techniques likewise capitalizes on this basic feature of the odds ratio and allows the relationship between more complex class and party systems to be summarized, thus complementing and facilitating greater sophistication in the representation of class and vote.

Research using these advances comes to rather different conclusions than those in the two-class, two-party tradition. Studies in the UK (i.e., Heath *et al.* 1995) have found little evidence of declining class voting and have concluded that 'trendless fluctuation,' or at most a 'one-off' change, best captures the pattern of association in class voting over time. The USA, by comparison has seen the growth of a new class–vote cleavage—between those who vote and those who do not—as a deficit of working class interest representation that has been associated with increased class differences in participation (Hout *et al.* 1995). Various studies reported in Evans (1999) show that although there is evidence for a linear decline in left versus non-left class voting (Nieuwbeerta and de Graaf 1999), it is not typical. For example, when examined over the long-term, levels of class voting in the UK were found to have increased during the 1940s and 1950s before falling back in the 1960s; while the 1990s have seen increased levels of class voting in new post-communist democracies as these societies have undergone the rigors of marketization. Only in certain Scandinavian countries is there robust evidence of a decline from an unusually high degree of class voting to levels similar to those in other Western democracies.

3. *New Directions*

The debate on class voting has been strong on evidence and weak on theory. Most scholars have assumed a sociological, relatively deterministic account of the transition to industrial and postindustrial politics, but there are those who have rejected these in favor of more voluntaristic models. Kitschelt (1994), for

example, argues that the electoral fortunes of European social democratic parties are largely determined by their strategic appeals, rather than by secular trends in the class structure—a line of reasoning that echoes Sartori's (1969) influential emphasis on the importance of organization, and especially parties, in the creation of class constituencies. From this perspective there is reason to believe that even in advanced industrial societies class voting might increase as well as decrease. It also implies that the adoption of class-relevant policy programs should be associated with an increase in the class basis of partisanship. Some evidence for this has been presented by Evans *et al.* (1999) who show a relationship over time between left-right polarization in parties' manifestos and the extent of class voting.

This is not to say that sociological changes have no impact, however. Changes in the relative *sizes* of classes have been thought to have implications for party strategy: most importantly, in a change to a 'catch-all' strategy by parties on the left in response to the shrinking class basis of support for those parties (Przeworski and Sprague 1986). In some political systems such moves leave open the space for left parties to attract support from marginalized working class groups; in others, such as first-past-the-post systems, we might expect that the start-up costs for electorally viable left parties would be too great. Unfortunately, many of these arguments still await rigorous tests using survey analysis. Research that tries to unravel *why* voters in different classes vote differently is in short supply, though Weakliem and Health (1994) provide evidence of the role of rational choice and inherited preferences, while Evans (1993) models the effects for future expected rewards and the likelihood of promotion conditioned by position in the life-cycle. A more explicit link between models of class voting and their dependence on advances in theories of voting behavior more generally is an area where further development might usefully occur.

See also: Class and Law; Class Consciousness; Class: Social; Interest Groups; Interest Groups, History of; Interest: History of the Concept; Political Sociology; Social Inequality in History (Stratification and Classes); Social Stratification; Voting, Sociology of

Bibliography

Alford R 1967 Class voting in Anglo-American political systems. In: Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments: Cross-National Perspectives*. Free Press, New York, pp. 67–94

Anderson D, Davidson P E 1943 *Ballots and the Democratic Class Struggle*. Stanford University Press, Stanford, CA

Erikson R, Goldthorpe J H 1992 *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Clarendon, Oxford, UK

Evans G 1993 Class, prospects and the life-cycle: explaining the association between class position and political preferences. *Acta Sociologica* **36**: 263–76

Evans G (ed.) 1999 *The End of Class Politics? Class Voting in Comparative Context*. Oxford University Press, New York

Evans G, Heath A, Payne C 1999 Class: Labour as a catch all party? In: Evans G, Norris P (eds.) *Critical Elections: British Parties and Voters in Long-term Perspective*. Sage, London, pp. 87–101

Franklin M N, Mackie T T, Valen H 1992 *Electoral Change: Responses to Evolving Social and Attitudinal Structures in Western Countries*. Cambridge University Press, Cambridge, UK

Goldthorpe J H 1996 Class and politics in advanced industrial societies. In: Lee D J, Turner B S (eds.) *Conflicts about Class: Debating Inequality in Late Industrialism*. Longman, London, pp. 196–202

Heath A F, Evans G, Payne C 1995 Modelling the class/party relationship in Britain, 1964–92. *Journal of the Royal Statistical Society*, *Series A* **158**: 563–74

Heath A, Jowell R, Curtice J 1985 *How Britain Votes*. Pergamon, Oxford, UK

Hout M, Brooks C, Manza J 1995 The democratic class struggle in the United States. *American Sociological Review* **60**: 805–28

Inglehart R 1997 *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton University Press, Princeton, NJ

Kitschelt H 1994 *The Transformation of European Social Democracy*. Cambridge University Press, Cambridge, UK

Lipset S M 1981 *Political Man: The Social Bases of Politics* (expanded and updated edition). Heinemann, London

Manza J, Hout M, Brooks C 1995 Class voting in capitalist democracies since World War II: dealignment, realignment or trendless fluctuation? *Annual Review of Sociology* **21**: 137–62

Nieuwbeerta P, de Graaf N D 1999 Traditional class voting in 20 postwar societies. In: Evans G (ed.) *The End of Class Politics?* Oxford University Press, New York, pp. 23–57

Przeworski A, Sprague J 1986 *Paper Stones: A History of Electoral Socialism*. University of Chicago Press, Chicago

Sartori G 1969 From the sociology of politics to political sociology. In: Lipset S M (ed.) *Politics and the Social Sciences*. Oxford University Press, Oxford, UK, pp. 65–100

Weakliem D L, Heath A F 1994 Rational choice and class voting. *Rationality and Society* **6**: 243–70

G. Evans

Voting: Compulsory

Compulsory voting (CV) is a system of laws and/or norms, mandating that enfranchised citizens turn out to vote, and usually specifying penalties for non-compliance. The number of countries using CV in at least some of their elections is greater than commonly recognized. One recent estimate is that twenty-four countries constituting roughly 17 percent of the world's democracies employ compulsory voting to some extent (Australia 2000).

CV raises questions that strike at the heart of democratic theory. For example, is it 'democratic' to compel citizens to turn out to vote via the threat of financial or social sanction? In what sense are elections conducted under such circumstances 'free and fair'? On the other hand, what legitimacy attaches to elected representatives and their policies when large proportions of the electorate abstain? Responding to this latter question, political scientists have long considered CV an effective mechanism to increase voter turnout (e.g., Gosnell 1930). Scholarly interest in CV was piqued by Arend Lijphart's 1996 Presidential Address to the American Political Science Association, providing a detailed review of CV and a compelling case for its adoption. CV raises numerous issues for students and practioners of electoral politics, which are the focus of this article. Real-world implementations of CV are considered first, followed by an assessment of the consequences of CV for electoral politics, policy, campaigning, and the administration of elections.

1. *Implementations of CV*

Western democracies utilizing CV for at least some of their elections include Australia, Austria (in selected *Länder*), Belgium, Greece, Italy, Liechtenstein, Luxembourg, and Switzerland (just one canton). The Netherlands also had CV up until 1970. In almost all of these Western democracies CV was instituted in the early twentieth century, shortly after the expansion of voter suffrage and the political organization of labor movements in the late nineteenth and early twentieth centuries. At the same time there were some experiments with CV in American states: North Dakota (1898) and Massachusetts (1918) actually amended their constitutions to permit CV, but the respective legislatures did not implement CV in their statutes (Gosnell 1930, pp. 206–7). Abraham (1955) also reviews American flirtations with CV. CV is widespread throughout Latin America; for instance, Argentina, Brazil, Costa Rica, Ecuador, Guatemala, Honduras, Panama, Peru, Uruguay and Venezuela have CV. Cyprus, Egypt, Fiji, Turkey, Thailand, Singapore and the Phillipines also have CV.

1.1 *Inducing Compliance*

Compliance with CV is most commonly induced with a system of small to moderate fines (e.g., Australia and Belgium), sometimes in conjunction with other penalties, although a substantial number of countries do not impose fines. Other penalties include ineligibility for elected office for a prescribed period (e.g., Argentina), or ineligibility for certain types of government employment (e.g., Venezuela), or disenfranchisement (e.g., Thailand). In other countries, CV has more the status of a norm, with sanctions for non-compliance taking the form of social embarrassment, with non-voting noted on official documents (e.g., Italy). In another class of countries CV appears as a constitutional provision or in statute, but with no mention of any penalty for non-compliance (e.g., Bolivia, Chile, Liechtenstein, Egypt). Many countries also have large classes of exemptions. For instance, Australia's federal CV statute exempts non-voters with 'valid and sufficient' reasons for not turning out; Australia's courts have rigorously denied non-voters' claims of indifference between the candidates or alienation from politics as valid and sufficient reasons (AEC 1999). Age exemptions are also common: for instance, Brazil makes voting optional for citizens between the ages of 16 and 18, citizens over the age of 70, and for illiterates.

2. *Consequences of CV*

CV has a direct impact on voter turnout, evident in both aggregate and individual-level analyses. The logic of this is extremely straightforward, with the sanctions of CV offsetting the costs of going to the polls; that is, CV makes the option of not turning out costly, overcoming the fact that turning out is a low benefit activity for many citizens (Lijphart 1997, p. 9).

2.1 *Voter Turnout: Cross-national Evidence*

Even a casual inspection of compendiums of aggregate turnout statistics reveals higher turnout among countries with CV. For instance, a recent collection of data on voter turnout in 171 countries finds turnout roughly six or seven percentage points higher in the 24 countries with CV than in countries without CV (IDEA 1997, p. 32). Multivariate statistical analyses typically find CV to have larger impacts on turnout, by controlling for other institutional and political variables that affect turnout. Lijphart's (1997) review finds CV associated with a boost in turnout rates of seven to sixteen percentage points (for examples of the studies reviewed, see Powell (1981), Jackman (1987), Jackman and Miller (1995), and Franklin (1999), the latter study being distinctive for including an individual-level analysis, exploiting survey data from European Union countries). Among Latin American countries, the estimated turnout boost associated with CV is roughly eleven to seventeen percentage points (Fornos 1996).

These results are striking considering (a) large cross-national differences in institutional and political characteristics of these countries that impact turnout (e.g., Jackman's 1987 study considered competitiveness of elections, electoral disproportionality, number

of political parties, unicameralism vs. bicameralism), and (b) considerable variability in the implementation of CV among those countries that nominally have CV. In addition, of all the institutional variables found to impact voter turnout, CV has far and away the largest effect.

2.2 Voter Turnout: Evidence from Within-country Comparisons

Hirczy (1994, p. 65) makes a compelling argument that cross-national analyses provide 'no causal proof that mandatory voting actually produces high turnout,' and indeed, the causal arrow may be reversed; i.e., a country that adheres to a norm of high turnout simply 'enshrines its civic norm in law.' A research design that overcomes this threat is to compare turnout within countries, before and after the implementation or repeal of CV, or across sub-national units with and without CV. An additional strength of this research design is that, within countries, many of the factors affecting turnout remain constant, regardless of CV.

Studies of this type find CV to have large effects on aggregate turnout. Prior to the implementation of CV in 1924, turnout in elections to Australia's House of Representatives averaged 64.2 percent; in the nine elections following the introduction of CV turnout averaged 94.6 percent, an increase of 30.4 percentage points ($t = 8.7$; author's calculations, using data in Hughes and Graham (1968)). In The Netherlands, the abolition of CV in 1970 caused a drop of roughly 10 percentage points to around 84 percent (Irwin 1974, Hirczy 1994). In addition, the removal of fines for nonvoting in Venezuela in 1993 saw turnout fall by roughly 30 percentage points (Lijphart 1997, p. 9). In Austria, cross-provincial and longitudinal variation in the use of CV permits a powerful assessment of the impact of CV. Turnout in eleven federal parliamentary elections between 1953 and 1987 averaged 92.7 percent in provinces without CV; among provinces with CV turnout averaged 95.7 percent, to yield a treatment effect of 3.0 percentage points ($t = 3.4$), this smaller treatment effect reflecting a 'ceiling effect' with turnout rates bounded from above at 100 percent. A reasonable conclusion based on these studies is that CV's effects are conditional on baseline levels of electoral participation; i.e., CV is likely to have bigger impacts on turnout when other factors predispose a country to low turnout, and vice versa (Hirczy 1994). This insight can be applied in cross-national studies of voter turnout, using interaction terms or multilevel statistical models to make the effects of compulsory voting conditional on other variables.

Public opinion surveys have also been used as proxies for the 'natural experiment' of removing CV. For instance, researchers in a number of countries have asked respondents to engage in thought experiments, in which respondents report their likely behavior in a counter-factual scenario of voluntary turnout. Results from studies of this type suggest modest falls in voter turnout in Australia—to about 88 percent from the current 96 percent figure (Mackerras and McAllister 1996)—but larger falls in Belgium (Hooghe and Pelleriaux 1998), Brazil (Power and Roberts 1995), and Venezuala, as large as 30 percentage points. Jackman (1999) urges caution in interpreting these estimates, arguing that (at least in the Australian case) survey response bias produces an overestimate of voluntary voter turnout, since would-be nonvoters are less likely to respond to a survey on politics in the first place.

2.3 Political Consequences

To gauge the political consequences of CV, most scholars start by noting the near universal connection between political participation and socioeconomic status. Noting that 'low voter turnout means unequal and socioeconomically biased turnout' (Lijphart 1997, p. 2), many scholars see CV as going a long way to removing this 'bias.' Comparative studies of turnout note that the relationship between socioeconomic status and voter turnout weakens as turnout increases (e.g., Powell 1986). Thus to the extent that CV generates higher voter turnout (and the research summarized in the previous sections suggests strongly that it does), CV also removes socioeconomic differences in electoral participation. Quite simply, when almost everyone votes, there can be no socioeconomic 'biases' in turnout.

It also follows that the high turnout rates engendered by CV also have partisan and ideological implications. Since support for parties of the left is correlated with socioeconomic status (notwithstanding a lengthy debate about a decline in 'class voting'), as socioeconomic based differentials in turnout diminish, support for parties of the left increases (Pacek and Radcliff 1995, Nagel 1988, McAllister and Mughan 1986). In turn, increased support for parties of the left results in higher welfare spending and more state interventions in the macroeconomy and labor markets (Hicks and Swank 1992, Castles and McKinlay 1979). Again, to the extent that CV increases turnout, and turnout increases support for left parties and their policy agendas, then CV can be shown to have policy implications. Put simply, CV helps mobilize voters who would otherwise abstain, generating a distribution of voters' policy preferences further to the left than would otherwise result, and so party competition takes place over an interval of the ideological spectrum shifted to the left from where it would otherwise be. CV is thus causally quite remote from outcomes such as left-party success or particular policy outcomes. Indeed, this causal link could even be reversed: i.e., it is only in countries with commitments to social democratic outcomes that CV is adopted, and

in fact, this accords with the pattern of adoption of CV among Western democracies in the early twentieth century.

2.4 Other Consequences

One criticism of CV is that it compels the participation of disinterested and hence poorly informed citizens who would otherwise abstain. Lijphart's (1997, p. 10) response to this is criticism is that CV 'may serve as an incentive to become better informed'; a cross-national study by Gordon and Segura (1997) finds a small though statistically significant increase in political sophistication in countries with CV, but otherwise the evidence for CV promoting greater civic awareness is scant. Another untested argument along these lines is that CV leads to 'higher quality' political campaigns, since the parties are less concerned with mobilizing voters and more with the conversion of voters that CV guarantees will go to the polls. Lijphart (1997, p. 10) speculates that this diminishes (a) the role of money in political campaigns, and (b) incentives for candidates to use attack advertising.

Students of Australian politics have speculated whether CV leads to higher levels of party identification than might otherwise result (Mackerras and McAllister 1999), consistent with a view of voters as 'cognitive misers,' looking for heuristics to help them deal with a forced choice among parties. On the other hand, Australia's minor parties are obvious beneficiaries and keen supporters of CV, providing an alternative for voters dissatisfied with major parties, but legally compelled to turn out. Another long-standing feature of CV is a higher rate of invalid ballots (e.g., Tingsten 1937) and 'donkey ballots' (where voters simply select the first name they see on the ballot); some instances of these phenomena are protests against CV itself.

Two final points. First, CV places an onus on citizens, but the state typically reciprocates with institutional mechanisms that make it easy to comply (e.g., weekend voting, ease of registration, widespread use of postal ballots): it would be rather odd if a state made voting compulsory but did not employ these other turnout-enhancing institutional mechanisms. According to Gosnell (1930, p. 209), 'fines and penalties under a system of compulsory voting are a minor matter. The important feature of the system is that voting is regarded as a civic duty and the government does everything to impress upon voters this point of view.' Second, the more seriously a country enforces its CV provisions, the more bureaucratic resources are required to maintain registration records and ensure compliance. In Australia, these two functions—the 'carrot and stick' of CV, as it were—are administered by the Australian Electoral Commission, a large and highly professional bureaucracy responsible for all aspects of Australian federal elections. Ensuring compliance with CV is just one of many AEC functions, and the bulk of its activities are to do with other aspects of election administration (e.g., redistricting, voter registration, public financing of campaigns, ballot design, staffing polling stations, vote tallying). Thus one (perhaps unintended) consequence of an aggressive system of CV is the centralization and professionalization of election administration. In turn, this may mitigate the problems that accompany decentralized and non-professional election administration, clearly evident in the aftermath of the 2000 US presidential election.

See also: Democracy; Democracy, History of; Democracy: Normative Theory; Electoral Systems; Political Representation; Representation: History of the Problem; Voting: Class; Voting: Issue; Voting, Sociology of; Voting: Tactical; Voting: Turnout

Bibliography

Abraham H 1955 *Compulsory Voting*. Public Affairs Press, Washington, DC

AEC 1999 *Compulsory Voting. Number 8 in Electoral Backgrounder*. Australian Electoral Commission, Canberra

Australia, Joint Standing Committee on Electoral Matters 2000 *The 1998 Election: Report of the Inquiry into the Conduct of the 1998 Federal Election and Matters Related Thereto*. Parliament of the Commonwealth of Australia, Parliament House, Canberra

Castles F, McKinlay R D 1979 Does politics matter: An analysis of the public welfare commitment in advanced democratic states. *European Journal of Political Research* **7**: 169–86

Fornos C 1996 Explaining voter turnout in Latin America. Master's thesis, Lousiana State University

Franklin M 1999 Electoral engineering and cross-national turnout differences: What role for compulsory voting? *British Journal of Political Science* **29**: 205–24

Gordon S B, Segura G M 1997 Cross-national variation in the political sophistication of individuals: Capability or choice. *Journal of Politics* **59**: 126–47

Gosnell H 1930. *Why Europe Votes*. University of Chicago Press, Chicago

Hicks A M, Swank D H 1992 Politics, institutions, and welfare spending in industrialized democracies, 1960–82. *American Political Science Review* **86**: 658–74

Hirczy W 1994 The impact of mandatory voting laws on turnout: A quasi-experimental approach. *Electoral Studies* **13**: 64–76

Hooghe M, Pelleriaux K 1998 Compulsory voting in Belgium: An application of the Lijphart thesis. *Electoral Studies* **17**: 419–24

Hughes C, Graham B D 1968 *A Handbook of Australian Government and Politics, 1890–1964*. Australian National University Press, Canberra

IDEA 1997 *Voter Turnout from 1945 to 1997: A Global Report on Political Participation*. International Institute for Democracy and Electoral Assistance, Stockholm

Irwin G 1974 Compulsory voting legislation: Impact on voter turnout in the Netherlands. *Comparative Political Studies* **7**: 292–314

Jackman R W 1987 Political institutions and voter turnout in the industrial democracies. *American Political Science Review* **81**: 405–23
Jackman R W, Miller R A 1995 Voter turnout in the industrial democracies during the 1980s. *Comparative Political Studies* **27**: 467–92
Jackman S 1999 'Non-compulsory voting in Australia? What surveys can (and can't) tell us. *Electoral Studies* **18**: 29–48
Lijphart A 1997 Unequal participation: Democracy's unresolved dilemma *American Political Science Review* **91**: 1–14
Mackerras M, McAllister I 1999 Compulsory voting, party stability and electoral bias in Australia. *Electoral Studies* **18**: 217–33
McAllister I, Mughan A 1986 Differential turnout and party advantage in British general elections, 1964–1983. *Electoral Studies* **5**: 143–52
Nagel H 1988 Voter turnout in New Zealand general elections, 1928–1988. *Political Science* **40**: 16–38
Pacek A, Radcliff B 1995 Turnout and the vote for left-of-centre parties: A cross-national analysis. *British Journal of Political Science* **25**: 137–43
Powell G B Jr. 1981 Party systems and political system performance: voting participation, government stability and mass violence in contemporary democracies. *American Political Science Review* **75**
Powell G B Jr. 1986 American voter turnout in comparative perspective. *American Political Science Review*
Power T, Roberts J 1995 Compulsory voting, invalid ballots, and abstention in Brazil. *Political Research Quarterly* **48**: 795–826
Tingsten H 1937 *Political Behavior: Studies in Election Statistics*. P. S. King and Son, London

S. Jackman

Voting: Issue

Who should pay what in taxes? Should immigration be strictly controlled? How effectively is the economy being managed? Should women have a guaranteed right to an abortion during the first trimester of pregnancy? Each of these is commonly understood as a political issue. A political issue is defined as a policy-based controversy relevant to the mass public's evaluation of either political parties or candidates for public office. The link between citizens and their elected leaders that arises from political issues lies at the very heart of democratic governance. This article explores that link.

1. *Position and Valence Issues*

Issues can be usefully divided into two basic types: position and valence (Stokes 1963). Position issues are those on which there is meaningful disagreement among members of the mass public with regard to alternative policies. Questions of taxes, immigration, abortion, health insurance, and environmental policy are among those that usually fall within the position issue rubric, with different people having quite different preferences about government policy. Valence issues are those on which there is general agreement, usually about a goal, and the main controversy concerns which party or candidate is more likely to achieve the desired result. Virtually everyone wants a well-managed economy and safe streets; the question is who is more likely to deliver the outcome.

Assessments of parties and candidates on valence issues tend to be retrospective, that is, based on past performance (Fiorina 1981). Because such assessments are highly sensitive to the state of the times, they often have a 'tidal' character, increasing or decreasing support for a party across virtually all groups in a country. A good economy tends to help incumbents, while a bad economy or large-scale social unrest hurts. Thus, valence issues can be very important in determining election outcomes. Retrospective assessments of performance are also relatively easy to make as they require little knowledge beyond who is in office and how things are going in the country. Thus, while retrospective assessments are widely used by all types of voters in evaluating governments, the least engaged voters tend to rely most heavily on short-term retrospective judgments to guide their choice (Sniderman et al. 1991).

Position issues intrinsically require more information from the electorate. At a minimum, people need to have some preference on an issue and some sense of where the parties or candidates stand on the issue, in order for issues to affect their political judgment. While there are some very knowledgeable voters, the knowledge level of the average citizen is fairly low (Delli Carpini and Keeter 1996). This does not rule out a role for position issues in elections, but it does suggest that the role issues play will not rely on a highly informed electorate (Campbell et al. 1960).

2. *Political Parties and Social Groups*

Political parties and social groups help voters make sense of political issues. For the most part, parties form around core social and ideological cleavages in society (Lipset and Rokkan 1967). They endure across elections and tend to maintain fairly consistent orientations towards issues. Given the low attention most people pay to politics, the long-standing consistency of parties on most issues provides a critical link for allowing voters to make policy-oriented vote choices. Parties also enhance the accurate use of valence-type performance issues, as they provide the over time continuity necessary for voters to reward good performance and to punish bad performance.

Social groups can also assist citizens in casting meaningful policy votes. Labor, business, religious, linguistic, and racial/ethnic groups traditionally have

been important sources of cues for voters. Feelings toward a group and awareness of the group's position on an issue can serve to alert the voter to the issue, and can help determine the voter's own position.

3. Low-information Rationality

The low-information level of voters does not imply an unthinking or irrational electorate. In general, voters attempt to make reasonable decisions based on the information they possess (Key 1966, Popkin 1991, Lupia and McCubbins 1998). This idea is encapsulated in the phrase 'low-information rationality,' which suggests that the main element limiting the quality of the electorate's judgments is the limited amount of information voters possess. Since information is costly to acquire and no single vote is likely to be influential, the strategy of making good use of limited information is a rational strategy for voters to adopt (Downs 1957).

This in turn implies that certain types of issues will be used more readily by voters. Issues that have a straightforward moral structure (such as abortion) or those that fit into a clear symbolic framework (such as the common man vs. the wealthy) are cognitively simpler and easier to process. Hence, voters are more likely to use them in making decisions (Converse 1964).

It is also important to realize that voters themselves differ widely. More knowledgeable voters and those concerned about a particular issue are more able to process issue information (Converse 1964, Krosnick 1991, Zaller 1992). Nevertheless, the predominant characterization of the electorate is one with low information and relatively little attention to politics. And even with increasing levels of education, there has been no discernable increase in levels of political knowledge (Delli Carpini and Keeter 1996).

4. The Directional and Proximity Models of Issue Voting

How do voters connect their issue preferences to party or candidate choice? Two models offer competing explanations for how position issues operate in electoral settings. The directional model argues that successful issue campaigns develop clear strong messages to attract voters to a party (Rabinowitz and Macdonald 1989, Macdonald et al. 1991). The proximity model argues that parties must support centrist policies to be electorally successful (Downs 1957, Davis et al. 1970).

4.1 The Directional Model

According to the directional model, voters operate in a simplified issue environment in which issues have just two sides. For example, a voter might be in favor of restricting immigration or opposed to restricting immigration, or she might not care. Voters differ in terms of which side of the issue they favor and the intensity with which they favor that side. When a party takes a position on the side of the issue a voter favors, it is evaluated favorably by the voter; when a party takes a position on the opposite side, the voter evaluates it unfavorably. Voters who are more intense about an issue react more strongly than less intense voters: they like parties or candidates on their side more and display more hostile feelings toward those they oppose. Similarly, when a party takes a more intense stand on an issue, it arouses more intense feelings. It garners more favorable evaluations from voters on the same side of the issue and more unfavorable evaluations from voters on the opposite side. In mathematical terms, the product of the voter's and the party's intensity determines how strongly a voter feels about a party, either favorably or unfavorably. This effect is mediated, however, if a party goes beyond an extremity threshold.

In directional theory, parties and candidates are constrained in their intensity by the need to avoid being viewed as irresponsible or extremist. In order to draw support based on an issue, parties and candidates must lay out relatively strong stands emphasizing the policy directions they favor; at the same time they must show sufficient moderation to suggest they will be responsibly pragmatic in office. In the directional model a hypothetical voter asks, 'Are you on my side of the issues?' and 'Can I trust you to be responsible?'

4.2 The Proximity Model

The proximity model contrasts with the directional model. According to the proximity model, voters consider an ordered set of policy options on an issue. With regard to the environment, for example, the options might be a series of increasingly tougher policies regulating factory emissions. The voter is assumed to have a preference for a particular policy among the set. This is different from the more diffuse, two-sided view ('favor or oppose new environmental regulations') that characterizes directional theory. Further, in proximity theory each party is assumed to have a distinct policy position on the issue that is known to the voter. Voters evaluate parties based on the distance between the party's position and their own position. In the proximity model a hypothetical voter asks, 'How close are your positions to mine?'

The proximity and directional models are most distinct with regard to center voters. According to proximity theory voters in the center are committed to centrist policies and support parties that take centrist stands. According to directional theory voters in the center are simply neutral between the two sides and don't care which position the parties take.

4.3 Issue Strategies

4.3.1 Two-party systems. In both directional and proximity theory, the voters most critical to determining strategy in two-party contests are those in the middle of the voter distribution. In the simplest case where a single dimension characterizes the competition the critical voter is the median voter. The median is literally the voter in the middle. If, for example, there are nine voters and they are arrayed from furthest left to furthest right, the fifth voter is the median. Note that the voter in the middle of the distribution is not necessarily in the center on the issue—that depends on where the voters are located.

In an unidimensional competition, winning the median voter will guarantee victory. In the more realistic multidimensional case a single critical middle voter is unlikely to exist. Nevertheless, the principle that the voters in the middle will decide the election continues to hold, as they are the key to capturing a majority of votes.

The need to court the same constituency, however, does not imply that the two theories dictate similar strategies. In directional theory, parties must be attentive to the balance of voters on the two sides of an issue. If a clear plurality favors one side, then the middle voters will be on that side and it is advantageous for the parties to be there too. If both parties are on the plurality side, then the party that takes the stronger stand without becoming irresponsible will attract the most voters. Just as a strong magnet is more effective than a weak magnet in drawing objects to it, so a party with strong cues is more effective in garnering the support of like-minded voters.

When the electorate is divided evenly on an issue, then parties are not constrained in which side they take, because support for the two sides will tend to balance out. In directional theory it can make strategic sense for parties to diverge on evenly divided issues: there is no net gain or loss in supporting one side or the other, and by taking a distinct stand a party can develop enthusiasm from a set of potential voters.

In proximity theory parties should always adopt a position close to the middle set of voters. This follows from the fact that proximity is the key to support. If Party A is close to the middle voters and Party B is not, then the middle voters will support Party A, as will all the voters who lie on the opposite side of Party A from B, giving a majority to Party A. Thus, according to the proximity model, any party seeking to maximize votes must stay close to the middle of the voter distribution. A party that deviates from the middle on any issue is open to defeat.

4.3.2 Multiparty systems. In multiparty systems proximity theory makes no clear predictions about party strategy. Directional theory suggests that the most successful parties will take relatively strong clear stands on some issues, while avoiding an extremist label. Extremist parties can be successful in multiparty systems in that they can attract enough voters to win seats in proportional elections; however, their appeal will be limited because of their extremism.

Centrist parties are expected to be weak in directional theory, because they fail to develop strongly favorable evaluations from any voters. Some parties that are centrist on left-right are well defined on other issues, such as agricultural policy, and these parties can be successful. Parties with no strong issue stands, however, will find that while few voters dislike them, they are not the first choice of many voters.

4.3.3 Issue salience and issue ownership. Some issues are more important in elections than others. Sometimes this happens by accident and sometimes by design. In any event, if a party has an advantage on an issue compared to its competitors, then it helps the party for that issue to be more salient in the election. This is true in both two-party and multiparty systems, and under both the directional and proximity models.

Related to the idea of issue salience is the idea of issue ownership (Budge and Farlie 1983, Petrocik 1996). This concept conveys the view that parties develop reputations for being especially credible on certain issues, and thereby come to 'own' them, for example, the Labour Party in Britain on social welfare issues and the Christian People's Party in Norway on questions of public morality. Whenever the issue a party owns becomes the focus of a campaign, the party benefits. The concept of issue ownership fits well with a directional view of issue voting as parties tend to take strong stands on 'their' issues. The manipulation of issue salience, however, is important to both directional and proximity theory, and is a key element in virtually all political campaigns.

4.4 Which Theory is Correct?

The empirical evidence in most advanced industrial societies generally has supported the directional over the proximity model. Directional theory has been more able to predict how voters evaluate parties and candidates, and has more accurately predicted the type of parties (noncentrist and nonextreme) that have succeeded in most advanced industrial democracies. As yet, however, there is no consensus in the research literature on which of the two theories better represents the relationship between issues and mass support (Lewis and King 2000). In some formulations the theories are difficult to disentangle, in particular in predicting vote choice, and recent work in proximity theory has incorporated features that provide more

overlap with directional theory (Hinich and Munger 1994). Some scholars have found elements of each theory attractive, and have blended the models in an effort to provide a fuller understanding of electoral politics (Rabinowitz and Macdonald 1989, Merrill and Grofman 1999).

5. *Conclusion*

In democracies voters are presumed to influence public policy through elections. Underlying this view is the assumption that voters' issue preferences influence their vote choice. Research has shown, however, that the majority of voters, even in highly educated advanced industrial democracies, do not follow politics closely and do not have high levels of political information. The impact of issues, therefore, requires voters to use their limited information effectively. They are aided in this process by the presence of political parties that provide coherence to political systems. Further, meaningful policy-based votes appear to require considerable stimulation by political elites (either parties or candidates) to activate issues.

Two competing models, the directional model and the proximity model, have been proposed to describe the way in which voter preferences and party positions interact to determine voter choice. The alternative theories have quite different implications for effective party strategies and the nature of mass control of public policy. The question of which theory, or combination of theories, most accurately describes the role of issues in electoral systems is a matter of current controversy.

See also: Media Effects; Political Communication; Populism; Public Opinion: Microsociological Aspects; Public Opinion: Political Aspects; Rational Choice in Politics; Voting, Sociology of

Bibliography

Budge I, Farlie D J 1983 *Explaining and Predicting Elections*. Allen and Unwin, London

Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. Wiley, New York

Converse P E 1964 The nature of belief systems in mass publics. In: Apter D E (ed.) *Ideology and Discontent*. Free Press of Glencoe, London

Davis O, Hinich M J, Ordeshook P 1970 An expository development of a mathematical model of the electoral process. *American Political Science Review* **64**: 426–48

Delli Carpini M X, Keeter S 1996 *What Americans Know about Politics and Why It Matters*. Yale University Press, New Haven, CT

Downs A 1957 *An Economic Theory of Democracy*. Harper, New York

Fiorina M P 1981 *Retrospective Voting in American National Elections*. Yale University Press, New Haven, CT

Hinich M J, Munger M C 1994 *Ideology and the Theory of Political Choice*. University of Michigan Press, Ann Arbor, MI

Key V O Jr. 1966 *The Responsible Electorate: Rationality in Presidential Voting, 1936–1960*. Belknap Press of Harvard University Press, Cambridge, MA

Krosnick J A 1991 The stability of political preferences: Comparisons of symbolic and nonsymbolic attitudes. *American Journal of Political Science* **35**: 547–76

Lewis J B, King G 2000 No evidence on directional vs. proximity voting. *Political Analysis* **8**: 21–33

Lipset S M, Rokkan S 1967 Cleavage structures, party systems, and voter alignments: An introduction. In: Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments*. Free Press, New York

Lupia A, McCubbins M D 1998 *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge University Press, Cambridge, UK

Macdonald S E, Listhaug O, Rabinowitz G 1991 Issues and party support in multiparty systems. *American Political Science Review* **85**: 1107–31

Merrill S III, Grofman B 1999 *A Unified Theory of Voting*. Cambridge University Press, Cambridge, UK

Petrocik J R 1996 Issue ownership in presidential elections, with a 1980 case study. *American Journal of Political Science* **40**: 825–50

Popkin S L 1991 *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. University of Chicago Press, Chicago

Rabinowitz G, Macdonald S E 1989 A directional theory of issue voting. *American Political Science Review* **83**: 93–121

Sniderman P M, Brody R A, Tetlock P E 1991 *Reasoning and Choice: Explorations in Political Psychology*. Cambridge University Press, UK

Stokes D E 1963 Spatial models of party competition. *American Political Science Review* **57**: 368–77

Zaller J R 1992 *The Nature and Origins of Mass Opinion*. Cambridge University Press, Cambridge, UK

S. Macdonald and G. Rabinowitz

Voting: Retrospective/Prospective

In the 1950s, leading democratic theorists, under the aegis of the American Political Science Association, extolled their model of a 'responsible party system,' wherein parties vying for office would propose to the voters the policies they intended to carry out if elected. Informed by coherent ideological principles, their policy programs would be specified clearly, and significantly different from each other. The winner of the election would consequently be that party or candidate whose policy proposals the majority of voters found most attractive.

Research carried out in the ensuing decades demonstrated the unrealistic nature of this ideal. Politicians blur, equivocate, and traffic in vague generalities. When they do proffer comprehensible proposals, the centrifugal pull of the median voter induces them to

offer much the same thing (Black 1958). Voters, in turn, typically lack the information, stable preferences, and cognitive sophistication that 'issue voting,' i.e., choosing on the basis of competing policy proposals, would require (Converse 1964).

While many observers fretted that these findings bode ill for the practice of democracy, others reasoned that it was this particular model of electoral competition, and not the voters and politicians, that was deficient. They argued that, rather than consider the merits of nebulous, hypothetical policy alternatives, voters decide to either retain or reject incumbent politicians on the basis of the objective conditions that obtained during their tenure in office (Downs 1957, Key 1966, Fiorina 1981). This theory of *retrospective voting* has three major features. First, voters need to acquire little information over and above that which they assimilate in their daily lives. Second, it is the results of policies that voters evaluate, not the policies themselves. Third, elections are primarily referenda on the performance of incumbent officeholders, and so provide little guidance as to the future course of public policy. As Key (1966) put it, '(the electorate) judges retrospectively; it commands prospectively only insofar as it expresses either approval or disapproval of that which happened before.'

There is any number of indicators of well-being to which voters might respond—the crime rate, environmental quality, combat deaths, etc. Research on retrospective voting, however, has focused primarily upon economic conditions. Virtually all of it can be placed in one of two categories: Time-series analyses of aggregate economic and voting data, and analyses of responses to opinion surveys, in either cross-sectional or panel form.

1. *Time-series Analyses*

In one of the most influential studies in the history of political science, Kramer (1971) demonstrated that the better the recent performance of the American economy, the larger the share of votes won by congressional candidates of the incumbent president's party. Subsequent time-series studies of aggregate economic and voting data found that the electoral fortunes of the president himself depend importantly upon economic fluctuations (Fair 1978), as do the fates of candidates of his party running in senatorial and gubernatorial elections.

Most subsequent research followed Kramer's lead in several important ways. Incumbency is defined by the party of the president, not by whether those running for lesser office are themselves officeholders. The analyses assume voters have short time horizons—support for the incumbent party is a function of economic performance in the year or so running up to the election. Finally, the economic variable they key upon is usually per capita income change, the variable Kramer found to have the largest impact upon election outcomes.

These studies have produced overwhelming evidence as to the existence of economic retrospective voting. But it is not a uniform, ubiquitous phenomenon. In the USA, favorable economic conditions clearly benefit the incumbent president, and in on-year elections, congressional candidates of the president's party also benefit . The economy has a smaller impact upon midterm elections, however, as many voters appear to be seeking to counterbalance the ideological orientation of the incumbent president (Alesina et al. 1993, Erikson 1990). Economic retrospective voting is also stronger and more consistent in some countries than in others (Paldam 1991). According to Powell and Whitten (1993), the strength of economic retrospective voting is a function of 'clarity of responsibility' in the institutions of government. Where it is lacking, voters have difficulty in deciding whom to reward and whom to punish, thus undermining their ability to vote on the basis of economic performance. Clarity of responsibility is obscured by multiple parties in the governing coalition, minority governments, weakly cohesive parties, opposition party control of a policy-making institution (either a separately elected presidency or house of the legislature), and opposition party chairs of parliamentary committees. These features arise most frequently in proportional representation regimes.

Despite the vast amount of research that has been done on aggregate voting and economic data, there are weaknesses in this approach that make the findings less robust, less definitive, and less cumulative than one would suppose. First, although the economic measures used in these studies are perceived to be quite precise, the more historical data—particularly for years prior to World War II—are not. Kiewiet and Udell (1998) report that the electoral impact of unemployment more than doubles when a more accurate, revised series is used in place of the official Bureau of Labor Statistics series.

Second, the data series are very short, and many studies have examined even smaller subsets of the available observations (e.g., postwar mid-term elections). Researchers also omit 'unusual' election years, and make idiosyncratic choices as to the beginning and ending years of their series. Even small changes in the data series can affect results dramatically. For these same reasons, aggregate-level economic retrospective voting models, though touted as forecasting tools, are not. In such models, error bars are usually around plus or minus 5 percent, and due to sample selection effects this is surely an under-estimate. While an improvement over guesses based upon historical averages, forecasts containing that much error are neither impressive nor useful.

A reasonable response to this problem is to analyze pooled, cross-section time series of disaggregated data. Few studies have adopted this approach, but those

that have underscore the value of more data even when observations are not independent. Rosenstone's (1982) analysis of state-level data finds the electoral effects of income change to be close to the average of the effects reported in various studies using aggregate time series data, but his estimates are far more precise. Peltzman's (1990) also analyzes state-level data, but his findings are at odds with the bulk of the aggregate-level findings. Specifically, he concludes that voters base their decisions on what transpires economically during the entirety of the incumbent president's entire administration, and not just the last year or so.

Peltzman's study points to another potential deficiency in the time-series literature. Most studies posit simple retrospective voting rules on the part of the electorate, the null hypothesis being that recent trends in economic conditions have no impact upon voting decisions. A few studies have found support for alternative hypotheses that imply greater 'rationality,' either in terms of expectations formation, as in the Peltzman study, or that voters punish incumbents for deviating from a prescriptive fiscal and monetary policies (Chappell and Keech 1985). However, in an explicit comparison of the 'naïve' retrospective model vs. a 'rational' model positing that voters differentiate between systematic growth and exogenous income shocks, Alesina et al. (1993) conclude that it is the naïve model that best comports with the data.

2. *Opinion Survey Analyses*

Survey researchers have sought to account for the time-series evidence of economic retrospective voting by identifying the precise ways in which individual voters' economic evaluations influence their voting decisions. The hypothesis receiving the most initial attention was that voters 'voted their pocketbooks,' reacting primarily to trends in their own personal economic circumstances. This seemed to be a natural extension of the simple retrospective voting hypothesis that has dominated the aggregate-level time series analyses, as it posits (a) low information requirements—voters base their decisions to support or oppose incumbents on the basis of information acquired in the course of their daily lives, (b) what matters to voters are objective conditions, not campaign promises, and (c) voters are most reactive to what they have experienced recently.

What researchers found, however, was that whether or not voters were personally doing well or poorly economically had a surprisingly weak impact upon their reported voting decisions. The choice to retain or to reject the present incumbents was instead conditioned much more strongly upon their assessments of recent trends in the nation's economy (Kinder and Kiewiet 1979). Although many scholars understood these findings as showing that voters vote according to what they perceive to be the interests of the country as a whole rather than on the basis of narrow self-interest, this inference is unfounded. Voters may be motivated purely by concern for their own individual welfare, but realize that the performance of the economy overall is a superior indicator of the efficacy of the incumbent's policies. Peltzman (1990) puts it this way:

> To a crude approximation, the president's policy affects mainly aggregate income rather than its personal distribution. If voters understood this, they would respond mainly to aggregate income surprises and ignore personal deviations from the aggregate. Voters care about these personal deviations, but they are mainly irrelevant in evaluating the *policy* of their agent.

Many studies of survey data surmised that the relationships they observed between economic concerns and voting decisions were too weak to account for the aggregate-level, time-series findings. Seldom, however, did they make the calculations necessary to determine if this were actually the case. The few studies that do detect no gross inconsistencies. Indeed, after correcting for measurement error in the survey data, Markus (1988) identifies individual-level effects that would produce interelection vote swings even larger than those observed by Kramer (1971) and Fair (1978).

There are other important ways in which survey analyses of economic retrospective voting have paralleled the aggregate-level time series literature. In a study of five major Western European democracies, Lewis-Beck (1988) finds that the strength of economic retrospective voting was inversely related to the number of parties in the governing coalition. Like Powell and Whitten (1993), he reasoned that multiparty regimes obscure responsibility, making it difficult for voters to determine to whom to attribute credit or blame. Second, other survey researchers have looked beyond simple retrospective voting for evidence of other ways in which economic concerns might affect voting decisions. The most prominent of these alternatives is the prospective voting hypothesis, i.e., that voters base their decision upon how they expect the economy to perform in the near future (MacKuen et al. (1992)). But here, too, direct comparisons tend to favor the retrospective hypothesis. Specifically, Anderson and Wlezian (1997) report that much of the correlation between economic expectations and vote choice arises from respondents' attitudes concerning the incumbent president coloring their expectations about the future. Retrospective evaluations, in contrast, appear to be more resistant to cognitive consistency pressures. Many supporters of Jimmy Carter may have been confident that the economy would turn around in 1981, but few could convince themselves that 1980 had been a good year.

Finally, survey research on economic retrospective voting, as in the case of time-series analyses, has been hampered by deficiencies in the data. Here, however, the problem is the small number of survey questions

available for analysis rather than a small number of observations. The only individual-level data concerning economic evaluations and vote choice that have been gathered on an ongoing, systematic basis or those contained in the National Election Studies, which began in the early 1950s. The questions they ask are reasonably good, as survey questions go, and we are indeed extremely fortunate to have them. Nonetheless, we are left with the nagging doubt that differently posed questions might reveal more accurately the extent of economic retrospective voting.

See also: Voting, Sociology of

Bibliography

Alesina A, Londregan J, Rosenthal H 1993 A model of the political economy of the United States. *American Political Science Review* **87**: 12–33

Black D 1958 *The Theory of Committee and Elections*. Cambridge University Press, Cambridge, UK

Chappell H W, Keech W R 1985 A new view of accountability for economic performance. *American Political Science Review* **79**: 10–27

Converse P 1964 The nature of mass belief systems in mass publics. In: Apter D E (ed.) *Ideology and Discontent*. The Free Press, London

Downs A 1957 *An Economic Theory of Democracy*. Harper & Row, New York

Erickson R S 1990 Economic conditions and the congressional vote: A review of the macrolevel evidence. *American Journal of Political Science* **34**: 373–99

Fair R 1978 The effect of economic events on votes for president. *Review of Economics and Statistics* **60**: 159–73

Fiorina M P 1981 *Retrospective Voting in American National Elections*. Yale University Press, New Haven, CT

Key V O 1966 *The Responsible Electorate*. Belknap Press of Harvard University Press, Cambridge, MA

Kiewiet D R, Udell M 1998 Twenty-five years after Kramer: An assessment of economic retrospective voting based upon improved estimates of income and unemployment. *Economics and Politics* **10**: 219–48

Kinder D, Kiewiet D R 1979 Economic discontent and political behavior: The role of personal grievances and collective economic judgments in congressional voting. *American Journal of Political Science* **23**: 495–527

Kramer G B 1971 Short-term fluctuations in U.S. voting behavior, 1896–1964. *American Political Science Review* **65**: 131–43

Lewis-Beck M S 1988 *Economics and Elections: The Major Western Democracies*. University of Michigan Press, Ann Arbor, MI

MacKuen M, Erikson R, Stimson J 1992 Peasants or bankers: The American electorate and the US economy. *American Political Science Review* **86**: 597–611

Markus G B 1988 The impact of personal and national economic conditions on the presidential vote: A pooled cross-sectional analysis. *American Journal of Political Science* **32**: 137–54

Paldam M 1991 How robust is the vote function? A study of seventeen nations over four decades. In: Norpoth H, Lewis-Beck M S, Lafay J-D (eds.) *Economics and Politics: The Calculus of Support*. University of Michigan Press, Ann Arbor, MI

Peltzman S 1990 How efficient is the voting market? *Journal of Law and Economics* **33**: 27–64

Powell G B, Whitten G D 1993 A cross-national analysis of economic voting taking account of the political context. *American Journal of Political Science* **37**: 391–414

Rosenstone S J 1982 *Forecasting Presidential Elections*. Yale University Press, New Haven, CT

Wlezien C, Franklin M, Twiggs D 1997 Economic perceptions and vote choice: Disentangling the endogeneity. *Political Behavior* **19**: 7–17

D. R. Kiewiet

Voting, Sociology of

Because democracies choose their governments via the ballot box, the bulk of the literature on voting behavior deals with its decisional component, focusing on who votes and for whom. This article presents a comparative assessment of research on electoral choice and electoral turnout, along with the main theories and empirical findings in the field.

1. Paradigms Explaining Voting Choice

Electoral geography was the first, and for a long time the only, clue for understanding political behavior, linking voters' choices to the social structures and the religious and political traditions of the place where they lived. Since the earliest studies of André Siegfried, Charles Gosnell, or Herbert Tingsten, this 'contextual' or 'ecological' approach, based on the analysis of aggregate electoral data, has gained in statistical sophistication, as illustrated by the school of Johnston in Britain or King in the United States. Its limitation lies in the fact that the unit of analysis considered is the group—people living in the same constituency, the same district, the same region, etc.—and that the correlations observed at the aggregate level do not necessarily hold true at the level of the individuals, as Robinson's (1950) famous article on the 'ecological fallacy' reminds us. Opinion polls provide a better tool for understanding individuals' behavior, asking them directly how and why they vote. Since World War II, survey-based research has given a new impetus to electoral sociology, especially in the United States, where the pioneer studies and the classic models explaining vote choice were developed.

The 'sociological' model originates from the local surveys conducted by Paul Lazarsfeld and his colleagues at the Bureau for Applied Social Research of the University of Columbia. Repeated interviews with a randomly selected panel of voters, first in the district of Erie County, Ohio during the 1940 Roosevelt–Wilkie presidential campaign, then in the town of

Elmira, New York, at the time of the Truman–Dewey contest reveal, to their surprise, that the campaign has but a limited effect on voters' choice. Mostly it reinforces long-standing and diffuse political preferences which reflect the norms of the groups they belong to. Constructing an index of political predisposition based on social class, religion, and rural/urban residence, the authors of *The People's Choice* can predict accurately how people vote. Those combining a rural residence, Protestant affiliation, and high social status tend to support the Republicans, while those who are working class, nonprotestant, and residing in an urban area mostly vote Democrat. The main conclusion of their book is that 'A person thinks, politically, as he is, socially. Social characteristics determine political preference' (Lazarsfeld et al. 1944). And this social anchoring accounts for the relative stability of voting choice, except among the cross-pressured or poorly integrated voters. However, in his later empirical and theoretical work such as *Voting* (Berelson et al. 1954), Lazarsfeld was to develop a more complex view of action in general and more particularly of voting behavior (his research evolution is related in Lautman and Lécuyer, 1998).

Reacting against what they see as a social determinism, Angus Campbell and his colleagues from the Survey Research Center (SRC) at the University of Michigan propose a psychological model focusing on individual decision-making. Reconstructing the chain of events that converge in what they call the 'funnel of causality,' they give special attention to the attitudes formed towards candidates, issues, and parties. On the basis of nationwide panel surveys conducted in the presidential elections from 1948 to 1956, they state in *The American Voter* (Campbell et al. 1960) that party identification is the key factor orienting vote choice. By that they mean an affective attachment to a specific party, formed early in life, generally passed over from parents to children, and reflecting the individual's social and religious background. It functions as a filter coloring their political perceptions and guiding their votes, especially for the least sophisticated: 'Like the automobile buyer who knows nothing of cars except he prefers a given make, the voter who knows simply that he is a Republican or Democrat responds directly to his stable allegiance without the mediating influence of perceptions he has formed on the objects he must choose between' (Campbell et al. 1960). Lazarsfeld portrayed voters as social conformists; the Michigan team sees them more as political believers. Yet their conclusions are very similar: the average voters are far from the democratic ideal of a rational and active citizen. Most of them are not interested in politics, display low levels of ideological awareness, and are unfamiliar with the candidates and the issues at stake. They vote according to long-standing political preferences shaped by their early socialization.

The third approach stems from Downs's seminal essay, *An Economic Theory of Democracy* (1957). It postulates, in the line of the neo-classic economic paradigm, that voters are rational and that their choice is based on a cost-benefit calculation. Voters cast their votes for the party or the candidate from which they expect the maximum 'utility' or benefits. The electoral decision between party A in office and opposition party B boils down to a simple equation, the difference between the benefits expected from the victory of party A (Ua) and the benefits expected from party B (Ub). Voters will favor party A if they see the difference between Ua and Ub as positive, party B if the difference is negative, and not bother to vote if they see no difference. Since Downs's pioneering work, elaborate econometric models have developed. Some correlate aggregate electoral results with global economic indicators such as purchasing power, inflation rates, and unemployment rates. Others use survey data to analyze individual pocket-book voting, based on the evaluation of candidates past (retrospective voting) or future performances (prospective voting). What they have in common is to shift attention from the people's demand to the supply side of politics, the specific goods offered on the electoral marketplace by competing parties and candidates.

2. *Alignment, Dealignment, or Realignment?*

These three models have been guiding electoral research agendas. In the wake of the sociological model attention was first given to the effects of social class and religion, even more so in Europe where party systems were historically built on the basis of social and religious conflicts and cleavages (see Lipset and Rokkan 1967). Rose's (1974) classic study of voting patterns in 15 democracies shows that class in Scandinavian countries and religion in the others are by far the most predictive factors explaining postwar electoral behavior, a working class position predisposing to a left-wing vote and regular church attendance to a right-wing vote. Then, in the 1960s, the Michigan paradigm became the leading model, inspiring nationwide research surveys in Europe, especially in Britain where Butler and Stokes (1969) explained long-term alignments of British voters by their party identification, itself rooted in their social background. What became known as the 'two class, two party' model does not fit as well in countries such as France with multiparty systems. There, Left/Right orientations are the substitute for party identification, equally rooted in religion and social class (Michelat and Simon 1977).

Neither model though can account for the increasing electoral change that appeared in the United States by the end of the 1960s, and in Europe at the end of the 1970's. On the basis of the SRC presidential surveys between 1956 and 1974, the authors of *The Changing American Voter* (Nie et al. 1976) revealed a decline of party ties. The proportion of 'independent' voters,

who define themselves as neither Republican nor Democrat, has risen from 23% in 1964 to 40% in 1974. Among those who still identify themselves with one side or the other the proportion who declare a 'strong' attachment or have a positive image of a party is declining, and the very link between party preference and voting is weakening. Voters seem more interested in politics, more aware of the issues at stake and more inclined to choose their candidates according to their political stands, whatever party they belong to.

Parallel studies in Europe detect similar trends of electoral volatility or instability of voting choice and partisan dealignment or gradual moving away from all parties (Crewe and Denver 1985), seen as a structural feature of the emerging postindustrial society. Social and geographical mobility is blurring the traditional class cleavages and loosening community ties. Progress in education and exposure to the media have increased the average levels of political sophistication. The rise of the permissive and individualistic set of values coined as postmaterialist by sociologist Inglehart (*The Silent Revolution*, 1977) encourage a new style of politics, more demanding and protest-prone, and promotes new post materialist issues—environmental, feminist—that cut through traditional party lines. All these trends converge to make citizens less dependent upon existing parties to make their choice. Voters are seen as 'autonomous,' 'strategic,' 'rational,' or 'reasoning' and the developing cognitive sciences are increasing studies on political reasoning mechanisms, the way voters process and organize information, the cues they rely on to make a decision, etc. (Sniderman et al. 1991). Short-term issue or candidate centered voting could well be taking the place of the former stable 'cleavage voting' with a partisan, religious, or social base (Franklin et al. 1992).

But a second research trend is taking an opposite stand, questioning the extent of this dealignment process by casting doubts on the way it is measured. The most heated controversies have developed around the supposed decline of class voting. Until now the simplest and most widely used indicator has been Alford's index, which measures the difference between the proportion of manual and nonmanual voters voting for the Left. For instance, if all the British working-class voters supported the Labour party and none among the middle classes in a given election, the value of the Alford index would be 100%, indicating perfect class voting. Measured by this index, the downward trend is undeniable. Class voting across post-World War II elections has decreased, for instance, by almost half in Britain, and more than two-thirds in Germany. But critics underline several weaknesses of such an index. Its dichotomous nature does not take into account the complexity of post-industrial occupational structure nor the complexity of party systems. It is statistically biased because it does not consider the changes in the marginal distributions of the two classes and of the two votes from one election to another. And it only measures dealignment, while various processes of electoral realignment along new cleavages such as private sector vs. public sector, self-employed vs. wage-earners, or educated vs. non- educated voters, seem to be taking place in postindustrial democracies, as shown by multiclass approaches using detailed occupational classifications and sophisticated measures such as odds-ratios or log-linear models. In Britain for instance, in the long run, Heath et al. (1991), contrary to previous findings, find no decline of class voting, just 'fluctuations without trend.' In France since the 1980s, the salaried middle classes have tended to support the Left, compensating for the drop of working-class support, attracted by the extreme right embodied by the National Front (Boy and Mayer 1997). In the United States, Hout, Brooks, and Manza dispute Clark, Lipset, and Rempel's claims about the declining political significance of social class, by showing the historical realignment in presidential post-War elections since 1968, as professionals and nonmanagerial white- collar workers moved from supporting the Republicans to voting for the Democrats (special issue of *International Sociology*, **8**(3), 1993, on class voting). Lastly, the new institutionalism approach attempts to reintegrate voters' choices in their political context, analyzing their response to changes in the electoral rules of the game, in the party structure, and in the strategies of political actors, especially concerning their uses of the media.

3. *The Puzzle of Electoral Participation*

An analysis of voting choice would not be complete without taking into account the alternative behavior: not voting. Whatever the election, a sizable proportion of eligible citizens are not registered or choose not to go to the polls. The proportion ranges between less than 5% in countries where voting is compulsory to 40% or more in Switzerland or in the United States. Whatever the country, the profile of these nonvoters, revealed by aggregate as well as survey data, is quite similar. Educated voters and members of the upper and middle classes participate more than the uneducated and the working classes, and middle-aged voters go more willingly to the polls than the very young or the very old. What has become known as the 'socioeconomic status (SES) model' (Verba and Nie 1972) postulates that well-integrated and upper-status individuals are more likely to participate because they have more resources to invest in politics—time, money, information, networks—and because they are more likely to develop the 'civic' attitudes (see Almond and Verba, *The Civic Culture: Political Attitudes and Democracy in Five Nations*, Princeton University Press, Princeton, 1963) that predispose to do so, such as interest in politics, a sense of duty, and a feeling of political efficacy. Political participation is

socially skewed. There are compensating factors, however, such as group consciousness or ethnic awareness. With the development of feminism, the difference between female and male turnout rates has disappeared, and 'black-political empowerment' in the United States since the Voting Right Acts has reduced the difference between whites and African–Americans, the latter participating even more than whites when differences in SES are taken into account (Verba and Nie 1972). On the whole, the increasing educational level of the electorate and the growth of the middle classes should gradually raise electoral participation in postindustrial societies. But paradoxically the opposite is happening. There has been a clear drop in turnout rates among American voters, from 63% of the eligible voters in the 1960 Kennedy–Nixon election to 50% in the 1988 Bush–Dukakis contest. A similar decline is developing in European democracies but the trend is recent, milder, and educational differences are far less marked (Topf 1995).

This 'participation puzzle' has revived interest in rational choice models. Because one vote has but an infinitesimal influence on the electoral outcome, the rational behavior, at first sight, would be not to bother going to the polls, and since Downs, a large body of literature has been devoted to what appeared as the 'voting paradox.' It can only be understood if one extends the notion of rationality and if one admits that voting, like most human actions, is not only guided by interest ('instrumental' rationality) but by principles, and values ('axiological' rationality) (Boudon 1997, Friedman 1996). At the beginning of the twenty-first century, along the same line, several studies try to understand the paradox of nonvoting. Some focus on the costs and benefits of voting turnout and particularly institutional factors such as registration formalities, compulsory voting, electoral rules, party competition, and all the political factors that facilitate or make meaningless the act of voting (Powell 1980). Others see nonvoting as a deliberate strategic behavior, expressing political discontent with the electoral choices available (Subileau and Toinet 1993). Studies of electoral turnout detecting nonvoters on official voting records, such as the series conducted in France by the National Institute for Statistics and Economics Studies (INSEE), are more reliable than survey data because voters are reluctant to confess what is considered as an uncivic behavior. Yet they confirm the tight imbrication of the social and political factors leading to nonvoting.

The rising number of emerging or transitional democracies and the multiplication of cross-national data bases such as the Eurobarometers, the Central and Eastern Eurobarometers, the Latino Barometer and the Russia Barometer, provide fresh opportunities to test in different political cultures the validity of the theories and concepts elaborated in Western democracies. They could also be an occasion to explore what most studies leave aside, the noninstrumental meanings the act of voting can take. Although the political outcome of elections is essential, they also have ritual, expressive, and social functions, detailed by Edelman in his classic essay about *The Symbolic Uses of Politics* (University of Illinois Press, Urbana, 1964).

See also: Citizen Participation; Citizenship: Sociological Aspects; Class Consciousness; Class: Social; Electoral Geography; Groups, Sociology of; Participation: Political; Race Identity; Rational Choice Theory in Sociology; Socialization, Sociology of; Values, Sociology of; Voting: Class; Voting: Issue; Voting: Tactical; Voting: Turnout; Women's Suffrage

Bibliography

Berelson B R, Lazarsfeld P F, Mac Phee W N 1954 *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago University Press, Chicago

Boudon R 1997 Le 'paradoxe du vote' et la théorie de la rationalité. *Revue Française de Sociologie* **38**(2): 217–27

Boy D, Mayer N 1997 *L'électeur A Ses Raisons*. Presses de Sciences Po, Paris

Butler D, Stokes D 1969 *Political Change in Britain: Forces Shaping Electoral Choice*. St. Martin's, New York

Campbell A, Converse P, Miller W, Stokes D 1960 *The American Voter*. Wiley, New York

Crewe I, Denver D T (eds.) 1985 *Electoral Change in Western Democracies*. St. Martin's, New York

Franklin M N, Mackie T T, Valen H et al. 1992 *Electoral Change: Responses to Evolving Social and Attitudinal Structures in Western Countries*. Cambridge University Press, Cambridge, MA

Friedman J (ed.) 1996 *The Rational Choice Controversy. Economic Models of Politics Reconsidered*. Yale University Press, New Haven, CT

Heath A, Curtice J, Jowell A et al. 1991 *Understanding Political Change: The British Voter 1964–1987*. Pergamon Press, Oxford, UK

Lautman J, Lécuyer B P (eds.) 1998 *Paul Lazarsfeld (1901–1976). La Sociologie de Vienne à New York*. L'Harmattan, Paris

Lazarsfeld P F, Berelson B R, Gaudet H 1944 *The People's Choice*. Duell, Sloan, and Pearce, New York

Lipset S M, Rokkan S (eds.) 1967 *Party Systems and Voter Alignments*. Free Press, New York

Michelat G, Simon M 1977 *Classe, Religion et Comportement Politique*. Presses de la FNSP/Editions Sociales, Paris

Norris P (ed.) 1998 *Elections and Voting Behavior: New Challenges, New Perspectives*. Aldershot, Brookfield, VT

Nie N H, Verba S, Petrocik J R 1976 *The Changing American Voter*. Harvard University Press, Cambridge, MA

Norpoth H, Lewis-Beck M S, Lafay J D (eds.) 1991 *Economics and Politics: The Calculus of Support*. University of Michigan Press, Ann Arbor, MI

Powell G, Bingham Jr 1980 Voting turnout in thirty democracies. In: Rose R. (ed.) *Electoral Participation*. Sage, Beverly Hills, CA

Robinson W S 1950 Ecological correlations and the behavior of individuals. *American Sociological Review* **15**: 351–7

Rose R 1974 *Electoral Behavior: A Comparative Handbook*. Free Press, New York
Sniderman P M, Brody R A, Tetlock P E 1991 *Reasoning and Choice. Explorations in Political Psychology*. Cambridge University Press, Cambridge, UK
Subileau F, Toinet M F 1993 *Les Chemins de L'abstention: Une Comparaison Franco-Américaine*. La Découverte, Paris
Topf R 1995 Electoral participation. In: Fuchs D, Klingemann H D (eds.) *Citizens and the State* (*Beliefs in Government*). Oxford University Press, Oxford, UK, Vol. 1
Verba S, Nie N H 1972 *Participation in America: Political Democracy and Social Equality*. Harper & Row, New York

N. Mayer

Voting: Tactical

1. Tactical Voting to Affect the Allocation of Seats

The best-known example of tactical voting to affect the allocation of seats occurs in single-member districts operating under plurality rule (the plurality rule requires that the candidate with the most votes wins the seat). Imagine a race in which a right-wing candidate, R, faces a centrist C and a leftist L. One of R's supporters, upon seeing a poll that places C and L in a close fight for first, with R a distant third, may decide that voting for the second most-preferred candidate, C, is more likely to do some good (electing C rather than L) than is voting for R (which would more likely waste the vote, as far as affecting the outcome is concerned). In single-member districts, then, tactical voting tends to diminish the number of votes cast for candidates placed third or lower in the polls, concentrating them on the top two candidates.

The tendency for voters in single-member districts to avoid wasting their votes on hopeless candidates was noted as long ago as 1869 by the English electoral reformer Henry Droop. But it was Maurice Duverger, writing in the 1950s, who most clearly articulated the systemic consequences of tactical voting. In Duverger's view, tactical voting arose prominently in single-member plurality elections but hardly at all in proportional representation (PR) systems. He, thus, articulated two main propositions: one, known as Duverger's Law, asserts that single-member plurality systems promote bipartism; the other, known as Duverger's Hypothesis, asserts that multimember PR systems promote multipartism. (A third proposition dealt with runoff systems.)

Theoretical work on tactical voting since Duverger has focused on five main topics: (a) tactical voting in electoral systems other than single-member plurality; (b) the logical foundations of Duverger's Law; (c) the transition from local to global bipartism in Duverger's Law; (d) tactical voting that deconcentrates, rather than concentrates, the vote; and (e) the role of elites in fomenting tactical voting.

1.1 Tactical Voting Beyond Single-member Plurality

Reacting to Duverger's apparent belief that tactical voting was unknown under PR systems, Leys (1959) and Sartori (1968) argued that tactical voting under PR was no different in kind from that found under plurality, differing only in the degree to which it came into play—and, hence, in the degree to which it tended to reduce the number of viable parties. Sartori's notion of a continuum of systems, from strong (in which tactical voting acts forcefully to depress the number of parties) to feeble (in which tactical voting is largely absent and thus puts little downward pressure on the number of parties), is now standard in the literature.

Subsequent work has made the Leys–Sartori viewpoint more precise, both theoretically and empirically. Steven Reed, for example, has extended Duverger's Law to Japan's former electoral system (used from 1925 to 1993, with a brief postwar gap), which gave voters a single vote but used multimember districts. To illustrate Reed's extension, consider a district with three seats. Candidates expected to place fifth or lower in such districts were sometimes seen as hopeless and were thus deserted by their supporters, essentially reducing the competition to a four-candidate fight for three seats. Reed discovered similar tendencies for elections in four-seat districts to reduce to five-candidate contests, and for elections in five-seat districts to reduce to six-candidate contests. More generally, under a wide range of electoral systems, the rule is that tactical voting tends to reduce competition in an M-seat district to, at most, $M+1$ candidates or lists (Cox 1997).

Although political scientists' main reason for interest in tactical voting is its tendency to concentrate votes locally, hence to restrict party systems nationally, further research into four questions has clarified some limits of this line of reasoning. These questions are considered in the next four subsections.

1.2 Logical Foundations of Duverger's Law

First, under what conditions does tactical voting in single-member plurality systems really lead to local bipartism? Recently, various scholars (Cox 1997, Palfrey 1989, Myerson and Weber 1993) have answered this question by articulating conditions under which avoidance of wasted votes in single-member districts leads, not just to some erosion of support for third- and lower-placed candidates, but to a nearly complete erosion of such support, leaving only two candidates with non-trivial vote shares. Two key conditions identified in this work are that voters be

concerned exclusively with the outcome of the current election and that they have 'rational expectations,' meaning that their beliefs about the candidates' likely order of finish in the poll are correct. As empirical circumstances approximate these idealized conditions less closely, one expects less actual erosion of the votes of trailing candidates, and thus less pressure toward local bipartism.

1.3 Local and Global Bipartism

Second, under what conditions does tactical voting in single-member plurality systems lead to global bipartism? The answer is that, although tactical voting does lead under certain circumstances to local bipartism, it does not directly promote bipartism at the national level. Recent work (e.g., Cox 1997) has argued that how local party systems link to form national party systems depends on national-level factors, such as the nature of executive power (presidentialism vs. parliamentarism), rather than on district-specific factors such as tactical voting.

Because local and global bipartism are produced by different factors, they may or may not coincide. Sometimes they do, as in the USA: two parties in each district, two parties nationally. But it is also possible to have local bipartism coupled with national multipartism. This bifurcation between local and global party systems occurs when different parties are viable in different parts of a country, as in India.

1.4 Tactical Voting that Deconcentrates the Vote

Third, does tactical voting always act to concentrate the vote at the local level, hence to restrict the party system? The answer here is 'no.' In electoral systems using multimember districts, tactical voters will transfer excess votes from strong candidates, in addition to transferring wasted votes from weak candidates (Tsebelis 1986). This sort of tactical voting deconcentrates the vote, rather than concentrating it, and does not reduce the number of viable parties in the system. Furthermore, in electoral systems using multiple votes, voters may vote for hopeless candidates in order to avoid giving votes to viable candidates who might beat the candidate they really want to win. This sort of tactical wastage of votes again deconcentrates, rather than concentrating, the vote.

1.5 Elite Fomentation of Tactical Voting

Fourth, is it really voters who play the lead role in tactical voting? Technically, voters that care only about who wins the current election care about their vote only when it breaks a tie (or creates one), because it is only then that their vote affects the outcome. But ties between any trailing candidate and one of the front-runners are much less likely than is a tie between the top two. Thus, purely instrumental voters vote for whichever of the two front-runners they most prefer, as their votes are then most likely to do some good.

But if citizens think this way about voting, why would they vote at all? The chances of a single vote actually affecting the outcome are tiny, which suggests that the purely instrumental voter might just as well stay home and save a trip to the polls.

Scholars who argue along these lines have suggested that the key actors in generating tactical voting are parties and candidates—who have more at stake in an election and can take actions that potentially swing many votes, hence have a reasonable chance of affecting the outcome. To date, however, there is no published elaboration of how such elite-led tactical voting would differ from the grass roots tactical voting envisioned in most of the literature.

2. Tactical Voting to Affect Government Formation

The second broad category of tactical voting is that by which voters seek to secure a better allocation of ministerial portfolios. Portfolio-maximizing tactical voting has three different subcategories.

One subcategory is strategic sequencing—or voting to affect which party gets the first opportunity to form a government. This sort of tactical voting may have been important in Israeli elections prior to the reform of the rules for electing the prime minister.

A second subcategory is strategic balancing—or voting so as to prevent any single party or coalition from securing control of all bodies in a system of separated powers. This sort of tactical voting is argued to play a role in German and US elections. In the US case, the argument is that some voters vote for one party's presidential candidate but the other party's congressional candidate, intentionally seeking to deny either party control of both policy-making branches of government (Alesina and Rosenthal 1995). In the German case, a similar argument has been made to explain the opposition's consistent success in midterm state elections, which determine control of the upper house at the federal level (Lohmann et al. 1997).

A third subcategory is strategic threshold insurance—or voting to keep a prospective coalition partner's vote above a legally defined threshold to qualify for PR seats. This sort of tactical voting has clearly been important in German elections and may become so in other countries that have parties hovering around the relevant threshold values.

3. Tactical Voting as Protest

The third broad category of tactical voting is protest voting—voting against one's most-preferred party, even though it stands a good chance of winning, in

order to chasten it and modify its behavior in future. Protest voting fits the general definition of tactical voting given at the outset of the article, if one includes the result of future elections in the 'better outcome' that the voter seeks. As an example, consider a voter who prefers that party A win the seat in a single-member district but is nonetheless unhappy with some aspects of A's performance. Such a voter may try to 'send a message' by either abstaining, casting an invalid vote, voting for a third party, or (if the voter is confident that A will nonetheless win) voting for A's main opponent. The point is not to defeat A, but to reduce A's margin of victory and hopefully scare it into better behavior in future.

See also: Majoritarianism and Majority Rule; Third Parties: United States; Voting: Class; Voting: Issue; Voting, Sociology of; Voting: Turnout

Bibliography

Alesina A, Rosenthal H 1995 *Partisan Politics, Divided Government, and the Economy*. Cambridge University Press, Cambridge, UK

Cox G W 1997 *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge University Press, Cambridge, UK

Droop H 1881 On methods of electing representatives. *Journal of the Statistical Society* **44**: 2

Duverger M 1954 *Political Parties*. Wiley, New York

Leys C 1959 Models, theories and the theory of political parties. *Political Studies* **7**: 127–46

Lohmann S, Brady D, Rivers D 1997 Party identification, retrospective voting, and moderating elections in a federal system—West Germany, 1961–1989. *Comparative Political Studies* **30**: 420–49

Myerson R, Weber R 1993 A theory of voting equilibria. *American Political Science Review* **87**: 102–14

Palfrey T 1989 A Mathematical Proof of Duverger's Law. In: Ordeshook P (ed.) *Models of Strategic Choice in Politics*. University of Michigan Press, Ann Arbor, MI

Reed S R 1990 Structure and behavior: extending Duverger's law to the Japanese case. *British Journal of Political Science* **20**: 335–56

Sartori G 1968 Political development and political engineering. In: Montgomery J D, Hirschman A O (eds.) *Public Policy*. Harvard University Press, Cambridge, MA

Tsebelis G 1986 A general model of tactical and inverse tactical voting. *British Journal of Political Science* **16**: 395–404

G. W. Cox

Voting: Turnout

Why do some people vote and others do not? Why do some countries have higher turnout rates than others? Would election outcomes change or governmental policies be different if turnout levels were higher? These are the primary questions that concern students of voter turnout.

1. Factors that Affect Turnout

With one caveat, a useful approach to explaining turnout rates of either individuals or countries entails considering the various costs and benefits of voting. The caveat is straightforward and reflects the observation that any single vote 'is lost in a sea of other votes' (Downs 1957, p. 246). Paul Meehl clearly sketches out the implication of 'this big brute fact,' namely that 'my chances of determining who becomes President [or influencing just about any election outcome] are of about the same order of magnitude as my chances of being killed driving to the polls—hardly a profitable venture … [Voting] is irrational as an instrumental action toward an egocentric end' (Meehl 1977, p. 11). To make any headway in explaining voter turnout, one must recognize that there are intrinsic or expressive benefits to be gained from voting that do not depend on the infinitesimally small probability of affecting an election outcome. Without this expanded conception, save for the small number of instances where there are penalties for nonvoting, there can be no coherent explanation for why so many people choose to cast ballots.

The significant factors that affect the costs and benefits of voting may be usefully characterized as either of the institutional or individual variety. There are three primary types of institutional variables. *Laws* that specify the registration process, the type of election districts, and the structure of the national legislature all influence turnout. *Political characteristics* like the degree of disproportionality in the translation of votes into seats and the number of political parties represent another set of institutional characteristics that affect turnout. The role of the last group of factors, which taps the connection between the *social structure* and the political process, is more uncertain.

With regard to individual characteristics, although any examination of the vast turnout literature will reveal a plethora of factors, four stand out. Although the size of the difference varies, in virtually every election context the better educated have higher turnout than those who are less educated; older people have higher turnout than the young; those who are more interested in politics have higher turnout than those whose interests lie elsewhere; and, in elections that pit parties against one another, strong partisans have higher turnout than people who weakly identify or don't identify with a party. The explanation for these patterns may be found in a careful consideration of the costs and benefits potential voters face.

2. The Costs of Voting

Focusing first on the costs that people must overcome in order to vote, previous research quite clearly demonstrates that the higher the costs, the lower the turnout. Two primary costs should be considered.

Although one can find exceptions, it is generally the case that in order to vote people must first be registered. This can entail dealing with bureaucratic requirements and be rather time consuming, especially in a country like the USA where individuals, as opposed to the government, must take the initiative to register, often well in advance of election day when the excitement that accompanies the approach of an election is usually absent. Turnout is higher in countries that impose lower registration costs on their citizens, and within a country that imposes higher costs, like the USA, turnout is lower among those who are less well equipped to bear those costs (i.e., the less educated).

All voters must bear the second significant cost of voting, deciding for whom or what to vote. Commonly referred to as information costs, people must gather or be given sufficient information to decide how they will vote. It is a cost that can be easily overlooked, because for 'people already well informed, keeping up with politics is scarcely dull, is almost effortless, and the notion that information costs something seems almost peculiar' (Converse 1975, p. 97). Yet for many, politics is barely a peripheral concern. For these less interested people, information costs are higher and their turnout is consequently lower.

Two other individual factors are helpful in mitigating information costs. The greater exposure to and experience with the political system that come with age fosters turnout by making voting decisions easier. In addition, partisans can rely on the relatively simple cues of party labels to make their choices, and consequently their turnout is also higher.

Information costs for potential voters are also reduced if they are undertaken by someone or something else. In this regard candidates and political parties are extremely useful. Voter mobilization efforts reduce information costs because they diminish the need for people to gather for themselves the necessary information in order to make their decisions. Consequently, factors that provide candidates and parties with incentives for broad-based mobilization efforts, like nationally competitive electoral districts, close elections, and lower levels of disproportionality in the translation of votes to seats, all have the effect of boosting turnout by bringing down information costs for potential voters.

3. *The Benefits of Voting*

Even in a political context where voting costs are minimized, they are never eliminated; 'voting is inherently costly' (Downs 1957, p. 265). It is therefore necessary to consider the benefits citizens derive from voting in order to explain why anyone casts a ballot at all. The 'big brute fact' that a single vote has virtually no effect on an election outcome leads to the observation that 'most of the action is, in fact, in the intrinsic values of voting *per se*' (Aldrich 1993, p. 266). This is likely to be the case even in the few countries that have mandatory voting laws, given that the penalties for nonvoting are minor and that enforcement is minimal. The higher turnout induced by these laws probably derives from the expressive benefit of fulfilling a civic obligation, rather than avoiding a sanction.

Though less strong in countries without mandatory voting, the notion that one has a duty to vote pervades most democracies. Schooling is one way this norm in inculcated, and provides another reason why the better educated have higher turnout, whether it derives from the positive feelings of having voted, or sidestepping the guilt that would accompany not performing one's civic duty. In addition, stronger partisans and those who are more interested in politics gain greater intrinsic satisfaction from voting because they are more engaged with the political system.

The association between institutional characteristics and the benefits of voting is less intuitive, but is empirically evident. Voting is more gratifying when there is a stronger connection between election results and government formation. The relationship is weaker in countries with more political parties, where post-election coalition building is far more frequent and complicated, and turnout is lower. Similarly, in countries with unicameral national legislatures, the consequences of election outcomes are more decisive for governing, and people are therefore more likely to perceive elections as meaningful and vote (Jackman 1987).

The last variable to consider is the 'linkage' between the social and political structures within a country: the degree to which social groups and political parties are aligned. In countries where substantial proportions of people with a given group (religious, racial, occupational, etc.) characteristical vote for a specific party, the benefits of voting ought to be enhanced because voting becomes an expression of group identification. Moreover, information costs are lower because people will have an easier time deciding how to cast their ballots. And where parties and groups are more closely aligned, parties have greater incentive to invest resources in mobilization efforts. At the empirical level, measuring the degree of linkage between parties and social groups has proved difficult, and reports of its turnout effects are inconsistent (Powell 1986, Jackman 1987), suggesting that the topic deserves further scrutiny.

4. *The Implications of Nonvoting*

Attention to social structure leads to the final question regarding the effects of nonvoting on election outcomes and government policy. In any country without universal electoral participation (i.e., all countries)

some groups will have higher turnout than others. Compared to their proportion of the population, some will be over-represented while others will be under-represented in the voting population. The potential political implications for the party system and election outcomes are sufficiently significant that in his presidential address to the American Political Science Association, *Unequal Participation: Democracy's Unresolved Dilemma*, Arend Lijphart (1997) advocated compulsory voting.

Two factors determine the implications of nonvoting. First, the difference between the preferences of voters and nonvoters must be considered. To the extent that what nonvoters want differs from the desires of voters, the representation of interests in the voting population will be a distortion of the interests of the larger population. If voters and nonvoters have similar preferences, then the degree of distortion will be less. The second important consideration concerns the relative size of the nonvoting population. If the number of nonvoters in an election is small compared to the number of voters, then even if nonvoters have a very different set of preferences, their small number will limit the potential for impact on the aggregate preferences of voters. However, where nonvoters are more numerous, their potential influence grows.

Several concrete examples illustrate the interplay between nonvoters' preferences, the size of the nonvoting population, and implications of nonvoting. In the USA, after the Voting Rights Act of 1965 and the widespread dismantling of legal and extralegal barriers to black suffrage in the southern states, large numbers of African-Americans entered the voting population. Their preferences were quite different from those of white voters, thereby spurring the political realignment in that region of the country. In contrast, when the USA (in 1919) and Switzerland (in 1971) extended the right to vote to women, similarly large-scale political change did not ensue. Although larger numbers of voters were added to the electorate, their preferences were not especially distinctive, and as a result the electoral consequences were minimal.

In the current political era, the potential implications of nonvoting are greatest in the USA because turnout is comparatively low and the connection between socioeconomic status and turnout is relatively strong. (In other countries, the relationship is weaker because aggregate turnout is higher.) The actual consequences, however, do not appear to be great. First, the relationship between socioeconomic status and partisan and policy preferences is modest. More importantly, when one considers the groups that comprise the nonvoting population, one finds that those groups with the most distinctive preferences (the poor and minorities) are outnumbered by others. The largest groups of nonvoters, the young and residentially mobile, are not politically distinctive, with the exception of the former's greater inclination toward political independence. Thus, 'nonvoters appear well represented by those who vote' (Highton and Wolfinger 2001, p. 10). The possibility that this conclusion applies to other countries is suggested by G. Bingham Powell's (1986) comparative analysis that finds that at the individual level within countries, age, not education, is the strongest demographic predictor of turnout.

To conclude, voter turnout is shaped by both individual and institutional factors. Explaining the connection between these factors and the decision to cast a ballot requires a consideration of the costs and benefits facing potential voters. Because the costs and benefits are not identical for all people, different groups vote at different rates. As a result, there is a possibility that the preferences of voters will not be representative of the entire population. To be sure, this sort of distortion is evident, but a number of factors suggest that in general its magnitude is not especially large.

See also: Citizen Participation; Electoral Geography; Electoral Systems; Participation: Political; Political Representation; Voting: Class; Voting: Compulsory; Voting: Issue; Voting, Sociology of; Voting: Tactical

Bibliography

Aldrich J H 1993 Rational choice and turnout. *American Journal of Political Science* **37**: 246–78

Converse P E 1975 Public opinion and voting behavior. In: Greenstein F I, Polsby N W (eds.) *Handbook of Political Science*. Addison-Wesley, Reading, MA

Crewe I 1981 Electoral participation. In: Butler D, Penniman H R, Ranney A (eds.) *Democracy at the Polls*. American Enterprise Institute for Public Policy Research, Washington, DC

Downs A 1957 *An Economic Theory of Democracy*. Harper, New York

Franklin M N 1996 Electoral participation. In: LeDuc L, Niemi R G, Norris P (eds.) *Comparing Democracies: Elections and Voting in Global Perspective*. Sage, Thousand Oaks, CA

Highton B 1997 Easy registration and voter turnout. *Journal of Politics* **59**: 565–75

Highton B, Wolfinger R E 2001 The political implications of higher turnout. *British Journal of Political Science* **31**

Jackman R W 1987 Political institutions and voter turnout in the industrial democracies. *American Political Science Review* **81**: 405–23

Lijphart A 1997 Unequal participation: Democracy's unresolved dilemma. Presidential address, American Political Science Association 1996. *American Political Science Review* **91**: 1–14

Meehl P E 1977 The selfish voter paradox and the thrown-away vote argument. *American Political Science Review* **71**: 11–30

Powell G B 1986 American voter turnout in comparative perspective. *American Political Science Review* **80**: 17–43

Rosenstone S J, Hansen J M 1993 *Mobilization, Participation, and Democracy in America*. Macmillan, New York

Wolfinger R E 1994 The rational citizen faces election day, or, What rational choice theorists don't tell you about American

elections. In: Jennings M K, Mann T E (eds.) *Elections at Home and Abroad: Essays in Honor of Warren E. Miller*. University of Michigan Press, Ann Arbor, MI
Wolfinger R E, Rosenstone S J 1980 *Who Votes?* Yale University Press, New Haven, CT

B. Highton

Vulnerability and Perceived Susceptibility, Psychology of

Virtually all major theories of health behavior are based on the assumption that people estimate their vulnerability or susceptibility to a disease and evaluate the costs and benefits of precautionary behavior before taking action. Research on the accuracy of perceived risk or susceptibility is summarized, followed by a discussion of the relation between perceived susceptibility and precautionary behavior. Methodological considerations about how to measure perceived susceptibility and how to investigate its role as a behavioral determinant will be discussed. Implications for health education practice will be briefly discussed.

1. *The Role of Vulnerability in Models of Health Behavior*

Vulnerability or perceived risk is generally conceived as consisting of two components: the *likelihood* and the *severity* of negative outcomes. Most of the prevailing models of health behavior assume that decisions about risk behaviors are based on conscious decisions in which perceived risks and benefits determine behavior (see *Health Behavior: Psychosocial Theories*).

The *theory of reasoned action* (Fishbein and Ajzen 1975) and the *theory of planned behavior* (Ajzen 1991) are frequently used models of health behavior in which the probability and severity of consequences for people's health of specific behavioral practices are assumed to be a prime determinant of attitudes toward precautionary behavior. These behavioral beliefs can also include risks for health and/or well-being.

The *health belief model* (HBM) is another frequently used framework to explain preventive health behavior (Janz and Becker 1984). The model aims to describe decisions about the costs and benefits of specific actions and distinguishes several factors that are assumed to determine the adoption of protective action. Among these are the perceived vulnerability to developing a specific health problem and the perceived severity of that problem. Precautionary behavior is assumed to be most likely when perceived severity and vulnerability are high, while the perceived benefits of precautionary behavior are substantial, and the costs of behavior change are low.

Weinstein's (1988) *precaution adoption process* is also based on behavioral decision theory, and assumes a series of steps or stages preceding the adoption of preventive action to reduce the threat of negative consequences for people's health. First, people have to realize that a specific risk exists. Second, they have to realize that the risk is significant and can affect people. Third, they have to realize that they are vulnerable to the risk. After these necessary requirements, behavioral change will be a function of the perceived severity of the consequences for health and the efficacy and costs of preventive behavior.

In *protection motivation theory* (Rippetoe and Rogers 1987) perceived vulnerability is also a major factor in the motivation to avoid risk. The theory combines perceived vulnerability to the negative event with appraisals of its severity and the efficacy of the recommended (preventive) action and self-efficacy. The latter factor refers to the ability to initiate and/or sustain a specific precautionary behavior.

Thus, most models of preventive health behavior incorporate the recognition of one's own risk-status or vulnerability as an important condition for adopting behaviors that reduce these risks. The construct of vulnerability basically has the same meaning in all of the models, and it has been assessed with questions that seem more or less interchangeable. One implicit assumption of all these models is that people are able to adequately assess the risks associated with their behavior (see *Health: Self-regulation*).

2. *Accuracy of Perceived Risk/Vulnerability*

There is a substantial amount of research showing that the estimation of risk tends to be a complex process that depends on factors such as the context in which the risk is presented, the way the risk is being described, and also on personal and cultural characteristics. Risks that are more cognitively 'available' due to personal experience or media coverage tend to be overestimated. Thus, estimates of the likelihood of risks such as contracting AIDS or being involved in an air crash tend to be too high, while estimates tend to be too low for more common and/or less sensational risks such as heart disease or being involved in a car accident. The magnitude of perceived risk is also affected by how the perceived probability of a negative outcome is measured. For instance verbal methods such as a scale ranging from 1 (very unlikely) to 7 (very likely) lead to different answers than numerical methods such as a percentage scale ranging from 0 to 100 percent. Verbal measures are better predictors of individual preferences than the more difficult numerical scores. Generally, people have a reasonable idea of the relative risks of various activities and behaviors, although their estimates of the magnitude of risks tend to be biased.

Most research on biases in perceived risk focuses on cognitive factors such as availability and tends to ignore motivational factors that may influence the perception of risk. Although people seem quite aware of the relative risk of specific activities or behaviors, things can change when this knowledge is applied to their own behavior. For instance, many smokers accept the association between smoking cigarettes and disease, but do not believe themselves to be personally at risk. Motivational explanations of this perceived invulnerability tend to focus on the need to reduce feelings of fear and anxiety. Support for the role of these mechanisms is provided by research showing more biased risk estimates in situations of increased threat (see *Health Risk Appraisal and Optimistic Bias*.

In sum: Research shows that estimations of perceived risk tend to be biased, but these biases primarily concern the magnitude of risk; people seem to have a reasonable idea of the relative risk of various activities. This picture can change when dealing with risks associated with people's own behavioral practices. Under conditions of threat motivational factors could come into play and result in an underestimation of risk.

3. *Perceived Risk or Vulnerability and Behavior*

Weinstein (1984) argued that people seem able and willing to incorporate knowledge about their family history, personality, and physical or physiological attributes into their perceptions of risk or vulnerability, but seem much poorer at recognizing the relationships between their own actions and the risks they run. Considerable research effort has been directed to understand the relation between perceived risk or vulnerability and protective behavior. Harrison et al. (1992) conducted a meta-analysis of studies that examined the relation between the four major components of the HBM and various health behaviors. They reported a weighted mean effect for the relation between perceived risk or vulnerability and behavior (risk-reducing practices) in the small to medium range. Sheeran and Abraham (1996), in their review of more than 30 years of research on HBM, point at a number of limitations, and argue that more clarity is needed on the precise role of perceived vulnerability. For instance, it could well be that specific levels of perceived severity and vulnerability function as a threshold variable before perceived vulnerability has an impact on behavior. This could explain the sometimes weak relationship between perceived vulnerability and behavior.

Gerrard et al. (1996) investigated the relationship between perceived vulnerability to HIV and precautionary sexual behavior, and conclude that the often dramatic changes in sexual risk behavior of high-risk groups were only marginally related to beliefs about being at risk. Cross-sectional research suggests that people who engage in more risky behavior tend to be more aware of the increased likelihood of contracting HIV than people who tend to behave less riskily, but only modestly so (see *Sexual Risk Behaviors*).

4. *Some Methodological Considerations*

The assumption shared by many models of health behavior that preventive behavior is the result of the wish to reduce one's risks has an explicit temporal order; that is, initial perceptions of risk are expected to influence subsequent behavior. This assumption not only implies that perceived risk or vulnerability influences future behavior, but also that it is shaped by prior behavior.

A lot of research relies on correlational data. Moreover, these correlations from cross-sectional studies are often interpreted as bearing upon the perceived risk–precautionary behavior hypothesis, while they actually provided information about the accuracy of risk perceptions. This is for instance the case if people engage in risky practices, claim their risk to be low, and intend to continue their risky behavior. In such a case risk perception is clearly inaccurate (see Weinstein and Nicolich 1993).

Another issue concerns the distinction between *unconditional* and *conditional* risk estimates. Unconditional risk estimates refer to the subjective likelihood that a negative consequence will occur, based on whatever factors individuals take into account (e.g., perceptions of control, the perceived efficacy of preventive actions). Conditional risk refers to the probability of adverse consequence for health if no preventive action is taken, or their probability if a specific (preventive) action is taken. A conditional risk estimate thus requires people to indicate their risk or vulnerability given their present behavioral practices, or changes in these practices.

Although most models of health behavior refer to conditional risk estimates, most research tends to rely on unconditional risk estimates (e.g., 'How likely is it that you will get…', followed by the health risk(s) under study). The major drawback of unconditional risk estimates is that it is unclear what set of factors people take into account when answering this general question. A conditional risk estimate is more likely to be related to behavioral intentions or expectations in a consistent and interpretable manner (van der Pligt 1998).

5. *Conclusions and Future Directions of Theory and Research*

A number of biases affect the perception of risk and vulnerability. Cognitive availability has an effect on the perceived magnitude of risk, but generally people

have a fairly accurate view of their relative risk. Motivational factors, such as the tendency to deny risks and/or avoid threatening information about risks, can also affect perceived vulnerability. Unfortunately, research findings concerning the impact of perceived risk on behavior are mixed. Generally, perceived risk is a modest predictor of preventive health behavior. Some argue that the hypothesis that perceived risk is a determinant of protective behavior should be rejected or at least thoroughly re-examined. This skepticism may be premature. The absence of a consistent empirical relationship between perceived risk and precautionary behavior seems partly related to shortcomings in the design of many studies investigating this relationship, and on the way perceived risk is measured. It could well be that at a certain point the status of 'being at risk' rather than the amount of risk becomes central in people's minds. Generally, the perception of personal vulnerability to health risks seems a necessary requirement for people to consider behavioral change, but it is not sufficient actually to induce people to change risky practices.

Programs aiming to increase the prevalence of protective behavior should thus not only provide risk information. It seems essential to develop interventions that also focus on how to change behavior. Future research should help to design interventions that increase risk awareness (without being too threatening), and also help to induce and maintain precautionary behavior (see *Health Education and Health Promotion*).

See also: Health Behavior: Psychosocial Theories; Health Education and Health Promotion; Health Risk Appraisal and Optimistic Bias; Health: Self-regulation; Sexual Risk Behaviors

Bibliography

Ajzen I 1991 The theory of planned behavior. *Organizational Behavior and Human Decision Processes* **50**: 179–211

Fishbein M, Ajzen I 1975 *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA

Gerrard M, Gibbons F X, Bushman B J 1996 Relation between perceived vulnerability to HIV and precautionary sexual behavior. *Psychological Bulletin* **119**: 390–409

Harrison J A, Mullen P D, Green L W 1992 A meta-analysis of studies of the health belief model with adults. *Health Education Research* **7**: 107–16

Janz N, Becker M 1984 The health belief model: A decade later. *Health Education Quarterly* **11**: 1–47

Rippetoe P A, Rogers R W 1987 Effects of components of protection-motivation theory on adaptive and maladaptive coping with a health threat. *Journal of Personality and Social Psychology* **52**: 596–604

Sheeran P, Abraham C 1996 The health belief model. In: Conner M C, Norman P (eds.) *Predicting Health Behavior*. Open University Press, Buckingham, UK pp. 23–61

Van der Pligt J 1998 Perceived risk and vulnerability as predictors of precautionary behaviour. *British Journal of Health Psychology* **3**: 1–14

Weinstein N D 1984 Why it won't happen to me: Perceptions of risk factors and susceptibility. *Health Psychology* **3**: 431–57

Weinstein N D 1988 The precaution adoption process. *Health Psychology* **7**: 355–86

Weinstein N D, Nicolich M 1993 Correct and incorrect interpretations of correlations between risk perceptions and risk behaviors. *Health Psychology* **12**: 235–45

J. van der Pligt

Vygotskij, Lev Semenovic (1896–1934)

1. Biography

Lev Semenovic Vygotskij was born on November 5, 1896 in Orsha as the second child to a secular Jewish family and spent his youth in Gomel in Byelorussia. He received private tuition and only entered school at the age of 15 to follow the last two years of a private Jewish gymnasium. Having finished school, he moved to Moscow to study law at the Moscow Imperial University and humanities at the Shanyavsky University. He finished his studies at Shanyavsky University with a Masters thesis on Shakespeare's *Hamlet* and returned to Gomel to engage in a wide variety of activities (Vygodskaya and Lifanova 1996).

His interests at the time were divided between two areas: on the one hand, since his youth he had been fascinated by literature and theater. During his university years he had pursued these interests, taking courses in literature and linguistics, attending the plays in Stanislavsky's Art Theater and Tairov's Chamber Theater, and writing theater and book reviews for literary journals. On the other hand, taking Pavel P. Blonsky's courses at Shanyavsky University on education and psychology had quickened his interest in problems of a psychological nature (Feigenberg 1996). His activities in Gomel reflected these two broad areas of interest. During these years (from 1917 to 1924), Vygotskij worked as a literary critic for local newspapers, taught different subjects (e.g., literature, philosophy, psychology) at various schools of different levels, co-founded a short-lived publishing house and an equally short-lived literary journal, and worked for the Gomel authorities as the official responsible for the cultural life of the town. In this latter capacity he contracted theater groups, held lectures on topics of general interest, and organized poetry evenings (Van der Veer and Valsiner 1991).

While in Gomel, Vygotskij became interested in the way the formal properties of a literary text evoke aesthetic experiences in the reader. This led him to deepen his interests in psychology, as did his experiences as a teacher of children and adults of highly

different abilities and penchants. He installed a psychological laboratory at one of the schools where he was teaching and conducted his first empirical psychological investigations. His account of these investigations at a congress led to an invitation by Luria (see *Luria, Aleksander Romanovich* (1902–77)) to come and work as a psychologist at Moscow University. His Ph.D. Thesis, *The Psychology of Art* (Vygotskij 1925/1971) gives a perfect overview of his early views.

Having arrived in Moscow in 1924, Vygotskij again displayed a wide variety of activities. During what was to be the last decade of his life he worked as a clinical psychologist, carried out empirical studies with children, held teaching positions at various universities, co-founded several scientific journals, headed research institutes, worked for publishing houses, and published a couple of hundred books and articles on widely diverging topics. His interests transcended narrow disciplinary boundaries. He began, for example, studying the semiotics of cinematic language with the film director Sergej Eisenstein, he developed a test to study the thinking of schizophrenics, he lectured on the problems of blind and deaf children, he initiated research into memory development, and he developed an interest in what is now called neuropsychology. In 1931, while already being a full professor of psychology, he began studying medicine. However, on June 11, 1934 Vygotskij died in Moscow of tuberculosis, a disease that had haunted him for the better part of his adult life.

The writings of Vygotskij were virtually unknown outside Russia until the publication of *Thought and Language* (Vygotskij 1962) in English stirred the international interest. Now a substantial part of his writings are available in translation. As a result, the discussion of the merits and shortcomings of Vygotskij's work has become a matter of interest to the international scientific community.

2. *Main Ideas*

Given his interest in literature and language, it comes as no surprise that Vygotskij (1962, 1994) attached great importance to language in his psychological thinking. Like other thinkers, he held that the use of language (speech, in his terminology), together with the manufacture and use of tools, is a distinguishing characteristic of humans. In his view, all truly human, higher mental processes involve the use of speech.

2.1 *From Dialogue to Self-regulation: The Sociogenetic Law*

The original function of language is communication, Vygotskij (1997a) argued. Children and adults communicate their intentions and steer or guide the behavior of the social other. This communicative process is at first very asymmetric in that the adult does most of the guiding and attributes meaning—at first often counterfactually—to the infants' utterances. Very soon, however, children become able to use words to accomplish things and, most importantly, to steer their own behavior. Self-regulation thus originates in dialogue.

A simple example is that of a child learning to cross a street. In a first stage, caregivers will tell the child to watch out carefully and children will simply follow these instructions. In the next stage, the child will say aloud to watch out and thus instruct him/herself. In the final stage, the child will merely think these instructions. Thus, a process that was originally shared between two persons, an interpsychological process, has become an individual, intrapsychological process. The child is applying to itself what was first applied by others and thereby mastered its own behavior through the use of conventional signs. With Baldwin, Janet (see *Janet, Pierre* (1859–1947)), Royce, and others, Vygotskij (1997a) claimed that all higher mental process are originally shared between persons and he coined the developmental transition from inter- to intramental processes the sociogenetic law.

This example is linked with another phenomenon that Vygotskij discussed in his writings, namely that of the intersection of thought and speech. In his view, thought and speech have different genetic roots and only merge to form verbal thought or intellectual speech at a later stage in phylogenetic and ontogenetic development. Thus, in human children both preintellectual speech (i.e., babbling) and preverbal thinking (e.g., infants expect to find an object where it was hidden previously) can be observed, while in the phylogenetic domain one sees animals solving simple problems (e.g., using an artifact to reach a goal) and displaying primitive communicative means (e.g., expressive movements). It is only in humans, however, that speech and problem solving merge to form the powerful capacity of verbal thought (see *Khler, Wolfgang (1887–1967)*). Problem solving now no longer proceeds in the graphic or visual field, Vygotskij (1997b) argued, but in a semantic field.

2.2 *Egocentric and Inner Speech*

Vygotskij's claim that speech develops in an intrapersonal situation and subsequently is used to steer the self, received additional support through his discussion of the phenomenon of egocentric speech. It was Piaget (see *Piaget, Jean* (1896–1980)) who first described the phenomenon of children who while carrying out some activity speak for themselves in a way that is often not intelligible to others. Piaget hypothesized that such speech does not take account of the viewpoint of the social other and reflects younger children's general inability to transcend their own point of view. In his view, children are at first

unable to see the world from another perspective than their own and only gradually realize, under the pressure of their peers, that theirs is a specific viewpoint.

Vygotskij criticized Piaget's interpretation of the phenomenon of egocentric speech and conducted a series of experiments to argue that children's egocentric speech is social in the sense that it addresses social others (Van der Veer and Valsiner 1991, pp. 363–9, Vygotskij 1987). He noted that egocentric speech is absent or greatly reduced when the child is alone (which suggests that it is meant for others) and that its incidence rises when the child is confronted with unexpected problems (which suggests that it has some function in the solution of problems). He also observed that egocentric speech becomes less intelligible as the child grows older. This set of results led him to posit that egocentric speech: (a) originates in normal, communicative speech and branches off at a later stage; (b) has as its function to steer the child's behavior when the need arises; and (c) becomes less and less intelligible for the outsider until it becomes proper inner speech. Egocentric speech, then, is an intermediary stage between normal, communicative speech and inner speech. Like communicative speech, it is audible, and like inner speech, it serves to guide the child's thinking.

Vygotskij went on to speculate about the properties of inner speech. He argued that its syntax must be special: it is speech that is abbreviated and tends toward predicativity, that is, omitting the subject of the utterance. He also suggested that inner speech has special semantic properties: personal sense dominates over dictionary meaning. He thus believed to have demonstrated that, in the process of becoming private, speech undergoes a series of transformations.

2.3 Word Meaning and Conceptual Development

Above it was said that children use conventional signs to regulate their own behavior. Towards the end of his life Vygotskij began emphasizing that the meanings of these signs or words change over time. Children use the same words as adults do to designate specific objects but the meaning of these words may be quite different. Thus, in the child the word 'farmer' may evoke romantic images of a person taking care of animals and growing crops, while adults may think of an entrepreneur who tries to maximize profit. Using various strategies, Vygotskij tried to prove experimentally that a mature understanding of many concepts is only reached in adolescence. In this connection, he introduced a distinction between mature, academic, or scientific concepts, on the one hand, and immature, or everyday concepts, on the other hand. Supposedly, the academic concepts express the true nature of things and are linked to other similar concepts. Together they form a coherent network covering a field of knowledge. Academic concepts are taught at school. Everyday concepts, on the other hand, often focus on irrelevant features of some subject and together form a disconnected whole. Everyday concepts are more or less independently acquired by children in informal settings.

Vygotskij's basic idea was that academic concepts and everyday concepts mutually enrich each other. The everyday concept of a farmer becomes enriched because the child learns the nonapparent fact that farmers form part of an economic market in which they try to realize certain goals. The academic concept of a farmer is enriched because the abstract notion of entrepreneurs dealing with cattle and crops is filled with concrete facts of daily farmer life and thereby gets body and flesh. Thus, academic concepts presuppose everyday concepts, build upon them, but once acquired they alter the everyday concepts in fundamental ways. The result of the interplay between everyday and academic concept is an elaborate and rich notion of the concept of farmer. Vygotskij (1935) speculated that the introduction of systems of academic concepts in school would restructure the child's whole style of thinking, that is, he assumed that the introduction of such concepts in specific subject areas would generalize to other areas. Education leads development, as he used to put it. Instruction in the school setting propels child development along lines that are each time specific for a certain culture or society.

2.4 Predicting Mental Growth: The Zone of Proximal Development

In several lectures delivered in 1933, Vygotskij transposed the sociogenetic law to the domain of cognitive prognosis. He claimed that if one were to measure what the child is able to do in cooperation with someone else (interpsychologically), one would be able to predict what the child can do independently (intrapsychologically) later on. At any given moment, both the child's dependent (e.g., with hints from an adult) and its independent performance on an IQ-test can be measured. The difference between the two IQ-scores yields an indication of the child's cognitive potential. Some children will be able to profit more from adult hints and will improve their independent score more substantially than others. These children show the richest developmental potential, Vygotskij (1935) argued. To measure what children can do independently is to measure what they can do already. To measure what children can do with the help of an adult is to measure what they will be able to do in the near future, in what Vygotskij coined their zone of proximal development. In this way, he criticized the contemporary use of IQ-tests and suggested a procedure for the assessment of cognitive potential.

3. *Relevance for Contemporary Science*

Vygotskij's ideas inspired an enormous and still growing amount of research, both in and outside the former Soviet Union.

Wertsch (1985) tried to corroborate Vygotskij's view that self-regulation develops in dialogue by investigating mother-child joint problem solving. He and others have also advanced proposals to fuse Vygotskij's ideas on dialogue and its relevance for self-regulation with those of Bakhtin.

Vygotskij's view of the nature and function of egocentric and inner speech has become a hotly debated topic and has been developed further by others (e.g., A. A. Leontiev, A. R. Luria, A. N. Sokolov). The Eastern European research is well represented in Sebeok (1968), Slobin (1966), Prucha (1983), and Wertsch (1978). Zivin (1979) gives a representative account of both the research tradition initiated by Vygotskij and Luria and its Western counterparts.

Vygotskij's concept of word meaning and his view of the dialectic interplay between academic and everyday concepts in education have also led to a whole research tradition. An excellent and critical account can be found in Wertsch (1985), while Van der Veer and Valsiner (1991) have traced how Vygotskij's research into the development of concepts was used by Western clinical psychologists to diagnose various groups of mental patients (e.g., schizophrenics and patients suffering from aphasia).

The concept of the zone of proximal development has inspired an immense number of publications. These can be divided roughly into two groups. The first group consists of publications that deal with the influence of education on development. It is analyzed, for example, how classroom discussions can promote a zone of proximal development in students, or what kind of teaching is needed to promote children's use of language for self-regulation. An overview of this kind of work can be found in Moll (1990). The second group consists of publications dealing with the assessment of learning potential. In this approach the zone of proximal development is used as an individual difference metric designed to provide information about individual students, that is, Vygotskij's original idea to measure aided and unaided performance on some IQ-test is operationalized and it is investigated whether such repeated measurements yield important prognostic information (cf. Campione et al. 1984).

All in all, it can be said that although many of Vygotskij's ideas are debatable (see especially Van der Veer and Valsiner 1991) and will have to be discarded or extended, they nevertheless have inspired a great and still growing number of researchers in the humanities. In itself this is remarkable for a man who published very little in languages other than Russian, who died more than 60 years ago and whose international fame began only 30 years after his death. The fact that Vygotskij's legacy still inspires contemporary research marks him as a true classic thinker of the humanities.

See also: Constructivism/Constructionism: Methodology; Constructivism in Cognitive Psychology; Cooperative Learning in Schools; Cultural Psychology; Culture in Development; Language and Literature; Luria, Aleksander Romanovich (1902–77); Memory: Collaborative; Pavlov, Ivan Petrovich (1849–1936); Piaget, Jean (1896–1980); Piaget's Theory of Child Development; Piaget's Theory of Human Development and Education; Psychologism, History of; Self-monitoring, Psychology of; Self-regulation in Adulthood; Self-regulation in Childhood; Vygotskij's Theory of Human Development and New Approaches to Education

Bibliography

Campione J C, Brown A L, Ferrara R A, Bryant N R 1984 The zone of proximal development: Implications for individual differences and learning. In: Rogoff B, Wertsch J V (eds.) *Children's Learning in the Zone of Proximal Development*. Jossey-Bass, San Francisco, pp. 77–91

Feigenberg I M (ed.) 1996 *L. S. Vygotskij: Nachalo puti*. Jerusalem Publishing Centre, Jerusalem

Moll L C (ed.) 1990 *Vygotsky and Education: Instructional Implications and Applications of Sociohistorical Psychology*. Cambridge University Press, Cambridge, MA

Prucha J 1983 Pragmalinguistics: East European Approaches. Benjamins, Amsterdam

Sebeok T A (ed.) 1968 *Current Trends in Linguistics. Vol. 1. Soviet and East European Linguistics*. Mouton, The Hague, The Netherlands

Slobin D I 1966 Soviet psycholinguistics. In: O'Connor N (ed.) *Present-day Russian Psychology; A Symposium by Seven Authors*. Pergamon Press, Oxford, UK

Van der Veer R, Valsiner J 1991 *Understanding Vygotsky: a Quest for Synthesis*. Blackwell, Oxford, UK

Vygodskaya G L, Lifanova T M 1996 *Lev Semenovich Vygotskij: Zhizn', Deiatel'Nost', Shtrikbi K Brtretu*. Smysl, Moscow

Vygotskij L S 1925/1971 *The Psychology of Art*. [trans. by Scripta Technic]. MIT, Cambridge, MA

Vygotskij L S 1935 *Umstvennoe razvitie detey v protsesse obucheniya*. Uchpedgiz, Moscow and Leningrad, Russia

Vygotskij L S 1962 *Thought and Language*. MIT, Cambridge, MA

Vygotskij L S 1987 The problem of speech and thinking in Piaget's theory. In: Rieber R W, Carton A S (eds.) *The Collected Works of L. S. Vygotsky. Vol. 1. Problems of General Psychology*. Plenum Press, New York, pp. 53–91

Vygotskij L S 1994 Tool and symbol in child development. In: Van der Veer R, Valsiner J (eds.) *The Vygotsky Reader*. Blackwell, Oxford, UK, pp. 99–174

Vygotskij L S 1997a The history of the development of higher mental functions. In: Rieber R (ed.) *The Collected Works of L. S. Vygotsky. Vol. 4. The History of the Development of Higher Mental Functions*. Plenum Press, New York, pp. 1–251

Vygotskij L S 1997b Preface to Koffka. In: Rieber R W, Wollock J (eds.) *The Collected Works of L. S. Vygotsky. Vol. 3.*

Problems of the Theory and History of Psychology. Plenum Press, New York, pp. 195–232
Wertsch J V 1978 *Recent Trends in Soviet Psycholinguistics*. Sharpe, New York
Wertsch J V 1985 *Vygotsky and the Social Formation of Mind*. Harvard University Press, Cambridge, MA
Zivin G (ed.) 1979 *The Development of Self-regulation through Private Speech*. Wiley, New York

R. van der Veer

Vygotskij's Theory of Human Development and New Approaches to Education

Vygotskij's legacy is in his emphasis on social, historical, and cultural aspects of development and learning, semiotic mediation, activity, and social interaction. After about 50 'dark years' of silence, from the early 1930s (immediately after his death) to the end of the 1970s, caused in part by the repressive political circumstances in the USSR, social sciences have been experiencing Vygotskij's renaissance. Since the 1970s, Vygotskij has become a 'root theoretician' for a family of approaches in social sciences including sociolinguistics, anthropology, psychology, education, and literacy studies, among others. In fact, it is probably not an overgeneralization to say that developmental psychology and education have currently been highly influenced by Vygotskij's work. This article will describe the main tenets of Vygotskij's theory of human development, current critique of these tenets, and new approaches to education stemming from a dialog with Vygotskij's work. The article is organized around three major interrelated issues: what develops, what are the processes of development, and what are the directions of development.

1. *What Develops?*

Vygotskij argued that human development is based on the intertwining of two lines: natural, based on processes of biological maturation, and cultural, based on mastery and use of cultural means. Psychologically, these two lines are represented by lower and higher mental functions, respectively, such as memory, attention, and intelligence. For example, mechanical memorization of unrelated bits of information mainly involves lower functions of memory. Meanwhile, memorization that is based on mental structuring of information in some hierarchical classes (e.g., ax, saw, and hammer are tools) or a story connecting items to remember involves higher mental functions of memory. Vygotskij insisted that higher mental functions are formed by a reorganization of lower mental functions mediated by a cultural sign (e.g., mnemonics). Higher mental functions allow people to control their behavior: 'in the elementary form [of memory], something is remembered; in the higher form, humans remember something … the basic characteristic of human behavior in general is that humans personally influence their relations with the environment and through that environment personally change their behavior, subjugating it to their control' (Vygotskij 1978, p. 51). Thus, according to Vygotskij, what develops is higher mental functions via a mastery of cultural signs (i.e., semiotic mediation), leading to a person's development of self-regulation and self-control.

Probably the most elaborate application of Vygotskij's stress on cultural mediation as the object of development has been in the 'sign' pedagogy developed by Russian educational psychologist Davydov (1986). Davydov and his colleagues rewrote an entire elementary school curricula (e.g., math, sciences, language) representing it as a series of dialectical contradictions of mental actions mediated by cultural signs (Davydov's unit of analysis). They also changed the curriculum sequence putting the abstract concepts—higher-level cultural mediators—ahead of concrete application of them (e.g., putting study of fraction and algebra ahead of integer arithmetic). For example, according to Davydov (1986), the fraction represents the contradiction between the permanence of the object's length and the variability of its measurement, depending on the unit mediating the measurement. The mathematical formula $A/c = N$ is a model of the fraction that allows first grade students to make comparisons involving any measurement unit.

Yet a growing literature has implicitly or explicitly criticized Vygotskij for making mental functions the object of development (Rogoff 1990). These scholars have argued that culturally mediated functions are not self-contained, but instead embedded in sociocultural activities. Lave (1988), introducing the notion of 'situated cognition' demonstrates, for example, that people's computation in school-like math tasks and grocery shopping is not psychologically equal even when the problems are mathematically isomorphic. The problem solving is mediated and structured differently in these two conditions because the problems are embedded in different sociocultural activities with different goals and different social relations. In response to these and other related findings, alternative objects of development have been proposed.

Rogoff (1990) proposes a sociocultural approach in which a person's participation in sociocultural activities and communities is the object of development. Development is seen as transformation of participation in a sociocultural activity—transformation that is socially valuable. For example, when a child whose parents had previously read books to her starts asking her parents to read a book, this change indicates a

transformation of the child's participation in the reading practice. Some have argued that mediation by cultural tools and scripts pointed by Vygotskij is only a part of the developmental story and should be considered together with aspects such as social relations, division of labor, community rules and norms, and relations with other practices and communities (Engeström et al. 1999). These features of social life were often ignored in cultural research and research on child development conducted by Vygotskij and his colleagues (Vygotskij 1978).

2. *What are the Processes of Development?*

According to Vygotskij's general genetic law of cultural development, any function in the child's cultural development appears on three planes: natural, social, and psychological. Vygotskij emphasized a dynamic and emergent character of development: Each plane emerges on the periphery ('in an embryonic state') of the previous plane. First, it appears on the natural plane, when children find themselves in the environment when the function is used by other people. For example, in human society, people use gestures in their communication with each other and they often are inclined to interpret some of a child's actions as gestures. When a very young child tries unsuccessfully to grasp a remote object, the caregiver may interpret it as a command gesture to bring the object to the child.

The second plane of development is social: the child is an active participant in the function socially distributed among people. Continuing Vygotskij's example of the development of the index gesture, at some point young children notice that they can control adults by stretching their arm toward a desired object—the adults often bring the remote object to the child. Thus, on the social plane, children actively participate in the social function of controlling other people rather than finding themselves in the function fully organized by other people as it was on the natural plane. Finally, on the third plane, the function transforms from external social into internal psychological. The child's command of others transforms into the index gesture controlling the child's own attention. Now the stretched arm controls not others but the child him or herself, whose own attention is mediated by the index gesture. The developmental cycle is completed: the cultural higher mental function of the index gesture that initially exists only outside the child on the natural plane becomes the child's own on the psychological plane (Vygotskij 1983). According to Vygotskij, development involves nonlinear and systemic processes of changing relationships and organizations of mental functions.

The issue of the relationship between development and education led Vygotskij to introduce the notion of 'the zone of proximal development' (ZPD) as the gap between actual and potential development (Vygotskij 1978). Vygotskij offered two major interpretations of ZPD. The first interpretation is the ZPD as a gap between the psychological (i.e., completed) plane of development, i.e., what children can do by themselves, and the social (i.e., potential) plane, i.e., what they can do with the help of more capable others (adults or peers).

This use of ZPD defines 'teachability' of the child in a specific activity or in problem solving. If an activity or problem can be accomplished by the child with the help of more capable others, this activity or skill is considered possible to teach the child. If, however, the activity or skill cannot be accomplished by the child with the help of more capable others, it is considered not useful to teach to the child.

Unlike Piaget, who believed that instruction should follow development, Vygotskij argued that guidance can, should, and does lead development. They would differently define what is currently called 'developmentally appropriate curriculum.' Piaget insisted that learning is essentially an individual endeavor and that adults can only facilitate by providing an enriched stimulating learning environment and opportunities for children to share and discuss their egocentric thinking with each other to promote disequilibrium in the child's thinking. Adults should not interfere in the child's individual thinking because it can only lead to imposition of the adult's ideas onto the child—what Piaget called 'sociocentrism.' In contrast, Vygotskij encouraged adults to provide guidance and help and to engage students in activities that are beyond their individual levels of competence ('performance before competence,' Cazden 1992).

Currently there are attempts in education to develop 'scientific instruction' using a medical model of 'educational interventions' (Pease-Alvarez C, personal communication, 19 July 1999). The essence of these efforts is to develop educational diagnostics of students' skill deficits and prescribe an appropriate dose of standardized guidance providing no less and more help to the student than is required (Newman et al. 1989). For that purpose, some researchers have tried to develop a 'ZPD test'—a standardized assessment of a student's teachability. However, it is doubtful that a reliable ZPD test could be developed because, as Newman et al. (1989) demonstrate, the notion of ZPD is relational. A student's teachability depends not only on the student but also on the teacher (and broader communities in which the child participates). Thus, no test of the child alone would accurately determine the child's teachability—the teacher always counts.

According to this neo-Vygotskian view, the ZPD is applied not only to the student but also to the teacher. Both the teacher and the student try to manage the uncertainty that their joint activity creates. During teacher–student instructional interaction, the student learns how to do the classroom activity while the teacher learns how to guide the student. Each engages in their own zones of proximal development, which are

mutually constituted by each other—the student's learning is shaped by the teacher's guidance, while the teacher's guidance is shaped by the student's ongoing learning.

The second interpretation of the ZPD offered by Vygotskij is about the activity in which the child is involved. Vygotskij argued that at different ages, different activities become leading forces in a child's development. For young children, play is the leading activity:

> play creates a zone of proximal development of the child. In play a child always behaves beyond his average age, above his daily behavior … As in the magnifying glass, play contains all developmental tendencies in a condensed form and is itself a major source of development … Action in the imaginative sphere, in an imaginary situation, the creation of voluntary intentions, and the formation of real-life plans and volitional motives—all appear in play and make it the highest level of preschool development. (Vygotskij 1978, p. 102)

He introduced the notion of 'leading activity' as an activity that determines the child's development. El'konin and Leont'ev elaborated this notion by creating a scale of leading activities from infancy to adulthood as the base of age periodization (Griffin and Cole 1984).

Bruner and his colleagues developed the didactic notion of 'scaffolding' in their attempts to operationalize Vygotskij's ZPD in their study of dyadic interactions involving adults tutoring children (Wood et al. 1976). Scaffolding involves a series of adult strategies that assume responsibility for children's learning, such as focusing the child on the task, simplifying the task to fit the child's current mastery, maintaining the child's motivation working on the task, managing the child's level of frustration and risk involved in the problem solving, and demonstrating an idealized version of the required act (Rogoff 1990). Like Vygotskij's ZPD, the scaffolding metaphor is aimed at explaining the gaining of skills in social interaction that the child may apply later on when working alone. It focuses on how the adult fine-tunes the extent of help he or she provides the child for each successful activity outcome while viewing the child as inept, 'One sets the game, provides a scaffold to assure that the child's ineptitudes can be rescued or rectified by appropriate interaction, and then removes the scaffold part by part as the reciprocal structure can stand on its own' (Bruner 1983, p. 60).

In a critique of the concept of scaffolding, Griffin and Cole (1984) emphasize the role of the child in his or her own development, 'The metaphor [of scaffolding] becomes more problematic when we focus not on the execution of a specific task but on the changes in the child' (p. 47). This concept makes it difficult to address the issue of the child's creativity. If adult support is the universal source of a child's development, 'then there is a strong sense of theology—children's development is circumscribed by adults' achieved wisdom' (p. 47). If Vygotskij's concept of ZPD seems to be too focused on the child, then the concept of scaffolding seems to be too focused on the role of the adult in guidance. As a consequence, both concepts are limited by dyadic interaction between more and less knowledgeable partners. Besides, cross-cultural research on guidance suggests that scaffolding is not universal guidance strategy (Rogoff 1990).

Unlike scaffolding, the concept of guided participation developed by Rogoff (1990) can be applied in diverse cultural activities because it focuses on transformation of participation guided not only by more knowledgeable partners but also by sociocultural tools, culturally defined goals and problems, and social arrangements of joint activity:

> Interaction with other people assists children in their development by guiding their participation in relevant activities, helping them adapt their understanding to new situations, structuring their problem-solving attempts, and assisting them in assuming responsibility for managing problem solving. This guidance of development includes tacit and intuitive forms of communication and distal arrangements of children's learning environments; it is often not designed for the instruction of children and may not involve contact or conversation. The model is one of routine arrangements and engagements that guide children's increasingly skilled and appropriate participation in the daily activities valued in their culture (Rogoff 1990, p. 191).

Lave insists that learning is inherent to activity and an aspect of any activity. Learning occurs even despite expectations and wills of experienced members of the community—it is not a matter of whether students learn in school but a matter of what they learn. The students might actually learn what they were not expected to learn and might not learn what was expected for them to learn. Learning is not an independent activity among other activities but, like development, it is an aspect of any activity in the world.

Lave and Wenger (1991) argue that learning is situated in communities of practice. Learning is always a question about membership in the community, about participation in the community practice. A novice is not simply a person who lacks some entities called 'skills,' but rather a newcomer who needs to negotiate her or his participation in the community practice. A person engages in ongoing negotiation of membership/participation in different communities of practice. Learning, as a process of negotiation and renegotiation of participation in the community of practice, is often not prime-time community business; it is going on in the periphery of community activity. Because the community is aware of newcomers, the peripheral processes of negotiation and renegotiation of participation have a legitimate character. Newcomers are anticipated and usually organized by the community. An analysis of situated learning allows Lave and Wenger to construct a new productive concept of 'legitimate peripheral parti-

cipation' that becomes the main definition of learning and development.

Learning as a communal process is both a descriptive and guiding approach. A view of learning as communal processes embedded in communal practices has inspired many educational practitioners and researchers to explore and define new forms of guidance that can be used in schools, such as instructional conversations, reciprocal teaching, cognitive apprenticeship, community of learners, practice and problem-based learning, building a professional community, and dialogic inquiry. This family of instructional approaches and models shares at least the following important principles: Learning is a communal process, learning is embedded in activities and practices in which it occurs, learning involves development and negotiation of new communal identities, students' guided initiation of discourse and definition of problems and goals are crucial for becoming an active member of a community of practice, ownership for guidance and learning should be shared among students and between the students and the teacher, and a community is based on practice and communication. Nonschool sociocultural research on learning shares many of the same principles.

3. *What is the Direction of Development?*

According to Vygotskij, development leads toward a build-up of higher mental functions with mediated cognition increasingly playing the central role among all other psychological functions that promote self-regulation in the individual. Vygotskij's primary focus in his sociohistorical theory of development was on how human society reproduces itself. Vygotskij often saw non-Western cultures as 'historical slices' of the development of Western civilization (Luria 1976, Vygotskij et al. 1993). When the developmental directionality was considered, Vygotskij was, at least, inconsistent with his own tenet that stressed the contextual and activity specificity of intellectual functioning (Cole 1988). Prioritizing self-regulation, mastery of independent working, literacy, schooling, scientific concepts, universal rationality, decontextualization, and systematic hierarchical thinking as the pinnacle of individual's development reveal Vygotskij's ethnocentrism (Rogoff 1990, Wertsch 1985).

A diversity of goals of different communities necessitates defining development in terms of progress toward more responsible participation in specific communities of practice rather than assuming that development is a generic process independent of the goals and institutions of the communities in which an individual develops. As Tharp and Gallimore (1988, p. 31) report, 'Boys in Micronesia, where sailing a canoe is a fundamental skill, will have a ZPD for the skills of navigation, created in interaction with the sailing masters. A girl in the Navajo weaving community will have experiences in a zone not quite like any ever encountered by the daughters of Philadelphia.' At the same time, the developing individual contributes to the further development of the practices (and goals and institutions) of the community.

Griffin and Cole (1984) argue that the zone of proximal development is not just about a gap between the child's present and the adult's past but also between the child's present and the society's future. There is an increasing tendency in the social sciences to consider the developmental telos as a social construction of cultural, institutional, communal, and personal values that have both stable and dynamic components that can conflict with each other. The historical time when positive values of literacy, school, science, rationality, intelligence, and the Western civilization were unquestionable in the social sciences is gone. Stemming from Vygotskij's sociohistorical theory, new approaches to development and learning are emerging.

See also: Cognitive Development: Child Education; Cognitive Development in Childhood and Adolescence; Human Development, Bioecological Theory of; Piaget's Theory of Human Development and Education; Situated Cognition: Contemporary Developments; Situated Cognition: Origins; Vygotskij, Lev Semenovic (1896–1934)

Bibliography

Bruner J 1983 *Child's Talk: Learning to Use Language*, 1st edn. W.W. Norton, New York

Cazden C B 1992 *Whole Language Plus: Essays on Literacy in the United States and New Zealand*. Teachers College Press, New York

Cole M 1988 Cross-cultural research in the sociohistorical tradition. *Human Development* **31**(3)

Davydov V V 1986 *Problemy razvivaushego obucheniay* [The Problems of Developing Learning]. Pedagogika, Moscow

Engeström Y, Miettinen R, Punamäki-Gitai R-L 1999 Perspectives on Activity Theory. Cambridge University Press, Cambridge, UK

Griffin P, Cole M 1984 Current activity for the future: The Zoped. In: Rogoff B, Wertsch J V (eds.) *Children's Learning in the 'Zone of Proximal Development'*. Jossey-Bass, San Francisco, Vol. 23, pp. 45–64

Lave J 1988 *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life*. Cambridge University Press, Cambridge, UK

Lave J, Wenger E 1991 *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK

Luria A R 1976 *Cognitive Development, its Cultural and Social Foundations*. Harvard University Press, Cambridge, MA

Newman D, Griffin P, Cole M 1989 *The Construction Zone: Working for Cognitive Change in School*. Cambridge Univeristy Press, New York

Rogoff B 1990 *Apprenticeship in Thinking: Cognitive Development in Social Context*. Oxford University Press, New York

Tharp R G, Gallimore R 1988 *Rousing Minds to Life: Teaching, Learning, and Schooling in Social Context*. Cambridge University Press, Cambridge, UK

Vygotskij L S 1978 *Mind in Society: The Development of Higher Psychological Processes* (Trans. Luria A R, Lopez-Morillas M, Cole M, Wertsch J). Harvard University Press, Cambridge, MA

Vygotskij L S 1983 *Istoria razvitiya vysshikh psikhicheskikh funktzii* [History of the Genesis of the Higher Mental Functions, 1st edn., Vol. 3. Pedagogika, Moscow

Vygotskij L S, Luria A R, Golod V I, Knox J E 1993 *Studies on the History of Behavior: Ape, Primitive, and Child*. Lawrence Erlbaum Associates, Hillsdale, NJ

Wertsch J V 1985 *Vygotskij and the Social Formation of Mind*. Harvard University Press, Cambridge, MA

Wood D, Bruner J S, Ross G 1976 The role of tutoring in problem solving. *Journal of Child Psychology & Psychiatry & Allied Disciplines* **17**(2): 89–100

E. Matusov

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

16343

 ISBN: 0-08-043076-7

W

Wage Differentials and Structure

At a point in time in a particular economy there will be a distribution of wage rates (average total compensation per hour) across all economically active members of the population. This distribution of wages is referred to as the wage structure of the economy. The value of the wage rate of each member of the labor force will depend systematically on a set of observable characteristics of that person as well as on other characteristics that are not observed. This relation between wages and worker characteristics as well as the distribution of worker characteristics determines the wage structure of the economy. The average wage of a person with a particular characteristic relative to the wage of a person with a different characteristic (e.g., the wage of a worker in Paris relative to one in Lyon), all other observed and unobserved characteristics held constant (*ceteris paribus*), is called a wage differential. This article discusses the effects of different worker characteristics on wage structure.

1. *Earnings and Wages*

There are several different ways of defining the wage rate, and the appropriateness of any particular definition depends on the questions in which one is interested. Table 1 shows the distribution of average hourly gross (pretax) wage rates of a random sample of 4,300 full-time workers in the United States in 1995. In principle, one would like to include nonwage compensation (employer contributions for fringe benefits and social insurance programs), but such data usually must be collected from employers rather than from individuals. A worker at the 10th percentile of the distribution earned \$6.07 per hour, which means that 10 percent of workers earned less than this amount. One way to summarize the wage structure is to compare the average wage in the highest quintile (from the 80th percentile up) to the average wage in the lowest quintile (from the 20th percentile down). These estimated averages for the US in 1995 were, respectively, \$33.15 and \$5.84, a ratio of 5.7.

An alternative to the use of hourly wage rates is the use of average weekly, monthly, or annual earnings of individuals. Total earnings per year for an individual are by definition equal to the average wage times hours per year. Data on individual incomes rather than hourly wages are often more readily available in most countries. An example of this is reported in Table 2 in which the ratio of the average annual income of workers in the top quintile to that in the lowest quintile is reported for a selection of countries. The study from which this study was taken includes data for 108 different countries. The dispersion of incomes within a country generally is higher than the dispersion of wages becauseannual hours tend to be higher for

Table 1
Estimated distribution of hourly wages of full-time workers in the United States in 1995

Percentile	*W*
10	\$6.07
20	7.83
30	9.26
40	10.97
50	12.83
60	14.50
70	17.03
80	20.62
90	27.40

Source: Panel Study on income dynamics data files

Table 2
Ratios of top quintile's share of income to bottom quintile's share for selected countries

Country	Ratio
United States	8.5
Canada	5.5
United Kingdom	4.0
France	6.3
Germany	5.4
Italy	4.9
Czech Republic	3.1
Ukraine	3.7
Japan	7.1
China	5.2
Hong Kong	9.5
India	5.0
Egypt	6.9
Morocco	7.0
Colombia	13.9
Brazil	23.1
Kenya	18.2
South Africa	32.1
Cote d'Ivoire	7.2

Source: Deininger and Squire 1996, Table1

persons with higher wage rates. Notice that for the US the ratio of the highest to the lowest quintile in terms of income is 8.5 compared to 5.7 for the wage rates of full-time workers.

Table 2 shows that there is substantial variation among countries in the dispersion of incomes—as there would be if the data referred to hourly wages rather than annual incomes. A few facts are illustrated by these data. First, there tends to be substantially less dispersion in individual wages and incomes within industrialized countries than in the developing countries. Second, there is much greater dispersion of incomes in the US than in other industrialized countries. Third, the degree of wage and income dispersion in the formerly socialist countries is relatively low.

2. *The Determinants of Individual Wages*

The hourly wage rate (W) of any individual worker in a particular economy at a moment in time will, in principle, depend on a set of observable variables (such as that person's level of completed schooling) as well as on a set of factors that generally are not observed (such as the individual's innate ability, a characteristic that is difficult to measure and usually not obtained in a census-type surveys). Let X represent all the variables determining wages that are observed and U represent all the unobserved determinants. We can then specify what economists call an earnings function in which the value of W for each individual depends on X and U according to

$$W = \beta X + \gamma U + \varepsilon \qquad (1)$$

where ε is a random error term (pure 'luck') that is unrelated to the values of X and U.

β is a vector of coefficients representing the effects of each of the observable variables on wages and γ a vector of coefficients representing the effects of the unobserved variables. If, for example, we thought that there were only two observable determinants of W, years of completed schooling (S) and age (A), βX might be specified as $\beta_1 S + \beta_2 A$, and the average effect of an additional year of schooling on W would be β_1.

In practice most estimated earnings functions feature the replacement of the numerical value of W with its natural logarithm. This means that a unit increase in the value of a particular X variable causes W to increase proportionately by the coefficient on that variable. In the case of years of schooling discussed above, with the semilogarithmic form β_1 represents $(\partial W/\partial)/W$ rather than $\Delta W/\Delta S$ if W is entered in numerical form.

From the earnings function given by Eqn. (1) it is clear that the wage structure, the distribution of **W** across the working population, depends on both the joint distributions of X and U (as well as the distribution of ε) and the parameters β and γ. Wage differentials, the *ceteris paribus* difference between the average wages of individuals with one characteristic versus another, are determined by the β parameters.

The various determinants of individual wages, both potentially observable and unobserved, can be classified into five groups. These include:

(a) skill characteristics,
(b) job location,
(c) job characteristics,
(d) discrimination/nepotism, and
(e) rents.

We will discuss the variables within these groups in turn.

2.1 *Skill Characteristics*

Generally, it is presumed that in most societies more skilled workers are more productive and that they earn higher wages (see *Human Capital*). Two fairly crude but readily observable variables relating to skill of each worker are their years of schooling and potential labor market experience. Some microdata sets allow a more detailed specification of education (what subjects were studied, how many resources were allocated to each year of schooling, etc.) and experience (actual vs. potential years, amount of on-the-job-training, etc.), but this sort of detail is rare.

An important source of differences in skill and productivity differences among individuals is variations in unobserved ability and motivation. Although a few surveys have included the scores on simple IQ tests, these factors are either largely or completely ignored, that is, they are considered U rather than X variables. This illustrates a general problem associated with the estimation of earnings functions. The true model is given by Eqn. (1) above, but the researcher is only able to estimate W as a function of observed variables, that is

$$W = bX + e \qquad (2)$$

where e now represents variation in W that cannot be explained by the observed variables. We know that for most empirical earnings functions e is very important, for the fraction of the variation in W jointly explained by the observable variables is usually fairly small—between 25 and 35 percent (see Juhn et al. 1993).

There is thus a large potential for empirical estimates of the effects on W of any individual observed variable to be biased upward or downward due to the correlation of that variable with unobserved variables. In the very simple model above in which schooling and age are the only observed variables, we want to estimate the 'true' effect of S on W from the model $W = \beta_1 S + \beta_2 A + \gamma U + \varepsilon$, but we are forced because of data limitations to estimate $W = \mathrm{b}_1 S + \mathrm{b}_2 A + e$. The use of b_1 as an estimate of the true effect β_1 is in general subject to specification error which arises if S is

correlated with γU (holding A constant). It seems reasonable to suppose that persons with relatively high levels of intelligence and/or motivation (a) will tend to have higher wage rates (i.e., $\gamma U > 0$) and (b) be more likely to have relatively high schooling levels. This would suggest that b_1 would be an upward-biased estimate of β_1. Interestingly, however, the consensus of the mass of research into this question (see the review by Card 1999) is that conventional estimates of the effect of schooling on wages are probably not significantly biased in an upward direction.

Another feature of empirical earnings functions is that their coefficients may change over time. This is especially true for coefficients concerning skill variables. For example, in the US the relative hourly wage of 30-year old male workers with a bachelor's degree (16 years of schooling) to those with just a high school degree (12 years of schooling) fell from 1.18 in 1973 to 1.11 in 1979, but it then rose during the 1980s to a value 1.41 in 1989 (see Bound and Johnson 1992, Table 1). Why these changes in relative wages occurred (in the US and in other countries with relatively flexible wage determination systems such as Canada and Britain) has received a great deal of attention from economists, but the answer is still a matter of controversy.

2.2 Job Location

A second set of observable variables that are usually included in empirical earnings functions are dummy (one/zero) variables describing the location of the worker within the country. Typically the locational variables represent the size of the area in which the person lives (whether or not, for example, it is a metropolitan area with a population in excess of a million) and indicators of residence in particular areas. For example, in a hypothetical study of the determinants of wages in the Great Britain, one would include a large city dummy variable (L_{large}) and separate dummy variables for persons employed in Greater London (L_{london}), other parts of Southern England (L_{south}), and Scotland (L_{scot}). This specification would be represented as $W = b_1 L_{\text{large}} + b_2 \mathrm{L}_{\text{london}} + \mathrm{b}_3 L_{\text{south}} + b_4 L_{\text{scot}} + \cdots$ with the other observable determinants of wages (the X's in Eqn. (2)) included. The estimated coefficient b_2 in this earnings function would represent the average *ceteris paribus* wage differential between a person working in London and another working in a large city in Northern England. If, as is generally the preferred approach, W is entered as a natural logarithm, the estimated coefficient b_2 would reflect the approximate proportionate wage differential associated with employment in London.

What is the economic interpretation of coefficients on location of employment? First, because of higher land prices, it is generally more expensive to live in areas with a highly concentrated population. In order for real wages (nominal wages divided by the price level) to be equal across areas, observed wages must be appropriately higher in expensive areas. This would apply especially to very large cities within individual countries—London, Paris, New York, Tokyo, etc.

A second reason for the existence of wage differentials by job location is the preferences of workers concerning the characteristics of different areas. For example, living in a large urban area involves greater stress and poorer air quality than living in a rural area, so, holding the cost of living constant, the typical worker might have to be paid a compensating wage differential to work in the large urban area. Similarly, some regions have better weather and other amenities (like public services) than others, and one would expect that wages would be higher in the more desirable than in the less desirable areas in order for potential migrants to be indifferent between different areas. For the US, the estimated effect on wages of an area having sunshine 80 percent of the time vs. only 30 percent is, as simple economic theory would predict, a reduction of about 7 percent (Blomquist et al. 1988).

This, of course, presumes that all (or most) workers in the economy have the same preferences for the amenities of different areas. To the extent that preferences differ across the population—some people prefer rural areas (warm weather) to large cities (cold), others the opposite—predictions concerning compensating wage differentials become somewhat less clear.

A final reason for the existence of wage differentials across locations is the effect of changes in the structure of labor demand. Suppose, referring to the example above, that there was for some reason a major boom in Scotland that caused the demand for labor there to increase permanently by 50 percent. It is unlikely that there would be an immediate large migration of workers from England to Scotland, so wages in Scotland would be bid up significantly. The earnings function for Great Britain based on data collected after the demand shift would yield an estimated coefficient b_{scot} that would now be positive. How rapidly the Scottish/English relative wage would return to one (the value of b_{scot} to return to approximately zero) depends on how long it takes for a portion of the British population to relocate to Scotland.

2.3 Job Characteristics

In addition to the location of employment, an important set of determinants of wage rates is the attributes of the job held by each individual worker. We can arbitrarily define each job attribute (the set of which we will call A) as a bad—the degree to which it is physically onerous, monotonous, a threat to the worker's health and safety, has low social status, etc. It would be expected that higher values of the elements of A would require a compensating differential. Put

Table 3
Estimated relative salaries of Ph.D.'s in the US by field of doctorate and type of employer, March 1995

Field of Doctorate	
Computer science	105
Mathematics/statistics	85
Agricultural science	82
Biology	86
Health sciences	97
Chemistry	85
Physical science	88
Economics	100
Psychology	83
Other social sciences	81
Engineering	97
Humanities	73
Type of Employment	
Academic tenure track	100
Academic nontenure track	87
Postdoctoral fellowships	72
Lower education	89
Business	129
Nonprofit	111
Federal government	115
State and local government	93

differently, holding all other (observed and unobserved) determinants of wages constant, the market equilibrium wage of each job would be higher for 'bad' jobs (those with high values of A) than for 'good' jobs (those with low values of A). Thus, for the specification $W = b_1A_1 + b_2A_2 + \cdots$, where the A_is are numerical representations of bad characteristics (like the injury rate of the job), we would expect to estimate positive b_is.

There are numerous examples of the operation of compensating differentials. (For an excellent introduction to the topic, see Chapter 8 of Ehrenberg and Smith 2000.) On the other hand, there are many instances in which the estimated effects of certain job attributes tend to be zero or have the 'wrong' sign (from the point of view of the theory) (see Brown 1980 for a refreshingly honest discussion of these problems).

An important set of applications of estimates of job amenities on wages is to the question of the economic value of life. If one of the As (say A_1) is the rate of accidental death associated with the occupation of the worker, then one can infer from the estimated value of $\partial W/\partial A_1$ how much individuals value—apart from the opportunity to earn and consume—place on the continuation of their lives. This information can, in principle, be used to do cost/benefit analyses of programs to reduce risks to life and health.

An example of wage differentials associated with differences in job characteristics is given in Table 3. These results illustrate some of the difficulties in the interpretation of the effect of job attributes. The sample refers to individuals working full-time in the US in 1995 who had received a Ph.D. in science or humanities. The regression equation relates the annual salary of each respondent to years since receipt of the doctorate, gender, race, and region as well as the field of receipt of the Ph.D. and the current type of employment.

The range of field effects (with economics equal to 100) is from 73 for humanities to 105 for specialists in computer science. This means that, holding gender, race, years of experience, type of employment, and region constant, the computer/humanities relative salary is 1.44. There are also substantial compensation differences across employer types. For example, Ph.D.s employed in the for-profit business sector earn, *ceteris paribus*, 29 percent more than those in tenure track academic positions.

There are various explanations of why relative compensation differentials of this magnitude exist—in this case among sets of individuals who have the same quantity of education. First, these differentials may reflect compensating differentials in the sense that some fields (economics, for example) are intrinsically boring relative to other fields (like humanities) and require higher monetary rewards in order to induce persons to study them. Some of the relative compensation levels by employer type—for example, the business/academia comparison—most likely reflect premiums that are necessary to get Ph.D.s to work in jobs that are perceived to be less rewarding in a nonfinancial sense.

A second potential explanation of differentials of this sort is that they reflect, in part, average ability differences of the individuals who are in each field or employment type. This is especially relevant for comparisons of employment type in which there is partial sorting on the basis of ability. For example, the compensation difference between academics in tenure track positions and those in both nontenure track positions and lower education (secondary schools and 2-year colleges) is very likely to reflect practices by which the 'better' Ph.D.s are granted tenure.

A third explanation of these compensation differentials is that some of them reflect temporary market factors, and there is a presumption that the high relative compensation of the computer science field in Table 3 is an example of this. This phenomenon is illustrated theoretically in Fig. 1. The demand for a particular type of labor is assumed to increase such that the demand curve shifts from D′ to D″. In the short run the supply curve, S_{SR}, is quite inelastic, reflecting the fact that the pool of persons qualified to work in this market is fixed. (In the case of computer scientists, some additional labor can be obtained from people not on the labor force or from closely related fields, but these possibilities are limited.) Thus, the short-run effect of an increase in the demand for this type of labor is to increase its relative compensation level. In the long run the supply of labor to this market

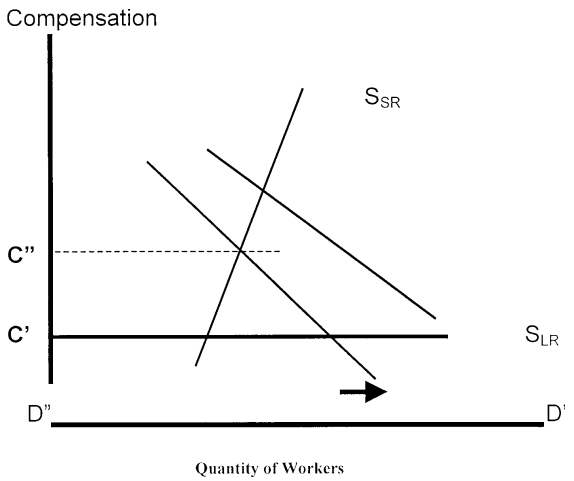

Figure 1
Adjustment of a labor market to demand shock in the short run and the long run

is very or completely elastic, for new entrants to the labor force can choose to train for this type of work. (In the case of computer specialists, young persons are induced to get Ph.D.s in that field rather than in mathematics or physics.) Over time, therefore, the relative compensation level in this market would be expected to return to its initial equilibrium value, C′, rather than stay at its temporary value, C″ in Fig. 1.

Analysis of wage differentials of various sorts can proceed along the above lines. A major example of this is the analysis of average compensation across industries. An example of this is Table 4 in which average monthly salaries across broad industrial groups are reported for six rather diverse countries. Wages in agriculture, trade, hotels and restaurants, and personal services tend to be relatively low; wages in mining, utilities, and finance tend to be relatively high. Some of these differentials reflect quality and compositional differences (that are not adjusted for in the table). But some of the differentials reflect differences in the amenities associated with work in the particular industries.

2.4 Gender and Ethnicity

A set of observable variables that are generally included in earnings functions are the demographic characteristics of the workers—his/her gender, race, or ethnicity. For example, the inclusion of a one/zero dummy variable for women as one of the *X*s in Eqns. (1) and (2) is very common. For the theoretically complete model given by Eqn. (1), which includes *all* observed and unobserved influences on wage rates, the interpretation of the (presumably negative) coefficient on women must be in terms of some sort of labor market discrimination against women. Why else would otherwise identical workers receive lower compensation? The same conclusion applies to conclusions concerning potential dummy variables on race and ethnicity for earnings functions in probably every country in the world (blacks in the US, French speakers in Quebec, Arabs in France, natives of Greenland in Copenhagen, etc.).

The problem with attributing the negative effects of gender and certain racial and ethnic distinctions entirely to labor market discrimination is that we are only able to estimate the model for which we have data, Eqn. (2) rather than the theoretically complete Eqn. (1) above. Thus, there is the possibility that the negative coefficients on gender and ethnicity reflect, wholly or in part, differences in the average value of omitted variables rather than some form of direct

Table 4
Average monthly salaries of employees by industry relative to manufacturing in 1997 for selected countries

Industry	Egypt	Mexico	Estonia	UK[a]	Australia	Romania
Agriculture	0.72	0.52	0.69	0.75		0.76
Fishing	0.66	1.75	1.02			0.67
Mining	1.95	1.90	1.23		1.62	1.67
Manufacturing	1.00	1.00	1.00	1.00	1.00	1.00
Utilities	0.95	1.34	1.35	1.35	1.28	1.84
Construction	1.08	0.86	1.04	1.05	1.12	0.98
Trade	1.04	0.85	0.87	0.96	0.90	0.73
Hotel & restaurant	0.73	0.78	0.65	0.78	0.79	0.64
Transport	1.20	1.20	1.24	1.16	1.15	1.32
Finance	1.59	2.15	2.15	1.41	1.10	2.84
Real estate	2.06	1.26	1.14	1.28	1.09	1.11
Public admin.	0.90	1.40	1.18	1.22	1.14	0.96
Education	0.78	1.59	0.78	1.51	1.27	0.83
Health	0.86	1.41	0.86	1.14	1.07	0.70
Personal services	0.80	0.79	0.81	1.14	1.00	0.83

Source: ILO 1999, Table 5A, a—Weekly salaries

labor market discrimination. Further, it is necessary to know exactly what is the cause of an observed wage differential if a society decides it wants to eliminate this differential (see Altonji and Blank 1999 for an extensive discussion of these issues).

2.5 Rents

Most of the above discussion proceeded as if wages are determined in markets that are similar in process to the markets for melons and equities. The fact that the commodities involved in labor markets are human beings requires that this 'model' be modified to some extent. Indeed, a set of institutions has emerged to provide modification of the outcomes that would follow from the unfettered equilibrium of the labor market. Among these institutions is trade unionism, which has, over time and across societies, taken several forms. Government policy—both directly through legislation determining wages and indirectly through its role as employer—also affects the wage structure.

At one level the existence of institutions such as trade unions creates another wage differential in which to be interested. In terms of Eqns. (1) and (2), being a union member or working in which wage rates are determined by a collective bargaining contract means that a worker could earn a higher wage than another worker who is not so situated. In this case, $W = \beta_1 U + \cdots$ is the complete wage equation, including all observed and unobserved variables, where U equals one with union representation and equals zero otherwise. We do not, of course, observe everything about the workers and his/job, so we estimate the equivalent of Eqn. (2), $W = \mathrm{b}_1 U + \cdots$

How good the estimator b_1 is of the 'true' effect of unionism β_1 depends, as in the estimation of other wage differentials, on the correlation of the omitted determinants of wages with U. In the US, the consensus estimate of the average effect of unionism on wages is about 15 percent (see Lewis 1986).

A more important effect of unionism and government policy is their influence on other wage differentials. Much of the above discussion of the structure of wages and of wage differentials reflected the implicit assumption that relative wages are, at least in the long run, free to adjust to their market-clearing levels. This assumption is all right for the analysis of labor markets in the US, Japan, and (during the past 20 years) the United Kingdom, but there are serious problems associated with its application to most other industrialized countries in which bargaining coverage is over 75 percent (vs. 18 percent in the US and 22 percent in Japan in the early 1990s). To the extent that union and/or government policy attempts to prevent increases in the dispersion of earnings, wage differentials will not behave as neatly as the elementary theory suggests. Instead, wage structures will be subject to a great deal of inertia, and they will adjust at best very slowly to shocks in demand.

That the wage structure is slow to change is not *per se* a bad thing. Indeed, I would join most economists in agreeing that stability of the distribution of income is a good thing. The problem with such policies, however, is that they often lead to heavy unemployment of those groups whose relative wages are protected.

See also: Comparable Worth in Gender Studies; Income Distribution; Inequality; Inequality: Comparative Aspects; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Supply; Labor Unions; Labor, Division of; Sex Differences in Pay; Wealth Distribution; Work, Sociology of

Bibliography

Altonji J G, Blank R M 1999 Race and gender in the labor market. In: Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*, Vol. 3C. Elsevier Science, Amsterdam, pp. 3143–3260

Blomquist G, Berger M, Hoehn J 1988 New estimates of the quality of life in urban areas. *American Economic Review* **78**(1): 89–107

Bound J, Johnson G 1992 Changes in the structure of wages in the 1980s: An evaluation of alternative explanations. *American Economic Review* **82**(3): 371–92

Brown C 1980 Equalizing differences in the labor market. *Quarterly Journal of Economics* **94**: 113–34

Card D 1999 The causal effect of education on earnings. In: Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*, Vol. 3A. Elsevier Science, Amsterdam, pp. 1801–63

Deininger K, Squire L 1996 A new data set measuring income inequality. *The World Bank Economic Review* **10**(1): 565–91

Ehrenberg R, Oaxaca R 2000 *Modern Labor Economics*, 7th edn. Addison-Wesley, Reading, MA

International Labour Office 1999 *Yearbook of Labour Statistics 1999*. ILO, Geneva

Juhn C, Murphy K M, Pierce B 1993 Wage inequality and the rise in returns to skill. *Journal of Political Economy* **101**(1): 410–42

Lewis H G 1986 *Union Relative Wage Effects: A Survey*. University of Chicago Press, Chicago

G. Johnson

War: Anthropological Aspects

Sun-tzu, the famous Chinese military expert, began his book *The Art of War* with the words 'Warfare is the greatest affair of state, the basis of life and death, the Way (Tao) to survival or extinction. It must be thoroughly pondered and analyzed.' These words, 2,500 years later, still hold true: war has grown more deadly, and more damaging to human existence.

In the twentieth century alone, over 250 formally declared wars took over one 100 million lives. Undeclared wars, including political repression, commu-

nal violence, and tribal genocide took millions more; for example between 50–100 million tribal people have been killed by forces and citizens of states in the twentieth century. If we expand the definition of war to include such conflicts as 'the war on drugs' and gang warfare, casualties figures rise, though accurate statistics are not available on these forms of violence.

As we enter the third millennium, one-third of the world's countries are engaged in some form of political violence. Whether these conflicts are called war or not often depends more on political rhetoric than on an accepted definition of the term. In addition, approximately two thirds of the world's security forces use human rights abuses to control their populations. The victims tend to label this violence war or dirty war, while the state tends to classify this as defense or counterinsurgency.

The world has not always been characterized by such high levels of violent warfare. Wars today are longer in duration, more deadly, and kill higher numbers of civilians than wars of preceding centuries. The reasons behind war and the ways in which it is waged change across cultures and time. The changing characteristics of war demonstrate that organized violence is not a fixed and eternal fact of biology, nor an inescapable feature of a Freudian psyche, but a human practice guided by norms of behavior and codes of conduct situated in cultural values.

1. *Defining War*

Neither the two world wars nor the several hundred local and regional wars since 1900 have brought us closer to a shared understanding of war. Most scholars accept a basic definition of war as the deployment of violence to force opponents to comply with one's will. War is organized, group-level, armed aggression rooted in hierarchies of dominance which assume winners and losers in a contest over resources, people, and power. Yet war is defined differently by the winners and the losers, by historical perspective, by soldiers and pacifists—and in each case the definitions are more politically charged than factually correct. For example, freedom fighter, terrorist, insurgent, rebel, traitor, and soldier are all terms variously applied to the same actors by different groups seeking to maximize their own political and moral justifications. Governments define war in their own interests, and militaries are loath to admit strategies that entail civilian casualties, torture, and human rights abuses. The most basic understanding of war is affected by differential and biased reporting; for example, casualty statistics for World War Two vary by millions, depending on the nationality and viewpoint of the researcher. Controlling the definitions of war are integral to the waging of war (Sluka 1992).

The ethnographic study of war and peace has added a new dimension in the understanding of political violence. This academic research has demonstrated that war is a far more complex reality than classical definitions positing a violent contest between two or more armed forces seeking a military, and thus political, victory (Warren 1993, Nordstrom and Robben 1995). Soldiers often battle unarmed civilians and not each other—evident from the ethnic cleansing of the Yugoslav forces in Bosnia and Kosovo or the two million deaths in Sudan's civil war. Paramilitaries, private militias, death squads, and roving bands of armed predatory gangs patrol warzones. Some operate at the behest of state forces while others are independent of all sovereign or rebel control. Mercenary forces are a global phenomenon today, and range from informal groups such as the Yugoslav mercenaries fighting in Central Africa to the formal Executive Outcome organization, comprised of former apartheid South African soldiers, who broker with governments as well as rebel groups. Battlezones are also home to looters, sex workers, criminals, and profiteers. Warzones are a bazaar of international arms and supplies merchants who reap billions of dollars yearly worldwide. International nongovernmental organizations are found in all warzones today, providing services ranging from conflict resolution to humanitarian and development aid. Finally, the fronts of wars are home to the inhabitants. Regardless of formal military regulations mandating the legal role of women, children, and the aged in war, all of these people fight for survival when they find themselves on the frontlines. Armed or unarmed, women defend homes and towns, children are forced to take up arms and fight, and the aged battle forced sieges. The unscrupulous sell out their neighbors for a few coins, and the altruistic set up medical clinics, schools and trade routes to provide critical resources under bombardment.

2. *The Development of War*

War is a fairly recent invention, in terms of the anthropological expanse of human existence. Humans, as a species, have lived 90 percent of their history without war. Social hierarchies and concepts of ownership appear necessary for the advent of war. The earliest form of human organization was the band: fluid egalitarian groups of nomads. The archeological record indicates that while interpersonal violence was known in bands—determined by puncture and crushing wounds from weapons—it was limited. It did not reach the level of formalized intergroup violence among contending warriors.

The first indications of organized warfare occur as ownership of animals, goods, and property create divisions within societies. With the historical development of tribal societies and protostates comes a differentiation in power, and the emergence of organized intergroup violence (Ferguson and Whitehead 1992). These societies did not have standing armies

and military institutions separate from general society; warrior status tended to be open to all able-bodied men, and, less commonly, to women. The early years of tribal war were not necessarily a dangerously lethal activity. For many tribal groups, preparations for war constituted an elaborate ritual process. The rules of engagement were often well delineated: contending factions would meet in full battle regalia and hurl challenges and possibly weapons. Casualties generally brought a halt to the aggressions. Here, it is the display and enactment of power, and not violence, that defines war. Among some communities—the archeological record suggests these were later developments—fighting was much more lethal, though the intent was seldom, if ever, genocide. The goal was to force surrender and extend control over people, property, or territory.

Formalized military institutions and standing armies develop with the rise of the 'state' as a form of political, economic, and social organization. The term state here is used in its anthropological sense—originating some 8,000 years ago, and not in the political science definition as developing in the mid 1600s. (The latter, the modern state, will be considered in the next section.) Chiefdoms are replaced by royal families or governing bodies. Social, gender, and often ethnic inequality is codified in laws of land ownership, labor rights, and inheritance. Dispute resolution becomes formalized into judicial systems, and the legitimate use of force is restricted to state leaders and institutions. Contemporary warfare—fought among contenders for power, privilege, and gain—emerges.

3. *The Changing Nature of War*

Contemporary warfare itself has changed dramatically over time and circumstance, giving lie to any notion that war is a 'natural' social phenomenon or a fixed product of overarching political organization (van Creveld 1991). The era of the modern state provides a good illustration (Holsti 1996). In Europe, the end of the Thirty Years War (from 1618 to 1648) coincided with the beginning of the modern state (marked by the Treaty of Westphalia). The Thirty Years War depopulated a large part of Central Europe. It was known for its sheer brutality: writers of the time speak of the wanton killing, torture, plunder, and destruction of anyone and anything who found themselves in the path of the aggressors. The levels of violence are attributed to the enduring impact of religious wars and the Inquisition, to the transformations wrought by urbanization and early industrialization, and to the upheavals marking the shift from kingly rule to the modern state.

Over the following two centuries the nihilism characterizing the Thirty Years War gave way to what has been called the gentlemen's war of the Enlightenment period. Formal warfare during this era often, though certainly not always, followed strict rules of conduct and engagement: soldiers fought soldiers in hand to hand combat on battlegrounds apart from human habitation. This was not a new era of war for humankind: Buddhist and Hindu scriptures 2,000 years BCE outlined similar 'gentlemen's wars' in Asia.

While military texts tend to focus on these formal military engagements between two contending armies, another form of warfare developed during this period: colonial repression of conquered peoples. In many ways these actions presaged the dirty war of contemporary times—wars that brutally targeted unarmed people in attempts to instill political acquiescence.

The colonial encounter gave rise to another distinct form of war: the guerrilla war, the mainstay of wars for independence worldwide. Guerrilla warfare was developed by nonstate actors challenging financially and technologically superior state forces. Classical guerrilla philosophy—institutionalized in the mid-twentieth century by military strategists such as Mao Zedong, Che Guevara, and Ho Chi Minh—postulates that guerrilla forces, by definition, have the support of the broad population, and it is this that gives them indefatigable strength, crucial resources, and moral political superiority. While in many cases this has proven true, it is by no means always so. Guerrilla groups such as the Khmer Rouge in Cambodia, Renamo in Mozambique, and the Contras in Nicaragua demonstrated that nonstate forces can also use repressive tactics in an effort to control populations.

4. *The Twentieth Century*

The twentieth century was characterized by sophisticated and far reaching developments in the international laws and institutions governing war and protecting peace. Despite these, this period was the bloodiest in human history. Overall, wars in the twentieth century were longer in duration, more lethal in the cost to human lives, and more destructive to societal systems than in preceding times. Simply pointing out that this era saw the advent of world war, high-tech and nuclear war, and modern paramilitary warfare does not convey the changes in the philosophy and conduct of war that occurred during this time (Keane 1996). The most dramatic example concerns the ethics of who may and may not be targeted in war. Over 80 percent of the casualties in World War One were soldiers. With the advances in modern technology and the idea that a country's citizens were now part of the war effort (given their role in producing the means of war), noncombatant casualties rise to 50 percent of all war-related deaths in World War Two. This trend escalates rapidly in the last half of the twentieth century: in the Vietnam war, more than 80 percent of all casualties were noncombatants, and at the start of the twenty-first century, civilians account for 90 percent of all war-related deaths worldwide. Not only has the line between combatant and noncombatant grown indistinct, the line between gender

and age in soldiering has too. Women die in equal numbers to men, and more children are killed in war today than soldiers. The use of child soldiers has grown in recent years: at present over three hundred thousand exist worldwide. This serious rise in civilian casualties parallels an increase in the number and sophistication of international organizations (such as the United Nations and the Organization of African States) and international legal bodies (such as the International Court of Justice and the War Crimes Tribunals) seeking to control destructive wars. Today there are 70,000 protocols protecting human rights.

5. *Theorizing War*

As the predominate means of waging war changes through time, so too do the definitions and theories of war (Simons 1999). In premodern times many, including early Christian and Buddhist societies, saw war as inescapable at times, but not as honorable; the post-war period was crafted as one of atonement. The notion of the honorable war develops in the west with the rise of the 'gentleman's war' of the modern state. Carl von Clausewitz, the famous Prussian military expert, codifies war as 'an extension of politics,' placing war directly in the rational politics of Enlightenment philosophy. Warfare, as rational, became justified—the most dangerous examples of this were seen in colonial conquests, often rationalized under the rubric of the 'evolution of civilization' by scholars.

These Enlightenment philosophies wed with the functionalist and structuralist schools in the early 1900s. Here, theoreticians investigated the 'functions' of war, and placed the causes of war in competition over scarce resources, overpopulation, and the increasing complexity of societies. At the same time, psychological and sociobiological theories were popular. These postulated aggression and self-interest as inherent to humans, and therefore as serving an evolutionary purpose. The fact that not all societies engage in war, and that the majority of the people in any society at war do not choose to fight was not addressed by these theories. Political theory within these schools was shaped by the advent of the world wars. After World War One, functionalist theories take an idealist cast that postulates the progress of civilization as one that will finally eschew war. In the wake of the vast destruction of World War Two, realist theory replaces idealism as the dominate theory in the social and political sciences. Here, war is seen as a natural effect of competition among sovereign states. In both schools the solution lies in creating strong state and international institutions to wage, and ideally to control, war.

As the period of the World Wars gave way to wars for independence, regional wars, and the Cold War, theories of war underwent another revolution. Critical to this shift is the fact that researchers around the world began to experience political violence directly, whether by intention or by accident. The reigning functionalist theories did not fit their observations (Foster and Rubinstein 1986). Clearly, the advent of nuclear war gave lie to ideals of 'victory' in war—for the first time all sides to a conflict could perish. Wars such as those fought by the USA in Vietnam and by the USSR in Afghanistan laid to rest old notions of the gentleman's war. Dirty wars such as those in Argentina, and genocides such as those conducted by the Nazis in World War Two and the Khmer Rouge in Cambodia challenged notions of the inherent functionality or rationality of war. The increases in noncombatant deaths undermined the claims of sociobiology—it became hard to argue that noncombatant deaths, widespread torture, systematic sexual violence, and the death of children were biologically or socially productive acts (Enloe 2000). Scholars also began to question Clausewitz's 'truism' that war is (predominately) an extension of politics. Certainly it is political, but the recognition of the vast sums of money made in wildcatting valuable resources in warzones and in selling war supplies worldwide made it necessary to integrate economics with politics in the war equation (Kaldor 1999). The rise of religious, ethnic, and identity factors in contemporary conflicts rendered it necessary to add social and cultural factors in with the politico-economic ones (Rupesinghe and Rubio Correa 1994). And studies of peaceful societies such as the Semai and the Quakers demonstrated that war is not inevitable, nor basic, to the human condition (Gregor 1996).

6. *The Future of War*

The dawn of the third millennium is marked by vast differences in war. While the superpowers spend trillions of dollars on high technology earth and space based weapons systems, the vast majority of today's war casualties are killed by small arms wielded by nonspecialists. The greatest dangers are the most accessible: there are estimated to be 500 million to one billion firearms in use today in the world; a lively international black-market sells every conceivable implement of war from AK47s to nuclear materials; recipes for chemical weapons can be found in basic texts; and computer specialists can wreck nationwide havoc by disrupting a country's basic infrastructural support systems. All of these facets of the war industry are set in global interactions (Castells 1998). Our theories of war must be revised to address these dynamics defining the contemporary world. Theories of war in the near future will in all likelihood address the complexities of war systems that spend billions on technological defenses (at the end of the twentieth century military spending worldwide reached 780 billion US dollars per year) while killing with inexpensive conventional weapons—and will delve into the cultural factors and economic gains as well as the political quests that underlie these realities. In the long

term, we should be prepared for the possibility that war, as we know it, may not define future conflict. While employing violence in the pursuit of dominance may continue to fuel war, violence may shift from physical killing to a different order of threat and inequality, and dominance might be reckoned along such nonmilitary factors as economics, environmental control, social viability, or a set of factors as yet unrecognized. War has not always been a part of the human condition, and perhaps future changes in sociopolitical organization and ethical systems will render war altogether obsolete. Effective research into the causes, solutions and future of war will hone combinations of theoretical inquiry with ethnography—helping to erase arbitrary distinctions between theory and data (Nordstrom 1997). The greatest advances will be in rethinking the very meanings of violence and aggression, going beyond simple biological and rudimentary social explanations to explore the complex interactions of violence and power, economics, survival, and identity both within and across local, regional, and transnational populations.

See also: Conflict and War, Archaeology of; First World War, The; Military and Politics; Military Geography; Military History; National Security Studies and War Potential of Nations; Second World War, The; Tribe; War: Causes and Patterns; War, Sociology of; Warfare in History

Bibliography

Castells M 1998 *End of Millennium*. Blackwell, London

Enloe C H 2000 *Maneuvers: the International Politics of Militarizing Women's Lives*. University of California Press, Berkeley, CA

Ferguson R B, Whitehead N L (eds.) 1992 *War in the Tribal Zone: Expanding States and Indigenous Warfare*. School of American Research Press, Santa Fe, NM

Foster M L, Rubinstein R A (eds.) 1986 *Peace and War: Cross-Cultural Perspectives*. Transaction Books, Oxford, UK

Gregor T (ed.) 1996 *A Natural History of Peace*. Vanderbilt University Press, Nashville, TN

Holsti K J 1996 *The State, War, and the State of War*. Cambridge University Press, Cambridge, UK

Kaldor M 1999 *New and Old Wars: Organized Violence in a Global Era*. Stanford University Press, Stanford, CA

Keane J 1996 *Reflections on Violence*. Verso, London

Nordstrom C 1997 *A Different Kind of War Story*. University of Pennsylvania Press, Philadelphia

Nordstrom C, Robben A C G M (eds.) 1995 *Fieldwork Under Fire: Contemporary Studies of Violence and Survival*. University of California Press, Berkeley, CA

Rupesinghe K, Rubio Correa M 1994 *The Culture of Violence*. United Nations University Press, Tokyo

Simons A 1999 War: back to the future. *Annual Reviews in Anthropology* **28**: 73–108

Sluka J 1992 The anthropology of conflict. In: Nordstrom C, Martin J (eds.) *The Paths to Domination, Resistance and Terror*. University of California Press, Berkeley, CA, pp. 18–36

van Creveld M 1991 *The Transformation of War*. The Free Press, New York

Warren K B (ed.) 1993 *The Violence Within: Cultural and Political Opposition in Divided Nations*. Westview Press, Boulder, CO

C. R. Nordstrom

War: Causes and Patterns

War involves large-scale organized violence between states or other political units. Although the conduct of war has changed in important ways over the millennia, war itself has been a recurrent phenomenon in international politics. It is one of the primary sources of change in international systems and an important factor in the evolution of the social and political organization of societies. Theorizing about the causes of war goes back to Thucydides' *History of the Peloponnesian War* between Athens and Sparta, but scholars are far from agreement on what causes war.

1. Patterns of Warfare

The current international system represents the most recent stage in the evolution and globalization of the system that originated in Europe about five centuries ago. Warfare in this system has historically been dominated by the 'great powers,' though the frequency of wars between these leading states has steadily declined, while their severity has increased. The period since 1945 has been characterized by both the longest period of great power peace in the last half millennium and a dramatic shift in the concentration of war from Europe to other regional subsystems and from international wars to internal wars, many of which have been intractable ethnonational or religious 'identity wars.' These recent trends have led some to argue that we have reached a turning point in the history of warfare. Some argue that major war between advanced industrial states has become obsolete, while others argue that traditional wars over power or ideology will give way to a 'clash of civilizations' defined in terms of religious or cultural identity (Huntington 1996). These arguments reflect different theoretical perspectives on the causes of war.

2. Theoretical Approaches

Carl von Clausewitz (1976) wrote in his influential book *On War* that war is a 'continuation of politics by other means,' suggesting that war is an instrument of policy for advancing state interests. This implies that war ultimately involves a *political* decision by state

political leaders, so to understand war one must understand why political leaders choose war rather than other strategies to achieve their ends. Technically, we must understand the *joint* decisions by rival states, because one side can usually avoid war if it is willing to make enough concessions.

2.1 The 'Levels of Analysis' Framework

Scholars previously emphasized monocausal explanations that identified a single primary cause of war, but political scientists have moved away from such explanations. Although they prefer parsimonious explanations that explain as much as possible with as little theoretical apparatus as possible, they generally recognize that there are many possible causes of war and that there is no single factor that is either necessary or sufficient for war. One analytic framework that they have found useful for categorizing the many possible causes of war is based on patterns of causation located at different 'levels of analysis': international system, nation-state, and individual. The first focuses on threats and opportunities to states that originate in their external environment and that affect the 'national interests' of the state as a whole. The second emphasizes the internal sources of foreign policy decision making that derive from either governmental structures or processes or from societal influences outside of the government. The third emphasizes the distinctive role of key individual decision makers in the processes leading to war.

2.2 Systemic-level Theories of War

Systemic-level causes of war include the anarchic structure of the international system (defined as the absence of a legitimate authority to regulate disputes and enforce agreements), the distribution of military and economic power among the leading states in the system, patterns of military alliances and international trade, and other variables deriving from the external environment of states. The leading systemic-level approach is 'realist theory,' which begins with the assumption of the primary role of sovereign states who act rationally to advance their security, power, and wealth in an anarchic international system. Given uncertainties regarding the current and future intentions of the adversary, political leaders focus on short-term security needs, adopt worst-case thinking, engage in a struggle for power, and utilize coercive threats to advance their interests, influence the adversary, and maintain their reputations.

At a very general level, realist theory posits two distinct paths to war. In one, the direct conflict of interests between states leads at least one side to prefer war to any feasible compromise. In the second, states prefer peace to war but are driven by the structure of the situation and by uncertainty regarding the intentions of others to take actions to protect themselves through armaments, alliances, and deterrent threats. These actions are often perceived as threatening by others (the 'security dilemma') and often lead to counteractions and conflict spirals which sometimes escalate to war.

The leading realist theory is balance of power theory. Although there are several versions of balance of power theory, most posit that the primary goal of states is to avoid hegemony, to prevent any single state from achieving a position from which it can dominate over others. This leads to the instrumental goal of maintaining a balance of power through the internal mobilization of military power, external alliances against potential aggressors, or the use of force if necessary. The theory predicts that this balancing mechanism almost always works successfully to avoid hegemony, either because potential hegemons are deterred by their anticipation of a military coalition against them or because they are defeated in war after deterrence fails.

Another theory that gives primary emphasis to the systemic-level sources of war, but that is associated with a liberal perspective that downplays the conflictual consequences of anarchy, emphasizes the potential for cooperation among states, and includes some domestic factors as well, is the liberal economic theory of war. The core of the theory, which originates with Immanuel Kant's *Eternal Peace* (1795/1977), is that trade promotes peace. Trade leads to economic benefits, but the economic interdependence generated by trade leaves states vulnerable to any disruption through war, and the fear of economic disruption and the loss of the gains from trade deter political leaders from taking actions that are likely to lead to war. Realists challenge this view and argue that because trade and interdependence are usually asymmetrical they often contribute to conflict rather than deter it, either because states may be tempted to exploit their trading partner's vulnerabilities or because domestic groups vulnerable to external economic developments demand protectionist measures, which can lead to retaliatory actions, conflict spirals, and war.

2.3 National-level Theories of War

Systemic-level theories, with their emphasis on the external forces that shape state decisions for war, posit that states in similar situations behave in similar ways. The implication is that factors internal to states have little impact on foreign policy decisions. There is substantial evidence, however, that decisions for war are often influenced by internal political and economic structures, political cultures and ideologies, and domestic political processes, and over the last decade international relations theorists have been giving more attention to domestic factors.

Regime type is particularly important, based on evidence that democratic regimes behave differently in

important respects than do authoritarian regimes. Although democracies get involved in wars as frequently as do authoritarian states, frequently fight imperial wars, and once involved in war often adopt a crusading spirit and fight particularly destructive wars, it is striking that democracies rarely if ever go to war with each other. This 'interdemocratic peace' is based on standard definitions of democracy (fair, competitive elections and constitutional transfers of executive power) and war (which is often distinguished from lesser conflicts by the threshold of a minimum of 1,000 battle-related deaths).

There are several interrelated explanations for interdemocratic peace. To be valid these explanations must account not only for the near absence of war between democracies but also for the fact that democracies get involved in wars just about as much as other states do. One model emphasizes the institutional constraints on democratic leaders—checks and balances, the dispersion of power, and the need for public debate—that enable governmental or societal groups to block attempts by political leaders to take the country into war. Related to this 'institutional model' is the 'political culture model,' which suggests that the norms of peaceful conflict resolution that have evolved within democratic societies are extended to relations between democratic states, and that these norms facilitate negotiated settlements.

Authoritarian leaders face fewer institutional or cultural constraints, and they often attempt to exploit the conciliatory tendencies of democracies. This undermines democratic political leaders' expectations that their conciliatory negotiating strategies will be reciprocated, reduces the internal constraints on their use of force, and provides incentives for democratic regimes to resort to force against authoritarian regimes both to protect themselves and sometimes to facilitate democratic transitions.

The institutional model of interdemocratic peace assumes that political leaders are more inclined to war than are their peoples, but this assumption does not always hold. Jingoistic public opinion, often exacerbated by the media, can force political leaders into wars that they would prefer to avoid or preclude them from making the concessions that might prevent war. There is a strong tendency for the use of force against external adversaries to generate a temporary boost in domestic support for political leaders in the form of a 'rally round the flag' effect. Political leaders anticipate this, and are sometimes tempted to undertake risky foreign ventures in an attempt to distract attention from domestic problems or to blame other states or groups for those problems. Many contemporary ethnic wars result in part from political leaders manipulating images of ethnic rivals and mobilizing their domestic publics against those rivals in order to serve their own narrow political interests. External scapegoating can backfire, however, if it results in a military defeat.

2.4 *Individual-level Theories of War*

Whereas systemic and national-level theories emphasize the role of international and domestic forces that lead to war and suggest that individual political leaders have little impact, other theories give significant causal weight to individuals, their beliefs about the world and specific adversaries, the psychological processes through which they acquire information and make decisions, and their personalities and emotional states. Some theories emphasize cognitive limitations and affective variables that impact most people in similar ways and result in standard patterns of deviations from ideal-type models of rational decision making. Other theories emphasize the variations among political leaders in the way they define state interests, perceive threats to those interests, assess the intentions of adversaries, evaluate the merits of alternative strategies to achieve those interests, use the lessons of history to shape current policies, and respond to the pressures and uncertainties of foreign policy crises. Misperceptions of the intentions and capabilities of adversaries and third states can be a particularly important cause of war.

3. *Conclusion*

Although the levels-of-analysis framework initially led scholars to focus on the question of which level of analysis was most important in the causes of war, and thus to emphasize single-level explanations, attention has recently shifted to the question of how variables at different levels interact in the processes leading to war.

See also: Alliances: Political; Balance of Power: Political; Cold War, The; Conflict and War, Archaeology of; Conflict Sociology; Deterrence; Foreign Policy Analysis; Imperialism: Political Aspects; Internal Warfare: Civil War, Insurgency, and Regional Conflict; Military and Politics; Military Geography; Military History; Military Sociology; National Security Studies and War Potential of Nations; Peace; Peace Movements; Realism/Neorealism; War: Anthropological Aspects; War Crimes Tribunals; War, Sociology of; Warfare in History

Bibliography

Clausewitz C von 1976 *On War* Howard M and Paret P. Princeton University Press, Princeton, NJ

Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon and Schuster, New York

Jervis R 1976 *Perception and Misperception in International Politics*. Princeton University Press, Princeton, NJ

Kant I 1977 Eternal peace. In: Friedrich C (ed.) *The Philosophy of Kant*. Modern Library, New York, pp. 430–76

Levy J S 1983 *War in the Modern Great Power System, 1495–1975*. University Press of Kentucky, Lexington, KY
Ray J L 1995 *Democracy and International Conflict*. University of South Carolina Press, Columbia, SC
Thucydides 1954 *History of The Peloponnesian War*. In: *The Landmark Thucydides*, ed. and trans. Strassler R B. Free Press, New York
Van Evera S 1999 *Causes of War*. Cornell University Press, Ithaca, NY
Vasquez J A 1993 *The War Puzzle*. Cambridge University Press, Cambridge, UK
Waltz K N 1959 *Man, the State, and War*. Columbia University Press, New York

J. S. Levy

War Crimes Tribunals

From Nuremberg to The Hague and Kigali, the latter half of the twentieth century witnessed the beginnings of an international framework for the prosecution and punishment of war crimes. Yet progress in this area has been sporadic and discontinuous, often seeming to reveal as much or more about what does *not* work as about what does. Contemporary efforts to establish a permanent international criminal court usually are traced to the Universal Declaration of Human Rights adopted by the United Nations General Assembly a half-century ago. These efforts continued through a summer of 1998 meeting of more than 100 nations in Rome that led to a tentative treaty to establish a permanent international criminal court (ICC). Continuing efforts include war crime tribunal activity in relation to the wars in Rwanda and the former Yugoslavia and ongoing efforts to ratify and implement the treaty for a permanent court through the United Nations. Although there is much anecdotal and growing empirical evidence of the need for a permanent institution to deal with war crimes, there is a lack of social science analysis and a resulting knowledge base that could help to broaden the foundation of support for an international criminal court by the public. The agenda of such a court increasingly is recognized as not only involving righting wrongs of the past but also preventing war crimes in the future.

1. *The Lessons of War*

A lesson of World War I was that, left to their own devices, vanquished as well as victorious nations often will do little to punish war criminals appropriately (Marrus 1997). Following World War I, the vulnerable politicians of the fledgling Weimar Republic insisted that the prosecution and punishment of German war criminals should be handled by the German Supreme Court in Leipzig. The resulting sentences ranged from two months to four years of imprisonment and were given to only six of 12 war criminals tried from a list of more than 900 suspects supplied by the Allied forces.

The proceedings at Nuremberg following World War II were a notable improvement. The results of Nuremberg are made more notable by the fact that public opinion tended to simply favor summary execution of the captured leaders of the Axis powers. Franklin Roosevelt's Secretary of War, Henry Stimson, argued instead for a trial before an international tribunal. Roosevelt's successor as president, Harry Truman, endorsed the trial model and persuaded the Allied powers to work through the United Nations in establishing the International Military Tribunal (IMT) to prosecute and punish 'crimes against peace,' 'war crimes,' and 'crimes against humanity.'

The resulting Nuremberg Indictment was regarded as an extension of the common law of the nations involved, although this common law was far from universally recognized at the time of Nuremberg. A defense posed in response to the resulting prosecutions was that the tribunal's charges were *ex post facto*, retroactive creations. Marrus (1997, pp. 565–6) notes that the Harvard criminologist and law professor, Sheldon Glueck, responded with the common law retort that

> Surely … Hitler, Himmler, Goering, Ribbentrop, Frank, Doenitz, and the rest ... knew full well that murder is murder, whether wholesale or retail, whether committed in pursuance of a gigantic conspiracy to disregard all treaties and wage lawless wars or of a smaller conspiracy evolved by a group of domestic murderers.

Twelve war criminals included under the Nuremberg Indictment received death sentences. Three acquittals also demonstrated that such a tribunal could provide reprieve as well as punishment. The development of a jurisprudence of 'crimes against humanity' was especially noteworthy as an innovation that addressed the persecution of specific groups, in this case including the German Jews as well as other groups. And, of course, Nuremberg helped establish in the public mind the responsibility of individuals to refuse illegal orders. The tribunal's work also brought into limited question some Allied war strategies and methods, including the firebombing of German cities.

But the Nuremberg Tribunal also received justifiable criticism. The presence on the tribunal of Russian judges and the impunity of Stalin and other Russian leaders belied their wartime atrocities. Some charges of 'crimes against peace' were clumsily conceived and pursued. In spite of Eleanor Roosevelt's successful work for the adoption of the Universal Declaration of Human Rights in 1948, this effort expressly rejected a standing tribunal to deal with future war crimes. Similarly, the impetus of the

Nuremberg Tribunal was to put the Nazi war crimes into the past, and the IMT did not become the model it might have for a continuing institutional response to the problems of war crime.

2. *The Balkans and Rwanda*

There certainly was no absence of need for the prosecution and punishment of war crimes following World War II and in our more recent history. A partial contemporary list would certainly include Argentina's 'dirty war,' South Africa's apartheid, Pinochet's Chile, the acts of communist Eastern European governments, as well as more recent events in Bosnia, Kosovo, and Rwanda (Neier 1998). During the Cold War, which lasted for most of the second half of the twentieth century, human rights violations were routinely ignored by the West in the process of supporting anti-communist governments. As the Cold War ended, this tendency was extended in the ostensible interest of promoting democratization. For example, in Pinochet's Chile and Botha's South Africa, amnesties became bargaining devices to promote democratic transitions. In Latin America and in South Africa, truth commissions became compensatory vehicles of reconciliation. The liability of this approach, of course, is that political and military mass murderers have escaped prosecution and punishment.

Events in the Balkans and Rwanda evolved against this international backdrop, with the difference that feminist scholars and activists became involved in exposing the systematic use of rape as an instrument of terror and intimidation in waging war. A United Nations commission assisted by human rights groups found that in the 1990s ethnic Serb military and paramilitary groups had systematically tolerated and encouraged the raping of Bosnian Muslim women as part of the effort to drive Muslims from their communities. This form of sexual violence was recognized as a deliberate tactic used to terrorize and humiliate a civilian population not only through the immediate assaults involved, but also through the creation of forced pregnancies. Rape was also instrumental in the genocidal campaign waged by Hutu troops against Tutsi women in the war conducted in Rwanda in 1994. Both in Bosnia and Rwanda, these tactics were recognized as premeditated acts of war aimed at women.

Commission and human rights work in the Balkans and Rwanda formed the foundation for the demand that rape be recognized internationally as a war crime, in a manner analogous to assaults against noncombatants, mistreatment of prisoners, torture, and other systematic, deliberate, and premeditated forms of violence. This innovation was crucial in the formation by the United Nations Security Council of two *ad hoc* criminal tribunals for the former Yugoslavia and Rwanda in 1993 and 1994 respectively. Justice Richard Goldstone served as the first war crimes prosecutor for the Balkans and Rwanda and played an important role in successfully treating sex crimes as a form of genocide and as crimes against humanity.

Each of the *ad hoc* tribunals currently has investigators, police officers, military analysts, legal advisers, trial lawyers, and judges who are involved in developing and trying cases, from the exhumation of mass graves through the preparation of formal indictments and court proceedings, with trials, dispositions, and appeal hearings held in Arusha, Tanzania and The Hague, Netherlands. The tribunal investigators and prosecutors play unique roles through their daily contacts with victims and witnesses. This involves building the trust and cooperation that is necessary in translating victims' and witnesses' stories of atrocities into the testimony that is needed to try cases. The most severe sentence that is available to the tribunals is life imprisonment.

The tribunals have indicted about 20 persons a year since their formation. Moving from indictment to arrest and prosecution is often a formidable process, since the tribunals must rely on the police and armed forces of cooperating nations to apprehend and take indicted persons into custody. As of spring 1999, 34 persons had been arrested and charged before the Rwanda tribunal, and 26 before the tribunal for the former Yugoslavia. Assuming custody and the initiation of a prosecution, the nature of the cases and severe prospective punishments creates further incentives for the accused to delay the beginning of a trial and then pursue a lengthy defense. Justice Goldstone's successor as the chief tribunal prosecutor, Justice Louise Arbour, attempted to speed up the process by trying accused in groups, but she was rebuffed by the Court. Justice Arbour is recognized for numerous reforms and innovations that have dramatically improved and expanded the role of the *ad hoc* tribunals. The tribunal in Rwanda was mired in United Nations bureaucracy and allegations of corruption until a shakeup by Arbour that led to increased arrests and the successful prosecution of a former prime minister held responsible for a genocide that included the slaughter of more than 700,000 Rwandans in 1994. At one point, Arbour was physically prevented from entering Kosovo by Yugoslavian border guards. Undeterred, she pursued the indictment of former Bosnian Serb President Radovan Karadzic and the General Ratko Mladic. When Arbour assumed her position there existed no clear precedent for indicting a sitting head of government for war crimes, although Admiral Karl Doenitz, as the Nazi head of state following Hitler's suicide at the end of World War II, had been convicted and imprisoned for war crimes at Nuremberg. Arbour established a clear precedent when she obtained a war crimes indictment against Serbian President Slobodan Milosevic. She gained further attention when she remarked in public that

NATO leaders could conceivably stand trial for air strikes that caused civilian deaths in Yugoslavia. This signal of evenhandedness is important because figures such as Milosevic have drawn comparisons between killings in Kosovo and the execution of Vietnamese villagers by American soldiers during the war in Southeast Asia. One of Arbour's most significant achievements involved bringing war crimes prosecutions fully into the present as a 'real-time' institutional response to atrocities and genocide virtually as they happened in Kosovo.

The scale of the tasks that confront war crimes tribunals of the kinds recently established is suggested by the situation in Rwanda where more than 100,000 people, almost all Hutus, have been held in jail on charges stemming from the massacres and other crimes involved in that country's genocidal conflicts. No legal system can put all of the guilty through full and formal trial procedures, but prosecuting high-level or representative perpetrators is essential to achieving a sense of justice. There is a clear contrast between the Western-influenced emphasis on time-and-resource-consuming procedural safeguards in the tribunal context and the expedited forms of criminal and civil dispute resolution practiced in settings such as the Rwandan courts. Some combination of jurisdictions is required, with tribunal responsibility for cases with greater potential impact nationally and internationally. This kind of issue has complicated calls for a more permanent court for war crimes.

3. *An International Criminal Court*

The Balkan and Rwandan tribunals have served as prospective models for a permanent international court. In July 1998, more than 100 countries met in Rome with American participation to negotiate a treaty creating a permanent war crimes tribunal that would act independently of the United Nations and its Security Council. The result was a vote approving in principle an International Criminal Court (ICC) with powers to try cases of genocide, war crimes, crimes against humanity, and crimes of aggression. The positive votes for the treaty to establish the court came from 120 nations, with 21 nations abstaining and seven nations opposed, including the USA. Unlikely allies of the USA in this opposition included Libya, Algeria, China, Qatar, and Yemen, with Israel also not signing.

American officials involved in the negotiation of the treaty were especially concerned about the risks the ICC could pose to its own military and political purposes and personnel. These officials were quick to note that with about 200,000 military personnel permanently stationed in about 40 countries around the world, the USA has much at stake. The risk for the USA, it is argued, is increased by expectations that this nation act quickly in crisis situations, and that the world relies on the USA's being able to do so without fear of arbitrary litigation and liability. The worry is that countries such as Cuba or Libya would use the court to harass the USA in its legitimate efforts. To address this concern, the US negotiators in Rome sought guarantees that American military forces and agents would not be prosecuted by the ICC without US consent. At one point in the deliberations an American motion to obtain such guarantees was refused by a vote of 113 to 17, with 25 abstentions. The concern was that if the USA achieved such a loophole, other nations would expect similar protection. Despite the strong support from most nations for the establishment of the ICC, its future remains in doubt. The treaty requires formal ratification by at least 60 nations before it can take effect.

The provisions of the treaty provide that the ICC could proceed with a case if the accused was from a signatory nation or the crime took place within a country that had ceded jurisdiction to the international body. The court could not take jurisdiction over a case unless a national court failed to act, leading advocates to argue that its existence could constitute an important institutional incentive for countries to prosecute their own war criminals. The court would be made up of an international panel of judges, a prosecutor's office, and an administrator; it would presumably rank along with other international institutions such as the World Trade Organization, the International Monetary Fund, and the World Bank. The jurisdiction of this court is over individuals, as distinct from the International Court of Justice in The Hague, which settles disputes between countries. A key premise is that an indictment by the court of a citizen or a leader of a country would be a serious matter that would restrict the movement of those involved and in this way immediately have a significant impact, even without arrest or prosecution.

Hundreds of citizen human rights and legal groups, such as the New York-based Human Rights Watch and the International Center for Human Rights and Democratic Development in Montreal, support the creation and establishment of the ICC. The failure of the USA to strongly support the ICC is a frustration to many of these groups. Eleanor and Franklin Roosevelt were prime movers in the adoption of the Universal Declaration of Human Rights, the founding document from which many treaties of this type have followed. Yet the USA has often lagged behind other nations in the ratification of such treaties. For example, it took the USA fully 40 years to ratify the Genocide Convention. So while the USA played a major role in establishing the Rwandan and Balkan tribunals, it continues to maintain a reluctant stance with regard to the ICC.

Perhaps the strongest argument that can be made against American approval for the ICC is the control the USA might lose over the use of war crimes prosecutions as a bargaining chip in its negotiations

over the complicated conflicts it seeks to control around the world. The perplexing complexity of such negotiations was revealed with all of its frustrations in the Balkan conflict. The Clinton administration sought a settlement that would avert a war over Kosovo through its negotiation of the Dayton Accords. At first it seemed that the United Nations war crimes tribunal had assisted this process by indicting Radovan Karadzic and elevating the position of Slobodan Milosevic. However, Milosevic soon proved to be no better a partner for peace than Karadzic, and the Serbian violence was allowed to continue, resulting in his belated indictment as well. The lesson may simply be that mass killings and genocide will continue to occur until a system of deterrents comes into effect through an institution such as the ICC.

It is of special note that two of America's closest allies, Canada and Great Britain, have taken a particular interest in the Rwandan and Balkan tribunals and the ICC, often seeming to lend more durable support than the USA. The Canadian Foreign Minister, Lloyd Axworthy, built on the leadership role played by the Canadian jurist, Louise Arbour, during her term as Chief Prosecutor of the Rwandan and Balkan tribunals. Both have emphasized that armed conflicts in the world now increasingly involve struggles within rather than between nations, with parallel increases in civilian as opposed to military casualties. They argue this requires new ways of thinking about and dealing with issues of national sovereignty, so that efforts to protect the latter do not become shields used by political and military leaders to cloak such crimes as murder, torture, rape, forced deportations, and expulsions. During the Kosovo crisis in particular, and the subsequent Milosevic indictment, the British government showed a firm resolve to use international criminal law to deal with intra-national crimes against humanity.

The Canadian-born and now British-based journalist Michael Ignatieff, and the British lawyer and human rights activist Geoffrey Robertson, have been among the most effective spokespersons for a permanent international criminal court. Ignatieff's (2000) book, *Virtual War*, makes a powerful argument for just how much the politics and technology of war have changed in recent decades. Ignatieff argues that it is our failure of moral imagination to keep pace with these changes that inhibits the movement to a permanent international criminal court. Robertson's (2000) book, *Crimes Against Humanity*, paints a historically and geographically sweeping picture of a world in which violent human rights abuses more often than not go unpunished. Nonetheless, Robertson argues that the Balkan and Rwandan tribunals signal a new 'age of enforcement' that is promising and already in progress. Many share Robertson's hope that these tribunals will continue to evolve into a permanent ICC.

4. *Research Priorities*

Many important questions remain regarding war crimes tribunals. In particular, it will be important to learn how the international prosecution of war crimes is organized and to contrast this with the work of national justice systems, such as Rwanda's, which may also prosecute and punish war criminals in large numbers. Answers to the following kinds of questions would be especially useful in expanding our knowledge base: How are the presumed worst offenders and offenses selected for international prosecution? How are prosecutors and judges selected for tribunal work? What are the patterns of conviction and punishment? What is the role of international public opinion in influencing the tribunals and the prospective permanent court? What are the effects on the prosecutors and judges who become involved in this work? What are the prospects of a more permanent international criminal court and how might its work differ from the current efforts of the *ad hoc* tribunals? The importance of international war tribunal work seems obvious, and our knowledge of this work is unfortunately modest.

See also: First World War, The; Second World War, The

Bibliography

Allen B 1996 *Rape Warfare: The Hidden Genocide in Bosnia-Herzegovina and Croatia*. University of Minnesota Press, Minneapolis, MN

Bassiouni M, Manikas P 1996 *The Law of the International Criminal Tribunal for the Former Yugoslavia*. Transnational Publishers, New York

Finkielkraut A 1992 *Remembering in Vain: The Klaus Barbie Trial and Crimes Against Humanity*. Columbia University Press, New York

Ignatieff M 2000 *Virtual War: Kosovo and Beyond*. Henry Holt, New York

Marrus M 1997 The Nuremberg trial: fifty years after. *The American Scholar* **66**: 563–70

Neier A 1998 *War Crimes: Brutality, Genocide, Terror, and the Struggle for Justice*. Times Books/Random House, New York

Prunier G 1995 *The Rwanda Crisis: History of a Genocide*. Columbia University Press, New York

Robertson G 2000 *Crimes Against Humanity: The Struggle for Global Justice*. Penguin Books, London

J. Hagan

War, Political Violence and their Psychological Effects on Children: Cultural Concerns

1. *The Problem*

Social science research on children and war can be traced back to at least the first World War but began

in earnest during World War II. For most of this time this work has been regarded as outside the mainstream of psychological research largely because it was believed that few children were exposed to war or political violence. This is beginning to change because today civilians are increasingly becoming the victims of war. World War I recorded 10 percent civilian casualties; World War II some 50 percent, but in all subsequent wars around 80 percent of casualties have been civilians.

While it is difficult to say how many of these casualties have been children, one estimate is that during the 1980s alone some 1.5 million children may have died while an additional 4 million may have been injured. Further, even more children (perhaps as many as 15 million) may have managed to escape physically but still have had their lives thrown into turmoil as they and their families fled to other countries or became 'internally displaced' within their own countries.

2. *Mental Health*

Despite this long history the field today still focuses mainly on its initial area of concern—the possibility that exposure to political violence is a traumatizing experience for children. Certainly there is evidence that war experiences can lead to clinical illnesses involving a wide variety of variety of symptoms typically referred to as post-traumatic stress disorder (see *Anxiety Disorder in Children*).

The empirical evidence however, suggests that such suffering is not inevitable. This conclusion is based on relatively sound evidence that has been accumulated over the last half century. The only weakness in this area, it could be argued, is that researchers have tended to rely too heavily on teacher and parental estimates of child stress levels, with not enough direct measures from children themselves. A further problem is the almost complete lack of studies in which children have been seen both before and after their exposure to political violence.

The realization that children show resilience even when exposed to political violence has led investigators to search for those factors which may be helping at least some children to escape the more severe forms of stress. Early candidates in this area included personal factors such as age, sex, and personality. Surprisingly, in the context of political violence, very few studies have examined these in any systematic way, which makes it difficult to draw clear conclusions.

Additional sources of social support have attracted speculation including the extended family, the child's peers, and the wider community. Here the evidence is almost entirely anecdotal but provides many interesting ideas that warrant empirical investigation. The cognitive-phenomenological model of coping may indicate a way forward in this area. For example there is some evidence that the child's religious and political beliefs may influence the way in which political violence is appraised. On the other hand, we appear to know almost nothing about children's sources of information regarding political violence and how they evaluate these sources.

Also problematic is the long-term effect of exposure to political violence. Evidence from holocaust survivors provides the largest source of information on this subject. Here it appears that claims of universal difficulties in later life have possibly been exaggerated. In contrast, evidence from children who become refugees is mildly optimistic, at least in the medium term, with the role of the child's pre-refugee family support emerging as an important factor. Again, much of this is based on weak evidence. Unfortunately studies in this area often lack adequate (or any) control groups, while samples are open to bias because they are obtained from clinical sources.

3. *Attitudes and Values*

Either because children are modeling themselves on the adults around them or because aggressive behavior is a form of coping, the expectation has been that exposure to political violence will lead to an increase in aggressive behavior, for example when interacting with peers.

Several well-designed studies have examined this question, but evaluating the evidence is difficult because no two studies are identical. Some have involved children exposed to long-term political violence, but only one of these employed observational methods. In another study the children were observed but had only been exposed to political violence for a short time. Therefore despite the unusually high quality of the evidence, no clear conclusions can be reached. If nothing else, this illustrates the urgent need for cross-national replication of research in this area.

There has also been much speculation that children's moral attitudes and values may be altered by exposure to war, however, once again there is little empirical evidence testing this hypothesis. What evidence there is appears to come almost exclusively from one society—Northern Ireland. The suggestion is that children in Northern Ireland may have an underdeveloped sense of the complexity of moral problems. This comes from research using measures of moral reasoning. However, when moral behavior (measured using paper and pencil tests) has been examined, no effect for political violence has been found.

Whatever the explanation, permanent changes in morality are almost certainly not linked to an increase in crime among young people at this time. Juvenile crime does increase with increasing political violence but the most likely explanation for this is the absence of authority figures, such as fathers or the police, combined with extreme levels of deprivation.

4. *Social Problems*

Of course the 'normal' problems that children face during childhood do not evaporate simply because of the onset of political violence, in particular the disadvantages associated with lower socioeconomic status. Therefore, even for children in societies where there is political violence, the major hurdles of everyday life remain to be surmounted. In addition, because political violence can damage the general infrastructure of society, for example, disrupting medical services or the food supply, children may be exposed to many risks other than the obvious problems of dealing with bombs and bullets, including increased risk of accidental injury or death. These conclusions are based on good evidence. Further evidence is also now becoming available to suggest that it is likely that there is a cumulative effect with the direct negative consequences of political violence added to the effects of economic and social disadvantage.

Part of this increased risk may be due, it has been suggested, to the fact that the whole family structure comes under stress as a result of political violence. In particular this may include changes in child-rearing tactics. Further, there is speculation that these changes may carry over to influence the way in which the next generation bring up their children. Unfortunately, despite the importance of this area, the amount of hard evidence is almost negligible. Only the speculation that changes in child-rearing styles accompany exposure to political violence has really been tested and even then only in a very limited way.

In many societies schooling is an important factor in the everyday life of children. Anecdotal evidence suggests that where schools can be kept open they may act as to protect children but this needs to be confirmed empirically. On the other hand, it is clear that political violence can and often does disrupt schooling, either because children do not attend school or because schools are forced to close. The evidence for this is reliable but is confined to developed societies.

5. *Child Combatants*

Children are not always passive victims of war and/or political violence but may be involved in several different ways. By far the largest of these consists of younger children who are involved at the periphery, often in such overt political activities as rioting or demonstrating. In some societies, however, for example in South Africa and Palestine, this form of activity involved large numbers of young people and their role became central to the political struggle. Older children are more likely to be involved in clandestine activities as members of a terrorist organization. A variety of hypotheses have been entertained as to why children become involved in this form of urban guerrilla warfare. These include the thrill of danger, extension of gang related activities common at this age, or the possibility that they are modeling themselves on adult behavior. A more worrying suggestion is that the role of adult involvement is more overt—the Godfather hypothesis.

This raises the question of the relationship of these children with their families: do families approve, encourage, admire, or do they disapprove? In terms of hard evidence we know very little about these or other questions. Basically we know that children are involved but we don't know why.

Membership of paramilitary groups or guerrilla armies probably involves smaller numbers of children and young people who, as noted above, are often slightly older than their stone throwing brother and sister. In this area it may be important to recognize two types of groups, anarchic/ideologues and nationalist/separatists. Probably more children and young people join the latter.

It is clear that it is not possible to fit these young people into any particular sociodemographic or psychological categories. In particular there is little evidence to support the common claim that guerrilla organizations attract a disproportionate number of psychopathic individuals. Nor is it likely that there is a particular 'terrorist type'—even with wide criteria such as action oriented, aggressive, and sensation seeking. Instead rewards may play a role such as prestige, glamour, excitement, and even material rewards. It is not clear however if these are the reasons people join in the first place or are these factors which sustain membership?

Rather than suggesting that individuals are motivated to become political activists because of certain psychological characteristics or because of the lure of rewards, it is hypothesized that children are socialized into this role in subtle ways, particularly via the indirect impact of institutions such as the family. Also there is the possibility that schools may influence children's political ideas in certain societies, especially through the teaching of history. Of course not all history is taught in schools and in more recent times media coverage of other political struggles has also been implicated in this socialization process as paradoxically have prisons. There is also some intriguing qualitative evidence that being a victim or witnessing another become a victim may play a role in stimulating young people to become politically involved.

Finally, the UN convention notwithstanding, it must be remembered that there are still many state-run armies around the world that include children in their ranks. In addition there are many young people, mostly boys (hence the term 'boy-soldiers') who are press-ganged into serving, usually with irregular forces.

As conflicts come to an end the fear is often expressed that child member of groups trained to kill or maim will not be able to be resocialized. Anecdotal evidence suggests that this is not necessarily true. One

possible explanation, which is yet to be tested empirically, is that a key element in membership of paramilitary groups may be the development of a morality of loyalty which in turn is related to the development of a relevant situated social identity.

This is obviously a very difficult to area in which to get hard evidence. The existing evidence therefore comes from atypical group members, is largely anecdotal, and has mostly been gathered by journalists rather than social scientists. It could be argued, however, that scholars could make more use of these as primary sources. Probably the only thing we can claim to have any firm evidence on is that children and young people who join nationalist/separatist guerrilla movements are not psychopaths and come from no particular social strata of society. Other hypotheses including ideas of 'terrorist personality types' or psychodynamic explanations will always be very hard to substantiate.

6. *Children and Peace*

One of the problems, it has been suggested, in bringing peace to societies that have experienced political violence is that the next generation will have either begun to believe that there is no future, or that they will be able to think of the future only in negative terms.

Given that we know little about the way in which children develop concepts of peace and war it could be argued that it is premature to try to educate children to be peace makers (see *Peace Promotion, Psychology of*). Despite this lack of basic knowledge this process has already begun in some societies and some people apparently believe it is effective. However, there is virtually no empirical evidence to substantiate their claim. A major problem would appear to be that too much of what passes for peace education focuses on interpersonal conflict as opposed to intergroup conflict. In future, curriculum designers need to produce a more effective peace education program and also to overcome the problem that peace education *per se* is not always politically acceptable.

It could also be argued that school-based peace education is always bound to be ineffective because it targets the wrong people in the wrong setting. For example, there is speculation, if not evidence, that children's ideas about peace and war may be more influenced by what they learn from their parents than from their schools. There is definitely evidence, which is now often forgotten, that learning about such things as peace and war involves emotions primarily and that providing facts may not alter these emotions. This is obviously an area which is in need of much more research which develops what is known and makes it amenable for use in applied settings.

The alternative to peace education which is most often advocated is bringing children together from opposing groups in order to foster positive intergroup attitudes—the contact hypothesis. This has been a well-researched area for many years and now boasts an extensive literature. What this literature indicates is that for contact to be even minimally effective it has to take place under highly prescribed conditions. However, what advocates of the contact hypothesis appear to be reluctant to accept is that while there is evidence that bringing groups together promotes interpersonal contact satisfactorily, it does not necessarily promote intergroup contact.

There is also good evidence that intergroup conflict can be reduced by manipulating the process of social categorization in order to alter group boundaries. This is a strategy that both social scientists and policy makers should consider more often. The evidence also suggests it is not possible to bring social categorization to an end entirely. Rather it is better to concentrate on altering the content of stereotypes or manipulating who gets put in which social category by altering intergroup boundaries.

See also: Coping across the Lifespan; Disasters, Coping with; Stress and Coping Theories; Violence and Effects on Children

Bibliography

Cairns E 1996 *Children and Political Violence*. Blackwell, Oxford, UK

Dawes A, Donald D 1994 *Childhood and Adversity: Psychological Perspectives from South African Research*. David Philip, Cape Town

Freud A, Burlingham D T 1943 *War and Children*. Medical War Books, New York

Garbarino J, Kostelny K, Dubrow N 1991 *No Place to be a Child*. Jossey-Bass, San Francisco

Leavitt L, Fox N 1993 *The Psychological Effects of War and Violence on Children*. Lawrence Erlbaum, Hillsdale, NJ

Machel G 1996 *Impact of Armed Conflict on Children*. United Nations Children's Fund and United Nations Department of Public Information, New York

Raviv A, Oppenheimer L, Bar-Tal D 1999 *How Children Understand War and Peace*. Jossey-Bass, San Francisco

E. Cairns

War, Sociology of

The sociology of war is a central topic in both political and historical sociology, since war is one of the most important policies states can pursue, and the outcomes of wars have often shaped both the formation and the dissolution of states. The literature on war is thus concerned with both its causes and its consequences.

Studies of the causes of war can be divided into three broad categories. The first type takes the system as a whole as the unit of analysis and focuses on how

characteristics of the interstate system affect the frequency of war. Debates focus on characteristics of the interstate system that are thought to increase or decrease war, such as global economic cycles, balances of power, and the increasing role of transnational organizations such as the United Nations. States are the unit of analysis in the second type, which explores the relationships among political, economic, and cultural features of particular states and their propensity to initiate wars. Social scientists disagree about the effects of political systems (democracy vs. autocracy) and economic systems (capitalist, socialist, or other) within states on war. The third type analyses war as an outcome of choices made by individual and small-group decision making. There is also no consensus on which model of individual decision making is most appropriate for the study of war. Is the decision to go to war based on a rational calculation of economic costs and benefits, or is it an irrational outcome of distortion in decision making in small groups and bureaucracies?

Theories of the consequences of war tend to focus either on its role in state formation, or on its causal impact on internal revolts and revolutions. Historical sociologists have shown that the frequency, duration, and timing of medieval and early modern warfare were the most important determinants of the size and structure of states (Tilly 1975, 1990, Ertman 1997). However, just as war can make states, it can break them too. For example, Skocpol (1979) argues that costly warfare often leads to fiscal crises and state breakdown, facilitating revolutions.

1. *Causes of War: The Interstate System*

Most studies of war that take the interstate system as the unit of analysis begin with assumptions from the 'realist' paradigm. States are seen as unitary actors, and their actions are explained in terms of structural characteristics of the system. The most important feature of the interstate system is that it is anarchic. Unlike politics within states, relations between states take place in a Hobbesian 'state of nature.' Since an anarchic system is one in which all states constantly face actual or potential threats, their main goal is security. Security can only be achieved in such a system by maintaining power. In realist theories, the distribution of power in the interstate system is the main determinant of the frequency of war.

Although all realist theories agree on the importance of power distribution in determining war, they disagree about which types of power distributions make war more likely. Balance-of-power theories (Morgenthau 1967) suggest that an equal distribution of power in the system facilitates peace and that unequal power distributions lead to war. They argue that parity deters all states from aggression and that an unequal power distribution will generally result in the strong using force against the weak. When one state begins to gain a preponderance of power in the system, a coalition of weaker states will form to maintain their security by blocking the further expansion of the powerful state. The coalitions that formed against Louis XIV, Napoleon and Hitler seem to fit this pattern.

Hegemonic stability theory (Gilpin 1981) suggests exactly the opposite, that unequal power in the system produces peace and that parity results in war. When one state has hegemony in the world system, it has both the incentive and the means to maintain order in the system. It is not necessary for the most powerful state to fight wars, since their objectives can be achieved in less costly ways, and it is not rational for other states to challenge a hegemon with overwhelming power. For example, the periods of British and US hegemony were relatively peaceful and World Wars I and II occurred during intervening periods in which power was more equally distributed. A related attempt to explain great-power war is power transition theory (Organski 1968). Power transition theory suggests that differential rates of economic growth create situations in which rising states rapidly catch up with the hegemonic state in the system, and that this change in relative power leads to war.

Debates about power transitions and hegemonic stability are of much more than theoretical interest in the contemporary world. Although the demise of the USSR has left the USA as an unchallenged military hegemon, its economic superiority is being challenged by the European Union and emerging Asian states (Japan in the short run, perhaps China in the long run). If power transition and hegemonic stability theories are correct, this shift of economic power could lead to great power wars in the near future.

Another ongoing debate about systemic causes of war concerns the effects of long cycles of economic expansion and contraction. Some scholars argue that economic contraction will increase war, since the increased scarcity of resources will lead to more conflict. Others have suggested the opposite: major wars will be more frequent during periods of economic expansion because only then will states have the resources necessary to fight. Goldstein's (1988) research suggests that economic expansion tends to increase the severity of great-power wars but that economic cycles have no effect on the frequency of war.

One significant change in the last half of the twentieth century which will require substantial revisions in realist systemic theories of war is the development and increasing power of transnational organizations (such as the United Nations), since their assumption that the interstate system is anarchical may no longer be valid. If the military power of the United Nations continues to grow, it could become more and more effective at preventing wars and suppressing them quickly when they do start. Of course, it remains to be seen whether powerful existing

states will choose to cede more power to such institutions.

Theoretical debates about the systemic causes of war have not been resolved, in part because the results of empirical research have been inconclusive. Each theory can point to specific cases that seem to fit its predictions, but each must also admit to many cases that it cannot explain. Part of the problem is that systemic theories have not incorporated causal factors at lower levels of analysis, such as internal economic and political characteristics of states. Since the effects of system-level factors on war are not direct but are always mediated by the internal political economy of states and the decisions made by individual leaders, complete theories of the causes of war must include these factors as well.

2. *Causes of War: Capitalism and Democracy*

One of the longest and most heated debates about the causes of war concerns the effects of capitalism. Beginning with Adam Smith ([1776]1976), liberal economists have argued that capitalism promotes peace. Marxists (Lenin [1917]1939), on the other hand, suggest that capitalism leads to frequent imperialist wars.

The Smithian liberal argument suggests that since capitalism has both increased the benefits of peace (by increasing productivity and trade) and the costs of war (by producing new and better instruments of destruction), it is no longer rational for states to wage war. The long period of relative peace that followed the triumph of capitalism in the nineteenth century and the two world wars that came after the rise of protectionist barriers to free trade are often cited in support of liberal economic theories (but the same facts can be explained by hegemonic stability theory as a consequence of the rise and decline of British hegemony).

In contrast, Marxists (Lenin [1917]1939) argue that economic problems inherent in advanced capitalist economies create incentives for war. First, the high productivity of industrial capitalism coupled with a limited home market due to the poverty of the working class result in a chronic problem of 'underconsumption.' Capitalists will thus seek imperial expansion to control new markets for their goods. Second, capitalists will fight imperialist wars to gain access to more raw materials and to find more profitable outlets for their capital. These pressures will lead first to wars between powerful capitalist states and weaker peripheral states, and next to wars between great powers over which of them will get to exploit the periphery.

With the increasing globalization of economics, and the transitions of more states to capitalist economies, the debates about the effects of capitalism, trade, and imperialism on war become increasingly significant. If Adam Smith is right, our future is likely to be more peaceful than our past; but if Marxist theorists are right, we may see economically based warfare on an unprecedented scale.

The form of government in a country may also determine how often it initiates wars. Kant ([1795]1949) argued that democratic states (with constitutions and separation of powers) will initiate wars less often than autocratic states. This conclusion follows from a simple analysis of who pays the costs of war and who gets the benefits. Since citizens are required to pay for war with high taxes and their lives, they will rarely support war initiation. Rulers of states, on the other hand, have much to gain from war and can pass on most of the costs to their subjects. Therefore, when decisions about war are made only by rulers (in autocracies), war will be frequent, and when citizens have more control of the decision (in democracies), peace will generally be the result.

Empirical research indicates that democratic states are less likely than nondemocratic states to initiate wars, but the relationship is not strong (Kiser et al. 1995). Perhaps one reason for the weakness of the relationship is that the assumption that citizens will oppose war initiation is not always correct. Many historical examples indicate that in at least some conditions citizens will support war even though it is not in their economic interests to do so, due to nationalism, religion, ethnicity, or other cultural factors.

Perhaps the most interesting current debate about democracy and war surrounds the proposition that democratic states never fight each other. There is clearly a strong empirical generalization to be explained, since all agree that democratic states rarely fight each other—depending on exactly how 'democracy' and 'war' are defined, some argue they never do (Weart 1998). However, scholars disagree about the causal mechanism responsible for this association. Some stress the role of political culture, arguing that the norms of toleration and mutual accommodation that prevent conflicts within democracies from escalating to violence have the same effect in limiting violent conflicts between democracies. These states consider each other part of the same 'ingroup' sharing the same values, and are thus very unlikely to fight. In contrast, others suggest that 'democratic peace' could be the result of rational self-interest. Democratic politicians may simply fear the negative impact that losing a war might have on their prospects for re-election. Further research, probably at the level of detailed case studies that can reveal precise causal mechanisms, will be necessary to resolve this debate.

3. *Causes of War: Decision Making*

Few theories of war focus on the individual level of analysis; their assumptions about individual decision making are usually implicit or undeveloped. Notable exceptions include rational choice theories (Bueno de Mesquita 1981, Bueno de Mesquita and Lalman 1992)

and arguments about organizational and small group decision making (Allison 1971).

Bueno de Mesquita begins by assuming that the decision to initiate war is made by a single dominant ruler who is a rational expected-utility maximizer. Rulers calculate the costs and benefits of initiating war, and the probability of victory, so wars will be initiated only when rulers expect a net gain from them. These assumptions generate several counterintuitive propositions. For example, common sense might suggest that states would fight their enemies and not their allies, but Bueno de Mesquita (1981) argues that war will be more common between allies than between enemies. Wars between allies are caused by actual or anticipated policy changes that threaten the existing relationship. The interventions of the USA in Latin America and of the USSR in Eastern Europe since World War II illustrate the process. Other counterintuitive propositions suggest that under some conditions a state may rationally choose to attack the stronger of two allied states instead of the weaker, and under some conditions it is rational for a state with no allies to initiate war against a stronger state with allies.

Other analyses of the decision to initiate war focus on how the social features of the decision-making process lead to deviations from rational choice. Allison (1971) argues that standard operating procedures and repertoires within states tend to limit the flexibility of decisions and make it difficult to respond adequately to novel situations. Others focus on the small groups within states (such as executives and their cabinet advisers) that actually make decisions about war. The cohesiveness of these small groups often leads to a striving for unanimity that prevents a full debate about options and produces a premature consensus. In spite of these promising studies, work on the deviations from rational choice is just beginning, and we are still far short of the general microlevel theoretical model of the decision to initiate war.

4. *Consequences of War: State Formation and Bureaucratization*

How and to what extent does war affect the formation and the structure of states? Tilly (1975, 1990) argues that 'war made states' in the early modern era. Warfare (along with the repayment of debts from past wars) cost far more than any other state policies in early modern Europe. Moreover, the timing and the nature of war shaped the structures of these developing states.

One of the most interesting ongoing debates concerns whether war facilitates or hinders bureaucratization. Weber ([1922]1978) argued that states involved in military competition with other states (e.g., Western Europe) would be more likely to bureaucratize than those that were more isolated (e.g., China, Japan). Although he does not use this terminology, he basically argues that states not facing the threat of war would be satisfied with existing administrative arrangements, whereas those competing militarily would be forced to adopt the more efficient bureaucratic form.

In contrast to this, Levi (1988) views war as a consistent impediment to bureaucratization. She argues that war raises the discount rates of rulers, causing them to pursue policies that provide immediate gains even if they are costly in the long term. Thus, rulers facing war would be unlikely to pay the high start-up costs of bureaucratization, but would instead do things like selling offices which would make bureaucracy much more difficult to implement.

Ertman (1997) suggests a related argument in which the timing of war is important. When states experienced early sustained warfare (prior to about 1450) they developed patrimonial administrations (due to lack of trained personnel and the dominance of 'cultural models' derived from feudal and Catholic institutions). These institutions were very difficult to bureaucratize due to the power of entrenched officials to block reform. Ertman argues that states that were able to avoid frequent war until later were able to develop more bureaucratic administrations.

Finally, wars that result in severe losses may facilitate bureaucratic reforms. One of the main barriers to bureaucratization is the entrenched officials in the state administration who have both the incentives and the power to block reform (Ertman 1997). These officials will not be dislodged by most wars, but their power will be broken by a severe loss at war, especially one that results in foreign occupation.

5. *Consequences of War: Revolt and Revolution*

Since the classic work of Simmel ([1908]1955, pp. 98–9) and Coser (1956, pp. 19, 95), the conventional wisdom has been that wars decrease the probability of revolts by increasing the internal cohesion of societies (the 'ingroup–outgroup' or 'conflict–cohesion' hypothesis). However, the results of empirical tests of the relationship between revolt and war have been mixed, at best.

In contrast to the 'conflict–cohesion' hypothesis, historical sociologists focusing on the early modern period (Tilly 1975, p. 74) argue that fighting wars increased the likelihood of revolts. Tilly argues that since subjects had few institutionalized mechanisms available to stop rulers from initiating wars contrary to their interests, the only option they had to try to limit war was revolt. In addition, Tilly (1975, p. 74) notes that since participating in wars weakens rulers, it increases the likelihood of a revolt.

Theda Skocpol (1979) has made the most compelling argument that war is a primary cause of major 'social revolutions.' Her argument about the origins of the French, Russian, and Chinese revolutions stresses the importance of factors that weaken the state,

making it more vulnerable to overthrow. The main factor that tends to weaken states is war, especially in the context of strong peasant communities and an alienated dominant class. This work has been the primary inspiration for the development of a 'state-centered' approach to historical sociology.

6. Conclusion

This short summary has only been able to scratch the surface of the voluminous literature on war. Future research should attempt to link the several topics discussed here, by bringing together the micro and macro causes of war, and by tying the causes more closely to the consequences of war. The increasing development of technologies of mass destruction and the rise of transnational political units will also challenge existing theoretical frameworks. Like most of sociology, the sociology of war is still in its infancy.

See also: Conflict and War, Archaeology of; Conflict: Anthropological Aspects; Conflict Sociology; Geopolitics; Military Geography; Military History; National Security Studies and War Potential of Nations; Peace; Peacemaking in History; War: Anthropological Aspects; War: Causes and Patterns

Bibliography

Allison G 1971 *Essence of Decision*. Little Brown, Boston

Bueno de Mesquita B 1981 *The War Trap*. Yale University Press, New Haven, CT

Bueno de Mesquita B, Lalman D 1992 *War and Reason*. Yale University Press, New Haven, CT

Coser L 1956 *The Functions of Social Conflict*. Free Press, Glencoe

Ertman T 1997 *Birth of the Leviathan*. Cambridge University Press, Cambridge, UK

Gilpin R 1981 *War and Change in World Politics*. Cambridge University Press, Cambridge, UK

Goldstein J 1988 *Long Cycles*. Yale University Press, New Haven, CT

Kant I [1795] 1949 Eternal peace. In: Friedrich C (ed.) *The Philosophy of Kant*. Modern Library, New York, pp. 430–76

Kiser E, Drass K, Brustein W 1995 Ruler autonomy and war in early modern Western Europe. *International Studies Quarterly* **39**: 109–38

Lenin V I [1917] 1939 *Imperialism*. International Publishers, New York

Levi M 1988 *Of Rule and Revenue*. University of California Press, Berkeley, CA

Morgenthau H 1967 *Politics Among Nations*. Alfred A. Knopf, New York

Organski J 1968 *World Politics*. Alfred A. Knopf, New York

Simmel G [1908] 1955 *Conflict and the Web of Group Affiliations*. Free Press, New York

Skocpol T 1979 *States and Social Revolutions*. Cambridge University Press, Cambridge, UK

Smith A [1776] 1976 *The Wealth of Nations*. University of Chicago Press, Chicago

Tilly C (ed.) 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ

Tilly C 1990 *Coercion, Capital, and European States, 990–1990*. Blackwell, Oxford

Weart S 1998 *Never at War: Why Democracies Will Not Fight One Another*. Yale Universtiy Press, New Haven, CT

Weber M [1922] 1978 *Economy and Society*. University of California Press, Berkeley, CA

E. Kiser

Warfare in History

This article divides the history of warfare into five periods: the classical era (including the wars of Greece and Rome); the Middle Ages (roughly from AD 500 to the Turkish conquest of Byzantium in 1453); the age of gunpowder (from 1453 to the American and French Revolutions beginning in 1776); the 'long nineteenth century' (from 1776 to 1918); and finally the contemporary period (covering World War II and the Cold War). Each of these periods will be defined by changes in three dimensions. First and most obviously, each period is associated with a general group of military technologies. Second, each period also involves a shift in the nature of military organization. Finally, this periodization demonstrates changes in the reasons why people have gone to war.

1. The Classical Period

1.1 Technology

The Classical period deserves consideration on its own merits and in the context of an historical debate over the influence of Greco-Roman warfare on later European patterns. In the field of technology, some historians argue that the Greeks and Romans began a pattern of using superior technology to compensate for having to fight wars against numerically superior enemies. The Persian Wars of the fifth century BC demonstrated the need for Greek city-states to combat the much larger armies of Persian kings Darius and Xerxes. Needing to slow the Persians, Spartan general Leonidas threw his much smaller force into a mountain pass near Thermopylae in 480 BC knowing that it would be destroyed to a man. Aware that his army would not return, Leonidas took with him only those men who had children to succeed them, an important consideration because the army and the citizenry were closely linked.

To avoid future Thermoplyaes, the Greeks developed both offensive and defensive technologies. Eventually, the typical Greek soldier wore nearly seventy pounds of armor and carried spears and swords that

were superior to Persian weapons. These technological solutions helped the Greeks win, but the unity that the Greek city-states showed against Persia soon devolved into internecine wars. During the most famous of these wars, the Peloponnesian War (431–04 BC), technology again proved critical. In this case, Athenian naval vessels attempted to balance the superior land power of their enemy, Sparta. Defensive technologies evolved in counterpoint as Athens, Syracuse, and other city-states built protective walls strong enough to defy even determined sieges.

The Romans also leaned heavily on technology to defeat more numerous enemies, though many Roman military technologies were not necessarily weapons systems. Building upon the Roman genius for engineering, Roman armies assisted in the construction of a system of roads (even today the Roman roads are roughly contiguous with the European highway system) that permitted the legions to move quickly from one point in the empire to another. The Romans also depended upon defensive works such as trenches and walls (Hadrian's Wall in the UK is the most famous example) to contain their enemies rather than fight them. Some historians argue that the Classical tradition bequeathed to later Western militaries the importance of looking for technological solutions to military problems.

1.2 Organization

The Greeks based their military upon the hoplite system. Hoplites (named for their shield, the hoplon) were citizens of the republic and rough social equals. Thus, Greek warfare had no concept directly analogous to the modern division of officers and enlisted men. Greek generals (such as they were) received their appointment by yearly election. Municipal assemblies also voted on when and where the city-state would go to war. As citizens, hoplites participated fully in these elections.

As nonprofessionals, the hoplites needed a fairly simple tactical system because so little time could be devoted to formal training. The result was a dense concentration called a phalanx. A hoplite's shield protected his own left side and his neighbor's right side. Thus the phalanx could only succeed when every hoplite stood his ground, lest the gap between himself and his neighbor be exploited. Once broken apart, a phalanx soon dissolved, resulting in many casualties by trampling.

The Romans trained and organized much more formally than the Greeks. Groups of about 120 men formed 'maniples,' thirty-five of which were then used to form legions of approximately 4,000 infantry and 300 cavalry (the exact sizes of maniples and legions varied significantly over time). This system foreshadowed the modern creation of platoons, companies, and brigades. The Romans used professional gladiators (named for their short, thrusting sword, the gladius) and evolved a concept of officers (usually wealthier citizens) and 'other ranks.' Increased sophistication and professional training allowed the Romans to develop a military that used various weapons, including soft iron javelins called pila. The soft iron bent on impact so that enemies could not throw them back. If victorious, however, Roman smiths could recover the pila after the battle and reshape them. The Roman system gave the legions flexibility, endurance, and striking power that few militaries could match.

The Greek and Roman systems of warfare mirrored their societies. Both preferred service by citizens, perhaps another legacy to modern nation-states. The Roman system extended citizenship to conquered people, but the effectiveness of its army noticeably declined as more non-Roman militia entered the Roman military. The citizen model was less prevalent in the armies of autocrats such as Alexander the Great. The Romans also gradually separated wealth from military service by introducing regular pay and even retirement benefits.

1.3 Purposes

Among the legacies left by the Greeks and Romans to later styles of warfare is the concept of pitched battle. The Greeks especially tended to favor violent, but short, battles intended to solve disputes in a single afternoon. The bloody nature of Greek warfare appalled many of their opponents, including the Persians, who disliked a style of warfare that left even the winners with high casualties. Despite its violence, Greek warfare did follow certain understood rules, including safe passage for heralds and ambassadors and amnesty for temples and other religious sites.

Roman warfare followed a similar pattern. Wars were generally fought with the single goal of annihilating the enemy. The Roman sack of Carthage in 146 BC is the most famous, but hardly the only example. Roman armies typically pursued retreating enemies with cavalry and often treated prisoners severely. In one case, Julius Caesar kept an opposing general on display in Rome for six years before executing him. The Roman preference for fast, deadly, decisive battles differed significantly from other cultures. The Aztecs, for example, believed that the purpose of warfare was to capture prisoners for later sacrifice.

2. The Middle Ages

2.1 Technology

The most important Medieval technology, odd though it may seem, was not a weapon, but the stirrup. The stirrup changed the traditional relationship between

cavalry and infantry. With stirrups allowing riders to grasp lances and shields firmly and with newer, more durable horseshoes allowing horses to run faster and farther, cavalry came to the forefront for most of this period. The Mongols proved to be the most successful of the horse-based military systems of the Asian steppes; by the late thirteenth century they controlled an empire that stretched from China to modern-day Ukraine. Their military relied almost exclusively on cavalry.

In Western Europe, cavalry proved to be no less important. For the aristocrats who could afford the extensive support systems that accompanied cavalry, horses meant power. The European aristocracy, as well as its Japanese counterpart, depended upon horses, armor, and swords for offensive power and increasingly larger castles for defensive power. The victory of mounted Norman knights over Anglo-Saxon infantry at the Battle of Hastings in 1066 underscored the value of cavalry. The Catholic Church even tried to ban weapons (such as the crossbow) that threatened the knight's status, though the church permitted the use of such weapons on infidels.

Technological changes also helped to end the period of dominance by the mounted knight. The Hundred Years War marks an important turning point. At the battles of Crécy (1346), Poitiers (1356), and Agincourt (1415), skilled English archers devastated French knights with the Welsh longbow, a weapon capable of rapidly delivering arrows powerful enough to penetrate armor at ranges up to 400 yards. Lacking effective missile weapons themselves, the French knights suffered terribly. At Crécy, the longbows allowed the English to defeat a French force three times its size, further underscoring the Western pattern of looking to technology to compensate for smaller numbers.

2.2 *Organization*

Military training and military service was concentrated in a relatively small group of wealthy aristocrats. In most Eurasian systems, peasants and serfs were not expected to engage in actual fighting, though they might participate in ancillary operations, particularly as archers or pikemen supporting the cavalry. The challenge to kings and other rulers, then, was to maintain the loyalty and fealty of their aristocratic vassals. In the Third Crusade (1189–92), for example, Frederick I of Germany, Philip II of France, and Richard I of England spent considerable time keeping their unruly knights and barons in check. Powerful, aristocratic military orders, relatively independent of royal authority such as the Knights Templars and the Teutonic Knights emerged from the Crusades as well.

In Japan a similar system developed, though loyalty was ordinarily based more on personal than contractual allegiances. The Japanese samurai class emerged in the twelfth century, replacing the conscript armies that had existed previously. The samurai also came to dominate Japanese politics, supplanting the traditional Japanese aristocracy. The failure of Mongol invasions of Japan in 1274 and 1281 removed any serious outside threat to the home islands, leaving the small samurai class in near-complete charge of military affairs. Civil wars and power struggles did occur, but they rarely drew Japanese peasants into the fighting.

In the Islamic world, the Abbasid Caliphate began in the ninth century to use slave soldiers called mamluks. Originally from the steppes of Asia Minor, the mamluks soon translated their military dominance into political dominance in Egypt and elsewhere in the Middle East. They were the only major Islamic force that proved capable of resisting the Mongols, defeating them in 1260 at the Battle of Ain Jalut in modern-day Israel. The mamluks, the samurai, the Mongols, and the European aristocracy, though different in many ways, shared a near-monopoly over military force in their societies, which were arranged in feudal, or at least fief-based, political systems. All of these systems were prone to succession crises and internal strife. None proved to be militarily superior to any other.

2.3 *Purposes*

Generally speaking, the motivations and purposes for warfare in this period were either religious, feudal, or both. The Islamic empires acquired tremendous territory, in part through military conquest, without significant technological advantages over their opponents. They did so by wedding religious ideology to their military forces. Under the reigns of Caliphs Umar (634–44) and Uthman (644–56), Muslim armies conquered Spain, North Africa, Persia, and the Indus River valley. Their fleets conquered Cyprus and raided Sicily and Rhodes. Qur'an readers accompanied Muslim armies, converting many men who may have joined more for plunder than for religion.

The Crusades (1095–1291) also reflect a religious basis for warfare in this period. In the eleventh century, the success of the Seljuk Turks against the Byzantines led Pope Urban II to call the First Crusade (1095–99) to defend Christianity against the incursions of Islam. The Crusades perfectly show the interplay of religion and feudal obligation. The Pope acted both as a religious leader and (more controversially) as a noble in his own right. The First Crusade (also known as the 'Peasants' Crusade') demonstrated that only knights could be relied upon to form the center of Christian armies. The peasant armies were too large to feed and too unsophisticated to stand up to the Turks. Of the 30,000 non-nobles who fought in the First Crusade, 25,000 died. Sending young nobles off to the Middle East also served a political purpose for the Pope,

removing warriors and weapons from the constant internal struggles of feudal Europe. Thus by the fifteenth century similar and roughly coequal military systems existed in Western Europe, the Islamic Empires, the Asian steppes, and Japan.

3. The Gunpowder Age

3.1 Technology

The Turkish capture of Constantinople in 1453 involved 70 heavy artillery pieces, including a 19-ton piece that fired a 1,500-pound projectile more than a mile. The appearance of these weapons was not new. Knowledge of gunpowder had moved from China to the Middle East to Europe. At the Battle of Crécy (1346) the English introduced very primitive gunpowder weapons to European warfare. Nevertheless, the dramatic Turkish success at a place Europeans considered sacred marked a watershed in military technology. Ironically, the Islamic empires subsequently proved slower to adopt gunpowder weapons than did their European adversaries. By the end of the eighteenth century, Europeans were the clear masters of gunpowder warfare on land and at sea.

Early small arms, like the arquebus, were clumsy, unreliable, and often more dangerous to the shooter than to his intended victim. Until the widespread adoption of the rifle in the nineteenth century the common command given to musketeers was 'level' rather than 'aim' since one could not expect to hit a specific target at any significant distance. Large artillery was even more dangerous to their users and so large that pieces often had to be cast at the siege site. By the sixteenth century Europeans had standardized ammunition sizes, developed training systems for musketeers, and invented more reliable firing systems for small arms. These changes made both large and small arms more mobile and effective.

Europeans developed gunpowder weapons much more quickly than any other society in this period. The simple availability of raw materials and craftsmen were important prerequisites; men who could cast church bells were in special demand as they could easily cast cannon as well. European nobles also had ample motivation to acquire new weapons for their struggles with other nobles. Finally, the European system proved more amenable to allowing private enterprises to develop and sell new weapons. The mamluks and the samurai showed a particular dislike for guns as they tended to undermine the military monopoly that kept them in power.

3.2 Organization

Guns threatened the very existence of a horse-based nobility whose expertise was in swords and lances. The age of the knight, already waning, came to an end. Non-noble musketeers replaced them. Because small arms were so inefficient when used singly, musketeers had to be arranged in lines and taught to fire and load their weapons in uson. Armies thus became much larger and more disciplined. Aristocrats who remained in military service became the backbone of an officer class that led armies mostly constituted of peasants.

Holland's Maurice of Orange (1567–1625) and Sweden's Gustavus Adolphus (1594–1632) deserve much of the credit for forming and training the new armies. They introduced regular drill and other reforms that eventually led to the classic European linear infantry formations (initially as many as 10 men deep, later two or three deep) that permitted one line to fire, then move to the back. This system maintained a near-continuous rate of fire, but required intense training (the first drill manuals appeared in 1607) and discipline in order to assure the safe and effective operation of guns. Every officer, furthermore, feared the specter of mutiny. Intense discipline, therefore, characterized most European armies; in many cases, the men feared their officers more than they feared the enemy.

Many European nations also developed powerful sail-based navies in this period. England's Henry VIII, excommunicated from the Catholic Church in 1534, sold Church lands to finance the construction of a fleet. Future monarchs continued to build upon Britain's command of the seas, introducing men-of-war in the seventeenth century that were durable enough to carry large cannon. England's *HMS Victory*, launched in 1759, was built from 2,000 oak trees, 27 miles of rope, and four acres of sails. It had 104 guns capable of firing 1,100 pounds of iron every 90 seconds. Seapower became critical to acquiring colonies, protecting merchant marines, and projecting power. Britain dominated the 'Great Age of Sail,' but Portugal, Spain, Holland, France, and others used naval vessels to increase their wealth and status.

3.3 Purposes

Three purposes dominated warfare: religion; dynastic ambition; and imperialism. The Thirty Years' War (1618–48) demonstrated the interplay of the first two. Lutheran-Catholic-Calvinist antagonisms partially explain the bloody nature of the war; Germany proportionately lost more people in the Thirty Years' War than in World War II. Religion, however, did not always determine the allegiances of the participants. Catholic France and Catholic Spain, for example, consistently fought one another for control of Italy and Flanders. Gunpowder weapons also led to a concentration of power as wealthy kings, in command of increasingly larger and more disciplined armies (not to mention artillery able to destroy castles), could compel their restless vassals to accept their rule.

The growth of monarchy meant that warfare often followed the dynastic goals of kings. Louis XIV in

France, Peter the Great in Russia, and Frederick the Great in Prussia all used gunpowder-based armies to extend their power inside their kingdoms and enforce their will outside. Often, dynastic wars spilled over into the empires Europeans were in the process of building with those same gunpowder weapons. In the Seven Years' War (1756–63), Britain seized France's Canadian and Indian colonies.

Finally, of course, Europeans translated their gunpowder weapon advantages into overseas empires. In this period, their greatest military advantage over non-Europeans was in areas accessible to their great warships. Navies allowed Europeans to enforce their will along the coastlines of Africa and Asia. On land, guns helped to make the Spanish, French, and British empires in the Americas possible. In Asia, 'gunpowder empires' such as the Ottomans and the Mughals also concentrated political power. Japan, after an initial period of experimentation with guns, effectively banned gunpowder weapons in 1587 because of the threat they posed to samurai dominance.

4. *The Long Nineteenth Century, 1776–1918*

4.1 *Technology*

This period, dominated by nationalism and industrialization, marked a tremendous advance in military technologies. Many existing technologies became much more sophisticated. Smoothbore muskets gave way in the nineteenth century to rifles, which have a groove cut in their barrels. Rifling and the conical bullet vastly increased the range and accuracy of small arms fire. After the American Civil War (1861–65) various forms of repeating rifles and machine guns emerged. Artillery also became more deadly. By 1918 the Germans had developed a gun (albeit an inaccurate one) with a range of 80 miles. Naval vessels became more deadly as well. Iron and steel replaced wood as the primary construction material and oil replaced coal as the main fuel. *HMS Dreadnought*, launched in 1906, revolutionized warfare by carrying ten 12-inch guns. *Dreadnought* set off a naval arms race that helped to cause World War I. Winston Churchill noted the importance of Dreadnought-class ships to his nation thus: 'the Admiralty asked for six, the Cabinet proposed four, and we compromised on eight.'

This period also introduced many new types of weapons systems, including (in rough chronological order), submarines, airplanes, poison gas, and tanks. By the end of this period air forces had developed all major roles for airpower in use today except midair refueling. These technologies brought war directly to civilians as never before. Many important communications technologies such as telegraphs (first used in the Crimean War, 1854–56), railroads, and radio changed the nature of military command and control. Advances in medical technology reduced the number of men who died of wounds and disease.

These technological changes affected Western and Westernized societies (including the USA and, after 1853, Japan) the most. Warfare between them, therefore, reached unprecedented levels of violence. Westernized societies also had tremendous advantages against non-westernized societies. European imperialists were thus able to use new military technologies to move inland in Africa and Asia. Most visibly, Japan, which Westernized very quickly after 1853, easily defeated China, which did not, in the Sino-Japanese War of 1894–95.

4.2 *Organization*

The most obvious change in the organization of armies and navies in this period was the increasing connection between military service and nationalism. National volunteers and conscripts replaced mercenaries, once common in Europe. Between 1776 and 1783, the American colonies used a militia with both local and national roots to defeat a British army partly dependent upon mercenaries. This contrast partially explains the fame of George Washington's crossing of the Delaware River in 1776: his national army captured 1,000 Hessian mercenaries in British service.

The American experience, though, is minor compared to the French Revolution. The logic of the Revolution and the wars that followed implied that national armies must be populated by national soldiers. The *levée en masse* of 1793 established the principle that citizens of all ages, male and female, bore some responsibility for national defense. Napoleon extended this principle and most European armies soon copied the French pattern. Nationalism inspired zeal and a willingness to endure. In 1814 Prussia created the *Landwehr* as a permanent peacetime militia and by the end of the nineteenth century most European nations had introduced national conscription. The militaries of states that were not also nations (Austria-Hungary, the Ottoman Empire) declined as a result.

The two most important military reforms of this period were Napoleon's corps system and the Prussian General Staff. A Napoleonic corps was large enough to operate independently since it contained formations of cavalry, artillery, and infantry that could act in concert. This system, however, depended on capable officers. Napoleon therefore introduced a professional officer corps open to talent in place of the aristocratic officer corps upon which Europe had come to depend. Later in the century, Prussia developed a General Staff system to plan wars. Detailed Prussian planning led to victory in the Wars of German Unification (1864 against Denmark, 1866 against Austria, and 1870–71 against France). These successes led all major Western militaries to copy the general staff idea.

4.3 Purposes

Military historians sometimes speak of a continuum between 'limited' and 'total' wars. Because warfare in this period tended to be tied to national goals and national identity, this period represented a general shift toward totality. The Wars of the French Revolution, the American Civil War, and World War I stand as cases in point. In these wars and others, huge proportions of national income were dedicated to prosecuting wars and fundamental social, political, economic, and cultural patterns were forever altered. The increased suffering (and intentional targeting) of civilians also led to totality.

Advanced technologies also made imperialism (often as an extension of nationalism) cheaper and easier for the Europeans. European militaries moved inland with steam-powered ships and railroads. Machine guns, furthermore, made military contests with natives one-sided. By 1912, the Vickers-Maxim gun could fire 250 rounds per minute. As poet Halaire Belloc noted in a turn-of-the-century (twentieth century) poem, 'Whatever happens, we have got/The Maxim Gun, and they have not.' Nevertheless, isolated cases such as the Boer Wars (1899–1902) demonstrated that modern weapons in the hands of skilled practitioners could make imperialism extremely bloody and costly. The Boers, white settlers in South Africa of Dutch ancestry, resisted the British with breech-loading guns and smokeless powder. It took the British empire more than two and a half years and almost 500,000 men to subdue less than 90,000 loosely organized Boers. The Boer Wars provided a taste of the changes to come in European military dominance after World War II.

World War I caused 20,000,000 deaths, destroyed four empires, and left in place the conditions for World War II. Only wars fought by national armies could have been so destructive. Warfare in this period combined the productive capacity of industrialization with the inspiration of nationalism. This combination, when matched with the classical legacy for pitched battle, produced a level of carnage that continues to baffle historians today.

5. The Contemporary Period, 1918–2000

5.1 Technology

Although this period has much in common with the previous period, one specific technology divides it from all that had come before. Atomic weapons have changed the very definition of warfare. Some scholars have argued that in the atomic age 'strategy,' as the term was commonly understood before 1945, has no real meaning because atomic weapons do not permit a society to apply a level of force proportionate to desired political ends. In effect, a society only has the military option of obliterating one's enemy and, in the process, taking the risk that a similar fate might befall your own society. Atomic weapons created the post-World War II superpower system that so informed not only the Cold War but dozens of regional conflicts as well. The challenge of fighting wars 'under the nuclear umbrella' led to American defeat in Vietnam, Soviet defeat in Afghanistan, and 'war by proxy' as the superpowers tried to fight each other through client states in areas as diverse as Southeast Asia, the Middle East, sub-Saharan Africa, and Latin America.

Of course, more than just atomic weapons are important to this period. World War II saw the development of effective long-range bombers and aircraft carriers, capable of bringing war directly into an enemy's homeland. Radar and sonar provided new technologies to deal with enemy air and submarine attacks. By the end of the war, jet aircraft and rockets were operational, though neither proved decisive. More sophisticated transport planes and landing craft allowed for airborne and amphibious warfare to come into maturity as well.

Since World War II, electronics and computers have revolutionized warfare yet again. Cruise missiles and other 'smart' weapons ideally provide more accurate delivery of weapons and, hopefully, fewer civilian casualties as a result. Satellites, helicopters, sophisticated avionics, and stealth technology once again have provided an upper hand to advanced societies, though the Vietnam and Afghanistan cases prove that a technological advantage is not always enough to ensure victory. Nevertheless, today Western societies still depend upon their technological advantages—much as the ancient Greeks did—to compensate for the smaller numbers of soldiers that they can place on the battlefield.

5.2 Organization

Initially, the organization of World War II armies was designed to avoid the totality of World War I. The German *blitzkrieg* system attempted to win 'lightning' victories that would not drain German manpower in the trenches. The Japanese, too, hoped to destroy the American fleet at Pearl Harbor and force the USA into a negotiated settlement. Nevertheless, totality soon emerged as the Axis powers sought to annihilate, not defeat, their enemies. More than ever before, entire societies became mobilized. Women participated directly and indirectly in numbers and ways never before envisioned. The pattern of increasing female participation has continued into the post-war period. The USA armed forces used 35,000 women in the Persian Gulf War, many of them at or very near the front lines.

The totality of World War II ruined Europe and Japan as first-rate military powers. The end of European hegemony also meant the virtual end of imperialism. Many militaries resorted to guerilla warfare to

try to defeat more technologically sophisticated enemies. Fighting underneath the 'technology threshold' has proven difficult for more advanced military systems. The Americans had great difficulty with the Vietnamese but few battlefield problems with the Iraqis in 1991 because the latter's military was much more like their own that the former's.

Since the end of the Cold War, the general trend has been to much smaller, but more sophisticated, militaries. Virtually all European states have eliminated or greatly reduced conscription. The USA abandoned conscription during the Cold War (in 1973) partially as a result of Vietnam War protest, but also to acknowledge that sophisticated weapons systems require dedicated professional volunteers to operate them.

5.3 Purposes

World War II was a contest of three ideologies: democratic capitalism, fascism, and communism. As a result, armies became politicized as never before. Young German males often left Hitler Youth organizations and entered directly into the Wehrmacht. In the Soviet Union, political commissars (including Nikita Khrushchev) served with Soviet military units and had as much decision-making authority as generals. Ideology created what John Dower called a 'war without mercy.' Though he was referring to the Pacific War between the USA and Japan, the same phrase could be applied equally to the war between Germany and the Soviet Union as well. The Holocaust and the German siege of Stalingrad, among other episodes, show the cost of the war to civilians. The former killed 6,000,000 Jews and other targets of the Nazi state. In the latter, Joseph Stalin refused to order the evacuation of the city even in its most dire moments. In the end, only nine Stalingrad children were reunited with both biological parents.

The Cold War interacted with the general decline of European hegemony to impact regional conflicts. Throughout the Third World, nationalist movements and civil wars became arenas where the superpowers armed clients but, conversely, often went to great lengths to reduce global tensions. In 1956 and 1973, for example, the superpowers brokered truces in the Arab–Israeli wars rather than risk a USA–Soviet confrontation. Regional wars like the Iran–Iraq War (1980–88) also became entangled with the Cold War; the weapons that the superpowers made available lengthened wars that had little relation to the global capitalist–communist struggle.

The end of the Cold War raises as many questions as it provides answers. The Persian Gulf War seemed to set the model for American hegemony as the lone remaining superpower. After building a coalition based around itself, the USA operated in the Gulf without the fears of Soviet response that had haunted Korea and Vietnam. The NATO air war over Kosovo, however, suggested that Russia (and perhaps China) could not be ignored. Multipolarity appears to be every bit as dangerous as bipolarity. The Western powers continue to look toward technological solutions (especially airplanes and cruise missiles) to solve military problems, suggesting that in some ways we have not come so far from the Greeks after all.

See also: Arms Control; Contemporary History; First World War, The; Folklore; Geopolitics; History of Technology; Home in Anthropology; Imperialism, History of; International Relations, History of; Military Geography; Military History; National Security Studies and War Potential of Nations; Peace; Peacemaking in History; Revolutions, History of; Sacrifice; Second World War, The; Violence, History of; War: Anthropological Aspects; War: Causes and Patterns; War, Sociology of

Bibliography

Black J 1998 *War and the World: Military Power and the Fate of Continents, 1450–2000*. Yale University Press, New Haven, CT

Hanson V D 1989 *The Western Way of War: Infantry Battle in Classical Greece*. Knopf, New York

Howard M 1976 *War in European History*. Oxford University Press, Oxford, UK

Keegan J 1993 *A History of Warfare*. Knopf, New York

Keegan J 1976 *The Face of Battle: A Study of Agincourt, Waterloo, and the Somme*. Viking, London

Keeley L H 1996 *War Before Civilization: The Myth of the Peaceful Savage*. Oxford University Press, Oxford, UK

Keen M (ed.) 1999 *Medieval Warfare*. Oxford University Press, Oxford, UK

McNeill W H 1983 *The Pursuit of Power: Technology, Armed Force and Society Since 1000*. Oxford University Press, Oxford, UK

Overy R 1995 *Why the Allies Won*. Norton, New York

Paret P (ed.) 1986 *Makers of Modern Strategy from Machiavelli to the Nuclear Age*. Princeton University Press, Princeton, NJ

Parker G (ed.) 1997 *The Thirty Years War*. Routledge, New York

Strachan H (ed.) 1998 *World War I, A History*. Oxford University Press, Oxford, UK

M. S. Neiberg

Water Pollution: Marine

This article will briefly define the term marine pollution and then go on to consider the historic growth of problems in this area and the perception of those problems. The main areas of concern in terms of marine pollution are then outlined and the international methods of regulating marine pollution problems are considered.

1. Introduction

The term pollution is used widely and rather loosely. In the context of the oceans the UN sponsored Joint Group of Experts on the Scientific Aspects of Marine Pollution (GESAMP) offered the following definitions (GESAMP 1982):

> **Contamination** is the presence of elevated concentrations of substances in water sediments and organisms.
> **Pollution** is the introduction by man, directly or indirectly, of substances or energy into the marine environment (including estuaries) resulting in deleterious effects such as harm to living resources, hazards to human health, hindrance of marine activities, including fishing, impairing quality for use of seawater and reduction of marine amenities.

These definitions are extremely valuable because they draw an important distinction between human-induced changes in the concentrations of a substance somewhere in the marine environment (contamination) and actual damage resulting from this introduction (pollution). The oceans in total are vast and have a huge capacity to dilute, and in some cases break down, a discharge if time and physical processes allow this dilution and degradation capacity to be utilized. Acknowledgment of this capacity should not imply complacency about problems arising from human impacts on the ocean, but rather help focus concern on the most important issues.

2. Historical Developments

Major public health problems in developing cities in the Middle Ages led to the construction of primitive sewers to discharge wastes to local estuaries. Such discharge of wastes improved public health but caused pollution in estuaries such as the Thames in London from the fourteenth century onwards, promoting the earliest development of sewage treatment systems. However, until the twentieth century technological development was such that only the immediate coastal environment was significantly contaminated by human activity. Despite this, the tragedy of the decimation of the great marine mammals by hunting demonstrates major and global human impacts on the marine environment from the eighteenth century onwards.

Concern over marine pollution grew from the 1950s onwards as it became clear that the ability of the marine environment to absorb discharges was not infinite and that real damage was being done by human activity. Incidents such as the mercury poisoning at Minamata in Japan, the global contamination of the oceans by radioactive debris from bomb testing and by persistent pesticides such as DDT, and the effects of the first large-scale oil spills from sinking of the new giant oil tankers changed for ever the public perception of the vulnerability of the oceans to damage from human activity.

Investigations into these incidents also began to reveal the complexity of marine contamination and pollution issues. The saga of DDT illustrates that some contaminants can very effectively become globally dispersed and furthermore can be concentrated up the food chain to cause major ecological damage far from source to nontarget organisms.

By contrast, while the effects of an oil spill are locally catastrophic in the short term, the bacteria in the marine environment have the capacity to degrade such hydrocarbon material and hence the problem is relatively localized and short term. It is also now known that the biggest sources of hydrocarbons to the marine environment are not the rare catastrophic oil spills, but rather the chronic low-level operational discharges of oil, although the impacts of the two different sorts of discharge are very different. These examples illustrate that contaminants of the marine environment can be broadly classified into several groups.

Nondegradable. Elements such as lead and cadmium cannot be degraded, although they may become stored in sediments and largely removed from biological interactions.

Slowly degraded. Some organic compounds such as DDT and PCBs are only very slowly degraded over timescales of decades or longer. Radioactive material introduced to the marine environment decays over timescales controlled by radioactive half-lives which range from seconds to millions of years. In the case of both persistent organic and radioactive contaminants, intermediate decay products can also be toxic. As with nondegradable material, these compounds may be buried in sediments and effectively taken out of the system long before they are fully degraded.

Readily degraded. The marine environment is a natural biological system containing a wealth of micro-organisms able to break down many organic compounds including substances such as sewage and oil. As noted later, however, the breakdown process itself and the products of that breakdown can cause serious environmental problems.

Since the 1950s there has been enormous growth in the scientific understanding of how the oceans function as a physical, chemical, and biological system. This has allowed a much better assessment of the local and global scale of marine contamination and pollution problems. However, it was still not until the early 1970s that the natural distributions of some important contaminants, such as cadmium, in the oceans were described. A basic understanding of the natural distribution of an element is clearly a vital prerequisite to assessing the likely contamination of the oceans from human impacts. In the case of entirely manmade substances such as DDT or artificial radionuclides, the situation is rather different since their detection in the

environment at any level implies contamination, although not necessarily pollution.

In recognition of the growing impact of human activity on the oceans, discharge controls on degradable and nondegradable materials have been introduced in many countries. These aim to minimize harmful impacts, based on the best available scientific understanding of likely effects. There is clear evidence that these controls have had beneficial effects in many areas. The resulting declines in levels of lead and some radionuclides in the oceans, for instance, have been documented. However, the pollution arising from the use of tributyl tin (TBT) as an antifouling agent in the 1980s provides a warning that our control procedures are still far from being entirely effective. Every year hundreds of new organic compounds are synthesized and used in industries around the world. The effects of the inevitable resultant discharge are rarely known with certainty and there is growing concern over their possible effects. The issue of possible estrogenic effects of chemicals discharged to the environment represents an example of a new and disturbing pollution problem that we are a long way from understanding.

3. *Current Concerns*

A series of reviews of the marine environment have discussed the changing nature of pollution threats to the oceans and also suggested a broad consensus on the nature of current threats to the oceans (McIntyre 1995, Goldberg 1995, Huber 1999). The major issues of current and likely future concern center on the effects of sewage and nutrient discharge, and the release of organic compounds including pesticides. The issue of marine litter is also still of concern, with evidence of damage to bird and marine life from the indiscriminate discharge particularly of plastic wastes from ships and the loss of drift nets. Regulations to reduce these problems exist, but policing such regulations on the high seas is almost impossible.

The global population is growing and the proportion of that population living near the sea is also growing. The waste discharges from that population will inevitably enter the marine environment, although this will hopefully be after basic treatment. Furthermore, the agricultural and industrial activity to sustain that population will inevitably result in discharges to the marine environment, even if these are minimized by the use of the best available technology. Thus it is the insidious effects of this population growth that may pose the biggest threat to the oceans.

3.1 *Sewage*

Sewage wastes contain many harmful micro-organisms. The public health effects of coming into contact with these organisms either directly, via bathing, or in foodstuffs such as shellfish, where they can become concentrated, is increasingly being recognized. Sewage wastes are rather quickly broken down by natural micro-organisms in the marine environment to release nutrients such as nitrogen and phosphorus. However, the breakdown of organic wastes such as sewage requires oxygen, or some other oxidizing agent. If the rate of breakdown of waste and associated oxygen consumption exceeds the rate at which oxygen is supplied, oxygen concentrations in the water will fall, leading, in the extreme case, to oxygen depletion and the loss of almost all animals as well as the production of noxious odors. This happened in the Thames estuary as London's population grew and overwhelmed primitive waste treatment facilities. Improvements in waste treatment since the 1950s have restored the health of the Thames. Similar stories of damage from waste discharge and subsequent recovery with improved waste treatment can be told for industrialized estuaries around the world. They illustrate that we have the technology to mitigate the effects of sewage discharges and to reverse the damage done.

3.2 *Nutrients*

Human perturbations have approximately doubled the fluxes of the nutrients nitrogen and phosphorus through the environment, while the perturbation for the nutrient silicon is much less. Thus fluxes and the relative proportions of the nutrients have changed as a result of human activity. Nitrogen and phosphorus fluxes are projected to continue to increase in the future. The resulting inputs to the oceans have already led to eutrophication problems in some areas, although on a global scale human impact has done little to increase the vast natural nutrient inventory of the deep ocean. Problems are therefore focused on the coastal environment. Eutrophication describes the process of increased nutrient load stimulating the growth of phytoplankton, the microscopic algae that form the base of the food chain. A modest stimulation of phytoplankton growth has the potential to increase the phytoplankton stock and thereby provide more food for animals further up the food chain. However, excessive nutrient inputs coupled to inadequate dilution can lead to excessive algal growth which can settle into deeper water after death and consume all the available oxygen leading to the death of animals living in deep waters. The extent of such deep-water oxygen depletion depends critically on the physical water exchange processes which promote oxygen supply to the deep waters. Thus in areas of restricted mixing, such as the Baltic or Adriatic Sea, there are serious problems, while in the much more open North Sea the problem is less significant, although all these areas receive major discharges of nutrients via rivers and the atmosphere.

Large-scale algal growth, or blooms, can also involve undesirable or toxic species and have a range

of highly undesirable effects. There is some evidence that such blooms may be increasing worldwide, although systematic studies to confirm this are not available. Indeed, on an oceanic scale, or even a regional scale, it is often difficult to quantify the level of human perturbation of the marine system, particularly since coastal and near-surface environments are characterized by great natural variability. Systematic studies of the marine environment are relatively recent, so we often do not have sufficient a time series of data to identify with confidence concentration trends for many contaminants.

3.3 Pesticides

Pesticide use has many benefits to human society, particularly in terms of increased food production. However, the release of chemicals into the environment which are designed to produce harmful biological effects is inevitably a cause for concern. The use of some pesticides such as DDT and TBT are now severely restricted, although the extent and effectiveness of such regulation varies widely from country to
country. In developing countries the use of other pesticides is continuing to increase. These will inevitably escape into the wider environment and the effect of low levels of these compounds individually or in mixtures is very poorly understood. The input of other potentially harmful organic compounds is also of continuing concern, as noted earlier.

3.4 Global Issues

In addition to these concerns over the direct contamination of the oceans, recently attention has turned to the role of the oceans within the larger global climate-regulating system. Heat and CO_2 exchange between the ocean and the atmosphere are key elements in this system. The emission of gases, such as dimethyl sulfide (DMS), from the ocean also plays a major role in climate regulation. The response of these exchange processes to climate change represents a series of poorly understood feedbacks in the global climate system. This makes predicting the effects of increasing concentrations of CO_2 and other greenhouse gases particularly difficult. Thus it is important to recognize that the oceans play a key role in the control of the whole planetary system, and hence that atmospheric pollution is linked to marine pollution.

3.5 Loss of Marine Habitats

Marine pollution is usually considered in terms of inputs of chemical wastes to the marine environment. However, there are two major issues that represent simpler and more direct threats to the marine environment. First, there is clear evidence that almost all fisheries worldwide are fully or overexploited. In the North Sea, for example, the clearest trend of environmental change is not of chemical contamination, but the vast decline in fish stocks. Globally half of the individual fish stocks are fully exploited and another 22 percent are overexploited.

The second issue is the physical degradation of coastal environments. In some cases this takes the form of total loss of habitat to reclamation. For instance, in the Humber estuary, the largest English estuary draining to the North Sea, more than 90 percent of the intertidal area has been lost to reclamation over the last 300 years. Similar losses have occurred in many other estuaries. In tropical areas half the mangrove forests have been lost to developments of some sort. These represent a massive loss of ecologically important wetlands and also loss of nutrient-processing potential. In other cases the habitat loss is not complete but involves a major reduction of ecological value. This problem is illustrated by the loss of many inshore coral reef systems to sedimentation. This occurs as changing land use patterns increase erosion and hence turbidity in the coastal waters, blocking light to the coral community and ultimately killing the corals. Eutrophication can have similar effects. Intertidal areas and coral reefs also represent important components of coastal defense against erosion and flooding. The destruction of these areas therefore represents economic as well as ecological losses.

4. Research and Regulation

Long-range transport is a feature of marine pollution issues, and hence national regulation alone is inadequate to protect the marine environment. There have been a series of initiatives focused on regional seas such as the Baltic, the North Sea, and the Mediterranean to achieve coordinated international action to manage the area. The United Nations Convention on the Law of the Sea entered into force in 1994, providing an international framework for the exploitation of the marine environment and also a requirement on states to 'use the best practical means at their disposal' to prevent and control marine pollution. While imperfect, these initiatives represent a growing realization that the oceans need to be managed at an international level to ensure their continued value for future generations.

There has also been a growing recognition that progress in managing the marine environment requires interdisciplinary scientific study involving chemists, biologists, physicists, and geologists. An important recent development has been the integration of social scientists in this work to describe the human pressures on the marine system and to consider future management options for the marine environment. An example is the recent valuation of the whole marine ecosystem at $20 billion dollars, although the meth-

odology used is controversial. Most of the value assigned to the marine environment in this analysis is associated with the cycling and storage of nitrogen and phosphorus, and hence intimately linked to nutrient and sewage pollution issues. Huber (1999) noted 'Environmental degradation impedes economic and social development and diminishes human welfare.' Thus the true cost of indiscriminate discharges to the marine environment are now beginning to be considered.

Control of marine pollution problems does not necessarily require sophisticated regulation but rather rigorous and effective planning control that assigns an appropriate value to the marine environment. It also implies controls not only of direct discharges to the marine environment, but also throughout the river catchment draining to the marine environment and of atmospheric sources upwind.

Thus several areas of real concern in terms of potential marine pollution problems can be identified, although the evidence for large-scale effects in the open ocean is modest. The oceans have a very large capacity to dilute and degrade wastes. Marine pollution problems are most evident where this capacity is not effective, such as in coastal areas with restricted circulation or for contaminants that float, such as oil or plastics. The assessment of the nature of these threats has changed somewhat over the years as we have come to understand the marine system better and also to regulate human discharges into the system. Indeed, in a recent report GESAMP noted: 'While the resolution of these issues is neither simple nor easy, in most cases the nature of the ocean's environmental problems is understood, the knowledge to solve them is available and the necessary management tools exist. What is lacking is the determination and political will to act' (Huber 1999).

See also: Climate Change and Health; Climate Change, Economics of; Desertification; Environmental Challenges in Organizations; Environmental Planning; Environmental Policy; Environmental Risk and Hazards; Environmental Vulnerability; Water Pollution: Terrestrial and Groundwater; Water Resources

Bibliography

Andrews J, Samways G, Dennis P, Maher B 2000 Origin, abundance and storage of organic carbon and sulphur in the Holocene Humber Estuary. In: Shennan I, Andrews J E (eds.) *Holocene Land–Ocean Interaction and Environmental Change Around the North Sea*. Geological Society of London Special Publication, London, pp. 145–70

Botsford L W, Castilla J C 1997 The management of fisheries and marine ecosystems. *Science* **277**: 509–15

Clark R 1989 *Marine Pollution*. Oxford Science, Oxford, UK

Constanza R, Arge R, De Groot R, Farber S, Grasso M, Hannon B, Limburg K, Naeem S, O'Neill R, Paruelo J, Raskins R, Sutton P, Den Belt M 1997 The value of the world's ecosystem services and natural capital. *Nature* **387**: 253–60

Goldberg E D 1995 Emerging problems in the coastal zone for the twenty-first century. *Marine Pollution Bulletin* **31**: 152–8

Henrichsen D 1994 Life on the margin. *People and the Planet* **3**: 3–35

Houghton J, Meira Filho L, Callander B, Harris N, Kattenberg A, Maskell K 1996 *Climate Change 1995*. Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK

Huber M E 1999 Oceans at risk. *Marine Pollution Bulletin* **38**: 435–8

Jickells T 1998 Nutrient biogeochemistry of the coastal zone. *Science* **281**: 217–22

McIntyre A D 1995 Human impact on the oceans: the 1990s and beyond. *Marine Pollution Bulletin* **31**: 147–51

North Sea Task Force 1993 *North Sea Quality Status Report*. Oslo and Paris Commissions, London

The Open University Course Team 1991 *Case Studies in Oceanography and Marine Affairs*. The Open University and Pergamon, Oxford, UK

O'Riordan T 1995 *Environmental Science for Environmental Management*. Longman, Harlow, UK

Richardson K 1997 Harmful or exceptional phytoplankton blooms in the marine ecosystem. *Advances in Marine Biology* **31**: 302–85

Turner B L, Clark W C, Kates R W, Richards J F, Mathews J T, Meyer W B 1990 *The Earth as Transformed by Human Action*. Cambridge University Press, Cambridge, UK

Turner R K, Adger W N, Lorenzoni I 1998 *Towards Integrated Modelling and Analysis in Coastal Zones: Principles and Practices*. LOICZ International Project Office, Texel, The Netherlands

United Nations Joint Group of Experts on the Scientific Aspects of Marine Pollution 1982 *The Review of the Health of the Oceans*. GESAMP Reports and Studies no. 15, UNESCO, Paris

United Nations Joint Group of Experts on the Scientific Aspects of Marine Pollution 1990 *The State of the Marine Environment*. GESAMP NEP Regional Seas Reports and Studies no. 115, UNEP, Nairobi, Kenya

T. Jickells

Water Pollution: Terrestrial and Groundwater

1. Introduction

Water is an essential component of life. Plants and animals (including human beings) are dominantly made up of water, and rely on water to regulate temperature, carry dissolved materials inside and between cells, remove waste products, etc. Organisms cannot survive without a supply of water—human beings need to consume approximately 1.5 litres of water per day, and this water needs to be fresh (i.e., lacking in appreciable quantities of dissolved minerals) and clean. Human civilization has traditionally been

found in areas where clean water is readily available for consumption and for agriculture, but, with increasing population densities and increasing industrialization, the freshwater resources on which civilization relies are becoming increasingly strained—an estimated 1,200 million people worldwide lack a satisfactory or safe water supply (Pickering and Owen 1997). Along with overconsumption and resulting water shortages, the introduction of substances hazardous to human health into terrestrial water bodies (or pollution) is the most serious threat to the world's freshwater supplies. This article examines the scale and impacts of pollution in terrestrial, freshwater bodies such as rivers, lakes, and groundwater (underground water, specifically that below the water table). Pollution of the marine environment is examined in *Water Pollution: Marine*.

2. *Water Pollution*

Pollution can be defined as 'the introduction by man into the environment of substances or energy liable to cause hazards to human health, harm to living resources and ecological systems, damage to structures or amenity, or interference with legitimate uses of the environment' (Holdgate 1979). The major pollutants in freshwaters vary across the world, but pollution by sewage wastes is a major problem in 'underdeveloped' nations, whereas pollution from industrial effluents is, arguably, the dominant problem in overdeveloped nations (although sewage and agricultural pollution may also be serious problems). Industrial effluents may contain a variety of hazardous components, such as radioactive materials, hydrocarbons, heavy metals, and toxic organic compounds. Thermal pollution, due to discharge of cooling waters, may also have a significant effect on the biology of water bodies. A summary of the major types of water pollutants and their impacts is given in Table 1.

Many of the substances that cause pollution of water bodies are naturally present in the environment, and organisms tolerate, or even actively require, certain concentrations of these substances. It is the accumulation of these substances beyond safe levels, or the additive effects of a series of pollutants, that leads to detrimental impacts on the environment. For example, nitrates are essential to plant growth, but excess nitrate may cause eutrophication in lakes and rivers and contamination of groundwater supplies (see Sect. 3.2). In some areas, water bodies may be naturally contaminated to toxic levels. In Bangladesh and West Bengal, large volumes of groundwater used for public water supply are polluted with naturally occurring arsenic, which adversely affects the health of millions of people. Arsenic concentrations of up to 1000 $\mu g\ l^{-1}$ have been measured, well above the limits set for Bangladesh drinking water (50 $\mu g\ l^{-1}$) or the maximum concentrations recommended by the World Health Organisation (10 $\mu g\ l^{-1}$). Consumption of this water has led to widespread death and disease (Nickson et al. 1998). Conversely, many organic pollutants, such as some pesticides and industrial chemicals, are entirely artificial and so have no natural background concentration in the environment.

Following their release, pollutants undergo a variety of physical, chemical, and biological processes which act to change their concentration, and often their chemical form. Of the physical processes, mixing and dilution are particularly important. Waste discharged into a fast-flowing river will rapidly be washed downstream and diluted, whereas wastes which seep into reservoirs or groundwater may have long residence times—the water in lakes may take between 10 and 100 years to be replaced, while for groundwater the timescales may be even longer. Consequently, pollutants may build up over time, providing a long-term contamination problem. Other processes which can dramatically affect the concentration of pollutants in water bodies are adsorption and removal onto sediments, and uptake and breakdown by organisms. Sediments and soil particles, particularly those rich in organic material or clays, effectively adsorb a range of polluting substances, such as heavy metals, many organic pollutants, and some radioactive contaminants. Organisms may break down pollutants, e.g., the microbial oxidation and decomposition of sewage sludges, or may influence the transport of the pollutant, e.g., the methylation of elements such as arsenic and mercury by micro-organisms into more volatile (easily evaporated) forms. The impact of any pollutant discharged into a water body will largely depend on:

(a) the scale of the input (i.e., is the substance(s) discharged in large enough amounts to have a discernible effect on biological function?);

(b) the chemical form of the pollutant (e.g., will it undergo rapid removal onto bottom sediments, will it tend to accumulate in organisms?);

(c) the physical, chemical, and biological characteristics of the receiving water body (e.g., is the water body vigorously mixed, is it acidic or alkaline, oxygen-rich or reducing, etc.?).

As discussed above, a diverse range of pollutants may be present in freshwater systems. A detailed synthesis of the environmental behavior and impacts of each of these individual pollutants is beyond the scope of this article, and the reader is referred to the Bibliography. Instead, the potential impacts and environmental behavior of these pollutants in terrestrial water bodies is perhaps best shown by considering three case studies which illustrate the impacts of three major waste categories on terrestrial water bodies: (a) wastes from mineral extraction and processing, (b) agricultural wastes, and (c) urban and industrial hazardous wastes.

Table 1
Major types of pollutant and their impacts in terrestrial water bodies

Pollutant	Major anthropogenic sources	Environmental impacts
Oxygen-consuming wastes	Sewage effluent, agricultural run-off, some industrial effluents	Depletion of dissolved oxygen in water bodies through decomposition by aerobic bacteria, leading to death of flora and fauna. Decomposition by anaerobic bacteria produces toxic gases such as hydrogen sulfide
Nutrients	Sewage effluent, agricultural run-off	Algal blooms and subsequent death of submerged vegetation. Large quantities of rotting organic material produced causes oxygen depletion
Hydrocarbons	Drilling operations, accidental spillage, waste disposal	Various biological impacts, including death of organisms by ingestion or smothering
Other organic pollutants (e.g. organochlorines, organic solvents)	Agricultural run-off, sewage effluent, waste incineration, landfill sites, industrial effluents	Some components are carcinogenic and may bioaccumulate. Biological impacts include depression of immune system and reduced reproductive success in organisms
Acids	Acid rain, acid mine drainage	Acidification of natural waters and subsequent increase in toxic metal concentrations. Decline in species richness, death of fish
Heavy metals	Ore mining and associated industries, various technological processes	Various toxic effects depending on metal present. May bioaccumulate through the food chain
Radionuclides	Nuclear weapons testing, deliberate and accidental release from nuclear facilities	Depending on exposure, may cause cell damage, genetic defects, and cancers
Heat	Coolant waters from industry	Changes in species composition, suffocation of fish, disruption of reproductive cycle in fish and other aquatic organisms

Source: Jackson and Jackson (1996).

3. *Case Studies*

3.1 Legacy of Current and Historical Mining Activity

None of the 40 main elements that are economically important to modern industrial society can be synthetically made. Instead, they must be mined, extracted, and purified, which inevitably generates waste material and frequently causes water pollution. Mining of coal and metal ores generates vast amounts of waste material that contains significant quantities of metals. These wastes are generally dumped to form spoil heaps around the mining area, or may be washed into rivers. Sulfide minerals, most commonly iron sulfide (FeS_2) or pyrite, are commonly associated with coal seams and metal ores, and exposure of these minerals to oxygen and water produces reactions of the following type:

$$2FeS_2 + 7O_2 + 2H_2O \rightarrow 2FeSO_4 + 2H_2SO_4$$
$$\text{pyrite} + \text{oxygen} + \text{water} \rightarrow \text{iron sulfate} + \text{sulfuric acid}$$

Hence, the pyrite oxidizes to form iron sulfate and sulfuric acid. This generates acid mine drainage, which lowers the pH of water bodies into which it flows, dramatically affecting ecosystem characteristics, and which may solubilize toxic heavy metals associated with the mine waste. Extraction and purification of the mined material may also cause large-scale pollution of water bodies. Use of mercury to extract gold from river sediments in the Amazon region of Brazil has caused poisoning in people living around and downstream from mining areas (e.g., Homewood 1991), while cyanide-contaminated water accidentally released from the Baia Mare goldmine in northwest Romania recently devastated freshwater ecosystems along the River Tisza in Hungary.

In countries which underwent early industrialization, such as the UK, the rapid development and subsequent decline of the mining industry means that pollution from abandoned mines is arguably a bigger threat to water quality than contemporary mining activity. In Cornwall, the former center of the UK tin mining industry, a land area of around $50\,km^2$ is covered by mine wastes derived from historical working of tin and copper deposits and large-scale extraction of kaolinite. Some of these wastes contain elevated concentrations of heavy metals and have a low pH from the oxidation of sulfide minerals, which promotes the leaching of heavy metals. The release of waste particles from the extraction and washing of mined materials has led to extensive siltation of rivers and estuaries, while groundwater percolating through disused mines has produced acid mine drainage. The most obvious example of this is the 1992 Wheal Jane incident, where 50×10^6 l of low pH, heavy-metal-enriched mine water was released into Restronguet Creek following overflow of groundwater from Wheal Jane (a recently closed mine). The biological impact of this event, however, was relatively minor, and the heavy metal input small in comparison to the existing amounts of heavy metals in the sediments derived from historical mining activity.

In the Mendip Hills, southwest England, lead concentrations in soils around some former Pb–Zn mine workings are over five percent, over 500 times higher than lead concentrations in uncontaminated soil. Much of the contaminated material is in the form of lead-rich particles, which have been mixed into the local soil and subsequently covered by vegetation. Hence, the danger of human exposure is dramatically reduced, provided the surface remains undisturbed. The neutral to alkaline pH of the soil, and the presence of the lead in a mineral form that is stable in the presence of oxygen, i.e., as a carbonate mineral rather than a sulfide mineral, means that little lead is mobilized or leached from the site. In mine waste dumps generally, toxic heavy metals may be stabilized and water contamination reduced through inhibiting leaching via the insertion of clay or geotextile liners, and maintaining pH conditions suitable for heavy metal containment (e.g., where sulfide minerals are present, constructing a low pH, low oxygen environment such as a wetland).

3.2 Groundwater and River Contamination from Intensive Agriculture

Humans have practiced agriculture for thousands of years. Shifting cultivation, rough grazing of livestock, and crop rotation systems were the dominant agricultural methods until the 1950s, when these were largely superceded by intensive agricultural techniques in Europe and North America. In intensive agriculture, traditional crop rotation is replaced by continuous cultivation of crops, which requires the addition of large amounts of nutrients to replace those lost during harvesting. This is commonly done through the application of phosphate and nitrate-containing inorganic fertilizers. Wash-off of these fertilizers into terrestrial water bodies may cause artificial eutrophication to occur, where enrichment of nutrients (particularly phosphate) causes excessive growth of surface algae. These so-called algal blooms prevent sunlight reaching deeper-dwelling aquatic species, and on death produce large quantities of organic matter which may cause oxygen depletion (see below). It is estimated that in the USA, 85 percent of large lakes near major urban areas are affected by artificial eutrophication (Jackson and Jackson 1996).

The extensive use of nitrogen fertilizers has also increased the nitrate content of surface and ground waters over the last 50 years. Nitrates are not adsorbed in soils, but are instead leached into surface water bodies and groundwater, waters which may subsequently be used for drinking water supplies. Excessive nitrate in drinking water can cause a severe blood disorder, 'blue baby syndrome' or methemoglobinemia, in babies under six months of age, and may possibly contribute to the synthesis of carcinogenic nitrosamines in the human body. In the UK, the maximum allowable nitrate concentration in drinking waters is $50\,mg\,NO_3\,l^{-1}$. In intensively farmed areas of the southern UK, such as East Anglia, however, peak winter nitrate values in rivers may exceed $100\,mg\,NO_3{?}l^{-1}$ (Alloway and Ayres 1993).

In addition to continuous crop cultivation, the development of intensive agriculture has led to increased use of factory farming methods for livestock. The organic wastes produced by large concentrations of livestock can have a serious effect on the oxygen content of water bodies. The bacterially driven breakdown of organic wastes released into natural waters consumes oxygen, e.g.

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O$$

$$\text{glucose} + \text{oxygen} \rightarrow \text{carbon dioxide} + \text{water}$$

Hence, the introduction of concentrated organic wastes can deplete water bodies of oxygen, with detrimental impacts on flora and fauna. In extreme cases, where the rate of utilization of oxygen is greater than the rate at which oxygen is replenished (e.g., in canals or slow-flowing rivers), all the available oxygen may be consumed. These conditions favor anaerobic bacteria, which produce toxic hydrogen sulfide and ammonia.

3.3 The Love Canal Disaster and Landfill of Hazardous Wastes

All industrial processes produce waste. Modern lifestyles are dependent upon a vast range of chemicals and their by-products, e.g., a typical office contains

plastics, artificial floor coverings, computer components, etc. Synthesizing and producing these components inevitably produces waste, which needs to be recycled or disposed. A fraction of the waste produced may contain substances hazardous to human health, and disposal of these wastes needs to be carefully controlled. A variety of recycling and disposal options exist, ranging from uncontrolled direct dumping in water bodies such as rivers to removal and recycling of the hazardous wastes for use in other industrial processes. Where it is uneconomic or unfeasible to recycle the toxic components, incineration at high temperature is often used to break down various toxic species, although the incineration process is never 100 percent efficient, and may increase the toxicity of some waste products. For example, incomplete combustion of polychlorinated biphenyls (PCBs) in wastes can lead to the formation of toxic dioxin-containing products. An alternative and widely used disposal method is disposal at landfill sites. Landfill sites receive large quantities of domestic and commercial waste, mixed with varying quantities of hazardous wastes. Current landfill design uses a multibarrier disposal approach, where a series of barriers are put in place which are designed to slow the release of toxic components.

Dilution and dispersion, chemical reactions and biological processes (e.g., bacteria may break down some toxic organic components into less toxic forms, acid wastes may be neutralized by contact with surrounding alkaline rocks), reduce the toxic chemical species to acceptable concentrations before they come into contact with the public via surface and groundwater. The safe and long-term disposal of hazardous wastes, however, requires a thorough understanding of the physical, chemical, and biological processes occurring in and around the landfill site. In particular, groundwater movement and groundwater quality need to be monitored to assess whether toxic chemicals are being leached into water supplies. An illustration of the importance of groundwater movement in the landfill disposal of toxic wastes is provided by the disaster at New York State's Love Canal.

Love Canal was initially dug in the 1890s by William Love to connect the upper and lower Niagara River and provide hydroelectric power. The canal was subsequently abandoned until 1942, when it was purchased by the then Hooker Electrochemical Company. Over the next 10 years, the company dumped 22,000 tons of toxic chemical waste at the site, including carcinogenic lindane, chlorobenzenes, and dioxin-contaminated trichlorophenol. Population pressure from the City of Niagara Falls led to Hooker selling the site to the city school board. Prior to the sale the wastes were covered with a protective clay layer, and a school, houses, and roads were subsequently built over the landfill.

The clay cap overlying the canal, coupled with the fact that the canal was cut into the underlying clay geology, largely isolated the waste by limiting water migration (clay tends to retard water flow, and many pollutants will adsorb on to clays, see Sect. 2). The construction of a major road along the southern end of the canal in the 1960s, however, significantly altered the site hydrology. The road blocked the natural through-flow of groundwater at the site, which, until then, had been southward towards the Niagara River. Heavy rain and snow melt in the winter of 1977 caused a rise in groundwater levels, and unable to flow southwards, this groundwater filled the canal, which eventually overflowed. Contaminants migrated through surface soil layers and along sewer pipes, eventually entering the basements of houses around the site. Some buried drums of waste also broke through the overlying soil. Following national press coverage in 1978, a federal state of emergency was declared, and the worst-affected houses were evacuated, purchased by the state, and destroyed. A full review of the disaster and its aftermath is given in Hoffman (1995).

Although the Love Canal disaster is an extreme case, it does illustrate the potential long-term threat to terrestrial water quality of inadequately managed hazardous wastes. Disposing of any waste underground needs a careful consideration of water infiltration and groundwater movement—for example one of the major limitations regarding deep geological disposal of nuclear waste is the extent to which contaminants buried underground may interact with, and contaminate, groundwater.

4. *Water Pollution in the Twenty-first Century*

The twentieth century was a period of rapid technological development and population growth. The second half of the century saw the development of intensive agricultural systems, widespread nuclear power generation, and the synthesis and use of a wide variety of new organic chemicals. The wastes arising from these, and other, processes contain a wide variety of polluting substances, some of which are extremely mobile and persistent in the environment. There is now, however, a much greater knowledge of the impacts of such pollutants, and the need for controlled disposal and subsequent environmental monitoring is widely recognized. Technological developments have taken place which produce cleaner effluents and allow remediation of contaminated areas, and a much stronger legal framework exists in many countries to control polluting processes which may pose a threat to water quality. Pressure on terrestrial water resources will inevitably increase over the twenty-first century as the world's population grows, however, and it remains to be seen whether the quality of this water can be maintained.

See also: Desertification; Environment and Development; Environmental and Resource Management; Environmental Vulnerability; Environmentalism: Philosophical Aspects; Environmentalism: Preservation and Conservation; Human–Environment Relationships; Resource Institutions; Tropospheric Ozone: Agricultural Implications; Waterfront Planning

Bibliography

Alloway B J, Ayres D C 1993 *Chemical Principles of Environmental Pollution*. Blackie Academic and Professional, London

Hoffman A J 1995 An uneasy rebirth at Love Canal. *Environment* **37**(2): 4–9, 25–31

Holdgate M W 1979 *A Perspective of Environmental Pollution*. Cambridge University Press, Cambridge, UK

Homewood B 1991 Mercury poisoning confirmed among Amazon villagers. *New Scientist* **132**: 18

Jackson A R W, Jackson J M 1996 *Environmental Science: the Natural Environment and Human Impact*. Longman, Harlow, UK

Nickson R, McArthur J, Burgess W, Ahmed K M, Ravenscroft P, Rahman M 1998 Arsenic poisoning of Bangladesh groundwater. *Nature* **395**: 338

Pickering K T, Owen L A 1997 *An Introduction to Global Environmental Issues*, 2nd ed. Routledge, London

A. Cundy

Water Resources

Water resources encompass the freshwaters of the world. The geographic distribution and flow of freshwaters affect supplies of water available for human use, and associated problems of surplus, scarcity, and pollution. Psychological and behavioral research has focused on the perception of water problems and decision-making models, while phenomenological research has explored the deeper meanings of water use. Social scientific research on water has evolved from description of regional water problems, to economic and institutional analysis of alternatives and, more recently, to political and cultural dimensions of water problems. Although flawed, Karl Wittfogel's research on social aspects of water development stimulated work of theoretical and practical value, which argues for continuing attention to comparative international research on water resources.

1. Freshwaters of the World

In the social sciences 'water resources' generally refer to the freshwaters of the world. Freshwater includes precipitation that falls over the continents as rain or snow, infiltrates into soils and groundwater aquifers, runs off into stream networks that drain the continents, evaporates from lakes and transpires from vegetation, and ultimately flows back into the deltas and estuaries along the continental margins. As coastal and marine resources involve different types of use, they are treated separately (see *Water Pollution: Marine*). Advances in global environmental science may lead to greater integration fresh and saline water research.

Freshwaters constitute only 2.5 percent of the world's water resources, the bulk of which (96.5 percent) are in the oceans. The largest proportion of freshwaters are frozen in ice caps and glaciers (68.7 percent), followed by groundwater (30.1 percent), lakes (0.26 percent), the atmosphere (0.04 percent), marshes (0.03 percent), rivers (0.006 percent), and biological water (0.003 percent) (Shiklomanov 1993). Biological waters are small in volume but have great social significance because they include the water in human bodies.

In addition to their physiological and ecological importance, freshwaters have been central to fundamental social debates about property, power, public interest, and the role of the state. Those debates, in turn, have influenced the course of environmental science, technology, and law; and those fields have deepened understanding of the distribution and flows of water from one location, physical state, and quality to another.

2. Water Distribution and Flows

The static percentages of freshwater mentioned above must be qualified in four important ways to understand their social significance. First, freshwaters cycle dynamically in space and time. For example, biological waters are replaced (through ingestion, perspiration, and excretion) on timescales of hours to days, while groundwater is only replaced (through discharge and recharge) on timescales of tens of thousands of years. These temporal cycles have profound implications for human water supply and pollution control.

Although biological water needs do not vary much within species, they do vary across geographic regions. The distribution of freshwater varies by latitude and by climatic processes within latitudinal zones, which range from arid to humid and which vary on seasonal, annual, and longer timescales.

These variations have implications for human water use that are analyzed through water budget analysis. A water budget compares the amount of water supplied by precipitation with the amount of water consumed by evaporation and transpiration from vegetation and crops—the largest use of water worldwide. The difference between precipitation and evapotranspiration defines the magnitude and duration of surplus and scarcity, which societies address by acquiring supplemental water, draining excess waters, or moving to more favorable habitats.

Freshwaters also vary in physical characteristics, chemical composition, and aquatic life. These charac-

teristics reflect the broad climatic and physiographic contexts that shape the landscape of stream networks, groundwater aquifers, and wetlands. Variations in freshwater distribution, characteristics, and landscapes condition human water experience, which is the primary concern of this article (Gleick 1998). People who live in snow covered alpine watersheds have different water supplies, quality, and problems from those living in temperate or tropical watersheds fed by rain. People living in downstream environments face different water opportunities and problems from those upstream. It is well known that downstream groups face water quantity and quality problems generated by upstream groups; but it is appreciated less widely that downstream groups often have the political and economic power to regulate and constrain upstream development. Many engineering projects of the twentieth century failed to meet their intended goals or gain public support, in large part due to inadequate understanding of such basic social and behavioral aspects of water. Increasingly it is recognized that social research represents an important frontier for water resources management (Baumann and Haimes 1988).

3. *Psychological and Behavioral Aspects of Water*

In addition to supporting life, water has distinctive psychological characteristics. At a physiological level, 'thirst' is experienced differently from 'hunger' and other needs. Psychological disorders (e.g., water phobias) include fears of immersion, swimming, and drowning. At a social level, perception of water-related risks and adjustments, have received attention in behavioral research on flood and drought hazards, beginning in the 1960s. Flood and drought hazards research evolved from an early emphasis on description of disaster losses and recovery strategies to research that seeks to explain decisions to locate in floodplains (White 1945).

Droughts develop more slowly than floods and involve different processes of perception, behavior, and adjustment (Saarinen 1966). Water-related disease hazards range from disease vectors in aquatic habitats to inadequate water supply for washing, and vectors that breed in wet areas (White et al. 1972). Reexamination of pioneering studies of hazards designed to assess the effects of increased knowledge have noted limited progress in reducing economic losses. Efforts to develop general decision-making models to address water problems developed a key analytical construct, known as the 'range of choice,' which attributed failings to undue narrowing of the alternatives considered (Wescoat 1992).

Criticism of behavioral research on water has focused on structural constraints, especially in developing countries. These criticisms led to research on social vulnerability to water problems. Another, smaller, line of phenomenological inquiry has questioned assumptions about human experience of water (Bachelard 1983, Cosgrove and Petts 1990).

4. *Cultural Representations of Water*

Social research on the representations of water (e.g., in texts, arts, and popular culture) is scattered across many fields. In the sphere of belief systems, it includes creation stories such as the *Enuma Elish* that begin with a watery chaos from which fundamental divisions emerge. Comparative studies of flood stories (Dundes 1988), water deities and rituals, and paradise gardens 'underneath which rivers flow' (*Qur'an*, passim) have continuing significance for many cultures.

In the modern era, numerous travel accounts proceed from source to sea on major rivers of the world. While tourists generally travel downstream, 'imperial' explorers often labored against the current, searching out navigable trade routes and resources (e.g., Rennell on the Ganges; and Lewis and Clark on the Missouri River). Some literary works, such as Henry David Thoreau's *Walden*, influenced contemporary and subsequent views of water. There has been less research on water in the visual arts, music and film. One exception is the use of photography in environmental perception research related water resources management and planning (Litton et al. 1974).

5. *Water as a Natural Resource*

Questions about managing water as a natural resource, and about whether 'water is different' from other resources, have been central concerns of social scientists. They are pursued in close association with engineering, natural sciences, and law.

Water management has been dominated by engineering sciences, which have addressed social ideas and problems implicitly if not explicitly from antiquity to the present. Debates on the origin of rivers were resolved with the development of watershed science in seventeenth century France, which made it possible to estimate water supplies. The expansion of irrigation agriculture advanced with the development of water budget analysis. Water pollution control evolved with advances in instruments to measure organic and inorganic contaminants.

This is not to say that traditional water management is unsophisticated, impractical, or ineffective. Humoral theories that originated in Greek and Islamic sciences are still practiced effectively in rural South Asia. Hindu temple tanks and modern irrigation works operate parallel, though not always in coordination, with one another (Lansing 1991). The contrast between traditional and modern breaks down at many points. Some traditional irrigation societies devised highly sophisticated water diversion and mea-

surement systems that required elaborate forms of social organization (e.g., *qanats* in Persia). Many modern water systems, by contrast, are supply-driven with little measurement or management of water. Irrigation efficiences in some areas of the western US are only 50 percent, compared with rates of 90 percent in Israel and Jordan.

The most acute water problems seem to occur in low-income areas subject to civil unrest, state corruption, vague water laws, and/or highly skewed land tenure regimes. These conditions have arisen in irrigation systems and peri-urban areas of developing countries. In those situations, the laws, policies, and regulations that do exist rarely are implemented.

Laws define aspects of water that society regards as a resource. From antiquity to the present, legal systems have defined certain aspects of water as natural rights, others as social rights, and still others as *res nullius*. Legal struggles reflect changing social values and water demands.

Social research on water laws has focused on four major themes: (a) access rights; (b) water rights transfers; (c) institutional fragmentation; and (d) transboundary conflicts. Research on access rights asks who (including animals) has rights to use water and what are the social and environmental consequences of different access regimes. Research on water transfers has focused on the impacts of property rights transfers from one type or location of water use to another (Saliba and Bush 1987). Research on institutions has highlighted fragmentation, gaps, and redundancies among water organizations; and it has more recently begun to focus on power relations among organizations and their social and environmental consequences (Emel and Roberts 1995). Transboundary conflicts encompass water pollution and floods, as well as water scarcity. During the 1990s, a large body of research focused on transboundary water problems in the Middle East (e.g., Biswas et al. 1997, Committee on Sustainable Water Supplies for the Middle East 1999).

It is not clear what influence social research has had on water laws to date. The slow pace of legal reform has led social scientists to concentrate on public policies where legal issues may be at stake but decisions will likely be made by water managers and users. One of the oldest policy debates revolves around questions of whether 'water is different' from other resources. Most social scientists grant that water has unique physical qualities and uses but some, principally economists, have argued that water is no different from other resources. By this argument, water is a means to ends that have social values that may be, indeed must be, compared with other values, if society and individuals are to make rational decisions to improve human welfare and environmental quality.

These arguments have stimulated detailed empirical research on the economics of water use. One set of researchers has quantified the price elasticity of demand for residential water use while others have focused on agricultural water values (Baumann et al. 1998). These sectoral studies were drawn together in models for optimization and multiobjective management of complex river basins in the 1960s. These models sought to calculate trade-offs among competing water uses for agriculture, flood protection, recreation, municipal, and industrial sectors.

The principal internal debate in these efforts pitted optimization models, based on neoclassical economic principles, against simulation models based on behavioral models of decision-making. After sobering recognition of the limited use of these models, and further technological advances, research in the 1990s began to develop decision support systems (DSS). That work built upon research in human–computer interactions, river basin operations, and adaptive environmental assessment and management.

More recently, environmental economists have sought to estimate 'nonuse' values of waters that remain in natural watercourses, the value of 'ecosystem services,' and the trade-offs among competing environmental objectives. As in other applications, these water resources efforts rely largely on contingent valuation and other indirect methods.

As is evident in the examples above, water resources research has been more concerned with pragmatic inquiry (in the philosophical and colloquial senses of the term) than in leading debates in the social sciences (Wescoat 1992). It strives to define the social dimensions of pressing water problems and evaluate the consequences of alternative courses of action. Early research focused on river channel problems (e.g., flooding, navigation, and pollution). This emphasis on channels, rather than basins or watersheds, reflected legal and technical constraints on the role of the state in natural resources management. A concurrent line of research sought to describe river basin problems and resources. Basin surveys became increasingly important in the early twentieth century in Europe and the US. The Tennessee Valley Authority (TVA), authorized in 1933, sought to address problems of severe economic poverty, land degradation, and bureaucratic fragmentation. It included regional resource surveys, the design of planned communities, and economic development programs. Although the TVA experiment was not replicated in the US, it had growing significance in postwar development programs in many countries (e.g., the Damodar, Gal Oya, Indus, and Mekong river basins).

Policy reforms in the 1930s were influenced by social research on water financing, pricing, and decision-making. The TVA framework was broadened to the scale of national water planning in the 1960s. These efforts culminated in the US with the 'Principles and Standards for Planning Water and Related Land Resources,' which established four evaluation criteria: national economic development, regional economic development, environmental impacts, and other social

effects (e.g., community impacts). Economic research focused on the first account, while other social scientists focused on the other three accounts.

One line of opposition to these economic and multiobjective approaches argued that water is different from other resources and that it has unique importance in supporting life. These views were advanced both by water developers arguing for irrigation projects and their environmentalist and antidam opponents. Indigenous values often conflict with all of these arguments and approaches, though oppositions between indigenous and modern values are sometimes presumed rather than studied closely. Although arguments for the inherent natural or cultural value of water have sometimes been self-serving, they effected a major shift from water development to environmental management in the late-twentieth century. In some cases, they have disputed the possibility of commensurating different water uses and values (Espeland 1998). Policy stalemate is sometimes taken to be evidence of incommensurable values. Analysis of institutional problems, including conflict and stalement, was a major research advance in the 1970s and 1980s. In earlier decades, water organizations increased in scale and scope to address increasingly complex water problems, but they often weakened local institutions. In response, some research focused on local irrigation organizations (Ostrom 1992), while others assessed bureaucratic behavior and reform (Coward 1980). As engineering projects failed to meet their planning objectives and had unanticipated social impacts, increasing attention was directed to 'institutional strengthening,' 'capacity-building,' and 'participatory development' (Uphoff 1992). After two decades of research and experimentation, the importance of institutional factors is well established, though the efficacy of institutional alternatives remains a high research priority.

Beyond these economic and institutional approaches, social scientists have examined an array of specific social impacts in ways that have redefined water resources management. Social research on large reservoirs has focused on displacement and resettlement of people whose homes and lands are flooded by reservoirs, examining issues of equity, state performance, and antidam movements that have stopped international aid for projects in India, Turkey, and China.

Recognition that women, especially in poor rural societies, spend disproportionate amounts of time and energy carrying water and managing domestic water and sanitation, led to social research on water and gender. Development assistance, by contrast, often addressed men in those societies. Feminist research has shed light on women's understanding of water resources (e.g., Feldhaus 1995), and gender was one of four top priorities identified by the UN Commission for Environment and Development in 1992 for improving water management. Movements have organized around environmental problems in aquatic and riparian habitats, including species extinction, wetlands preservation, groundwater contamination, and river restoration. During the 1990s, a large number of watershed movements arose in the US to address issues ranging from nonpoint source pollution to riparian habitat protection and landscape identity. Social scientists have sought to understand how these movements articulate environmental values (intrinsic, instrumental and inherent) with political strategies, organizational dynamics, and environmental consequences.

A growing line of research focuses on political aspects of water management that behaviorist approaches failed to explain. These political issues include acute water problems faced by the poor, corruption and violence in water administration, inequitable access to water, vulnerability to flood hazards, and devolution of state authority to nongovernmental organizations. Brown and Ingram (1987) have attempted to explain problems faced by low-income ethnic minorities by contrasting 'community' and 'commodity' values in water. A broader, but fundamentally flawed, political framework was put forward by Wittfogel (1981).

6. *Water and Power: Wittfogel's Theory of Hydraulic Civilizations*

Wittfogel argued that irrigation agriculture had a formative role in the development of complex bureaucratic social organization, initially in the large arid floodplain environments of Asia which constituted a distinctively 'Oriental mode of production' that has, throughout history, threatened democratic societies. This contorted web of arguments was refuted soundly on empirical and theoretical grounds. Archaeological research on early irrigation systems showed that large-scale irrigation agriculture often followed complex state formation and had no consistent relationships with 'despotism' as Wittfogel defined it.

Notwithstanding these flaws, Wittfogel's work had two lasting implications for social research on water resources. First, it provoked three waves of research that included leading social scientists. The first wave included archaeologists and social historians who challenged the historical argument (e.g., Butzer 1976). The next wave of students focused on structures of power in modern irrigation systems. A third generation of students became involved in the applied institutional experiments described earlier (Coward 1980). The theoretical contributions of each group was sharpened by their response to, and continuation of, debates provoked by Wittfogel.

Wittfogel's second legacy was to draw together structural aspects of water management that were previously neglected or treated separately. His frame-

work encompassed environmental conditions, technologies, economic patterns, property regimes, state organization, social class, and modes of social power. Discerning the relationships among these structures of water management remains a major challenge for water researchers.

7. Conclusion

Of the natural resources, water seems especially prone to social experimentation. Its fluid characteristics make it less subject to private control than land, and thus more open to collective control. It is more subject to social control than diffuse resources like the atmosphere, which eludes most (purposeful) human management efforts. Water resources thus lie at a vital intersection between public and private control. The historical interactions between public and private water management, and social research on them have yielded insights of broader significance for the social and behavioral sciences.

One important observation on the twentieth century's water research concerns the changing significance of social research in different parts of the world. In the late-nineteenth century, US and Australian water specialists traveled around the world for models and lessons of water development in Asia, the Middle East, and Europe. The subsequent dominance of US and European researchers was based on advances in the theory and practice of integrated river basin development. But the most creative social research on water management in the late-twentieth century occurred in developing areas that were open to institutional experiments in irrigation, domestic water supply, and sanitation (Lansing 1991, Uphoff 1992). The most creative transboundary water research focused on Middle Eastern water problems. US and European research remained innovative in the fields of floodplain management, urban water management, and water resources effects of global climate change.

These observations point toward a number of opportunities for future inquiry, including: (a) collaborative research on transnational water problems; (b) research on social scales of water control from the local to global; (c) critical assessment of trends from public water management toward private and market-based water management; and (d) research on experiments in 'science-based policy' and 'adaptive management' of complex water ecosystems (National Research Council 1999).

See also: Conservation: Waterways; Conservation: Wetlands; Desertification; Environmental and Resource Management; Environmental Economics; Environmental Planning; Environmental Security; Environmentalism: Preservation and Conservation; Food Security; Land Reclamation; Water Pollution: Terrestrial and Groundwater; Waterfront Planning

Bibliography

Bachelard G 1983 (1942) *Water and Dreams: An Essay on the Imagination of Matter*. (Trans. Farrell E R). Pegasus Foundation, Dallas, TX

Baumann D, Boland J, Hanemann W M 1998 *Urban Water Demand Management and Planning*. McGraw-Hill, New York

Baumann D, Haimes Y (eds.) 1988 *The Role of Social and Behavioral Sciences in Water Resources Planning and Management*. American Society of Civil Engineers, New York

Biswas A K, Kolars J, Murakami M, Waterbury J, Wolf A 1997 *Core and Periphery: A Comprehensive Approach to Middle Eastern Water*. Oxford University Press, Delhi

Brown F L, Ingram H 1987 *Water and Poverty in the Southwest*. University of Arizona, Tucson, AZ

Butzer K W 1976 *Early Hydraulic Civilization in Egypt: A Study in Cultural Ecology*. University of Chicago, Chicago

Committee on Sustainable Water Supplies for the Middle East 1999 *Water for the Future: The West Bank and Gaza Strip, Israel, and Jordan*. National Academy Press, Washington, DC

Cosgrove D, Petts G (eds.) 1990 *Water, Engineering and Landscape*. Belhaven, London

Coward Jr E W (ed.) 1980 *Irrigation and Agricultural Development in Asia*. Cornell University Press, Ithaca, NY

Dundes A (ed.) 1988 *The Flood Myth*. University of California Press, Berkeley, CA

Emel J, Roberts R 1995 Institutional form and its effect on environmental change: The case of groundwater in the Southern High Plains. *Annals of the Association of American Geographers* **85**: 664–83

Espeland W N 1998 *The Struggle for Water: Politics, Rationality, and Identity in the American Southwest*. University of Chicago Press, Chicago

Feldhaus A 1995 *Water and Womanhood: Religious Meanings of Rivers in Maharashtra*. Oxford University Press, Delhi

Gleick P 1998 *The World's Water: The Biennial Report on Freshwater Resources, 1998–1999*. Island Press, Washington, DC

Kromm D E, White S E 1992 *Groundwater Exploitation in the High Plains*. University of Kansas Press, Lawrence, KS

Lansing S 1991 *Priests and Programmers: Technologies of Power in the Engineered Landscape of Bali*. Princeton University Press, Princeton, NJ

Litton R B, Tetlow R J, Sorenson J, Beatty R A 1974 *Water and Landscape. An Aesthetic Overview of the Role of Water in the Landscape*. Water Information Center, Port Washington, WI

Maass A, Anderson R L 1978 *...and the Desert Shall Rejoice: Conflict, Growth and Justice in Arid Environments*. MIT Press, Cambridge, MA

National Research Council 1999 *Downstream: Adaptive Management of Glen Canyon Dam and the Colorado River Ecosystem*. National Academy Press, Washington, DC

Ostrom E 1992 *Crafting Institutions for Self-Governing Irrigation Systems*. ICS Press, San Francisco

Saarinen T 1966 *Perception of the Drought Hazard on the Great Plains*. Research paper no. 106, Department of Geography, University of Chicago, Chicago

Saliba B C, Bush D B 1987 *Water Markets in Theory and Practice: Market Transfers, Water Values and Public Policy*. Westview Press, Boulder, CO

Shiklomanov I 1993 World fresh water resources. In: Gleick P (ed.) *Water in Crisis: A Guide to the World's Fresh Water Resources*. Oxford University Press, New York, pp. 13–24

Uphoff N 1992 *Learning from Gal Oya: Possibilities for*

Participatory Development and Post-Newtonian Social Science. Cornell University Press, Ithaca, NY
Wescoat Jr J L 1992 Common themes in the work of Gilbert White and John Dewey: A pragmatic appraisal. *Annals of the Association of American Geographers* **82**: 587–607
White G F 1945 *Human Adjustment to Floods*. Research paper no. 29, Department of Geography, University of Chicago, Chicago
White G F, Bradley D, White A U 1972 *Drawers of Water: Domestic Water Use in East Africa*. University of Chicago Press, Chicago
Wittfogel K W 1981 (1957) *Oriental Despotism: A Comparative Study of Total Power*, 1st edn. Vintage, New York

J. L. Wescoat Jr.

Waterfront Planning

Waterfront planning considers the future of the shoreline rivers, lakes or oceans, especially the redevelopment of abandoned port lands in urban areas.

1. *Evolution of the Port–City Interface*

Waterfront redevelopment became a major planning activity in almost every port city in the 1970s and 1980s. Transport geographers such as Hoyle (1988) suggest that redevelopment is only the latest stage in the evaluation of the port–city interface (Table 1). The primitive cityport of the nineteenth century (Stage I) used medieval technology to unload sailing ships: nets, pulleys, and carts. Many ports were rebuilt in the late nineteenth century (Stage II) to accommodate steamships with warehouses, linear quays, and rail terminals. The Port of London Authority (1908) was an early public agency which consolidated privately-owned docks and built modern facilities. Its example was followed by other cities so that much of the urban waterfront in the mid-twentieth century was publicly owned, devoted to port industrial uses and not accessible to the general public.

The modern industrial cityport (Stage III) emerged quite quickly in the 1960s and 1970s due to powerful technological changes. Passenger terminals disappeared as jet aircraft replaced ocean liners. Fleets of small fishing boats were replaced by factory vessels. General freight switched to container ships. Huge bulk carriers and supertankers needed ocean terminals. Port-related industries in multistorey facilities departed for suburban sites with expressway access, or to other countries.

New port facilities required less labor, and more land. Container ports were approximately 40 times faster than conventional cargo facilities, using a fraction of the labor (Chilcote 1988), but needed 10 times the land for staging fast rail or highway access. Competition from new ex-urban port facilities was fierce (Stage IV). In a few years, Port Elizabeth, NJ replaced Manhattan, Oakland replaced San Francisco, Tilbury replaced London, and Boston and Philadelphia lost much of their cargo trade. By 1980, almost every maritime city had underutilized port lands available for redevelopment (Stage V) on the

Table 1
Stages in the evolution of the city–port interface

Stage	Symbol ○ city ● port	Period	Characteristics
(I) Primitive cityport		Ancient–medieval to 19th century	Close spatial and functional association between city and port
(II) Expanding cityport		19th–early 20th century	Rapid commercial and industrial growth forces port to develop beyond city confines, with linear quays and break–bulk industries
(III) Modern industrial cityport	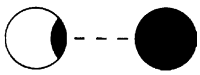	mid-20th century	Industrial growth (especially oil refining) and introduction of containers and ro-ro facilities require separation and increased space
(IV) Retreat from tthe waterfront	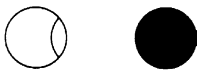	1960–1980s	Changes in maritime technology induce growth of separate maritime industrial development areas
(V) Redevelopment of the waterfront		1970–1990s	Large-scale modern port consumes large areas of land- and water-space; urban renewal of original core

Source: Hoyle et al. (1988)

front doorstep of its central business district (CBD). The largest ports had enormous derelict areas; London had over 5000 acres available in 1980, in docks that had been the busiest in the world only 20 years before. Whole communities of dock workers became unemployed.

2. Opportunities and Problems for Waterfront Planning

The good news was that the abandoned port lands were in large, publicly-owned blocks close to the CBD of many cities. Most waterfront sites did not need expropriation of private property or displacement of residents, common problems in previous urban renewal projects. The waterfront became available during a period when many cities were looking for lands to expand the financial and service industries in their CBD (see *Urban Planning: Central City Revitalization*). Best of all, waterfront sites have a strong sense of place, with long views, milder microclimates and opportunities for long, uninterrupted walks that one rarely finds elsewhere in most cities.

Waterfront planners faced enormous difficulties in redeveloping these sites. The port authorities and railways withheld their lands and interjurisdictional conflicts impeded planning. Conflicting demands for economic development, parks, and jobs increased political difficulties. The sites were often polluted and encumbered with industrial structures like grain elevators, which have heritage value but are difficult to reuse. Port lands had inadequate local utilities and were often cut off from the CBD by expressways and railway lines, requiring large early infrastructure investments.

3. Waterfront Planning Practice

While most nineteenth century cities considered their waterfronts as industrial districts to be exploited for maximum economic development, some shoreline was improved for public uses. The 1909 *Plan of Chicago* was a leader in this approach, reserving the entire shoreline of Lake Michigan for public parks. Gracious waterfront parks and promenades were sometimes added during infrastructure projects, using fill to extend the land adjacent to the downtown. New York's Battery Park, London's Victoria and Albert Embankments, and the Seine esplanade in Paris were welcome improvements. Waterfront projects by F. L. Olmsted, especially the Charles/Muddy River improvements in Boston, would prove influential as a result of their natural systems approach.

New planning techniques evolved for 1970s waterfront redevelopment projects. Most projects were implemented by a redevelopment authority, although the level of government which controlled the agency depended upon who was willing to make the early infrastructure grants. Good implementation practice required these agencies to manage three areas: politics, finance, and urban design.

3.1 Politics

Excellent long-term political management is required for waterfront redevelopment except in rare cases such as the Barcelona Olympics or Lisbon World's Fair, where deadlines and national pride pushed the projects forward in record time. Most urban waterfronts take decades to rebuild because the sites are large and complex. It often took five to ten years to build the development coalition needed to have a plan approved. Good relations with local governments was an advantage, since they often controlled infrastructure needed for redevelopment. Unilateral action by a senior government often led to long-term delays.

Good political management included:

(a) good relations and no surprises for sponsoring governments,

(b) board of directors well connected to all levels of governments,

(c) strong links to local government at staff and board levels,

(d) good relations with local residents, and

(e) linking private development with public benefits, especially waterfront parks.

3.2 Finance

Waterfront redevelopment requires start-up capital investment for land assembly, site clearance, environmental remediation and infrastructure. These costs make the projects unattractive to the private sector, since it takes years before political consensus is reached and land is ready for building. Substantial government start-up grants were required for most waterfront projects in the 1970s, while many governments were reducing funds for urban redevelopment. Many agencies therefore pursued investment by developers using some form of public–private partnership. The agency could assemble and clear land, negotiate planning approvals and install infrastructure. Private developers were solicited to complete the project. Private investment started late, built slowly, and cycled with the local property market. Flagship projects, such as the World Financial Center (New York) or Canary Wharf (London) often did not arrive until an agency had demonstrated credibility with site improvements, small developments and changing the waterfront image. Big projects also waited for a positive point in the real estate cycle, or risked bankruptcy like Canary Wharf.

Good financial planning practice included:

(a) access to long-term funding for infrastructure,

(b) streamlined developer selection process,

(c) incorporating recessions into financial plans, and

(d) some infrastructure and social housing projects waiting for recessions.

New York's Battery Park City Authority set a good example in the use of bonds to finance capital investment.

3.3 Urban Design

The most important element of waterfront urban design is to provide continuous, public access to the water's edge. A high-quality pedestrian promenade along the shoreline is appreciated by the public and adds value to adjacent property. Political controversies erupted in London and Boston when developers seized ownership of the water's edge, while the Toronto agency was dismantled after delays in delivering waterfront parks.

Urban design for waterfront redevelopment should facilitate over, several decades, implementation of:

(a) small development increments,
(b) tight phasing plans,
(c) simple infrastructure that can be phased,
(d) adapt existing infrastructure and buildings,
(e) continuous public access to the water's edge,
(f) view corridors down streets to the water.

Comprehensive schemes designed by one architect were discredited by early urban renewal projects. The 1979 plan for Battery Park City in New York was influential for its approach and urban design guidelines. The plan's extension of the Manhattan street grid allowed flexibility, easy phasing, small increments and simple infrastructure. The agency's commitment to creating quality public space before development made it a leader in urban waterfront design.

The waterfront redevelopment agencies learned from each other as they developed a new planning model. Baltimore, Boston, and Toronto were influential in the 1970s; the first two for festival marketplaces developed by the Rouse Corporation, and the latter for public programming facilities in recycled industrial buildings. These ideas were quickly copied around the world. A waterfront formula emerged: a festival marketplace, an aquarium, office buildings, condominium apartments for young urban professionals and perhaps a hotel. Like most formulae for planning, it did not replicate well in other countries and cultures, leading to poststructuralist criticism of global corporations overwhelming the local population while providing few benefits for the displaced industrial workers (Harvey 1989).

The best waterfront planning pays close attention to a local sense of place and waterfront-related uses (Breen and Rigby 1994, 1996). While an authentic place is hard to define, plans which renovate buildings, keep waterfront uses (fishing fleets and marinas), and provide walkways and parks seem to be the most successful. Vancouver's Granville Island demonstrated adaptive reuse of modern industrial buildings, while London's Hays Galleria and Venice's Arsenal renovated historic warehouses. At a regional scale, the best waterfront planning uses ecosystem principles to preserve shoreline environmental integrity, while providing public access and trails.

Completing redevelopment of big sites like the London Dockland may take 20 years. Meanwhile, a new generation of waterfront sites is emerging. Large waterfront railway yards, such as San Francisco's Mission Bay, New York's Riverside South and Vancouver's False Creek may offer a different planning challenge. Since these lands are typically owned and developed by large private companies, planning initiative is falling more to the private sector than did the 1980s public–private partnerships. While these new developers may require less public subsidy, it remains to be seen whether public benefits such as affordable housing and open space can be secured from this new generation of waterfront redevelopment.

See also: Environmental Planning; Planning, Politics of; Regional Planning, History of; Transportation Planning; Urban Planning: Central City Revitalization; Urban Planning: Growth Management; Urban Planning: Methods and Technologies; Urban Planning (Western), History of

Bibliography

Breen A, Rigby D 1994 *Urban Waterfronts: Cities Reclaim Their Edge*. McGraw Hill, New York

Breen A, Rigby D 1996 *The New Waterfront: A Worldwide Urban Success Story*. McGraw Hill, New York

Bruttomesso R 1993 *Waterfronts: A New Frontier for Cities on Water*. International Centre Cities on Water, Venice

Chilcote P 1988 The containerization story. In: Hershman M (ed.) *Urban Ports and Harbor Management*. Taylor and Francis, New York

Gordon D 1997 *Battery Park City: Politics and Planning on the New York Waterfront*. Gordon and Breach, New York

Hall P 1991 *Waterfronts: A New Urban Frontier*. University of California Institute of Urban and Regional Development, Berkeley, CA

Harvey D 1989 *The Condition of Post-Modernity*. Johns Hopkins University Press, Baltimore, MD

Hoyle B, Pinder D, Hussain M 1988 *Revitalising the Waterfront: International Dimensions of Dockland Development*. Bellhaven, London

Wrenn D 1983 *Urban Waterfront Development*. Urban Land Institute, Washington, DC

D. Gordon

Watson, John Broadus (1878–1958)

J. B. Watson the founder of the behaviorist school of psychology was born on January 9, 1878 in Travelers Rest, a poverty stricken village in South Carolina. His

father deserted the family when he was 12. His mother, a deeply religious woman, sold what little they had and moved to Greenville, a larger town, in the hope of providing better education for the children. In 1894, through the help of their church, John was admitted to the Baptist Furman College, from which he graduated with an M.A. degree in 1900. He was then admitted to the University of Chicago where 3 years later he became the youngest person to receive a Chicago Ph.D. He remained there until 1908 as Instructor and assistant to James Angell, the eminent philosopher. In 1904 he married Mary Ickes who was the youngest sister of Harold Ickes, the influential Secretary of the Interior in Roosevelt's 'new deal' cabinet. Harold disapproved of the marriage, and may have played a part in the problems Watson was to face later. In 1908 he moved to Johns Hopkins University as Assistant Professor and Chair of Psychology. Almost all of Watson's seminal work was produced in the relatively short span of the ensuing 12 years. In 1920 his academic career came to a premature end when he developed a love affair with a young student, Rosalie Rayner. Anticipating that the affair may lead to unwanted publicity for the university, Watson decided to inform the president of the university. At the interview the president demanded his immediate resignation which Watson instantly wrote. The anticipated scandal broke in newspapers some three months later, when his wife Mary filed for divorce. As soon as the divorce proceedings were complete, on December 31, 1920, Watson and Rosalie were married. Watson was penniless, having left everything he owned to Mary and Rosalie had been disowned by her parents, who disapproved of the marriage. For a time they stayed with friends in New York. Eventually, Watson was offered a position in the advertising agency, J. Walter Thompson Company. In accord with the company policy that all employees must have personal experience of selling, Watson was sent to New Orleans to sell fishermen's boots door-to-door. Upon returning to headquarters his ideas and proposals were so successful that within two years he had risen to the position of Vice President. Having reached the top of his profession, fallen, and risen to the top of another profession, Watson was to suffer yet another tragedy. In 1935 Rosalie died of pneumonia. She was 37. He was never to recover from this loss. He died a disappointed, depressed man on September 25, 1958 at, ironically, the Johns Hopkins Hospital in Baltimore.

In harmony with the spirit of his times Watson believed in the supremacy of science, and its power for resolving our social and personal problems. Psychology, he thought, must become a 'natural science.' His seminal paper, *Psychology as the behaviorist views it*, sometimes referred to as 'the behaviorist manifesto,' was published in 1913. Its essential arguments were that to be a natural science psychology must adhere to the principles that are basic to all sciences, including the principle that what is studied must be public, i.e. observable and replicable by more than one person. Therefore, the subject matter of psychology should be behavior and nothing else. These views came under severe attack from the psychological establishment for two main reasons: first, they implied that the work of the past several decades, based on the method of introspection, had to be discarded or at best, disregarded; and second, they conflicted with the established view that psychology studies the mind, the psyche, feelings, thinking, and the like. If psychology only studied behavior, what was to be made of such concepts? Watson's answer was that these concepts had no place in science. Thus, the controversy over behaviorism arose not as much from the proposal that the science of psychology should study behavior, as it did from Watson's insistence that it should study nothing else.

Watson wrote many scholarly articles, monographs, books, and popular articles developing his view of the behaviorist science of psychology. He believed the function of psychology is to 'predict and control' behavior which, when fully achieved, would result in the good society. He wrote 'behaviorism ought to be a science that prepares men and women for understanding the principles of their own behavior ... I wish I could picture for you what a rich and wonderful individual we should make of every healthy child if only we could let it shape itself properly and then provide for it a universe in which it could exercise that organization (1925). With reference to the introspectionists' efforts to obtain verbal reports of phenomena that occur 'inside' a person, he asserted that such events are no different from events outside the body and they are, just as behavior, fully amenable to objective, scientific investigation. Language is a form of behavior and, therefore, subject to behaviorist study. Personality is the result of 'what we start with and what we have lived through (1919). Since, as far as it was known at the time, what we start with could not be changed, personality could only be developed or changed by appropriately arranging the environment. Watson's statement 'Give me a dozen healthy infants, well-formed, and my own specified world to bring them up in and I'll guarantee to take any one at random and train him to become any type of specialist I might select—doctor, lawyer, artist, merchant-chief, and yes, even beggarman and thief, regardless of his penchants, tendencies, abilities, vocations, and race of his ancestors' is quoted in almost every textbook. However, the very next sentence is 'I know I am going beyond my facts and I admit it, but so have the advocates of the contrary and they have been doing it for many thousands of years (1925). The apparently deliberate omission of this sentence by his critics unfairly suggests an extreme claim that Watson did not make.

After his departure from academic life Watson continued to lecture, supervise research, and write on

psychology. Between 1922 and the death of Rosalie he conducted a course on behaviorism at New School for Social Research, corresponded with Bertrand Russell, Yerkes, Titchener and others, many of whom contributed to his course as guest lecturers. His book *Behaviorism* (1925) was originally written as lectures for this course. His research with 'Little Albert' where he and Rosalie conditioned the fear response in the boy to furry objects is often cited, but his research with a large number of other infants is less known. In later studies he supervised the elimination of the fear response in doctoral research by Mary Cover Jones. His views on child rearing were extreme, arguing that the development of emotional ties with infants should be avoided and that parents should not hug and kiss their children. He advised parents to follow this, in a popular book written with Rosalie Watson (1928). Some years later he changed his position on this and expressed regret that he had advocated it.

The roots of behaviorism go back much earlier than Watson. Lloyd Morgan's book, *Introduction to Comparative Psychology* (1894) sets the beginning of 'animal psychology' which necessarily required research into behavior. The assumption that the principles discovered through research into animal behavior will also hold for humans began to gain strength. Further, Pavlov's discoveries of conditioning, Darwin's theory of evolution indicating a continuity across species, and Thorndike's theory of learning based on research into animal behavior contributed to a *Zeitgeist* for the emergence of behaviorism. Watson's doctoral research which was entitled *The psychic development of the white rat correlated with the growth of its nervous system* (published as *Animal Education*) emerged in this context. His major contributions were to articulate the prevailing ideas, explicitly to formulate behaviorism, and to elaborate the ideas and principles that follow from it.

Watson regarded himself primarily a researcher and often referred to himself as 'a laboratory man.' He was interested in broad range of phenomena and in finding new ways of investigating them. It is little known that he was a pioneer of what later came to be established as Ethology. Beginning in 1906 he spent two summers on the Dry Tortugas islands, off Florida Keys, studying the behavior and homing patterns of noddy and sooty terns. Between 1907 and 1915 he published eight papers on the homing behavior and spectral sensitivity of birds. He was also a pioneer of what is now known as behavioral pharmacology. In his last year at Johns Hopkins he had begun to investigate the behavioral effect of ingested alcohol, with the collaboration of Edward Thorndike at Columbia University. The project was abandoned when he resigned. In his second profession in advertising he continued to develop innovative ideas, and pioneered the technique of associating an attractive image with the product even if they are not related intrinsically. The first of these, showing pictures of attractive women in an advertisement for Pond's face creams was very successful, and the technique became established, lasting to the present.

In his relatively short span of years in academia Watson became one of the most influential individuals in psychology. He was president of the American Psychological Association in 1915, editor of *The Psychological Review, The Journal of Animal Behavior,* and *The Journal of Experimental Psychology* that he founded. His influence on psychology has been broadly profound and enduring, but seldom recognized. Throughout his career Watson maintained the practice of not replying to his critics. Nevertheless, the controversy over behaviorism continued unabated, and in time detracted from recognition of Watson's scientific researches. This, combined with his departure from academia in mid-career, produced the unfortunate result that all of his ideas about the science of psychology, personality, language, and his researches into conditioning were taken and developed by his contemporaries and by later psychologists, notably by B. F. Skinner, without due acknowledgment. Nevertheless, the signs of Watson's influence in the shaping of contemporary experimental psychology remain evident.

Bibliography

Morgan C L 1894 *Introduction to Comparative Psychology*. W Scott, London

Watson J B 1913 Psychology as the behaviorist views it. *Psychological Review* **20**: 158–77

Watson J B 1919 *Psychology from the Standpoint of a Behaviorist*. Lippincott, Philadelphia, PA

Watson J B 1925 *Behaviorism*. Norton, New York

Watson J B, Watson R R 1920 Conditioned emotional reactions. *Journal of Experimental Psychology* **3**: 1–14

Watson J B, Watson R R 1928 *Psychological Care of Infant and Child*. W W Norton, New York

P. Harzem

Wayfinding and Navigation Behavior

Navigation is the purposeful movement through an environment, such as walking through a neighborhood or driving through a town. The task of navigating is achieved through the combination of locomotion and wayfinding. Locomotion refers to the physical movement through space, perceiving objects and where they are in the environment, and moving toward a visible place or object while avoiding various obstacles. Examples of locomotion include walking through a large shopping mall from store to store, or driving a car through a city center. Wayfinding involves the thought processes and thinking that enables movement to distant locations—locations that cannot be

seen from the current position in the environment. Wayfinding activities include choosing routes for travel, planning trips, and making estimations of distance, estimations of travel time, and orientation estimates to distant, unseen locations.

Social scientists are interested in navigation and wayfinding from various perspectives. Psychologists study navigation and wayfinding as a cognitive representation, exploring the various ways people may mentally represent the world around them (Tversky 2000). Landscape architects explore navigation and wayfinding to gain insight into the design of environments such as buildings, neighborhoods, and towns that are aesthetically pleasing and easy to navigate (Lynch 1960). Economists study navigation and wayfinding to understand people's shopping patterns and behaviors, and from this determine locations for distribution of goods and services that optimize profits (Fotheringham and Curtis 1992). Developmental psychologists test children's navigation and wayfinding abilities to understand which skills are acquired at what age, and from this develop curriculum materials (Hart and Moore 1973). Anthropologists study navigation and wayfinding behavior in previous and present cultures to explain hunting and foraging behavior as well as settlement patterns. Geographers' interest in navigation and wayfinding behavior interleaves with many other social scientists, studying all of the examples listed above. In addition, geographers explore methods for visualizing this spatial information in the form of paper and digital maps, written and spoken directions, onboard vehicle navigation aids, and virtual reality spaces (MacEachren 1995).

This article distinguishes between different wayfinding tasks, describes the strategies most frequently used to wayfind, and describes various systems of reference used to maintain orientation while wayfinding.

1. *Wayfinding Tasks*

Allen (1999) identified three categories of wayfinding tasks that he termed the commute, the explore, and the quest. The first, and most common type, is the commute. A commute involves traveling between two known places along a familiar route. This form of wayfinding is routine, where success in arriving at a destination is measured in terms of travel time (shortest time) and the amount of effort to negotiate the route (the fewest stops and turns). Little attention is required to accomplish this wayfinding task, which is most often associated with a person's travel from home to work, home to school, or home to a store. This wayfinding behavior is studied, for example, to monitor and control automobile traffic, and plan for new or upgraded roads.

The second wayfinding task is the explore. An explore involves travel through an unfamiliar area with the intent of learning about the unknown environment. The start and destination of an explore are often known places, typically the same place, but routes traveled while on an explore are new. Explores are not routine, and require strategic planning of places to visit and routes to travel to avoid becoming lost. This wayfinding task would be typical of tourists visiting a city for the first time, and would be considered in the design of tour maps and tourist information to assist the visitor.

The third wayfinding task is the quest. The quest, like an explore, begins at a familiar place, but ends at an unfamiliar destination. In this case, the destination is known to exist, but has not previously been visited. Information about the spatial relationship between the beginning and ending points of a quest are provided in the form of directions, typically maps or verbal directions. Cartographers have placed a great deal of effort in the design of navigation aids for quests, from simple verbal directions, to the standard paper road map, to the design of sophisticated vehicle navigation aids for automobiles.

2. *Wayfinding Strategies*

Though there are only three classes of wayfinding tasks, there are various strategies used to carry out these tasks. These strategies include piloting, following marked trails, habitual locomotion, dead reckoning, path integration, and lastly the use of internal representations. These strategies, though presented here as separate individual strategies, can be used together to carry out a wayfinding task. In addition, a particular strategy can be used to carry out more than one wayfinding task.

Piloting relies on the use of landmarks for orientation. The immediate goal in piloting is to traverse from the present landmark to the next landmark. The present landmark serves as a basis for orientation by providing a link back to the familiar environment, and information on where to go next. The spatio-temporal sequence of landmarks provides information on wayfinding progress, indicating how much of the trip has been completed and how much remains. Piloting is a strategy typically used in explores by tourists, or a person recently locating to a new city.

Following marked trails entails wayfinding that relies on a system of trail markers to highlight a route. Marked trails can include a road network with road signs and mile markers, multiple walking paths through a wooded area, or color-coded lines along the floor of a corridor in an office complex. Marked trails are becoming critical in the design of large build environments, such as hospitals, to aid people in finding a particular unit, such as X-ray or pediatric cardiology, while under stressful situations (Carpman and Grant 1997).

Habitual locomotion, on the other hand, is a strategy that develops after repeated navigation

through an environment, resulting in what has been referred to as route knowledge. Generally, habitual locomotion develops from piloting strategies, and after repeated exposure, the sequence of landmarks and the actions to be taken at each landmark are remembered. With increasing exposure to a route, habitual locomotion results in the development of commutes.

Path integration strategies, unlike habitual locomotion or piloting, are not based on the current landmark, or on the sequence of landmarks, but instead are based on orientation to the starting point of a route. Path integration enables travel directly back to the starting point along a novel route, after having traveled a complex route directed away from that point. This strategy necessitates constant knowledge of the direction or orientation back to the starting point. This wayfinding strategy is difficult, and is not common for humans in everyday wayfinding tasks, but instead is a common behavior in nonhuman species, such as honey bees and rats, studied extensively by behavioral psychologists (Gould 1986, Tolman 1948). It is also a strategy used in the sport of orienteering, where participants use a map and compass to navigate an unknown wilderness course.

Dead reckoning is a wayfinding strategy that entails first determining the direction to a destination, and then systematically searching until the destination is reached. Methods for orientation for dead reckoning generally include reliance on solar and lunar information, auditory information, and vestibular and proprioceptive information. Some cultures studied by anthropologists are able to use wind and water currents for orientation. Reliance on dead reckoning requires that the navigator perform periodic updates on position, making small corrections in distance and direction estimates while moving through the environment.

Wayfinding based on use of an internal representation requires reliance on a 'cognitive map' in which relative positions of landmarks, routes between landmarks, and the orientation and distances between landmarks are represented mentally. This spatial knowledge is referred to as survey or configurational knowledge. This is the most complex and sophisticated wayfinding strategy. Geographers and psychologists have long studied internal representations and the implication these representations have on navigation and wayfinding behavior; (for overviews see Downs and Stea 1973, Golledge 1999, Kitchin and Freundschuh 2000, Moore and Golledge 1976).

3. *Systems of Reference*

People rely on systems of reference, or reference frames, to stay oriented during navigation and wayfinding. Without reference frames, one would be lost. A reference frame can be as simple as knowing the starting point of an exploration and the direction of travel, or tying pieces of cloth to trees in the woods to mark a traversed path. In contrast, reference frames can be as complex as measurement of latitude and longitude positions, or the use of other abstract coordinate systems such as the Universal Transverse Mercator Grid. An effective reference system has a known origin (starting point) and a system for representing position, and direction (e.g., left/right, north/south, bearing of 90°).

There are three types of reference frames: egocentric, allocentric, and geocentric. An egocentric frame of reference is one in which the body is used as the point of reference. In this case, wayfinding decisions are based on the body's current location while moving through the environment. When a decision point is reached—a point in the environment where a choice is made to go straight, to turn left or right, or to turn around—the resulting choice is based on one's current location rather than on some other object or place.

When orientation is based on objects or places, such as objects in a room, or a plaza in a city, orientation is based on an allocentric reference frame. These objects and places that are used for orientation are called landmarks. A landmark is an object or place that stands out from its surroundings, holding a special significance for the navigator. Examples of landmarks include a chair, a street intersection, a unique building, a road, a region, or a mountain. Not only do landmarks serve as organizing features for the navigator, they can also have certain actions associated with them. For example, a street intersection may prompt a driver to turn left to go home, or turn right to go to a store. Reliance on an allocentric reference frame enables the navigator to travel novel routes, to take a different route back to the starting point.

A geocentric reference frame relies on abstract reference systems for orientation. These systems are called abstract because they do not exist in the environment. They are products of human invention, superimposed on the environment and used to specify location with a coordinate. Unlike egocentric and allocentric reference frames, geocentric reference frames must be learned through education and training. Examples of geocentric reference frames commonly used for orientation include the Geographical Grid (latitude and longitude) and the Universal Transverse Mercator system. Like other reference systems, the basis for location in a geocentric reference system is an origin. For the Geographical Grid, the origin is the intersection of the equator (0° latitude) and the line of longitude that bisects the Royal Observatory in Greenwich, UK (0° longitude). All measurements for latitude and longitude with the USA (and other Western countries) are made in reference to that origin (for a full discussion of reference systems, see Muehrcke and Muehrcke 1992). The basis for measurement using the Geographical Grid is the

calculation of angles, directions, distances, and the curvature of the earth. Technology has simplified the measurement of latitude and longitude with the development of global positioning systems (GPS). GPS can measure location on the earth's surface quickly, and with reasonable accuracy. The low cost of GPS systems has enabled a proliferation of these measurement devices, enabling almost anyone to find their location on the earth.

Bibliography

Allen G L 1999 Spatial abilities, cognitive maps, and wayfinding: Bases for individual differences in spatial cognition and behavior. In: Golledge R G (ed.) *Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes*. Johns Hopkins University Press, Baltimore, MD, pp. 46–80

Carpman J, Grant M 1997 Wayfinding in healthcare facilities. In: Marberry S (ed.) *Healthcare Design: An Introduction*. Wiley, New York

Downs R M, Stea D (eds.) 1973 *Image and Environment*. Aldine, Chicago

Fotheringham A S, Curtis A 1992 Encoding spatial information: The evidence for hierarchical processing. In: Frank A U, Campari I, Formentini U (eds.) *Theories and Methods of Spatio-temporal Reasoning in Geographic Space*. Springer-Verlag, Berlin, pp. 269–87

Golledge R G (ed.) 1999 *Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes*. The Johns Hopkins University Press, Baltimore, MD

Gould J L 1986 The locale map of honeybees: Do insects have cognitive maps? *Science* **232**: 861–3

Hart R A, Moore G T 1973 The development of spatial cognition: A review. In: Downs R M, Stea D (eds.) *Image and Environment*. Aldine, Chicago, pp. 246–88

Kitchin R, Freundschuh S M (eds.) 2000 *Cognitive Mapping: Past, Present and Future*. Routledge, London

Lynch K 1960 *Image of the City*. MIT Press, Cambridge, MA

MacEachren A M 1995 *How Maps Work: Representation, Visualization and Design*. Guilford Press, New York

Moore G T, Golledge R T (eds.) 1976 *Environmental Knowing*. Dowden, Hutchinson and Ross, Stroudsberg, PA

Muehrcke P, Muehrcke J 1992 *Map Use; Reading, Analysis, and Interpretation*. JP Publications, Madison, WI

Tolman E C 1948 Cognitive maps in rats and men. *Psychological Review* **55**: 189–208

Tversky B 2000 Levels and structure of spatial knowledge. In: Kitchin R, Freundschuh S M (eds.) *Cognitive Mapping: Past, Present and Future*. Routledge, London, pp. 24–43

S. M. Freundschuh

Wealth Distribution

Wealth represents a stock of accumulated assets; income represents a flow of current output. Families not only receive income over the course of a year but also save part of their income in the form of housing, savings accounts, stocks, bonds, and the like. Such accumulated savings are referred to as wealth. The first part of this article explains why wealth, like income, has an important bearing on the well-being of families and individuals. The second part develops the concept of household wealth; the third discusses some of the problems inherent in its measurement; the fourth shows how the structure of household wealth has changed over time; and the fifth presents time trends in the inequality of wealth. Data drawn from the USA will be utilized to illustrate parts of the discussion in the text.

1. Wealth and Well-being

Why are we interested in household wealth? Most studies use income as a measure of family well-being. Though certain forms of income are derived from wealth, such as interest from savings accounts and bonds, dividends from stock shares, and rent from real estate, income and wealth are by no means identical. Many kinds of income, such as wages and salaries and transfer payments, are not derived from household wealth, and many forms of wealth, such as owner-occupied housing, produce no corresponding income flow.

Moreover, family wealth by itself is also a source of well-being, independent of the direct financial income it provides. There are four reasons for this. First, some assets, particularly owner-occupied housing, provide services directly to their owner. This is also true for consumer durables, such as automobiles. Such assets can substitute for money income in satisfying economic needs. Families with the same money income but differing amounts of housing and consumer durables will have different levels of welfare.

Second, wealth is a source of consumption, independent of the direct money income it provides. With the possible exception of consumer durables, assets can be converted directly into cash and thus provide for immediate consumption needs. This is also true for the equity in owner-occupied housing, because second mortgages and home equity loans are another source of credit.

Third, the availability of financial assets can provide liquidity to a family in times of economic stress (such as occasioned by unemployment, sickness, or family break-up). In this sense, wealth is a source of economic security for the family over and above the income that it directly provides.

Fourth, large fortunes can be a source of economic and social power that is not directly captured in annual income. Large accumulations of financial and business assets can confer special privileges to their holders. Such large fortunes are often transmitted to succeeding generations, thus creating family 'dynasties.' They can also enable individuals to influence the political process through large donations to candidates

running for public office. Thus, wealth holdings provide another dimension to household welfare over and above income flows.

2. What is Household Wealth?

The conventional definition of household wealth includes assets and liabilities that have a current market value and that are directly or indirectly marketable. A typical list of assets includes owner-occupied housing and other real estate; cash, checking and savings accounts; bonds and other financial instruments; corporate stock; equity in unincorporated businesses; and trust fund equity. This measure will be referred to as 'marketable household wealth' (or simply 'household wealth'), since it represents those assets over which the family or individual has control. This notion corresponds to wealth as a store of value and is used in conjunction with a standard national accounting framework.

Table 1 presents the balance sheet for the US household sector in 1989. The wealth accounts are first divided into two major categories: assets and liabilities. Assets consist of items owned by households which have a positive value to them, while liabilities are debts, which by definition have a negative value. The difference between total assets and total liabilities

Table 1
Balance sheet of total household wealth in the USA, 1989 (billions of dollars)

Panel A: Marketable net worth	Value of component	Percent of total assets
I. Total assets	20,268	100
A. Tangible assets	8,960	44
1. Owner-occupied housing	6,118	30
2. Other real estate and undeveloped land	2,842	14
B. Financial assets	4,817	24
1. Checking accounts, cash, and currency	497	2
2. Time deposits, money market funds, CDs, IRAs, Keogh plans, and other retirement accounts	3,280	16
3. Bonds and financial securities	690	3
4. Life insurance cash surrender value	351	2
C. Equities	5,490	27
1. Stocks and mutual funds	1,394	7
2. Unincorporated business equity	3,478	17
3. Trust equity	618	3
D. Miscellaneous assets[a]	1,001	5
II. Liabilities	3,040	15
A. Mortgage debt	2,470	12
B. Other debt	571	3
III. Net worth	17,228	85
Memo:		
A. Home equity[b]	4,367	22
B. Pension fund reserves	2,857	14
C. Consumer durables	1,934	10
Panel B: Other concepts of household wealth	**Value of component**	**Ratio to net worth**
Marketable wealth	17,228	1.00
Marketable wealth plus consumer durables	19,162	1.11
Marketable wealth minus home equity	12,861	0.75
Augmented wealth	28,670	1.66
A. Pension wealth	3,540	0.21
B. Social security wealth	7,902	0.46

Sources: 1989 Survey of Consumer Finances (tape); Federal Reserve Board Flow of Funds; and Wolff (1992).
a Gold and other precious metals, royalties, jewelry, antiques, furs, loans to friends and relatives, future contracts, and miscellaneous assets. b Gross value of owner-occupied housing less outstanding mortgage debt.

is called net worth or, simply, wealth. In 1989, total household assets amounted to 20.3 trillion dollars, total liabilities were 3.0 trillion, and household net worth was 17.2 trillion.

Assets are divided into three main groups. The first, tangible assets, include stocks of real goods that are both 'reproducible,' such as houses, and 'nonreproducible,' such as land. In 1989, total tangible assets owned by households were worth 9.0 trillion dollars, which amounted to 44 percent of total household assets. The major tangible asset owned by households is owner-occupied housing (that is, houses that are lived in by their owner). This accounted for 30 percent of total household assets. The second component consists of land, apartment buildings, other rental property, and other real estate held by households. This group made up 14 percent of all assets.

The second major group is referred to as financial assets. These are assets which function as money or are readily convertible to money. They include cash, demand deposits (checking accounts), and time and savings deposits, which function directly as a means of exchange. They also include bonds, notes, and financial securities issued mainly by corporations and the government. Such securities are 'promissory notes,' by which the borrower agrees to pay back the lender a certain amount of money (the principal plus interest) at a certain date. The fourth component in this group is the 'cash surrender value' of life insurance funds. Certain life insurance plans (called 'full life insurance') allow individuals to contribute to a savings fund, and the amount the individual can withdraw from this savings fund is referred to as its cash surrender value.

In 1989, financial assets constituted 24 percent of household assets. Demand deposits and currency together amounted to only 2 percent of total assets, while time and savings deposits together with CDs, money market funds and pension accounts constituted 16 percent. Bonds and other securities amounted to 3 percent, while the cash surrender value of insurance plans was only 2 percent.

The third major group consists of equities or ownership rights. A corporate stock certificate issued by a company represents ownership of a certain percentage of the company's assets. Unincorporated business equity refers to small businesses (such as a farm or a store) owned directly by individuals, in contrast to corporations which are owned through stock shares. A person's equity in an unincorporated business is the value of the business. The third item, trust fund equity, refers to bank deposits, securities, corporate stock, and other financial instruments, which are held in a special legal arrangement called a trust fund. In a typical trust, the actual assets are managed by a specially named administrator (often a bank), and the income earned from the assets in the trust are remitted to individual beneficiaries. Trust funds held for individuals are counted as part of household wealth.

As a group, equities comprised 27 percent of total household assets in 1989. Corporate stock shares made up 7 percent of total household assets, unincorporated business equity 17 percent, and trust fund equity 3 percent.

On the liability side, the major form of household debt is home mortgages. A mortgage is a loan, issued normally by a bank and usually for a period from 15 to 30 years, that is used to finance the purchase of real property, particularly houses to live in. In 1989, mortgage debt comprised 81 percent of total household debt. The remaining 19 percent consisted of other household debt, including automobile and other consumer loans and credit card debt. Together, total household debt amounted to 15 percent of the value of total household assets.

Three other entries are shown in the bottom of Table 1. The first of these is home equity, defined as the gross value of owner-occupied housing less outstanding mortgage debt. This represents the amount of money a family could net from the sale of the house (after paying off its outstanding mortgage balance). Home mortgages amounted to 29 percent of the value of owner-occupied housing. As a result, home equity made up a smaller proportion of total assets than the gross value of owner-occupied housing—22 vs. 30 percent.

The second entry is the reserves (assets) of pension funds. These funds are accumulated mainly by corporations in order to pay pension benefits for their workers when they retire. In 1989, they amounted to 14 percent of total household assets. Technically speaking, these reserves are not part of marketable household wealth, since households have no control over the assets in these funds and cannot directly draw on these funds. However, these reserves are held for the benefit of individuals and therefore have some of the characteristics of a trust fund. The third category, called consumer durables, includes automobiles, washing machines and other appliances, furniture, and other long-lasting household equipment. Cars are the major component of household durables, and altogether durables amounted to 10 percent of total assets.

A theme that regularly emerges in the literature on household wealth is that there is no unique concept or definition of wealth that is satisfactory for all purposes. One alternative definition is the sum of marketable wealth and consumer durables. This concept is useful because consumer durables provide consumption services directly to the household, though they are not easily marketed and the resale value of these items typically far understates the value of their consumption services to the household. As shown in Panel B, the sum of marketable wealth and consumer durables amounted to 19.2 trillion dollars in 1989, which was 11 percent greater than marketable wealth.

A second is defined as marketable net worth minus net equity in owner-occupied housing. This definition

represents a more 'liquid' concept than marketable wealth, since one's home is difficult to convert into cash in the short term. It thus reflects the resources that may be directly available for consumption or various forms of investments in the short term. This concept totaled 12.9 trillion dollars in 1989, 75 percent of marketable wealth.

One of the major developments since the end of World War II has been the enormous growth in both public and private pension systems. Even though such pension funds are not under the direct control of individuals, they are a source of future income to families and thus may be perceived as a form of family wealth. Moreover, as Martin Feldstein (1974) has argued, insofar as families accumulate 'traditional' wealth to provide for future consumption needs, the growth of such pension funds may have offset private savings and hence traditional wealth accumulation.

A wider definition of household wealth will thus add some valuation of pension rights, from both public and private sources, to marketable wealth. Such a measure provides a better gauge of potential future consumption. Augmented household wealth is defined as the sum of household marketable wealth, social security wealth, and pension wealth. Social security wealth is defined as the present discounted value of the stream of future social security benefits, and pension wealth is defined as the present discounted value of future pension benefits (see Wolff 1992, for a discussion of measurement issues). As is apparent from the definitions, these two forms of wealth are not fungible or marketable, since individuals cannot convert these assets into cash.

What is the rationale for including these two forms of wealth? The main motivation is that, from the standpoint of an individual, a future guaranteed flow of income is in many ways like owning a financial asset. Indeed, there are marketable assets called 'annuities' which have precisely the characteristic of providing a steady stream of income after a certain time (or age) is reached. The anticipated social security (or pension) benefits that an individual will receive after retirement is comparable to such an annuity. From the individual's point of view, then, a guaranteed stream of future pension or social security benefits may be as much a form of wealth as money put into a savings account that will be drawn on after retirement.

According to my estimates, total pension wealth in 1989 was 3.5 trillion dollars and total social security wealth was 7.9 trillion dollars (see Panel B). Together, they amounted to 66 percent of marketable net worth. Social security wealth by itself was thus larger than any other component of household wealth, including owner-occupied housing and total bank accounts.

3. *Problems in Measuring Household Wealth*

There have been four principal sources of data for developing household wealth estimates: (a) estate tax data; (b) household survey data; (c) wealth tax data; and (d) income capitalization techniques. Each has its characteristic advantages and disadvantages.

3.1 *Estate Tax Data*

Estate tax data was the first major source of data used for wealth analysis. When someone dies, that person's assets are said to comprise his or her estate, and estate tax records are actual tax returns filed for probate. Such data have a great degree of reliability, since they are subject to scrutiny and audit by the state. Their main limitation in the USA, at least, is that the threshold for filing is relatively high, so that only a small proportion of estates (typically 1 percent or so) are required to file returns. In the UK the threshold is considerably lower, so that the majority of estates file tax returns. This is also true for some states in the USA, as well as Washington, DC, which have their own estate tax requirements.

Another difficulty with this data source is that the sample consists of decedents. Since most researchers are interested in distribution of wealth among the living population, a technique based on 'mortality multipliers' is used to infer the distribution of wealth among the actual population. If mortality rates were the same for each group, then the wealth of decedents would constitute a representative sample of the living population, and researchers could use the estate data directly. However, mortality rates are much greater for older people than younger people, so that different 'weights' must be given to the estates filed by decedents of different ages. Mortality rates are also higher for men than for women (that is, women, on average, live longer than men), and also higher for black individuals than whites. As a result, estates are assigned a weight based on the inverse of the group's mortality rate, and these are then used to generate the size distribution of wealth in the living population.

Estimates of wealth inequality based on this technique are, as might be expected, quite sensitive to the precision of the mortality multipliers. The estimates can have a very large standard error, particularly for the young, since there are very few of them in the sample. This means that the results are very sensitive to who happens to die in a given year and may therefore not be very reliable for young adults. There are two other problems associated with this technique. First, insofar as mortality rates are inversely correlated with wealth (that is, the rich tend to live longer), the resulting multipliers can be biased (see Shorrocks 1975). Most studies do try to correct for this problem by using mortality rates for the wealthy that are lower than for the population as a whole, due to the longer expected life span of the wealthy.

Second, the distribution of wealth estimated by this technique is for individuals rather than for families. Changing ownership patterns within families (for example, joint ownership of the family's house) can

affect estimated wealth concentration. For example, marital customs and relations have changed over the century. Married women now inherit more wealth and have higher wealth levels than they did in 1900 or 1930. This reduces individual concentration even if household wealth inequality does not change. For example, between 1929 and 1953, Robert Lampman (1962) reported that the percentage of married women among top-wealthholders increased from 9 to 18 percent.

As a result, since most researchers are interested in the distribution of family wealth rather than individual wealth, additional imputations must be performed to infer family wealth from estimates of individual wealth holdings. These typically involve making assumptions about marriage patterns—in particular, the correlation between the wealth holdings of spouses (whether, for example, rich husbands tend to be married to rich wives or poor wives).

Another problem with this data source is under-reporting and nonfiling for tax avoidance. Though the returns are subject to audit, the value of cash on hand, jewelry, housewares, and business assets are difficult to ascertain. Their value is typically understated in order to reduce the tax liability of the estate. Moreover, *inter vivos* transfers (that is, transfers of wealth between living individuals), particularly in anticipation of death, can bias estimates of household wealth among the living. If older people pass on wealth to their children just before they die, then their estates would tend to under-represent the wealth of comparably aged individuals still living.

Estate tax data have been used by Atkinson and Harrison (1978) for the UK, and Lampman (1962), Smith (1984), and Wolff (1996a) for the USA. The long-term time-series on wealth concentration for the UK and the USA are based on estate tax data.

3.2 Household Survey Data

Household surveys are questionnaires that are given to a sample of households in a population. Their primary advantage is to provide considerable discretion to the interviewer about the information requested of respondents. Their major drawbacks are that information provided by the respondent is often inaccurate (response error), and, in many cases, the information requested is not provided at all (nonresponse problems). Another problem is that, because household wealth is extremely skewed, the very rich (the so-called 'upper tail' of the distribution) are often considerably under-represented in random samples. An alternative is to use stratified samples, based typically on income tax returns, which oversample the rich. However, studies indicate that response error and nonresponse rates are considerably higher among the wealthy than among the middle-class. Moreover, there are problems in 'weighting' the sample in order to reflect the actual population distribution. Household surveys have been used extensively in many countries (see Wolff 1996b for a survey).

3.3 Income Capitalization Techniques

The third source of wealth data is based on the 'income capitalization' technique, which is usually applied to a sample of income tax returns. The earliest use of this technique on US data was a 1939 study by Stewart. In this procedure, certain income flows, such as dividends, rents, and interest, are converted into corresponding asset values based on the average asset yield. For example, dividends are capitalized into corporate stock holdings by dividing dividends reported in an income tax return by the average ratio of dividends to corporate stock in the economy as a whole. This technique when applied to a large sample of tax returns can provide an estimate of the size distribution of wealth.

This source also suffers from a number of problems. First, only assets with a corresponding income flow are covered in this procedure. Thus, owner-occupied housing, consumer durables, and idle land cannot be directly captured. Also, in the USA, state and local bonds cannot be estimated, because this source of interest income is exempt from federal income taxes. Second, the estimation procedure rests heavily on the assumption that asset yields are uncorrelated with asset levels (that is, for example, that large stock holders will receive the same average return on their stock holdings as small stock holders). Any nonzero correlation between asset holdings and yields can produce biased estimates. Third, the observational unit is the tax return. Various assumptions must be made in order to construct family wealth estimates from tax unit wealth.

3.4 Wealth Tax Data

A fourth source is wealth tax return data, which are available in a number of European countries such as Germany, Sweden, and Switzerland. These countries assess taxes not only on current income but also on the stock of household wealth. Though there is typically a threshold for paying wealth taxes, population coverage can be considerably greater than that of estate tax returns. However, the measurement problems are similar to those of estate tax data. The filer has a great incentive to understate the value of his or her assets, or even not to report them at all, for tax avoidance. Moreover, the assets subject to tax do not cover the full range of household assets (for example, consumer durables, pensions, and life insurance policies are often excluded). In addition, the observational unit is

the tax return unit, which does not directly correspond to the family unit. Wealth tax data have been used extensively by Spant (1987) for an analysis of wealth trends in Sweden.

4. *Changes in the Composition of Household Wealth*

There have been some significant changes in the composition of marketable wealth in the USA over the last two decades. One of the most salient trends is the rising indebtedness of American families, with the debt–equity ratio leaping from 15.1 to 18.9 percent between 1983 and 1997. The principal source of these increases is not rising consumer debt, like credit card balances. In fact, nonmortgage debt as a fraction of total assets fell from 6.8 to 5.2 percent over the period from 1983 to 1997.

Rather, the primary source is rising mortgage debt, including home equity loans and second mortgages, which climbed from 6 to 11 percent of total assets. Indeed, mortgage debt as a share of the value of homeowner's property increased from 21 to 36 percent. Whereas the total market value of homes remained almost constant as a share of total assets over this period, net home equity—the value of the house minus any outstanding mortgages—plummeted from 24 to 19 percent of total assets.

A second notable change is that pension accounts rose from 1.5 to 9.2 percent of total assets between 1983 and 1997. However, this increase almost exactly offset the decline in liquid assets, from 17.5 to 9.8 percent. As a result, the explosion in the use of various pension type accounts, like IRAs, 401(k) plans, and other thrift plans, rather than increasing overall family savings, appears, instead, to have allowed households to substitute tax-free pension accounts for taxable savings deposits.

A third important change is the sharp rise in the share of corporate equities in total assets, from 9.0 to 14.5 percent, reflecting the bull market in corporate stocks. However, this result does not reflect the full extent of the growth in corporate stock holdings, because stocks are also indirectly held in mutual funds, pensions accounts, and trust funds. If we include these, then corporate stocks, both directly and indirectly owned by households, almost doubled, from 11.3 to 20.6 percent of total assets.

5. *The Recent Rise in Household Wealth Inequality*

The final part of the analysis both highlights the sharp increase in wealth inequality in the USA since the early 1980s and illustrates the sensitivity of wealth inequality estimates to different definitions of household wealth.

Table 2
Composition of marketable household wealth in the USA, 1983 and 1997 (percent of gross assets)

	1983	1997
Principal residence (gross value)	30.1	29.3
Other real estate (gross value)	14.9	10.6
Business equity	18.8	17.3
Liquid assets[a]	17.5	9.8
Pension accounts[b]	1.5	9.2
Financial securities	4.2	3.1
Corporate stock and mutual funds	9.0	14.5
Net equity in personal trusts	2.6	3.6
Other assets[c]	1.3	2.6
Total	100.0	100.0
Debt on principal residence	6.3	10.7
All other debt[d]	6.8	5.2
Total debt	13.1	15.9
Memo		
Debt/equity ratio	15.1	18.9
Net home equity/total assets	23.8	18.6
Principal residence debt/house value	20.9	36.4
Stocks directly or indirectly owned/ total assets[e]	11.3	20.6

Sources: 1983 and 1995 Survey of Consumer Finances (tape) and the Federal Reserve Board Flow of Funds.
a Checking accounts, savings accounts, time deposits, money market funds, certificates of deposits, and the cash surrender value of life insurance. b IRAs, Keogh plans, 401(k) plans, the accumulated value of defined contribution pension plans, and other retirement accounts. c Gold and other precious metals, royalties, jewelry, antiques, furs, loans to friends and relatives, future contracts, and miscellaneous assets. d Mortgage debt on all real property except principal residence; and credit card, installment, and other consumer debt. e Includes direct ownership of stock shares and indirect ownership through mutual funds, trusts, and IRAs, Keogh plans, 401(k) plans, and other retirement accounts.

The calculations shown in Table 3 indicate an extreme concentration of wealth in 1995. The top 1 percent of families (as ranked by marketable wealth) owned almost 39 percent of total household wealth. The figures also show that wealth inequality, after rising steeply between 1983 and 1989, increased at a slower pace from 1989 to 1995. The share of wealth held by the top 1 percent rose by 3.6 percentage points from 1983 to 1989. Between 1989 and 1995, the share of the top percentile grew by a more moderate 1.1 percentage points. Projections to 1997, on the basis of asset price movements between 1995 and 1997, suggest that wealth concentration may have risen sharply again over these two years, largely from the surge in stock prices (and the fact that stock shares are highly concentrated in the hands of the rich). The share of the top 1 percent appears to have grown by another 1.6 percentage point.

The inclusion of automobiles in household wealth (the only consumer durable available in the wealth survey data) lessens the estimated share of the richest

Table 3
The percentage share of total household wealth owned by the richest 1 percent of households in the United States, 1983–1997

Year	Marketable net worth	Net worth plus autos	Net worth minus net home equity	Augmented wealth
1983	33.8	32.6	42.9	19.0
1989	37.4	36.1	46.9	20.3
1992	37.2	35.9	45.6	NA
1995	38.5	36.4	47.2	NA
1997	40.1	NA	48.6	NA

Sources: 1983, 1989, 1992, and 1995 Surveys of Consumer Finances, and Wolff (1992). Projections to 1997 are based on the change in asset prices between 1995 and 1997. Note that automobiles are the only consumer durable included in the survey data.

1 percent by about 1 percentage point in 1983 and by 2 percentage points in 1995. This reflects both the fact that automobiles are more widely distributed in the population than other assets and that the importance of vehicles in the portfolio of poorer and middle-income families has risen over time. However, the data also show similar time trends—a sharp rise in wealth inequality between 1983 and 1989 followed by a modest increase from 1989 to 1995.

Household wealth excluding home equity is even more concentrated than marketable wealth, with the richest 1 percent (as ranked by this concept of wealth) owning 47 percent of the total. The trend in inequality is similar, with the share of the top 1 percent gaining 4.0 percentage points between 1983 and 1989 and another 0.3 percentage points in the ensuing six years. Projections to 1997 suggest another rise in the share of the richest percentile, by 1.4 percentage points.

Table 3 also shows the trend in the share of augmented household wealth owned by the top 1 percent of wealthholders as ranked by augmented wealth from 1983 to 1989. The addition of pension and social security wealth has had a significant effect on measured wealth inequality because pension and social security wealth, particularly the latter, have grown much faster over time than marketable wealth and because social security wealth is more evenly distributed in the population than marketable assets. The gap in the share of the top 1 percent between marketable wealth and augmented wealth was almost 15 percentage points in 1983 (34 vs. 19 percent). However, wealth concentration based on augmented wealth still shows an increase between 1983 and 1989 (more current estimates are not available).

Future research might most appropriately be focused on understanding the factors behind changing wealth inequality both in the USA and other advanced nations (household wealth data are not currently available for less developed countries). How have relative wealth holdings changed among age groups, among income classes, across racial and ethnic lines within a country, between families with children and those without children, and across gender lines? How have wealth holdings changed among countries of the world? What are the respective roles of household savings and intergenerational wealth transfers in explaining wealth differences among individuals and between groups? What are the reasons for rising indebtedness among American families, and how does this compare to other countries? These are some of the salient research issues that occupy current researchers in the field.

See also: Consumption, History of; Consumption, Sociology of; Equality and Inequality: Legal Aspects; Income Distribution; Income Distribution: Demographic Aspects; Inequality; Leisure, Sociology of; Savings Behavior: Demographic Influences; Social Demography: Role of Inequality and Social Differentiation

Bibliography

Atkinson A B, Harrison A J 1978 *Distribution of Personal Wealth in Britain*. Cambridge University Press, Cambridge, UK

Feldstein M 1974 Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* **82**: 905–26

Lampman R J 1962 *The Share of Top Wealth-Holders in National Wealth, 1922–56*. Princeton University Press, Princeton, NJ

Lipsey R E, Tice H (eds.) 1989 *The Measurement of Saving, Investment, and Wealth*. Chicago University Press, Chicago

Shorrocks A F 1975 The age–wealth relationship: a cross-section and cohort analysis. *Review of Economics and Statistics* **57**: 155–63

Shorrocks A F, Davies J B 1999 The distribution of wealth. In: Atkinson A B, Bourguignon F (eds.) *Handbook of Income Distribution*. North-Holland, Amsterdam

Smith J D (ed.) 1975 *The Personal Distribution of Income and Wealth, NBER, Studies in Income and Wealth*. Columbia University Press, New York, No. 39

Smith J D 1984 Trends in the concentration of personal wealth in the united states 1958–1976. *Review of Income and Wealth* Series **30**(December): 419–28

Spant R 1987 Wealth distribution in Sweden: 1920–1983. In: Wolff E N (ed.) *International Comparisons of the Distribution of Household Wealth*. Oxford University Press, Oxford, UK

Stewart C 1939 Income capitalization as a method of estimating the distribution of wealth by size group. In: *Studies in Income and Wealth*. National Bureau of Economic Research, New York, Vol. 3

Wolff E N (ed.) 1987 *International Comparisons of the Distribution of Household Wealth*. Oxford University Press, Oxford, UK

Wolff E N 1992 Methodological issues in the estimation of retirement wealth. In: Slottje D J (ed.) *Research in Economic Inequality*. JAI Press, Greenwich, CT, Vol. 2, pp. 31–56

Wolff E N 1996a *TOP HEAVY: A Study of Increasing Inequality of Wealth in America*, Updated and expanded edition. The New Press, New York

Wolff E N 1996b International comparisons of wealth inequality. *Review of Income and Wealth* Series **42**(4): 433–51

Wolff E N 1998 Recent trends in the size distribution of household wealth. *Journal of Economic Perspectives* **12**(3): 131–50

E. N. Wolff

Weber, Max (1864–1920)

Max Weber was born Karl Emil Maximilian Weber in Erfurt, Thuringia on April 21, 1864. His father, Max Weber (1836–97) was a lawyer from a family of wealthy industrialists and merchants with international connections, whose core business was a linen factory in Oerlinghausen, in Westphalia. His uncle transformed this factory, which had been a traditional family concern, into a model modern business. 'The Protestant Ethic and the Spirit of Capitalism' contains thinly concealed elements of this history. Weber's father was a younger son in this family. His father pursued a legal and political career, serving in the German Reichstag 1872–84. Politicians and intellectuals, particularly historians, frequented the Weber household in Berlin. His mother, Hélène (née Fallenstein) was highly educated, religious, and active in poor relief.

Weber entered the University in 1882, studying law, as well as economics, ancient history, philosophy, and theology, and entered into the student life of the time, joining a fraternity. He served the first year of officer training in the army at Strasbourg, where relatives of his mother's were prominent at the University. He returned to Berlin in 1884, and concentrated on law and legal history. Here he became a hard working scholar with a sedate lifestyle, simultaneously pursuing his legal education and taking exams, completing military service, and continuing his academic studies. Levin Goldschmidt and August Meitzen instructed him in *Staatswissenschaft* and he became a doctoral student under Goldschmidt, also briefly clerking at the Royal District Court of Charlottenburg, and performing military service in Posen.

His dissertation, 'Development of the principle of joint liability and the separate fund in the public trading company from the household and trade communities in the Italian cities' (1889), based on hundreds of Italian and Spanish sources, was well-received. He joined the *Verein für Sozialpolitik* (Association for Social Policy) out of a concern with the 'Polonization' of the Eastern regions of Germany, and became involved politically in the *Evangelisch-sozialer Kongress*. He continued his academic work, receiving his Habilitation under Meitzen for 'Roman agrarian history and its importance for state and civil law' in 1891. During this period, the *Verein für Sozialpolitik* employed him for a study of 'The conditions of the agricultural workers in East Elbian regions of Germany.' The study, a huge empirical survey designed as the basis of policy recommendations, marked a new direction in his career. The recommendations were controversial, but made his reputation as an agricultural economist, and enabled him to secure a Professorial appointment in *Nationalökonomie* at Freiburg in 1894. From there he was appointed to a similar position at Heidelberg.

A mental breakdown after the death of his father ultimately led him to relinquish his teaching duties, but eventually he returned to an active role as an editor and producing scholar. His major 'methodological' essays (posthumously published as his *Gesammelte Aufsätze zur Wissenschaftslehre* in 1922) included part one of 'Roscher and Knies' published in 1903, 'Objectivity in Social Science and Social Policy' in 1904, part two of 'Roscher and Knies' and 'Critical Studies in the Logic of the Cultural Sciences' in 1905 and 1906, and, in 1907, 'Critique of Stammler.' With the 1904–5 publication of 'The Protestant Ethic and the Spirit of Capitalism,' his work began to take several new directions. During this period Weber traveled to the US, where he spent several months, appearing as an Agricultural Economist at the 1904 St. Louis World's fair, and visiting American relatives. His experiences were the basis for some of his subsequent writings. He became fascinated by the revolution of 1905 in Russia, learning the language in three months in order to follow the unfolding events, and wrote extensively on them, from a point of view that emphasized matters of social structure. Over the next few years he completed his work on ancient agrarian economic life, agreed to take over the editing of a work which evolved into the text now known as *Economy and Society*, which Weber was never to complete. He became engaged publically in questions of university policy, and also pursued inquiries into industrial psychology in the family textile works, an expression both of his interest in the rationalization of work and his recognition of the role of such psychophysical factors as fatigue in work processes.

Weber also played an active, though characteristically combative role, in two academic organizations, the Social Policy Association and the German Sociological Society. He was one of the founders of the German Sociological Society in 1909, but left soon after because of his views on the separation between

scholarship and advocacy. These views, which later became famous, if not well understood, under the label 'value-free science,' were developed and elaborated in a complex and long-running dispute within the Social Policy Association. This dispute began largely as a political dispute over matters of policy but it escalated quickly and was transformed by Weber and his brother Alfred into an accusation that their opponents in this dispute had confused considerations of value with objective policy considerations, a point that was particularly crucial for Weber, who rejected the bland statism of his monarchical socialist older colleagues. Weber articulated an elaborate rationale for his position that followed from his notion of the nonrationality of ultimate value choice. He also argued that maintaining this separation was part of the duty of university professors.

Weber's role in what was to become *Economy and Society* expanded and he became author rather than editor, a task which together with the series of studies that comprised 'The Economic Ethics of the World Religions,' occupied him until his death. Neither project achieved final form. During World War I Weber, a reserve officer, worked as a hospital administrator, until a reorganization made his services unnecessary. When the outcome of the war became clear, he wrote a series of articles in the *Frankfurter Zeitung* on the future constitution. During this period he gave two speeches with parallel titles to the Munich Free Students Association, 'Science as a Vocation,' apparently in late 1917, and 'Politics as a Vocation' in early 1919, after which both were published. The speeches were an aggressive and telling challenge to the cultural and political optimism of the time, and in the Weimar era remained relevant, especially to the long-running discussion of the 'Crisis of the Sciences.' After the war Weber served as a member of a group invited to participate in *Reichsamt des Innern* (Office of the Interior) meetings which produced the general outlines of the constitution of what became the Weimar Republic (Mommsen 1984), became a professor of economics at Vienna in 1918 and Munich in 1919, and unsuccessfully pursued a career in politics. His untimely death, June 14, 1920 in Munich, came when his influence among intellectuals was at its peak.

1. *Weber's Early Work*

Weber's habilitation dissertation, although he later dismissed it as a 'one of the sins of his youth' (quoted in Tenbruck 1987), together with an essay on 'The Social Basis of the Decline of Ancient Civilization' (1896) marked him out as an original scholar. The dissertation used his knowledge of medieval land tenure law to illuminate the Roman agrarian situation. It depended on the plausible but unprovable assumption that two of the different methods of mapping used by the Roman Colonial land surveyors corresponded to different types of land tenure: untaxed, roughly measured lease rights from public land, and precisely measured, private taxed lots. He argued that the size of the private lots suggested that they were too small for subsistence, and that the smallholders on this land had from the first depended on access to public land to survive, but that this access was eventually cut off by the rich, who could pay for leases and work the land with slave labor. He reasoned that the poor but free peasants were forced, by the inadequacy of their holdings, to assume labor obligations to landlords as rent, and later to be legally bound to the soil. When new slaves were no longer available, barracked slaves, who had enabled production for the market, had to be raised in status, freed sufficiently to encourage them to reproduce, making them essentially into serfs. Increasingly, money transactions were replaced by transactions in kind, the coastal commerce on which the slave produced surpluses depended vanished, and the cities diminished in power and in fiscal, and consequently political, importance for the state. The new rural economic order was manorialism *de facto*, so this was in effect an account of the origins of the medieval agrarian order.

2. *The 1903–1907 Methodological Essays*

Contemporary disputes in ancient history, notably the 'Bücher–Meyer controversy,' over a stage-theory of the development of the *oikos*, were formative for Weber. His most important methodological article, 'Objectivity in Social Science and Social Policy,' was published in 1904, on the occasion of the transfer of control of the *Archiv für Sozialwissenschaft und Sozialpolitik* to a team of editors of which he was the leading member. Weber argued that to construct a historical explanation, or even an object of explanation, requires conceptualization, that he examined in terms of the concept of ideal-types, which he distinguished from mere classifications that simply sorted objects into categories. Ideal-types typically diverged from reality, or were applied to empirical cases from which they diverged, as in the case of ancient capitalism, but served the purposes of the historian by virtue of the fact that they allowed the historian to ask questions that are meaningful in the present. Thus, while ideal-types were indispensable to the historical disciplines, these disciplines were also prone to a particular kind of error. The mistake in each case is that something that is properly regarded as an idealization of a phenomenon is treated as a real force or a teleological process. In Weber's hands, this was a powerful means of methodological criticism. Not only did it serve to indicate how the developmental models of his near predecessors and contemporaries, including Marx, used the idealization to constitute the phenomenon it explained, it called into question the project of final conceptualizations of history or social life, as well as

the cognitive status of such conceptual constructs as 'the state.'

The line between the parts of historical understanding that are relative and those that are nonrelative and factual is difficult to draw. In 'The Logic of the Cultural Sciences,' he spoke of the 'real causal processes' which our concepts give an intelligible form to. In 'Objectivity' he argued that 'genetic ideal-types,' are themselves relative to a culture-bound historical interest: the calculation of probabilities is fully objective, but the selection and conceptualization of the conditions cannot be. These are inherent limitations of a social science that must operate with concepts whose meaningfulness derives from sources outside of the science itself: sciences which speak 'the language of life' a language which is itself necessarily bound to a historical moment and its values (Weber 1949).

This argument takes on a special significance in connection with his account of values themselves. Weber argued that scholarship could at best enable judgments about the adequacy of means for ends, the consistency of ends, the separation between and the difference between factual and valuative considerations, but could not constitute the basis of either value judgments or the endorsements of policies, at least to the extent that the valuative purposes of the policy could not be given scholarly justification. Ultimate value choices, worldviews, religions, and national cultures were alike in that they were ultimately matters of commitment, not reason, and consequently questions of 'superiority' were misplaced.

3. *The Protestant Ethic Essay*

These basic methodological ideas, and especially the idea of ideal-types, figure in Weber's 1904–5 essay, *The Protestant Ethic and the Spirit of Capitalism*. The core problem of the essay was to account for the disruption and replacement of traditional economic life, and he argued that it was in part the indirect effect of Protestantism. Reformation theology, and particularly Calvinism, altered the problem of salvation. The doctrine held that the Elect, those predestined for salvation, are called to worldly tasks and an ascetic lifestyle: worldly monasticism. But, Weber reasoned, the doctrine of predestination was psychologically intolerable, and produced deep anxiety, which had to be allayed by pastoral assurances that faithful work in a calling was a sign of election, creating a novel and powerful psychological sanction for conduct. The theological premise thus had the indirect effect of producing a personality that, in the appropriate setting, becomes the austere, abstemious, rationalizing capitalist.

There are many controversial features of this account, some of which result from the fact that the doctrinal element of the ideal-type that Weber constructed only imperfectly corresponded to the theological doctrine of much of Protestantism in Europe. Nevertheless, the account is deeply compelling, in part because the depictions of the believer accord so well with a character type that unquestionably did leave a profound mark on western culture. Weber closed the essay with some prophetic remarks about the course of western culture, and an important analysis of the cultural situation created by the secularization of the 'ethic.' He observed that the rosy blush of its 'laughing heir,' the Enlightenment, has faded, and that the inexorable demands of the modern, rationally organized economy, enforce the organization of work into 'callings' but strip them of their religious meaning. He commented that even the accumulation of wealth takes on the character of sport, and the people in the machine of capitalism became divided beings, 'Specialists without spirit, sensualists without heart (1958a).

In his late lectures on economic history, Weber restated the thesis with some different emphases, notably with respect to the rational organization of free labor, which Weber came to see as especially crucial, and indeed to be the distinguishing feature of modern capitalism in contrast to booty capitalism, trade, and political capitalism, all of which were found in the ancient world. The crucial moment was when 'the production of goods was emancipated from all the bonds of inherited intelligence, and came under the dominance of freely roving intelligence' (1961), the moment in which the craftsman was replaced by the worker, and the owner of the firm supplied tools, exacted discipline, and assumed the risks. This change was, Weber thought, a precondition for the mechanization of work with the goal of saving labor costs, and in this sense rationalizing labor. And together with this was the rationalization of the use of capital generally, through rational accounting methods, rational price setting for the purpose of profit, and backed by calculable law. The worldwide expansion, and continued success, of these practices did not rest on the foundation of religious belief. Weber also provided an overview of other factors in the development of capitalism, such as calculable law, and other contingencies, like the availability of coal in Europe, without which, he argued, 'the revolution in the means of work … might have stopped and modern capitalism in its characteristic form never would have appeared' (1961).

4. *Economy and Society and the Economic Ethics of the World Religions*

In the last decade of Weber's life he produced a major methodological text; 'The Categories of an Interpretative Sociology,' a major systematic work, the incomplete *Economy and Society*, and the bulk of his mature substantive work, 'Economic Ethics of the World Religions.' The essays that make up this work are overtly concerned with the question of why a

rational capitalism based on calculability originated exclusively in the West. But their larger theme is the general problem of the cultural specificity of the West, that is to say the peculiarity, in world terms, of Western political, economic, and legal institutions. But the uniqueness of the West requires an unusual explanation.

The general development of religious ideas corresponds to the development of particular social strata, with peasants having a strong affinity for magic, aristocratic classes for a kind of self-justifying, charismatic self-conception. These affinities tend to be relatively stable and do not work to transform the social structures in which they arise. The problem of explanation, consequently, is to identify the point of openness in which a very specific kind of innovation, namely those that have transformative effects on individuals or groups, might arise. The general promise of religion is health, wealth, and happiness. And the basic affinities to strata express these needs in forms specific to the strata.

The exceptional cases are the salvation religions, which offer rebirth or redemption, and more generally religions that are doctrinally open to rationalization, and that produce potentially transformative innovations through rationalization. Religious rationalization, however, while producing new coherence or consistency in religious doctrines, characteristically serves to produce discrepancies with other 'spheres' from which religious life is differentiating under the same pressure of circumstance. A general problem which all salvation religions must face up to is the problem of the relationship between destiny and merit, the problem of theodicy, which these religions develop in extraordinarily elaborate ways—ways which on the one hand 'rationalize' the discrepancies by making the fact of evil or bad luck intelligibly consistent with the religious ideology, but also in ways that are consequential for practice. The Chinese solution, for example, was to reconcile the two by recourse to popular magical practice, which in turn had negative consequences: a 'deep repugnance to undertaking any change in the established conduct of life because supernatural evils are feared' (1961). Ancient Judaism, in contrast, devised solutions that excluded and repudiated magic, and this carried over into Christianity and consequently into the western worldview in a way that favored the eventual development of science and technology.

Weber pointed to many similar indirect effects of general institutional conditions, conditions that do not directly determine possibilities, but nevertheless greatly reduce the probability of developments in a particular direction. Rationalization, for example, means the attainment of greater consistency and completeness, but the two take various different forms for intellectuals and practical persons, such as merchants, and the direction that rationalization takes in a given situation or in the face of given problems is governed by many circumstantial contingencies. Weber saw the various forms and processes of rationalization as interacting with one another, that is to say saw rationalization in one sphere as creating difficulties which require rationalization to relate to spheres, and thus producing a kind of dynamic ongoing process of mutually reinforcing rationalization. Occasionally Weber appears to slip into the error of regarding the rationalization process as itself a historical force, but these rare cases ordinarily can be interpreted in causal, and typically 'interest' terms, as when for example the rationalization of modern law on the continent is interpreted as driven by the status interests of law professors.

5. *Economy and Society*

Economy and Society consists of a typology of forms of social action, especially institutional action, together with extensive historical commentary on the types. Weber said that the claim to be made on behalf of the typology was its usefulness; and the elaboration of the typology is an extended demonstration of the uses to which the types could be put in analyzing actual historical forms, particularly of institutional structure, and characterizing their historical course of development as the product of the basic properties of the type, always with the caveat that these were idealizations rather than 'real' types with a teleological character.

The most famous of these typologies involves beliefs legitimating authority and the forms that were characteristically produced by these beliefs. The categories were charismatic, traditional, and rational–legal. These were ideal types, which almost never, if ever, appeared in reality in a pure form. Traditional authority was based on unwritten rules believed to have been handed down from time immemorial; rational legal authority rested on the belief in the validity of written rules produced according to written procedures, and charismatic authority was the authority of the extraordinary person. Weber had in mind of course such figures as Napoleon and Jesus Christ, but he found many more modest examples in Indian gurus and medieval battle leaders, and in each of these cases obedience rested not on either written or unwritten rules but on the direct influence of the exceptional individual himself. Weber was careful to point out such things as the attenuated charismatic element in the modern monarchy and the jury.

Charismatic authority, Weber observed, is inherently unstable, though its temporary effects, and thus historical consequences, may be great. Gradually the critical charismatic leader must support his followers materially and this marks the beginning of the transformation of charisma into the everyday. Charismatic regimes also face problems of succession, and the 'routinization' of charisma through ritual in order

to pass on charisma to successors soon becomes transformed into something akin to traditional authority, or alternatively into a bureaucratic system.

The value of the ideal type is not so much that it serves to predict as that it enables one to ask relevant questions or identify relevant causal processes. But this process is itself governed by our cognitive purposes rather than something inherent in the historical world itself. Much of Weber's discussion of the historical case material is indeed about the logical unfolding of implications of basic ideas such as charisma or rational legal authority. However, there is no sense in which this unfolding is causally inevitable. It is rather that concepts like charisma are useful in making processes intelligible precisely because the elaboration of the idea corresponds to the kind of developments that historical agents actually bring about in the course of their actions employing their own ideas, beliefs, and so forth.

6. *Weber as a Political Thinker*

Weber was self-consciously part of the post-Bismarck generation, and an ardent nationalist who was deeply concerned about the problem of succession of leaders. Bismarck's career had shown how forceful and successful leadership produced popular acquiescence and even enthusiasm, but with the effect of making Parliament weak and irrelevant, and undermining party loyalty. Weber saw this form of popular leadership, exemplified for him by both the unelected Bismarck and the elected Gladstone, as a key to the nature of modern politics: modern regimes were mass democracies in fact, regardless of the forms of constitutions, because they required mass legitimacy and support to be effective. The selection of leaders, and beyond this the administrative machinery of parties, were constrained by the demands of effective electoral politics. He was concerned on the one hand that interest-faction parties, especially parties that inevitably were bureaucratically organized, would be effective electorally, but that this would assure the predominance of party hacks rather than figures who could act boldly as leaders. He resnded by suggesting a presidential office in which the president required support in the form of a plebiscite rather than party support, and thus stood above and outside of party structures. Although this analysis supported a liberal and democratic republican form of government, Weber nowhere based his argument for this form of state on traditional liberal constitutional concerns, such as the preservation of rights or the expression of the will of the people, notions he regarded as mythological, but instead based it on the problems of selecting great leaders in the face of the modern need for popular support for the tasks of nation-building.

Many interpreters have noted that the Nietzschean language of greatness that Weber employed, as well as the specific concerns about freedom he expressed, are rooted in a morality of the value of 'personality' that is potentially in conflict with constitutional liberalism, and especially to its egalitarian elements. When he observed that in rational work organizations 'each man becomes a little cog in the machine and, aware of this, his one preoccupation is whether he becomes a bigger one,' and when he asked if 'a portion of mankind (can be kept) free of this parceling out of the soul' (quoted in Mayer 1956) his concern was with a form of freedom accessible only to the few. Yet Weber's argument proceeds on more than one level. On the one hand he presented the state as a tool whose distinctive and necessary means is violence, and traced the origins of the historical conditions that constrain the possible present forms of state. But he made the point that the state is a means not necessary merely to the achievement of his own values, but to the achievement of virtually all this-worldly values: only someone like a Tolstoyan Saint has no need of the state. For the rest, there are no alternatives.

7. *Weber's Reception*

Weber was not employed as a professor during his most productive years, and exerted relatively little influence through the training of students. His primary personal influence was through a salon that he and his wife held in Heidelberg which attracted many of the best minds, and particularly the best young minds, of the university, across various disciplines. This group included Georg Lukács, Emil Lask, and Karl Jaspers. Jaspers portrayed him as an existentialist hero: 'an idealized modern man who resolutely exposes himself to intellectual risks and lives through intellectual tensions passionately without accepting final resolutions' (Schluchter 1996).

The Nazi period cut off serious discussion of Weber, who was described in the intellectual circles that supported the regime with the derogatory term 'liberal,' a characterization that shaped the revival of his thinking in the postwar period. Weber's 'liberal' admirers, such as Jaspers, promoted the idea that Germany would have been saved from Nazism had Weber been accepted as political leader, and as a consequence, after the Second World War, Weber was reinterpreted as a precursor of the new federal republic and its form of liberal democracy. This distortion was attacked relentlessly by Mommsen in his dissertation (1984) which noted Weber's attitude toward the normative premises of conventional liberal constitutionalism and suggested that his devotion to the idea of great leadership placed his thoughts uncomfortably close to the fascist *Führerprinzip* leadership principle of Nazi political thought. This suggestion produced a controversy that broke out in a visible way in the 1964 commemoration of the centennial of his birth, when the thesis was radicalized by Marcuse and Habermas.

Mommsen later stressed that he did not believe that Weber's own intentions had been antidemocratic in the Nazi manner, but that nevertheless 'it crossed the boundaries of the traditional understanding of democracy and was amenable to reinterpretation in an antidemocratic sense' (1984).

Weber's thought appears early in the US, but not as a founding figure of sociology, since he had no 'system' of the sort that other German and American thinkers had, and was only intermittently identified as a sociologist. Economists first translated his work: Frank Knight his *General Economic History*, based on student lecture notes and some added material from earlier work, Talcott Parsons his 'Protestant Ethic,' the argument of which was popularized in Britain by R. H. Tawney. After the Second World War, the American social sciences, and especially sociology, were exported, in part as an alternative to Marxism, and Weber's ideas, understood primarily in American terms, became influential, in the UK and in Germany itself. The characteristic emphasis was on value-free science and the rejection of ideological thinking, and on specific sociological concepts, such as charisma and bureaucratization, and on Weber's rejection of the Marxist claim of primacy of material over spiritual or institutional explanation. Weberian approaches typically were seen as emphasizing the practical demands of bureaucracy, or the autonomous role of interests other than economic interests, such as ideal interests. The notion of the Protestant ethic fit with analysis of modernization and development of the sort that were part of the American response to third world and anticolonial movements, and were generalized and applied to such countries as Japan.

Weber came to be regarded as a distinguished forebear of antipositivist sociology by virtue of his use of the terms 'meaning' and 'understanding' to differentiate his own approach to sociology. The refinement and appropriation of Weber's idea of an interpretative sociology by phenomenologists such as Alfred Schütz became one source of a phenomenological approach to sociology. Weberian ideas were also conveyed through the writings of such figures as Karl Mannheim and Norbert Elias, but in forms increasingly distant from Weber's own thought. This set the stage for a series of rediscoveries within sociology of 'the original' Weber so that the 1980s and 1990s saw within sociology, and especially in the work of such writers as Wolfgang Schluchter, and more generally in connection with the *Max Weber Gesamtausgabe* project, a much more disciplined return to Weber's original concerns, and a vindication of the importance of such basic problems to Weber as the problem of Western uniqueness.

Although in his final post and through much of his teaching career Weber was an economist and economic historian, his influence on economic history has been limited. But within political science and sociology, by the end of the twentieth century, a great many of the fragments of Weber's thought had become part of the standard conceptual tool kit. It is doubtless a result of Weber's gift as a clarifier of concepts that these tools have shown such as an astonishing staying power. In many cases, such as the concept of charisma and Protestant ethic, the concepts passed into common usage.

See also: Academy and Society in the United States: Cultural Concerns; Action Theory: Psychological; Agricultural Change Theory; Agriculture, Economics of; Antiquity, History of; Authority, Social Theories of; Boas, Franz (1858–1942); Buddhism; Bureaucracy, Sociology of; Capitalism; Civil Society/Public Sphere, History of the Concept; Civilization, Concept and History of; Cultural Psychology; Culture and Economic Development: Cultural Concerns; Culture: Contemporary Views; Democracy; Democracy, History of; Economic History; Economics, History of; Elites, Anthropology of; Elites: Sociological Aspects; Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity: Anthropological Aspects; Financial Institutions in Economic Development; Gender, Class, Race, and Ethnicity, Social Construction of; Generations, Sociology of; Global History: Universal and World; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: East Asia; Historiography and Historical Thought: Jewish Tradition; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Ideal Type: Conceptions in the Social Sciences; Idealization, Abstraction, and Ideal Types; Judaism; Law: Change and Evolution; Linguistic Anthropology; Methodological Individualism in Sociology; Migration, Economics of; Migration, Theory of; Migrations, Colonizations, and Diasporas in Archaeology; Organizations, Sociology of; Peasants and Rural Societies in History (Agricultural History); Policy History: State and Economy; Political Economy, History of; Power in Society; Power: Political; Religion: Evolution and Development; Religion, Sociology of; Sapir–Whorf Hypothesis; Social Class and Gender; Social History; Social Stratification; Society/People: History of the Concept; Sociology, History of; Sociology: Overview; State Formation; State, History of; State, Sociology of the; States and Civilizations, Archaeology of; Status and Role: Structural Aspects; Urban History; Weberian Social Thought, History of; Work and Labor: History of the Concept

Bibliography

Baier H, Lepsius M R, Mommsen W J, Schluchter W, Winckelmann J (eds.) 1982–present *Max Weber—Gesamtausgabe*. Mohr, Tübingen

Hennis W 1988 *Max Weber: Essays in Reconstruction* [trans. Tribe K]. Allen and Unwin, London/Boston
Käsler D 1988 *Max Weber: An Introduction to His Life and Work* [trans. Hurd P]. Polity Press, Cambridge, UK
Lehmann H, Roth G (eds.) 1993 *Weber's Protestant Ethic: Origins, Evidence, Context*. Cambridge University Press, Cambridge and New York
Mayer J P 1956 *Max Weber and German Politics: A Study in Political Sociology*, 2nd rev. and enlarged edn. Faber and Faber, London
Mommsen W J 1984 *Max Weber and German Politics, 1890–1920* [trans. Steinberg M S]. University of Chicago Press, Chicago
Schluchter W 1981 *The Rise of Western Rationalism: Max Weber's Developmental History* [trans. Roth G]. University of California Press, Berkeley, CA
Schluchter W 1989 *Rationalism, Religion, and Domination: A Weberian Perspective* [trans. Solomon N]. University of California Press, Berkeley, CA
Schluchter W 1996 *Paradoxes of Modernity: Culture and Conduct in the Theory of Max Weber* [trans. Solomon N]. Stanford University Press, Stanford, CA
Swedberg R 1998 *Max Weber and the Idea of Economic Sociology*. Princeton University Press, Princeton, NJ
Tenbruck F H 1987 Max Weber and Eduard Meyer. In: Mommsen W J, Osterhammel J (eds.) *Max Weber and His Contemporaries*. Allen and Unwin, London
Weber M 1946 *From Max Weber: Essays in Sociology* [trans. H. H. Gerth and C. W. Mills]. Oxford University Press, New York
Weber M 1949 *Max Weber on The Methodology of the Social Sciences* 1st edn. [trans. Shils E A, Finch H A eds.]. Free Press, Glencoe, IL
Weber M 1951 *Religion of China: Confucianism and Taoism*. Free Press, Glencoe, IL
Weber M 1952 *Ancient Judaism* [trans. Gerth H H, Martindale D eds.]. Free Press, Glencoe, IL
Weber M 1958a *The Protestant Ethic and the Spirit of Capitalism* [trans. Parsons T] student edn. Scribner, New York
Weber M 1958b *The Rational and Social Foundations of Music* [trans. Martindale D]. Southern Illinois University Press, Carbondale, IL
Weber M 1958c *The Religion of India: The Sociology of Hinduism and Buddhism*. Free Press, Glencoe, IL
Weber M 1961 *General Economic History*. Collier, New York
Weber M 1963 *The Sociology of Religion* [trans. Fischoff E]. Beacon Press, Boston
Weber M 1975 *Roscher and Knies: The Logical Problems of Historical Economics* [trans. Oakes G]. Free Press, New York
Weber M 1976 *The Agrarian Sociology of Ancient Civilizations* [trans. Frank R I]. New Left Books, London; Humanities Press, Atlantic Highlands, NJ
Weber M 1977 *Critique of Stammler* [trans. Oakes G]. Free Press, New York
Weber M 1978 *Economy and Society: An Outline of Interpretive Sociology* [trans. Roth G, Wittich C eds.]. The University of California Press, Berkeley and Los Angeles, CA
Weber M 1981 Some categories of interpretive sociology [trans. Graber E]. *The Sociological Quarterly* **22**: 151–80
Weber M 1988 *Max Weber: A Biography* [trans. Zohn H]. Transaction Press, New Brunswick, NJ
Weber M 1994 *Political Writings* (Lassman P, Speirs R eds.). Cambridge University Press, Cambridge, UK
Weber M 1995 *The Russian Revolutions* [trans. Wells G, Baehr P]. Cornell University Press, Ithaca, NY

S. P. Turner

Weberian Social Thought, History of

1. *From Weber to Weberianism*

In portraits of individual thinkers, the adjective 'Weberian' is as widely used as 'Marxian,' 'Durkheimian' or 'Freudian.' The expression 'Weberian Social Thought,' however, is an awkward one if taken to imply the existence of a unified, coherent body of work, a determinate set of beliefs, or a group of committed followers. Weber's admirers can be as jealous of his work as Marx's are of the work of Marx. Despite Weber's clearly enormous influence on twentieth century social and political analysis, there has never been a Weberian school of sociology comparable to that which grew up around Durkheim's work. There has also never been a Weberian school of political science, or a Weberian movement to match the power and intensity of Marxism or psychoanalysis. The reasons for this are as follows: the immense range of Weber's interests; the fragmented character of his work as a whole and the difficulty of reducing it to a small number of basic formulae; Weber's own disdain towards intellectual hero worship; and the intellectual and political circumstances in which his work was made available after his death in 1920. His statement that science was tied to the course of progress, and his expressed hope that his own work would one day be made obsolete, merely add irony to an already complex and fragmented legacy.

Weber's social thought does not contain a specific 'theory of' the family, state, education, history, and so on, although these have been the objects of subsequent efforts at reconstruction. It is rather a set of sensibilities, political commitments, and principles of analysis. These constitute the elements from which the various branches of 'Weberian social thought' are made up.

Politically Weber was both a liberal and a nationalist. His nationalism was the product of its time and has had little impact on his subsequent admirers, save as an embarrassing inconvenience. What Ralf Dahrendorf (1967) called his 'strange liberalism' expressed itself not in an explicit theory of human rights or justice—towards which abstractions he was skeptical—but in the search for those institutional arrangements which would ensure 'individual freedom of movement' and a politically educated citizenry in a rationalized, bureaucratized world, whose problems socialism threatened to exacerbate. Weber's political commentaries combine a concern with parliamentary or party machinations, an analysis of the major social

and economic groups of a society (considered in terms both of group interests and of group status), and the geopolitical situation of modern nation states. In these commentaries the value of freedom and the question of the qualities of human beings which a culture breeds are more important than the need to defend particular institutional arrangements on principle.

The focus of Weber's substantive sociology is the economic and social history of Europe, the major world religions, and large scale civilization complexes. The principal tool for this is the ideal type, the one-sided exaggeration of empirical reality from the point of view of a phenomenon's 'cultural significance.' Weber's ideal typical descriptions of social processes are notable for sure-footed shifts between levels of abstraction—achieved by the use of ideal types of varying degrees of generality—and for the deployment of the same type in a variety of historical settings without loss of nuance or specificity. His sociology knows neither rigid classifications of phenomena, nor social or historical laws, nor decisive individual events. He rejects holistic theories of society, anti-individualist or teleological philosophies of history, and cultural/political utopianism. The task of the social scientist is both to respect science's limits and to fulfill its vocation by promoting culturally significant historical knowledge and clarity concerning current conditions of political action, while offering no prescriptions for action. The Weberian social scientist is thus not a public intellectual in the sense that he is a nation or a people's or a class's conscience, but rather one who possesses specialist knowledge and makes that knowledge available to government and to an educated public when circumstance permits. In Raymond Aron's terms (1988), the Weberian social thinker is 'isolated and available.'

Weber's comparative study of civilizations and world religions convinced him that the secularization and rationalization of social relations in the West had not relieved individuals of the burden of previous religious struggles or conflicts, but had rather turned them inward. The centrality of 'struggle' in his work refers not merely to tragic and ineliminable conflicts between social and political interest groups or world views, but also to individually experienced conflicts between the demands placed upon modern human beings by a disenchanted world and their inward spiritual needs.

As Weber was influenced by, and integrated the thought of numerous predecessors—notably Machiavelli, Kant, Marx and Nietzsche—so the list of his successors is long and varied. Karl Mannheim, Carl Schmitt, Hans Morgenthau, Raymond Aron, Alfred Schutz, Ralf Dahrendorf or Barrington Moore may all be considered practitioners of 'Weberian social thought' in one form or another, though in each case key elements of Weber's work are neglected or rejected. Many others, notably German neo-Marxists have accepted the broad terms of Weber's diagnosis of the modern world—particularly his 'disenchantment' thesis—without endorsing either his response or the tools with which he constructs it. Significantly, Catholic social thought has found nothing congenial either in Weber's philosophy of social science or in his political writings, which decisively reject both holistic forms of social analysis and organicist social ethics.

2. *Weberian Social Thought in Germany*

The history of Weberian social thought is paradoxical and discontinuous, particularly in Germany and the USA. Following Weber's death in 1920, the hasty attempts by Marianne Weber to make both his work and life more broadly available in Germany met with mixed results. While Weber's work—in particular the lecture 'Science as a Vocation' of 1919—continued to be widely discussed, and many considered him to be the greatest of all German sociologists. However, in the intellectual climate of the Weimar Republic, Weber's research program, which refused to attribute agency to social wholes, and which endorsed the comparative analysis of social structures and cultural systems, met with broad indifference. The attractions of dialectical, utopian or messianic grand narratives, exemplified by Weber's one time protégés Georg Lukacs and Ernst Bloch, or by Oswald Spengler, and of holistic sociological analysis, proved greater to the generation which matured after World War I. With the advent of national socialism in 1933, Weberian social thought met with open hostility in universities which were oriented to holistic, Volkish doctrine and practice, and in which analysis and the drawing of categorical distinctions was considered typical of 'the Jewish intellect.' In the quarter century following its publication in 1922 *Wirtschaft und Gesellschaft*, considered by many to be Weber's *magnum opus*, had sold no more than 2000 copies.

After 1945 the advent of a Federal Republic oriented officially to a US image of democratic values, and the denazification of German universities, further weakened the link between German sociology and its classical roots. The formal and substantive influence of the American occupying forces culminated in an intellectual climate combining two basic needs for an empirical and methodologically sensitive social inquiry; and for what Helmut Schelsky (1959) called 'a neutral background ideology with a non-antagonistic character.' Neither these, nor the widespread postwar belief that social science ought to be a direct participant in the planned reconstruction of European society, were conducive to a revival of Weberian social thought. Wolfgang Mommsen's study ([1959] 1984) then strengthened an association between Weber and an outdated, bellicose nationalism. A measure of redress came in Heidelberg in 1964 with the conference to mark the centenary of Weber's birth, though notably all the keynote speakers were at the time

working outside Germany. In the 1960s and 1970s, while Weber's diagnosis of modernity—particularly the accounts of rationalization and disenchantment—remained an indispensable resource for second and third generation critical theorists, German social thought of the period was predominantly neo-conservative. This was exemplified by the neo-Aristotelianism in political philosophy and Gehlen's theory of institutions in sociology, neo-Marxist and neo-Freudian. The 'left Nietzscheanism' of the late 1960s did little to alter the predominant image of 'Weberian social thought' as a mode of cultural and political analysis within universities. In the year 2001 Weberian social thought in Germany is less a distinct style of thinking which rivals Marxism, rational choice theory or structuralism, than a set of analytical/conceptual tools or resources, a starting point, deployed by different scholars in a wide variety of settings. For the rest, Weber's work continues to be the object of hermeneutic puzzlement. Notable here are efforts to reconstruct the meaning of his work as a whole (Tenbruck 1980, Hennis 1988) and to systematize it into a coherent developmental history of the West (Schluchter 1981), both of them expressions of uncertainty as to what 'Weberian social thought' is. Hennis' interpretation, which plays down the developmental aspects in favor of Weber's 'vision of man,' claims that at the center of Weberian social analysis lies the study of the way in which individuals in different social locations conduct their lives. Though regarded as somewhat idiosyncratic in his native Germany, this has proved popular in Anglo-Saxon scholarship and has led to claims that Weber's work has affinities with contemporary diagnoses of post-modernity and even with some branches of cultural studies.

3. *Weberian Social Thought in Britain and America*

Paradoxically, Weberian social thought has flourished as readily in the English-speaking world as it has in Germany. Responsibility for this is frequently attributed to Talcott Parsons. Parsons contributed the first English-language monograph devoted to Weber's work as a whole, volume 2 of *The Structure of Social Action* (Parsons 1937) and also published a translation of *The Protestant Ethic and the Spirit of Capitalism* in 1930, and part of *Wirtschaft und Gesellschaft* in 1948. But this would be to ignore the role of numerous German *émigré* scholars in the USA in the 1930s and 1940s, such as Alexander von Schelting and Albert Salomon, and the presence of Karl Mannheim in the UK after 1933. The most accessible general collection of Weber's writings, *From Max Weber* (1946), was translated and edited by the *émigré* Hans Gerth with C. Wright Mills, and was followed in 1949 by Edward Shils' translations of three pieces on the philosophy of social science and culture, *The Methodology of the Social Sciences* (Weber 1949).

Nevertheless, more than one scholar loyal to European traditions of social thought has observed that in the course of the twentieth century the ideas of German thinkers in particular—one thinks here of Heidegger, Nietzsche and Freud—have been trivialized as they have been transferred to North America, and that that those of Weber are no exception. Hence the claim that the Weberian social thought which flourishes in the Anglo-American world might be interpreted as a bloodless version of the original which inspired it. This is certainly the view of Wilhelm Hennis (1988), who has laid responsibility for this at the door of Talcott Parsons. Parsons' functionalist account of institutions, scientist account of the nature of action, and evolutionary conception of history represent, so it is alleged, a misuse of Weberian ideas. In particular, it blinds us to the tragic sensibility that animates both Weber's conception of Western history and politics and his relationship to his own scientific activity. The Americanized Weber is, it is claimed, a Weber to whom the grand historical dramas and passions that unfold on the pages of Marx and Nietzsche meant little.

Yet against this charge, it may be recalled that Weber himself was an admirer of both the political and cultural achievements of three Protestant cultures outside Germany, the Netherlands, the UK and the USA. Weberian social thought in North America, which has ignored the tragic element in Weber, may be seen with caution as a partial reciprocation, an acknowledgment of how Weber's own thought might have appeared in North American conditions, freed of the specifically European burdens which Weber took it upon himself to bear. A central Weberian motif is after all, the need of the social scientist to face up to 'the demands of the day,' whatever they may be.

In addition, in should be noted that more than one Weberian baton has passed through the hands of Anglo-Saxon scholarship. As well as Salomon, Schelting and Gerth, versions of Weberian social thought have been promoted by Hans Morgenthau, Gunther Roth and Reinhard Bendix in the USA, and by John Rex and other conflict theorists in the UK. Morgenthau's realist theory of international relations, in which priority in political analysis is given to the power political interests of large nation states and the conduct of statesmen, bears the imprint of Weberian thought throughout. Secondly, Weber's historical sociology of domination, mediated by Bendix and Roth, has exerted a major influence on the analysis of the pattern and dynamic of development in Western and non-Western societies alike. The work which has grown up around Barrington Moore, Theda Skocpol, Bendix, and, arguably, Perry Anderson, owes a good deal more to Weber's own sociology of domination than to Parsons' version of it. These writers, however, unlike Weber, have generally organized their material

according to a theme, pattern or formula—dictatorship or democracy, Kings or people, the Absolutist State, the priority of external over internal pressures in revolutionary situations. And these historical sociologists have only fleetingly employed the distinctively Weberian rhetorical device—to be found even in his most antiquarian studies—of making implicit reference to current political circumstance. In the UK, the version of sociology which came to be known as conflict theory, and associated in particular with the work of John Rex and Ralf Dahrendorf, may be taken as a conscious effort to deploy Weberian categories in the analysis of industrial labor and race relations. A characteristically Weberian feature of such work was that it combined empirical inquiry, theoretical and methodological self-consciousness, and a sense of the social and political urgency of the problem addressed.

4. *Weberian Social Thought in France*

Weberian social thought has never taken root in France, as is made clear by the late appearance of French translations of Weber's work: *Science as a Vocation* and *Politics as a Vocation* in 1959, *The Protestant Ethic and the Spirit of Capitalism* in 1964, *The City* in 1982, *The Sociology of Law* in 1986. *Economy and Society*, while translated into Spanish in 1944 and Italian in 1961, has yet to appear in full in a French edition. This is partly explained by the entrenchment of national intellectual traditions in the first half of the twentieth century, exemplified by the oft-noted mutual indifference of Weber and Durkheim. Within sociology, the Durkheimian School remained important in the 1920s, and the reception of Weber's work was sporadic. Maurice Halbwachs reviewed Weber's *Protestant Ethic thesis* in 1925, and Raymond Aron's *German Sociology* of 1935 devoted a chapter to Weber. But the former remained a committed Durkheimian while Aron's text did no more than make a French audience aware of Weber's existence. The other potential conduit for Weberianism in France in the 1920s, Georges Gurwitch, knew German social and political thought primarily via the legal philosopher Emil Lask, and following World War II became openly hostile to what was seen as Weber's interpretative arbitrariness, nominalism, and lack of dialectical acumen.

Paradoxically, the role of the USA in postwar France, and doubts about the scientific status of sociology and attendant search for new voices appeared to make possible a revival of the fortunes of Weberian thought, with many younger French sociologists spending time in America and encountering both Weber and survey research. Yet the prestige of Durkheimian sociology gave way not to Weberianism but to Hegelian/Marxian and later Heideggerian and Nietzschean philosophy, an exclusion reinforced by the notorious intensity and intimacy of a French intellectual life dominated by Paris. With the exception of Raymond Aron and Julien Freund, no major French intellectual of the postwar period could be described as a Weberian social thinker, and even Aron, who maintained a constant dialogue with Weber and whose public role was the Weberian one of political educator oriented to 'the demands of the day,' offered the French public only a restricted version of Weberian social thought. Oriented primarily to the relationship between ethics and politics in the lives of political actors, it largely avoided the sociological analysis of the dynamic of material interest that found its way even into Weber's most journalistic political commentaries. And where Weber's political commentaries emphasized the tragedy of individual choice, Aron sought to reconcile conflicting values. It may be argued that in the 1950s this Weberian social thought for statesmen found its way into the thinking of many who passed through Aron's hands at the *Institut des Sciences Politiques* and *Ecole Nationale d'Administration*.

In the 1960s a more social, less overtly political form of Weberianism developed around the work of figures such as Alain Touraine and Jean Marie Vincent, who deployed Weberian categories in the study of industry and organization, and of new forms of social conflict. This work also allowed for a Weberian critique of the bureaucratizing tendencies within the French State. Since 1980, with an increasing awareness of the importance in Weber's work of the themes of life conduct and style of life, the work of figures such as Pierre Bourdieu and Michel Foucault has been held to display affinities with that of Weber. However, while Bourdieu's theories of 'cultural capital' have some affinities with Weber's account of status and prestige, his work as a whole owes as much to Marx and Durkheim. Michel Foucault's work on power, discipline, and the conduct of life bears thematic comparison with Weber but his archaeological and genealogical historical method, and libertarian conception of politics, differ from Weber's in basic respects.

5. *Weberian Social Thought in Other Countries*

Outside the European mainstream 'Weberian social thought' has been enormously influential without gaining a distinct foothold or producing significant modification or extension. In Italy, those elements of Weber's work which are of importance have been grafted on to an already established tradition of realist social and political analysis which overlaps with Weber's approach at numerous points. In the Soviet Union before World War II and in state socialist countries after it, Weberian social thought did not flourish. In the 1920s, while Russian historians such as Neusychin discussed Weber's account of feudalism and of the origins of capitalism with some sym-

pathy—albeit through aligning Weber with Marx—the ethical and political dimension of Weber's thinking—exemplified by *Science as a Vocation*—remained a closed book. This state of affairs, in which Weber's work was either shown to have affinities with Marxism or rejected for its combination of neo-Kantianism and undialectical positivism, prevailed throughout the Soviet bloc, and to a considerable extent in the West too, until the mid-1970s, until renewed interest among Western scholars in Weber's work as a whole began to have an effect in state socialist countries. In countries such as Poland—which before 1945 had a strong indigenous tradition of non-Marxist social thought—Weber's work has exerted a growing influence since the 1970s, firstly on discussions of ethics and politics, and since 1989 on discussions of the ethical and religious significance of economic conduct in rapidly transforming societies. One of the more remarkable episodes in the history of Weberian social thought concerns Japan. The opening of Japan following the Meiji restoration led to numerous intellectual exchanges with an imperial Germany that was rapidly developing while retaining many elements of a feudal/traditional ethos. Weber's co-editor Emil Lederer spent two years in Tokyo in 1923–1925, as did Karl Löwith in 1936–1941. The effect of these, was that the Protestant ethic thesis was discussed as early as 1910 in Japan, and Japanese translations appeared of the *General Economic History* (1927), of *Science as a Vocation*, *The Meaning of Ethical Neutrality* (both 1937), and *Politics as Vocation* (1939) before they appeared in English.

6. *Weberian Social Thought After 1989*

The history of Weberian social thought since the end of World War II has been largely one of the way in his work has been mediated and refashioned to suit changing political, economic, and intellectual circumstance. The fruits of those efforts have been ambiguous, the elements of Weber's own thought being developed in different directions and not always to Weberian purposes. Parsonian functionalism and Schutzian phenomenology are notorious cases in point, along with the fragmented appearance of Weber's work in anthologies, and the concentration of empirically informed Weberian social thought in the areas of industry and organization. Yet following the collapse of communism in 1989, with the geo-political tensions in Europe which gave rise to nationalist wars and the unification of Germany, and the rapid development of capitalism and the trappings of democracy in China and East Asia, there may be grounds for arguing that the history of Weberian social thought has come full circle. Also, that the current problems and dilemmas facing Europeans and non-Europeans alike are more directly Weberian than they have been at any time since his death.

Questions asked via Weberian categories during the cold war—are Western societies evolving towards a rationalized and disenchanted 'modernity'? What are the appropriate terms in which to understand the nature of human action? Has ideology come to an end? What are the main dimensions of power and authority within industrial enterprises and bureaucratic organizations?—may begin to yield to ones which recall the atmosphere of political and social urgency which animates Weber's own texts, is a united Germany with its capital in Berlin a threat to its Eastern neighbors, or for that matter its Western ones? What social consequences follow a state of affairs in which the European community confronts its poorer neighbors as a land of seasonal opportunity and as a fortress? Is there a difference between Protestant and Catholic social thought on European integration? What effects do new forms of commerce and enterprise have on relationships between urban and rural areas in societies with large peasant populations? What of the relationships between majority and minority populations? Do the institutional arrangements in the new democracies foster or hinder the political education of their citizens? What are the consequences of the rapid transition to a market economy for family structures, welfare provision and, above all, for the manner in which individuals are able to conduct their lives? What is the relationship between religion, ethics, and economic or political conduct? Can these questions be posed in the same way in Europe and the rest of the world, or is it still necessary to take into account the different religious and civilizational parameters through which they are addressed?

It would be un-Weberian to predict the future of Weberian social thought, but one might surmise that until 2020 the search for the core elements of Weber's own world view will wane as its protagonists reach the end of their careers. At the same time, the questions which Weber asked of his own time, questions which twentieth century social and political science have been either able to avoid or reluctant to address, may cast a well-defined shadow over the social and political world of the first half of the twenty-first century.

See also: Aristotelian Social Thought; Aron, Raymond (1905–83); Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Charisma and Charismatic; Charisma: Social Aspects of; Civilization, Concept and History of; Civilizational Analysis, History of; Civilizations; Economic Sociology; Law, Sociology of; Legitimacy; Legitimacy, Sociology of; Marxist Social Thought, History of; Nostalgic Social Thought; Parsons, Talcott (1902–79); Political Sociology; Pragmatist Social Thought, History of; Religion, Sociology of; Social Question: Impact on Social Thought; Totalitarianism: Impact on Social Thought; Utilitarian Social Thought, History of; Weber, Max (1864–1920)

Bibliography

Aron R 1988 *Power, Modernity, Sociology*. E. Elgar, Aldershot, UK
Bendix R 1978 *Kings or People*. University of California Press, Berkeley, CA
Collins R 1986 *Weberian Sociological Theory*. Cambridge University Press, Cambridge, UK
Dahrendorf R 1967 *Society and Democracy in Germany*. Doubleday, Garden City, NY
Gerth H, Wright Mills C (eds.) 1948 *From Max Weber*. Routledge and Kegan Paul, London
Hennis W 1988 *Max Weber: Essays in Reconstruction*. Unwin Hyman, London
Hirschhorn M 1988 *Max Weber et la Sociologie Francaise*. Editions L'Harmattan, Paris
Lassman P, Velody I (eds.) 1989 *Max Weber's 'Science as a Vocation.'* Unwin Hyman, London
Meja V, Misgeld D, Stehr N (eds.) 1987 *Modern German Sociology*. Columbia University Press, New York
Mommsen W [1959] 1984 *Max Weber and German Politics 1890–1920*. University of Chicago Press, Chicago
Moore B Jr. 1966 *The Social Origins of Dictatorship and Democracy*. Beacon Press, Boston
Oakes G, Vidich A J 1999 *Collaboration, Reputation and Ethics in American Academic Life*. University of Illinois Press, Urbana, IL
Parsons T 1937 *The Structure of Social Action*. Free Press, New York
Schelsky H 1959 *Ortsbestimmung der deutschen Soziologie*. Diedrichsverlag, Dusseldorf, Germany
Schluchter W 1981 *The Rise of Western Rationalism*. University of California Press, Berkeley, CA
Schwentker W 1998 *Max Weber in Japan: Eine Untersuchung zur Wirkungsgeschichte 1905–1995*. Mohr Siebeck, Tübingen, Germany
Tenbruck F H 1980 The problem of thematic unity in the work of Max Weber. *British Journal of Sociology* **31**: 313–51
Weber M 1930 *The Protestant Ethic and the Spirit of Capitalism*. Allen and Unwin, London
Weber M 1947 *The Theory of Social and Economic Organisation*. Free Press, New York
Weber M 1949 *The Methodology of the Social Sciences*, trans. by E Shils. Free Press, New York
Weber M 1986 *Sociologie du droit*. Presses universitaires de France, Paris

C. Turner

Weight Regulation: Psychological Aspects

The scientific challenge of weight regulation is to understand the processes which lead some people to maintain a normal, stable weight, and others to become unhealthily fat or dangerously thin.

Body size usually is characterized in terms of body mass index (BMI) (computed as the ratio of weight measured in kilograms to the square of height measured in meters). A BMI below 18.5 is defined as underweight, from 18.5 to 24.9 as normal weight, from 25 to 29.9 as overweight, and over 30 as obese. The predominant weight disorder in the industrialized world is obesity and now almost one in five adults are obese, and twice as many are overweight (Seidell and Rissanen 1998). Anorexia nervosa and bulimia nervosa fall at the other extreme of weight disorders, and while they don't match the prevalence of obesity, together they represent the most common psychiatric diagnoses for young women today (Hoek 1995).

1. Weight and Weight Control

Weight control problems appear to be very different at the two extremes. Most obese people are aware of their weight problem and try to regulate their food intake, but often they find it impossible to exert enough control, or to maintain control over a long enough period, to achieve a normal body weight. In contrast, patients with eating disorders appear to exert excessive control over their body weight, successfully defying biological and social pressures to eat. In between the two extremes lies the normal-weight population. At one time it was assumed that these would be the 'normal' eaters, that is, not worried about weight and regulating their energy balance without any deliberate effort. However, this view is looking increasingly questionable. In Britain, for example, the proportion of the population who are obese has doubled since the mid-1980s, and concerns about weight control are widespread, even among the normal-weight, suggesting that weight regulation is neither effortless nor effective. The normal-weight population probably includes a minority who maintain a normal weight without deliberate effort, a larger group (most often men) whose weight slowly increases over time, but who make little attempt to regulate it, and the largest group (in which women predominate) who are chronically attempting to control their food intake in order to regulate their weight closer to the contemporary ideals for body shape.

Research into psychological aspects of weight regulation is predominantly in three areas: (a) the psychological processes leading to the development of eating disorders, (b) the psychological processes leading to obesity, and (c) the causes and consequences of dietary restraint.

2. Weight Regulation in Eating Disorders

Psychological research on eating disorders has a long history dating back to observations in the eighteenth century that extreme loss of weight could have a psychogenic origin. More recently, the formal psychiatric diagnosis of anorexia nervosa was established, with key features being extreme weight loss, a morbid fear of fatness and deliberate restriction of food intake to achieve a low body weight. Bulimia nervosa was

characterized later as a distinct but related disorder, sharing a morbid fear of fatness, but with an eating pattern consisted of binge eating followed by self-induced vomiting or purging to achieve weight control (Garfinkel 1995). Much of the psychological research on eating disorders has been concerned with the origin of fear of fatness, and although there are a number of aetiological theories (see Brownell and Fairburn 1995), the most widely accepted view is that eating disorders emerge in response to the cultural pressures for a slim physique in vulnerable individuals. Recent evidence suggesting that there is a genetic element to eating disorders is consistent with the idea that there is an inherited predisposition which depends upon events in the environment to emerge as a disorder (Bulik et al. 2000).

2.1 Weight Control in Anorexia Nervosa

From the perspective of weight regulation, the most interesting question is how anorexic patients successfully maintain a low body weight, when everyone else finds it so difficult. There has been surprisingly little progress in understanding this process. Implicit in the etiological accounts is the idea that eating disordered patients are exceptionally highly motivated. Weight gain is so much feared that any degree of suffering is acceptable to avoid it. Cognitive factors may play a part, with clinical evidence indicating that anorexic patients use an elaborate array of cognitive strategies to render the idea of food and eating undesirable. A third factor is personality; anorexic patients tend to be both perfectionist and persevering, either of which might facilitate weight control. However, none of these factors appears to capture the secret of the extraordinary, albeit pathological, weight control capacities of patients with anorexia nervosa.

2.2 Weight Control in Bulimia Nervosa

The phenomenology of bulimia nervosa illustrates a different weight control strategy, involving giving in to the desire to eat during binge episodes, but avoiding the consequences of eating (weight gain) by inducing vomiting (Garfinkel 1995). Many bulimic patients start inducing vomiting as a temporary means of gaining control over their weight, but in practice both the sense of loss of control and the size of the binges, often increase. One explanation for this is that appetite is modified by the postingestional calorific consequences of feeding, and if these are attenuated by vomiting then appetite increases to compensate. Inducing vomiting after eating is certainly an effective strategy for weight control, but in terms of adverse psychological and physical effects, the costs usually exceed the benefits.

3. Weight Regulation in Obesity

The problem of obesity needs to be investigated on several levels in order to understand (a) why it has become so much more common over recent decades, (b) why some people are so much fatter than others, and (c) how obesity treatments could be made more effective.

3.1 The Increasing Prevalence of Obesity

Most accounts of the shifts in obesity prevalence over time draw not on changes in social or psychological conditions, but on the combination of an environment which favors a positive energy balance and an appetite control system which evolved when the food supply was limited. There is now an abundant, highly palatable, food supply in industrialized countries. At the same time, transportation systems and mechanisation have reduced the energy demands of everyday life. If human beings evolved to store energy supplies, then it is not surprising that so many people are now overweight (Hill and Peters 1998).

3.2 Emotional Theories of Obesity

Psychologists' first incursions into understanding the causes of obesity at an individual level focused on underlying psychopathology, with suggestions that eating served unconscious needs. These ideas came largely from clinical observations by psychodynamically-orientated therapists. Empirical studies have shown that levels of psychopathology are no higher in the obese than the normal-weight population, and although there is evidence that severely obese adults have raised levels of depression and anxiety, this appears to be a response to the multiple handicaps and stigmatization associated with obesity (Sobal 1991). Unfortunately, popular discourse continues to equate obesity with psychological problems, leading to greater problems of stigmatisation.

3.3 Psychobiological Theories of Obesity

This perspective was followed by psychophysiological models which combined investigations of the psychophysiology of obesity and food intake with assessments of eating behaviour. Obesity was hypothesized to be caused by low responsiveness to internal cues for satiety combined with high responsiveness to external food cues. Laboratory studies showed differences in responsiveness to food between obese and normal-weight adults, and the theory was reinforced by parallel observations on animal models of obesity (Schachter 1968). At around the same time, the idea of

a 'set-point' for weight was put forward, suggesting that each individual had a particular level of fat stores which was defended by up-regulating appetite if they were depleted, and down-regulating it if they were filled. At the time, the 'fact' that most people had a fairly stable body weight over their lifetime was adduced as support for the set-point theory. Obese people were hypothesized to have a higher than normal set-point, and would show abnormalities of appetite when they lost weight to below their own set point weight. However, subsequent research suggested that it was not obesity *per se* nor even weight loss which was linked with abnormal appetite, but deliberate weight control (Herman 1978). Obese people who were not bothering with weight control appeared to respond normally to food and satiety cues, while normal-weight dieters showed a pattern of responses which resembled the dieting obese. These observations led to tremendous interest in research on the effects of dietary restraint (see below) but a period of relative stagnation in research into the psychological process in obesity.

3.4 Psychological Treatment for Obesity

Psychological treatments for obesity developed primarily from a behavioral analysis of the problem. Obese people were hypothesized to have abnormal eating habits, which tended to promote a positive energy balance. The earlier formulations of the so-called 'obese eating style' incorporated rapid eating (preventing the normal processes of satiation emerging during the course of the meal), chaotic eating (eating in a wide range of different places), as well as the idea of externally-cued eating. None of these ideas survived rigorous examination, but in the meantime, they had proved to be useful as the basis for behavioral treatment. Treatment incorporated stimulus control (ensuring that exposure to cues for eating were controlled), slow eating, and limiting the number of places in which eating occurred, along with the standard behavioral procedures of self-monitoring, goal-setting, and self-reinforcement. The first results were very positive (Wing 1998). In time, cognitive elements were added to the programs, including modifying self-defeating cognitions and dealing with the lapse-relapse sequence. Cognitive behavior therapy (CBT) is now regarded as the gold standard for obesity treatment, and CBT training has been identified as a necessity for other health professionals. However, it remains true that many patients drop out of treatment, few lose substantial amounts of weight, and many regain the weight that they have lost, so there is still a great deal of scope for improvement.

Current thinking on obesity treatment has recovered from a period of despondency about treatment efficacy by acknowledging that weight losses are likely to be relatively small, and taking account of the fact that even small weight losses can have significant benefits for health (Allison and Pi-Sunyer 1995). There is a recognition that obesity is a chronic disorder requiring continuing care. There is a renewed interest in the psychological aspects of obesity, particularly the problems of body image and social stigmatisation, and some suggestion that treatments which offer a broader-based approach produce better overall outcomes on health and wellbeing. Most importantly, the difficulties which almost all obese people experience in losing weight has led to greater interest in prevention, in the hope that interventions could be initiated before overweight is established and health and well-being are compromised.

4. Weight Regulation and Dietary Restraint

While health risks associated with overweight emerge only at a BMI of over 25, many women aspire to a much slimmer figure as promoted across all forms of media (Brownell 1991). In search of the slim ideal, many women, and even pre-adolescent girls, try to restrict what they eat in order to control their size. The concept of dietary restraint was put forward in 1975, to describe an eating pattern which consisted of attempted control over eating, interspersed with regular failures of control (Herman 1978). Several psychometric measures have been developed, starting with the original Restraint Scale, but subsequently distinguishing the concepts of attempted control (now often called restrained eating) and failures of control (called counter-regulation or disinhibition). A raft of experimental studies showed that restrained eaters were less responsive to satiety, more responsive to palatability, and more likely to show over-eating in response to negative affect. Binge eating was also conceptualized in relation to restraint, following the realisation that most binge eaters were also chronic dieters (Wardle and Beinart 1981). These observations turned existing ideas on their head in suggesting that attempts to regulate food intake might be the cause of, and not simply a response to, problems of eating control.

4.1 Restraint and Inhibition

The association between restraint and regulatory problems is now well-established, but there is still uncertainty about the mechanism, and especially about why restrained eaters should ever actually eat *more* than unrestrained eaters. The cognitive model of restraint proposes that after a high calorie preload (in the laboratory) or a dieting transgression (in everyday life), the dieter 'reasons' that they might as well take advantage of the situation and leave dieting until tomorrow. The psychobiological model suggests that after a long history of imposed cognitive control, in which normal biological controls have been overridden and food intake patterns are chaotic and

irregular, sensitivity to satiety or hunger signals becomes blunted and learned patterns of hunger and satiety are disrupted. The emotional model suggests that the dieter copes with negative emotional states with dissociation, and as part of this, loses contact with their restrictive intentions. There is some evidence for each of these processes, but as yet there have been no definitive studies that distinguish between them.

4.2 Implications for Treatment

When the paradoxical effects of dietary restraint were invoked to explain binge eating in bulimia, this led to cognitive-behavioral treatments focusing not only on helping bulimic patients to take a more positive view of their body, but also to relinquish rigid dietary restraint. Likewise in obese binge eaters, loss of control over eating has been interpreted as a consequence of strict dieting. This view was supported by the weight history of many obese patients, among whom any period of losing weight on a diet seemed to be followed inexorably by regain to higher than initial weight, as well as by the poor outcome of most obesity treatment programs, where long-term follow-up showed that many patients ended up fatter than when they had started. Caution was counseled in offering conventional dietary treatment to obese binge eaters, and some psychologists promoted so-called 'un-dieting' treatments, where obese patients could be helped to stop their long-established habits of trying (and failing) to restrict what they ate (Goodrick et al. 1998).

4.3 Deliberate Weight Control

Recently, there have been signs of a renaissance of interest in deliberate weight control in the face of the epidemic scale of the obesity problem. Evidence is emerging that dieting is not inexorably associated with problems of control (Howard and Porzelius 1999). Among the obese, binge-eaters have not been found consistently to be more restrained than nonbinge-eaters, and contrary to expectations, obese binge-eaters fare no worse than nonbinge-eaters in conventional dietary treatments. In the normal-weight population, individuals who are either dieting or at least 'watching their weight' appear to consume healthier diets than nondieters, and probably gain less weight over time. There has also been a realization that most research on dietary restraint in the normal population was based on comparisons between restrained and nonrestrained eaters, but since restrained eaters probably become restrained because they were susceptible to weight gain, the causal processes may be more complex than was thought.

On present evidence it seems likely that restraint can increase the risk of loss of control, but there may be riskier and less risky forms. Flexible restraint has been associated with better weight control than rigid restraint. Eating regimens with regular meal patterns and a balanced nutrient intake are probably better than short-term drastic restriction. Negative body image is very often the motive for attempts at weight loss, but may in itself increase the difficulty of controlling weight, so interventions which improve body image could facilitate weight control. Higher levels of physical exercise consistently are associated with better weight control, both in the normal population and in maintaining weight loss after treatment, and deserve a higher profile in psychological treatments.

5. The Future

In recent years one of the most influential developments has been the observation that individual differences in weight are under substantial genetic control. Identical twins are highly similar in weight at all stages of life, and consistently more similar than fraternal twins. The results of twin studies are supported by adoption studies which have shown that the BMI of adoptive children is correlated strongly with the BMI of their biological parents and barely at all with the adoptive parents' BMI. These findings have important implications for psychological research into weight control, pointing to the need to investigate eating behavior in 'at risk' populations (Faith et al. 1997). If shared family environment plays much less of a part in determining weight than has been believed, then this raises the question of what aspects of the individual environment promote weight regulation and control.

In the past, psychological research on weight regulation has tended to run a parallel course to research on biological processes, but the best progress is likely to come from integrative approaches in which the control of appetite and weight is seen as the product of a network of interactions among the elements from different domains. There have been tremendous advances in understanding the biological basis of obesity, and a growing interest in new pharmacological treatments, all of which offer great opportunities for psychological research as we seek to understand how biology and behavior are integrated to achieve weight regulation.

See also: Body Image and Gender; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Eating Disorders, Determinants of: Genetic Aspects; Health: Self-regulation; Hunger and Eating, Neural Basis of; Obesity, Behavioral Treatment of; Obesity, Determinants of: Genetic Aspects; Self-efficacy and Health; Self-regulation in Adulthood

Bibliography

Allison D B, Pi-Sunyer F X 1995 *Obesity Treatment: Establishing Goals, Improving Outcomes, and Reviewing the Research Agenda*. Plenum, New York

Bray G A, Bouchard C, James W P T 1998 *Handbook of Obesity*. Marcel Dekker, New York
Brownell K D 1991 Dieting and the search for the perfect body: Where physiology and culture collide. *Behavior Therapy* **22**: 1–12
Brownell K D, Fairburn C G 1995 *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford Press, New York
Bulik C M, Sullivan P F, Wade T D, Kendler K S 2000 Twin studies of eating disorders: A review. *International Journal of Eating Disorders* **27**: 1–20
Faith M S, Johnson S L, Allison D B 1997 Putting the behavior into the behavior genetics of obesity. *Behavior Genetics* **27**: 423–39
Garfinkel P E 1995 Classification and diagnosis of eating disorders. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford Press, New York
Goodrick G K, Walker S C, Kimball K T, Reeves R S, Foreyt J P 1998 Non-dieting versus dieting treatment for overweight binge-eating women. *Journal of Consulting and Clinical Psychology* **66**: 363–8
Herman P 1978 Restrained eating. *Psychiatric Clinics of North America* **1**: 593–607
Hill J O, Peters J C 1998 Environmental contributions to the obesity epidemic. *Science* **280**: 1371–4
Hoek H W 1995 The distribution of eating disorders. In: Brownell K D, Fairburn C G (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*. Guilford Press, New York
Howard C E, Porzelius L K 1999 The role of dieting in binge eating disorder: Etiology and treatment implications. *Clinical Psychology Review* **9**: 25–44
Schachter S 1968 Obesity and eating. *Science* **161**: 751–6
Seidell J C, Rissanen A M 1998 Time trends in the worldwide prevalence of obesity. In: Bray G A, Bouchard C, James W P T (eds.) *Handbook of Obesity*. Marcel Dekker, New York
Sobal J 1991 Obesity and nutritional sociology: A model for coping with the stigma of obesity. *Clinical Sociology Review* **9**: 125–41
Wardle J, Beinart H 1981 Binge eating: A theoretical review. *British Journal of Clinical Psychology* **20**: 97–109
Wing R R 1998 Behavioural approaches to the treatment of obesity. In: Bray G A, Bouchard C, James W P T (eds.) *Handbook of Obesity*. Marcel Dekker, New York

J. Wardle

Welfare

Welfare is the result of an expanding production of care, financial benefits, or other services for ensuring the material and cultural conditions for the reproduction of humans as biological beings, as economic producers and members of a work force, and as social beings and citizens (the conditions of integration in society and of social cohesion). These functions for satisfying a set of needs are performed through various informal or institutionalized arrangements (ranging from family support to public assistance and social insurance funds) and by various types of actors (families, employers, nonprofit organizations, public administrations, etc.).

By identifying the present-day producers of welfare, the first part of this article will show that what we now commonly call the welfare state represents a combination of the factors of production of welfare, wherein the state plays a major but not exclusive role. The second part will describe the combinations corresponding to different types of welfare states. Finally, recent trends in welfare and reforms thereof in response to the 'crisis of the welfare state' will be examined.

1. *The Producers of Welfare*

There are three major producers of welfare: civil society, the market, and the state. Depending on its history, level of economic development, and political and social structures, each country has worked out its 'mode' of welfare production, characterized by a particular combination of these three pillars in what has been called the 'welfare triangle.'

1.1 *Society*

Forms of solidarity based on belonging to a neighborhood, community, or family represented the first means of social protection and care for individuals, before state institutions replaced face-to-face solidarity with public welfare systems. The family and nonprofit organizations, in particular, are major producers of welfare in civil society.

1.1.1 The Family. The family used to be the major supplier of welfare to individuals. But even in welfare states, it remains an influential producer of welfare, supplying as it does much of the food, housing, education, and care to children and the elderly. These unremunerated activities performed by families—and mainly by women—have always represented a basic guarantee for the individual against the contingencies of life. When old-age pension funds did not exist and the person had neither an inheritance nor an estate, the best guarantee for old age was to have many children who could later satisfy the basic needs of their aging parents, when the latter would no longer be fit to work. Family care, a fundamental sort of welfare in preindustrial societies, was also one of the preconditions for the development of wage labor: 'Participation in wage labor presupposes a support-system of unwaged labor outside the labor market' (Taylor-Gooby 1991, p. 100).

Even though the modern welfare state has assumed many of the family's traditional roles (especially in providing income support, education, and care to the young and old), the family still produces welfare.

During the 1980s and 1990s, the spread of joblessness and the aging of the population have increased its role. For one thing, unemployed young people live longer with their parents. For another, governments have, given the increasing number of dependent older persons, encouraged families to increase their help and care for aging parents—the intention being to curb the growth in public expenditures devoted to caring for the elderly. In families with four generations, two of them in retirement, young retirees quite often have to handle the needs of both their unemployed adult children and their dependent elderly parents. However, the family's role as a producer of welfare varies from one type of welfare state to another. Since welfare in southern Europe, for example, tends to be organized along the lines of providing monetary benefits, families there have to make up for the lack of adequate services.

1.1.2 Nonprofit organizations. Nonprofit organizations are important producers of welfare, even in developed welfare states. The friendly or mutual aid societies of the past were the harbingers of the public social security systems of our times. At present, nonprofit organizations cooperate closely with the state in order to produce welfare. In many continental European countries, the state provides financing, and nonprofit organizations deliver services. In Germany for instance, most of the assistance provided to families and the elderly comes through federations of nonprofit organizations. The same holds for France. Nor should we set the nonprofit sector at odds with the marketplace. Many for- and non-profit hybrids have come into being. In France for example, certain home services for the elderly are provided through associations, which take the elderly's place as an employer of home helpers and bill them the costs.

1.2 The Market

In the welfare vacuum in the early nineteenth century, security and well-being depended on one's estate, on one's status as property-owner. The person who owns 'a capital or an estate thus has insurance whereas those who own nothing depend on others in case of need' (Hatzfeld 1989, p. 27). At the turn of the twentieth century, a new type of security arose, based on social law (mainly labor law) instead of property rights. It served as the foundation for social insurance and social security systems.

On the periphery of these national welfare systems, the market continues, through private insurance funds and for-profit social services, satisfying part of the population's welfare needs. Several national welfare systems allow for public as well as private, for-profit activities. Such is the case for health in France and Germany, and for education in the United Kingdom.

Since the 1980s, 'neoliberal' supporters of free enterprise have advocated fully or partly privatizing social insurance and, in particular, retirement funds (World Bank 1994). However, privatization risks splitting the population in two: those who have the means of contracting out to a private insurance policy and those who will depend on public assistance for minimal support.

Companies, too, produce welfare for their employees. To provide employees with (supplementary) health or pension benefits, they usually turn to the marketplace and private insurance funds. When such benefits are provided under collective agreements, representatives of labor and employer organizations run the programs. This happens in the countries described as 'neocorporatist' (Germany, France and The Netherlands). In some countries, such as the United States or Japan, company programs provide basic access to welfare. The phrase 'corporate welfare' has been used to describe the Japanese system.

1.3 The State

Before the late nineteenth century, the state primarily intervened in society in order to maintain law and order rather than to provide social security for all. It played the role of a police on the periphery of the 'self-regulating' marketplace (Polanyi 1944). Public assistance was, in the main, repressive.

Modern welfare states have arisen out of the massive industrialization and urbanization trends, and their consequences for the emerging working class. Working class poverty motivated increasing state interventions in social policy. As wage-earning developed, more and more people had a single source of income: the sale of their labor in the marketplace. Whenever their strength failed, when illness, joblessness or old age felled them, their lives plunged into insecurity. In this context, security could no longer depend on owning property or on individual responsibility. It would be linked to the labor law and the status of wage-earner. As sociohistorical studies have established, the welfare state was built up on a new linkage, characteristic of 'wage-earning society,' between the dependent status as wage-earner and an extensive system of protection against risks (Castel 1995). Wage-earning employment and social security formed a system. The state would regulate it and vouchsafe the wage-earners' new 'social rights.' Through its legislative and regulatory interventions, the state gradually made insurance for covering the risks of health and old age compulsory for employers and their employees; and it organized social protection.

As Ewald (1986) has shown, the technique of insurance led to a new conception of responsibility and risk, which has served as the foundation for the development of the welfare state. This conception broke with the liberal doctrine of individual responsi-

bility and fault; and insurance offered a way out of the blind alley of paternalistic company-run welfare policies. It has thus become possible to avoid the quest to find a fault and the party causing it. The insurance technique replaces this quest with the question of how to collectively share the costs of risks. It has served as the basis for another conception of security in industry, and introduced a new rationality and a different paradigm of risk and security. On this basis, welfare states would develop after World War II. The welfare state represented a social pact between wage-earners and employers, a pact backed by the state and, to a large degree, organized by it in various ways from country to country.

Modern welfare states guarantee a comprehensive set of social rights to all citizens. These rights have thus become a major part of citizenship. This universality differentiates contemporary welfare from the social laws of earlier times, which mainly targeted the poor or the working class. According to Briggs (1961), the welfare state modifies the interplay between forces in the marketplace in three main ways. First of all, it guarantees a minimum income to individuals and families. Second, it reduces insecurity by enabling individuals and families to cope with the risks related to illness, old age, unemployment, and disability. Third, it guarantees all citizens, regardless of class or social status, access to certain standard of social services and facilities. As a consequence, the idea of the welfare state centers around this reference to the necessity of counterbalancing the market economy's negative effects. For this reason, 'decommodification' was one of the principal indicators used by Esping-Andersen (1990) to draw up a classification of 'welfare regimes' that is now solidly established in the literature.

2. *Three Types of Welfare Regimes*

Initially, comparative research on welfare resorted to a single model of the welfare state, under which one quantity was used to measure the contribution to welfare, namely public expenditures on it. Countries were thus classified on a scale based on these expenditures. This sort of approach to understanding the welfare state led to the 'industrialization' or 'convergence' thesis (Wilensky 1975), which maintained that the major determinant underlying the development of the welfare state is economic growth, with its bureaucratic and demographic implications. Accordingly, the degree of welfare in a country was causally linked to the level of economic development there. This implied that all developed countries would gradually converge toward similar social security programs. This single-minded 'developmentalist' view of progress proved unable to account for the diversity and specific characteristics of welfare programs in different countries, and even less able to explain the timing of the introduction of such programs. Reducing all forms of public interventions to the single quantity of expenditures on welfare does not enable us to understand the specific ways that, from country to country, the state itself has been built up through close interactions with civil society (Rosanvallon 1990). As it turned out, the same level of expenditures might correspond to considerable differences in welfare states, in both their objectives and forms of organization. For these reasons, it was necessary to identify and analyze different types of institutional models of the welfare state.

The typology of welfare states worked out by Esping-Andersen (1990) has replaced the foregoing approach. This scholar identified 'three worlds of welfare capitalism'—social democratic, conservative-corporativist, and liberal or residual—a typology based on three criteria:

(a) the level which the system has reached in 'decommodifying' needs, i.e., the degree to which individuals or families can maintain a socially acceptable standard of living, independent of their participation in the market;

(b) the system's impact on social stratification;

(c) the relationship between state, market, and family.

The social democratic welfare state, typical of Scandinavia, is the most 'universalist' of the three systems. It offers not just a high level of protection against risks, but also a good supply of services. This type of system ranks highest on the scale of decommodification, since its extensive 'social rights' grant individuals the most freedom from dependency on the market. Under this model, taxes serve to cover the costs of social services and benefits. The main aim is an equal redistribution of wealth. Hence, this type of system has a unifying effect on social stratification.

The conservative-corporativist welfare state is based on the wage-earning relationship. Benefits and services are granted as a function of one's occupational category. The aim is not to redistribute wealth but to partly replace wage-earners' income so as to maintain their status when they are at risk. This sort of system is financed not through taxes but through payroll contributions. Despite its generous benefits, it ranks lower on the scale of decommodification than the social democratic welfare state. Eligibility for benefits depends on access to the labor market and maintenance in the wage-earning status. When employment difficulties crop up, whole groups of young people, women, etc., lose their eligibility and fall back into the safety net of public assistance. Usually found in continental Europe outside Scandinavia, this type of system sometimes goes under the name of 'continental welfare state.'

The liberal or residual welfare system assigns the most weight to the market economy and the least to the state. As a consequence, it ranks lowest on the decommodification scale. The state intervenes only as

a last resort, and it serves mostly means-tested benefits tied to the person's income. Very limited welfare is provided with the intention of favoring the person's return to the labor market. In terms of social stratification, this system causes intense inequality between those who can take out private insurance and those who have to live on fixed, restricted allocations, and who are often stigmatized for receiving them. The United States of America is frequently mentioned as resembling, in many ways, this type of system.

As we can see, each of these welfare systems corresponds to a particular combination of welfare producers. Each combination of the three pillars—civil society, the market, and the state—in the welfare triangle represents a specific system of social protection. This combination cannot be reduced to state interventionism; instead, it assigns the state a degree of influence that varies along with that of the other two pillars of the welfare state.

3. *The Challenge: Reforming Welfare States*

Since the late 1970s, fierce criticism has increasingly been voiced about the crisis of contemporary welfare states. So far, however, reforms have met with strong institutional inertia. Most systems of social protection still have much the same structure as when they were set up four or five decades ago. The crisis of the welfare state was, at first, diagnosed in terms of the fact that all demands for welfare have gradually converged on the state, which, thus overburdened, has proven ineffective or even inefficient.

Several proposals of various sorts have been made for reforming the welfare state in line with this diagnosis. One sort seeks to reduce radically the state's role; the actions of Thatcher in Great Britain and of Reagan in the United States have been interpreted in terms of this 'retrenchment' (Pierson 1994), as the state withdraws from welfare to the profit of the market. The supporters of another sort of reform have argued for 're-embedding' solidarity in society (Rosanvallon 2000), so as to readjust the state's role by reinforcing the pillar of society in the welfare triangle, whence the idea of a 'welfare mix.' Rein and Rainwater (1986) have argued that the welfare state is too narrow a framework to cope with details about the economic well-being of individuals in a society. Accordingly, reforms must reckon with all the producers of welfare, and with how they form new combinations in the welfare triangle.

From a similar perspective, current thought about coping with the crisis of European welfare states has taken as a starting point the specific institutional arrangements in each country (Bonoli and Palier 2000). 'Each of Europe's welfare regimes confronts a different set of adjustment problems depending on their particular characteristics and thus differs in their paths to reform' (Ferrera et al. 2000, p. 431). For instance, the conservative-corporativist type of welfare system tends to be caught up in the vicious circle of 'welfare without work' (Esping-Andersen 1996). This syndrome incapacitates the system from dealing with the challenge of an aging population and its consequences for old-age pension funds. Such welfare states have usually implemented generous schemes for allowing wage-earners to withdraw from the labor market well before the age of retirement. They have thus favored the emergence of an 'early exit culture' that makes it impossible for pension reform to work. For this type of system, improving the ratio of the actively working to the retired population entails shifting from passive income support to active job policies for targeting aging wage-earners so as to maintain their employability and keep them in the work force (Guillemard 2001).

In conclusion, as this example shows, although European welfare states face the same challenges (in particular, demographic aging), they have to adopt reforms as a function of their particular problems. Imaginative solutions have to be tailored to the specific situation. They will come out of a complex process that cannot be reduced to a single procedure.

See also: Poverty, Sociology of; Social Security; Underclass; Welfare: Philosophical Aspects; Welfare Programs, Economics of; Welfare State; Welfare State, Geography of; Welfare State, History of

Bibliography

Bonoli G, Palier B 2000 How do welfare states change? Institutions and their impact on the politics of welfare state reform in Western Europe. *European Review* **8**(3): 333–52

Briggs A 1961 The welfare state in historical perspective. *Archives Européennes de Sociologie* **2**: 221–58

Castel R 1995 *Les Métamorphoses de la Question Sociale*. Fayard, Paris

Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Polity Press, Cambridge, UK

Esping-Andersen G 1996 Welfare without work: The impasse of labour shedding and familialism in continental European social policy. In: Esping-Andersen G (ed.) *Welfare States in Transition*. Sage, London

Ewald F 1986 *L'Etat Providence*. Grasset, Paris, pp. 66–87

Ferrera M, Hemerijck A, Rhodes M 2000 Recasting European welfare states for the 21st century. *European Review* **8**(3): 427–46

Hatzfeld H 1989 *Du Paupérisme à la Sécurité Sociale*. Presses Universitaires de Nancy, Nancy, France

Guillemard A M 2001 Continental welfare states in Europe confronted with the end-of-career inactivity trap. Working Papers Series. Center for European Studies, Harvard University, Cambridge, MA

Pierson P 1994 *Dismantling the Welfare State? Reagan, Thatcher and the Politics of Retrenchment*. Cambridge University Press, Cambridge, UK

Polanyi K 1944 *The Great Transformation*. Rhinehart, New York

Rein M, Rainwater L 1986 *Public/Private Interplay in Social Protection: A Comparative Study*. Sharpe Inc., New York

Rosanvallon P 1990 *L'Etat en France*. Seuil, Paris

Rosanvallon P 2000 *The New Social Question: Rethinking the Welfare State*. Princeton University Press, Princeton, NJ
Taylor-Gooby P 1991 Welfare state regimes and welfare citizenship. *Journal of European Social Policy* **1**(2): 93–105
Wilensky H 1975 *The Welfare State and Equality*. University of California Press, Berkeley, CA
World Bank 1994 *Adverting the Old Age Crisis: Policies to Protect the Old and Promote Growth*. Oxford University Press, Oxford, UK

A.-M. Guillemard

Welfare: Philosophical Aspects

'Welfare' is not an easy term to define with precision, even restricting attention to human beings. In the abstract, it refers broadly to the happiness, well-being, or good of an individual, society, political association, and so on. More concretely, it also refers to particular ingredients and sources of happiness or good. Life itself is an essential element of any human being's welfare, for example, and the means of basic subsistence, including food and shelter to some minimal extent, can be viewed as vital sources of human welfare whose provision should be guaranteed by government, as in the term 'welfare state.' The sources and ingredients of human happiness are so various, however, that 'welfare' in the philosophical sense must not be limited to basic needs or the means of satisfying them. In this article, four different interpretations of welfare are discussed critically. 'Welfarist' moral philosophies are also contrasted with 'nonwelfarist' moral philosophies. The first step is to dispel some confusion surrounding the term 'utility' and its relationship to welfare.

1. *Welfare and Utility*

Modern economists, decision theorists, and game theorists typically distinguish between welfare and utility. Utility in their terminology refers to the value of a mathematical function—strictly speaking, any function in a family of informationally-equivalent functions—that *represents* the preferences of a person or of society such that, for any pair of alternatives in the feasible set of outcomes, a larger value is assigned to the preferred alternative. The relevant family of informationally-equivalent utility functions is determined by the given assumptions of measurability and interpersonal comparability of utility. In this regard, measurability must not be confused with comparability across persons. A measurability assumption (such as ordinal measure, cardinal measure, or ratio-scale measure) specifies a particular degree of precision with which any person's preference ranking can be represented by utility numbers—the more precise, the smaller the family of informationally-equivalent utility functions. Such an assumption is logically independent of any comparability hypothesis, according to which the utility representations of different persons' preferences are mutually dependent. Given comparability, the utility function which can be used to represent a person's preferences depends on the utility function selected from the family available to represent another person's preferences. In the absence of comparability, however, there is no constraint of this sort on which members of the relevant families can be used to represent the different persons' preferences.

To illustrate, suppose that utility numbers are ordinal and noncomparable. Given the ordinal measure, a person's utility function is unique only up to positive monotonic transformations. Such transformations preserve the order of relative priority of the utility numbers assigned to the alternatives. Given noncomparability, it is permissible to transform any person's utility function without altering the utilities of other people. Thus, a person's ordinal noncomparable utility ranking generally provides no more information than that contained in his own preference *ordering*—a complete, reflexive, and transitive relation—over the alternatives.

If utility numbers are assumed to be cardinal and noncomparable, however, then a person's utility function is unique up to positive linear (strictly speaking, affine) transformations: the cardinal noncomparable family of informationally-equivalent functions is more restricted than its ordinal counterpart, which admits positive nonlinear transformations. A cardinal noncomparable utility ranking represents preference information that is richer than his preference ordering, namely, the relative *intensity* with which he prefers one option to another. If his revealed attitudes toward risk are independently assumed to measure how much subjective importance he assigns to his different options, for instance, the Von Neumann–Morgenstern method of cardinalization is appropriate and expected utility functions are valid cardinal noncomparable representations of his preference intensities (see, e.g., Binmore 1994, pp. 259–78, 304–15).

To take a final illustration, if utility numbers are interpersonally comparable, then one person's utility function cannot be subjected to any transformation unless every other person's utility function is similarly transformed. A cardinal comparable utility ranking provides information not only about the person's preference intensities, for example, but also about how intense his preferences are in comparison to those of other people. To construct such a utility ranking, however, there must be a method of making interpersonal comparisons apart from a method of cardinalization (Sen 1979). For instance, a person might extend sympathy to the other members of his society by adopting their personal preference intensities while imagining himself in their respective positions at any

pair of alternatives. He must also be able to render the different persons' intensities dimensionally compatible in order to compare them. By reducing them to a common unit, he could form a suitable extended preference reporting that when he places himself in their respective positions, he believes that one person i's preference for x over y is of greater intensity than another person j's preference for y over x, for any two persons and any pair of alternatives. His personal preferences, a subrelation of his extended ones, might continue to be decisive for his actions. But if he can extend sympathy to others in this way, his preferences can be represented by a cardinal comparable utility function. At the same time, different people may disagree with respect to interpersonal comparisons, and thus form conflicting extended preferences. *A fortiori*, different people may have conflicting personal preferences and conflicting cardinal comparable utility functions.

In general, a utility function encapsulates information about measurability and interpersonal comparability of the underlying preferences. As a device for representing this preference information (including orderings, intensities, and interpersonal comparisons), utility in this technical sense is quite distinct from any substantive idea of welfare. Indeed, utility in this sense can conflict with welfare if the preferences being represented are not valid welfare judgments: larger utility values can be assigned to alternatives which, though preferred by the agent, are worse in terms of his welfare or better than alternatives he rejects. Some have argued that the term utility should thus be confined to the technical meaning so as to avoid needless confusion resulting from the identification of utility with 'good' (e.g., Broome 1991a, 1991b). Perhaps so. But there are reasons to proceed with caution in this. First, utility continues to be associated with welfare—'being useful or beneficial'—in our everyday language. Even within economics, this everyday association still surfaces occasionally. More confusion may be caused by trying to banish it than by living with it.

Second, the identification of utility with welfare has a long history within both economics and philosophy (Sen 1987, 1991). Bentham, J. S. Mill, and other classical utilitarian philosophers held that utility means happiness or enjoyment in the sense of pleasure and absence of pain, for example, and that actions ought to be evaluated in terms of their consequences for general happiness or social welfare, maximization of which is the sole ultimate standard of good (see *Consequentialism Including Utilitarianism*). For these thinkers, any action or other thing is 'useful' or 'has utility' if and only if it has a tendency to promote pleasure net of pain. They deny (what is commonly asserted) that something may give much pleasure yet be useless, or that something may be useful even though it tends to promote more pain than pleasure. Earlier thinkers such as Hume and Helvetius also may have held this hedonistic idea of utility if not the distinctive utilitarian doctrine that general utility maximization is the *sole* ultimate test of good conduct.

As Mill (1965, pp. 45–6) elaborated the hedonistic idea, three kinds of utilities can be produced by human actions, namely: 'utilities fixed and embodied in outward objects' (such pleasures are created by 'investing external material things with properties which render these serviceable to human beings'); 'utilities fixed and embodied in human beings' (again, such enjoyments are made possible by 'conferring on human beings qualities which render them serviceable to themselves and others'); and, finally, 'utilities not fixed or embodied in any object, but consisting of a mere service rendered; a pleasure given, an inconvenience or a pain averted, during a longer or shorter time, but without leaving a permanent acquisition in the improved qualities of any person or thing; the labor being employed in producing an utility directly, not (as in the two former cases) in fitting some other thing to afford an utility.' Useful material objects and educated human beings are *means* to pleasure, of course. After repeated experience of their consequences, however, these more or less permanent sources of pleasure can become inseparably associated with the idea of happiness itself, and thus be seen as concrete ingredients of the good (Mill 1969, pp. 234–9).

Perhaps because utilitarians were so influential in the development of classical economics, the hedonistic interpretation of utility was also widely accepted by economists prior to the 'ordinalist revolution' of the 1930s and 1940s. Classical economists were preoccupied with the production and distribution of material wealth, however, and so tended to focus their attention on the kind of utilities embodied in material objects susceptible to accumulation. As used by Mill (1965, pp. 48–54), for example, the term 'productive labor' is confined to those types of actions that produce utilities embodied in material objects. Actions are classed as 'unproductive labor' (however useful to society) to the extent that they produce the other two kinds of utilities. Moreover, 'productive consumption' (i.e., investment) includes consumption by productive laborers of things which are necessary to maintain and increase their own health, strength, capacities of work, or population. But it does not include any consumption by unproductive laborers (those who do not produce material wealth) or consumption of luxuries by the productive themselves (however pleasing these forms of consumption are to the community). In short, although they did not deny that other kinds of utilities existed, economists in the classical tradition (from Smith and Ricardo to Marshall and Pigou) tended to identify utility with material welfare and to be especially concerned about the needs of those members of the community who produced the kind of utilities embodied in material wealth.

In the post-Second World War period, utilitarian ethics have declined in influence and hedonistic conceptions of utility have largely been abandoned by economists and philosophers. Whereas philosophers (including utilitarians) have typically gone on to identify utility with some nonhedonistic idea of welfare, however, economists have generally adopted the technical idea of utility in its purely ordinalist form—utility is merely a convenient mathematical tool for representing interpersonally noncomparable preference orderings (see *Rational Choice Explanation: Philosophical Aspects*). It is not obvious that philosophers should follow economists in viewing utility as a merely technical term, distinct from welfare. (For a controversial argument that this shift in the meaning of utility within economics—the product of the 'ordinalist revolution'—does not imply 'progress' because it has involved 'losses in understanding human welfare,' in particular, 'material welfare,' including any person's objective interpersonally comparable material needs; see Cooter and Rappoport 1984.) But perhaps they should, without necessarily accepting the narrow restriction to pure ordinalism. Even so, it would be incorrect to conclude that because technical utility values are distinct from, and can conflict with, substantive welfare values, there is some inescapable contradiction between utility in the technical sense and welfare. This leads to a third reason for caution in insisting on a divorce between utility and welfare, to wit, the technical idea of utility may remain a formal aspect of any adequate conception of welfare if judgments about welfare are a species of preferences that can be represented by utility functions. In that case, technical utility, as a formal component of the idea of good, could be seen as a vestige of earlier utility concepts that identified utility with welfare.

Whether or not judgments about welfare can be represented by utility functions, the question remains: what are such judgments about?

2. *Conceptions of Welfare*

At least four different conceptions of welfare continue to be debated in the literature.

2.1 *Welfare as Pleasure*

Although currently out of favor, hedonistic conceptions have not vanished entirely. Hedonism is typically dismissed as crude and naive, however, both as a psychology and as a theory of welfare. Expectation of pleasure is not the sole motivational force in human nature, it is said, and human good involves myriad things other than pleasure. No rational and moral person would be happy, for example, if his life consisted solely of a stream of supreme pleasures guaranteed by artificial means.

Perhaps such arguments are decisive against some versions of hedonism. But, as already discussed, the classical hedonism of Mill, Bain, and others does not deny that many things are means to pleasure which can also become valued as ends-in-themselves, in which case they are seen as concrete ingredients of happiness. Classical hedonism is not a purely subjective doctrine. It insists that some kinds of pleasures are embodied in material objects and human qualities, and can only be experienced by people who possess those objects and qualities, which must first be produced. Particularly intense intellectual and moral pleasures can only be experienced by agents who possess the relevant intellectual and moral capacities, for example, and those capacities must be developed in part by the agents themselves.

2.2 *Welfare as Preferences*

Perhaps currently most influential, preference-based approaches typically conceive personal welfare in terms of personal preferences and social welfare in terms of social or ethical preferences that are constructed in some way from the set of personal preferences. Nevertheless, there is an ambiguity in the key notion of 'preference-satisfaction': it may mean the fulfillment of preference by a suitable shift in the way the world is; or it may mean the satiation of preference by a suitable shift in the subject's conception of the way the world is. Proponents of preference-based approaches typically insist on the latter 'satiation' interpretation because of their implicit subjectivism. Fulfillment without satiation (e.g., posthumous fulfillment of preferences) does not count. From this subjectivist perspective, hedonism can be viewed as a special case, in which a claim is made that expectation of pleasure is the sole *ultimate* motivation of preference. But defenders of preference-based conceptions usually reject that claim as implausible.

Two versions of the preference-based approach should be distinguished. One defines 'preferences' as actual or hypothetical 'choice behavior,' which is in principle observable under certain conditions (see, e.g., Harsanyi 1992, Binmore 1994, 1998). A person's preference for x over y is 'revealed' by his choice of x when y is available, independently of his desires or other motivations for the choice. In contrast, the other version defines preferences as 'desires' rather than 'choices' (see, e.g., Hare 1981, Griffin 1986). A preference for x over y implies greater desire-satisfaction (though not necessarily pleasure) at x than at y. Nevertheless, the agent could—and perhaps should—still make a counterpreferential choice of y over x in this case. Of course, the notion of 'counterpreferential choice' makes no sense under the choice-based approach. But that approach is silent about the motivations people have for making the choices it views as comprising personal or social

welfare. It needs to be supplemented by a theory of motivation, not only to explain why people do make those choices but also to justify why those choices ought to be seen as conducive to welfare. For its part, the desire-based approach does not lack a theory of motivation. But its theory may well be inadequate. It is arguable that people are not always motivated to choose the options that they expect will bring them the greatest desire-satisfaction, nor should they be. Perhaps they should be motivated to sacrifice their greatest desires because of strategic considerations, for example, or moral commitments (Sen 1977).

However preferences are interpreted, the main appeal of preference-based conceptions seems to reside in their implicit subjectivism: the welfare of any individual or society is judged from the subjective viewpoint of that particular individual or society. The contents of welfare judgments may vary significantly across different persons or social groups (see *Diversity and Disagreement*). Nevertheless, the appeal of any purely subjective interpretation of welfare is quite limited. People may be deluded, for example, or ill-informed, or resigned to their customary positions as slaves or 'untouchables' in highly inegalitarian societies (see *Adaptive Preferences: Philosophical Aspects*; *Irrationality: Philosophical Aspects*). Their personal preferences may thus be poor indicators of personal welfare. Moreover, people may be unable to imagine themselves in others' shoes, or they may clash over interpersonal comparisons, or display undue bias against others' interests. As a result, they may fail to construct ethical preferences that can serve as social welfare judgments.

Defenders of preference conceptions typically respond to these difficulties by suitably laundering preferences. Conditions are specified that a person's choices or desires must satisfy before they can be accepted as conducive to his own welfare or that of his society. Thus, for example, Harsanyi (1992) suggests that a person's choices are conducive to his own welfare only if they are informed and conform to the expected utility axioms, whereas his choices are conducive to social welfare only if, in addition, they are ethical. Ethical choices must conform to the equiprobability model of moral choice. The person must forget his own position (defined broadly to include family history and genetic makeup) and imagine himself occupying the position of any person in society with equal probability. He must also extend sympathy to each person j by making the same 'self-oriented' choices j would make while in j's shoes. He can then infer interpersonal comparisons of cardinal utility which—on the independent assumption that all persons share a fundamentally similar human nature (including attitudes toward risk used to measure the relative importance of different choices), which is revealed to anybody who fights against his biases by adopting the 'impersonal' perspective of the equiprobability model—are identical to those inferred by anybody else who goes through this procedure. Finally, he must aggregate impartially over the cardinal comparable utilities he experiences at each person's position to construct social welfare judgments, which turn out to be (average) utilitarian in form. (For difficulties surrounding the interpretation of Harsanyi's theorems as derivations of utilitarianism, see, e.g., Sen 1986, pp. 1122–4.)

Even if such hypothetical preferences have appeal as conceptions of welfare, the subjectivism which is usually thought to be the main strength of preference-based approaches is lost unless agents actually affirm the hypothetical preferences under the relevant conditions. To preserve an element of subjectivism, individuals and societies must willingly adopt those preferences before they can be said to be conducive to the welfare of those particular individuals or societies. If actual adoption of the preferences is a necessary condition of good, however, attention is redirected to theories of evolution and education which explain how people can be motivated to develop the intellectual and moral capacities required to make the relevant choices or form the relevant desires (see *Evolution, Natural and Social: Philosophical Aspects*). It is not clear whether a credible theory of this sort can be supplied to anchor the desire-based variants. But whether or not human capacities can be developed such that human beings invariably obtain the greatest desire-satisfaction from making the choices that comprise personal or social welfare, defenders of choice-based versions of the preference conception require a credible theory of motivation.

2.3 Welfare as Resources

In resource-based conceptions, personal welfare is defined in terms of an objective list of resources broadly defined, whereas social welfare consists implicitly of the effective production and fair distribution of the relevant resources (see *Justice: Philosophical Aspects*). Rawls (1971, 1993) can be read as offering such a conception, for example. He refers to 'primary goods' as 'things which every rational man is presumed to want' and 'things citizens need as free and equal persons,' and provides a (perhaps incomplete) list that includes 'basic rights and liberties,' 'freedom of movement and free choice of occupation against a background of diverse opportunities,' 'powers and prerogatives of offices and positions of responsibility in the political and economic institutions of the basic structure,' 'income and wealth,' and 'the social bases of self-respect' (1971, pp. 60–5, 1993, pp. 178–90). Dworkin (1981, 1990) spells out another approach of this sort.

There is little doubt that human welfare depends on the possession of such valuable things as those listed by Rawls or Dworkin. Data on the extent to which a person possesses these items may provide helpful

indicators of different aspects of his good, even if the indicators cannot be combined into a single index of personal welfare. Similarly, data on the social distributions of these things may provide indicators of different dimensions of social welfare, with the same caveat.

Nevertheless, resource-based conceptions face at least two difficulties. First, it seems that the relevant resources are conducive to welfare only because they are objects of suitably laundered preferences. As Rawls suggests, for example, primary goods are things which every person would prefer to have if he were rational and reasonable. Furthermore, since a person's holdings of one resource (say, political responsibilities) can interfere with his holdings of another (say, wealth), it seems that his preferences must settle which balance of resources is proper for his welfare. Similarly, it seems that social or moral preferences are needed to settle which of many competing social distributions of resources are permissible if not essential to social welfare. If this is right, resources may well be essential ingredients of welfare but they properly figure as objects of preferences.

The second difficulty is that resource-based conceptions do not take account of interpersonal variation in human capacities to convert resources into the achievement of ends (Sen 1990). This problem leads to another way of conceiving welfare.

2.4 Welfare as Capabilities

Sen (1985a, 1985b, 1990) argues that human well-being should be conceived in terms of 'capabilities'—or 'positive freedoms' (see *Freedom: Political*)—to achieve valuable human 'functionings' given by some objective list. He points out that different people may have unequal capabilities to convert resources into achievements. Physically handicapped people may need more income than the able-bodied need to achieve a given level of mobility, for example. People who are genetically vulnerable to particular diseases may need more medicines than the less vulnerable need to achieve a given standard of health. Women may need more food than men need to achieve a given standard of nutrition. Racial minorities may need more educational and job opportunities than white males need to achieve a given level of career success if racist attitudes are prevalent. The elderly may need more wealth than the young need to achieve a given standard of living. These are essentially empirical questions.

Such examples suggest that capability-based conceptions can remedy a serious deficiency in resource-based approaches to welfare. But, as with resources, it seems arguable that certain capabilities are properly viewed as essential components of welfare because they are objects of suitably laundered preferences. Some capabilities are in effect things that every rational or moral person is presumed to want or choose. To the extent that anybody lacks these benchmark freedoms, reasonable people may agree that some form of special compensation ought to be provided so that the person can function 'normally,' for example, income subsidies, extraordinary medical assistance, affirmative action in hiring decisions, and so on. Furthermore, beyond the benchmark capabilities which everyone is presumed to need, it seems that a person's preferences must determine which balance of extra capabilities is proper for his welfare since capabilities to achieve some functionings can presumably interfere with capabilities to achieve others. Again, since one person's capabilities might interfere with another's, social or moral preferences are needed to settle which of many competing social distributions of capabilities (including benchmark capabilities) is permissible if not essential to social welfare.

If this is right, it seems that some version of the preference-based conception of welfare may have appeal as a way to combine insights from the various approaches discussed. The upshot is that preferences must be suitably laundered not only to satisfy certain formal conditions of rationality and morality but also certain substantive conditions. In particular, preferences can be welfare judgments only if they give due weight to certain valuable resources, making due allowance for interpersonal variation in capabilities. Moreover, a credible theory of motivation is required to explain how individuals can come to hold the relevant preferences, even if hedonism is rejected.

If some version of the preference approach has such potentially broad appeal, *welfarist* moral philosophies might adequately capture the phenomenon of human welfare. This will be the case if the relevant preferences can invariably be represented by technical utility functions, and if utility information is all that matters for the construction of social or ethical preferences.

3. Welfarist Philosophies

Welfarist philosophies are distinguished by their exclusive reliance on utility information. Perhaps the most well known examples of such philosophies are the standard versions of utilitarianism. But there are many different forms of welfarist philosophies. They can be clarified by means of the concept of a *social welfare functional* (SWFL) (Sen 1986, pp. 1111–28). For a given set of alternatives, any SWFL maps a given set of personal utility functions (one for each person in society) into a social or ethical preference which can itself be represented by a technical utility function. Since each person's utility function is an element of some family of informationally-identical utility functions as determined by the given assumptions of measurability and comparability, however, the SWFL is required to satisfy an invariance re-

quirement: it must generate the same social or ethical preferences if any given set of personal utility functions is subjected to transformations permitted by the measurability and comparability assumptions.

Welfarism is a condition of strong neutrality imposed on the form of a SWFL. It requires that if each person's utility values with respect to one pair of options x and y are equal to his utility values with respect to another pair w and z—subject to transformations permitted by the given measurability and comparability assumptions, then the SWFL must generate the same social or ethical preference over the one pair as over the other. The only information that matters for a welfarist SWFL is utility information. Nonutility information has no influence.

Arrow's (1963) well-known 'impossibility theorem' shows that any welfarist SWFL must generally be dictatorial if personal utility functions are ordinal and noncomparable. But a dictatorial procedure cannot generate social or ethical preferences that give positive weight to each person's utilities: the dictator sympathizes with nobody but himself in determining social preferences. (For variations and extensions of Arrow's theorem, including its extension to the case of cardinal noncomparable utilities, see Sen 1986, pp. 1073–106, 1114–15.) It follows that welfarism can have appeal as a philosophy of human welfare only if utility information is sufficiently rich to support the interpersonal comparisons required to construct ethical preferences. As Harsanyi's equiprobability model of moral choice illustrates, one person can be assumed to replicate the operation of a welfarist SWFL over a set of comparable personal utility functions, without that person becoming an Arrovian dictator.

Although various forms of welfarist philosophies making use of comparable utilities have been investigated in the literature, doubts remain that any adequate philosophy of human welfare can be put into a welfarist form.

4. Beyond Welfarism

Even if some version of the preference-based conception turns out to have broad appeal, a nonwelfarist philosophy—one that relies to some extent on nonutility information—may be required to capture that conception of human welfare. Myriad forms of nonwelfarist philosophies are conceivable. Such a philosophy may take the form of a SWFL that does not satisfy the relevant strong neutrality condition. The SWFL might make use of nonutility information about the alternatives—some alternatives might belong to an individual's 'protected sphere' of choice, for example, or some might involve 'primary goods'—to generate ethical preferences that can be represented by utility functions. Or it might rely on nonutility information to distinguish between different kinds of utility functions representing different kinds of preferences, some of which are held to be inherently more valuable than others (Riley 1988, 2001). The relevance of any particular nonutility information is determined by the conception of welfare under consideration. In the latter case, for example, the information that different kinds of utility functions ought to be identified and arranged in a lexical hierarchy is supplied by the idea of welfare itself, which is taken to have this sort of multidimensional structure.

Nonwelfarist philosophies may also eschew the form of an SWFL. Harsanyi's (1992) version of rule utilitarianism takes the form of a two-stage game, for example, in which a cooperative game is played at the first stage to select an optimal moral code and a noncooperative game is played at the second stage where players are restricted to strategies permitted by the optimal code. This game structure gives a special status to whatever outcome is assumed to prevail in the absence of agreement at the first stage. Also, the rules of any selected code will govern right and wrong actions independently of any strong neutrality condition. Alternatively, a nonwelfarist philosophy might take a value pluralistic form (see *Value Pluralism*), in which welfare is only one of plural irreducible values in ethical evaluations. Sen (1985a, 1985b) proposes a dualistic structure in which the value of 'well-being' cannot be reduced to that of 'agency,' for instance. By implication, incompleteness or even inconsistency might arise in all-things-considered ethical preferences if well-being comes into conflict with agency.

See also: Economics and Ethics; Equality: Philosophical Aspects; Ethics and Values

Bibliography

Arrow K J 1963 *Social Choice and Individual Values*, 2nd edn. Wiley, New York

Barry B 1973 *A Liberal Theory of Justice*. Clarendon Press, Oxford, UK

Binmore K 1994 *Playing Fair*. MIT Press, Cambridge, MA

Binmore K 1998 *Just Playing*. MIT Press, Cambridge, MA

Broome J 1991a Utility. *Economics and Philosophy* **7**: 1–12

Broome J 1991b Utility—ideas and terminology—reply. *Economics and Philosophy* **7**: 285–7

Cooter R, Rappoport P 1984 Were the ordinalists wrong about welfare economics? *Journal of Economic Literature* **22**: 507–30

Dworkin R 1981 What is equality? Part 2: Equality of resources. *Philosophy and Public Affairs* **10**: 283–345

Dworkin R 1990 Foundations of liberal equality. In: McMurrin S (ed.) *The Tanner Lectures on Human Values*. University of Utah Press, Salt Lake City, UT, Vol XI

Griffin J 1986 *Well-being*. Clarendon Press, Oxford, UK

Hare R M 1981 *Moral Thinking*. Clarendon Press, Oxford, UK

Harsanyi J C 1992 Game and decision theoretic models in ethics. In: Aumann R J, Hart S (eds.) *Handbook of Game Theory*. Elsevier, Amsterdam, Vol 1

Mill J S 1965 Principles of political economy. In: Robson J M (ed.) *Collected Works of J.S. Mill*. University of Toronto Press

and Routledge & Kegan Paul, Toronto and London, Vols II–III pp. 48–54
Mill J S 1969 Utilitarianism. In: Robson J M (ed.) *Collected Works of J. S. Mill*. University of Toronto Press and Routledge & Kegan Paul, Toronto and London, Vol X pp. 234–9
Rawls J 1971 *A Theory of Justice*. Havard University Press, Cambridge, MA
Rawls J 1993 *Political Liberalism*. Columbia University Press, New York
Riley J 1988 *Liberal Utilitarianism*. Cambridge University Press, Cambridge, UK
Riley J 2001 *Maximizing Security: A Liberal Utilitarian Theory of Justice and Rights*. Clarendon Press, Oxford, UK
Sen A K 1977 Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* **6**: 317–44
Sen A K 1979 Interpersonal comparisons of welfare. In: Boskin M J (ed.) *Economics and Human Welfare*. Academic Press, New York
Sen A K 1985a *Commodities and Capabilities*. Elsevier, Amsterdam
Sen A K 1985b Well-being, agency and freedom. *J. Philos*. **82**: 169–221
Sen A K 1986 Social choice theory. In: Arrow K J, Intriligator M D (eds.) *Handbook of Mathematical Economics*. North-Holland, Amsterdam, Vol III pp. 1073–1128
Sen A K 1987 *On Ethics and Economics*. Basil Blackwell, Oxford, UK
Sen A K 1990 Justice: means versus freedoms. *Philosophy and Public Affairs* **19**: 111–21
Sen A K 1991 Utility: ideas and terminology. *Economics and Philosophy* **7**: 277–83

J. Riley

Welfare Programs, Economics of

All more developed nations have created some set of means-tested public assistance or welfare programs, designed to protect vulnerable groups from economic destitution. Among the groups that are often targeted by such programs are elderly, poor children, disabled persons, and low-income families, especially single-mother families. In many cases, these programs are designed to provide an income supplement to those who cannot work or earn income. Other public assistance programs may be designed to support low-wage workers and help make it economically viable for them to continue in employment. As we shall see, there is often a conflict between the goal of subsidizing income versus encouraging work, which often spills over into controversy about the design and operation of welfare programs. In part because of this controversy, welfare programs have been the focus of ongoing public and political discussion in many nations. This has provided an incentive for extensive research among social scientists about the effects of such programs.

1. Traditional Welfare-program Design

The design of a traditional cash-benefit program can be seen in Fig. 1, which shows the budget constraint under a typical cash-benefit program. This welfare program pays a benefit (G), often called the guarantee level, to those at the lowest levels of income. In Fig. 1 we show G paid to those who work zero hours, and assume that those who do not work have little other income. Of course, some families might have other income that would make them ineligible for welfare benefits, even if they were not working. As hours of work increase and earnings rise (at a given wage, w), the welfare benefit declines at a rate t, the benefit reduction rate or benefit tax rate, until at some income level (Y^* in Fig. 1, at H^* hours of work) benefits disappear entirely. This is point A in Fig. 1 and is often referred to as the break-even point.

Several aspects of the design of this welfare program are important. Note first that the maximum benefit payment occurs at the lowest hours of work. On the one hand, this assures that the poorest families receive the most aid. On the other hand, this creates disincentives to work, by increasing the income available to nonworkers from 0 to G. Note second that the rate at which income increases for families below the break-even point is lower than the rate at which earnings rise. For every hour worked between 0 and H^*, earnings increase by the hourly wage (w), but income increases only by $w(1 - t)$. This is because public benefits are reduced at a rate t for every hour worked. This is equivalent to a reduction in the hourly wage; for every hour worked, these families earn only $w(1 - t)$ rather than w. Standard analysis of income and substitution effects suggests that wage changes have theoretically ambiguous effects on work incentives, but the empirical literature indicates that wages are positively correlated with labor supply among low-wage workers. This means that a lower wage will lower labor supply.

The enactment of a welfare program similar to that

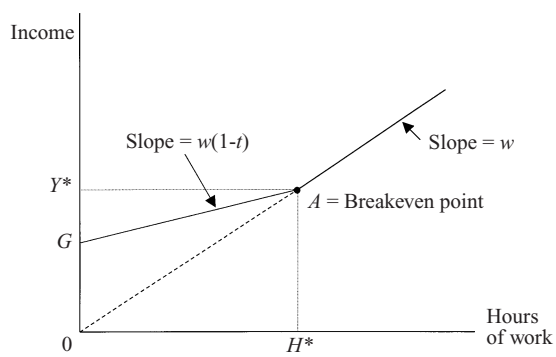

Figure 1
Budget constraint under a traditional cash-benefit program

described in Fig. 1 is expected to reduce labor supply, by increasing the number of persons who choose the zero hours point, as well as by reducing hours of work among those who would have worked somewhere between 0 and H^* hours in the absence of a welfare program. In addition, because of the 'kink' in the budget constraint at point A, some people whose hours would be greater than H^* in the absence of the welfare program will find it preferable to drop to a point below H^*.

One way to view the trade-offs inherent in this type of welfare program is to realize that this program essentially contains two policy parameters—the maximum level of benefits, G, and the benefit reduction rate, t. But there are often at least three goals for these programs: to provide adequate income support to those who are not expected to work, to provide work incentives for those who can work, and to keep government spending at a reasonable level. If there is an increase in the benefit level G, this improves income support, but reduces work incentives and increases government spending. If the benefit reduction rate t is lowered, this moves the break-even point up the income distribution, to higher levels of income and higher hours of work. The result is an increase in work incentives for those at zero or low hours of work, but this will also induce more welfare participation among those near the new (higher) break-even point. It will also increase government expenditures, by expanding eligibility to more families.

A wide variety of empirical studies have attempted to estimate the impact of benefit levels and benefit reduction rates on labor supply. In fact, in the 1970s, the US government funded a series of randomized experiments known as the negative income tax experiments, which compared the impact of different levels of G and t (Robins 1985, Munnell 1987). (A cash-assistance program as outlined in Fig. 1 that is run through the tax system is known as a negative income tax; in this case one receives a payment if income is below a certain level, and pays taxes if income is above a certain level.) Other researchers have used variation in benefit levels and effective tax rates across states in the USA to look at the impact of these parameters. Moffitt (1992) summarizes this literature. His conclusions are that the primary cash support program in the USA for low-income families (Aid to Families with Dependent Children or AFDC) created nontrivial work disincentives. (AFDC was replaced with a block grant known as Temporary Assistance to Needy Families or TANF in 1996.) A midpoint estimate of the impact of AFDC is that it reduced work by about 5.4 hours per week, a 30 percent reduction in work effort.

There are circumstances under which these tradeoffs between income support and work incentives may be less problematic. For instance, for a population from whom little or no work is possible or expected, it may be easy to set a high G (allowing greater income support), along with a high t (creating a break-even point at a low level of hours). The limited work incentives in this design are unimportant if there is no desire to encourage work among the eligible population. An alternative way to avoid these problems is to institute a mandatory work program for welfare participants who are judged to be able to work. For example, a wide variety of states in the USA have experimented with what are typically called 'workfare' programs, which mandate that work-eligible recipients must work a certain number of hours in publicly created jobs in order to maintain eligibility for their benefits. These and other work-incentive programs are discussed below.

Researchers have also noted that these tradeoffs between work and cash support may be less problematic in a situation where there is 'stigma' associated with being on welfare. Moffitt (1983) has modeled this situation, showing that if persons believe there is a cost to being on welfare (such as the stigma of being looked down on by one's neighbors), this will make persons less likely to use it, and will shift welfare participation only to the neediest cases. If stigma is greater among those who are more able-bodied and work-ready, and less among needier and less work-able persons, this will reduce the work disincentives of the program, while protecting its income-support aspects. Of course, if stigma is also high among needy persons who are not able to work, this could make the program less effective as a social safety net. It may be difficult to target efforts to increase participation among the neediest without also reducing stigma and increasing participation among the less needy as well.

Stigma is only one reason why participation rates (often called take-up rates) among those eligible for a program are often far below 100 percent. There may be costs associated with participating, related to the regular reporting and certification process. There may also be a lack of information about the program among eligibles. It is not unusual to find that close to half of those eligible for a particular public-assistance program do not participate in it (Blank and Ruggles 1996).

While much of the research literature has focused on the labor supply effects of traditional welfare programs, other concerns have also been raised about such programs. For instance, welfare programs which provide support for mothers with children may increase the incidence of divorce or out-of-wedlock childbearing. The literature on these effects is much more controversial, and also more limited in terms of empirical sophistication. In general, there is little evidence of strong benefit effects on family fertility or female headship, but there are also methodological problems with much of the literature on this topic. (Moffitt 1992 also discusses these issues. A more recent review occurs in Robins and Fronstin 1996.) Disagreement over the impact of the negative income tax experiments on divorce and separation has also

resulted in more recent estimates suggesting the impacts were relatively small (Cain and Wissocker 1990).

2. *In-kind Benefits*

So far, the discussion has presumed that welfare programs pay out cash benefits, which are theoretically indistinguishable from other sources of income. Many welfare programs provide in-kind benefits, however, which are available only for particular types of consumption. For instance, in the USA, the food-stamp program provides special coupons which recipients can use in lieu of cash at the grocery store. Similarly, housing assistance programs typically subsidize rent, or child-care assistance programs subsidize child-care costs. (For a description of the research relating to food stamps, see Ohls and Beebout 1993; for a similar survey of the literature on housing subsidies in the USA, see Newman and Schnare 1992.)

In most cases, the same issues that arise in traditional cash-benefit programs also arise in in-kind programs. These in-kind programs provide subsidies, which are then reduced in some manner as incomes rise. This raises the same questions about income support, work incentives, and incentives for family formation.

In-kind benefits are a more constrained form of a subsidy than cash benefits, however. Figure 2 shows the income constraint that emerges in the presence of a food subsidy. On the X-axis is food consumption and on the Y-axis is all other consumption. The food-stamp program shifts the income constraint from F_1–C_1 to F_2–C_2'. Note that the line segment C_2'–C_2 is *not* part of the new budget constraint. The additional resources expand food purchases only, and cannot be used to expand other forms of consumption.

If these in-kind subsidies are relatively modest, they should cause little distortionary behavior. Individuals who were choosing point A on the income constraint F_1–C_1, now move to point A' on the income constraint F_2–C_2', somewhere within the northeastern quadrant to the right and top of point A. Note that at A', a family consumes both more food and more other goods, that is, they substitute some of their earlier (cash) spending away from food to other goods. If, however, the food subsidy is relatively large or the family spends little cash on food, there is some possibility that the family would prefer to be on the C_2'–C_2 segment of the budget constraint, rather than on the F_2–C_2' segment. In this case, the in-kind program constrains expenditures and forces families to choose the consumption point C_2', where they consume more food than they would if they had received a cash benefit rather than an in-kind benefit. Indeed, research in the USA, which compares food stamps to a program that provides an equivalent subsidy in cash, concludes that families who receive cash reduced their food expenditures by between 20 cents to 25 cents per dollar of food-stamp benefits, suggesting that the in-kind form of the benefit forced families to purchase a higher quantity of food than they would with an equivalent amount of cash (Fraker et al. 1995). In contrast, for elderly or disabled food-stamp recipients, there was little difference in their consumption when given food stamps vs. an equivalent amount of cash.

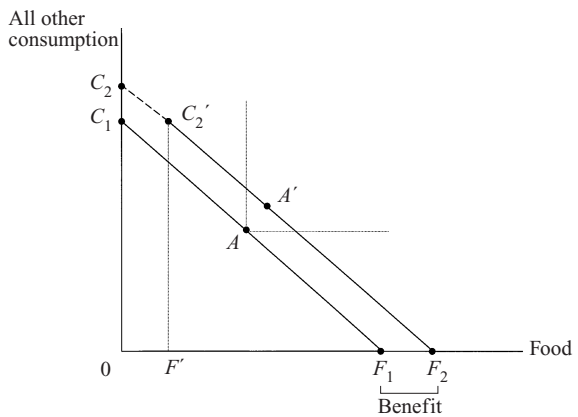

Figure 2
Income constraint with in-kind benefits

Economists have traditionally believed that in-kind programs are inefficient, since a cash transfer of an equivalent amount could make a family at least as well off, and might make them better off if they want to allocate more of their spending to another category of consumption. Those who support in-kind benefits tend to make two arguments in response. First, there is often more widespread support for in-kind programs than for cash programs. That is, the median taxpayer might be more willing to provide food stamps than general cash assistance. This raises the question of whose preferences are being maximized in poor support programs. While lower-income people might prefer less constrained subsidies, taxpayers might prefer to provide constrained assistance, in order to limit the extent to which public dollars are used to subsidize behavior that taxpayers believe is inappropriate. If welfare programs are meant to be responsive to taxpayer concerns, in-kind programs might be a reasonable way of providing support. Second, many have noted that in-kind programs often come attached to their own 'lobby.' For instance, food stamps are heavily supported by agricultural interests, while public health insurance is heavily supported by the health industry. This may provide a level of political support and protection for these more targeted in-kind programs—support that general cash-assistance programs often lack. In the midst of regular waves of budget-cutting fervor at both the state and national levels, public assistance programs are often vulnerable.

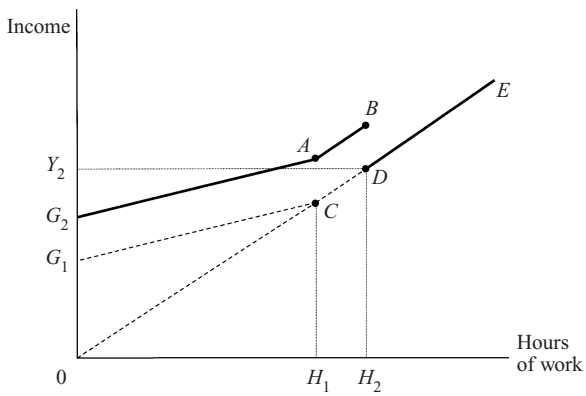

Figure 3
Notched budget constraint under a traditional cash-benefit program with added in-kind benefits

A series of targeted, smaller, in-kind programs may be more politically viable and enduring in the long run than a single, large cash-assistance program (such as a negative income tax).

In-kind programs often raise another problem as well. Many of them are designed to phase out more abruptly than cash programs (although cash programs may be designed to phase out quickly as well). For instance, housing-subsidy programs often pay substantial benefits up to a particular income level and then cease, rather than slowly lowering benefits to a continuous break-even point. This creates a 'notch' in the budget constraint.

Figure 3 graphs a budget constraint in which a traditional cash-benefit program is supplemented by an in-kind benefit program for some specific good (this could be a program providing supplemental assistance to pay for food, housing, child care, or any other specified good.) In Fig. 3, the in-kind program provides a benefit of G_2–G_1, which is available to all families with incomes below Y_2 and who work less than H_2 hours at the specified wage w. Once income ses above Y_2, the in-kind benefit ceases. In this case, the cut-off point for in-kind benefits (D) is above the break-even point for receipt of cash benefits (C). The result is a very oddly shaped budget line, G_2–A–B–D–E, with a notch at B–D when the in-kind benefits cease. Standard economic analysis suggests that a number of families will choose point B and resist increasing their hours above H_2, since that means the end of their in-kind subsidy. This causes a 'welfare trap,' in which increases in work can actually produce a decline in income. Depending on the size of the notch, such a trap might be a real problem for policy makers who want to increase people's employment and economic self-sufficiency.

Public-assistance systems with multiple-benefit programs layered on top of each other typically have some notches or welfare traps imbedded in them. In addition, high marginal tax rates can result when multiple programs have phase-out ranges in the same distribution of earnings, each acting as a separate tax on earnings. Blinder and Rosen (1985) discuss the pros and cons of 'notched' budget constraints. Under some circumstances, it may be preferable to create a notch rather than to reduce benefits more continuously, particularly if a slow benefit reduction would greatly increase potential eligibility for the program and a fast benefit reduction would create high marginal tax rates over a range of income.

3. *Programs that Subsidize Work*

Increased concern with the work disincentives imbedded in traditional welfare programs has led to a series of changes in the structure of public assistance programs in the USA, and similar changes are occurring in other countries. This concern is driven by a variety of factors. Increasing labor-force participation among women at all skill levels has made government support for mothers who stay at home less politically popular. In the USA growing public-assistance caseloads and a growing concern with the problems faced by low-income single-mother families have led to increasing concern about finding ways to help these families permanently to escape poverty and avoid welfare 'dependency.' (Blank 1997a discusses many of these issues.) In Europe, sustained high unemployment rates have also increased caseloads and increased public spending on welfare-related programs. The result has been a growing emphasis on work-oriented welfare reform. This has taken several forms.

3.1 *Welfare-to-work Employment and Training Programs*

In the 1990s, major efforts were made to meld job-search and employment programs with traditional welfare programs. In some states, this took the form of mandatory welfare, or work for your benefits, as noted above. Other states in the 1980s began to experiment with serious job search and job placement programs aimed at welfare recipients. The result of these efforts has been a revolution in the way that public assistance offices operate and what services they deliver. In addition, a variety of very well-designed evaluations of these efforts reinforced the belief that such policies could increase work and earnings, and decrease public-assistance spending. (The interaction between these experimental results and political decision making is reviewed in a set of articles in the 1991 *Journal of Policy Analysis and Management* **10**.)

These evaluations indicated that even rather minimal efforts at job-search assistance seem to produce employment and earnings gains among women on welfare (Gueron and Pauly 1991). The returns from

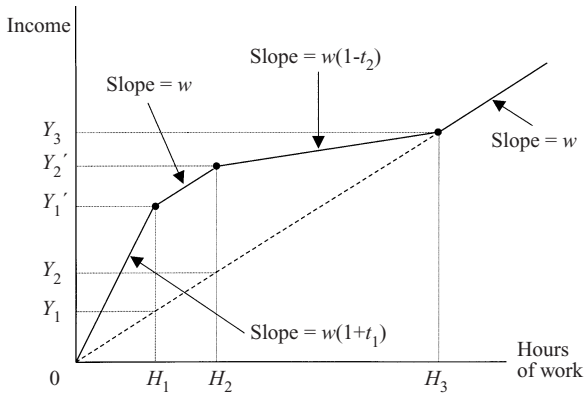

Figure 4
Budget constraint under a wage-subsidy program

these programs are usually greater than the costs to the government budget, making them cost-effective policies. The resulting earnings increases are not huge—at most around $1,000 per year—but they represent a large percentage increase in earnings among women, many of whom had little earned income to start with.

Since 1996, US states have had almost full discretion to design their own cash public-assistance programs. The result is a wide variety of experiments, many of which meld work and welfare together even more closely. For instance, many states have lowered their benefit reduction rates, increasing work incentives for those on welfare. Many states have also expanded their emergency-assistance programs (now often called 'diversion activities'), which help welfare applicants deal with immediate problems, such as an ill family member or a housing eviction, without entering them permanently on the welfare rolls. All states are mandating that a growing share of their caseload participate in job-search and job-placement programs. Many states have also instituted serious sanctions, cutting benefits among families who do not respond to the mandate to enter a job-search program. (For a discussion of these changes, see Blank 1997b, Gais and Nathan 1998, or Gallagher et al. 1998.)

This movement to a more work-oriented welfare system has reduced caseloads and increased labor-force participation by single mothers. It has also raised a host of new policy-related questions, such as whether former welfare recipients will be able to maintain their jobs during an economic downturn, and whether women who initially enter very low-wage positions will be able to increase their earnings over time, becoming more economically stable and self-sufficient.

3.2 Wage Subsidies to Support Work

In the face of increasing demands that welfare recipients move into work as quickly as possible, the question has emerged, 'How do we design a safety net that supports work?' If one believes that work—even in unstable jobs and at low wages—is preferable to welfare, then providing the public support that enables steady workers to support their families becomes increasingly important.

A major innovation in the USA, which has been very important in increasing the returns to work and raising labor-force participation among adults in low-income families, is a wage-subsidy program called the Earned Income Tax Credit (EITC). The structure of the EITC is very different from traditional welfare programs. It provides a subsidy to low-wage workers in low-income families, based on family income but conditioned on work.

Figure 4 graphs the budget constraint produced by the EITC. In contrast to traditional welfare programs (where a nonworker receives the maximum level of benefits), a nonworker gets zero benefits from the EITC. As work hours increase, the EITC subsidizes the wage at a rate t_1. Hence, the effective wage rate faced by the worker rises from w to $w(1+t_1)$. When earned income reaches Y_1 (at H_1 hours of work), the subsidy reaches its maximum level ($Y_1'-Y_1$). It stays at this level while earned income increases from Y_1 to Y_2, providing a flat subsidy to the worker. As earned income rises past Y_2, the subsidy is phased out at a rate t_2, which lowers the wage rate to $w(1-t_2)$. The subsidy ends entirely at the break-even point when income reaches Y_3.

Note several things about the design of the EITC. First, its maximum subsidy levels occurs at incomes between Y_1 and Y_2 (hours between H_1 and H_2). In order to reach the maximum subsidy, a person must work some (substantial) amount of time. Second, the work incentives imbedded in the EITC vary across different segments of the EITC budget constraint. For nonworkers or very low-hour workers the impact is positive, as the EITC raises wage rates. For workers with earnings greater than Y_1, the impact is likely to be more negative, particularly in the benefit reduction range when the effective wage rate is reduced.

These theoretical observations are consistent with the empirical research to date, which suggests that the EITC has had a strong positive effect on work effort among single mothers, who tend to be nonworkers or low-hour workers (Eissa and Leibman 1996), but has had a small negative effect on the work effort of wives, whose family income is more likely to be higher, putting them in the benefit reduction range (Eissa and Hoynes 1998). These labor-supply effects may be somewhat tempered by the complex design of the EITC. It is possible that few low-income workers fully understand its structure or can easily calculate how the EITC affects their own incentives.

It is also worth noting that the EITC and the minimum wage reinforce each other in useful ways. The minimum wage provides an income base to workers, which means that the EITC subsidy need not

be as large. This is important, because the larger the maximum subsidy the more difficult it is to do the benefit reduction. At a high subsidy level, if a relatively low benefit reduction rate (t_2) is chosen, the break-even point for the EITC would occur far up into the earnings and income distribution, substantially increasing the number of persons who receive the subsidy and making it less targeted to low-income workers. On the other hand, if a high benefit reduction rate is chosen quickly to reduce the subsidy, then the work disincentives in that part of the budget constraint are greater. The minimum wage assures that low-income workers can reach a target level of earnings without the problem of how to 'claw back' a very large EITC subsidy. Similarly, the EITC reinforces the higher take-home pay provided by the minimum wage, and allows the level of the minimum wage to remain lower, while still increasing income among low-income workers. This presumably reduces the potential disemployment effects of the minimum wage.

The importance of recent increases in the USA in the EITC and the minimum wage together can be easily calculated. The earnings of a single mother with two children, working full-time at the minimum wage and receiving the EITC, would have risen from $9,856 in 1989 to $14,056 in 1998 (in 1998 dollars), a 40 percent increase due entirely to these policy changes. This is a substantial addition to earnings for low-income families. It is not surprising that these changes appear to be causally related to the increase in labor-force participation among single mothers (Meyer and Rosenbaum 1998).

3.3 *Additional Work-support Programs*

While job-search assistance programs and wage-subsidy programs are often discussed as a primary way to encourage work, most countries also operate a variety of other programs that help or hinder the movement of less skilled persons into employment.

Particularly for many single mothers, the cost of childcare can substantially increase the expense of working (or, alternatively, reduce the take-home pay generated by employment). Public subsidies to offset child-care expenses among working low-income mothers can increase work incentives. If childcare is viewed as an hourly expense (call it c), the returns to work are $(w - c)$. If a subsidy (s) is provided, then returns to work rise to $(w - c(1 - s))$. This is the same as an increase in the effective wage and should increase work effort. There is clear research evidence that childcare assistance can increase labor supply among low-income mothers (Anderson and Levine 1999).

The availability of health care is also an issue for working low-income families in the USA. Families that receive public assistance have historically been eligible for Medicaid, the publicly provided health-insurance plan. Leaving public assistance has usually meant losing Medicaid. Since few low-wage jobs offer health insurance, this often means going without insurance, and may lead a number of families to stay on welfare in order to continue receiving Medicaid. The evidence on the magnitude of this effect is mixed (Winkler 1991, Moffitt and Wolfe 1992), but recent welfare reforms have mandated that states offer ongoing Medicaid coverage to persons leaving welfare for at least some period of time. This is one area, however, where the USA lags behind many European nations. Their nationalized health-insurance systems make health-insurance coverage a nonissue in the choice between work and welfare.

Other types of programs may also be important in supporting income and work under particular circumstances. The regular collection of child support from noncustodial parents is one way to guarantee additional income among single mothers whose earnings may be low. Subsidized public transportation systems, with routes that run through low-income neighborhoods, may lessen travel costs and travel time for potential workers. Strong enforcement of anti-discrimination laws may also improve the availability of jobs for certain populations. In short, a wide variety of policies can be important in assisting low-income families into employment.

4. *Conclusions*

Given the multiple goals associated with welfare programs, it is not surprising that there has been constant controversy about the appropriate design of such programs. It is clear that there are a number of alternative policy choices which may be attractive for different groups among the poor population.

For groups that not expected to work and for whom income support is the primary consideration, such as the infirm and elderly, traditional welfare programs may be highly effective and desirable. Indeed, the Supplemental Security Income program in the USA, designed to provide income support to elderly and disabled individuals, is designed as a traditional welfare program, and continues to receive widespread public support.

For groups that expected to work, a wide variety of programs can encourage work. Many of the changes in US welfare programs over the 1990s were designed to provide better work incentives and a stronger set of supports to working low-income families (as opposed to cash assistance to nonworkers). The expanded EITC wage subsidy appears to have been particularly effective in increasing labor-force participation among low-wage single mothers with young children. And the shift toward mandated participation by welfare recipients in job-search and job-preparation programs has added to these labor-force participation increases.

In the USA, as in many other nations, welfare programs are multilayered. Both the state and federal governments often have some program responsibility. These programs include a mix of cash and in-kind

support programs. They include a mix of more and less work-oriented programs aimed at different groups in the population. Some are run at the local level, while others are run through the national tax system. The effect is a quite confusing array of programs, which can produce overlapping incentive effects on individuals who are eligible for multiple programs. But such a mix of programs is also a response to the multiple goals and the desire to differentiate between different groups among the poor. And many of these programs have their own community of political advocates who support them and lobby for maintaining and/or expanding them. This may produce a network of programs that in sum provide much greater overall support to lower-income families than would be achieved by a more streamlined and unified welfare system.

See also: Equality and Inequality: Legal Aspects; Income Distribution; Inequality; Inequality: Comparative Aspects; Poverty, Culture of; Social Welfare Policies and Gender; Wealth Distribution; Welfare; Welfare State; Welfare State, History of; Work, History of; Work, Sociology of

Bibliography

Anderson P M, Levine P B 1999 Child Care and Mother's Employment Decisions. In: Card D, Blank R M (eds.) *Finding Jobs: Work and Welfare Reform*. Russell Sage Foundation, New York

Blank R M 1997a *It Takes A Nation: A New Agenda for Fighting Poverty*. Princeton University Press, Princeton, NJ

Blank R M 1997b Policy watch: The 1996 welfare reform. *Journal of Economic Perspectives* **11**: 169–77

Blank R M, Ruggles P 1996 When do women use aid to families with dependent children and food stamps? *Journal of Human Resources* **31**: 57–89

Blinder A S, Rosen H S 1985 Notches. *American Economic Review* **75**: 736–47

Cain G G, Wissocker D A 1990 A reanalysis of marital stability in the Seattle-Denver income maintenance experiment. *American Journal of Sociology* **95**: 1235–69

Eissa N, Hoynes H W 1998 The earned income tax credit and the labor supply of married couples. Unpublished manuscript, University of California, CA

Eissa N, Liebman J B 1996 Labor supply response to the earned income tax credit. *Quarterly Journal of Economics* **111**: 605–37

Fraker T M, Martini A P, Ohls J C 1995 The effects of food stamp cashout on food expenditures: An assessment of the findings from four demonstrations. *Journal of Human Resources* **30**: 633–49

Gais T L, Nathan R P 1998 *Overview Report: Implementation of the Personal Responsibility Act of 1996*. Nelson A. Rockefeller Institute of Government, Albany, NY

Gallagher L J, Gallagher M, Perese K, Schreiber S, Watson K 1998 *One Year After Federal Welfare Reform: A Description of State Temporary Assistance for Needy Families (TANF). Decisions as of October 1997. Occasional Paper no. 6, Assessing the New Federalism Project*. Urban Institute, Washington, DC

Gueron J M, Pauly E 1991 *From Welfare to Work*. Russell Sage, New York

Meyer B, Rosenbaum D 1998 Welfare, the earned income tax credit, and the labor supply of single moms. Unpublished manuscript, Northwestern University

Moffitt R A 1983 An economic model of welfare stigma. *American Economic Review* **73**: 1023–35

Moffitt R A 1992 Incentive effects of the United-States-welfare system: A review. *Journal of Economic Literature* **30**: 1–61

Moffitt R A, Wolfe B 1992 The effect of the Medicaid program on welfare participation and labor supply. *Review of Economics and Statistics* **74**: 615–26

Munnell A 1987 *Lessons from the Income Maintenance Experiments*. Federal Research Bank of Boston, Boston

Newman S J, Schnare A B 1992 *Beyond Bricks and Mortar: Re-examining the Purpose and Effects of Housing Assistance*. Urban Institute, Washington, DC

Ohls J C, Beebout H 1993 *The Food Stamp Program: Design Trade-offs, Policy, and Impacts*. Urban Institute, Washington, DC

Robins P K 1985 A comparison of the labor supply findings from the four negative income tax experiments. *Journal of Human Resources* **20**: 567–82

Robins P K, Fronstin P 1996 Welfare benefits and birth decisions of never-married women. *Population Research and Policy Review* **15**: 21–43

Winkler A E 1991 The incentive effects of Medicaid on women's labor supply. *Journal of Human Resources* **26**: 308–37

R. M. Blank

Welfare State

A welfare state is a state that is committed to providing basic economic security for its citizens by protecting them from market risks associated with old age, unemployment, and sickness. Protection is provided by a range of social policies, including health care, unemployment benefits, and old-age pensions. Changing norms about the need for protection and the appearance of new risks mean that the policies included under the welfare state rubric have shifted over time. While health care, unemployment, accident, and sickness benefits have formed the core programs, other policies such as social services, child-care policies, and family allowances are included in the study of the welfare state. Although early analysts of welfare states emphasized the importance of education, subsequent work has largely excluded education.

The term 'welfare state' first emerged in the UK during World War II as part of an effort to boost wartime morale and to contrast the British state with the Nazi warfare state (see Flora and Heidenheimer 1987). The welfare state symbolized the solidarity of the British people, which would be sustained after the war by a new set of social policies available to all British citizens. The Beveridge Report of 1944 supplied the blueprint for the British welfare state, with its vision of a state that would guarantee freedom from want to all its citizens. British sociologist T.H. Marshall ([1949] 1964) provided an influential early

analysis that linked welfare state development to the emergence of full democracy. He traced history as the progressive development of rights: civil rights in the eighteenth century, political rights in the nineteenth century, and social rights in the twentieth century. For Marshall (1964), the social rights embodied in the welfare state were the means by which the equal status of citizenship could sharply limit, although not override, the inequalities created by the market.

Subsequent efforts to identify the origins of the welfare state took issue with this formulation. The seeds of contemporary welfare states were traced back to Bismark's compulsory social insurance laws of the 1880s. The motivation for creating welfare state programs in Germany had little to do with democracy and much to do with social control. The rhetoric of citizenship, solidarity, and equality was notably missing in nineteenth century Germany, casting doubt on Marshall's (1964) neat progression of rights outside of the UK. The diverse motivations that characterized the origins of welfare states were reflected in the variety of mechanisms that financed and delivered social benefits. The Bismarkian model of social insurance provided benefits to workers on the basis of dedicated taxes—called contributions—paid into state-mandated schemes. Benefits were linked to contributions and thus tended to be status-preserving rather than equalizing. The Beveridge model relied on general taxation and provided flat-rate benefits to all citizens. By setting a high minimum level of benefits, the aim was to reduce inequality.

One enduring stream of welfare state research has sought to develop typologies to classify cross-national variation in the scope and design of welfare state programs. Richard Titmuss's (1958) classification of welfare states into residual, institutional, and industrial-achievement models was an early influential categorization. Building on Titmuss (1958), sociologist Gøsta Esping-Andersen (1990) developed the welfare state typology most widely used in contemporary welfare state research. The typology identifies three welfare state regime types—liberal, conservative-corporatist, and social democratic—that differ by the role they assign to the state, market, and family for ensuring well-being. Each model is further distinguished by the quality of social rights it creates and by the system of social stratification it produces.

The *liberal* welfare state presumes minimal interference of the state with the market. Welfare benefits are designed to limit their impact on the market. Accordingly, benefits for working-age adults tend to be minimal and are often subjected to a means test in which recipients must show they have no other source of support. Social rights are weak and states do little to alter market-created stratification. Private social benefits provided by firms are common and serve to deepen stratification. Countries that exemplify this model include the USA, Canada, and Australia. The *conservative* model provides generous social rights but seeks to preserve market-based status differentials. Benefits are commonly provided through social insurance in which the size of benefits is linked to previous contributions. Reflecting the influential role that the Catholic Church played in these countries, conservative welfare states rely heavily on the traditional family to promote well-being. This aim is reflected in the minimal provision of social services. Examples of conservative welfare state regimes include Germany, Austria, France, and Italy. The *social democratic* model, found in Scandinavia, is based on principles of universalism in which generous benefits are provided to all citizens. These welfare states seek to promote equality by ensuring strong social rights. Unlike the conservative welfare states, social democratic welfare states provide extensive social services designed to relieve the family—and women in particular—of its traditional responsibilities. Likewise, social democratic welfare states intervene extensively in markets in order to promote social protection.

In reality, most countries incorporate features of more than one model. For example, the USA provides social insurance for the elderly with a system that has egalitarian features and is based on contributions. Moreover, the UK, the original welfare state, fits poorly within the typology. Some analysts have proposed a 'fourth world of welfare' characteristic of Australia and New Zealand. Although these states provided minimal formal social benefits, they have sought to ensure social rights and equality through a system of wage-setting labor courts and high tariffs. Students of Southern Europe have emphasized the distinctiveness of welfare provision in such countries as Italy, Spain, and Greece. These welfare states have been called 'familial,' reflecting the prominent role assigned to families in providing social welfare. Despite the many qualifications and amendments to Esping-Andersen's (1990) typology, it continues to provide an important reference point in the welfare state literature.

Considerable research has sought to account for the development of welfare states and for the variation among them. The earliest work links welfare state development to the level of economic growth. Based on a mutivariate quantitative analysis of 64 countries, Wilensky (1975) showed that the initial emergence of welfare state provision was a function of economic growth. Among the richest countries, however, he demonstrates that the correlation between wealth and welfare state provision breaks down and instead a variety of political factors account for differences. A second line of research focuses on these political factors in the richest countries. Most influential has been the 'power resources' argument, which views working-class power as the key determinant of welfare state development. Because social policies empower workers vis à vis employers, power resources analysis sees the welfare state as part of the 'democratic class struggle' (Korpi 1983). The power resources school

relies on cross-national quantitative data and regression analysis. Historically oriented researchers have criticized the power resources school for overlooking the importance of other social forces, among them business and the middle classes. A third major line of analysis, the historical institutional approach, relies on qualitative comparative case analysis to account for differences in welfare state development (Immergut 1992). This work highlights the role of political institutions in shaping welfare state origins and development (Weir et al. 1988). In the 1990s, these three lines of explanation were joined by a fourth school of feminist analysis (Orloff 1993). Feminist studies of welfare states highlight the role of ideology about gender roles and families as a central factor in determining the type of welfare state. Feminist analyses have broadened the scope of inquiry to social services relevant to the family and policies related to women's rights, often neglected in welfare state studies.

Since the 1980s, welfare states have faced new pressures due to economic shifts and changes in families and gender roles. Persistently high rates of unemployment in Europe, low wages in the USA, growing numbers of female-headed households, and increasing female labor force participation have altered the social and economic assumptions on which welfare states were built. Moreover, as social programs, especially pensions, have matured, they have become very costly. Demographic changes, notably aging populations and smaller workforces, exacerbate these pressures. These conditions have prompted new questions about welfare states. One line of inquiry examines whether welfare states face serious retrenchment as a consequence of these shifts and whether the politics of welfare state retrenchment differ significantly from the politics of creating welfare states (Pierson 1994). Analysts agree that the broad popularity of most welfare state programs makes them difficult to dismantle. There is much less agreement about whether the modifications that have been made in welfare states are significant and whether new social and economic conditions have altered the ability of welfare states to promote equality and social solidarity.

A second new line of inquiry asks whether the increased pressures on welfare states—economic globalization, in particular—are causing them to converge or whether long-standing national differences in type and amount of social provision will endure (Scharpf and Schmidt 2000). This work examines the institutional foundations of welfare states with special attention to the relationship between welfare states and economic arrangements. One key question animating this work is whether social policies are a hindrance to economic growth or whether social policy can contribute to economic competitiveness in an open economy. Questions about welfare state convergence also appear in an emerging literature about the impact of the European Union on national systems of social provision.

In addition to the research about how and why welfare states are changing, a third line of argument makes policy and normative arguments about how welfare states should be restructured. In the 1990s, the Organisation for Economic Cooperation and Development issued a series of influential studies arguing for more 'active' welfare states, which emphasize policies that help beneficiaries work instead of providing passive compensation. These arguments reflect new assessments of risk, security, and welfare states. The earliest generation of welfare state theorists argued for universal programs available to all citizens (Titmuss 1958). Such programs, they believed, would build social solidarity and would be politically strong due to their broad base of beneficiaries. Since the 1980s, there has been growing disagreement about how welfare states should be organized and about what values they should promote. Some analysts defend the classic political logic of the welfare state and argue for an updated universalism that addresses the needs of working families and emphasizes education (Skocpol 2000). Others emphasize the benefits of risk and the need to target social benefits more efficiently. Likewise, arguments about the proper role of the state vs. civil society and about the proper balance of rights vs. responsibility have animated academic as well as policy debates.

Most welfare state research has focused on the developed world because it is only these countries that have recognized and enacted a range of social rights broadly applicable to all citizens. As the developing world has changed, with some regions such as East Asia growing richer and others, such as Africa, growing poorer, there has been increasing interest in studying systems of social protection in these regions. The focus on social protection broadens the field of inquiry beyond traditional welfare state studies to examine how a wide range of market, government, and social arrangements affect well-being.

See also: Age Policy; Economic Globalization and Gender; Education and Income Distribution; Family Health; National Health Care and Insurance Systems; Retirement and Health; Retirement, Economics of; Social Security; Welfare: Philosophical Aspects; Welfare Programs, Economics of; Welfare State, Geography of; Welfare State, History of; Youth Culture, Sociology of

Bibliography

Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Princeton University Press, Princeton, NJ

Flora P, Heidenheimer A J (eds.) 1987 *The Development of the Welfare State in Europe and America*. Transaction Books, New Brunswick, N J

Immergut E M 1992 *Health Politics: Interests and Institutions in Western Europe*. Cambridge University Press, Cambridge, UK
Korpi W 1983 *The Democratic Class Struggle*. Routledge & Kegan Paul, London
Marshall T H 1964 Citizenship and social class. In: Marshall T H *Class, Citizenship, and Social Development*. Doubleday, Garden City, NY, pp. 65–122
Orloff A S 1993 Gender and the social rights of citizenship: The comparative analysis of gender relations and welfare states. *American Sociological Review* **58**(3): 303–28
Pierson P 1994 *Dismantling the Welfare State? Reagan, Thatcher and the Politics of Retrenchment*. Cambridge University Press, Cambridge, UK
Scharpf F, Schmidt V (eds.) 2000 *From Vulnerability to Competitiveness: Welfare and Work in the Open Economy*. Oxford University Press, New York
Skocpol T 2000 *The Missing Middle: Working Families and the Future of American Social Policy*. W. W. Norton, New York
Titmuss R M 1958 *Essays on the Welfare State*. Allen and Unwin, London
Weir M, Orloff A S, Skocpol T (eds.) 1988 *The Politics of Social Policy in the United States*. Princeton University Press, Princeton, NJ
Wilensky H L 1975 *The Welfare State and Equality: Structural and Ideological Roots of Public Expenditures*. University of California Press, Berkeley, CA

M. Weir

Welfare State, Geography of

The term 'welfare state' describes the state institutions and policies that are devoted to the delivery of human and social services, poverty amelioration, and enhancement of human well-being. In general, the welfare state is intended to ensure that the basic needs of the population are met with regard to food, housing, education, income, and health. The provision of social welfare serves two interrelated purposes. The first is to provide a social safety net through which citizens of a country should not drop. This role reflects a belief in the social rights of the citizenry and the responsibility of the society to each of its members. The second purpose of the welfare state is to provide a labor force with the skills, human capital, and basic well-being necessary to make it economically productive. The bundle of services provided through the welfare state, the means of provision, and the outcomes or effects of provision are not fixed. Rather, the welfare state is temporally and geographically variable. This variability reflects the interaction between changing ideas about human well-being, how best to attain it, and the specific conditions, practices, and resources that are available for meeting well-being in different places. This variability forces us to consider the geography of the welfare state and, specifically, two critical questions: (a) what are the processes by which unevenness in welfare provision is created? And (b) at what scale are welfare services provided?

These issues are considered under the following headings. Section 1 provides a basic characterization of the forms and functions of the welfare state. Section 2 considers reasons for this variation, including differences in the ability to provide social and human services, differences in the ways that states conceptualize their social role with regard to the well-being of the people who live within their territories, and the role of international organizations. The final section anticipates future changes in the welfare state in both developed and developing countries.

1. Forms and Functions of the Welfare State

The term 'the welfare state' is really a misnomer, as welfare states may take a variety of forms, and these forms have important implications for the citizens, communities, and economies of countries. Among the many typologies developed to characterize the forms of welfare states, those of Gøsta Esping-Andersen (1990) and Diane Sainsbury (1996) describe the framework of welfare provision within a country.

1.1 Typologies of the Welfare State

Esping-Anderson focuses on the ways in which ideas and beliefs regarding welfare and the state's role in meeting economic needs have shaped the form of welfare provision. He identifies three basic 'welfare state regimes':

Liberal welfare states respond to market and labor force imperatives. Many benefits, such as health insurance and pensions, are linked with employment. Means testing is used to determine eligibility for state services, and relatively modest cash and voucher benefits are provided for those deemed eligible.

Conservative/corporatist welfare states rely on state provision of services, rather than on market or private provision. These states often manifest normative ideals of a nuclear family characterized by a male breadwinner and a woman who tends to the family.

Social democratic welfare states promote a vision of the state as the guarantor of social rights. These states promote equality of benefits at high levels as a way of minimizing the effects of social class and income. Welfare benefits are used to equalize the ability of all citizens, regardless of income, to participate in the political community.

This study was based on 18 countries characterized by advanced market economies in the 1980s. As such, it reflects a tendency in the literature to conceptualize welfare states as being limited to industrialized democracies, rather than to consider the welfare apparatus in all countries. It is unclear, then, whether this typology will hold for countries at different stages

of development or undergoing transition to market economies. This typology also assumes countries develop their welfare apparatus in response to national political cultures, ideologies, and internal political pressures. The experience of countries beyond the 18 studied is clearly at odds with that assumption, as the development of their welfare policies and institutions is shaped by external influences from international aid and lending organizations and by a globalized economy dominated by transnational corporations. As such, the typology developed by Esping-Anderson may not fully describe the geography of the welfare state.

Sainsbury's study is significant in helping to understand the differential impacts of the welfare state on citizens and in understanding the ways in which gender ideologies shape the structure of social provision. As noted by Esping-Andersen, some welfare state regimes are based on normative ideals of family and gender roles. Sainsbury's framework clarifies the bases of entitlement to welfare provision under different regimes, drawing on her characterization of gender ideologies within countries. From this starting point, she traces the effects of welfare regimes on gender relations and the gendered division of labor in the family and society. As with Esping-Andersen, Sainsbury's research is based on developed, Western countries. Sonia Alvarez (1998) extends this approach in her analysis of gender policies in Latin American countries. Both studies demonstrate that the effects of gender equality reforms in welfare policy vary widely from country to country based on the ways that gender ideologies are embedded in their political economies, creating a differentiated geography of welfare provision and well-being.

1.2 Decisions about Welfare Provision

The welfare state can be conceptualized as a framework that guides three types of decisions with respect to welfare provision: decisions about services to be provided, institutional arrangements for provision, and the scale of provision and decision-making.

Decisions about the services to be provided and the conditions under which they are provided are the basis of Esping-Andersen's typology. The differences between these regime types center on three questions: (a) whether welfare is primarily a matter of private provision through employment or is a universal right; (b) what constitutes 'basic needs'; and (c) the extent to which welfare services are used to redistribute income and equalize the material bases of quality of life for all citizens. To these questions, geographers add a fourth: at what scales will decisions about welfare be made and services provided?

Liberal welfare regimes, such as that in the United States, place employment at the center of needs provision. Minimum wage provisions are intended to ensure workers' ability to pay for housing, food, clothing, and transportation. Pensions and health care are typically provided through a mix of employer and employee contributions. Eligibility for other services is often linked to employment, job training, or the search for employment. For example, the Personal Responsibility and Work Opportunity Reconciliation Act (PRWOR) is the 1996 legislation that restructured welfare provision in the United States. PRWOR links public assistance to participation in programs to make recipients more employable. Education is central to this model, and is notable for being the sole welfare service for which universal access is guaranteed; indeed, many liberal welfare regimes do not characterize education as a welfare service. While some degree of redistribution of wealth may occur through welfare provision in liberal regimes, the redistribution is minimal. Welfare services are only intended to meet basic needs and provide recipients the opportunity to enter the labor force. The slogan 'welfare as a hand up, not a hand out' is characteristic of this approach. By contrast, conservative and social democratic regimes rely more directly on state provision of services, and so are characterized by a large social sector for health, housing, pensions, and income support, regardless of employment status. As noted previously, what differentiates these two regime types is that conservative regimes such as the United Kingdom, Germany, and France place more emphasis on the nuclear family and on gender ideologies that reinforce women's labor force position in contingent and low wage jobs. In addition, the conservative emphasis in welfare provision is on guaranteeing basic needs, although a considerably broader spectrum of services is typically considered 'basic' as compared with liberal regimes. Social democratic regimes, such as those in Norway and Sweden, tend to have a larger redistributive goal than the other two regimes. As such, they strive toward universal provision at very high levels and across a much wider variety of services.

The institutional arrangements for welfare provision are implied in Esping-Andersen's typology, but these arrangements require additional attention. They are the subjects of growing political debate across most welfare regimes, and they reflect fundamental beliefs as to the appropriate relationships between the state, market, and civil society. In practice, institutions within all three of these spheres play a role in welfare provision, with the state playing an important regulatory or coordinating role. Liberal regimes conceive of the state as the provider of last resort for most welfare services, with the previously noted exception of services that prepare and maintain a productive labor force. Welfare, to a considerable degree, is the responsibility of individuals and firms, and becomes an issue for state intervention only when one or the other of these does not meet their responsibilities. In that situation, the state has regulatory powers that may be used to create rules and incentives for firms

and individuals to conform to expectations with regard to employment and welfare, and to impose penalties for noncompliance. The primary state role is one of developing workforce capacity, creating incentives for individuals to work, and maintaining a regulatory environment that guides decisions made by firms about benefits and work conditions for employees. In addition, states attempt to develop a voluntary sector through which many social services are provided; these efforts involve tax deductions for charitable contributions and the provision of funds to some nonprofit, nongovernmental organizations that are involved in direct service provision. These organizations are an important part of what Jennifer Wolch (1990) terms 'the shadow state.' While not institutionally part of the state apparatus, their existence and activities are regulated by the state and they perform vital services for the state.

Since the 1980s, many conservative and social democratic regimes have moved closer to the liberal form of welfare provision. Many developing countries, whose welfare apparatuses are not as well developed as those in industrialized countries, have also moved to this model due to structural adjustment and austerity programs imposed by international lending and aid organizations. At the same time, many new jobs in countries around the world do not provide the support for welfare provision that was assumed with liberal welfare regimes. These changes have shifted the boundaries between state, market, and civil society in ways that place the responsibility for maintaining welfare on civil society. However, these changes represent only a *shift* in the role of the state, not a *declining* role, as these changes are enabled and regulated through state action (Staeheli, et al. 1997).

The scales at which decisions about welfare provision are made constitute the final element that shapes the geography of the welfare state. Much of the discussion about the welfare state has focused on the national scale, even though service provision is often local. Late twentieth-century trends in welfare provision, though, highlight the limitations of such a focus. Many industrialized nations have begun to devolve much welfare decision-making to local and regional scales. Developing countries, as well, face pressures for devolution. For example, the United Nations Development Program argues in its Poverty Report that moving decision-making to local levels is an important element in creating governance structures that are more responsive to the needs of the populace, are more accountable, and will be more effective in poverty reduction (UNDP 2000). At the same time, increased globalization of the economy shifts economic decisions to scales above the nation-state. Finally, supranational and international institutions also shift the locus of decision-making. The European Union, for example, is attempting to harmonize social welfare provision and labor regulations, and international human rights conferences call for the development of standards of well-being and social provision. The combined effect is a reconfiguration of the scales at which decisions are made and welfare ensured. These trends highlight the importance of understanding the geography of the welfare state as multiscalar, in which decision-making with regard to various aspects of welfare provision is shifted between scales. While some aspects of decision-making and welfare provision may be devolved, other aspects may move outside the gambit of the nation-state to supranational and international scales.

2. *Sources of Variation in Welfare Provision*

The structure of the welfare state within any region or country is the result of the interplay of many pressures and contextual factors. Some of these pressures may be specific to a particular place or country; these pressures might include gender and family ideologies. Other pressures might come from higher scales, such as the development of an international human rights regime that promotes a different ideology of gender and family. The interaction between processes, conditions, and practices that operate at local, national, and global scales gives rise to geographic variation in the welfare state and human well-being. While the sources of variation are many, three are of particular importance: the resources upon which states can draw, differences in ideology and political culture that shape the ways states conceptualize their roles, and pressures from international organizations.

The resources available to countries are clearly related to their overall level of wealth; at the same time, the need for social spending is also related to wealth. As a result, rich countries may have relatively small expenditures for welfare as a proportion of GDP, while still meeting basic needs. Thus, the Republic of Korea has a relatively small public budget and a low level of expenditure for social needs. As a relatively wealthy country, however, the overall level of well-being is high. By contrast, Pakistan and Indonesia have relatively large public budgets as a proportion of GDP. The proportion of the public budget, however, that is directed to social needs is very low, with predictable implications for overall well-being (UNDP 2000).

The above statements make clear that well-being and the success of welfare projects are affected by the pre-existing levels of well-being, by expenditure levels, and by the priorities established by national governments. The interaction between state action and contextual conditions in different localities gives rise to the uneven geography of welfare and well-being. Here the importance of the typologies developed by Esping-Andersen, Sainsbury, and others becomes clear, as those typologies reflect priorities and beliefs as to social rights. Social democratic and conservative regimes posit welfare as a social right of citizenship;

liberal regimes do not to the same degree. Also at work are differences in the ways that regimes conceptualize the responsibilities of citizenship. Conservative regimes assign families a high level of responsibility for individual and community well-being once basic needs are met, whereas social democratic regimes do not. The balance between rights and responsibilities for welfare, then, reflects ideologies related to the idea of dependence and the roles of families and communities (Fraser and Gordon 1994). Some countries take well-being to be a social responsibility, and do not link citizenship to dependence. Others take well-being to be an individual responsibility and link citizenship with independence. The question facing developing countries and countries undergoing economic and politic transitions is which path they will follow. Here the role of international organizations becomes clear.

International organizations can influence, and in some cases determine, the path that countries will follow as they develop their welfare apparatus. The role of the World Bank and International Monetary Fund in structural adjustment policies is well known. These organizations have made emphasis on productive activities a condition of development assistance. This has led to a redirection of public expenditure away from basic needs provision to public expenditures that support job growth (Corbridge 1993). While education may be a priority in this regard, health care, housing, and basic human needs may not. There is a widespread belief that responsibility for social reproduction has been shifted to families, and most particularly, to women (Lustig 1995).

Other international organizations propose poverty reduction strategies that rely on local communities, rather than national government support for social safety nets. The UNDP (2000), for example, argues:

> The foundation of poverty reduction is self-organization at the community level. Such self-organization is the best antidote to powerlessness, a central source of poverty.... What the poor need, therefore, is not resources for safety nets but resources to build their own organizational capacity.

To the extent that such statements actually guide the development of welfare state apparatuses in poor countries, they would result in a high priority placed on local communities, rather than state or business support, as a basis for poverty reduction. This would represent a fourth model, not anticipated by Esping-Andersen. It is a model, however, that is based on a gender ideology in which women play critical, and often uncompensated, roles in social provision. This new model is also characterized by heightened variability in both institutional structures at the local level and welfare outcomes, as diverse structures interact with diverse local conditions. The result is a model of welfare provision with a complex geography, and in which well-being is clearly related to the jurisdictions within which people live and work.

3. *Anticipating the Future Welfare State*

There is a widespread belief that the commitment of states to social welfare provision is declining in countries around the world. The 1999/2000 World Development Report, however, does not show a consistent trend toward decreased public expenditure as a percentage of GDP, nor does it show a consistent trend toward retrenchment of public expenditures on social services. Indeed, of 47 countries reporting expenditures in both 1980 and 1997, 36 had either maintained or increased spending on social services (World Bank 1999, Table 14).

It is clear, however, that the pattern of expenditures is changing, even if the total level is not. The United Nations and World Bank have called for new approaches to social welfare and poverty amelioration. They urge a reorientation of welfare policy to integrate poverty amelioration with other goals of economic development, sustaining environments, better governance, and building community assets. The attempts to build social institutions and orient welfare policy to enhancing the capacity of communities to provide for their members are consistent with the language of welfare reform in developed countries. Increasingly, the social capital within communities is being looked to as a resource in welfare provision. Under these conditions, the welfare state may take on a larger coordinating role, providing institutional frameworks, with a smaller role in direct service provision.

The outcome of this emerging form of the welfare state is not clear. It seems reasonable to expect, however, that heightened variability and unevenness will characterize both the geography of institutional practices and the geography of outcomes. It may be that this unevenness will reflect the resources, capacities, and intentions of communities and localities in ways that make the welfare state more responsive and effective. However, as Cope (1997) suggests, it may also create new forms of uneven development, with rich areas getting richer and poor areas getting poorer.

See also: Poverty Policy; Social Security; Social Stratification; Social Welfare Policies and Gender; Social Welfare Policy: Comparisons; Welfare; Welfare: Philosophical Aspects; Welfare Programs, Economics of; Welfare State; Welfare State, History of

Bibliography

Alvarez S 1998 Feminismos latinamericanos. *Revista Estudos Feministas* **6**: 265–84

Cope M 1997 Responsibility, regulation, and retrenchment: The end of welfare? In: Staeheli L A, Kodras J E, Flint C (eds.) *State Devolution in America: Implications for a Diverse Society*. Sage Publications, Thousand Oaks, CA, pp. 181–205

Corbridge S 1993 *Debt and Development*. Blackwell Publishers, Oxford

Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Princeton University Press, Princeton, NJ
Fraser N, Gordon L 1994 A genealogy of 'dependency': Tracing a keyword of the U.S. Welfare State. *Signs* **19**: 309–36
Kodras J 1997 Restructuring the state: Devolution, privatization and the geographic redistribution of power and capacity in governance. In: Staeheli L, Kodras J, Flint C (eds.) *State Devolution in America: Implications for a Diverse Society*. Sage Publications, Thousand Oaks, CA
Lustig N (ed.) 1995 *Coping with Austerity: Poverty and Inequality in Latin America*. The Brookings Institution, Washington, DC
Pinch S 1997 *Worlds of Welfare: Understanding the Changing Geographies of Social Welfare Provision*. Routledge, London
Sainsbury D 1996 *Gender, Equality and Welfare States*. Cambridge University Press, Cambridge, UK
Staeheli L, Kodras J, Flint C 1997 *State Devolution in America: Implications for a Diverse Society*. Sage Publications, Thousand Oaks, CA
United Nations Development Program 2000 *Overcoming Human Poverty*. http://www.undp.org/dpa/publications/index.html
Wolch J R 1990 *The Shadow State: Government and Voluntary Sector in Transition*. The Foundation Center, New York
World Bank 1999 *World Development Report, 1999/2000 Entering the 21st Century*. World Bank, Washington, DC

L. Staeheli

Welfare State, History of

1. The Origin and Use of the Term Welfare State

The term was coined in England in the 1940s. It is to be found in *Citizen and Churchman* (1941) by William Temple, Archbishop of York, where it is contrasted with the Power State or Warfare State. After the publication of the Beveridge Report in 1942, the term soon came to be applied to the hopes for postwar reconstruction and by 1948 was used regularly for the postwar reforms. The meaning of 'welfare' had not suffered the deterioration that had afflicted the word 'charity.' Although Beveridge himself disliked the term for what he called its 'Santa Claus' connotations and preferred the older term 'social service state,' 'welfare state' normally carried a favorable meaning.

Welfare had a different set of connotations in American English. 'Public Welfare,' as understood in the USA then and since, excluded the elements of social insurance, with its claims to benefit as of right, which were so prominent in the British construction of the term (Lindeman 1937). 'Welfare state,' once imported from Britain, thus carried a stronger association of paternalism (Skocpol and Ikenberry 1983).

In that respect Germany was closer to the USA. In the early years of the Federal Republic, when many English terms were naturalized into German, it seemed natural to render 'welfare state' with the similar word *Wohlfahrtsstaat*. Such similarities can be misleading. *Wohlfahrt* was associated with charitable gifts, and *Wohlfahrtsstaat* had acquired a pejorative connotation during the crisis of the Weimar Republic, as when in 1932 Chancellor von Papen accused politicians of the early years of the Weimar Republic of attempting to turn the state into a *Wohlfahrtsstaat*, and thereby weakening the moral strength of the nation. The Federal Republic wished to return to the Weimar tradition in social policy without falling foul of the associations of the term *Wohlfahrtsstaat*. One possible course was to use the term in its positive English sense and to resort to *Versorgungsstaat* to convey the pejorative meaning. The alternative was to revive a nineteenth-century term, *Sozialstaat* (Ritter 1989). The issue has never been settled decisively, although *Sozialstaat* is the more usual.

If the term dates from the 1940s, its definition also owes much to the changes of that period. According to Asa Briggs, 'A welfare state is a state in which organized power is deliberately used (through politics and administration) in an effort to modify the play of market forces in at least three directions—first by guaranteeing individuals and families a minimum of income irrespective of the market value of their work or their property; second, by narrowing the extent of insecurity by enabling individuals and families to meet certain "contingencies" …; and third, by ensuring that all citizens without distinction of status or class are offered … a certain agreed range of social services' (Briggs 1961). These are usually taken to relate at least to medical treatment, housing provision, and education. Briggs added a reference to the best standards available, but it has since been generally accepted that such an aspiration is unrealistic.

2. The 'Origins of the Welfare State'

Such an emphasis on the 1940s leaves the historian with the task of relating these innovations to what had gone before, i.e., to consider the 'origins of the welfare state.' Three national examples briefly indicate how differently the origins of the welfare state have been constructed.

2.1 British Historiography

Since 1961 this has operated in terms of the transformation of old institutions by new principles, forged at moments of political crisis (Fraser 1973). The first of these institutions was the Poor Law, which established the responsibility of each local parish for the relief of the poor (1603) and was reformed in 1834 on more explicitly deterrent lines in response to population growth and agricultural change. The subsequent history has been presented in terms of new social problems caused by industrialization, the Victorian

origins of collectivism in public health and factory regulation, the crisis of the Poor Law at the end of the nineteenth century, and the consequent Edwardian legislation (1905–14). This put an end to the possibility of a gradual extension of Poor Law services on a nondeterrent basis and provided for needy and deserving groups in other ways. It introduced free school-meals, a school medical service, tax-financed old age pensions for the poor, and national health and unemployment insurance, supported by a national network of labor exchanges. Alongside these measures 'New Liberalism' articulated a theoretical justification for state intervention in society and economy. The scope of this legislation was limited in respect of population covered and of benefits provided, but it has been regarded as having laid 'the foundations of the British welfare state.'

The interwar years saw housing provision added to the prewar measures, but social policy responded slowly and piecemeal to economic and social problems. It adapted rapidly once World War II had created new problems and a new spirit of citizenship. The postwar legislation, roughly but far from exactly based on the proposals of the Beveridge Report (1942), universalized national insurance, extended its benefits to cover a greater range of contingencies ('from the cradle to the grave'), and established a mainly tax-financed national health service. This was accompanied by tax-financed child allowances, the extension of housing provision, and free secondary education. Fundamental was a government commitment to a policy of full employment, based to some extent on a belief in the efficacy of Keynesian demand management that was, however, far from unequivocal. The transformation of old social services by new principles of universalism had been presented in the Beveridge report as a characteristic British revolution. Historians followed a similar line, which fitted into the then dominant 'Whiggish' historiography with its emphasis on the continuity of British history and the adaptability of its political leaders. In the influential view of the sociologist T. H. Marshall the welfare state had achieved the extension of citizen rights from civil rights via political rights in the past to a new body of social rights (Marshall 1950). The blueprint of the welfare state as established in the 1940s was for a long time presented as a major British contribution to postwar social reconstruction in the wider world.

2.2 German Historiography

This has located the origins not merely of its own welfare state but of the welfare state as such in the establishment of compulsory insurance for workers by the German Empire under Bismarck in the 1880s. It has both emphasized the pioneering role of Germany and the continuity of its insurance institutions despite drastic political change. This emphasis on the role of Bismarck, notorious for his hostility to the claims of the working-class movement, as the originator of the welfare state has presented historians with the problem of reconciling these two aspects of his policy. Until recently it was suggested that wooing the workers with the benefits of social policy was an intrinsic aspect of Bismarck's attempt to suppress social democracy. This view had the merit of presenting the welfare state as the achievement of the democratic forces while recognizing that the policies usually were not what these had been demanding. The publication of documents on Bismarck's social policy has cast doubt on this view of his aims and there is a need for re-interpretation (Tennstedt and Winter 1993, 1995).

The transition from the establishment of institutions with limited scope to the conscious acceptance of the people's welfare as an essential object of the state is located by German historiography in the Weimar Republic. It found expression in the text of the constitution but also across a range of its institutions, several of which had developed from war-time innovations. Collective bargaining backed by compulsory arbitration, subsidized housing, the preservation of prewar social insurance for both blue and white collar workers, and unemployment insurance as part of labor market policy, all suggest that the Weimar Republic was a welfare state in deed as in word. In German history the breakthrough to the welfare state was therefore achieved not after World War II but after World War I. However, the Weimar welfare state was unable to live up to its expectations (Abelshauser 1987). The political acceptance of its new role was insufficiently broadly based. As the earlier quotation from von Papen suggests, repudiation of the Weimar Republic in the years of economic crisis was connected closely with its social policy.

Although marginalized and only just saved from drastic changes in 1942, the social insurance system survived the Nazis. In the immediate postwar years of economic ruin, forced migration, and impoverishment it faced perhaps its greatest test. All accumulated funds had been lost. In the Russian zone of occupation it was reconstructed radically along lines originally proposed by the left-wing critics of Weimar policy. It was no more sacrosanct in the West. The French quickly followed the Russian example. Drastic British proposals followed, based partly on Social Democratic ideas from the Weimar years, partly on Beveridgean lines. They involved retrenchment and thereby aroused opposition well beyond the threatened vested interests. Opposition rallied around the defense of the Bismarckian inheritance in social insurance, which was presented as one of Germany's proud contributions at a time when the nation badly needed a boost to national self-esteem. With the onset of the Cold War the proposals were dropped as the Western occupation powers gave priority to working with German public opinion. It was left to the Federal Republic to deal with the problems of funding social

security; under CDU leadership the traditional structures were almost completely reinstated. Significant reform was delayed until 1957. In that year the introduction of the 'dynamic' pension, linked to average wage levels, represented an imaginative response to the problems of the Federal Republic: not just the problem of inflation but that of the low pension levels which contrasted sharply with the rising standard of living enjoyed by the working population. At a time when the legitimacy of the Federal Republic was fragile, this pension reform contributed significantly to the acceptance of the regime. In 1958–61 proposals to remodel the sickness insurance system and contain its spiralling costs were defeated. The Bismarckian prototype had acquired a secure political profile.

Thus, in the German Federal Republic the 1940s were a period of defence and restoration. The emphasis on the significance of the Bismarckian reforms for the long-term development of the German welfare state and the tendency to identify it narrowly with contributory social insurance were forged in the political conflicts of those years. This emphasis was reinforced in the early 1980s by the centenary celebrations of the Bismarckian reforms. Much of the German historiography of the welfare state was created in the 1980s (Ritter 1983, 1986, 1989).

These examples indicate that the history of the welfare state has been perceived differently in different countries, in terms of periodization and of emphasis. This reflects not only its actual development in different states but certainly also the political purposes which historiography has served.

2.3 The Swedish Model

These points must be made more briefly for the influential Swedish model of the welfare state. There, it was developments in the 1930s that had the significant long-term effects that they singularly failed to have in Britain or Germany. The expansion of the earlier limited social security provisions took a significantly different form. Unlike Germany, where insurance contributions and benefits differed according to levels of income, Swedish benefits were paid, as in Britain, at a flat rate for all. But they differed from both the German and the British provisions in that retirement pensions and sickness insurance were financed from general taxation. They were set at a level that provided pensioners with an adequate income, something that the British pensions, tied to contributions that even the lowest-paid could afford, entirely failed to do. Significant differences between the policies of these three welfare states also extended to industrial relations and taxation.

The Swedish welfare state has been presented as the achievement of class mobilization and labor solidarity. Standard accounts have stressed the long period of Socialist-dominated government and cooperation between government and labor organizations in the shaping of labor market policies (Baldwin 1990).

3. A General History of 'The Welfare State'

If it exists at all, the history of the welfare state, as distinct from the history of individual welfare states, must be sought on a higher level of generalization. A welfare state is a state of a certain kind, distinguished from earlier kinds by the addition of a new set of objectives and institutions. Its history must be sought within the wider context of the history of the state. It presupposes the unification of law across the state territory, the establishment of bureaucracies, and the elimination or control of intermediate forms of authority. It is, however, also characterized by new systems of domination, consisting of distributing elites, service bureaucracies, and social clienteles. The process by which these objectives were both established and extended to include the whole nation has been driven by the advent of representative democracy, but it also drew on older concerns to maximize the human resources of the state. In addition, these new state objectives have been connected intimately with the development of modern industrial capitalism.

The state became a welfare state because it dealt increasingly with the social consequences of the particular way in which modern industrial capitalism was established. These consequences, often described as externalities, resulted from the narrow definition of the legal obligations of capitalist entrepreneurs, which contrasted greatly with the obligations imposed on entrepreneurs in the older corporatist economy. This emancipation of the entrepreneur was a deliberate act of state policy, undertaken in the interest of increasing 'the wealth of nations' and thus the power of states over against other states. It undoubtedly had that effect. Enterprise took new and unforeseen forms once it was freed from old regulations. To limit obligations towards workers to the short-term purchase of their labor power, towards the community to the payment of local taxes, and to resort to competition without responsibility for those driven out of the market, all this encouraged innovation and facilitated capital accumulation. But it created problems that resulted from economic decisions while being external to the economic process.

3.1 Combinations and/or Transfer Payments

These problems could be dealt with in at least two ways. (a) By combination, intended to limit free competition over wages, or to pool the economic resources of individuals, or both (trade unions and mutual aid societies). To what extent that was possible depended on the attitude of the state and on the

capacity of individuals for organization. (b) By state action in whatever territorial units—national, provincial, or local—it exercised its power. That meant taxation and a limited redistribution of the resources originally distributed by the market. The administrative capacity that made it possible to introduce graduated taxes and differential tax allowances as instruments of redistribution is an important aspect of this history.

The legalization of combinations to increase wages in turn made it increasingly possible for individuals to satisfy their needs through the market, thereby creating opportunities for new kinds of enterprise. The legalization of mutual aid associations created structures of redistribution, which could operate as alternatives to the bureaucratic structures of the state. In countries in which the development of bureaucracy had not advanced far, or where it was viewed with mistrust, subsidies from taxes became a way of making mutual associations more effective. Such subsidies were not unconditional but subjected the association to a greater or lesser degree of control. An extreme example of such control was the compulsory establishment of combinations on the order of the state, such as the German sickness funds and trade associations to administer accident insurance. The higher the level in the administrative hierarchy on which the state operated to redistribute resources, the greater the degree of uniformity it could impose.

3.2 Goods and Services

So far the focus has been mainly on the redistributive role of the welfare state. The problem of externalities has also been addressed by the public provision of goods and services, which the market supplied only in ways deemed politically unacceptable. Such goods and services, considered necessary for the common good as politically defined, have included the infrastructure of public health, the provision of healthy and environmentally approved housing, of medical treatment, and of schools, but this list is far from exhaustive. Here, too, subsidies could take the place of full state provision, e.g., in respect of houses, food, or schools.

3.3 Regulation

Since externalities had been allowed to arise because the state had regulated the conditions of production in minimal ways, it also began to introduce new regulations. Under such regulations the market was permitted to supply goods and services only on conditions acceptable to the state. Instead of providing housing, the state regulated the conditions of tenancies. Instead of running its own schools, it regulated and inspected those provided by other bodies. And, as in Thatcherite Britain, it regulated the sanitary infrastructure while relying on the market to provide it. Such regulation could be accompanied by subsidies. However, such regulation suffers from the same considerations that applied to the general conditions of production in the first place. By regulating the actions of the market, the state is liable to hinder innovation and to make the market an inefficient mechanism for economic growth.

Finally, the state has been able to regulate the economy itself in such a way as to minimize externalities. Examples are the state regulation of wages and the regulation of the economy so as to reduce cyclical fluctuations. At all stages of this process, the question has to be faced whether the object is being achieved at the cost of economic innovation and efficiency. For that reason there are limits to what regulation can achieve.

A radical alternative has been for the state to take over the running of the total economy. That rules out a welfare state, defined as a state that modifies the play of market forces, and it creates a state so different that it is better not to describe it as a welfare state. That is not to deny, however, that 'communist' states provide social services and organize transfer payments in the interest of social security. This article does not deal with the history of states with a command economy, confining itself to the history of what is sometimes called the liberal welfare state.

Even there the importance of the welfare elements of the state can vary according to the extent to which entrepreneurs did in practice externalize their costs. Where company welfare was built into the entrepreneurial ethic, as in Japan, state expenditure on social security has been restricted compared with other advanced industrial countries.

4. The History of Welfare States Since 1945

Despite this attempt and others (Flora and Heidenheimer 1982) to construct explanatory models of the welfare state as a general phenomenon of European and subsequently extra-European history, the established tradition of nation–state history has led historians to study developments both before and after 1945 in terms of individual states. They have given prominence to the political circumstances under which social policy was constructed and emphasized the importance of national traditions.

Some few generalizations across countries are feasible. Thus, the experience of the 1930s had demonstrated that institutions of social security were seriously vulnerable to periods of mass unemployment. Such periods greatly increased the demands for transfer payments, while also reducing the income from which these were made. This was particularly true of contributory insurance, but it also applied to finance from general taxation. This lesson had been learned, and in the 1940s macro-economic policy

became a policy component of every state, whether it had passed through the experience of a war economy or, like Sweden, had not. In some countries this period saw major reforms and extensions in social security, as in Britain, France, and Sweden. In others, major reforms had to wait until the 1960s and early 1970s after the restructuring of the economy had yielded its rewards, as in The Netherlands and Italy.

Even where significant legislative innovation occurred in the immediate postwar years, subsequent expansion of the economy provided the setting for a period of further change in the 1960s and 1970s. Indeed, the most convincing historical generalizations that can be made about the welfare states since 1945 refer to the effect of major changes in the international economy. A long period of economic recovery in Europe led by the 1960s to a remarkable growth in national income. This in turn both raised expectations and facilitated the introduction of more generous provisions. Levels of benefits were raised, flat-rate pensions complemented by graduated ones that mirrored differences in original earnings and appealed to the better off, while those disadvantaged by or excluded from earnings-based insurance systems received compensatory entitlements.

By contrast, the setback to economic expansion in the mid-1970s, triggered by the oil-price crisis, caused a fall in employment, except where special measures for job creation had been implemented (as in Sweden). This, in turn, led to rises in social security payments and falls in the revenue available for the purpose. Steady increases in the gross national product, which had spurred ambitious changes in entitlements in the previous decade and a half, could no longer be assumed. Nor had recent rapid growth in social expenditure been due only to deliberate changes in entitlement. It had also been caused by such demographic, cultural, and technological changes as an increase in the proportion of the elderly, a rise in single-parent families, and the growing cost of medical treatment. To the fiscal crisis that occurred in every country sooner or later, even in Sweden, was added a crisis of belief in the assumptions on which welfare states had been based. Yet these assumptions had been deeply implanted in the citizens, and many vested interests created. Measured by the radical debate that broke out, the actual changes in welfare states were not great, even in Great Britain, where the call for the rolling back of the state had been loudest.

5. *Historiography of Welfare States*

There are several ways in which the original traditions of welfare state history have been modified.

5.1 *A Gender-conscious Historiography*

Just as the feminist movement has since the 1960s exerted a political influence on the agenda of policy making itself, so it has also changed historical perceptions. Feminist historians have highlighted the way in which social policies were shaped by assumptions about the gender divisions of labor, power, and social responsibility. They have shown how little the basic assumption that each adult woman was dependent on a male earner was borne out in reality. Contributory insurance, with entitlements dependent on the performance of waged labor over most of the adult life span, had been at the center of welfare state history. It looked very different through the eyes of feminist historians, but these have not merely confined themselves to criticism. By drawing attention to policies in support of motherhood, they have reconstructed the outlines of welfare state history (Pedersen 1993, Koven and Michel 1993). By making historians more aware of the implications of gender, this emphasis has yet unrealized potentials. Male roles, once so taken for granted as to be largely invisible to historians, have in turn become objects of historical study.

5.2 *Welfare State History 'from Below'*

The history of social policy traditionally has been written through the eyes of 'policy-makers' and administrators, and occasionally of political philosophers, trade union officials, and radical politicians, authority figures all. Welfare state history has been particularly resistant to the search for history as experienced by those without power and influence. Yet the construction of social policy and its administration is that of a relationship, in which the recipients of 'benefits' are at least half the story. Like all forms of history from below, attempts to recreate their experience has been hindered by their inarticulacy, at least in forms that have survived as records. But once historians search, they are able to find evidence, not merely for recent times but at least in England going back well into the past (Lees 1998, Vincent 1991).

5.3 *Studying the 'Mixed Economy of Welfare'*

One result of paying more attention to the lives of the poor has been to downgrade the importance of state provisions and to draw attention to the mixture of resources that made up the economy of the poor. Since public authorities frequently did not intend to provide those in need with a full subsistence—and, when they did, were not good at judging the amount required—this should be unsurprising. It has made this history less state-centered (Katz and Sachsse 1996, Thane 2000). This trend has been reinforced by a change of social policy in several countries in the 1980s, Britain and the USA in particular. The replacement of direct state provision of services by contractual relationships with charities and profit-making providers has given prominence to the concept of the 'mixed economy of welfare.'

Third, at a time when the clients of the welfare state are drawn from all citizens, contemporary studies of social policies have come to pay attention to the mixture of resources at the disposal of citizens, particularly of the elderly, whether these were from public, occupational, or commercial sources. Since that mixture differs in different countries, comparative studies have moved from an exclusive preoccupation with state provision towards a concept of welfare regimes more broadly defined (Esping-Andersen 1990). The influence of comparative studies by social scientists on historiography is the final issue to be considered.

5.4 Comparative Studies

The historical study of welfare states was for a long time overparticularized. Attempts to identify categories suited for generalizations were left to social scientists specializing in the quantitative analysis of policy outcomes. This has been a dynamic field of study since the 1970s, but it developed in a particularly ahistorical form. One reason was its dependence on information provided by such international agencies as the United Nations and the Organization for Economic Co-operation and Development. Little of this went further back than the 1960s, so that the historical perspective was limited. Second, the questions asked and the explanatory power of the answers provided were constrained by the availability of such quantitative evidence, which was at first largely limited to figures of expenditure. The narrowness of these comparative studies was only slowly overcome as new evidence became available, e.g., the Luxembourg Income Study in the 1980s, which provided internationally comparable income-distribution data at microlevel, and as researchers began to approach national governments and other agencies directly to construct their own database. Third, the choice of appropriate categories for comparison across different national contexts is notoriously difficult and not always successful.

Evidence of this kind was, in any case, unable to provide explanations that advanced beyond the discovery of correlations. Intentions were ignored, because unsuited to standardization, and important dimensions of history were thereby ruled out of court. The most important result of these correlations was to group welfare state regimes into three main types: liberal, corporatist, and social democratic. It is more obvious that these describe ideal types than that they describe clearly differentiated clusters of indicators. Australians have demurred at the inclusion of their country in the same category as the USA; Germans have felt unhappy at being grouped with Italy and France; and the case of Britain is recognized widely as a hybrid. It is certainly not possible to write anything like a common history of the several welfare states grouped into any one of these types, with the possible exception of the four social democratic ones, who all happen to be Scandinavian.

Yet, by their very limitations, comparative empirical studies have helped to identify so-called anomalies and to encourage the intensive historical investigations of individual cases (Castles 1989, which includes Japan and Israel). The gulf between the social scientific and the historical study of welfare states has narrowed, as indicated by sharply focused comparative histories of pensions and of public health or by a five-volume history of individual welfare states addressing similar questions and written on a common editorial plan (Baldwin 1990, 1999, Flora 1986/2000).

See also: Capitalism; Labor Movements, History of; Social Welfare Policies and Gender; Social Welfare Policy: Comparisons; State, History of; Welfare State; Welfare State, Geography of

Bibliography

Abelshauser W 1987 *Die Weimarer Republik als Wohlfahrtsstaat* (*The Weimar Republic as a Welfare State*). Steiner, Wiesbaden/ Stuttgart, Germany

Ambler J S 1991 *The French Welfare State: Surviving Social and Ideological Change*. Cambridge University Press, New York

Baldwin P 1990 *The Politics of Social Solidarity. Class Bases of the European Welfare State 1875–1975*. Cambridge University Press, Cambridge, UK

Baldwin P 1999 *Contagion and the State in Europe 1830–1930*. Cambridge University Press, Cambridge, UK

Briggs A 1961 The welfare state in historical perspective. *Archives Européennes de Sociologie* **2**: 221–58

Castles F G (ed.) 1989 *Comparative History of Public Policy*. Polity Press, Cambridge, UK

Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Polity Press, Cambridge, UK

Flora P (ed.) 1986/2000 *Growth to Limits. The Western European Welfare States since World War II*, 5 vols. de Gruyter, Berlin

Flora P, Heidenheimer A J (eds.) 1982 *The Development of Welfare States in Europe and America*. Transaction, New Brunswick, NJ

Fraser D 1973 *The Evolution of the British Welfare State*. Macmillan, London

Katz M B, Sachsse C (eds.) 1996 *The Mixed Economy of Social Welfare*. Nomos, Baden-Baden, Germany

Koven S, Michel S (eds.) 1993 *Mothers of a New World. Maternalist Politics and the Origins of Welfare States*. Routledge, London

Lees L H 1998 *The Solidarities of Strangers. The English Poor Law and the People, 1700–1948*. Cambridge University Press, Cambridge, UK

Lindeman E C 1937 Public welfare. In: Seligman E R A (ed.) *Encyclopaedia of the Social Sciences*. Macmillan, New York

Lowe R 1993 *The Welfare State in Britain since 1945*. Macmillan, London

Marshall T H 1950 *'Citizenship and Social Class' and Other Essays*. Cambridge University Press, Cambridge, UK

Pedersen S 1993 *Family Dependence and the Origins of the Welfare State: Britain and France, 1914–45*. Cambridge University Press, Cambridge, UK

Ritter G A 1983 *Sozialversicherung in Deutschland und England. Entstehung und Grundzüge im Vergleich*. Beck, Munich, Germany
Ritter G A 1986 *Social Welfare in Germany and Britain. Origins and Development*. Berg, Leamington Spa, UK/New York
Ritter G A 1989 *Der Sozialstaat* (*The Welfare State*). Oldenbourg, Munich, Germany
Skocpol T, Ikenberry J 1983 The political formation of the American welfare state in historical and comparative perspective. *Comparative Research*. **6**: 87–148
Tennstedt F, Winter H 1993, 1995 Die Anfänge des Sozialstaats im Deutschen Reich von 1871, teil 1 und 2 (The beginning of the welfare state in the German Empire of 1871 Parts 1 and 2). *Zeitschrift für Sozialreform* **38**: 362–92 and **41**: 671–706
Thane P 2000 *Old Age in English History. Past Experiences, Present Issues*. University Press, Oxford, UK
Vincent D 1991 *Poor Citizens. The State and the Poor in the Twentieth Century*. Longman, London

E. P. Hennock

Well-being and Burnout in the Workplace, Psychology of

Job-related well-being is a specific aspect of subjective well-being and comprises as positive features phenomena such as job satisfaction, job attachment, involvement, and job morale; impaired job-related well-being includes experiences such as job tension, depression, burnout, and alienation from work (Warr 1990). The well-being concept refers both to minimal and temporary changes in employee mood and affect and to long-term disturbances of mental health (e.g., chronic anxiety or psychosomatic complaints). Research has paid more attention to long-term job-related well-being than to more momentary fluctuations in mood and affect. It is typical for work-related research to examine mental health and well-being in normally working samples who are not psychiatric patients.

1. *Conceptualizations of Well-being and Burnout*

Generally, it is assumed that well-being is not a unidimensional concept (Diener et al. 1999). This holds also for job-related well-being. However, authors disagree about the exact number and content of the separate dimensions of job-related well-being. Based on the conceptualization of arousal and pleasure as two independent dimensions, Warr (1987) proposed a model of job-related well-being comprising three dimensions: the first dimension runs from displeasure to pleasure, the second from anxiety to comfort, and the third from depression to enthusiasm. Additionally, feelings of competence and aspiration also can be regarded as facets of job-related well-being (Warr 1990).

Burnout is a specific and severe form of disturbed job-related well-being, originally mainly observed in human service professionals. The burnout syndrome comprises as key dimensions emotional exhaustion, depersonalization (i.e., negative and cynical attitudes towards one's clients or patients), and reduced personal accomplishment (Maslach and Jackson 1981). Researchers have observed that burnout does not develop only in human service professions, but also that it can occur in practically all kinds of jobs.

2. *Theoretical Models on Job-related Well-being and Burnout*

In the late twentieth century, researchers put much emphasis in the development of theoretical models on the relationship between work situation factors and job-related well-being (Kahn and Byosiere 1992). Among the most prominent models are the job characteristics model (Hackman and Oldham 1976), the job demand–job control model (Karasek and Theorell 1990), and the vitamin model (Warr 1987).

In their job characteristics model, Hackman and Oldham (1976) specified five core job characteristics which they assumed to be crucial for high internal work motivation and employee satisfaction: skill variety, task identity, task significance, autonomy, and feedback from the job. These job characteristics are expected to impact critical psychological states (i.e., experienced meaningfulness of the work, experienced responsibilities for work outcomes, and knowledge of the results of work activities), which in turn impact on the motivation and satisfaction as outcome variables. Furthermore, Hackman and Oldham proposed that moderators such as growth need strength, contextual satisfaction, knowledge, skills, and abilities have an effect on the relationship between core job characteristics and critical psychological states and also on the relationship between critical psychological states and outcome variables.

Similarly to the job characteristics model, the job demand–job control model (Karasek and Theorell 1990) received much attention within work and organizational psychology. In the original version, the job demand–job control model describes two dimensions relevant for job-related well-being: job demands and job control. In this model, job demands refer to high workload and time pressure; job control includes decision authority and intellectual discretion. It is assumed that the combination of high job demands and low job control resulting in high-strain jobs has severe negative effects on employee well-being. The model assumes that a combination of high job demands and high job control results in active jobs which have more beneficial effects for employee well-being. More recently, other work situation factors, such as social support and informational control, have been incorporated into the job demands–job control

model (Karasek and Theorell 1990, Terry and Jimmieson 1999). These extended models state that the effects of high-strain jobs will be most negative under conditions of low social support. High informational control, however, is predicted to attenuate the negative impact of high-strain jobs.

In his vitamin model, Warr (1987) proposed that job-related well-being is affected by job characteristics in a way similar to the effects of vitamins on physical health. Basically, the vitamin model suggests that some job characteristics have a linear effect whereas others have a nonlinear effect on well-being. Among the job characteristics assumed to be linearly related to well-being are physical security, a valued social position, and availability of money. Job characteristics assumed to show a curvilinear relationship with well-being include opportunity for personal control, opportunity for skill use, externally generated goals (i.e., demands), variety, environmental clarity, and opportunity for interpersonal contact. More specifically, the vitamin model assumes that well-being increases as these latter job characteristics increase; however, at a very high level, these job characteristics will cause a decrement in well-being.

Specific models have been developed which aim at explaining the emergence and development of burnout (Schaufeli and Buunk 1996). These models refer to individual, interpersonal, and organizational processes. For example, with respect to individual processes, burnout has been explained as resulting from a pattern of wrong and unrealistic expectations, as progressive disillusionment, as a failure in keeping up an idealized self-image, as a narcissistic disorder, as a disturbed action process, and as loss of coping resources. At the interpersonal level, the specific emotionally demanding relationship between caregiver and clients in addition to unsatisfactory social comparison and unbalanced social exchange processes are regarded as being the main causes of an individual developing burnout symptoms. With respect to organizational processes, individual burnout has been described as an effect of an unhealthy organization.

3. *Empirical Studies About Predictors of Job-related Well-being and Burnout*

Stressful working conditions were found to be negatively related to employee well-being, in both cross-sectional and longitudinal studies (Kahn and Byosiere 1992, Zapf et al. 1996) (see *Stress in Organizations, Psychology of*). Such stressful working conditions include physical stressors (e.g., noise), role stressors (e.g., role ambiguity, role conflict), work overload or time pressure, and more specific stressors such as monotonous work and attentional demands. Similarly, two out of the three burnout factors, i.e. emotional exhaustion and depersonalization, were found to be substantially correlated with job stressors such as lack of role clarity, role conflict, role stress, stressful events, workload, and work pressure (Lee and Ashforth 1996).

Moreover, other aspects in the work situation such as specified in the job characteristics model are positively related to employee well-being, particularly job satisfaction (Fried and Ferris 1987). Among these job characteristics, job control plays an important role. Job control is not only directly related to employee well-being but also acts as a buffer in the stress process. This buffer effect implies that in work situations with high-control stressful working conditions have a less negative impact on employee well-being than in work situations with low control. However, unlike depressive symptoms or anxiety, burnout seems to be unrelated to job control (Lee and Ashforth 1996).

Other work-related variables which play an important role in employee well-being are social support and group process characteristics, particularly a positive team climate and high cohesiveness within the team (Lee and Ashforth 1996, Sonnentag 1996). Persons who receive a high level of high social support from their supervisors and co-workers and also persons who experience a good team climate and a high cohesiveness within their teams enjoy a better well-being than persons who lack social support and who suffer from an unfavorable group process.

Additionally to these general findings about stressful working conditions, control at work and social support, research revealed that there is substantial individual variation in the relationship between work situation variables and individual well-being. Particularly an internal locus of control seems to protect one's well-being, also in relatively unfavorable work situations (Kahn and Byosiere 1992).

Moreover, job-related well-being is not only affected by a person's work situation or an interaction between work situation variables and individual characteristics such as locus of control. A number of studies have shown that personality dispositions such as negative affectivity are directly related to job-related well-being (Warr 1996). Also life satisfaction was found to have an effect on job-related well-being.

Taken together, job-related well-being and burnout have multiple causes, which lie partially in the person's work situation and partially in his or her personality. It seems that particularly the combination and interplay of various factors are responsible for a person's job-related well-being

4. *Empirical Studies About the Consequences of Job-related Well-being and Burnout*

Low job-related well-being and burnout are not only unpleasant experiences. They may threaten the individual's functioning within the work organization. Empirical research shows positive correlations be-

tween job-related well-being and job performance (Warr 1996). However, these correlations are relatively small in size and do not necessarily imply that well-being affects performance. It might be that performance has an effect on job-related well-being or that both well-being and performance are influenced by other factors such a leadership behavior. Moreover, studies revealed that impaired job-related well-being and burnout show small to moderate correlations with absenteeism and turnover intentions, two phenomena associated with high organizational costs (Lee and Ashforth 1996, Warr 1996).

5. *Intervention Programs*

Different kinds of interventions aim at the protection and improvement of job-related well-being. Most interventions are oriented towards the individual, i.e., to the way an individual perceives or deals with unfavorable working situations. Other individual oriented interventions are designed to directly influence the individual's psychological and physical well-being, for example, by progressive muscle relaxation, biofeedback, cognitive-behavioral interventions or meditation techniques (Kahn and Byosiere 1992, Murphy 1996). Empirical evaluations of these programs show that a combination of physical relaxation exercises and cognitive-behavioral interventions work best, particularly with respect to general aspects of well-being and the emotional exhaustion component of burnout (Murphy 1996, Schaufeli and Buunk 1996).

Another approach to improve job-related well-being consists of workplace interventions. During such workplace interventions, jobs are redesigned, e.g., by increasing employee control at work, by reducing specific job stressors or by introducing self-managed work groups. Empirical studies show substantial beneficial effects of such workplace interventions on job-related well-being (Kahn and Byosiere 1992). Since workplace interventions tackle the source of disturbances of well-being, they are to be preferred to individual-oriented intervention programs.

See also: Burnout, Psychology of

Bibliography

Diener E, Suh E M, Lucas R E, Smith H L 1999 Subjective well-being: Three decades of progress. *Psychological Bulletin* **125**: 276–302

Fried Y, Ferris G R 1987 The validity of the job characteristics model: A review and meta-analysis. *Personnel Psychology* **40**: 287–322

Hackman J R, Oldham G R 1976 Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance* **16**: 250–79

Kahn R L, Byosiere P 1992 Stress in organizations. In: Dunnette M D, Hough L M (eds.) *Handbook of Industrial and Organizational Psychology*, 2nd edn. Consulting Psychologists Press, Palo Alto, CA, Vol. 3, pp. 571–650

Karasek R, Theorell T 1990 *Healthy Work: Stress, Productivity and the Reconstruction of Working Life*. Basic Books, New York

Lee R T, Ashforth B E 1996 A meta-analytic examination of the correlates of the three dimensions of job burnout. *Journal of Applied Psychology* **81**: 123–33

Maslach C, Jackson S E 1981 The measurement of experienced burnout. *Journal of Occupational Behavior* **2**: 99–113

Murphy L R 1996 Stress management techniques: Secondary prevention of stress. In: Schabracq M J, Winnubst J A M, Cooper C L (eds.) *Handbook of Work and Health Psychology*. J. Wiley, Chichester, UK, pp. 427–41

Schaufeli W B, Buunk B P 1996 Professional burnout. In: Schabracq M J, Winnubst J A M, Cooper C L (eds.) *Handbook of Work and Health Psychology*. J. Wiley, Chichester, UK, pp. 311–46

Sonnentag S 1996 Work group factors and individual well-being. In: West M A (ed.) *Handbook of Work Group Psychology*. Wiley, New York, pp. 345–67

Terry D J, Jimmieson N L 1999 Work control and employee well-being: A decade review. In: Cooper C L, Robertson I T (eds.) *International Review of Industrial and Organizational Psychology*. Wiley, Chichester, UK, pp. 95–148

Warr P 1990 The measurement of well-being and other aspects of mental health. *Journal of Occupational Psychology* **63**: 193–210

Warr P 1996 Employee well-being. In: Warr P (ed.) *Psychology at Work*, 4th edn. Penguin, London, pp. 224–53

Warr P B 1987 *Work, Unemployment, and Mental Health*. Oxford University Press, Oxford, UK

Zapf D, Dormann C, Frese M 1996 Longitudinal studies in organizational stress research: A review of the literature with reference to methodological issues. *Journal of Occupational Health Psychology* **1**: 145–69

S. Sonnentag

Well-being and Health: Proactive Coping

1. *Proactive Coping as a Blend of Coping and Self-regulation*

Proactive coping refers to a set of processes through which people anticipate or detect potential stressors and act in advance to prevent them or to mute their impact (Aspinwall and Taylor 1997). As such, proactive coping is a blend of two extensively studied processes, *coping* and *self-regulation*. Coping consists of activities undertaken to master, reduce, or tolerate environmental or intrapsychic demands perceived as representing threat, harm, or loss (Lazarus and Folkman 1984). Self-regulation is the process through which people control, direct, and correct their own actions as they move toward or away from various goals (Carver and Scheier 1998). Proactive coping combines these two processes by examining people's emotions, thoughts, and behaviors as they anticipate

and address potential sources of adversity that might interfere with the pursuit of their goals (see Aspinwall 2001, Aspinwall and Taylor 1997).

1.1 Contrast to Other Approaches to the Study of Stress and Coping

The typical approach to the study of stress and coping examines how people try to solve existing problems and to manage the psychological distress they bring. The proactive coping approach, in contrast, focuses on people's efforts to identify potential sources of stress and take steps ahead of time either to prevent them or to mute their impact. For example, whereas a study of stress and coping might ask how students would manage if they had three final exams on the same day, a study of proactive coping might ask what students in such a position did ahead of time to alter their situation (for example, by asking one of their professors for an extension or by dropping one of the classes to avoid such a situation).

In the domains of aging and work, the proactive coping concept has been used to examine situations in which people, rather than being passively shaped by their environments, respond to actual or potential environmental changes by gathering information about new environments and selecting and modifying environments to meet their needs. For example, studies of proactivity in aging take as their starting point people's efforts to learn about their residential options, to select a particular retirement community in the first place (for example, by selecting one with opportunities for health-enhancing recreational activities), and to develop social resources once they arrive (Kahana and Kahana 1996). Similarly, studies of workplace entry suggest that new employees engage in several strategies to learn about their new environments and to develop social resources and that these strategies are often related to subsequent job performance and satisfaction (Ashford and Black 1996).

As these examples illustrate, proactive coping efforts may be applied to short-term, discrete situations or to developmental challenges and losses that are anticipated by the individual. In both cases, activities critical to understanding people's outcomes in educational, residential, and work settings occurred prior to or early in the course of some event or transition and served to change their experience of the situation. This signal property of proactive coping—the possibility of changing or altering one's outcomes in potentially stressful situations—is one of the principal distinctions between proactive coping and anticipatory coping. *Anticipatory coping* refers to efforts to brace oneself for the expected consequences of a known, imminent stressor. In proactive coping, the outcome of a situation may be unknown, and there is also room for one's actions to affect one's outcomes. Thus, proactive coping may involve a wide range of processes involved in trying to understand and alter various future possibilities (see Aspinwall 1997, Gollwitzer and Bargh 1996).

2. The Proactive Coping Model

The proactive coping model (Aspinwall and Taylor 1997) examines the personal, social, and situational factors that may influence the nature and quality of efforts to detect and respond to potential stressors. As shown in Fig. 1, proactive coping is thought to have five inter-related stages: (a) resource accumulation, (b) recognition of potential stressors, (c) initial appraisals of potential stressors, (d) preliminary coping efforts, and (e) the elicitation and use of feedback about the success of one's efforts.

Proactive coping starts with the accumulation of specific resources and skills that facilitate the detection of potential problems based on either warning signs from the environment or internal processes of reflection. Once a potential problem has been detected, it must then be appraised. At this stage, the predominant questions are, What is it? and What is it likely to become? Correspondingly, one must mentally simulate a variety of possible directions the situation and one's efforts to address it might take. Additionally, as illustrated by the star burst at right, the need to regulate negative emotional arousal evoked by detecting a potential stressor may compromise one's efforts to appraise it. Appraisal of a situation as potentially problematic may promote greater sustained attention to it, whereas efforts to reduce the threatening implications of the information may reduce attention to it, potentially compromising the accuracy of the appraisal. In the next stage of the model, initial appraisals of potential stressors give rise to preliminary coping efforts. These efforts, in turn, may be more or less successful in eliciting information about the nature of the problem and the appropriateness of one's preliminary efforts to address it. This information may, in turn, qualify one's appraisals and lead to the modification of coping efforts.

2.1 Using the Model to Understand the Relation of Individual Differences to Stress-related Outcomes

The model and its feedback loops provide a framework for understanding how individual differences, such as optimism and neuroticism, may be related to proactive efforts to anticipate and manage potential stressors and thus to long-term physical and mental health outcomes. For example, dispositional optimism, the expectation of more good than bad outcomes in the future, has been shown to predict increased attention to potential threats to health, as well as more elaborated appraisals of such threats (see Aspinwall 2001

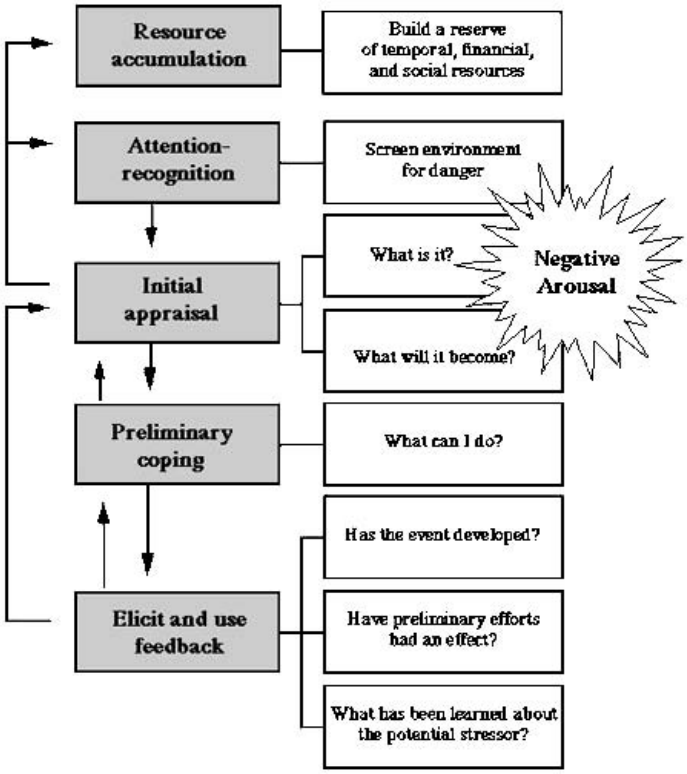

Figure 1
Aspinwall and Taylor's (1997) five-stage model of proactive coping (source: Aspinwall and Taylor 1997; reproduced by permission of American Psychological Association from *Psychological Bulletin* **121**: 417–36)

for review). In dozens of studies, optimism has also been linked to more active and less avoidant ways of managing problems. To the extent that active coping is more likely than avoidant coping to elicit information about their problems, optimists' subsequent coping efforts should be better informed and thus more likely to succeed. Through such processes, the successful proactive management of problems could account in part for optimists' success in a wide range of life domains.

In contrast, other individual differences may reliably impede the acquisition and use of information about potential problems and ways to manage them proactively. Some individual differences, such as neuroticism, anxiety, and repression, are linked to heightened attention to negative information, but also to difficulties in processing such information. Further, these traits typically predict more avoidant forms of coping, strategies that are less likely to elicit information about potential stressors. Over time, then, such ways of handling problems and information about them may lead to coping efforts that do not correspond well to the nature of the problem. Thus, the feedback loops in the model provide a way of thinking about the relation of individual differences to the accuracy of people's initial appraisals of problems, and for thinking about the ways that preliminary coping efforts may be differentially successful in eliciting useful information about problems.

2.2 *Other Influences on Proactivity*

Many other personal, social, and situational factors are likely to influence proactive coping (Aspinwall and Taylor 1997). For example, busy, demanding, unpredictable environments may make potential stressors more difficult to detect early in their course, leading to further resource depletion as problems may grow larger before they are detected. Further, such environments may interfere with the development of coping skills and social resources (Taylor et al. 1997). At the individual level, differences in the value placed on future, as opposed to current, outcomes may also play a role in important forms of proactivity, such as preventive health behavior and conservation efforts (Strathman et al. 1994).

3. Applications to Health

Successful proactive coping may promote physical and mental health in many ways. First, addressing a problem earlier in its course requires fewer resources and may be more likely to be successful. The proactive coping model may be useful in understanding preventive behavior and early responses to physical symptoms, including emotional and other barriers to early detection. Second, successful proactive efforts to manage both large and small problems should serve to reduce an individual's total exposure to stress and thus be linked to better physical and mental health. Even if problems cannot be prevented altogether, proactive efforts to address potential problems are unlikely to be wasted, as they may contribute to one's knowledge base for use in future efforts. Third, in studies to date, a central aspect of proactivity seems to be the development of social resources, which themselves have robust links to good health outcomes.

4. Future Directions

Continuing research on proactivity in major life transitions, as well as in everyday behavior, should foster an increased understanding of the social cognitive and developmental underpinnings of coping. To date, the coping literature has been characterized by excessive reliance on retrospective self-reports of coping over long intervals and by coping measurements that sum together many different kinds of activities undertaken to solve problems, regulate emotions, and shift priorities in the face of adversity. Careful delineation of the component activities of proactivity, such as regulation of attention, planning, mental simulation, and the elicitation and use of feedback about the success of one's coping efforts, how such components develop (Friedman and Scholnick 1997) and influence lifespan development (Freund and Baltes 1998), and how they are affected by personal, social, and environmental factors will add greatly to the understanding of successful proactivity and, ultimately, inform interventions to foster proactivity in health, educational, and work settings (Schwarzer 2000).

Additionally, increased attention to the role of social factors in successful proactivity may prove fruitful. In the aging and workplace examples, people entering new environments acted quickly to cultivate social resources. Social resources may play a key role in proactive coping, as people use the opinions and outcomes of others to make predictions about what the future might hold for themselves (Aspinwall 1997) and actively enlist others' assistance in appraising and managing potential stressors. Further, understanding how people prevent losses of social resources by altering their behavior to counteract expected discrimination (Miller and Myers 1998) or by responding to incipient problems in close relationships may prove to be important in understanding mental and physical health outcomes over time.

Finally, the study of proactivity may be expanded to include thoughts and actions in the service of not only preventing adverse outcomes, but also creating future opportunities (Schwarzer 2000). Future efforts to understand both kinds of proactivity may increase our understanding of people's capacity to shape their environments and outcomes through their capacities for future-oriented thinking and action.

See also: Adulthood: Developmental Tasks and Critical Life Events; Coping across the Lifespan; Coping Assessment; Depression, Hopelessness, Optimism, and Health; Life Management, Developmental Psychology of; Stress and Coping Theories

Bibliography

Ashford S J, Black J S 1996 Proactivity during organizational entry: The role of desire for control. *Journal of Applied Psychology* **81**: 199–214

Aspinwall L G 1997 Future-oriented aspects of social comparisons: A framework for studying health-related comparison activity. In: Buunk B P, Gibbons F X (eds.) *Health, Coping, and Well-being: Perspectives from Social Comparison Theory*. Erlbaum, Mahwah, NJ

Aspinwall L G 2001 Dealing with adversity: Self-regulation, coping, adaptation, and health. In: Tesser A, Schwarz N (eds.) *The Blackwell Handbook of Social Psychology: Intraindividual Processes*. Blackwell, Malden, MA

Aspinwall L G, Taylor S E 1997 A stitch in time: Self-regulation and proactive coping. *Psychological Bulletin* **121**: 417–36

Carver C S, Scheier M F 1998 *On the Self-regulation of Behavior*. Cambridge University Press, New York

Freund A M, Baltes P B 1998 Selection, optimization, and compensation as strategies of life management: Correlations with subjective indicators of successful aging. *Psychology and Aging* **13**: 531–43

Friedman S L, Scholnick E K (eds.) 1997 *The Developmental Psychology of Planning: Why, How, and When Do We Plan?* Erlbaum, Hillsdale, NJ

Gollwitzer P M, Bargh J A (eds.) 1996 *The Psychology of Action: Linking Cognition and Motivation to Behavior*. Guilford, New York

Kahana E, Kahana B 1996 Conceptual and empirical advances in understanding aging well through proactive adaptation. In: Bengston V L (ed.) *Adulthood and Aging: Research on Continuities and Discontinuities*. Springer, New York

Lazarus R S, Folkman S 1984 *Stress, Appraisal, and Coping*. Springer, New York

Miller C T, Myers A M 1998 Compensating for prejudice: How heavyweight people (and others) control outcomes despite prejudice. In: Swim J K, Stangor C (eds.) *Prejudice: The Target's Perspective*. Academic Press, San Diego, CA

Schwarzer R 2000 Manage stress at work through preventative and proactive coping. In: Locke E A (ed.) *The Blackwell Handbook of Principles of Organizational Behavior*. Blackwell, Oxford, UK

Strathman A, Gleicher F, Boninger D S, Edwards C S 1994 The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology* **66**: 742–52

Taylor S E, Repetti R L, Seeman T 1997 Health psychology: What is an unhealthy environment and how does it get under the skin? *Annual Review of Psychology* **48**: 411–47

L. G. Aspinwall

Well-being (Subjective), Psychology of

When people evaluate the quality of their lives, we label this 'subjective well-being.' Their appraisals might come in the form of life satisfaction judgements, as pleasant emotional reactions to life events, or as feelings of fulfillment. Thus, subjective well-being (SWB) includes both cognitive judgements and affective reactions in which people evaluate their lives and circumstances. It can vary from abject depression and complete dissatisfaction at the low end, to elation and consummate satisfaction at the high end. In addition to reactions to one's life taken as a whole, subjective well-being includes evaluations of specific components of one's life such as work, marriage, health, recreation, and religion. SWB includes global assessments of a person's life from childhood to the present, and also more specific evaluations that are focused on a narrower time frame such as the past month. In sum, SWB includes all the components about how people think of and evaluate their lives. SWB includes everyday concepts such as happiness, morale, zest for life, and peace of mind.

1. *SWB and Quality of Life*

A pivotal early finding in the field was Bradburn's (1969) discovery that pleasant emotions and unpleasant emotions form separable entities, and different factors influence each. In other words happy emotions and unhappy ones are not polar opposites: people's experience of one does not tell us much about how much of the other they feel. The importance of this finding is that even if we understand totally what causes depression, anxiety, and unhappiness, we may have little knowledge of what causes positive feelings of joy, affection, and contentment. Thus, we must understand the positive end of well-being to gain a full understanding of quality of life. The balance between pleasant and unpleasant emotions in a person's life is termed 'hedonic balance.'

In the 1970s another major component was added to SWB: life satisfaction. It was discovered that although satisfaction with life was related to hedonic balance, it formed a separate factor. Therefore it is now commonplace for SWB researchers to measure life satisfaction, pleasant affect, unpleasant affect, and satisfaction with specific life domains.

The classic Greek philosophers inquired as to what constitutes the 'good life,' and Asian thinkers such as Confucius explored similar questions. Many valued characteristics such as virtue, wisdom, and pleasure have been proposed as essential components of the quality of life. In the field of subjective well-being, researchers examine how people themselves evaluate their lives, using whatever standards and characteristics they choose. In this approach the factors that determine whether people will view their life as good are studied, and researchers do not impose their own set of values in judging that life. Thus, SWB is a democratic approach to quality of life in that it grants recognition to people's own evaluations of their lives.

As the wealthier nations achieved economic prosperity during the twentieth century, people became increasingly concerned with quality of life that goes beyond material prosperity. Social indicators measuring factors such as education, health, and equality arose in order to yield a measure of well-being that transcended economic security. At the same time, after World War II an interest in subjective well-being variables led survey researchers to measure quality of life from people's internal frame of reference. Pollsters asked people how satisfied and happy they were. In early survey research there was an emphasis on the effects of demographic variables on SWB. For example, Campbell, Converse, and Rodgers (1976) examined samples of Americans, and determined how much demographic factors such as age, sex, income, education, and ethnic group correlated with reports of SWB. To their surprise—and this finding has been replicated numerous times—demographic variables accounted only for a modest portion of SWB. If demographic variables such as income do not account for much of the variation in SWB, what does?

2. *Temperament*

People's personalities have an important influence on whether they are happy or unhappy (see Diener and Lucas, chap. 11 in Kahneman et al. 1999). Studies of monozygotic vs. dizygotic twins who were reared together or apart from birth demonstrate that SWB has a substantial heritable component. Because they share all their genes in common, identical twins reared apart are more similar in their levels of SWB than are fraternal twins who are raised together. Although heritability tends to be strong, the influence of shared early home environment appears to be weak. Other types of studies tend to support the importance of personality to SWB. For example, there is a substantial consistency across situations in people's satisfaction, as well as considerable stability of SWB in adulthood. Furthermore, personality traits such as extraversion

and neuroticism show substantial correlations with levels of pleasant and unpleasant affect, respectively (Headey and Wearing, 1989).

3. *Adaptation*

When good and bad events happen to people, it is not surprising that they react with happiness or unhappiness. Over time, however, they usually adapt, returning toward their earlier level of SWB (see Diener et al. 1999 for a review). Often people may adapt to very good or very bad events in a relatively short time. Adaptation combined with the effects of personality led Headey and Wearing (1989) to propose the 'dynamic equilibrium' model in which personality determines the long-term baseline level of SWB, but in which events can temporarily move people above or below their baseline. However, it must be noted that people do not completely adapt to all life circumstances; some environmental factors can move people's baseline up or down. For example, people who are born with multiple disabilities are less happy as adults, even though they have had many years to adapt to their conditions. Similarly, people in very poor nations are also less happy than those in wealthy nations. Thus, happiness seems to arise from a combination of the effects of personality with long-term conditions that have a pervasive effect on people's lives. Recently, the idea has emerged that SWB also depends on cognitive habits related to how people perceive and interpret events.

4. *Goals*

People have various values and goals that they believe are important. When they make progress toward their goals, they tend to be happier, and when they fail to move toward their goals, they are less happy. One way that people can adapt to bad conditions is to change their goals. For example, if a woman is failing at her goal to become president, she might focus on doing her job particularly well, and on the accomplishments within that position, and therefore the disappointment due to the failure would be likely to diminish. On the other hand, happy people tend to adopt goals for which they possess adequate resources to be successful. Thus, it is found that resources such as money, physical attractiveness, or social skills are most related to SWB when they are relevant to achieving the person's goals (see Diener et al. 1999).

Ryff and Singer (1998) propose that purpose and meaning in life are essential to mental health. Several researchers have concluded, however, that the content of a person's goals can influence SWB. Brunstein et al. (1998) found that goal strivings must be congruent with a person's motives to produce SWB, and Kasser and Ryan (1996) proposed that goals must be intrinsic rather than imposed to enhance SWB. Emmons (1986) found that success at goal strivings is related to positive affect, whereas conflict between one's goals is related to negative affect. Thus, the relation between goal striving and SWB is more complex than originally believed.

5. *Cultural Differences in SWB*

Certain variables predict life satisfaction in some cultures, but not in others (Diener and Suh 2000). For example, self-esteem is a strong correlate of satisfaction in highly westernized, individualistic cultures, but not in collectivistic societies where the group is more important in defining who one is. Similarly, consistency and acting in congruence with the self are stronger predictors of SWB in individualistic than in collectivistic cultures. Also, individualists on average heavily use their own emotional feelings to judge their life satisfaction, whereas collectivists are more likely to weight normative prescriptions for happiness, and the views of others. For collectivists, SWB must include the belief that significant others evaluate one's life well. The explanation for these differences between nations is that various cultures emphasize different values and goals, and therefore tend to make certain information chronically salient so that it is used when people judge their lives.

Not only are there cultural differences in the correlates of SWB, but average SWB levels differ across nations as well. For instance, high levels of SWB are currently reported in northern European nations, whereas much lower levels are reported in eastern European countries. Several factors seem to account for the differences between nations in SWB: wealth versus poverty, political stability versus instability and social disruption, and culture. In terms of culture, some societies seem to emphasize a positive approach to life, and the desirability of happiness. In these 'positivity cultures' there are higher levels of satisfaction than are predicted based on wealth or objective factors alone.

6. *Changing Focus*

In the early days of research on SWB, the focus was on large surveys, and the determination of what demographic factors are most correlated with 'happiness.' Recently attention has shifted to the influence of temperament, in combination with environmental circumstances. However, another new direction is to examine the processes that determine feelings of well-being. For example, the goals people select may be as important to satisfaction as are objective life circumstances. Similarly, people may focus on the desirable elements in their lives or the problems, and this can influence their SWB. People's memory structures for good versus bad events, their interpretation of events, and the information they use when judging

their life circumstances have recently become important topics of study.

7. *Methodological Issues*

Because the area of subjective well-being has roots in survey research, broad, representative samples of many nations are available, and this has been a major strength of the field. Questions arise, however, in considering the validity of the measures of SWB. Virtually all measurement of variables in this area have been based on global self-reports: on one occasion people report on numbered scales how happy or satisfied they are. Fortunately, research has shown that there is a degree of validity to these survey instruments. The SWB scales correlate with nonself-report measures such as smiling, the reports of family and friends, interviewer ratings, daily reports of mood, and whether individuals can quickly recall more good events than bad events from their lives. Nonetheless, certain obvious biases can enter into self-report measures of SWB. For example, people cannot recall with accuracy their moods over a long period of time. For this reason, researchers in the field have begun to supplement the global self-report measures with experience sampling measures of SWB, in which people are asked at random moments about their moods and level of satisfaction. The most desirable way to measure SWB is with a battery of assessment devices, including experience sampling, informant reports, and memory measures because these measurement methods complement each other in terms of protecting against specific measurement artifacts, and also because together they yield a multimodal assessment of various manifestations of well-being: experience (global self-reports and experience sampling), behavior (reports of informants and nonverbal behavior), and memory (the recall of good and bad events).

One of the challenges in studying SWB is to determine what causes it and what is caused by it. Because it is difficult to do experimental work with random assignment to conditions in this field, most work is correlational and cross-sectional. This leaves open the question of whether many variables are causes or consequences of SWB. For example, if people feel in control of their lives and report that their marriage is strong, are these causes of happiness, results of being happy, or caused by some third variable such as temperament? Diener (1984) referred to this issue as the question of 'top-down' vs. 'bottom-up' influences on SWB.

One method used to gain some insight into causal networks influencing SWB is longitudinal panel designs in which people are measured repeatedly over time. From such data we can conclude, for example, that marriage on average causes SWB rather than simply the alternative explanations that happy people marry more often, and stay married more. Similarly, there is longitudinal evidence suggesting that income leads to somewhat higher SWB rather than it being the case that happy people earn more money. In a few areas there are experimental or quasiexperimental studies. For instance, lottery winners seem to be somewhat happier than controls, suggesting that money may have a causal influence on SWB. In addition, negative income tax experiments (in which randomly selected welfare recipients are given greater monetary benefits) suggest that increased income can also heighten stress. Thus, there is beginning to be some examination of causality in this field, but most conclusions remain based on self-report measurement at a single point in time, and thus do not speak to issues of causality.

8. *Future Research*

Because SWB is a relatively new field, and scientists have only recently begun research in this area, there are many unanswered questions. One issue is, given the influence of temperament, the degree to which happiness can be learned. For instance, can people learn cognitive styles of perceiving and interpreting the world that will heighten their joy and contentment? A related question is: what is the optimal level of SWB? People cannot and should not be happy every moment; the unpleasant emotions serve useful functions. But what degree of pleasant vs unpleasant emotions is optimal for healthy functioning? Similarly, what level of life satisfaction will allow people to remain motivated and perform well, and yet view their lives in positive terms?

The possibility of attaining happiness through modern drugs brings the above question into focus. On the one hand, people could conceivably make themselves very happy on an ongoing basis with drugs that will be created in the foreseeable future, pharmacopeias without the side-effects of current drugs. Would this method of achieving happiness undercut motivation and interfere with our other values? For example, might people become too happy, and might they then ignore good social relationships and personal responsibilities?

The scientific study of SWB is in its infancy, and therefore researchable questions abound in virtually every area. We now have developed, however, the basic methodological tools to analyze empirically how people feel about their lives, and why.

See also: Chronic Illness: Quality of Life; Neuroticism; Personality and Adaptive Behaviors; Personality Structure; Quality of Life: Assessment in Health Settings; Self-esteem in Adulthood; Self-evaluative Process, Psychology of; Social Comparison, Psychology of

Bibliography

Argyle M 1987 *The Psychology of Happiness*. Methuen, London

Bradburn N M 1969 *The Structure of Psychological Well-being*. Aldine, Chicago, IL
Brunstein J C, Schultheiss O C, Grassman R 1998 Personal goals and emotional well-being: The moderating role of motive dispositions. *Journal of Personality and Social Psychology* **75**: 494–508
Campbell A, Converse P E, Rodgers W L 1976 *The Quality of American Life: Perceptions, Evaluations, and Satisfactions*. Russell Sage Foundation, New York
Diener E 1984 Subjective well-being. *Psychological Bulletin* **93**: 542–75
Diener E SuhE (eds.) 2000 *Subjective Well-being Across Cultures*. MIT Press, Cambridge, MA
Diener E, Suh E M, Lucas R E, Smith H 1999 Subjective well-being: Three decades of progress. *Psychological Bulletin* **125**: 276–302
Emmons R A 1986 Personal strivings: An approach to personality and subjective well-being. *Journal of Personality and Social Psychology* **51**: 1058–68
Headey B, Wearing A 1989 Personality, life events, and subjective well-being: toward a dynamic equilibrium model. *Journal of Personality and Social Psychology* **57**: 731–9
Kahneman D, Diener E, Schwarz N (eds.) 1999 *Well-being: The Foundations of Hedonic Psychology*. Russell Sage Foundation, New York
Kasser T, Ryan R M 1996 Further examining the American dream: Differential correlates of intrinsic and extrinsic goals. *Personality and Social Psychology Bulletin* **22**: 280–7
Myers D G 1992 *The Pursuit of Happiness: Who is Happy—And Why*, 1st edn. W Morrow, New York
Meyers D G Diener E 1996 The pursuit of happiness. *Scientific American*: 70–72
Ryff C D, Singer B 1998 The role of purpose in life and personal growth in positive human health. In: Wong P T P, Fry P S (eds.) *The Human Quest for Meaning: A Handbook of Psychological Research and Clinical Application*. Earlbaum, Mahwah, NJ, pp. 213–35
Strack F, Argyle M, Schwarz N (eds.) 1991 *Subjective Well-being: An Interdisciplinary Perspective*, 1st edn. Pergamon Press, Oxford
Veenhoven R 1993 *Bibliography of Happiness: 2472 Contemporary Studies on Subjective Appreciation of Life*. Erasmus University Rotterdam, Department of Social Sciences, Center for Socio-Cultural Transformation, Rotterdam, The Netherlands

E. Diener

West Africa: Sociocultural Aspects

This article offers an overview of social worlds, values, and material and nonmaterial cultures of the region south of the Sahara from Mauritania to Cameroon. Attention is paid, in particular, to modes of social solidarity and the cultural dynamics of community formation in West African settings.

1. *The People*

West Africa is mainly peopled by Black Africans. Some groups of Berber and Arab descent inhabit the Saharan extremes. A small Lebanese diaspora, distributed mainly along the coast, occupies a role in commerce not unlike the Indian traders of eastern and southern Africa. Malaria deterred European settlement.

2. *Linguistic Diversity*

A striking aspect of West Africa is its linguistic diversity (Dalby 1977). Nigeria, a country of c. 100–120 million people, covering only about a seventh of the region, is estimated to have between 250 and 400 languages. (The huge expanse of eastern and southern Africa, by contrast, is dominated by closely related Bantu languages.) Language diversity in West Africa mainly reflects *in situ* development of long-settled farming communities, rather than long-distance migration. European languages, especially English and French, became established as public languages during the colonial period.

3. *The Agrarian Base*

West Africa divides into three broad environmental and cultural zones—an agro-pastoral desert edge dotted by ancient towns involved in trans-Saharan trade, a coastal rain forest zone shaped by external (sea-borne) commerce (including the Atlantic Slave Trade), and an intermediate zone of drier forest margins and wetter savannas, with a history of endogenous agrarian development dating back several thousand years.

The intermediate zone is the heartland of West African civilization and culture. It is here that local crops, notably white yam (*Dioscorea rotundata*), African rice (*Oryza glaberrima*), and a range of millets ('hungry rice,' pearl millet, and one race of sorghum), were first domesticated. The intermediate zone also provided the oil palm (*Elaeis guineensis*)—the source of oil used in cooking, a mild alcoholic beverage (palm wine), raw material for soap and medication, and ribs, trunks and leaves from which bags, hammocks, roofs, building frames, and bridges are made.

In the eastern part of West Africa, along the forest–savanna transition and in the river valleys of the savanna, yam-oil palm farming systems prevail. Away from the riverine environments sorghum and pennisetum millet tend to replace yam as staples. The pattern repeats in the western half of West Africa, but with African rice taking the place of white yam (though today African rice is partly replaced by Asian rice, perhaps first introduced by the Portuguese in the era of the slave trade but spread widely only in the last century).

The early rise of a relatively intensive heartland agriculture went hand in hand with other socio-technical changes, as evidenced by ancient iron working and terracotta art work, suggesting permanent

settlement and complex division of labor in the intermediate zone at an early date.

The sacred importance and symbolic significance accorded intermediate-zone crops, even where they have been squeezed out of farming by commercially more productive introductions, is a notable aspect of West African material culture. Wild palms supply a rich red oil that remains an essential ingredient in many rituals, ceremonies, and remedies, greatly favored over the more common plantation substitute. The tasteless white rices of the coastal wetlands are considered no substitute for the nutritionally superior red rices of the interior. From Ghana to eastern Nigeria, yam survives as a preferred staple despite cheaper root crop substitutes. In the *Arrow of God*, Nigerian novelist Chinua Achebe appropriately made the disruption of a new yam festival in eastern Nigeria a striking metaphor for the far-reaching disruption of West African social and cultural values caused by colonial conquest (Achebe 1964).

4. *Lineage Society: 'Wealth in People'*

West African agriculture depends on the hoe. The soils of the wetter savannas and adjacent forests are not suited to plough cultivation. Due to fly-borne disease, pastoralism is well developed only along the desert margins. Elsewhere it is not easy to accumulate wealth in livestock. Successful hoe agriculture depends on abundance of human labor. A central social value in the West African heartland is 'wealth in people.' This has many manifestations. The importance of a chief, for example, is expressed in terms of wives, children, and domestic dependents, rather than through holdings of land and livestock.

But in hoe agriculture, timing and coordination of labor are as important as numbers of workers, especially in strongly seasonal savanna cultivation systems. Interhousehold labor cooperation, and the social values, crop types, and farming practices conducive to such cooperation figure strongly in West African agrarian social systems. In the old days the young people of an entire community might pool their labor for farm operations. Today it is more likely that they will form cooperative gangs to tackle back-breaking and time-constrained jobs on a rotational basis. Where agriculture is commercialized labor gangs ply for hire.

Labor cooperation is a practical manifestation of an egalitarianism among the young typical of West African heartland social organization. This egalitarianism is frequently fostered through initiation into so-called 'secret' societies, widespread and still powerful in parts of the region. Such societies typically tend to support social values among the young, strikingly in contrast with hierarchical values prevailing in the world of the elders. From Guinea to Cameroon the senior grades within secret societies form a ladder along which elders compete to advance.

French Marxist anthropologists in the 1960s and 1970s wrote about this typical opposition between youthful egalitarians and ranked elders, as 'lineage' society, or a 'lineage mode of production,' clearly differentiated from feudal relations found in medieval Europe and elsewhere (Seddon 1978). It was conceived that elderly elites, drawing on the language of kinship and descent for legitimation, reproduced themselves at the expense of the labor of cadets. According to some authors this amounted to a kind of class conflict. The explanation was controversial. An alternative explanation is that typical agrarian societies in the West African heartland are the complex outcome of a number of competing principles of social solidarity, none of which, by itself, can sustain a satisfactory community life over the longer term.

5. *Agrarian Social Solidarity*

As suggested, two principles of social solidarity often encountered in West African heartland social formations are egalitarianism and hierarchy. They offer quite distinct approaches to the problems of keeping social groups together (Douglas 1993). Neither, alone, can sustain a community except in the short term. Compromise is an essential prerequisite for survival.

In a land-abundant setting, hierarchs may be able to do little to prevent youthful egalitarians from forging their own breakaway communities. But breakaway youthful egalitarians quickly run into problems of leadership. Youth labor gangs often have their own 'government,' and spend large amounts of time in dispute settlement, to the detriment of the work in hand. Charismatic leadership sometimes helps the egalitarians to go-it-alone for a time, but as with sectarians everywhere, splits lie around the corner. The more settled authority of the elders offers clear advantages. The typical West African agrarian compromise seems to have allowed for the emergence of spheres of partial autonomy for women and youth within a more complex social totality presided over by elders (and ancestors). Achieving such coordination and compromise without recourse to strong centralized political institutions was often the work of the 'secret societies' (for both men and women) found particularly in the Upper Guinean section of the Atlantic Coast in the western half of West Africa, and more widely in the region. In communities welded by initiation and secrecy, a social contract renewed or reinforced typically manifests itself from time to time in dramaturgical terms (often as masquerade display). The arts of masking are among the greatest manifestations of West African material culture and societal values.

The test of any mode of social solidarity is the extent to which it permits humans to flourish. A central fact about many decentralized ('stateless') societies of the

West African heartland is that they often sustained denser populations than adjacent territories with a history of strong (and often strongly oppressive) central institutions (Middleton and Tait 1958).

6. *Society and Culture at the Margins: Coast and Desert Edge*

International trade rather than agriculture was a mainstay of life at the coast and on the desert edge. It is in these regions that we first see most clearly the emergence of a trade-related West African urbanism, with its own distinctive institutions and cultural values. Along the desert edge, long-distance trade was associated with the early spread of Islam, and with Islam came literacy, science, and an enhancement of bureaucratic capacity. A process of modernization and reform took shape among desert-edge trade-related towns and polities in the late eighteenth century, and by the mid-nineteenth century larger regional formations, such as the Sokoto Caliphate (Last 1969), covered a large part of what is today northern Nigeria. Two interior languages—Hausa and Maninka (Dioula)—developed some importance as media of regional communication, along with written Arabic for letter writing and record keeping.

At the coast, trade was stimulated by European contacts from the fifteenth century. Christianity gained several footholds, and small ports and polities oriented to the Atlantic trade developed, at times ruled by mercantile elites of Afro-European descent. Coastal trade pidgins, drawing first on Portuguese and later on English, developed into a distinctive family of mutually comprehensible English-oriented pidgins and creoles, today perhaps as widely spoken in coastal urban centers from The Gambia to Cameroon as Hausa or Dioula, the trade languages of the savanna, are spoken in interior West Africa. The English-oriented pidgins and creoles have considerable importance in coastal West Africa as means of communication among the nonelite, and in popular music and entertainment (Todd 1984).

Globalization is no new phenomenon in desert-edge or coastal West African communities. From an early date children of trading elites might be sent to Cairo or Liverpool for their schooling. Nelson's victory against Napoleon's fleet at the Battle of the Nile in 1803 was known in Kumasi within weeks; that the news made sense at all implies a sophisticated world map in the West African interior. The typewriter was in use among Christian converts in the interior of Yoruba-speaking country (now western Nigeria) within decades of its invention. The colonial telegraph, and now the modern satellite phone, are among means through which globally dispersed West African families have maintained and developed their wide networks of social contacts in the modern period.

Conversion to the 'world religions,' Islam and Christianity, tells us about growing global awareness even at early dates. Coastal linguistic creolization was part of a larger process of cultural creolization that helps explain the international outlook and mobility of many West Africans today (Hannerz 1987). Diaspora communities of West Africans, often involved in trade, are found in most world cities today, and not just in the capitals of the former colonial powers. It is not uncommon to find street children or unemployed youth in Dakar, Douala, Freetown, or Lagos with a better grasp of politics and world issues than well-educated teenagers in Britain, America, or France.

7. *The Impact of External Trade on Social Values*

Arguments and compromises over social values took different forms under the impact of long-distance trade from those already described for heartland agrarianism. We need to distinguish internal and legitimate trade from the slave trade. Vigorous internal trade, exchanging forest and savanna products (notably cattle, kola nuts, and salt) was a stimulus to social and cultural innovation in a number of communities located on major north–south internal trade routes.

Typical adaptations to long-distance trading (Cohen 1969) included the development of culturally distinct diasporas (Hausa cattle dealers in western Nigeria, or Fula kola dealers in Sierra Leone for example). At times agrarian elites maintained some control of diaspora trading groups through intermarriage. At other times diaspora communities have been confined to special settlements (Zongo, Sabon Gari) at the edge of villages or towns.

In other cases, women or youth have been free to exploit the potential of semi-autonomous social spheres to build up commercial activities. Igbo young men from eastern Nigeria have long adapted the gang-like structures of the labor company to itinerant trading (in for example second-hand clothing, cloth or motor spares). Agrarian labor cooperatives often double, or develop, as rotational savings and credit associations for the purposes of commerce.

Women traders are prominent throughout the West African costal and forest zone, but especially so in Ghana and southern Nigeria. Large-scale female traders, sometimes with little or no formal education, are to be encountered traveling the world, often in groups, perhaps as likely collecting car spares in Taiwan as exporting a consignment of red palm oil to London. West African women's involvement in commerce typically began with local and regional foodstuffs trade neglected by the colonial commercial houses. The extensive yam trade from the Middle Belt interior to burgeoning colonial coastal cities of Ghana and Nigeria would be an example. In patrilineal

communities women are life-long brokers between kin groups. Taken in conjunction with the gender solidarity fostered by initiation, this helps set up an extra household social space in which women's commerce has tended to flourish.

But we need to be careful not to impute modern ideas about individualism and enterprise to the development of West African trading relationships with the wider world. Frequently overseas or trans-desert trade was the monopoly of rulers who kept a tight rein on contacts and commercial information. Those controlling the trade-based polities of coast and desert-edge were more likely to be monopolists and autocrats than agents of an enterprise culture. In the nineteenth century the spread of firearms allowed trade-oriented rulers extensive use of socially unaccountable force. A not dissimilar militaristic adventurism has reasserted itself in some parts of West Africa (notably in Liberia and Sierra Leone) as the twentieth century ends, prompted by an international political vacuum and a surfeit of cheap weapons from the ending of the Cold War.

The mode of social solidarity that tends to predominate under war-lord autocracy is not individualism but fatalism, a fact illustrated in the 1997 elections in Liberia, where a war-weary population voted for the strongest of the faction leaders from a destructive seven-year civil war, reasoning the most powerful autocrat was the best guarantee of a kind of peace. Fatalism is little appreciated as a basis for social solidarity, but may be preferable to the mayhem of a bush civil war. Here we have an insight into why social life on the West African frontier so often seems to have been characterized by quietism, and a subsistence-oriented conservatism, rather than by the expansive individualism associated with the making of the American frontier (Kopytoff 1987). The attitude is 'we make no demands other than to be left to get on with our lives.' The old heartland agrarian social mix continues to have a certain basic appeal where the alternative is the postmodern violence of late twentieth century conflict in Liberia and Sierra Leone.

8. *Social Consequences of the Slave Trade and Colonial Rule*

Perhaps between 10 and 15 million West Africans were sold into Atlantic slavery from the sixteenth to nineteenth centuries, and 8–10 million into trans-Saharan slavery. In addition to the adverse demographic impact, in a region dependent on labor-intensive hoe agriculture, the conversion of human beings into commodities dealt an enormous blow to the trust essential to the functioning of many social institutions (Inikori 1982). It is hard to overestimate the continuing consequences of the slave trade for the evolution of social values in West Africa. Fatalism, fear of entrepreneurial individualism, distrust of chiefs, and undue respect for autocrats in return for basic security are among the symptoms of the damage. Distrusts inherited from the era of the slave trade remain a factor in the nihilism of much of the current youth-led violence blighting the region. How ordinary West Africans continue to come to terms with the harsh social legacy of the slave trade is also very apparent in popular religion, in ideas about chiefs as shape-shifting witches, or through divination and possession as ways of addressing old betrayals and coming to terms with disastrous breakdowns in family or community solidarity (Shaw 1997).

Colonial protection in West Africa arose from initiatives to eliminate the slave trade, beginning with the assumption in 1807, by the British Crown, of administrative responsibility for Sierra Leone, an eighteenth century West Coast settlement for freed slaves founded by British Christian philanthropists. Ending the slave trade became the pretext for a series of moves by industrializing European powers to control resource extraction and guarantee markets for their own manufacturers. By the end of the nineteenth century West Africa had been carved up by the Germans, French, and British (Crowder 1968).

Colonial rule at the beginning of the twentieth century was militaristic, often formally racist, and based on a protectionist principle of imperial preference. Although colonial regimes built railways and some roads, and Christian missions under colonial protection introduced school education and Western medicine, colonial autocracy did nothing to repair the adverse social legacy of the slave trade.

9. *Postcolonial Social Trends*

Postcolonial leaders in the region have continued the path of colonial autocracy, paying more attention to external opportunities than local accountability. Aid donors and Western governments hope to see a flowering of individualism and rational choice to support a flourishing modern market society, but without taking full account of the extent to which the slave trade and colonialism prevented the development in West Africa of the informal checks and balances between state, market, and civil society upon which modern market society depends. West African people still retain the option of retreating towards heartland survivalism. But even this option is no longer as open to all comers as once it was.

The population is ever more youthful; more than 50 percent of all West Africans are under the age of 18. The colonial migratory labor options that once took teeming Sahelian youth to the coastal plantations and cities have begun to fail in modern times (Cooper 1996). Under recession conditions, migrants no longer gain enough wealth to return to intermediate zone homes to assume the status of rural elders. Not all have villages to which to return. Some are the landless

descendants of former domestic slaves. Others have begun to succumb to HIV/AIDS. Many are now permanently hustling the city streets, locked into a vicious circle of poverty, petty crime, and drugs. Despairing of the future, some of these younger elements have been willing recruits (from the streets of Monrovia, Abidjan, and other coastal cities) to militia groups fighting the 1990s wars for diamonds and timber in rural Liberia and Sierra Leone. Embittered by their lack of social recognition they hit out at everything in their path, including those who seek refuge in the old agrarian compromise. This is the new 'masterless class,' in awe only of its own charismatic and demagogic leaders. Human rights mean little, and atrocity multiplies.

There is a widening recognition that West African youth is in crisis, but as yet less recognition that the crisis is not only one of crime, drugs, and weapons but also one of basic social values (Richards 1996). Youthful egalitarianism has slipped free of the restraints of a wider and more complex set of social compromises. Enclaved within bush war zones by the ferocity of their own violence, and by civilian fear and hatred, but still able to accumulate forest and mineral resources readily saleable overseas, the masterless egalitarians have a capacity for prolonged survival. The fission that normally ends egalitarian adventures may not come into play if armed enclave groups, lacking other exit options, find themselves locked in place by the unremitting hostility of the wider society. Perhaps the violence can be tamed only by a search for new compromises between the distinct, and currently disengaged, modes of social solidarity of youth and elders, men and women.

But some would say that youthful egalitarianism is not the only challenge to established social compromise in contemporary West Africa. The disengaged individualism of newly enriched and internationally footloose commercial elites also poses urgent problems. Yet despite current difficulties, the vibrancy and inventiveness of West African society and culture continues to impress. Social renewal in West Africa must reach beyond old solutions, and turn away from ineffectual inducements to value change—the casually proffered carrot-and-stick of the international community. Perhaps some of the urgently required thinking about complex social values, and new forms of social solidarity, will emerge from debate, fostered by some West African and African-American scholars and activists, about moral accountability for the evils of the slave trade.

Bibliography

Achebe C 1964 *Arrow of God*. Heinemann, London

Cohen A 1969 *Custom and Politics in Urban Africa*. Manchester University Press, Manchester, UK

Cooper F 1996 *Decolonization and African Society*. Cambridge University Press, Cambridge, UK

Crowder M 1968 *West Africa Under Colonial Rule*. Hutchinson, London

Dalby D 1977 *The Language Map of Africa*. International African Institute, London and the Adjacent Islands

Douglas M 1993 *In the Wilderness*. JSOT Press, Sheffield, UK

Hannerz U 1987 The world in creolization. *Africa* **57**: 546–59

Inikori J E (ed.) 1982 *Forced Migration*. Africana Pub. Co., New York

Kopytoff I (ed.) 1987 *The African Frontier*. Indiana University Press, Bloomington, IN

Last M 1967 *The Sokoto Caliphate*. Longmans, Harlow

Middleton J, Tait D (eds.) 1959 *Tribes Without Rulers*. Routledge & Paul, London

Richards P 1996 *Fighting for the Rain Forest*. James Currey for the International African Institute, Oxford, UK

Seddon D (ed.) 1978 *Relations of Production*. Frank Cass, London

Shaw R 1997 Production of witchcraft, witchraft as production: memory, modernity, and the slave trade in Sierra Leone. *American Ethnologist* **24**(4): 856–876

Todd L 1984 *Modern Englishes: Pidgins and Creoles*. Blackwell/Andre Deutsch, Oxford, UK

P. Richards

Western European Studies: Culture

The ontological and epistemological status of culture has received much attention in recent social science literature (Sewell 1999). This article assumes a parsimonious position on these debates. Culture as a concept has a cognitive and material dimension. As a form of cognition, culture is the realm of shared meanings that permits mutually informed social interaction within bounded spaces. This anthropological view of culture points to culture as a medium of social life and measure it in the 'thick description' of shared social practices (Geertz 1973). Culture as material object refers to expressive forms such as art, music, literature, architecture—more recently, state institutions and legal codes. Material culture is analytically distinct from but not unrelated to the anthropological concept of culture. This entry traces the development of both senses of culture in post-war European Studies. While individual nation states have diverse intellectual traditions that articulate themselves differently in cultural studies, this entry will focus on commonalties rather than differences.

1. Culture in the Shadows: West European Studies to the 1980s

Until the mid-1980s, culture as either an analytic frame or causal variable was conspicuously absent from the social science research on Europe. Apart from Elias' studies of court society and the civilizing process that linked modern notions of manners to the

formation of the nation state (Elias 1994) and Kantorowicz's study of the king's 'two bodies' (1957) (both of whom were 'rediscovered' in the 1980s), culture, for the most part, remained in the shadows of social science analysis. The work of British cultural historian Raymond Williams (1958) that followed in a tradition of 1930s European Marxist cultural analysis was a rare exception to the general trend. Williams' cultural and historical sociology was the precursor to what became known as British Cultural Studies in the 1970s (for example, Hoggart 1961).

Two dimensions of culture's transparency accounted for its absence in European Studies. First, intellectuals did not question Eurocentrism—the idea that Europe was the seat of civility and apex of modern civilization. Weber's 'Introduction' to *The Protestant Ethic and the Spirit of Capitalism* (1958) suggests the embeddedness of Eurocentrism. The hegemonic position of the European cultural heritage in the West rendered it moot as an analytic category among sociologists and political scientists. European culture with a capital 'C' was elite culture and left to the purview of art historians and literary critics.

Second, culture was absent, but traditions were not. There was an implicit assumption in social science that traditions, however under specified, played a role in European political life. The collective *mentalité* of European society—and social science—viewed culture and tradition as intrinsic to each other. Analysts rarely parsed this linkage and produced anthropological accounts that demonstrated that peasants and modern political practices were antithetical to each other. Banfield's (1967) study of a southern Italian village that ascribed its political backwardness to deeply rooted traditions is a classic in this genre. Customs and superstitions that had plagued them for hundreds of years trapped Banfield's Italian peasants in a premodern and unchanging landscape of political practices. Wylie's *Village in the Vaucluse*, (1974) a kinder study in a similar vein portrayed French peasants whiling away their time at boules and drinking at local cafés while high politics was conducted in a distant Parisian capital.

The assumptions that undergrid Eurocentrism and traditionalism rendered culture so transparent as to make it invisible. Moreover, these assumptions were fundamentally conservative. In the late 1970s and early 1980s, social history engaged the attention of first historians and later sociologists—usually influenced by Marxism (see the introductory summary in Bonnell and Hunt 1999). At first, the politically engaged scholars who emerged in the 1970s viewed culture as suspect—a position that changed as the decade progressed. In the new historical anthropology of Europe, peasants were subversive (Ginzberg 1980); religious events were riotous (Davis 1975); oppressed apprentices murdered cats to get control of their work situation (Darnton 1985); and communist comrades could also be Christian (Kertzer 1980).

The combination of cultural conservativism and Socialist politics were not the only entities that contributed to the lack of attention to culture as an analytic frame. Social scientists working in the positivist mode favored in the 1950s and 1960s considered culture as too imprecise a concept to operationalize. As a variable, culture appeared to lack explanatory power. Where culture was taken into account we have the classic studies of Almond and Verba (1989), who measured political culture in sharply defined variables all pointing to the fact that the more Protestant a group was, the more democratic they were likely to be (for summaries, see Berezin 1994, 1997b).

2. *Analysis Follows History: The 'Cultural Turn' in European Studies*

No matter what ideological position one espouses or analytic method one favors, content always defines culture, whereas trajectory defines history. By 1990, social history became the 'new cultural history' (Hunt 1989). Europe was in many instances the site of the emerging 'cultural turn' (Bonnell and Hunt 1999). Specific historical changes in the last quarter of the twentieth century contributed to the new salience of culture in analyses of European society.

Political scientists and historians frequently refer to the period between roughly 1950 and 1970 as the period of the 'post-war settlement.' That period was characterized on the political front by the consensus supported by the US of the containment of communism and on the economic front by the corporatist bargain that kept labor and business in accord. The European welfare state contained class conflict while valorizing class politics. European parties represented genuine ideological differences. A spirit of prosperity prevailed.

This bargained equilibrium with its cultural and ideological concomitants began to break down in the 1970s. The student revolts of the late 1960s and the period of terrorism, 'iron years' as the Italians labeled them, pointed to an emerging 'end of ideology' politics (Lumley 1990). By the 1980s, social science followed history. New Social Movements (NSMs) emerged that prioritized values over interests (i.e., peace, environment) and identity politics (i.e., gender, ethnicity) superceded class politics as traditional European parties began to fractionate (Cohen 1985).

The fall of the Berlin Wall and the collapse of communism in 1989 coupled with the commitment to European union signaled the end of the 'old' Europe. In academia, Eurocentrism gave way to multiculturalism (Joppke and Lukes 1999), nationalism to postnationalism. Postmodernity replaced modernity and the Enlightenment, the cornerstone of Eurocentricism, came into question (Wagner 1994). Inglehardt (1977) following in the research style of Almond and

Verba, documented 'postmaterialist' values. Social scientists turned to culture in an attempt to make sense of a society and politics that traditional modes of analysis could neither explain nor comprehend.

Transformation was not merely cultural and conceptual. Political and economic shifts fueled the retreat to identity and post-materialism. The 1970s and economic recession in Europe was the death knell of class collaboration. In the wake of European recessions, socialist technocratic parties emerged as well as out and out conservative parties such as Margaret Thatcher's neoliberalism. Traditional left and right parties began to lose their constituencies. Kertzer's (1996) analysis of the re-reorganization of the Italian Communist Party underscores the schizophrenia of changing European values. Formerly the strongest Communist Party in Western Europe, the party needed an infusion of new ideas in the early 1990s. But it could not decide upon its new name or the design of its flag. The Italian Communist Party finally agreed to change its name to the Democratic Party of the Left. Its flag incorporated 'postmaterialist' values by superimposing an environmentally correct tree over its traditional hammer and sickle—thus wedding green to red, new to old.

3. *Analytic Transformation*

Social and political change brought culture out of the shadows and into the full light of social scientific analysis. Identifiable analytic shifts occurred between 1980 and 2000. Among the multiple possibilities for categorizing these trends, this entry identifies four as salient: first, the move from class to identity; second, the move from connoisseurship to material culture; third, the new focus on culture and institutions; and lastly, an emerging interest in democracy, civil society and political culture. These shifts, presented here as discrete categories of analysis, in practice often shade into each other.

3.1 *Class to Identity*

Class was the initial venue that wed culture to political mobilization and social closure. Influenced by the early, more culturalist and consciousness-based writing of Karl Marx, scholars tried to understand why and how working class culture became a vehicle of class based political action. The seminal study in this vein was E. P. Thompson's *The Making of the English Working Class* (1963). Influenced by this work, other social historians began to focus on class cultures. Stedman-Jones (1983) focused on British languages of class; Sewell (1980) examined how a corporate idiom shaped the political culture of French artisans; and Calhoun (1982) argued that established traditions forged solidaristic social ties within local communities that generated class cohesion among eighteenth-century British workers. Hebdige (1979) showed how British youth culture was a form of class identity and social rebellion.

Bourdieu (1984) redirected the focus of class studies from Karl Marx to Max Weber by incorporating the idea of status. His insight that capital could be cultural and social as well as purely economic launched a genre of studies on social boundary making. Bourdieu (1977) argued that habitus shaped the perceptions of a group to include and exclude based on cultural and social cues. Lamont (1992) built on this work by studying middle class, and later working class, men in the United States and in France to test empirically how boundaries and the conception of 'one of us' were formed.

Sexuality, gender, race and nationality have become the basis of identity studies on Europe. Foucault's (1990) three-volume work on the history of sexuality that questioned the entire concept paved the way for studies of the body (Laqueur 1990). Hunt's (1992) monograph on the 'family romance' of the French Revolution linked Freudian analysis of sexuality to issues of gender and politics. Borneman's (1992) study of kinship in the former East and West Berlin linked issues of gender and national identity. Soysal's (1994) discussion of post-nationalism argued that citizenship, a category of nationality, is a malleable identity category that transcends the materiality of space.

With the exception of British media studies, the European ideology of connoisseurship relegated studies of art, music, film, architecture, and literature to the relevant humanities disciplines. Since the mid-1980s, social scientists have reinvented art as material culture and produced a broad range of studies. For example, Griswold (1986) studied Renaissance revivals on the British stage. Biernacki (1995) studied the cultural significance of divergent factory regimes among woolen workers in Britain and Germany. Politics and cultural production is an emerging area of interest. Mukerji (1997) wrote on the relation between formal gardens and French state building.

Consumption as an economic process that intersects with cultural production has informed a whole genre of studies of Americanization and globalization in European studies. DeGrazia (1989) studied how American film culture defined and penetrated European films in the 1920s and 1930s. French gastronomy (Ferguson 1998) and furniture design (Auslander 1996) were intimately connected to issues of class autonomy and national identity and citizenship. Consumption is now global and studies document the invasion of foreign capital and cultural products on European soil. McDonalds and the culture of fast food are a frequent target. A recent special edition of *Le Monde Diplomatique* entitled 'L'Amerique dans les Tetes' ('America in the Head') provides an example of the concern with the Americanization/globalization of European culture.

3.2 Institutional Studies of Culture

Institutional studies of culture that focus upon how the state mobilizes cultural and political institutions, such as schools, the military, language consolidation, nationality codes, as well as symbolic entities such as art, music, and architecture to create identities has arguably the longest trajectory in European cultural studies. Weber's (1976) *Peasants Into Frenchman* is an early work in this genre that provides a history of how the French state mobilized local cultural institutions to create a state centered French national culture. Mosse (1991) looked at the phenomenon of what he termed the 'nationalization of the masses' to document how territories that had become nation states became bounded cultural political spaces in the early twentieth century. In 1983, Hobsbawm's and Ranger's influential anthology *Inventing Traditions* emphasized the fact that all national cultures were the products of carefully designed state rituals and symbolic practices—beginning with the British monarchy.

The intersection of state institutions and cultural practices dominated European political cultural studies from the mid-1980s to the present (Steinmetz 1999). In the wake of Ozouf's (1988) influential book on rituals and the French Revolution, Hunt (1984) explored the link between the 'poetics of power' and the 'sociology of politics' to describe the 'unexpected invention of [French] revolutionary politics' (p. 3). Scholars followed Hunt's lead—turning to, for example, Nazi Germany (Fritzsche 1998) and fascist Italy (Berezin 1997a). Scholars who focus on membership issues in nation states demonstrated how states institutionalized political culture in diverse approaches to citizenship law (Brubaker 1992). In a series of articles, Laitin (1989) and Laitin and Sole (1994) argued that language consolidation in European nation states was analogous to a tipping game where it was rational to learn a national language if one wished to participate in consolidated national language markets.

3.3 Democracy, Civil Society and Political Culture

The studies of democracy and culture, a relatively new preoccupation, divide into two groups: first, scholars influenced by Jurgen Habermas' theories of the public sphere and those influenced by the theory of social capital. Habermas' *The Structural Transformation of the Public Sphere* (1989) posited that the origins of democratic civil society lay in the exchange of information and ideas in public space from eighteenth-century coffee houses to modern newspapers. Multiple groups of citizens could meet and discuss and arrive at democratic consensus.

This new version of political culture was heavily dependent on the valorization of public discourse and was particularly appealing to social scientists contemplating the reconstruction of civil society in the former Eastern Europe. In an influential article that rethought T. H. Marshall's (1964) conception of citizenship, Somers (1993) argues that citizenship rights rather than attributes of individuals are deeply embedded in a set of institutional practices that are highly dependent upon the permeability of local public spheres. In contrast to much of the theoretical literature that Habermas' notion of the public sphere has generated, Zaret (2000) provides the first extended empirical study that examines the modes of communication available to a public in early modern England.

Putnam's *Making Democracy Work* (1992) building on Alexis de Tocqueville's theory of democratic association introduced social capital to studies of political culture. Putnam studied the success of the Italian regional governments in Italy over a 20-year period. He attributed the more efficient governments found in the north, in contrast to the south, to a political culture of association that produced a high level of trust and cooperation among participants in governments. He terms this talent for government 'social capital'—meaning that a citizenry that participated in multiple networks of associations developed an elevated capacity for acting democratically. To date, Putnam's work has generated volumes of research on American democracy, but little empirical work on Europe itself. This will likely change as the results of trans-European collaborative studies on social capital become available.

4. New Turns

The analytic trends that I have outlined are in progress and will continue to occupy researchers for years to come. They raise questions for which the answers are decidedly not in yet. But what is new on the horizon? In addition to the areas I have outlined, four additional themes are just beginning to occupy scholars. First, European identity will begin to supersede national identities as a focus of research. In part, this will emerge as a response to the European Commission's push to create not only a European market but also a European space (Kastoryano 1998). Some of these attempts, as the search for a common European history that does not focus on war, will be strained. But the temporal effects of transborder life will no doubt contribute to the formation of a European consciousness. The second area of interest will no doubt be the new technologies of communication, particularly the internet. At the time of writing, all political parties in Europe have websites. The ramifications and efficacy of a virtual public sphere is a whole terrain that has not been explored and will no doubt take off as the internet had done. A third emerging area is the field of Europe/America comparisons. This is relatively novel, as in the past America was viewed as 'exceptional.' As Europe becomes more consumer oriented, and as it continues to deal with the in-

corporation of immigrant groups, America becomes a more likely comparative venue than it has been in the past.

The last area is the paradoxically growing importance of 'memory' studies (Nora 1996–8). The European absorption with its past, particularly the period of Nazism, genocide, fascism and collaboration, has taken on a surprising urgency as the continent struggles to shape a vision *of* and program *for* its future. The tendency began in 1985 in Germany with the public controversy known as the historians' conflict (*Historikerstreit*) (Baldwin 1990). The concern with memory has continued as scholars attempt to come to terms with Vichy France (Rousso 1991) and Nazi Germany (Maier 1988). An offshoot of the obsession with the past has been the emergence of a field of 'compensation' studies that examines whether nation states can come to terms with the horrors of state sponsored genocide in monetary terms.

In short, to write about Europe in the year 2001, is to write about a continent in the midst of vast political, social, and technological transformation. Underneath this truism, lies the observation that cultural studies thrive in times of upheaval and uncertainty. The last 20 years of European studies suggest that political and social problems that standard notions of modernity and rationality cannot explain forces culture to the forefront of everyone's research agenda. There are no signs that this trend is likely to reverse itself at any time in the near future.

Bibliography

Almond G, Verba S 1989 *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Sage, Newbury Park, CA

Auslander L 1996 *Taste and Power*. University of California Press, Berkeley, CA

Baldwin P (ed.) 1990 *Reworking the Past*. Beacon, Boston, MA

Banfield E 1967 *The Moral Basis of a Backward Society*. Free Press, New York

Berezin M 1994 Fissured terrain: Methodological approaches and research styles in culture and politics. In: Crane D (ed.) *The Sociology of Culture*. Basil Blackwell, Oxford, UK

Berezin M 1997a *Making the Fascist Self: The Political Culture of Inter-war Italy*. Cornell University Press, Ithaca, NY

Berezin M 1997b Politics and culture: A less fissured terrain. *Annual Review of Sociology* **23**: 361–83

Biernacki R 1995 *The Fabrication of Labor*. University of California Press, Berkeley, CA

Bonnell V E, Hunt L (eds.) 1999 *Beyond the Cultural Turn*. University of California Press, Berkeley, CA

Borneman J 1992 Belonging in the Two Berlins. Cambridge University Press, Cambridge, UK

Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK

Bourdieu P 1984 *Distinction*. Harvard University Press, Cambridge, MA

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Calhoun C 1982 *The Question of Class Struggle*. University of Chicago Press, Chicago, IL

Cohen J L 1985 Strategy or identity: new theoretical paradigms and contemporary social movements. *Social Research* **52**: 663–716

Darnton R 1985 *The Great Cat Massacre and Other Episodes in French Cultural History*. Vintage, New York

Davis N Z 1975 The rites of violence. In *Society and Culture in Early Modern France*. Stanford University Press, Stanford, CA

De Grazia V 1989 Mass culture and sovereignty: The American challenge to European cinemas, 1920–1960. *Journal of Modern History* **61**: 53–87

Elias N 1994 *The Civilizing Process*. Basil Blackwell, Oxford, UK

Ferguson P P 1998 A cultural field in the making: Gastronomy in 19th century France. *American Journal of Sociology* **104**: 597–641

Foucault M 1990 *The History of Sexuality*. Vintage, New York

Fritzsche P 1998 *Germans into Nazis*. Harvard University Press, Cambridge, MA

Geertz C 1973 *Thick Description: Toward an Interpretive Theory of Culture. The Interpretation of Cultures*. Basic Books, New York

Ginzburg C 1980 *The Cheese and the Worms*. Johns Hopkins University Press, Baltimore, MD

Griswold, W 1986 *Renaissance Revivals*. University of Chicago Press, Chicago, IL

Habermas J 1989 *The Structural Transformation of the Public Sphere*. MIT Press, Cambridge, MA

Hebdige D 1979 *Subculture*. Methuen, London

Hobsbawm E, Ranger T (eds.) 1983 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK

Hoggart R 1961 *The Uses of Literacy*. Beacon, Boston, MA

Hunt L 1984 *Politics, Culture and Class in the French Revolution*. University of California Press, Berkeley, CA

Hunt L (ed.) 1989 *The New Cultural History*. University of California Press, Berkeley, CA

Hunt L 1992 *The Family Romance of the French Revolution*. University of California Press, Berkeley, CA

Inglehart R 1977 *The Silent Revolution*. Princeton University Press, Princeton, NJ

Joppke C, Lukes S 1999 *Multicultural Questions*. Oxford University Press, New York

Kantorowicz E H 1957 *The King's Two Bodies*. Princeton University Press, Princeton, NJ

Kastoryano R 1998 *Quelle Identite pour L'Europe?* Presses de Sciences Po, Paris

Kertzer D I 1980 *Comrades and Christians: Religion and Political Struggle in Communist Italy*. Cambridge University Press, Cambridge, UK

Kertzer D I 1996 *Politics and Symbols: The Italian Communist Party and the Fall of Communism*. Yale University Press, New Haven, CT

Laitin D D 1989 Linguistic revival: Politics and culture in Catalonia. Comparative Studies in Society and History **31**: 297–317

Laitin D D, Sole C 1994 Language and the construction of states: The case of Catalonia in Spain. *Politics and Society* **22**: 5–29

Lamont M 1992 *Money, Morals and Manners*. University of Chicago Press, Chicago, IL

Laqueur T 1990 *Making Sex: Body and Gender from the Greeks to Freud*. Harvard University Press, Cambridge, MA

Lumley R 1990 *States of Emergency: Cultures of Revolt in Italy from 1968 to 1978*. Verso, New York
Maier C S 1988 *The Unmasterable Past: History, Holocaust and National Identity*. Harvard University Press, Cambridge, MA
Marshall T H 1964 Citizenship and social class. In *Class, Citizenship and Social Development: Essays by T H Marshall*. University of Chicago Press, Chicag, IL
Mosse G 1991 *The Nationalization of the Masses*. Cornell University Press, Ithaca, NY
Mukerji C 1997 *Territorial Ambitions and the Gardens of Versailles*. Cambridge University Press, Cambridge, UK
Nora P 1996–8 *Realms of Memory. 3 vols*. Columbia University Press, New York
Ozouf M 1988 *Festivals and the French Revolution*. Harvard University Press, Cambridge, MA
Putnam R 1992 *Making Democracy Work*. Princeton University Press Princeton, NJ
Rousso H 1991 *The Vichy Syndrome. Historical Memory in France since 1944*. Harvard University Press, Cambridge, MA
Sewell W H 1980 *Work and Revolution in France*. Cambridge University Press, New York
Sewell W H 1999 The concept(s) of culture. In: Bonnell V E, Hunt L (eds.) *Beyond the Cultural Turn*. University of California Press, Berkeley, CA
Somers M R 1993 Law, community, and political culture in the transition to democracy. *American Sociological Review*. **58**: 587–620
Soysal Y N 1994 *Limits of Citizenship*. University of Chicago Press, Chicago, IL
Stedman-Jones G 1983 *Languages of Class*. Cambridge University Press, Cambridge, UK
Steinmetz G (ed.) 1999 *State/Culture*. Cornell University Press, Ithaca, NY
Thompson E P 1963 *The Making of the English Working Class*. Vintage, New York
Wagner P 1994 *A Sociology of Modernity*. Routledge, New York
Weber E 1976 *Peasants Into Frenchmen*. Stanford University Press, Stanford, CA
Weber M 1958 *The Protestant Ethic and the Spirit of Capitalism*. Charles Scribner's Sons, New York
Williams R 1958 *Culture and Society 1780–1950*. Chatto & Windus, London
Wylie L 1974 *Village in the Vaucluse*. Harvard University Press, Cambridge, MA
Zaret D 2000 *Origins of Democratic Culture*. Princeton University Press, Princeton, NJ

M. Berezin

Western European Studies: Environment

The study of social and political aspects of the natural environment in Western Europe has been a major growth area since the early 1970s. All social science, disciplines, in particular history, geography, economics, psychology, sociology, and political science, have seen the development of major environmental research fields, although the relative importance of each discipline has varied between countries and over time. Cross-national variances are generally the result of specific disciplinary traditions; variations over time are mainly the outcome of changes in the importance of particular environmental discourses.

There are important differences in the conception of 'the environment' between different social science disciplines. While the interaction between human society and the natural environment is the most common element, many approaches, in particular within geography and environmental psychology, also include the entire physical (e.g., the built environment) and social environment.

The terms used to delineate this field of activity have also reflected major social and political changes. The label 'environmental' to define an area of policy or politics first came into use in the 1960s in the USA and subsequently also became popular in Western Europe. Radical critiques of environmentalism led to increased references to 'ecology,' 'ecological,' and less frequently, 'ecologism.' Since the success of the German Green Party on an anti-nuclear, ecological platform in the early 1980s, the term 'green' has come to be used in a wide variety of settings, in the 1990s referring to virtually any activity that had previously been labeled 'environmental.'

The great variety and scope of social science work in this field escapes easy compartmentalization and makes it difficult to do justice to all approaches involved. There are four broad areas of research that are crucial to the agenda of environmental sociology and political science, but that can also be used to discuss the contribution of other disciplines: environmental philosophy and political theory, environmental attitudes and behavior, social movements and green parties, and environmental policy.

1. *Environmental Philosophy and Political Theory*

Political theorists took a relatively long time to take up the challenge of modern environmental politics. Up to the late 1970s, 'ecological thought' had barely been recognized as a subdiscipline of political thought and philosophy. Ecological 'thinkers' mainly came from outside academia, or at least from outside the realm of political theory and philosophy. However, this period of relative neglect was followed by a major expansion in the 1990s when ecological political thought became a major area of academic activity in Western Europe, particularly in the UK and The Netherlands. The aim was to establish a theoretical basis for ecologism as a separate political theory with a distinct identity rather than as an extension of established traditions such as liberalism, anarchism, or socialism. One of the major inspirations was the work on deep ecology and ecocentrism that had emanated from North America and Australia. Among its important European antecedents were the writings of the Norwegian philosopher Arne Naess (Dobson 2000).

The main challengers of ecocentric political theory in Western Europe are various approaches broadly associated with a libertarian vision of socialism; anarchism, and more recently, liberalism also play a role. Following on from the politics of the student movement, a revival of interest in Marxist thinking in the 1970s also extended to ecological questions. The ideas of the Frankfurt School, anarchist approaches as well as more traditional Marxist views dominated continental theoretical debates in the 1970s and 1980s. Continental theoretical discourses gave more prominence to sociological theories rather than environmental ethics. The work of Ulrich Beck and his theory of the *Risk Society* (Beck 1986) was particularly influential. However, the resonance of Niklas Luhmann's (1986) equally sweeping, if rather more pessimistic, attempt to employ systems theory to devise a theory of ecological communication was mainly limited to the German-speaking academic world.

Apart from the formulation of a green theory informing current political practice, the history of ecological thought is a major research area. West European historians, philosophers, economists, and sociologists have devoted considerable attention to the study of environmental aspects of the work of philosophers, scientists, economists, social theorists, and other writers preceding the arrival of 'modern' ecological thinking (Pepper 1984, Wall 1994). In particular, from the perspective of the 'sociology of science,' the history of ecology as a scientific discipline has been of great interest, involving also the role of scientific 'ecologists' in the definition of research and political agendas. The link between 'ecology' and other political ideologies has been another major focus of debate. Among the most contentious topics is the analysis of the history and current development of ecological thought in the context of authoritarian political traditions, including the ideology of the German Nazis (Bramwell 1989, Ferry 1992).

However, rather underdeveloped is the political sociology of environmental ideology and knowledge, analyzing, for example, why particular theories or knowledge claims have greater resonance than others in certain countries and specific historical periods. Among the few studies addressing these issues is Jamison et al.'s (1990) analysis of the changing 'knowledge interests' of environmental groups.

2. *Environmental Attitudes and Behavior*

Environmental attitude research started as a serious academic activity in the 1970s. The first wave of studies was mainly policy oriented, intended to measure the degree of environmental concern and support for environmental policies that had started to be formulated and implemented in most European countries in the late 1960s and early 1970s. Sociologists were at first mainly interested in the socio-demographic background and political orientation of 'environmentalists.' Most of the political analysis of environmental concern in the late 1970s and 1980s was inspired by Ronald Inglehart's theory of 'postmaterialist value change.' This analysis interpreted the rise of environmental concern as part of a far broader 'silent revolution' that was producing a new political agenda dominated by 'left-libertarian' issues (Lowe and Rüdig 1986).

The 1980s and early 1990s saw a major expansion of environmental attitude research. With the rise of global environmental issues and green consumerism, research priorities focused on the multidimensional nature of public attitudes, the role of 'environmental knowledge,' the conditions for behavioral change to save the environment, and the acceptance of specific environmental policies. In particular, the perceived gap between expressed environmental concern and the reluctance of the public to engage more deeply in environmental behavior or to accept stronger environmental policy measures, such as higher energy prices, became a key topic. In theoretical terms, the role of post-materialism as a guiding concern waned, and other approaches, in particular rational choice theory, became more important. More attention also began to be paid to public discourses and personal narratives of environmental questions, with focus groups and qualitative studies becoming a more popular research tool (Diekmann and Jaeger 1996, Macnaghten and Urry 1998).

Research about environmental attitudes and behavior has also played a major role in other social science disciplines. Within environmental psychology, a wholly separate research tradition on environmental attitudes was established, grounded in a more holistic approach to human interaction with the socio-physical environment, with particularly interesting contributions to the analysis of the perception of different environmental problems and the barriers of environmental consumer behavior (Pol 1993).

A key research focus for sociologists, psychologists, and other social science disciplines has been the concept of risk. 'Risk perception' and 'public attitudes to technology' became major research topics in the wake of environmental protest against nuclear power and other 'big' technologies, particularly in Germany. Following on from the analysis of public perception of risk was the development of 'risk communication' strategies, aimed at increasing public acceptance of specific technologies (Renn and Zwick 1997).

Also in the economic sciences, the analysis of environmental attitudes and behavior has played an important role, in particular in cost–benefit analysis based on the allocation of monetary values to environmental goods. This is done with surveys asking hypothetical questions about the public's willingness to pay for particular goods, as well as the analysis of existing behavior to elicit 'revealed preferences' about

trade-offs between environmental and other goods. Such contingent valuation methods became particularly influential in UK policy discourses in the 1990s, but they also attracted strong criticism on ethical and methodological grounds (Foster 1997).

Social science research on environmental attitudes and behavior has attracted other controversies. Research on risk perception and risk communication has been attacked for its instrumentalization to manipulate public opinion to increase the public acceptance of particular technologies. The rise of relativist, 'social constructionist' approaches in the environmental social sciences has offered contradictory opportunities to use research results to influence political debate. For example, a 'cultural theory' of environmentalism which sees environmental concern as a response to cultural change rather than to 'real' environmental problems could be seen as undermining the legitimacy of environmental groups. Other approaches, particularly from the sociology of science, have targeted the claims of science in legitimating public policy, seeking to establish the credibility of lay persons facing environmental threats to make relevant knowledge claims (Lash et al. 1996, Redclift and Benton 1994).

3. *Social Movements and Green Parties*

Environmental concern and green consumer behavior do not directly engage the individual in the political process. This is different where people support particular pressure groups and parties, or join in protest activity, to campaign for specific environmental policies or for a general green worldview. The analysis of environmental politics thus includes environmental political behavior at the individual level, but it also focuses on the behavior of groups and parties.

The politicization of the environment in Western Europe was chiefly the role of a new generation of protest movements that started to emerge in force in the 1970s. This new phenomenon attracted the academic attention of a fairly wide range of social scientists. At first, descriptive case studies of specific incidents of protest or environmental conflict were the main approach. But by the late 1970s and early 1980s, the notion of 'new social movements' had emerged as the dominant paradigm. The 1980s and 1990s saw a very substantial development of this field. The cross-national comparison of environmental social movements became a major focus; the reception of key elements of US social movement theory, in particular the 'Resource Mobilization' approach, led to more sophisticated research designs and theoretically informed empirical studies.

Methodologically, the study of protest events became a very influential approach, particularly for cross-national comparisons. Another key feature of empirical research was the study of individual movement participants and group members, their social backgrounds and motivations. Studies of environmental groups as organizations, based on national and cross-national surveys of such groups, focused *inter alia* on network formation and interrelationships with other political actors (Dalton and Kuechler 1990, Della Porta and Diani 1999, Rootes 1999, Rucht 1991).

One of the most significant aspects of the politicization of the environment is the rise of ecological and green parties. After party formation in the 1970s and 1980s, such parties entered parliament in most West European countries during the 1980s and 1990s (the main exceptions being Norway, Denmark, and Spain); by the late 1990s, green parties had entered national government in Italy, Finland, France, Germany, and Belgium.

Green parties have attracted substantial research efforts, both in the form of major national case studies and cross-national comparisons. Much of the cross-national analysis has focused on the influence of aggregate factors such as electoral systems and economic conditions on the green parties' development. There is also a wealth of empirical national and cross-national studies on green ideology, the internal structure of green parties, their members and activists, and green voting (Delwit and De Waele 1999, Kitschelt 1989, Richardson and Rootes 1995).

4. *Environmental Policy*

The social sciences continue to play a fairly marginal role in the general academic literature on environmental policy, with approaches from the natural sciences and technical disciplines dominating the field. Within the social sciences, most of the publications addressing questions of environmental policy come from geography, economics, and the legal professions, while sociology and political science play a more minor role.

In Western Europe, environmental policy started to become an important research topic in the early 1970s. The policy optimism of the early years was often disappointed, and one important research theme was the question of the implementation deficit that prevented policies being as successful as intended. Also popular were cross-national comparisons between 'leaders' and 'laggards' in environmental policy innovation, often involving the USA and one or more European countries, or, somewhat later, the UK and Germany.

The 1970s and 1980s also saw increasing environmental policy activism within the European Union (EU), and the study of policy development and implementation of EU policy became a major growth area. With the rise of global environmental issues and the creation of international environmental regimes in the 1980s and 1990s, the international level became an important focus of policy research, with international

relations approaches playing a more influential role in the literature.

The 1980s and 1990s brought a major shift in the debate about policy instruments, away from traditional 'command-and-control' regulation to new economic instruments. Policy discourses became increasingly dominated by economic ideas. With the 1992 Earth Summit, the idea of 'sustainable development' was placed firmly on the policy agenda, and many European countries played a lead role in furthering this debate at local, national, and EU levels (Rüdig 1999).

Environmental policy research in Western Europe is well established but there remain many key challenges to be addressed. One of the them is the lack of interdisciplinary work, with public policy specialists, international relations experts, sociologists of science, and many other social science disciplines researching in parallel without taking much notice of each other. Some of the shortcomings of research are related directly to this failure to break down disciplinary boundaries. One such area that has received comparatively little attention is issue-making and agenda-building processes. The application of discourse analysis to public policy research is likely to play an important role in future research (Fischer and Hajer 1999), as are studies of social movements. The role of science and the media could make an important contribution toward improving the understanding of the formation of environmental policy agendas.

With the increasing role of European and international environmental policy processes, the interlinkages between the international, EU, and national policy processes is a major research challenge. In the past, the study of international relations, EU politics, and national public policies has been addressed by separate academic communities; in the environmental field, major attempts have been made to combine international and national processes—for example, in research on the effectiveness of international environmental regimes. In Western Europe, in particular, environmental policy is a 'multilevel' process, and the study of the interlinkages between EU and national as well as subnational policy processes has become a major research area (Weale et al. 2000).

Finally, the question of the impact of environmental policy has to be addressed. In Western Europe, there is a large body of empirical material on differences in policy approaches in different European countries (Hanf and Jansen 1998, Jänicke and Weidner 1997). Comparative environmental policy impact studies in Western Europe suggest that economically successful nations also had successful environmental policies. The only institutional element associated with environmental policy success is a general feature of economic policy making—neo-corporatism. There is, however, no firm evidence so far on the relative effectiveness of particular environmental institutions and policy instruments (Rüdig 1999).

5. *Conclusions*

Social science research on environmental issues is thus well established in Western Europe. In terms of the wider political discourse, environmental economics and other 'policy' oriented approaches have had the most impact. Within a purely academic ambit, green political theory has been the most outstanding success story. Environmental sociology and political science have established very strong research traditions, but have tended to struggle to be recognized as 'relevant' knowledge producers by environmental activists and policy makers alike. In an area defined increasingly by highly technical, 'scientific' policy discourses, the social sciences face a stark choice between either leaving the field to other disciplines, or successfully challenging the legitimacy of rival knowledge producers, thus establishing the reflection about the social and political premises of environmental thought and action as a key element in the protection of environmental public interests.

See also: Ecological Economics; Ecology, Deep; Environmental Economics; Environmental Policy; Environmental Psychology: Overview; Environmental Risk and Hazards; Environmentalism: Philosophical Aspects; Environmentalism, Politics of; Environmentalism: Preservation and Conservation; Green Parties; Social Movements: Environmental Movements

Bibliography

Beck U 1986 *Risikogesellschaft: Auf dem Weg in eine andere Moderne*. Suhrkamp, Frankfurt am Main, Germany [1992 *Risk Society: Towards a New Modernity*. Sage, London]

Bramwell A 1989 *Ecology in the Twentieth Century: A History*. Yale University Press, New Haven, CT

Dalton R, Kuechler M (eds.) 1990 *Challenging the Political Order: New Social and Political Movements in Western Europe*. Polity Press, Cambridge, UK

Della Porta D, Diani M 1999 *Social Movements: An Introduction*. Blackwell, Oxford, UK

Delwit P, De Waele J-M (eds.) 1999 *Les Partis verts en Europe*. Éditions Complexe, Brussels, Belgium

Diekmann A, Jaeger C C (eds.) 1996 *Umweltsoziologie*. Westdeutscher Verlag, Opladen, Germany

Dobson A 2000 *Green Political Thought*, 3rd edn. Routledge, London

Ferry L 1992 *Le Nouveau Ordre Écologique: L'Arbre, l'Animal et l'Homme*. Bernard Grasset, Paris [1995 *The New Ecological Order*. University of Chicago Press, Chicago, IL]

Fischer F, Hajer M A (eds.) 1999 *Living with Nature: Environmental Politics as Cultural Discourse*. Oxford University Press, Oxford, UK

Foster J (ed.) 1997 *Valuing Nature? Economics, Ethics and Environment*. Routledge, London

Hanf K, Jansen A-I (eds.) 1998 *Governance and Environment in Western Europe: Politics, Policy and Administration*. Longman, Harlow, UK

Jamison A, Eyerman R, Cramer J, Læssøe J 1990 *The Making of the New Environmental Consciousness: A Comparative Study of the Environmental Movements in Sweden, Denmark and the Netherlands*. Edinburgh University Press, Edinburgh, UK
Jänicke M, Weidner H (eds.) 1997 *National Environmental Policies: A Comparative Study of Capacity Building*. Springer, Berlin
Kitschelt H 1989 *The Logics of Party Formation: Ecological Parties in Belgium and West Germany*. Cornell University Press, Ithaca, NY
Lash S, Szerszynski B, Wynne B (eds.) 1996 *Risk, Environment and Modernity: Towards a New Ecology*. Sage, London
Lowe P, Rüdig W 1986 Political ecology and the social sciences: The state of the art. *British Journal of Political Science* **16**: 513–55
Luhmann N 1986 *Ökologische Kommunikation*. Westdeutscher Verlag, Opladen, Germany [1989 *Ecological Communication*. Polity Press, Cambridge, UK]
Macnaghten P, Urry J 1998 *Contested Natures*. Sage, London
Pepper D 1984 *The Roots of Modern Environmentalism*. Croom Helm, London
Pol E 1993 *Environmental Psychology in Europe: From Architectural Psychology to Green Psychology*. Avebury, Aldershot, UK
Renn O, Zwick M M 1997 *Risiko- und Technikakzeptanz*. Springer, Berlin
Redclift M, Benton T (eds.) 1994 *Social Theory and the Global Environment*. Routledge, London
Richardson D, Rootes C (eds.) 1995 *The Green Challenge: The Development of Green Parties in Europe*. Routledge, London
Rootes C (ed.) 1999 *Environmental Movements: Local, National, Global*. Frank Cass, Ilford, UK
Rucht D (ed.) 1991 *Research on Social Movements: The State of the Art in Western Europe and the USA*. Westview Press, Boulder, CO
Rüdig W (ed.) 1999 *Environmental Policy*. Edward Elgar, Cheltenham, UK, 2 Vols
Wall D (ed.) 1994 *Green History*. Routledge, London
Weale A, Pridham G, Cini M, Konstadakopulos D, Porter M, Flynn B 2000 *Environmental Governance in Europe: An Ever Closer Ecological Union?* Oxford University Press, Oxford, UK

W. Rüdig

Western European Studies: Gender and History

Gender and history have combined in the last three decades to produce new understandings of how events, social structures, and fundamental elements of culture operate in Western Europe. Historical analysis has come to recognize gender as relational systems of power in which both men and women operate and to which both contribute as agents and members of larger cultural constructs. The principle developments in gender and history emerged as social history concerns focused on women gave way to post-structuralist and postmodern critiques of historical practice. The linguistic or cultural turn has shaped the premises and practices of gender and history so as to reveal how gender saturates categories of culture once presumptively believed to be scientifically neutral.

Rooted initially in the women's movements in America and Western Europe as they emerged out of the new social movements of the 1960s, gender history began as women's history. Most early explorations of women's historical experience operated within a social history paradigm, seeking out neglected materials to recuperate lives long deemed irrelevant to traditional modes of historical research. While some advocated pursuing women's history in isolation as a political choice highlighting the patriarchal qualities of apparently objective history, others argued for recuperating the causes and frameworks of female erasure in the context of larger historical narratives (see Bock 1989 for a consideration of both positions). The notion that women were best analyzed in the wider cultural context came to predominate, in part because of epistemological concerns within history as a discipline.

1. *Fundaments of Gendered Discourse*

As social relations are carried out in language and culture, gender and history have been informed by post-structuralist understandings of language and postmodern analyses of culture. While these overlap significantly, particularly in seeing culture as symbolic and expressed in language, the emphases are distinct. Theorists and historians who have studied language argue that attention must be paid to the historical specificity of linguistic constructions and to the general gender inflections language encodes. The questioning of a tacit acceptance that language is 'neutral' revealed how all communication is endowed with the power of social custom and historical practice. The 'cultural turn' similarly challenges the neutrality of older explanatory models, positing that all inquiry must admit that culture constructs meaning. Culture is the primary object of inquiry in this formulation, rather than language.

1.1 *Taking the 'Cultural Turn'*

The 'new cultural history' emerged from dissatisfaction with the limitations of social history and engagement with post-structuralist critiques. Historical analysis revealed huge variations in social categories over time and place. Instead of a coherent understanding of the social in the past, the picture just became more and more complex. Post-structuralists like Jacques Derrida and Michel Foucault challenged the tenets of scientific truth and objective reality posited by structuralism. Historians accordingly turned to culture as an explanatory device. While 'culture' is a slippery term, new cultural history in-

cluded concerns about the methodologies of historical practice, an epistemological skepticism with respect to narrative truth claims, and an interest in culture as a symbolic or semiotic system (Bonnell and Hunt 1999). The cultural turn opened up new ways to examine problems and issues related to gender as an artifact of culture and a determinant of symbolic configurations.

1.2 Gender, Language, and Historical Analysis

Post-structuralist analysis contributed to an awareness of language as constructed by context, while post-structuralist feminism articulated how gender difference is formulated in language rather than in the referent, and materiality is produced by linguistic formations. Seemingly 'natural' things (like bodies) are actually complex constructions. Gender is created by interactions of individuals, institutions, and social norms. The repetitive quality of perceived sexual difference is produced by discreet, contingent experiences. These experiences are inscribed in language and shape notions of sexual difference. For instance, the term 'biology,' in describing sexual difference, is taken to be neutral, scientific fact. But 'biology' has a history which belies its apparent neutrality. It was an early nineteenth-century term to describe sexual differences in early 'nature vs. nurture' debates concerning human development. 'Biology' in some circles was utilized to normalize gender bias: women's differences were in need of explanation. The persistent association of men with culture and women with biology points to the power embedded in language (Bock 1989).

Analysis of the discursive constructs of gender in historical terms has been formulated in the context of postmodern theories of language and subject formation. Luce Irigaray (1985) has attempted to rewrite sexual difference without the binary of masculine–feminine, arguing for a 'labial politics' which privileges female embodiment as a way of revaluing the binary and hierarchical notions of gender. Moi (1985) has expressed concern that Irigaray's position is essentialist and ahistorical, but Irigaray's poetic rendering of the problem has inspired interrogations of the relational positioning of masculine and feminine values in history.

Anglo-American theorists have generally been more materialist, but efforts have been made to utilize French linguistic deconstruction in historical analysis. Joan Scott (1988) argues that gender as formulated in language can be analyzed by drawing on deconstruction to dismantle hierarchical binaries of language in historical settings. Social relationships based on notions of sexual difference include within them power dynamics, the experience of which can be traced over time. Expressions of gender and power occur in cultural symbols, in the conceptualization of institutions, in kinship systems, and in subjective identity. Gender becomes a way to understand how society has been organized around the power differentials embedded in notions of difference.

The focus on discursive formations to some has risked the effacement of agency and the dispersal of explanatory models of historical change. Prompted by such fears, many historians have self-consciously paid attention to issues of agency. Concerns remain that the role of gender in producing larger historical change is not apparent, and efforts continue to articulate causal relationships between expressions of gender in culture or language and events or persons.

2. Gender and Identity Formation

That gender historically has been a component of identity seems difficult to dispute. In the twentieth century, identity has been located largely in psychosexual development, while other components of identity (work, family, religious affiliation) are seen as cultural overlays. Whether through differentiating the subject through knowledge of genital difference (Sigmund Freud) or through language in the mirror stage (Jacques Lacan), gender difference is considered foundational to notions of identity.

2.1 Histories of Sexualities

In the psychosexual models of identity, biological sex and gender constitute identity even though historians have noted that sexuality has not always been regarded as foundational. Much historical work has been devoted to understanding the meanings attached to sexuality and sexual behaviors in the past as key areas in which gender operates. The resulting explorations have at once brought into question the psychosexual account of the meaning of sexuality and introduced new paradigms in which sexuality is seen as experienced in relation to society and power.

The history of sexuality reflects the problem of how to treat the sex/gender system historically. The separation of sex from gender allows for a more precise rendering of the effects of embedded assumptions. 'Sex' as anatomical difference between male and female bodies is theoretically distinct from 'gender' as the social processes of becoming male or female. The history of sexuality has focused on gender in this sense because mapping sex and gender separately has proved ontologically unsatisfying. Unlike other domains in which gender was first recuperative and then theorized, the history of sexuality was framed initially in discursive terms and then the particulars of past sexual experience were recovered.

Michel Foucault's *The History of Sexuality* (1978) situated the history of sexuality in an analytics of power. Outlining a model of power as operating in elliptical, subversive, and counter-intuitive ways, Foucault moved away from narrative forms which did

little to suggest how or why attitudes toward sexuality changed over time. Examining the cultural and social forces which constrained and policed expressions of sexuality, Foucault found that a proliferation of discourse brought the body and its functions under the pervasive view of larger structures of authority, namely the church and the state. Foucault challenged the standard paradigm that Victorian prudery replaced a more frank and free sexual climate which lasted until the seventeenth century. Instead, Early Modern Europeans lived amidst a whole host of structures which directed sexuality, while the Victorians talked about sex and talked about not talking about it all the time. They moved sex into the home behind closed doors and shrouded it in secrecy, but Victorians who tried to shield children from sexual knowledge and restrict sexuality to heterosexual monogamous relationships had the effect of proliferating discourse about sex. The proliferation of discourses brought the body and its functions—ironically, given the purported interest in making sex more private—under the more pervasive view of larger structures of power and authority, namely the church and the state. Foucault's understanding of sexuality as having a history and that this history was integrally intertwined with institutionalized power structures opened up historical analysis to the implications of sexuality with respect to both gender and sex.

The absence of actual agents and questions about the particulars of Foucault's revised narrative encouraged explorations of the relationship between gender and sex. Much work has been devoted to recuperating the human and historical specificity of the discursive environment Foucault described by locating behavior previously ignored or undetected. Explorations of sexual rituals such as courtship and marriage have revealed how social constructions of sexuality operated in the past (Evans 1993). Topical issues include deviance, the implications of sexually transmitted disease, and sexual coercion and violence. Historical specificity has made it clear that the emphasis on modern notions of identity as a product of sexuality are quite recent. The recuperation of a multiplicity of sexualities which are not linked necessarily to underlying identity formations has provided ammunition for opponents of tradition-based sexual policies. The claims of universal 'right' and 'wrong' lose their explanatory power in the face of historical variation.

2.2 Gendered Embodiment

The history of sexuality has been linked to new conceptualizations of the body. Western thought for much of modernity has downplayed the importance of the materiality of corporeal existence. The thinking subject was typically represented as pure intellect, as if the mind operated without a body. Gender and history have called into question the disembodied thinking subject and the corporeal body as a stable, material object.

Since Descartes' distinction between intelligence and thought (*res cogitans*) and corporeal existence (*res extensa*), the mind–body split he posited allowed for a notion of the body as a machine which obeyed laws of 'nature.' That 'nature' resolutely defined certain types of bodies (white, male) as normative while all others were lesser, inferior, or uninteresting remained largely unquestioned. Historians demonstrated that scientific disciplines including medicine, anatomy, and anthropology produced a hierarchized model of sexual and racial difference based on the normative white male body. The menstruating, lactating, reproducing female body was the epitome of difference and inferiority. Historically, pressure to police, control, and contain the persistent physicality of the female body was chronic.

Coming out of feminist concerns about social difference as inscribed on female bodies, the historians of the body have moved on to question the assumption that the body is always and everywhere constitutive of the self in the same ways. Feminist scholarship in the 1970s frequently engaged with the question of 'essentialism' as opposed to social construction. The central debate was whether men and women had 'essential' biological differences or were different because of socially constructed ideas and norms (Riley 1988). The body was a particularly fierce battleground, since so much of female oppression seemed tied to bodily functions. While some argued that female corporeality defined in terms of reproduction caused women's oppression, others were less pessimistic. Cultural feminists and activists in the women's health movement embraced the female body as the source of specifically feminine ethical sensibilities. The distinctive qualities claimed for the female body signaled the limitations of the debate as few initially questioned the epistemological grounding of biological difference.

Because of the existence of biological difference between male and female bodies, essentialism was hard to put to rest. How could one theorize the body at all without reference to essential differences? One direction was to explore understandings of the body over time. Historians of the body found that the meanings attached to the seemingly stable facticity of the body were not really stable at all. Even the idea that there are two sexes was relatively new. From ancient Greece until the late seventeenth century, the dominant model was that there was only one sex—male—and that women were simply the inversion of it (Laqueur 1990). While some have questioned whether the single sex model was entirely dominant, much scientific thought organized its understanding of the body and its functions on the one sex model. The cultural addenda to the one sex model were astonishing in their range and importance, particularly for notions of sexuality. For instance, cultural wisdom

held that women needed penetrative sex to keep all their organs in the proper place and in working order.

Another mode of inquiry began to question the very existence of corporeal embodiment. The materiality of the body can be seen as a function of text and physicality. This is not to say that bodies do not exist. Rather, the body is always mediated by context, the social constructions, and the notions of difference (of age, race, gender, mobility, and so forth) in which it exists. If embodiment is no more or less real than the contexts which create it, the question remains whether the conditions of constraint in which female bodies seem to operate can be dismantled (Grosz 1994).

2.3 Instability and Masculinity

In many domains of knowledge, male bodies remain the unspoken norm, but efforts to historicize not just the male body but the ideas which are attached to it as appropriately 'masculine' have suggested how difficult the normative position is to maintain in practice. Gender male is a product of men either succeeding or failing to embody appropriate behaviors. Since many of these behaviors are negatively defined as being 'not feminine,' masculinity is often difficult to assert and maintain effectively, and masculinity as a result is often highly compensatory. A brief examination of how concerns over masculinity have shaped larger ideological and historical configurations suggests how compensatory strategies of masculinity have had significant effects on Western European society.

The study of masculinity was initially resisted by some feminist scholars who came out of women's studies. Many feared that masculinity was just another way of studying men—that historians who emphasized masculinity were in a sense highjacking one of the central purposes of women's studies, which was to restore women to the history written largely by men and privileging, usually uncritically, the male subject. But, as Natalie Zemon Davis (1976) noted, consideration of women in isolation from men was rather like imagining a class society consisting only of peasants. Ignoring men is no better intellectually or logically than ignoring women. The two operate in conjunction with each other to create historically specific definitional frames for both masculinity and femininity. Studying one alone necessarily distorts the picture.

It turns out that masculinity is often marked by the deliberate absence of women, suggesting that gender still constructs masculinity even without gender female physically present. John Tosh (1994) has argued that masculinity is always, in Western society at least, a matter of public demonstration at home, at work, and in all-male associations. The three elements are mutually dependent and reinforcing, but also highly unstable. Being able to establish and provide for his family was a crucial marker of male adulthood. The male head of household was expected to maintain order and discipline in his family, and the failure to do so marked his masculine status as fundamentally suspect (Davidoff and Hall 1987). As the model of marriage became less authoritarian and the idea of separate spheres gained acceptance, masculinity at home was often asserted by husbands with even greater vigor. Although the notion of separate spheres has been challenged, the persistence of gender differentiation was perceived as an important element in masculine self-definition.

As Tosh (1994) has pointed out, work was even more complex to negotiate than family. Work for nineteenth-century British middle-class men had to support the family, but it also had to be dignified and the man's commitment to it complete. If work was supposed to be a vocation for middle-class men, for working-class men, the mechanization of work represented a serious threat to his sense of self. Without work defined as 'meaningful' to confirm their masculinity, male workers often insisted on more sharply distinctive notions of appropriate male and female behavior. When women were workers, they were emphatically assigned to different tasks that were defined as inferior in terms of the skill levels required. Masculinity in the workplace thus depended on the inferiority of women, expressed either as female incapacity or as lesser status (Rose 1988).

Tosh's third element, male associations, includes a range of institutions, from boy scouts to colleges to men's dining clubs. The homosociality of these and myriad other all-male settings provided practical supports for masculine privilege in terms of contacts and access to other men of power in the case of many adult organizations. More significantly, collective homosociality produced the notion that masculinity was located solely in the company of men. The persistence of women's discomfort in the face of such expressions of masculinity remained and is still a powerful element shaping gender configurations.

Because masculinity is defined against femininity, the relationality of the terms makes them unstable. Historians, for instance, have found that masculinity was at times associated with homosexuality as an appropriate masculine behavior. In Renaissance Florence, homosexual behavior was widespread despite being heavily penalized. The configuration of the older man as the active penetrator and the younger man as the passive penetrated partner were conceptualized to support notions of gender hierarchy which posited men as active and women as passive. Denunciations of men for assuming the incorrect role were frequently accompanied by language which subsumed the transgressor into negative female roles (Rocke 1996). Homosexual behavior could support masculinity if it conformed to gender norms, but was increasingly figured as a threat to appropriate masculine behavior in the nineteenth century. The emergence of the New Woman and the increasingly visible homosexual subculture which deliberately played with the signifiers of

masculinity made for the assertion of masculinity as patriarchal and resolutely heterosexual (Showalter 1990).

Masculinity, gender historians have revealed, is dependent on a range of variables which are themselves culturally constructed. Because masculinity is not always easy to inhabit and because the stakes for men in losing the privileges which generally accompany masculinity are very high, it is as much a source of anxiety as of power. The recent visibility of mass movements that emphasize masculinity and support it with emotional language and highly orchestrated collective bonding experiences suggests that the anxiety is still very much present.

2.4 Gender Roles as Performance

Recent efforts to explore the history of identity formation have incorporated post-structuralist theories to tease out how the self was defined in gendered terms and how the gendered self related to others and to entities such as the state. Continental post-structuralist feminism has largely followed Jacques Lacan's psychoanalytic theory, developing the analysis of discourse in a mode which remains highly attenuated from material concerns. Lacanian psychoanalysis feminists, led by Julia Kristeva (1980), have re-interpreted the mirror stage of development to emphasize the abjection of the female subject. If the unspoken male subject in Lacan's formulation comes to his identity through engagement with the mirror image of the self, for the female subject, this experience is violent and alienating.

Judith Butler's articulation of gender as performance captures a number of elements of post-structuralist theory which have then been utilized to examine identity formation in the past. Butler (1990) argues that gender is best understood as a performance in which the subject is constituted in relation to the regulatory practices of society. The norms of these practices are themselves culturally constructed around heterosexuality defined by masculine and feminine binaries. Gender is actually wholly performative around those binaries such that it exists only in constituting a particular identity. While some historians expressed concern with the postmodern tendency to deconstruct without offering positive constructions, notions of identity as performative have helped recuperate historically specific moments of identity formation. Understanding gender as a signifying practice helps reveal how it is integral to other expressions of identity in specific historical settings.

3. Politics of Several Varieties

Going back to the ancient Greeks in the Western European tradition, politics has been a highly gendered pursuit. Politics was one of the main arenas in which men achieved honor and glory in the ancient world, and despite the overlay of Christian asceticism, politics has remained that, despite the attempt to present political participation as a gender neutral act on behalf of the public good. Throughout much of history, men of appropriate class standing defined themselves in terms of their eligibility for public political life. Even though politics is no longer formally restricted to male participants in Western Europe, old exclusivities die hard.

3.1 High Politics and Political Theory

The replacement of monarchical systems of governance with more or less democratic ones in Western Europe was long assumed to be a gender neutral change for the better. The rights-bearing political subject was, however, always male until the twentieth century, a fact noted but generally unquestioned by historians. Recent analysis has revealed that gender difference has been foundational to conceptions of politics throughout the Western tradition (Pateman 1988). Ancient theorists, particularly Aristotle, described political participation as reliant on the qualities of mind and body of men. Female inferiority was based on mental and physical weakness which required the restriction of women as much as possible to the home. That this restriction was in place largely to protect paternity and inheritance dropped out of the theoretical descriptions of society. Both Aristotle's teleological analysis of politics and the persistence of anxiety about paternity and female sexuality enabled the political arrangements to remain almost entirely male in expression.

The gendered logic beneath the political was obscured by the elaboration of misogynous overlays concerning women, beginning especially in the late Middle Ages with the emergence of the *querelle des femmes*. Except in France, where women were explicitly barred from monarchical inheritance, the occasional female monarchs only underscored the rule that men did politics and women did not. This was true in the Italian and Dutch republics as much as in monarchical states, and remained generally unquestioned until the French Revolution.

The positing of universal citizenship based on inherent rights presented a philosophical and practical challenge to all-male political arrangements. Even when other notions of hierarchy and difference were falling, however, gender remained. Joan Landes (1988) has argued that political participation as articulated in the French Revolution was foundationally masculinist rather than contingently or accidentally so. Not simply a matter of women being overlooked, revolutionaries made an active effort to differentiate the ability to be public and political based on sexual difference. The assumption remained that sexual difference rendered women incapable of proper political activity. The

increased language about domesticity and women being properly 'private' in the nineteenth century was a result of the reassertion of politics as properly gendered male. Ironically, the ancient justification for female exclusion from public affairs—concerns about her chastity and female sexual voraciousness—was completely overwhelmed by the assumption of female sexual passivity. White middle-class women were increasingly seen as asexual, although this was constructed against assumptions that women of other races and classes were sexually voracious and threatening (Hammonds 1999).

As this recounting of gender and politics has suggested, participation in high politics has long been organized and informed by notions of exclusion which are consistently centered on women as problematic. The persistence of the gendering of politics in the twentieth century can be seen in the long fight that was necessary to secure female suffrage in most Western European states and in the paucity of women elected to high office and in positions of leadership in political parties. The awareness that this persistence is not 'natural' but constructed has prompted some elected officials and feminists to challenge the overwhelmingly male character of high politics, but the association of women with domesticity remains strong.

4. *The Politics of the Personal*

The gendering of high politics, of course, is dependent on the gendering of domains regarded as outside of politics. The personal or the private are just as much elements of the political as matters of electoral politics and state policy. This was an important lesson of the women's movement, but the implications of it remain controversial in many areas. The fractures in social policy often go in many directions, as a number of areas related to sexuality and sexual expression suggest. The relationship of work to family is just one of the most prominent places in which public politics have become highly contested over what used to be considered private issues.

4.1 *Social Policies for Work and Family*

Work and family have long been seen as quintessentially opposites: work is public while family is private. But work and family are no longer seen as naturally occurring or organized in light of the historical analysis of gender. The language of 'natural' gender roles remains strong with respect to work and family, but the linking of the terms as historically inseparable has been combined with the imbrication of family and work issues in the welfare state. Even in the domain of social policy, understandings of gender as historically specific and constituted in cultural terms can inform policy development.

Gender history has focused on the cultural determination of work and family in Western Europe. Examinations of tax records and census documents reveal that the gendering of work does not follow the expectation that men will do heavy work while women do the less physically demanding jobs. In fact, work has long been assigned to men and women along culturally determined lines which were a function of the household as the unit of production for much of Western European history. Efforts to understand the historical relationship between work and family led to the emergence of two important paradigms. First, historians of women posited that women enjoyed a relatively positive working experience—with greater opportunity and higher wages—before industrial capitalism. Some even referred to early modern women's work as a 'golden age' in Western Europe. Second, historians argued whether continuity in the face of patriarchy or change in conjunction with larger historical trends best defined the history of women's work (Sharpe 1995). The controversy has produced efforts to historicize the workings of patriarchy so that the components of continuity are understood for their flexibility. This has enhanced present-day understandings of the persistence of gendered workplaces, occupational structures, and wage scales. At the same time, concerns about seeing the history of women's work as always the same produced a critique that emphasizes the interactive nature of gender in the workplace. Men's and women's roles shift in relation to each other and to larger economic changes such as the rise of industrial capitalism.

The gendering of occupations outside the home seems to follow shifts in competition between men and women for jobs. The chronology and geography of these shifts remains under exploration, as does the direction of change. Do changes in the family enable changes in workplace gender configurations or does the gendering of work force changes in the family (Berg 1993)? The question remains an open one, but is relevant to Western European controversies over the welfare state and gender patterns in employment. The present trends toward the separation of sex from marriage, increased divorce rates compared to earlier historical periods, child-rearing increasingly handled outside the family, and adjustments in the gendering of labor, signal significant transformation in work and family organization. The male breadwinner model persists, but along with it is increased pressure on men to participate in the emotional life of the family. If men are to be more involved in the family, the recognition of women's unemployment suggests that women are no longer automatically considered to be wives and mothers with paid employment as a secondary identity.

The points of articulation between policy and cultural expressions are the subject of recent analysis. Diane Sainsbury (1999) has argued that different and competing models of the welfare state are actually

both the result of gender and familial ideologies. The deficiencies in much welfare state analysis are a result of the failure to recognize gender as a crucial organizing framework in policy. Language and culture once again are the loci in which gender and history provide the bases for analysis.

5. Implications and Conclusions

Because gender is operative in all times and places in which human beings interact, gender and history provide the means and methodologies to understand the developing global community. Postmodern analytical techniques when applied to problems of gender in history in Western Europe have produced understandings of identity, politics, and sexuality in terms of their interrelations. Historical practices have just begun to articulate how specific experiences in different times and places relate to each other. Given that human communities are always male and female, and thus always marked by negotiations over expressions of masculinity and femininity, over the meanings of the body, and over the valuation of work and family, gender and history encompass issues global in scale in which to situate local understandings and problems.

See also: Family and Kinship, History of; Foucault, Michel (1926–84); Gender, Economics of; Gender History; Labor Markets, Labor Movements, and Gender in Developing Nations; Postmodernism: Philosophical Aspects; Sexuality and Gender; Western European Studies: Culture; Western European Studies: Society

Bibliography

Berg M 1993 What difference did women's work make to the Industrial Revolution? *History Workshop Journal* **35**: 22–44

Bock G 1989 Women's history and gender history: Aspects of an international debate. *Gender and History* **1**(1): 7–30

Bonnell V, Hunt L (eds.) 1999 *Beyond the Cultural Turn*. University of California Press, Berkeley, CA

Braidotti R 1994 *Nomadic Subjects: Embodiment and Sexual Difference in Contemporary Feminist Theory*. Columbia University Press, New York

Butler J 1990 *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, London, UK

Davidoff L, Hall C 1987 *Family Fortunes: Men and Women of the English Middle Class, 1780–1850*. University of Chicago Press, Chicago, IL

Davis N Z 1976 'Women's history' in transition: The European case. *Feminist Studies* **3**: 83–103

Evans D T 1993 *Sexual Citizenship: The Material Construction of Sexualities*. Routledge, New York

Foucault M 1978 *The History of Sexuality*, *Volume* 1: *An Introduction* [trans.] Hurley R. Vantage, New York

Grosz E 1994 *Volatile Bodies: Toward a Corporeal Feminism*. Indiana University Press, Bloomington, IN

Hammonds E 1999 Toward a geneology of black female sexuality: The problematic of silence. In: Price J, Shildrick M (eds.) *Feminist Theory and the Body: A Reader*. Routledge, New York, pp. 93–104

Irigaray L 1985 *This Sex Which is Not One*, trans. Porter C, Burke C. Cornell University Press, Ithaca, NY

Kristeva J 1980 *Pouvoirs de l'horreur: Essai sur l'abjection*. Éditions du Seuil, Paris, France

Landes J 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca, NY

Laqueur T 1990 *Making Sex: Bodies and Gender from the Greeks to Freud*. Harvard University Press, Cambridge, MA

Moi T 1985 *Sexual/Textual Politics*. Methuen, London

Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA

Riley Denise 1988 *'Am I that Name?' Feminism and the Category of 'Women' in History*. University of Minnesota Press, Minneapolis

Rocke M 1996 *Forbidden Friendships: Homosexuality and Male Culture in Renaissance Florence*. Oxford University Press, New York

Rose S O 1988 Gender antagonism and class conflict: Exclusionary strategies of male trade unionists in nineteenth-century Britain. *Social History* **13**: 191–208

Sainsbury D 1999 *Gender and Welfare State Regimes*. Oxford University Press, Oxford

Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York, pp. 28–50

Sharpe P 1995 Continuity and change: Women's history and economic history in Britain. *Economic History Review* **48**: 353–69

Showalter E 1990 *Sexual Anarchy: Gender and Culture at the Fin de siècle*. Viking, New York

Tosh J 1994 What should historians do with masculinity? Reflections on nineteenth-century Britain. *History Workshop Journal* **38**: 179–202

K. Crawford

Western European Studies: Geography

1. The Changing Discipline of Human Geography

This section presents a brief history of geographical thought in order to situate the approaches that human geographers have developed in studying Western Europe (for a fuller history, see Johnston 1991). During the nineteenth century, geography in Western Europe sought to describe and explain the spatial distributions of the natural world and of people and their activities. However, association between strong strands of environmental determinism that sought to explain human characteristics in terms of natural environmental conditions, exemplified by the work of Ellen Semple, and the politically discredited doctrine

of *Lebensraum* in Nazi Germany, produced a shift from explanatory concerns to regional description. While there were variations between national 'schools,' with the French under the influence of Paul Vidal de la Blache most powerfully exemplifying the regional descriptive approach, there was also a degree of commonality of approach to geographical scholarship across Western Europe.

Then, in the 1960s, influenced by earlier developments in North America summarized as the 'quantitative' and 'conceptual' revolutions, human geography in Western Europe began to be redefined. The emphasis switched to spatial science and explanation of generalized spatial patterns, drawing on ideas from neoclassical economics, and from qualitative regional description to quantitative—especially statistical—analysis of spatial patterns. Initially 'revolution' was concentrated in a few Departments in the UK and Sweden. From there, it diffused to other parts of Western Europe. These changes were contested, however. Much Western European geography retained a strong interest in the uniqueness of places, a focus that later re-emerged, albeit in a modified way. It soon became clear that spatial science had major limitations. Initially, behavioral geographers in Western Europe, as in North America, responded to these by seeking more realistic assumptions about the knowledge and motives that underlay peoples' spatial behaviors (Pred 1967). This did little to remedy the explanatory weaknesses, however. Consequently, human geographers sought stronger explanations grounded in more powerful abstractions of the processes that generated spatial patterns. As a result, Marxian and neoMarxian political economy soon became the focal point of 'critical' geography within Western Europe (Carney et al. 1980) and more generally (Harvey 1982). However, there were again strong counter-pressures, with humanistic geography (Ley and Samuels 1978) emphasizing human agency, meaning, intentionality, and individual life-worlds, rather than the unfolding structural logic of capital in explaining spatial variations in human activity.

Around the late 1970s, however, some important changes emerged in the social sciences, with implications for theory and practice in human geography and the ways in which Europe was studied. As a result, human geography in Western Europe became a pluralist discipline, encompassing many epistemological positions and substantive focuses, with strong links to cognate social sciences. Most importantly, there was an elaboration of seminal ideas about the centrality of space to social and economic life and growing recognition of the significance of spatiality in many of the social sciences. This generated a constructive dialogue within Western Europe across national and disciplinary boundaries. It involved social scientists seeking to infuse their theories with spatial sensitivity, and geographers (such as Massey 1984) exploring relations between social process and spatial form. 'The difference space makes' to the constitution of societies, and the ways in which economic, social, and political processes operate within the structural limits that define capitalist economies, became cross-disciplinary research frontiers.

There was growing recognition that relationships between social processes and spatial forms are reciprocal, contingent, and indeterminate. This led to exploration of the ways in which such relationships actually do evolve and the spatiotemporally variable forms that they take, directing attention to the institutions through which societies and their geographies are constituted and reproduced. The prime focus of concern became middle-level theories and concepts as human geography underwent an 'institutional turn.' Linked to this, poststructural approaches resulted in greater sensitivity to the limitations of, and absences from, metanarratives. This was associated with a broader 'cultural turn' in human geography and the social sciences, and heightened awareness of cultural variation in the ways in which capitalist societies are constituted and of the variable meanings that people, places, and events can assume because of this (Hudson 2001).

Paralleling developments of the 1980s in terms of sociospatial relationships, in the 1990s the social sciences in Western Europe increasingly acknowledged the significance of connections between the social and natural worlds in understanding societies and their geographies. As such, another 'traditional' geographical concern was given a new twist and became central to broader social science debates. One consequence of this was growing interest in Actor–Network theories, which seek to bridge the divide between social and natural worlds and explore how human subjects and nonhuman objects become contingently linked in networks in particular time/space contexts (Thrift 1996).

These growing concerns with relationships between society, space, and nature coalesced in debates about processes of globalization, links between 'the global and the local' and their impacts on and in Europe (Amin and Thrift 1994). This led to recognition that changes at 'global' and 'local' levels are reciprocal and complex rather than simply one-way and one-directional. While processes of globalization influence 'local' change in Europe, those processes are affected by 'local' place-specific effects in Europe (from the formation of the European Union to local social movements protesting about environmental pollution or cultural destruction). Some see 'globalization' as homogenizing, eliminating differences between places under the relentless pressures of capital accumulation and technological advances in transport, information technology, and telecommunications as the world of places dissolves into a world of flows (Castells 1996). In contrast, human geographers in Europe have conceptualized it as a complex interplay of processes that both link and help define varying spatial scales in

a complex mosaic. Globalization is seen as the latest phase of combined and uneven development, enhancing the distinctiveness of places and their importance in the contemporary world.

2. Changing Emphases in the Geographical Study of Western Europe

2.1 The Changing Geopolitical Map of Europe and the Meaning of 'Western Europe'

The dramatic changes in the geopolitical map of Europe in the postwar period have attracted the attention of geographers (such as Williams 1998). These changes include:

(a) the division of Europe between East and West following the partition of Berlin and the lowering of the imaginary but highly influential 'iron curtain' in 1948;

(b) the creation of the European Coal and Steel Community in 1951 and then the European Economic Community (EEC) in 1957 by the Benelux countries, France, Germany, and Italy;

(c) the formation of the European Free Trade Area (EFTA) in 1961 by nations that either did not want or were not allowed to join the EEC (Austria, Denmark, Finland, Ireland, Norway, Switzerland, the UK);

(d) the subsequent revision of the boundary between the EEC and EFTA, and the admission of nonEFTA states to the EEC; and

(e) the collapse of state socialism in the east in 1989, the destruction of the Berlin Wall, and a new phase of expansion of the European Union (EU), as it had become in the process of deepening economic integration, in the 1990s.

By 1995 the EU had expanded to 15 member states. Currently, there are several other applicant nations in Central, Eastern, and Southern Europe (Cyprus, the Czech Republic, Estonia, Hungary, Poland, Slovenia, and Turkey). Other nations (Bulgaria, Latvia, Lithuania, Malta, Romania) may become applicants for accession in the foreseeable future (Hudson 2000).

Three cleavage planes have been decisive in this ongoing partitioning of geopolitical space, in creating boundaries between different groupings of nation-states and in the allocation of nations to groups. First, there was a deep division for four decades between nations embracing capitalism and those embracing state socialism. Second, within the former group, there was a split between those committed to parliamentary democracy and those (temporarily) characterized by dictatorship (Greece, Portugal, and Spain). Third, there is continuing competition between models of national capitalisms and varying relationships between national economies and (welfare) states (Hudson and Williams 1999). There has also been competition between national states and interests and supranational interests represented in the emergent EU. The evolving postwar geopolitical map of Europe has thus been the product of highly contested processes.

2.2 New Multiscalar Geographies of Governance and Regulation: Restructuring the National State

The changing architecture of governance and regulation within Europe has been a related focal point of study. In postwar Western Europe, there were strong pressures to check the spread of communism, prevent future wars between its major states, and create a more effective political voice in the international arena. This led to the construction of supranational European institutions, beginning a process of redefining the architecture of governance and regulation in Europe.

Within the EU, there has been an upward drift of power, a 'hollowing out' as part of processes of national state restructuring (Jessop 1997). The increased regulatory competencies of the EU have been interpreted in one of three ways. First, for some, shifting state regulation 'up' a spatial scale signifies evidence of an emergent superstate. Second, it is argued that political changes in the 1980s and 1990s produced a complex hybrid form of decision-making processes in the EU, combining intergovernmentalism and cooperative federalism. Third, others view the EU as one element within a more complex multilevel and multidimensional system of governance. Whatever the interpretation, however, the EU is seen as weakened by a deep democratic deficit (Painter 1999).

While the EU can be seen as the most elaborated example of an emergent multilevel system of governance and regulation, there are strong continuities with the past, especially in the continuing significance of national states. Some new forms of economic and financial globalization ('stateless monies,' for example) do take matters beyond the effective control of many national states. However, the view that the significance of the national state is being eroded via 'hollowing out' has been challenged on various grounds. In many respects, national states in the EU retain considerable powers (Anderson 1995), and these states are perhaps the paradigmatic case for 'hollowing out' theorists.

3. Europeanization, Globalization and the Impacts of these Processes within Europe

The development of the EU as a homogenized economic space, together with the extension of capitalism into Central and Eastern Europe, created new spaces in which companies can reorganize their activities. This has become another focus of geographic research in and on Europe (Hudson 1999a). The

prospect of the Single European Market generated a surge in intraEU foreign direct investment and enhanced cross-border acquisitions within the EU. These often led to rationalization, plant closure, and job loss. Investment by Western European companies in CEE also increased somewhat, although heavily spatially concentrated at national and subnational scales, further widening spatial inequalities.

Europe-wide strategies have frequently been part of broader corporate globalization strategies. Creating bigger European companies via acquisition or merger has often been necessary to enable the enhanced spatial scale of corporate organization. There has also been an increase in strategic alliances between these bigger companies, especially in manufacturing industries characterized by high entry costs, continuing importance of economies of scale and scope, rapidly changing technologies, and substantial operating risks (Dicken and Oberg 1996). Processes of acquisition, merger, and strategic alliance have also reshaped parts of the service sector, ranging from banking and financial services to retailing, in response to competitive pressures and possibilities created by technological changes and deregulation (Hudson 1999a).

In the 1990s, transnationals based in the USA, Canada, and south east Asia increasingly entered strategic alliances with European companies. There was also a surge in acquisition activity by nonEU transnationals in Europe, along with increased direct inward investment in anticipation of completion of the Single Market. Inward Foreign Direct Investment (FDI) to the EU has been concentrated in banks and financial services and in existing major metropolitan centers. Simultaneously, technological advances in computing and telecommunications, deregulation of previously nationally protected markets in services, and resultant pressures on companies to cut costs to remain competitive, led to decentralization of back-office functions to peripheral locations. There was some further inward investment in manufacturing, but with a rather different geography. Investment in activities such as European headquarters offices and research and development (R&D) has been in major cities in economically strong core regions, but has been limited in extent. Most has been in routine production activities in consumer good sectors such as automobiles, electronics, and 'white goods.' It has involved both acquisition of existing companies and factories and inward investment in new 'greenfield' manufacturing plants, the latter typically in peripheral locations in both Eastern and Western Europe (Hudson 1999a).

3.1 The Rise of the Regional Economy and the Rediscovery of the Urban Economy in Europe

Some researchers in Western Europe have increasingly identified the region as the crucial territorial and institutional framework for economic success, linked in part to debates about national state 'hollowing out' and regional devolution. There have been two strands to this work. First, stimulated by Bagnasco's (1977) seminal study, there has been great interest in industrial districts, regionally concentrated mono-industrial production systems, producing a range of labor-intensive consumer goods (such as clothing, shoes, or ceramics). They are organized via relatively egalitarian horizontal networks of small firms, built around trust and linked into complex social divisions of labor. These districts are also characterized by cooperative capital–labor relations, and by supportive institutional structures in the local state and civil society. The net result is flexible production systems, able to respond swiftly to changing tastes and consumer demands. Since the 1970s, there have been a number of studies of similar industrial districts in various parts of Europe (for example, see Maskell et al. 1998) while some scholars have conceptualized key 'command and control' nodes within global cities, such as the City of London, as a particular form of industrial district (Amin and Thrift 1994).

The second strand focuses upon a different type of territorially clustered network structure, with close links between companies and the institutions of state and civil society, but based upon hierarchical and asymmetrical power relations between the constituent firms. These are production systems driven by major firms, with others linked to them in tiered supply chains, with Baden-Württemberg perhaps the 'classic' European example (Herrigel 1996). A more recent variant of this regional economic structure is linked with inward investment, especially from Japan and south east Asia, as assembly plants attract component suppliers to locate nearby. This has led to contested claims about a 'new' type of higher quality of investment in 'embedded' branch plants, more integrated into the surrounding regional economy than traditional 'global outpost' branch plants and so with greater regional developmental potential. Such new factories are also seen by some as bringing new, enriching, and more satisfying multiskilled jobs but by others as offering intensified, multitasked jobs that further disempower workers (Hudson 1999a).

Emphasis on the importance of regional cultures and institutions in an associational economy underpinning regional economic success has led researchers to highlight the significance of concepts such as regionally specific social capital, trust, untraded interdependencies, tacit knowledge, learning, and appropriate institutional structures. Important studies in this regard include Amin and Thrift (1994) and Storper (1995), although others are more skeptical about this 'regional cultural and institutional turn' (Gertler 1997, Hudson 1999b). Nonetheless, such ideas have also been increasingly influential in urban studies as geographers have developed new perspectives on cities (for example, see Amin and Thrift 2001).

4. Structural Economic Change, Service Sector Growth, and Feminization of Labor Markets

Service sector growth as part of a changing social division of labor has been the most significant structural change in the European economy and led to claims about the emergence of postindustrial and postFordist economies. There have also been important changes within the service sector. Initially, service sector expansion was closely linked with growing state involvement over much of Europe. More recently, the dominant pattern has been shrinkage of the public sector and expansion of private sector services. The most significant feature of the latter has been growth of private sector business and financial services, often linked to financial product innovations and deregulation of financial markets (Tickell 1999).

The increasing importance of the service sector has also been associated with three important changes in European labor markets. First, further 'feminization' of the labor market has led to some growth of well-paid, full-time secure professional employment for women but more generally growth has been in less desirable jobs. Second, women form a disproportionately large part of the 'flexible workforce,' produced by the growth of new forms of employment contract and part-time and casual work (Perrons 1998). Third, service sector growth, especially in business and financial services, has led to renewed focus on urban environments as key nodes of economic dynamism (for example, see Keil and Ronnenberger 2000). This has led to increasing interest in the characteristics of urban environments seen as conducive to high order service sector growth and the attraction of dynamic new services and cultural industries within an emergent 'knowledge economy.' Echoing earlier work on successful regions, this again emphasizes the significance of 'soft' infrastructure and informal social relationships as well as the necessary 'hard infrastructure' of transport, telecommunications, and informatics networks (for example, see Amin and Thrift 2001).

5. New Geographies of Identities

European identity and the criteria used to denote 'Europeans' have emerged as a focus of interest. EU member states have clearly become *de facto* multiethnic and multicultural societies, raising questions about cultural and ethnic variation and issues of (multiple) identities at national, regional, and local scales within the EU and about the character and location of the boundaries of the EU (Amin 2001). Boundary definition is closely tied to processes of 'othering' and the criteria by which 'Europeanness' is to be judged. As the EU expands, particularly eastwards into areas that border on or are part of Asia as commonly understood, these issues become more pressing.

One conception of a singular European identity would see it constructed through a process analogous to that involved in creating national identities. However, the project is now to transcend those identities. Previously the 'imagined community' was national; in the new imagination it is to be European (or perhaps more accurately EU). For many people, there is now undoubtedly greater awareness of other cultures and lifestyles elsewhere in Europe. Such shared activity and communication spaces do not necessarily translate into a shared European consciousness, however. The EU's cultural and educational policy initiatives have sought to encourage the formation of a common consciousness and shared identity beyond existing local and national affinities. This view of European identity and 'Europeanness' sits comfortably with the preferences of many members of political élites within Europe. However, it lacks a firm grounding in popular consciousness (Smith 1995).

There are pressures from national states and their citizens to resist further erosion of national identity and sovereignty. The Danish rejection of the Euro and the electoral success of extreme right-wing political parties in countries such Austria, Belgium, and France, for example, suggests the resurgence of more insidious and xenophobic nationalisms and racisms (White 1999). Furthermore, there are strong pressures from 'suppressed nations' seeking their own national states, so that territory corresponds with identity, as events in the Caucasus and the Balkans in the 1990s made painfully clear. At the same time, however, localist and regionalist movements challenge national states and seek greater autonomy from the center, asserting their particular cultural and territorial identities over the national. Often this builds upon past histories of national states suppressing subnational variation and culture in the process of building national identities (see, for example, Fernández Rodríguez 1985). Such tendencies are expressed widely over Western Europe (Anderson 1995). These regionalist and substate nationalist tendencies often emerged in a complex relationship to processes of globalization, Europeanization, and the emergence of the EU. Often they were seen as a way of resisting global forces via forging alliances between the regional and supranational EU levels, bypassing the national state. Thus complex geographies of hybrid identities were forged, with allegiance to territories at different scales, which became a focus of interest for human geographers and social scientists (Amin 2001).

6. Geographies of Social Inequality

One consequence of processes of political and economic restructuring is that Europe has become a more sharply divided place, characterized by deepened sociospatial divisions. Processes of economic restructuring have etched the contours of inequality

more deeply into the European landscape, at several spatial scales: between countries, between cities, and between regions within countries; and within cities and regions (Bennett et al. 2000). There are growing divisions between the long-term unemployed—structurally excluded from the labor market—and the employed, but also within the employed workforce (in terms of wages and terms, conditions, and security of employment). There has been growing recognition of gender divisions (Perrons 1998) while issues of ethnic and racial divisions have often been exacerbated by EU expansion and by immigration and fears of 'the other,' especially those of different skin color (White 1999). There has also been a growing recognition that spatial divisions intertwine in complex ways with those of class, ethnicity, and gender (Hadjimichalis and Sadler 1995).

7. *Social Exclusion, Inclusion and 'Third Sector' Local Economic Initiatives*

Growing unemployment, poverty, and inequality in the EU have led to growing concern—academic and political—with new ways of creating work and delivering services to disadvantaged communities. The focus of this attention has been on the social economy and 'third sector' economic initiatives, typically constructed at local level and designed to tackle social exclusion. While social exclusion encompasses a variety of dimensions beyond lack of waged employment, in practice attention has increasingly been focused upon tackling problems of unemployment, especially long-term and amongst the young, as these have become stubbornly persistent in Western Europe. The emphasis is upon creating new forms of socially useful work as an alternative to both private sector and state employment that will provide meaningful jobs for the unemployed and/or provide services that are no longer provided by state or markets to deprived communities (Lipietz 1995). There are concerns, however, that this may simply provide cut-price alternatives to public sector welfare services while reproducing inequalities from mainstream economy and society within the 'third sector' (Amin et al. 1999).

7.1 *Sustainable Societies and Places*

Sustainability in various guises—of natural environments, economies and societies, cities and regions—has become a further research focus within Western Europe, not least because it has become a pressing practical and political issue. Initially the emphasis was upon investigating ways of enhancing natural environmental sustainability but it was increasingly recognized that this must be situated in the context of transition to methods of production and transportation and lifestyles and consumption patterns with a lighter ecological footprint. This encompassed both technological innovation to generate materially more efficient methods of production and movement and institutional and regulatory changes to increase the efficiency of resource use and reduce waste. Environmental and social justice were increasingly seen to be intimately linked. Such issues are likely to become more important on geographical and social science research agendas in Western Europe in the twenty-first century, at a variety of spatial scales from the global to the regional and urban (Hudson 2001).

8. *Conclusion*

There were significant changes in the practice of human geography in Western Europe in the second half of the twentieth century, in how geographers studied Europe and in what they studied as Europe itself changed. Furthermore, growing interest in the social sciences with relationships between society, space, and nature—geography's established mainstream concerns—led to exciting developments across national and disciplinary boundaries in the study of Europe. These in turn had wider ramifications within geography and parts of the social sciences within and beyond Europe.

See also: Europe: Sociocultural Aspects; European Integration; Geographic Thought, History of; Geography; Humanistic Geography; Regional Geography; Urban Geography; Western European Studies: Society

Bibliography

Amin A 2001 Immigrants, cosmopolitans and the idea of Europe. In: Wallace H (ed.) *Whose Europe?* Macmillan, London

Amin A, Cameron A, Hudson R 1999 Welfare as work? The potential of the UK social economy. *Environment and Planning A* **31**: 2033–51

Amin A, Thrift N (eds.) 1994 *Globalization, Institutions and Regional Development in Europe*. Oxford University Press, Oxford, UK

Amin A, Thrift N 2001 *Cities: Re-Thinking Urban Theory*. Polity, Cambridge, UK

Anderson J 1995 The exaggerated death of the nation state. In: Anderson J, Brook C, Cochrane A (eds.) *Global Worlds*. Oxford University Press, Oxford, UK

Bagnasco A 1977 *Le Tre Italie*. Il Mulino, Milan

Bennett K, Beynon H, Hudson R 2000 *Coalfields Regeneration: Dealing with the Consequences of Industrial Decline*. Policy Press, Bristol, UK

Carney J, Hudson R, Lewis J (eds.) 1980 *Regions in Crisis: New Directions in European Regional Theory*. Croom Helm, London

Castells M 1996 *The Rise of the Network Society*. Blackwell, Oxford, UK
Dicken P, Oberg S 1996 The global contest: Europe in a world of dynamic economic and population change. *European Urban and Regional Studies* **3**(2): 101–20
Fernández Rodríguez F 1985 *La España de las Autonomias*. Taravilla, Madrid
Gertler M 1997 The invention of regional culture. In: Lee R, Wills J (eds.) *Geographies of Economies*. Arnold, London, pp. 47–58
Hadjimichalis C, Sadler D (eds.) 1995 *Europe at the Margins: New Mosaics of Inequality*. Wiley, Chichester, UK
Harvey D 1982 *The Limits to Capital*. Arnold, London
Herrigel G 1996 Crisis in German decentralized production: Unexpected rigidity and the challenge of an alternative form of flexible organization in Baden Württemburg. *European Urban and Regional Studies* **3**(1): 33–52
Hudson R 1999a The new economy of the new Europe: Eradicating divisions or creating new forms of uneven development? In: Hudson R, Williams A (eds.) *Divided Europe: Society and Territory*. Sage, London, pp. 29–62
Hudson R 1999b The learning economy, the learning firm and the learning region: A sympathetic critique of the limits to learning. *European Urban and Regional Studies* **6**(1): 59–72
Hudson R 2000 One Europe or many? Reflections on becoming European. *Transactions of the Institute of British Geographers, New Series* **25**(4) (forthcoming)
Hudson R 2001 *Producing Places*. Guilford, New York (forthcoming)
Hudson R, Williams A 1999 Re-shaping Europe: The challenge of new divisions within a homogenized economic space. In: Hudson R, Williams A (eds.) *Divided Europe: Society and Territory*. Sage, London, pp. 1–28
Jessop B 1997 Capitalism and its future: Remarks on regulation, government and governance. *Review of International Political Economy* **4**: 561–81
Johnston R J 1991 *Geography and Geographers*, 4th edn. Arnold, London
Keil R, Ronneneberger K 2000 Francfort-sur-le-Main; capitale fatale dans une région métropolitanisée. In: Benko G, Lipietz A (eds.) *La Richesse des Régions: La Nouvelle Géographie Socio-Économique*. Presses Universitaire de France, Paris, pp. 225–45
Ley D, Samuels M S (eds.) 1978 *Humanistic Geography: Prospects and Problems*. Croom Helm, London
Lipietz A 1995 *Green Hopes: The Future of Political Ecology*. Polity, Cambridge, UK
Maskell P, Eskelinen H, Hannibalsson I, Malmberg A, Vatne E 1998 *Competitiveness, Localised Learning and Regional Development*. Routledge, London
Massey D 1984 *Spatial Divisions of Labour*. Macmillan, London
Painter J 1999 New geographies of democracies in contemporary Europe. In: Hudson R, Williams A (eds.) *Divided Europe: Society and Territory*. Sage, London, pp. 107–31
Perrons D 1998 Maps of meaning: Gender inequality in the regions of western Europe. *European Urban and Regional Studies* **2**(2): 99–120
Pred A 1967 *Behavior and Location*. Department of Geography, University of Lund
Smith A 1995 *Nations and Nationalism in the Global Era*. Polity, Cambridge, UK
Storper M 1995 The resurgence of regional economies, ten years later: The region as a nexus of untraded interdependencies. *European Urban and Regional Studies* **2**(3): 191–223
Thrift N 1996 *Spatial Formations*. Sage, London
Tickell A 1999 European financial integration and uneven development. In: Hudson R, Williams A (eds.) *Divided Europe: Society and Territory*. Sage, London, pp. 63–78
White P 1999 Ethnicity, racialization and citizenship as divisive elements in Europe. In: Hudson R, Williams A (eds.) *Divided Europe: Society and Territory*. Sage, London, pp. 210–30
Williams A 1998 The European Union: Cumulative and uneven integration. In: Unwin T (ed.) *A European Geography*. Longman, London, pp. 129–48

R. Hudson

Western Europe Studies: Health

The health domain has served as a focal point for investigating the emergence and extension of state capacities for social governance in Western Europe. Such studies have considered, on the one hand, the desire and ability of nation–states to govern the health sector, and, on the other, the amenability of the health sector to governance. Governance, in this case, refers to efforts by governments to intervene in the private provision of medical services. Not only have governments throughout Western Europe underwritten the costs of medical care, they have introduced standards and mechanisms for the control of the quality, availability, and cost of medical services. Consequently, governance of the health sector has entailed government regulation of medical practitioners, and hence, political conflict between doctors and the state (see *Health Care Systems: Comparative*; *Health Policy*; *National Health Care and Insurance Systems*).

1. Health Politics

1.1 Interest-group Politics

Political scientists have regarded the health area as a test case for the theory of political pluralism. Pluralist theory focuses on interest groups as the key actors in the political process (see *Interest Groups*; *Interest Groups, History of*; *Pluralism*). Citizens, the theory assumes, form interest-groups when confronted with problems, and use these groups to lobby for the resolution of these problems. As a result, both the array of organized groups and the pattern of governmental activity should, in theory, reflect these popular demands. Studies of the role of medical associations in health politics have called this model into question, however. Medical associations appear to have had a strong influence on health policy decisions. To the

extent that medical associations are able to veto any proposal to which they are opposed—by threatening to go on strike or by questioning the medical soundness of government programs—one can speak of 'undue' or excessive political influence, or even of professional dominance. Although successful interest-group influence is part and parcel of the pluralist paradigm, such one-sided influence would constitute a fundamental challenge to the pluralist ideal of politics as a compromise amongst competing interests.

1.2 History of Health Insurance: the Nineteenth Century

Research on Western European health politics has investigated the conflicts surrounding public health insurance programs. Beginning in the late eighteenth and early nineteenth centuries, many European governments introduced laws for registering and subsidizing voluntary mutual aid societies or sickness funds (Friendly Societies, *Hilfskassen*, *Mutuelles*, *Società di Mutuo Soccorso*), which provided their members with cash benefits and medical treatment. National mutual aid society legislation was enacted in Britain in 1793, Spain in 1839, Belgium in 1849, France in 1834, Italy in 1886, Sweden in 1891, Denmark in 1892, Switzerland in 1911. Early laws—those passed before 1850—were most concerned with monitoring the activities of these popular associations. After 1850, government legislation often included funding for the mutual aid societies. In the late nineteenth and early twentieth centuries, lawmakers made the subsidies conditional on improvements in insurance coverage and medical benefits, requiring, for example, that funds accept all applicants and that they provide access to medical treatment, rather than only the more common cash benefits, which were to cover income-loss during illness.

After Germany introduced compulsory public health insurance (*gesetzliche Krankenversicherung*) for industrial workers in 1883, many other European nations followed. Austria introduced compulsory health insurance in 1888, Luxembourg in 1901, Norway in 1909, Britain in 1911, the Netherlands in 1913 (for cash benefits only) and in 1943 (for medical benefits), France in 1930, Greece in 1937, Spain in 1942, Italy in 1944, Belgium in 1944, Portugal in 1946, Sweden in 1946, Finland in 1964, Denmark in 1971. Early compulsory health insurance laws often built upon the basic structures provided by the voluntary mutual funds, leaving health insurance administration to the preexisting funds, but making membership for certain population groups compulsory.

The politics of mutual aid society and compulsory health insurance legislation are diverse. Many different types of political parties, spanning the political spectrum, introduced legislation intended to expand health insurance coverage. In France and Germany, conservatives proposed mutual aid society legislation against the opposition of liberals, whereas in Sweden and Switzerland, for example, the situation was just the reverse. German socialists opposed conservative Chancellor Bismarck's compulsory health insurance, but, as they found they were able to use the sickness funds as an organizational resource, they later came to champion the insurance system. Indeed, it is precisely such alliances between political parties or social movements and the sickness funds that explain some of the partisan variation regarding statutory (that is, required by law) health insurance. Whereas in Germany, the territorial sickness funds (*allgemeine Ortskrankenkassen*) became a social democratic and union stronghold, in Belgium, the sickness funds were closely allied to the liberal, Catholic, and the socialist political parties. In the Netherlands, many sickness funds were started by doctors; in Sweden, they were tied to the Temperance (anti-alcohol) and Free Church (opposed to the State Church) movements, which, in turn, tended to be supporters of the liberal party, as were many Swiss sickness funds.

Some of the most important political conflicts surrounding nineteenth century and early twentieth century legislation concerned the issue of whether insurance coverage should be voluntary or compulsory, whether employers or workers or both should pay contributions and in what amounts, what proportion of the population—often defined in terms of occupational groups—should be covered, and how to achieve portability of insurance benefits. These issues tended to be of more direct concern to labor unions, employers, and the mutual aid societies than to doctors, and can thus perhaps be better viewed as 'class' issues than 'professional' issues. Not surprisingly, studies of this early period—which include contributions from history, sociology, and politics—have tended to be concerned with the causes and consequences of the rise of social insurance, and of the welfare state, more broadly. A key question has been whether the rise of working-class organizations and parties, as well as differences in working-class political mobilization—the degree to which members of the working-class actually voted for workers' parties or joined trade unions, whether these organizations were divided by religious cleavages, and whether parties and trade unions cooperated—can explain the emergence of different programs of social protection in different European nations. A second line of questioning has been whether fear of class unrest is the sole motivation for government initiatives in this period, or whether states had other considerations related to their need for political legitimacy, particularly as parliamentary democracy came to be established in some nations but not in others (see *Class: Social*; *Christian Parties: European*; *Employers' Associations*; *Labor Unions*; *Social Movements: Resource Mobilization Theory*; *Social Democracy*; *Labor Movements, History of*; *Social Question: Impact on Social Thought*; *Social*

Security; *Social Welfare Policy: Comparisons*; *Socialism*; *Voluntary Associations, Sociology of*; *Welfare*; *Welfare State*; *Welfare State, History of*).

1.3 The Interwar Period: Doctors vs. Sickness Funds

As the growth of both voluntary and compulsory health insurance expanded the market for medical care in the first third of the twentieth century, conflicts arose between doctors and both mutual aid societies and government insurance authorities. Whether public or private, the insurance carriers (those responsible for administering health insurance) sought to reduce the costs of medical treatment by restricting the fees that doctors could charge to insurance patients. These insurance conflicts coincided with doctors' achievement of professional status, and encouraged the efforts underway to organize members of the profession into effective lobby associations (see below). Now health, politics had indeed become very much characterized by professional politics, and scholars studying twentieth-century health legislation have focused largely on the opinions, strategies, and success of organized medical professions in shaping government health programs to their liking.

For medical organizations, the key issue was to avoid becoming economically dependent on health insurance practice, and, especially, to maintain the status of doctors as independent practitioners in face of efforts to turn doctors into employees of the sickness funds. In Germany, the sickness funds relied on individual contracts with doctors and direct third-party payment (i.e., the sickness funds paid the doctors directly) to keep fees low; only doctors that agreed to the conditions of the funds were admitted to insurance practice. This type of system is also known as a 'closed panel' or, in contemporary language, as a 'preferred provider' health insurance system. When the German system was to be extended to larger sectors of the population through the *Reichversicherungsordnung* (Imperial Insurance Code) of 1911, the two main doctors' associations, the *Leipziger Verband* and the *Deutsche Ärzteverein*, led a successful strike that culminated in the *Berliner Abkommen* (Berliner Accord) of 1913. This agreement established a bipartite board of representatives of both the medical profession and the sickness funds to govern entry to sickness fund practice, thereby ending the funds' unilateral control of access to the insurance market.

In Germany, doctors fought within an established insurance system for ever better conditions of practice. Other medical associations looked to the German example, and fought for these measures from the start. In France and Switzerland, for example, a series of quite similar issues occupied health insurance debates from the beginning of the century through the 1960s. Doctors preferred a fee-for-service payment method (which allowed doctors to charge patients a separate fee for each service performed); they opposed third-party payment (*tiers payant*) and wished to receive payment directly from patients, leaving the patients to submit their bills to the sickness funds for reimbursement (*tiers garant*); medical associations were willing to negotiate collective fee schedules (lists of maximum fees, called *Gebührenordnungen* or Conventions) with insurance authorities, but they were opposed to individual contracts and asserted the individual discretion of physicians to charge more in difficult cases, or for all cases, if the physician was particularly well-known. Realizing that these conflicts were inherent in insurance practice, medical associations fought also against the extension of compulsory health insurance systems to entire populations, preferring instead that government or government-subsidized insurance remain voluntary, or at least compulsory only for those with low incomes. Insurance carriers, on the other hand, did not wish to be responsible for undefined and uncontrolled costs, and lobbied for government regulation of medical fees, and the extension of compulsory coverage.

1.4 The Postwar Period: National Health Systems

After the Second World War, most Western European governments massively expanded their health insurance coverage, many insuring entire populations as a right of citizenship. Not only health insurance, but the health system as a whole was the subject of reform efforts. In 1946, a Labour government introduced the British National Health Service (NHS) to provide healthcare to all citizens through publicly-owned hospitals and publicly-paid doctors. In order to overcome opposition from the British Medical Association—which largely represented the interests of the general practitioners—plans for salaried practice for outpatient care was dropped; instead general practitioners would be under contract to the NHS, and paid on a capitation basis (a yearly flat-fee per patient enrolled with a particular doctor, regardless of how much or how little treatment is received), as they had been by the Approved Insurance Societies (Friendly Societies approved to carry national health insurance), since the National Insurance Act of 1911. Similar ideas of universal provision of healthcare through salaried hospital doctors and national health hospital plans were discussed in other nations as well, such as France (in 1944) and Sweden (in 1948), but failed due to medical opposition.

In a second wave of reform in the 1960s and 1970s, however, many European governments introduced controls on doctors' fees, salaried service for hospital physicians, and national plans for health systems. In France, the decrees of May 12, 1960, promulgated by de Gaulle, established a system of departmental conventions (lists of maximum fees) for the control of

doctors' fees, which was replaced by a national convention in 1971. The Debré reform of 1958 established new mechanisms for hospital planning, created full-time salaried posts for hospital physicians, called for phasing out the private practice privileges of hospital chiefs of staff, and reformed medical education. In Sweden, a Social Democratic government banned private medical practice from public hospitals in 1969, and, in conjunction with the reform, hospital physicians agreed to be paid on a salary basis. In Italy, an agreement between the Italian Communist Party and the Christian Democrats established the Italian National Health Service (*Servizio Sanitario Nazionale*) in 1978. In 1986, the Spanish socialist government converted the national health insurance to a national health service, as well.

Even in nations where no formal changes in type of health system were made, efforts were made to introduce more effective measures of cost-containment, as well as hospital financing and planning. The German Cost-Containment Law of 1977 (*Krankenversicherungs-Kostendämpfungsgesetz*) provided for tripartite discussion of healthcare costs amongst representatives of government, sickness funds, and the medical profession, and for limits on the collective reimbursement made from the sickness funds to health insurance doctors. Subsequent reform efforts (most importantly in 1981, 1992, and 1997) have focused on further improvements in cost-containment and in extending these mechanisms to the hospital sector. In the Netherlands, too, legislation governing healthcare prices and hospital planning was introduced in 1982 that built upon a preexisting framework of negotiations between sickness funds and healthcare providers.

Since the 1990s, one can observe a shift in emphasis from direct government regulation of health prices and planning to experimentation with 'market' reforms. In the Netherlands, reforms have aimed to increase competition amongst insurance carriers. In the United Kingdom and Sweden, efforts have been made to introduce 'internal markets' and to increase 'public competition,' by creating mechanisms for the purchase of health services by local healthcare authorities and general practitioners (in the UK), and units of local government (in Sweden), (see *Health Care Markets: Theory and Practice*; *Health Economics*; *Public Management, New*).

2. Health Governance

2.1 Corporatism

The power of the medical profession is a central theme not only in studies of government efforts to introduce and to reform public health programs, but also in research on the effectiveness of different health policies. Here, the theory of corporatism has proved to be a useful organizing principle for comparing government health care strategies and their effects (see *Corporatism*). Proponents of a corporatist approach examine the ways in which medical doctors are organized in different countries. In Germany, for example, all doctors treating public insurance patients are required to belong to Associations of Sickness Fund Doctors, which in turn are represented by a peak association (*Kassenärztliche Bundesvereinigung*). They are thus not only highly organized, but organized into a single association. This legacy of the early days of German health insurance has allowed the government to leave many aspects of health policy—such as cost-containment measures—to arenas of collective negotiation between representatives of the sickness funds and the medical profession. Consequently, some policy analysts view negotiation—or governance through association—as an attractive alternative both to direct government intervention ('State') or purely economic mechanisms ('Market').

Scholars investigating the impact of government intervention in the medical area have also raised the question of the meaning of health reforms for the freedom and power of the medical profession. Is the right to a private practice an integral part of medical freedom—as many doctors' associations have argued—or is the defense of the private market an outmoded vestige? Although many different aspects of the resources and privileges of doctors have been subsumed under the general term 'power,' investigations of the health area indicate that there may be several dimensions of medical power—not all of which vary together.

2.2 Theories of the Professions

Theories of the professions—an offshoot of sociological research on stratification—try to explain why some occupations are more privileged than others. Debate has concerned both the indicators of 'privilege' and its causes. The 'functionalist' approach views the essence of professionalism in the demonstration of the efficacy of a particular body of knowledge which generates public legitimacy for the autonomy of judgment of its practitioners. The 'historicist' approach, by contrast, argues that the emergence of a domain of knowledge and the association of that domain with an occupation grouping is an arbitrary and political process which can only be understood historically. The health field has proved a fertile ground for this debate, because of the apparent unparalleled success of doctors in this 'professionalization project.' Western European studies have advanced the historicist argument, in particular, as research has detailed the great variation in the boundaries of professional knowledge, the organization of medicine, and the character of medical privilege across Europe (see *Func-*

tionalism in Sociology; *Historicism*; *Medical Sociology*; *Medical Profession, The*; *Parsons, Talcott (1902–79)*; *Professionalization/Professions in History*; *Professions, Sociology of*; *Health Policy Social Stratification*; *Social Inequality in History (Stratification and Classes)*).

Theories of the professions emphasize the process by which the medical profession became unified into a single, recognizable profession, and how members of this profession were able to upgrade their collective status, achieving recognition from both governments and the public for their claims to expertise. Debate continues, however, about the extent to which health expertise translates into success in the political arena. If one compares the success of medical professions in defending their right to set their own fees, and to an unregulated private practice—their 'economic autonomy,' so to speak—variation in Western Europe does not appear to be accounted for by variation in professional monopoly. In Switzerland, for example, government regulation of medical licensing, and even against medical quackery, has been relatively weak, yet the Swiss Medical Association has been highly successful in blocking the introduction of compulsory national health insurance and controls on doctors' fees. An alternative explanation for such variation may lie in the properties of political institutions and their effects on political decision-making (see *Institutionalism*). Professional associations—in particularly those representing medical specialists—appear to be more universally successful in other areas of decision-making, however, such as individual decisions about treatment (clinical autonomy), or political debates about more purely 'medical' matters (such as funding of medical research or issues of medical malpractice). Whether an issue becomes defined as 'medical,' 'economic,' or 'social,' however, is very much a matter of politics, and hence at the core of professional strategies.

2.3 States and Social Control

Following the lead of Norbert Elias and Michel Foucault—as well as, of course, the classical sociologists Durkheim and Weber—scholars have tried to reconstruct the meaning of health intervention for the relationship between state and society (see *Durkheim, Emile (1858–1917)*; *Elias, Norbert (1897–1990)*; *Foucault, Michel (1926–84)*; *Weber, Max (1864–1920)*). The idea is to understand both why states become involved in regulating the health of citizens, and how beliefs associated with concepts like health, disease, and contagion are formed. The direction of analysis is not only from state to society, but examines how social practices involving health and hygiene—for example, the set of manners that demarcate the boundaries of appropriate public behavior—establish culturally-specific concepts (also called 'collective representations') of state and society. In this way, the 'politics of the body' are thought to be construct images of the body politic, and hence to influence citizens' attitudes and behaviors vis-à-vis social collectivities and political authorities. Consequently, work in this vein has interpreted health beliefs and health practices as part of a system of social and political domination. This type of 'constructivist' or 'postmodern' approach has often been adopted by proponents of a 'gender' approach. Here, ideas about the body, and the relationship of doctors to patients of both sexes are thought to play a role in establishing definitions of gender differences (see *Control: Social*; *Conventions and Norms: Philosophical Aspects*; *Deconstruction: Cultural Concerns*; *Gender and Feminist Studies*; *Gender and Feminist Studies in History*; *Gender and Health Care*; *Gender History*; *Body, History of*; *Norms*; *Postmodernism in Sociology*; *Postmodernism: Methodology*; *Social Constructivism*).

2.4 State Sovereignty

Finally, studies of healthcare governance have been part of a more general debate about the rise and future of the nation–state. Healthcare has been used as an example by which one can observe the origins and development of the institutions and mechanisms of governance peculiar to Western Europe. The complexities of governing the healthcare sector have provided empirical substance to a more general debate about whether classical state sovereignty is on the wane. The growth of the welfare state—precisely through the expansion of programs like national health insurance and national health services—has involved governments increasingly in negotiations with social actors—medical professions, labor unions, employers, and the like—thereby, in the eyes of some scholars, eroding states' hard-won 'internal sovereignty' (the independence achieved by absolutist rulers from medieval corporations, estates, cities, or other subnational political actors). The international integration of national economies, and, especially, the regional integration of the Western European economies is, analogously, thought by some to threaten states' 'external sovereignty' (the autonomy wrested by modern European nation–states from supranational authorities, such as the Catholic Church) (see *Absolutism, History of*; *Luhmann, Niklas (1927–98)*; *Nations and Nation-states in History*; *Rokkan, Stein (1921–79)*; *State Formation*; *State, History of*; *State and Society*; *State, Sociology of the*).

Whereas the problems of 'domestic' health politics and health governance have been investigated quite thoroughly, far less research has focused on the emerging problems of international or, especially, European health governance. It will be a challenge of future research to refine the concepts and issues that have emerged from national contexts (albeit in comp-

arative perspective)—such as the medical profession, professional autonomy, corporatism, national healthcare systems—in light of international cooperation and integration. Further, while a focus on Western Europe has promoted interdisciplinary research—with political scientists and historians, for example, using sociological theories of the professions as a departure for their own research—an equally pressing challenge will be to look beyond the borders of Western Europe for alternative approaches to the definition and study of 'health.' Lastly, as the study of health, like many policy areas, becomes increasingly influenced by an economics perspective—with market mechanisms favored by policy-makers and economic models of decision-making by social scientists—the remaining social and behavioral sciences will need to articulate compelling arguments for their continued contribution to the study of health.

Bibliography

Alber J, Bernardi-Schenkluhn B 1992 *Westeuropäische Gesundheitssysteme im Vergleich—Bundesrepublik Deutschland. Schweiz, Frankreich, Italien, Großbritannien*. Campus, Frankfurt, Germany

Baldwin P 1999 *Contagion and the State in Europe, 1830–1930*. Cambridge University Press, Cambridge, UK

Burrage M, Torstendahl R (eds.) 1990 *Professions in Theory and History: Rethinking the Study of the Professions*. Sage, London

Flora P A 1986 *Growth to Limits: The Western European Welfare States since World War II*, 3 vols. Walter de Gruyter, Berlin

Foucault M 1963 *Naissance de la Clinique*. Presses Universitaire de France, Paris [1973 *The Birth of the Clinic*. Pantheon Books, New York]

Freddi G, Björkman J W (eds.) 1989 *Controlling Medical Professionals: The Comparative Politics of Health Governance*. Sage, London

Hafferty F W, McKinlay J B (eds.) 1993 *The Changing Medical Profession: An International Perspective*. Oxford University Press, Oxford, UK

Hollingsworth J R 1986 *A Political Economy of Medicine: Great Britain and the United States*. Johns Hopkins University Press, Baltimore, MD

Immergut E M 1992 *Health Politics: Interests and Institutions in Western Europe*. Cambridge University Press, Cambridge, UK

Marone J A, Goggin J M (eds.) 1995 European health policies: Welfare states in a market era. *Journal of Health Politics, Policy and Law* **20**: 557–785

Moran M 1999 *Governing the Health Care State: A Comparative Study of the United Kingdom, the United States and Germany*. Manchester University Press, Manchester, UK

Köhler P A, Zacher H F (eds.) 1981 *Ein Jahrhundert Sozialversicherung in der Bundesrepublik Deutschland, Frankreich, Großbritannien, Österreich und der Schweiz*. Duncker und Humblot, Berlin [1982 *The Evolution of Social Insurance. 1881–1981. Studies of Germany, France, Great Britain, Austria and Switzerland*. Frances Pinter, London and St. Martin's Press, New York]

Ramsey M 1984 The politics of professional monopoly in nineteenth-century medicine: The French model and its rivals. In: Geison G (ed.) *Professions and the French State, 1700–1900*. University of Pennsylvania Press, Philadelphia, PA

de Swaan A 1988 *In Care of the State: Health Care, Education and Welfare in Europe and the USA in the Modern Era*. Polity Press, Cambridge, UK (Blackwell, Oxford, UK)

Tennstedt F 1976 *Sozialgeschichte der sozialmedizin*. In: Blohmke M V, Ferber C, Kisker K P, Schaefer H (eds.) *Handbuch der Sozialmedizin*, Vol. 3. Ferdinand Enke Verlag, Stuttgart, Germany, pp. 385–492

E. M. Immergut

Western European Studies: Religion

1. Church and State

The study of Western Europe has been until modern times the study of Christendom, and the evolution of fundamental western institutions such as the state and civil society cannot be adequately understood without an analysis of the Church and its institutions. Historical studies of European institutions have been necessarily preoccupied with the development of relationships between the state and the church. During the Roman Empire, the Christian community had been a cult whose millenarian doctrine sharply separated the religious community from the secular world. Eschatological expectations about the Second Coming meant that Christian leaders had little interest in shaping this world. The conversion of Constantine, the foundation of Constantinople as a Christian city and the acceptance of Christianity as the official religion of the empire brought about the end of Christian chiliasm. In the late fourth and fifth centuries, the papacy emerged as a governmental institution that exercised this-worldly rulership.

The collapse of the Roman Empire and the advance of barbarian kingdoms in the fifth and sixth centuries created the foundations of feudalism. As Christianity came to adjust to the existence of the secular institutions of this world, such as kingship, it was forced to reconsider the relationship between the church and the state. Medieval Christian theology provided basic models of power, legitimacy, corporation, and immunity that have been, and remain an influence in western political institutions. Throughout the medieval period, Christian theologians struggled to reconcile two problematic principles: the state was an institution of violence, but it was necessary for social order, and the Church, whose foundation was a divine act, depended upon the state to secure its soteriological purpose. What was common to medieval belief and what came to shape the growth of constitutional theory was the assumption that power should exist within a normative framework (Canning 1996).

The medieval Church inherited the impulse of primitive Christianity to disengage from political life,

but this impulse could not be comprehensively applied since the Church was ultimately dependent on state institutions (Wolin 1961). The Protestant Reformation created a new vision of independent, self-governing communities in which the sanctified individual would be liberated from the tyranny of secular powers and the corruption of ecclesiastical institutions. It was a vision of the transformation of power to godly rule that inspired the American Revolution. Lutheranism embraced a nostalgic vision of the purity of the primitive Church with which to challenge Catholic institutions, but it came to recognise that the congregations of the 'visible Church' would need discipline in order to prepare for the 'invisible Church.' The sins of this world were the pretext for a theory of the state as a moral authority, but Luther went further to recognise that the state was necessary to release Christian souls from the tyranny of the Church in the world. The fundamental instability of this relationship between the sovereignty of the state and the divine mission of the Church produced an endless cycle of reformation and counter reformation that structured political life in Europe for centuries.

Sociologists have studied the pastoral counsel of Reformation Protestantism as an important cultural foundation for the growth of capitalism (Weber 1930). Although the 'Protestant Ethic thesis' is controversial, there is evidence that Protestant divines contributed both to the ethos of entrepreneurial capitalism and the social conditions that promoted the rise of democracy. Catholicism has therefore been seen as a conservative tradition whose opposition to Protestant rationalism gave rise to the Counter Reformation and to Baroque culture and institutions. The stark interiors of northern Protestantism were clearly demarcated from the exuberant and luxurious interiors of southern Catholic Europe.

The authority of the Church universal was further compromised by the growth of nationalism and the nation-state. The Peace of Westphalia (1648) produced a political system of nation states and forced the churches to operate within national boundaries. The French Revolution created the modern notion of national citizenship and a secular republican tradition that profoundly challenged European Catholicism. Modernization has been necessarily associated with secularism and the growth of scientific criticism of revelation as the foundation of Christian belief. Radicals like Marx and Engels regarded socialism as a necessary condition for the effective criticism of religion as part of the ideology of a dominant class. In addition, the growth of secularism had paradoxical consequences for European Judaism by offering Jews the possibility of assimilation, when the French National Assembly granted civil rights to Jews in 1791. Marx had argued that 'the Jewish Question' could not be solved merely by the political emancipation of Jewry but only by a social transformation of society. The spectre of assimilation broke the Jewish community into reform and counter-reform traditions, but European anti-Semitism continued to exclude Jews from public life and laid the foundations for the Holocaust. Mass emigration of Jews to the USA has profoundly changed the religious demography of Europe.

The dramatic decline of institutional religion in the twentieth century gave rise to the secularization thesis in the 1960s (Wilson 1966). The contrast between the USA where religion appeared to flourish and the secularity of most European societies was a topic of considerable interest. In America, the consequence of post-war migration was to convert religious identity into a sign of American membership (Herberg 1955), whereas in Europe religious affiliation was more persistently associated with class membership (Thompson 1963). Contemporary research on religious institutions in Europe has been less triumphal in its assessment of secularization in recognising the continuing importance of religion for what sociologists have called 'collective representations' (Durkheim 1954).

Whereas post-war studies of religion were preoccupied with the study of church–state relations, the Protestant–Catholic divide, and the survival of religion under communism, contemporary research has been interested in the impact of globalization, the growth of non-Christian religions and the rise of new religious movements. While the Christian churches have institutionally declined, religious cultures continue to flourish in European society.

2. *Secularization Thesis*

It has been a basic assumption of the sociology of religion that the Christian churches in Europe have, from the nineteenth century, been subject to a profound and ineluctable process of secularization. While there is clear evidence of secularization in the sense that membership of, and participation in, Christian churches have declined, religious identity continues to play a generic role in national identity and consciousness. In *The Social Teaching of the Christian Churches*, Troeltsch (1931) argued that the oscillation between church and sect that had shaped much of European history had come to an end with the final erosion of the universal church. Religious life would become a matter of private belief and practice. The growth of an urban industrial society had undermined the social, cultural, and intellectual conditions that made religious attachment and belief possible (MacIntyre 1969). While sects continue to flourish, there is incontrovertible evidence of the institutional decline of mainstream Christianity (Wilson 1976). Within this general pattern of decline, there are however discernible differences between the predominantly Catholic and Protestant regions and states

(Martin 1978). Catholicism has remained a dominant cultural force in Italy, France, Spain, Portugal, and Poland. In France an estimated ninety percent of the population is Roman Catholic and two percent is Protestant. In Ireland, 92 percent are Catholic and 80 percent attend church weekly.

Catholicism, prior to political liberalization in the late twentieth century, was central to the expression of nationalism in continental Europe and remained a major counter force to communism. Recognition of the importance of Catholicism for moral leadership of the working class produced a notion of 'hegemony' in the study of European politics (Gramsci 1971). The social and cultural influence of Catholicism has been closely associated with its control over education. In Ireland, for example, while the state provides over 85 percent of funding, Catholic boards of managers generally control schools. The dominance of the Catholic Church on the European right guaranteed that regional, party, and class divisions were often drawn along religious lines.

After the Second World War, Catholicism also played an important cultural and political role in relation to atheist communism. The Polish Solidarity movement decisively demonstrated the capacity of Catholicism to survive communism and to act as a conduit of social protest and change, because it is embedded in national identity. Irish national identity and republicanism have also been thoroughly merged within a Catholic tradition. Despite the Catholic Emancipation Act (1829), the Protestant–Catholic divide in Northern Ireland has remained an obstinate fact of political life. Owing to the political alienation of the population to the pre-1922 British state, the majority equate being Catholic with being Irish. With the end of the Cold War and the fall of communism, Catholicism may play a diminished role in the articulation of nationalism and national identity. In Spain, General Franco rose to power in 1936 following his attack on the socialist government. Franco's regime was decidedly Catholic and supported traditional values against godless atheism, but the collapse of the Franco regime following his death in 1975 has resulted in the diminution of the public authority of the Catholic Church. Economic prosperity and growing multiculturalism in Eire has also brought about the beginning of a partial divorce between state and church.

In Protestant societies, the relationship with the state has been more remote, and hence the political influence of the churches has been obscured. While the Catholic Church has largely resisted Protestant infiltration in France and Italy, the Protestant countries have been religiously more diverse and the churches have enjoyed a privileged rather than a monopolistic position. In the Lutheran traditions of Scandinavia, the churches have been incorporated into the state, where religious functionaries became a component of the state bureaucracy. In Norway, Article 2 of the constitution proclaims both the existence of religious freedom and recognises the Evangelical Lutheran religion as the official religion of the state. In practice, the separation of church and state is recognised, despite the fact that the cabinet has the right to appoint bishops. In the UK, the Church of England functioned as a national church with the monarch as head, but religious tolerance and pluralism have been accepted principles of Lockean liberalism. In the UK there has been no real equivalent to Bismarck's *Kulturkampf* to secure cultural homogeneity. The political transition from a confessional state to religious pluralism has been typical of English political gradualism in which discriminatory laws against Catholics and Jews were pragmatically abandoned rather than explicitly rejected by an assertion of religious equality. While the devolution of powers to regional parliaments in Scotland and Wales has weakened the political significance of the Church of England, Anglicanism will no doubt remain an important ingredient of Englishness. The UK has been a society of 'tolerant discrimination' rather than a fully-fledged melting pot in the American tradition (Boyle and Sheen 1997).

Prior to the unification of Germany in 1990, religion played an important part in defining the cultural differences between East and West Germany. In the united Germany, the constitutional right of return to Germany by ethnic Germans (*Aussiedler*) has insured cultural homogeneity and implicitly reinforced the role of religion as an essential element of being German. Around eight percent of the population are foreigners who are not permitted citizenship, and their children may have difficulty in obtaining naturalisation. The Weimar Republic rejected a state church, but Hitler destroyed the liberal constitution in 1933 and established National Socialism on the basis of one people, one state, and one leader. In the 1949 constitution, the separation of church and state was recognised, but religious education in state schools was a constitutional obligation. In contemporary Germany, approximately 45 percent of the population are members of the Evangelical Church and around 37 percent are Catholic. Because Christian traditions play an important part in national identity, there is some tension between Catholics and Protestants (for example over Christian symbols in state schools).

3. *Migration and Multiculturalism*

The history of religion in western Europe has been dominated by two issues, namely church–state relations and secondly the cultural divisions between Protestantism and Catholicism. This historic pattern has been shaken by two inter-related processes, namely the secularization of society and the globalization of the European economy. Post-war European economic prosperity has combined with a greying population to

produce a multicultural society that satisfies its labor market needs by migrant labor. The working-age population of Europe is declining rapidly and by 2030 it is estimated that the ratio of working taxpayers to nonworking pensioners in Germany and Italy will drop to 1:1. Young migrants, whose fertility rate is typically higher than the host population, are filling the gap between workers and non-workers. This structural dependency on foreign labor has drawn in significant numbers of non-Christian migrants, whose presence is permanently changing the cultural map of Europe. In many European countries, the foreign migrant community is 10 percent of the host population. The most significant group, both in numbers and influence, is Muslim. There are 10 to 13 million Muslims in Europe, and in Germany the foreigners will make up 30 percent of the population by 2030, most of whom will find employment in cities such as Munich and Frankfurt.

Global aging and labor shortages in the developed world will insure that immigration and religion remain on the political agenda of European societies. Migrants from Pakistan to the UK, from Turkey to Germany, from the Middle East and North Africa to France, and from Indonesia to the Netherlands have produced a diasporic politics that has raised fears about the impact of Islamic fundamentalism on democratic institutions in Europe. In Germany, Turks and Kurds entered the labor market in the 1960s and 1970s, and these 'guest workers' are now a consistent more or less permanent second generation, amounting to approximately two million people. While many of these migrants are secular, Islamic organisations play an important part in their social and political organisation. In France, there is a strong nationalist feeling that North African Muslims (*maghrebins*) cannot assimilate to French republican culture. The *hijab* case (*l'affaire des foulards*) in 1989 caused a divisive public debate about the desire of Muslim girls to wear the *hijab* in state schools. The French intellectual Left regard secular schools as important for personal liberation from religious ideology, while the Right interpret the hijab as an attack on French national custom. In English culture, there has been an established tradition of distrust towards Islam that can be illustrated for example by hostility to translations of the Qur'an in the sixteenth century (Matar 1998). In recent legislation, there has been some accommodation to the beliefs and practices of other religions, such as the acceptance of customs relating to the wearing turbans by Sikhs, animal slaughter, and solemnization of marriages. However, the 'Rushdie Affair' involving the publication of *The Satanic Verses* in 1989 by the novelist Salman Rushdie and the *fetwa* placed on him polarized British public opinion, and reinforced the public perception of Muslims as fanatics whose culture is fundamentally incompatible with parliamentary democracy and liberal values. The Rushdie Affair confirmed the view of many intellectuals that European culture is fundamentally racist and that its world-view is deeply Orientalist in which the Muslim has become 'the Other' (Said 1978). In France, a poll by *Le Monde* in 1989 revealed that about 70 percent of the French population associate Islam with fanaticism. Attempts by Muslim intellectuals living in Europe to counteract these negative stereotypes have had little impact on the general population (Ahmed 1992). The difficulties of finding a peaceful solution to the Middle East crisis and the apparent failure of the Oslo agreement as illustrated by the Palestinian Intifada will continue to fuel hostility towards Islam (Turner 1994).

4. *Conclusion: Religion and Civic Culture*

Latin Christianity had created a common religious and political culture in medieval Europe. The Reformation and the division of Europe broke this dominant culture into competing states with distinctive national religious traditions. The growth of nationalism in Europe had very diverse consequences for the churches, but religious symbols, often combined with epic literature and folk culture, have been indispensable for the creation of nations as 'imagined communities' (Anderson 1983). This vitality of religious symbols in nation formation has been constructed on the basis of conservative hostility to both secularism and communism. The dominant transformations of religion in the twentieth century have been brought about by the secularization of Europe and the decline of communism as a significant atheist alternative to theological belief systems. In the twenty-first century, the religious map of Europe is changing once more as globalization has created cultural hybridisation. The new component of this ideological cartography is Islam, but there are other changes that have come to the attention of social scientists, namely the emergence of new religious movements and alternative expressions of religious consciousness involving syncretism (Hamilton 1995). Europe governments have frequently attempted to curb the development of such sects by legislative means. There is considerable public anxiety about New Age groups, Scientology and Jehovah's Witnesses. In 1995 the British home secretary refused entry to the Rev Sun Myung Moon who had planned to enter the UK to hold services for the Unification Church. In Germany, the federal government has identified 25 sects that are seen to be a threat to 'democratic values.' These religious problems of a multicultural society are now a generic aspect of European politics, and are indicative that the traditional Protestant–Catholic–Jewish divisions of European politics have been further complicated by global hybridity (Robertson 1992). Changes in the nature of the study of religion as a European institution are thus reflections of the growth of global religious cultures.

See also: Development and the State; Western European Studies: Culture; Western European Studies: Society

Bibliography

Ahmed A S 1992 *Postmodernism and Islam. Predicament and Promise*. Routledge, London

Anderson B 1983 *Imagined Communities*. Verso, London

Boyle K, Sheen J 1997 *Freedom of Religion and Belief. A World Report*. Routledge, London

Canning J 1996 *A History of Medieval Political Thought 300–1450*. Routledge, London

Durkheim E 1954 *The Elementary Forms of the Religious Life*. Allen and Unwin, London

Gramsci A 1971 *Selections from the Prison Notebooks*. New Left Books, London

Hamilton M B 1995 *The Sociology of Religion. Theoretical and Comparative Perspectives*. Routledge, London

Herberg W 1955 *Protestant Catholic Jew*. Doubleday, New York

MacIntyre A 1969 The debate about God: Victorian relevance and contemporary irrelevance. In: MacIntyre A, Ricoeur P (eds.) *The Religious Significance of Atheism*. Columbia University Press, New York and London

Martin B 1978 *A General Theory of Secularization*. Basil Blackwell, Oxford

Matar N 1998 *Islam in Britain 1558–1685*. Cambridge University Press, Cambridge, UK

Robertson R 1992 *Globalization Social Theory and Global Culture*. Sage, London

Said W E 1978 *Orientalism*. Pantheon Books, New York

Thompson E P 1963 *The Making of the English Working Class*. Penguin, Harmondsworth, UK

Troeltsch E 1931 *The Social Teaching of the Christian Churches*. Macmillan, New York

Turner B S 1994 *Orientalism, Postmodernism and Globalism*. Routledge, London

Weber M 1930 *The Protestant Ethic and the Spirit of Capitalism*. Allen and Unwin, London

Wilson B 1966 *Religion in Secular Society*. Watts, London

Wilson B 1976 *Contemporary Transformations of Religion*. Oxford University Press, London

Wolin S S 1961 *Politics and Vision. Continuity and Innovation in Western Political Thought*. George Allen and Unwin, London

B. S. Turner

Western European Studies: Society

Scholars of Europe have long been driven by an infrequently stated question: what confluence of social factors accounts for the fact that this small subcontinent came effectively to dominate the world by the end of the nineteenth century? This question predates the advent of 'Western European studies,' but Western European studies have nevertheless had to come to terms with it.

Weber's assertion that Western Europe's position in the world was the product of a unique and unrepeatable set of economic, cultural, and political factors that together produced a novel form of advanced industrial society forms a frame for the present examination of how Western European studies have approached 'society.' The strengths as well as the fault lines of the field are still shaped by Weber's claim, even as the society Weber took as his model has changed almost beyond recognition. Some contemporary scholars go so far as to question whether the very foundations of European dominance—industrial capitalism and the nation–state—can any longer meaningfully be said to exist in a Europe characterized by post-industrialism and supra-national integration. Others insist that both the nation–state and the set of complex social relations that developed along with the rise of industrial capitalism, however modified by late twentieth century transformations, continue to form the basis for a distinctive Western European society.

1. Models of Western European Social Development

Beginning in the mid 1960s, Western Europe's extraordinary world historical success was taken up with renewed vigor among social scientists. This new wave of socio-historical scholarship set itself against earlier modernization literature (see *Modernization, Sociological Theories of*) and its frequently unhistorical assumptions regarding the portability to the developing world of the modern Western nation–state form. It returned to Marx's and Weber's insistence that attention to historical specificity and nuance be drawn into the service of effective theory building.

1.1 Determinants of European Dominance

A catalog of determinative societal features, the unique combination of which is found in Western Europe, includes:

(a) the formation of highly centralized, authoritative, and bureaucratic national states;

(b) the homogenization and integration into nations of disparate regional subcultures that, crucially, were already less stratified than outside Western Europe;

(c) the development of citizenship;

(d) the rise of capitalist industrialization and the attendant destruction of traditional ways of life; and

(e) conflict between aristocratic, commercial/bourgeois, urban working, and rural peasant classes.

1.2 State Building

That the formation of strong states preceded—historically and logically—the construction of coher-

ent nations has become widely accepted. Powerful, centralized states evolved out of the feudal system of 'parcellized sovereignty,' as Weber characterized it. For Anderson (1974), these absolutist regimes of Western Europe represent a 'paradox' of social and political development central to the idea of Western modernity: absolutism protected aristocratic privilege and property, yet also provided the opportunity for rising urban commercial classes to promote their interests. Tilly (1975) elaborates the characteristics of absolutist 'stateness' as the formal autonomy of government, differentiation from nongovernmental organizations (such as the church), centralization, and internal coordination, and agrees that the benefits of strong states were not the monopoly of the states themselves: rights and privileges also accrued incrementally to subject classes. These social and political rights—the unplanned offspring of absolutism—were protected by the increasingly bureaucratic apparatus of states, forced by the logic of their own development to recognize procedural norms. Tilly also highlights how extraction of taxes from the bourgeoisie helped build state structures.

With extraction comes repression. One of Moore's (1966) many contributions is to insist that violence, both on behalf of and against ruling classes, be given its due place in understanding Western state-building processes. Where absolutism eventually led to distinctively Western liberal democratic outcomes, violence was a key part of the process. In England, enclosures eliminated a traditional peasant way of life, and the Civil War beheaded a king. The French Revolution made war on the landed elites. (This model does not hold for all contemporary democracies in Western Europe. As Tucholsky 1973 said in 1930, 'Due to inclement weather the German revolution was held in music.') The winner—ultimately throughout Western Europe—was the bourgeoisie, whose interests no longer had to vie with traditional elites for state protection.

1.3 Nation Building

Strong states only developed into nation–states where social and cultural integration occurred within their boundaries. This process of nation building has been the subject of an enormous and diverse literature (see *Ethnonationalism, Cultural Concerns*). Why did it happen in Western Europe?

Some commentators ask whether this is even the right question, suggesting that a previous generation of scholarship depicting nation building as part of a vast process of modernization exaggerates the depth and extent of national integration. Grillo (1980) argues that there are few, if any, states in Western Europe where integration is complete, and finds the 'European repertoire' from which the concept of nation is constructed to be arbitrary. Rogers (1991) shows that even in France, the archetypical case of a homogeneous, integrated nation, enormous local and regional socio-cultural differences persisted late in the twentieth century.

New research, however, continues to demonstrate that Western European societies are indeed more fully integrated than most others. In a comparative study of demographic integration, Cotts Watkins (1991) looks at how the process of national integration in the nineteenth century was accelerated by the expansion to the national level of previously local networks and institutions, such as transportation, communications, education, and labor organizations. Diverse patterns of fertility began to converge, and in turn spurred the processes of state building and market integration.

Again, why in Western Europe? Anderson (1983) offers an explanation linked to the development and dissemination of national print languages that in their adoption superseded both Latin and spoken vernaculars. These languages, combined with the distributive technologies of capitalism, allowed for dramatically increased communications among nascent bourgeois classes. These classes formed the basis for assimilative 'imagined communities,' or nations: constructed, perhaps even artificial, yet nations nonetheless.

Elias (1939), largely unread and unrecognized until the 1970s, takes a very different approach to the question of European particularity in constructing homogeneous nations. Elias places his startlingly detailed empirical account of the development of manners within a broader theoretical discussion of civilizing and state-building processes that subjected entire classes to new codes of 'civilized' social conduct. By his account, bourgeois civil society could develop within the confines of the absolutist state only because it adopted certain recognizable social features of court society.

1.4 Social Dimensions of Citizenship

The realm of civil society in which citizenship developed is contained within the classical nation-state form. The rights and obligations of citizenship are enjoyed by members of the 'nation,' subject to the structuring influence of the 'state.'

Habermas (1962) makes a distinctly social case for the development of citizenship within a uniquely Western European public sphere. He notes that 'public' was long synonymous with 'the state,' but that the structure of absolutist societies in England and France allowed for a parity of educated persons—bourgeois as well as aristocratic—to engage one another as social equals outside the confines of state structured (court) relations. These social innovations had direct political consequences, but Habermas argues that the public sphere could only emerge

through the impetus of a new form of society that allowed citizenship fully to develop.

The very social bases of citizenship, according to Pateman (1988), rest upon civil society's gendered foundations, firmly embedded in the theoretical premises of the social contract itself. This has far-reaching consequences for the conduct of social life, including derogating the importance of the familial. The establishment of the originary social contract on gendered grounds exposes supposedly 'universal' categories—liberty and equality, for example—as inherently particularistic and misogynist.

Gauthier's (1977) critique of the social contract is even more damning. For Western European societies, the social contract is far more than a theory; according to Gauthier contractarianism is an ideology, a deep structure that orders every aspect of social life. This ideology justifies rights, duties, and institutions, both between persons and between persons and society. The central institution justified by the contract is the market, acquisitive participation in which is the hallmark of Western European humanity. For Gauthier, the central problem with the social contract as ideology is that it is designed both by and for bourgeois men. Not only women, but workers as well have been denied full access to its institutions, and can therefore be regarded as less than wholly human.

2. *Consolidation and Inclusion*

If the socio-historical processes leading to the construction of Western European nation–states were unique, equally so was the nature of the consolidation of those states after two world wars. Western European studies have lavished attention on the nominally inclusive quality of this period. The explosion of industrial capitalist development was more regulated in Western Europe than elsewhere in the First World. Economic prosperity allowed for the institutionalization of social services that would have been considered utopian by nineteenth century social democrats. 'Abstract,' 'exploited' labor, previously subject to the destructive market forces that Marx and Weber agreed typified capitalist society, might now become 'social' labor, secure in a network of meliorating and compensatory social, economic, and political compromises.

The term 'bounded capitalism' is frequently used in Western European studies to denote a distinguishing set of institutional attributes, including unemployment insurance, healthcare systems, social security, industrial policy and relations, macroeconomic policy, national financial institutions, and social bargaining. As Gourevitch (1986) argues, despite important differences among Western European countries in the extent to which they adopted these institutions, bounded capitalism became the norm rather than the exception in the postwar period.

2.1 *The Welfare State*

The postwar Western European social compact built around bounded capitalism, in contrast to the purely abstract universal inclusiveness of social contract theory and the socially exclusive attributes of contractarianism discussed above, promised fully to integrate once excluded classes. Indeed, one of the welfare state's initial successes was to mitigate the demographically disruptive effects of postwar industrialization by absorbing displaced rural populations. In addition, welfare policy helped redistribute social risks from the community and the family to the state.

Modernization theory holds that welfare filled a 'functional' vacuum created by the encroachment of the market into household and community based economies. The functionalist argument continues to find adherents. Kaelble (1987), for example, argues that the generous welfare states of Northern Europe developed out of necessity where older people could not rely on the family for support. Given the extended family structure of Southern Europe, in contrast, less state support was necessary. Modifying functionalism, Korpi (1983) finds that generous social policy developed in Sweden because groups who needed it were able to mobilize themselves in its support. The comparative strength and 'resource power' of national working-class movements, in this analysis, are decisive in determining diverse welfare outcomes. Marxist critiques, in turn, suggest that the origins of mid-twentieth century social policy lay in the conservative cooptation of trade union demands by Bismarck and other late nineteenth century *Realpolitikers*, thereby further legitimating class domination. These origins at the confluence of private capital, wage labor, and state policy, says Offe (1984), are reproduced in persistent structural contradictions between the incompatible needs of capitalist accumulation, social security, and state regulation.

Another approach to the political economy of welfare states situates social policy regimes within an analysis of class coalitions and other institutional factors. Esping-Andersen (1990) delineates three principle types:

(a) 'social democratic' (Scandinavia), which achieved a high degree of 'decommodification' (i.e., protection of citizens from the insecurity of the market) and extensive cross-class support;

(b) 'conservative' (continental Europe), where somewhat lower levels of decommodification were nevertheless fairly comprehensive, and supported by a broad social consensus; and

(c) 'liberal' (Anglo-Saxon countries), where individualist social traditions prevented political agreement on generous benefits.

Rhodes and Mény (1998) add that continental policies were strongly influenced by church and social reformers. An important contribution of such approaches is to suggest that, during the heyday of

welfare capitalism, the social contract was effectively rewritten in some parts of Western Europe, and only modified in others.

2.2 Interest Representation

The literature on interest representation that dominated scholarly discourse from the 1950s to the 1970s, based on an Anglo-Saxon pluralist model, took interests, their representation, and the mediating role they played in the social structure as given. The pluralist model, derived from liberal individualist principles, assumed that all interests were theoretically equal, and equally capable of being represented within the social system. More recent studies, based on European evidence, have tended to problematize interest formation and representation.

An important result has been the corporatist model of interest representation (Schmitter 1979). Corporatism describes forms of associational behavior in which organizations are less competitive than in pluralism, and where certain types of interests—especially labor and capital—are status prioritized within the system of interest representation. The corporatist model is most useful in describing the inclusive, mid-century phase of Western European development. Its applicability to late twentieth century social structures is significantly more controversial (see below, Sect. 3).

For Touraine et al. (1984) the conflict between the workers' movement and the interests of capital lies at the very roots of European modernity. If market capitalism is characterized in part by its socially disintegrative force, then the outcome of the labor/capital conflict will alter—for better or worse—the structure of social solidarities and stratifications. Until the economic crisis of the 1970s, Western European labor and capital got along better than in any other part of the world. To the extent that this was achieved largely outside the bounds of a formal plebiscitary process, some argue that the outcome came at the expense of full democracy. As Berger (1981) puts it, interest groups in Western Europe perform crucial functions that were once the exclusive province of political parties and governments, which suggests a fundamentally new and expanded role for 'society' as against markets and politics. If this was once so, the economically and politically driven process of European integration has since altered the structure of interest representation.

2.3 New Social Movements

Another way to look at the putative rise of society is as a fusion of the political and nonpolitical spheres. The so-called New Social Movements (NSMs) that began in Western Europe with student unrest in the 1960s moved a wide array of 'social' issues into the 'political' realm, and mobilized new classes and generations of social-political actors.

NSMs are considered by most commentators to be new above all in comparison with the classical social movements, especially labor. These older movements were defined by their relation to the power structure (disadvantaged), their goals (class benefit), and their methods (institutional pressure) (Tarrow 1989, Tilly 1978). NSMs, in contrast, often work outside formal channels of political pressure, are driven by profuse mobilizing agendas not necessarily linked to the direct advantage of the social actors themselves, and are frequently made up of the economically and socially privileged.

NSMs are seen as reformist and parochial, without the socially transformative potential that labor represented. Offe (1987) sees the movements emerging out of growing affluence and the interests of social actors 'peripheral' to the labor market whose mobilization is enabled by the generosity of the welfare state. Tarrow (1989) suggests that the fused social/political structures of the 1970s and 1980s created opportunities for the expression of new grievances. Habermas (1981) sees NSMs as a defensive reaction to the encroachment into private life of threatening capitalist and bureaucratic structures. Kreisi et al. (1995) argue that 15 years' worth of NSM activity corresponds directly to altering party alignments in different countries, that is, the movements arose due to political opportunity structures. Inglehart (1990) sees these opportunity structures as opening because of intergenerational change and 'postmaterialism.'

3. Decline of the Nation-state?

The coherence and authority of European nation-states, and the way in which social life is ordered within them, have been subjected to powerful tests since the 1970s. Restructured economies, the decline of the labor movement, European integration, immigration, welfare and demographic crises, increased social inequality, and a host of other challenges have assaulted the stability of the boom years and eroded the social compact that characterized postwar bounded capitalism.

Some scholars ask whether the structural integrity of the nation–state, and the modified social contract premised upon it, can survive such supranational pressures. Wieviorka (1992) suggests that three basic components of Western European social life—industrial society, an egalitarian state, and national identity—have become distorted to such an extent that dangerous obverse tendencies are gaining ground: 'postindustrialism' has produced social selfishness and declining solidarities; state institutions find it difficult to act on egalitarian principles due to fiscal pressures; and nationalism has become increasingly linked with xenophobia. Others, most notably Milward (1992),

argue that Western European responses to recent challenges, at both the national and supranational levels, have actually strengthened the nation–state's social legitimacy and authority.

While the denationalization of economic production and the decline of Fordist industrial structures may or may not have produced a 'postindustrial' society in Western Europe, it is widely agreed that exogenous change in the form of European integration and globalization have reduced the abilities of national governments to effect socially bargained, market-correcting economic policies and to counteract increasing social inequality. The extent to which these changes have undermined the bases for corporatism is the subject of a lively debate (Scharpf 2000).

3.1 Crisis of the Welfare State

Welfare states throughout Western Europe have cut their benefits, and face fiscal and demographic crises of varying degrees. Aging populations create a shift from system contributors to recipients, deindustrialization and 'Eurosclerosis' have produced high unemployment, and the labor market has been transformed by the growing proportion of nonstandard work and the mass entrance of women into the workforce.

Central to recent literature on the welfare crisis are issues of gender and family. Feminist scholars have criticized the political economy approach to welfare state studies because in concentrating its attention on the state and market, it ignores the family and the structural ramifications of women's unpaid domestic work, and in focusing on 'equity' without considering gender, it falsely universalizes equality. At an even more fundamental level, Orloff (1993) asks whether welfare state scholarship is correct to assume that welfare states actually promote overall social well-being and security, since its premises are putatively sexist.

Because of two trends, even 'mainstream' Europeanist welfare state scholars must now acknowledge the issue of women's role in the labor market and the family. First, many women entering the workforce take part-time or fixed term employment, and are penalized because benefits are geared toward men working full time (Rhodes and Mény 1998). Nonstandard work is also becoming more common among men. Second, fertility rates have plummeted since the 1960s. Rates are lowest in countries with low levels of female employment and stingy welfare benefits (e.g., Italy and Spain). From a rationalistic economic point of view, argues Becker (1981), families decide to have fewer children as it becomes more expensive to raise them. By extension, in Scandinavia, where the welfare state is supportive of women's work and child rearing (through day-care facilities, healthcare, etc.), fertility rates have not dropped as dramatically through the late 1990s.

3.2 Immigration and Exclusion

As populations age and fertility declines, immigration becomes increasingly crucial to Western European societies. Since at least the 1960s, 'guest workers,' at the invitation of governments, have filled holes in the job market and replaced falling birth rates. Their role has been complex: needed for their labor, but resented for their 'otherness,' guest workers and other immigrants have challenged norms of inclusion. In spite of Western Europe's liberal, pluralist, universalistic premises, cultural differences have at times been racialized as part of an exclusionary ideology and derided as impediments to civic incorporation.

Migration patterns, especially in Western Europe, were established in the 1960s by two flows: guest workers, most of whom came from southern and eastern Europe (including Turkey) at the implicit or explicit invitation of states, and immigrants from former colonies. With the significant exception of southern Europe, which changed in the 1980s from a region of emigration to one of immigration, these patterns still largely hold through the end of the century (Miles and Thränhardt 1995).

Rates of naturalization among European countries vary considerably. Brubaker (1992) attributes France and Germany's sharply different rates to fundamentally different conceptions of sovereignty and assimilation. The French, Brubaker says, understand the nation as the creation of the state, while for the Germans the nation is the basis of the state. Soysal (1994), in contrast, argues that postnational international norms have led to 'deterritorialized' notions of rights. She challenges the assumption that citizenship is imperative to membership, since many social rights are enjoyed by noncitizens.

The rights of immigrants, regardless of how they are acquired or conferred, are often shown to be in conflict with actual social treatment. Racist conceptions of otherness tied to immigration, according to Taguieff (1988), are 'differentialist' in their invocation of fundamental cultural distinctions between 'natives' and 'foreigners'; this contrasts with an older 'inegalitarian' notion of biological differences. Indeed, racism has accompanied immigration throughout the region, even while some commentators (e.g., Wieviorka 1992) see important national variations. Others see Western European-wide patterns of racism and exclusion emerging. Balibar and Wallerstein (1988) argue that the most effective defense against this trend is a solid conception of universal civic equality, which should replace more problematic notions of national identity and belonging.

3.3 Social Policy

The extraordinary process of European integration that began modestly in the early postwar years and

accelerated toward the sweeping treaties of the 1990s was primarily economic and secondarily political in motivation, structure, and outcome. Yet, an accompanying social aspect of European unification has begun to receive scholarly attention, namely, the nature of social policy within the authoritative structure of the European Union (EU).

Europeanists ask whether the elevation of economic and political authority to the EU level will find a correlate in social policy. Leibfried and Pierson (1995) argue that a 'social Europe' already exists, and is actively engaged in policies of redistribution, thanks in part to the efforts of trans- and cross-national pressure groups percipient enough to recognize that in certain issue areas power resides in Brussels, not national capitals. Ireland (1995) notes that on immigration issues, EU administrators have taken an activist stance to protect foreign workers and their families, stepping in where national governments have failed to uphold human rights. But Streeck (1995) cautions, first, that EU intervention in a few areas does not reflect an ideologically driven reformist agenda. Second, the EU is beholden to market interests, and will withdraw any move toward social intervention if it threatens market dominance. Streeck's view implies that corporatism at the national level has been defeated by European integration (for a counter argument see Scharpf 2000); indeed, he claims that labor was present in the integration process only insofar as it represented a European-wide market for labor.

3.4 Collective European Identity

Does a transnational European social constituency exist that might eventually force the EU to take more aggressive steps to alleviate market produced inequities and nationalistic exclusionary practices? More abstractly, is the idea of vesting citizenship and collective identity in 'Europe' far fetched? Given the horrific record of twentieth century European nationalisms in protecting human rights and maintaining peace, might superseding the nation-state not be an evolutionary advance?

Scholars occupied less with policy analysis than with theoretical questions of identity and membership have recently addressed these questions. Habermas (1996) argues that the nation–state is mortally ill, and has been in decline since it began to repudiate and exclude everyone regarded as foreign. The process of moving from national to supranational regimes and governance should be viewed as part of the unfinished historical process of modern nation building. The initial steps of nation building in Europe, when the nation replaced disintegrating traditional ties of early modern society, Habermas asserts, is precisely analogous to the present situation in which national sources of solidarity and the moral bases of society are eroding. But what is to replace problematically national sources of identity and belonging? Habermas suggests that pan-European attachment to a shared political culture in the form of an enlarged and fully inclusive public sphere may develop, and that it is the only alternative to increased particularism in multicultural states.

National identities historically evolved through simultaneous inclusion and exclusion, such that every 'us' was defined against an opposing 'them.' Delanty (1995) questions how a collective European identity might develop without asserting an exclusionary 'Fortress Europe,' geographically defined by the shifting borders of the EU. The contemporary idea of Europeanism is not morally neutral, suggests Delanty, but rather constructed through technocratic and media-driven agendas that can neither replace national nor constitute collective identity. Smith (1997) articulates a view widely held among Western Europeanists: firmly entrenched national identities will continue to trump supranational European identity because the affective bonds that support nationalism do not exist for Europe as a whole.

Whether and how such pan-European bonds might develop will occupy scholars in coming decades. The social bases of European nation–states did not grow organically; they were constructed along with the nation-states themselves. The interplay between national form and social content has been based in part on national, regional, and global power factors. These dynamics are fluid, as the process of European integration demonstrates. If economic and political integration can find an inclusive social correlative, exclusionary nationalisms may indeed be mitigated.

See also: Ethnonationalism: Cultural Concerns; Modernization, Political: Development of the Concept

Bibliography

Anderson P 1974 *Lineages of the Absolutist State*. Verso, London

Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London

Balibar E, Wallerstein I 1988 *Race, Nation, Classe: Les identités ambiguës*. Éditions La Découverte, Paris

Becker G 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA

Berger S (ed.) 1981 *Organizing Interests in Western Europe*. Cambridge University Press, Cambridge, UK

Brubaker R 1992 *Citizenship and Nationhood in France and Germany*. Harvard University Press, Cambridge, MA

Delanty G 1995 *Inventing Europe: Idea, Identity, Reality*. St Martin's, New York

Elias N 1939 *Über den Prozess der Zivilisation*, 2 Vols. Haus zum Falker, Basel, Switzerland

Elias N 1994 *The Civilizing Process: The History of Manners and State Formation and Civilization*. Blackwell, Oxford, UK

Esping-Andersen G 1990 *The Three Worlds of Welfare Capitalism*. Polity, Cambridge, UK

Gauthier D 1977 The social contract as ideology. *Philosophy and Public Affairs* **6**(2): 130–64

Gourevitch P 1986 *Politics in Hard Times: Comparative Responses to International Economic Crises*. Cornell University Press, Ithaca, NY
Grillo R D (ed.) 1980 *'Nation' and 'State' in Europe: Anthropological Perspectives*. Academic Press, London
Habermas J 1962 *Strukturwandel der Öffentlichkeit*. Hermann Luchterhand, Darmstadt, Germany
Habermas J 1981 *Theorie des kommunikativen Handelns*. Suhrkamp, Frankfurt, Germany
Habermas J 1988 *The Theory of Communicative Action, Vol. II: Lifeworld and System: A Critique of Functionalist Reason*. Beacon, Boston
Habermas J 1989 *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. MIT Press, Cambridge, MA
Habermas J 1996 *Die Einbezieung des Anderen: Studien zur politischen Theorie*. Suhrkamp, Frankfurt, Germany
Habermas J 1998 *The Inclusion of the Other: Studies in Political Theory*. MIT Press, Cambridge, MA
Inglehart R 1990 *Culture Shift in Advanced Industrial Societies*. Princeton University Press, Princeton, NJ
Ireland P 1995 Migration, free movement, and immigration integration in the EU: A bifurcated policy response. In: Leibfried S, Pierson P (eds.) *European Social Policy: Between Fragmentation and Integration*. The Brookings Institution, Washington, DC
Kaelble W 1987 *Auf dem Weg zu einer Europäischen Gesellschaft: Eine Sozialgeschichte Westeuropas, 1880–1980*. C. H. Beck, Munich, Germany
Korpi W 1983 *The Democratic Class Struggle*. Routledge & Kegan Paul, Boston
Kreisi H, Koopmans R, Duyvendak J W, Giugni M 1995 *New Social Movements in Western Europe: A Comparative Analysis*. University of Minnesota Press, Minneapolis, MN
Leibfried S, Pierson P (eds.) 1995 *European Social Policy: Between Fragmentation and Integration*. The Brookings Institution, Washington, DC
Miles R, Thränhardt D 1995 *Migration and European Integration: The Dynamics of Inclusion and Exclusion*. Pinter, London
Milward A S 1992 *The European Rescue of the Nation-state*. University of California Press, Berkeley, CA
Moore B 1966 *The Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon, Boston
Offe C 1984 *Contradictions of the Welfare State*. MIT Press, Cambridge, MA
Offe C 1987 Challenging the boundaries of institutional politics: social movements since the 1960s. In: Maier C S (ed.) *Changing the Boundaries of the Political: Essays on the evolving Balance Between the State and Society, Public and Private in Europe*. Cambridge University Press, Cambridge, UK
Orloff A S 1993 Gender and the social rights of citizenship: The comparative analysis of gender relations and welfare states. *American Sociological Review* **58**: 303–28
Pateman C 1988 The fraternal social contract. In: Keane J (ed.) *Civil Society and the State: New European Perspectives*. Verso, London
Rhodes M, Mény Y (eds.) 1998 *The Future of European Welfare: A New Social Contract?* St Martin's, New York
Rogers S C 1991 *Shaping Modern Times in Rural France: The Transformation and Reproduction of an Aveyronnais Community*. Princeton University Press, Princeton, NJ
Scharpf F 2000 Globalization: The limitations on state capacity. *Swiss Political Science Review* **6** (2) Available on the World Wide Web at http://www.ib.ethz.ch/spsr/debates/debat_global/art-2-2.html
Schmitter P 1979 Still the century of corporatism? *Review of Politics* **36**: 85–131
Smith A D 1997 National identity and the idea of european unity. In: Gowan P, Anderson P (eds.) *The Question of Europe*. Verso, London
Soysal Y 1994 *Limits of Citizenship: Migrants and Postnational Membership in Europe*. University of Chicago Press, Chicago
Streeck W 1995 From market making to state building? Reflections on the political economy of European social policy. In: Leibfried S, Pierson P (eds.) *European Social Policy: Between Fragmentation and Integration*. The Brookings Institution, Washington, DC
Taguieff P-A 1988 *La force du préjugé: Essai sur racism et ses doubles*. Éditions la Découverte, Paris
Tarrow S 1989 *Struggle, Politics and Reform: Collective Action, Social Movements and Cycles of Protest*. Cornell University Press, Ithaca, NY
Tilly C (ed.) 1975 *The Formation of National States in Western Europe*. Princeton University Press, Princeton, NJ
Tilly C 1978 *From Mobilization to Revolution*. Addison-Wesley, Reading, MA
Touraine M, Wieviorka M, Dubet F 1984 *Le mouvement ouvrier*. Fayard, Paris
Touraine M, Wieviorka M, Dubet F 1987 *The Workers' Movement*. Cambridge University Press, Cambridge, UK
Tucholsky K 1973 *Schnipsel*. Rowohlt, Reinbek bei. Hamburg, Germany
Watkins C S 1991 *From Provinces into Nations: Demographic Integration in Western Europe 1870–1960*. Princeton University Press, Princeton, NJ
Wieviorka M 1992 *La France raciste*. Seuil, Paris

K. W. Moore

Whig

Originally a term of abuse, 'Whig' is a term used to describe a diverse, yet identifiable, party or grouping which was historically important in Great Britain from the late seventeenth century to the mid-nineteenth century. Although now extinct, its ideas and sympathies left important legacies to nineteenth century political life and culture. It is also used to refer to a party, also now extinct, which flourished in the United States in the second quarter of the nineteenth century. Although the Whigs in both countries shared a set of attitudes and dispositions rather than any coherent political doctrines or ideology, the term 'Whig' may still be used in contemporary political discourse.

1. *The British Whigs*

The Whigs were the dominant force in parliamentary politics in the eighteenth century and they remained important in the first part of the nineteenth century. Founded in opposition to Charles II, they embraced the principles of the 'Glorious' English revolution of

1688 and were supportive of the Hanoverian succession of 1714. The Whigs rejected royal absolutism and by extension any claims by the monarch to exercise personal authority. The Whigs' central ideas emphasised the virtues of limited government, the importance of a balanced or mixed constitution based on parliamentary power and representation based on property not individual rights. Whig thought was especially influenced by the writings of John Locke (1632–1704) and the classical republicanism of James Harrington (1611–77). Whig ideology thus stood in contrast to Tory views about the prerogative rights of the Crown on the one hand and radical populism and democratic individualism on the other. Under Sir Robert Walpole (1676–1745) the Whigs' loyalty to these ideas introduced a new and relatively coherent party character into parliamentary politics, although the ties which bound the Whig Party were likely to be personal and familial rather than doctrinal. Given that formal party organization was not a feature of parliamentary politics until the mid-nineteenth century, social ties were especially important in binding the Whigs together. Whigs were largely, but not exclusively, drawn from the aristocratic and landed interest which to some extent explained their conviction that there was a natural link between a strong property-owning class and the protection of liberty.

The Whig Party, although a recognizable entity in British politics, was never wholly united. In addition to personal rivalries and generational differences, there were severe splits over the American revolution of 1776 and the French revolution in the eighteenth century and parliamentary reform in the nineteenth century. The French revolution saw the Whigs split into two antagonistic groups in 1794: those who opposed the violence inherent in the revolution and who supported war with France and those who looked more favorably on the revolution and could see no cause for war with France. The criticism of the French revolution was led by Edmund Burke (1729–97) who in his *Reflections on the Revolution in France* (1790) and his *Appeal from the New to the Old Whigs* (1791) provided a defence of gradual constitutional change which influenced later conservative thought. The pro-French party was associated with Charles James Fox (1749–1806).

The new political agenda created by industrialization and democratization in the nineteenth century further weakened the Whigs. As the expansion of the franchise created new pressures to organize outside parliament as well as within it, the Whigs found their ideas and their politics eclipsed by radicals and liberal reformers who stressed the rights of individuals and looked to the state to effect social, economic, and political reform. With the formation of the Liberal Party in the 1850s, the distinctive character of the Whigs was eroded. In the party realignments over the Irish issue in 1885–6 the Whigs finally lost their identity and the word largely went out of political currency.

2. *The American Whigs*

The theorists of the American revolution shared much of the intellectual heritage of the English Whigs in their emphasis on constitutionalism, balanced government and liberty, and their admiration for classical republicanism. American whiggery also parallelled the aristocratic and landed character of English whiggism in its appeal especially to the plantation owners of the south. And whiggism had a strongly Protestant appeal in the United States as in England.

A separate and distinctive Whig Party did not however arise in the United States until 1834 when a loose coalition of opponents of President Andrew Jackson (1767–1845) and his policies emerged around the leadership of Henry Clay. The name 'Whig' was deliberately taken to signal the new party's opposition to Jackson's assertive view of presidential power (which evoked memories of monarchical rule) and support for national development in contrast to Jackson's agrarian populism.

The coalition which formed the Whig Party included the National Republican Party, supporters of Henry Clay's American System, the anti-Masonic party, and anti-Jackson Democrats. Although the Whigs could not agree on a presidential candidate in 1836, in 1840 they united around William Henry Harrison who defeated Martin van Buren. When Harrison died after only a month in office, John Tyler, a strong supporter of states rights, succeeded as President. Tyler became increasingly distanced from the Whig Party's more nationalistic senior figures Clay and Daniel Webster. The Whig Party won the presidency again in 1848 with Zachary Taylor. The Whigs were the majority party in the Senate between 1841 and 1844 and controlled the House of Representatives between 1841–2 and 1847–8.

Millard Fillmore, Taylor's vice-president who succeeded to the presidency on Taylor's death in 1850, was the last Whig president. Regional and policy divisions, especially over free soil and slavery, weakened the Whigs in the 1850s and the Party ran its last presidential candidate, Winfield Scott, who lost against Franklin Pierce in 1852. Factions of the old Whig Party were subsequently absorbed into the Democratic Party and the Republican Party which in 1860 nominated a former Whig, Abraham Lincoln, as its presidential candidate.

3. *Contemporary Usage*

Although rarely used in contemporary political debate the term 'whig' is occasionally used to refer to liberals who are suspicious of democratic populism and abstract political theories, believing instead that political liberty is best promoted by the preservation of property and reforms rooted in the specific institutions of a society. It is also still used to refer to a school of historical writing which sees historical development as

a benign, purposive, and progressive process. Although this so-called 'whig view of history' is most commonly associated with nineteenth century writers such as Thomas Babington Macaulay (1800–1859), it may be used more broadly to refer to any evolutionary historical theory.

See also: Conflict/Consensus; Issue Constraint in Political Science; Locke, John (1632–1704); Parliamentary Government

Bibliography

Elofson W M 1996 *The Rockingham Connection and the Second Founding of the Whig Party 1768–1773*. McGill Queens University Press, Montréal

Guttridge G H 1979 *English Whiggism and the American Revolution*. AMS Press, New York

Holt M F 1999 *The Rise and Fall of the American Whig Party: Jacksonian Politics and the Onset of the Civil War*. Oxford University Press, New York

Howe D W 1979 *The Political Culture of the American Whigs*. University of Chicago Press, Chicago

Jones J R 1970 *The First Whigs: The Politics of the Exclusions Crisis 1678–1683*. Oxford University Press, London

Mitchell L G 1971 *Charles James Fox and the Disintegration of the Whig Party*. Oxford University Press, Oxford, UK

Poage G R 1936 *Henry Clay and the Whig Party*. University of North Carolina Press, Chapel Hill, NC

Southgate D 1993 *The Passing of the Whigs*. Gregg Revivals, Aldershot, UK

G. Peele

White-collar Workers, History of

1. Preliminary Remarks

Although now in many developed industrial societies white-collar workers represent, at least in numerical terms, the single most important group within the general workforce, their delimitation and their specific social profile in relation to blue-collar workers and public servants remains ambiguous and controversial (Kocka 1980, 1981, Lockwood 1989). With hardly any other social stratum was and is the complex interrelation between the actual appearance of a group, its self-definition and its treatment as an object of investigation so clear as with the white-collar employees.

The sociological interest in these groups has a long tradition which dates back into the late nineteenth century. It arose mainly for two reasons: (a) the exceptional rapid growth of the various groups, labeled as 'white-collar workers,' *Angestellte*, *employŭs*, *impiegati*, *empregados*, etc. and (2) the fact that these groups, with regard to some objective features of their working conditions, aspects of their legal status, as well as their consciousness and their behavior, differed markedly from other social groups already established. Interest was nowhere so strong as in the German-speaking countries, where, as a field of the emerging social sciences, a special *Angestelltensoziologie* developed. This pecularity was mainly due to the existence of a radical labor movement and of a strong bureaucratic tradition. The institution of the privileged 'civil servant' served as a role model for many white-collar workers and attributed much to the relative depth of the distinction in relation to the blue-collar workers in the respective countries. However, comparative historical research shows that the terms 'white-collar workers,' *Angestellte*, *employŭs*, *impiegati*, and *empregados* are in no way interchangeable. The delimitation and, particularly, its social 'meaning' vary strongly historically and from country to country. Though much of the special social profile of the various groups labeled as 'white-collar' has faded away, there still remain some important differences, especially in relation to blue-collar workers.

2. Definition by Functional Criteria

In many countries in everyday language as well as in scientific usage the term white-collar worker refers to a group of dependent workers which, compared to manual or blue-collar workers, is occupied not directly in the process of manufacturing. Rather it is situated in functional as well as in spatial terms *before* the process of manufacture, *beside* it and *behind* it. One encounters white-collar workers in all economic sectors, even agriculture, at least, as it is organized commercially and as a large-scale business. However, the actual domain of the employees is the service

Table 1
White-collar workers in the US, GB, and Germany 1882–1950 (in percent of the total labor force)

	USA	GB	Germany
1882			6.1
1900	12		8.5 [1895]
1910	15	18.7	11.7 [1907]
1921	19 [1920]	21.2	16.0 [1925]
1931	22 [1930]		17.1 [1933]
1939	24 [1940]		18.9
1950	30.9	30.9 [1951]	19.9
1970			35.7
1987			50.1

Source: Behringer 1985 p. 515, Kocka 1980 p. 42f, Kunig p. 172
Note: These numbers include white-collar employees in private businesses as well as in the public sector

Table 2
Percentage of women in white-collar occupations in Germany, 1882–1982

Year	All sectors	Office and public service	Sales personnel	Engineers and technical personnel
1882	11.7		19.0	
1895	14.7	4.5	30.0	
1907	19.7	16.8	42.0	
1925	29.7	34.5		4.4
1933	34.6	36.0	77.9	
1939	39.8	47.0	79.1	6.0
1950 (FRG)	38.6	40.0	78.3	8.6
1970	49.2	50.3	81.5	12.0
1982	52.4	64.6	80.5	11.1

Source: Kunig, p. 172.

sector; that is, trade, banking, insurances, where they always made up the predominant part of the workforce. In addition, large groups of employees are to be found in industry and the public sector. Since the end of the nineteenth and above all in the course of the twentieth century employees have increased their share in all sectors considerably in all developed societies compared to that of manual workers and public servants (see Table 1).

With respect to their productive functions white-collar workers differ from manual workers in that they—varying with the economic sector in which they are employed—perform directing, overseeing, constructive, analyzing, administrative, commercial, and information-processing tasks. These criteria may lead, in many cases, to a clear distinction between manual workers and white-collar employees in private enterprises. They are hardly suitable for a clear distinction between officials or civil servants (*Beamte*) and employees (*Angestellte*) in the public sector. The functions of these two categories normally overlap. But the upper ranks of the hierarchy are staffed by employees with the status of civil servant (Humphreys 1958). *Beamte* and *Angestellte* differ by legal status rather than by types of functions.

3. *Origin and Growth*

The origins of some white-collar professions may be traced back to the premodern period. Among them figure, for instance, the commercial employees which at that time still had a good chance to reach the status of a self-employed salesman later in life. Nevertheless as a mass stratum one has called white-collar employees rightly the 'latest child of industrial capitalism.' Two main developments contributed to the rapid growth of the group: (a) the shifting balance between the 'first' (agriculture) and the 'second sector' (industry) in favor of the 'third,' the service sector (Fourastiй 1963)—a development that took place first in Great Britain while other industrial nations followed suit in the course of the nineteenth and twentieth century; (b) the rise of 'Organized Capitalism' (Hilferding 1981) since the end of the nineteenth century, a phase characterized by the growing importance of organization, planning, information-gathering, bureaucratic control, and the creation of structures within and outside industry to serve these ends. The relative growth of the share of the employees within the public sector, however, is determined not only by changes in the function of the modern state but also to a large extent by the political will either to expand or to restrict the spread of particular privileges attached to the status of civil servant (*Beamte*). Such privileges are, for example, the protection against dismissal and pension schemes, etc.

As recent social data show, even in the most advanced contemporary industrial societies the expansion of white-collar jobs has not yet come to an end. But there are clear signs that the rate of expansion slowed down significantly in the second half of the twentieth century. As with all social growth, the creation of white-collar jobs as a strategy to tackle problems arising from the development of economy and society as a whole is subject to the law of diminishing returns. While in the early stages of industrial capitalism white-collar employees were instrumental in the process of rationalizing production—a fact that caused widespread resentment among manual workers—in further stages they were themselves subjected to strategies of cost-cutting and rationalizing. This development began very soon after the turn of the twentieth century with the introduction of specific office machines as the typewriter, the Hollerith-machine—an early form of large-scale data-processing—and other changes in the structure and organization of office work. This process definitely has gained a new quality in recent years. In particular, the spreading of the computer, the communication revolution, and the integration of office work into controllable networks has opened new chances for quantitative control of employee activities once considered to be typical only for manual work. Whether

this tendency is outweighed by other trends towards 'lean management' and more autonomy on the job remains to be seen.

4. Status and Privilege

The Swedish–German sociologist Croner (Croner 1951) developed his 'theory of delegation,' according to which most early white-collar jobs in industry emanated through a process of division of labor and of delegation out of functions originally performed by the employer. This theory has been questioned by empirical research. Especially in modern industries many tasks performed by white-collar workers actually represent new functions without any historical precedent. It seems highly questionable that *all* early functions classified as white-collar carried with them high prestige. Especially, in the area of small-scale shop-keeping there existed extremely exploitive working conditions for clerks with low pay, no job security, long working-hours, etc. (Anderson 1976, Crossick 1977). Nevertheless, it holds true that many white-collar occupations enjoyed at least some form of privilege in comparison with manual workers, and continue to do so. Some of the differences resulted from the specific nature of their work that could not be subjected easily to quantitative methods of control. One of the traits that distinguished most—not all—white-collar jobs from blue-collar functions was a fixed income, the 'salary,' that, at least in principle, did not vary with working hours or the amount of work carried out. Because of the fact that knowledge of business transactions within the specific firm constituted the main qualification of many white-collar employees, the accumulation of such individualized knowledge gave members of the group a higher job security than manual workers, especially in the early stages of industrialization. Even in later days, the office never became the place for a general 'hire and fire' climate. As many positions of white-collar workers were integrated into large, delicately differentiated bureaucratic hierarchies, aspirations for upward mobility never vanished completely. Besides, in many countries there were, traditionally, legal privileges attached to certain white-collar occupations. These entitlements included protection against dismissal, paid vacation, continuation of wage payment in the case of illness, company pensions schemes, etc.

The idea that the class of white-collar workers as a whole would face 'proletarianization,' as put forward especially by writers in the Marxist tradition, has given way to a much more nuanced picture. At earlier stages of the development legal privileges attached to certain types of white-collar workers were generalized among larger groups. Moreover, the French sociologist Michel Crozier has argued that the consequences of rapid rationalization were often absorbed by the constant influx of female workers. Though the equation 'proletarianization-feminization' does not come out, there is some degree of truth in it. New styles of management—the movement towards 'lean management' since the 1980s, for instance—have amplified autonomy and reopened scope, at least, for certain groups among the white-collar population. On the other hand there can be no doubt that a process of narrowing and even blurring of differences between blue-collar and white-collar workers took place in the course of the twentieth century. Important groups within the white-collar population, especially clerical workers and sales personnel, have been overtaken by skilled manual workers in terms of income, job-security, even prestige. Many of the rights once attached exclusively to certain white-collar positions have, step-by-step, been granted to the workforce as a whole, making white-collar workers, at least in some countries such as Germany, important, if involuntary, harbingers of the modern social state. Though, on average, white-collar workers still appear to be privileged in relation to the manual workforce, the 'average' white-collar worker largely has become a statistical fiction due to the immense variety of positions and professions that nowadays characterize these categories.

5. Organizational Behavior

Depending on the structure of the trade union movement within the respective countries, white-collar workers tend to organize exclusively or are integrated into general unions (Behringer 1985, Bain and Pollins 1965, Bain 1972; Jenkins and Sherman 1979). In most cases union membership among white- and blue-collar workers overlaps at least to a certain degree. The main area in which both groups differ is trade union membership in general. White-collar workers are considerably less well organized than manual workers with the one exception of the public sector. Hence, their growing weight among the workforce represents one of the most important factors in the crisis of the trade union movement in developed countries in general.

6. National Differences and Pecularities

While in most industrialized countries the division between white- and blue-collar workers was never clearly marked or valid only in regard of relatively small groups, in some European countries such as Austria, Czechoslovakia, Switzerland, and Germany until the middle of the twentieth century, blue- and white-collar workers constituted distinguished entities, separated in many aspects, legal as well as organizational and ideological. Social legislation, es-

pecially pensions schemes, had been enacted before and after the First World War. They had an important impact, strengthening existing differences between white und blue-collar workers. The dividing line between the two groups structured large parts of the fabric of these societies—legislation, trade union movement, self-help organizations, political affiliations, etc. By the 1950s, the distinction between white and blue collar had begun to erode gradually, a development that was supported by the expansion of former white-collar privileges to the general workforce. Another important aspect in which German-speaking countries before the Second World War differed from other societies was the sheer size of the legally defined white-collar population which included sales personnel as well as general managers.

7. *Social Theory*

White-collar workers were introduced into the realm of modern social science at the end of the nineteenth century. The starting point was twofold: (a) the notion of a growing number of occupations, employments, and functions that shared many traits of the working-class while at the same time differing in important aspects from manual workers as well as from the higher social strata; (b) the exceptional rapid growth of these categories in the decades around the turn of the century which even outdid that of the manual workers.

The most intensive debates on white-collar workers took place in German-speaking countries from about 1890 up to the 1930s against the background of the formation of strong labor movements (largely blue-collar) and of a political system, dominated by conservative elites eager to defend the status quo and in search of numerically important allies among the general population. In the first stage of the debate, which took place before the First World War, status quo-oriented scholars identified white-collar workers as part of a new socially stable and conservative 'middle estate' and, at the same time, encouraged the state to engage in a social-protective policy (e.g., white-collar specific insurances and labor laws) while Marxist theorists predicted a general proletarianization of these groups. The 1920s saw a convergence of opinions, reinforced by the sceptical notion that the 'new middle class' in German-speaking countries seemed to figure prominently among the mass base of National Socialism. The label of a potential for right-wing extremism has stuck to the group ever since, not least through the intellectual influence of German social scientists, many of whom were forced to emigrate in the early 1930s (Fromm 1984, Speier 1986). The notion showed up again in many theories of Fascism developed in the 1950s and 1960s, generalizing what was seen to be the experience of the interwar years (Lipset 1960). In contrast, social historians who took up the debate in the 1970s tried to demonstrate the pecularities of the German and Austrian experience (e.g., Kocka 1980) while latest research has challenged the notion of a predominantly right-leaning, 'new middle class' in the 1920s altogether (Falter 1991).

The debate over the role of white-collar workers in the mass base of National socialism represented the climax and at the same time a turning point. Interest in interpretations of the white-collar group as a whole has subsided (but see Schulz 2000) in Germany as well as in other countries where it never figured prominently.

Not surprisingly, with regard to the heterogeneity of the group, social science attributed contradictory attitudes and social functions to these groups, often departing from specific traits of subgroups. One important theme over the decades has been the impact the expansion of white-collar occupations has had on the emancipation of women, whether it offered new opportunities for social mobility or whether it only helped to reproduce traditional inequalities (Suhr 1930). There has been much research on the impact of rapid technological change on office work through the decades, some with more pessimistic, others with more optimistic overtones. Studies of the service sector, for instance, stress its relative resistence to change as a result of the need to maintain a workforce even when there is no constant demand, a fact which has been labeled 'guard labor.' Other scholars have pointed to the partial regain of autonomy on the work-place as part of the automation movement in the 1970s which was said to have led to the formation of a 'new worker' (Mallet 1963). Actually, the debate referred to members of the technical staff supposedly holding key positions for potential labor disputes. Another important context was the rise of modern consumer society since the First World War in Europe and the United States. The concentration of large numbers of white-collar employees in larger cities led to the identification of them with the urban population as a whole, regarding them as pioneers of modern consumption and new forms of leisure (Coyner 1977). This notion dates back to the influential study first published in 1924, of the German social writer, Siegfried Kracauer (Kracauer 1998) and translated into many languages ever since. It has been taken up in various studies of the historic development of modern consumption and leisure, leaving out, with regard to the white-collar population, those members of the group living in smaller and middle-size cities.

Though, in the second half of the twentieth century, interest for white-collar employees in the social sciences has flared up from time to time, there seems to be no future for any revival of a wholesale white-collar worker sociology as it developed in some European countries during the interwar years. One of the most important prerequisites, the existence of a self-conscious socialist working-class, no longer exists. With

its disappearance the interest of social science, policy-makers, and the different groups that make up this strata to develop a broad definition based upon the distinction between 'us,' the 'white-collar,' and 'them,' the 'blue-collar workers,' seems to have vanished. What remains is a remarkable gap between the numerical importance of the group in question and its relative neglect by social scientists in the last decades.

See also: Bourgeoisie/Middle Classes, History of; Income Distribution; Professionalization of Social and Behavioral Scientists: United States; Professionalization/Professions in History; Professions, Sociology of; Social Inequality in History (Stratification and Classes); Social Mobility, History of; Work and Labor: History of the Concept; Work, History of; Working Classes, History of

Bibliography

Anderson G I 1976 *Victorian Clerks*. University Press, Manchester

Bain G S 1972 *The Growth of White-collar Unionism*, 2nd edn. Clarendon Press, London

Bain G S, Pollins I I 1965 The history of white-collar unions and industrial relations: A bibliogaphy. *Bulletin of the Society for the Study of Labour History* **11**: 20–25

Behringer P 1985 *Soziologie und Sozialgeschichte der Privatangestellten in Großbritannien*. Frankfurt, Germany

Bonham J 1954 *The Middle Class Vote*. Faber of Faber, London

Coyner S J 1977 Class consciousness and consumption: The new middle class during the Weimar Republik. *Journal of Social History* **10**: 310–31

Croner F 1951 *Tjδnstemannakaren i det moderna Samhδllet*. Stockholm (German, revised edn. Frankfurt 1954)

Crossick G (ed.) 1977 *The Lower Middle Class in Britain 1870–1918*. St. Martin's Press, New York

Crozier M 1971 *The World of the Office Worker* (trans. Landau D). University of Chicago Press, Chicago

Falter J W 1991 *Hitlers Wδhler*. Beck, Munich

Fourastiй J 1963 *Le grand espoir du XXe siècle*. Gallimard, Paris

Fromm E 1984 *The Working Class in Weimar Germany—a Psychological and Sociological Study* (trans. Weinberger B, ed. and with an introduction by gang Bonss W). Bierg, Leamington Spa, UK

Gerstl J E, Ilutton S P 1966 *The Engineers: The Anatomy of a Profession. A Study of Mechanical Engineers in Britain*. Tavistock Publications, London

Giddens A 1981 *The Class Structure of the Advanced societies*, 2nd edn. Hutchison, London

Hilferding R 1981 *Finance Capital: A Study of the Latest Phase of Capitalist Development* (ed. with an introduction by Bottomore T from translations by Watnick M, Gordon S). Routledge, London

Humphreys B V 1958 *Clerical Unions in the Civil Service*. Blackwell & Mott, London

Jenkins C, Sherman B 1979 *White-collar Unionism. The Rebellious Salariat*. Routledge, London

Kater M H 1983 *The Nazi Party—a Social Profile of Members and Leaders, 1919–1945*. Blackwell, Oxford, UK

Kocka J 1980 *White-collar Workers in America: 1890–1940: a Social-political History in International Perspective*. Sage, London

Kocka J (ed.) 1981 Angestellte im europδischen Vergleich. Die Herausbildung angestellter Mittelschichten seit dem spδten 19. Jahrhundert

Koenig M 1991 *Die Angestellten unterwegs. Vom Berufsstand zur modernen Gewerkschaft 1890–1990*. Bund-Verlag, Kuln

Kracauer S 1998 The Salaried Masses–Duty and Distraction in Weimar Germany (trans. Hoare Q and with an introduction by Mulder-Bach I). Verso, New York

Lederer F 1979 *Kapitalismus, Klassenstruktur und Probleme der Demokratie in Deutschland 1910–1940. Ausgewδhlte Aufsδtze von Emil Lederer herausgegeben von Jurgen Kocka*. Vandenhoeck und Ruprecht, Goettingen, Germany

Lipset S M 1960 *Political Man: The Social Bases of Politics*, 1st edn. Doubleday, Garden City, NY

Lockwood D 1989 *The Blackcoated Worker—a Study in Class Consciousness*, 2nd edn. University Press, Oxford, UK

Mallet S 1963 *La nouvelle classe ouvriure*. Paris (Spanish: *La nueva condicion obrera*. Madrid 1969).

Mills C W 1951 *White-collar—the American Middle Classes*. Oxford University Press, New York, London

Reader W J 1966 *Professional Men: The Rise of the professional Classes in 19th Century England*. Weidenfield & Nicolson, London

Schulz G 2000 *Die Angestellten seit dem 19. Jahrhundert*. Oldenbourg, Munchen, Germany

Sobel R 1989 *The White-collar Working Class. From Structure to Politics*. Proeger, New York

Speier Il 1986 *German White-collar Workers and the Rise of Hitler*. Yale University Press, New Haven and London

Suhr S 1930 *Die weiblichen Angestellten*. Zentralverband der Angestellten, Berlin

Touraine A 1969 *La société post-industrielle*. Paris

M. Prinz

Williams Syndrome

In the final decades of the twentieth century, cognitive neuroscience paid particular attention to a group of disorders known as microdeletion syndromes. These are neurodevelopmental disorders in which the missing genes on a chromosome can be identified and the resulting gene-behavior relations explored. They differ from syndromes like Down syndrome in which an entire extra chromosome is present, as in Trisomy 21. One microdeletion disorder that caused particular excitement among cognitive neuroscientists is Williams syndrome (henceforth WS). This is because WS results in an unusually uneven cognitive profile. Language and face processing are seemingly spared, whereas other higher cognitive functions (spatial cognition, number, planning, and problem solving) are seriously impaired. Initial characterizations of the syndrome at the cognitive level seemed to hold the promise of relatively straightforward gene-cognition

mappings. WS was, and continues to be, hailed as the prime example of some intact, innately specified cognitive modules in the face of general intellectual impairment (e.g., Pinker 1994, 1999). However, more recent evidence fails to support this view, favoring instead a more dynamic, neuroconstructivist approach to genetic disorders. In this entry, the cognitive processes underlying the purportedly spared domains will be examined, together with a focus on the infant starting state and the developmental trajectory leading to the phenotypic outcome.

1. *Williams Syndrome: Genetic Profile*

Williams syndrome is a rare developmental disorder and occurs in approximately 1 in 20,000 live births. It is caused by a submicroscopic deletion on chromosome 7q11.23. The deleted region contains some 20 genes, about 17 of which have been identified (Ewart et al. 1993, Frangiskakis et al. 1996, Meng et al. 1998, Tassabehji et al. 1996). Only a few of these genes are expressed in the brain and are therefore of special interest to the cognitive neuroscientist. Others affect the physical development of patients, particularly with respect to impairments to the cardio-vascular system.

Initial excitement came from the discovery that in the majority of cases of WS one copy of the elastin gene (ELN) (Ewart et al. 1993) and one copy of the Limkinase1 gene (LIMK1) (Frangiskakis et al. 1996, Tassabehji et al. 1996) were consistently deleted. ELN is important for elasticity of the heart, skin, blood vessels, and lungs. Its deletion was therefore rapidly linked to the facial dysmorphology and supra-valvular aortic stenosis (SVAS) found consistently in individuals with WS (Ewart et al. 1993). LIMK1 is expressed in the developing brain and its deletion was claimed to explain the typical pattern of spatial impairments found in the cognitive profile of such individuals (Frangiskakis et al. 1996, Mervis et al. 1999). Despite the claims about the role in WS of the ELN and LIMK1 genes in facial dysmorphology and spatial cognition respectively, direct one-to-one genotype/phenotype mappings are highly unlikely in cognitive neuroscience, as a subsequent study by Tassebehji and collaborators showed (Tassabehji et al. 1996). Three patients with SVAS were examined who had partial deletions on chromosome 7 in the same region as the clinical groups with WS. The study showed that despite the ELN and LIMK1 deletions, none of these patients had the facial dysmorphology typical of WS, nor did they display the uneven WS cognitive profile of impaired visuo-spatial cognition and enhanced linguistic capacities. They all had an even cognitive profile within the normal range. The results indicate that the ELN deletion does not alone explain the facial dysmorphology found in WS. They also suggest that the LIMK1 deletion is either irrelevant to the development of spatial cognition or that its expression interacts with a number of other genes to contribute to the spatial impairment. But it is clear that one-to-one mapping between specific genes and higher cognitive outcomes does not hold.

2. *Williams Syndrome: Phenotypic Profile*

The pioneering work of Bellugi and her collaborators initially pointed to some clear-cut dissociations in the cognitive architecture of WS. Language and face processing appeared to be surprisingly preserved in the face of both general retardation and particularly serious problems with visuo-spatial cognition, number, planning, and problem solving (Bellugi et al. 1994). Researchers in the field of WS have been fairly cautious about their claims, referring to relative strengths and weaknesses rather than absolute ones (Bellugi et al. 1999, Karmiloff-Smith 1998, Klein & Mervis 1999, Mervis 1999, Tager-Flusberg et al. 1998, Vicari et al. 1996, Volterra et al. 1996). By contrast, secondary sources cited in writings by linguists, developmental psychologists, neuropsychologists of adult brain damage, and philosophers have often used Williams syndrome to bolster claims about innate and independently functioning modules, some of which are purportedly intact and others impaired (e.g., Pinker 1994, 1999). This stems from the view that the pattern of behavioral performance found in the phenotypic outcome is a direct window on the purported innately specified, modular structure of the cognitive architecture of the brain (Baron-Cohen 1998, Leslie 1992, Temple 1997). Such reasoning treats the genetically impaired brain as if it were a normal brain with parts intact and parts impaired, ignoring the dynamic role of genetic mutation in interaction with environmental input in fostering overall brain growth. This has been particularly the case with studies of autism and Williams syndrome, in which cognitive impairments in older children and adults have been used to make claims about gene expression, in the absence of studies of the starting state in infants.

Recent studies of infants, children, and adults with WS strongly suggest that the starting state cannot be simply inferred from the phenotypic outcome (Paterson et al. 1999). Infants with WS were compared with chronological (CA) and mental-age (MA) matched infants with Down syndrome (DS), as well as MA and CA-matched typically developing controls. Despite the fact that WS adults perform significantly better than DS adults on vocabulary tasks, WS infants are as seriously impaired on vocabulary tasks as are DS infants. Moreover, despite WS adults having significantly worse problems with judging numerosities in adulthood than DS adults and control groups, WS infants perform normally, like the CA-controls, and significantly better than DS infants. In other words, the patterns obtaining in infancy turn out to be the opposite of the patterns in adulthood, pointing to the

importance of the dynamics of developmental trajectories over time, rather than a static view of the infant starting state and the phenotypic outcome.

Even in cases where the phenotypic outcome seems to display spared performance, in-depth analyses suggest that people with WS process inputs via different cognitive processes (Karmiloff-Smith 1998). In the domain of vocabulary, individuals with WS do not obey the same lexical constraints as normal controls when learning new words. In the domain of syntax, WS adults tend to display patterns typical of much younger children rather than intact performance (Klein and Mervis 1999). For example, it has been claimed that individuals with WS have intact regular past tense formation of verbs, alongside impaired associative lexical processes (Clahsen and Almazan 1998). Recent research challenges this claim in that once verbal mental age is taken into account, WS patients display no selective deficit for irregular verbs (Thomas et al. 2001). Specifically, the WS data can be placed on the normal developmental pathway found in much younger subjects. These various results are consistent with the hypothesis that the WS language system is seriously delayed because it has developed under different constraints.

WS language is not simply delayed, however. Several studies now suggest that there is an imbalance in WS at different times in development between phonology and semantics (Karmiloff-Smith 1998, Thomas et al. 2001). For example, when WS participants monitor sentences for a target word, they do not show sensitivity to certain sentential violations, suggesting that in WS semantic information may become available too slowly to be integrated with the online processing of syntax. A study of reading in WS came to similar conclusions about the weak role of semantics in learning to read new words. The WS group displayed equal levels of reading for both concrete and abstract words. By contrast, the controls found concrete, imageable words much easier to read. In general, imageability effects have been shown to be weaker in people with WS (Karmiloff-Smith 1998). A more recent study by Vicari and his colleagues demonstrated that, compared to normal controls, word learning is superior in WS if the auditory presentation of a word is accompanied by the simultaneous presentation of a photograph depicting the object (Vicari et al. 2000). This seems to be because, unlike normal controls, people with WS are defective in spontaneously forming a visual (semantic) image of auditorily presented words. Finally, when learning new spoken words and despite a vocabulary test age of 8, people with WS behave like 4–5 year olds, and do not show the pattern typical from 6 years onwards in the normal population. Like very young children, WS patients are less influenced by the semantics of the real words that the nonce terms resemble. Rather, they rely more on phonology. Taken together, a variety of studies suggest that, unlike typical development, phonological representations are at times stronger than semantic representations in their influence on the way in which WS language develops.

Although it is becoming increasingly clear that vocabulary development does not follow a normal developmental trajectory in WS and that semantics places a weaker constraint on WS language development than in typical controls, it remains possible that WS syntax is intact, as many have claimed (e.g., Pinker 1994, 1999). There are, however, a number of lines of evidence to doubt this. First, vocabulary levels are usually better than syntactic levels in WS on various standardized tasks, although both are significantly below chronological age (Karmiloff-Smith 1998). Second, even in very simple imitation tasks, participants with WS show impairment with complex syntactic constructions such as embedded relative clauses. These and various other findings in different laboratories are hardly consistent with the view that WS syntax is intact. Even in an area of relatively simple syntax—grammatical concord over sentence elements—which young, normal French-speaking children acquire easily and early, people with WS show impairment even in adulthood (Karmiloff-Smith 1998). Although the WS children learn the local gender marker (correct article) for a nonce term easily (in fact, more easily than control children), their capacity for gender agreement across sentence elements such as agreement on adjectives or pronouns is seriously impaired. Studies of Italian-speaking children have also revealed that grammatical gender is a particular problem, with WS children displaying errors never encountered in normal development (Volterra et al. 1996). Several studies (e.g., Klein and Mervis 1999) now suggest that the problems that people with WS have with semantics and syntax are often camouflaged by their good verbal memory. This again demonstrates that overt behavior is not necessarily an index of underlying cognitive competence.

A similar conclusion holds for the domain of face processing. Despite reports that WS face processing is intact (Bellugi et al. 1994, Rossen et al. 1996), in-depth studies of face processing suggest that individuals with WS use different strategies from normal controls (Karmiloff-Smith 1998). Several studies (Deruelle et al. 1999, Karmiloff-Smith 1998, Udwin & Yule 1991) have replicated earlier work, revealing normal or near normal behavioral scores on standardized tasks like the Benton Facial Recognition Test (Benton et al. 1983) and the Rivermead Behavioural Memory Test (Wilson et al. 1985). However, these studies have seriously challenged the notion that the behavioral success displayed in WS face processing capacities is subserved by normal cognitive processes. Where normal controls use predominantly configural strategies for processing upright faces and featural strategies for processing inverted faces, the WS patients tend to process the featural details of both upright and inverted faces. These different strategies have been

further explored in brain imaging studies (Mills et al. 2000). When older children and adults with WS have to match faces in an event related potential study, they display temporal processes found at no age in normal controls. They also tend to process faces bilaterally or predominantly with the left hemisphere, whereas normal controls show a right hemisphere bias for faces. All of these findings point to atypical face processing in WS, despite the normal behavioral scores.

3. Concluding Comments

Developmental cognitive neuroscience must take development seriously. The WS brain is 20% smaller than normal brains and qualitatively different in terms of brain anatomy (Bellugi et al. 1999), brain chemistry (Rae et al. 1998), and computational processing (Mills et al. 2000). This holds throughout embriogenesis and postnatal brain development and means that interaction with environmental stimuli will be subtly different. Given a very different brain, it is unsurprising that even when overt behavior seems normal, as in some aspects of WS language and face processing, these skills actually turn out to be underpinned by cognitive processes that are different from the normal case (Karmiloff-Smith 1998).

Williams syndrome is an excellent model for the neurocognitive study of genetic disorders, because of its strikingly unusual cognitive profile. The syndrome is especially important because of the way in which in-depth research highlights the need to go beyond both observable behavior and static descriptions of snapshots of developmental outcomes, to the charting of neurocognitive trajectories from infancy onwards.

See also: Cognitive Development in Infancy: Neural Mechanisms; Infant Development: Physical and Social Cognition; Prefrontal Cortex Development and Development of Cognitive Function

Bibliography

Baron-Cohen S 1998 Modularity in developmental cognitive neuropsychology: Evidence from autism and gilles de la Tourette syndrome. In: Burack J A, Hodapp R M, Zigler E (eds.) *Handbook of Mental Retardation and Development*. Cambridge University Press, Cambridge, UK, p. 335

Bellugi U, Wang P P, Jernigan T 1994 Williams Syndrome: An unusual neuropsychological profile. In: Broman S H, Grafman J (eds.) *Atypical Cognitive Deficit in Developmental Disorders: Implications for Brain Function*. Lawrence Erlbaum Associates Inc., Hillsdale, NJ, pp. 23–56

Bellugi U, Lichtenberger L, Mills D, Galburda A, Korenberg J R 1999 Bridging cognition the brain and molecular genetics: evidence from Williams syndrome. *Trends in the Neurosciences* **22**(5): 197–207

Benton A L, Hamsher K, de S Varney N R, Spreen O 1983 *Contributions to Neuropsychological Assessment*. Oxford University Press, New York

Clahsen H, Almazan M 1998 Syntax and morphology in Williams syndrome. *Cognition* **68**(3): 167–98

Deruelle C, Mancini J, Livet M O, Casse-Perrot C, de Schonen S 1999 Configural and local processing of faces in children with Williams syndrome. *Brain and Cognition* **41**: 276–98

Ewart A K, Morris C A, Atkinson D, Weishan J, Sternes K, Spallone P, Stock D A, Leppert M, Keating M T 1993 Hemizygosity at the elastin locus in a developmental disorder Williams syndrome. *Nature Genetics* **5**: 11–6

Frangiskakis J M, Ewart A, Morris C A, Mervis C B, Bertrand J, Robinson B F, Klein B P, Ensing G, Everett L A, Green E D, Proschel C, Gutowski N J, Noble M, Atkinson D L, Odelberg S J, Keating M T 1996 LIM-kinase1 hemizygosity implicated in impaired visuospatial constructive cognition. *Cell* **86**: 59–69

Karmiloff-Smith A 1998 Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences* **2**: 389–98

Klein B P, Mervis C B 1999 Contrasting patterns of cognitive abilities of 9 and 10-years-olds with Williams syndrome or Down syndrome. *Developmental Neuropsychology* **16**(2): 177–96

Leslie A M 1992 Pretence autism and the theory-of-mind-module. *Current Directions in Psychological Science* **1**: 18–21

Meng X, Lu X, Zhizhong L, Green E D, Massa H, Trask B J, Morris C A, Keating M T 1998 Complete physical map of the common deletion region in Williams syndrome and identification and characterization of three novel genes. *Human Genetics* **103**: 590–9

Mervis C B 1999 The Williams syndrome cognitive profile: Strengths weakness and interrelations among auditory short-term memory language and visuospatial constructive cognition. In: Winograd E, Fivush R, Hirst W (eds.) *Ecological Approaches to Cognition: Essays in Honor of Ulric Neisser*. Lawrence Erlbaum Associates Inc. Mahwah, NJ, pp. 193–227

Mervis C B, Morris C A, Bertrand J, Robinson B F 1999 Williams syndrome: Findings from an integrated program of research. In: Tager-Flusberg H (ed.) *Neurodevelopmental Disorders*. MIT Press, Cambridge, MA, pp. 65–110

Mills D L, Alvarez T D, St. George M, Appelbaum L G, Bellugi U, Neville H 2000 Electrophysiological studies of face processing in Williams syndrome. *Journal of Cognitive Neuroscience* **12**(11): 47–64

Pani J R, Mervis C B, Robinson B F 1999 Global spatial organization by individuals with Williams syndrome. *Psychological Science* **10**(5): 453–8

Paterson S J, Brown J H, Gsodl M K, Johnson M H, Karmiloff-Smith A 1999 Cognitive modularity and genetic disorders. *Science* **286**(5448): 2355–8

Pinker S 1994 *The Language Instinct*. Penguin, Harmondsworth, UK

Pinker S 1999 *Words and Rules*. Weidenfeld & Nicolson, London

Pober B R, Dykens E M 1996 Williams syndrome: An overview of medical cognitive and behavioural features. *Mental Retardation* **5**: 929–43

Rae C, Karmiloff-Smith A, Lee M A, Dixon R M, Grant J, Blamire A M, Thompson C H, Styles P, Radda G K 1998 Brain biochemistry in Williams syndrome: Evidence for a role of the cerebellum in cognition? *Neurology* **51**: 33–40

Rossen M, Klima E S, Bellugi U, Bihrle A, Jones W 1996 Interaction between language and cognition: Evidence from Williams syndrome. In: Beitchman J H, Cohen J H,

Konstantareas M M, Tannock R (eds.) *Language Learning and Behavior Disorders: Developmental Biological and Clinical Perspectives*. Cambridge University Press, New York, pp. 367–92
Tager-Flusberg H, Boshart J, Baron-Cohen S 1998 Reading the windows to the soul: Evidence of domain-specific sparing in Williams syndrome. *Journal of Cognitive Neuroscience* **10**: 631–9
Tassabehji M, Metcalfe K, Fergusson W D, Carette M J A, Dore J K, Donnai D, Read A P, Proschel C, Gutowski N J, Mao X, Sheer D 1996 LIM-kinase deleted in Williams syndrome. *Nature Genetics* **13**: 272–3
Tassabehji M, Metcalfe K, Karmiloff-Smith A, Carette M J, Grant J, Dennis N, Reardon W, Splitt M, Read A P, Donnai D 1999 Williams syndrome: Use of chromosomal microdeletions as a tool to dissect cognitive and physical phenotypes. *American Journal of Human Genetics* **64**: 118–25
Temple C M 1997 Cognitive neuropsychology and its applications to children. *Journal of Child Psychology and Psychiatry* **38**: 27–52
Thomas M S C, Grant J, Gsödl M, Laing E, Barham Z, Lakusta L, Tyler L K, Grice S, Paterson S, Karmiloff-Smith A 2001 Can atypical phenotypes be used to fractionate the language system? The case of Williams syndrome. *Language and Cognitive Processes* **16**(2/3): 143–76
Udwin O, Yule W 1991 A cognitive and behavioural phenotype in Williams syndrome. *Journal of Clinical and Experimental Neuropsychology* **13**: 232–44
Vicari S, Carlesimo G, Brizzolara D, Pezzini G 1996 Short-term memory in children with Williams syndrome: A reduced contribution of lexical-semantic knowledge to word span. *Neuropsychologia* **34**: 919–25
Vicari S, Bellucci S, Carlesimo G A 2001 Procedural learning deficit in children with Williams syndrome. *Neuropsychologia* **39**(7): 665–77
Volterra V, Capirci O, Pezzini G, Sabbadini L, Vicari S 1996 Linguistic abilities in Italian children with Williams syndrome. *Cortex* **32**: 663–77
Wilson B, Cockburn J, Baddeley A 1985 *Rivermead Behavioural Memory Test*. Thames Valley Test Co., Reading, UK

A. Karmiloff-Smith

Wisdom: Philosophical Aspects

Wisdom is not a central topic of contemporary philosophy. Today's philosophers seem to confirm Nietzsche's observation: 'When philosophers meet among themselves they start casting off all sorts of wonderful rubbish; above all … they hang up "the love of wisdom" like stuffy robes of office' (Nietzsche 1980a, p. 511). Common sense, on the contrary, still expects wisdom to be provided by philosophy, and the discipline's social reputation is partly based on this expectation. But as contemporary philosophy seems scarcely inclined or able to fulfill this demand, common sense draws on forms of wisdom from non-occidental cultures or from esoterics. While 'wisdom' is a term rarely used in contemporary philosophy, the literature about wisdom from the esoteric and the anthropological–cultural side is flourishing.

There is nothing wrong with sensing deficiencies in the academic field and turning to other sources instead. Yet in the second part of this article I will demonstrate that contemporary philosophy can in fact provide a great many elements of wisdom appropriate for our times. The first part is devoted to an accurate reconstruction of the main historical turns in the philosophical understanding of wisdom.

1. *Philosophy—A Concept Introduced by Plato*

As is commonly known, philosophy is related through its name to wisdom: 'philo-sophy' (in Greek: *philo-sophia*) means 'love of wisdom' or 'quest for wisdom' (in Greek: *sophia*). But this concept of *philosophia* corresponds neither to the original understanding of philosophical thinking nor to that of wisdom. The concept of *philosophia* was first created by Plato in the first half of the fourth century BC.

1.1 *The Pre-Platonic Understanding of Philosophy and Wisdom*

Philosophers before Plato neither deployed the term 'philosophy' nor described their enterprise with respect to wisdom (*sophia*). Rather, they characterized their doctrines as 'histories'. Moreover, the Greek understanding of wisdom—as paradigmatically expressed in the topical talk of the 'Seven Sages of Greece'—focused on practical, not theoretical knowledge. In everyday language, a wise person was just an expert of some kind or other (a craftsman, artist, doctor, military leader etc.); in order to be included in the catalogue of the Seven Sages his work had, in addition, to be of extraordinary political relevance. So Thales, the only ancient philosopher ever to figure among the Seven Sages, was included in this group not as a philosopher, but as an astronomer and engineer who brought military success to his community by predicting an eclipse and redirecting a river.

1.2 *Plato's Turn*

What, then, is the specific sense of Plato's introduction of the concept of *philo-sophia*—an innovation so successful that his concept is commonly taken to be *the* concept of philosophy altogether?

Three shifts are characteristic. First, the emphasis in the understanding of *sophia* as well as *philosophia* is put on theory instead of praxis. Philosophers observe the nature of things not for the sake of practical usage but for that of pure cognition. The new concept of philosophy is strictly opposed to the old practical aims of *sophia*. Second, a border line is drawn between

philosophy and wisdom: philosophers can only *strive* for wisdom—without ever actually reaching wisdom. This is why philosophy is defined as *philo*-sophy (*love* of wisdom), not as wisdom proper (*sophia*). 'Philosophy' implies an element of modesty and limitation. Thirdly, this limitation cannot be overcome—complete wisdom is reserved for the Gods only. By philosophizing we do not surpass philosophy to actual wisdom.

1.3 Strategic Implications

Plato's innovation was directed not only against the traditional and the everyday understanding of wisdom but above all against his contemporary competitors, the Sophists. In everyday language at Plato's time—where compositions with '*philo*' had long since become popular—'philosophizing' (*philosophein*) was a term without any extravagant meaning; it just meant studying, educating oneself, being inclined to reflective activity. So Plato's declaration that wisdom is impossible for humans and reserved to the Gods alone discredited not only the mythical Seven Sages but also his compatriots' devotion to philosophical activity and above all the Sophists, who at the time of Plato roved from city to city offering, for payment, lessons in rhetoric, virtue, poetry, mathematics, music or astronomy. As their name (*Sophists*) indicates, they were considered to be bearers of wisdom (*sophia*). The idea that they were in fact frauds—Plato was constantly trying to discredit the Sophists as dealers in fake knowledge, and as being the morally and intellectually dubious counterparts of the real philosopher—was the logical result of Plato's redefinition: If wisdom is altogether unattainable for humans, then the Sophists who masquerade as teachers of wisdom, must obviously be frauds.

2. Philosophy's Claim to Represent True Praxis and Wisdom

The introduction of modesty and limitations in matters of wisdom with the new concept of philosophy was an efficient strategy in distancing the Sophists, but once this was achieved the proponents of the new concept soon turned to more ambitious claims. Philosophy was successively considered capable of representing the utmost fulfillment of both practical and theoretical wisdom.

2.1 Philosophical Theory as the Highest Praxis

The move from a practical to a theoretical emphasis was meant as an empowerment of philosophy in matters of true praxis. The traditional concept of praxis, so characteristic of the original understanding of wisdom, was criticized for having in fact provided only a concept of pragmatic utility, one ignoring the very kind of praxis pertaining to the full accomplishment of human life. This higher and true sense of praxis was to be fulfilled precisely by philosophy in the new sense of theoretical activity. This tendency is obvious already in Plato, who always conceived of philosophy as the best way of human life, with respect to both the philosophizing individual and societies complying with philosophical insights. Aristotle (in his emphatic eulogy of theory in *Nicomachean Ethics* X, 7–9) declared the theoretical life to be the best and highest form of human praxis and life altogether. Significantly enough, he explained the supremacy of theoretical life by contrasting its self-sufficiency with the purposeful orientation of ordinary practical life: according to him, philosophical theory is the only activity 'loved for its own sake; for nothing arises from it apart from the contemplating, while from practical activities we gain more or less apart from the action.'

2.2 Philosophy as Wisdom Proper

Although Plato had initially declared wisdom to be reserved for the Gods, he made the border between human philosophy and divine wisdom permeable. Ultimately philosophizing should result in a *homoiosis theo*, a 'becoming like the divine' (Plato, *Theiatetos*, 176 b 1 f). Aristotle definitely dropped the Platonic reserve. According to him, philosophy and wisdom are one—his appraisal of philosophical *theoria* in *Nicomachean Ethics* X, 7–9 is one of *sophia*. In theoretical life, he says, we realize 'the divine in our midst' and hence our 'true Self'. With Aristotle, '*philosophia*' is no longer a modest term inferior to *sophia*. Hegel's phrase that philosophy should 'lay aside the title "*love* of knowing" and be *actual* knowing' (Hegel 1977, p. 3) is Aristotelean in spirit.

3. Reaccentuating the Practical

3.1 Theory's Open Flank

But philosophy's effort to completely absorb the older practical aspects of wisdom has an open flank. In his theoretical attitude the philosopher may well possess the utmost theoretical knowledge—but this is not to possess all kinds of knowledge: genuinely practical insight, ethical knowledge and situational prudence (*phronesis*) are lacking. And, although theory may represent the highest form of praxis, it cannot as such replace all forms of praxis. Hence, specifically practical knowledge—lying outside theoretical *sophia*—requires respect and clarification in addition. This applies all the more, given that practical knowledge is (as, quite properly, Aristotle was the first to demonstrate) of an intrinsically different type from theor-

etical knowledge. So the new concept of philosophical theory, as impressive as it may be, remains in need of supplementation. Genuinely practical knowledge (which is so decisive for all our everyday orientation) requires a different approach.

3.2 *Hellenistic and Roman Philosophy*: *The Return of the Sage as a Practical Expert*

Following Greek philosophy's classical period, a return to conceiving the sage as an expert in practical matters took place in Hellenistic and Roman philosophy (with the Epicureans and Stoics). The sage once again became a figure of outstanding practical knowledge and exemplariness, knowing how to lead the best life. From now on wisdom was seen as the archetype of accomplished living. Theoretical capacities were to serve this goal.

This understanding of wisdom prevailed even in modern times. So Montaigne defined the sage as someone who appreciates that 'our great and glorious masterpiece is to live appropriately' (Montaigne 1958, p. 851). Kant too still declared that 'the teacher of wisdom through doctrine and example,' 'the practical philosopher,' 'is the real philosopher' (Kant 1992, p. 537). Even today we still expect a wise person to embody the right mode of human living.

4. *Middle Ages*

The Christian philosophy of the Middle Ages had trouble with wisdom. In the Bible the wisdom of this world had been declared to be foolishness before God (*Paul*, 1 *Cor*. 3:19; similar 1 *Cor*. 1:20 and *Romans* 1:22). This did, on the one side, have its advantages over pagan thinking, but on the other side it also called into question the very possibility of reasonable reflection and hence of Christian philosophy altogether. At the least it imposed severe limitations. Some philosophers tried to settle the conflict by developing the doctrine of 'double truth.' But distrust of purely human knowledge—be it wisdom, reason or science—prevailed. Luther was later to renew the accusation that reason is 'the devil's whore' (Luther 1964, p. 164, cf. also Luther 1967, p. 82).

5. *Philosophy in League with Science—Not with Wisdom*

In modern times, philosophy's ally has been science not wisdom. When philosophers, occasionally, do still use the term 'wisdom' what they actually have in mind is science. So Descartes uses 'wisdom' as just another name for the sum total of science (Descartes 1998, p. 65). Even Kant, though affording wisdom the highest rank, emphasizes that 'science … is the narrow gate that leads to the *doctrine of wisdom*' (Kant 1993, p. 171). Husserl was to resume this modern emphasis on science instead of wisdom by declaring 'science has spoken, from now on wisdom has the task of learning' (Husserl 1965, p. 63).

6. *Modern Philosophy's Practical Intentions*

But modern philosophy has a strong practical side too. It comes in two flavors: as technical purposiveness along the lines of science, and as moral-practical orientation taking up older elements of wisdom.

6.1 *Technical–Practical Side*

From the start modern science was addressed to technical ends. It was in this respect that Bacon and Descartes commended it. Marx then claimed that philosophy should become altogether practical—and thus realize itself by vanishing as philosophy. Even if this project has failed, the prospect of technically reshaping reality has proven widely successful.

6.2 *Moral–Practical side*

Complementary to this scientific-technical strain, to which wisdom is irrelevant, practical aspects of wisdom survive in modern philosophy's other strain, the practical one devoted to ethics and morals. It is here, if anywhere, that 'wisdom' can be a substantial term in modern times. Descartes had already conceived of ethics as 'the final stage of wisdom' (Descartes 1983, p. XXIV), and Kant, having defined wisdom as 'the idea of the practical use of reason's complete conformity to law' (Kant 1983), afforded this practical wisdom the highest rank, declaring 'the practical philosopher' to be 'the real philosopher' (Kant 1992, p. 537). In the moral-practical respect wisdom retains the highest esteem.

Enlightenment philosophers, as well as their opponents, agreed that true morality—and by consequence wisdom in its practical sense—is the most important thing for us humans. Rousseau certainly emphasized, in contrast to Kant, the simplicity of moral knowledge, but even this claim to simplicity was shared by a great many philosophers of the Enlightenment. In this they could even see the great advantage of practical over theoretical knowledge: 'It took centuries to know part of nature's laws. A single day suffices to recognize man's duties.' (Voltaire 1967, p. 342) With respect to this simplicity distinguishing practical wisdom, Wolff, D'Holbach and Voltaire proclaimed Confucius, the exemplary 'simple sage', to have been the wisest person ever.

But obviously this modern line of revalidating practical wisdom was not very fortunate. Why else would we feel disoriented and lament the absence of wisdom appropriate to our times and problems?

7. Protest Against Wisdom—Nietzsche

In the course of time modern philosophers came to protest more and more against all traditional understanding of wisdom. Nietzsche is the most prominent example.

He generally suspected the traditional ideal of wisdom to have been simply a symptom of weakness. The old philosophers' 'claim that they possessed wisdom' appears to him to have always been 'a screen behind which the philosopher saves himself because he has become weary, old, cold, hard—as a premonition that the end is near, like the prudence animals have before they die: tey go off by themselves, become still, choose solitude, hide in caves, and become *wise*' (Nietzsche 1974, p. 315). The Stoic is the paradigm case for Nietzsche. His wisdom consists in making himself insensible of the tensions of life and acquiring an armor-plated attitude. Hence Nietzsche's ruthless formula: 'The stoic type. Or: the perfect ox' (Nietzsche 1980b, p. 125).

The attitude of the 'genuine philosopher' would, in Nietzsche's opinion, be completely different: he 'lives "unphilosophically" and "unwisely," above all imprudently, and feels the burden and the duty of a hundred attempts and temptations of life—he risks himself constantly, he plays the wicked game' (Nietzsche 1989, p. 125). 'The wisest person would be the richest in contradictions, who has, at it were, organs of touch for all kinds of person: and in their midst his great moments of grandiose harmony—in us too the lofty coincidence!' (Nietzsche 1980a, p. 182).

Ultimately, however, 'wisdom' remains an improper expression in Nietzsche's view—it is too strongly impregnated by its traditional meanings. Only in a critical manner can we still refer to or draw on wisdom. The first possibility is exemplified by Nietzsche's phrase: 'There is more reason in your body, than in your best wisdom.' (Nietzsche 1966, p. 146) The second is represented by a striking claim of Nietzsche: 'My wisdom has long gathered like a cloud; it is becoming stiller and darker. Thus does every wisdom that is yet to give birth to lightning bolts … For these men of today I do not wish to be light, or to be called light. These I wish to blind. Lightening of my wisdom! put out their eyes!' (Nietzsche 1966, p. 401).

Languid wisdom or lightning wisdom, wisdom as a sedative or as dynamite—this is Nietzsche's alternative in matters of wisdom. If wisdom is referred to, one should be on one's guard as to which type of wisdom is involved.

8. Epistemological Contours

8.1 Shifting from Practical to Theoretical Philosophy

The disappointment with contemporary philosophy's neglect of wisdom is flawed by one-sidedness. One assumes elements of wisdom must be found in the field of practical philosophy, where it would indeed be hard to discover any. Conversely, however, insights of contemporary theoretical philosophy provide a great many elements giving contours to a contemporary concept of wisdom. My thesis is that the design of wisdom today is to be drawn from the epistemological development in recent decades.

8.2 Recent Insights: Theoretical Reason with the Characteristics of Practical Reason

Shifting to epistemology in matters of contemporary wisdom makes all the more sense given that in recent decades theoretical reason has been shown to be characterized by traits formerly considered typical of practical reason alone. Since Aristotle the difference between the two types of reason had been considered fundamental: theoretical reason considers permanent things, whereas practical reason deals with changeable things. Correspondingly, theoretical reason demands distance and provides perfect cognition, whereas practical reason requires involvement and allows only context-dependent cognition. In recent decades, however, theoretical reasoning has been shown to bear the same basic traits as practical reasoning: theoretical cognition too represents relative, not absolute cognition; it too is context-bound and its validity is tied to specific (axiomatic, historical, social, etc.) frameworks. In short, theoretical reasoning also exhibits the characteristics of practical reasoning. This explains why the shape of contemporary wisdom can be inferred from innovations on the theoretical side.

9. Elements of Modern Epistemology Relevant for a Contemporary Concept of Wisdom

As this view is unusual, I will set out in the following the relevant elements in some detail.

9.1 From Nietzsche to Quine: The Uncertainty of Fundaments

Nietzsche can be regarded as a forerunner of the contemporary condition in having pointed out that our theoretical cognition is of a piece with practical and pragmatic cognition: 'one may certainly admire man as a mighty genius of construction, who succeeds in piling up an infinitely complicated dome of concepts upon an unstable foundation, and, as it were, on running water. Of course, in order to be supported by such a foundation, his construction must be like one constructed of spiders' webs: delicate enough to be carried along by the waves, strong enough not to be blown apart by every wind' (Nietzsche 1979, p. 85). In the twentieth century Neurath said quite similarly, 'we are like mariners who must rebuild their ship on the

open sea, without ever being able to disassemble it in dock and reassemble it with the best components' (Neurath 1932/3, p. 206). Popper's diagnosis reads: 'just when we believed that we were standing on firm and safe ground, all things are, in reality, insecure and unstable' (Popper 1992, p. 65); 'the basis is unstable' (Popper 1969, p. 76). Quine later adopted Neurath's phrase as the motto of *Word and Object* (Quine 1960, p. VII).

9.2 *Scientific Pluralism* (*Quine*)

The inevitability of scientific pluralism is a decisive insight of contemporary epistemology that was first formulated by Quine. Even in the strong realm of science no one, singularly valid theory is to be expected, but a multitude of equally valuable alternative candidates has to be taken into account. This is a consequence of the empirical under-determination of theory shown by Quine in his seminal essay 'Two Dogmas of Empiricism' from 1951: 'Total science, mathematical and natural and human, is … under-determined by experience' (Quine 1980, p. 45). Empirical tests can only tell us that something is in need of being altered in the theory, but not exactly which element is to be changed. 'Any statement can be held true come what may. Conversely, by the same token, no statement is immune to revision' (Quine 1980, p. 43).

In *Word and Object* (Quine 1960), Quine resolutely drew the consequences of scientific pluralism: 'we have no reason to suppose that man's surface irritations even unto eternity admit of any one systematization that is scientifically better or simpler than all possible others. It seems likelier … that countless alternative theories would be tied for first place' (Quine 1960, p. 23). 'We can never do better than occupy the standpoint of some theory or other, the best we can muster at the time' (Quine 1960, p. 22).

9.3 *The Impossibility of any Meta-position* (*An Insight Common to Hermeneutics*, *Poststructuralism*, *and Analytic Philosophy*)

This also means that it is impossible to assume some meta-position which would allow us to regard the multitude of scientific theories from a superior point of view and select the 'truly right' one among them. There is no God's-eye standpoint, no view from nowhere—neither in theoretical nor in practical matters. The impossibility of such a meta-stance is a decisive insight of recent philosophy that is common to all (superficially so variant) strands: to Gadamer's hermeneutics (according to which we cannot achieve a metahistorical standpoint towards history), to Lyotard's poststructuralism ('Simplifying to the extreme, I define postmodern as incredulity toward metanarratives,' Lyotard 1984, p. xxiv), and, of course, to analytic philosophy, where this insight was first formulated not by Quine, but by Wittgenstein who declared it his 'leading principle' that there is no 'metaphilosophy' (Wittgenstein 1974, p. 116).

9.4 *Wittgenstein: Limits of Justification*

Wittgenstein has developed insights crucial to the present understanding of cognition and also relevant to every conception of contemporary wisdom altogether.

First, Wittgenstein demonstrated limits of justification with respect to the certainties of common sense. These are neither based on cognition nor could they ever be secured by cognitive justification. For these certainties are by themselves constitutive of all our activities of justification: 'All testing, all confirmation and disconfirmation of a hypothesis takes place already within a system. And this system is not a more or less arbitrary and doubtful point of departure for all our arguments: no, it belongs to the essence of what we call an argument. The system is⋯the element in which arguments have their life' (Wittgenstein 1972, p. 16).

Second, Wittgenstein showed how this structure affects all our decisions: 'Nothing we do can be defended absolutely and finally. But only by reference to something else that is not questioned' (Wittgenstein 1984, p. 16). Decisions necessarily have recourse to something which they take for granted: 'the questions that we raise and our doubts depend on the fact that some propositions are exempt from doubt, are as it were like hinges on which those turn' (Wittgenstein 1972, p. 44). 'If I want the door to turn, the hinges must stay put' (Wittgenstein 1972, p. 44).

So far this means that not only the validity of scientific theories (Quine), but also of our everyday certainties and all our argumentation cannot be conclusively secured, they possess only conditional validity. In the regress of justifications one does not reach some incontrovertible ground, but only postulates, which in principle are once again questionable. While admitting good and varied justifiability, there is no such thing as an absolute, ultimate justification.

Finally, Wittgenstein has also shown how all this applies even to the whole framework of justification and argumentation (to 'rationality'): 'We can't talk of reasons for thinking … We can describe the game of thinking, but not the reasons why we think. "Reason" only applies within a system of rules … It is nonsense to ask for reasons for the whole system of thought' (Wittgenstein 1980, p. 88). What Wittgenstein wants to draw our attention to is that the quest for reasons (for example, for reasons for the whole business of reasoning) suggests leading behind this game, but obviously it does not do so: it asks the very question that typifies this game itself and thus remains within its realm.

Hence there is no possibility of justification which absolutely vouches for itself. This is expressed in the following sentence (which, significantly enough, sounds like the dictum of a sage): 'You must bear in mind that the language-game is so to say something unpredictable. I mean: it is not based on grounds. It is not reasonable (or unreasonable). It is there—like our life' (Wittgenstein 1972, p. 73).

10. Epistemic Contemporariness and Wisdom

These epistemic advances, which characterize the present state of affairs in philosophy, surprisingly also correspond to contemporary standards of everyday consciousness. The latter is molded to a high degree by the acknowledgment that all validity is relative and by an awareness of the plurality of modes of experience, cultures, forms of life and even worlds. Owing to this correspondence, the new epistemic philosophical standards are also apt to serve as maxims for contemporary everyday orientation. They match our expectations and needs in matters of wisdom, and they do so in a manner specifically appropriate to the contemporary situation.

The present understanding of wisdom typically implies circumspection, awareness of limitation, and competence in dealing with uncertainty. The sage is someone who is attentive to neighboring and commonly neglected aspects. She is aware of the context and brings it to bear regulatively. Whereas an ignorant person confidently attributes himself a grasp of the whole, the wise person objects to such totalization. She recognizes the limited nature of all singular perspectives, is aware of differing viewpoints and attentive to the open potential for additional alternatives. She warns against misrecognizing particularities, infringements and overbearingness. She advocates justice at the level of detail and openness on the whole. So the wise person is an expert in uncertainty too. Wisdom is particularly needed in situations where justice is required with respect to diverging demands without a meta-rule being available from which one could deduce the right decision. In these cases, even though it be cognized, but only found, the wise person will have a sense for what is right. Sometimes, however, it will be important just to fend off the wrong thing while leaving open what's right. Such composure also distinguishes the wise person.

Contemporary epistemology, which emphasizes the limits of all justification and demands their respect, is thus properly tailored to the aspects of wisdom just set out. It inherently works towards a consciousness which bears traits of wisdom in high measure.

This is why a state of consciousness that has incorporated those epistemic standards and which knows how to practice them can provide orientation for our time and satisfy our demands in matters of wisdom. Whoever has made this genuinely philosophical possibility her own will be in no great need of extra-philosophical and esoteric doctrines of wisdom that have become so popular in recent years.

See also: Pluralism and Toleration; Relativism, Pluralism, and Skepticism (Philosophical Perspective): Cultural Concerns; Value Pluralism; Wisdom, Psychology of

Bibliography

Aristotle 1984 *Complete Works of Aristotle*. Princeton University Press, Princeton, NJ

Burkert W 1960 Platon oder Pythagoras? Zum Ursprung des Wortes 'Philosophie'. *Hermes* **88**: 159–177

Descartes R 1983 *The Principles of Philosophy*. Reidel, Dordrecht, The Netherlands

Descartes R 1998 *Regulae ad Directionem Ingenii*. Rodopi, Amsterdam

Diogenes L 1931 *Lives of Eminent Philosophers*. Harvard University Press, Cambridge, MA, Vol. 2

Diogenes L 1972 *Lives of Eminent Philosophers*. Harvard University Press, Cambridge, MA, Vol. 1

Goodman N 1978 *Ways of Worldmaking*. Hackett, Indianapolis, IN

Hegel G W F 1977 *Phenomenology of Spirit*. Clarendon Press, Oxford, UK

Husserl E 1965 *Philosophie als strenge Wissenschaft*. Klostermann, Frankfurt, Germany

Kant I 1983 *Anthropologie in pragmatischer Hinsicht*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany [1978 *Anthropology from a Pragmatic Point of View*. Southern Illinois University Press, Carbondale, IL]

Kant I 1992 The Jäsche logic. In: *Lectures on Logic*. Cambridge University Press, Cambridge, UK

Kant I 1993 *Critique of Practical Reason*. MacMillan, New York

Luther M 1964 Wider die himmlischen Propheten, von den Bildern und Sakramenten. In: *Werke. Kritische Gesamtausgabe* (*Weimarer Ausgabe*). Böhlau, Weimar, Germany, Vol. 18

Luther M 1967 Wider der Vernunft Hindernis. In: *Werke. Kritische Gesamtausgabe* (*Weimarer Ausgabe*). Böhlau, Weimar, Germany, Vol. 6

Lyotard J-F 1984 *The Postmodern Condition: A Report on Knowledge*. University of Minnesota Press, Minneapolis, MN

Montaigne M 1958 *The Complete Essays of Montaigne*. Stanford University Press, Stanford, CA

Nietzsche F 1966 *Thus Spoke Zarathustra*. In: *The Portable Nietzsche*. Viking Press, New York

Nietzsche F 1974 *The Gay Science*. Vintage Books, New York

Nietzsche F 1979 On truth and lies in a nonmoral sense. In: Breazeale D (ed.) *Philosophy and Truth. Selections from Nietzsche's Notebooks of the early 1870's*. Humanities Press, New York

Nietzsche F 1980a *Nachgelassene Fragmente. Juli 1882 bis Herbst 1885*. In: Colli G, Montinari M (eds.) *Sämtliche Werke. Kritische Studienausgabe in 15 Bänden*. Deutscher Taschenbuch Verlag, Munich, Germany, Vol. 11

Nietzsche F 1980b *Nachgelassene Fragmente. Herbst 1885 bis Anfang Januar 1889*. In: Colli G, Montinari M (eds.) *Sämtliche Werke. Kritische Studienausgabe in 15 Bänden*. Deutscher Taschenbuch Verlag, Munich, Germany, Vol. 13

Nietzsche F 1989 *Beyond Good and Evil*. Vintage Books, New York

Neurath O 1932–3 Protokollsätze. *Erkenntnis*. **3**: 204–14
Plato 1961 *The Collected Dialogues of Plato*. Pantheon Books, New York
Popper K 1960 *The Logic of Scientific Discovery*. Hutchinson, London
Popper K 1969 *Logik der Forschung*. Mohr, Tübingen, Germany
Popper K 1992 The Logic of the social sciences. In: *In Search of a Better World*. Routledge, London, New York
Quine W 1960 *Word and Object*. MIT Press, Cambridge, MA
Quine W 1980 Two dogmas of empiricism. In: *From a Logical Point of View*. Harvard University Press, Cambridge, MA
Toulmin S 1990 *Cosmopolis*. University of Chicago Press, Chicago
Voltaire 1967 Article on 'philosophie'. In: *Dictionnaire philosophique*. Garnier, Paris
Welsch W 1989 Weisheit in einer Welt der Pluralität. In: Oelmüller W (eds.) *Philosophie und Weisheit*. Schöningh, Paderborn, Germany, pp. 214–59
Welsch W 1995 *Vernunft. Die zeitgenössische Vernunftkritik und das Konzept der transversalen Vernunft*. Suhrkamp, Frankfurt, Main, Germany
Whitehead A 1929 *Process and Reality*. MacMillan, New York
Wittgenstein L 1972 *Über Gewissheit—On Certainty*. Harper, New York
Wittgenstein L 1974 *Philosophical Grammar*. University of California Press, Berkeley, Los Angeles
Wittgenstein L 1980 *Wittgenstein's Lectures. Cambridge, 1930–1932*. Blackwell, Oxford, UK
Wittgenstein L 1984 *Culture and Value*. University of Chicago Press, Chicago

W. Welsch

Wisdom, Psychology of

1. Some Historical Background to the Psychological Study of Wisdom

Since the beginnings of human culture, wisdom has been viewed as the ideal endpoint of human development (Baltes and Staudinger 2000a). Certainly, the psychological study of wisdom is still rather young compared with its philosophical treatment when considering that the very definition of philosophy is 'love or pursuit of wisdom.' Historically, wisdom was conceptualized in terms of a state of idealized being (such as Lady Wisdom), as a process of perfect knowing and judgment as in King Solomon's judgments, or as an oral or written product such as wisdom-related proverbs and the so-called wisdom literature (see also *Wisdom: Philosophical Aspects*). Important to recognize is that the identification of wisdom with individuals (such as wise persons), the predominant approach in psychology, is but one of the ways by which wisdom is instantiated. In fact, in the general historical literature on wisdom, the identification of wisdom with the mind and character of individuals is not the preferred mode of analysis. Wisdom is considered an ideal that is difficult to be fully represented in the isolated individual.

Throughout history, the interest in the topic of wisdom has waxed and waned (Baltes and Standinger in press). In general, two main lines of argument have been in the center of the historical evolution of the concept of wisdom: the distinction between philosophical and practical wisdom, and the question of whether wisdom is divine or human.

In the Western world, these two issues (philosophical vs. practical; divine vs. human) were at the center of wisdom-related discourse during the Renaissance. An initial conclusion of this debate was reached during the later phases of the Enlightenment. Recently, in conjunction with value pluralism and the need for orientation characteristic of postmodern times, interest in the concept of wisdom has been revived. Finally, archeological–cultural work dealing with the origins of religious and secular bodies of wisdom-related texts in China, India, Egypt, Old Mesopotamia and the like has revealed a cultural and historical invariance with regard to wisdom-related proverbs and tales. This relative invariance gives rise to the assumption that concepts such as wisdom with its related body of knowledge and skills have been culturally selected because of their adaptive value for humankind (see also).

Among one of the major reasons for the emergence of the psychological study of wisdom in the late 1970s, early 1980s was the search for the potential of aging or, more specifically, the search for domains or types of intellectual functioning that would not show age-related decline. Life experience and wisdom as well as professional expertise, everyday problem solving or practical intelligence were selected as topics of investigation.

2. Psychological Approaches to the Definition of Wisdom

A first approach to the definition of wisdom from a psychological perspective is its treatment in dictionaries. The major German historical dictionary, for instance, defined wisdom as 'insight and knowledge about oneself and the world … and sound judgment in the case of difficult life problems.' Similarly, the Oxford Dictionary includes in its definition of wisdom 'Good judgment and advice in difficult and uncertain matters of life.'

In a next step, psychologists further specified the content and formal properties of wisdom-related phenomena. These initial efforts for the most part were theoretical and speculative. G. Stanley Hall in 1922, for example, associated wisdom with the emergence of a meditative attitude, philosophic calmness, impartiality, and the desire to draw moral lessons that emerge in later adulthood. Furthermore, writers have emphasized that wisdom involves the search for the moderate course between extremes, a dynamic be-

tween knowledge and doubt, a sufficient detachment from the problem at hand, and a well-balanced coordination of emotion, motivation, and thought. In line with dictionary definitions, such writings refer to wisdom as knowledge about the human condition at its frontier, knowledge about the most difficult questions of the meaning and conduct of life, and knowledge about the uncertainties of life, about what cannot be known and how to deal with that limited knowledge (for an overview, see Sternberg 1990).

2.1 Implicit (Subjective) Theories about Wisdom

Most empirical research on wisdom in psychology so far, has focused on further elaboration of the definition of wisdom. Moving beyond the dictionary definitions of wisdom, research explored the nature of everyday beliefs, folk conceptions, or implicit (subjective) theories of wisdom. The pursuit of answers to questions such as 'what is wisdom?', 'how is wisdom different from other forms of intelligence?', 'which situations require wisdom?', 'what is a wise act?', and 'what are the characteristics of wise people?' have been at the center of psychological wisdom research during the 1980s (for a review, see Sowarka 1989, Sternberg 1990).

These studies in principle build on research initiated by Clayton (e.g., Clayton and Birren 1980). In her work three dimensions were typical of wise people: (a) affective characteristics such as empathy and compassion, (b) reflective processes such as intuition and introspection, and (c) cognitive capacities such as experience and intelligence. The focus of a study conducted in 1985 by Sternberg was the location of wisdom in the semantic space marked by other constructs such as creativity and intelligence. Wisdom was found to be described by six dimensions: reasoning ability; sagacity; learning from ideas and environment; judgment; expeditious use of information; and perspicacity. The greatest overlap was found between intelligence and wisdom. The sagacity dimension, however, was specific to wisdom. In later theoretical work, Sternberg defined wisdom as balancing intrapersonal, interpersonal, and extrapersonal interests to achieve a common good through a balance between adaptation to, shaping of, and selection of environments (Sternberg 1998).

Another major study on subjective theories of wisdom was conducted by Holliday and Chandler (1986). A factor analysis of the attributes judged to be 'most prototypical' of a wise person and wise behavior revealed two factors. One dimension was labeled 'exceptional understanding of ordinary experience,' i.e., combining qualities of the mind with practical virtues of leading a good life. The second factor was labeled 'judgment and communication skills,' i.e., qualities like comprehending, weighing consequences, and giving good advice.

Another study in the tradition of implicit-theory research involved asking subjects to nominate wise people and subsequently characterize nominees (Sowarka 1989). Two findings were of special importance. First, it seemed that the characterization of wisdom and wise persons was a task readily performed by elderly research participants. Second, subjects emphasized the notion that the persons they had nominated as wise displayed 'excellent character.'

Finally, there is initial evidence on the implicit theories about wise acts (Oser et al. 1999). Wise acts seem to be characterized by the following seven features: (a) paradoxical, unexpected; (b) morally integer; (c) selfless; (d) overcoming internal and external dictates; (e) striving towards equilibrium; (f) implying a risk; (g) striving towards improving the human condition.

From this research on implicit theories of wisdom and wise persons, it is evident that people in Western samples hold fairly clearcut images of the nature of wisdom. Four findings are especially noteworthy. First, in the minds of people, wisdom seems to be closely related to wise persons and their acts as 'carriers' of wisdom. Second, wise people are expected to combine features of mind and character, and to balance multiple interests and choices. Third, wisdom carries a very strong interpersonal and social aspect with regard to both its application (advice) and the consensual recognition of its occurrence. Fourth, wisdom exhibits overlap with other related concepts such as intelligence, but in aspects like sagacity, prudence, and the integration of cognition, emotion, and motivation, it also carries unique variance.

2.2 Explicit Theories and Assessment of Wisdom

A more recent line of empirical psychological inquiry on wisdom addresses the question of how to measure behavioral expressions of wisdom. Within this tradition, three lines of work can be identified (Staudinger and Baltes 1994): (a) assessment of wisdom as a personality characteristic, (b) assessment of wisdom in the Piagetian tradition of postformal thought, and (c) assessment of wisdom as an individual's problem solving performance with regard to difficult problems involving the interpretation, conduct, and management of life.

Within personality theories, wisdom is usually conceptualized as an advanced if not the final stage of personality development. Wisdom, in this context is comparable to 'optimal maturity.' A wise person is characterized, for instance, as integrating rather than ignoring or repressing self-related information, by having coordinated opposites, and by having transcended personal agendas and turned to collective or universal issues. Ryff and Whitbourne, for example, have undertaken the effort to develop self-report questionnaires based on the Eriksonian notions of personality development, especially integrity or wis-

dom. Recently, wisdom, in the sense of self-development and maturity, was operationalized using various scales of the California Q-sort and was found to predict life satisfaction in old age (Ardelt 1997).

Central to Neopiagetian theories of adult thought is the transcendence of the universal truth criterion that characterizes formal logic. This transcendence is common to conceptions such as dialectical, complementary, and relativistic thinking. Such tolerance of multiple truths, that is, of ambiguity, has also been mentioned as a crucial feature of wisdom. Empirical studies in this tradition by Labouvie-Vief (1994) or Kramer et al. (1992), found that, at least up to middle adulthood, performances on such measures of adult thought evince performance increases.

Besides these measures of wisdom as a personality characteristic, or as a feature of mature thought, there is also work that attempts to assess wisdom-related performance in tasks dealing with the interpretation, conduct, and management of life. This approach taken by the Berlin wisdom paradigm is based on lifespan theory, the developmental study of the aging mind and aging personality, research on expert systems, and cultural–historical definitions of wisdom (Dittmann-Kohli and Baltes 1990, Baltes and Smith 1990). By integrating these perspectives, wisdom is defined as 'an expert knowledge system in the fundamental pragmatics of life permitting exceptional insight, judgment, and advice involving complex and uncertain matters of the human condition.'

The body of knowledge and skills associated with wisdom as an expertise in the fundamental pragmatics of life, entails insights into the quintessential aspects of the human condition, including its biological finitude and cultural conditioning. Wisdom involves a fine tuned coordination of cognition, motivation, and emotion. More specifically, wisdom-related knowledge and skills can be characterized by a family of five criteria: (a) rich factual knowledge about life, (b) rich procedural knowledge about life, (c) lifespan contextualism, (d) value relativism, and (e) awareness and management of uncertainty (see Baltes et al. 1992 for an extensive definition).

To elicit and measure wisdom-related knowledge and skills, in this approach participants are presented with difficult life dilemmas such as the following: 'Imagine, someone receives a phone call from a good friend who says that she/he can't go on anymore and has decided to commit suicide. What would the person be thinking about, how would the person deal with this situation?' Participants are then asked to 'think aloud' about such dilemmas. The five wisdom-related criteria are used to evaluate these protocols. The obtained scores are reliable and provide an approximation of the quantity and quality of wisdom-related knowledge, and judgment of a given person (see Staudinger and Baltes 1994). When using this wisdom paradigm to study people who were nominated as wise according to nominators' subjective beliefs about wisdom, it was found that wisdom nominees also received higher wisdom scores than comparable control samples of various ages and professional backgrounds.

Part of the Berlin paradigm also is a general framework outlining the conditions for the development of wisdom as it is instantiated in persons (Baltes et al. 1992). The model presents a set of factors and processes which need to 'cooperate' for wisdom to develop. First, there are general individual characteristics related to adaptive human functioning such as intelligence and personality. Second, the model presumes that the development of wisdom is advanced by certain expertise-specific factors, such as a strong motivation to find out about life, practice with difficult life situations, and being guided by a mentor. Third, the model implies the operation of macro-level facilitative experiential contexts. For instance, certain professions and historical periods are more facilitative than others. Further, given the experiences certain social–cognitive processes (life planning, life management, and life review) are assumed to be critical for the development of wisdom-related knowledge and judgment.

The empirical work based on this ontogenetic model and the measurement paradigm, produced outcomes consistent with expectations. Contrary to work on the fluid mechanics of cognitive aging, older adults performed as well as young adults (for a review see Staudinger 1999a). It seems that wisdom-related knowledge and judgment emerges between the ages of 14 and 25 years. Furthermore, when age was combined with wisdom-related experiential contexts, such as professional specialization specifically involving training and experience in matters of life (e.g., clinical psychology), even higher levels of performance were observed. In line with the historical wisdom literature, that portrays wisdom as the ideal combination of mind and virtue, it was found that wisdom-related performance was best predicted by measures located at the interface of cognition and personality, such as a judicious cognitive style, creativity, and moral reasoning. Neither intelligence nor personality independently of each other made a significant contribution to wisdom-related knowledge and judgment. The predictive pattern is of an almost complementary nature when wisdom-related performance in adolescence is considered.

2.3 Is there Wisdom-related Potential?

Researchers of wisdom are usually quite aware that it is a courageous undertaking to try to study wisdom empirically; a complex and content-rich phenomenon which, as many scholars have claimed, defies attempts at scientific identification. However, research on explicit theories of wisdom has shown that it is possible to measure wisdom in terms of a personality characteristics (standardized or open-ended), a characteristic of

adult thought, as well as performance (judgment, advice) on difficult life tasks.

Given the fact, that wisdom-related performance had been successfully operationalized, the question arose of whether it was possible to modify this performance or more specifically to increase wisdom-related knowledge and judgment. It should also be noted that, not surprisingly, the overall level of wisdom identified was below or around the theoretical mean of the scale. Thus, leaving a lot of room for improvement. So far, to our knowledge three studies have been conducted to test these ideas. One was carried out within the framework of Neopiagetian notions of wisdom. It was demonstrated that the level of reflective judgment in adolescence could be raised by presenting examples of higher-level responses (Kitchener et al. 1993).

Within the Berlin wisdom paradigm two different approaches have been successful in activating wisdom-related potential. The first study took into account the fact that wisdom may be too complex a knowledge system to be traceable in an isolated individual. Rather, it may be the case that higher performance levels are found if two minds can interact before responding to the wisdom dilemma. And indeed, when two individuals usually interacting about life problems in everyday life (real dialog) had a chance to do so before they individually reflected and responded to wisdom tasks, the performance level was increased by one standard deviation (Staudinger and Baltes 1996). In line with notions of symbolic interactionism, performance increase was also identified when participants thought about what other people might say while thinking about the problem (virtual dialogue). The second study focused on one of the five wisdom-related criteria, value relativism, and adopted a very successful memory training technique, the method of loci, for use in the realm of wisdom-related performance (Baltes and Staudinger 2000a). Participants were trained to think about life problems located at different regions of the world. At post-test, participants trained in the knowledge activating strategy outperformed the control group by more than half a standard deviation.

3. *Conclusion and Future Directions*

The concept of wisdom represents a fruitful topic for psychological research: (a) the study of wisdom emphasizes the search for continued optimization and the further evolution of the human condition, and (b) in a prototypical fashion, it allows for the study of collaboration between cognitive, emotional, and motivational processes. We expect that future research on wisdom will be expanded in at least three ways: (a) the further identification of social and personality factors and life processes relevant for the ontogeny of wisdom, (b) the exploration of wisdom as a meta-heuristic aimed at orchestrating mind and virtue towards human excellence (Baltes and Staudinger 2000a), and (c) the contribution of wisdom research to building a psychological art of life (Staudinger 1999b).

See also: Adult Cognitive Development: Post-Piagetian Perspectives; Adult Development, Psychology of; Ego Development in Adulthood; Expertise, Acquisition of; Personality Development in Adulthood; Problem Solving (Everyday), Psychology of; Virtue Ethics; Wisdom: Philosophical Aspects

Bibliography

Ardelt M 1997 Wisdom and life satisfaction in old age. *Journal of Gerontology* **52B**: 15–27

Baltes P B, Smith J 1990 Toward a psychology of wisdom and its ontogenesis. In: Sternberg R J (ed.) *Wisdom: Its Nature, Origins, and Development*. Cambridge University Press, New York, pp. 87–120

Baltes P B, Smith J, Staudinger U M 1992 Wisdom and successful aging. In: Sonderegger T (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, Vol. 39, pp. 123–67

Baltes P B, Staudinger U M 2000a Wisdom: a metaheuristic (pragmatic) to orchestrate mind and virtue towards excellence. *American Psychologist* **55**: 122–36

Baltes P B, Staudinger U M in press *Wisdom: The Orchestration of Mind and Character*. Blackwell, Boston

Clayton V P, Birren J E 1980 The development of wisdom across the life-span: a reexamination of an ancient topic. In: Baltes P B, Brim O G (eds.) *Life-span Development and Behavior*. Academic Press, New York, Vol. 3, pp. 103–35

Dittmann-Kohli F, Baltes P B 1990 Toward a neofunctionalist conception of adult intellectual development: wisdom as prototypical case of intellectual growth. In: Alexander C, Langer E (eds.) *Higher Stages of Human Development*. Oxford University Press, New York, pp. 54–78

Holliday S G, Chandler M J 1986 Wisdom: explorations in adult competence. In: Meacham J A (ed.) *Contributions to Human Development*. Karger, Basel, Switzerland, Vol. 17, pp. 1–96

Kitchener K S, Lynch C L, Fischer K W, Wood P K 1993 Developmental range of reflective judgement: the effect of contextual support and practice on developmental stage. *Developmental Psychology* **29**: 893–906

Kramer D, Kahlbaugh P E, Goldston R B 1992 A measure of paradigm beliefs about the social world. *Journals of Gerontology*: *Psychological Sciences* **47**: 180–189

Labouvie-Vief G 1994 *Psyche & Eros. Mind and Gender in the Life Course*. Cambridge University Press, New York

Oser F K, Schenker C, Spychiger M 1999 Wisdom: an action-oriented approach. In: Reich K H, Oser F K, Scarlett W G (eds.) *Psychological Studies on Spiritual and Religious Development*. Pabst, Leugerich, pp. 105–21

Sowarka D 1989 Weisheit und weise Personen: Common-sense-konzepte älterer Menschen [Wisdom and wise persons: common-sense views from elderly people]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* **21**: 87–109

Staudinger U M 1999a Older and wiser? Integrating results on the relationship between age and wisdom-related performance. *International Journal of Behavioral Development* **23**: 641–64

Staudinger U M 1999b Social cognition and a psychological approach to an art of life. In: Blanchard-Fields F, Hess B T (eds.) *Social Cognition, Adult Development, and Aging*. Academic Press, New York, pp. 343–75
Staudinger U M, Baltes P B 1994 The psychology of wisdom. In: Sternberg R J (ed.) *Encyclopedia of Intelligence*. Macmillan, New York, pp. 1143–52
Staudinger U M, Baltes P B 1996 Interactive minds: a facilitative setting for wisdom-related performance?*Journal of Personality and Social Psychology* **71**: 746–62
Sternberg R J 1985 Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology* **49**: 607–27
Sternberg R J (ed.) 1990 *Wisdom: Its Nature, Origins, and Development*. Cambridge University Press, New York
Sternberg R J 1998 A balance theory of wisdom. *Review of General Psychology* **2**: 347–65

U. M. Staudinger

Witchcraft

The scientific study of witchcraft has been concentrated in two research fields clearly separated in time and space: historians' studies of the enigmatic outburst of witches' trials in early modern Europe, and anthropological studies of the role of witchcraft in notably African and Melanesian societies during colonial and postcolonial periods. In her introduction to a conference volume commemorating the 30-year anniversary of Evans-Pritchard's classical study *Witchcraft, Oracles and Magic among the Azande* (southern Sudan), Douglas (1970 p. xiii) noted with some apparent surprise the deep chasm between these two fields of study: 'The anthropologists of the 1950s developed insights into the functioning of witch beliefs which seemed about as relevant to the European experience as if they came from another planet. Dangerous in Europe, the same beliefs in Melanesia or Africa appeared to be tame, even domesticated; they served useful functions and were not expected to run amuck.' In retrospect, one might wish the second half of this quotation to be still true. The growing obsession over the last decades in many parts of Africa and Melanesia with new forms of witchcraft seems to reflect, on the contrary, widespread popular worries that it has 'run amuck.' Yet Douglas's trenchant observation did highlight striking differences between the trajectories of witchcraft studies by historians and anthropologists and, therefore, offers a challenging starting point for any effort to come to a more general survey of the topic.

The differences between the two fields are, indeed, so great that some authors even objected to the use of the same term. Crick (1979), for instance, warned that the term witchcraft was so saturated with implications stemming from European history that anthropologists working on other areas had better avoid the term altogether. One problem with such a puristic stance is that this English term (or the corresponding French, Spanish, and Portuguese terms) have been adopted widely by people throughout the Southern hemisphere. A more substantial objection might be that despite huge variations the representations concerned revolve around basic conceptions that are surprisingly similar. In many parts of the world notions that are now commonly translated as witchcraft refer not only to hidden human aggression, but also to more specific representations about people being able to double themselves, leaving their bodies at night and flying off to nocturnal meetings with their accomplices.

Especially anthropologists, confronted by a bewildering proliferation of imaginaries in this field, have tried to clarify things by formulating strict definitions and distinctions. The best-known is Evans-Pritchard's proposition—on the basis of the distinctions the Zande, the subjects of his first book, supposedly made—to distinguish witchcraft and sorcery: the first referring to an inborn quality, the second to the use of acquired means. But even for neighbouring areas in Africa this distinction proved difficult to maintain in view of the fluidity and dynamics of witchcraft discourses. More recently, some authors (see Geschiere 1997) suggested that this search for clear-cut definitions and classifications risked to make anthropologists blind to the surprising dynamics of these representations and their chameleonic qualities which are crucial to understand how they can adapt themselves so easily to modern changes. In this treacherous field of study, researchers may have to resign themselves to working with fuzzy notions that may exhibit a common core but at the same time engender highly variable and constantly changing elaborations. The differences between the historians' and the anthropologists' versions of witchcraft are illustrative of the widely varying implications these notions can acquire.

The historians' view paid more attention to the dynamics of these representations. After all, a central question to them was why persecutions of witches—quite rare throughout what was sometimes called the 'dark' Middle Ages—acquired epidemic forms in several parts of Western and Central Europe, precisely at the dawn of modern times (sixteenth and seventeenth century). Historians like Robert Mandrou or Norman Cohn explained this as the effect of a longer process of the diabolization of local witchcraft beliefs. Popular worries about perpetrators of evil within the local community (*maleficium*) increasingly became linked to ecclesiastically inspired visions of a Satan's cult. In the confessions of people accused of witchcraft—increasingly obtained under torture by the henchmen of ecclesiastical and secular courts—the Devil emerged as the leading figure at their nocturnal meetings. Some historians related the quite sudden proliferation of witch-hunts after 1500, therefore, to the disciplinary action of the State and the Church that

ever more affected the local communities. Muchembled (1993), for instance, sees a clear distinction in the archives between two discourses: the one of the witnesses who accuse their neighbours of destroying corps, cattle, and even human lives, and the one of the judges who force the accused to relate all this to their dealings with Satan. Thus, Muchembled construes a direct link between the witch trials and the rise of the absolutist State combined with the religious struggles inspired by Reformation and Counter Reformation; he sees these trials as an emanation of the 'forced acculturation' of popular culture by a 'triumphant centralization.' However, in the beginning of the 1970s other historians, mainly working on England, had already advanced another explanation, emphasizing rather the increased willingness of notably rural people to accuse their fellow villagers. Thomas (1971), for instance, related this to the deepening inequalities within rural communities. Richer villagers increasingly resisted older obligations to share with their poorer fellowmen, translating feelings of guilt about this into witchcraft accusations. In this view, State and Church were rather drawn into these affairs by pressures from below.

Thus, there seemed to be a more or less clear contrast between a continental pattern—that for some time also affected Scotland due to dynastic links with Denmark—in which State and Church authorities played a key role with their obsession with the Devil; and a quite different one, prevailing in England, in which locals pressurized the authorities to do something about the *maleficium* of the witches in their midst. More recently, however, this contrast was nuanced by several authors (Briggs 1996, Barry et al. 1996), who emphasized that on the continent as well, tensions and accusations within the local communities played a key role, while direct involvement of State or Church authorities in the persecution of witches was the exception—certainly in the most absolutist States (France, Spain)—rather than the rule. Moreover, they signaled that in England as well, Satan came to play an increasing role in the witches' confession, be it belatedly and less prominently. Yet it is clear, despite such different emphases, that in the historians' interpretations witchcraft could not remain a local issue but had to be studied in interaction with broader power structures, notably the State and the Church. After all, the historians' very sources—mainly court archives—were produced by the latter. An attempt by Ginzburg (1990) to relate the imaginary of the witches' sabbath to a set of folk beliefs—going back to Neolithic times as some sort of substratum throughout Eurasia—received a lukewarm reception and clearly fell outside the historians' main paradigm.

Anthropologists, in contrast, for a long time continued to address witchcraft as a local issue, trying to resolve its riddles by in-depth studies of local relationships and conceptions. Yet their approaches were also subject to quite surprising turns. Evans-Pritchard's Azande study, often quoted as a classic in the field, hardly proved to be a trendsetter. Its focus was on issues of cognition—on why the Azande combined very commensensical explanations with interpretations in terms of occult aggression. But it inspired a series of micropolitical studies in which witchcraft was analyzed—in line with the functionalist ideas that dominated the discipline at the time—in relation to social order and integration. In a series of in-depth monographs of local societies in what at the time was called British Central Africa, anthropologists of the 'Manchester school' studied witchcraft accusations as a form of social control. Typical for this approach was Max Marwick's characterization of witchcraft as 'a social strain gauge': as a measure of sociopolitical tensions, but also as a way to restore social order. It was this interpretation that produced the in retrospect somewhat amazing view of witchcraft as 'tame' and 'domesticated' (see quote from Douglas above).

It is questionable whether this view applied even during the colonial period under the *pax britannica* (c.q. *gallica*). But it is clear that it hardly prepared anthropologists to confront the quite spectacular dynamics of representations of the occult after Independence. It is difficult to decide whether accusations and rumors, let alone actions, in terms of witchcraft increased after Independence (just as the old debate among anthropologists on whether witchcraft 'increased' or 'decreased' after the colonial conquest had no conclusive ending). But it is quite clear that it came out into the open, in many parts of postcolonial Africa. Witchcraft became a hotly debated issue, not only in the street rumors of *Radio Trottoir*, but also in newspapers, on TV, etc. It was striking, moreover, that rumors and references to the occult forces proliferated especially in the more modern sectors of society: in relation to modern forms of entrepreneurship, national politics, sport, and the new institutions for healthcare and education. The view, to many still self-evident, of witchcraft as a 'traditional' relict that will automatically disappear with modernization—'witchcraft will disappear with electric light' as a Dutch missionary in Cameroon claimed in 1971—became difficult to maintain in view of the popular tendency to address precisely such modern developments in terms of these occult discourses. Neither did the anthropological image of witchcraft as a local phenomenon prove to be of much use for understanding the spectacular increase in scale of these rumors that acquired national or even transnational dimensions. No wonder that after the 1960s anthropologists tended to abandon the topic to theologians and journalists. An approach in terms of social order proved to be of little help in analyzing the new transformations of these local beliefs.

In many parts of Africa, for instance, there was increasing popular concern about new forms of witchcraft that were explicitly related to the *nouveaux*

riches who were supposed to accumulate wealth by exploiting the labor of their witchcraft victims. Often people made a contrast between older forms of witchcraft in which the victims were 'eaten' and the new one of 'selling people,' that is, turning them into some sort of zombies and selling them to work on 'invisible plantations.' The fear of such new threats triggered a frantic search for new forms of protection, for instance, urgent appeals to the State authorities to intervene. The pressing character of the witchcraft issue was further highlighted by the lethal witch-hunts that marked the end of Apartheid in the northeastern part of South Africa—young men, encouraged by *nganga* (witch-doctors) and claiming to act as local representatives of the ANC, executing dozens of 'witches' (Niehaus 1993). The half-hearted implications of State authorities in such witch-hunts in many parts of the continent made it all the more urgent to gain deeper insight into the modern dynamics of witchcraft discourses.

No wonder that since the early 1990s there has been a new wave of witchcraft studies in Africa and Melanesia, notably by anthropologists concentrating on the relation with popular views of 'modernity' —that is, the dreams of the modern, in all its varieties, that effect ever wider strata of the population (cf. Comaroff and Comaroff 1993, Geschiere 1997). This linking of witchcraft and modernity, according to some even a new paradigm, implied attention to particular aspects that could explain the continuing resilience of these local beliefs—notably their circular character, their moral ambiguity, and their striking capacity to relate local realities of the family and the home to global processes of change.

The circularity of witchcraft discourses is expressed most vividly in the paradox that the main protection against witchcraft often is to be sought in the very same field. In many parts of Africa, for instance, the *nganga* (healer) is supposed to be able to offer protection and healing only because of his (or her) overdeveloped witchcraft. In some societies there is a consistent effort to conceptually or even institutionally separate the *nganga*'s (or the chief's) occult power from the evil forces of the witches. Then witchcraft seems to be, indeed, unequivocally evil. But in practice, such distinctions between positive and negative forms of occult power are often highly precarious and controversial. In many African societies, for instance, *nganga* remain highly ambivalent figures—healers but at the same time suspect because of their dealings with extremely dangerous forces. The *nganga* is, therefore, also a striking example of the moral ambiguity of these forces. Basically these forces are certainly evil, but—as in the *nganga*'s case—they can be canalized and used in a constructive way. In this more positive sense, witchcraft—it is in such a context that this English translation of African notions becomes, indeed, problematic—can even be seen as a precondition for accumulating wealth and power. This applies especially to the new modes of enrichment that seem to bring out ever more starkly the ambivalence of witchcraft conceptions: the rumors, mentioned above, about the new witchcraft of wealth that turns its victims into zombie-laborers express horror about such treacherous attacks, but also fascination with the new riches this acquired. It is these ambiguities that give witchcraft discourses their elasticity and their surprising capacity to relate the global to the local. In most parts of Africa and Melanesia these discourses still are rooted strongly in the local setting, notably in the intimacy of kinship: witches are supposed to have a special hold over their kin or in any case to work through an ally 'within.' Yet, the same discourses are invoked to cope with the unsettling modern developments: the vagaries of the world-market, the dazzling riches of the happy few and the recurrent mishap of the many in their pursuit of the new forms of wealth. In such contexts, witchcraft rumors can attain truly global dimensions referring, for instance, to the circulation of zombies on a world scale or to the appropriation of extremely potent 'medicine' from Europe or India.

Such anxieties about hidden aspects of modern developments are certainly not special to Africa or Melanesia (cf. the seminal studies by Taussig 1987 and Kapferer 1997 on, respectively, Amazonia and Sri Lanka). The new anthropological interest in the paradoxical link between witchcraft and modernity inspires even an alternative reading of developments in what are thought to be typically 'modern' societies. Already in the 1970s, the anthropologist Favret-Saada (1977) highlighted the surprising resilience of witchcraft practices in the French countryside in a truly innovative analysis of how words are charged with power, but also of how the researcher herself became unavoidably involved. In a more direct relation to modern forms of politics and communication, Geschiere (in Meyer and Pels in press) drew attention to intriguing parallels between witch-doctors and 'spin-doctors' in respectively African and American politics—both manipulating esoteric expertise and 'public secrets' that shroud the exercise of power. In her detailed analysis of rumors on Satanism and child abuse in present-day England, La Fontaine (1998) demonstrated the relevance of anthropological witchcraft studies on Africa and Melanesia in such a modern context. The broader interest of new directions in the study of witchcraft, often still seen as the archetype of the traditional, is precisely that they help to debunk simplistic opposition between modern and traditional, science and magic, transparent and opaque, or even more simply between the West and the rest. Comparative studies, like the one mentioned in this paragraph, show that even in the West modernity retains its own forms of enchantment and opacity.

See also: Magic, Anthropology of; Religion: Evolution and Development; Religion: Morality and

Social Control; Religion, Phenomenology of; Trials: Cultural

Bibliography

Barry J, Hester M, Roberts G (eds.) 1996 *Witchcraft in Early Modern Europe*. Cambridge University Press, Cambridge, UK

Briggs B 1996 *Witches & Neighbours*. Fortuna Press, London

Comaroff J, Comaroff J (eds.) 1993 *Modernity and its Malcontents*. Chicago University Press, Chicago

Crick M 1979 Anthropologists' witchcraft: symbolically defined or analytically undone? *Journal of the Anthropological Society of Oxford* **10**: 139–46

Douglas M (ed.) 1970 *Witchcraft Confessions and Accusations*. Tavistock, London

Favret-Saada J 1977 *Les mots, la mort, les sorts*. Gallimard, Paris

Geschiere P 1997 *The Modernity of Witchcraft, Politics and the Occult in Postcolonial Africa*. University Press of Virginia, Charlottesville, VI

Ginzburg C 1990 *Ecstasies: Deciphering the Witches' Sabbath*. Hutchinson Radius, London

Kapferer B 1997 *The Feast of the Sorcerer*. Chicago University Press, Chicago

La Fontaine J S 1998 *Speak of the Devil, Tales of Satanic Abuse in Contemporary England*. Cambridge University Press, Cambridge, UK

Meyer B, Pels P (eds.) in press *Magic and Modernity*. Stanford University Press, Stanford, CA

Muchembled R 1993 *Le roi et la sorcière: l'Europe des bûchers*. Desclée, Paris

Niehaus I 1993 Witch hunting and political legitimacy: continuity and change in Green Valley, Lebowa 1930–1991. *Africa* **63**(4): 498–529

Stephen M (ed.) 1987 *Sorcerer and Witch in Melanesia*. Melbourne University Press, Melbourne, Australia

Taussig M 1987 *Shamanism, Colonialism and the Wild Man*. Chicago University Press, Chicago

Thomas K 1971 *Religion and the Decline of Magic*. Weidenfeld & Nicolson, London

P. Geschiere

Wittgenstein, Ludwig (1889–1951)

1. Life

Ludwig Josef Johann Wittgenstein was born in Vienna on April 26, 1889 as the youngest of eight children of Karl and Leopoldine (née Kalmus) Wittgenstein. His father was a protestant, his mother a catholic, and the children were brought up in catholic faith. Wittgenstein was of partly 'Jewish' descent (three-quarters according to Nuremberg laws), and he accepted this classification to the point of self-attributing character traits that counted as Jewish in the Habsburg monarchy of the turn of the century. (Weininger's *Geschlecht und Charakter* impressed him deeply.) The family had become one of the wealthiest of the Habsburg Empire in the iron and steel business, and young Ludwig got his early schooling from private teachers until his father decided to send him to the *Realschule* (technical gymnasium) at Linz in 1903, no doubt hoping him to become an engineer, where he passed his '*Matura*' in 1906.

The Wittgenstein home in Vienna was very *fin de siècle*. On the positive side, the house was open to artists and composers, the family exercising generous Maecenatism; Ludwig read intensively in his youth, and there is documented influence from Goethe, Schopenhauer, Heinrich Hertz, Karl Kraus, Wolfgang Köhler, and others. On the negative side, the children were attended more by staff than by their parents, and the sons had definitely self-destructive tendencies. At least two of Ludwig's brothers, possibly a third one, committed suicide, a tendency which Ludwig had to combat himself for most of his life. Looking back at his childhood, he considered it to have been unhappy.

Having graduated from the Linz *Realschule*, Ludwig went to the Technische Hochschule Charlottenburg (now Technische Universität Berlin) where he studied mechanical engineering and was awarded his 'Diplom' in 1908. He went to Manchester in the same year to do research in aeronautics; however, he began to develop a growing interest in pure mathematics and read Bertrand Russell's *Principles of Mathematics* of 1903 and Gottlob Frege's *Grundgesetze der Arithmetik* of 1893. Fascinated by these books' enterprise of deriving mathematics from logic, he visited Frege in Jena in 1911 who encouraged him to study with Russell in Cambridge. He was a student in Cambridge from 1911 to 1914, but although attending some classes in philosophy (notably with George Edward Moore), his work had more or less the character of an extended dialogue with Russell on questions of logic and the foundations of mathematics.

During his Cambridge time, Wittgenstein experienced the first of a number of deep emotional friendships with young men. Although he was homosexual, the relation with David Pinsent probably remained platonic until it was ended by the latter's death in World War I. One love relationship, with Francis Skinner in Cambridge in the 1930s, is known to have been fulfilled sexually, but Wittgenstein broke it off, just as if unfit for happiness in human bonds, and in addition, deeply ashamed of his disposition.

He volunteered for the Austrian army in 1914 and served, with distinction, throughout the war; extant diaries testify to ongoing work that was to lead up to his only philosophical book published during lifetime, the *Logisch-Philosophische Abhandlung*, which was actually finished in summer 1918 and first appeared, very poorly edited, in Ostwald's *Annalen der Naturphilosophie* for 1921. The book became famous only after an English translation, under the title *Tractatus Logico-Philosophicus*, had been published in 1922 when Wittgenstein, after having trained as an el-

ementary school teacher, was teaching children in villages in Niederösterreich—evidently in order to live a morally meritorious life. However, he had to resign in 1926 because of punishing a schoolboy too severely, and he would probably have given way to despair had not his sister, Margarete Stonborough, asked him to build her new Vienna house together with his friend, the architect Paul Engelmann. (The building now houses the Bulgarian embassy.)

Philosophical contact with Cambridge had never been cut off completely. Back in Vienna, Wittgenstein had some philosophical conversations with members of the 'Vienna Circle,' notably Moritz Schlick and Friedrich Waismann. Thus, although a lecture by L. E. J. Brouwer is said to have given the last push, it was not completely out of the blue that Wittgenstein returned to Cambridge in 1928. Having given away his inherited fortune to his brothers and sisters after the war, he had to apply for a fellowship (submitting the now famous *Tractatus* as a doctoral dissertation). This was his way of earning his living until his death, except for holding a chair in Cambridge from 1939 to 1947 and living on savings now and then. He held regular classes in Trinity College, laconically entitled 'Philosophy' or 'Philosophy for Mathematicians.' A difficult person and not popular with most of his colleagues, his teaching success with a small number of students was enormous and accounted for his burgeoning philosophical ideas to be spread far ahead of the posthumous publication of his constantly rewritten notebooks and typescripts.

Except during wartime, Wittgenstein regularly went to Vienna in summer and for Christmas to see his family. Also, he several times went to Skjolden on the Sognefjord in Norway where he had had a little solitary house built before World War I. In 1939, with the assistance of John Maynard Keynes, he succeeded in applying for British citizenship which allowed him to travel in '*Großdeutschland*' even as an Austrian 'Jew' and to arrange for two of his sisters to be declared 'Aryan' in exchange for transferring large valuta parts of the family money to the German Reichsbank. During World War II, he served different jobs outside the university that he considered more important than doing philosophy.

Wittgenstein's second book, the *Philosophical Investigations*, was more or less finished in 1946—finished in the sense of Wittgenstein himself considering it as printable. But a general mood of dissatisfaction with philosophical results seems to have got an increasing grip on him. Even though he continued writing and discussing philosophical questions, he was suspicious of finished answers. Asking the proper questions was, in his view, the important thing—an attitude which he had developed in the 1930s and which he expressed more and more sharply. In this mood, Wittgenstein did not press for publication. Living outside Cambridge for much of his remaining time, he alternately sought company and solitude until cancer of the prostate, diagnosed late in 1949, had worsened so much that traveling became impossible. He died in Cambridge on April 29, 1951.

2. *Philosophical Work*

Wittgenstein's early (*Tractatus*) philosophy can be regarded usefully as an attempt at answering two questions: Why are logical truths true? How can we recognize that they are? He did this by inventing a theory of meaning which, in turn, was based on an ontological theory of the most general features of reality. Starting from the observation that logical truths, as he had come to know them from the work of Gottlob Frege and Bertrand Russell, are always syntactically complex, Wittgenstein first stipulated that there is but one way of constructing complex sentences from components, viz., by combining the components in a complex which is a 'truth function' of the components (if every single component is either true or false, this is sufficient for the complex sentence to be definitely true or definitely false). The components may be complex themselves; but according to a second stipulation, the combinatory process originally starts from components that are elementary sentences, i.e., sentences which are not truth-functional complexes of anything else. If a complex sentence has been constructed from n elementary sentences, there are 2^n different combinations of each of them being either true or false; this is the postulate of mutual independence of the elementary sentences which is characteristic of 'Logical Atomism.' If the complex sentence comes out as true in every such combination, then it is logically true (e.g., 'If it is neither raining nor snowing, then it is not raining' is true, no matter whether or not it is raining and whether or not it is snowing). This is manifest for anybody who understands the complex sentence; thus logical truth 'shows itself,' and both questions above are answered if every elementary sentence is either true or false. Here a third stipulation comes in: Every elementary sentence is a 'picture' of an elementary 'state of affairs' which either exists, in which case the elementary sentence is true, or does not exist, in which case the corresponding elementary sentence is false. Existing elementary states of affairs make up the 'world' as it is. By a picture, Wittgenstein means an isomorphic correlation of the simple signs (which he called 'names') whose concatenations form the elementary sentences and of simple 'objects' which make up the elementary states of affairs. Given a set of elementary states of affairs and a set of elementary sentences, normally there will exist several possibilities to correlate simple signs and simple objects while preserving isomorphy; the actual correlation, according to a further stipulation, is chosen by the speaker who correlates sentences with states of affairs by thinking a state of affairs as a sentence's sense. Since 'names' occur in many different sentences, this think-

ing of the senses of whole sentences reduces the alternative correlations and thus determines certain objects as the meanings of given names which can form other sentences; so the understanding of new sentences can be explained.

So far, Wittgenstein's stipulations add up to a referential and speaker-oriented theory of the meaning of sentences that explains logical truth. However, he added restrictions two of the most provocative of which are the following: First, the states of affairs that make up the world and are necessary for language to be meaningful are all natural ones; there are no moral or aesthetic states of affairs. Evaluative pronouncements are, therefore, nonsensical. Second, the fact that a state of affairs is correlated with a sentence as the latter's sense, or the fact that a speaker correlates sentences with states of affairs (and in consequence, names with objects) by this or that 'projection rule' is not itself a state of affairs, which means that any description of this correlation is nonsensical too. Wittgenstein accepted this consequence by declaring all *Tractatus* sentences nonsensical, even if helpful, in the book's last propositions. On which Frank P. Ramsey is said to have commented 'What you cannot say you cannot say, and you cannot whistle it either' (an allusion to Wittgenstein's popular gift of whistling entire pieces of classical music). A third restriction often goes unnoticed, viz. that the *Tractatus* theory of meaning is restricted to declarative sentences, excluding questions, imperatives, optatives, and interjections. Wittgenstein hoped these might be incorporated into his framework by proper analysis, without, however, attempting to do so; the same holds for his failure to even look for real examples of elementary sentences.

In a period of hard work from 1929 until about 1936 which it has become customary to call 'intermediate,' Wittgenstein renounced Logical Atomism, i.e., the supposition that elementary sentences are mutually independent; all candidates (e.g., simple color statements) proved to exhibit meaning relations with other ones of the same kin (what is green is not red). So the idea of systematic dependency being co-responsible for meaning emerged. This fell in line with the new idea of devising a nonrepresentational (nonsymbolic) conception of meaning: pieces in chess have a meaning (a role), without representing anything, simply in virtue of their place in the rules of the game. If linguistic elements played a role in a system of rules for their use in social interaction, this might well suffice for their meaningfulness. However, the appeal to rules implied a problem of its own. If rules are identified by means of their linguistic expressions, there must be meaningful expressions for there to be meaning (an obvious interpretation regress), and if rules have to be agreed upon there must be rules to confer validity on the agreement procedures (an obvious ruling regress known from constitutional and international law). Sticking to the idea of meaning to be found in systems of rules, Wittgenstein invented one of his most famous tools, the 'language-game,' which he elaborated in the posthumous masterpiece of his 'late philosophy,' the *Philosophical Investigations*. A language game may be understood as a set of regularities in behavior, involving several people, where the use of certain elements (to be identified as linguistic on further inspection) on the one hand depends on what has been done and what the other circumstances are, and on the other hand is relevant for further things to be done subsequently, either by the 'speaker,' 'audience,' or by third parties. These regularities in behavior have to be rich enough to differentiate, e.g., statements from commands; thus 'The window is open,' if playing the role of a command, will evoke responses different from when it is used as an answer to a question.

Commands do not qualify as such simply by being followed by someone else's doing something. They occur only where the speaker has the required kind of authority, and the addressee, rather than actually doing something, has to do it. (This is why utterances with given meanings are possible only in a social setting that provides for the characteristic rights and duties—languages are embedded in 'forms of life.') Both facts can, in the event, be analyzed as certain rules being followed by the people concerned, and it is here that Wittgenstein's solution to the double-regress problem about rules sets in. Rather than asking how to identify or how to establish rules, he asked how behavior looks like if it is rule-following behavior. If behavior is rule-following, the rule it follows can be extracted from it in the following way. There is a description of the behavior that makes use of the rule although the people who follow the rule do not make any use of the rule themselves. With Wittgenstein, a behavior *B* follows the rule *R* if *B* conforms to *R*, if it is learnable and if it goes without saying, i.e., is mutually expected among the people concerned. This is not new to present-day sociology, of course, but it served Wittgenstein to cut off both regresses sketched above. Nobody among the people who follow the rule need be able to identify it, and it need not be enacted, or agreed upon.

Anchoring linguistic meaning in practices that are complex in such a way that they can truly be described as rule-following has the consequence that there is no linguistic meaning outside a community. Whether or not an utterance is a command depends on others to expect a certain person now to do something. Unless they expect the person to do so, the speaker has not uttered a command. Meaning an utterance in a certain way is not something a speakers achieve by exhausting their very own mental resources and capacities (it is not due to their capacity of 'intending' in a traditional sense). Saying something meaningful, as well as meaning something, depend on others to be prepared to understand the speaker in a given way. This implies one of Wittgenstein's most famous theses, viz. that there cannot be a language private to one speaker, i.e.,

a language whose signs have their meanings no matter whether or not others understand them or react to them. However, Wittgenstein argued his critique independently, thus providing indirect evidence for his social account. The only rule available to a 'private meaner' would stipulate something like: '*S* means *x* if and only if I remember *S* to have meant *x* in the past'; since what is to be explained, viz. '*S* means *x*,' occurs in the explanation, the only available rule turns out to be void.

Whereas this famous 'private language argument' has attracted much philosophical debate, a side effect has proved to be more fruitful. Wittgenstein used the language of sensations as his example and so proved, by implication, that it is not possible to report on one's own sensations by way of privileged inner observation. (For in an unproblematic situation of observation, using the right words is a test of the observer's linguistic competence; if others cannot interfere, reports on one's own sensations would have private meanings in the sense defined above.) The idea that there is privileged knowledge of one's own sensations (and other 'conscious' psychological states like moods, emotions, impressions, and the like) originates in the fact that speakers have a special authority in expressing them, and Wittgenstein took two ingenious steps. First, he pointed out that the speaker's privilege is just one case among others of utterances (often prefaced by 'I know' as in 'I myself know best what is good for me') which are privileged according to our social rules rather than epistemically. Second, he interpreted them, rather than as reports, as being meaningful in the same way as nonverbal expressive behavior is, viz. in virtue of playing a recognizable role in a behavioral pattern. Thus, 'Ouch' or 'I am in pain' are meaningful insofar as they occupy the place of crying in the series injury/crying/winning attention (a series that exhibits nicely the structure, which is characteristic of commands, of authority/imperative/expected behavior, or the structure of precondition/utterance/expected reaction that characterizes meaningful utterances in general).

By parity of reasoning, the symmetry of verbal and nonverbal expressive behavior implies a last important result. Just as verbal utterances owe their meanings to others' reactions, so nonverbal behavior expresses what it does in virtue of others' socially established response. Now if one cannot make head or tail of the notion of an inexpressible psychological state (as Wittgenstein argued at length), a person's psychological state is what he or she could express at the moment, and then what that person is feeling, thinking, wishing, etc. is co-defined by how others would react should he or she exhibit expressive behavior. Wittgenstein is, thus, the first philosopher to have touched the topic of what has come to be called 'Social Anti-individualism.' In an unduly neglected paper, Fleming (1978) has likened Wittgenstein's view of the interpretation of expressive behavior to the interpretation of pictures. Pictures owe their content to a civilization's way of seeing them. Their content is manifest because it is defined by established interpretation. Just so, established responses to other persons' behavior are operative in determining their psychological situation. This is what distances Wittgenstein from Logical Behaviorism. Rather than being inferred from behavior, psychological facts are directly observable; and rather than just being ascertained by observation, they are dependent on external interpretation.

Although Wittgenstein developed his 'middle' and later philosophy in opposition to and correction of his *Tractatus* ideas, there is considerable continuity both as regards content and methodology. In point of content, Wittgenstein is sure to be the first philosopher in history who developed his philosophical ideas from thoughts about how it is that the elements of language have meaning. In this respect, our linguistic competence plays a role, with Wittgenstein, not unlike that of our epistemic abilities in Kant's far-reaching deductions outside epistemology. As regards methodology, Wittgenstein insisted, both in the *Tractatus* and in later writings including the *Philosophical Investigations*, that good philosophy has a peculiar task, i.e., that of clarifying obscurities that have been and are constantly being produced, either when scientists misunderstand what they are doing or when scientific or everyday questions and experiences are misunderstood and linguistically distorted by bad philosophy. This continuity has often been stressed, but for at least two reasons it should not be over-estimated. First, the task of clarifying was all that was left over for philosophy in the *Tractatus* since its theory of meaning reserved all meaningful statements for the natural sciences. Second, the assignment was a piece of explicit methodology which Wittgenstein did not obey in practice. In contrast, Wittgenstein's later conception of meaning admits to philosophy every useful language-game (e.g., that of describing language-games), the assignment of the task of clarifying being motivated by his experience of how much linguistic distortion his own early philosophy had suffered from; and clarification meant method rather than methodology—there is sure to exist no other corpus of philosophical writing so densely packed with repeated demonstrations of how translating philosophical questions and theses into everyday language provides clarity. As an original contribution to our picture of what philosophy can achieve, Wittgenstein developed this clarificatory way of handling philosophical thinking into intellectual therapy. The real aim is achieved when the philosopher, by way of rectifying his distorted use of language, has come to see that there has never been any philosophical problem except for the riddles he got entangled with by falling prey to the temptation of highbrow speaking. Wittgenstein went so far as to say that after such therapy, there was no place in philosophy for controversial theses or

theories. Whereas many take these declarations at face value, others have argued that in view of the substantial and controversial achievements of his later philosophy, the declarations should be either interpreted in a restrictive way or taken as overstating an otherwise valuable methodological warning.

3. *Impact*

Wittgenstein's influence on the intellectual life of the twentieth century has not been uniform or coherent; early and late philosophy, written work and teaching contact, work and personality have appealed to different interests. The general public has been and still is mostly interested in his disconsolate life; following the arbitrary turns of fashion, he has been canonized in a way that would, to his utter disgust, have nowadays turned him into an unwilling victim of talk shows. Furthermore, Wittgenstein's prose style has been all the rage since the *Tractatus* was published, and whereas every Tom, Dick, and Harry arranged their texts in decimal notation before 1953, quoting aphorisms from the *Philosophical Investigations* became imperative for feuilletonists in Anglo-Saxonia from that date, and in the German speech area after 1960 when Suhrkamp had timidly published a joint edition of both of Wittgenstein's masterpieces (together with a '*Beiheft*' in order to secure respectability for an unknown Austrian–British philosopher by documenting the interest of poets like Ingeborg Bachmann).

As regards more serious questions, research has been affected differently by the *Tractatus* and by subsequent work, and by the latter differently depending on whether people were confronted with Wittgenstein's teaching in Cambridge or with his writing. The *Tractatus* gained a great popularity soon after the publication of the bilingual edition in 1922, and it exerted considerable influence on some of the members of the 'Vienna Circle' who had gathered around Moritz Schlick and were striving for a really scientific philosophy. The most enduring effect was to be the 'verification theory of meaning' which, as stated in the *Tractatus*, identified any proposition's meaning with the set of distributions of 'true' and 'false' to its elementary sentences which make the proposition true; elaborated by the circle's most influential member, Rudolf Carnap, it made its way into logic, philosophical semantics (the present-day truth-condition semantics), and is even influential in linguistic semantics. Karl Raimund Popper's disclaimers notwithstanding, the *Tractatus* theory that only those sentences have sense which risk to wreck on experience is a clear forerunner of his falsifiability theory of cognitive significance, influential in sociology and psychology since the 1940s.

In his philosophizing since 1929, Wittgenstein influenced some bright Cambridge students who learnt, from his live example, that being sober, honest, and clear was more important than, and at least as attractive as, producing high-faluting and sloppy nonsense. Following the advice to rephrase any philosophical sentence in everyday idiom, they began to form a veritable movement, Ordinary Language Philosophy, which had its high tide in the 1950s and early 1960s. The most famous and most important representative was Gilbert Ryle, Oxford based but on friendly terms with Wittgenstein, whose seminal *Concept of Mind* of 1949 inaugurated Logical Behaviorism, to be superseded by Functionalism; so it is a safe guess that without Wittgenstein, present-day research on the mind-body problem would look entirely different. Further effects resulted from the publication of the *Philosophical Investigations*, jointly in German and English, in 1953; the impact in philosophy was heavy, but surprisingly small in linguistics (the reason being, perhaps, that John L. Austin's speech act theory won the day). At present, the most interesting developments, in the behavioral sciences outside philosophy, seem to be going on in Cognitive Science and in Ethnomethodology. In Cognitve Science, Hubert Dreyfus and others have argued forcefully that intelligent achievements cannot be accounted for by supposing that subjects are following explicit rules but have to follow them implicitly in Wittgenstein's sense. This has been condensed into the prediction that the architecture of the brain's working is to be regarded as connectionist rather than computational. In Ethnomethodology, Peter Winch and others have suggested that without a Wittgensteinian conception of forms of life, an adequate understanding of alien societies cannot be attained.

On the whole, however, Wittgenstein's influence, on the behavioral sciences, has remained marginal outside philosophy. In particular, social anti-individualism is still awaiting proper conversion into research strategies, just as is an adequate use of Wittgenstein's conception of language games in linguistic pragmatics.

See also: Atomism and Holism: Philosophical Aspects; Communicative Competence: Linguistic Aspects; Cultural Expression and Action; Ethnomethodology: General; Expression and Interaction: Microfoundations of Expressive Forms; Expressive Forms as Generators, Transmitters, and Transformers of Social Power; Individualism versus Collectivism: Philosophical Aspects; Interactionism and Personality; Interactionism: Symbolic; Language and Philosophy; Language and Social Inferences; Language and Society: Cultural Concerns; Language, Continental Philosophy of; Logical Positivism and Logical Empiricism; Meaning and Rule-following: Philosophical Aspects; Positivism, History of; Science and Technology Studies: Ethnomethodology; Semantics; Theory of Mind; Truth and Credibility: Science and the Social Study of Science; Truth, Verification,

Verisimilitude, and Evidence: Philosophical Aspects; Verbal Communication, Psychology of; Word Meaning: Psychological Aspects

Bibliography

Baker G P, Hacker P M S 1980 *An Analytical Commentary on the 'Philosophical Investigations,' Vol. I: Wittgenstein, Understanding and Meaning*. Blackwell, Oxford, UK

Baker G P, Hacker P M S 1985 *An Analytical Commentary on the 'Philosophical Investigations,' Vol. II: Wittgenstein, Rules, Grammar and Necessity*. Blackwell, Oxford, UK

Drudis-Baldrich R 1992 *Bibliografia Sobre Ludwig Wittgenstein*. Aporia, Madrid

Fleming N 1978 Seeing the soul. *Philosophy* **53**: 33–50

Fogelin R J 1976/1987 *Wittgenstein*. Routledge and K. Paul, London and New York

Frongia G, McGuinness B F 1990 *Wittgenstein: A Bibliographical Guide*. Blackwell, Oxford, UK

Glock H-J 1996 *A Wittgenstein Dictionary*. Blackwell, Oxford, UK

Griffin J P 1964 *Wittgenstein's Logical Atomism*. Oxford University Press, Oxford, UK

Hacker P M S 1990 *An Analytical Commentary on the 'Philosophical Investigations,' Vol. III: Wittgenstein, Meaning and Mind*. Blackwell, Oxford, UK

Hacker P M S 1996 *An Analytical Commentary on the 'Philosophical Investigations,' Vol. IV: Wittgenstein, Mind and Will*. Blackwell, Oxford, UK

ter Hark M 1990 *Beyond the Inner and the Outer: Wittgenstein's Philosophy of Psychology*. Kluwer, Dordrecht, The Netherlands

Kenny A 1973 *Wittgenstein*. Allen Lane, London

Lapointe F H 1980 *Ludwig Wittgenstein: A Comprehensive Bibliography*. Greenwood Press, Westport, CT

McDonough R M 1986 *The Argument of the Tractatus*. State University of New York Press, Albany, NY

McGuinness B F 1988 *Wittgenstein: A Life, Band I, Young Wittgenstein*. University of California Press, California

McGuinness B F, Schulte J (eds.) 1989 *Logisch-Philosophische Abhandlung*. Frankfurt a.M., Suhrkamp

Monk R 1990 *Ludwig Wittgenstein, The Duty of Genius*. Free Press, London

Mounce H O 1981 *Wittgenstein's Tractatus: An Introduction*. Blackwell Oxford, UK

Philipp P 1996 *Bibliographie zur Wittgenstein-Literatur*. Wittgenstein Archives, Bergen, Norway

von Savigny E 1988–1996 *Wittgenstein's 'Philosophische Untersuchungen.' Ein Kommentar für Leser*. Frankfurt a.M., Klostermann

St. Hilmy S 1987 *The Later Wittgenstein: The Emergence of a New Philosophical Method*. Blackwell, Oxford, UK

St. Shanker V A, Shanker S G 1986 *A Wittgenstein Bibliography*. Croom Helm, London

Schulte J 1989/1992 *Wittgenstein*. State University of New York Press, Albany, NY

Schulte J 1993 *Experience and Expression*. Oxford University Press, Oxford, UK

Stern D 1995 *Wittgenstein on Mind and Language*. Oxford University Press, Oxford, UK

Wittgenstein L 1922/1961 *Tractatus*. Routledge, London

Wittgenstein L 1953/1967 *Philosophical Investigations*. Blackwell, Oxford, UK

Wittgenstein L 1984 *Werkausgabe*. Frankfurt a.M., Suhrkamp

von Wright G H, Nyman H, Schulte J (eds.) in press *Philosophische Untersuchungen*. Frankfurt a.M., Suhrkamp

E. von Savigny

Women and Psychiatry

To understand the role of women in psychiatry, a context is necessary. This will be provided by tracing the history of women in medical roles.

1. *The History of Women in Health Care*

Throughout history women have been healers and caretakers, playing multiple roles, as pharmacists, nurses, midwives, abortionists, counselors, physicians, and 'wise women,' as well as witches.

From 4000 BCE there were women who studied, taught, and practiced medicine, including Moses' wife Zipporah. The first known woman physician, Merit Ptah, practiced in 2500 BCE. The Chinese record, in 1000 BCE, women engaged in typical physician-like activities.

The Roman, Soranas, believed women were divinely appointed to care for women and children. His criteria for practitioners included literacy, anatomic understanding, a sense of patient responsibility, and ethical concerns, particularly regarding confidentiality. However, Roman women physicians were rarely permitted to perform in the same capacities and positions as men.

Women founded and staffed early hospitals in 260–220 BCE; the deaconesses of the early Christian church developed the first public health systems (Shryock 1959). Middle Eastern hospitals were founded in the sixth and seventh centuries CE. By the eighth century Arab medicine began to flourish. Since Muslim men were forbidden from having contact with women who were not family members, women cared for women. In Islamic Spain, women received medical educations and were surgeons before the twelfth century.

European medical education formally began in Salerno, Italy, in the ninth century. By the tenth and eleventh centuries, women apparently studied there (Corner 1937). When medical degrees were first awarded in 1180, they went only to men. Women received a special diploma until the fifteenth century. A woman physician, Trotula, was an important eleventh century academic and teacher in Salerno. She wrote major texts on obstetrics and gynecology and headed a department of women's diseases. Her contributions were significant because, at that time, women provided all obstetrical care and many had no training. However, many of her accomplishments, and those of other women, were attributed to their husbands and sons.

An early twelfth century abbess, Hildegarde of Bingen, was a healer, philosopher, composer, and scientific scholar who provided scientific data on nervous system regulation and an understanding of normal and abnormal psychology.

Initially, women were ambivalently received when they expanded their traditional midwifery roles, and they were prevented from obtaining medical training. However, in England, the need for physicians during the Black Death of 1349 opened the way for women to study medicine.

In 1423 Constanza Calenda became the first woman formally to obtain a medical degree. She lectured in medicine at the University of Naples. Although, in the fourteenth and fifteenth centuries women practiced medicine in Italy, France, England, and Germany, their practice was usually limited to bleeding, administration of herbs and medicines, and reducing fractures, in addition to midwifery. In some parts of Europe, women also practiced as physicans, surgeons, orthopedists, and ophthalmologists. It was generally believed that women were better able to treat women's diseases. Many served kings, royal families, and even the military.

Despite this success, by the end of the fifteenth century the movement to exclude women from practicing gained momentum; they were specifically limited to functioning as midwives. This coincided with the misogyny of Johann Sprenger and Heinrich Kraemer, in the *Malleus Maleficarum* of 1486, which laid out much of the philosophy of the inquisition. Witch-hunting focused on the 'spiritual and mental' inferiority of women. Even when active witch-hunts subsided, their effects remained and women were eliminated from many medical roles.

In the seventeenth century, when examinations and licensing procedures were required for physicians, women were rarely allowed to sit for required examinations.

2. *Women in Eighteenth and Nineteenth Century American and European Medicine*

In colonial America, the healing role of women was critical to survival, since medical care of families was the responsibility of the women. There were no medical schools and relatively few trained physicians. Training was obtained by going abroad or apprenticing.

In 1765 Dr. John Morgan founded the first 'regular' American medical school. By excluding women, it began a tradition of barring them from formal medical training in the USA. Still, many women set up 'irregular' practices without diplomas, having apprenticed in homeopathic or eclectic traditions. Toward the latter part of the nineteenth century, US women began to study in European medical schools.

In 1847, Elizabeth Blackwell became the first woman accepted to a 'regular' US medical school. When she received her MD from the Geneva Medical School, the New York State Medical Association censured the school. In 1850 Harriet K. Hunt, an 'irregular' practitioner in Boston, applied for admission to Harvard Medical School. She was admitted, but denied her seat when the all-male class threatened to leave if women were admitted. It was not until 1946 that Harvard Medical School finally admitted women.

By mid-nineteenth century, three all-female medical colleges were founded in the USA: the Women's Medical College of Pennsylvania, one in Boston (now Boston University), and another in Cincinnati. Several US universities, including Syracuse, Michigan, Iowa, and California accepted women (Luchetti 1998).

During the Civil War, Dr. Mary Edwards Walker was awarded the Medal of Honor, the most prestigious military award, for her performance as a surgeon. She was, however, subjected to enormous criticism of her qualifications. The medal was rescinded in 1917 after a change in guidelines limited it to men in combat. It was reinstated in 1977 (Luchetti 1998).

In 1878 the National Eclectic Medical Association became the first medical society in the USA to accept women for membership, but the American Medical Association (AMA) continued to be closed to women.

Negative attitudes about women physicians were supported by a 'new science' rationalizing the inferior status of women on biological grounds. An 1873 book stated, 'higher education for women produces monstrous brains and puny bodies.' It quoted an 1848 obstetrics textbook stating that women have 'a head almost too small for the intellect but just big enough for love.'

The road continued to be difficult even for those women who managed to obtain training. Ann Preston, founder of the Women's Medical College of Pennsylvania, described the entry of women onto the wards of the Pennsylvania Hospital, 'We entered in a body, amidst jeerings, groaning, whistlings, and stamping of feet by the men students … On leaving the hospital, we were actually stoned by those so-called gentlemen' (in Alsop 1950, pp. 354–5).

Denied hospital appointments, US women began to open their own hospitals and clinics. In 1857 Elizabeth Blackwell, her sister Emily, and Marie Zakrzewska established the New York Infirmary for women and children. In 1862 Marie Zakrzewska founded a similar hospital in Boston. They also founded the Women's Medical College of the New York Infirmary in 1868. They believed that women received an inferior education and they sought to correct this problem by establishing rigorous entrance exams and longer clinical training than other medical schools.

There were many important successes for women at the end of the ninteenth century. Dr. Mary Putman became a member of the New York Academy of Medicine in 1880. In 1875 Emmeline Cleveland, Dean of the Pennsylvania Women's Medical College, be-

came the first woman physician in the USA to perform major surgery.

The role of women in medicine, however, continued to be hotly debated. In 1881 Rachel Bodley, Dean of the Women's Medical College of Pennsylvania, surveyed the 244 living women graduates of the school and found that, despite persistent belief that training for women was wasteful, the overwhelming majority were in active practice, and those who had married reported that their profession had no adverse effect on their marriages, nor had marriage interfered with their work. Many early women physicians were from medical families and they joined family members in practice. By 1876, 0.8 percent of US physicians were women. By 1890, 35 US medical schools admitted women, and in 1900 6 percent of physicians were women (Luchetti 1998).

During this period, women physicians graduated from some European universities in Italy, Switzerland, and the UK. There was a woman chair in obstetrics in Italy.

In 1866 Elizabeth Garrett-Anderson opened a dispensary for women and children in London during a cholera epidemic. People were grateful for medical attention, even from women. In 1874 Sophia Jex-Blake, Elizabeth Garrett-Anderson and others opened the London School of Medicine for Women. The medical hierarchy threatened to blackball any male who taught there, and no teaching hospital would allow women to practice. For several years women could not get the needed training for registration in medicine.

In 1877 the Royal Free Hospital in London allowed women access to its facilities. Dublin accepted women in the early 1880s and the Scottish schools followed in 1895. The Royal College finally opened its exams to women in 1910.

In Canada the situation was no different. Dr. James Barry (1797–1865), Inspector General of the British Army Medical Department and subsequently Inspector General of Hospitals for Canada, was discovered to be a woman after her death. She was the first woman doctor to practice medicine in Canada.

3. *Women in Twentieth Century Medicine*

Despite many changes, negative attitudes about women in medicine persisted. In 1905, the President of the Oregon State Medical Society stated:

> hard study killed sexual desire in women, took away their beauty, brought on hysteria, neuroasthenia, dyspepsia, astigmatism and dysmenorrhea … Educated women could not bear children with ease because study arrested the development of the pelvis at the same time it increased the size of the child's brain and therefore its head. This causes extensive suffering in childbirth. (in Bullough and Voght 1973)

In 1900, women comprised 10 percent of the students at 18 medical schools, 50 percent at Boston University, and 40 percent at Tufts University. Johns Hopkins opened its medical school in 1893 with the financial help of women, who exacted the condition that women be admitted on the same terms as men, based on qualifications. By 1917, Dr. Florence Sabin became the first woman full professor at that institution (Luchetti 1998).

Abraham Flexner, in his Carnegie Commission Report on the status of US medical education (1910), stated: 'Medical education is now, in the United States and Canada, open to women upon practically the same terms as men. If all institutions do not receive women, so many do, that no woman desiring an education in medicine is under any disability in finding a school to which she may gain admittance.' Further, he stated, 'Now that women are freely admitted to the medical profession, it is clear that they show a decreasing inclination to enter it.'

Flexner's Report established medicine as an academic discipline in the USA, with high standards for training and practice. This resulted in the closing of many medical schools, often those that admitted women. Thus, the number of female physicians in the USA leveled off at between 4 and 7 percent until the 1970s, a lower percent than in the vast majority of other countries.

Other countries did not experience this abrupt drop and continued to report higher percentages of female physicians. In 1965, for example, 7 percent of US physicians were women, in the Soviet Union 65 percent, in Poland 30 percent, in the Philippines 25 percent, in the German Federal Republic 20 percent, in Italy 19 percent, in the UK and Denmark 16 percent, and Japan 9 percent (Lopate 1968).

Medicine has been viewed as a male profession in the USA more than in most other countries. Some hypothesize that this occurred because medicine has had higher prestige and income than many other professions. It is also likely that science requirements were not considered appropriate for US women who were thought to be suited to caretaking, but not instrumental activities. The practice of medicine represents a melding of roles, the caretaking functions of physicians fit the 'traditional' views of women's roles, but the technological and instrumental aspects do not. Although cultures conceptualize these issues differently, in Cuba where gender stereotypes are disparaged, traditional roles for women in health care persist and 30–40 percent of physicians are women, while virtually all nurses and midwives are women.

The choices of specialty has remained constant since women were admitted to medical schools in the USA. Women have characteristically entered pediatrics, internal medicine, psychiatry, family practice, and, more recently, obstetrics and gynecology, anesthesiology, dermatology, radiology, pathology, and many medical subspecialties. The surgical fields remain sparsely populated by women.

Regardless of the specialty, academic and administrative appointments as well as other decision-making positions are almost exclusively held by men; women tend to be involved in direct patient care (90 percent of female physicians in the USA). In the 1970s, the fact that women assumed this role was used as an argument for increasing their numbers in medical school. They could be counted on to meet the health care needs of the population.

In countries where women have made significant progress in the health fields, changes have occurred most often in times of war, physician shortages, or major cultural reorganization. In Russia, midwives proved effective doctors in the Russo-Turkish War of 1870, thus beginning the influx of women into medical schools. After the 1917 revolution, the prestige of medicine declined in Russia, and women were admitted to medical school in great numbers. By 1940 62 percent of Soviet physicians were women. As elsewhere, Russian women held a disproportionately small number of senior positions (Nadelson and Notman 1995).

The large numbers of women in medicine in the former USSR, in low prestige positions, reduced the differentiation between doctors and paramedical workers. The *feldscher* (semiprofessional health worker or physician's assistant) was a female role, and apparently remains so in Russia and the other former Soviet republics.

The rise of female health professionals in China occurred along with the reorganization of the medical care system and the society. About half of Chinese physicians are women. In 1950, 50,000 midwives were reeducated as local health workers or 'barefoot doctors' (usually women peasants who receive basic medical training and provide medical care to meet the needs of fellow workers).

The increase in numbers of women physicians in the early 1970s in the USA coincided with the civil rights and women's movements. At the beginning of the twenty-first century almost 50 percent of graduating medical students and over 20 percent of all physicians are women. There continue, however, to be disparities in the leadership roles of women in medicine. A 2000 report (Nonnemaker 2000) indicated that although women graduating from medical schools were more likely to enter academic careers, their numbers at higher academic levels were significantly lower than expected and their rates of advancement were slower than men's.

4. *Women in Psychiatry*

There are historical references to the care of the mentally ill from early classical times. Mental illness was generally framed in terms of religion and there were sanctuaries performing religious rites to treat people. As pagan faiths lost favor, there was a deterioration in the quality of treatment afforded the mentally ill, because mental illness was increasingly seen as caused by supernatural powers, evil spirits, or demons (Zilboorg 1941). It took many centuries before there was institutional care comparable to ancient times (Zilboorg 1941).

The first 'morotrophium' or house for 'lunatics' is recorded in the Byzantine world in the fourth century CE. These institutions began to spread to other areas of the world, generally privately funded. Pilgrimages were made to shrines believed to eject evil spirits. One of the early morotrophiums (house for lunatics) was St. Dymphna, a home for mental healing run by women in 1200 CE.

Over the next several centuries asylums were established in many parts of Europe and the Middle East, often with the support of women. In the sixteenth century, the Sisters of Charity took in the mentally ill when hospitals refused to admit them (Tomes 1994). By the seventeenth century some French institutions began to train nurses to care for these patients. In general, however, care was punitive and often inhumane.

A seventeenth-century Swiss woman physician, Barabara von Roll pioneered in the treatment of mental illness and opened the field of 'psychosomatic medicine' by documenting the connection between physical and mental health. (Brooke 1997).

By the eighteenth century mental illness received greater attention in Europe and the USA. Although the focus was primarily on nosology and categorization, some therapeutic approaches included bloodletting, purgatives, dousing in ice water, and placement of the patient in a spinning chair until the patient lost consciousness. Benjamin Rush, the founder of US psychiatry, was an advocate of the latter method since he believed that congested blood in the brain caused mental illness, and the movement would relieve it (Alexander and Selesnick 1966).

During the nineteenth century, wives of US physicians entertained asylum patients in their homes, to 'calm a fevered mind' (Tomes 1994). They were often responsible for the 'moral' atmosphere of the institutions and advocated respect and attention to promote recovery. Dorothea Dix pioneered efforts to care for the mentally ill and reform inhumane practices.

In the late nineteenth century, about 200 women physicians staffed US asylums (Lunbeck 1994). In 1880, Alice Bennett became the first woman chief of a women's division, Norristown State Hospital. Bennett's unpopular and outspoken opposition to mechanical restraint and neglect of the chronically insane eventually caused her to lose her position.

By the twentieth century increasing numbers of women worked with the mentally ill. The development of psychoanalysis in Europe attracted many women who were important in the development of the field. In 1923, the American Psychoanalytic Association admitted two women, Lucille Dooley and Marion Kenworthy. Karen Horney challenged psychoanalytic

theory, and a school of psychoanalysis evolved around her work.

Over the next few decades, many notable women psychoanalysts emigrated from Europe to the USA. Prior to World War II, 30 percent of European and 20 percent of US psychoanalysts were women (Chodorow 1986). They contributed substantially to the growth and influence of psychoanalysis in the USA. Among the early women leaders were Clara Thompson, Hilda Bruch, Helene Deutsch, Phyllis Greenacre, Judith Kestenberg, Marianne Kris, Margaret Mahler, Edith Jacobson, Stella Chess, Frieda Fromm-Reichmann, and many others who pioneered as researchers and clinicians.

Grete Bibring became the first woman to head a department of psychiatry and become a full professor at Harvard Medical School. She was at Beth Israel Hospital in Boston. Viola Bernard was the first woman APA vice-president; she also led other psychiatric organizations and made pioneering contributions to social psychiatry.

Psychosomatics evolved in the 1930s with the significant contributions of women, such as Therese Benedek and Helen Flanders Dunbar. Marion Kenworthy was notable in exerting an influence on the military, successfully facilitating an increase in the number of mental health units on army bases (Deutsch 1959). In addition, she was a founding member and president of the Group for the Advancement of Psychiatry and became the first woman president of the American Psychoanalytic Association in 1958.

After World War II, women leaders emerged in organizations and academia, including Evelyn Ivey, Eveoleen Rexford, Helen Beiser, Jeanne Spurlock, Mildred Mitchell-Bateman, and June Jackson-Christmas. It is notable that the latter three were African-American.

Much of the leadership in the development of child psychiatry came from women, including those mentioned above. Melanie Klein developed internationally accepted theory concerning early life psychodynamic processes. Anna Freud contributed significantly to the understanding of ego development.

The 1970s marked another increase in numbers of women in psychiatry, coinciding with the women's movement in the USA. Still, women had few leadership roles although 20 percent of psychiatrists were women. In the 1980s and 1990s, women's leadership increased. Carol Nadelson was elected the first woman APA president in 1985, followed by Elissa Benedek and Mary Jane England. Carolyn Robinowitz became senior American Psychiatric Association Deputy Medical Director, and the first woman psychiatrist medical school dean. The APA has had several women officers, including its first women assembly speaker, Donna Norris. Other psychiatric organizations have had a few women leaders but have not developed female leadership in proportion to the increasing numbers of trained women psychiatrists.

Neuroscientist Nancy Andreasen was the first woman editor of the American Journal of Psychiatry. Judith Rappaport became the head of child psychiatry at the National Institutes of Mental Health. In Canada, Judith Gold was the first women president of the Canadian Psychiatric Association. In Australia, Beverly Raphael, the first woman psychiatry chair, became the first woman president of the Royal Australia New Zealand College of Psychiatry. Sarah Romans became the first woman psychiatry chair in New Zealand. Felice Lieh Mak, chair of psychiatry in Hong Kong, was the first woman elected president of the World Psychiatric Association, and Astrid Heiberg became the first woman professor of psychiatry in Norway. Women psychiatrists continue to expand their leadership roles in many countries of the world, and to become significant researchers in evolving areas of psychiatry. However, women's roles in organizational and academic leadership continue to be small.

5. *Female Leadership in Medicine and Psychiatry: The Twenty-first Century*

Although the number of women trainees in psychiatry in the USA approaches 50 percent, and there are substantial numbers of women in junior faculty positions in psychiatry, as in all areas of medicine, there are many fewer women at senior ranks than would be projected by the size of entering cohorts.

Women continue to make slower progress in academic psychiatry as in the rest of academic medicine. They often lack research training and rarely have the mentorship that will enable them to advance their careers. They are, less often than their male colleagues, offered research opportunities during training or at entry level academic positions, and they tend to be assigned clinical, teaching, and administrative tasks that are time-consuming and do not allow them to develop the necessary skills to enable them to progress academically. The pressures are greater for women interested in academic careers since they are in their peak childbearing years when they must also commit themselves to their career development.

Women in academic psychiatry, as in other fields of medicine, are less likely to have peer-reviewed grants or publications than men. Senior women in psychiatry in the USA are relatively productive in terms of research and publications, but they constitute only 5 percent of professors and 2 percent of department chairs in psychiatry. Controlling for relevant variables, women psychiatrists also earn less than men (Leibenluft 1999). This is not substantially different in other countries and in other medical fields.

It is interesting to note that the citation rate for women's articles has been reported to be higher than it is for men (Sonnert and Holton 1996). One report also found that both men and women full-time faculty were equally career motivated and that marital and family

status were not correlated with publication numbers. Being a parent, however, caused greater problems in academic productivity for women than for men (Carr et al. 1998). Women also had less institutional support. Compared to men with children, women with children had fewer publications. Lack of on-site and backup childcare, and other family leave policies presented significant problems for women faculty with children. Women with children had greater dependent responsibilities but similar professional time allocated responsibilities. There were no significant differences between the sexes for faculty without children (Carr et al. 1998).

Women experience a longer time to promotion than men. The difference in research productivity and publications, however, does not account for the large gender difference in promotion to senior ranks (Tesch et al. 1995, Tesch and Nattinger 1997).

Although childbearing and rearing remain key reasons for womens slower progress in their careers, there are other obstacles. Women's commitment, qualifications, and capabilities are questioned and doubted more than men's. Discrimination and the subtle harassment and demeaning of women continues to exist. Women are often related to in transferential roles. They are perceived and treated by many men as mothers, daughters, sisters, etc., making it difficult to also be a colleague and peer (Moss-Kanter 1977, Milwid 1982). This is also observed in group situations where women's opinions have been observed to spark negative and resistant responses, and their credibility in positions of leadership is often questioned (Nadelson 1987, Reich 1989).

A 1965 report in the *Harvard Business Review* suggested that these views are deeply entrenched. They reported that half of the men surveyed were not comfortable working for a woman; one fifth felt that women were temperamentally unfit to be leaders, and half of both men and women did not feel that women could succeed. Twenty years later the majority of men no longer felt that women were temperamentally unfit for management but the women continued to feel pessimistic about acceptance in the corporate world (Bowman et al. 1965, Sutton and Moore 1985).

Interestingly, an AAMC report (Bentsen 1998) stated that medical school deans, when questioned about the qualities necessary for their success, weighed interpersonal skills higher than any others, including academic and clinical experience, in their list of important qualities necessary for medical school leadership. They also highly ranked personality traits, managerial experience and taking pride in the accomplishments of others. Many of these traits are those characteristically attributed to women.

Men and women also have different perceptions of the need for female leadership. A survey at a medical school searching for a dean revealed that 99.9 percent of the women students and faculty and only 50 percent of the men thought that more women leaders were needed in medicine (Scadron et al. 1982). More current data on women in medical leadership implies that this attitude has not substantially changed. The usual reasons given for the lack of female leadership are that the numbers of women available are not sufficient, and/or that they are not motivated because of family responsibilities. The evidence suggests, however, that men and women in similar cohorts have different career trajectories, and that women are not promoted at the same rate as men, despite comparable qualifications (Tesch et al. 1995).

Clinical psychiatry has a rich history of female contributions and leadership. If the field is to continue to evolve and expand its knowledge base, women must play a vital role in research. That 50 percent of US trainees are women suggests that the talents and potential contributions of this substantial group will be lost if only 50 percent, rather than all of them, are seen as potential future researchers.

See also: Equality of Opportunity; Klein, Melanie (1882–1960); Medical Profession, The; Medicine, History of; Psychiatry, History of

Bibliography

Alexander F G, Selesnick S T 1966 *The History of Psychiatry*. Harper and Row, New York

Alsop G F 1950 *History of the Women's Medical College of Pennsylvania*. Lippincott, Philadelphia, PA

Barton W E 1987 *The History and Influence of the American Psychiatric Association*. APPI, Washington, DC

Bonner T N 1992 *To the Ends of the Earth*. Harvard University Press, Cambridge, MA

Bowman G W, Worthy N B, Greyser S A 1965 Are women executives people? *Harvard Business Review* **43**: 14–28; 164

Brooke E 1997 *Medicine Women*. Quest, Wheaton, IL

Bullough V, Voght M 1973 *Bulletin of the History of Medicine* **47** 66–82

Carr P L, Ash A S, Friedman R H, Scaramucci A 1998 The relationship of family responsibilities and gender to the productivity and career satisfaction of medical faculty. *Annals of Internal Medicine* **129**: 532–8

Chodorow N 1986 Varieties of leadership among early women psychoanalysts. In: Dickstein L J, Nadelson C C (eds.) *Women Physicians in Leadership Roles*. American Psychiatric Press, Washington, DC, pp. 45–54

Corner G W 1937 The rise of medicine at Salerno in the twelfth century. In: *Lectures on the History of Medicine: A Series of Lectures at the Mayo Foundation and the Universities of Minnesota, Wisconsin, Iowa, Northwestern and the Des Moines Academy of Medicine, 1926–1932*. Saunders, Philadelphia, PA pp. 371–99

Deutsch A 1959 *The Story of GAP*. Group for the Advancement of Psychiatry, New York

Flexner A 1910 *Medical Education in the United States and Canada: A Report to the Carnegie Foundation for the Advancement of Technology*. Carnegie Foundation, New York. Repr. 1960 Washington Science and Health Publications, Arno Press, New York

Hacker C 1974 *The Indomitable Lady Doctors*. Clarke, Irwin, Toronto, ON, Canada

Leibenluft E 1999 Women in academic psychiatry. In: Jerald K, Silberman E, Pessar L (eds.) *Handbook of Psychiatric Education and Faculty Development*. American Psychiatric Association, Washington, DC, pp. 95–107
Lopate C 1968 *Women in Medicine*. Johns Hopkins University Press, Baltimore, MD
Luchetti C 1998 *Medicine Women*. Crown, New York
Lunbeck E 1994 *The Psychiatric Persuasion*. Princeton University Press, Princeton, NJ
Milwid M E 1982 Women in male-dominated professions: A study of bankers, architects, and lawyers. Ph.D. dissertation, Wright Institute, Berkeley, CA
Moss-Kanter R 1977 *Men and Women of the Corporation*. Basic, New York
Nadelson C C 1987 Women in leadership roles: Development and challenges. *Adolescent Psychiatry* **14**: 28–41
Nadelson C C, Notman M T 1995 Women as health professionals. In: Reich W T (ed.) *Encyclopedia of Bioethics*, 2nd edn. Macmillan, New York, pp. 1713–20
Nonnemaker L 2000 Women physicians in academic medicine: New insights from cohort studies. *New England Journal of Medicine* **342**: 399–405
Reich R R 1989 Wake-up call. *Ms Magazine* **18**: 32–3
Scadron A, Axelrod M, Witte M, Aren C, Meitz E G, Greenberg E A 1982 Attitudes in the medical academic community toward women physician leaders. Paper presented at the AAMC Annual Meeting, October 27, Washington, DC
Shryock R H 1959 *The History of Nursing: An Interpretation of the Social and Medical Factors Involved*. Saunders, Philadelphia, PA
Sonnert G, Holton G 1996 Career patterns of women and men in the sciences. *American Scientist* **84**: 63–71
Sutton C D, Moore K K 1985 Executive women 20 years later. *Harvard Business Review* **63**(5): 42–66
Tesch B, Wood H M, Helwig A L, Nattinger A B 1995 Promotion of women physicians glass ceiling or sticky floor? *Journal of the American Medical Association* **273**: 1022–5
Tesch B, Nattinger X 1997 Career advancement and gender in academic medicine. *Journal of the Irish College of Physicians and Surgeons* **26**: 172–6
Tomes N 1994 *The Art of Asylum Keeping: The Thomas Story Kirkbride Origins of American Psychiatry*. University of Pennsylvania Press, Philadelphia, PA
Zilboorg G 1941 *A History of Medical Psychology*. Norton, New York

C. Nadelson

Women's Health

1. Introduction

Women's health is a broad term referring to physical and mental health problems that are of exclusive concern for women, and which are more common in women or which differ in presentation, severity, or consequences in women compared to men. Women's health is often defined in terms of reproductive health and safety for younger women and in terms of diseases that appear in the female reproductive organs. However, the most prevalent diseases in women are cardiovascular diseases. The universally lower social and economic status experienced by women, compared to that of men, also contributes to poor health and lack of access to care among women. As part of this lower status, violence against women has become a worldwide public health concern.

1.1 Demographic Imperatives

Substantial changes in women's health and life expectancy have occurred over the past 100 years. At the beginning of the twentieth century, in much of the Western industrialized world, life expectancy at birth for women was less than 50 years. Life expectancy at birth for women in most industrialized countries is now well into the late 70s and the early mid-80s, on average about seven to eight years longer than men. By age 65 and over, this gender gap narrows, and by age 85, life expectancy for women is very close to men's. Life expectancy is increasing in all countries. The worldwide birth rate is slowing down, which means that the proportion of older women will increase substantially as well as numbers.

Since most chronic diseases and disabilities occur in older women, in many respects women's health is concerned with older women. However, in developing countries, infectious diseases and maternal mortality remain high relative to those of industrialized countries. Since most of the world's women live in developing countries such as Asia, Africa, and Latin America, the impact of morbidity and mortality from these causes is significant from a global perspective. Cultural practices, views about health, and women's economic and social roles also greatly influence their physical and psycholsocial health. Since women live longer than men in most parts of the world, an array of poorly understood cultural, social, and behavioral consequences follow upon those changes. The increasing interconnectedness of the world by telecommunications and faster travel means that women's traditional roles may be rapidly influenced and changed dramatically by the alternatives presented to women through these channels. It is well documented that education for women can improve their health, increase women's reproductive autonomy, and increase women's life expectancy. The Global Burden of Disease report (Murray and Lopez 1998) has projected a decline in death from maternal mortality and an increase in noncommunicable diseases in women from 28.1 million death (1990) to 49.7 million in 2020. These changes are primarily due to reductions in early death in infancy, childhood, or in young adulthood from infectious diseases and improved prenatal care.

2. Socioeconomic Factors and Women's Health

2.1 Women's Socioeconomic Status

Worldwide, women's incomes are lower, and they are less educated than men, limiting their ability to meet

personal and family economic and social demands (UN 1995). In most of the world, women have primary responsibility for childcare and for maintaining the household (Haw 1995). Most women work either as employees or in subsistence agriculture as providers of food and shelter. In 1994, approximately 45 percent of the world's women aged 15–64 were economically active (International Labor Organization 1994) outside the home. The percentage of households headed only by women in Western Europe increased from 24 percent in 1980 to 31 percent in some areas of the world, it now exceeds 50 percent. In Africa, 60 percent of rural women live below the poverty line (Haw 1995). Total social and economic demands on women are therefore often higher than for men, and their ability to cope with these demands is often constrained by their socially defined roles, creating psychological as well as social conflicts. For example, women's family roles as primary parent may be in conflict with the need to earn a living. Some industrialized countries in Scandinavia and Europe offer flexible work policies, maternal and family leave, but this is far from being the rule.

2.2 Women's Health and Low Socioeconomic Status (SES)

SES is inextricably linked with poor health across the life span in both men and women. The effects of SES on women's health compared to men's health have not been well studied. Health effects of SES in women are likely to be confounded by marital status and by spousal income. Women who are not married are often of lower SES and women who are single heads of households are almost universally of lower SES than women who are married. Thus, it is difficult to separate the effects on health of women's social class from the effects of their spouse's social class. Most infectious and chronic diseases are more prevalent in women of lower socioeconomic status than in women of higher SES. In the United States in 1997 among women aged 65 and over, 12 percent of white women, 27 percent of African American women, and 26 percent of Hispanic women were living below the Federal poverty line (US DHHS 1999).

3. Medical Care

Medical care is a factor in women's health in several ways: (a) access to care is more limited for many women because of their lower economic status, (b) treatment of disease in women may be characterized by less aggressive follow up or by differential lack of recognition of the disease in women compared to men, (c) knowledge about the efficacy of treatments in women is limited by the past lack of clinical trails including women. Federal policies now require inclusion of women in medical research in many developed countries. Women in developed countries use more medical care services than men and are the major decision makers for healthcare utilization for their families. In developing countries, medical care is a more limited resource and so may be preferentially available to men. Reproductive health policies in many developed and developing countries restrict the rights of women to use contraceptives or to seek abortion for unwanted pregnancies. Prenatal care is limited as well, leading to higher maternal mortality rates in women of lower SES and in less affluent countries. Education of women in developing countries has been shown to reduce the birthrate and increase the use of contraceptive measures (World Health Report 2000).

4. Cultural and Ethnic Factors in Women's Health

Being a member of an ethnic group that is not the dominant culture in a country is often associated with a higher risk of disease and death. This happens at least in part because of the lower SES and ethnic discrimination often associated with immigration or minority status. In some instances, cultural or ethnic factors may be protective if traditional culture provides extensive social support and networks or includes better health or dietary practices than the dominant culture. In many cultures, women are denied education, employment, and basic human rights, and immigration may offer the opportunity to improve their circumstances. However, women from more traditional societies may be more dependent on their male partners, and immigration may result in greater isolation and substantial stress related to cultural dissonance of their expected role and the role of women in the dominant society (Carballo et al. 1996).

4.1 Culture and Healthcare

Immigrant and minority women usually face economic, legal, linguistic, and cultural barriers in obtaining healthcare. The higher vulnerability of immigrant and minority women to sexual abuse and violence increases their risk for sexually transmitted diseases (STDs) and posttraumatic stress disorders; issues which are rarely addressed in the healthcare settings in which these women find themselves. Even if a healthcare provider is sensitive to these issues, women may be reluctant to reveal what has happened due to fear and cultural taboos.

4.2 Ethnicity

The social environment encountered by women of color may account for a substantial part of ethnic differences in health. For example, Cooper et al. 1999) have described the higher rates of hypertension in African Americans compared to whites as arising

from their interaction with the social environment. Cross-national comparisons of Jamaicans and African Americans have shown that life expectancy amongst Jamaicans is six years longer than African Americans (74 compared to 70). African American women are at higher risk of cardiovascular diseases and hypertensive diseases. Obesity in African American women is consistently over 50 percent. The prevalence of most genetic factors is relatively low and none have been identified that contribute significantly to ethnic differences in health.

5. *Psychosocial Factors*

Some social and psychological factors significantly contribute to women's health status either directly as risk factors for disease, or indirectly by creating vulnerability to disease. Higher levels of social support have often been linked to better health outcomes in both men and women. Social support may modify the effects of other variables upon the risk of coronary heart disease. This may occur because social support may directly reduce reactivity to acute stressors, or by modifying behavioral risk factors. Social support may be an important factor in buffering the effects of stressors on cardiovascular diseases and blood pressure. Women may respond differently to social support than men, or may access different kinds of social support.

Social isolation has been linked to a higher risk of all-cause and cardiovascular mortality rates in both men and women (Kaplan and Salonen 1988). Social isolation may be more common in older women, although this issue has not been studied globally. Certainly, poorer survival among men leads to a disproportionate number of older women living alone. In the United States in 1990, 35.3 percent of all women aged 60+ were living alone (Casper and Barnett 2000). In countries where the sex difference in survival is greater, social isolation may be greater for women in old age. Social isolation in women is also linked to poorer SES.

5.1 *Psychological Factors*

Gender differences in the influence of traits such as hostility and anxiety on health have not been examined extensively. Matthews and Owens (1998) found that anger suppression and hostile attitudes were associated with higher levels of carotid atherosclerosis in middle-aged women. Knox et al. (1998) reported that coronary heart disease was more prevalent in women with higher levels of hostility. This association was modified by the presence of social support.

5.2 *Health Behaviors and Practices in Women*

Health behaviors such as preventive mammography screening or vaccinations and lifestyle factors such as smoking or obesity or exercise are important determinants of health in women. According to a 1998 survey of health behaviors, about 25 percent of US women aged 55+ are physically inactive and this number increases with age. Obesity is around 50 percent for African American women. Smoking prevalence has declined, but about 15 percent of women in the 55+ age group still smoke. Behavior or lifestyle factors also are associated with compliance with treatment for existing disease. For example, a study of adherence to cholesterol guidelines among women with heart disease revealed that older women, African American women, and married women were less likely to use lipid-lowering medications. Women who exercised, did not smoke, and drank moderate amounts of alcohol were more likely to use lipid-lowering medications. In the United States in 1997, the median percent of women who reported they had had a mammogram in the past year was 73.7 percent (range: 56.5–83.6 percent) 85 percent had a cervical smear and 66.4 percent had both a clinical breast exam and a mammogram. Screening behavior may increase with age as perceived risk increases. Women aged 55+ tend to seek both breast and cervical screening more often than younger women do. Acculturation and ethnicity also can affect health behaviors: Latin-American women who are less acculturated tend to smoke and drink less, exercise less, but are more often obese. Similar patterns are seen for African American women.

5.3 *Violence Against Women*

Violence against women, including rape and domestic assault, has also become a worldwide concern and a public health issue. Violence is related to the lower social and economic status of women compared to men. The World Health Organization (1996) has reported that the prevalence of physical abuse against women ranges from 20 percent to 50 percent based on 40 studies covering 24 countries on four continents. Rape and other forms of violence co-occur frequently. As well as having immediate and lasting physical health effects, such occurrences can affect women's psychological well-being; including self-esteem, poorer perceived health, obsessive disorders, eating problems, difficult interpersonal relationships and depression (Russo and Denious 1997).

6. *Menopause*

Although menopause has been more of a concern in developed countries, as maternal mortality declines and life expectancy increases in the developing world, the rate of growth in numbers of postmenopausal women is projected to increase. At this time, approximately 25 million women worldwide pass through menopause every year. By the year 2030, there will be 1.2 billion menopausal and postmenopausal women,

with an added 47 million new entrants each year (Hill 1996). By the decade 2020–2030, there may be nearly 5 million women reaching age 50 in sub-Saharan Africa, over 7 million in India and over 10 million in China. In the future, the vast majority of postmenopausal women will not be European in ethnicity. However, competing risks from the AIDS epidemic in Africa may well attenuate the numbers of women reaching middle age. As women age, their overall mortality rates and cardiovascular or cancer mortality rates tend to converge with men's rates, but do not generally exceed them. This convergence is often attributed to declining estrogen levels in women. The menopause is also accompanied by a number of other changes in both physical and mental health in women and by changes in social and cultural roles played by women.

Much of the menopause research had included only women of European background who are urban and affluent and relatively healthy (Kaufert 1996). In a number of cultures, menopause frees women from the threat of pregnancy and may even increase their status. Symptoms of menopause may vary by culture as well. The most consistently reported symptoms are hot flushes, ranging from nearly 60 percent in an affluent European sample to seven percent in Karachi. Reports of these symptoms are substantially lower in Japanese women (Lock 1998) compared to US or Canadian women. As well, symptoms associated with menopause by Japanese women are very different and may be more related to aging in general than to menopause *per se*. Menopause is often portrayed as a catalytic or even catastrophic event for women's health. For example, the increase in cardiovascular disease seen in women after menopause is primarily attributed to decrease in estrogen levels. However, other diseases such as breast or colorectal cancer, and dementia, also increase more rapidly after menopause. This fact is not consistent with the view that menopause is the major reason for decline in the health status of women. If she lives long enough, every women undergoes menopause, but most women do not become ill or dysfunctional after menopause. The great variations in cultural and socioeconomic factors influencing health status among women also suggest that the effects of biological menopause on health are no doubt contributory, but do no solely determine an older women's health status. Symptoms of the menopause vary across cultures, again suggesting social and cultural influences on the experience and importance of menopause in health.

7. *Disability and Aging*

Women in areas of the world other than the US and Europe more frequently suffer from lifelong nutritional deprivation, chronic and repeated exposures to infectious diseases, and poorer access to healthcare. Their reproductive experience is also very different, involving multiple births, possibly longer periods of breast feeding, and greater exposures to sexually-transmitted diseases. Older women in the non-European world more often face widowhood, divorce, abandonment, and poverty, and may carry the burden of caring for the elderly and grandchildren. Women spend a greater proportion of their older years with limitations in physical activity than do men with activities of daily living (ADLS) and instrumental activities of daily living (IADLS). The need for caregivers providing assistance with ADLS or IADLS also increases in women with age at a greater rate than it does for men. The percentage increase with age of women with unmet assistance needs is greater than for men, reflecting their greater longevity but also their higher levels of disability (US DHHS 1999). Such factors have a substantial social and economic impact upon the families of these older women.

8. *Conclusions*

The global response to the growing imperative of women's health in developed countries has included national policies which provide for increased research funding, establishment of wider availability of clinical care programs targeting women, and increased allocation in the private sector to treatments targeting women's diseases. Many countries and the World Health Organization have established a number of centers targeting women's health. Although women around the world often live longer and are healthier at younger ages than men, this longevity is often accompanied by greater levels of disability and health problems related to women's lower SES and restrictive traditional roles. In the most economically deprived areas of the world, women's lower status places them at higher risk of death, female infanticide, genital mutilation, socially sanctioned murder or suicide (of unwanted wives, for example), unwanted pregnancies, and sexually transmitted diseases such as AIDS. Globally, womens' health issues range from the subtle problems of menopause to the most glaring issues of outright violence and economic and social injustice.

See also: Care and Gender; Chronic Illness, Psychosocial Coping with; Gender and Cardiovascular Health; Gender and Health Care; Gender and Physical Health; Gender and Reproductive Health; Gender Role Stress and Health; Men's Health; Prostitution; Violence as a Problem of Health

Bibliography

Bureau of Statistics, ILO 1994 Estimates and projections of the economically active population: Provisional results

Carballo M, Grocutt M, Hadzihasanovic A 1996 Women and migration: A public health issue. *World Health Statistics Quarterly* **49**: 158–64

Casper M L, Barnett E 2000 *Women and Heart Disease: An Atlas of Racial and Ethnic Disparities in Mortality*. CDC.
Cooper R S, Rotimit C N, Ward R 1999 The puzzle of hypertension in African-Americans. *Scientific American* **February**
Haw C E 1995 The family life cycle: A forgotten variable in the study of women's employment and well-being. *Psychological Medicine* **25**: 727–38
Kaplan G A, Salonen J T 1988 Social connections and mortality from all causes and from cardiovascular disease: Prospective evidence from eastern Finland. *American Journal of Epidemiology* **128**: 370–80
Kaufert P 1996 The social and cultural context of menopause. *Maturitas* **23**: 169–80
Knox S S, Siegmund K D, Wiedner G, Ellison R C, Abelman A, Paton C 1998 Hostility, social support, and coronary heart disease in the National Heart, Lung and Blood Institute. *American Journal of Cardiology* **82**(10): 1192–6
Kramarow E, Lentzner H, Rooks R, Weeks J, Saydah S 1999 *Health, United States 1999 with Health and Agency Chartbook*. National Center for Health Statistics, Hyattsville, MD
Lock M 1998 Menopause: Lessons from anthropology. *Psychosomatic Medicine* **604**: 410–9
Matthews K A, Owens J F 1998 Are hostility and anxiety associated with carotid atherosclerosis in healthy postmenopausal women? *Psychosomatic Medicine* **605**: 633–8
McDonough P, Williams D R, House J S, Duncan G J 1999 Gender and the socioeconomic gradient in mortality. *Journal of Health and Social Behaviour* **40**: 17–31
Murray C J, Lopez A D 1997 Alternative projections of mortality and disability by cause 1990–2020. *Global Burden of Disease Study*. *Lancet* **349**(9064): 1498–504
Russo M F, Denious J E 1997 Intimate violence and black women's health. *Womens' Health* **3**: 315–48
United Nations. Women in a Changing Global Economy; 1994 World Survey of the Role of Women in Development
WHO 2000 *World Health Report*, 2000. WHO, Geneva
WHO Technical Report 1996. Violence against women, *Fact Sheet N128* WHO, Geneva

M. N. Haan

Women's Religiosity

In Western societies, women are more religious than men on every measure of religiosity. Why this should be has attracted the sporadic attention of clergy, sociologists and psychologists, but it has never been a major interest of either sociologists of religion or feminists. The sociology of religion has only belatedly included gender on its agenda (Woodhead 2000). Feminists have been more concerned with why there are so few women in the priesthood than with why there are so many in the pews. They have also been concerned with documenting the historical antiwomen stance of Christianity, the pre-Christian spirituality of women, and the spirituality of hitherto ignored but highly gifted individuals, rather than documenting the ordinary religiosity of millions of churchgoing women. Feminist theories that religion is bad for women need to come to terms with, but in fact often ignore, women's opting for formal religion in greater numbers than men. This article reviews the possible explanations for women's greater religiosity, focusing particularly on mainline churchgoing in the west in the twentieth century. More detailed reviews may be found in Beit-Hallahmi and Argyle (1996, Chap. 8), Francis (1997), and Walter and Davie (1998).

1. Data

Whether it is a matter of churchgoing, private prayer, or the content of religious belief, women appear more religious than men, and have been for at least 200 years (Douglas 1978). Within Western Christianity, mainline churches demonstrate the greatest fall-off in attendance by men; independent, Pentecostal and Orthodox churches have been able to resist the trend. The extent to which secularization is caused by men reducing religious observance more than women varies by country (Woodhead 2000).

The dimension of religious life where the difference between women and men is biggest is private devotion—prayer and Bible reading. This raises the question whether men engage in religious practices when they are publicly acceptable or even required, but tend not to bother with private devotions when there is no social pressure. It is possible that secularization, defined as the decline in the social importance of religion, tends at the same time to lead to religion's feminization. Islam and Judaism, in which religion is still part of public life, involve men more than women in public worship.

Independent of degree of belief, *what* men and women believe differs. Women, if they are asked to describe the God in whom they believe, concentrate rather more on a God of love, comfort, and forgiveness, while men tend to refer to a God of power, planning, and control. How can such differences be explained?

1.1 Material and Social Deprivation

Neo-Marxist deprivation/compensation theories see religion as a way in which oppressed people cope with their misery and suffering. If, as feminists argue, women are more oppressed than men, we would expect more women than men to turn to religion. There certainly are examples of this, especially in the early periods of industrialization both historically in the West and in currently industrializing societies. In advanced industrial societies, however, there is less evidence for this theory of female religiosity, churchgoing being related to parenthood and employment status rather than to measurable indices of gendered deprivation.

1.2 Lack of Status

Women's systematic exclusion from better paid jobs, from political office, and from other parts of the public world has reduced their social status. Christianity's proclamation that God loves every individual equally may be more immediately attractive to those who are not accorded respect in society. Once inside the church, however, women may find themselves even more deprived of status and power vis-à-vis male members. Yet women remain in greater numbers than men. Ozorak (1996) found that many women valued their churches for the social and emotional support which existed irrespective of their exclusion from formal power; for those more concerned with connection and relationship than with hierarchy and power, the female fellowship of the church empowered them even though the office-holding men monopolized the glory.

Conservative churches and synagogues that emphasize traditional gender roles within the family appeal to many women: they provide status for the wife and mother, and provide hope that conversion may turn their menfolk into responsible husbands and fathers. This is no mean attraction for a woman struggling to rear a family in the slums of a newly industrializing city, whether she be a Methodist in Manchester in 1800 or a Pentecostal in Sao Paolo in 2000.

1.3 Guilt

A number of psychological studies show that women experience more guilt feelings than men. If one of the main purposes of Christianity is to assuage guilt, then we would expect it to attract those who feel a need for forgiveness. Some churches actively foster guilt feelings, which they then offer to resolve. Churches have often portrayed women as more guilty, more profane, than men. Insofar as Christianity induces guilt, we may say that religion not only compensates for a societally created deprivation of status and self-respect, but actually fosters the very deprivation that it then compensates for. This is perhaps the nearest to a feminist theory of why women should be attracted to a religion that oppresses them. Though there is plenty of historical evidence of this kind of process, some recent studies have not found that contemporary churchgoers have high levels of guilt.

1.4 Birth and Child-rearing

Christian faith is supposed to provide comfort against anxiety and existential terror; psychological studies point out that women generally have higher levels of fear than men. This could be related to their more direct involvement in the inherently scary businesses of birth and death and their day-to-day responsibility for child rearing. Men have comparable familiarity with sickness and death only on active service during wartime—a time when they too are more likely to turn to religion. What motherhood does for women, perhaps only war can do for men?

There have been very few studies into the spirituality of childbirth, so it is too early to say whether experiencing the 'miracle' of birth accounts in part for female religiosity. Much more research has been done into religiosity in relation to childrearing. The general pattern is that having children to look after raises men to the religious level of women; it is possible that the vulnerability and dependency of young children raises religious issues that men can normally avoid. It is therefore parenthood, not motherhood, that raises questions of existential anxiety and a desire to involve children in religion.

1.5 Workforce Participation

A major thread in theories of secularization is that the more people are immersed in the institutions of modernity, notably by working for bureaucratic and/or capitalist organizations, the more secular they are likely to become. Women who stay at home are therefore protected from the secularizing influence of the modern marketplace. This hypothesis (classically formulated by Luckmann) has led to a number of empirical surveys aiming to determine whether, and if so why, women who go out to work are less religious than those who stay at home.

The findings of these studies vary by country. Taking a paid job in, for example, Australia where there is a secular ethos in the workplace is likely to have a more secularizing effect than taking a job in the American bible belt. Since both data and hypotheses remain inconclusive, a constructive avenue of research might be to look at the *type* of paid work done by women. Women's paid work is far more likely than men's to be of a caring nature, involving skill in interpersonal relations and 'emotional labor.' If what protects women (and some men) from the secularity of modernity is not family life but interpersonal care, then those employees engaged in caring work should be compared with those engaged in other kinds of work. If research were to confirm that women paid to care are more religious than those paid to do other things we would still be left with a puzzle: are they more religious because the caring nature of the work protects them from secularity, or do they enter caring work because they were more religious to start with?

1.6 An Ethic of Connection

Gilligan (1982) considers that women operate according to a flexible ethic of compassion and connection rather than abstract justice. Woodhead (2000) has used this insight to critique the 'privatization thesis,' namely that in an alienating, plural modern world, people retreat to the private sphere where they find identity, meaning, and … religion. She suggests that this thesis makes sense from a male perspective,

but women are less involved in the alienating public sphere. For them, religion has less to do with providing meaning and more to do with articulating the social relationships that are so important to their lives. Traditional churches do this by encouraging the family, but they can also encourage other kinds of relationships. A religion whose founder emphasized love and compassion is more likely to appeal to Gilligan's women. Indeed, male clergy have often struggled to make Jesus appeal to macho men. Their concern finds an echo in Thompson's (1991) finding that what is correlated with Christian religiosity is not census-bureau gender but a 'feminine' worldview, that is, one that is affectionate, sympathetic, sensitive to the needs of others, eager to soothe another's feelings, tender, and loving toward children.

1.7 Death

The extent to which religion is a way of helping humans deal with death has been extensively debated. What is clear is that women encounter death disproportionately as carers, and also as mourners. In the times of high infant mortality, women may have been more affected than their husbands by the deaths of their children, though men were more likely than women to lose a spouse—as nineteenth century gravestones bear ample witness. Today, with women living longer than men and marrying younger than them, wives are very likely to outlive their husbands. Spousal bereavement is now a much more common experience for women than for men. Whether or not this helps to explain women's greater religiosity, it seems very likely to lie at the root of the most common form of afterlife belief in the modern West, namely a belief in reunion with loved family members. Women who place a high value on personal relationships are unlikely to give them up when death intervenes, and this may possibly account for women being more likely than men to experience the presence of one who is dead.

Insofar as St Paul is correct that life after death is central to Christianity, and insofar as women have a greater interest in there being life after death, we may postulate that some women's religiosity may be rooted in their experience of death and bereavement. On the other hand, it may be the other way around—that more women than men believe in life after death simply because of their general religiosity.

See also: Buddhism and Gender; Feminist Ethics; Feminist Movements; Feminist Theology; Feminist Theory; Religion and Gender; Religion and Health; Religion: Family and Kinship; Religion: Mobilization and Power; Religion: Morality and Social Control

Bibliography

Beit-Hallahmi B, Argyle M 1996 *The Psychology of Religious Behaviour, Belief and Experience*. Routledge, London

Douglas A 1978 *The Feminization of American Religion*. Alfred Knopf, New York

Francis L 1997 The psychology of gender differences in religion: A review of empirical research. *Religion* **27**: 81–96

Gilligan C 1982 *In A Different Voice*. Harvard University Press, Cambridge, MA

Ozorak E 1996 The power but not the glory: How women empower themselves through religion. *Journal for the Scientific Study of Religion* **35**: 17–29

Thompson E 1991 Beneath the status characteristics: Gender variations in religiousness. *Journal for the Scientific Study of Religion* **30**: 381–94

Walter T, Davie G 1998 The religiosity of women in the modern West. *British Journal of Sociology* **49**: 640–60

Woodhead L 2000 Feminism and the sociology of religion: From gender-blindness to gendered difference. In: Fenn R (ed.) *The Blackwell Companion to the Sociology of Religion*. Blackwell, Oxford, UK

G. Davie and T. Walter

Women's Suffrage

The campaign for women's suffrage has been the most visible feature of that broader historical movement for ending women's subordination to men that we now call feminism. Demands for women's vote and full citizenship have arisen as part and parcel of efforts to democratize political participation, during the transition from absolutist monarchies to democratic republics. These efforts first began in western Europe and North America where citizenship, including the right to vote, was accorded first to adult men, initially as propertied heads of households and increasingly, from the mid-nineteenth century onward, as male individuals. Advocates of women's suffrage demand that the franchise be accorded without distinction of sex, in parallel with other campaigns that argued for ending exclusions based on property qualifications (i.e., class) or race. They argued that women constitute half the human race and must not be excluded from political decision-making.

1. Women's Suffrage and Democratization

In some European societies, critiques of women's subordination—by women and men alike—antedated by several centuries the demands for women's suffrage that erupted with the advent of the French revolution in 1789. From that time forth, however,voting quickly became a symbol of full citizenship and personal independence—for women as well as for men. Women's suffrage advocates saw the vote not as a goal in itself but rather as the key to all other reforms, a means for women to acquire the political power necessary to improve or transform their disadvantaged situation in the short run (primarily through improved educational and economic opportunities), and, in the

long run, of dismantling the laws, structures, attitudes, practices, and knowledge that undergirded male dominance. Women's suffrage advocates argued that women's participation in community decision-making could be effective at various levels: in school boards, in workers' or business councils, in local, regional, or provincial affairs. But the most desirable and seriously contested form of voting in the newly emerging representative governments was at the national, federal, or parliamentary level. It was there that questions about the national budget and taxation, and about war, peace, and social and economic policy were being decided. It was also there—where power, authority, and control over collective resources were most concentrated—that the strongest resistance to woman suffrage would be found. The problem suffrage advocates faced in these new male-dominated regimes was to convince already established and exclusively male bodies to acknowledge women's capabilities and equality and to admit them to full citizenship, not as a privilege but as a right. In virtually all countries women suffrage advocates included some sympathetic men. In some countries, however, antisuffrage forces mobilized substantial numbers of supporters of both sexes on the grounds that women should not be permitted in, or were unsuited for, public affairs. In their view, public business was exclusively men's business, and women, as legal dependents, should consider their interests represented by husbands, fathers, or guardians.

Despite very substantial resistance, the twentieth century ultimately witnessed the mass enfranchisement of women throughout the world, along with proclamations of (and sometimes even the granting of) their equal legal, educational, and economic rights. The United Nations Convention on the Political Rights of Women (1952) firmly underscored women's equal right to vote. Only in authoritarian societies where political representation and community decision-making have not been permitted has the demand for woman suffrage been muted. At the beginning of the twenty-first century, only a few sovereign states in the Middle East, such as Saudi Arabia (where no one votes) and Kuwait, have unconditionally refused to enfranchise women. What political power the vote actually provides to women or men in different countries, however, varies greatly, depending on the size and strength of political parties, the modes of election, and many other factors.

2. *History of Women's Suffrage*

The first explicit arguments for women's participation in political decision-making were expressed at the outset of the French Revolution, initially by the Marquis de Condorcet in 1787 and 1790, and in a number of anonymous pamphlets that turned these arguments into demands to be admitted to exercise the *droit de cité*, denouncing masculine aristocracy and challenging men's right to make the rules for women. As revolutionary French assemblies proceeded to declare the Rights of Man, to draft constitutions and reshape old laws for a new society, individual women such as Olympe de Gouges and men such as Pierre Guyomar insisted on women's full inclusion in the decision-making process. After four years of resisting eloquent arguments for women's inclusion, revolutionary political leaders shut women out of political life on the grounds that women and government should not be mixed, an argument that had centuries-old roots. These demands were widely repeated in the 1830s and again, in early 1848, when leaders of the revolutionary Second Republic decreed universal (manhood) suffrage, and a group of Parisian women immediately organized to protest their exclusion. Deliberate choices of words from 'male' to the more subtle '*en âge viril*' signaled the exclusion of women from political affairs, and campaigns ensued to change the wording of electoral laws, as from 'male' to 'person' or to stipulate 'and women' following 'men,' or to add the term 'of both sexes' to qualify the masculine 'Français.' Despite these early beginnings, however, French women only acquired the vote in 1944; the French campaign for woman suffrage began earlier than elsewhere and lasted far longer. Resistance to women in positions of authority was deeply embedded in France. There the 'Rights of Man' were constantly qualified on grounds of 'public utility' when it came to women's rights.

In the English-speaking world, governing elites began to re-examine their electoral laws in the late eighteenth- and early nineteenth centuries. Demands for women's suffrage were strongly expressed during the 1820s. The Reform Act of 1832 in England added the word 'male' to qualify electors, even as it broadened the franchise greatly, and in the next few decades a number of other European countries moved to do the same. The campaign for a Second Reform Act in the late 1860s mobilized women to seek inclusion. John Stuart Mill brought their demands to change 'male' to 'person' to the floor of Parliament during his brief tenure as representative of Westminster; although parliamentary suffrage was not forthcoming, qualified women did obtain a local vote in 1869. The campaigns in England would continue into the twentieth century, when they rose to a new peak, thanks to the combined efforts of the National Union of Woman Suffrage Societies (NUWSS) and the more militant activities of the Women's Social and Political Union (WSPU). The spectacular and extensive campaigns of the WSPU from 1906 through 1913, featuring mass parades, public demonstrations, occasional incidents of violence against state property, and the imprisonment and force-feeding of suffrage militants by British authorities gained world media attention and set a benchmark by which all other campaigns were subsequently judged. The cause of woman suffrage

became a decisive political issue for both England and Ireland, and Irish suffrage advocates did not soon forgive their parliamentary deputies, who in 1912 sacrificed the cause of British women's suffrage in order to obtain Irish home-rule. Ironically, Irish women did receive the vote along with English women who met certain property or educational qualifications in 1918; 'universal' suffrage including all women did not become a reality in the United Kingdom until 1928.

In the United States, democracy without women began to emerge in the 1820s, as various states slowly lifted property qualifications on white male voters. The demand for the vote was a key feature of the 'Declaration of Sentiments' issued by the first women's rights convention, held in Seneca Falls, New York, in 1848. Because US electoral laws were made at the state level, not at the federal level, suffrage advocates faced the challenge of organizing suffrage campaigns on a state-by-state basis, following their exclusion from the federal amendment that enfranchised black men after the Civil War. The territory of Wyoming enfranchised women in 1869, and Utah territory in 1870, but these proved exceptions until new western states granted votes to women in the 1890s. Resistance to woman suffrage remained strong in the eastern, central, and southern states. In the United States, racial and immigration issues complicated the problem still further and from the passage of the Fourteenth Amendment to the Civil Rights Movement of the 1960s, even black men's vote in southern states was often qualified by literacy requirements, poll taxes, or threats of physical violence. Finally, in the early twentieth century, suffrage leaders decided to attempt the federal amendment approach again; following a huge prosuffrage campaign, reminiscent of the British suffrage movement, Congress passed the nineteenth amendment, which was successfully ratified by a majority of states and became law in 1920. Democracy, even without women, was not an unproblematic development, even in the US; indeed, few nation-states had wholeheartedly embraced the principle of full democratic participation in government by the end of the nineteenth century, even for all men.

The first nation-states to enact full woman suffrage were New Zealand (1893) and two Australian states (1894, 1899). Originally British colonies, these South Pacific outposts of the British Empire—like the western states of the USA—proved more favorable to women's citizenship than the older, more settled, and more authoritarian societies of Europe and Asia.

In some twentieth-century continental European countries, where the demographic balance had tilted and adult women began to outnumber men significantly, the prospect of convincing men in power to enfranchise women became even more complicated. This complication was compounded by concern among ruling elites over Marxist-socialist enthusiasm for enfranchising all women along with all men, irrespective of social class (endorsed by the Second International in the 1890s); the Socialist Women's International endorsed unrestricted suffrage for women in 1907. It was further compounded by longstanding concerns among otherwise liberal, even progressive men that women would vote as the priests dictated, to the detriment of progressive (often anticlerical) secularizing regimes (especially after the Roman Catholic Church muted its earlier opposition to woman suffrage following World War I). Meanwhile, Finnish women obtained the vote alongside their menfolk in 1906, when Finland obtained a degree of independence from the Russian Empire; by 1913 Norwegian women were also fully enfranchised, followed by Danish women in 1915. Suffrage campaign strategies would differ markedly, depending on whether men had already been enfranchised in a particular setting, and to what degree.

3. *The International Woman Suffrage Alliance*

The International Woman Suffrage Alliance (IWSA), founded in 1902, brought national organizations of European and North American women together to support each other's campaigns in a more politically astute, systematic fashion. By 1913, the IWSA had affiliated societies from most European countries (including Bulgaria, Hungary, Romania, Russia, and Serbia), and from the far-flung corners of the British Commonwealth—Australia, Canada, and South Africa. The 1920s would see a remarkable extension of IWSA's network, to include delegations from Latin America, Egypt, China, and other non-Western countries. But the organization's leadership became increasingly torn between the need to focus on suffrage in countries where women did not yet have the vote, and the need to advance progressive women's agenda for change in the countries where they could already vote. Issues of married women's nationality, combating state-regulated prostitution and the traffic in women and children, and fostering peace were the focus of considerable activity by IWSA's board and member organizations. Controversy over protective legislation for women workers vs. absolute equality in the workplace led to schism in the organization in 1926, and to a reorientation of focus.

4. *Women's Suffrage from the 1920s to the 1950s*

In some countries, women sometimes gained the vote with far less of a struggle than in France, England, or the United States. After World War I, for instance, new governments accorded woman suffrage as part of universal suffrage in a number of European countries, defeated during the war, that had acquired new, progressive regimes—Weimar Germany, the Austrian Republic, Hungary, etc. The Russian Revolution in

1917 brought women the vote, but the subsequent rule of the Communist Party ensured that they would have no choice in whom to vote for. Western and Central European political parties began to develop an interest in women voters, at the same time attempting to keep them at arm's length from party decision-making positions or candidacy for office. The various socialist parties elected the most women to office, but even then their leadership was reluctant to allow women representatives much leeway in deviating from the party line, which continued to insist on noncooperation with 'bourgeois' feminists. Catholic political parties were equally reticent when it came to incorporating women into party leadership. In a time when adult women very often outnumbered men, all parties were understandably interested in whether or not women might vote as a bloc. Several initiatives to form women's political parties and coalitions (e.g., England, Sweden, Austria) did not succeed, and overall it seems to be the case that during the early years, women's voting patterns resembled those of men, and often they were somewhat more conservative. This did not deter fascist governments from striking out against elections in general, and in particular, against the ostensible threat of women's active participation in political life.

In Latin American countries, where voting, even for men, was often restricted by birth, property, literacy, marital status, or racial qualifications, suffrage became a priority for women in some countries and not in others. Woman suffrage became an important issue in Cuba, Uruguay, and Brazil, and in Mexico during the late 1920s and early 1930s, but seemed less urgent in other Latin American states, where social justice issues were paramount concerns for political women. Miller (1991, p. 96) observes that 'effective universal suffrage, male and female, did not exist anywhere until after World War II.' But Lavrin (1995) emphasizes the breadth and importance of woman suffrage campaigns in Argentina, Uruguay, and Chile, accompanied by the emergence of women's parties.

Only after 1945 did women's suffrage find a place in the political cultures of most Asian countries. Indonesia and Japan enfranchised women in 1945, while Chinese women gained the vote in 1947. India enfranchised women in 1949, following independence (though limited suffrage for some women had existed since 1919). Successful decolonization campaigns have led to women's enfranchisement in many other countries. (A comprehensive chronology of women's suffrage can be found in Daley and Nolan 1994).

5. *Scholarship on the Women's Suffrage Movement*

Scholarship on the history of the women's suffrage movement has expanded substantially since the 1960s, thanks in part to the development of social history and the inclusion of women's movements in the study of political movements and collective action and also to the dramatic growth of women's studies as an academic field. Contextualization of the various national and local suffrage movements has been an important objective of recent studies, along with comparative studies of the international women's movement (Rupp 1997). Scholars have also published comprehensive biographical studies of some of the internationally prominent suffrage movement leaders, including Elizabeth Cady Stanton, Harriot Stanton Blatch, Carrie Chapman Catt, Millicent Garrett Fawcett, Hubertine Auclert, and Aletta Jacobs, and have pressed for the recognition of women's suffrage leaders in national and international biographical dictionaries. Others have begun to examine the cultural history (theatre, literature, arts, campaign propaganda, public manifestations, even trials) of the most expansive woman suffrage campaigns. The literature on the British and American suffrage movements is voluminous; only the most recent works will be listed below (bibliographies in these works will lead readers to the earlier and classic works; see also the bibliographies in Offen 2000 and Daley and Nolan 1994). Documentary films have brought the British and American suffrage campaigns to the attention of English-speaking television audiences (MacKenzie 1988, Ward et al. 1999), and commemorative museum exhibits organized in London and various American states have attracted sizeable audiences. In other countries the movement for women's suffrage has also begun to generate a growing historical literature, in a variety of languages, even as political scientists debate the results that woman suffrage has had in various political contexts. Comparative analyses of women's suffrage movements are still in their early stages (see Bolt 1993 and Bader-Zaar 1994), and a global assessment of the long-term significance and consequences of women's suffrage remains to be undertaken.

See also: Citizenship: Sociological Aspects; Education and Gender: Historical Perspectives; Feminist Movements; Feminist Theory: Liberal; Gender and Feminist Studies in Political Science; Nation-states, Nationalism, and Gender; Political Sociology; Social Movements and Gender; State, Sociology of the; Voting, Sociology of

Bibliography

Bader-Zaar B 1994 *Vergleichende Aspekte der Geschichte des Frauenstimmrechts in Grossbritannien, den Vereinigten Staaten von America, Österreich, Deutschland und Belgien, 1860–1920*. Ph.D. Dissertation. University of Vienna

Bolt C 1993 *The Women's Movements in the United States and Britain from the 1790s to the 1920s*. University of Massachusetts Press, Amherst, MA

Bosch M, Kloosterman A (eds.) 1990 *Politics and Friendship: Letters from the International Woman Suffrage Alliance, 1902–1942*. Ohio State University Press, Columbus, OH

Capel R M 1992 *El Sufragio femenino en la Segunda República*. Comunidad de Madrid, Madrid
Cohen Y, Thébaud F (eds.) 1998 *Féminismes et identités nationales: Les Processus d'intégration des femmes au politique*. Centre Jacques Cartier, Lyon, France
Corbin A, Lalouette J, Riot-Sarcey M (eds.) 1997 *Les Femmes dans la cité*. Créaphis, Grâne, France
Daley C, Nolan M (eds.) 1994 *Suffrage and Beyond: International Feminist Perspectives*. Auckland University Press, Auckland; New York University Press, New York; Pluto Press, London
DuBois E C 1997 *Harriot Stanton Blatch and the Winning of Woman Suffrage*. Yale University Press, New Haven, CT
DuBois E C 1998 *Woman Suffrage and Women's Rights*. New York University Press, New York
Fagoaga C 1985 *La Voz y el voto de las mujeres: El sufragismo en España, 1877–1931*. Icaria, Barcelona, Spain
Fauré C 1997 *Encyclopédie politique et historique des femmes: Europe, Amerique de Nord*. Presses Universitaires de France, Paris
Grimshaw P 1987 [orig. 1972] *Women's Suffrage in New Zealand*. Auckland University Press, Auckland; Oxford University Press, Oxford, UK
Hardemeier S 1997 *Frühe Frauenstimmrechtsbewegung in der Schweiz (1890–1930): Argumente, Strategien, Netzwerk und Gegenbewegung*. Chronos, Zurich, Switzerland
Hause S C, Kenney A R 1984 *Women's Suffrage and Social Politics in the French Third Republic*. Princeton University Press, Princeton, NJ
Holton S S 1986 *Feminism and Democracy: Women's Suffrage and Reform Politics in Britain, 1900–1918*. Cambridge University Press, Cambridge, UK
Holton S S 1996 *Suffrage Days: Stories from the Women's Suffrage Movement*. Routledge, London
Joannou M, Purvis J (eds.) 1999 *The Women's Suffrage Movement: New Feminist Perspectives*. Manchester University Press, Manchester, UK
Lavrin A 1995 *Women, Feminism, and Social Change in Argentina, Chile, and Uruguay, 1890–1940*. University of Nebraska Press, Lincoln, NB
MacKenzie M 1988 *Shoulder to Shoulder: A Documentary. The Stirring History of the Militant Suffragettes*. Vintage Books, New York [Accompanies the documentary film]
Miller F 1991 *Latin American Women and the Search for Social Justice*. University Press of New England, Hanover and London
Murphy C 1989 *The Women's Suffrage Movement and Irish Society in the Early Twentieth Century*. Temple University Press, Philadelphia
Offen K M 2000 *European Feminisms, 1700–1950: A Political History*. Stanford University Press, Stanford, CA
Oldfield A 1992 *Woman Suffrage in Australia: a Gift or a Struggle*. Cambridge University Press, Cambridge, UK
Prentice A, Bourne P, Brandt G C, Light B, Mitchinson W, Black N 1988 *Canadian Women: A History*. Harcourt Brace Jovanovich, Toronto
Rupp L J 1997 *Worlds of Women: The Making of an International Women's Movement*. Princeton University Press, Princeton, NJ
Sepe C, Izzi Di Paolo P (eds.) 1997 *Il voto alle donne cinquant'anni dopo*. Ufficio Progretti Donna, Commune di Roma
Smith P 1996 *Feminism and the Third Republic: Women's Political and Civil Rights in France, 1918–1945*. Clarendon Press, Oxford, UK
Van Wingerden S A 1999 *The Women's Suffrage Movement in Britain, 1866–1928*. Macmillan, Houndmills, UK
Ward G C, Saxton M, Gordon A D, Dubois E C, Burns K 1999 *Not for Ourselves Alone: The Story of Elizabeth Cady Stanton and Susan B. Anthony. An Illustrated History*. Knopf, New York [Accompanies the documentary film]

K. Offen

Word Classes and Parts of Speech

There is a long tradition of classifying words, for the purpose of grammatical description, into the ten word classes (or parts of speech) noun, verb, adjective, adverb, pronoun, preposition, conjunction, numeral, article, interjection. While each of these terms is useful, and they are indispensable for practical purposes, their status in a fully explicit description of a language or in general grammatical theory remains disputed. Although most of the traditional word class distinctions can be made in most languages, the cross-linguistic applicability of these notions is often problematic. Here I focus primarily on the major word classes noun, verb, and adjective, and on ways of dealing with the cross-linguistic variability in their patterning.

1. *The Classification of Words*

Words can be classified by various criteria, such as phonological properties (e.g., monosyllabic vs. polysyllabic words), social factors (e.g., general vs. technical vocabulary), and language history (e.g., loanwords vs. native words). All of these are classes of words, but as a technical term, word class refers to the ten traditional categories below (plus perhaps a few others), most of which go back to the Greek and Roman grammarians. In addition to the terms, a few examples are given of each word class.

Noun	*book*, *storm*, *arrival*
Verb	*push*, *sit*, *know*
Adjective	*good*, *blue*, *Polish*
Adverb	*quickly*, *very*, *fortunately*
Pronoun	*you*, *this*, *nobody*
Preposition/adposition	*on*, *for*, *because of*
Conjunction	*and*, *if*, *while*
Numeral	*one*, *twice*, *third*
Article	*the*, *a*
Interjection	*ouch*, *tsk*

(In this article, the more general term 'adposition' will be used rather than preposition, because many languages have postpositions rather than prepositions, and word order is irrelevant in this context.)

The special status of the classification above derives from the fact that these are the most important classes of words for the purpose of grammatical description, equally relevant for morphology, syntax, and lexical semantics. This makes the classification more interesting, but also more complex and more problematic than other classifications of words. Besides the term word class, the older term part of speech (Latin *pars orationis*) is still often used, although it is now quite opaque (originally it referred to sentence constituents). The term word class was introduced in the first half of the twentieth century by structuralist linguistics. Another roughly equivalent term, common especially in Chomskyan linguistics is 'syntactic category' (although technically this refers not only to lexical categories such as nouns and verbs, but also to phrasal categories such as noun phrases and verb phrases).

The main two problems with the maximal word-class above are (a) that some of the classes intersect (e.g., the English word 'there' is both a pronoun and an adverb), and (b) that the different classes do not have equal weight; while most languages have hundreds of verbs and thousands of nouns, there are far fewer pronouns and conjunctions, and often only a handful of adpositions and articles. The solution that is often adopted explicitly for the second problem is to make a further subdivision into major word classes (nouns, verbs, adjectives, adverbs) and minor word classes (all others). (Alternative terms for major and minor classes are content words/function words and, especially in Chomskyan linguistics, lexical categories/functional categories.) This distinction is discussed further in Sect. 2. The solution to the first problem that is implicit in much contemporary work is that pronouns and numerals are not regarded as word classes on a par with nouns, verbs, prepositions, and so on. Instead, they are regarded as semantically highly specific subclasses of the other classes. For instance, there are nominal pronouns (e.g., *he, who*), adjectival pronouns (e.g., *this, which, such*) and adverbial pronouns (e.g., *here, thus*). Similarly, there are adjectival numerals (*five, fifth*), adverbial numerals (*twice*), and nominal numerals (*a fifth, a five*). Some languages also have verbal pronouns and verbal numerals. Accordingly, this article will not deal with pronouns (see *Pronouns*) and numerals (see *Numeral Systems*).

2. *Content Words and Function Words*

In all languages, words (and entire word classes) can be divided into the two broad classes of content words and function words. Nouns, verbs, adjectives, and adverbs are content words, and adpositions, conjunctions, and articles, as well as auxiliaries and words classified as 'particles' are function words. While there is sometimes disagreement over the assignment of words and even entire word classes to these two broad categories, their usefulness and importance is not in doubt. Content word classes are generally open (i.e., they accept new members in principle) and large (comprising hundreds or thousands of words), and content words tend to have a specific, concrete meaning. They tend to be fairly long (often disyllabic or longer), and their text frequency is fairly low. By contrast, function word classes are generally closed and small, and function words tend to have abstract, general meaning (or no meaning at all, but only a grammatical function in specific constructions). They tend to be quite short (rarely longer than a syllable), and their text frequency is high. This is summarized in Table 1.

The reason why auxiliaries are not included in the traditional list of word classes is probably merely that they are not prominent in Greek and Latin grammar, but in many languages these 'function verbs' are very important (English examples are *be, have, can, must, will, should*). The class 'particle' is really only a waste-basket category: function words that do not fit into any of the other classes are usually called particles (e.g., 'focus particles,' such as *only* and *also*, 'question particles,' such as Polish *czy* in *Czy mówisz po polsku*? 'Do you speak Polish?,' or 'discourse particles' such as German *ja* in *Das ist ja schön!* 'That's nice! (expressing surprise).'

The precise delimitation of function words and content words is often difficult. For instance, while the conjunctions *if, when, as,* and *because* are unequivocally function words, this is less clear for words like *suppose, provided that, granted that, assuming that*. And while the adpositions *in, on, of, at* are clearly function words, this is less clear for *concerning, considering, in view of*. In the case of adpositions, linguists sometimes say that there are two subclasses, 'function adpositions' and 'content adpositions,' analogous to the distinction between content verbs and function verbs (= auxiliaries). Another widespread view is that word-class boundaries are not always sharp, and that there can be intermediate cases between full verbs and auxiliaries, between nouns and adpositions, and between nouns/verbs and conjunctions. Quite generally, function words arise from content words by the diachronic process of grammaticalization (see *Grammaticalization*), and since grammaticalization is generally regarded as a gradual diachronic process, it is expected that the resulting function words form a gradient from full content words to clear function words. When grammaticalization proceeds further, function words may become clitics and finally affixes, and again we often find intermediate cases which cannot easily be classified as words or word-parts.

3. Defining Nouns, Verbs, and Adjectives

In the following, the emphasis will be on the content word classes nouns, verbs, and adjectives (for adverbs, a problematic class, see Section 5.3 below). The properties of the function words are more appropriately discussed in other contexts (e.g., auxiliaries in the context of tense and aspect, conjunctions in the context of subordinate clauses, and so on).

Before asking how nouns, verbs, and adjectives are defined, it must be made clear whether a definition of these word classes in a particular language (e.g., English or Japanese) is intended, or whether we want a definition of these classes for language in general. The widely known and much-maligned definitions of nouns as denoting 'things, persons, and places,' of verbs as denoting 'actions and processes,' and adjectives as denoting 'properties' is, of course, hopelessly simplistic from the point of view of a particular language. In most languages, it is easy to find nouns that do not denote persons, things, or places (e.g., *word, power, war*), and verbs that do not denote actions or processes (e.g., *know, lack, exist*), and many languages also have adjectives that do not denote properties (e.g., *urban, celestial, vehicular*). However, if the goal is to define nouns, verbs, and adjectives in general terms that are not restricted to a particular language, these simplistic notional definitions do not fare so badly.

In the first part of the twentieth century, the structuralist movement emphasized the need for rigorous language-particular definitions of grammatical notions, and notionally based definitions of word classes were rejected because they patently did not work for individual languages or were hard to apply rigorously. Instead, preference was given to morphological and syntactic criteria, e.g., 'if an English word has a plural in *–s*, it is a noun,' or 'if a word occurs in the context *the … book*, it is an adjective.' But of course this practice was not new, because words like *power* and *war* have always been treated as nouns on morphological and syntactic grounds. Some older grammarians, neglecting syntax, defined nouns, verbs, and adjectives exclusively in morphological terms, and as a result nouns and adjectives were often lumped together in a single class in languages like Latin and Greek, where they do not differ morphologically. But the predominant practice in Western grammar has been to give priority to the syntactic criterion. For instance, adjectives in German have a characteristic pattern of inflection that makes them quite unlike nouns, and this morphological pattern could be used to define the class (e.g., *roter/rote/rotes* 'red (masculine/feminine/neuter)'). However, a few property words are indeclinable and are always invariant (e.g., *rosa*, as in *die rosa Tapete* 'the pink wallpaper'). These words would not be adjectives according to a strictly morphological definition, but in fact everybody regards words like *rosa* as adjectives, because they can occur in the same syntactic environments as other adjectives.

Thus, there is universal agreement among linguists that language-particular word classes need to be defined on morphosyntactic grounds for each individual language. However, two problems remain. (a) The generality problem: how should word classes be defined for language in general? Morphological patterns and syntactic constructions vary widely across languages, so they cannot be used for cross-linguistically applicable definitions. (b) The subclass problem: which of the classes identified by language-particular criteria count as word classes, and which only count as subclasses? For instance, English has some property words that can occur in the context *is more … than*, e.g., *beautiful, difficult, interesting*. Another group of semantically similar words (e.g., *pretty, tough, nice*) does not occur in this context. Nobody takes this as evidence that English has two different word classes where other languages have just a single class (adjectives), but it is not clear why it does not count as sufficient evidence.

The solution to the generality problem that is usually adopted (often implicitly, but cf. Schachter 1985 and Wierzbicka 2000) is that one defines word classes on a language-particular basis, and then the word class that includes most words for things and persons is called 'noun,' the word class that includes most words for actions and processes is called 'verb,' and the word class that includes most words for properties is called 'adjective.' However, the subclass problem has not been solved or even addressed satisfactorily, and the use of word-class notions in a general or cross-linguistic sense remains problematic.

Table 1
Content words and function words

	Content words	Function words
word classes	nouns, verbs, adjectives, adverbs	adpositions, conjunctions, articles, auxiliaries, particles
class membership	open	closed
class size	large	small
meaning	concrete, specific	abstract, general (or none)
text frequency	low	high

4. *Characterizing Nouns, Verbs, and Adjectives*

Despite the theoretical problems in defining word classes in general, in practice it is often not difficult to agree on the use of these terms in a particular language. This is because nouns, verbs, and adjectives show great similarities in their behavior across languages. Their most common characteristics are briefly summarized in this section.

4.1 *Nouns*

In many languages, nouns have affixes indicating number (singular, plural, dual, see *Grammatical Number*), case (e.g., nominative, accusative, ergative, dative), possessor person/number ('my,' 'your,' 'his,' etc.), and definiteness. Some examples follow.

(a) Number. Khanty (Western Siberia) *xot* 'house,' *xot-yyn* 'two houses' (dual), *xot-yt* 'houses' (plural).

(b) Case. Classical Arabic *al-kitaab-u* 'the book' (nominative), *al-kitaab-i* 'the book's' (genitive), *al-kitaab-a* 'the book (accusative).'

(c) Possessor person/number. Somali *xoolah-ayga* 'my herd,' *xoolah-aaga* 'your herd,' *xoleh-eeda* 'her herd,' *xooli-hiisa* 'his herd,' etc.

Syntactically, nouns can always be combined with demonstratives (e.g., *that house*) and often with definiteness markers (*the house*), and they can occur in the syntactic function of argument (subject, object, etc.) without additional coding. Thus, in a simple two-argument clause we can have the *child*$_N$ *caused the accident*$_N$, but not **smoke*$_V$ *causes ill*$_A$. (Here and in the following, the subscripts *N*, *V* and *A* indicate nouns, verbs and adjectives.) Verbs like *smoke* and adjectives like *ill* need additional function-indicating coding to occur in argument function (*smok-**ing** causes ill-**ness***). Because reference is primarily achieved with nouns, it is nouns that can serve as antecedents for pronouns (compare *Albania's destruction of itself* vs. **the Albanian destruction of itself* (impossible)). Finally, nouns are often divided into a number of gender classes which are manifested in grammatical agreement (see *Grammatical Gender*).

4.2 *Verbs*

In many languages, verbs have affixes indicating tense (present, past, future), aspect (imperfective, perfective, progressive), mood (indicative, imperative, optative, subjunctive, etc.), polarity (affirmative, negative), valence-changing operations (passive causative, see *Valency and Argument Structure in Syntax*), and the person/number of subject and object(s) (see *Grammatical Agreement*). Semantic notions that are more rarely expressed morphologically are spatial orientation and instrument. Some examples follow.

(a) Tense. Panyjima (Australia) *wiya-lku* 'sees,' *wiya-larta* 'will see,' *wiya-rna* 'saw.'

(b) Subject person/number. Hungarian *lát-ok* 'I see,' *lát-sz* 'you see,' *lát* 's/he sees.'

(c) Valence-changing. Turkish *unut-* 'forget,' *unut-ul-* 'be forgotten' (passive), *unut-tur-* 'make forget' (causative).

(d) Spatial orientation. Russian *vy-letat*' 'fly out,' *v- letat*' 'fly in,' *pere-letat*' 'fly over ,' *vz-letat*' 'fly up.'

Syntactically, verbs generally take between one and three nominal arguments, e.g., *fall* (1: patient), *dance* (1: agent), *kill* (2: agent, patient), *see* (2: experiencer, stimulus), *give* (3: agent, patient, recipient). Nouns and adjectives may also take arguments, but they are not nearly as rich as verbs, and nouns that correspond to verbs often cannot take arguments in the most direct way (compare *Plato defined beauty* vs. **Plato definition beauty* (impossible); additional coding is required: *Plato's definition of beauty*. Verbs always occur as predicates without additional coding, whereas nouns and adjectives often need additional function-indicating coding when they occur as predicates, namely a copular verb (cf. *Halim works*$_V$ vs. **Halim a worker*$_N$ (impossible), **Halim hard-working*$_A$ (impossible); here the copula *is* is required).

4.3 *Adjectives*

In a fair number of languages, adjectives have affixes indicating comparison (comparative degree, superlative degree, equative degree), and in a few languages, adjectives are inflected for agreement with the noun they modify. Some examples follow.

(a) Comparison. Latin *audax* 'brave,' *audac-ior* 'braver' (comparative), *audac-issimus* 'bravest' (superlative).

(b) Comparison. Tagalog (Philippines) *mahal* expensive;' *sing-mahal* 'as expensive as.'

(c) Agreement. Hindi *acchaa* 'good' (masculine singular), *acchee* (masculine plural), *acchii* (feminine singular/plural).

In many languages, adjectives show no inflectional properties of their own. Syntactically, a peculiarity of adjectives is that they can typically occur in comparative constructions (whether they show affixes marking comparison or not), and they can be combined with degree modifiers of various kinds that do not co-occur with verbs and nouns (e.g., *very hot*$_A$, *too difficult*$_A$, cf. **work*$_V$ *very*, **too mistake*$_N$ (impossible)). Adjectives generally occur as nominal modifiers without additional coding (cf. *a bald*$_A$ *man*), whereas nouns and verbs mostly need additional function-indicating coding when they occur as modifiers (**a beard*$_N$ *man*/*a man with a beard*, **a shave*$_N$ *man*/*a man who shaves*).

5. *Difficulties of Classification*

The general properties of nouns, verbs, and adjectives that were sketched in Sect. 5 are sufficient to establish these classes without much doubt in a great many languages. However, again and again linguists report on languages where such a threefold subdivision does not seem appropriate. Particularly problematic are adjectives (Sect. 5.1) but languages lacking a noun–verb distinction are also claimed to exist (Sect. 5.2), and Sect. 5.3 discusses adverbs, which present difficulties in all languages.

5.1 *The Universality of Adjectives*

In contrast to nouns and verbs, adjectives are sometimes like function words in that they form a rather small, closed class. For instance, Tamil (southern India) and Hausa (northern Nigeria) have only about a dozen adjectives. Interestingly, in such languages the concepts that are denoted by adjectives in the small class coincide to a large extent (Dixon 1977): dimension ('large,' 'small,' 'long,' 'short,' etc.), age ('new,' 'young,' 'old,' etc.), value ('good,' 'bad'), color ('black,' 'white,' 'red,' etc.). Other concepts for which English has adjectives (e.g., human propensity concepts such as 'happy,' 'clever,' 'proud,' 'jealous,' and physical property concepts such as 'soft,' 'heavy,' 'hot') are then expressed by verbs or by nouns. For instance, in Tamil, 'heavy man' is *ganam-uḷḷa manuṣan*, literally 'weight-having man,' and in Hausa, 'intelligent person' is *mutum mai hankali*, literally 'person having intelligence.'

But even more strikingly, many languages appear to lack adjectives entirely, expressing all property concepts by words that look like verbs or like nouns. For instance, in Korean, property concepts inflect for tense and mood like verbs in predication structures, and they require a relative suffix (see *Relative Clauses*) when they modify a noun, again like verbs (cf. (b) (i), (ii)) below.

(a) Predication

(i) Event

salam-i *mek-ess-ta*
person-NOMINATIVE eat-PAST-DECLARATIVE
'the person ate'

(ii) Property

san-i *noph-ess-ta*
hill-NOMINATIVE high-PAST-DECLARATIVE
'the hill was high'

(b) Modification

(i) Event

mek-un *salam*
eat-RELATIVE person
'a person who ate'

(ii) Property

noph-un *san*
high-RELATIVE hill
'a high hill'

While languages where all property words can be classified as verbs are very common, languages where all property concepts are nouns are less widely attested. A language for which such a claim has been made is Ecuadorian Quechua: in this language, property concept words can occur in argument position and take the same inflection as nouns (cf. (a)(i), (ii) below), and nouns can occur as modifiers without additional coding, like property words (cf. (b) (i), (ii) below).

(a) argument position

(i) Thing

wambra-ta-mi
child-ACCUSATIVE-FOCUS
wajta-rka
hit-PAST.3RD.SINGULAR
'he hit the child'

(ii) Property

jatun-ta-mi

big-ACCUSATIVE-FOCUS

wajta-rka

hit-PAST.3RD.SINGULAR

'he hit the big one'

(b) modification

(i) Thing

rumi wasi

stonehouse

'stone house'

(ii) Property

jatun wambra

big child

'big child'

Thus, it is often said that while nouns and verbs are virtually universal, adjectives are often lacking in languages. However, it is generally possible to find features that differentiate a property subclass within the larger class to which property words are assigned. For instance, Korean property verbs do not take the present-tense suffix *-nun*, and Ecuadorian Quechua thing words do not combine with the manner adverb suffix *-ta* (e.g., *sumaj-ta* 'beautifully,' but not **duktur-ta* 'in a doctor's manner'). Here the subclass problem arises: on what grounds do we say that Korean has two classes of verbs (non-property verbs vs. property verbs), rather than two word classes (verbs and adjectives)? Since this question is difficult to answer, some linguists have claimed that most languages have adjectives after all, but that adjectives have a strong tendency to be either verb-like or noun-like (e.g., Wetzer 1996).

5.2 *The Universality of the Noun–Verb Distinction*

For a few languages, it has been claimed that there is no (or only a very slight) distinction between nouns and verbs, for instance for several North American languages of the Wakashan, Salishan, and Iroquoian families, as well as for a number of Polynesian languages. For instance, in Samoan (a Polynesian language), full words referring to events and things show intriguingly similar behavior. Both thing (or person) words and event words seem to occur in the same predication structures (a) and in argument positions (b) below.

(a) Predication

(i) Thing

sa fōma'i le fafine

PAST doctor the woman

'the woman was a doctor'

(ii) Event

sa alu le fafine

PAST go the woman

'the woman went'

(a) Argument

(i) Thing

e lelei le fōma'i

GENERIC good the doctor

'the doctor is good'

(ii) Event

e lelei le alu o le vasi

GENERIC good the go of the bus

i Apia

to Apia

'it's good that the bus goes to Apia'

Clearly, the similarity of thing-words and event-words in such languages is quite striking and differs dramatically from what is found in the better-known European languages. But thing-words and event-words do not behave exactly alike in Samoan; the pattern above is asymmetrical in that *fōma'i* means 'be a doctor' and 'person who is a doctor,' but *alu* does not mean both 'go' and 'person who goes,' but rather 'the fact of going.' Upon closer examination, it has usually turned out that major word classes which can be called nouns and verbs can be distinguished even in the problematic languages.

5.3 *The Problem of Adverbs*

Adverbs are the most problematic major word class because they are extremely heterogeneous in all languages, and unlike for nouns, verbs, and adjectives, no semantic prototype can be identified easily for them (cf. Ramat and Ricca 1994). The most that can be said in general about adverbs is that they serve to modify non-nominal constituents (verbs or verb phrases, adjectives, other adverbs, sentences). Perhaps the

concept of adverb should not be taken too seriously, because there are very few properties that adverbs of different kinds share. Five broad subclasses of adverbs are often distinguished: setting adverbs (locative: *here, there, below, abroad*; temporal: *now, then, yesterday, always*), manner adverbs (*quickly, carefully, beautifully*), degree adverbs (*very, too, extremely*), linking adverbs (*therefore, however, consequently*), and sentence adverbs (*perhaps, fortunately, frankly*) (see Quirk et al. 1985 for the most comprehensive semantic classification of adverbs).

Setting adverbs, degree adverbs, and linking adverbs are relatively small, closed classes, and they often share properties with function words. Sentence adverbs are rare in most languages, and their great elaboration is probably a peculiarity of the written languages of Europe (Ramat and Ricca 1998). The only sizable subclass of adverbs that has equivalents in many languages is the class of manner adverbs. Many languages have a productive way of forming manner adverbs from adjectives (e.g., English *warm/warmly*, French *lent* 'slow,' *lentement* 'slowly'). But this also makes manner adverbs problematic as a major word class, because one could argue that adjective-derived manner adverbs are just adjectives which occur with a special inflectional marker to indicate that they are not used in their canonical noun-modifying function. This point of view is non-traditional, but it seems quite reasonable, and it is strengthened by the fact that in quite a few languages, adjectives can be used as manner adverbs without any special marking.

One of the main features that unifies the various subclasses of adverbs in languages like English and French is that four of the five classes contain adjective-derived words ending in *-ly/-ment* (only setting adverbs are almost never of this type). This is certainly no accident, but it should also be noted that this is probably a feature typical of European languages that is hardly found elsewhere.

6. *Theoretical Approaches*

While the identification and definition of word classes was regarded as an important task of descriptive and theoretical linguistics by classical structuralists (e.g., Bloomfield 1933), Chomskyan generative grammar simply assumed (contrary to fact) that the word classes of English (in particular the major or 'lexical' categories noun, verb, adjective, and adposition) can be carried over to other languages. Without much argument, it has generally been held that they belong to the presumably innate substantive universals of language, and not much was said about them (other than that they can be decomposed into the two binary features [±N] and [±V]: [+N, −V] = noun, [−N, +V] = verb, [+N, +V] = adjective, [−N, −N] = adposition) (see *Linguistics: Theory of Principles and Parameters*).

Toward the end of the twentieth century, linguists (especially functionalists) became interested in word classes again. Wierzbicka (1986) proposed a more sophisticated semantic characterization of the difference between nouns and adjectives (nouns categorize referents as belonging to a kind, adjectives describe them by naming a property), and Langacker (1987) proposed semantic definitions of noun ('a region in some domain') and verb ('a sequentially scanned process') in his framework of Cognitive Grammar. Hopper and Thompson (1984) proposed that the grammatical properties of word classes emerge from their discourse functions: 'discourse-manipulable participants' are coded as nouns, and 'reported events' are coded as verbs.

There is also a lot of interest in the cross-linguistic regularities of word classes, cf. Dixon (1977), Bhat (1994) and Wetzer (1996) for adjectives, Walter (1981) and Sasse (1993a) for the noun–verb distinction, Hengeveld (1992b) and Stassen (1997) for non-verbal predication. Hengeveld (1992a) proposed that major word classes can either be lacking in a language (then it is called rigid) or a language may not differentiate between two word classes (then it is called flexible). Thus, 'languages without adjectives' (cf. Sect. 6) are either flexible in that they combine nouns and adjectives in one class (N/Adj), or rigid in that they lack adjectives completely. Hengeveld claims that besides the English type, where all four classes (V−N−Adj−Adv) are differentiated and exist, there are only three types of rigid languages (V−N−Adj, e.g., Wambon; V−N, e.g., Hausa; and V, e.g., Tuscarora), and three types of flexible languages (V−N−Adj/Adv, e.g., German; V−N/Adj/Adv, e.g., Quechua; V/N/Adj/Adv, e.g., Samoan).

The most comprehensive theory of word classes and their properties is presented in Croft (1991). Croft notes that in all the cross-linguistic diversity, one can find universals in the form of markedness patterns; universally, object words are unmarked when functioning as referring arguments, property words are unmarked when functioning as nominal modifiers, and action words are unmarked when functioning as predicates. While it is not possible to define cross-linguistically applicable notions of noun, adjective, and verb on the basis of semantic and/or formal criteria alone, it is possible, according to Croft, to define nouns, adjectives, and verbs as cross-linguistic prototypes on the basis of the universal markedness patterns.

For a sample of recent work on word classes in a cross-linguistic perspective, see Vogel and Comrie (2000), and the bibliography in Plank (1997). Other overviews are Sasse (1993b), Schachter (1985), and further collections of articles are Tersis-Surugue (1984) and Alpatov (1990).

See also: Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Speech Production,

Neural Basis of; Speech Production, Psychology of; Word, Linguistics of; Word Meaning: Psychological Aspects; Word Recognition, Cognitive Psychology of

Bibliography

Alpatov V M (ed.) 1990 *Časti reči: Teorija i tipologija*. Nauka, Moscow

Bhat D N S 1994 *The Adjectival Category: Criteria for Differentiation and Identification*. Benjamins, Amsterdam

Bloomfield L 1933 *Language*. Holt, New York

Croft W 1991 *Syntactic Categories and Grammatical Relations: the Cognitive Organization of Information*. The University of Chicago Press, Chicago

Dixon R M W 1977 Where have all the adjectives gone? *Studies in Language* **1**: 19–80

Hengeveld K 1992a Parts of speech. In: Fortescue M, Harder P, Kristoffersen L (eds.) *Layered Structure and Reference in a Functional Perspective*. Benjamins, Amsterdam, pp. 29–56

Hengeveld K 1992b *Non-verbal Predication: Theory, Typology, Diachrony*. de Gruyter, Berlin

Hopper P J, Thompson S A 1984 The discourse basis for lexical categories in universal grammar. *Language* **60**: 703–52

Langacker R W 1987 Nouns and verbs. *Language* **63**: 53–94

Plank F 1997 Word classes in typology: recommended readings (a bibliography). *Linguistic Typology* **1**: 185–92

Quirk R, Greenbaum S, Leech G, Svartvik J 1985 *A Comprehensive Grammar of the English Language*. Longman, London

Ramat P, Ricca D 1994 Prototypical adverbs: On the scalarity/radiality of the notion of adverb. *Rivista di Linguistica* **6**: 289–326

Ramat P, Ricca D 1998 Sentence adverbs in the languages of Europe. In: van der Auwera J (ed.) *Adverbial Constructions in the Languages of Europe*. de Gruyter, Berlin, pp. 187–275

Sasse H-J 1993a Das Nomen—eine universale Kategorie? *Sprachtypologie und Universalienforschung* **46**: 187–221

Sasse H-J 1993b Syntactic categories and subcategories. In: Jacobs J et al. (eds.) *Syntax: An International Handbook of Contemporary Research*. de Gruyter, Berlin, Vol. 1, pp. 646–86

Schachter P 1985 Parts-of-speech systems. In: Shopen T (ed.) *Language Typology and Syntactic Description*. Cambridge University Press, Cambridge, UK, Vol. 1, pp. 3–61

Stassen L 1997 *Intransitive Predication*. Oxford University Press, Oxford

Tersis-Surugue N (ed.) 1984 *L'opposition verbo-nominale dans diverses langues du monde*. (Special issue of *Modèles linguistiques* VI.1.), Presses Universitaires de Lille, Lille, France

Vogel P M, Comrie B (eds.) 2000 *Approaches to the Typology of Word Classes*. (*Empirical Approaches to Language Typology, Vol. 23*.) de Gruyter, Berlin

Walter H 1981 *Studien zur Nomen-Verb-Distinktion aus typologischer Sicht*. Fink, München

Wetzer H 1996 *The Typology of Adjectival Predication*. de Gruyter, Berlin

Wierzbicka A 1986. What's in a noun? (Or: How do nouns differ in meaning from adjectives?) *Studies in Language* **10**: 353–89

Wierzbicka A 2000 Lexical prototypes as a universal basis for cross-linguistic identification of 'parts of speech'. In: Vogel P M, Comrie B (eds.) *Approaches to the Typology of Word Classes*. de Gruyter, Berlin, pp. 285–317

M. Haspelmath

Word, Linguistics of

In thinking about words, the first question is how to divide an utterance up into them. The simple answer is that words are demarcated by spaces, just as they are on this page. But this simple answer depends on the existence of writing. In speech, we do not normally leave a space or pause between words. Most languages throughout history have not been written down. Surely we do not want to say that only written languages have words, and, even with written languages, spacing does not provide an entirely satisfying answer. For example, English compounds can be spelled in three ways: open, closed, or hyphenated, and some items can be spelled in any of these three ways without being affected in any detectable way: *birdhouse*, *bird-house*, *bird house*. We do not want to say that the first spelling is one word, the last spelling two words, and the middle one neither one word nor two, which we would have to do if we accepted spaces as criterial. The better conclusion is that spelling conventions are not a completely reliable clue to whether something is a word or not.

Some linguists avoid the problem by claiming that the whole notion 'word' is theoretically invalid, just an artifact of spelling. In their favor is the fact, which people are always surprised to learn, that not all languages have a word for 'word.' The classical languages, biblical Hebrew, classical Greek, and classical Latin, for example, all have terms that are systematically ambiguous among 'speech,' 'word,' and 'utterance' but none has terms that distinguish clearly among these notions and certainly none has a special term that means just 'word.' Even in the opening verse of the Gospel of John, 'In the beginning was the Word,' it is still not clear just what is the meaning of the Greek word *λόγος* [*logos*] that we conventionally translate as 'word.' Many scholars believe that the best translation is 'thought' or 'reason.'

The classical languages are not alone. The anthropologist Bronislaw Malinowski declared that the distinction between word and utterance was not an obvious one to most peoples, that 'isolated words are in fact only linguistic figments, the products of an advanced linguistic analysis' (Malinowski 1935, p. 11), suggesting that there was no reason for most languages to distinguish between words and utterances. Nonetheless, most modern linguists believe that all languages do have words, whether their speakers are aware of the units or not. The question then becomes how to figure out what a word is and how to identify words in a way that is valid for all languages, written or spoken, and, since language is first and foremost spoken, our answer must not depend on writing.

The earliest explicit discussion that we have of the notion 'word' and of words in the speech stream is in the work of Aristotle. Aristotle made a distinction between an utterance or sentence, for which he used the term *λόγος* [*logos*] (confusingly, the term that we

usually think of as meaning 'word'), and a word, which he called: *μέρος λόγου* [meros logou], literally 'piece of an utterance.' Aristotle's term is still with us, since it was translated literally into the Latin term *pars orationis*, from which in turn we derive the English term 'part of speech' by a similar literal translation. He defined this new entity as a component of a sentence, having a meaning of its own and not further divisible into meaningful units (*De interpretatione* 2–3). In other words, Aristotle saw the speech stream as being divided up into atomic pieces, each being what we now call a word. The problem of defining words and identifying them was thus one and the same for him. Aristotle, incidentally, identified only two types of words, or parts of speech, which correspond roughly to our own nouns and verbs.

It is difficult for us to appreciate the importance of Aristotle's realization that the speech stream could be divided into pieces, because we are aware of words. We have been taught to think of language in terms of words and if we think about speech, it is not as a stream, but as words lined up one after the other, like pearls separated by knots on a string. But in a culture without writing or in a culture such as that of ancient Greece, in which writing played an extremely small part in the life of even the most educated people, this conception is not at all obvious. After all, the philosophical school of Aristotle was called peripatetic because his method of teaching consisted of discussion, conducted while walking about in the Lyceum of Athens. This intellectual style was purely oral, not written, and for both Plato and Aristotle, language was spoken, not written. Indeed, Aristotle's teacher Plato had a deep distrust of writing, which he called inhuman. This role of spoken language in classical Greece must be underscored because spoken language, unlike modern written English, is not a string of pearls. We do not speak in single words with spaces between, and so the discovery that speech could be broken down into smaller component words was indeed remarkable.

Was it an accident that Aristotle made his discovery about words in one of the first societies with writing? Many people would say that written language, even an alphabet, is a prerequisite for linguistic awareness: we cannot begin to analyze the stream of language until we see it recorded in writing. This claim faces a number of obstacles, most prominently the fact that the most advanced system of linguistic analysis known prior to the advent of modern linguistics in the nineteenth century was that of the Sanskrit grammarians, the best known of whom was Panini. Panini flourished around 500 BCE. His grammar was embedded within an oral tradition of grammatical analysis of the Sanskrit sacred texts that must have begun long before him. Strikingly, we have no evidence at all of writing in Sanskrit until at least several hundred years after Panini, leading to the conclusion that the Paninian tradition of linguistic analysis, which recognized not only words, but also the internal parts of words, and used the notion of 'zero' long before the Hindu mathematicians who introduced it to the rest of the world, existed before the advent of writing in India.

So writing is not a prerequisite to recognizing that the stream of language can be broken down into words. Even so, having language written down as an object must have assisted in the recognition of the words. The evidence from early writing is tantalizing, but mixed. On the one hand, the earliest writing systems, notably Sumerian cuneiform and Egyptian hieroglyphics, had no consistent means of dividing the text into words. Similarly, ancient Semitic and Greek texts were usually written continuously, with no spaces between words, and the practice extended into medieval times for Latin. Even among modern languages, Chinese, to this day, is always written without any spaces between words. On the other hand, later cuneiform scripts did mark the boundaries of words and the Etruscans, whose alphabet was based on the Greek alphabet, marked these from quite early on by means of a centered dot, a practice continued by the Romans: when one looks at the inscriptions on such monuments of Imperial Rome as Trajan's column, inscribed in the same Roman forms that we use most commonly today on our word processors and which are consequently so strikingly readable, one sees these same dots separating the words. In the Semitic scripts that descend from the first alphabet there are special forms of certain letters that appear only at the ends of words. The Greek letter sigma has two forms, one of which is used only at the ends of words. Even when there were no spaces between words, these languages could not be written without knowing where a word ended. Finally, the Aramaic scripts (notably Syriac, in which many important documents of the early Christian church were written, and which is still important for Christian scholars today) were the first to develop a cursive form, where all the letters of a word are joined together into a single unit, again impossible without knowing where a word begins and ends.

So there is evidence from early in the history of writing that, although a writing system can record language without breaking it down into words, it is equally possible to represent the breaks between words. What permits people to do this reliably? How do we know where one word ends and the next one begins in spoken language?

For some languages, the answer lies in sound, specifically word stress or word accent. In the simplest case, every word in the language is stressed on the same syllable, so that one can quite easily recognize the words directly from the stresses. Hungarian and Czech, for example, have inexorable word-initial stress, while Persian has word-final stress and Polish has stress on the second-to-last syllable. In all of these cases, even if a person does not speak the language, he or she can usually tell directly from the stress where the

word breaks are. Even a computer, properly programmed, would be able to break up the speech stream of these languages into words without knowing what any of these words mean.

Of course, not all languages are quite so simple. In English and Russian, for example, although every word has a main stress, so one can determine the number of words in an utterance from the number of main stresses, the stress patterns of these words are not easily predictable; some words are stressed on the first syllable (e.g., '*sym*pathize'), while others are stressed on the last (e.g., 'kanga*roo*') and others on the second-to-last ('re*mem*ber') or even the third-to-last ('a*no*maly'). A fluent speaker just knows for every word where the stress lies and speakers seldom make mistakes. No one says 'sympath*ize*,' or 'kang*a*roo,' or 'ano*ma*ly.' So an English speaker has no trouble telling where the word breaks are, but the task is quite overwhelming for a learner or for a simple computer program. In the most problematic type of languages for a stress-based method of word division, as with French or many of the languages of India, there is no word stress, so that phonology cannot be used to detect word division in a simple way.

But French still has words, so there must be some nonphonological way to isolate words. One clever test that linguists use for word divisions involves interruption. Take a sentence like *Theformerleaderremainedinsidethe WhiteHouse*, which is here written without spaces in order to emphasize the difficult nature of the task. Where can this sentence be interrupted? We may understand this question in two ways. One involves pauses: where is it possible to pause naturally? The other involves insertion: where can we insert elements in the string without affecting its structure. Curiously, the two give us slightly different answers. The natural-pause criterion identifies words as follows: *The former leader remained inside the White House*. Insertion gives us the following breaks (with inserted elements in parentheses): *The* (*recalcitrant*) *former* (*American*) *leader* (*stubbornly*) *remained* (*ensconced*) *inside* (*precisely*) *the* (*stately*) *White House*. The resulting sentence may not be elegant, but it is English. We do not normally pause after the definite article 'the' in English, because the article is what linguists call a clitic, an element that cannot be pronounced as a word all by itself, though other criteria such as insertion may identify it as a word. We cannot either pause or insert an element between the two elements of 'White House' because, despite the space in the standard orthography, 'White House' is a single compound word. We may then use the twin criteria of possible natural pause and insertion in order to identify the breaks between words in a spoken utterance. These criteria, though seemingly simple, depend on the *possibility* of pause and insertion. A learner thus has to compare numerous utterances before being able to break an utterance up reliably into words.

Where to pause or insert may seem obvious (though again our familiarity with written language may give us false confidence), but there are places in our sentence where, although linguistic analysis tells us there are breaks between elements, these breaks are below the word level, and where, though a machine might be tempted, both pause and insertion are impossible for speakers of the language. For example, 'leader,' 'remained,' and 'inside' can easily be divided into 'lead-er,' 'remain-ed,' and 'in-side,' yet no English speaker would insert an element or pause at these divisions. Here we are dealing with the meaningful parts of words, which linguists call morphemes, the divisions between which are not easily recognized by speakers.

Aristotle identified two types of words, nouns and verbs. School grammars permit more: adjectives, adverbs, pronouns, prepositions, and conjunctions. Linguists call these different types lexical categories and distinguish between open and closed categories. Open categories are those to which new members may be added easily, either by word formation or by borrowing from other languages. In English, the open categories are noun, verb, and adjective. We add new nouns at a tremendous rate, verbs and adjectives more slowly. The other categories are closed: for example, no new pronoun has been added to English since 'them' was borrowed from Old Norse over a millennium ago.

The lexical categories differ from language to language. The class of adjectives is much smaller in many languages and in some, like the Dravidian languages spoken widely in South India, it has only a dozen or so members. Even the class of verbs may be very small in some languages (e.g., Bengali). English also has more prepositions than most languages; some Austronesian languages of Indonesia have only one. Languages may even lack certain lexical categories. English is unusual in having a category of adverbs separate from adjectives and many linguists believe that the distinction between the two is not even valid for English. It has also been claimed that some languages do not distinguish nouns from verbs.

How to define the categories is another question. The traditional definitions are based on meaning: a noun denotes a person, place, or thing; a verb denotes an action or state; an adjective denotes an attribute. But these definitions are problematic. Nouns like 'love' or 'confusion' denote states, not things, as do adjectives like 'solid' and 'angry,' while relational nouns or verbs like 'mother,' 'son,' and 'include' fall under none of the standard meaning-based definitions. Modern linguists prefer instead to define the categories grammatically, in terms of how they function in a sentence. A noun is thus the head or main word of a subject or object phrase, an adjective/adverb a modifier, and a verb the main word of a predicate phrase.

Over the last century, the major research preoccupation of linguists concerned with words has been with

analyzing their internal structure or morphology. Almost all languages have internally complex words, though a few, notably Vietnamese, have very simple word structure, limited largely to compounds, words formed by combining words, like English 'shoelace' or 'boxcar.' In most languages, the majority of complex words are formed with suffixes like the '-er' and '-ed' in 'leader' and 'remained' above. Somewhat less frequent across language are prefixes like 'un-' in 'undo,' 'unpack,' and 'unroll.' In English and many other languages, prefixes and suffixes may combine, resulting in words like 're-institut-ion-al-iz-ation' or the infamous 'anti-dis-establish-ment-ari-an-ism.' Less common means of forming complex words may involve internal sound changes like 'mouse/mice' or 'goose/geese' or even changes in stress like that between the verb 're*ject*' and the noun '*re*ject.' Another mechanism is to repeat part of a word: in ancient Greek, the past tense of a verb is produced by repeated the first consonant; the stem '*graph*' (meaning 'write') becomes '*gegraph*' in the past tense, and similarly for other verbs.

Linguists distinguish two grammatical functions of complex words, inflection and derivation. The word inflection descends from the Latin verb meaning 'bend.' The idea is that a word bends or adjusts its shape to fit its context in a sentence. In English, the vast majority of verbs inflect, or adjust their shape, for person and tense by suffixation in a simple fashion: the past tense/past participle form of most verbs is formed by adding the suffix '-ed' as in 'remain-ed,' the third person singular present adds '-s' ('remains') and the present participle adds '-ing' ('remaining'). Elsewhere, the verb is morphologically simple ('remain'). We may say that the four verb forms ('remained,' 'remains,' 'remaining,' and the uninflected 'remain') constitute a set, with the grammar deciding the context in which each member of the set will appear. We call this set of forms a lexeme and we use the uninflected form (here 'remain') to name the set. For most English nouns, there are two inflected forms in the set of each lexeme, the singular and the plural.

English inflection is very simple, but other languages are much more complex. In many of the Indo-European relatives of English, nouns and adjectives inflect differently depending on their function in a sentence, the way English pronouns do ('I, me, my, mine'). A Russian noun or adjective will usually have ten inflectional forms. Verbs can be much more complex. Some languages have noun classes or genders, a dozen or more in the Bantu languages that dominate Africa, and in these languages many verbs have a distinct inflectional form for every gender. Verbs also often inflect for tense and for the person and number of the subject, with the result that verb inflection may become baroque, with over two hundred forms for every verb in ancient Greek, a thousand or more in Navajo, and a theoretically infinite number for each verb in Turkish. Experimental studies have shown that the inflected forms of a single lexeme are very closely tied to one another in the mind.

The other mechanism for forming complex words is derivation, which results in new lexemes, rather than contextually limited forms of the same lexeme. Thus, 'bakes,' 'baked,' and 'baking' are all inflected forms of a single verb lexeme, 'bake,' but 'baker' and 'bakery' are different words, nouns, distinct lexemes from 'bake,' though formed by the same general method (suffixation) as the inflected forms. And each of these nouns in turn may be inflected: 'bakers, bakeries.' Two different lexemes like *bake* and *baker* will normally not be as tightly connected in a person's mind as are the forms of a single lexeme.

This brings us to the general question of how words are stored in the human mind, how we retrieve them when we speak or write, and how we recognize them when we understand language. This is the frontier of linguistic research on words and it involves cooperation among linguists, psychologists, and neuroscientists. The results of this line of research are very exciting, though not well enough established to allow for firm conclusions. What we can say is that the question of what is a word will continue to be fundamental, even as the methods of research on language become more and more sophisticated.

See also: Lexical Semantics; Lexicology and Lexicography; Lexicon; Morphological Case in Linguistics; Morphology in Linguistics; Word Classes and Parts of Speech

Bibliography

Anderson S R 1992 *A-morphous Morphology*. Cambridge University Press, Cambridge, UK

Aronoff M 1976 *Word Formation in Generative Grammar*. MIT Press, Cambridge, MA

Aronoff M 1994 *Morphology by Itself*. MIT Press, Cambridge, MA

Bolinger D 1963 The uniqueness of the word. *Lingua* **12**: 113–36

Bybee J L 1985 *Morphology: A Study of the Relation Between Meaning and Form*. Benjamins, Amsterdam

Cardona G 1994 Indian linguistics. In: Lepschy G (ed.) *History of Linguistics*. Longman, London, Vol. 1

Daniels P T, Bright W 1996 *The World's Writing Systems*. Oxford University Press, Oxford, UK

Harris Z 1946 From morpheme to utterance. *Language* **22**: 161–83

Lyons J 1977 *Semantics*. Cambridge University Press, Cambridge, UK

Malinowski B 1935 *Coral Gardens and Their Magic; A Study of the Methods of Tilling the Soil and of Agricultural Rites in the Trobriand Islands*. American Book Company, New York

Matthews P H 1991 *Morphology*. Cambridge University Press, Cambridge, UK

Miller G 1991 *The Science of Words*. Scientific American Library, New York

Robins R H 1979 *A Short History of Linguistics*. Longman, London
Spencer A 1991 *Morphological Theory*. Blackwell, Oxford, UK
Spencer A, Zwicky A 1998 *Handbook of Morphological Theory*. Blackwell, Oxford, UK

M. Aronoff

Word Meaning: Psychological Aspects

Word meaning can be studied from multiple perspectives including linguistic, philosophical, and psychological. This article focuses on the issues of meaning of most interest to psychologists: how to characterize the knowledge that constitutes people's mental representation of meanings, and how to describe the use of this knowledge in interpreting words that are heard or read.

1. Approaches to Word Meaning

When people think casually about the topic of *word meaning*, they typically think of dictionaries. They view dictionary entries as constituting a statement of a word's meaning. They also think of the dictionary as the ultimate authority on the correct meaning: in any dispute, they defer to the dictionary definition as determining what is right. In doing so, they adopt implicit assumptions that word meanings belong to a language, as opposed to its users, and that there are experts who alone are qualified to report the true meaning of each word.

Elements of this sort of view appear in some linguistic and philosophical approaches to word meaning. For instance, language is sometimes treated as an abstract system to be analyzed, without reference to the users of the system. And Putnam's (1975, Chap. 12) influential philosophical approach to meaning appeals to a 'division of linguistic labor,' suggesting that correct use of some common words can be determined only by experts who have special knowledge of the entities in question.

However, modern *lexicographers*—those who write dictionaries—do not hold such a view themselves. Their job in providing dictionary definitions is not to report the 'true' meaning of a word that is divorced from what ordinary speakers of the language use the word to mean. The meanings associated with words change over time, as their use by speakers of the language shifts. For instance, 'nice' currently has a highly positive connotation but in earlier times referred to someone who was ignorant or foolish; 'hussy' meant a housewife, and 'mistress' was simply a feminized form of 'mister' (Hock and Joseph 1996, Chap. 7). Just a generation ago, 'mouse' referred only to a small rodent; now it is used more often to refer to a computer peripheral. The goal of lexicographers must therefore be to capture something about how a community of speakers uses a word at a particular time in the history of the language (see, e.g., Laird 1978).

Psychologists interested in word meaning take a perspective more similar to that of lexicographers than to either the layperson's intuitive view or the view reflected in the kinds of linguistic and philosophical approaches mentioned. Psychologists view meanings as changeable, as belonging to individual language users, and as shared across communities of language users. But psychologists, unlike lexicographers, are particularly interested in mental representation and in the cognitive processes that draw on mental representation. Their primary goals in the study of meaning are to characterize the knowledge that constitutes word meanings and to describe how this knowledge is used in interpreting words heard or read.

2. Meaning as Concepts

A standard way of describing what constitutes a word meaning, for a psychologist, is to say that meanings consist of *concepts*. Concepts are coherent nonlinguistic knowledge about some entity or set of entities (e.g., a person could have a concept of his or her brother Max and also a concept of what brothers are in general). A strict one-to-one correspondence does not exist between the words known by a person and his or her concepts, since a person can have a concept without having a word to label it. (Children in the early stages of language learning presumably are often in this situation.) Also, a person occasionally may know a word but have little or no meaning associated with that word. In most cases, though, the words stored in a person's memory will have meanings associated with them, and the argument is that the stored meaning for each word consists of a concept. Thus, for any given person, the meaning of 'brother' consists of his or her concept of what a brother is, the meaning of 'run' consists of his or her concept of what running is, and so on. For many common words, the associated concept is assumed to be similar for most people, and so the meaning is shared.

The idea that word meanings are concepts, though, does not delve very far into the question of the nature of meaning. If meanings are concepts, and psychologists want to understand the nature of meanings, they need to know in more detail what concepts consist of. Psychologists have devoted much effort to the study of concepts, both for their own sake and in order to understand word meaning (e.g., see *Natural Concepts, Psychology of*).

An early answer to the question of what concepts are like was suggested by linguists pursuing their own interests in word meaning. The proposal from linguistics was that the meaning of a word is a set of *defining features*; that is, features true of all things appropriately called by that name and together separating

those things from all things called by other names. For instance, defining features for the word 'bachelor' might be 'adult,' 'male,' and 'unmarried.' This proposal was closely related to analyses by anthropologists suggesting that kin terms such as 'mother' and 'grandmother' could be analyzed in terms of elements of meaning such as such as sex (male vs. female) and generation (one removed vs. two removed). Psychologists initially adopted the defining features view as a view of the nature of concepts, not of word meaning. Under the perspective that word meanings consist of concepts, though, the circle back to the original linguistic theorizing is closed, with the difference that psychologists considered defining features to constitute the mental representation of the meaning, not a meaning detached from individual minds.

By the mid 1970s, psychologists were finding this analysis of concepts and meanings problematic because for many words, it was impossible to determine defining features. Rosch and Mervis (1975) suggested that concepts consist, instead, of sets of features having varying strengths of association to the category. For instance, most fruits are juicy but a few are not; many fruits are sweet, but some are not; some have a single pit while others have many seeds. The most common features, like *sweet* and *juicy* for fruit, are true of prototypical examples but are not defining features. If word meanings are represented by concepts, then according to this analysis, they are represented by concepts that include *prototypes* and probabilistically associated features, but not defining features.

Refinements and modifications to this view of concepts have been made. Most notable is the idea that concepts are not simply unstructured collections of features; instead, they are held together by a 'conceptual glue' that provides information about the relation of features to one another and to the entities in question (e.g., Murphy and Medin 1985). For instance, a person's concept of birds might include the information that they all have wings, that most fly, and that some nest in trees, but these are not unrelated pieces of information. Instead, the concept includes the knowledge that having wings has a particular causal relation to flying and being able to fly has a particular causal relation to being able to nest in trees. If concepts include such information, then word meanings, likewise, do.

3. *Multiple Meanings: Polysemy*

Research on concepts has thus by now provided a fairly detailed (though still evolving) view of what concepts consist of. In that sense, word meanings are also understood in some detail. However, saying that meanings are concepts leaves some important issues unresolved. These issues revolve around the fact that many words have more than one meaning. In a small number of such cases, a word has two unrelated meanings, such as 'ear' meaning either a sense organ or a part of a corn plant. In these cases, it is often historical accident that the two meanings are expressed through the same word. During the earlier history of the language, two more distinctly different words may have been used, but they evolved into the same form as changes to the sound or spelling patterns of the language took place. Of greater interest for theories of meaning are cases of *polysemy*, where a single word has two or more meanings that are related in some way. For instance, the word 'run' means something somewhat different, but not entirely so, in each of the following phrases: 'run a race,' 'run errands,' 'run for election,' 'run a company,' 'the machine is running,' 'the water is running,' 'the movie runs three hours.' In other cases of polysemy, a word may be used in more radically different (yet still clearly related) ways. For instance, the word 'newspaper' most typically refers to a physical object consisting of printed pages of newsprint, but in sentences like 'The newspaper is being sold' or 'The newspaper says Jones is leading in the polls,' the same word is being used to refer to the corporate entity that produces the physical object, or to the writers whose writing appears in the physical object. Sometimes a word is even used as a part of speech it has not been used as before, as 'siren' is in 'The police sirened the car to a stop' (Clark and Clark 1979). The challenge for a psychological view of meaning is to explain how mental representations accommodate the different senses a word may have, and how the intended sense is understood.

One approach to addressing this challenge is to say that people have a single meaning stored in their memory for a given word, and variations on that meaning are created by the context it occurs in. A person would take into account his or her stored concept for the word along with other information in the sentence and discourse context to determine the meaning. In some cases, there may be regularities in what kinds of meanings can be constructed that allow people to quickly arrive at the relevant meaning. For instance, there may be a general rule that says that names of objects can be used to stand for people associated with those objects, making figuring out the meaning of cases like 'The newspapers says Jones is leading in the polls' straightforward (Nunberg 1979). In other cases, there may be no standard rule, but other aspects of the context, including knowledge of the background and goals of the person who said or wrote the sentence, are sufficient for the reader or hearer to figure out the right meaning.

This view assumes that the word's meaning itself changes in the contexts. An alternative way of framing what happens is to say that *interpretation* of the word is modified by the context, but the word meaning itself stays the same. For instance, the meaning of 'newspaper' is always about a physical object made of

newsprint, but a person can draw on that meaning, along with any standard rules and various clues from the sentence and discourse context, to figure out that the speaker is using the word 'newspaper' to refer to something somewhat different. This alternative framing is more consistent with the idea that meaning consists of concepts, because it distinguishes between meaning as stored knowledge and the use of that stored knowledge to construct an interpretation on an occasion of use.

Whichever framing is used, it is unlikely that all variations in interpretation/meaning are constructed anew in interpreting a word in a particular context. One reason for thinking so is that it has been impossible to determine 'core' meanings from which all other meanings can be derived. For instance, there is no set of features associated with the central meaning of 'run' that seem to be true of all the senses of 'run.' In some cases, senses of a word seem to develop in *chains* (Lakoff 1987, Case Study 2), where one sense derives from another, and another from that one, until there are senses having little or nothing in common with the typical sense. For instance, the concept associated with 'box' in its most central sense is something like a rectangular container made of cardboard for storing solid objects. However, there have come to be cardboard juice boxes (for drinking, not storing, and for liquids, not solids). From those developed plastic (not cardboard) reuseable juice boxes, and from them, plastic juice boxes in the shape of various animals (not rectangular). A plastic bear-shaped juice box has none of the properties associated with typical boxes, yet people willingly call it 'juice box.' Thus it seems likely that people learn and store senses for 'box' that are distinct from and in addition to their stored knowledge about ordinary cardboard boxes. Another reason for thinking that at least some of the variations are stored in memory is that some are encountered repeatedly (as in the case of 'juice box'), so it would be inefficient to have to derive them anew each time.

If, as this last point suggests, a number of different senses are stored in memory, one might wonder what part of the stored knowledge should be identified as the concept that constitutes the word's meaning. One way to resolve the problem may be simply to say that each word has multiple concepts associated with it, with the concepts in some cases differing only subtly from one another. Just as a person's concept of a sparrow vs. of a finch may be distinct but only slightly different, so may be his or her concepts associated with a polysemous word like 'run.'

4. *Meaning Across Languages*

So far, all the examples discussed have been of English words and their associated concepts. Speakers of English generally have the intuition that the distinctions they make between chairs and sofas, bottles and jars, and so on are so obvious that the same distinctions must be observed by speakers of other languages: the words would differ but the concepts associated with them would be the same. However, the situation is not that simple. Words that are traditionally translated as equivalent by dictionaries are not always fully equivalent on closer examination. For instance, Spanish speakers apply the word '*botella*' to a much narrower range of objects than English speakers apply the word 'bottle' to. On the other hand, English speakers apply the word 'on' to a more restricted range of spatial relations than Spanish speakers apply the word '*en*' to. Apparently, the concepts associated with the words are not identical. Such differences can be found across many languages and for words labeling many different kinds of entities.

This observation raises at least two important questions, one theoretical and one practical. The theoretical one concerns the relation between language and thought. A controversial proposal in psychology and linguistics has been that the language a person speaks influences the way he or she perceives the world (see Lucy 1992). If the meaning of 'bottle' and '*botella*' are not the same, or the meaning of 'on' and '*en*' are not, do people in some way perceive the objects or spatial relations differently depending on their language? For instance, if five objects are all called 'bottle' in English, but require two different names in Spanish, does the Spanish speaker see them as more different than the English speaker does? The available evidence suggests not (Malt et al. 1999). However, the *cross-linguistic differences* may influence other aspects of thought, such as how early a child learns a word or what features he or she first focuses on (e.g., Bowerman 1996).

The practical issue concerns *second language learning*. Second language instruction has typically taught vocabulary as paired associate learning: students learn that 'chair' is '*silla*' in Spanish, 'bottle' is '*botella*,' and so on. If the words are not truly equivalent in their meaning, then learners' use of the words may not be like native speakers'. For instance, learners may use the word '*botella*' for the same set of objects that they would use 'bottle' for in English and not realize that their use is inappropriate. It may take second language learners extensive experience with the language to fine tune their meanings and their word use to match native speakers, if they ever do.

See also: Meaning and Rule-following: Philosophical Aspects; Memory for Meaning and Surface Memory

Bibliography

Bowerman M 1996 Learning how to structure space for language: A cross-linguistic perspective. In: Bloom P, Peterson M A, Nadel L, Garrett M F (eds.) *Language and Space*. MIT Press, Cambridge, MA, pp. 385–436

Clark E V, Clark H H 1979 When nouns surface as verbs. *Language* **55**: 767–811

Hock H H, Joseph B D 1996 *Language History, Language Change, and Language Relationship: An Introduction to Historical Linguistics*. Mouton de Gruyter, Berlin
Laird C 1978 Language and the dictionary. In: Guralnik D B (ed.) *Webster's New World Dictionary of the American Language*, 2nd edn. Collins, Cleveland, OH
Lakoff G 1987 *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. The University of Chicago Press, Chicago
Lucy J A 1992 *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. Cambridge University Press, Cambridge, UK
Malt B C, Sloman S A, Gennari S, Shi M, Wang Y 1999 Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language* **40**: 230–62
Murphy G L, Medin D L 1985 The role of theories in conceptual coherence. *Psychological Review* **92**: 289–316
Nunberg G 1979 The nonuniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy* **3**: 143–84
Putnam H 1975 *Mind, Language, and Reality*. Cambridge University Press, New York
Rosch E, Mervis C B 1975 Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* **7**: 573–605

B. C. Malt

Word Order

As is well known, words in clauses and phrases are not strung together in a random fashion but are ordered in a communicatively significant way. Thus, for example, *John hit Mary* does not mean the same thing as *Mary hit John*. Moreover, the word orderings that are communicatively significant are not cross-linguistically uniform. Thus, whereas in English neither of the events described in the above two italicized clauses could be expressed with the verb in initial position, i.e., as *Hit Mary John*, this is precisely the order of words that would be used in Samoan or Berber. The nature of the word order patterns found within and across languages and the factors underlying their occurrence and distribution has been a topic of intensive study from a number of different perspectives. This article provides an overview of functional—typological research on word order concentrating on three issues, namely, the classification of languages in terms of their basic word order, the determination of the degree of language-internal word order variation, and the attempts to formulate a unifying principle of order capable of accounting for both cross-linguistic and language internal word order patterns.

1. Basic Word Order Typology

1.1 The Clause

Ever since Greenberg (1963) it has been the custom to classify the clausal word order of a language in terms of the location of the verb and its arguments relative to each other in transitive clauses (as opposed to intransitive or ditransitive ones) into one of the following six types: SOV, SVO, VSO, VOS, OVS, OSV, where S stands for subject, O for object, and V for verb. The transitive clauses which have served as the basis for word order classification are active, positive, declarative clauses with two nominal (not pronominal) participants, a definite human agent, and a definite nonhuman patient such as *The man killed the snake*, or *The girl extinguished the fire*. Greenberg took the order obtaining in such clauses to be the basic order of a language. For languages exhibiting more than one arrangement of the S, O, and V in the relevant type of clauses, the basic order was identified as the statistically dominant order.

Nowadays it is generally recognized that not all languages fall into one of the six word-order types. First of all, there are languages which simply lack transitive clauses with two overt NP participants, such as Puget Salish. Second, in many other languages, though clauses with two overt nominal participants do exist, they are quite rare, as both of the arguments must, or tend to, be expressed by pronominal affixes on the verb or by clitics elsewhere in the clause. For example, in her analysis of narrative texts in the South American language, Yagua, Payne (1990) found only 45 transitive clauses with two overt nominal participants, which constitutes less than 10 percent of the transitive clauses (463) and only 3 percent of all the clauses (1516) in the texts considered. In such languages, it may not always be possible to discern a clearly dominant order among the verb and its nominal arguments. Dryer (1997) suggests that the basic order should be at least twice as frequent as the order or orders with which it contrasts. In terms of such a criterion, none of the transitive orders in Yagua qualify as basic, as the frequencies noted by Payne were: VSO (19), SVO (15), and OVS (11). The third type of languages which may fall outside the six-way typology are languages in which, while transitive clauses with two overt nominal participants are quite frequent, no single order may unequivocally be dominant, as the ordering of constituents may be determined not by their grammatical relations but rather by their pragmatic status and discourse role.

In view of the above, several modifications of Greenberg's typology are currently in use. One involves using double classifications such as VSO/VOS or SVO/VSO. Another involves augmenting the six-way typology by a seventh type, namely, no basic order. A third modification consists in reanalyzing the six orders in terms of a three-way opposition of verb position into verb-initial, verb-medial, and verb-final. The most radical modification is that suggested by Dryer (1997), who argues for replacing the six-way typology by two separate two-way typologies of SV vs. VS and OV vs. VO, which together define four types: VS & VO, SV & VO, SV & OV, and VS & OV. While

this typology also concerns the order of nominals in transitive clauses, it differs from all of the above in being based on the presence of only one nominal participant, as opposed to two. Accordingly, it is more widely applicable and easier to apply than the six-way typology.

Despite the problems posed in classifying some languages in terms of their basic transitive order, the cross-linguistic investigations conducted since the 1960s identify the same basic word order preferences consistently; the most common basic order is SOV (+/−45 percent), followed closely by SVO (+/−35 percent), and then VSO (+/−10 percent). The three object-before-subject orders are all highly uncommon as basic orders, with VOS occurring slightly more frequently than OVS, and OSV being extremely rare.

Turning to the basic order in other than transitive clauses, with few exceptions (e.g., Salinan, Woods Cree, Makushi), the basic order of the sole argument of the intransitive verb is the same as that of the S argument of a transitive clause. Consequently, the classification of a language's transitive order also covers that of intransitive clauses. This also holds to a large extent for ditransitive clauses, i.e., clauses featuring, in addition to an agent and patient, a recipient (e.g., *Mother gave the baby a toy*.) Again with relatively few exceptions (e.g., Nadeb, Trumai, Yoruba) recipients tend to be placed on the same side of the verb as the O in transitive clauses, and typically occur contiguous to the patient. According to Primus (1997), in the basic order recipients are never separated from the patient by the agentive argument.

1.2 The Phrase

Order within phrases is characterized typically in terms of a two-way typology originating in the dependency tradition and the work of the nineteenth-century German typologists. This two-way typology is based on the division of the constituents of phrases into head/modifier pairs such as: verb/object, verb/adverbial, auxiliary/main verb, adposition/noun phrase, noun/adjective, noun/genitive, noun/relative clause, etc. (where the first member of the pair is the head and the second the modifier). As reflected in Vennemann's (1972) natural serialization principle and its analogs, languages are assumed to show a preference for serializing constituents with reference to either the head > modifier or the modifier > head schema. Those that exhibit basic transitive VO order are expected to display head > modifier order, and those that have basic OV order, modifier > head patterns. Languages that conform to these expectations are termed consistent. Consistency, however, is viewed currently not as the cross-linguistic norm, but rather as an ideal to which languages approximate to various degrees. And indeed very few languages are absolutely consistent in ordering all their modifiers either to the left or the right of the head. Thus, for example, of the 142 languages in Greenberg's Appendix II, only 68 (48 percent) exhibit consistent ordering in relation to just four head/modifier pairs, i.e., the verb and object, adposition and noun phrase, noun and genitive, and noun and adjective. Needless to say, the number of consistent languages drops once further head/modifier pairs are considered.

Several refinements to the expectation of consistent ordering have been suggested. One of these is the principle of cross-category harmony (PCCH) formulated by Hawkins (1983). The PCCH states that languages display a preference to generalize the order obtaining in one head/modifier category to that of other categories. Departures from this preference in one head/modifier category are in turn likely to induce comparable departures in other head/modifier pairs. Thus, for example, Hawkins notes that the continuum in verb position, verb-initial, verb-medial, and verb-final is reflected in the position of the noun in relation to its modifiers, VSO languages being more consistently noun-initial than SVO ones, and non-rigid SOV languages (SOV languages which allow for the placing of some modifiers to the right of the verb) less consistently noun-final than rigid SOV ones.

Another refinement to the preference for consistent ordering has been introduced by Dryer (1992) who argues that this preference actually involves not head modifier pairs but rather lexical and phrasal categories. Dryer suggests that languages tend to be either right-branching, in which phrasal categories follow non-phrasal categories, or left-branching, in which phrasal categories precede non-phrasal categories. This is argued to hold irrespective of which is considered to be the head and which the modifier. For non-branching categories, no word-order predictions are made. Thus, according to Dryer's branching direction theory, whereas, for example, a genitive phrase (e.g., of John) and relative clause (whom I met) should occur on the same side of the noun, a demonstrative phrase need not, as it is not a branching category; nor need an adjective pattern as the genitive phrase or relative clause, since it is not a fully recursive phrasal category. While Dryer's word order predictions receive strong cross-linguistic support, their correctness is heavily dependent on specific theories of constituent structure which are not uncontroversial.

The notion of consistent ordering in its various guises relates to the order of head/modifier pairs and says nothing about the ordering of modifiers relative to each other. This latter aspect of phrasal order is still under investigation. One subset of modifiers whose order has received detailed attention is that of the demonstrative, numeral, and adjective within simple NPs. Rijkhoff (1997) accounts for the cross-linguistic linearizations of these three modifiers with reference to three principles, namely, the principle of domain integrity, the principle of head proximity, and the principle of scope. All three principles are elaborations

of what is known as Behagel's (1932, p. 4) first law, which is 'what belongs together semantically is also placed together syntactically.' Of the 24 logically possible linearizations of the demonstrative, numeral, adjective, and noun, the three principles define the following eight word-order patterns as potential basic orders.

(a)	(b)
DemNumAdjN	AdjNNumDem
DemNumNAdj	NumAdjDem
DemNAdjNum	DemAdjNNum
NumNAdjDem	NAdjNumDem

Rijkhoff's cross-linguistic data reveal that all these orders, and only these, are attested as basic orders.

2. *Word Order Variation*

Languages are classified not only in regard to their basic order, but also with respect to the amount of word order variation that they display. However, though labels such as rigid or fixed-order language, or free or flexible-order language are commonplace, there is little consensus on what they actually mean. This is not altogether surprising as it is far from clear how word-order variation could be measured.

The major problem in devising a measure of word-order variation is the sheer number of potential variables involved. An alternative order may be conditioned by a wide range of factors—syntactic, semantic, and pragmatic—relating to the nature of the text and/or genre (e.g., written vs. spoken, narrative vs. expository vs. procedural), the properties of the clause (e.g., its illocutionary force or main vs. subordinate status), the features of the verb (e.g., its voice, tense, or aspect) and the characteristics of its arguments and adjuncts such as referentiality, definiteness, givenness, humanness, animacy, semantic role, and morphological and syntactic complexity. Needless to say, there is no obvious reason why changes in order attributable to any one factor or set of factors should be viewed as being more significant than others.

The above notwithstanding, some idea of the word-order flexibility of a language can be gathered by reducing the number of variables and considering the number of word-order permutations in clauses or head/modifier pairs exhibiting the same characteristics as those used for determining the basic order. In terms of such a measure, English emerges as a fairly rigid word order language at the clausal level, as the only variation on its basic SVO order that it permits is OSV. Polish, on the other hand, is a highly flexible word order language, since in addition to its basic SVO order it also has all the other five permutations of the S, O, and V. Applying the above measure of word-order flexibility to a sample of 171 1anguages, Siewierska (1997) found no variation on the basic transitive order in 22 percent of the languages, only one variant in 35 percent, two variants in 17 percent, three variants in 12 percent, four variants in 6 percent, and five variations in 8 percent. Thus languages with no or little word-order flexibility appear to be considerably more common than those with highly flexible order.

Some generalizations can also be made with respect to the type of word order variations that languages are most and least likely to exhibit, namely (a) verb initial languages are least likely to exhibit verb final orders, and verb final languages, verb initial orders, (b) in languages of all word order types, SVO is a more likely variable order than OVS. Thus, an SOV language with fairly inflexible order may be expected to have an OSV variant, or possibly a SVO one, but not a VSO or OVS variant. And a VSO language may be expected to exhibit SOV order only if it also has VOS, SVO, and OVS.

The degree of word order flexibility of a language with respect to the ordering of the verb and its arguments is often seen to be tied to the existence of agreement and case marking. The argument is that, if the subject is distinguished morphologically from the object by case and/or agreement marking, word order is 'freed' from performing this function and can thus be used for other communicative ends. This, however, is not necessarily so. Cross-linguistic data cited by Siewierska (1997) reveal that, contrary to what is often suggested, it is not the presence of case and agreement marking that has an affect on word-order flexibility but rather their absence. Rigid order is associated strongly with lack of agreement and to a lesser extent with lack of case marking, but neither the presence of case nor agreement marking correlates with flexible order.

The investigation of word-order variation carried out to date suggests that languages exhibit more variation in the ordering of constituents at the level of the clause than at the level of the phrase. Significant differences in word-order flexibility have also been observed in regard to specific head/modifier pairs. Bakker's (1997) investigations of the possibility of alternative orderings of the noun and a single modifier among the languages of Europe suggests that the modifiers of the noun can be placed in a hierarchy of decreasing likelihood of exhibiting an alternative order to that of their basic order, namely: adjective > genitive > relative clause > numeral > demonstrative. Whether this will be. confirmed globally remains to be determined.

3. *Discourse-pragmatic vs. Processing, Explanations of Order*

Both the cross-linguistic preferences in regard to the distribution of basic orders as well as the type of variations on the basic order that languages tend to favor have typically been accounted for in discourse-pragmatic terms. Until recently, the dominant principle assumed to underlie linearization has been the

Topic > Comment principle, which defines a preference for placing constituents encoding more topical, predictable, given information before those encoding less topical, unpredictable, new information. The universal applicability of this principle has been called into question by the existence of languages which appear to favor placing the comment not after, but before, the topic. Givón (1988) has sought to reconcile the existence of two opposing principles of order by formulating the principle of Task Urgency, which is: 'Attend first to the most urgent task,' where a communicative task is identified as more urgent if the information to be communicated is either less predictable or more important. Givón suggests that when the topic of discourse is a highly predictable referent, the most urgent task is to present the comment, and when the topic corresponds to a less predictable or unpredictable discourse referent, the most urgent task is to state the topic. Thus the principle of Task Urgency predicts a cross-linguistic preference for placing either less predictable > more predictable and/or more important > less important information.

An alternative unifying principle of order has been posited recently by Hawkins (1994), who sees order as being determined primarily by processing considerations and not discourse-pragmatic ones. Building on the long-recognized tendency to place short constituents before longer ones, Hawkins hypothesizes that the human parser displays a preference for linearization patterns which allow the quickest recognition of immediate constituents of syntactic groupings. This is captured in the Early Immediate Constituent recognition principle (EIC), which is: 'The human parser prefers to maximize the left-to-right IC (immediate constituent)-to-word ratios of the phrasal nodes that it constructs'. The EIC predicts that information about immediate constituency should be placed in the earlier rather than the later part of a string. For relatively heavy categories, such as sentential and NP complements, this prediction entails a preference for short > long linearization in right-branching (VO) languages and for long > short in left-branching (OV) languages. But in the case of single-word categories, an overall preference for placing the light category to the far left is defined.

Strong cross-linguistic evidence for the EIC has been cited from several types of data: real-time psycholinguistic experiments of alternative orderings, of ICs, language-internal text frequency counts, native speaker acceptability judgments, and the cross-linguistic distribution of basic orders. These data show that not only is there an unequivocal preference for orderings defined as optimal by the EIC, but also that there is a progressive decline in the frequency of non-optimal orders directly proportional to the degree of their non-optimality defined by the EIC. Thus, for example, while the EIC defines the presence of both prepositions in OV languages, and of postpositions in VO as non-optimal, the EIC score for the former, e.g., for *with the friends played* is considerably lower (20 percent) than for the later, e.g., *played the friends with* (65 percent). And though both orders are cross-linguistically uncommon, in Dryer's (1992) data, OV languages with prepositions are three times less common than VO languages with postpositions (1.5 percent vs. 4.6 percent).

Whether an EIC account of word order is indeed superior to one based on discourse-pragmatic principles, as Hawkins argues, remains to be seen.

See also: Grammar: Functional Approaches; Word Classes and Parts of Speech; Word, Linguistics of

Bibliography

Bakker D 1997 Flexibility and consistency in word order patterns in the languages of Europe. In: Siewierska A (ed.) *Constituent Order in the Languages of Europe*. Mouton de Gruyter, Berlin

Behaghel O 1932 *Deutsche Syntax: eine geschichtliche Darstellung. Band IV: Worststellung-Periodenbau*. Carl Winter, Heidelberg, Germany

Dryer M 1992 The Greenbergian word order correlations. *Language* **63**: 81–138

Dryer M 1997 On the six-way word order typology. *Studies in Language* **21**: 69–103

Givón T 1988 The pragmatics of word order: Predictability, importance and attention. In: Hammond M, Moravcsik E A, Wirth J R (eds.) *Studies in Syntactic Typology*. John Benjamins, Amsterdam, The Netherlands

Greenberg J A 1963 Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg J (ed.) *Universals of Language*. MIT Press, Cambridge, MA

Hawkins J A 1983 *Word Order Universals*. Academic Press, New York

Hawkins J A 1994 *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge, UK

Payne D 1990 *The Pragmatics of Word Order. Typological Dimensions of Verb Initial Languages*. Mouton de Gruyter, Berlin

Primus B 1997 The relative order of the patient and recipient in the languages of Europe. In: Siewierska A (ed.) *Constituent Order in the Languages of Europe*. Mouton de Gruyter, Berlin

Rijkhoff J 1997 Order in the noun phrase of the languages of Europe. In: Siewierska A (ed.) *Constituent Order in the Languages of Europe*. Mouton de Gruyter, Berlin

Siewierska A 1997 Variations in major constituent order: A global and a European perspective. In: Siewierska A (ed.) *Constituent Order in the Languages of Europe*. Mouton de Gruyter, Berlin

Vennemann T 1972 Analogy in generative grammar, the origin of word order. In: Heilmann L (ed.) *Proceedings of the Eleventh International Congress of Linguistics*, Vol. 2. Il Mulino, Bologna, Italy

A. Siewierska

Copyright © 2001 Elsevier Science Ltd. All rights reserved.

 ISBN: 0-08-043076-7

Word Recognition, Cognitive Psychology of

The term 'word recognition' has two senses. Cognitive psychologists use it to refer to a component process of mind that transforms the printed or spoken features of a word into a linguistic representation. Cognitive psychologists must infer component processes from measures of behavior. Word recognition is a measured task performance. Laboratory volunteers perform a task in which they discriminate words from nonwords (the lexical decision task), and measurements of their performance are used to test hypotheses about the nature of word recognition. At one time, performance in the lexical decision task was assumed to provide an unambiguous view of the word-recognition component in operation. For example, 'there is no way to perform this task without [word recognition], we can estimate the time for this operation by measuring the subject's reaction time to make the decision' (Forster 1976, p. 260). As a performance phenomenon, word recognition may or may not imply a distinct component process. Nevertheless, it is intuitive that reading and listening include a component process of word recognition, as the examples that follow may illustrate.

1. Component Processes and the General Linear Model

1.1 Intuitive Word Recognition

Nineteenth century eye movement studies discovered that a reader's eyes make a series of jumps, while moving across a line of text, pausing about a quarter of a second when fixated. Eye movements thus implied the 'fixation' (recognition) of individual printed words in natural reading. A skilled reader can recognize thousands of printed words with no noticeable effort. A dyslexic child, on the other hand, does not easily acquire this skill. Recognizing a printed word, as a particular word, can be effortful to the point of frustration. Dyslexia may plague an otherwise bright and articulate child, which appears to implicate word recognition as the crux of reading. Reading is not exclusively the recognition of printed words, but it is word recognition that distinguishes reading from natural language. In effect, to become a reader is to master this special skill.

A different example illustrates word recognition in spoken language. When listening to speech in an unfamiliar language, people inevitably wonder why speakers 'talk so fast.' All languages are spoken at equivalent rates, but in a familiar language, we perceive breaks between words (although no physical breaks exist), which 'slows down' perception. Unfamiliar languages sound fast because the listener lacks the ability to segment continuous speech and to recognize the words.

In the examples of dyslexia and foreign speech, our intuition that a component of word recognition may exist derives from the absence of competent word recognition. However, the absence of a hypothetical competence tells us little or nothing about whether the hypothesis is valid. This requires scientific investigation, and a logic for the analysis of complex behavior.

1.2 Nearly Decomposable Systems

Herbert Simon proposed a logic of analysis for complex systems, such as cognitive systems. Complex systems can be partly decomposed, if the components interact linearly. The internal dynamics of cognitive components may be complex (nonlinear). If the interactions between components are, nevertheless, linear, then it is possible to identify these components. Cognitive components, described thus, work as a chain of single causes. Single causes entail the familiar notion of domino causality. Push the first domino in a chain of standing dominoes and each will fall in its turn. Linear interactions, between the dominoes, add up to the total behavioral effect. If behavior is the sum of its parts, the total effect can be reduced to component causes. This formal logic justifies behavioral studies to infer underlying cognitive components.

1.3 Additive Factors Method

The most well-known methodological tool to individuate cognitive components is the additive factors method, proposed by Saul Sternberg. Experiments that include several experimental conditions (i.e., factorial designs) allow simultaneous manipulation of several factors, which provides the opportunity for interaction. If the effects of two or more manipulations are strictly additive, the manipulated variables satisfy the assumption of selective influence—they selectively influence distinct components. For example, the total time from first to last domino, in the previous, idealized chain of falling dominoes, is the sum of each domino's falling time. In this idealization, separate experimental manipulations that slow falling times of individual dominoes, but do not change the falling times of dominoes that precede or follow an affected domino, satisfy the assumption of selective influence. A manipulation selectively effects a single domino's falling time, without affecting other dominoes' falling times. Such effects simply add time to the total falling time of the entire chain. Alternatively, when nonadditive interactions are observed, manipulations have not satisfied the assumption of selective influence. Factors that are not additive, influence (at least) one common component. Thus, to Sternberg's lasting credit, his method includes an empirical failure point: ubiquitous nonadditive interaction effects.

2. Word Recognition as Information Processing

The logic of the additive factors method also fits the metaphor of cognition as information processing. For the chain of dominoes, substitute the guiding analogy of a flow chart of information processing, like unidirectional flow charts of computer programs. Information flows from input (stimulus) to output (behavior) through a sequence of cognitive components. In word recognition, input representations from a sensory process—visual or auditory features of a word—are transformed into an output representation—the identity of the word—that, in turn, becomes the input representation for a component downstream (i.e., a component of response production or sentence processing). In this tradition, empirical studies of word recognition pertain to the structure and function of the lexicon. The lexicon is a memory component, a mental dictionary, containing representations of the meanings, spellings, pronunciations, and syntactic functions of words. Structure refers to how the word entries are organized, and function refers to how words are accessed in, or retrieved from, the lexicon. Two seminal findings illustrate the distinction: semantic priming effects and word frequency effects. Both effects are found in lexical decision performance.

2.1 The Lexical Decision Task

In the lexical decision task, a person is presented, on each trial, with a target string of letters, and must judge whether the target string is a correctly spelled word in English (or some other reference language). Some trials are catch trials, which present nonwords such as 'glurp.' (One may also present words and nonwords auditorally, to examine spoken word recognition.) The participant presses a 'word' key to indicate a word and a 'nonword' key otherwise. The experimenter takes note of the response time, from when the target stimulus appeared until the response key is pressed, and whether the response was correct. Response time and accuracy are the performance measures.

2.2 Semantic Priming and the Structure of the Lexicon

Word pairs with related meanings, such as 'doctor' and 'nurse' or 'bread' and 'butter,' produce semantic priming effects. Semantic priming was discovered by David Meyer and Roger Schvaneveldt, working independently (they chose to report their findings together). Lexical decision performance to a word is improved by prior presentation of its semantically related word. Prior recognition of 'doctor,' as a word, facilitates subsequent recognition of 'nurse'; lexical decisions to 'nurse' are faster and more accurate, compared with a control condition. This finding is commonly interpreted to mean that semantically related words are structurally connected in the lexicon, such that retrieval of one inevitably leads to retrieval of the other (in part or in whole).

2.3 Word Frequency and the Function of Lexical Access

Word frequency is estimated using frequency counts. The occurrence of each word, per million, is counted in large samples of text. Lexical decision performance is correlated with word frequency. Words that occur more often in text (or in speech) are recognized faster and more accurately than words that occur infrequently. This finding is interpreted in a variety of ways. The common theme is that lexical access functions in a manner that favors high-frequency words. In one classical account, proposed by John Morton, access to a lexical entry is via a threshold. Word features may sufficiently activate a lexical entry, to cross its activation threshold, and thus make that entry available. Common, high-frequency words have lower threshold values than less common words. In a different classical account, proposed by Kenneth Forster, the lexicon is searched in order of word frequency, beginning with high frequency words.

2.4 Challenges to the Information Processing Approach

Additive interaction effects are almost never observed in word recognition experiments, and, while it is not possible to manipulate all word factors simultaneously in one experiment, it is possible to trace chains of nonadditive interactions across published experiments that preclude the assignment of any factors to distinct components. Moreover, all empirical phenomena of word recognition appear to be conditioned by task, task demands, and even the reference language, as the examples that follow illustrate.

The same set of words, which produce a large word frequency effect in the lexical decision task, produce a reduced or statistically unreliable word frequency effect in naming and semantic categorization tasks. All these tasks would seem to include word recognition, but they do not yield the same word recognition effects. Also, within the lexical decision task, itself, it is possible to modulate the word frequency effect by making the nonwords more or less word-like (and, in turn, to modulate a nonadditive interaction effect between word frequency and semantic priming). Across languages, Hebrew produces a larger word frequency (familiarity) effect than English, and English than Serbo–Croation.

Consider the previous examples together, within the guidelines of additive factors logic. Word recognition factors cannot be individuated from each other, and they cannot be individuated from the context of their occurrence (task, task demands, and language). The

limitations of additive factors method are well known. Because additivity is never consistently observed, we have no empirical basis for individualizing cognitive components. The *de facto* practice in cognitive psychology is to assume that laboratory tasks and manipulations may differ from each other by the causal equivalent of one component ('one domino'). But how does one know which tasks or manipulations differ by exactly one component? We require *a priori* knowledge of cognitive components, and which components are involved in which laboratory tasks, to know reliably which or how many components task conditions entail. Notice this circularity, pointed out by Robert Pachella: the goal is to discover cognitive components in observed laboratory performance, but the method requires prior knowledge of the self same components.

Despite these problems, most theorists share the intuition that a hypothetical component of word recognition exists. When intuitions diverge, however, there may be no way to reconcile differences. Theorists who assume that reading is primarily an act of visual perception discover a visual component of word recognition; theorists who assume that reading is primarily a linguistic process discover a linguistic component of word recognition, in the same performance phenomena. Repeated contradictory discoveries, in the empirical literature, have lead to a vast debate concerning which task conditions provide an unambiguous view of word recognition in operation. The debate hinges on exclusionary criteria that may correctly exclude task effects and bring word recognition into clearer focus. Otherwise, inevitably, one laboratory's word recognition effect is another laboratory's task artifact.

3. *Connectionism*

Context effects seem to occur at all scales at which words may be viewed. For example, just as semantically related prime words (and appropriate sentence contexts) produce benefits for word recognition; words themselves, as contexts, affect letter and phoneme identification. A briefly presented letter is more accurately identified if it is presented within a word than the same letter presented in a nonword, or presented alone. Also, an ambiguous initial consonant, which could be either a /d/ or a /t/, is more likely to be identified as /d/ in the context of /_ash/ (where /dash/ is a word and /tash/ is not), and as /t/ in the context of /_ask/ (where /task/ is a word and /dask/ is not).

3.1 *Interactive-Activation Models*

James McClelland and his colleagues constructed interactive activation models to simulate the previous context effects. The original interactive activation model simulated context effects on letter identification, and was extended in the another model to simulate context effects on phoneme identification. Interactive activation models are connectionist models. In a connectionist model, constraints on response options are implemented as excitatory and inhibitory connections among nodes that behave as idealized neurons. The original interactive activation model included a hierarchy of representations, implemented in three node levels: visual feature nodes, letter nodes, and word nodes. Most important, letter nodes and word nodes have reciprocal excitatory connections, which allow feedback from word representations to excite letter representations. As a consequence, letter nodes that are inadequately activated by feature nodes may benefit from word feedback. This is the hypothetical basis of word context effects on letter (and phoneme) identification.

3.2 *PDP Models*

In interactive activation models, the connection strengths between nodes are preset by the modeler. Parallel distributed processing (PDP) models include learning algorithms capable of covariant learning. Covariant learning shapes a matrix of connection weights to reflect statistical relations between the input and output patterns of a training set. Systematic relations, at a variety of grain sizes, between English spelling and phonology (or phonology and semantics, etc.), may all be construed as statistical relations. For example, consonant spellings are more strongly correlated with consonant pronunciations than are vowel spellings with vowel pronunciations, but in both cases there are statistically dominant and subordinate relations. Regular words, with dominant relations, are named more quickly than exception words, with subordinate relations. Likewise, body-rime relations may be consistently correlated (the body _uck always indicates the same rime, in English, as in the word 'duck'); but other body-rime relations are less strongly correlated (_int pronounced as in 'mint'); and still others are only weakly correlated (_int pronounced as in 'pint')—a rank order that is also corroborated by readers' naming times. The emphasis of PDP models, on learning the relation between spelling and phonology, coincides with the commonly observed failure of dyslexics to sound out words, and pronounceable pseudowords such as 'glurp.' Not all scientists agree, but the most common form of dyslexia appears to be a specific failure to learn the fine-grain consonant and vowel relations between spelling and phonology.

Covariant learning tracks all grain sizes of covariation, simultaneously, in the connection weights of a PDP model. Even word frequency approximates a relative correlation among whole-word spellings and whole-word pronunciations, and high-frequency

words are named faster than low-frequency words. The outcome of covariant learning is determined by statistical relations in the training set, and so a description of the training set is an integral part of the theory. In effect, covariant learning attunes a network to constraints inherent in a literate culture, for example, the relation between spelling and phonology. Consequently, the description of the cultural artifact, implicit in the training set, is as important for cognitive theory as hypothetical, internal, cognitive processes. Debra Jared used this aspect of PDP models to derive a nonintuitive empirical test using a naming study. The aggregate statistical relation of spelling to phonology, in body-rimes of high-frequency words, predicted a statistical advantage for some high-frequency words over others. Previously, word recognition of higher frequency words was assumed to be immune to such statistical relations. Nevertheless, Jared corroborated the prediction: faster naming times to high-frequency words with consistent body-rime relations. She did not observe this effect in the lexical decision task, however.

3.3 Feedback Consistency Effects

Body-rime consistency effects, in lexical decision performance, were predicted using a combination of interactive activation and covariant learning. Resonance models include learning algorithms that induce symmetrical covariant relations, consistent in both feed forward and feedback directions. Consequently, they predict that symmetrical, consistent relations, between spelling and phonology (and phonology and spelling), imply faster and more accurate word recognition. For example, the body-rime (and rime-body) relation in 'duck' is a symmetrical, consistent relation. All words with the body _uck are pronounced to rhyme with 'duck'; all words with the rime /_uk/ are spelled with the body _uck, as in 'duck.' Inconsistent relations, including inconsistent feedback relations, from phonology to spelling, add time and variability to performance of word recognition. The predicted feedback consistency effect is highly nonintuitive. From an information processing view, processes should always flow forwards, as from spelling to phonology. It should not matter in visual word recognition that a pronunciation may have more than one spelling (or in spoken word recognition that a spelling may have more than one pronunciation).

Feedback consistency effects have been found in performance of English and French lexical decision by Greg Stone, Johannez Ziegler and their colleagues. Words such as hurt (in English), with phonological rimes (/_ûrt/) which could be spelled in multiple ways (_urt, _ert, _irt) yield slower lexical decision times and more errors than words with rimes spelled in only one way. Also, a symmetrical feedback consistency effect is found in performance of auditory lexical decision (in English and French). What is feed forward for visual presentation is feedback for auditory presentation, and vice versa—a parsimonious qualitative symmetry. Moreover, once feedback consistency is taken into account, reliable feed forward consistency effects emerge in visual and auditory lexical decision performance, effects previously thought to be unreliable.

4. Recurrent Network Models and Nonlinear Dynamical Systems Theory

Interactive activation models and PDP models were actually proposed as first steps toward strongly nonlinear models that combined their features. This combination was realized in fully recurrent 'neural' networks (resonance models). Recurrent networks are attractor networks simulated as nonlinear iterative maps. An iterative map takes its output at one time-step as input on the next time-step, until it reaches a stable pattern of node activity (an attractor pattern).

Nonlinear iterative maps approximate solutions of systems of nonlinear differential equations. Thus, recurrent network models, as dynamical systems, invoke the mathematical framework of nonlinear dynamical systems theory. A few recurrent network models of word recognition have been implemented, notably by Alan Kawamoto (and colleagues), Michael Masson, and Stephen Grossberg (and colleagues).

Nonlinear dynamical systems theory concerns the complex behavior of systems, produced by components that interact nonlinearly (nonadditively). A few pioneering studies have corroborated empirical signatures of nonlinearity, including multistability in performance of printed word naming, hysteresis in performance of spoken word identification, and 1/f noise in lexical decision.

4.1 Multistability

Multistability is a generic phenomenon of nonlinear dynamical systems. In this case, the same stimulus word reliably produces more than one naming response. Homograph words are an obvious example. Homographs, such as 'wind,' have two or more pronunciations, and are thus multistable by definition. A word like 'pint,' with a statistically subordinate pronunciation (consider 'hint,' 'lint,' and 'mint'), is a more subtle example. Readers produce two systematic pronunciations to 'pint': the correct pronunciation and an error pronunciation that rhymes with 'mint.' Moreover, when readers are encouraged to produce rapid pronunciations, they are much more likely to produce the systematic mispronunciations. In the latter manipulation, a quantitative change in the time available for a naming response produces a qualitative change in the response itself. Qualitative changes consequent on quantitative manipulations are precisely the kind of phenomena that are addressed by nonlinear dynamical systems theory.

4.2 Hysteresis

Hysteresis effects are more elaborate signatures of nonlinearity, which extend the concept of multistability. Betty Tullerand her colleagues have demonstrated hysteresis effects in perception of artificial speech. As in the previous example, of homograph words, an identical pattern of artificial speech may yield multiple (multistable) perceptions. A typical experiment manipulates the presentation order of speech stimuli. Stimuli are constructed to change quantitatively and incrementally in acoustic properties along a continuum, between the words 'say' and 'stay', for example. Each run of an experiment presents the continuum, one stimulus at a time, running from 'say' to 'stay', and back again (or vice versa). Hysteresis is observed when some intermediate range of stimuli is perceived as 'say'—if preceded by 'say' stimuli—but the identical range is perceived as 'stay' if preceded by 'stay' stimuli. This intermediate range is multistable; and the context effect demonstrates hysteresis.

The hysteresis pattern is a generic pattern that is observed widely in physical, chemical, biological, and cognitive systems. Prior to mathematical developments in this century, however, it was considered a nuisance effect. As a nuisance effect, it motivated one of the first methods in psychology—Fechner's method of limits in nineteenth century psychophysics—effectively, a technique to get around the problem of hysteresis. Generic constructs such as hysteresis and multistability provide an analogy to understand ambiguity resolution in natural language. Words are pronounced differently, and have different meanings, in different contexts. Context sensitivity is a defining feature of multistable phenomena (as hysteresis demonstrates), and the mathematical framework of nonlinear systems yields a simplifying formal perspective on context sensitivity in natural systems.

4.3 1/f Noise

David Gilden and his colleagues have observed 1/f noise in the fluctuation of trial-by-trial response times from several cognitive tasks, including a word recognition task. 1/f noise is observed in the residual 'error' variance of individual participants' trial-by-trial response times (the variability that remains after 'treatment effects' are removed). If we graph each residual time, in the trial order of the experiment, the data points fluctuate between fast and slow times. The connected data points form a complex waveform, which may be viewed as a composite of waves spanning a range of frequencies. 1/f noise is an inverse correlation between the frequency of the composite waves and their power (amplitude).

The phenomenon of 1/f noise can be difficult to grasp, because it goes against the grain of a typical psychological analysis. Typically, error variance is discarded, rather than analyzed for structure. In Gilden's data, the error variance is analyzed and found to resemble the mathematically generic pattern of 1/f noise (a construct from fractal geometry). 1/f noise is a signature of processes that have no characteristic measurement scale. It contradicts additive factors logic, which assumes that we may partition response times into additive, independent sources of variance. This practice strictly requires that the response time in each trial is independent of the response times in other trials. The assumption of independence is at the heart of the linear statistical models used to identify cognitive components. The presence of 1/f noise contradicts this assumption. Response time data do not have 'joints' that may reduce to component causes.

4.4 Challenges for Connectionism

A laudable feature of the information processing approach was its explicit logic and method of analysis, derived from the general linear model, and implemented in additive factors method. No comparable logic has gained general acceptance within the connectionist approach (although connectionist models have been used productively in tests of nonintuitive predictions). In large part, connectionism has inherited the methodology of information processing psychology. However, empirical analyses that assume the general linear model, and theories implemented as strongly nonlinear dynamical systems, are incompatible at their root. They entail contradictory notions of cause and effect.

The domino causality of information processing analysis individuates single causes in additive effects—cause and effect relations that allow the morphological reduction of behavior to underlying component causes. In contrast, the circular causality of strongly nonlinear systems allows that all the components of a system may be present in qualitatively different behaviors of the system. Causal properties emerge in the interaction of components, which are not reducible to causal properties of the components themselves. Circular causality requires a strategic (not morphological) reduction. In a strategic reduction, the same, generic, nonlinear phenomena may be observed at multiple levels of a system. One must concede, however, that higher-level phenomena do not reduce to lower-level causes.

The previous contradiction between linear methods and nonlinear models raises questions, as yet unanswered, for the cognitive psychology of word recognition. For example, is it more productive, for scientific purposes, to view word recognition performance as a product of the nervous system (which also appears as a nonlinear dynamical system), a product of inscrutable components of mind (which have combined in nonlinear interaction), an emergent product of interaction between readers and texts, or

some other possibility that is not articulated in the previous alternatives? Answers to such questions await a generally accepted and reliable logic of nonlinear analysis, appropriate to cognitive performance, connectionist models, and nonlinear dynamical systems theory.

See also: Connectionist Models of Language Processing; Lexical Processes (Word Knowledge): Psychological and Neural Aspects; Priming, Cognitive Psychology of

Bibliography

Carroll D W 1994 *Psychology of Language 2nd edn*. Brooks/Cole, Pacific Grove, CA

Farmer J D 1990 A Rosetta Stone for connectionism. *Physica D* **42**: 153–87

Forster K I 1976 Accessing the mental lexicon. In: Wales R J, Walker E (eds.) *New Approaches to Language Mechanisms*. North-Holland, Amsterdam

Frost R, Katz L (eds.) 1992 *Orthography, Phonology, Morphology, and Meaning*. North Holland, Amsterdam

Plaut D C, McClelland J L, Seidenberg M S, Patterson K 1996 Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review* **103**: 56–115

Port R F, van Gelder T (eds.) 1995 *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, MA

Rayner K, Pollatsek A 1989 *The Psychology of Reading*. Prentice Hall, Englewood Cliffs, NJ

Van Orden G C, Stone G O, Pennington B F 1990 Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review* **97**: 488–522

Van Orden G C, Pennington B F, Stone G O 2001 What do double dissociations prove? *Cognitive Science* **25**

M. A. Moreno and G. C. van Orden

Work and Labor: History of the Concept

The concept of work has at least three different but complementary meanings in contemporary Western societies. As an action, it refers to the content and the performing of an activity; as a product, it is part of the organization of capitalist societies; finally, as a status, it constitutes a linchpin of social organization. Work, as the mainspring of economic activity, is also at the crux of the functioning of the political and social world. Affecting the slightest aspects of life in society, it can be viewed as a 'total social fact' (Mauss 1966). This pivotal position has made work a notion that has long been taken for granted as universal, to the detriment of its variability in time and space (Kocka and Offe 2000). Yet the contemporary meaning given to the concept is the result of an eminently historical process, including in capitalist societies in which work constitutes the yardstick. Although various meanings of the word 'work' existed in the nineteenth century (Cottereau 1994), it is on the particularly restrictive concept of labor, in the sense of quantifiable productive activity, that Western capitalist societies have been built.

The development of the concept was concurrent with the evolution of the social sciences. Labor, conceptualized in political economy after Adam Smith in terms of value, came to be seen as the mediator between the individual and the collective. It thus became a focal point of sociological analysis. The transformations of productive activity and their consequences in industrial society have certainly contributed to shaping the main directions of development in social sciences, but the descriptive as well as prescriptive discourse of the latter has also played an essential role since the early twentieth century in establishing the social sphere on the basis of work.

The history of the concepts of work and labor will be examined from this angle of a dual process in which the social sciences and the societies they study mutually engender one another. Thus delimited, the history will not be a total one, but situated in time and space, that is, confined to the question of the reduction of the concept of work to the concept of labor in capitalist European societies from the nineteenth century onward. This rather Eurocentric history is also the history of the statistic and legal codification of labor into a social status, providing access to citizenship within the welfare state. Finally, since the 1970s it has been the history of re-examining the concept, closely tied to a reassessment of paradigms of action in the social sciences.

1. The Semantic Reduction of Work to Labor

Labor, more precisely wage labor, a dominant form in social–economic relationships in today's societies, has not always had this value as a standard. In the Middle Ages, when idleness was the privilege of the aristocracy, labor was synonymous with poverty and social devaluation. Without going back over the history of work (see *Work, History of*), it is important to underline the contradictory heritage that characterizes the notion in the nineteenth century. On the one hand, work was associated with the idea of alienation. This association finds its origins in the Greek conception of slavery, the function of which was to satisfy the vital needs of reproduction of the human species (Arendt 1958). It was further developed by Marx's critique of wage labor, a modern form of alienation through work. On the other hand, work rhymes with accomplishment and self-expression. This opposite aspect, based on the Protestant formalization of an occupation (*Beruf*) as a vocation (Weber 1930), drawing on the Hegelian conception of work as the

essence of man (Hegel 1977/1805), considerably fed the ideology of the welfare state. This tension between alienation and liberation was to mark the concept of labor in the long term and run through theoretical debates in the social sciences even until today (Méda 1996).

1.1 The Invention of Abstract Work

The opposition between work considered as the essence of human existence and the reality of alienating labor provided, in particular, the basis of the Marxist critique of the political economy (Marx 1968/1867) that marked a major turning point in the conceptualization of work. Political economy, of which labor is a key concept, produced an initial swing of the concept of work toward that of labor. This swing occurred in the context of liberalization of work, consecrating the idea of the market through the promulgation of a set of legal provisions abolishing hereditary dependency, establishing the free circulation of individuals, freedom of commerce and free competition. On the theoretical level, Adam Smith's work (1776) some years earlier consecrated labor as a value, as a source of national wealth, and saw freedom of work as the means to maximize this wealth. From Adam Smith to John Stuart Mill, British theorists of political economy made labor a product, a commodity sold according to the laws of supply and demand, on a market held to be the dominant form of regulation. Determined by the principles of economic liberalism, this conception associates work with a manifestation of individual freedom, it is a specific commodity having an exchange value. Boiled down to a commodity that each person theoretically should be able to sell as he wishes by way of freely agreed contracts, work became, through contractual mediation, a general legal and abstract category, separate from the individual who produced it. This was the invention of abstract work, quantifiable and measurable in time and money; this was labor, a founding principle of capitalism (Biernacki 1995).

Marx embraced this conceptualization of abstract work, but introduced at least two new elements. The first was labor power. Marx viewed labor less from the perspective of the resulting product but from the angle of the labor power (*Arbeitskraft*) required to achieve it. Biernacki (1995) sees in this shift the expression of the structuring nature of the German work culture on the development of Marx's thought. But beyond cultural specificities, taking into account labor power provokes a considerable shift in the problematics of work. Work is no longer reduced to the mere economic dimension of selling and circulating goods, it is henceforth associated as well with the condition of the producer in a capitalist market system. The second contribution of the Marxist critique is the introduction of nonwork, a notion previously absent from political economy (Marx 1982, p. 208). Seen from the angle of the economic dependence of the worker, who can only meet his needs through the sale of his labor power, labor cannot be envisaged independently from nonwork and its different individual consequences. Marx thus contributed to a conceptualization of labor as dependent work, that is, seen from the producer's rather than the product side. This Marxist critique of political economy was to be, in the second half of the nineteenth century, a fertile ground for the development of the social sciences. Their particular contribution was, among others, to objectify nonwork by breaking it down into categories of social organization such as old age, disability, illness, unemployment, etc.

1.2 Wage Labor: A New Expression of Dependent Work

The concept of work mobilized in this endeavor for social objectivation meets the dual criteria of quantification and dependence. The reformist social sciences of the early twentieth century and the new categories they contributed to producing thus reinforced the process of dissociation between on the one hand work/*œuvre*, characterized by the independence of the producer and the quality of the product, and on the other hand abstract work associated with dependence and quantity. This social differentiation grew more marked up until World War II, accompanied at the same time by an increased hold of abstract labor over work to the point of assimilating Work with Labor, to the detriment of other possible meanings of the concept. This process went along with the spread of wage labor which was legally codified—in the form of a contract or through labor legislation—and statistically codified—in the form of socio-professional categories—as a response to standardization and mass production in economic activity.

The concept of employment, at the heart of the legal codification of labor, is informed by a dual meaning: from a macroeconomic standpoint it means trading a given quantity of work for a given salary on the market; from the political and social standpoint, it means exchanging individual subordination for security, that is, the worker's subordination to his employer and the security offered by labor legislation and social protection. From this standpoint, employment is not so much determined by the nature of economic activity as by the nature of the social relationships that frame this activity. Employment can thus be defined as socialized labor, insuring protection for the producer, in response to the Marxist critique. But this protection is highly variable geographically, reflecting the different avenues of society building around the integrative virtues of work (Esping-Anderson 1990). As Biernacki (1995) has shown in the case of England and Germany, the historical forms of economic objectivation of labor specific to each country—as a product

for the English, as labor power for the Germans—affect the way in which the concept of labor has been mobilized in the production of the political and social order.

2. Labor and the Constitution of the Social World

Labor has become the nerve center of the social world. As participation in the common good, it was used at the end of the nineteenth century in the definition of a link that could bind together individuals in the broadened framework of the nation-state. Consequently, it was codified as a vector of individual belonging to the collective, thereby insuring the group's cohesion. But although the social world reflects the actual organization of the group, it is also the fruit of competing endeavors at formalization and cannot be apprehended independently from the discourse and knowledge that institute it. Social science discourse as it was institutionalized at the turn of the century is in this respect exemplary. It played an essential role in indexing the social world on labor and in defining labor's relationships to the public concern (Durkheim 1964, Weber 1930). During the first half of the twentieth century, this scholarly activity focused on the production of general categories to organize the industrial society on the level of the nation-state. In other words, it was an activity mainly aimed at stating principles of homogenization and generalization, as if the transformations of the late nineteenth century had at the same time generated social sciences with strong holistic orientations and the organization of society on a large scale, on the basis of principles of standardization referring to labor within the framework of national institutions.

2.1 Labor as a Status

The close connection between work and social status is one of the main features of the concept of labor. This type of connection doubtless already existed in feudal societies and guild organizations in cities, for instance, but its specificity lies in its relation, since the late nineteenth century, to demands of social justice, political equality, and democracy. Labor has thus become the democratic means of access to social status and political integration, a function that for centuries was ensured by property and heredity. As a provider of income and the principle on which the redistribution of wealth is based, labor has become the central social relationship in capitalist societies. A powerful determining factor of each person's position in the social hierarchy, it helps produce individual identities and shapes the ways in which citizenship is exercised. This link between work and politics, in the sense of organization and production of the society on the whole, was not, however, a foregone conclusion even two centuries ago (Arendt 1958). But it has become such a natural fact within the European nation-state that it today seems to be taken for granted historically (Zimmermann et al. 1999).

2.2 Labor, Class and Nation

The socialization of work and nonwork, based on the critique of a strictly economic conception of labor, has been held up to be the promise of the nation-state's order and political legitimacy. This conjunction between providing security for wage-labor—through social protection and labor legislation—and consolidating the nation-state has led to a considerable growth of the social value attached to paid work. The established expression 'social security' suggests to what extent the socialization of nonwork is as much a question of collective as of individual security. This collective security is based on the capital/wage labor dichotomy, in a framework where the social regulation of labor is a partner of economic regulation of the market more than a substitute for it. The encounter of the nation-state, economic organization and labor institutions under the auspices of the welfare state thus appears as the fruit of a historical compromise between capital and labor. This compromise, at the basis of Fordism, is founded on the norm of full employment taken to be a constituent principle of the social world. From such a perspective, the social world can be seen as a collective order produced on the basis of wage labor and codified by a set of rights and duties attached to the salaried class (Castel 1995).

The history of this collective order is, from 1890 to 1960, one of establishing a coherence between individual identities and economic and political practices on the basis of the principles of class and nation (Wagner 1994). It is a movement of widespread standardization aiming to reduce uncertainty as well as the variety of possible actions and interpretations. Statistical and legal methods were the preferred instruments of this standardization and establishment of equivalences. Through the institution of nomenclatures, they gave shape to a national body made homogenous that can be observed and planned thanks to the objective nature of figures. By granting social and political rights guaranteed by the state, they stipulate individual belonging to this body. But today these standardization principles are put to the test by the demands of flexibility and the multiplication of work statuses.

2.3 Transformations in Work and Reconfiguration of the Social Sciences

The quest for generality a century ago has been replaced by the quest for diversity. This holds true for economic and social life as well as for the field of social

sciences. Over the past few years, the issue has been one of producing new equivalencies between situations and persons, able to place increasingly individualizing practices into a collective order that guarantees equity. The demand for flexibility provides an exemplary illustration. Whereas for the employer it is a powerful tool for dealing with the ups and downs of the economy, for the employee it introduces an uncertainty that the rhetoric of risk and predictability at the basis of social protection (Evers and Nowotny 1987) can no longer necessarily control. One of the current challenges is thus to make individual careers secure in the new context of uncertainty that weighs on work (Supiot 1999). But the categories established on the norm of full employment and the model of society that goes with it are having difficulty integrating these imperatives for diversity and flexibility. It is more generally the reduction of the concept of work to that of labor that should be incriminated, to such an extent that an examination of the ongoing transformations entails a need to rediscover the diversity of meanings of work.

In 1958, Arendt undertook a radical critique of this reduction that led to labor invading the political sphere, eventually crushing it in favor of the 'social.' Against the institution of labor as a general category of human action, Arendt reminds us that it is merely one aspect among others of *Vita activa*. She emphasized the major historical differentiation between:

(a) labor, an activity of *anima laborans* governed by the human race's needs of subsistence and reproduction, producing ephemeral things destined to be consumed;

(b) work, an activity of *homo faber*, producing lasting objects aimed at stabilizing human life;

(c) action, *praxis*, which produces things that are inseparable from the actor and which only occur through interaction. Action thus constitutes the eminently public and political aspect of *Vita activa*.

Whereas Arendt deplored the progressive absorption first of work, then of action, by the reductionist concept of labor, the topicality of the question of the relationship between work and action has since been renewed.

Along the lines of the revival of action theories in the social sciences, the concept of work has opened up toward action. Until the 1960s, continental European sociology of work was mainly concerned with the status conferred on individuals by their jobs. Today it has developed the issue of work as action, mobilizing a broader signification of the concept of work. The paradigms of action lead to taking into account a greater diversity of meanings of work (women's domestic work, voluntary and civic work, for instance). They also lead to envisaging work from the perspective of the interactions and conventions that preside over coordination among agents ('L'économie des conventions' 1989). Such an approach indicates a return to work in itself, to work as an action that refers to particular people and situations (Boltanski and Thévenot 1989). After the major movement of standardization that characterized the history of the concept of labor and sealed its hegemony, the concept of work today is being redesigned to integrate the quality of products and people's capacities, and has opened up to the plurality and diversity of configurations of action (Salais and Storper 1997, Boltanski and Chapiello 1999).

By the same token, the uncertainties that weigh on national employment institutions today, under the pressure of globalization and the constitution of supranational political entities (Habermas 1999), have found considerable resonance in debates on postmodernity that run through the social sciences. After the supremacy of the national level, today new scales of regulation and organization of the collective sphere are being explored, both from a practical and theoretical standpoint. Here again, the connection between economic and political transformations and paradigmatic changes in the social sciences is a close one. To such an extent that, where some see the end of work (Rifkin 1995) or the nation-state approaching, others announce both the end of society and the social sciences. But far from this alarmist diagnosis, today seems rather to be an era of revising both the categories of organization of society built around work and the nation, and their principles of intelligibility as they were defined by the social sciences.

See also: Alienation, Sociology of; Arendt, Hannah (1906–75); Capitalism; Class Consciousness; Class: Social; Differentiation: Social; Durkheim, Emile (1858–1917); History and Memory; Labor History; Labor Movements and Gender; Labor Unions; Marx, Karl (1818–89); Smith, Adam (1723–90); Social Class and Gender; Social Security; Weber, Max (1864–1920); Work: Anthropological Aspects; Work, History of; Work, Sociology of; Working Classes, History of

Bibliography

Arendt A 1958 *The Human Condition*. The University of Chicago Press, Chicago

Biernacki R 1995 *The Fabrication of Labor. Germany and Britain, 1640–1914*. University of California Press, Berkeley, CA

Boltanski L, Chiapello E 1999 *Le nouvel esprit du capitalisme*. Gallimard, Paris

Boltanski L, Thévenot L 1989 *Justesse et justice dans le travail*. Presses Universitaires de France, Paris

Castel R 1995 *Les métamorphoses de la question sociale. Une chronique du salariat*. Fayard, Paris

Cottereau A 1994 Théories de l'action et notion de travail. Note sur quelques difficultés et quelques perspectives. [Theories of action and the notion of work – Difficulties and prospects]. *Sociologie du Travail* Sp.Iss.SI **36**: 73–89

Durkheim E 1964 *Division of Labor in Society*. Free Press of Glencoe, New York

L'économie des conventions 1989 *Revue Economique* **40** (2): Special issue
Esping-Anderson G 1990 *Three Worlds of Welfare Capitalism*. Polity Press, Cambridge, UK
Evers A, Nowotny H 1987 *Über den Umgang mit Unsicherheit. Die Entdeckung der Gestaltbarkeit von Gesellschaften*. Suhrkamp, Frankfurt/Main
Habermas J 1999 Der europäische Nationalstaat unter dem Druk der Globalisierung. *Blätter für Deutsche und Internationale Politik* **4**: 425–36
Hegel G W F 1977/1805 *Phenomenology of Spirit*. Clarendon Press, Oxford, UK
Kocka J, Offe C 2000 *Geschichte und Zukunft der Arbeit*. Campus, Frankfurt/New York
Mauss M 1966/1950 *Sociologie et anthropologie*. Presses universitaires de France, Paris
Marx K 1968/1867 Das Kapital. Kritik der politischen Ökonomie. In: Marx K, Engels F *Werke*. Dietz, Berlin, Vol. 23
Marx K 1982 Oekonomisch-philosophische Manuskripte. In: Marx K, Engels F *Gesamtausgabe* Dietz, Berlin, Vol. 2
Méda D 1996 New perspectives on work as value. *International Labor Review* **135**: 633–43
Rifkin J 1995 *The End of Work: The Decline of the Global Labor Force and the Dawn of the Post-Market Era*. G P Putnam's Sons, New York
Salais R, Storper M 1997/1993 *Worlds of Production. The Action Frameworks of the Economy*. Harvard University Press, Cambridge, MA
Smith A 1776 *An Inquiry into the Nature and Causes of the Wealth of Nations*. Whiteston, Dublin, Ireland
Supiot A (ed.) 1999 *Au-delà de l'emploi. Transformations du travail et devenir du droit du travail en Europe*. Rapport pour la Commission Européenne. Flammarion, Paris
Wagner P 1994 *A Sociology of Modernity. Liberty and Discipline*. Routledge, London/New York
Weber M 1930 *Protestant Ethic and the Spirit of Capitalism*. G Allen and Unwin, Ltd., London
Zimmermann B, Didry C, Wagner P (eds.) 1999 *Le travail et la nation. Histoire croisée de la France et de l'Allemagne*. Maison des Sciences de l'Homme, Paris

B. Zimmermann

Work: Anthropological Aspects

Work is defined as a continuous human activity aimed at producing goods and services. It is a phenomenon that exists in all human societies. In contrast to sociologists, anthropologists primarily have studied work in nonindustrial/noncapitalist societies (hunters and gatherers, cultivators, pastoralists, craftsmen). Unlike historians, they are interested mainly in contemporary societies. Anthropologists recently have started studying work in capitalist enterprises, and the transformation from noncapitalist to capitalist work. Key concerns of the anthropology of work include the interaction between humans and nature, tool use, and the institutional and cultural embeddedness of work. In these contexts, work is treated as part of a system. In another perspective, work is studied as action and performance. The anthropology of work has been delineated by Applebaum (1992), Gamst (1995), and Wallman (1979).

1. *The Origins of the Anthropology of Work*

Many leading scholars became interested in work in the wake of European industrialization. Their main focus was industrial work, but they made comparative reference to pre-industrial work. Marxpraised the high productivity and discipline of capitalist industrial work, but criticized the alienation it caused. Pre-capitalist work was less productive (e.g., the work of craftsmen); nonetheless it had aesthetic character. In a comparative study on religion in Europe, India, and China, Weberwrote about the development of the work ethic and the rationality of work.

Bücher (1893) studied the division of labor and cooperation, and formulated a theory of 'independent household economy,' in which work is subordinate to the autarchy of the household. Bücher (1896) turned his attention to performance: How was it possible to perform constant, strenuous work in pre-industrial societies? In his view, the answer lay in the aesthetics and rhythms of work. In contrast to Marx and Weber, Bücher also drew upon contemporary ethnographic literature, placing him at the forefront of the anthropology of work. He did not, however, conduct any fieldwork personally.

Malinowski, the founder of modern anthropological fieldwork, was a student of Bücher's in Leipzig. Bücher and other German scholars (including Hahn and Weber) inspired his interest in the topic of work, which became the primary focus of his research on gardening practices on the Trobriands (Malinowski 1935). On the one hand, he was interested in the issue of performance: How do the Trobrianders succeed in performing sustained, strenuous work? His findings conclude that generally they were motivated by an interest in food, but were also interested in the aesthetics and social organization of gardening. Malinowski also examined the relationship between work and magic. Unlike Weber, he did not consider magic a hindrance to successful gardening: Magic is compatible with rational work, helping people to cope with the vicissitudes of gardening.

Two of Malinowski's students, Firthand Richards, continued his line of study. In her classic work on the Bemba in Northern Rhodesia, Richards (1939) stressed more than Malinowski the separation of work and needs: Work is not related merely to food requirements, but is embedded in cultural traditions, in this case the warrior past of the Bemba.

Based on intensive fieldwork, the studies produced by Malinowski, Firth, and Richards are still worth reading, mainly due to the diversity of issues and abundance of material. At a theoretical level they were

functionalists, taking into account economic, ecological, institutional, and cultural factors. Subsequent studies on work, however, belonged to more specialized subdisciplines of anthropology.

2. Theoretical Approaches

2.1 Ecological and Technological Approaches

Ecological anthropology investigates work as interaction between man and nature. The studies focus on hunters and gatherers, cultivators, and pastoralists. Quantitative measurements on energy input and output are characteristic of this approach. Lee (1979) investigated the hunting and gathering activities of the Kung San of the Kalahari (Botswana). One interesting finding was that the women—for less work—yielded more calories collecting nuts and plants than did the male hunters.

The Kung spent an average of 17 hours per week (i.e., 880 hours per year) hunting and gathering. The average calorific value of the food yield was 2,355 calories, i.e., it exceeded the daily requirements of the Kung, who are of relatively short stature. These findings triggered off a series of debates in anthropology and other fields. Based on the findings of Lee and other researchers, Sahlins (1972) argued that hunters and gatherers led a life of affluence consisting of little work. Other anthropologists challenged the thesis, provoking a debate that has yet to be resolved.

As a rule, farmers work more hours than hunters and gatherers. This is particularly true of intensive agriculture (Netting 1993). The annual person-hours in shifting agriculture is 480–960, but it is 3–4 times higher in intensive agriculture. This type of agriculture generally is practiced on small, permanently cultivated areas, often artificially irrigated and situated in high-density population areas. The work requires long working hours, and specialized knowledge and skills.

From an evolutionist perspective, many scholars view intensive agriculture as primitive in comparison to the mechanized agriculture practiced in industrial countries, where labor productivity is higher. Netting, however, argues that labor-intensive agriculture is more suited to densely populated regions. It is also more sustainable. In the USA, the energy input–output ratio for rice production is 1:1.55, while in the Philippines it is 1:7.

Herdsmen are often described as lazy in comparison to farmers. This verdict is based on the premise that only manual activities involving tool use constitute work. But in his study on camel herding and goat herding, Spittler (1998) shows that long working hours, complex knowledge and skills, and the strain of a harsh environment (wilderness) are all part of their work.

Technological approaches in studies on work concentrate on tool use. Craftsmen are the main focus of these studies, not hunters and gatherers, or farmers and pastoralists. In anthropology, the study of tools is a traditional one. This is due partly to the accessibility of collections in ethnological museums, and partly to theoretical paradigms. For many years, tool production was seen as the essential human characteristic. Technical and social evolution was also linked to tool development.

In effect, this fixation on tools was a hindrance to research on work. The fact that many types of work require few or no tools was overlooked. Above all, too little notice was taken of the fact that the starting point for any study on work should be human action, not tools. This is especially true of preindustrial tools, the use of which requires considerable skill, strength, and stamina. Bücher (1896) identified this fact, but found no following at first. Later, two Hungarian folklorists, Fél and Hofer (1974), wrote an excellent analysis of work with agricultural tools. Their detailed accounts of working with hoes, scythes, whips, etc. show that such work entails a number of factors (family traditions of work, general work ethic, emic theory of the body and of tool use, and the social life of tools) which cannot be deduced from the study of the tools themselves.

2.2 Institutional Approaches

With the one important exception of housework, work in industrial societies generally is organized in capitalist or public companies. One major feature here is the recruitment of free labor. There exist other types of work institutions. Slavery is a historically widespread form of organizing agricultural and industrial labor, e.g., in ancient Greece and Rome, in precolonial Africa, and in colonial America. Hunters and gatherers often work in loosely organized bands. The most common form of work organization among farmers, herdsmen, and craftsmen is the domestic economy, also known as the domestic mode of production or family economy. The concept was introduced by Bücher (1893), developed by Chayanov (1923), and established by Sahlins (1972) in the field of anthropology.

Domestic economies produce either for private consumption or for the market. The aim is not primarily profit, but to provide for the household. In the pure family economy studied by Chayanov, labor is recruited exclusively from the ranks of the nuclear family. A domestic economy involving additional labor (relatives, servants, slaves) can grow to a considerable size, comprising 100 or more workers (e.g., the ancient *Oikos*).

For a long time, the family was seen as a unit in which decisions were made in the interests of the entire family, a claim that is being challenged increasingly. Gender studies show that the burden of work is often distributed unequally, with women performing the greater share. The same applies to the age-related

division of labor, which is not always decided on the basis of natural abilities. Children are sometimes exempted from work, but more often are forced to work particularly hard. Both work and the fruits of work are often distributed unequally within the family, with elderly men and women enjoying the most privilege.

The term domestic or family economy is misleading, since a household, or family, is not exclusively—or even primarily—geared toward production, performing a range of other tasks as well. The members of a household are usually provided for, both materially and psychologically, irrespective of their work output. Children are reared and marriages arranged. The members of the household cohabit. Due to the diversity of tasks involved, work cannot be separated from other activities as it is in capitalist enterprises. On the contrary, it remains embedded in the multifunctional institution of the household.

In the purely family economy, the number and combination of the workers is determined by biological reproduction within the family. It depends on the family cycle and demographic contingencies. The output of individual workers fluctuates considerably, depending on the number of consumers. Such imbalances can be rectified by a temporary or permanent increase or reduction in the number of workers (e.g., by employing servants or farming out family members as servants). Greater flexibility is also achieved if households work together on a cooperative basis. Since Bücher, anthropologists frequently have studied the formation of work groups extending beyond the individual household. Due to internal competition, such work groups often lead to increased productivity.

In family economies, the division of labor is mainly on the basis of gender and age. This limited division of labor leads to a lack of specialization, since each member has to accomplish a number of tasks. The tasks are learnt at an early age in an informal process of teaching and learning. Work discipline is enforced by parental authority, but also through specific family sanctions (e.g., inheritance law).

What is the level of work efficiency in domestic economies? Inadequate specialization and restricted sanctions are often identified as negative factors. Family workers cannot be recruited or fired. On the other hand, family solidarity and early training can improve work performance. In areas where a large number of workers is no advantage, e.g., in agriculture, domestic economy proves competitive even under capitalist conditions (Chayanov 1923, Netting 1993).

In contrast to domestic economy, apprenticeship prepares people for a specialized activity. Apprenticeships occur mainly in the crafts (e.g., smiths, potters, weavers). Through the apprenticeship itself and the voluntary enrollment of apprentices, work is rationalized more than in domestic economies (Coy 1989). Similar to domestic economy, trial and error, demonstrations, and repetition are more important in acquiring skills than written instructions, explanations, and questions (Keller and Keller 1996).

From a methodological point of view, this implies that work can only be researched partially on the basis of interview. Precise observation and participation are required. These findings apply to both the study of work and to other areas. They challenge the predominance of the linguistic paradigm and the interview as its methodological correlate.

2.3 Cultural Approaches

In the 1970s and 1980s, a cultural (albeit not entirely new) perspective began to appear in anthropological studies of work (Wallman 1979, Godelier 1984, Applebaum 1992). Weber's analyses of the religious foundations of rationality and the work ethic were an excellent precedent. Now, however, new problems emerged and the approach was radicalized.

Semantic analyses of the definition of work in various cultures show that the meaning of any linguistic equivalent to 'work' can be expressed in broad or narrow terms. In most societies, the most arduous subsistence activities are defined by this one word. However, work is not always confined to useful activities, but can include what is good and beautiful. By contrast, intellectual work is rarely placed in the same category as physical work.

Cultural analysis goes beyond semantic analysis and the specification of taxonomies. It investigates models or commonsense theories of work. For example, work is seen variously as the transformation of objects (the prevailing view in industrial societies), as fighting with nature, or as caring for plants and animals. Commonsense theories also express claims about the part that work (in comparison to chance, magic, divine influences) plays in the outcome of action, and how human work capacity (physical strength, knowledge, skills) is created. Cultural analyses also re-explore the question raised by Marx: What are the implications of work for human identity?

Concepts of work differ from culture to culture, but also change over the course of history. However, many analyses of the history of the concept of work since classical antiquity do not address the question of whether the philosophers' concepts were shared by the working population. In contrast, Applebaum (1992) includes in his study contradictory interpretations of work between philosophers and working people.

Ecological, technological, institutional, and cultural approaches are mutually complementary. Although materialistic (ecological, technological) and idealistic (cultural) approaches were once opposed sharply, they have been combined fruitfully in several recent studies. Descola (1986), for example, in his study about the Achuar Indians in Ecuador, combines the structuralist theory of Lévi-Strauss with ecological approaches.

3. New Directions and Future Research

3.1 Work in Capitalist Societies

Since the 1980s, anthropological research on work has shifted increasingly to capitalist societies. The *Anthropology of Work Review*, founded in 1980, focuses almost exclusively on work in capitalist societies. What is the anthropological content of these studies? What distinguishes them from technological, sociological, or psychological work studies? Anthropology focuses on culture. Work culture and the culture of organizations are, therefore, one of the main areas of study. As a result, the anthropological definition of culture is often modified considerably, especially in the case of corporate culture. The methodological approach of these studies is even more important than their cultural perspective. Most anthropological studies are based on participant observation. Many researchers complete an apprenticeship, and often practice their acquired trade for several years. As a result, they practice participant observation far more radically than most classical anthropologists.

Many domains of work in modern societies interest anthropologists. Some concentrate on craftsmen, since their trades are affiliated most closely to classical anthropology. In a study on blacksmiths based on personal job experience, Keller and Keller (1996) showed how closely learning and working practice are based on nonverbal cognitive influences and experiences.

Anthropologists' traditional predilection for exotic themes is confirmed by their interest in kinds of work with a romantic flair (cowboys, and more recently, locomotive engineers, truck drivers) or fringe appeal (hustlers). Recently, however, they have turned their attention to executive management. Industrial work and its transformation is another topic of interest, and some anthropologists spend many years in industrial employment (Burawoy 1979). In recent years, the modern service and information sector has attracted major interest, including, for example, fast food outlets, computer firms, repair services, and science laboratories.

Despite the diversity of occupations and different theoretical approaches, these studies have a lot in common due to their uniform methodological approach. Anthropologists proceed from the actor's perspective, not from the perspective of economic, technical, or cultural systems. They show that an actor's skills are not limited to the specifications of a job description. They also include physical strength, manual skills, knowledge, stamina, energy, and learning ability. In most jobs, more is learned through acquired experience and discussions with colleagues than through formal training.

Performance at work is ensured only partially by checks, incentives, and sanctions. Other motives also play a role, although not necessarily a specific work ethic. In the minimum scenario, people work simply because it is preferable to 'killing time' over a long day. All the investigations show that consistent work performance possesses an inherent structure that is not governed by checks or sanctions. Furthermore, the work world is never determined exclusively by work; it is a life world, providing scope for personal relationships, emotions such as joy and frustration, and a range of other external influences.

3.2 Work in Comparative Perspective

The classical writers of the nineteenth century (Marx, Bücher, Weber) assumed that pure work first evolved in modern times as a result of capitalist wage labor or a process of rationalization. Still today, most social scientists would agree that precapitalist work is embedded in society and culture, and combined with other activities. They also argue that there is more freedom, because there is less rigid control.

However, if we review the research findings of the 100-year history of the anthropology of work, we tend to arrive at the opposite conclusion. Precapitalist, preindustrial work seems determined largely by ecological and technical factors. The actors are merely peripheral. By comparison, the studies of work in capitalist society reveal autonomous actors who not only perform their work but also display emotions, cultivate personal relationships, and see their work as a life world.

Are the old theories wrong? The new research findings partially correct the dichotomously constructed theories on the antagonism between precapitalist and capitalist work. To a large extent, however, these results are research artifacts. Most anthropologists working in capitalist societies accept the structural conditions of work as given, and concentrate on the scope for individual action (but, see, Burawoy 1979). This view is endorsed by the methods of participant observation. The economic and technical framework conditions are already known or being studied by other sciences.

By contrast, the anthropologist, whose research centers around peasants, herdsmen, or foragers, has to investigate all aspects of their work. This includes ecological, technical, and institutional structures. Anthropologists do not measure working time in capitalist societies, since the data are usually otherwise available. In precapitalist societies, however, time-consuming methods are often required for its measurement. On the other hand, the actor's point of view is rarely examined (but see Malinowski 1935, Fél and Hofer 1974, Spittler 1998). The method is also important in this context. Participant observation, which favours the actor's perspective, is practiced more in capitalist than in noncapitalist societies.

The conclusion is that comparison is difficult, since the investigations are based on different presup-

positions. In noncapitalist societies, anthropologists should investigate work performance and the work world far more closely than in the past; in the case of Western society, they should concentrate more on structural conditions, including the latest developments in the international division of labor. Although the social division of labor was once a central theme of the social sciences (Marx, Bücher, Durkheim), anthropologists have largely ignored it, except for the gender-specific division of labor.

Other major comparative research deals with processes of transformation from noncapitalist to capitalist societies. Extensive anthropological literature is now available on the transformation of work in developing countries and Eastern Europe. These transformation processes show that the labor process in capitalist institutions differs considerably from that of noncapitalist institutions. Conflicts about time schedules, labor recruitment, and labor relations not only express the different interests of workers and management but are indicators of a conflict between work cultures. Therefore, we may expect that the transformation process will not result in the enforcement of a pure capitalist work culture but in a hybrid type of labor process.

See also: Economic Anthropology; Economic Sociology; Hunter–Gatherer Societies, Archaeology of; Hunting and Gathering Societies in Anthropology; Labor History; Labor, Division of; Work and Labor: History of the Concept; Work, History of; Work, Sociology of

Bibliography

Applebaum H 1992 *The Concept of Work. Ancient, Medieval and Modern*. State University of New York Press, Albany, NY

Bücher K 1893 *Die Entstehung der Volkswirtschaft*. Verlag der H. Laupp'schen Buchhandlung, Tübingen (1901 *Industrial Evolution*. Holt, New York)

Bücher K 1896 *Arbeit und Rhythmus*. Reinicke, Leipzig, Germany

Burawoy M 1979 The anthropology of industrial work. *Annual Review of Anthropology* **8**: 231–66

Chayanov A V 1923 *Die Lehre von der bäuerlichen Wirtschaft*. Parey, Berlin (1966 *The Theory of Peasant Economy*. Irwin, Homewood, IL)

Coy M W (ed.) 1989 *Apprenticeship: From Theory to Method and Back Again*. State University of New York Press, Albany, NY

Descola P 1986 *La nature domestique. Symbolisme et praxis dans l'écologie des Achouar*. Editions de la maison des sciences de l'homme, Paris (1994 *In the Society of Nature. A Native Ecology in Amazonia*. Cambridge University Press, Cambridge, UK)

Fél E, Hofer T 1974 *Geräte der Átányer Bauern*. Hos, Copenhagen, Denmark

Gamst F (ed.) 1995 *Meanings of Work*. State University of New York Press, Albany, NY

Godelier M 1984 *L'idéel et le matériel. Pensée, économies, sociétés*. Fayard, Paris (1986 *The Mental and the Material*. Verso, London)

Keller C M, Keller J D 1996 *Cognition and Tool Use: The Blacksmith at Work*. Cambridge University Press, Cambridge, UK

Lee R B 1979 *The Kung San. Men, Women, and Work in a Foraging Society*. Cambridge University Press, Cambridge, UK

Malinowski B 1935 *Coral Gardens and their Magic*. George Allen & Unwin, London, 2 Vols.

Netting R McC 1993 *Smallholders, Householders. Farm Families and the Ecology of Intensive, Sustainable Agriculture*. Stanford University Press, Stanford, CA

Richards A I 1939 *Land, Labor and Diet in Northern Rhodesia*. Oxford University Press, London

Sahlins M 1972 *Stone Age Economics*. Aldine-Atherton, Chicago

Spittler G 1998 *Hirtenarbeit. Die Welt der Kamelhirten und Ziegenhirtinnen von Timia*. Koeppe Verlag, Cologne, Germany

Wallman S (ed.) 1979 *Social Anthropology of Work*. Academic Press, London

G. Spittler

Work, History of

The concept of work designates certain human activities; however, those activities that are considered to be work and those that are not, vary among different cultures and epochs. Each historical society has its own definition of work, and a variety of different conceptions are juxtaposed to one another within each society. How work is defined is closely associated with the value placed on it, and this in turn is connected with workers' positions in society. The history of work must therefore bring together two themes: the history of the meaning of work and the history of work as a practice. The relation between gender and work cuts across these two themes and shows how strongly they are connected.

1. History of the Concept of Work

The state of research on the history of the concept of work and the meaning of work suffers from three shortcomings. First, it is based upon the reflections of intellectual elites. The ways in which other social groups—above all, the workers themselves—understood and evaluated work have not left much of a trace in historical source material, nor has much research been done on this tissue. The history of the concept of work is a history of ideas, and—in a very few approaches to this field—a history of mentalities as well (Le Goff 1980). Second, research has tended to ascribe a unitary—or, at least, a predominant—concept of work to the respective historical society under investigation. Contradictory or conflicting attitudes have been almost neglected. Third, the history of

the concept of work has remained eurocentric. Despite the pronounced interest in the work ethic of non-European cultures and religions that was already evident in the work of Max Weber, and in spite of the growing interest in global history, a systematic intercultural comparison remains a desideratum (Godelier 1980).

1.1 Concepts of Work in Pre-industrial Societies

The history of the concept of work in Western historiography begins as a rule with the great philosophers of Greek Antiquity (Aristotle, Plato) (Conze 1972). They employed the term work (*ponos, poiesis*) to designate work done manually, and described thereby activities that required physical exertion and strength, that presupposed the use of tools, and the aim of which was to turn out a particular product. The meanings which coalesced to define the word *ponos* included labor, toil, hardship, and suffering. This concept of work had a parallel in the Indo-Germanic languages, which as a rule contain several terms for work and working: *labor* and *opus* in Latin, *travailler* and *ouvrier* in French, *labor* and *work* in English. The first word stem means great effort, travail, and affliction, activities dictated by necessity or constraint. The second word stem refers to the results of this effort—the attainment of a certain goal, the creation of a product.

Most historians agree that philosophical and political thought in Antiquity was characterized by contempt for work. Work was an evil, the activity of slaves, while the élite engaged in art, philosophy, and politics. The connection between abhorrence for manual labor and slavery is confirmed by the contrast provided by non-European cultures. Slave labor did not play an important role in the historical societies of East Asia, in whose great religious-philosophical systems—foremost among them Buddhism and Confucianism—work (the fieldwork of peasants and the handiwork of artisans) had a positive connotation. But European Antiquity was also characterized by contradictory attitudes toward work. Critical schools of thinking such as the Cynics esteemed physical and mental labor and contrasted them to idleness, which they regarded as a vice. In ancient Rome as well, particularly during the Republic, peasant farm work performed by free men was not seen in a totally negative light.

In the Judeo-Christian tradition, the ambivalence of the concept of work is expressed with particular clarity. In the Old Testament God himself appears as a worker, and man was originally meant to care for God's creation. 'And the Lord God took the man, and put him into the garden of Eden to dress it and to keep it' (Genesis 2:15). However, with man's fall from grace, work became penance and punishment: 'In the sweat of thy face shalt thou eat bread.' (Genesis 3:19). The New Testament carried forth this ambivalence. Christian writings of the Middle Ages constitute an 'ideological arsenal containing weapons for the defence of every position, in favour both of labour and nonlabour (…) ranging from (…) contempt and condemnation to a tendency to respectful evaluation' (Le Goff 1980, p. 77). From the tenth century on, a new trifunctional or tripartite social schema took hold; it divided society into 'priests' (*oratores*), 'warriors' (*bellatores*), and 'workers' (*laboratores*). This model assigned all working people to the lowest level of the social hierarchy, but nevertheless made them an indispensable segment of society. The Reformation strengthened the tendencies to enhance the status of work. Luther and Calvin understood work as serving God. The work that each Christian performed was considered a 'calling' or a 'divine calling,' and thus assumed a quasi-religious character. The adherents of certain Protestant denominations—above all, the Puritans—regarded success they achieved in their work as confirmation that they numbered among God's elect. Since the publication of Max Weber's study (1904–5) it is generally accepted that the higher valuation of work during the Reformation legitimated and promoted the spread of a bourgeois work ethic, even if it has important roots in medieval thought.

1.2 Concepts of Work in Modern Western Society

In early modern Europe, work was redefined in two respects. The first definition exhibits anthropological traits. Work came to appear as active human intervention in nature for the purpose of assuring the ongoing existence of the human species. At the same time, work appeared to be a social process based upon collaboration among human beings. This new idea of work played a central role in Enlightenment philosophy and the writing of cultural history done during this period. Man was seen as ruling over nature; division of labor and cooperation as the foundation of human socialization; and increasing domination of nature and further division of labor as the very epitome of social progress. Tools were considered the basis upon which human dominion over nature rested. Man thus appeared to be the 'tool-making animal,' as Benjamin Franklin put it. In the nineteenth century, Karl Marx and other thinkers of the early Industrial Age carried on this intellectual tradition. In his classic formulations in Volume I of *Das Kapital* Marx described this dual character of work: 'Labor is, in the first place, a process in which both man and Nature participate, and in which man of his own accord starts, regulates and controls the material re-actions between himself and Nature (…) By thus acting on the external world and changing it, he at the same time changes his own nature' (Marx 1867, p.192).

The second redefinition originated in the seventeenth century in the field of political economy, which linked the concept of work with the idea of value. Work was defined as an activity that created value,

and value-producing activities were considered to be all those that were for pay or that generated income. This led to equating work with paid labor. In industrial capitalism, paid labor became the very epitome of work.

These new definitions were associated with the greatly enhanced status of work. 'The modern age has carried with it a theoretical glorification of labor and has resulted in a factual transformation of the whole society into a laboring society' (Arendt 1958, p. 4). Paid labor became the basis of self-esteem, social recognition, and political participation—in short, the admission ticket to civil society. However, attitudes toward work were ambivalent in modern society as well. The classic works of political economy of the eighteenth and nineteenth centuries, beginning with Adam Smith, feared the brutalizing and stultifying effects of extreme division of labor and repetitive mass production. For Marx, work in capitalism was exploitation and alienation; his communist utopia vacillated between self-realization by means of work and liberation from it. These ambivalencies were expressed in increasingly multifaceted views of work. The esteem accorded to work referred to a specified type of work in contrast to its polar opposite, which was assessed negatively; mental vs. physical work, planning and supervision vs. actually performing tasks, skilled vs. unskilled work, a steady job vs. 'casual labor,' 'proper' work vs. housework. A complex hierarchy of work is an essential component of the modern concept of work.

2. *Work as Practice*

Which activities were taken into consideration by scholars compiling the history of work as practice is closely connected to the development of the concept of work. Three perspectives emerge in the historiography of work: work as production and application of technology; work as economic activity; and work as culture.

2.1 *Work as Production*

The conception of work as dominion over nature that had emerged in the seventeenth century began to exert an influence on the history of work in the nineteenth century. Benjamin Franklin's characterization of man as the 'tool-making animal' was cited approvingly by Marx. Tools, or, in more general terminology, instruments of labor, were for Marx, fundamental to the periodization of history and a yardstick of human progress: 'It is not the articles made, but how they are made and by what instruments, that enables us to distinguish different economic epochs' (Marx 1867, p. 194). Proceeding on this basis, Marxian historians and archeologists developed a world historical perspective on work that focused on two (or three) 'revolutions in the forces of production' (Childe 1936, Bernal 1954).

The first of these was called the Neolithic Revolution. Its point of origin was the Near East (Southern Anatolia, Mesopotamia, the Jordan Valley) and between about 8000 BC and 3000 BC it spread over large portions of South, East, and South-east Asia, North Africa, Europe, and Central America. The essence of this revolution was the transition to agriculture and animal husbandry as the most important basis of human existence, and in the emergence of agrarian (farming) societies. According to this view, this was the beginning of goal-oriented human intervention in nature and was thus, in an alternate formulation, the beginning of work. 'Early men did not work at all in the true sense (…) real work, steady work, labor for one's livelihood, came into being when agriculture was invented' (Williams 1976, p. 335).

The second great revolution in the forces of production was, from this perspective, the Industrial Revolution, which began during the second half of the eighteenth century in Western Europe and encompassed large parts of the world over the course of the nineteenth and twentieth centuries. Now, the tools employed in performing work were no longer operated manually by the human hand, but rather by a machine that was driven by artificially produced energy. Thus, production became independent of the skillfulness, endurance, and physical strength of human beings. In capitalist fantasies, as in Andrew Ure's *Philosophy of Manufacturers*, 1835, machinery even functioned as the most perfect means to control and to discipline the workforce, as the final victory over the obstinate spirit of the workers. Factory, mass production and conveyor belt became the symbols of work in industrial society.

The increasingly important role played by microelectronics and the computer in industrial production since the 1960s appeared, in this perspective, as the third great revolution in the history of work. Now, human beings withdrew from the process of controlling production and turned over mental activities as well to a machine, the computer. Automation, or flexible communication between humans and computers—a 'man–machine dialog'—seems now to be the defining characteristic of a new world of work. The amount of human labor necessary for direct production is on the decline. The production process is increasingly becoming the activity of a 'collective worker' (Marx), in which technology and science play an important role.

This 'grand narrative' was subjected to critique and developed further by scholars working in the history of technology, a field that became over the course of the twentieth century an independent academic subdiscipline (cf. *History of Technology*). Its early practitioners were engineers and technicians with their own narrow understanding of technology, who were not interested in human agency, but in the interaction of

material objects. As early as the 1930s (Bloch 1967) though, economic, social and cultural historians were beginning to get involved in this field, and making the effort to conceive of technology as an integral part of society, and thus to conceive of work in an all-encompassing sense. In contrast to Marxist 'history of the forces of production,' they emphasized the numerous small technical innovations made in all historical epochs. An evolutionary perspective predominated in the history of technology. What the two approaches have in common is the tendency to represent the history of work as developing along a path rising in linear fashion, and as the history of progress.

2.2 Work as Economic Activity

At around 1900, with the newly awakened interest in economic and cultural history, there was also increased demand for a history of work, which then proceeded to emerge as a not-very-clearly delimited subdivision of general economic and social history of various different cultures and epochs. Since then, the history of work as an economic activity has been focusing on work as wage labor; on working hours, wages, and working conditions; and on the history of the modern work ethic and work discipline.

Wage labor was widespread in many pre-industrial societies. In Europe, it took on increased significance from the late Middle Ages. Over the course of the Early Modern Period, the population became increasingly dependent upon wage labor. In the economically most highly developed regions of Europe, the 'proletarization' of the populace was a pre-condition and not a consequence of the Industrial Revolution. In the industrialization process of the nineteenth century, wage labor in its 'pure' form ultimately established itself as the predominant type of labor relationship. Among the features that typically characterized it were the standardization and regulation of work, including a high intensity of labor, a regular working rhythm, and a highly pronounced working discipline; a high degree of specialization of work along occupational lines; a clear demarcation of time spent at work and time off (over the course of a day, a week, a year, and a lifetime); the separation of the workplace from home; and a complex system of institutional and legal regulation.

The developmentf wage labor, however, was not a linear process. Recent research has shown that diversity of labor relationships was characteristic of the period from the late Middle Ages to the eighteenth century. A growing number of men and women earned their daily bread from a combination of agriculture, crafts and trades, and commercial activities, and by combining self-employment with wage labor. A high degree of flexibility and increased readiness to undertake labor migration were preconditions called for by this mixed economy. Flexibility, however, was not only a voluntary practice, but was enforced by the instability of the economy, by necessity and misery. The French historian Fernand Braudel suggested that commercial capital and industrial capital, small workshops and large-scale enterprises, rural home industry and centralized manufacturing during this period should not be regarded as sequential phenomena, but rather as a pool of possibilities, as co-existing labor relationships and modes of production, among which men and women could choose and switch.

An ongoing point of discussion is the question of why men and women of the Early Modern Period increased the intensity of their work. Was this attributable to the exigencies of hardship in an increasingly polarized or proletarianized society, as mercantilist economists of the seventeenth century presumed? Was it the capitalist entrepreneurs who used shop regulations and penalties to turn up the pressure upon their employees? (Thompson 1967). Was the transformation of the world of work simply a part of a general disciplining process—including workhouses, prisons, and schools—to which men and women were forced to submit (Foucault), or to which they—as part of the process of civilization (Elias)—subjected themselves? Or was it the transformation of material culture—namely, the brave new world of consumer goods that expanded so rapidly in the Early Modern Period—that made working harder acceptable, or even desirable? Are the beginnings of consumer society thus also the historical roots of the laboring society? (de Vries 1994) (see *Consumption, History of*). Which role was played by the changing perception of time, by which—from the fourteenth century on—Europeans began to adapt to the abstract rhythm of mechanical clocks? No matter how these questions are answered, it is clear that the role of the Industrial Revolution and mechanized factory labor as factors giving rise to the modern world of work have been devalued, whereas more emphasis is now placed on the long history of capitalism.

The industrialization process, too, became a field of revisionism. As it is conceived now, it was not represented by the factory alone; handicrafts, skilled work done by hand, and small-scale enterprises continued to play a key role (Sabel and Zeitlin 1997, Samuel 1977). The diverse branches of the service sector, such as transportation and shipping, retailing, education, entertainment, and the widely varied world of office work would very soon be employing more men and women than the manufacturing sector. In most industrial states, a majority of the workforce has never been employed in the secondary (industrial) sector. And ultimately, factory work has proved to be more varied than it had long appeared to outside observers.

The concentration of research on the 'ideal type' of paid labor in modern society has caused too little attention to be paid to forms of labor that deviate from that model. This applies, for instance, to various

forms of unfree or forced labor. In European history since the Middle Ages, slave labor has not been widespread, but as a basis of the Atlantic economic system from the sixteenth to the nineteenth century, it was indeed an essential component in the emergence of industrial capitalism. However, forced labor was widespread in the totalitarian political systems of the twentieth century. Seen from a global perspective, 'free' and 'pure' wage labor has (still?) not become the predominant form of work. Today, both agriculture as well as a combination of different activities and forms of work—as was characteristic of European history during the Early Modern Period—provides work and a living to a majority of the world's population.

2.3 Work as Culture

Since the 1980s, a new way of looking at the history of work has become widely accepted. A growing number of historians consider work as culture. As far as the concrete practice of research is concerned, this means a micro-historical perspective focusing on the workplace itself. Particular attention has been paid to two types of workplaces: first, the small workshops of artisans in the Early Modern Period and in the nineteenth century, that represent, on one hand, 'traditional' work, and, on the other hand, their increasing integration into capitalist relationships of production (Kaplan and Koepp 1986); and secondly, the factories of industrial enterprises of the late nineteenth and early twentieth centuries, as symbols of the modern industrial world of work (Lüdtke 1993). The chief focus of these micro-historical studies is upon day-to-day working processes themselves and on social relationships in the workplace. They show the enormous diversity of practical labor and thus constitute a much-needed corrective to the generalizations yielded by the macro-historical perspective. Scholars investigating the cultural history of work have also attempted to reconstruct the meaning of work for, 'common, everyday people' by using 'ego documents' such as autobiographies and diaries; iconographic and pictorial sources; and oral-history interviews (cf. Passerini 1982).

There has also been great interest in the symbolic and ritual forms in which solidarities and conflicts are carried out and resolved on the workplace (cf. Darnton 1984). In investigating these phenomena, the micro-historical perspective also extends beyond the immediate confines of the workplace. Work gives rise to social relationships that retain their significance beyond the workplace, the term 'world of work' is used to connote this expanded meaning.

Often, work's significance becomes clearly evident only when it is absent. The implementation of the working norms of industrial capitalism also led to a redefinition of being out of work. Between 1880 and 1910, a new term describing involuntarily not working, became widespread: unemployment (*Arbeitslosigkeit, chômage*) (Topalov 1994). The first microanalysis of work and unemployment that became one of the classics of twentieth-century social science is the study *Die Arbeitslosen von Marienthal*. In this small Austrian factory town in 1932, the textile mill that dominated the village's economic life shut down, so that practically the entire population was thrown out of work overnight. As a result, the rhythms of everyday life became slower and slower, and the unemployed were able to summon less and less energy for activities that they had previously pursued enthusiastically during their free time (Jahoda et al. 1933).

This draws attention to the fact that leisure is, on one hand, the antithesis of labor, though on the other hand, it is also part and parcel of the world of work. In pre-industrial societies, there was no strict dividing line between working and not working. The peasants of the Occitanian village Montaillou, who in about 1300 were interrogated by envoys of the Inquisition about their way of life, interrupted their labors at the slightest opportunity—to chat with neighbors, to hang around the village square, to have some fun playing a game (LeRoy Ladurie 1975). It was only Industrial Society that drew a clear distinction between working and not working, thereby changing the character of not only the world of work but of leisure as well. The concept of leisure refers less and less to merely not working and rather more to an ensemble of activities pursued by specific classes at certain times, such as games, sports, watching TV, going to bars or symphony concerts. In this sense, leisure seems to some historians to be a cultural practice that was not 'invented' until the Early Modern Period (Burke 1995).

3. Work and Gender

The relation between work and gender is a field of scholarship in which the reciprocities between the concept and practice of work, between work as economic activity and as culture, can be seen with particular clarity. Early studies of women's work conducted up into the 1970s almost exclusively treated gainful employment on the part of the women. Initial interest was focused upon factory labor, and then increasingly on 'specifically feminine' occupations like maid, governess, midwife, teacher, secretary, and salesgirl. The aim of this research was to reintegrate women into the male-defined world of paid labor. Beginning in the 1970s, feminist historians increasingly criticized the prevailing modern concept of work. Housework became the subject of theoretical discussions and of empirical investigations. Increasingly, studies on the history of women's work focused both on gainful employment and domestic activities (Tilly and Scott 1978, Simonton 1998).

The issue of the division of labor between the sexes has become a much-favored field of interdisciplinary

research. Anthropologists, historians, and sociologists have shed a great deal of light on the variability of the division of labor. One and the same activity can be a typical man's job in one culture and woman's work in another. Which work is done by men and women does not depend on the 'nature' of their respective sex, but upon socially constructed gender roles. Recent theoretical approaches oppose biological or 'biosocial' as well as cultural determinism. Emphasis is in the variety of options taken in the sexual division of labor and the contesting definition and redefinition of norms.

In European history of the late Middle Ages and the Early Modern Period, the economic significance and the symbolic meaning of the respective activities assumed decisive explanatory power. One research approach emphasizes the exclusion of women from market-oriented activities in agriculture, crafts, and trades. Work that promised to yield profits, was learned and practiced according to formalized rules, coalesced into formalized occupations and had organized politically along occupational lines became the domain of men. Early 'classical' studies of women's work and some recent research proceed under the assumption that up to the beginning of the modern period women also took part in artisanal market production and were organized in guilds—either in exclusive women's guilds or mixed guilds (Wiesner 1989). According to this perspective, it was not until later that there occurred a gradual exclusion of women, be that as a consequence of 'male bonding,' the spread of capitalism, or the Industrial Revolution. Other historians have stressed the continuities of the gender-based division of labor since the twelfth century, which placed women in the domestic sphere or in low-skilled, low-status and low-paying occupations (Bennett 1988). Most historians emphasize the fluctuating boundaries between men's and women's work, which are continuously redefined and renegotiated (Honeyman and Goodman 1998).

Another research approach emphasizes the cooperation between men and women. In this perspective, the 'working couple' formed the economic backbone of the agrarian and the artisanal economy (Pahl 1984, Wunder 1992). This can be seen even more clearly in the 'family economy' of the lower classes. Women played a central role in efforts to exploit and combine as many of the available income opportunities as possible. The 'industrious revolution' which preceded the industrial one in the view of many historians, was based on the cooperation of men and women in the family economy (de Vries 1994). Conflict and cooperation are not necessarily antagonistic but rather complementary perspectives on the gender-specific division of labor in pre-industrial Europe.

With respect to the nineteenth century, the focus of research has been on the 'male breadwinner model' and the emergence of 'separate spheres,' which made paid labor the exclusive domain of the male and, conversely, explained domesticity and care of children as being appropriate to the female nature (Janssens 1997, Hausen 1981). This was an effective model, even if there are diverging assessments of its effective range and the extent to which it actually did determine practice. Indeed, it was a short-lived model. The proportion of women pursuing paid labor began to climb again in the 1960s, including highly qualified professional work and lifelong careers. The history of the division of labor between the sexes shows that 'the division of labor is an effect of the social hierarchy, not its cause.' (Godelier 1980, p. 170).

The feminist critique of the concept of work in modern society has considerably widened the boundaries of the history of work. Since generally applicable definitional criteria are not in sight, they have to be developed from case to case. 'In this widest sense' the English historian Penelope J. Corfield wrote, '"work" is whatever people consider as work (…) the boundaries between "work" and "non-work" are in practice very often obscure' (Corfield and Keene 1990, p. 208).

See also: Alienation: Psychosociological Tradition; Alienation, Sociology of; Artisans and Guilds, History of; Capitalism; Feminist Economics; Industrialization, Typologies and History of; Labor Unions; Labor, Division of; Sex Differences in Pay; Slavery as Social Institution; Work and Labor: History of the Concept; Work: Anthropological Aspects

Bibliography

Arendt H 1958 (2nd edn. 1998) *The Human Condition*. University of Chicago Press, Chicago

Bennett J M 1988 History that stands still: women's work in the European past. *Feminist Studies* **14**: 269–83

Bernal J D 1954 *Science in History*. Watts, London

Bloch M 1967 *Land and Work in Medieval Europe*. (Selected papers by Marc Bloch) Routledge and Kegan Paul, London

Burke P 1995 The invention of leisure in early modern Europe. *Past and Present* **146**: 136–50

Childe V G 1936 *Man Makes Himself*. Watts, London

Conze W 1972 Arbeit. In: Brunner O, Conze W, Koselleck R (eds.) *Geschichtliche Grundbegriffe. Lexikon zur politisch)-sozialen Sprache in Deutschland*, Bd. 1. Klett-Cotta, Stuttgart, Germany

Corfield P J, Keene D (eds.) 1990 *Work in Towns 850–1850*. Leicester University Press, Leicester, UK

Darnton R 1984 *The Great Cat Massacre and Other Episodes in French Cultural History*. Basic Books, New York

de Vries J 1994 The industrial revolution and the industrious revolution. *The Journal of Economic History* **54**: 249–70

Eggebrecht U, Flemming J, Meyer G, v. Müller A, Oppolzer A, Paulinyi A, Schneider H 1980 *Geschichte der Arbeit. Vom Alten Ägypten bis zur Gegenwart*. Kiepenbheuer & Witsch, Köln, Germany

Godelier M 1980 Work and its representations: a research proposal. *History Workshop Journal* **10**: 164–74

Hausen K 1981 Family and role-division: the polarisation of sexual stereotypes in the nineteenth century—an aspect of the

dissociation of work and family life. In: Evans R J, Lee W R (eds.) *The German Family*. Croom Helm, London, pp. 51–83
Honeyman K, Goodman J 1998 Women's work, gender conflict, and labour markets in Europe 1500–1900. In: Shoemaker R, Vincent M (eds.) *Gender and History in Western Europe*. Arnold, London
Jahoda M, Lazarsfeld P, Zeisel H 1933 *Die Arbeitslosen von Marienthal. Ein soziographischer Versuch*. S. Hirzel Verlag, Leipzig, Germany
Janssens A (ed.) 1998 *The Rise and Decline of the Male Bradwinner Family?* Cambridge University Press, Cambridge, UK
Joyce P (ed.) 1987 *The Historical Meanings of Work*. Cambridge University Press, Cambridge, UK
Kaplan S L, Koepp C J (eds.) 1986 *Work in France. Representations, Meaning, Organization, and Practice*. Cornell University Press, Ithaca, NY
Le Goff J 1980 *Time, Work & Culture in the Middle Ages* (Pour un autre Moyen Age: temps, travail et culture en Occident). The University of Chicago Press, Chicago
LeRoy Ladurie E 1975 *Montaillou, Village Occitan de 1294 à 1324*. Editions Gallimard, Paris
Lüdtke A 1993 *Eigen-Sim. Fabrikalltag, Arbeitererfahrungen und Politik vom Kaiserreich bis in den Faschismus*. Ergebnisse Verlag, Hamburg, Germany
Marx K 1867 (1968) *Das Kapital. Kritik der politischen Ökonomie*. Marx-Engels Werke Band 23, Dietz Verlag, Berlin
Pahl R E (ed.) 1988 *On Work. Historical, Comparative and Theoretical Approaches*. Basil Blackwell, Oxford, UK
Pahl R E 1984 *Divisions of Labour*. Basil Blackwell, Oxford, UK
Passerini L 1982 Work ideology and working class attitudes to fascism. In: Thompson P (ed.) *Our Common History: The Transformation of Europe*. Pluto Press, London, pp. 54–78
Sabel C F, Zeitlin J (eds.) 1997 *World of Possibilities. Flexibility and Mass Production in Western Industrialization*. Cambridge University Press, Cambridge, UK
Samuel R 1977 The workshop of the world: steam power and hand technology in mid-Victorian Britain. *History Workshop Journal* **3**: 6–27
Simonton D 1998 *A History of European Women's Work. 1700 to the Present*. Routledge, London
Thompson E P 1967 (1993) Time, work discipline and industrial capitalism. In: Thompson E P (ed.) *Customs in Common*. Penguin Books, London, pp. 352–403
Tilgher A 1931 *Work. What it has Meant to Men Through the Ages (Homo Faber)*. (Homo faber. Storia del concetto di lavoro nella civiltà occidentale). George G. Harrap, London
Tilly L A, Scott J W 1978 *Women, Work, and Family*. Holt, Rinehart and Winston, New York
Topalov C 1994 *Naissance du Chômeur 1880–1910*. Édition Albin Michel, Paris
Ure A 1835 *The Philosophy of Manufacturs; or, An Exposition of the Scientific, Moral, and Commercial Economy of the Factory System of Great Britain*. C. Knight, London (3rd edn 1969 B. Franklin, New York)
Weber M 1904/05 (English edn. 1976) *The Protestant Ethic and the Spirit of Capitalism*. Scribner, New York
Wiesner M E 1989 Guilds, Male bonding and women's work in early modern Germany. *Gender & History* **1**: 125–37
Williams R 1976 *Keywords. A Vocabulary of Culture and Society*. Fontana Press, London, p. 335
Wunder H 1992 *'Er ist die Sonn', sie ist der Mond.' Frauen in der Frühen Neuzeit*. C. H. Beck, Munich, Germany

J. Ehmer

Work, Sociology of

Work is central to the panhuman existence. In the twin disciplines of sociology and sociocultural anthropology, the study of work has thus long been a central interest, and it pervades much of the disciplines' published research. 'Work' has perhaps the widest meaning of any word in English. It is a culture-bound, scholarly concept, and folk idea regarding both a process and its accomplished state. In this article, the value-laden meanings and the dimensions of 'work' are examined. Also a brief history and salient conceptual issues of the sociology of work are reviewed.

1. *A Brief History of the Sociology of Work*

Few ethnographic monographs about nonwestern societies exclude information on work. In the twentieth century, the twin disciplines depicted the centrality of work in community studies in the USA and Canada. Employment and its access are a vital interest of the fieldworker studying a community, often from an ecological perspective. Pioneering research on work and workers in North America apart from community studies includes that of Whiting Williams, beginning in 1918, on the human relations of employees and employer. With the advent of the experiments and observations regarding workers in the Hawthorne plant of Western Electric, by 1930, the twin disciplines entered their phase of industrial social science, often having a social psychological bent. Studies of work received legitimization in 1946, when the American Sociological Association formed its Section on Industrial Sociology and again, in 1951, when Delbert Miller and William Form wrote the first textbook, *Industrial Sociology*. Work-studies are a variety of empirical emphases not having a characteristic conceptual framework or methodology. Workers display the ordinary range of social behavior. Organizations of work, from corporations to cliques on the job, are components of the greater human society. Nevertheless, in studying modern societies, work relations are part of most conceptualizations of social structures, and social information is most frequently collected in a workplace.

The research on work broadened to encompass many emphases and overlapping subtopics, from gender issues to division of labor. Work research gradated into investigations conducted across disciplines such as on organizations, social control, industrial relations, and role theory. What, then, is 'work'?

2. *Work as Culture-bound Idea*

Knowledge does not exist supra culture; thus, one must first examine the Anglo-Saxon cultural context of work. Work is the broadest English word for both

the process of doing something and the state of something done. Work is viewed ambivalently. It is a moral necessity and the deity's curse of travail upon humankind ('in the sweat of thy face...'). As etymologically rooted and in folk views, work is conceived broadly, as all endeavors.

Many social scientists, however, attempt an analytically useful narrowing of definition to the creating of goods and services valued in a society. But valued by whom? As Melville Herskovits explains, the phenomena of value regarding things exchanged are understandable only as a part of a specific culture or a subculture (Herskovits 1952). Adriano Tilgher informs, because economic value is always relative, so too is the work creating it (1958).

The social value of work then, depends upon the socialized view of the observer, varying even within a specific society. Managers of ante-bellum Southern railroads and their corporate slaves had differing views on the nature of value created.

As with other issues in the sociology of work for example, the management-oriented studies spawned by Hawthorne, definition of value in work resolves into whose ox is being gored, or fed (Carey 1967). A social disruption, however, could be eufunctional for at least part of a society. Thus an intensive strike after which workers lose their employment is not always dysfunctional to labor overall. Locomotive engineers and switchmen decisively lost the bitter strike on the Burlington road in 1888. Management's 'victory', however, proved so costly to shareholders that other railroads quickly began recognizing the labor organizations of the on-train operating crafts. If not definable by value, work could be usefully viewed in various dimensions.

3. *The Dimensions of Work*

Work has a personal dimension in defining the self and obtaining positive affect from community members. This dimension has a strong hold on all humans, in nonmarket and market societies. In the Horn of Africa, the Wayto hippopotamus hunter seeks his village's recounting of his prowess as a killer of game; the Qemant plowman seeks community recognition as a good husbandman.

For the street-corner man of Washington, DC, job experiences and consequent job fears constantly assault him. Terrified of an opportunity to work, he stands defeated by experience, with self-worth destroyed and fear thereby validated. Work greatly affects the integration of personality.

Work also has a social dimension, concerning interpersonal relations' as much as productive output. Many behaviors at work are comprehensible at the collective level. Individual self-interest is dampened by behavioral prescriptions including norms and etiquette. Moreover, in industrial society, with its atrophied bonds of kinship and limited ties of community, work in groups provides opportunities for the close interactions that all primates require, including humanly exchanging positive affect and releasing tensions in a 'support group.'

Each culture structures work in its own particularized dimension of time. To coordinate industrial transportation, standard time zones were created. For modern production and communication, the microsecond was created. Even in so-called timeless foraging and agrarian societies, space cannot be used effectively without some kind of reckoning of time for coordinating activities. A number of temporal modes including frequency, synchronization, and duration time an industrial society. Without such timing, society's work becomes dysfunctionally independent instead of functionally interdependent. The clock, more than the steam engine, was the serpentine machine in the garden of agrarian America.

Anglo-Saxon folk and intellectual thought about work exclude much laboring activity. Some activities are not considered work when the time and exertion seem sporadic and little planned, the activities are only somewhat or unrelated to an institutional product (purpose), or these activities are commonplace. For utility in theory, a concept of human work must be valid across cultures and not bound by Anglo-Saxon culture. Need work be purposeful and gainful, and how is livelihood cross-culturally viewed?

4. *Work as Concept*

Perhaps useful is contrasting a biological exertion involving energetic activity of motor acts and thought with their expression in culturally patterned work. A purely mental activity such as planning a task becomes work. Work, as solely a motor act apart from related thought probably does not occur.

Because work is social, even when done in isolation from other humans, it is enmeshed in the collective conventions of a society. A person's work activities ordinarily intertwine with activities of others in role sets for producing, exchanging, and consuming goods and services. A work role is a cluster of attributes and expectations for a work status, a social position in a society occupied by a person. Even when work is performed apart from other persons in a set, it is enacted regarding a reference group for such activities. Shipwrecked, Robinson Crusoe remained a member of his tinkering British society.

Should exertion be purposeful to qualify as work? Purpose overlaps with relative social value. The issue remains, purpose from whose perspective?

Canadians each perform the same exertion in felling trees: an employed lumberjack, a thieving 'timber pirate', and a gang of vandals. In economic parlance, the first two persons purposefully 'add value', the third group does not. To the extent that defining work

depends on purpose, this characteristic is (sub-) culturally relative.

Work could be defined as gainful; a livelihood earned in a nonmarket, transitional, or market economy. Broadly considered, gainful work does not require market exchange or money. It could be coerced. The gain could be either an increment or a saving of something valued, incorporeal or material. A shaman reciprocates séances for sundry services from community members. A chief collects part of all produce and redistributes some to the chiefdom's members in kind and some as political administration.

In modern societies, an individual has multiple gainful statuses of varying duration. A woman might be a monetarily unpaid homemaker, corporate executive paid in the market, communities volunteer given positive affect, and administrator of a food cooperative compensated in kind. Livelihood work thus supports a human's life (one's biological requisites) and life way (one's sociocultural requisites). With such a broad conception of livelihood, less kinds of exertion become excluded as work.

5. *Work in Modern Societies*

Although a conception of work must be cross-culturally valid, much of its investigation concerns industrial and industrializing societies. In industrial society, few remain primary food or craft producers. Instead, people consume goods and services created outside their local community. They gainfully work mainly for increments and savings to enhance their standard of consumption. Devices enormously multiply exertion in work with a consequently great personal productivity. By hoe, a Plains Indian tilled 0.1 of an acre annually, but, with machines, a Plains farmer tills 2,000 acres annually. Although a concept of work does not depend upon money, in today's market economy, it is the acme of activity. Sufficient money allows great independence from kinship bonds and community ties.

Because monetarily paid work is a central identifier of the self; the characteristic used as a person's overall social label and assessment is occupational status. After the ritualistic 'How do you do?' comes the meaningful, 'What do you do?' Work furnishes a person's basis of social power and economic and other welfare.

In its concomitants, loss of work could mean the severing of one's mooring in society. The largest medium of monetary savings, the house, can be lost. In the USA, vital health insurance is terminated and retirement credits are not accrued. Also concomitantly, the job delimits the kind of neighborhood along with the personal safety and property security obtainable.

It determines the quality of housing inhabited. It influences the kind of education and opportunity for children. It also restricts who will be friends, neighbors, and co-workers. Above all, the job allows a reassuring, predictable routine in an otherwise hectic life.

6. *Social Control of Work*

Coherence in science rests on a central notion. In the twin disciplines, the notion might be social control, their most widely used concept but about which no agreed definition exists. The term could mean an underlying social process, a social mechanism delimiting behavior, or a scholarly model for analysis. Most conceptions, however, concern norms (a culture's prescriptive 'shoulds' and 'oughts') and related sanctions, and the ways in which social organizations change while preserving some societal stability (Gibbs 1994).

Control maintains social relations not equally beneficially to all the involved actors, depicted in the innumerable studies of work and reactions to controllers of work.

Industrial relations constitute contests for social control between employers and employees. Social control of work has a number of forms. These include bureaucratic, by formal rules and procedures; technological, by devices limiting or monitoring action; and occupational, by grouped workers defining, in some part, the terms, conditions, and content of work. The pivotal question in studying control is, 'in what ways are persons made to act in a manner not serving personal self-interest but, instead, society's collective interest?'

6.1 *Employee Participation*

In the 'new industrial relations', the strong control of work by managers is reclad with 'cooperative' programs such as: quality circles, team concept, and employee involvement. For unionized workers, participation in the programs necessarily broadens the character of their labor-management relations. A cooperative program, if fully developed and controlled through bilateral negotiations, could foster mutual benefits. Participation programs, however, are frequently created in trendy public relations, (naively?) overlooking the opposing goals and highly consequential exercises of social control in the relations between labor and management. A large business organization possesses great social power, alone and in coordination with similar firms united in associations having special access to government.

Unorganized employees, and sometimes those organized, are at a great disadvantage regarding power relations and abilities to exercise social control in the market. In its traditional adversarial role of countering union social control, through cooperative programs management can usurp union responsibilities and pose as a benefactor of employees. Cooperation could allow

a transition from powerful collective to powerless individual bargaining for employees. A true employee participation means management must surrender some of its social control of workers. Genuine co-operation in work is fully bilateral, thus a partnership. It can only be effected by near equals in social control. This has not yet been the case. The global market pressures on wages and conditions of employment are inexorably downward.

6.2 Social Control of Technological 'Progress'

Business practice in North America, includes a sectarian, *laissez-faire* ideology promoting technological 'progress.' A tenet of such practice is a view that workers, their families, and fellow community members must suffer the consequences of progress without redress or complaint. Technology, however, is not a neutral force, or impartial hand of advance in an industry, for shaping society and inclining an individual's life. Technological change for work receives guidance from visible hands, according to goals of private firms, public agencies, and other organizations including unions and advocacy groups. In business, only rarely does management effect social control solely through technology. Ultimate control comes through the bureaucratic mechanisms of a firm, thus from managerial norms. The high value North Americans place on technological change masks what seldom changes, the social controls shaping technology and its policy and, thereby, society.

7. Gender

In foraging societies, occupational Jacks- and Jills-of-all-trades exist, respectively, hunters and gatherers. As technologies of plant and animal domestication free some persons from primary food getting, a true division of labor, with occupational specialization, develops in addition to the universal natural division of labor, by age and sex.

Despite a true division—including specialization of work technology within segments of a society or separation of a particular technology into its component techniques, or both—division of labor by sex remains everywhere.

A specific culture assigns to each biological sex a number of gender roles that are mainly learned culture. The roles are normative results of socialization, including for work. Not only do the conventional roles delimit what is 'male' and 'female', but also what is appropriate work for each sex.

Because a great overlapping range of somatic and mental characteristics exists for males and females, a similar range of tasks performed equally well by each sex obtains. Cross-culturally, almost no work roles are exclusively men's roles. Among the Hopi, women do the construction of multistory houses; among the Dahomey, women were the assault troops; and, in many African societies, women have apical political statuses. Whatever a man can do a woman can do equally well, with few exceptions, mainly related to somatic strength.

Nevertheless, work-role and occupational sex typing is dominant even in industrial democracies. At Christmas time, a boy customarily receives a black doctor's kit and a girl a pink nurse's kit. Such socialization for conventional aspiring to statuses of work is multi-stranded, from throughout the fabric of a culture. Some feminist scholarship explores the relations among capitalism, patriarchy, and job segregation. Although these neo-Marxist studies are illuminating, they do not explain similar segregation in the patriarchal, plow-agricultural societies, which might additionally feature other sex typing such as, clitorectomies, infibulations, and familial honor killings of females.

In industrial societies, occupational sex typing of women not only limits their kinds of work and levels of compensation allowable but also the range of beneficial concomitants of work. Perhaps the overcrowding hypothesis accounts for the lower compensation achievable by women compared to men.

With large numbers of women channeled and competing in relatively small numbers of 'pink-collar' work statuses, their compensation rates are driven downward in the market. Restricted compensation opportunities foster women's dependence on men, with increased exploitation by men.

Focus on the workplace ignores the importance of an unpaid wife at home supporting a successful husband at work. Such economic exchanges between social unequals are at odds with the folk category of the self-made man. Not only is the workplace a site of male exploitation of women, sexual and otherwise, but also it is a locale for behaviors from flirting to consensual sex. Not just sexual harassment but additionally sexual pursuits have a workplace arena, not surprising among highly sexual primates. The full range of organizational sexuality, its relations to job benefits and harm, and its resistance to social control are among little studied dimensions of work (Williams et al. 1999).

8. The New Nature of Work

In late industrial societies, the future of work could include not enough of it for permitting the earning of middle-class compensation. Competition in the global market engenders a massive de-industrialization with the loss of adequate jobs for those with high-school levels of education or less.

American production workers must compete with Chinese prison and South Asian child labor. In today's job market, employee loyalty has become a one-way street. Overnight, jobs can be eliminated in the thousands by a single firm. Some terminations are psychologically brutal.

The era of lifetime jobs gradually erodes, replaced by insecure contingent work, including part-time, contract and temporary persons—all workers who are not full-time with full benefits. Conditions of employment not obtained by contingent workers include job training and career advancement with a firm. Such workers become employed only with an immediate demand for them. Of course, some persons desire contingent work. Automation and reorganization of work created a controversy about the existence of deskilling, under capitalism, a continuous lowering of levels of skill throughout a national economy, leading to an unskilled proletariat.

De-skilling in neo-Marxist theory is a form of social control of workers through a minute division of labor, with organizing work into simple, repetitive tasks involving few mental reflections. Although a good number of crafts and some professions were deskilled during the twentieth century, no deskilling throughout an overall industrial economy exists.

Trade unions contest management for social control in some workplaces. In America, for unionized employees in private industry, the overall pay is one-third higher and for blue-collar workers about 70 percent more than in nonunion companies. Unionization provides a wage advantage to workers.

The future of the labor movement, however, is clouded. In the USA and Canada, trade unions have diminished in union density, organizing success, level of labor actions, and political influence. Reasons given for this decline include: union sloth; the law as enacted and then interpreted by the courts; globalized competition; 'cooperative' programs; and a neo-liberal philosophy embraced by the large political parties. For union success, new forms and policies in the labor movement must be created including, multi-employer bargaining, removal of legal impediments to organizing, and coordination with community, advocacy, religious, and ethnic organizations to further mutual goals (Clawson and Clawson 1999).

See also: Alienation, Sociology of; Capitalism: Global; Community Sociology; Industrial Sociology; Labor Markets, Labor Movements, and Gender in Developing Nations; Labor Movements and Gender; Labor, Division of; Organizations, Sociology of; Strikes: Sociological Aspects

Bibliography

Abbott A 1993 The sociology of work and occupations. *Annual Review of Sociology* **19**: 187–209

Applebaum H (ed.) 1984 *Work in Non-market and Transitional Societies*. State University of New York Press, Albany, NY

Bouvier P 1984 Le travail [Work]. *Sociétés* **1**(2): whole issue

Bouvier P 1989 *Le travail au quotidien: Une démarche socio-anthropologique* [Everyday Work: A Social Anthropological Presentation]. Presses Universitaires de France, Paris

Bouvier P 1994 *Le travail* [Work], 2 edn. Presses Universitaires de France, Paris

Carey A 1967 The Hawthorne Studies: a radical criticism. *American Sociological Review* **32**: 403–16

Clawson D, Clawson M A 1999 What has happened to the US labor movement? Union decline and renewal. *Annual Review of Sociology* **25**: 95–119

Gamst F C (ed.) 1975 Golden anniversary special issue on industrial ethnology. *Anthropological Quarterly* **50**(1): whole issue

Gamst F C (ed.) 1995 *Meanings of Work: Considerations for the Twenty-first Century*. SUNY Press, Albany, NY

Gamst F C, Helmers S 1991 Die kulturelle Perspektive und die Arbeit [The Cultural Perspective and Work]. *Zeitschrift für Ethnologie* **116**: 25–41

Gibbs J P 1994 *A Theory about Control*. Westview Press, Boulder, CO

Helmers S (ed.) 1993 *Ethnologie der Arbeitswelt: Beispiele aus europäischen und auereuropäischen Feldern* [*Ethnology of the Work World: Examples from European and non-European Locations*]. Holos Verlag, Bonn, Germany

Herskovits M J 1952 *Economic Anthropology*. Knopf, New York

Kalleberg A L, Berg I 1987 *Work and Industry: Structures, Markets, and Processes*. Plenum, New York

Miller D C, Form W H 1951 *Industrial Sociology*. Harper and Row, New York

Miller J 1988 Jobs and work. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA

Pennbridge J N 1985 Industrial Anthropology: A Selected Annotated Bibliography. *Society for Applied Anthropology*, Monograph no. 14

Simpson I H 1989 The sociology of work: Where have all the workers gone? *Social Forces* **67**: 563–81

Tilgher A 1958 [1930] *Work: What it has Meant to Men through the Ages*. Henry Regnary, Chicago

Wallman S (ed.) 1979 *The Social Anthropology of Work*. Academic Press, New York

Williams W 1918 *Human Relations in Industry*. US Department of Labor, Washington, DC

Williams C L, Guiffre P A, Dellinger K 1999 Sexuality in the workplace: Organizational control, sexual harassment, and the pursuit of pleasure. *Annual Review of Sociology* **25**: 73–93

F. C. Gamst

Working Classes, History of

1. The Working Classes: Origins of the Concept

The concept of the working class emerged in Western Europe during the nineteenth century. Originally, the term was used as a plural, denoting a heterogeneous conglomerate of social groups performing various forms of wage labor. The working classes were considered to be a fraction of a larger 'proletariat,' that is, the estate of people without property, beyond honor. The wage earners were only a part of this amorphous mass. According to the French nobleman Granier de Cassagnac (1838, p. 30), the proletariat formed 'the lowest rank, the deepest stratum of society,' that consisted of four groups: 'the workers, the beggars, the thieves and public women'. 'The worker is a proletarian, because he works in order to

live and earns a wage; the beggar is a proletarian, who does not want to work or cannot work, and begs in order to live; the thief is a proletarian, who does not want to work or beg, and, in order to make a living, steals; the prostitute is a proletarian, who neither wants to work, nor beg, nor steal, and, in order to live, sells her body.' As capitalist relations crystallized more and more, the desire arose among the 'respectable' workers to distinguish themselves from the rest of the 'proletariat.'

Over the course of the nineteenth century, social theorists and the emerging labor movements began to redefine these notions: (a) the working classes became homogenized as a singular working class; (b) this working class was identified with the proletariat; and (c) this proletariat/working class was marked off in three ways: from the petty bourgeoisie, from the so-called lumpenproletariat (the criminal and unsteady 'dangerous classes'), and from the unfree workers (slaves and serfs). An early example of this development was the *Manifesto of the Communist Party* of 1848, which emphatically distanced the proletariat/working class on the one hand from the 'lower middle class, the small manufacturer, the shopkeeper, the artisan, the peasant,' and on the other hand from the '"dangerous class," the social scum, that passively rotting mass thrown off by the lowest layers of old society' (Marx and Engels 1973, p. 77). A similar sharp distinction was drawn between 'free' workers and the slaves in the South of the United States and in some colonies.

Moreover, the notion of a working class was a gendered abstraction, based on the assumption that workers' families exist mainly on the wages earned by the male head of the household and perhaps the additional income of other household members. From 1900 or so scholars of various theoretical orientations seemed to have reached an agreement on what a worker is. In 1926 the German social-christian sociologist. Briefs summarized this *communis opinio* as follows: 'The worker is *personally free*, i.e., his physical and spiritual-moral powers are completely at his own disposal. […] He *has no property*, i.e., he has no exclusive material power over capital as a secure basis with relative permanency. […] He has neither a stock of consumer goods that enable him to live, nor permanent interests of capital. […] He lives in economic circumstances in which means of subsistence can be obtained only through *economic returns*. […] He is forced to offer personal capacities with an economic exchange value in return for means of subsistence (Briefs 1926, p. 149).

2. *Origins in the 'Core Countries'*

Detailed surveys of the living and working conditions of the working classes date from the first decades of the nineteenth century. A relatively early example was the *Tableau de l'état moral et physique des ouvriers employés dans les manufactures de coton, de laine, et de la soie*, written by Villermé (1840), a medical doctor with a great interest in the social and physical condition of 'the people'. Many writings of this kind were intended partly as moral and political indictments of social abuses. One of the best known instances is Engels's (1845) *Die Lage der arbeitenden Klasse in England* [*The Condition of the Working Class in England*], which gave a minute description of the working classes in English industry, mining, and agriculture. During the second half of the nineteenth century social surveys became a quite common genre, especially in industrializing Western Europe. Highlights include the many studies of Frédéric Le Play and his school which covered all of Europe, beginning with *Les ouvriers européens* (1855); Henry Mayhew's *London Labour and the London Poor* (1851–61) with the significant subtitle 'a cyclopaedia of the conditions and earnings of those that will work, those that cannot work, and those that will not work'; Charles Booth's *Life and Labour of the People of London* (1892–97); Alphons Thun's detailed monograph *Die Industrie am Niederrhein und ihre Arbeiter* (1879) on the industrial and working classes in West Germany; and numerous publications of the German Verein für Sozialpolitik (Social Policy Association), including Max Weber's famous study of agricultural laborers in East Prussia (1894).

All these works were mainly contemporaneous, though they have by now themselves become important historical sources. Truly historical studies of working classes followed at the beginning of the twentieth century. A pathbreaking effort was the trilogy written by John Le Breton Hammond and his wife Barbara Bradby. *The Village Labourer, 1760–1832* (1911), *The Town Labourer, 1760–1832* (1917), and *The Skilled Labourer, 1760–1832* (1919). Together these studies provide a rich survey of the British working class during the Industrial Revolution. In other countries similarly wide-ranging works usually were published much later, generally after the Second World War.

3. *Classical Debates*

The study of the working classes has always been central to social and labor history. But different aspects have been studied and emphasized.

The concept of the working class has been used in many different ways. Somewhat simplifying, one might distinguish three main patterns of interpretation. The structural approach uses the notion of the working class in a restrictive sense as a term merely characterizing the common socioeconomic location of a group of wage earners. The historical study of a particular working class then boils down to the reconstruction of their joint 'class position.' How this

class position is analyzed depends on theoretical assumptions (Marxist, Weberian, or otherwise) and on the various ways in which research is operationalized (models of social stratification).

A second approach focuses on agency and emphasizes culture, mentalities, consciousness, and social and organizational behavior. Numerous variations on this theme have developed over time. Perhaps the most influential approach was presented by Edward P. Thompson in his *The Making of the English Working Class* (1963). Here, he argued that. 'The notion of class entails the notion of historical relationship. Like any other relationship, it has a fluency which evades analysis if we attempt to stop it dead at any given moment and anatomize its structure.' Class, according to Thompson is an outcome of collective experience, and workers are active participants in their own class formation.

While the first approach has been criticized for its 'objectivism' and the second approach for its 'culturalism,' a third approach considers working-class formation as 'an outcome of temporal conjuntures between multiple causal structures' (Sewell 1990, p. 73). In this view, class formation results from a combination of economic and technological changes on the one hand and of cultural, political, and discursive developments on the other hand.

The debate on working-class formation is paralleled by an extensive debate on working-class devolution: Does it still make sense to speak of working classes in advanced capitalist countries? Or have these classes dissolved? And if there has been a process of devolution, how and when did it start? Naturally, everything depends again on one's conceptualization of the working class. What does one consider as decisive: structural issues (free wage labor, etc.), collective behavior (e.g., trade unions, strikes), or cultural features? Whichever position one takes, it is evident that the condition of wage earners in advanced societies has changed profoundly during the twentieth century. Research has not given equal attention to all these developments. Especially the cultural shifts after the Second World War remain underexplored.

Most early contributions to working-class history focused on work, industrial relations, and labor markets. At a later stage the field of research was expanded and historians began to pay serious attention to other aspects of working-class existence, including labor processes, family life, religion, etc. Over the course of time the first approach has become more sophisticated and the second approach has gained influence. The historiography of the workers as a class is now increasingly merged with the historiography of race, ethnicity, and gender. While earlier historians considered race and gender as mere adjuncts of class, it is now generally becoming accepted that class, race, and gender are thoroughly interpenetrating, and that they shape each other. As a result of these developments, working-class history increasingly overlaps into women's studies, ethnic studies, anthropology, institutional economics, sociology, or social psychology.

Now that there exists a huge and still growing body of literature on the various historical aspects of working-class existence first attempts are made to integrate the history of various aspects of working-class life and to write comprehensive syntheses. The most spectacular example of such an endeavour is the ongoing project on the '*History of the Workers and the Labor Movement in Germany since the End of the Eighteenth Century*,' co-ordinated by Gerhard A. Ritter (see Kocka 1990, 1991, Ritter and Tenfelde 1992, Schneider 1999, Winkler 1984, 1985, 1987).

4. *The Rise of the 'Periphery'*

The growth of working classes outside the advanced capitalist countries has, especially since the Second World War, caused an increased interest in the working classes of Africa, Asia, and Latin America. From a contemporary point of view this field was first explored by anthropologists and students of industrial relations, who were interested in the problems accompanying the integration of rural migrants into the emerging wage-earning economies. Early examples were A. L. Epstein's ethnographic *Politics in an Urban African Community* (1958), on copper miners in Northern Rhodesia, and industrial relations studies driven by the question how processes of social change could be controlled and how the 'commitment' of the workers could be intensified (e.g., Moore and Feldman 1960; see also Breman 1999).

Simultaneously, historiographical studies emerged; initially these were often based on official sources and looked at the workers through the eyes of employers and authorities. Such was for instance the case in J. Norman Parmer's thorough *Colonial Labor Policy and Administration* (1960) on the rubber plantation industry in Malaya during the last decades before the Second World War. Since the 1960s and 1970s the historiography of the 'peripheral' working classes has changed too. First there was an upsurge of classical labor movement history. Many of the major early studies of this genre were written by radical intellectuals, like Jean Chesneaux's *Le mouvement ouvrier en Chine de 1919 à 1927* (1962) or Guillermo Lora's *Historia del movimiento obrero Boliviana* (1967–70). Second, gradually the perspective broadened and everyday life, mentalities, etc. of working classes received more attention, just as had been the case in the historiography of the high-income countries. Early attempts in this direction were still heavily influenced by Western models of historical development. Morris D. Morris's *The Emergence of an Industrial Labor Force in India: A Study of the Bombay Cotton Mills, 1854–1947* (1965) is a major example of this approach, despite its rich empirical basis. Later attempts

often tried to develop a less Eurocentric approach. Pathbreaking works include Charles van Onselen's *Chibaro* (1976) on mine labor in Southern Rhodesia or Ranajit Das Gupta's *Labour and Working Class in Eastern India* (1994) on plantation workers, miners, and textile workers in Assam, Bengal, and elsewhere. These new developments have been covered by a series of consecutive collections of essays (Sandbrook and Cohen 1975, Gutkind et al. 1978, Cohen et al. 1979, Munslow and Finch 1984, Agier et al. 1987, Amin and Van der Linden 1996). While the traditional historiography of the high-income countries focused primarily on artisanal and factory labor, the new working-class historiography of the peripheral countries reflected the peculiarities of colonial exploitation and unavoidably paid much attention to the labor in agriculture (including plantations), mining, and the transport sector (docks, railways).

5. *Global Insights*

The rise of working-class history in low-income and medium-income countries has made necessary a rethinking of old concepts. First, in the periphery, the larger part of the wage earners is not 'free' in an economic sense. According to the classic definition, wage earners, as 'free persons' are free in a double sense because they alone market their own labor as a commodity and have no other commodity for sale. A 'free' worker is, strictly speaking, someone without obligations to any particular employer. He or she can cancel employment at will and apply for a job with another employer. In reality, however, such a completely uncommitted relationship with the boss is rare, occurring for example, among some forms of casual labor. Many workers are tied to their employer, for example through debts or nonmaterial obligations. There can also be all sorts of cultural or geographic reasons why it is difficult to change employers. Somewhat better-placed personnel can be tied to a firm through superannuation provisions. Frequently, the accommodation of workers is part of the labor contract, etc. In short: what formally seems like a free labor relation is in practice often much less free. A 'free' worker, moreover, does not possess his or her own means of production at all. But in reality there are many types of labor relations in which the employee contributes at least part of the means of production himself. One thinks here of factory workers who own (part of) their own tools, but one can also think of agrarian share-cropping arrangements. Still clearer is the case of self-employed workers, who formally are wholly independent, but who in fact often form a very flexible labor force for enterprises. We can also mention here the case of home workers.

Second, there is wage labor integrated throughout in households and families whose survival always remains partly dependent on other activities. Not only does a great deal of subsistence labor take place every day in every family (especially, but not exclusively performed by women), but there are usually also several forms of activities remunerated with money. Already in the early 1970s Allen observed, for example: 'Cash crops and wage employment are simply alternative methods of coping with the encroachment of capitalism on the traditional way of life. A fall in the demand for labour increases the emphasis on subsistence or cash-crop production, whereas a drought or a bad harvest, for instance, forces Africans to enter the labour market (Allen 1972, p. 181). Within these arrangements with multiple coping methods the 'roles' different family members have are often not fixed permanently, but instead signify a transient social relationship, which can be replaced rather quickly by other sources of income.

Third, the new working class does not mainly or exclusively exist in the industrial sector. The agricultural sphere is proportionally more important in this sphere rapidly advancing proletarianization has created a large stratum of agricultural laborers and share tenants.

Fourth, in the so-called Third World and even in the East- and Southeast Asian newly industrializing countries, permanent employment relationships have always been the exception rather than the rule. Insecure employment relationships, underemployment, and long-term unemployment are the reality there.

6. *Implications for Future Research*

The new insights offered by Third World historians have implications for the advanced countries as well.

First, the dividing line between wage labor and small entrepreneurship is much more obscure than was originally thought. In both the 'Third' and 'First' worlds there are gray zones of (ostensible) self-employment, in which individuals formally work on their own but in reality are dependent on one or two customers. This type of self-employment has been widespread in peripheral capitalism for decades, but lately, it has been increasing rapidly in the centers as well, such as in the construction business. Generally, there seems to be a historical connection in highly developed capitalism between phases of rising unemployment, on the one hand, and stagnation or the growth of 'small-scale' self-employment, on the other.

Second, the border between wage laborers and the marginal groups is not nearly as obvious as older theories would have us believe. In times of desperation, regularly employed wage earners may indeed exhibit behavior characteristic of the *Lumpenproletariat*, such as begging and prostitution, as a way to survive. They will also steal at the workplace, which represents thoroughly criminal behavior from the entrepreneurial viewpoint. Yet, the experience in the Third World has drawn our attention to a lower social class that can be characterized neither as the industrial reserve army in

a Marxist sense nor as the classic *Lumpenproletariat*. Inevitably we need to come up with a new concept to apply to insecure employment relationships in this respect.

Third, the concept of the 'free' laborer is less precise than is usually assumed. Fourth and last, the strict differentiation commonly made in advanced countries between urban and rural life must be revised. Contrary to what modernization theories like to tell us, the ties of urban migrants to their home villages often do not weaken but strengthen over time. Most likely, the reason for this is the lack of social insurance systems. Therefore, the villages become safety nets during times of economic crisis. Recently, though, many villages have not been able to fulfill this task, since the rise of a commodity economy has drained their capacity for subsistence.

What it means to be classified as 'working class' is changing; the category is becoming especially multifarious. We could even flip our intellectual problem and ask why a 'full-time proletarian' in capitalism should have existed or can now even exist at all. From the entrepreneurial standpoint, part-time proletarians are cheaper by far. Part-time proletarians (who belong to a household in which there are other sources of income, e.g., from agricultural work or from self-employment) lower the bottommost, barely acceptable wage threshold because they have other forms of real income (Wallerstein 1983).

See also: Alienation, Sociology of; Bourgeoisie/Middle Classes, History of; Class Consciousness; Class: Social; Industrial Society/Post-industrial Society: History of the Concept; Labor History; Labor Movements, History of; Labor Unions; Labor, Division of; Marx, Karl (1818–89); Work and Labor: History of the Concept; Work, History of; Work, Sociology of

Bibliography

Agier M, Copans J, Morice A (eds.) 1987 *Classes ouvrières d'Afrique noire*. Karthala-Orsthom, Paris

Allen V L 1972 The meaning of the working class in Africa. *The Journal of Modern African Studies* **10**: 169–89

Amin S, Van der Linden M (eds.) 1997 *"Peripheral" Labour? Studies in the History of Partial Proletarianization*. Cambridge University Press, Cambridge, UK

Breman J 1999 Industrial labor in post-colonial India, part 1/part 2. *International Review of Social History* **44**: 249–300, 451–483

Briefs G 1926 Das gewerbliche Proletariat. In: *Grundriss der Sozialökonomik, Part IX*. Mohr (Paul Siebeck), Tubingen, Germany, pp. 142–240

Cohen R, Gutkind P C W, Brazier P (eds.) 1979 *Peasants and Proletarians: The Struggles of Third World Workers*. Hutchinson, London

Granier de Cassagnac A 1838 *Histoire des classes ouvrières et des classes bourgeoises*. Desrez, Paris

Gutkind P C W, Cohen R, Copans J (eds.) 1978 *African Labor History*. Sage, Beverly Hills, CA

Kocka J 1990 *Weder Stand noch Klasse. Unterschichten um 1800*. Dietz, Bonn

Kocka J 1991 *Arbeitsverhältnisse und Arbeiterexistenzen. Grundlagen der Klassenbildung im 19. Jahrhundert*. JHW Dietz, Bonn

Marx K, Engels F 1973 Manifesto of the Communist Party (1848). In: Marx K (ed.) *Political Writings, Vol. 1: The Revolutions of 1848*. Penguin Books, Harmondsworth, UK, pp. 67–98

Moore W E, Feldman A S (eds.) 1960 *Labor Commitment and Social Change in Developing Areas*. Social Science Research Council, New York

Munslow B, Finch H (eds.) 1984 *Proletarianisation in the Third World*. Croom Helm, London

Parmer J N 1960 *Colonial Labor Policy and Administration. A History of Labor in the Rubber Plantation Industry in Malaya, c. 1910–1941*. Augustin, Locust Valley, NY

Ritter G A, Tenfelde K 1992 *Arbeiter im Deutschen Kaiserreich 1871–1914*. Dietz, Bonn

Sandbrook R, Cohen R (eds.) 1975 *The Development of an African Working Class: Studies in Class Formation and Action*. Longman, London

Schneider M 1999 *Unterm Hakenkreuz. Arbeiter und Arbeiterbewegung 1993–1939*. Dietz, Bonn

Sewell W H Jr 1990 How classes are made: Critical reflections on E. P. Thompson's theory of working-class formation. In: Kaye H J, McClelland K (eds.) *E. P. Thompson: Critical Perspectives*. Polity, Cambridge, MA, pp. 50–77

Thompson E P 1963 *The Making of the English Working Class*. Gollancz, London

Villermé L-R 1840 *Tableau de l'état physique et moral des ouvriers employés dans les manufuctures de coton, de laine et de soie* (Introductions J P Chaline, F Dérnier). Etudes et Documentation Internationales, Paris

Wallerstein I 1983 *Historical Capitalism*. Verso, London

Winkler H A 1984 *Von der Revolution zur Stabilisierung. Arbeiter und Arbeiterbewegung in der Weimarer Republik 1918–1924*. Dietz, Bonn

Winkler H A 1985 *Der Schein der Normalität. Arbeiter und Arbeiterbewegung in der Weimarer Republik 1924–1930*. Dietz, Bonn

Winkler H A 1987 *Der Weg in die Katastrophe. Arbeiter und Arbeiterbewegung in der Weimarer Republik 1930–1933*. Dietz, Bonn

M. Van der Linden

Working Memory, Neural Basis of

In experimental psychology, the term 'working memory' has been used to refer to the temporary storage of information that is necessary for cognitive processing underlying the comprehension and learning of new material, reasoning, mental arithmetic, etc. This theoretical construct emerged in the 1970s out of earlier experimental behavioral work on short-term memory (see Baddeley 1986 for a review). The term working memory has been defined variously to refer not only to the transient maintenance of information but also to its processing. For instance, one of the most popular working memory models in cognitive psychology assumes the existence of a central executive system

supplemented by 'slave' systems that are specialized for the maintenance and processing of verbal and visual material (see Baddeley 1986). Such an expanded concept of a working memory system becomes tantamount to the entire neural processing of information that occupies current awareness. Naturally, any system defined so broadly will be the result of processing in multiple areas of the brain. This article will focus on research that aims to identify the role of several cortical structures in what has been termed working memory at the cognitive level of analysis.

1. The Frontal Cortex and Working Memory

There is considerable evidence that the frontal cortex plays a major role in working memory although its precise role remains a matter of debate. The frontal cortex occupies a very large area of the primate cerebral cortex. Its posterior part, which lies immediately in front of the central sulcus, consists of the motor and premotor cortex (see Fig. 1). The vast area that lies further forward is often referred to as the 'prefrontal' cortex. The prefrontal cortex has expanded enormously in the primate brain and comprises many distinct areas. These areas differ in terms of their cytoarchitecture (i.e., the arrangement of the neurons within them) and have their own distinct connections with other parts of the brain (Pandya and Yeterian 1985). These anatomical facts imply that there must be some differences in the local neuronal computations occurring within particular areas since both the local circuitry and the inputs and outputs are distinct.

The first evidence that suggested that the prefrontal cortex plays a major role in working memory came from work in the 1930s by Jacobsen and co-workers who examined the performance of monkeys with lesions restricted to the prefrontal cortex on the delayed response task (Jacobsen 1936). This work showed that frontal cortical damage rendered animals incapable of performing the delayed response task which was introduced by Hunter in 1913 to study transient memory in animals. This task typically is conducted in the following manner. The experimenter attracts the monkey's attention and, in full view of the animal, a reward (i.e., a piece of food) is placed under one of two identical stimuli (e.g., cups, plaques, etc.). The stimuli are then hidden from the animal's view by an opaque screen and, after a delay of a few seconds, the screen is raised and the animal is allowed to choose one of the two stimuli to obtain the reward if, of course, the choice is correct. In interpreting the impairment on delayed response by monkeys with frontal cortical damage, Jacobsen (1936, p. 52) argued that the behavior of such animals 'instead of being directed by a balance of immediate, recent and temporally remote stimulation, is dominated by present external stimuli and the stable habit systems derived from earlier training; the mnemonic influence of past experience is lacking.'

The demonstration that adequate performance of the delayed response and related tasks is severely impaired by lesions of the prefrontal cortex has been

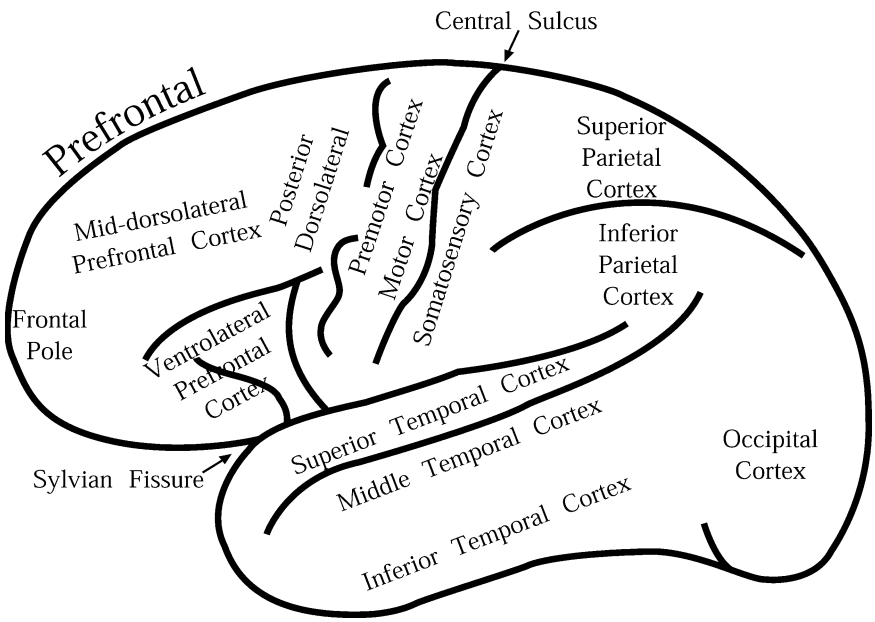

Figure 1
Outline of the left hemisphere of the human brain to illustrate the location of some of the major cortical regions

the starting point of a large amount of experimental work on its possible role in working memory. The lesion investigated in Jacobsen's work involved the entire prefrontal cortex. Subsequent work in the monkey by Pribram, Mishkin, Gross, Weiskrantz, and others showed that the lateral surface of the prefrontal cortex was critical for delayed response performance. Furthermore, the deficit on the spatial version of the task was localized around the sulcus principalis (a prominent sulcus that runs longitudinally in the middle of the lateral frontal cortex of the monkey). It also became clear that the delay was a critical variable for the impairment on these tasks (see Petrides 1994 for a review).

By the early 1970s, the method of recording the activity of single neurons in alert monkeys as they performed various cognitive tasks was perfected and a number of studies appeared that recorded the activity of single neurons in the prefrontal cortex during delayed response performance. Fuster and Alexander (1971) in the United States and Kubota and Niki (1971) in Japan were the first to record such neuronal activity. This work demonstrated that certain neurons in the lateral prefrontal cortex respond specifically when a particular location is indicated as the one to be remembered and continue to discharge during the delay period until the choice has been made. Thus, these neuronal responses appear to maintain temporarily the processed information. This early work and several later studies by Goldman-Rakic, Miller, and co-workers clearly established the existence within the prefrontal cortex of sustained neuronal activity during the delay period of various tasks (see Fuster 1997 for a review).

Thus, both the lesion and the single-neuron work indicated the involvement of the prefrontal cortex in some aspect of working memory. But what exactly is the role of the prefrontal cortex in working memory? One of the theoretical interpretations of the above data was to assume that the prefrontal cortex is a critical region for the short-term maintenance of information and to conceive the different prefrontal cortical areas as specialized working memory modules for the maintenance of different types of information (Goldman-Rakic 1996). For instance, it was argued that the dorsolateral prefrontal area is specialized for spatial working memory (i.e., the maintenance of spatial information), whereas the ventrolateral prefrontal area is a specialized module for nonspatial visual working memory (Goldman-Rakic 1996).

An alternative theoretical model of the role of the different lateral prefrontal cortical areas in working memory was proposed by Petrides (1994), based on work that further analyzed the nature of the impairments on memory tasks by human patients and monkeys with prefrontal lesions. It was argued that the existence of delay neurons in the prefrontal cortex, although implying that some neuronal computations are occurring there during the delay, does not require the conclusion that the primary role of the prefrontal cortex is the short-term maintenance of information. For instance, neurons with sustained activity during the delay in memory tasks can be observed in virtually every part of the cerebral cortex, as well as in several subcortical structures. In addition, there is the major puzzle that patients with damage to the lateral prefrontal cortical region can perform well on several standard short-term memory tests, such as simple recognition, digit span, and story recall. These patients, however, are impaired on the self-ordered working memory tasks in which successful performance cannot be based simply on the maintenance of information in short-term memory but requires, in addition, the executive process of monitoring that information. In the self-ordered working memory tasks, the subjects are presented with different arrangements of the same set of stimuli and, on each trial, they have to select a different stimulus until all have been selected once. Successful performance requires the subjects to compare carefully those stimuli that they have already selected with those still remaining to be selected. In other words, events in working memory must be monitored closely. It was, therefore, argued that the problem of patients with prefrontal lesions lies not in memory per se, but in the application of executive processes on certain memory tasks (Petrides 1994, 2000).

In work on the monkey, where lesions can be restricted with precision within particular regions of the brain, it was demonstrated that damage to the mid-dorsolateral part of the prefrontal cortex results in severe impairments on these self-ordered tasks and related tasks that require the monitoring of externally ordered stimuli (see Petrides 1994). Further evidence of the importance of the mid-dorsolateral prefrontal cortex to the monitoring of information within working memory comes from functional neuroimaging studies on normal human subjects that have consistently shown increased activity in the mid-dorsolateral prefrontal region when the monitoring requirements of a memory task are high (Petrides 2000). In a study by Courtney et al. (1998) in which monitoring and manipulation were minimized, activity related specifically to the maintenance of an active mnemonic representation of spatial information was observed in a region that was posterior to the mid-dorsolateral prefrontal region, suggesting that posterior dorsolateral frontal cortex may play a role in the focussing of attention on information that is held in posterior cortical regions.

2. *The Temporal and Parietal Cortex and Working Memory*

A large number of studies with patients who had sustained lesions limited to posterior cortical regions have documented impairments in short-term memory,

thus providing evidence that the temporary maintenance of information may be occurring in posterior cortical regions (see Vallar and Shallice 1990). There are several reports of patients with a selective impairment in auditory–verbal short-term memory after lesions in the inferior parietal region of the left hemisphere. These patients may be able to maintain and repeat single numbers, letters, or words presented auditorily, but their span (i.e., the number of items that they can maintain after a single presentation) is severely limited (e.g., Warrington and Shallice 1969). Patients with severely impaired short-term memory spans for visual verbal material usually have lesions in the occipitotemporal region of the left hemisphere (e.g., Shallice and Saffran 1986). De Renzi et al. (1977) have investigated visuo-spatial span using the Corsi blocks test. In this test, the patient observes the examiner tap, in sequence, a number of randomly arranged blocks and has to reproduce this sequence. Patients with right posterior hemisphere lesions were particularly impaired on this task, suggesting that the occipitoparietal cortex of the right hemisphere may be a key region for the maintenance of visuo-spatial information (De Renzi et al. 1997).

This notion that posterior cortical regions may play a role in the maintenance of information in working memory receives additional support from work in the monkey. For instance, there is considerable evidence that the anterior part of the inferior temporal cortical region, which is known to be the final stage in the occipitotemporal neural system underlying visual object perception, is also critically involved in short-term visual memory. This evidence comes both from investigations of the effects of lesions (Mishkin 1982) and from studies in which the activity of single neurons was recorded in monkeys as they performed visual short-term memory tasks (Fuster and Jervey 1982, Miller et al. 1991). Similar evidence exists in other sensory modalities. Lesions of certain parts of the superior temporal cortex have been shown to lead to a serious impairment in auditory short-term memory (e.g., Colombo et al. 1990).

How can the evidence for impairments in working memory both from posterior cortical regions and prefrontal cortex be conceptualized? In one theoretical model, the temporary storage and processing of incoming and recalled information is seen as an integral part of neuronal computation in modality specific and multimodal posterior association cortical areas (Petrides 1994). In other words, neural activity in these posterior cortical areas underlies not only perceptual processing and long-term storage of information, but also short-term memory of recently processed information which may be currently provided by the environment or re-activated from long-term storage. The prefrontal cortical areas are considered to play a role in working memory as distinct executive processing modules that allocate attention to different types of information, and monitor the contents of information in posterior cortical regions. Thus, if working memory were to be viewed as the multicomponent workspace of current cognitive activity, then this workspace can be construed as the local computations occurring within specialized posterior cortical regions for the maintenance and processing of modality-specific information, within frontal cortical regions for different aspects of executive processing and the bi-directional interactions between them.

See also: Cognitive Control (Executive Functions): Role of Prefrontal Cortex; Prefrontal Cortex; Prefrontal Cortex Development and Development of Cognitive Function; Working Memory, Psychology of

Bibliography

Baddeley A D 1986 *Working Memory*. Oxford University Press, Oxford, UK

Colombo M, D'Amato M R, Rodman H R, Gross C G 1990 Auditory association cortex lesions impair auditory short-term memory in monkeys. *Science* **247**: 336–8

Courtney S M, Petit L, Maisog J Ma, Ungerleider L G, Haxby J V 1998 An area specialized for spatial working memory in human frontal cortex. *Science* **279**: 1347–51

De Renzi E, Faglioni P, Previdi P 1977 Spatial memory and hemispheric locus of lesion. *Cortex* **13**: 424–33

Fuster J M 1997 *The Prefrontal Cortex. Anatomy, Physiology and Neuropsychology of the Frontal Lobe*. Raven Press, New York

Fuster J M, Alexander G E 1971 Neuron activity related to short-term memory. *Science* **173**: 652–4

Fuster J M, Jervey J P 1982 Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *Journal of Neuroscience* **2**: 361–75

Goldman-Rakic P S 1996 Architecture of the prefrontal cortex and the central executive. *Annals of the New York Academy of Science* **769**: 71–83

Jacobsen C F 1936 Studies of cerebral function in primates. I. The functions of the frontal association areas in monkeys. *Comparative Psychology Monographs* **13**: 1–60

Kubota K, Niki H 1971 Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology* **34**: 337–47

Miller E K, Li L, Desimone R 1991 A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* **254**: 1377–9

Mishkin M 1982 A memory system in the monkey. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* **298**: 85–95

Pandya D N, Yeterian E H 1985 Architecture and connections of cortical association areas. In: Peters A, Jones E G (eds.) *Cerebral Cortex*. Plenum Press, New York

Petrides M 1994 Frontal lobes and working memory: Evidence from investigations of the effects of cortical excisions in nonhuman primates. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*. Elsevier, Amsterdam, Vol. 9, pp. 59–82

Petrides M 2000 Frontal lobes and memory. In: Boller F, Grafman J (eds.) *Handbook of Neuropsychology*, 2nd edn. Elsevier, Amsterdam, Vol. 2, pp. 67–84

Shallice T, Saffran E M 1986 Lexical processing in the absence of explicit word identification: Evidence from a letter-by-letter reader. *Cognitive Neuropsychology* **3**: 429–58
Vallar G, Shallice T (eds.) 1990 *Neuropsychological Impairments of Short-term Memory*. Cambridge University Press, Cambridge, UK
Warrington E K, Shallice T 1969 The selective impairment of auditory verbal short-term memory. *Brain* **92**: 885–96

M. Petrides

Working Memory, Psychology of

Working memory (WM) can be thought of as the desktop of the human brain, a set of cognitive functions that allows humans to keep track of what they are thinking, what they are doing, or where they are moment to moment. It holds information long enough to make a decision, to acquire vocabulary, to mentally image the layout of a home, to support creative thinking, or remember what to do next. This article describes the theories that have been developed to explain these human abilities and some of the experimental evidence for these theories that has been derived from empirical studies of healthy adults and children and of individuals who suffer from cognitive deficits as a result of brain damage.

1. *Why is the Concept of Working Memory Useful?*

The concept of working memory (WM) has figured prominently in theories and explanations for human on-line cognition, as well as offering a basis for understanding how humans accomplish a wide range of everyday tasks. It can be thought of as a mental tool that aids survival; if an organism is to interact successfully with its environment, then clearly there must be some means to record, on a temporary basis, its own most recent actions and to have some means to plan actions for the immediate future. This gives a significant advantage over organisms entirely driven by automatic responses to external stimuli that signal basic needs. Examples of how this mental tool is used in everyday interaction with the environment highlight its utility. Humans use WM every time they have to solve a problem, to perform a mental calculation, to recall where an object has just been laid down, or to repeat a strange foreign word that has just been heard.

From these examples, it is clear that the concept is much broader than the more traditional idea of short-term memory. Indeed, the concept of WM arose from the need to provide a theoretical account of a wide range of everyday tasks that require both temporary memory and on-line cognition. Until the early 1970s the focus of research on this topic was on retention of random sequences of words, letters, or digits, with the everyday example of recalling telephone numbers while dialling. However a mental faculty would not have evolved to deal with these relatively artificial tasks, and therefore WM is intended to encapsulate on-line cognition and temporary verbal and visuo-spatial memory in both laboratory tasks and in everyday life.

2. *Theoretical Views of Working Memory*

Most memory researchers consider WM to be a useful concept that uses relatively few assumptions to explain a wide range of experimental evidence as well as important aspects of everyday cognition: it is a system with limited capacity, which loses information over periods of seconds, whose contents are subject to displacement by new input, and which interacts and is supported by other parts of the cognitive system such as stored knowledge or the products of perception. Most theories of WM include these features, however, there remains a debate as to its precise characteristics. Some theories view memory as a largely unitary system with WM as the sum total of what is currently activated from the store of knowledge, skills, expertise, and past events. Others view WM as separate from the long-term store of knowledge and events, but even among these latter theories there is some discussion as to whether the cognitive functions are supported by a single flexible system that provides both temporary storage and on-line processing, or comprises multiple components, each specialized for a particular function. A recent book edited by Miyake and Shah (1999) provides a comprehensive discussion of the different theories of WM.

The first type of theory, that WM is simply currently activated knowledge, can account for a wide range of data from experimental psychology with healthy adults, including the ability to bring rapidly to mind information related to areas of personal expertise. Theories of this kind can very successfully explain why chess masters can play several games simultaneously, expert waiters can recall details of meal orders from large groups, and why burglars can recall better than householders, or indeed the police, details of houses related to their crime. Expertise extends the apparent capacity of WM, but there is some recognition among proponents of this theory type of the need for a temporary memory system to deal with completely novel or unfamiliar material. Such theories additionally assume inhibitory processes that serve to prevent activation of information that is not strictly relevant to the task in hand. A breakdown in these inhibitory processes has been suggested as one of the corollaries of normal aging, with elderly people more prone to be misled by irrelevant information or to wander off the main topic in a conversation.

The strict view that WM is no more than the sum total of currently activated experience and memories for events runs into difficulty when considering dissociations that are apparent from studies of healthy adults performing more than one task at the same time. There is substantial evidence that healthy adults can perform two quite demanding tasks concurrently (such as holding a sequence of digits and following rapid random movements of a target). There is very little impact on the performance of either task, and this can be accomplished without any particular expertise or training. Complementary studies have shown that placing a heavy demand on temporary memory places few if any restrictions on retrieval of stored knowledge about past events or that is linked to expertise. These findings suggest that the human cognitive system may comprise several components, each specialized for particular kinds of tasks and which may operate in parallel. The findings also point to the idea that WM is best considered as being separate from the system that stores knowledge of the world and of experiences.

The unitary view of memory also has difficulty explaining patterns of impairment and sparing in individuals who suffer from cognitive deficits as a result of brain damage. The cognitive impairments are often selective, for example, with very poor performance on immediate memory tasks (such as repeating back a random sequence of digits) coupled with normal performance on long-term retention (such as recalling events that occurred weeks, months, or years previously). The converse dissociation also occurs, namely normal immediate memory contrasting with very poor retrieval from long-term memory or difficulties in long-term learning. Theories assuming that working memory comprises temporary activation of relevant information from long-term memory would predict that if long-term memory access is impaired then temporary activation should also result in poor performance. Theories that assume WM to be a functionally separate part of the cognitive system can account more readily for the neuropsychological dissociations as well as the data patterns from healthy adults, and therefore may be more useful.

3. *The Case for Multiple Memory Systems*

During the first half of the twentieth century the prevailing view in North America was of human memory as primarily the result of observable learning of associations, and theories of learning dominated. During the same period in Europe, memory as a cognitive concept was more salient, for example in the work of Bartlett, Broadbent, and Zangwill. By the mid-1950s, the latter became the more generally agreed view, and a cognitive approach to memory gave rise to major new avenues of research. Very quickly it became clear that memory was best thought of as comprising at least two elements, with one focusing on temporary or short-term retention and the other on long-term retention. This reflected a dichotomy proposed in the seventeenth century by the British philosopher John Locke who suggested that the 'faculty of retention' comprised 'contemplation' or 'keeping an idea in mind,' and that this was separate from 'memory' or the 'storehouse of these ideas.' This distinction appeared again in the late nineteenth century when the North American philosopher and psychologist William James referred to primary memory which he contrasted with retention.

During the 1960s, the contrast between primary memory and what then became secondary memory was demonstrated experimentally. For example, when healthy adult volunteers are asked to retain sequences of words for immediate recall in any order, omissions tend to occur among words from the middle part of the list and much better recall is observed for items close to the beginning or close to the end of each list. If instead, the volunteers are asked to recall the items following a delay, performance on the last few items is no better than it is for items in the middle of the list. However, initial items in the list retain their advantage despite the delay. These contrasting patterns were thought to support the idea that two memory systems are reflected in recall of different parts of the list: the early list items are thought to be rehearsed and transferred into a longer term memory system. As the list progresses, rehearsal becomes more difficult and therefore few items are transferred to a long-term store. The final items are thought to be held in a temporary memory system whose contents are lost unless recall is immediate. The advantage for the final items with immediate recall is known as the 'Recency Effect.'

Further support for the dichotomy comes from amnesic patients, that is people who, following brain damage, suffer from severe memory impairment that is characterized by an inability to learn new material. When such patients are tested for immediate free recall of word lists, their pattern of recall lacks the advantage for the initial items in the list but retains a normal recency effect. This suggests damage to the memory system responsible for retention of the early items but normal functioning of the memory system responsible for the recency effect.

A second experimental paradigm demonstrated that different memory codes are used in immediate and delayed recall. In one set of experiments, volunteers were given lists of words that were similar in meaning (e.g., tall, broad, wide, fat, huge, big, great) or were similar in sound, that is they were phonologically similar (e.g., man, mad, mat, map, cap, can, cat), for immediate recall or for recall after a delay. Recall of both lists was compared with memory for lists of words that were neither semantically nor phonologically similar (e.g. car, lock, spoon, book, chair, pen, dog). With immediate recall, phonologically

similar words were recalled more poorly than were words from the other lists, a phenomenon known as the acoustic similarity or the phonological similarity effect. It is interesting to note that this effect appears even when the words were read by the volunteers, suggesting that the visually presented words were translated into a sound-based or phonological code in memory. In contrast, with delayed recall, the semantically similar words led to poorest recall. Because phonological similarity affected short-term but not long-term retention, while semantic similarity showed the converse, this led to the conclusion that acoustic or phonological codes were a signature of a temporary memory system (for visual or auditory presentation) and semantic codes were characteristic of the use of a longer term memory system.

4. *Developing the Concept of Working Memory*

The idea of primary and secondary memory provided a useful framework to account for the accumulating body of experimental data in support of a two-component memory system. However, it was limited by the fact that virtually all of the research relied on experiments with presentation and recall of words, letters, or digits over relatively short time periods. It is clear that we can also retain information about the appearance of objects and scenes as well as their names and descriptions. Therefore, from the late 1960s, primary memory began to be seen as a system that could support temporary retention of different kinds of material as well as offering a means to process and manipulate that material. The dominant or 'modal' concept then became that of a short-term, or WM processing and storage system. This 'modal model' retained the notion that information was transferred from the senses, via WM to secondary, or long-term memory. It also incorporated the assumption that the system has a maximum capacity that could be employed by storage, by processing, or by elements of both.

However, a number of features of this model came under pressure from the 1970s onwards. The recency effect, once thought to be a major source of support for a temporary memory system, was also demonstrated in longer term recall. For example, when recalling visits to movies, where the car was parked, or attending rugby matches, individuals can often recall details of very recent events but have rather poorer recollection of experiences that are more remote. This demonstrated that recency was not unique to a temporary memory system. Also, there were demonstrations that semantic coding, thought previously to be the prerogative of long-term storage, could be used to help support short-term retention. For example, immediate free recall of real words is better than recall of nonsense words, suggesting that the additional stored information about the real words (such as their meaning) can support immediate memory. These findings did not wholly undermine the traditional view of a short-term memory, because there was other evidence that a unitary model of memory could not readily explain, such as that mentioned above from dual task performance and from brain damaged individuals. However, the findings of long-term recency effects suggested that recency might be a property of any human memory system (short- or long-term), and the evidence for short-term semantic coding showed that the memory systems interact, with any one memory task probably drawing on several different aspects of the cognitive system depending on task demands.

Another core assumption of the modal model has come under increasing strain in recent years, namely the idea that information from vision or hearing could reach long-term memory only through WM. For example, one famous neuropsychological case is that of patient 'PV' who, following a stroke in the sylvian region of the left hemisphere, presented with a very severe deficit in retaining sequences of words or digits for immediate recall, but had no difficulty in retrieving details of autobiographical events and proved well able to learn new material. The pattern for PV and other similar patients suggests that access to long-term memory does not depend on normal function in short-term verbal retention, either for new learning or for recall of past events (e.g. Vallar and Shallice 1990).

Additional evidence comes from patients who suffer from a neuropsychological disorder known as unilateral spatial neglect. This arises most commonly from damage to the right parietal cortex of the brain, and such patients typically are unable to report details of scenes or objects to their left, despite having normal vision. A few patients show unilateral neglect only for information in their mental image or immediate memory for scenes. For example, one patient 'NL' (Beschin et al. 1997) could quite readily report details of everything in his immediate environment. However, if he was asked to close his eyes, he could only report from memory details that were on his right. Moreover, if asked to recall details of a familiar scene such as the square in his home town he could recall only those details that were on his imagined right. This was not a problem of accessing his long-term memory, because when asked to imagine himself standing at the opposite end of the square, he could then report details that were now on his imagined right but had been omitted before from a different imagined viewpoint. This pattern demonstrates a clear dissociation between the processes of perception, which for NL were perfectly normal, the contents of his long-term memory, which again was normal, and the operation of his visual working memory which was dramatically impaired. It suggests that WM may not act as a gateway between perception and long-term memory as assumed by the modal model and by many recent introductory textbooks on memory.

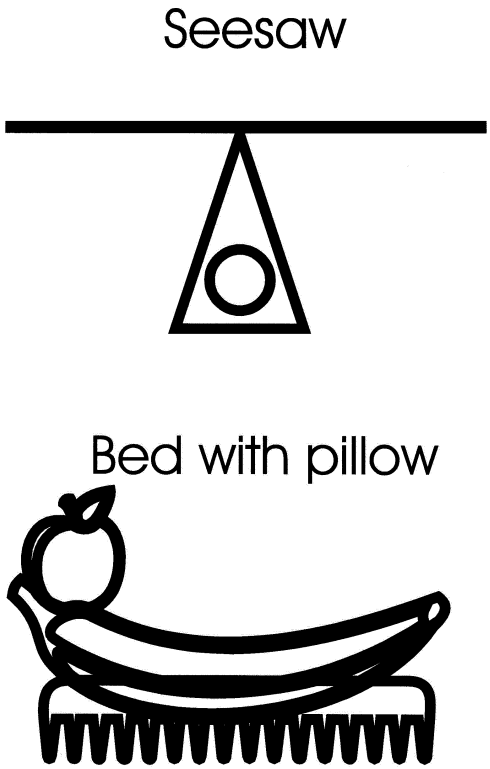

Figure 1
Example drawings plus names generated by volunteers in a mental discovery task involving combinations of geometric shapes (upper) and object shapes (lower)

Studies of healthy adults are pointing in a similar direction. Research over the last decade has explored the possible constraints on creative thinking or mental discovery. In one form of these experiments, volunteers are asked to imagine a set of shapes such as a triangle, a horizontal line, and a circle. They are then to manipulate and combine these shapes in their mental imagery to form a recognizable object which they then name and draw. There is no correct answer except that all the shapes must be included, and they must retain their original shape (e.g., a circle is not to be used as an oval). One possible response generated by a volunteer in these experiments is shown in the upper part of Fig. 1. This demonstrates that working memory can be used for mental discovery. However, if instead, volunteers are asked to combine the shapes of real objects rather than generic shapes, they have much greater difficulty in doing so. One successful drawing from a volunteer is shown in the lower half of Fig. 1 which is an attempt to combine the shapes of an apple, a banana, and a comb. The problem with object shapes was not their complexity, but the fact that they have a meaning associated with them. This made it difficult, for example, to think of an apple as anything other than an apple, because this requires reinterpretation of the shape. This suggests that the contents of working memory have meaning, and they can only have meaning if the perceived shape is interpreted in the context of stored knowledge. That is, it appears that information from the senses goes through and activates the contents of long-term memory before it becomes available for retention or manipulation in working memory, quite the reverse of the traditional flow of information in the modal model of memory.

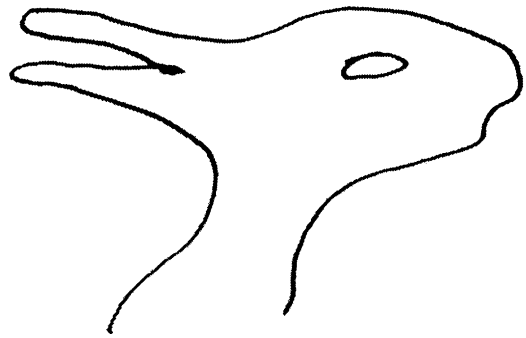

Figure 2
Example of the duck-rabbit ambiguous figure

A final, rather convincing demonstration that the contents of working memory are interpreted comes from experiments with ambiguous figures. Figure 2 can be interpreted in one of two ways, as a duck or as a rabbit, and if we continue to inspect the figure then our interpretation 'flips' between one and the other. However, in some ingenious experiments, Chambers and Reisberg (1985, see Cornoldi et al. 1996) showed this figure for only a brief time to a large number of volunteers who were asked to report from their memory what the figure depicted. All of the volunteers reported it as either a duck or a rabbit, but none of them could then reinterpret their image in memory and see the figure as anything other than their first interpretation. Nevertheless, they could do so when asked to draw their image on paper moments later. In other words, again, the contents of working memory are interpreted and under some circumstances are very difficult to reinterpret unless we can use the processes of perception with a physical drawing. This offers quite convincing evidence that perception feeds into long-term memory which activates relevant knowledge for manipulation with working memory (for discussions see Cornoldi et al. 1996, Helstrup and Logie 1999).

A second assumption of the modal model was that of WM as a single, flexible, limited capacity system that could undertake both temporary storage and some processing of the material being stored. This theme remains in one contemporary strand of WM research that has focused on measures of individual differences in cognitive capacity. In a typical task,

healthy adult volunteers are presented with a series of sentences. The requirement is to recall the last word of each sentence after all the sentences have been presented. The number of sentences is gradually increased until the participant can no longer accurately recall all of the final words, and the maximum number of words that can be recalled is taken as a measure of the individual processing and storage capacity. The measure in various forms has been shown to correlate well with other measures of cognitive ability such as reading comprehension or mental arithmetic performance. These correlations are then used to argue that WM is a mental capacity for processing and storage that can be applied to a wide range of both laboratory and everyday cognitive tasks. In this sense, the sentence 'span' (or WM capacity) measure has been quite useful, and offers the basis for a standardized test of cognitive ability. Set against its pragmatic value, the underlying theoretical assumptions are that there is a single flexible cognitive system supporting performance. One possible alternative view is that the task taps several aspects of cognition, and that this same range of cognitive functions is employed for other complex cognitive tasks such as reading comprehension. A further limitation is that the test provides only a measure of memory in the context of processing (retention of the final words while reading). In a recent test of the above assumptions, a separate measure was taken of processing capacity by asking volunteers to verify each of the sentences for their semantic content without a memory requirement. Memory ability was measured independently by asking volunteers to recall lists of individual words. Participants were then asked to combine verification with recall of all of the final words, and the task demands were set at the maximum that had been determined for each task component separately. If WM were a single flexible, limited capacity system, we might expect performance to be very poor when these two demanding task elements were combined. Results showed that despite the high processing demand and high memory demand, performance on both elements of the task remained very close to those obtained when the subjects were performing the tasks independently.

These results point to the notion that temporary memory might be served by different aspects of the cognitive system from those used to support on-line processing. This argument in turn suggests that WM might be best thought of as the orchestrated operation of multiple cognitive instruments, each one with a specific role for temporary memory or for processing.

5. *Working Memory as a Multiple Component Workspace*

The most influential model of multicomponent WM was proposed in 1974 by two British scientists, Alan Baddeley and Graham Hitch. The model included some elements of the modal model by incorporating processing and memory, however, these were viewed as distinct. Three components were originally proposed to provide, respectively, temporary verbal storage, temporary visuospatial storage, and a coordinating or 'executive' function. The verbal storage component was originally named the articulatory loop, although subsequently it has been referred to as the phonological loop, and to consist of two components, a passive phonologically based temporary store and a mental verbal rehearsal process. The visuospatial component was originally referred to as the visuospatial scratch pad, but has subsequently been thought to comprise a specialist temporary memory store (or visual 'cache') for retaining the visual appearance of objects and scenes linked to a system (an 'inner scribe') for retaining, and possibly rehearsing, sequences of movements and more dynamic properties of scenes. The coordinating mechanism is referred to as the central executive, but in more recent years has been thought of as a collection of executive functions, including the coordination of multiple task performance, the allocation of attention, and the implementation of strategies and heuristics used in problem solving and mental discovery or when learning and retrieval are demanding of attention.

A schematic representation of a recent version of the multiple component model is shown in Fig. 3. The diagram illustrates the various specialist subcomponents in addition to the suggested flow of information from perception through a long-term knowledge base into working memory. In the latter sense the contents of WM comprise some of the products of perception and of subsequent mental processing. WM can then be considered as a multicomponent workspace that stores and manipulates information which has been activated from the long-term store, rather than as a primary route into long-term memory from the environment. Detailed discussion of the evidence for the features of the model can be found in Logie (1995), Miyake and Shah (1999) or Richardson et al. (1996).

This conception of WM fits neatly with the notion that people are strategic in how they use their cognitive functions. For example, recalling a list of words can be accomplished by rehearsing the sounds of the words, thinking about their meaning, organizing them into groups, or imagining the objects or concepts to which they refer. As a result, a phenomenon such as the phonological similarity effect (see above) with visually presented words only appears if the participants rely on mentally rehearsing the sounds of words prior to recall. There is evidence that some participants in these experiments spontaneously use other strategies that might rely on visual or semantic codes, and for them the phonological similarity effect does not appear (e.g. Logie et al. 1996). The same argument holds for other well-established phenomena such as the recency effect. Such strategic use of different aspects of WM would not arise if the material did not

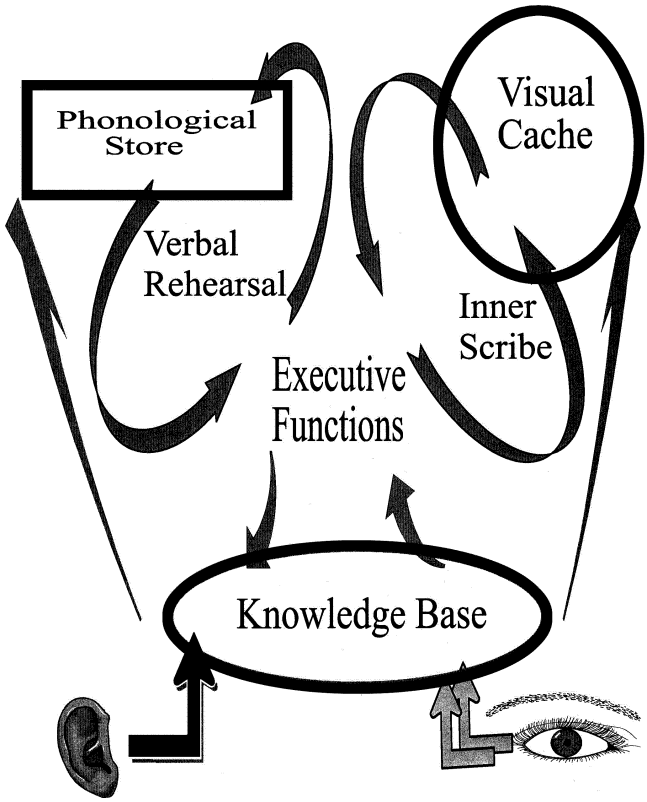

Figure 3
Schematic diagram of a recent version of the multiple component model of working memory

activate knowledge stored in long-term memory prior to it being held in WM for the period until recall is required.

6. *Applying the Concept of Working Memory*

The idea of a multiple component WM has proved useful in understanding some of the cognitive deficits suffered by individuals with brain damage as well as accounting for important aspects of healthy cognition. Discussed above are some of the insights into cognitive deficits arising from the use of the model, and how studies of patients with such deficits help the understanding of normal cognition. The application of WM theory to some aspects of mental discovery and creative thinking has also been mentioned. There have been several other areas of application, including the use of the phonological loop in explaining important aspects of vocabulary learning and mental arithmetic. For example, one feature of the phonological loop is that it can retain a recently heard word, mentally rehearse the word, and then generate the word as articulate speech. A number of studies have examined the ability of young children to repeat aloud new or nonsense words that they have just heard. Their ability to do so correlates well with their general competence in language and predicts their language ability several years later. Patients with severe verbal short-term memory deficits also appear to have difficulty acquiring new foreign vocabulary. These findings have led to the suggestion that the phonological loop is a cognitive function that has evolved to support the acquisition of language, and that has the emergent property that it can support temporary verbal memory.

Among the acquired skills that appear to depend on the phonological loop are counting and mental arithmetic. Keeping track of a running total in counting relies on subvocal rehearsal of the counting sequence, while keeping track of partial solutions in mental arithmetic again appears to rely on subvocalization of the subtotals involved, unless the solution is already well learned. For example, for most people the answer to $4+5=?$ would be available without the need for mental calculation, because it would be recalled as an arithmetic fact. However, the answer to $23+19+48+34=?$ would require a series of operations with temporary memory for the subtotals (held in the phonological loop) as each part of the sum is completed. Evidence for these conclusions arises from

studies in which volunteers perform counting or mental arithmetic while suppressing articulation through repeating an irrelevant word. Accuracy in counting and in arithmetic is disrupted by this procedure but is not disrupted by performing other irrelevant tasks such as tapping the table in a regular pattern or even when being presented with irrelevant speech sounds. The phonological loop therefore seems to support accuracy in counting and arithmetic.

The visuospatial resources of WM can support a wide range of other everyday cognitive tasks such as remembering where we are on the page while reading, maintaining and updating a representation of objects in our immediate environment, mental scanning to plan a route around a familiar town, or imagining the configuration of furniture in an empty room.

7. Conclusion

WM is an evolving concept which may be subject to change in the light of new evidence or questions as to its detailed characteristics. Nevertheless, it allows the interpretation of a great deal of evidence and phenomena from both laboratory tasks and from everyday experience.

Bibliography

Baddeley A, Della Sala S 1996 Working memory and executive control. *Philosophical Transactions of the Royal Society of London B* **351**: 1397–403

Beschin N, Cocchini G, Della Sala S, Logie R H 1997 What the eyes perceive, the brain ignores: A case of pure unilateral representational neglect. *Cortex* **33**: 3–26

Cornoldi C, Logie R H, Brandimonte M A, Kaufmann K, Reisberg D 1996 *Stretching the Imagination: Representation and Transformation in Mental Imagery*. Oxford University Press, New York

Della Sala S, Logie R H 1993 When working memory does not work. The role of working memory in neuropsychology. In: Boller F, Spinnler H (eds.) *Handbook of Neuropsychology*. Elsevier Science Publishers BV, Amsterdam, Vol. 8, pp. 1–63

Helstrup T, Logie R H (eds.) 1999 Working memory and mental discovery. Special issue of the *European Journal of Cognitive Psychology*. The Psychology Press, Hove, UK

Just M A, Carpenter P A, Keller TA 1996 The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychological Review* **103**: 773–80

Logie R H 1995 *Visuo-Spatial Working Memory*. Lawrence Erlbaum Associates, Hove, UK

Logie R H, Della Sala S, Laiacona M, Chalmers P, Wynn V 1996 Group aggregates and individual reliability: The case of verbal short-term memory. *Memory and Cognition* **24**(3): 305–321

Miyake A, Shah P (eds.) 1999 *Models of Working Memory*. Cambridge University Press, New York

Richardson J T E, Engle R W, Hasher L, Logie R H, Stoltzfus E R, Zacks R T 1996 *Working Memory and Human Cognition*. Oxford University Press, New York

Vallar G, Shallice T (eds.) 1990 *Neuropsychological Impairments of Short-term Memory*. Cambridge University Press, Cambridge, UK

R. H. Logie and S. Della Sala

Workplace Environmental Psychology

Workplace environmental psychology focuses on the role of physical environments in workers' experience, performance, and interaction in buildings that serve as workplaces, especially offices and factories. This subfield of environmental psychology (see *Environmental Psychology: Overview*) studies the influences of physical features of workplaces on workers and their relationships, and how workers and groups use, modify, and adapt to their environments. Also a subfield of industrial– organizational psychology (see *Industrial–Organizational Psychology: Science and Practice*), workplace environmental psychology overlaps with engineering psychology (see *Engineering Psychology*), but deals with larger-scale environmental features, including some studied by architectural psychology (see *Architectural Psychology*).

Workplace environmental psychology developed during the 1980s, though the underlying empirical research dates from the early 1900s (Sundstrom 1986).

1. Scope

Table 1 illustrates the scope of workplace environmental psychology. It emphasizes the individual and interpersonal levels of analysis, although the scope also extends to the organizational level (see Fischer 1997, Sundstrom 1987).

1.1 Individual Level of Analysis

The individual or psychological level of analysis focuses on objectively measurable ambient conditions—temperature, air quality, illumination, and sound—and features of workstations, or areas designed to accommodate one worker. It also focuses on features of supportive environments such as restrooms and walkways. Research and theory of the subfield concern the influences of these features on individual performance, job satisfaction, health, and other outcomes via psychological processes listed in Table 1.

1.2 Interpersonal Level of Analysis

The interpersonal or social–psychological level of analysis examines the social implications of psychological influences of workplaces, such as an adverse effect of noise on attention to social cues. The

Table 1
Scope of workplace environmental psychology

Levels of analysis	Facets of physical environments	Key processes and outcomes
Individual worker	Ambient conditions: ● temperature ● noise and sound ● air quality ● illumination	Psychological processes: ● arousal ● stress ● overload ● adaptation
	Workstations: ● seating ● equipment ● personalization	Outcomes: ● performance ● satisfaction ● health
	Supporting environments	
Interpersonal relationships and groups	Workspaces: ● status markers ● proximity ● enclosure	Interpersonal processes: ● communication ● regulation of interaction ● group formation
	Room layout Gathering places Group Workspaces	Outcomes: ● group cohesion ● team effectiveness

interpersonal level of analysis also applies to features of workspaces (individually assigned workstations) that communicate status, personal identity, or other meanings, and to features of work settings that limit or facilitate access among workers for casual, face-to-face interaction, meetings, or conferences. This level of analysis extends to elements of physical settings with roles in the development and effectiveness of work teams (interdependent collections of individuals who share responsibility for specific outcomes for their organizations; see *Teamwork and Team Training*).

Features of physical environment shown in Table 1 can serve multiple functions at both individual and interpersonal levels, and one function can be served by multiple features. For example, a fully enclosed ('private') office can symbolize rank, limit noise, and aid in regulating interaction. Each element of a workplace can have a role in a larger system, so analyzing them calls for a systems perspective.

1.3 Organizational Level of Analysis

Because of its roles in interpersonal relationships, the physical environment is a factor in organizational development and effectiveness, although relevant research is sparse. Facets of physical environments can support, reinforce, and complement organizational structure. Consequently, interventions to promote organizational development include changes in work environments. Designing buildings for work calls for mapping organizational structure to optimize the fit (Becker and Steele 1995).

2. State of the Art

Workplace environmental psychology, a relatively new subfield, draws on parent fields, environmental and industrial–organizational psychology, for research and theory. Empirical research on topics listed in Table 1 includes over 100 studies conducted in offices, factories, and other workplaces, and hundreds more in laboratories.

The research shows that the physical workplace, as a whole, represents a significant factor in individual job satisfaction, health, and performance, and group formation and cohesion. Findings on specific facets of the workplace are summarized in the following sections at the individual and interpersonal levels of analysis, along with accepted theoretical explanations.

3. Individual Workers and Ambient Conditions

Ambient conditions are the most intensively studied facets of the workplace. Evidence concerning individual satisfaction and health comes mainly from research in workplaces, including postoccupancy evaluations (assessments of new buildings using multiple measures; see *Architectural Psychology*). Evidence concerning performance comes mainly from laboratory experiments.

3.1 Temperature

Thermal comfort in workplaces depends on ambient air temperature and relative humidity, air movement,

radiant heat sources, activity level, clothing, and individual differences. In offices, temperature represents a common source of dissatisfaction with the physical environment. Up to 20 percent of office workers report feeling too cold or too warm at temperatures the majority finds comfortable.

Effects of temperature on performance reflect arousal (generalized physiological/psychological excitation) which increases during initial exposure to heat, and eventually falls below normal after continued exposure. The arousal hypothesis (see Sundstrom 1986) predicts optimum performance at moderate arousal and poorer performance if arousal is too high or low. Optimum arousal is higher for simple or practiced tasks, and lower for complex tasks and learning.

Consistent with the arousal hypothesis, laboratory experiments show that warm temperatures briefly boost performance of simple tasks. Other research shows no effect on performance of moderately demanding cognitive tasks. Exposure to heat for 30 minutes or more without rest periods degrades performance of complex and demanding tasks such as manual tracking, vigilance (monitoring displays for change or irregular signals), and dual tasks (simultaneous primary and secondary tasks). Decrements in dual tasks appeared in secondary tasks, often as errors of omission, suggesting that heat contributes to sensory/cognitive overload and prompts the usual coping responses: narrowing of attention and ignoring low-priority inputs. Early studies in factories found adverse effects of heat on output and accident rates ameliorated by ventilation.

3.2 Noise and Sound

Noise (or unwanted sound; see Kryter 1994) ranks high among sources of dissatisfaction with offices, particularly coworkers talking and telephones ringing. Dissatisfaction correlates with peak sound levels above background, not with background levels, consistent with evidence that noise is most irritating when irregular or unpredictable. Some research finds overall job satisfaction significantly and inversely correlated with annoyance by noise.

Loud background noise initially elevates arousal, which diminishes with adaptation (decrease over time in responsiveness to a feature of the environment, via perceptual change, altered metabolism, or active coping). Performing a complex or demanding task during noise slows adaptation. Laboratory research shows that adaptation can have psychological costs or adverse aftereffects in the form of temporarily impaired capacity immediately after working during loud, unpredictable, uncontrollable noise.

Individual health may suffer when prolonged exposure to noise is combined with an additional source of job stress. Working in a noisy environment correlates with physiological symptoms of stress among employees dissatisfied with their jobs, working on rotating shifts, or facing high job demands (see Evans et al. 1994).

Noise affects performance through arousal, overload, distraction, and masking of helpful sounds. Consistent with the arousal hypothesis, continuous loud noise over 100 decibels (dB) produces a brief increment in performance of some clerical tasks. Short-term exposure to noise has no effect on performance of some moderately complex tasks, such as reading. However, noise as loud as 90dB degrades performance of complex motor tasks, vigilance, mental tasks, and dual tasks. Decrements in performance of complex tasks intensify when the loud noise comes in irregular or intermittent bursts.

Music—sound intended for enjoyment—accompanies work in some offices and factories. Of published studies conducted in workplaces, many found no effect of music on performance. Those that assessed employees' attitudes about music found them generally favorable. A few studies reported positive effects of music on output or quality; none found decrements. Laboratory research found no effects of music on performance of clerical tasks or reading, but some beneficial effects on vigilance. Performance increments are consistent with the capacity of music to increase arousal or postpone its usual decline during prolonged, monotonous work.

3.3 Air Quality

Key variables related to air quality in workplaces include ventilation (flow of fresh air) and air pollution (gas, dust, mist, vapor, fiber, or other ingredients besides the natural constituents of air). In factories, a small fraction of complaints about working conditions concern air quality. In offices that allow smoking, a substantial fraction of nonsmokers report annoyance with tobacco smoke. In office buildings, poor ventilation has been associated with elevated rates of respiratory illness.

3.4 Illumination

Laboratory research indicates that the intensity of illumination influences performance through its effects on visibility of details. In tasks requiring visual discrimination, performance increases with intensity of lighting, but performance gains diminish as lighting intensity increases. As light becomes brighter, glare (uncomfortably bright light in the visual field) can pose problems. Research in offices has generally found employees satisfied with the lighting. Employees tend to prefer 'natural' light from windows, a preference that could reflect other desirable features of windows such as the view outside.

4. Individuals and Features of Workstations

Analysis of relationships between individuals and workstations relies on models of the human–machine system (see *Engineering Psychology*). Related research yields guidelines for designing displays, controls, and other features such as seating and positioning of equipment.

Research on a common type of equipment in many workstations, computer video displays, illustrates a typical sequence. When video displays entered workplaces, early studies found problems with visibility and complaints of symptoms such as headaches and eyestrain. New designs that reduced or resolved the problems soon appeared in offices; later research found fewer problems. Similar research/redesign/re-evaluation cycles occurred for other common elements of workstations, such as keyboards and chairs. Early studies of office chairs found complaints of discomfort and symptoms of strain. New chair designs incorporated guidelines based on the research; later research found most office employees satisfied with chairs that usually had padded seats, lower back support, and adjustable seat height.

Employees may personalize their workspaces by decorating them with photos, memorabilia, artwork, plants, and other unique objects—as allowed by company policies and features of the workspace. Some research in offices found satisfaction with the work environment positively correlated with personalization of workspaces.

5. Interpersonal Effects of Ambient Conditions

5.1 Noise

Noise interferes with conversation by masking the sounds of coworkers' voices. In noisy workplaces, coworkers move closer and shout to be heard. (They may also talk less.) In uncomfortably loud noise, people pay less attention to social cues, possibly as a response to a form of overload from noise. Participants in a laboratory simulation assigned lower salaries to fictitious job applicants in noise than in quiet, but this finding has not been replicated in actual workplaces.

5.2 Temperature

Research from social psychology has found a correlation of moderately uncomfortable heat and aggressive behavior, but this finding has no demonstrated parallel in workplaces. A multination study found national, ambient temperature positively correlated with reported role overload (Van de Vliert and Van Ypersen 1996).

6. Interpersonal Relationships and Workplaces

6.1 Workspaces as Status Markers

Characteristics of workspaces that signify rank, or status markers, include a private office, privileged location, exterior windows, floor space, guest seating, and others. Research in offices shows that employees consistently recognize status markers within their organizations. Status markers communicate the occupant's rank to visitors. Research in offices has found that among those with supervisory responsibility, status markers are associated with satisfaction with the work environment.

6.2 Proximity and Communication

In offices and factories where individuals have assigned workspaces, research has consistently found the frequency of casual or informal conversation correlated with the physical proximity of the employees' workspaces. The relationship of formal communication (required by work roles) with proximity of workspaces is more complex. Coworkers whose jobs call for frequent, face-to-face interaction tend to be assigned to workspaces near one another. As walking distance between workspaces increases, preferred means of communication shift from face-to-face conversation to telephone calls or written messages, except for certain kinds of meetings seen as requiring face-to-face interaction.

6.3 Physical Enclosure and Regulation of Interaction

Perceived privacy (selective control over access by others to oneself) in a workspace correlates with enclosure by walls or partitions. Privacy facilitates regulation of workers' interaction. Some research in offices found privacy correlated with job satisfaction among employees of similar rank.

Enclosure of workspaces is important to performance because it restricts accessibility, limits noise and visual distraction, reduces interruptions, and possibly limits overload. Enclosure may also limit social facilitation (motivating effect of the presence of others) and spontaneous conversation with coworkers located nearby. Research points to a complex relationship of enclosure with performance depending on task complexity and personality traits like ability to 'screen' social input (see Oldham et al. 1995).

In the 1970s and 1980s many organizations adopted 'open-plan offices,' with minimal enclosure of workspaces and few signs of rank. Reduced enclosure was expected to aid communication. Some evaluations of open-plan offices found no change in communication;

some found increases only in casual conversation; others found declines in confidential conversation, friendship opportunity, and supervisor feedback. Privacy declined, usually along with satisfaction with the environment.

6.4 Room Layout and Face-to-face Conversation

Research in offices finds greater 'psychological distance' in face-to-face meetings over a desk than without a desk, although meetings last just as long in either arrangement. Recent research found 'stand-up' meetings as effective, but faster and less personal, than 'sit-down' meetings.

In laboratory research on face-to-face meetings of three or more people, conversation follows lines of eye contact. Leaders emerge in places affording eye contact with the most members. Circular tables avoid this 'head of the table' effect.

7. Physical Environments and Work Groups

Informal groups tend to form among employees in adjacent or nearby workspaces in offices and factories. Work groups physically enclosed or isolated from their organizations develop cohesion (attraction/commitment of members to the group) often along with unique group norms (shared habits and attitudes).

In the 1990s, many organizations adopted team-based organizations (with groups as primary performing units). Facilities supportive of effective work teams ideally incorporate colocated workstations, shared group workspace, conference space, informal gathering places, and other features, depending on the type of group (Wineman and Serrato 1999).

Some organizations use 'virtual teams' (work groups with geographically dispersed members who collaborate via electronic communication; see Mittleman and Briggs 1999). Facilities for virtual teamwork include videoconference facilities and computer-based 'group decision support systems,' which can also be used face to face.

8. Workplace Trends and Evolution

Issues for workplace environmental psychology evolve along with the technology of workplaces. Factory automation requires fewer workers, who spend more time at computer workstations. Increasing use of manufacturing teams calls for more conference space. As technology makes knowledge work more portable, more of it occurs in workers' homes as 'telecommuting' (Christensen 1988). Work at home reduces office costs, but raises issues concerning home workspaces. Numbers of full-time office occupants may decline. Other trends point to increasing needs temporarily to accommodate part-time employees, contractors, visitors from elsewhere in the organization, and employees of supplier or customer organizations. Increasing use of work groups brings corresponding increases in needs for conference space in offices.

As workplaces evolve more toward conference centers, workplace environmental psychology may focus more on physical support for work groups and collaboration.

See also: Group Processes in Organizations; Industrial–Organizational Psychology: Science and Practice; Job Analysis and Work Roles, Psychology of; Job Design, Psychology of; Leadership in Organizations, Psychology of; Organizational Behavior, Psychology of; Performance Evaluation in Work Settings; Psychological Climate in the Work Setting; Stress in Organizations, Psychology of; Teamwork and Team Training; Well-being and Burnout in the Workplace, Psychology of

Bibliography

Becker F, Steele F 1995 *Workplace by Design: Mapping the High-performance Workscape*. Jossey-Bass, San Francisco

Christensen K (ed.) 1988 *The New Era of Home-based Work*. Westview, Boulder, CO

Evans G W, Johansson G, Carrere S 1994 Psychosocial factors and the physical environment: Inter-relations in the workplace. *International Review of Industrial and Organizational Psychology* **9**: 1–29

Fisher G N 1997 *Individuals and Environment: A Psychological Approach to Workspace*. De Gruyter, Berlin

Kryter K 1994 *The Handbook of Hearing and the Effects of Noise*. Academic Press, New York

Lipnack J, Stamps J 1997 *Virtual Teams: Reaching across Space, Time, and Organizations with Technology*. Wiley, New York

Mittleman D, Briggs R 1999 Communication technologies traditional and virtual teams. In: Sundstrom E et al. (eds.) *Supporting Work Team Effectiveness*. Jossey-Bass, San Francisco, pp. 246–70

Oldham G, Cummings A, Zhou J 1995 The spatial configuration of organizations: A review of the literature and some new research directions. *Research in Personnel and Human Resources Management* **13**: 1–37

Sundstrom E 1986 *Work Places: The Psychology of the Physical Environment in Offices and Factories*. Cambridge University Press, New York

Sundstrom E 1987 Work environments: Offices and factories. In: Stokols D, Altman I (eds.) *Handbook of Environmental Psychology*. Wiley, New York, pp. 733–82

Sundstrom E, Altman I 1989 Physical environments and work-group effectiveness. *Research in Organizational Behavior* **11**: 175–209

Van de Vliert E, Van Ypersen N W 1996 Why cross-national differences in role overload? Don't overlook ambient temperature. *Academy of Management Journal* **39**: 986–1004

Wineman J, Serrato M 1999 Facility design for high performance teams. In: Sundstrom E (ed.) *Supporting Work Team Effectiveness*. Jossey-Bass, San Francisco, pp. 271–98

E. Sundstrom

Workplace Safety and Health

1. Definition of Workplace Safety and Health

Workplace safety and health refer here to all practical actions taken at the shopfloor level to promote the safety and health of workers. Such actions should be theory-driven because accident theories help to pinpoint dangerous factors in the company, and they offer directions for practical solutions. The focus of these actions is the behavior and attitudes of the workers subject to the highest risk of occupational accidents and diseases.

2. Short History of Occupational Safety and Health

The history of occupational safety and health starts in the fourth century BC, when Hippocrates described the occurrence of lead poisoning among miners. Later, another Roman physician, Galen, wrote about various occupational diseases (Fraser 1989). However, the first written source to mention a disease as actually being occupational in origin was Agricola's *De Re Metallica* (1556), which ends with a note on diseases and accidents among miners. At that time a woman could be married to as many as seven miners, one after the other. Eleven years later (1567) Paracelsus published *Von der Bergsucht und anderen Bergkrankheiten*, describing the diseases of smelter workers and metallurgists (Hunter 1955).

Finally, in 1700, Ramazzini published a book entitled *De Morbis Artificum Diatriba*, in which he examined the health conditions of over 50 occupations. Because of his success, he is called the Father of Occupational Medicine.

Observations on occupational diseases did not lead to legislative action until 1802, when the Health and Morals of Apprentices Act was confirmed in England. This act was then improved by the Factory Acts in the middle of the nineteenth century. The first compulsory social security system for wage-earners was also set up about that time, in Germany in 1883. The Workmen's Compensation Act followed in Britain in 1897. It provided for automatic compensation to an injured worker. In 1919 occupational safety and health became a truly worldwide concern when the International Labour Organisation (ILO) was founded in Geneva to establish International Conventions and International Recommendations for the development of standards and procedures for occupational health and accident prevention (Fraser 1989).

3. Theories on Occupational Accidents

A feature common to different definitions of an occupational accident is the unexpected and unintentional nature of the incident. When physical injury is added, an occupational accident can be defined as an unexpected and unintentional series of events leading to the physical injury of a person at work. The other precondition for an occupational accident is a contractual relationship between the employer and the employee (Figlio 1985).

Until the end of the nineteenth century, people thought that accidents were brought on by their sins or were 'God's will.' This idea was based on the fact that most hazards were caused by natural phenomena. People had very limited means with which to resist the forces of nature. The onset of industrialization in the twentieth century meant that production moved from homes to factories. Since people had built these environments, most of the hazards were also man-made. Thus the idea of accident prevention fell on fruitful ground.

3.1 Theories on Accident Causes

The first scientific theory attempting to explain accidents was the theory of accident proneness. Greenwood and Woods (1919) insisted that 'accidents occur to a limited number of individuals who have a special susceptibility to accidents' (preface). Accident proneness was presumed to be a stable personality trait. The theory assumed that some people have an inherent tendency to be involved in accidents (Shaw and Sichel 1971).

The theory of accident proneness has been criticized on both theoretical and empirical grounds. The studies published have displayed conceptual confusion, however, surrounding the meaning of accident proneness. Accident proneness is understood as a unitary trait, as a general characteristic, and as an innate and unmodifiable characteristic, and it is thus used as an explanatory concept. The statistical criticism of the accident proneness theory culminates in the question of whether the entire population has the same risk to be involved in an accident or not. Because the accident proneness theory is based on an incorrect assumption of homogeneous exposure to risk, it should be substituted by the concept of accident liability, which is based on the assumption of the inequality of exposure (McKenna 1983).

According to the danger factor theory, an accident happens when a worker and a danger factor meet so that the worker injures him/herself. The accident is a

stochastic event, the probability of which is determined by the number of danger factors. The most critical danger factors are those with the highest energy content (Skiba 1973). This theory is mostly based on practical experience.

A more sophisticated version of the danger factor theory is the ergonomic approach to the causes of accidents. The ergonomic approach assumes that disturbances in the flow of information increase the risk of accident occurrence. The exchange of information between a worker and his or her environment, or between a worker and co-workers or supervisors are preconditions to accident avoidance (Saari 1984).

3.2 Theories on Risk Taking

The recent theories related to accidents are those on risk taking. They have been developed in research on traffic safety. Most of the authors have thought that their theories can be applied to risk-taking situations in worklife as well.

The risk compensation theory assumes that people have a constant level of risk that they are ready to accept. If the risk level decreases due to safety measures (e.g., seat belts in cars), people tend to adjust their goals (e.g., try to arrive sooner at their destination) so that the risk level returns to exactly the same level as before. The effects of safety measures are thus eliminated in the long run (O'Neill 1977). Because the assumption of exact compensation is criticized, the concept of risk compensation is often replaced by the concept of behavioral adaptation.

According to the risk homeostasis theory, the risk of traffic accidents is determined by the homeostatic system, which is analogous to the room temperature system. The main assumption of the theory is that the target level of risk is the only factor determining accident rates. Perceived benefits and costs of risk taking and cautious behavior are the factors controlling the level of target risk. The risk homeostasis theory argues that the effect of a safety measure depends on its motivational nature for the people who use it. The effect of a measure (e.g., mandatory seat belt use) which does not motivate its users will be eliminated in the long run as the drivers take more risks. However, safety can be enhanced by measures that motivate people to desire safety (Wilde 1982). If such motivation can be achieved, the effect of the measure may be greater than technical calculations show.

A complementary theory for the risk homeostasis theory is the risk motivation theory (Trimpop 1994). This model assumes that both personality and situational factors influence one's risk perception, which is divided into physiological, emotional, and cognitive components. Each one has its own target level of risk, which contributes to the total target level of risk. Based on the rational calculation of costs and benefits, a risk taker is motivated to act. These acts influence the environment, which then gives either positive or negative feedback. The actor can then perceive the risks in a new and better way.

The theory of zero risk received its name from the assumption that drivers of motor vehicles adapt to the risks involved in driving to the point that they do not generally feel any risk in a traffic situation or their subjective risk assessments approach zero. People avoid the feeling of risk just as they avoid pain. There is a risk threshold, however, above which risk is experienced as aversive. A driver tends to satisfy his or her motives in traffic, as well as in other areas of life. This need to satisfy motives pushes him or her to drive faster and more hazardously. Both excitatory and inhibitory motives influence the decision making of a driver. The most hazardous excitatory motives are 'extra motives' (i.e., those outside traffic, such as the saving of time and effort) which prompt the driver to increase speed. The increased risk of an accident is related to the strength of these extra motives (Näätänen and Summala 1976).

If the psychological portrayal of a human being with respect to accident theories is analyzed, a change can be seen. The first theories assumed that a human being is a passive accident victim who has an inherent tendency to be involved in accidents. In the more recent theories, a human being is an actor who actively searches for information from the environment. The theories, however, accept that a human being is not perfect and does make mistakes and errors. As technology becomes more reliable, a human being will be the greatest risk factor for him/herself (see *Health Risk Appraisal and Optimistic Bias*).

4. The Behavior Modification Approach

The approach of behavior modification is based on operant learning behavior. Behavior is seen as a function of its relation to various prior and current events. The preceding events are called antecedents, and the events following the performance are consequences. If an act has increased in rate after a particular consequence has reliably followed it, the consequence is reinforcing. But if the rate has decreased, the consequence is aversive. Last, if a behavior disappears when a prior reinforcing consequence is continued, the situation is called extinction (Sulzer-Azaroff et al. 1994).

4.1 Interventions

The training of workers is usually the first form of intervention included in a behavior modification program. The aim of training is to teach safe work habits. The training must be as illustrative as possible, for example, showing photographs of workers using

safety glasses. It is also possible to show unsafe scenes and ask workers to point out the unsafe actions. The important point in training is to repeat the safe habits in different ways for the workers. Several studies, however, have shown that training alone is not sufficient to improve safety performance (Sulzer-Azaroff et al. 1994).

The alternative method for traditional classroom teaching is an informational safety campaign. The first prerequisite for an effective informational campaign is to set a clear objective with a specific message that is positive, simple, plausible, and comprehensible. The campaign material should be designed to attract attention. Posters should be placed at crucial points of action, for example, on the wall of the canteen at a construction site. The effect of a campaign is usually short-lived. The types of intervention used in informational safety campaigns have not influenced the number or the seriousness of accidents (Saarela et al. 1989).

The other alternative method is to utilize small groups in accident prevention. This idea is based on experiences from quality circles, which are problem-solving groups largely used in the manufacturing industry. As in the case of quality circles, small accident prevention groups consist of members of the same organization meeting regularly to solve safety-related problems. In a Finnish shipyard the small groups had from 4 to 13 members, and they met 5 to 11 times during one year. The two tasks of the groups were to eliminate the obstacles to good order and safety and establish new housekeeping practices—behavior modification. The program was successful both in improving housekeeping and in reducing accidents (Saarela 1990).

Goal setting is the other main method with which to improve performance. The first requirement for successful goal setting is that the workers who are to be the subjects of the intervention participate in the setting of the goals. Difficult goals inspire better performance than easy goals do. The best goals are difficult, but achievable. The utmost goal should be divided into subgoals that are easy to understand. The first goals should be rather concrete, such as 'we should use hard hats all the time at the construction site' (Locke and Latham 1990).

4.2 Positive Reinforcement

Positive reinforcement refers to an increase in the frequency of a response following the presentation of a positive reinforcer. Whether or not a particular event is a positive reinforcer is determined empirically. The effectiveness of the reinforcer depends on several factors. Reinforcement given immediately after the response is more effective than feedback given after a long delay. The greater the amount of a reinforcer, the more frequent the response. Reinforcers that are highly valued lead to greater performance. A continuous schedule of reinforcement, where a response is reinforced each time it occurs, is the most effective (Kazdin 1975).

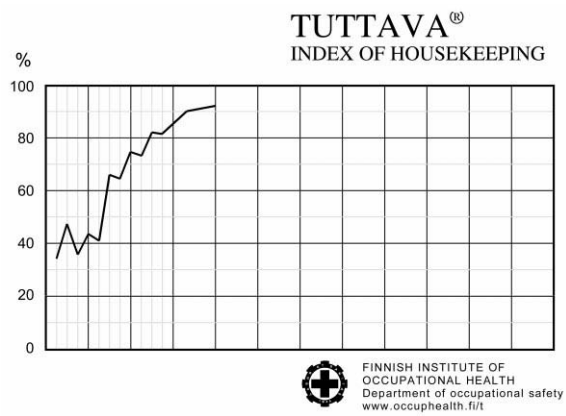

Figure 1
TUTTAVA information board

Positive reinforcement should be given soon after the observation of a desired behavior. It ought to be given regularly, for example, once a week. The feedback should be visible and given to all workers at the same time. One way to fulfill these requirements is to place a special information board on the wall so that a curve illustrates the development of the safety score on the experimental area from one observation to another. One example of such an information board is the TUTTAVA-board developed at the Finnish Institute of Occupational Health (Fig. 1), TUTTAVA being a Finnish acronym for a program to help improve production, quality, and safety.

One form of positive reinforcement is the use of a token economy. A token is a tangible object, for example, a lottery ticket, which can become a conditioned reinforcer. The use of tokens may shorten the delay between the behavior and the consequence, reinforce the behavior any time and any place, and provide a visible record for evaluation. The weakness of the token economy is that it pictures behavior as a simple series of stimuli and reactions. The token economy is, however, rather inexpensive for companies to operate (Brown 1978).

4.3 Measurement of Safety Behavior

It is useful to start a behavior modification program with accident analysis. The analysis of accidents during the previous three to five years pinpoints the hazardous sites of the workplace. It is worthwhile classifying accidents as unpreventable and preventable, that is, as whether or not workers could avoid a similar accident if they changed their behavior. The

researcher should then discuss the actions that could prevent accidents with the on-site personnel. Using the results of this analysis, the researcher collects 10–20 safety practices which should be phrased positively (Komaki 1986).

The next step is to prepare the measurement of safety performance. Generally an observation method is used for this purpose because the object of measurement is the external behavior of the subjects. The researchers should select observation periods at random (Tarrants 1980). When the observation is done at different times of the day, the workers behave normally and do not avoid the most hazardous tasks. The observation can be done from a single observation point or by walking through the observation area; the latter method is more common.

The observation sheet includes at a minimum a list of the desired safety practices. Usually it is constructed with only a dichotomized classification. The behavior of the observed worker is either safe or unsafe, as defined by the criterion of behavior included on the observation sheet. The third possible category is 'not observed,' under which the practice does not occur (Komaki 1986).

One example of an observation method is ELMERI (Fig. 2), ELMERI being the Finnish acronym for a workplace safety and health observation method. It has been developed through cooperation with safety authorities, safety managers, and safety representatives from selected workplaces. The observation method consists of 26 items, divided into 7 groups: (a) safety behavior, (b) order and tidiness, (c) machine safety, (d) industrial hygiene, (e) ergonomics, (f) walkways, and (g) first aid and fire safety. For each item, the observer makes a decision as to whether the item is correct according to the safety criteria defined in the handbook of the observation method. The alternatives for the observer are 'correct,' 'not correct,' or 'not observed.' The 'not observed' alternative is used, for example, when there are no machines with permanent means of access in the workplace. It takes about 15 minutes to observe one work station with the ELMERI method, and only a few hours of training are needed to reach an acceptable level of interobserver reliability. It has been shown that a high safety index from ELMERI is related to a low accident rate (Laitinen et al. 1998).

The walkthrough observation rounds should be repeated once a week during five to ten weeks. The suggested time for each round is 30–60 minutes. The minimum number of observations during one observation round is 100. The results of the observations are calculated as a safety score as follows:

$$\text{safety score} = \text{correct}/(\text{correct} + \text{not correct}) \times 100.$$

The higher the value of the safety score, the better the safety situation within the workplace (Komaki 1986, Tarrants 1980).

The reliability of the observation is based on interobserver agreement. Therefore, two observers should independently assess the behavior of the same workers. It is desirable to conduct reliability checks when both the observers and the observed workers are unaware that checks are being conducted. The reliability is calculated by dividing the smaller frequency by the bigger frequency and multiplying by 100. The reliability should be at least 80 percent to be acceptable (Kazdin 1975).

4.4 Experimental Designs

The effect of the intervention is demonstrated by the experimental design. It has two tasks. First, it should show that the change in behavior is permanent. Second, it should prove that the change in behavior is due to the intervention. There are three main types of experimental designs.

4.4.1 Reversal or ABAB design. The reversal design demonstrates the effect of the intervention by alternating the presentation and removal of the program over time. The purpose of the design is to demonstrate a functional relationship between the target behavior and the intervention. The first task in a reversal design is to measure the baseline rate of behavior, which describes the behavior in a normal situation before the intervention. The baseline period (referred to as phase A) is continued until the rate of the response becomes stable.

Then, in the experimental phase (referred to as phase B), the intervention is carried out. This phase continues until the behavior reaches a stable level or diverges clearly from the baseline level. Now the change in behavior is evident, but the cause of change is unclear.

In the reversal phase (phase A) the intervention is withdrawn. The target behavior usually returns to or near the original baseline level. The purpose of the reversal phase is to determine whether the behavior would have remained unchanged if the intervention had not been introduced. When the behavior reverts to the baseline, it is possible to reinstate the intervention (phase B). The design is called the ABAB design because the phases A and B are alternated (Kazdin 1975).

4.4.2 Control group design. Another way to demonstrate the effect of intervention is the classical experimental design with experimental and control groups. The intervention is done with one group (the experimental group) and not done with the other group (the control group). Immediately before and after the intervention the behavior of the workers is assessed. The basic requirement of the study design is the similarity of the groups. The best procedure to control for systematic differences between the groups

ELMERI observation form

Company: _____ Date: _____ Observer: _____
Work station: _____

Items	Correct	Total	Not correct	Total	Not observed	Total
1. SAFETY BEHAVIOR						
1.1. Use of PPE and risk taking						
2. ORDER AND TIDINESS						
2.1. Work tables						
2.2. Shelves						
2.3. Surfaces of machines etc.						
2.4. Waste container						
2.5. Floors and platforms						
3. MACHINE SAFETY						
3.1. Design, construction and condition						
3.2. Control devices and emergency stops						
3.3. Machine guards						
3.4. Permanent means of access						
4. INDUSTRIAL HYGIENE						
4.1. Noise						
4.2. Lighting						
4.3. Air quality						
4.4. Thermal conditions						
4.5. Chemicals						
5. ERGONOMICS						
5.1. Design of work station						
5.2. Manual material handling						
5.3. Repetitive work						
5.4. Physical variability						
6. WALKWAYS						
6.1. Design and markings						
6.2. Order and condition						
6.3. Visibility and lighting						
7. FIRST AID AND FIRE SAFETY						
7.1. Electric distribution box						
7.2. First aid kit						
7.3. Fire extinguisher						
7.4. Emergency exits						
	Total		Total			

ELMERI index = $\frac{\textit{correct}}{\textit{correct} + \textit{not correct}}$ x 100 = ---------- x 100 = %

Notes

Finnish Institute of Occupational Health 1997 Occupational Safety Inspectorate

Figure 2
ELMERI observation form

is to allocate the workers to groups at random (Kazdin 1975).

At the shopfloor level a random classification of workers is not always possible for production reasons.

Then two departments, which are as similar as possible, are selected as experimental and control groups. In addition to nonrandomness, overlapping of the information between the departments may be a problem in the comparison of departments.

4.4.3 Multiple-baseline design. The third experimental design to show the effect of interventions is the multiple-baseline design. The aim of this design is to demonstrate that the change in behavior is associated with the introduction of the contingency at different time points. This design requires the use of two or more interventions. After the baseline reaches a stable rate, one of the interventions is introduced, while baseline conditions are continued for other interventions. When the rates are stable once again, the second intervention is introduced. The effects of each intervention are established only after it has been introduced.

One problem that can occur with the multiple-baseline design is that changes in behavior take place too early. If behavior changes before the intervention is introduced, it is not clear whether the intervention is responsible for the changes or not (Kazdin 1975). Another problem is that the effects of the different interventions could be confused.

4.5 Analysis of Intervention Data

Intervention studies often require the use of statistical tests appropriate for single-case experimental designs. Conventional t and F tests can be used in the statistical analysis of intervention data if the data are not serially dependent. This practice requires that the autocorrelation (i.e., correlation between data points separated by different time intervals) not be significant (Kazdin 1984). The other statistical method is to use tests for time series. The best for a short time series ($n < 10$) is Kendall's tau, which analyses the mean trend.

It is possible to complement the quantitative data of the effects of the intervention study by qualitative data. Qualitative methods are the most useful for analysing the meanings and symbols of the culture or subculture. In an intervention study, the most useful qualitative method is perhaps a discourse analysis. With this method, a researcher can analyse the possible change in the attitudes of the workers participating in the intervention.

In the reports of behavior modification studies, graphs should be used to illustrate both the study design and the results. It is easy to show the development of a safety score with a line graph (Gillan et al. 1998). In some cases it is possible to use the same graph to give feedback to workers and to illustrate the results in the report, as the TUTTAVA curve does (Fig. 1).

High-quality intervention studies help to avoid the wasting of time, money, and effort. Shannon et al. (1999) have presented eight criteria for an intervention study of high quality. First, the objectives of the study or its underlying hypotheses are clearly presented. In the experimental design the subjects are randomly allocated to experimental and control groups. The study has external validity and therefore its results can be generalized to other workplaces. The outcome measurement with acceptable reliability and validity should be appropriate to the objective of the intervention. Qualitative data can be used to supplement quantitative data. Diffusion of intervention manipulation from the experimental group to the control group (overlapping) or improper randomization are the main threats to internal validity. Appropriate statistical analysis should be applied in the analysis of the data set. The conclusions should address the program objectives and the limitations of the study and be supported by the analysis. These criteria are useful both in the assessment of other studies and in the planning of one's own study.

5. Future Recommendations for Workplace Safety and Health

The fatality rate in Finland has fallen from 7.2 to 3.2 deaths per 100,000 workers from 1975 to 1994 (Statistics Finland 1997). Most of this positive development is due to technical solutions to prevention and improved safety inspections. Nowadays most of the technical solutions are in use and simple accidents can be prevented. Only the more complex accidents need to be tackled nowadays; they are due to the interaction between technical failure and human error.

Occupational accidents cause disturbances in production. When a serious accident occurs, all workers stop their work and go to help. According to Finnish calculations of accident costs, this participation forms the highest costs to the employer. Some customers therefore demand that the number of occupational accidents should stay under the acceptable limit.

Traffic research in Sweden has produced an idea called Vision Zero. According to this idea, no one should be killed or sustain an injury resulting in permanent impairment (Tingvall 1997). Some workplaces, at least in Finland, have already accepted the idea of Vision Zero. In other words, they follow the principle that no accident is acceptable and that all accidents are preventable.

These trends increase the expectations for safety modification behavior. Workplaces require more effective ways of changing the attitudes of workers to a more positive direction and activating them towards work safety. Alternatively, a deeper understanding of human behavior based on the advances of social and behavioral sciences makes it possible to respond to these increased requirements.

See also: Occupational Health; Stress at Work; Stress Management Programs; Worksite Health Promotion and Wellness Programs

Bibliography

Brown R L 1978 Adapting token economy systems in occupational safety. *Accidient Analysis and Prevention* **10**: 51–60

Figlio K 1985 What is an accident? In: Weindling P (ed.) *The Social History of Occupational Health*. Croom Helm, London, pp. 180–206

Fraser T M 1989 *The Worker at Work*. Taylor and Francis, London

Gillan D J, Wickens C D, Hollands J G, Carswell C M 1998 Guidelines for presenting quantitative data in HFES publications. *Human Factors* **40**: 28–41

Greenwood M, Woods H M 1919 *The Incidence of Industrial Accidents upon Individuals with Special Reference to Multiple Accidents*. Medical Research Committee, Industrial Fatigue Research Board, Report No. 4. Her Majesty's Stationery Office, London

Hunter D 1955 *Diseases of Occupations*. Hodder and Stoughton, London

Kazdin A E 1975 *Behavior Modification in Applied Settings*. Dorsey Press, Homewood, IL

Kazdin A E 1984 Statistical analyses for single-case experimental designs. In: Barlow D H, Hersey M (eds.) *Single Case Experimental Designs*. Pergamon, New York

Komaki J L 1986 Promoting job safety and accident prevention. In: Cataldo M F, Coates T J (eds.) *Health and Industry. A Behavioral Medicine Perspective*. Wiley, New York

Laitinen H, Rasa P-L, Räsänen T, Nykyri E 1998 An observation method as a tool for workplace risk assessment. *Technology, Law and Insurance* **3**: 57–61

Locke E A, Latham G P 1990 *A Theory of Goal Setting & Task Performance*. Prentice Hall, Englewood Cliffs, NJ

McKenna F P 1983 Accident proneness: A conceptual analysis. *Accidient Analysis and Prevention* **15**: 65–71

Näätänen R, Summala H 1976 *Road-user Behavior and Traffic Accidents*. North-Holland, Amsterdam

O'Neill B 1977 A decision-theory model of danger compensation. *Accidient Analysis and Prevention* **9**: 157–65

Saarela K L 1990 An intervention program utilizing small groups: A comparative study. *Journal of Safety Research* **21**: 149–56

Saarela K L, Saari J, Aaltonen M 1989 The effects of an informational safety campaign in the shipbuilding industry. *Journal of Occupational Accidents* **10**: 255–66

Saari J 1984 Accidents, and disturbances in the flow of information. *Journal of Occupational Accidents* **6**: 91–105

Shannon H S, Robson L S, Guastello S J 1999 Methodological criteria for evaluating occupational safety intervention research. *Safety Science* **31**: 161–79

Shaw L, Sichel H S 1971 *Accident Proneness*. Pergamon Press, Oxford, UK

Skiba R 1973 *Die Gefahrenträgertheorie*. Forschungsbericht Nr. 106. Bundesanstalt für Arbeitsschutz and Unfallforschung, Dortmund, Germany

Statistics Finland 1997 *Accidents at Work*. 1994 SVT Labour market 1997:1. Tilastokeskus, Helsinki, Finland

Sulzer-Azaroff B, Harris T C, McCann K B 1994 Beyond training: Organizational performance management techniques. *Occupational Medicine: State of the Art Reviews* **9**: 321–39

Tarrants W E 1980 *The Measurement of Safety Performance*. Garland STPM Press, New York

Tingvall C 1997 The Zero Vision. A road transport system free from serious health losses. In: von Holst H, Nygren Å, Thord R (eds.) *Transportation, Traffic Safety and Health*. Springer, Berlin

Trimpop R M 1994 *The Psychology of Risk Taking Behavior*. North-Holland, Amsterdam

Wilde G J S 1982 The theory of risk homeostasis: Implications for safety and health. *Risk Analysis* **2**: 209–25

S. Salminen

Worksite Health Promotion and Wellness Programs

1. Introduction and Historical Development

Organizational efforts to improve employees' health began with an almost exclusive concentration on worksite safety and progressed rather slowly to broader areas of concern, which today include health promotion. In the 1920s, concerns over employees' safety and health grew with industrialization. This was, to a large extent, due to organized labor's concerted efforts to improve unhealthy and dangerous work conditions. These concerns led to the development of a broad range of occupational safety and health initiatives which, for the most part, dealt with physical, chemical, and biological hazards in the work environment (Wolfe et al. 1994).

Health promotion initiatives gained a new lease of life in the late 1970s and 1980s (for a typical review, see Parkinson et al. 1982). Shain et al. (1986), based on a sample of historical case studies, suggested that health promotion programs are really just a modern expression of an old idea, namely employer paternalism. It was criticized that these programs communicated the view that deteriorating health is a moral failing for which the individual worker is responsible. With the move to a more comprehensive approach to worker health, the adoption of the health promotion philosophy by a number of employers in the 1970s represented a move away from the previous industrial medicine (occupational safety and health) orientation in three ways. Firstly, the primary concern shifted from preventing injury and disease to the active promotion of health. Secondly, the newly-adopted health promotion model heavily emphasized employee or personal behavior instead of workplace conditions as the primary focus. Thirdly, health promotion programs which traditionally had been initiated by

management and tended to be used primarily by white-collar staff were now extended to blue-collar workers as well (Wolfe et al. 1994).

2. Definition and Domain of Workplace Health Promotion Programs

The terminology used to label employee health promotion initiatives varies. Terms such as fitness programs, worksite health promotion programs, employee health management programs, health promotion programs, health enhancement programs, disease prevention programs, and wellness programs are only some of the most commonly used. Over the years, not only the name and content of workplace health promotion (WHP) programs have changed, they have also gradually broadened. For example, Elias and Murphy (1986) regard health promotion as 'art of a broad spectrum encompassing all health services and activities. This spectrum consists of health promotion, health maintenance, and health restoration.' Katzman and Smith (1989) defined health promotion programs as 'various combinations of in-house activities initiated by a company with the intent of enhancing the overall physical and psychological health of its employee population.' Kizer et al. (1992) refer to the definition of Parkinson et al. (1982) as they regard health promotion as 'a combination of educational, organizational, and environmental activities designed to support behavior conducive to the good health and wellbeing of employees and their families.' Wolfe et al. (1993) define WHP programs as 'ongoing organizational activities designed to promote the adoption of personal behaviors and organizational practices conducive to maintaining and/or improving employee physiological, mental or social well-being.'

The common vision of WHP programs is that of an integrated approach of a total package of activities, targeting not only the individual and organizational levels, but also environmentally-focused strategies and community-based resources. Such an interactive approach would offer the possibility of remedying problems of a physical and mental nature, promoting and maintaining good physical and mental health among employees, and developing an environment and tapping into the community to facilitate successfully sustaining positive changes achieved by the WHP programs.

3. Why Promote Health at the Worksite?

There are a number of advantages and disadvantages associated with the use of the worksite for health promotion initiatives. From the employees' perspective, they are convenient, affordable, and provide the opportunity for social support. These strengths may also be disadvantages where the timing of meetings may interfere with other valued activities, and/or individuals may not wish to disclose certain behaviors in front of their peers. For employers, potential benefits include better employee morale and public relations (Cataldo and Coates 1986, Everly and Feldman 1985); improvements in employees' performance; positive influences on morale, absenteeism, turnover, recruitment and productivity (Robison et. al. 1992); as well as reduced healthcare costs, particularly in the USA (e.g., Harvey et al. 1993), and reduced absenteeism (Bertera 1990). However, as many programs focus on individuals rather than organizational structures, there is a danger of 'victim blaming' and the development of ineffective programs.

One of the attractive aspects of WHP programs is their potential to reduce the rapidly increasing healthcare and health-related costs incurred by 'unhealthy' organizations. With WHP programs, substantial savings can be achieved by the organization. For example, cost savings may be associated with reductions in the cost of heathcare insurance, disability payments, and workers' compensation. In addition, a healthier workforce would absent themselves less frequently from work, thus reducing health-related costs. Of perhaps greater importance than a health promotion program's ability to control potential healthcare costs is its potential to improve employees' performance. Research has found that such programs can positively influence morale, absenteeism, turnover, recruitment, and productivity (Robison et al. 1992). Research has shown that generally, employees are satisfied with health promotion programs (Davis et al. 1984), and programs contribute to improvements in employee morale and job satisfaction (Pelletier 1991, Wolfe et al. 1993).

From a theoretical perspective, the worksite offers an excellent opportunity to maximize behavioral change through long-term individual programs, and the potential for environmental policies and manipulations to foster and maintain such change. However, not all programs have utilized the full potential of the worksite. For example, worksites which provide health screening clinics with no provision for followup or support through other environmental changes gained no advantage over similar interventions conducted in other settings (e.g., Erfurt et al. 1990). Accordingly, while the worksite provides an excellent setting for health promotion, its benefits can only be maximized through programs that utilize its environmental and social resources to the full.

4. The Environment—Challenges for the Working World in the Twenty-first Century

Clearly, a range of new challenges are confronting the world of work today. The world of work is undergoing major changes—a process which will continue for some time to come. A few random examples include genetics, the widespread of use of new technologies,

Table 1
Major types of differences in the workforce on a global basis

Supply of workers — Plentiful in some areas, scare in others
Skill level — High in some areas, low in other
Level of productivity — High levels lacking in some cultures
Level of motivation — Work ethic weaker in some culture
Attitudes towards leadership — Participating approach expected in some cultures, undesired in others
Attitudes towards achievement — High achievement not important in some cultures where poverty is not looked upon negatively
Reaction to discipline — less accepted among some cultures
Expectations of the job security — High levels of job security expected and required in some cultures
Attitudes towards status and rank — Significant class distinctions in some cultures
Attitudes towards women and people of various races — Sexism and racism more prevalent in some cultures

and demographic change. More general socioeconomic developments are also generating tremendous pressure on innovation, which also impacts on the field of work and health. Some of the key issues to be confronted are:

4.1 Globalization

As more firms become global in their operations, organizations will have to take into account the diversity in the workforce when deciding on wellness issues and programs. Some issues involved are differences in customs, language, and communication styles; differences in attitudes toward time; and the effects of labor laws, ethical standards, and the political climate.

The global workforce is anything but homogeneous. Employees' attitudes and value systems vary greatly in a wide variety of important ways. Table 1 summarizes some of the major differences in the workforces around the world. For example, about 7 percent of the Western population suffers from hypertension in such a measure that it needs to be treated in order to prevent cardiovascular problems. Another 7 percent has moderately heightened blood pressure which also needs to be attended to. Only a small number of people are being treated for this condition. Several studies which offered non-medication therapy (e.g., reduced smoking, losing weight, better nutrition, relaxation exercises, and stress management) reported success in the hypertension programs (Baier et al. 1992, Fielding 1982, Glasgow and Terborg 1988).

4.2 Unemployment

Millions of workers at all occupational levels have lost their jobs in recent years as a result of reorganizations and the inability of some industries to remain competitive in the world market. Retrenchment is traumatic, not only for the incumbent but also for those who remain in the job. Organizations involved in layoffs have found reduced morale, job commitment, and job involvement among their remaining workers (Seppa 1996).

4.3 Increasing Use of Technology

Of the 20 fastest growing occupations, one-half are related to the health and computer fields. Bezold et al. (1986) have cited data indicating that computers and robots will probably affect 7 million factor jobs and 38 million office jobs. The projected effects of this will include job displacement, deskilling, and lower paying jobs.

4.4 Changes in Employment Practice

With the advances in personal computers, telecommunications, and fax machines, flexibility has come to characterize the workplace. In recent years, flexibility in work scheduling (e.g., short-term and part-time employment) has become more widespread. Telecommuting—the decentralizing of work—is already in effect at more than 600 US companies. Increases in productivity of up to 30 percent have been reported (Becker and Steele 1995).

4.5 Change in Organizational Size and Structure

Due to the huge changes in company structures, new production and management concepts are introduced. The 'classic' company with its stable and predictable structures is heading towards a new flexibility involving ever smaller core workforces, new forms of work, and greater mobility between jobs. Overall, there is an increase in the number of small- and medium-sized enterprises. These companies are subject to special conditions, particularly with regard to occupational safety and health and workplace health promotion.

4.6 Changing Demographics

The demographic development must be observed as regards its effects on workplace health promotion.

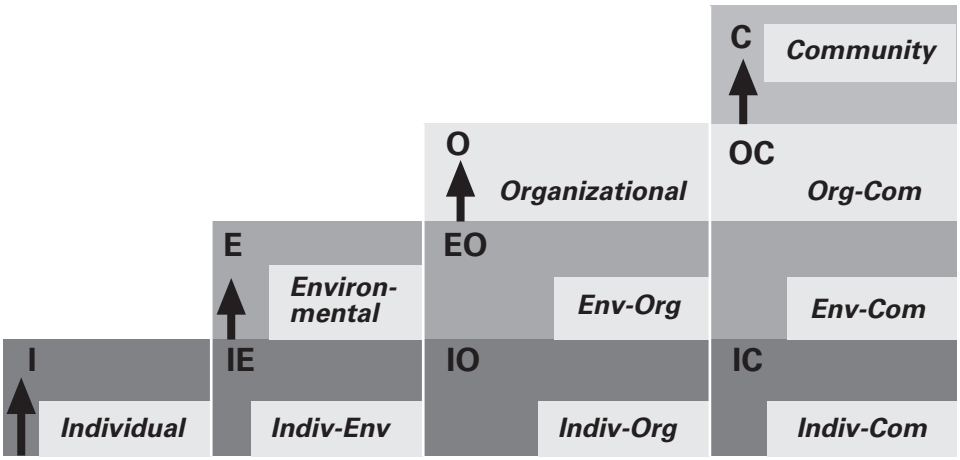

Figure 1
Workplace health promotion grid: a multitier strategy

4.6.1 Gender composition. According to Fullerton (1985), of every 10 new jobs between 1985 and 1995, six would be filled by women, and the proportion of women would continue to increase to 46 percent of the workforce. Because of role demands and constrained occupational opportunities, this trend may have an adverse impact on physical and mental health.

4.6.2 Ethnic composition. Another emerging demographic trend is that of an increased participation of minority-group members in the workforce. Data from the USA identifies a strong association between ethnicity and health status. Blacks have higher age-adjusted mortality rates for heart disease, cancer, liver disease, diabetes, and pneumonia than whites. They are also more likely to die from violence (Markides 1983). In the UK, rates of ill health and mortality amongst ethnic minorities differ from those of the native (white) population, but there are also marked differences between them (see Balarajan and Raleigh 1993). As differences in health behaviors are associated with both race and gender, health promotion programs will do well to take these factors into consideration. According to Narang and Murphy (1994) and Doyle (1991), ethnic minorities experience more problems in gaining access to health services such as cancer screening and antenatal care, and have lower survival rates following diagnosis of cancer or AIDS, than their white counterparts (Primm 1987). They are also more likely to experience environmental stressors such as poverty, poor housing, and poor working conditions.

4.6.3 Age composition. The change in health structure has, on the one hand, resulted in a higher proportion of older employees in the working population: on the other hand, this change also requires longer working lives to safeguard the 'agreement between the generations'. This demographic development means that the group of older employees is growing in relative terms. While at risk from poorer health and from sometimes outdated qualifications, these older employees offer a large but frequently untapped potential of experience and qualifications.

5. The WHP for the Twenty-first Century—A Multitier Strategy

It is generally agreed by health promotion researchers that in the twenty-first century, mental health problems in the workplace are prevalent. These problems have high human and economic costs (both direct and indirect) and are related to overutilization of general medical care and increasing healthcare costs. Unfortunately, mental health problems are not yet sufficiently addressed at the worksite. Today, the objective of WHP is therefore more than just protection against accidents and occupational diseases. It includes all the activities that serve to maintain and promote the physical and mental performance and motivation of the employees. A 'biobehavioral,' multidimensional perspective to WHP programs is recommended—that of examining occupational-related stress and/or health factors in combination with life-style factors (e.g., sedentary behavior, smoking, poor diet). Such life-style factors affect the health of the individual worker and collectively affect the wellbeing and

productivity of the entire workforce. These concerns should be addressed collectively through comprehensive disease prevention and/or health promotion efforts.

In modern companies, strong involvement of the employees in this process is essential, taking into consideration the changes in the communication and information conditions in society and within companies. Successful WHP is not feasible without intensive communication and constructive cooperation at all levels. The future success of organizations is dependent on having well-qualified, motivated, and healthy employees. Worksite health promotion has a significant role to play in preparing and equipping people and organizations to face these challenges. In view of the multiple issues faced by organizations in the twenty-first century, a multitier approach is suggested. Workplace health promotion programs need to be implemented strategically and with some permanence. To ensure that this happens, there is a need to disregard the notion of creating 'programs,' which are temporary by definition. Change must begin with policy; thus, a need to involve both the micro (individual and environmental) and macro (organizational and community) levels are essential. These will include:

1. *Individually-focused strategies* of health risk appraisal, clinical preventive services for employees and their dependants, and employee life-style change (including physical exercise, good nutrition, avoidance of tobacco products and excessive alcohol consumption, safe driving practices, safe sexual behavior, safe work practices, and the application of sunscreens when exposed to direct sunlight).
2. *Environmentally-focused strategies* of ergonomic workplace design, industrial hygiene and occupational safety (e.g., removal of chemical and environmental hazards, maintenance of good indoor air quality), and effective facilities design and management practices.
3. *Organizational efforts* to promote employee wellbeing and healthful workplaces, for example, policies that establish smoke-free worksites; efforts to facilitate flexible work schedules; provision of employee health benefits and insurance plans that incorporate disease prevention and health promotion services; managerial efforts to cultivate socially supportive and responsive organizational climates and to offer worksite violence prevention programs; and corporate compliance with occupational safety and health regulations.
4. Beyond the workplace, *community programs and resources* that enhance the health of employees and their dependants, including the development of 'healthy city' programs and effective healthcare delivery services at local and regional levels. The community also includes employee groups, the private sector, healthcare providers, health insurance companies, unions, educational institutions, and the federal and state governments.

6. *The Future of Work and Health Programs*

Based on the above, one of the goals of future WHP programs should be understanding the rapidly changing nature of work and its impact on the health of employees. In addition, sophisticated program evaluation activities need to be incorporated in the planning phase of such programs. Employees must have a primary role in the conceptualization of the research questions and methodologies in order to anticipate limitations and ensure that the findings can be assimilated into corporate policies. A change towards a culture in which it is seen as normal and positive to be concerned with improving health and maintaining good health at the workplace causes a change of attitude and social norms with regard to healthy behavior. An important part of the health program is that information reaches employees through different channels, thus raising the level of knowledge of health promotion and, importantly, influencing the existing culture with regard to health-promoting activities at the workplace. Evaluation should be an integral part of the program. In order to attain good evaluation, the aims of the program should be described in terms of clear, quantifiable, and feasible goals. These goals should be formulated in such a way that they are recognizable and in accordance with the needs of the participants.

Many organizations are currently investing considerable resources in WHP, with the expectation that it will produce substantial organizational dividends. This presents a window of opportunity for health and organizational scientists to develop WHP programs based on sound scientific principles that will meet both public health goals and organizational objectives. This will extend the benefits of WHP programs to reach not only the working population, but also the population at large.

See also: Community Environmental Psychology; Health Education and Health Promotion; Health Interventions: Community-based; Minority Access to Health Services: United States; Occupational Health; Workplace Environmental Psychology; Workplace Safety and Health

Bibliography

Baier C A, Grodzin C J, Port J D, Leksas L 1992 Coronary risk factor behavior change in hospital personnel following a screening program. *American Journal of Preventive Medicine* **8**: 115–22

Balarajan R, Raleigh V S 1993 *Ethnicity and Health in England*. HMSO, London

Becker F, Steele F 1995 *Workplace By Design: Mapping the High-Performance Workscape*. Jossey-Bass, San Francisco, CA

Bertera R L 1990 Planing and implementing health promotion in the workplace: A case study of the Du Pont company experience. *Health Education Quarterly* **17**: 307–27

Bezold C, Carlson R J, Peck J C 1986 *The Future of Work and Health*. Auburn House, Dover, MA
Cataldo M F, Coates T J 1986 *Health and Industry: A Behavioral Medicine Perspective*. Wiley, New York
Davis M F, Rosenberg K, Iverson D C, Vernon T M, Bauer J 1984 Worksite health promotion in Colorado. *Public Health Reports* **99**: 538–43
Doyle Y 1991 A survey of the cervical screening service in a London district, including reasons for non-attendance, ethnic responses and views on the quality of the service. *Social Science and Medicine* **32**: 953–7
Elias W S, Murphy R J 1986 The case for health promotion programs containing healthcare costs: a review of the literature. *American Journal of Occupational Therapy* **40**: 759–63
Erfurt J C, Foote A, Heirich M A, Gregg W 1990 Improving participation in worksite wellness programs: Comparing health education classes, a menu approach, and follow-up counselling. *American Journal of Health Promotion* **4**: 270–8
Everly G S, Feldman R H L 1985 *Occupational Health Promotion: Health Behavior in the Workplace*. Wiley, New York
Fielding J E 1982 Effectiveness of employee health improvement programs. *Journal of Occupational Medicine* **24**: 907–16
Fullerton H N Jr 1985 The 1985 Labor Force: BLS's latest projections. *Monthly Labor Reviews* **108**(11): 17–25
Glasgow R E, Terborg J R 1988 Occupational health promotion programs to reduce cardiovascular risk. *Journal of Consulting and Clinical Psychology* **56**(3): 365–73
Harvey M R, Whitmer R W, Hilyer J C, Brown K C 1993 The impact of a comprehensive medical benefit cost management program for the City of Birmingham: Results at five years. *American Journal of Health Promotion* **7**: 296–303
Katzman M S, Smith K J 1989 Evaluation of occupational health promotion programs. *Employee Assistance Quarterly* **4**: 27–46
Kizer K W, Folkers L F, Felten P G, Neimeyer D 1992 Quality assessment in worksite health promotion. *American Journal of Preventive Medicine* **8**: 123–7
Markides K 1983 Mortality among minority populations: A review of recent patterns and trends. *Public Health Reports* **98**: 252–60
Narang I, Murphy S 1994 An assessment of ante-natal care for Asian women. *British Journal of Midwifery* **2**: 169–74
Parkinson R S et al. (eds.) 1982 *Managing Health Promotion in the Workplace: Guidelines for Implementation and Evaluation*. Mayfield, Palo Alto, CA
Pelletier K R 1991 A review and analysis of the health and cost-effective outcome studies of comprehensive health promotion and disease prevention programs. *American Journal of Health Promotion* **5**(4): 311–15
Primm B 1987 AIDS: A special report. In: Dewart J (ed.) *The State of Black America 1987*. National Urban League, New York
Robison J I, Rogers M A, Carlson J J, Mavis B E 1992 Effects of a 6-month incentive-based exercise program on adherence and work capacity. *Medicine and Science in Sports and Exercise* **24**(1): 85–93
Seppa N 1996 Keeping a job can be stressful. *APA Monitor* **27**: 38
Shain M, Suurvali H, Boutilier M 1986 *Healthier Workers: Health Promotion and Employee Assistance Programs*. Lexington Books, Lexington, MA
Wolfe R, Parker D, Napier N 1994 Employee health management and organizational performance. *Journal of Applied Behavioral Science* **30**(1): 22–42
Wolfe R, Slack T, Rose-Hearn T 1993 Factors influencing the adoption and maintenance of Canadian, facility-based worksite health promotion programs. *American Journal of Health Promotion* **7**: 189–98

J. T. S. Ho

World Systems Theory

World Systems Theory (WST) is the intellectual descendant of Marxist theories about imperialism and of Latin American Dependency Theory. It uses a holistic methodology that claims to incorporate all the social sciences and to look at the entire globe as a unit. It is meant to overcome the narrow specialization by discipline and geographic area that has yielded much specific knowledge but few general theories to explain the extreme differences in wealth between various parts of the world. WST does this by saying that under capitalism poorer or 'peripheral' countries have been systematically underdeveloped by the rich and powerful 'core' countries, and as such, they stand little or no chance of escaping without some sort of world revolution. In between peripheral and core countries are 'semiperipheral' ones that act as a kind of global middle class stabilizing the world capitalist system. On the other hand, it is precisely in semiperipheral countries that the strains and contradictions of the world system are most evident, and therefore, that is where revolutions against the world system are most likely. Because of its ideological radicalism, WST has been somewhat marginalized, but if divorced from its ideological bias, it has the potential to become a valuable tool for the analysis of social change.

1. *The Intellectual History of the Theory*

The theory was developed primarily by the sociologist Immanuel Wallerstein (1974, 1979) in the 1970s, and it quickly became popular in sociology and anthropology, as well as among some historians and political scientists trying to explain both historical inequality between nations and the international situation of the 1970s, particularly the war the United States lost in Vietnam. But it has older antecedents.

The notion that capitalist growth and therefore successful industrialization can only occur through the ruthless exploitation of an impoverished working class was Marxist. Though Karl Marx himself concentrated primarily on the internal workings of class conflict within rather than between different societies, he also stressed the global nature of capitalist expansion and its effects on colonial or semicolonial societies. This became an important aspect of later Marxist theories of international relation. Rosa Luxemburg's and V.I.

Lenin's work as well as that of many non-Marxist critics of late nineteenth and early twentieth century European imperialism stressed the internationally exploitative nature of capitalism and its seeming need to control ever growing markets and resources. Lenin's theory of *Imperialism* (1939), written in 1916, claimed that World War I was the logical outgrowth of the competition between capitalist powers for colonies, because without dependencies, the internal contradictions of capitalism would produce growing worker immiseration and eventually, revolutions. The consequence of this theory, if it is to be believed, would be that revolutions outside the core capitalist societies can deprive the core of easily exploitable human and material resources, and therefore lead to worldwide revolution. This belief remains at the heart of Wallerstein's ideas, except that whole societies act as members of a class. Class conflict within societies remains important, but its ultimate direction is determined by the global struggle between exploiting and exploiting societies. For Wallerstein, no revolutionary movement within a single society, or even several of them, can succeed in bringing down capitalism without a global revolution.

The theory's other intellectual parent was early twentieth century right wing nationalism. Joseph Love's work (1996) traces the development of nationalist economic theories of underdevelopment from the German Friedrich List in the nineteenth century, through the Romanian fascist-corporatist Mihail Manoilescu, and to the Latin American right in the 1930s. The relative backwardness of these societies (they were what Wallerstein would call 'semi-peripheral' at the time they produced the relevant theorists) was explained mostly by the dominant market position of the most developed capitalist powers of those times, first England, and later, the major West European powers and the United States. The implications were that in order to develop, the poorer societies had to shield themselves from the world market, to somehow exclude the world system. Nationalism justified such policies as essential in order to protect the self-interest and integrity of the less developed nations.

World War II discredited fascist-corporatist notions of development, but the theory retained its attractiveness, and it gradually migrated leftward. In Latin America, Raul Prebsich, Celso Furtado, and Fernando Henrique Cardoso became the chief spokesmen for what came to be called 'Dependency Theory.' The explanation of backwardness remained essentially the same, that the developed capitalist powers impeded the development of those who were more backward, but it no longer seemed necessary to carry out protectionist theories within the framework of an authoritarian, ultranationalist, fascist-corporatist state (Cardoso and Faletto 1979). It should be noted that Cardoso, who eventually became Brazil's president, abandoned most of the implications of his theory long before achieving power in his own county. Meanwhile, the theory became increasingly popular within the revolutionary left.

The problem was that a high degree of protectionism in an increasingly powerful and advanced world economy required heavy state intervention to block market forces and popular demand for the consumer goods produced by the world capitalist system. Therefore, some sort of authoritarian, state centralizing mechanism was required to carry out such policies rigorously enough to permit escape from the world system. If the theorists were right, the capitalist core, now led by the United States, would use its economic, diplomatic, and even military power to force open recalcitrant nations. This was the theoretical explanation of American intervention in Vietnam and throughout the Third World. Thus, to motivate their people and mount a defense, the leaders of anti-capitalist regimes would have to use nationalism as the glue binding their people together in struggle, because they would be unable to provide the kind of consumer rewards offered by world capitalism. For those hostile to right wing, authoritarian nationalism, this led inescapably to the conclusion reached by Wallerstein and his followers, that Leninism in its most autarkic, closed Stalinist or Maoist forms was the only way to protect antisystemic nations until the world revolution achieved its goal of destroying the world capitalist system and replacing it with a more equitable, redistributive world socialist system.

This created an intellectual quandary recognized by many of WST's early followers who did not relish the ideological conclusions reached by Wallerstein and those who remained with him. Could something be saved from the comparative, holistic methodological approach advocated by Wallerstein without having to accept the extreme leftism he and his closest followers, such as Christopher Chase-Dunn (1989) were also promoting? The problem became all the more acute in the late 1980s and 1990s as communism collapsed and Marxism–Leninism was largely discredited throughout the world.

2. *The Uses of World Systems Theory*

Despite its shortcomings a world systemic approach to the study of social change remains a powerful antidote to narrow specialization. It also addresses problems that will remain with us far into the future, and it offers insights absent in other theoretical models.

2.1 *World Systems Theory is not a Substitute for but a Supplement to Economics*

First, we are obliged to admit that WST cannot claim to be a substitute for sound economics. While many development economists recognize that perfectly free

trade policies by poorer economies may indeed have detrimental effects, few if any accept the notion that autarky is a reasonable development policy, whatever the type of political regime. The import-substitution policies followed by many Latin American countries from World War II until the 1980s, as well as by India and many other countries, slowed economic growth, led to entrenched, corrupt domestic interest groups that blocked innovation and reform, and ultimately provoked economic and financial crises that threatened economic catastrophe. On the other hand, trade protectionism and developmental policies that fostered export-led industrialization in East Asia produced robust economic growth (Ray 1998). But beyond this, the truly extreme forms of withdrawal from the world capitalist system, that is the autarkic communist states, have all produced economic disasters. Furthermore, economic historians have shown that in the nineteenth century, as least, if not before, capitalist development did not rely on colonial markets, so that the economic argument in favor of imperialism turned out to be largely specious (Bairoch 1993). Yet, no matter what their strengths, there remains something unsatisfying about purely economic models of economic development.

In the hands of comparativists who place the societies they analyze in a global context and understand capitalism's transformative power, but who also recognize the importance of domestic social structures and the variations caused by specific historical experiences, world systemic ideas help account for some of the differences in rates of economic modernization. Peter Evans (1995) has done this in explaining the success of certain state led industrialization projects. Though somewhat suspicious of core power, especially American hegemony in the contemporary world capitalist system, Evans recognizes that development within that system remains possible for those who know how to combine effective bureaucratic leadership with local entrepreneurial talent, and who remain oriented to succeeding within the world system and using its advanced technology and market opportunities rather than revolting against capitalism.

Historical studies such as the one by Resat Kasaba (1988) also show how WST can provide important insights as long as it is not used dogmatically. He proves that in the mid-nineteenth century local merchants within the Ottoman Empire were able to use growing trade relations with the capitalist world to strengthen their own position against both the Ottoman state and the interests of the hegemonic core powers. But later, the Ottoman bureaucracy, helped by some powerful foreign, core financial institutions regained its strength, and this reduced the economic power of the domestic merchant class. Far from being 'compradors,' that is, subservient tools of the capitalist core (which is what conventional WST and most Marxists contemptuously call them), these local entrepreneurs might have spearheaded an economic and social modernization that would have made Ottoman society more open, tolerant, and progressive. The world capitalist system did indeed play an important role in peripheralizing late Ottoman society, but not in the expected way, and only because it was successfully manipulated by an increasingly nationalistic, Turkish, anticommercial state elite.

2.2 World Systems and Premodern Social Formations

Wallerstein believes that the capitalist world system began in the sixteenth century, with the European explorations and expansion around Africa to Asia and across the Atlantic to the Americas. But some of those inspired by WST have extended his model to precapitalist social and economic formations, arguing that notions of core and periphery based on economic and cultural exchanges explain social dynamics in a much wider array of societies than merely modern ones.

Janet Abu-Lughod (1989) argues that there was a large pre-European world system in the thirteenth and fourteenth centuries, linking China, India, and the Middle East, and that it was only bad luck that subsequently weakened this particular system and allowed Europe to become dominant. Her book, while historically somewhat at odds with Wallerstein, is faithful to his general world view.

Thomas Hall (1989) has used his research on the Southwestern United States from the fourteenth to the late nineteenth centuries to show that its many societies were interconnected in ways analogous to those specified by Wallerstein long before the arrival of Europeans, and such links continued in modified form under Spanish and Mexican rule. Looking at the shifting geographical distribution of ancient cores and peripheries has led Hall and others to reinterpret archeological data to better understand prehistorical social change.

3. Theoretical Limitations and Methodological Promises

WST's methodological approach reaffirms an old but often neglected concept. All societies have operated within larger systems, and without placing them in that context, our understanding of how they work is severely limited. There can be no general laws of social behavior that do not take this into account.

Beyond this, WST has shed light on the nature of modernization, even though that is a term it avoids in order to show it rejects the functionalist theories proposed by 'Modernization Theory' in the 1950s and 1960s. WST has shown that there is indeed a capitalist world system, and its rules, largely determined by the core powers, affect everyone. Furthermore, this capi-

talist world is itself an evolving, changing structure. It is subject to periodic waves of prosperity followed by economic downturns. Even economists need to take such large systemic factors into account in order to perfect their models. So do the other social scientists who, by and large, limit comparisons to discrete cases rather than incorporating system wide variables.

3.1 Limitations

A world systemic approach to social science, however, must admit the limitations of the theory as it now stands. Aside from its tendency to neglect economic theory, mentioned above, WST has also relegated technology and science to an almost insignificant role. Thus, Wallerstein, Abu-Lughod, and most others refuse to admit that Western European capitalism's early dynamism depended on the religious and scientific transformations that occurred during the Renaissance, Reformation, and Enlightenment. Max Weber remains a better guide to this process of change than any of the contemporary practitioners of WST, and by neglecting him, and the cultural issues he raised, they have weakened their paradigm.

Ideas generated by the core, from science and technology to various political ideologies such as nationalism and Marxism, have affected the rest of the world as much if not more than the economic and military power of the core. These ideas have been transformed in many ways within the periphery, and have stirred up particularly disruptive ideological currents of resentment and antimodernism in what Wallerstein has called the semiperiphery. There is no theoretical reason why WST could not use its model to expand our understanding of how this has worked. There is no need to continue to adhere to the simplistic materialism so many of its practitioners now follow.

It is precisely because the core, still led by the United States, retains its cultural vitality and historically unusual ability to innovate that the world capitalist system is still thriving. The repeated assertions that world capitalism is or will soon be in some sort of terminal 'crisis' has marred WST's ability to understand the present, just as the failure to take into account the cultural specificity of Western Europe has weakened its explanation of the rise of the modern world. But these theoretical deficiencies should not obscure the merit of using a world systemic approach.

3.2 Methodological Promises

When he first presented his theories, Wallerstein declared himself to be a disciple of Fernand Braudel and the French *Annales* school. Indeed, Braudel's vast comparative sweep in which entire regions, and ultimately the globe, are studied as a system of interdependent parts, is entirely compatible with WST.

Many more recent first-rate social scientists have adopted a similar, world systemic approach, though usually without acknowledging their methodological kinship with it. For example, the historian Anthony Reid (1988/1993) has presented a daringly new vision of Southeast Asian economic, cultural, and political history by putting it into the kind of comparative and global perspective that WST has always championed.

Ken Jowitt's analysis of the failed attempt to create a Leninist world system (1992) also adopts a world systemic approach without admitting it. He points out that the communists actually were more successful at creating a cultural and ideological world system of their own, with a powerful core struggling to keep its peripheries under control, than at creating a system based on economic links. He then predicts that the collapse of Leninism will have a globally destabilizing effect because it will leave the resentful losers in the world capitalist system without any constructive or predictable revolutionary alternative.

Sociological theories about the spatial distribution of cities, or trade, or global networks could all benefit from adopting a world systemic perspective. So far, judging from the literature in the articles on these subjects assembled by Neil Smelser and Richard Swedberg in *The Handbook of Economic Sociology* (1994) this has not happened. But if it did, it could yield important results. On the other hand, some geographers interested in these issues as well as by problems posed by environmental degradation already have used a WST in a fruitful way (see Taylor 1993).

Other examples and promising avenues could be spelled out, but the important point is that WST's potential is far from exhausted. Flexibly used, it still has lot of new insights to offer, and it is likely that in the early twenty-first century, in response to increasing globalization, continuing inequality, and uneven prosperity, it will undergo a creative rebirth that will take it beyond its sectarian origins.

See also: Capitalism; Center–Periphery Relationships; Class: Social; Dependency Theory; Economic Sociology; Finance, Geography of; Financial Institutions in Economic Development; Globalization, Anthropology of; Globalization: Political Aspects; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Historical Sociology; Imperialism, History of; Imperialism: Political Aspects; Marxism in Contemporary Sociology; Modernization, Sociological Theories of; Nationalism, Sociology of; Revolutions, Theories of; Social Change: Types

Bibliography

Abu-Lughod J L 1989 *Before European Hegemony: The World System A.D. 1250–1350*. Oxford University Press, New York

Bairoch P 1993 *Economics and World History: Myths and Paradoxes*. University of Chicago Press, Chicago

Cardoso F H, Faletto E 1979 *Dependency and Development in Latin America*. University of California Press, Berkeley, CA

Chase-Dunn C K 1989 *Global Formation: Structures of the World-Economy*. Basil Blackwell, Cambridge, MA
Evans P B 1995 *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press, Princeton NJ
Hall T D 1989 *Social Change in the Southwest, 1350–1880*. University Press of Kansas, Lawrence, KS
Jowitt K 1992 *New World Disorder: The Leninist Extinction*. University of California Press, Berkeley, CA
Kasaba R 1988 *The Ottoman Empire and the World Economy: The Nineteenth Century*. State University of New York Press, Albany, NY
Lenin V I 1939 *Imperialism, The Highest Stage of Capitalism: a Popular Outline*. International Publishers, New York
Love J L 1996 *Crafting the Third World: Theorizing Underdevelopment in Romania and Brazil*. Stanford University Press, Stanford, CA
Ray D 1998 *Development Economics*. Princeton University Press, Princeton, NJ
Reid A 1988/1993 *Southeast Asia in the Age of Commerce, 1450–1680*. Yale University Press, New Haven, CT, 2 Vols.
Smelser N J, Swedberg R (eds.) 1994 *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ
Taylor P J (ed.) 1993 *Political Geography of the Twentieth Century: A Global Analysis*. Belhaven Press, London
Wallerstein I M 1974 *The Modern World System*. Academic Press, New York
Wallerstein I M 1979 *The Capitalist World Economy: Essays*. Cambridge University Press, Cambridge, UK

D. Chirot

World Trade Organization

1. Background/Objectives

The World Trade Organization (WTO) was formed in 1995 as an outgrowth of earlier arrangements set up by the governments of 23 countries under the General Agreement on Tariffs and Trade (GATT) in 1947. The objectives set for the WTO largely mirror those which emerged in the postwar years for the GATT. Prior to 1947, when the GATT was formed, there had been substantial discussion between governments of the appropriate postwar financial and trade arrangements which would govern the global economy. In the area of international finance, the International Monetary Fund had been formed with a mandate to maintain exchange rate parities, and to do this by extending loans to countries coming under exchange rate pressure. These, in turn, were to be repaid when such pressures subsequently receded. In the area of international development and infrastructure financing, the World Bank had been founded with the objective of providing financing for roads, bridges, dams, and other forms of infrastructure through concessional financing.

These arrangements were put in place at the 1944 Bretton Woods Conference but in the trade area things were left hanging for some time further. By early 1946, however, there was movement to also negotiate a charter for a wide-ranging new international trade organization which would lay down rules concerning the conduct of international trade and allowable interventions in trade by national governments. This negotiation was initiated in Havana, but as 1947 advanced it became clear to a subset of countries that this negotiation was in some difficulty in being too broad, and having a focus which was excessively interventionist. In the fall of 1947, a number of countries therefore decided to engage in a separate negotiation which, among other things, would provide a framework for the recording of the results of an initial tariff cutting conference which had been planned. The GATT was born of this separate negotiation, and as a temporary arrangement pending the ratification of the wider International Trade Organization (ITO).

The ITO charter was not ratified by the US Congress in 1951 on the grounds that it contained too many compromises with free trade principles. So the GATT, born as a temporary arrangement, was left as the major arrangement in the postwar years between national governments available to guide the conduct of global international trade policy. At first seen as a minimalist organization, with only limited expectations of it, it subsequently grew through a series of successful negotiating rounds which were fueled in the1950s by the creation of the European Community and, to a lesser degree, the entry of Japan into the GATT in 1955. This Round-based liberalization reached its high point in the 1960s with the Kennedy Round (1964–7) of trade negotiations, during which there were substantial reductions in tariff barriers, largely attributed at the time to the cooperation negotiated with the framework of the GATT. The later Kennedy Round (1973–9) was to a large degree an echo of the Kennedy Round, and the ambitious and still controversial later Uruguay Round (1986–94) was driven by the perceived need to broaden trade negotiations beyond now low tariffs.

2. Basic Elements of the WTO Structure

In essence, the GATT as it evolved after 1947 and became permanent in the World Trade Organization in 1994 embodies three key elements which each contribute to oversight of the conduct of world trade and the ways in which national governments allow each other to intervene in that trade. The first is a series of mutually agreed limits (or rules) on the use of trade-restricting measures by national governments. These were initially set out in the General Agreement itself, but have been interpreted and elaborated on both by further negotiation and through the dispute settlement process within the GATT. The second is progressive liberalization, to be achieved through negotiations on trade barriers and reductions in trade barriers. These negotiating Rounds, guided by reciprocal exchanges

of concessions between member governments, occurred in the 1960s and 1970s (in the Kennedy and Tokyo Rounds), and in the 1980s through to 1994 (in the Uruguay Round).

The third element in the GATT structure, and subsequently incorporated into the WTO, is an institutional presence maintained through the operation of its Secretariat. Importantly, the Secretariat largely oversees the conduct of dispute settlement panels, which, in effect, define the enforcement mechanism for the rules of the trading system within the WTO system. These dispute settlement panels have accelerated greatly in their use in recent years, culminating in decisions in a number of key areas, including recently in banana trade (see the statement by the WTO Director General to the WTO General Council, March 8, 1999).

3. *Broader Objectives*

While these three elements of rules, progressive liberalization, and dispute settlement define the structure of the WTO, at a broader level its objectives reflect a firm belief that it is important to restrict the use of new trade measures globally so as to prevent outbreaks of retaliatory trade wars that some people think characterized the 1930s. At this time, there was a rapid increase in trade barriers in the United States mandated under the 1930 Smoot-Hawley Tariff Act, and through the retaliation which followed in Europe. In the 1930s depression, world trade volumes shrank rapidly, some suggest by as much as 70 percent over a two year period (Kindleberger 1973). A central objective therefore set for the GATT was to prevent the world economy from going back to anything reminiscent of the 1930s, fueled by breakdown in international cooperation in the trade area.

Another broad objective relates to the potential gains from progressive global trade liberalization. It is widely believed in many circles that international trade represents the engine of growth for many economies around the world. While this is still contested by some, those involved with the WTO often point to the much more rapid growth of trade than income in the world economy since 1945, and the accelerating growth of trade in recent years. In the mid-1990s, for instance, trade growth was approximately three times income growth in the global economy, and many attribute this to the liberalization achieved within the trading system under the WTO in these years (*UNCTAD Handbook of International Trade and Development* 1999, 1993, 1995).

While these two overarching objectives, preserving international cooperation and going further with new liberalization, are generally agreed to be those underpinning the GATT/WTO system, there has been much debate over other objectives for global trade policy. One concerns the relationship between trade and development and whether the rules of the trading system through WTO should be designed so as to help fuel the growth and development of developing countries, and facilitate their wider participation in the trading system, and hence the achievement of growth and development. The developing countries generally believe that growing trade is an important part of their developmental process, and have aggressively pursued their interest in this set of issues in the WTO in recent years. At the same time, many cling to longstanding positions that the WTO (and the GATT) has *de facto* operated as a rich man's club, and that poorer countries do not actively participate in it because they are effectively excluded from it because of the structure of GATT bargaining (Srinivasan 1998).

Another area which has taken on a high profile recently concerns the use of trade policy to achieve nontrade objectives. These considerations came into play in the Uruguay Round with the introduction of intellectual property into the trade negotiations. The idea was that international norms and standards could be agreed in intellectual property protection which could then be enforced through dispute settlement allowing retaliatory actions covering trade in goods. The discussion of international trade policy instruments in this case was not so much to achieve liberalization and new growth of trade as to enforce arrangements in a nontrade area through the use of trade threats and sanctions. Some trade economists have expressed concerns over these developments since they raise the prospect of higher rather than lower trade barriers from the WTO because of the linkage of trade to nontrade areas.

Beyond intellectual property, a similar structure of linkage prevails in other related proposals which are now before the WTO. These were at the heart of the conflict at the 1999 Seattle Ministerial meeting last autumn, and involved the proposed use of trade measures to achieve environmental objectives and the use of trade measures to enforce labor standards (UNCTAD 1999).

In summary, the WTO reflects a structure based on three pillars which is widely believed by many economists to have been extremely successful in most of the postwar years in both promoting world trade and preventing any increases in trade barriers. The broadening of the WTO agenda in the Uruguay Round, however, has converted it into much more than a trade organization. It now deals with the new issues which were introduced into the Uruguay Round (services, investment, and intellectual property) and threatens to broaden further to also cover environmental issues and labor standards (Crome 1998).

4. *Current Issues in the WTO*

The WTO currently faces a range of issues, many of which have their origins in previous negotiations and, to a large degree, the Uruguay Round, while

others span the openness of WTO processes (transparencies), the political acceptability of WTO decisions and NGO participation, even to the survivability of the WTO itself. Shorter term, key issues lie in the area of agriculture; a key sector for global trade, insofar as it was effectively excluded from the GATT/WTO system in 1947 by provisions included in GATT Articles 11 and 16. GATT Article 11 specifies that all use of import quotas is to be banned, but special provisions apply to allow these to be used in agriculture; GATT Article 16 has similar provisions applying to export subsidies. The reasons for these exemptions for agriculture reflect a belief in 1947 that the agricultural sector had been vital to the war effort and cultural identification with the family farm and the agricultural sector as central to national identity justify protection. In addition, some countries (such as Japan) held the view that national security objectives implied that countries should be allowed to be self-sufficient in agriculture through trade protection.

For all these reasons, agriculture was initially outside the framework of the GATT structure and subsequent developments in the trading system including a special GATT waiver for the United States in 1955 and the 1957 European Common Agricultural Policy further intensified this situation (Ritson and Harvey 1997). Europe, which was a net food importer of large proportions in the 1950s, first excluded imports under the Common Agricultural Policy. Then with high prices in their domestic support programs and stockpiles, Europe became a significant net food exporter. This led to global export subsidy wars and major conflict more generally in agriculture in the 1980s.

The Uruguay Round represented the first attempt to deal with agriculture in a substantive way in the GATT, and introduced new disciplines. These left the original concessions in place but put in place an overarching structure which aimed to reduce levels of support to farmers, to cut export subsidies, and to convert all existing trade restrictions into transparent border measures (tariffication). Agriculture, however, was left as an area after the Uruguay Round for further negotiation (Hathaway 1987). Approaching the Seattle Ministerial meeting, agriculture and services largely define the so-called 'built-in agenda' for future negotiations under the WTO. (For information on built-in agenda, see WTO Secretariat's note. Implementation of the provisions for Review, Future Work or Negotiations in the WTO agreement and related decisions and Declarations of 7 May 1998 (WTO document WT/L/271).) New conflicts cover such matters as to whether or not there should be a ban on export subsidies, a ban which given the European structure of agricultural supports would be difficult for them to implement. The issue of and the rationale for agricultural protection on existence value grounds is a further source of conflict.

Another area of current discussion is textiles and clothing. In the Uruguay Round, the developed countries agreed to terminate their use of textile and clothing trade restrictions under the Multifibre Arrangement (Trela 1994). These restraints go back to the early 1960s, when a series of short-term agreements between the US and Asian exporters became subject to mutually agreed trade restraint on the grounds of allowing time for domestic adjustments which would accompany free trade to take place in the United States. The initial time agreed was one year, but this was first renewed and then converted into longer term agreements of five years, and subsequently multilateralized to include other OECD countries in 1974 (Whalley and Hamilton 1996).

In 1994 under the Uruguay Round decisions, the developed countries agreed to terminate the Multifibre Arrangements (MFA), but at the same time have been effectively tightening their trade restrictions they are committed to eliminate, despite an agreed phase-out plan for the MFA. Many developing countries are therefore skeptical over what may happen in the year 2004. Developing country negotiators have compared the developed country position on MFA elimination as akin to walking off a cliff (Nguyen et al. 1993). How this is to be resolved is yet another issue.

New issues also pose future issues for the WTO despite their inclusion in the Uruguay Round agreement. These include services, investment, and intellectual property. In the case of intellectual property, the phase-in period is long with special delays on implementation given for certain types of developing countries. Its full impact will be felt in the twenty-first century. In services, a whole structure of restraint on trade measures was agreed, but very little substantive liberalization to include in the structure, and hence, what happens next is very much an issue of debate.

Future issues also include the trade and environment and later standards issues which have been at the heart of a lot of the discussion and debate in Seattle. In the environmental area, the issue is the extent to which trade-restricting measures can be used to achieve environmental objectives. The trade and environment conflict was greatly accelerated in the early 1990s because of a dispute between the United States and Mexico over the use of trade restrictions on tuna which had been caught in ways which generated large incidental kills of dolphin (the activities of the GATT (1990)). The measures mandated by US courts in response to actions taken under US law by various environmental NGOs in the US were subsequently not deemed to be GATT compatible by a GATT dispute panel (Seattle Ministerial Conference—Trade and Environment: Proposals on the Possible Implications for Developing Countries). The GATT and subsequently the WTO acquired a reputation among environmental groups as being environmentally unfriendly and hostile to environmental considerations. This perception persists and was in part behind some of the conflict which prevailed in Seattle. How this is

all to be worked out in the future is now a central issue for the WTO, who in turn would perhaps like to see the trade and environment issue diffused with the environmental problems dealt with in another forum.

The trade and labor standards is also another high-profile issue because of the claim in some circles and primarily in Europe that imports occur from factories where child labor is used in unacceptable conditions: long working days, poor lighting, children who could be subject to deformities as a result of their activities in carpet factories and other locations. The free traders, however, see this as a dangerous form of new protectionism which will be supported by labor groups in OECD countries who have an incentive to show opposition to imports from countries with low wages.

All of these, then, are current issues, and current matters for the GATT and the WTO. The overarching issue, however, remains one of the contribution of the WTO to global economic performance. In the 1930s, Ragnar Nurkse posed the question of whether or not trade was the engine or handmaiden of growth; did trade drive growth as being a major, perhaps even sole, determinant within national economies or did trade merely grow because the wider economies themselves were growing? This question posed by Ragnar Nurkse has never been adequately resolved in the research community, but strong views prevail on these matters nonetheless. Free traders generally believe that strong economic performance requires an open trade regime. This is despite the performance in Japan in the 1950s and 1960s and in some other countries that have adopted more protectionist trade for them, and the fact that some countries have liberalized and not grown. On the other hand, it seems clear that countries such as Korea who have generally moved towards strong outward orientation and Taiwan have grown rapidly as a result of their trade orientation.

The stance one takes on the issue of the role of trade in future global growth largely determines how much importance one attaches to the WTO. As a vehicle for preventing the world economy reverting to the 1930s, as part of the mechanism which avoided the major cataclysmic recessions in the 1940s and 1950s comparable to the 1930s, and as a vehicle for fueling substantial liberalization through the 1960s and 1970s, it seems clear that the GATT and the WTO have much pride that they can claim in their actions. However, issues such as whether or not the system favors rich countries over poor countries, whether the system is inequality enhancing at the expense of the poor, and whether the system is closed and highly nontransparent and hence not in accordance with principles of good governance, remain. How these issues are to be worked out will largely determine both the acceptability and the profile of the WTO in the years which follow. The attempts that follow in the next few years to move forward from the Seattle Ministerial of December, which in the short term has proved to be destructive for the WTO, will largely determine both the acceptability and the profile of the WTO in the future.

See also: International Business; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; International Trade: Geographic Aspects

Bibliography

Crome J 1998 *The Present Outlook for Trade Negotiations in the World Trade Organization*. World Bank, Geneva

Hathaway D E 1987 *Agriculture and the GATT: Rewriting the Rules*. Institute of International Economics, Washington, DC

Kenen P B (ed.) 1994 *Managing the World Economy: Fifty Years After Bretton Woods*. Institute for International Economics, Washington, DC

Kindleberger C P 1973 *The Great Depression, 1929–39*. Allen Lane Penguin Press, London

Nguyen T, Perroni C, Wigle R M 1993 An evaluation of the Draft Final Act of the Uruguay Round. *Economic Journal* **103**(421): 1540–9

Ritson C, Harvey D 1997 *The Common Agricultural Policy*. CABI International, Wallingford, UK and New York

Srinivasan T N 1998 *Developing Countries and the Multilateral Trading System: from the Gatt to the Uruguay Round and the Future*. Westview Press, Boulder, CO

Trela I 1994 Phasing out the MFA in the Uruguay Round: Implications for Developing Countries. Report for an UNCTAD project on the impact of the Uruguay Round on developing countries

UNCTAD Handbook of International Trade and Development 1999, 1993, 1995. UNCTAD, Geneva and New York

UNCTAD 1999 *Seattle Ministerial Conference—Trade and Environment*: *Proposals on Their Possible Implications for Developing Countries*, Mimeo

Whalley J, Hamilton C 1996 *The Trading System After the Uruguay Round*. Institute of International Economics, Washington, DC, pp. 178–80

Wyndham White E 1964 GATT and the Kennedy Round. *Review of Industry and Technology* **May**

J. Whalley

Wright, Sewall (1889–1988)

Sewall Wright was born December 21, 1889, in Melrose, Massachusetts, and died in Madison, Wisconsin, March 3, 1988. Before his third birthday, the family moved to Galesburg, Illinois, where his father, Phillip G. Wright, had accepted a teaching position at Lombard College. Wright graduated from Lombard College in 1911 and, after a year at the University of Illinois, went to Harvard University as a student of William E. Castle. After receiving the

doctorate in 1915, he was employed as senior animal husbandryman by the US Department of Agriculture (USDA) in Beltsville, Maryland. He remained there until 1925, when he accepted a position as Professor at the University of Chicago. Upon retirement in 1954 he moved to the University of Wisconsin, where he stayed for the rest of his long life. Wright remained intellectually and physically active until his death at age 98, caused by his slipping on an icy sidewalk during one of his regular long walks.

1. *Early Life*

Wright was a precocious child. He could read and kept a diary before starting school. He also learned arithmetic and astonished his first-grade teacher by extracting cube roots. Later he and his brothers printed the first poems of Carl Sandburg, who was a student in their father's composition class. At Lombard he took several courses from his father, including calculus. But, despite his later reputation as a formidable mathematician, Wright never took advanced courses; his mathematics was largely self-taught. His interest in genetics was kindled by reading Punnett's article on Mendelism in the Encyclopedia Britannica.

Between his junior and senior years of college, he used his mathematical and surveying skills to work on a railroad construction in the Standing Rock Reservation in South Dakota. It was an exciting time for him. In his 90s he enjoyed telling about the rough frontier life with cowboys, Indians, and outlaws, and he still recalled a large number of words in the Sioux dialect. This work was cut short by a lung infection. While confined to a caboose, he used the time to study Tait's book on quaternions. As a result of the damaged lung, he later had trouble getting life insurance, something that he found increasingly amusing as he continued to live far beyond the usual life-expectancy.

Wright's father, Philip, was a man of many talents. He taught courses in astronomy, mathematics, economics, surveying, English composition, and physical education. He loved poetry and music, and was disappointed that his son did not take to them. Later he moved to the Brookings Institute where he published a number of books on economics. Wright had two brothers who also attained distinction. Quincy became a leader in international law and Theodore was an aeronautical engineer and acting president of Cornell University.

A full length biography of Wright is available (Provine 1986), with full details of his life and scientific accomplishments. His scientific papers span the period from 1912 to 1988, an astonishing 76 years. His bibliography comprises more than 200 articles, including a four volume set published during his retirement (Wright 1968–78).

2. *Animal Breeding*

Wright's thesis at Harvard analyzed interaction of coat colors in guinea pigs, a subject which he pursued throughout his USDA and Chicago years. He enjoyed guinea pigs and spent many hours taking care of them. Only the non-availability of an animal house prevented his continuing this after his move to Wisconsin. While at Harvard, he also began his theoretical studies of population genetics. On moving to the USDA, he inherited a colony of guinea pigs that had been inbred for many generations. His analysis of the effects of inbreeding and crossbreeding set a standard for careful analysis; the paper could be written today with hardly a word changed. During this time he invented the inbreeding coefficient, with a simple algorithm now standard textbook material (Wright 1922a). Also during this period, he formulated the basic principles for analysis of quantitative traits. Wright, along with R. A. Fisher (see *Fisher, Ronald A (1890–1962)*), converted animal breeding into a quantitative science.

Wright was especially proficient in manipulating correlation and regression coefficients. One of his earliest papers (Wright 1917) foreshadowed the analysis of covariance. His major statistical contribution was developing the method of path analysis. This is a way of measuring the relative contribution of different causal paths to a quantity of interest, such as size or performance. This was presented in a diagram in which causal paths were designated by arrows and correlations by curved, two-headed arrows (Wright 1921). A path coefficient, measuring the influence of a path, is a standardized partial regression coefficient. Wright formulated a set of simple rules that made the analysis easy to apply. This method enjoyed great popularity among livestock breeders, although recently it has been largely replaced by more sophisticated, computer-driven techniques that permit analysis of large bodies of data, and measurement of the precision of the estimates.

The technique is most useful for nonexperimental situations and in recent years has found its greatest use in the social sciences. Wright himself used the method in the late teens to analyse 510 correlations of 42 variables involving corn yields and prices and hog production. He could not get this monumental paper published because an animal husbandryman was not supposed to know about economics. It was not published until several years later, and only after the intervention of Henry A. Wallace, son of the then Secretary of Agriculture.

3. *Evolution*

After moving to Chicago, Wright changed his emphasis from animal breeding to evolution, although he still retained his interest in the physiological genetics of guinea pigs. Three people, R. A. Fisher (1929) and

J. B. S. Haldane (1932) in Britain, and Sewall Wright (1931) in the USA are mainly responsible for the 'new synthesis,' the welding of Mendelian inheritance and Darwinian natural selection into a coherent mathematical theory. Wright's particular contribution was the 'shifting balance theory.' He thought that a major difficulty in evolution was the frequent inability of a well-adapted organism to evolve into a better adapted one without passing through maladaptive intermediates. In Wright's metaphor, a population at an adaptive peak could not move to a higher peak without passing through less well adapted intermediate stages.

Wright thought that a way out of this dilemma was to have a population broken up into partially isolated subgroups. Within a small subgroup random gene-frequency changes can occur and occasionally these may produce a combination of genes that is close to that of the genotypes at the higher peak. Then, selection could take over and carry the population to the peak. This highly adapted subpopulation would then expand and send migrants to upgrade the whole population, and the process could start over.

Wright's theory became very popular among biologists. Its emphasis on gene interactions and its offering a way of putting together adaptive combinations of genes that might be deleterious in other combinations had a wide appeal. On the other hand, the theory was less popular with mathematical theorists. Wright's model required a delicate balance of population size, selection, and migration, which would be rare. It also required that the population spend much of its time, during the trial and error period, in a less than optimally adapted state. Finally, there is a question as to whether the theory is really needed; perhaps mass selection usually suffices for adaptive evolution.

Wright's theory was steadily refined during the rest of his life, but the essence remained unchanged. The theory became almost an obsession with him and he delivered many talks on the subject and wrote paper after paper. It is fair to say that, as to the importance of his theory, the jury is still out. Wright's latest paper (1988) was spirited defense of his idea, while still noting the importance of the contributions of Fisher and Haldane, and of Kimura (see *Kimura, Motoo (1924–94)*) who argued for the importance of mutation-driven, neutral evolution of molecular differences. He concluded that 'all are valid.'

4. *Retirement Years*

After moving to Wisconsin, Wright spent the first few years finishing the analysis of his extensive guinea pig studies. These were masterful analyses, but by that time the molecular revolution had changed the nature of developmental genetics, and Wright's methods were out of vogue. When this was completed, in his 70s he began his monumental four volume set, the last volume of which was published when he was 88 (Wright 1968–78). Although his shifting balance theory is conspicuous throughout, the book is far more than this. It summarizes, in four densely packed volumes, not only his life's work, but the whole field of population genetics. For the last decade of his life Wright's eyesight became progressively worse, and he used a reader that projected a greatly enlarged picture onto a television screen. He continued to read and to enjoy conversations, but in his last years he found it increasingly difficult to keep up with molecular advances and turned more to reading history and biography.

As a person, Wright was quiet, shy, and introverted. He had no small talk, and was hard to converse with. Yet, once the conversation turned to a subject on which he had an interest—and there were many—he could go on at great length. His lectures were always full of details and regularly ran overtime. He had a charming sense of humor and was excessively generous with his time. His students and colleagues held him in great affection.

5. *Social Sciences and Philosophy*

Wright's greatest influence in the social sciences is his method of path analysis. It was taken up by several social scientists, such as Dudley Duncan in sociology and Arthur Goldberger in economics, and has become a standard part of the statistical methodology. One of Wright's earliest applications of path analysis used adopted children to separate genetic from environmental influences on IQ scores, and there have been many follow-up studies extending his methods to more complex analyses of behavioral traits with better data. Finally, with the increasing interest in behavioral evolution, especially among psychologists, mathematical theories of evolution are becoming increasingly common and Wright takes a place in this, along with Fisher and Haldane.

Wright was unusual among biologists in taking an active interest in philosophy. He had a somewhat Leibnitzian view. He disliked any idea of emergence. He found no sharp borders between different levels of complexity, such as between embryo and adult or between mind and no mind. Thus if the mind does not emerge by magic, it must trace its development into a continuous process back to the embryo, the egg and sperm, and ultimately the DNA molecules. He called his concept 'dual-aspect panpsychism.' Mind is everywhere; so is matter.

The biological community either disagreed with Wright's view, or was indifferent. Most regarded mind as a natural outgrowth of appropriately organized matter. Wright held on to his dualistic view, however, and found a few kindred spirits among philosophers, particular his long time colleague at the University of Chicago, Charles Hartshorne.

Bibliography

Fisher R A 1929 *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford, UK

Haldane J B S 1932 *The Causes of Evolution*. Longmans Green, London

Provine W B 1986 *Sewall Wright and Evolutionary Biology*. University of Chicago Press, Chicago

Wright S 1917 The average correlation within subgroups of a population. *Journal of the Washington Academy of Sciences* **7**: 532–5

Wright S 1921 Correlation and causation. *Journal of Agricultural Research* **20**: 557–85

Wright S 1922a Coefficients of inbreeding and relationship. *American Naturalist* **56**: 530–38

Wright S 1922b The effects of inbreeding and crossbreeding in guinea pigs *Bulletin* **1090**: 1–63, **1091**: 1–60

Wright S 1931 Evolution in Mendelian populations. *Genetics* **16**: 97–139

Wright S 1968–78 *Evolution and the Genetics of Populations*. 4 Vols. University of Chicago Press, Chicago

Wright S 1988 Surfaces of selective value revisited. *American Naturalist* **131**: 115–23

J. F. Crow

Writing, Evolution of

Writing is mankind's principal technology for collecting, manipulating, storing, retrieving, communicating and disseminating information. Writing may have been invented independently three times in different parts of the world: in the Near East, China and Mesoamerica. In what concerns this last script, it is still obscure how and when the Olmec symbolic system, first attested from 600 to 500 BC, was transmitted to the Mayas and other Mesoamerican cultures (Marcus 1992). The earliest Chinese inscriptions dated of the Shang Dynasty, c. 1400–1200 BC, consist of oracle texts engraved on animal bones and turtle shells (Keightley 1985). The highly abstract and standardized signs suggest prior developments presently undocumented.

Of these three writing systems, therefore, only the earliest, the Mesopotamian cuneiform script, invented in Sumer, present-day Iraq, c. 3200 BC, can be traced without any discontinuity over a period of 10,000 years, from a prehistoric antecedent to the present-day alphabet. Its evolution is divided into four phases: (a) clay tokens representing units of goods were used for accounting (8000–3500 BC); (b) the three dimensional tokens were transformed into two-dimensional pictographic signs. Like the former tokens, the pictographic script served exclusively for accounting (3500–3000 BC); (c) phonetic signs, introduced to transcribe the name of individuals, marked the turning point when writing started emulating spoken language and, as a result, became applicable to all fields of human experience (3000–1500 BC); (d) with two dozen letters, each standing for a single sound of voice, the alphabet perfected the rendition of speech. After ideography, logography and syllabaries, the alphabet represents a further segmentation of meaning.

1. *Tokens Precursor of Writing*

The direct antecedent of the Mesopotamian script was a recording device consisting of clay tokens of multiple shapes (Schmandt-Besserat 1996). The artifacts, of mostly geometric forms such as cones, spheres, disks, cylinders and ovoids, are recovered in archaeological sites dating 8000–3000 BC (Fig. 1). The tokens, used as counters to keep track of goods, were the earliest code—a system of signs for transmitting information. Each token shape was semantic, referring to a particular unit of merchandise. For example, a cone and a sphere stood respectively for a small and a large measure of grain, and ovoids corresponded to jars of oil. The repertory of some three hundred types of counters made it feasible to manipulate and store information on multiple categories of goods (Schmandt-Besserat 1992).

The token system had little in common with spoken language except that, like a word, a token stood for one concept. Unlike speech, tokens were restricted to one type of information only, namely, real goods. Unlike spoken language, the token system made no use of syntax. That is to say, their meaning was independent of their placement order. Three cones and three ovoids, scattered in any way, were to be translated 'three baskets of grain, three jars of oil.' Furthermore, the fact that the same token shapes were used in a large area of the Near East, where many dialects would have been spoken, shows that the counters were not based on phonetics. Evidently, the goods they represented were expressed in any language. The token system showed the number of units of merchandize in one-to-one correspondence, in other words, the number of tokens matched the number of units counted: x jars of oil were represented by x ovoids. Repeating 'jar of oil' x times in order to express plurality is unlike spoken language.

2. *Pictography: Writing as Accounting Device*

After four millennia, the token system led to writing. The transition from counters to script took place simultaneously in Sumer and Elam, present-day western Iran when, around 3500 BC, Elam was under Sumerian domination. It occurred when tokens, probably representing a debt, were stored in envelopes until payment. These envelopes made of clay in the shape of a hollow ball had the disadvantage of hiding the tokens held inside. Some accountants, therefore, impressed the tokens on the surface of the envelope

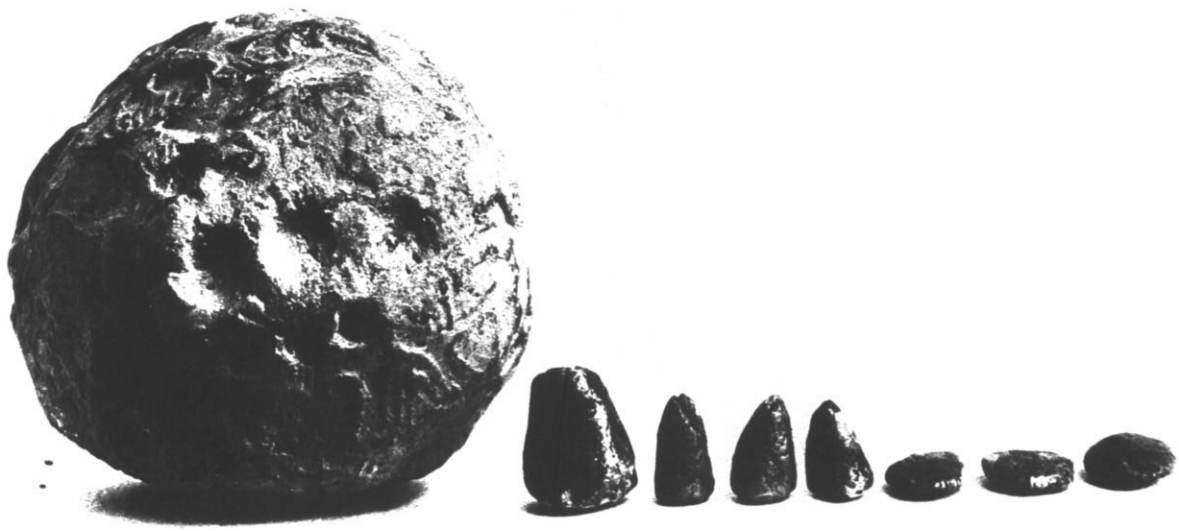

Figure 1
Envelope, tokens and corresponding markings, from Susa, Iran (Courtesy Musée du Louvre, Département des Antiquités Orientales)

Figure 2
Impressed tablet featuring an account of grain, from Godin Tepe, Iran (Courtesy Dr. T. Cuyler Young, Royal Ontario Museum, Toronto)

before enclosing them inside, so that the shape and number of counters held inside could be verified at all times (Fig. 1). These markings were the first signs of writing. The metamorphosis from three-dimensional artifacts to two-dimensional markings did not affect the semantic principle of the system. The significance of the markings on the outside of the envelopes was identical to that of the tokens held inside.

About 3200 BC, once the system of impressed signs was understood, clay tablets—solid cushion-shaped clay artifacts bearing the impressions of tokens—replaced the envelopes filled with tokens. The impression of a cone and a sphere token, representing measures of grain, resulted respectively in a wedge and a circular marking which bore the same meaning as the tokens they signified (Fig. 2). They were ideo-

Figure 3
Pictographic tablet featuring an account of 33 measures of oil, from Godin Tepe, Iran (Courtesy Dr. T. Cuyler Young, Royal Ontario Museum, Toronto)

grams—signs representing one concept. The impressed tablets continued to be used exclusively to record quantities of goods received or disbursed. They still expressed plurality in one-to-one correspondence.

Pictographs—signs representing tokens traced with a stylus rather than impressed—appeared about 3100 BC. These pictographs referring to goods mark an important step in the evolution of writing because they were never repeated in one-to-one correspondence to express numerosity. Besides them, numerals—signs representing plurality—indicated the quantity of units recorded. For example, '33 jars of oil' were shown by the incised pictographic sign 'jar of oil, 'preceded by three impressed circles and three wedges, the numerals standing respectively for '10' and '1' (Fig. 3). The symbols for numerals were not new. They were the impressions of cones and spheres formerly representing measures of grain, which then had acquired a second, abstract, numerical meaning. The invention of numerals meant a considerable economy of signs since 33 jars of oil could be written with 7 rather then 33 markings.

In sum, in its first phase, writing remained mostly a mere extension of the former token system. Although the tokens underwent formal transformations from three- to two-dimensional and from impressed markings to signs traced with a stylus, the symbolism remained fundamentally the same. Like the archaic counters, the tablets were used exclusively for accounting (Nissen et al. 1993). This was also the case when a stylus, made of a reed with a triangular end, gave to the signs the wedge-shaped 'cuneiform' appearance (Fig. 4). In all these instances, the medium changed in form but not in content. The only major departure from the token system consisted in the

Figure 4
Economic cuneiform tablet (Courtesy Texas Memorial Museum, The University of Texas at Austin)

creation of two distinct types of signs: incised pictographs and impressed numerals. This combination of signs initiated the semantic division between merchandise and numerosity.

3. Logography: Shift from Visual to Aural

About 3000 BC, the creation of phonetic signs—signs representing the sounds of speech—marks the second phase in the evolution of Mesopotamian writing, when, finally, the medium parted from its token antecedent in order to emulate spoken language. As a result, writing shifted from a conceptual framework of real goods to the world of speech sounds. It shifted from the visual to the aural world.

With state formation, new regulations required that the names of the individuals who generated or received the registered merchandise were entered on the tablets. The personal names were transcribed by the mean of logograms—signs representing a word in a particular tongue. Logograms were easily drawn pictures of words with a sound close to that desired (for example in English the name Neil could be written with a sign showing bent knees 'kneel'). Because Sumerian was mostly a monosyllabic language, the logograms had a syllabic value. A syllable is a unit of spoken language consisting of one or more vowel sounds, alone, or with one or more consonants. When a name required several phonetic units, they were assembled in a rebus fashion. A typical Sumerian name 'An Gives Life' combined a star, the logogram for An, god of heaven, and an arrow, because the words for 'arrow' and 'life' were homonyms. The verb was not transcribed, but inferred, which was easy because the name was common.

Phonetic signs allowed writing to break away, finally, from accounting. Inscriptions on stone seals or metal vessels deposited in tombs of the 'Royal Cemetery' of Ur, c. 2700–2600 BC, are among the first texts that did not deal with merchandise, did not include numerals and were entirely phonetic. The inscriptions consisted merely of a personal name: 'Meskalamdug,'

PUABI, QUEEN

Figure 5
Name and title of Puabi carved on a seal recovered in the Royal Cemetery of Ur (U10939) (Source: Amiet 1980)

or a name and a title: 'Puabi, Queen' (Fig. 5). Presumably, these funerary texts were meant to immortalize the name of the deceased, thereby, according to Sumerian creed, insuring them of eternal life. Other funerary inscriptions fostered even further the emancipation of writing. For example, statues depicting the features of an individual bore increasingly longer inscriptions. After the name and title of the deceased, followed patronymics, the name of a temple or a god to whom the statue was dedicated (in a dative form), and, in some cases, a plea for life after death, including a verb. These inscriptions introduced syntax and, therefore, brought writing yet one step closer to speech.

After 2600–2500 BC, the Sumerian script became a complex system of ideograms mixed more and more frequently with phonetic signs. The resulting syllabary—system of phonetic signs expressing syllables—further modeled writing on to spoken language. With a repertory of about 400 signs, it could express any topic of human endeavor. Some of the earliest syllabic texts were royal inscriptions, and religious, magic and literary texts.

The second phase in the evolution of the Mesopotamian script, characterized by the creation of phonetic signs, not only resulted in the parting of writing from accounting, but also its venturing out of Sumer to neighboring regions. The first Egyptian inscriptions, dated to the late fourth millennium BC, belonged to royal tombs (Bard 1992). They consisted of ivory labels and ceremonial artifacts such as maces and palettes bearing personal names, written phonetically as a rebus, visibly imitating Sumer. For example, the Palette of Narmer bears hieroglyphs identifying the name and title of the Pharaoh, his attendants and the smitten enemies. Phonetic signs to transcribe personal names, therefore, created an avenue for writing to spread outside of Mesopotamia. This elucidates why the Egyptian script was instantaneously phonetic. It also explains why the Egyptians never borrowed Sumerian signs. Their repertory consisted of hieroglyphs representing items familiar in the Egyptian culture and that evoked sounds in their tongue.

The phonetic transcription of personal names clearly played an important role for the dissemination of writing to the Indus Valley where, during a period of increased contact with Mesopotamia, c. 2500 BC, writing appears on seals featuring individuals' names and titles (Possehl 1996). In turn, the Sumerian cuneiform syllabic script was adopted by many Near Eastern cultures who adapted it to their different linguistic families and in particular, Semitic (Akkadians and Eblaites); Indo-European (Mitanni, Hittites, and Persians); Caucasian (Hurrians and Urartians); and finally, Elamite and Kassite. It is likely that Linear A and B, the phonetic scripts of Crete and mainland Greece, c. 1400–1200 BC, were influenced also by the Near East (Chadwick 1990).

4. *The Alphabet: The Segmentation of Sounds*

The invention of the alphabet about 1500 BC ushered in the third phase in the evolution of writing in the ancient Near East. The first, so-called Proto-Sinaitic or Proto-Canaanite alphabet, which originated in the

PROTO-CANAANITE	EARLY LETTER NAMES AND MEANINGS	
	alp	oxhead
	bêt	house
	gaml	throwstick
	digg	fish
	hô(?)	man calling
	wô (waw)	mace
	zê(n)	?
	ḥê(t)	fence?
	ṭê(t)	spindle?
	yad	arm
	kapp	palm
	lamd	ox-goad
	mêm	water
	naḥš	snake
	cên	eye
	pi't	corner?
	sa(d)	plant
	qu(p)	?
	ra'š	head of man
	tann	composite bow
	tô (taw)	owner's mark

Figure 6
Proto-Sinaitic Alphabet (source: Roaf 1990)

region of present-day Lebanon, took advantage of the fact that the sounds of any language are few. It consisted of a set of 22 letters, each standing for a single sound of voice, which, combined in innumerable ways, brought an unprecedented flexibility to transcribe speech (Healey 1991). This earliest alphabet was a complete departure from the previous syllabaries. First, the system was based on acrophony—signs to represent the first letter of the word they stood for—for example an ox head (*alpu*) was 'a,' a house (*betu*) was b (Fig. 6). Second, it was consonantal—it dealt only with speech sounds characterized by constriction or closure at one or more points in the breath channel, like b, d, l, m, n, p, etc. Third, it streamlined the system to 22 signs, instead of several hundred.

The transition from cuneiform writing to the alphabet in the ancient Near East took place over several centuries. In the seventh century BC the Assyrian kings still dictated their edicts to two scribes. The first wrote Akkadian in cuneiform on a clay tablet; the second Aramaic in a cursive alphabetic script traced on a papyrus scroll. The Phoenician merchants established on the coast of present day Syria and Lebanon, played an important role in the diffusion of the alphabet. In particular, they brought their consonantal alphabetic system to Greece, perhaps as early as, or even before 800 BC. The Greeks perfected the Semitic alphabet by adding letters for vowels—speech sounds in the articulation of which the breath channel is not blocked, like a, e, i, o, u. As a result the 27-letter Greek alphabet improved the transcription of the spoken word, since all sounds were indicated. For example, words sharing the same consonants like 'bad,' 'bed,' 'bid,' 'bud,' could be clearly distinguished. The alphabet did not subsequently undergo any fundamental change.

5. *The Modern Alphabets*

Because the alphabet was invented only once, all the many alphabets of the world, including Latin, Arabic, Hebrew, Amharic, Brahmani and Cyrillic, derive from Proto-Sinaitic. The Latin alphabet used in the western world is the direct descendant of the Etruscan alphabet. The Etruscans, who occupied the present province of Tuscany in Italy, adopted the Greek alphabet, slightly modifying the shape of letters. In turn, the Etruscan alphabet became that of the Romans, when Rome conquered Etruria in the first century BC. The alphabet followed the Roman armies. All the nations that fell under the rule of the Roman Empire became literate in the first centuries of our era. This was the case for the Gauls, Angles, Saxons, Franks and Germans who inhabited present-day France, England and Germany.

Charlemagne (800 AD) had a profound influence on the development of the Latin script by establishing standards. In particular a clear and legible minuscule cursive script was devised, from which our modern day lower case derives. The printing press invented in 1450 dramatically multiplied the dissemination of texts, introducing a new regularity in lettering and layout. The Internet catapults the alphabet into cyberspace, while preserving its integrity.

6. *Writing: Handling Data in Abstraction*

Beyond the formal and structural changes undergone by writing in the course of millennia, its evolution also involved strides in the ability to handle data in abstraction. At the first stage, the token system antecedent of writing, already abstracted information in several ways. First, it translated daily-life commodities into arbitrary, often geometric forms. Second, the counters abstracted the items counted from their context. For example, sheep could be accounted for by an administration, independently from their actual location. Third, the token system separated the data from the knower. That is to say, a group of tokens communicated directly specific information to anyone initiated in the system. This was a significant change for an oral society, where knowledge was transmitted by word of mouth from an individual to another, face to face. Otherwise, the token system represented plurality concretely, in one-to-one correspondence. Three jars of oil were shown by three tokens, as it is in reality. At the same time, the fact that the token system used specific counters to count different items was concrete—it did not abstract the notion of item counted from that of number. (Certain English numerical expressions such as twin, triplet, quadruplet and duo, trio, quartet, to refer to particular sets are comparable to concrete numbers.)

When tokens were impressed on the envelopes to indicate the counters enclosed inside, the resulting markings could no longer be manipulated by hand. In other words, the transmutation of three-dimensional counters into two-dimensional signs constituted a second step in abstraction. By doing away with tokens, the clay tablets marked a third level of abstraction since the impressed markings no longer replicated a set of actual counters. The invention of numerals, which separated the notion of numerosity from that of the item counted, was a crucial fourth step in abstraction. The signs expressing the concept of oneness, twoness, etc., allowed plurality to be dealt with in fully abstract terms. In turn, the phonetic units marked a fifth step of abstraction, since the signs no longer referred to the objects pictured, but rather the sound of the word they evoked.

Phonetics allowed writing to shift from a representational to a conceptual linguistic system. That it to say it enabled writing to leave the realm of real goods in order to enter the world of words and the ideas they stand for. Finally, the process that started with ideograms expressing concepts and phonetic signs

referring to the sound of monosyllabic words reached the ultimate segmentation of meaning with letters. As Marshall McLuhan (1997) defined it, the alphabet consisted of semantically meaningless letters corresponding to semantically meaningless sounds. The alphabet brought data handling to a final double-stepped abstraction.

7. *Conclusion: The Stability of Writing Systems*

The origin of the Chinese script and the development of Mesoamerican writing are still obscure. The Mesopotamian script, however, offers a well-documented evolution over a continuous period of 10,000 years. The system underwent drastic changes in form; gradually transcribed spoken language more accurately, and handled data in more abstract terms. The most striking universal feature of all writing systems, however, is their uncanny endurance, unmatched among human creations. The Chinese script never needed to be deciphered because the signs have changed little during the 3400 years of its recorded existence. It also always remained ideographic, merely inserting rebus-like phonetic complements in some characters. The Mesoamerican Maya phonetic glyphs preserved the symbolism initiated by the Olmecs over one millennium. Finally, when the last clay tablet was written in the Near East, c. 300 AD, the cuneiform script had been in use for three millennia. It replaced an age-old token system that had preceded it for over 5000 years; it was replaced by the alphabet that has prevailed for 3500 years.

See also: Brain, Evolution of; Genes and Culture, Coevolution of; Letter and Character Recognition, Cognitive Psychology of; Number Systems, Evolution of; Reading Skills; Spelling; Writing Instruction; Writing Process, Psychology of; Writing Systems; Writing Systems, Psychology of

Bibliography

Amiet P 1980 *La Glyptique Mesopotamienne Archaïque*. Editions du Centre National de la Recherche Scientifique, Paris Plate 90 1182

Avrin L 1991 *Scribes, Script and Books*. American Library Association, Chicago, IL

Bard K A 1992 Origins of Egyptian writing. In: Frieman R, Adams B (eds.) *The Followers of Horus*. Egyptian Studies Association Publication 2 Oxbow Monograph 20, Oxford 297–306

Chadwick J 1990 *The Decipherment of Linear B*. Cambridge University Press, Cambridge, UK

Ferioli P, Fiandra E, Fissore G G (eds.) 1996 *Administration in Ancient Societies*. Ministero per I beni Culturali e Ambientali Ufficio Centrale per I beni Archivisti, Rome

Günther H (ed.) 1994 *Schrift und Schriftlichkeit*. de Gruyter, Berlin

Healey J F 1991 *The Early Alphabet*. University of California Press, Berkeley, CA

Keightley D N 1985 *Sources of Shang History, The Oracle-Bone Inscriptions of Bronze Age China*. University of California Press, Berkeley, CA

Marcus J 1992 *Mesoamerican Writing Systems*. Princeton University Press, Princeton

McLuhan M 1997 *Media Research*. Overseas Publishers Association, Amsterdam, The Netherlands

Nissen H J, Damerow P, Englund R K 1993 *Archaic Bookkeeping*. University of Chicago Press, Chicago, IL

Possehl G L 1996 *Indus Age: The Writing System*. University of Pennsylvania Press, Philadelphia PA

Roaf M 1990 *Cultural Atlas of Mesopotamia*. Equinox, Oxford 150

Schmandt-Besserat D 1992 *Before Writing* 2 vols. University of Texas Press, Austin, TX

Schmandt-Besserat D 1996 *How Writing Came About*. University of Texas Press, Austin, TX

Senner W M (ed.) 1989 *The Origins of Writing*. University of Nebraska Press, Lincoln

Watt W C (ed.) 1994 *Writing Systems and Cognition*. Kluwer Academic Publishers, Dordrecht

D. Schmandt-Besserat

Writing Instruction

The term 'writing' covers a huge range of activities, from, at one extreme, the motor processes in forming letters, to, at the other extreme, the creative mental processes involved in writing a novel. For present purposes, the term 'writing' will be restricted to referring to the cognitive and social processes involved in writing, and 'instruction' to referring to educational practice as it has been informed by research into these processes.

Contemporary approaches to writing instruction have their origins in the 1970s (see Nystrand et al. 1993 for a review of the development of writing studies during this period). Before this period, the overwhelming emphasis was on the text as a linguistic product, with research and instruction concentrating largely on identifying the distinctive features of exemplary texts and teaching these to students. In contrast, the defining feature of more recent research has been its emphasis on understanding the processes by which text is produced, and in particular, on the processes involved in creating the knowledge embodied in text. Writing, in this view, is not just a means of communication, but a vital tool for learning and thinking. During this period, understanding of the nature of the processes involved has expanded and, in some respects, changed considerably. Research has moved from an initial, almost exclusive, concern with the cognitive processes

within individuals, through a broader view of text as the product of social processes, to a mixed view, in which writing is seen as the product of an interaction between conflicting processes differing in the extent to which they involve adapting to the social context.

1. Writing as Problem Solving

Cognitive research on writing has been dominated by a view of writing as a form of problem solving, in which writing skill depends on the ability to think strategically about how best to achieve communicative goals. The most widely cited model in this area is Hayes and Flower's (1980) descriptive model of the cognitive processes in writing. Based largely on protocols of writers thinking aloud as they composed, the model identifies three basic components of the writing process, each of which can be further analyzed into more basic processes until they 'bottom out' in the basic cognitive operations studied by cognitive and experimental psychologists. The three processes are: planning, which involves setting goals, and generating and organizing ideas to satisfy them; text generation, which involves encoding ideas in language; and revision, which involves reading and editing the text that has already been produced. These basic processes are controlled by a monitor or central executive which is responsible for coordinating the processes as they operate on information stored in long-term memory, in response to the demands of the task environment. The key feature of the model is that the basic components of planning, translation, and revision are recursively applied processes rather than stages in a sequence of activities. Thus, each of the processes can, in principle, be called on just as much while jotting down notes before writing as while writing or rewriting the text itself. In consequence, the model is able to capture the dynamic interplay between different components at different points in writing. Hayes and Flower's subsequent work, with a variety of different collaborators, involved more detailed studies of the different subcomponents of the writing process. A useful review of the basic findings from this research, and a heavily revised version of the original model, can be found in the chapter by Hayes in a volume edited by Levy and Ransdell (1996).

Perhaps the most important consequence of this research was that it enabled a characterization of differences between expert and novice writers. (See Hayes and Flower 1986, for a review of their own findings, and Bereiter and Scardamalia 1987, for an influential model examining the development of children's writing, and procedures whereby this could be developed.) Bereiter and Scardamalia characterize the difference between novice and expert writing as a contrast between a knowledge-telling model of writing and a knowledge-transforming model of writing. Knowledge telling is typically employed by children and less experienced writers and is essentially a 'think-say' method of composing in which ideas are retrieved associatively from memory and then translated directly into text. Typically, there is little differentiation between components of the writing process, so that text is produced in a single draft with little preplanning or revision of the text. Even when explicitly instructed to plan before writing or to revise a text, planning tends to be restricted to the direct listing of content as it comes to mind, and revision to cosmetic changes to surface features of the text. The result is a text, which, although it may cover relevant material, is lacking a distinctive point of view or evidence of reflection upon ideas. By contrast, the kind of knowledge-transforming strategy employed by expert writers involves actively designing a text to satisfy communicative goals. Experts spend more time planning, both before and during writing, and develop more elaborate representations of the rhetorical problem to be solved, with the result that ideas are not just retrieved directly from memory but are actively constructed and evaluated with respect to this richer representation of the rhetorical problem. Similarly, revision becomes a goal-directed process involving global changes to the structure and content of the text, as well as correction of its surface features. The result is more reflective text, which shows more evidence of design, and in which ideas are selected and presented in the service of higher-level communicative goals. Moreover, in the process of working out how to communicate effectively with a reader, writers are also actively transforming their own knowledge.

According to this view of writing expertise, one of the main factors affecting writing performance is cognitive overload, arising from the fact that a complex set of processes has to be carried out in a limited-capacity working memory. In particular, the demands of translating ideas into well-formed text may consume resources required for higher-level reflective thought. Although this is particularly true for children, for whom even the basic mechanics of forming letters may be resource consuming, it is a pervasive problem stemming from the nature of the process itself. Research in this area has suggested that individual differences in performance are associated with differences in working memory capacity, and also that writers with greater knowledge of the different components of the writing process (metacognitive knowledge) are better able to manage the conflict between processes, and to switch attention between planning, translation, and revision at appropriate points. More specifically, Kellogg (1994) has shown that outlining before writing improves the quality of text compared to producing text in a single draft, and that this is because it enables writers to focus attentional resources on either planning or translation at different points in writing, rather than dividing up their mental resources among both processes. (For a

review of Kellogg's work on drafting strategies, see Kellogg 1994; for a model of working memory in writing, see Kellogg's chapter in Levy and Ransdell 1996.)

The second main strand of research in this area has focused on the reflective processes involved in the knowledge-transforming strategy itself. Bereiter and Scardamalia describe these in terms of an interaction between a content problem space, in which the writer works out the content to be included in the text, and a rhetorical problem space, in which the rhetorical problem is represented, and in which goals for the text are developed. In this view, while novice writers have the procedures for turning content into text, they lack the executive procedures required to translate problems encountered in rhetorical space back into goals to be achieved in content space. Accordingly, the crucial ingredient of writing expertise that the novice needs to learn is how to incorporate communicative goals into the writing process. Bereiter and Scardamalia suggest two broad ways in which this can be done.

First, the processes involved can be made visible by providing models of the writing process. For example, a teacher thinking aloud as they compose a text can demonstrate the formulation of goals during writing, and the search for, and evaluation of, content relevant to these goals. Second, children can be helped to incorporate more goal-directed thinking into their own writing using the technique of 'procedural facilitation.' This involves providing external prompts suggesting goals, possible courses of action, or relevant evaluatory criteria which the writer can consult moment by moment as they write a text. Gradually, as such procedures are internalized, this scaffolding can be removed, leaving the writer capable of carrying out the appropriate procedures unaided.

Teaching programs designed to increase metacognitive awareness of the components of writing and using methods such as modeling and procedural facilitation to develop more goal-directed thinking have shown strong effects on children's writing compared with control groups. (For examples of effective programs based on these ideas, see Englert et al. 1991, and the chapter by Graham and Harris in the collection edited by Levy and Ransdell 1996.)

2. *Writing as Social Action*

In both Hayes and Flower's original model and in Hayes's more recent revision of the model, the social nature of the writing process is represented by the inclusion of the audience and topic specification as features of the task environment, and by the inclusion of audience knowledge and more general discourse knowledge as part of the information stored in long-term memory (see Hayes's chapter in Levy and Ransdell 1996). One might expect, therefore, that there would be a complementary relationship between cognitive and social research on writing, with cognitive research focusing on the processes by which information is manipulated, and social research focusing on such factors as the relationship between the writer and their audience, or on elaborating on the nature of the conventions which make up the writer's discourse knowledge. In fact, social research has developed in part as a reaction against cognitive models of writing, and has led to the development of alternative theoretical approaches to the writing process which have had a strong influence on approaches to writing instruction.

Social critiques of problem-solving models take a variety of forms (see Cooper and Holzman 1989 for a fairly representative example). But at the heart of them all there is an underlying disquiet with the way in which problem-solving models treat individual elements of text as being constructed out of a stable set of cognitive processes and representations which can be explicitly manipulated by the writer. In particular, they question the extent to which content knowledge is stored independently of discourse knowledge, and dependent on explicit goal-directed processing, if it is to be realized as a complete social act. Instead, they treat each part of the text as an indivisible 'social act,' which cannot be further broken down into components which can be explicitly manipulated.

Within social approaches, then, the elements of a text are defined as social acts, and analyzed in terms of the functions they serve within a particular context, rather than in terms of the processes by which they are cognitively constructed. The direct determinant of the form of the text is the social environment in which writing takes place, and variations in writing performance are treated as variations in the way the writer responds to that environment. In this view, writing is not a uniform set of mental procedures, but rather a repertoire of practices, which varies widely across different contexts. This has a range of implications for research and instructional practice. First, research into writing takes a fundamentally different form, more like that of natural history, and involves predominantly qualitative methods, aimed at identifying the different kinds of act which are employed in specific contexts, relating these to the environment in which they occur, and examining how they vary across different contexts. Second, learning to write is no longer treated as a matter of learning to write in general, but rather of learning what practices or social acts are appropriate in a variety of particular contexts. Finally, the teaching of writing becomes a matter of creating authentic contexts that draw out approximations of the relevant social acts, which can then be shaped with practice into the appropriate form.

The most dominant of the various social approaches to writing is the field of genre studies. The fundamental idea here is that texts are not produced by individuals working in isolation, but rather by individuals as

members of communities employing a characteristic set of discursive practices to achieve the goals shared within the community. A genre in its most general form is a repeated set of social acts produced in response to a recurrent situation. It specifies the shared set of communicative purposes within a group or academic discipline, and shapes the schematic structure of text, as well as influencing choices of content and style. It provides conventions about what issues are appropriate, the forms of reasoning which are typically employed, and assumptions about the different roles of readers and writers within the community. Because these practices are conventional in form and vary widely across groups, novice writers cannot work them out from first principles, or from their experience within other discourse communities, but instead have to internalize them by immersion within the community. As they gain experience and begin to be able to use the genre as a guide to their own text production, so the genre also becomes a tool enabling them to construct their knowledge in the way characteristic of the particular discipline to which they are trying to gain membership. (See Freedman and Medway 1994 for an overview of work in the field.)

Research in this field has had a number of effects on writing instruction. First, it has led to a move away from the teaching of writing as a general skill by 'writing experts' toward the teaching of writing within disciplines by disciplinary insiders. Second, the content of teaching focuses on explicating the goals of the discipline and identifying the social acts within the text which enable the discipline to achieve its goals. Third, in line with the principle that text is a response to the social environment, a great emphasis is placed on designing tasks and creating environments which recreate as authentically as possible the social conditions and purposes of the community as it is in the 'real world.' (See Swales 1991 for a detailed example of the ways in which a genre approach can be applied to writing instruction.)

3. *Writing as a Knowledge-constituting Process*

Problem-solving models and social approaches to writing have both had a significant impact on writing instruction, and in many cases writing instruction draws eclectically on insights from both traditions. However, the theoretical conflict between them has not yet been resolved, with the result that research has tended to progress in parallel, with the two traditions using fundamentally different vocabularies to describe the writing process. There have been some attempts to combine the two traditions (see, for example, Flower 1994 for a relatively cognitive view, and Brandt 1992 for a more social-interactive approach), but these have not been generally accepted. More recently, however, there have been signs of a more integrated approach based on an alternative cognitive framework.

This has developed from a strand of research focusing on the implicit processes involved in text production itself, rather than on higher-level problem solving or social negotiation. (See Torrance and Galbraith 1999 for a collection of papers describing some of the empirical research involved, and for discussions of their theoretical implications for models of the writing process.) Galbraith, for example, has proposed a dual process model of writing, in which the more implicit and involuntary kinds of processing implied by social approaches to writing are combined with the more explicit and deliberate processes proposed by problem-solving models. In particular, this model invokes connectionist principles—involving parallel constraint satisfaction within a distributed representation of knowledge—to explain how 'social acts' can be formulated in the text as it is produced, without the need for the deliberate manipulation of separate components of information assumed by traditional problem-solving models. In this view, text production is an active knowledge-constituting process in its own right. Furthermore, and crucially, although the model accepts that the more explicit problem-solving processes assumed in traditional cognitive models of writing play an important role, it stresses that these operate on the output of the knowledge-constituting process, and highlights the potential for conflict between these two different processes.

This emphasis on text production as an active knowledge-constituting process, and on the conflicting nature of the writing process as a whole, is compatible with what has become a highly influential, although not universally accepted, approach to writing instruction, associated, in particular with the work of Elbow (1998; and see Belanoff et al. 1991 for a more recent review of work in this tradition). Elbow advocates the use of a multiple drafting strategy, in which the writer is encouraged to start writing by producing a spontaneous draft of their thoughts as they come to mind, without preplanning or attempting to impose organization on the text. Once a draft has been completed, revision involves identifying the main ideas that have emerged in the text, and these are used as the basis for a further draft of the text. This process of alternating between spontaneous drafts and more reflective revision continues until the text achieves a satisfactory form. Writers are also encouraged to engage in regular 'free-writing' exercises, as a means of providing them with direct experience of the generative properties of this form of writing, and to participate in peer feedback groups to gain direct experience of the effect of their writing on their readers.

This more recent research on text production can be seen, in part, therefore, as providing a theoretical rationale for some long-standing practices within writing instruction. But, in addition, in promising a deeper understanding of the implicit processes involved in text production, it may help to resolve the

division between social and cognitive approaches to writing.

See also: Problem Solving and Reasoning, Psychology of; Situated Cognition: Contemporary Developments; Social Constructivism; Working Memory, Psychology of; Writing Process, Psychology of; Writing Systems, Psychology of

Bibliography

Belanoff P, Elbow P, Fontaine S I (eds.) 1991 *Nothing Begins with N: New Investigations of Freewriting*. Southern Illinois University Press, Carbondale, IL

Bereiter C, Scardamalia M 1987 *The Psychology of Written Composition*. Lawrence Erlbaum Associates, Hillsdale, NJ

Brandt D 1992 The cognitive as the social: An ethnomethodological approach to writing process research. *Written Communication* **9**: 315–22

Cooper M, Holzman M 1989 *Writing as Social Action*. Boynton/Cook, Portsmouth, NH

Elbow P 1998 *Writing without Teachers*, 2nd edn. Oxford University Press, New York

Englert C S, Raphael T E, Anderson L M, Anthony H M, Stevens D D 1991 Making strategies and self-talk visible: Writing instruction in regular and special education classrooms. *American Educational Research Journal* **28**: 337–72

Flower L 1994 *The Construction of Negotiated Meaning*. Southern Illinois University Press, Carbondale, IL

Freedman A, Medway P (eds.) 1994 *Genre and the New Rhetoric*. Taylor and Francis, London

Hayes J R, Flower L 1980 Identifying the organization of writing processes. In: Gregg L W, Steinberg E R (eds.) *Cognitive Processes in Writing*. Erlbaum, Hillsdale, NJ

Hayes J R, Flower L 1986 Writing research and the writer. *American Psychologist* **41**: 1106–13

Kellogg R T 1994 *The Psychology of Writing*. Oxford University Press, New York

Levy M, Ransdell S (eds.) 1996 *The Science of Writing*. Lawrence Erlbaum Associates, Mahwah, NJ

Nystrand M, Greene S, Wiemelt J 1993 Where did composition studies come from? An intellectual history. *Written Communication* **10**: 267–333

Swales J M 1991 *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge, UK

Torrance M, Galbraith D (eds.) 1999 *Knowing What to Write*. Amsterdam University Press, Dordrecht, The Netherlands

D. Galbraith

Writing Process, Psychology of

The psychology of the writing process treats the cognitive and emotional dimensions of expressing private thoughts as public written symbols, including their neural basis and development. The kinds of writing done in schools, the workplace, the home, and elsewhere are diverse. Signing one's name, completing a form, drafting a routine letter, and composing a novel differ in their cognitive demands, but these and other writing tasks draw on common production systems. In this article, three aspects of the psychological processes underlying the production of an extended coherent text will be considered in turn. These are thinking, language, and memory.

1. Writing as Thinking

Writing involves systems for formulating, executing, and monitoring text. The seminal model of the writing process proposed that formulation involves planning content and generating sentences based on knowledge drawn from the writer's long-term memory (Hayes and Flower 1980). Planning a text entails generating ideas, organizing them, and setting goals to be achieved in the text. Translating ideas or sentence generation includes the subprocesses of speech, such as selecting appropriate lexical entries and assembling the constituents in the proper order. Reviewing the text involves reading and editing operations that detect faults at multiple levels of text structure, ranging from local mechanics to the coherence of the whole text. The writer must monitor whether the processes are successful and decide which writing process to engage at a given moment.

If one focuses on the reflective cognition seen in planning and reviewing, then writing looks like thinking. Composition studies in rhetoric and education in the 1970s stressed the similarity of writing to problem solving. Later composition studies shifted the focus to social cognition and stressed that thinking is situated within a community of other writers and readers.

1.1 Problem Solving

Planning, translating, and reviewing are constrained by the task environment and the writer's long-term memory. The topic, the type of text, the intended readership, and the text produced thus far constitute elements of the task environment, as do the writer's tools, collaborators, and deadlines. The task environment and long-term memory together pose the problems of content (what to say) and rhetoric (how to say it) that must be solved in order to produce a text.

Mature writers seek solutions to both the content and rhetorical problems posed by the writing task at hand, using a strategy called knowledge transforming (Bereiter and Scardamalia 1987). The writer gains insight into the ideas being formulated as a consequence of expressing them as text, actively transforming what they know about a topic as a result of

trying to write about it. Unlike the reflective problem solving of mature writers, young children fail to elaborate their thinking through the act of writing. They instead engage in knowledge telling. The task environment automatically cues children to retrieve an idea from long-term memory and then translate it into sentences. Writing continues in this manner until retrieval fails to provide relevant material. During adolescence, knowledge transforming begins to emerge as a complement to knowledge telling.

Data based on verbal protocols and other forms of introspection indicate that adult writers recursively attend to planning, translating, and reviewing processes. Although planning generally occurs more frequently in the beginning stages of a first draft and reviewing later, all three operations may be called upon at any point. One effective strategy for coping with the multiple demands of writing on working memory is to delay concerns about how ideas are expressed and focus on planning. This increases the number of ideas generated during planning (Glynn et al. 1982). Another effective strategy is to prepare an outline of ideas and their organization before composing a first draft. Advance outlining permits the writer to focus on translating ideas into sentences and can increase the fluency and quality of writing at least for short texts on familiar topics.

There are individual differences in the strategies preferred by adult writers (Kellogg 1994). Some prefer to plan and review in advance in careful detail, perhaps creating diagrams, outlines, and notes, before starting a first, relatively polished draft. This has been called by several names, such as the Planner or Mozartian strategy. Such writers transform their knowledge about the topic by tackling the content and rhetorical problems as part of prewriting. Others prefer to discover ideas and develop their thinking on a topic by actually writing about them in a rough draft and then later doing extensive revision. This goes by the name of Discoverer, Reviser, or Beethovian strategy. Surveys of faculty and graduate students have revealed that planning carefully and revising little is associated with higher productivity in academic writing. However, it is unknown whether particular kinds of texts are generated more readily with a discover and revise approach (e.g., creative writing), or why some writers prefer one approach to another while others adapt their strategy to the task at hand (e.g., personality traits).

Writing exposes not only the content of our thoughts but also our ability to think, remember, and articulate. Avoidance of writing can be a means of coping with the anxiety and other emotional reactions to expressing thoughts as public symbols. In writer's block, text production halts altogether, although reading and thinking about the task may continue. The prewriting phase can be lengthy and a useful time for reflection for any writer. But in writer's block, the prewriting activities become a way to procrastinate and no longer serve any useful purpose. Procrastination, excessive worrying, impatience, perfectionism, and evaluation anxiety characterize writer's block.

1.2 Social Cognition

Besides an act of problem solving, writing is inherently a social act. Texts are created to be read within a discourse community (Rafoth and Rubin 1988). The social aspects may be explicit, as in collaborative writing, or they may be implicit, as in the way an imagined reader shapes the writer's ideas and manner of expression.

Expert writers conform their texts to the expectations of their readers. This point is well illustrated in the publication style requirements of the American Psychological Association. The style demands an allegiance to the empirical method and so defines the psychologist's task as a writer. Conclusions must be hedged so as to avoid stepping beyond the limits of the data, creating an air of uncertainty atypical in the humanities. Disagreements are couched in terms of differences about appropriate methods or interpretations of data, never in the personal terms at times seen in literary criticism.

It is often through revision that writers make their texts accessible to readers. Immature writers fail to take their audience into account and revise effectively. Reading and editing the text are highly effortful for young writers, in part because of working memory limitations. The lack of thorough revision also stems from a failure to establish clear plans in the first place. Without explicit goals that include the needs of potential readers, young writers cannot detect problems in meeting those goals. Revising becomes little more than editing errors in spelling, punctuation, and other writing mechanics. Besides needing explicit representations of what they wrote and what they intended, writers further need a representation of how the text appears to their readers. Skilled writers can see their work from another person's perspective. Taking the perspective of another individual is an important milestone in social cognitive development in many domains, including written language production.

2. Writing as Language Production

Writing is an act of language production, too. However, speech production has received far more attention in psycholinguistics than writing, largely because the field of linguistics stressed the spoken over written word during the twentieth century. Psychological studies of literacy are also lopsided in that reading outweighs writing (for example, see *Reading Skills*). This asymmetry derives from the methodological difficulties of studying language production compared with comprehension.

In experiments on language comprehension, the inputs to the system can be controlled and the outputs carefully measured. What someone says or writes in response to an experimenter's prompt is, by contrast, unpredictable. Rather than experimentation, the natural observation of speech errors or slips of the tongue guided theories of speech production, and slips of the pen have likewise informed us about written spelling. Observations of pauses in handwriting—both their location and duration—have provided insights into writing as language production. Another alternative to experimental manipulations has been the use of neuropsychological case studies that correlate lesions in brain structure with difficulties in speaking, writing, or both.

From slips of the tongue and neuropsychology, much is known about the translation of ideas and their motor execution in speech (for example, see *Speech Errors, Psychology of*). Five levels of processing must be distinguished in sentence production. The message level corresponds to the planning of the ideas or propositions to be conveyed. Writers may start with only inchoate feelings or images that must be mapped onto propositional representations suitable for sentence production. Functional processing reflects the translation of propositions into the words or lexical items along with the grammatical roles to be played by various phrases. Positional processing continues translation by retrieving the lexical (e.g., phonological segments) and grammatical morphemes (e.g., inflections), and positioning these constituents in a hierarchical syntactic structure. Next, the phonological component completes translation in speech by specifying all aspects of the sound structure of the sentence, including its prosodic features such as stress patterns. Unlike speech, writing requires a graphemic representation (Caramazza 1991). Graphemes can be computed either from the visual appearance or orthographic code of a whole word or by applying rules for converting phonology to orthography in spelling out the word. The order of graphemes for spelling a word and the motor programs for controlling the musculature in typing or handwriting must be specified. In handwriting, for example, the size of the letters and their allographic form must be selected.

In crafting a written sentence, more care may be given to correctness and precision than in speech. Whereas sentence fragments are characteristic of speech, written language is expected to be well formed. The motor execution process, particularly with handwriting, is markedly slower than typical rates of speech production, and the writer is freed from the social awkwardness of pauses, allowing for both more intensive formulation and editing than would be found in most speech. The production of a single sentence in writing, then, is an intricate affair, dependent on multiple neurological structures. The inability to express a thought in writing, agraphia, usually occurs to some degree as a consequence of significant brain damage, regardless of its location. Specific disorders of central language systems, as in aphasia, or of motor systems, as in Parkinson's disease, cause agraphia, as does widely spread pathology, as in Alzheimer's disease.

But written language production entails more than the production of a sentence, more than appropriately arranged letters and words. A text is made up of multiple sentences that must be coherent at both a local level and a global level (for example, see *Memory for Text*). Local coherence is achieved by cohesive ties that link one sentence to another. To illustrate, anaphoric pronouns form a link between the pronoun and the noun to which it refers in the previous sentence (e.g., The *dog* chased the thief down the street. *He* barked fiercely all the way.). Local cohesion is necessary but not sufficient, however. All the sentences in a text must relate to one another so that the text is also globally coherent. An extended text must be organized both at the level of its microstructure and its macrostructure.

3. *Writing as Memory*

Without knowledge, writers cannot succeed at their task. Diverse kinds of knowledge must be available in long-term memory, retrieved when needed, and maintained in working memory as the writer plans, edits, and so on. In particular, writers must know much about their topic, content knowledge, and about the requirements of a well-structured text, discourse knowledge.

3.1 *Long-term Memory*

Content knowledge guides the generation of informative and interesting ideas. To develop a theme, writers must be able to retrieve ideas from long-term memory throughout the course of prewriting, drafting, and revising. An inadequate stock of content knowledge is one reason why student writing fails and professional writing succeeds. Just as expertise in a domain improves a reader's ability to comprehend a text, it also improves their ability to write in an interesting manner. For example, a high degree of knowledge about baseball enables writers to generate many more game relevant propositions in a narrative about a half inning of play (Voss et al. 1980). Although less knowledgeable writers are able to generate macrostructure propositions well, they fail to supply the necessary detail for the microstructure of the text. Including sufficient detail improves the reader's comprehension and opinion of the text.

Discourse knowledge guides the selection of relevant ideas and the organization of these ideas. The writer must know how to link clauses together using conjunctions, how to use pronouns so that their

reference is unambiguous, and how to choose between the use of the same word or a synonym in linking clauses. These and other cohesive devices must be learned for a writer to produce text that is locally coherent. Although it is often emphasized that people acquire the syntactic rules needed to support spoken language before a child enters school, learning to use cohesive links among clauses in a written text develops throughout the primary grades.

To create global coherence in narrative or other kinds of text, each of the individual sentences must help to develop the story, theme, or argument. Narrative schemas are acquired at a young age and are heavily practiced throughout childhood in listening to and telling stories. Knowing the typical sequence of events in a story helps the writer to select ideas that need to be included and provides an organizational format for expressing them. Teachers draw on the writer's knowledge of discourse by giving narrative assignments, and students show their strongest writing skills in the narrative mode. In contrast, schemas for argumentation and persuasion are weakly developed in all but the highest ability college students, and the quality of the average student's expository and persuasive writing is correspondingly poor (Britton et al. 1975). Regardless of the type of discourse being written, the more one knows about the requirements of global coherence, the better prepared one is to actually produce an effective and memorable text.

3.2 Working Memory

Representations must be not only retrieved from long-term memory, but they must be maintained in an active state for writers to plan, translate, and review. The details of how verbal, visual, spatial, and central executive components of working memory support writing processes are still unclear. But it is known that if working memory resources are diverted away from writing, then performance suffers. For example, when adult writers try to maintain a set of five words in working memory and concurrently transcribe orally presented sentences in French, they make subject–verb agreement errors (Fayol et al. 1994). The limitations of working memory are especially problematic for children learning to write. Planning suffers because they often forget the ideas recently generated and so repeat them again. Sentence generation suffers because the number of linguistic units that may be coordinated is limited. Furthermore, handwriting or typing are not as effortless in children as they are in adults. Children must divert working memory resources from formulation and monitoring to motor execution.

Finally, the extensive interactions among planning, translating, and reviewing processes found in mature writing are highly effortful. For example, in planning a text, an experienced writer sets rhetorical goals that are shaped by the intended audience. Such high-level planning has consequences for the selection of ideas to be included and even the choice of words in a given sentence. It is not the case that the writer plans the ideas of the text and then automatically translates them into sentences. The formulation and monitoring processes typically require working memory resources even when motor execution has been automatized. Given its demands, learning to write effectively requires instruction and practice throughout childhood and young adulthood. Whereas learning to speak occurs universally to a remarkable degree of mastery by five years of age, the demands of literacy greatly extend the language acquisition period.

See also: Literacy, Cognitive Psychology of; Speech Production, Psychology of; Writing Systems, Psychology of

Bibliography

Bereiter C, Scardamalia M 1987 *The Psychology of Written Composition*. Lawrence Erlbaum, Hillsdale, NJ

Bock K 1996 Language production: Methods and methodologies. *Psychonomic Bulletin and Review* **3**: 395–421

Boice R 1985 Cognitive components of blocking. *Written Communication* **2**: 91–104

Britton J, Burgess T, Martin M, McLeod A, Rosen H 1975 *The Development of Writing Abilities*. Macmillan Education, London

Caramazza A 1991 *Issues in Reading, Writing, and Speaking: A Neuropsychological Perspective*. Kluwer Academic, Dordrecht, The Netherlands

Fayol M, Largy P, Lemaire P 1994 When cognitive overload enhances subject–verb agreement errors. *The Quarterly Journal of Experimental Psychology* **47A**: 437–64

Garrett M F 1980 Levels of processing in sentence production. In: Butterworth B (ed.) *Language production*. Academic Press, London, pp. 177–220

Glynn S M, Britton B K, Muth D, Dogan N 1982 Writing and revising persuasive documents: Cognitive demands. *Journal of Educational Psychology* **74**: 557–67

Hayes J R, Flower L S 1980 Identifying the organization of writing processes. In: Gregg L W, Steinberg E R (eds.) *Cognitive Processes in Writing*. Lawrence Erlbaum, Hillsdale, NJ, pp. 3–30

Kellogg R T 1994 *The Psychology of Writing*. Oxford University Press, Oxford, UK

Lorch M P 1995 Disorders of writing and spelling. In: Kirshner H S (ed.) *Handbook of Neurological Speech and Language Disorders*. Marcel Dekker, New York, pp. 295–324

Madigan R, Johnson S, Linton P 1995 The language of psychology: APA style as epistemology. *American Psychologist* **50**: 428–36

McCutchen D 1996 A capacity theory of writing: Working memory in composition. *Educational Psychology Review* **8**: 299–325

Rafoth B A, Rubin D L (eds.) 1988 *The Social Construction of Written Communication*. Ablex, Norwood, NJ

Traxler M J, Gernsbacher M A 1993 Improving written communication through perspective-taking. *Language and Cognitive Processes* **8**: 311–34

Voss J F, Vesonder G T, Spillich G J 1980 Text generation and recall by high-knowledge and low-knowledge individuals. *Journal of Verbal Learning and Verbal Behavior* **17**: 651–67

R. T. Kellogg

Writing Systems

The term 'writing system' specifically denotes a set of symbols which is used for the graphic (written) recording of language. Its use is commonly synonymous with that of the term 'script,' and the two are used interchangeably; however, the notion 'writing system' may additionally include the conventions (or spelling practices) which dictate how the symbols of such a system can be used. Forms of graphic expression which may have some semantic content—such as cave drawings, petroglyphs, icons, and even sophisticated picture messages—but which do not or could not record a linguistic utterance directly, are thus excluded from the realm of *writing*. This is not to suggest that all writing systems are capable of recording language with equal precision, but that by definition writing systems are fundamentally linguistic.

1. *Types of Writing System*

Writing systems are of three fundamental types. Logographic systems consist of symbols, called logograms, which represent entire words. In opposition to the two remaining types, a logographic system is said to be nonphonetic to the extent that its minimal elements record the pronunciation of an entire word, rather than representing the individual phonetic components (consonants and vowels) which make up the word. The reality of the phonetic opacity of a logographic system becomes obvious when, for example, one works with an ancient language written with logograms. It may be possible to identify the meaning of a logogram by its context and other clues, but if the word which it represents is not also attested by a more revealing, phonetic writing system, its pronunciation will remain uncertain or unknown.

In contrast to logographic scripts, syllabic and segmental writing systems are said to be phonetic. A syllabic system consists of symbols, syllabograms, which have the value of a single syllable (although the syllables of a word spelled with a syllabic script may not actually line up with the syllabograms used in a one-to-one fashion). Syllabaries most commonly possess symbols of the value V (vowel) and CV (consonant + vowel); CCV, CVC, VC, among still other permutations, also occur, though less frequently.

Segmental writing systems are those which are composed of symbols representing individual sounds, and are themselves of two types. Consonantal scripts (or alephbeths) have symbols for consonants only. Alphabets are characterized by both consonant and vowel symbols.

The relationship of the three types of writing system to one another is in large measure of a historical, developmental nature. Consequently, the ensuing discussion will roughly follow a chronological scheme. As will be seen, most writing systems are not purely logographic or phonetic, but display some degree of hybridization.

2. *The Birth of Writing*

On the basis of presently available evidence, it is clear that human beings first began to record language in the ancient Near East. Until quite recently, the very earliest examples of writing were to be found among the archaeological remains of ancient Mesopotamia. However, recent discoveries may place the appearance of writing in Egypt at approximately the same time as its first attestations in the Fertile Crescent.

2.1 *Mesopotamian Writing*

Writing appears among the Sumerians of southern Mesopotamia in the late fourth millennium BC, and it is the Sumerians who are usually credited with the invention of writing, although it appears to be attested not much later in Syria (at Tell Brak) and in Iran (in texts preserving the poorly understood language of Proto-Elamite). The events which led to the advent of Sumerian writing are of course unknown, although there are various theories regarding its appearance. In the view of some scholars, the Sumerian invention of writing was a deliberate act, probably occasioned by the needs of palace administrators, and most likely in the city of Uruk where it is first attested. Others would see the process as being more of an evolutionary, in effect accidental, nature.

Regardless of the nature of the crucial step which led to the appearance of a system of writing on the Sumerian Plain, it seems undeniable that a precursor of the script is to be found in the clay tokens which had been in use for record keeping in Mesopotamia since at least 8000 BC. Some of these tokens actually resemble the commodity which they depict; others are of an arbitrary shape. By the latter half of the fourth millennium BC, some tokens had begun to be used as stamps, being impressed into moist clay tablets to produce administrative ledgers. Other tokens, those of more complex design, began to be etched freehand in the clay using a stylus. Schmandt-Besserat (1996) has argued that the appearance of Sumerian logographic writing was a natural evolutionary step following from the practice of inscribing token images (a process

linked, she argues, to the development of symbols for abstract numerals).

As suggested by their conjectured association with tokens, the earliest Sumerian symbols are pictographic; in other words, the symbols have a picture-like appearance. The curved lines of the pictographic characters began, however, to give way to more angular, conventionalized shapes. Such conventionalization of symbols is a common process in the evolution of writing systems, but in this instance that process and the direction of development were dictated by the medium and instruments of writing. The chief medium used throughout the history of Sumerian writing was clay. Sumerian scribes incised their pictographic characters in wet clay tablets using pointed styli. These, however, were replaced by a stylus with a triangular-shaped head which was simply pressed down into the clay, allowing for more efficient production of the symbols. Characters produced in this way have a characteristic wedge-shape (quite pronounced by the middle of the third millennium BC), and are hence called cuneiform (from the Latin *cuneus*, meaning 'wedge').

The value of early Sumerian symbols is logographic: each symbol represents an entire word. This logographic system is characterized by extensive homophony and polyphony. Homophony refers to the existence of multiple logograms having the same pronunciation, as they spell phonetically identical but semantically different words (homophones). Polyphony is, in a sense, the very opposite; the term denotes the practice of using a single logogram to spell two or more conceptually related but phonetically different words. For example, a single logogram is used to spell both DINGIR 'god' and AN 'heaven' (by Sumerological convention, Sumerian words are transcribed using capital letters).

A subset of the Sumerian logograms did double duty as semantic markers, called determinatives. These were written in conjunction with another logogram and were used, in effect, to provide the reader with a clue as to the meaning of that logogram, although their use is both conventionalized and optional. For example, the logogram GIŠ, 'wood,' is used as a determinative with logograms denoting items made of wood and tree names, such as ^{giš}ERIN, 'cedar' (determinatives are commonly transcribed as superscript). Similar devices also occur in the logographic Egyptian and Maya scripts, and in a somewhat different form in Chinese writing as well.

Sumerian is an agglutinating language with complex word structure involving the joining of multiple affixes in sequence. Logographic spelling thus obscures much word-internal morphology. Perhaps it was chiefly this inherent shortcoming which led to a significant typological alteration in the Sumerian logographic writing system—the introduction of syllabic symbols (by about 3000 BC). The method employed in devising such symbols was that of a rebus. Just as in an English rebus, a picture of the sun, for example, could be used to 'spell' the word 'son,' so Sumerian logograms began to be used to represent syllables having the same sound. Sumerian scribes used the syllabic symbols not only for representing grammatical elements but also as phonetic complements, syllabic characters used to distinguish between polyphonous logograms. For example, the logogram KA was used to spell 'mouth' as well as related notions such as 'word,' 'speak,' 'voice,' and still others; attaching the phonetic complement *-ma* (i.e., KA-*ma*) signals that the logogram is being used to spell INIM 'word.'

Late in the third millennium BC, texts begin to appear in the Mesopotamian archaeological world which are written in the cuneiform script of the Sumerians, but which are composed in a completely different language, the Semitic language Akkadian. The Assyrians and Babylonians, speakers of Akkadian, greatly extended the use of syllabic symbols, but continued to utilize Sumerian logograms and determinatives as well. The Akkadian cuneiform script was transmitted to other peoples of the Near East, providing writing systems for the Hittites and Luvians, the Hurrians and Urartians, the Elamites, and, in a significantly streamlined form, the Persians. The latest known Akkadian cuneiform texts were produced in the first century AD.

2.2 Egyptian Writing

While the pictographic symbols of the Egyptian hieroglyphic script are clearly graphically distinct from those of Sumerian, and the internal operation of the two scripts shows appreciable differences, similarities are undeniable. Both possess a logographic component; both have a phonetic component, which is commonly used for phonetic complementation; both expand the use of symbols through the rebus principle; both use a subset of logograms as determinatives. As it has appeared that the Egyptian script is first attested a bit later than the Sumerian (about 3100–3000 BC), as there is evidence for a precursor to the Sumerian logograms, and as Egypt was otherwise being influenced by Mesopotamia at the end of the fourth millennium BC, it has been commonly conjectured that at least the idea of writing, if not a writing system itself, came to the Egyptians from Mesopotamia. One possible difficulty with this scenario is that the Egyptian writing system appears well developed at its earliest attestation, perhaps suggesting some earlier, undocumented period of use.

Quite recently, moreover, small tags of bone and ivory, about postage-stamp size, have been discovered on which are carved symbols that, in some cases, bear similarity to later-known hieroglyphs. These have been dated to around 3400–3200 BC by the German archaeologist, Günter Dreyer, who discovered them,

and the symbols have been interpreted by Dreyer as writing. The dating of the tags and, perhaps more crucially, the interpretation of their markings as writing requires further investigation.

The most striking difference between the Egyptian and Sumerian systems is the nature of the phonetic component. In Egyptian writing the phonetic symbols do not have syllabic value, but represent consonants—either single consonants or sequences (although not necessarily phonetically contiguous strings) of two or three consonants. Twenty-six monoconsonantal symbols occur (including two variant forms), representing the individual consonants of the language; there are almost 100 biconsonantal symbols, and somewhat fewer triconsonantal. Biconsonantal and triconsonantal phonetic symbols are often themselves accompanied by one or more monoconsonantal phonetic complements. For example, the triconsonantal symbol *nfr*, spelling the Egyptian word meaning 'good,' is commonly used in conjunction with one or two monoconsonantal complements, as in the spelling sequence *nfr f r*.

3. *Consonantal Writing Systems*

During the early second millennium BC, a Semitic people living and laboring in Egypt developed a script for their own language by exploiting a single component of the Egyptian writing system. With the inspiration provided by the monoconsonantal Egyptian symbols, these people designed a fully phonetic and greatly simplified writing system which consists entirely of symbols for individual consonants. The first evidence of the script, called Proto-Sinaitic, came to light in the early twentieth century with Sir Flinders Petrie's excavations of Egyptian turquoise mines at Serabit el-Khadim in the Sinai; hence the name. Quite recently, additional examples of Proto-Sinaitic writing have been discovered near the ancient city of Thebes, at Wadi el-Hol, perhaps dating back to as early as 1900 BC.

The method which the adapters employed in turning the Egyptian script to their own use was dubbed the acrophonic principle by the British Egyptologist Sir Alan Gardiner. The values assigned to the phonetic characters of this earliest Semitic script—characters of pictographic form—were determined by the initial consonant of the Semitic word for the symbol depicted, rather than by any phonetic value associated with a corresponding Egyptian symbol. For example, the Egyptian hieroglyphic symbol resembling waves of water and used in the Egyptian system as a monoconsonantal symbol with the value /n/ was assigned the value /m/ in the Proto-Sinaitic script, as the word for 'water' in the Semitic language of that script began with /m/. The Proto-Sinaitic character for /n/, on the other hand, was based on the Egyptian symbol representing /dŽ/, a hieroglyph depicting a serpent, and received its Semitic value from the initial consonant in the Proto-Sinaitic word for 'snake.'

Proto-Sinaitic is one particular form (the earliest by the interpretation of an origin within Egypt) of a more broadly, if meagerly, attested consonantal script called Proto-Canaanite, found at several sites in Syria-Palestine. This Proto-Canaanite script is the immediate ancestor of the Phoenician script (appearing in the eleventh century BC), which is in turn the source of the ancient Hebrew (*ca.* ninth century) and Aramaic (*ca.* eighth century) writing systems (the Hebrew language began to be written with the Aramaic script during the sixth century BC, although the Samaritan sect continued to use the old Hebrew letters, and continues its liturgical use to the present day). The Aramaic script gave rise to the writing system of Syriac (*ca.* first century AD, and late enough to be influenced by the Greek alphabet) and, through the particular Aramaic script used by the Arab Nabateans, to the Arabic writing system (*ca.* fourth century AD).

Proto-Canaanite writing also served as the inspiration and model for the script of the ancient Levantine city of Ugarit. The Ugaritic writing system, however, departs markedly from the other Canaanite scripts in that its symbols are of cuneiform shape. It thus represents something of a hybrid writing system, combining the consonantal nature of Proto-Canaanite with the graphic form of the Mesopotamian scripts, although some scholars have argued that the shapes of the Ugaritic consonantal characters were themselves modeled on the Proto-Canaanite symbols but executed with the cuneiform stylus. The Ugaritic writing system also differs from other Canaanite scripts to the extent that it does not consist wholly of consonantal characters, but includes in its graphemic inventory three CV syllabic symbols, representing a glottal stop followed by the vowels /a/, /i/, and /u/.

Yet a different Semitic writing tradition would descend from the Proto-Canaanite script. In the south of the Arabian peninsula a distinct consonantal script had appeared by at least the sixth century BC, and certainly had begun to develop several hundred years earlier. An interesting feature of the Old South Arabian script is that the order of its letters departs from that of Phoenician and the consonantal script of Ugarit. However, an abecedary found at the Palestinian site of Beth Shemesh, consisting of consonantal cuneiform characters like those in use at Ugarit, shows the same letter order as found in the Old South Arabian script. The consonantal script of South Arabia spread northward to be adapted for use by early North Arabian languages (distinct from Arabic) and westward to the horn of Africa, where it was employed for writing the Semitic language of Ethiopic. Interestingly, the consonantal script of Ethiopic had undergone a significant metamorphosis by the fifth century AD, becoming a syllabary of almost 200 CV symbols.

4. Mediterranean Syllabaries

4.1 Crete

At about the time a Semitic people resident in Egypt were fashioning a consonantal script out of Egyptian hieroglyphs, three writing systems emerged among the Minoans of Crete. Probably the first to appear was the pictographic script called Cretan Hieroglyphic. Dating back to about the eighteenth century BC is a second Minoan script, attested on only a single document, the Phaistos Disk. A third script, also appearing in the eighteenth century and likely to be descended from Cretan Hieroglyphic, is the one that was dubbed Linear A by Sir Arthur Evans, the British archaeologist who excavated the Minoan city of Knossos early in the twentieth century. The name refers to the linear nature of this script's symbols in contrast to the pictographic quality of symbols in the other two Minoan scripts. None of these ancient Minoan scripts has yet been deciphered by modern investigators, although recent work on Linear A may suggest that it records an Indo–European language of the Anatolian subfamily (an idea first proposed in the mid-twentieth century by the British classicist and linguist L. R. Palmer). Cretan Hieroglyphic is possibly a syllabic script, as is probably that of the Phaistos Disk; Linear A is certainly a syllabary.

In his excavations at Knossos, Evans unearthed documents written in yet a different script. Bearing an obvious similarity to Linear A, this writing system was labeled Linear B. In 1952 the Linear B script was successfully deciphered by Michael Ventris, a British architect working in collaboration with the classicist John Chadwick. The language of the Linear B documents turned out to be Greek—that dialect used by the Mycenaean Greek culture of the second-half of the second millennium BC. The identification of the language of the Linear B script was made particularly difficult by the seeming unsuitability of the script for spelling that language. The great majority of the symbols of this syllabic writing system are of the type CV. As Greek has many word-internal CC sequences and word-final consonants, spelling practices were devised which in some cases required the use of CV symbols for spelling consonants only (with no following vowel), and in some cases resulted in a consonant not being spelled at all. The basic strategy for spelling consonant sequences is one which coopts a phonological property of the Greek language, namely the relative sonority of consonants. It is the general case that a consonant is written when it precedes a consonant of greater sonority (such as a stop occurring before a fricative), utilizing the CV symbol whose vocalic component is identical with the vowel which follows the consonant sequence, but is not written if the ensuing consonant is of less sonority (such as a fricative before a stop). In either case, the second consonant is written (with the appropriate CV symbol). Word-final consonants are not written and this practice appears to be a further expression of the sonority-based strategy.

4.2 Cyprus

There is evidence for writing on the eastern Mediterranean island of Cyprus as early as the middle of the second millennium BC. The writing system in use is preserved in distinct forms called Archaic Cypro-Minoan, and Cypro-Minoan I, II, and III (although the last type is actually attested in the documentary remains of the city of Ugarit, on the neighboring Levantine coast, and not on Cyprus itself). In spite of the fact that none of the three has yet been deciphered, the formal similarities between these scripts and those of Minoan Crete clearly suggest that the Bronze Age Cypriot writing system had its origin in the Cretan scripts.

Cyprus was a key locale lying along the trade routes which stretched from Greece to the east, and certainly for this reason there had been a Greek presence in Cyprus during the Mycenaean era; however, the Greek population of Cyprus increased dramatically as refugees fled the conflagration which brought an end to the Mycenaean civilization in the Greek homeland. The earliest remains of Greek writing in Cyprus come from the middle of the eleventh century BC, and the system of writing used was that of the Cypriot Syllabary. This syllabary, while probably based graphically upon Cypro-Minoan 1, is obviously closely affiliated with the Mycenaean script. Like Linear B, the Cypriot Syllabary is composed predominantly of symbols having the value CV. The same problems of representing word-internal consonant sequences and word-final consonants which Linear B scribes faced also plague the Cypriot Greek writing system; and the solutions to these problems are virtually identical to those utilized by the Mycenaeans, being based upon the relative sonority of Greek consonants. The notable difference between the two systems in this regard is that the Cypriot writing system makes much more limited use of the CV-deletion strategy than does Linear B. In most instances in which the first of two consonants is of greater sonority, the Cypriot strategy is to write that consonant (and not to refrain from writing it, as in Linear B) using a CV symbol whose vocalic component is identical to the vowel which precedes the consonant sequence.

5. The Alphabet

5.1 The Greek Origin

It was the Greeks who created the first segmental writing system to represent not only consonants but vowels as well. The raw material for the Greek

alphabet was provided by the consonantal script of the Phoenicians. As there was both a significant Phoenician presence on the island of Cyprus, and a vibrant tradition of Greek literacy, Cyprus is a *de facto* likely candidate for the place at which the Greek alphabet came into existence, probably in the early eighth (or possibly late ninth) century BC. There are, moreover, particular features of the early Greek alphabet which are not motivated within that segmental script, but which are motivated within the Cypriot Syllabary. For example, the biconsonantal symbol *xi*, representing the sequence of /k/ + /s/, is unnecessary within a segmental script that has both a *k*-symbol (*kappa*) and an *s*-symbol (*sigma*). Within the Cypriot Syllabary, however, *ksV* symbols are idiosyncratically required by the strategy of consonant cluster spelling which is based upon the relative sonority of consonants. Cypriot Greeks appeared to have transferred this and other features from their syllabic script to the newly devised alphabet.

For vowel symbols, the Greeks utilized Phoenician consonantal characters which were not required for writing Greek. Thus, Phoenician *aleph*, symbol for a glottal stop (which sound Greek did not possess phonemically) was assigned the Greek value /a/ (the Greek letter *alpha*). Phoenician *he*, spelling the glottal fricative /h/, was used by the Greek adapters to record mid front (*e*) vowels (*epsilon*; for writing Greek /h/, Phoenician *ḥet* was employed, the symbol for a voiceless pharyngeal fricative /ħ/). To write mid back (*o*) vowels (*omicron*), the Greeks tapped Phoenician *ayin*, which represented a voiced pharyngeal fricative /ʕ/. Phoenician *yod*, spelling the palatal glide /y/, and *waw*, representing /w/, were assigned the Greek values /i/ (*iota*) and /u/ (*upsilon*) respectively.

5.2 The Etruscan Alphabet

Following rapidly upon its creation, the alphabet was widely disseminated throughout the Greek world, developing many local varieties. In the south of Italy, Greek colonists passed their western Greek alphabet on to a non-Indo-European people of Italy, the Etruscans, in the seventh century BC. As the Etruscan language lacked voiced stops, the Greek letters representing such sounds, *beta* (/b/), *delta* (/d/), and *gamma* (/g/), were not needed for that purpose. The first two letters would disappear from the Etruscan alphabet, although *gamma* was retained and used for spelling the voiceless stop /k/ when it occurred before a front vowel. The Etruscans used Greek *kappa* for /k/ before /a/ and Greek *qoppa* for /k/ before the high back vowel /u/. In turn, the Etruscans transmitted the alphabet to other peoples of Italy.

5.3 The Roman Alphabet

Among those places to which the Etruscan writing system came was the city of Rome—a singly significant step in the history of the alphabet. For a time, the Romans continued the Etruscan practice of using three different symbols to represent the velar stop /k/ in its various phonetic contexts. However, *C* (from Greek *gamma*) was eventually used regularly for /k/ as well as for /g/ (*K*, Greek *kappa*, continued to be used for a very few words and *Q*, Greek *qoppa*, was employed before *V* for spelling /k^{w}/). By the third century BC, a variant form of *C* had been introduced, namely *G*, for writing the voiced velar stop /g/. However, as *G* was assigned the alphabetic position held by Greek *zeta* (spelling the early Greek sequence /zd/, later /z/), which was not required for spelling Latin, it has been argued that *zeta*, not *gamma*, provided the raw material for the creation of the new letter *G*. Later, in the first century BC, with the influx of much Greek vocabulary, both *Y* and *Z* (for Greek /ü/ and /z/ respectively) were appended to the Roman alphabet.

5.4 The Alphabet Abroad

The alphabet would make its presence known, directly and indirectly, throughout the European continent and beyond. The fragmentary remains of Celtic writing from continental Europe are attested in alphabetic scripts—Etruscan, Roman, or Greek (aside from the Hispano-Celtic inscriptions written in a curious script which is part syllabic, part alphabetic). Moreover, the 'native' writing systems of both the ancient Celts of Britain and Ireland, and of the Germanic peoples of Europe are alphabetic traditions, although formally distinct from both the Greek and Roman scripts, and strikingly so in the case of Celtic Ogham, with its characters of oblique and right-angle lines. Nevertheless, it is undoubtedly the case that both Ogham and the Germanic runes found their inspiration and model in the Mediterranean alphabets. When in the fourth century AD the Gothic Bishop Wulfila translated portions of the Bible into his Germanic language, he used an alphabetic script of his own design, based chiefly on the Greek alphabet.

The spreading of the Christian faith and the effort to ground believers in the biblical texts were—far beyond Wulfila—fundamental factors in the evolution of alphabetic writing systems. The Roman Catholic Church played a key role in the introduction and eventual establishment of the Roman alphabet across much of the European continent. Carried by the colonizing European nations, this writing system would be transported around the world, being utilized for writing European languages on all six of the populated continents, and adapted for spelling indigenous languages of Asia and Africa, as well as providing raw material for the symbols of the Cherokee syllabary of North America.

In the ninth century AD, Greek missionaries led by Saint Cyril and Saint Methodius developed writing

systems, based on their own Greek alphabet, for translating the Bible into the language of south Slavic converts. One of these scripts, the Cyrillic alphabet, thrived and is the predecessor of the alphabetic scripts of various modern Slavic languages, most notably Russian.

The alphabets of both Armenian and Georgian are traditionally attributed to the fifth-century cleric, Saint Mesrop. These writing systems were based at least partially on the Greek alphabet, but their creator may have also drawn upon a Semitic consonantal script.

6. *Writing in South and East Asia*

6.1 *The Indian Syllabaries*

The earliest attested writing system of India is the script of the Indus Valley civilization, dating back to *ca.* 2500 BC. Many of the symbols are of a pictographic nature, and the number of symbols identified suggests that this writing system is logographic, possibly partially syllabic. It is, however, presently poorly understood and does not appear to have been the source of the Indian syllabaries which appear in the third century BC. These scripts, Kharoshthi and Brahmi, probably developed from the Aramaic consonantal writing system. Of the two, only Brahmi appears to have survived late antiquity; its perhaps best-known form being Devanagari, the writing system of Sanskrit, as well as that of Hindi and other modern Indo-Aryan languages.

Devanagari differs notably from the Mediterranean and Mesopotamian syllabic writing systems in that rather than having a distinct symbol for each possible consonant + vowel combination, Devanagari consists of a basic set of CV symbols having the vowel value -a (i.e., Ca symbols). These CV symbols can be assigned different vowel values (Ci, Cu, etc.) by the addition of diacritic strokes (this is essentially the same strategy utilized by the Ethiopic syllabary, which traces its origins to a Semitic consonantal writing system, as apparently do the Indic scripts). Devanagari also has the capacity to transform basic CV symbols into ones representing more complex consonantal components (CCV, CCCV, etc.) through the use of conventionalized ligatures. Such ligatures consist essentially of a Ca symbol (representing the final member of the consonant cluster) to which a form, often highly abbreviated, of one or more other Ca symbols is attached. Word-final consonants are spelled utilizing the appropriate Ca symbol beneath which a diacritic mark, called virama, is placed, revealing to the reader that the symbol is to be read with no vowel component.

The Indian syllabaries were spread through much of Central and Southeast Asia, giving rise to writing systems for numerous languages. Not only does the Tibetan script trace its roots to India, but also the writing systems of Javanese, Thai, Burmese, Lao, and Khmer.

6.2 *Chinese Writing*

The earliest evidence for writing in China is provided by the Shang-dynasty oracle-bone inscriptions of *ca.* 1200 BC. There is no clear evidence that writing developed in China under the influence of any then existing writing system; as in Mesopotamia (and perhaps in Egypt), writing seems to have appeared *de novo*. Like the earliest Mesopotamian writing system, the early Chinese script is pictographic with each

Table 1
The Greek alphabet

Symbol	Value	Symbol	Value
Α, α	a	Ξ, ξ	ks
Β, β	b	Ο, ο	o
Γ, γ	g	Π, π	p
Δ, δ	d	Ϻ	s
Ε, ε	e	Ϙ	k
Ϝ	w	Ρ, ρ	r
Ζ, ζ	zd	Σ, σ	s
Η, η	ē	Τ, τ	t
Θ, θ	t^h	Υ, υ	u
Ι, ι	i	Φ, φ	p^h
Κ, κ	k	Χ, χ	k^h
Λ, λ	l	Ψ, ψ	ps
Μ, μ	m	Ω, ω	ō
Ν, ν	n		

Table 2
The Ugaritic cuneiform consonantal script

Symbol	Value		Symbol		Value
𐎀	å	[ʔa]	𐎛	ḏ	[ð]
𐎁	b		𐎜	n	
𐎂	g		𐎝	ẓ	[ð’]
𐎃	ḫ	[x]	𐎞	s	
𐎄	d		𐎟	ʿ	[ʕ]
𐎅	h		𐎠	p	
𐎆	w		𐎡	ṣ	[s’]
𐎇	z		𐎢	q	[k’]
𐎈	ḥ	[ħ]	𐎣	r	
𐎉	ṭ	[t’]	𐎤	ṯ	[θ]
𐎊	y		𐎥	ǵ	[ɣ]
𐎋	k		𐎦	t	
𐎌	š		𐎧	i	[ʔi]
𐎍	l		𐎨	ù	[ʔu]
𐎎	m		𐎩	ṡ	[s] (?)

Table 3
The basic CV symbols of Devanagari

Symbol	Value	Symbol	Value	Symbol	Value
क	ka	च	ca	ट	ṭa
ख	kha	छ	cha	ठ	ṭha
ग	ga	ज	ja	ड	ḍa
घ	gha	झ	jha	ढ	ḍha
ङ	ṅa	ञ	ña	ण	ṇa
त	ta	प	pa		
थ	tha	फ	pha		
द	da	ब	ba		
ध	dha	भ	bha		
न	na	म	ma		
य	ya	श	śa		
र	ra	ष	ṣa		
ल	la	स	sa		
व	va	ह	ha		

symbol representing an entire word. Moreover, as in both Mesopotamia and Egypt, the Chinese writing system is one which multiplies the value of its characters utilizing the rebus principle; and one which utilizes semantic markers to minimize graphic ambiguity. In China, however, these markers were commonly combined with the associated characters to produce compound symbols of two components—one signaling pronunciation, the other meaning. The fundamental logographic nature of Chinese writing has been preserved to the present day, and the inventory of logographic characters has steadily increased. While approximately 60,000 distinct characters are currently identified, about 2,500 occur in common use.

6.3 Japanese Writing

The modern Japanese writing system is quite complex, probably the most so of modern scripts. The earliest attested examples of Japanese writing (*ca.* eighth century AD) are executed in Chinese script. However, by the end of the ninth century, two Japanese syllabaries had appeared, evolving out of the practice of using certain Chinese characters as phonetic symbols. Both of these syllabaries remain in use alongside kanji, essentially logographic symbols originating in Chinese characters; in the contemporary script, a single such symbol can have multiple values. The syllabary called katakana is utilized for scientific terms, for foreign loans, for indicating emphasis, among still other uses. Modern hiragana is used for spelling grammatical particles and affixes associated with kanji.

6.4 Korean Writing

As in the case of Japanese, the Korean language was first written with Chinese characters. It was in the fifteenth century AD, according to Korean tradition, that the Korean alphabetic writing system of Han'gul was invented. Credit for this script—the world's only alphabetic writing system which may not trace its roots to the Greek alphabet—is given to Sejong, a Korean king. Han'gul is a highly efficient and phonetically transparent script. Its various consonantal symbols are purported to have been designed to represent schematically the articulatory position of the tongue in producing the consonant sounds of Korean. While Han'gul is alphabetic, consonant and vowel characters are written together in such a way as to form orthographic blocks corresponding to phonetic syllables.

7. Writing in Mesoamerica

There is at least a third region of the world, in addition to China and the Near East, where human beings appear to have designed a means of writing without prior knowledge of any such thing, namely Mesoamerica. While the study of ancient Mesoamerican writing is still at an early stage, significant strides have been made since the 1970s. Of the several Mesoamerican scripts currently known, the best attested and best understood is that of the Maya, who were writing by at least the third century AD. In its fundamental characteristics, the Maya writing system is much like those of the Sumerians and Egyptians. The script's pictographic symbols, or glyphs, are logographic and syllabic in value. The latter type is used both in conjunction with logograms as well as for the phonetic spelling of words. In addition, the Maya system makes use of determinatives, like the Sumerian and Egyptian systems, to provide semantic guidance to the reader.

See also: Reading Nonalphabetic Scripts; Writing, Evolution of; Writing Instruction; Writing Process, Psychology of; Writing Systems, Psychology of

Bibliography

Coulmas F 1989 *The Writing Systems of the World.* Blackwell, Oxford, UK

Coulmas F, Ehlich K 1996 *A Bibliography on Writing and Written Language*. De Gruyter, Berlin

Daniels P T, Bright W (eds.) 1996 *The World's Writing Systems.* Oxford University Press, Oxford, UK

Diringer D 1968 *The Alphabet: A Key to the History of Mankind*, 3rd edn. Funk and Wagnalls, New York
Driver G R 1976 *Semitic Writing from Pictograph to Alphabet* [rev. edn. Hopkins S A (ed.)]. Oxford University Press, Oxford, UK
Gelb I J 1963 *A Study of Writing*, 2nd edn. University of Chicago Press, Chicago
Günther H, Ludwig O (eds.) 1994 *Schrift und Schriftlichkeit*. De Gruyter, Berlin
Hooker J T 1990 *Reading the Past: Ancient Writing from Cuneiform to Alphabet*. The British Museum, London
Robinson A 1995 *The Story of Writing*. Thames and Hudson, London
Sampson G 1985 *Writing Systems: A Linguistic Introduction*. Stanford University Press, Stanford, CA
Schmandt-Besserat D 1996 *How Writing Came About*. University of Texas Press, Austin, TX
Senner W M (ed.) 1989 *The Origins of Writing*. University of Nebraska Press, Lincoln, NE
Woodard R D 1997a Writing systems. In: Comrie B, Matthews S, Polinsky M (eds.) *The Atlas of Languages*. Bloomsbury, London
Woodard R D 1997b *Greek Writing from Knossos to Homer*. Oxford University Press, Oxford, UK
Woodard R D 2002 *Cambridge Encyclopedia of the World's Ancient Languages*. Cambridge University Press, Cambridge, UK

R. D. Woodard

Writing Systems, Psychology of

Although the psychological study of *reading* and the psychology of *language* are well established domains of research and theory, the *psychology of writing systems* and their uses, *literacy*, are relatively recent additions to the cognitive sciences. The central goal of the psychology of writing systems is to bring these two, hitherto independent, fields of study into conjunction, allowing both to be seen in a new light.

Writing systems have been neglected in psychology for two reasons. First, twentieth century linguistics has focused almost exclusively on speech, consigning writing merely to a way of recording. The study of writing was left to specialist domains such as epigraphy and classical studies. Nonetheless, important studies of writing were produced, that by Gelb (1963) providing the foundation for more recent work. Second, only the alphabetic writing system was the object of psychological studies of reading, leaving the impression that other writing systems were primitive and scarcely worth careful study.

In the period since the 1960s, the two fields, the psychology of reading and the psychology of language, have begun to converge to produce several lines of new research and theory. These relations make up the major sections of this article.

1. Relation Between Speech and Writing

Making marks which can serve mnemonic and communicative purposes is as old as human culture itself. Writing systems are a subset of such marks which, typically, represent aspects of linguistic form. Sampson (1987) provided an abstract classification of types of writing system. Some systems, *semasiographic* systems, represent meanings without any direct link to spoken form. Historical examples include the so-called 'picture writing' of American Indians, more modern ones include systems of highway signs and mathematical symbols. Some writers (Coulmas 1989, Harris 1995) consider such nonspeech sign systems as full writing systems, claiming that what makes a writing system is its adequacy in conveying information visually rather than its purported links to speech.

Logographic scripts are linguistically based but the spoken elements represented are meaningful units, primarily words and morphemes. A clear example of such a script is Chinese in which morphemes which sound identical are often represented by entirely dissimilar graphic characters, although some theorists (de Francis 1989) insist that any full writing system is phonologically based, at least to some extent.

Syllabaries are scripts which represent spoken patterns by signs for syllables. Thus *ka*, *ke*, *ki*, *ko* would be written with the same graphic sign. Arabic and Hebrew are modern examples.

Finally, *alphabets* employ a separate mark to represent the consonant and vowel segments of a syllable, roughly, the phonemes of the language. These categories of writing systems are generic types; no living writing system is based exclusively on one type of sign. Thus in English script, signs represent both phonemes by means of letters and words by means of spaces between letter groups. Japanese writing uses Chinese logographs (Kanji) to represent the stems of content words and a completely different syllabic script (Kana) to represent grammatical morphemes and inflections (Taylor and Taylor 1995).

In one sense, the history of scripts reflects the discovery of increasingly subtle aspects of speech, going from meanings to analytical constituents of speech. Yet contrary to the evolutionary view (Gelb 1963), many researchers (Gaur 1987, Sampson 1987, Olson 1994, Harris 1995) have argued that these so-called progressive changes merely reflect the impact of borrowing a script invented for one language to represent a completely different language. The consequence of such borrowing was the discovery of previously ignored properties of speech, word signs, syllable signs, and phoneme signs.

2. Writing Systems and Reading

As different types of writing systems bear distinctively different relations to speech, learning to read and write

them call upon different cognitive activities. Semasiographic systems found in Ancient Mesoamerica, namely the Mixtec and Aztec systems used until the conquest for record keeping, depended highly on rich factual knowledge; one had to know what the inscription was about before one could read it. Further, it depended heavily upon interpretation as the inscriptions allowed variable readings. Indeed, there is no categorical difference between the writing and drawing. Nineteenth-century Canadian Blackfoot Indians, for example, depicted 'the year that horses got drowned' by means of a circle (representing a lake) and four stick-drawings of horses superimposed on the circle. Such a chronicle records an important cultural event but not the linguistic form of its description. That is why it allows such variable readings.

All other so-called full writing systems map on to or represent spoken forms and learning to read each of them requires a somewhat different orientation to the script. Scribner and Cole (1981) found that their Vai subjects in Liberia in learning to read a syllabic script became adept at analyzing their speech into syllabic units but failed to analyze their speech into words. Others have found that the ability to manipulate phonemes was tied to learning to read an alphabet. Read et al. (1986) found that Chinese readers of traditional character scripts could not detect phonemic segments whereas those who could read Pinyin, an alphabetic script representing the same language, could do so.

Ample empirical evidence exists to show that it is acquaintance with an alphabetic script that is critical to the acquisition of awareness of the phonological properties of the language (Morais et al. 1979, Vernon and Ferreiro 1999). Concepts of words as lexical entities, similarly, have been shown to be tied to concepts of words as written entities (Olson 1994). Concepts of sentences, needed for the mastery of punctuation, are acquired at even more advanced levels of literacy. Conceptual development in children is, in part, the consequence of the acquisition of these systems for representing language (Vygotsky 1986).

Much confusion about the cognitive and social implications of writing and reading arises from a failure to recognize that modes of reading and interpretation do not inevitably arise from simply learning to identify words and letters. The history of writing is largely a history of inventing devices, such as punctuation and text structures, as well as rules for interpretation that have taken sometimes millennia to develop. A dictionary itemizing all the words of a language is a recent, literate invention. The very idea of fixity of a text is not inevitably attributed to writing and printing but to a specialized set of conventions for interpreting texts and institutional structures for preserving them (Johns 1998). However, once procedures had been developed for the systematic interpretation of Scriptural texts, it was found that such procedures were also suitable for the interpretation of the 'book of nature.' Scientific epistemology, it has been argued, is a form of applied hermeneutics (Olson 1994).

3. *Theories of Reading Alphabetic Scripts*

Assumptions about the relation between script and language have given rise to considerable, well-publicized, differences of opinion and differences in reading theory. One group emphasized the recognition of meaning of written expressions in ways analogous to the recognition of meaning in oral forms. That is, written language is seen as an alternative to spoken language each with their peculiar and distinctive perceptual and cognitive processes but with a common goal of meaning construction, hence the label Whole Language.

The other group, credit for which is often given to Alvin Liberman and Isabel Liberman of the Haskins Group at Yale University, is that reading is essentially a metalinguistic activity in which readers must find mappings from the written form to aspects of their implicit knowledge of speech (Olson 1994, Adams et al. 1998). Evidence on the importance of word and phoneme segmentation, two forms of metalinguistic knowledge, to learning to read and write tends to favor the metalinguistic theory of reading (Stanovich 1986, Lundberg et al. 1988), that is, the theory that reading is not a matter of mapping signs to meanings but rather a matter of mapping visual signs to aspects of oral speech and hence on to meaning. It is this metalinguistic theory of reading that helps to explain not only why learning to read brings aspects of speech into consciousness but also why reading and writing may contribute to a new and distinctive mode of thought in certain social and institutional contexts.

4. *Writing Systems and Thinking*

Nineteenth-century theorists took the view that writing and literacy had the general effect of making thought more logical and formal. The recognition of the fact that there are many types of writing systems and that even those different types may be read in very different ways (Saenger 1997) has made such general claims problematic, as has the dispute as to the meaning of literacy, some identifying it with the basic skills of reading and writing (Scribner and Cole 1981) and others with the evolved products of particular literate cultural traditions (Goody 1968, Olson 1994).

Although there is wide agreement on the 'psychic unity of mankind' ('They are just like you and me,' as Jack Goody once put it), and there is wide agreement that all languages are *in principle* capable of expressing all nuances of meaning, nonetheless patterns of thought vary widely from culture to culture and context to context, so much so that it is not uncommon

to despair of finding interesting cognitive differences attributable to writing or to a written tradition. Mathematics, the epitome of a written tradition, is sometimes seen as basically equivalent to practical computations, so-called street math.

On the other hand, cultural historians have asserted the cultural and psychological implications of writing and literacy, arguing, for example, that the medium of representation is the most important aspect of the message (McLuhan 1962), that writing separated the knower from the known (Ong 1992), that scriptal changes caused the origins of modern silent reading (Saenger 1997), and that writing allowed the formulation of modern conceptions of logic and mathematics (Goody 1968).

The metarepresentational theory of reading gives new credence to these more radical views. Writing and reading a linguistically based script, on this view, are an exercise not only in meaning, but also in consciousness and discourse about language, that is, about units of sound, units of meaning, words and structural units, sentences. Literacy on this view is seen as instrumental to discourse about word meaning, sentence meaning, and to other aspects of form which may be *mentioned* rather than *used* in discourse. This metarepresentational shift invites talk not only about particular words and sentences independently of their use, but also about other metarepresentational properties, such as *truth*, *logical form*, and *aesthetic goodness*. Furthermore, it allows mathematics to be clearly distinguished from 'street math,' the former being concerned with number rather than quantity. In the former, a number is expressed in the singular: *five* is more than *four*. In the latter, quantity is expressed as plural: I have seven apples, *five* are red. Only the former is metarepresentational, that is, mentioned rather than used, and consequently an object of thought in its own right.

Furthermore, writing is instrumental to thinking in general as a form of metalinguistic knowledge, that is, knowledge 'about' the lexical, grammatical, and logical properties of the language.Vocabulary knowledge, for example, is both greatly extended by reading (Anglin 1993), and reflective knowledge about words, as required for judgments of synonymy, for example, serves as a major aspect of measured intelligence in a literate society (Stanovich 1986). Formal reasoning tasks which require judgments of validity rather than truth or agreement, too, appear to be related to literacy, especially the specialized literacy of schooling (Ong 1992).

In order to account for such effects ranging from phonological awareness to awareness of sentence meaning (as opposed to what a speaker means by the sentence), Olson (in press) has proposed that writing be seen as quoted speech, and that reading and writing be viewed as learning to deal with quoted utterances rather than with ordinary direct utterances. This associates writing with a special feature of language known as reflexivity and as roughly equivalent to the use–mention distinction and language–metalinguistic distinction. Thus, in writing and reading one is not merely using an expression but quoting and thereby mentioning a linguistic form. The special interpretation called for in such a context depends upon the special properties of the form, its 'linguistic meaning' or 'sense,' rather than its referential meaning as well as allowing for judgments of similarity of meaning and judgments of logical entailments. Some writers such as Havelock (1982) and Ong (1992) argue that 'all elaborate, linear, so-called 'logical' explanation depends on writing' (Ong 1992 p. 312). Admittedly, not every culture and/or every reader exploits these distinctions as they rely heavily on particular institutional arrangements such as schools, professions, and academic disciplines. However, no reader reading a word- and sentence-segmented alphabetic text can fail to acquire the special metalinguistic concepts of sound, word, and sentence. The remainder comes with more specialized study, normally in schools.

5. *Writing as Normative Practice*

Writing does not merely transcribe speech but presents speech in terms of an idealized model. When learning to write, one is taught not to transcribe speech but rather to edit the speech by removing false starts, correcting errors, and adding needed qualifiers to make the transcription more complete, correct, and formal than the more spontaneous oral utterance. Writing in a literate culture is a normative practice, which is to say it involves the correct, authorized forms of words, sentences, and texts. Admonitions to 'speak grammatically' are in fact admonitions to speak as one would write. Although linguists abhor so-called prescriptive grammars, claiming that all utterances of a native speaker are grammatical, schools are regularly charged with teaching 'proper English.' In this way, writing comes to serve as a model for speaking.

Reading texts and composing them provide a special forum for coordinating the content of a domain with the syntactic and rhetorical demands of the written models available. van Dijk and Kintsch (1983) showed how readers coordinate linguistic form with conceptual structure in interpreting texts and Scardamalia and Bereiter (1986) showed how in writing, students coordinated their ideas with available discourse forms.

6. *Writing and Knowledge*

Although all of the peoples of the world know a great deal about nature, themselves, and the social world, writing systems in the modern world have largely preempted such local knowledge, overwriting it with institutional forms of knowledge based on textual archives and specialized interpretive practices. Have-

lock (1982) suggested that writing was responsible for divorcing the knower from the known, allowing the latter to accumulate and become the subject of research and revision when the social conditions were suitable. Contemporary forms of knowledge are without exception text based, with access to these forms limited to those willing to undergo years of specialized training in the formulation and interpretation of the principle documents of the tradition. Stock (1983) called such traditions 'textual communities' and Smith (1990), a sociologist, characterizes the organization of power in contemporary society in terms of objectivized forms of text-based knowledge: 'The primary mode of action and decision in … business, government, the professions … are textual. The realities to which action and decision are oriented are virtual realities vested in texts and accomplished in the distinctive practices of reading and writing' (Smith1990 pp. 61–2). Psychologists have only begun to explore the implications of living and working in such text-based environments.

7. *Literacy and Social Development*

Because literacy plays such a prominent role in modern societies, it is often assumed that the route to social development is through teaching people to read and write, a view supported by UNESCO. Research and practice have shown that in order to bring about cultural and social transformation, literacy must be seen as an activity embedded in social and cultural practice. Literacy, bureaucratic institutional structures with explicit procedures and accountability, and democratic participation are mutually reinforcing. Rather than being seen simply as a goal, literacy has come to be seen as a means to fuller participation in the institutions of the society, whether in law, science, or literature (Street 1984), as well as a means for their transformation.

See also: Knowledge Activation in Text Comprehension and Problem Solving, Psychology of; Language and Thought: The Modern Whorfian Hypothesis; Readability, Applied Psychology of; Reading Skills; Speech Production, Psychology of; Writing, Evolution of; Writing Instruction; Writing Process, Psychology of; Writing Systems

Bibliography

Adams M J, Treiman R, Pressley M 1998 Reading, writing and literacy. In: Damon W, Seigel I E, Renninger K A (eds.) *Handbook of Child Psychology*, 5th edn. *Child Psychology in Practice*. Wiley, New York, Vol. 4

Anglin J M 1993 Vocabulary development: a morphological analysis. *Monographs of the Society for Research in Child Development* **58**: No. 10: 1–166

Coulmas F 1989 *The Writing Systems of the World*. Blackwell, Oxford, UK

de Francis J 1989 *Visible Speech: the Diverse Oneness of Writing Systems*. University of Hawaii Press, Honolulu, HI

Gaur A 1987 *A History of Writing*, paperback edn. British Library, London (original work published in 1984)

Gelb I J 1963 *A Study of Writing*, 2nd edn. University of Chicago Press, Chicago

Goody J 1968 *Literacy in Traditional Societies*. Cambridge University Press, Cambridge, UK

Harris R 1995 *Signs of Writing*. Routledge, London

Havelock E 1982 *The Literate Revolution in Greece and its Cultural Consequence*. Princeton University Press, Princeton, NJ

Johns A 1998 *The Nature of the Book: Print and Knowledge in the Making*. University of Chicago Press, Chicago

Lundberg I, Frost J, Petersen O 1988 Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly* **23**: 264–82

McLuhan M 1962 *The Gutenberg Galaxy*. University of Toronto Press, Toronto, Canada

Morais J, Cary L, Alegria J, Bertelson P 1979 Does awareness of speech as a sequence of phones arise spontaneously? *Cognition* **7**: 323–31

Olson D R 1994 *The World on Paper*. Cambridge University Press, New York

Olson D R in press The written word: inaugural lecture as University Professor, University of Toronto. *Redefining Literacy*

Ong W 1992 Writing is a technology that restructures thought. In: Downing P, Lima S D, Noonan M (eds.) *The Linguistics of Literacy*. John Benjamins, Amsterdam

Read C A, Zhang Y, Nie H, Ding B 1986 The ability to manipulate speech sounds depends on knowing alphabetic reading. *Cognition* **24**: 31–44

Saenger P 1997 *Space Between Words: the Origins of Silent Reading*. Stanford University Press, Stanford, CA

Sampson G R 1987 *Writing Systems: a Linguistic Introduction*, revised edn. Hutchinson, London

Scardamalia M, Bereiter C 1986 Research on written composition. In: Wittrock M (ed.) *Handbook on Research on Teaching*, 3rd edn. Academic Press, New York

Scribner S, Cole M 1981 *The Psychology of Literacy*. Harvard University Press, Cambridge, MA

Smith D 1990 *Texts, Facts, and Femininity*. Routledge, New York

Stanovich K E 1986 Matthew effects in reading: some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly* **21**: 360–407

Stock B 1983 *The Implications of Literacy*. Princeton University Press, Princeton, NJ

Street B 1984 *Literacy in Theory and Practice*. Cambridge University Press, Cambridge, UK

Taylor I, Taylor M M 1995 *Writing and Literacy in Chinese, Korean and Japanese*. John Benjamins, Amsterdam

van Dijk T A, Kintsch W 1983 *Strategies of Discourse Comprehension*. Academic Press, New York

Vernon S A, Ferreiro E 1999 Writing development: a neglected variable in the consideration of phonological awareness. *Harvard Educational Review* **69**: 395–415

Vygotsky L S 1986 *Thought and Language*. MIT Press, Cambridge, MA

D. R. Olson

Copyright © 2001 Elsevier Science Ltd. 16643
All rights reserved.

 ISBN: 0-08-043076-7

Wundt, Wilhelm Maximilian (1832–1920)

1. Life and Times

Wilhelm Maximilian Wundt was born on August 16, 1832 in Neckarau, Germany. From 1844, he attended secondary schools in Bruchsal and Heidelberg. He enrolled at Tübingen for medical study in 1851, then went to Heidelberg, where he earned his doctorate in 1856 with a work 'On the Behavior of Nerves in Inflamed and Degenerated Organs.' In Heidelberg, neuroanatomist Friedrich Arnold and the clinician and pathologist Ewald Hasse guided his interest toward neuroanatomical and neurophysiological problems. During the summer of 1856, he worked under Emil Du Bois-Reymond (1818–96) at the famous physiological laboratory founded in Berlin by Johannes Peter Müller (1801–56). Returning to Heidelberg, he became Hasse's assistant. He completed postgraduate work in physiology in 1857.

In 1858, Hermann Helmholtz (1821–94) was made professor at Heidelberg. Wundt was made professor of Anthropology and Medical Psychology in 1864. For 16 years, he worked alongside Helmholtz at the Physiological Institute in Heidelberg. During this period, his orientation shifted from experimental physiology to physiological psychology, and finally to philosophy. From 1864 to 1868 he was a representative to parliament in the state of Baden, where he advocated social integration for workers. He was appointed professor of Inductive Philosophy (philosophy of science) in Zurich, Switzerland in 1874. On October 18, 1875 he accepted a professorship in Leipzig in Philosophy ('with a natural scientific orientation'). In Leipzig, he founded the world's first Institute for Experimental Psychology, which soon attained international renown. Students came to Leipzig from many countries, including Bulgaria, the UK, Japan, Canada, Poland, Romania, Russia, and Hungary, and especially from the USA. Many students went on to establish institutes in their home countries, using the Leipzig Institute and its course of study as their model. In 1881, Wundt founded one of the world's first journals of psychology, *Philosophical Studies*. Wundt died on August 31, 1920 in Großbothen near Leipzig, and was buried in the professors' section of the South Cemetery in Leipzig.

What kind of an era did Wundt live in? In the year of his birth (1832), the celebrated writer of German classical literature Johann Wolfgang von Goethe died. In 1848, the revolution that sought to achieve democratic process in Germany failed. In 1871, the Second German Reich was born after the violence of Prussia's wars against Denmark (1864), Austria (1866) and France (1870–1), and Austria was excluded from the German states' centuries-long common political development. By the time Wundt died (1920), the Second German Reich had lost World War I in 1918 and the Weimar Republic had been established. Looking back, Wundt wrote in his memoirs: 'Lived experience is the most immediate of what the gods have bestowed, but knowledge is the better portion of what he has been granted' (Wundt 1920, p. iv).

2. Work and Influence

2.1 An Overview

Early in his career, Wundt focused on the experimental analysis of physiological and then psychophysiological problems. Somewhat later, he attempted to provide a philosophical foundation for his experimental work. The titles of his works published in Heidelberg vividly illustrate his path: *Contributions to the Science of Muscle Movements* (1858), *Contributions to the Theory of Sense Perception*' (1859–62), *Lectures on the Mind of Humans and Animals* (1863), *Textbook of Physiology* (1864), *Physical Axioms* (1866), *Handbook of Medical Physics* (1867), and *Principles of Physiological Psychology* (1874). In 1902, one of his first biographers summarized: 'Evidently, it was involvement with questions of sensory physiology that first awakened in Wundt the need for a philosophical orientation' (König 1902, p. 23).

In Heidelberg, Wundt also developed the fundamentals of his ethnopsychology (*Völkerpsychologie*), stimulated by the work of Moritz Lazarus (1824–1903) and Hajim Steinthal (1823–99) in their *Journal for Ethnopsychology and Linguistics* published from 1859. Decades later in Leipzig, Wundt elaborated these fundamentals into a 'general psychology' and a 'general developmental psychology.' After 1875, he wrote the works that would bring him world renown as a methodologist, experimental psychologist, and ethnopsychologist. This progression, too, can be traced in the titles of his books: *System of Philosophy* (1889), *Logic: An Investigation of the Principles of Knowledge and Methods of Scientific Research* (finally a three-volume work, 1906–8), *Introduction to Philosophy* (1901), and the 10-volume work *Ethnopsychology* (1900–20). Between 1908 and 1911, he also expanded *Principles of Physiological Psychology*, his standard work on experimental individual psychology,' to three volumes. His students in Leipzig carried out the psychological experiments that would fill the 20 volumes of his *Philosophical Studies* and those of the subsequent *Psychological Studies*.

2.2 Wundt's Major Work—Individual Psychology and Ethnopsychology

Wundt's major work was devoted to psychology, which was for him the basis of philosophy. His psychology had a twofold basis, consisting of a scientifically oriented experimental 'individual psy-

chology' and a humanistically oriented comparative 'ethnopsychology.' About his individual psychology, he wrote, 'The concept of individual psychology covers investigations whose subject are the psychological processes of individual human consciousness, inasmuch as these possess a *typical*, generally valid significance for normal consciousness' (Wundt 1908, p. 162). He added, with regard to methodology, 'In truth, there are … for individual psychology … *two* aids available: casual self-observation and experimental method. The former assumes both a preparatory and a supplementary function; the latter, however, is the sole tool for analysis of elementary psychological processes' (Wundt 1908, p. 163). 'Ethnopsychology,' in contrast, had '… the task of complementing individual psychology with information about the general mental products of social life' (Wundt 1908, p. 163). 'The principle domains of ethnopsychology … are: *language*, *art*, *myth* and *customs*. They signify for the consciousness of cultures what imagination, feeling, and will signify for individual consciousness' (Wundt 1908, p. 232). Clarifying the main aspects of ethnopsychology he continues, 'So the psychological value of language consists primarily in expressing laws of *thought*, therefore various forms of language correspond to specific steps in the development of this basic mental function. In the same way, *myth* is intimately connected with the activities of fantasy in *art*' (Wundt 1908, p. 232). Finally, he describes the psychological value of (social) customs, writing, 'As myth does for art and religion, so do *customs* possess a central significance in their relation to two other important areas of life' (Wundt 1908, p. 233). 'One of these areas is *society* … It is society that first makes possible the formation of customs, the prevailing embodiments of norms of will within a society … The other area deriving from customs is *law*' (Wundt 1908, p. 234). He adds finally: 'Language, art, myth and customs are thus not the sole objects of ethnopsychology, but they are the ones that from the beginning express social life' (Wundt 1908, p. 234). On methodology in ethnopsychology, he wrote, 'The fundamental method of ethnopsychology … is … *comparison* in the ways it is generally used in the humanities' (Wundt 1908, p. 240). And he adds, '… ethnopsychology … is conducted … by assembling for comparison a collection of facts about psychological viewpoints, and through this close and direct connection to psychological analysis, comparison itself leads to the establishment of psychological laws … In this manner, the correct use of the comparative method here gains a significance similar to that of experimental procedure in individual psychology' (Wundt 1908, p. 240).

With these ideas, Wundt created a conceptual framework that would have great consequence in the history of psychology: a scientific psychology of elementary psychological phenomena and a humanistic psychology of higher psychological phenomena. Scientific psychology is an 'individual psychology' since it gets its data from the experimental investigation of individuals. Humanistic 'ethnopsychology' is a 'general psychology,' since it examines general psychological objects. As such, it is also a 'general developmental psychology' since those objects also develop throughout history. Methodologically, ethnopsychology gets its data from the ethnological analysis of historically developed products of mind (i.e., language, art forms, myths, customs). From these, conclusions can be drawn about the psychological processes involved in their emergence. Such processes include thinking, feeling, or willing. The view that a methodical humanistic approach was the only way to study higher psychological processes was momentous in the history of psychology. According to this concept, there could be no experimental psychology of thinking, emotion, or volition. From today's standpoint it should be noted that Wundt did recognize that psychological methodology could not be solely experimental. It must also be a quasi-experimental and historical methodology. On the other hand, his view that higher psychological processes were inaccessible to experimentation proved to be false. Wundt's methodological quarrel at the beginning of the twentieth century with members of the Würzburg school around Oswald Külpe (1862–1915), which concerned the role of experimentation in the psychology of thinking, had already shown that his critique of their method as 'interrogation experiments' was untenable.

3. *Wundt and the Genesis of the Discipline of Modern Psychology*

Wundt made fundamental contributions to the genesis of the discipline of modern psychology in the nineteenth and twentieth centuries, in Germany and beyond. This genesis may be described generally in terms of six phases and their corresponding developmental features.

First, Wundt had a lasting impact on the development of the subject matter and methods of empirical psychology since the beginning of the nineteenth century with his 'individual psychology.' This is also true to a certain extent for his 'ethnopsychology', if historically developed products of mind are regarded as historical empirical experience.

Second, Wundt was influential in institutionalization, that is, founding institutes for experimental psychology and anchoring them in the university system, as well as other forms of institutionalization, such as the founding of scholarly journals and scholarly societies. Active in scientific policy making, Wundt made a definitive contribution to institutionalization by founding the world's first psychological laboratory, the Institute for Experimental Psychology at the University of Leipzig in 1879. He

also founded one of the first experimental journals in psychology, *Philosophical Studies* in 1881.

Third, Wundt can be said to have contributed to the temporary division into major schools during the period between about1890 and 1940, inasmuch as his institute became a model for psychological institutes in many countries. But he did not found an actual school in the strict sense of the Berlin school of Gestalt psychology, the Würzburg school of cognitive psychology, or the Vienna school of psychoanalysis.

The fourth phase is the development of applied psychology, that is, through the inclusion and development of fields of praxis outside the universities and higher education after about 1890. Wundt had an ambivalent, rather negative relationship to the development of applied psychology. In his opinion, psychology should attempt a transition to practical applications only after its foundation has been systematically constructed (Wundt 1913).

Fifth, with his twofold psychology and with numerous students from many countries, Wundt made a definitive contribution to the development of a pluralistic system of psychology and to the gradual integration of the major schools after about 1930.

The final phase is professionalization, that is, the establishment, social anchoring, and state certification of standards for academically educated psychologists active outside of the university system around the time of World War I. With his early work in developing a curriculum for the education and training of psychologists, Wundt contributed indirectly to professionalization. This curriculum, with further contributions from other psychologists, was adopted in 1941 in the form of the first state diploma examination regulations in Germany.

The genesis of the discipline described above was initially, in the nineteenth century, determined primarily by experimental psychology. This constituted the first phase of methodological development in which experimental methods, the technical apparatus and the mathematics of experimental physics and of experimental physiology were transferred to psychology. Wundt's 'individual psychology' is an expression of this development. It thus becomes understandable why many of the 'pioneers' of modern psychology, Wundt's contemporaries in nineteenth-century Germany, were trained as experimental physicists or experimental physiologists. They include Ernst Heinrich Weber (1795–1878), anatomist and physiologist, Gustav Theodor Fechner (1801–87), doctor and physicist, and Hermann von Helmholtz (1821–94), physiologist and physicist.

During the second half of the nineteenth century a second phase of methodological development began. More or less standardized nonexperimental methods were added step by step to the experimental procedures. These methods were initially analytical documentary procedures that Wundt called methods of 'comparison' within his 'ethnopsychology.' Later they also included the preexperimental empirical methods based on the paradigms of assessment, achievement, arrangement, and interpretation. These methods contributed significantly to the development of psychodiagnostic procedures. Both groups of nonexperimental procedures today constitute elements of 'historiographic' and 'quasi-experimental' methodology in psychology. Wundt's work and its influence may be categorized within these two phases of nineteenth-century methodological development: his experimental 'individual psychology' falls within the first phase and his nonexperimental 'ethnopsychology' within the second.

4. *Review and Outlook—What was Transitory, What Endures*

Four achievements of Wundt's life's work must be especially emphasized. First, he contributed, in the politics of science, to the genesis of the discipline of modern psychology, that is, to its independence as an empirical psychology with its own subject matter and methods. Second, he constructed a system of psychology as an experimental 'individual psychology' and as a humanistic 'ethnopsychology.' Third, he founded the world's first Institute for Experimental Psychology, which gave impetus to the institutionalization of psychology and fourth, he founded one of the first scientific journals for psychology, *Philosophical Studies*.

Critically it should be noted that Wundt's concept of the experimental inaccessibility of higher psychological processes played a limited role for a time in the discipline's development. However, even some of his contemporaries did not share this idea. They included Gustav Theodor Fechner (1801–87), Carl Stumpf (1848–1936), Hermann Ebbinghaus (1850–1909), Georg Elias Müller (1850–1934) and Oswald Külpe (1862–1915). Nor did many experimental psychologists of the next generation, such as Narziß Ach (1871–1946), Karl Bühler (1879–1963), Max Wertheimer (1880–1943) or Otto Selz (1881–1944), share Wundt's view.

Wundt's writings and activities attained worldwide influence, especially because of his many students. It was them, most of all, who spread the tradition of experimental 'individual psychology.' Granville Stanley Hall (1844–1924), who established the second psychological laboratory in the USA at Johns Hopkins University in Baltimore in 1883 and in 1887 founded the *American Journal of Psychology*, was a prime example. Another was James McKeen Cattell (1860–1944), Wundt's first assistant in Leipzig, who occupied the first chair in psychology in the USA at the University of Pennsylvania in 1887. He became known primarily as the cofounder of the field of psychological diagnostic testing. Edward Bradford Titchener (1867–1927) who left Oxford, England in

1890, spent two years in Leipzig and went on to promulgate Wundt's teachings as structuralism at Cornell University in the USA. Another student was Dimitrij Nikolaevich Uznadze (1887–1950), founder of the Georgian school of psychology and known for his theory of set. Notable among Wundt's German students were Emil Kraepelin (1856–1926), one of the founders of modern psychiatry, psychopharmacology, and experimental psychodiagnostics; Ernst Meumann (1862–1915), founder of experimental pedagogy; Theodor Lipps (1851–1914), an exponent of experimental esthetics; Hugo Münsterberg (1863–1916), founder of industrial psychology and cofounder of applied psychology ('psychotechnics'); and Karl Marbe (1869–1953), one of the founders of forensic psychology.

See also: Developmental Psychology; Disciplines, History of, in the Social Sciences; Empiricism, History of; Ethnology; Experimentation in Psychology, History of; Hall, Granville Stanley (1844–1924); Kraepelin, Emil (1856–1926); Psychology: Historical and Cultural Perspectives; Psychology: Overview

Bibliography

Bringmann W G, Tweney R D (eds.) 1980 *Wundt Studies*. Hogrefe, Toronto, Canada

Bringmann W G, Lück H E, Miller R, Early C E (eds.) 1997 *A Pictorial History of Psychology*. Quintessence, Chicago, IL

Eckardt G, Sprung L (eds.) 1983 *Advances in Historiography of Psychology*. VEB Deutscher Verlag der Wissenschaften, Berlin, Germany

König E 1902 *W. Wundt als Psycholog und als Philosoph* [W. Wundt as Psychologist and Philosopher], 2nd edn. Frommanns, Stuttgart, Germany

Popplestone J A, White McPherson M 1999 *An Illustrated History of American Psychology*, 2nd edn. The University of Akron Press, Akron, OH

Sprung L, Sprung H 1981 Wilhelm Maximilian Wundt—ancestor or model? *Zeitschrift für Psychologie* **189**: 237–46

Sprung L, Sprung H 1999 Rückblicke auf ein schwieriges Jahrhundert—Zur Geschichte der Psychologie im 20. Jahrhundert in Deutschland. In: Hacker W, Rinck M (eds.) *Zukunft gestalten. Bericht über den* 41. *Kongreß der Deutschen Gesellschaft für Psychologie in Dresden* 1998. Pabst, Lengerich, Germany, pp. 123–43

Viney W, Brett King D 1998 *A History of Psychology. Ideas and Context*, 2nd edn. Allyn and Bacon, Boston, MA

Wundt W M 1863/1990 *Vorlesungen über die Menschen- und Thierseele* [Lectures on the Mind of Humans and Animals] 2 Vols. VEB Deutscher Verlag der Wissenschaften, Berlin, Germany

Wundt W M 1900–09 *Völkerpsychologie* [Ethnopsychology] 10 Vols. Engelmann, Leipzig, Germany

Wundt W M 1906–8 *Logik. Eine Untersuchung der Prinzipien der Erkenntnis und der Methoden wissenschaftlicher Forschung* [Logic: An Investigation of the Principles of Knowledge and Methods of Scientific Research] 3 Vols. Enke, Stuttgart, Germany

Wundt W M 1908–11 *Grundzüge der Physiologischen Psychologie* [Principles of Physiological Psychology]. 6th edn. Engelmann, Leipzig, Germany

Wundt W M 1910–21 *Kleine Schriften* [Small Papers] 3 Vols. Engelmann, Leipzig, Germany, pp. 515–43

Wundt W M 1913 Die Psychologie im Kampf ums Dasein [Psychology in the Struggle of Existence]. In: Wundt W M (ed.) 1921 *Kleine Schriften*, Vol. 3. Kröner, Stuttgart, Germany

Wundt W M 1920 *Erlebtes und Erkanntes* [Lived Experience and Knowledge]. Kröner, Stuttgart, Germany

L. Sprung

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Western European Studies: Economics

1. Background

In many respects, economic Western Europe looks very different today from what it was in the Cold War years of the early 1980s. In these two eventful decades, the Western European economy underwent radical transformations, with deep structural changes and complex institutional reforms affecting the working of both markets and policy authorities.

Since the second half of the 1980s, after a long period of sclerosis, an ambitious process of economic integration was launched and successfully implemented. It first took the form of market enlargement in real and financial assets aiming at the perfect mobility of goods and productive factors. The idea of a single European market was significantly reinforced by the decision to introduce a common currency and to design a new set of European institutions and procedures. Consequently, economic convergence within the area increased while, following German reunification, integration with Eastern European countries also proceeded at a very rapid pace. Projects for a further widening and deepening of European integration, ranging from a more complete economic and political union to the Eastward enlargement of membership, are now in an advanced stage of elaboration or just about to be implemented.

The decisive steps toward greater integration were taken under the influence (which occasionally became a threat) of world trade globalization, the new technological revolution, and the generalized liberalization of capital movements. In Europe, accepting these external challenges brought about a decisive retreat of the nation-state from the effective control of a series of domestic policy instruments and the public ownership of industrial concerns. Traditional theories on market failure were replaced by a number of alternative approaches based on incentives and mechanism design. State-owned enterprises have been affected by a massive wave of privatizations, while new procedures to deal with anticompetitive behaviors were elaborated by the European Union. The European Commission has now acquired increasing powers in this respect and some commentators have argued that latecomer Western Europe has one of the most developed antitrust policies in the world.

In this process, regulation was critically scrutinized as a source of market inefficiency and systemic degeneration. Politicians and academic scholars often wondered whether Government intervention had become a powerful obstacle to employment and economic growth. The pressure to have common labor legislation and economic regulations going from a harmonization of tax rules to the redefinition of social benefits opened new fields of investigation on the traditional theme of the optimal relationship between markets and authorities.

Major transformation processes such as these often impose heavy adjustment costs on societies or specific groups. Western Europe was no exception to this rule of history. It would probably be an exaggeration to talk of a long recession of the European economy in the 1990s, but it remains true that, in comparative terms, the century closed with a long series of weak economic performances that have not been reversed by the introduction of the new currency. The unwelcome consequence of this rapid transition toward a more integrated and competitive economic system was the upward convergence of unemployment rates and the increase of social inequalities. Double-digit figures were a permanent feature of the European labor market in the 1990s, and remained quite unaffected from cyclical fluctuations in output and investment.

Quite understandably, integration, unemployment, and regulation were among the economic issues that attracted a great deal of intellectual energies. Although the richness of Western European studies in economics is obviously far greater, considerable work was done on these general topics, and this explains the decision to restrict this discussion to them.

2. Integration

Generally speaking, integration in Western Europe has been interpreted as the outcome of sovereign actions of politicians and bureaucrats, rather than a market-induced process (Streit and Mussler 1994). Ever since the Treaty of Rome (1957), a long series of 'interventions from above' produced substantial results in the fields of trade policy, agriculture, the monetary system, competition, industrial policy, and social cohesion. The Maastricht Treaty (1992) strictly followed this functional approach, extending the Community powers over new sectors and policies, and establishing a rigid set of criteria for the creation of a monetary union. Policies for redistribution and compensation within the area were also significantly reinforced after Maastricht. Many more recent features (like Sweden's large majority rejection of the monetary union in September 2003 that strengthened the anti-Euro sentiments of the Northern fringe of Scandinavian countries and Britain) corroborates the view that European integration is an elitist project with rather limited popular support.

Before the approval of the Delors report (1986), agriculture and exchange rates dominated the agenda. The Community Agriculture Policy turned out to be a 'subsidized folly' (Petersmann 1993), burdening the European taxpayer and consumer with extremely high costs. It still absorbs most of the Community (modest) financial resources and instead of winning

approval it provoked loud protests even among its beneficiaries. It turned out to be a permanent source of conflict, both within and outside the Community, and was particularly detrimental to new trade creation and growth opportunities with Less Developed and Eastern European Countries in transition.

The European Monetary System (EMS) worked as an exchange rate mechanism, which was built on fixed but frequently adjustable parities, enhanced central banks cooperation, and an anti-inflationary bias. Flexibility, new financial facilities, and a strong dollar were the main factors for its first successful stage (1979–87), while rigidities, frictions, a weak dollar, and the economic consequences of German reunification marked its tumultuous end (1992). In terms of a monetary union, the EMS was unsuccessful and steps in that direction (the creation of a European monetary fund and the issue of a 'parallel' currency) did not evolve.

However, the actual working of the EMS stimulated fresh analysis on the interaction between monetary arrangements on the one hand and the functioning of real markets and the efficacy of policy decisions on the other. European economists refined such concepts as credibility, reputation, time consistency, or the games between governments and central banks, and applied them as analytical tools to explain the new road to integration. They also showed that asymmetries and rigidities were the result of choice rather than constriction. New theoretical models were framed to speculate on the advantages of 'tying one's hands' in the conduct of domestic monetary policy: the EMS, so it was argued, allowed member states to import credibility and acquire stability from the core country (i.e., Germany) and its central monetary authority. Therefore, by the end of the 1980s, the loss of monetary independence was almost complete, thus easing the way toward greater monetary integration (Giavazzi and Pagano 1988).

The EMS collapsed due to its incapacity to target simultaneously fixed exchange rates and use interest rate policy in pursuit of other domestic economic objectives (i.e., employment growth). While this is possible when capital movements are prohibited, in a global financial market the 'trinity' becomes inconsistent. Thus, a monetary union or, conversely, flexible exchange rates, remained the alternate options to overcome the 'open economy trilemma,' given the freedom of capital flows. After some discussion and popular reactions, Western European countries decided for the former (Giavazzi and Giovannini 1989).

The European Monetary Union (EMU) remains the world's most advanced experiment in moving beyond national economic sovereignty. The transition to monetary union was accompanied by widespread skepticism and criticism. Particularly, American observers thought that the likely effects of monetary union on macroeconomic fundamentals would be adverse even though they might be somewhat compensated by closer political ties (Feldstein 1997).

Debates on EMU focused on what prerequisites real markets of goods and productive factors ought to possess for monetary integration to generate net welfare gains. The Maastricht schedule of events and convergence criteria, relying on arbitrary quantitative ratios, was attacked for having nothing to do with the economics of optimal currency areas. Requisites for the 'sustainability' of EMU were strongly influenced by the prevailing German view and were defined in terms of nominal convergence of inflation and interest rates, monetary independence, and fiscal policy restraints.

Economists emphasized that the real properties for monetary integration simply did not exist in Western Europe: local markets were not efficient enough to guarantee low adjustment costs and offset the loss of exchange rate changes. Mobility, flexibility, and social integration were other indispensable factors that Western Europe did not possess. Common monetary policies with no exchange rate adjustments rendered the absorption of asymmetrical shocks more difficult. The inclusion of restrictions on fiscal policy in a treaty that set the timetable for monetary union was a further source of considerable debate. Many felt that the road to monetary unification needed to be supplemented by some forms of fiscal coordination in order to absorb country-specific shocks and avoid migration of taxpayers. However, while nation-states were all quite reluctant to surrender or harmonize their fiscal stance, Germany imposed stringent rules in order to discipline high deficit countries and avoid negative spillovers that would threaten the dogma of price stability and interest rate convergence. The economic rationale of this 'stability pact,' however, was thought to be very weak, particularly if rigidly applied over the business cycle.

On the whole, the (almost) unanimous verdict was that Europe did not constitute a 'natural' monetary union, and that imposing a new single currency would be detrimental mainly for cyclical unemployment. On the positive side of the balance sheet, the European countries' significant degree of openness to mutual trade remained, mainly intraindustry. This feature rendered adjustment through exchange rates less effective and less compelling. Moreover, a single currency was thought to provide further expansion to intra-European trade and investment, because of lower transaction costs, less discrimination, and no currency risks. The ensuing regionalization of the European economy was expected to reduce the likelihood of country-specific asymmetric shocks.

Therefore, within the economics profession, EMU is still receiving favorable support, despite the difficult start of its new currency and widespread

manifestations of anti-Euro public opinions. On the one hand, it created a safety net against such outcomes as trade wars, competitive devaluations, and discriminatory practices that are always likely to be implemented as it was experimented in the aftermath of the 1992 EMS crisis. On the other, dynamic efficiency gains might be obtained by integration in terms of more perfect competition, economies of scale, productivity growth, and a more efficient allocation of loanable funds.

Building up on these advantages, new studies show that currency integration is indeed a very dynamic process that cannot be (exclusively) predetermined by some set of exogeneous criteria (whether the Maastricht ones or other more theoretically sound ones). In Europe, the lack of mobility and flexibility is a regional, not national, phenomenon, and there is no reason why monetary integration should make things worse. The classic Mundell criteria have thus been substantially revised. Research efforts have recently been directed to the study of the intertemporal balancing of the benefits against the costs (Padoan 2000). Monetary integration must then be analyzed in an endogeneous growth model. In the process toward more complete integration, public confidence in the system may grow, some new benefits may emerge, and existing benefits may rise; alternatively, expected costs and the occurrence of asymmetric shocks may diminish as integration proceeds.

Recent manifestations of anti-Euro sentiments have been sustained by bad economic performance in the Euro area. At the current stage of European monetary integration, Europe oddly resembles more a multi (rather than an optimal) currency area, while economists have raised serious doubts on whether European policy rules have been efficiently designed and properly managed.

Finally, empirical research has shown that Eastern and Western Europe have integrated faster than it was commonly expected. Eastern European firms are now participating in Western Europe-centered production hubs, and have actively organized cross-bordered corporate networks. Intraindustry trade and input sourcing boomed together with foreign direct investments, just showing the growing complexity and maturity of economic integration between the European Union and the transition economies.

3. *Unemployment*

Western European unemployment has increased dramatically since the early 1970s, moving up from 2% to more than 10%. Compared to other countries, European unemployment is suffering from higher duration rates (relating to those who are unemployed for one year or more), higher rates of open unemployment (which considers discouraged workers and involuntary part-time jobs), and low participation rates among particular classes of individuals, such as prime-age men, married women, and the elderly. Economists often talk of the existence of a 'dual' labor market in Western Europe, one formed by very protected (male) workers (the so-called 'insiders'), the other by an army of unemployed for which market mechanisms and incentives no longer seem to be functioning (the 'outsiders').

Debates on rising European unemployment have focused on three main issues: (i) the effects of globalization with growing imports from low-wage countries; (ii) the effects of the German way to monetary union; and (iii) the effects of a rigidly constrained labor market structure. According to the former, increased international trade from LDCs may have something to do with European unemployment. Although analysts often disagree of the real magnitude of these effects, demand for unskilled labor has been dramatically reduced in developed countries as a consequence of globalization, and changes in the structure of foreign trade provide a plausible explanation (Davis 1998). As to the second argument, meeting the Maastricht criteria for EMU membership has led in some countries to substantial short-term increases in unemployment. This reflected the contraction of aggregate demand due to tight monetary and fiscal policies that all countries implemented in order to meet the deadline. However, the greatest attention has been devoted by far to the study of the labor market, which has been mainly developed in comparative terms. Particularly, the North America versus Europe dichotomy has been featured in a variety of works (Siebert 1997, Nickell 1997, Layard et al. 1991).

Conventional wisdom states that European unemployment is now comparatively higher because of the more powerful interactions between local factor market institutions and global goods markets. In other words, greater openness of the European economy to trade and external competition in the goods market was unaccompanied by greater openness and competition in the factor markets, where the downward wage rigidity is the main source of unemployment. Local institutional features (such as the minimum wage, workers immobility, or high firing costs) introduced important asymmetries that affect the level of wages and have a negative impact on labor demand and supply.

The contrast between Western Europe and North America is more complex than this stylized picture. Economists, for instance, have begun to question the description of America as a flexible wage economy and of Europe as a rigidly constrained and overly institutionalized economy. Furthermore, it is quite inconvenient to talk of Western European unemployment since the structure of labor markets is hardly homogeneous, while their performance is even more markedly different (Siebert 1997). European rates of job turnover are quite satisfactory while overall wage

flexibility is not dramatically lower compared to North American standards (Nickell 1997). Consequently, economic analysts have been trying to disaggregate the European labor market in order to understand as to what are its main structural weaknesses that generate high unemployment rates. One of the more challenging questions is whether there are any specific institutional features that are more responsible than others and what practical remedies can be introduced.

The review of the empirical evidence shows that Western Europe has strong legislation both on employment protection (relating to the legal framework governing hiring and firing) and other aspects of the labor market (from working time to fixed-term contracts and minimum wages). Unemployment benefits are also fairly generous even though, beyond the short term, they tend to adjust toward the lower North American coverage. While extremely variable across countries, nonwage labor costs (which are mainly responsible for the so-called tax wedge) appear extremely burdensome. Hence labor productivity is financing social insurance benefits to a greater extent than in other labor markets. Finally, in most European countries, trade unions play an important role in wage determination and even when union density is quite low the coverage rate is always substantial. One possible generalization from these institutional features is that, in Western Europe, the wage elasticity of labor needs more time to determine the necessary quantitative adjustments, while the long-term response of wages to unemployment is satisfactory and even higher than in other contexts (Siebert 1997).

Among the remedies, together with greater wage flexibility, the high fiscal wedge is considered to be the most important single explanation to European unemployment, regardless of how the tax burden is redistributed. Extension of social insurance is the main reason for the increase in the overall marginal tax wedge, which has particularly depressed the demand for labor. Lowering payroll and income taxes has thus become a very popular recommendation for reducing unemployment. Moreover, there seems to be an increasing awareness that job protection rules must be softened in order to increase firms' incentives to create new labor opportunities. Serious implications are also associated with unemployment benefits, especially if they have no time limit and are combined with a decreased willingness to search for a new occupation. The so-called 'reservation wage' (which guarantees minimum social welfare and unemployment benefits) needs to become more negotiable, thus reducing distortions in the market mechanism that are familiarly indicated under the heading 'unemployment traps.' As far as trade unions are concerned, centralization of wage bargaining is thought to increase efficiency if the process is coordinated with strong employer associations and negotiated within the more general macroeconomic compatibilities imposed by policy authorities. However, centralization leads, almost by definition, to less wage differentiation: in Western Europe, equity reasons have again prevailed to market clearing equilibrium, mainly at the expense of low-skilled workers. Finally, increasing investments in education are generally recommended as a remedy against structural unemployment.

4. *Regulation*

The great claim for perfect competition is that it leads to efficient results in the allocation of resources. From the production side of the economy, a perfectly competitive market solves two problems. First, production is allocated in such a way that desired output is produced at a minimum cost (i.e., efficient cost structure). Second, given that prices are equal to marginal costs, desired output is supplied by a sufficient number of (viable) firms (i.e., efficient price structure). Thus, production efficiency is a desirable objective of public policy.

Unfortunately, the competitive model may fail to be satisfied in the sense that markets are neither always complete (some may even be missing) nor are all markets perfectly competitive. The theory of market failures and the rationale for government intervention have been pioneered by such European economists as Lindahl (1919), Pigou (1920), and Wicksell (1896). Then, starting in the 1930s, in Western Europe the question of how deeply the government should be involved in the national economy has strongly been debated with economists divided between advocates (Robinson 1933) and critics (Hayek 1944) of government intervention. After World War II, market failures together with the difficulties of price controls, subsidies, taxation, and nonprice regulation were the major considerations that, with slight differences across countries, turned out the answer in favor of regulation and/or granting governments ownership and operation of selected activities. Some have argued that the role of government should have been that of acting as a 'welfare state' and thus modifying the actions of the market that would not produce vital (or public) goods and services at an acceptable cost (Briggs 1961). In this vein, most Western European countries adopted the so-called *public interest* view of regulation in which positive theories are based on normative considerations: to improve (maximize) social welfare the government needs to respond to market imperfections, and the extent of its action, subjected to various constraints, varies accordingly with the extent of those imperfections. Moreover, abstracting both from positive and normative rationales, state ownership of certain 'strategic' industries like steel, defense, airlines, automobiles, and banks has been

perceived as necessary to promote economic growth. However, this interpretation is perhaps too idealistic. As we shall see later, it is often inappropriate to assume that a benevolent government would make decisions solely in the interests of citizens.

From the 1980s, the very idea of a single European market has shifted policy from using regulation and state-owned enterprises (SOEs) to a greater reliance on the market mechanism and on incentive systems to improve consumer welfare as well as pursue industrial and employment goals. The promotion of competition has been one of the pillars of the European Union since the treaty of Rome, and governing bodies of the Union have acquired increasing powers to deal with noncompetitive behavior, both by private firms and public agencies, that distort competition and inhibit integration of markets across Europe. Empirical analysis raised concerns about the cost, price, and quality performance of regulated firms, and the policy arena became increasingly aware of the possibility of regulatory failures. Theory advancements also showed how, although market failures remain an argument for government intervention, the 'unnecessary' and/or the 'wrong' intervention may worsen the situation.

The existence of market failures, however, is not a sufficient reason to grant government intervention in markets. This insight was provided some 20 years before by Ronald Coase, a British economist working in the US. In the seminal paper 'The Problem of Social Cost,' he paved the way for the economic profession to realize that some market failures may be solved by redefining property rights and creating markets. Moreover, in choosing regulatory policies one should take into account the fact that informational limitations may raise the possibility of strategic behavior both by the regulated firm and of the regulator itself. Thus, rather than on command-and-control provisions, regulatory enforcement should be based on incentive mechanisms. This was made clear by a French economist, Dupuit (1844). In his pioneering contribution on marginal cost pricing, he was perhaps the first to take up the problem of how to design incentive-compatible regulatory mechanisms. These issues, which were largely absent in the traditional regulatory theory, have been brought up by developments in the literature on game theory, principal-agent, information economics, and the new institutional economics. The introduction of informational and institutional constraints showed how second-best optimization of social welfare is not straightforward while the set of feasible regulatory instruments is rather limited (an excellent account of most of the issues of the 'new regulatory economics' is provided by Laffont and Tirole 1993).

In the last two decades, European Union public policy reform has been mainly concerned with liberalizing prices and access to restricted markets as well as transferring SOEs to the private sector, and/or contracting out activities previously run directly by governments (i.e., privatization). These phenomena have been reinforced by developments in the product and capital markets, changes in the level and structure of demands, and a sharp decline in the cost of certain products and services. Although the pace and extent of liberalization have been quite different across countries, the policy objectives of the European Union can be summarized as follows:

- setting rules that make access to industries possible for a plurality of efficient providers,
- unbundling vertically integrated monopolies, thus creating new markets to replace transactions that previously took place within firms,
- restructuring tariff structures in ways that reflect underlying costs,
- ensuring that noneconomic public interest may be attained in a competitive environment,
- designing regulatory mechanisms that reflect the possibility of 'capture' by regulated firms or other interest groups,
- replacing regulation through public ownership with arm's length regulation,
- ensuring independence of the regulator from the executive branch of the government, and
- imposing constraints on regulator's discretion.

Looking at selected industries, such as services and utilities, makes it possible to highlight some of the features of the reform process.

Regulation of *network industries*, such as the local loop in telecommunications, electricity distribution, water pipelines, and rail track, has been based on two kinds of market failures: natural monopoly and externalities. In a natural monopoly a single supplier could serve the entire market at a unit cost lower than any industry configuration with two or more firms (Baumol et al. 1982). Hence, the case for natural monopoly is based on the fixed and costly infrastructures needed to deliver services to users. Externalities are effects of jointness of production/consumption not priced by the market. Here, externalities are mainly related to network effects: the demand for the product or service increases with the increase in the size of the network, since there are benefits in being connected to a larger network.

Restricted entry and extended price regulation have been sought to prevent: (a) that the single unregulated firm would not produce the desired output level, making impossible the fulfillment of social or universal service, (b) unstable prices, and (c) excessive investments. However, as Coase (1960) put it, 'all solutions have costs, and there is no reason to suppose that governmental regulation is called for simply because the problem is not handled well by the market or the firm.' Take, for example, the case of telecommunications. Here, the existence of both scale and scope economies has been used to justify vertical integration of the provider of the local service (a true natural monopolist) with potentially competitive

activities such as long distance and mobile services. In these multiproduct industries socially optimal second-best (Ramsey) prices, which permit a natural monopoly to break even, resulted in cross-subsidization. In other words, prices of competitive activities (i.e., long-distance calls) higher than those based on 'stand-alone' costs subsidized the costs of noncompetitive activities (i.e., local access). Hence, high prices in potentially competitive services prevented consumers from reaping the benefits of technical progress.

Increasingly aware of these pitfalls, the European Union moved toward regulatory reform that involved the privatization of state monopolies, the elimination of entry restrictions, and the adoption of incentive-based regulatory techniques. An example is the so-called 'Full competition' directive (96/19), which required the elimination of all remaining legal barriers to entry in telecommunications markets by January 1, 1998.

Where competition *within* the market is not feasible an alternative competition *for* the market has been introduced (Demsetz 1968). For example, auctions for the franchise right to operate the mobile telecommunication services have been used in the United Kingdom, Germany, Italy, and the Netherlands. Provided that (a) inputs are competitive available to all bidders, and (b) there is no collusion among bidders, auctions compel prospective entrants to reveal private information on their services, thus reducing the agency problem that characterizes the relationship between the government and the regulated firm. Regulated firms almost always have superior information concerning costs, demand conditions, products quality, and technological possibilities than the regulator. Indeed, this asymmetric information plagues the whole (multi)principal–(multi)agent relationship of the regulatory problem. Informational constraints limit both the control of the regulator over the firm and the set of feasible instruments to exert such control.

In traditional rate-of-return regulation, prices are set so that the regulated firm can cover its costs plus a reasonable rate of return for capital. Here, the regulator suffers heavily for informational limitations on costs and demands. Thus, where possible, rate-of-return regulation has been displaced by more fruitful incentive-based techniques. For example, to increase the available information, the regulator may make use of external information, such as the performance of other firms as a benchmark by which to compare the performance of the regulated firm (Shleifer 1985). *Yardstick competition* is increasingly used by European agencies in assessing efficiency and/or evaluating performance. It has strongly been recommended by the European Commission for the regulation of interconnections prices in telecommunications. Another type of incentive-price regulation—one that avoids continuous and detailed information on costs and demands—is *price–cap regulation*. In price–cap regulation, the regulator fixes caps on the prices that firms can charge for their products (or basket of products) for a given period of time. Then, firms choose to price at or below the caps—that are reviewed over time. Since firms have the incentive to minimize costs, price–cap regulation induces price structures to be more in line with underlying costs than rate-of-return regulation.

Traditionally, in Europe, regulatory objectives such as quality of service, price control, and universal service obligations were thought to be more easily implemented by state-owned firms than through regulatory schemes. However, several European governments—most notably Britain, France, Germany, Italy, and Spain—have launched large privatization programs in the 1990s, which have been widely used to raise revenue and to reduce public debt without raising taxes or cutting other government services. Nevertheless, in some industries, state control through share ownership remains widespread and the dominant role of incumbents (former monopolists) is still an obstacle to full competition.

Positive theory suggests that politicians are reluctant to privatizations because they have to give up their ability to operate SOEs in ways that provide economic benefits for their political constituencies (Shleifer and Vishny 1994). Regulators have the power to generate and redistribute rents across interest groups and/or may use this power to gain, or maintain, support from their political principals. At the same time, regulated firms have incentives to 'capture' regulators and influence regulatory outcomes (Stigler 1971). Indeed, the European Union has strongly criticized regulatory agencies of member states for delaying the implementation of its directives aimed at liberalizing markets.

Bibliography

Baumol W J, Panzar J C, Willig R D 1982 *Contestable Markets and the Theory of Industry Structure*. Harcourt Brace Jovanovich, New York

Briggs A 1961 The welfare state in historical perspective. *European Journal of Sociology* **2**: 221–59

Coase R 1960 The problem of social cost. *Journal of Law and Economics* **3**: 1–44

Davis D R 1998 Does European unemployment prop up American wages? National labor markets and global trade. *American Economic Review* **88**(3): 478–93

Demsetz H 1968 Why regulate utilities? *Journal of Law and Economics* **11**: 55–65

Dupuit J 1844 De la Mésure de l'Utilité des Travaux Publics. *Annales des Ponts et Chaussees* 8. Translated as On the Measurement of the Utility of Public Works. In: Arrow K, Scitovsky T (eds.) *AEA Readings in Welfare Economics*. Allen & Unwin, London

Feldstein M 1997 The political economy of the European economic and monetary union: Political sources of an

economic liability. *Journal of Economic Perspectives* **11**: 23–42
Giavazzi F, Giovannini A 1989 *Limiting Exchange Rate Flexibility: The European Monetary System*. MIT Press, Cambridge, MA
Giavazzi F, Pagano M 1988 The advantages of tying one's hands: EMS discipline and Central Bank credibility, *European Economic Review*
Hayek F A 1944 *The Road to Serfdom*. Chicago University Press, Chicago
Laffont J J, Tirole J 1993 *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA
Layard R, Nickell S, Jackman R 1991 *Unemployment: Macroeconomic Performance and the Labor Market*. Oxford University Press, Oxford
Lindahl E 1919 Die Gerechtigkeit der Besteuering, translated as Just Taxation—A Positive Solution. In: Musgrave R A, Peacock A T (eds.) 1958, *Classics in the Theory of Public Finance*. Macmillan, London
Nickell S 1997 Unemployment and labor market rigidities: Europe versus North America. *Journal of Economic Perspectives* **11**: 55–74
Padoan G C (ed.) 2000 *L'euro e i Mercati Reali*. Il Mulino, Bologna
Petersmann E-U 1993 Constitutional principles governing the EEC's commercial policy. In: Maresceau M (ed.) *The European Community's Commercial Policy After 1992: The Legal Dimension*. Kluwer, Boston
Pigou A C 1920 *The Economics of Welfare*. Macmillan, London
Robinson J 1933 *The Economics of Imperfect Competition*. Macmillan, London
Shleifer A 1985 A theory of yardstick competition. *RAND Journal of Economics* **16**: 319–27
Shleifer A, Vishny R W 1994 Politicians and firms. *Quarterly Journal of Economics* **109**: 995–1025
Siebert H 1997 Labor market rigidities: At the root of unemployment in Europe. *Journal of Economic Perspectives* **11**: 37–54
Stigler G 1971 The economic theory of regulation. *Bell Journal of Economics* **2**: 3–21
Streit M E, Mussler W 1994 The economic constitution of the European community: From Rome to Maastricht. *Constitutional Political Economy* **5**: 319–53
Wicksell K 1896 Ein Neues Prinzip der Gerechten Besteuerung, translated as A new principle of just taxation. In: Musgrave R A, Peacock A T (eds.) 1958 *Classics in the Theory of Public Finance*. Macmillan, London
Wyplosz C 1997 EMU: Why and how it might happen. *Journal of Economic Perspectives* **11**: 3–22

P. F. Asso and A. Cognata

Copyright © 2004 Elsevier Ltd.
All rights reserved.

 ISBN: 0-08-043076-07

X

Xenophobia

Although the term xenophobia is well established in the language of politics, the media, and the sciences, as a concept it has only a weak theoretical foundation. This is due not least of all to the ambiguity of the term. The word xenophobia comes from the Greek words *xénos*, meaning both 'the stranger' and 'the guest,' and *phóbos*, meaning 'fear.' Strictly speaking, then, xenophobia means 'fear of the stranger,' but it is usually taken to mean 'hatred of strangers.' This contradiction is particularly clear in German, where xenophobia and 'Fremdenfeindlichkeit' (i.e., hostility towards strangers) are used consistently as synonyms. The semantic discrepancy between fear and hatred of strangers can be illustrated in the following way: the stranger whom one fears can be welcomed as a guest, but not so the stranger whom one hates.

1. *The Origin of the Term*

The concept of the 'stranger' exists in premodern European societies as well as in practically every existing culture worldwide. Usually strangers are considered to be foreign, strange, unfamiliar, uncommon, unusual, and extraordinary. One cannot expect a relationship with a stranger to be good from the outset: only time can tell whether it will be a friendly or inimical one. As a result, attitudes towards strangers are characterized by ambivalence. This is particularly the case for strangers who live 'amongst us'—in our village, our city. While the concept of the stranger can be thought of as a universal category, this is not the case for xenophobia. Xenophobia is a concept of recent coinage: it was first used by Anatole France in his 1901 novel *Monsieur Bergeret à Paris*. The term anti-Semitism of the Dreyfus affair, which shook domestic politics in France at the turn of the twentieth century, provided a social and political background for xenophobia, as did the virulent form of nationalism that was emerging at the time (Villard 1984). In 1906—the very year of Dreyfus's rehabilitation—xenophobia was listed for the first time in a French dictionary: the *Nouveau Larousse illustré*. But only many years later could the word be found in English and other dictionaries. While the scientific world hardly took notice of the concept, 'xenophobia' was all the more present in the press. Attempts to account for the phenomenon of xenophobia included the explanation that it is the result of prejudice against strangers (*Athenaeum*, March 13, 1909) or that it is related to imperialist attitudes (*Nation*, May 11, 1912). The peoples of foreign countries—for example, the Afghans (*Mail*, May 24, 1922)—were accused of xenophobia, and so were nationalist movements in Egypt, China, Persia, and Turkey (*Nation*, December, 20, 1919). Finally, the phrase 'xenophobic outburst' was coined to describe the rising wave of anti-Semitism that was the precursor of German fascism.

While 'xenophobia' slowly became an established term in the French and English-speaking worlds, it was rarely used by speakers of German. The reason is that a conceptual pair already existed in the German language: 'Überfremdung' ('overstrangerization,' infiltration with foreign elements, foreign control) is a direct expression of xenophobic attitudes, and 'Fremdenfeindlichkeit' (hostility against strangers) clearly describes such attitudes. 'Überfremdung' had already emerged as a concept in the German-speaking world towards the end of the nineteenth century. The use of the term reveals anxieties about a loss of national identity due to a massive increase of immigrants.

2. *Theoretical Framework*

One reason why xenophobia was given virtually no theoretical foundation is probably because the phenomenon cannot be discussed without taking into account normative concepts that focus on attempts to mark 'one's own' off from 'the other.' Race, nation, ethnicity, and culture are such concepts. In social and political reality, xenophobia manifests itself in accordance with the division of the world into one's own race, nation, ethnic group, and culture, and other races, nations, ethnic groups, and cultures. This type of symbolic and normative worldview promotes self-centeredness, and constitutes precisely the kind of cognitive framework within which xenophobia is spawned, articulated, and disseminated. Since fear and hatred, the two emotional states contained in xenophobia, are qualities that rely on subjective experience, they require cognitive signposts and social values to allow them to focus on their object. Focus on an object is less relevant when one fears something, than when one hates. Hatred of what is strange or foreign, therefore, always calls for naming and objectification. Labeling is the first step towards constructing

the object against which xenophobic hate can be directed. Social valuing of the object then modulates the intensity of xenophobia; it triggers the change from fear of the other—i.e., latent xenophobia—to hate of the other—i.e., virulent xenophobia.

Two scientific fields in particular deal with xenophobia: social psychologists examine forms of ingroup/outgroup behavior in which xenophobia can appear in its latent form, while researchers in the social sciences and humanities explore and interpret the social construction of xenophobia as well as its effect in modern societies.

3. *The Concept of Xenophobia in Social Psychology and Sociobiology*

In social psychology, research on xenophobia primarily explores groups rather than societies. Social psychological approaches start by dealing with questions such as 'how is group membership established?' and 'how does the ingroup distinguish itself from the outgroup?' Research shows that the self has to align with the group in order to establish group membership. To allow such alliances to be formed, group-related categories and identities must be available and used. If they do not yet exist, they are generated through the process of alliance itself. Intergroup comparisons allow the emergence of the kind of psychological distinctiveness which can lead to upgrading of the ingroup and downgrading of the outgroup.

Intergroup behavior is usually valued as a response to real or imagined group interests. At the heart of this theory is the proposition that group members' intergroup attitudes and behavior will tend to reflect the objective interests of their group vis-à-vis other groups. Where these interests conflict, their group's cause is more likely to be furthered by a competitive orientation towards the rival, which often is extended easily to include prejudiced attitudes and even overtly hostile behavior. At the same time, the success of the ingroup in achieving the goal is likely to be furthered by very positive attitudes towards other ingroup members, thereby engendering high morale and cohesion. Where, on the other hand, the groups' interests coincide it is more functional for the group members to adopt a cooperative and friendly attitude towards the outgroup. If this is reciprocated, a positive joint outcome is more probable (Sherif 1967). The social psychologists associated with Tajfel (1981) even proved experimentally that the mere fact of perceiving the existence of group categories suffices to generate an ingroup response: neither past hostility nor an objective conflict is necessary. Mere categorization is sufficient to produce intergroup discrimination. Categorizations that generate intergroup discrimination are intimately related to education and the use of stereotypes (Macrae et al. 1996) and prejudices (Lippmann 1922).

According to social psychology, then, xenophobia is a specific ingroup response that uses categorization in the form of stereotypes and prejudices, articulates real or imagined group interests, and pursues the goal of enhancing ingroup cohesion through discrimination against the outgroup.

Both in social psychology (LeVine et al. 1971) and in other scientific disciplines, as in sociobiology (Reynolds et al. 1987), xenophobia is either directly equated with ethnocentrism or at least interpreted as the logical extension of an ethnocentric frame of mind. The concept of ethnocentrism was first used by William G. Sumner, one of the key representatives of American Social Darwinism. In his definition of ethnocentrism, Sumner (1911) took up the concepts of 'codes of amity' (referring to internal cooperation) and 'codes of enmity' (referring to self-defence) defined by Herbert Spencer in the *Principles of Ethics*. His definition is still accepted today in several scientific disciplines, for which ethnocentrism is the technical name used for the feelings of cohesion that generate internal fellowship and devotion to the ingroup, which in turn lead to a sense of superiority to any outgroup.

4. *The Concept of Xenophobia in the Social Sciences*

In the social sciences, the focus of research is on social values, norms, historical developments, and structural factors, rather than on biological or psychological constants which provoke xenophobic behavior. The social sciences ask how these factors support the existence of latent xenophobia, as well as allow xenophobic outbursts within specific social and political contexts. Unlike sociobiologists, social scientists do not consider xenophobia to be a universal phenomenon; instead, they describe it, first, as one among several possible forms of reaction generated by anomic situations in the societies of modern states. Second, social scientists agree that xenophobia grows out of the existence of essentialist symbolic and normative systems that legitimate processes of integration or exclusion. Xenophobic behavior, therefore, is based on already existing racist, ethnic, religious, cultural, or national prejudice. However, social scientists usually fail to make clear distinctions between such systems of categorization and xenophobia.

Attempts were made at an early stage to understand differences between manifestations of xenophobia in different nations. Lapierre (1928), for example, gave four possible answers to the question of why racist prejudice and xenophobia are stronger in England than in France. Apart from the answer—common at

the time—that biological differences may play a role, he formulated three hypotheses that are still in use today. First, differences between the two countries may result from different forms of colonial involvement. Second, they may be determined by differences between the processes of political socialization in metropolises. Third, possibly, no radical difference exists between the two countries with respect to this issue, because manifestations of xenophobia can be interpreted as short-term, periodically variable reactions to immigration and group competition. Post-World War II research on xenophobia and racism has continued to develop these three approaches. The social sciences, however, show a preference for the second hypothesis, which is why political socialization is given particular attention. Research has indeed shown that in states which are comparable with regard to immigration, both the intensity of xenophobia and racism and their forms and organization develop differently (Banton 1996). Thus, empirical research has revealed that in France, the social distance between native North Africans and the French is significantly smaller than the social distance that separates Germans from native Turks in Germany. An important indicator is the number of mixed marriages in these two countries (Todd 1994). This discrepancy is not thought to be caused by a supposed greater inclination towards xenophobia among the Germans than among the French. Differences between processes of political socialization rooted in the conditions of the nation–state are considered more meaningful. The French model of the nation is based more on a universalist–assimilationist approach, while the German one is inspired more by a differentialist–culturalist approach. In the first model, barriers against naturalization are low because the principle of *jus solis* is still partly valid. In the second model, where the concept of the nation is rooted in the principle of *jus sanguinis*, the barriers against acquiring citizenship are high. In the French model, particular cultural identities are submitted to discrimination, while the German model favors such identities. At the same time, cultural categories are formulated negatively and constitute the basis for latent xenophobia and racism. This is a phenomenon that is currently particularly visible not only in Germany, but also in East-European and Southeast-European countries. Positive expressions of cultural categories, however, lead to multiculturalism. From the point of view of social sciences, racism and multiculturalism, therefore, are two sides of the same coin. Both forms of discourse are particularly developed in societies where there is a large social gap between ethnic groups.

At the beginning of the 1990s, xenophobic outbursts led to an increase in acts of racist violence in several Western countries. This was the culmination of a sudden rise of xenophobia in the West that began in the 1960s, and that can be distinguished from the old form of racism which led to Nazism and Fascism, with respect to both ideological roots and causes. Some researchers believe that it is possible to speak of a 'new racism' (Barker 1981) that began to develop slowly in the postwar area, as racism was no longer based on biological but on cultural differences.

Two interrelated causes are thought to have provoked the resurgence of xenophobic and racist movements. First, new migration patterns developed during the postcolonial era as a consequence of the gradual internationalization of the labor market. In the 1950s and 1960s, under the influence of recruiting programs such as the American *bracero* program and the *Gastarbeiter* (guest workers) policies developed by Germany and Switzerland, the kind of spontaneous migration which links low-wage and high-wage countries soon developed. In the countries of immigration, social groups with little access to resources considered newcomers as competitors who contended with them for jobs and public services (housing, education, social welfare). This easily fostered a social and political climate that generated xenophobia and racism—i.e., defensive reactions against migrants—as well as nationalism—i.e., demands that the state provide better protection against foreigners for its own population. This is the background that gave strength to those right wing parties which certainly now argue from a 'new racist' point of view and are particularly attractive to social groups who feel threatened by migrants.

The second cause that is believed to reinforce xenophobic and racist trends is globalization. Growing internationalization and more flexible markets of all kinds (for products, services, and labor) have led to intensification of competition between and within states. In turn, such competition has put pressure upon states, inducing them to reduce their services. Neoliberalism is the ideological foundation for globalization, and requires not only lean production but also a lean state. The reduction of state services in the areas of social welfare, education, and healthcare also hits those segments of the population who already live on the margins of society. The social groups whose livelihoods are menaced by the dismantling of the welfare state are the main breeding ground for xenophobic and racist ideologies, since these groups are placed in direct competition with migrants for welfare services (Wrench and Solomos 1993). Of course, outbursts of racist violence in various countries illustrate that xenophobia need not be caused by strong immigrant flows of the cultural other. Social decline of specific groups and right-wing political organization are sufficient preconditions for the emergence of xenophobia.

See also: Authoritarian Personality: History of the Concept; Ethnic and Racial Social Movements; Ethnic Cleansing, History of; Ethnic Conflicts; Ethnic Conflicts and Ancient Hatreds: Cultural Concerns;

Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Sociology of; Prejudice in Society

Bibliography

Banton M B 1996 The cultural determinants of xenophobia. *Anthropology Today* **12**(2): 8–12

Barker M 1981 *The New Racism: Conservatives and the Ideology of the Tribe*. Junction Books, London

Lapierre R T 1928 Race prejudice: France and England. *Social Forces* **13**: 102–11

LeVine R A, Campbell D T 1971 *Ethnocentrism: Theories of Conflict, Ethnic Attitudes and Group Behavior*. Wiley, New York

Lippmann W 1922 *Public Opinion*. Harcourt and Brace, New York

Macrae C N, Stangor C, Hewstone M 1996 *Stereotypes and Stereotyping*. Guilford Press, New York

Reynolds V, Falger V, Vine I 1987 *The Sociobiology of Ethnocentrism. Evolutionary Dimensions of Xenophobia, Discrimination, Racism and Nationalism*. Croom Helm, London

Sherif M 1967 *Group Conflict and Co-operation: Their Social Psychology*. Routledge & Kegan Paul, London

Sumner W G 1911 *War and Other Essays*. Yale University Press, New Haven, CT

Tajfel H 1981 *Human Groups and Social Categories. Studies in Social Psychology*. Cambridge University Press, Cambridge, MA

Todd E 1994 *Le Destin des immigrés. Assimilation et ségrégation dans les démocraties occidentales*. Seuil, Paris

Villard P 1984 Naissance d'un mot grec en 1900. Anatole France et les xénophobes. *Mots* **8**: 191–95

Wrench J, Solomos J (eds.) 1993 *Racism and Migration in Western Europe*. Berg, Oxford, UK

H.-R. Wicker

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

16652

 ISBN: 0-08-043076-7

Y

Y-chromosomes and Evolution

For several decades now, genetic studies have been used as a complement to more traditional anthropological approaches in the study of human origins and diversification (see *Modern Human Origins*). Most genetic analyses have used autosomal and mitochondrial (mtDNA) markers and such data provide a global picture of human evolution (Cavalli-Sforza et al. 1994, Ruiz-Linares 1999). Two major findings are that African populations show a higher level of genetic diversity when compared with other continents and that the greatest differentiation between human populations is seen between African and non-African populations. Both these observations are in agreement with paleoanthropological data showing that humans originated in Africa. Interestingly, the overall level of genetic differentiation between human populations is small relative to the diversity observed within populations, suggesting that the differentiation of human populations occurred relatively recently, possibly around 100,000–200,000 years ago. Based on paleoanthropological evidence two general models of human origins have been proposed: a multiregional and a Garden of Eden model. Under the first scenario the diversification of human populations is very old, tracing back to the initial differentiation of regional *Homo erectus* populations. The Garden of Eden model proposes that the differentiation of human population is recent (of the order of 100,000 years) and dates to the migration out of Africa of a fully modern human species. Genetic data have been found to be mostly consistent with a recent out-of-Africa model.

Until recently, little was known about the extent and distribution of diversity in the Y-chromosome. This problem stemmed from a basic difficulty in identifying genetic variation on this chromosome. None of the traditional genetic markers are found on the Y and attempts at identifying molecular variation progressed slowly until the late 1990s. A major breakthrough came about in 1997 when a novel technique known as Denaturing High Pressure Liquid Chromatography (DHPLC) was developed and allowed the rapid screening of a large number of DNA fragments from the Y-chromosome, leading to an increase in the speed of identification of new molecular variants. Recent developments resulting from the Human Genome Project, such as the availability of a draft sequence of the human genome and the activities of the so-called SNP (single nucleotide polymorphism) consortium are leading to an even faster identification of molecular variants on the Y-chromosome. The recent availability of a large number of genetic markers on the Y-chromosome has led to an increasing interest in the use of this chromosome in human evolutionary studies, both to explore human origins and ancient diversification as well as to examine recent population history.

1. *Transmission of the Y-chromosome*

Part of what makes the Y-chromosome of special interest for human population genetic studies is its peculiar mode of transmission. The passage of the Y from one generation to the next is immediately apparent: it determines the male sex. In other words, it is only transmitted from fathers to their male offspring. The fact that the Y-chromosome is transmitted between generations only by one sex means that if one reconstructs a genealogy of ancestors, only a single line of inheritance needs to be followed. Another type of DNA that shows a uniparental mode of inheritance is mitochondrial DNA (mtDNA), which is transmitted only by females (see Mitochondria and evolution). Thus, the Y-chromosome and mtDNA allow respectively the reconstruction of the male and female lines of inheritance or lineages (Jobling and Tyler-Smith 1995).

2. *Variability of the Y-chromosome*

Although the Y-chromosome determines the male sex, very few genes have been identified on the Y and none of the classical genetic markers are encoded by this chromosome. There is considerable cytogenetic variability of the Y but such variation is difficult to interpret for evolutionary studies. Similarly, most of the variants identified by molecular methods before DHPLC consisted of complex molecular rearrangements that were difficult to characterize and interpret in a population setting. The increasing number of variants that are now being identified can be categorized into two basic types: biallelic and multiallelic. The biallelic markers consist of nucleotide substitution or small insertion/deletions while the multiallelic markers generally consist of regions of the DNA in which a variable number of Short Tandem Repeats (STR) of a

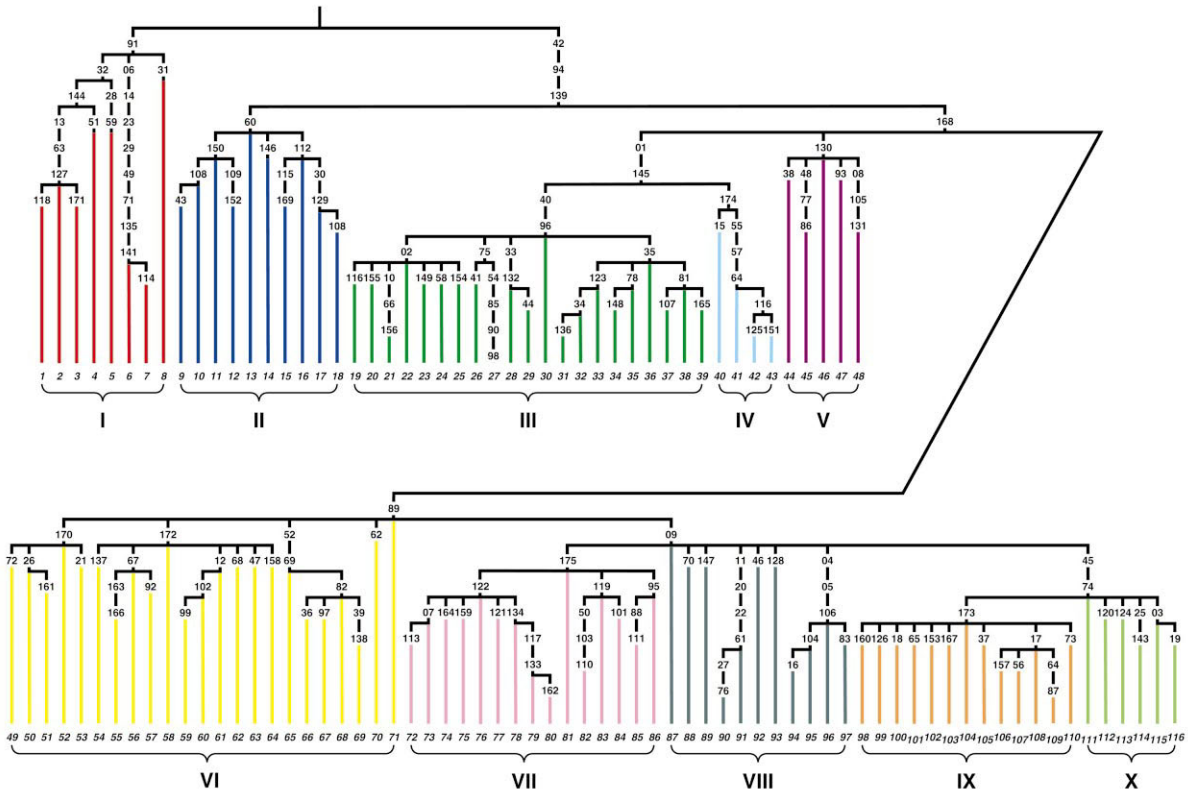

Figure 1
Phylogenetic tree relating Y-chromosome lineages from worldwide human populations (source: Underhill et al. 2000. Reproduced with permission)

DNA motif occur (for example a variable number of the tetranucleotide CATG). One basic difference between these two types of markers is that biallelic polymorphisms generally represent single evolutionary events while mutations at the STR loci occur at a relatively higher rate and the mutational process is frequently reversible. This means that identical STR alleles might have a different evolutionary history, as the same allele might be generated by independent mutational events, thus potentially erasing an evolutionary signal when probing too far back in time. However, the combination of these two types of markers actually increases the power of evolutionary analyses based on the Y-chromosome. Slowly evolving biallelic markers can be used to define the basic relationships between Y-chromosomes in an unambiguous form. The rapidly evolving STR markers can then be used to assess the relationships of Y-chromosomes within the basic lineages defined by the biallelic markers. Both kinds of markers can also be used to estimate the age of Y-chromosome lineages based on different evolutionary models, with the consistency of the two estimates thus providing greater confidence in such estimates (Pritchard et al. 1999, Thomson et al. 2000).

3. *The Origin and Diversification of Human Populations*

3.1 *Human Origins*

As biallelic markers on the Y-chromosome generally represent events that occurred only once in human evolution it is possible to examine the relatedness of Y-chromosomes worldwide and represent their ancestral relationships on a phylogenetic tree. The analysis of 167 genetic markers in 1,062 males from 21 populations around the world recently allowed the identification of 116 Y-chromosome types (Underhill et al. 2000). Based on the common ancestry of certain types, it is possible to group human Y-chromosomes into ten basic lineages (Fig. 1). Interestingly, the most ancient splits on the tree separate lineages I and II, which are

almost exclusively seen in African populations, from all other lineages which have a wider geographic distribution. This observation is in agreement with an origin of human populations in Africa followed by a migration out of that continent of a sample of African men. Descendants of these initial migrants would have founded small populations outside of Africa that would have remained small in size until the end of the last glaciation (around 16,000 years ago) when they would have expanded in numbers, spreading geographically to all inhabitable areas. Such an expansion is clearly evidenced in the Y-chromosome by the large proportion of mutations that are specific to a particular chromosome type. Estimates of the age of Y-chromosome lineages indicate that their initial differentiation is about 59,000 years old with a migration out of Africa around 44,000 years ago (Pritchard et al. 1999, Thomson et al. 2000). These findings are more in agreement with a model of recent African origins of modern humans; however the dates obtained with these data are considerably younger than those estimated from mtDNA and autosomal markers (Horai et al. 1995, Goldstein et al. 1995). This inconsistency might be due to natural selection (see *Natural Selection*) acting with variable intensity in different genetic systems or could relate to demographic differences between men and women. Overall, the Y-chromosome shows a reduced level of variation when compared with other genetic systems and such variation is highly structured by geography (Seielstad et al. 1998). This makes the Y particularly suitable for the study of the history of specific geographic regions.

3.2 Human Population Diversification

The analysis of Y-chromosome diversity at 22 biallelic markers in 1,007 males from 25 populations of Europe and the Middle East identified 20 different lineages (Semino et al. 2000). However, over 95 percent of the men examined have one of 10 lineages, several of which are closely related. Examining the distribution of these lineages in European populations it was observed that two closely related lineages showed East–West gradients in frequency, with one of them being particularly frequent in the Basque and the other in Eastern Europe (Poland, Hungary, and Ukraine). Phylogenetic and genetic dating analyses suggest that these two types of Y-chromosomes were brought to Europe by people of the Aurignacian culture, which spread to Europe about 35,000 to 40,000 years ago. The East-West gradient is possibly explained by the demographic contraction caused by the last glacial maximum (20,000 to 13,000 years ago) that would have led human populations to seek temporary refuge in more hospitable areas from which they would have subsequently expanded. A third Y-chromosome lineage is about 20,000 years old and is restricted to Europe (where it is seen at highest frequency amongst central Eastern Europeans) although it is closely related to Y-chromosomes frequent in the Middle East. The age and distribution of this lineage suggest that it was introduced by people of the Gravettian culture that appeared in Europe around 20,000 years ago. These people found refuge in central Europe during the last glacial maximum, after which they expanded into Western Europe. Finally, a group of younger and closely related haplotypes show a gradient in frequency from the Middle East into Europe and might have been spread by the people that introduced the practice of agriculture into Europe about 10,000 years ago. Y-chromosome data thus indicate that two Paleolithic and one Neolithic migrations are at the origin of most of the current European population as manifested by three main population clusters seen when lineage frequencies are compared between populations (Fig. 2). The three clusters observed (Western Europe, Eastern Europe, and the Middle East) correspond to the two major glacial refuges and to the area of origin of agriculture.

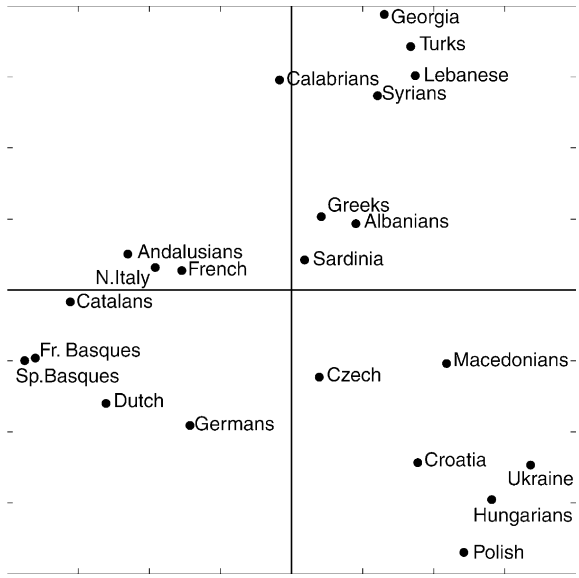

Figure 2
Principal component analysis of Y-chromosome lineage frequencies in European populations (source: Semino et al. 2000. Reproduced with permission)

4. Male and Female Migration in Human Evolution

A very interesting use of genetic data is in teasing apart male and female demographic history throughout human evolution, by comparing population diversity of Y-chromosome, autosomal, and mtDNA markers in native populations from around the world. In an

initial evaluation it was observed that the level of genetic variability across populations based on Y-chromosome markers was much greater than that observed with mtDNA markers (Seielstad et al. 1998). This means that the distribution of Y-chromosome variants is geographically more localized than that of mtDNA variants. A possible explanation for this global pattern could be that about 70 percent of human populations practice patrilocality, that is women tend to move into the areas inhabited by the males they marry and consequently move more, on an evolutionary scale, than men. A subsequent refined analysis restricted to South American populations failed to confirm the generality of a higher migration rate of women in native populations (Mesa et al. 2000). A detailed evaluation of the proposal that females have generally had a higher migration rate than males awaits the analysis of more genetic markers and populations. However, it is possible that important regional differences in male/female migration rates might be detected and that these differences relate to the frequency with which patrilocality is practiced in different parts of the world.

5. *Maternal and Paternal Origins of Recently Founded Human Populations*

The comparison of mtDNA and Y-chromosome markers is also informative when applied to populations established more recently and such data can be used to refine information obtained from traditional historical sources. Particularly, analyzing different genetic systems can help establish the relative proportion of males and females from particular geographic areas that founded a new population. In the province of Antioquia in Colombia it was seen that about 90 percent of Y-chromosomes were European while 90 percent of the mtDNA lineages were Native American (Carvajal-Carmona et al. 2000). This indicates that the population was founded mostly by European males and Native females. This finding is consistent with the geographic isolation of the province and with historical records indicating that in early colonial times, on average only about 20 percent of the Spanish migrants to the New World were women. However, this result is inconsistent with autosomal markers, which indicate that this population is predominantly of European ancestry. This inconsistency between bi and uniparental markers might relate to the collapse of the Native American population during the first century of Spanish colonization and to the fact that Spanish immigration was a long-term process. These demographic factors might have resulted in a progressive dilution of the founding Native American component as estimated with autosomal markers.

Curiously, the analysis of rapidly evolving STRs indicates that the population of Antioquia seems to have a relatively important Jewish ancestry as evidenced by the finding in this population of Y-chromosome STR lineages seen at high frequency only in Jewish populations. However, no record is available of a Jewish migration to Antioquia. Furthermore, almost simultaneously with the Spanish expansion to America, Jews were expelled from Spain or forced to convert to Christianity. The Y-chromosome findings suggest that amongst the Spanish founders of Antioquia a relative large proportion of converts was represented. No record of such contribution is available probably because the migration to the colonies of converts was considered illegal under colonial rule.

6. *Conclusion*

Developments resulting from the Human Genome Project have recently catapulted the use of Y-chromosome markers into the forefront of the study of human population origins and diversification. The large number of markers currently available together with novel highly efficient technologies enable analyses of an unprecedented resolution and scale. Evolutionary analyses are facilitated by the fact that slowly evolving markers allow the unambiguous assessment of the evolutionary relationship between Y-chromosomes. Rapidly evolving markers can refine analyses within specific lineages or populations. These studies illuminate not only questions related to the origin of our species and its early diversification but also allow the probing of more recent demographic events. The synthesis of genetic data with information obtained from sources including geology, paleoanthropology, archaeology, and historical demography is allowing a refined reconstruction of human evolution stretching from our origins as a species all the way to the exploration of quite recent historical events.

See also: Chronology, Stratigraphy, and Dating Methods in Archaeology; Ethnic Identity/Ethnicity and Archaeology; Evolution, History of; Evolutionary Approaches in Archaeology; Human Evolutionary Genetics; Migration, Theory of; Migrations, Colonizations, and Diasporas in Archaeology; Mitochondria and Evolution; Natural Selection; Paleodemography: Demographic Aspects

Bibliography

Carvajal-Carmona L G, Soto I D, Pineda N, Ortiz-Barrientos D, Duque C, Ospina-Duque J, McCarthy M, Montoya P, Alvarez V M, Bedoya G, Ruiz-Linares A 2000 Strong Amerind/White sex bias and a possible sephardic contribution among the founders of a population in Northwest Colombia. *American Journal of Human Genetics* **67**: 1287–95

Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

Goldstein D B, Ruiz Linares A, Cavalli-Sforza L L, Feldman M W 1995 Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 6723–7

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N 1995 Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 532–6

Jobling M A, Tyler-Smith C 1995 Fathers and sons: The Y-chromosome and human evolution. *Trends in Genetics* **11**: 449–56

Mesa N R, Mondragon M C, Soto I D, Parra M V, Duque C, Ortiz-Barrientos D, Garcia L F, Velez I D, Bravo M L, Munera J G, Bedoya G, Bortolini M C, Ruiz-Linares A 2000 Autosomal, mtDNA, and Y-Chromosome diversity in Amerinds: Pre- and post-Columbian patterns of gene flow in South America. *American Journal of Human Genetics* **67**: 1277–86

Pritchard J K, Seielstad M T, Perez-Lezaun A, Feldman M W 1999 Population growth of human Y-chromosomes: A study of Y-chromosome microsatellites. *Molecular Biology and Evolution* **16**: 1791–8

Ruiz-Linares A 1999 Microsatellites and the reconstruction of the history of human populations. In: Goldstein D B, Schlotterer C (eds.) *Microsatellites. Evolution and Applications*. Oxford University Press, Oxford, pp. 183–97

Semino O, Passarino G, Oefner P J, Lin A A, Arbuzova S, Beckman L E, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti A S, Cavalli-Sforza L L, Underhill P A 2000 The genetic legacy of Palaeolithic *homo sapiens sapiens* in extant Europeans: A Y-chromosome perspective. *Science* **290**: 1155–9

Seielstad M T, Minch E, Cavalli-Sforza L L 1998 Genetic Evidence for a higher female migration rate in humans. *Nature Genetics* **20**: 278–80

Thomson R, Pritchard J K, Shen P, Oefner P J, Feldman M W 2000 Recent common ancestry of human Y-chromosomes: Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences* **97**: 7360–5

Underhill P A, Shen P, Lin A A, Jin L, Passarino G, Yang W H, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd J R, Mehdi S Q, Seielstad M T, Wells R S, Piazza A, Davis R W, Feldman M W, Cavalli-Sforza L L, Oefner P J 2000 Y-chromosome sequence variation and the history of human populations. *Nature Genetics* **26**: 358–61

A. Ruiz-Linares

Youth Culture, Anthropology of

The study of youth has often hovered on the relative margins of more central anthropological preoccupations even as it has been inspired by work in such disciplines as sociology and cultural studies. During the late 1980s and over the course of the 1990s, research on this subject, both within anthropology and cognate disciplines, was catalyzed by a growing interest in popular culture and globalization, fields in which youths have often played pivotal roles. In the process of this reanimation, the study of youth within anthropology has moved away from a concern with adolescence as a life course stage to an interest in youths as cultural producers and consumers.

1. Defining Youth

The range of persons anthropologists include in the category of youths has often varied depending on the cultural context in which this position is being defined and on what question the researcher is asking in respect to this field. In a rare early example of a systematic ethnographic focus on youths, during the 1920s Margaret Mead conducted research centering on a group of girls, ranging in age between 10 and 20 years, who were residing in three villages on the Samoan island of Tau (Mead 1928/1968). In carrying out this research, Mead was concerned to answer the question of what 'coming of age' meant in Samoa and how this differed from the experience of growing up in the USA. She claimed that unlike the emphasis on storm and stress which then (and has since) dominated American characterizations of adolescence, growing up in Samoa was a simple, easy matter, a process which involved no special conflicts or dramatic transitions. Beyond the obvious signs of physical maturation, she suggested that there were few differences between Samoan girls still approaching puberty, those who were just passing through this transition, and those already several years beyond it. Given this emphasis, it is hardly surprising that Mead was not concerned with defining the boundaries of adolescence too precisely or distinguishing it sharply from either childhood or young adulthood.

In contrast, Alice Schlegel and Herbert Barry's (1991) much later survey of adolescence in ethnographic accounts of 186 nonindustrial societies was more expressly directed at delimiting this stage cross-culturally. Schlegel and Barry argued that while the social duration of adolescence varied cross-culturally and between genders, it was universally recognized as a social stage of the life cycle and tended to fall between the ages of 11 and 17 or so. When social adulthood was delayed far beyond puberty, a further stage intervened between adolescence and adulthood, a stage which Schlegel and Barry called 'youth.' While adolescence was universal, the stage of youth was only recognized in a minority of societies.

However, this distinction between adolescence and youth as two distinct social stages is not otherwise common in the anthropological literature on youth

culture. In part this reflects an ambiguity in the boundaries assigned to adolescence which is especially common in the Western industrialized settings which have often been the locales for youth culture research but in part this also reflects a set of particular analytical priorities. For example, a cross-cultural collection of articles on youth cultures edited by Vered Amit-Talai and Helena Wulff (1995) included groups ranging in age between 10 and 30 years with considerable variation in the scope of autonomy and social responsibilities represented among them. However, this collection, unlike Schlegel and Barry's or Margaret Mead's studies, was not directed at defining youth (or adolescence) as a stage of human development. Rather it was concerned with the ways in which young people creatively engage with, are affected by, and in turn help to shape the statuses, images, products, styles, and orientations identified, sometimes very broadly, with youth. In short, it was concerned with the experiences and attributions of youth rather than when and how youths become adults. As such it reflects a reorientation of the study of youth which has been as much inspired by influences outside anthropology as within it.

2. *The Cross-disciplinary Antecedents of Youth Culture Studies*

While Margaret Mead was very successful as a popularizer of anthropology, she was not equally successful in provoking fellow anthropologists to continue her focus on adolescents. For many years following Mead's work in Samoa, youths largely remained marginal figures in the anthropological record. Concerted interest in this field was much more evident in other social science disciplines than it was in anthropology. During the 1930s, a little over a decade after Mead's initial research in Samoa, the sociologist William Foote Whtye began participant observation fieldwork in the North End of Boston. Whyte's (1943/1981) classic ethnographic study of the organization of peer relations among second generation young Italian men in a 'slum area' exerted a seminal influence primarily on the development of US urban studies. Yet it was and remains one of the most thorough and sympathetic portraits of youth street life ever produced, effectively countering the notion of social disorganization which had dominated the US sociological literature on inner cities until that point. Not long after Whyte's study was published, the concept of youth culture appears to have been introduced by another American sociologist, Talcott Parsons (1942/1964) and was later taken up more systematically by James Coleman (1961) to denote the circumscription of American youth peer interaction. It was, however, the work of British cultural theorists during the 1970s and early 1980s which did most to develop youth culture as a central field of research on class, subculture, and consumption.

During this period, the Centre for Contemporary Cultural Studies (CCCS) at the University of Birmingham, sometimes referred to as the 'Birmingham School,' launched an ambitious collaborative study of British youth subcultures (Hall and Jefferson 1976). For the CCCS scholars, youth appeared to offer a special vantage point from which to consider the more general dislocation and fragmentation of the British working class as the structure of Britain's system of production, labor force, income distribution, and lifestyles was transformed over the course of the post-World War II period. The emergence of a variety of highly stylized British youth subcultures (punks, mods, skinheads, etc.) appeared to crystallize the growing importance of mass culture, consumption, and leisure practices as bases for increasingly fragmented working class identities. The gaze of the CCCS was self-consciously selective, focusing on young British nonconforming white working class males. The larger population of 'ordinary' working class youth who accommodated to rather than visibly resisted institutional controls, middle class youth, and girls, with some notable exceptions, were not accorded similar attention, gaps which were well noted by later critics. The members of the 'Birmingham School' were also criticized for reading ideological undertones of class resistance and consciousness into what sometimes appeared to be fairly trivial instances of adolescent rebellion. In spite of an emphasis on case studies, the CCCS published work was notably slim in empirical findings. Here again, there was a notable exception in Paul Willis' (1977/1981) ethnography, Learning to Labour, which followed a group of boys, the 'lads,' through their last year of high school and onto the industrial shop floor. Yet, in spite of these limitations, the Birmingham School's focus on popular culture, consumption practices, resistance and cultural agency, and their attention to youth as a nexus for broader social change, continue to resonate in research on youth culture both within anthropology as well as in other disciplines.

3. *Anthropology of Youth Culture: Key Features*

Current anthropological research on this subject is characterized by a determined shift away from socialization and human development approaches, which have often entailed less of a focus on youths than on the adults they will eventually become or the adults attempting to direct and socialize them. Instead, contemporary anthropologists are more likely to focus on young people in their own right. While much of this research pays heed to the marginality generally associated with youth, it does not assume that this social location produces a simple helplessness or passivity.

Instead, it often features an insistence on respecting the capacity of juveniles to be active agents creatively shaping their environment even as they contend with a variety of legal, political, economic, and social constraints imposed on them by virtue of their age and liminal status.

These emphases have emerged within the context of a broader reorientation of anthropology away from the notion of unitary and enduring cultural formations which was more common in Mead's day to a contemporary conception of culture that can more easily accommodate internal diversity, fragmentation, and transience. This has made it easier for anthropologists to be comfortable with a notion of youth as eliciting specific kinds of knowledge and cultural competence that are different from rather than incomplete versions of adult cultures. It has also rendered more persuasive the claim that youth cultures can have broader analytical and social significance even when they are highly ephemeral and situationally specific. During the 1980s and 1990s, critical reappraisals of the conventions of ethnographic accounts, in particular of the ways in which the textual authority exercised by anthropologists can silence or distort the perspectives of the people they are studying, provided a broader context for a reconsideration of the ways in which ethnographic representations had effaced the presence and experiences of children and youth. In a cogent illustration of the dialectical nature of intellectual exchange, an essay by George Marcus (1986) in a volume which was particularly influential in this broader anthropological reassessment, focused on Paul Willis' seminal CCCS ethnographic study of working class boys in school. In turn, volumes of this kind and the critical introspection they prompted influenced subsequent anthropological studies and representations of both children and youth.

Notwithstanding this exchange, the study of youth culture remained on the relative fringes of mainstream preoccupations within late twentieth century anthropology. There were more ethnographers conducting research in this field than twenty years previously but their work still only had limited recognition within the larger discipline of anthropology. This is in spite of the fact that youth studies often focused on issues with general salience in contemporary anthropology, particularly the impact of globalization, migration, and popular culture. Hilary Pilkington's (1994) study of Moscow youth focused on the ways in which the production and consumption of music, dance, and dress styles mixed and hence mutually transformed imported Western and local Russian influences. Mark Liechty's (1995) study of youth in Kathmandu explored the media appropriation, dissemination, and reinterpretation of the Western concept of teenager. Livio Sansone (1995) examined the changing and increasingly global nature of youth culture among young Creoles of Surinamese origin living in Amsterdam.

Notwithstanding the broad dimensions of the processes they consider, many of these ethnographies of youth are self-consciously and intentionally intimate in focus. Helena Wulff's (1988) South London study examined a wide range of issues including the symbolic dimensions of black British ethnicity, consumption practices, friendship, femininity, and institutions of family, school, and youth club, but it deliberately and carefully focused on the relationships and lives of 20 girls between 13 and 16 years of age. Wulff's focus was based on her concern to relate cultural process concretely to individual lives and social relationships, a goal more accessible through a concentration on small-scale cultural processes or 'microculture.' Similarly, nearly a decade later, in another study of black British youth in London, Claire Alexander argued that in focusing on a small number of male informants she was 'hoping to portray an alternative vision of black youth; not as a unified and homogenous, externally defined and structurally constrained entity, but a collection of individual lives, choices, and experiences' (Alexander (1996, p. 18). Their capacity to marshal a wealth of detail about relatively few individuals allows these small-scale studies to move beyond stereotypical notions of youthful resistance or undifferentiated peer culture to reveal the fluidity and creativity of cultural processes among young people.

4. *Gaps in the Anthropology of Youth Culture*

In spite of nearly two decades of critiques, the influence of the pioneering work of the CCCS continues to be expressed in a predominant emphasis on leisure activities and peer relations within many youth culture studies. Ironically, given the CCCS interest in class, the focus on leisure has often been at the expense of systematic analyses of the political economy of youth. Problems of unemployment and the gulf between globalized media images of Western consumption practices and the limited economic resources of many youths are alluded to in a number of studies. However, the nature of the economic roles performed by youths, while still students or nominally minors, as workers in service or informal sectors has not received much sustained attention. We are still much more likely to see youths dancing, singing, clubbing, and consuming fashion and media than working. Yet the globalization of production and the growth of service industries has often relied on the availability of cheap, part or full time youth labor. We are also not very likely to see much of the interaction between youths and adults, particularly with their parents. In the concern to move away from adult-based socialization models of adolescence, researchers often appear to have created ethnographic landscapes remarkably devoid of the adult figures who nonetheless figure so largely in the lives of the youths being studied. Not only are such

portraits ethnographically incomplete but, without including adults, it is difficult to see how these studies can adequately address issues of power and subordination, however sensitive they may be to the association between youth and social marginality. Finally, one has to wonder whether the marginality of their youthful subjects has been extended by proxy to this subfield of anthropology. If so, it would be a strange irony in a discipline that has prided itself on its particular attention to poor and marginalized peoples.

See also: Adolescence, Sociology of; Adolescent Vulnerability and Psychological Interventions; Childhood: Anthropological Aspects; Childhood Health; Youth Culture, Sociology of; Youth Gangs; Youth Movements; Youth Movements, History of

Bibliography

Alexander C E 1996 *The Art of Being Black: The Creation of Black British Youth Identities*. Clarendon, Oxford, UK

Amit-Talai V, Wulff H (eds.) 1995 *Youth Cultures: A Cross-Cultural Perspective*. Routledge, London

Coleman J S 1961 *The Adolescent Society*. Free Press of Glencoe, New York

Hall S, Jefferson T (eds.) 1976 *Resistance Through Rituals: Youth Subcultures in Post-War Britain*. Hutchinson, London

Liechty M 1995 Media, markets and modernization: youth identities and the experience of modernity in Kathmandu, Nepal. In: Amit-Talai V, Wulff H (eds.) *Youth Cultures: A Cross-Cultural Perspective*. Routledge, London, pp. 166–201

Marcus G E 1986 Contemporary problems of ethnography in the modern world system. In: Clifford J, Marcus G E (eds.) *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, CA, pp. 165–93

Mead M 1928/1968 *Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilization*. Dell, New York

Parsons T 1942/1964 *Essays in Sociological Theory*. Free Press, Glencoe, IL

Pilkington H 1994 *Russia's Youth and its Culture: A Nation's Constructors and Constructed*. Routledge, London

Sansone L 1995 The making of a black youth culture: Lower class young men of Surinamese origin in Amsterdam. In: Amit-Talai V, Wulff H (eds.) *Youth Cultures: A Cross-cultural Perspective*. Routledge, London, pp. 114–43

Schade-Poulsen M 1995 The power of love: raï music and youth in Algeria. In: Amit-Talai V, Wulff H (eds.) *Youth Cultures: A Cross-Cultural Perspective*. Routledge, London, pp. 81–113

Schlegel A, Barry III H 1991 *Adolescence: An Anthropological Inquiry*. The Free Press, New York

Whyte W F 1943/1981 *Street Corner Society: The Social Structure of an Italian Slum*. University of Chicago Press, Chicago

Willis P 1977/1981 *Learning to Labour: How Working Class Kids Get Working Class Jobs*. Columbia University Press, New York

Wulff H 1988 *Twenty Girls: Growing Up, Ethnicity and Excitement in a South London Microculture*. University of Stockholm,

V. Amit

Youth Culture, Sociology of

Youth culture refers to the cultural practice of members of this age group by which they express their identities and demonstrate their sense of belonging to a particular group of young people. The formation of youth culture thus implies boundary drawing. Both the youth culture's distinction from mainstream (adult) culture and its internal coherence help define the members' personal identity and enhance feelings of belonging. With the extension of the life stage 'youth' in the later decades of the twentieth century (Buchmann 1989), the boundaries of youth as an age group have been increasingly blurred. The issue of what constitutes authentic and autonomous cultural expressions of youth (as opposed to the mainstream adult cultural practice) has therefore become more complex. Likewise, the internal differentiation of juvenile cultural practice and the proliferation of youth cultures from the late twentieth century raise the question of the relationship between structural location, socialization, and cultural practice. After a brief history of the concept and the presentation of the major theoretical currents, these issues will be discussed.

1. A Brief History of the Concept

The notion of youth culture dates back to the start of the twentieth century. With the development of youth movements in Germany, such as the *Wandervogel*, representatives of the so-called *Reformpädagogik* noticed the distinct ways of cultural articulation shown by this group of young people, for which they coined the term 'youth culture.' The juvenile cultural orientations and styles of behavior were interpreted as signs of authentic and autonomous cultural practices. This fundamental issue—whether the social group of youth may be characterized by a common authentic lifestyle developed within the self-defined community of peers—would accompany the theoretical debates within the sociology of youth over large stretches of the twentieth century. The first input to this scientific discourse can be traced back to the 1920s, when researchers from the Chicago School of sociology engaged in many empirical studies on criminal milieux, especially on delinquent youth groups (e.g., Sutherland 1924). Detailed investigation of groups of juvenile delinquents revealed similarities in their outlook on life, values, and styles of behavior. To characterize these

commonalities, the notion of youth culture was introduced into the scientific literature. But it was not until the 1940s that juvenile cultural practices would be conceived of explicitly in terms of a subculture (Hollingshead 1949). The use of this concept implied, first, the assumption that there is global cultural differentiation between youth and the dominant (adult) society, and, second, the idea of the strong internal coherence of juvenile cultural expressions.

2. *The Cultural Differentiation of Youth from Adult Society*

The two basic assumptions of the early conceptions of youth culture mentioned above can be linked to two influential theoretical traditions within sociology. The first refers to Mannheim's concept of generation (Mannheim [1928] 1997). The second theoretical tradition, linked especially to the name of Eisenstadt (1956), focuses on the peculiarities of the transition between age groups within modern society and related questions of socialization.

2.1 *The Legacy of Mannheim's Concept of Generation*

Even popular conceptions of youth, especially those advanced by the mass media in the second half of the twentieth century, would often refer to the concept of generation to characterize a juvenile way of life. Descriptions of youth, such as the '68 generation,' the 'beat generation,' the 'no-future generation,' or the latest version of 'generation X' document the attractiveness of this concept.

The major input to the scientific debate on generations goes back to Mannheim ([1928] 1997) who developed the generational approach. Mannheim maintains that the formation of a generation is most likely to occur during the life stage of adolescence. During these formative years, the experience of social circumstances and historical events will have especially profound and long-lasting effects. However, the common location in the historical processes, or the generation location (*Generationslagerung*), is not sufficient for the formation of a generation as an actuality. Although the common location in history provides access to a specific range of potential experiences, the generation as actuality (*Generationszusammenhang*) presupposes the '*participation in the common destiny* of this historical and social unit' (Mannheim [1928] 1997, p. 46 [Mannheim's emphasis]). Mannheim acknowledges the possibility that groups of youths exposed to the same generation location may '*work up the material of their common experiences in different specific ways*' (Mannheim [1928] 1997, p. 47 [Mannheim's emphasis]). Hence, these groups constitute separate generation units (*Generationseinheit*). Mannheim's approach has provided two major theoretical contributions to the sociology of youth culture. The first input is the conceptualization of generations in relation to social processes and historical events as opposed to biological age boundaries. The second insight relates to differences in the ways in which young people within the same generation location cope with their common experiences and thus develop different cultural orientations and styles of behavior. Unfortunately, much of the empirically oriented research on youth culture in the second half of the twentieth century, while incorporating the concept of generation, did not make use of Mannheim's fine distinctions.

2.2 *The Role of the Juvenile Subculture in the Transition to Adulthood*

In Mannheim's theoretical frame, the role of socialization processes in adolescence is related to the profound and long-lasting effects of historical events. By contrast, structural functionalism as developed by Parsons and further elaborated by Eisenstadt (1956) focuses on the role of age-homogeneous groups (peer groups) as agents of socialization. According to structural–functional theory, these groups fulfill the function of facilitating the young people's transition to adulthood in modern society. In order to become full members of modern society (i.e., adults), adolescents must undergo a fundamental reorientation. As children, they are predominantly integrated into the family, a sphere of social participation characterized by particularistic relations. The positions and roles in this action sphere are mostly ascribed, and behaviors are scarcely evaluated on the basis of achieved criteria.

In the adult world, however, the range of social participation is much broader, demanding much more elaborate action capacities. Moreover, most spheres of social participation are structured according to universalistic principles. Action outcomes are therefore attributed to individual competencies. In this context, formal and informal age-homogeneous groups of youths may be regarded as 'interlinking spheres' helping to bridge the gap between the demands of the children's world and those of the adult world. Norms and values predominant in peer groups usually show a specific combination of particularistic, as well as universalistic, elements. Hence, participation in peer groups and their culture provides a great opportunity for juveniles gradually to acquire action competencies compatible with universalistic principles. Given the relative absence of direct social control by adults, age-homogeneous groups provide a social space in which autonomous and self-responsible action patterns may be learned without being immediately exposed to adult sanctions. The flipside of such a sanction-free space may be the development of peer cultures that are in more or less stark contrast to the norms and values of the adult world.

Against this background, theoretical thinking about youth culture in the 1950s and early 1960s was largely dominated by the idea that youth as a social group adheres to common values, goals, and behaviors distinct from those of the adult world. Most influential in this respect was the work by James Coleman (1961), who described the independence, the internal homogeneity and coherence of the world of youth, for which he coined the term 'peer culture.' Empirical evidence, however, did not support the global differentiation of youth culture from that of the dominant (adult) society. Various forms and styles of juvenile cultural expression have always existed, displaying some common elements. With the onset of the general critique of structural functionalism within sociology, the focus on the functional necessity of adolescents' orientation towards peers and their culture gradually faded and gave way to theoretical thinking that paid greater attention to the internal differentiation of youth culture.

3. *The Structure of Social Inequality in Modern Society and Class-specific Youth Cultures*

The reorientation within the sociology of youth culture became especially evident with the research conducted at the Centre for Contemporary Cultural Studies (CCCS) in Birmingham, UK (Hall and Jefferson 1975, Willis 1990). Although the concept of subculture remained essential to their understanding of juvenile cultural expressions, this research tradition not only linked it to the dimension of age (generation), but also stressed the connection to the dimension of social inequality, in particular to social class. Within this theoretical framework, juvenile subcultures are regarded as the outcome of class-specific socialization processes. Hence, youth culture is internally differentiated and, most likely, several youth cultures coexist at any given period of time within modern society. The CCCS model assumes that the cultural sphere of modern society is made up of two independent cultures, hierarchically related to each other. The bourgeois culture is the dominant one and the working class culture is the discriminated one. Within these cultural entities, differentiation by age, ethnicity, and sex may occur. The differentiated subcultures show a varying degree of overlap with the superordinate culture, incorporating common, as well as distinct, cultural elements. Moreover, they are characterized by relatively high internal consistency and relative stability, encompassing all life spheres.

Theoretically, the aforementioned characteristics of subcultures are linked to socialization processes. As members of a particular social class, individuals learn the class-specific cultural repertoire and become equipped with particular resources, with which they try to solve the problems of daily life. Given the multiplicity of membership in modern society (e.g., class, gender, age), particular tasks may occur that demand new ways of coping. With respect to the age group of youth, juveniles have to cope with their indeterminate social position—no longer children, yet not full adults. This particular social situation predisposes them to invent cultural practices that help them to define who they are (i.e., identity construction) and to develop social bonds and a sense of belonging. In this influential theoretical tradition, juvenile subcultures are interpreted as class-specific answers to the particular demands of the life stage 'youth.' Although all young people are confronted with these demands, the social milieu greatly determines the material and symbolic resources available to them, thus resulting in class-specific cultural practice.

4. *Structural and Cultural Change Within Late Modern Society: Youth Cultures as Youth Styles*

The basic assumption of the cultural model developed by the researchers of the CCCS maintained that the precise social location was essential to the definition of juvenile subcultures. The cultural practice of youth was conceived to be the symbolic expression of their social existence. Observers of youth culture in the 1980s noticed, however, the proliferation of juvenile cultural expressions characterized by a looser coupling with young people's social positions (Cavalli and Galland 1995, Chisholm et al. 1995, Thornton 1996). The weakening of the linkage between social location and cultural practice, furthermore, was accompanied by the growing importance of the element of style in young people's cultural expressions, i.e., the expressive and esthetic presentation of one's identity and social belonging.

Although the research tradition of the CCCS, focusing on class-specific differentiation of youth cultures, paid particular attention to the element of style (Hebdige 1979), the full meaning of style as the main ingredient of youth culture was not yet recognized. In this tradition, style was conceived mainly to be a particular means by which the unresolved problems and contradictions inherent in the respective adult class cultures could be expressed symbolically. More recent conceptions of the role of style in youth culture stress the intricate relations to several structural changes within late modern society. In particular, three facets of social change have been identified: first, the growing structural differentiation, resulting in the blurring of age boundaries and the changing significance of social standing; second, the increasing importance of the social sphere of leisure; and third, the role of the media in constructing and disseminating symbols of youth style. The three facets of social change and their implications for juvenile cultural practice are discussed below.

Although still hotly debated, social change in late modern society has been associated with the growing

differentiation of the system of social inequality and the concomitant diversification of individual life course patterns (Buchmann 1989). Advocates of this view link these changes to the expansion of the educational system, which provided access to medium-level and higher education for the broader masses. The profound changes within the institution of education resulted in a substantial extension of the life stage 'youth,' and especially in the blurring of age boundaries between youth and adulthood (Buchmann 1989). The massive expansion of the welfare state is regarded as another major element of these changes, making life less dependent on the family. With the broad application of information and communication technologies, profound changes in the economy and the labor market occurred, transforming work roles and occupational careers. Hence, the pathways from school to work, the major role and status change in the transition to adulthood, became increasingly diverse. Under conditions of increasingly differentiated social positions and life course patterns, the ways in which individuals express their identities and demonstrate their sense of belonging is affected as well. The quasi-automatic integration into social milieux and collectivities can no longer be taken for granted. Rather, the individual's identity and his or her group membership become more dependent on action capacities and competencies. Cultural practices, especially those including stylistic elements, therefore advance to particularly suitable means to symbolically present one's self and express one's social affiliation. Hence, juvenile cultural practice crystallizes into particular youth styles.

The structural transformation of late modern society described above also brought about a shift in the relevance of various spheres of life. In particular, the social sphere of leisure became much more important, both as an element of lifestyle and as a means of cultural distinction. Characterized by a strong component of social visibility, leisure-time activities provide ample opportunities to satisfy the needs of the expressive and esthetic presentation of one's self and one's social belonging. With regard to youth especially, cultural practices such as music, dancing, movies, visual arts (e.g., comics), particular sports (e.g., skateboarding), and fashion (e.g., clothing and hairstyles) are preferred means of expressing a distinct way of life that is recognized by others as a sign and signal of a particular identity and group membership (Austin and Willard 1998, Danesi 1994, Epstein 1998). In fact, late twentieth/early twenty-first century youth cultures are mostly integrated into leisure time and oriented toward mass media and consumption. They are to a large degree leisure styles. The heterogeneity of juvenile cultural scenes and styles observed at the beginning of the twenty-first century can be attributed both to increasing degrees of freedom in the individual's scope of action, due to the structural changes in late modern society discussed above, and the great abundance of cultural materials provided in the sphere of leisure. According to the particular needs of social representation, young people may assemble and reassemble stylistic elements of various origins in ever new ways to form distinct styles of juvenile cultural practice.

For the swift circulation and dissemination of the cultural signals of youth styles, media and computer culture play an ever greater role (Kellner 1994). The formation of youth styles and scenes may be regarded as an interactive process with the media. Styles for clothing and hair generated in informal groups of youths, for example, are often quickly appropriated by the media and the market, and turned into fads and fashion. As a result, larger segments of youth may get the opportunity to participate in this style, thus contributing to its dissemination. And, vice versa, cultural materials provided by the media may be appropriated and reinterpreted by young people and utilized in the construction of juvenile styles. The mutual appropriation and simultaneous transformation of cultural signs and signals by groups of young people and the media is the major reason for the rapid fluctuation and swift turnover of youth styles at the beginning of the twenty-first century. As soon as juvenile cultural expressions have become integrated into the mainstream (adult) cultural world, they can no longer serve as means of cultural distinction. However, this is essential for presenting one's identity and expressing a sense of belonging to a particular group of youth.

5. *Future Directions in the Sociology of Youth Cultures*

Scholars of youth culture agree that in recent years the field has been dominated by detailed descriptive studies of the proliferating forms of juvenile cultural expressions. As valuable as these studies are for gaining insights into contemporary juvenile cultural orientations and styles of behavior, theoretically based issues and matters should again become more prominent in future research. To date, it is largely unknown which structural or cultural factors are responsible for the predominance of class-related juvenile cultural expressions and, vice versa, for the predominance of generation-related juvenile cultural styles. Furthermore, the impact of juvenile cultural innovations on cultural change in society at large remains largely unknown. Likewise, the role of the media in the construction, dissemination, and appropriation of juvenile lifestyles is not well understood.

See also: Adolescent Behavior: Demographic; Adolescent Development, Theories of; Delinquency, Sociology of; Generations in History; Generations: Political; Generations, Relations Between; Generations, Sociology of; Life Course in History; Youth Culture, Anthropology of; Youth Gangs; Youth Movements; Youth Sports, Psychology of

Bibliography

Austin J, Willard M N (eds.) 1998 *Generations of Youth: Youth Cultures and History in Twentieth-Century America*. New York University Press, New York
Buchmann M C 1989 *The Script of Life in Modern Society: Entry into Adulthood in a Changing World*. University of Chicago Press, Chicago
Cavalli A, Galland O (eds.) 1995 *Youth in Europe*. Pinter, London
Chisholm L, Budiner P, Kruger H H, du Bois Reymond M (eds.) 1995 *Growing Up in Europe: Contemporary Horizons on Childhood and Youth Studies*. Walter de Gruyter, Berlin
Coleman J S 1961 *The Adolescent Society*. Free Press, New York
Danesi M 1994 *Cool: The Signs and Meaning of Adolescence*. University of Toronto Press, Toronto, ON
Epstein J (ed.) 1998 *Youth Culture: Identity in a Postmodern World*. Blackwell, Oxford, UK
Eisenstadt S N 1956 *From Generation to Generation*. Free Press, New York
Hall S, Jefferson T (eds.) 1975 *Resistance through Rituals: Youth Subcultures in Post-War Britain*. Unwin Hyman, London
Hebdige D 1979 *Subculture: The Meaning of Style*. Methuen, New York
Hollingshead A B 1949 *Elmtown's Youth*. Free Press, New York
Kellner D 1994 *Media Culture*. Routledge, New York
Mannheim K [1928] 1997 The problems of generations. In: Hardy M (ed.) *Studying Aging and Social Change: Conceptual and Methodological Issues*. Sage, Thousand Oaks, CA, pp. 22–65
Sutherland E M 1924 *Principles of Criminology*. University of Chicago Press, Chicago
Thornton S 1996 *Club Cultures: Music, Media and Subcultural Capital*. Wesleyan University Press, Hanover, NH
Willis P 1990 *Common Culture: Symbolic Work at Play in the Everyday Culture of the Young*. Open University Press, Milton Keynes, UK

M. Buchmann

Youth Gangs

1. Introduction

Youth gangs have been a part of the American culture since the eighteenth century (Sante 1991, Sheldon 1898). Gangs are diverse in American society; not all gangs are the same. Traditionally, researchers have identified poverty as a common bond between members of youth gangs (Miller 1958, Cohen 1960, Moore 1978, Taylor 1989, Curry and Spergel 1992). Many gangs today are still largely populated by young people from inner city neighborhoods, characterized by high unemployment, high educational dropout rates, and a general feeling of hopelessness. Yet, today, not all gangs share an impoverished background. Youth gangs are springing up in affluent suburbs and in rural areas (Goldstein and Soriano 1994, Spergel and Curry 1993, Evans et al. 1999). At the start of the twenty-first century, many traditional views about what constitutes a gang are still prevalent. Defining gangs has been, and continues to be, a problem in the USA. The practice of defining a gang has become highly subjective, ranging from disorganized noncriminal to highly organized criminal organizations (Taylor 1989). This diversity continues to be debated by professionals in social work, law enforcement, education, and community organizations.

The following sections include discussions on the difficulty of defining youth gangs (see Sect. 2), as well as covering the diverse ethnic membership of youth gangs in America (see Sect. 3). Myths of gangs and gang members are discussed because society tends to view gangs from a myth-based perspective. In order to understand the problem, define the problem, and prevent the problem of gangs, a reality-based approach must be found (see Sect. 5). These discussions rely on available gang studies for a broader perspective of the gang problem in the USA today. Following the conclusion, an extensive reference list is provided to assist with further review of the topic.

2. Defining Gang

Historically, defining the term *gang* has been a problem in the USA. Early definitions did not focus on criminal activity, but rather on delinquent behaviors. Consensus of definition does not exist today. Many researchers have created their own versions of what constitutes a gang (Klein 1968, Huff 1989, Taylor 1990, Monti 1993). In fact, every organization that comes in contact with gangs has created an operational definition of the term to suit its own needs. Although law enforcement's focus, understandably, is on gangs that are breaking the laws and mores of society, there isn't one definition for all jurisdictions to follow (Cromwell et al. 1992).

Neighborhoods, families, and communities also come in contact with gangs and must find a way to define existing groups. Groups of youth in many communities across America may appear similar to gangs, yet they have no criminal past or future. One researcher called these types of groups adjunct, advocated political groups, and underscored that often they comprise productive young citizens of a particular community (Taylor 1993).

Gangs vary by membership and activity. Generally, researchers agree that there are different degrees of gang involvement and activity (Fagan 1989, Huff 1990, Taylor 1990). Although some young gang members earn a living through drug trafficking, illegal weapons sales, and other criminal activities, others join gangs for noncriminal reasons. For example, a young person may join a neighborhood gang to get protection in their neighborhood; another may join to gain respect they would otherwise not receive at home or in the community; and still others may join for the camaraderie. Whatever the reason, gang members are as diverse as the gangs themselves.

What constitutes a gang to some communities is clearly not always tied to criminal behavior. Some will choose to define gangs by categorizing them as organized or unorganized criminal groups. Yet others will be swayed by images that are often fictional (see Sect. 4).

3. *Overview of US Youth Gang History*

The earliest record of youth gangs in the USA may have been as early as the eighteenth century, at the end of the American Revolution in 1783 (Sante 1991). Some of these gangs were known as Smith's Vly gang, the Bowery Boys, the Broadway Boys, and the Long Bridge Boys (Osman 1992). These gang members were not necessarily criminals who committed violent crimes, although they did fight with rivals.

Redfield (1941) believed the first gangs in the USA migrated from Mexico in the nineteenth century after the Mexican Revolution in 1813. It was around this time that the first criminal gangs were formed. These were youth gangs involved in criminal activity, due in part to the economic situation and the increase in population in urban areas.

The Irish immigrants were the first to form criminal gangs in New York City in the 1820s. The 1850s saw an increase in gang membership as urban government corruption was rampant. During the early years after the civil war ended, migration patterns increased for industrial centers like New York, Chicago, and Detroit. The Irish, Jewish, and Italian immigrants who came were impoverished and formed gangs based on ethnicity. It was also at this time that drugs were introduced to gang life.

The early part of the twentieth century was dominated by economic turmoil. Due to the increase in population of the urban areas and the decline in the economy, the gap between the rich and poor grew. According to Thrasher (1927), there were 1,313 gangs in Chicago in the 1920s. Many of these were ethnic gangs. The increase in the number of immigrants in urban areas continued throughout this century. In particular in the early 1940s large numbers of Puerto Ricans entered New York City. This fact, along with a growing African-American population from the south, contributed to the large minority populations in northern cities. While Irish, Italian, Jewish, and other Eastern European ethnic groups were establishing their communities, Puerto Ricans and African-Americans, including strong pockets of black West Indians, became a strong presence. Racial conflict was clear in northern big cities such as Detroit, Michigan, where one of the worst race riots in American history took place in 1943. Groups of white youth gangs roamed the city attacking black citizens (Shelden et al. 1997). Around the same time in Los Angeles, the Zoot Suit Riots of 1943 were underway. In these riots, white residents and visiting soldiers harassed and beat up young Chicano men who dressed in the popular zoot suit style of clothing (Moore 1978, Shelden et al. 1997).

Following World War II, gangs became more organized. As the population in ethnic ghetto neighborhoods shifted from largely Eastern European in the early 1950s toward African-American and Latino in the 1960s and 1970s, so did the ethnicity of gangs in those areas. Weapons, illicit narcotics, and sexual encounters came with delinquent and criminal street youth gangs in the 1950s and 1960s. Urban cities experienced poverty, crime, and inferior education. Some researchers contend that young people, under-educated and without access to good jobs, become frustrated with their lives and join gangs as an alternative to boredom, hopelessness, and poverty (Shaw and McKay 1942, Cloward and Ohlin 1960, Taylor 1990). Today, the composition of youth gangs is changing with the onset of affluent suburban and rural youth gangs (Korem 1994, Evans et al. 1999).

During the 1960s, the connection between criminal gangs and youth gangs varied from neighborhood to neighborhood, and city to city. Female gangs evolved from male gangs in some of these cities. The media focused on youth culture and its music, dancing, and desire to have its own identity. Youth culture defines itself within the context of hip hop, grunge, and other youth experiences and expressions. Many times gangs were blamed for anything negative or delinquent (Miller 1975). The issue of social class is underscored by how some communities responded to the challenge of gangs. Many middle class communities supported recreational programs for youth. Unfortunately, the concept of positive youth development only applied to the middle class youth in suburbia. Inner city youths were not included in positive youth development, instead youth detention was used to regain and maintain social control.

The Watts Riots of 1965 did for African-American gangs what the Zoot Suit Riots did for the Chicano gangs in 1943. Basically, the outcome was that the media began to portray the youths in a negative light, and the youths began to see themselves differently. They chose to see themselves as defiant rather than defeated, and to 'redefine exclusion as exclusivity' (Shelden et al. 1997).

For the most part, the 1960s saw a decline in gang violence. By the early 1970s, they were back on the front page. At this time, drug use in gangs began to decrease while the use of violence began to increase. The gangs of the 1970s, with greater use of guns, were more lethal in their encounters than previously (Horowitz and Schwartz 1974, Taylor 1989). By the 1980s, the urban ghettos became overcrowded and the youth gangs transformed themselves into entrepreneurial organizations with the crack cocaine epidemic of that decade (Sanchez-Jankowski 1991, Taylor 1989).

The youth gang problem in the USA today affects communities of all sizes. Although youth gangs are

more prevalent in urban areas like Detroit, New York, Chicago, and Los Angeles, they can also be found in affluent suburban and rural areas (Korem 1994, Evans et al. 1999). According to the 1996 National Youth Gang Survey, approximately 31,000 gangs and 846,000 gang members were active in the USA in that year. The survey respondents reported that, especially in rural areas, the gang problem began to effect them in 1994. This study also found that the proportion of Caucasian gang members was especially high in rural areas.

In looking at youth gangs, it is imperative to include not only those in urban centers, but also those in suburban and rural communities. Gang membership today is no longer exclusive to school drop-outs from impoverished, overpopulated urban areas. It can also include affluent middle-class suburban youths and youths from underpopulated rural areas.

While it is easy to see the daily newscast about Latino and African-American gangs, the gangs of skinheads and White Supremacist are not always defined as gangs. Remembering the diversity of gangs, including youth gangs, is important. Highly structured, corporate gangs can be found all around the country; but not all gangs are corporate-style gangs, some are loosely held organizations, having virtually no leadership (Taylor 1990).

The idea of defining the term gang is good in theory. The problem is in gaining consensus among the organizations and communities that are effected by youth gangs. Today, gangs are identified based on the needs of the organization that comes in contact with them. Whether a highly organized criminal corporation or a loose-knit group of friends that comes together for camaraderie rather than criminal activities, gangs are too diverse for one definition. Whatever the definition of choice, in defining gangs it is paramount that the distinction is made between criminal and noncriminal gangs. Juvenile youth gangs could be considered delinquent depending on how their respective communities view delinquency. However, criminal behavior is not automatically part of youth gangs. Some researchers have promoted the notion that all gangs are criminal. One of the shortcomings of literature on gangs is the focus on criminality. Delinquency, continued lifetime criminal behavior, is well documented in many academic studies. Yet, there is little if any attention on youth gangs, gangs, or gang members involved in noncriminal activities. This lack of focus does not conclude there are no other gang types; it does, however, point to the posturing and approach of how society has viewed youth gangs in America.

4. *Role of the Media and Cinema*

The debate about what constitutes a gang has been underscored by the portrayal of gangs and gangsters in the media and cinema. Entertainment has become big business as an industry that promotes and sells gangsters, action heroes, and violence.

The media has been developing a bio-sketch of a gang member since the Zoot Suit Riots of 1943 and using it to keep fear of diversity in the minds of all who are willing to be taken in. The Watts Riots of 1965 gave the media more ammunition for its exaggerated view of a gangster. Thanks to their portrayal, many people today believe that a gang member is a Latino or African-American illiterate youth, who comes from a female-headed household, in an impoverished urban area. The suburban gangster is not seen as a threat because he has been portrayed as a literate youth who comes from an intact family with Christian family values, from an affluent community. Given these choices, who would you be afraid of?

American cinema has a long history of gangster movies. From 'West Side Story' to the more hardcore portrayal of criminal street gangsters today, the cinema has literally been a training ground for gang wannabes. Classic movies such as 'The Godfather' give an excellent example of Thrasher's (1927) evolution of street gangs, rising from neighborhood playgroups and evolving into successful criminal street gangs. From the portrayal of rebelling youth (i.e., 'Rebel Without A Cause,' Blackboard Jungle,' 'Wild Ones'…), to star-crossed love (i.e., 'West Side Story'), and the hardcore reality of street gangsters (i.e., 'Menace II Society,' '187,' 'Heat,' 'Boyz in the Hood'), the youth of America did not have to look far for gangster role models.

Society, and in particular some media forces, has been ready to assign responsibility for brutal acts exclusively to so-called vicious gangs. But the reality is that gang members run the spectrum of angels to demons. According to the Bureau of Justice, only 10 percent of gang members are hardcore, violent members. Researchers have found that most are peripherally involved in violence; more important, there is evidence to show that only a small percentage of gang members are responsible for the violence.

5. *Gang Myths*

In order to gauge properly the gang influence in a community, gang myths must be addressed and dispelled. Although there are many myths in existence today, the following are a selection that tend to crop up in research (Goldstein and Huff 1993, Taylor 1990, 1993).

Myth 1: All street gangs are turf oriented.

Reality 1: There are gangs that claim ownership to a particular territory, but this is not the exclusive reason for gang membership. Others include Scavenger gangs (loose-knit, immoral, noncriminal), Commercial gangs (focused on material gain), and Corporate gangs (well organized, highly structured, focused on financial gain by criminal action).

Myth 2: Females are not allowed to join gangs.

Reality 2: Females are joining gangs in record numbers. One female gang supported research, showed females in autonomous gangs involved in organized criminal activities.
Myth 3: There are no gangs in my neighborhood.
Reality 3: Today, no neighborhood, regardless of economic status, is immune to gang membership. Gangs can be found in rural areas, suburban areas, as well as urban areas.
Myth 4: Gang members wear baggy clothes and athletic team baseball hats.
Reality 4: Baggy clothing has become the 'cool' style of dress and not a uniform that all gang members wear.
Myth 5: All gangs have a single leader and a tight structure.
Reality 5: Some gangs are loosely held organizations, having virtually no leadership.
Myth 6: Gangs are a law-enforcement problem.
Reality 6: Gangs are a problem for every member of society including parents, teachers, and police.
Myth 7: I know a gang member when I see one.
Reality 7: This statement opens the door to racism. Using traditional ideas of gang membership would mean that only Latino and African-American youths would be targeted. It is important to remember that youth gang members are diverse in color, style of dress, activities, and background.

6. *Conclusion*

Historically, youth gangs began as a group of young people coming together with a common bond. Today, ethnicity seems to be the bond that is focused on by the media, the cinema, and society. The youth gang problem continues to grow in the USA. It is no longer an urban issue, as shown by the current and rapid onset of suburban and rural gangs. The gang structure is also changing, with gangs of different sizes engaged in activities ranging from non-criminal to criminal. At the start of the twenty-first century, society needs to define the term gang, taking into account the facts that there are different types of gangs ranging from non-criminal to criminal; loosely knit to highly structured; coming from diverse ethnic backgrounds; experiencing different reasons for joining, as well as choice of activities pursued. Youth gangs should be studied and defined by the behavior that is associated with gangs. As difficult as it will be to come to one definition of the term gang, some consensus must be reached if there is any hope of identifying specific prevention and intervention programs.

See also: Adolescent Behavior: Demographic; Gangs, Sociology of; Groups, Sociology of; Poverty and Child Development; Poverty, Sociology of; Social Psychology: Sociological; Urban Poverty in Neighborhoods; Violence and Effects on Children; Violence: Public; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements; Youth Movements, History of

Bibliography

Cloward R A, Ohlin L E 1960 *Delinquency and Opportunity: A Theory of Delinquent Gangs*. The Free Press, New York

Cohen A K 1960 *Delinquent Boys: The Culture of the Gang*. The Free Press, Glencoe, IL

Cromwell P, Taylor D, Palacios W 1992 Youth gangs: a 1990s perspective. *Juvenile and Family Court Journal* **43**(3): 25–31

Curry G D, Spergel I A 1992 Gang involvement and delinquency among Hispanic and African-American adolescent males. *Journal of Research in Crime and Delinquency* **29**: 273–91

Evans W P, Fitzgerald C, Weigel D, Chvilicek S 1999 Are rural gang members similar to their urban peers? Implications for rural communities. *Youth and Society* **30**(3): 267–82

Fagan J E 1989 The social organization of drug use and drug dealing among urban gangs. *Criminology* **27**: 633–69

Goldstein A P, Huff R 1993 *The Gang Intervention Handbook*. Research Press, Champaign, IL

Goldstein A P, Soriano F I 1994 Juvenile gangs. In: Eron L D, Gentry J H, Schlegel P (eds.) *Reason to Hope: A Psychosocial Perspective on Violence and Youth*. American Psychological Association, Washington, DC

Horowitz R, Schwartz G 1974 Honor, normative ambiguity and gang violence. *American Sociological Review* **39**(April): 238–51

Huff C R 1989 Youth gangs and public policy. *Crime and Delinquency* **35**: 524–37

Huff C R 1990 *Gangs in America*. Sage, Newbury Park, CA

Klein M W 1968 Impressions of juvenile gang members. *Adolescence* **3**: 53–78

Korem D 1994 *Suburban Gangs; the Affluent Rebels*. International Focus Press

Miller W B 1958 Lower class culture as a generating milieu of gang delinquency. *Journal of Social Issues* **14**: 5–19

Miller W B 1975 *Violence by Youth Gangs and Youth Groups as a Crime Problem in Major American Cities*. US Department of Justice, NCJ 137446, Washington, DC

Monti D J 1993 Origins and problems of gang research in the United States. In: Cummings S, Monti D J (eds.) *Gangs* State University of New York Press, Albany, NY

Moore J W 1978 *Homeboys: Gangs, Drugs and Prison in the Barrios of Los Angeles*. Temple University Press, Philadelphia, PA

National Youth Gang Survey 1996 Office of Juvenile Justice and Delinquency Prevention, NCJ 173964, Washington, DC

Osman K 1992 *Gangs*. Lucent Books, San Diego, CA

Redfield R 1941 *Folk Culture of Yucatan*. University of Chicago Press, Chicago, IL

Sanchez-Jankowski M S 1991 *Islands in the Street: Gangs and American Urban Society*. University of California Press, Berkeley, CA

Sante L 1991 *Low Life: Lures and Snares of Old New York*. Vintage Books, New York

Shaw C R, McKay H D 1942 *Juvenile Delinquency and Urban Areas*. University of Chicago Press, Chicago, IL

Shelden R G, Tracy S K, Brown W B 1997 *Youth Gangs in American Society*. Wadsworth Publishing

Spergel I A, Curry D G 1993 The national youth gang survey: A research and development process. In: Goldstein A P, Huff C R (eds.) *The Gang Intervention Handbook* Research Press, Champaign, IL

Taylor C S 1990 *Dangerous Society*. Michigan State University Press, East Lansing, MI
Taylor C S 1993 *Girls, Gangs, Women and Drugs*. Michigan State University Press, East Lansing, MI
Thrasher F M 1927 *The Gang*. University of Chicago Press, Chicago, IL

C. S. Taylor

Youth Movements

Youth movements are the organized, conscious attempts by young people to bring about or resist societal change. The formal organization and coordinated activities of youth movements distinguish them from brief episodes of youthful unrest and collective behavior. Rooted in generational tensions and specific sociohistoric conditions, the defining characteristic of youth movements is that they are staffed and carried out largely by young people—typically between the ages of 17–30—who join together to protest adult authority and take it upon themselves to transform society. Ranging from a small number of members to hundreds of thousands of active supporters, youth movements have included student movements, cultural (literary, artistic, musical, ethnic, religious, countercultural) movements, peace and antiwar movements, nationalist and political movements, and environmental–ecology movements. Young people spontaneously may initiate their own movements for change, or youth movements may be sponsored by adults to supplement the activities of larger social and political movements (see *Age, Sociology of*; *Collective Behavior, Sociology of*; *Social Movements, Sociology of*).

1. *The Problem of Youth and Youth Movements*

Throughout history, adults have defined young people as 'a problem.' Wedged between childhood and adulthood, it has long been recognized that young people's idealism, energy, impetuousness, and desire for independence draw them to action and put them at odds with adults. Although incidents of youthful unrest and collective behavior have been recorded since antiquity, full-fledged, organized youth movements did not emerge until the early nineteenth century in Europe. The impetus for youth movement activity stemmed from the forces of modernization and from the Enlightenment, which ushered in the new values of freedom, equality, and self-determination (see *Modernization, Sociological Theories of*). Increasing in number and geographic location from the early nineteenth through the twentieth centuries, youth movements occurred on a global scale during the 1960s and 1980s. Youth movements have toppled governments and have been a force for democracy and societal reform as well as violence, terrorism, and bloody revolution.

The study of youth movements ranges from personal and journalistic accounts to detailed descriptions, analyses, and research investigations by historians, sociologists, political scientists, anthropologists, psychologists and psychiatrists throughout the world. As a result, information about youth movements, while plentiful, is fragmented and scattered among many disciplines and countries.

Several theoretical perspectives have influenced the way youth movements are perceived and studied. Advocates of a life-course approach—based largely on Freudian psychodynamic theory—attribute youth movements to generational rebellion, young people's life-cycle characteristics and needs, and deep-seated emotional conflicts between youth and adults (Erikson 1968; Feuer 1969) (see *Life Course: Sociological Aspects*). From a cohort-generational perspective, youth movements are viewed as products of a rapidly changing social order and unique growing-up experiences that exacerbate age-group tensions and relations, and, at times, may heighten young people's generational consciousness and organized behavior. In addition, because of their varying social locations in society, young people may disagree over the direction that change should take and form generation units or competing groups within the larger generational movement (Mannheim 1952; Ortega 1962) (see *Generations, Sociology of*).

According to socialization theory, young people involved in youth movements are not rebelling against their parents but are carrying out values learned in the home. Schools, universities, peer groups, and the media also act as influential sources of socialization that promote or reinforce youthful activity (Flacks 1971; Keniston 1971) (see *Socialization, Sociology of*). Historical conditions are the focus in another perspective on youth movements. According to this view, the formation, structure, and demise of youth movements are affected strongly by specific local, national, and international trends and events. In general, societal dislocations, inequities, and widespread discontent in conjunction with new opportunity structures and resources explain why youth movements occur during certain periods and not others (Esler 1971; Tilly 1975) (see *Historical Sociology*).

2. *Historical Generational Patterns of Youth Movement Activity*

A survey of modern history indicates that youth movements erupt periodically and became global in scope by the 1960s. Moreover, youth movements tend to cluster during particular eras, followed by periods

Table 1
Historical generations, protest issues, and geographical location of youth movements

Historical generation (dominant period)	Protest issues	Geographical location of youth movements (number of countries)
Young Europe 1815–48 1860–90	Nationalism and extension of Enlightenment values (freedom, equality, self-determination) to educated elites	Europe (6) ($N = 6$)
Post-Victorian 1890–1920	Mass democracy and self-determination; growth and spread of ideological conflict between capitalism and socialism	Europe (5) Asia (2) Latin America (4) ($N = 11$)
Great Depression 1930–40	Crisis in capitalism in West and rise of communism; call for national independence and self-determination	Europe (13) North America (1) Asia (2) Latin America (1) ($N = 17$)
1960s Generation 1960–70	Counterculture spreads throughout world; participatory democracy; anticolonialism and liberation; civil rights, social reforms	Europe (15) North America (2) Asia (11) Africa and Middle East (16) Latin America (16) ($N = 60$)
1980s Generation 1980–90	First universal age of human rights and self-determination; environmental destruction and nuclear power; mininationalism (ethnic, religious, cultural); democratization in East Europe and China	West Europe (17) East Europe (8) North America (2) Asia (10) Africa and Middle East (8) Latin America (11) ($N = 56$)

Source: Braungart and Braungart 1993, 1997

of relatively little youth unrest. A cluster of youth movements represents a historical generation. Over the last nineteenth and twentieth centuries, there have been five identifiable historical generations or waves of youth movement activity: Young Europe, Post-Victorian, Great Depression, 1960s Generation, and 1980s Generation. Historically, most youth movements have formed over issues of citizenship, including freedom, equality, self-determination, nationalism, and cultural expressiveness (Braungart and Braungart 1989b, 1993) (see *Citizenship: Sociological Aspects*). To provide an overview of the growth of youth movements during this period, the five historical generations are described briefly in Table 1.

Young Europe was the first wave of youth movement activity where young people fought for Enlightenment values. Napoleon's defeat in 1815 ushered in an age of nationalism, romanticism, and liberalism that inspired university students to organize movements for national independence (see *Nationalism, Sociology of*; *Revolutions, Theories of*). First, in Germany and then elsewhere in Europe, students called for an end to absolutism and the *ancien regime* in favor of the modern nation-state. The Post-Victorian Generation reflected another period of upheaval in modern history. A new era of mass democracy and self-determination encouraged young people to organize against colonialism and imperialism. *Fin de siècle* youth mobilized to challenge the last of the great empires (Victorian, Austro–Hungarian, Ottoman, Ch'ing, Meiji, Romanov), and young socialists fought nineteenth-century liberals. The Great Depression Generation reacted against the economic failures of capitalism and waged intense ideological struggles involving totalitarian regimes (communist, fascist), authoritarian states, and liberal democracies. The call for national independence reached developing countries by the 1930s.

The 1960s Generation, rooted in the post-World War II baby boom, faced several international challenges, including the East–West Cold War and the growing economic gap between rich and poor countries. In unprecedented numbers, young people demanded freedom, equality, and peace, while countercultural lifestyles and behavior spread rapidly around the world (see *Counterculture*; *Youth Culture, Sociology of*). The 'revolution of rising expectations' generated youth movement activity in developing countries. The 1980s Generation mobilized over issues relating to human rights, freedom, and democracy. The Cold War was under attack by young people in Western and Eastern Europe; youth in Asian countries demanded greater democracy; the North–South economic gap worsened and spawned youthful protests in developing countries; and ethnic and religious confrontations involving young people erupted in all global regions. What started in Berkeley in 1964 had reached Beijing by 1989.

3. *Intergenerational and Intragenerational Dynamics of Youth Movements*

An examination of several hundred youth movements in history indicates a number of common patterns in generational activity. Specifically, much of the momentum for youth movements is derived from the intergenerational conflict between young people and adult authorities and the intragenerational conflict among competing generation units or youth groups. When the research is pulled together to provide an understanding of youth movement activity, the principal dynamics of generational conflict are identified and suggest a synthesis of the various theoretical perspectives.

Intergenerational conflict is the central dynamic of youth movement activity and plays a significant role in mobilizing young people to transform society. The generational tensions, misunderstanding, and conflict between youth and adults stem from two sources: (a) the life-course differences, needs, and orientations of young people that set them apart from adult needs and interests; and (b) the distinct growing-up experiences of each cohort in a rapidly changing society. During routine times, youth and adults manage to cooperate or at least coexist. However, during certain extraordinary periods, specific historical conditions—especially a large educated youth cohort, institutional discontinuities, and opening-up eras—heighten young people's dissatisfaction with society and desire for gradual reform or total change. These sociohistorical conditions, in conjunction with effective leadership and opportunities for youthful mobilization, facilitate the formation and spread of youth movements.

The dynamics of age-group mobilization involve the intensification of young people's peer and generational identification (see *Identity: Social*). Adults and adult-run institutions are 'deauthorized' or chastised for the mistakes and disappointments they created in society, and young people 'authorize' or empower themselves to bring about or resist the desired change. As a youth movement becomes more active and visible, adult authorities typically react by attempting to curtail youthful protests, using either relatively peaceable methods or violence and repression. The wider adult population may support or oppose the youth movement activity and strongly influences its success or failure.

Once a youth movement challenges adult authority, intragenerational conflict erupts within the youth generation. That is, although young people may concur that society needs reform, youth groups are likely to disagree and compete over the direction, means, and extent of the proposed changes. During each historical generation, youth movements fighting for radical change are countered by youth groups defending the status quo or championing the status quo *ante*. Whether initiated by young people or sponsored by adults, extreme utopian as well as ideologically moderate generation units may organize that reflect an array and intensity of social and political views (revolutionary, progressive, moderate, conservative, and reactionary). The various opposing generation units compete over the control of the larger generational movement (Braungart and Braungart 1989a, 1989b, 1997).

4. *The Impact of Youth Movements*

As youth movements unfold, the dynamics of inter- and intragenerational conflict sustain the momentum and largely determine how the movement is played out—as a positive source of societal renovation and renewal or fraught with injury, destruction, and death. Historically, youth movements have ranged from being mildly disruptive to thoroughly destabilizing; they have been short- or long-lived; and they have been a significant force for extending democracy and citizenship as well as for totalitarian repression and genocide. Over the nineteenth and tweentieth centuries, youth movements have become a powerful means for young people to bring about or resist change in every part of the world.

With their personal identity and future tied to the vitality of the nation or state, young people readily are recruited into movements to reform or revolutionize society by charismatic leaders who are successful in connecting youthful life-course needs to movement objectives. Specific conditions and opportunities account for why youth movements erupt during particular periods in history. In all likelihood, future generations of young people will mobilize over issues that challenge their contemporaries. The question is not whether a new wave of youth movement activity will arise, but when and over what issues?

Bibliography

Braungart R G, Braungart M M 1989a Generational conflict and intergroup relations as the foundation for political generations. In: Lawler E J, Markovsky B (eds.) *Advances in Group Processes*. JAI Press, Greenwich, CT, Vol. 6

Braungart R G, Braungart M M 1989b Political generations. In: Braungart R G, Braungart M M M (eds.) *Research in Political Sociology*. JAI Press, Greenwich, CT, Vol. 4

Braungart R G, Braungart M M 1993 Historical generations and citizenship: 200 years of youth movements. In: Wasburn P C (ed.) *Research in Political Sociology*. JAI Press, Greenwich, CT, Vol. 6

Braungart R G, Braungart M M 1997 Why youth in youth movements? *Mind & Human Interaction* **8**: 148–71

Erikson E H 1968 *Identity: Youth and Crisis*, 1st edn. Norton, New York

Esler A 1971 *Bombs, Beards, and Barricades: 150 Years of Youth in Revolt*. Stein and Day, New York

Feuer L S 1969 *The Conflict of Generations*. Basic Books, New York

Flacks R 1971 *Youth and Social Change*. Markham, Chicago

Keniston K 1971 *Youth and Dissent*, 1st edn. Harcourt Brace Jovanovich, New York

Mannheim K 1952 The problem of generations. In: Kecskemeti P (ed.) *Essays on the Sociology of Knowledge*. Routledge and Kegan Paul, London

Ortega Y, Gasset J 1962 *Man and Crisis*, 1st edn. Norton, New York

Tilly C 1975 Revolutions and collective violence. In: Greenstein F I, Polsby N W (eds.) *Handbook of Political Science*. Addison-Wesley, Reading, MA, Vol. 3

R. G. Braungart and M. M. Braungart

Youth Movements, History of

While 'youth,' the age between childhood and adulthood, varies from one culture or period of history to another in terms of form, duration, and degree of regulation, the 'young generation' is always an object of the older generations' focused attention, as adults try to prepare young people for their role as full (male or female) members of their society, passing on to them the generally accepted values, norms, and viewpoints by regimes of education and discipline. As history shows, this constellation has again and again led to individual and collective group conflicts. Since ancient times, young people have often found themselves in the front lines of diverse political and/or social movements. However, the 'youth movement' as a form of the 'young generation' more or less violently revolting against the adult world is a 'modern' phenomenon. It first appeared in the 'old' European states around 1800, emerged in such 'young' states as the USA and Australia in the early twentieth century, and has increasingly turned up in countries of the so-called Third World since 1950.

1. Youth Movements as an Expression of Generation Constellations in the 'Modern Age'

Collective revolt by groups of young people against the conditions in which they live may aim for more or less general political, social, or cultural change, essentially telling the older generation to 'move over'; or it may target specific social trends which the younger generation considers outrageous, threatening, or inhuman, such as repercussions of political upheaval, the loss of civil rights, the destruction of the social environment, or the deterioration of economic prospects. In this sense 'youth movements' need to be clearly distinguished from 'youth welfare' and 'youth work,' which are activities led or dominated by adults to socialize the younger generation. These include the youth organizations of established parties and youth organizations with a national or ideological background (such as, for instance, most of the scout organizations), not to mention all types of national youth organizations. It may sometimes be hard to draw a line between a 'youth movement' and ventures for 'youth welfare' in certain instances. As a rule, however, modern 'youth movements' are expressions of a specific form of conflict between generations and as such have spread throughout the world. Since the early nineteenth century, the predominant style elements have been those of male societies, whereas female members have tended to play a minor role. It is only in recent years that this has been changing.

While generation conflict constellations do represent one of the most important roots of youth movements, they are not simply reflections of the natural relations between parental and child age groups. According to the sociologist Karl Mannheim (1928), generation constellations reflect the distinct mental dispositions of each generation, that is the intellectual horizons and emotional needs that the members of an age group develop in their youth in a specific historic context. As a result, the various age groups in a society experience, perceive, and pass judgment in distinctly different ways. The 'Modern Age' uprooted the tradition of passing on experience, as the acceleration of change in the industrial societies of the nineteenth century devalued the knowledge and experience of the older generations as irrelevant, increasingly leading to widespread disregard for the elderly who had previously been held in high esteem. This was accompanied by the development of a youth cult, as industrial production, urban culture, and the capitalist economy began to give preference to youth over old age as stronger, more dynamic, more ready to learn, and more open to change. The older generations meanwhile held sway at the hubs of political power where decisions about future developments are made, rigorously defending their authority and interests. Educational reforms, increasing possibilities for young people to participate politically and to articulate their interests, supported by the universally accessible mass

media, combined to promote burgeoning self-confidence and corresponding expectations among younger and younger age groups. Advertising for consumer goods added to 'youth emancipation,' so that, particularly in times of upheaval, latent constellations between age groups developed into a conscious and distinct frame of reference and identity. Such constellations have been particularly apt to cause individual groups in the 'young generation'—often among students—to develop a strong sense of identity and radical concepts, making massive demands on the 'responsible' older generations including shows of aggression to the extent of using violence, propelled by the belief in a revolutionary mission.

2. *On the History of 'Youth Movements' Since the Nineteenth Century*

Nineteenth century Europe abounds in examples of such generation constellations and the youth movements that they engendered. These include the student fraternities in Germany around 1820 who fought against the reactionary politics of the royal houses after the defeat of Napoleon; Guiseppe Mazzini's 'Young Europe,' which he founded in Switzerland in 1834 along with other national 'youth movements' (including, beside Giovina Italia, La jeune France, Junges Polen and Junges Deutschland); the Narodniki movement of the Russian intelligentsia in the 1860s (where a plan was discussed to exclude all people over the age of 25 from public life in order to enable radical reform); and finally, at the end of the nineteenth century, the 'Young Turks' in northern Greece under Ottoman rule, who did manage to overthrow the Ottoman Empire and come to power in 1909, Ataturk later emerging from their midst.

While the young activists in these movements (male dominated throughout) had mainly political goals, notwithstanding some bohemian excursions, and went as far as to plan and go through with coups and assassinations, the German Youth Movement of the 'Wandervögel' (viz., 'migrant birds') and the 'Freideutsche' (viz., 'free Germans'), which emerged in the urban centers of the German Empire as well as in Switzerland and Austria after 1900, is a unique example of a highly romantic youth movement which was primarily critical of modern civilization, abjured any political involvement, and developed an independent youth culture staking out an autonomous sphere of freedom for itself. Its members were mainly members of the educated classes, whose intellectual leadership was in decline as a result of the rise of a new technical and economic elite. The educated classes propagated 'simple living' to counterbalance the supposedly demoralizing influence of urban mass culture, sought out the folk traditions of bygone days, and dreamed of creating a 'New Man' through a general reform of everyday life. This 'New Man' was to be the leavening in the dough of society to promote greater humanity.

The movement's highly idealistic motto was coined in 1913 at an open-air festival on the 'Hohe Meißner,' a mountain near Kassel: 'The 'Freideutsche Jugend' seek to shape their lives by themselves, responsible to themselves, and true to themselves.' In hindsight, history has shown that anti-political idealism and the fantasy of 'breeding' an improved version of humankind were to be abused by racist demagogues and ideologues in the 1930s to mobilize a new 'young generation' (born in the years around World War I) for totally inhuman and destructive goals, leading directly to World War II and National Socialist genocide. As a movement of youth culture, the German Youth Movement did however manage to promote, firstly, women's emancipation in the ranks of the autonomous girls' movement and, secondly, migration to Palestine and the beginnings of the Kibbutz movement by empowering the Jewish youth groups which split off from other youth groups in response to the latter's frequently nationalist tilt.

This brings the history of youth movements well into the first half of the twentieth century, when throughout the world independent movements of the respective younger generation engaged in at least temporary social and political activism in increasing numbers, criticizing the political situation or giving impulses for political, social, and cultural change. There are many constellations where national politics were influenced by the specific experience of a whole generation. Examples include the Japanese reform circles socialized during the Meiji-Era (the first Japanese constitution being instituted in 1889) and the Spanish 'Generation of 1898,' whose formative experience included the shock of defeat in the Spanish–American War. The greatest influence during the second third of the twentieth century, particularly in Europe, but also for example in Australia, seems to have been that of the 'Generation of 1914' (Wohl 1980), which has been called the 'lost generation.' The way in which this generation dealt with the experience of World War I, particularly in the nations which had lost the war, created explosive constellations of immense historic consequence: Not coincidentally, the fascist movements, particularly in Italy and Germany, styled themselves as revolutionary movements of a 'disinherited' young generation, only to form a dangerous alliance then, with the younger 'Great Depression Generation,' the former leading the latter.

The American 'New Deal Generation' on the other hand, while belonging to this same age group, had totally different aims, seeking a deep-rooted democratization of American society and an expansion of social welfare institutions. While they were propelled by unmistakable self-confidence and a sense of mission, this generation did not develop any discernable forms of youth movement. However, the political

opening of American society that they promoted was to pave the way for the coming independent social movements that were to be carried mainly by young people, including the Civil Rights Movement and Women's Liberation. Beginning in the 1920s, liberation movements emerged in other parts of the world as well, demanding civil rights, national independence, and cultural autonomy for nations or cultures that had lived under colonialism or other forms of oppression (such as the black population in South Africa and the Aborigines in Australia). Whether such movements can be connected specifically to the young generation needs to be studied case by case, but there is no doubt that large groups of young people were among the followers of charismatic leaders or groups.

It is just as difficult to decide whether the subculture of young street gangs and peer group cliques (such as hooligans or skinheads) found in almost all industrialized states since the early twentieth century—primarily among the underprivileged social classes—are actually part of youth movements as defined here. They, too, clearly react in their way to their actual living environment, aim to provoke the older generations, and sometimes encounter massive persecution (as did the 'Edelweiss Pirates' under German National Socialism).

Since World War II, and particularly since the end of the 1950s, when a generation of young people without any first-hand experience of war came into its own, waves of youth movements have become quite common and have developed new forms, peaking in the second half of the 1960s, when student movements began simultaneously in a number of countries, following in the tracks of such predecessors as the Easter March movement and such counter-culture groups as the 'Beatniks' in the USA, the 'Angry Young Men' in Great Britain, and the young 'Existentialists' in France. Beginning in the early 1960s, generation conflicts specific to the different countries began to mesh with the international involvement of young people with problems they considered to be of worldwide importance: the massive rearmament during the 'Cold War,' the growing atomic power potentials, the exploitation and poverty in the so-called 'Third Word,' the military intervention by the major powers in regional conflicts (particularly the involvement of the USA in Vietnam), the destruction of renewable natural resources, and the increasing danger of global ecological catastrophes due to reckless industrialization, the disregard for civil rights in many countries despite UN declarations, the continued inequality between men and women, and so on. Some of the initiatives originally organized regionally managed to become established as large international organizations, such as Amnesty International and Greenpeace. Such initiatives were always attempts to assert new perspectives and political and cultural ideals against those of the older generation (coining such phrases as 'Don't trust anybody over 30'), in the extreme case of the Chinese 'Cultural Revolution' and later the famous events of the Tienanmen Square in Beijing 1989 causing a great deal of bloodshed. The young generation's interest in criticizing prevailing conditions and in replacing the established authorities gave all kinds of ideologues and demagogues a chance to preach new catechisms to the young or to install themselves as their new leaders. The often overblown idealistic expectations pursued with youthful pathos were usually watered down or twisted out of shape in public debate, and in the end were confronted with both the complexity of social reality and with the fact that as their proponents got older and 'wiser,' their attitudes changed, with corresponding consequences for the movements.

3. *Inner Structure, Style, and Mental Consequences of Youth Movements*

Aside from providing a venue for general social self-criticism, reform, and renewal in times of upheaval and an outlet to the young in generation conflicts, youth movements also play a central role in the lives of the young people who have been involved in them at one time or another (sometimes only a small portion of their age group), shaping their life-course and lifestyle. To come back to K. Mannheim's generation model (1928): the development of an emotional bond with a generation at a young and impressionable age as provided by youth movements is an especially intensive experience which creates long-term mental dispositions that give those involved many points of reference with which to define themselves in historical terms and in relationship to the various age groups living together. Such dispositions may include political and philosophical values as well as principles about how people should treat each other, which lifestyle is appropriate, which esthetic qualities are desirable, and so on. This introduces the question of the inner structure of youth movements.

Youth movements, which sometimes develop out of small activist groups or even secret societies (such as the Italian 'Carbonari' in the early nineteenth century), allow young people looking for orientation and adventure the opportunity to broaden their horizon substantially beyond the everyday confines of their families, schools, and apprenticeships. In their 'conspiracy' of like-minded people their own age, usually motivated by charismatic leaders who are often not much older, members of youth movements develop their own morality and dynamics structured by special codes of communication beyond the blatantly stated goals of the movement. Developing their own language, manners, esthetics, initiation rites, rituals, as well as rules to exclude 'defectors' and 'traitors,' they document the difference between the movement and its social surroundings, particularly with regard to the

styles, morality, and values of the older generation. Elements whose importance is not to be underestimated for most of the movements are short populist manifestos on the one hand, and esthetic elements such as songs and other musical elements on the other, which create an exciting atmosphere at ritual celebrations and give the participants a sense of security. Looking and acting provocative in public, using strategies ranging from clothing, beard, and hair styles to massive shows of militancy, is meant to shock outsiders and strengthen the feeling of belonging among insiders, to the point of producing the feeling of being part of a 'chosen' few. Individuals have been known to initiate spectacular actions in such a hyped-up, extremist, sectarian climate, such as assassinating a social figurehead or mounting a kamikaze attack, which then provides the movement with its heroes or martyrs.

The spread of the mass media in the twentieth century has produced a public arena for youth movements. Television in particular has publicized the aims of youth movements worldwide, touching off direct spin-offs, winning the sympathy of 'fellow travelers,' and generating a wide spectrum of both critical and positive resonance. The activists, in turn, have quite conspicuously played up to the media to promote their own goals. This strategy, however, has frequently backfired. Many of the innovative impulses and manners that were provocative at first have been copied massively in a toned down version, effectively taking the 'sting' out of them and making them generally fashionable. It therefore seems rather simplistic to say programmatically that 'young people make their own history' (Gillis 1981). While youth movements have certainly provided impulses for social change again and again and have been able to get things moving by provoking the older generations, they are, of course, subject to the complex reality of society and its generation structures, specific constellations of which, as we have seen, had to be given before the time was historically 'ripe' for political revolution and/or cultural change.

Research should continue to look for the links between mental dispositions and structural constellations, particularly in the context of the history of generations and in regard to the long-term repercussions in the collective biography of individual age groups involved in youth movements in their respective societies. International comparative research has hitherto been sadly neglected. Papers given in 1990 at the International Historical Congress in Madrid (Fauvel-Rouif 1992) have unfortunately remained a very first step in this direction.

See also: Islam: Middle East; Islamic Fundamentalism; Middle East and North Africa: Sociocultural Aspects; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Sociology of; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Politics; Religion: Nationalism and Identity; Social Movements and Gender

Bibliography

Coutrot A 1971 *Jeunesse et Politique Guide des recherches*. Colin, Paris

Deniger M A et al. 1986 *Jeunesses: des illusions tranquilles*. VLB, Montréal, Canada

Dowe D (ed.) 1986 *Jugendprotest und Generationenkonflikt in Europa im 20*. Jahrhundert Verlag Neue Gesellschaft, Bonn, Germany

Eisenstadt S N 1964 *From Generation to Generation: Age Groups and Social Structure*. Free Press, New York

Ellemers J E 1957 Recente uitingen van de jeugd: een gemeenschappelijke verklaringsgrond? *Sociologische Gids* **4**: 97–110

Fauvel-Rouif D 1992 *La jeunesse et ses mouvements. Influence sur l'évolution des sociétés aux XIXe et XXe siècles*. CNRS, Paris

Feuer L S 1969 *The Conflict of Generations. The Character and Significance of Student Movements*. Heinemann, New York

Fogt H 1982 *Politische Generationen. Empirische Bedeutung und theoretisches Modell*. Westdeutscher Verlag, Opladen, Germany

Gillis J R 1981 *Youth an History: Tradition and Change in European Age Relations, 1770 to the Present*. Academic Press, New York

Heer F 1974 *Revolutions of our Time: Challenge of Youth*. Weidenfeld and Nicolson, London

Kett J F 1977 *Rites of Passage: Adolescence in America, 1790 to the Present*. Basic Books, New York

Laqueur W Z 1962 *Young Germany: A History of German Youth Movement*. Basic Books, London

Mannheim K 1928 Das Problem der Generationen. *Kölner Vierteljahreshefte für Soziologie* **7**: 157–85, 309–30

Nagy L 1984 *250 Millions de Scouts*. Favre, Lausanne, Switzerland

Roseman M 1995 *Generations in Conflict: Youth Revolt and Generation Formation in Germany, 1770–1968*. Cambridge University Press, Cambridge, UK

Speitkamp W 1998 *Jugend in der Neuzeit*. Vandenhoeck & Ruprecht, Göttingen, Germany

Spitzer A B 1973 The historical problem of generations. *American Historical Review* **78**: 1353–85

Springhall J (ed.) 1986 *Coming of Age: Adolescence in Britain, 1860 to 1960*. Gill and Macmillan, Dublin, Ireland

Springhall J 1977 *Youth, Empire and Society: British Youth Movements, 1883–1940*. Croom Helm, London

Wohl R 1980 *The Generation of 1914*. Weidenfeld and Nicolson, London

J. Reulecke

Youth Sports, Psychology of

Youth sports are adult-organized sports programs for children and youth, typically between the ages of 7 and 18 years, which have designated coaches,

organized practices, and scheduled competitions (Gould 1982). There is an estimated 22 million youth in the USA who participate in nonschool youth sports programs. There are equally high percentages of children and youth in Australia, Brazil, Canada, and many Asian and European countries, who participate in youth sport programs through clubs or schools (De Knop et al. 1996). Most of these programs are aimed at mass participation rather than focused on developing elite athletes.

The psychology of youth sports, as an area of scholarly inquiry, is relatively new. Research in this field did not start to proliferate until the 1980s. The research developed out of concern by professionals for the educational value of youth sports participation and the well-being of the children involved. The topics that have been most often addressed on the psychology of youth sports include psychological readiness for youth sports participation, the benefits and detriments of youth sport participation, participant motivation and attrition, and adult involvement in youth sports.

1. *Psychological Readiness for Youth Sport Participation*

Although the question of when children should begin competing in sports appears simplistic on the surface, it is very complex. Readiness, in youth sports, is viewed as the ability of a child to handle successfully all the demands presented in competitive sports. Physicians and educators in the first half of the twentieth century believed that the biological maturation of children was the sole determiner of their readiness to perform sport skills. Contemporary researchers, however, view biological, motoric, and cognitive readiness as interacting with psychological readiness to handle the demands of competitive sport. Factors that have been considered include motivational and cognitive readiness to compete, physical maturation and body size, the knowledge and skill level of the child, and the type of sport. Unfortunately, these factors have not been studied together systematically in the area of youth sports.

1.1 *Motivational and Cognitive Readiness*

The research literature has long shown that most children are not capable of consistently competitive or cooperative behavior until the age of five years (Greenberg 1932). Competitive sport involves setting abstract standards and goals for one's performance and this is usually not possible for children under five years of age. Children also need to have sufficient emotional competence to handle stressful situations inherent in competitive sports. Youth sports participation also requires a number of cognitive skills. These include the child's ability to process a considerable amount of information about the sport, such as rules, skill techniques, and strategies. Most children before the age of five have attention spans that are too short to stay on task in sports games. In addition, children are not accurate judges of their competence until they have developed the ability to understand that winning and losing are the products of interactive causal factors such as physical skill, strategy, preparation, effort, opponent difficulty, and luck. This cognitive ability may not be well developed in children before the age of 12 years (Kaley and Cloutier 1984, Nicholls 1978). Thus, children who are not motivationally and cognitively ready to meet the demands of a competitive sports program may not experience the same success and enjoyment as those who have this readiness and could be likely candidates for sport withdrawal. Gibson (1977) found that Australian boys, aged 10 to 13 years, who participated in adult-style football games, lost interest in participation over the course of the season. Whether modified youth sports programs (e.g., those with simplified rules, less emphasis on competition, shorter practices, equal playing time) would attenuate negative experiences for children who are not yet emotionally and cognitively ready for adult-styled games has not been systematically studied.

1.2 *Physical Maturation Readiness*

Even when children have the motivational and cognitive readiness to learn sports skills and to compete, some may not have the physical maturation to be reasonably successful with other agemates in their chosen sports program. Because children mature at different rates, some 10-year-old children will be more like a 12-year-old child physically and some will be more like an eight-year-old child. When sports programs do not classify children by physical maturity or weight (e.g., weight classifications in wrestling), they usually favor the early maturer (boys in particular). The later maturing boy will usually have less strength, endurance, and skeletal maturity than his average peers. This individual variation in biological maturity has a direct influence on the child's self-perceptions in sport and chances of success. For instance, teams that win the Little League World Series comprise players who are maturationally two to three years older than their chronological age.

1.3 *Knowledge and Skill Level*

Children may be physically and psychologically mature enough to compete in organized sports, but unless they have the requisite skills and some knowledge of the activity, positioning, and game rules, they will have little early success. By six to eight years of age, most fundamental motor skills, such as running, jumping, catching, and striking, are reasonably well

developed in the majority of children (Branta et al. 1984). However, a number of six to eight-year-old children have not yet developed movement patterns mature enough to participate successfully in sport. Whether children with immature movement patterns can still benefit from youth sports participation has not been studied, but may quite well depend on the instructional focus and game modifications in a particular sport program. Different coaches and programs place different emphases on the motor, cognitive, and competitive skills of a sport.

1.4 Type of Sport

Different sports require different levels of physical and cognitive/emotional maturity. For instance, to play baseball/softball effectively, a child must be able to throw the appropriate distance, run, bat, and catch, as well as remember where to make the play for a specific situation. Competitive swimming and running require heavy physiological demands on the cardiovascular and respiratory systems. Each sport has its own unique set of rules, strategies, and skills to be learned. Unfortunately, the criteria of readiness for specific sports have not been studied systematically. Tests of readiness exist in school-related subjects, but none exist for youth sports. Most youth sport programs, however, modify the rules, equipment, and game to accommodate the skill levels and lack of readiness of age-grouped participants. These modifications also need systematic and longitudinal study to determine whether they are of any benefit to those participating.

2. Psychological Benefits and Risks of Youth Sport Participation

For children who are ready to participate in competitive sports, there are a number of potential benefits. Proponents of youth sport programs maintain that youth sports provide an effective medium for psychosocial development such as the development of social skills, cooperativeness, persistence, achievement motivation, emotional control, responsibility, self-confidence and self-esteem, sportsmanship, and leadership. Critics, on the other hand, maintain that youth sports place excessive psychological and physical stress on children, diminish the value of play as a free and spontaneous outlet for expression, and tempt participants to violate moral norms. However, little systematic study has been conducted to determine any causal changes, positively or negatively, from participation in youth sports.

2.1 Benefits

The majority of studies examining the benefits of youth sports participation have been correlational. These studies indicate that regular participation in organized sport activities is associated with increases in physical self-esteem, feelings of physical well-being, and peer acceptance (Brustad 1993). Whether participation in sport results in positive psychological effects for children or whether children who already possess these positive characteristics gravitate toward sport is still uncertain. A few studies have examined the influence of coaching behavior on children's self-esteem and ability perceptions and found that coaches' feedback style can influence the self-esteem and perceived competence of children across a season of competition (Smoll and Smith 1996). However, research on the benefit of youth sports to moral development, sportsmanship, and other socialization outcomes has received no consistent support other than from self-report studies. Self-report studies suggest that the positive outcomes outweigh the negatives.

2.2 Risks

The most frequently cited psychological risks to youth sports participation are competitive stress and anxiety, lowered self-esteem, and reduced moral competency. Participation in youth sports has the potential to be stressful to children because of the uncertainty of competitive outcomes and demands imposed on players. However, most of the research on the anxiety levels of children before and after competition found that children typically report rather low levels of stress, and that these levels of stress are no higher than other evaluative activities typically found in school programs (e.g., academic tests, music competitions). Children who tend to have high competitive trait anxiety, low self-esteem, and low performance expectations, report greater stress than children who are low competitive trait anxious, high in self-esteem, and high in their performance expectations. The experience of stress does not imply that competition is necessarily detrimental to children if there are no consequences beyond a temporary unpleasantness associated with the child's emotional state. However, consequences of stress that lead to loss of sleep, loss of appetite, or dropping out of the sport, are considered detrimental to children.

The question of whether youth sports has an influence on children's moral behavior is difficult to ascertain because the experiences children have in sport are not uniform in terms of the social relationships and interactions that shape behavior. The research in this area has also been correlational in nature. The sport and delinquency research has shown a negative relationship between sport participation and delinquent behavior, however, explanations for the relationship are unclear. Most of the contemporary research on moral reasoning and aggression in sport has been conducted by Brenda Bredemeier and her colleagues (Shields and Bredemeier 1995). This research has shown that reasoning about moral issues in

sport is significantly higher for nonathletes than for athletes but not so for moral problems in nonsport settings. Shields and Bredemeier suggest that athletes use a concept they call 'game reasoning,' a unique type of moral reasoning that is legitimized within the sport context. The divergence between life reasoning and game reasoning becomes greater with age and experience in sport. The genesis, developmental course, and behavioral implications of game reasoning await further research.

3. *Participant Motivation and Attrition*

The importance of participant motivation in youth sport relates to beliefs that participation patterns and experiences in youth sports strongly influence overall active lifestyles in adolescence and adulthood and that attrition is the consequence of undesirable psychological experiences in youth sports. The dropout problem in adolescence is common in many countries, especially among girls. Much of the early research in participation motivation and attrition has been descriptive. Investigators found generally that young athletes participate in sports primarily to have fun, for excitement, to improve their skills, to be with friends and make new friends, and for success. These participation reasons can be categorized into enjoyment, competence, affiliation, and competition motives, respectively. These motives tend to generalize cross-nationally. The most common motives, worldwide, for taking part in youth sports are instrinsic (e.g., for enjoyment and affiliation reasons), rather than competitive (De Knop et al. 1996). Although, children are likely to cite multiple motives, there are few age, gender, experience, and sport-type differences.

Fewer studies have documented the reasons for children's discontinuation in youth sports. Those studies have indicated that competiveness of the program, conflicts with coaches, insufficient success, other interests, and lack of fun are the major reason for discontinuing in a sport. However, the most common motive across studies for withdrawal from a sport had to do with conflicting interests on the part of the child, especially during adolescence. The major problems with these studies are their inability to discern whether a conflict of interests is due to a dissatisfaction with their sport experience or is due to more practical reasons. A second problem is the inconsistency among studies in the definition of the term 'dropout.' Some studies have defined a dropout as one who discontinues in the sport in which he or she was previously participating. However, a number of studies have found that these individuals do not drop out of sport altogether, but rather transfer from one sport to another. Even the small percentage of those who have withdrawn from sport altogether have mostly returned to their sport sometime later (Weiss and Gould 1986). Thus, the motives for withdrawal from sport may be based on practical, age-related, or negative reasons.

More recently, theoretically-based studies have emerged to supplement the initial descriptive research on this topic. The theoretical frameworks most commonly used to explain children's sport involvement and withdrawal are competence motivation theory (Harter 1978), achievement goal-orientation theories (Nicholls 1984), and cognitive evaluation theory (Deci and Ryan 1985). These theories regard self-perceptions of competence as central influences on a child's motivated behavior. Children who perceive themselves to be highly competent in a sport should maintain interest in participating and mastering that sport. In contrast, children who perceive themselves to have low competence in a sport should not maintain persistence and interest. In addition, goal-orientation theories also emphasize that children approach their sport participation with different goals and ways of judging their competence. Children having a task-oriented approach to sport participation hold a self-referenced view to judge their achievement. Children having an ego-oriented approach use social comparison information to judge their success. Task-oriented children should be motivated to participate and remain in sport as long as they have an appropriate level of personal challenge. Ego-oriented children should be motivated to persist in sport only if they perceive that they have more ability than their socially comparative others. These theoretical predictions have been supported by a number of empirical studies. Thus, children tend to persist in sport if they perceive they have the competence to be successful, either in their personal challenges or as compared with others in their program.

Very few studies on participation motivation have been conducted longitudinally or prospectively. It is reasonable to assume that a child's motives for participation will be shaped over time by a number of intrapersonal and contextual factors. Second, retrospective studies on discontinued participation in youth sports do not allow for causal inferences. One prospective study (Ommundsen and Vaglum 1991) examined the relationship of low perceived competence in soccer and social skills with later dropout from boys' soccer over a 16-month period. Low perceived soccer and social competence negatively influenced soccer persistence. More prospective studies are needed looking at additional factors such as type of program, type of sport, personal characteristics of the coaches, and achievement orientations of the children.

4. *Adult Involvement in Youth Sports*

There is general consensus that the benefits, negative experiences, and children's motivation surrounding youth sports are influenced strongly by the adults

involved, coaches and parents in particular. Unfortunately this area of youth research has received the least attention. Again, the majority of the work has been correlational and assessed from the child's perspective. In terms of parental influence, parental support is linked to greater initiation and continued participation in sport, as well as higher levels of intrinsic motivation. Parental pressure is linked to greater sport anxiety and lower enjoyment. Parental judgments of their child's competence are related strongly to the child's self-perceptions.

Much of the research on the influence of coaching behavior on children's self-perceptions and motivation has been conducted by Ronald Smith and Frank Smoll (Smith and Smoll 1996). In one of the few experimental studies on coaching behavior, Smith et al. (1979) showed that young athletes, whose coaches were trained to use a positive approach to coaching (frequent encouragement, positive reinforcement, corrective feedback), had significantly higher self-esteem ratings over the course of a season compared with players whose coaches received no additional training. Low self-esteem children who played for the trained coaches showed the greatest increases in self-esteem.

5. *A Biopsychosocial Perspective on Youth Sports*

Researchers in the psychology of youth sports are calling for a more biopsychosocial approach (Smith and Smoll 1996). For instance, in explaining children's motivation and attrition in sport, the influences of training/conditioning, physical growth, and maturation should be considered in concert with self-perceptions and adult influences. Other questions from this perspective include: What are the long-term biopsychosocial effects of competition in youth sports at various intensity levels? and How do the factors of motor development, biological maturity, physiological function, and psychosocial development interact to determine readiness for and success in sport competition? To examine these questions adequately, youth sports researchers have called for longitudinal, systematic research programs that consider development in the design of studies and that focus on a number of interrelated questions.

See also: Academic Achievement Motivation, Development of; Cognitive Development in Childhood and Adolescence; Developmental Sport Psychology; Sport and Exercise Psychology: Overview; Sport Psychology: Training and Supervision

Bibliography

Branta C, Haubenstricker J, Seefeldt V 1984 Age changes in motor skills during childhood and adolescence. *Exercise and Sport Sciences Reviews* **12**: 275–305

Brustad R J 1993 Youth in sport: Psychological considerations. In: Singer R N, Murphey M, Tennant L K (eds.) *Handbook of Research on Sport Psychology*. Macmillan, New York, pp. 695–717

Cahill B R, Pearl A J (eds.) 1993 *Intensive Participation in Children's Sports*. Human Kinetics, Champaign, IL

Deci E L, Ryan R M 1985 *Intrinsic Motivation and Self-determination in Human Behavior*. Plenum, New York

DeKnop P, Engström L-M, Skirstad B, Weiss M R (eds.) 1996 *Worldwide Trends in Youth Sport*. Human Kinetics, Champaign, IL

Duda J L 1987 Toward a developmental theory of children's motivation in sport. *Journal of Sport and Exercise Psychology* **9**: 130–45

Gibson B 1977 Participation in junior Australian rules football. *Sports Coach* **1**: 15–25

Gould D 1982 Sport psychology in the 1980s: Status, direction, and challenge in youth sports research. *Journal of Sport Psychology* **4**: 203–18

Greenberg P J 1932 Competition in children: an experimental study. *American Journal of Psychology* **44**: 221–48

Harter S 1978 Effectance motivation reconsidered toward a developmental model. *Human Development* **21**: 34–64

Kaley R, Cloutier R 1984 Developmental determinants of self-efficacy predictiveness. *Cognitive Therapy and Research* **8**: 643–56

Nicholls J G 1978 The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child Development* **49**: 800–14

Nicholls J G 1984 Achievement motivation: Conception of ability, subjective experience, task choice, and performance. *Psychological Review* **91**: 328–46

Ommundsen Y, Vaglum P 1991 The influence of low perceived soccer and social competence on later dropout from soccer: A prospective study of young boys. *Scandinavian Journal of Medicine and Science in Sports* **1**: 180–8

Shields D L L, Bredemeier B J L 1995 *Character Development and Physical Activity*. Human Kinetics, Champaign, IL

Smith R E, Smoll F L, Curtis B 1979 Coach effectiveness training: A cognitive-behavioral approach to enhancing relationship skills in youth sports. *Journal of Sport and Exercise Psychology* **1**: 59–75

Smoll F L, Smith R E (eds.) 1996 *Children and Youth in Sport: A Biopsychosocial Perspective*. Brown & Benchmark, Madison

Weiss M R, Gould D 1986 *Sport for Children and Youths*. Human Kinetics, Champaign, IL

D. L. Feltz

Youth to Adulthood across Cultures, Transition from

1. *Transition and Psychological Functioning*

Adolescence is a period in life that begins with biological maturation during which people accomplish a series of developmental tasks raised by the many biological–physical changes, cognitive changes, and

social–emotional changes. It begins roughly at 10 to 13 years and ends—depending on cultural and educational conditions—at 18 to 22 years of age. The following period in life, early adulthood, is a time of establishing personal and economic independence, career development, and of selecting social contexts.

In earlier historical times transitions from youth to adulthood were fairly clearly marked with rites. The purpose of initiation rites in preindustrial societies was to cushion the emotional disruption arising from the transition from one life status to another. In the case of the transition to adulthood, these rites formally ended dependence on parents and served as an introduction to the life of an adult. For example, according to some psychoanalytic perspectives, ceremonies related to sexuality and separation from mothers are necessitated by the conflict between fathers and sons over who will dominate the women. To make clear the adults' supremacy and to ensure the allegiance of the young males, adults typically put the youth through a series of trials at which they were superior to the youth. Most trials included tests of strength, endurance, power, and courage, and—depending on tribe—mutilations. Female adolescents were also exposed to physical pain during the transition to adulthood. For example, genital mutilations were meant to sensitize them to the greater role that sex will play in their lives. These mutilations also made them recognizable to the members of the tribe. The ceremonies often ended in giving the new adult member a new name, usually that of a close ancestor. Seebald (1992) delineates three functions that rites of passage serve: (a) informing the individual of his or her new rights; (b) announcing the new status to the community; and (c) anchoring youth to the group emotionally by instilling a profound sense of belonging to that community.

Entry into adulthood is far more complex for adolescents today, largely due to the need for many more years of formal education and hence prolonged dependence on parents. No specific rituals exist in present Western societies to aid youth through this change. Two general types of transition, intrinsic and role transitions, are distinguished. An intrinsic transition involves new skills and competencies and a qualitative change of the representation of the world. A role transition is exclusively bound to chronological age and social roles, and involves a new status, mainly legitimized by the social environment. Both types of transitions can be described as opening opportunities for further development. If, however, a lack of developmental opportunities occurs on a continual basis and if it occurs within important transition domains, then it is detrimental to the individual.

Psychological stress during transition occurs when people are confronted by demands that exceed their capacities. The fact that something important is at stake, and that the situation is uncontrollable, leads them to perceive that the situation as more of a threat than a challenge. Hence, psychological functioning during transitions consists in an interplay between demands and resources on personal, social, and environmental levels. If the demands exceed a person's resources, the level of functioning decreases. A continuous misfit between demands and resources leads to suboptimal or problematic transition from youth to adulthood.

1.1 Four Principles Underlying Transitions

Four principles are crucial in understanding development in a lifespan perspective, especially the transition from youth to adulthood. First, individual development is carried forward in such a way that the earlier development has implications for a person's pattern of functioning at a later time (Rutter 1994). So the question is, which processes and structures are instrumental in increasing stability in behavior and which of them are instrumental for changing this behavior. In both these cases, two patterns are influential: On the one hand, biological processes, social and societal expectations and individual actions play a role, and on the other hand, a person's objective means.

The accentuation principle is a second component. According to Elder and Caspi (1990) stressful or challenging experiences tend to emphasize and strengthen pre-existing characteristics. Negative critical life events and the lack of opportunities have a consequential negative impact, especially on vulnerable individuals. These events accentuate pre-existing characteristics rather than give rise to new behavior. So personal consistency and behavioral stability is increased.

Third, the increasing stabilization of development across the lifespan is founded in the person's self-concept. As many developmental psychologists have pointed out, experiences have effects in life only through a cognitive transformation. People think about themselves and what happens to themselves. The patterns of individual interpretations and thoughts are condensed in a self-concept (Harter 1996). The self-concept has a clear bias toward assimilation, i.e., people act on the world in congruence with their self-concept. Once the self-concept is established it works as an organizer of subsequent experiences and is rather robust against change.

The fourth principle takes into account that individual development always occurs in a specific environment. The environment is crucial in its influence and in the degree to which it is contingent on the individual's actions. Psychological theories about learning and development were for a long time almost exclusively concerned with this respect, i.e., the contingencies between environment and action. The fundamental learning principle assumes that the contingent positive effects of a person's behavior increase the likelihood of the specific behavior.

1.2 Transition as a Turning Point and in a Lifespan Perspective

It follows from such a conceptualization that development is almost stable and invariant across time (Caprara and Cervone 2000). However, there are turning points which call for new behavior (Caspi and Moffitt 1993). A turning point removes a person from the known structure. On the one hand the individual does not have prior experiences to deal with the new situation, but is forced to act in it. On the other hand, the social environment has not provided the specific individual with experiences, but is inevitably connected to the individual. In this respect, a new learning pattern must be established that has the potential for a persistent change. Rutter (1996) differentiates among three categories of possibly lasting turning point effects: (a) those that discontinue or open opportunities; (b) those that involve a lasting change in the environment; and (c) those with a lasting effect on a person's self-concept. A large amount of empirical evidence supports the characterization of early menarche, schooling, geographical displacement, teenage pregnancy and parenthood, marriage, army experiences, and work as typical examples of turning points. It is noteworthy that a turning point in general provokes a chain of experiences. It is not a singular event, but a chain of events that causes an existing developmental path to change direction.

There is another body of literature which refers to lasting changes during transition, i.e., the lifespan perspective in developmental psychology. Within this framework change and stability across the course of a life has been demonstrated as multidirectional in ontogenetic change and is markedly influenced by sociocultural conditions in a given historical period. Empirical evidence exists for a continuous dynamic interplay between gains and losses in adaptive capacities across the life course. Individual plasticity was here demonstrated as a key concept, i.e., the within-person variability and potential for different forms of behavior and development (Baltes 1987, 1997).

2. Domains of Transition: Developmental Tasks

Havighurst (1948) introduced the concept of developmental tasks in order to describe transitions. Developmental tasks are learning tasks across a lifespan. Successfully solved developmental tasks in the context of real demands enable people to learn new skills and competencies which are necessary for the satisfactory and constructive mastering of life within a given society. Success in solving tasks at earlier stages increases the probability of successfully solving subsequent tasks. People are expected to solve a series of developmental tasks across different life periods in order to be accepted in society. However, not all developmental tasks of a specific life period have to be solved at the same time. Within which time range and in which order developmental tasks are solved is subject to individual differences as well as to cross-cultural variations and sociohistorical shifts.

2.1 Three Sources of Developmental Tasks

2.1.1 Physical maturation. Physical maturation is the basis of developmental tasks and varies little across cultures. For example, in most cultures puberty gives rise to the search for new relationships with same- and opposite-sex peers.

2.1.2 Societal expectations. Societal expectations are the main source of cultural relativity of developmental tasks, typically reflected in age norms. This means that a person is expected to solve a specific developmental task within a certain period. Therefore, the individual's developmental stage can be described as one in which he or she is able to deal with culturally set demands. Cultural norms are also reflected in 'early,' 'on time,' and 'late' development in various transition domains, as, for example, deciding on a job or preparing for a family. Expectations of the timing in role transitions exist for the individual as a private person (e.g., reaching emotional and economic independence from parents, taking responsibility for further decisions) as well as a public person (e.g., age-related norms for different civic rights).

Societal expectations of time ranges in solving developmental tasks are subject to historical change. For example, less than 10 percent of youth attended senior high school in the USA at the end of the nineteenth century compared to almost 90 percent 100 years later. Another example: two-parent farm families were by far the most prevalent family type over decades (Hernandez 1993). A decrease of this family type started in the nineteenth century, and in 1990, less than 5 percent of children lived in two-parent farm families. In parallel, two-parent nonfarm families with a full-time working father and a full-time homemaker mother rapidly increased, resulting in the most prevalent family type between 1910 and 1970. This family type almost disappeared in the 1990s, being replaced by dual-earner families. Since the 1970s, one-parent families have become more prevalent. For the rest of the Western world, the pattern of family structures also changed, but in different time frames compared to the USA. It goes without saying that the family structures in non-Western settings changed, too, but in ways different from those described above. These and other patterns of societal change such as the divorce rate, numbers of births to unmarried women, number of siblings, and education had lasting effects

on the timing of the transition from youth to adulthood (Grob et al. 2001).

2.1.3 Individual goals and values. Individual goals and values are the third source of developmental tasks. In general, people try to organize their actions in congruence with their purposes and goals. There are goals that are mainly the same for all individuals, for example to decide on a career or to reach independence from parents, and others which differ across individuals in obvious and important ways, for example becoming expert in petrography.

2.2 Developmental Tasks in the Transition From Youth to Adulthood

For adolescence, Havighurst proposed nine developmental tasks, i.e. (a) accepting one's physique and a masculine/feminine role; (b) forming new relations with peers of both sexes; (c) achieving emotional independence from parents and other adults; (d) achieving assurance of economic independence; (e) selecting and preparing for an occupation; (f) developing intellectual skills and the concepts necessary for civic competence; (g) desiring and achieving socially responsible behavior; (h) preparing for marriage and family life; and (i) building conscious values in harmony with an adequate scientific view of the world. The eight developmental tasks for young adulthood were (a) deciding on a partner; (b) living with a partner; (c) starting a family; (d) raising children; (e) maintaining a (family) household; (f) starting a professional career; (g) taking societal responsibility; and (h) finding an adequate social network.

The superordinate theme across adolescence and young adulthood, however, is what Erikson (1968) calls identity and intimacy. During adolescence individuals begin to search for who they are, what they are all about, and where they are going in life. As part of this search for an identity, the adolescent experiments with a variety of roles, some sexual, others ideological, and still others vocational. Each of Erikson's eight stages of development across the lifespan centers around a salient and distinct emotional concern stemming from biological pressure from within the individual and from sociocultural expectations. Concerns or conflicts may be resolved in a positive or healthy manner or in a negative or unhealthy way. The characteristics for the transition from youth to young adulthood are described by Erikson with two central themes, i.e., identity versus identity confusion in adolescence and intimacy and isolation in young adulthood.

The empirical evidence on how adolescents manage the transition to adulthood stems among other fields in research on identity. Four identity stages were identified which are: identity diffusion, foreclosure, moratorium, and achievement (Marcia et al. 1992). Individuals who achieved identity, for example, are committed to adult roles by having passed developmental work. Identity statuses are seen in developmental sequences, but not necessarily in the sense that the one is the prerequisite of the other. Identity achievers appear to be healthier than other individuals, and score higher on achievement motivation, moral reasoning, career maturity, and social skills with peers (Kroger 1993). The actual formation of an adult identity does not occur until late adolescence. Identity achievement is primarily reached in the sphere of occupation, whereas commitment in the area of political and ideological identity is often not evident for a considerable proportion of young adults. Thus, the challenge of identity forming in given domains is not resolved at one point in time, but continues to re-emerge again as the individual moves through late adolescence and early adulthood.

3. Further Research

Further research is needed to identify which factors generate discontinuity and continuity in transitions from youth to adulthood within *and* across cultures. Factors that may result in discontinuities in the transitional processes are: (a) variations of social expectations at different historical times due to economic and societal shifts; (b) changes in the specific life contexts and life experiences of adolescents and their families; and (c) the emergence of new developmental tasks at different ages and times as, for example, increased pressure on self-determination and individualization. However, there are also factors that may generate continuity, such as; (a) the genetic disposition of the individual, (b) the continuing influence of early experience, and (c) a contingently rewarding social environment. Determining which parts of the pattern—i.e., biological, historical, social, and individual conditions—contribute to which aspect of successful development in various cultural settings and within cultural subsettings is a task for future longitudinal studies.

See also: Adolescence, Sociology of; Adolescent Behavior: Demographic; Adolescent Development, Theories of; Adolescent Health and Health Behaviors; Adolescent Vulnerability and Psychological Interventions; Adolescent Work and Unemployment; Adolescents: Leisure-time Activities; Adult Development, Psychology of; Adulthood: Developmental Tasks and Critical Life Events; Antisocial Behavior in Childhood and Adolescence; Childhood and Adolescence: Developmental Assets; Cognitive Development in Childhood and Adolescence; Friendship: Development in Childhood and Adolescence; Identity

in Childhood and Adolescence; Life Course in History; Social Competence: Childhood and Adolescence; Socialization in Adolescence; Youth Culture, Anthropology of

Bibliography

Baltes P B 1987 Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. *Developmental Psychologist* **23**: 611–26

Baltes P B 1997 On the incomplete architecture of human ontogeny: selection, optimization and compensation as foundation of developmental theory. *American Psychology* **52**: 366–80

Caspi A, Moffitt T E 1993 When do individual differences matter? A paradoxical theory of personality difference. *Psychological Inquiry* **4**: 247–71

Caprara G V, Cervone D 2000 *Personality: Determinants, Dynamics, and Potentials*. Cambridge University Press, New York

Elder Jr. G H, Caspi A 1990 Studying lives in a changing society: sociological and personological explorations. In: Rabin A I, Zucker R A, Frank S, Emmons R A (eds.) *Studying Persons and Lives*. Springer, New York

Erikson E 1968 *Identity: Youth and Crisis*. Norton, New York

Grob A, Krings F, Bangerter A 2001 Life markers in biographical narratives of people from three cohorts: a life span perspective in its historical context. *Human Development* **44**: 171–90

Harter S 1996 Historical roots of contemporary issues involving the self-concept. In: Bracken B A (ed.) *Handbook of the Self-Concept*. Wiley, New York

Havighurst R J 1948 *Developmental Tasks and Education*. Longman, New York

Hernandez D J 1993 *America's Children: Resources from Family, Government, and the Economy*. Russell Sage Foundation, New York

Kroger J (ed.) 1993 *Discussions on Ego Identity*. Lawrence Erlbaum, Hillsdale, NJ

Marcia J E, Waterman A S, Matteson D R, Archer S L, Orlofsky N 1992 *Ego Identity. A Handbook of Psychosocial Research*. Springer, New York

Rutter M 1994 Continuities, transitions and turning points in development. In: Rutter M, Hay D (eds.) *Development Through Life*. Blackwell, London

Rutter M 1996 Transitions and turning points in developmental psychopathology: as applied to the age span between childhood and mid-adulthood. *International Journal of Behavioral Development* **19**: 603–23

Seebald H 1992 *Adolescence: A Social Psychological Analysis*. Prentice-Hall, Englewood Cliffs, NJ

A. Grob

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7

Z

Zionism

The political structures of Zionism were largely formed through the institutions which emerged in the wake of the first Zionist Congress, convened by Theodor Herzl in Basel in 1897; they were inspired by Herzl's own *Der Judenstraat* (1896) and by earlier tracts like Moses Hess' *Rom und Jerusalem* (1862) and Leon Pinsker's *Autoemancipation* (1882). Yet Zionism's origins are more complex and reflect the impact of modernization and secularization on a traditional Jewish identity which was basically religious. On the one hand, Zionism draws on the traditional Jewish messianic beliefs in a redemptive salvation which would return Jews from their exile to Zion (= Jerusalem) and the Land of Israel. Yet this belief, which greatly enhanced the ability of Jewish individuals and communities to survive two millenia of exile and persecutions, was at its base religious and looked for divine intercession in the historical process, and not for human, practical action; it was thus, basically, passive. It also never led, with some few yet notable exceptions, to concrete steps in the here and now aimed at bringing Jews to the Holy Land or to the establishment of a political commonwealth. Modern, racist antisemitism, which grew towards the end of the nineteenth century in many European societies, did indeed put in question the hopes for the disappearance of anti-Jewish prejudices with the advent of the Enlightenment and secularization, yet did not by itself lead to the emergence of a political movement among the Jewish populations in Europe: those fleeing antisemitism, especially in the Czarist Empire in the 1880s, looked for emigration to the West—mainly to North and South America, but also to England and South Africa.

In other words, neither traditional Jewish religiosity nor modern antisemitism can adequately explain the emergence of Zionism. It can be understood only in the context of modernity—political, social, and economic—which totally transformed Jewish–Christian relations in Europe, which were based on a delicate, and not always stable but culturally deeply anchored, balance which enabled Jews to find a niche for themselves in Christian Europe. For despite theological condemnation, persecution, discrimination, and occasional expulsions, Christian Europe was able to tolerate the existence of a small Jewish minority in its midst. Provided they accepted their marginality and subjugation, as well as their exclusion from public spheres of social life, Jews were allowed a relative free exercise of their religion and could find an economic sphere for highly circumscribed yet permitted activities.

The emergence of the modern nation-state in the wake of the French Revolution totally changed the parameters of this relationship. On one hand, the modern nation-state was based on the ideas of popular sovereignty and equality among all citizens; on the other, it drew on notions of common descent and put at its symbolic core a commonalty of language, history, myths, and institutions: this confronted both Jews and non-Jews with new challenges. It was not anymore a question of religious tolerance or even equality before the law, but the question of identity and belonging: thus liberalism and nationalism posed a new set of questions. While the ideas of the French Revolution emancipated the Jews as individuals and made them, for the first time, into equal citizens and thus broke down the physical and spiritual barriers of the ghetto, new questions of identity arose for which there were no normative answers in traditional, orthodox Judaism. How should Jews behave with regard to Jewish holidays, customs and laws of marriage, preservation of cultural traditions within a society which, for the first time, is now open to them? At the same time, the emergence of modern nationalism in central and eastern Europe posed equally new questions to Jewish individuals and communities: should they view their identity in purely religious or confessional terms, give up their ethnic and cultural identity, and strive to become 'Germans (or Magyars) of the Jewish confession'? And if they choose this option—as many of them did—would their surrounding society, deeply imbued with ideas of national continuity, history, and cultural identity, really accept them as equals and cease to view them as aliens on their national territory? These issues became especially acute in the areas of the Czarist and Habsburg Empires, where most European Jews lived in the nineteenth century: these empires were also challenged by nationalist movements, and Jews living on the fault lines of these various movements were further torn apart by competing claims: should Jews living in Prague identify with German or Czech nationalism, those living in Budapest identify with German or Magyar culture, or those living in Warsaw support Polish nationalism or Russian rule?

Precisely because secularized, emancipated Jews were looking for an anchoring in their surrounding society, when most of those societies in Central and

Eastern Europe were riven with competing claims for allegiance and identity, it became clear that modernity and even equality do not supply answers to questions of this sort: should one send one's child to German or Polish or Hungarian or Czech schools? Whatever the choice it sometimes raised more questions than it answered. The emergence of the Hebrew *Haskala* (Enlightenment) in Central and Eastern Europe raised a new challange. Under the impact of Herderian ideas, Hebrew became once again a vehicle for literary and cultural creativity beyond its traditional use as a language of prayer and rabbinical discourse. The study of the history of the Jewish people appeared for the first time among Jewish secularized intellectuals, with scholars like the German-Jewish historian Heinrich Graetz following a historiography inspired by Ranke and Michelet making Jewish people much more aware of their own history than the mostly atemporal religious discourse tended to do. This development led to a modern, literary-historical and national new interpretation of the Bible which was now seen by nonpracticing Jews not as a religious text but as a repository of poetic, ethical, literary, and historical sources.

The Haskala created a new Jewish intelligentisia—relatively free from religious observance, yet deeply aware of its history, cultural heritage, and specific contribution to world history: thus the Exodus from Egypt was turned from a story of divine redemption to the first revolt of slaves in history and the Maccabeans became the first fighters for religious freedom. All this was a specific Jewish response both to ideas of political liberalism and national romanticism emerging from the complex heritage of the French Revolution.

This was the context within which the Zionist Organization, founded by Herzl, flourished, developed, and found its adherents. Initially Herzl believed that the acquisition of a Jewish homeland in Palestine could be achieved through diplomatic means and the instrumentality of the European Great Powers bearing upon the Ottoman Empire to grant his organization a Charter for Jewish settlement. It became clear, however, that this was not to be, and that without a demographic change in the population structure of Palestine, no real results could be achieved. The few and sporadic Jewish settlements established in Palestine prior to the emergence of political Zionism became the nucleus of a much wider and organized attempt to create a social, economic, and political Jewish infrastructure in Palestine. After World War I, Palestine became a British Mandate under the League of Nations, with a vague and ambiguous British commitment to support 'a Jewish national home' in the country (the 1917 Balfour Declaration). Between 1918 and 1939 the Jewish population in Palestine grew, through immigration, from 60,000 to almost 600,000, accompanied by a sustained effort at creating an agricultural Jewish class, representative institutions, and relatively autonomous economic sector. Zionism always viewed it as an 'anomaly' of Jewish life that Jews lived not only in exile and under conditions of political powerlessness, but also that their social and economic structure differed widely from that of the populations in whose midst they were living: because Jews were not allowed to own land, they became a mainly urban population; because certain economic activities were barred to them, they were limited to a narrow range of financial or petty trading activities. One of the first aims of Zionism was to change this so-called 'inverted pyramid': hence the emphasis on settling on the land, supported by strong social solidarity and national institutions: it was this which gave the Zionist Left, greatly inspired by Russian socialist revolutionary and populist (*Narodnik*) ideas, its great impetus towards creating new forms of communal living (kibbutzim, moshavim) and the emergence of the Labour Federation (Histadruth) as an overarching 'Community of Labor' combing trade union functions with aims of social transformation and economic development.

As a national movement anchored in its relationship to what it considered the historical homeland of the Jewish people in Palestine, the Zionist movement found itself confronted by the emergence of Arab nationalism, both in Palestine and in the surrounding countries: this at a time when with the demise of the Ottoman Empire in World War I, Arab Nationalism became a major force among Arabic-speakers in the Middle East. Arabs in Palestine saw in Zionism and the influx of new immigrants a danger to their status in Palestine, which in Arab national thought was seen as integral a part of the Arab homeland as the surrounding countries of Syria, Lebanon, Egypt, etc.

The fear of becoming a minority in their own land drove Palestinian Arabs in 1921 and 1929 to anti-Jewish riots and then, during 1936–9, to the Arab Revolt against the British which was aimed at stopping Jewish immigration at a time when the tides of antisemitism in Europe made Palestine sometimes into the only option available to Jews fleeing Nazi persecution. After 1945, Jewish resistance to the British—which still did not allow free Jewish immigration to Palestine, even after the Holocaust—caused the issue to be brought to the United Nations which in 1947 proposed the partition of British Mandatory Palestine into two states, one Jewish and one Arab. While the Jewish community in Palestine and the Zionist organization accepted the idea of partition, it was totally rejected by the Arabs of Palestine and by all member countries of the Arab League. This led to the War of 1948, when the Arabs of Palestine and Arabs armies that came to their help were defeated in their attempt to prevent the establishment of a Jewish State in a partitioned Palestine and the Palestinian refugee problem was created.

Arab countries continued to reject the acceptance and legitimacy of Israel, and this led to the wars of

1956, 1967, and 1973; Israeli occupation in the war of 1967 of the West Bank and Gaza caused the Palestinian intifada against Israeli military rule. President Sadat's visit to Jerusalem in 1977 led to the first Arab–Israeli peace treaty, in which Egypt accepted the legitimacy of Israel in return for the return of Egyptian territories occupied by Israel. Further agreements with the Palestinian Liberation Organization (PLO) in 1993 and with Jordan in 1994 paved the way for a wider reconciliation and for the possibility of an eventual emergence of a Palestinian state living side by side with Israel—basically the 1947 UN formula.

On the ideological level, Zionism encompasses a variety of tendencies—from rightwing nationalistic Zionism (mainly represented by the Likud Party) through religious Zionism to a variety of socialist and social-democratic parties (now mainly represented in the Labour Party). The mass immigration of almost a million Jews from the lands of the former Soviet Union since the late 1980s, as well as the immigration of almost 60,000 Jews from Ethiopia during the same decade, suggest the basic dilemmas out of which Zionism grew are still a reality in some regions of the world. In the democratic West, on the other hand, an identification with Israel is the only hallmark of a Zionist Jewish identity. In Israel itself, relative security and the feeling that the main tasks of Zionism have been fulfilled, albeit under conditions of warfare and siege, now give rise to a less nationally-centered public discourse, sometimes called post-Zionist, which suggests an attempt to substitute the national ideology by one focused more on ideas of civil society. The more the peace process succeeds, the more these new and not always coherent ideas will become much more central to the public discourse of the country.

See also: Immigration; Middle East and North Africa: Sociocultural Aspects; Migration out of Europe; Religion: Mobilization and Power; Religion: Nationalism and Identity; Religious Nationalism and the Secular State: Cultural Concerns

Bibliography

Avineri S 1981 *The Making Modern Zionism: The Intellectual Origins of the Jewish State*. Basic Books, New York

Berkowitz M 1998 *Zionist Culture*. The University of North Carolina Press, Chapel Hill, NC and London

Halpern B, Reinharz J 1998 *Zionism and the Creation of a New Society*. Oxford University Press, New York

Hertzberg A 1959 *The Zionist Idea*. Atheneum, New York

Katz J 1986 *Jewish Emancipation and Self-Emancipation*. Jewish Publication Society, Philadelphia, PA

Shimoni G 1995 *The Zionist Ideology*. University Press of New England, Hanover, NH

S. Avineri

Zionism, History of

1. Historiographic Premises

Israel is one of those modern societies whose institutions were clearly shaped by an ideological movement. Thus, Israel is similar to other new states whose political institutions directly derive from the nationalist movement that won their independence. But, in addition, a wide variety of social, economic, and cultural institutions of the new Jewish state were initially developed by the Zionist organization and its associated bodies. In these respects, Zionism had an impact unusual among nationalist movements. It must rather be compared with social revolutionary and radical reform movements. Anthony D. Smith contrasts 'ethnocentric nationalism' in which the dimensions of 'power and culture inhere in [a single] cultural group' and 'polycentric nationalism' in which the group's chief desire is to 'join the "family of nations" becoming a nation like all others in a condition of dignified equality.' He further refines this differential with respect to 'diaspora nationalism' citing Garveyism, Zionism, Lebanese, Liberians, Greeks, and Armenians as 'classic cases' (Smith 1983).

A national liberation movement typically arises among a people oppressed by foreign rulers in its own land. Elie Kedourie argues that this conceptual approach actually has roots in the French revolution and the toppling of the *ancien regime*. (See Kedourie's classic essay, especially his comments on 'Politics in a New Style' in *Nationalism* (Kedourie 1985).) To this generalization, the Zionist movement and the rise of Israel are exceptions. Jews did not face the common nationalist situation of being oppressed in their own land by a foreign ruler. Rather, the Jewish people lived as a minority in numerous countries throughout the world, among many different ruling nations and under different regimes. The oppressions they suffered were different in kind, varying from intense to imperceptible, and no single foe could plausibly be held responsible for all Jewish frustrations.

Hence, when Jewish nationalism arose, rival ideologies such as Jewish liberalism, socialism, autonomism, and anti-Zionism did not succumb to it (see Laqueur 1992). They engaged instead in a sharp polemic through which their own positions became more fully elaborated and well defined. Nor did the whole nationalist movement concentrate upon the primary objective of national sovereignty. Numerous nationalist factions arose, each insisting that its goal was most important and all others secondary (see Halpern 1969, Chap. 1, Shimoni 1995).

2. Origins of Zionism

In the eighteenth and nineteenth centuries, the Jewish communities of Europe were traditional in character. This period witnessed a complex phenomenon known

as Jewish emancipation: the process by which the Jews, especially those in Western and Central Europe, were propelled towards new social and economic opportunities and became engaged in gentile society. The prospect of emancipation gave rise to the ideology of Jewish liberalism, which sought to redefine the Jews as a religious community, aware of and responsive to the norms of gentile society. Jewish liberalism was especially strong in Western Europe—England, France, and Germany—where the numerical insignificance of the Jewish population, the speed of social and economic modernization, and the relative ease of legal emancipation all appeared to legitimize the pro-emancipation viewpoint.

While Jewish liberalism dominated the thought of the emancipated Western Jewish leadership, there was a broad body of Jewish public opinion that objected to the loss of a specific Jewish identity. Both traditionalists and defectors from the camp of Jewish liberalism shared this perspective. The typical opponents of Jewish liberalism were born and educated in areas outside the Western liberal mainstream, in Eastern and South Eastern Europe, where traditional values held strong, antisemitism was a palpable reality, and Jewish integration did not appear to be a viable option.

Although Zionism was, in part, a response to modern antisemitism, a number of factors forced the retention of a residual Jewish communal consciousness and national identity. As integral to Zionism as its negative reaction to antisemitism, was its positive assertion of belief in messianic prophecy and the historic destiny of the Jewish people (see Schweid 1985, Chaps 1, 3). Two leading precursors of Zionism, Rabbi Zvi Hirsch Kalischer (1795–1874) and Rabbi Judah Solomon Hai Alkalai (1798–1878), were moved to create a Jewish settlement in Palestine by sociopolitical forces and a deep religious conviction that the time for the biblical injunction of the 'ingathering of the exiles' had arrived—the prophecy was to be literally acted upon.

Similarly, the socialist Moses Hess (1812–1875) rediscovered Jewish nationalism after a distinguished radical career in Central Europe. His book *Rome and Jerusalem* (1862), was an eloquent appeal for the reassertion of Jewish national identity. The work of Hess, and of his contemporary, the Hebrew publicist and editor Perez Smolenskin (1840–1885), demonstrates that even before the Russian Jewish crisis of 1881–1882, Western Jewish liberalism was under attack from Jews with a nationalist perspective.

The term Zionism was probably coined by Nathan Birnbaum (1864–1937), an Austrian leader of Hovevei Zion [Lovers of Zion], a loose federation of groups devoted to the promotion of national resettlement of the Jews in their ancestral homeland. The Hovevei Zion viewed resettlement as the solution to the Jewish question, which it variously defined as the problem of antisemitism or the problem of Judaism in a modern, secular world. Significantly, the term Zionism reflects both traditional sentiment (the longing for Zion) and a modern political orientation.

3. *Ideological and Political Evolution*

The assassination of Tsar Alexander II by revolutionaries, the pogroms of 1881–1882 and 1903–1905, and the ensuing the Russian Revolution provoked despair and panic and eventually prompted the massive flight of East European Jewry. Many young Jews considered options such as emigration to and the colonization of Palestine. An immediate rivalry developed between the Amerikantsy, represented by such groups as Am Olam [Eternal People] who saw the United States as the obvious safe haven, and the Palestintsy who desired resettlement in Palestine.

In 1882 scattered groups of Palestine-minded students united to found the Bilu movement, a secular Zionist group dedicated to creating exemplary rural colonies in the Land of Israel. Their most celebrated publicist was the young journalist Moshe Leib Lilienblum (1843–1910). Although the movement failed to generate mass migration to Palestine, it did lay much of the ideological groundwork for Russian Jewish pioneers in this period (Frankel 1984, Chap. 2).

The movement for a new exodus received ideological justification from Leon Pinsker (1821–1891), who in 1882 published an influential pamphlet entitled Auto-Emancipation. Pinsker headed the Russian movement known as Hibbat Zion [Lovers of Zion] from 1884 until his death. The efforts of the Russian Zionists were largely ineffectual. They managed to establish a few agricultural colonies in Palestine, but many of them became dependent on the largesse of the important French Jewish philanthropist Baron Edmond de Rothschild (1845–1934). Though deeply committed to the development of the Land of Israel, Rothschild was not a Zionist (Halpern and Reinharz 2000, Chap. 4).

There was a significant growth in the Jewish population of Palestine in the decades before World War I, largely due to Zionist immigration. Estimates for the beginning of the nineteenth century indicate that only about 7,000 Jews lived in Palestine, about 2.4 percent of the whole population, and mostly concentrated in the four Holy Cities: Jerusalem, Tiberias, Safed, and Hebron. At the end of the century the number of Jews had grown to about 43,000, about 8.1 percent of the total population. Well over 90 percent of these Jews lived in the cities. 5,000 Jews lived in Jaffa, half of the total population of the city. Furthermore, by 1890 Jews were already a majority of 60 percent of the population of Jerusalem, and their proportion was due to grow even more in the following years. Before World War I they comprised half of the Jewish population of the country, but these were

mostly Orthodox Jews who had little to do with the developing Zionist movement (Friesel 1972, 1990).

Zionist ideas also found fertile ground on American soil (Raider 1998). As early as the beginning of the nineteenth century, prominent American Jews expressed the view that the Jewish people should be allowed to return to the Land of Israel. Mordecai M. Noah (1785–1851), a well known publicist and former consul in Turkey, following an ill-fated attempt to found a Jewish colony called 'Ararat' near Buffalo, New York in 1825, turned to Palestine as an ultimate haven for the Jews. So, too, did Warder Cresson (1798–1860), a former American diplomat who converted to Judaism and established a settlement near Jerusalem.

4. *Modern Political Zionism*

The fortunes of the fledgling Zionist movement changed dramatically with the appearance of Theodor Herzl (1860–1904), the father of modern political Zionism. Herzl was an Austrian Jewish writer and journalist and, until the Dreyfus Affair of 1894, a typical, assimilated, middle-class Jew. Shocked to the core by the virulent display of antisemitism surrounding the Dreyfus Affair, Herzl subsequently turned his efforts to the implementation of a Zionist solution to the so-called 'Jewish problem.'

In 1896, after unsuccessful attempts to enlist the support of Franco–Jewish philanthropists, Herzl sought a wider audience with the publication of *Der Judenstaat* (The Jews' State). In 1897 he launched a Zionist weekly called *Die Welt* (The World) and presided over the First Zionist Congress in Basel, Switzerland. A total of six congresses were held between 1897 and Herzl's death in 1904, and they created the organizations and institutions of the Zionist movement: the World Zionist Organization, the Jewish Colonial Trust, and the Jewish National Fund.

Herzl pursued his goals through diplomacy aimed at the rulers of European states and the Ottoman sultan. In 1903, the Zionist movement underwent a crisis when the British government offered Herzl a tract of land in East Africa for Jewish colonization. Herzl recommended the acceptance of this territory to his followers as a temporary 'safe haven' for the Jews. This 'Uganda Project,' as it came to be known, ultimately failed, foundering on the hostility of the Russian Zionists led by Menahem Mendel Ussishkin (1863–1941). In the event, the militant rejection of Herzl's flirtation with territorialism assured the Zionist movement's enduring commitment to settlement in Palestine alone (Vital 1975, 1982).

Some years after Herzl's premature death in 1904, the leadership of the movement was captured by the 'practical Zionists' who subordinated the long-range, diplomatic and social goals of Zionism to the task of building up the Jewish settlement in Palestine. Another important influence was that of Ahad Haam (the pseudonym of Asher Zvi Ginsburg), a Russian Jew who emerged as the foremost critic of both Hibbat Zion and Herzlian Zionism. Ahad Haam (1856–1927) rejected the concept of Zionism as a mass movement, arguing instead for a concept generally characterized as 'spiritual Zionism.' In his view, the Jewish people required cultural revival and modernization, objectives best carried out by a small elite based in Palestine. This was a clear rejection of the idea of immediate, mass resettlement. Although never part of the Zionist mainstream, the writings of Ahad Haam had a profound influence on many of the younger leaders of the Zionist movement (Zipperstein 1993).

The differences between Herzl and Ahad Haam reflected the differences between West and East European Jewry, stemming from their contrasting political and cultural situations. The multiethnic character of some of the states and political parties of Eastern Europe theoretically presented the Jews with the possibility of a secular national culture within the context of modern society. However, East European Jewry was forced to face the brutal nature of antisemitism in a much more violent fashion than West European Jewry, and because of the exigency of this problem East European Zionists were pressed to develop a program of immediate, practical solutions that involved alternatives that did not require the restoration of sovereignty in Palestine. In contradistinction to East European Zionism, Western Zionism until the rise of the Nazis was largely a philanthropic movement devoted to the welfare of persecuted Jews in Eastern Europe. It did, however, contribute significantly to the intellectual and ideological fermentation of Zionism in the East and the West.

In the decades that bridged the nineteenth and twentieth centuries, socialist ideas merged with East European Zionism to produce a number of socialist Zionist parties, some of them with a Marxist orientation, some not. The left-wing Zionist groups produced many of the members of the Second Aliyah, the wave of Jewish immigration to Palestine between 1904 and 1914 that had a profound impact on Zionist politics and ideology (Halpern and Reinharz 2000).

5. *Zionist Policy and the British Mandatory*

The outbreak of World War I opened up new possibilities for the Zionist movement. Some of the leading Zionists, among them Chaim Weizmann (1874–1952) and Vladimir (Zeev) Jabotinsky (1880–1940), decided that the future of the movement laid in the Entente camp. Their hopes were indeed realized with the great political success of the Balfour Declaration in 1917, wherein the British government expressed its commitment to the establishment of a Jewish national home in Palestine. The Balfour Declaration, which generated tremendous excitement among diaspora Jews, was received by the Arab

community in Palestine with profound misgivings. But the hopes and fears the Declaration stirred respectively in each camp were equally exaggerated.

After World War I, when the British Mandatory regime's policy unfolded, it became clear that it had neither the will nor the intention to establish a Jewish national home in Palestine in the sense envisaged by the Zionist Movement. Instead, during the interwar years, the British government issued a series of White Papers stressing that the Balfour Declaration had not intended to create a sovereign Jewish state in Palestine. The principle of the 'double obligation' to Jews and to Arabs in equal measure, became the cornerstone of British policy in Palestine in the years from 1922 to 1939. Unappeased, however, Arab political leaders continued to view Zionism as an extension of Western imperial aims and Arab opposition to the Jewish community in Palestine intensified (Stein 1985).

In the 1920s and 1930s, Palestine's Jewish society-in-the-making (known in Hebrew as the *Yishuv*), grew gradually in numbers and developed a political structure, communal institutions, and an economic foundation. There were 85,200 Jews in Palestine in 1922 (11.1 percent of the total population), and 175,100 in 1931 (17 percent of the total population). Due mainly to significant increases in the 1930s, the number of Jews reached in 1947 to 630,000, close to one-third of the total population. In 1946, over three-quarters of the Jewish population lived in cities. With over 180,000 inhabitants, Tel Aviv became the major Jewish center of the country (Friesel 1990, Hersch 1948).

Chaim Weizmann, the architect of the Balfour Declaration, emerged as the leader of the Zionist movement in the years after World War I (Reinharz 1993). Despite Weizmann's secure diplomatic standing, the new Zionist hierarchy led to a somewhat anomalous situation. On one hand, the Zionists in the cities and the settlements in Palestine, many of them religious, others secular, an influential segment imbued with a particular socialist ideological agenda, lived all according to the specific dynamics dictated by the life in Palestine. On the other hand, the individuals who retained ultimate political control of the Zionist movement lived in a totally different milieu thousands of miles away. The ambivalent attitude of the British authorities toward the Yishuv and Zionism only accentuated this division.

The geographic distance separating the Zionist leadership in London from the movement in Palestine and elsewhere caused misunderstandings, ideological quarrels, and eventually a power struggle in world Zionist politics. A rift emerged between the traditional European-based Zionist leaders and the stewards of the movement in the United States. Louis D. Brandeis (1856–1941), an associate justice of the United States Supreme Court and head of the American Zionist movement, believed that the political era of Zionism had ended with the Balfour Declaration. By contrast, Chaim Weizmann believed that the political struggle was far from over, and he continued to view Zionism, as officially represented by the Jewish Agency for Palestine, as representing the national will of the Jewish people. In fact, Brandeis and Weizmann did not differ significantly in their programmatic positions. However, they did differ over a principle of fundamental importance: the notion of exile and the Land of Israel as the authentic home of the Jewish people. Weizmann eventually emerged supreme in this clash of ideologies, defeating Brandeis at a convention of the American Zionists held in Cleveland, Ohio in 1921 (Halpern 1987).

6. *Labor Zionism and Palestine*

In the 1930s the Labor Zionist movement rose from its minority status in the World Zionist Organization to a position of political dominance. Despite Labor's initially meager size, the movement had always played a critically important role in Zionist efforts aimed at immigration, land acquisition, and permanent Jewish settlement, particularly the process of organizing and training young Zionist pioneers abroad. Notwithstanding the British Mandatory's restrictive policies, the Arab community's steadfast and even violent opposition to Jewish settlement, and the economic fragility of the region as a whole, the Labor movement gradually developed a sound socioeconomic, country-wide infrastructure.

At the heart of the Labor enterprise stood the Histadrut, the General Federation of Jewish Labor, created in 1920 by a conference of socialist Zionist parties in Palestine. The Histadrut commanded a broad network of agricultural settlements, educational institutions, trade unions, self-defense groups, a workers' press and a workers' health organization. This web of social, economic, cultural, and political interests consolidated Labor's position in the World Zionist Organization and enabled the Yishuv to cope with different crises through broad measures of cooperative organization.

Labor never attained a political majority in the Jewish community in Palestine, and had always to form coalitions with other parties, especially with the Mizrahi national-religious party and the middle-class liberals, known as the Progressives. The ensuing political experience of power-sharing, which continued for most of the time of the British Mandate in Palestine, created patterns of political collaboration that were to continue later on, after the establishment of the State of Israel (Halpern and Reinharz 2000, Medding 1990).

Labor Zionism's most serious political challenge came from Revisionists, the right-wing dissident movement created by Vladimir Jabotinsky in 1925. Jabotinsky originally created the Revisionist party to counteract what he viewed as the corruption of Herzlian political Zionism under Chaim Weizmann's

leadership. He also emphasized the primacy of private capital in developing the Yishuv and insisted on maximal nationalist and antisocialist principles.

A critical juncture factor in determining the Yishuv's future was reached in the late 1920s and early 1930s when a coalition of Zionist and non-Zionist forces united behind the objective of attaining a secure Jewish national home in Palestine. To this end, the World Zionist Organization initiated a series of negotiations with organized Jewish bodies in various countries, which culminated in 1929 in the creation of what came to be known as the Expanded Jewish Agency and was regarded as an expression of the growing interest in Palestine of the entire Jewish people (Halpern 1969, Chap. 6).

The Arab riots of 1929, which resulted in the massacre of Jews in Hebron and other places, dramatically changed the Yishuv's internal debate over the issues of Jewish sovereignty and security. The pretext for the riots was a clash over conflicting Arab and Jewish rights in Jerusalem. In fact, the tension between Arabs and Jews had been steadily mounting since the previous wave of riots in 1921. Shortly thereafter, the uncertainty of the Mandatory's short- and long-range goals was further compounded by the eruption again of Arab riots in 1936–1939. The latter initially prompted the Royal Peel Commission of 1937 to recommend the partition of Palestine and the creation of a small Jewish state—a plan reluctantly adopted by the Zionists and rejected by the Arabs.

But the stillborn Peel proposal ultimately led to a complete reappraisal of Mandatory policy and a retreat from the Balfour Declaration-inspired policy. In its place came the MacDonald White Paper of 1939, which acquiesced to Arab demands. The White Paper established a limit on Jewish immigration of 75,000 between 1939 and 1944 (when immigration would cease), curtailed drastically new purchases of land by Jews, and established a ten-year transition period after which Palestine should become an independent country, meaning, an Arab state with a Jewish minority of about 30 percent of the population. The Jewish Agency and Zionists everywhere believed this to be a clear breach of past British pledges. In the face of this turnabout, exacerbated by the onset of World War II, David Ben-Gurion (1886–1973), the leader of the Zionist movement after 1935, declared a new Zionist policy agenda: 'We must support the [British] army as though there were no White Paper, and fight the White Paper as though there were no war' (Shapira 1992, Chap. 7).

7. *World War II, the Holocaust, and Combative Zionism*

During the course of World War II, Britain received assistance from the Zionist organization, including the Haganah, the military arm of the Jewish Agency. The British military received assistance from the Palmah (Hebrew acronym for *plugot mahaz*, shock troops), for example, in the campaign against the Vichy French authorities in Syria in 1941. The Palmah had originally been formed to help defend Palestine against a possible German or Italian invasion of Palestine.

While the British Mandatory was ambivalent about the Haganah and the Palmah, it was plainly negative toward the Revisionist fighting force known as the Irgun Zvai Leumi (National Military Organization, also known by the acronym Ezel). Created in the midst of the 1936–1939 riots, the Irgun was dedicated to undertaking retaliatory action against Arab terrorists. When Menahem Begin (1913–1994) assumed command of the Irgun in 1943, he was convinced that the only effective strategy for gaining Jewish control of Palestine was to exert so much pressure through selective terrorism—such as the bombing in 1946 of the King David Hotel in Jerusalem—that the British would eventually be forced to choose between total repression or complete withdrawal.

Zionist terrorism was unacceptable to the leaders of Jewish Agency. After the assassination in Cairo, in 1944, of Lord Moyne, the British Minister for the Middle East, by members of another Zionist terrorist organization, the Stern Group or Lohamei Herut Israel [Freedom Fighters of Israel, also known by the acronym Lehi), the Haganah and the Palmah were given orders to 'neutralize' the Irgun. This led to the so-called 'Open Season' during which Haganah and Palmah operatives interned and handed over right-wing Zionist activists to the British, causing considerable bitterness in the Yishuv. As this clash reveals, global events at the height of World War II brought about a steady radicalization of Zionist ideology and activity in the Yishuv.

The outbreak of World War II resulted in the complete dislocation of Jewish life on the European continent that had been the backbone of much of Jewish life since the eighteenth century. The constitutional pattern of the World Zionist Organization and the Expanded Jewish Agency became unfeasible and in practice responsibility for the Zionist enterprise fell to the Yishuv and American Jews. In May 1942, a wartime conference was held in lieu of a Zionist Congress at the Biltmore Hotel in New York City. The Biltmore Conference symbolized the strengthening of the position of Zionist leaders in the Yishuv and America and united most of world Zionism around the call for a Jewish state in Palestine (Laqueur 1992). This juncture also signaled the displacement of the Zionist movement's veteran leadership, notably Chaim Weizmann and Stephen S. Wise (1874–1949), and the ascendance onto the world stage of a younger generation led by David Ben-Gurion in Palestine and Abba Hillel Silver (1893–1963) in America (Teveth 1987, Raider et al. 1997).

The election of the British Labour party in 1945 raised expectations that Palestine's partition and

Jewish independence would soon follow. These hopes were dashed, however, when British Foreign Secretary Ernest Bevin announced, in effect, that the White Paper policy of 1939 would continue and that Britain was committed to a Jewish 'home' in Palestine, but not a Jewish state. Set against the catastrophic tragedies that befell European Jewry during the war—but, in fact, before the full magnitude of European Jewish destruction was known—British policy was received with mixture of anger and alarm by Zionists of all political persuasions.

8. *Campaign for Jewish Statehood*

The years immediately following World War II were characterized by turmoil in Palestine, intensive Zionist activity worldwide, and ongoing diplomatic initiatives in the Western capitals, especially in London, Washington, and New York. The Jewish Agency spearheaded the postwar campaign for the establishment of the Jewish state. The Agency used three interrelated approaches to the attainment of this goal: first, political negotiation, aimed at securing international recognition for a Jewish state; second, the practical work of immigration, land development, and the building of a strong autonomous Jewish community; and third, the organization of clandestine or illegal Jewish immigration to Palestine during the British Mandate known in Hebrew as *Aliyah Bet* (Ofer 1990).

The latter reached a climax in 1947 against the backdrop of overflowing displaced persons camps in Cyprus when the British decided on a policy of returning illegal Jewish immigrants to their original ports of departure. The most notorious example was the Exodus-1947, a ship carrying nearly 5,000 Jewish refugees. Intercepted off the Palestine coast, the Exodus was compelled to return to Hamburg after a pitiful ordeal in which the British forced the passengers off the ship and beat those who resisted with clubs and hoses—all in the glare of the world's media. The affair demonstrated Britain's inability to deal with the ever-increasing flow of illegal Jewish immigrants and hastened its decision to relinquish the Mandate. It also highlighted the Zionist leadership's postwar strategy for winning the battle for international public opinion. The boatloads of beleaguered European Jewish refugees and determined Zionist activists gradually eroded sympathy for British policy and underscored the urgency of postwar Jewish reconstruction (Laqueur 1992, Sachar 1976).

On November 29, 1947 the United Nations adopted a resolution calling for the establishment of a Jewish state in Palestine. The sudden collapse of the Mandate, the persisting hostility of the Arab states, and particularly the absence of and ensuing conflict with the Palestinian Arabs abruptly and totally altered the conditions under which the policy of the Jewish state would thenceforth have to be formulated. When the state was formally proclaimed on May 14, 1948, the major functions of the World Zionist Organization and the Jewish Agency were immediately assumed by the nascent Israeli government.

The Zionist movement before the rise of Israel proclaimed, in addition to the goal of national independence, the following objectives: to develop Hebrew as a spoken language and as the foundation of a Jewish national consensus; to transfer to Palestine all Jews who could not or did not wish to live in diaspora countries; to establish a Jewish community in Palestine free from the peculiar social, economic, and cultural problems that beset the Jewish status as a minority people scattered throughout the world; and to carry out the transformations in the Jewish social and economic distribution, to create the appropriate social institutions, and to foster the cultural changes that were the necessary means for attaining the above ends. By World War II, the Jewish community in Palestine had indeed secured a socioeconomic infrastructure in conformity with the ideal of a self-sustaining national home and developed institutions capable of realizing its political aims. With the creation of the State of Israel in 1948, the ideal of Jewish national independence was institutionalized in its ultimate form: Jewish political sovereignty.

See also: Ethnic and Racial Social Movements; Ethnic Groups/Ethnicity: Historical Aspects; Ethnicity, Sociology of; Historiography and Historical Thought: Jewish Tradition; Judaism; Judaism and Gender; Nationalism: General; Religion: Nationalism and Identity; Religious Nationalism and the Secular State: Cultural Concerns; Social Movements, Sociology of

Bibliography

Bauer Y 1970 *Flight and Rescue: Brichah*. Random House, New York

Eisenstadt S N 1985 *The Transformation of Israeli Society*. George Weidenfeld and Nicolson, London

Frankel J 1984 *Prophecy and Politics: Socialism, Nationalism, and the Russian Jews, 1862–1917*. Cambridge University Press, Cambridge

Friesel E 1972 Migrations. *Encyclopaedia Judaica* **16**: 1520–4

Friesel E 1990 *Atlas of Modern Jewish History*. Oxford University Press, Oxford, UK, pp. 118–19

Halpern B 1987 *A Clash of Heroes: Brandeis, Weizmann, and American Zionism*. Oxford University Press, Oxford, UK

Halpern B 1969 *The Idea of the Jewish State*, 2nd edn. Harvard University Press, Cambridge, MA

Halpern B, Reinharz J 2000 *Zionism and the Creation of a New Society*. University Press of New England, Hanover and London

Hersch L 1948 The Jewish population in Palestine. In: Hersch L (ed.) *The Jewish People: Past and Present*. Marstin Press, New York, Vol. 2, pp. 40–4
Kedourie E 1985 *Nationalism*. Hutchinson, London
Laqueur W 1992 *A History of Zionism*, 2nd edn. Holt, Rinehart and Winston, New York
Medding P Y 1990 *The Founding of Israeli Democracy, 1948–1967*. Oxford University Press, New York
Ofer D 1990 *Escaping the Holocaust: Illegal Immigration to the Land of Israel, 1939–1944*. Oxford University Press, Oxford, UK
Penslar D 1991 *Zionism and Technocracy: The Engineering of Jewish Settlement in Palestine, 1870–1918*. Indiana University Press, Bloomington, IN
Raider M A 1998 *The Emergence of American Zionism*. New York University Press, New York
Raider M A, Sarna J D, Zweig R W 1997 *Abba Hillel Silver and American Zionism*. Frank Cass, London
Ravitzky A 1996 *Messianism, Zionism, and Jewish Religious Radicalism*. University of Chicago Press, Chicago
Reinharz J 1985 *Chaim Weizmann: The Making of a Zionist Leader*. Oxford University Press, New York, Vol. 2
Reinharz J 1993 *Chaim Weizmann: The Making of a Statesman*. Oxford University Press, Oxford, UK, Vol. 2
Reinharz J, Shapira A (eds.) 1986 *Essential Papers on Zionism*. New York University Press, New York
Sachar H M 1976 *A History of Israel: From the Rise of Zionism to Our Time*. Alfred A. Knopf, New York, pp. 282–3
Schweid E 1985 *The Land of Israel: National Home or Land of Destiny*. Associated University Presses, London
Shapira A 1992 *Land and Power: The Zionist Resort to Force, 1881–1948*. Oxford University Press, New York
Shavit Y 1988 *Jabotinsky and the Revisionist Movement, 1925–48*. Frank Cass, London
Shmuel A, Jehuda R, Shapira A 1998 *Zionism and Religion*. University Press of New England, Hanover and London
Shimoni G 1995 *The Zionist Ideology*. University Press of New England, Hanover
Smith A D 1983 *Theories of Nationalism*. Holmes & Meier Publishers, New York
Stein K W 1985 *The Land Question in Palestine, 1917–1939*. University of North Carolina Press, Chapel Hill, NC
Teveth S 1987 *Ben-Gurion: The Burning Ground, 1886–1948*. Houghton Mifflin Company, Boston
Vital D 1975 *The Origins of Zionism*. Oxford University Press, Oxford, UK
Vital D 1982 *Zionism: The Formative Years*. Oxford University Press, Oxford, UK
Yahil L 1987 *The Holocaust*. Oxford University Press, Oxford, UK
Zipperstein S J 1993 *Elusive Prophet: Ahad Haam and the Origins of Zionism*. University of California Press, Berekeley, CA

J. Reinharz and M. A. Raider

Zooarchaeology

Zooarchaeology is the subdiscipline of archaeology that focuses on the analysis of animal remains from archaeological sites. The term has a less euphonious twin, archaeozoology, which has less currency in English than it does in other languages. While some have distinguished among the kinds of approaches that might be treated under these two terms (Reitz and Wing 1999), research identical in form and content is published under both names.

A thorough history of zooarchaeological research has yet to be written. It would be a daunting task to do so, since highly relevant literature in a diverse variety of languages extends well back into the nineteenth century. Zooarchaeology has continued to be an inherent part of archaeological research for well over a century for a wide variety of reasons: faunal data can help place archaeological sites in time, provide detailed information on past human diets, provide detailed knowledge of the environments in which past peoples lived and on human impacts on those environments, and, less securely, provide insights into the nature of the social organization of the people involved.

Even though zooarchaeology has always been an integral part of prehistoric archaeology, zooarchaeology as a distinct and recognizable subdiscipline extends no further back than to about 1970. Prior to that time, most research dealing with animal remains from archaeological sites was conducted by people whose primary training and interests lay outside of archaeology itself, usually within zoology or paleontology. The research done by these practitioners was frequently of superb quality and enduring value, but was often poorly-integrated with other aspects of the archaeological research involved.

This situation began to change in the mid-1960s, as archaeology as a whole began to shed its largely descriptive roots and turned toward more explicitly scientific approaches to the past. Within a decade or so, there came into existence a set of scientists with intensive training in both biology and archaeology whose research focused on zooarchaeological matters, and who began developing techniques, methods, and theory relevant to answering those questions (e.g., Grayson 1984, Klein and Cruz-Uribe 1984). Today, most well-regarded archaeology Ph.D. programs in the USA now include zooarchaeology as part of their training program.

Compared with other object-based research realms in archaeology, including lithic and ceramic analysis, zooarchaeology benefits tremendously from the fact that the units studied are provided by biology, and that their biological meaning is never in doubt. These units tend to be of two prime sorts. First, there are taxonomic units, drawn from the Linneaen system—species, genera, and families, for instance. Second, there are anatomical units—the separate parts of the vertebrate skeleton, for instance, or of the insect exoskeleton. As a result, while zooarchaeologists might argue over the identification of a particular specimen, they do not argue about the meaning of their classification systems. This is often not the case in lithic or ceramic analysis—as anyone who has worked

with European Neanderthal stone tool assemblages or with ceramic Poverty Point objects in southeastern USA can attest.

1. Vertical Applications

Much zoological research is directed toward understanding change through time, and relies heavily on the analysis of data from stratified archaeological sites or from sites which, if not stratified, can be arrayed in time. As a result, such research has a strong temporal, or vertical, component, and can differ in approach from zooarchaeological research that focuses on a particular site or set of sites at a precise point in time, and is thus primarily spatial or horizontal in emphasis. These different approaches may be driven largely by the research problems that intrigue a particular investigator, but they may also be driven by the kinds of sites available for analysis. Because vertical zooarchaeological research tends to focus on change through time, it often draws more heavily on biological theory than do horizontally-oriented analyses, which tend to depend more heavily on less-well developed social theory. Even though this is the case, vertical analyses often have horizontal components, horizontal analyses often have vertical components, and either approach can make use of either kind of theory.

In recent years, vertical zooarchaeological research has revolutionized our understanding of the relationship between prehistoric peoples and their biotic environments. It has done this by showing the impacts that even small-scale societies can have on their faunal landscapes, and by showing that those impacts can force changes in the behavior of the people themselves.

Those impacts can be dramatic. For instance, it has long been clear that people caused the extinction of some 11 species of large, flightless moas in New Zealand within a few hundred years of the human colonization of these islands some 1,000 years ago. While it is not known what combination of human hunting, habitat alteration (largely by burning), and human-introduced predators (rats and dogs) caused moa extinction, human activities were certainly to blame (Anderson 1989).

During the past two decades, however, zooarchaeological work throughout Polynesia has shown that bird extinctions followed human arrival in island settings with deadly consistency. While large, flightless birds were particularly vulnerable, smaller, flighted species were lost in high numbers as well (Martin and Steadman 1999), and over half the species that comprised the Polynesian avifauna may have been lost after the initial human colonization of these islands. As is the case with New Zealand's moas, the exact causes of each extinction is unknown, but some combination of human hunting, human-induced vegetational change, and human-introduced predators was involved (Kirch and Hunt, (eds.), 1997).

Zooarchaeological research has detected impacts of this sort, though not necessarily of this magnitude, in other parts of the world as well. In some of these places, the branch of evolutionary ecology known as foraging theory has been used to understand not only why particular species declined in abundance under conditions of human predation, but also the impact of such declines on the people who caused them.

Among other things, foraging theory is designed to answer two apparently simple questions. First, given that resources are not often distributed randomly across the landscape but are instead clumped into patches (like ducks at oases), which of those patches will a predator chose to visit and how long should it stay there? Second, given that multiple prey types, or species, exist in a given patch, which of these prey types should a predator bother to pursue upon encountering it? The former issues are treated with what is called the patch-choice model, the latter with what is called the prey-choice or diet-breadth model, but combinations of the two also exist (Smith and Winterhalder 1992, Stephens and Krebs 1986). Developed in biology for non-human predators, foraging theory has seen wide application in contemporary human contexts, with impressive successes (Smith and Winterhalder 2000).

In recent years, it has begun to be applied to archaeological contexts as well. These applications have assumed that human prey choice is driven by energy returns, and that up to a point, larger animals provide higher returns than smaller ones. From this, it follows that vertical analyses of archaeological faunal assemblages should show that these animals face the steepest decline in the face of human predation. Zooarchaeological research has shown that they do just that.

In the San Francisco Bay area of California, for instance, Broughton (1999) has shown that such large-bodied animals as deer formed an ever-decreasing part of human diets between 2,600 and 2,000 years ago, while such smaller and more costly animals as sea otters became increasingly important. After 2,000 years ago, this pattern changed, and deer began to become more common in the diet—but only because sea otters themselves had become rarer, and it now paid to go further afield to hunt deer. As all this was happening, sturgeon, the largest fish available in the area, not only declined in abundance through time, but even declined in size. All this is predicted by biological theory under conditions of heavy predation.

In fact, it seems that wherever zooarchaeologists look for such effects, and have access to sufficiently fine-grained data, they find them. The moas provide another example. It has been realized for many years that after moas became extinct, human diets in New Zealand became dominated by small prey species, including fish and shellfish. Recently, Nagaoka (2000) has used foraging theory to provide the details of this process.

Foraging theory predicts that, as prey items that provide high energy returns decrease in abundance, the number of species included in the diet may increase in response to the resultant energetic shortfall. Using data from the Shag Mouth archaeological site on the east coast of New Zealand's South Island, Nagaoka has shown that, as the abundance of moas decreased through time here, not only did higher cost, smaller vertebrates become more important in the diet, but the number of species included in the diet increased as well. She has also shown that as moa populations dwindled, the places people went to procure food changed substantially. As time went on, for instance, people began to bring only the meatiest parts of moas back to Shag Mouth, suggesting that they had to go further and further to hunt those moas that still existed. In addition, during the earliest occupations of the site, most prey species came from either inland or coastal settings. Only late in the history of the site did people begin to exploit offshore resources, mainly barracuda, heavily, and this only after the human activities had caused depletion of larger vertebrates available nearer-by. All this occurred within a period of less 200 years, between about AD 1300 and AD 1500.

Zooarchaeologists have thus shown some of the kinds of impacts that even small-scale societies can have on the faunal landscape, and the influence that those impacts can in turn have on the people themselves. These kinds of theory-driven studies are fairly new. More traditional are zooarchaeological analyses that show the influence that 'natural' (i.e., non-anthropogenic) environmental changes can have on human adaptations, but these make far heavier use of, and greater contributions to, biological theory than they did in even the fairly recent past.

For instance, the site of Grotte XVI in southwestern France contains a series of archaeological deposits that date to between 36,000 and 12,000 years ago. Zooarchaeological analyses of the large mammal bones from these deposits showed that the percentage of reindeer that people hunted increased steadily through time, from 47 percent of all the large mammals at about 36,000 years ago, to 94 percent of those mammals at about 12,000 years ago. Comparing the dated archaeological record at Grotte XVI with a reconstructed summer temperature curve for southern France demonstrates that each of these increases was associated with a decrease in summer temperature (Grayson et al. 2001), culminating in a human diet whose large mammal component was almost entirely reindeer. Biogeographic theory shows that the challenges posed to mammals by temperature extremes, either hot or cold, can lead to mammalian communities that are dominated by a small number of species, and Grotte XVI, as well as other sites, show this process in operation through time.

Thus, vertical zooarchaeological analyses have begun to show the complexities of human relationships with their faunal landscapes. In some cases, people caused the abundances of the species on which they preyed to decline dramatically, and both people and prey paid the price. In others, people were forced to react to environmental changes for which they were not responsible, and the results can be readily seen in their diets, and probably in other aspects of their lives as well.

2. *Horizontal Applications*

Horizontal approaches to zooarchaeology focus on fine-scale analyses of faunal material across space within, and occasionally between, sites, though other aspects of faunal data play a role here as well. These distributions are usually used to test hypotheses about social organization, hunting patterns, or other aspects of past human behavior, or to infer aspects of such behavior from the distributions themselves.

Ongoing work at the ca. 12,500 year-old site of Verberie, located in the Oise River Valley north of Paris, provides a simple yet compelling example of such an analysis. Unlike Grotte XVI, which is a cave, Verberie is an open-air site consisting of a series of thin depositional layers, each of which represents a distinct occupation of short duration. These occupations vary in extent from 45 to 250 square meters, and are marked by stone-lined fireplaces, around which stone artifacts are concentrated. The site is also quite rich in well-preserved bone, most of which is of reindeer (Audouze et al. 1981, Audouze and Enloe 1997, Enloe and Audouze 1997). The Verberie bones are distributed across each occupational layer, but many are found concentrated around relatively bone free, circular zones. These concentrations tend to be marked by parts of the reindeer skeleton that do not have much meat associated with them in the living animal, suggesting that they had been left behind after butchering and removal of the meatier parts. In addition, these concentrations are located several meters from the hearths, which are in turn located several meters from one another. Taken as a whole, however, the occupation floors at Verberie contain all parts of the caribou skeleton. Since reindeer are large animals that are not likely to have been transported far when whole, the Verberie occupations would appear to represent hunting camps.

One of the most powerful tools zooarchaeologists have available for studying social organization is provided by what is known as refitting. There are two sorts of bone refitting, the first of which involves reassembling broken bones. Refits of this sort can tell us about the stratigraphic integrity of a site, since fragments of the same bone found in different strata demonstrate that specimens have moved up or down the deposits. They can also provide much more precise information as to the nature of the faunal assemblage as a whole, since small shaft fragments often cannot be

identified to species on their own, but can be identified once reassembled (e.g., Marean and Frey 1997, Marean and Kim 1998).

In the second approach, developed by Enloe (1991), detailed measurements and morphological comparisons allow bones that are highly likely to have come from the same individual animal to be refit to one another. If such refit body parts are found on spatially disparate parts of an undisturbed occupation surface, they have the potential of telling us about such things as food-sharing and other aspects of past social organization.

Bone refitting of this second sort done by Enloe at Verberie shows that bones belonging to the same animal either did not travel very far from the areas in which they are concentrated (and at which initial butchering apparently occurred), or, if they did, travelled to, but not between, the hearths. This situation is quite different from that found at the roughly contemporaneous site of Pincevent, located not far from Verberie. Here, bones traveled from hearth to hearth, suggesting food sharing among the people using those hearths (Enloe 1991). All this, along with a wealth of other evidence from the site, suggests that Verberie, unlike Pincevent, served as a briefly-occupied hunting camp—a place from which hunting was conducted and to which the results of the hunt were brought for processing. Enloe was also able to use tooth eruption data to show that in all likelihood the Verberie reindeer had been killed during the Fall, and thus presumably during migration.

The work done at Verberie provides a remarkably detailed view of what happened at this particular place at about 12,500 years ago. Small groups of people came here to hunt reindeer during autumn. Hearths were constructed and stone tools fashioned or refashioned around those hearths; animals were hunted nearby, then brought back to be butchered, with the meat transported elsewhere.

Not all horizontal archaeology is as firmly-grounded at that which has been done at Verberie and Pincevent. A good deal of it, in fact, is hardly compelling at all, consisting largely of poorly-controlled assertions about the behavioral meaning of a particular set of faunal material. However, when horizontal zooarchaeology is well-done, as at Verberie and Pincevent, the degree of detail it can provide about ancient human behavior can be close to astonishing.

3. Taphonomy

The differences between vertical and horizontal approaches to zooarchaeology are often differences in degree, not kind. Multiple horizontal analyses at a single stratified site can be combined to look at change through time, and many vertical analyses examine the distribution of specimens across space as well.

Both approaches require that we understand the nonhuman factors that may have impacted the faunal material that is under study. Carnivores, for instance, can introduce bones into sites occupied by people, and can modify both the condition and the location of bones deposited by people. Specimens can be destroyed by a roof-fall in a cave, or by chemical action anywhere; they can be moved by water, trampled by large animals (including people), attacked by roots, and in general be affected by a depressingly broad variety of processes that can dramatically transform the assemblage that was originally deposited. Because of this, the area known as taphonomy, the study of the processes that intervene between the time an animal dies and scientists study what is left of it in the lab, has become extremely important in zooarchaeology (Lyman 1994). Most zooarchaeologists are proficient at conducting taphonomic analyses. Without such analyses, there can be a high risk of interpreting as archaeologically meaningful patterns that have been produced in other ways.

In short, the thriving subdiscipline of zooarchaeology provides us with information that can be acquired in no other way on human diet, on human interactions with the environment, on those environments themselves, and on a surprising variety of past human behaviors that are themselves invisible but that leave faunal residues from which they can now at times be fairly confidently inferred.

See also: Bioarchaeology; Environmental Archaeology; Food Production, Origins of

Bibliography

Anderson A 1989 *Prodigious Birds*. Cambridge University Press, Cambridge

Audouze F, Cahen D, Keeley L H, Schmider B 1981 Le site Magdalénien du Buisson Campin à Verberie (Oise). *Gallia Préhistoire* **24**: 99–143

Audouze F, Enloe J G 1997 High resolution archaeology at Verberie: Limits and interpretations. *World Archaeology* **29**: 195–207

Broughton J M 1999 Resource depression and intensification during the late Holocene, San Francisco Bay. *Anthropological Records* **32**

Enloe J G 1991 Subsistence Organization in the Upper Paleolithic: Carcass Refitting and Food Sharing at Pincevent. Ph.D. Dissertation, University of New Mexico

Enloe J G, Audouze F 1997 Le Rôle de l'environnement dans la vie des chasseurs Magdaléniens du Bassin parisien. In: J P Fagnart, Thevenin A (eds.) *Le Tardiglaciare en Europe de Nord-Ouest*. Éditions du Comité des Travaux Historiques et Scientifiques, Paris, pp. 177–86

Grayson, D K 1984 *Quantitative Zooarchaeology*. Academic Press, Orlando, IN

Grayson D K, Delpech F, Rigand J-P, Simek J 2001 Explaining the development of dietary dominance by a single ungulate taxon at Grotte XVI, Dordogne, France. *Journal of Archaeological Science* **28**: 115–25

Kirch P V, Hunt T L (eds.) 1997 *Historical Ecology in the Pacific Islands*. Yale University Press, New Haven

Klein R G, Cruz-Uribe K 1984 *The Analysis of Animal Bones from Archaeological Sites*. University of Chicago Press, Chicago

Lyman R L 1994 *Vertebrate Taphonomy*. Cambridge University Press, Cambridge, UK

Marean C W, Frey C J 1997 Animal bones from caves to cities: Reverse utility curves as methodological artifacts. *American Antiquity* **62**: 698–716

Marean C W, Kim S Y 1998 Mousterian large mammal remains from Kobeh Cave: Behavioral implications for Neanderthals and early modern humans. *Current Anthropology* **39**: S79–S113

Martin P S, Steadman D W 1999 Prehistoric extinctions on islands and continents. In: R H MacPhee (ed.) *Extinctions in Near Time*. Kluwer Academic/Plenum Publishers, New York

Nagaoka L A 2000 Resource Depression, Extinction, and Subsistence Change in Prehistoric Southern New Zealand. Ph.D. Dissertation, University of Washington, WA

Reitz E J, Wing E S 1999 *Zooarchaeology*. Cambridge University Press, Cambridge, UK

Smith E A, Winterhalder B 1992 *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, New York

Stephens D W, Krebs J R 1986 *Foraging Theory*. Princeton University Press, Princeton, NJ

Winterhalder B, Smith E A 2000 Analyzing adaptive strategies: Human behavioral ecology at twenty-five years. *Evolutionary Anthropology* **9**: 51–72

D. K. Grayson

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7